SOLUTION BRIEF

MAXIMIZING AI DEVELOPMENT & DELIVERY WITH VIRTUALIZED NVIDIA A100 GPUS

Supermicro Systems with NVIDIA HGX A100 4-GPU Optimized for NVIDIA Virtual Compute Server and Multi-Instance GPU

AS -2124GQ-NART

SYS-220GQ-TNAR

TABLE OF CONTENTS

SUPERMICRO

As a global leader in high performance, high efficiency server technology and innovation, we develop and provide end-to-end green computing solutions to the data center, cloud computing, enterprise IT, big data, HPC, and embedded markets. Our Building Block Solutions® approach allows us to provide a broad range of SKUs, and enables us to build and deliver application-optimized solutions based upon your requirements.

Executive Summary

A flexible AI processing environment makes it easy to incorporate AI into IT workflow. Many applications involve small to large AI models and different data batch sizes. Simultaneously, AI product development and IT deployment require a wide range of AI processing capabilities on demand. NVIDIA® Virtual Compute Server (vCS) and Red Hat® Virtualization running on Supermicro's systems with NVIDIA HGX A100 provide that flexibility, cost effectiveness, and responsiveness to run AI workloads for developers and deployment.

We describe in this paper the capabilities and flexibility of Supermicro's systems with NVIDIA HGX A100 and NVIDIA vCS combined with Red Hat Enterprise Linux® to virtualize NVIDIA A100 GPUs to run independent AI workloads. Multiple developers, some needing a small GPU and some needing multiple GPUs, can share the same HGX systems. The same systems also support AI workloads incorporated into IT applications at the same time.

SUPERMICRO HGX SERVERS USING NVIDIA A100 WITH NVLINK

AS -2124GQ-NART, with AMD EPYC™ CPUs

SYS-220GQ-TNAR with Intel® Xeon® CPUs

Supermicro servers with NVIDIA HGX A100 deliver highest-performance A100 GPUs with 600GB/s NVLink PEER-to-PEER connections. Choices of 3rd Gen AMD EPYC™ or 3rd Gen Intel® Xeon® Scalable processors, system memory up to 8TB, PCIe Gen4 connectivity, NVMe drives, 200Gbit/s network connectivity, A100 GPUs with either 40GB/80GB memory, redundant power IPMI/Redfish v1.8 monitoring/ management, TPM 2.0, hardware Root of Trust 1.0 security.

Glossary	
vCS	Virtual Compute Server
vGPU	Virtual GPU, used for vCS
MIG	Multi-Instance GPU, A100 feature
Multi-vGPU	Aggregation of virtual GPU, vCS deployment with MIG disabled
Multi-GPU	Aggregation of physical GPU, bare-metal deployment

Virtualizing Supermicro Systems with NVIDIA HGX A100 for AI Workloads

Depending on the size of the AI model and batch data, an AI developer might need a fraction of a GPU or multiple GPUs. NVIDIA vCS enables easy sharing of the same GPU resources running on the Red Hat Enterprise Linux (RHEL) KVM hypervisor for multiple developers. vCS assigns to each virtual machine (VM) a fraction of a GPU or multiple GPUs connected by NVIDIA® NVLink™ in a Supermicro system with NVIDIA HGX A100. By virtualizing the shared GPU, vCS seamlessly allocates the GPUs based on user needs. This flexibility is easier to manage than managing multiple NVIDIA HGX servers running independent operating systems. Red Hat Virtualization provides system redundancy with auto VM failover and scalability.

NVIDIA A100 GPUs in Supermicro systems are the state-of-the-art AI processing engines, with 3rd generation Tensor Cores supporting sparsity acceleration, 6,912 CUDA Cores, and up to 80GB of fast GPU HBM2e memory. Multiple NVIDIA A100 GPUs can be combined into a single large processing unit using NVIDIA GPUDirect® Peer-to-Peer over NVLink 3.0. Each NVIDIA A100 could also be partitioned into 2 to 7 Multi-Instance GPUs (MIG), where each MIG behaves as an independent smaller GPU.

