Cahners
MICROPROCESSOR REPORT
www.MPRonline.com
THE INSIDER'S GUIDE TO MICROPROCESSOR HARDWARE

IBM's Power4 Unveiling Continues

New Details Revealed at Microprocessor Forum 2000

By Kevin Krewell {11/20/00-03}

Microprocessor Report gave IBM's Power4 its Technology Award for 1999 (see *MPR 2/07/00-01*, "Best New Technology: Power4") because of the impressive range of innovations it offered, including thread-level parallelism, extreme memory bandwidth, multiprocessor scalability, and multichip packaging. As IBM revealed at Microprocessor Forum 1999, the company packed two processor cores onto one die, with a shared L2 cache that had enough bandwidth to feed both processors and a high-bandwidth chip-to-chip bus, which is pretty impressive stuff (see *MPR 10/06/99-02*, "Power4 Focuses on Memory Bandwidth"). IBM then packed four of those chips into one module, placing a total of eight 64-bit processor cores—680 million transistors—into 20 square inches. IBM Distinguished Engineer Charles Moore unveiled more details of the Power4 architecture at **Microprocessor Forum 2000**.

IBM has taken a balanced approach to server design, adding sufficient memory bandwidth to feed multiple 1GHz (or better) superscalar cores. In the x86 business, processor frequency has taken center stage in the race to higher performance, but PC system buses and memory bandwidth have typically lagged processor frequencies. The Athlon processor is an excellent example of a processor core that was held back by insufficient memory bandwidth. Once Athlon was mated to an on-chip L2 cache (Thunderbird) and, more recently, the AMD-760 DDR SDRAM chip set, it made significant (10% or more) performance advances. IBM has designed Power4, up front, to gain maximum performance from memory—including L1, L2, and L3 caches and main memory.

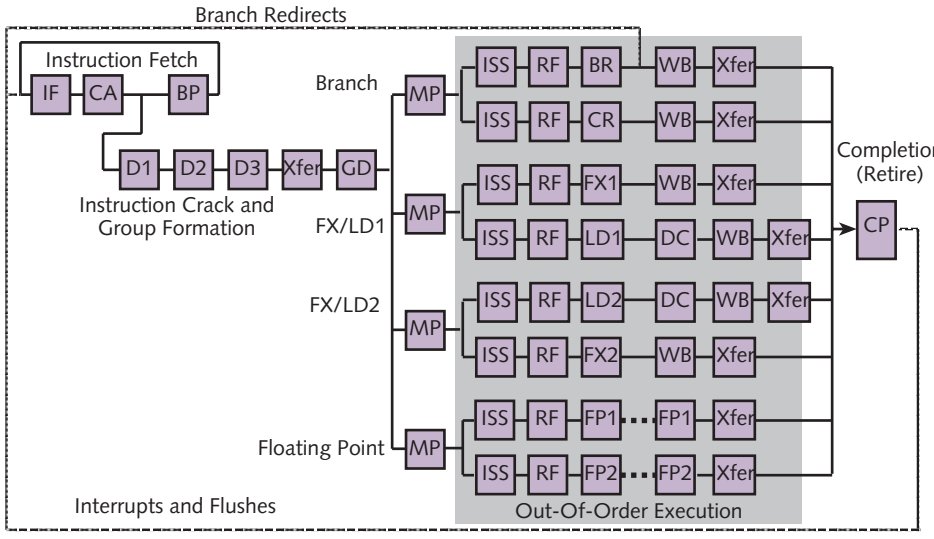

Figure 1. At **MPF2000**, Charles Moore of IBM described the 12-deep and 8-wide Power4 processor pipeline. The area in gray is the out-of-order execution core.

Scalability is a key attribute of large modern servers, and IBM has provided instantaneous SMP multiprocessing by placing two processor cores on a die and four die on a module, for a total of eight processor cores per module. Four modules can then be connected, using four parallel high-speed rings for 32-way SMP multiprocessing.

