High Performance Ethernet for Computing and Storage Systems

Yan Zhuang, Huawei Technologies

IEEE 802.3 New Ethernet Adhoc

June 22nd, 2022

Contributors and Supporters

- José Duato, Polytechnic University of Valencia, Member of the Royal Spanish Academy of Sciences
- TruongThao Nguyen, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki, Japan
- Fazhi Qi, Computer Center, Institute of High Energy Physics, CAS
- Shan Zeng, Computer Center, Institute of High Energy Physics, CAS
- Wei Zhang, Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences)
- Liang Guo, CAICT
- Jie Li, CAICT
- Zhaogeng Li, Baidu
- Xuequn Wei, JD.com
- Hao Quan, Meituan
- Leon Bruckman, Huawei

Acknowledgement

The authors would like to thank Sudheer Chunduri, Scott Parker, Pavan Balaji, Kevin Harms and Kalyan Kumaran from ALCF, Argonne National Laboratory, Argonne National Laboratory, for their good analysis in the paper “Characterization of MPI Usage on a Production Supercomputer” and permission to share their work in this contribution.

The authors also would like to thank SNIA for their work and would like to share it in this contribution.

Agenda

- Background activities
- Market Drivers
  - Computing Market
  - Storage Market
- Ethernet challenges for high performance applications
  - High Throughput
  - Low Latency
- Technology Feasibility
- Summary
- Q&A?

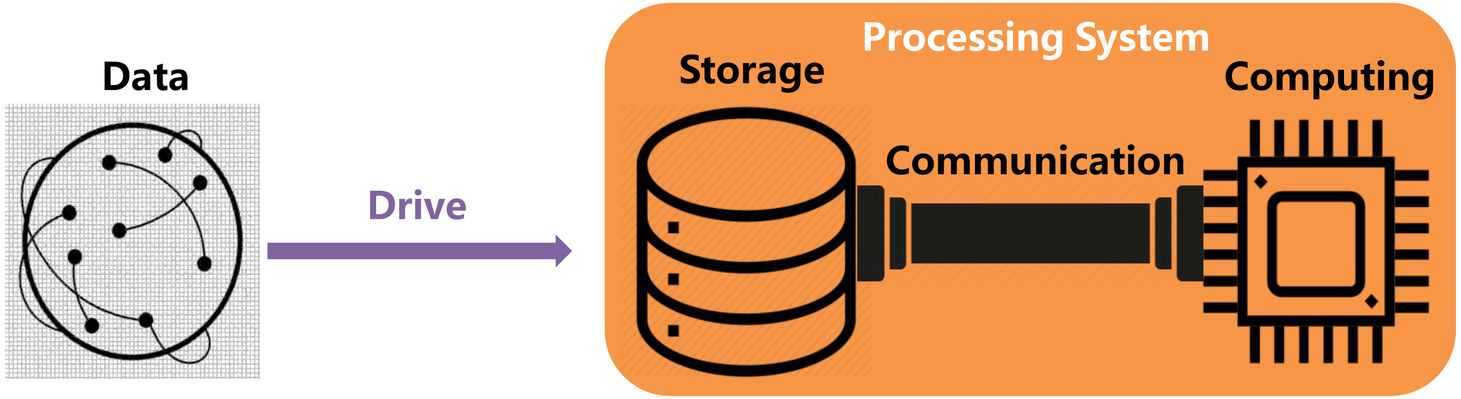
Background activities

- Start the 1st discussion in IEEE 802.3 NEA to provide the market drivers in high performance storage and computing applications and show two general directions (i.e. low latency and high throughput) (please refer to zhuang_nea_01a_210407.pdf).
- A session focused on short frame discussions that provide some technology feasibility to support short frames in Ethernet was held.
  - Some technical and performance questions were asked and more data was requested.
- What do we want to discuss in this meeting?
  - Updated marketing trends, provide further information on technical gaps and technology feasibility

MARKET DRIVERS

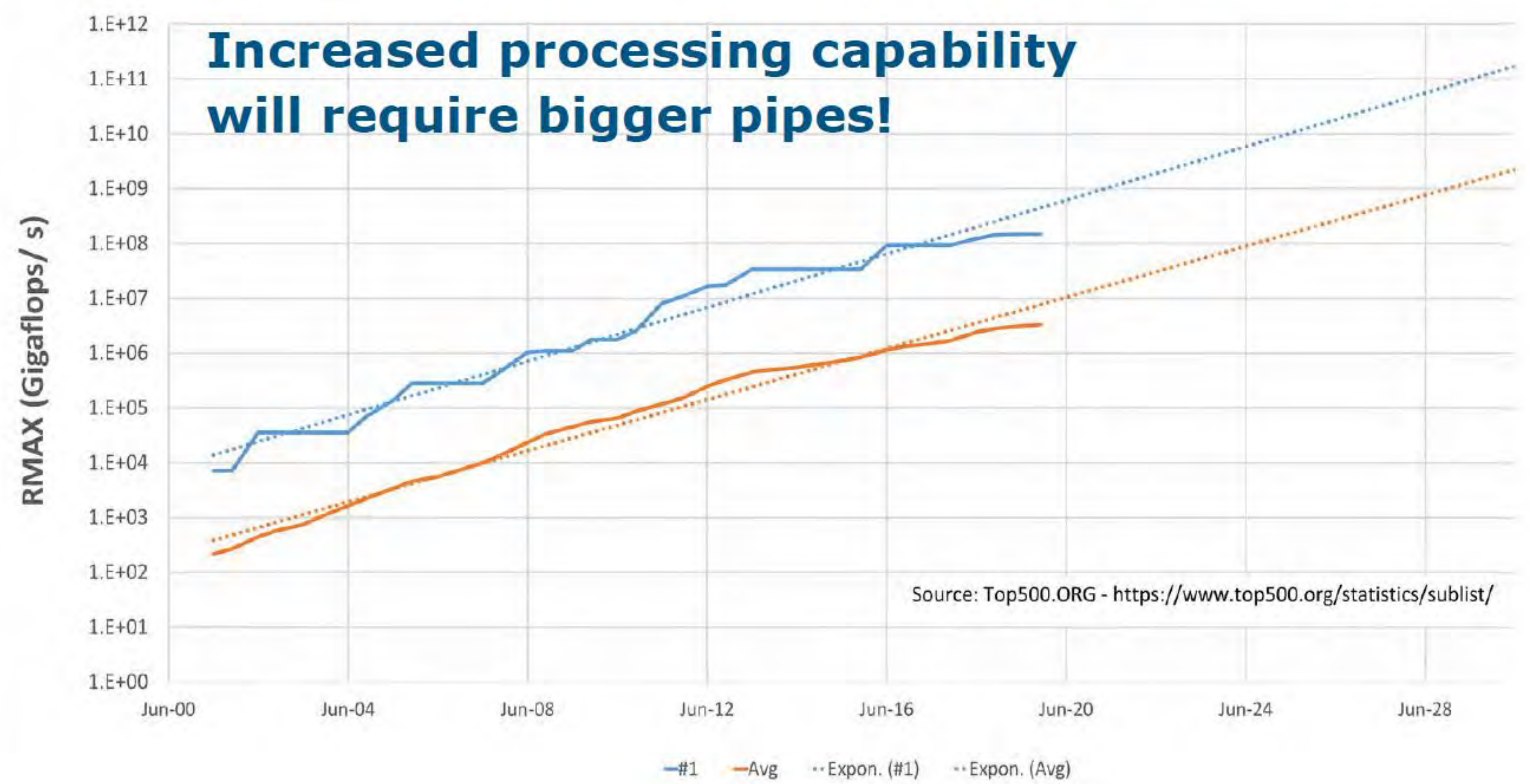
The Market that we are talking about and its Ethernet trends

- Computing Market
- Storage Market



HPC Market

HIGH PERFORMANCE COMPUTING



Increased processing capability will require bigger pipes!

Source: Top500.ORG - https://www.top500.org/statistics/sublist/

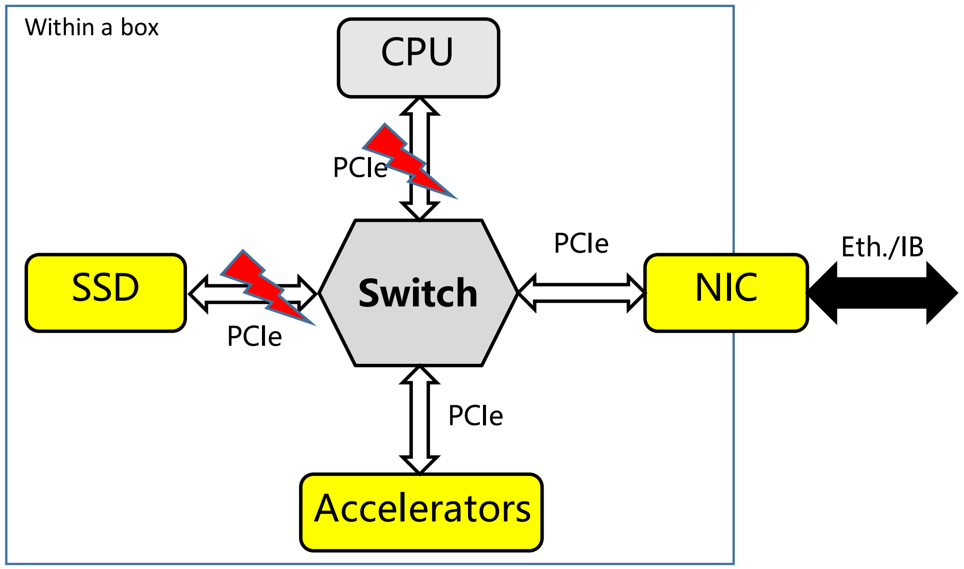
Source: B400G CFI: https://www.ieee802.org/3/ad_hoc/ngrates/public/calls/20_1029/CFI_Beyond400GbE_Rev7_201029.pdf

HPC now is stepping to more AI-focused use cases for science computation as well as data analysis. Meanwhile, OTT companies (like AWS, IBM as well as Huawei etc al.) are providing HPC cloud to their customers for high performance computing. These two markets are trending to merge somehow to provide computing-intensive services for their customers by cloud access. AI Market trends can be found in the B400G CFI as well.

