The IBM Blue Gene/Q Interconnection Network and Message Unit

Pooja Nilangekar (poojan@umd.edu)
CMSC714

Overview

- Detailed description of IBM Blue Gene/Q network and message unit.
- Highly parallel Message Unit (MU) acts as a network interface.
- Network logic & MU are integrated consume only 8% of the chip’s area.
- Network is configured as a 5D torus.

Interconnection Network

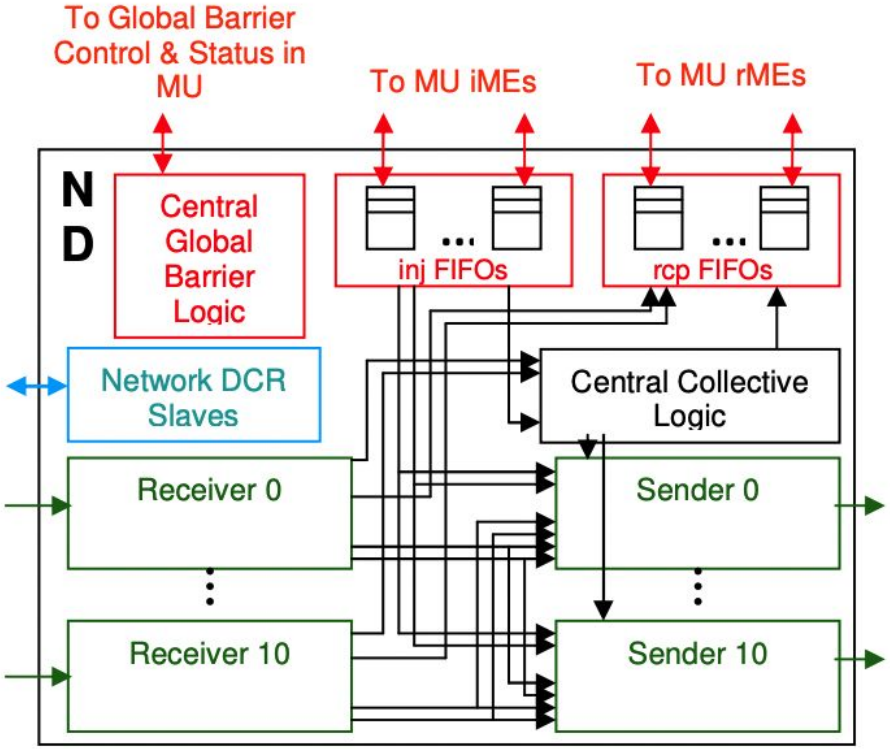

Figure 1. The BG/Q Network Device (ND) Router Logic

Why use a 5D torus

- High nearest neighbor bandwidth while increasing bisection bandwidth and reducing maximum number of hops.
- Permits partitioning a large machine into independent submachines; applications running in different partitions do not have to affect one another.
- The torus permits most links within the midplane to be electrical rather than optical, thus reducing cost.

Interconnection Network

- Virtual Channels for point-to-point and collective traffic.
- Performance improvements for point-to-point routing.
- Supports broadcast down a line of torus dimension using point-to-point virtual channels.
- BG/Q incorporates one pass double precision floating point sums.
- BG/Q supports collective operations over MPI sub communicators, provided they’re contiguous subrectangles in the torus.

Interconnection Network

- Barrier Support - initiated via writes to memory mapped I/O registers in MU and completion is detected by an MU memory mapped I/O read.
- Network Router Arbitration - The network router logic implements a distributed arbitration mechanism.
- Each sender broadcasts its link available and token available signals for each virtual channel to all receivers and injection FIFOs.
- Performance Counters, Protocols, RAS and Physical Design - On chip performance counters, Reed Solomon Codes, ECC. (only 3% of chip)
- Programmable chip link for partitioning.

Message Unit

- Provides low latency and high throughput, enough to keep all the links busy.
- Supports direct puts (RDMA write), remote gets (RDMA read) and memory FIFO messages.
- The MU has a multitude of engines; one injection Messaging Engine (iME) for each network injection FIFO and one reception Messaging Engine (rME) for each network reception FIFO. The iMEs and rMEs share multiple master ports on the BG/Q memory system crossbar switch.
- The MU has extensive logic and checks to separate user from system traffic, and to prevent user-space errors from interfering with system messaging.
- All internal buffers and data paths in the MU are ECC protected, providing very high resistance to soft errors.

