



# SUPERMICRO SUPERBLADE® POWERED BY AMD EPYC™ PROCESSORS EXCEL AT SCALING DISTRIBUTED AI AND ML TRAINING



## TABLE OF CONTENTS

Executive Summary .....	1
System Configurations .....	2
AI/ML Workloads on the SuperBlade .....	4
SuperBlade AI/ML performance .....	8
Use Cases .....	11
Conclusion .....	11

## SUPERMICRO

Supermicro (Nasdaq: SMCI) is the leading innovator in high-performance, high-efficiency server and storage technologies and a premier worldwide provider of advanced server Building Block Solutions® for Enterprise Data Center, Cloud Computing, Artificial Intelligence, and Edge Computing Systems. Supermicro is committed to protecting the environment through its “We Keep IT Green®” initiative by providing customers with the most energy-efficient, environmentally friendly solutions available on the market.

accelerated SBA-4119SG blade provides end-to-end AI/ML/DL solutions from the edge to the data center. This system integrates one 200G HDR InfiniBand switch and two 25G Ethernet switches in a highly scalable HPC environment that processes distributed ML workloads without losing flexibility or performance. This “cloud-in-a-box” configuration is also ideal for AI platforms where the experiment and deployment of ML models can be easily managed and scaled. A Supermicro SuperBlade achieved one of the best overall [MLPerf Inference v1.0 performance scores](#).

## Executive Summary

Extremely large AI and ML workloads can push the boundaries and capabilities of a multi-server platform. Even the most advanced CPUs and GPUs require coordination among multiple independent servers. Combining fast CPUs and GPUs with a fast and efficient networking architecture is critically important for these massive training workloads. A well-designed system scales when needed and is also available for smaller workloads requiring only a single system. This white paper describes how the Supermicro® SuperBlade® powered by 3<sup>rd</sup> Gen AMD EPYC™ processors excels at scaling distributed AI and ML training.

## Supermicro 8U SuperBlade Overview

The Supermicro 8U SuperBlade system hosts up to 20 individual SuperBlade servers in a single 8U enclosure. The GPU-

Horovod is an open-source framework for scaling deep learning training across hundreds of GPUs in parallel. It is a distributed, scalable deep learning training framework based on the ring allreduce algorithm that leverages High Performance Computing (HPC) techniques, such as MPI, Data Parallel, etc., to scale across multiple devices and nodes in both on-premise and cloud deployments efficiently. In addition, it enables running GPU-enabled AI/ML frameworks such as TensorFlow, Keras, PyTorch, and Apache MXNet. This paper describes the testing performed running image classification on eight GPU-enabled Supermicro 8U SuperBlade servers using the ResNet50 benchmark, demonstrating high throughput with distributed workloads across multiple nodes.

## System Configurations

The Supermicro 8U SuperBlade Enclosure (SBE-820H-822) supports both CPU-only and GPU-enabled blades. The tests described in this whitepaper used the following components:

Components	Part Description	QTY
SBE-820H-822	8U SuperBlade Enclosure with 8x2200W PSUs	1
Management	SuperBlade Chassis Management Module (CMM)	1
Networking	25GbE Switches	2
InfiniBand Switch	40-port 200G HDR InfiniBand Switch	1

Table 1 – SuperBlade chassis components

Supermicro SuperBlade® systems are available in both CPU-only (Table 2) and GPU-enabled (Table 3) configurations.


	Components	Part Description	QTY
	SBA-4114S-T2N	Supermicro SuperBlade (AMD-powered single-socket)	10
	CPU	AMD EPYC™ 7713 64-Core processor	10
	Memory	32GB DDR4-3200 2Rx4 ECC REG DIMM	80
	Storage (M.2)	2TB M.2 NVMe SSD	20
	Storage (U.2)	2TB U.2 NVMe SSD	20

Table 2 – CPU-only blade components


	Components	Part Description	QTY
	SBA-4119SG	Supermicro SuperBlade (AMD-powered single-socket)	8
	CPU	AMD EPYC™ 7713 64-core processor	8
	Memory	32GB DDR4-3200 2Rx4 ECC REG DIMM	64
	Storage	2TB M.2 NVMe SSD	8
	GPU	NVIDIA® A100™ PCIe 40GB	8

Table 3 – GPU-enabled blade components.

