VOLUME 10, NUMBER 3
MARCH 5, 1996

MICROPROCESSOR REPORT

THE INSIDERS' GUIDE TO MICROPROCESSOR HARDWARE

Intel's MMX Speeds Multimedia

Instruction-Set Extensions to Aid Audio, Video, and Speech

By Linley Gwennap

The first major extension to the x86 instruction set since 1985 will greatly improve the venerable architecture's handling of emerging multimedia applications. Collectively known as MMX, these 57 new instructions accelerate calculations common in audio, 2D and 3D graphics, video, speech synthesis and recognition, and data communications algorithms by as much as 8×. Overall, users will see a 50–100% performance improvement or more on these types of programs when using MMX instructions, as Figure 1 shows.

Intel plans to implement MMX throughout its product line in 1997. The first instantiation of MMX will be the P55C, a Pentium derivative due in 4Q96. MMX will also be included in Klamath (see page 3), a cost-reduced Pentium Pro that we expect to debut in 1H97. By the end of 1997, these two devices (and their successors) will displace most or all of Intel's non-MMX processors. AMD plans to incorporate MMX in its future processors (see MPR 1/22/96, p. 4), and we expect Cyrix will follow suit.

MMX is designed to have no impact on the operating system, making it compatible with existing x86-based OSs. Applications can take advantage of MMX in two ways: either by calling MMX-enabled drivers, such as a graphics driver, or by adding MMX instructions to critical routines. Most applications will take the driver route.

Intel has been working with dozens of key software and hardware vendors for months to help them add MMX to their applications and drivers. Today's public disclosure of the instruction set will enable any programmer to begin recoding their software for the new instructions. The number of MMX-enabled drivers and applications will build quickly once P55C systems are released.

These new instructions will provide PC users with a highly visible performance boost on many of today's most performance-critical applications. This boost should foster increased growth in the PC market and give Intel a leg up on competitors, such as PowerPC, that are lagging in adopting similar technology.

Simple Software Model

The mark of a good organization is learning from past mistakes. In designing the MMX software model, Intel took pains to avoid creating any new modes or new user state that would complicate an already complex architecture. MMX instructions can be used in any processor mode and at any privilege level. They generate no new interrupts or exceptions. These features eliminate the need for changes in the operating system to allow use of the new instructions.

From the programmer's view, there are eight new MMX registers (MM0–MM7) along with new instructions that operate on these registers. But to avoid adding new state, these registers are mapped onto the existing floating-point registers (FP0–FP7). When a multitasking operating system (or application) executes an FSAVE instruction, as it does today to save state, the contents of MM0–MM7 are saved in place of FP0–FP7 if MMX instructions are in use.

The obvious drawback is that programs cannot use both FP and MMX instructions within the same routines, as both share the same register set. This is rarely an issue, since most programs don't use FP at all, and those that do typically

Continued on page 6

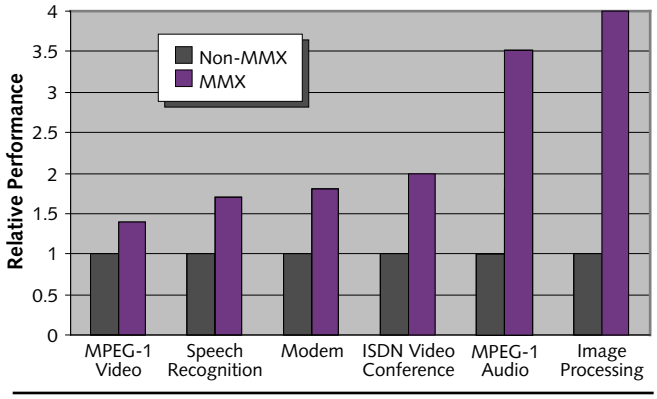


Figure 1. MMX improves performance on most multimedia applications by 50–100%, according to simulations of the forthcoming P55C processor. MPEG-1 performance refers to decoding; image processing is pixel manipulation, as in Photoshop. (Source: Intel)

AT A GLANCE

MICROPROCESSOR REPORT

Publisher and Editorial Director
Michael Slater
E-mail: mslater@mdr.zd.com

Editor in Chief
Linley Gwennap
E-mail: linley@mdr.zd.com

Senior Editor
Jim Turley
E-mail: jturley@mdr.zd.com

Senior Analyst
Yong Yao
E-mail: yyao@mdr.zd.com

Editorial Assistant: Suzanne Gifford

Editorial Board
Dennis Allison, Rich Belgard, Brian Case, Dave Epstein, John Novitsky, Bernard L. Peuto, Nick Tredennick, John F. Wakerly

Editorial Office
480 San Antonio Rd., Suite 210
Mountain View, CA 94040
Phone: 415.917.3050 **Fax:** 415.917.3093

Microprocessor Report is published every three weeks, 17 issues per year. Rates are: N. America: $495 per year, $895 for two years. Europe: £375 per year, £645 for two years. Elsewhere: $595 per year, $1,095 for two years. Additional copies in the same envelope: $175 per year in North America, $225 elsewhere. Back issues are available.

Published by
MICRODESIGN RESOURCES
President: Michael Slater

Business Office
874 Gravenstein Hwy. So., Suite 14
Sebastopol, CA 95472
Phone: 707.824.4004 **Fax:** 707.823.0504
Subscriptions: 707.824.4001
E-mail: cs@mdr.zd.com
World Wide Web: www.chipanalyst.com

Winner, Computer Press Award, 1993, 1994

Printed on recycled paper with soy ink.

THE EDITOR'S VIEW

Intel's Geography Lesson

Product "Rivermap" Includes Klamath, Deschutes, and Merced

If you have seen Intel's latest processor roadmap, you know that a river runs through it. In fact, it resembles nothing so much as a whitewater rafting itinerary, listing destinations such as Klamath, Deschutes, and Merced. These code names have replaced the traditional "Pxx" nomenclature that Intel used for a decade. Recent leaks from the blue HQ have provided insight into how the waters will flow.

The Klamath River (with bodacious Class V white water, if you like to get wet) flows across the Oregon border into California, perhaps symbolic for a processor designed at Intel's Hillsboro (Oregon) site with help from the Santa Clara (California) team. Sources indicate that Klamath is a cost-reduced version of Pentium Pro intended to bring that level of performance to mainstream PCs in 1997.

Although based on the P6 core used in Pentium Pro, Klamath will add the MMX multimedia extensions (see cover story). Sources indicate that Klamath will also gain enhancements to improve performance on 16-bit code, speeding Windows 95. Such enhancements might include a segment-descriptor cache (see MPR 7/31/95, p. 1), a Pentium feature omitted from Pentium Pro.

We believe Klamath will be a single-chip device, breaking out of the two-chip package used by Pentium Pro. This design will give OEMs the flexibility to build systems with a greater variety of external cache sizes and speeds than Intel might want to provide. To avoid a significant performance loss, Klamath will probably include larger primary caches, perhaps as big as 32K each (four times the current size but only twice the size of the P55C's caches), along with an interface to special high-speed external SRAMs. Such a design would probably deliver nearly the same PC performance as a Pentium Pro at equivalent clock speeds.

We expect Klamath to reach clock speeds of 233 and perhaps 266 MHz using Intel's 0.28-micron CMOS version of its P854 process (see MPR 7/10/95, p. 16). This would be slightly faster than the 0.35-micron BiCMOS version of the Pentium Pro, allowing Klamath to outperform Pentium Pro but at a lower cost. The first Klamath products should appear early in 1997.

Intel prefers to launch a new product in a known process, in this case P854, then use that proven design to bring up a new process. By designing Klamath in pure CMOS, Intel is paving the way for a quick shrink to its forthcoming 0.25-micron process, which does not support BiCMOS. This 0.25-micron part, according to our guides, is called Deschutes. In the real world, specifically Oregon, the Deschutes River flows north into the Columbia, splitting the state in half.

Because the 0.28-micron version of P854 uses the same metal pitches as the 0.35-micron version, the Klamath die will be a bit larger than the 196-mm^2 Pentium Pro CPU. The 0.25-micron process, however, will have smaller metal pitches and perhaps a fifth metal layer, bringing the die size of Deschutes down to around 120 mm^2, enough to allow high-volume, low-cost production.

The smaller transistors and shorter metal traces will boost Deschutes' clock speed to 300 and probably 333 MHz by 2H97. It will therefore debut at premium price points, but the processor's low cost structure will allow it to become Intel's volume product in 1998 and 1999. The smaller transistors will also allow Deschutes, probably in a low-voltage version, to succeed P55C in high-end notebooks.

The successor to Deschutes, we believe, is a muscle-bound P6 (Rogue? Umpqua? Willamette?) that we know only as the P68. We speculate this device will expand the instruction-decoding capabilities and reorder buffer of the P6 core, perhaps even adding a function unit, but will retain the basic pipeline and overall architecture of the P6, reducing design and verification time. Even with relatively minor changes, a design like this could outperform Klamath by 50% or so on a per-clock basis, albeit with a much larger die. We expect the P68 to debut in late 1997 and hold the high end throughout 1998.

Our river tour ends on the Merced, which flows from California's Yosemite National Park directly toward Intel's blue headquarters complex in Santa Clara, although it stops halfway in the San Joaquin Valley. Intel's Merced will be the first chip to implement the 64-bit architecture jointly developed with HP (see MPR 1/22/96, p. 4) and, if all goes well, should appear in late 1998. The project formerly known as P7 has occasionally been referred to as Tahoe (a lake on the California-Nevada border), but that code name is more properly applied to the whole Intel-HP relationship.

While Intel's Santa Clara team works on Merced and its Hillsboro architects focus on the P68, HP is said to be developing what would have been called the P8 under the old nomenclature. (No word on its river name.) By adding HP's resources to its two main design teams, Intel will bring to market an unprecedented number of new cores in 1997 and beyond. With Intel's key competitors only now bringing their Pentium-class chips to market, they will have to double their efforts to keep pace with Intel's aggressive 1997 plans. M

Apple Broadens Mac OS Licensing

Finally moving toward open availability of Mac OS, Apple has granted Motorola a license for its Macintosh operating system. The arrangement is unique in giving Motorola the ability to sublicense Mac OS with its PowerPC motherboards. For the first time, a company can enter the Macintosh clone business without negotiating a licensing agreement with Apple.

This announcement, coupled with the completion of the PowerPC Platform Specification, previously known as CHRP, could enable a wide variety of vendors to enter the Mac clone market. To date, Apple's insistence on controlling the types of systems and geographical areas that its licensees could sell into has scared away all but a handful of small companies. Under the new agreement, vendors can simply purchase motherboards from Motorola with a Mac OS license and sell them however they please.

One drawback is that Motorola cannot sublicense the OS outside of a PowerPC motherboard. Although some vendors may be willing to purchase boards for a low-volume entry into the Macintosh market, this provision prevents a high-volume manufacturer from building its own boards and undercutting both Motorola's and Apple's costs. Another issue: Apple continues to insist that all Mac clone vendors submit their systems to Apple for certification, which could be a more subtle way to control the clone market.

This agreement is a positive step toward increasing the overall market share of Mac OS to a viable level. To continue the momentum, Motorola must aggressively market its motherboards, and Apple must consider a further expansion of its licensing program, both to allow other vendors to build their own motherboards and to address the concern of potential clone vendors that don't want to have their products certified by their biggest competitor.

Digital First to Exceed 10 SPECint95

Putting some distance between Alpha and Pentium Pro, Digital today formally announced 366- and 400-MHz versions of its 21164 processor, already the fastest in the world. The new clock speeds are enabled by a move to 0.35-micron CMOS. The chip is the first to post a double-digit SPECint95 score: the 400-MHz version is rated at 11.2 SPECint95 and 15.2 SPECfp95 (base). Its closest competitor, the 200-MHz R10000, delivers 8.1 SPECint95 and 10.5 SPECfp95 (base).

The new device, initially called the 21164A (see MPR 10/23/95, p. 4), is not only faster than its predecessor but smaller and cooler as well. At 209 mm^2, the chip's estimated manufacturing cost is down to $190. Power dissipation for the 2.0-V part is 20 W at 400 MHz. Both of these figures are a bit high for a mainstream PC but are comparable to Pentium Pro's ratings and actually quite good for a processor at this performance level.

Both speed grades are sampling now, with volume production slated for 2Q96. Coming down a bit from its usual $3,000 price point, Digital is listing the 400-MHz part at $1,913 and the 366-MHz version at $1,602, both in quantities of 1,000. This price/performance is on par with that of other high-end microprocessors and better than some.

These clock speeds are only a starting point for the 0.35-micron part. As Digital gains experience with its new process, it expects to deliver parts as fast as 500 MHz by the end of 1996. The company is also developing its next-generation device, the 21264, which it hopes to ship by the end of 1997. These plans give Alpha a good chance at staying in its traditional performance leadership role for at least the next couple of years.

By staying in the lead, Alpha systems continue to be attractive in the most performance-sensitive markets, such as scientific and technical computing. To broaden its markets, however, Digital must offer this strong performance at a lower price and expand its software base through products such as FX!32 (see page 11). Without these changes, Alpha will dominate its niches but not surge in volume.

Mobile Pentium Heats Up to 133 MHz

Striving to keep notebook performance close to that of desktop PCs, Intel has rolled out a 133-MHz member of its Mobile Pentium family. Like the rest of the family, the chip operates with a 2.9-V supply rather than the 3.3-V supply favored by desktop Pentiums. The lower voltage reduces power dissipation, simplifying cooling and extending battery life in a portable system.

The new part, known as the Pentium-133 VRT, is rated at 3.3 W (typical), about 10% more than any other Mobile Pentium. Although Intel considers this variance to be within the "thermal envelope" of most Pentium notebooks, it could cause cooling problems for systems that can barely handle today's Mobile Pentiums, which are already hotter than other notebook processors (see page 27). The 133-MHz device carries a 1,000-piece list price of $371, a 16% premium over the comparable non-VRT part.

Intel also added a 100-MHz Pentium to its VRT lineup. The two new parts are the first Mobile Pentiums to use a 66-MHz system bus; most first-generation Pentium notebooks were unable to support this speed, sticking to 50- and 60-MHz buses. The 100-MHz part comes in two versions: a 0.6-micron part dissipating 3.0 W costs $218, while the 0.35-micron version, rated at just 2.25 W, is priced at $271. For comparison, a standard Pentium-100 sells for $198.

Coming on the heels of the Pentium-166 (see MPR 1/22/96, p. 4), the Pentium-133 VRT keeps the notebook market one step behind the mainstream desktop. Because many buyers now use their notebooks as their primary system, Intel would like to keep this gap as small as possible.

Unfortunately, notebook users are unlikely to see a further performance upgrade until the P55C debuts late this year. While faster VRT Pentiums are certainly possible, their power dissipation would not fit within the existing envelope without another voltage drop, but the lower voltage would prevent the chips from achieving a higher clock speed. The notebook market is just beginning to digest Pentium, however, so it will probably take the rest of this year for the Mobile Pentium line, which now spans 75 to 133 MHz, to fully establish itself in the notebook market.

Pentium Systems Get OverDrives

To help early Pentium buyers keep up with the latest advances, Intel has released the first of a series of Pentium OverDrive processors. Like the company's earlier 486 OverDrive parts, the Pentium OverDrive chips are retail products, but the new chips drop into existing Pentium sockets, boosting CPU performance by 60–70%. These chips are targeted at older Pentium systems at speeds up to 100 MHz.

Four parts were announced: a 125-MHz OverDrive processor for upgrading 75-MHz Pentium systems, a 150-MHz chip for Pentium-90 systems, and a 166-MHz part for Pentium-100 sockets. Intel also developed a single 120/133-MHz OverDrive chip that upgrades either 60- or 66-MHz systems. The low-end part conforms to the original P5 Pentium pinout and includes its own voltage converter. All four parts include an integral DC fan on top of the ceramic PGA package. The three 3.3-V parts run at 2.5× the rate of the external bus; the 5-V upgrade runs at 2× the bus frequency.

