OPENFABRICS ALLIANCE

15th ANNUAL WORKSHOP 2019

TO HDR AND BEYOND

Ariel Almog, Software Architect

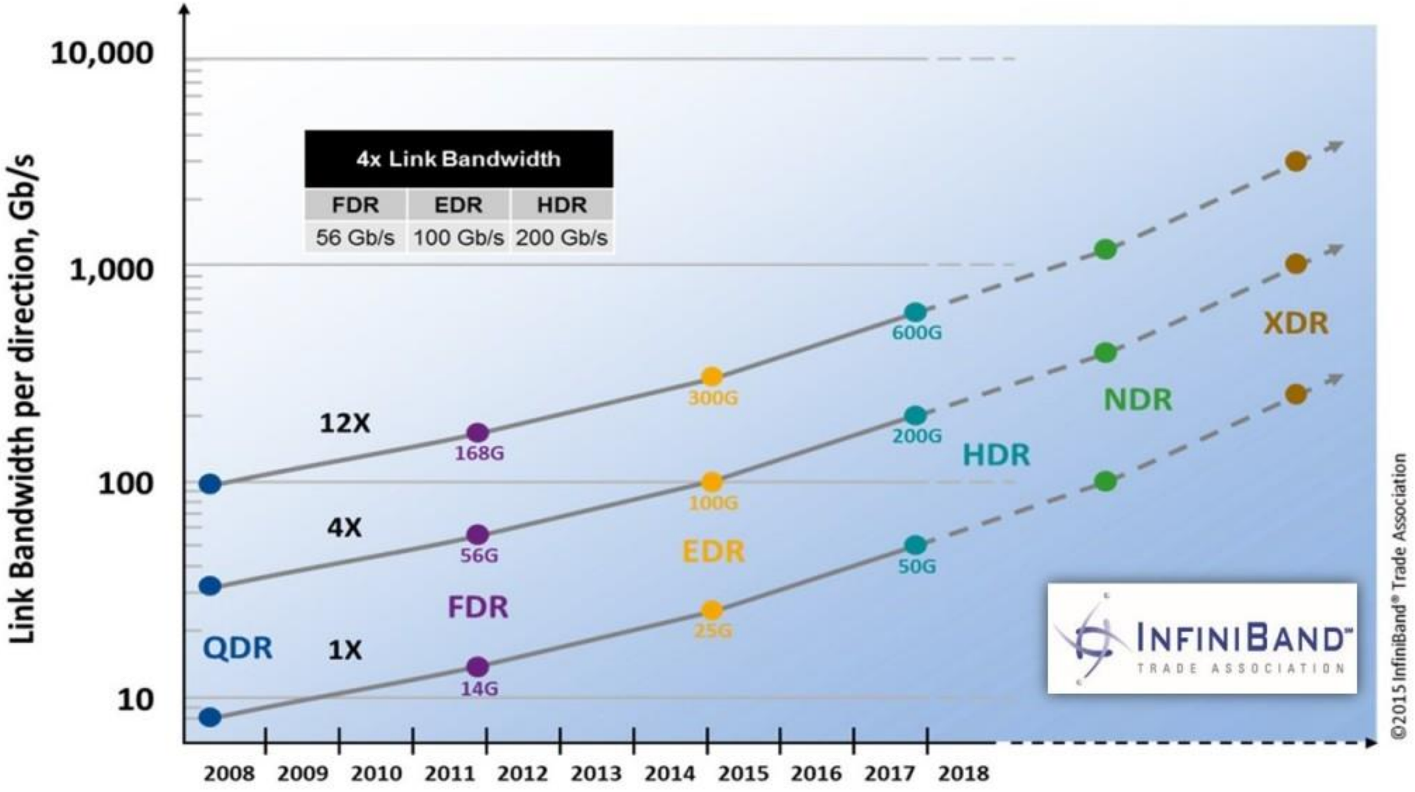
Mellanox Technologies

March, 2019

Mellanox TECHNOLOGIES

NETWORK DATA RATE IS EXPANDING

- **Network data rate is duplicated every 3 years**
  - SDR – 2.5 Gbps per lane, 2001
  - DDR – 5 Gbps per lane, 2005
  - QDR – 10 Gbps per lane, 2007
  - FDR – 14 Gbps per lane, 2011
  - EDR – 25 Gbps per lane, 2014
  - HDR – 50 Gbps per lane, 2018
- **Each new rate exposes new types of modules**



4x Link Bandwidth		
FDR	EDR	HDR
56 Gb/s	100 Gb/s	200 Gb/s

TYPES OF INTERCONNECT

Direct Attach Coax (DAC)

Copper Wires.
Directly Attaches one system to another
Key feature = Lowest Cost
Limit = 3m @ 25G rates

Optical Transceiver

Converts electrical signals to optical.
Transmits blinking laser light over optical fiber.
Key feature = long reach.
Limit = Higher cost, higher power

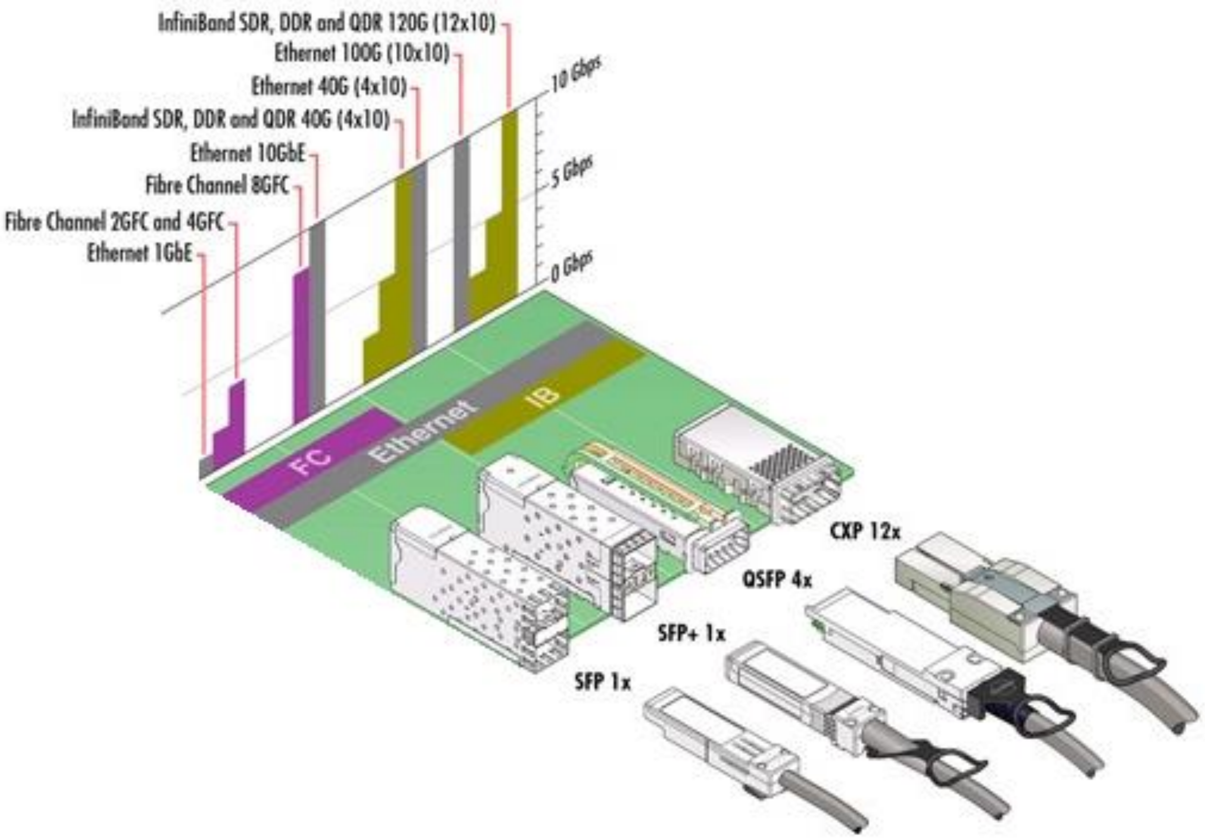
"Transceiver"
1/4/8-chennels Transmit
1/4/8-channels Receiver

