10.4 Creating the BlueGene/L Supercomputer from Low-Power SoC ASICs.

Arthur A. Bright[1], Matthew R. Ellavsky[2], Alan Gara[1], Ruud A. Haring[1], Gerard V. Kopcsay[1], Robert F. Lembach[2], James A. Marcella[2], Martin Ohmacht[1], Valentina Salapura[1],

[1]IBM, Yorktown Heights, NY
[2]IBM, Rochester, MN

Traditional supercomputer design, using large powerful networks coupled to state-of-the-art processors, is hitting power and cost limits. The major drivers for cost are increased system complexity and stringent power requirements. Current directions in integration, power, and technology are driving toward using multiple modest cores on a single chip rather than one high-performance processor. This is especially significant as W/FLOPS ratio will not significantly improve with future technologies. These considerations motivated the pursuit of BlueGene/L as a massively parallel system design, based on embedded PowerPC processors with a relatively modest clock rate. By adopting a cost-effective SoC approach, a low-cost design with low power consumption is realized, allowing for aggressive system integration.

The architecture of the BlueGene/L Compute (BLC) chip is illustrated in Fig. 10.4.1. The BLC chip integrates two symmetric PowerPC 440 embedded processor cores, each with an attached dual-pipeline floating-point unit (FPU), a full L1/L2/L3 3-level memory hierarchy, an interface to external DDR DRAM, and five different communication network interfaces on a single chip. Two of the chip-to-chip communications networks in the BlueGene/L system, the Torus and the Collective networks [1], utilize high-speed serial I/Os and incorporate sophisticated routers on each chip. An overview of characteristics of the BLC chip is given in Fig. 10.4.2.

The BLC ASIC incorporates a full complement of SoC features, including hard and soft cores, custom logic, SRAM and embedded DRAM. The ASIC is implemented in a 0.13μm generation CMOS process, and has 95M transistors on a 123mm^2 die. The chip operates at up to 700MHz. The floor plan for the chip is driven by several constraints:

• To meet the low latency requirement between the two PowerPC processor cores and L2 caches, the processor cores are placed surrounding the two L2 cache units. The L2 cache units communicate at high bandwidth with several other units, requiring a large number of wires. At 7 layers of metal, the wiring capability of the technology could accommodate these buses.

• Because of the large number of wide buses in the L3 cache area and the need to keep the controller logic compact for low latency, the four eDRAM macros are arranged with two on each side of the L3 controller logic.

• The high I/O count for the external DDR DRAM interface require the use of most of the available area array pin locations in the lower half of the chip, including many superimposed on the eDRAM macros.

• Technology constraints on I/O placement drove the 1.5V high-speed serial communication I/Os to be placed in one quadrant, and the remaining units with 2.5V off-chip interfaces in the other three quadrants.

The utilization of chip area and a power break-down are illustrated in Fig. 10.4.3. Over half of the chip is consumed by the hard cores and embedded DRAMs. Other fixed components, such as I/O cells, decoupling capacitors, fuse macros, and SRAMs, occupy another quarter. About 10% of the area is used for custom logic.

To implement 4MB L3 cache, 4 embedded DRAM macros are used with a 1024b wide data port in aggregate. Operating at 1/4 frequency of the processor cores, the bandwidth of the embedded DRAM is matched to the 32B/cycle total data load capability of both floating-point units simultaneously. Thus, a large amount of cache is provided with a low access-latency of just over 40 processor cycles. The access latency can be largely hidden using aggressive L2 and L3 pre-fetching schemes. All memory is ECC protected.

The BlueGene/L supercomputer runs off a single master clock distributed over the whole system. Therefore, individual chips run synchronously and chip-to-chip communications only have to allow for phase differences with limited drift due to temperature and voltage variations. The oscillator for the BLC chip is received at 700MHz, and is used as the reference signal for an on-chip PLL with an output frequency of 1.4GHz, as illustrated in Fig. 10.4.4. The PLL output is divided down to several frequencies that maintain a well-defined frequency and phase relationship to each other.

