BROADCOM

WHITE PAPER

Unleashing the Power of AI/ML

Why Ethernet Reigns Supreme over InfiniBand for Large-Scale Networks

WHITE PAPER

Unleashing the Power of AI/ML

Why Ethernet Reigns Supreme over InfiniBand for Large-Scale Networks

TABLE OF CONTENTS

Introduction

Ethernet is the leading networking technology across various industries, including data centers, service providers, and enterprise networks. In contrast, InfiniBand technology is primarily used in high-performance computing (HPC) applications, where low unloaded latency is critical. However, in the case of artificial intelligence/machine learning (AI/ML) training clusters that require high bandwidth, reliable transport, and predictable low tail latency to minimize job completion time, Ethernet is more effective than InfiniBand when compared under the same switch bandwidth and SerDes speed. Moreover, Ethernet offers several advantages over InfiniBand, such as greater bandwidth, higher radix, multi-vendor standard solutions, lower cost, power-efficient economics, better throughput, and overall superior performance.

The proliferation of large language models (LLMs) has led to a significant surge in the size of AI/ML training models and jobs. For instance, the GPT-3 model, released in 2020, comprises 175 billion parameters and 300B tokens, a measure of the training dataset. GPT-4, released in 2023, has more than 1.7 trillion parameters. It's expected that even larger LLMs will be developed in the future, with GPT-5 anticipated to surpass its predecessor in parameter count. LLMs require a large AI/ML cluster for training.

A typical back-end AI/ML cluster network includes hundreds to thousands of AI/ML accelerators, CPUs, NVMe storage devices, one or two tiers of network switches, and NICs connected to the GPU or a PCIe switch. It's also worth noting that certain GPUs possess integrated NIC functionality.

To evaluate technology for an AI/ML training network, it is important to understand key requirements for supporting AI/ML workloads:

- **High-speed data transfer:** AI models require large amounts of data to be processed, which must be transferred through the network at high speeds. The network should have a fast and reliable connection to the data sources.
- **Scalability:** AI workloads can vary in size and complexity, and the network should be able to scale up or down depending on the workload requirements. This may involve adding or removing nodes from the network or adjusting the amount of resources allocated to individual nodes.

- **Robustness and reliability:** AI training can be a time-consuming and resource-intensive process. The network should be designed to minimize the risk of downtime or failure, and to recover quickly in the event of an issue.

MINIMIZING LLM TRAINING TIME HAS A DIRECT IMPACT ON BUSINESS ECONOMICS.

Training time for large models can take several days. It is essential to minimize the training time—also referred to as job completion time—as that has direct impact on business economics.

Broadcom provides two scheduled fabric solutions for AI/ML networks. The scheduled fabric can originate either at the switch or at the endpoint:

- **Switch scheduled fabric** solutions leverage the Jericho/Ramon family, optimized for AI/ML features and performance.
- **Endpoint scheduled fabric** solutions leverage the Tomahawk® family, with several enhancements to improve AI/ML performance.

In both solutions, the endpoint can be one of the following four options: Broadcom® NIC, Customer NIC, Merchant NIC, or Native Ethernet interface from the GPU/Accelerator.

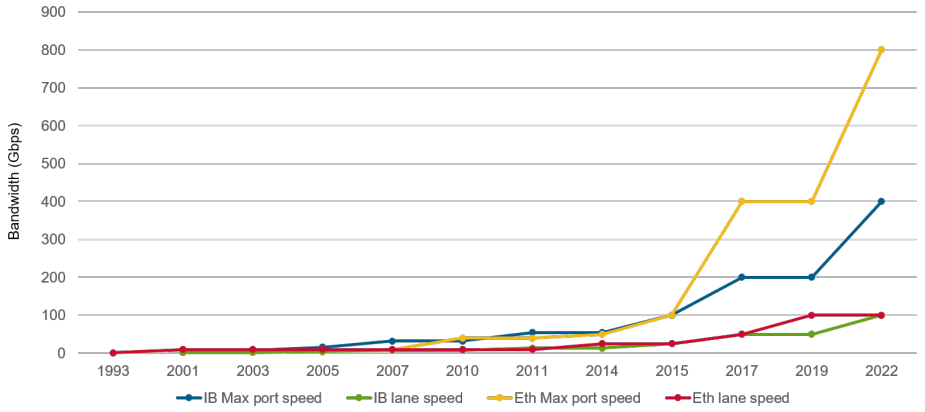
In the following sections, we will analyze the InfiniBand switch versus the scheduled fabric solutions across various attributes.