## Networking

The density-optimized 8U SuperBlade® supports advanced networking options to include 10G, 25G, 100G EDR and 200G HDR InfiniBand switches. In addition, each Supermicro SuperBlade chassis includes networking switches for both internal and external connections, as shown in Figure 1.

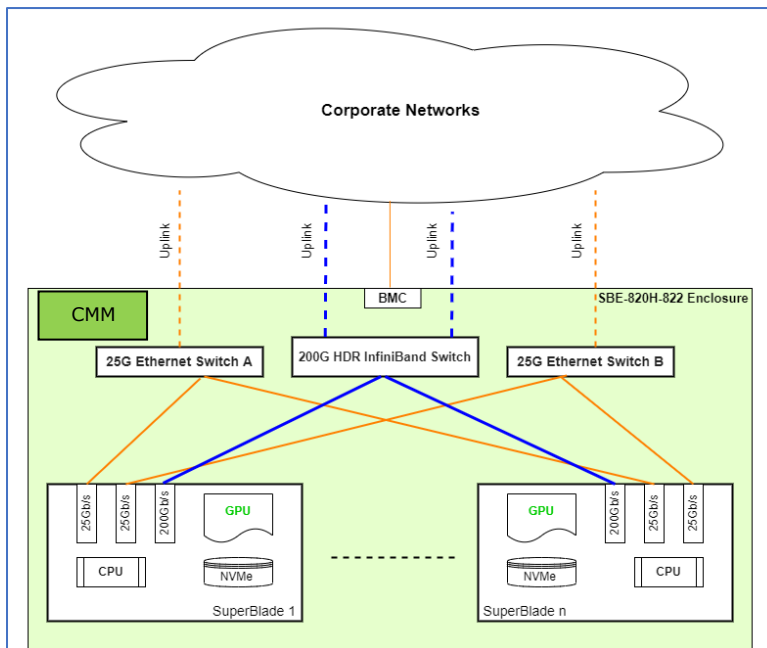


Figure 1 - Networking component topology

Figure 1 illustrates the following components:

- **200G HDR InfiniBand Switch:** Mainly used for high bandwidth HPC communications between the blades. This switch can also boost performance by fully leveraging Remote Direct Memory Access (RDMA) to bypass the CPU. It also supports NVIDIA GPUDirect using the NVIDIA Ampere architecture, provided that the OFED driver is installed and the Subnet Manager is running on one of the blades.
- **2x 25G Ethernet Switches:** Primarily used for data traffic or NIC bonding for each blade for redundancy and higher bandwidth. You can also configure the uplinks on the switch modules with different VLANs to reduce routing latency in large data center environments.
- **CMM Module:** Supermicro Chassis Management Module (CMM) is primarily used to access the chassis for initial configuration, monitoring components, and firmware upgrades. Once each blade is configured, you can directly access that blade's BMC via the CMM from any routable subnets.

## NVIDIA® Certifications

The AMD-based GPU blade SBA-4119SG is an NVIDIA-Certified Systems™ (NCS) 2.3, as shown in Table 4. These certifications demonstrate the flexibility of the Supermicro SuperBlade for AI/ML applications in either the data center or at the edge. <https://www.nvidia.com/en-us/data-center/data-center-gpus/qualified-system-catalog/>

System Category	GPU Type	Certifications
Datcenter (compute only)	NVIDIA A100, A30	Single node, Multinode, GDS certified with NVIDIA Driver 460.73.01 Datcenter
Datcenter (compute and graphics)	NVIDIA A40, A10	

Table 4 – NCS 2.3 certifications

## AI/ML Workloads on the SuperBlade

The Message Passing Interface (MPI) is the primary tool for making multiple ML sessions run in parallel by scaling a single-GPU training script across multiple GPUs in a single- or multi-node configuration. The Horovod MPI wrapper provides a straightforward interface for distributing ML workloads across the HPC environment with performance on par with both bare metal systems and Docker containers, as described [here](#). Supermicro performed the benchmark tests described in this whitepaper using the Horovod Docker configuration shown in Table 5.

