InfiniBand and OpenFabrics Successes and Future Requirements:

Matt Leininger, Ph.D.

Sandia National Laboratories
Scalable Computing R&D
Livermore, CA

25 September 2006
Outline

• DOE InfiniBand Clusters
• InfiniBand Cluster Experiences
• Early Experiences using OpenFabrics Enterprise Distribution
• Future Petascale InfiniBand Cluster Requirements
• How do we reach these goals?
DOE/ASC has Evaluated Several Generations of InfiniBand

2001-2002: Nitro I & II: IB blade reference designs (SNL) 2.2 GHz Xeon processors, small clusters, funded early MPI/IB work, and Cadillac (LANL) 128 node cluster

2003: Catalyst: 128 nodes 4X PCI-X IB (SNL), Blue Steel: 256 dual nodes 4X PCI-X (LANL), 96 nodes 4X PCI-X Viz Red RoSE (SNL)

2004: Catalyst: Added 85 nodes 4X PCIe IB, 288 port IB switch(SNL), ~300 nodes 4X PCIe Viz Red Rose (SNL)

2005: Thunderbird and Talon: 4,480 and 128 dual 3.6 Ghz nodes, 4X PCIe IB (SNL) Lustre/IB production @ SNL Red RoSE

2006: 2,048 + 1,500 nodes IB (LANL) 1,800 nodes (LLNL) SNL, LANL, LLNL have ~11,000 nodes of IB today and more to come
DOE Goals for InfiniBand

- To accelerate the development of an Linux IB software stack for HPC
  - High performance (high bandwidth, low latency, low CPU overhead)
  - Scalability
  - Robustness
  - Portability
  - Reliability
  - Manageability
- Single open source SW stack, diagnostic and management tools supported across multiple (i.e. all) system vendors
- Integrate IB SW stack into mainline Linux kernel at kernel.org
- Get stack into Linux distributions (RedHat, SuSE, etc.)

OpenFabrics was formed around these goals

DOE ASC PathForward program has been funding OpenFabrics development since early 2005
Sandia Thunderbird
Architecture

System Parameters
- 14.4 GF/s dual socket 3.6 GHz single core Intel SMP nodes DDR-2 400 SDRAM
- 50% blocking (2:1 oversubscription of InfiniBand fabric)
- ~300 InfiniBand switches to manage
- ~9,000 InfiniBand ports
- ~33,600 meters (or 21 miles) of 4X InfiniBand copper cables
- ~10,000 meters (or 6 miles) of copper Ethernet cables
- 26,880 1 GB DDR-2 400 SDRAM modules
- 1.8 MW of power, 400 tons of cooling
- Up to 2000 nodes Linpack efficiency was ~82%

#5 in Top500
38.2 Tflops on 3721 nodes
71% efficiency
LLNL Peloton Clusters

System Parameters
• Three clusters 1024, 512, and 256 nodes
• Quad socket dual core AMD Opterons
• 4X DDR PCIe InfiniBand – Full fat-tree
• DDR2 667 DRAM
• <3 us, 1.8 GB/s MPI latency and Bandwidth over IBA 4x
• Support 800 MB/s transfers to Archive over quad Jumbo Frame Gb-Enet and IBA links from each Login node. Some solutions have 10GbE
• No local disk, remote boot and SRP target for root and swap partitions on RAID5 device for improved RAS
• Software for build and acceptance 3 provided software stack distributions (all based on several common key components: RedHat, Lustre, OpenFabrics, MPICH or OpenMPI)
Sandia Thunderbird (4,480 nodes; 8,960 processors)
- Production computing resource (~1 yr)
- Running Cisco/Mellanox VAPI proprietary software stack
- CiscoSM and Cisco diagnostics
- MVAPICH1 and OpenMPI
- Starting to test OFED-1.0 with RHEL4U4 and OpenMPI

LLNL Peloton (~1024+512+256 nodes; ~14,400 processors)
- First large IB cluster build using only OFED software
- First cluster of 512 nodes is undergoing hardware and software “burned in”
- Using an OFED-1.0 based stack with RedHat
- Using OpenSM and OpenFabrics diagnostics/management tools
InfiniBand Successes