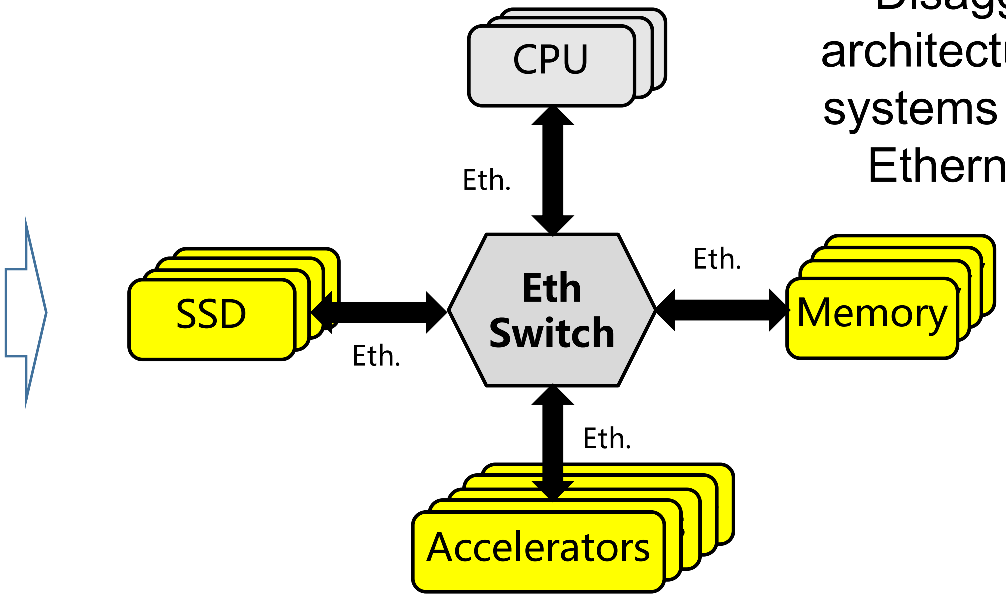
Disaggregation: Direct Interconnects to components

- The primary objective of computing infrastructures is to provide high/large computing power to their customers to deal with various computing-intensive tasks.

As Is



To Be



Disaggregated architecture/pooling systems with Direct Ethernet attach

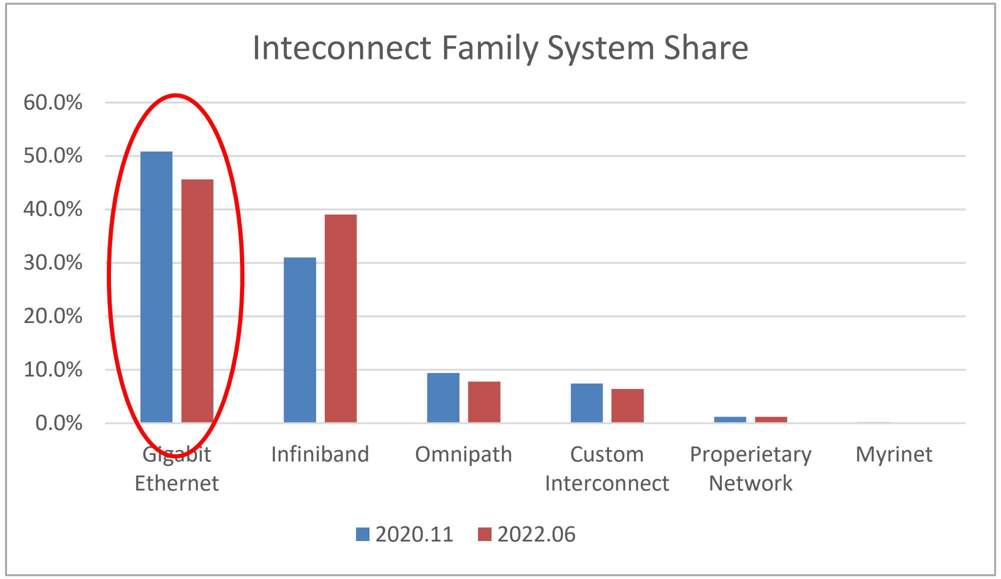
- Moore's Law is coming to an end, increasing the number of accelerators (e.g. GPUs, TPUs…) in a server comes at the expense of proportionally increased power consumption.
- PCIe interconnects become a bottleneck due to the limited bandwidth and increasing components.

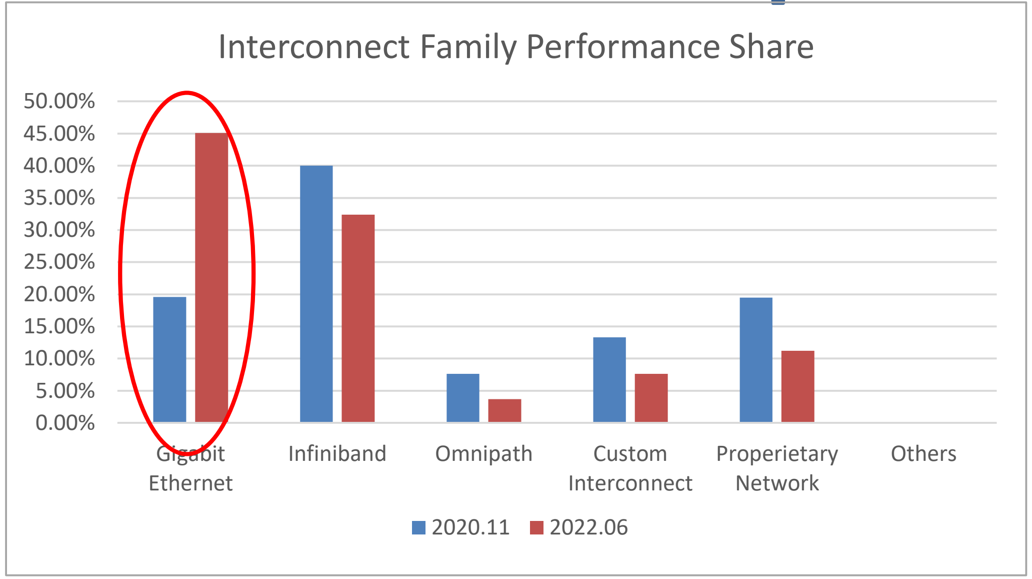
- Flexible resources allocation by pooling resources.
- Fast access by high bandwidth Ethernet.
- Extensible Ethernet network for good system scalability.

*Please refer to systems like Habana Gaudi and "Server Disaggregation" for more information.

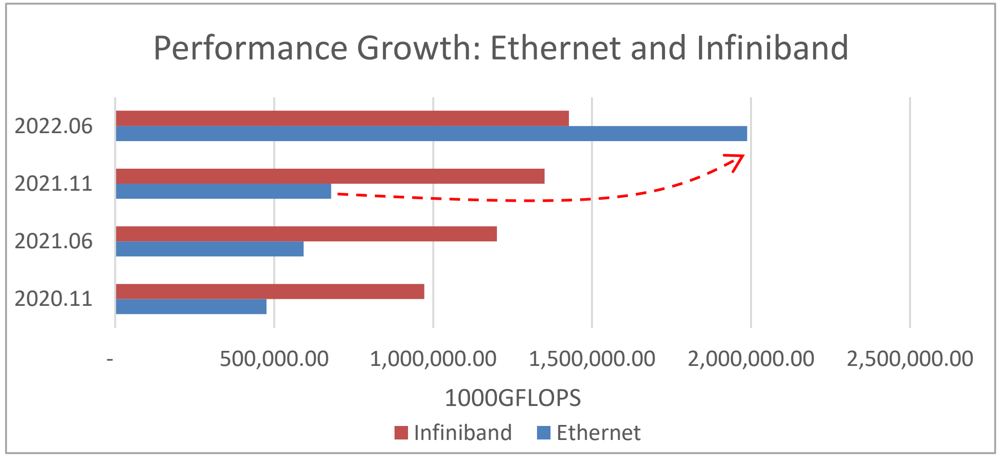
How to effectively "sum up" the increasingly heterogeneous nodes to provide more computing power!

Ethernet Interconnects Trends in Top500





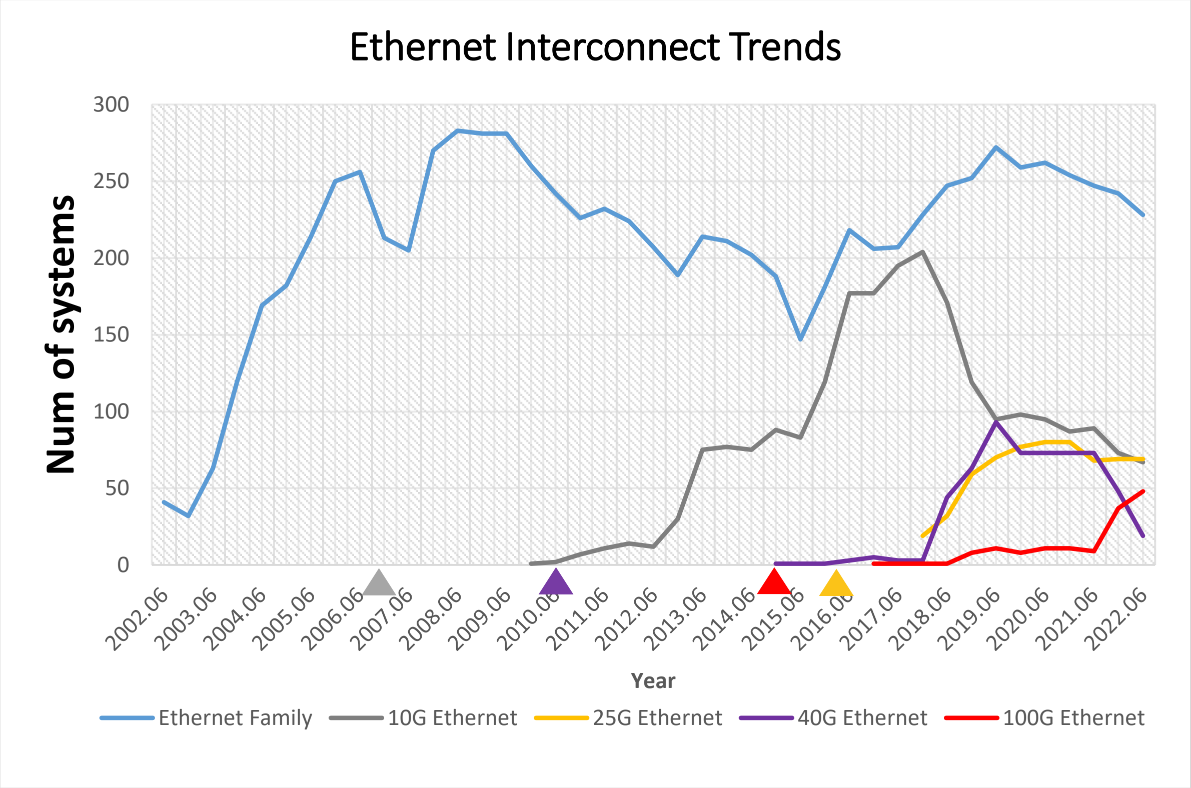
In the latest Top500 list, Ethernet Interconnects reduces its share from 50.8%@Nov, 2020 to 45.6%@June, 2022, while its contributed performance increases from 19.6% performance share (i.e. Rmax) to 45.1% which exceeds IB ☺.

