True Scale and Parallel File Systems over RDMA

OpenFabrics Software User Group Workshop

Tom Elken, Intel Corporation

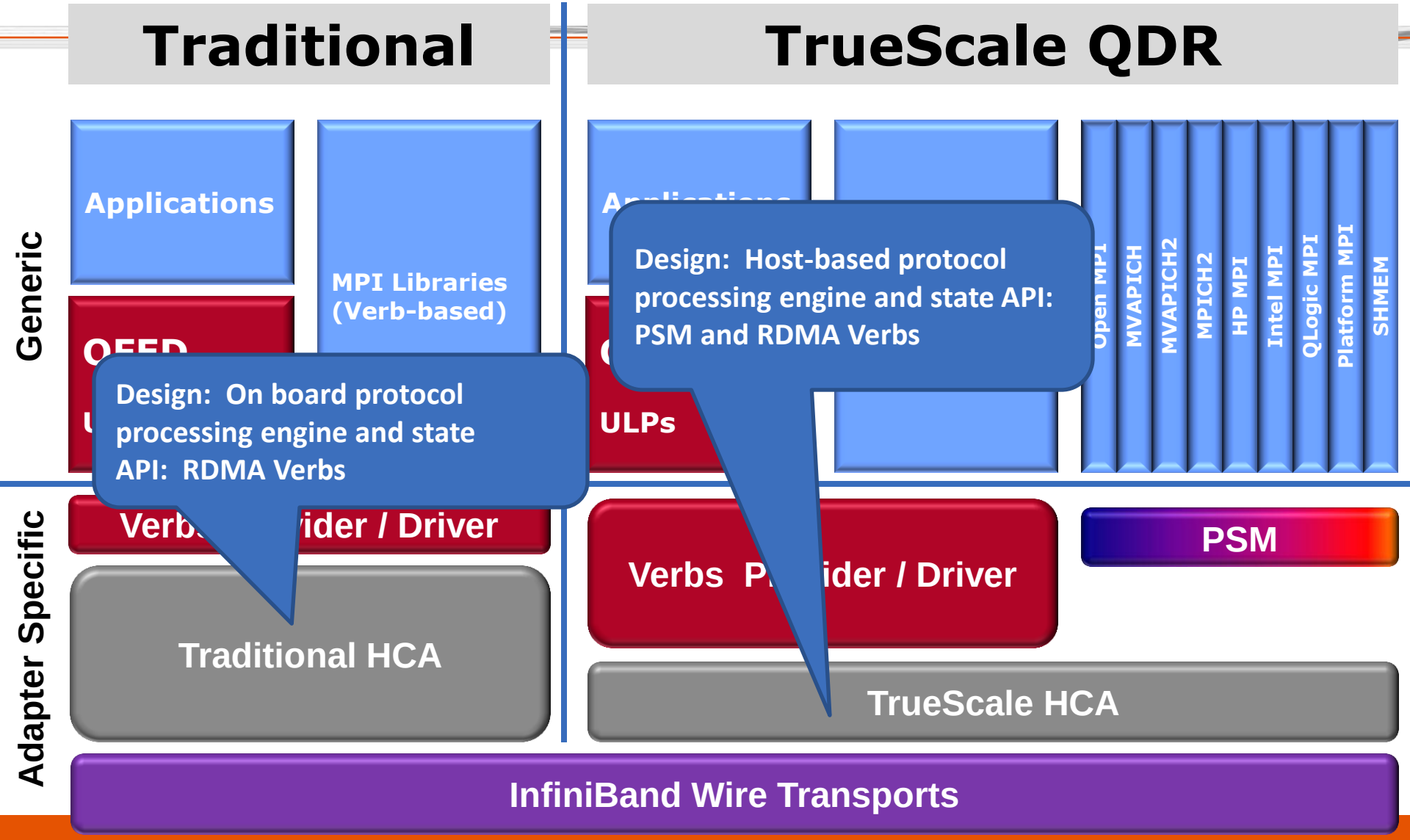
#OFSUserGroup

Overview

- True Scale architectural roots: PSM and verbs
- Parallel File Systems using RDMA
  - Lustre
  - GPFS
- Customer examples
  - Fermilabs, Batavia, IL
  - McGill University, Canada
  - Crihan, France
  - Vestas, Denmark

