OFA Technical Advisory Group

Lloyd Dickman

Monterey Workshop

April 4, 2011

Technical Advisory Group Members

- Lloyd Dickman, QLogic – Chair
- Diego Crupnicoff, Mellanox
- Parks Fields, LANL
- Michael Kagan, Mellanox
- Matt Leininger, LLNL
- Bill Magro, Intel
- Bob Pearson, System Fabric Works
- Bob Woodruff, Intel

Interviews

- **TACC**
- **NASA Ames**
- **NYSE Technologies**
- **Oracle**

Note: Comments received were generally InfiniBand transport centric.

OFED Packaging - #1 Tactical Concern

- End-users receive OFED technology via multiple methods:
  - Use a standard Linux distro (e.g., RH, SuSE, …)
  - Use an integration supplier's distro (e.g., ClusterCorp, …)
  - Use a vendor's distro (e.g. Oracle, QLogic, Mellanox, Voltaire)
  - Develop a site-specific stack composed of OFED directly downloaded from OpenFabrics.

- Linux distro vendors have differing preferences
  - RedHat – Pull kernel code from kernel.org and user code from OpenFabrics. No OFED kernel patches. Pick and choose features to include.
  - SuSE – Pull complete OFED release from OpenFabrics.

OFED Packaging - Requests

- Prefer a conventional Linux community RPM mechanism – current scripts touch many unrelated packages and have numerous interdependencies
- Offer OFED components ala carte
  - System manager can mix and match capabilities to site specific requirements. Customers want to build site-specific stacks and may use different distributions may be used for various cluster functions – e.g., Lustre, MPI, …
  - Upgrade portions of the system separately, rather than install an entirely new OFED system release. → Minimize OFED component interdependencies.
- Support additional distro's (e.g., need Real Time Linux kernels)
- OpenFabrics to push all features upstream (e.g., RoCE)
- Process to certify/validate updated drivers and firmware in between OFED releases

Support

- OpenFabrics needs to do a better job at disseminating knowledge across the community. The email reflectors are a start, however , it was strongly advocated that a “knowledge base” be introduced as part of the new website. This could be implemented as a wiki.

Broaden User Base

- Existing APIs and ULPs limit wider adoption. Increase the constituencies that benefit from OFED capabilities.
  - Easier API for end-user applications. E.g., simplify memory registration.
  - Increase middleware adoption to address the needs of new markets
    - E.g., memcached as a new OFED ULP for a broad range of markets
    - E.g., AMQP for the financial services market
  - Consider providing an ultra performance sockets capability, something beyond IPoIB or SDP. May not need to use the TCP/IP stack, nor be absolutely Berkeley sockets compliant. E.g., async sockets.

Resiliency: Commercial and HPC

- APM – Is it fully supported by all ULPs and connection manager? RDMA-CM scaling?

- Explore methods for subnet manager to report back into applications – e.g., upcoming MPI3 call-back mechanism

Large Fabrics

- Support Large Subnets
  - Increase address space
    - OK for RoCE and iWARP
    - InfiniBand needs to extend
- Large Fabrics Require
  - Distributed subnet management
  - Distributed connection manager
  - Parallelized tools
  - Tolerance to frequent change
- Support Multiple Subnets
  - A single, flat network is not operationally manageable
  - Route subnets independently

Fabric Configuration and Management

- Broaden OFED Acceptance
  - OFED management should easily **integrate into existing management frameworks** from various vendors.
- Support Increasingly Large InfiniBand Fabrics: Several suggestions
  - **Rapid Fabric Reconfiguration.** In such a large system, switch changes are highly disruptive taking many minutes for the fabric to reconfigure itself. Objective would be to have this take under 10 seconds, and ideally make it transparent to the application such that few fabrics errors bubble up the application/user.
  - **VL15 Congestion.** Large fabrics inject substantial numbers of MAD packets into the subnet managers which cause end-point congestion spreading into the fabric. VL15 traffic is very bursty and targets a small number of end-points. Consider ways to reduce the amount of VL15 traffic as well as to more evenly distribute routes involving MAD packets. Dropped VL15 traffic is troublesome and expensive. Suggest moving more functions to reliable and flow-controlled connections.
  - **Topology Awareness.** A manual process today. Important for mesh/torus topologies. To provide support, OFED will need to add a matrix describing relative communication costs for resource managers to reference. Ibnetdiscover currently provides the basic topology information from which this matrix can be constructed.

Tools

- Provide visual tools for operations staff to more easily mange large fabrics
- Provide CCA configuration and monitoring tools
- Provide tools to validate cluster correctness

Virtualization

- Need SR-IOV support
- Include virtual environments in the list of supported OSs

Proposed Roadmap – How to Enable?

- OpenFabrics community is volunteer driven
  - Is the self-interest of community members sufficient to evolve OFA technology?
- Funding models to consider
  - Can OFA invest own funds to evolve technology?
  - Should OFA seek external funding?

OPENFABRICS ALLIANCE

Thank you