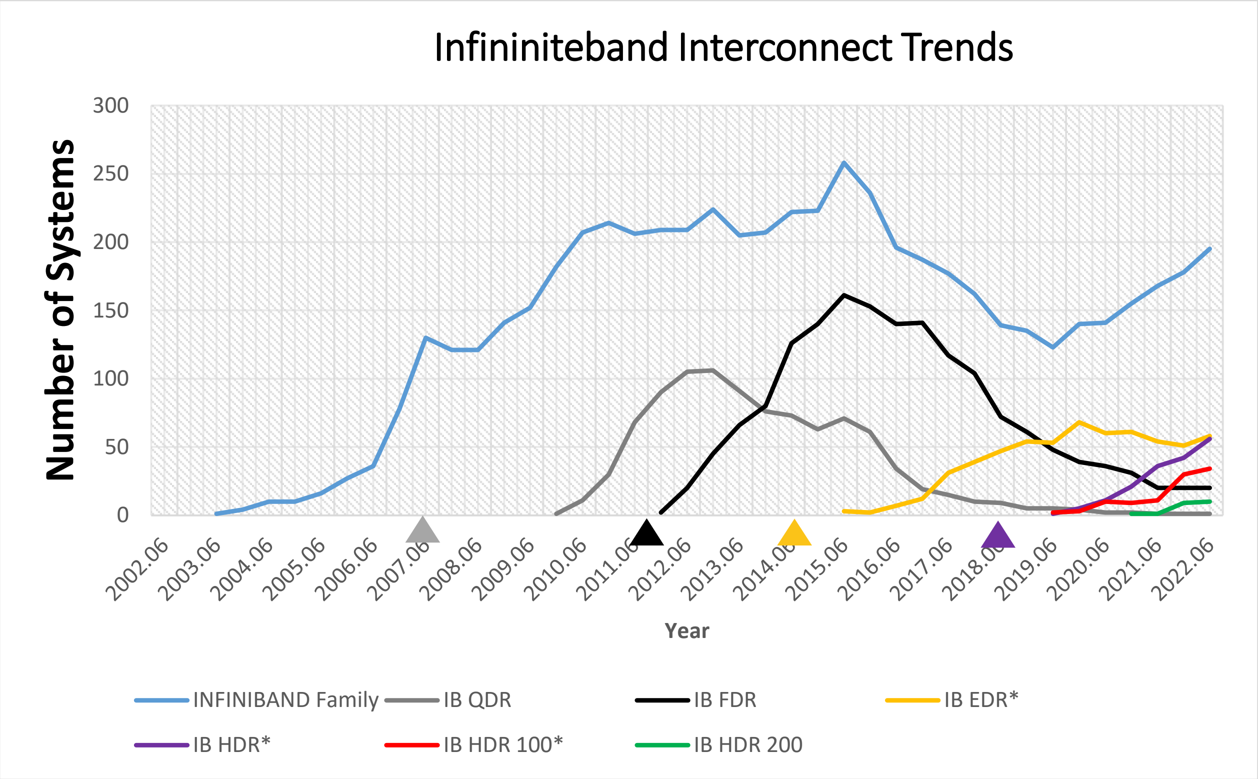


- The R_{max} of Ethernet increases by 1,308,115,360 GFLOPS which is almost triple the total performance of last year, which might be mostly contributed by the newly launched Slingshot-11 systems ☺ (contributed1,332,485,500GFlops in total) .
- In the latest list (released in June 2022), 3 newly launched Slingshot-11 based systems takes positions in Top10 (rank 1, 3 and 10), while Slingshot is based on Ethernet with some specialized enhancements.

Note: all data from https://www.top500.org/statistics/list/ Note: Rmax is a system's maximal achieved performance

Interconnect Upgrades in HPC





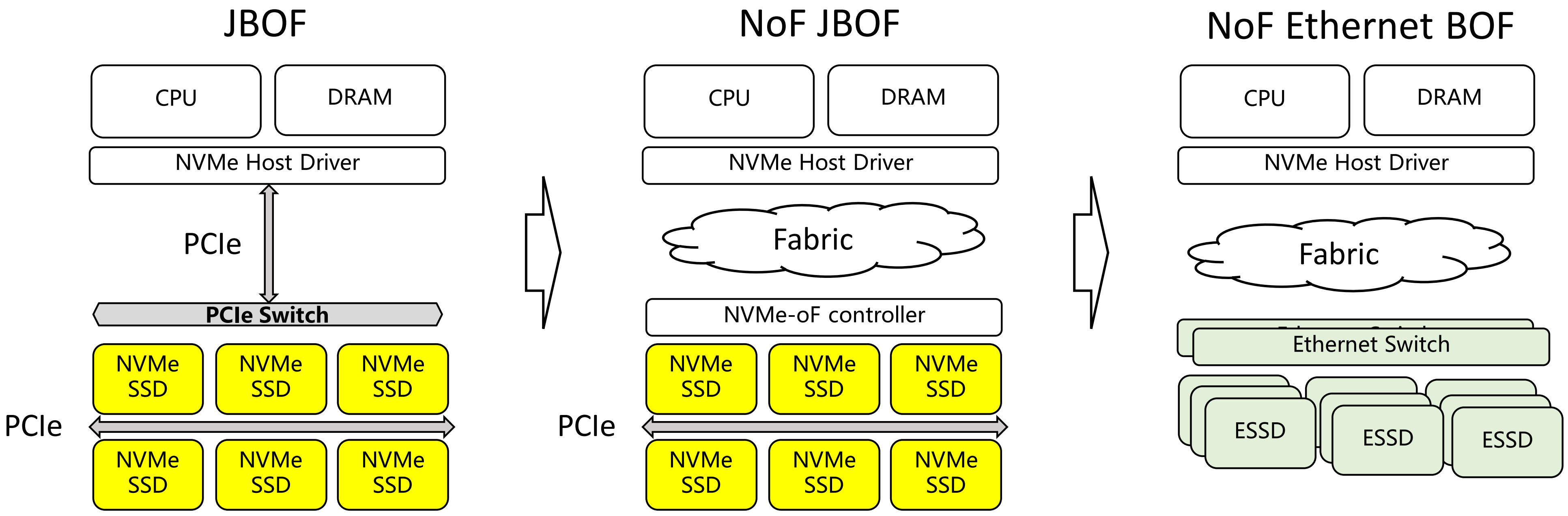
GbE	10GbE	25GbE	40GbE	100GbE
802.3ab-1999	802.3an-2006	802.3bq-2016 802.3by-2016	802.3bq-2016 802.3ba-2010	802.3ba-2010 802.3bj-2014 802.3cu-2020

QDR: 10Gb/s	FDR: 14Gb/s	EDR: 25Gb/s	HDR: 50Gb/s	NDR: 100Gb/s	XDR: 250Gb/s
2007	2011	2014	2018	2021	??

- Ethernet starts earlier in HPC networking while IB applies its high rate into HPC networking much faster. And also for standard Ethernet, the performance is uncompetitive while some specialized Ethernet (e.g. Slingshot-11) even outperforms IB interconnects.
- HPC networking is stepping into Exascale or higher computing performance with much more data exchange, aimed for more efficiency interconnects and Ethernet has its chance and possibility.

Note: all data is from: https://www.top500.org/statistics/list/.

Storage: JBOF to NVMe-oF(NoF) Ethernet BOF

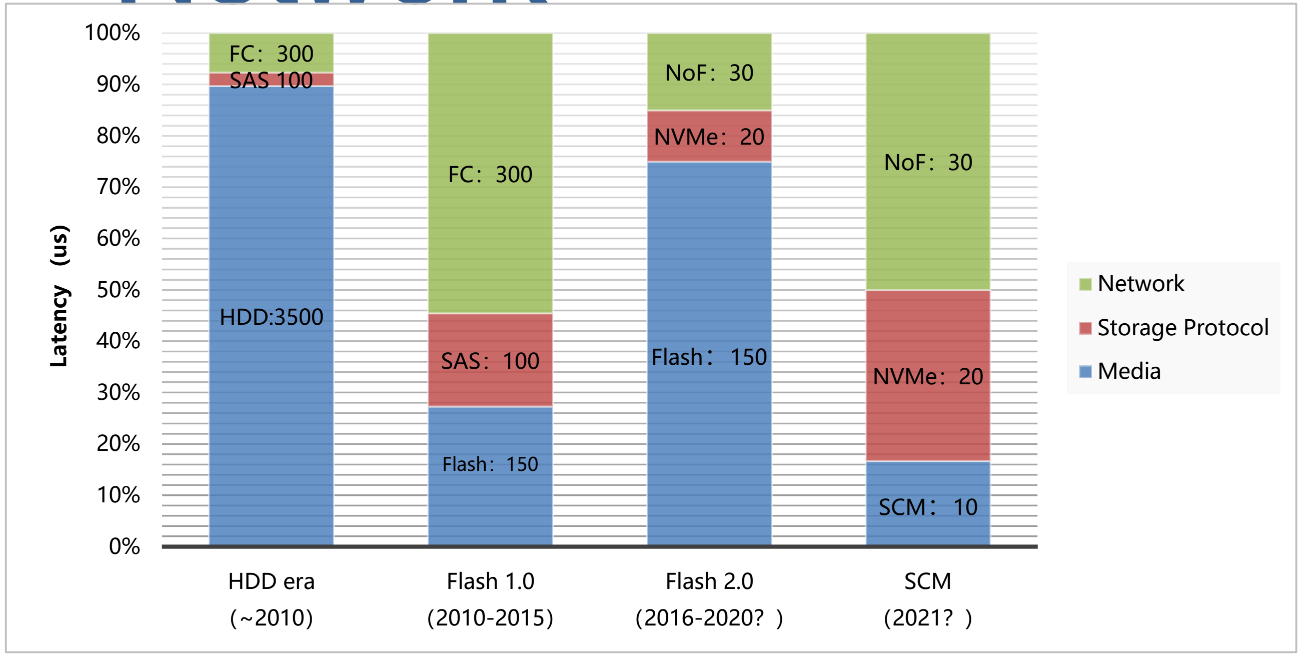


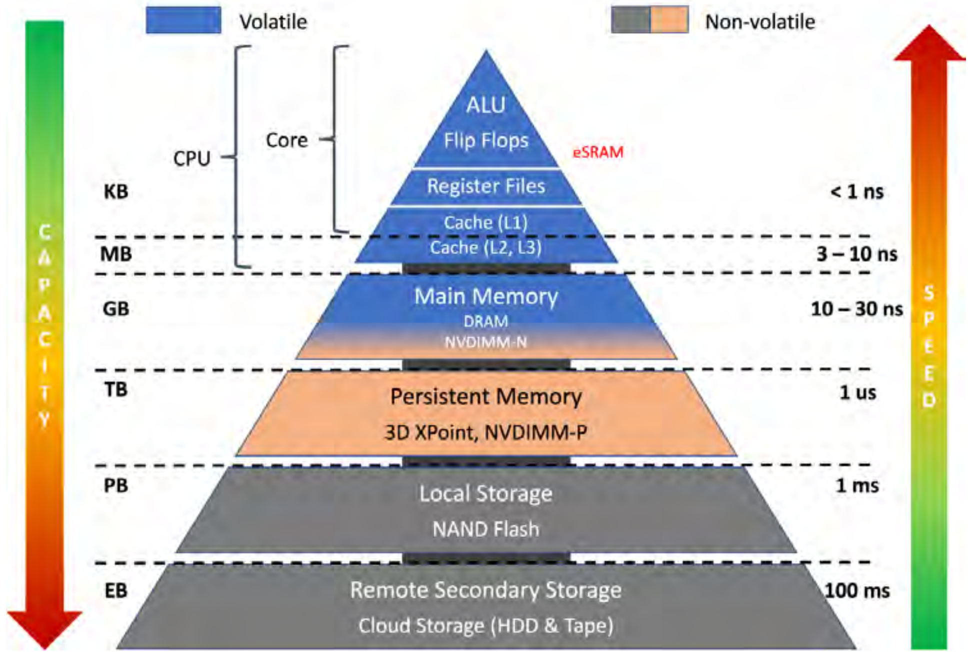
- From JBOF to NoF EBOF:
  - A simple backplane design that offers high density and full utilization of flash attached.
  - Scalable Ethernet switching and extensibility for more SSD nodes.
  - Less power with direct connected to Ethernet

Shahar Noy from Marvell also presented “Storage - Ethernet as the Next Storage Fabric of Choice”, OCP summit 2019

Notes: JBOF = Just a Bunch of Flash.