Comparison of Architectural Approaches

Traditional vs. TrueScale QDR



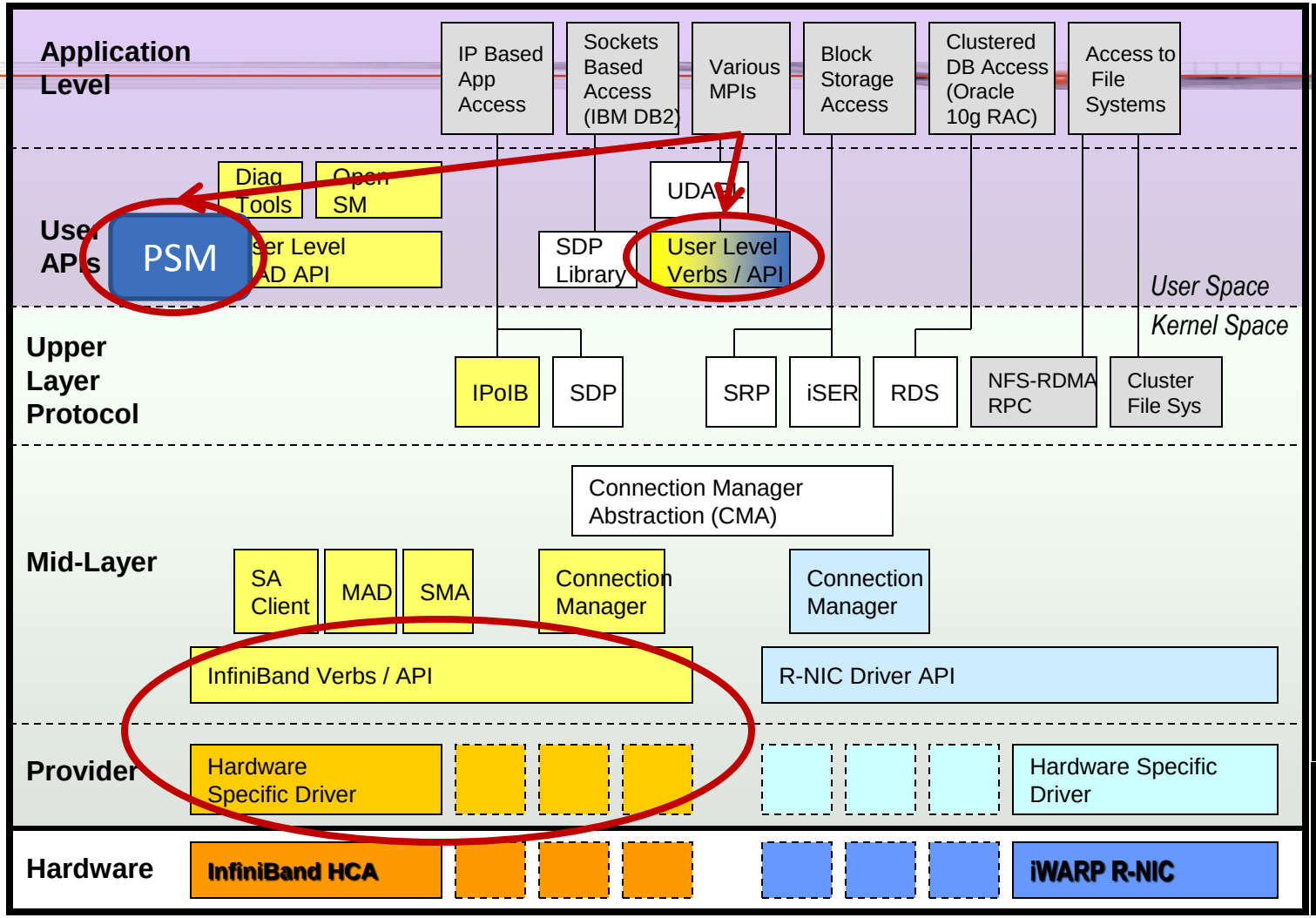
What is PSM?

