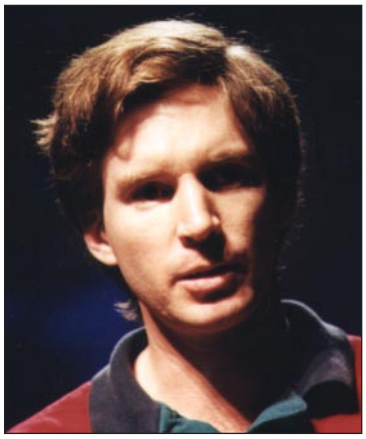
Alpha 21364 to Ease Memory Bottleneck

Compaq Will Add Direct RDRAM to 21264 Core for Late 2000 Shipments

by Linley Gwennap

With processor speeds rapidly approaching the gigahertz mark, the shortcomings of today's memory architectures are becoming all too apparent. Waiting a few hundred nanoseconds to retrieve data from main memory is tolerable for a 100-MHz CPU, but this delay turns into hundreds of cycles for a 1-GHz processor. Compaq's Alpha processors are likely to be the first to reach that speed, and at this month's Microprocessor Forum, the company previewed its solution to this problem, embodied in the 21364.

The 21364, code-named EV7, will use multiple Direct RDRAM channels to pump information from main memory right into the processor, greatly reducing latency. The 21364 certainly won't be the first processor to include a DRAM controller; recent examples range from Cyrix's MediaGX to Sun's forthcoming UltraSparc-3. But the Alpha design presents a unique and elegant combination of a large on-chip L2 cache, direct access to DRAM, and a high-speed interprocessor connection. We believe this combination will become a common design for high-performance multiprocessor systems in the next decade.



MICHAEL MUSTACCHI

Compaq's Pete Bannon explains how the 21364's revolutionary system interface helps TPC-C.

The 21364, which will wrap this new system interface around a 0.18-micron 1-GHz version of the current 21264 CPU core, is still early in its design phase. At the Forum, Alpha architect Pete Bannon said he expects the chip to tape out in late 1999, with shipments in late 2000 or early 2001. Thus, its main competition will come from future high-end processors such as UltraSparc-4 and Intel's Merced. The Alpha chip's projected workstation scores of 60 SPECint95 (base) and 100 SPECfp95 (base) set a high bar for competitors to leap, but the 21364 will truly soar in multiprocessor servers, where its lead could be greater.

It's the Memory, Stupid!

Although Alpha inventor Dick Sites left Digital about a year ago, the 21364 is covered with his fingerprints. In a column titled "It's the Memory, Stupid!" (see MPR 8/5/96, p. 18), Sites related that in a database study using the TPC-C benchmark, the CPU was stalled waiting for main memory on three out of every four cycles. Given that the CPU at the time was a 400-MHz 21164, we can only imagine the number of stalls that would be encountered by a 1-GHz 21264. Sites concluded his column by saying, "Over the coming decade, memory subsystem design will be the *only* important design issue for microprocessors."

At the time, memory-system design and processor design were considered to be independent. This is clearly no longer the case. Sites realized that to solve the memory bottleneck, the processor must communicate directly to the memory, without intermediaries.

Consider the circuitous path required to access main memory in a typical system today. The processor connects to a system bus running at a fraction of the CPU speed. A few CPU cycles are typically lost synchronizing with the slower bus. The processor must then arbitrate for this bus and send the address to the memory controller. After receiving and processing the address, the memory controller reads the requested data from the DRAM array. The memory controller must then process the data and transmit it back to the processor across the slow system bus before the CPU can use it.

All of these activities—arbitration, transmission, processing, reception—occur at the system-bus speed, which may be one-third to one-fifth of the CPU speed. While each of these overhead activities may take only a few bus cycles, they can easily add up to dozens of CPU cycles. Even an out-of-order processor will quickly come to a halt during such a lengthy delay.

Things get even worse in multiprocessor systems. Most of today's MP designs share a single bus among among 4, 8, or even 16 processors and one or more memory controllers. If that bus is busy when a processor needs data, that processor must wait for its turn, extending its stall. For these reasons, transaction-processing performance often improves by only small amounts when more processors or faster processors are added to a system without improving the bus or memory bandwidth.

