SOLUTION BRIEF

SUPERMICRO AND SCONTAIN CREATE A BLUEPRINT FOR EMPOWERING CONFIDENTIAL AI

SCONE Secure Cloud Infrastructure Platform Enables Encrypted AI Data Processing

Supermicro Enterprise AI Server SYS-322GA-NR

Table of Contents

Executive Summary

In an era where artificial intelligence (AI) and machine learning (ML) demand unprecedented computational power while stringent security requirements protect sensitive data, the collaboration between Supermicro and Scontain represents a pivotal advancement in secure cloud infrastructure. Supermicro, a global leader in high-performance, application-optimized server solutions, partners with Scontain, the German innovator behind SCONE—the premier confidential computing platform. This alliance combines Supermicro's robust hardware ecosystem, including the SYS-322GA-NR server optimized for AI workloads, with Scontain's cutting-edge software layer, which enables seamless confidential execution environments.

Confidential computing facilitates processing data without seeing it. Often, companies enlist the services of external providers to operate components of their computing environment. This creates additional data surfaces that are exposed to potential insider attackers. These insider attacks can be mitigated by masking both the data and the encryption keys. Using the n-eyes principle, SCONE enables the management of applications by an external provider. This results in a cloud platform that meets the highest security standards, allowing it to be used in, for example, the German healthcare market, which prohibits anyone with access to the computing infrastructure from being able to see any patient data.

By integrating Supermicro's GPU-accelerated systems with SCONE's process-based Trusted Execution Environments (TEEs), the partnership delivers a unified solution that accelerates AI training and inference without compromising data integrity. This collaboration leverages Intel® Trust Domain Extensions (Intel TDX) for CPU-level confidentiality and NVIDIA H200 and RTX PRO™ GPUs for high-throughput AI processing, enabling enterprises to deploy mission-critical AI applications in untrusted cloud environments with full assurance.

Security and Compliance Challenges for Processing Sensitive Data

Cloud computing has revolutionized data processing [1], including training/fine-tuning and inference AI/ML workloads [2], but it introduces profound security and compliance risks, particularly for sectors that handle sensitive information, such as eHealth and finance. Traditional cloud setups expose data to threats from privileged insiders, malicious administrators, or supply-chain vulnerabilities, even during computation—"data in use" remains a glaring blind spot in encryption paradigms that safeguard data at rest and in transit.

In eHealth, patient records and genomic data must comply with regulations like HIPAA, where breaches can erode trust and incur massive fines. Financial institutions grapple with PCI-DSS and similar standards, facing risks from model poisoning or data exfiltration during AI-driven fraud detection or algorithmic trading. Multi-tenant clouds amplify these issues, as shared infrastructure heightens the attack surface, while the rise of generative AI exacerbates concerns over intellectual property leakage and regulatory scrutiny under frameworks like the EU AI Act.

Positioned as a comprehensive, end-to-end Cloud Confidential Computing platform, SCONE redefines secure AI/ML deployment. It combines CPU TEEs [3] via Intel TDX for encrypted memory isolation, confidential GPUs with NVIDIA H200 and RTX PRO™ for protected AI/ML acceleration, and SCONE's secure orchestration for Kubernetes-native workflows. This stack ensures data sovereignty by attesting to the integrity of code, data, and execution environments, allowing organizations to process sensitive workloads in public clouds while maintaining granular control over access and compliance. The result: a scalable, performant framework that bridges high-performance computing (HPC) with ironclad security, enabling AI innovation without the trade-offs of on-premises silos.

System Design Overview

Insider Attacks and Ransomware Attacks

Supermicro and Scontain's systems are designed to defend against a highly capable adversary operating within complex cloud virtualization environments. The adversary might try to gain access to administrators' credentials and is aware of all vulnerabilities. In this scenario, we assume an attacker has gained full control over the system software stack—including the operating system (OS) and hypervisor—and may even perform simple physical attacks, such as memory probing. Additionally, we assume the cloud network is untrusted, allowing the adversary to drop, inject, replay, or alter packets and manipulate routing paths. These conditions reflect the classical Dolev-Yao adversary model [4].

