WHITE PAPER

SUPERMICRO SSE-T8032S: AN IDEAL SOLUTION FOR MODERN AI NETWORK FABRICS

Reduce Time to Solution Using Supermicro Switches for AI Training

TABLE OF CONTENTS

Executive Summary 1
AI Data Centers 2
GPU Communications in an AI Data Center 2
Network Fabric in AI Data Center 3
Supermicro SSE-T8032S Switch 3
Supermicro Enterprise SONiC 3
Summary 7
More Information 7
Appendix 8

Executive Summary

The Artificial intelligence (AI) revolution is here. The digital revolution over the last few decades has taken the next giant leap in the form of Artificial intelligence and Machine Learning (ML). AI and ML have immense potential to disrupt and transform entire industries. Supermicro is already seeing several breakthrough innovations emerge in healthcare, banking, automotive, retail, etc., to name a few sectors. Advancements in large language models (LLM) and training and inferencing technologies are breaking barriers and enabling transformative innovations. Supermicro is a leader in the AI infrastructure space and has been constantly rolling out advanced solutions that have enabled these innovations. This paper will cover the infrastructure challenges posed by AI workloads on the network fabric and how Supermicro's 400G 25.6Tbps high-speed Ethernet switch SSE-T8032S running Supermicro Enterprise SONiC is uniquely equipped to tackle them. The results from the performance and benchmarking tests run on a cluster of AMD instinct MI300X nodes interconnected by an SSE-T8032S switch fabric and Broadcom 400G Network interface cards (NIC) discussed in this paper further validate that the SSE-T8032S switch running Supermicro Enterprise SONiC is a great choice for modern high-performance AI network fabric.

AI Datacenters

AI workloads have specific, unique requirements. The unprecedented growth in LLMs in terms of the number of parameters and size of the data sets demands a new type of infrastructure with specialized hardware like AI accelerators and GPUs. Large clusters of GPUs, in many cases, tens of thousands of them, are interconnected to form a unified high-performance computing environment to perform massively parallel computations needed to train these massive LLMs. The AI workloads running on such large, distributed computing environments must constantly communicate and synchronize to complete the job in real time. One of the key metrics used to measure the executing efficiency of the AI infrastructure is the job completion time (JCT), which effectively is the overall time taken between the job assignment and collective completion. The network fabric connecting these distributed GPUs must be best designed to optimize the job completion time of the training jobs to improve overall efficiency and resource utilization of the AI infrastructure.

GPU Communications in an AI Data Center

AI workloads are massive in scale, and techniques such as data parallelism and model parallelism enable them to be broken into mini-batches and run across multiple GPU nodes. The GPU nodes constantly communicate using the MPI collective communications library (CCL) primitives to synchronize and exchange data, which demands an efficient network fabric in the AI infrastructure. All-reduce and all-gather are some of the most commonly used CCL primitives in AI networks. All-reduce is used to aggregate and then redistribute data from cluster nodes as they coordinate and work on completing the job. For each iteration, the nodes perform a reduce operation and update all the nodes with the computed result. These data-intensive communication operations can cause all nodes to transmit simultaneously, causing network in-cast scenarios and leading to increased congestion in the network.