Port Speed, Radix, and Bandwidth

Ethernet has a rich and a vibrant ecosystem. The pace of port speed, radix and bandwidth is at least one generation ahead of InfiniBand switches.

THE PACE OF ETHERNET PORT SPEED, RADIX AND BANDWIDTH IS AT LEAST ONE GENERATION AHEAD OF INFINIBAND SWITCHES.

Figure 1: Pace of Port Bandwidth



The maximum port speed and radix of switches are critical for building cost-efficient and power-efficient networks. Figure 1 shows that the lane speed for Ethernet is ahead of InfiniBand over time, and there is a significant lead in port bandwidth for Ethernet compared to InfiniBand.

The commercially available InfiniBand switch has a bandwidth of 25.6Tb/s. In contrast, Broadcom began the mass production of Tomahawk 5, the world's only 51.2Tb/s switch, in 2023. This chip has been deployed in hyperscale data centers worldwide for compute, storage, and AI cluster connectivity. Table 1 details the market availability of InfiniBand and Ethernet switches.

THE TOMAHAWK 5 CHIP HAS BEEN DEPLOYED IN HYPERSCALE DATA CENTERS WORLDWIDE FOR COMPUTE, STORAGE, AND AI CLUSTER CONNECTIVITY.

Table 1: InfiniBand and Ethernet Switch Availability

Company	Switch	Sample Date	Bandwidth	Port Speed	Logical Ports per Node	400G Ports per 2-T Clos
	Switch-IB-2	*Nov. 2015*	*3.6T*	*100G*	*64*	
	Quantum	*Nov. 2017*	*8T*	*200G*	*64*	
	Quantum-2	*Nov. 2021*	*25.6T*	*400G*	*64*	*2048*
BROADCOM	*Tomahawk*	*Sept. 2014*	*3.2T*	*100G*	*128*	
	Tomahawk 2	*Oct. 2016*	*6.4T*	*100G*	*652*	
	Tomahawk 3	*Dec. 2017*	*12.8T*	*400G*	*128*	
	Tomahawk 4	*Dec. 2019*	*25.6T*	*400G*	*256*	
	Tomahawk 5	*2022*	*51.2T*	*800G*	*256+*	*8192*

The Tomahawk line of switches is expected to introduce port speeds of 1.6T in the next couple of years. The increased radix of Ethernet switches allows for a larger cluster size in a 2-tier Clos network.

The Jericho/Ramon family is not limited by the radix of an individual packet processing switch, but is limited by the fabric switch. With the Jericho/Ramon family, the entire cluster is a scheduled fabric, and it acts as a single traffic domain. The Jericho3AI device, announced earlier this year, has 14.4T (36 x 800G) of Ethernet interface and 14.4T+ of fabric bandwidth. Ramon3 is a 51.2T cell-based spine switch. With Jerico3AI/Ramon3 we can build a large cluster of 32K x 800G ports, which today's InfiniBand cannot match.

Addressing and Forwarding

An InfiniBand network is divided into subnets. Each subnet can have a maximum of 64K devices addressed, using a 16-bit forwarding identifier known as LID. However, the addressable nodes within a subnet are limited to 48K, as the remaining nodes are used for multicast. There is no known InfiniBand implementation operating at such a large scale. In contrast, there is no addressing limitation for Ethernet, and the size of the cluster depends on the address tables in the switches.

