Lunch and Learn: InfiniBand, OFED, Mellanox, and their use on Xcellis Systems

Bijan Tehrani (bijan.tehrani@quantum.com)

For Internal Quantum Engineering Use Only

POWER WHAT'S NEXT

InfiniBand, Mellanox OFED, and Xcellis

- What is InfiniBand(IB)?
- What are the uses of InfiniBand/who uses it?
- What is OFED?
- What IB software does MLNX_OFED provide?
- What IB software does Linux provide?
- What parts of IB configuration are integrated in Xcellis Platform software and what parts are not?
- Q/A

InfiniBand is a Hardware Transport Protocol

- InfiniBand Trade Association (IBTA) http://www.InfiniBandta.org/
  - Founded in 1999, the InfiniBand® Trade Association (IBTA) is chartered with maintaining and furthering the InfiniBand™ Architecture specification defining **hardware transport protocols** sufficient to support both reliable messaging (send/receive) and memory manipulation semantics (e.g. remote DMA) without software intervention in the data movement path. These transport protocols are defined to run over Ethernet (RoCE) as well as InfiniBand fabrics.

(This presentation's InfiniBand diagrams and information uses excerpts from http://www.mellanox.com/related-docs/whitepapers/InfiniBandFAQ_FQ_100.pdf and IBTA architecture document (available from www.infiniBandta.org website)

InfiniBand Defines Asynchronous Communication

- Message-based asynchronous communication
  - A consumer application queues up work requests (WRs) to be executed by the IB hardware by adding work queue elements (WQE) in its work queue (WQ)
  - Work queues are typically created in pairs (Queue pair, or QP), one for send and one for receive operations.
    - Send Q holds instructions for transferring data from consumer memory to remote consumer memory.
    - Receive Q holds instructions about where to put data that is received from remote consumer.

InfiniBand Provides Remote Direct Memory Access (RDMA)

- Operations performed by Work Requests
  - SEND operation
    - Consumer specifies a block of data in memory to be send to destination in WQE added to send Q.
    - Requires remote consumer to already have a WQE in its receive Q with memory address where the received data is to be placed
  - RDMA operation
    - WQE specifies the address in the remote consumer's memory where the data is to be placed as well.
    - It does not need a WQE in receive Q of the destination.

InfiniBand RDMA Operations

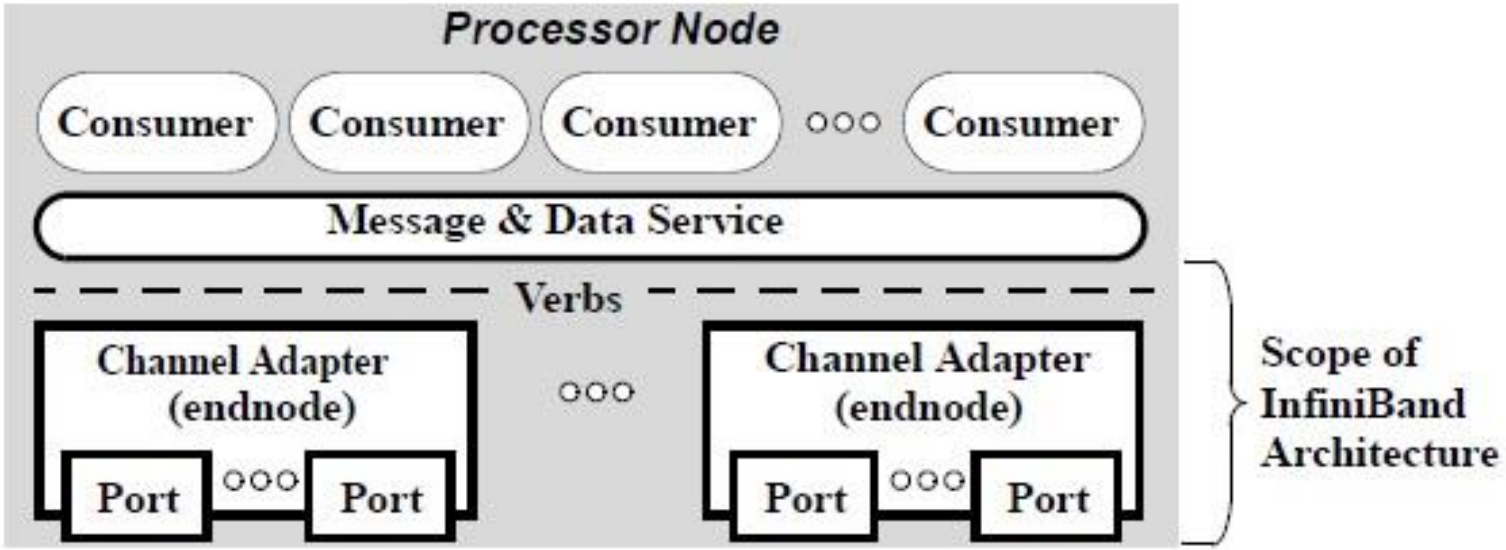
- RDMA Operation Types
  - RDMA-WRITE : Request hardware to transfer data from consumer memory to remote consumer memory
  - RDMA-READ: Request hardware to transfer data from remote consumer memory to consumer memory
- Using Queue Pairs (QP) by consumer:
  - Each consumer may create and use multiple QPs
  - Consumer and Remote consumer may be on the same node

