WHITE PAPER

SUPERMICRO RACK SCALE AI NETWORKING WITH ETHERNET

Building AI Networks Using Ethernet Technologies

TABLE OF CONTENTS

Executive Summary

Supermicro is a leading and trusted technology partner at the forefront of the data center industry. Supermicro provides optimal Total Cost of Ownership (TCO) solutions across various domains, including Data Centers, AI, and HPC applications. It has an extensive range of servers, storage solutions, and switches in its portfolio.

Supermicro is a leading supplier of systems for Generative AI, working with our partners up and down the technology stack. As such, customers trust Supermicro with tested designs and solutions that simplify the complexity of these types of deployments. While the components in the AI cluster typically deliver very high performance, other considerations go beyond the speeds and feeds to allow for optimal use of these resources in each solution.

This paper will discuss some AI networking alternatives and show some tested designs based on industry leading components. If we look at the overall market forecast, 650 Group estimates that in 2027, the overall AI Networking market will be $10B, with Ethernet accounting for $6B of that total, with the remainder being InfiniBand. As such, this document will address some key points in each, as both will remain very important.

Considerations for the Networking Component in an AI Training Cluster

To understand the impact of the network, we will first briefly overview the general process that training these large models follow. The general process follows a distribution of a subset of Large Language Models (LLMs) and data to be trained onto a cluster of systems – each working on its own portion of the overall data set. The systems will then perform compute-intensive operations of deriving tensors from extremely large sparse matrices based on the model at hand and the data they were given. When each node completes the work, it needs to exchange the information with all other systems in the given cluster, and the individual system waits for all other nodes to receive all system outputs. These nodes then merge all that data with their own and then proceed to the next iteration of calculations. These outputs are constantly assessed until the job is completed.

Many factors come into play when considering optimizing the overall Job Completion Time (JCT) in AI training. The iterative process of parsing out the data sets, having many parallel systems train on that subset of data, then merging the results between the cluster members and repeating them during the training often involves thousands of systems and many weeks of computation. In this paper, we will focus on that portion of the process where the systems must communicate over an interconnect fabric between their peers. Examples of these communications include sharing data sets, iterative results, and overhead synchronization operations, among many others. At the 2022 OCP Global Summit, Alexis Bjorlin from Meta shared the information below on the testing performed to quantify the impacts of time spent in these communication mechanisms, as opposed to the core training on the models. Figure 1 shows the results below.

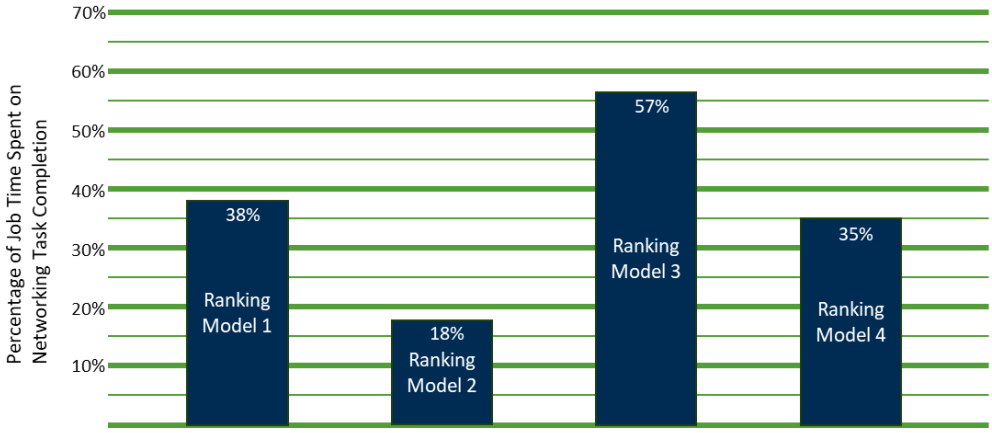

Figure 1 - From OCP Keynote by Alexis Bjorlin (MEta) at 2022 OCP Global Summit

While the results show a varying level of impact, the overall conclusion is that the importance of optimizing the system for network I/O can have a meaningful impact on the overall JCT.

Traditional AI Fabric with Infiniband

This communication has often utilized an InfiniBand (IB) fabric, which initially was optimal for High Performance Compute (HPC) markets, and many of its characteristics complemented the needs within these AI fabrics. Some examples of these optimizations are:

- Remote Direct Memory Access (RDMA)
- Lossless transmission and reception via end-to-end buffer management
- Low overhead to minimize latency in the fabric
- Adaptive Routing to restore communication rapidly
- Message Aggregation (not exclusive to IB but integrated in today's silicon)

RDMA

Generally, the RDMA concept has been around in IB for many years, where systems could bypass the kernel data structures and buffers and minimize the intermediate data movement between these layers on a data transmission on the sender side and the reception on the receiver side. Whereas traditional TCP/IP and UDP/IP data transmission make use of the kernel network drivers and stack, which all use the host CPU, the RDMA bypasses those layers and places data directly from the application memory to the wire and likewise from the wire directly to the application memory. In modern AI systems, the "application" points to the GPU systems' memory space, bypassing kernel memory. In a system with many GPUs but only a few CPUs, we don't want the CPU copying and processing of frames to become the bottleneck. Figure 2 below shows an illustration of the concept.

