

CMSC 714  
Lecture 12  
High Performance Networks

Alan Sussman

# Notes

- OpenMP assignment due Monday
  - don't compile for profiling (with `-pg`) and for OpenMP at same time
  - For fine-grained timing, use **`omp_get_wtime()`** – works similarly to **`MPI_Wtime()`**
  - other questions?
- Readings for next week posted
  - Don't forget to send questions

# Infiniband

- Designed to support I/O and network connectivity, from a single PCB to a cluster network to a LAN
  - over copper (twisted pairs) and fiber
- Targeted at cluster networks, SANs, and even embedded systems
  - scalable, and provides RAS – “bandwidth out of the box”
  - idea is to extend the on-processor I/O bus to off-chip network
- Switched point-to-point I/O fabric
  - endpoints (host machines, I/O devices, ...) connect to switches, which route connections to other endpoints
  - link speed from 2.5Gb/sec (1X) to 30Gb/sec (12X) by adding more wires – parallel transfers – newer standards use higher link speeds for higher transfer rates
- Protocols described in terms of standard network layers
  - physical, link, network, transport

# Infiniband Layers

- Physical

- defines electrical and mechanical characteristics – cables, connectors, pins, etc.

- Link

- packet layout - management and data
- switching - uses local IDs in Local Route Header of a packet
- QoS through Virtual Lanes
- credit based flow control
- data integrity – error correction both for each link (VCRC) and end-to-end (ICRC)

- Network

- route packets across subnets – uses IPv6 addresses (128 bits) in Global Route Header of a packet

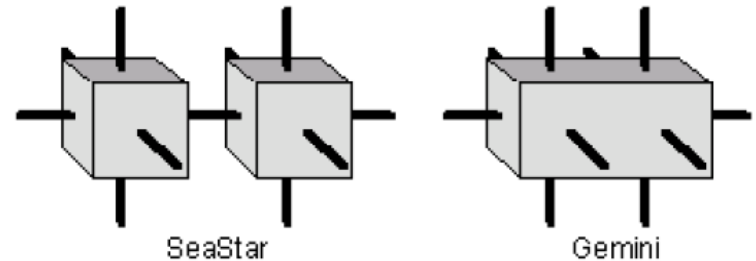
# Infiniband Layers (cont.)

- **Transport**

- in-order packet delivery – sequence numbers
- segmenting data into packets
- channel multiplexing
- transport services – reliable/unreliable connection/datagram
- all implemented in hardware

# Cray Gemini

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



# Gemini Block Structure

- Each node has HyperTransport3 connections (up to about 8GB/s) with a 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 2 links to y neighbors
- Direct data transfer between nodes without OS intervention (specify address, id, and size)