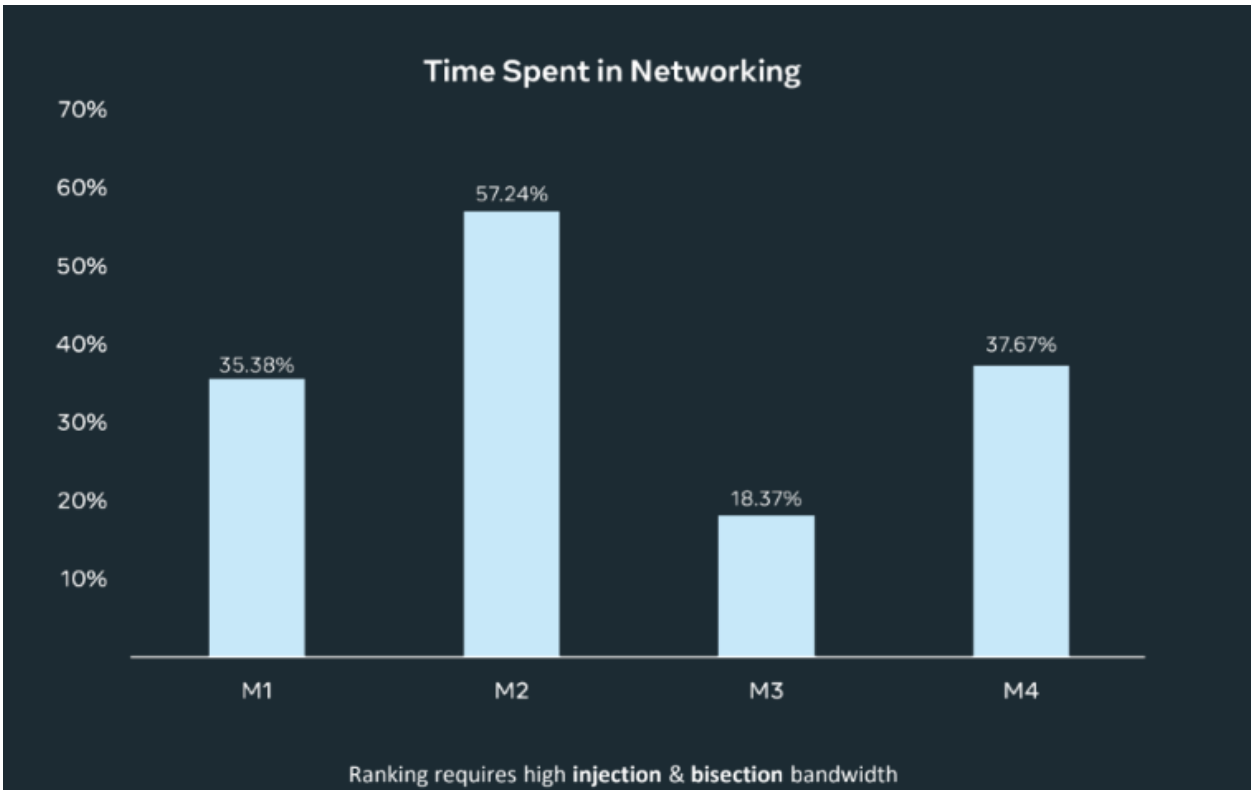


Figure 1 – Network performance in large AI Clusters (Source: OCP2022 Keynote by Meta)

Unlike traditional data centers, the workloads in the AI infrastructure are very different. The AI workloads constantly exchange messages between the GPU nodes in the cluster to collectively complete the job. These message exchanges generate increased bursts of significant inter-GPU traffic. These messages initiate and terminate within the cluster between the same set of GPUs, causing the flow entropy to be very low. This low entropy bursty traffic is more prone to congestion and sub-optimal data forwarding that can cause link saturation and packet loss. Congestion or loss of packets in the network fabric can slow compute nodes, increasing job completion time and possibly blocking job completion altogether. For instance, Figure 1 above shows the overall time taken by the network in training four different AI models (M1, M2, M3, M4) within Meta's AI data center. It shows that in large clusters, as much as 40-50% of the total time is consumed by the network transport that connects the GPU nodes, of which about 33% of the time is wasted waiting for the network. This clearly makes network fabric the most critical piece in the AI infrastructure puzzle.