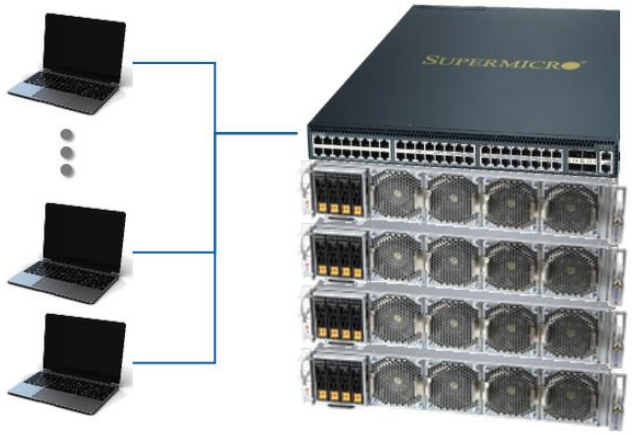

Figure 1. vCS and Red Hat Enterprise Linux running on Supermicro servers with NVIDIA HGX A100, supporting multiple developers, with on-demand GPU for AI

vCS offers flexibility to support one or more users to run multi-vGPUs as an aggregated large GPU. vCS aggregates the vGPUs using GPUDirect Peer-to-Peer in the server and RDMA GPUDirect across servers.

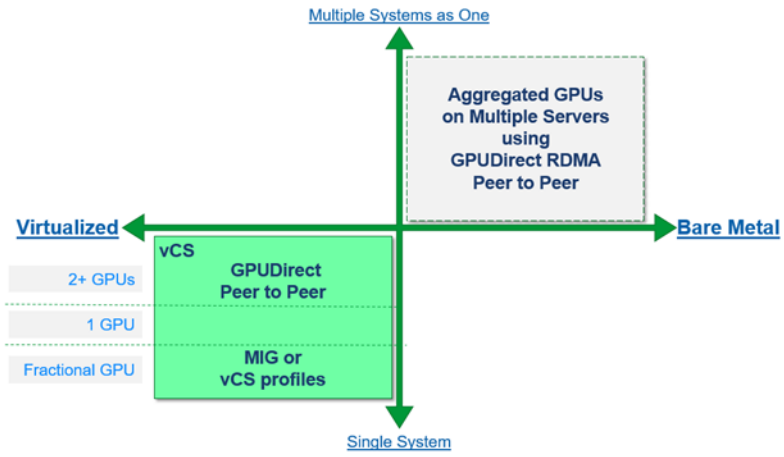

Figure 2. Two ways to use Supermicro servers with NVIDIA HGX A100: (left hand side) vCS virtualizes GPU resources into 1/10 of an NVIDIA A100 to all GPUs in one server; (right hand side) aggregating GPUs on multiple servers for very large jobs, using GPUDirect RDMA Peer-to-Peer running on servers with bare-metal OS (instead of vCS).

In other situations, when users need just fractions of an A100 GPU's power. vCS can virtualize the GPU into as small as 1/20 of a GPU for users. There are two vCS modes. In Enabled MIG Mode, vCS dedicates 1/7 to the entire GPU to a user. In Disabled MIG Mode, where all the partitions are the same, a user could get a virtual GPU that is 1/20 of an NVIDIA A100 GPU to an aggregation of 16 vGPUs from vCS.

	A100 MIG Mode Enabled	A100 MIG Mode Disabled
Max partitions	7	10 (A100/40GB) 20 (A100/80GB)
Partition Type	SPACE-SLICED	TIME-SLICED
Partition Sizes	Different sizes, as long as they add up to 1 GPU	All the same size per GPU
Largest vGPU	One A100	16 vGPU
Compute resources	Dedicated	Shared
NVIDIA NVLink Support	No	Yes
Heterogeneous Profiles	Yes	No

Figure 3. vCS supports MIG Mode and disabled MIG Mode, with different functions.

Red Hat Virtualization

NVIDIA vCS runs on top of a hypervisor on the Supermicro server with NVIDIA HGX A100. The Red Hat Virtualization hypervisor allocates CPU and main memory resources for virtual machines to support each user, while vCS allocates the GPU resources. Each virtual machine can run a Linux OS. Red Hat Enterprise Linux provides a robust, enterprise-class hypervisor, offering virtualization that scales from one to racks of servers. Key features include:

- System scheduler
- Storage and network management
- Hot plug of virtual resources
- User and group based authentication and security
- Red Hat Virtualization Manager
- Red Hat Virtualization hypervisor

Red Hat Virtualization

Supermicro servers with NVIDIA HGX A100 are validated to run Red Hat Enterprise Linux. RHEL supports KVM 8.3, which supports NVIDIA vCS.