The clock-design methodology for the BLC chip is based on an ASIC clock-distribution methodology which efficiently routes low-skew trees to latches. The clock tree is designed and maintained separately from the rest of the logic on the chip. Clock-gating signals, test-control signals, and frequency dividers are kept within the clock tree and drive idealized clock splitters. In the physical design stage, idealized clock splitters are converted to real clock splitters, and balancing, re-powering and skew minimization are performed. While this approach is very flexible and efficient, it only allows for clock gating at a subsystem level.

The ASICs approach to physical design relies on automated placement and wiring tools to achieve performance and area with as little intervention as possible. However, some parts of the BLC chip design demand exceptional performance and timing uniformity. This is achieved by assembling selected components into "bit stacks", i.e. clusters of custom placed components. A portion of a bit stack on the BLC chip is shown in Fig.10.4.5. The illustrated design is used for receiving an off-chip signal for either the Torus or the Collective network. The incoming data arrives with arbitrary phase and is aligned to the on-chip clock using a delay line implemented as a chain of inverters. PowerSpice is used to determine logic cell strength, clock fan-out strategies, and decoupling. Simulation results are close enough to the static timing that no post wiring changes are required.

In conclusion, the BLC chip integrates processors, memory and interconnect subsystems on a single low-power chip. Figure 10.4.6 is a micrograph of the chip, with major units indicated. Each BlueGene/L node, which includes a BLC chip with associated 512MB of DRAM, delivers a peak performance of 5.6 GFLOPS at less than 13W. This low-power approach allows air cooling even at a very high system packaging density, resulting in unmatched system performance per watt, per square meter of floor space, and per dollar.

Acknowledgement:
This work has benefited from the cooperation of many individuals in IBM Research (Yorktown Heights, NY), IBM Engineering & Technology Services (Rochester, MN), and IBM Microelectronics (Burlington, VT and Raleigh, NC). The BlueGene/L project is supported and partially funded by the Lawrence Livermore National Laboratories on behalf of the United States Department of Energy, under LLNL Subcontract No. B517552.

References:
[1] The BlueGene/L Team, "An Overview of the BlueGene/L Supercomputer," *Proc. of Supercomputing Conference*, Nov., 2002.
[2] The BlueGene/L Team, "Cellular Supercomputing with System-On-A-Chip," *in Proc. of ISSCC Dig. Tech. Papers*, vol. 2, pp. 152-153 Feb., 2002.
[3] Alan Gara, "The BlueGene/L Processor for Massively Parallel Supercomputing," *Fall Processor Forum,* Oct., 2004.

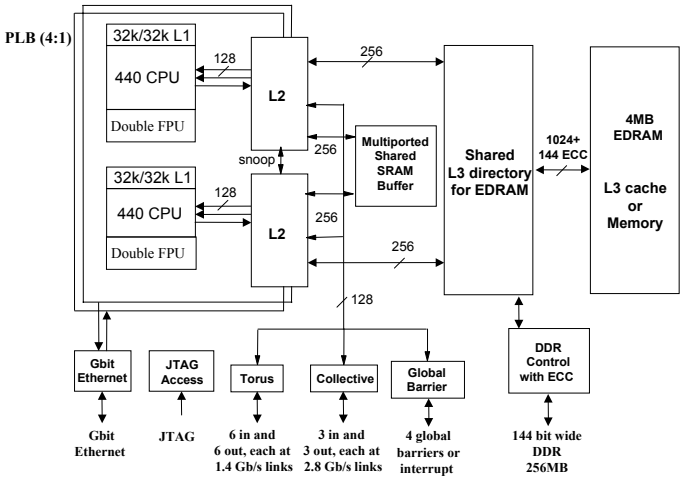

Figure 10.4.1: BlueGene/L compute chip schematic.

System Features		BG/L
Node Properties	Node Processors	2* 440 PowerPC
	Processor Frequency	700MHz
	L1 Cache (private)	32KB/processor
	L2 Cache (private)	14 stream prefetching
	L3 Cache size (shared)	4MB
	Main Store	256MB/512MB/1GB
	Main Store Bandwidth	5.6GB/s
	Peak Performance	5.6GF/node
Torus Network	Bandwidth	6*2*175MB/s=2.1GB/s
	Hardware Latency (Nearest Neighbor)	200ns (32B packet) 1.6us(256B packet)
	Hardware Latency (Worst Case)	6.4us (64 hops)
Collective Network	Bandwidth	3*2*350MB/s=2.1GB/s
	Hardware Latency (round trip worst case)	5.0μs

Figure 10.4.2: BlueGene/L properties.

