WHITE PAPER

DEEP LEARNING BENCHMARKS ON SUPERMICRO'S 4U 8 GPU SYSTEM BASED ON DUAL 3RD GEN AMD EPYC™ PROCESSORS

TABLE OF CONTENTS

SUPERMICRO

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

Demonstrating the performance of the Supermicro AS -4124GS-TNR, a 4U dual-processor 8 GPU server with up to 8TB of memory, and 160 Lanes of PCI-E 4.0, shows the generation over generation performance improvements of the new 3rd Gen AMD EPYC 7003 Series Processors on Deep Learning benchmarks

Executive Summary

Artificial Intelligence is being adopted in various industries worldwide. The choice of systems to perform these complex tasks is critical and requires understanding how the different system components act together. A series of benchmarks have been created that allow those who evaluate systems and architectures to determine which combination of CPUs and GPUs are the best fit for their workloads.

AI workloads require optimized systems and need to incorporate the proper hardware and tuning the software to deliver maximum performance at a given price point. A solution that provides value to end-users consists of the choice of CPUs, GPUs, and the proper software stack. Various numbers of cores, communication latency between cores, GHz, and which generation of CPU architectures can influence benchmark performance of real-world AI applications.

A comparison will be run for this benchmark that compares 2nd Gen AMD EPYC™ processors to 3rd Gen AMD EPYC processors. AMD provides a wide range of processors with different numbers of cores and speed levels. Any AI/DL/ML application will depend heavily on the GPUs selected. Supermicro has run benchmarks that use different CPU generations and NVIDIA V100 and A100 GPUs. The CPU controls the management and assignment of work to the GPUs, while the GPU does the heavy lifting of transforming, loading, and analyzing the data. This is the training phase of AI deep learning, as well as inferencing.

SUPERMICRO SERVER FOR AMD EPYC AI BENCHMARKS

A+ Server 4124GS-TNR

Supermicro servers are designed for maximum application performance while minimizing power consumption.

- Multi-GPU optimized thermal designs for highest performance and reliability
- Advanced GPU interconnect options for best efficiency and lowest latency
- Leading GPU architectures including NVIDIA® HGX platform with NVLink™ and NVSwitch™

1. System Configuration

Supermicro designs and delivers a wide variety of servers and storage systems to enterprises worldwide. For these benchmarks, the AS-4124GS-TNR was the system of choice and features dual AMD 2nd Gen or 3rd Gen EPYC processors, up to 160 PCI-E Gen 4 lanes and with up to 8 PCI-E GPUs.

The benchmarks that are used in this paper are widely available, as is the software stack. The table below describes the software stack used for these benchmarks and the URLs' location where the software can be downloaded. Table 1 and Table 2 list these components.

Software		
Items	**Version**	**Source**
OS Ubuntu	18.04.4 TLS	https://al.mirror.kumi.systems/ubuntureleases/18.04.4/
NVIDIA Driver	450.51	https://www.nvidia.com/download/driverResults.aspx/162630
CUDA Version	11	https://developer.nvidia.com/cuda-downloads
Docker	20.10.3	https://docs.nvidia.com/datacenter/cloud-native/container-toolkit/install-guide.html
Nvidia-docker	2.5.0	
Mellanox driver	5.1	http://content.mellanox.com/ofed/MLNX_OFED-5.1-2.5.8.0

Table 1 - Software Specifications

Hardware		
Items	**Model**	**Description**
System	AS-4124GS-TNR	Supermicro GPU system SKU
MB	H12DSG-O-CPU	Motherboard BIOS 2.0
CPU	2x AMD EPYC 7313	3rd Gen AMD EPYC CPU
Memory	1 TB HMAA8GR7AJR4N-XN	System DRAM
Storage	SAMSUNG MZQLB7T6HMLA-00007	7 TB NVMe SSD drive
GPU	4xNVIDIA-A100PCIe-40GB	NVIDIA A100 40GB PCIe Gen4 GPUS
Network	MT28800 ConnecX-5	Mellanox 100 Gbps AOC

Table 2 - Hardware Specifications

Deep Learning Benchmark

There are many ways to benchmark a GPU system with a Deep Learning workload. Many types of workloads can be run as benchmarks, and a comprehensive list, with details, methodologies, and required software components, is maintained on github.com.