NVIDIA vCS and AI Workflows

While NVIDIA vCS virtualizes the GPUs in one Supermicro GPU system, Using Red Hat Virtualization and management tools, multiple Supermicro servers running NVIDIA vCS can be deployed to provide a data center virtualized infrastructure to support scalable AI workflows.

Each AI developer can quickly spin up a virtual machine with the GPU resource he/she would need. The developer modifies the AI model, trains it with different data batches, and then tests AI inference. The developer can use one virtual machine or multiple VMs. Multiple developers can do their development independently of each other's, as NVIDIA vCS would allocate system and GPU resources as needed.

When the AI model is trained and incorporated into the end application and ready for deployment, NVIDIA vCS can also be used to operate and scale the AI application one or more virtual machines, with allocated GPU resources to run the AI inference.

Multi-Tenancy could also be supported using vCS and the underlying virtualization hypervisor. Furthermore, Kubernetes infrastructure could be added to drive containers running in the vCS virtual machines. The vCS VMs would run as worker nodes in the Kubernetes infrastructure.

NVIDIA vCS and Red Hat Virtualization running on Supermicro systems with NVIDIA HGX A100 are very flexible and scalable.

Supermicro Systems with NVIDIA HGX A100 and NVIDIA Certification

Supermicro systems with NVIDIA HGX A100 4-GPU technologies are NVIDIA-Certified. These systems have passed the NVIDIA certification tests to run as a single server and run multiple servers using 200-gigabit networks, supporting the software containers and AI frameworks from the NVIDIA NGC catalog. NGC Support Services are available to help customers develop and deploy AI and HPC systems running on these Supermicro servers.

These Supermicro systems can run bare metal using either Canonical Ubuntu or Red Hat Enterprise Linux as the operating system. The NGC software runs in containers under the Linux OS. Alternatively, these systems can run Red Hat Virtualization and NVIDIA vCS – this environment allows customers to manage multiple systems to provide Virtual Machines, allocating virtual GPUs, to run the NGC software containers.

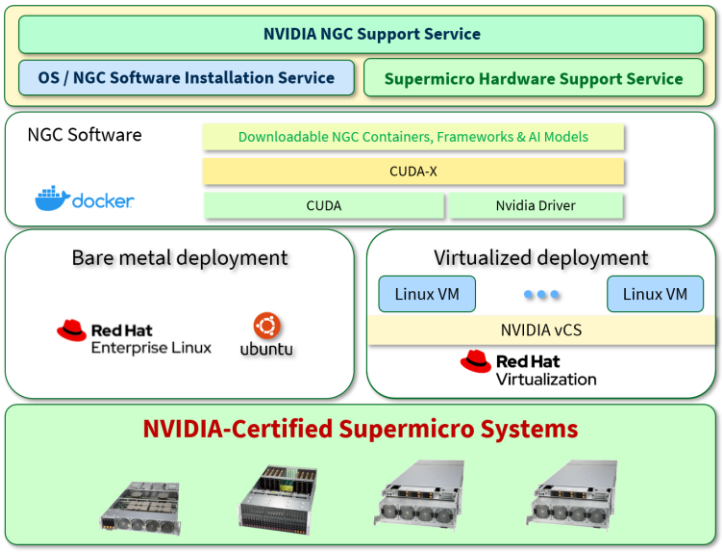

Figure 4. Virtualized and Bare metal deployment options for NVIDIA-Certified Supermicro systems.

For an NVIDIA HGX A100 4-GPU systems, Supermicro offers both Intel and AMD CPUs. For example, a customer may choose Intel CPUs because of application fit, while another might choose AMD CPUs because he wants more CPU cores for AI pre-training data processing.

