White Paper

Industry
Cybersecurity

Supermicro Delivers AI-Accelerated Network Security with Intel® Xeon® 6 SoC

Supermicro and Intel deliver an AI-optimized infrastructure combining Intel® Xeon® 6 SoC with advanced hardware acceleration to enable scalable, real-time multimodal cybersecurity and fraud detection across distributed enterprise and edge deployments.

Authors

David Lu
AI Software Architect

Jay L. Vincent
Principal Solution Architect
Intel

Toby McClean
Principal Solutions Architect
Supermicro

Table of Contents

Executive Summary

As enterprises modernize cybersecurity and fraud prevention workflows, they increasingly require AI platforms capable of analyzing multimodal data in real time across text, image, audio, and video transported over the network infrastructure. Supermicro and Intel are enabling this transformation with an optimized infrastructure solution designed for scalable AI inference and security-focused analytics.

This solution combines Supermicro server platforms with Intel Xeon 6 SoC to accelerate both natural language processing and multimodal anti-fraud detection. The platform leverages Intel hardware acceleration features including Intel® Media Transcode Accelerator, Intel® Advanced Vector Extensions 512 (Intel® AVX-512) and Intel® Advanced Matrix Extensions (Intel® AMX) for select AI inference workloads.

The solution demonstrates two benchmark scenarios:

- Bidirectional encoder representations from transformers (BERT) benchmark for transformer-based Natural Language Processing (NLP) inference
- Anti-fraud detection benchmark for end-to-end multimodal fraudulent content detection across text, image, audio, and video

The Opportunity: AI for Network Security and Fraud Detection

Modern security operations centers, service providers, and enterprise platforms face rapidly growing volumes of content and events that must be classified, verified, and scored in real time. Traditional rule-based tools often struggle to keep pace with the sophistication and scale of modern attacks and fraudulent content.

At the same time, AI-enabled security services increasingly depend on multimodal pipelines. One of the biggest threats to AI agents is malicious code and instructions hidden in HTML pages and PDF documents. Fraudulent activity may be embedded not only in text, but also in screenshots, scanned documents, manipulated audio, synthetic video, or combinations of these modalities. Organizations therefore need infrastructure that can support multiple AI services operating together as a coordinated workflow.

Supermicro and Intel address this need by delivering a high-performance platform for multimodal inference and orchestration. The architecture supports flexible deployment from enterprise edge to data center environments and is optimized for low latency, high throughput, and efficient resource utilization. In addition, this type of inferencing is easy to scale and orchestrate independent of specialized accelerator which is essential for reliability.

Solution Overview

Hardware and Software Optimization

This solution is optimized for Intel-based multimodal inference leveraging processor based hardware acceleration.

Intel® Xeon® 6 SoC

Intel Xeon 6 SoC provides the compute foundation for AI inference orchestration and multimodal processing. Key hardware features used in this solution include:

- Intel AVX-512 for vectorized compute acceleration
- Intel AMX for matrix operations in AI workloads
- Intel Media Transcode Accelerator for media processing tasks including video-related workloads

These technologies create a powerful foundation for multimodal AI security analytics by addressing the computational bottlenecks that limit real-time threat detection. Intel AVX-512 processes multiple data elements simultaneously in single instructions, dramatically speeding natural language processing of security logs and audio signal analysis for voice recognition or anomaly detection.

Intel AMX provides dedicated matrix operations through specialized 2D processing arrays that excel at the matrix multiplications central to deep learning models, delivering up to 8x performance improvements[1] for computer vision and transformer-based threat analysis.

The integrated Intel Media Transcode Accelerator offers hardware-accelerated decode/encode that offloads video processing from CPU cores, enabling real-time analysis of multiple surveillance streams without computational overhead.

These technologies work synergistically to create end-to-end acceleration pipelines that can simultaneously correlate threats across text logs, surveillance footage, audio feeds, and sensor data in real-time. For security organizations, this translates to faster threat detection, reduced infrastructure costs, and the ability to process exponentially more data streams with existing hardware investments, ultimately improving security posture while lowering operational expenses.