New Storage Media Needs Microseconds Network





Source: ‘ Storage Hierarchy: Capacity & Speed’ diagram reprinted with permission from SNIA, ‘Introduction to SNIA Persistent Memory Performance Test Specification’, ©2020, www.sina.org.

- Storage media is getting faster from SSD to SCM (Storage Class Memory) (aka. PM, Persistent Memory) with latency down to ~10us or even below 1us.
- Note that even with 10us latency, the network takes 50% of the total latency (with 10us SCM) and will perform even worse with more advanced media (e.g. 1us PM).

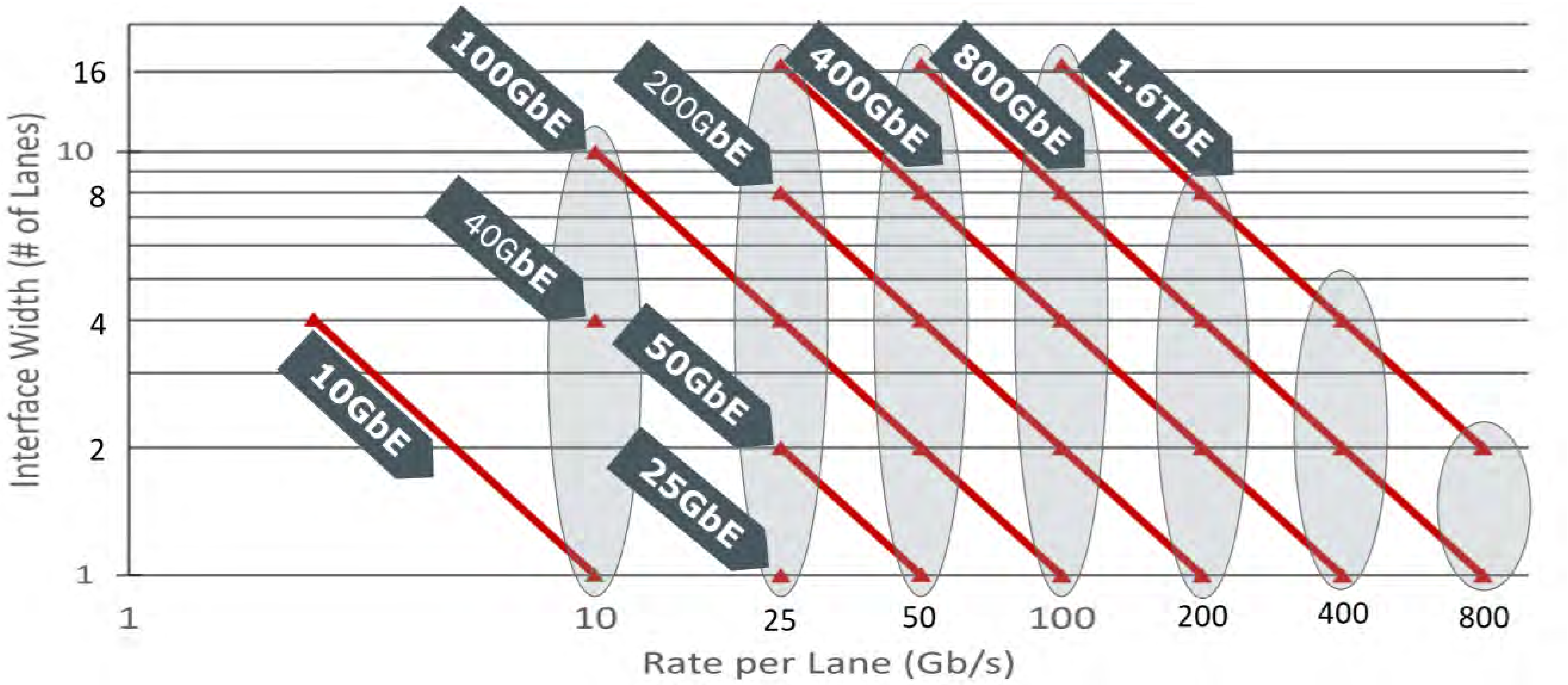
Network is stepping into ~10us or even ~1us level to offer proportional gains!

*this page was presented on April 7th, 2021 in zhuang_nea_01a_210407.

Challenges for Ethernet

- Why Ethernet?
- The way for Ethernet to go -
  - High Throughput
  - Low latency improvement, even with small steps.

Why Ethernet ?



Source: https://www.ieee802.org/3/ad_hoc/ngrates/public/calls/20_0430/dambrosia_nea_02a_200430.pdf

Development of High Speed Technologies

Deep ecosystem involvement and rich connections.

Challenge 1 : High Throughput

We are focusing on higher bandwidth, however:

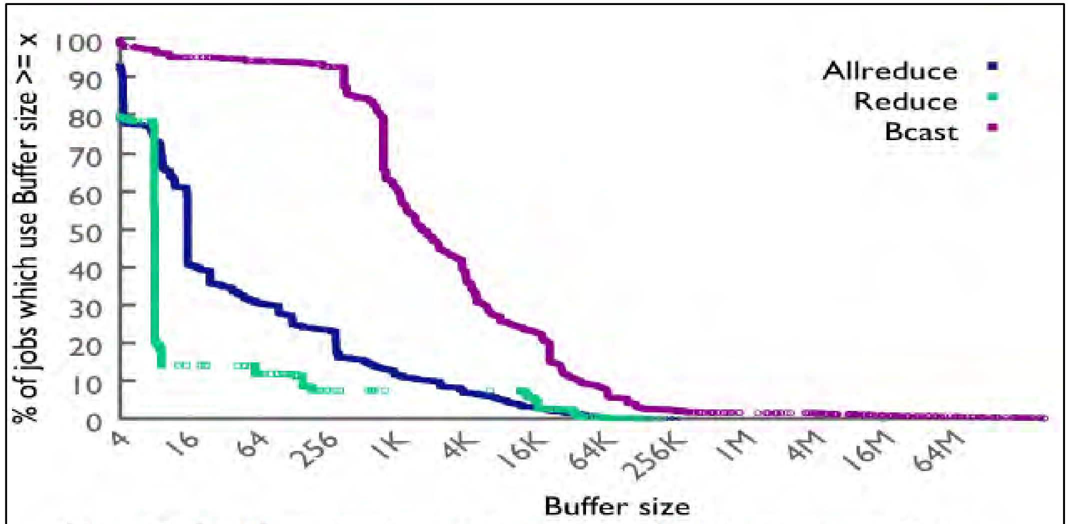
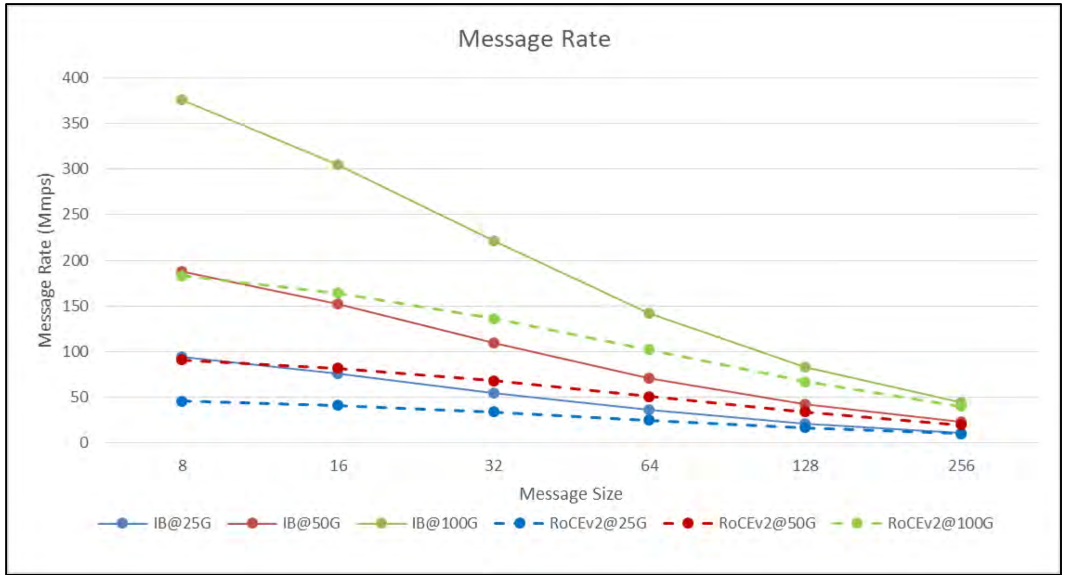


Fig. 17: MPI collectives total accumulated bytes on Cetus

S. Chunduri, S. Parker, P. Balaji, K. Harms and K. Kumaran, "Characterization of MPI Usage on a Production Supercomputer," SC18: International Conference for High Performance Computing, Networking, Storage and Analysis, Dallas, TX, USA, 2018, pp. 386-400, doi: 10.1109/SC.2018.00033. (© 2018 IEEE).

MPI is the communication protocol for most HPC applications. It provides support for collective communication, which includes Reduce, AllReduce and Broadcast operations. For HPC applications, over 85% of the Reduce messages are below 64 bytes and 70% of the Allreduce messages are below 64 bytes.



Analysis of the message rates of IB and Ethernet @ 25G/50G/100G

- As we can see from the graph, Ethernet provides lower message rates when compared with IB port @ the same rate.
- The Next Plaform also posted an article of “How Cray Makes Ethernet Suited For HPC And AI With Slingshot” August 16, 2019, by Timothy Prickett Morgan, in which it provides message rate comparison between Ethernet and other interconnects.