NVIDIA Virtual Compute Server (vCS) and NGC

NVIDIA vCS supports NVIDIA NGC. All the containers, pre-trained AI models, Helm Charts, GPU Operators in the NGC Catalog can run inside the virtual GPU setup by the vCS. With NVIDIA certification, NGC Support Services are available for the Supermicro servers with NVIDIA HGX to help customer accelerate their use of the NGC capabilities. NGC Support Services provide problem resolution to customers using NGC containers and AI models running on the Supermicro servers with NVIDIA HGX.

More vCS Choices

In addition to solutions using Supermicro servers with NVIDIA HGX A100 and Red Hat Virtualization described here, Supermicro also offers other choices to provide NVIDIA vCS capabilities. The CPU choices allow the best fit for AI workloads and cost effectiveness. For considerable AI training, it is best to use HGX systems, whereas, for small inference jobs, the NVIDIA A10 might be more cost-effective. The server choices allow the best fit into customer IT infrastructure and needs. Supermicro Twin and Ultra servers might be best if there is a strong need for more CPU processing in addition to GPU operations. Choose the hypervisor that fits best into the existing customer virtual environment. Supermicro offers all these choices to enable the best fit for customer needs.

Virtualized GPU	vCS (compute only)	vCS (compute only)	vCS for Compute, vWS for 3D graphics acceleration	vCS for Compute, vWS for 3D graphics acceleration
GPU Choices	NVIDIA HGX A100 4 GPU, with 40GB or 80GB	NVIDIA A100 / PCIe (choice of 40GB or 80GB) or NVIDIA A30 / PCIe	NVIDIA A40 / PCIe or NVIDIA A10 / PCIe	T4 / PCIe
Server Choices	AS -2124GQ-NART AS -4124GO-NART SYS-220GQ-TNAR	AS -4124GS-TNR	AS -4124GS-TNR	AS -4124GS-TNR, Twin servers, Ultra servers
Virtualization Hypervisor Choices	VMware, Red Hat	VMware, Red Hat	VMware, Red Hat	VMware, Red Hat

Example Applications

Here are example applications using the virtualized machine learning/HPC infrastructure. Specific customer solutions need to situation need to be adjusted to match customer needs.

Virtualized GPU	Number of Simultaneous Users (VMs)	CPU Cores	System Memory	Storage	NVIDIA A100-GPU	GPU System
AI/HPC Development – Small Jobs (Single GPU)	Up to 80 (4C)	128	512GB	100TB	4	AS -2124GQ-NART or SYS-220GQ-TNAR
	Up to 160 (4C)	256	1024GB	200TB	8	2x AS -2124GQ-NART or 2x SYS-220GQ-TNAR
AI/HPC Development – Large Jobs (multiple GPUs)	32 (20C)	64 (AMD) 40 (Intel)	1024GB	400TB	16	4x AS -2124GQ-NART or 4x SYS-220GQ-TNAR
	64 (20C)	128 (AMD) 80 (Intel)	1024GB	800TB	32	8x AS -2124GQ-NART or 8x SYS-220GQ-TNAR
AI Inference	Up to 80 (4C)	128	256GB	10TB	4	AS -2124GQ-NART or SYS-220GQ-TNAR
	Up to 160 (4C)	256	512GB	20TB	8	2 x AS -2124GQ-NART or 2x SYS-220GQ-TNAR

Conclusion

Supermicro systems with NVIDIA HGX A100 offer a flexible set of solutions to support NVIDIA vCS and NVIDIA A100 GPUs, enabling AI developments and delivery to run small and large AI models. Using the highest performing NVIDIA A100 GPUs, developers minimize their valuable time to run their AI models, delivering fast and cost effective AI features into new and existing products and services.

Supermicro offers these as integrated solutions, including systems, software, and support. Please call your Supermicro representative for more information.

References

1. NVIDIA Multi-Instance GPU and NVIDIA Virtual Compute Server Technical Brief
2. NVIDIA Virtual Compute Server for Red Hat Enterprise Linux with KVM Deployment Guide
3. NVIDIA Virtual Compute Server (VCS) Power the Most Compute-Intensive Workloads with Virtual GPUs
4. NVIDIA Virtual GPU Software Documentation
5. NVIDIA Certified Servers
6. Supermicro AS -4124GO-NART User's Manual
7. Supermicro AS -2124GQ-NART User's Manual