IBM POWER4: a 64-bit Architecture and a new Technology to form Systems

Rui Daniel Gomes de Macedo Fernandes

Departamento de Informática, Universidade do Minho
4710-057 Braga, Portugal
ruif@net.sapo.pt

Abstract. The IBM POWER4, a new 64-bit architecture for the POWER family, includes now Symmetric MultiProcessing capabilities (SMP), referred as a distributed switch. In this communication we take a closer look into its main features and microarchitecture improvements and we will try to place its position on the high-end servers.

1 Introduction

The POWER4 is the latest feature of the POWER architecture from IBM. Initially introduced in 1990 with the POWER1 they were 32-bit, RISC, with 8KB instruction cache (I-Cache) and either a 32 KB or 64 KB data cache (D-Cache) [1]. They had a single floating-point unit capable of issuing one compound floating-point multiply/add (FMA) operation each cycle, with a latency of only two cycles. Later in 1993, with the POWER2, the most significant improvement was a floating-point unit (FPU) that was enhanced to contain two 64-bit execution units. Thus, two FMA instructions could be executed each cycle. A second fixed-point execution unit was also provided.

At the same time, IBM announces the PowerPC 601, which was the first processor that implements PowerPC Architecture[1][2], from a partnership between IBM, Motorola and Apple. The PowerPC Architecture basically includes most of the POWER instructions, however, some instructions that were less executed were excluded, and some new instructions were added such as support for symmetric multiprocessor (SMP) systems. In fact, the 601 did not implement the full PowerPC instruction set, and was a bridge from POWER to the full PowerPC Architecture implemented in more recent processors. In 1997, with the first RS64 processor with 64-bit architecture, 2-way associative 4MB L2 cache, 64KB L1 instruction cache and 64KB data cache, one floating-point unit, one load-store and one integer unit, allows systems design to use up to 12 processors. Later, this processor family had new designs (RS64-II, III and IV) that mainly increase the L1 cache from 64K to 128K, the L2 cache became 4-way set associative with an increase in size from 8MB to 16 MB, hardware multithreading was also introduced, permitting systems design up to 24 processors.

In 1998, the POWER3 processor brought together the fundamental design of the POWER2 microarchitecture, with the PowerPC Architecture: two floating-point execution units, while being a 64-bit, SMP-enabled. Note that earlier IBM Power/Power PC processors were divided into server and scientific ones - POWER and RS64. Today, the POWER4 was designed to address both commercial and technical requirements, full compatible and extended of 64-bit PowerPC Architecture, operating from 1.1 to 1.3 GHz.

[1] A PowerPC Architecture is a specification that implements RISC

The recent POWER4 has new enhancements and a new interconnection topology, known as a distributed switch, within which a set of chip are designed together to realize a system [3].

In the following sections we will present the main features of the POWER4 processor, mainly parallelism, processors interconnection, memory hierarchy and memory bandwidth, as well as the chip himself. Later, the Intel IA-64 architecture is reviewed to better follow the options and goals that were taken by the two companies, trying to place its target application.

2 POWER4 Chip: Microarchitecture Overview

Mainly designed to achieve High Performance Computing (HPC), which mean in our days, multithreading and parallelism [4], Figure 1 shows the components of the POWER4 chip. The design features 2 processors cores per unit, each one having an L1 cache (64KB for instructions, 32KB for data, per processor).