Network Fabric in an AI Data Center

As the compute clusters scale, the AI data center's performance greatly depends on a highly robust network that can enable seamless communication within the cluster. If the network fabric in the AI data center is not architected right, it can lead to performance bottlenecks, causing sub-optimal functioning of the expensive GPU resources. As more GPUs are inter-connected in larger clusters, the increased message exchange and synchronization between them need a higher aggregate cluster bandwidth for all the GPUs to communicate effectively. This mandates the network fabric connecting the compute clusters be of high bandwidth and non-blocking. The distributed nature of the AI workload execution translates to multiple iterations of computation, updating, and synchronization of the mini-batches to successfully complete the job. In such a scenario, even a single slow node could end up causing the entire cluster to slow down, making it very sensitive to packet loss or delays in the network. So, it is not just the individual packet latency but the tail latency of the network that also needs to be low for all the nodes to successfully exchange messages and proceed to the next iteration without having to wait unnecessarily on other GPU nodes. All these requirements warrant a need for a high bandwidth, non-blocking network fabric with low tail latency and efficient congestion control mechanism to meet the stringent Quality of Service (QoS) demands posed by the distributed computing challenges in the AI infrastructure. Ethernet already has all these capabilities needed in an AI network fabric built into it. RDMA over Converged Ethernet (RoCE) technology enables RDMA high throughput and low latency message transfer capability on Ethernet. A standard CLOS network fabric running high-speed ethernet with RoCEv2 is thus an ideal solution for an AI network of any scale.

Supermicro SSE-T8032S Switch

Supermicro SSE-T8032S switch platform offers 64 400G (2*400G with 32 Physical OSFP) ports in a 1 RU form factor. The forwarding engine of the switch is built with a Broadcom Tomahawk 4 ASIC that can support 25.6Tbs of switching capacity. Broadcom Tomahawk 4 provides several advanced features like better in-cast absorption, lower end-to-end latency, advanced load balancing capabilities, and efficient congestion management mechanisms. The Supermicro Enterprise SONiC has a comprehensive set of all the features that the AI networks need. The single chip architecture of SSE-T8032S powered by Tomahawk 4 running Supermicro Enterprise SONiC makes it an excellent switch platform with the correct port density in a 1 RU form factor to build a non-blocking low latency CLOS network fabric to interconnect the current generation GPUs in the AI data centers.

Supermicro Enterprise SONiC

The SSE-T8032S switches come with the Supermicro Enterprise SONiC network operating system. Software for Open Networking in the Cloud (SONiC) is an open-source Debian Linux based network operating system that has seen increased

adoption in the industry across major cloud providers and enterprises. Supermicro SONiC is hardened with more features and usability features to deploy and manage at scale. The Supermicro Enterprise SONiC has all the needed software support to realize all the advanced hardware capabilities of the Tomahawk 4 switch, such as RoCE and congestion control technologies. Let's look at RoCE and the congestion control aspects that make SSE-T8032S running Supermicro Enterprise SONiC suitable for modern AI network fabrics.

RDMA over Converged Ethernet (RoCE):

Remote Direct Memory Access (RDMA) is a network data transfer technology employed in large, distributed computing that allows applications to read from and write into the host adapters directly. RDMA bypasses the operating system kernel and CPU, immensely bringing down data transfer latency. Distributed AI applications with a lot of synchronous messaging benefit greatly from RDMA technology and, hence are a widely deployed protocol in AI infrastructure for the compute nodes to communicate between them. With RoCE, all the performance benefits and optimization of the RDMA technology are now available in ethernet transport, making Ethernet even more suitable for a scale-out AI network fabric. Large AI network fabrics are prone to congestion, and unfortunately, RDMA does not have congestion control built into it natively and relies on the transport protocol. Ethernet does have a suite of congestion management technologies that can be deployed alongside RoCE for congestion control.

Priority Flow Control (PFC):

Priority flow control is a link level flow control mechanism used to pause the transmission of the network's flows, causing congestion. Traffic is classified into queues based on priority and assigned a watermark. When the watermark is breached, a PFC message is generated towards the upstream device, signaling to pause the flow transmission until the receipt of a subsequent message to resume. The upstream device resumes transmission when the PFC is de-asserted. PFC thus mitigates congestion by pausing the congesting queues instead of dropping packets.