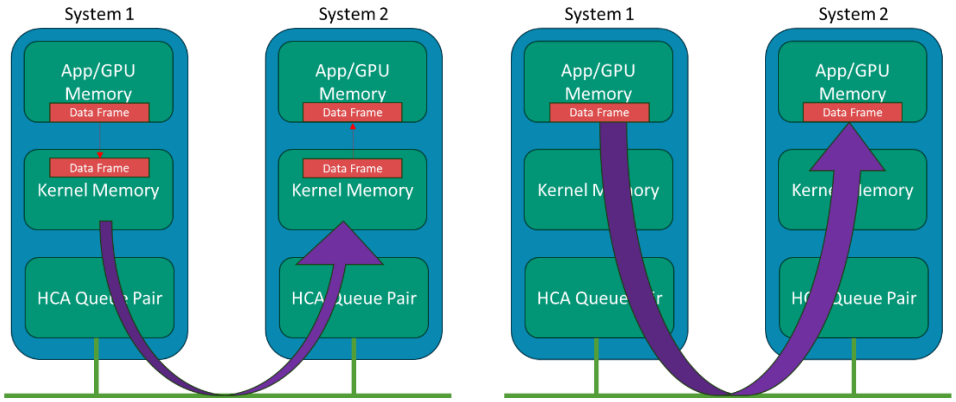

Figure 2 - With and Without RDMA

Lossless Transmission & Low Latency

Traditional packet networking involves some assumption of a percentage of the data being lost in transit due to overflowing buffers, packets failing integrity checks due to bit errors, and others. InfiniBand has built-in queue management to handle the cases where buffers could be overrun, and the Forward Error Correction (FEC) mechanisms today handle most cases of bit errors in transit. These lead to lossless behavior on the fabric.

The IB header is small and rigid enough to allow for a very low latency of traffic that would transit any node, and when combined with the RDMA discussed earlier, the result is a low-latency traffic solution.

Adaptive Routing

Infiniband architecture mandates an item called a "Subnet Manager" which has overall responsibility for monitoring the behavior in the fabric via periodic sweeps and messaging between the fabric elements. By itself, there would be a risk of sub-optimal behavior in accounting for link and node outages, but this is handled via adaptive routing, where the forwarding of packets also takes into consideration the queue depth on egress interfaces. If filling up based on congestion, it will choose another ECMP member to send the packet to. This provides a much faster means to re-route traffic around links affected by these events. Adaptive Routing in an IB fabric does introduce the chance of out of order packet arrival – but modern HCA silicon can correct these. Figure 3 below shows a simple illustration of this situation where no adaptive routing is present.

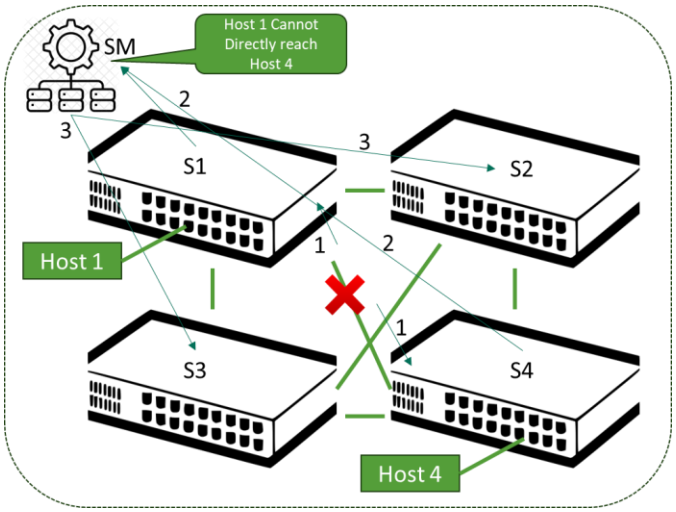

Figure 3 - Without Adaptive Routing - Subnet Manager Coordinates on Link loss or link congestion

Message Aggregation

One important area of parallel processing systems is the means to share the outputs of each node within the larger system with its peers. While methods to aggregate these outputs are not new, the extension of these into AI networks is something that came along as these systems grew. The overall discussion behind methods here is beyond this paper's scope, but we will attempt to summarize the key points based on how this occurs today. To help illustrate the issue, we will use Figure 4 below.

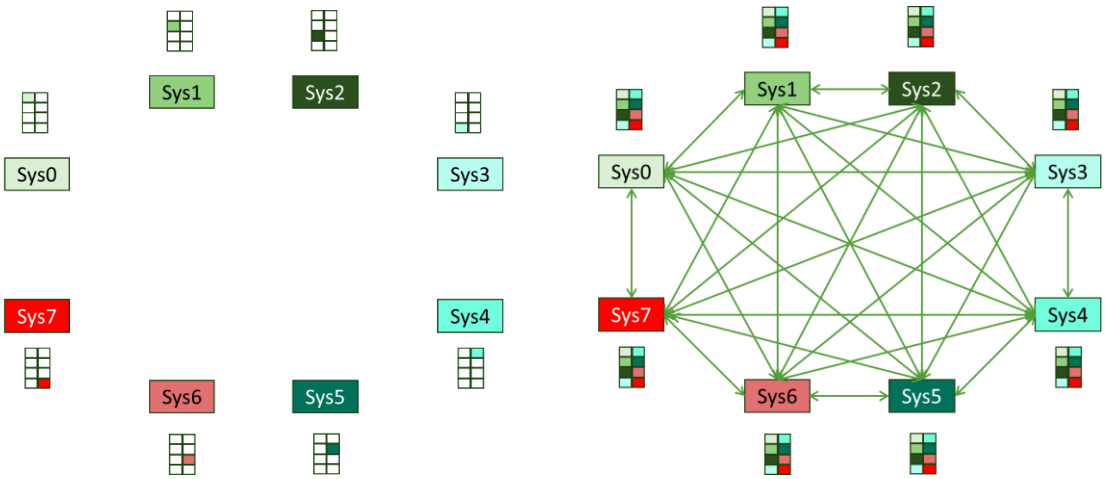

Figure 4 - Messages between nodes that occur at each iteration in the learning model

This leads to many individual nodes communicating on the fabric with many others, and each message has its own connection setup, transfer, and teardown on the hosts and fabric. The idea of developing a method to simplify this through some intermediate aggregation was added to the InfiniBand architecture2 (although the mechanism does not mandate IB). The concept of message aggregation and a more efficient means to minimize the counts of interconnects would be increasingly important as the node count scales into the thousands and tens of thousands of nodes in a cluster. Intermediate nodes performing this message aggregation task, along with the duties of transporting data itself, were conceived and are utilized today.

