WHITE PAPER

SUPERMICRO® SYSTEM COMBINES AMD EPYC™PROCESSORS AND NVIDIA GPUS TO ACHIEVE CONSISTENT DEEP LEARNING PERFORMANCE WITH LINEAR SCALING

With the Supermicro A+ Ultra Server 2023US-TR4

TABLE OF CONTENTS

SUPERMICRO

Supermicro is a global leader in high performance, green computing server technology and innovation. We provide our global customers with application-optimized servers and workstations customized with blade, storage, and GPU solutions. Our products offer proven reliability, superior design, and one of the industry's broadest array of product configurations, to fit all computational needs.

Executive Summary

Selecting the right systems to process a deep learning workload can be very challenging. Deep learning benchmarking is a way to understand how a system performs with specific hardware and software environments, which are the essential references for data scientists and MLOps to choose systems that fit their IT infrastructure and meet the performance expectations. Supermicro is a leading GPU platform manufacturers and runs MLPerf benchmark to understand overall system performance. MLPerf is a benchmark that can boost potential customers' confidence in using Supermicro systems to solve a specific deep learning problem. MLPerf sets a deep learning benchmark standard for large CPU/GPU systems whose configuration often includes 8 or more GPU's, creates results with a relevant Deep Learning metric. Using frameworks provides the ability to produce repeatable measurements at a reasonable cost (which is essential for a GPU product). This paper presents the benchmark results for Supermicro AS -2023US-TR4, and AMD EPYC™ CPU based Supermicro high-end enterprise GPU server platform.

1. System Configuration

The Supermicro A+ Ultra Server 2023US-TR4 is a high-end enterprise server with Dual AMD EPYC™ 7002 Series Processors that supports 2 FHFL NVIDIA V100 for PCIe GPUs with flexible network configurations. This system provides excellent performance in virtualization, hyperconverged storage, and cloud computing configurations and is also a cost-effective server to process AI workloads with NVIDIA GPUs.

Table-1 High-level system configuration

Components	Part Descriptions	Quantity
Motherboard	H11DSU-iN with PCIe Gen3	1
System Model	AS -2023-US-TR4	1
CPU	AMD EPYC™ 7552 CPU (PSE-ROM7552-0076 48C/96T 2.2G)	2
RAM	MEM-DR464L-CL02-ER32 64 GB DDR4	16
NVMe	M.2 Samsung PM 983 3.8 TB	1
Storage	HDD-A8000 2Gb/s 7.2K 8TB	6
Network	Intel X710-BM2 10GbE SFP+	2
Storage controller	AOC-S3108L -H8iR	1
Riser	RSC-W2-66	1
GPU	NVIDIA V100 32GB PCIe	2

Table 1

REFERENCES

MLPerf white paper: https://arxiv.org/pdf/1910.01500.pdf

Memory and Batch Size https://towardsdatascience.com/how-to-break-gpu-memory-boundaries-even-with-large-batch-sizes-7a9c27a400ce

Supermicro Vertical Solutions: https://verticalsolutions.supermicro.com

© 2021 Copyright Super Micro Computer, Inc. All rights reserved

Figure 1 and **Figure 2** provide the system overview from different perspectives, the detailed system information is HERE.

AS -2023US-TR4

(Rear View – System)

Slot	Description
1	PCI-E 3.0 x16 Slot (FH, 10.5"L) (CPU1)
2	PCI-E 3.0 x8 Slot (FH, 10.5"L)** (CPU1)
3	Internal PCI-E 3.0 x8 Slot (LP, HL (6.6") (CPU1)
4	PCI-E 3.0 x8 Slot (LP, HL (6.6") (CPU2)
5~8	PCI-E 3.0 x8 Slot (

*Full redundancy based on configuration And application load.