The new processors are compatible with nearly all Pentium systems and motherboards. Intel publishes a list of compatible systems; the company claims the major stumbling blocks for incompatible systems are lack of mechanical clearance for the fan, insufficient cooling, and a BIOS with timing-dependent code.

All Pentium OverDrive processors are packaged for retail sale and are intended as end-user upgrades. The upgrades for 60-, 66-, and 75-MHz systems are shipping now at a suggested retail price of $399, about 25% more than the 1,000-piece price for a standard Pentium-133. The upgrades for 90- and 100-MHz systems will be available in May for $499 and $679, respectively.

Intel expects to deliver a 180-MHz upgrade for Pentium-120 and -150 systems as well as a 200-MHz part for Pentium-133 and -166 systems, but not until late in 1997. Note the diminishing returns as the original CPU speed increases; Pentium-150 and -166 systems will see little improvement from these OverDrive processors.

The OverDrive product line, inaugurated with 486 upgrades, has not been wildly successful. Sales of DX4 OverDrive chips were disappointing, as 486 users swapped their motherboards for systems with 72-pin SIMMs and PCI rather than just a new CPU. As those Pentium systems begin to fall behind in performance, Intel stands a better chance of selling a second processor to those early adopters.

Netpower Moves to Pentium Pro

Joining the growing list of RISC vendors that have adopted Pentium Pro, Netpower (Santa Clara, Calif.) has announced its Calisto family of workstations based on Intel's high-end processor, starting at $5,295 for a system with no monitor. The startup, founded by veterans of MIPS Computer and other RISC vendors, had been pushing MIPS-based systems running Windows NT (see MPR 3/29/93, p. 5). Although it will continue to market MIPS boxes, Netpower says it has ceased development of new MIPS products. A key factor in the decision was Pentium Pro's ability to match the integer performance of the R10000.

In addition, the NT-on-MIPS market has not developed as expected. Silicon Graphics, the largest maker of MIPS systems and owner of MIPS Technologies, has taken a negative position toward NT, and no other MIPS vendor has the size to support NT on its own. Relatively few applications are available for the NT/MIPS combination, and the platform has clearly lost the momentum it had as the first RISC architecture to support NT.

In the past months, both Intergraph and Data General have adopted Pentium Pro as their primary platform, replacing moldy RISC architectures (Clipper and the 88000, respectively). Netpower is certainly not of the magnitude of these other vendors, but its decision indicates that the RISC/x86 pendulum continues to swing toward Intel.

Processor Modules Aid Notebook Design

MicroModule Systems (Cupertino, Calif.) has introduced a revolutionary new way to design notebook PCs. The company, which manufactures multichip modules, has defined a standard MCM pinout for these systems. In this scheme, the MCM contains the processor, secondary cache (if present), and system-logic chip set. It has two main connections, one directly to DRAM and the other to PCI.

Thus, the notebook system itself can contain standard memory chips and peripheral devices, but the processor core can be dropped in at the last minute to configure the system for the desired price/performance point. This technique would simplify notebook makers' manufacturing and inventory process and provide easy access to next-generation CPUs (e.g., P6) when they are introduced.

Other advantages of this approach are a reduction in the board space required for the processor core, which is now smaller due to the MCM packaging; the possibility of higher operating frequencies between the CPU and the cache; and improved thermal management.

MicroModule has arranged to obtain the necessary bare die: Mobile Pentium, Mobile Triton, and cache SRAM. Modules with these chips are now sampling, with production slated for June. The company also plans to offer a version with PicoPower's Vesuvius chip set, but this product is about three months behind the initial device. All the products in this family will be pin-compatible to allow easy upgrades.

Continued on page 10

MMX

Continued from page 1

use these calculations to generate data, while MMX is typically used in separate routines that display data. For 3D graphics, Intel recommends that geometry calculations remain in floating point while MMX is used to accelerate 3D rendering routines.

Without a mode bit, there is no foolproof way to prevent FP instructions from corrupting MMX data, and vice versa. Intel has taken some precautions, however, to trap the most common foolish situations. When data is loaded from memory into any MM register, it marks all the FP registers as busy, causing any subsequent FP instruction to trap. At the end of an MMX routine, the programmer must insert an EMMS instruction to restore the registers for FP use.

The converse situation is mostly covered. Taking advantage of the fact that the MMX registers are 64 bits wide while the FP registers are 80, MMX instructions always set the 16-bit exponent to NaN (not a number) while storing the result in the 64 fraction bits. Thus, although there is no way to trap an MMX instruction that is executed during a sequence of FP instructions, any subsequent FP calculation on the modified data will produce a floating-point exception. Alas, an FST instruction could store the corrupted data in memory; Intel could not find an easy way to plug this hole.

Single Instruction, Multiple Data

The new instructions, listed in Table 1, use a SIMD (single instruction, multiple data) model, operating on several values at a time. In this respect, they are similar to multimedia instructions in the Motorola 88110, HP's PA-7100LC, and Sun's UltraSparc. Using the 64-bit MMX registers, these instructions can operate on eight bytes, four words, or two double words at once, greatly increasing throughput.

Figure 2 shows the three new data types: packed byte, packed word, and packed double word. (RISC devotees should note that Intel words are 16 bits, and double words are 32 bits.) These data types are particularly suited to multimedia because many algorithms work on small data sizes.

For example, audio data is usually stored in 8-, 12-, or 16-bit samples; the average person cannot appreciate further precision. Video is represented in pixels, commonly encoded as RGB (red, green, blue) triplets. Each of the three color values can be stored in 4, 6, or 8 bits; the last provides 16 million possible colors, more than most people can discern.

Most of the new mnemonics begin with "P" for packed; for example, PADD means packed add. The opcodes all begin with the byte 0F, as do existing long jump, set byte, and Pentium-specific instructions. MMX uses previously reserved values for the second byte (none of which is used by other x86 vendors). The next two (or more) bytes provide the two operands, using the same encodings as other x86 instructions, except the target registers are the MMX registers, not the integer registers (EAX, etc.).

For example, the MOVD and MOVQ instructions can move data to and from memory using the same multitude of addressing modes as the standard MOV instruction; they also move data from one MM register to another. The MOVD instruction can even exchange data with the integer registers. Likewise, PADD performs register-to-register or memory-to-register operations, just like the integer ADD instruction. One exception is that register-to-memory mode is not supported in MMX.

Group	Mnemonic	Opcode*	Description
Data Transfer, Pack, Unpack	MOV[D,Q]	6E/7E,6F/7F	Move [double,quad] to/from MM register
	PACKUSWB	67	Pack words into bytes with unsigned saturation
	PACKSS[WB,DW]	63,6B	Pack [words into bytes, doubles into words] with signed saturation
	PUNPCKH [BW,WD,DQ]	68,69,6A	Unpack (interleave) high-order [bytes, words, doubles] from MM register
	PUNPCKL [BW,WD,DQ]	60,61,62	Unpack (interleave) low-order [bytes, words, doubles] from MM register
Arithmetic	PADD[B,W,D]	FC,FD,FE	Packed add on [byte, word, double]
	PADDS[B,W]	EC,ED	Saturating add on [byte, word]
	PADDUS[B,W]	DC,DD	Unsigned saturating add on [byte, word]
	PSUB[B,W,D]	F8,F9,FA	Packed subtraction on [byte, word, double]
	PSUBS[B,W]	E8,E9	Saturating subtraction on [byte, word]
	PSUBUS[B,W]	D8,D9	Unsigned saturating subtraction on [byte, word]
	PMULHW	E5	Multiply packed words to get high bits of product
	PMULLW	D5	Multiply packed words to get low bits of product
	PMADDWD	F5	Multply packed words, add pairs of products
Shift	PSLL[W,D,Q]	F1/71,F2/72, F3/73†	Packed shift left logical [word, double, quad]
	PSRL[W,D,Q]	D1/71,D2/72, D3/73†	Packed shift right logical [word, double, quad]
	PSRA[W,D]	E1/71,E2/72†	Packed shift right arithmetic [word, double]
Logical	PAND	DB	Bitwise logical AND
	PANDN	DF	Bitwise logical AND NOT
	POR	EB	Bitwise logical OR
	PXOR	EF	Bitwise logical XOR
Compare	PCMPEQ[B,W,D]	74,75,76	Packed compare if equal [byte, word, double]
	PCMPGT[B,W,D]	64,65,66	Packed compare if greater than [byte, word, dbl]
Misc	EMMS	77	Empty MMX state

Table 1. Intel's MMX multimedia extensions include 57 new opcodes. Brackets indicate a set of options where only one may be chosen for a given instruction. B=byte, W=word, D=double word, Q=quad word. *All opcodes start with 0F followed by the extension byte shown here. †Opcodes with 71, 72, and 73 as the second byte end with a third byte: Dr (PSRL), Er (PSRA), or Fr (PSLL), where "r" is the first operand.

Pack and Unpack Instructions Shuffle Bytes

In many cases, byte or word data is already stored in consecutive locations in memory and thus can be operated on by the new instructions. If the data is stored as aligned 32-bit values, however, it may be necessary to rearrange it to the packed format. The PACKxxDW instruction reads two double words from memory and combines them with two double words in a register, resulting in four packed 16-bit words.

If the original item exceeds the maximum value expressible in 16 bits, it is saturated: items that are too small are set to the smallest possible value, and items that are too large are set to the largest possible value. There are two options for calculating the saturation values, signed and unsigned, depending on how the source data is expressed. Similarly, the PACKxxWB instruction converts packed word data to packed byte data.

These operations can be reversed to unpack data. The mellifluous mnemonic PUNPCKxBW converts packed bytes into packed words, zero-extending the bytes, as Figure 3(a) shows. It can also interleave two sets of byte data into word data, as Figure 3(b) demonstrates. Similarly, PUNPCKxWD and PUNPCKxDQ convert packed words to double words and packed double words to quad words, respectively.

New Instructions Calculate in Parallel

Once the data is packed, calculations proceed in parallel. Each MMX calculation combines two 64-bit operands and produces a 64-bit result, so packed-byte instructions calculate eight results in parallel. Similarly, packed-word instructions generate four results, and instructions that operate on packed double words produce two results. Because most x86 instructions generate only one result at a time, this parallel calculation ability is the key to MMX's performance gains.

The performance boost is even greater on the P55C, due to the way MMX instructions are issued. The current Pentium design, while nominally two-way superscalar, executes only one FP calculation at a time and cannot pair these instructions with integer operations. The P55C allows pairing of MMX and integer instructions and can even pair MMX instructions with each other as long as they use different function units. Thus, the P55C can calculate up to 16 results (of one byte each) per cycle, helping to generate the performance gains seen previously in Figure 1.

The calculation instructions are all similar and, for the most part, straightforward. As Figure 3(c) shows, the PADDW (packed add word) instruction performs a parallel add of each of the four words in the source operand with the corresponding word in the destination operand, storing the result in the specified destination. Any carry out of a 16-bit addition is ignored. The results of the integer flags (carry, overflow, zero, etc.) are unchanged by any MMX calculation.

Both the add and subtract instructions can operate on packed bytes, words, or (in some cases) double words. Mnemonically, the suffixes B, W, or D are appended to indicate the data type. A set of logical operations (AND, AND NOT, and OR) operate on a bit-by-bit basis and thus do not require a data-type suffix. They are identical to the existing logical instructions, except that they operate on MMX registers instead of integer registers.

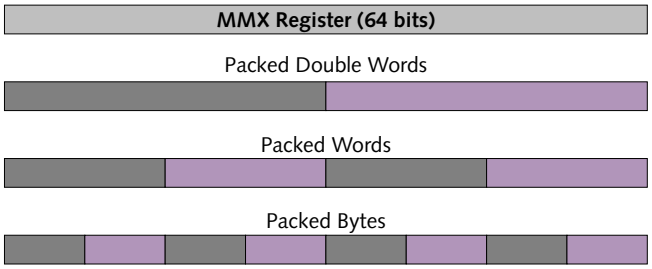


Figure 2. MMX adds three new data types: packed byte, packed word, and packed double word.

New shift instructions differ from integer shift instructions in that each packed data element is treated individually. For example, PSLL (packed shift logical left) shifts items left while filling the lower bits of each with zeroes. The logical right shift fills the upper bits with zeroes, while the arithmetic right shift inserts sign bits. The shift count can be specified by an immediate value or an MMX register. These instructions can be used to quickly multiply or divide signed and unsigned data by powers of two.

Saturating and Unsaturating Arithmetic

The add and subtract instructions have three variations. The default (no suffix) option is simple, nonsaturating arithmetic. The other two options apply saturating arithmetic; as with the saturating PACK instructions, any overflow causes the result to be "clamped" to its maximum value, and underflows set the result to the minimum value. The suffix S indi-

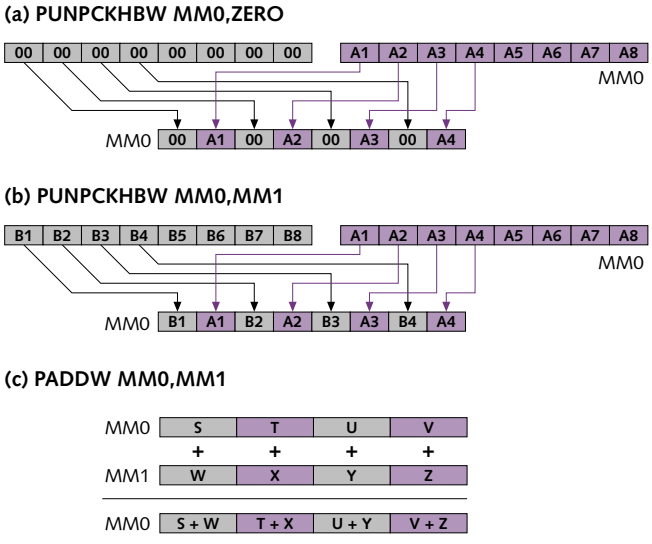


Figure 3. (a) Unpack operation converts packed bytes to words with zero extension. (b) PUNPCK can also interleave bytes from two MM registers. (c) Parallel add instruction calculates four 16-bit sums simultaneously.

cates signed saturating arithmetic; the most significant bit in each field is treated as a sign bit.

The third case is unsigned saturating (US) arithmetic, typically used for pixel operations. When two intensities, for example, are added, the result can never be whiter than white or blacker than black; saturating arithmetic handles this automatically, avoiding the long series of overflow and underflow checks needed with traditional instruction sets. In fact, a single PADDUSB instruction could replace 40 non-MMX x86 instructions.

These options are not completely orthogonal, a fact that should not surprise any x86 programmer. Although the nonsaturating form supports bytes, words, and doubles, the saturating forms cannot handle 32-bit data. The combination of the extra logic required to perform saturation with the longer carry chain of the 32-bit adder failed to meet the cycle-time requirement of the P55C.

For most situations, the saturating and nonsaturating forms are equivalent. In fact, it is dangerous to use the nonsaturating form for values that might cause an overflow, as it has no overflow trap. Of the nonsaturating adds and subtracts, only the 32-bit versions are typically used, to compensate for the lack of 32-bit saturating instructions.

Two Instructions Perform 16-bit Multiplication

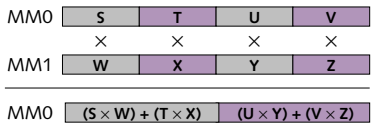
Simple multiplication is handled by PMULHW and PMULLW. These instruction operate only on 16-bit values. Because the result of a multiplication can be twice the width of its operands, PMULHW stores only the high-order word of the result in the destination register. (Of course, it generates and stores four results in parallel.) In some situations, this 16-bit result will provide adequate precision.

