Solution Brief

Vendors, Enterprises, Service Providers
Networking and Security ISVs

Palo Alto Networks Automates Cybersecurity with Machine Learning

Palo Alto Networks uses 3rd Gen Intel® Xeon® Scalable processors and Intel software in the cloud for its Cloud Delivered Security Services to boost machine learning and inferencing performance

The stakes are high for enterprise cybersecurity teams because of the fast-evolving and insidious nature of malware attacks. Malware is morphing faster than a human can respond with thousands of new variants emerging daily. Attackers can modify their existing attacks using machine learning (ML)-based automation to create new variants that bypass signature-based security.

Cybersecurity technology and service providers are also using ML to stay ahead of these automatically generated attacks by automatically detecting the new variants, quickly creating a new signature, and pushing that out to all the security devices within minutes. Another key benefit of ML for security is detecting and identifying IoT sensors to verify authenticity.

Palo Alto Networks has built ML capabilities into its Cloud-Delivered Security Services, a suite of eight individual security services that share data and so can be deployed in various combinations to deliver a complete range of cyber security services anywhere in the enterprise.

While ML provides fast response to threats, it is computation intensive and requires significant CPU cycles to deliver a low latency response. ML uses AI inferencing to detect new threats and understand them. With inferencing being the most compute-intensive part of Palo Alto Network's AI pipeline, the company collaborated with Intel to use Intel technologies to optimize the inference performance. Palo Alto Networks used 3rd Generation Intel® Xeon® Scalable processors and Intel ML software frameworks to deliver the desired outcome. The results of this collaboration can be seen in the benchmark test results discussed later in the paper which showed as much as a six times reduction in mean inference time[1].

Palo Alto Networks Cloud Delivered Security Services

Palo Alto Networks offers a comprehensive range of security features. The Cloud-Delivered Security Services are a suite of solutions that work together to create a network effect that better serves its 85,000+ customers by automatically coordinating intelligence and issuing prevention steps to mitigate against known, unknown and evasive threats in real-time.

Each of the Cloud-Delivered Security Services is designed to complement and enhance others in the suite allowing customers to bolster existing cybersecurity defenses with a specific security service, or to deploy a complete cybersecurity system. All the security capabilities aid and support the zero trust model for network security initiatives.

Cloud-Delivered Security Services include:

- **Advanced Threat Prevention:** Help stop known exploits, malware, malicious URLs, spyware, and command and control (C2) attacks, while utilizing industry-first prevention of zero-day attacks.
- **Advanced WildFire:** Improve file safety by automatically identifying known, unknown, and highly evasive malware quickly using the software's extensive threat intelligence and malware prevention engine.
- **Advanced URL Filtering:** Improve safe access to the internet and prevent web-based attacks with the industry's first real-time prevention of known and unknown threats, stopping malicious URLs faster than other vendors.
- **DNS Security:** Gain more threat coverage and stop a significant amount of the malware that abuses DNS for command-and-control and data theft, without requiring changes to network infrastructure.
- **Enterprise DLP:** Minimize risk of a data breach, stop out-of-policy data transfers, and enable compliance consistently across your enterprise.
- **SaaS Security:** With next-generation cloud access security broker (CASB) natively integrated into Palo Alto's SASE, the software offers proactive SaaS visibility, comprehensive protection against misconfigurations, real-time data protection, and best-in-class security.
- **IoT Security:** Safeguard every "thing" and implement zero trust device security instantly with the industry's smartest security for smart devices.
- **AIOPs:** AIOps for NGFW redefines firewall operational experience by empowering security teams to proactively strengthen security posture and resolve firewall disruptions.

Palo Alto Networks Uses ML for Automated Security Response

The Cloud-Delivered Security Services rely on ML to power automated sharing of intelligence across the services, creating a network effect between different facets of security to enhance all of them, from malware prevention to web security, providing a greater ability to thwartDNS attacks and intrusion prevention.