**GPU kit needed to support up to 10.5" length, default supports up to 6.6" length.

Figure 1

AS -2023US-TR4

(Angled View – System)

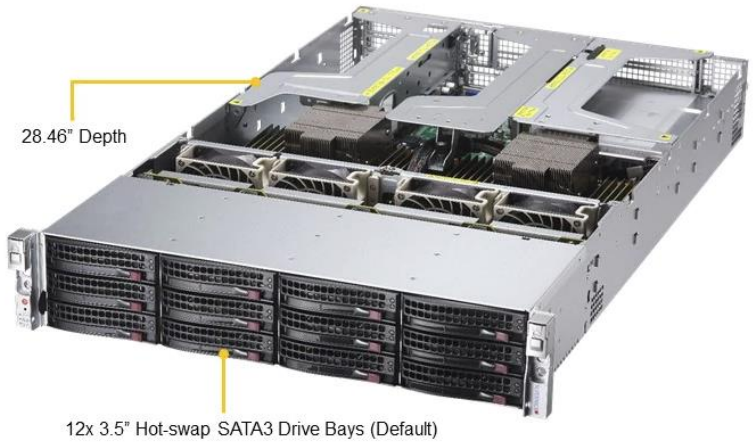


Figure 2

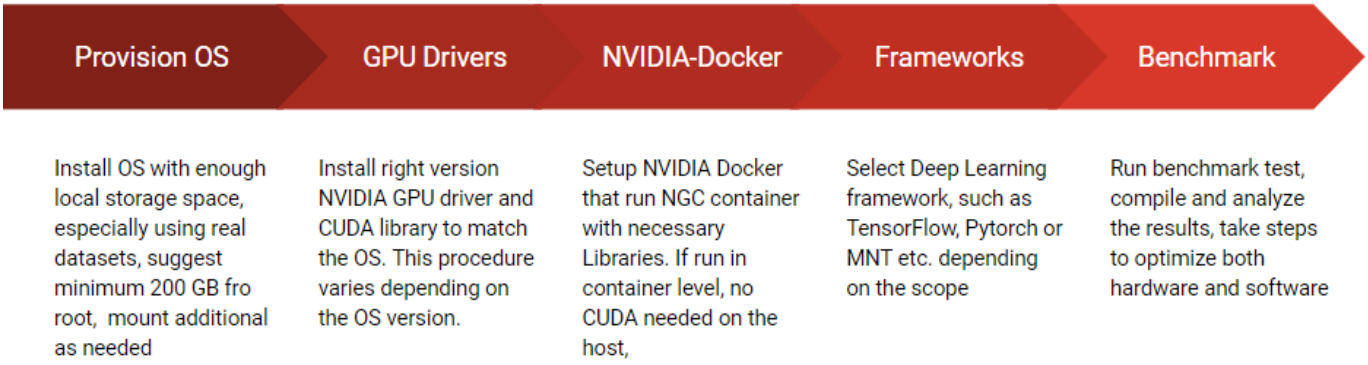
© 2021 Copyright Super Micro Computer, Inc. All rights reserved

2. Benchmarking Methodology

TensorFlow and Pytorch are two predominant deep learning software libraries, or Machine Learning/Deep Learning frameworks widely adopted by image classification and object detection applications. Both frameworks were used in this benchmarking exercise.

For details on TensorFlow please refer: https://www.tensorflow.org/. Similarly, details on PyTorch can be referenced at https://pytorch.org/

The following is a typical **workflow** to set up and run a **deep learning training benchmark** in **Docker containers** on a **single node**:



Configuration

There are many factors that directly and indirectly impact the efficiency of the deep learning workloads. To provide a comprehensive landscape showcasing how the server works under different hardware and Deep Learning software, a matrix of parameters were used in this test. With the Convolutional Neural Network(CNN network) the benchmark was performed on both bare-metal and on NVIDIA NGC containers.

Table-2 shows the details of these parameters.

Testing Variables	Descriptions
Batch Size	32 up to 1024 training samples depending on the limit of the GPU memory
CNN Model	GoogLeNet, AlexNet, ResNet, VGG
Framework	TensorFlow, PyTorch
Precision	Single precision FP32 and Half precision FP16
DataSets	synthetic, ImageNet, COCO
Number of GPU	1 and 2 respectively

Table-2

© 2021 Copyright Super Micro Computer, Inc. All rights reserved

There are two benchmarking methods – one with a synthetic dataset and one with a real dataset (such as ImageNet dataset). For performance benchmarking, a synthetic dataset is commonly used since processing a real dataset is expensive and time-consuming. For instance, the ImageNet real dataset size for image classification is 150 GB and it takes a considerable amount of time to preprocess, especially with the TensorFlow framework. Tests using synthetic datasets minimizes data preprocessing time. Also, it helps with the performance comparison of tests that were run on different Deep Learning network models and between tests run on CPU/GPU on a particular framework such as PyTorch.