Supermicro Platform Advantages

Supermicro server platforms provide a flexible hardware foundation for deploying AI workloads in security, networking, and edge environments. Their system-level integration supports scalable deployment of Intel Xeon processors for demanding inference applications.

For benchmarking, Supermicro SYS-112D-36C-FN3P configured with Intel Xeon 6 SoC was used:

Supermicro SYS-112D-36C-FN3P is a short length 1U rackmount edge server consisting of the following:

- One Intel® Xeon® 6553P-B processor (Intel Xeon 6 SoC) with 36 cores, 2.6 GHz base frequency, 144 MB cache with a 235-watt thermal design point
- One 128GB DDR5 6400MHz ECC RDIMM
- One 960GB NVMe PCIe SSD m.2 1DWPD TLC D
- Three 100Gb QSFP28 ports
- Two 800W redundant power supplies

Benchmark 1: BERT Inference Performance

The first benchmark uses a BERT model to showcase real-world security analytics scenarios to demonstrate efficient inferencing for natural language processing (NLP) workloads. The representative use cases include:

- Alert or log message classification
- Security-related text triage
- Suspicious content categorization
- Policy and threat intelligence text analysis

These use cases are highly sensitive to latency for high volume processing of logs and messages from firewalls and intrusion detection applications that require the processing of more than 50,000 log messages per second to identify and categorize anomalies and threats to prioritize responses in a sub 100ms latency.

For Benchmarking Steps, please see the Appendix.

Table 1. BERT Benchmark Result

	Latency(ms) under PyTorch 2.7.1		Latency(ms) under OpenVINO 2024.6.0	
Batch Size	**FP32**	**BF16**	**BF16**	**INT8**
1	120.85	34.64	8.91	9.35
2	240.38	65.58	11.52	11.23
4	478.59	127.79	16.58	14.41
8	956.94	249.84	26.75	22.00

The result is tested on one Intel Xeon 6 SoC's CPU core, with 512 input tokens:

The benchmark results demonstrate compelling evidence for CPU-based inferencing using Intel Xeon 6 SoC in security-critical edge deployments when using a MiniBERT model and OpenVINO optimization. These specialized optimizations achieved 8.91ms of latency with FP32 precision representing 13x improvements over PyTorch 2.7.1. In addition to CPU optimization benefits, gains in memory efficiency are observed when using INT8 quantized models which reduce the memory footprint by approximately 75% compared to FP32 while maintaining improved performance.

Sizing Your Deployment

To achieve optimal latency on a unpredictable pattern of inputs for resource constrained edge deployments we focused on optimizing smaller batch sizes. Based on the test results we were able to achieve 22 ms of latency for a batch size of 8 and a quantization of INT8.

The following table projects total number of tokens that can be processed and the throughput per second as the core count is scaled up for various size deployments.

Table 2. BERT Model Core Scaling and Sizing*

	Cores	Tokens	Throughput (Tokens/Sec)
Baseline	1	512	23,273
Extra Small (Sm. Branch/IOT Device)	20	81,600	3,700,000
Small Location (Med. Branch/Industrial Edge)	40	163,200	7,400,000
Medium Location (Lg. Branch/Factory Floor)	64	261,120	11,900,000
Large Location (Div. HQ/Industrial Plant)	72	294,400	13,400,000
Extra Large (World HQ)	216	884,160	40,200,000

** Table 2 Note:*

Actual performance should be verified through testing with higher core counts and specific platform configuration.

Projections based on:

- *Batch size of 8*
- *Quantization of INT8*
- *OpenVINO Optimizations*
- *Per core latency of 22ms*

Results may vary based on the following factors:

- *Memory Bandwidth limitations as core count increases*
- *Limitations on shared L3 cache size*
- *Numa limitations and locality of memory*

Benchmark 2: Anti-Fraud Detection

The second benchmark is an anti-fraud detection application demonstrating an end-to-end AI pipeline for identifying fraudulent content across multiple data types including:

- Text
- Image
- Audio
- Video

This benchmark is designed to reflect real-world enterprise security and trust-and-safety scenarios where content must be analyzed using multiple coordinated AI services rather than a single standalone model. This multi-modal approach is effectively deployed in CPU's to increase efficiency and reduce the need for a large deployment of hardware accelerators.