Explicit Congestion Notification (ECN)

ECN manages congestion by proactively signaling the transmitting host of the potential congestion in the network without pausing the traffic. Unlike PFC, in the case of ECN, the congestion signaling is propagated all the way to the end host and not just to the upstream device. During congestion, as the egress queue starts building up in an ECN capable device, it starts to probabilistically mark the Congestion Experienced (CE) bits in the DSCP packets in the IP header of the data packets to notify the destination host of the potential congestion. As congestion continues to build up, more packets are marked until all the packets towards the destination are marked to signal congestion. On seeing the CE bit set in the data frames, the destination host would then send a Congestion Notification Packet (CNP) to notify the sender of the flows about the congestion. The sender would react to CNP by reducing the data transmission rate to clear the congestion. Unlike PFC, ECN handles congestion by slowing down the frame transmission and shaping the flow instead of pausing transmission.

Data Center Quantized Congestion Notification (DCQCN):

DCQCN is basically a means of managing congestion by deploying both ECN and PFC simultaneously to get the best of both. In DCQCN, as congestion in the network builds up, ECN is configured in a way to get triggered first and start marking the packet to signal the host to slow down the transmission. If that does not happen on time and the congestion continues to build up, PFC kicks in to pause the offending flows. The goal essentially is to use ECN as your first line of defense to tackle congestion without having to stop data transmission as best as possible before resorting to PFC as a backup mechanism. The effectiveness of

DCQCN depends on having the right PFC and ECN parameters such that PFC does not prematurely start pausing packets before ECN attempts to throttle down the transmission rate to handle the congestion.

