SUPERMICRO | NVIDIA

Smart AI Combos: Hardware and infrastructure for top performance

An introduction to optimizing flexibility and value

Executive Summary

Artificial Intelligence (AI) is booming. Advances in hardware and software are enabling AI to transform industries worldwide, from financial services, to manufacturing, healthcare, and many others. To move to the next phase, organizations must now optimize AI performance, scalability, and demonstrate a clear return on investment. Developing a consistent platform and infrastructure strategy for deploying and expanding AI is a key step for cost-effective growth.

Whether it's for certified AI platforms and rack-based solutions, for servers, and clusters, motherboards and other building blocks, or for reference architectures, adopting an open, modular, best-of-breed approach helps enterprises choose the best combinations of processors, connectors, and form factors to create flexible, powerful foundations for AI-driven transformation.

This white paper will familiarize readers with what's needed: from the basics to new technologies, and resources for selecting flexible, "future proof" AI hardware, platforms, and infrastructure. It will help technology and business buyers make the best choices for maximum performance and efficiency, accelerating time to value and unlocking the power of AI today and tomorrow.

Introduction

Optimizing AI outcomes and value requires optimized infrastructure

From public and private clouds to the edge and embedded systems, AI applications are proliferating everywhere: in drug design, product optimization smart assistants, self-driving cars, robo-advisors, conversational bots, email spam filters, assembly lines, part inspections — the list grows longer daily. Many organizations are now working hard to advance successful early efforts into wider production and to demonstrate the value of AI. Others are advancing proof-of-concepts and pilots.

IT and business leaders tasked with cost-effective AI growth must successfully manage two kinds of optimization:

SUCCESS FACTOR 1

Optimizing *systems* for specific AI workloads

In AI, one type or size of a system does not fit all. AI applications have different requirements, response times, batch throughput requirements, continuous streaming, different use cases of different models, architectures, frameworks, platforms, and operating environments. Machine learning models, for example, require loading, transforming, and processing extremely large datasets to glean critical insights. Deep learning demands a system with large amounts of memory, massive computing power, and fast interconnects for scalability. Inferencing, computer vision, automated assistants, and other AI workloads each have their own stringent, vastly different requirement for number of CPU cores, and GPU capabilities.

Creating systems and environments best suited to specific AI applications requires careful choosing of many elements, including processor capabilities, memory capacities, storage, networking, and a variety of software.

SUCCESS FACTOR 2

Optimizing AI *infrastructure* to optimize AI outcomes

AI applications and systems rarely exist in a vacuum. AI workloads depend heavily on infrastructure — including existing infrastructure — to deliver optimum results and time to value. The reason is simple: To drive AI, data centers must handle massive amounts of data, both structured and unstructured. Doing so requires infrastructure with high compute power, fast memory access, generous storage capacity, energy efficiency, and scalability. That, in turn, demands powerful hardware including GPUs, flash memory, and high-bandwidth networks. The right platforms, frameworks, data sets, and even pre-trained AI models are also key. Poorly optimized infrastructures can lead to suboptimal outcomes, wasted expenses, and outright failure.

Key Requirements

Optimizing AI workloads and infrastructure boils down to two key challenges for IT and business leaders. Any products or solutions must support these critical requirements:

Managing price/performance

As we've seen, particulars differ but a high-performance computing environment is critical for all types of AI. At the same time, no organization has unlimited budget for simply ripping and replacing existing infrastructure with new products. So, infrastructure cost management is a crucial discipline in AI, especially in scaling proof-of-concepts into enterprise deployments.[1]

> Even more than in other IT efforts, efficient use of resources is essential for balancing AI availability, manageability, and affordability.

Simplifying complexity and integration

Different AI workloads have vastly different requirements for hardware, software, interconnectivity, networking, and data. Choosing the best combinations of these elements is a huge and complex task. So is creating an end-to-end AI solution from disparate products and vendors. Things get even tougher when new systems must be integrated with existing infrastructures, including a large installed base of different CPUs and GPUs, each with different capabilities and performance characteristics. Little wonder that Gartner calls legacy integration a top barrier to AI implementation, with just 53% of AI projects making it from pilot to production due to challenges here.[1]

> The takeaway: Factor in ease and cost of integration from the start of the selection process.