The benchmark uses a chained inference workflow across the following five microservices:

- Audio recognition
- Image OCR
- Image classification
- Text classification
- Video decoding

The message detection scheduler orchestrates these services using Envoy and a custom Lua filter to support recursive analysis and service composition.

Anti-Fraud Detection Value

This benchmark demonstrates:

- End-to-end multimodal workflow orchestration
- Efficient use of CPU acceleration using Intel AMX
- Real-time response for complex content analysis
- Flexible deployment of modular AI services
- Practical applicability to enterprise fraud detection and network security environments

The following table illustrates the setup for each test case and the resulting performance indicators.

Table 3. Anti-Fraud Detection Benchmark Result

TestCase	Runtime	CPU Cores	Service Batch Size	Request Per Second	Latency(ms)
vllmxft4corebatch8	vLLM+xFT	4	8	9.3	1719.90
vllmxft8corebatch8	vLLM+xFT	8	8	16.18	989.04
pictureclslcore	OpenVINO	1	4	18.02	444.03
picturecls2core	OpenVINO	2	4	33.53	477.14
paddleocrlcore	OpenVINO	1	1	17.98	111.23
paddleocr2core	OpenVINO	2	1	29.91	133.74
funasrlcore	PyTorch	1	1	2.23	449.27
funasr2core	PyTorch	2	1	3.07	325.88
videolcore	FFmpeg	1	1	0.16	6075.78
video+mediahw	FFmpeg	1	1	0.56	1775.02

The result is tested on one Intel Xeon 6 SoC's CPU core:

- **vLLMxft**: result is reasonable and scales per the number of CPU cores available, accelerated by Intel AMX through intel xFasterTransformer+ vLLM. **(baichuan2-7B INT8, input is 40 token text)**
- **Picture classification:** result is reasonable and scales per the number of CPU cores available, accelerated by Intel AMX through OpenVNIO. **(clip-vit-base-patch16, image size: 4KB, 96*96)**
- **OCR:** result is reasonable and scales per the number of CPU cores available, accelerated by Intel AMX through OpenVNIO. **(PaddleOCR v4, image size: 304KB, 72200*1150)**
- **FunASR:** result is reasonable, accelerated by Intel AMX through PyTorch. **(FunASR, Audio size: 131KB,4s)**
- **Video decoding:** Intel Media Transcode Accelerator media processing performance by4× compared to without the use of the accelerator. **(Video size: 4096x2160, 20807kbps, 25frams/second, 17.3 MB, length 7s)**
  - For large video files, the acceleration is even more significant, reaching 5–6×.

The benchmark results demonstrate a compelling use case in enterprise anti-fraud detection, showcasing how Intel Xeon 6 SoC delivers a unified platform capable of processing diverse AI workloads simultaneously. Unlike traditional approaches that sought discrete accelerators for each modality, Intel Xeon 6 SoC successfully handles text classification (16.18 RPS), image recognition (33.53 RPS), OCR processing (29.91 RPS), audio analysis (3.07 RPS), and video decoding (0.56 RPS with hardware acceleration) on a single processor. This consolidated approach eliminates the architectural complexity of orchestrating multiple specialized accelerators while maintaining performance levels suitable for real-world AI-enabled enterprise security operations.

Sizing Your Deployment

Based on the scaling factors derived from the above benchmark results, the following table highlights the core counts required to achieve various performance targets based on the number of requests per second for various deployment sizes and locations.

