Power4 Focuses on Memory Bandwidth

IBM Confronts IA-64, Says ISA Not Important

by Keith Diefendorff

Not content to wrap sheet metal around Intel microprocessors for its future server business, IBM is developing a processor it hopes will fend off the IA-64 juggernaut. Speaking at this week's Microprocessor Forum, chief architect Jim Kahle described IBM's monster 170-million-transistor Power4 chip, which boasts two 64-bit 1-GHz five-issue superscalar cores, a triple-level cache hierarchy, a 10-GByte/s main-memory interface, and a 45-GByte/s multiprocessor interface, as Figure 1 shows. Kahle said that IBM will see first silicon on Power4 in 1Q00, and systems will begin shipping in 2H01.

No Holds Barred

On this project, Big Blue is sparing no expense. The company has brought together its most talented engineers, its most advanced process (0.18-micron copper silicon-on-insulator), and its best packaging, reliability, and system-design know-how. The sheer scale of the project indicates that IBM is mindful of the threat posed by IA-64 (see MPR 5/31/99, p. 1) and signals that the company is prepared to fight for the server market that it considers its birthright.

After years of building their own processors, IBM, HP, and others have been forced to watch as systems based on commodity Intel microprocessors have chipped away at their market. HP recognized the futility of continued resistance and threw in the towel. But IBM sees that with more and more of the critical system-performance features moving onto the processor, the loss of control over the processor silicon would rob it of the ability to assert its superior technology and to differentiate itself from the pack.

Although the IBM PC Company has already elected to go with IA-64 for its Netfinity servers, IBM apparently believes it cannot strategically afford to do the same for its high-end (high-margin) server businesses, where it makes a large portion of its revenues today and which it expects will grow rapidly along with the Internet. Therefore, the company has decided to make a last-gasp effort to retain control of its high-end server silicon by throwing its considerable financial and technical weight behind Power4.

After investing this much effort in Power4, if IBM fails to deliver a server processor with compelling advantages over the best IA-64 processors, it will be left with little alternative but to capitulate. If Power4 fails, it will also be a clear indication to Sun, Compaq, and others that are bucking IA-64, that the days of proprietary CPUs are numbered. But IBM intends to resist mightily, and, based on what the company has disclosed about Power4 so far, it may just succeed.

Looking for Parallelism in All the Right Places

With Power4, IBM is targeting the high-reliability servers that will power future e-businesses. The company has said that Power4 was designed and optimized primarily for servers but that it will be more than adequate for workstation duty as well. The market IBM apparently seeks starts just above small PC-based servers and runs all the way up through the high-end high-availability enterprise servers that run massive commercial and technical workloads for corporations and governments.

Much to IBM's chagrin, Intel and HP have also aimed IA-64 at servers and workstations. IA-64 system vendors such as HP and SGI have their sights set as high up the server scale as IBM does, so there is clearly a large overlap between the markets all these companies covet. Given this, it is surprising that they have come to such completely different technical solutions.

Intel and HP have concluded there is still much performance to be found in instruction-level parallelism (ILP). Hence, they have mounted an enormous effort to define a new parallel instruction-set architecture (ISA) to exploit it (see MPR 5/31/99, p. 1). Evidently, they expect a significant speedup from machines that can issue six or more instructions per cycle (any less wouldn't justify a new ISA).

IBM, in contrast, believes the place to find parallelism in server code is not at the instruction level but at the thread level and above. It doesn't believe there's enough ILP in individual threads of server code to fill a large number of instruction-issue slots. Even if there were, IBM says that EPIC-style architectures like IA-64 are contraindicated. Although high-ILP processors may reduce processor busy time, IBM points out that they do nothing to reduce processor wait time, which is the far larger problem. In fact, it says EPIC architectures exacerbate this problem by burdening the

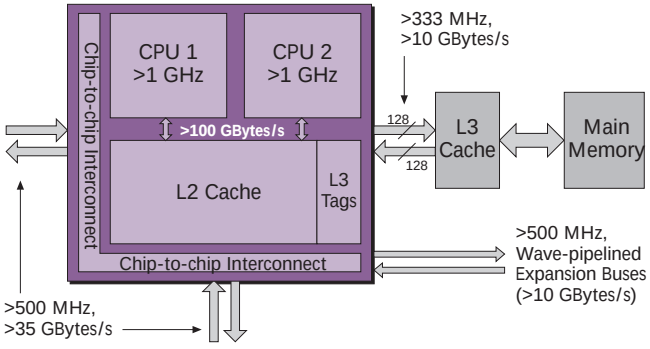


Figure 1. Power4 includes two >1-GHz superscalar cores with more than 100 GBytes/s of bandwidth to a large shared-L2 cache and more than 55 GBytes/s of bandwidth to memory and other Power4 chips.

memory system with a large number of conditionally executed instructions that are eventually discarded.

