12.1 Cellular Supercomputing with System-On-A-Chip

G. Almasi, G. S. Almasi, D. Beece, R. Bellofatto, G. Bhanot, R. Bickford, M. Blumrich, A. A. Bright, J. Brunheroto, C. Cascaval, J. Castaños, L. Ceze, P. Coteus, S. Chatterjee, D. Chen, G. Chiu, T. M. Cipolla, P. Crumley, A. Deutsch, M. B. Dombrowa, W. Donath, M. Eleftheriou, B. Fitch, J. Gagliano, A. Gara, R. Germain, M. E. Giampapa, M. Gupta, F. Gustavson, S. Hall, R. A. Haring, D. Heidel, P. Heidelberger, L. M. Herger, D. Hoenicke, R. D. Jackson, T. Jamal-Eddine, G. V. Kopcsay, A. P. Lanzetta, D. Lieber, M. Lu, M. Mendell, L. Mok, J. Moreira, B. J. Nathanson, M. Newton, M. Ohmacht, R. Rand, R. Regan, R. Sahoo, A. Sanomiya, E. Schenfeld, S. Singh, P. Song, B. D. Steinmacher-Burow, K. Strauss, R. Swetz, T. Takken, P. Vranas, T. J. C. Ward, J. Brown[1], T. Liebsch[1], A. Schram[1], G. Ulsh[1]

IBM TJ Watson Research, Yorktown Heights, NY / [1]IBM Enterprise Server Group

Large powerful networks coupled to state-of-the-art processors traditionally dominate supercomputing. As technology advances, this approach is likely to be challenged by a more cost-effective system-on-a-chip approach, with higher levels of system integration. The scalability of applications to architectures with tens to hundreds of thousands of processors is critical to the success of this approach. Significant progress has been made in mapping numerous compute-intensive applications, many of them grand challenges, to parallel architectures. Applications hoping to execute efficiently on future supercomputers of any architecture must be coded in a manner consistent with an enormous degree of parallelism.

The conventional approach of building tightly-coupled clusters of large symmetric multiprocessing (SMP) nodes suffers lower application efficiencies due to the distance to memory growing as technology advances. Because many modern scientific applications (for example the multi-grid conjugate gradient (MGCG) application mentioned below) have little cache reuse, performance is often impossible to recover through cache techniques. System-on-a-chip level integration allows for low latency to all levels of memory including local main store, significantly lessening this problem.

The BlueGene/L (BG/L) program is developing a peak nominal 180TFLOPS (360TFLOPS for some applications) supercomputer to serve a broad range of science applications. BG/L generalizes QCDOC [1], the first system-on-a-chip supercomputer that-which is expected in 2003. BG/L consists of 65,536 nodes, and contains four integrated networks: a 3D torus [2], a combining tree, a Gb Ethernet network and JTAG.

The 3D torus interconnect is organized as 64x32x32 nodes. Every node is connected to 6 bi-directional torus links, each with a bandwidth of 350MB/s in each direction. For general communication between nodes, throughput and latency are optimized through adaptive, minimal path, virtual cut-through routing [3]. Two virtual channels provide fully-dynamic adaptive routing for high throughput [4], while two additional channels are reserved for guaranteed deadlock-free routing and low-latency, priority routing. Each node sources and sinks a global binary combining tree, allowing any node to broadcast to all others with a 2μs latency and 1.4GB/s bandwidth. Hardware provides reductions in the tree such as integer addition and maximum. Each sub-tree of 64 compute nodes is serviced by a dedicated I/O node with a Gb Ethernet link resulting in an aggregate system bandwidth of 1Tb/s to a large RAID disk system. The physical architecture of the BG/L system is closely tied to the 3D torus. Eight nodes on a circuit card form a 2x2x2 cube. Sixty-four of these 8-way cards share a midplane to form an 8x8x8 cube. Sixty-four racks, each with two 16"x22" midplanes, make up the full torus. The machine can be electrically partitioned into independent computers, each with their own independent networks. The BG/L machine will have spare rows of nodes that can be swapped in utilizing the partitioning functionality to achieve high reliability and accessibility. There is also a dedicated RAS network based on 100Mb Ethernet and JTAG.

Each 15W node (Figure 12.1.3) consists of a single ASIC and 9 SDRAM-DDR memory chips, totaling 256MB. The ASIC uses a CMOS 0.13μm technology. In the ASIC diagram, the gray blocks are standard system-on-a-chip offerings from an ASIC library. The white blocks require a new design effort, while the hatched blocks are developed from existing designs. Each of the two symmetric 700MHz PowerPC 440 cores delivers 2.8GFLOPS, although the normal mode of operation dedicates one processor to message handling. The ASIC contains the network components and the memory caches. The L2 caches are small, and provide prefetch storage for the L1 caches of the processor cores. The L3 cache consists primarily of 4MB of on-chip embedded DRAM. There is a 16B error-correcting DDR SDRAM controller integrated into each node. This physically small node coupled with a high-density interconnect allows 5.6TFLOPS peak performance in a single rack, which is anticipated to consume 15kW.

