INTEL SECURE AI

PROTECTING THE DATACENTER, CLIENT, AND NETWORK EDGE

EXECUTIVE SUMMARY

Next-generation enterprise AI is poised to become one of the most disruptive technology platforms introduced in decades, if not this century. It has the potential to unlock newfound efficiencies within enterprise IT and OT operations and incubate new business value. However, as with any leapfrog infrastructure advancement, bad actors will attempt to exploit vulnerabilities for financial gain and other advantages. Thus, organizations must defend against AI weaponization across the entire network expanse—from datacenters on-premises or in the cloud, to clients at the network edge.

Enterprise AI LLMs and the underlying algorithms represent a significant investment and are only valuable if they can be trusted. That is why they must be protected. Used for generative AI or specialized analyses, they contain intellectual property and business-critical or regulated data. Data protection is a complex endeavor that requires prescriptive considerations that safeguard data at rest, in motion, and in use to ensure refinement for higher levels of accuracy over time. In addition to AI model and data protection, fortified defense tools are also required to combat AI-enriched attacks that can occur anywhere along workload runtimes. It is a two-sided coin: the need to provide AI for security and security for AI to ensure end-to-end protection.

Moor Insights & Strategy believes that Intel's approach to secure AI provides a complete framework for combating adversarial AI anywhere it occurs. Intel's strategy is sound, and its products provide essential foundational technologies and security solutions that will help protect enterprise AI deployments at scale. As proof, the company's efforts in security for AI enable a host of independent software vendors to leverage deep silicon features to detect cyber-attacks with greater efficacy. This effort also helps protect models and data input that run on infrastructure from edge to cloud. To this end, Intel enables mission-critical confidential AI in the cloud and datacenter.

On the other hand, both Intel Threat Detection Technology (Intel TDT) at the client level and independent attestation with Intel Tiber Trust Services for workload verification demonstrate the company's commitment to AI for security. Furthermore, Intel's continued investment in building an ecosystem through best-in-class partnerships, extensive internal testing, external bug bounty programs, and groundbreaking research through Intel Labs also enables an adaptive platform to secure AI at scale.

THE RISE OF THE AI PERSONAL COMPUTER

Next-generation AI workloads began with large language models utilizing computational processing in the cloud. This workflow is highly scalable, but it demands enterprises make a significant investment in computational and networking infrastructure by procuring hardware and software or using public cloud provider services. These deployments represent a significant investment in capital and operational expense, and it is not surprising that the evolution of AI workloads would eventually move to the client for cost considerations and result in the birth of the AI PC.

The long-term potential of AI PC's is exciting. It includes the ability to unlock new use cases to increase end-user productivity, reduce operational cost models, and run smaller language models closer to data creation points for improved economics. However, the eventual proliferation of devices and the hybrid nature of future AI workloads spanning cloud and on-premises datacenters to network edges will significantly expand the threat surface. Consequently, this will require additional provisions to manage AI at scale, including a DevOps framework that incorporates AI tools and machine learning to facilitate the visibility and security needed for models that are downloaded to AI PCs and other edge devices from the cloud.

From an endpoint protection standpoint, Intel TDT was one of the first applications of AI introduced on client and edge devices. Its security architecture leverages Intel's CPU telemetry and AI models running on the integrated Intel GPU to scan for threats and ransomware attacks. Today, billions of Intel-based PCs and future Intel AI PCs are protected. Security is often a layered approach, and Intel TDT offers deeper protection through pre-integrations with Microsoft Defender and other leading EDR solution providers.

It is also worth highlighting that Intel collaborates with security ISVs to identify new features and use cases that can take advantage of its client AI XPU hardware. This process involves optimization techniques designed to improve privacy, shorten mean time to issue remediation, and deliver better security outcomes leveraging data at the network edge. The result is not only a stronger security posture for Intel AI PCs but also a new and cost-effective option to run AI workloads locally versus public cloud services.

Intel has a long track record of delivering security on client devices through innovations at the silicon level. MI&St believes that the company's prior success in confidential computing provides a unique advantage that can be effectively leveraged to deliver what is required to secure AI.

AI Security Considerations In The Datacenter

To secure AI and do so end-to-end, the underlying models, algorithms, data, and clients must all be safeguarded, as well as the AI lifecycle itself, spanning development, training, validation, fine-tuning, and deployment. Only a handful of companies have the depth of silicon, breadth of infrastructure and device design-in, deep investment in ecosystem collaborations, and trust attestation functionality to secure AI at scale. Intel possesses all these capabilities and is uniquely positioned to provide organizations with what is required in today's modern, highly distributed operational environment, from cloud to on-premises datacenters.

There is a broad misconception that only GPUs are relevant in AI workload processing. Today's most widely deployed AI and ML inference applications run on CPUs. To this end, Intel continues to demonstrate its capabilities in delivering confidential AI through broad partnerships with public cloud providers, hyperscalers, and industry leaders, such as NVIDIA, that utilize Intel Xeon Scalable processors and Intel Software Guard Extensions (Intel SGX), Intel Trusted Domain Extensions (Intel TDX), and Intel Advanced Matrix Extensions (Intel AMX) as a critical part of deployed AI infrastructure.

