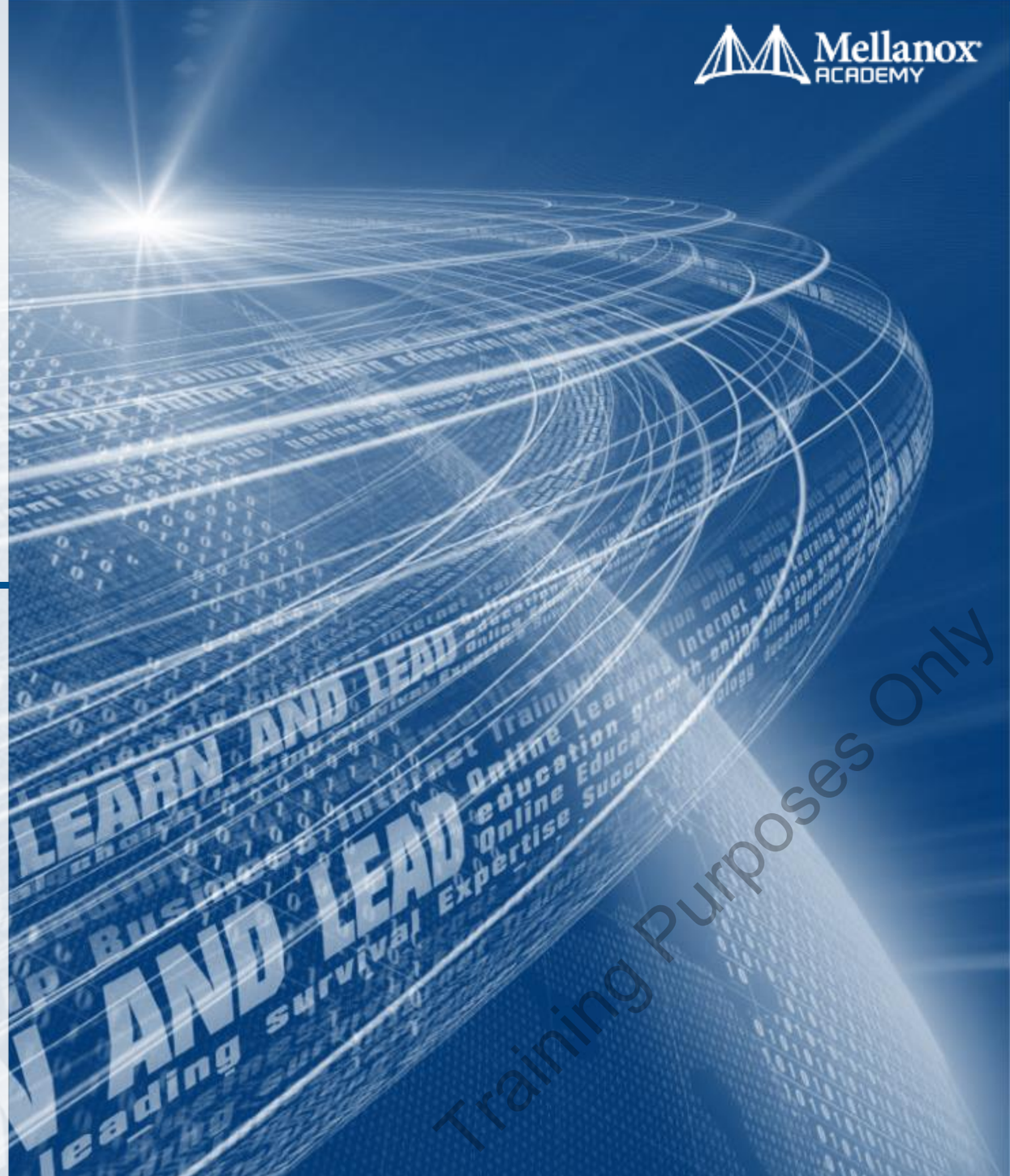
Introduction to InfiniBand

Unit 1



1. InfiniBand Trade Association (IBTA)
2. What is InfiniBand?
3. Why InfiniBand?
4. InfiniBand Key Features
5. InfiniBand Fabric Components

By the end of this unit, you will be able to:

- Describe the benefits of InfiniBand fabrics
- List the most common usages of InfiniBand fabrics
- Describe the main features of InfiniBand fabrics
- List the major network components of InfiniBand fabrics

Mellanox Technologies Company Overview

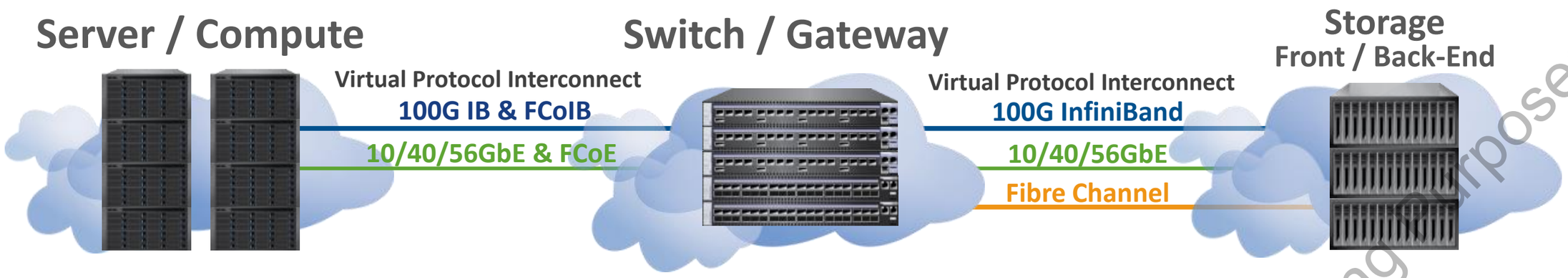
- A leading supplier of end-to-end InfiniBand and Ethernet interconnect solutions and services for servers and storage
- Founded in 1999
- Company headquarters:
  - Yokneam, Israel; Sunnyvale, California
  - ~2000 employees* worldwide

InfiniBand is an open standard interconnect protocol developed by the **InfiniBand® Trade Association (IBTA)**

- Was released in 2000
- Provides a solution starting from the **hardware** layer to the **application** layer
- InfiniBand Software is developed under the OpenFabrics Open Alliance
- The InfiniBand technology is based on the InfiniBand Specification

What is InfiniBand?

- Interconnect technology for interconnecting processor nodes and I/O nodes to form a system area network
- The architecture is independent of the host operating system (OS) and processor platform
- Open industry-standard specification
- Defines an input/output architecture
- Offers point-to-point bidirectional serial links



- InfiniBand is a technological advancement over Ethernet

Cloud Computing

Data Centers

HPC

Financial Companies

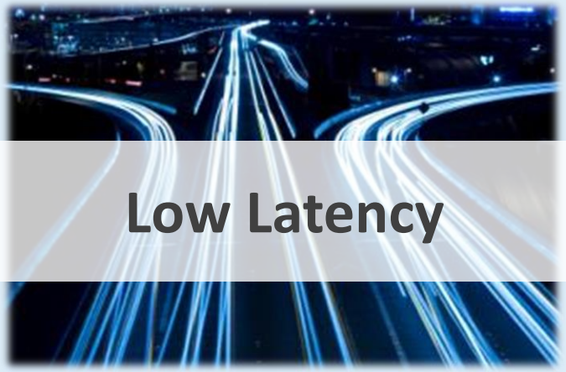
Big Data

Keys to Success

Low Latency
High Throughput
Parallel Computing

InfiniBand Key Features

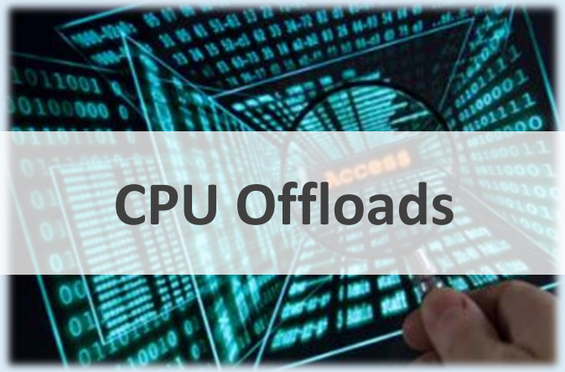
High Bandwidth



Low Latency

Quality of Service

Scalability/Flexibility



CPU Offloads

Simplified Management

InfiniBand Bandwidth

Result of single lane speed multiplied by the number of combined lanes (width).

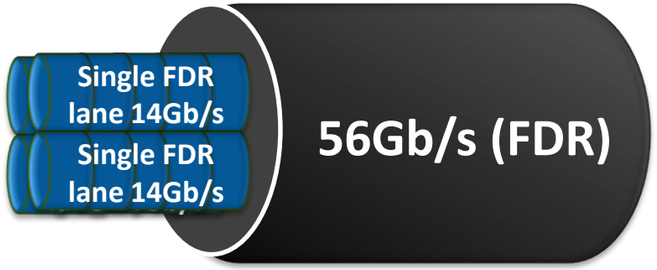
10 gb/s	20 gb/s	40 gb/s	56 gb/s	100 gb/s
2.5*4	5*4	10*4	14*4	14*4
Single Data Rate	Double Data Rate	Quadruple Data Rate	Fourteen Data Rate	Enhanced Data Rate
SDR	DDR	QDR	FDR	EDR
2002	2005	2008	2011	2013