For full 32-bit precision, the second half of the result is generated by PMULLW. The results of the two instructions must then be combined using PUNPCKWD, which interleaves the two destination registers. The following code multiplies the four words in MM1 by the four words in MM2, storing the four products as two double words in MM1 and two double words in MM2:

```
MOVQ  MM0, MM1          ;Make copy of MM1
PMULHW  MM0, MM2        ;Calculate high bits in MM0
PMULLW  MM1, MM2        ;Calculate low bits in MM1
MOVQ  MM2, MM1          ;Make copy of low bits
PUNPCKHWD  MM1, MM0     ;Merge first two dwords
PUNPCKLWD  MM2, MM0     ;Merge second dwords
```

This code calculates four products in 6 cycles on a P55C, whereas a non-MMX Pentium requires 10 cycles to complete a single $16 \times 16 \rightarrow 32$-bit integer multiplication.

(a) PMADDWD MM0,MM1



(b) PCMPEQW MM0,MM1

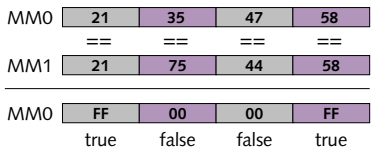


Figure 4. (a) Packed multiply-add sums two pairs of products with a single instruction. (b) Packed compare-if-equal compares packed words and generates a packed Boolean output, where all zeroes indicates false and all ones indicates true.

Multiply-Add Speeds Signal Processing

The packed multiply-add instruction differs from the other calculation instructions in that the data type of the result is different from that of the source. As Figure 4(a) shows, PMADDWD multiplies two pairs of 16-bit words, then sums each pair, producing two 32-bit results. It executes in just three cycles on a P55C and is fully pipelined.

Multiply-add is at the heart of many audio and video algorithms, such as the fast Fouriér transform (FFT). This procedure multiplies two vectors and accumulates the sum of the products. Using PMADDWD, a P55C can multiply and accumulate four vector entries per cycle (assuming the loop has been unrolled three times) while freeing the second pipe to perform loads, stores, index calculations, and branches. This is eight times the peak FP performance of a Pentium (which has no FP multiply-add instruction), not counting the advantage of executing instructions in the second pipe.

One drawback to MMX is the lack of a multiply or multiply-add for 32-bit operands. A fast 32-bit multiplier consumes four times more die area than a 16-bit multiplier, and Intel felt this feature was not worth the extra area. Besides, multiplication of 32-bit data can be performed using the standard integer multiply instruction. Although this instruction takes 10 cycles in the Pentium core and is not pipelined, it requires 4 cycles on Pentium Pro (and presumably Klamath) and, more important, is fully pipelined.

The integer multiplier, however, operates on the integer registers, not the MMX registers, and it cannot perform parallel calculations like the MMX units. Furthermore, there is no integer multiply-add instruction in x86. Because 16-bit precision is inadequate for advanced audio algorithms, such as wavetable sound, and for most 3D geometry calculations, the lack of a 32-bit multiply-add prevents these types of routines from taking advantage of MMX.

Parallel Comparisons Eliminate Branches

The MMX extensions include parallel compare operations that seem awkward at first but will produce big performance savings, particularly for Klamath and its successors. The PCMPEQW instruction, for example, compares two packed words; the fields in the result are set to zero if the comparison is false (not equal, in this case) or all ones if the comparison is true (equal), as Figure 4(b) shows.

This function is useful when combining or overlaying two images. For example, a common video technique known as chroma keying allows an object (such as the weatherman) in front of a blue screen to be superimposed on another image (such as the weather map). In a digital implementation, this technique requires combining two images such that any blue pixels in the first image are replaced by the corresponding pixels in the second image.

Assume that X[i] is the first image, Y[i] is the background image, and the result is put back into X[i]. Using traditional x86 code, a single iteration might look like this:

```
CMP  X[i], BLUE          ;Check if blue
JNE  next_pixel          ;If not, skip ahead
MOV  X[i], Y[i]          ;If blue, use second image
```

In this case, three instructions are needed per pixel.

Using MMX instructions, this sequence can be recoded as follows, assuming 16-bit pixels:

```
MOV  MM1, X[i]           ;Make a copy of X[i]
PCMPEQW  MM1, BLUE       ;Check four pixels in X[i]
PAND  Y[i], MM1          ;Zero out non-blue pixels in Y
PANDN  MM1, X[i]         ;Zero out blue pixels in X
POR  MM1, Y[i]           ;Combine two images
```

Note that this sequence assumes all pixels are in MMX registers. The compare instruction generates four results in register MM1, setting each to zero if the corresponding pixel is not blue. The PAND combines this result with Y[i], zeroing any pixels corresponding to non-blue values in X[i]. Conversely, the PANDN zeroes the blue pixels in X[i].

At first glance, this routine appears to be about 2.5× faster than the non-MMX routine, processing four pixels in five instructions. The actual performance will be even better, however, because the second routine eliminates a branch. Although modern processors predict branches, they mispredict perhaps 10–20% of the time. Pentium's misprediction penalty is 4–5 cycles, while Pentium Pro (and presumably Klamath) takes an average of 15 cycles to recover from a misprediction. Thus, eliminating branches in this way significantly improves performance.

Outshined by VIS But Ahead of Others

In many ways, MMX is quite similar to the VIS instruction set (see MPR 12/5/94, p. 16) developed by Sun for UltraSparc. Both MMX and VIS pack 8-, 16-, and 32-bit data into 64-bit registers for parallel operations, including addition, multiplication, comparisons, and logical operations. Both use the floating-point registers to store these values. Both perform saturating and unsaturating arithmetic.

To this baseline feature set, VIS adds some highly specialized instructions. For example, PDIST calculates the sum of the absolute values of the differences of two sets of eight pixels. This instruction vastly accelerates the motion estimation process in MPEG and other video-compression algorithms, allowing UltraSparc to perform real-time MPEG-1 encoding. MMX will accelerate motion estimation compared with non-MMX processors, but not as much as VIS.

For More Information

For more information on MMX, obtain the document *Intel MMX Technology Programmer's Reference Manual* by calling Intel at 800.628.8686, or access the Web at *www.intel.com/pc-supp/multimed/mmx/index.htm*, or contact your local Intel sales office.

UltraSparc also includes instructions to accelerate the discrete cosine transform (used in video decompression), pixel masking, and 3D rendering. As with other SPARC instructions, the VIS instructions use three-operand encoding rather than the two-operand MMX style, which would alleviate some of the awkwardness seen in the above MMX code examples.

The VIS instructions also operate on 32 registers instead of the limited set of 8 MMX registers. A 4 × 4 matrix of constants, commonly used in digital filters, fits within the VIS register set but requires extra memory accesses in MMX. MMX has one advantage in its multiply-add instruction; it takes two instructions to perform this task under VIS.

While MMX may not go quite as far as VIS, it provides a much wider range of multimedia-oriented instructions than any other popular instruction set. HP's recent processors include some parallel arithmetic (see MPR 1/24/94, p. 16) but operate on only two 16-bit quantities at once.

To date, the other leading desktop RISCs—PowerPC, MIPS, and Alpha—have somehow failed to implement multimedia instructions, despite the significant performance benefits and minimal cost. One advantage that PowerPC has over current Intel chips is in floating-point performance, which can be used to speed audio processing and speech recognition, for example. For these multimedia applications, MMX processors should improve Intel's position.

Both NexGen and Cyrix have been developing their own multimedia extensions to the x86 instruction set. AMD's purchase of NexGen and its subsequent licensing agreement with Intel (see MPR 1/22/96, p. 5) ensure that company's processors will move to MMX, starting with the K6. Cyrix has said its M2 processor, due in early 1997, will include its own multimedia extensions. We expect Cyrix will eventually switch to MMX, although perhaps not in the first version of the M2.

MMX to Appear Mainly in Drivers

Programming in MMX is challenging. Taking full advantage of the SIMD architecture often requires unrolling loops and carefully arranging instructions, yet there is no compiler support planned other than allowing in-line MMX assembly code. Intel plans to provide libraries of routines for common multimedia functions as well as an assembler and debugger that support MMX. Over time, third parties will also supply MMX tools and library code.

Most multimedia applications will take advantage of the new instructions simply by calling MMX-enabled drivers or including the new library routines. One advantage of relying on drivers is that an application can automatically take advantage of a hardware accelerator for 3D graphics, sound, or MPEG decoding if one is installed. This model, of course, pushes the coding effort onto the driver writers.

A few applications will have to incorporate MMX instructions directly. These include programs that do image processing (e.g., Photoshop) or speech recognition, since APIs for these tasks are not yet defined. Fortunately, a significant speedup can often be obtained by simply modifying a few critical inner loops.

One problem is managing separate versions of each application for MMX and for non-MMX systems, which will be the majority of the installed base for several years. Software can check bit 23 of the CPUID to determine if a processor implements MMX. Again, simply relying on drivers eliminates this problem for the application.

Improved Multimedia Fuels PC Sales Growth

Although the installed base of MMX processors is nonexistent today, it will grow rapidly. We project that more than half of Intel's 1997 processor shipments, and virtually all thereafter, will contain MMX, totaling more than 30 million processors by the end of 1997. The 50–100% performance gain will motivate multimedia software vendors to use MMX; those that don't will be uncompetitive. Intel expects dozens of drivers and applications to be shipping with MMX code when P55C systems first appear; this number will quickly increase during the course of 1997.

While the two are not directly connected, MMX is clearly designed to accelerate native signal processing (NSP). NSP will be used mainly in low-end systems to perform multimedia tasks; more expensive PCs are likely to include hardware accelerators. The P55C will significantly increase the baseline multimedia capabilities of low-end systems without accelerator chips. As Klamath reaches the mainstream in 1998, it will offer another big performance boost, possibly eliminating accelerator chips even in midrange systems.

It's not every day that you get a sizable step up in performance with minimal die cost, but that's what MMX promises. Once the combination of MMX-based processors and applications reaches the market, it should increase the growth of PC sales, particularly in the already hot consumer market, where multimedia is used most today. Even businesses will see the benefit as more use their PCs for videoconferencing and similar tasks.

One problem will be measuring this new performance level. Current benchmarks (SPEC95, Winstone, etc.) do not measure improvements of this type. The computer industry, with prodding from Intel, will most likely come up with new benchmarks to solve this problem, but perhaps not in time for the P55C's debut. Fortunately, the increase in graphics and video performance is something the buyer can see. M

Most Significant Bits

Continued from page 5

The technology outlined by MicroModule is compelling. The issues are the same as with other MCMs: price and availability. The company must be able to deliver these modules with only a small premium over the cost of the discrete components; the notebook market is slightly less price sensitive than the desktop market but is unwilling to pay significantly more for similar functions.

MicroModule must also demonstrate the ability to deliver these modules in volume. The first usage will probably come from small vendors or in low-volume product lines. The adoption of this technology would be hastened by the existence of a second source.

Intel is developing similar modules but has not yet committed to offering them as a product. Intel is unlikely to use MCM technology or match MicroModule's pinout. But if Intel chooses to offer processor modules, or if MicroModule can get its products into high-volume production, it could change the way notebook systems are designed.

First USB Microcontroller Comes from Intel

Intel has announced its first microcontroller with an integrated USB (universal serial bus) interface. The 82930A is a derivative of the company's MCS 251 family, the 16-bit upgrade of the venerable 80C51 architecture. The chip is intended to control intelligent USB peripherals such as keyboards and telephony products.

From a programmer's perspective, the new part is similar to other members of the 16-bit 251 family, which is upward binary-compatible with a plethora of 8-bit 80C51 microcontrollers and development tools. The 82930A includes a full USB master/slave interface, four transmit and receive FIFOs, 1K of on-chip RAM, a UART, timers, a 16-bit address bus, and an 8-bit data bus. In 10,000-unit volumes, the part is priced at $6.

The 82930A represents the first of a promised range of USB-capable microcontrollers, chip sets, and interface logic from Intel. The company's 430VX and 430HX Pentium core-logic chip sets (see MPR 2/12/96, p. 5), for example, are the first to include USB interfaces. For Pentium Pro systems, VIA's Apollo P6 chip set (see MPR 2/12/96, p. 6) also includes a USB interface, which Intel's Orion does not. We expect USB to become an integral part of most new PCs by the end of 1996, with USB-compatible keyboards, modems, and other peripherals to follow.

P-Rating Clarification

In the Chart Watch of our February 12 issue, we listed "P ratings" for PowerPC microprocessors. These are our estimates of what Pentium speed class each chip falls in, based on the vendor's SPECint95 ratings. It has no relationship to the recently announced P-rating standard (see MPR 2/12/96, p. 11) and does not reflect any testing by MDR Labs. M

Alpha Runs x86 Code with FX!32

Digital's Emulation Strategy Could Help Boost Alpha/NT System Sales

by Jim Turley

Digital is giving its Alpha NT systems the ability to execute standard Win32 application programs written for x86 systems. The new technology, dubbed Digital FX!32, is still in development but is expected to begin shipping with all Alpha Windows NT systems in 2Q96, and it will be available free of charge to all existing Alpha customers.

FX!32 uses a combination of run-time emulation and background binary translation to decipher x86 binaries and convert them to native Alpha code. Digital claims translated applications can run as fast on a midrange Alpha chip as they would on a high-end Pentium system. The company has not yet achieved its performance targets, however, nor has it demonstrated that FX!32 delivers software compatibility.

The strategy can be viewed either as proof of Alpha's performance superiority over x86 processors or as proof of the hopelessness of bucking the x86 software tidal wave. Either way, FX!32 is an interesting technical feat and a bold strategic move on Digital's part. The ability to run "shrink-wrapped" software should help boost Digital's NT system sales by allowing its customers to run common PC productivity applications. With acceptable x86 performance, the company could make its Alpha/NT platform more attractive than alternative RISC-based NT systems to buyers concerned about the availability of commercial software.

Emulation, Translation Intertwined

Digital discovered early on that it was impossible to perform a straightforward single-pass translation of x86 object code. Offering translated binaries for popular applications would have been the easiest solution for its customers, but this technique presented serious legal and technical hurdles.

For example, it is impossible to examine an x86 application and locate all the branch targets or even determine which portions are code and which are data. The problem lies with the x86's variable-length instruction encoding and its irregular use of extension words, prefixes, suffixes, and segment-override bytes. The tendency of compilers to intersperse static data among code strings also complicates attempts at automated disassembly.

Without actually executing the program, automated methods of extracting the code stream are prone to error. Products such as FlashPort from Echo Logic or Hunter Systems' Xdos sidestep this problem by relying on hand-tuned translator "hint" files developed for individual applications.

Digital hoped to avoid any kind of user intervention or manual tweaking and make FX!32 totally automatic. In the end, the company relies on the applications themselves to assist FX!32 in separating code from data.

Applications Get Off to Slow Start

The FX!32 translator requires at least two passes through each application to complete a partial translation. Rather than try to pick out the executable portions of a program, FX!32 begins by emulating x86 instructions at run time.

As Figure 1 shows, the FX!32 software consists of three component parts: the runtime emulator, the background translator, and the transparency server. The first time an x86 application is executed, the transparency server determines that the file is an untranslated x86 program and hands it off to the emulator.

The program runs the first time in emulation. No permanent x86-to-Alpha translation is performed during this phase. Instead, x86 instructions are emulated one by one. All operating system calls are handled natively. Because FX!32 works only with x86 applications written for the 32-bit Win32 API, trapping NT system calls is a simple matter. This approach is somewhat different from the method used by Insignia Solutions' SoftWindows or Apple's 68K-to-PowerPC emulator (see MPR 5/30/94 p. 12), which look at certain sequences of instructions.

The FX!32 emulator keeps track of which portions of the application it has emulated. The results are kept in a log file, which Digital calls the execution profile. This profile enables the later translation stage.