- PSM = Performance Scaled Messaging
- API designed specifically for HPC
  - Analyzed needs of various MPI Channel Interfaces
  - Designed an API to match the needs of those Interfaces
- Carefully selected division of responsibility
  - MPI handles higher level capabilities
    - The wide array of MPI functions and user-facing functionality, etc
  - PSM focused on interconnect specific details
    - Data movement strategies and tuning
    - Advanced features – QoS, dispersive routing, resiliency, etc
  - All MPIs ported to PSM can take advantage of PSM advanced features
- Implemented as a user space library
- Open source version of PSM included in OFED (since 1.5.2)

qib driver, implements verbs

- Integrated into standard Linux OFED stack
  - Hardware specific driver with support for True Scale QLE72XX and QLE73XX adapters
  - Driver is qib -- named for “QLogic IB” -- implements IB verbs for Intel True Scale
  - driver module: ib_qib.ko
  - HCAs named, e.g.: qib0, qib1
  - As with PSM, verbs is on-load; protocol processing is on host CPU cores

Software Environment is Linux/OFED



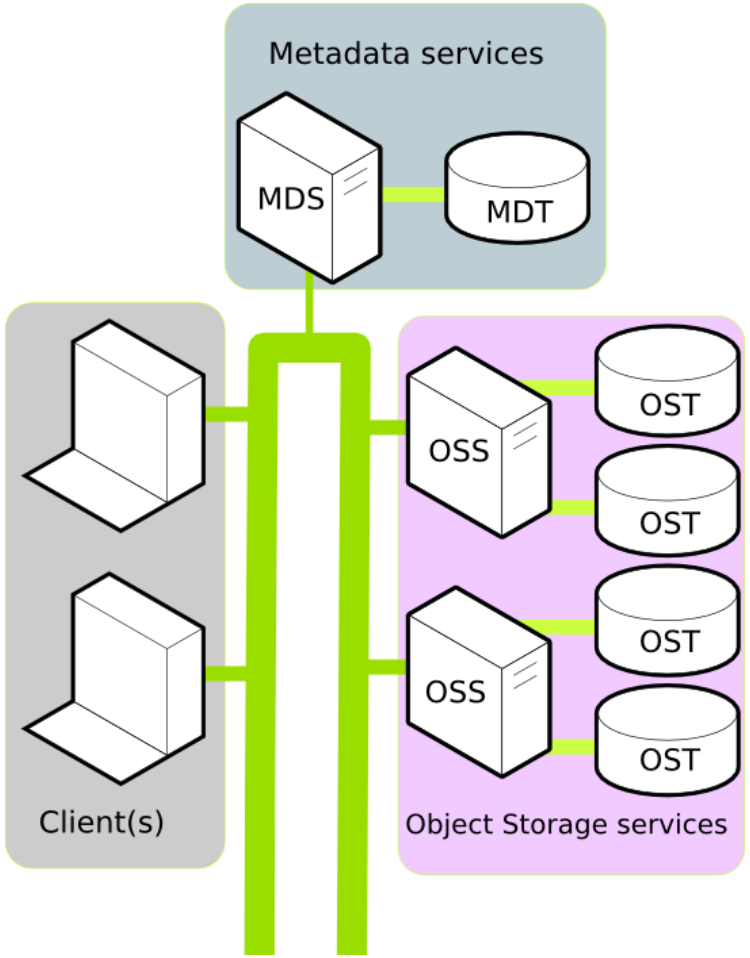
SA	Subnet Administrator
MAD	Management Datagram
SMA	Subnet Manager Agent
PMA	Performance Manager Agent
IPoIB	IP over InfiniBand
SDP	Sockets Direct Protocol
SRP	SCSI RDMA Protocol (Initiator)
iSER	iSCSI RDMA Protocol (Initiator)
RDS	Reliable Datagram Service
UDAPL	User Direct Access Programming Lib
HCA	Host Channel Adapter
R-NIC	RDMA NIC

Key: Common; InfiniBand; iWARP; Apps & Access Methods for using OF Stack

Lustre

- Clustered File System
- Only runs on Linux
- Open Source: GPLv2
- Community represented by OpenSFS