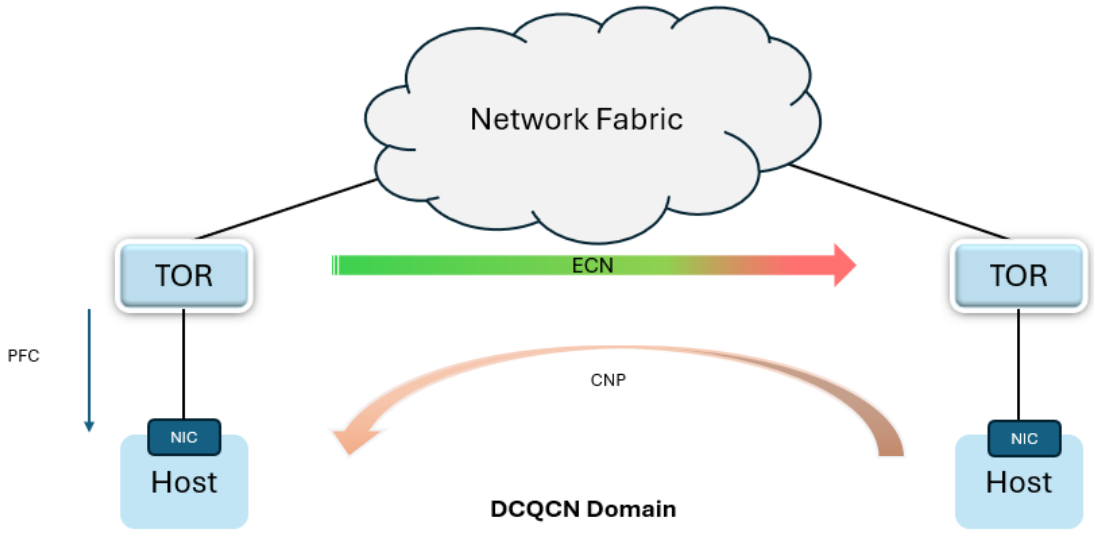


Figure 2 - Data Center Quantized Congestion Notification

SSE-T8032S Fabric Test

Supermicro SSE-T8032S interoperates with all the leading current generation GPU and NIC products to create high-performance network fabric in the AI data centers. We will now focus on one of the recent benchmarking tests performed on an AI cluster built using AMD GPUs and Broadcom GPUs NICs. A small-scale AI cluster was emulated using Supermicro SSE-T8032S switches to connect Supermicro AS-8125GS-TNMR2 GPU Systems using Broadcom's latest 400G Network interface cards (NIC) to evaluate GPU cluster performance. Supermicro AS-8125GS-TNMR2 is a high-performance GPU accelerator platform that integrates eight fully connected AMD Instinct MI300X GPU modules using an AMD infinity fabric. The topology below shows how a simple two-tier network fabric was built using SSE-T8032S to interconnect these 32 GPUs into a small cluster. RDMA was provisioned on the Broadcom 400G NICs to enable high throughput and low latency message transmission between the GPU nodes. DCQCN congestion control feature was configured both on the SSE-T8032S nodes and Broadcom NICs, interconnecting the 32-node AMD Instinct MI300X cluster to enable a loss-less fabric for the RDMA endpoints to communicate efficiently using RoCEv2.

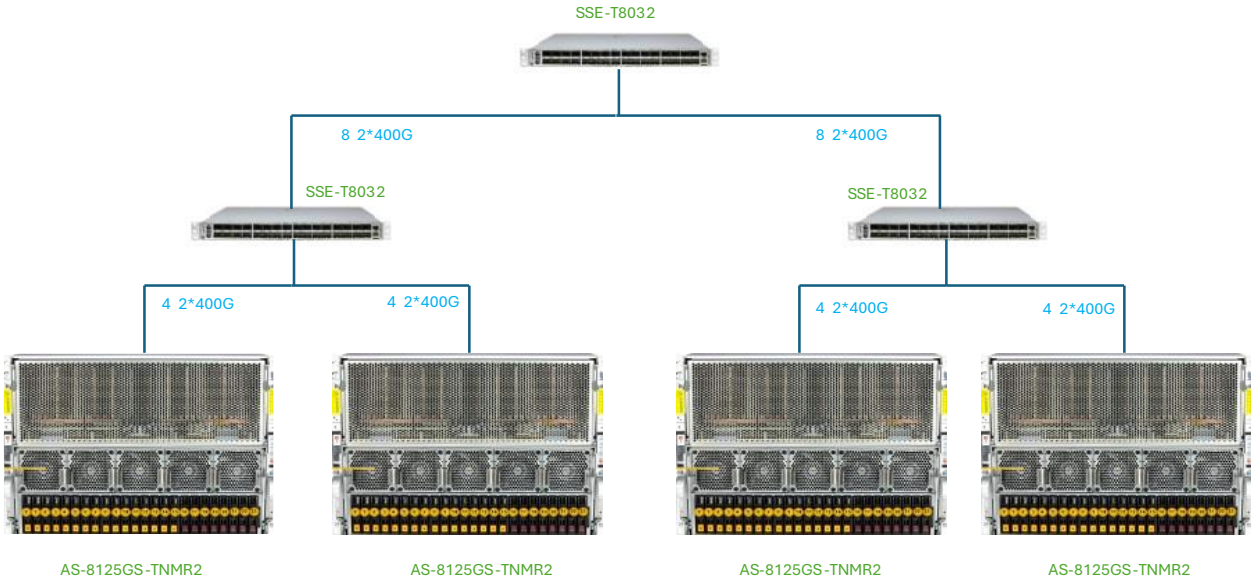
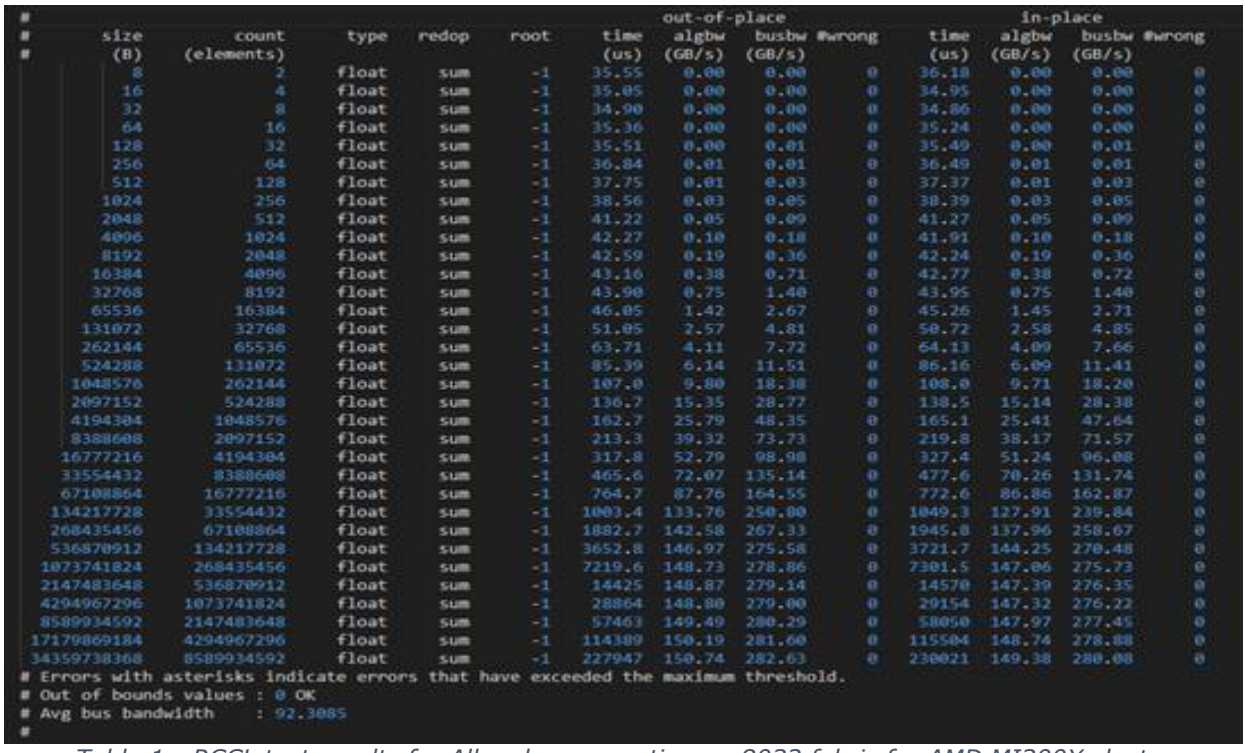