The first time an x86 application is run on an Alpha system, its performance will likely be disappointing because it is completely emulated; no optimization is performed and

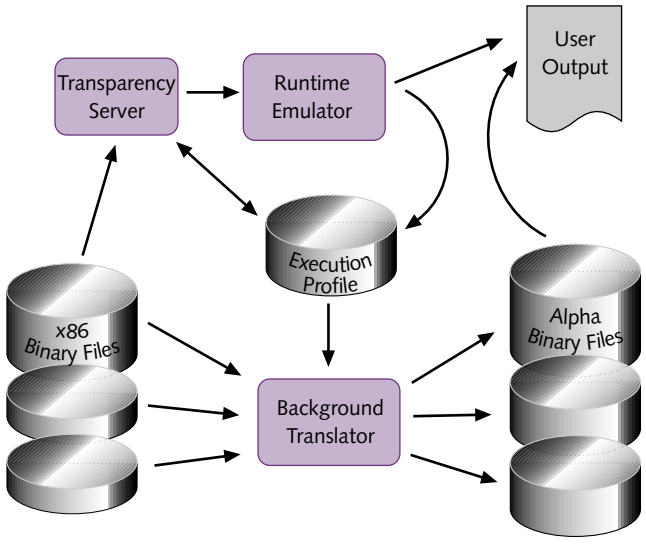


Figure 1. Digital's FX!32 emulation technology consists of a runtime-emulation stage and a later background-translation stage. The transparency server manages the executable images.

Price and Availability

The Digital FX!32 emulation/translation software will be available free of charge when it is released around mid-1996. Distribution media have not been determined. For more information, contact Digital Semiconductor (Hudson, Mass.) at 800.322.2712 x32 or 508.628.4750 x32; fax 508.626.0547.

performance may be only a small fraction of full native speed. Compute-intensive applications are particularly vulnerable; conversely, software that relies heavily on NT system calls may run reasonably well. Unfortunately, Digital is not prepared to release any performance indicators until FX!32 testing is further along.

In effect, the first pass through an application is merely a training run, intended only to help FX!32 determine which portions of the application are executable and where the branch targets are. It is not until the user exits the application for the first time that the real translation begins.

Later Passes Speed Translated Performance

Using the execution profile generated during the first run, FX!32 begins translating the application in earnest. This translation is always performed as a background process after the application has quit. Commercial x86 applications frequently include several executable (EXE) and library (DLL or LIB) files; each of these files is translated into a corresponding Alpha object file. From the point of view of the operating system, each translated file is stored as a dynamic link library (DLL) file on disk.

Only code that was actually executed during the first run will be translated. The exception to this rule is simple branches. In its current version, when the background translator encounters a two-way branch, it translates both forks of the branch even if one fork was never explored in the emulation run. However, multiway switch constructs often remain partially untranslated. Digital expects this artifact to be eliminated in the final release of FX!32.

After all previously executed portions of the software are translated and stored as DLLs, the application is ready to run again. This time, FX!32 examines the application's execution log and, seeing that certain files that make up the application have already been translated, it loads and runs the translated DLLs in place of the original x86 object files whenever possible. Assuming the second run of the application exercises exactly the same features as the first run, the application is now sufficiently translated into Alpha code to run noticeably faster. In this instance, no run-time emulation would be required; the application would run entirely from translated Alpha DLLs.

If the application veers into previously unexplored sections of code, the FX!32 server will again invoke the emulator to work through untranslated portions at run time while updating the application's execution profile. After the user exits the application, the background translation process starts afresh.

Portions of the application that are never exercised—such as little-used features the user never exploits—may never be translated. Different users might generate different binary translations of the same application because of their different usage habits.

Generally, once an x86 file has been translated the first time, it is never retranslated or updated. However, if the application's usage changes substantially—such as when a new program feature is exercised for the first time—the FX!32 server may determine that the application needs to be optimized and retranslated. A frequent consequence of retranslation is that multiway switch constructs are optimized and translated more completely.

The decision to retranslate is based on the growth of the application's execution profile. If the profile grows beyond a certain threshold after the file's initial translation, the server initiates a retranslation. Although this threshold is set to a nominal value when FX!32 is delivered, users can adjust the threshold to suit their tastes. Adjusting the threshold up or down will decrease or increase, respectively, the amount of optimization and retranslation FX!32 performs.

Only 32-Bit Windows Applications Work

All calls to the operating system are trapped and executed natively, which provides the best possible performance. Like emulated Macintosh code, the more time the application spends in system calls, the faster it runs.

FX!32 works only with Win32 applications, that is, with code written for the 32-bit programming model used by Windows NT and Windows 95. FX!32 does not translate legacy 16-bit programs written for Windows 3.x or DOS.

However, Windows NT itself includes emulation capability for many 16-bit applications. The current release of NT 3.51 for RISC processors ships with an equivalent to SoftWindows, allowing those systems to run 16-bit, standard-mode x86 applications through emulation. Future releases of NT will add support for enhanced-mode 16-bit code as well.

Between NT and FX!32, Alpha users will have access—one way or another—to most Windows applications. Because SoftWindows is an emulator, however, the performance of 16-bit software will not be up to Digital's hopes for FX!32's native translation of 32-bit applications.

Digital does not consider this to be a major drawback, and will continue to rely on Microsoft and Insignia to provide compatibility with 16-bit code. This is probably a sound strategy for Digital's market; we believe a number of factors are pushing software developers to adopt the Win32 API, steadily reducing the need to support 16-bit applications as time goes on. Intel's Pentium Pro, for instance, delivers notoriously poor performance on 16-bit code, and Microsoft itself no longer releases Win16 applications.

Modest Memory, Disk Overhead

Because translation is carried out on a file-by-file basis, FX!32 is able to take advantage of shared code libraries. That is, when multiple applications share a single library file, as is often the case with Microsoft products, FX!32 will translate the shared library once, and the translated version will then be shared among all the appropriate applications.

Obviously, the translated DLLs require additional disk space. The x86 instruction set yields very good code density—for all the same reasons that make it difficult to disassemble—while Alpha's fixed-length 32-bit instruction words result in fairly bulky binaries. Digital estimates the translated binaries may be about twice the size of their x86 originals. The exact ratio depends on a number of factors, including the ratio of code to data (data does not get translated) and the percentage of code actually exercised and converted. Note that this is in addition to the original x86 files, which are not deleted.

The memory requirements of FX!32 are relatively modest. The transparency server, which determines which applications have been translated and which haven't, uses only 140K of RAM. The run-time emulator requires about 2M of RAM, which is fairly small, considering the task at hand. SoftWindows running on a Power Macintosh, for example, requires several megabytes of RAM, although it must emulate the entire PC/Windows environment as well. When the background translator is running, it requires approximately 1.5M of memory plus enough for both the original and translated versions of the file being translated. Considering the amount of RAM Digital ships with its Alpha NT systems, FX!32 makes modest demands on the system.

Performance Still a Big Question Mark

Unfortunately, Digital has not released performance numbers for any translated applications. The company hired National Software Testing Labs (NSTL) to conduct application benchmark testing, but results are not available, casting some suspicion on the current condition of the product.

Digital believes FX!32 might someday achieve 70% of native Alpha performance on translated applications—a laudable goal, but a difficult one to gauge, since native Alpha versions do not exist for virtually any Windows applications (the whole reason for FX!32's existence). Digital cautions the 70% goal has not been achieved in the current implementation of FX!32 and offers no timeline for reaching that particular milestone.

Correlating Alpha performance to Pentium performance involves a number of variables, but a specific example can be helpful. Judging from SPEC95 results, a 21164-300 delivers 7.75 SPECint95 (base) versus 4.76 for a Pentium-166, or 63% of the 21164's performance. Thus, 70% of native Alpha performance would be slightly faster than native Pentium-166 performance. If FX!32 were to meet its performance goal, then in this instance Alpha would execute Pentium applications faster than today's fastest Pentium.

Granted, the Alpha chip is running at nearly double the clock rate of the Pentium and has a huge L2 cache. And these numbers ignore the fact that the Alpha processor, much less the entire system, costs twice as much as the Pentium. The role of FX!32 is not to displace PCs but to make Digital's Alpha machines more attractive than competing RISC-based NT systems.

It's clear that Digital won't lure away customers who are interested primarily in PC applications—that is not the company's intent. Instead, FX!32 offers a booster to the company's NT line by removing the emotional obstacle of PC compatibility. It may now be feasible to satisfy, for example, the engineer who requires access to native Alpha applications as well as to a few productivity applications that are available only for x86 processors.

So far, only Apple has pulled off a transparent binary conversion from one architecture to another. In Apple's case, the wide disparity in performance between PowerPC and the 68K made it easier to mask the inefficiencies of a run-time emulator. Pentium is a closer match for Alpha, at least in integer performance, making it tougher to run in emulation mode and still get acceptable speed. By allowing FX!32 to take its time and work its binary translation in the background, Digital has given its product a potential performance advantage over the Apple or Insignia approaches. The cost is more disk space and a certain amount of delayed gratification when running an application for the first time.

Adding FX!32 to its NT systems will not help Digital penetrate the mainstream PC market, nor will it boost the company's Unix system sales. FX!32 will appeal only to the portion of the market that is considering Windows NT. In that segment, Digital competes with systems based on PowerPC, Pentium, and Pentium Pro processors, with MIPS support fading fast. By offering better floating-point performance than the x86 systems and better x86 compatibility than the other RISC systems, Digital noses ahead on the features checklist. FX!32 won't make buyers choose NT over Unix, but it may help them choose Alpha over PowerPC.

Digital's biggest competition is coming from Pentium Pro systems. Their integer performance is as good as Alpha's, the selection of vendors is greater, prices are much lower, and x86 compatibility is a given. Finally, Pentium Pro's much-maligned 16-bit performance is not an issue here because FX!32 doesn't translate 16-bit applications.

FX!32 is an interesting technical exercise, but so far, a completely unquantified one. In its current state, FX!32 does not meet its goal of running at 70% of native Alpha speed. For some applications, such as e-mail and word processing, reaching top speed may not be a big issue. It is important that Digital demonstrate that FX!32 can work transparently with any Win32 application without side effects, bugs, or special hand waving. If the company can clear that hurdle, FX!32 will have reached its most important goal: allowing users to simply install and run shrink-wrapped PC applications on their Alpha systems. M

Motorola, TI Extend 16-Bit DSP Families

568xx, 320C2xx Girded for Battle over High-Volume Applications

by Jim Turley and Phil Lapsley

Two new DSP families from Texas Instruments and Motorola are prepared to capture a large chunk of the growing 16-bit DSP market. Almost simultaneously, the two vendors revealed plans to dramatically lower the price of midrange fixed-point DSP performance to under $10 in volume. While Motorola focuses on ease-of-use and programmability, TI provides a compatible growth path including on-chip RAM, ROM, and—for the first time—flash memory.

Motorola 56800 Borrows from 56100

Hot on the heels of its high-end 56300 announcement (see MPR 12/4/95, p. 14), Motorola revealed its new low end, the 56800 family. These new 20-MIPS DSPs are interesting primarily because of their programming model. The 56800 is a hybrid of a DSP and a traditional 16-bit microcontroller. The chips should integrate easily into control-oriented applications by providing a number of on-chip peripheral functions that microcontroller users have come to expect.

Digital signal processors have become less of a boutique item and more widespread, fostering a need to combine DSP cores with conventional microcontrollers. The combination provides true signal-processing capability with control-oriented functions normally associated with microcontrollers. The company is targeting control-oriented applications in answering machines, motor control, automotive noise suppression, and portable communicators.

The 56800 borrows heavily from its architectural predecessor, the 56100, but makes a number of changes. The two core designs share a similar overall architecture and internal data paths, but the 56800 family has fewer registers. As Figure 1 shows, the 56800 core has three 16-bit general-purpose data registers, whereas the 56100 has four, and only one address-modifier and one offset register in place of a set of eight. The 56800 enlarges its program counter to 19 bits, up from the 56100's 16 bits.

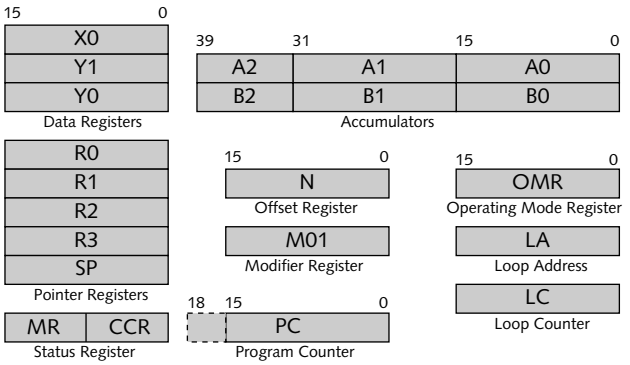


Figure 1. The 56800 retains the 56100's 40-bit accumulator registers but drops one general-purpose data register. The 56800's program counter was extended to 19 bits for a bigger code space.

New Design Appeals to CPU Programmers

By reducing and rearranging the register set, Motorola was able to encode the 56800's instructions more efficiently. One benefit programmers will applaud is the ability to reference memory-resident operands rather than adhering to a strict load/store model. The instruction set is also considerably more orthogonal, with few of the irregular register restrictions of the 56100. All in all, the instruction set looks more like that of a good 16-bit microprocessor than that of a DSP.

The 19-bit program counter aids applications that exceed 64K words of code. The most significant three bits are stored in the chip's status register, allowing a simple form of code bank-switching. The 56800 also implements a conventional software stack, a concession to microprocessor programmers and a boon for compilers.

Other changes not apparent from the programmer's model are the extension of the bit-manipulation unit to 16 bits and a 16-bit barrel shifter. Like the 56100, the 56800 has multiple on-chip data and address buses, and its data memory is dual ported, allowing it to fetch one instruction and make two data references simultaneously.

The company has teamed with software tool-developer Tartan to bring out a suite of C and, in an unusual move for a DSP architecture, C++ programming tools. Like the 56800's internal structure, the availability of a C compiler is intended to make microprocessor programmers feel more at home. Unfortunately, the software is not as far along as Motorola's chips. A beta version of the C compiler isn't due until 2Q96, with the C++ compiler and source-level debugger following by several months.

On the outside, the new 56800 devices even look like microcontrollers. The two initial devices, the 56L811 and 56L812, are nearly identical. Both chips carry a synchronous serial port, a pair of asynchronous serial-peripheral interfaces, three timers, and 32 general-purpose I/O lines. Eight of the I/O pins can interrupt the processor, giving the 56800 chips an unusual amount of interrupt flexibility for a DSP.

Both the '811 and the '812 have 16-bit external address and data buses. They differ only in the type and amount of on-chip memory: the '811 has a 1K program RAM and 2K of data RAM, while the '812 has 22K of program ROM, 2K of data ROM, and 2K of data RAM. Motorola plans more versions near the end of the year, including derivatives with flash memory.

TI Tackles Cost, Adds Flash

Like Motorola, TI created its new low-cost family by leveraging an existing design. In TI's case, the venerable 320C2x core has been updated to create the 320C2xx family. The result is source-code compatible with the 'C2x at the assembler level but offers higher performance, lower prices, and a range of new features.

TI is currently shipping the 320C203 and 'C209 at speeds of 28.6 MIPS. TI has announced plans for no fewer than six initial members of the 320C2xx family with different amounts of memory and peripheral functions; four of the chips will be pin-compatible. For ASIC development, TI is inviting high-volume customers to design around the new DSP core, which it calls the 320C2xLP. Fabricating customer ASICs around a commercial DSP core is an unusual move that could give TI a big leg up in the highest-volume markets.

Table 1 summarizes the differences among the six members of the 320C2xx family. As the table shows, all versions except the 'C209 deliver 40 MIPS peak performance, much faster than their 12.5-MIPS 'C2x predecessors. The first five chips run with a 3.3-V nominal supply, versus the 3.0-V supply used by the two Motorola devices.