The Solution

Open, modular, best-of-breed hardware optimizes AI performance, cost, and flexibility

A wide range of fast-changing use cases and environments require a "future-ready" platform approach. Whether enterprises are shopping for a server, cluster, building blocks or a certified solution, to meet the requirements of good price/performance and easy integration, look for systems and infrastructure choices that offer:

Modularity and interoperability

These key pillars let organizations pursue a vendor-agnostic approach to AI by using products and services from leading vendors. The good news here: A huge portfolio of products and building blocks offer modular, optimized, and certified platforms built on best-of-breed hardware and software (more on this in "Putting It Together", below). Enterprises can start small with products that provide multiple expansion slots or capabilities in the same footprint or product line.

[1] Gartner "P-19019 AI in Organisations", Claudia Ramos, Erick Brethenoux, 2020

Openness

The secret sauce of interoperability is openness. More specifically, open choices help avoid lock-in by proprietary systems and provide flexibility for the future. What happens if a GPU or CPU needs more horsepower but can’t be upgraded? What happens when enterprises choose an architecture optimized for a narrow range of workloads instead of capabilities that can work with a broad range of algorithms? New, open approaches using the same API let enterprise go from a lower-end processor to a high-end processor, within and across systems.

Ultimately, selecting systems that offer modularity and openness lets organizations sidestep complexity, integration headaches, and runaway infrastructure costs. Importantly, this approach eliminates the need to rip and replace; hardware can be added and upgraded easily and brought into the defined environment in the data center. This flexibility provides a hugely valuable “future proofing” for adding additional processors that can take advantage of the latest algorithms, including demanding new workloads like HPC or more advanced analytics.

Flexibility is a huge, non-negotiable requirement.

Key Hardware Components

Now that enterprise understands its importance and what's needed for optimized AI systems and infrastructure, let's take a brief look at the hardware. Whether deployed in an on-premise data center, in a public or private cloud, or building or buying a flexible solution, carefully matching best-of-breed components is crucial for delivering the best balance of performance and cost-effectiveness for advancing an organization's unique AI goals.

CPUs and general-purpose processor

The computing layer is the heart of power-hungry AI. Modern, performance-optimized x86-based central processing units (CPUs) offer high core counts and myriad matched options for memory, I/O, data persistence, and networking. General-purpose compute resources can provide an economical option for some small AI/ML models and inferencing. Benefits include reduced initial cost, lower OpEx and power consumption, and easy upgrades.

GPUs and special-purpose processors

Graphic processing units (GPU), field-programmable gate arrays (FPGAs), and application-specific integrated circuits (ASICs) all power a wide range of AI workloads. Each has its strengths and drawbacks: ASICs are the least expensive on a per-unit basis, FPGAs the most. Both are costly in resources and time to program. For most data centers, GPUs are the best choice for AI servers. GPUs offer excellent performance and ease of programming, saving valuable development resources.

GPU-accelerated data centers deliver breakthrough performance with fewer servers and less power, resulting in faster insights with dramatically lower costs. Their highly parallel structure is especially suited to AI model training as part of an AI pipeline. New models, featuring the latest NVIDIA Ampere architecture-based GPUs, are optimized for AI/Deep learning as well as for HPC, 5G, and data analytics.

Interconnectivity

Lightning-fast processors are of little use if data cannot swiftly travel between other processors, storage, networks, and other key components. So, pay careful attention to the interconnectivity options for AI systems.

Internal connections

High-speed networking plays an integral role in scaling application performance across the entire data center — the new unit of computing for AI and HPC. NVIDIA is paving the way with software-defined networking, In-Network Computing acceleration, remote direct-memory access (RDMA), and the fastest speeds and feeds.

PCI-E 4.0 and 5.0

These newer PCI-E standards provide a higher-bandwidth connection to GPUs, SSDs, and other peripherals. 3rd Gen Intel Xeon Scalable processors and 3rd Gen AMD EPYC processors support PCI-E 4.0, while future Intel Xeon and AMD EPYC processors will support PCI-E 5.0. All generations of PCI-E are backward compatible, so there's no reason not to upgrade.

NVIDIA HGX

NVIDIA HGX combines extremely fast interconnections, at 600 GB/s using NVLink® between multiple GPUs, in a single server platform for the highest possible performance. While PCI-E-based NVIDIA GPUs offer more versatility for adding GPUs to existing servers, communication between GPUs is slower than with HGX, since shared data must pass through PCI-E to the CPU before being consumed by a different GPU.

