The Gemini System Interconnect

Robert Alverson, Duncan Roweth, Larry Kaplan Cray Inc

2010 18th IEEE Symposium on High Performance Interconnects

Introduction:

- Improvement to Seastar network for Cray HPCs
- System-on Chip (SoC) constructs 3D torus network > 100,000 nodes
- Built for fast MPI
- 2-node Opteron allowing for 10 connections per block
- Adaptive routing and ECC memory add layer of fault tolerance to prevent job termination in the event of limited hardware failure

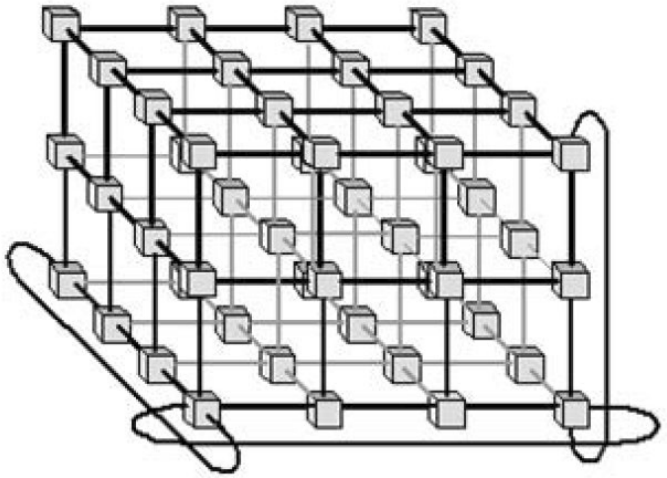
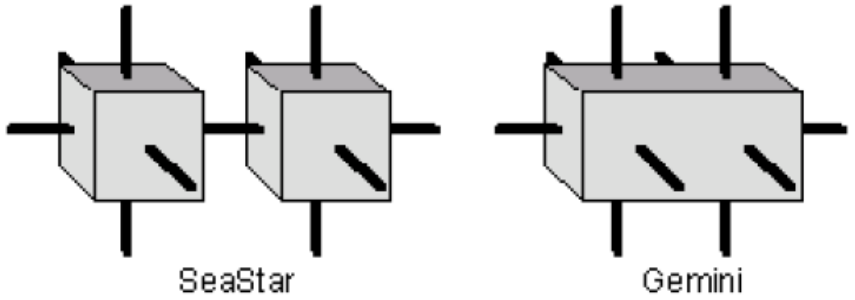


Figure 1: 3D Torus network



Gemini Block Layout:

- Each node has HyperTransport3 (5.2 MT/s) and dedicated NIC
- Each block contains a router and supervisor processor (L0) connected to Hardware Supervisory System (HSS)
- Router has 8 links to x/z and 4 links to y neighbors
- Direct data transfer between nodes without OS intervention (specify address, id, and size)
- Implemented in TSMC 90 process on a 232.8 mm^2 die size

