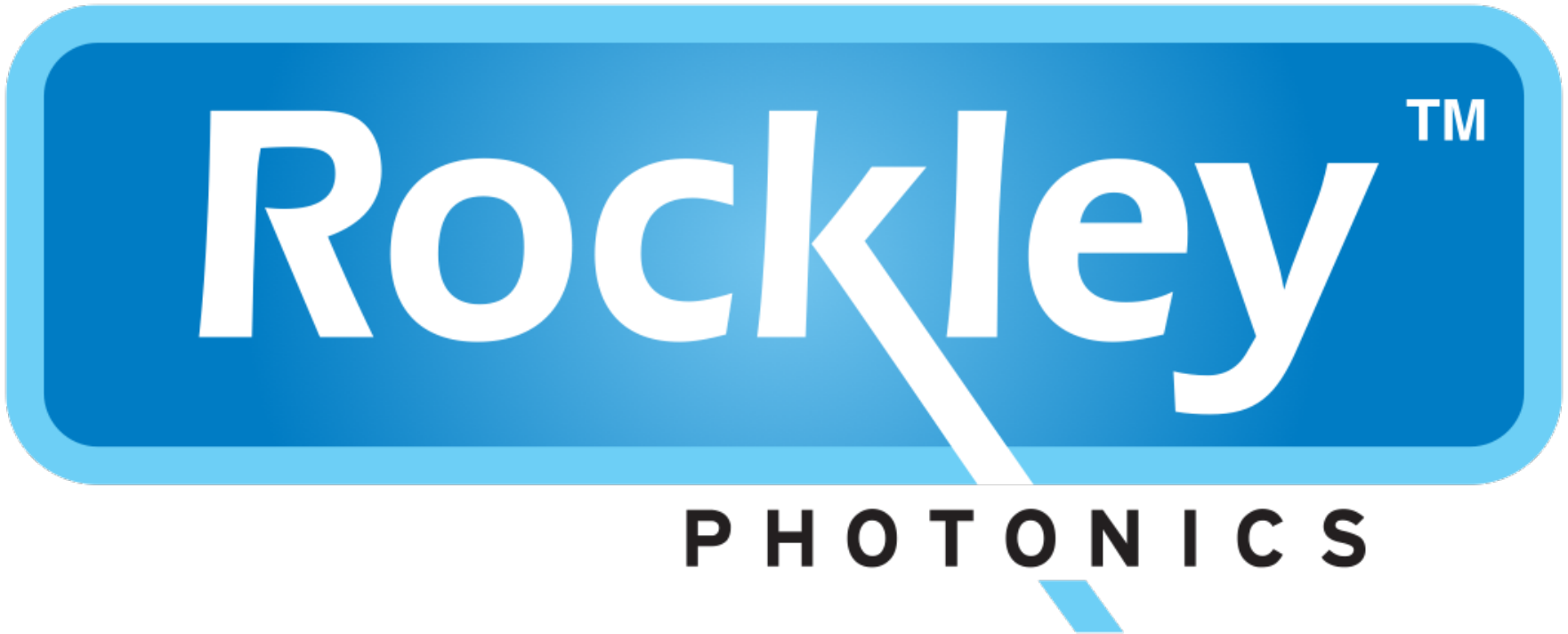

Rockley™
PHOTONICS

HiPINEB 2018
Panel Discussion

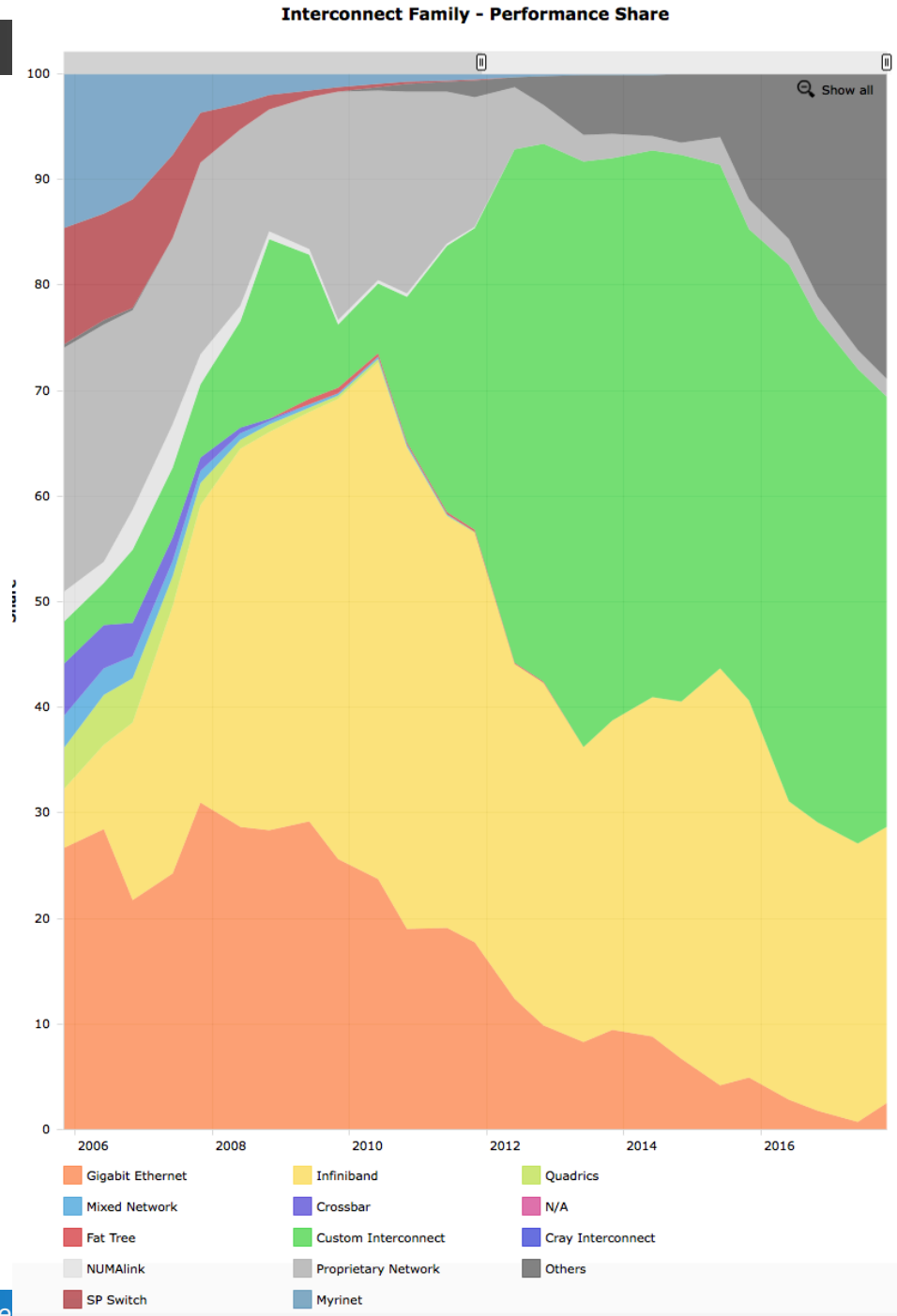
Cyriel Minkenberg

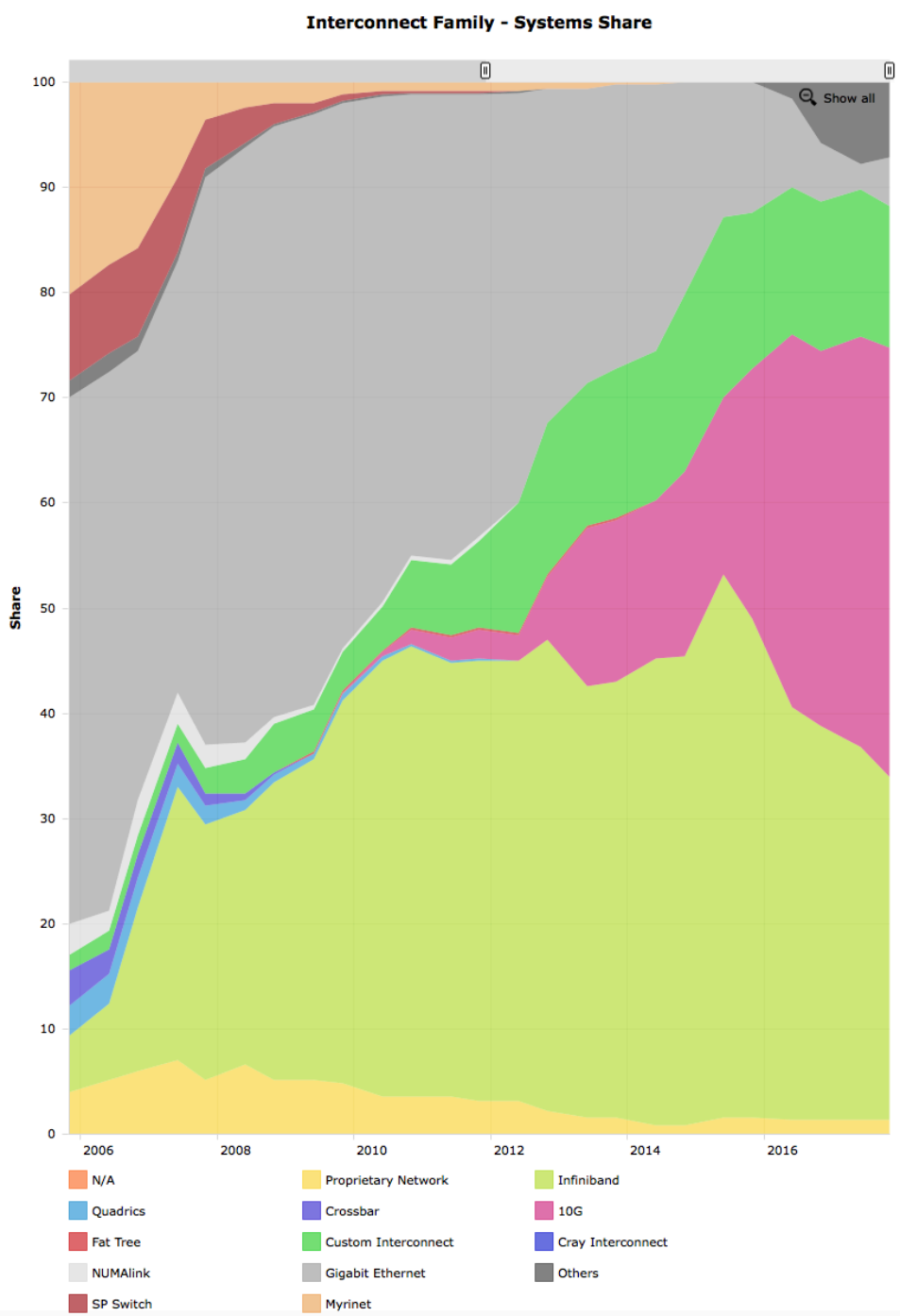
Rockley Photonics