3. The Benchmarking Configuration and Results

Tests were conducted on Supermicro A+ Ultra Server 2023-US-TR4 system running CentOS 8 using synthetic datasets.

Table 3 presents the test environment configuration.

Configuration Items	Descriptions
System Model (hardware)	AS -2023-US-TR4
System BIOS	Supermicro 2.2a
CPU	AMD EPYC™ 7552 CPU (PSE-ROM7552-0076 48C/96T 2.2G)
GPU	2xTesla V100 32GB PCIe
GPU BIOS	88.00.98.00.01
OS	CentOS 8
FrameWork (Software)	TensorFlow 1.14.0
Docker Container	Docker version 19.03.12 Build 48a66213fe
CDUA	Version 11.0
NVIDIA Driver	Version 450.51.05
NGC Container	NVIDIA/TensorFlow:20.08-tf1-py3
Dataset	synthetic
Benchmark type	Training

Table 3

© 2021 Copyright Super Micro Computer, Inc. All rights reserved

4. Batch Size and GPU memory*

Batch size, along with other hyperparameters, has a significant impact on training performance and accuracy. The possible range of batch sizes depends on GPU memory. For instance, ResNet50 model with single-precision (FP32), the batch size only goes up to 256 with the current system based on the V100 32GB cards. Among the models with current GPU memory capacity, batch size 512 is the preferable size across all tests (except ResNet152 where batch size 256 is preferred).

5. Bare metal systems versus Containerized systems

The Bare metal configuration requires the CUDA toolkit to be installed on the node, whereas the containerized setup only needs to install the NVIDIA Driver and NVIDIA Docker image for Tensorflow or PyTorch. In this benchmark, TensorFlow 1.14.0 Python package was installed by pip, the package manager written in Python, onto the host directly and the TensorFlow 1.14.0 Docker image was pulled down from the docker hub repository using the "Docker pull" command.

As Figure-3 shows, running benchmarks within a container or on a Bare metal host showed no significant performance difference. However, the container is a portable and reusable benchmark image for all systems. Therefore, the document will use the containerized test results.

