Reed Electronics Group

MICROPROCESSOR REPORT

www.MPRonline.com

THE INSIDER'S GUIDE TO MICROPROCESSOR HARDWARE

POWER5 TOPS ON BANDWIDTH

IBM's Design Is Still Elegant, But Itanium Provides Competition

By Kevin Krewell {12/22/03-02}

On large multiprocessing systems, often the dominant attributes needed to assure good incremental processor performance include memory coherency issues, aggregate memory bandwidth, and I/O performance. The Power4 processor, now shipping from IBM, nicely balances integration with performance. The Power4 has an eight-issue superscalar core, 12.8GB/s of memory bandwidth, 1.5MB of L2 cache, and 128MB of external L3 cache. The recently introduced Power5 steps up integration and performance by integrating the distributed switch fabric between memory controller and core/caches. (See *MPR 10/14/03-01*, "IBM Raises Curtain on Power5.") The Power5 has an on-die memory controller that will support both DDR and DDR2 SDRAM memory. The Power5 also improves system performance, as each processor core is now multithreaded. (See *MPR 9/08/03-02*, "IBM Previews Power5.")

IBM has packed four processor die (eight processor cores) and four L3 cache chips on one 95- × 95mm MCM, seen in Figure 1. The MCM has 4,491 signal I/Os and a ceramic substrate with 89 layers of metal. Higher integration has reduced component count from that of Power4, consolidating the memory controller on the processor die, and the L3 cache is now implemented in one 36MB chip. The eight processor cores located on the module form a tightly coupled symmetrical multiprocessing (SMP) group that can be extended to 32 processors by connecting four modules. Two MCM modules form what IBM terms a book.

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 2 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 are eight bytes wide, offering more than 4GB/s of bandwidth per bus. The fast buses are used to enable what IBM terms aggressive cache-to-cache transfers. This tightly coupled multiprocessing is designed to allow processing threads to operate in parallel over the processor array. With eight modules, Power5 supports up to 128-way multithreaded processing.

The bus structure and distributed switch fabric also allow IBM to create a 64-way (processor cores) configuration.

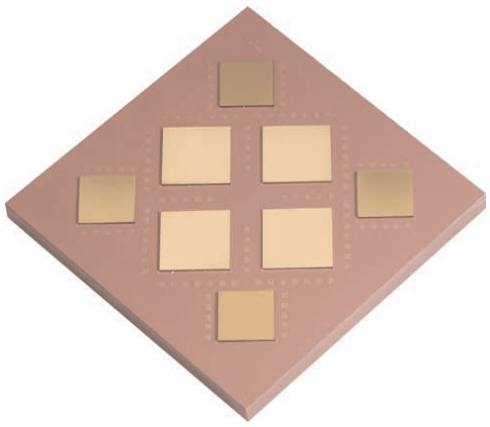

Figure 1. A Power5 module with four processor die and four L3 cache chips.

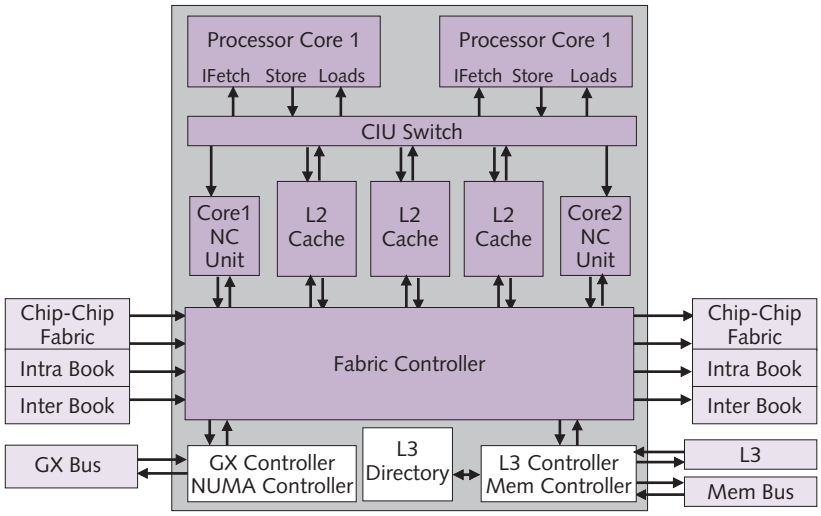

Figure 2. The Power5 block diagram closely resembles its predecessor, the Power4, but offers greater integration. The intra-MCM buses now run at full-processor speeds, but also narrower than in the Power4. The inter-MCM buses still run at half-clock speed. The number of buses has increased, increasing the aggregate bandwidth of the Power5 processor over the Power4. The GX bus is used for I/O interconnection, either through a high-performance switch to interconnect systems or through a hub to PCI-X adapters.