Dynamic Scheduling Is Better, Says IBM

Power4 engineers cite a number of arguments in favor of dynamic scheduling over EPIC-style static scheduling for servers. One issue is cache misses; dynamic machines constantly remake the instruction schedule, thereby avoiding many pipeline stalls on cache misses. EPIC machines, because of their in-order execution and static instruction groupings, are less adaptive. EPIC does allow the compiler more freedom to boost loads, and a register scoreboard like the one in Merced allows some run-time adjustments, but cache misses can be hard to predict at compile time and EPIC machines will generally take less advantage of run-time information than reordering superscalar machines.

Another issue IBM raises is the impracticality of code profiling. According to IBM, profiling large server applications is often difficult, and the results not that valuable. But EPIC compilers rely heavily on profiling information to schedule predication and speculation. Wen-Mei Hwu, speaking at last year's Microprocessor Forum, spelled out several other EPIC-compiler challenges. IBM believes many of these will not be solved for a long time.

If EPIC compilers for traditional code are a challenge, dynamic just-in-time compilers (JITs) for Java will be a nightmare. EPIC compilers must search a large code window to discover ILP and must perform complex code transformations to exploit predication and speculation. Thus, EPIC compile time can be long, making it hard to amortize at run time. Java performance is a serious issue for IBM, which is committed to Java for server applications and has the second-largest cadre of Java programmers in the world, next to Sun. Sun probably agrees with IBM's concerns about EPIC, as its new MAJC architecture (see MPR 9/13/99, p. 12) has many features that are radically different from IA-64 for just these reasons.

IBM is also concerned that EPIC binaries are too tightly coupled to the machine organization. Although Intel and HP have taken steps to ensure that IA-64 code will function across generations, IBM says that an EPIC instruction schedule is so dependent on the machine organization that, in practice, it will restrict hardware evolution.

But IBM's primary objection to EPIC isn't that it's bad, it's just that it's so unnecessary. IBM sees no difficulty in building dynamically scheduled processors that can exploit most of the ILP in the vast majority of server applications. It also sees no difficulty—now or in the future—in building dynamically scheduled POWER processors that can fully tax any practical memory system. Therefore, IBM concludes that the memory system is the real determinant of server performance, not the instruction set. Thus, staying with POWER imposes no real penalty and avoids a pointless ISA transition.

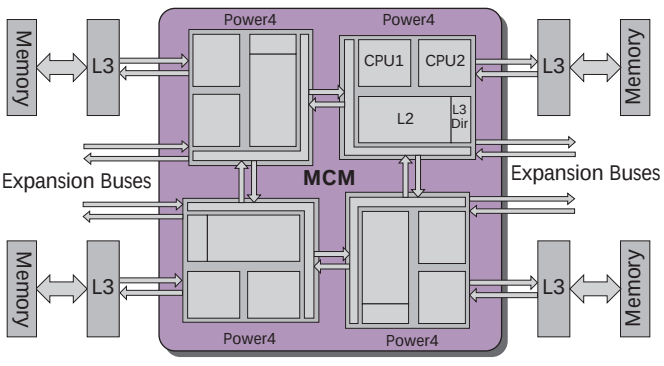


Figure 2. Four Power4 chips will be offered in a single MCM package as an eight-processor SMP with total bandwidths of 40 GBytes/s to memory and 40 GBytes/s to other modules.

Chip-to-Chip Interconnect Shares L2

As a result, IBM has focused on system design rather than on instruction-set design. The technology, and most of the silicon, in a Power4 chip is dedicated to delivering data to a large number of processors as quickly as possible. The key element IBM uses to accomplish the task is the shared L2 cache. Power4's on-chip L2 is shared directly by the two on-chip processors and by processors on other chips via a high-speed chip-to-chip interconnect network, as Figure 2 shows.

Details on the physical structure of the network have not yet been disclosed, pending patent applications. Kahle did, however, describe some of its features. The network logically appears to each processor as a simple low-latency bus, while the actual physical network provides the high bandwidth and nearly contention-free throughput of a full crossbar switch, but without the complexity.

The chip-to-chip data paths shown in Figure 2 each include multiple 16-byte-wide point-to-point buses arranged in a ring-like topology that IBM describes only as a distributed switch. The switch is implemented entirely on the Power4 die, with no external chips required.