Figure 3 - Topology for RCCL test on 8032 fabric for AMD MI300X cluster

ROCm communication collective library (RCCL) benchmarking tests were run on this cluster to emulate various CCL communication, such as all-reduce, all-to-all, scatter, etc., to measure the cluster's performance. The standard RCCL tests emulate the CCL operations supported on the ROCm framework by scheduling jobs of different sizes and measuring the GPU performance. The tests report three metrics – operation time, algorithm bandwidth, and bus bandwidth. While operation time and algorithm bandwidth are an excellent criterion for point-to-point operations like send, receive, etc., but are not an accurate measure of performance in case of collective operations like all-reduce, all-to-all, etc., which involve multiple GPU nodes talking to the same GPU at the same time. Algorithm bandwidth decreases as the number of ranks in the cluster increases; hence, bus bandwidth is used in operations like all-reduce. Bus bandwidth is more reflective of the actual bandwidth. It is computed by applying an additional calculation to the algorithm bandwidth to factor in the number of ranks and the collective operation type. The bus bandwidth equates to 2(n-1)/n * algorithm bandwidth for an all-reduce operation. Table 1 below captures the RCCL test result run for an all-reduce operation in the lab setting. For the 32-node AMD MI300X GPU cluster connected by a two-tier network fabric with SSE-T8032S, the bus bandwidth was measured at 280GBps. Please see the appendix for details on the hardware and software components used to construct this test scenario.



