Blue Gene/L, a 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, 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, R.B. Tremaine, M. Tsao, P. Vranas, T.J.C. Ward, and M. Wazlowski
IBM TJ Watson Research
J. Brown, T. Liebsch, A. Schram, and G. Ulsh
IBM Manufacturing Division

Abstract

Large powerful networks coupled to state-of-the-art processors have traditionally dominated 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 efficiently execute on future supercomputers of any architecture must be coded in a manner consistent with an enormous degree of parallelism.

The BG/L program is developing a peak nominal 180 TFLOPS (360 TFLOPS 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 is expected in 2003. BG/L consists of 65,536 nodes, and contains five integrated networks: a 3D torus[4], a combining tree, a Gb Ethernet network, barrier/global interrupt network and JTAG.

The 3D torus interconnect is organized as 64x32x32 nodes. Every node is connected to 6 bi-directional torus links, each with an expected bandwidth of 350MB/s in each direction. For general communication between nodes, throughput and latency are optimized through adaptive, minimal path, virtual cut-through[3] routing. Two virtual channels provide fully dynamic adaptive routing for high throughput[2], 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 an expected 4usec hardware latency and 1.4 GB/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 Gbit 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. A midplane forms 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 consists of a single ASIC and 9 SDRAM-DDR memory chips, totaling 256MB. The ASIC uses IBM CMOS CU-11 0.13mm technology. Each of the two symmetric 700MHz PowerPC 440 cores delivers 2.8GFLOPS, although the normal mode of operation will dedicate one processor to message handling. The ASIC contains the network components and

Proceedings of the IEEE International Conference on Cluster Computing (CLUSTER'02)
0-7695-1745-5/02 $17.00 © 2002 **IEEE**

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 design coupled with a high density interconnect allows for 5.6TFLOPS peak performance in a single rack, which is anticipated to consume 15kW.

References

[1] QCDOC: A 10-teraflops scale computer for lattice QCD. *Nucl.Phys.Proc.Suppl.*, 94:825–832, 2001.

[2] W. J. Dally. Virtual-channel flow control. *IEEE Trans on Parallel and Distributed Systems*, 3(2):194–205, 1992.

[3] P. Kermani and L. Kleinrock. Virtual cut-through: A new computer communication switching technique. *Computer Networks*, 3:267–286, 1979.

[4] S. Scott and G. Thorson. The Cray T3E network: Adaptive routing in a high performance 3D torus. In *Proceedings of HOT Interconnects IV*, 1996.

Proceedings of the IEEE International Conference on Cluster Computing (CLUSTER'02)
0-7695-1745-5/02 $17.00 © 2002 **IEEE**