Physically, each chip-to-chip bus is unidirectional and operates on a synchronous latch-to-latch protocol. The low-voltage signals transfer data at a rate of over 500 MHz, giving each Power4 chip an aggregate sustainable chip-to-chip bandwidth of over 35 GBytes/s. Such high bandwidth keeps the network utilization low, which, according to queuing theory, minimizes network latency. The bus architecture is designed so that when four Power4 chips are located in close proximity and each die rotated 90˚, the buses between chips route directly. This keeps the wires very short and therefore allows the buses to be very wide and very fast.

As Figure 3 shows, the shared-L2 cache is divided into three multiported, independently accessible slices. A 100-GByte/s switch connects the L2 slices to the on-chip processors as well as to off-chip processors through the chip-to-chip interconnect ports. A shared-intervention protocol is used to enforce cache coherence and to move data into the L2 on the chip that used it last. The goal of the design is to get the right data into the right L2 at the right time and, from a coherency perspective, make sure it is safe to use.

IBM has not disclosed the size of the L2 cache on each Power4 chip, but, based on 170 million transistors and the floor plan in Figure 3, we estimate that the L2 is about 1.5M. We also expect it to be at least eight-way set-associative, as IBM rarely builds on-die cache of less. Due the large size of the L2 and the reliability requirements for high-availability servers, the L2 is protected from manufacturing defects by row and column redundancy and protected from run-time soft errors by ECC.

A Memory Bandwidth Behemoth

Each Power4 chip provides an L3-cache port separate from the chip-to-chip ports. The L3 port is 16 bytes wide in each direction and operates at a 3:1 clock ratio, providing over 10 GBytes/s of memory bandwidth. The L3 cache tags are kept on the processor die so cache coherency actions can take place at on-chip cache speeds. From the size of the L3 directory shown in Figure 3, we estimate that each Power4 chip can support up to 32M of external L3 cache.

IBM did not describe the L3 architecture, but Figure 2 shows it to be an inline design. This application is a perfect fit for IBM's embedded-DRAM process, which the company has used before to construct integrated-cache chips. With its latest 0.18-micron CMOS-7SF merged-logic/DRAM process, IBM could easily construct a very large set-associative ECC cache with a high-speed interface to the Power4 chip and an interleaved ECC memory controller to drive the main-memory DRAMs.

To help convert Power4's copious memory bandwidth into low-latency memory accesses, the chip implements eight software-activated prefetch streams. These prefetch streams use spare bandwidth to continuously move data through the memory hierarchy and into the L1. Up to 20 cache lines can be kept in flight at a time. Once the prefetch pipe is filled, the memory system can theoretically deliver new data from main memory to the core every cycle.

Chip Multiprocessing Boosts SMP Performance

Placing its bet behind the theory that the most important parallelism in server workloads is above the instruction level, IBM has optimized the Power4 system for shared-memory symmetric-multiprocessing (SMP) performance, as opposed to uniprocessor performance. Instead of spending its transistors on a single monolithic CPU, IBM has opted for two smaller CPUs on each Power4 chip.

The theory is this: above some point, say four instructions per cycle, ILP becomes hard to find, leading to diminishing returns on transistors spent to recover it. This implies that a single monolithic CPU will not scale linearly with transistor count. On the other hand, with efficient data sharing, two processors can be made to scale almost linearly, at least when there are enough independent threads available to keep both cores busy, which is usually the case with server workloads. Thus, for a given transistor budget, two smaller CPUs should outperform one big one.

The key is efficient data sharing, which is what Power4 is all about. The latency and bandwidth between on-chip CPUs and a shared multiported L2 cache can be many times what is achievable with discrete CPUs. For discrete CPUs with separate on-chip L2 caches, shared data must be shuffled between chips across external wires. For discrete CPUs with an external shared L2, every L2 access from both CPUs goes off chip.

In either case, to match the speed of on-chip data sharing, the discrete CPUs would require external buses that are far wider and faster than physics allows. For any given number of wires connecting processors, higher levels of SMP can be achieved with two cores on a chip than with one core. Furthermore, containing all the memory traffic between two CPUs and their L2 on a chip takes an enormous load off the external buses, simplifying the chip-to-chip interconnect.

If this theory is valid, it alone would be enough to justify the chip multiprocessing (CMP) approach IBM has taken with Power4. But CMP has secondary benefits as well. For one, a small simple CPU will generally run at higher clock rates than a large complex one. For another, it is easier to design and replicate a simple CPU than it is to design a complex one.

"Simple CPU" Is a Relative Term

For the CMP approach to work, each CPU must be powerful enough to exploit most of the ILP that exists in single threads. Although IBM is not ready to release details of the Power4 CPU microarchitecture, it has given a few clues to suggest that each of Power4's two CPUs will exceed the power of any single microprocessor that exists today.