Multi-tenancy

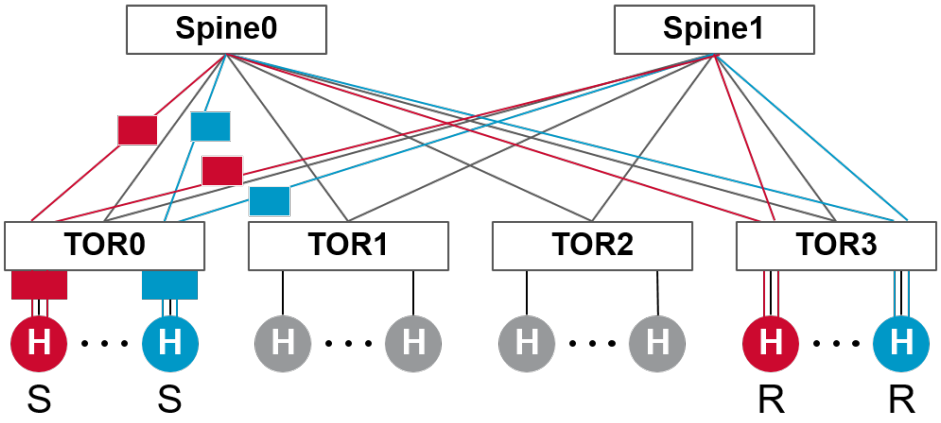
With Ethernet, multi-tenancy and multi-jobs can be easily handled through technologies such as VxLAN. InfiniBand has no mechanisms for tenant isolation. Instead, isolation is done through subnets, implying an InfiniBand router is required to connect these subnets, increasing latency and costs.

Multipathing and Routing

InfiniBand uses destination-based routing, where the intermediate switch performs LID lookup to forward the packet. InfiniBand load balancing is done by assigning multiple LIDs to a single port. This uses a concept called LID mask count (LMC), which allows multiple paths to be created between two end nodes. An LMC value of x allows 2^x paths between two end nodes. An LMC value can be between 0 and 7, so the maximum number of multipaths is 2^7, or 128 paths. The number of LIDs assigned to each port in a subnet is 2^x. This means a subnet enabled for multipathing with an LMC value of x can only use $48K/2^x$ end nodes.

Load Balancing and Congestion Management

Figure 2: Perfect Load Balancing in Jericho Scheduled Fabric



THE JERICHO FABRIC ACHIEVES THE LOWEST TAIL LATENCY DUE TO ITS PERFECT LOAD BALANCE AT A VERY HIGH NETWORK LOAD.

The Jericho family of scheduled fabric-initiated devices supports perfect load balancing independent of the hashing scheme. Packets are split into cells; cells are sprayed across all the available paths and reassembled at the destination switch. The Jericho fabric achieves the lowest tail latency due to its perfect load balance at a very high network load. In addition, the Jericho fabric provides the following features:

- End-to-end credit protocol for congestion avoidance; alternatively, InfiniBand only provides reactive congestion control
- Hardware-based link failure detection and rerouting
- Lower power, using its most efficient cell switch chips
- Large clusters in a single traffic and management domain

The Tomahawk line of switches has extensive support for ECMP/WCMP and dynamic load balancing, distributing flows among links based on the dynamic load. The packet order is still maintained. Adaptive routing is also supported to avoid congestion hot spots in the network.

On the InfiniBand side, the Quantum-2 switches support SHIELD, which can reroute in the presence of link failures. However, the Jericho fabric supports link failure detection and rerouting in less than 10 ns. The Tomahawk family supports global load balancing (GLB), which is about 10x faster than the InfiniBand capability, as measured by large cloud customers.

Large-Scale Challenges

InfiniBand supports link-level credit flow control. Link-level flow control is like priority flow control (PFC), except it is more granular. The limitations of PFC in large-scale networks are well documented with congestion spreading and deadlocks, and would exist in InfiniBand network as well with link-level credits. On the other hand, the Jericho3/Ramon3 family supports receiver-driven protocol, an end-to-end credit-based protocol that provides the sender credits based on the receiver's ability to drain. This also manages the incast very well, which InfiniBand networks do not handle efficiently.

THE ETHERNET SWITCHES AVAILABLE TODAY PROVIDE A RICH SET OF VISIBILITY METRICS.

