



TRANSFORM YOUR BUSINESS WITH THE NEXT GENERATION OF ACCELERATED COMPUTING

Supermicro servers optimized for NVIDIA A100 GPUs are solving the world's greatest HPC and AI challenges



Executive Summary

Escalating data complexities, coupled with the rise of digital intelligence is driving a significant paradigm shift across a diversity of vertical markets. Computing tasks are exploding in complexity, placing immense pressure on organizations to invest in superior systems that deliver massive power and flexibility. Implementing accelerated computing solutions from Supermicro and NVIDIA is crucial to become innovative and succeed in today's dynamic environments.

TABLE OF CONTENTS

Executive Summary	1
Enhancing Compute Capabilities to Power the Most Complex Workloads.....	2
Adopting Right-Sized Solutions for Every Environment	3
Transforming Business Performance with Unparalleled Solutions	3
Fueling the Next Generation of Business Innovation	9
Conclusion	9

SUPERMICRO

Supermicro is a global leader in high performance, green computing server technology and innovation. We provide our global customers with application-optimized servers and workstations customized with blade, storage, and GPU solutions. Our products offer proven reliability, superior design, and one of the industry's broadest array of product configurations, to fit all computational needs.

Enhancing Compute Capabilities to Power the Most Complex Workloads

Transformation is a key to capitalize on global trends such as the demand for high-performance computing (HPC) and the expansion of [artificial intelligence \(AI\)](#). The democratization of technologies like HPC, AI, machine learning, and deep learning, driven by the arrival of highly dense computing solutions, helps today's organizations overcome their largest and most complex problems with greater agility. These capabilities have gained popularity in multiple industries—including Life Sciences, Oil and Gas, Manufacturing, Retail, and Government—as organizations strive to unlock their data's full value and convert those insights into action.

With troves of information being generated, organizations are racing to boost their intelligence, solve critical business challenges, and drive competitive advantage. However, legacy technologies pose a mounting problem for organizations attempting to execute compute-intensive processes like HPC in the data center, high-performance analytics, and deep learning training and inference. A new breed of computing is essential to run these complex data pipelines.

Advancements in the HPC and AI space are causing a shift to accelerated computing, backed by GPUs' extreme processing abilities and demanding network infrastructures to gain industry-leading performance. These breakthrough capabilities provide high efficiency and capacity to optimize any workload. As a result, more businesses than ever are looking to deploy GPU-accelerated platforms to operate faster and more intelligently.

As industries recognize that advanced technologies are necessary to support the upsurge of HPC and AI usage, the next generation of accelerated computing has emerged, offering the right balance of compute and performance. Advanced GPU-based systems, such as [Supermicro's versatile 4U system with NVIDIA HGX 8-GPU](#) are changing the game for modern enterprises, delivering unprecedented speed and productivity to wrangle insight from large datasets while reducing time-to-completion.

The introduction of cutting-edge compute acceleration hardware along with state-of-the-art high-speed network fabrics is allowing businesses to redefine how they operate in order to tackle some of their biggest concerns head-on:

- **Faster data movement** – Greater memory and bandwidth to move massive volumes of data being processed on one server and communicate the data between servers.
- **Infrastructure optimization** – Easily adapt to changing requirements as workloads increase in size, scope, and complexity.
- **Performance at scale** – High throughput and low latency to ensure peak levels of efficiency and improve value.
- **Technology standardization** – Migration from isolated servers to centralized IT frameworks as AI becomes more mainstream, in addition to the use of universal accelerators to fuel more diverse workloads.

Numerous industries are realizing the undeniable value of accelerated computing, as organizations transform their technology infrastructures to pursue AI innovation with the utmost confidence.

In the Manufacturing space, accelerated computing is dramatically enhancing research and development (R&D) efforts. Organizations rely on robust infrastructure for techniques such as modeling and simulation to streamline the design and testing phases of R&D and deliver high-quality products at a faster rate of completion. State-of-the-art manufacturers are implementing robotics to automate assembly line processes, increase operational productivity and output, and run entire smart factories in large scale deployments. Additionally, critical HPC and AI insights are being applied to capabilities like industrial inspection and predictive maintenance to ensure safety and compliance.

