HPC Customer Requirements for OpenFabrics Software

Matt Leininger, Ph.D.

Sandia National Laboratories
Scalable Computing R&D
Livermore, CA

16 November 2006

I'll focus on software requirements (well maybe)

The HPC community has many hardware requirements
see presentations from Sonoma Workshop and joint
OFA-IBTA workshop (http://openfabrics.org/conference.html)

DOE Goals for InfiniBand

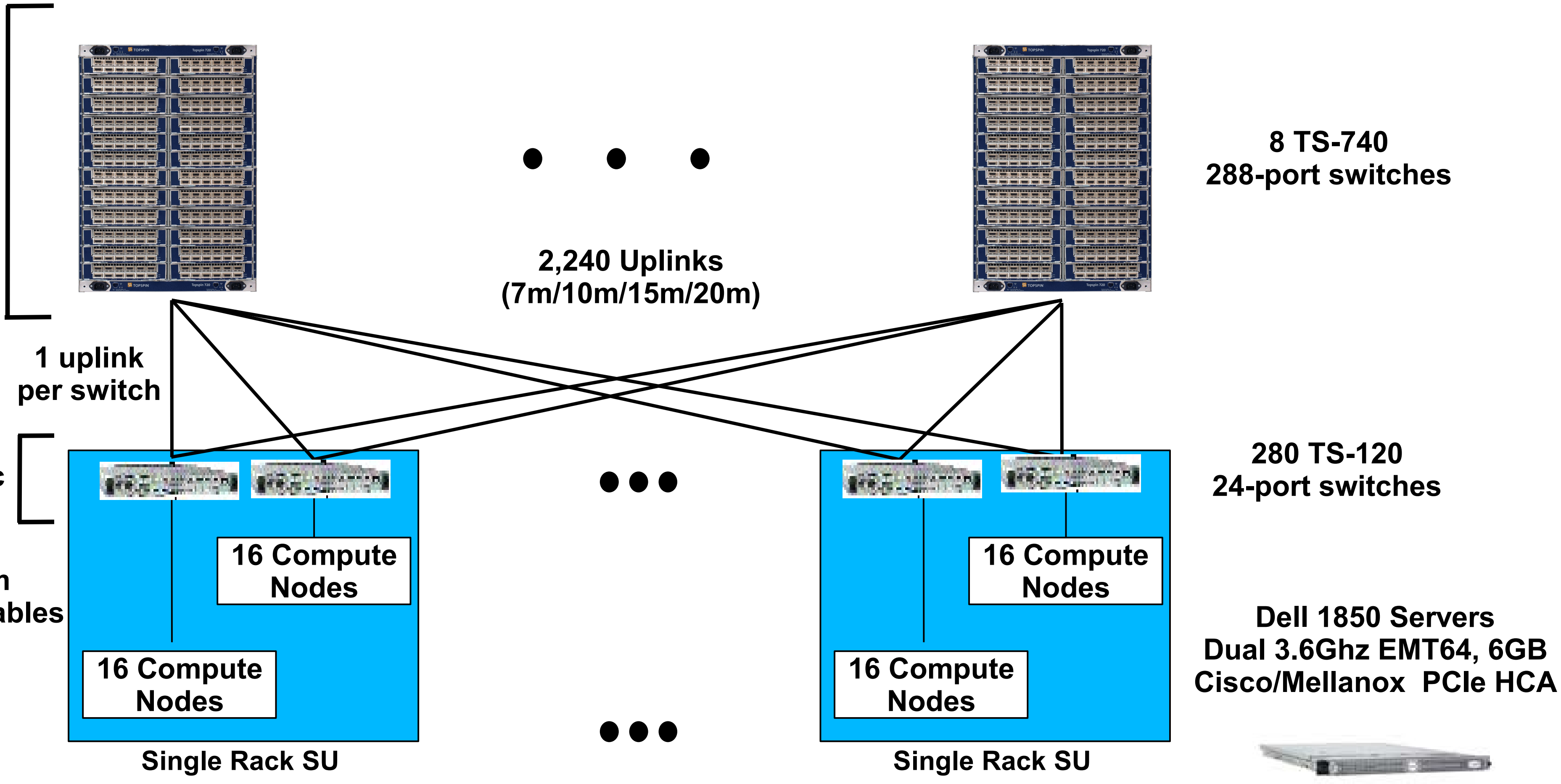
- To accelerate the development of an Linux IB software stack for HPC
  - High performance (high bandwidth, low latency, low CPU overhead)
  - Scalability
  - Robustness
  - Portability
  - Reliability
  - Manageability
  - Single open source SW stack, diagnostic and management tools supported across 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 Cluster

8,960 Processor, 65TF/s



Core Fabric

8 TS-740
288-port switches

2,240 Uplinks
(7m/10m/15m/20m)