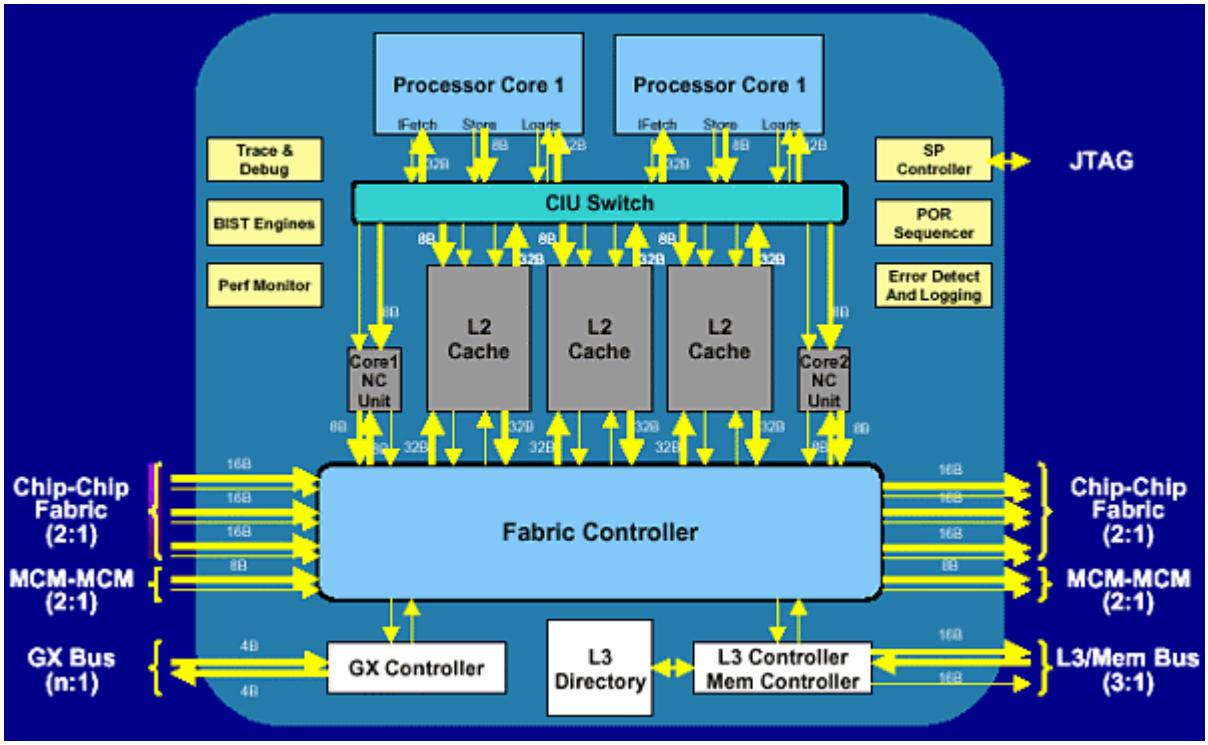

Figure 1 – POWER4 chip logical view (courtesy of IBM)

2.1 Memory-Level Parallelism, System Bandwidth

The 2 processors share a unified L2 cache (~1,5MB) which is implemented as 3 separate controllers connected to the cores via a Core Interface Unit (CIU) that is a cross bar switch between the L2 and the processors, with each one operating concurrently. Each processor uses two separate 32-byte buses to connect the CIU for data fetching and data loading, as well as a separate 8-byte bus to save the results. The L2 cache has a bandwidth of 124.8 GBytes/s per chip. Note that POWER4 on-chip L2 is shared directly by two on-chip processors and by processors on other chips via a high speed chip-to-chip interconnect

network, as Figure 2 shows. Each processor has a non-cacheable unit (NC), logically as a part of the L2, responsible for handling instruction-serializing functions and perform any noncacheable operation in the storage hierarchy. Data flows coming from the memory and L2 and L3 caches, and the buses of the chips are controlled by the Fabric Controller. The L3 controller and the L3 directory are located on the chip, the L3 memory (32MB 200 MHz DDR SDRAM) is off the chip. For connection with the L3 cache working at 1/3 of the processor's speed and with the memory there are two 16-byte buses (one on and one off)[2]. The throughput of L3 interface is about 11.1 GBytes/s per chip, 55.5 GBytes/s for 4-chip module. One of the most interesting characteristics of L3 cache is the share via switched buses among the several modules (Figure 3). The L3 memory controller on chip has 8 coherency processors, 8 snoop cast-out queues and a separate directory snoop port. Such high bandwidth keeps the network utilization low, which according to queuing theory, minimizes network latency. The main memory can be 0-16 GB.