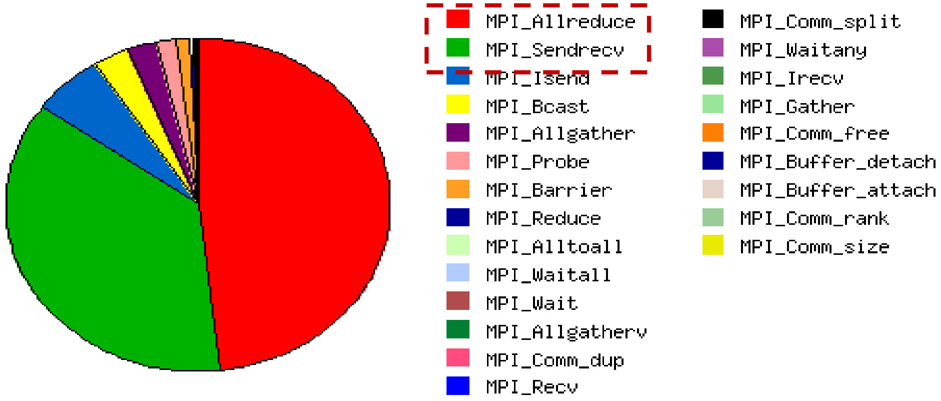
Address Feedback from last presentation

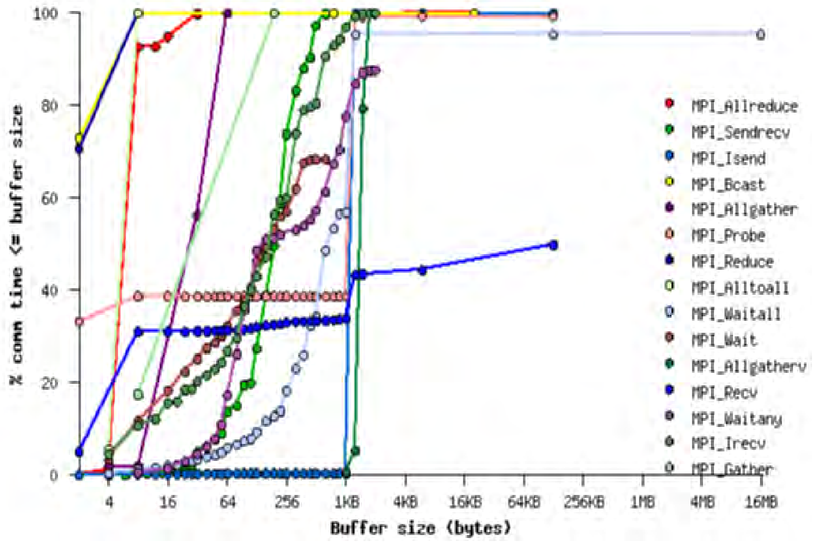
- Questions1: What is the traffic profile for these short-frame applications? Note that consecutive short frames might introduce large costs to support short frame forwarding/processing.
  - Answer: It is always a mixture of normal frames and short frames.
- Question2: How much performance can we get by making this small step? Maybe the performance gain is small just by saving a few bytes.
  - Answer: Let’s see an analysis of typical applications.

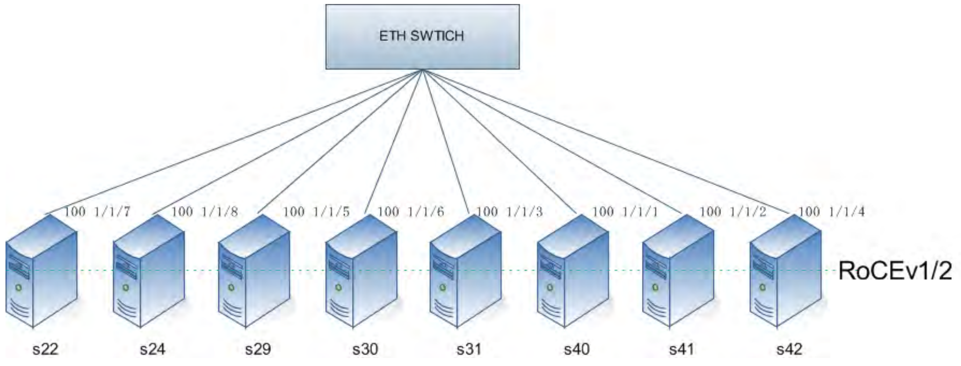
JCT improvements with less overhead

Traffic Profiles: SU2 mini-DFT is an well-known open-source software tools provided by Standard ford University (https://su2code.github.io) for performing Partial Differential Equation (PDE) analysis and solving PDE-constrained optimization problems, which is be used for applications like computational fluid dynamics (CFD) and aerodynamic shape optimization.

We ran a CDF application on SU2 platform over RoCEv1 and RoCEv2 network protocols separately to see the operation distribution and JCT (Job Completion Time) improvement as shown below.



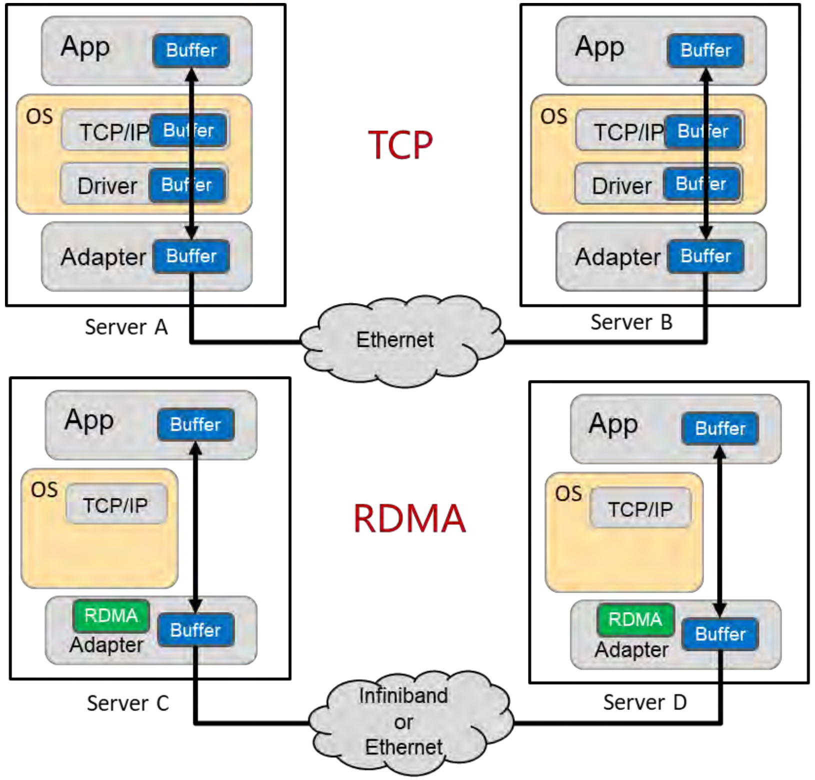




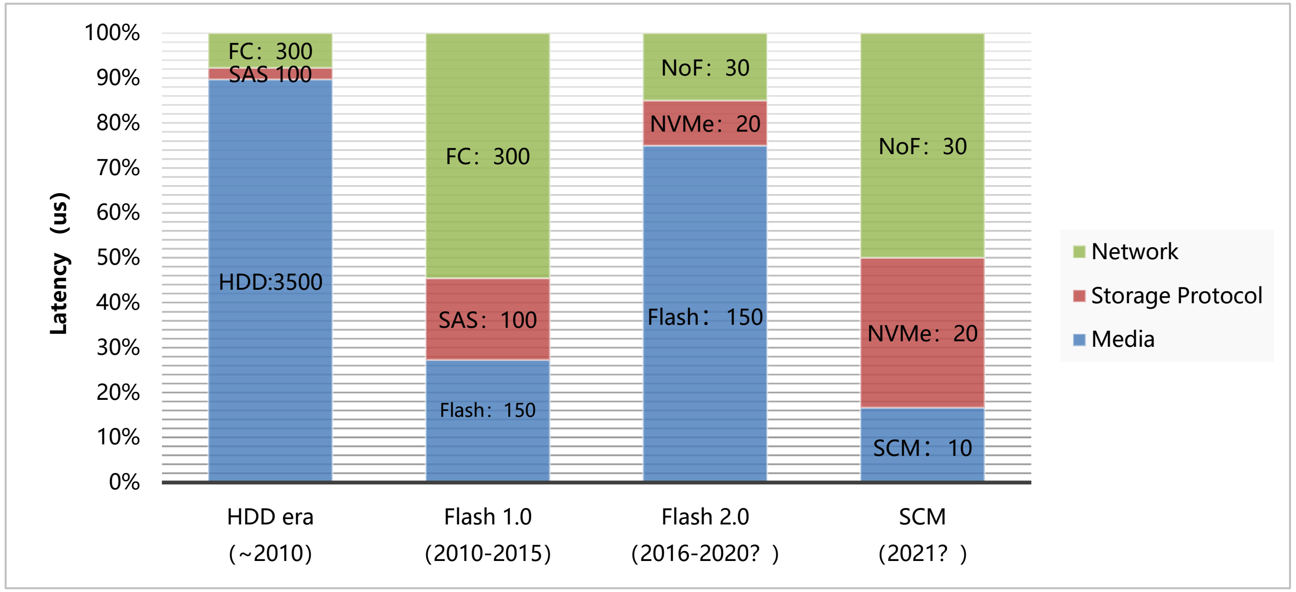
Compared with RoCEv1, RoCEv2 saves 12 bytes header overhead.
JCT (Job Completion Time) improvement: 13.7%.

Challenge 2 ： Low Latency in Storage

Transport gets faster: from TCP to RDMA



Storage gets faster: from 100us SSD to 1us PM/SCM



Storage media and stack gets faster and faster.
Even with 10us latency, the network takes 50% of the total latency (with 10us SCM) and will even perform worse with more advanced media (e.g. 1us PM).

- By using RDMA, time to transfers improves around 5x times which leads to below 10us round-trip.
- With the new storage media and storage protocols, the network latency percentage increases as part of the total latency. With 10us latency SCM, the network takes 50% of the total latency and will even perform worse with more advanced media (e.g. 1us PM).
- For large parallel application executed on the next generation high performance computing (HPC) systems, MPI communication latency should be lower than 1us*.

