Trevor Bekolay, 6796723, umbekol0

Assignment 4, Question 4

Blue Gene/L Overview

I will discuss the Blue Gene/L architecture from a bottom up approach, starting at the smallest pieces and working my way up.

Compute Nodes

Each compute node is a self contained "system on a chip." Each chip is composed of two PowerPC 440 Processors (clocked at 700 MHz), two double precision floating point units (FPUs), a shared 4 MB EDRAM L3 cache, and a 1 Gb Ethernet connection. Unlike many supercomputers, the Blue Gene/L tries to keep these compute nodes relatively power efficient; each chip consumes only 10-13 watts.

Compute Node Networking

In addition to the 1 Gb Ethernet connection, every compute node is connected to two different networks. The first network is a 3-dimensional torus that connects together all of the compute nodes (but not the I/O nodes). Each compute node has 6 incoming and 6 outgoing connections, with aggregate bandwidth of 2.1 GB/s. The latency is between 1 and 10 µs, depending on how far away the other compute node is.

The second network is a tree that connects all compute nodes and I/O nodes. It's specialized to provide quick I/O. Each compute node has 3 incoming and 3 outgoing connections, with bandwidth of 2.8 Gb/s per link. The latency of a one way tree traversal is 2.5 µs.

I/O Nodes

The I/O nodes use the same "system on a chip" as the computer nodes. The difference is that the I/O nodes use their 1 Gb Ethernet connection to communicate with I/O devices; NAS, etc. The number of I/O nodes in a Blue Gene/L system can range from one per 128 compute nodes to one per 8 compute nodes.

Compute Cards

A compute card contains two compute nodes and up to 2 GB of RAM.

Node Cards

A node card contains 16 compute cards and up to two I/O cards (compute cards with I/O nodes).

Racks

A rack contains 32 node cards (1024 compute nodes).

A Blue Gene/L system may contain up to 64 racks.

Scalability

One of the main reasons that the Blue Gene/L system is so scalable is the relative simplicity found in each compute node. The entire system, with the exception of main memory (which is stored on a compute card), is stored on one chip. Not only does this make each compute node easy to assemble, it is easy to add new compute nodes as needed (in increments of 1024); when a node has to be replaced, one can replace the node without removing or moving any cables. Further, each compute node consumes relatively little power, so a system is easier to scale up, as the main cost is done initially, not over time with power consumption.

Another significant innovation that improves scalability is the Blue Gene/L reliability, available, security (RAS) mechanisms. In addition to compute and I/O nodes, the Blue Gene/L contains service nodes that perform system management services in a way transparent to the compute and I/O nodes. When a compute node fails, the application running on that node is restarting from a checkpoint on a different node. These and other mechanisms ensure that the system can handle an extremely large number of nodes.

General Purpose?

The Blue Gene/L system has been first on the list of the top 500 supercomputers worldwide; to be on this list, it must be a general purpose computer. Yet, its speed is always measure in FLOPS (floating point operations per second) despite the fact that most applications a normal user would encounter deal largely in integer operations. Of course, Blue Gene/L can do integer operations, but it is without a doubt specialized for floating point operations, which are used in most complicated scientific applications. Even IBM's own report (http://www.research.ibm.com/journal/rd/492/gara.html) notes that "we chose not to build a machine that would necessarily be appropriate for all applications." Certainly, the Blue Gene/L system would not be appropriate for a normal home user, however, it is general purpose in the context of scientific applications, which are what supercomputers are generally used for.

Sources: http://en.wikipedia.org/wiki/Blue_Gene ,
http://www.redbooks.ibm.com/redbooks/pdfs/sg247352.pdf ,
http://www.skatelescope.org/US_SKA_Technology_Day06/Liebsch.pdf ,
http://www.research.ibm.com/journal/rd/492/gara.html