From the floor plan shown in Figure 3 and the transistor count, we estimate that each CPU core (including L1 caches) contains about 30 million transistors, three times as

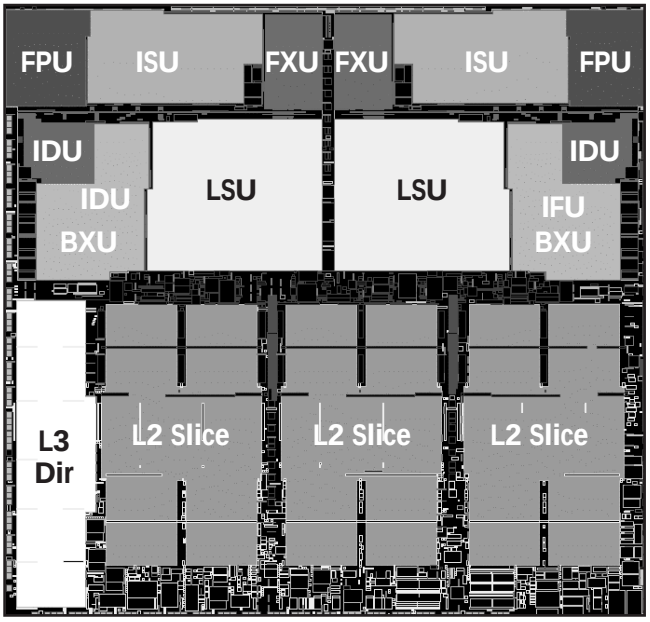
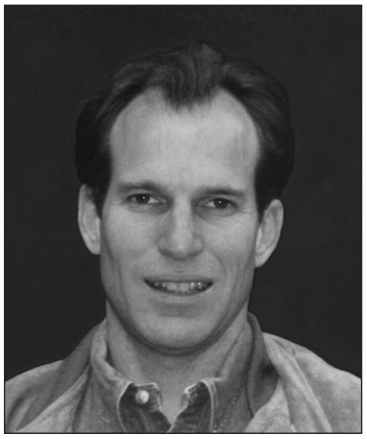


Figure 3. With 170 million transistors, a Power4 chip will occupy about 400 mm^2 in IBM's seven-layer-metal 0.18-micron CMOS-8S2SOI process, as this floor plan shows.

many as in Pentium III. In addition, each Power4 CPU will run at "over 1 GHz," which probably means at least 1.1 GHz. To achieve these frequencies, IBM set a design goal of 8 to 10 gate delays between pipeline stages, which, for a RISC-style ISA, probably indicates an integer pipeline of about 10 stages and a load pipeline of about 12; IBM has not confirmed these estimates. We expect each Power4 CPU to be like Power3 and have two fully pipelined double-precision floating-point multiply-add units and two complete load/store units.

Even though IBM disdains IA-64's EPIC approach, it appears to be stealing a page from Intel's playbook. In the same way that Intel usurped RISC principles to implement its x86 CISC architecture in P6, IBM plans to expropriate VLIW principles to implement its RISC architecture in Power4.

IBM only vaguely described the mechanism, but apparently in the early stages of the pipeline, the Power4 CPU groups instructions into VLIW-like bundles. These bundles are dispatched to issue queues, where individual instructions are held until their dependencies are resolved and then issued to the execution units. The pipeline beyond the issue stage is noninterlocked; so, once issued, nothing stops an instruction from completing, but all instructions in a bundle must complete before the bundle is retired.



Jim Kahle, chief architect of Power4, described its high-bandwidth interface at the Forum.

Unlike conventional superscalar implementations that track individual instructions from dispatch through completion, the Power4 CPU tracks bundles only. According to IBM, this mechanism, along with data-flow sequencing through the noninterlocked pipelines, dramatically simplified the Power4 implementation, cutting the percentage of control logic in half compared with that of the four-issue Power3 design (see MPR 11/17/97, p. 23). This brought the control complexity of Power4 more in line with that of a VLIW machine while preserving the advantages of dynamic scheduling.

IBM said that the out-of-order-completion resources in the Power4 CPU are deep enough to hide the full latency of an L2 cache hit, which is probably 8–10 cycles. Also, to a greater extent than on any previous Power or PowerPC processor, Power4 will exploit the architecturally specified weak-storage-ordering model to reorder memory transactions and hide memory latency.