Active Optical Cable

2 Transceivers with optical fiber glued in.
Key feature = Lowest Cost Optical
Limit = 100m

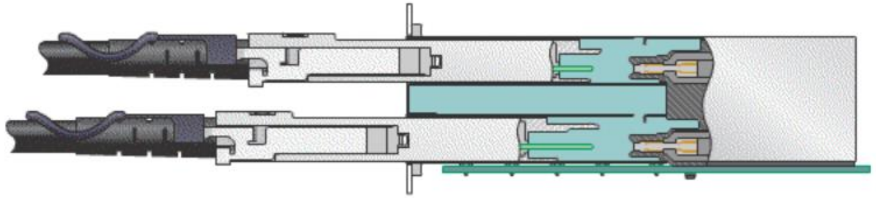
4X/1X FORM FACTORS

- **Defined by SFF Committee / MSA (Multiple**
- **QSFP – Quad Small Formfactor Pluggable**
  - 4 electrical lanes
  - QSFP28 : 25G-28G per channel
  - QSFP+ : 10G-14G per channel
- **SFP – Small Formfactor Pluggable**
  - 1 electrical lane
  - SFP28 : 25G per channel
  - SFP+ : 10G per channel



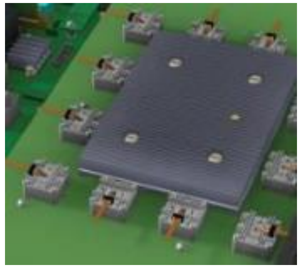
FORM FACTORS – WHAT'S NEW

- **QSFP-DD (QSFP Double Density)**
  - 8X electrical lanes connector
  - Backward compatible to QSFP modules
  - Up to 12W, 36 in 1U



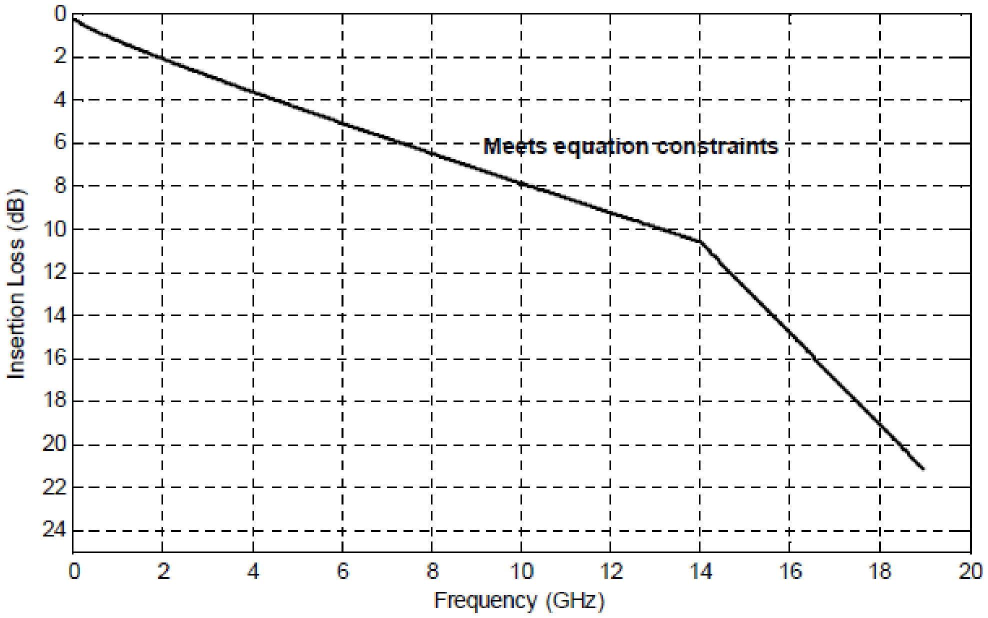
- **OSFP (Octal Small Form Factor Pluggable)**
  - 8X electrical lanes connector
  - Wider than QSFP
  - Up to 15W, 36 in 1U

- **COBO (Consortium for On-Board Optics)**
  - 8x on board optics
- **All 3 form factor target to share the same management interface**
- **QSFP-DD showed already POCs of systems, copper**
- **Other**
  - QSFP / SFP will be used also for 50G/lane (based on QSFP28/SFP28)
  - SFP-DD – recently initiated
  - uQSFP up to 5W, 4 lanes, 72 in 1U (SFP width)



COPPER CABLES – DAC (DIRECT ATTACH CABLE)

- **“Simple” copper connection between two ends of the link.**
  - No active electronics or optics – simplest construction
  - Zero power consumption – no active elements
- **Copper cables properties:**
  - **AWG (American wire gauge)**
    - 26 AWG
    - 28 AWG
    - 30 AWG
  - **Attenuation / loss**
    - In dB units per frequency
  - **Cable Length**
    - For EDR up to 5m (typical 3m)
    - For HDR / 200GE up to 3m (2m in IB spec)
- **Copper cables frequency response limits**
  - Can support rates below the max rate*

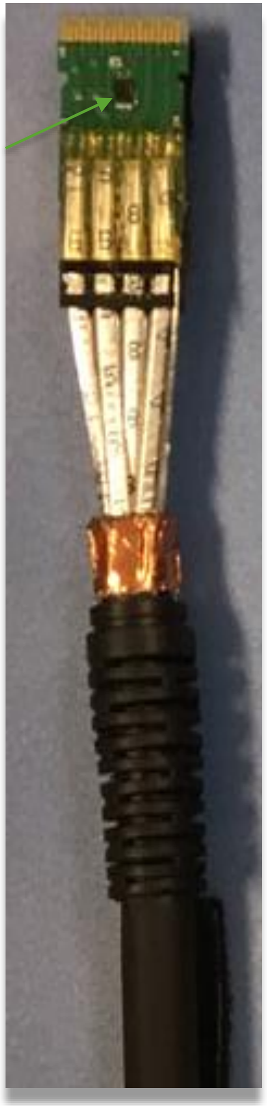


COPPER CABLES – DAC (DIRECT ATTACH CABLE)