Flash Memory Makes 'F206/207 Stand Out

The 320F206 and 320F207 are almost unique among DSPs in having flash memory on the chip. Even at the highest clock rates, the 32K×16 on-chip flash memory can be accessed with no wait states. The flash is also easy to use—it can be erased and reprogrammed by the CPU under software control or via the chip's JTAG test port. Either way, no special programming voltages are required, just the chip's usual supply. This rare combination of features makes the flash memory ideal for large program stores or for maintaining volatile data.

TI is nearly alone in offering flash memory in a commercial DSP. AT&T's DSP1600 family also offers flash, but at more than $1,000 per device, the AT&T chip is clearly intended only for prototyping.

The flash-based 'F206 and 'F207 are particularly interesting for the role they play in TI's new roadmap for high-volume DSPs. With four pin-compatible versions of the family available, TI customers can start their designs with the RAM-based 'C203 or 'C205, migrate to the ROM-based 'C204 or flash-based 'F206 as volumes increase, and finally end up with a custom ASIC based on the 'C2xLP core. Such a clearly defined series of stepping stones from prototype to high volume should be an important advantage for TI and its customers.

Prices Designed to Attract New Customers

For both vendors, the prices of these new 16-bit families are as important as the chips' technical features. For example, TI is offering the 'C203 at 28.6 MIPS for only $6.85 in 1,000-unit quantities; Motorola will say only that its new DSPs are priced at "less than $10" in 100,000-piece lots. Both prices set new lows for 16-bit DSPs with respectable performance.

Price and Availability

Motorola's first 56800-family device, the 56L811, is sampling now. Production is scheduled to begin in 2Q96; pricing will be less than $10 in quantities of 100,000. The 56L812 will begin sampling in 3Q96. For more information, contact Motorola (Austin, Texas) at 800.845.6686; fax 512.891.3877; or via the Motorola DSP Web site at *motserv.indirect.com/dsp/dsphome.html*.

Texas Instruments' 320C203 and 'C209 are currently in production at 28.6 MIPS; availability of the remaining parts is listed in the table below. For more information, contact TI (Denver, Col.) at 800.477.8924 x4500; fax (outside North America) 303.294.3747; or via the Web at *www.ti.com/dsps*.

Clearly, both companies are gunning for the same high-volume customers, seeking designs in disk drives, consumer electronics, telecommunications equipment, and the like.

Motorola's hardware developments are on roughly the same schedule as TI's, with chips already sampling and production slated to begin at midyear. The company has also put a lot of effort into making the 56800 series appealing to microcontroller and first-time DSP users, an interesting tactic. However, with only two products announced, the 56800 roadmap is not yet clear to us. Although the high-volume customers Motorola is courting will certainly be interested in a clearer roadmap, the current family should satisfy many vendors' immediate needs.

TI's product family seems to be more fully realized, although some members are a bit further off. The 320C2xx family offers a compelling price/performance proposition and a clear and attractive growth path. Both product families are undoubtedly headed for fierce competition in the coming months, promising an interesting range of low-cost, DSP-based consumer items to come. M

Phil Lapsley of Berkeley Design Technology (Fremont, Calif.) is the editor of Buyer's Guide to DSP Processors, *available through MicroDesign Resources by calling 707.824.4004.*

	C203	C204	C205	F206	F207	C209
Speed (MIPS)	40	40	40	40	40	28.6
RAM (words)	544	544	4.5K	4.5K	4.5K	4.5K
ROM (words)	None	4K	None	None	None	4K
Flash (words)	None	None	None	32K	32K	None
Sync Serial	1	1	1	1	2	None
Async Serial	1	1	1	1	1	None
Package	TQFP	TQFP	TQFP	TQFP	TQFP	TQFP
Pins	100	100	100	100	144	80
Samples	2Q96	2Q96	3Q96	2Q96	4Q96	Now
Production	3Q96	3Q96	4Q96	4Q96	1Q97	Now

Table 1. The six initial members of TI's 320C2xx family are nearly identical, and four members are pin-compatible. The chips differ mainly in the type and amount of on-chip memory they contain.

Competition Heats Up in 3D Accelerators

Market Hungers for a Unified Multimedia Platform from Microsoft

by Yong Yao

This is the second of two articles on PC three-dimensional graphics. The first part (see MPR 1/22/96, p. 14) provides background for understanding 3D graphics. This article discusses 3D products from 3Dfx, 3Dlabs, Chromatic, Nvidia, Oak, Rendition, S3, Trident, and VideoLogic.

With the PC market moving quickly to 3D graphics, a torrent of new products has appeared to improve performance on 3D software. Although 3D accelerators have been common for years in workstations, the new products share a common goal of bringing high-performance 3D to PCs without adding significant cost. As this market is just emerging, the new chips display a variety of technical and business models for achieving this goal. The key business challenge is to pull together the chip design, manufacturing, and sales.

From the technology side, 3D chip makers must be concerned about a variety of issues. Choosing the right DRAM technology is critical to achieving the required bandwidth. Overall rendering speed must exceed the performance of the host CPU and be competitive with other accelerators; performance should also scale well as the host processor or memory system is upgraded. To meet the required price point, roughly $250 for an add-in card, vendors must consider combining the frame buffer, texture buffer, and Z-buffer. Another cost-reduction tactic is to combine 3D acceleration with other standard functions, such as 2D and video acceleration and possibly audio.

The key issue remaining is software. There is no 3D software standard for PCs today. Some of the early players, such as Creative Labs, are trying to establish proprietary software schemes, and any developer selling chips today has to seek ISVs to write applications specifically for its chips. In short, vendors must have a plan to attract applications. For most, this plan is to develop drivers for Microsoft's Direct3D API and hope that Microsoft ships it soon and that ISVs code to this emerging standard.

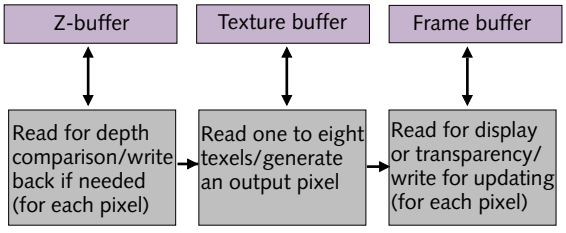


Figure 1. Data flow in a simplified rendering pipeline for 3D hardware accelerators in most of today's graphics workstations.

There is an urgent need for a unified multimedia software platform. OpenGL is a great 3D API, but it carries some overhead and runs slowly under Windows 95. Microsoft does not have a plan to fine-tune OpenGL for Windows. Although Silicon Graphics is working on its strategy for pushing the OpenGL technology to PCs, Direct3D is likely to prevent OpenGL from becoming the main 3D API for the PC. The overall multimedia software platform will be Microsoft's DirectX, which includes Direct3D. The problem is that Direct3D is late, and it is not clear how soon accelerator vendors will have drivers for it. Until we see whether Direct3D meets its goals, sales of PCs with 3D accelerators, and the number of 3D titles for PCs, will be limited.

Hardware Accelerators Share a Common Goal

Adding 3D capability to a PC is not just an incremental increase in functionality. Like the move from a text-only interface to today's graphical user interface (GUI), 3D is a feature that will enable a new set of applications. These applications include 3D animation, virtual reality, architectural walk-through, mechanical design, and, of course, games. Using 3D offers viewers more freedom, enabling them to go around and behind objects, much as in real life. But 3D graphics is an order of magnitude more complex and compute intensive than 2D graphics.

Although none of the 3D vendors covered are taking the same approach toward 3D hardware implementation, all share the same goal: to deliver Jurassic Park–like 3D performance while meeting PC price points. The design trade-off is between affordability and realism.

There are many ways to reduce the cost of a traditional workstation 3D subsystem without significantly impacting its graphics performance. For instance, Figure 1 illustrates the data flow in a simplified rendering pipeline for a typical workstation 3D accelerator. When performing texture-mapped 3D graphics, memory can be allocated in each of the three buffers: the frame buffer, texture buffer, and Z-buffer. For decent 3D graphics, each of these buffers must be 2M or larger. Even using commodity DRAM, these three buffers alone will cost more than $160, the cost of an entire 3D add-in card for the retail PC market, preventing this approach from being effective in the PC market.

Simply getting rid of these three buffers or combining them without careful design will not meet the 3D performance requirements. Products like those from Oak and VideoLogic address the high cost of memory by eliminating the need for an external Z-buffer without losing too much performance. Other products discussed below innovate in different ways to achieve the common goal.

Features in Common or for Differentiation

The products discussed in this article all have a glueless PCI interface and share the following 3D features:

- Alpha blending: zero, one, and source alpha
- Filtering: point-sampled and bilinear filtering
- Primitives: points, lines, and triangles
- Raster capabilities: two-op, line draw, and polygon fill
- Shading capabilities: flat and Gouraud shading
- Texture formats: 1-, 2-, 4-, and 8-bit palletized textures
- Textures: animated textures, texture mapping with perspective correction, and MIP mapping

Other common 3D features are double buffering and dithering. Table 1 lists additional 3D features that are not common among all the products covered.

Products that integrate a 2D accelerator all deliver conventional Windows acceleration and provide the logic required for a standard VGA-compatible graphics card. Products that integrate a video accelerator all perform YUV-to-RGB color space conversion and image scaling, features that are useful for accelerating software-based MPEG video playback. Besides integration and rich feature sets, there are other implementation strategies for product differentiation.

One way to resolve the cost of multiple buffers is to perform Z-buffering without a Z-buffer. In general, Z-buffering provides simpler database generation by allowing interpolation of surfaces for hidden-surface removal to create more complex and realistic objects with fewer polygons than non-Z-buffering. For cost-sensitive systems, some 3D applications can run without the Z-buffer. For instance, most of today's games do not require a Z-buffer. But the best of both worlds is to support Z-buffering without paying the cost of a Z-buffer.

Another way to reduce cost is to hold the frame buffer, texture buffer, and Z-buffer in a single memory. That memory can then be allocated dynamically under software control. For example, if an application does not need a Z-buffer, the available memory can be used for textures or display.

Other products differentiate themselves by handling the setup part of the 3D geometry processing in hardware, which has several advantages. It removes the performance bottleneck in the geometry process, since in some cases the 3D setup can consume a great deal of computing power. This strategy also reduces PCI traffic, since instead of sending all 3D primitives over the PCI bus, only 3D vertices need to be transferred.

We believe more companies will realize the benefit of having a 3D hardware accelerator handle the setup part of the 3D geometry process. Designs that rely on the host processor to perform setup calculations for each triangle quickly become CPU-bound, as well as PCI-bound, when rendering speed increases.

Programmable 3D: Chromatic and Rendition

Programmability makes it easy to adopt new software without worrying about hardware obsolescence as standards evolve. Programmable processors, also called media processors, present OEMs with a new way of accelerating multimedia functions like 3D graphics, potentially offering more features for less cost. In contrast, a hardwired 3D architecture may suffer as APIs, data formats, and algorithm technology change over time.

Multimedia companies taking the programmable approach include Chromatic, IBM, Philips, and Rendition. Since IBM's Mfast (see MPR 12/4/95, p. 1) and Philips' TM-1 (see MPR 11/13/95, p. 22) are far from being available, this section discusses only Chromatic's Mpact and Rendition's Verite. Mpact is based on a VLIW (very long instruction word) architecture, whereas Verite is RISC-based. These two companies write their own software to perform 3D acceleration on their programmable engines. Essentially, the programmable engines provide the computing power required for this software to run. The functions of these products depend on how the software is written.

Mpact Leads Multimedia Integration

The Chromatic Mpact media processor (see MPR 10/23/95, p. 23) is designed to efficiently process "natural data types" such as graphics, video, and sound. For these data types, the media processor can perform up to two billion integer operations per second, about ten times more than today's

	3Dfx Voodoo	3Dlabs Permedia	Chromatic Mpact	Nvidia NV3	Oak OTI64311	Rendition Verite	S3 Virge	Trident T3D9695	VideoLogic PowerVR
Anti-aliasing	Yes	No	Yes	No	Yes	Yes	No	No	No
Atmospheric effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Via Apha	Yes
Memory required	2M	2M	2M	1M	1M	2M	2M	2M	2M
Memory type	EDO	SGDRAM	RDRAM	SGRAM	EDO	EDO	EDO	SGRAM	EDO
Sub-pixel positioning	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
Texture compression	Yes	No	No	Yes	Yes	Yes	Yes	No	No
Texture buffer needed	Yes	No	No	No	Yes	No	No	No	Yes
Trilinear filtering	Yes*	No	Yes	No	Yes	Yes	Yes	Yes	No
Z-buffer needed for Z-buffering	Yes	No	No	Yes	No	No	No	No	No

Table 1. 3D feature comparison among the nine newly announced 3D products shows 3Dfx and Oak provide the most complete feature set, while 3Dlabs' Permedia and Trident's 9695 deliver only the bare necessities. *3Dfx uses an equivalent called advanced filtering. (Source: vendors)

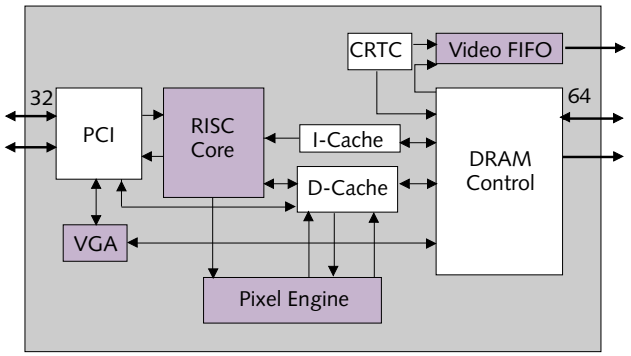


Figure 2. The internal block diagram of Rendition's Verite, a RISC-based programmable 3D accelerator, is similar to that of a highly integrated microprocessor, with on-chip cache, DRAM control, and a PCI interface.

high-end Pentium processors. To fill a 792-bit-wide internal data path, which is linked via high-speed connections to different levels of on- and off-chip memory, the media processor relies on a 500-Mbyte/s (peak) Rambus interface.

Mpact does not handle floating-point computing; it relies on the host processor to execute floating-point calculations. Thus, its 3D performance is somewhat limited by the host CPU's floating-point capability. For example, Mpact cannot handle the setup part of 3D geometry processing, which requires floating-point operations.

For 3D, Chromatic adopts a unified buffer architecture, combining the frame buffer, texture buffer, and Z-buffer within a single memory, which is also used for 2D and video functions. The Mpact feature set is detailed in Table 1.

Chromatic calls its Mpact software Mediaware. It is modularized and can perform multiple functions: video, 2D, 3D, audio, fax/modem, telephony, and videophone. The Mediaware software modules enable these seven functions and provide various levels of capability within each function. The modules contain assembly-language microprograms for the media processor as well as for the operating system and needed drivers.

Chromatic has an interesting business model. The company relies on its semiconductor partners to manufacture and sell the Mpact chips. So far, Toshiba and LG Semicon are the only two vendors that have publicly announced support for the Mpact architecture. The agreements with its partners do not prevent Chromatic from bringing additional manufacturers on board. Chromatic will enjoy some royalty revenue from its partners' sales, but its main source of income will be from the sale of Mediaware modules directly to PC OEMs. These modules will be installed on a PC's hard disk, allowing an OEM to produce several models with varying multimedia functions, all using the same basic motherboard design.

Chromatic has begun alpha testing of both the Mpact media processor and the Mediaware software; Mpact-based motherboards are up and running today. Production-ready Mpact silicon and Mediaware modules are slated for 3Q96. Chromatic claims that five of the top six home PC makers worldwide are working on designs using the Mpact product. For the ISV community, besides making those developers aware of Mpact's capabilities, Chromatic depends on Microsoft to push various types of software development.

Verite Uses RISC-based Design

Instead of using a VLIW instruction set, Rendition adopted a RISC-style instruction set for its Verite product, illustrated in Figure 2. The Verite RISC core handles geometry setup and filtering functions, while its pixel pipeline handles pixel-level texture mapping, blending, and other 3D operations that must occur at hardware speeds. The key architecture differentiation is that Verite offloads the host processor from doing the setup part of geometry operation.