Below is a summary of the Supermicro GPU servers with details about the form factor, numbers of CPUs, GPUs, and internal data paths:

Supermicro GPU servers

System	Height, Nodes	#CPUs/Node	Max # GPUs	PCI-E or HGX
SYS-420GP-TNAR(+)	4U,1	2 (Intel)	8	HGX
A+ 4124GO-NART(+)	4U,1	2 (AMD)	8	HGX
SYS-420GP-TNR	4U,1	2 (Intel)	10 (dual root)	PCI-E, NVLink Bridge
A+ 4124GS-TNR	4U,1	2 (AMD)	8 (dual root)	PCI-E, NVLink Bridge
A+ 2124GQ-NART-LCC	2U,1	2 (AMD)	4	HGX + Liquid Cooling
SYS-220GQ-TNAR(+)	2U,1	2 (Intel)	4	HGX
A+ 2124GQ-NART(+)	2U,1	2 (AMD)	4	HGX
SYS-220GP-TNR	2U,1	2 (Intel)	6	PCI-E
SYS-210GP-DNR	2U,2	1 (Intel)	3 per node	PCI-E
A+ 2114GT-DNR	2U,2	2 (AMD)	3 per node	PCI-E
SYS-120GQ-TNRT	1U,1	2 (Intel)	4	PCI-E

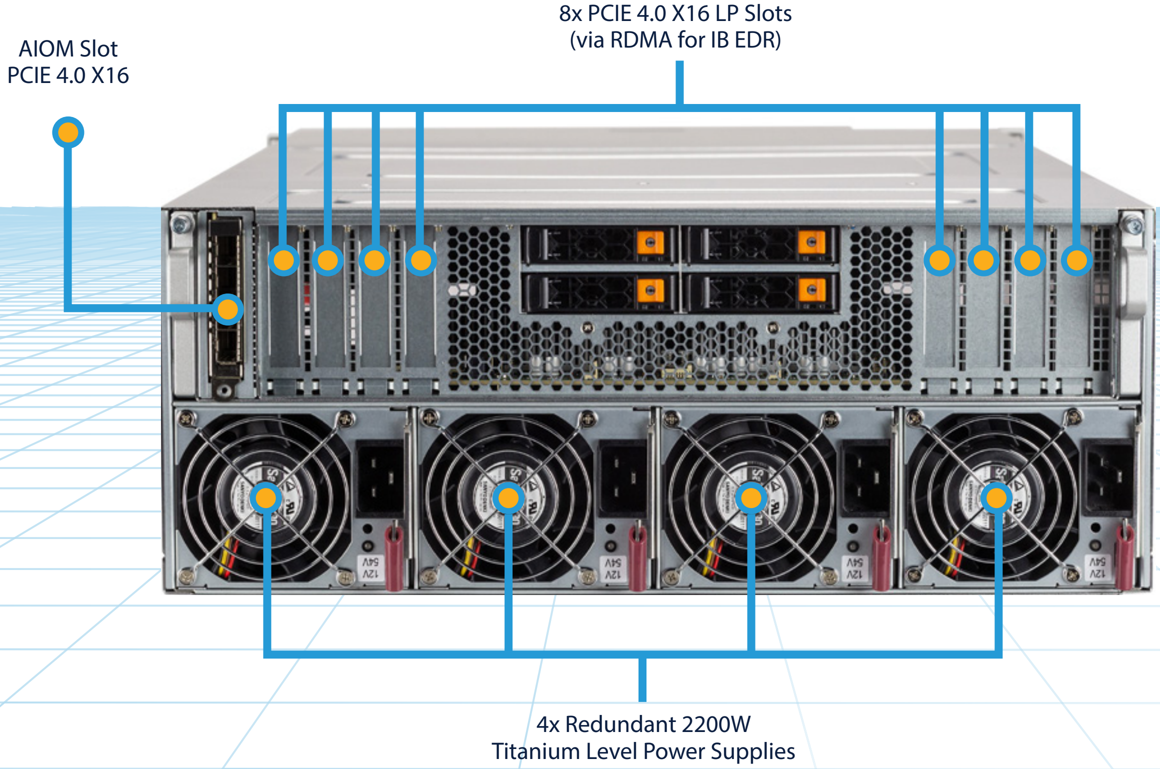
Note: Supermicro X12 and H12 Generations Only