The typical distance supported on InfiniBand's link level retry (LLR) support is 30 meters. In a large-scale AI network, the link distances are on the order of hundreds of meters. So, although LLR reduces latency by retransmission when there are link errors, the mechanism is practically useless in an AI/ML cluster.

Telemetry

Network fabrics should be able to troubleshoot link failures, anomalies, link utilization, packet size distribution, and traffic monitoring (IPFix/SFlow). The Ethernet switches available today provide a rich set of visibility metrics. Unfortunately, the InfiniBand switches do not; they only provide some basic counters. This lack of in-depth metrics means an InfiniBand network may not yield much granular visibility for troubleshooting and debugging failure or congestion events.

Total Cost of Ownership: Cost and Power

Since InfiniBand lags behind Ethernet in fabric bandwidth and port speed, the network cost of building an equivalent cluster is significantly better with Ethernet than InfiniBand. This is illustrated in Table 2, which shows a direct comparison of solutions in two different cluster sizes.

WITH THE CURRENT GENERATION OF ETHERNET, WE CAN ACHIEVE A 4x SYSTEM SCALE FOR 2-TIER WITH 3x FEWER SWITCHES THAN INFINIBAND.

Table 2: Scale Comparison for Equivalent Cluster Sizes

	Ethernet			InfiniBand		
	TH5	Tiers	Cables	Quantum-2	Tiers	Cables
256 Nodes of 200G	*1*	*1*	*256*	*6*	*2*	*512*
32K Nodes of 200G	*192*	*2*	*64K*	*640*	*3*	*96K*

With the current generation of Ethernet, we can achieve a 4x system scale for 2-tier with 3x fewer switches than InfiniBand.

Power from networking fabrics plays a key role. InfiniBand switches have lower radix and port speeds compared to Ethernet switches, which means the overall power for an InfiniBand network is significantly higher than Ethernet fabric for an equivalent large-scale AI network with high port speeds. This is because the InfiniBand switches commercially available today would require more tiers and optics, which significantly increases the overall network power[1].

[1] Hugo Touvron*, et.al, "Llama 2: Open Foundation and Fine-Tuned Chat Models", July 2023, https://arxiv.org/abs/2307.09288

ETHERNET SWITCHES CAN HAVE A DEDICATED MANAGEMENT NETWORK AND DO NOT CONSUME PREMIUM BANDWIDTH TRAFFIC FOR MANAGEMENT PURPOSES.

Management

InfiniBand management uses in-band datagrams referred to as MAD frames for subnet management, telemetry, and operational state query. MAD frames have a minimum MTU size of 256. Though MAD frames occupy a dedicated QP, they still adversely impact the network bandwidth when a large number of queries are done by the control plane application. High network load can also cause MAD frames to sit in the switch buffer for a longer time, thus adversely impacting the control plane convergence time[2].

Ethernet switches can have a dedicated management network and do not consume premium bandwidth traffic for management purposes. The MTU size for the management frame is 9k (jumbo frames), reducing the need for multiple messaging.

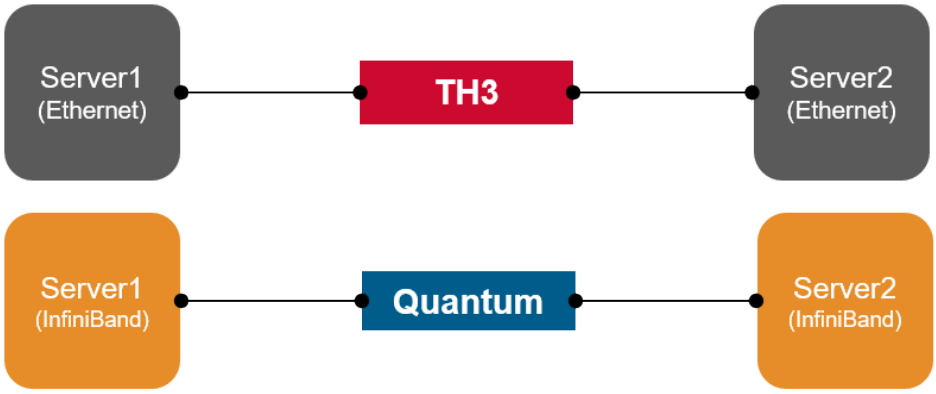
Performance

AI/ML network performance is compared using micro-benchmarks, collective benchmarks, and application benchmarks. In this section, we will provide a sample comparison between Ethernet and InfiniBand using micro-benchmarks and collective benchmarks.