Putting the "Direct" in RDRAM

Although Intel has been the most vocal proponent of the new Direct RDRAM technology from Rambus, the x86 vendor will initially connect the RDRAMs to the system logic (north bridge) rather than the processor. This method maintains the current PC structure, simply substituting RDRAM for SDRAM.

The 21364, in contrast, will connect the new RDRAM directly to the microprocessor itself. This revolutionary approach bypasses the inefficiencies of the system logic while

taking advantage of the high bandwidth of the Rambus design. Compaq estimates the memory latency of the 21364 will be 90 ns; in contrast, the company's current TurboLaser servers require 240 ns. In fact, because up to 10 processors share the TurboLaser's main memory, any given processor often takes longer than 240 ns to receive data. In contrast, each 21364 processor will have its own local memory, reducing such conflicts.

For the 21364, however, not all memory accesses will be satisfied by the local memory. Because the total memory in the system is divided among several processors, a processor must occasionally access memory controlled by another processor. In fact, if software is not rewritten to concentrate accesses in the local memory, remote accesses will be fairly frequent.

Improving the latency of these remote accesses requires a high-bandwidth, low-latency connection between multiple processors in a system. Instead of using a single bus or crossbar, the 21364 designers opted for a set of high-speed point-to-point connections. This system allows 21364 processors to be connected in a mesh, as Figure 1 shows. Processor-to-processor latency is just 15 ns, or 30 ns per round-trip. Thus, accessing memory in one of four adjacent processors takes 120 ns, just 38% longer than accessing local memory.

In the 16-processor mesh shown in the figure, all but one of the processors can be accessed within three hops. Even the worst case of four hops takes 200 ns, still faster than the best-case access in today's TurboLaser system. Compaq estimates the average memory latency of a 21364 system running TPC-C will be less than half that of the current system.

Technically, this approach is a nonuniform memory access (NUMA) design. In a classic NUMA system, however, remote memory can take an order of magnitude longer to respond than local memory, forcing software to take these delays into account. The speed of the 21364's interconnect keeps memory latency within a factor of two or three, simulating a standard shared-memory model. NUMA-optimized software will see a small but tangible performance gain.

Based on Existing 21264 CPU

Taking Dick Sites' exhortation literally, the 21364 team is spending all its efforts on memory subsystem design and essentially none on CPU design. The team plans to reuse the 21264 CPU core (see MPR 10/28/96, p. 11), leveraging as much of the physical design as possible.

The 21264 is a four-issue superscalar machine that can reorder up to 80 instructions at a time, more than any other processor announced to date. To facilitate high clock speeds, the CPU is unique in dividing the integer units into two clusters, each with its own copy of the register file. The chip also uses a unique branch predictor that combines two distinct prediction schemes.

The only significant change contemplated for the CPU core is a slight modification of the branch predictor. The team's experience with the 21264 has identified some minor changes that could improve the prediction accuracy. Otherwise, the core will be left essentially alone.

Other "core" changes actually involve buffers around the outside of the core. The 21264 can buffer up to 8 L1 miss requests at a time; the 21364 will extend this to 16 requests. Similarly, the victim buffer, which holds dirty cache lines on their way back to main memory, will be increased from 8 to 32 entries. The new chip's 32 entries will be divided into 16 for the L1 data cache and 16 for the L2 cache.

Massive On-Chip Cache

The 21264 core was originally designed for a 0.35-micron process, in which the chip measures 298 mm^2. The 21364, however, will be built in a 0.18-micron process, which boosts the CPU speed from 600 MHz to 1 GHz and reduces the size of the core to about 100 mm^2. This reduction creates plenty of room to add more stuff.

Specifically, the new stuff includes a six-way-associative 1.5M on-chip L2 cache, as Figure 2 shows. This cache will cycle at the speed of the CPU, delivering 128 bits of data every nanosecond. At this speed, the bandwidth to the L2 cache is a stunning 16 Gbytes/s, well beyond the impressive 4 Gbytes/s achieved by the 21264. Compaq could take better advantage of the on-die cache by increasing the width of the interface, perhaps delivering a full cache line per cycle, but this would require modifying the CPU core.