Side-channel attacks [5, 6] are well studied, but in practice, they are much more challenging to perform. Nevertheless, the SCONE [7] platform provides mitigation against L1-based side-channel threats [8] and L2-based side-channel threats with AEX-Notify [9]. In addition, it is hardened against Iago attacks [10]. To address Spectre-related vulnerabilities [11, 12], we employ LLVM-based techniques such as speculative load hardening. Denial-of-service attacks are also out of scope, as they can be trivially executed by any entity controlling the infrastructure, such as the OS or hypervisor.

SCONE enables the encryption of credentials so that the encryption key is accessible only to, e.g., the confidential backup program. The confidential backup program can use the credentials to store an encrypted backup on an external S3 object

store, but an adversary with root access cannot access the credentials. Files encrypted with SCONE are automatically checked for ransomware before they are sent to the object store.

Building Blocks

SCONE: Confidential Computing with Minimal Trusted Computing Base (**TCB)**

Shielded Execution

At the software core, SCONE [7] provides a shielded execution framework that enables unmodified applications to run inside TEE enclaves (e.g., Intel SGX/TDX). Built on Intel SGX/TDX and compatible with emerging confidential GPU technologies, SCONE encrypts data and code in use, shielding them from host OS, hypervisors, or cloud providers. Its hallmark is a minimal Trusted Computing Base (TCB) enabling granular trust at the microservice level. This isolates individual components, drastically reducing the attack surface compared to full-VM TEEs, while facilitating straightforward audits through remote attestation reports that verify enclave integrity without exposing contents. SCONE's architecture supports unmodified applications, wrapping them in secure wrappers that handle encryption keys and attestation seamlessly. For AI workloads, it extends protection to GPU offloads, ensuring models and datasets remain confidential during tensor operations on NVIDIA H200 hardware. This process-based approach not only minimizes overhead but also integrates natively with CI/CD pipelines [13], empowering developers to build compliant applications effortlessly.

Trustless Cloud-Native Deployment with K8s

System orchestration is streamlined through Kubernetes (K8s) compatibility, enabling cloud-native deployments without trusting the underlying OS or K8s control plane. SCONE acts as a secure overlay, injecting enclave protections into container runtimes such as Docker or containerd, so that workloads execute in isolated TEEs regardless of the host environment. This "zero-trust" model means no modifications to applications, libraries, or manifests are required—SCONE's wrappers handle encryption, key management, and attestation transparently. For multi-node clusters, SCONE facilitates attested sidecar proxies for service mesh integration (e.g., Istio), ensuring end-to-end confidentiality in microservices architectures. Deployment is as simple as annotating K8s pods with SCONE policies, supporting autoscaling for AI jobs while preserving data sovereignty across hybrid clouds.

SCONE has been used and extended in several EU projects. Most recently in projects IPCEI-CIS, NearData (No 101092644), CloudSkin (No 101092646) and AISprint (No. 101016577).

Supermicro SYS-322GA-NR: Dual Processors, Quad NVIDIA GPUs, and Intel TDX

The hardware foundation is Supermicro's SYS-322GA-NR, a powerful 3U rackmount server engineered for dense, GPU-intensive AI and HPC deployments. It accommodates dual Intel Xeon 6900-series processors (with P-cores up to 128 cores/256 threads per CPU and 500W TDP), delivering up to 256 cores of parallel processing power fortified by Intel SGX and TDX for hardware-rooted confidentiality. Memory scales to 6TB of DDR5-6400 RDIMM across 24 slots, ensuring ample capacity for large-scale datasets. GPU support is unparalleled, with up to 10 PCIe 5.0 x16 slots enabling configurations of 4x NVIDIA H200 NVL (141GB HBM3e each) or 4x RTX PRO™ 6000 Blackwell Server Edition (96GB GDDR7 each) GPUs for versatile AI model training/finetuning/inferencing.

Storage includes up to 14 hot-swap E1.S NVMe bays for ultra-fast data access, while high-I/O bandwidth and power efficiency (optimized cooling for 6kW+ TDP) make it ideal for edge-to-cloud scaling. Integrated IPMI for remote management and redundant power supplies ensures reliability in production environments.

