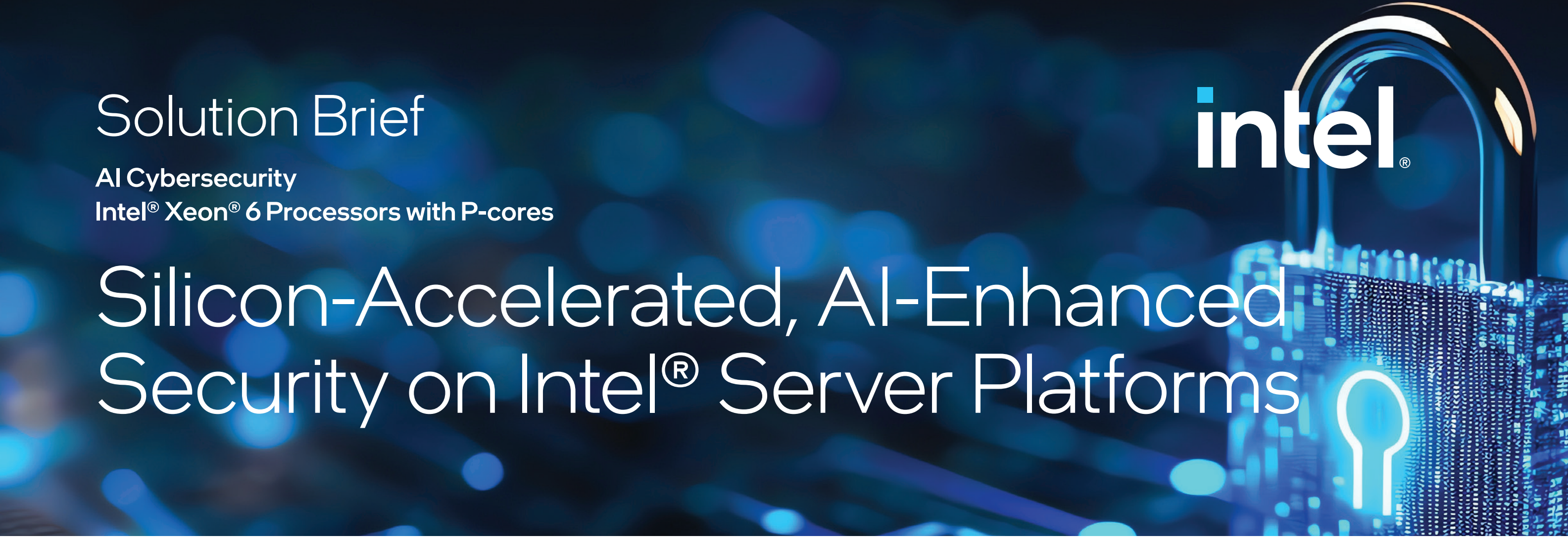

Solution Brief

AI Cybersecurity
Intel® Xeon® 6 Processors with P-cores

Silicon-Accelerated, AI-Enhanced Security on Intel® Server Platforms

Intel® Xeon® 6 processors with P-cores enhance performance and ROI for AI-enabled cybersecurity. Acceleration technologies built into the platform combine with optimizations across the commercial and open source ecosystem to deliver value from AI-enabled security tools and practices.

The stakes involved in enterprise cybersecurity continue to grow, with increasingly prevalent, sophisticated and damaging attacks crippling businesses and other organizations. As a result, the global average price of a data breach in 2024 has reached nearly $5 million, up 10% since 2023, the largest-ever annual increase.[1] The heightened threat environment comes as attack surfaces expand across workloads that are more distributed than ever and which operate with unprecedented data volume, velocity and variety.

AI adoption — and generative AI (GenAI) adoption in particular — increases an organization's attack surface and is now widely used by attackers to find and exploit vulnerabilities. In response to that escalation, defensive measures are evolving to incorporate AI, which enhances IT security postures by drawing automated insights from operational data at massive scale and high speed. As a result, security teams that use AI and automation extensively for attack prevention have been shown to reduce the cost per breach by more than $2 million on average.[1]

$4.88M
Global average cost of a data breach in 2024[1]

$2.22M
Average savings from security AI and automation[1]

Next-generation server platform for AI and network security (NetSec)

Intel® Xeon® 6 processors with P-cores expand on the capabilities of predecessors for AI-powered security usage models such as network analytics, GenAI semantic analysis and anomaly detection. The unmatched prevalence of Intel processors both on-premises and in public clouds makes them a compelling mainstream target platform for running AI-based security countermeasures without the requirement for add-on hardware.