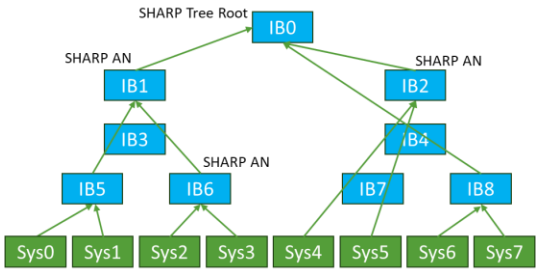

Figure 5 - Simple Illustration of a SHARP tree

By employing these techniques of message reduction, the number of messages required between nodes scale is far better than the N*(N-1)/2 of the full mesh.

Ethernet as an Alternative to InfiniBand

Many customers of Supermicro AI solutions continue to deploy with InfiniBand as the fabric of choice. We embrace these designs with our industry partners to bring the best solutions to our customers. Some of those, however, along with many new customers, are evaluating or already using Ethernet fabrics as an alternative to IB for some of the following reasons:

- Much larger and open ecosystem
- Attractive economics with merchant silicon and open software options
- Unmatched interconnect speeds and growth trajectory
- Larger scale domains possible in Ethernet – even further with routed traffic
- Existing expertise and tooling in deploying, managing, and observing Ethernet – no desire to learn a new fabric if they have no IB experience
- Many of the above IB technical capabilities are also addressed in Ethernet – more being added rapidly

Before we delve into the technical review of some elements in Ethernet, we will first show a high-level traffic comparison between Ethernet and IB as it exists today from the perspective of I/O performance. On the Ethernet side, we have RDMA by using RDMA over Converged Ethernet (RoCE that will be discussed below), which provides a lossless setup via Priority-Based Flow Control (PFC that will be discussed below) alongside Explicit Congestion Notification (ECN) to do these traffic performance measurements. Figure 6 shows the results over a variety of message sizes. These tests were done with EDR InfiniBand, and the comparison was 100GE. The general takeaway was that if we have a general-purpose Ethernet domain (switches, adapters, etc.) that would not suffice for the needs of an AI network that was originally on IB, but an Optimized lossless Ethernet and RoCE very closely approximated the performance of InfiniBand.

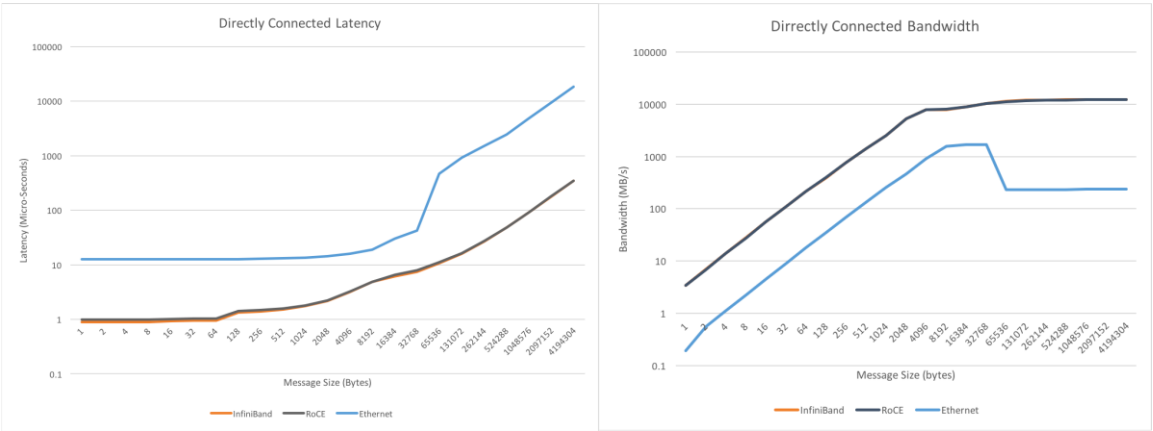

Figure 6 - Ethernet and IB Test[3] Presented at SC16

The above figure shows that the raw performance from I/O is consistent, with Ethernet having a slight edge. To get the full picture in a practical AI networking scenario, however, we also need to account for the various optimizations present in the IB world, which would not show on a raw performance graph. A list of these technologies includes:

- Direct GPU to GPU communications
- Raw speed & latency
- Lossless behavior & congestion avoidance
- Message aggregation – non-specific to the fabric deployed
- Link failure and re-routing
- Path segmentation
- Load balancing

Direct GPU to GPU Communications

As discussed earlier in this paper, IB uses the RDMA to directly map I/O to/from the wire to the memory on the GPU. In the Ethernet domain, this has been adopted by using RDMA but transported over Converged Ethernet – which means this Ethernet is now converging to transport traffic that traditionally has been on its own fabric (i.e., InfiniBand). RDMA has been enhanced to offer more scale and physical distance by adding a method to route this traffic in version 2 of the specification. The result is commonly referred to as RoCEv2 and is ubiquitous in the Ethernet domains and on the Network Interface Cards (NICs) in the systems. An illustration is in Figure 7 below.

