Reed Electronics Group

MICROPROCESSOR REPORT

www.MPRonline.com

THE INSIDER'S GUIDE TO MICROPROCESSOR HARDWARE

IBM Makes Designer Genes

BlueGene/L Supercomputer Processor Inspired by Embedded SoCs

By Tom R. Halfhill {10/11/04-01}

Designing the world's fastest supercomputer by drawing inspiration from embedded processors seems like imitating a Vespa when building a Formula 1 racer. Aren't lowly embedded chips supposed to be on the receiving end of hand-me-down technology? As we've seen in the past few years, however, embedded processors are blazing the trail for multicore designs, hardware multithreading, massively parallel processor arrays, high-speed on-chip interconnects, and other advanced design strategies.

So perhaps it's no surprise that IBM Microelectronics would pattern a new supercomputer processor after an embedded system-on-chip (SoC), even to the point of recycling a five-year-old processor core previously found only in embedded parts. Moreover, IBM readily acknowledges the new processor's ancestry, taking pride in a design that combines performance with parsimony.

The new dual-core supercomputer processor springing forth from the embedded gene pool is called BlueGene/L. At last week's **Fall Processor Forum** (FPF) in San Jose, California, IBM revealed new details about this fascinating chip. It's destined for an awesome supercomputer of the same name, which will harness the power of 65,536 processor chips (containing 131,072 PowerPC processor cores) and 32 terabytes (TB) of main memory. When the first BlueGene/L supercomputer is finished next year, IBM expects it to deliver peak performance of 360 trillion floating-point operations per second (teraflops). And it will run at only 700MHz, an embedded-realm clock frequency that would provoke snickers from PC users.

IBM is developing the BlueGene/L supercomputer for three customers so far: Lawrence Livermore National Laboratory in California; ASTRON, an astronomy organization in the Netherlands; and the National Institute of Advanced Industrial Science and Technology (AIST) in Japan, which is acquiring a smaller version of the computer. BlueGene/L will tackle a variety of compute-intensive scientific problems—including, perhaps, the so-called "grand challenge" science projects, such as the union of fundamental forces and simulating life in silicon (IVIS: In Vivo–In Silico). Other possible applications are nuclear-weapon research, hydrodynamics, quantum chemistry, molecular dynamics, climate modeling, and financial modeling.

To put BlueGene/L's performance in perspective, the world's fastest supercomputer is currently NEC's 5,120-processor Earth Simulator in Japan, which delivers 35.8 teraflops in the Linpack benchmark, according to the supercomputer scoreboard at *www.top500.org*. California Digital's Thunder supercomputer at Lawrence Livermore, with 4,096 Intel Itanium-2 processors, ranks second, with 19.9 Linpack teraflops. Hewlett-Packard's 8,192-processor ASCI-Q AlphaServer at Los Alamos National Laboratory ranks third, with 13.8 Linpack teraflops. On September 29, IBM announced that a BlueGene/L prototype sustained 36.01 Linpack teraflops during internal testing at IBM's lab in Rochester, Minnesota. That unofficial benchmark edges out the first-place Earth Simulator, which occupies 100 times as much space and consumes 28 times as much power as IBM's prototype.

Note that the BlueGene/L prototype has 8,192 dual-core processor chips—only 12% as many chips as envisioned. When IBM builds out the machine to its full complement of 65,536 chips with 131,072 processor cores, its theoretical peak

performance will surpass 360 teraflops. Sustained Linpack performance won't reach that stratosphere; nevertheless, BlueGene/L will be a significant leap forward in computing power.

Leveraging the Virtues of Simplicity

During his FPF presentation, IBM's Alan Gara, chief architect of BlueGene/L, explained why it makes sense to use embedded SoCs as inspiration for a supercomputer processor. For one thing, although the cost and power-consumption budgets for a world-class supercomputer are lavish, a machine with more than 131,000 processors must still strive to control both budgets. Second, an embedded-derived design reduces the system's complexity, which already verges on the overwhelming because of the machine's vast scale. Third, embedded-design principles reduce development risk and time to market—the competition among supercomputer architects is fierce, and fame is fleeting. Fourth, it's easier to maintain reliability, availability, and serviceability (RAS) with a leaner, simpler supercomputer. Finally, a machine based on a proven PowerPC processor core is programmable with familiar development tools.