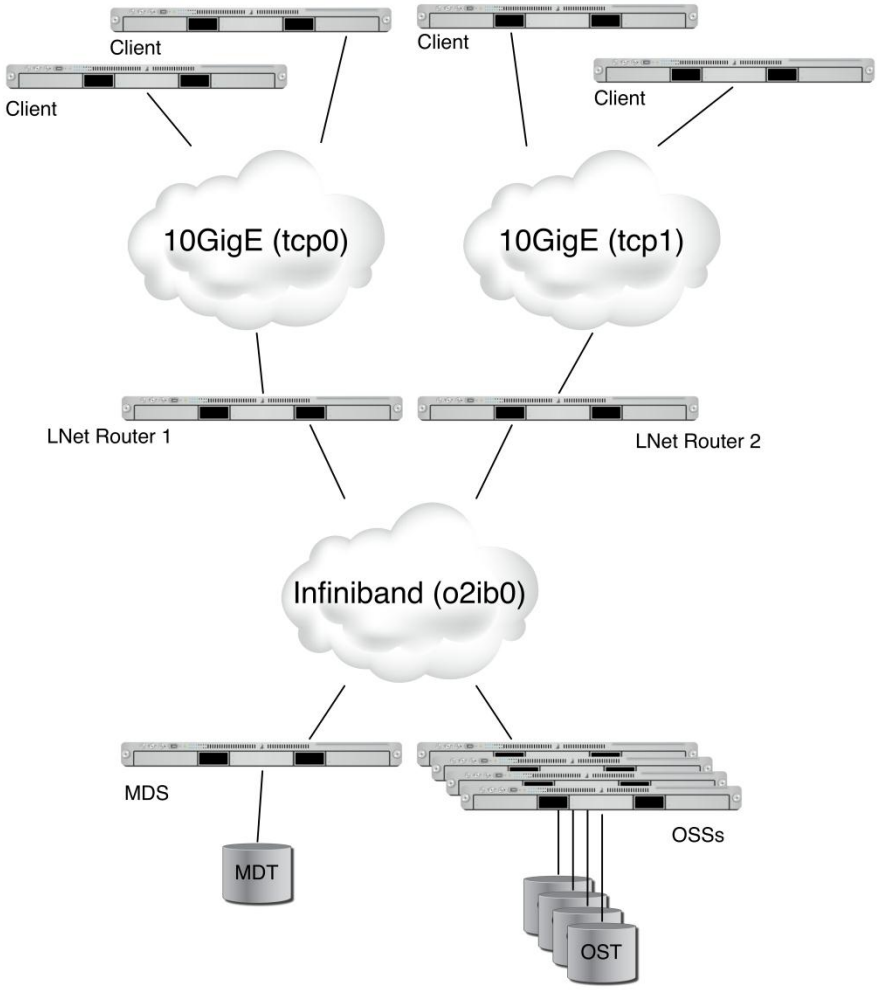
Lustre and LNet



Node-specific Lustre code
ptlrpc
LNet
LND (o2iblnd)
OFED

Lustre with Multiple Subnets/Fabrics

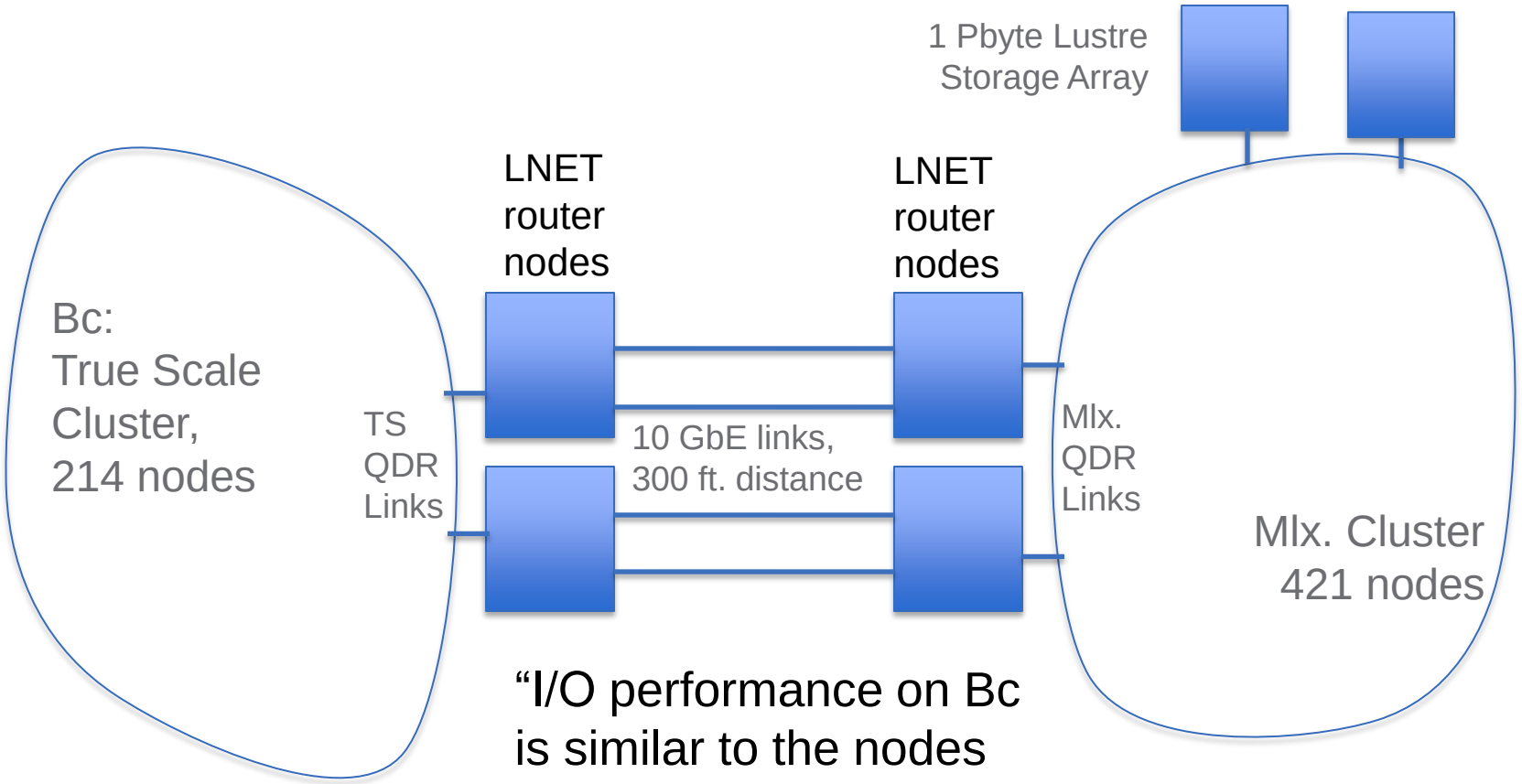
- LNet supports running as an “LNet Router”
- This allows for bridging LNet packets between different subnets
- Subnets can be different fabrics (i.e. Ethernet and Infiniband)



Fermi Labs

- Bc ="B-sub-C“ Cluster uses True Scale QDR InfiniBand
- 214 compute nodes, also Lustre clients
  - 32 AMD cores (4 sockets x 2.8 GHz 6320 CPUs) per node
  - 32 GB of system memory.
- Application: LQCD (lattice quantum chromodynamics) calculations.
- All jobs are parallel, using MPI (Open MPI, MVAPICH, or MVAPICH2) over InfiniBand.
- Operate a ~ 1 PByte Lustre filesystem that is accessed by roughly 900 clients.

Access to the Lustre File System at Fermilab



"I/O performance on Bc is similar to the nodes on the other side of the routers."