In the Retail space, organizations are leveraging GPU technologies to capture a deluge of inventory, customer, and transactional data. Smart retailers have the ability to analyze endless streams of data, which facilitates

competitive advantages such as demand forecasting, personalized recommendations and tailored advertising, and inventory logistics. Accelerated computing enables retailers to optimize complex analytics workloads, delivering actionable insights to hone operations from warehouses to physical stores, online platforms, and applications.

In the Healthcare and Life Sciences space, organizations are progressively adopting new forms of compute to fuel scientific research and discovery and heighten patient care. Laboratories, research departments, and care facilities must invest in ongoing innovations to optimize their work, deriving precise results in real- or near-time for various uses. For example, accelerated computing platforms enable hospitals to support multiple departments running different use cases simultaneously—radiology can conduct image analysis, while a separate department uses the same platform to rapidly analyze health histories and create highly specialized treatment plans.

Adopting Right-Sized Solutions for Every Environment

Right now, there is a massive demand for accelerated computing technologies that can be tailored to fit specific use cases and industry requirements. Organizations need a new class of solutions that offers greater resilience and elasticity, from data center to edge to cloud.

Supermicro and NVIDIA help industries build infrastructure to carry out today's workloads and scale for tomorrow's challenges. With [Supermicro's adaptable, fastest-to-market servers](#) powered by groundbreaking [NVIDIA A100™ Tensor Core GPU technology](#), organizations can satisfy their specific vertical needs with ease. In addition to business-centric applications for enterprise, these solutions provide the right building blocks for diverse tasks—from running inference on developed models to HPC, to high-end training requests:

- **AI infrastructure** – Bringing AI to scale from data center to edge, aiding AI adoption with the ideal IT infrastructure for breakthrough innovation.

- **Deep learning applications** – Accelerating AI training and inference workloads for fastest time-to-market, supporting the end-to-end AI lifecycle from experimentation and prototyping model training to deployment in production.
- **HPC applications** – Extending next-generation HPC in the data center, industrializing AI workflows with proven architecture for competitiveness, responsiveness, and efficiency.
- **High-performance analytics** – Turning massive data sets into insights, utilizing pre-built, pre-optimized AI tools to solve bigger and more complex data science problems faster.
- **Enterprise-ready utilization** – Expanding the reach of accelerated computing, running simultaneous mixed workloads to maximize flexibility and ROI by adopting right-sized GPUs with guaranteed quality of service for every job.

Together, Supermicro and NVIDIA accelerate organizations to modernize their infrastructure. Our first-to-market hardware innovations combined with leading-edge software which can be modified and configured as organizations evolve, along with standardized tools that simplify deployment and management by centralizing IT, regardless of the solution. Supermicro's broadest portfolio of GPU systems empower organizations to choose the server that matches their desired workloads, density, I/O, and cost needs, backed by a customizable number of NVIDIA GPUs cores. These solutions are turnkey, so organizations can implement a single platform to put servers into operation immediately.

Transforming Business Performance with Unparalleled Solutions

Optimized and Fully Validated Hardware with NVIDIA GPUs

In enterprise computing, storage, networking, and green technology, Supermicro leads the market with the most extensive portfolio of [high-performance hardware](#). They offer the broadest selection of innovations for fully application-optimized servers, storage, embedded systems, and workstations.

Supermicro quickly leverages cutting-edge solutions to pioneer a new era of computing based on the new [NVIDIA A100™ Tensor Core GPUs](#). The broad, flexible and robust system portfolio employs the latest GPU acceleration technology stack and delivers peak levels of performance, scalability, and serviceability to facilitate a wide range of HPC and AI initiatives. Coupled with high core counts and PCI-E Gen4 lanes that dual AMD EPYC processors provide, each platform is designed for performance and cost-effectiveness that modern IT environments require to execute diverse workloads, from studying molecular behavior for drug discovery to honing financial models for insurance approvals.

As the engine of the NVIDIA data center platform, A100 GPUs on a right platform enable unprecedented agility to reduce time-to-insight and time-to-market. Supermicro has integrated this universal accelerator into [fully validated GPU systems](#), developing a comprehensive architecture that can accommodate increasingly complex tasks.

