Packaging the Blue Gene/L supercomputer

P. Coteus
H. R. Bickford
T. M. Cipolla
P. G. Crumley
A. Gara
S. A. Hall
G. V. Kopcsay
A. P. Lanzetta
L. S. Mok
R. Rand
R. Swetz
T. Takken
P. La Rocca
C. Marroquin
P. R. Germann
M. J. Jeanson

As 1999 ended, IBM announced its intention to construct a one-petaflop supercomputer. The construction of this system was based on a ***cellular architecture****—the use of relatively small but powerful building blocks used together in sufficient quantities to construct large systems. The first step on the road to a petaflop machine (one quadrillion floating-point operations in a second) is the Blue Gene®/L supercomputer. Blue Gene/L combines a low-power processor with a highly parallel architecture to achieve unparalleled computing performance per unit volume. Implementing the Blue Gene/L packaging involved trading off considerations of cost, power, cooling, signaling, electromagnetic radiation, mechanics, component selection, cabling, reliability, service strategy, risk, and schedule. This paper describes how 1,024 dual-processor compute application-specific integrated circuits (ASICs) are packaged in a scalable rack, and how racks are combined and augmented with host computers and remote storage. The Blue Gene/L interconnect, power, cooling, and control systems are described individually and as part of the synergistic whole.*

Overview

Late in 1999, IBM announced its intention to construct a one-petaflop supercomputer [1]. Blue Gene*/L (BG/L) is a massively parallel supercomputer developed at the IBM Thomas J. Watson Research Center in partnership with the IBM Engineering and Technology Services Division, under contract with the Lawrence Livermore National Laboratory (LLNL). The largest system now being assembled consists of 65,536 (64Ki[1]) compute nodes in 64 racks, which can be organized as several different systems, each running a single software image. Each node consists of a low-power, dual-processor, system-on-a-chip application-specific integrated circuit (ASIC)—the BG/L compute ASIC (BLC or *compute chip*)—and its associated external memory. These nodes are connected with five networks, a three-dimensional (3D) torus, a global collective network, a global barrier and interrupt network, an input/output (I/O) network which uses Gigabit Ethernet, and a service network formed of Fast Ethernet (100 Mb) and JTAG (IEEE Standard 1149.1 developed by the Joint Test Action Group). The implementation of these networks resulted in an additional link ASIC, a control field-programmable gate array (FPGA), six major circuit cards, and custom designs for a rack, cable network, clock network, power system, and cooling system—all part of the BG/L core complex.

System overview

Although an entire BG/L system can be configured to run a single application, the usual method is to partition the machine into smaller systems. For example, a 20-rack (20Ki-node) system being assembled at the IBM Thomas J. Watson Research Center can be considered as four rows of four compute racks (16Ki nodes), with a standby set of four racks that can be used for failover. Blue Gene/L also contains two host computers to control the machine and to prepare jobs for launch; I/O racks of redundant arrays of independent disk drives (RAID); and switch racks of Gigabit Ethernet to connect the compute racks, the I/O racks, and the host computers. The host, I/O racks, and switch racks are not described in this paper except in reference to interconnection. **Figure 1** shows a top view of the 65,536 compute processors cabled as a single system.

[1]The unit "Ki" indicates a "kibi"—the binary equivalent of kilo (K). See *http://physics.nist.gov/cuu/Units/binary.html*.

0018-8646/05/$5.00 © 2005 IBM

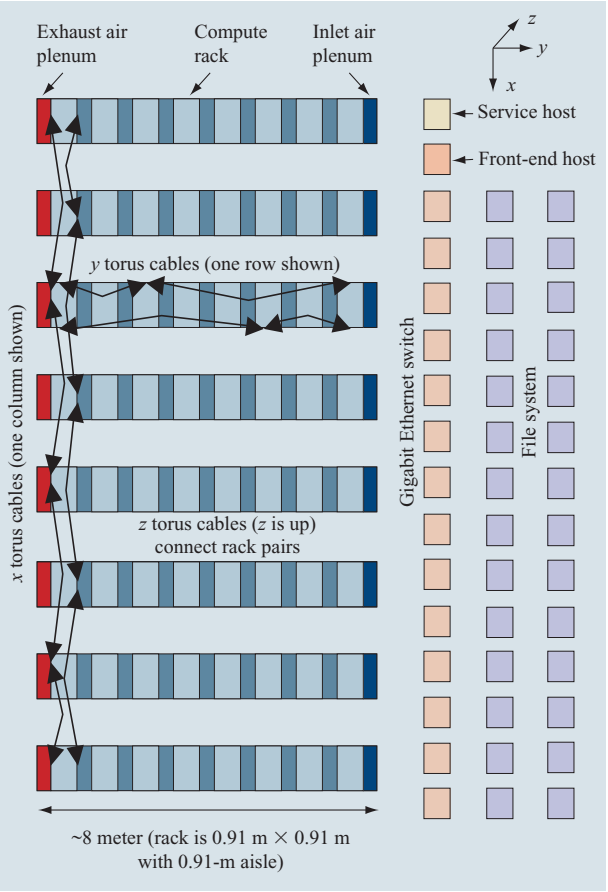


Figure 1

Top view of a conceptual 64-rack Blue Gene/L system.

Cooling system overview

BG/L compute racks are densely packaged by intention and, at ~25 kW, are near the air-cooled thermal limit for many racks in a machine room. No one component had power so high that direct water cooling was warranted. We chose either high or low rack power, airflow direction, and either computer room air conditioners (CRACs) or local rack air conditioners. By choosing the high packaging density of 512 compute nodes per midplane, five of six network connections were made without cables, greatly reducing cost. The resultant vertical midplane had insufficient space for passages to allow front-to-back air cooling. The choices were then bottom-to-top airflow or an unusual side-to-side airflow. The latter offered certain aerodynamic and thermal advantages, although there were challenges. Since the populated midplane is relatively compact (0.64 m tall × 0.8 m deep × 0.5 m wide), two fit into a 2-m-tall rack with room left over for cables and for the ac–dc bulk power supplies, but without room for local air conditioners. Since most ac–dc power supplies are designed for front-to-back air cooling, by choosing the standard CRAC-based machine-room cooling, inexpensive bulk-power technology in the rack could be used and could easily coexist with the host computers, Ethernet switch, and disk arrays. The section below on the cooling system gives details of the thermal design and measured cooling performance.

Signaling and clocking overview

BG/L is fundamentally a vast array of low-power compute nodes connected together with several specialized networks. Low latency and low power were critical design considerations. To keep power low and to avoid the clock jitter associated with phase-locked loops (PLLs), a single-frequency clock was distributed to all processing nodes. This allowed a serial data transfer at twice the clock rate whereby the sender can drive data on a single differential wiring pair using both edges of its received clock, and the receiver captures data with both edges of its received clock. There is no requirement for clock phase, high-power clock extraction, or clock forwarding, which could double the required number of cables. Lower-frequency clocks were created by clock division, as required for the embedded processor, memory system, and other logic of the BG/L compute ASIC. The clock frequency of the entire system is easily changed by changing the master clock. The signaling and clocking section discusses the estimated and measured components of signal timing and the design and construction of the global clock network.

Circuit cards and interconnect overview

The BG/L torus interconnect [2] requires a node to be connected to its six nearest neighbors ($x+$, $y+$, $z+$, $x-$, $y-$, $z-$) in a logical 3D Cartesian array. In addition, a global collective network connects all nodes in a branching, fan-out topology. Finally, a 16-byte-wide data bus to external memory was desired. After considering many possible packaging options, we chose a first-level package of dimensions 32 mm × 25 mm containing a total of 474 pins, with 328 signals for the memory interface, a bit-serial torus bus, a three-port double-bit-wide bus for forming a global collective network, and four global `OR` signals for fast asynchronous barriers. The 25-mm height allowed the construction of a small field-replaceable card, not unlike a small dual inline memory module (DIMM), consisting of two compute ASICs and nine double-data-rate synchronous dynamic random access memory chips (DDR SDRAMs) per ASIC, as shown in **Figure 2**. The external memory system can transfer one 16-byte data line for every two processor cycles. The ninth two-byte-wide chip allows a 288-bit error-correction code with

four-bit single packet correct, a double four-bit packet detect, and redundant four-bit steering, as used in many high-end servers [3]. The DDR DRAMs were soldered for improved reliability.

A node card (**Figure 3**) supports 16 compute cards with power, clock, and control, and combines the 32 nodes as $x, y, z = 4, 4, 2$. Each node card can optionally accept up to two additional cards, similar to the compute cards, but each providing two channels of Gigabit Ethernet to form a dual-processor I/O card for interface-to-disk storage. Further, 16 node cards are combined by a midplane ($x, y, z = 8, 8, 8$). The midplane is the largest card that is considered industry-standard and can be built easily by multiple suppliers. Similarly, node cards are the largest high-volume size.

To connect the midplanes together with the torus, global collective network, and global barrier networks, it was necessary to rebuffer the signals at the edge of the midplane and send them over cables to other midplanes. To provide this function at low power and high reliability, the BG/L link ASIC (BLL or *link chip*) was designed. Each of its six ports drives or receives one differential cable (22 conductor pairs). The six ports allow an extra set of cables to be installed, so signals are directed to either of two different paths when leaving a midplane. This provides both the ability to partition the machine into a variety of smaller systems and to "skip over" disabled racks. Within the link chip, the ports are combined with a crossbar switch that allows any input to go to any output. BG/L cables are designed never to move once they are installed except to service a failed BLL, which is expected to be exceedingly rare (two per three-year period for the 64Ki-node machine). Nevertheless, cable failures could occur, for instance, due to solder joint fails. An extra synchronous and asynchronous connection is provided for each BLL port, and it can be used under software control to replace a failed differential pair connection.

Midplanes are arranged vertically in the rack, one above the other, and are accessed from the front and rear of the rack. Besides the 16 node cards, each with 32 BG/L compute ASICs, each midplane contains four link cards. Each link card accepts the data cables and connects these cables to the six BLLs. Finally, each midplane contains a service card that distributes the system clock, provides other rack control function, and consolidates individual Fast Ethernet connections from the four link cards and 16 node cards to a single Gigabit Ethernet link leaving the rack. A second integrated Gigabit Ethernet link allows daisy-chaining of multiple midplanes for control by a single host computer. The control–FPGA chip (CFPGA) is located on the service card and converts Fast Ethernet from each link card and node card to other standard buses, JTAG, and I2C (short for "inter-integrated circuit"); it is a multimaster bus used to connect integrated circuits (ICs). The JTAG and I2C buses of the CFPGA connect respectively to the ASICs and various sensors or support devices for initialization, debug, monitoring, and other access functions. The components are shown in **Figure 4** and are listed in **Table 1**.

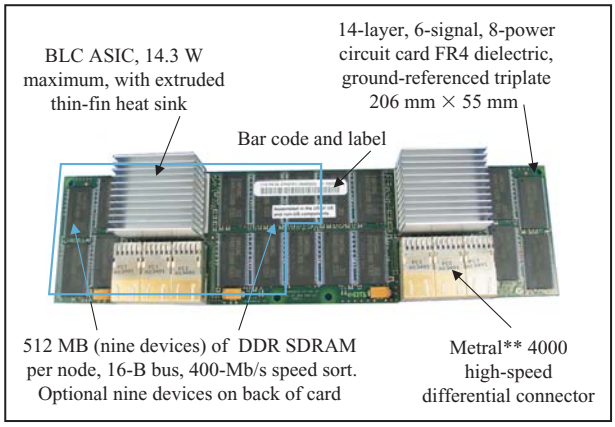


Figure 2

Blue Gene/L compute card.

Figure 3

Blue Gene/L node card.

Power system overview

The BG/L power system ensures high availability through the use of redundant, high-reliability components. The system is built from either commercially available components or custom derivatives of them. A two-stage power conversion is used. An $N+1$[2] redundant bulk power conversion from three-phase 208 VAC to 48 VDC (400 V three-phase to 48 V for other countries, including

[2]We require six to operate and have one extra. Normally we use all seven, unless one fails.

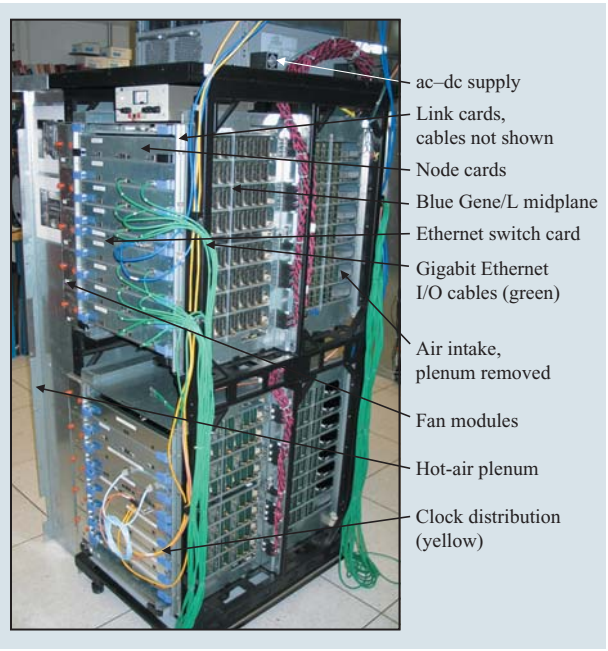


Figure 4

Blue Gene/L compute rack.

Europe and China) is followed by local point-of-load dc–dc converters. The current, voltage, and temperature of all power supplies are monitored. A service host computer is used for start-up, configuration, and monitoring; a local control system is used for each midplane for extreme conditions requiring immediate shutdown. In addition, a current-monitoring loop is employed in concert with soft current limits in the supplies to allow for processor throttling in the event of excessive current draw. Attention is paid to electromagnetic emissions, noise immunity, and standard safety tests (power line disturbance, lightning strikes, Underwriters Laboratories safety tests, etc.). The power distribution systems, both 48 V to the local converters and the local converters to the load, are engineered for low loss and immunity to rapid fluctuations in power demand. There are 128 custom dual-voltage 2.5-V and 1.5-V power converters and 16 custom 1.5-V power converters in each rack, which source about 500 amperes (A) from the 48-V rail under peak load, with minor requirements for 1.2 V, 1.8 V, 3.3 V, and 5.0 V. Some voltages are persistent, while others can be activated and controlled from the host.

Reliability overview

Before moving to the detailed sections, a prediction of the hard failure rate of 64 BG/L compute racks will be made. The term *FIT*—standing for *failures in time*, or failures in parts per million (ppm) per 1,000 power-on hours (POH)—is used. Thus, the BG/L compute ASIC with a hard error rate of 20 FITs after burn-in results in a system failure rate of 20 fails per 10×9 hours $\times$ 65,536 nodes = 1.3 fails per 1,000 hours = 1 fail per 4.5 weeks.

Hard errors are sometimes handled with a redundant unit that seamlessly replaces the failed unit and sends an alarm to the host for future service. This is the case with the ac–dc supplies, dc–dc converters, fans, and a single nibble (four-bit) fail in an external DRAM. Other hard fails, such as a compute ASIC, cause a program fail and machine stop. Most can be handled by repartitioning the BG/L system by programming the BLLs to use the "hot spare" BG/L racks to restore the affected system image.

Table 1 Major Blue Gene/L rack components.

Component	*Description*	*No. per rack*
Node card	16 compute cards, two I/O cards	32
Compute card	Two compute ASICs, DRAM	512
I/O card	Two compute ASICs, DRAM, Gigabit Ethernet	2 to 64, selectable
Link card	Six-port ASICs	8
Service card	One to twenty clock fan-out, two Gigabit Ethernet to 22 Fast Ethernet fan-out, miscellaneous rack functions, 2.5/3.3-V persistent dc	2
Midplane	16 node cards	2
Clock fan-out card	One to ten clock fan-out with and without master oscillator	1
Fan unit	Three fans, local control	20
Power system	ac/dc	1
Compute rack	With fans, ac/dc power	1

Table 2 Uncorrectable hard failure rates of the Blue Gene/L by component.

Component	*FIT per component*[†]	*Components per 64Ki compute node partition*	*FITs per system* (K)	*Failure rate per week*
Control–FPGA complex	160	3,024	484	0.08
DRAM	5	608,256	3,041	0.51
Compute + I/O ASIC	20	66,560	1,331	0.22
Link ASIC	25	3,072	77	0.012
Clock chip	6.5	~1,200	8	0.0013
Nonredundant power supply	500	384	384	0.064
Total (65,536 compute nodes)			5,315	0.89

[†] T = 60°C, V = Nominal, 40K POH. *FIT* = Failures in ppm/KPOH. One FIT = 0.168×16^{-6} fails per week if the machine runs 24 hours a day.

After repartitioning, the state of the machine can be restored from disk with the checkpoint-restart software, and the program continues. Some hard fails, such as those to the clock network or BLLs, may cause a more serious machine error that is not avoided by repartitioning and requires repair before the machine can be restarted.

Table 2 lists the expected failure rates per year for the major BG/L components after recognition of the redundancy afforded by its design. The reliability data comes from the manufacturer estimates of the failure rates of the components, corrected for the effects of our redundancy. For example, double-data-rate synchronous DRAM (DDR SDRAM) components are assigned a failure rate of just 20% of the 25 FITs expected for these devices, since 80% of the fails are expected to be single-bit errors, which are detected and repaired by the BG/L BLC chip memory controller using spare bits in the ninth DRAM.

BG/L cooling system

Introduction to BG/L cooling

BG/L is entirely air-cooled. This choice is appropriate because, although the 25-kW heat load internal to each rack is relatively high for the rack footprint of 0.91 m × 0.91 m (3 ft × 3 ft), the load is generated by many low-power devices rather than a few high-power devices, so watt densities are low. In particular, each of the ASICs used for computation and I/O—though numerous (1,024 compute ASICs and up to 128 I/O ASICs per rack)—was expected to generate a maximum of only 15 W (2.7% higher than the subsequently measured value), which represents a mere 10.4 W/cm^2 with respect to chip area. A multitude of other devices with low power density are contained in each rack: between 9,216 and 10,368 DRAM chips (nine per ASIC), each generating up to 0.5 W; 128 dc–dc power converters, each generating up to 23 W; and a small number of other chips, such as the BLLs. The rack's bulk-power-supply unit, dissipating roughly 2.5 kW and located on top of the rack, is not included in the 25-kW load figure above, because its airflow path is separate from the main body of the rack, as described later.

Side-to-side airflow path

As in many large computers, the BG/L racks stand on a raised floor and are cooled by cold air emerging from beneath them. However, the airflow direction through BG/L racks is unique—a design that is fundamental to the mechanical and electronic packaging of the machine. In general, three choices of airflow direction are possible: side-to-side (SS), bottom-to-top (BT), and front-to-back (FB) (**Figure 5**). Each of these drawings assumes two midplanes, one standing above the other parallel to the yz plane, but the orientation of node cards connecting to the midplane is constrained by the airflow direction. SS airflow requires node cards lying parallel to the xy plane; BT airflow requires node cards lying parallel to the xy plane; FB airflow requires node cards lying parallel to the x-axis.

FB airflow, for which air emerges from perforations in the raised floor as shown, is traditional in raised-floor installations. However, FB airflow is impossible for the BG/L midplane design, because air would have to pass through holes in the midplane that cannot simultaneously be large enough to accommodate the desired airflow rate of 1.4 cubic meters per second (3,000 cubic feet per minute, CFM) and also small enough to accommodate the dense midplane wiring.

Each of the two remaining choices, BT and SS, has advantages. The straight-through airflow path of BT is advantageous when compared with the SS serpentine path, because SS requires space along the $y+$ side of each rack (above the SS raised-floor hole shown) to duct air upward to the cards, and additional space along the

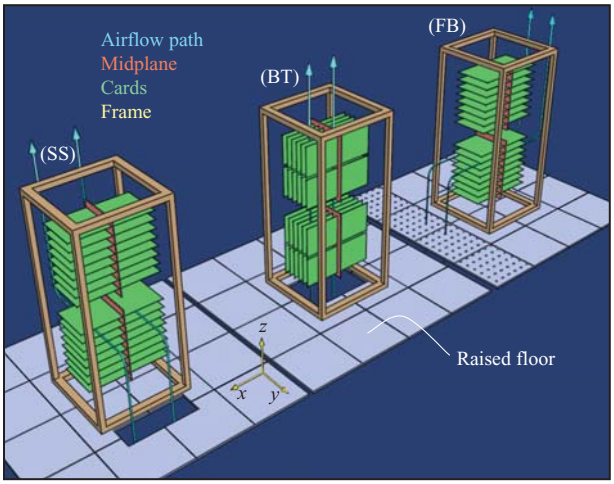


Figure 5

Three alternative airflow directions for the Blue Gene/L compute rack. (The front of the rack is the $+x$ face. (SS = side-to-side; BT = bottom-to-top; FB = front-to-back.)