Layering for Frequency

Each Power4 CPU implements the same ISA as IBM's current RS/6000 and AS/400 systems and is also fully PowerPC compatible. IBM did, however, make some improvements that will be invisible to programs. The company is finally acknowledging that some of the complex instructions retained from the original 1990 POWER definition may not have been such great ideas. These instructions hinder the ability to run dynamically scheduled wide-issue processors at high frequency.

Convinced, however, that instruction-set stability is critical to its customer base, IBM didn't take the radical step of expunging these instructions from the ISA. Instead, it has introduced instruction-set layering into Power4. In this strategy, the hardware is optimized for the simple instructions, making no frequency compromises for complex ones. Slightly complex instructions, such as the base-register-update form of loads and stores, are cracked into two simple instructions by the instruction decoders. Moderately complex instructions, such as the string ops, are executed by a simple nonbranching microcode engine. The most complex instructions, such as the old POWER instructions that were removed in PowerPC, trap to software emulation routines. In this way, existing binaries run unmodified, but new binaries created by compilers aware of the layering may run faster by exploiting the faster alternatives.

Systems of All Sizes

The dual-CPU Power4 chip will serve as the basic building block of a wide range of RS/6000 and AS/400 server systems. The first systems will probably be eight-way SMPs built with four Power4 chips mounted on a multichip module (MCM), as Figure 4 shows. This design point is the sweet spot for Power4 chips, as it utilizes most of the chips' features in their most optimal configuration and balance.

The MCM, designed by IBM for Power4 systems, is not your garden-variety MCM. Since, according to our calculations, each 1.5-V Power4 chip will dissipate over 125 W, the MCM has to dissipate over half a kilowatt. It must also deliver 350 A of noise-free current and transmit thousands of 500-MHz signals among Power4 chips and out to memory.

The solution is a multilayer glass-ceramic substrate with copper interconnect layers. Glass ceramic provides a dielectric constant (k) of about 5, 45% lower than conventional alumina-ceramic (Al_2O_3) substrates ($k \approx 9$). The copper interconnect layers offer significantly lower resistance than the refractory-metal layers (tungsten or molybdenum) used in alumina-ceramic packages.

The processor die are flip-chip mounted into the MCM with a staggering 5,500 100-µm C4 solder balls spaced on 200-µm centers. Of the 5,500 connections, approximately 2,200 are signal I/Os; the rest provide power and ground. An advanced direct-attach technique improves heat transfer from the silicon to the MCM-package substrate.

As Figure 4 shows, the MCM is mounted on a massive metal carrier that physically attaches it to the motherboard and to its air-cooled heat sink. Since the land-grid-array style package is too large and too expensive to be reflow soldered,

we suspect IBM may be using the metallized-particle interconnects (MPI) offered commercially by Thomas & Betts or the CIN::APSE fuzz-button connectors offered by Cinch.

These types of connectors can require as much as 60 grams of force per pad to make reliable electrical contact across such a large package. Thus, with 5,200 pads, the MCM would require a total of about 700 pounds of force to insert. This may explain the thickness of the metal carrier, which must be extremely flat and rigid to evenly distribute that much force while maintaining the necessary planarity. (MPI connectors have a compliance of about 250 microns.)

Elastic I/O Connects MCMs

Each Power4 chip has two 16-byte-wide L3/memory buses as well as multiple expansion buses that are routed off the MCM through approximately 3,400 signal pads. The expansion buses, among other things, allow multiple MCMs to be connected together to form larger systems.

IBM calls its expansion buses elastic I/O, due to their unique ability to decouple latency from bandwidth. With traditional buses, the maximum bandwidth of the channel is determined by its latency, which is limited by the end-to-end channel delay and by the worst-case timing skew across the width of the channel. But IBM's elastic I/O uses a low-voltage source-synchronous wave-pipelining technique with per-bit de-skew to eliminate the dependence on channel latency. With IBM's scheme, multiple bits are kept in flight on each wire at the same time, and the per-bit de-skew allows arbitrarily wide buses to operate at high clock frequencies.

The two eight-byte-wide intermodule buses operate at more than 500 MHz, giving each chip a bandwidth of about 8 GBytes/s for a total of about 32 GBytes/s between modules. This bandwidth is probably sufficient to build a four-MCM SMP (32-processor) system with memory-access times sufficiently uniform to support classical SMP workloads without retuning the software for nonuniform memory access (NUMA). In addition to the intermodule buses, the expansion buses include separate buses for I/O and NUMA, bringing the bandwidth of each chip's expansion buses above 10 GBytes/s.