Ganging thousands of small processors together is productive, because most supercomputer applications have great inherent parallelism and are highly scalable. Grand science projects tend to involve repetitive number crunching on huge datasets. IBM argues that existing supercomputers, based on increasingly complex superscalar processors, are reaching their limits of cost, power consumption, compute performance, and RAS. As other papers presented at FPF demonstrated, the vendors of server processors, PC processors, and embedded processors are rapidly moving toward multicore chip designs. It's logical for supercomputer processors to follow the same path.

IBM's foundation for BlueGene/L is the PowerPC 440 processor core. Introduced at Microprocessor Forum 1999, the PowerPC 440 is a three-way superscalar design with out-of-order execution, seven-stage integer pipelines, dynamic branch prediction, and a single-cycle 32-bit multiplier. It was the first officially announced embedded-processor core to crack 1,000 Dhrystone MIPS and the first to implement IBM and Motorola's Book E extensions. (See *MPR 10/25/99-03*, "IBM PowerPC 440 Hits 1,000 MIPS.") Its nominal clock frequency five years ago in IBM's 0.18-micron copper CMOS process was 555MHz, so the core isn't working hard to reach the target clock speed of 700MHz in BlueGene/L. In fact, IBM is fabricating BlueGene/L chips in 0.13-micron copper CMOS, not in the latest 90nm process. This chip is a relatively conservative design with room to grow.

Floating-point performance is paramount in a supercomputer, so IBM souped up the integer-only PowerPC 440 core with what it calls a "double hummer" FPU. IBM simply duplicated a single-pipelined 64-bit FPU—a standard component from IBM's Blue Logic IP library—to create twin 64-bit FPUs. Also cloned was the floating-point register file, so there are two independent sets of 32 registers, 64 bits wide. Because the double-hummer FPU can crunch 64-bit numbers at the same speed it can manipulate 32-bit numbers, most programs will probably use 64-bit math.

Programmers can direct a floating-point operation to execute in either pipeline, simultaneously execute the same operation in both pipelines, or simultaneously execute two different operations on different operands. However, some restrictions apply. Not all the standard floating-point instructions can execute simultaneously in the twin pipelines. To compensate, IBM added some still-undisclosed new instructions to the PowerPC instruction set. Some of those instructions can execute in parallel, as if they are a single instruction. IBM says the new instructions are optimized for non-SIMD operations typically found in linear algebra.

The supercomputer's theoretical peak performance of more than 360 teraflops is based on executing a 64-bit fused multiply-add (FMA) instruction in each FPU pipeline per clock cycle. Each double-hummer FPU has two pipelines, and each BlueGene/L chip has two PowerPC 440 cores, so peak performance at 700MHz is 2.8GFLOPS per core and 5.6GFLOPS per chip. As Figure 1 shows, the next building block in the system is a compute card with two chips. Sixteen compute cards will plug into a