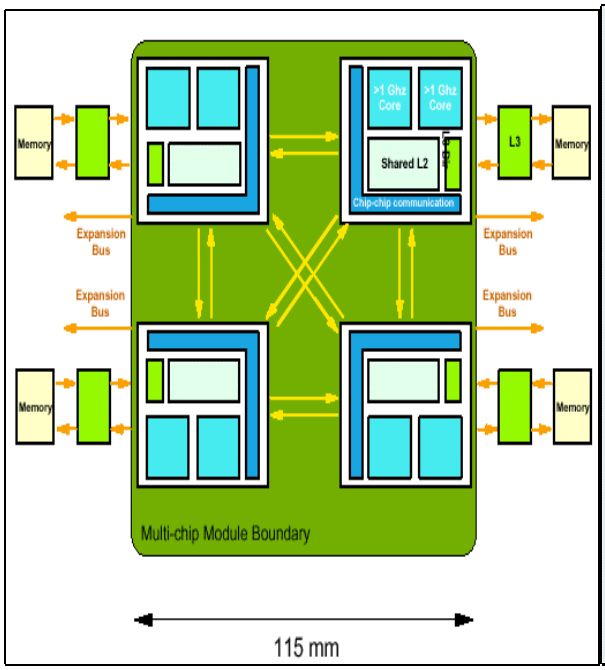

Figure 2 – Four POWER4 chips can be packed into one module forming a 8-way SMP

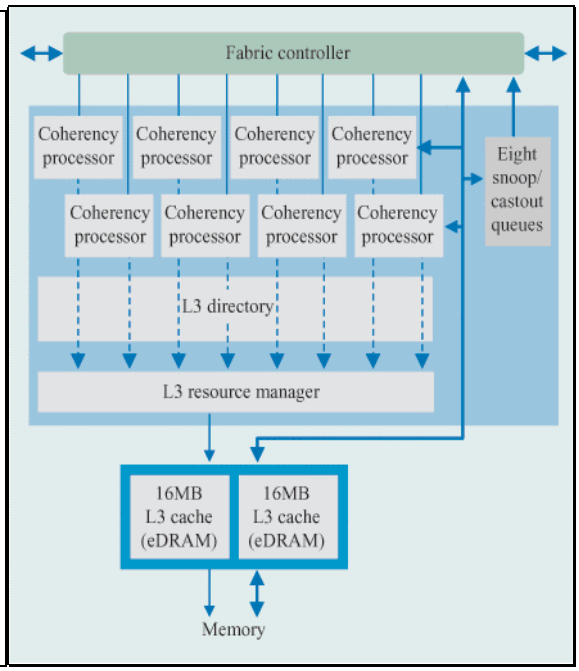

Figure 3 – L3 Logical view

2.2 Thread-Level Parallelism

The 4-byte GX Bus (one on and one off) running at 1/3 of the processors clock frequency is responsible for controlling the flow of information into and out of the system. That is, the point at which we would directly attach an interface to a switch for clustering multiple POWER4 nodes. Usually a four POWER4 chips can be package on a single module to form a 8-way SMP (tightly coupled SMP) as shown in Figure 2. Four such modules can be interconnected to form a 32-way SMP (hierarchical SMP and clusters). To accomplish this, each chip contains five primary interfaces. To communicate with POWER4 chips on the same module, there are logically four 16-byte buses. Physically, these four logical buses are implemented with six buses (three on and three off)[2]. To communicate with POWER4

[2] One on and one off, is the same as in and out from the chip point of view, as shown in Figure 2

chips on other modules, there are two 8-byte buses (one on and one off) with 41.6 GBytes/s chip interconnect. We can also locate some units that perform logical functions like *trace and debug*, *built-in self-test*, *performance-monitoring,* an interface to the *service processor* used to control the overall system, a *power-on reset* sequencing logic, and a *error detecting and logging* circuitry.

The interconnection topology appears like a bus-based system from a perspective of a single chip. From the perspective of the module, it appears like a switch. When interconnecting multiple modules, the intermodule buses act as a repeater, moving requests and responses using a ring topology (that IBM describes only as a distributed switch), that is, each chip always send requests/commands and data on its own bus but snoops all buses. 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 its complexity.

2.3 POWER4: Instruction-Level Parallelism

The POWER4 has a speculative superscalar out-of-order execution design. Up to eight instructions, with a sustained completion rate of five, can be issued each cycle. With this organization it is possible to have more than 200 instructions in flight at any given time. So, instruction-level parallelism is achieved by eight execution units, each one capable of issuing an instruction per clock cycle. As shown in Figure 4, these execution units are divided: two identical floating-point execution units (FP1, FP2) each capable of starting a fused multiply and add; two load load/store units (LD1, LD2); dual fixed-point units (FX1, FX2); a branch execution unit (BR) and an execution unit to perform operations on the condition register (CR).

The load/store unit employs a special mechanism to ensure memory consistency to support out-of-order execution of instructions.

