CTF for Data Center Network

Lily Lv (Huawei)

Paul Congdon (Huawei)

High Performance Applications Growing Fast in the Data Center

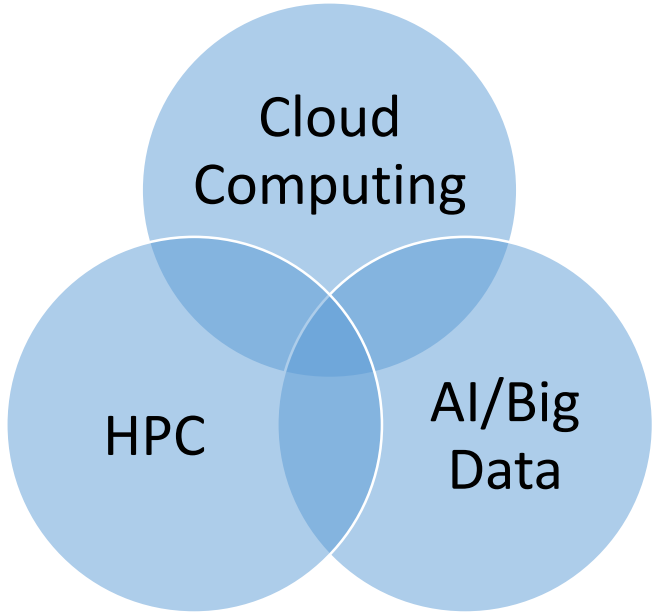
- **HPC(High Performance Computing), AI(Artificial Intelligence)/Big Data and cloud computing are hot growth areas.**
- **The converging of the 3 areas in data center shows great value in various verticals and becomes the trend.**
  - HPC is available as a cloud service in many public offerings (AWS, Azure, Alibaba etc). It grows with 17.6% CAGR(Compound annual growth rate) , even 2.5 times faster than on-prem HPC.
  - HPDA(High performance data analytics) and HPC-based AI are fast emerging markets, with 16% and 31% CAGR respectively.

	2019	2020	2021	2022	2023	2024	CAGR
HPC cloud	$2,466	$3,910	$4,300	$5,300	$6,400	$8,800	17.6%
On-Prem HPC	$27,678	$23,981	$26,774	$31,872	$36,138	$38,214	6.7%

Source: Hyperion Research, November 2020

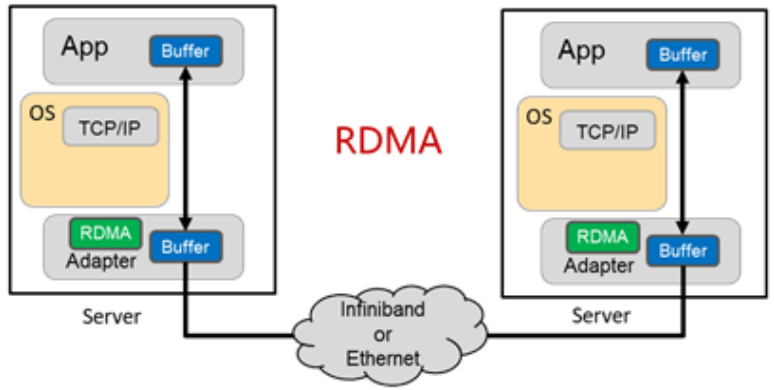
	2018	2019	2020	2021	2022	2023	2024	CAGR '19-'24
HPC Server Revenues	13,683	13,713	11,846	13,295	15,817	17,942	19,044	6.8%
HPDA Server Revenues	3,153	3,598	3,932	4,737	5,467	6,480	7,478	15.8%
HPC-Based AI (ML, DL & Other)	747	918	1,094	1,399	1,810	2,745	3,555	31.1%

Source: Hyperion Research, 2020

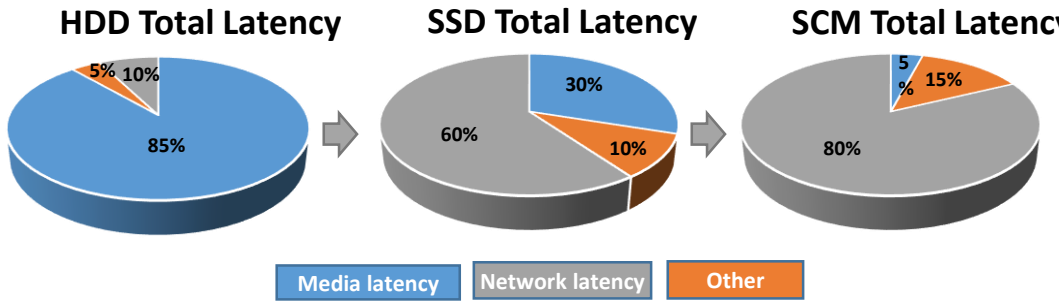


Latency is Critical in Data Center Network

- **Besides large computing power, high performance applications require low latency.**
  - E.g. Synchronization of large parallel software operation is critical to job completion time
- **New technologies are emerging to reduce system latency, such as RDMA, NVMe, new storage media etc.**