InfiniBand Channel Adapter Hardware

- InfiniBand Channel Adapter equipment mainly provides
  - Queuing services to consumers for requesting work to be performed to SEND or to RDMA data between consumer and remote consumer memories
  - Send and receive IB packets on the wire
- Two types of Channel Adapters are identified
  - Host Channel Adapter (HCA) equipment is used by host endnode
    - ConnectX-3 VPI card in Xcellis node
  - Target Channel Adapter (TCA), is integrated and used by target endnode
    - NetApp array with IB interface controllers

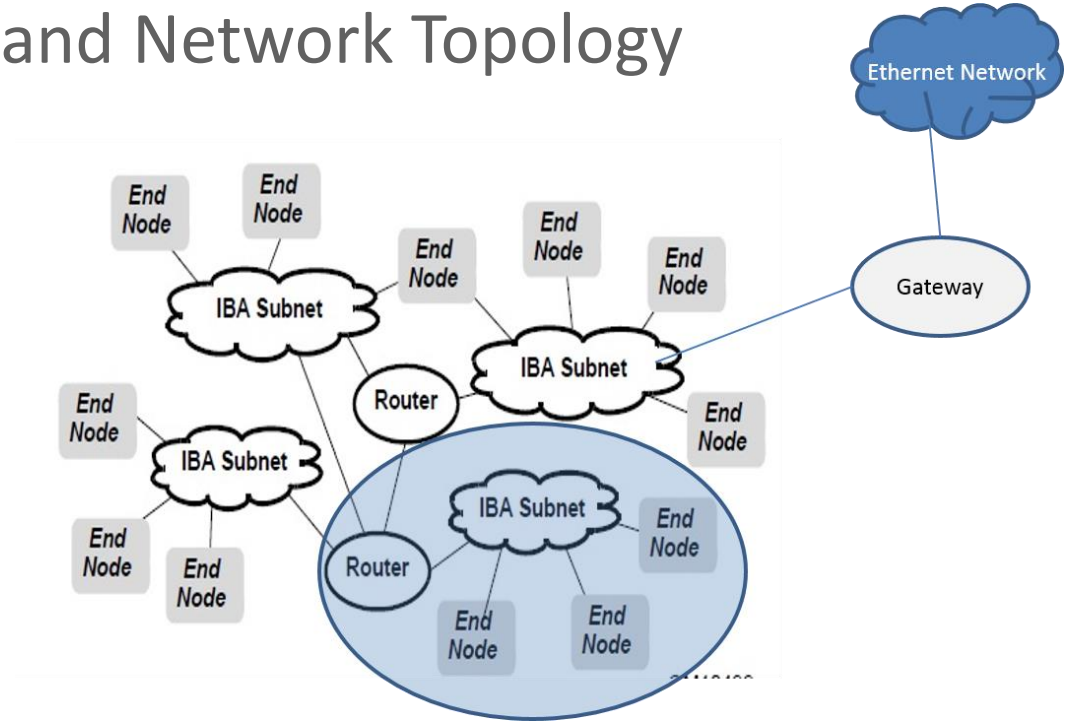
What Is InfiniBand?

