

DELIVERING AI TO ENTERPRISES, WITH SUPERMICRO SYSTEMS, NVIDIA AI ENTERPRISE, AND VMWARE

Supermicro Systems Supporting Scalable Virtualized AI Solutions



TABLE OF CONTENTS

Executive Summary	1
Artificial Intelligence Applicable Across Industries	2
Incorporating Artificial Intelligence into Enterprise Applications	2
Enterprise Challenges to Adopting AI	3
Easing into AI, Leveraging Existing Virtual Infrastructure	3
NVIDIA AI Enterprise Software Suite	4
AI Software Stack	5
Enterprise Support	6
Virtualized GPU Options	6
Mixing CPU Cores and GPU Cores	7
Supermicro NVIDIA-Certified Systems, also VMware Certified	7
Conclusion and References	8

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

Enterprises are increasingly using AI in their IT workflows. AI techniques have proven to be very effective in identifying patterns, finding defects in manufacturing, recognizing objects and images, generating effective recommendations, and many other functions. By incorporating these new AI techniques, enterprises can significantly improve their operational efficiency, grow their top lines, and optimize their bottom lines.

There are several challenges to incorporating these new technologies, including identifying the AI frameworks and tools, optimized systems, and implementing IT management software from development to deployment. NVIDIA AI Enterprise, an end-to-end cloud-native suite of AI and analytics software, addresses these challenges. NVIDIA AI Enterprise brings AI to existing VMware vSphere infrastructure that many enterprises use. Developers and IT administrators can incorporate AI training and inference in small steps. With included AI and data science tools and frameworks, enterprises can start with small investments and grow that capability over time. In addition, enterprises can deploy NVIDIA AI Enterprise software optimized and certified by NVIDIA to run on VMware vSphere with Supermicro NVIDIA-Certified Systems[™]. The servers are also VMware certified to support GPUs.

November 2021



As a leader in Green Computing, Supermicro is offering multiple choice of energy efficient server options, so customers can best optimize on power, cooling, space, and cost to fit their infrastructure while bringing AI to their enterprise.

Artificial Intelligence Applicable Across Industries

Since the advancement of GPUs to accelerate deep neural network processing and the fast-paced evolution of the technology in the last few years, deep learning artificial intelligence has evolved into multiple techniques used across every industry. The techniques include image and object recognition, text and speech understanding, finding correlations and anomalies in data. Here are some examples in select industries:

- In Finance, text and speech understanding can be applied to news and information to extract potential trading signals.
- In Healthcare, image processing is applied to find diseases in MRI, X-Ray, or other medical imaging techniques.
- In Retail, image processing can be used to build automated customer purchases in automated stores.
- In Manufacturing, object recognition can be applied to identify components and to find manufacturing defects.

The applications are pervasive, and most industries are beginning to take advantage of these artificial intelligence capabilities.

Incorporating Artificial Intelligence into Enterprise Applications

Within an industry, a corporation or enterprise can apply artificial intelligence in the various functions within the organization, for example:

- Marketing and sales departments use collaborative filtering to make cross product selling to customers.
- Human resources departments use text processing to identify talents.
- Manufacturing and services departments use computer vision techniques to identify defects in parts and components.

Each department within the enterprise can take advantage of AI techniques to make jobs more efficient and more cost effective, producing more reliable and accurate results.



Multiple departments can share the compute resources to maximize results and minimize infrastructure costs by leveraging virtualized data centers and NVIDIA AI Enterprise.



Enterprise Challenges to Adopting AI

For many enterprises, there could be significant challenges to adopting artificial intelligence. Since the recent Al advancements are mostly new and fast changing, enterprises have to find the people and know-how with domainspecific understanding to successfully integrate into AI their existing applications.

With NVIDIA AI Enterprise running on Supermicro NVIDIA-Certified Systems, enterprises can overcome these challenges to start integrating AI into enterprise workflows. Supermicro NVIDIA-Certified Systems and NVIDIA AI Enterprise

Challenges	Solutions	Additional Assistance
People	Data scientists	Best-in-class AI frameworks and tools included in the NVIDIA AI Enterprise suite
Knowhow	AI and domain knowledge	NVIDIA AI Enterprise Support Services
Consistent Tools	Containers with consistent and tested tools & libraries	Regular updates & support
Performance	NVIDIA Ampere-based GPUs and networking solutions	Supermicro NVIDIA-Certified Systems
Risk	NVIDIA AI Enterprise on existing VMware infrastructure	Supermicro NVIDIA-Certified Systems
Scaling Deployment	VMware scaling and future Kubernetes support	Supermicro NVIDIA-Certified Systems, networking

software together minimize risks because the entire solution is pretested through NVIDIA's rigorous certification process. In addition, the certified solutions assure performance and deployment scaling as the systems and software are designed to provide high performance and can be scaled using server clusters.

Supermicro servers are designed for high performance and cluster scaling, with built-in PCI-Express Gen-4 I/O to support high bandwidth, low latency networking devices such as the NVIDIA ConnectX-6 and ConnectX-7, fast storage such as the latest Gen4 NVMe disks, as well as the NVIDIA GPUs. As a result, data scientists and developers can get the most productivity using the Supermicro systems to run their AI development and deployment.