Configuration Items	Descriptions
System SKU	Supermicro SuperBlade SBA-4119SG (8 nodes)
# of nodes	8
Chassis	8U
System BIOS	SMBIOS 3.3.0
CPU	3 <sup>rd</sup> Gen AMD EPYC™ 7003 Series Processor (7713 64-Cores, 2.0GHz, 225W TDP)
GPU	NVIDIA® A100™ PCIe 40GB VBIOS 94.02.5C.00.04
CUDA	11.4
OS	Ubuntu 20.04.2
Docker engine	20.10.8
NVIDIA driver	460.73.01
NVIDIA docker	2.6.0
Horovod Docker	v0.23.0
Dataset	Synthetic
Benchmark type	Training

*Table 5 – Benchmark configuration*

Figure 2 provides a hardware and software stack graphical depiction of the information listed in Table 5. It consists of the Horovod Stack at the top, followed by SuperBlade hardware configuration at the bottom.

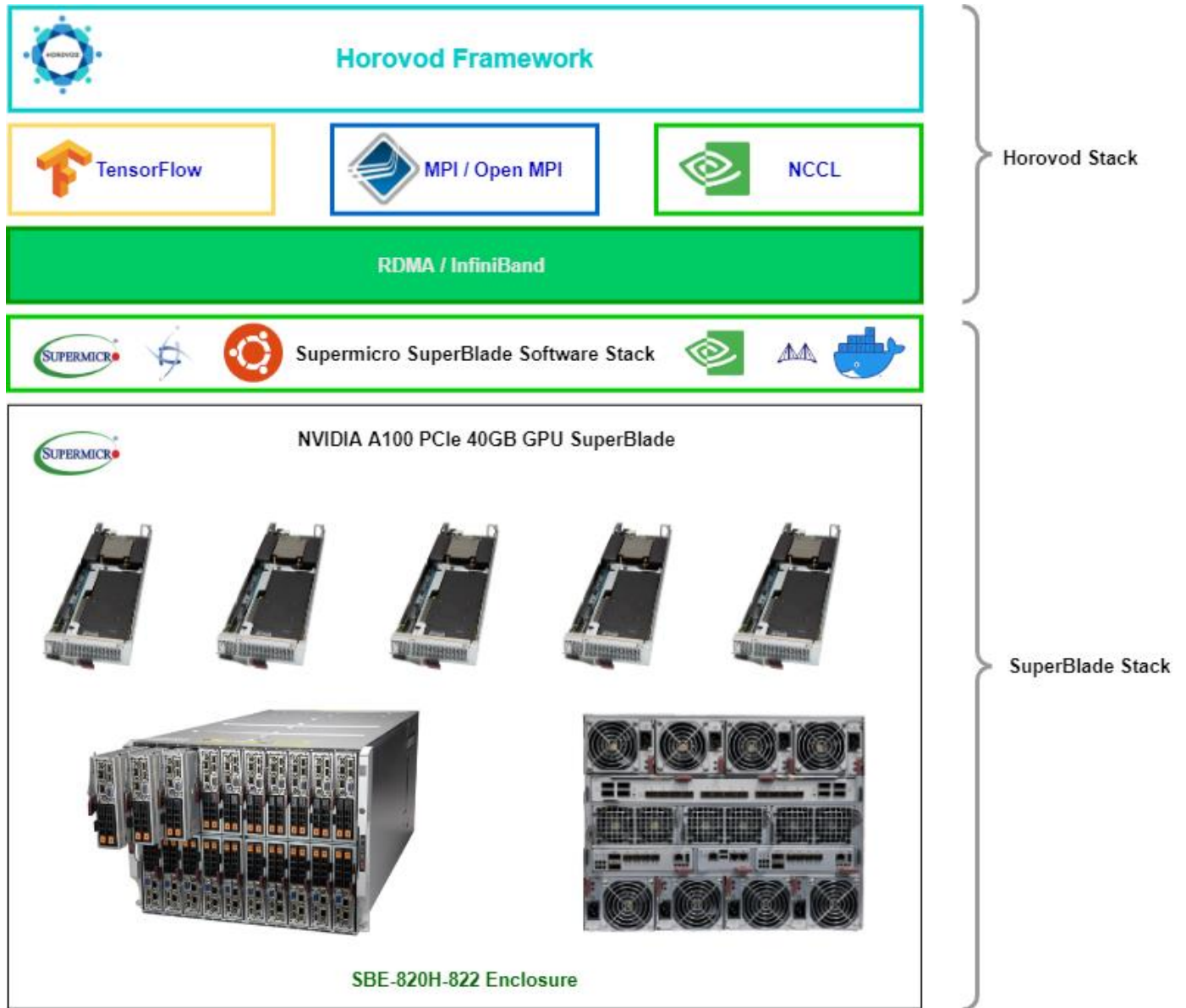


Figure 2 – Hardware and software stack

## Running Horovod with Docker

A prebuilt Horovod Docker image is available from the [Docker hub](#). The Docker file can be customized to fit any specific test environment. Figure 3 summarizes the steps to configure Horovod for the tests described in this whitepaper.

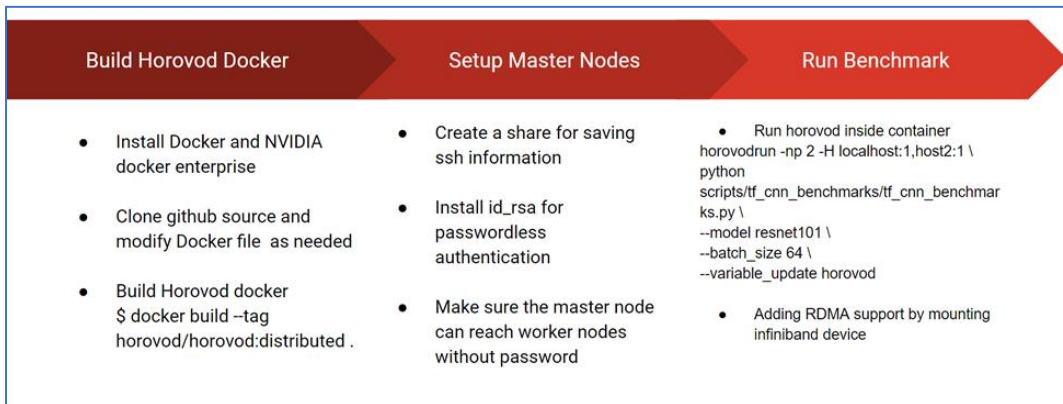


Figure 3 – Horovod configuration summary

Here are the steps described below to run Horovod on the eight GPU enabled Supermicro SuperBlade servers:

1. Initiate the Horovod container.

```
docker run -it --network=host --cap-add=IPC_LOCK --device=/dev/infiniband  
- gpus all \  
horovod /bin/bash
```

2. Run the benchmark on a single node inside the Horovod container.

```
horovodrun -np 1 -H localhost:1 \  
python scripts/tf_cnn_benchmarks/tf_cnn_benchmarks.py \  
--model resnet101 \  
--batch_size 64 \  
--variable_update horovod
```

3. Start the benchmark on two nodes inside the Horovod container.

```
horovodrun -np 2 -H localhost:1,host2:1 \  
python scripts/tf_cnn_benchmarks/tf_cnn_benchmarks.py \  
--model resnet101 \  
--batch_size 64 \  
--variable_update horovod
```

4. Add hosts as described in steps 1 to 3.

## SuperBlade AI/ML Benchmark Performance

The Multi-node benchmark with TCP tests was performed with Horovod distributing multiple AI/ML workloads with a batch size of 256 FP32 across all eight nodes using image database samples from GoogleNet, ResNet50, ResNet101, and Inception3. Figure 4 uses several popular benchmarks to show how the number of images trained per second scales as blades are added.