The access time of the L2 cache is a leisurely 12 cycles. This is due in part to the team's reluctance to change the 21264 core, which allows this many cycles to access its off-chip L2 cache. The long access time also enables the tag lookup to occur before the data access. By taking the time to locate the correct tag, the chip needs to power only a single set and sector within the cache array on each cycle, greatly reducing the amount of power dissipated by the cache.

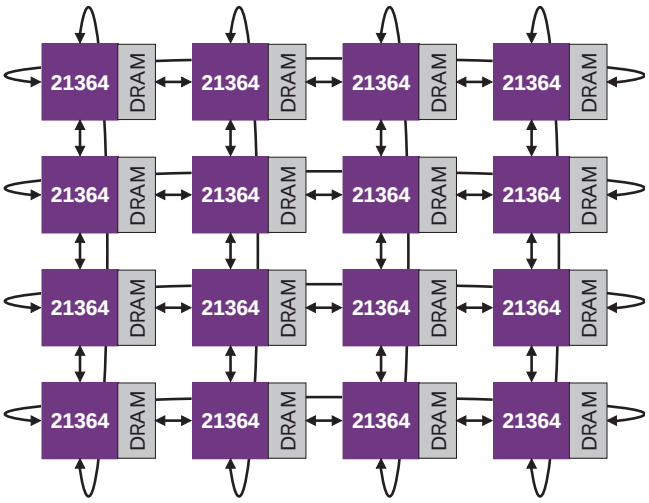


Figure 1. A large 21364 system will consist of a mesh of interconnected processors, each with its own local memory. In this 16-CPU example, any processor can access any other processor's memory with no more than four "hops."

Because the chip is estimated to consume a blazing 100 W, even with this power-saving feature, driving the entire cache every cycle would be impractical.

With two levels of on-chip cache, the 21364 harkens back to the 21164 (see MPR 9/12/94, p. 1) in its cache design. Digital itself repudiated this two-level design when it introduced the 21264, saying that the 21164's caches were too small, the L2 cache was too slow, and the split-level scheme created needless overhead.

The 21364 design addresses many of these issues. Due to the limitations of the 21164's 0.5-micron process, that chip sports only 8K primary caches and 96K of L2 cache. The 21364 will have 8 times more primary cache and 16 times more secondary cache. With only 112K of on-chip cache, the 21164 required an external L3 cache for reasonable performance. In contrast, the 21364's combination of a large L2 and a fast path to main memory eliminates the need for an L3 cache. Thus, the 21364 has the same number of cache levels, and the same overhead, as most other processors.

One potential problem is the L2 access time. Whereas the 21164 was thought to be deficient with a six-cycle access, the 21364 requires twice that. Even Intel's Mendocino (see MPR 8/24/98, p. 1), not exactly a paragon of performance, needs only eight cycles to access its on-chip L2 cache. Its impressive 80-entry reorder buffer notwithstanding, the 21364 is unlikely to queue enough instructions to avoid stalling during a 12-cycle L1 cache miss.

Improving the L2-cache latency, or even eliminating the L2 in favor of large primary caches, would have required opening the 21264 core design, which Compaq did not want to do. Granted, circuit design at 1 GHz is challenging, but the long L2 latency will certainly impact performance.

The data array for this massive cache consumes about 150 mm^2 in the 0.18-micron process; including tags and control logic, the complete cache measures nearly 200 mm^2. To improve yield, the data array is protected by redundant rows and columns. The cache array also contains enough extra bits for full ECC protection, correcting spontaneous single-bit errors.