24 February 2018

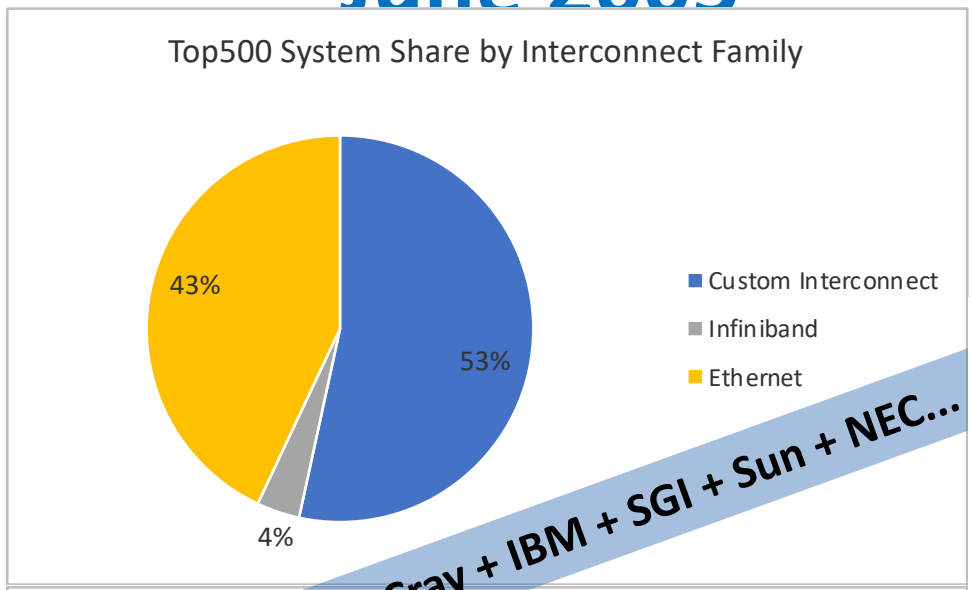
Question 1

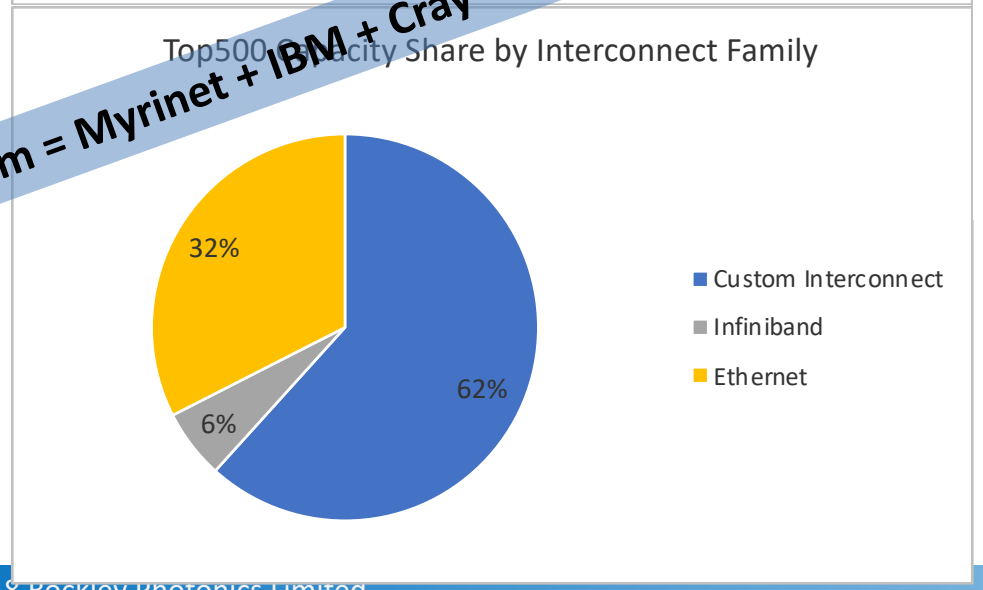
1. About twenty years ago, high-performance networking appliances for high-performance computers featured higher bandwidth and lower latency over standardized appliances. For this reason, top supercomputers of the time preferred these propriety networks over Ethernet. Is it the same today as it was then?