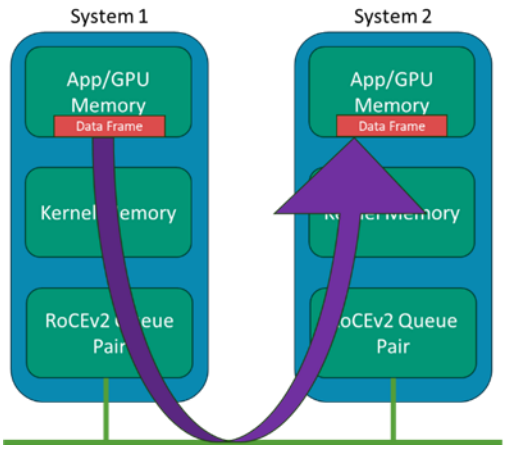

Figure 7 - Simple Illustration of RoCEv2

Raw Speed and Latency

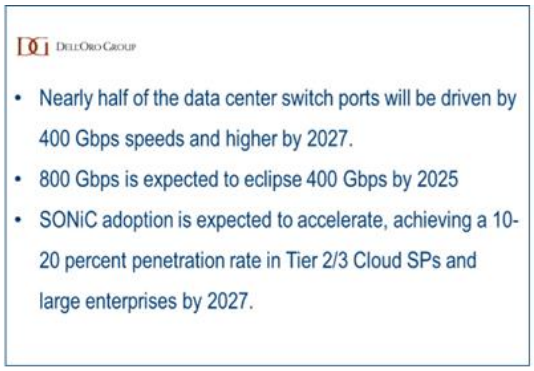

Figure 8: Del'Oro 5 Year Data Center Switching Report June 2023

Dell'Oro estimates on the Ethernet port market are on the left, and the 800GE ports becoming dominant over 400GE in 2025 shows the rate of current innovation in speed, with high-speed ports being dominant in the DC by 2027. As we look further at silicon manufacturer roadmaps, we see 1.6T Ethernet ports gaining rapid penetration in 2027. This pace of growth is not showing any meaningful decline as of late 2023.

Within Ethernet networks, we define a port-to-port latency for a frame to ingress, be scheduled, and egress a giving switching element. In this area, a typical Ethernet element may vary considerably (ranging from ~250ns to ~3000ns) depending on the hardware and what is happening to the frame as it transits the device. On a modern ethernet switch with equivalent speed ingress to egress ports, cut-through forwarding provides for very low latency, as shown in Figure 8.

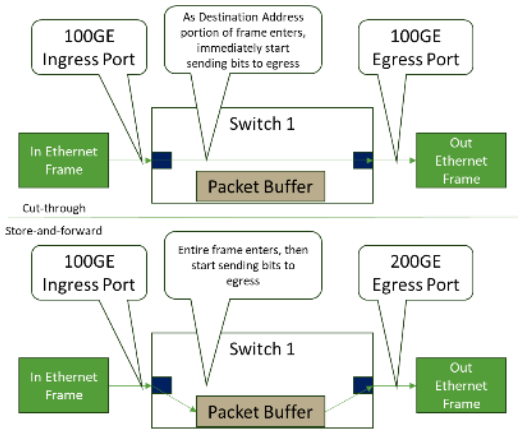

Figure 9 - Cut-Through Switching

InfiniBand can be in the range of ~200ns to 300ns for the same hop, meaning this area on fabric IB has an advantage as this transport does not have the same range of application types. Getting the latency as low as possible is desired for many AI networking needs. There is also an element many use in discussions around "tail latency," which is the latency of the last message in a chain of messages and is important in AI clusters as all nodes wait for all data to be updated before moving to the following setup. This may mean nodes are idle while others await the last messages, meaning that is one very important metric over the whole fabric. This level of visibility is more impactful than looking at a switching element latency, and this is where the Ethernet industry is optimizing between the elements themselves and the methods to balance and distribute the traffic at the fabric level. In that realm, the many Ethernet toolsets that exist to optimize these flows bring a latency discussion to a general comparable result between these technologies.

Lossless Behavior, Congestion Avoidance, & Path Segmentation

The concepts of enabling lossless behavior and avoiding congestion on Ethernet is a longer discussion and beyond the scope of this document (as is lossless methods in IB).

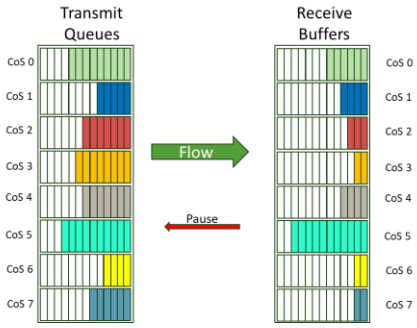

Figure 10 – Priority Flow Control

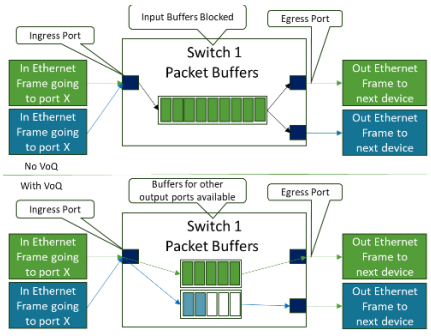

Figure 11 – Virtual Output Queueing

The individual enablers include Per-Priority Flow Control (PFC) (example in Figure 10) to allow a pausing of traffic by class of service.