Supermicro is the first to benchmark a system's performance under different Neural Network applications, followed by benchmarking the GPU system with a real dataset. For comprehensive and a more controlled comparison of Deep Learning workloads, an increasing number of manufacturers and end customers are adopting the MLPerf suite, which covers a wide variety of AI/ML/DL workloads. Supermicro is committed to making the MLPerf benchmark as part of specifications for all GPU-capable systems.

The benchmarks that Supermicro ran are varied and need to be discussed separately.

1) Deep Learning performance with different Neural Network applications

Figure 1 shows the throughput of different Deep Learning Neural Network applications. The benchmark was run with the NVIDIA NGC container that the Deep Learning platform supplies, performs, and bare-metal systems. More details about Deep Learning Neural Network applications are available here for technical information about the Neural Network Applications.

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

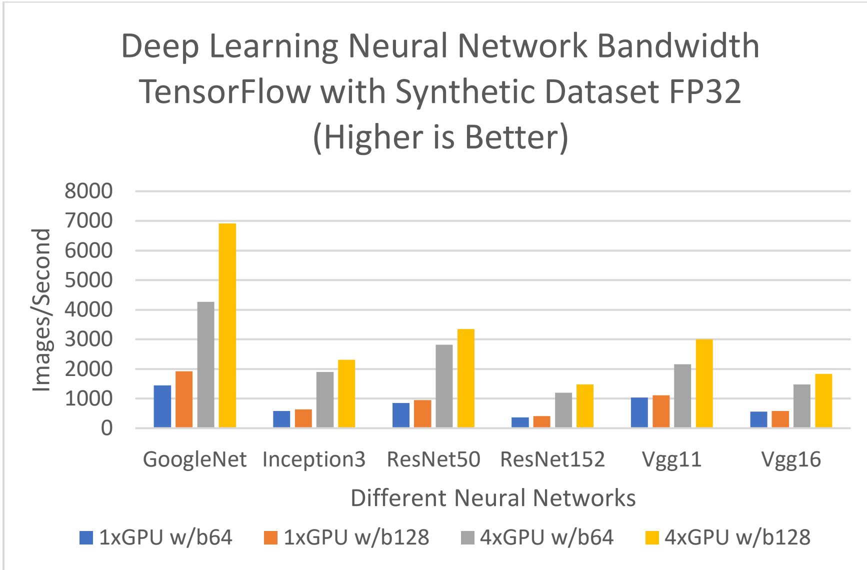

Figure 1 – Results Running the Deep Learning Neural Network Application Bandwidth AS-4124GS-TNR with 3rd Gen AMD EPYC 7313 CPUs

To be compliant with the with NVIDIA NGC.Ready fundamental performance criteria, the Google Neural Machine Translation(GNMT), and ImageNet Classification are used in this benchmarking system.

2) TensorFlow ImageNet Learning performance with ResNet50v1.50,single-precision FP32

There are many ways to benchmark a system in a given domain. Synthetic benchmarks are constructed to generate a specific workload on the underlying system and use its application to generate the data. Real-world workload benchmarks use actual data loaded into an application to produce results. Figure 2 and Figure 3 show the Deep Learning benchmark results with Synthetic and Real datasets, respectively.

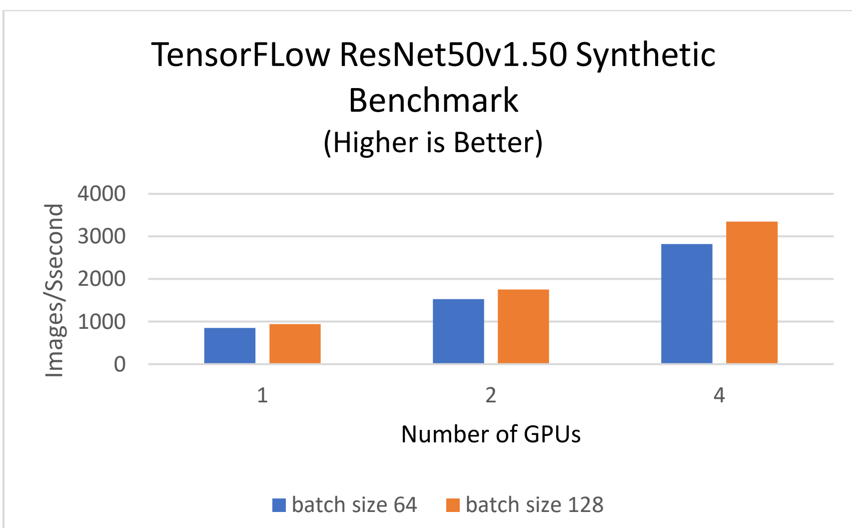

Figure 2 - Results running the TensorFlow Benchmark on an AS-4124GS-TNR with 3rd Gen AMD EPYC 7313 CPUs