- **Straight Copper configurations:**
  - QSFP28 <-> QSFP28
  - SFP28 <-> SFP28
  - QSFP-to-SFP port adapters & cables

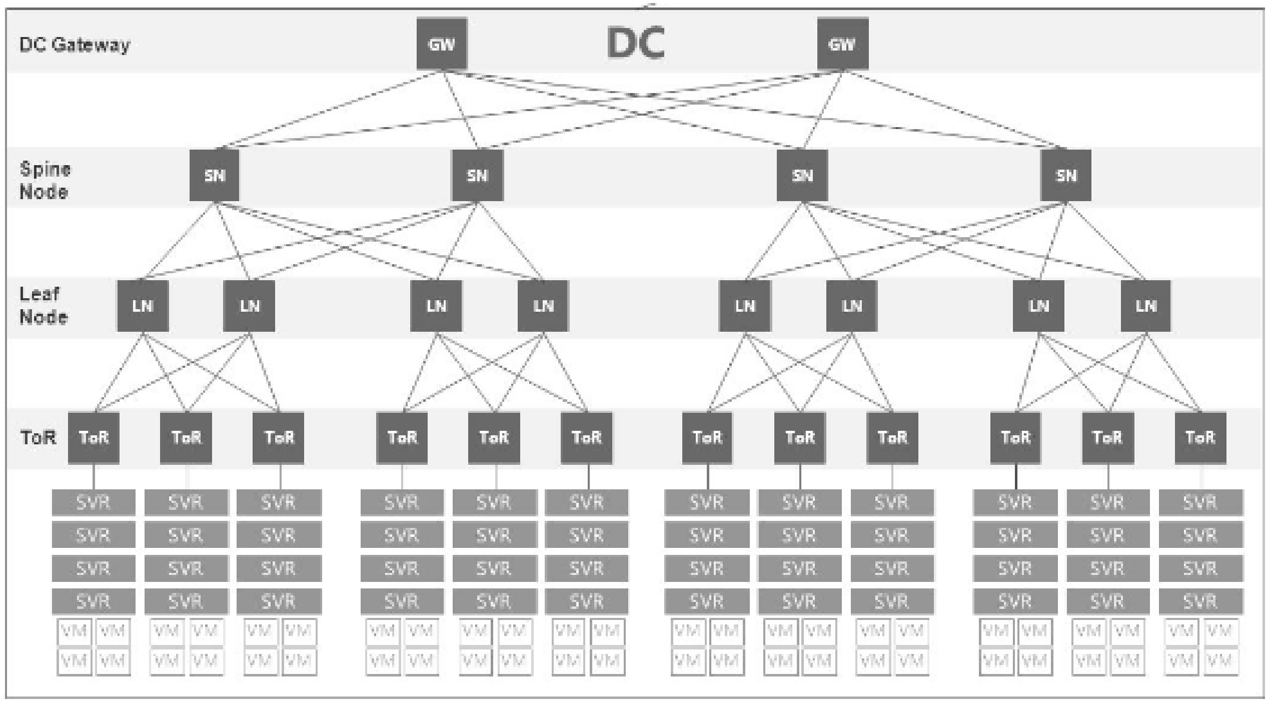
- **Splitter configurations:**
  - QSFP28 <-> 2 x QSFP28 (half populated)
  - QSFP28 <-> 4 x SFP28

- **Copper cable has an EEPROM**
  - Identification
  - Vendor Name
  - Part Number
  - Serial Number
  - Date of production
  - Max supported rate
  - Length
  - Attenuation
  - Near-End / Far-End configuration


Cable's EEPROM

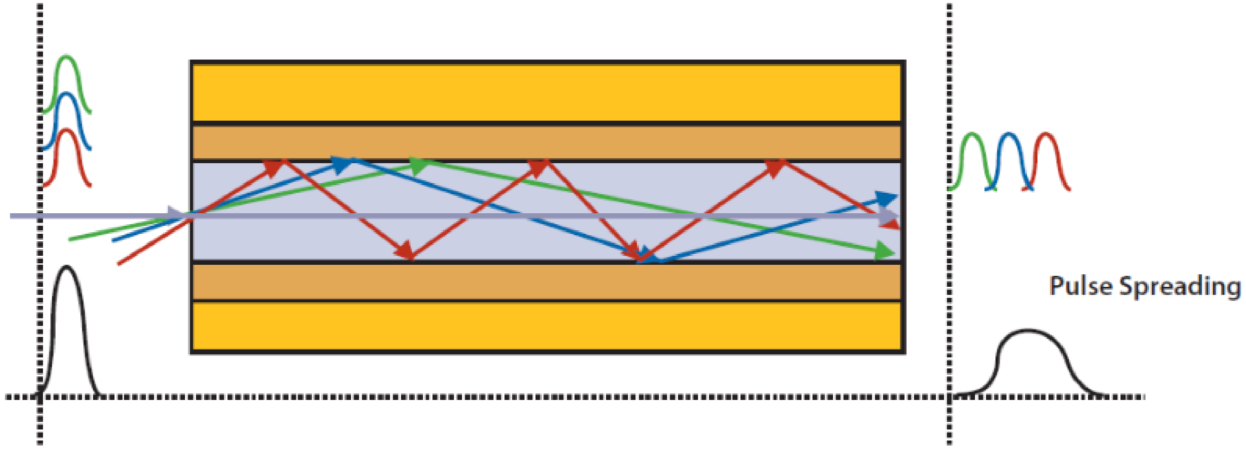
WHY IS OPTICAL COMMUNICATION NEEDED

- **Reach - Typical datacenter layout**
  - Minimum reach inside the rack (between TOR to HCA) – 3m
  - Between racks to ‘Leaf’ switch – up to 30m
  - Between Leaf switch to Spine switch – up to 100m



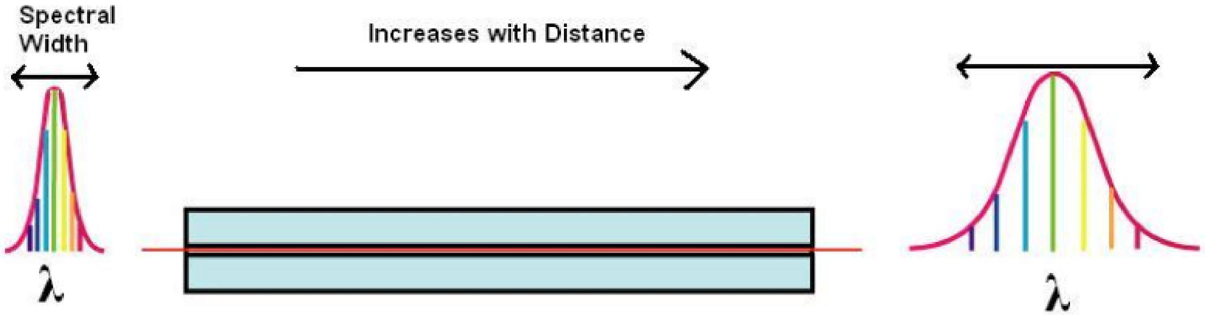
MULTIMODE VS SINGLEMODE

- **Modal dispersion – each mode propagates in a different speed through the core => Pulse spreading.**