This is further extended with Virtual Output Queueing (VoQ) to ensure no flow can head-of-line block other flows by filling port buffers when that flow is congested, as shown in Figure 11.

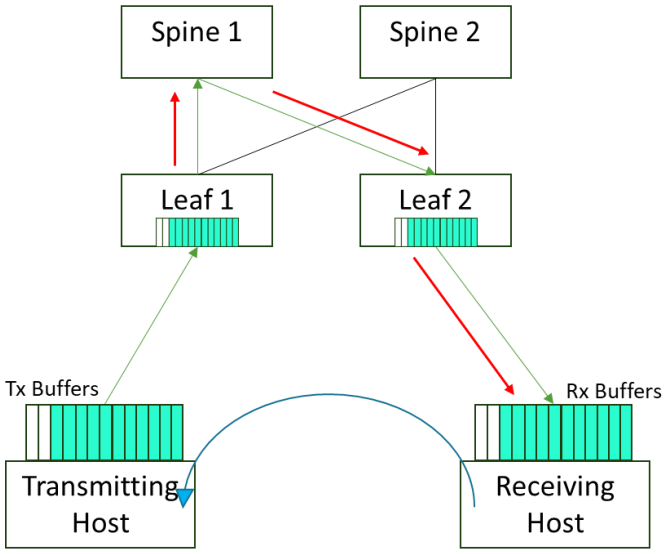

Figure 12 – Explicit Congestion Notification

Another key to managing congestion to allow lossless behavior is shown in Figure 12, where an Explicit Congestion Notification (ECN) is signaled by setting a Congestion Experienced (CE) value in the DSCP portion of the IP header on these intermediate switches to allow the receiver to inform the sender to slow down so that traffic is not lost. The mechanisms to monitor, detect, and signal all of this are built into modern switching silicon and are already present in a RoCEv2 aware network fabric. These methods will also monitor congestion on certain links and are used as inputs to other protocols to control traffic on alternative links for effective path segmentation and load balancing (covered below).

Message Aggregation – On the Fabric vs. On the GPU systems

In some customer discussions, message reduction is viewed as necessary as the AI cluster scale grows. The idea of doing this function on fabric nodes, however, does create some hesitation, whereas the concept of in-network-computing and performing app level transformations on devices that were reserved for and managed by a team purposed with fabric management and availability. To restate, the idea of keeping the network nodes just focused on networking concepts was often an item of conversation. The idea is that message reduction does not mandate IB but could extend to Ethernet (although we are unaware of any implementation in the market today – only early experimentations by some groups).

Another method to review is to look at having the system nodes to this reduction themselves – for scale out, ownership of operations and maintenance, etc. There are methods in the industry to allow for these optimized messaging mechanisms, as shown in Figure 13 below.

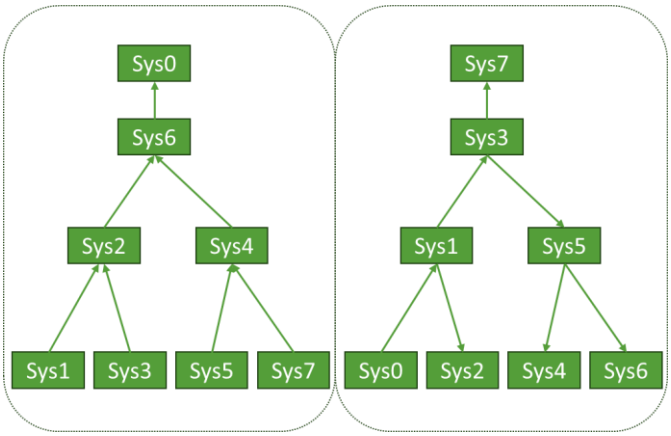

Figure 13 - Message Reduction on Systems

Considering the 8-node small cluster above for illustrative purposes, the concept is to form 2 non-overlapping binary trees, and the message flows between them are aggregated in the direction of the arrows. It can be seen that the 14 message exchanges here are larger but half as frequent than the 28 messaging connections shown in Figure 4 above for a fully meshed system. As the clusters expand to thousands or tens of thousands of nodes, this math and all the possible permutations of traffic to manage and monitor will remain far more manageable. Thus, the advantage of the earlier mentioned SHARP can be realized at the system level instead of on intermediate transit nodes. Again, some customers will still deploy and advocate for SHARP on IB, and Supermicro is happy to work with either situation regardless of the technology used.

Intra-DC Border Gateway Protocol (BGP) as a means to approximate IB adaptive routing and Load Balancing on the fabric

Ethernet has used Layer2 mechanisms over its history (Spanning Tree and other more modern methods) to ensure a loop-free path in the fabric for hosts to communicate. When a link or node is lost, the directly connected device will remove entries from MAC tables using the failed pathway. Traditionally, timers would need to expire for elements in the fabric to look for alternate paths (Static Routing). Alongside the loss of link/device cases, the congestion case was generally not handled until the introduction of PFC, ETS, DCQCN with ECN, WRED, etc., on the Ethernet fabric. To enhance both the link and node failure recovery, along with effectively adapting for congested links, many end users are now making use of methods like BGP routing within the DC to bring more active adjustments to the fabric in a means somewhat close to what Adaptive Routing on InfiniBand performs depending on if equal-cost or unequal-cost load balancing is utilized. These mechanisms are very effective at tracking and adjusting traffic patterns and re-establishing pathways without human interaction. Figure 14 below shows an example of the topology (from a public paper by Facebook). In this scenario, the leaf's will advertise the reachability of the connected systems with the spines (and other leaf's learn also) such that effective balancing can occur when combined with information about link utilization and other metrics. With the EVPN type of BGP, we can interconnect multiple leaf-spine fabrics to extend the size of the clusters (as a possible alternative to InfiniBand Dragonfly+ topologies, for example) and allow direct connectivity between nodes as required with Ethernet Dragonfly+, Super Spine, and/or Super Spine Plane designs.