Unit	Size (cells)	Size %	Active power	Power %
Clock tree + access	264k	0.5%	1.15W	8.91%
CPU/FPU/L1	14,700k	28.7%	7.54W	58.45%
Torus network	4,963k	9.7%	0.67W	5.19%
Collective network	2,350k	4.6%	0.25W	1.94%
L2/L3/DDR control	18,310k	35.7%	2.60W	20.16%
Others	10,720k	20.9%	0.49W	3.80%
Leakage			0.20W	1.55%
Total	51,307k	100%	12.90W	100%

Figure 10.4.3: Area usage and power consumption of the BLC chip.

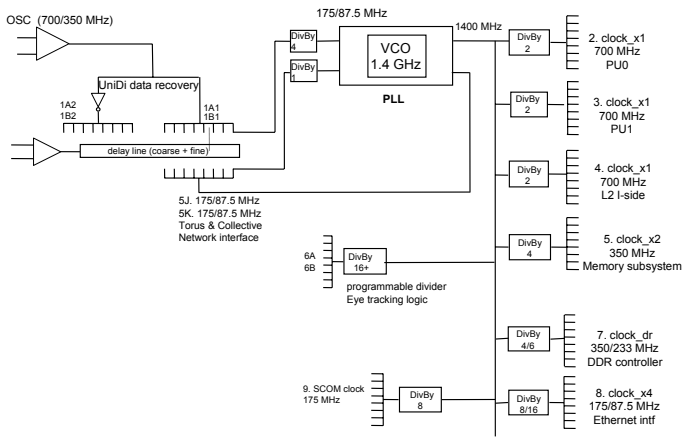

Figure 10.4.4: Clock tree for the BlueGene/L chip.

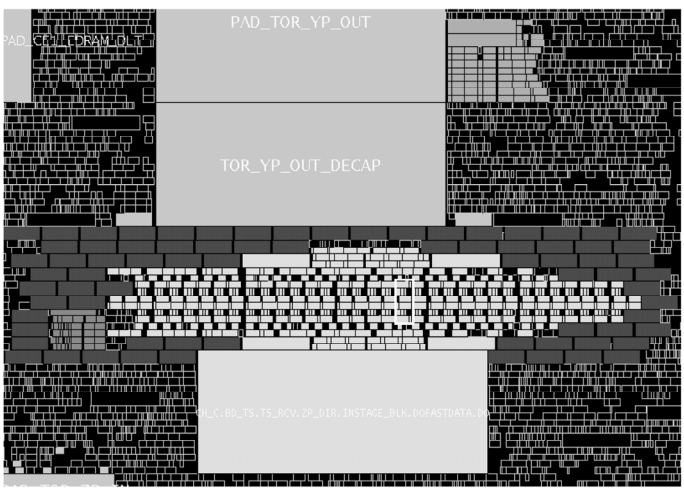

Figure 10.4.5: Bit stack for the high-speed data reception macro.

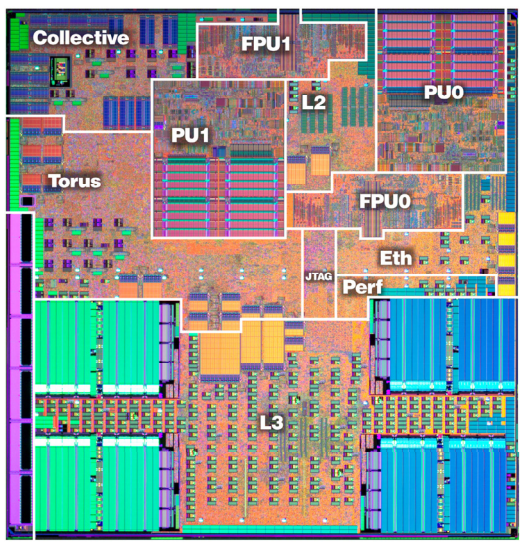

Figure 10.4.6: BlueGene/L Compute chip DD2.0 die micrograph.