The connecting buses between the MCMs exploit an enhanced version of the distributed switch from the Power4 processor. All chip interconnections operate at half-processor frequency and scale with processor frequency. Intra-MCM buses have been enhanced from Power4 to allow operation at full-processor speeds. The inter-MCM buses continue to operate at half-processor speeds. IBM has not revealed specific target clock speeds for the Power5, but we expect the 130nm version of Power5 to be slightly faster than the 1.7GHz Power4 IBM ships today. When the Power5 shrinks in IBM's 90nm process, we expect it will ship at clock speeds faster than 2.0GHz.

The architecture also supports logical partitioning of the processing elements. Partitioning allows the collection of processors to be subdivided into smaller groups; each group is isolated and can run a separate operating system. Partitioning is a mainframe technique that recently has been applied to more-conventional microprocessor systems. The system partitioning has a number of advantages in a data center, including improved fault isolation and tolerance, and partitioning allows IT managers to invest in a single hardware MP system without needing to commit in advance to a set of applications and a corresponding operating system. For example, the partition can support arbitrary combinations of Linux and AIX versions. IT managers can also consolidate numerous servers into one box. This technique is also available for SPARC, x86, and Itanium processors. IBM partitions processors, using a combination of hardware and what the company calls a hypervisor; IBM's Power architecture can also create multiple protected partitions, even on one processor.

Sun's Product Delays Limit System Bandwidth

Sun's UltraSPARC IV continues to use the FirePlane System Interconnect infrastructure, released in 2000 with the UltraSPARC III. (See *MPR 11/10/03-02*, "UltraSPARC IV Mirrors Predecessor.") The aged design is based on SDRAM, not double-data-rate (DDR) SDRAM. The US III system relies on wide datapaths to provide the necessary bandwidth. The FirePlane design uses an interconnect point-to-point data network with a hierarchical address bus. The 4.8GB/s bidirectional interface matches the 4.8GB/s of memory bandwidth and has a capability of 150 million snoops per second.

The other SPARC design from Fujitsu is moving to a new system bus for the SPARC64 VI, which it calls Jupiter. (See *MPR 11/24/03-01*, "Fujitsu Makes SPARC See Double.") The new Jupiter bus design is based on the company's experience with mainframe design. It adds a new cache-state MOWESI protocol, with the W state modified by others and unmodified by the owner. The 128-way SPARC64 V systems will become 256-processor-core systems with the dual-core SPARC64 VI. The Jupiter bus is unidirectional, eliminating arbitration and bus turnaround time.

AMD Opteron scales well for two-way and four-way, with three coherent HyperTransport links between

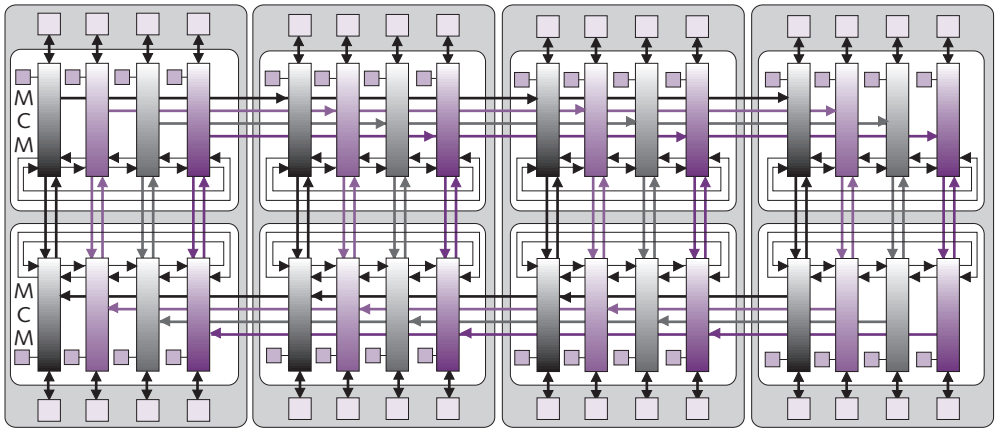

Figure 3. A Power5 module configuration with four MCMs connected. The basic Power5 building block is two MCM's that form a book.