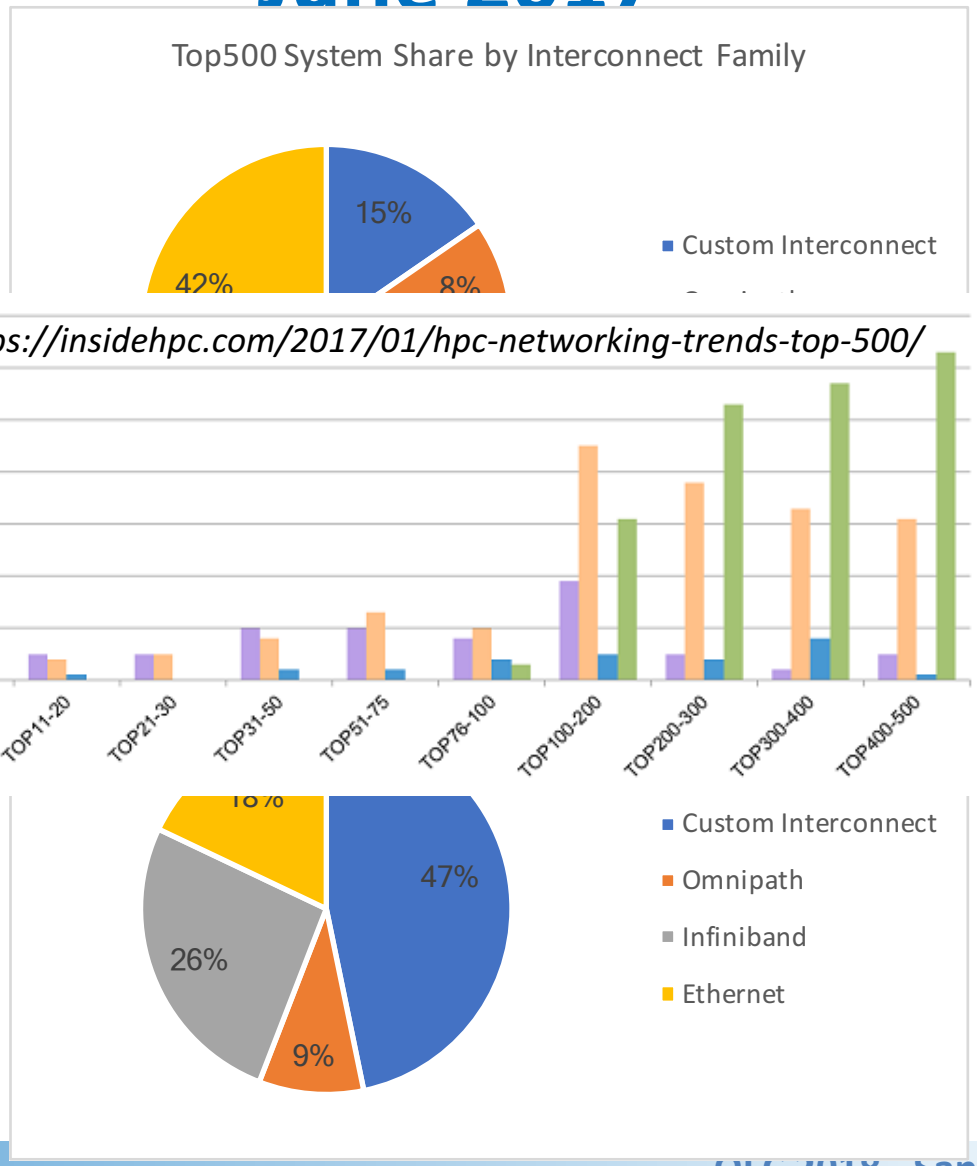
June 2005





Custom = Myrinet + IBM + Cray + IBM + SGI + Sun + NEC...

June 2017



Based o

n top500.org

Diversity dwindling into IB/Ethernet duopoly

- Infiniband has cut heavily into custom, and continues to do so
  - IBM moving to IB
  - Omni-Path is IB-ish
  - Tianhe is IB-ish (?)
- Only remaining true custom interconnect are coming from Cray
- Ethernet
  - Still mopping up the low end
  - High-end Ethernet not competitive with Infiniband in $/Gbps
  - Lacks basic features to make it suitable as a high-end HPC interconnect (Proper link-level flow control, Proper congestion management, Virtual channels, Direct-index routing tables, Adaptive routing & load balancing, Source routing, Valiant routing, ...)

Is it the same as 12 years ago?

- No
  - Custom interconnect share & diversity has tanked
  - IB has displaced most custom interconnects

- Yes
  - IB is basically a single-vendor standard
  - High-end HPC still has specific requirements that Ethernet does not meet

Question 2

- Recently, there is a lot of excitement around machine-learning and artificial intelligence applications running on high-performance computing platforms. How does this trend impact the use and the architecture of high-performance computer networks?