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

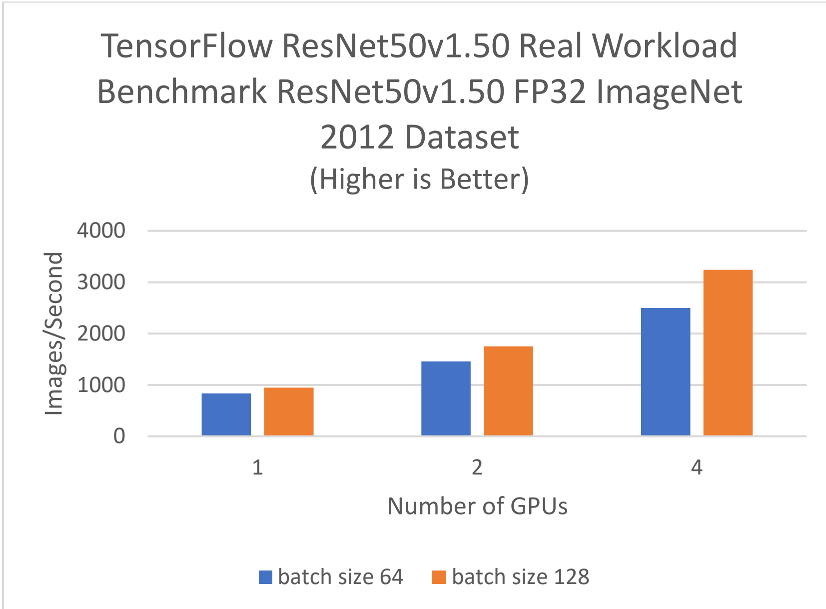

Figure 3 - Results running the TensorFlow Benchmark on an AS-4124GS-TNR with 3rd Gen AMD EPYC 7313 CPUs

To understand the basics of how this benchmark was set up and run, please look at the instructions posted here. This document contains step-by-step procedures for using an NGC container benchmarking with the NVIDIA GPUS system with both synthetic datasets and real datasets. There are also instructions for running the ResNet512 benchmark with TensorFlow. Visit this page to find out more about Deep Learning benchmark results and logs.

3) PyTorch GNMT Real Dataset Deep Learning performance with single-precision FP32

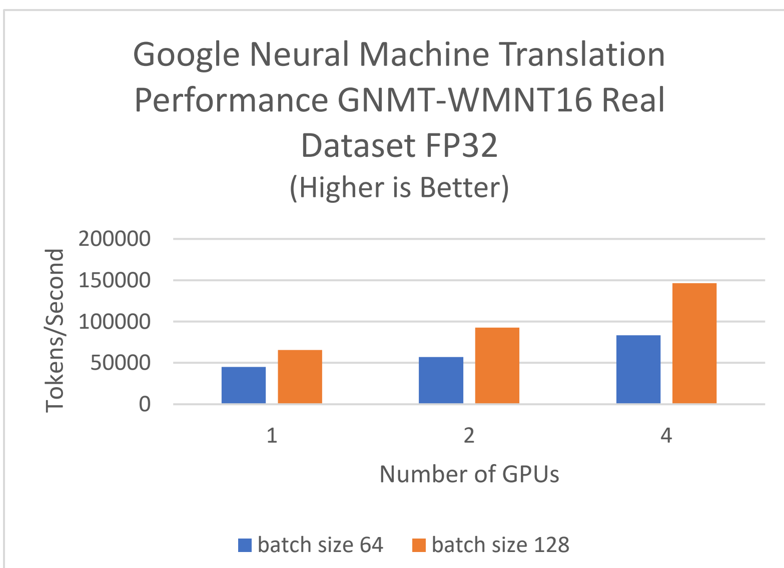

Figure 4 - Results running the Google Neural Machine Translation Performance on an AS-4124GS-TNR with 3rd Gen AMD EPYC 7313 CPUs

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

Deep Learning Benchmark Comparison using Different Workloads

Comparing benchmarking results with a similar base system and the same Deep Learning workload can give a business overview of how a GPU system throughput is improved by introducing more advanced CPUs and GPUs.