Verite is a single chip integrating the programmable 3D engine, a VGA-compatible 2D accelerator, and a digital video accelerator. An important piece missing in Rendition's design is audio capability. Most consumers interested in 3D will use their 3D machines for games. Without high-performance sound, these games will not be too exciting. Verite must depend on the host processor or an audio chip for sound processing.

Verite's internal RISC processor can issue multiple instructions per cycle and has an expanded graphics instruction set. These graphics instructions, coupled with the large register file (128 × 32 bits), significantly reduce the instruction count of graphics rendering algorithms.

The Verite RISC is similar to other RISC microprocessors. Its RISC engine has the following features:

- Interlocked 32-bit integer processor
- Single-word, multiple-operation instructions
- Linear 32-bit address space
- Combinatorial 32 × 32 multiply
- Delayed load with large register file
- Register file can be used as memory scratchpad
- Delayed non-annulling branches
- Load/store double-word, word, unsigned half-word, and unsigned-byte operands

Compared with Chromatic's Mpact, the Verite programmable engine is tuned mainly for 3D graphics. It has less flexibility in supporting other programmable features than the Mpact processor.

Like the Chromatic chip, the Verite processor combines the frame buffer, Z-buffer, and texture buffer. Unlike Mpact, which is based on RDRAM, the Verite media processor connects to a single bank of relatively inexpensive EDO DRAM. Because of its built-in DMA engine, Verite can also read textures from the PC's main memory.

The potential cost savings from these key features are attracting quite a number of multimedia board makers and title developers to the Verite platform. Creative Labs has committed to Verite for the next generation of its 3D Blaster entertainment platform; it has ported the Creative Graphics Library (CGL) API to Verite. Number Nine is also working

on a Verite-based board. Software vendors like Domark, Mindscape, and Terminal Reality are all planning to release next-generation game titles that take full advantage of the Verite performance.

Hardwired 3D: 3Dfx, 3Dlabs, and Nvidia

Often a hardwired product has higher performance than an equivalent programmable one. A programmable product has advantages in flexibility, but only if the hardwired vendors cannot revise their designs quickly enough to keep up with the evolution of standard algorithms and APIs.

Like Chromatic and Rendition, 3Dfx, 3Dlabs, and Nvidia are all fabless startups that aim to provide 3D solutions to the PC market. These three companies, however, are taking very different hardware approaches from those of Chromatic and Rendition. Their products are fixed in functions and features and depend on host-resident driver software to track changing APIs.

3Dfx Does Voodoo Graphics

The only "pure" 3D company today is 3Dfx. Its first product, Voodoo, does nothing but accelerate various 3D operations. The Voodoo chip set consists of two devices: Pixelfx and Texelfx. The Pixelfx chip includes a triangle engine, a pixel pipeline, and interfaces to PCI and a Z-buffer. The Texelfx chip performs texturing functions such as trilinear filtering, MIP mapping, transparency, and translucency; it connects directly to texture memory. The Voodoo chip set has some features, such as detailed and projected textures, texture morphing, and texture compositing, that are not offered by most existing PC 3D implementations. In terms of 3D performance, Voodoo is one of the highest-performing chip sets for PCs, as Table 2 shows.

Voodoo divides various 3D tasks among the host CPU and its two chips. It is the CPU's responsibility to handle backface culling, transformations, lighting, viewpoint clipping, polygon clipping, and texture management. The Voodoo chip set deals with triangle setup, rasterizing polygons, pixel texturing, texture smoothing, texture combining, pixel fog, depth testing, and alpha blending. Like Rendition's Verite, Voodoo offloads the host processor by taking care of triangle setup, one of the most compute-intensive parts of the 3D geometry operation.

Recognizing that gaining PC OEM business is tougher than selling to arcade and game-console vendors, 3Dfx takes a two-step approach. The company starts by winning designs in the arcade market, where its product's cost is insignificant compared with the selling price of coin-operated arcade machines. For such systems, 3D performance is critical. Most of the company's key contributors are ex-SGI employees who have experience designing high-performance 3D graphics. 3Dfx attracts the arcade-system vendors by allowing them to use high-end Pentium PCs to replace their expensive proprietary engines.

Another advantage to attacking the arcade market first is that the arcade is an existing market for 3D, whereas 3D is an emerging technology for PCs. In addition, most current high-end 3D titles are developed for arcade applications. Those title developers will have more incentive to develop 3D titles for 3Dfx's platform, since, with slight modification, their products can work on both coin-op arcade machines and a growing base of multimedia PCs.

The second step is to push the Voodoo 3D architecture into the PC market. In the mainstream PC market, there is no room for standalone 3D chip sets; what PC makers need is a complete multimedia solution that must include at least 2D and video along with 3D. To address the incompleteness of the initial Voodoo product, 3Dfx recently established a

	3Dfx Voodoo	3Dlabs Permedia	Chromatic Mpact	Nvidia NV3	Oak OTI64311	Rendition Verite	S3 Virge	Trident T3D9695	VideoLogic PowerVR
2D acceleration	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Video acceleration	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Audio capability	No	No	Yes	Yes	No	No	No	No	No
On-chip RAMDAC	No	No	No	No	Yes	No	Yes	Yes	No
TV/VCR interface	No	No	No	No	Yes	No	No	Yes	No
Telephony	No	No	Yes	No	No	No	No	No	No
Fax/modem	No	No	Yes	No	No	No	No	No	No
Pixel-fill rate (million)	45-90	25	TBD	66	33	25	17	37.5	TBD
Polygon-rendering (50-pixels, textured)	655,000	500,000	TBD	500,000	660,000	150,000	340,000	750,000	550,000
Price for chip	$75	<$50	$25-$50	$45	TBD	$75	$40	<$40	<$50
Price for software	Free	Free	$10-$40	Free	Free	Free	Free	Free	Free
Price for board (est)	$350	$250	$150	$90	TBD	$300	$200	$200	$250
Packaging (PQFP)	208/240	256 (BGA)	240	208	208	240	208	208	208
Production date	June 96	2Q96	3Q96	Oct. 96	2H96	1Q96	Feb. 96	4Q96	3Q96
Contact number	(415) 934-2425	(408) 436-3456	(415) 254-0729	(408) 720-7132	(415) 962-9550	(415) 335-5900	(408) 980-5400	(415) 691-9211	(415) 875-0606

Table 2. Functions, performance, price, and availability among the nine newly announced 3D products. These figures show the 3Dfx Voodoo and Trident 9695 as the best performers, but the performance numbers could be misleading because some vendors may quote numbers with certain features disabled. TBD indicates data not available. (Source: vendors)

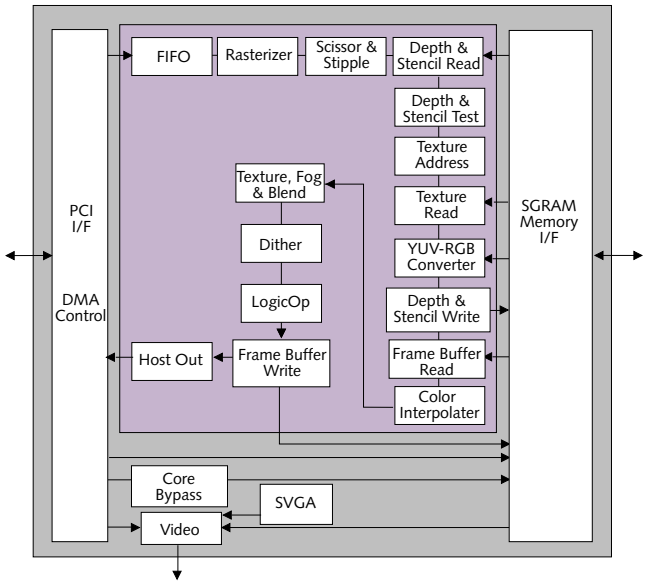


Figure 3. The Permedia internal architecture and graphics pipeline contains a variety of specialized function blocks.

partnership with Alliance Semiconductor. In the short term, this partnership will focus on ensuring compatibility between the Voodoo chip set and Alliance's ProMotion graphics and video controllers. In the long term, the joint effort aims to combine Alliance's DRAM technology, 2D graphics expertise, and PC market experience with 3Dfx's 3D technology into an integrated multimedia solution.

Permedia Aims at Mainstream PCs

Although 3Dlabs was formed from a buyout of Du Pont Pixel in April 1994, it has a long history. If one counts the original Du Pont development and its investment of more than $40 million, the firm has accumulated 10 years of experience in designing hardware and software for 3D graphics. The Permedia device is the company's third-generation product, following Glint and the 3D Blaster chip, which it designed for Creative Labs. Permedia is less expensive than Glint and faster than the 3D Blaster chip.

The Permedia feature set was refined based on feedback from more than 200 software developers for 3D Blaster. New features include chroma-keying and specular lighting. Permedia also adds both 2D graphics and video acceleration to the 3D Blaster feature set, and it replaces that chip's VL-bus with a PCI interface. The chip's internal architecture is illustrated in Figure 3.

The company picked SGRAM (synchronous graphics DRAM) as its memory technology with the anticipation that synchronous memory technologies will dominate PCs in 1997 and 1998, because these chips can scale to frequencies of 100 MHz and beyond. Permedia uses 50-MHz parts now but will enhance performance by taking advantage of faster SGRAMs as they become available.

Like Rendition's Verite, Permedia integrates a VGA-compatible 2D graphics accelerator and video-acceleration functions. Using SGRAM block fills, the Permedia 2D graphics engine can achieve a 2D fill rate of 1.6 Gbytes/s, which is much faster than Verite's rate. In addition, Permedia has a built-in DMA engine that can download textures and command streams from main memory via the PCI bus at rates up to 100 Mbytes/s.

3Dlabs has an interesting marketing strategy, which is to use Creative Technology, a $1.2 billion company, as its board-development partner to test out its technology before it pursues the general market. As one of the first companies to enter the PC 3D business, 3Dlabs needed a partner that could drive ISVs to provide custom support.

The company also licenses its technology to companies that want to integrate 3D into their system silicon, although 3Dlabs expects most of its revenue to come from chip sales. Its licensing support also includes channel, fab capacity, sales force, and technology exchange. To make this business model work, its chip designs are done completely in VHDL. They are modular and abstract.

3Dlabs will maintain two product lines: Glint and Permedia. Glint is for the professional workstation market, optimized for performance, where OpenGL is the main API. Permedia is for the mass consumer market, optimized for low cost and performance, where Direct3D is the emphasis.

The company has realized the benefit of moving the 3D setup operation from the host CPU to external hardware. It has been developing a separate geometry-engine chip for months now. The first silicon of this chip is now working in the lab. Future devices will implement a unified memory architecture, putting the texture buffer, Z-buffer, and frame buffer into the system's main memory. Dynamically allocating memory for various function blocks within the system from a central memory may be the ultimate way to resolve increasing demand for memory while keeping the overall system cost low.

Nvidia Supports Multiple Polygon Models

Among the products discussed in this article, Nvidia's NV1 is the only one being shipped in volume today. The NV1 design (see MPR 7/10/95, p. 13) requires only a frame buffer, the same buffer used for the 2D graphics function. The Z-buffer is optional.

In Nvidia's architecture, textures are stored in system memory and for rendering are transferred across PCI via bus-mastering DMA. System memory is swapped out as needed by the application. Texture mapping starts with the texture and computes the destination pixels to which these source texels should be mapped. This algorithm, called forward rendering, accesses the source texels in consecutive order by scanline, making these accesses amenable to bursting over PCI. This method allows Nvidia to keep its textures in system memory and still maintain the bandwidth required by the internal rendering engine.

In contrast, traditional 3D texture-mapping algorithms start from the destination pixels on the display space, calculate the mapping to the texture space, read the texture, and apply it to the destination pixel. This algorithm, called inverse rendering, makes burst accesses difficult, because the accesses to texture memory are not in consecutive order. It is not feasible to store textures in system memory if there is no mechanism to get them to the rendering engine quickly.

Nvidia has admitted that if it could market NV1 all over again, it would stress that the NV1 part accelerates triangles as well as curves. Because the company put some initial emphasis on its propriety quadratic texture map (QTM) algorithm, which is similar to NURBS, many ISVs believe NV1 uses a proprietary algorithm, something most ISVs don't want to support.

Actually, NV1 can perform well using regular polygons like triangles and rectangles. The difference is that the NV1 architecture can handle one additional class of primitives —curves—that cannot be rendered by other current 3D implementations. Applications that use the curve support of the QTM algorithm have found better photorealism than those running with that algorithm disabled.

But of the software vendors Nvidia is working with today, none is planning to use the QTM algorithm. Most current software models curved surfaces using many triangles. To improve appearance without using QTM, software can modify the light intensity of pixels so the polygons appear curved and smooth. This method has the advantage of working on a variety of 3D chips, unlike the proprietary QTM.

Nvidia Moving to Second Generation: NV3

Like the other 3D startups, Nvidia has a unique business model. The company arranged for SGS-Thomson to allocate foundry capacity to Nvidia's accelerators, granting SGS an exclusive right to sell the DRAM version of NV1 while Nvidia sells the VRAM version. Nvidia began shipping NV1 products last November. The company claims 100,000 units have shipped, as of 1/31/96. The company has sold more of its VRAM version than SGS-Thomson has of the DRAM version, because the early adopters of the new chip care more about performance than cost.

Sega has already bought into Nvidia's story. Its Virtua Fighter game software was demonstrated at Comdex '95 running under Windows 95. So far, the NV1 is the only chip available that accelerates all the APIs contained in Microsoft DirectX, using only 1M of memory.

The NV1 chip provides 2D acceleration, 3D graphics, wavetable MIDI synthesis, MPEG video acceleration, and a precision digital/analog joystick interface. No host CPU intervention is needed for the joystick, and with a Pentium-133 CPU, the audio engine takes only 1% of the host's power for multitrack simultaneous MIDI and digital mixing.

The built-in DMA and 600-Mbyte/s on-chip packet bus together deliver the data stream required by the multimedia engine. The chip maximizes 2D scrolling performance beyond the refresh rate of the monitor at any resolution. One drawback is that the NV1 is not quite Sound Blaster-compatible. For compatibility, an external device is needed.

Nvidia is now working on its next-generation media accelerator, NV3, which will offer 100% Sound Blaster compatibility. The key feature of NV3 is not its Sound Blaster compatibility, but its superior performance (see Table 2). The NV3 part also integrates the custom RAMDAC that is a separate chip for NV1, reducing system cost. NV3 is a functional superset of NV1, maintaining 100% software compatibility with NV1. Recently, Nvidia announced a partnership with Lexicon to deliver Lexicon's home theater sound into the PC using the NV3 product. The company hopes to ship the NV3 this fall.

Home-grown 3D: Oak, S3, Trident, VideoLogic

These four companies have been providing 2D graphics, video, and other peripheral chips to the PC market for years. With the increasing demand for 3D in PCs, these companies cannot afford to stand still. If they do, they will not only miss many 3D business opportunities but will also lose their existing business to those aggressive 3D startups. The advantages these companies have are their 2D/video technologies and their existing customer bases, as well as their knowledge of the PC business. Their disadvantage is that some of them lack 3D design experience.

This section briefly discusses the OTI64311 from Oak, S3's Virge, the T3D9695 from Trident, and PowerVR from VideoLogic and NEC. Like the Nvidia chips, some of these devices combine graphics and video acceleration with 3D features. We describe these chips as media accelerators.

Oak Launches Its First 3D Chip

Oak, the world's biggest supplier of CD-ROM controllers, also produces high-quality audio and video products. The company recently assembled a 30-person team for its 3D business unit. Its 3D designers come from Kubota, SGI, Sun, and Martin Marietta. Some key technologies were acquired from Kubota, which recently exited the workstation graphics business. In this way, Oak has quickly established its own 3D knowledge base.