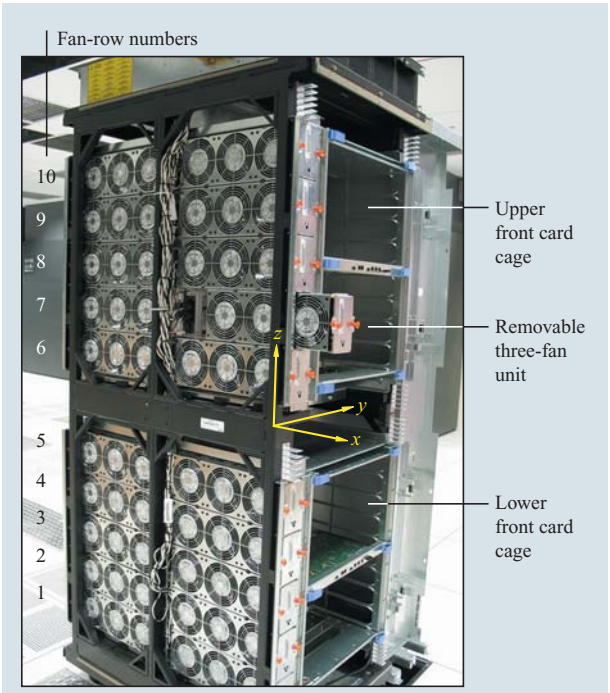


Figure 6

Blue Gene/L fan array.

$y-$ side of each rack to exhaust the air. However, the SS scheme has the advantage of flowing air through an area A_{SS} that is larger than the corresponding area A_{BT} for BT airflow, because a rack is typically taller than it is wide. This advantage may be quantified in terms of the temperature rise ΔT and the pressure drop Δp of the air as it traverses the cards. ΔT is important because, if air flows across N identical heat-generating devices, the temperature of air impinging on the downstream device is $[(N-1)/N]\Delta T$ above air-inlet temperature. Pressure drop is important because it puts a burden on air-moving devices. Let $(\Delta T, \Delta p)$ have values $(\Delta T_{SS}, \Delta p_{SS})$ and $(\Delta T_{BT}, \Delta p_{BT})$ for SS and BT airflow, respectively. Define a temperature-penalty factor f_T and a pressure-penalty factor f_P for BT airflow as $f_T \equiv \Delta T_{BT}/\Delta T_{SS}$ and $f_P \equiv \Delta p_{BT}/\Delta p_{SS}$, respectively. It may be shown,[3] using energy arguments and the proportionality of Δp to the square of velocity (for typical Reynolds numbers) [4], that $f_P f_T^2 = (A_{SS}/A_{BT})^2$. Typically $A_{SS}/A_{BT} \approx 2$, so $f_P f_T^2 \approx 4$. This product of BT penalties is significantly larger than 1. Thus, side-to-side (SS) airflow was chosen for BG/L, despite the extra space required by plenums.

Air-moving devices

To drive the side-to-side airflow in each BG/L rack, a six-by-ten array of 120-mm-diameter, axial-flow fans (ebm-papst** model DV4118/2NP) is positioned downstream of the cards, as shown in **Figure 6**. The fan array provides a pressure rise that is spatially distributed over the downstream wall of the rack, thereby promoting uniform flow through the entire array of horizontal cards. To minimize spatially nonuniform flow due to hub-and-blade periodicity, the intake plane of the fans is positioned about 60 mm downstream of the trailing edge of the cards. The fans are packaged as three-fan units. Cards on each side of a midplane are housed in a card cage that includes five such three-fan units. In the event of fan failure, each three-fan unit is separately replaceable, as illustrated by the partially withdrawn unit in Figure 6.

Each three-fan unit contains a microcontroller to communicate with the CFPGA control system on the service card (see the section below on the control system) in order to control and monitor each fan. Under external host-computer control, fan speeds may be set individually to optimize airflow, as described in the section below on refining thermal design. The microcontroller continuously monitors fan speeds and other status data, which is reported to the host. If host communication fails, all three fans automatically spin at full speed. If a single fan fails, the other two spin at the same full speed.

Complementary tapered plenums

BG/L racks are packaged in rows, as shown in **Figure 7**. For example, the BG/L installation at Lawrence Livermore National Laboratory has eight racks per row; the installation at the IBM Thomas J. Watson Research

[3] S. A. Hall, private communication.

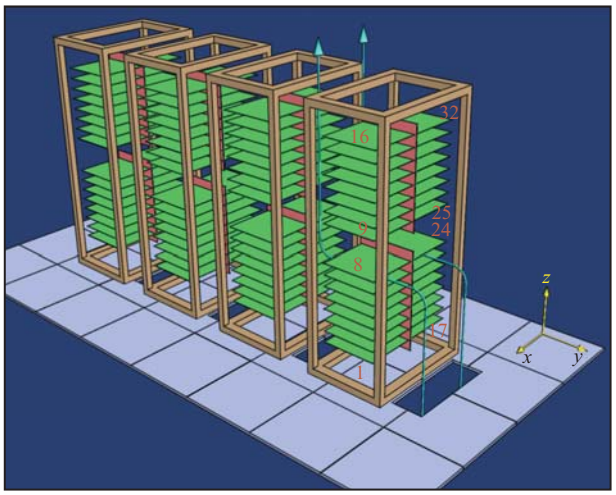


Figure 7

Conceptual row of Blue Gene/L racks.

Center has four racks per row. The racks in each row cannot abut one another because, for the SS airflow scheme, inlet and exhaust air must enter and exit through plenums that occupy the space above the raised-floor cutouts. To make the machine compact, these plenum spaces should be as small as possible.

Two alternative plenum schemes are shown in **Figure 8**. In Figure 8(a), hot and cold plenums are segregated, alternating down the row. Cold plenum B supplies cold air to its adjoining racks 1 and 2; hot plenum C collects hot air from its adjoining racks 2 and 3, and so on. This scheme uses plenum space inefficiently because, as a function of vertical coordinate z, the plenum cross-sectional area is not matched to the volumetric flow rate. As suggested by the thickness of the arrows in Figure 8(a), the bottom cross section of a cold plenum, such as B, carries the full complement of air that feeds racks 1 and 2, but because the air disperses into these racks, volumetric flow rate decreases as z increases. Thus, cold-plenum space is wasted near the top, where little air flows in an unnecessarily generous space. Conversely, in a hot plenum, such as C, volumetric flow rate increases with z as air is collected from racks 2 and 3, so hot-plenum space is wasted near the bottom. Clearly, the hot and cold plenums are complementary in their need for space.

Figure 8(b) shows an arrangement that exploits this observation; tapered hot and cold plenums, complementary in their shape, are integrated into the space between racks. A slanted wall separates the hot air from the cold such that cold plenums are wide at the bottom and hot plenums are wide at the top, where the respective flow rates are greatest. As will be shown, complementary tapered plenums achieve higher airflow

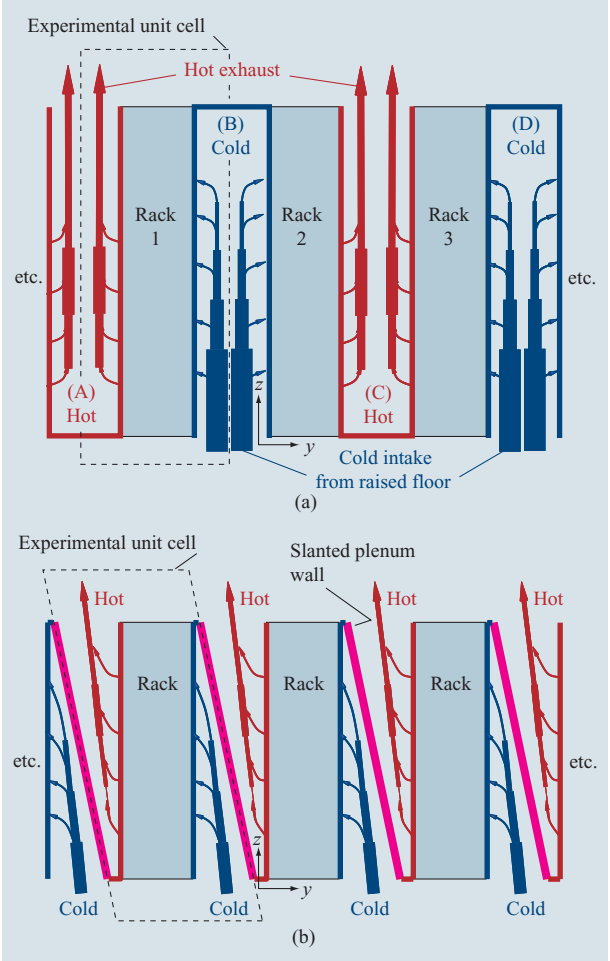


Figure 8

(a) Segregated, constant-width plenums. This is an inefficient use of space. (b) Complementary tapered plenums. This is an efficient use of space.

rate and better cooling of the electronics than segregated, constant-width plenums of the same size. Consequently, complementary tapered plenums are used in BG/L, as reflected in the unique, parallelogram shape of a BG/L rack row. This can be seen in the industrial design rendition (**Figure 9**).

To assess whether the slanted plenum wall of Figure 8(b) requires insulation, the wall may be conservatively modeled as a perfectly conducting flat plate, with the only thermal resistance assumed to be convective, in the boundary layers on each side of the wall. Standard analytical techniques [5], adapted to account for variable flow conditions along the wall, are applied to compute heat transfer through the boundary layer.[4] The analysis is done using both laminar- and turbulent-flow

[4]S. A. Hall, private communication.

Figure 9

Industrial design concept for Blue Gene/L. The slanted form reflects the complementary plenums in the unique side-to-side airflow.

assumptions under typical BG/L conditions. The result shows that, of the 25 kW being extracted from each rack, at most 1.1% (275 W) leaks across the slanted wall. Thus, the plenum walls do not require thermal insulation.

First-generation thermal mockup tests tapered plenums

To demonstrate and quantify the advantage of complementary tapered plenums, a full-rack thermal mockup with movable plenum walls was built [**Figure 10(a)**]. This early experiment reflects a node-card design, a rack layout, and an ASIC power level (9 W) quite different from those described subsequently with the second-generation thermal mockup (see below) that reflects the mature BG/L design. Nevertheless, the early mockup provided valuable lessons learned because it was capable of simulating the "unit cells" shown in both Figures 8(a) and 8(b), a feat made possible by movable walls whose positions were independently adjustable at both top and bottom. The unit cell in Figure 8(a) is not precisely simulated by the experiment because the vertical boundaries of this unit cell are free streamlines, whereas the experimental boundaries are solid walls, which impose drag and thereby artificially reduce flow rate. This effect is discussed below in connection with airflow rate and plenum wall drag.

The two walls have four positional degrees of freedom (two movable points per wall), but in the experiments, two constraints are imposed: First, the walls are parallel to each other; second, as indicated in Figure 10(a), the distance between the walls is 914.4 mm (36 in.), which is the desired rack-to-rack pitch along a BG/L row. Thus, only two positional degrees of freedom remain. These are parameterized herein by the wall angle θ and a parameter β that is defined as the fraction of the plenum space devoted to the hot side, $w_H/(w_H + w_C)$, where w_H and w_C are horizontal distances measured at the mid-height of the rack from node-card edge to hot and cold plenum walls, respectively. Fan space is included in w_H.

In the experiment, the only combinations of θ and β physically possible are those between the green and orange dashed lines in **Figure 10(b)**. Along the green line, the top

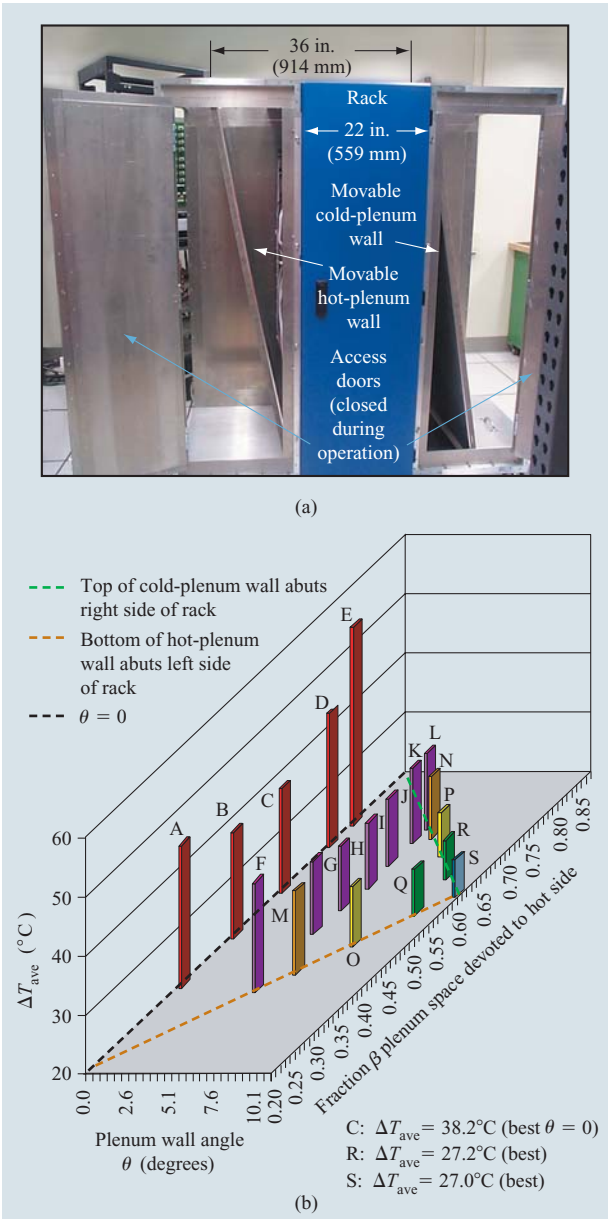


Figure 10

(a) First-generation thermal mockup with adjustable plenum walls. This was used for testing the concept of complementary tapered plenums. (b) Results from the first-generation thermal mockup. For various combinations of θ and β, the bars show the average rise (ΔT_{ave}) of mockup-ASIC case temperatures above the 20°C air-inlet temperature. There are 51 ASICs in the sample.

Table 3 Experimental results with first-generation full-scale thermal rack.

Case	*Type of plenum*	θ (degrees)	β	ΔT_{ave} (°C)	ΔT_{max} (°C)	*Standard deviation* (°C)
C	Constant width	0	0.495	38.2	47.9	3.93
R	Tapered	8.9	0.634	27.2	34.5	3.84
S	Tapered	10.5	0.591	27.0	36.7	4.34
∞	Infinitely wide	–	–	20.1	25.9	3.22

of the cold plenum wall abuts the upper right corner of the rack; along the orange line, the bottom of the hot plenum wall abuts the lower left corner of the rack. At apex S, both conditions are true, as shown in Figure 10(a), where the wall angle θ is maximized at 10.5 degrees.

The rack in Figure 10(a) contains an array of mockup circuit cards that simulate an early version of BG/L node cards based on early power estimates. ASICs and dc–dc power converters are respectively simulated by blocks of aluminum with embedded power resistors that generate 9 W and 23 W. DRAMs are simulated by 0.5-W surface-mount resistors. A sample of 51 simulated ASICs scattered through the rack are instrumented with thermocouples embedded in the aluminum blocks that measure ASIC case temperatures.

Each bar in Figure 10(b) represents an experimental case for which the mockup rack was run to thermal equilibrium. A bar height represents the average measured ASIC temperature rise ΔT above air-inlet temperature. With segregated, constant-width plenums [Figure 8(a)], only Cases A through E are possible (constant width being defined by $\theta = 0$). Of these, Case C is best; statistics are given in **Table 3**.

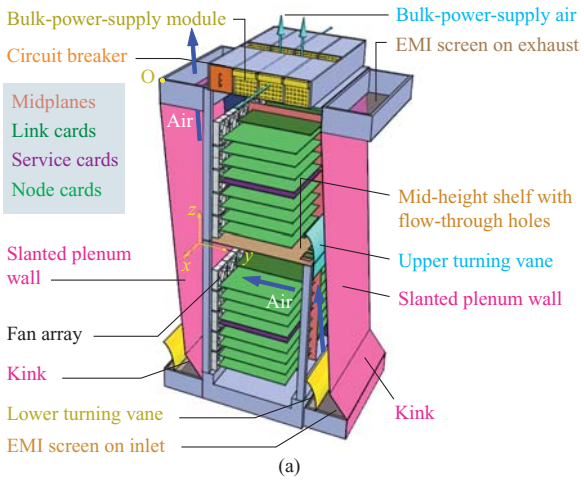
In contrast, with complementary tapered plenums [Figure 8(b)], all cases (A through S) are possible. Of these, the best choices are Case R (lowest ΔT_{ave}) and Case S (lowest ΔT_{max}). Choosing Case R as best overall, the BG/L complementary tapered plenums apparently reduce average ASIC temperature by 11°C and maximum ASIC temperature by 13.4°C, compared with the best possible result with constant-width plenums. This result is corrected in the section on airflow rate and plenum-wall drag below in order to account for drag artificially imposed by the plenum walls when the apparatus of Figure 10(a) simulates the unit cell of Figure 8(a).

Because slanting the plenum walls is so successful, it is natural to ask what further advantage might be obtained by curving the walls or by increasing the rack pitch beyond 914 mm (36 in.). Any such advantage may be bounded by measuring the limiting case in which the hot wall is removed entirely ($w_H = \infty$) and the cold wall is removed as far as possible ($w_C = 641$ mm $= 25.2$ in., well beyond where it imposes any airflow restriction), while still permitting a cold-air inlet from the raised floor. This limiting case represents the thermally ideal, but unrealistic, situation in which plenum space is unconstrained and the airflow path is straight rather than serpentine, leading to large airflow and low ASIC temperatures. The result, given in the last row of Table 3, shows that ΔT_{ave} for infinitely wide plenums is only about 7°C lower than for Case R. Curved plenum walls or increased rack pitch might gain some fraction of this 7°C, but cannot obtain more.

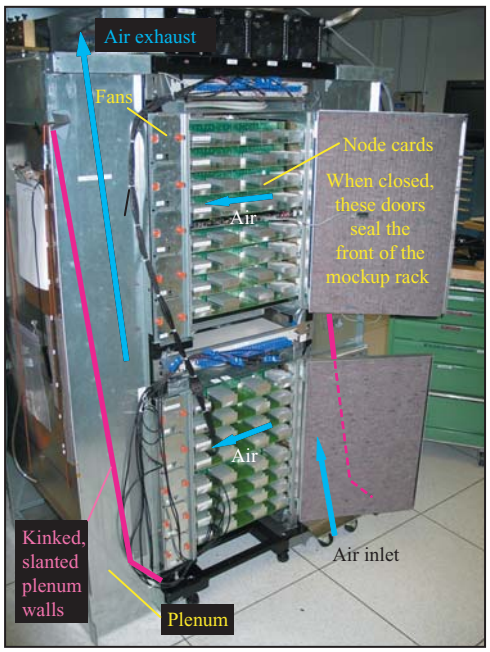
Second-generation thermal mockup reflects mature design

Figure 11(a) is a scaled but highly simplified front view of a BG/L rack (0.65 m wide) and two flanking plenums (each 0.26 m wide), yielding a 0.91-m (36-in.) rack pitch. It is precisely 1.5 times a standard raised floor tile, so two racks cover three tiles. The plenum shown is a variation of the tapered-plenum concept described above, in which the slanted wall is kinked rather than straight. The kink is necessary because a straight wall would either impede exhaust from the lowest row of fans or block some of the inlet area; the kink is a compromise between these two extremes. This dilemma did not occur in the first-generation thermal mockup, because the early BG/L design (on which that mockup was based) had the bulk power supply at the bottom of the rack, precluding low cards and fans. In contrast, the final BG/L design [Figure 11(a)] makes low space available for cards and fans (advantageous to shorten the long, under-floor data cables between racks) by placing the bulk power supply atop the rack, where it has a completely separate airflow path, as indicated by the arrows. Air enters the bulk-supply modules horizontally from both front and rear, flows horizontally toward the midplane, and exhausts upward. Other features, including turning vanes and electromagnetic interference (EMI) screens, are explained in the next section.