Intel testing of Lustre on True Scale by Lustre team

- What We Did
  - Tested the performance of the verbs implementation in OFED+ 7.2.x with the Lustre Networking Layer in Lustre 2.1.4
  - Test the performance of the verbs implementation in the OFED provided by RHEL 6.4

Low level RDMA performance

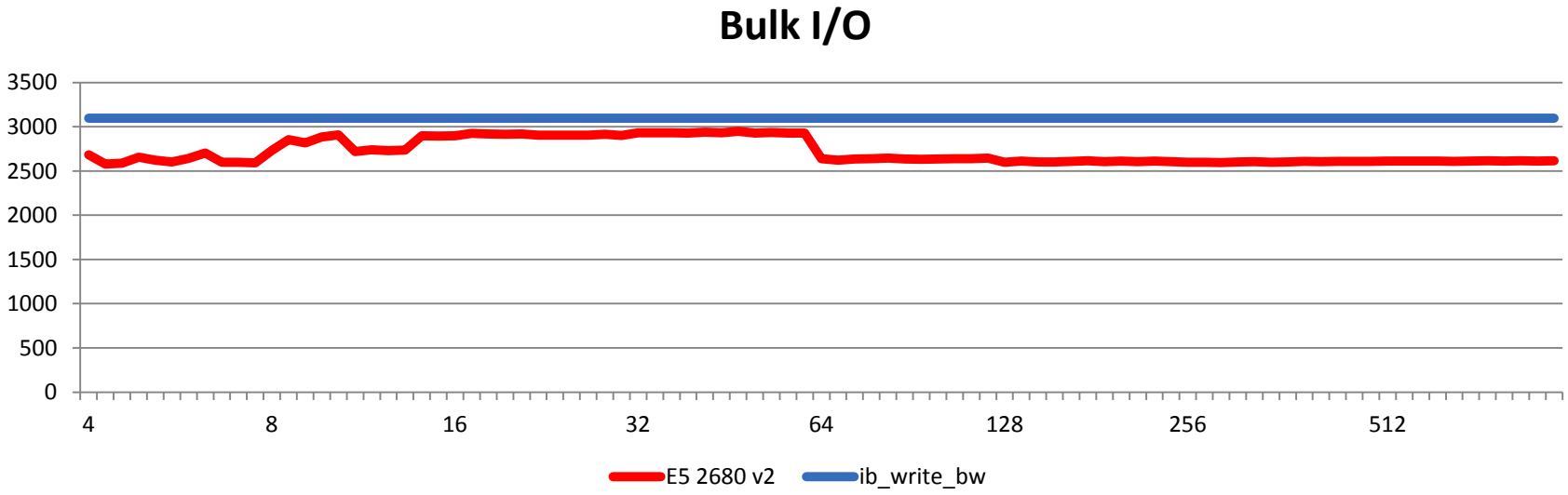
- We used the OFED *ib_read_bw* micro benchmark
- TrueScale QDR w/ IFS 7.2 -> 3167 MB/sec

- Lustre normally gets around 90-95% of the ib_read_bw number, between storage and client

Configurations - Summary

- Default configuration:
  - HT off
  - QIB default
  - LNET default
  - Operating System default
- Tuned and improved configuration:
  - HT off
  - QIB:
    - options ib_qib singleport=1 pcie_caps=0x51 krcvqs=4 rcvhdrcnt=4096
  - LNET:
    - options ko2iblnd peer_credits=128 peer_credits_hiw=64 credits=1024 concurrent_sends=256 ntx=2048 **map_on_demand=32** fmr_pool_size=2048 fmr_flush_trigger=512 fmr_cache=1
  - Operating System:
    - Use tuned-adm profile throughput-performance in RedHat 6.4