The company's first 3D chip, the OTI64311, will combine 2D and video with 3D graphics rendering. Leveraging the company's 2D and MPEG technologies, the OTI64311 is a highly integrated single-chip media accelerator. The chip has a port for connecting directly to an external NTSC/PAL decoder for TV and VCR output and also includes a 170-MHz RAMDAC and dual clock generators. This level of integration should greatly reduce system cost and the board space needed for multimedia acceleration.

The OTI64311 incorporates several innovative memory and computation techniques to enable high-performance 3D graphics with standard EDO DRAMs. The key architectural feature of the OTI64311 is the way Z-buffering is implemented. With a proprietary memory-efficient

algorithm, the OTI64311 can perform a normal Z-buffering operation without paying the Z-buffer cost. The 3D features of the OTI64311 media accelerator are listed in Table 1.

S3 Ready to Ship Virge in Volume

Virge stands for Video and Rendering Graphics Engine; it is S3's first hardware accelerator to include 3D support. Virge is a single chip that consists of 3D functions plus the company's popular Trio64 graphics/video chip. The 3D features included in Virge are listed in Table 1.

The chip uses either EDO DRAM or VRAM for the graphics buffers. S3's multiplexed buffering allows Z-buffering without the cost of dedicated Z-buffer memory. This patented buffering scheme can share memory with the frame buffer in the system while minimizing additional memory requirements. A 220-MHz RAMDAC is also integrated into the Virge chip.

Virge is pin-for-pin compatible with S3's Trio64V+, one of the world's most widely implemented 64-bit 2D graphics controllers. This compatibility enables add-in-card developers and system makers to upgrade to 3D acceleration without redesigning their boards. Virge also incorporates the company's proprietary bus for a direct connection to live video or an MPEG decoder.

The business model for S3 has been to go after big customers such as Compaq and Microsoft. The company claims more than 15 hardware manufacturers and more than 10 game developers are supporting its Virge product.

Trident Announces Third Generation: T3D9695

Trident is one of the world's leading graphics/video chip suppliers. Its product strategy has always been to go after the volume market with the best balance of performance and price. Trident gained its experience from its first 3D chip, the T3D2000, which was announced in June 1995. It launched a software developer program that provides developers with evaluation boards based on the T3D2000 as well as documentation and software-development tools for porting to its 3D technology. The T3D9695 is the third generation of products with 3D-acceleration functions. The second-generation chip, the T3D9692, is currently selling.

The newest device integrates 2D, 3D, and video acceleration along with a VMI port for connecting an MPEG decoder and a video port for connecting an NTSC/PAL encoder for TV and VCRs. All these features are packed in a 208-pin PQFP package and manufactured using 0.5-micron CMOS. The key differentiators of the T3D9695 are:

- Pin-compatibility with its popular predecessors, such as the T3D9692 and TGUI9680, provides designers with the flexibility to use a single board design for multiple graphics solutions.
- Like Chromatic's Mpact, it uses a single buffer as frame buffer, texture buffer, and Z-buffer.
- Its high integration and low power dissipation are well suited to multimedia notebook PCs.
- The TV interface helps certain TV-related multimedia consumer products.
- It can work with UMA core-logic chip sets from Opti, SiS, UMC, and VIA, which will help bring basic 3D functions into low-cost PCs.

Trident expects OEMs to build a graphics, video, and 3D accelerator board for less than $200 using the T3D9695. The company is considering reducing the price of future products to the $25 range by cutting some little-used features, such as CAD support.

It is worth mentioning that the T3D9695 is one of the best video-quality chips among those covered in this article. Trident's proprietary edge-smoothing algorithm enables both horizontal and vertical interpolation for the video stream. The chip supports motion video from a CD-ROM at up to 30 frames per second without the need for additional hardware accelerators.

VideoLogic Teams with NEC for PowerVR

VideoLogic has been providing video chips and add-in video cards to the PC market. It is now entering the 3D race with semiconductor giant NEC as its partner. Under its agreement with NEC, VideoLogic is responsible for design and development while NEC handles production and sales. Both companies put effort into product design and marketing. The two have teamed with Namco, a leading arcade-system and game-title developer, to bring arcade-quality software to the PC platform. Namco's first game title running on PowerVR was demonstrated at Fall '95 Comdex. NEC says its sales target is two million PowerVR chips installed in PCs worldwide by the middle of 1996.

PowerVR is the 3D architecture originated by VideoLogic. The key to the architecture is abolishing the Z-buffer and reducing the memory-bandwidth required. The algorithm developed by VideoLogic is based on a so-called infinite plane (surface based) that does hidden-surface removal on the fly. This method sidesteps memory-bandwidth limitations that arise from accessing the Z-buffer.

Because PowerVR has lower bandwidth requirements, it is possible to render large polygons, which are required for some special effects. For example, shadows and searchlights are fundamentally large objects, as they represent light rays traveling long distances. A full shadow, where any object can cast shadows on any other object, and real-time searchlights bring a new dimension to games. In addition to eliminating the cost of a Z-buffer, PowerVR removes the bottleneck associated with accessing such off-chip memory.

Although a frame buffer and a texture buffer are still required in the PowerVR design, their bandwidth requirements are greatly reduced. Each pixel requires an access to the frame buffer only twice: once to be written to the frame buffer, and the second time for being transferred to the display. This is true even for transparent pixels.

Using a "deferred" texturing approach, the PowerVR chip reduces both texture- and frame-buffer bandwidth.

Unlike the conventional rendering system, where every polygon is textured, the PowerVR approach textures only final visible pixels. This technique makes the memory bandwidth independent of the polygon rendering rate. The texture buffer will be accessed one to eight times depending on the complexity of the texture mapping: point mapping requires only one access, whereas trilinear MIP mapping requires eight accesses.

The PowerVR design has the following advantages:

- Cost savings due to elimination of the Z-buffer
- Low memory bandwidth requirements because of the elimination of the Z-buffer, updates of only visible pixels, and the deferred texture approach
- Memory bandwidth requirements independent of polygon-per-second performance figures, since not all intermediate pixels and texels are considered

NEC will manufacture three chips using the PowerVR technology. An image synthesis processor (ISP) and a texture shading processor (TSP) will be used in arcade systems. Performance can be increased by adding more ISPs. For PC applications, another chip, called PCX1, combines one ISP, one TSP, and a PCI interface into a single die. The performance figures of the PCX1 chip are listed in Table 2. A future single-chip solution is also planned for console applications.

Choosing the Right Solutions for 3D Applications

With a dozen 3D graphics accelerators about to appear on the market and the lack of a unified software platform, picking the right 3D solution isn't easy. A 3D board manufacturer or a multimedia PC maker must take the following issues into consideration while choosing its 3D strategy:

Soft 3D. The Pentium-166 is already in the market. Even with the Pentium-133, Microsoft has recently found that under Direct3D, PCs are capable of performing many 3D tasks without any 3D hardware acceleration. Intel will roll out its long-awaited P55C in the second half of this year. With its MMX instructions (see cover story), larger internal cache, higher clock frequency, and pipeline enhancements, the P55C processor can certainly bring host-based 3D performance to another level.

We expect the P55C will be introduced in 4Q96, followed quickly by Klamath, a low-cost Pentium Pro with MMX, in early 1997. Depending on the type of CPU and the expected shipment date of the PC under consideration, software-only 3D may provide adequate performance, eliminating the cost of a hardware 3D accelerator.

Integration. If the design goal is high integration without relying on add-in cards for multiple functions, the OEM should consider Chromatic's Mpact, Nvidia's NV1, and Trident's T3D9695. The Mpact media processor and Mediaware modules support 2D, 3D, audio, MPEG video, fax/modem, telephony, and videophone. The NV1 media accelerator supports 2D, 3D, audio, and video. Products based on the NV1 will have a migration path to the next level of performance and integration using the NV3. The T3D9695 supports 2D, 3D, video, and a TV interface, but not audio. The T3D9695 was designed for applications such as multimedia notebooks and TV-related consumer products.

Performance. If high performance is desired to differentiate its products, the OEM should consider 3Dfx's Voodoo, 3Dlabs' Permedia, Nvidia's NV3, Oak's OTI64311, and Rendition's Verite. These products are now either in their engineering-sample stages or their final verification stages. Testing or simulation results have been promising. The companies that produce these products have engineering teams that are capable of designing high-performance 3D products. None, however, has been benchmarked in production systems.

Cost. If the target market segment is price-driven, the system maker should consider products from Chromatic, Nvidia, Oak, Trident, and VideoLogic. Although the Mpact media processor needs a frame buffer, a texture buffer, and a Z-buffer, these three buffers are physically in a single unified memory. Because of Mpact's high integration, products based on it can get rid of many discrete components.

Because of their DMA engines and forward texture mapping, Nvidia's NV1 and NV3 eliminate the need for a dedicated texture buffer. Although they do not have fax/modem, telephony, and videophone support, the NV1 and NV3 media accelerators integrate four key multimedia functions: audio, video, 2D, and 3D.

The Oak chip's architecture supports Z-buffering without the cost of a Z-buffer. Both Oak and Trident have a track record for providing low-cost solutions to the PC market. Since the PowerVR architecture gets rid of the Z-buffer completely, it is a good solution for 3D add-in cards. It is not as good for a system solution, however, because of its lack of support for 2D and video functions.

Risk. As mentioned above, PC 3D graphics is an emerging technology. Most products discussed in this article have not yet been fully debugged. Comparatively, 3Dlabs' Permedia, Nvidia's NV1, S3's Virge, and Trident's T3D9695 are lower in risk. The NV1 is in production, and Virge has been proved to work. Permedia and the T3D9695 are third-generation products for 3Dlabs and Trident, respectively, reducing their risk.

Chromatic, 3Dfx, and Rendition are close to shipping their 3D products, but these companies are startups that have not yet produced any volume products. Although Oak and VideoLogic have been in business many years, the 64311 and PowerVR are the companies' first 3D products.

With so many companies working on competing products, it is likely that at least some will deliver attractive 3D accelerators in time for volume PC shipments later this year. Assuming Direct3D meets its objectives, software developers will be able to add exciting 3D capabilities to their games and other applications, spurring consumer demand for 3D performance. A plethora of new products and new companies will emerge as 3D becomes prevalent in PCs by the end of 1997. M

LITERATURE WATCH

DEVELOPMENT TOOLS

EDA tools let you track and control CMOS power dissipation. Knowing where your chip is dissipating power is important for both IC and chip-based system design. Jim Lipman, *EDN Europe,* 1/96, p. 7, 7 pp.

IC analysis tools help manage power. Power consumption has become an important parameter. Now designers have a choice of tools to help them cope. Lisa Maliniak, *Electronic Design,* 1/22/96, p. 71, 6 pp.

Pin multiplexing yields low-cost logic emulation. Instead of letting logic simulators bog down on complex designs, debug using reprogrammable logic emulators. Lisa Maliniak, *Electronic Design,* 1/22/96, p. 65, 3 pp.

Designers cross over to language-based tools. Graphical tools for high-level design languages (HDLs) help designers cope with complexity, but performance may suffer. Mike Donlin, *Computer Design,* 1/96, p. 69, 6 pp.

DSPS

Optimize DSP design with an extensible core. Maximize your DSP circuit's cost, power, and form factor with a building-block approach that lets you choose and match the right cells from a library of ASIC functions. Bob Caulk, *Electronic Design,* 1/22/96, p. 81, 4 pp.

$5 buys a 20-MHz, 16-bit fixed-point DSP controller. Zilog's 289323 combines a Clarkspur DSP core with a full set of peripherals. *Computer Design,* 1/96, p. 118, 1 pg.

GRAPHICS/VIDEO

JPEG parameters determine compression-system performance. The wide availability of JPEG software and hardware simplifies image-compression system design. Debora Grosse, Unisys; *EDN,* 1/18/96, p. 141, 6 pp.

Integrated MPEG ICs lower set-top, DVD costs. Vendors including LSI, SGS-Thomson, IBM, C-Cube, and AT&T offer low-cost MPEG-2 decoders. John Mayer, *Computer Design,* 1/96, p. 57, 4 pp.

MISCELLANEOUS

High-capacity, removable storage drives shake floppy foundation. More than a dozen products based on magnetic and optical technologies are ready to take a shot at replacing the 1.4M floppy. Maury Wright, *EDN,* 1/18/96, p. 41, 9 pp.

PowerPC poised to dominate VME. PowerPC has become the dominant VME RISC processor by offering the best mix of performance, popularity, and embeddability. Jeff Child, *Computer Design,* 1/96, p. 133, 5 pp.

PERIPHERALS

Cryptographic techniques. Code-generation and decoding ICs provide rolling-code encryption techniques that let noncryptographic experts design in security. Doug Conner, *EDN,* 1/18/96, p. 57, 6 pp.

Advancing the art of industrial video imaging: CCD cameras and frame grabbers. The industrial market for video-imaging applications is wide and varied. John Gallant, *EDN,* 1/18/96, p. 73, 6 pp.

PROCESSORS

Future comm processors will fuse DSP and RISC. One vision of future embedded communications processing calls for a single processor with a hybrid architecture, combining the best features of today's RISC and DSP machines. Dan Mansur, *Electronic Design,* 1/8/96, p. 99, 3 pp.

Embedded-controller architectures suit all needs. For cost-sensitive applications, 4- and 8-bit microcontrollers continue to dominate designer's thoughts. Dave Bursky, *Electronic Design,* 1/8/96, p. 53, 8 pp.

On-chip instructions boost multimedia. An UltraSparc architect explains how that chip accelerates real-time video and graphics. Marc Tremblay, *Electronic Design,* 1/8/96, p. 131, 3 pp.

PROGRAMMABLE LOGIC

Efficient RAM-based FPGAs ease system design. Enhanced FPGAs with distributed or dedicated RAM boost system throughputs. Dave Bursky, *Electronic Design,* 1/22/96, p. 53, 5 pp.

Enhanced FPGA family delivers 125,000 gates. By setting a new standard for gate count, the Xilinx XC4000EX allows integration of system functions. Dave Bursky, *Electronic Design,* 1/26/96, p. 141, 2 pp.

Programmable logic continues to push density/power thresholds. AT&T's 2CxxA family, Altera's EPF10K50, Xilinx's XC9500, and Atmel's ATF1500 offer increased density over previous CPLDs. Mike Donlin, *Computer Design,* 1/96, p. 32, 2 pp.

SYSTEM DESIGN

Invisible computing: now you see it, now you don't. With the convergence of several technologies, embedded computing is fast becoming the foundation of mass-market products. Clifford Meth, *Electronic Design,* 1/8/96, p. 84, 6 pp.

Previewing system solutions of the future. Tomorrow's communication systems will demand more of their embedded controllers, leading experts to advocate a two-processor solution. Trey Oprendek, *Electronic Design,* 1/8/96, p. 104, 3 pp.

Choosing lithium primary-cell types. Although they share common attributes, each of the various chemistries for lithium cells lends itself to different uses. Mark Schimpf, *Electronic Design,* 1/8/96, p. 141, 3 pp.

Supervisory ICs establish system boundaries. Although they don't have the glamour of CPUs, supervisory ICs perform unseen critical tasks in reset, memory protection, and watchdog functions. Bill Schweber, *EDN,* 1/8/96, p. 23, 5 pp.