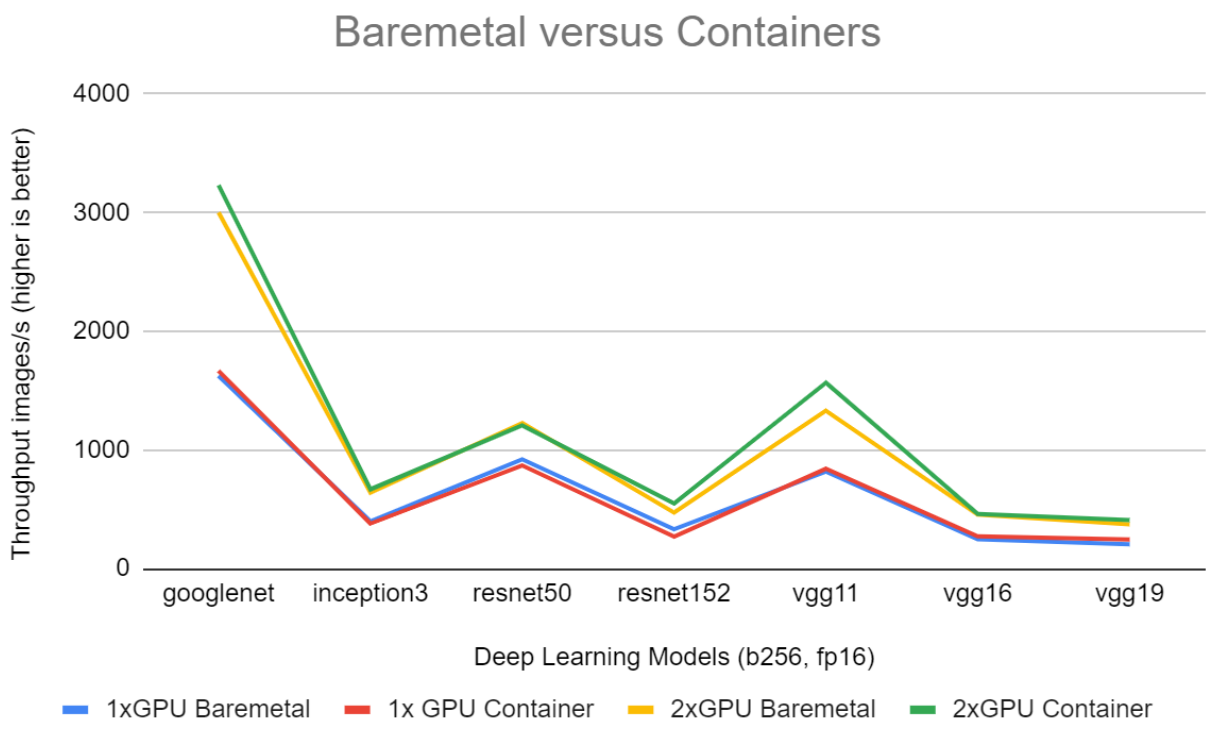


Figure 3

*Refer to following URL for information https://towardsdatascience.com/how-to-break-gpu-memory-boundaries-even-with-large-batch-sizes-7a9c27a400ce

© 2021 Copyright Super Micro Computer, Inc. All rights reserved

6. One GPU versus Two GPUs

With multiple GPUs installed, the system performance can scale depending on the type of workload, dataset size, and the number of GPUs that can be selected prior to a run.

Figure-4 depicts the scalability of a Supermicro GPU platform with 8 Tesla V100 32GB GPUs.

As shown in Figure-1, 2 Tesla V100 GPUs could be installed in slot1 and slot5 in AS -2023US-TR4 and connected to CPU1 via the Riser; when architected in this way, it is a Deep Learning cluster design. As we can see in Figure-5, the 2xGPU almost linearly doubles the single GPU. Due to GPU memory limitations, we were not able to produce any performance data for ResNet152 with batch size 512.