Modal dispersion in step-index multimode fiber

- **Chromatic dispersion – light source is never monochromatic. Each color propagates at a different speed**



MULTIMODE VS SINGLEMODE

Parameter	Multimode	Singlemode
Dispersion	Modal + Chromatic	Chromatic only
Light coupling	Easy	Challenging
Fiber cost	A bit higher	A bit lower
Light source cost	Lower	Higher
Typical wavelength	850 nm	1310 nm / 1550 nm
Overall transceiver cost	Lower	Higher
Reach	100s of meters	Many KMs
WDM	Yes*	Yes

KEY PARAMETERS OF OPTICAL TRANSCEIVER

- **BER – Bit Error Rate**
  - Bit Error = deciding on ‘0’ when actually ‘1’ was transmitted (or vise versa)
  - BER = what is the ratio between bit errors and good bits
  - Typical BER requirement is 10^{-12} (1 error in every 10^{12} transmitted bits)
  - Optical Link budget defines the optical power we have to spare while keeping a minimum BER requirement.

- **FEC – Forward Error Correction**
  - Adds extra/redundant information to a transmission so that a receiver can “ recover ” from small errors
  - Today, done at the host only, not in the transceiver
  - Costs latency (more processing to do on the bit stream, even when there are no error)

ETHERNET PROTOCOLS CONVENTIONS

$$\boxed{\text{XXX}}\,\text{BASE-}\boxed{\text{M}}\boxed{\text{E}}\boxed{\text{N}}$$

- **XXX - MAC speed:**
  - 10 / 50 / 100 / 200 / 400
- **M - Media type:**
  - C - copper, K - backplane, S - MMF optics, L - 10km SMF optics, D - 500m SMF optics, F - 2km SMF
- **E - Encoding:**
  - R – 64/66 (and all new protocols), X – 8/10 encoding (1G / legacy 10G “XAUI”)
- **N – Number of physical (PMA) lanes**
  - 1 (not written) / 2 / 4 / 8
- **Examples: 100GBASE-SR4, 400GBASE-DR4, 25GBASE-CR**
- **AUIs:**
  - chip⇔module / chip⇔chip: 25GAUI, 400GAUI-8, CAUI-4
- **Most protocol names follow the above scheme, however spec wise it’s just a name.**

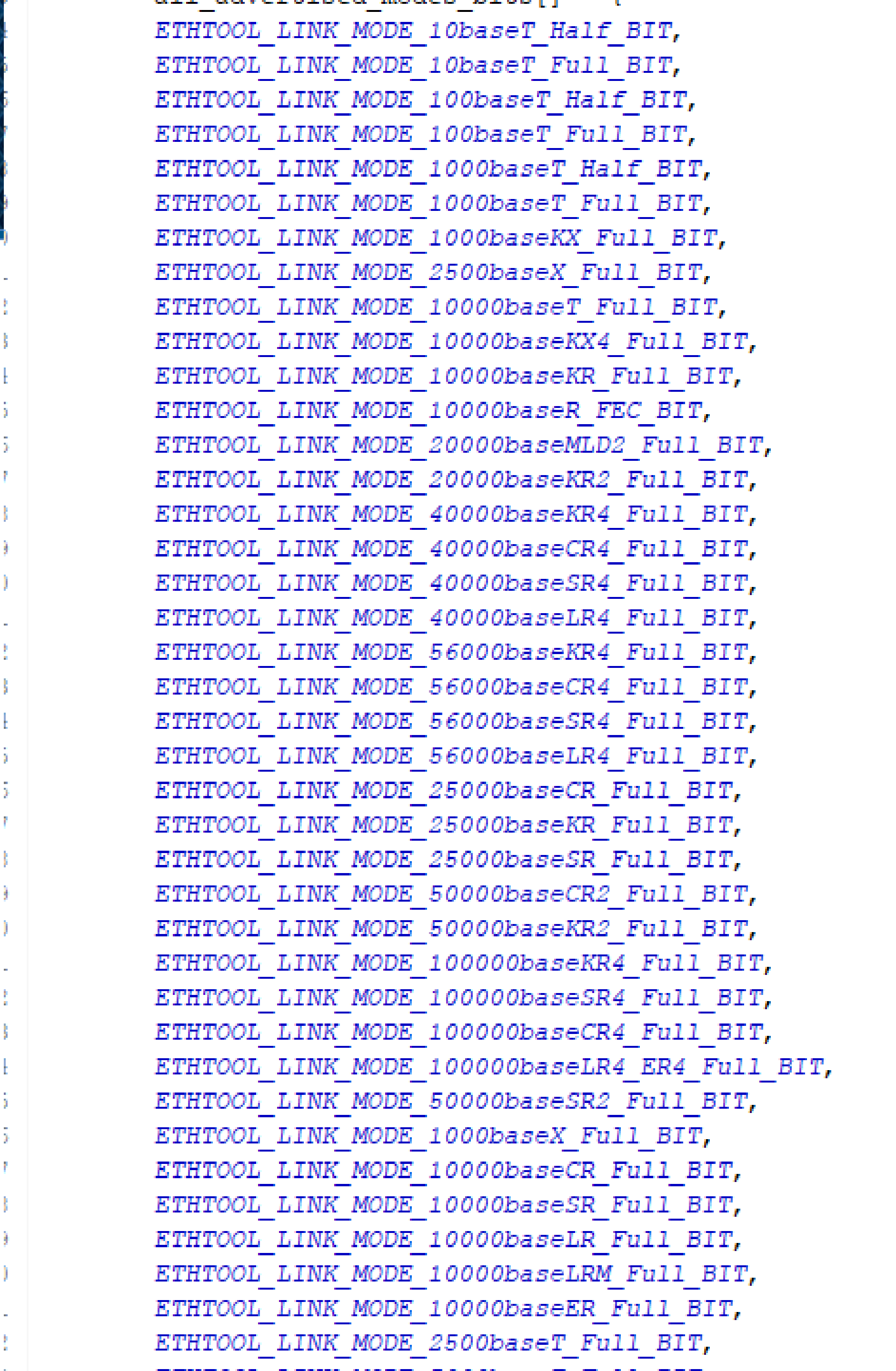
ETHERNET PROTOCOLS (BUT THERE ARE MANY MORE...)