* Reference from: J. Tomkins, “Interconnects: A Buyers Point of View,” ACS Work-shop, June 2007

Challenge 2 ：Low Latency in HPC

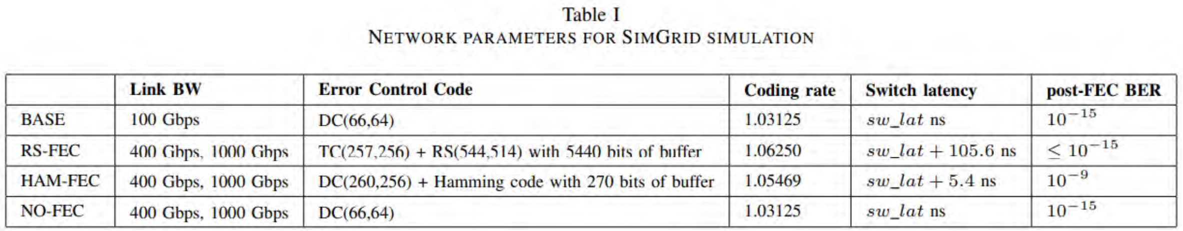
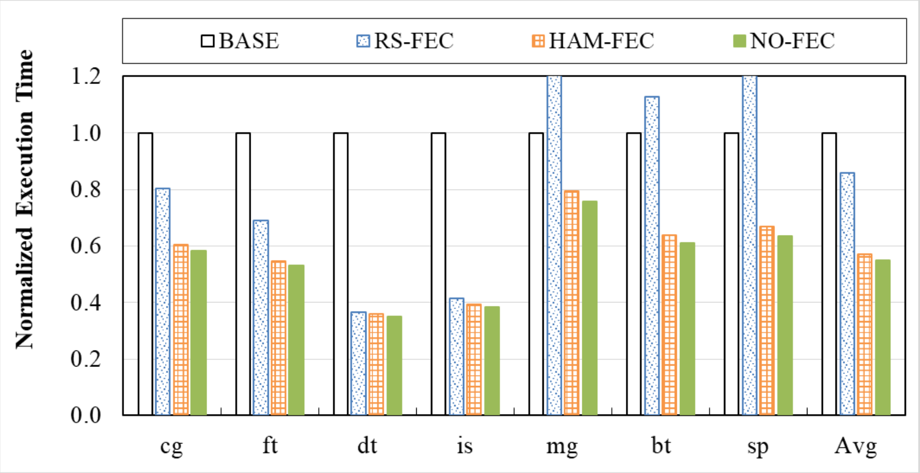


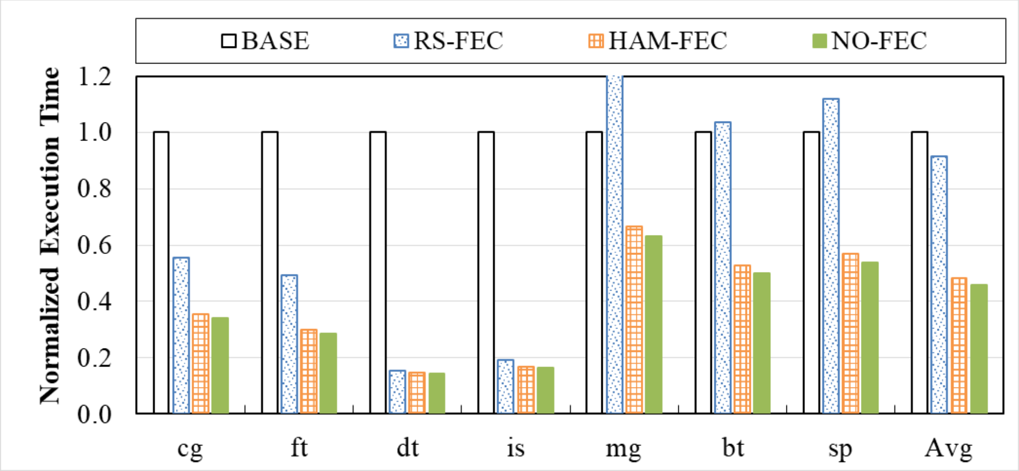
Table I
NETWORK PARAMETERS FOR SIMGRID SIMULATION

	Link BW	Error Control Code	Coding rate	Switch latency	post-FEC BER
BASE	100 Gbps	DC(66,64)	1.03125	sw_lat ns	10^{-15}
RS-FEC	400 Gbps, 1000 Gbps	TC(257,256) + RS(544,514) with 5440 bits of buffer	1.06250	$sw\_lat + 105.6$ ns	$\leq 10^{-15}$
HAM-FEC	400 Gbps, 1000 Gbps	DC(260,256) + Hamming code with 270 bits of buffer	1.05469	$sw\_lat + 5.4$ ns	10^{-9}
NO-FEC	400 Gbps, 1000 Gbps	DC(66,64)	1.03125	sw_lat ns	10^{-15}

Simulation settings:

- SimGrid with MVAPICH2 for MPI and runs NAS Parallel Benchmarks (including cg, ft, dt, is, mg, bt, sp).
- Network: 256 switches in 3 topologies (4D-torus, dragonfly and random), one host per switch with 100GFlops computation speed.
- Assume switch latency to 60ns as reported by 200G IB HDR*.



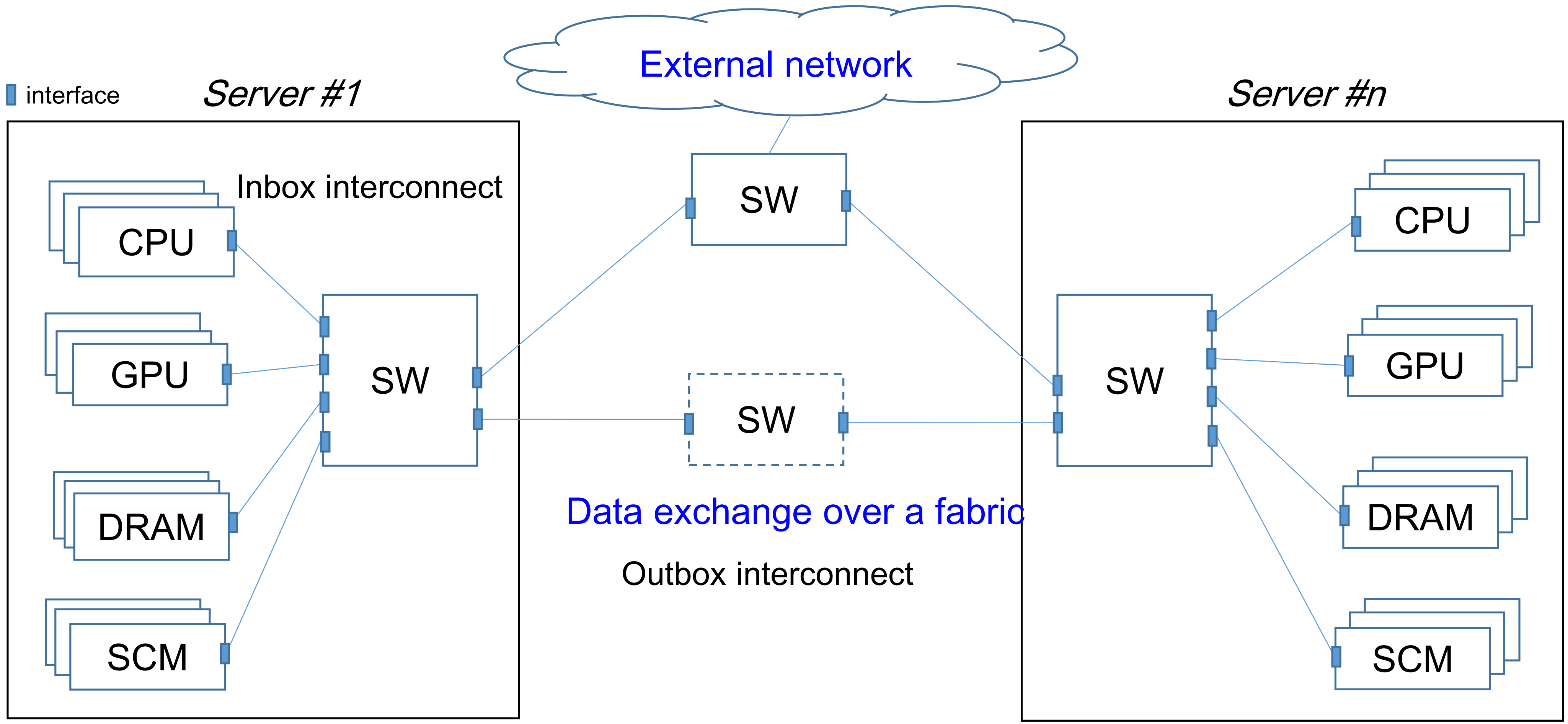


T.T. Nguyen, H. Matsutani and M. Koibuchi, “Low-Reliable Low-latency Networks Optimized for HPC Parallel Applications," 2018 IEEE 17th International Symposium on Network Computing and Applications (NCA), Cambridge, MA, USA, 2018, pp. 1-10, doi: 10.1109/NCA.2018.8548063. (© 2018 IEEE).

- The parallel applications in the figure includes: (i) latency-sensitive (e.g. bt, mg, sp), (ii) throughput-sensitive (e.g. is, dt), (iii) both (e.g. cg and ft).
- With 100ns latency improvement of FEC, for the latency-sensitive applications (mg, bt, sp), we can see the low-latency FEC (HAM-FEC) outperforms the RS-FEC by 62-66% and 90-97% in case of 400Gbps (on the left) and 1000Gbps (on the right).

*Mellanox Technologies LLC, “Introducing 200g hdr infiniband solutions,” http://www.mellanox.com/related-docs/ whitepapers/WP Introducing 200G HDR InfiniBand Solutions.pdf, 2018.

Conceptual interconnects within a system



$$Interconnect\, latency = static\ latency + dynamic\ latency$$

Latency is accumulated

$$Interconnect\ latency = \boxed{Static\ latency} + Dynamic\ latency$$

$$Endpoint_{tx_{phy_{delay}}} + Endpoint_{rx_{phy_{delay}}} + \boxed{processing\ time\ per\ hop^{*}} * hops + \sum_{i} link\ latency_{i}$$

$$processing\ time\ per\ hop = rx_{phy_{delay}} + tx_{Phy_{delay}} + Switching\ time$$

A. Dynamic latency is related to queuing time due to congestion and latency caused by retransmission due to packet loss.

B. Link latency = 5ns/m * reach. If the anticipated communication latency is less than 1us, then the reach is limited (might be below 20m).

C. Switching Time = Pipeline Time + SWB time.

② Outbox interconnects:

- 2-layer CLOS topology, there are 3 hops, which is $4 * (rx_{phy_{delay}} + tx_{phy_{delay}})$
- 3-layer CLOS topology, there are 5 hops, which is $6 * (rx_{phy_{delay}} + tx_{phy_{delay}})$

① Inbox interconnects (heterogeneous interconnects needs protocol transition)

$$phy\ latency = (1 + hops) * (rx_{phy_{delay}} + tx_{phy_{delay}})$$

* Note: for simplicity, we assume all interfaces with the same rate, no convergence.