Micro-benchmark Comparison

Figure 3 documents the setup for the micro-benchmark comparison. The OSU Latency and OSU bandwidth benchmarks were run in the lab.

Figure 3: Micro-benchmark Setup for Performance Comparison



Test Setup Description	
Parameter	**Test Condition**
Devices	*Server 1 - 100G NIC* *Server 2 - 100G NIC*
Connectivity	*Ethernet (RoCEv2) connected via TH3*
	InfiniBand connected via Quantum
Measurement	*Latency - message transfer latency between two servers*

Test location: *Broadcom® labs* **Benchmark:** *Osu_latency*

[2] Sjur Tveito Fredriksen, 2017, Thesis: Designing an InfiniBand Metric Collector and Exploring InfiniBand Management Overhead and Scalability, https://www.duo.uio.no/bitstream/handle/10852/59275/1/msc-sjurtf.pdf

Figure 4 and Figure 5 show the latency and bandwidth comparison.

ETHERNET CLEARLY PROVIDES COMPARABLE PERFORMANCE ON BOTH MICRO-BENCHMARKS.

Figure 4: OSU Latency Micro-benchmark Results

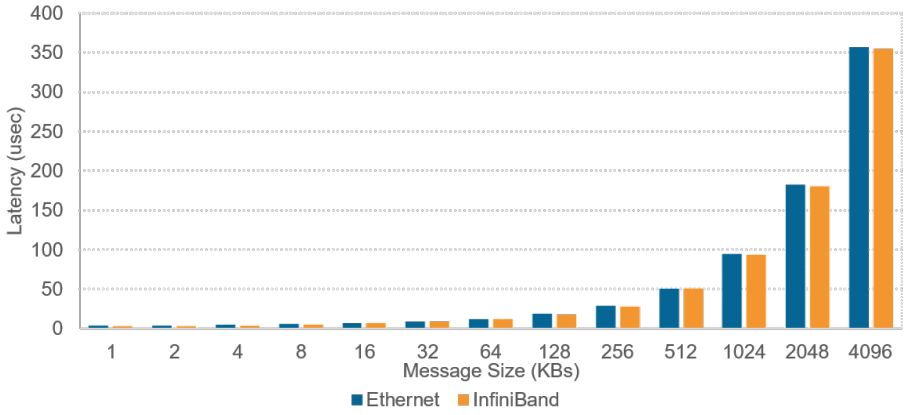
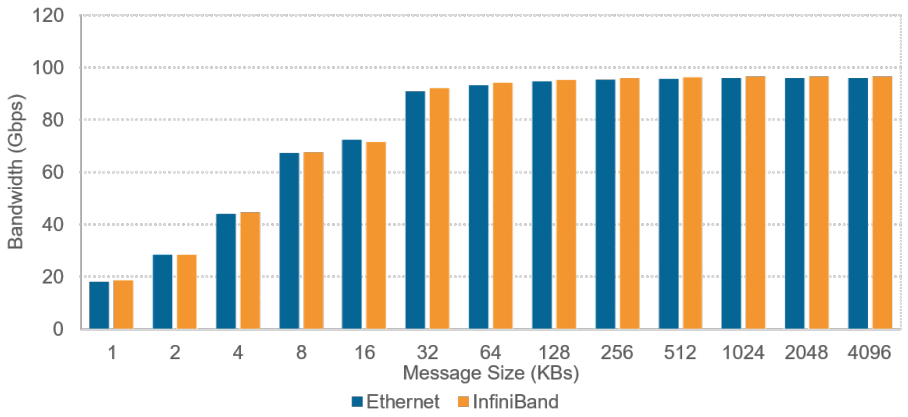