Easing into Artificial Intelligence, Leveraging Existing Virtual Infrastructure

With VMware vSphere, NVIDIA AI Enterprise makes it easy to start AI development in small steps for enterprises that have virtualized their IT infrastructure. An enterprise can start with a single server with a single GPU to run NVIDIA AI Enterprise. This AI Enterprise system operates as a member in the VMware vSphere environment. vCenter Server manages the AI Enterprise server as another member of the infrastructure, immediately enabling AI workloads for the enterprise.

Depending on the size of the AI model and batch data, an AI developer may



need a fraction of a GPU or full utilization of a GPU. This approach minimizes risk and maximizes the opportunity to create AI value. Over time, as the enterprise begins to realize the potential of AI enabled applications, the IT department in the enterprise can add servers running NVIDIA AI Enterprise to scale the applications.



NVIDIA AI Enterprise Software Suite

NVIDIA AI Enterprise is a software suite that runs in the VMware vSphere environment running on Supermicro NVIDIA-Certified Systems. The software suite includes AI tools and frameworks, cloud native deployment, and infrastructure optimization software to enable rapid AI development and deployment in VMware infrastructures.



NVIDIA AI Enterprise is optimized, certified, and supported on VMware to achieve near bare-metal performance with virtualization of AI workloads on NVIDIA-Certified Systems, of which Supermicro offers a wide myriad of optimal choices. By making it a low risk and simple approach to integrate AI into the existing enterprise virtualization environment, NVIDIA AI Enterprise enables an end-to-end software stack approach to start using AI in the enterprise. Enterprise developers can initially run small trials until they feel comfortable expanding to more extensive deployment. At that point, the solution is very scalable to deployment in multiple racks.

The design of NVIDIA AI Enterprise targets two personas:

- Data scientists and developers who use and incorporate AI, and
- IT administrators who need to ensure operational uptime of the systems

NVIDIA AI Enterprise provides the AI and data science tools and frameworks, which data scientists and developers need for their AI applications. By using Supermicro NVIDIA-Certified Systems, the IT administrators get assurance of reliable systems with high performance. These Supermicro systems are also VMware certified to support GPUs. Using VMware, the IT administrators are familiar with their existing infrastructure.



AI Software Stack

The NVIDIA AI Enterprise software suite offers the best-in-class tools and frameworks that AI practitioners and data scientists need, from data preparation (NVIDIA RAPIDS), training neural networks (TensorFlow and PyTorch), inference (NVIDIA TensorRT), to scaling inference operations (NVIDIA Triton Inference Server.) The deployment can start on one or more GPUs within a server. As the datasets grow, multi-node scaling can level up for many organizations.

Optimized for VMware vSphere environments, the suite delivers near bare-metal performance across multiple nodes, streamlining the production of new AI applications and services.





Enterprise Support

Using the NVIDIA AI Enterprise software suite, enterprise customers get enterprise-grade support for the entire system, from AI software to the virtualization and system hardware, including NVIDIA data center GPUs and network accelerators, all tested and optimized in the Supermicro systems. As a one-stop shop,



Supermicro offers the entire system, with VMware and NVIDIA AI Enterprise with support.

Virtualized GPU Options

NVIDIA AI Enterprise provides multiple virtualized GPU options. By picking the appropriate vGPU profile, each CPU in the system can be partitioned into virtual machines accordingly.

For the NVIDIA A30 Tensor Core GPUs and NVIDIA A100 Tensor Core GPUs, there is the choice of using Multi-Instance GPU (MIG) partitioning, which provides hardware isolation for each hardware partition relative to other partitions in a single A30 or A100 GPU.

Alternatively, the GPUs can be _ partitioned as time-slices using

	NVIDIA A100 MIG Mode Enabled	NVIDIA A100 MIG Mode Disabled	NVIDIA A30 MIG Mode Enabled	NVIDIA A30 MIG Mode Disabled	NVIDIA A40 (no MIG)
Max partitions	7	10 (A100/40GB) 20 (A100/80GB)	4	6 (A30/24GB)	32 (*VDI) (A40/48GB)
Partition Type	SPACE- SLICED	TIME-SLICED	SPACE- SLICED	TIME-SLICED	TIME-SLICED
Partition Sizes	Different sizes, as long as they add up to 1 GPU	All the same size per GPU	Different sizes, as long as they add up to 1 GPU	All the same size per GPU	All the same size per GPU
Largest vGPU	One A100	One A100	One A30	One A30	One A40
Compute resources	Dedicated	Shared	Dedicated	Shared	Shared
Virtualized 3D Graphics Support	No	No	No	No	Yes
NVIDIA [®] NVLink [®] Support	No	Yes (SXM4)	No	No (PCIe only)	No (PCIe only)
Heterogeneous Profiles	Yes	No	Yes	No	No

NVIDIA vGPU included in the NVIDIA AI Enterprise suite, which would allow users to invoke their use by sharing the GPUs with the VMware hypervisor allocating the GPU slices for use as needed by users in their respective virtual machines. With the NVIDIA A40 GPUs, customers can also use the GPUs for virtualized graphics applications. One popular use of the NVIDIA A40 is to deploy them for virtual desktop applications during the day, where users can run high-end 3D graphics applications in the centralized infrastructure. Then at night, while users no longer use the systems, the NVIDIA A40 GPU systems can be automatically



redeployed to run AI workloads, such as AI training or inference. Customers have a wide range of choices to optimize the use of GPUs.

