High Performance Interconnection Networks for Cluster Computing

Ron Brightwell
Sandia National Laboratories
Albuquerque, NM
What are the future trends in high-performance networking and what are the implications of these trends?

- Large-scale clusters (> 1024 nodes) will be more common
  - Significant increase in components
  - More components requires more reliability, availability, serviceability (RAS) features in the network for predicting/detecting component failure
  - Performance at scale will become more important

- Latency bottleneck will move from I/O bus to software stack
  - Will drive the need for better one-sided communication model standard than MPI-2
Are quantitative measures of latency and bandwidth enough to characterize a network interconnect? What other ways should we be evaluating interconnects?

- Ability to overlap computation and communication (using MPI)
  - Efficiency of higher-level protocols
  - Collective communication performance
    - More effective hardware support for global operations
- Resource usage/utilization of the network stack
  - How much NIC memory or CPU did I use?
- Operating system interactions
  - Memory registration, validation/translation overhead
- Bit error rate
- Functionality (e.g., connectionless, ordered, etc.)
- How about application performance and scalability?
Will the “status quo” in networking continue?
Ethernet as the low-end solution, with IB, Quadrics, and Myrinet “relegated” to high-end and more costly clusters?

• Yes
  • Ethernet will continue as the low-end solution
  • Quadrics and Myrinet will continue to dominate the large-scale HPC cluster market
    – Vertical versus horizontal solution seems to be the differential
  • IB scares me
    – Too much complexity
    – Anticipated latency benefit may not be significant relevant to PCI Express, HyperTransport
What assumptions must interconnects make about the underlying architecture (or what assumptions would they like to make) PCI-X? PCI-Express, HyperTransport?

- Good question for the hardware vendors 😊
In five years, how will today’s interconnects evolve and/or compete in high-end computing?

- They will drive new user-level APIs
  - Better support for application-level RMDA operations
  - Integration of compute requirements with visualization and data mining requirements
- Performance may not be the primary issue
  - New functionality – compute offload capability
  - Reliability and RAS support
Will IB replace Myrinet or Quadrics as the costlier high-performance interconnect for high-end clusters?

• Maybe someday, but not anytime soon
• Software stack is targeted for data center and SAN's
  – Subnet management
    • Centralized subnet manager
    • Limited scalability
    • Not targeted for large-scale clusters
    • Assumes random topology
    • Connection establishment extremely too slow
  – Linux drivers
    • Multiple drivers for various IB layers (Verbs, IPoIB, VIPL, SDP)
    • Large memory footprint
• Clearly not targeting large-scale HPC use
What features and improvements are needed in communication subsystems to build next generation clusters? (network hardware, communication layers, libraries, programming models, etc.)

• Better RAS capability
• Performance monitoring API analogous to PAPI
• Lower latency one-sided standard than MPI-2 will be needed
• More complex network functionality will be needed for non-compute intensive applications like data mining, visualization, etc.

• Smarter users 😊
  – A significant amount of effort in delivering performance is wasted at the application-level