1 uplink
per switch

Edge Fabric

280 TS-120
24-port switches

4,480 1m
intra-rack cables

16 Compute Nodes

16 Compute Nodes

Single Rack SU

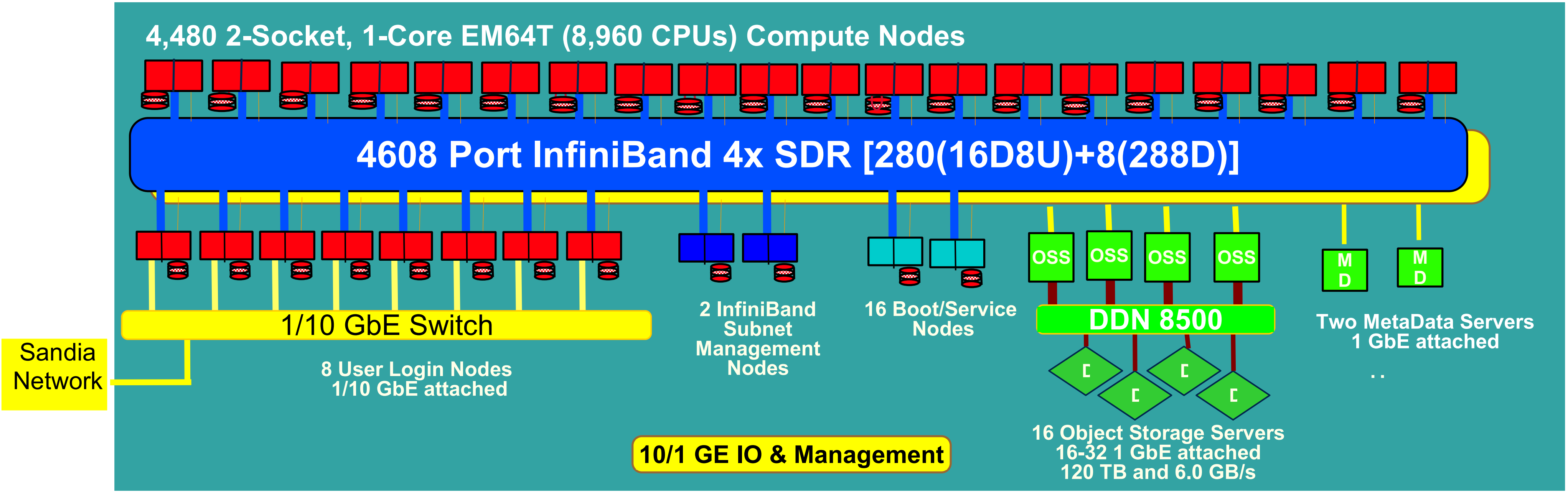
16 Compute Nodes

16 Compute Nodes

Single Rack SU

Dell 1850 Servers
Dual 3.6Ghz EMT64, 6GB
Cisco/Mellanox PCIe HCA

Sandia Thunderbird Architecture


4,480 2-Socket, 1-Core EM64T (8,960 CPUs) Compute Nodes
4608 Port InfiniBand 4x SDR [280(16D8U)+8(288D)]
OSS
OSS
OSS
OSS
M D
M D
1/10 GbE Switch
DDN 8500
Sandia Network
8 User Login Nodes 1/10 GbE attached
2 InfiniBand Subnet Management Nodes
16 Boot/Service Nodes
Two MetaData Servers 1 GbE attached
16 Object Storage Servers 16-32 1 GbE attached 120 TB and 6.0 GB/s
10/1 GE IO & Management

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

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

Thunderbird Linpack

Nodes	Stack	Runtime	Memory	Result	Efficiency	Date
3721	MVAPICH,VAPI	7.35 hrs	68%	38.27 TF	71.42%	2005
4347	OpenMPI,OFED	6.72 hrs	65%	52.57 TF	83.98%	2006
4347	OpenMPI,OFED	8.37 hrs	70%	52.71 TF	84.20%	2006
4347	OpenMPI,OFED	9.44 hrs	73%	53.00 TF	84.66%	2006

The efficiencies at large scale were possible because of

- OpenFabrics (OFED 1.0)
- OpenMPI 1.1.2
- Memfree HCA firmware
- Stunt mode Linux (no RAID, no HD, no IPMI, no PFS, no random daemons)

Thunderbird Infiniband Software

- Sandia Thunderbird Production Computing (4,480 nodes; 8,960 processors)
  - Past year
    - RHEL4
    - Running Cisco/Mellanox VAPI proprietary software stack
    - CiscoSM and Cisco diagnostics
    - MVAPICH1
  - Currently upgrading production environment
    - OFED v1.0/1.1
    - OpenMPI v1.1.2 or v1.2, MVAPICH 0.9.7/0.9.8
    - RHEL4U4
    - OpenFabrics management and diagnostic tools
- LLNL Peloton Production Computing (1,100+570+280; ~14,000 processors)
  - OFED v1.1, MVAPICH1 and OpenMPI, RHEL4U3, OFA management and diag. tools

SNL, LANL, LLNL has more than 12,000 InfiniBand nodes
and continue to deploy more clusters