- An InfiniBand Processor node



InfiniBand Architecture Model

- An InfiniBand Network Topology

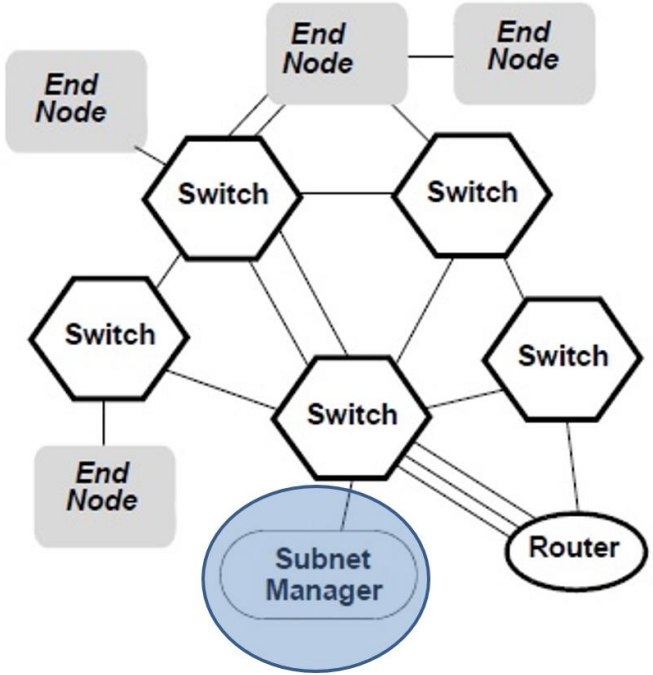


InfiniBand Architecture Model Elements

- End nodes Hosts with IB HCA, targets with IB TCA)
- Switches – physically connecting end nodes
- Subnet – collection of switches and end nodes
  - A subnet of up to ~40,000 end nodes can run efficiently
  - No need to break down to multiple smaller subnets
- Routers - Connect InfiniBand subnets
  - needed when number of nodes in a subnet > 40,000
- Gateway – Connects InfiniBand subnet to other protocol networks (e.g. InfiniBand to Ethernet)

InfiniBand Architecture Subnet Manager

- Subnet Manager – manages an IB Subnet



InfiniBand Architecture Subnet Manager

- Subnet Manager
  - Can run on any entity: end nodes or switches
  - One SM must be active at all times, others will be standby
  - Configures local subnet and ensures its continued operation
  - Sets up primary & secondary paths between every end point
- Subnet Manager Agent
  - Each entity runs an SM Agent to allow SM to communicate with the various InfiniBand components
  - Implements Software-Defined Networking (SDN) – controls and handles devices, data traffic, flexibility, and scalability.

Comparing InfiniBand to TCP/IP Networking

InfiniBand	TCP/IP/Ethernet
Hardware-centric Architecture	Software-centric Architecture
Data Delivery: Application-centric Point-to-Point Remote Direct Memory Access (**RDMA**) between end-points (zero-copy)	Data Delivery: Network-centric indirect access among end-points through OS Networking (kernel/user level coping)
Low-latency due to send/receive offload via InfiniBand adapter	Latency due to CPU usage for network processing through OS networking stack
High-throughput: 56-100Gb/s	Catching up: 100Gb/s