Table 4. CPU Core Sizing by Location

RPS: Request per Second	Service	Text	Image	OCR	Audio	Total Cores Required
	Scaling Factor (RPS/Core)	2.3	17	15	1.5	
Extra Small (Sm. Branch/IOT Device)	Performance Target (RPS)	11.5	85	75	7.5	20
	Cores Required	5.0	5.0	5.0	5.0	
Small Location (Med. Branch/Industrial Edge)	Performance Target (RPS)	23	170	150	15	40
	Cores Required	10.0	10.0	10.0	10.0	
Medium Location (Lg. Branch/Factory Floor)	Performance Target (RPS)	36.8	272	240	24	64
	Cores Required	16.0	16.0	16.0	16.0	
Large Location (Div. HQ/Industrial Plant)	Performance Target (RPS)	41.4	306	270	27	72
	Cores Required	18.0	18.0	18.0	18.0	
Extra Large (World HQ)	Performance Target (RPS)	124.2	918	810	81	216
	Cores Required	54.0	54.0	54.0		

Why Supermicro and Intel

Supermicro and Intel provide a scalable, efficient, and flexible platform for AI-enabled network security and anti-fraud detection.

The following table uses the core scaling tables (Table 2 and Table 4) to show the ideal Intel Xeon 6 SoC SKUs and the associated Supermicro platform SKUs available for deploying both the Intrusion Detection NLP Mode and the Multi-Modal Anti-Fraud Detection Model for various location sizes.

Supermicro 1U IoT Server

Supermicro 2U IoT Server

Table 5. Platform Sizing for CPU-Based Intrusion Detection and Anti-Fraud Detection*

	Total Cores Required	Target Intel® Xeon® 6 SoC	SoC Configuration	Supermicro Platform	Platform Configurations
Extra Small (Sm. Branch/IOT Device)	20	Xeon 6516P-B	20 Cores 40 Threads 2.3GHz Base Freq. 145 W TDP	Single SYS-112D-20C-FN3P	Single Socket 1U Short Length 4 x DDR5 DIMM Slots 2 x 2.5" NVMe SSDs 2 x 100G Networking 2 x 800W Redundant PS
Small Location (Med. Branch/ Industrial Edge)	40	Xeon 6716P-B	64 Cores 128 Threads 2.3 GHz Base Freq. 305 W TDP	Single SYS-112D-40C-FN8P	Single Socket 1U Short Length 4 x DDR5 DIMM Slots 2 x 2.5" NVMe SSDs 8 x 25G Networking 2 x 2000W Redundant PS
Medium Location (Lg. Branch/ Factory Floor)	64	Xeon 6756P-B	40 Cores 80 Threads 2.5 GHz Base Freq. 235 W TDP	Single SYS-212D-64C-FN8P	Single Socket 2U Short Length 8 x DDR5 DIMM Slots 2 x 2.5" NVMe SSDs 8 x 25G Networking 2 x 2000W Redundant PS
Large Location (Div. HQ/ Industrial Plant)	72	Xeon 6776P-B	72 Cores 144 Threads 2.3 GHz Base Freq. 325 W TDP	Single SYS-212D-72C-FN8P	Single Socket 2U Short Length 8 x DDR5 DIMM Slots 2 x 2.5" NVMe SSDs 8 x 25G Networking 2 x 2000W Redundant PS
Extra Large (World HQ)	216	3 x Xeon 6776P-B	72 Cores 144 Threads 2.3 GHz Base Freq. 325 W TDP	Three SYS-212D-72C-FN8P	Single Socket 2U Short Length 8 x DDR5 DIMM Slots 2 x 2.5" NVMe SSDs 8 x 25G Networking 2 x 2000W Redundant PS

** Table 5 Note:*

Actual performance should be verified through testing with higher core counts and specific platform configuration.

Projections based on:

- *Batch size of 8*
- *Quantization of INT8*
- *OpenVINO Optimizations*
- *Per core latency of 22ms*

Results may vary based on the following factors:

- *Memory Bandwidth limitations as core count increases*
- *Limitations on shared L3 cache size*
- *Numa limitations and locality of memory*

Key benefits include:

- High-performance infrastructure for multimodal AI inference
- CPU acceleration using Intel AMX for balanced workload execution
- Support for real-time service orchestration and response
- Flexible deployment from edge to enterprise data center
- Optimized support for AI model preparation, inference, and monitoring

By combining modular software services with infrastructure acceleration, this solution helps organizations deploy advanced AI capabilities for security analytics, fraud detection, and intelligent content inspection.