Hardware acceleration for AI inference makes it possible for software executing on the CPU to inspect the massive amounts of network data generated in production enterprise environments and identify threats dynamically using AI. This ability improves the organization's overall security posture, while realizing the economic advantages of executing on Intel Xeon processors, which are ubiquitous in both data centers and public cloud infrastructure.

Purpose-built hardware accelerators for a wide variety of enterprise workloads, including AI inference, cryptography and in-memory analytics, are built into the Intel Xeon 6 processor. The platform provides open shared infrastructure for AI, security and general-purpose workloads, helping increase the value of capital investments. Extensive hardware features for security as well as optimizations across the open source and commercial ecosystem help extend the performance and cost savings delivered by the balanced Intel Xeon 6 platform:

- **High-throughput AI on the CPU**. Built on the latest Intel 3-nanometer (Intel 3) technology with up to 86 Performance-cores and improvements to instructions and built-in accelerators.
- **Memory enhancements for scale-up security models**. Faster, higher-bandwidth memory with up to eight channels of DDR5-6400 MT/s or 8000 MT/s MCR DIMM memory and a larger L2 cache.
- **High-performance I/O for efficient data movement**. Up to 88 lanes PCIe 5.0, four UPI 2.0 links at up to 24 GT/s and up to 64 lanes CXL 2.0 Type 3 per socket.

The next generation of Intel® Deep Learning Boost (Intel® DL Boost) enhances hardware acceleration for AI workloads on Intel Xeon 6 processors, for outstanding performance across cybersecurity usage models. Intel DL Boost includes some Intel® Advanced Vector Extensions 512 (Intel® AVX-512) instructions as well as Intel® Advanced Matrix Extensions (Intel® AMX). Intel AVX-512 vectorizes processing with 32 registers of 512 bits each, enabling the CPU core to perform the same operation on multiple pieces of data in a single clock cycle.

Intel AMX, a built-in hardware accelerator for the matrix multiplication at the heart of deep learning workloads, builds on a common development history with Intel AVX-512 but delivers up to 8x more operations per clock cycle.[2]

> *"The modern data center is facing unprecedented demands, requiring smarter, more efficient solutions to overcome challenges such as space, power and cooling constraints. At Cisco, we understand these complexities and are committed to optimizing infrastructure — from the data center to the edge — for peak performance and efficiency.*
>
> *Seamless consolidation of diverse enterprise workloads — across cloud-native, virtualized and hyperconverged environments — combined with robust network security, is more critical than ever. Leading organizations like Huntington Bank are choosing Cisco UCS with Intel® Xeon® 6 to scale smarter, faster and more efficiently, with a focus on zero-touch provisioning and automated lifecycle management."*
>
> **– Jeremy Foster, GM and SVP, Cisco Compute**
>
> CISCO

Intel AMX uses 2D register files called tiles to hold larger amounts of data than conventional registers, and Tile Matrix Multiplication (TMUL) operates on those tiles to compute larger matrices in a single operation. With the Intel Xeon 6 processors with P-cores, Intel AMX adds support for FP16 datatypes, which provides the ability to run GPU-trained FP16 models faster on the CPU.

This brief introduces the potential for improved performance across a range of AI-enabled cybersecurity use cases on Intel Xeon 6 processors with P-cores, including extended detection and response (XDR), semantic analysis using large language models (LLMs) and anomaly-based threat detection. The brief uses performance results to help demonstrate the value of Intel's role in the security ecosystem, optimizing components and solutions to enhance performance synergies between hardware and software.

AI-driven data analytics for extended detection and response (XDR)

XDR solutions automate threat detection and response by collecting and analyzing data from wide-ranging sources such as device, network and application telemetry, as well as threat intelligence feeds, security and operations tools and log files from throughout the environment. AI and machine learning dramatically improve XDR capabilities by intelligently correlating that data to help identify threats and attacks, including zero-day exploits. Back-end real-time analytics based on these large datasets are a core requirement for a capable XDR solution.