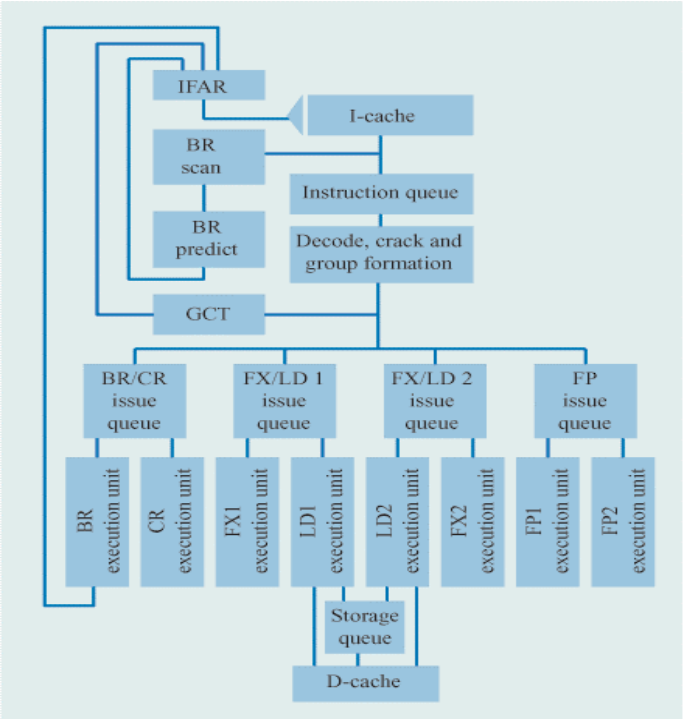

Figure 4 – High Level block diagram (Courtesy IBM)

2.3.1 Branch Prediction

As we discuss later, the POWER4 has a 17 stages pipeline designed for high-frequency. With such level of pipeline it will be necessary to have a schema to implement mechanisms of branch-prediction. POWER4 uses a multi-level branch-prediction that is a set of three branch history tables. The first, called the *local predictor*, is similar to a branch-history table (BHT) that produces 1-bit predictor that indicates to be *taken* or *not taken*. The second, called the *global predictor*, produces another 1-bit branch-direction predictor based on the previous eleven fetch groups of instructions. A third table, called *selector table*, is used to select between the local and the global predictions.

2.3.2 Decode, Crack, and Group formation

The instruction cache (I-cache) is capable of delivering up to 8 instructions per clock according to the address given by the Instruction-Fetch Address Register (IFAR) the contents of which are determined by the branch prediction unit. Groups of instructions are formed. A group contains up to five internal instructions referred to as IOPs. In the decode stages the instructions are placed sequentially in a group, the oldest instruction is placed in slot 0, the next oldest one in slot 1, and so on. Slot 4 is reserved for branch instructions only. Then, the POWER4 cracks PowerPC instructions into a greater number of simpler instructions which then combine into groups and are executed. If an instruction is split into 2 instructions we consider that *cracking*. If an instruction is split into more than 2 IOPs then we term this a *millicoded* instruction. Each instruction slot feeds separate issue queues for the FP, BC, CR, FX and LD execution units. Table 1 summarizes the depth of each issue queue and the number of queues available for each type of queue.

Table 1 – Issue queues

Queue type	*Entries per queue*	*Number of queues*
Fixed-point and load/store	9	4
Floating-point	5	4
Branch execution	12	1
CR logical	5	2

2.3.3 Instruction Execution Pipeline

The POWER4 has a 17-stage instruction pipeline as shown in Figure 5. With a closer look, there are some important remarks: when a branch instruction is mispredicted, there is at least a 12-cycle branch-mispredict penalty; floating-point instructions require six execution cycles, when there are one dependent on a prior floating-point, that cannot issue within six cycles; however, as it is the case with other execution units, other floating-point instruction may issue during this period; to execute two fixed-point instructions with one dependent on the other, must have at least one dead cycle between their issue cycles.