Tuned lnet selftest performance from low to high concurrency



- **Intel(R) Xeon(R) CPU E5-2680 v2 @ 2.80GHz, 10-core**
- Tuned configuration

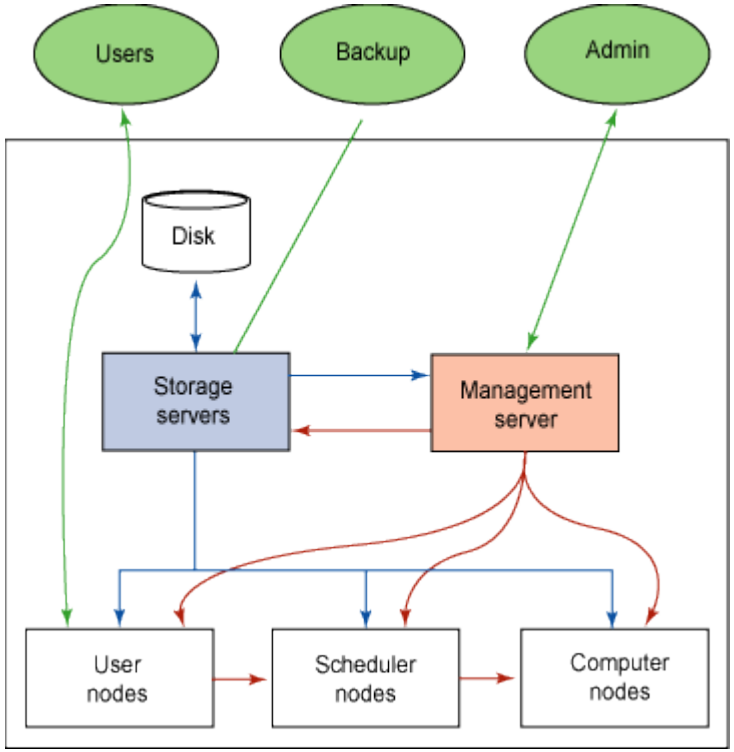
- Super Compact server – 4 server in 2U
- Only 1 PCI Express

IBM GPFS (now Spectrum Scale)

- Clustered File System
- Runs on Linux & Windows
- Can run over multiple fabrics, incl.:
  - Ethernet
  - IPoIB
  - RDMA
- Closed Source, owned by IBM
- GPFS User Group available (gpfsug.org)

GPFS Cluster

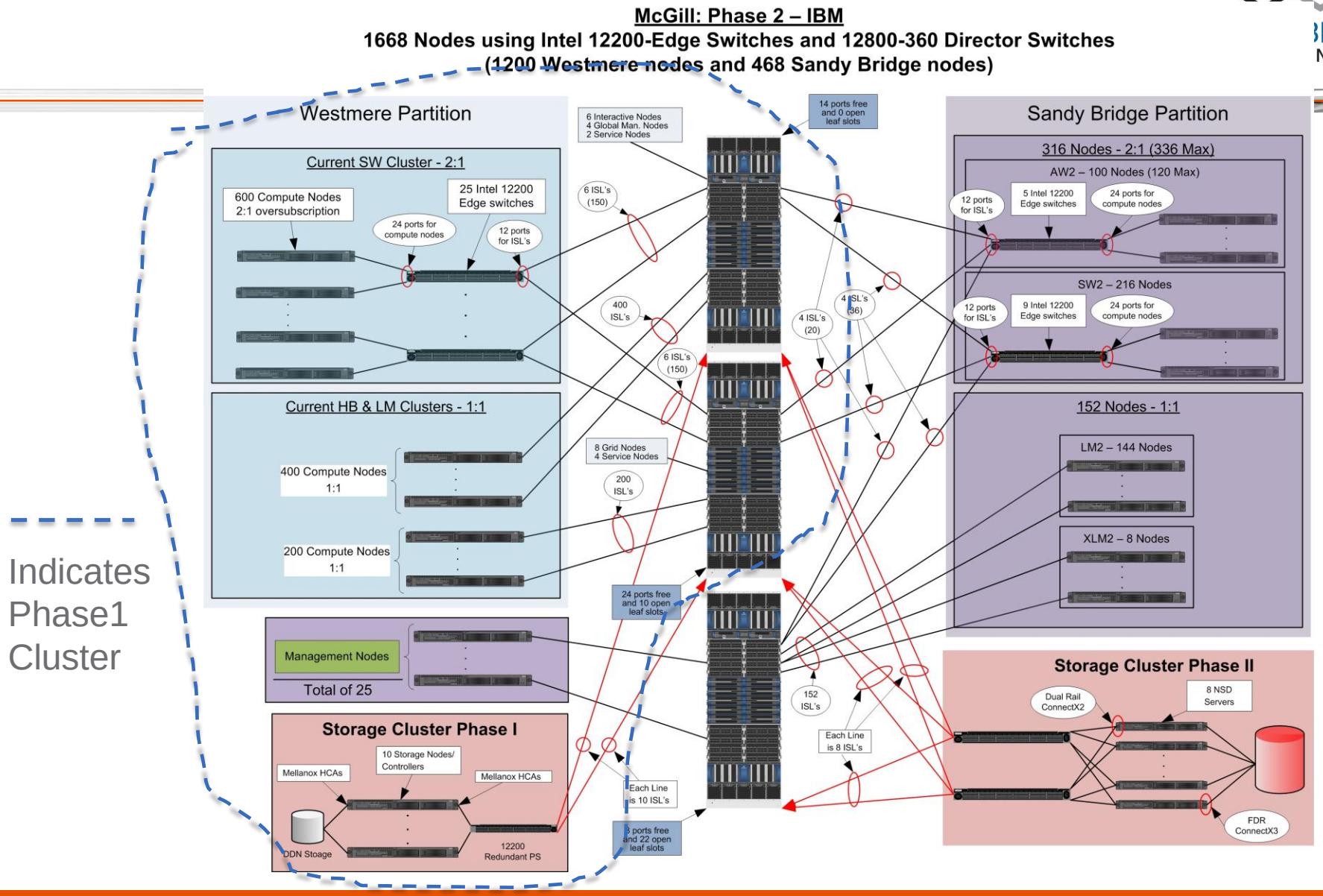
- NSD = Network Shared Disks
- NSD Server = the server fronting those disks
- Token management server