Supermicro systems powered by the A100 can leverage the blazing network speeds of [NVIDIA Mellanox ConnectX® SmartNICs leveraging PCI-E Gen4 lanes with AMD EPYC processors](#) to [quickly scale to thousands of GPUs](#) or, using new multi-instance GPU (MIG) technology, can be partitioned into seven isolated GPU instances to run different jobs. NVIDIA A100 third-generation Tensor Cores with Tensor Float provides up to [20X faster speeds](#) over the previous generation without requiring any code changes, plus up to 6X higher out-of-the-box performance for AI training (Figure 1). Large AI models like BERT can [be trained in just 37 minutes](#) on a cluster of 1,024 A100s. For inference workloads, A100 servers offer up to [7X higher performance with MIG](#) (Figure 2).

Figure 1: BERT Training

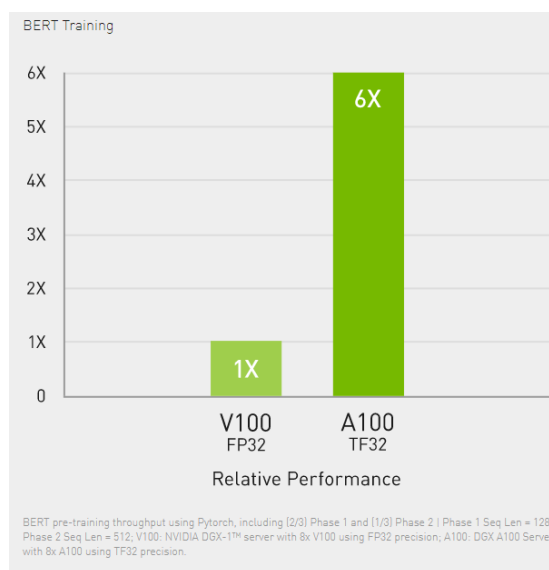
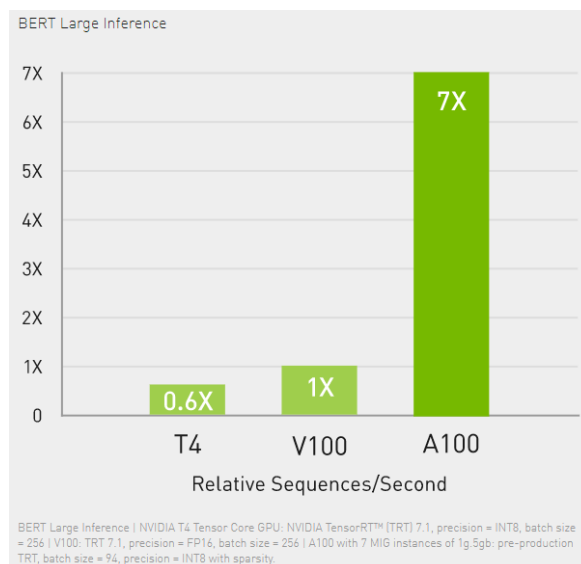
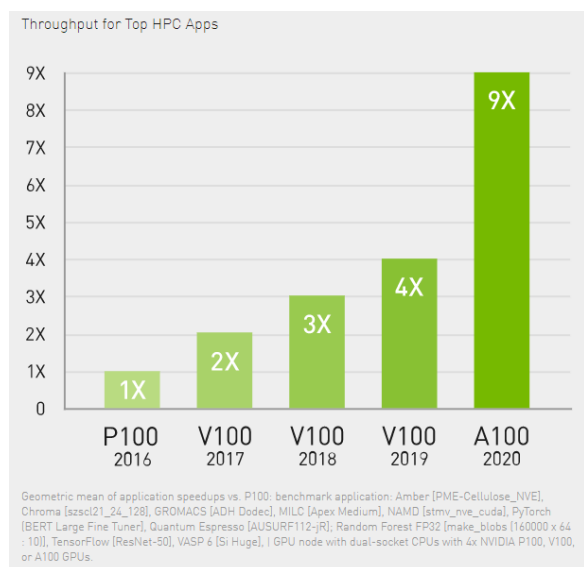


Figure 2: BERT Large Inference



The A100 also delivers [9X higher HPC performance](#) than comparable GPU technologies in 2016 (Figure 3). Now, organizations can harness greater throughput for compute-intensive applications to drastically reduce the rate of completion.

