Programming Models Implications for Interconnects

Ron Brightwell
Scalable System Software
Computation, Computers, and Mathematics Center
Sandia National Laboratories
Albuquerque, NM

IAA Interconnects Workshop
July 21, 2008

Sandia is a Multiprogram Laboratory Operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy Under Contract DE-ACO4-94AL85000.

National Nuclear Security Administration

Sandia National Laboratories

Panelists

- Me
- Paul Hargrove, Lawrence Berkeley National Lab
- Vinod Tipparaju, Oak Ridge National Lab
- Scott Pakin, Los Alamos National Lab

Questions to Consider

- What critical capabilities are currently missing or insufficient?
- What are they key performance metrics?
- What important scalability limitations need to be addressed or overcome?
- How will the need for application reliability and resiliency impact the network?

Challenges in Supporting MPI

What critical capabilities are currently missing or insufficient?

- Scalable network resource management
  - Unexpected message buffering should be independent of number of peers
- Scalable and efficient flow control
  - MPI library should not have to do credit-based flow control
  - Don't penalize well-behaved applications
- Data movement should be independent of process/thread scheduling
  - Network should make progress on outstanding communications without being kicked
- Low-level network instrumentation
  - MPI profiling interface does not capture enough information
  - Hardware performance counters and the ability to map them to MPI
  - Timestamps
- Standard interface for exposing network routing/topology
  - Hierarchy is becoming increasingly important

What critical capabilities are currently missing or insufficient? (cont'd)

- Message completion notification
  - Polling memory host locations is inefficient and problematic
- OS/NIC memory management
  - Explicitly locking memory pages used for network transfers encourages bad behavior
- Meaningful benchmarks
  - Currently an oxymoron
- Non-contiguous transfers
  - Packing/unpacking is costly

What are they key performance metrics?

- Latency
  - With no send-side or receive-side copies
- Bandwidth
  - With independent progress
- Overhead
  - Efficiently overlap communication with computation and communication
- Small message rate
  - Without copies or message coalescing
- Collective communication performance
  - For a wide range of message sizes
- Bisection bandwidth
- System balance

How will the need for application reliability impact the network?

- Endpoint virtualization
  - Primitives for efficiently updating routing information on-the-fly
- Maintaining network state
  - Have the network be responsible for managing and recovering state

Improving MPI

- Enhanced collective operations
  - Non-blocking
    - Mitigate load imbalance
  - Persistent, early binding, channels, etc
    - More explicit resource management possibilities
- Application optimization hints
  - Mechanism to allow MPI library to do the right thing
  - Assuming somebody actually knows what that is
- Real one-sided data movement
  - Current RMA functions are complex and inefficient
  - More RDMA-like functionality?

IEEE Symposium on High-Performance Interconnects (Hot Interconects)

- August 27-28 at Stanford
- 14 regular research papers
  - On-chip interconnects
  - Memory subsystem interconnects
  - Routing and performance evaluation
  - Network processing
- 5 Industry research papers
  - HP, SiCortex, Quadrics, Myricom, Dawning
- Invited industry talks

http://www.hoti.org