- Large IB clusters are running DOE production capacity computing workloads
- OFED-1.0 (and updates) available in RHEL4U4 and SLES10
- Several DOE clusters are now using OFED or have plans to upgrade to OFED
- OFED-1.0 RHEL4U4 is undergoing testing this week up to 1024 nodes (2048 processors)
- Stability and ease of use of OFED has been much better than proprietary stacks
- Sandia Thunderbird will likely move OFED into production with RHEL4U5
- OpenMPI is proving to be more scalable than MVAPICH
- OpenFabrics diagnostics and management tools provide good basic (static) information about the IB fabric
InfiniBand Technical Issues

- Current solutions are working but there is still room for much improvement

- InfiniBand Scalability
  - Current scaling is reasonable for small capacity jobs (< 256-512 nodes)
  - Need single job scalability up to 4000 nodes (multipath routing, etc.)

- Network congestion information is still difficult to extract from fabric
  - IB Congestion control architecture is not yet supported
  - Vendors will need to implement congestion control agents in switches

- OpenFabrics SRP and iSER are not production worthy yet, thus DOE production storage purchases have remained Fiber Channel rather than InfiniBand.

- Current OpenFabrics stack is missing a performance manager (PM)
  - Error rates are more helpful than raw errors numbers
  - PM could also provide information about fabric congestion
OpenFabrics Support Issues

- Vendors need to fully support OpenFabrics software in production environments
  - Stop pushing proprietary solutions (still happening)
  - There is no long term value add for customers in proprietary stacks
  - Production environments are multi-vendor (e.g. SNL has Voltaire, Cisco, Silverstorm, Mellanox, and Qlogic)
  - Make sure OpenSM supports your switches and any advanced features (performance manager, congestion manager, etc.)
  - Customers are willing to pay for OpenFabrics support to meet their performance, stability, robustness requirements
  - OFED is a reasonable start but we need vendors to stand behind the OFED product
OpenFabrics Enterprise Distribution (OFED)

• Goals to improve and control the quality of the OF software stack
  - Performance
  - Compliance
  - Stability
  - Reliability
  - Diagnostic tool set
  - Industrial-strength support and rigorous regression testing

• OFED has created a collaborative testing and productization environment
  - Single OpenFabrics software stack supported by all IB vendors and Linux distributions

• OFED is a good first step however ...
  - OFED process has been difficult
  - Collaboration is working but it is a constant struggle
  - Need a more robust and smoother release plan and process in order to succeed
  - Release process will only get more complicated with the addition of iWARP
  - Use experiences from the OFED 1.0 and 1.1 release to improve process
Petascale InfiniBand Cluster Requirements

- 4000 nodes with multi-core CPUs feasible in next 2-3 years
- Need InfiniBand to scale beyond the current 512 nodes to 4K nodes
- Scalability to this level will require:
  - < 1us pt2pt latency
  - High performance RC/UD send/recv (near line rate)
  - High message injection rate (15-20M/s) for small/medium sized messages
  - Hardware and OF software support for congestion control architecture
  - OF Performance manager and support in vendor switches
  - Fully adaptive routing (addition to IB spec.)
  - Cheap reliable fiber for 4X/12X DDR and QDR (match the cost of copper)
  - High performance (near line rate - SDR, DDR, QDR) native IB-IB routing
  - Reliable multicast (up to a minimum of 128 peers)
  - SRQ shared receive block (improve flow control for SRQ)
  - Reduce/eliminate memory registration overhead

- Achieving these goals will be a collaborative effort between OFA, IBTA, and HPC community
How can OpenFabrics and IBTA Collaborate?

- Complete IB-IB routing specification
- Add fully adaptive routing to InfiniBand specification
- Have joint memberships between IBTA and OpenFabrics?
- More joint workshops
- Continued discussions ...
How do we move forward?

• OFA is making good strides in the development and hardening of an OpenFabrics stack
  – Single multi-vendor software stack included in Linux distributions
• The use of OFED in a production environment is not 6 month out, it is today
  – Need stronger commitment from vendors (from CEO to field engineers)
• Streamlined OFED testing and release process
  – Perhaps using better web-based collaborative tools
• Continue to develop a strong collaboration between OFA, HPC customers, and IBTA
For more information

Matt Leininger mlleini@sandia.gov