Global byte/FLOP ratio is falling off a cliff

- Consider e.g. IBM Summit
  - Compute: 40 TFLOP per node
  - ~4,600 nodes
  - Network: 200 Gbps = 25 GBps per node
  - Bytes/FLOP < 0.000625

- NVIDIA DGX-1
  - 1 PFLOP (FP16)
  - 4x IB EDR = 400 Gbps = 50 GB/s
  - Byte/FLOP = 0.00005 (FP16)

Source: https://www.olcf.ornl.gov

Source: nvidia.com

SYSTEM SPECIFICATIONS

GPUs	8X Tesla V100	8X Tesla P100
Performance (GPU FP16)	1 petaFLOPS	170 teraFLOPS
GPU Memory	128 GB total system	
CPU	Dual 20-Core Intel Xeon E5-2698 v4 2.2 GHz	
NVIDIA CUDA® Cores	40,960	28,672
NVIDIA Tensor Cores (on V100 based systems)	5,120	N/A
Maximum Power Requirements	3,200 W	
System Memory	512 GB 2,133 MHz DDR4 LRDIMM	
Storage	4X 1.92 TB SSD RAID 0	
Network	Dual 10 GbE, 4 IB EDR	
Software	Ubuntu Linux Host OS See Software Stack for Details	
System Weight	134 lbs	
System Dimensions	866 D x 444 W x 131 H (mm)	
Packing Dimensions	1,180 D x 730 W x 284 H (mm)	
Operating Temperature Range	10–35 °C	

Driven by network cost

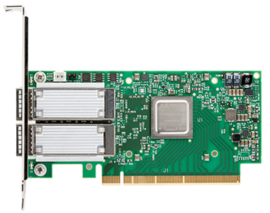
- Assume network costs only $1/Gbps: 0.1 B/F implies $32,000 per node
- ~$150M just for the network
- Network actually costs >> $1/Gbps...
- Should be closer to $0.25/Gbps (including NIC, cables, switches, everything)