OpenFabrics Support Issues

- Vendors need to fully support OpenFabrics software in production environments
  - Production and R&D 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
  - Customers need the ability to track changes in OF stack and customize (an OFED release) for their computing environment and requirements
  - The OFED build/patch scripts make it overly difficult to develop site customized OFED-based stacks

MPI Requirements

- High message injection rates (10-15M/s) today – 20M/s ++ in near future
- User-space multicast, atomic, gather/scatter API for use in optimized collectives
- Reliable multicast would be great (ok, this is HW)
- More vendors working on and contributing to Open-MPI
- HPC Optimized routing
- Multiple HCA per node (multi-rail)
- Optimized MPI datatypes – noncontiguous data transfers
- Latency target of < 1us pt2pt through single switch
  - Need SW fast path to be "really fast"
- Topology information to MPI to optimize communications
- Thread-safe MPI (MPI_THREAD_MULTIPLE)
- Performance and logistic improvements for memory registration
  - See Pete Wyckoff's (OSC) paper
- Improved flow control for SRQ – shared received block

Issues with SRQ

- SRQ allows multiple QPs to use one pool of (fixed-size) buffers
  - Avoids the complexity of figuring out how to distribute buffers across QPs
  - Doesn't help determine
    - Appropriate number of buffers
    - Appropriate size of buffers
    - Rate at which buffers should be replenished
- These are highly dependent on application behavior
    - Application message passing patterns are not fixed
    - Dependent on input data, number of processes, application algorithm, etc.
- Requires complexity inside MPI implementation to always insure that buffers will be available

Proposal: Shared Receive Block

- SRQ without individual message buffer boundaries
- Use a large block of memory for incoming messages
- Messages flow one right after the other into the block
- A running offset allows the next message to start where the previous message ended
  - May want to align start of message
- Essentially a first-fit block allocation for incoming messages
- Portals uses this type of strategy to deal with MPI unexpected messages

Benefits of This Approach

- Efficient use of memory dedicated to the network
  - Small messages don't waste memory
  - An 8 KB buffer isn't consumed by an 8-byte message
- Larger block of memory reduces the rate at which memory resources are consumed
- Significantly reduces the need for complex user-level strategies that try to insure message buffers are always available
- Locking down and translating a large block may be more efficient than more dynamic strategies

Fabric Mangagement Requirements

- DOE HPC is starting to move away from proprietary SM's and tools
- OpenSM has scaled to 4,500 node of Sandia Thunderbird
- Performance manager
  - Sysadmins want to know how many errors on link over past hour, etc.
- User-space API for congestion control parameters and fabric info
- HPC optimized routing algorithms in OpenSM
- QoS support
  - Partition fabric for compute, I/O, visualization, etc.
  - User-space API to interact with QoS settings
- CLI for OpenSM
- Daemon mode for OpenSM – may need some cleaning up

I/O Requirements

- Improved support and features for SRP (already in the works)
  - Labs (SNL, LLNL, ORNL) are willing to help test
  - Seeing 525MB/s with OFA SRP – got 600MB/s with IBGD on same HW
  - Immediate improvements to SRP will directly impact DOE storage solutions
- Have SRP now but need to move to network agnostic ULPs
  - iSER/iSCSI high performance, robust, feature rich open source initiator
- Improved IP performance
  - Getting by with IPoIB (UD) now
  - IPoIB Connected mode needed
  - SDP enhancements and easy of use
- Common RDMA stack for InfiniBand and Ethernet
  - OpenFabrics is moving in this direction – more RDMA Ethernet companies need to contribute code to OFA
- DOE Labs already have institutional parallel file systems
  - IB-IB routing and IB-IP routing
- Lustre, GPFS, Panasas, PVFS, pNFS, etc.

Virtualization Requirements

- Growing interest in Xen virtualization
- Xen + low over head RDMA features is causing us to rethink where we can use virtualization
- Application specific OS using Xen
- And other areas.....

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
- Streamlined OFED testing and release process
- Facilities to test, validate, and evaluate developer snapshots/releases
- Continue to develop a strong collaboration between OFA, HPC customers, and IBTA

Can't Resist Talking About Hardware

Petascale InfiniBand Cluster Requirements

- Sandia Cplant, LLNL MCR, and LANL Pink Clusters (1500-2000 processors)
  - Commodity HW, high speed interconnect, Linux, and other open source SW
  - Showed that is was possible to bring scientific computing to the masses
- Sandia Thunderbird
  - Pushing scientific simulations
  - Scalability, InfiniBand, OpenFabrics, OpenMPI, and Linux to 4000 nodes
- Future – Petascale InfiniBand Linux clusters
  - 4000 nodes with multi-core CPUs feasible in next 2-3 years
- Scalability to this level will require:
  - < 1us pt2pt latency and increased message injection rate (15-20M/s) for small messages
  - Hardware and OF software support for congestion control architecture
  - 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)
  - More requirements presented at Sonoma Workshop and joint OFA-IBTA workshop (http://openfabrics.org/conference.html)

Achieving these goals will be a collaborative effort between OFA, IBTA, and HPC community

For more information

Matt Leininger
mlleini@sandia.gov
leininger2@llnl.gov