1) NVIDIA GPU – A100 vs. V100

The NVIDIA A100 GPU increases the Deep Learning training throughput and adds more features, such as TF32 Tensor Core and Multi-Instance-GPU (MIG). MIG virtualizes a physical GPU into seven different instances that isolate AI/ML/DL workloads and maximize GPU utilization without interference from another GPU. To learn more about TF32 and MIG, please read more at this link. Figure 5 shows the same Deep Learning workload on two different types of GPUs, and the throughput is up to 20-40%.

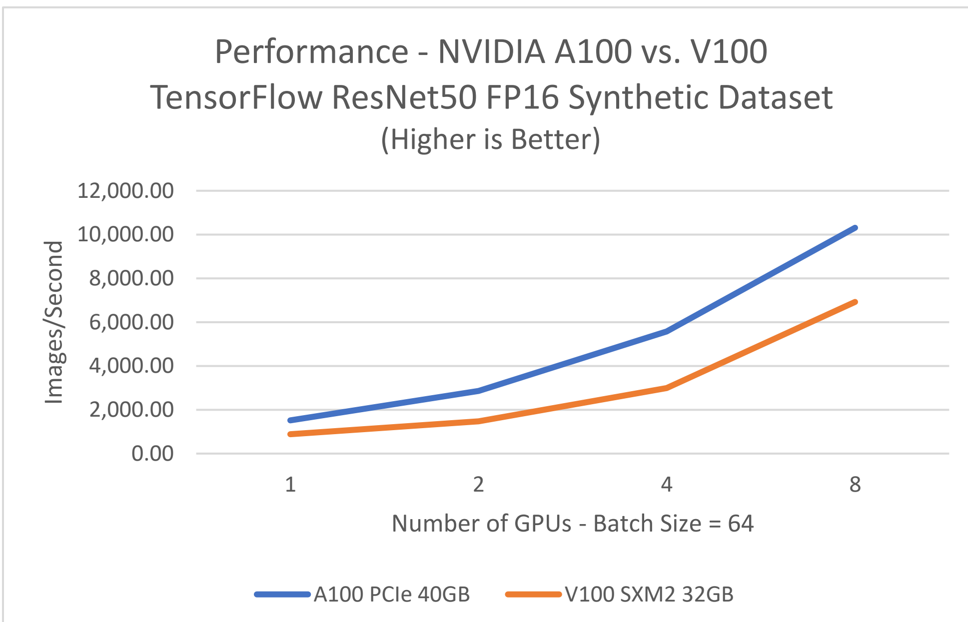

Figure 5 – Comparison of NVIDIA A100 and V100 GPU Results

2) AMD CPUs – 3rd Gen AMD EPYC (formerly codenamed "Milan") vs. 2nd Gen AMD EPYC (formerly codenamed "Rome")

Figure 6 indicates that the 3rd Gen AMD EPYC 7313 can perform better than the 2nd Gen AMD EPYC 7H12 on a comparable base system, training a model with ImageNet Synthetic dataset using single-precision FP32.

The 3rd Gen AMD EPYC 7313 has many new BIOS options critical to these new processors' high performance. As in previous generations of AMD EPYC processors, the setting of IOMMU and NPS are two of them that could significantly impact OS installation and overall performance. Please refer to the NVIDIA design guide, DG-10105-001, for PCIe servers. Tuning the CPUs to get datasets ready for Deep Learning applications is critical to both system designers and the end-user customers. Please look at the AMD resource guide for more information.

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

Benchmark results may vary if test conditions are different. The synthetic dataset indicates theoretical performance. Figure 6 is the benchmark results with the synthetic dataset. The real dataset, however, would more accurately represent Deep Learning application performance.