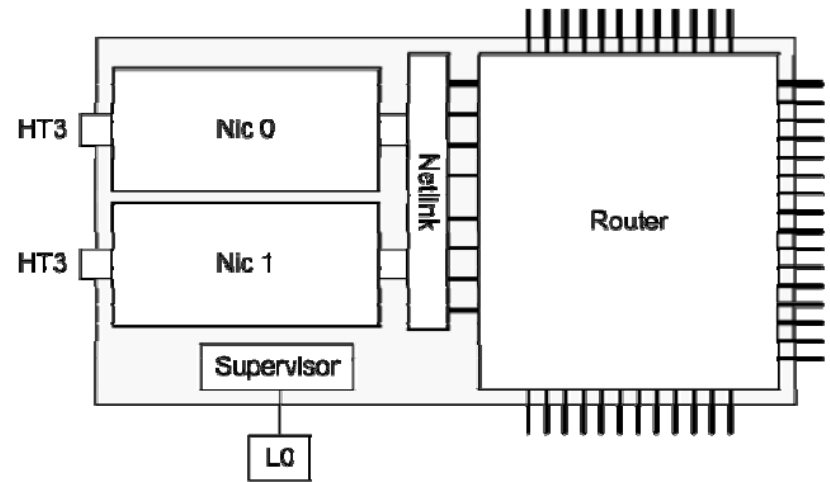


Figure 3: Gemini block structure

# Gemini NIC

- **Fast Memory Access (FMA)**
  - Puts, Gets done directly on NIC (up to 64 bytes)
  - Translated from processor loads/stores into full 58 bit network addresses
  - Has its own sync/barrier methods
- **Block Transfer Engine (BTE)**
  - Asynchronous transfers between local and remote memory
  - No guarantee of order, but can use fence operations for synchronization
  - Up to 4 GB w/o CPU involvement (after setup)
  - Higher bandwidth but also higher latency than FMA
- **Completion Queue (CQ)**
  - Notification mechanism for FMA and BTE
- **Atomic Memory Operation (AMO)**
  - Multiple processes accessing the same variables (e.g., atomic remote add, conditional swap, to build higher level collective and sync functions)
  - Prevents program locking
  - Dedicated AMO cache reduces load on host memory



# 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)

- **Bandwidth**

- NIC transfers 64 bytes every 5 cycles in each direction
- 8.3 Gbytes/s
- Improved bandwidth as PPN increases

- **AMO Performance**

- Atomic adds
- Single AMO all performed on AMO cache
- Achieved 45 – 100 million updates per second

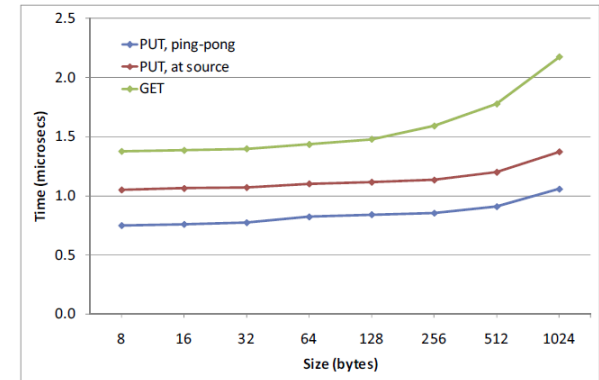


Figure 7; Gemini put and get latencies as a function of transfer size

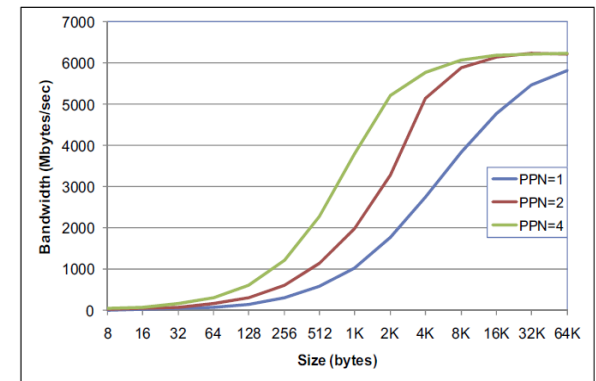


Figure 8: Gemini FMA put bandwidth as a function of transfer size for 1, 2 and 4 processes per node

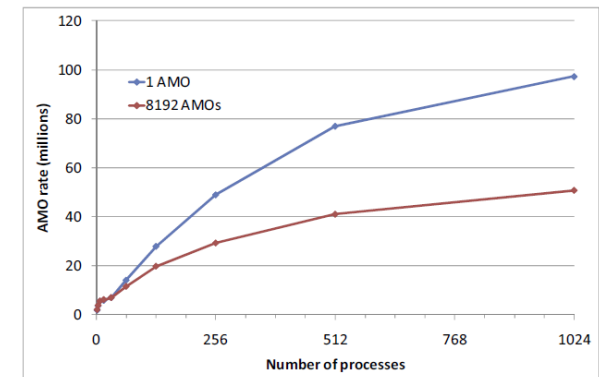


Figure 9; Gemini AMO performance