The true value proposition emerges when considering the operational realities of enterprise fraud detection systems. Increased complexity of configuring discrete and workload specific acceclerators create significant deployment and operational challenges including multi-vendor hardware management, complex CUDA driver coordination, substantial power requirements (1.2-2.8kW vs 200-400W), and expensive cooling infrastructure. Intel Xeon 6 SoC's unified architecture enables enterprises to deploy comprehensive fraud detection capabilities, reducing total cost of ownership by:

- reducing the expense to purchase accelerators
- reducing the investment for amount of memory required to support accelerators
- reducing system power consumption and cooling required for the systems
- improving space allocation for smaller server form factors, and
- simplifying software platforms and development support required for them.

More critically, the integrated approach allows for dynamic resource allocation between modalities based on real-time threat patterns, a capability that's extremely difficult to achieve with discrete accelerators.

For security applications the need to process individual network packets, log entries, or authentication requests in real-time for IoT security gateways, network appliances, or embedded security systems, a sub-10ms response time provides the necessary response for immediate threat detection and mitigation. The smaller model footprint also enables multiple security models to run concurrently on the same server system, allowing comprehensive threat detection across different attack vectors simultaneously.

For enterprise security teams, this translates to a fraud detection system that can simultaneously analyze suspicious phone calls (audio), verify document authenticity (OCR), classify transaction images, process natural language communications, and examine video evidence—all within a single, manageable platform. The sub-500ms latency for most operations enables real-time fraud prevention, while simplified architecture reduces the specialized expertise required for deployment and maintenance. Rather than managing a complex integration of discrete accelerators, security teams can focus on refining detection algorithms and responding to threats, making Intel Xeon 6 SoC an ideal foundation for enterprise-grade multimodal fraud detection systems that prioritize operational efficiency alongside security effectiveness.

Conclusion

As fraud and security threats become increasingly multimodal and complex, enterprises need AI platforms that can process diverse data streams with speed, accuracy, and scalability. Supermicro and Intel address this challenge with an optimized solution built on Supermicro systems configured with Intel Xeon 6 SoC.

With support for both BERT NLP inference and end-to-end Anti-Fraud Detection workflows, the platform demonstrates how AI microservices, service orchestration, and hardware acceleration can be combined to enable practical, production-ready security AI deployments.

From a total cost of ownership perspective, CPU-based inference presents significant advantages for distributed security deployments. Edge locations do not typically have the power, cooling and cost budgets for power-consuming discrete accelerator based inferencing. A CPU-centric solution will reduce power consumption and expense by reducing the accelerator footprint. Additional savings would also be achieved with the simplified software stack. For organizations deploying hundreds or thousands of edge security nodes across retail locations, branch offices, or industrial facilities, these savings compound dramatically while still delivering the real-time threat detection capabilities essential for modern cybersecurity operations.