To quantify and improve the BG/L thermal performance, the second-generation thermal mockup, shown in **Figure 11(b)**, was built to reflect the mature mechanical design and card layout shown in Figure 11(a). The thermal rack is fully populated by 32 mockup node cards, called *thermal node cards*, numbered according to the scheme shown in Figure 7. Service cards are present



(a)



(b)

Figure 11

(a) Blue Gene/L rack and plenum layout. (b) Second-generation Blue Gene/L full-rack thermal mockup.

to control the fans. Link cards, whose thermal load is small, are absent.

A thermal node card, shown in **Figure 12**, is thermally and aerodynamically similar to a real node card (Figure 3). Each thermal node card contains 18 thermal compute cards on which BLC compute and I/O ASICs are simulated by blocks of aluminum with embedded resistors generating 15 W each (the maximum expected ASIC power). DRAM chips and dc–dc power converters are simulated as described for the first-generation mockup. Extruded aluminum heat sinks glued to the mock ASICs are identical to those glued to the real BG/L ASICs. The y dimension of the heat sink (32 mm) is limited by concerns over electromagnetic radiation; its x, z dimensions (20 mm × 29.2 mm) are limited by packaging. Fin pitch is 2.5 mm, fin height is 17 mm, and fin thickness tapers from 0.8 mm to 0.5 mm base to tip.

Temperatures are measured as described for the first-generation mockup, and so represent case temperatures, not junction temperatures. Experimentally, since only 128 thermocouple data channels are available, a mere fraction of the 1,152 ASICs and 128 power converters in the rack can be monitored. Power converters were found to be 10°C to 40°C cooler than nearby ASICs, so all data channels were devoted to ASICs. In particular, ASICs 0 through 7 in downstream Column D (Figure 14) were measured on each of 16 node cards. The 16 selected node cards are 1–8 and 25–32 in Figure 7, such that one node card from each vertical level of the rack is represented.

Column D was selected for measurement because its ASICs are hottest, being immersed in air that is preheated, via upstream Columns A–C, by an amount $\Delta T_{\text{preheat}}$. Thus, temperature statistics below are relevant to the hottest column only. Column A is cooler by $\Delta T_{\text{preheat}}$, which may be computed theoretically via energy balance at the rack level, where the total rack airflow, typically 3,000 CFM (1.42 m^3/s), has absorbed (before impinging on Column D) three quarters of the 25-kW rack heat load. Thus, $\Delta T_{\text{preheat}} = 11.3°\text{C}$, which agrees well with direct measurement of ASIC temperature differences between Columns A and D.

An ASIC case temperature T is characterized below by ΔT, the temperature rise above inlet air temperature T_{in}. ASIC power P_{asic} is fixed in the experiment at $(P_{\text{asic}})_0 \equiv 15\,\text{W}$; "node power" P_{node} (dissipation of ASIC plus associated DRAM) is fixed at $(P_{\text{node}})_0 \equiv 19.5\,\text{W}$. For arbitrary values of T_{in}, P_{asic}, and P_{node}, ASIC case temperature may be conservatively estimated as $T = \max(T_1, T_2)$, where

$$T_1 \equiv T_{\text{in}} + \left[\frac{P_{\text{node}}}{(P_{\text{node}})_0}\right]\Delta T; \qquad T_2 \equiv T_{\text{in}} + \left[\frac{P_{\text{asic}}}{(P_{\text{asic}})_0}\right]\Delta T. \quad (1)$$

Refining BG/L thermal design

Using the measurement plan described above, the second-generation thermal mockup [Figure 11(b)] was used to investigate numerous schemes for improving the BG/L thermal performance. **Table 4** summarizes the three most important improvements—turning vanes, low-loss EMI screens, and fan-speed profiling—by comparing four

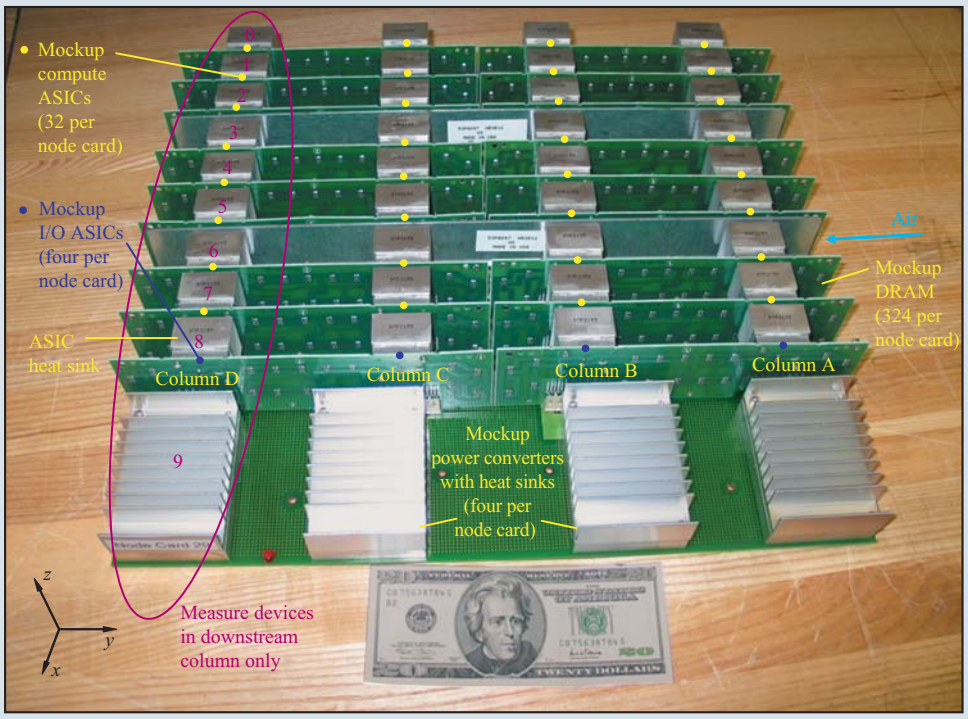


Figure 12

Mockup of Blue Gene/L thermal node card.

Table 4 Experimental results with second-generation full-scale thermal rack.

Case no.	*Turning vanes*	*EMI screen* (% open)	*Fan speeds*	ΔT_{ave} (°C)	ΔT_{max} (°C)	*Standard deviation of* ΔT (°C)
1	None	61 (high loss)	All max speed (6,000 RPM)	40.9	88.0	14.18
2	Optimized			37.9	54.2	6.81
3		95 (low loss)		31.9	48.9	6.87
4			Optimally profiled	32.0	41.8	4.22
∞	N/A; plenums removed	N/A; plenums removed	All max speed	22.0	29.2	3.24

experimental cases (Cases 1–4) in which these three improvements are progressively added.

For Cases 1–4, ΔT_{ave} and ΔT_{max} are plotted in **Figure 13(a)**. **Figures 13(b)–13(d)** show details behind the statistics; each is a paired comparison of two cases, where colored bars and white bars respectively represent individual ASIC temperatures before and after an improvement is made. Explanations of these comparisons, as well as an explanation of Case ∞ in Table 4, are given below.

Case 1 compared with Case 2 (turning vanes)

Figure 13(b) emphasizes the importance of the turning vanes shown as yellow and blue curved sheets in Figure 11(a). Without these vanes, the temperatures of the lower two or three cards in each card cage (e.g., Cards 1–3 and 25–26) are unacceptably high—as much as 34°C higher than when optimized vanes are installed. The reason is that the BG/L card stack [Figure 11(a)] is recessed about 85 mm from the upstream face of the rack, and the inlet air cannot negotiate the sharp corner to enter the low cards. Instead, separation occurs, and

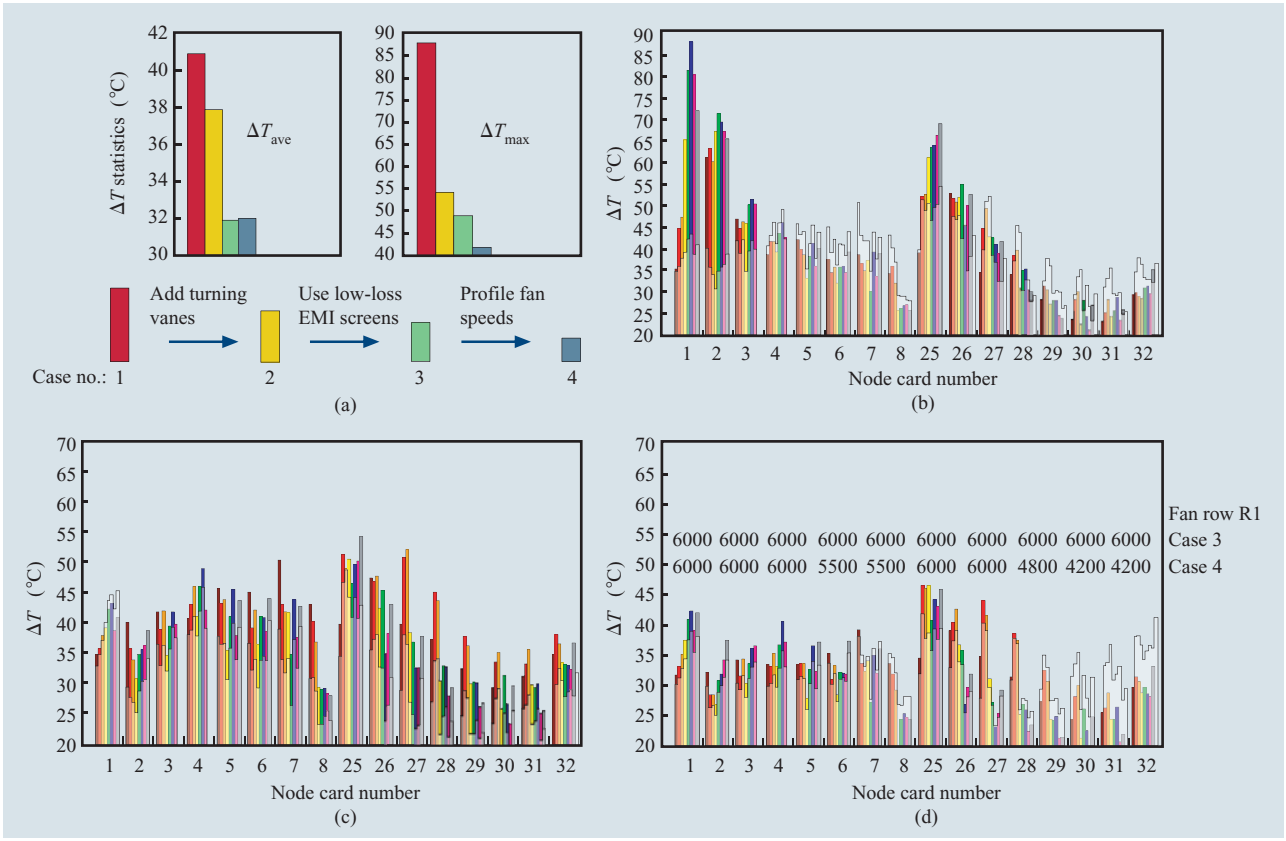


Figure 13

(a) Major thermal improvements obtained with second-generation full-rack thermal mockup. (b) Colored bars = Case 1. White bars = Case 2. (c) Colored bars = Case 2. White bars = Case 3. (d) Colored bars = Case 3. White bars = Case 4.

stagnation regions form upstream of these cards, starving them of air. The turning vanes prevent this by helping the air turn the corner. As shown in Figure 11(a), the upper and lower vanes are quite different. The lower vane rests entirely in the plenum space. In contrast, the upper vane rests entirely in the rack space, turning air that passes through holes in the mid-height shelf. [One hole is visible in Figure 11(a)]. ASIC temperatures in lower and upper card cages are sensitive to the geometry of the lower and upper vanes, respectively. In each case, the optimum vane shape turns just enough air to cool the low cards without compromising temperatures on higher cards. Tests with elliptical and easier-to-manufacture kinked shapes show that elliptical shapes provide better performance. Thus, in BG/L, both vanes are elliptical in cross section, but of different size. The upper vane is a full elliptical quadrant; the lower vane is less than a full quadrant.

Case 2 compared with Case 3 (EMI screens)

Figure 13(c) shows the importance of low-pressure-loss (i.e., high-percentage-open) EMI screens. As shown in Figure 11(a), air flowing through a BG/L rack traverses two EMI screens, one at the bottom of the cold plenum, the other at the top of the hot plenum. Using simple, 61%-open square-hole screens with typical BG/L operating conditions, the measured drop in static pressure across the pair of screens is $\Delta p_{\text{screens}} = 150$ Pa, in agreement with empirical results [4]. This is the largest pressure drop in the system—far more than that across the node cards (20 Pa), or that incurred traversing the hot plenum bottom to top (60 Pa). Consequently, improving the EMI screens dramatically reduces pressure loss and improves cooling: When the 61%-open screens are replaced by 95%-open honeycomb screens, $\Delta p_{\text{screens}}$ drops by a factor of 6, to 25 Pa. As shown in Figure 13(c), the corresponding drop in average ASIC temperature is 6°C.

Case 3 compared with Case 4 (fan-speed profiling)
Figure 13(d) shows the thermal benefit derived by profiling fan speeds; that is, running fans near cooler ASICs at less than the maximum speed, 6,000 revolutions per minute (RPM), to balance flow and reduce the temperature of hotter ASICs. In all experiments, the six fans in each of the ten horizontal rows (see Figure 6) were run at the same speed. Thus, the speed profile is described by ten speeds, one per fan row. For Cases 3 and 4 in Figure 13(d), these ten speeds are overlaid on the figure such that the speed of a fan row aligns roughly with the measured node cards that it cools. For example, the two rightmost numbers for Case 4 (4,200, 4,200) represent the topmost two rows of fans, which are directly downstream of node cards 30–32. The optimal fan-speed profile shown for Case 4 was determined by trial and error—speeds were reduced where Case 3 temperatures were low (e.g., Cards 30–32), and were left at maximum speed where Case 3 temperatures were high (e.g., Cards 25–27). The result is a 7°C reduction in maximum ASIC temperature.

Case ∞
Case ∞ in Table 4 is analogous to Case ∞ in Table 3: Plenums are removed from the thermal mockup to simulate the unrealistic, limiting case of infinitely wide plenums. In Table 4, comparing Case ∞ to Case 4 shows that the BG/L 0.26-m-wide plenums cause the ASICs to be 10°C hotter, on average, than they would be if the plenums were infinitely wide. Further optimization of plenum geometry, turning vanes, etc., might gain some fraction of this 10°C, but cannot obtain more.

Airflow rate and plenum-wall drag
Volumetric flow rate V of air through the thermal mockup rack is measured by scanning the air-exit plane (top left of Figure 12) with a velocity probe. The result for conditions as in Case 2 of Table 4 are shown in **Figure 14**, in which the (x, y) origin is point O in Figure 11(a). Integrating this velocity plume over area yields $V = 1.25\ \text{m}^3/\text{s} = 2{,}640$ CFM, which corresponds to an average air velocity of $\bar{v} = 6.66$ m/s. In practice, the BG/L rack true flow rate depends on room conditions—the characteristics of the computer-room air conditioners, their positions with respect to the rack, and how they are shared among racks. To permit airflow balancing at the room level in spite of such variations, an airflow damper is included in each BG/L rack at the top of the hot plenum.

Aerodynamic drag of the slanted plenum walls, artificially present when the apparatus of Figure 10(a) simulated the unit cell of Figure 8, may be estimated from Figure 14. Without this drag, the velocity deficit approaching the slanted wall at $y = 0$ in Figure 14 would not exist. Instead, the velocity plateau would continue to

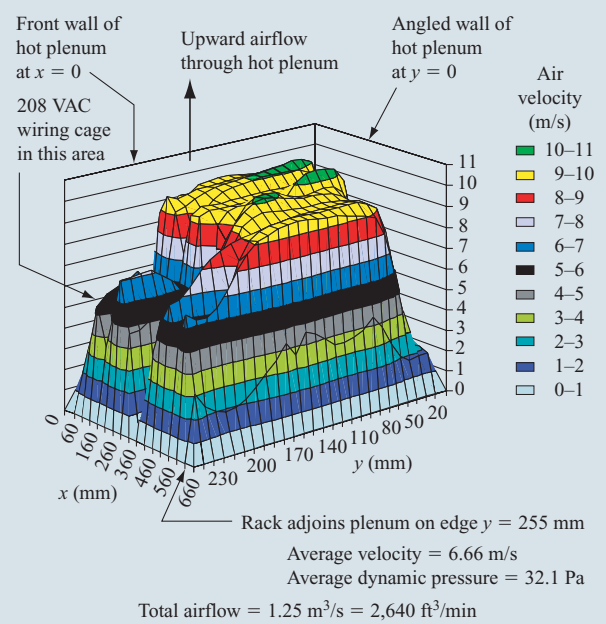


Figure 14

Velocity profile of air exiting the top of the hot plenum.

$y = 0$, increasing average velocity $\bar{v}$ by 13.5%. Therefore, Case C of Table 3 does not fairly represent the constant-width plenums of Figure 8(a), because $\bar{v}$ is experimentally 13.5% too low (assuming similarity between the two thermal mockups). Since ΔT for laminar flow over a heated plate, such as the ASIC heat-sink fins, is inversely proportional to $\sqrt{\bar{v}}$ [5], assuming all velocities scale with $\bar{v}$, the measured ΔT for Case C is 6.5% too high. Thus, $\Delta T_{\text{ave}} = 38.2$°C for Case C is 2.5°C too high, and the advantage of complementary tapered plenums over constant-width plenums, previously stated in the first-generation thermal mockup section as 11°C, is more accurately estimated as 8.5°C.

Signaling and clocking

Signaling overview
As described in [2], there are five interconnection networks in Blue Gene/L: a 3D torus, a global collective network, a global barrier and interrupt network, an I/O network that uses Gigabit Ethernet, and a service network formed of Fast Ethernet and JTAG. This section describes the design and performance of the signaling technology used to implement the torus (six full-duplex ports, each one bit wide), the global collective network (three full-duplex ports, each two bits wide), and the global interrupt network (four full-duplex ports, each one bit wide).

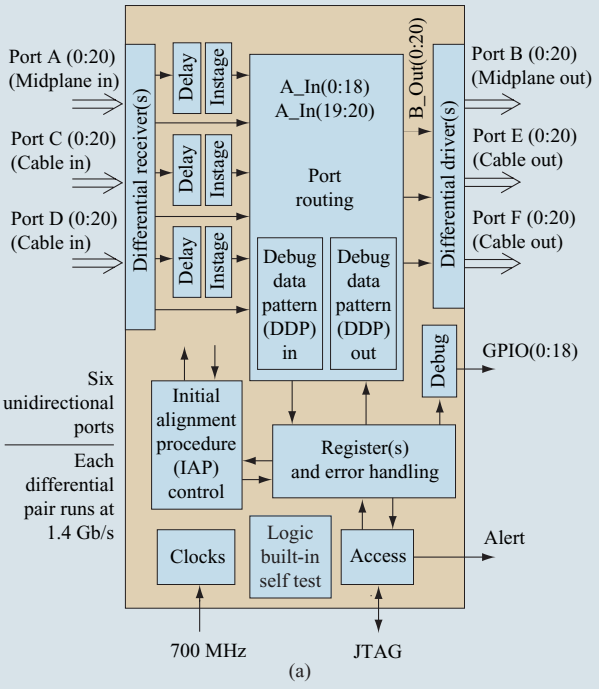


Figure 15

(a) Blue Gene/L link ASIC (BLL); (b) Blue Gene/L link card.

The main design goals of the signaling and interconnect circuits used in BG/L were stated in the Introduction. These are low power, low latency, and small silicon area to enable the required number of I/O cells to fit on the processor chip. Also, it was decided early in the system design cycle that the signaling data rate should be at least twice the processor frequency in order to provide sufficient network performance. Differential signaling for the torus and network links was chosen for a robust design with low error rates. For the torus network, there is a bidirectional serial link consisting of two differential pairs, one for each signal direction, connecting each pair of adjacent nodes. Since each BG/L node has six nearest neighbors, six torus drivers and six torus receivers are required on each ASIC. For the collective network, there are two differential pairs carrying signals in each signal direction per port. There are three collective ports on each BLC ASIC, resulting in another six drivers and six receivers, although not all are used in a typical system; some nodes use only two collective ports, and the terminal nodes use only a single port. On average, about two collective ports are used per node.