Primarily due to shared-memory bandwidth constraints, neither Power4's nor any other known technology will allow SMP systems to scale beyond a few dozen processors. For applications, such as transaction processing, that are amenable to software partitioning, larger Power4 systems can be constructed in NUMA configurations. Power4 chips have integrated support for large NUMA configurations as well as for IBM's logical partitioning (LPAR) feature, now also supported by Sun in its Enterprise 10000 systems. IBM envisions large Power4 NUMA nodes combined into even larger systems, using the clustering technology developed for its S/390 mainframes and its RS/6000SP multiprocessor systems.

Going the other direction in system size, IBM says it plans to offer the Power4 chip in a single-chip module for small dual-processor SMP servers. Presumably, it could also offer a single-processor system using partially good die. Partially good die is one more advantage of CMP construction. The redundancy of two identical CPUs can, in theory, be exploited to reduce manufacturing scrap, thereby reducing average manufacturing cost. This effect can be substantial for a large die, especially in a new, immature process. But IBM has given no indication it intends to exploit this capability.

All Hands to Battle Stations

"Power4" is actually somewhat of a misnomer. The name denotes a part that is simply the next-generation processor in the Power, Power2, Power3 series. But the name vastly understates the size and importance of this project to IBM. Previous Power chips were designed in relative isolation by the small RS/6000 group in Austin. Although viable products, these chips ran far below industry norms for clock frequencies, and the systems offered no compelling technical advantages. As a result, RS/6000 systems have slipped in market share against Sun, HP, and the myriad Xeon-based systems, disappearing almost completely from the workstation market.

Power4 is an entirely different beast, overpowering all previous Power projects. The only similarity between Power4 and its predecessors is the instruction set. The level of investment is of an entirely different order of magnitude. For Power4, the very best people and technology have been marshaled from every corner of the massive company.

High-frequency circuit-design methods were contributed by IBM Yorktown, which developed the techniques used to design the 637-MHz Alliance G6 mainframe microprocessor, until recently the highest-speed microprocessor shipping from any company. IBM Burlington developed the wave-pipelining technology for the expansion buses. The packaging technology was developed by experts with roots in IBM's Hudson Valley mainframe group. The RS/6000 group in Austin, working jointly with the AS/400 group in Rochester, did the system design. The CPU core was

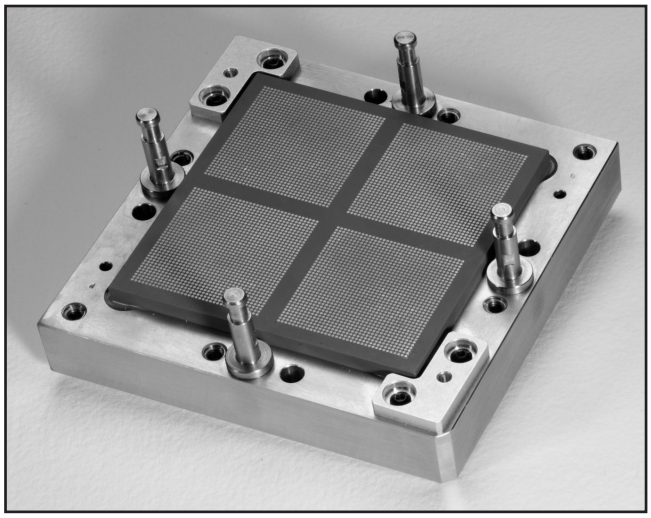


Figure 4. The Power4 MCM includes eight processors and some 5,200 I/O pads in a glass-ceramic package about 4.5" on a side.

developed by chip architects from the Power3 and Somerset groups in Austin, with help from IBM's Austin Research Labs and its T.J. Watson Research Labs in Yorktown.

Reliable All the Way Down to the Silicon

The CMOS-8S2SOI process was developed in IBM's East Fishkill process-development labs. This 1.5-V seven-layer-metal process is a variation of IBM's 0.18-micron copper CMOS-8S (see MPR 9/14/98, p. 1), which IBM will put into production later this year. The 8S2 derivative has 15% shorter channel lengths (L_g < 0.12 µm) and is built on a silicon-on-insulator (SOI) wafer (see MPR 8/24/98, p. 8). According to IBM, the low parasitic capacitance of SOI transistors boosts logic speed by over 25% compared with an equivalent bulk process, while also reducing power consumption.

A major constraint placed upon the development of CMOS-8S2SOI was very high reliability. Most processor manufacturers design their gate dielectrics to a Grade 3 failure-rate specification of 1,000 FITs (failures per billion hours). IBM, however, says this isn't good enough for duty in continuous-availability servers, because internal error-detection features extensive enough to compensate for IC-process-reliability problems would add cost and sacrifice considerable speed. As a result, IBM specifies its processes to a 10-FIT failure rate, two full orders of magnitude better than most companies.