External connections

Smart adapters and switches reduce latency, increase efficiency, enhance security, and simplify data center automation to accelerate end-to-end application performance. The industry-leading NVIDIA® ConnectX® family of smart network interface cards (SmartNICs) offer advanced hardware offloads and accelerations. NVIDIA Ethernet adapters enable the highest ROI and lowest Total Cost of Ownership for hyperscale, public and private clouds, storage, machine learning, AI, big data, and telco platforms. ConnectX SmartNICs provide an extremely accurate time-synchronization service for data center applications and underlying infrastructure.

A Supermicro SYS-420GP-TNAR System with many I/O connections labeled.

(Rear View – System)



Scalability

For many organizations, scalability is THE biggest AI challenge. Success depends heavily on having the right enabling technologies. Among them:

- optimal PCI-E topology within a server
- optimal networking across servers, using protocols such as RDMA
- algorithms that can scale out using these protocols
- management frameworks that make it easy to deploy and manage workloads at various scales

It's a full-stack problem and solution.

Other aspects of scalability:

- ability to use the same GPU for different scale jobs, from the use of multiple fractional instances of a GPU (Multi-Instance GPU) for inferencing to using multiple GPUs within one host or even across multiple hosts for large-scale data analytics or DL training
- ability to add systems in the future with the assurance that the architecture can scale and incorporate these new systems without bottlenecking

Scaling GPU servers to multiple racks for extremely large AI applications can be easily accomplished.

Putting It Together

Service? Custom? Platform? Solution?

The final task is understanding and selecting the best way to choose and deploy the most flexible, cost-effective, and optimized systems and infrastructure that will best advance your organization's AI goals and needs.

Cloud services

Some organizations choose these pay-as-enterprise-go offerings for their initial foray into AI. However, many soon discover that this approach can quickly get expensive and reduce efficiency. Plus, you're basically stuck with the vendor's choice of hardware and software.

Custom systems

Building or buying "one-off" AI solutions can make sense in some circumstances, such as for a specific workload or use case – or for enterprises looking to experiment, deploy quickly, and plan only limited adoption of AI. Drawbacks include multiple AI siloed and duplicated infrastructure and spending.

AI infrastructure

For most organizations, building a broader AI architecture and infrastructure for an AI platform is a smarter choice. That's especially true for enterprises that are more experienced with AI and plan to deploy widely. Taking this route promises to deliver greater long-term business value and increase ROI as capabilities build.

This approach makes most sense for organizations with:

- large and wide scope of use cases
- high-level strategic commitment to AI
- need for long-term capacity planning
- availability of in-house infrastructure to build upon
- expertise and experience with AI
- requirement to use the latest AI technology as soon as available
- custom servers (CPU, Memory, GPU, Storage, Network) requirements

Certified solutions

Seeking to reduce the huge complexity for organizations assembling their own AI systems, leading vendors have introduced highly flexible turn-key solutions. With the continued rollout of advanced applications and workloads, customers require manageable, secure, and scalable servers for their data centers.Supermicro's lineup of high-performance servers supporting NVIDIA GPUs and data processing units (DPUs) includes a growing number of NVIDIA-Certified Systems(™), with many more currently undergoing the certification process. Each server/GPU configuration earns its own certification.

Validated and pre-certified systems bring many benefits:

- faster to implement if working with known vendors with lots of knowledge
- wide range of CPU and GPU options for on-prem installations
- consistent APIs from the edge to the core
- on-prem data centers and private clouds can take advantage of the latest hardware and software immediately. Organizations don't have to wait for a public cloud to make new instances available.
- customizable for specific AI workload requirements
- flexible enough to manage changing workloads at no or low cost
- fast updating of the latest software components on the organization's maintenance schedule
- lower environmental impact and e-waste because specific components can be re-used