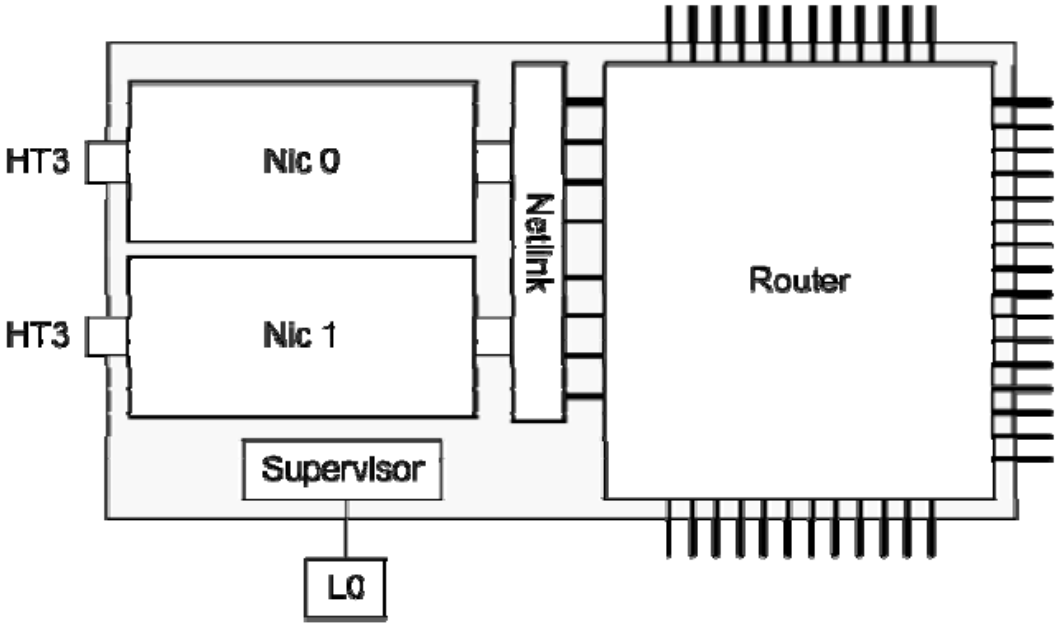


Figure 3: Gemini block structure

Gemini NIC Layout:

- Fast Memory Access (FMA)
  - Puts, Gets stored directly on NIC (64 bytes)
  - Translated from processor stores into full 58 bit network addresses
  - Features it's own sync/barrier methods
- Block Transfer Engine (BTE)
  - Asynchronous transfers between local and remote memory
  - No guarantee of order, but can use fence operations
  - Up to 4 GB w/out CPU involvement
- Completion Queue (CQ)
  - Notification mechanism for FMA and BTE
- Atomic Memory Operation (AMO)
  - Multiple processes accessing the same variables
  - Prevents program locking
  - Dedicated AMO cache reduces load on host NIC
- Synchronization Sequence Identification
  - Packet tracking system
  - Set of packets all have same SSID, can be delivered in any order
  - CQ isn't notified until all finish

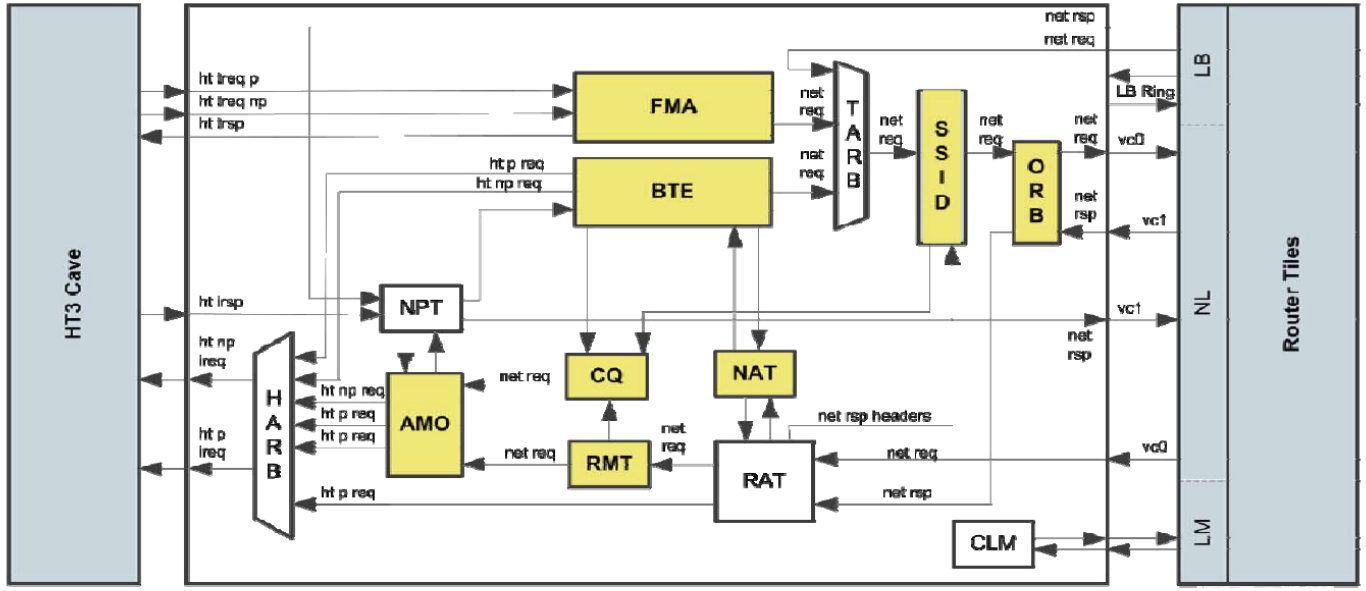
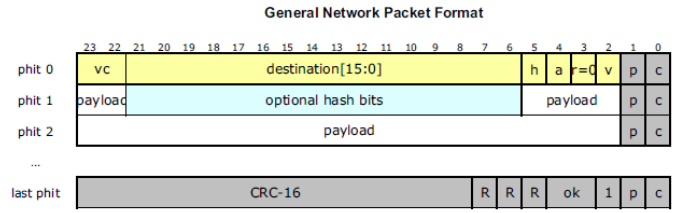
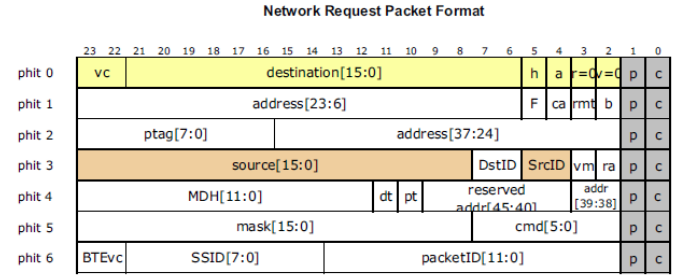