Detailed Design

Scontain has engineered a secure, confidential AI/ML system leveraging Trusted Execution Environments (TEEs), such as Intel TDX and NVIDIA H200 GPUs provided by Supermicro. This architecture ensures the protection of sensitive data and computations throughout the ML lifecycle, including inference, training, and fine-tuning. Secure VMs powered by TEEs integrate seamlessly with NVIDIA H200 GPUs via encrypted bound buffers, enabling high-performance confidential computing without exposing plaintext data to untrusted hosts or infrastructure. To facilitate robust remote attestation of ML applications, a custom kernel module is implemented within the Guest OS of each confidential VM. This module performs dynamic measurements of running ML workloads, generating verifiable quotes that attest to the integrity and authenticity of the applications.

- Configuration and Attestation Service (CAS): Acts as the root of trust for the entire system. CAS operates within its own TEE and can be directly attested by end users, ensuring transparency and verifiability. It manages security policies, handles attestation workflows, and provisions secrets (e.g., decryption keys) upon successful verification.

- Kernel Module for Attestation: Deployed in the Guest OS of confidential VMs, this module measures ML applications at runtime. It interacts with CAS to obtain signing keys, generates cryptographic quotes based on application hashes, and supports the overall remote attestation chain.

- Security Policies: User-defined configurations that specify allowable measurements, attestation requirements, and secret provisioning rules for confidential VMs and ML applications.

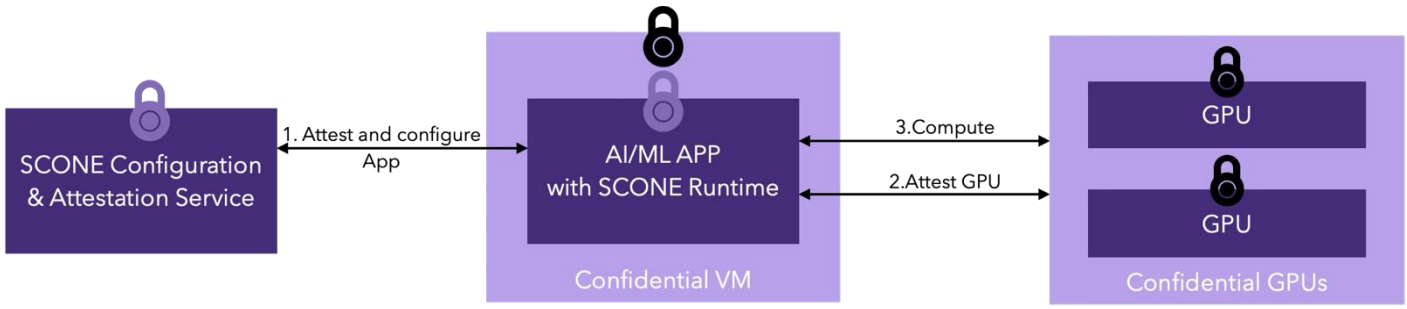

Figure 1 - System Architecture

System Workflow

Figure 1 illustrates the high-level workflow of the system architecture. The setup of the confidential AI/ML workflow begins with the user defining a security policy for their AI/ML application and attesting to the Configuration and Attestation Service (CAS) [14], which operates within a Trusted Execution Environment (TEE) to establish trust. The user then submits the policy to CAS. The workflow starts by spawning a confidential VM using Intel TDX, a VM-specific policy extension containing the VM ID is generated and uploaded to CAS. As the VM boots, its hardware measures and attests the firmware, which in turn attests the Guest OS and kernel, including a custom attestation kernel module, and the kernel module measures the running ML application—during inference or training—generates a cryptographic hash, signs it to produce a quote, and sends it to CAS. CAS verifies the quote against the policy and, if valid, provisions configuration data and secrets, such as decryption keys (step

1). This module initiates a request to the GPU driver to attest to confidential GPUs [15] (step 2). A successful GPU attestation enables the ML application to securely process data within the TEE and the NVIDIA H200 GPU environment (step 3).