IB Widths (combined lanes) are: **1X, 4X**

InfiniBand Bandwidth Continued

Width

$\times 4 = 56\text{Gb/s}$

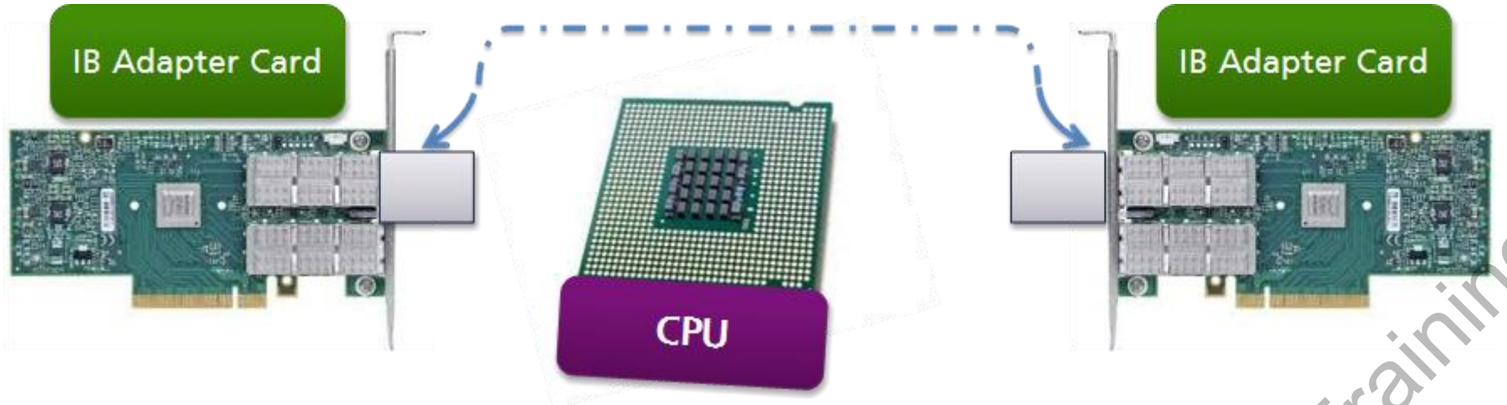


What is the throughput that a QDR network adapter card can support?

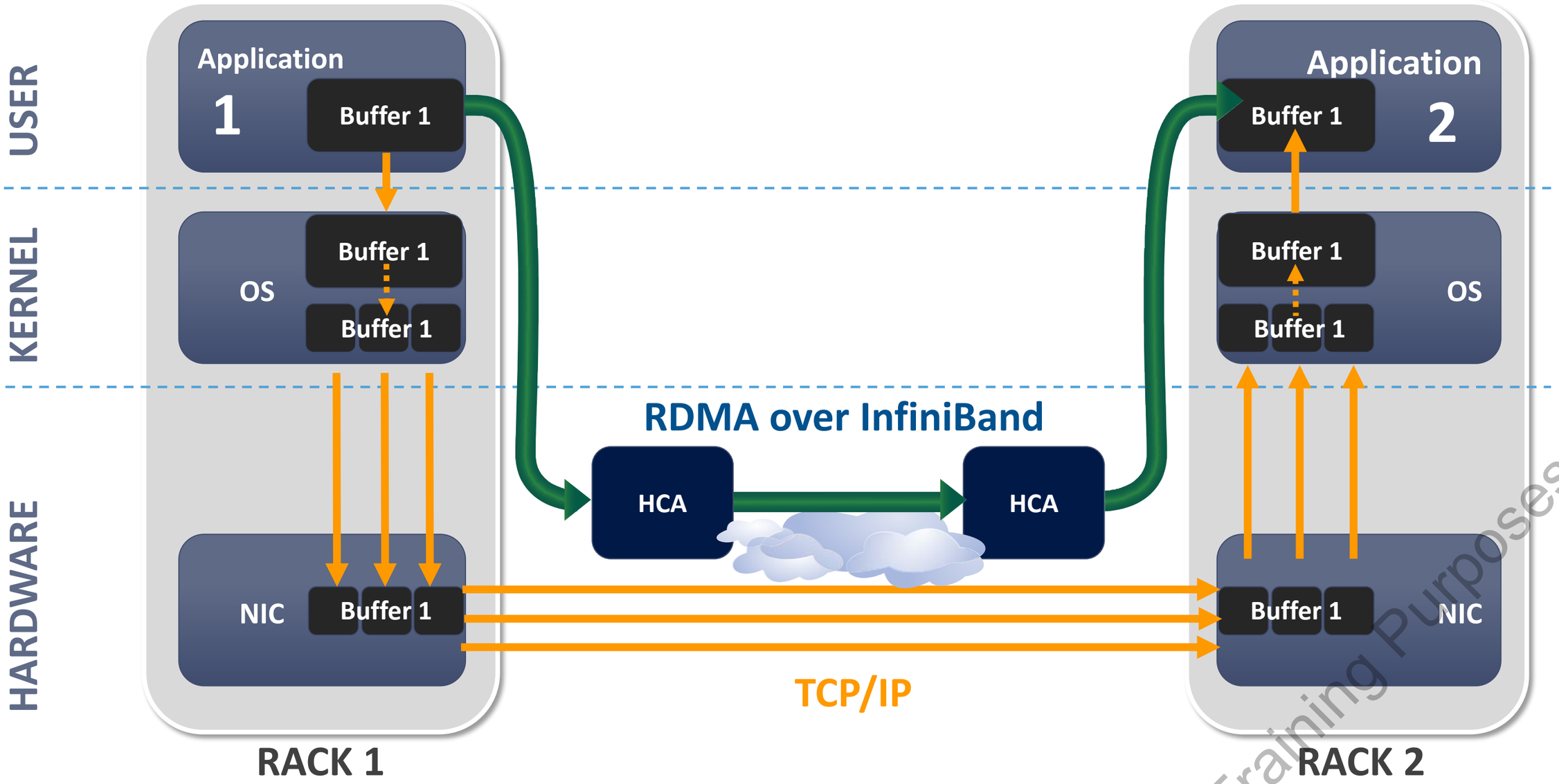
?

- The IB architecture supports packet transportation with minimal CPU intervention
- This is achieved thanks to:
  - Hardware-based transport protocol
  - Kernel bypass
  - RDMA support