InfiniBand Message Delivery Communication Model

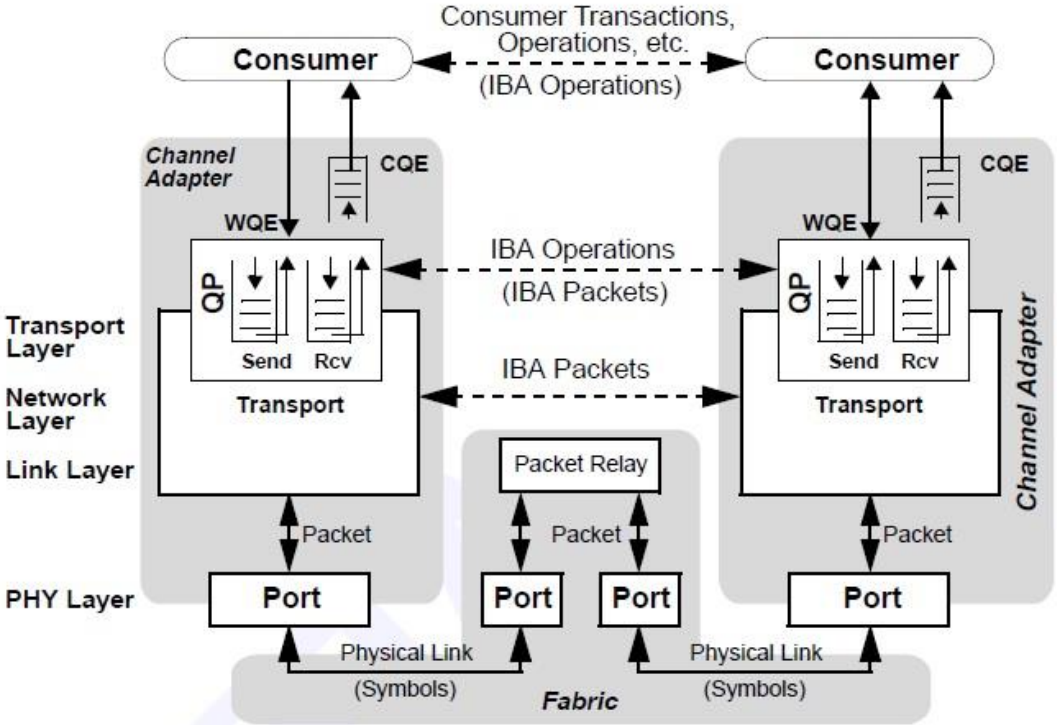
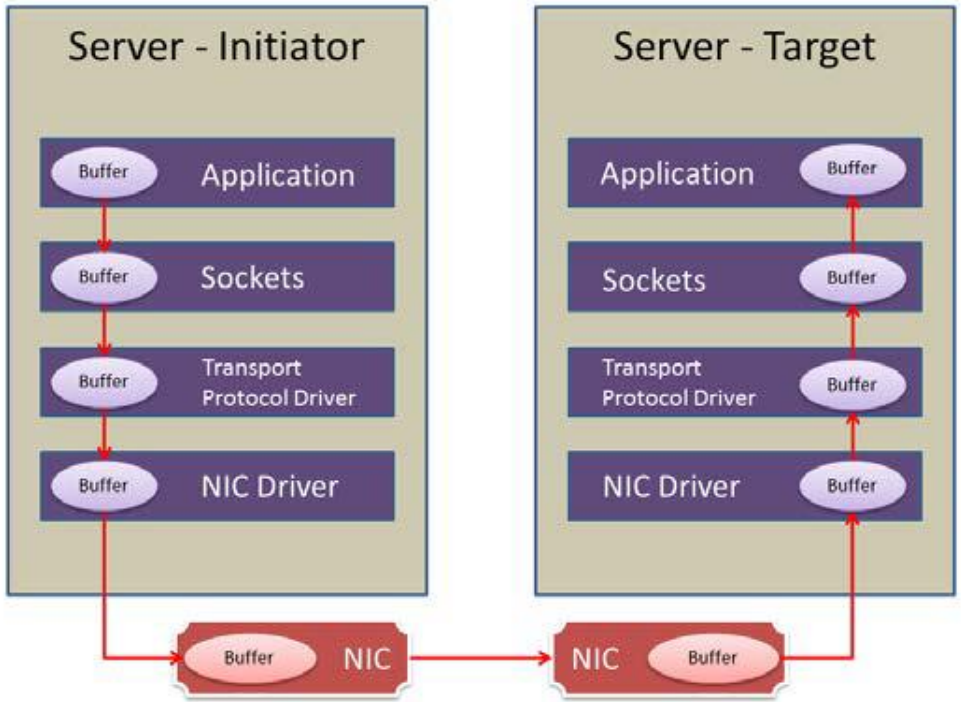


Figure 13 IBA Communication Stack

TCP/IP Packet Delivery Communication Model



InfiniBand, Mellanox OFED, and Xcellis

- What is InfiniBand(IB)?
- **What are the uses of InfiniBand/who uses it?**
- What is OFED?
- What IB software does MLNX_OFED provide?
- What IB software does Linux provide?
- What parts of IB configuration are integrated in Xcellis Platform software and what parts are not?
- Q/A

What are the uses of InfiniBand

- InfiniBand technology provides messaging service for RDMA operations (main function of the IB HCA) between consumers at two end of communication over InfiniBand defined network
- Almost any other I/O and network communication protocol modules may be implemented canonically on top of IB RDMA service.
- These modules may be user-level or kernel-level
- OFA defines and maintains reference implementation of many such well-known protocol modules under the name OFED
- Mellanox has contributed to, and continues to contribute to, adopt and release OFA software under its own MLNX_OFED software bundle