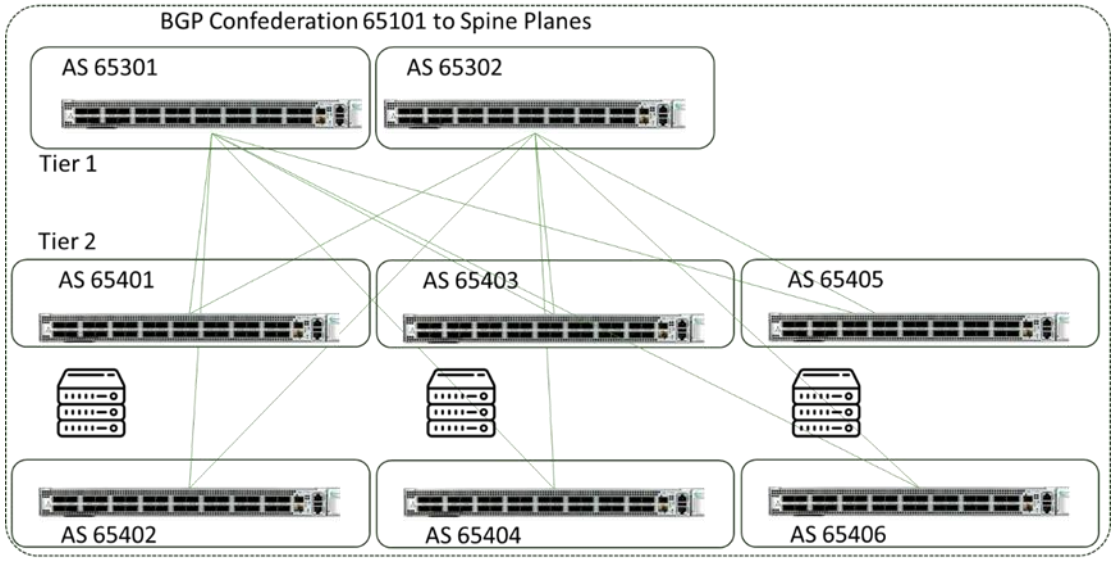

Figure 14 - BGP example from Meta4 for ethernet Leaf-Spine architecture for maximum availability and traffic balancing

Futures – Ultra Ethernet Consortium

Ethernet has long existed in many areas of a typical IT infrastructure, and the open nature of the technology is adaptable to new traffic types. This, however, does not mean existing devices can actively participate in these new capabilities, but they are always made to be compatible with any additions.

Due to the rapid growth of generative AI needs within many companies, Ethernet is again being expanded to allow for even new types of methods to enable solutions to be engineered with further optimizations to enhance AI performance and minimize the network impact on JCT. At the time of this paper, the industry is rapidly converging on "Ultra Ethernet Protocol," which allows for better link utilization and traffic management mechanisms in AI and other frameworks. The interested reader can go to https://ultraethernet.org to learn more. Many of Supermicro's partners and suppliers are participants in this consortium.

A note on other referenced fabrics:

With the interest and rapid growth of AI interconnect systems, there are many alternatives in various phases of development and availability in the industry. Supermicro already offers solutions for some of these, and we are open to wherever the market takes us in the future with our customers. Some common areas that are growing in interest include:

NVIDIA NVLink™

As the industry slowly grew to newer and faster forms of interconnect to/from the host CPU complex and links between accelerator devices, NVIDIA developed a mechanism to overcome the current limitations in the industry in NVLink. This technology can be reviewed with NVIDIA and is very common within a system or some number of systems within certain distances of each other (often intra-rack) – there may be needs to transport even further over IB or Ethernet. Building out a many thousand node cluster with just NVLink may be a challenge, but the reader can search the NVIDIA website.

AMD Infinity Fabric™

AMD has developed an architecture for inter and intra "Graphics Compute Die" interconnections via Infinity Fabric Links. These links interconnect not only a pair of the Die within a GPU but also the linkages between them. These GPU interconnections are for very high bandwidth in between with a general bi-directional loop topology. To extend beyond the local grouping, the PCIe interfaces are used with Ethernet as the transport.

Compute eXpress Link (CXL)

CXL is an emerging interconnect technology that builds over the PCIe foundation to allow new deployment models. There are many elements around maintaining the coherency of the cached elements in a system or set of systems, offering a dynamic method of interconnecting various devices between the host and/or themselves, allowing for a disaggregation of elements within a system, and many other advantages. The same question as above around the effective diameter of these elements comes into play, with some companies engineering solutions to aggregate over wider diameters, often using Ethernet again. CXL is growing in v3 to allow for tiering of elements and routing between. As such, the underlying Ethernet framework would still need to be robust, with many features outlined in this document.

Peripheral Connect Interconnect (PCIe) Switching

A common traditional method was extending the PCIe lanes from a system to an integrated or top-of-rack switching element. This works for smaller numbers of components and systems as the lane count on today's PCIe switching silicon is well below almost any interconnected GPU cluster without encapsulating within Ethernet again.