					out-of-place				in-place			
size (B)	count (elements)	type	redop	root	time (us)	algbw (GB/s)	busbw (GB/s)	#wrong	time (us)	algbw (GB/s)	busbw (GB/s)	#wrong
8	2	float	sum	-1	35.55	0.00	0.00	0	36.18	0.00	0.00	0
16	4	float	sum	-1	35.05	0.00	0.00	0	34.95	0.00	0.00	0
32	8	float	sum	-1	34.90	0.00	0.00	0	34.86	0.00	0.00	0
64	16	float	sum	-1	35.36	0.00	0.00	0	35.24	0.00	0.00	0
128	32	float	sum	-1	35.51	0.00	0.01	0	35.49	0.00	0.01	0
256	64	float	sum	-1	36.84	0.01	0.01	0	36.49	0.01	0.01	0
512	128	float	sum	-1	37.75	0.01	0.03	0	37.37	0.01	0.03	0
1024	256	float	sum	-1	38.56	0.03	0.05	0	38.39	0.03	0.05	0
2048	512	float	sum	-1	41.22	0.05	0.09	0	41.27	0.05	0.09	0
4096	1024	float	sum	-1	42.27	0.10	0.18	0	41.91	0.10	0.18	0
8192	2048	float	sum	-1	42.59	0.19	0.36	0	42.24	0.19	0.36	0
16384	4096	float	sum	-1	43.16	0.38	0.71	0	42.77	0.38	0.72	0
32768	8192	float	sum	-1	43.90	0.75	1.40	0	43.95	0.75	1.40	0
65536	16384	float	sum	-1	46.05	1.42	2.67	0	45.26	1.45	2.71	0
131072	32768	float	sum	-1	51.05	2.57	4.81	0	50.72	2.58	4.85	0
262144	65536	float	sum	-1	63.71	4.11	7.72	0	64.13	4.09	7.66	0
524288	131072	float	sum	-1	85.39	6.14	11.51	0	86.16	6.09	11.41	0
1048576	262144	float	sum	-1	107.0	9.80	18.38	0	108.0	9.71	18.20	0
2097152	524288	float	sum	-1	136.7	15.35	28.77	0	138.5	15.14	28.38	0
4194304	1048576	float	sum	-1	162.7	25.79	48.35	0	165.1	25.41	47.64	0
8388608	2097152	float	sum	-1	213.3	39.32	73.73	0	219.8	38.17	71.57	0
16777216	4194304	float	sum	-1	317.8	52.79	98.98	0	327.4	51.24	96.08	0
33554432	8388608	float	sum	-1	465.6	72.07	135.14	0	477.6	70.26	131.74	0
67108864	16777216	float	sum	-1	764.7	87.76	164.55	0	772.6	86.86	162.87	0
134217728	33554432	float	sum	-1	1003.4	133.76	250.80	0	1049.3	127.91	239.84	0
268435456	67108864	float	sum	-1	1882.7	142.58	267.33	0	1945.8	137.96	258.67	0
536870912	134217728	float	sum	-1	3652.8	146.97	275.58	0	3721.7	144.25	270.48	0
1073741824	268435456	float	sum	-1	7219.6	148.73	278.86	0	7301.5	147.06	275.73	0
2147483648	536870912	float	sum	-1	14425	148.87	279.14	0	14570	147.39	276.35	0
4294967296	1073741824	float	sum	-1	28864	148.80	279.00	0	29154	147.32	276.22	0
8589934592	2147483648	float	sum	-1	57463	149.49	280.29	0	58050	147.97	277.45	0
17179869184	4294967296	float	sum	-1	114389	150.19	281.60	0	115504	148.74	278.88	0
34359738368	8589934592	float	sum	-1	227947	150.74	282.63	0	230021	149.38	280.08	0