For network interconnections within each 512-node midplane, organized as an $8 \times 8 \times 8$ array of BG/L processors, 100-Ω differential printed circuit card wiring was used. Signals are driven and received directly on the processor ASIC. As described in the following section, these printed wiring interconnections connect processors on the same compute card, processors on different compute cards plugged into the same node card, or processors on different node cards on the same midplane. Thus, in addition to a circuit card wiring length of 50–700 mm, there can be two, four, or no card edge connectors in the signal interconnection path.

BLL cable redrive

When torus, global collective, and global interrupt interconnections span midplanes, either in the same rack or from rack to rack, the signals must be redriven to propagate through cables of up to 8.6 m in length, depending upon the system size and configuration. To implement this redrive, the BLL ASIC, the second unique chip designed for the system, was used. The BLL ASIC implements a switching or routing function that enables an array of racks to be partitioned into a variety of smaller configurations. Like the BLC chip [6], the BLL is fabricated by IBM Microelectronics using a 0.13-μm complementary metal oxide semiconductor (CMOS) process (CU-11). The BLL fits on a 6.6-mm $\times$ 6.6-mm die and contains more than 3M logic gates and about 6.7M transistors. It uses 312 signal I/O connections. Both the internal logic and all of the I/O are powered from a 1.5-V supply. The BLL chip is packaged in an IBM 25-mm $\times$ 32-mm ceramic ball grid array (CBGA) module with 474 total connections. The BLL is shown in **Figure 15(a)**.

The chip contains three send and three receive ports; signals received at each input port can be routed to any of the output ports. Both the link and compute ASICs use the same I/O circuits to drive and receive high-speed signals. However, on the link ASIC, these signals are grouped together in ports containing 21 differential pairs (17 data signals, a spare signal, a parity signal, and two asynchronous global interrupt signals). **Figure 15(b)** is a photograph of a link card with six BLLs; four of these link cards are used per midplane. The link card has 16 cable connectors, each attached to a BLL driving or

receiving port. The cable connectors are manufactured to IBM specifications by InterCon Systems, Inc. Each of these connectors attaches to a cable with 22 differential pairs using 26 American wire gauge (AWG) signal wire manufactured by the Spectra-Strip** division of Amphenol Corporation. The extra differential pair is used for a sense signal to prevent an unpowered chip from being driven by driver outputs from the other end of the cable.

Asynchronous signaling

The asynchronous global interrupt signals are driven and received over cables via the same I/O circuits used for the high-speed data links, but without the clocked data capture units. On each of the node and link printed circuit cards, the individual global interrupt channels are connected in a dot-OR configuration using normal single-ended CMOS drivers and receivers. This configuration is adequate to synchronize all processors in a 512-node midplane with a latency of less than 0.2 μs. A large (64K-node) system should have an interrupt latency of approximately 1 μs, limited by propagation delays.

Clock distribution

To aid in understanding the signaling and data-capture circuitry designed for BG/L, the system clock distribution is briefly described here. All BLC and BLL chip clocks are derived from a single high-frequency source that runs at the frequency of the processor core, i.e., 700 MHz. Differential clock buffers and clock interconnect are used to distribute this single source to the BLC and BLL ASICs. Individual node clocks are coherent and run at the identical frequency, although they are not necessarily equal in phase. A high-frequency source centralized in the array of racks is divided using a clock splitter and distributed to secondary clock-splitter cards via differential cables approximately 4.5 m long. These secondary cards, identical to the first except that the cable input replaces the clock source, in turn distribute the clock to tertiary clock splitters that in turn send one clock to each midplane. On the midplane, the service card distributes clocks to the 20 other cards in the midplane. Node and link cards, in turn, using the same clock splitter, provide clocks to all ASICs on the card. The maximum depth of the clock network is seven stages.

All clock paths to the ASICs have similar delay, having passed through the same number of cables, connectors, buffers, etc. With the use of low-voltage positive-emitter-coupled logic (LVPECL) clock driver chips based on bipolar technology, the delay through the clock buffer itself is nearly independent of voltage. This minimizes clock jitter due to voltage fluctuations at the different clock-splitter chips. One principal source of remaining jitter is due to temperature differences that occur slowly enough to be tracked by the data-capture scheme. The other main source of clock jitter is the result of noise at the receiver that affects the switching point, particularly since the clock rise time is degraded by line attenuation. This noise is reduced by common-mode rejection at the differential clock receivers. However, clock duty cycle symmetry is dependent upon careful length matching and symmetrical termination of the differential printed circuit card wiring. Clock symmetry is limited by the ability to match each differential pair using practical printed circuit wiring geometry, vias, and connector, module, and chip pins.

Measurements on a prototype of the BG/L clock distribution and at BLLs in a test setup have shown the root-mean-square (RMS) clock jitter to be between about 10 ps and 14 ps. Clock symmetry at the ASICs has been measured at about 45% to 55%, worst case. An experiment was performed in which excess jitter was added between the clock inputs of sending and receiving BLLs on separate cards when connected through an 8-m cable. For an added RMS jitter of 40 ps, the worst-case reduction in measured eye size at the receiving port data-capture unit was two inverter delays, or about 50 ps.

Synchronous data capture

A key part of BG/L signaling technology is its data-capture circuitry. Because of the large number of high-speed connections that make up the torus and collective networks of the system, a low-power signaling technology is required. As described above, the clock distribution system generates all clock signals from a single coherent source. However, there is an undetermined phase relationship between clock edge arrival times at the different ASICs that are at either end of each network link. Standard circuit solutions for asynchronous detection, such as PLL clock extraction, would require high power and introduce additional jitter. Source synchronous signaling (sending clock with data) requires an unacceptable increase in the number of cables and printed circuit wires, which would greatly increase the size and complexity of the system package. Therefore, to meet the unique needs of BG/L, a new low-power data-capture technique for use in all BG/L high-data-rate signaling was developed.

Figure 16 illustrates the logic of the BLL data-capture macro. The incoming received signal is passed through a chain of inverters which makes up a multi-tapped digital delay line. Two sets of latches, one clocked on the rising and one on the falling edge of the local clock, then sample the signal state at each inverter tap. The outputs from adjacent latches are compared via an exclusive-or (XOR) circuit to determine the position of the data transitions in the delay line. The aggregate of these comparisons forms

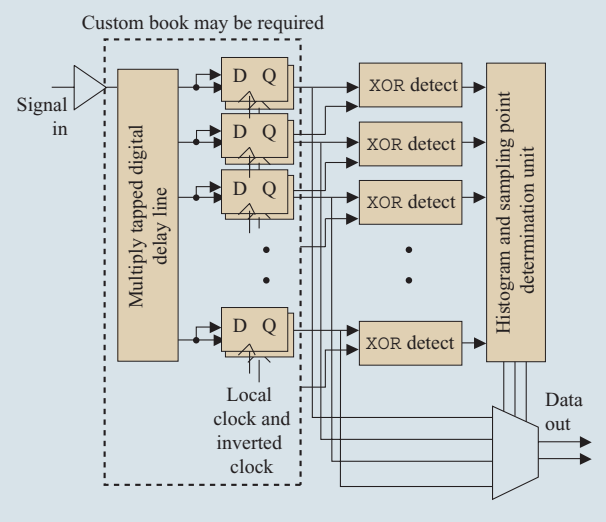


Figure 16

Blue Gene/L serial data-capture macro. (D = flip-flop data in, Q = flip-flop data out.)

a clocked string that is used to generate a history, which determines the optimal data sampling points. These optimal sampling points are found from the history string by looking for regions where the data never changes between adjacent delay taps; the control unit then sets the sampling point in the middle of this valid data region. The history is updated every *slow* clock period. This *slow* clock is generated by reducing the *fast* clock input frequency by a programmable divisor (selectable between 16 and 4,096).

With this data-capture method, the unknown phase offset between different ASIC clocks can be compensated by initializing the link with a data precursor, or training pattern, that provides a specific sequence with sufficient transitions to reliably determine the valid data window. This phase can then be tracked, and the capture point can be adjusted to compensate for long-term phase drift. The characteristic time response of this phase tracking is determined by the number of slow clock periods required for an update. There must be enough data transitions recorded between updates to the sampling point to ensure good eye determination. The update delay is programmable via a control register (selectable between 1 and 256). Setting this interval too low wastes power and might make the sampling point determination unreliable if too few data transitions are seen. Too large an update interval might not be able to track fast enough to compensate for transient thermal effects, power-supply variations, etc.

The circuitry contained within the dashed line in Figure 16 has been designed using a bit-stacking technique to force a compact physical layout that ensures the achievement of critical delay uniformity. The delay line must be long enough to contain an appreciable fraction of a data bit at the lowest speed of operation and shortest delay line. The design uses 32 inverters to achieve a delay between 0.99 ns and 1.92 ns, depending upon process tolerance, voltage, and temperature. To ensure that the data is optimally located in the fine-delay line, it is preceded by a second, coarse-delay line. The coarse-delay line, which consists of 16 stages, is used to center the received data within the fine-delay line of the data-capture unit. The coarse-delay line adds an additional delay time of one or two bits at the nominal data rate of 1.4 Gb/s. Minimal insertion delay is critical for good performance, since a dominant source of on-chip jitter is expected to be voltage-dependent delay variation as the V_{dd} and ground supply rails fluctuate because of noise.

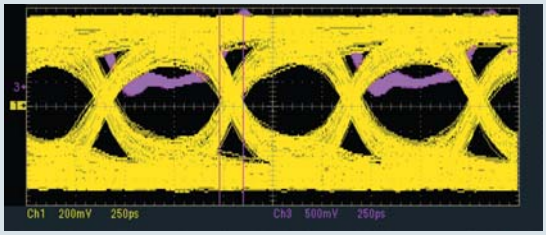
Differential data I/O measurement, timing, power, and BER

The differential driver used for the BG/L high-speed networks is a modified CMOS design with two output levels. If the previous data bit is the same as the current data bit, the higher drive impedance output is used. This single bit of pre-emphasis provides approximately 4.5 dB of compensation for high-frequency cable and interconnect attenuation. The differential receiver is a modified low-voltage differential signaling (LVDS) design with on-chip termination.

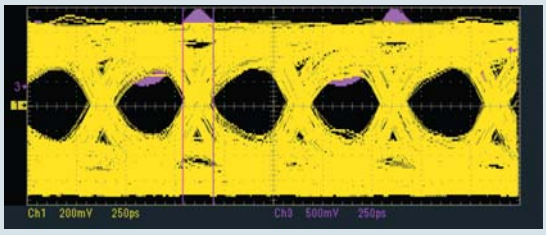
Figure 17 shows eye diagrams measured at the receiving chip input pins. The cable net includes about 10 cm of printed circuit wire from the driving chip to the cable connector on each end. The printed wiring connection includes traversal of two card edge connectors. The purple traces are the reference clock supplied by a pulse generator to the oscilloscope to provide a stable trigger; the same pulse generator is used as the system clock source for this experiment. **Figure 18** shows the measured eye parameters extracted from the data-capture unit for these same interconnection conditions; these are the minimum eye sizes, measured in units of inverter delay, for all data-capture units across the entire port. The lines shown are best fits to the measured data points. The reciprocal of their slopes gives an average delay of 48.2 ps to 50.4 ps for each inverter stage in the delay lines of the data-capture unit. This delay is indicative of a chip performance toward the slow end of the process distribution. The eye patterns show signal integrity and timing margins that are sufficient for reliable operation of 1-m card and 8-m cable paths up to about 2 Gb/s; this exceeds the performance requirements of the BG/L system.

A simplified jitter and timing budget for the BLLs is shown in **Table 5**. Since the data is captured as soon as it is latched, the synchronous logic after the latches does not affect the timing jitter. The principal sources of jitter and timing uncertainty are listed as follows:

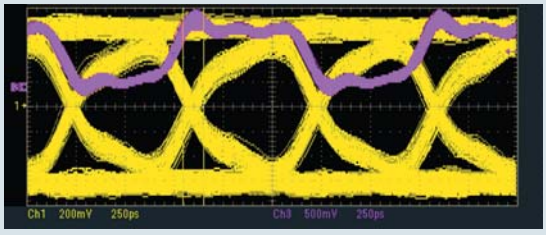
1. Intersymbol interference due to attenuation and dispersion on the interconnections.
2. Quantization, which is the effect of requiring a valid eye of at least three inverter taps for the purpose of tracking drift.



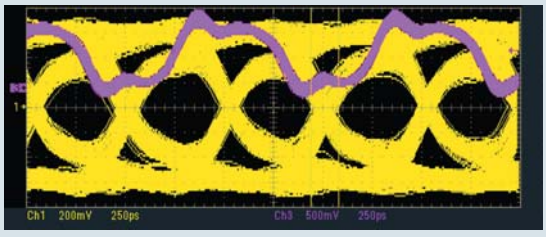
(a) 8-m cable at 1.6 Gb/s.



(b) 8-m cable at 2.0 Gb/s.



(c) 86-cm printed circuit wiring at 1.6 Gb/s.



(d) 86-cm printed circuit wiring at 2.0 Gb/s.

Figure 17

Measured eye diagrams for BG/L high-speed signaling over card and cable connections.

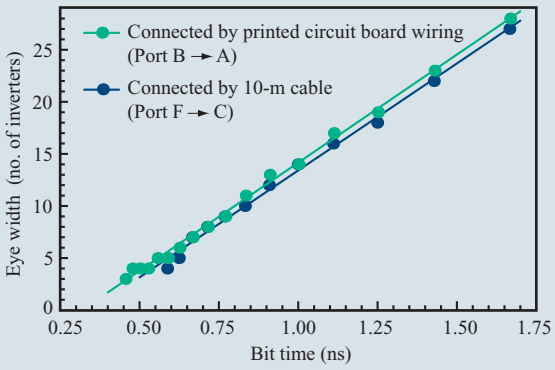


Figure 18

Minimum eye size compared with bit time (data-capture ports A and C).

3. Power-supply noise resulting in variation of delay-line value. The contribution that should be included here is due to the difference in delay between the clock and the data delay to the sampling latch. A significant (~80%) part of the delay, which is common to both paths, does not affect the jitter, since both the clock buffers and the data delay line are affected by the same amount. The delay from the data pin to the center of the inverter chain should be equal to the delay of the clock pin to the latch clock of the center of the inverter chain. The third term is from the effect of noise on a finite rise time signal.
4. Jitter on the local clock plus sending clock. This is jitter in the clock over approximately 50 ns, which is approximately the worst-case signal delay.
5. Asymmetries in the clock fan-out for the latches and the clock duty cycle.
6. Asymmetry in the inverter delay rise and fall.
7. Sending-clock asymmetry, which effectively results in a smaller bit time.
8. V_{dd}/V_{ss} noise on the sending side, which affects the jitter being introduced into the signal. This includes all logic drivers or muxes that follow the last latch. The delay line on the sending side (which is used to move the sampling delay line to the eye and to optimally choose the relationship between sending and receiving data) is a large source of error here.
9. Sending-side jitter caused by power-supply noise on the clock for the last latch before data is sent.

Estimates for each of the above contributions are detailed in Table 5. There are typically two values listed, for both best- and worst-case inverter delays. To ensure

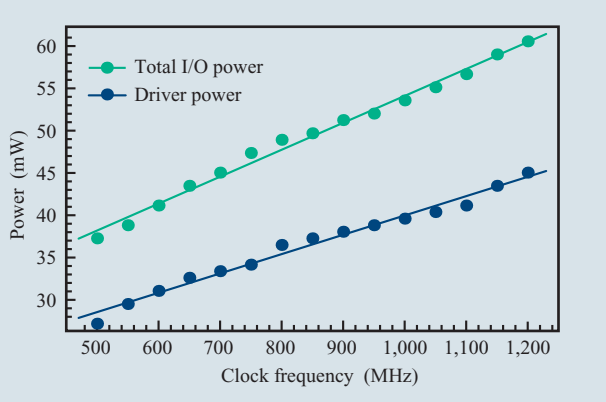


Figure 19

I/O power per bit for BG/L signaling circuits.

timing closure, we assume the worst case that all errors add linearly.

By running continuous data transfer over multiple links with various interconnection types on the BLL hardware, statistics on random data errors for both card and cable paths were accumulated. Only eight total errors were seen in over 4,900 hours of operation at data rates from 1.4 Gb/s to 2.0 Gb/s. The total bit error rate (BER) estimated from these experiments is less than 2×10^{-18}. The BER observed at data rates of 1.4 Gb/s to 1.7 Gb/s is less than 3×10^{-19}.

Low power was one of the main design goals for the signaling circuitry used in BG/L. The I/O power, which includes contributions from the output logic, driver, termination, receiver, and data-capture unit, was extracted from total current measurements as a function of clock frequency and circuit activity. These results are summarized in **Figure 19**. The power per bit varies approximately linearly from about 37 mW at 1 Gb/s to 54 mW at 2 Gb/s. Driver power as shown includes the contribution from the output logic and the net termination. The remaining power is consumed in the data-capture unit. We estimate about 45 mW total I/O power per link at our nominal data rate of 1.4 Gb/s. These results are in excellent agreement with predictions based on simulation of the I/O circuit designs.

The six bidirectional torus links, the differential driver, and the differential receiver require 0.42 mm^2 of chip area. The data-capture logic, decoupling capacitors, state

Table 5 Timing error terms for unidirectional data capture.

Term	*Worst and best case* (ps)	*Method of determination*
1. Intersymbol interference	160	Simulated unidirectional signaling
2. Quantization	180, 90	Three inverter delays
3. V_{dd}/V_{ss} noise	185, 100	Three contributions: • Worst-case delay from middle to end of delay line = 10% × ½ delay line • Common-mode voltage term for clock and delay line = 20% of total delay to middle of delay line × 10% • Slew rate of clock contribution = 10% or 100 ps
4. Local clock jitter	30	30 ps (measured) difference
5. Clock fan-out skew	30	(Assumed since bit-stacked layout)
6. Rise/fall delay line asymmetry	(0/2) (0/2)	Both ends of eye are reduced by the asymmetry of invert_bal_e, simulated 2 ps
7. Sending-side clock duty cycle asymmetry.	30	Estimated
8. Sending-side V_{dd}/V_{ss} noise on driver + delay line	5 + 35 for 1.4 Gb	Driver effect simulated. For delay line, a 5% supply shift results in ~5% of one bit time
9. Sending-side V_{dd}/V_{ss} noise on clock network	25	5% supply shift results in ~25 ps on a 500-ps clock network fan-out
10. Clk rcvr offset	10	30-mV offset, 1 V/300-ps edge rate
11. Data rcvr offset	15	200 mV/100 ps
Total (linear sum)	(705/530)	

machines, and control logic add another 0.63 mm^2, for a total of slightly more than 1.0 mm^2.

Circuit cards and interconnect

Design philosophy

The circuit cards and interconnect were designed to physically implement the major system communication buses in as efficient a manner as possible. Rather than accept pin placement of the compute and link ASICs and wire them together, we chose to work backward from the high-level package to the chip pin placement.

First, the bus widths and basic compute card form factor were determined in a process that traded off space, function, and cost. Having the fewest possible compute ASICs per compute card would give the design the lowest-cost field-replaceable unit, decreasing the cost of test, rework, and replacement. However, having two compute ASICs per compute card instead of one gave a more space-efficient design due in part to the need to have an odd number of DDR SDRAM chips associated with each compute ASIC.

With the two-ASIC compute card determined, the bus width of the torus network was decided primarily by the number of cables that could be physically connected to a half-rack-sized midplane. Since most midplane-to-midplane connections are torus connections, the torus network was the primary determining factor in the number of cables escaping each midplane and each rack. With the torus bus width set at one bit bidirectional between each nearest-neighbor ASIC in a 3D logical grid, the collective network and interrupt bus widths and topology were determined by card form factor considerations, primarily the compute card. With the external DRAM memory bus and torus bus widths selected, the number of pins per compute ASIC was then determined by the choice of collective network and interrupt bus widths plus the number of ports escaping each ASIC. Choosing the number of collective ports per ASIC and the number of collective ports between the various card connectors was a tradeoff between global collective network latency and system form factor.

The final choice of three collective ports per ASIC, two bidirectional bits per collective port, and four bidirectional global interrupt bits per interrupt bus was set by the need to fit the compute ASIC into a 32-mm $\times$ 25-mm package. Having the smaller chip dimension limited to 25 mm decreased the pitch between node cards sufficiently that 1,024 compute ASICs could be fitted into a single rack. The compute card form factor, ASIC package size, and widths of the various buses were thus determined to yield the maximal density of compute ASICs per rack.