Benchmarks

Supermicro and Scontain have rigorously validated the SYS-322GA-NR with SCONE, benchmarking end-to-end workflows to confirm < 5% overhead in AI training on H200 GPUs. Security validations include penetration testing against TDX exploits and full attestation chain verification, guaranteeing production-grade reliability.

Use Cases and Value Proposition

Enables secure AI in eHealth, finance, and multi-stakeholder scenarios

This solution unlocks secure AI across high-stakes domains. In eHealth, it powers secure multi-stakeholder machine learning [17, 18] and federated learning [19] for collaborative diagnostics on patient data without centralization, enabling AI/ML models to train/fine-tune on distributed genomic datasets while preserving privacy. For finance, confidential GPUs accelerate risk modeling and anomaly detection on transaction logs, mitigating insider threats during real-time processing. In multi-stakeholder scenarios, such as supply-chain AI/ML consortia, it supports attested multi-party computation, allowing competitors to co-train models on private data without exposing their data. Comprehensive protection spans the data lifecycle: at rest via encrypted NVMe drives and TDX-secured memory; in transit through TLS-wrapped SCONE channels with mutual attestation; and during processing in GPU enclaves that shield computations from the host. This eliminates “data in use” vulnerabilities, with hardware-enforced isolation ensuring that even root-level attacks cannot access plaintext.

Easy application deployment while ensuring compliance (e.g., GDPR)

Deployment simplicity is a cornerstone—applications run unmodified in K8s, with SCONE automating compliance artifacts, such as GDPR-mandated data protection impact assessments, via auditable attestation logs. This “compliance-by-default” approach reduces engineering overhead by 70% compared to custom TEE integrations, enabling rapid iteration while meeting global standards such as GDPR, HIPAA, and the EU AI Act.

Conclusion

As the cornerstone for cloud confidential computing, the SYS-322GA-NR + SCONE stack is purpose-built for AI training/inferencing—handling trillion-parameter LLMs with confidential NVIDIA H200 and NVIDIA RTX PRO GPU acceleration—and HPC simulations requiring secure, high-fidelity data processing. It scales from proof-of-concept clusters to exascale deployments, democratizing secure AI for enterprises worldwide.

The joint design minimizes the trusted computing base (TCB) by relying on hardware-enforced TEEs and granular attestation, reducing the attack surface while supporting unmodified ML applications. It enables secure multi-tenant cloud deployments for sensitive workloads across domains such as healthcare, finance, and research. Future enhancements could include support for additional TEE technologies or federated learning scenarios to further extend confidentiality across distributed environments.

References

[1] Do Le Quoc, Franz Gregor, Jatinder Singh, and Christof Fetzer. Sgx-pyspark: Secure distributed data analytics. In The World Wide Web Conference (WWW), 2019.

[2] Do Le Quoc, Franz Gregor, Sergei Arnautov, Roland Kunkeland, Pramod Bhatotia, and Christof Fetzer.secureTF: A Secure TensorFlow Framework. In Proceedings of the 21st International MiddlewareConference (Middleware), 2020.

[3] Jatinder Singh, Jennifer Cobbe, Do Le Quoc, and Zahra Tarkhani. Enclaves in the clouds: Legal considerations and broader implications. Communications of the ACM, 2021.

[4] D. Dolev and A. C. Yao. On the security of public key protocols. In Proceedings of the 22nd Annual Symposium on Foundations of Computer Science (SFCS), pages 350–357, 1981.

[5] Ferdinand Brasser, Urs Müller, Alexandra Dmitrienko, Kari Kostiainen, Srdjan Capkun, and Ahmad-Reza Sadeghi. Software grand exposure:{SGX} cache attacks are practical. In 11th {USENIX} Workshop on Offensive Technologies (WOOT), 2017.

[6] Wenhao Wang, Guoxing Chen, Xiaorui Pan, Yinqian Zhang, XiaoFeng Wang, Vincent Bindschaedler, Haixu Tang, and Carl A Gunter. Leaky cauldron on the dark land: Understanding memory side-channel hazards in sgx. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS), 2017.