ClickHouse is an open source, real-time, columnar database that uses SQL queries to automatically perform and report on analytics. It is a popular choice for XDR solutions, with its high-speed operation supporting resource-intensive analytics workloads. ClickHouse can accelerate data analytics with Hyperscan, a high-performance regex matching library developed by Intel that is optimized for performance using Intel AVX-512 technology. Testing with a customer dataset and customer-defined typical XDR test cases demonstrates that enabling Hyperscan results in a 3.19x reduction in search query latency.[4]

Faster XDR Analytics

ClickHouse XDR analytics with Hyperscan versus without Hyperscan

UP TO **3.19x** lower query latency[3]

Novel security usages enabled by large language models (LLMs)

GenAI and LLMs in particular have begun to offer advanced capabilities across enterprise software that approach human capabilities for many semantic tasks, such as generating, summarizing or translating language, including the ability to engage in conversation and perform some reasoning tasks. Common LLMs available in the community include various generations of GPT-J and LLaMA. In the security domain, they hold significant promise for analysis and interpretation of massive data feeds coming into security operations centers from every part of the environment.

LLMs have potential to enhance the entire range of security tools and techniques, from security information and event management (SIEM) platforms to data loss protection and threat-hunting initiatives. For example, they can augment human security teams by investigating incidents, enhancing phishing detection and analyzing the attack surface in real time in response to human-language queries. With potentially trillions of parameters, the scale of LLMs makes optimization critical so they can run with acceptable speed and latency.

Intel® Extension for PyTorch is an open source library that optimizes the popular PyTorch deep learning framework for features and capabilities of Intel Xeon processors, including Intel DL Boost. The library provides comprehensive feature support and optimizations for the LLM domain on Intel® architecture through a dedicated module called ipex.llm. These tools help accelerate inference and reduce the memory footprint across many LLMs and a range of datatypes, to meet the performance and accuracy requirements of enterprise security usage models.

Threat and anomaly detection, enhanced by deep learning

Deep learning models are adept at establishing baseline, expected behaviors on enterprise networks and automatically detecting anomalies and variations from baseline that may indicate malicious behavior. This approach has become common in threat detection and response tools across the cybersecurity landscape. Intel has worked with ecosystem partners to develop and optimize solutions in this space on TensorFlow, ONNX and other popular frameworks. Intel solutions can accelerate deep-learning-based model MalConv for malware detection, and can help address network security use cases, such as preventing command and control (C2) attacks.

Detecting phishing attacks based on email or SMS is an ongoing challenge for enterprise security organizations. Deep learning has proven adept at detecting these advanced attacks, as the basis for anti-phishing capabilities that are now built into many security products. The BERT model provides higher accuracy but requires longer inference time than other models. To boost that performance, Intel Network Builders implements the Hugging Face BERT base model using Intel Extension for PyTorch, resulting in 1.81x faster mean inference time in a processor generation-to-generation comparison[4] and making BERT-based anomaly detection more viable at scale.

Enhanced AI Anomaly and Phishing Detection

BERT Base Model with Intel® Extension for PyTorch (IPEX) Intel® Xeon® 6 Processor vs 4th Gen Intel® Xeon® Scalable Processor

UP TO **1.81x** faster mean inference time[4]

View the latest performance data at www.intel.com/PerformanceIndex

Rigorous security for AI initiatives

As the leader in hardware-based security, Intel has invested significant resources and intelligence in building solutions that help protect sensitive data, applications and infrastructure. These solutions enable organizations to take advantage of new possibilities without compromising data privacy. Forbes ranked Intel #1 on its 2023 “America’s Most Cybersecure Companies” list,[5] and an independent study by ABI Research shows that Intel leads the silicon industry in product security[6].

Intel® Security Engines, built into Intel Xeon processors, deliver enhanced security capabilities that allow even the most sensitive data to be available for AI analysis, training or processing — all while remaining private and confidential. These hardware-enabled security features are designed to control who accesses sensitive data, add layers of protection between your data and threats and protect your platform from the ground up.