	Phy (PMD)	Description	Number of wavelengths	Number of Fibers / Copper channels per RX/TX
Passive	50GBASE-CR	Copper up to 3m	-	1 pairs
	100GBASE-CR2		-	2 pairs
	100GBASE-CR4		-	4 pairs
	200GBASE-CR4		-	4 pairs
MMF*	50GBASE-SR	50G per λ	1	1
	100GBASE-SR2	50G per λ parallel	2	2
	200GBASE-SR4	50G per λ parallel	4	4
SMF 500m	100GBASE-DR	100G per λ	1	1
	200GBASE-DR4	50G per λ PSM4	4	4
	400GBASE-DR4	100G per λ PSM4	4	4
SMF 2km-10km	50GBASE-FR / LR	50G per λ	1	1
	200GBASE-FR4 / LR4	50G per λ WDM4	4	1
	400GBASE-FR8 / LR8	50G per λ WDM8	8	1

EXPOSING NEW MODULE IN OS

- **Linux configuration is done through ethtool**
  - Ethernet interfaces and partial support in Infiniband
- **`ethtool -s` *`devname`* `speed` *`N`* `[duplex half|full] [port tp|aui|bnc|mii] [autoneg on|off] [advertise` *`N`*`]`**
  - Current appearance shall be changed as It doesn’t scale
    - For example, for 200 Gbs (4 * 50) around 20 permutation can be found
  - AN and advertised is limited
  - Duplex
- **`--eeprom-dump`**
  - Support for e2prom parsing for new type of msa

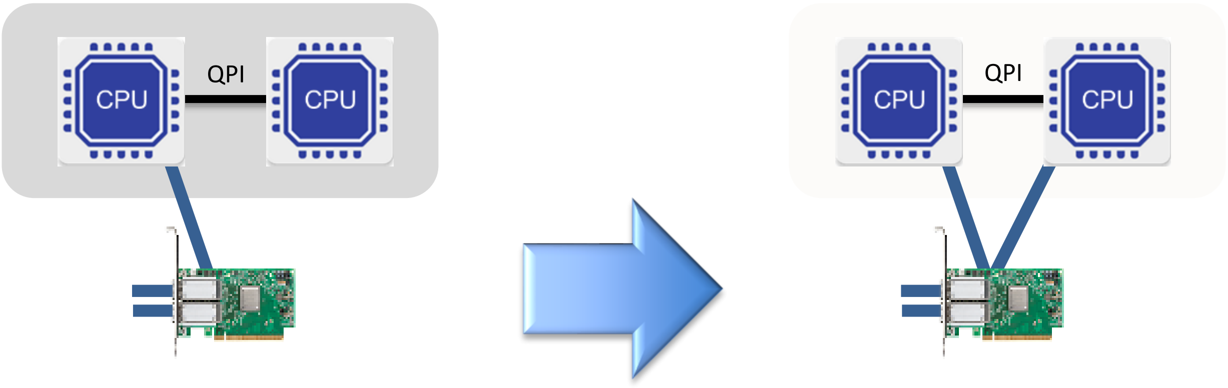
- **Other OSs (Windows, freebsd)**

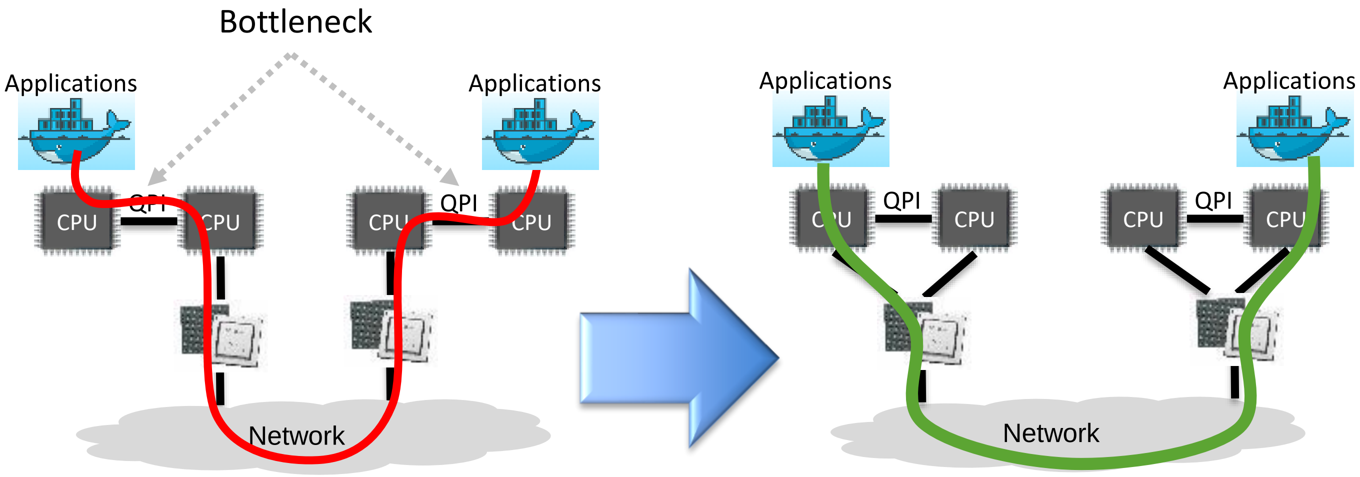


```
ETHTOOL_LINK_MODE_10baseT_Half_BIT,
ETHTOOL_LINK_MODE_10baseT_Full_BIT,
ETHTOOL_LINK_MODE_100baseT_Half_BIT,
ETHTOOL_LINK_MODE_100baseT_Full_BIT,
ETHTOOL_LINK_MODE_1000baseT_Half_BIT,
ETHTOOL_LINK_MODE_1000baseT_Full_BIT,
ETHTOOL_LINK_MODE_1000baseKX_Full_BIT,
ETHTOOL_LINK_MODE_2500baseX_Full_BIT,
ETHTOOL_LINK_MODE_10000baseT_Full_BIT,
ETHTOOL_LINK_MODE_10000baseKX4_Full_BIT,
ETHTOOL_LINK_MODE_10000baseKR_Full_BIT,
ETHTOOL_LINK_MODE_10000baseR_FEC_BIT,
ETHTOOL_LINK_MODE_20000baseMLD2_Full_BIT,
ETHTOOL_LINK_MODE_20000baseKR2_Full_BIT,
ETHTOOL_LINK_MODE_40000baseKR4_Full_BIT,
ETHTOOL_LINK_MODE_40000baseCR4_Full_BIT,
ETHTOOL_LINK_MODE_40000baseSR4_Full_BIT,
ETHTOOL_LINK_MODE_40000baseLR4_Full_BIT,
ETHTOOL_LINK_MODE_56000baseKR4_Full_BIT,
ETHTOOL_LINK_MODE_56000baseCR4_Full_BIT,
ETHTOOL_LINK_MODE_56000baseSR4_Full_BIT,
ETHTOOL_LINK_MODE_56000baseLR4_Full_BIT,
ETHTOOL_LINK_MODE_25000baseCR_Full_BIT,
ETHTOOL_LINK_MODE_25000baseKR_Full_BIT,
ETHTOOL_LINK_MODE_25000baseSR_Full_BIT,
ETHTOOL_LINK_MODE_50000baseCR2_Full_BIT,
ETHTOOL_LINK_MODE_50000baseKR2_Full_BIT,
ETHTOOL_LINK_MODE_100000baseKR4_Full_BIT,
ETHTOOL_LINK_MODE_100000baseSR4_Full_BIT,
ETHTOOL_LINK_MODE_100000baseCR4_Full_BIT,
ETHTOOL_LINK_MODE_100000baseLR4_ER4_Full_BIT,
ETHTOOL_LINK_MODE_50000baseSR2_Full_BIT,
ETHTOOL_LINK_MODE_1000baseX_Full_BIT,
ETHTOOL_LINK_MODE_10000baseCR_Full_BIT,
ETHTOOL_LINK_MODE_10000baseSR_Full_BIT,
ETHTOOL_LINK_MODE_10000baseLR_Full_BIT,
ETHTOOL_LINK_MODE_10000baseLRM_Full_BIT,
ETHTOOL_LINK_MODE_10000baseER_Full_BIT,
ETHTOOL_LINK_MODE_2500baseT_Full_BIT,
```