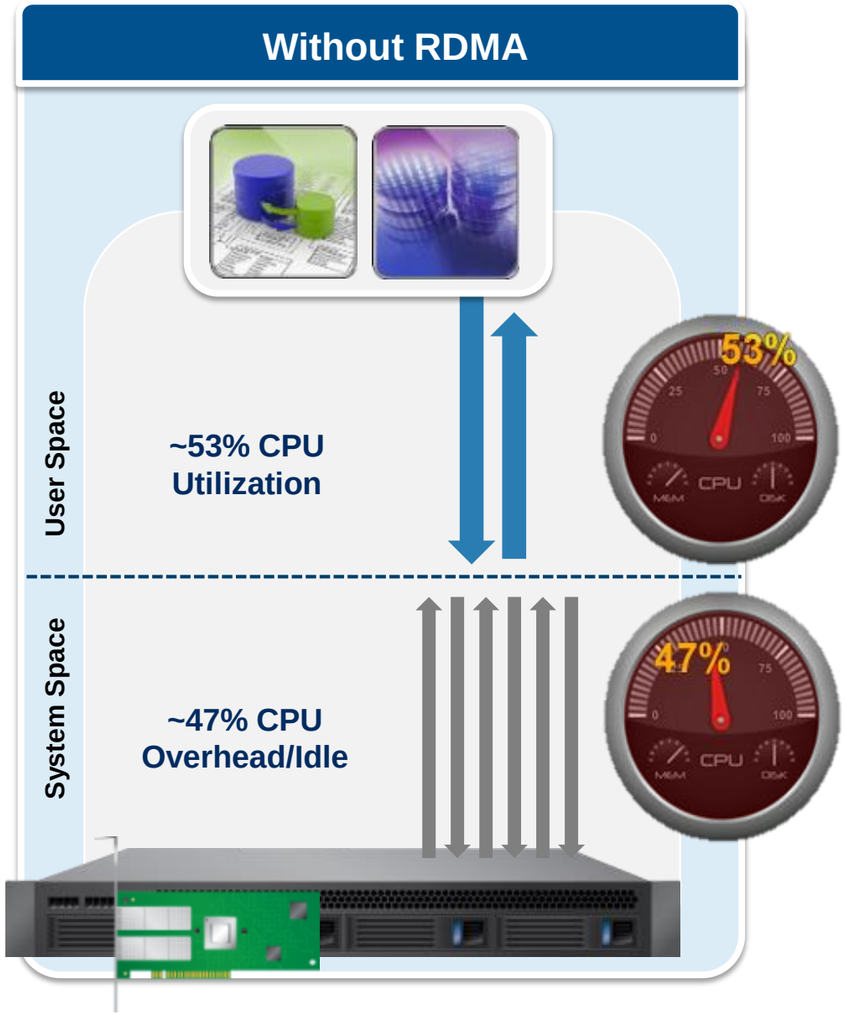
Mellanox adapter cards

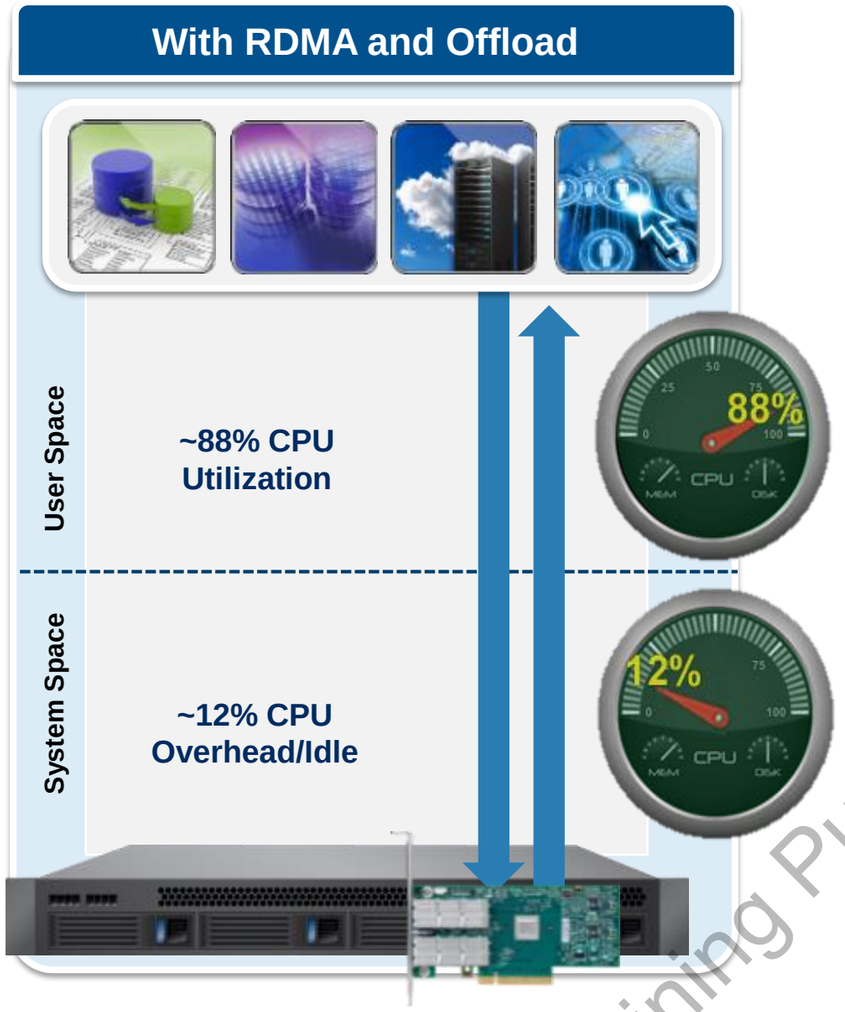


RDMA – How does it Work

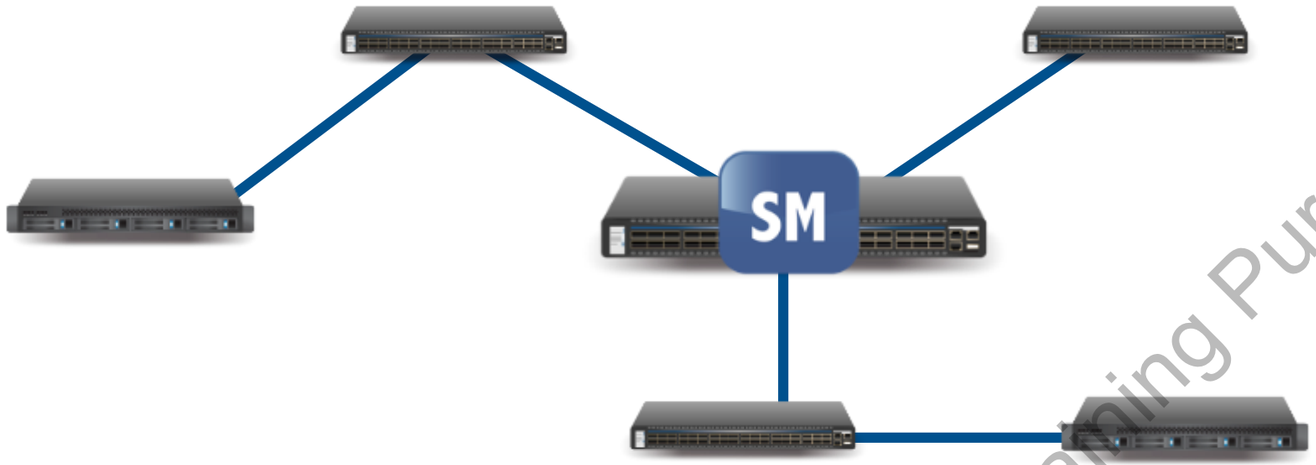


CPU Offload with RDMA





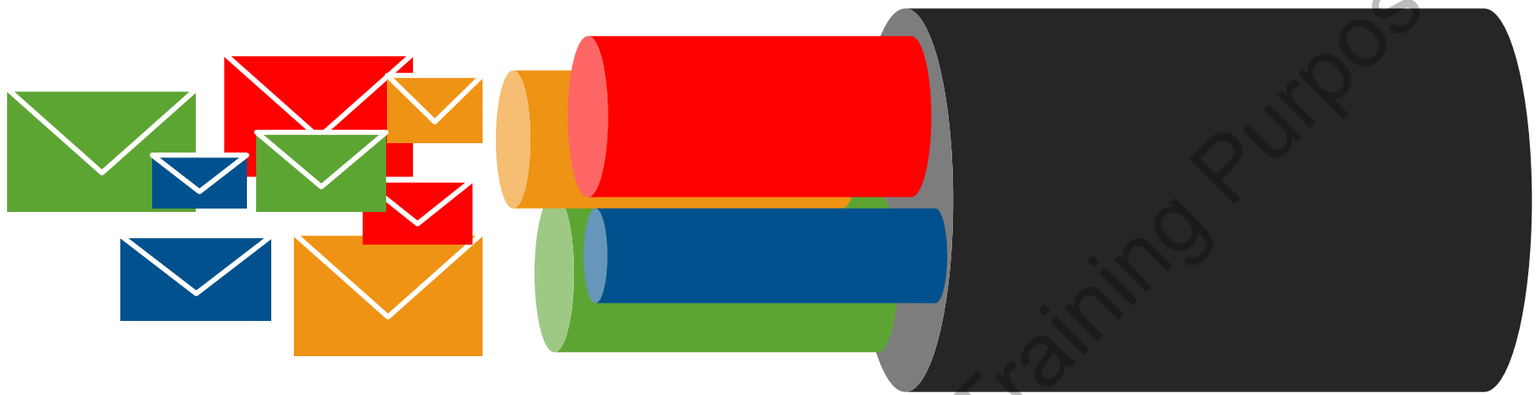
- Subnet Manager (SM) - a program that runs and manages the fabric
  - Any IB fabric has its own single (master) SM
  - The SM makes the fabric management very simple:
    - Plug & Play end nodes environment
    - Centralized route manager



The ability to provide **different priority** to different:

- Applications
- Users
- Data flows

- QoS implementation can be achieved by:
  - Defining I/O channels at the adapter level
  - Defining Virtual Lanes at the link level
- Allows the control of congestion on the network



Scalability and Flexibility

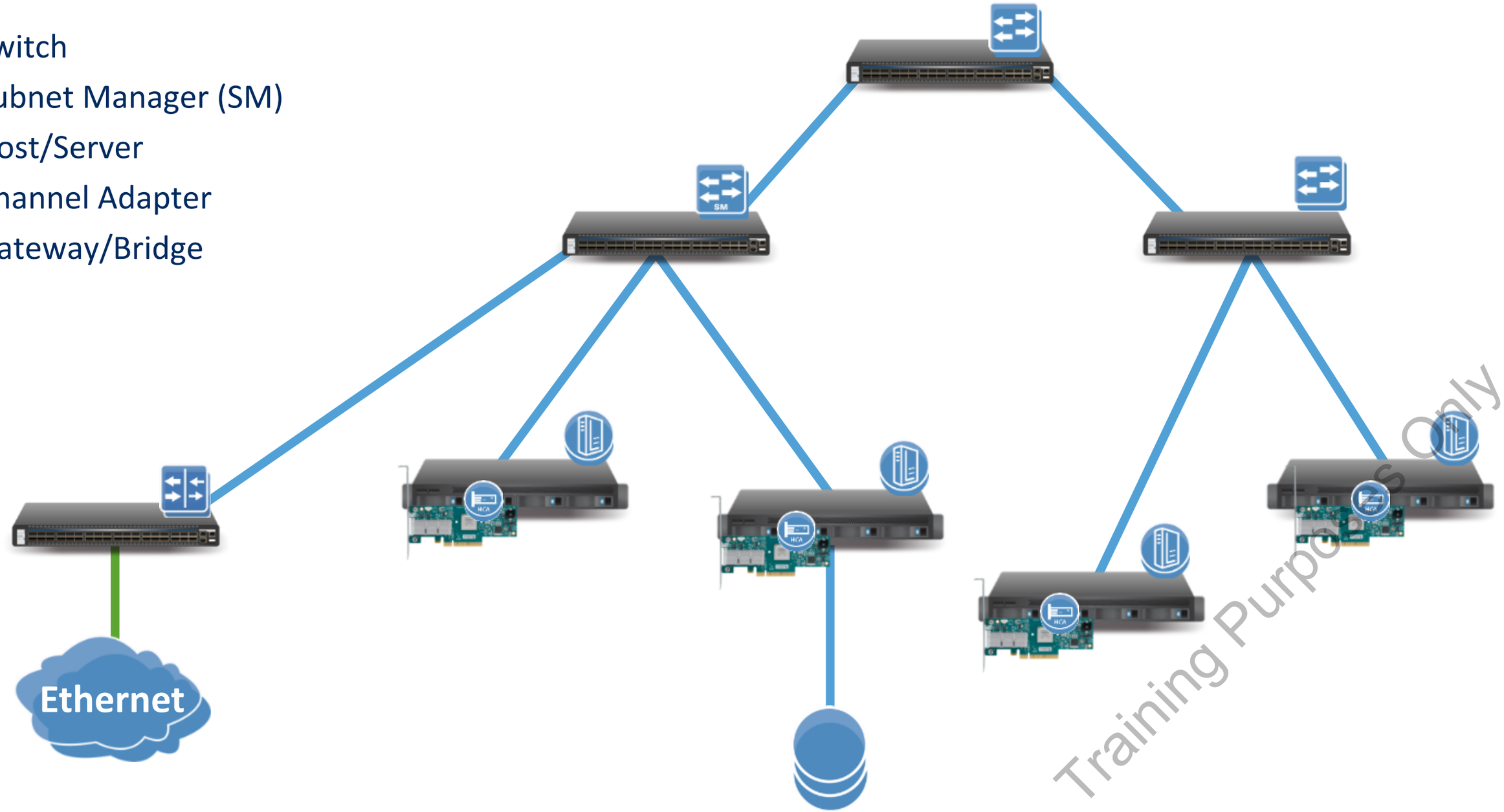
- IB devices offer easy scaling for data centers with great flexibility
- IB enables a scaling of up to:

48,000

Nodes in
a single subnet

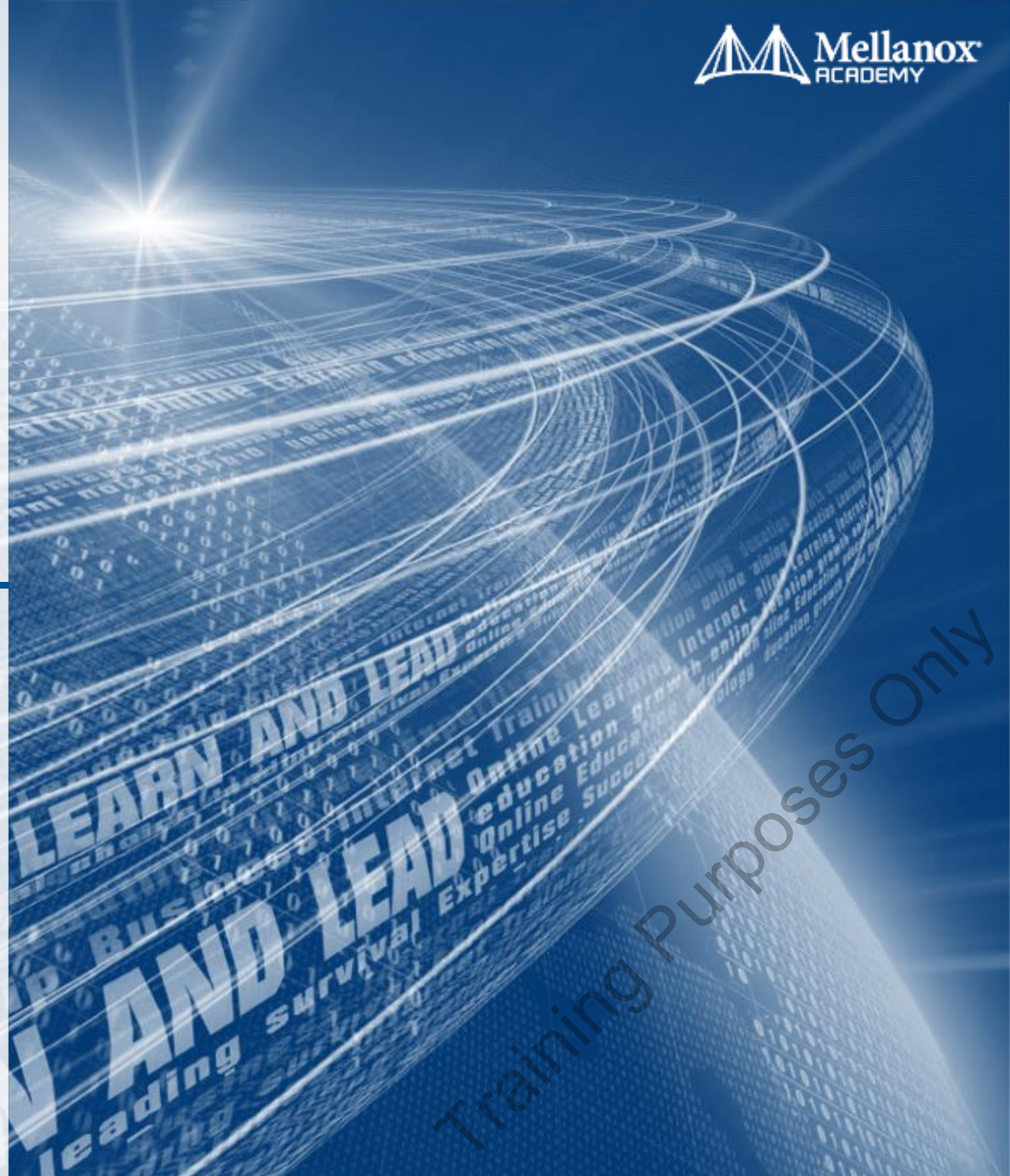
InfiniBand Fabric Components

- Switch
- Subnet Manager (SM)
- Host/Server
- Channel Adapter
- Gateway/Bridge



Introduction to InfiniBand Architecture and Management

Unit 2



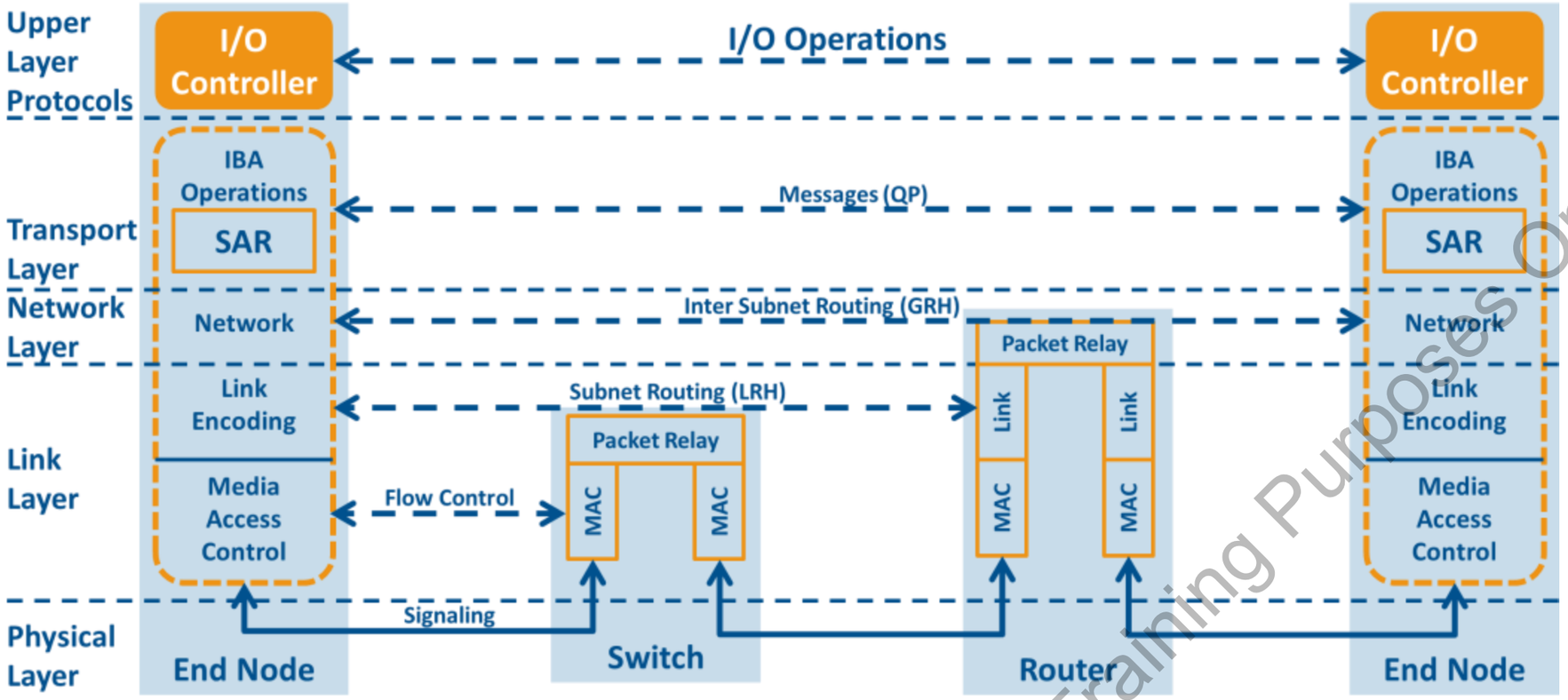
Outline

1. InfiniBand Network Stack
2. IB Architecture Layers
3. Data Packet Structure
4. The Subnet Manager (SM)
5. Fabric Addressing
6. Fabric Segmentation
7. Fabric Routing
8. Subnet Management Model
9. Node Identifiers
10. Virtual Lanes
11. Common Routing Algorithms
12. OpenSM

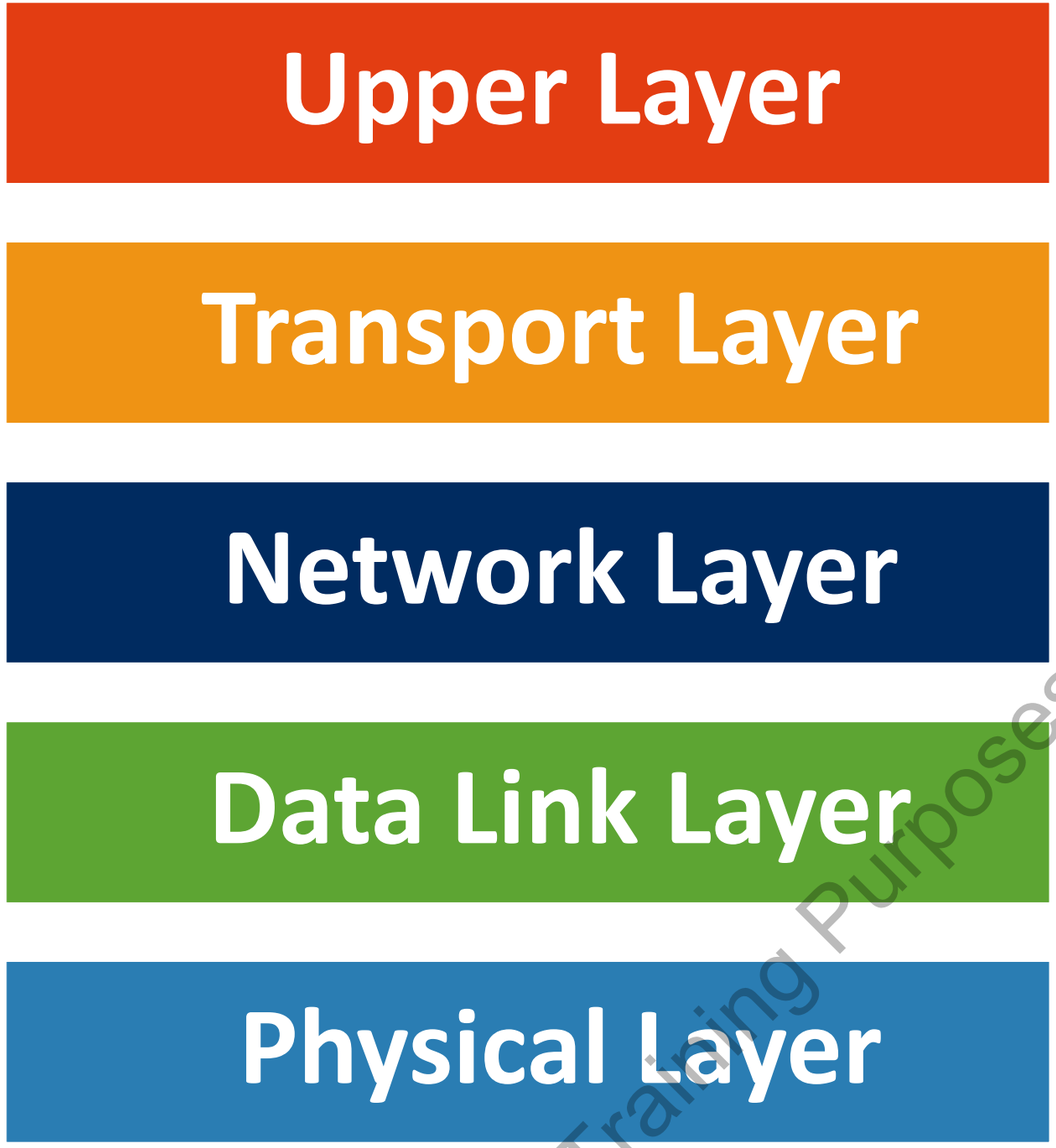
By the end of this unit, you will be able to:

- List the five main layers of InfiniBand architecture
- Describe the main responsibilities of each layer
- Identify the main mechanisms/features of each layer
- List the common IB topologies and their pros and cons
- Describe the primary addressing and routing elements
- Describe the basic management concepts

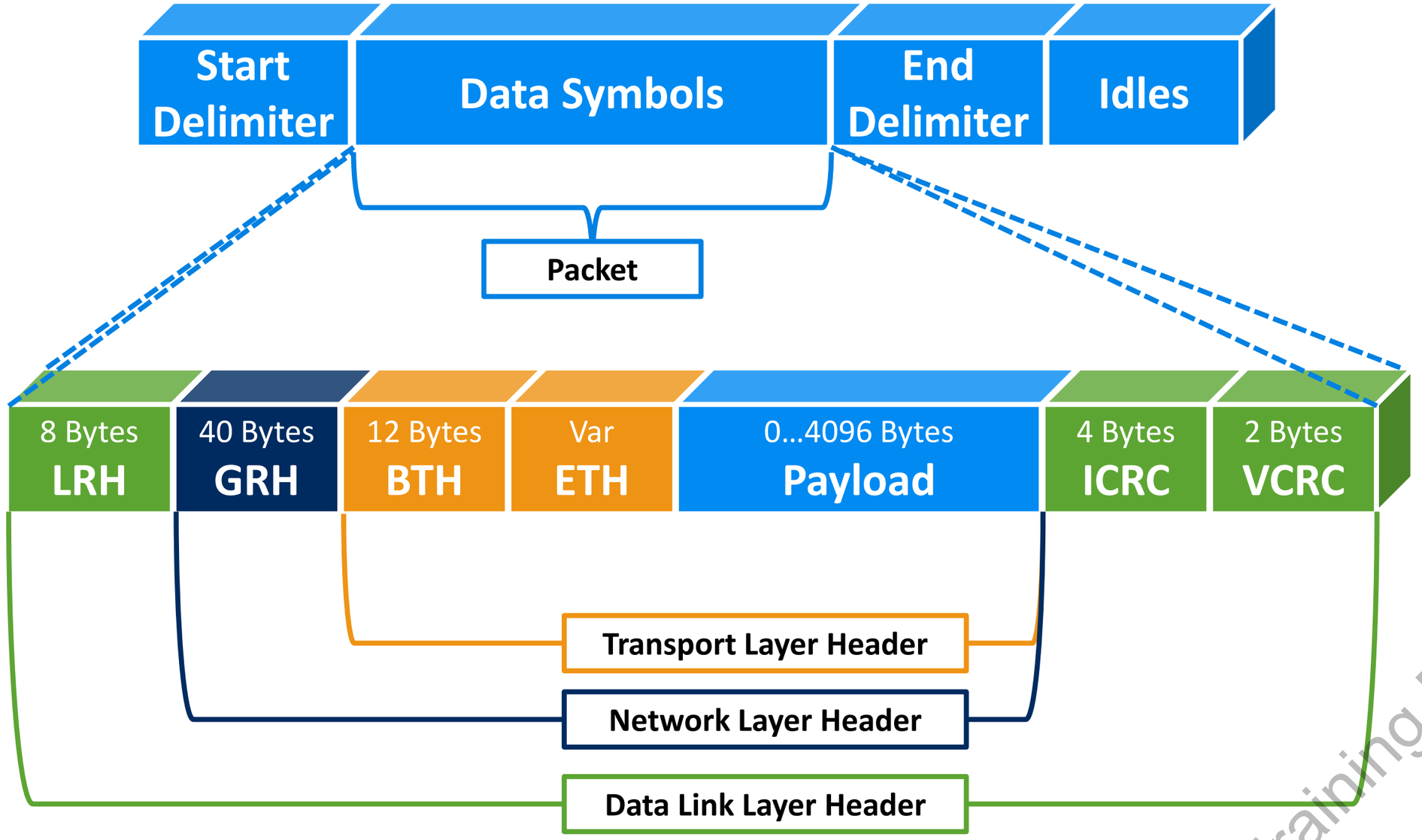
- InfiniBand uses a multi-layer processing stack to transfer data between nodes
- CPU offloads functions
- Offers greater adaptability through a variety of services and protocols



- Each layer describes its specific:
  - functions
  - Protocols
  - devices
- Each layer:
  - Provides **services** to the layer above
  - Issues **service requests** to the layer below



Data Packet Structure



Subnet Manager - Rules & Roles

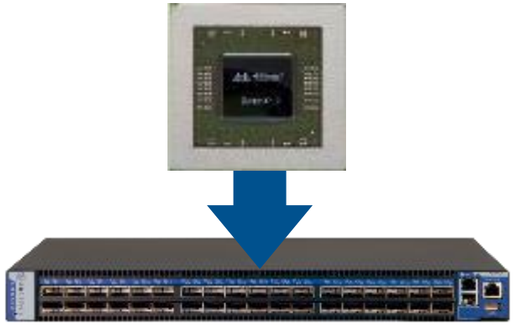
The Subnet Manager (SM):

- Managing all the elements in the fabric
- Discovering the topology
- Assigning LIDs to devices
- Calculating and programming switch forwarding tables
- Monitoring changes in subnet

- Can be implemented on any node in the fabric
- Only **one** master SM is allowed in a subnet

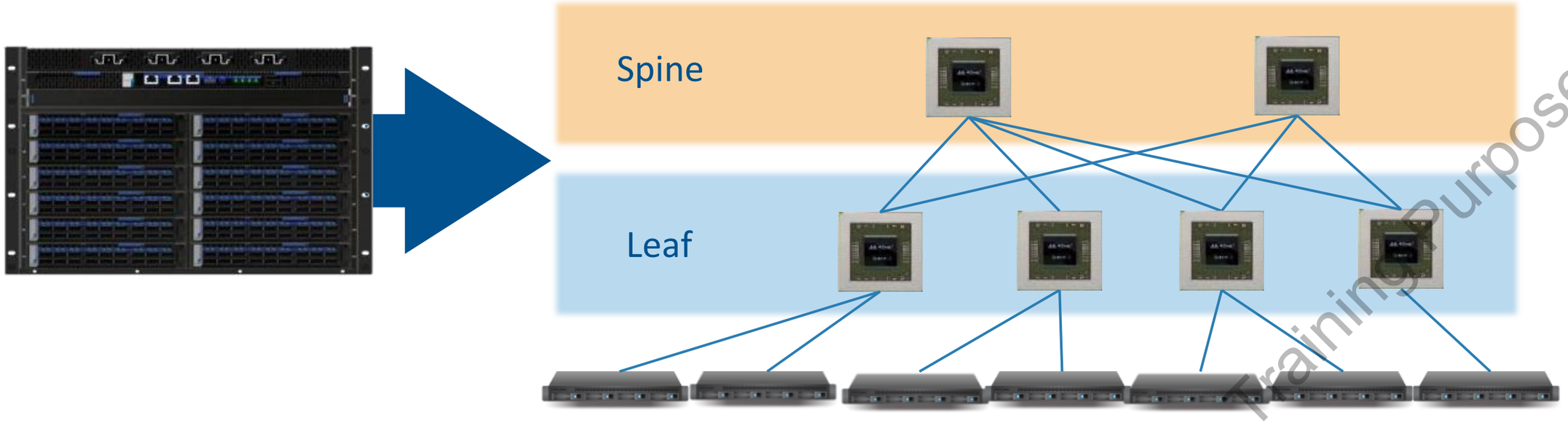
IC (Integrated Circuit) – Switch ASIC

- Switches that contain up to 36 ports (as for today) have 1 IC



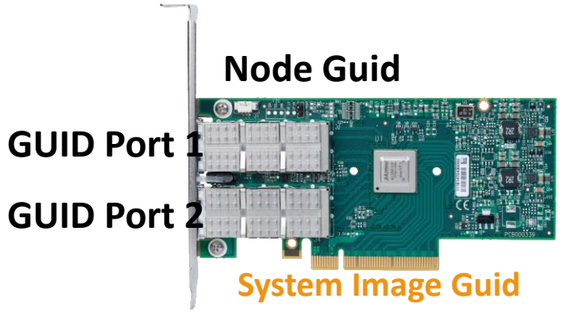
SX6036 – contains 36 ports (1 IC)