To predict application performance on BG/L, both single node and network performance are considered. Often computations can be broken down into a compute phase followed by a communications phase. Let T1 and T2 be the time (in ns) for the compute and communications phases, and suppose that an average of f floating point ops/processor/ns are executed during the compute phase. For P processors, the sustained application performance is P*f*T1/(T1+ T2) ops/ns. A detailed, near cycle accurate, simulator of the torus allows prediction of T2. T1 and f can be estimated through an understanding of how an algorithm computational requirements map onto the hardware, including factors such as memory bandwidth, etc.. Using this methodology, BG/L performance is predicted for two demanding applications: a 3D FFT, which stresses the network, and a sparse linear matrix solver, MGCG (multi-grid conjugate gradient), which stresses single node performance. Performance estimates are given in Figure 12.1.4.

A 3D 64b complex FFT of size N^3 is mapped on a machine of size X*Y*Z. Each node contains all x-coordinate values, each node in an x-y plane contains N/(X*Y) y-coordinate values, and each node in a z-row contains N/Z z-coordinate values. Local 1D FFTs are performed followed by transposes in which each processor sends messages to all processors in either the same x-y plane, or the same z-row. The simulator predicts that the network is approximately 85% efficient for such transposes. Measurements on a Power3 CPU indicate at least 30% efficiency (relative to peak) is achievable for 1D FFTs. Therefore, it is estimated that a 64k-node BG/L is approximately 29 times faster than Livermore's Blue-Pacific machine.

MGCG, which iteratively solves a 7-point discretization of a diffusion equation on a 3D uniform mesh, requires communication with 6 "neighbors" as well as a global reduction on the tree. The neighbors may be either nearest neighbors on the torus, or randomly distributed over the torus. For nearest neighbors, the network is 83% efficient with the limitation being the FIFO fill/drain bandwidth. For random neighbors on a symmetrical machine, the network is ~80% efficient.

Technology scaling yields dramatic increases in gate density, which permits multiple processing cores per chip. Three main issues constrain effectiveness with which highly-integrated nodes can harness the added computational power of multiple processing cores: node bandwidth, network traffic and power. The required network bandwidth for each node will grow because the network will need to support more processors, each running at a faster speed. As the total number of nodes in the system increases, the network will also need to carry additional cut-through traffic, further increasing the difficulties. As the number of cores is increased, the node power will also increase. Aggressive power-saving techniques will be required to utilize the processing and cost advantages offered by a multiprocessor node. A local memory system based on commodity DRAM will continue to be a critical component for high performance computing. The bandwidth requirements to this memory system will rise dramatically as process size shrinks due to the increasing number of faster processing cores.

References:

[1] QCDOC: A 10-TERAFLOPS SCALE COMPUTER FOR LATTICE QCD. Nucl.Phys.Proc.Suppl.94:825-832,2001

[2] SL Scott and GM Thorson, "The Cray T3E Network: Adaptive Routing in a High Performance 3D Torus," In Proceedings of HOT Interconnects IV, 1996

[3] P. Kermani and L. Kleinrock, "Virtual Cut-Through: A New Computer Communication Switching Technique," Computer Networks 3, 267-286, 1979

[4] WJ Dally, Virtual-Channel Flow Control. IEEE Trans on Parallel and Distributed Systems 3, No. 2, 194-205, 1992

