WHITE PAPER

Securing AI Workloads with Intel® TDX, NVIDIA Confidential Computing and Supermicro Servers with NVIDIA HGX™ B200 GPUs: A Foundation for Confidential AI at Scale

SEPTEMBER 2025

TABLE OF CONTENTS

EXECUTIVE SUMMARY

As artificial intelligence (AI) and machine learning (ML) workloads grow in complexity and sensitivity, organizations face increasing pressure to ensure both performance and data confidentiality. This white paper explores how Intel's Trust Domain Extensions (TDX) and NVIDIA Confidential Computing with Supermicro's HGX B200-based systems together provide a powerful, secure, and scalable platform for next-generation AI infrastructure. By combining hardware-based security with cutting-edge secure GPU acceleration, businesses can unlock new opportunities in AI while maintaining compliance and protecting intellectual property.

INTRODUCTION

The AI revolution is transforming industries—from healthcare and finance to manufacturing and government—by unlocking new levels of automation, personalization, and decision-making. Yet, the infrastructure required to support large-scale AI models brings challenges in performance, scalability, and, critically, data security.

AI workloads frequently process sensitive and regulated data. Ensuring confidentiality during both training and inference is essential. Still, persistent security risks, siloed data environments, and regulatory complexity often limit an organization's ability to leverage its data fully—and realize its competitive advantage.

As a rapidly evolving field, Generative AI (GenAI) presents new privacy and security challenges, compounding existing risks. These include long-standing threats like data leakage and intellectual property theft, as well as emerging issues such as hallucinations, prompt injection, and data poisoning. Failure to mitigate these risks can result in significant reputational, financial, and operational damage—alongside the opportunity cost of missed innovation. Unsurprisingly, security and privacy are now cited as top barriers to GenAI adoption in enterprise environments.

Traditional security models fall short, particularly in multi-tenant cloud environments, where data may be exposed to privileged infrastructure layers, including hypervisors and system administrators. At the same time, the demand for high-throughput and low-latency compute continues to rise, driven by the scale and complexity of modern AI workloads.

This document introduces a platform solution that addresses both challenges:

- Intel® CPUs with Trust Domain Extensions (TDX) provide hardware-based Confidential Computing capabilities to protect sensitive data during AI execution—even in untrusted or shared environments.
- NVIDIA Confidential Computing expands the protection of sensitive data and models during processing by the NVIDIA Blackwell GPU.
- The Supermicro system with NVIDIA Blackwell GPUs, accelerated by the NVIDIA Blackwell architecture, delivers exceptional GPU-accelerated performance, making it ideal for training and inference at scale.

Combined, these three technologies form a secure, high-performance end-to-end capability for confidential AI—enabling enterprises to protect data, meet compliance mandates, and scale their AI strategies with confidence.

POWERING A NEW ERA OF CONFIDENTIAL AI WITH CONFIDENTIAL COMPUTING

Intel® Trust Domain Extensions (TDX) is a foundational technology driving the next generation of Confidential Computing. It enables hardware-isolated, encrypted execution environments that protect sensitive data—even within untrusted infrastructure—by ensuring both confidentiality and integrity during processing. Intel (TDX) empowers organizations to process sensitive data—such as financial records, healthcare information, and proprietary AI models—securely in both cloud and on-premises environments. By establishing hardware-enforced isolation at the virtual machine level, TDX fosters greater trust, supports regulatory compliance, and accelerates innovation across data-driven industries— while delivering performance that scales with enterprise and AI workloads.