- Switches that contain more than 36 ports have several Ics
- In this Example – 72 ports switch, using 6 identical ASICs:



GUID – **G**lobally **U**nique **Id**entifier

- A 64-bit unique address that is burned onto the HW by the vendor
- Persistent through reboots
- Types of GUIDs:
  - Node GUID
  - Port GUID
  - System Image GUID
- Switches that belong to the same chassis have the same System GUID



LID – **Local Id**entifier

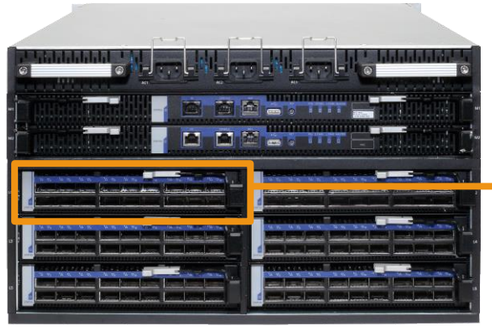
- 16-bit L2 address
- Assigned by the Subnet Manager when port becomes active
- Not persistent through reboots
  - Usually maintained as possible though (LID Cache)
- **HCAs** – each port has a LID
- **Switches**
  - All Switch Ports share the switch LID
  - Grid director switches have multiple LIDS – one for every switch in the chassis

```
[root@v-sup16 ~]# ibstat
CA 'mlx4_0'
        CA type: MT4099
        Number of ports: 2
        Firmware version: 2.10.2280
        Hardware version: 0
        Node GUID: 0x0002c90300455ec0
        System image GUID: 0x0002c90300455ec3
        Port 1:
                State: Active
                Physical state: LinkUp
                Rate: 40
                Base lid: 7
                LMC: 0
                SM lid: 15
                Capability mask: 0x02514868
                Port GUID: 0x0002c90300455ec1
                Link layer: InfiniBand
        Port 2:
                State: Down
                Physical state: Polling
                Rate: 40
                Base lid: 6
                LMC: 0
                SM lid: 1
                Capability mask: 0x02514868
                Port GUID: 0x0002c90300455ec2
                Link layer: InfiniBand
```

Node LID

All Ports Share The IC Node LID

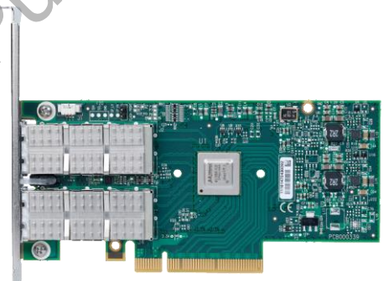
GRID Director
Multiple switches topology sharing the same chassis



Switch 1 node LID

LID Port 1

LID Port 2



GID – a 128-bit field in the Global Routing Header (GRH) used to identify a single end port or a multicast group

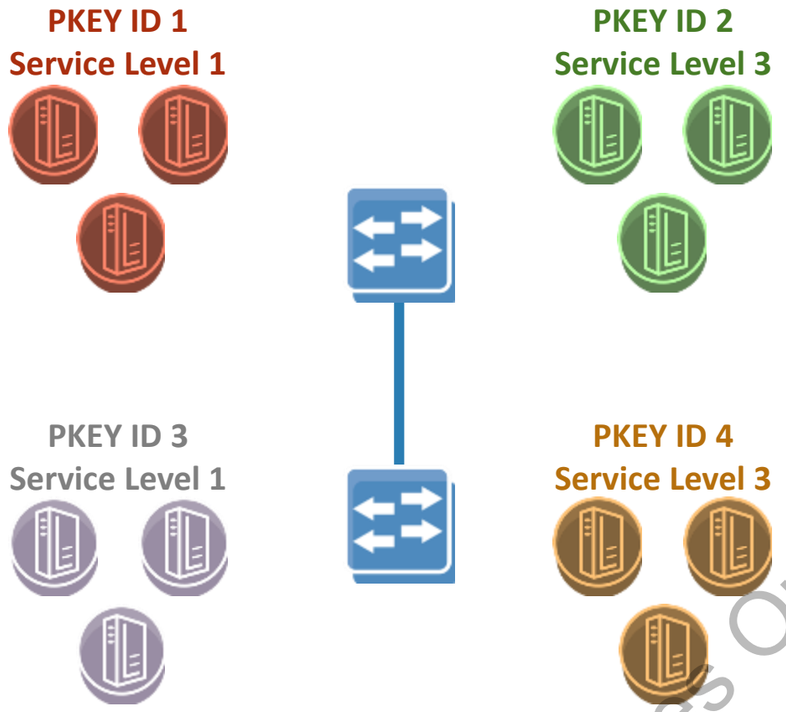
- GIDs are globally unique (across multiple subnets)
- GID's structure:
  - Based on the Port GUID combined with the Subnet Prefix: a 0- to 64-bit identifier
  - IPv6 type header

Default GID: fe80:0000:0000:0000:0002:c903:0045:5fd1

Subnet Prefix: 0-64-bit identifier

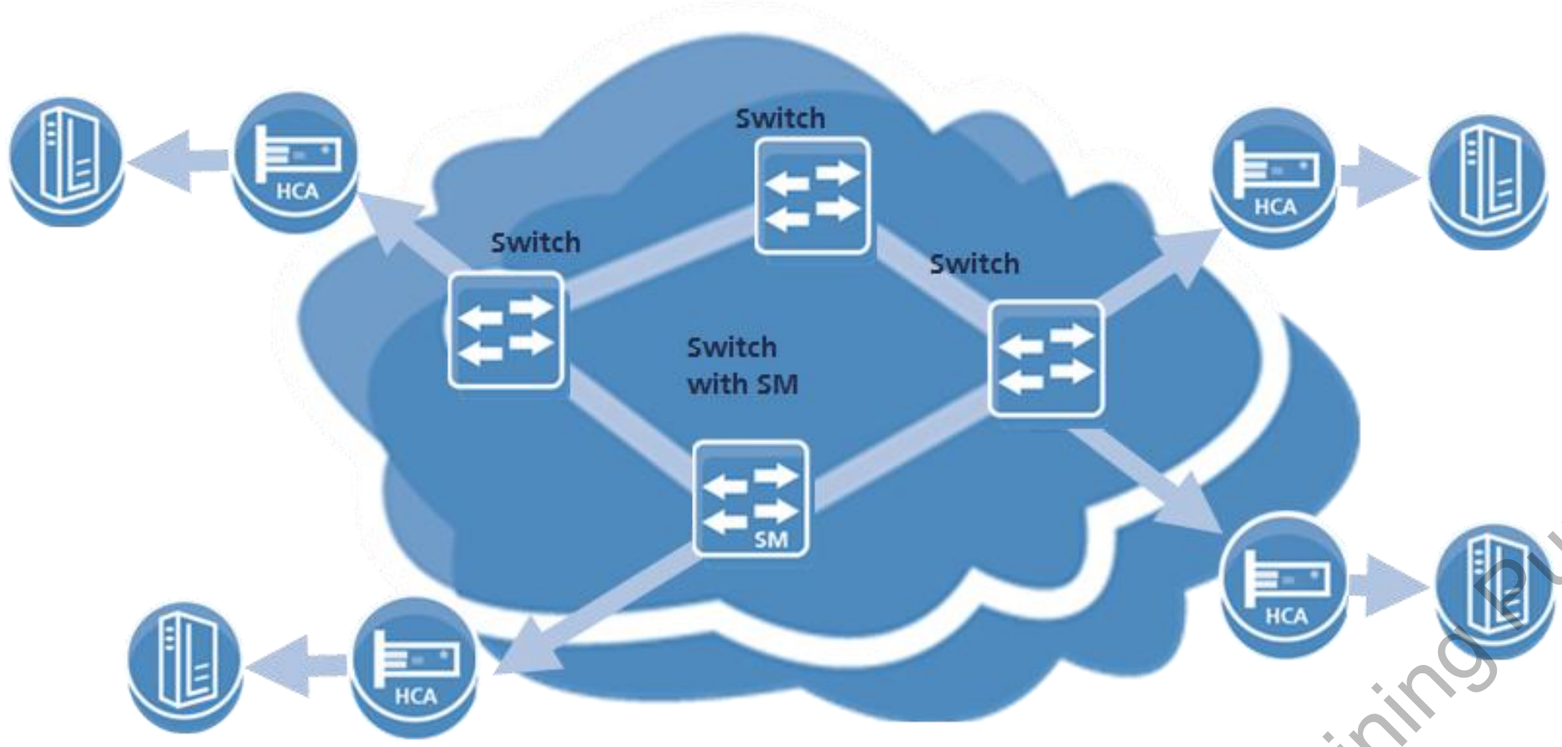
Port GUID

- Partition – describes a set of end nodes within the fabric that may communicate
  - Ports in different partitions are unaware of each other
    - Limited membership
    - Full membership
  - Ports may be members of multiple partitions at once
  - PKEY – partition identifier (a field in the BTH header)



For which purposes should you use partitioning in the fabric?