Figure 4: Gemini NIC

Gemini Router Layout:

- Following a Packet:
  - Packet arrives at input link
  - Input buffer makes routing decision to output column
  - Sent to row bus that contains intersection with column
  - Routing decision is refined to which output port
  - Column channel sends packet to output buffer
- Packets are divided into 24-bit “phits”
- Fault tolerance embedded in CRC protects 64 bytes and headers
- ECC memory protects larger data
- Data is automatically re-sent if CRC readback error
- If Gemini routes go offline system can re-route the network





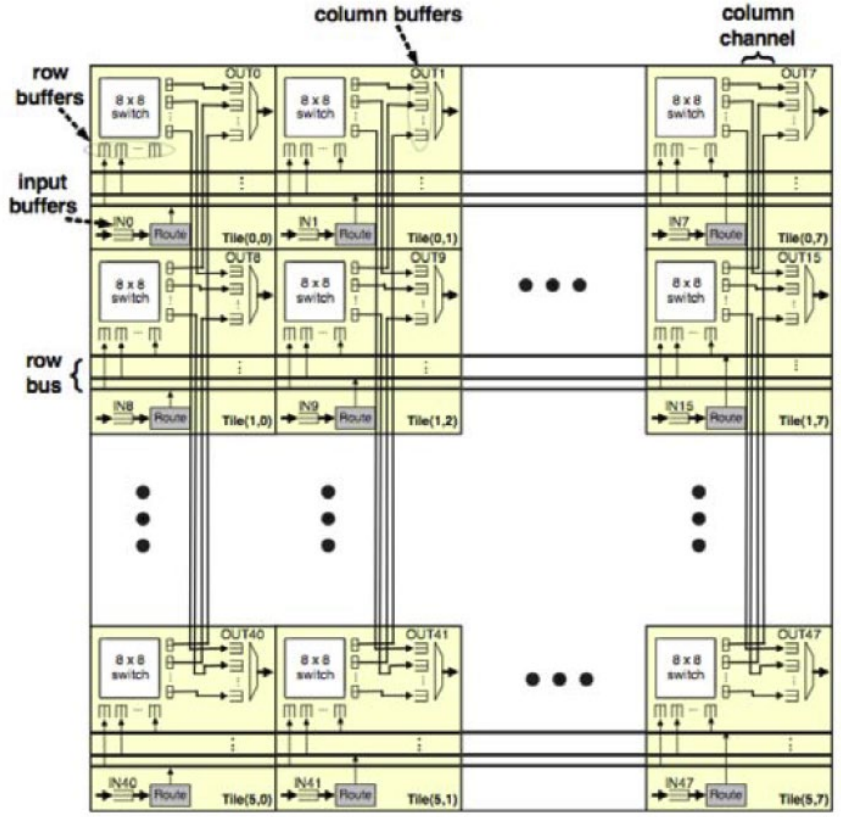


Figure 5: Gemini router

Performance:

- Clock Speeds
  - NIC 650 MHz
  - Router 800 MHz
  - SERDES 3.1 to 6.25 GHz
  - HyperTransport 1600 – 2600 MHz
- Latency
  - End-point 700 ns
  - 1.5 micro or less for small MPI (HyperTransport reads)
- Bandwith
  - NIC transfers 64 bytes every 5 cycles in each direction
  - 8.3 Gbytes/s
  - Improved bandwith as PPN increases
- AMO Performance
  - Atomic adds
  - Single AMO all performed on AMO cache
  - Achieved 45 – 100 million updates per second

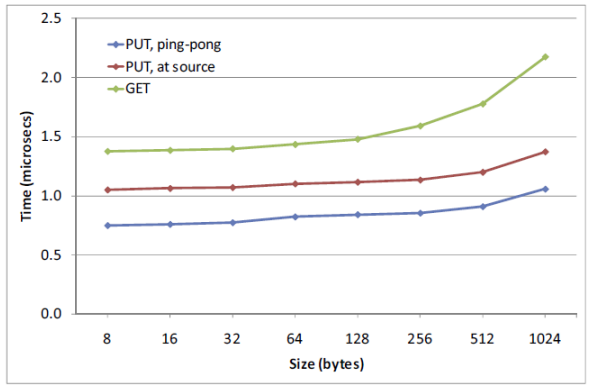


Figure 7; Gemini put and get latencies as a function of transfer size

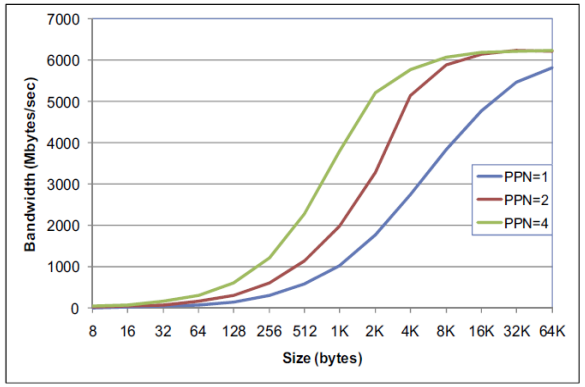


Figure 8: Gemini FMA put bandwidth as a function of transfer size for 1, 2 and 4 processes per node

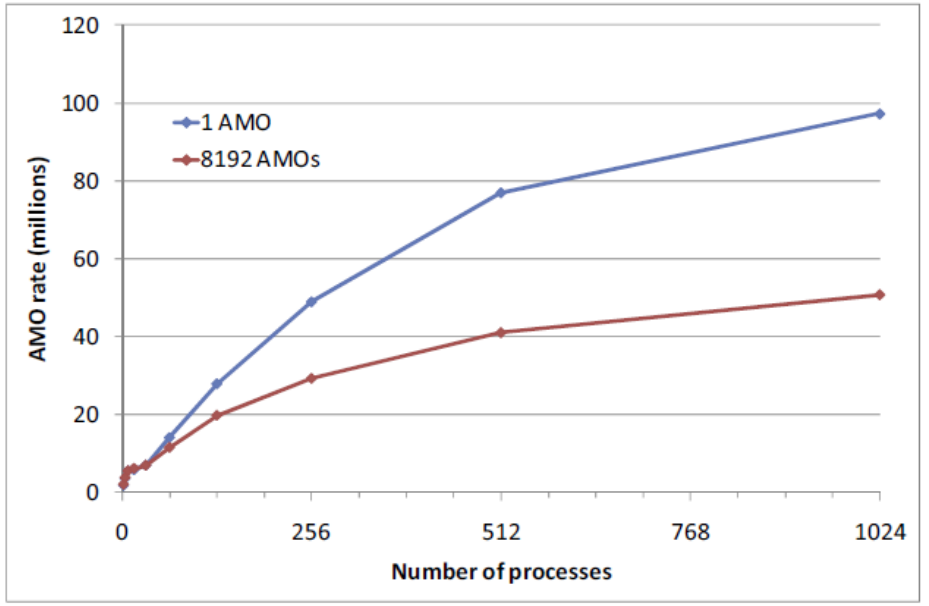


Figure 9; Gemini AMO performance