TABLE OF CONTENTS

SOLUTION BRIEF

INFERENCE APPLICATION OPTIMIZED SERVER - SUPERMICRO MAX IO

EXECUTIVE SUMMARY

A proliferation of AI-enabled services will mark this era of pervasive intelligence. There is an increased demand for services like image and speech recognition, natural language processing, visual search, and personalized recommendations. The growing complexity of networks and increasing storage creates a great demand for system performance, efficiency, and responsiveness that are critical to the computing needs of today and the future. Supermicro's inference platform offers these capabilities to power the next generation of AI products and services in the cloud. Supermicro's platform enables AI implementation in the Data Center, at the network's edge, and in the autonomous machines.

By utilizing Supermicro hardware and NVIDIA NGC docker platform, inferencing server can be deployed with a cost-effective and high-performance solution.

NVIDIA NGC DOCKER

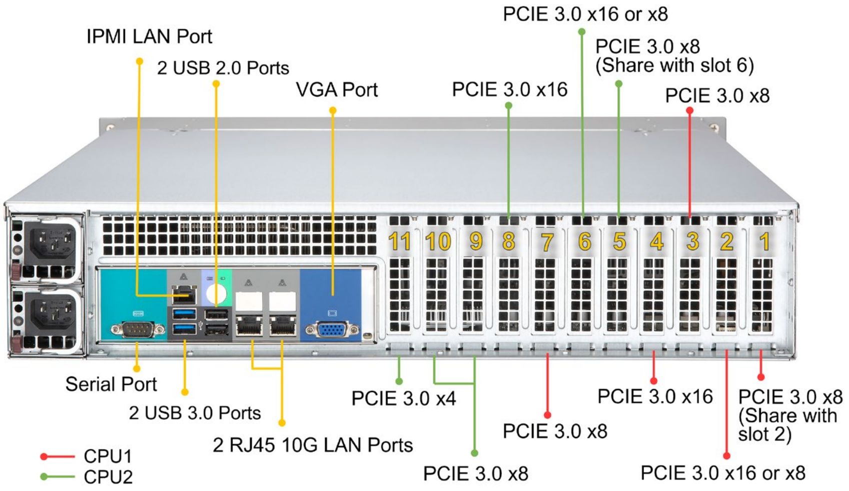
Figure 1. Supermicro/NVIDIA Inference Solution

Super Micro Computer, Inc.
980 Rock Avenue
San Jose, CA 95131 USA
www.supermicro.com

SUPERMICRO MAX IO SERVERS

Figure 2. Supermicro SYS-2029P-TXRT

Supermicro MAX IO solutions are resource optimized, cost-effective and ideal for space-constrained applications. Supermicro's highly dense yet compact server designs provide excellent compute, networking, storage and I/O expansion. Supermicro MAX IO solutions support a range of Intel technologies for a diverse portfolio of use cases. Support for high-performance enterprise-level New 2nd Gen Intel® Xeon® Scalable processors enables customers to meet the requirements of applications including Industrial Automation (IPC), Medical Imaging, Intelligent Transport, Digital Signage, Digital Security and Surveillance, Network, Storage appliances, Edge Computing and other applications.



FEATURES OF MAX IO SERVERS:

- **Dual Socket P (LGA 3647) support 2nd Gen Intel® Xeon® Scalable processors (Cascade Lake/Skylake).**
- **16 DIMMs; up to 4TB 3DS ECC DDR4-2933MHz RDIMM/LRDIMM, Supports Intel® Optane™ Persistent Memory**
- **CPU TDP support Up to 205W, 3 UPI up to 10.4 GT/s**
- **2 PCI-E 3.0 x16 slots**
  2 PCI-E 3.0 x16 slots (or 4 PCI-E 3.0 x8 by MUX)
  4 PCI-E 3.0 x8 slots
  1 PCI-E 3.0 x4 (in x8 slot)
  1 PCI-E 3.0 x4 M.2 slot
- **2x 10GBase-T LAN ports via Intel X550-AT2**
- **16 Hot-swap 2.5" SAS/SATA drive bays, 1 slim DVD-ROM drive bay**
- **1000W High efficiency (96%) Titanium Level Redundant PSU**

In order to show Supermicro MAX IO and NVIDIA inferencing servers are a perfect match. We use perf_ client benchmark tool to run some inferencing benchmark upon NVIDIA NGC inferencing server with the following setup.