Intel Technologies Implemented

The Intel technologies that are used by the Palo Alto Network ML solutions include capabilities built into the 3rd Generation Intel® Xeon® Scalable processors and specialized software frameworks that run on the CPU. The processor delivers the following capabilities:

Intel® Deep Learning Boost: A group of acceleration features that provides performance increases[2] to inference applications built using leading deep-learning frameworks such as PyTorch, TensorFlow, MXNet, PaddlePaddle, Caffe, and OpenVINO™. The foundation of Intel Deep Learning boost is VNNI, a specialized instruction set that reduces multiple separate instructions into a single instruction.

Intel® Advanced Vector Extensions 512 (Intel® AVX-512): A 512-bit vector processing instruction set that can accelerate performance for demanding workloads and usages like AI inferencing.

In addition to the CPU instruction sets, Intel has software technologies that include optimized ML industry and networking frameworks:

Intel® oneAPI Deep Neural Network Library (oneDNN): oneDNN is an open source cross-platform performance library of building blocks for deep learning applications. The library is optimized for Intel processors, Intel® Processor Graphics and Xe Architecture graphics.

Intel® Neural Compressor: Intel Neural Compressor is an open source Python* library running on Intel processors and GPUs, that deliver unified interfaces across multiple deep-learning frameworks for popular network compression technologies, such as quantization, pruning, and knowledge distillation. The library is available across popular deep-learning frameworks such as TensorFlow, PyTorch, MXNet, and Open Neural Network Exchange (ONNX) runtime.

Traffic Analytics Development Kit (TADK): A collection of optimized libraries and tools covering the needs of a typical end-to-end AI/ML pipeline used in networking applications. TADK has a modular design supporting custom extensions and customer specific libraries to be included in the overall pipeline. TADK also includes sample open-source application integration (NGINX, FD.io VPP, ModSecurity) as well as sample trained models focusing on traffic classification and web application firewall use cases.

The benchmarking data below shows how these capabilities provide performance improvements for Palo Alto Networks Cloud-Delivered Security Services.

Performance data

Figure 1 shows the results of performance testing of Palo Alto Networks machine learning, command and control (MLC2) attack model in Google Cloud Platform (GCP) environment (server and software configurations are in Appendix A and Appendix B). The tests included cloud instance types using the last two generations of Intel® Xeon® Scalable processors. Testing used 32-bit single-precision floating-point format (FP32) both with and without oneDNN turned on and eight-bit integer 8 (INT8) format with oneDNN turned on.

The results in Figure 1 show the mean inference time can be dramatically improved under TensorFlow by applying Intel oneDNN and Intel Neural Compressor. The mean inference time was more than six times faster comparing Saved_Model to INT8 model under TensorFlow 2.7 for GCP n2-std-8 instance with 3rd Generation Intel® Xeon® Scalable processors.

The performance tuning with oneDNN and using the Intel Neural Compressor is easy and straightforward. Intel DL Boost, which contributes to the observed performance improvement, is a standard and universally available feature in 2nd Generation Intel® Xeon® Scalable processors and 3rd Generation Intel® Xeon® Scalable processors without the need to utilize an auxiliary hardware accelerator.