CONNECTX-5 SOCKET DIRECT - SYSTEM USAGE

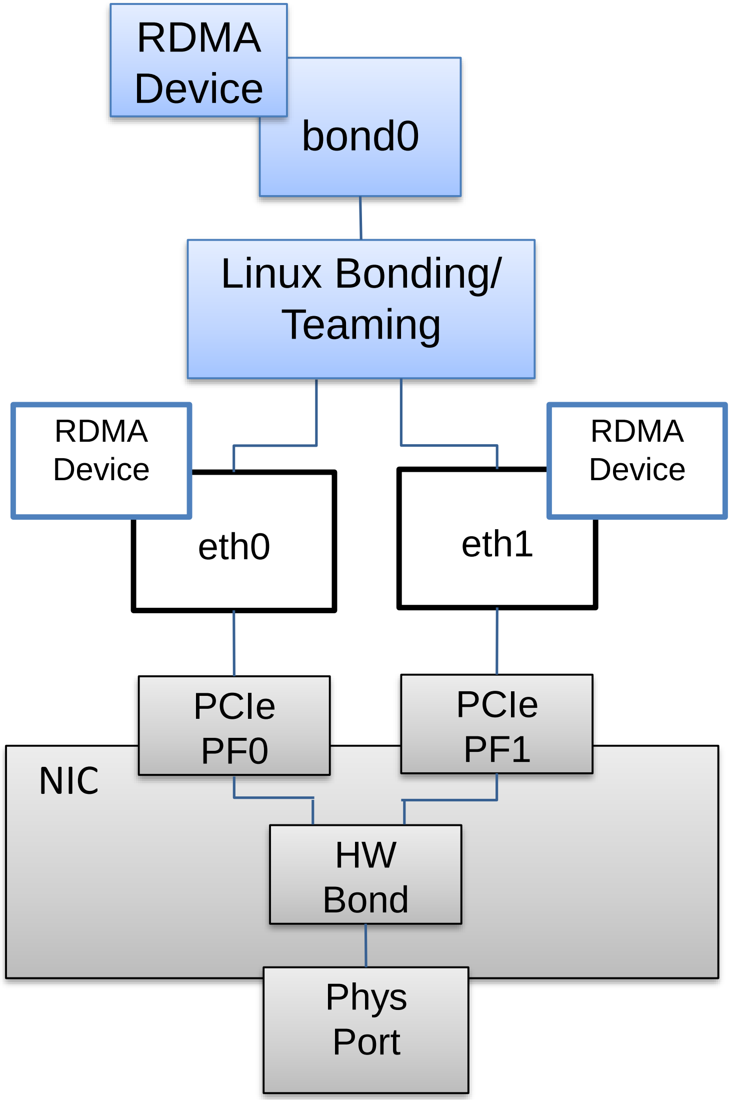
- 100Gb/s network adapters
  - Use two PCIe x8 slots
  - Adapter and PCI extender connected by harness
- Both CPU's directly connect to the network
  - Improved performance
  - Enables GPU / peer direct on both slots
- Each PCIe bus is connected through different NUMA node
- For OS, exposed as 2 or more net_device each with it's own associated RDMA device
- Application enjoy direct device to local NUMA access
- Ordering OPN
  - MCX556M-ECAT_S25
  - MCX556M-ECAT_S35A
    - With active auxiliary card



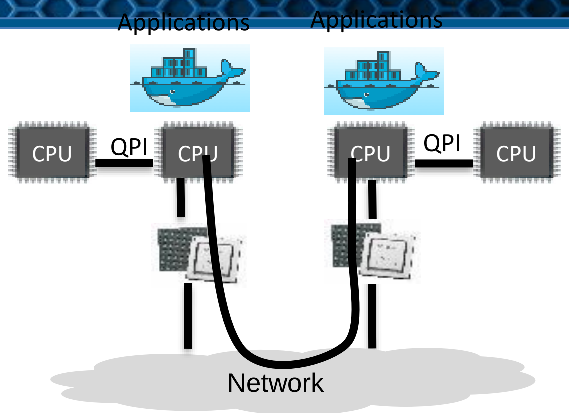


MULTI-PCI SOCKET NIC - TRANSPARENCY TO THE APP

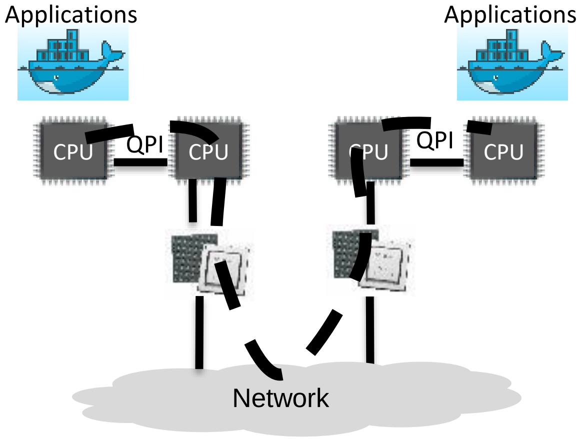
- **Application use & feel – would like to work with single net I/F**
- **Use Linux bonding with RDMA device bonding**
- **For TCP/IP traffic**
  - On TX, select slave according to TX queue affinity
  - On RX, use accelerated RFS to educate the NIC which slave to use per flow
- **For RDMA/User mode ETH (Verbs/DPDK) traffic select slave according to:**
  - Explicit - Transport object (QP) logical port create affinity attribute
  - Or transport object creation thread CPU affinity attribute
  - QPn namespace is divided across slaves
    - On receive use QPn to slave mapping
      - From BTH or from Flow Steering action
- **Don’t share HW resources (CQ, SRQ) on different CPU sockets**
  - each device has it’s own HW resources