[7] Sergei Arnautov, Bohdan Trach, Franz Gregor, Thomas Knauth, Andre Martin, Christian Priebe, Joshua Lind, Divya Muthukumaran, Dan O'Keeffe, Mark L. Stillwell, David Goltzsche, Dave Eyers, Rüdiger Kapitza, Peter Pietzuch, and Christof Fetzer. SCONE: Secure Linux containers with Intel SGX. In the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI), 2016.

[8] Oleksii Oleksenko, Bohdan Trach, Robert Krahn, Mark Silberstein, and Christof Fetzer. Varys: Protecting SGX enclaves from practical side-channel attacks. In Proceedings of the USENIX Annual Technical Conference (USENIX ATC), 2018.

[9] Scott Constable, Jo Van Bulck, Xiang Cheng, Yuan Xiao, Cedric Xing, Ilya Alexandrovich, Taesoo Kim, Frank Piessens, Mona Vij, and Mark Silberstein. {AEX-Notify}: Thwarting precise {Single-Stepping} attacks through interrupt awareness for intel {SGX} enclaves. In 32nd USENIX Security Symposium (USENIX Security 23), pages 4051–4068, 2023.

[10] Stephen Checkoway and Hovav Shacham. Iago attacks: Why the system call api is a bad untrusted rpc interface. In Proceedings of the Eighteenth International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), 2013.

[11] Paul Kocher, Jann Horn, Anders Fogh, Daniel Genkin, Daniel Gruss, Werner Haas, Mike Hamburg, Moritz Lipp, Stefan Mangard, Thomas Prescher, Michael Schwarz und Yuval Yarom. Spectre attacks: Exploiting speculative execution. In 40th IEEE Symposium on Security and Privacy (S&P'19), 2019.

[12] G. Chen, S. Chen, Y. Xiao, Y. Zhang, Z. Lin, and T. H. Lai. SgxPectre: Stealing Intel Secrets from SGX Enclaves Via Speculative Execution. In IEEE European Symposium on Security and Privacy (Euro S&P), 2019.

[13] Robert Krahn, Nikson Kanti Paul, Franz Gregor, Do Le Quoc, Andrey Brito, André Martin, and Christof Fetzer. Tical: Trusted and integrity-protected compilation of applications. In 2024, the 19th European Dependable Computing Conference (EDCC), 2024.

[14] Franz Gregor, Wojciech Ozga, Sébastien Vaucher, Rafael Pires, Sergei Arnautov, André Martin, Valerio Schiavoni, Pascal Felber, Christof Fetzer, et al. Trust management as a service: Enabling trusted execution in the face of byzantine stakeholders. In 50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), 2020.

[15] NVIDIA. NVIDIA Confidential Computing. https://www.nvidia.com/en-in/data-center/ solutions/confidential-computing/, 2025.

[16] Franz Gregor, Robert Krahn, Do Le Quoc, and Christof Fetzer. Sinclave: Hardware-assisted singletons for tees. In Proceedings of the 24th International Middleware Conference, 2023.

[17] Wojciech Ozga, Christof Fetzer, et al. Perun: Confidential multi-stakeholder machine learning framework with hardware acceleration support. In IFIP Annual Conference on Data and Applications Security and Privacy, pages 189–208. Springer, 2021.

[18] Wojciech Ozga, Do Le Quoc, and Christof Fetzer. Perun: Secure multi-stakeholder machine learning framework with gpu support. arXiv preprint arXiv:2103.16898, 2021.

[19] Do Le Quoc and Christof Fetzer. Secfl: Confidential federated learning using tees. arXiv preprint arXiv:2110.00981, 2021.

Further Information

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

https://www.supermicro.com/en/products/system/iot/3u/sys-322ga-nr

https://scontain.com/

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.

Learn more at www.supermicro.com

SCONTAIN

Scontain GmbH is one of the leading companies in confidential computing domain. Scontain supports its customers to build confidential applications with the help of their SCONE platform. It has a strong partnership with cloud companies, e.g. Deutsche Telekom and Microsoft Azure.

Learn more at: www.scontain.com