InfiniBand, Mellanox OFED, and Xcellis

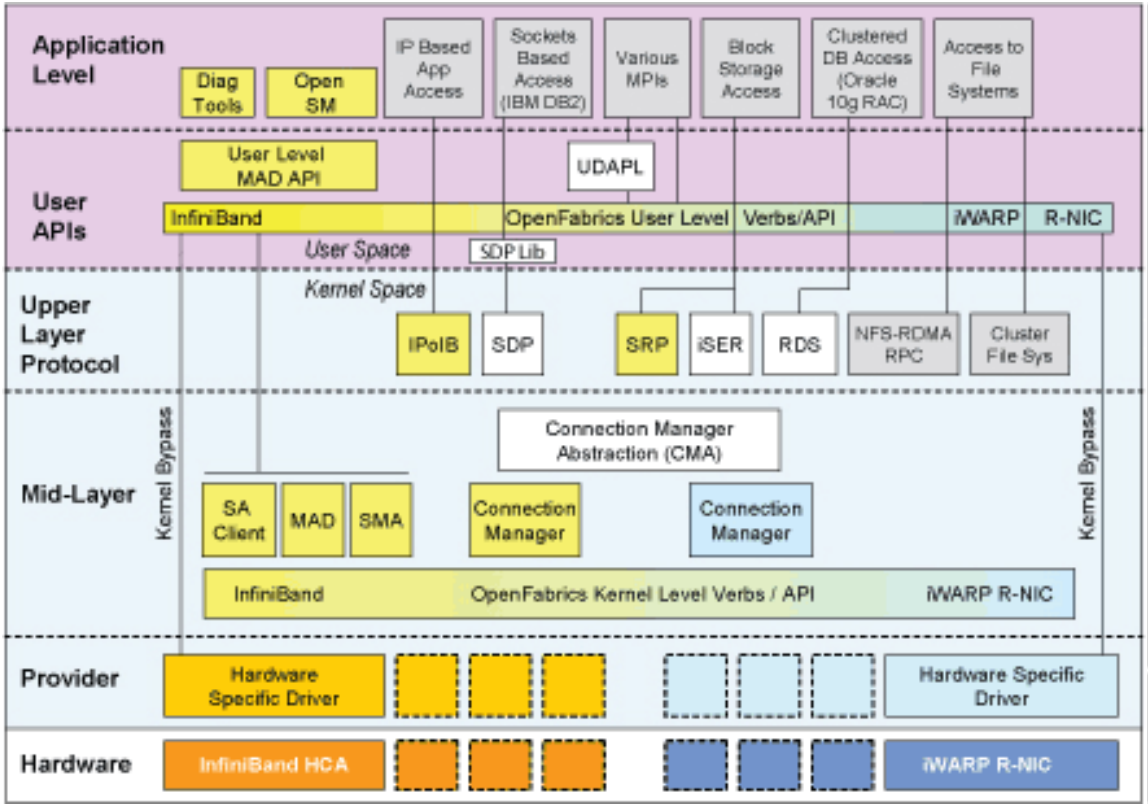
- What is InfiniBand(IB)?
- What are the uses of InfiniBand/who uses it?
- **What is OFED?**
- What IB software does MLNX_OFED provide?
- What IB software does Linux provide?
- What parts of IB configuration are integrated in Xcellis Platform software and what parts are not?
- Q/A

What is OFED (a.k.a OFS)?

- OpenFabrics Alliance(OFA)https://www.openfabrics.org defines OFED
  - OpenFabrics Enterprise Distribution (OFED™)/OpenFabrics Software is open-source software for RDMA and kernel bypass applications.
  - OFS includes kernel-level drivers, channel-oriented RDMA and send/receive operations, kernel bypasses of the operating system, both kernel- and user-level application programming interface (API) and services for parallel message passing (MPI), sockets data exchange (e.g., RDS, SDP), NAS and SAN storage (e.g. iSER, NFS-RDMA, SRP) and file system/database systems.

  (quoted from the OFED Overview page of the website)
  - Openfabrics.org maintains all past and present reference implementations of OFED for major operating systems