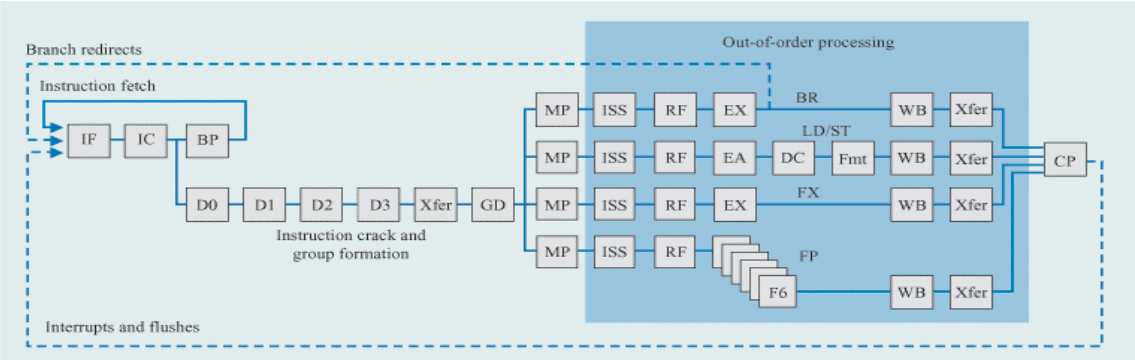

Figure 5 – POWER4 instruction execution pipeline (courtesy of IBM)

2.4 Storage Hierarchy, Hardware Data Prefetch

The POWER4 memory hierarchy consists of three levels of cache: L1- direct map, four 32-byte sectors for instruction, two-way 128-byte line data cache; L2- eight-way, 128-byte line; L3- eight-way set-associative, four 128-byte sectors (for compatibility with the L2 cache); and the main-memory subsystem (DRAMs 0-16GB). With such a memory bandwidth it is necessary to grant low-latency memory accesses. The chip implements a mechanism of prefetch streams, actually eight software-activated prefetch streams. Figure 6 shows the sequence of prefetch operations. 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. Special logic to implement data prefetching exists in the processor load/store unit (LSU) and in the L2 and L3. The direction to prefetch, up or down, is determined by the actual load address within the line that causes the cache miss.

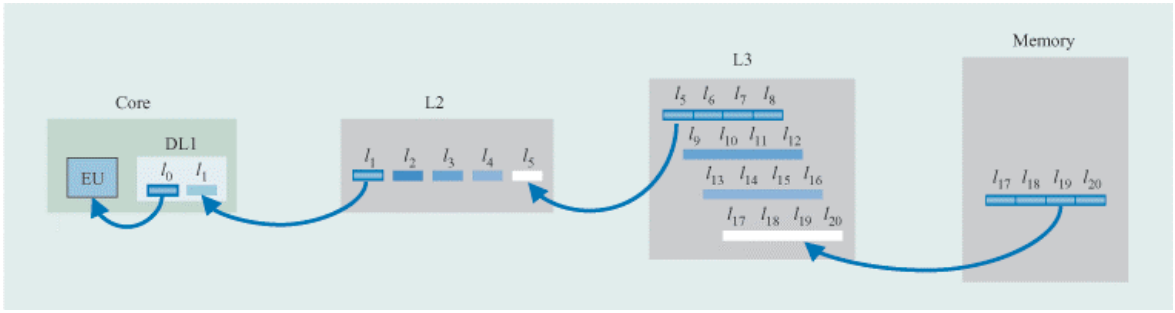

Figure 6 – POWER4 hardware data prefetch (courtesy of IBM)

3 POWER4 and IA-64 –Different Approaches to Processor Architectures

If we look at the microarchitecture of these two processors, we can easily conclude that both companies, IBM and Intel, have taken two different and opposite options concerning 64-bit based processors for similar goals. Both talk about high-availability systems, and both are interested in workstations. Intel has focused the efforts on exploiting single-thread Instruction Level Parallelism (ILP), with less concern on memory bandwidth. With the IA-64, it was the right time to work over Instruction Set Architecture (ISA); after all, it is a

new 64-bit architecture that could not use the old ISA 32-bit based and easily get high levels of performance. With this ISA approach, Intel intends to give a solid platform to which it can later add Thread Level Parallelism (TLP) and high-bandwidth interfaces.

At the IBM side, we can see focus on massive memory bandwidth and TLP, with moderate attention to ILP. To IBM, memory bandwidth is the limiting factor today and predicts that it will 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 a new technology replace the actual use of transistors (probably with the *quantum dots*[3] technology). Thus, the ISA is not the main factor for the performance of a processor.

Another radical difference between these two processors concerns the scheduling, dynamic on the POWER4 and EPIC-static scheduling for servers. In the presence of cache misses, the POWER4 dynamically remakes the instruction schedule, thereby avoiding pipeline stalls on cache misses. IA-64 machines, because of their in-order execution and static instruction groupings, are less adaptive. If EPIC compilers for traditional code are a challenge, dynamic just-in-time compilers (JIT's) for Java will be a nightmare. Java performance is a serious issue for IBM, which 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 MACJ architecture has many features that are radically different from IA-64 for just these reasons).