Figure 5: OSU Bandwidth Micro-benchmark Results

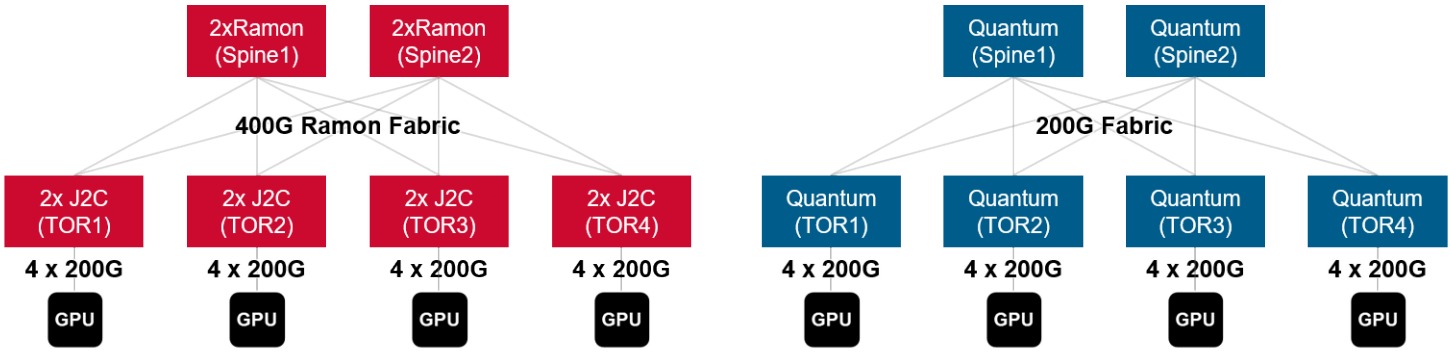


The data in Figure 4 and Figure 5 demonstrates that Ethernet clearly provides comparable performance on both micro-benchmarks.

Collective Benchmark Comparison

A few hyperscale customers compared NCCL test performance between the Jericho fabric and InfiniBand. Figure 6 documents the test setup.

Figure 6: Test Setup for NCCL Benchmark at a Hyperscaler, Jericho Fabric vs. InfiniBand



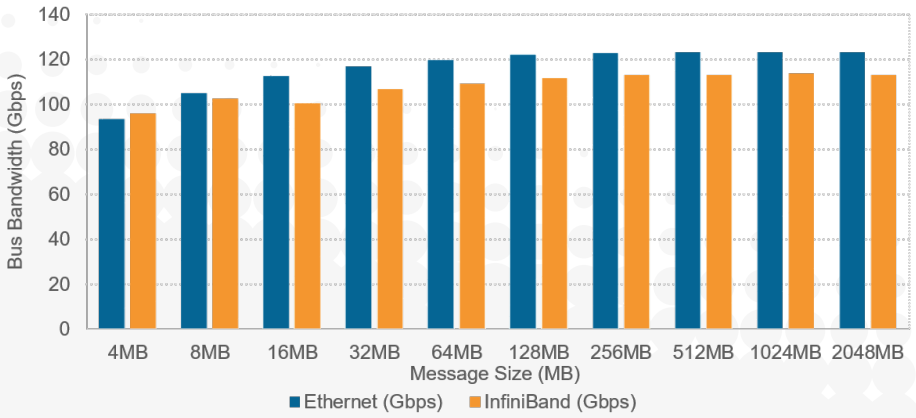
Test Setup Description, Ethernet	
Parameter	**Test Condition**
Devices	• *Total of 4 GPU servers; each server includes the following:* – *8x A100 GPUs* – *4x CX-6 NICs (200G)* • *4x TORs, each with 2x J2C cards* • *2x Spines, each with 2x Ramon fabric*
Connectivity	• *200GbE between server and TOR* • *400G between TOR and Ramon spine*
Test	*MOE-T5: message size from 60 MB to 100 MB*
Measurement	*Bandwidth: average transfer data rate*