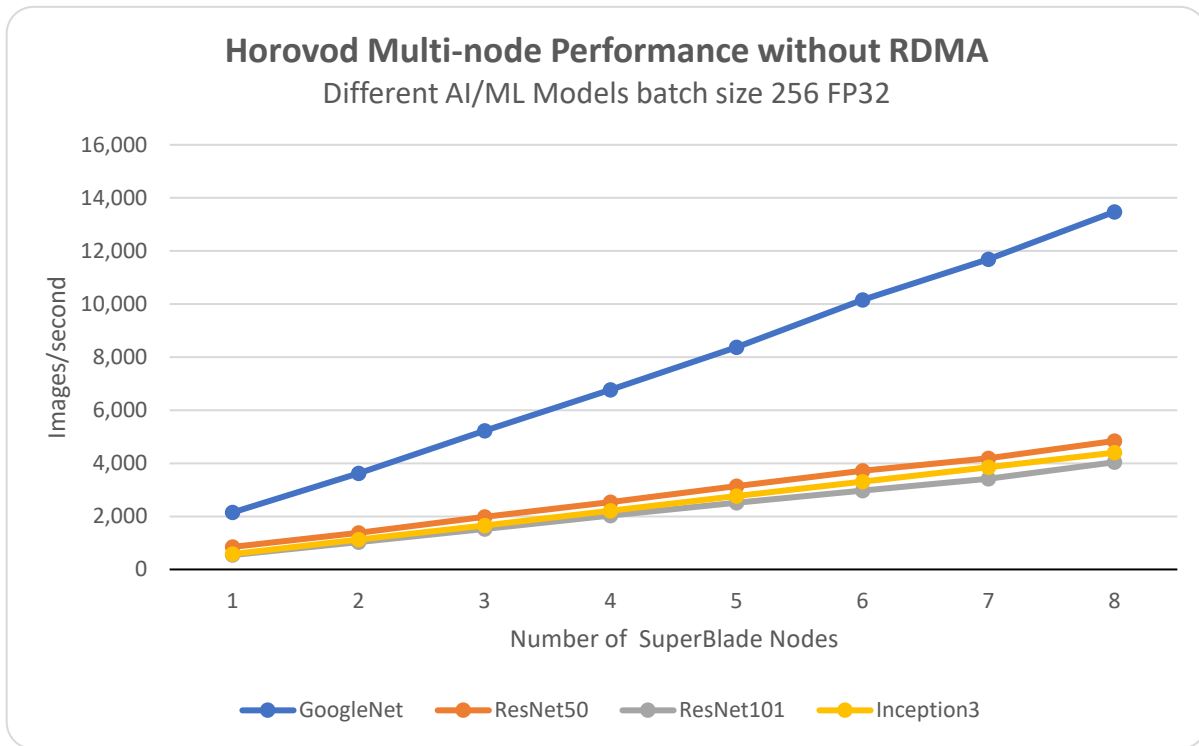


Figure 4 – Adding nodes boosts benchmark performance. (Source: internal Supermicro testing)

From the above testing, it is evident that adding SuperBlade nodes delivered scale-out performance gains. For example, Figure 4 shows that 2 SuperBlade nodes process 3,622 GoogleNet images/second, and the number scales up to 13,475 GoogleNet images/second on 8 SuperBlade nodes.



## Optimized Multi-node SuperBlade Benchmark Performance with RDMA

The Horovod Docker image also includes a built-in InfiniBand driver that supports the Mellanox Quantum RDMA features. Enabling RDMA in the Horovod Docker container boosts AI/ML training performance by allowing the data path to bypass the CPU and directly access system memory.