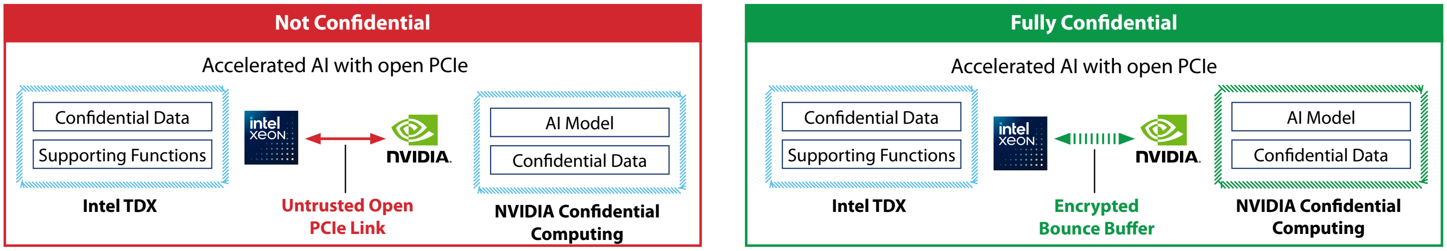
TDX extends these protections to entire virtual machines by creating Trust Domains (TDs)—hardware-enforced isolation boundaries that shield workloads from the host OS, hypervisor, and other VMs. Through built-in attestation mechanisms, TDX provides cryptographic evidence that workloads are running in trusted, verified environments. Whether deployed in multi-tenant public clouds or on-premises data centers, TDX empowers organizations to process sensitive data with a high degree of assurance, reducing the trusted computing base and enabling verifiable trust.

Key Features:

- **Hardware-Isolated Trust Domains** - Creates secure, hardware-backed enclaves (Trust Domains) that are isolated from the host, hypervisor, and other tenants.
- **Protection from Privileged Software** - Shields workloads even from system administrators and infrastructure software, dramatically reducing the trusted computing base.
- **Attestation and Verifiable Integrity** - Enables cryptographic attestation of the execution environment, allowing enterprises and partners to validate the trustworthiness of workloads before releasing sensitive data or executing critical operations.
- **Confidential AI Workloads** - Enables secure training and inference on sensitive datasets—without sacrificing the performance needed for modern AI pipelines.
- **Compliance-Ready Architecture** - Supports stringent data protection regulations such as GDPR, HIPAA, CCPA, and DORA, helping enterprises meet audit and assurance requirements.
- **Enterprise-Class Performance** - Built on Intel's latest silicon platform enhancements, TDX is optimized to deliver strong isolation with minimal overhead, enabling scalable deployment of real-time, latency-sensitive applications.
- **Secure GPU Connectivity with TDX** - Enables secure integration of GPUs with TDX-protected virtual machines using techniques like bounce buffers to control data flow outside the trust boundary. When combined with TDX Connect, this approach supports policy-based orchestration of accelerator access. It ensures that sensitive data is only exposed to verified, authorized components—empowering high-performance AI workloads without compromising confidentiality.

NVIDIA's Blackwell architecture revolutionizes Confidential Computing and Secure AI with Trusted Execution-enabled GPUs that deliver industry-first GPU-based confidential computing capabilities directly within the accelerator. NVIDIA B200 maintains nearly identical performance compared to unencrypted workloads across models of every size, including large language models (LLMs). NVIDIA Confidential Computing ensures that sensitive data, proprietary models, and inference results remain encrypted and isolated within secure GPU enclaves, protected from operating systems, hypervisors, and cloud providers.

The combination of Confidential Computing technologies on Intel Xeon and NVIDIA Blackwell is creating a new standard for secure, cloud-native computing by extending confidential computing seamlessly from CPU to GPU. The first key advancement is the encrypted bounce buffer, which keeps data encrypted even during transfers between TDX-protected environments and NVIDIA accelerators—an essential capability for protecting sensitive AI and analytics pipelines. This innovation reflects the strategic collaboration between Intel and NVIDIA, ensuring that confidential computing now safeguards workloads not just in the CPU, but also throughout GPU-accelerated inferencing and training. Alongside this, foundational features such as attestation and emerging services like TDX Connect signal the next phase of scalable, integrated trust, promising new ways to attest, manage, and securely deploy workloads across cloud and hybrid infrastructures at enterprise scale.



Confidential Computing with Intel and NVIDIA on the Supermicro HGX B200 system is ideally suited for cloud, edge, and hybrid environments, where data privacy and performance are both mission critical. It allows enterprises to confidently deploy AI and data-intensive workloads in shared infrastructure—without compromising security, compliance, or efficiency.