- Node
- Manager (SM)
- Agent
- Management Datagram (MAD)

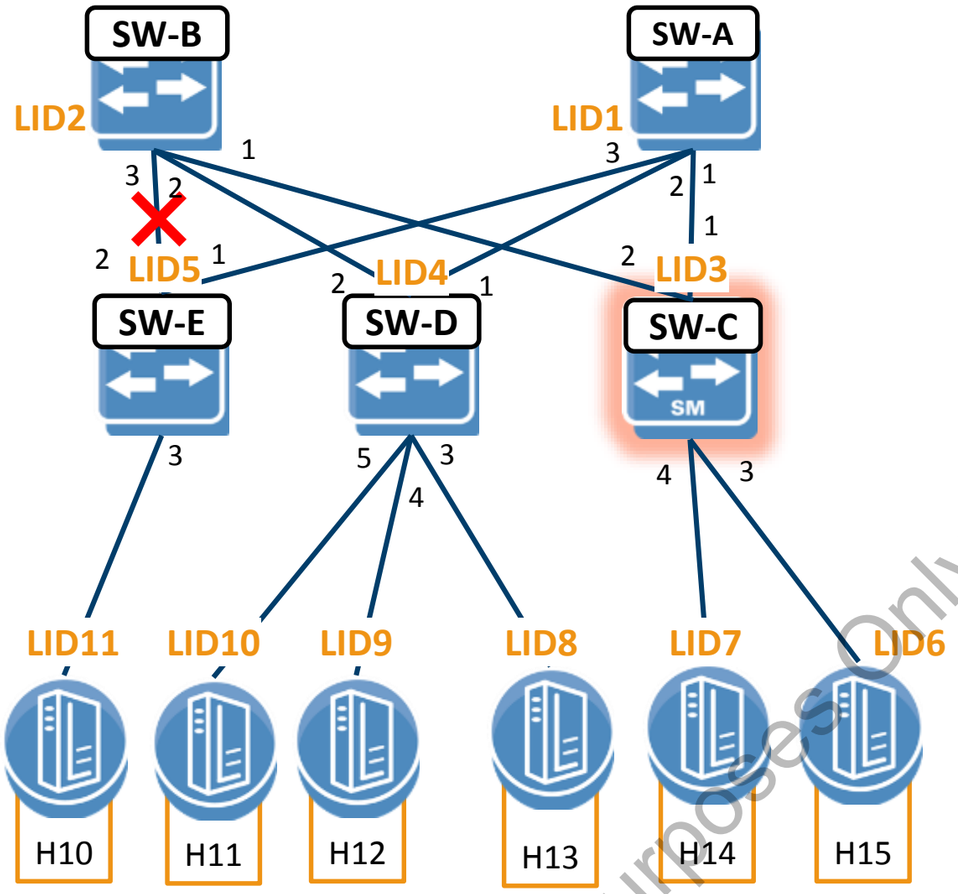


The Subnet Manager (SM)

Every IB fabric **requires** one master SM

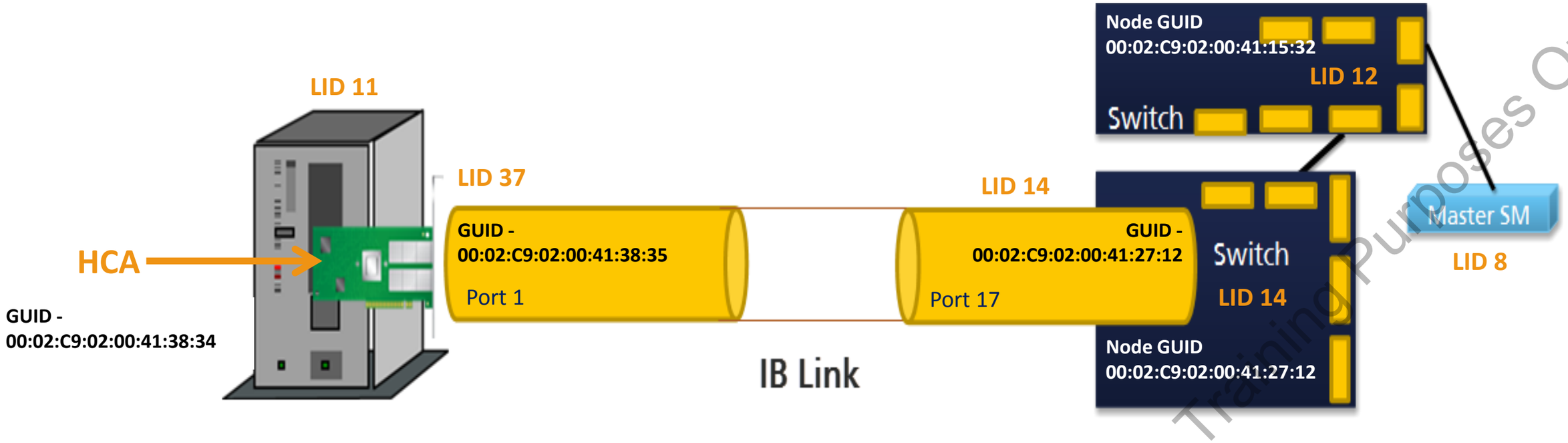
- Identifies all the nodes in the fabric
- Assigns the **local identifiers** to the nodes
- Setting the forwarding tables in the switches (Linear Forwarding Table - LFT)
- Scans for changes in the fabric and reprograms the management elements
- May run from any node
- Each node in the Fabric requires a **Subnet Manager Agent (SMA)**

Linear Forwarding Table of SW-C	
Dest. LID	Best Route/ exit port
1	1
2	2
3	0
4	1
5	1
6	3



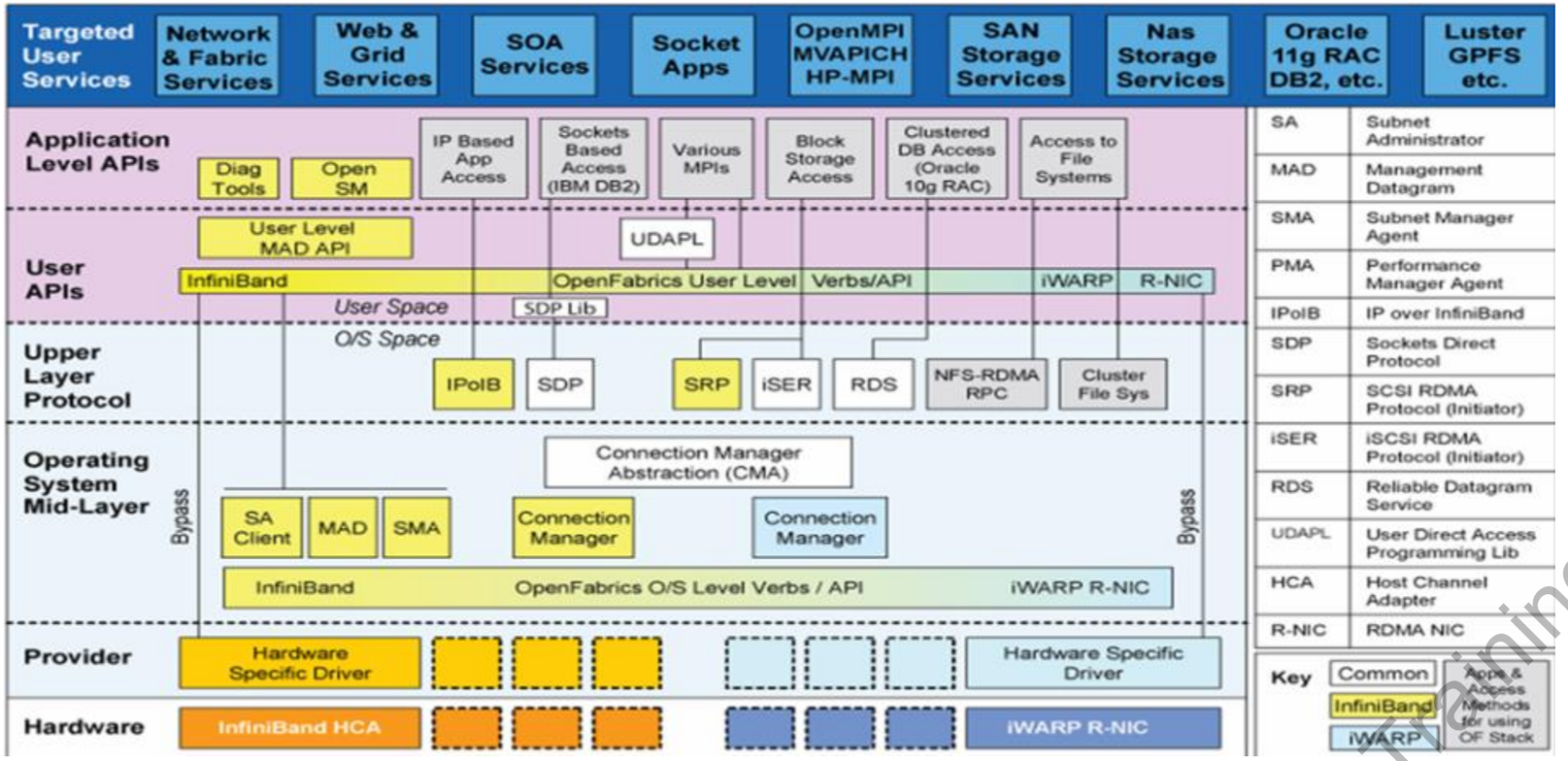
Node Identifiers – Port Basic Identifiers

- Port number
- GUID
  - HCAs have Node GUID while Switches have Switch GUID
  - Switches that belong to the same chassis have the same System GUID
- LID
  - HCAs: each port has its own LID
  - Switches: all the ports of the same module of the switch (IC) share the same LID



OFED and OFED Utilities

- Open-source software stack for RDMA and kernel bypass applications
  - Provides high performance computing sites and enterprise data centers with **flexibility** and **investment protection**
  - The OFED architecture defines a means of interaction and creates a common language between different protocols, drivers and kernels in order to establish RDMA connectivity.