Although the Power4 employs some of the same RAS (reliability, accessibility, serviceability) concepts used by IBM mainframes (see *MDR 11/13/00-02*, "IBM Delivers 64-Bit Mainframe"), the Power server line doesn't address quite the same customer constituency and does not need to go to the same extremes. The S/390 has redundant instruction fetch and execution units and a functional unit to check every instruction execution, all of which use about 40% of the die area. And while Power4 doesn't go to excess, devoting somewhat less than 10% of the die to RAS features, it is very competitive with products from other RISC vendors. The S/390 is used mainly by financial institutions and must be absolutely foolproof; while the Power4 must still provide high availability and reliability, it is not held to quite the same extreme standard of fault tolerance.

MICHAEL MUSTACCHI

Charles Moore, IBM Distinguished Engineer, presenting at MPF 2000.

The Power4 uses error detection and correction codes (ECCs) to protect the L2 and L3 caches and main memory. ECC is also used on the buses that connect memory to the processor. Other soft-error-sensitive arrays are covered by parity. The L1 cache is protected by parity, rather than ECC, to improve access speed and because the L1 is a write-through cache; a parity error will force a reload from the L2 (which is ECC-protected). Power4 uses a deferred error strategy, relying on software for recovery. It also has fast BIST capability, including access to all chip latches, registers, and array contents, for testability.

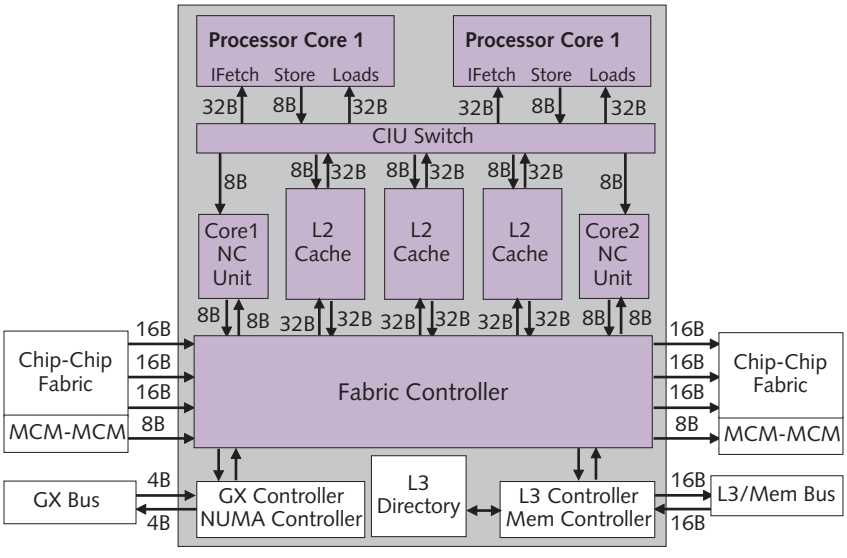

Figure 2. The processor cores are connected to the L2 slices through the CIU switch. The L2 is connected to a fabric controller that facilitates scalability.

Power Gets RISC-ier

To hit gigahertz and beyond frequencies, the Power4 operates on simplified Power instructions. Complex PowerPC and Power instructions are "cracked" into multiple simple instructions. Simpler instructions streamline the execution pipeline, which brings us back to one of the original concepts of RISC. The penalty for cracking Power instructions into a simple subset is a longer decode stage (three cycles). In the decode stages, instructions are bundled to reduce the instruction-tracking burden through the out-of-order execution stages. The transfer stage (Xfer) accounts for wire delays, traversing the die to the group dispatch (GD) stage, where instructions are next dispatched to the proper execution pipeline.