Figure 3: Throughput for Top HPC Apps



Density Optimization

Supermicro's A100 servers are the most inclusive accelerated computing solutions available, ensuring that customers can maintain high processing power and memory bandwidth without incurring overwhelming costs. Our GPU systems solve major pain points that businesses face as they attempt to innovate and scale, by providing flexible, density-optimized infrastructure. Supermicro's latest generation of GPU systems powered by [AMD EPYC™ 7002 series processors](#) with up to 128 PCIe Gen4 lanes per socket leverage outstanding core density and unparalleled I/O capacity to maintain operational consistency. AMD EPYC processors deliver maximum performance to propel modern data center workloads, so customers can manage deployments and escalate computing demands easily and cost-effectively.

Backed by this powerful hardware, A100 servers can scale from 8 to 64 cores (16 to 128 threads per socket)—greater than any other x86 vendors in the industry—to increase density and reduce capital, power, and cooling expenses.

NGC Certified Solutions

The A100 GPU is part of the complete NVIDIA data center solution that incorporates hardware, networking, software, libraries, optimized AI models, and applications from [NGC](#), all built on Supermicro servers. These data center platforms simplify the deployment of AI systems, allowing organizations to deliver real-world results and deploy solutions into production.

NGC includes pre-packaged toolkits, pre-trained models, frameworks with latest updates, and other essential tools to streamline HPC and AI projects—such as optimizing models and launching them into production quickly. Hardware solutions require sufficient software support to add productivity and simplify IT deployment and management. NGC-ready systems from Supermicro give customers a strategic advantage to rapidly transform and tackle their most demanding workloads.

NGC supports virtualized environments with [NVIDIA Virtual Compute Server \(vCS\)](#), which enables data centers to accelerate server virtualization with the latest NVIDIA data center GPUs. By migrating compute-intensive workloads—including AI, deep learning, and data science—to virtual machines, businesses can share physical GPU resources as well as assessing different instances for GPU usage. This capability provides extreme elasticity for customers to take advantage of A100 acceleration across geographic locations.

Advanced Networking/Storage

Supermicro systems combine [NVIDIA Mellanox ConnectX® SmartNICs along with NVIDIA Mellanox Quantum and Spectrum switches](#) with GPUs to run complex workloads more securely from a variety of locations, with the flexibility to scale out to multiple servers. The network adapters provide outstanding speed and functionality previously unseen on the market, leveraging the [best TCO](#)

[for 25G+ deployments](#) in data centers and the cloud. SmartNIC is the ideal blend of hardware and programmable acceleration to enable secure, high-performance communication in Ethernet and InfiniBand environments. This includes expediting and virtualizing access to storage, such as NVMe supported natively on A100 servers to allow the fastest response times for all types of enterprise workloads.

[NVIDIA® NVLink™ and NVIDIA NVSwitch™, the high-speed multi-GPU interconnection](#), enhance reliability and provide higher bandwidth, more links, and improved scalability for multi-GPU system configurations to ensure seamless communication between components. NVLink moves large volumes of data at breakneck speeds, whether processed on one server or shared between multiple GPUs on the server. A single A100 GPU supports up to 12 third-generation NVLink connections for a total bandwidth of [600 gigabytes per second](#), the greatest throughput in the industry to date. With the latest version of NVLink and NVSwitch technologies, Supermicro servers can deliver up to [5 petaflops of AI performance](#) in a dense, single 4U system.

[Supermicro® Advanced I/O Module \(AIOM\)](#) further enhances multi-GPU communication. This cost-effective solution delivers energy-efficient ethernet to expand network connectivity. The current AIOM form factor is designed with a variety of networking options and I/O enhancing features for server, data center, and edge platforms.

Thermal Design/Liquid Cooling

To drive further efficiency, Supermicro offers multi-GPU optimized thermal designs that provide groundbreaking density in terms of pricing, space, and capacity per rack. With 1U, 2U, and 4U rackmount A100 servers, customers can select the ideal platform to support their HPC and AI applications. Systems with smaller form factors feature heavy-duty PWM fans with optimal speed control to maintain operating ambient temperature. For larger form factors, Supermicro liquid cooling solutions can [reduce TCO by as much as 40%–50%](#) when compared to forced-air cooling. Liquid cooling is essential for organizations

utilizing higher rack density to capitalize on the rise of AI, where GPUs often run hotter than traditional CPUs. Additionally, because liquid cooling is exceptionally efficient, data centers can increase their processing per square foot beyond what is possible with forced-air cooling, while improving their power usage effectiveness.