MULTI-HOST SOCKET DIRECT^{TM} BENCHMARKS

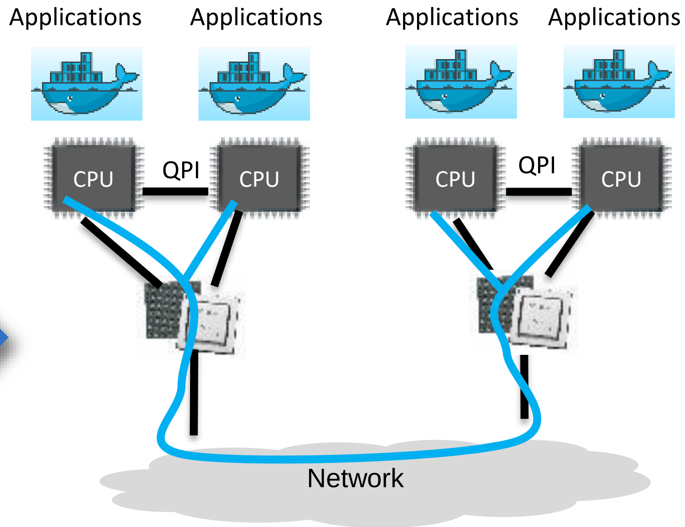


Local CPU Test Setup



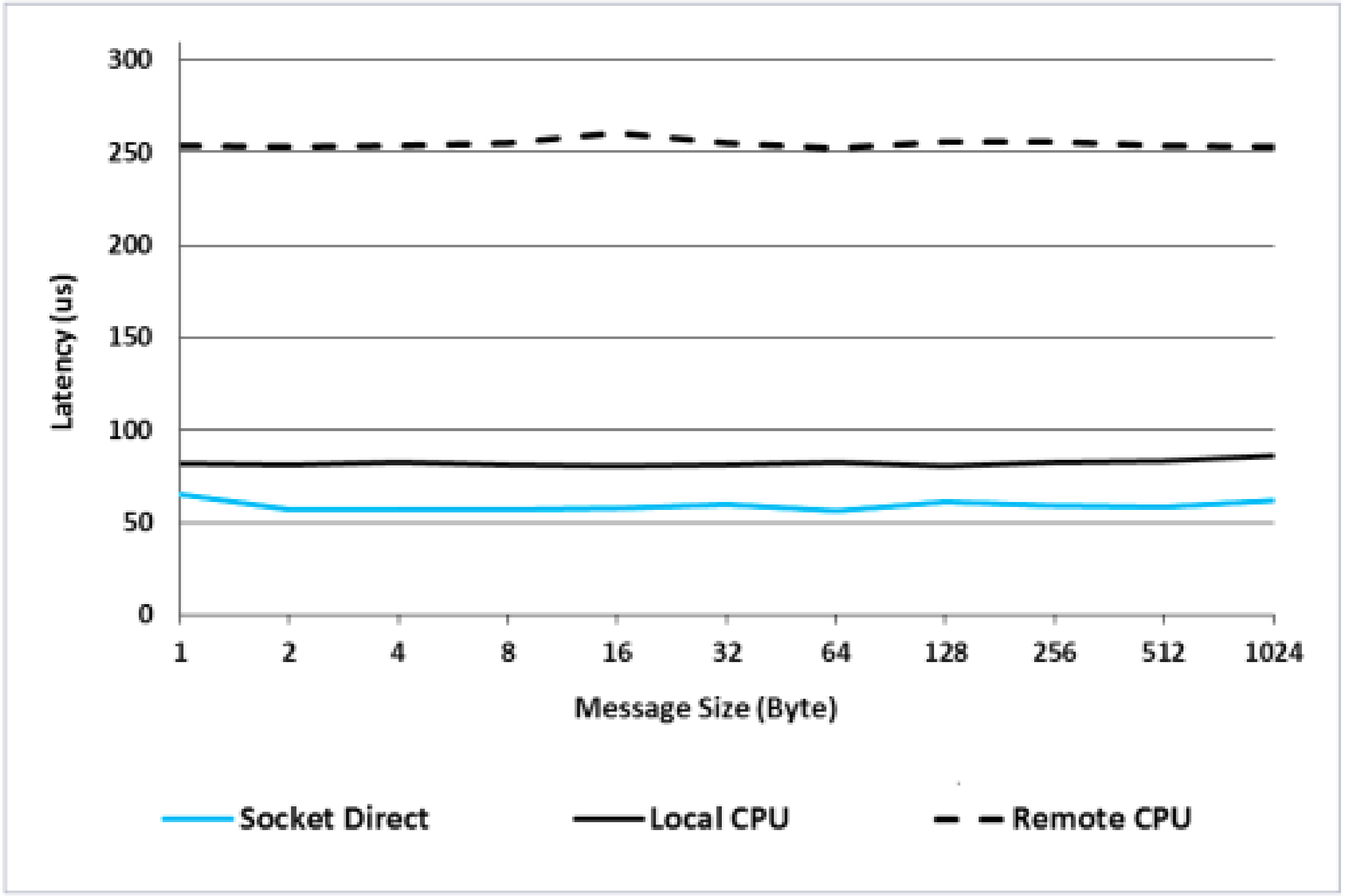
Remote CPU Test Setup

- Reduced Latency
- Reduced CPU Utilization
- Better Throughput
- Increased available QPI bandwidth