As Figure 1 shows, Power4 has eight execution pipelines: two for loads/stores, two for integer execution, one for branch, one for condition register operations, and two for floating point. Before issuing the bundle to the execution pipeline, the mapper (MP) stage identifies and allocates the required resources and determines register dependencies. When the bundle is moved to the execution pipelines, the instructions are held in queues and issued in the issue-cycle stage (ISS). The out-of-order execution core uses dataflow concepts, such as noninterlocking pipelines and self-initiated issue logic, and instructions are bundled and scheduled (and retired) as bundles to minimize control logic. Within the inner core, issue queues feed the eight execution pipelines. The CR unit is responsible for performing logical operations on the condition register. With the numerous execution pipelines, a large register rename pool, and deep pipelines, the Power4 is capable of supporting more than 200 instructions in flight. The inner core of the processor supports out-of-order execution, but instruction bundles are retired as a unit. Bundles are retired (completed) in order (CP stage), and all instructions in the bundle must have completed execution before the bundle is retired. The Power4 pipeline has 12 stages, from instruction fetch to write-back, for an integer operation.

One key aspect of fast processors that isn't very glamorous, but is essential to fast operation, is the clock generator. Clock

circuit design and clock distribution can be a significant challenge in a die as large as the Power4. IBM used a gridded H-tree structure and has managed to design a clock circuit on Power4 that keeps clock uncertainty between any two latches on the chip to less than 22ps.

The Power4 has a reasonably long 12-stage pipeline, and long pipelines are susceptible to IPC (instructions per cycle) degradations, because instruction branches cause the pipeline to be flushed and then (of necessity) refilled, wasting execution cycles and computational resources. To improve branch predictions and minimize flushing/refilling, Power4 has put significant effort into the branch predictor. The branch predictor has a local, a global, and a (local/global) select predictor, each with a 16KB table. Branch prediction accuracy should be in the mid-90% range for typical commercial workloads. And when branch prediction does fail, IBM also has a fast, selective flush to minimize the cycle penalty by discarding only instructions on the mispredicted path.

Power4 Is Rich With Cache and Bandwidth

The L1 cache has been optimized for minimum latency and maximum availability. The 32KB data cache is triple ported to allow two loads and one store every cycle. The separate 64KB instruction cache is single ported. The L1 instruction cache hardware supports up to eight prefetch streams. To support the large number of instructions and threads in flight, the cache supports a total of 11 outstanding misses (8 for data and the remaining 3 for instructions). Each L1 has a dedicated port into the L2, which has greater than 100GB/s of bandwidth. IBM also included substantial address translation resources, although details were not available at Microprocessor Forum 2000.

The on-chip L2 cache consists of three slices of approximately 0.5MB each, as Figure 2 shows. The three are connected to the "fabric" that provides the interprocessor connectivity for multiprocessing. Within each cache slice are 4 coherency processors that control data movements and reloads. The 12 coherency processors give the entire L2 cache the ability to support 12 outstanding L2 misses. An arbiter in each slice controls the requests to the coherency processors from the two on-chip processor cores' store queues, L3 requests, the multiprocessor fabric controller, and the processors' read requests. The processor store queues can also gather stores by delaying (pacing) the store in the hope that a store to another location in the cache line can be included in the store cycle to memory. The L2 cache is coupled to the fabric for snoops and includes an outbound snoop queue and a cast-out queue to the fabric. The L2 implements a sophisticated seven-state cache coherency protocol. In contrast, Intel's Pentium III Xeon uses a four-state (MESI) protocol, and AMD's Athlon implements a five-state (MOESI) protocol. These additional states provide enhanced coherency protocols, resulting in improved cache efficiency and reduced memory traffic.

IBM also provides an eight-way set-associative tertiary (L3) memory controller on chip, but die space limitations dictated that that L3 memory should be off chip, especially considering that the L3 is 32MB, as described at Microprocessor Forum 2000. The L3 controller and directories are on the processor die to support fast coherency checking, and the L3 controller has eight coherency processors and eight snoop/cast-out queues. The L3 directory has separate snoop ports for performance and implements a five-state cache coherency protocol. Main memory and the L3 are accessed through the same channel, and, although this channel operates at one-third of the processor speed, it offers more than 10GB/s of bandwidth.