Errors with asterisks indicate errors that have exceeded the maximum threshold.
Out of bounds values : 0 OK
Avg bus bandwidth : 92.3085
#

Table 1 - RCCL test results for All-reduce operation on 8032 fabric for AMD MI300X cluster

Summary

AI is a transformative technology that is continuously evolving at a rapid pace. AI workloads need a new class of scalable, robust, and efficient data center infrastructure to meet the compute and power demands of these large-scale distributed systems. A robust AI infrastructure calls for a very robust and efficient network fabric to interconnect the massively distributed GPU resources, reduce job completion time, and drive peak performance of the overall system. This makes network fabric the most critical component in the AI data center. The right kind of network can enable the utilization of expensive GPUs in the infrastructure and significantly drive down the overall spending on AI infrastructure. Supermicro SSE-T8032S, Tomahawk 4 based 64 port 400G switch running Supermicro Enterprise SONiC, is an excellent switch with all the features needed to build a large scale, high throughput, low latency network with efficient congestion management techniques to meet the needs of the current generation AI data center. The SSE-T8032S switches also come as part of the Supermicro rack scale solutions - a fully validated solution for the AI infrastructure with extremely low lead times. Using Supermicro switches in the infrastructure helps in having a more tightly integrated overall offering. Pre-validated designs and the cluster orchestration tools from Supermicro radically accelerate day 0 deployments, aiding in faster infrastructure build-out.

For More Information

Supermicro | Products | Networking | SSE-T8032S

http://benchmark.supermicro.com/report/HPC/amd-mi300x-cluster-rccl-tests

GitHub - ROCm/rccl-tests: RCCL Performance Benchmark Tests

Appendix – Component Details

1. Cluster Hardware Components

Item Number	Description	Quantity
AS -8125GS-TNMR2	H13DSG-OM, 8 x GPU-FRU-MI300X-OAM	4
SSE-T8032S	32 x 2x400G OSFP Ports, 1RU, Front airflow, AC PSUs (2 leaf, 1 spine)	3
CBL-NTWK-0976-15M-J	[NR] 800G OSFP to 2x400G QSFP112 Y-cable, 26 AWG,1.5M, JPC	32
CBL-NTWK-1107-15M-J	OSFP Cable 800G 28AWG -1.5M	16

2. Server Components

Hardware	Description	Quantity
CPU	AMD EPYC 9654 96-Core Processor, @2.40GHz 96Cores/192Threads	2
GPU	AMD Instinct MI300X 192GB 750W	8
Memory	Micron MTC40F204WS1RC48BB1 MHFF 4800 MT/s 96 GB	24
Drive	Samsung MZQL23T8HCLS-00A07 3.8TB	4
Network	Broadcom 57608 PCIe Gen 5.0 x16 low-profile standard form factor single QSFP-DD connector	8
PSU	PWS-3K06G-2R	6
PSU	PWS-DF009-2F	2
Fans	N/A	10

3. Software Components

Item	Version Information
Linux OS	Ubuntu 22.04.4 LTS
Kernel	5.15.0-102-generic
ROCm Software	6.1.0
BMC	01.02.32
Redfish	1.11.0
CPLD	F2.65.17
BIOS	1.1
BKC	24.06.10
OpenMPI	Open MPI v5.0.4a1
UCX	v5.0.4
Broadcom Driver	netxtreme-bnxt_en-1.10.3-229.2.43.0

4. Switch SSE-T8032S Specifications

Item	Detailed Information
Ports	64x400G in 32 OSFP + 2 x 10G SFP+
Switch Capacity	25.6Tbps aggregated switching capacity
Packet Buffer	112MB memory packet buffer
OS	SONiC-OS-4.2.1-Enterprise_Advanced
CPU	INTEL X86 Xeon 8-core CPU

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.