After all, both companies may have correct approaches but with different goals: to IBM, in servers, memory bandwidth and TLP may matter more than ILP or ISA; to Intel, ILP and ISA may be important – just to a different market.

The market that IBM apparently tries to cover, starts over small PC-based trough the high-end high-availability enterprise servers that run massive commercial and technical workloads for corporations and governments. Recently IBM presented a downsized version of POWER4 specially focused on the PowerPC chip for smaller servers, graphics workstations, and desktop computers[5]. Intel with IA-64, is initially focused on low-end to midrange industry standard servers, where price/performance is more important. Also, Intel with high focus on ISA rather than TLP and memory bandwidth, could have its eye on the PC's market.

Java can also make some difference on the target market for both processors. Most Java code is heavily multithreading, playing directly to the strengths of POWER4. IBM is making large investments in Java technology – everything from Java class libraries for server applications to faster compilers and virtual machines.

Unlike other companies, that for one reason or other have dropped their processor development, IBM resists, mainly because of the more and more critical system-performance placed on the processor. So, losing the control over it would rob the ability to have full control over the technology and to differentiate itself from other companies. Essentially, to IBM the POWER4 will feed the high-end (high-margin) server business that is expected to continuous grow rapidly along with the Internet and the e-business. But, IBM will have to compete over the best IA-64 processors, especially with the last Mckinley implementation. By now, some companies are already bucking IA-64, if POWER4 fails it seems to be a clear indicator for big companies like Sun, HP-Compaq and others, that proprietary CPU's may be out of fashion [6].

[3] Quantum dots are nanometer-scale "boxes" for selectively holding or releasing electrons. That means computers based on quantum physics would have quantum bits, or "qubits," that exist in both the on and off states simultaneously, making it possible for them to process information much faster than conventional computers.

4 A Perspective to the Future

As usual, new enhancements are expected in order to provide even greater performances increases over time. The current design introduces parallelism that allows the processor to continue processing in the presence of cache misses. Later in 2004, we can expect "POWER5" - besides the usual clock that is expected to run well above 2 GHz. Clock frequency will increase, together with larger caches and memory, allowing CPU improvements on parallel instruction execution and processor interconnections. There will be "Fast Path" hardwiring (on-chip hardware acceleration) of some common tasks like TCP/IP processing; later maybe high-level database or bioinformatics routines. Two years after, we should see "POWER6", with expected 4-way multithreading, wide parallel execution and even more hardwiring capability (maybe direct on-chip memory controllers and fast vector FP units).

5 Conclusions

Clearly, the POWER4 processor has born to achieve HPC. Simultaneously, with this processor IBM intends to hold the large high-end servers, preventing encroachment from others competitors especially from Intel with is last implementation of the IA-64 (Itanium-2). Despite this goal, it seams that the biggest bet from IBM when developing processor microarchitecture takes concern with the thread-level-parallelism, memory bandwidth and multi chip module that can be further interconnected to form 16-, 24-, 32-way SMP's. Mainly, because of Java big investments, and necessarily because of the more and more globalization of the internet and the e-business, that impose huge workloads and multiprocessing.

In the next few years, we expect to see the new developments concerning performance and probably specialized processors for specific applications, and as soon as new technology improves new enhancements that probably impose a redefinition of all micro architectures.

References

[1] Scott Vetter: IBM Redbooks: The POWER4 Processor Introduction and Tuning Guide, (2001) 1-4

[2] Tendler J.M., Dodson J. S., Fields J. S. Jr., Le H., Sinharoy B.: POWER4 system microarchitecture, IBM Journal Research & Development, Vol. 46, No 1, (2002).

[3] C. May (Ed.), E. M. Silha, R. Simpsom, and H. S. Warre, Jr. (Ed.): The PowerPC Architecture: A Specification for a New Family of RISC Processors, Morgan Kaufmann Publishers, Inc., San Francisco, (1994).

[4] Breckenridge Trey: Presentation for High Performance Computing, Engineering Research Center, Mississippi State University, (2002).

[5] Halfhill Tom R.: IBM Trims Power4 Adds AltiVec, Microprocessor Watch, Issue #101, (2002).

[6] Diefendorff Keith: Power4 Focuses on Memory Bandwidth, Microprocessor Report, Vol. 13, Nr. 13, (1999).