Mixing CPU Cores and GPU Cores

NVIDIA AI Enterprise offers flexibility in how to partition GPU resources among the virtual machines. For example, some applications may need more CPU cores to do work, while GPU cores run compute-intensive operations, such as matrix multiplications for AI. The table shows some examples of the choices that can be

for

applications. The Supermicro

different

made

AI Use Case in a VM Examples	CPU Cores	GPU Cores / GPU Memory	GPU Type
Data Scientist Al Development	1-8	1 / 10 GPU to full GPU 1 / 20 GPU to full GPU	A100 (40GB) A100 (80GB)
AI Training / Inference using pre- trained model	1-8	1 / 7 GPU (MIG) to full GPU	A100 (40GB) A100 (80GB)
Data Scientist AI Development	1 - 8	1 / 6 GPU to full GPU	A30
AI / Inference Training using pre- trained model	1-8	1 / 4 GPU (MIG) to full GPU	A30
Mixed Workloads (Virtual Desktop by Day, Compute at Night)	1-8	1 / 32 GPU to full GPU	A40

systems provide wide options of CPU cores that could be matched to the system with the maximum number of supported GPUs.

Supermicro NVIDIA-Certified Systems, also VMware Certified

Supermicro offers a wide range of systems that are NVIDIA-Certified to support PCI-E Gen 4 based NVIDIA A30, A40, and A100, as well as the NVIDIA HGX-A100[™] 4-GPU systems. Choosing from these optimized Supermicro systems, customers can select the systems that provide the best performance while optimizing their energy usage and data center cooling.

Supermicro systems range from 1U to 4U support, supporting 1 to 8 GPU, depending on needs. NVIDIA network accelerators, e.g., NVIDIA[®] ConnectX[®]-7 SmartNIC are also supported. ConnectX-7 and NVIDIA[®] Blue-Field[®] DPUs (Data Processing Units) provide fast, low-latency network connectivity to the Supermicro systems. As a result, enterprises can select the most appropriate systems with needed power, cooling, and GPU support.

Supermicro Server	Rack Height	Depth	Max GPU	GPU Type	Network Acceleration	Disks	Dual CPU
SYS-120U-TNR/+	1U	29.1"	1	NVIDIA A100, A30, A40	CX6/7, BF2-DPU	12	3 rd Gen Intel® Xeon® Scalable
SYS-220U-TNR/+	2U	28.2"	2	NVIDIA A100, A30, A40	CX6/7, BF2-DPU	24	3 rd Gen Intel® Xeon® Scalable
SYS-120GQ-TNRT/+	1U	35.2"	4	NVIDIA A100, A30, A40	CX6/7, BF2-DPU	2	3 rd Gen Intel® Xeon® Scalable
SYS-220GP-TNR/+	2U	30.18"	4	NVIDIA A100, A30, A40	CX6/7, BF2-DPU	10	3 rd Gen Intel® Xeon® Scalable
AS -2124GQ-NART/+	2U	32.7"	4	NVIDIA HGX A100-4GPU	CX6/7, BF2-DPU	4	3 rd Gen AMD EPYC™
SYS-220GQ-TNAR/+	2U	32.7"	4	NVIDIA HGX A100-4GPU	CX6/7, BF2-CPU	4	3 rd Gen Intel® Xeon® Scalable
SYS-740GP-TNRT	4U/Tower	26.5"	4	NVIDIA A100, A30, A40	CX6/7, BF2-DPU	8	3 rd Gen Intel® Xeon® Scalable
AS-4124GS-TNR	4U	29"	8	NVIDIA A100, A30, A40	CX6/7, BF2-DPU	16	3 rd Gen AMD EPYC™



Conclusion

Supermicro NVIDIA-Certified Systems offer solutions flexible to support **NVIDIA** AI Enterprise in the VMware environment, enabling AI developments and delivery to run small and large AI models for AI training and inference. Using the highest performing NVIDIA A100, A30, and A40 GPUs, enterprise developers minimize their valuable time to run experiments and development on their models. AI



delivering fast and cost effective AI features into new and existing products and services. In addition, by selecting the appropriate Supermicro systems that provide the best performance, I/O, power, and cooling, from the many choices Supermicro offer, IT managers can deliver the best AI capable additions to their VMware infrastructure.

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

References

- 1. NVIDIA AI Enterprise
- 2. NVIDIA-Certified Systems
- 3. NVIDIA Virtual GPU Software Documentation
- 4. Supermicro GPU Servers supporting the Ampere architecture
- 5. Supermicro Ultra Servers

© 2021 Copyright Super Micro Computer, Inc. All rights reserved. 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 product names, logos, and brands are property of their respective owners.