Ethernet Delay constraints in spec

Sublayer delay constraints (maximum: ns)

Sublayers		25GbE	40GbE	100GbE	200GbE	400GbE
RS,MAC and MAC control (round-trip)		327.68	409.6	245.76	245.76	245.76
xxGBASE-R PCS		143.36	281.6	353.28	801.28	800
BASE-R FEC	BASE-R FEC	245.76	614.4	1228.8	--	--
	RS-FEC	-	-	409.60	--	--
BASE-R PMA		163.84	102.4	92.16	92.16	92.16
BASE-R PMD	CR/CR-S	20.48	102.4	CR4: 20.48 CR10:97.28	40.96	20.48
	KR/KR-S		51.2	KR4: 20.48 KP4 PMA/PMD: 81.92	40.96	20.48
	SR/LR/ER		25.6	20.48	20.48	20.48
BASE-T PHY		1024	640	--	--	--

The FEC sublayer takes over 60% of the whole latency. For 200GbE, the PCS/FEC sublayer takes over 67.9% of the whole PHY delay constraints, while for 400GbE, it takes 69%.

Taking KP4 FEC with150ns as an example, in a 2-layer CLOS network, the end2end FEC delay would be greater than 0.6us regardless of the other sublayers and link transmission time.

Technology Feasibility — Possible directions to go

- High Throughput for small frames
- Low Latency Ethernet

Consider to support short frames

- Minimum frame size is a legacy from the CSMA/CD days, while high speed PHYs are full-duplex.
- Market Drivers:
  - HPC has a large portion of short frames below 64 bytes (as stated) .
  - HPC applications throughput can be substantially improved by supporting “small frames”.
- Possible changes to be considered:
  - Remove the limitation of 64B minimum frame size or change it to a smaller number.
  - Provide mode negotiation before enabling short frame capability, so as to be compatible with existing Ethernet interface.

Low Latency Interconnects in the industry

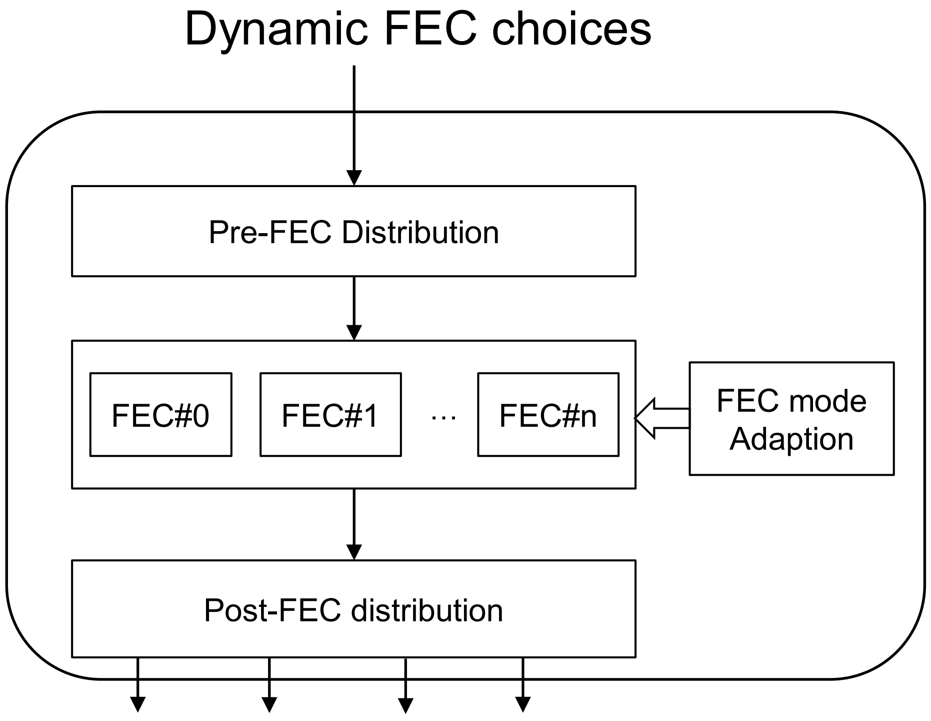
- **IB :** Provide Low latency FEC for NDR (PAM4), HDR(PAM4), FDR(NRZ).
- **GEN-Z :** Provide and consider low latency FEC options (as indicated in 1120_liaison_ieee_genz.pdf to IEEE P802.3ck)
- **Ethernet Technology Consortium :** Published ‘Low Latency Reed Solomon Forward Error Correction’ specification.
- **Ethernet: (previous IEEE 802.3 discussions)**
  - Technology feasibility: light FEC was discussed in 802.3cd previously.

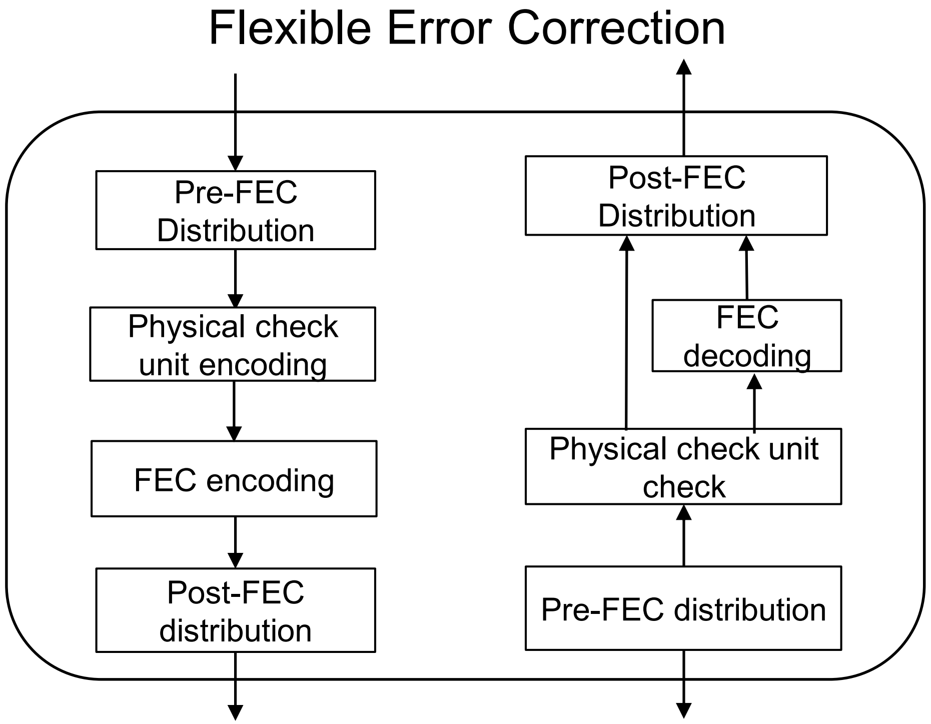
Ways to provide Low Latency Ethernet

- Use light FEC for low latency Ethernet, with:
  - ✓ Better Channels? Consider relevant channels for targeted applications only
    - e.g. 50G Serdes over .3ck channels or shorter reach
    - e.g. Infiniband differentiates short trace and long trace
    - Define better channels for low latency
  - ✓ Better Serdes?
    - e.g. 100G Serdes @ 50Gpbs for low latency
  - ✓ NRZ to improve SNR for 50G/lane?
    - We don't have 50G NRZ in the standard and might need new PMDs
  - ✓ Relax BER requirements for high performance applications?
- Flexible FEC architecture to provide low latency
  - Since in practice we have better channels/Serdes, it might be possible to have flexible FEC choices based on the performance of actual Serdes and channels in use (similar consideration is proposed in welch_3df_logic_220425.pdf).
- Simplified PMA to reduce latency (e.g. eliminating VLs if EIO and Optical lanes match)

Flexible FEC architecture

- A Flexible FEC architecture to reduce latency:
  - Dynamic FEC choices for different channels (incorporate multi-level FEC choices)
  - Or/and new encoding for flexible error corrections (i.e. adding error marking within a FEC codeword)





Summary: High Performance Ethernet with Low Latency and High Throughput

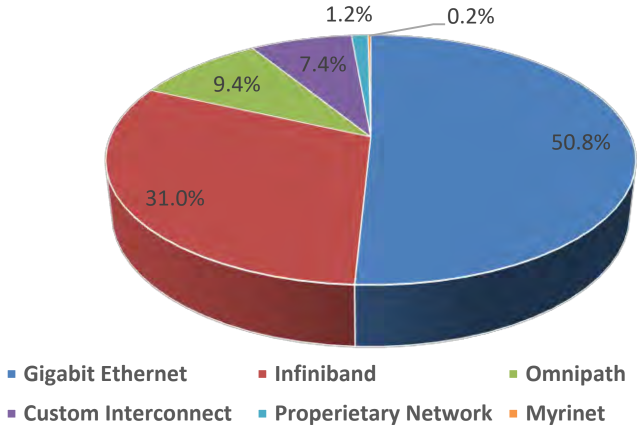
- Markets need low latency interconnects
  - Computing market (HPC, AI and hybrid cloud)
  - Storage market
- High performance computing applications benefit from short frames
- Industry already started efforts targeting low latency
  - IB/Gen-Z/ETC
- Propose to build consensus towards a call for interest for a High Performance Ethernet Enhancements project.
- Please contact us (zhuangyan.zhuang@huawei.com and/or leon.bruckman@huawei.com) if you are interested in participating or have any suggestions/comments.

Thanks!

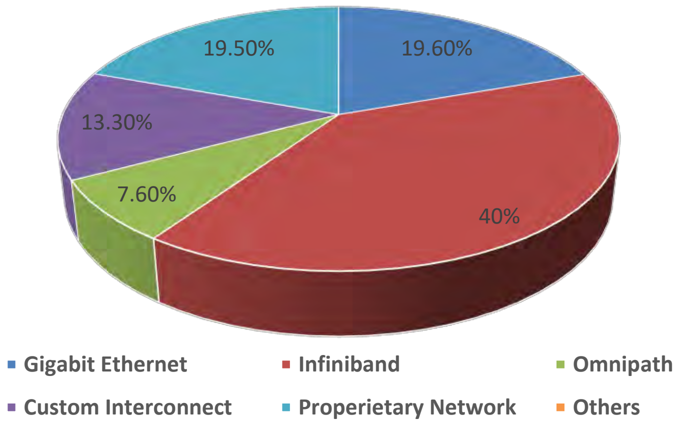
Ethernet Interconnects in Top500

Nov 2020

Interconnect Family System Share

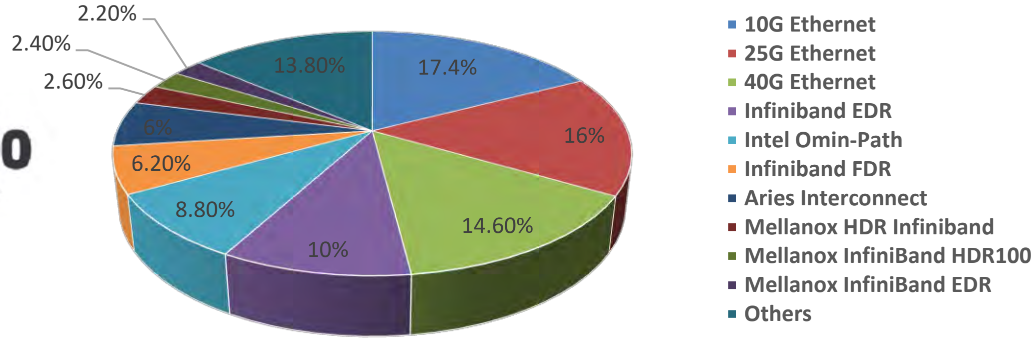


Interconnect Family Performance Share



Ethernet Interconnects takes the largest share (50.8%) of the Top 500 systems, while it only contributes 19.6% performance share (i.e. Rmax) compared with IB interconnects contributing 40% performance share with 31% system share.