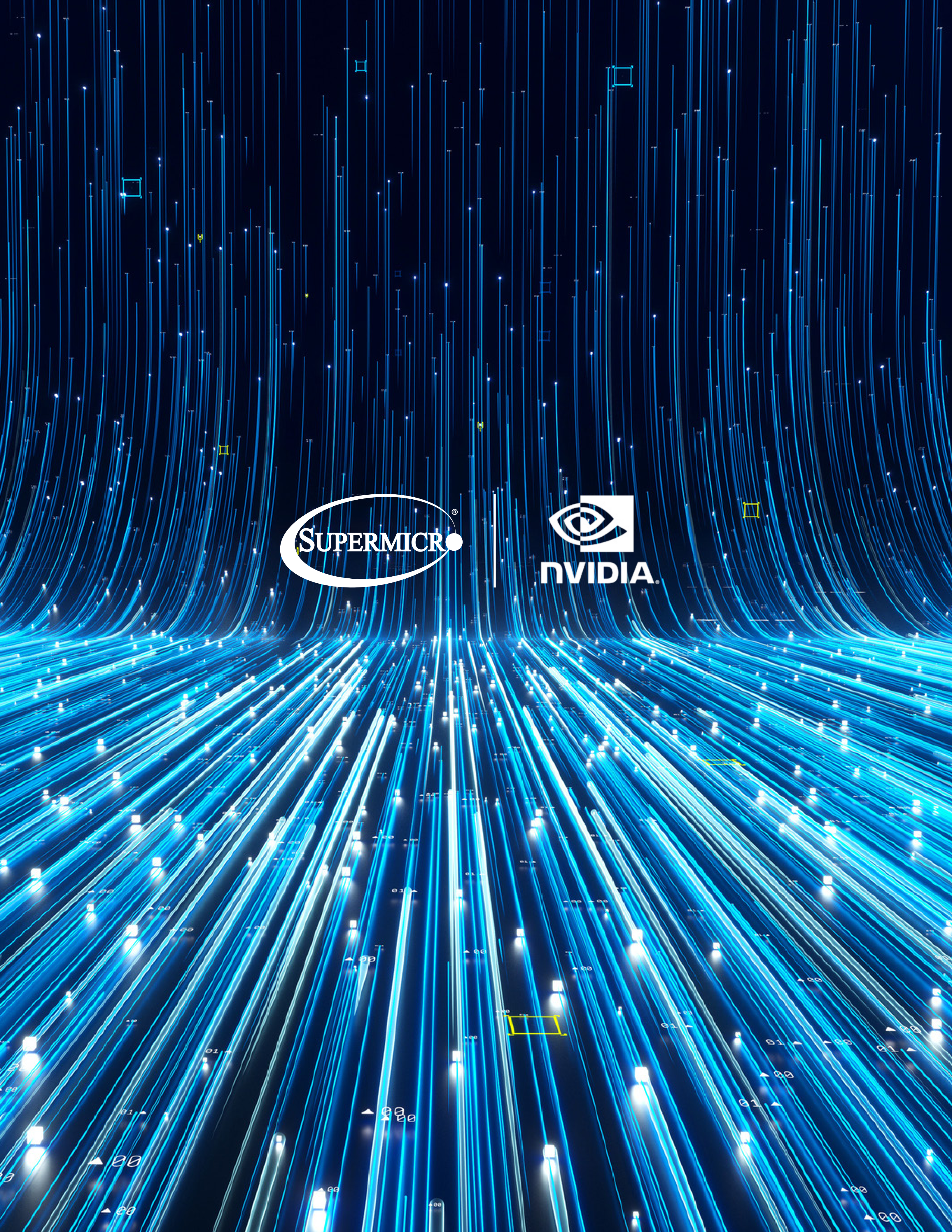
Summary

Enterprises in every industry are looking to develop and deploy AI to improve business results across their entire organization. Scalability and flexibility will become even more important as AI becomes more pervasive. A consistent platform and infrastructure strategy is vital for optimally deploying, scaling, and operating AI applications in the next chapter of your digital transformation journey.

Supermicro and NVIDIA help equip enterprises for success in a wide range of needs, environments, and AI workloads. A global leader in enterprise computing, storage, networking, and green computing technology, Supermicro offers turnkey, pre-defined, pre-tested, and validated rack-level solutions for the most demanding workloads found in advanced data center environments.

Learn more about optimizing your organization's AI with complete racks pre-configured with the latest servers, storage, networking equipment, cabling, software configurations, and management infrastructure designed and built by a global in-house staff of data center experts.

https://www.supermicro.com/en/products/GPU


SUPERMICRO
NVIDIA