With the compute card form factor and basic system arrangement determined, the circuit card connectors, card cross sections, and card wiring were defined next. These were chosen with a view to compactness, low cost, and electrical signaling quality. With all of the high-speed buses being differential, a differential card-to-card connector was selected. With signaling speeds defined to be under 4 Gb/s (final signaling rate 1.4 Gb/s), a press fit connector was determined to be electrically sufficient. A connector with two differential signal pairs per column of pins was selected because this allowed the signal buses to spread out horizontally across nearly the entire width of each card-to-card connection. This resulted in fewer card signal layers required to escape from under the connectors, and it allowed signal buses to cross fewer times within the cards, also reducing card signal layer counts. These considerations narrowed the connector choice down to only a few of the contenders. The final choice of the six-row Metral** 4000 connector was made primarily by the convenient pin placement, which maximized routing of wide differential pairs underneath the connector. Once again, form factor was key.

Fundamental to the circuit card cross sections was a requirement of high electrical signaling quality. All cards had two power layers in the center of the card cross section sandwiched inside two ground layers. The rest of the card was built outward by adding alternating signal and ground layers on both sides of the central power core. This resulted in signals that were always ground-referenced. This design practice, which decreases the return path discontinuities and noise coupling that might result from such discontinuities, has been tried with success in other recent machines [7].

The largest card, the midplane, ended up with 18 layers, but all other cards were confined to 14 total layers. One card, the node card, required additional power layers to distribute the 1.5-V core voltage from the dc–dc power converters to the array of compute cards with an acceptably small dc voltage drop in the card. Fortunately, the wiring on this card was regular enough to permit having one less signal layer, and a central signal and ground pair of planes were each reallocated to be additional 1.5-V power planes. Card linewidths for the 100-Ω line-to-line differential impedance were chosen to provide some layers with wide (190-μm to 215-μm or 7.5-mil to 8.5-mil) 1.0-ounce copper traces for low resistive loss for those high-speed nets that had to travel long distances.

Card thickness was minimized by having short connections on other layers be narrow (100-μm- or 4-mil-wide) 0.5-ounce nets, the narrowest low-cost linewidth. Card dielectrics were low-cost FR4. Once the final signaling speed was determined to be 1.4 Gb/s, there was no electrical need for a lower-loss dielectric. At these

Table 6 Torus dimensionality of BG/L circuit cards.

Card	*x torus length*	*y torus length*	*z torus length*	*ASICs per card*
Compute	1	2	1	2
Node	4	4	2	32
Midplane	8	8	8	512

signaling speeds, copper cross section mattered more than dielectric loss. Card sizes were determined by a combination of manufacturability and system form factor considerations. The node cards were designed to be near the maximum card size obtainable from the industry-standard low-cost 0.46-m × 0.61-m (18-in. × 24-in.) raw card blank. The larger midplane was confined to stay within the largest panel size that could still be manufactured affordably by multiple card vendors. Card thickness was similarly targeted to remain less than the 2.54-mm (100-mil) "cost knee" whenever possible. This was achieved on all cards but the midplane.

Card wiring was defined to minimize card layers and thus minimize card cost. The most regular and numerous high-speed connections—the torus—were routed first. The regular x, y, z logically 3D grid of torus interconnections was routed in the cards first, and the connector signal pin positions were chosen to minimize torus net signal crossing. The global collective network and global interrupt bus were laid out next, with the exact logical connection structure of these two networks being determined in large part by the way in which they could be most economically routed on the minimal number of card layers. Electrical quality was maintained by carefully length-matching all differential pairs and by enforcing large keep-out spaces between lines for low signal coupling. With the major high-speed buses routed for all cards, the compute and link ASIC pinouts were then chosen to make the ASIC package pin positions match the physical signaling order of the buses in the cards. This again minimized signal crossing and enabled fewer vias, better electrical performance, fewer wire crossings, less route congestion, fewer card layers, and lower card cost.

The layout of the 16-byte-wide DDR SDRAM memory bus was also performed prior to choosing the compute ASIC pin placement, again to optimize the package escape on the card and minimize routing layers. These single-ended traces were designed for 55-Ω impedance. This choice allowed differential and single-ended lines of the same line width on the same layer and increased the wiring utilization.

It was difficult to enforce exact pin placement on the ASIC because the locality of the major functional units and the pin placements had other constraints [8]. In many cases, compromises could be made to move pin placement close to what was desired, but one wiring layer more than that required for a simple fan-out was required in our ground-referenced ceramic first-level package. This was considered a reasonable cost tradeoff for the greatly simplified circuit card designs.

With the high-speed buses routed, high-speed clocks were then added, again with minimal crossing and minimal vias. Finally, all low-speed nets were routed in a space-efficient manner with multiple via transitions between different routing layers in order to use the remaining available wiring space. All cards were hand-routed.

Network implementation

The physical implementation of three of the major BG/L communications buses (torus, global collective network, and global interrupt) is described below. The fourth network, I/O, shares subsets of the global collective network wires to permit groups of compute ASICs to send I/O traffic to their designated I/O ASIC. This is described below in the section on the global collective network. Each I/O ASIC has a direct electrical Gigabit Ethernet link to an RJ45 jack on the node card faceplate (tail stock), permitting the I/O ASIC to access an external disk storage network. The Ethernet and JTAG control network is the fifth BG/L network and is described in the section on control.

Torus network

The torus network connects each compute ASIC to its six nearest neighbors in a logically 3D (x, y, z) coordinate space. Each torus link between ASICs consists of two unidirectional differential pairs which together form a bidirectional serial link. The data rate for each torus link is two bits per processor cycle, send and receive, in each of the six directions. Two compute ASICs are assembled on each compute card, 16 compute cards are plugged into each node card to give 32 compute ASICs per node card, and 16 node cards are plugged into each midplane to give 512 compute ASICs per midplane. The logical x, y, z size of the 3D torus (or mesh) on each card is listed in **Table 6**.

For the smaller compute and node cards, the torus connections were hardwired as a 3D mesh, with each card itself being unable to complete a looping torus without its neighboring cards being present. The smallest unit in which the torus network can wrap is the midplane. For example, given eight ASICs in series in the x dimension on a midplane, there is a software-configurable link that can wrap ASIC number 8 back to ASIC number 1, completing a torus loop. This configurable connection is made in the BLLs on the link cards and permits software-controlled partitioning of the system, as described below in the section

on partitioning design. A torus larger than 512 compute ASICs can be created by switching the link ASICs so that they connect each midplane via cables to each of its six logically nearest neighbors in 3D space. A total of 128 differential pairs, or 64 bidirectional serial links, connect each of the six faces ($x+$, $y+$, $z+$, $x-$, $y-$, $z-$) of an $8 \times 8 \times 8$ cubic midplane to its logical neighbor. Each of these nearest-neighbor midplane connections requires eight electrical cables, for a total of 48 cables connected to a midplane. As described below in the section on split partitioning, up to 16 additional cable connections per midplane may be used to increase the number of ways that midplanes can be connected into larger systems, for a maximum of 64 cables connected to each midplane.

For each 22-differential-pair cable, 16 pairs are allocated to the torus signals. The largest BG/L system currently planned is a 128-midplane (64-rack) system, with eight midplanes cabled in series in the x dimension, four midplanes in the y dimension, and four midplanes in the z dimension. This results in a single system having a maximal torus size of 64 (x) by 32 (y) by 32 (z) compute ASICs, with 65,536 ASICs total. Since the z dimension cannot be extended beyond 32 without increased cable length, the largest reasonably symmetric system is 256 midplanes, as 64 (x) by 64 (y) by 32 (z).

Global collective network

Single midplane structure

The global collective network connects all compute ASICs in a system. This permits operations such as global minimum, maximum, and sum to be calculated as data is passed up to the apex of the collective network, then broadcasts the result back down to all compute ASICs. Each ASIC-to-ASIC collective network port is implemented as a two-bit-wide bidirectional bus made from a total of four unidirectional differential pairs, giving a net ASIC-to-ASIC collective network bandwidth of four bits per processor cycle in each direction. Significant effort was made to minimize latency by designing a collective network structure with the minimum number of ASIC-to-ASIC hops from the top to the bottom of the network. It will be shown that there are a maximum of 30 hops required to traverse the entire network, one way, in the 64-rack, 65,536-compute-ASIC system. Since the structure of the collective network is much less regular than that of the torus, it is diagrammed in **Figure 20** and explained below.

The collective network implementation on the 32-compute-ASIC node card is shown in Figure 20(a), where each circle represents a compute card with two compute ASICs. Each ASIC has three collective network ports, one of which connects to the other ASIC on the same compute card, while the other two leave the compute card through the Metral 4000 connector for connection to ASICs on other cards. The narrow black connections represent the minimum-latency collective network connections. There are eight ASIC transitions for information to be broadcast from the top of the node card to all nodes on a card. Additional longer-latency collective network links, shown in red, were added in order to provide redundancy. Global collective network connectivity can be maintained to all working ASICs in a BG/L system when a single ASIC fails—no matter where in the system that ASIC is located. Redundant reordering of the collective network requires a system reboot, but no hardware service. As shown at the top of the figure, there are four low-latency ports and one redundant collective network port leaving the node card and passing into the midplane.

In Figure 20(a), each half circle represents one half of an I/O card. There may be zero, one, or two I/O cards present on any given node card, for a total of up to 128 Gb links per rack, depending on the desired bandwidth to disk for a given system. Compute cards send I/O packets to their designated I/O ASIC using the same wires as the global collective network. The collective network and I/O packets are kept logically separate and nonblocking by the use of class bits in each packet header.[5]

The collective network is extended onto the 512-compute-ASIC midplane as shown in Figure 20(b). Here, each oval represents a node card with 32 compute ASICs. Up to four low-latency ports and one redundant collective network port leave each node card and are wired to other node cards on the same midplane. The low-latency collective network links are shown as black lines, and the redundant links as red lines. The local head of the collective network on a midplane can be defined by software as a compute ASIC on any of the three topmost node cards. In a large system, minimum latency on the collective network is achieved by selecting node card J210 as the local midplane apex. In this case, there are seven ASIC transitions to go from this node card to all other node cards. When the eight-ASIC collective network latency on the node card is added, the result is a maximum of 15 ASIC-to-ASIC hops required to broadcast data to all nodes on a midplane or collect data from all nodes on a midplane.

Multiple midplane structure

Between midplanes, the global collective network is wired on the same cables as the torus. One extra differential pair is allocated on each cable to carry collective network signals between midplanes. Thus, there are six

[5]D. Hoenicke, M. A. Blumrich, D. Chen, A. Gara, M. E. Giampapa, P. Heidelberger, V. Srinivasan, B. D. Steinmacher-Burow, T. Takken, R. B. Tremaine, A. R. Umamaheshwaran, P. Vranas, and T. J. C. Ward, "Blue Gene/L Collective Network," private communication.

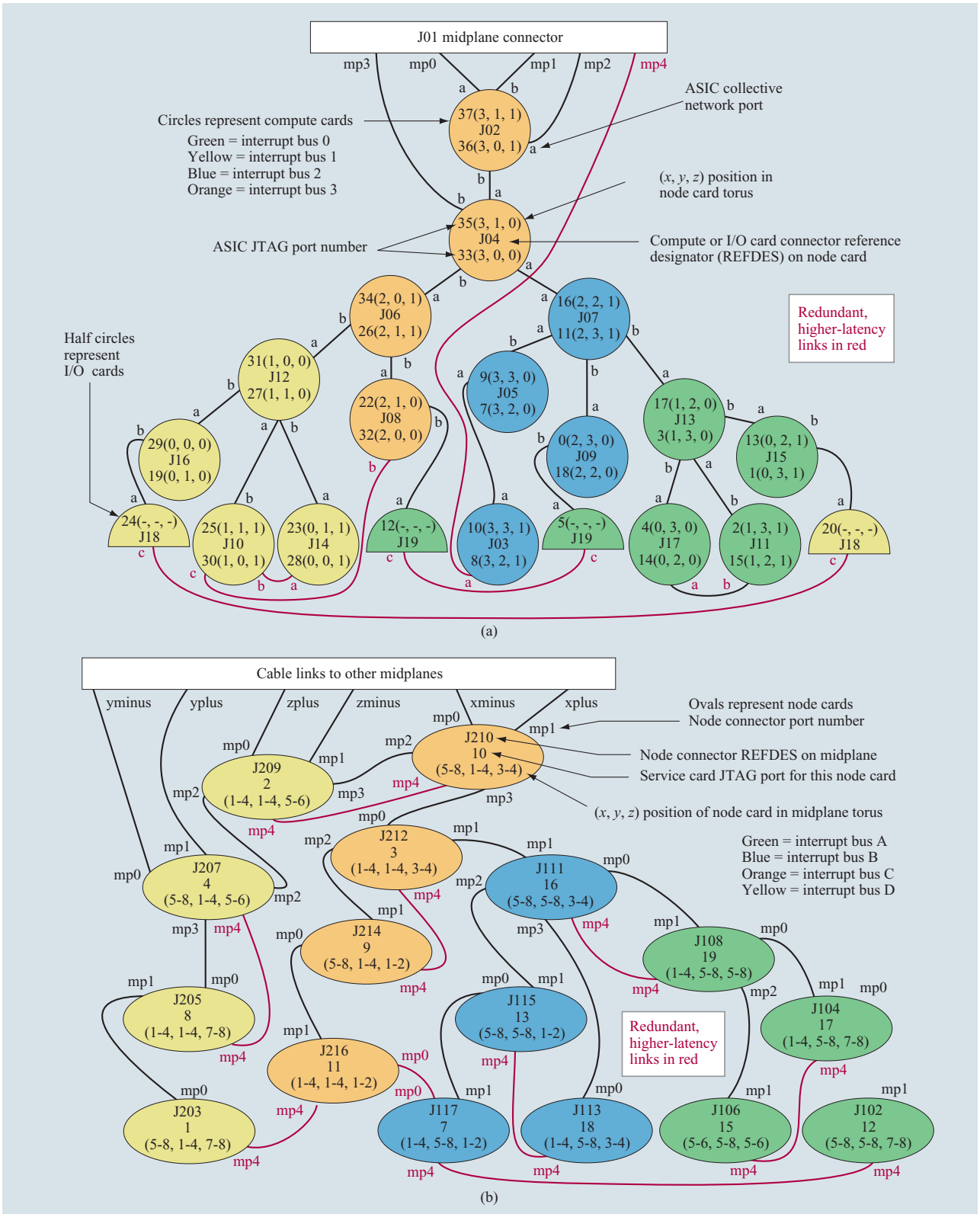


Figure 20

Blue Gene/L global combining network (a) on the node card; (b) on the midplane.

bidirectional collective network connections from each midplane to its nearest neighbors, labeled *yminus, yplus, xminus, xplus,* and *zminus, zplus* according to the torus cables on which the inter-midplane collective network connections are carried [Figure 20(b)]. Control software sets registers in the compute ASIC collective network ports to determine which of these six inter-midplane collective network connections is the upward connection and which, if any, of the remaining connections are logically downward connections. For a 64-rack, 65,536-compute-ASIC BG/L system, a maximum of 15 ASIC-to-ASIC collective network hops are required to traverse the global collective network, one way, from the head of the collective network to the farthest midplane. Add this to the 15 intra-midplane hops explained earlier, and there are a total of 30 ASIC-to-ASIC hops required to make a one-way traversal of the entire collective network.

Global interrupt bus

Single midplane structure

The global interrupt bus is a four-bit bidirectional (eight wires total) bus that connects all compute and I/O ASICs using a branching fan-out topology. Unlike the global collective network, which uses point-to-point synchronous signaling between ASICs, the global interrupt bus connects up to 11 devices on the same wire in a "wire-ORed" logic arrangement. As explained above, asynchronous signaling was used for high-speed propagation. The global interrupt bus permits any ASIC in the system to send up to four bits on the interrupt bus. These interrupts are propagated to the systemwide head of the interrupt network on the OR direction of the interrupt bus and are then broadcast back down the bus in the RETURN direction. In this way, all compute and I/O ASICs in the system see the interrupts. The structure of the global interrupt bus is shown in **Figure 21**.

Figure 21(a) shows the implementation of the global interrupt bus on the 32-compute-ASIC-node card. Each oval represents a compute or an I/O card containing two ASICs. Each line represents four OR and four RETURN nets of the interrupt bus. Either eight ASICs on four cards, or ten ASICs on five cards, are connected by the same wire to a pin on the node card CFPGA. For the OR nets, a resistor near the CFPGA pulls the net high and the compute and I/O ASICs either tristate (float) their outputs or drive low. Any one of the eight or ten ASICs on the net can signal an interrupt by pulling the net low. In this way a wire-OR function is implemented on the node card OR nets. The node card CFPGA creates a logical OR of its four on-node card input OR pins and one off-node card down-direction input OR pin from the midplane, and then drives the result (float or drive low) out the corresponding bits of its off-node card up-direction OR nets. Conversely, the node card CFPGA receives the four bits off the midplane from its one up-direction RETURN interrupt nets and redrives copies of these bits (drive high or drive low) onto its four on-node card and one off-node card down-direction, or RETURN, interrupt nets. In this way, a broadcast function is implemented. No resistors are required on the RETURN nets.

Figure 21(b) shows the details of the global interrupt bus on the 512-compute-ASIC midplane. Each shaded rectangle represents a node card with 32 compute and up to four I/O ASICs. The interrupt wiring on the midplane connects CFPGAs in different node and link cards. The same wire-OR logic is used on the OR or up-direction nets as was used on the node cards between the CFPGAs and the ASICs. RETURN or down-direction signaling is broadcast, as described earlier for the node card.

The midplane carries five different bidirectional interrupt buses. The four lower interrupt buses (A–D) connect quadrants of node cards. Here the down-direction port of each quadrant-head CFPGA connects to the up-direction ports of the CFPGAs on the three downstream node cards. All eight bits of each bus are routed together, as in the node card. The fifth, or upper interrupt bus, shown in Figure 21(b), connects the up-direction ports of the quadrant-head CFPGAs to the down-direction ports of the link card CFPGAs. The bits of this upper interrupt bus are not routed together. Rather, the CFPGA on each of four link cards handles one bit (one OR and one RETURN) of the four-bit bidirectional midplane upper interrupt bus. These four link card CFPGAs are collectively the head of the global interrupt network on the midplane. If the BG/L system is partitioned as a single midplane, the link card CFPGAs receive the OR interrupt signals from all ASICs and CFPGAs logically below them on the midplane, and then broadcast these interrupt bits back downward on the RETURN or down-direction interrupt nets.