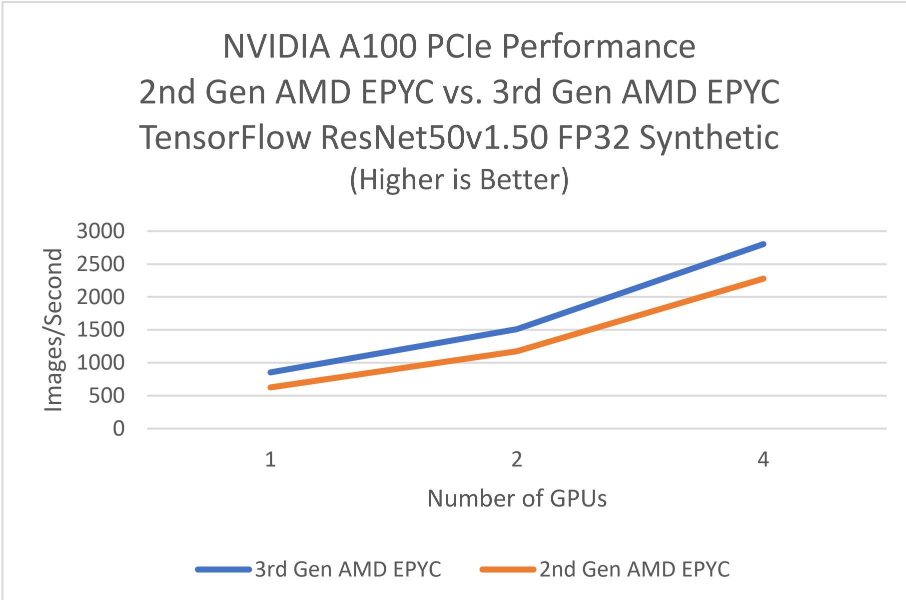

Figure 6 - Comparison of performance of 3rd Gen AMD EPYC vs. 2nd Gen AMD EPYC

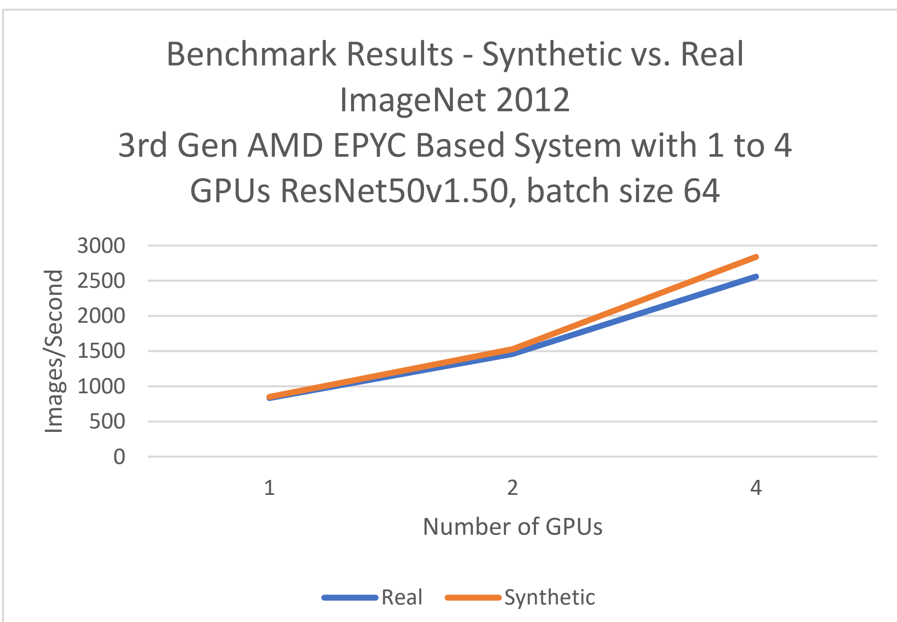

Figure 7 – Performance difference between Real and Synthetic Benchmarks with increasing number of GPUs.

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

Figure 7 shows the benchmark results between the Synthetic and Real benchmarks with the ImageNet 2012 datasets on the 3rd Gen AMD EPYC CPU. It can be up to a 33% throughput difference between synthetic and real ImageNet 2012 dataset, using ResNet50 v1.50 in a multi-GPU system. Therefore, benchmarking a system with a real Deep Learning related dataset is essential to assess performance accurately.

Conclusion

The benchmark results clearly show that the AMD EYPC 7313 processors improve the NVIDIA GPU system's throughput with Deep Learning workloads. A 15-30% generational increase in the synthetic benchmark test is seen with EPYC with the same NVIDIA A100 GPUs in a similar Supermicro chassis. The benchmark results also demonstrate that A100 PCIe can outperform V100 SXM2 in the comparable GPU systems up to 40%. Combined with the NVIDIA A100 GPU, Supermicro AMD CPU-based GPU systems are very flexible, competitive, and offer exceptional customer experiences. To learn more information on Supermicro GPU systems, please visit https://www.supermicro.com/en/products/GPU/

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

Supermicro, Server Building Block Solutions, and We Keep IT Green are trademarks and/or registered trademarks of Super Micro Computer, Inc.

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

All other brands, names, and trademarks are the property of their respective owners.