Message Unit Logic

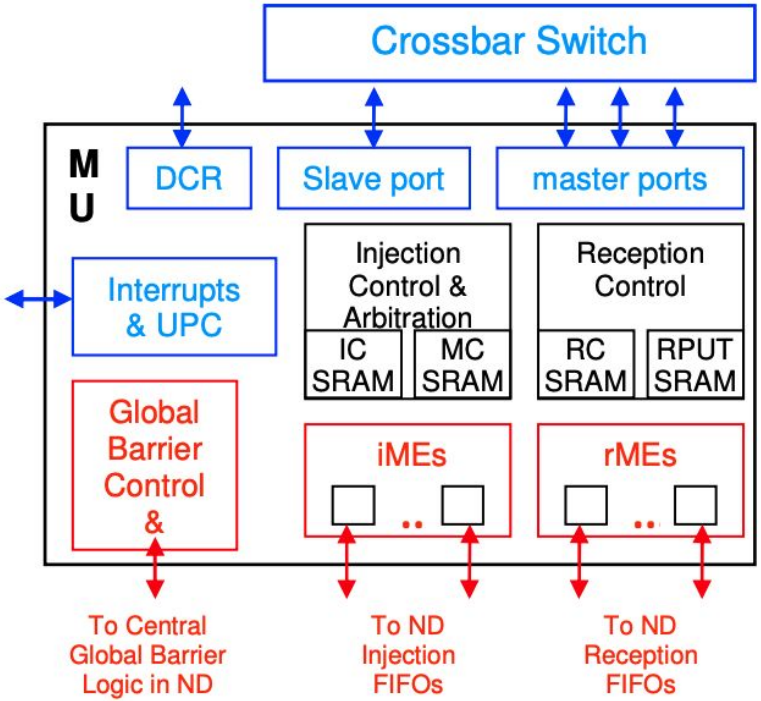

Figure 2. The BG/Q Message Unit (MU) Logic

Interface to Local Memory System

- Master Ports - All MU reads and writes pass through the globally shared L2 cache and the MU depends on the L2 to manage coherency across the node’s memory system. The master port bandwidth is sufficient to keep all network links simultaneously busy. The master ports are shared by requestors consisting of the iMEs, the rMEs, and the message descriptor fetch logic.
- Slave Port - The MU contains one slave port that communicates exclusively with the cores at 800 MHz via the crossbar switch. Software uses the slave to write and read the MU’s memory mapped IO (MMIO) registers and SRAM locations.

Messaging Unit

- The MU hardware processes each descriptor by splitting and packaging the message into network packets, and injecting those packets into the network injection FIFOs.
- The injection control logic arbitrates the next message descriptor to fetch for message injection into the network.
- Messages arrive at the MU in FIFO order. The MU receives the packets and writes them into the appropriate location in the memory system. There are three types of packets:
  - Memory FIFO - Raise an interrupt when the last packet is received.
  - RDMA write - directly write the payload to the memory system
  - RDMA read - generates RDMA write packets to transmit the data back to the sender.

Messaging Unit

- The MU's Device Control Register (DCR) unit provides a convenient way to access the configuration and interrupt registers.
- The MU also contains a universal performance counter (UPC) module that connects directly to a UPC ring. The UPC module provides useful performance counts.
- The global barrier (and interrupt) network is embedded in the BG/Q torus and is accessed via the MU.
- The MU has extensive logic to protect the system from user program errors.
- All internal buffers and data paths in the MU are ECC protected, providing very high resistance to soft errors. Most other latches are protected with parity for single bit error detection

Software Support

- Provides highly optimized Cinlines via the System's Programming Interface (SPI).
- The SPIs are a thin software abstraction layer of the BG/Q network and MU hardware.
- MU Kernel SPI
- Message Unit (MU) SPIs

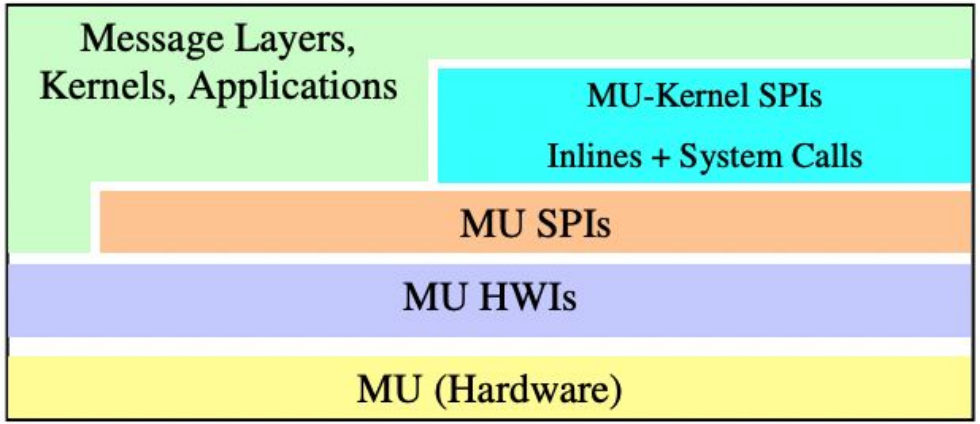

Figure 3. SPI and HWI components.

Performance Measurements

- Experiments on 512 BG/Q pass 2 prototype nodes.
- C language benchmarks such as ping-pong, nearest neighbor, broadcast and allreduce.
- Ping-pong benchmark - latency increases when number of hops increase.
- Two types of latencies - H/W (induced by network and MU), Software (injecting the descriptor, monitoring for completion)
- Nearest Neighbour throughput - efficiency of 88.4% of the raw link throughput and 98.3% of the peak 90% effective data utilization of the links.
- The achievable bandwidth is limited by the DDR memory interface instead of the MU and the network.

Performance Measurements

- All-to-All performance - On a 512-node prototype achieved 95% - 97% of peak theoretical performance.
- Collective operation performance - Due to more complex logic for collective operations the percent of peak is slightly lower for collective communication than point-to-point communication.