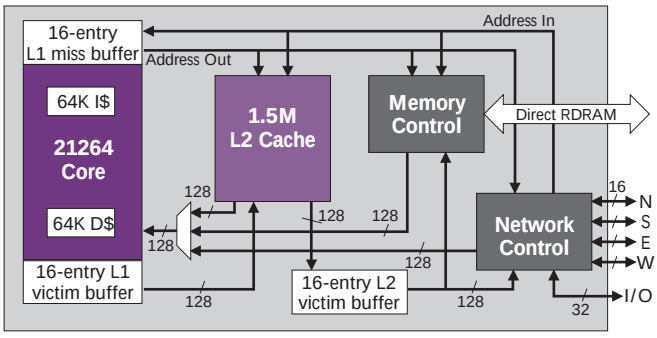


Figure 2. The 21364 combines the 21264 CPU core with a large L2 cache, a Rambus memory controller, and a network interface that connects the chip to four other processors (North, South, East, and West) and to I/O devices.

How Many Direct RDRAM Channels ... ?

How many Direct RDRAM channels does it take to satisfy a 1-GHz processor? Apparently, plenty. Although Compaq's Bannon coyly refused to enumerate the channels, he said the total bandwidth to main memory will be 6.0 Gbytes/s, which corresponds to four standard Direct RDRAM channels. The 21264, the current bandwidth leader, supports a maximum of 2.6 Gbytes/s, although the initial systems deliver only about 1.6 Gbytes/s.

One Direct RDRAM channel can sustain 1.6 Gbytes/s of read bandwidth, or about 1.5 Gbytes/s with a random mix of reads and writes (see MPR 10/27/97, p. 25). The channel has 16 data lines operating at 800 Mtransfers/s, using both edges of a 400-MHz clock. Counting the necessary address, control, power, and ground pins, a single channel requires 45 pins.

Compaq has licensed the necessary interface logic from Rambus but is designing its own memory controller. Unlike the simple Rambus memory controller that will be used in PCs, the 21364 controller can monitor the full 128 pages that can be open on each channel. As a result, the Alpha chip can maintain more than 500 open pages at once, reducing memory latency on accesses to those pages.

With four channels, the 21364 will support up to 1G of main memory with the initial 64-Mbit RDRAMs. This limit is small for a high-end server, but remember that each CPU has its own memory, so a 16-processor server could hold up to 16G. Future 256-Mbit RDRAMs will allow up to 4G per 21364, and even with the smaller memory chips, repeaters can be used, if necessary, to increase memory capacity.

All of this memory is, of course, protected by ECC. Furthermore, Compaq says it has figured out how to cross-connect the RDRAMs to handle single-chip failures without bringing down the memory system. This feature will not be available in standard PC servers, as Compaq is not licensing its solution.

One problem with a distributed memory system is cache coherency. Without a central memory bus, processors cannot snoop other transactions and substitute their own data if necessary. The 21364 will implement a directory-based coherency scheme that uses part of the local memory to store the standard MESI coherency information. If a block of memory is shared among processors, they must use the interprocessor connections to update each other when data changes.

Processors Communicate Directly

The 21364 will have four interprocessor ports. Each port consists of two 16-bit connections that operate at up to 800 Mtransfers/s, similar to Direct RDRAM. To facilitate high-speed operation, the connections are unidirectional and point-to-point, and they use low-voltage signaling. The ports are source synchronous and do not need to be synchronized across the entire system, simplifying clock distribution.

Communication consists of packets that each contain a header (indicating the destination node, memory address

within the node, and other information) and 128 bits of data. Derating the peak bandwidth to account for the overhead leaves about 1.2 Gbytes/s of data bandwidth in each direction. Thus, the total bandwidth of the four ports is nearly 10 Gbytes/s (assuming all ports are transmitting and receiving at the same time).

The network controller, shown in Figure 2, handles all incoming and outgoing data. It automatically forwards packets that are not addressed for that node, so the CPU doesn't need to get involved. The network controller assumes that the processors are arranged in a mesh; although other arrangements could have been supported, this restriction simplifies the routing circuitry.

Because of the variable distance to other processors in the mesh, data may not be returned in the order in which it was requested. The out-of-order CPU core easily accommodates this situation.