To meet this stringent specification, the 8S2 gate oxide had to be made 3.6 nm thick (T_{ox} at 1.5 V), 20% thicker than the gate oxide in Intel's 0.18-micron 1.5-V P858 process (see MPR 1/25/99, p. 22), which it will use for Merced and McKinley. IBM had to develop other means to compensate for the losses of transistor drive current and of switching speed that result from the thicker gate oxide. SOI and copper were key to achieving these goals. Copper also improved the reliability of the on-chip interconnects; because the metal is nearly impervious to electromigration, it can sustain higher currents for longer periods without failing.

Even with this level of processes reliability, IBM still included a number of RAS (reliability, availability, and serviceability) features in Power4. IBM isn't ready to reveal all of Power4's RAS features, but it did confirm that the part has traditional features such as ECC on the L2, L3, and main memory. It also said that the Power4 has an independent on-chip full-speed test processor and logic analyzer that can be used during manufacturing and system operation to verify functionality and isolate failures. External testers are simply not viable for gigahertz chips with the amount of on-board logic, memory, and I/O that Power4 has.

Systems Still a Long Way Off

Although Power4 looks good at this point, a lot can happen between now and system shipments. Even though IBM feels it has invested enough in Power4 to ensure its success, the company is not invulnerable to technical glitches. IBM has, however, taken a number of risk-management steps, including the fabrication of a large test chip to validate Power4's critical technologies. IBM reported on that chip at this summer's Hot Chips. The company has also scheduled more than ample time between first silicon, due 1Q00, and system shipments, scheduled for 2H01. As a result, technical risk probably isn't IBM's biggest concern.

Cost is also not an issue. In CMOS-8S2, 170 million transistors, half of them cache, should fit on a 400-mm^2 die. While large, such a die is manufacturable for IBM; it is actually 15% smaller than HP's current PA-8500 (475 mm^2), which has the same amount of cache. Even assuming $400 for the MCM and conservative estimates of defect density and wafer costs, the MDR Cost Model projects a manufacturing cost of under $2,500, hardly unreasonable for an eight-processor module. Besides, in large servers the leverage of the CPU is so enormous that price is rarely an issue.

The real issue for IBM is competition. Compared with today's server microprocessors, of course, there is no contest. Even next year's Foster, Merced, UltraSparc-4, and 21364 aren't likely to be a match for Power4. The real challenge will come from the next generations of these processors, which are due out in late 2001 or 2002. Unfortunately, not enough is publicly known about them to make solid comparisons.

Today, Sun is the most direct competitor for IBM's server business. In the past, Sun has thrived, despite relatively low performance processors, by concentrating on high memory bandwidth and robust multiprocessor systems. With Power4, however, IBM may have Sun outgunned, as it is difficult to imagine anyone creating a system with much higher bandwidth than Power4. If Sun can deliver its 1.5-GHz UltraSparc-5 in late 2001, as planned, it might compete with Power4, but there is some question about Texas Instrument's (Sun's UltraSparc foundry) desire to match IBM's leading-edge IC processes, given its own focus on low-cost DSPs.

Performancewise, Compaq's Alpha processors are everyone's most feared competitor. The current 667-MHz four-issue out-of-order 21264 is the industry's performance leader. By the time Power4 arrives, the 21264 will have been replaced by the 21364 (see MPR 10/26/98, p. 12). This part will use the 21264 core but boost frequency to 1 GHz with a 0.18-micron process, add a 1.5M on-chip L2, a 6-GByte/s memory port, and 13 GBytes/s of chip-to-chip bandwidth.

In some ways, the system architecture of the 21364 is similar to Power4's. Both employ out-of-order superscalar microarchitecture, large on-chip caches, a dedicated memory port, and a high-speed point-to-point interconnect network between chips. The 21364, however, doesn't offer chip-level multiprocessing, and the topology of the interconnect network is different. The 21364's flat mesh has an elegant symmetry, but it doesn't match Power4's raw bandwidth numbers. Since the topologies are different, however, the bandwidth numbers are difficult to compare.

The 21464, due out sometime in 2002, will be a multithreaded version of a new core, designed to exploit the thread-level parallelism (TLP) that Power4 exploits with on-chip multiprocessing. CMP and multithreading each have

advantages and disadvantages, and it will be interesting to see which approach offers better performance. This assumes, of course, that Compaq will remain committed to Alpha after Merced and McKinley ship, and that it can find a fab capable of matching IBM's.

