TABLE OF CONTENTS

SOLUTION BRIEF

SCALING AI TRAINING WITH SUPERMICRO DESIGNED NVIDIA GPU SYSTEMS

Supermicro NGC-Ready Systems

EXECUTIVE SUMMARY

Time to market is key to success for today's AI development. By using Supermicro designed NVIDIA GPU systems that have all the latest AI stack installed and supported, data scientists and AI developers can start testing and training their AI models for product development and research.

New AI models, such as BERT, GPT-2, or R-CNN require the compute power of multiple GPUs to solve them. Strong scale-up nodes can be built by using NVIDIA® NVLink™ and NVSwitch™ interconnect technology within a single server, and GPUDirect RDMA across servers. Supermicro offers a set of scalable multi-GPU systems using the fastest NVLink/NVSwitch GPU interconnects supporting 4, 8, and 16-GPU, and high speed Mellanox interconnect between servers. The more GPUs that are aggregated, the shorter the AI training time.

Other times, each GPU can be shared by multiple data scientists and developers. Scaling from fractional GPU use to aggregating thousands of GPUs, NVIDIA GPU Cloud (NGC) is the most cost effective way to get these multi-GPU systems ready for AI development.

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

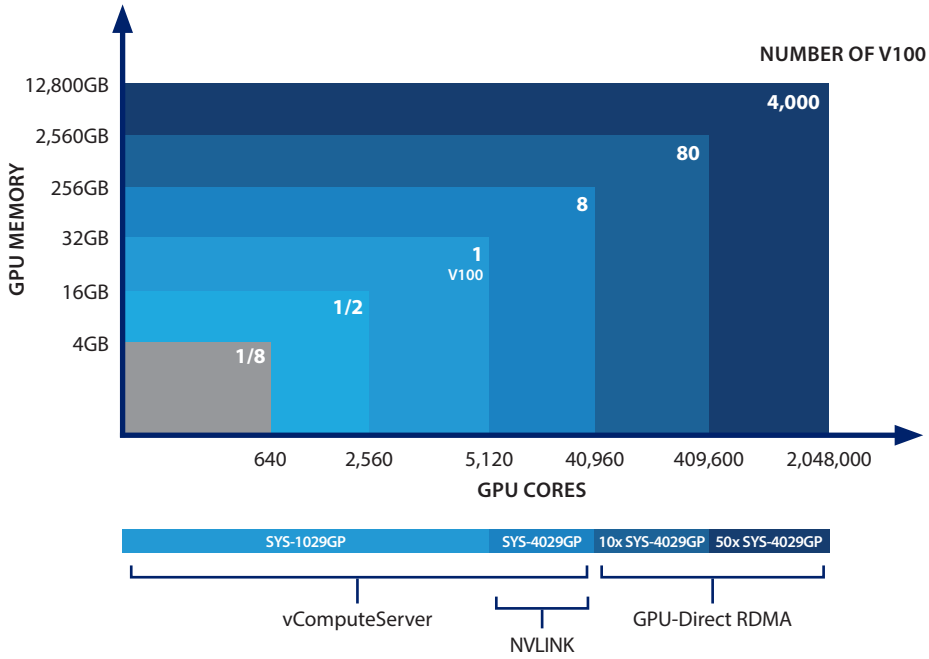


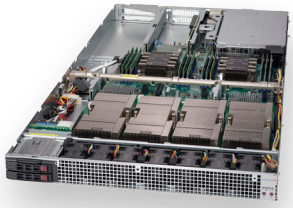
Figure 2. *GPU scaling for more CUDA cores and bigger GPU memory.*

AI READY NVIDIA GPU CLOUD CONTAINERS

NGC is the hub for GPU-optimized software that contains quality assured, enterprise-grade containers for AI and HPC applications, pre-trained models, model scripts, helm charts and industry SDKs. This allows data scientists, developers and IT managers to access, build, and deploy apps across various platforms, from on-premise to cloud to edge. The software can be pre-installed on the systems to enable the following:

- Latest optimized and secure AI models, many of which are pre-trained
- Containers run anywhere, supports CI/CD infrastructure
- Full 24x7 support

HIGHEST PERFORMANCE MULTI-GPU SYSTEMS



SYS-1029GP-TVRT

SYS-4029GP-TVRT

SYS-9029GP-TNVRT

Supermicro offers a scalable set of multi GPU systems with NVIDIA V100 GPUs interconnected with NVIDIA® NVLink™ and NVSwitch™. These systems offer an increasing number of aggregate CUDA cores and GPU memory for scaling larger AI models and batch data.

SUPERMICRO SXM-GPU SYSTEMS	CONFIGURATION WITH ADDITIONAL OPTIONS
SYS-1029GP-TVRT With 4 NVIDIA V100 SXM2 GPUs 300GB/s NVIDIA NVLink™ GPUDirect RDMA support Aggregate 20,480 CUDA cores, 128GB GPU memory	1U Rackmountable, 35.2" depth Dual 2nd Gen Intel Xeon Scalable Processors Max 3TB ECC memory 4 PCI-E 3.0 x16 slots for I/O 2 SAS/SATA drives 2000W (1+1) Redundant Titanium Level power IPMI, Redfish
SYS-4029GP-TVRT With 8 NVIDIA V100 SXM2 GPUs 300GB/s NVIDIA NVLink GPUDirect RDMA support Aggregate 40,960 CUDA cores, 256GB GPU memory	4U Rackmountable, 31.7" depth Dual 2nd Gen Intel Xeon Scalable Processors Max 6TB ECC memory 6 PCI-E 3.0 x16 slots for I/O 16 SAS/SATA (optional 8 NVMe) drives 2200W (2+2) Redundant Titanium Level power IPMI, Redfish
SYS-9029GP-TNVRT With 16 NVIDIA V100 SXM3 GPUs 300GB/s NVIDIA NVLink and NVSwitch fabric GPUDirect RDMA support Aggregate 81,920 CUDA cores, 512GB GPU memory	10U Rackmountable, 27.75" depth Dual 2nd Gen Intel Xeon Scalable Processors Max 3TB ECC memory 16 PCI-E 3.0 x16 slots for I/O 2 SAS/SATA drives 3000W (5+1) Redundant Titanium Level power IPMI, Redfish

For AI models and data batches that require more CUDA cores and GPU memory than that available from the largest 16-GPU system, GPUDirect RDMA and NCCL are used to scale the GPU cores and memory over 100Gigabit InfiniBand or Ethernet fabrics. NGC software makes deployment of multiple GPU systems easy. Tens or hundreds of these systems can be aggregated to run the biggest AI models and data batches. Combined with high performance NVMe-fabric storage and networking, these systems build complex AI systems with ease and speed.

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

No part of this document covered by copyright may be reproduced in any form or by any means — graphic, electronic, or mechanical, including photo-copying, recording, taping, or storage in an electronic retrieval system — without prior written permission of the copyright owner.

Supermicro, the Supermicro logo, Building Block Solutions, We Keep IT Green, SuperServer, Twin, BigTwin, TwinPro, TwinPro², SuperDoctor are trademarks and/or registered trademarks of Super Micro Computer, Inc.

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

© Copyright 2020 Super Micro Computer, Inc. All rights reserved.

Printed in USA ♻ Please Recycle 14_Scaling-AI-Train_2020_01-2