Test Setup Description, InfiniBand	
Parameter	**Test Condition**
Devices	• *Total of 4 GPU servers; each server includes the following:* – *8x A100 GPUs* – *4x CX-6 NICs (200G)* • *4x TORs with Quantum InfiniBand switches* • *2x Spines with Quantum switches*
Connectivity	• *200GbE between server and TOR* • *400G between TOR and Ramon spine*
Test	*MOE-T5: message size from 60 MB to 100 MB*
Measurement	*Bandwidth: average transfer data rate*

Test location: *Customer labs* **Benchmark:** *NCCL tests*

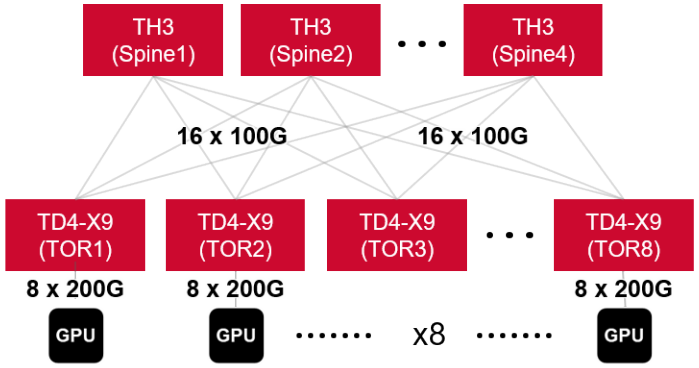
Figure 7 shows the NCCL all-to-all bandwidth results. The Jericho fabric, with its superior load balancing and congestion management, delivers a 10% improvement over InfiniBand. When running large training jobs, this 10% improvement can translate into *several days* of reduced job completion time.

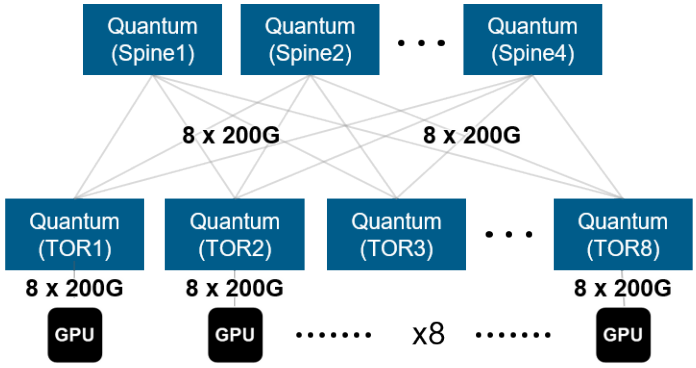
Figure 7: NCCL All-to-All Bandwidth, Jericho Fabric vs. InfiniBand



Another hyperscaler compared NCCL test performance between the Tomahawk family and InfiniBand. Figure 8 documents the setup.

Figure 8: Test Setup for NCCL Benchmark at a Hyperscaler, Tomahawk Family vs. InfiniBand





Test Setup Description, Ethernet	
Parameter	**Test Condition**
Devices	• *Each GPU includes 8x A100 GPU cards* – *GPU includes Nvidia CX-6 NIC* • *TD4-X9 switches in TOR* • *TH3 switches in Spine*
Connectivity	• *CX-6 connects to TOR switch via 1x200G Ethernet* • *TOR connects to spine by 100G links* • *4-ECMP group, each with 4x 100G from Spine-to-TOR* • *1-ECMP group of 16x 100G from TOR-to-Spine*
Measurement	*Bandwidth: average transfer data rate*