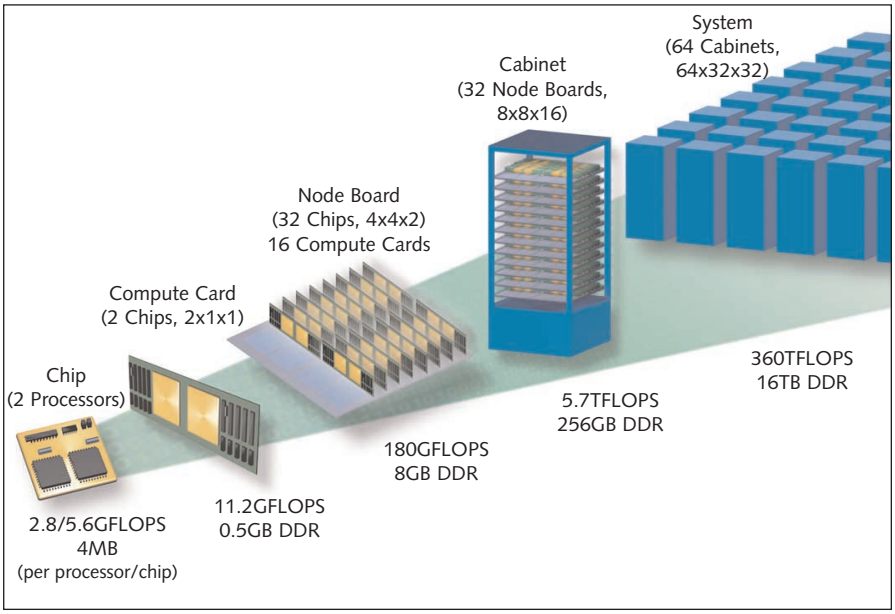

Figure 1. When installation is finished at Lawrence Livermore National Laboratory in 2005, the first BlueGene/L supercomputer will have 65,536 BlueGene/L chips with a total of 131,072 PowerPC 440 processor cores. Peak performance will exceed 360 32-bit teraflops. Source: IBM

single node board, and 32 node boards will fill one cabinet; the finished machine will have 64 cabinets. It all adds up to about 367 teraflops.

On-Chip Memory Is Deep and Wide

Except for the double-hummer FPUs, the PowerPC 440 cores in BlueGene/L are the same cores IBM licenses to customers for their embedded SoCs and ASICs. In fact, it's the same core whose encrypted model is available as a free download for evaluation purposes from IBM's website. (See *MPR 4/26/04-02*, "IBM Loosens Up CPU Licensing.") Outside the core, however, the BlueGene/L chip starts looking less and less like an embedded SoC—or, at least, like any SoC we've seen for everyday embedded applications. The chip's memory system and I/O resources are definitely supercomputer class.

To begin with, IBM has supplemented the standard CPU caches with two additional levels of caches and a shared block of SRAM, all integrated on chip. Each CPU core has 32KB instruction and data caches and a 2KB L2 cache. That's not a typo—the L2 caches really are 2KB, not 2MB. They are surprisingly small for a modern microprocessor, especially for a supercomputer processor, but IBM says they're really a pair of stream buffers for the on-chip L3 cache. Each L2 cache can buffer 16 L3-size cache lines, at 128 bytes per line. Datapaths between the CPUs and L2 caches are 128 bits wide and run at half the core speed, so maximum bandwidth is 5.5GB/s per port. The dual CPUs can snoop both L2 caches to maintain coherency, and the cache latency is 11 clock cycles. However, the L1 caches are not coherent, so programmers must manage L1 coherency in software.

Each L2 cache has a pair of 256-bit pathways (one for each CPU core) to the shared on-chip L3 cache, which has 4MB of embedded DRAM. This huge cache is divided into two banks with 128-byte lines and is eight-way set-associative. L3 cache latency at 350MHz ranges from 28 clock cycles to 40.

On the back-side of the L3 cache is a 128-bit DDR SDRAM interface (144 bits, including ECC) that also runs at half the core frequency. It provides 5.5GB/s of bandwidth to main memory, with a latency of 86 clock cycles. Each BlueGene/L chip can address 256MB to 1GB of memory, depending on DRAM densities. The 65,536-chip BlueGene/L supercomputer at Lawrence Livermore will use 512Mb DRAM chips, for a system total of 32TB.

Each chip also has 16KB of dual-ported shared SRAM, which the CPUs can access over a 128-bit pathway. The SRAM connects to a JTAG interface, which is useful for debugging programs and for real-time monitoring during the computer's operation. Processors can exchange control information more efficiently by using the SRAM as a mailbox, and the JTAG interface keeps this traffic off the other networks carrying program instructions and data. Figure 2 shows a block diagram of a BlueGene/L chip.