Battle With IA-64 Takes Shape

The most serious competition will surely come from IA-64, not just in HP systems but also from the collective mass of other server vendors that have lined up behind that architecture. The first IA-64 processor, Merced (see MPR 10/6/99, p. 1), will ship in systems starting in 2H00 and will still be the prevailing IA-64 processor when Power4 arrives in 2H01. Merced is a single six-issue sub-gigahertz processor with a small on-chip L2 and less than a tenth of Power4's chip-to-chip bandwidth, so it isn't likely to match that chip's server performance.

Power4's first real IA-64 challenge will come from McKinley, due in late 2001. Intel and HP say that McKinley will be far superior to Merced. According to some sources, McKinley will run at 1.2 GHz and deliver twice the performance and three times the bandwidth of Merced. McKinley may outrun Power4 on single-thread benchmarks, bit it lacks CMP and presumably has far less system bandwidth.

The great unsolved mystery is why Intel/HP and IBM arrived at such polar-opposite solutions. Intel and HP have obviously focused their efforts on exploiting single-thread ILP, with less concern for TLP or memory bandwidth. At the opposite extreme, IBM has focused on massive memory bandwidth and TLP but paid only moderate attention to ILP.

Intel obviously believes there is enough latent ILP lying around to justify a departure from the most dominant architectural franchise in the history of mankind. Intel says it has made the switch to a new ISA at this time to give it a solid platform to which it can later add TLP and high-bandwidth interfaces. It believes that others will eventually be forced to make this same ISA transition to avoid leaving a wealth of parallelism on the table.

IBM, on the other hand, clings to a far less pervasive ISA, seeing little rationale for more than minor tweaks. IBM says that memory bandwidth is the limiting factor today and predicts that it will only get worse over time. The company believes that the parallelism achievable with superscalar, multithreading, and multiprocessing can saturate any practical memory system, now and until quantum dots replace transistors. Thus, the whole issue of the ISA is simply a moot point.

Something is obviously amiss; both camps cannot be right. There are a number of possible explanations for the disagreement. One is that the companies are pursuing different markets. This explains some of the differences, but not all. If Intel were solely focused on low-end to midrange industry-standard servers, where price/performance is more important, that would explain the traditional busing and packaging technologies of Merced, and probably McKinley as well.

But this is not a completely satisfactory explanation. Although IBM may be biased more toward the high end than Intel is, HP's target market is right in line with IBM's. Intel and IBM both speak about servers with similar numbers of processors, both talk about high-availability systems, and both are interested in workstations. Given these similarities, it is hard to see how the workloads of the systems Intel and IBM both seem to covet could possibly be large enough to justify such disparate views on computer architecture.

Intel, of course, could have its eye on an even more distant market: PCs. While Intel is initially deploying IA-64 at the high end, where it is easier to flesh out, it may really be optimizing the architecture for future duty in PCs. This explanation makes some sense. After all, IBM may be correct: in servers, memory bandwidth and TLP may matter more than ILP or ISA. But Intel could also be correct: ILP and ISA may be important—just to a different market.

If this explanation is correct, it presents IBM with both a big opportunity and a big problem. With Intel's real attention elsewhere, IBM has a chance to bring its considerable resources and technology to bear exclusively on the server market, possibly establishing a strong market position before IA-64 gains a full head of steam. The risk IBM takes, however, is that the momentum Intel will gain in the broader markets could eventually undermine and overwhelm Power4-based servers, despite any technical superiority.

Another partial explanation for their differences may be Java. IBM is making large investments in Java technology—everything from Java class libraries for server applications to faster compilers and virtual machines. Most Java code is heavily multithreaded, playing directly to the strengths of Power4. Not coincidentally, Sun's Java architecture MAJC (see MPR 8/23/99, p. 13) is also optimized for TLP over ILP. Like Power4, MAJC uses CMP and, like IBM, Sun does not envision high-ILP cores; MAJC is optimized for four-instruction issue.

Power4 Not the End of Line

Even if Power4 is wildly successful in IBM servers, its overall impact on the market will be limited. IBM has no current plans to sell Power4 chips commercially, so other server vendors do not have it as an option. Even if IBM were to sell Power4 chips, it would be too late to derail IA-64. IA-64 appears destined to become the basis of industry-standard servers, and Power4 will always be vulnerable to it.

To prevent encroachment from IA-64, IBM must not only acquire the performance lead with Power4, it must hold it. And this performance lead must be convincing to make its market position unassailable. Of course, IBM is planning for just that. Its roadmap shows frequency increases of 25% every year, with performance growing at three times that rate before jumping dramatically with the mid-decade introduction of a new Power5 design. Considering the strength of the Power4 design and the technology muscle IBM is putting behind it, it may be a long time, if ever, before IA-64 infiltrates the large servers that are at IBM's heart. M