OFED Utilities - Working with CLI Commands

- In order to view the full IB commands list
  - ib [tab*2]
- Help tags
  - <-h>
  - <--help>
  - <?>
- Tab key
- Up arrow key
- Right mouse click
- man command ()

```
[root@mtlacad01 ~]# ibportstate -h

Usage: ibportstate [options] <dest dr_path|lid|guid> <portnum> [<op>]

Supported ops: enable, disable, on, off, reset, speed, espeed, fdr10,
        width, query, down, arm, active, vls, mtu, lid, smlid, lmc,
        mkey, mkeylease, mkeyprot


Options:
  --config, -z <config>   use config file, default: /etc/infiniband-
diags/ibdiag.conf
  --Ca, -C <ca>           Ca name to use
  --Port, -P <port>       Ca port number to use
  --Direct, -D            use Direct address argument
  --Lid, -L               use LID address argument
  --Guid, -G              use GUID address argument
  --timeout, -t <ms>      timeout in ms
  --sm_port, -s <lid>     SM port lid
  --show_keys, -K         display security keys in output
  --m_key, -y <key>       M_Key to use in request
  --errors, -e            show send and receive errors
  --verbose, -v           increase verbosity level
  --debug, -d             raise debug level
  --help, -h              help message
  --version, -V           show version

Examples:
  ibportstate 3 1 disable                       # by lid
  ibportstate -G 0x2C9000100D051 1 enable       # by guid
  ibportstate -D 0 1                    # (query) by direct route
  ibportstate 3 1 reset                 # by lid
  ibportstate 3 1 speed 1                       # by lid
  ibportstate 3 1 width 1                       # by lid
  ibportstate -D 0 1 lid 0x1234 arm             # by direct route
```

- ibstat
  - Displays basic information obtained from the local IB driver
  - Output includes LID, SMLID, port state, link width active, and port physical state
  - Similar to the ibstatus utility but implemented as a binary rather than a script
  - Can run on a specific HCA and port:
    ibstat <hca name> <port>
    ibstat <lid> <port>

```
[root@ib-cert-sv02 ~]# ibstat
CA 'mlx4_0'
        CA type: MT26428
        Number of ports: 1
        Firmware version: 2.9.1000
        Hardware version: b0
        Node GUID: 0x0002c903004c46e0
        System image GUID: 0x0002c903004c46e3
        Port 1:
                State: Active
                Physical state: LinkUp
                Rate: 40
                Base lid: 12
                LMC: 0
                SM lid: 2
                Capability mask: 0x0251086a
                Port GUID: 0x0002c903004c46e1
                Link layer: InfiniBand
```

Basic IB Information 2

- ibstatus
  - Displays basic information obtained from the local IB driver
  - Output includes LID, SMLID, port state, link width active, and port physical state
  - Similar to ibstat but also includes GIDs
  - Can run on a specific HCA and port:
    ibstatus <hca name> <port>
    ibstatus <lid> <port>

```
[root@ib-cert-sv02 ~]# ibstatus
Infiniband device 'mlx4_0' port 1 status:
        default gid:
fe80:0000:0000:0000:0002:c903:004c:46e1
        base lid:        0xc
        sm lid:          0x2
        state:           4: ACTIVE
        phys state:      5: LinkUp
        rate:            40 Gb/sec (4X QDR)
        link_layer:      InfiniBand
```

- LID is given in Hexadecimal format

0 X 0 0 2 5

- LID is entered into the IB command in decimal format

0*4096 + 0*256 + 2*16 + 5*1 = 37

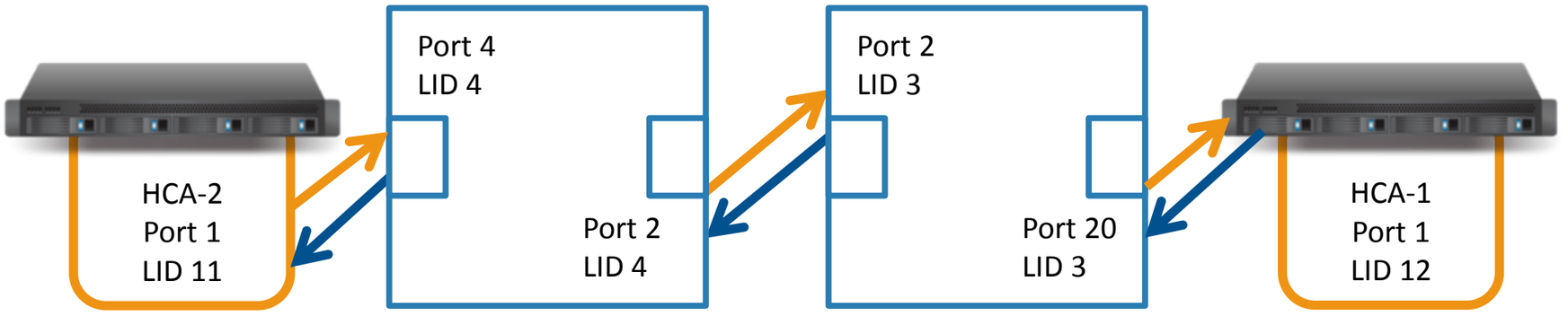
- ibping
  - Server side: ibping -S
  - Client side: ibping -L <LID>
    - Can be activated only between hosts
    - Run a simple ping-pong test over InfiniBand that measures latency
    - “Client-server" command – needs to run between 2 nodes at the same time

Server Side

```
[root@ib-cert-sv01 ~]# ibping -S
```

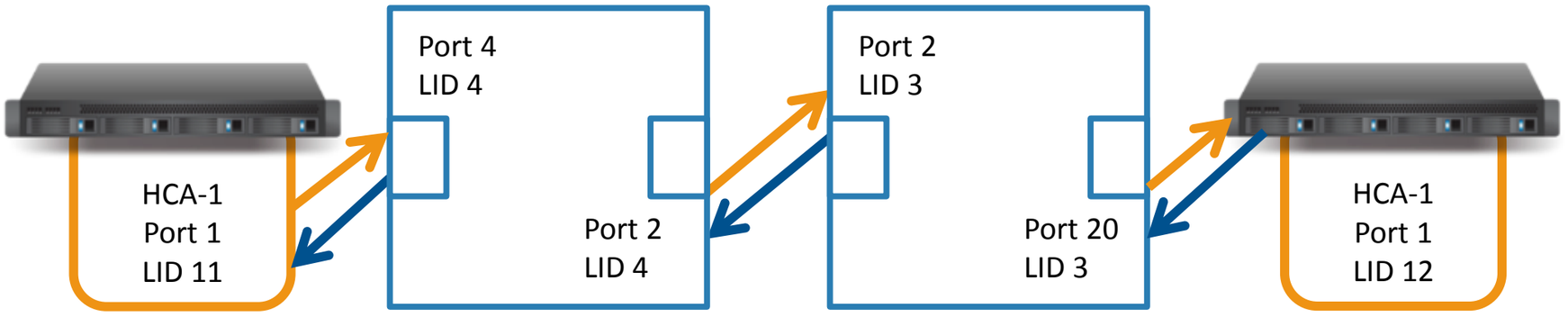
Client Side

```
[root@ib-cert-sv02 ~]# ibping -L 12
Pong from ib-cert-sv01.lab.mtl.com (Lid 11): time 0.141 ms
Pong from ib-cert-sv01.lab.mtl.com (Lid 11): time 0.085 ms
Pong from ib-cert-sv01.lab.mtl.com (Lid 11): time 0.082 ms
Pong from ib-cert-sv01.lab.mtl.com (Lid 11): time 0.056 ms
Pong from ib-cert-sv01.lab.mtl.com (Lid 11): time 0.070 ms
```



- ibtracert
  - Shows the path between 2 LIDs
  - Displays:
    - PortNum
    - GID
    - Device Name
    - LID information regarding each device in the path

```
[root@ib-cert-sv02 ~]# ibtracert 11 12
From ca {0x0002c90300e5ffe0} portnum 1 lid 11-11 "ib-cert-sv01 HCA-1"
[1] -> switch port {0x0002c90300850c71}[4] lid 4-4
"MF0;6036A17:SX6036/U1"
[2] -> switch port {0x0002c90300838481}[2] lid 3-3
"MF0;6036B19GW:SX6036/U1"
[20] -> ca port {0x0002c903004c46e1}[1] lid 12-12 "ib-cert-sv02 HCA-1"
To ca {0x0002c903004c46e0} portnum 1 lid 12-12 "ib-cert-sv02 HCA-1"
```



According to the following output please mark the correct facts.

1. This HCA has 2 ports
2. The SM lid is 0
3. Port 1 LID is 15
4. This HCA has only 1 port
5. This HCA's ports support IB and Ethernet networks
6. This HCA has 2 active ports

```
   ot@l-supp-07 ~]# ibstat
CA 'mlx4_0'
        CA type: MT26428
        Number of ports: 2
        Firmware version: 2.9.1000
        Hardware version: b0
        Node GUID: 0x0002c903004ae6fe
        System image GUID: 0x0002c903004ae701
        Port 1:
                State: Active
                Physical state: LinkUp
                Rate: 40
                Base lid: 15
                LMC: 0
                SM lid: 33
                Capability mask: 0x02510868
                Port GUID: 0x0002c903004ae6ff
                Link layer: InfiniBand
        Port 2:
                State: Down
                Physical state: Disabled
                Rate: 10
                Base lid: 0
                LMC: 0
                SM lid: 0
                Capability mask: 0x00010000
                Port GUID: 0x0202c9fffe4ae6ff
                Data Link layer: Ethernet
```

Unit Summary

- InfiniBand Network Stack
- IB Architecture Layers
  - Physical layer
  - Data Link layer
  - Network layer
  - Transport layer
  - Upper layer
- Data Packet Structure
- Data Packet Flow
- The subnet manager (SM)
- Fabric addressing
  - GUID – Globally Unique Identifier
  - LID – Local Identifier
- Fabric routing – GIDs
- Partitioning
- Management elements
- Subnet Management Model
- Node identifiers
  - Port identifiers (HCA, port number, GUID, LID)
- Virtual lanes
- Common Routing Algorithms
- OpenSM Standard Main Functionalities

Unit Summary

- Introduction to the InfiniBand Trade Association (IBTA)
- What is InfiniBand?
- Why InfiniBand?
- InfiniBand key features
  - InfiniBand Bandwidth
  - Low Latency
  - CPU Offloads
  - Simplified Management
  - Quality of Service
  - Scalability and Flexibility
- InfiniBand Fabric Components
- Common IB Network Topology Icons