RDMA improves transfer time around 5x times compared with TCP/IP



Faster storage media has higher IOPS and lower latency.

Network latency becomes the bottleneck.

Latency is Critical in the Data Center Network

- **Types of latency in data center network**
  - Dynamic latency = queuing delay (congestion) + retransmission delay (packet loss)
  - Static latency = switch forwarding + packet processing + link latency
- **Dynamic latency attracts a lot of the industry's interest (congestion control, flow control etc.), however, static latency becomes significant in high performance scenarios, such as HPC.**

CTF Benefits HPC Network

- **HPC network operates at the nanosecond level.**
  - E2E latency is only several micro-seconds.
  - Per hop latency is required as low as possible, hundreds of nano seconds, or even lower.
- **CTF is applicable in HPC network**
  - Predictable traffic load in HPC network, less congestion on switches
  - Well structured fabric with same type of switches.

Regular Topologies: Two typical HPC network topologies

Fat Tree is a rearrangeably non-blocking structure, which provides an oversubscription ratio of 1:1 to all servers.

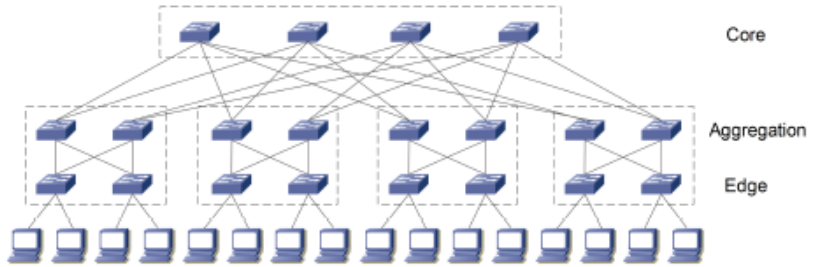


FIGURE 2. A simple 3-level Fat-Tree topology.

Source: Rethinking the Data Center Networking: Architecture, Network Protocols, and Resource Sharing

Dragonfly is a well-balanced network with no oversubscription.

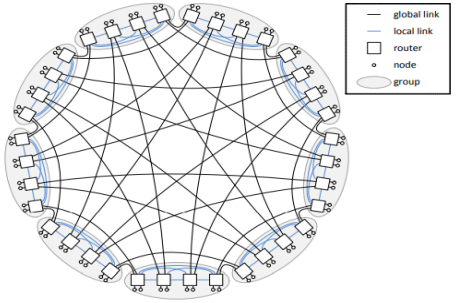
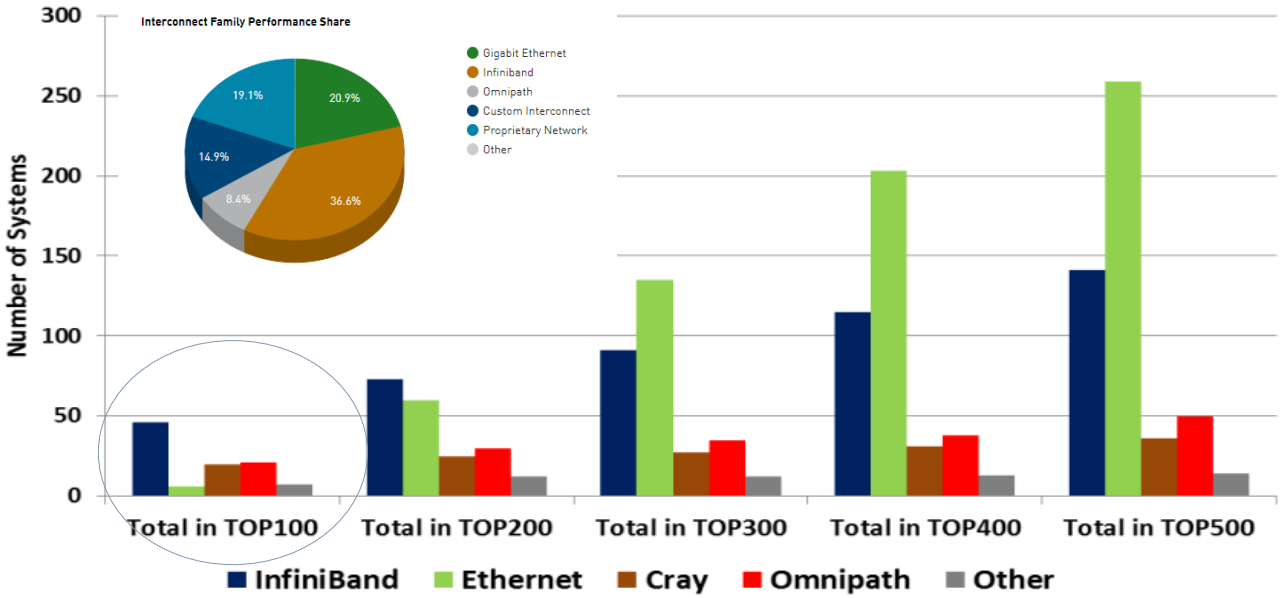