Intel Xeon 6 technology strengthens foundational cryptographic primitives by increasing the Advanced Encryption Standard (AES) encryption key size to 256 bits (previously, 128 bits). This addresses Post-Quantum Cryptography (PQC) symmetric encryption requirements set by NIST CNSA 2.0 requirements. Intel® QuickAssist Technology (Intel® QAT) supports AES-256, adding PQC to the foundations of cloud, on-prem, and networking solutions.

Intel confidential computing solutions are designed to protect data in use with isolation, encryption and control, and verification capabilities. Both Intel® Software Guard Extensions (Intel® SGX) and Intel® Trust Domain Extensions (Intel® TDX) now use memory encryption fortified with AES-256, designed for secure scalability by supporting up to 2048 Intel TDX unique keys.

Intel Xeon 6 processor also introduces the first generation of Intel® TDX Connect for select devices, extending Confidential AI solutions to include discrete AI accelerators. Intel SGX, Intel TDX, and Intel TDX Connect also support Local Attestation, allowing customers to deploy trusted confidential workloads across their tailored deployment scenarios.

Silicon-enhanced security and network analytics with Intel® Ethernet

Intel® Ethernet 800 Series Network Adapters, including the new Intel Ethernet E830, when paired with Intel Xeon 6 processors with P-cores, offer a robust foundation for enhancing network security for AI systems, incorporating features such as modern standards-based cryptographic security anchored by a silicon root of trust (RoT) in compliance with NIST SP 800-193 platform firmware resiliency guidelines.

By embedding cryptographic keys and secure identity within the Ethernet adapter, the Intel Ethernet hardware RoT can authenticate and authorize the execution and update of embedded firmware protecting the AI platform from software tampering. Intel Ethernet adapters meet FIPS 140-3 level 1 requirements and offer secure boot capabilities which help maintain the confidentiality, integrity, and availability of critical AI workflows, while safeguarding the platform against sophisticated cyber threats.

Intel Ethernet adapters can help enhance network analytics on Intel Xeon 6 processors for faster thread detection and mitigation by implementing a fully programmable and offloaded packet processing pipeline that can parse and steer specific network packets to individual cores on Intel Xeon 6 processors for decision-making.

Intel Ethernet delivers advanced integrations and rigorous qualifications with industry-leading network data processing frameworks like DPDK and XDP that provide high-performance and highly customizable security and firewall implementations.

By accelerating security-related tasks through programmable hardware, these adapters boost performance and reduce latency. Simultaneously, capabilities like silicon RoT, and open source integrations enable the combined Intel Xeon 6 processor and Intel Ethernet platform to deliver a silicon-enhanced AI security platform that enables organizations to proactively identify and mitigate potential threats while adopting AI.

Conclusion

Intel Xeon 6 processors with P-cores advance the state of AI-powered enterprise security, with a robust balanced platform and built-in hardware accelerators for key workloads. Software optimization and other enablement for the open source and commercial solution ecosystems help deliver the maximum benefit of these advances to the industry, countering the escalating cyberthreat landscape with next-generation capabilities to detect and respond to today's threats and the novel ones yet to come.

Learn More

www.intel.com/xeon

networkbuilders.intel.com

[1] IBM, "Cost of a Data Breach Report 2024." https://www.ibm.com/reports/data-breach.

[2] Intel®, "Architecture Day 2021." https://edc.intel.com/content/www/us/en/products/performance/benchmarks/architecture-day-2021/.

[3] See [7N27] at intel.com/processorclaims: Intel® Xeon® 6. Results may vary.

[4] See [7N26] at intel.com/processorclaims: Intel Xeon 6. Results may vary.

[5] Cybercrime Magazine, June 12, 2023. "Forbes Names America's Most Cybersecure Companies 2023. Intel Ranked No. 1." https://cybersecurityventures.com/forbes-names-americas-most-cybersecure-companies-2023-intel-ranked-no-1/.

[6] ABI Research, 2024. "Embracing Security as a Core Component of the Tech You Buy." https://www.intel.com/content/www/us/en/security/security-as-a-component-of-tech.html.

No product or component can be absolutely secure.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

You may not use or facilitate the use of this document in connection with any infringement or other legal analysis concerning Intel products described herein. You agree to grant Intel a nonexclusive, royalty-free license to any patent claim thereafter drafted which includes subject matter disclosed herein.

The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

0225/DK/MESH/PDF 356896-001US