OFED Software Components



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

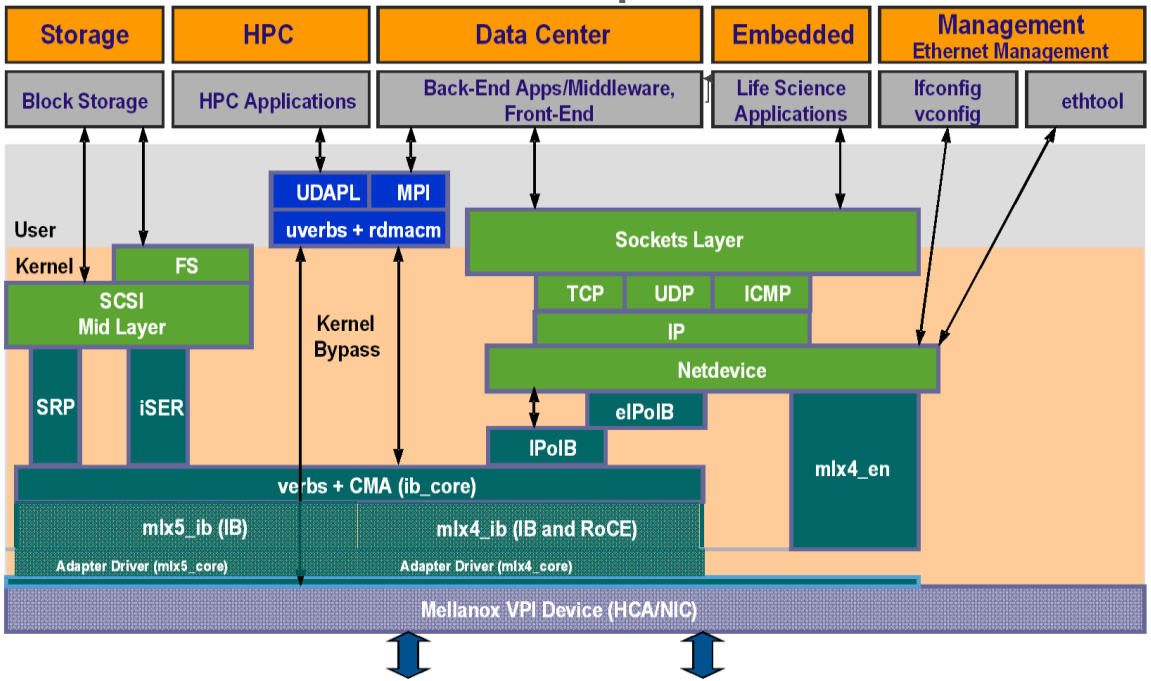
Key: Common; InfiniBand; iWARP; Apps & Access Methods for using OF Stack

InfiniBand, Mellanox OFED, and Xcellis

- What is InfiniBand(IB)?
- What's the use of InfiniBand/who uses it?
- What is OFED?
- **What IB software does MLNX_OFED provide?**
- What IB software does Linux provide?
- What parts of IB configuration are integrated in Xcellis Platform software and what parts are not?
- Q/A

MLNX_OFED

- MLNX_OFED is Mellanox Implementation of (parts of) OFED defined software components



MLNX_OFED

- MLNX_OFED software includes
  - HCA drivers
    - Mlx5: mlx5_ib, and mlx5_core (includes Ethernet driver)
    - Mlx4: mlx4_core, mlx4_ib, mlx4_en (Ethernet driver)
  - Mid-layer core
    - Verbs, MADs, SA, CM, CMA, uVerbs, uMADs
  - **Upper Layer Protocols (ULPs)**
    - IPoIB (IP over InfiniBand)
    - iSER (iSCSI Extensions for RDMA)
    - SRP (SCSI over RDMA)
    - SDP (Sockets Direct Protocol)
    - uDAPL (User Direct Access Programming Library)

InfiniBand, Mellanox OFED, and Xcellis

- What is InfiniBand(IB)?
- What's the use of InfiniBand/who uses it?
- What is OFED?
- What IB software does MLNX_OFED provide?
- **What IB software does Linux provide?**
- What parts of IB configuration are integrated in Xcellis Platform software and what parts are not?
- Q/A

Linux InfiniBand Support

- Linux has integrated core InfiniBand components, majority of which are contributed by Mellanox
  - mlx4_core, mlx4_ib, mlx4_en, mlx5_core, mlx5_ib
- However when MLNX_OFED is installed, it possibly installs its own, newer versions, of these modules that are compatible with rest of MLNX_OFED software version being installed.

InfiniBand, Mellanox OFED, and Xcellis

- What is InfiniBand(IB)?
- What's the use of InfiniBand/who uses it?
- What is OFED?
- What IB software does MLNX_OFED provide?
- What IB software does Linux provide?
- **What is Supported by Xcellis Platform for InfiniBand?**
- Q/A

Xcellis and InfiniBand HCA Support

- Xcellis currently only supports Mellanox ConnectX-3 VPI card (CX-3).
- ConnectX-4 VPI (CX-4) cards are under consideration for addition to Xcellis future releases

Xcellis and MLNX_OFED Software Support

- When a supported IB HCA is present in the system the entire MLNX_OFED software will be installed
- Xcellis supports management of CX-3 card port protocol mode setting: Ethernet or InfiniBand
- Xcellis supports Ethernet mode ports of CX-3 entirely as any other Ethernet ports in the system
- Xcellis supports management of InfiniBand protocol modules insertion and removal from the kernel for use when ConnectX-3 VPI card port is in InfiniBand mode