The 21364 includes a fifth port for connecting to an I/O ASIC. This port also consists of two unidirectional buses, but these are 32 bits wide. The wider buses allow clock speeds of around 200 MHz, simplifying the ASIC design, while still delivering a total of about 3 Gbytes/s. This ASIC can bridge to PCI or other appropriate I/O buses. Note that in a multiprocessor server, only one or two processors might have I/O devices connected; the others can access these I/O devices via the interprocessor network.

Big Chip With Big Appetites

As a high-end processor, there isn't anything small about this chip. Even in a 0.18-micron process, the die size will be about 350 mm^2; since the chip is so far from tapeout, the exact size is yet to be determined. Yield will be slightly better than a standard processor of that size, as nearly half the chip consists of the L2 cache array that is protected (at least partially) by redundancy.

To contain the four Direct RDRAM ports, four interprocessor ports, the I/O port, and enough power and ground pins to source nearly 70 A of current at 1.5 V, the 21364 will require a package with roughly 1,000 pins. Again, the exact number will be determined when the design is finalized. The expensive package contributes to an estimated manufacturing cost of $380, according to the MDR Cost Model.

As part of the Digital dissolution (see MPR 11/17/97, p. 1), Intel is obligated to manufacture Alpha processors for Compaq. The Alpha team, however, is not satisfied with Intel's 0.18-micron process, which is rather vanilla (see MPR 9/14/98, p. 1). Samsung, the other current Alpha fab, has recently discussed adding copper and SOI (see MPR 8/24/98, p. 8) to its 0.18-micron process; Intel's process includes neither of these features.

Compaq says only that it is investigating various options for fabbing the 21364. Using Intel's process, the company believes it can attain the rated 1-GHz clock speed. In a more aggressive process, the Alpha chip could be as much as 20–30% faster.

Price & Availability

Compaq sells Alpha systems but not Alpha processors. Alpha processors are available from Samsung, but that vendor did not quote price or availability for the 21364. We do not expect volume shipments of the 21364 to occur before 4Q00. For more information on the 21364, access *www.digital.com/semiconductor/alpha*.

Head to Head Against Merced

Given that the 21364 is still two years away from shipments, comparing it to its competition is difficult. The Alpha chip's feature set and clock speed, however, are quite impressive, particularly for high-end servers. Sun's UltraSparc-3 (see MPR 10/27/97, p. 29) also includes an on-chip memory controller, gaining the advantage of improved memory latency. Due to its reliance on standard SDRAM, however, that processor has less than half the memory bandwidth of the 21364. The SPARC chip has roughly the same amount of cache-coherency bandwidth but is not optimized for large meshes of processors.

UltraSparc-3 is due in late 1999, and by the time the 21364 emerges, Sun plans to deploy UltraSparc-4, which is due to hit 750 MHz. If both chips hit their clock-speed and performance goals, the 21364 is likely to deliver more performance than UltraSparc-4 in both workstation and server configurations. Judging by the current performance of the 21264, a 1-GHz 21264 should reach about 50 SPECint95 and 80 SPECfp95 (base); the 21364's faster memory interface should improve these scores slightly. Sun's future performance is more questionable, as the company has yet to demonstrate the Cheetah core that will drive UltraSparc-3 and UltraSparc-4.

The 21364 may face tough competition from Merced (see MPR 10/26/98, p. 16). Certainly, the IA-64 chip will be used by many more vendors, including Compaq itself. But Merced will be hard pressed to exceed the SPEC ratings planned for the 21364. Furthermore, the Intel chip has a more traditional system interface with memory bandwidth less than half that of the 21364, so it isn't likely to perform as well in large MP configurations.

Intel is developing a second IA-64 processor, codenamed McKinley, scheduled for late 2001 shipments. Intel says this part will have much better performance and much more bus bandwidth than Merced. Compaq plans to counterattack with a new Alpha core, known as EV8, or Araña, that is rumored to be multithreaded.

The competition between IA-64 and Alpha should be quite interesting. While Intel is concentrating on a new instruction set and new processor cores, the Alpha team has chosen to focus on the system interface. The fastest CPU in the world won't shine without improved memory latency and bandwidth. The 21364 design shows that the Alpha team has learned this lesson well. M