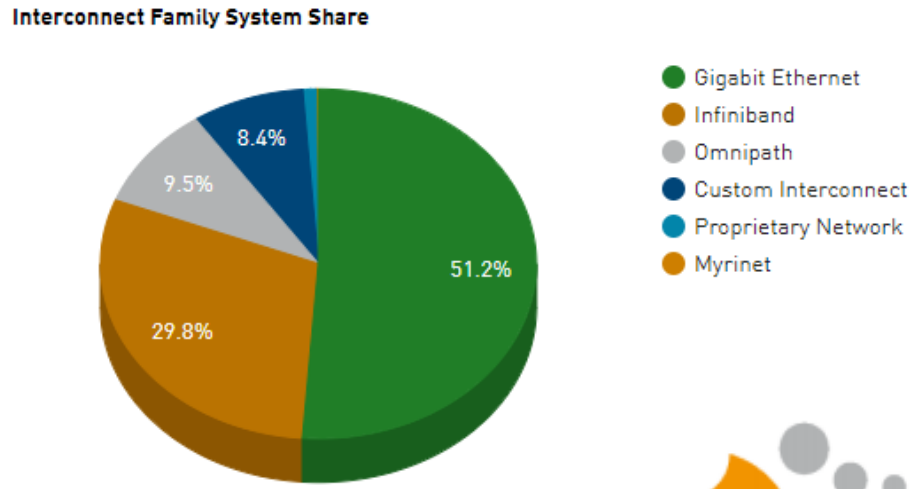
Fig. 1: Sample Dragonfly topology with h=2 (p=2, a=4), 36 routers and 72 compute nodes.

Source: On-the-Fly Adaptive Routing in High-Radix Hierarchical Networks

InfiniBand is the ‘first-choice’ in HPC Today

Although 51% supercomputers use Ethernet fabrics (more percentage than IB and other proprietary technologies), InfiniBand is the dominant interconnect in TOP100





TOP500 The List.

TOP500 is a list of the world’s 500 most powerful computer systems

InfiniBand is the ‘first-choice’ in HPC Today

- InfiniBand switch per hop latency is much lower than Ethernet switch
  - Ethernet chip latency is hundreds of nano seconds. The latency increases with frame size increasing if no cut-through.
  - InfiniBand chip latency can be lower than 100ns
  - Cut-through is an important feature in InfiniBand, to keep the per hop latency low.

Ethernet (non-CT)

	BRCM THK
Port	128*25G

One 25GbE Port to One 25GbE Port Test

Frame Size(Bytes)	64	128	256	512	1024	1280	1518	2176	4096	9216
Latency(ns)	511	528	556	567	717	793	872	1082	1694	3334