RECENT IC ANNOUNCEMENTS

PART NUMBER	VENDOR	DESCRIPTION	PRICE/QUANTITY	AVAILABILITY
MICROPROCESSOR				
WARP11	SGS-Thomson 617.259.0300	Fuzzy-logic Weight-Associative Rule Processor (WARP) has 16 inputs and outputs, handles 256 rules, and operates at 40 MHz; in PLCC-68.	$18.12/5,000	Prod.—Now
INTERFACE				
DB87144	Databook 508.762.9779	Single-chip host interface controller supports CardBus (PCI on PCMCIA) and proposed Zoomed video specification; complies with PCMCIA 2.1.	$21/1,000	Samples—Now Prod.—2Q96
CL-CD1865	Cirrus Logic 510.226.2140	Eight-channel communications controller has 24 eight-byte FIFOs and improved serial data rates; compatible with CD180 and CD1864.	$25/1,000	Samples—Now Prod.—2Q96
AD421	Analog Devices 617.937.1428	Loop-powered 16-bit DAC sends 4–20-mA signals to microcontrollers or control devices; with voltage regulator for 3-V, 3.3-V, or 5-V outputs.	$6.95/1,000	Prod.—Now
MEMORY				
MPC27T416 MPC27T415	Motorola 512.933.7726	Fast synchronous SRAMs for L2 cache subsystems, with 9-, 10-, and 12-ns access times; organized as 16K×16 ('416) or 16K×15 ('415).	$28.88/10,000	Samples—2Q96 Prod.—3Q96
IDT7281 IDT7282	IDT 800.345.7015	Dual asynchronous FIFOs are housed in small, 56-lead TSSOP packages; with 15-, 20-, and 25-ns access times; 512×9 ('281) or 1K×9 ('282).	$6.40/5,000	Prod.—Now
HN62W5016	Hitachi 800.285.1601	Burst-mode masked ROM has 16-Mbit capacity and can be configured for either ×16 or ×32 organization; 40-ns access time and 3.3-V supply.	$9/10,000	Prod.—Now
ISD2560	ISD 408.369.2422	ChipCorder memory IC is designed for voice recording and playback using 256-level voltage format per cell, eliminating external A/D or D/A.	$13.55/1,000	Prod.—Now
MISCELLANEOUS				
ML4878	Micro Linear 408.433.5200	Backlight chip for liquid-crystal displays reduces external component count and reduces leakage current to extend battery life; in SSOP-14.	$3.95/1,000	Prod.—Now
LTC1480	Linear Technology 408.432.1900	RS-485 transceiver operates from a single 3.3-V supply while providing 10kV ESD protection; part draws 300 µA supply current; in SO-8 package.	$2.55/1,000	Prod.—Now
X25045	Xicor 408.432.8888	Microprocessor supervisor for Intel microcontrollers provides reset generation, 512 bytes of EEPROM, and watchdog timer; in SOIC-8 package.	$1.50/10,000	Prod.—Now
78Q8392L	Silicon Systems 800.624.8999	Transceiver chip for coaxial Ethernet is compliant with Ethernet II standards 802.3, 10BASE5, and 10BASE2; part draws 50 mA while transmitting.	$2.60/1,000	Prod.—Now
78Q2250	Silicon Systems 800.624.8999	Line transceiver for ATM provides full transmit/receive facility with line-interface and clock-recovery functions at 155.52 Mbits/s over UTP wiring.	$32/1,000	Prod.—Now
VSC7125	Vitesse 805.388.3700	Integrated Fibre Channel transceiver operates from a single 3.3-V supply while drawing 650 mW at 1.0625 Gbps; in H-GaAs IV process.	$37/1,000	Prod.—Now
PROGRAMMABLE LOGIC				
MACH111 MACH231	AMD 800.222.9323	Fast CPLDs have guaranteed 5-ns ('111) or 6-ns ('231) maximum pin-to-pin delays; with 32 or 64 macrocells and 44 or 84 pins.	$17/1,000	Prod.—Now
XC4000EX	Xilinx 408.559.7778	Programmable logic family extends to 125,000 usable gates with selectable RAM cells; parts are PCI compliant and operate at 66 MHz.	$695/100	Samples—Now Prod.—4Q96

PATENT WATCH

by Rich Belgard, Contributing Editor

The following U.S. patents related to microprocessors were issued recently. Please send comments via e-mail to belgard@umunhum.stanford.edu

5,467,460

M&A for minimizing data transfer to main memory from a writeback cache during a cache miss
Issued: November 14, 1995
Inventor: Piyush G. Patel
Assignee: Intel
Filed: April 28, 1994
Claims: 4
A cache memory having at least two modified bits for each block of data coupled to a writeback buffer circuit for transferring a fraction of the data block when a cache miss occurs. The data array of the data cache is partitioned into two halves; each block of data has two dirty bits. When a cache miss occurs, a replacement algorithm determines which of the lines in a given set shall be replaced. The contents of the chosen line in the data cache is copied to a writeback buffer circuit. The line of data from external memory is then written into the data cache, clearing the two dirty bits in the data cache in the process. If only one dirty bit is set, only half of a block of data is written back into the data cache.

5,467,455

Data-processing system and method for performing dynamic bus termination
Issued: November 14, 1995
Inventors: James G. Gay, et al
Assignee: Motorola
Filed: November 3, 1993
Claims: 29
A data-processing system and a method for performing dynamic bus-signal termination uses a dynamic bus-termination circuitry with a device. The circuitry is enabled when data is incoming to the device and is disabled when data is outgoing from the device to selectively reduce unwanted signal reflection at the signal end of a bidirectional bus. The disabling allows the circuitry to be removed or tristated from any connection with the bus when not needed (i.e., data outgoing) to reduce loading. The disabling of the termination circuitry also aids in reducing the power consumption of the part when either the bus is sitting idle or the part is in a low-power mode of operation.

5,465,373

Method and system for single-cycle dispatch of multiple instructions in a superscalar processor system
Issued: November 7, 1995
Inventors: James A. Kahle, et al
Assignee: IBM
Filed: January 8, 1993
Claims: 7
A method and system for permitting single-cycle instruction dispatch in a superscalar processor system that dispatches multiple instructions simultaneously to a group of execution units for execution and storing of results. Multiple intermediate storage buffers are provided, and each time an instruction is dispatched to an available execution unit, one of the intermediate storage buffers is assigned to any destination operand within the dispatched instruction, permitting the instruction to be dispatched within a single cycle by eliminating any requirement for determining and selecting the register.

5,465,336

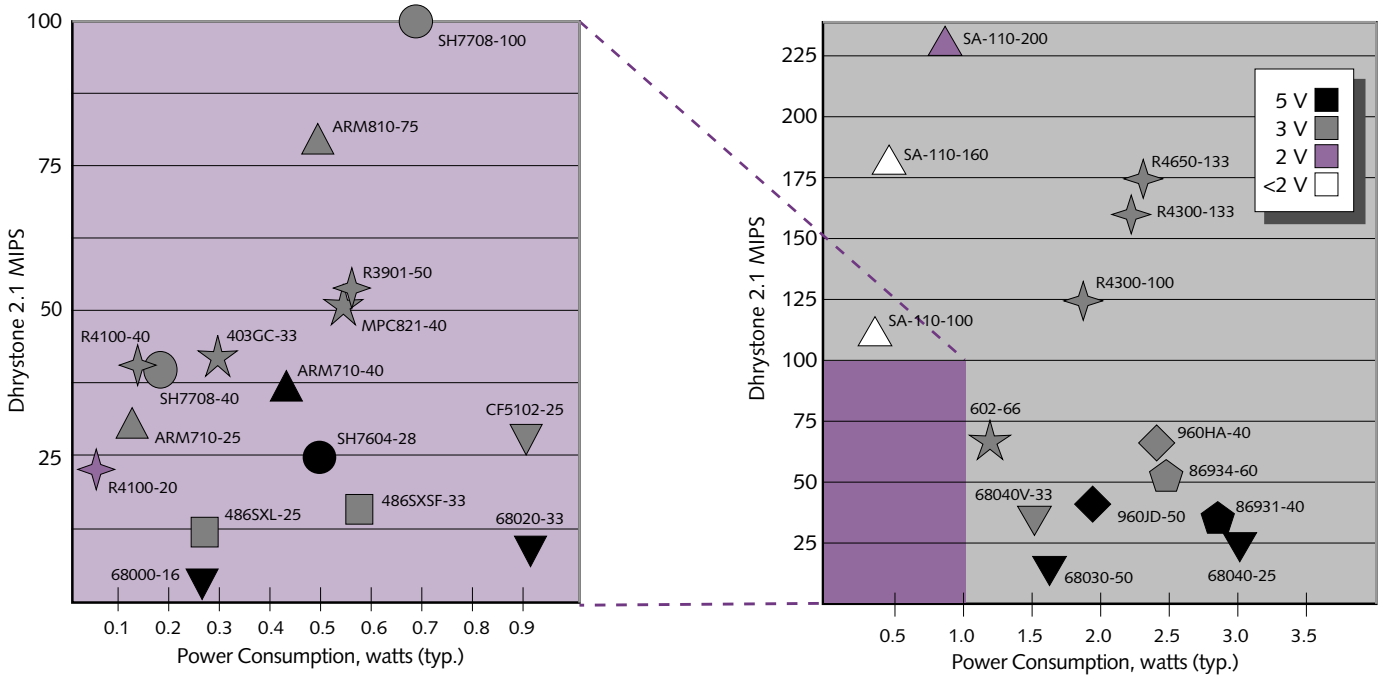
Fetch and store buffer that enables out-of-order execution of memory instructions in a data-processing system
Issued: November 7, 1995
Inventors: Benjamin T. Imai, et al
Assignee: IBM
Filed: June 30, 1994
Claims: 2
A method and device for handling fetch and store requests in a data-processing system. A fetch and store buffer consists of a store queue, a fetch queue, a register, a comparator, and a controller. The store queue and the fetch queue receive requests from one or more execution units. When the fetch queue receives a fetch request from an execution unit, it sets a mark in a field associated with the request, indicating the store queue entries present at the time the fetch request is entered. It then removes the mark from the field when the associated store queue entry is drained.

The controller sends a queued fetch request to the memory unit when it is ready to accept a request. The comparator determines if there is a dependency between the request in the memory unit address register and any store queue entries marked in the gated request's field. When a dependency is determined by the comparator, the controller drains the store queue entries marked in the pending fetch request's field from the store queue prior to draining the fetch queue entries.

OTHER ISSUED PATENTS

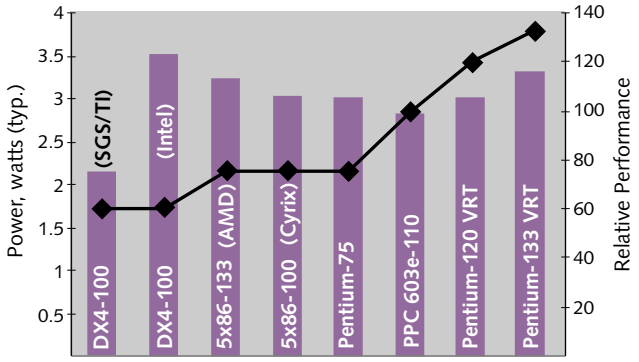
5,469,547 *Asynchronous bus interface for generating individual handshake signal for each data transfer based on associated propagation delay within a transaction*
5,467,318 *Address generating and decoding apparatus with high operation speed*
5,465,377 *Compounding preprocessor for cache for identifying multiple instructions which may be executed in parallel* M

CHART WATCH: MOBILE PROCESSORS



This Chart Watch covers low-power processors for portable and battery-powered systems. The table and the chart in the upper-right show the performance/power ratio for a number of embedded CPUs and notebook processors; the chart above is an inset for the lowest-power of these processors.

The chart on the right compares Pentium and Pentium-class processors, including relative performance (diamonds) and typical power consumption (bars).



	SA-110	ARM710	SH7708	PPC 602	R4100	R3901	CF5102	486SXSF	486SXL
Vendor	Digital	VLSI	Hitachi	Motorola	NEC	Toshiba	Motorola	Intel	National
Clock rate	200 MHz	40 MHz	100 MHz	66 MHz	40 MHz	50 MHz	25 MHz	33 MHz	25 MHz
I/D cache	16K/16K	8K	8K	4K/4K	2K/1K	4K/1K	2K/1K	8K	1K/0K
FPU?	No	No	No	Yes	No	No	No	No	No
MMU?	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	No
Bus width	32 bits	32 bits	32 bits	64 bits	32 bits	32 bits	32 bits	32 bits	16 bits
Bus frequency	66 MHz	40 MHz	25 MHz	33 MHz	20 MHz	50 MHz	25 MHz	33 MHz	25 MHz
MIPS	215 MIPS	36 MIPS	100 MIPS	65 MIPS*	40 MIPS*	52 MIPS	27 MIPS	16 MIPS*	12 MIPS
Voltage§	2.0/3.3 V	5 V	3.3 V	3.3 V	3.3 V	3.3 V	3.3 V	2.7 V	3.3 V
Power (typ)	900 mW	424 mW	700 mW	1,200 mW	120 mW	550 mW	900 mW	515 mW	260 mW
MIPS/watt	239	85	143	54	333	95	30	31	46
MIPS/mm²	4.30	1.04	2.37	1.30	1.60	0.72	n/a	n/a	0.41
Transistors	2,100,000	570,295	800,000	1,000,000	450,000	510,000	n/a	n/a	256,000
IC process	0.35µ 3M	0.6µ 2M	0.5µ 3M	0.65µ 4M	0.5µ 3M	0.65µ 3M	0.6µ 3M	0.8µ 2M	0.65µ 3M
Die size	50 mm²	34 mm²	44 mm²	50 mm²	25 mm²	72 mm²	n/a	n/a	30 mm²
Est mfg cost	$18*	$9*	$12*	$14*	$8*	$9*	$9*	$15*	$8*
Availability	2Q96	Now	Now	Now	Now	Now	Now	2Q96	Now
Price (10K)	$49	$28	$50	$45*	$25‡	$27†	$25*	$72†	$15

† list price in 1,000s ‡list price in 100,000s §core/bus voltage n/a: information not available (Source: vendors except *MDR estimates)

Mobility Moves into San Francisco

With one in four new PCs sold being portable and more than eight million Americans telecommuting, times are good for vendors of portable, mobile, and wireless products. To give them all a place to meet, **Mobile Forum '96** will be held at the Hyatt Regency San Francisco Airport in Burlingame. The two-day show starts March 18 with presentations from the presidents of Ericsson and AT&T Wireless Services and the FCC's director of strategic policy. The next day deals with technical and marketing issues for wireless and wireline communications, voice, infrastructure, and applications.

Admission costs $1,495 and includes all events and meals. For more information, contact Technologic Partners (New York) at 212.343.1900 or e-mail *events@technologicp.com.*

Humans Congregate in British Columbia

The **Conference on Human Factors in Computing Systems** will be held April 13–18 at the Vancouver Trade Center, Vancouver, Canada. The annual forum is hosted by the ACM and includes six days of tutorials, workshops, and special-interest groups.

Early registration (before 3/7) for the conference costs US$390 for ACM members; $504 for nonmembers. Each tutorial costs $255 extra. For more information, contact the CHI conference office (Maitland, Florida) at 410.263.5382; fax 407.628.3186.

Los Angeles Serves Up Digital Video

The University of California at Los Angeles (UCLA) Extension is offering a three-day short course in **Digital Video Technology,** to be held near the UCLA campus on April 10–12. The course covers many aspects of acquiring, transmitting, processing, and displaying photographic-quality still and moving pictures. It is intended for engineers, designers, and programmers.

The fee for the course runs $1,295, including all course materials, and provides 1.8 CEU credits. A BS degree or equivalent is required. For more information, or to register, contact UCLA (Los Angeles) at 310.825.3344; fax 310.206.2815.

E3 Reprise Brings It All Home

The second annual **Electronic Entertainment Expo** (E3) comes once again to the L.A. Convention Center on May 16–18. The show focuses on CD-ROM and multimedia software and hardware, electronic media, games, accessories, on-line entertainment, virtual reality, and video-game platforms. Four conference tracks deal with the business, creative, marketing, and financial aspects of the electronic-entertainment market. E3 also has some of the most interesting demos of any trade show.

Admission to the exhibits costs $35 before April 12; $50 at the door. Conference registration costs $120. For more information, or to register, fax E3 (Norwood, Mass.) at 617.440.0359 or access the Web at *www.mha.com/e3.*