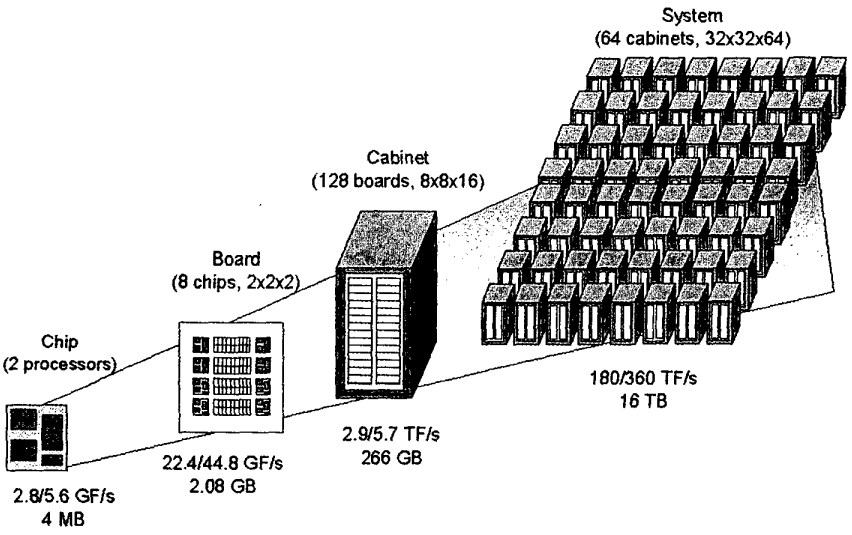

Figure 12.1.1: BlueGene/L packaging hierarchy.

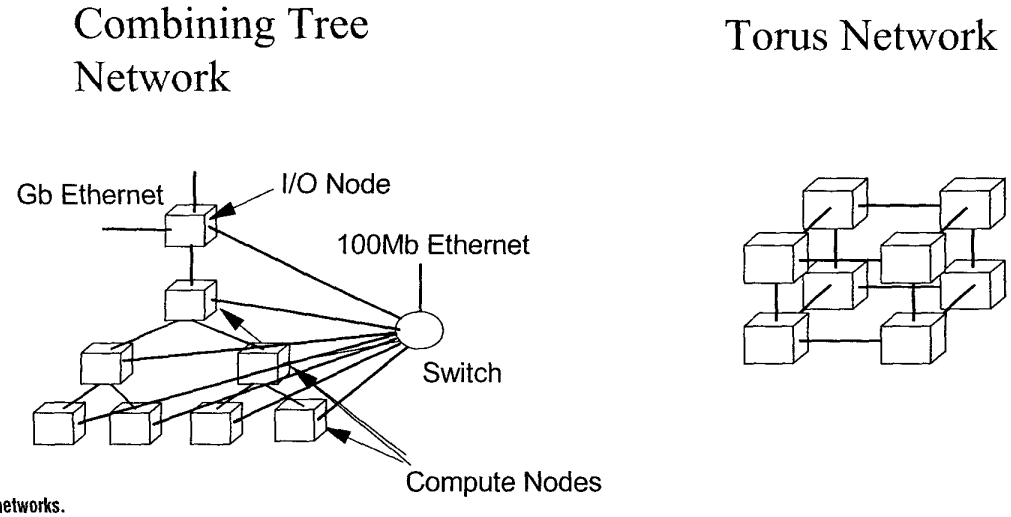

Figure 12.1.2: BG/L internal networks.

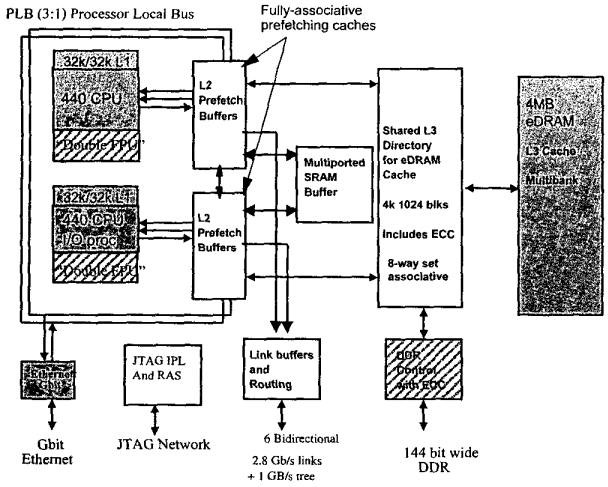

Figure 12.1.3: BG/L single ASIC node.

For the FFT, there are 2M elements per node in the fixed size/node case and a total of 1 billion elements in the fixed total size case. For MGCG, the sub-mesh size is 32^3 elements/node in the fixed size/node case and there are a total of one billion elements in the fixed total size case. MGCG with random neighbors does not map well onto the asymmetrical 64k machine since it creates a highly non-uniform network traffic pattern. Note: (a) For reasons of symmetry, the 64k system has 1million elements/node.

Application	Machine Size	TeraOp/sec Fixed Size per Node	TeraOp/sec Fixed Total Size
FFT	512	0.27	0.27
	4k	1.64	1.56
	32k	9.15	8.07
	64k	13.72 (a)	11.93
MGCG Nearest Neighbor	512	0.32	0.33
	4k	2.60	2.65
	32k	20.81	20.81
	64k	41.63	40.99
MGCG Random Neighbor	512	0.30	0.32
	4k	2.14	2.39
	32k	14.41	14.41

Figure 12.1.4: Estimated sustained BG/L performance.