Proprietary Networks Unite the CPUs

Several I/O interfaces tie the chip into the supercomputer's large array of processors. Indeed, each chip is a node in five independent networks: a 3D Torus fabric, which provides point-to-point connections among the processors; a hierarchical-tree collective network for global operations; a low-latency network dedicated solely to carrying global interrupts and operation barriers; a Gigabit Ethernet network for file I/O; and a control network for booting the system, monitoring its operation, and running diagnostics. Figure 3 illustrates the topologies of these five networks.

For conventional file I/O, the BlueGene/L chip has a Gigabit Ethernet interface. The other network I/O interfaces are proprietary. For the 3D Torus network, each chip has six input links and six output links. Each link is one bit wide and clocked at 1.4GHz, providing 2.8Gb/s of bidirectional bandwidth per link, or about 2.1GB/s of aggregate bandwidth per chip.

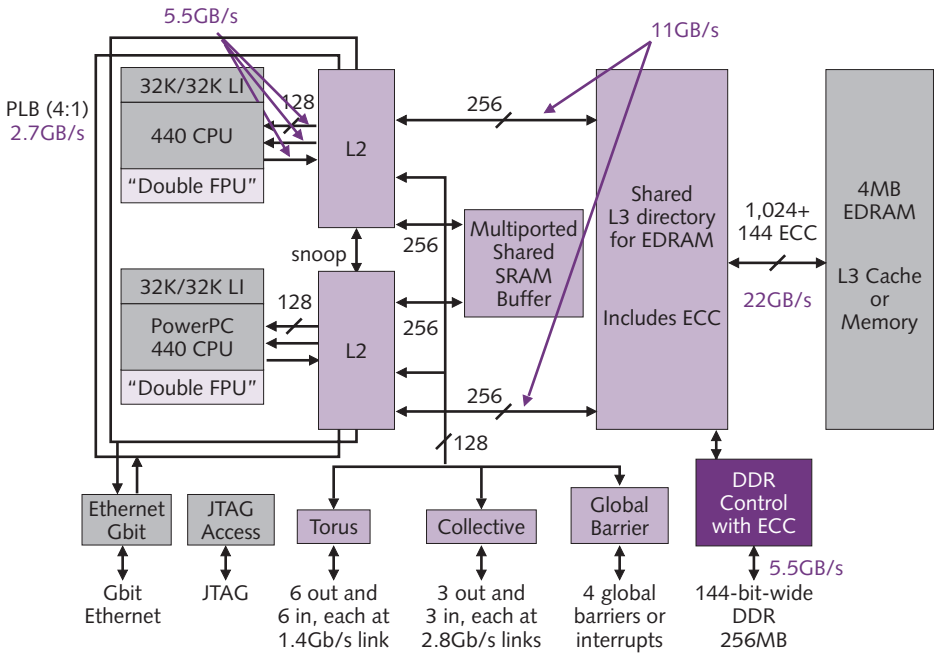

Figure 2. Each BlueGene/L chip has dual PowerPC 440 processor cores with IBM's special "double hummer" FPUs. Note the deep on-chip memory system: three levels of CPU caches, including 4MB of embedded DRAM in the L3 cache, plus 16KB of shared SRAM for exchanging control information among the cores. ECC parity protects data integrity throughout the memory system. Wide datapaths connect the memory hierarchy and typically run at 350MHz, half the CPU core frequency.