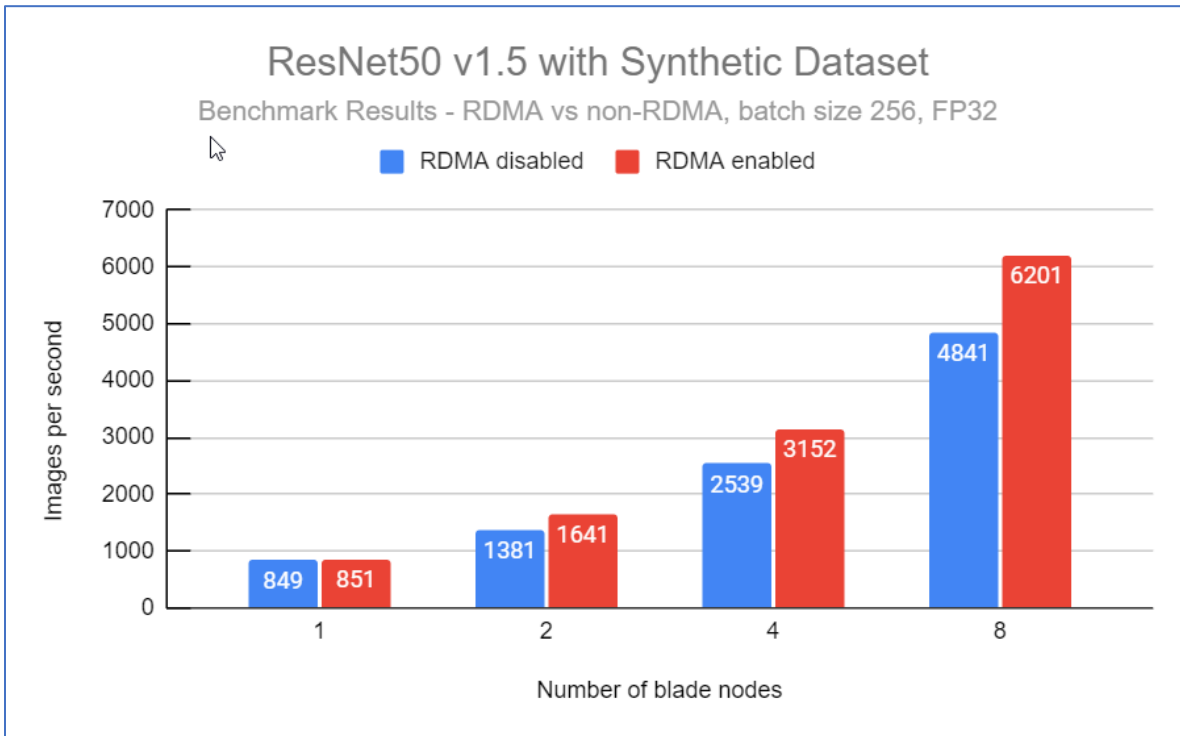


Figure 5 – Enabling RDMA boosts ResNet50 performance (Source: internal Supermicro testing)

For example, figure 5 above shows how enabling RDMA boosts ResNet50 benchmark performance by up to **30%**.

*Note: Other ML models (e.g., VGG16, ResNet101, and GoogleNet) produce similar benchmark results.*

Figure 6 shows each of the eight blades with a single GPU connected to the CPU in that blade. Horovod is a data-parallel distributed framework; it divides the whole dataset and copies it to each GPU.