HARDWARE

- System: Supermicro 2029P-TXRT
- CPU: 2x CLX8270
- Memory: 12x 32GB 2666MHz
- GPU: 8x Tesla T4

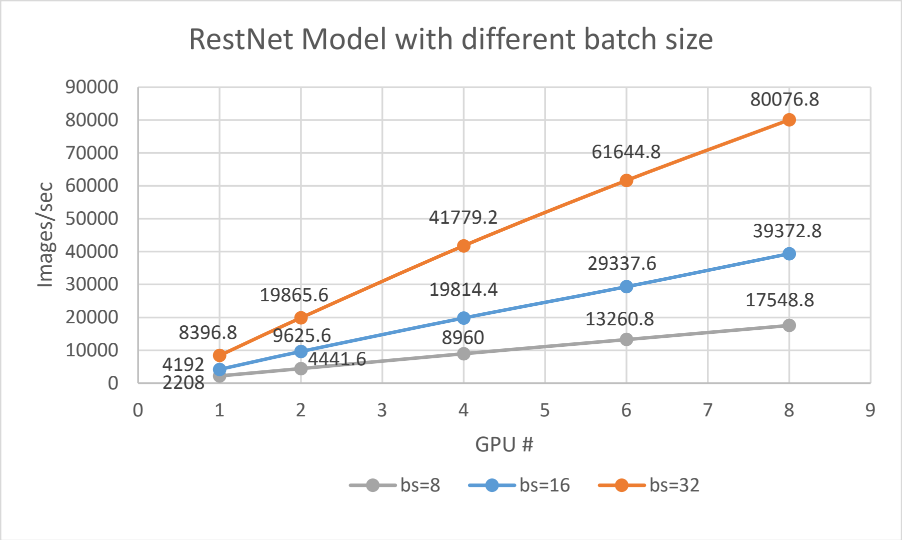
SOFTWARE

- OS: Ubuntu 18.04.3 LTS
- Driver version: 418.87.00 with CUDA 10.1
- Docker version 19.03.5
- TRTIS version: 0.10.0
- TRTIS container version: 19.12

PERF_CLIENT TESTING PARAMETERS

- Instance per GPU: 1
- Batch size: 8, 16, 32
- Latency: 200 ms
- Measurement window: 5000 msec
- Concurrency: various from 0 to maximum 100 threads, testing result only shows the concurrency number that has the maximum throughput.

The following data is running based on ResNet-50 model with input image size as 224*224 three channels, more information please refer to reference[1]. Final result Image per second on Y-axis is calculated by multiplying Inferences/second and batch size.



Testing results show that even when MAX IO is populated with up to 8 pieces of T4, inference server can get nearly linear Inferencing scalability on ResNet-50 with different batch sizes including 8, 16 and 32. With 8 T4 populated in MAX IO from slot 0 to slot 7, we have 6 out of 8 T4 running in PCI-E x8 bandwidth. However, inferencing benchmark result on ResNet50 with small to medium batch sizes shows that we can fully utilized T4's computing power without being capped by PCI-E x8 bandwidths.

[1] *Source:* https://github.com/NVIDIA/tensorrt-inference-server/blob/master/docs/examples/model_repository/resnet50_netdef/config.pbtxt

CONCLUSION

GPU scalable inferencing capability is just one of the applications that MAX IO can provide, by offering 11 PCI-E slots in a 2U form factor, MAX IO can also be used in applications that needs lots of IOs in single node such as storage control node and network node.

This scalable design will allow our customer to customize their configuration based on the actual application.

ABOUT SUPER MICRO COMPUTER, INC.

Supermicro® (NASDAQ: SMCI), the leading innovator in high-performance, high-efficiency server technology is a premier provider of advanced server Building Block Solutions® for Data Center, Cloud Computing, Enterprise IT, Hadoop/Big Data, HPC and Embedded 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.

www.supermicro.com

Printed in USA ♻ Please Recycle 14_NVIDIA-MAX IO-Inference_2020_02-1