Interconnect System Share

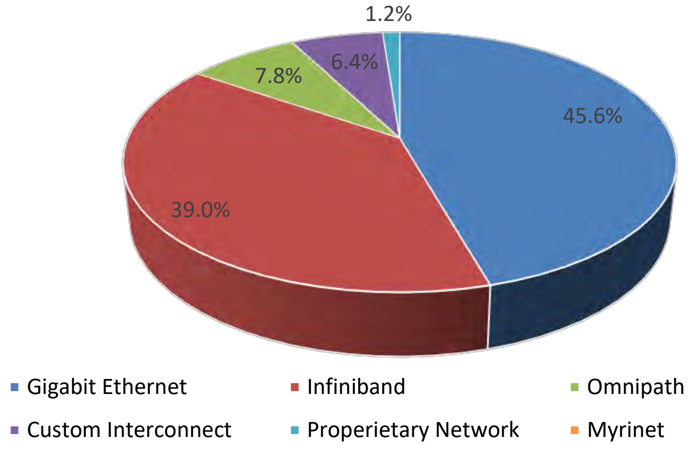


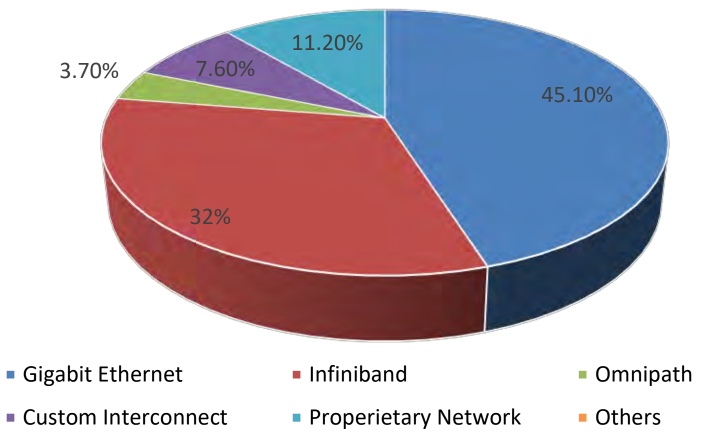
Infiniband interconnects are mainly EDR and stepped into HDR200 in Nov. 2020, while Ethernet interconnect stays still at 10GE (17.4%), 25GE (16%) and 40GE (14.6%).

Note: all data from https://www.top500.org/statistics/list/ Note: Rmax is a system's maximal achieved performance

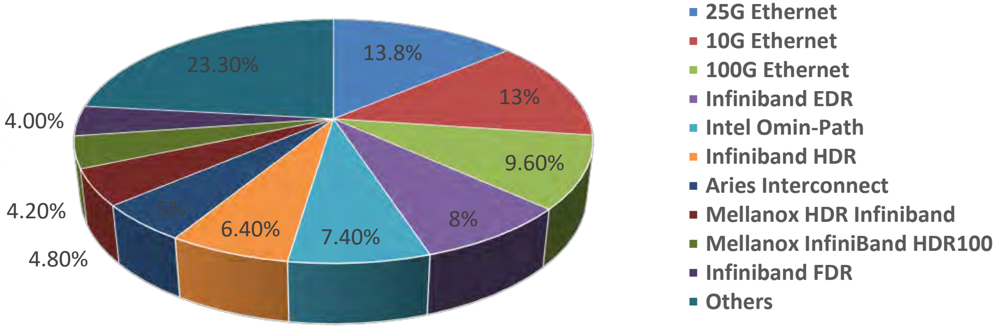
Ethernet Interconnects in Top500

June. 2022


Interconnect Family System Share


Interconnect Family Performance Share

Ethernet Interconnects reduces its share from 50.8%@Nov, 2020 to 45.6%@June, 2022 in the Top 500 systems, while its contributed performance increases from 19.6% performance share (i.e. Rmax) to 45.1% which exceeds IB ☺.


Interconnect System Share

Infiniband interconnects are largely stepped into HDR in June 2022, while Ethernet interconnect moves to 25GE (13.8%), 10GE (13%) and 100GE (9.6%).

Note: all data from https://www.top500.org/statistics/list/ Note: Rmax is a system's maximal achieved performance

Minimum Frame size in spec

- 4.2.3.3 Minimum frame size
  - The CSMA/CD Media Access mechanism requires that a minimum frame length of minFrameSize bits be transmitted. If frameSize is less than minFrameSize, then the CSMA/CD MAC sublayer shall append extra bits in units of octets (Pad), after the end of the MAC Client Data field but prior to calculating and appending the FCS (if not provided by the MAC client). The number of extra bits shall be sufficient to ensure that the frame, from the DA field through the FCS field inclusive, is at least minFrameSize bits.
- 4.4.2 MAC parameters

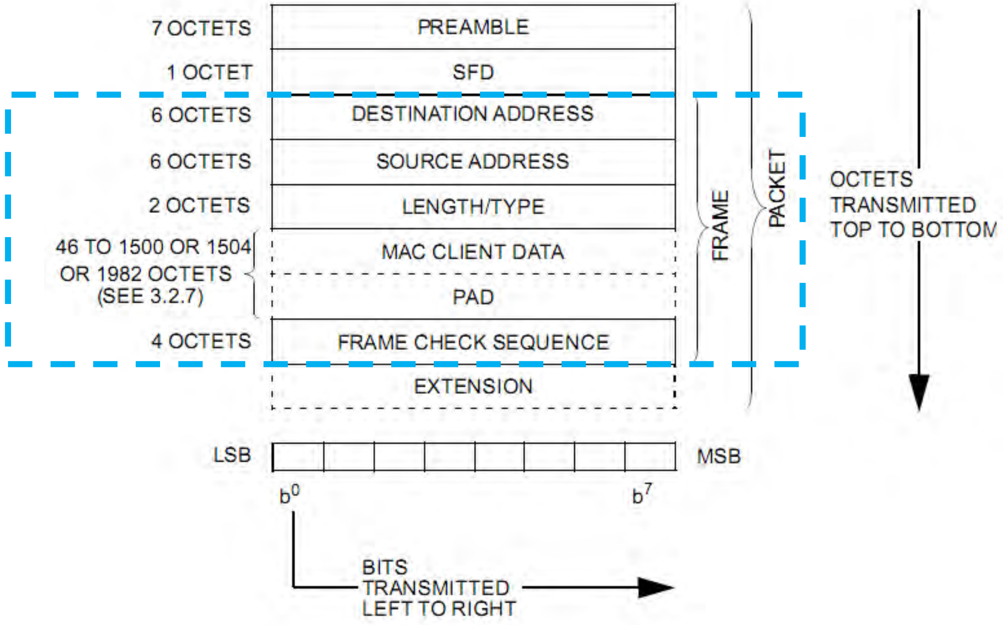
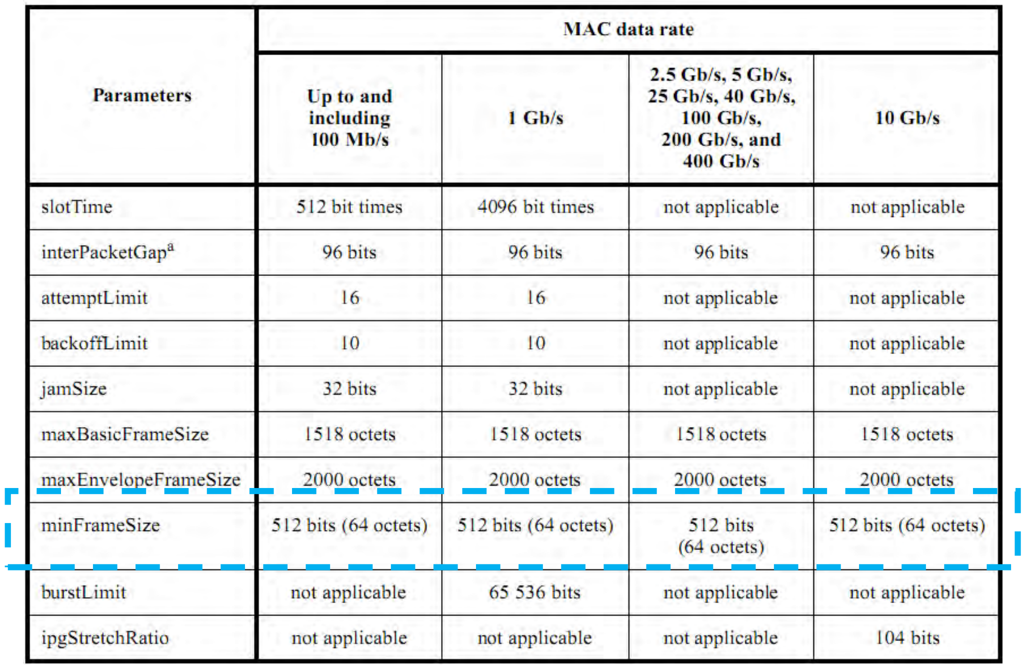


Figure 3–1—Packet format

Table 4–2—MAC parameters



Parameters	MAC data rate			
	Up to and including 100 Mb/s	1 Gb/s	2.5 Gb/s, 5 Gb/s, 25 Gb/s, 40 Gb/s, 100 Gb/s, 200 Gb/s, and 400 Gb/s	10 Gb/s
slotTime	512 bit times	4096 bit times	not applicable	not applicable
interPacketGap[a]	96 bits	96 bits	96 bits	96 bits
attemptLimit	16	16	not applicable	not applicable
backoffLimit	10	10	not applicable	not applicable
jamSize	32 bits	32 bits	not applicable	not applicable
maxBasicFrameSize	1518 octets	1518 octets	1518 octets	1518 octets
maxEnvelopeFrameSize	2000 octets	2000 octets	2000 octets	2000 octets
minFrameSize	512 bits (64 octets)	512 bits (64 octets)	512 bits (64 octets)	512 bits (64 octets)
burstLimit	not applicable	65 536 bits	not applicable	not applicable
ipgStretchRatio	not applicable	not applicable	not applicable	104 bits