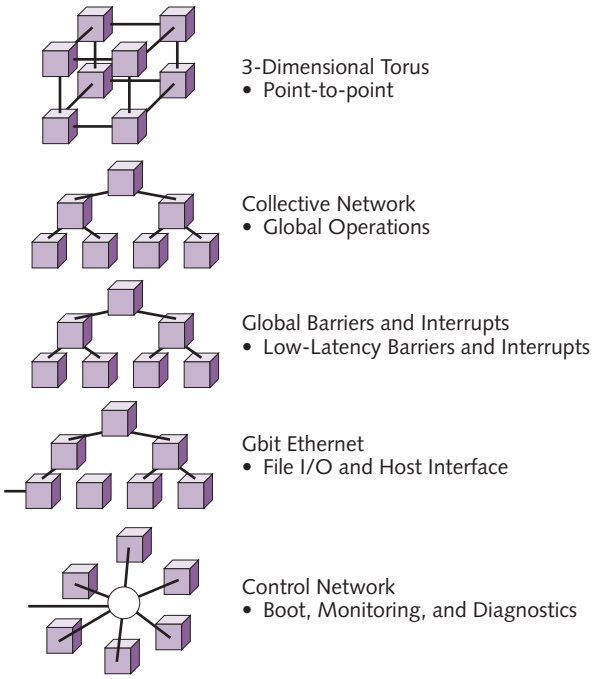

Figure 3. The finished BlueGene/L supercomputer will have 65,536 BlueGene/L processor chips, and each chip is a node in five independent networks. Each network serves a different purpose, such as data sharing (local and global), interrupt control, file I/O, and central control. Separate I/O interfaces on the chips tie the networks together.

Note that the Torus links operate at twice the clock speed of the CPU cores—a remarkable feature. This is the first microprocessor we've seen that drives an I/O interface faster than the CPU core clock frequency. It's a critical interface, because the Torus network is the main fabric stitching the 65,536 chips together, supporting point-to-point communications with a latency of about 100ns per hop. Like the Torus interface, the Torus network protocol is proprietary, with packets ranging in length from 32 bytes to 256. The protocol includes cyclic redundancy checking (CRC) and provisions for retransmitting faulty packets.

Each chip has three input and three output links for the collective network, and each link is eight bits wide and clocked at 350MHz, providing 350MB/s of bandwidth per link. This proprietary network broadcasts messages to the entire processor array, and it also connects the processors to the file I/O system. It's built on a fault-tolerant tree topology, with redundant pathways between the computer's node cards.

A separate JTAG interface carries control signals to each chip's SRAM mailbox, monitors the operation of the processors, and allows the operators to run diagnostics. Along with the Gigabit Ethernet and DDR memory interfaces previously described, these proprietary network interfaces constitute the chip's full complement of I/O resources. It has none of the industry-standard I/O interfaces found in other high-performance embedded processors, such as PCI-X, HyperTransport, PCI Express, or SPI-4.2. Clearly, IBM has designed the BlueGene/L chip for a single-minded purpose: scientific supercomputing.

Cell Block	Size (Cells)	Size %	Active Power	Power %
Clock Tree + Access	264K	0.5%	1.15 W	8.91%
CPU, FPU, L1 Cache	14,700K	28.7%	7.54 W	58.45%
3D Torus Network	4,963K	9.7%	0.67 W	5.19%
Collective Network	2,350K	4.6%	0.25 W	1.94%
L2, L3, DDR Ctrl	18,310K	35.7%	2.60 W	20.16%
Other Blocks	10,720K	20.9%	0.49 W	3.80%
Leakage	—	—	0.20 W	1.55%
Total	51,307K	100%	12.9 W	100%

Table 1. BlueGene/L power consumption is reasonable for a microprocessor of its capabilities fabricated in a 0.13-micron process. Nevertheless, at 12.9W active power per chip, the finished BlueGene/L supercomputer will consume more than 845kW in the processors alone, not counting other components and main memory, which will be 32TB. If Berkeley residents in the neighborhood of Lawrence Livermore National Laboratory notice their lights dimming next year, maybe they will blame IBM this time instead of Enron.

Supercomputing Justifies Big Iron

The BlueGene/L chip isn't small by embedded standards, but it's not outsize by any means, especially for a cell-based design. The square die measures 123mm^2 and contains 95 million transistors. We figure the 4MB of embedded DRAM in the on-chip L3 cache accounts for a third of those transistors, even before adding the SRAM transistors in the L1 and L2 caches.