Supermicro Recommended Designs for AI Cluster Networks

Supermicro designs for AI networks are similar to the networking component across many of our partner vendors. Supermicro makes and markets an 800GE switch with 2x400GE possible per port to support these designs, but similar models from the market leaders will fit in, given the correct port speeds and configurations. The foundational technical software features and functionality deployed differ, but at Supermicro, we use the industry standard open Software for Open Networking in Cloud (SONiC). All the Ethernet features and functionality referenced in this paper are included in SONiC. To gain more familiarity with SONiC, we encourage tutorials and deploying a virtual switch environment. Supermicro will also be releasing a white paper on an optimized SONiC configuration for various example clusters in the near future. More customers and partners are deploying Supermicro switches due to the short lead times, the desire for a fully validated solution rack from fewer component vendors inside, and the bench strength of all things AI that we provide. We will show a very brief overview below:

SSE-T7132 & SSE-T8032 Ethernet Switches – 7132 Available today, 8032 in Early 2024

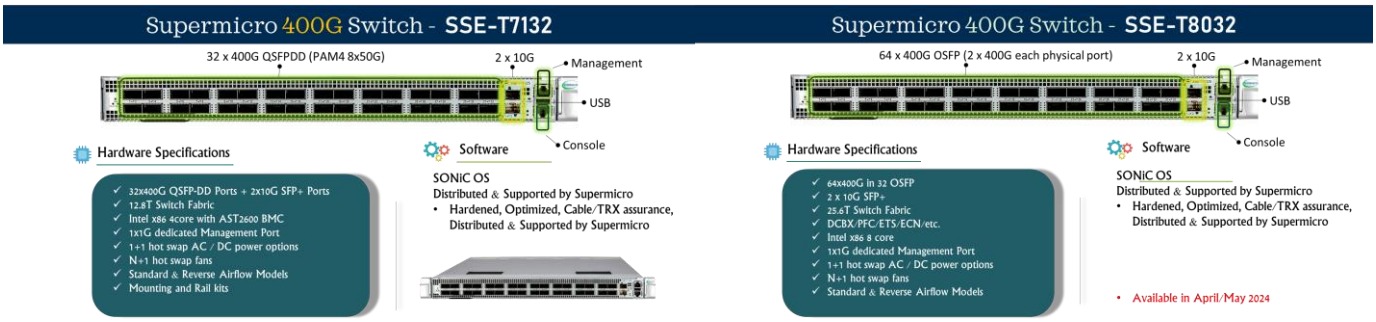

Figure 15 - Supermicro SSE-T7132 and SSE-T8032 Ethernet Switches

Above are the key points of the SMC SSE-T8032 Ethernet switch are used as both the leaf and spine in our AI network cluster designs. In these designs, Supermicro will use direct 800G presented as 2x400GE links between all elements via 400GE breakouts for connection to the NICs on our NVIDIA H100 systems (with H200 systems announced). Supermicro has similar designs for both AMD and Intel GPU offerings. Supermicro has a roadmap for regular updates to these designs. Please contact your Supermicro sales team to learn more. Each NVIDIA H100/H200 is connected to a 400G NIC interface on the fabric.

These designs are based upon a local pod concept, with the leaf devices placed at the center of a 3-rack pod cluster. Within this cluster are eight systems, with each rack having a maximum estimated power of 25-30kW for our H100/H200 in 8U server design. Our upcoming additions allow for H200 GPU in a 4U Direct Liquid Cooling design, which will be covered in other papers. The logic for the placement is to optimize the connectivity for copper connections while primarily keeping the more costly fiber connections to leaf-spine connections. Figure 16 below shows the Pod.

Figure 16 - Supermicro AI Cluster with GPU Servers with NVIDIA HGX H100/H200 GPUs and Switches

The connections between the leaf and the eight systems are eight sets of OSFP 800G in 2x400G to a pair of QSFP112 400G on the NIC. The leaf-spine connections are OSFP 800G each, as shown in Figure 17 below. To scale out the number of systems in the cluster (the example shown is a 128 system cluster for a 1024 H100/H200 GPU cluster), we need only to increase the spine count.