From a data security perspective, Intel provides AI workload and application isolation with Intel SGX and virtual machine isolation with Intel TDX. Both are confidential computing technologies that support a trusted execution environment where data and code can operate within an isolated space. Intel SGX employs hardware-based memory encryption that isolates specific application code and data in memory. Intel SGX facilitates this functionality by allocating private regions of memory, often called secure enclaves, which are protected from processes that run at higher privilege levels.

Intel TDX provides virtual machine (VM) isolation from VM managers and hypervisors within public cloud deployments. This capability creates trust domains that have the potential to safeguard a broad range of software. Together, Intel SGX and Intel TDX offer a complete, end-to-end AI model and data protection platform spanning both on-premises and public cloud datacenters. Finally, Intel AMX is an integrated accelerator designed to improve the performance of deep-learning training and inference on Intel Xeon Scalable processors. It is ideally suited for applications like natural language processing, recommendation systems, and image recognition and has the potential to improve business outcomes and lower operational and capital expenditures tied to AI workload processing.

Trust attestation or verification is also an important element within a complete, end-to-end AI security architecture. Intel Tiber Trust Services are designed to establish trust within AI collaboration frameworks, protect workloads, and continually verify confidentiality and integrity in real time. This service does so with a consistent attestation workflow that simplifies the often complex task of managing hybrid and multi-cloud deployments. Furthermore, Intel Tiber Trust Services support NIST Zero Trust Architecture principles while also incorporating SLAs that provide high degrees of uptime and support services.

The success of all of Intel's efforts can be measured in terms of improved security controls, as well as the higher resilience of PC and datacenter applications. Furthermore, organizations can unlock the power of data, facilitate controlled cloud migrations, collaborate with partners while maintaining privacy and compliance standards, and ensure zero trust—all from cloud to network edge with ease. MI&S believes that Intel Tiber Trust Services, as well as the company's success with Intel TDX and Intel SGX all combine to provide a formidable AI workload protection suite cross-domain.

Intel's Secure AI Ecosystem

Intel continues to support and nurture a robust and secure AI ecosystem anchored by partnerships that further the AI model and data protection. The collaboration in cybersecurity is broad and deep, spanning endpoint protection, identity and access management, zero trust, network, application, infrastructure, and data security. However, three specific AI partnerships are noteworthy:

- Fortanix and Intel are collaborating to encrypt data in use through the isolation of data and code in Intel's 4th Gen Xeon Scalable processors and Intel SGX-protected enclaves. It also utilizes Intel AMX to accelerate inferencing in the cloud to improve performance and deliver better AI application outcomes.
- Google Cloud recently announced new confidential VM support using 4th Gen Xeon Scalable processors and Intel TDX and Intel AMX. The combination of extensions implements confidential computing at the VM level to migrate workloads to a trust domain while improving the performance of deep-learning training and inference on the CPU through pre-integrations into PyTorch and TensorFlow.

- Intel is collaborating within HiddenLayer's Technology Alliance Program to further customer innovation, provide enhanced security, and improve business outcomes through the HiddenLayer AISec Platform. AISec is a GenAI protection suite that provides detection and response capabilities to safeguard prompts, defend against adversarial AI attacks, and protect supply chain vulnerabilities. Specifically, HiddenLayer relies on Intel SGX to enable the smallest possible trust boundary to facilitate an encrypted ML model scanner designed to protect models.

INTEL LABS

Intel Labs was founded over two decades ago with a mission to deliver innovation through global research. Its charter and organizational efforts are expansive, addressing emerging challenges across computing, connectivity, cloud-to-edge infrastructure, and AI. These efforts undoubtedly further the company's strength in providing a secure AI platform, as evidenced by its focus on cognitive AI and graph neural networks.

Recent research initiatives, such as advanced deep learning investigations with Habana Labs and Hugging Face and responsible AI research with the Canadian-based Mila research institute, provide Intel with a strong foundation to further refine and expand the capabilities of its secure AI offering.

CALL TO ACTION

Next-generation enterprise AI applications will unlock newfound efficiencies within enterprise operations. Still, organizations must defend datacenters, on-premises or in the cloud, as well as clients at the network edge. Intellectual property that includes large language models, underlying algorithms, and business-critical data is constantly at risk as bad actors attempt to find weaknesses in emerging use cases that include generative AI.

MI&S believes that Intel's approach to secure AI provides a complete framework for combating adversarial AI anywhere it occurs—providing both security for AI and AI for security. The success of these efforts is evidenced by the company's continued deep security investments in its technology, expansive solution co-development activity, deep partnerships, hardened testing, and practical applied research that delivers highly optimized security outcomes.

CONTRIBUTOR

Will Townsend, Vice President & Principal Analyst, Networking & Security Practices at Moor Insights & Strategy

PUBLISHER

Patrick Moorhead, CEO, Founder and Chief Analyst at Moor Insights & Strategy

INQUIRIES

Contact us if you would like to discuss this report, and Moor Insights & Strategy will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts but must be cited in-context, displaying author's name, author's title, and "Moor Insights & Strategy". Non-press and non-analysts must receive prior written permission by Moor Insights & Strategy for any citations.

LICENSING

DISCLOSURES

This paper was commissioned by Intel. Moor Insights & Strategy provides research, analysis, advising, and consulting to many high-tech companies mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

DISCLAIMER

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Moor Insights & Strategy disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Moor Insights & Strategy and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Moor Insights & Strategy provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.