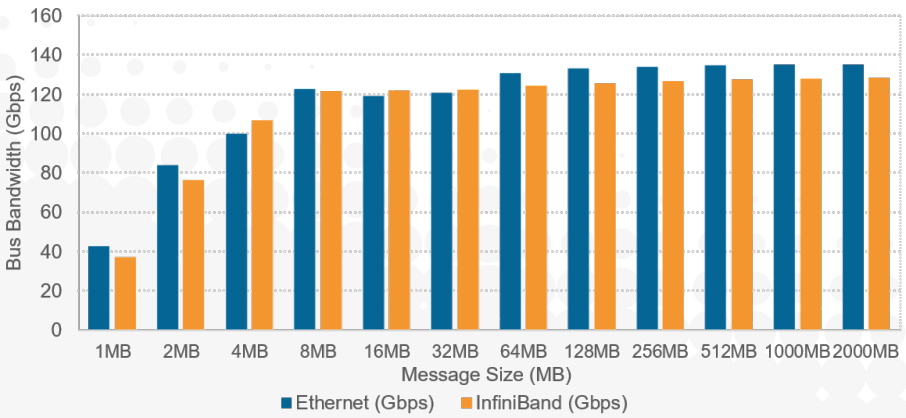
Test Setup Description, InfiniBand	
Parameter	**Test Condition**
Devices	• *Each GPU includes 8x A100 GPU cards* – *GPU includes Nvidia CX-6 NIC* • *Quantum switches in TOR/Spine*
Connectivity	• *CX-6 connects to TOR switch via 1x200G InfiniBand*
Measurement	*Bandwidth: average transfer data rate*

Test location: *Customer labs* **Benchmark:** *NCCL tests*

Figure 9 shows the NCCL bandwidth results. The Tomahawk family delivers comparable performance to that of InfiniBand.

The scheduled fabric solutions deliver equivalent or better job completion times than InfiniBand.

Figure 9: NCCL Bandwidth Test, Tomahawk vs. InfiniBand



SCHEDULED FABRIC SOLUTIONS DELIVER EQUIVALENT OR BETTER JOB COMPLETION TIMES THAN INFINIBAND.

Table 3 shows a comparison between Ethernet and InfiniBand across various attributes.

Table 3: Ethernet vs. InfiniBand, Comparative Metrics

AI Cluster Connectivity Features/Attributes	InfiniBand	Scheduled Fabric	
		EP	Switch
Fabric Bandwidth Pace	🔴	🟢	🟢
Port Speed	🔴	🟢	🟢
Dynamic Load Balancing	🟡	🟢	🟢
Perfect Load Balancing	○	🟡	🟢
End-to-End Congestion Management	○	🟢	🟢
Fabric Management	🟢	🟢	🟢
Telemetry	○	🟢	🟢
Tail Latency Performance	🔴	🟡	🟢
Total Cost of Ownership, Power and Cost	🔴	🟢	🟢
Multi-job, Multi-tenancy	🔴	🟢	🟢
Multi-vendor Support	🔴	🟢	🟢

BROADCOM SCHEDULED FABRIC SOLUTIONS ARE ALIGNED WITH THE VISION OF UEC AND ARE OPTIMIZED TO PROVIDE SUPERIOR AI/ML NETWORKING PERFORMANCE.

Conclusion

Ethernet has all the essential features required for a top-performing AI/ML training cluster, such as high bandwidth, efficient end-to-end congestion management, load balancing, fabric management, and more cost-effective than InfiniBand. Furthermore, Ethernet has a diverse ecosystem of numerous silicon vendors, OEMs, ODMs, cables, optics, software, and continuous innovations. The recently launched Ultra Ethernet Consortium (UEC) aims to standardize features for high-performant networks for large-scale AI/ML and HPC networks, furthering Ethernet technology's deployment[3,4] and democratizing an already vibrant ecosystem.

The Broadcom scheduled fabric solutions are aligned with the vision of UEC and are optimized to provide superior AI/ML networking performance.

[3] Jag Brar and Pradeep Vincent, "First Principles: Superclusters with RDMA—Ultra-high Performance at Massive Scale", Feb. 2023, https://blogs.oracle.com/cloud-infrastructure/post/superclusters-rdma-high-performance
[4] Leah Shalev et.al "A Cloud-Optimized Transport Protocol for Elastic and Scalable HPC", IEEE Micro, Volume: 40, Issue: 6, 01 Nov.-Dec. 2020

BROADCOM
connecting everything®

For more information, visit our website at: www.broadcom.com

AI-ML-Ethernet-WP100 December 4, 2023