SUPERMICRO SERVERS WITH NVIDIA BLACKWELL GPUS: AI PERFORMANCE REDEFINED

The Supermicro HGX B200-based system is engineered for extreme AI performance at scale, delivering one of the most advanced and mature designs on the market. Built on decades of collaboration with NVIDIA and powered by the cutting-edge Blackwell architecture, it supports up to eight NVIDIA Blackwell GPUs in an HGX baseboard configuration, offering exceptional compute density and scalability for next-generation AI workloads.

With NVIDIA Blackwell architecture and Intel TDX–enabled CPUs, the system brings together the latest secure GPU acceleration and CPU security capabilities to deliver end-to-end protection and performance. This unique combination positions the HGX B200 as ideally suited for high-performance AI in regulated or multi-tenant environments.

Delivered as an out-of-the-box solution, Supermicro provides a broad portfolio of pre-validated SKUs optimized for these deployments. Stocked and ready to ship, most configurations can be delivered within days, enabling organizations to accelerate time-to-value while retaining complete control of their infrastructure.

A broad portfolio of Supermicro SKUs optimized for confidential AI use cases is now available and ready for immediate deployment. These product offerings are pre-validated and stocked for rapid shipment, with most configurations deliverable within days, enabling organizations to accelerate time-to-value while maintaining complete control over sensitive data and infrastructure.

System Highlights:

- **Dual Intel® Xeon® 6900/6700 Series CPUs:** High-core-count processors optimized for AI & HPC workloads

		GPU Support	
		HGX H100/H200	**HGX B200**
Generation	**X13 (5th Gen Xeon)**	SYS-821GE-TNHR	SYS-A21GE-NBRT
		SYS-421GE-TNHR2-LCC	SYS-421GE-NBRT-LCC
	X14 (Xeon 6)		SYS-A22GA-NBRT
			SYS-422GA-NBRT-LCC
			SYS-422GS-NBRT-LCC

- **NVIDIA HGX B200 8-GPU Baseboard:** Up to 60 TB/s of bandwidth and 1.4TB of direct GPU memory
- **PCIe 5.0 and MRDIMM Support:** High-speed interconnects and memory for data-intensive applications
- **Liquid and Air-Cooling Options:** Flexible thermal management for diverse deployment environments

Designed for generative AI, large language models (LLMs), and high-performance computing (HPC), the Supermicro system with NVIDIA HGX B200 servers as a cornerstone of next-generation AI infrastructure—delivering the scalability, compute density, and efficiency required for today's most demanding workloads.

AI USE CASES ENABLED BY CONFIDENTIAL COMPUTING AND SUPERMICRO SERVERS WITH NVIDIA HGX BLACKWELL GPUS

Private AI Training on Regulated Data	Secure Multi-Party Model Training / Federated Learning	Confidential Inference-as-a-Service	Confidential Fine-Tuning of Foundation Models
Unlocks high-value data that was previously off-limits due to compliance or trust concerns	Enables collaboration without compromising IP or confidentiality	Enables trust in AI-as-a-service offerings, attracting privacy-sensitive clients	Retains ownership of sensitive domain knowledge embedded in models
Accelerates time-to-insight with richer, more representative training data	Expands model scope and robustness by training on diverse datasets	Differentiates your platform by offering verifiable confidentiality	Enables model customization without needing to trust a third-party host
Strengthens compliance posture under GDPR, HIPAA, and emerging regulations like DORA & the EU AI Act	Reduces legal and operational friction in data sharing agreements	Expands market access in sectors with high privacy expectations	Drive faster innovation without compliance bottlenecks
Improves model accuracy and relevance by allowing access to previously siloed or encrypted datasets	Builds competitive alliances while preserving each party's data sovereignty	Mitigates data leakage risk, protecting against insider threats or cloud misconfigurations	Preserves competitive advantage by protecting data during the fine-tuning process

Regulatory-Compliant AI Deployment
Simplifies compliance with verifiable controls (e.g., attestation, access logs) Reduces regulatory risk and potential fines Streamlines audits and vendor reviews, accelerating go-to-market

TESTIMONIAL

GMI Cloud, a premier NVIDIA Reference Platform Cloud Partner, has built its AI infrastructure around Supermicro HGX GPU systems and Intel® Trust Domain Extensions (TDX) to deliver confidential computing at scale, from the infrastructure layer up to the application layer.