Titanium Power Supply Standard

With the added sustainability of [Supermicro's Titanium Level Power Supply \(PS\)](#), businesses can "Keep It Green" to realize even more significant cost savings. The Titanium PS Standard supplies ample power for GPU systems of all sizes—from server to storage components—with the [industry's highest efficiency rating of 96%](#). This solution includes redundancy for power-hungry GPU systems on multiple platforms. The ability to use a single PS saves exorbitant amounts of time and resources on developing additional sources, meeting safety regulations, and remaining compliant.

Rack Integration

By integrating dense compute power and optimal storage and networking options, Supermicro leads the industry in accelerated computing performance. Supermicro's rack solution connects power and storage to guarantee the highest throughput for data transfer. This design is a key differentiator for customers looking to integrate accelerated computing components to fully optimize AI and deep learning workloads. The chassis incorporates shared hardware with a common PS in a standalone system, which mitigates the amount of space required while dramatically boosting efficiencies. This is a huge advantage for organizations that want to deploy multiple servers without increasing rack space.






Customer-Centric Solutions

Supermicro is committed to developing [total computing innovations](#) that are customer-centric. Supermicro's unmatched solutions portfolio with the addition of cutting-edge NVIDIA GPUs and networking technology is changing the game for today's businesses. Now, customers have a one-stop-shop where they can select

and build the ideal platform to fit their requirements. As accelerated computing continues to transform data centers, Supermicro will provide the very latest system advancements to optimize A100 utilization at every scale. These new systems will significantly boost performance across HPC, data analytics, AI training, and inference.

A100 GPUs power a wide range of Supermicro systems, including the groundbreaking [NVIDIA HGX A100](#), creating the world's most powerful accelerated server platforms for AI featured in the [Supermicro NVIDIA HGX A100 Delta Platform](#) and [Supermicro NVIDIA HGX A100 Redstone Platform](#). Additionally, the [newly unveiled PCIe A100 form factor](#) offers high versatility across multiple configurations, including the [Supermicro NVIDIA A100 PCIe AMD Platform](#), [Supermicro NVIDIA A100 Ultra AMD Platform](#), and [Supermicro NVIDIA A100 Blade AMD Platform](#) to expedite data-heavy workloads.