Cost estimate

- Assume full bisection IB Folded Clos with radix 36
- Requires three tiers
  - Per node: 1 HCA + 5 switch ports + 1 DAC + 2 AOCs
  - 1 x 100G HCA port = $450
  - 5 x 100G switch port = 5x $350 = $1,750
  - 1 x 100G DAC = $100
  - 2 x 100G AOC = 2 x $500 = $1,000
  - Times x2 rails
  - Total $\underline{\$6{,}600}$ per node ($33/Gbps)
  - Times x4,600 nodes ≥ **$\underline{\$30{,}000{,}000}$**

Mellanox MCX456A-ECAT ConnectX-4 VPI Adapter Card EDR IB (100Gb/s) Port QSFP28 PCIe3.0 x16 ROHS R6

Condition: New MPN: MCX456A-ECAT



Availability:	Ships same day
List Price:	~~$1,509.00~~
Our Price:	$920.00
Savings:	(You save $589.00)
Technology:	InfiniBand
Speed:	EDR & 100GbE
Ports:	Dual
Connector Type:	QSFP28
PCI:	PCIe3.0 x16
Condition:	New
ECCN:	5A991

Download MCX456A-ECAT Product Brief

Sources: store.mellanox.com, fs.com

Mellanox MSB7790-EB2F Switch-IB EDR InfiniBand 1U Switch 36 QSFP28 Ports 2 Power Supplies (AC) Unmanaged Short depth P2C Airflow Rail Kit RoHS6

Condition: New MPN: MSB7790-EB2F

Availability:	In Stock
List Price:	~~$21,328.00~~
Our Price:	$13,005.00
Savings:	(You save $8,323.00)
Switch Family:	SB7700
Condition:	New
Ports:	36
Connector Type:	QSFP28
Speed:	EDR
Technology:	InfiniBand
ECCN:	5A991

Download MSB7790-EB2F Product Brief

Quantity: 1

Add Mellanox Support? -- None --

Add to Cart

Get a Quote?

FREE SHIPPING. Orders over $499.

Save the MSB7790-EB2F

Email LinkedIn

No se puede mostrar la imagen en este momento.

30m (98ft) Generic Compatible 100G QSFP28 Active Optical Cable

#65872 ★★★★★

Available, Ship on Tue. Mar. 13

US$ 570.00

Hardware: Custom interconnect for AI/ML?

- Network architects have bigger fish to fry...
- Networks are heavily $$$ constrained
- Developing a custom interconnect for AI/ML costs even more $$$
  - Can this be economically justified?
  - Funding needs to come from government agencies
  - May add bells & whistles, but feeds and speeds still constrained
- We need to learn to deal with global bandwidth constraints

Protocols: Programmability to the rescue?

- Main networking “innovation” in recent years has been programmability
  - OpenFlow, P4, Barefoot Tofino,
  - Innovium Teralynx, Broadcom Tomahawk 3, Mellanox Spectrum 2
- Enables protocol innovation

Software

- GPU-based architectures will have to deal with low B/F
- Problems need to be embarrassingly parallel
- OR algorithms need to be communication avoiding

Network data throughput

- Assumption: Networking hardware is the bottleneck, not communication stack
- Ways to obtain higher data throughput (hardware)
  - Reduce overhead on the wire
  - Increase goodput
  - Increase signaling rate per channel
  - Increase number of channels

Question 3

- Given that the data-driven applications such as machine-learning and artificial intelligence applications will have their bottlenecks in data throughput, can standardized networking hardware such as 10 Gigabit Ethernet and Infiniband be configured efficiently to support their communication requirements? If so, how? If not, what features are needed?

- The solution is not in the network
- Workloads need to be architected around network constraints
- IB and Ethernet standards are what they are
  - Feeds and speeds will continue to increase
  - Resistance to add new features is IMMENSE
- Drastic reduction in $/Gbps of optics could alleviate the bottlenecks

- Layer 1: Feeds and speeds will continue to increase
  - 400GE ratified Dec. 6, 2017
  - 100G per channel is in the pipeline
  - But none of this will not close the gap with compute per node
  - Optics will (have to) become pervasive AND a lot cheaper
- Layer 2 is pretty much cast in stone
  - Highly unlikely that application-specific requirements will make it in the standard
- Layer 3 and up
  - Pick from existing protocols
  - Invent your own and use programmable networking hardware (and be prepared to rewrite networking software)