IBM says the on-chip memory and DDR controller account for about 36% of the chip's cells, but only about 20% of the chip's active power consumption. The dual CPU cores account for about 29% of the chip's cells and 58% of the power. Total power consumption is less than 13W per chip. We estimate that the whole supercomputer, when fully loaded with memory, will require nearly 2 megawatts (MW) of power. That's a lot—but NEC's Earth Simulator requires 6MW, and BlueGene/L will be much faster. Table 1 shows a breakdown of the PowerPC chip's cell blocks and power consumption.

Some critics are questioning the need for such large supercomputers. There's a growing movement in favor of distributed supercomputing, also known as grid computing. By linking thousands or even millions of ordinary PCs together in a network, it's possible to assemble the equivalent of a huge parallel-processing supercomputer. The PCs may be dedicated for this purpose or do the number crunching as a background task, harvesting CPU cycles that would otherwise be wasted. Grid computing takes advantage of the same inherent parallelism in scientific applications that supercomputers like BlueGene/L exploit.

Grid computing is a proven concept. Perhaps the largest example is Oxford University's Cancer Drug Discovery Project, which has linked more than a million PCs together to analyze potential cancer-fighting treatments. Better known is the SETI @ Home Project (Search for

Extraterrestrial Intelligence). It has more than 500,000 active Internet users, who volunteer to let their PCs run a background program that analyzes radio signals received from outer space. Another well-known project is System X at the Virginia Polytechnic Institute and State University (Virginia Tech). In only 90 days last year, a team of students, faculty, administrators, and volunteers linked 1,100 Macintosh G5 desktop computers into a terascale cluster that, for a while, ranked as the world's third-fastest supercomputer.

Now grid computing is becoming a fad of sorts. One Saturday last April at the University of San Francisco (USF), a "flash mob" of students and other enthusiasts gathered in a gym and linked 669 PCs together to build a supercomputer for the weekend. (If this sounds frivolous, just recall the days when college students competed to stuff themselves into Volkswagens.)

However, grid computing has several limitations that preserve the need for big-iron supercomputers like BlueGene/L. Network latencies are much greater, especially for grid nodes linked over the public Internet. For some applications, such as SETI, it doesn't matter much—the little green men can wait. For other applications, such as real-time weather forecasting, it could matter a lot—hurricanes don't wait. In addition, ad hoc networks may lack the redundancy and error checking implemented throughout a system like BlueGene/L.

A dedicated supercomputer can deliver more compute performance per watt than a grid can, but most of the power consumed by a grid would go to waste anyway, and the power consumption may be more dispersed, so it's not a major factor. More important, for a research center like Lawrence Livermore, is security. Some applications, like nuclear-weapon simulations, don't lend themselves to dissemination over public networks. Although supercomputer clusters cobbled together by flash mobs have a certain populist appeal, they're hard to manage—nothing much was accomplished by the USF experiment. And an often-overlooked drawback is the cost of the pizza.

BlueGene/L is IBM's bid to reclaim the top spot in big-iron supercomputing. The PowerPC chip IBM has designed as the fundamental building block is an intriguing combination of low-power embedded-processor cores, pumped-up FPUs, generous on-chip memory, and special-purpose I/O interfaces. There's nothing quite like it. BlueGene/L definitely deserved its place in the Cool Technology session at FPF. ◇

For More Information

IBM Microelectronics currently has no plans to offer the BlueGene/L chips for sale on the merchant market. Chips are in production now, and the first BlueGene/L supercomputer is scheduled for completion in 2005. For more information about BlueGene/L, see *http://researchweb.watson.ibm.com/bluegene/*.

The Top 500 supercomputer scoreboard is updated periodically and maintained by the University of Mannheim, the University of Tennessee, and the National Energy Research Scientific Computing Center at Lawrence Berkeley National Laboratory. To view the latest list, visit *www.top500.org* and click on "Current List" on the navigation bar near the top of the web page.