GPFS choices on an IB fabric

- Control traffic, aka RPC communications, options:
  - Ethernet, or
  - IPoIB
    - indicated by IP address or the host name of the NodeName that is being used to define the GPFS cluster
- Bulk Data traffic:
  - IPoIB
  - RDMA

GPFS Customer: McGill University (Montreal)



Performance Summary: GPFS

Test results as seen at McGill University with IBM using IOR benchmark

1 client and 4 servers (RDMA)	Write	Read
	GB/s	GB/s
No tuning, original IB driver (GB/s)	2.7	0.9
With GPFS tuning and enhanced driver 4 MB GPFS block size and 64K RDMA	6.5	6.5
Improvement	2.4X	7.2X

GPFS Parameters as tuned at McGill (Phase 1)

verbsRDMA : enable
verbsPorts : qib0/1
verbsRdmasPerNode : 112
verbsRdmasPerConnection : 14
verbsRdmaMinBytes : 8192
verbsRdmaMaxSendBytes : 65536
verbsRdmaTimeout : 14

GPFS Customer: McGill University (Montreal)

- Phase 1 Result: Exceeded the GPFS performance requirements for 1,200 node IBM iDataPlex cluster with DDN storage at McGill University

- Testimonial from OEM:
  "QLogic's GPFS performance is just as good as <Major Competitor>'s
  89%+ HPL efficiency
  Expect it to be a Top 100 system (in the 70s)
  It's nice to see a vendor actually achieve
  the #s they project."

GPFS Customer Crihan (Computer Resource Centre of Upper Normandy, France)

- News from customer / sales engineer …
- GPFS has been re-designed quite extensively in 3.5, in order to achieve more performance, but requires more threads than before.
- Good thread settings for this configuration:
  - nsdSmallThreadRatio=1,nsdminworkerthreads=96,nsdmxworkerthreads=96
- for the full config performance to the NSD nodes:
  - 3.2GB/s with GPFS 3.3 (without setting the number of threads)
  - 2.0GB/s with GPFS 3.5 (without setting the number of threads)
  - 4.8GB/s with GPFS 3.5 (when the number of threads is set),
  - 4.8 GB/s =maximum performance of the disk bay.

GPFS Customer – Vestas (mfg. of wind turbines, Denmark)

- Using 14 NSD servers, with 12 active disks on each NSD server, Achieved the requirement of 20.5 GiB/s
- Parameters
  - verbsRdma enable
  - verbsPorts qib0
  - verbsRdmaMaxSendBytes 65536
  - verbsRdmasPerNode 128
  - verbsRdmasPerConnection 16
  - maxMBpS 4000
  - idleSocketTimeout 0

Summary

- True Scale is an on-load architecture: PSM and verbs protocol processing is on the CPU cores
- SW optimization has improved verbs performance
- Worked closely with two major parallel file systems -- Lustre and GPFS -- to tune software and parameters to get good compute + file system performance …
- … leading to these success stories.

Thank You

OPENFABRICS ALLIANCE

OpenFabrics Software
User Group Workshop