Wrapping All the Threads Together

The eight processor cores on the module form a tightly coupled symmetrical multiprocessing (SMP) group that can be

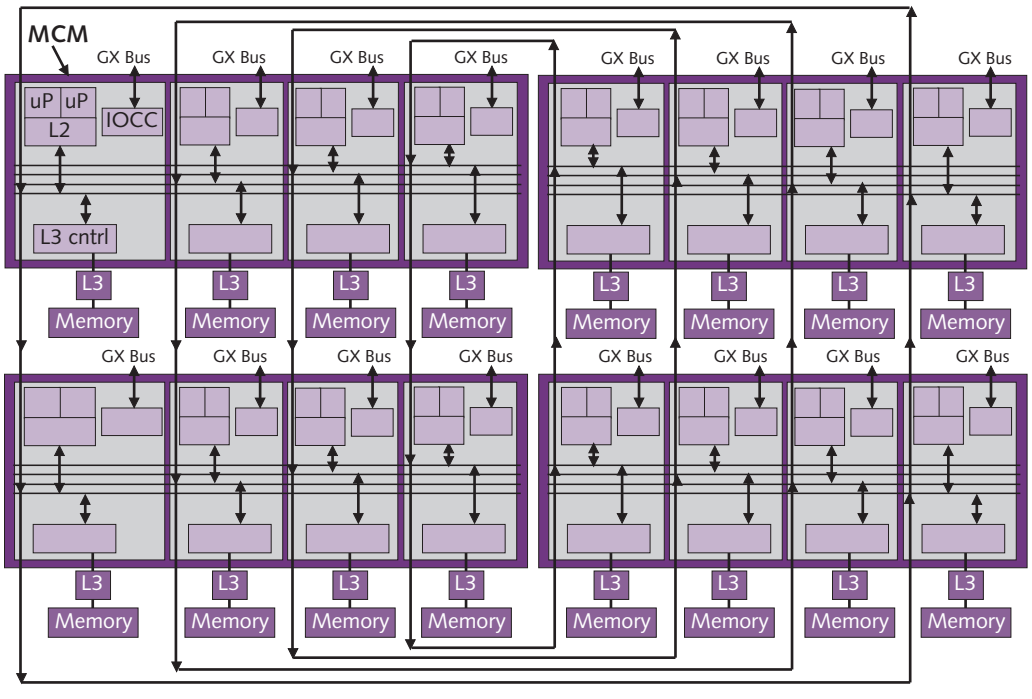

Figure 3. Processor die are connected in the MCM (multichip module) through four hybrid switch/bus structures. The virtual buses can then connect together up to four modules (32 processor cores) without glue logic.

extended to 32 processors by connecting four modules. To keep 32 processor cores connected, IBM has a hybrid switch/bus design that can be visualized as four logical buses per module, as Figure 3 shows. The four buses can be connected to the four buses in an adjoining module, and four modules can be arranged in a ring topology without any glue logic. These MCM-to-MCM buses run at half the processor speed, but at 8 bytes wide, offering more than 4GB/s of bandwidth per bus. The fast buses are used to enable "aggressive" cache-to-cache transfers. This tightly coupled multiprocessing allows processing threads to operate in parallel over the processor array. With four modules, Power4 supports up to 32-way multithreaded processing.

Beyond 32 processors, 2GB/s coherent links attach to the processor die through the "GX" bus, and the same links can be used for clustered systems. The architecture also supports logical partitioning of the processing elements.

With a fast processor core (1GHz+), glueless scalability, and scalable bandwidth, the Power4 will form an impressive server when it is introduced. IBM will continue to leverage this architecture and scale it with even faster processors (1.5GHz and later 2.0GHz+), faster buses, expanded coherency capabilities, larger caches, and more threads closer together. With this level of system sophistication, performance, and commitment, IBM's Power architecture is living up to its name. ◇