Source: Tolly, February 2016

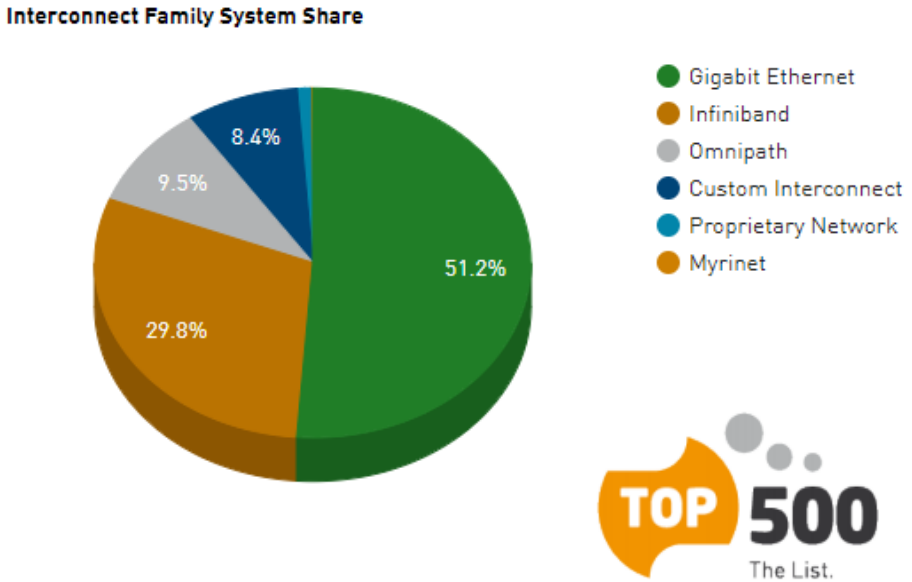
IB (with CT)

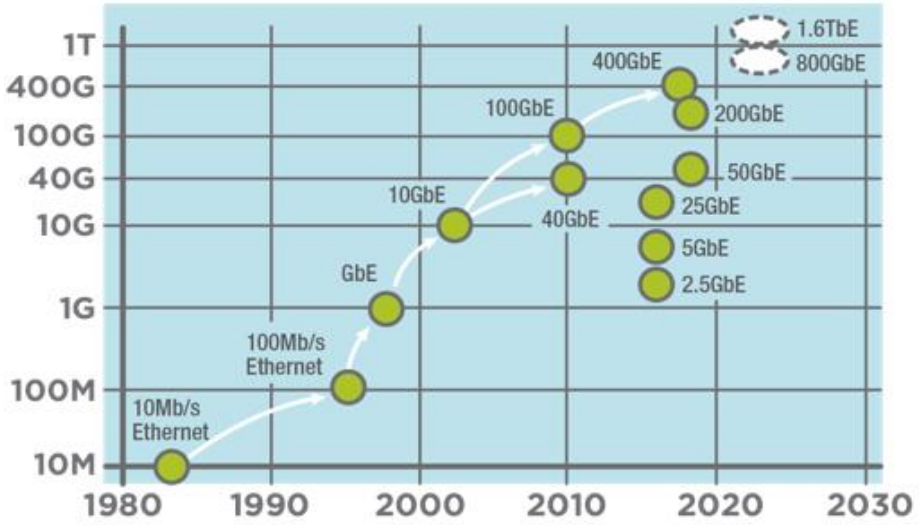
	MLNX Switch-IB	**MLNX Switch-IB2**
Port	144*25G	144*25G
Latency	90ns	90ns

Source: https://www.mellanox.com/news/press_release/

Necessity of CTF in Ethernet

- Ethernet has great opportunity to become more competitive in HPC market.
  - TOP 500 shows Ethernet Interconnects already takes the largest share (51%)
  - Ethernet has its own advantages
    - Ethernet is ubiquitous technology.
      - Cost-effective solution
      - Relatively easy to deploy and manage
      - Leading technology development
    - Ethernet provides large bandwidth connectivity
      - up to 400G, 100G for single lane
      - towards 800G, 200G for single lane





Source: Ethernet Alliance 2020 Roadmap

Necessity of CTF in Ethernet

- People see the appeal of Ethernet and want it for HPC
  - Converged network in their data center reduces complexity of management
    - Ethernet based RoCEv2(RDMA over Ethernet v2) is good enough to compete with FC(Fiber channel) in traditional storage network
    - The first choice for high performance computing network and storage network is still InfiniBand.
- The obvious gap of Ethernet is latency
  - Per hop latency gap is significant compared with InfiniBand
  - CTF is a good method to improve per hop latency

Why Need Standard?

- There are commercial switches already supporting CTF with proprietary solution. Hence CTF is technical feasible.
- However, proprietary solution may have interoperation problem: CTF may be invalid if interconnected switches with different CTF mechanism.
- Standardize CTF could enhance Ethernet eco-system advantage to help Ethernet extend the market.

Thanks!