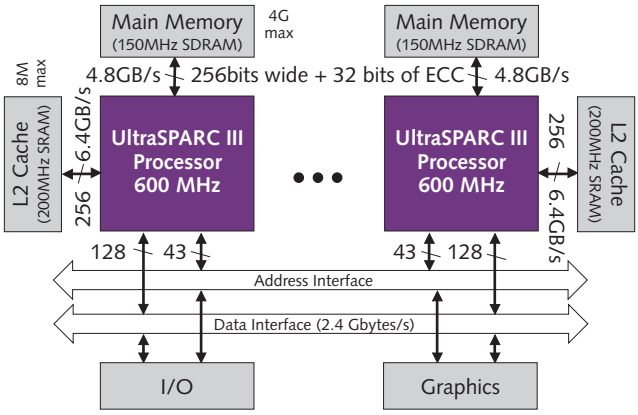

Figure 4. Sun's UltraSPARC III bus design will be extended to the UltraSPARC IV.

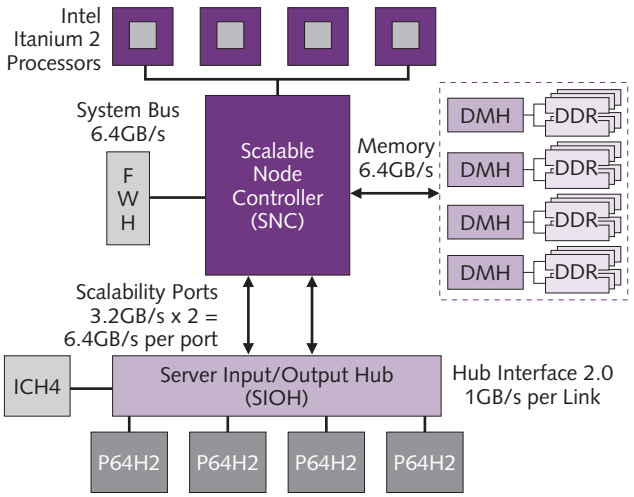

Figure 5. The Itanium 2 chip set from Intel balances the bandwidth of memory, the shared system bus, and the scalability ports, all at 6.4GB/s of raw bandwidth. Larger eight-way and above systems are built by having an OEM-specific switch that connects each four-way node through the scalability ports.

processors, and can be configured as an eight-way. Beyond four-way, however, cache snoop traffic and nonuniform memory accesses begin to hurt scalability. AMD lacks the off-the-shelf solution for scalability beyond eight-way. The recent agreement with Sun Microsystems could give AMD access to Sun's scalable technology.

Itanium 2 Leverages OEM Technology to Scale

Intel's Itanium 2 processor uses a shared front-side bus for up to four processors, as Figure 5 shows. Beyond four-way, Intel provides a proprietary scalability port to connect clusters of four-way configurations. Intel itself doesn't provide a switching solution for systems larger than four-way; it is the province of the various Intel OEMs to provide their own "secret sauce."

HP has its Superdome design that started with PA-RISC processors and is now migrating to Itanium 2 processors. SGI is building systems based on technology it acquired with Cray. Unisys also offers scalable solutions, even for Xeon processors. The Itanium 2 bus runs a double-pumped 200MHz bus across four processor nodes, with a 128-bit wide bus to provide raw bandwidth of 6.4GB/s. Intel basically took Pentium 4 technology and scaled it out for Itanium. HP also adopted the bus for its PA-RISC processor to ease the transition from PA-RISC to Itanium. To provide significant increases in performance, however, Intel is scaling frequency and on-die L3 cache size, just as it has done in its 32-bit XeonMP line.

Intel has committed to this interface through the Montecito processor in 2005. Not until the Tanglewood processor in 2006 or later will Intel be able to offer a more up-to-date system design. (See *MPR 10/06/03-01*, "IDF Delivers Extreme Surprises.") That is not to say that a system vendor couldn't create a switched-bus design, but it's not very practical with the Itanium 2 support Intel provides today. Despite those limitations, Itanium 2 systems have shown good scalability on SPEC-rate benchmarks with both the HP Integrity Superdome server, up to a 64-way configuration, and the SGI Altix 3000 server, also up to 64 ways.

The design issues in scalable single-image servers include cache-coherency traffic for snoops or the bandwidth constraints of tables in directory-based coherency. Cache snooping can use up significant bus bandwidth and can limit scalability. Directory-based schemes require active resources in the chip sets and extensive experience. IBM's complete system solution, though not new, still provides an elegant, scalable, and space-efficient solution in the 64-bit server market.

At present, the biggest/fastest IBM Power4 system on the SPEC benchmark website, the eServer pSeries 690 Turbo (1,700MHz, 32 CPUs), lags behind a 32-way Itanium 1.5GHz HP Superdome in both SPEC_INT2000 and SPEC_FP2000 rate charts. The Power5 design should boost individual processor performance and aggregate system bandwidth in a way that Intel will not be able to match until the Montecito and Tanglewood processors appear. Power5 should provide IBM with an edge during 2004 and 2005. ◇