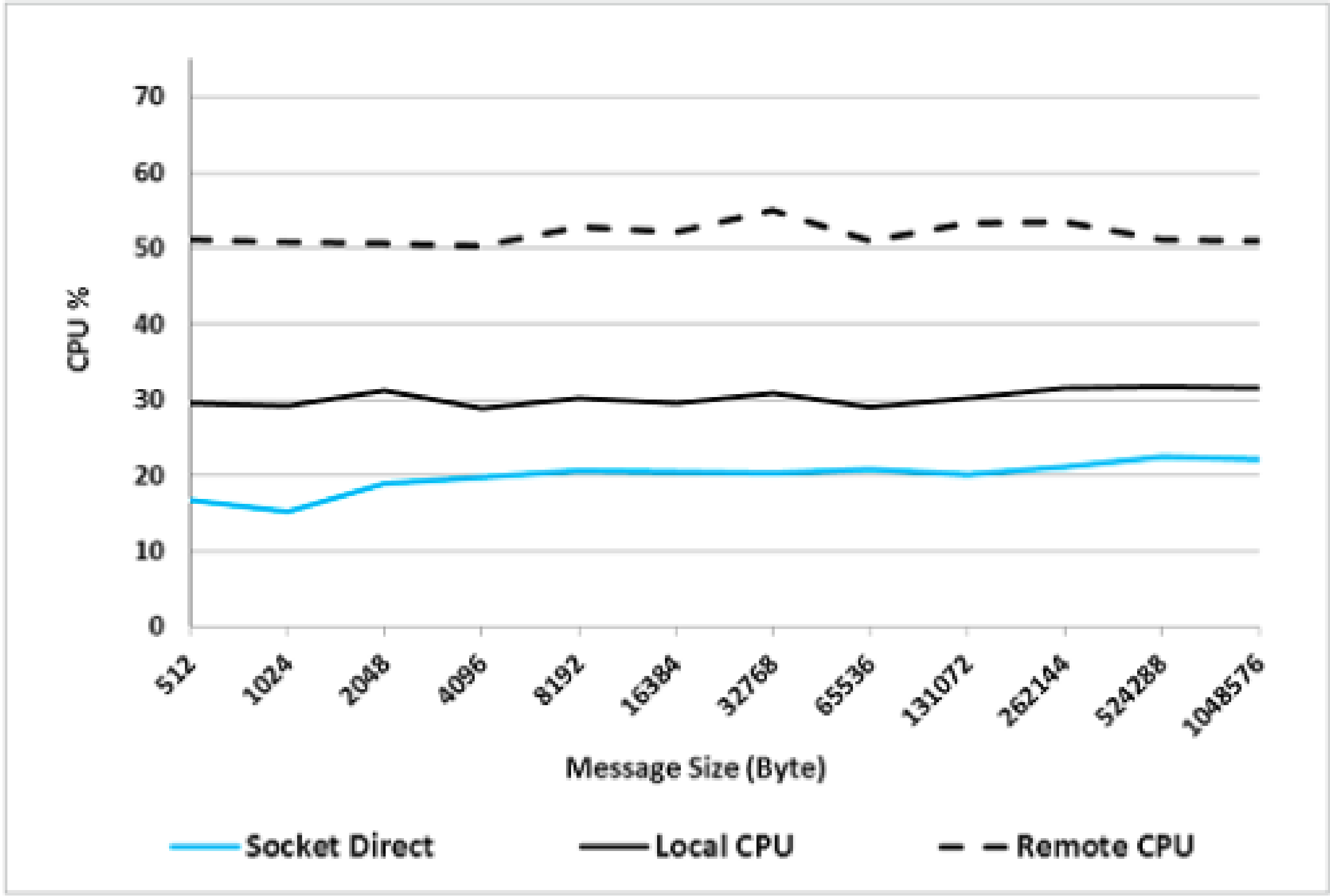
Socket Direct Test Setup

SOCKET DIRECT ADAPTER – AVERAGE LATENCY



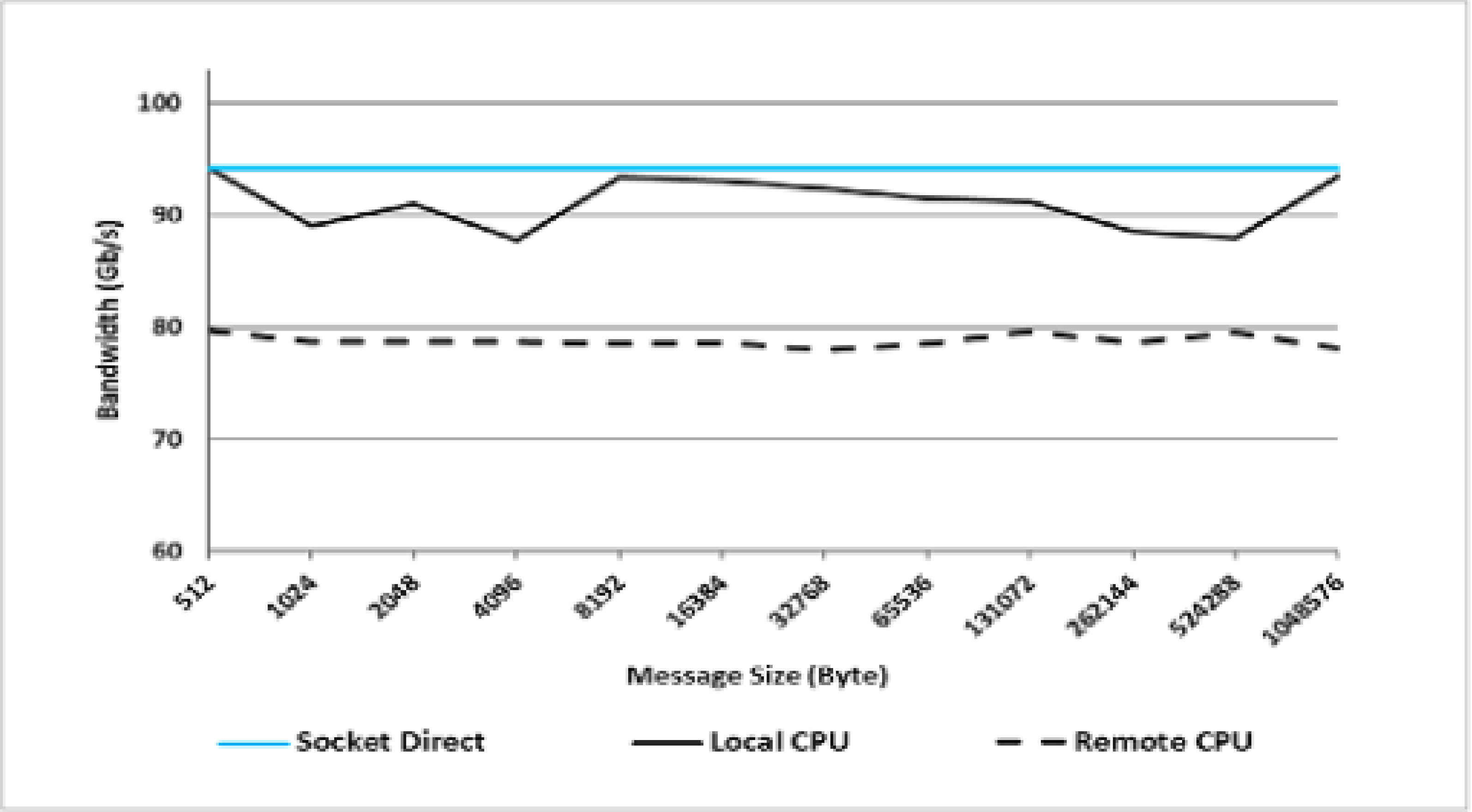
Lower is better

SOCKET DIRECT ADAPTER – CPU UTILIZATION



Up to 50% CPU Utilization Improvement

SOCKET DIRECT ADAPTER – NETWORK THROUGHPUT



Higher is better

16% Network Throughout Improvement

BENCHMARK SETUP DETAILS

Component	Description
Gen3 System	Dell PowerEdge R730
CPU	Intel(R) Xeon(R) CPU E5-2687W v4 @ 3.00GHz
Number of cores	24
Distribution name	Red_Hat_Enterprise_Linux_Server_release_7.3
Driver version	MLNX_OFED_LINUX-4.0-0.1.2.0
Firmware	12.18.1000
MTU	1500B
PCIe	Gen3
Width	x16 / x8
Mellanox adapter	ConnectX-4 MCX456A-ECAT / MCX456M-ECAT

OPENFABRICS ALLIANCE

15th ANNUAL WORKSHOP 2019

THANK YOU

Ariel Almog, Software Architect

Mellanox Technologies

Mellanox TECHNOLOGIES