Multiple midplane structure

If multiple midplanes are configured in a single system, the global interrupt bus is connected from midplane to midplane using the same topology as the global collective network. Global interrupt signals are wired on the same cables as the torus, with one extra differential pair per cable allocated to carry global interrupt signals between midplanes. There are six bidirectional global interrupt connections from each midplane to its nearest neighbors, labeled yminus, yplus, xminus, xplus, and zminus, zplus according to the torus cables on which the inter-midplane collective network connections are carried. Control software sets registers in the link card CFPGA to determine which of these six inter-midplane collective

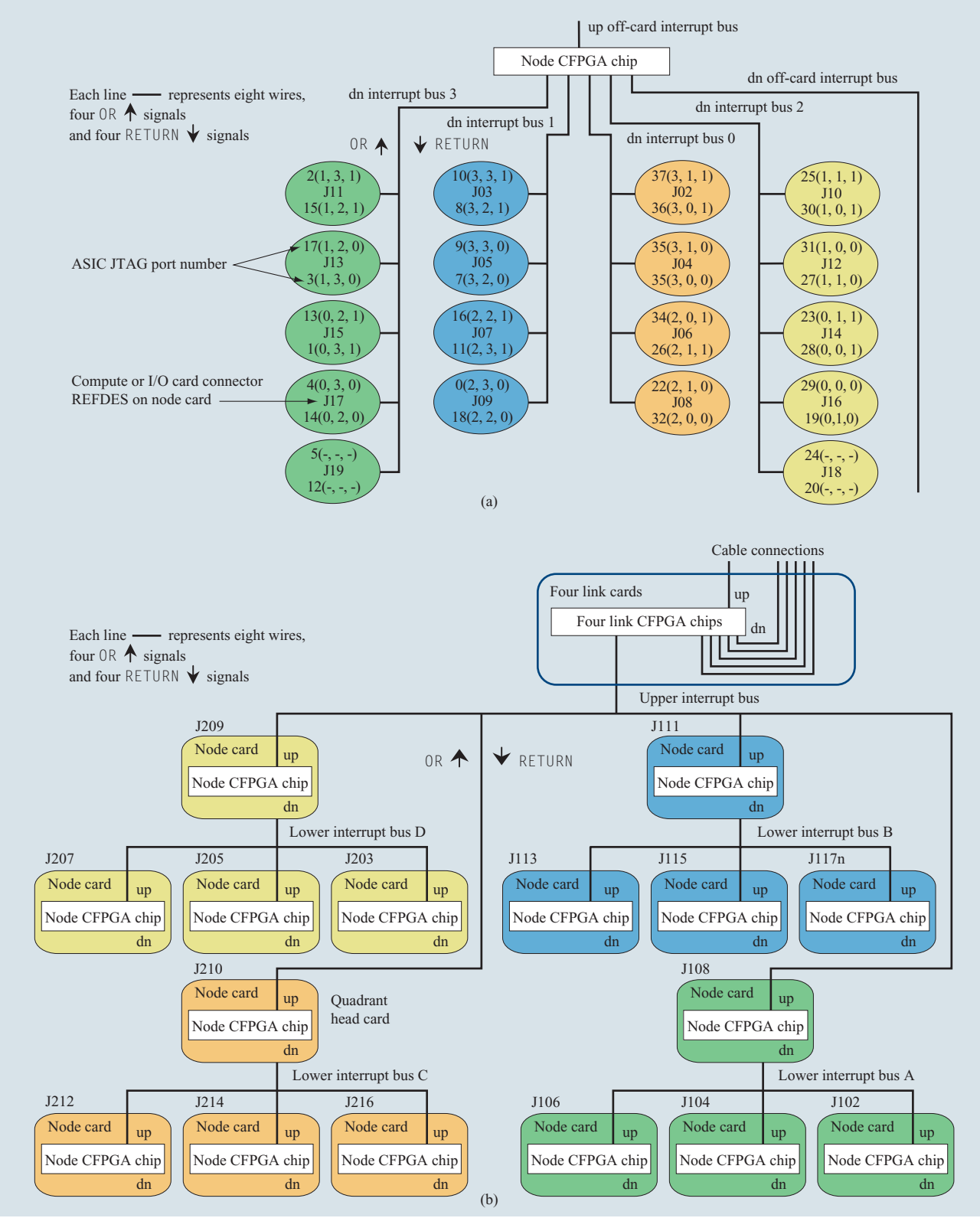


Figure 21

Blue Gene/L global interrupts (a) on the node card; (b) on the midplane. (CFPGA = control–FPGA.)

network connections is the upward interrupt connection and which, if any, of the remaining cable connections are logically downward interrupt connections.

Partitioning design

The smallest torus in the BG/L system is the 512-compute-ASIC midplane of logical dimension $8 \times 8 \times 8$ (x, y, z). BLLs and data cables permit the torus, global collective, and global interrupt networks to be extended over multiple midplanes to form larger BG/L systems. The BLLs also act as crossbar switches, permitting a single physical system to be software-partitioned into multiple smaller logical systems. Partitioning permits more efficient use of the hardware and allows system software to swap in redundant hardware when a node, compute, or I/O card failure occurs.

Regular partitioning

Once a given BG/L multirack system is cabled, the system partitions and midplane-to-midplane data connections are set by configuring the BLLs. There are four link cards per midplane and six BLLs per link card. On each link card, two BLLs switch the x torus signals, two BLLs switch y, and two chips switch z, with the collective and interrupt networks carried as sideband signals on the same cables. This allows all networks to be partitioned, or repartitioned, by simple configuration of the BLL. **Figure 22(a)** shows several of the configuration modes. When in Mode 1, the BLL includes the ASICs of the local midplane in the larger multi-midplane torus. Torus data traffic passes in from cable port C, loops through the local midplane torus via ports B and A, and continues out cable port F to the next midplane. Mode 2 isolates the local midplane torus, completing the local torus in a one-midplane loop while continuing to pass cable data so as to preserve the larger multi-midplane torus of which other midplanes are a part. This switching function from Mode 1 to Mode 2 provides the regular partitioning function of BG/L and permits any one midplane to be made a separate local torus while preserving the torus loop of the remaining midplanes. The y and z torus dimensions have only this regular partitioning.

Split partitioning

Regular partitioning alone does not permit larger multirack systems to be evenly divided into smaller logical systems that are one half or one quarter the size of the largest system. Split partitioning was therefore added in the x torus dimension to provide greater system-partitioning flexibility. Comparing the x and y torus logical diagrams in **Figure 22(b)**, it can be seen that the y torus contains only cable links that connect the midplanes in a single torus loop. By contrast, the x torus has additional "split-partitioning" connections, shown in green in the x torus logical diagram. These split-partitioning cables are plugged into ports D and E of the BLL diagram. The BLL can be programmed in Modes 3 through 9 to use these split-partitioning cables. For example, Modes 3 and 4 of the BLL connect split-redirection port D to the torus network, instead of regular redirection port C.

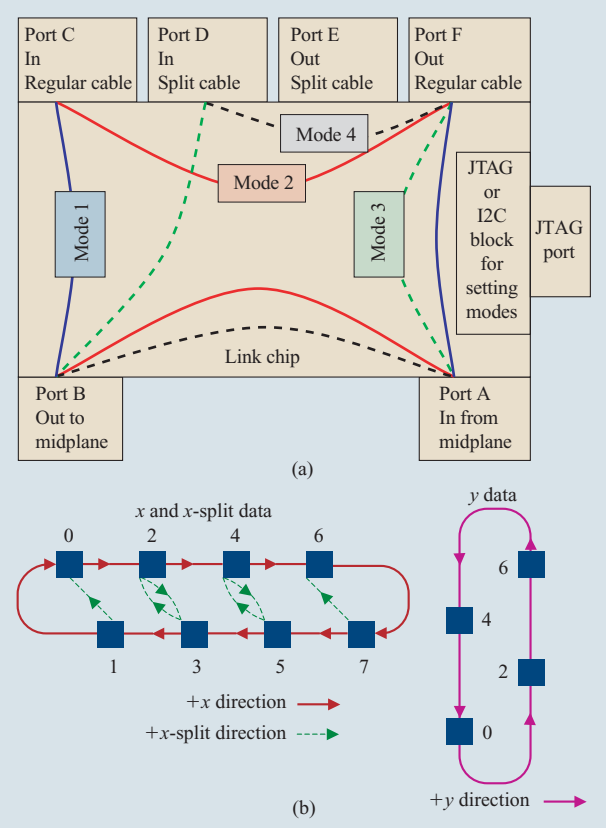


Figure 22

(a) Blue Gene/L link chip switch function. Four different modes of using the link chip. (b) Split partitioning: Blue Gene/L torus cable connections with and without split-redirection cables.

System partitioning and redundancy

The net result of the BLL partitioning function is the ability to flexibly divide one physical configuration of a multirack system into multiple logical partitions of various sizes. This is useful for both running different software simultaneously on efficiently sized partitions and for substituting other (redundant) midplanes while faulty sections are serviced. **Figure 23** shows the floorplan of a 64-rack BG/L system, with each colored square representing a rack and each color representing a different logical partition. If code had been running in a blue-colored $32 \times 32 \times 32$ x, y, z partition on the rightmost four columns of racks when a compute ASIC failed in

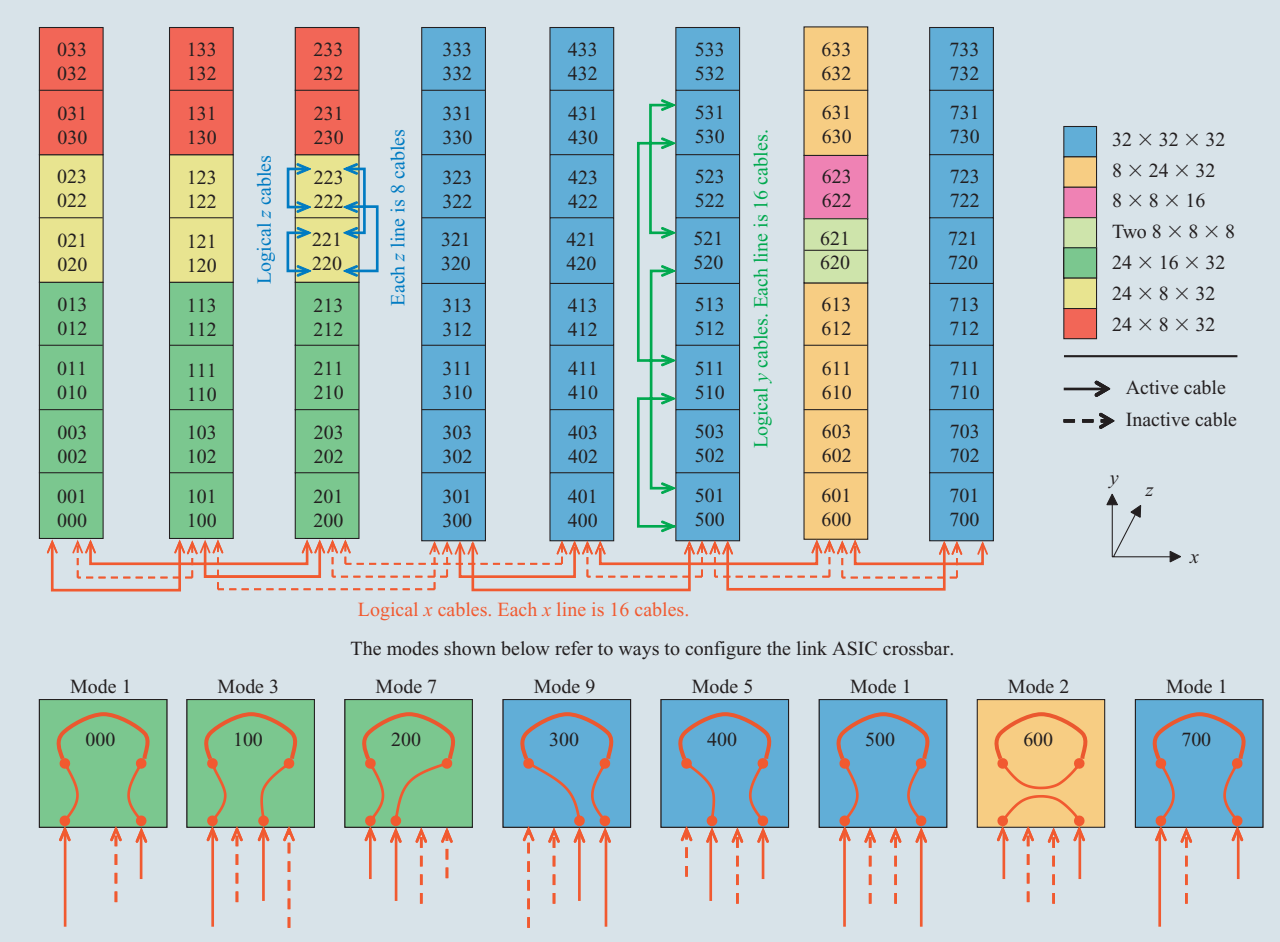


Figure 23

Examples of Blue Gene/L system partitioning.

midplane 620 of the second rightmost column, the control host could then reconfigure the system as shown. The $32 \times 32 \times 32$ partition is now allocated to other columns of racks, and midplanes 620 and 621 are isolated in their own local toruses. Applications can be restarted from their last checkpoint to disk and run in the other partitions, while the error in midplane 620 is serviced [2].

Power distribution system

Power system overview

Figure 24 shows the BG/L power system, which uses bulk ac–dc conversion to an intermediate safe extra-low voltage (SELV) of 48 VDC, then local point-of-load dc–dc conversion. Each rack is individually powered with either 208 VAC or 400 VAC three-phase power from a single 100-A line cord. Above the rack, six 4.2-kW bulk ac–dc converters, with a seventh for redundancy, distribute 48 VDC power through a cable harness to both sides of the midplanes; the midplane wiring carries the 48 V to the link, node, service, and fan cards that plug directly into it. The 48-V power and return are filtered to reduce EMI and are isolated from low-voltage ground to reduce noise.

When the circuit breaker for the ac–dc converter is switched on, power is applied to the midplanes and fans. The link, node, and service cards are designed to allow insertion and removal, while the midplane retains 48 V power once the local dc–dc converters have been disabled. Fan modules are designed to be hot-plugged. The fan, bulk power supplies, and local power converters all contain an industry-standard I2C bus, which can be used to power on, power off, and query status (voltage, current, temperature, trip points, fan speed, etc.) as well as to read the vital manufacturer product data.

The link, node, and service cards contain high-efficiency dc–dc converters that produce the required high current and low voltage where required. This minimizes

voltage drops, improves regulation and efficiency, and reduces possible electromagnetic compatibility (EMC) issues. The converters are very reliable, with a computed mean time between failure (MTBF) of at least 1M hours, or 1,000 FITs. However, even at this rate, there is one supply fail, on average, every five days in a 64-rack system. System power failures can be reduced to negligible levels by using redundant converters. BG/L uses two custom dc–dc power converters with integrated isolation power field-effect transistors (FETs). The eight link cards per rack each contain $1 + 1$ redundant 1.5-V converters, while the 32 node cards per rack each contain $3 + 1$ redundant dual-voltage 2.5-V–1.5-V converters, with peak current designed to match the peak demands of the memory system and core processor functions, respectively.

Bulk power assembly

The bulk power assembly consists of the bulk power enclosure (BPE) and the seven hot-pluggable ac–dc bulk power modules (BPMs). Each BPM is capable of delivering 4.2 kW, and only six are required to provide the 25-kW maximum for operation of a BG/L rack, making it $n + 1$ redundant. If one module fails, it may be removed and replaced with a spare while the system is in operation. The BPE houses the input EMC filter, circuit breaker, and power distribution buses.

The BPM is power-factor-corrected and achieves an 88% conversion efficiency over the operating range. It utilizes a conventional full-wave bridge rectifier, boost converter, full-bridge forward converter, isolated output full-wave rectifier, and diode-ORing to provide redundant operation. These 127-mm × 127-mm × 384-mm commercial units by Cherokee International were modified by adding an I2C-addressable monitoring and control system. A similar I2C interface is employed on the dc–dc converters and the fan assemblies. The host computer communicates through the CFPGA control network (see the section on the control network), and can control output voltage and monitor current, temperature, and vital product data (VPD) for each module. Converter status is continually monitored and service operations scheduled accordingly.

The BPM module is somewhat different, depending upon the line voltage of the installation. For the 208-VAC case, each BPM has a three-phase input, with the ac input connections of all modules wired in parallel. For the single-phase 400-VAC BPM, the BPE is wired to selectively distribute line power to the modules to approximately balance the phase loading. Additionally, the input line power in a system of multiple racks has phases distributed to further balance their loading.

A unique seven-bay BPE with integral breakers, filters, and distribution bus was designed to sit on top of the frame, since the bulk supply had the lowest mass density. The airflow is unusual; fans in the power modules draw cool air from both the front and rear of the rack, and discharge warm air vertically. This modest 2.5 kW of power is easily handled by the machine room air conditioning.

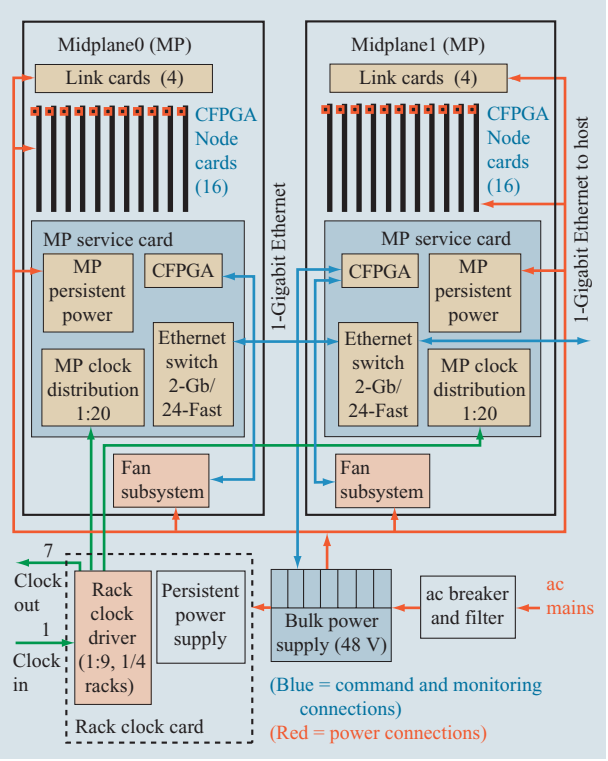


Figure 24

Blue Gene/L power, cooling, and control system. (CFPGA = control–FPGA chip.)

Choice of intermediate distribution voltage

Distribution of a dc power bus inside the machine has numerous advantages. It allows the use of the power-factor-corrected bulk power modules and thus removes the burden of line isolation and input voltage correction from the load converters. This improves EMC considerations, overall efficiency, and thermal management.

While a high-voltage distribution (350 V) is used in the IBM pSeries* and z900 servers [8], with the advantage of reduced current and small distribution cables, it also adds complications. The major problems are high-voltage safety requirements of spacing and insulation that must be carefully followed, as well as a very limited selection of high efficiency dc–dc converters and other components. At the 48-V SELV level, distribution of the voltage through the midplanes and cards is readily possible.

There are a large number of very-high-efficiency bulk ac–dc modules and dc–dc converters to choose from, as well as direct drive of fans for this voltage. The increased distribution current-carrying capacity was handled by distributing the power over numerous cables, connected to both sides of each midplane. It was the most cost-effective solution for this machine.

Distributed converters—redundancy

Because there are a very large number of dc–dc converters used on both the link and node cards, even the use of converters with an MTBF exceeding one million hours resulted in unacceptably high failure rates. Redundant converters were required. When the load voltage is several volts or higher, simple diode-ORing circuitry can be used to allow the parallel operation of many converters. During a fault condition, the diode provided isolation of the failed converter from the working converters that would supply the additional current. This worked as long as the supply voltage was large compared with the voltage drop across the diode. For the 1.5-V BG/L ASIC core voltage, another technology was required. By using synchronous rectifier converters, which drive low-resistance FET switches synchronized with the converter transformer drive, conversion efficiency of more than 92% can be achieved. A series FET switch was added to the power converter output to isolate the converters during a converter fail. A self-contained sensing circuit in each converter determines whether the output current is reversing and immediately disconnects the converter from the bus. The enabled output FET provided a very low-resistance path during normal operation while keeping the efficiency of the converters at 88%.

Voltage domains—redundancy and hot plugging

When a card fails in spite of the built-in redundancy, the power-supply system partitioning permits hot-plugging of a new card while other system functions remain active. As explained above, the link, node, and service cards can be replaced while the midplane retains 48 V power. If the service card has no failure, it remains active and provides control voltages for the CFPGA, clocking, and other control functions on the link and node cards. The dc–dc converter outputs on all other cards also remain on. Only the local dc–dc converters on the failed card are disabled, and the failed card is then replaced. In the example of a node card failure, all link ASICs can continue passing cable data traffic while the failed node card is replaced.

Decoupling and dc voltage drop design

The power-supply system function is to keep the V_{dd}-to-ground (Gnd) supply voltage at the active components constant to within acceptable tolerances. At low frequencies, the dc–dc power converters can respond to changing current needs, but the distribution system must be designed so that the dc voltage drop and heating between the dc–dc converters and the active components is acceptably low. At higher frequencies, the decoupling design must perform three main tasks: supply the temporary current demands of the active components above the frequency at which the dc–dc power supplies can respond, limit plane resonances within the circuit cards, and limit return current noise coupling when signals change reference planes. All of these tasks are geared toward limiting the variation in the supply voltage seen at the active components.

In the BG/L highest-current voltage domain, voltage drop was a concern. All of the dc–dc voltage regulators on high-current voltage domains monitor voltage sense lines and adjust their output to keep the voltage at these sense points constant. The sense points in the node and link card ASIC voltage supply domains are placed near the center of the active chips that are drawing the power. Several potential sense points were designed into the cards, and zero-ohm resistors were populated to connect only the sense point that minimized the variation in dc voltage drop across all ASICs. The only voltage domain for which the voltage variation was a potentially significant fraction (more than a few percent) of the supply voltage was the highest-current voltage domain in the system—the 1.5-V core voltage domain of the main compute ASICs. As discussed earlier in the section on card design, signal and ground layers in the node card were reallocated to the 1.5-V power domain to reduce the variation in dc drop to acceptable levels. In addition, there was one area in which the magnitude of the current density became a heating and reliability problem. Near the 1.5-V output of the dc–dc regulators that supply the 1.5-V ASIC core voltage on the node card, metal had to be added to the 1.5-V domain on several card layers to reduce the maximal current density and improve long-term card reliability.