SUPERMICRO NVIDIA A100 SYSTEMS

Model #	AS -2124GQ-NART	AS -4124GO-NART	AS -4124GS-TNR	AS -2024US-TRT	SBA-4119SG
System Model					
System Platform	NVIDIA HGX A100 4x GPU Redstone Platform	NVIDIA A100 8x GPU Delta Platform	NVIDIA A100 8x GPU PCIe AMD Platform	NVIDIA A100 2x GPU PCIe Ultra AMD Platform	NVIDIA A100 1x GPU PCIe per node Blade AMD Platform
Chassis	2U Rackmount	4U Rackmount	4U Rackmount	2U Rackmount	8U Rackmount (up to 20 nodes with 20 NVIDIA A100 PCIe GPU)
Processors	Dual AMD EPYC™ 7002 Series Processors, Up to 280W TDP	Dual AMD EPYC™ 7002 Series Processors, Up to 280W TDP	Dual AMD EPYC™ 7002 Series Processors, AMD Socket SP3, Up to 280W TDP	Dual AMD EPYC™ 7002 Series Processors, AMD Socket SP3, Up to 280W TDP	AMD EPYC™ 7002 Series Processors, AMD Socket SP3 225W/240W/280W @ 35C 200W/170W/155W/120W @ 42C
GPU	4x NVIDIA HGX A100 GPU 40GB	8x NVIDIA HGX A100 GPU 40GB	8x NVIDIA A100 GPU 40GB	2x NVIDIA A100 GPU 40GB	1x NVIDIA A100 GPU 40GB
Memory	Up to 32 DIMM Slots Up to 8TB DDR4 3200 MHz DIMMS	Up to 32 DIMM Slots Up to 8TB DDR4 3200 MHz DIMMS	32 DIMM Slots Up to 8TB 3DS ECC DDR4, 3200 MHz RDIMM/ LRDIMM	32 DIMM Slots Up to 8TB ECC DDR4, 3200 MHz SDRAM 8-Channel Memory Bus	8x 288-PW DDR4 DIMM Slot 1 DPC with Up to 2TB DDR4 3200 MHz ECC RDIMM
Drives	4 Hot-Swap 2.5 Drives Bay (SAS/SATA/NVMe Hybrid)	6 NVMe U.2 (4 from PCIe Switch & 2 from CPU) 2 NVMe M.2	Hot-Swap 2.5" Up to 24 x 2.5" SAS/SATA Drive Bays 4x 2.5" SATA Supported Natively 4x 2.5" NVMe Supported Natively	Hot-Swap 3.5" Drive Bays: 12 SATA3 by Default or 8" SATA3 + 4 NVMe via Optional Kit or 12 SAS3 via Optional SAS Kit	1 PCIe Gen 4 NVMe/ SATA M.2
I/O	4 PCIe 4.0 x 16 1 PCIe 4.0 x 8	8 PCIe x 16LP from PCIe Switch 2 PCIe x 16LP from CPUs AIOM Support	9 PCIe 4.0 x 16 (FHFL) Slots or 10 PCIe 4.0 x 16 (FHFL) Slots without NVMe Devices	2 PCIe 4.0 x 16 (FH, 10.5" L) Slots 1 PCIe 4.0 x 16 (FH, 9.5" L) Slot 1 PCIe 4.0 x 16 (LP) Slot 1 PCIe 4.0 x 8 (FH, 9.5" L, in x16) Slot 1 PCIe 4.0 x 8 (Internal LP in x16) Slot	1x PCIe 4.0 x 16 Slot for Optional Network Mezzanine, Up to 2x PCIe 16 Support, 1 Double-Width or 2 Single width FHFL GPU
Cooling Fans	4x Removable Heavy-Duty Fans	4x Removable Heavy-Duty Fans	8x 11.5K RPM Heavy-Duty Fans	4x Heavy-Duty PWM Fans with Optimal Fan Speed Control	Share with 16 Heavy-Duty System Fan in Chassis
Power Supply	2x 2200W Titanium Level PWS with 2x 3000W Redundant Upgrade Options	4x 2200W Titanium Level PWS with 4x 3000W Redundant Upgrade Options	2+2 2000W Titanium Level PWS Redundant Power Supplies	1600W Titanium Level PWS Redundant Power Supplies	Up to 8x Hot-Swap High-Efficiency Titanium 2000W Redundant Power Supplies

Fueling the Next Generation of Business Innovation

Supermicro and NVIDIA are delivering the next generation of accelerated computing to transform your business. As a global leader in high-performance, high-efficiency server technology, Supermicro provides the most comprehensive building blocks for a new breed of computing solutions, from data center to edge to the cloud. Leveraging NVIDIA's end-to-end technologies from AI to networking and full-stack offerings, our solutions combine first-to-market innovations with the massively parallel processing power of NVIDIA GPUs, so our customers can achieve unmatched performance to solve the world's most challenging problems.

Together, Supermicro and NVIDIA empower customers with application-optimized solutions to execute demanding workloads with extreme speed, reliability, sustainability, and cost savings. These robust solutions are equipping businesses for rapid growth, paving the way for future innovation.

Conclusion

As the expansion of HPC and AI poses mounting challenges to IT environments, Supermicro and NVIDIA are equipping organizations for success, with world-class solutions to empower business transformation. We are continually testing and validating advanced hardware featuring optimized software components to support a rising number of use cases.

NVIDIA compute and networking acceleration with Supermicro's extensive portfolio of hardware solutions lets organizations to handle the most demanding workloads with high flexibility, scalability, and sustainability. We are committed to helping organizations achieve breakthroughs in HPC and AI innovation. With engineering resources around the world, customers can communicate with experts to quickly resolve problems, discover new optimizations, and evolve.

This is your opportunity to transform. [Visit us online](#) to learn how Supermicro's NVIDIA A100 solutions unlock revolutionary performance.



AS-4124GO-NART with HGX A100 8-GPU



AS-2124GQ-NART with HGX A100 4-GPU



AS-4124GS-TNR with 8 A100 PCIe GPUs