128 GPU SYSTEMS (1024 H100/200 GPU) NETWORK - 400G NIC + **SSE-T8032 64X400G SWITCH (48 TOTAL SWITCHES)**

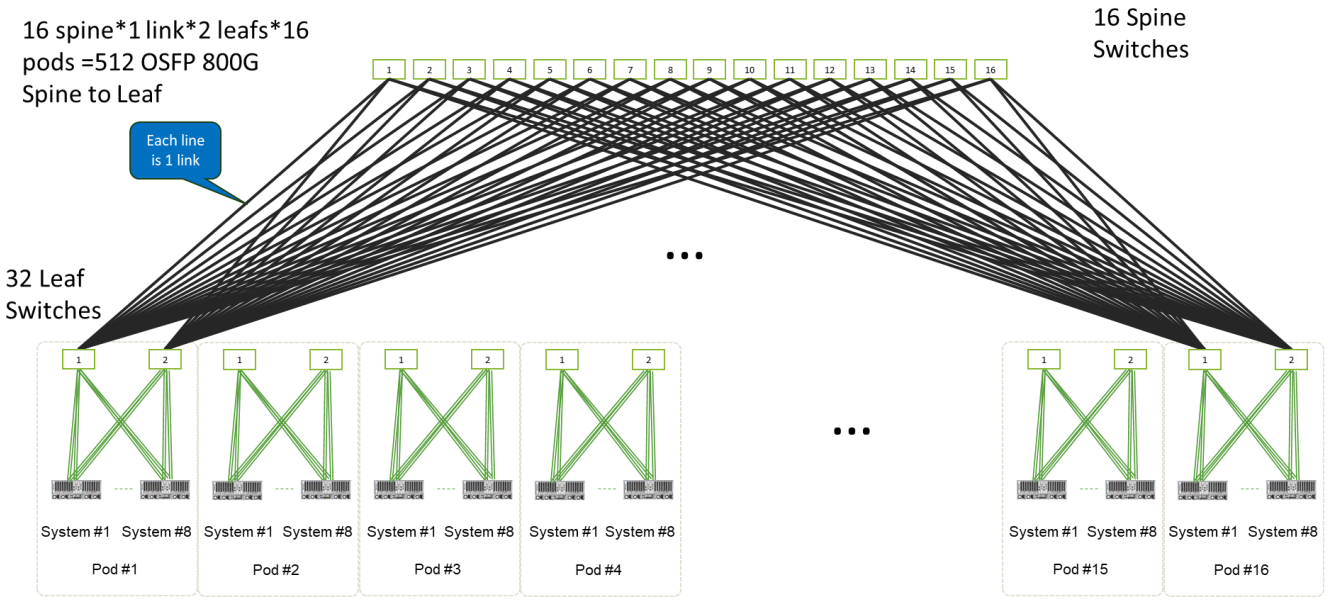

Figure 17 - 128 system 1024 GPU design with 400G to each GPU – similar designs exist for Intel and AMD

There is also a need for a management network and storage network access (where NVMeoF is more commonly used), with a section showing the storage network outlined in Figure 18.

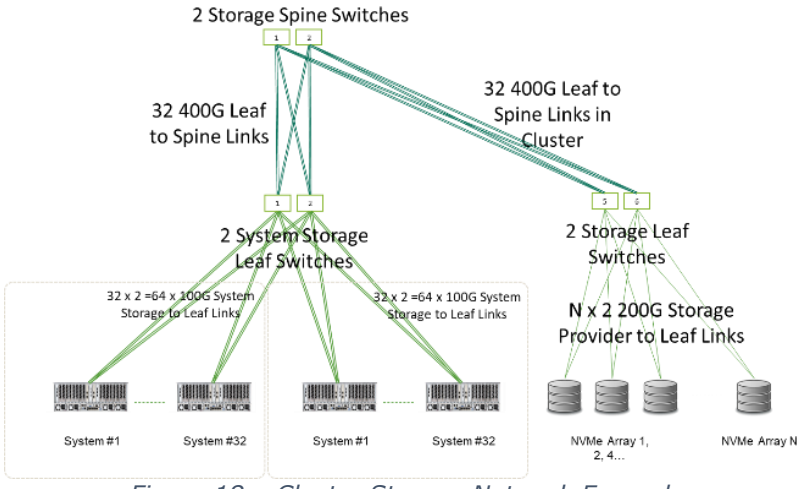

Figure 18 – Cluster Storage Network Example

Supermicro has a toolset to orchestrate the complete cluster, including the servers, networks, and storage via the Supermicro SuperCloud Composer (SCC). Details of the SCC product are available at:
https://www.supermicro.com/en/solutions/management-software/supercloud-composer

While this paper overviews these designs, please contact your Supermicro sales teams for the proper details and more to help with Supermicro AI solutions.

Summary

Many components of an AI cluster network in this paper discuss the possibility of using Ethernet in these designs and the tradeoffs. This paper is written from the lens of customers leaning towards Ethernet already. A summarized table (aside from cost differences) below includes a column on Supermicro switch offerings.

	NDR Infiniband	Ethernet	SMC Product
Speed	Up to 400Gbps	Up to 400/800Gbps with 1600Gbps coming near term	SSE-T7132 (200G NIC/400G Spine), Soon SSE-T8032 (64p 400G NIC/Spine)
AI Training Messaging Customization	Nvidia Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)	Not Optimal to do on network – ideal on GPU systems themselves	N/A
Remote Memory Access	RDMA	RoCEv2	Yes
Congestion Control & Avoidance	Adaptive Routing	ECN, DCQCN, BFD, UDLD, Others	Yes
Self-Healing Networking	Self-Healing Networking	Traffic engineering & Link State Notifications	Yes
Silicon Chip Latency*	~200-300nS	600-1200nS	570nS
Load Balancing	Yes	BGP and ECMP	
Path Segmentation and Crediting	Virtual Lane (VL)	Virtual Output Queues (VoQ)	Yes
QoS	Yes	Yes - Markings in Ethernet frame	Yes
Switching Vendor Count	1	Many	N/A
Scale	48k max nodes	Unlimited	Unlimited
Lossless	Yes	PFC / Buffering	Yes – 70MB Packet Buffer
NIC (Pluggable) on GPU Server	Nvidia	Nvidia, Broadcom	Nvidia, Broadcom

Figure 19 – Summary of InfiniBand and Ethernet as Pertaining to AI networks

References:

1. https://650group.com/blog/infiniband-and-ethernet-switch-markets-thrive-with-ai-ml-support/
2. Scalable Hierarchical Aggegation Protocol (SHArP): A Hardware Architecture for Efficient Data Reduction, by Richard L. Graham, et al. Mellanox 2016 IEEE
3. Super Compute 2016 Comparison of High Performance Network Options, Erickson et al. http://sc16.supercomputing.org/sc-archive/tech_poster/poster_files/post149s2-file3.pdf
4. Running BGP in Data Centers at Scale, Abhashkumar, et al. https://research.facebook.com/file/5208380302511734/Running-BGP-in-Data-Centers-at-Scale_final.pdf

SUPERMICRO

As a global leader in high performance, high efficiency server technology and innovation, we develop and provide end-to-end green computing solutions to the data center, cloud computing, enterprise IT, big data, HPC, and embedded markets. Our Building Block Solutions® approach allows us to provide a broad range of SKUs, and enables us to build and deliver application-optimized solutions based upon your requirements.