Decoupling was needed on all voltage domains to meet the time-dependent current needs of the active components [7]. At low frequencies, below 20 kHz, the dc–dc power converter 25-μs feedback loop enabled the converter to respond to changing current demands. In this range, the converter appears essentially as a resistive feedback resistance (R_feedback) between the V_{dd} power and Gnd planes. At higher frequencies, the dc–dc converter feedback loop can no longer respond, and output capacitors inside the converter are needed to provide changing current up into the range of 50+ kHz. Above 50 kHz, on-card capacitors are needed to supply the current as the active component demand changes. Several types of capacitors are used. Capacitors with large capacitance, larger package size, and higher

inductance supply the current at mid-frequencies from 50 kHz to around 1 MHz, while ceramic surface-mount capacitors with smaller capacitance, small package size, and lower inductance are used to supply current at higher frequencies, from 1 MHz up through ~50 MHz. On-card decoupling capacitors are unable to provide temporary current through the inductive chip package to the active chip inside the package above frequencies in the range of ~50 MHz. Therefore, on-module decoupling capacitors were added for the high-frequency, low-voltage compute ASICs to hold the supply voltage seen by the ASICs within specification from 50 MHz up until the on-chip decoupling can take over.

The power-supply system response of each power domain was predicted by modeling the frequency response of the dc–dc converter and of the various decoupling capacitors in the system and by calculating the frequency-dependent impedance seen by the active chips at their V_{dd} and Gnd supply rails. This response was compared with the impedance required by the active chips to meet their current needs while holding the supply voltage within specification. Assumptions were made about each chip-frequency-dependent current demand, $I(f)$. The ratio of $I(f)$ to the allowed voltage tolerance gave the V_{dd}–Gnd impedance, $Z(f)$, that the chip required as a function of frequency. The required impedance of the chip, $Z(f)_{required}$, and the power-supply-system predicted impedance, $Z(f)_{supply}$, were both plotted, and modifications were made in the decoupling design until $Z(f)_{supply}$ was lower than $Z(f)_{required}$ at all frequencies of interest.

Figure 25 shows the design calculation for the decoupling on the 1.5-V compute ASIC core voltage domain. Models for the dc–dc regulator, bulk tantalum capacitors, high-frequency ceramic surface-mount card capacitors, and on-module AVX LICA** capacitors are included. For capacitors, the effective capacitance differs little from the specified capacitance. However, the effective input inductance for physically small capacitors is primarily due to the pads, trace, and vias that connect the capacitor to the power and ground planes, with only a small inductive contribution from the capacitor package itself. Capacitor models used in the simulation therefore use the effective resistance (R), inductance (L), and capacitance (C) of each device, including the effects of both the capacitor and its connections to the card planes. The pink data points show the predicted impedance between 1.5 V and ground at the chip, $Z(f)_{supply}$, and the red data points show the required impedance, $Z(f)_{required}$. The required impedance, $Z(f)$, of the primary 1.5-V core voltage domain was determined by limiting the allowed V_{dd}–Gnd dynamic voltage error to 5% of nominal, or 75 mV. When dc voltage drops were added, this would keep total V_{dd}–Gnd voltage excursions

dc–dc power supply	E case	C case	B case	1206	0805	LICA 0603	IDC on-module	on-module
0.17	3	0	0	0	56	0		
0.085	1.5	0	0	0	28	0	4	0
0.06	10	10	10	5	2.2	1.8	0.2	0.2
0.0011	**0.04**	**0.04**	**0.04**	0.1	0.1	0.1	0.05	0.05
5,000	**470**	**100**	**47**	**10**	**4.7**	**1**	**0.025**	**1**
0.0023								

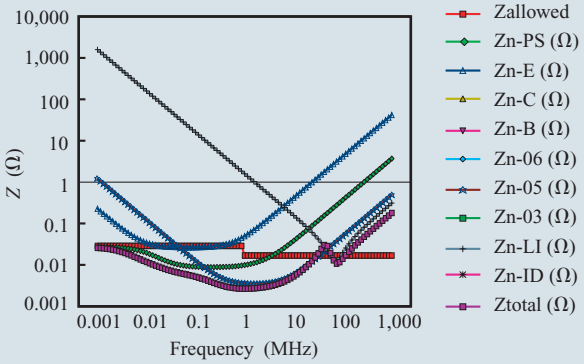


Figure 25

Blue Gene/L power supply decoupling, V_{dd}–Gnd impedance. (Zallowed = impedance budget; Zn = nominal or design impedance; Ztotal = total of all impedances.)

to within 100 mV of nominal to ensure proper circuit operation. On the basis of this analysis, four on-module capacitors and 28 on-card ceramic capacitors were used near each ASIC. For each two-ASIC card, three bulk capacitors and a portion (17%) of a dc–dc regulator were used.

Additional decoupling capacitors and ground stitching vias were used to control card plane resonances and to limit return-current discontinuities and noise coupling. Resonances were controlled by reducing the size of the voltage planes so that the planes did not cover the entire circuit card in cases in which active components of a particular voltage were not present in certain regions. In addition, every voltage plane had at least one ceramic capacitor between that voltage plane and ground within each square inch of card area. Return current discontinuities and noise were controlled by having a 100% ground-referenced design, by adding stitching vias to tie ground planes together where signals change layers, and by well-decoupling all I/O voltages to ground at all drivers and receivers.

Power system performance

The power consumed by the BG/L machine varies drastically depending upon the application being run.

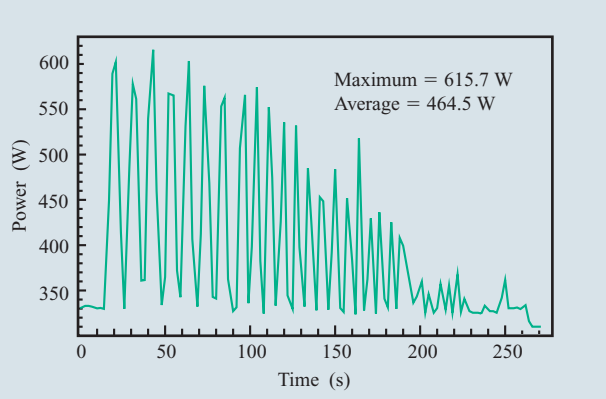


Figure 26

Node card power plotted against time under an extreme case of application load. The working memory size has been set to just exceed the L3 cache.

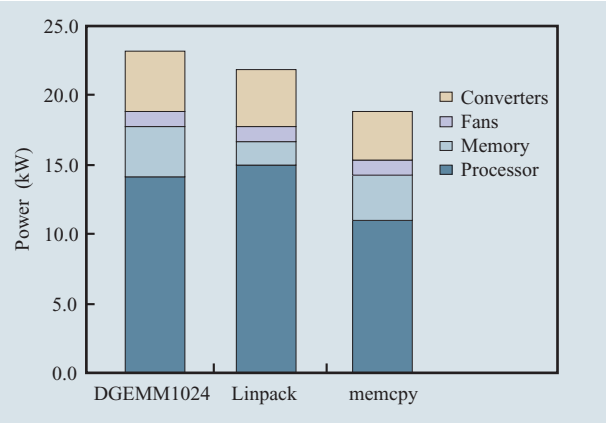


Figure 27

Power consumed by a Blue Gene/L compute rack at 700 MHz, for the cases of maximum total power (DGEMM1024), maximum processor power (Linpack), and maximum memory power (memcpy).

Figure 26 shows an example of the temporal power variation on a single 32-way node card for an application that alternates between periods of floating-point calculation and memory access. The wide-band decoupling network described above in the section on decoupling and dc voltage drop design handles these large current excursions, keeping the voltage variations within the acceptable noise margin, ensuring error-free performance.

There are three high-power operating modes: maximum total power, maximum processor power, and maximum memory power. These three measured cases are shown in **Figure 27**, where all sources of consumed power within the compute rack have been included, including power dissipated by the bulk power supplies. The rack is designed to deliver 25 kW of power to the rack, or 27 kW total including the losses of the bulk power supply. Power efficiency was one of the design goals of BG/L. The measured sustained performance is approximately 200 Mflops/W on a 512-W midplane partition including all power dissipated in the rack; the theoretical peak performance is 250 Mflops/W.

Control system

Philosophy

In the largest planned Blue Gene/L machine, there are 65,536 compute ASICs and 3,072 link ASICs. A large BG/L system contains more than a quarter million endpoints in the form of the ASICs, temperature sensors, power supplies, clock buffers, fans, status LEDs, and more. These all have to be initialized, controlled, and monitored.

One approach would be to put a service processor in each rack that was capable of communicating with each card (and each device on the card) and capable of communicating outside the rack where it can obtain configuration and initial program load (IPL) data. This service processor would run an operating system (OS) with application software capable of these functions. Both components of software are in nonvolatile memory. This approach is unwieldy and potentially unreliable for high-scale cellular systems in which many relatively small cells are desired.

For this machine, the service processor was kept out of the rack. A single external commodity computer (the *host*) is used as the service node. This is where the OS is run, along with the application code that controls the myriad devices in the rack. The host connects to the BG/L hardware through an Ethernet and CFPGA circuitry. To be more precise, the host communicates with Internet Protocol (IP) packets, so the host and CFPGA are more generally connected through an intranet (multiple Ethernets joined by IP routers). The CFPGA circuitry, in turn, drives the different local buses, such as JTAG, I2C, Serial Peripheral Interface (SPI), etc. to devices (such as power supplies) or chips (the compute and link ASICs) on a node, link, or service card. In a simple sense, the Ethernet and CFPGA circuitry forms a bus extender. A bus physically close to the host has been extended into the BG/L rack and has fanned out to many racks and hence many cards. In this manner, centralized control of the large cellular-based machine is achieved, and at the same time, the host is not limited to being a single computer. Using IP connectivity for this "bus extender" allows for

the flexibility of having the host be a set of computers interconnected with commodity networking fabrics.

Software development on this host uses all standard tools.

An exception: One local control function

In the BG/L control system design, one significant exception was made to the philosophy of keeping all service control function in the remote service host: over-temperature shutdown. Should the network connection between the compute racks and the service host ever fail, the system should still be able to shut itself down in the case of overheating. Each compute and I/O ASIC contains a thermal diode monitored by an external temperature sensor. There are also temperature sensors inside several other devices on the link, node, and service cards. If any of these sensors reports an over-temperature condition, the sensor logic in the link, node, or service card FPGA shuts down the local dc–dc power converters on the card. This sensor logic is housed inside the same FPGA chip as the CFPGA function, as shown in **Figure 28**.

A CFPGA implementation

A CFPGA chip is used on every major BG/L circuit card in a midplane. Figure 28 shows the implementation of the CFPGA circuitry. The host computer sends IP/UDP (User Datagram Protocol) packets onto an Ethernet intranet that then routes the packet to the CFPGA. The CFPGA comprises an Ethernet physical layer chip (PHY and its associated isolation transformer) and an FPGA chip. The PHY chip turns the Ethernet wire signaling protocol into a small digital bus with separate clock wires. Logic in the FPGA analyzes the bit stream (Ethernet packet) coming from the network (request packet). If the request packet has good cyclic redundancy check (CRC) and framing, and has the correct Ethernet and IP destination addresses for the CFPGA, data is extracted and acted upon, and a reply Ethernet packet is formulated and transmitted back to the host.

On the other side of the CFPGA are the interfaces with target devices. In BG/L these devices are

- JTAG: Compute ASICs, link ASICs (BLLs).
- I2C: Temperature sensors, power supplies, fan modules, and others.
- Via the control register: Global interrupt network logic and system clock management controls.

The CFPGA acts on a request by transferring data received from the host to a target, while simultaneously collecting return data from the target device for incorporation into the reply packet.

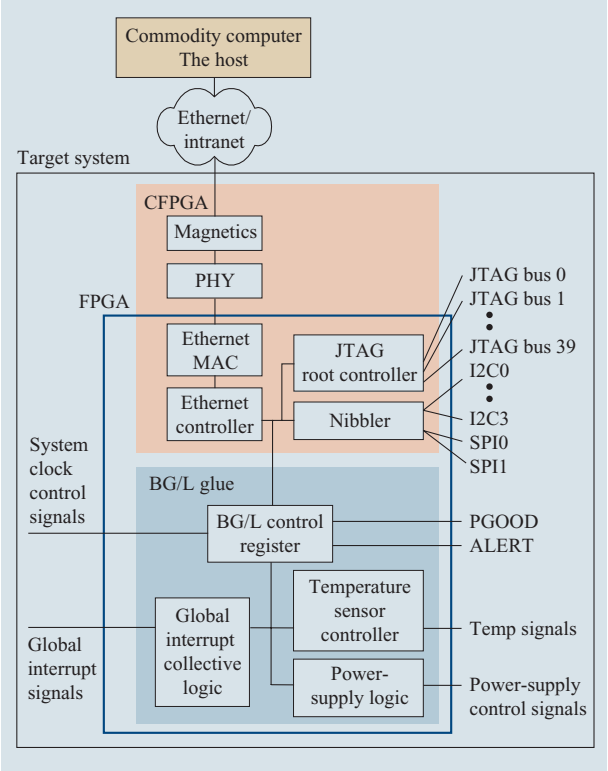


Figure 28

Control–FPGA chip implementation.

An essential feature of this control system is that all intelligence is placed in the software running on the host. The CFPGA circuitry is kept as simple as possible to increase its robustness and ease its maintenance. The host software typically matures with time, while the CFPGA circuitry remains unchanged. This claim has so far been demonstrated.

Midplane implementation

The fact that the CFPGA is not bulky enables widespread use within a rack. One is used on every major circuit card of a BG/L midplane. The high-bandwidth buses are JTAG, since they control a compute ASIC. These chips require a large amount of host data for three reasons: 1) performing in-system chip testing via manufacturing scan chains; 2) sending IPL data; and 3) retrieving software-debugging data. The first two put a requirement on the control system for ample bandwidth from the host to the target chip, and the last for ample bandwidth in the reverse direction.

The JTAG controller in the CFPGA has a broadcast function—the same data is sent out of a subset of its ports (ignoring any return data). With this function, a single

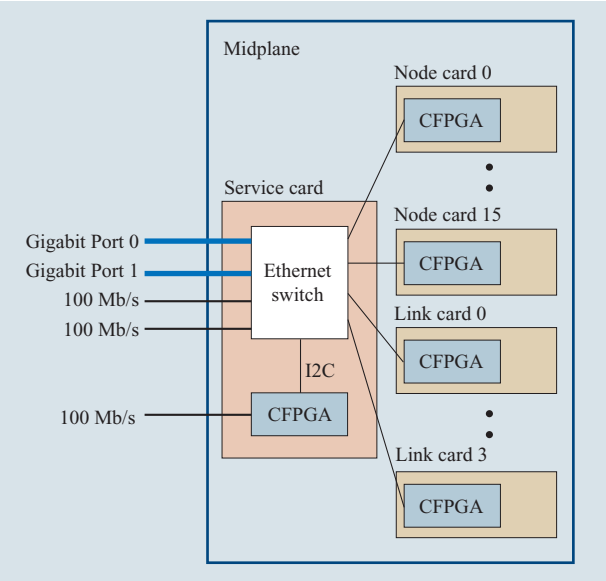


Figure 29

Midplane control system implementation. (CFPGA = control–FPGA.)

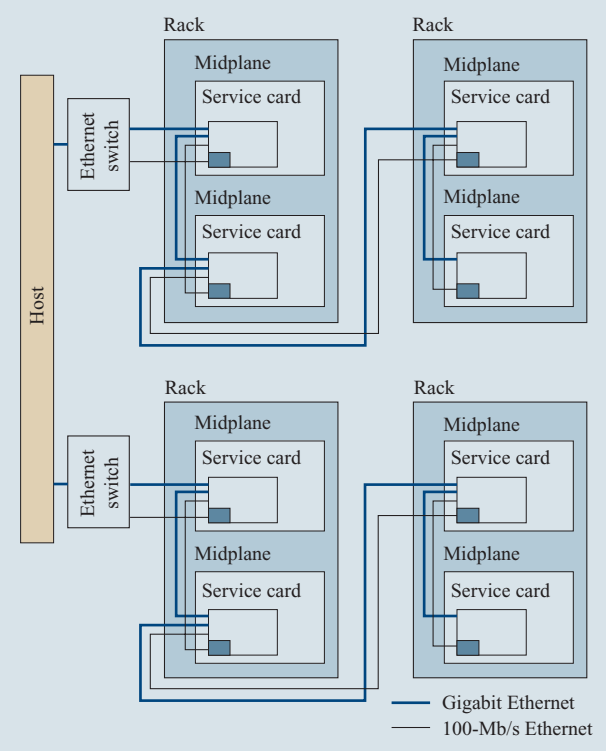


Figure 30

Typical multi-midplane control system configuration.

CFPGA for one midplane would be able to push data into all 576 compute ASICs with a 576:1 increase in speed. However, this does not help in retrieving data from the compute ASICs. Good performance on parallel software debugging is desired. The solution is to use multiple CFPGAs within a midplane with the appropriately sized Ethernet connecting them to the host. The JTAG bit rate is of the order of 10 Mb/s, while the Ethernet is at 100 Mb/s. This allows ten CFPGAs to operate in parallel with a single 100-Mb/s connection to the host. Operating larger numbers of CFPGAs in parallel is achieved by increasing the bandwidth to the host, for example, by using a Gigabit Ethernet switch to drive many 100-Mb/s Ethernet ports.

Figure 29 shows the implementation of the control system for a midplane. A CFPGA is placed on every card. An Ethernet switch on the service card combines 100-Mb/s Ethernet links from the 20 CFPGAs (16 node cards and four link cards) into Gigabit Ethernet, with two Gigabit Ethernet ports connecting to the outside world for ease in connecting multiple midplanes to the host. The Ethernet switch is managed by the service card CFPGA. This enabled simplified development and maintenance of this software component, and it allows the host to retrieve information from the Ethernet switch, e.g., switch tables, for diagnostic purposes.

Rack and multirack implementation

An essential feature of this control system is that a CFPGA speaks only when spoken to. In other words, it transmits an Ethernet packet (reply) only when asked to do so by a request packet. This feature enables the control system to scale up to many racks of midplane pairs. The only entity on the Ethernet that can cause traffic flow is the host; this prevents congestion on the network, which would harm performance. With this feature, a large system can be controlled using standard commodity Ethernet equipment, which is cost-effective and reliable.

Figure 30 shows a typical installation. A separate 100-Mb/s network connects all of the service card CFPGAs. This allows the service card Ethernet switches to be managed independently of the traffic flow across those switches. The traffic for other CFPGAs travels across multiple Gigabit networks to the host. This allows ample control system bandwidth.

Software

Each CFPGA is controlled by a single entity. Software is free to implement these CFPGA control entities in any way that is convenient. Each CFPGA control entity is responsible for managing the reliable flow of commands and replies.

The communication between a CFPGA chip and its control entity is very simple. The CFPGA has a

command sequence number. When the CFPGA chip receives a valid command and the sequence number in the command matches the sequence number in the CFPGA chip, the chip acts on the command and prepares the reply packet. As a side effect of the command execution, the sequence number is incremented, and this new number is stored in the reply packet buffer. When the reply has been assembled, the reply packet is returned to the control entity. The control entity is able to match the replies with the requests because replies have a sequence number one value larger than the sequence number used in the command.

To avoid dropped or duplicated packets in either direction, a simple retry mechanism is used. If the control entity does not receive a timely reply, it re-sends the last command using the original sequence number. The control entity does not know whether the command or the reply packet was dropped, but the CFPGA chip determines this. If the command packet was dropped, the CFPGA chip sees the second copy and takes the expected actions. If the CFPGA chip receives a command with the wrong sequence number, the CFPGA chip re-sends the contents of the reply buffer. This lets the control entity receive the expected reply so that it can sequence to the next command. This mechanism guarantees that the CFPGA chip will not act on duplicate command packets.

Once the control entity provides a reliable transport with the CFPGA chip, the code that interacts with devices connected to the CFPGA chip can communicate with the control entity using a wide range of protocols.

The BG/L system uses a multithreaded proxy process that provides the control entity for multiple CFPGA chips. Applications that must communicate with CFPGA-attached devices use Transmission Control Protocol (TCP) sockets to send commands to the proxy and receive replies for the commands. The proxy supports the sharing of devices by allowing applications to execute a series of commands to a particular device with no danger of intervening traffic. By allowing an application to perform an entire "unit of work" on a device, the application can ensure that the device is always left in a well-known state between operations. This allows cooperating applications to share devices.