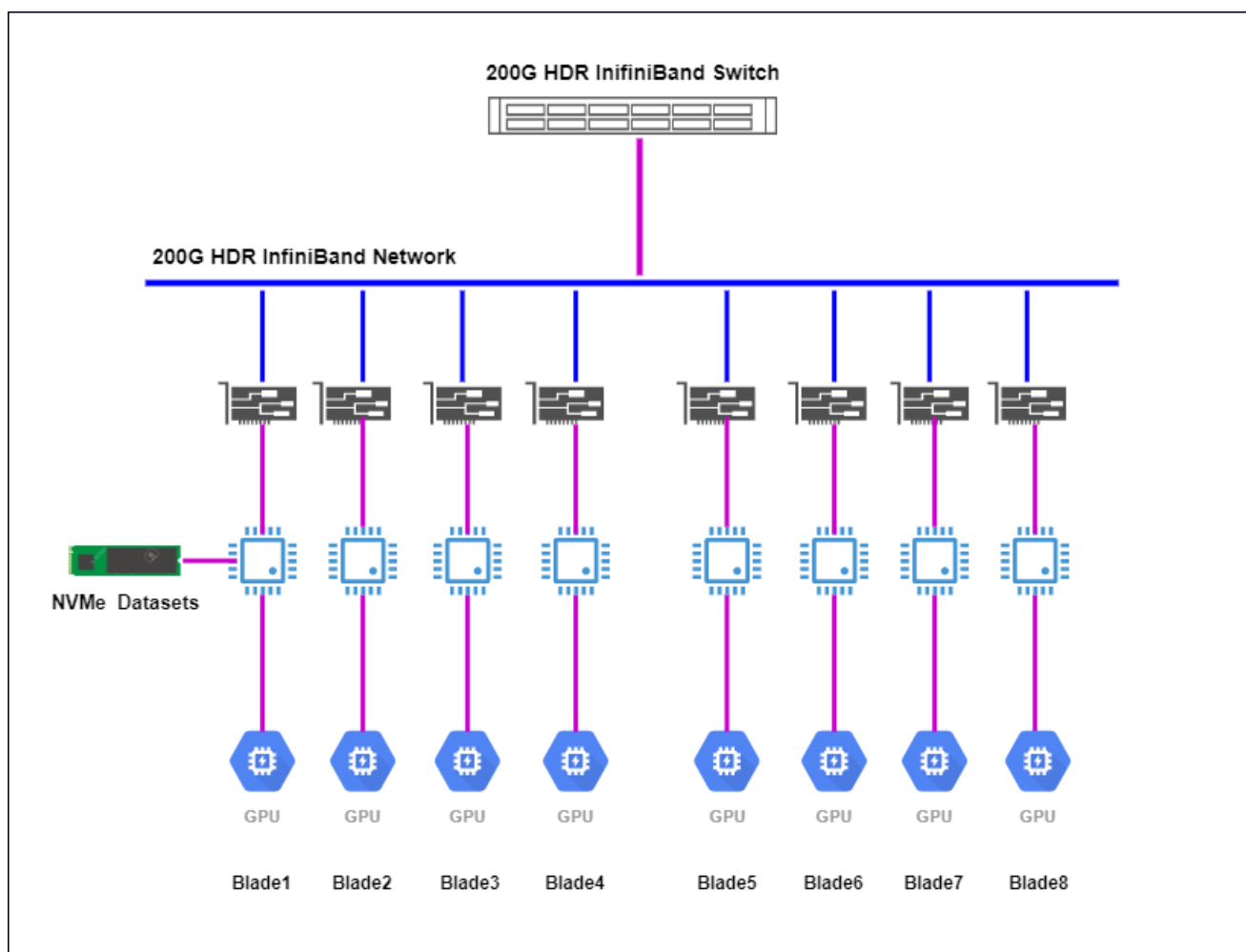


Figure 6 – Eight-blade SuperBlade system architecture

One GPU acts as the centralized parameter server that processes and redistributes the training results from different nodes. SuperBlade performs well in HPC environments because of its density and flexibility. The 8U SuperBlade, when fully populated with 20 nodes, is flexible to accommodate both Inferencing and Training simultaneously. For example, two of the 20 nodes can be used for inferencing, and the remaining 18 nodes can be used for Training models.

## Use Cases

Supermicro SuperBlade can improve performance-intensive computing, which is used to process large volumes of data or execute complex instruction sets in the fastest way possible. This is commonly used in artificial intelligence (AI), modeling and simulation, and Big Data and Analytic use cases. The Supermicro GPU-based SuperBlade accelerates solutions across retail, healthcare, financial service, transportation, automotive, media and entertainment, and manufacturing.

Having data insights at hand, digital innovators disrupt competitors' business models aggressively by innovating at a much faster pace. Some of the other areas where SuperBlade can provide benefits are:

- Data driven Enterprise Intelligence and AI everywhere
- Micro-segmented, hyper-personalized online shopping platforms
- GPS-driven ride-sharing companies
- Recommendation-driven streaming channels
- Adaptive learning-based educational tech companies
- Conversational AI-driven work scheduling

SuperBlade offers excellent operational efficiency by effectively allowing enterprises to automatically streamline processes, monitor for potential breakdowns, apply fixes and more efficiently facilitate the flow of accurate and actionable data throughout companies. In addition, this process improvement permits enterprises to offload maintenance work to machines so they can spend more time doing what they do best - which is innovating.

## Conclusion

Supermicro SuperBlade powered by 3<sup>rd</sup> Gen AMD EPYC processors delivers exceptional performance and TCO across multiple HPC and AI/ML workloads. GPU-accelerated SuperBlade servers are ideal for running converged AI/ML and HPC workloads, as evidenced by the excellent Horovod AI/ML workload performance described in this white paper and the ability to scale training across multiple nodes. In addition, Supermicro SuperBlade can accommodate simultaneous non-ML or AI/ML training and inferencing workloads. Supermicro GPU accelerated SuperBlade servers deliver high performance with high density, high throughput, and excellent scalability.