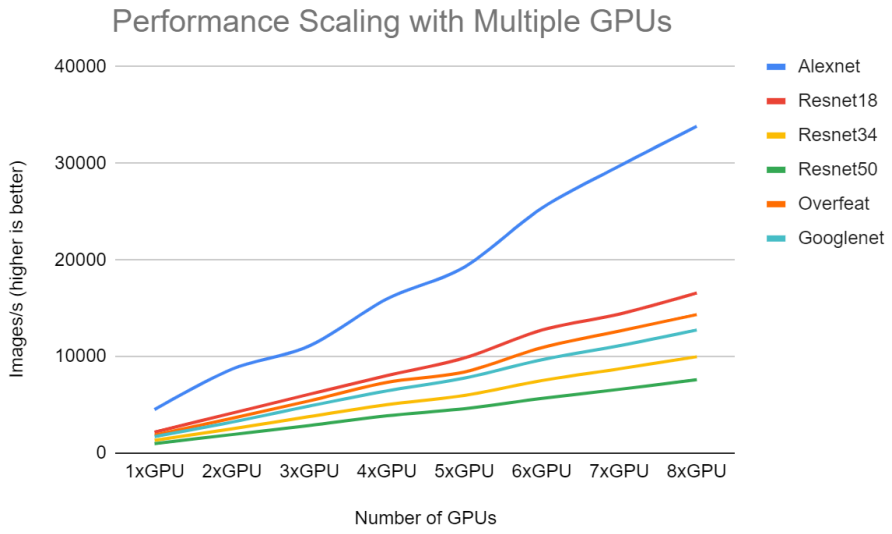


Figure 4

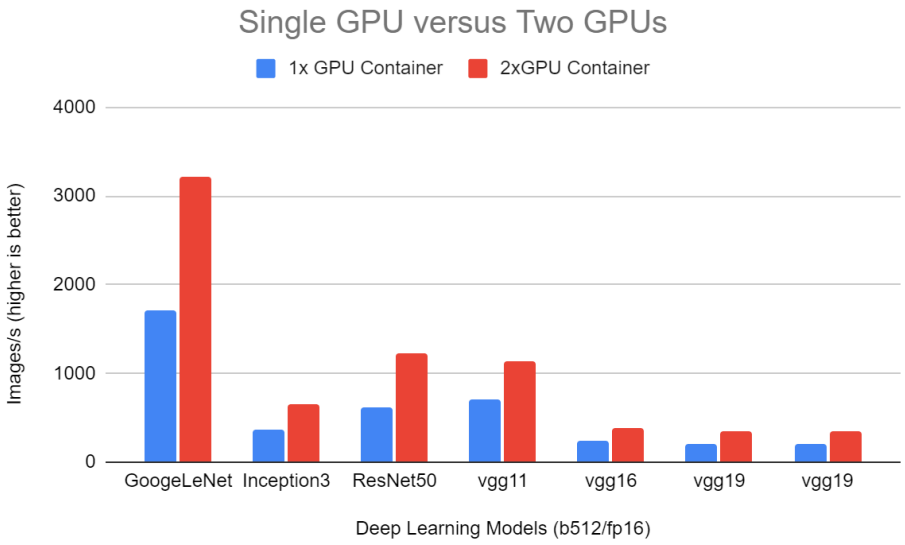


Figure 5

© 2021 Copyright Super Micro Computer, Inc. All rights reserved

7. Optimization with NVIDIA NGC

NVIDIA GPU Cloud, aka NGC, provides optimized containers that gear towards NVIDIA GPU. In this case, the server has two NVIDIA V100s for PCIe 32GB GPU. By leveraging NGC containers from a software perspective, you can see from the Figure-6 shows TensorFlow increases the AI workload performance. Tuning Neural Network is a complex process. Data scientists mathematically optimize models suitable for various scenarios. With AS-2023US-TR4 HPC architecture design, the test can take advantage of the newer TensorFlow frameworks.

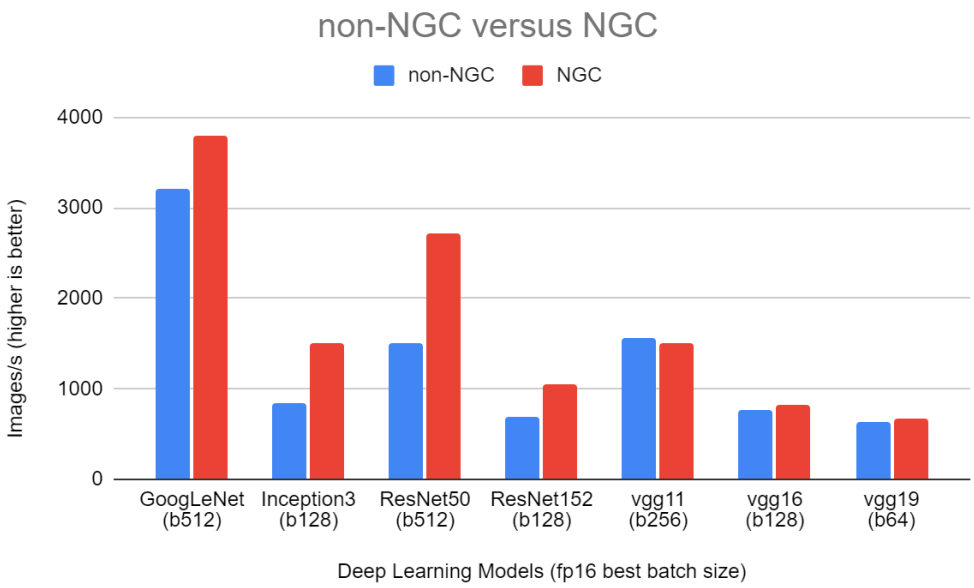


Figure 6

8. Conclusion

The benchmark results demonstrate very strong AI deep learning performance for the Supermicro A+ Ultra server 2023US-TR4 with the AMD EPYC™ CPU and 2 Tesla high performance GPUs. Whether it's Googlenet, Resnet50, or VGG19, the system shows very scalable performance from 1 to 2 GPUs. Besides, there is little performance impact running AI training in a container versus bare metal. Thus, running the NVIDIA NGC containers is the quickest way to become productive for the various image processing AI training workloads.

The 2U A+ Ultra server is modular and versatile and supports up to 2 high performance GPUs. While larger systems accelerate deep learning models, this is also true for many AI training applications, especially those using image processing neural networks. This Ultra server with 2 GPU offers the right performance and modularity. To support multiple users on the image recognition workload, system engineers can scale by adding servers with the same configuration in the rack.

If you are running smaller AI training or inference workloads, the 2U A+ Ultra server offers the right performance, modularity, and versatility.

All rights reserved. AMD, the AMD Arrow logo, EPYC, and combinations thereof are trademarks of Advanced Micro Devices, Inc.