The proxy has been very successful in supporting a wide range of applications. The applications have been written by a range of teams in geographically distant sites. Since the applications communicate with the hardware using TCP, it has been very easy to develop application code with a minimal amount of early hardware. The early BG/L hardware has been used by developers located on most continents and in many time zones. This decoupling of locality has been of tremendous value in the BG/L project.

The proxy is the most flexible and complex mechanism for communicating with the CFPGA-attached devices. This proxy supports applications ranging from assembly-level verification using JTAG boundary scan to low-level chip-debugging applications that use JTAG (Cronus, an internal low-level debug tool, and IBM RiscWatch), to ASIC configuration and IPL applications, to system environmental monitors, to component topology discover processes, and to high-level communication streams that provide console terminal support for Linux** kernels running on the BG/L nodes.

Simpler versions of control entities exist for less-demanding applications. In some cases, the control entity is tightly bound to the application; this minimizes the complexity of the application and can be very valuable for development and certain diagnostics applications.

It is anticipated that other software can and will be developed as more is learned about the ways in which the BG/L system is used. The very simple UDP-based command-and-response protocol used to communicate with the CFPGA chips places a very light burden on software and allows a wide range of design choices.

Summary

We have shown how the considerations of cost, power, cooling, signaling, electromagnetic radiation, mechanics, component selection, cabling, reliability, service strategy, risk, and schedule were used to design the Blue Gene/L system package. The result is a relatively low-cost, high-density rack within the limits of air cooling that can be easily scaled and maintained. Host computers and remote storage are easily added. The Blue Gene/L interconnect, power, cooling, and control systems have been shown to complement one another, reflecting the coordination and unified goals of the design team.

Acknowledgments

This project could not have been completed without the assistance of a large number of individuals. The Blue Gene/L project has been supported and partially funded by the Lawrence Livermore National Laboratory on behalf of the United States Department of Energy under Lawrence Livermore National Laboratory Subcontract No. B517552; Lynn Kissel (LLNL) supplied Figure 23 and was an excellent liaison to LLNL power, packaging, and cooling personnel. The management of George Chiu and the technical assistance of Alina Deutsch in cable selection and modeling, and of Barry Rubin in connector model creation, was invaluable, as was the technical support of Christopher Surovic in assembling the prototypes and the thermal rack. Ruud Haring and Arthur Bright helped us with the ASIC pin locations and the module designs. James Speidel, Ron Ridgeway, Richard Kaufman, Andrew Perez, Joseph Boulais, Daniel

Littrell, Robert Olyha, Kevin Holland, Luis Sanchez, Henry Pollio, and Michael Peldunas from the Central Scientific Services (CSS) Department of the IBM Thomas J. Watson Research Center assisted in the card assembly, physical design, and microcode development; we also relied upon William Surovic, Matteo Flotta, Alan Morrison, James Polinsky, Jay Hammershoy, Joseph Zinter, Dominick DeLaura, Fredrik Maurer, Don Merte, John Pucylowski, Thomas Ruiz, Frank Yurkas, and Thomas Hart from CSS mechanical. The efforts of a large number of individuals in the Engineering and Technology Services area are represented here, including Scott LaPree, Christopher Tuma, Max Koschmeder, David Lund, Michael Good, Lemuel Johnson, and Greg Lamp, who collaborated on the mechanical design; James Anderl and Ted Lin, who assisted with thermal issues; Todd Cannon, Richard Holmquist, Vinh Pham, Dennis Wurth, Scott Fritton, and Jerry Bartley, who helped complete and document the card designs; Don Gilliland and Charles Stratton, who refined the EMI shielding and filtering; and Brian Hruby and Michael Vaughn, who assisted in power-supply validation, while Brian Stanczyk helped refine the cable design; Marvin Misgen kept us compliant for safety. Norb Poch provided installation planning. Darryl Becker provided detailed design information to the IBM Microelectronics Division for the custom ground-referenced first-level packages for both ASICs. Robert Steinbugler, Jerry Muenkel, and Ronald Smith of IBM Raleigh provided the industrial design, while Kevin Schultz ensured that we were compliant with system status monitors. Last, but by no means least, Steve Strange, Dennis Busche, Dale Nelson, and Greg Dahl of IBM Rochester procurement helped secure suppliers and expedite procurement.

References

1. F. Allen, G. Almasi, W. Andreoni, D. Beece, B. J. Berne, A. Bright, J. Brunheroto, C. Cascaval, J. Castanos, P. Coteus, P. Crumley, A. Curioni, M. Denneau, W. Donath, M. Eleftheriou, B. Fitch, B. Fleischer, C. J. Georgiou, R. Germain, M. Giampapa, D. Gresh, M. Gupta, R. Haring, H. Ho, P. Hochschild, S. Hummel, T. Jonas, D. Lieber, G. Martyna, K. Maturu, J. Moreira, D. Newns, M. Newton, R. Philhower, T. Picunko, J. Pitera, M. Pitman, R. Rand, A. Royyuru, V. Salapura, A. Sanomiya, R. Shah, Y. Sham, S. Singh, M. Snir, F. Suits, R. Swetz, W. C. Swope, N. Vishnumurthy, T. J. C. Ward, H. Warren, and R. Zhou, "Blue Gene: A Vision for Protein Science Using a Petaflop Supercomputer," *IBM Syst. J.* **40**, No. 2, 310–327 (2001).
2. N. R. Adiga, G. Almasi, Y. Aridor, R. Barik, D. Beece, R. Bellofatto, G. Bhanot, R. Bickford, M. Blumrich, A. A. Bright, J. Brunheroto, C. Cascaval, J. Castaños, W. Chan, L. Ceze, P. Coteus, S. Chatterjee, D. Chen, G. Chiu, T. M. Cipolla, P. Crumley, K. M. Desai, A. Deutsch, T. Domany, M. B. Dombrowa, W. Donath, M. Eleftheriou, C. Erway, J. Esch, B. Fitch, J. Gagliano, A. Gara, R. Garg, R. Germain, M. E. Giampapa, B. Gopalsamy, J. Gunnels, M. Gupta, F. Gustavson, S. Hall, R. A. Haring, D. Heidel, P. Heidelberger, L. M. Herger, D. Hill, D. Hoenicke, R. D. Jackson, T. Jamal-Eddine, G. V. Kopcsay, E. Krevat, M. P. Kurhekar, A. Lanzetta, D. Lieber, L. K. Liu, M. Lu, M. Mendell, A. Misra, Y. Moatti, L. Mok, J. E. Moreira, B. J. Nathanson, M. Newton, M. Ohmacht, A. Oliner, V. Pandit, R. B. Pudota, R. Rand, R. Regan, B. Rubin, A. Ruehli, S. Rus, R. K. Sahoo, A. Sanomiya, E. Schenfeld, M. Sharma, E. Shmueli, S. Singh, P. Song, V. Srinivasan, B. D. Steinmacher-Burow, K. Strauss, C. Surovic, R. Swetz, T. Takken, R. B. Tremaine, M. Tsao, A. R. Umamaheshwaran, P. Verma, P. Vranas, T. J. C. Ward, M. Wazlowski, W. Barrett, C. Engel, B. Drehmel, B. Hilgart, D. Hill, F. Kasemkhani, D. Krolak, C. T. Li, T. Liebsch, J. Marcella, A. Muff, A. Okomo, M. Rouse, A. Schram, M. Tubbs, G. Ulsh, C. Wait, J. Wittrup, M. Bae, K. Dockser, L. Kissel, M. K. Seager, J. S. Vetter, and K. Yates, "An Overview of the BlueGene/L Supercomputer," *Proceedings of the ACM/IEEE Conference on Supercomputing*, 2002, pp. 1–22.
3. M. L. Fair, C. R. Conklin, S. B. Swaney, P. J. Meaney, W. J. Clarke, L. C. Alves, I. N. Modi, F. Freier, W. Fischer, and N. E. Weber, "Reliability, Availability, and Serviceability (RAS) of the IBM eServer z990," *IBM J. Res. & Dev.* **48**, No. 3/4, 519–534 (2004).
4. S. F. Hoerner, *Fluid Dynamic Drag: Practical Information on Aerodynamic Drag and Hydrodynamic Resistance*, Hoerner Fluid Dynamics, Bakersfield, CA, 1965. Library of Congress 64-19666; ISBN: 9991194444.
5. F. P. Incropera and D. P. DeWitt, *Fundamentals of Heat and Mass Transfer*, Fifth Edition, John Wiley & Sons, Hoboken, NJ, 2002; ISBN 0-471-38650-2.
6. A. A. Bright, R. A. Haring, M. B. Dombrowa, M. Ohmacht, D. Hoenicke, S. Singh, J. A. Marcella, R. Lembach, S. M. Douskey, M. R. Ellavsky, C. Zoellin, and A. Gara, "Blue Gene/L Compute Chip: Synthesis, Timing, and Physical Design," *IBM J. Res. & Dev.* **49**, No. 2/3, 277–287 (2005, this issue).
7. T. M. Winkel, W. D. Becker, H. Harrer, H. Pross, D. Kaller, B. Garben, B. J. Chamberlin, and S. A. Kuppinger, "First- and Second-Level Packaging of the z990 Processor Cage," *IBM J. Res. & Dev.* **48**, No. 3/4, 379–394 (2004).
8. P. Singh, S. J. Ahladas, W. D. Becker, F. E. Bosco, J. P. Corrado, G. F. Goth, S. Iruvanti, M. A. Nobile, B. D. Notohardjono, J. H. Quick, E. J. Seminaro, K. M. Soohoo, and C. Wu, "A Power, Packaging, and Cooling Overview of the IBM eServer z900," *IBM J. Res. & Dev.* **46**, No. 6, 711–738 (2002).

Received May 12, 2004; accepted for publication July 26, 2004; Internet publication March 23, 2005

Paul Coteus *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (coteus@us.ibm.com)*. Dr. Coteus received his Ph.D. degree in physics from Columbia University in 1981. He continued at Columbia to design an electron–proton collider, and spent 1982 to 1988 as an Assistant Professor of Physics at the University of Colorado at Boulder, studying neutron production of charmed baryons. In 1988, he joined the IBM Thomas J. Watson Research Center as a Research Staff Member. Since 1994 he has managed the Systems Packaging Group, where he directs and designs advanced packaging and tools for high-speed electronics, including I/O circuits, memory system design and standardization of high-speed DRAM, and high-performance system packaging. His most recent work is in the system design and packaging of the Blue Gene/L supercomputer, where he served as packaging leader and program development manager. Dr. Coteus has coauthored numerous papers in the field of electronic packaging; he holds 38 U.S. patents.

H. Randall Bickford *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (bick@us.ibm.com)*. Mr. Bickford is a member of the Package Design and Analysis Group. He received a B.S. degree in chemical engineering from the University of New Hampshire in 1971 and an M.S. degree in materials science and mechanical engineering from Duke University in 1977. He joined IBM in 1971 and has worked at the IBM Thomas J. Watson Research Center since 1978. Mr. Bickford's research activities have included development of miniaturized packages for the Josephson technology program, bonding techniques and advanced structures for low-cost packaging, chemical modification of fluoropolymer materials, patterned dielectric planarization, and copper metallization. He has worked in the area of systems packaging since 1995, focusing on high-performance circuit card improvements through enhanced interface circuit designs and layout, and the analysis of signal integrity issues for critical interconnection net structures, most recently for the Blue Gene/L supercomputer. Mr. Bickford is the co-holder of 18 U.S. patents.

Thomas M. Cipolla *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (tcipolla@us.ibm.com)*. Mr. Cipolla is a Senior Engineer in the Systems Department at the Thomas J. Watson Research Center. He received a B.A. degree from Baldwin-Wallace College, a B.S.M.E. degree from Carnegie-Mellon University in 1964, and an M.S.M.E. degree from Cleveland State University in 1973. He was with the General Electric Company Lamp Division, Cleveland, Ohio, and G. E. Corporate Research and Development, Schenectady, New York, from 1969 to 1984. He joined the IBM Research Division in 1984. His most recent work is in high-density electronic packaging and thermal solutions. Mr. Cipolla has received numerous awards for his work at IBM; he holds 53 U.S. patents.

Paul G. Crumley *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (pgc@us.ibm.com)*. Mr. Crumley has worked in the IBM Research Division for more than 20 years. His work and interests span a wide range of projects including distributed data systems, high-function workstations, operational processes, and, most recently, cellular processor support infrastructure.

Alan Gara *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (alangara@us.ibm.com)*. Dr. Gara is a Research Staff Member at the IBM Thomas J. Watson Research Center. He received his Ph.D. degree in physics from the University of Wisconsin at Madison in 1986. In 1998 Dr. Gara received the Gordon Bell Award for the QCDSP supercomputer in the most cost-effective category. He is the chief architect of the Blue Gene/L supercomputer. Dr. Gara also led the design and verification of the Blue Gene/L compute ASIC as well as the bring-up of the Blue Gene/L prototype system.

Shawn A. Hall *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (sahall@us.ibm.com)*. Dr. Hall received his Ph.D. degree in applied mechanics from the California Institute of Technology in 1981. He joined IBM in 1981 as a Research Staff Member at the IBM Thomas J. Watson Research Center. He has worked in numerous fields, primarily in development of machines and processes, but with occasional forays into software. His other fields of interest include laser and impact printing, fiber-optic packaging, rapid prototyping (using polymer to build 3D objects from CAD models), 3D graphics software, software and algorithms for real-time analysis of musical sound, micro-contact printing (a low-cost alternative to optical lithography), and computer cooling and mechanical packaging. He worked on cooling and mechanical design, and designing and testing rack-level thermal prototypes to develop and optimize the airflow path for the Blue Gene/L project. Dr. Hall holds nine U.S. patents.

Gerard V. Kopcsay *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (kopcsay@us.ibm.com)*. Mr. Kopcsay is a Research Staff Member. He received a B.E. degree in electrical engineering from Manhattan College in 1969, and an M.S. degree in electrical engineering from the Polytechnic Institute of Brooklyn in 1974. From 1969 to 1978 he was with the AIL Division of the Eaton Corporation, where he worked on the design and development of low-noise microwave receivers. He joined the IBM Thomas J. Watson Research Center in 1978. Mr. Kopcsay has worked on the design, analysis, and measurement of interconnection technologies used in computer packages at IBM. His research interests include the measurement and simulation of multi-Gb/s interconnects, high-performance computer design, and applications of short-pulse phenomena. He is currently working on the design and implementation of the Blue Gene/L supercomputer. Mr. Kopcsay is a member of the American Physical Society.

Alphonso P. Lanzetta *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (ptlanz@us.ibm.com)*. Mr. Lanzetta is a Staff Engineer in the System Power Packaging and Cooling Group at the IBM Thomas J. Watson Research Center. He has worked in the areas of mechanical packaging design and printed circuit board design since joining IBM in 1988. He has worked closely with his Blue Gene/L colleagues on nearly a dozen cards, six of which comprise the final system. Mr. Lanzetta is the co-holder of 21 U.S. patents.

Lawrence S. Mok *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (lmok@us.ibm.com)*. Dr. Mok received a B.S. degree in electrical engineering from the University of Tulsa, and M.S. and Ph.D. degrees in nuclear engineering from the University of Illinois

at Urbana–Champaign. He joined the IBM Thomas J. Watson Research Center in 1984. He has worked on various topics related to electronic packaging, especially in the thermal management area, including the cooling and power design of massively parallel computers and thermal enhancement for laptops. Dr. Mok has published several papers and holds 43 U.S. patents.

Rick Rand *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (rarand@us.ibm.com)*. Mr. Rand is a Senior Engineer in the System Power Packaging and Cooling Group. He received a B.E.E. degree from The Cooper Union for the Advancement of Science and Art, and an M.S.E.E. degree from the University of Pennsylvania. He has worked in the fields of high-speed pulse instrumentation and medical electronics. His other areas of interest at IBM have included supercomputer power and cooling systems, VLSI design for high-end servers, high-resolution flat-panel display (LCD) technology, optical communications, and optical inspection. He was a major contributor to the successful IBM Scalable POWERparallel* (SP*) processor systems, and is currently working on the Blue Gene supercomputer project. Mr. Rand has published seven technical articles and holds six U.S. patents.

Richard Swetz *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (swetz@us.ibm.com)*. Mr. Swetz received a B.E. degree in electrical engineering from Manhattan College in 1986 and an M.S. degree in electrical engineering from Columbia University in 1990. From 1986 to 1988, he was employed by the General Instrument Corporation, Government Systems Division, designing analog circuitry for radar systems. He joined IBM in 1988. Mr. Swetz has worked on various projects including the IBM Scalable POWERparallel (SP) series of machines and the scalable graphics engine.

Todd Takken *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (taken@us.ibm.com)*. Dr. Takken is a Research Staff Member at the IBM Thomas J. Watson Research Center. He received a B.A. degree from the University of Virginia, and an M.A. degree from Middlebury College; he finished his Ph.D. degree in electrical engineering at Stanford University in 1997. He then joined the IBM Research Division, where he has worked in the areas of signal integrity analysis, decoupling and power system design, microelectronic packaging, parallel system architecture, packet routing, and network design. Dr. Takken holds more than a dozen U.S. patents.

Paul La Rocca *IBM Engineering and Technology Services, 3605 Highway 52 N., Rochester, Minnesota 55901 (plarocca@us.ibm.com)*. Mr. La Rocca received a B.S. degree in mechanical engineering from Rutgers University in 1983. He joined IBM later that year in Endicott, New York, as a Facilities Design Engineer specializing in the design of heating, ventilation, and air conditioning system (HVAC) and control systems. He moved to packaging development in 1987 as a thermal engineer working on blower and cooling design for rack-mounted systems. He later transferred to IBM Rochester as a project manager for entry systems development, and in 1998 joined the System Package Design and Integration Team. Mr. La Rocca has been involved in all aspects of system-level packaging design including mechanical and power system development, thermal and acoustic design and optimization, and classical testing.

Christopher Marroquin *IBM Engineering and Technology Services, 3605 Highway 52 N., Rochester, Minnesota 55901 (marroqui@us.ibm.com)*. Mr. Marroquin received his B.S. degree in mechanical engineering from the University of Michigan in 1999. He is currently working toward his M.B.A. degree through the University of Minnesota. He began working for IBM in 1999 in mechanical design and integration. He has developed many server products while working at IBM, taking the design from early concept to a manufactured entity. He worked on the detailed mechanical design and system layout and is leading the IBM Engineering and Technology Services Power Packaging and Cooling Team for Blue Gene/L. He also coordinated the Blue Gene/L power, packaging, cooling, and classical testing. He is currently leading the installation effort for this large supercomputer system. Mr. Marroquin holds one U.S. patent, with several other patents pending.

Philip R. Germann *IBM Engineering and Technology Services, 3605 Highway 52 N., Rochester, Minnesota 55901 (pgermann@us.ibm.com)*. Mr. Germann is an Advisory Signal Integrity Engineer. He received a B.S. degree in physics from South Dakota State University in 1977, joining IBM as a Hardware Development Engineer for the IBM iSeries* and IBM pSeries servers. He completed his M.S. degree in electrical engineering from the University of Minnesota in 2002, with an emphasis on transmission lines and time-domain package characterization using time domain reflectometry (TDR). He moved from server hardware development to Engineering and Technology Services and has worked on packaging challenges ranging from overseeing card designs for a large-scale telecommunication rack-mounted switch (six million concurrent users) for a start-up company, to analyzing and designing a compact, high-performance personal computer (PC) fileserver on a peripheral component interface (PCI) form factor which was fully functional in a single design pass. Mr. Germann is the signal integrity and packaging leader for high-performance hardware, including the Blue Gene/L supercomputer, for IBM Engineering and Technology Services.

Mark J. Jeanson *IBM Engineering and Technology Services, 3605 Highway 52 N., Rochester, Minnesota 55901 (mjeanson@us.ibm.com)*. Mr. Jeanson is a Staff Engineer currently working as card owner and project leader for IBM Engineering and Technology Services. He joined IBM in 1981 after briefly working at the Mayo Clinic. He received an A.A.S. technical degree in 1986 from Rochester Community College and a B.S. degree in electrical engineering in 1997 through the IBM Undergraduate Technical Education Program (UTEP) program. He worked in manufacturing as an Early Manufacturing Involvement and Failure Analysis (EMI/FA) Engineer and was involved in more than ten bring-ups of the IBM AS/400*, iSeries, and pSeries servers. After receiving his B.S. degree, he joined the IBM Rochester development team with a focus on high-end processors and memory cards. Mr. Jeanson is currently the team leader for Blue Gene/L card hardware and support of IBM Legacy iSeries and pSeries server cards.