Appendix

Code snippets for setting up the BERT Benchmark for efficient inferencing of NLP workloads[DP13.1]:

```
# Download the minibert models
python3 -m netsec _ ai download --model-name google/bert _ uncased _ L-4 _ H-256 _ A-4 --num-labels
2 minibert2cls

# Generate datasets
python3 dataset/bert-2cls/datagen.py

# IPEX optimize minibert2cls model and save as PyTorchTrace
python3 -m netsec _ ai convert -t PyTorchTrace models/minibert2cls/pt/pt -d CPU
python3 -m netsec _ ai compress -c ipex -p bf16 models/minibert2cls/pt/pt -d CPU

# Convert minibert2cls pytorch model to OpenVINO, saved in models/minibert2cls/ov/ov
python3 -m netsec _ ai convert -t OpenVINO models/minibert2cls/pt/pt

# Use nncf to compress minibert2cls OpenVINO model to int8, saved in models/minibert2cls/ov/
ov-nncf-int8
python3 -m netsec _ ai compress -c nncf models/minibert2cls/ov/ov

# Benchmakr the minibert on CPU
numactl -C 1 python3 -m netsec _ ai evaluate -i 100 -b 4 models/minibert2cls/ov/ov -p -d CPU
numactl -C 1 python3 -m netsec _ ai evaluate -i 100 -b 4 models/minibert2cls/ov/ov-nncf-int8 -p
-d CPU
numactl -C 1 python3 -m netsec _ ai evaluate -i 100 -b 4 models/minibert2cls/pt/pt _ traced _ cpu
-p -d CPU
numactl -C 1 python3 -m netsec _ ai evaluate -i 100 -b 4 models/minibert2cls/pt/pt-ipex-bf16 -p
-d CPU
```

Code snippets for setting up the Anti-Fraud Detection Benchmark for multi-modal analytics in a single instance:

```
# Download the Anti-Fraud Detls -ection models
python3 -m netsec _ ai download --model-name baichuan-inc/Baichuan2-7B-Base baichuan2-7b
python3 -m netsec _ ai download --model-name openai/clip-vit-base-patch16 clip
python3 -m netsec _ ai download --model-name funasr _ zh funasr _ zh
python3 -m netsec _ ai download --model-name PaddleOCR paddleocr# Generate datasets
python3 dataset/baichuan2-7b/datagen.py
python3 dataset/clip/datagen.py
python3 dataset/funasr/datagen.py
python3 dataset/paddleocr/datagen.py
python3 dataset/ffmpeg-video/datagen.py

# Fine-tuning is time-consuming. For quick testing, just simply copy the original model (cp -rf
models/baichuan2-7b/pt/pt models/baichuan2-7b/pt/pt _ merged)
cp -rf models/baichuan2-7b/pt/pt models/baichuan2-7b/pt/pt _ merged

# Convert CLIP models to OpenVINO format
python3 -m netsec _ ai convert -t OpenVINO models/clip/pt/pt

# Build and start all Anti-Fraud Detection microservice, text classification select vllmxft,
running on CPU default.
cd ${NETSEC _ AI _ PATH}/applications/anti-fraud-detection./docker _ images _ build.sh build all
vllmxft
./docker _ images _ build.sh run   all vllmxft
cp benchmarking _ script.sh benchmarking.json ./
bash benchmarking _ script.sh
```

Configuration Notes

Intel platforms provide a diverse portfolio of inference engines including OpenVINO, IPEX and xFasterTransformer to unlock full hardware potential, delivering abundant options to match distinct performance benchmarking requirements .

- **xFasterTransformer**
  xFasterTransformer is an exceptionally optimized solution for large language models (LLM) on Intel platforms, refer to: https://github.com/intel/xFasterTransformer#vllm.
- **Intel® Extension for PyTorch***
  Enhance standard Pytorch with up-to-date features optimizations for an extra performance boost on Intel hardware, refer to: https://github.com/intel/intel-extension-for-pytorch.
- **OpenVINO**
  OpenVINO is an open-source toolkit for optimizing and deploying AI inference and boosting deep learning performance on Intel platforms with Intel AVX-512 and Intel AMX instruction, refer to: https://github.com/openvinotoolkit/openvino.

References

[1] What is Intel® Advanced Matrix Extensions (Intel® AMX)?

Learn More

Built to Accelerate - Supermicro and Intel Edge AI

www.supermicro.com

Supermicro SYS-112D-20C-FN3P | 1U IoT Server

Supermicro SYS-112D-40C-FN8P | 1U IoT Server

Supermicro SYS-112D-36C-FN3P | 1U IoT Server

Supermicro SYS-212D-64C-FN8P | 2U IoT Server

Supermicro SYS-212D-72C-FN8P | 2U IoT Server

www.intel.com

Intel® Xeon® 6 Processor

Disclaimers

0726/MK/PDF 369290-001US