Xcellis InfiniBand Management Support

- Xcellis InfiniBand service menu allows for:
  - Port Protocol setting: changing MLNX CX3 IB card port mode
    - Ethernet mode (default), the port is run as Ethernet via MLX4_EN Ethernet driver module.
    - InfiniBand mode: MLNX_OFED automatically uses IPoIB module on InfiniBand mode ports, creating ib# ports for IP address configuration. Other uses via ULPs pushed into kernel
  - Upper Layer Protocol (ULP) setting: Allows ULP protocol modules to be inserted onto or removed from IB stack in the kernel

Xcellis fully supports managing 40GbE Interfaces

- From Xcellis InfiniBand service menu
  - Change IB card port mode to Ethernet
  - Reboot (this happens when quitting IB service menu after changes)
- From Xcellis Network service menu, or StorNext GUI use the created 40GbE interface for IP address configuration and use

Xcellis does not manage configuration of IPoIB interfaces

- Xcellis service menu allows switching to InfiniBand mode to use IPoIB interfaces on IB network. But no further involved in configuring IPoIB interfaces.
- IPoIB creates ib# interfaces associated with CX-3 ports. However, association is not consistent, the assigned names depends on the order the ports were switched to InfiniBand mode.
- IPoIB interfaces are quite different from IP over Ethernet interfaces in configuration(other than IP address assignment) and in behavior.

Why does Xcellis not have IPoIB management support

- IPoIB interfaces do not support virtual interface creation
  - IPoIB instead has subinterface support, with quite different configuration steps for addition and removal of them
- IPoIB connected or datagram, is a configuration setting (via SET_IPOIB_CM=yes/no in /etc/infiniband/openib.conf), not a user choice. Users can only use IPoIB in the set mode.
- Bonding of IPoIB interfaces does not achieve higher bandwidth, Mellanox only recommends active/backup bonding.

Why does Xcellis not have IPoIB management support

- IPoIB interfaces MTU setting process is different, and is managed by IB Subnet Manager)
- IPoIB allows running Ethernet on top of it, via eIPoIB module. This creates Ethernet-like interfaces on the system. Must not allow eIPoIB even in eXpress deals
- Refer to section IP over InfiniBand(IPoIB) of the MLNX_OFED user manual for details of all of the above settings
- All these peculiarities of IPoIB interfere with StorNext networking management of Ethernet based interfaces

Xcellis supports configuration of SRP (SCSI over RDMA)

- Connect SRP target storage to IB network
  - Follow the steps described by the SRP storage vendor for setting up the SRP target storage (and necessary setups for the host side)
- From Xcellis InfiniBand service menu
  - Change port mode to InfiniBand
  - Add SRP module to be inserted in kernel
  - Reboot (this happens when quitting IB service menu after changes)
- Use created SCSI devices related to SRP target storage

Xcellis does not support configuration of iSER

- iSER (iSCSI Extensions for RDMA)
  - Permits data to be transferred directly into and out of SCSI buffers, eliminates TCP/IP processing
  - It provides access management to storage, by virtue of using same management tools as iSCSI, e.g. iscsiadm.
  - However, compared to SRP it is complicated to use
    - It requires TCP/IP network for management communication with target in addition to using SRP
    - On an IB network that translates to using IPoIB on both iSER communication ends (initiator and target)
      - ISER is implemented on various storage targets such as TGT, LIO, SCST and out of scope of this manual. (From MLNX_OFED user manual)

Xcellis does not support configuration of iSER

- Some references about iSER
  - For iSER detailed information refer to section 3.2.2
    - http://www.mellanox.com/related-docs/prod_software/Mellanox_OFED_Linux_User_Manual_v4.0.pdf
  - *How to Configure LIO enabled with iSER* by Mellanox:
    - https://community.mellanox.com/docs/DOC-1472
  - *iSER on InfiniBand Network* by IBM research:
    - https://www.research.ibm.com/haifa/satran/ips/iSER-in-an-IB-network-V9.pdf