GMI Cloud provides secure, high-performance AI infrastructure for large-scale model training and inference, tailored to finance, tech, and AI-native enterprises with strict compliance and data residency needs. The breadth of Supermicro's validated portfolio allows GMI Cloud to bring new GPU generations online swiftly while maintaining enterprise-grade availability.

By integrating TDX into both its Inference Engine and Cluster Engine, and deploying Supermicro systems with NVIDIA HGX B200, GMI ensures high-performance, secure AI services with container-level confidentiality, and GPU-optimized runtimes with peak tokens-per-second for their secure real-time agents and chat services. As Yujing Qian, VP of Engineering at GMI Cloud, states: "By building our infrastructure on Supermicro NVIDIA HGX based systems with Intel TDX, we enable enterprise-grade isolation, attestation, and performance… providing a secure, scalable foundation for AI-native companies, financial services, and global technology leaders."

CONCLUSION

As AI adoption accelerates across every industry, organizations must balance performance at scale with trust, security, and compliance. The combination of the Supermicro HGX B200 system built on NVIDIA Blackwell with Confidential Computing and Intel® Trust Domain Extensions (TDX) delivers a high-performance, confidential AI platform built for the demands of modern, data-driven enterprises.

With support for up to eight NVIDIA Blackwell GPUs, the Supermicro-based system with NVIDIA HGX B200 GPUs offers industry-leading compute density, energy efficiency, and scalability—essential for training and deploying large language models, generative AI workloads, and high-performance computing applications. When paired with Intel TDX, this platform enables end-to-end data protection, extending hardware-enforced confidentiality and integrity to entire virtual machines. This ensures sensitive data—whether in the cloud, at the edge, or on-premises—remains protected throughout the AI lifecycle.

From privacy-preserving AI training and federated learning to secure inference-as-a-service, this joint solution empowers enterprises to unlock the full value of their data without compromising security or regulatory posture.

Backed by decades of deep engineering collaboration between Intel, Supermicro, and NVIDIA, this solution reflects a shared commitment to innovation, open standards, and enterprise-grade reliability. It positions customers to move faster, operate more securely, and innovate with confidence in the AI era.

Discover how our cutting-edge systems support TDX and are optimized for confidential AI workloads. Designed with security and performance in mind, our solutions empower organizations to harness the full potential of trusted execution environments and privacy-preserving AI. Visit GPU Servers For AI, Deep / Machine Learning & HPC | Supermicro to explore technical specifications, use cases, and how we're advancing secure innovation.

SUPER MICRO COMPUTER, INC.

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 on your requirements.

www.supermicro.com

INTEL

Intel (Nasdaq: INTC) is an industry leader, creating world-changing technology that enables global progress and enriches lives. Inspired by Moore's Law, we continually work to advance the design and manufacturing of semiconductors, helping to address our customers' greatest challenges. By embedding intelligence in the cloud, network, edge, and every kind of computing device, we unleash the potential of data to transform business and society for the better. To learn more about Intel's innovations, go to newsroom.intel.com and intel.com.

www.intel.com

NVIDIA

NVIDIA accelerated computing platforms power the new era of computing, performing exponentially more work in less time with much lower energy consumption than traditional CPU-based computing. Accelerated computing revolutionizes energy efficiency across industries by harnessing NVIDIA GPUs, CPUs, and networking, all optimized through NVIDIA enterprise software solutions.

www.nvidia.com

SUPERMICRO®

Printed in USA ♻ Please Recycle 14_2025-AI-Workloads_Intel-TDX_R03 SEPTEMBER 2025