OFED User Libraries available on Xcellis

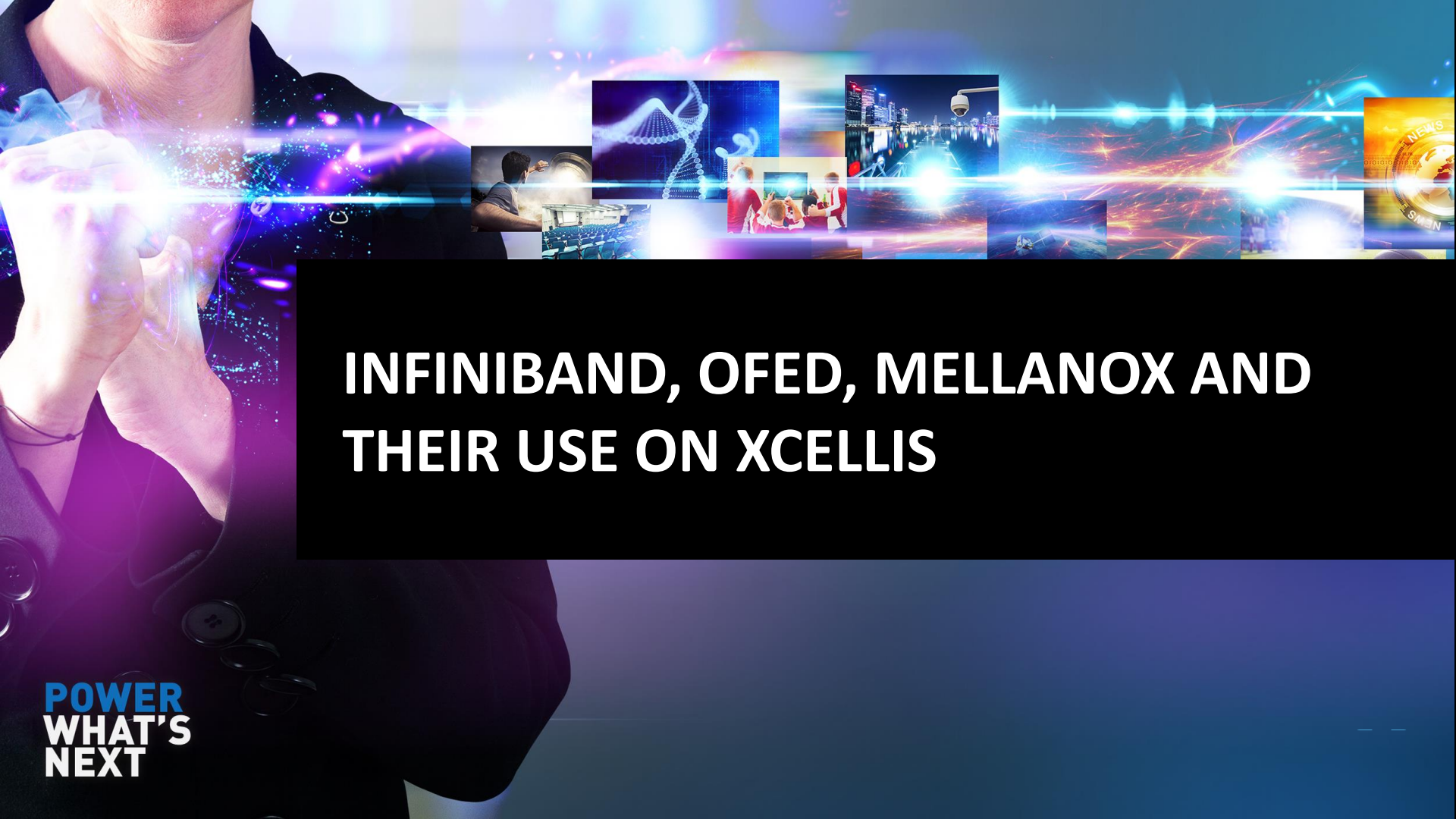
- MLNX_OFED installs support and user application writers may use
  - Iverb library: Provides user access to IB RDMA service
    - Start with man ibv_create_qp(3) for creating a Queue Pair from a user application. This demonstrates the power of having direct access to RDMA, main service provided by an IB HCA
      - StorNext over RDMA protocol (SNRP) in future?
  - uDAPL : User Direct Access Programming Library
  - MPI: Message Passing Interface library

InfiniBand, Mellanox OFED, and Xcellis

- What is InfiniBand(IB)?
- What's the use of InfiniBand/who uses it?
- What is OFED?
- What IB software does MLNX_OFED provide?
- What IB software does Linux provide?
- What parts of IB configuration are integrated in Xcellis Platform software and what parts are not?
- **Q/A**

InfiniBand, Mellanox OFED, and Xcellis

- What is InfiniBand(IB)?
- What's the use of InfiniBand/who uses it?
- What is OFED?
- What IB software does MLNX_OFED provide?
- What IB software does Linux provide?
- What parts of IB configuration are integrated in Xcellis Platform software and what parts are not?
- Q/A


INFINIBAND, OFED, MELLANOX AND THEIR USE ON XCELLIS
POWER
WHAT'S
NEXT

Quantum®

POWER WHAT'S NEXT