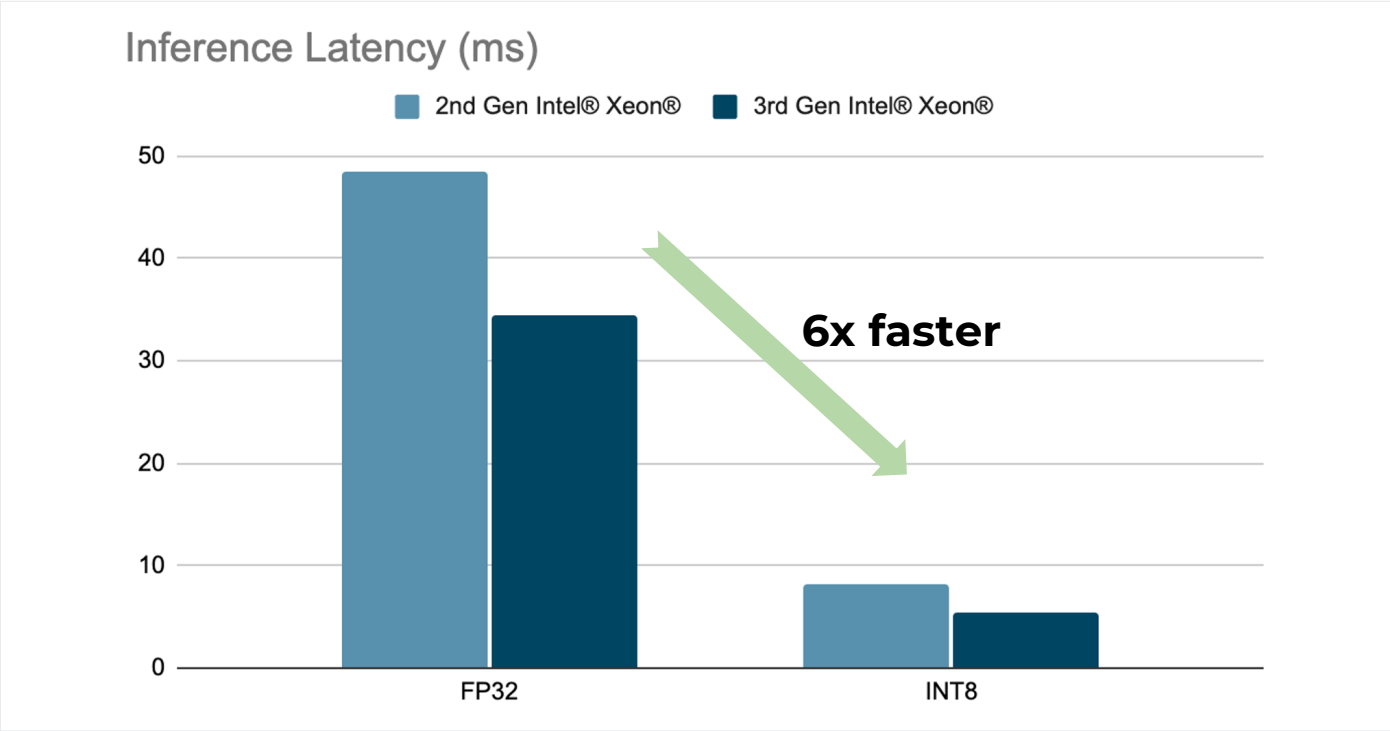

Figure 1. Mean inference time for MLC2 model using servers powered by servers using different generation Intel architecture CPUs (lower is better).

Conclusion

ML technology is a boon to enterprise cyber security teams that want to automate their malware prevention processes to keep up with the high and increasing volume of work. But ML inferencing takes CPU cycles, and this can slow down security application response – defeating the purpose of using the technology. Palo Alto Networks is a leader in using ML technologies and has collaborated with Intel to maximize inferencing performance by using technologies built into 3rd Generation Intel® Xeon® Scalable processor and using Intel ML software frameworks.

Learn More

Palo Alto Networks Cloud-Delivered Services

Intel® Network Builders

Intel® Xeon® Scalable processors

Notices & Disclaimers

[1] See Appendix A for cloud instance configuration.

[2] https://www.intel.com/content/www/us/en/artificial-intelligence/deep-learning-boost.html

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Your costs and results may vary.

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

0223/LV/H09/PDF ♻Please Recycle 354628-001US

Appendix A - Platform Configuration

Name (instance type)	n2-std-8 CLX	n2-std-8 ICX
Time	January 10 06:27:51 UTC 2022	January 10 06:27:51 UTC 2022
Manufacturer	Google	Google
Product Name	Virtual Machine	Virtual Machine
BIOS Version	Google V 1.0	Google V 1.0
OS	Ubuntu 20.04.3 LTS	Ubuntu 20.04.3 LTS
Kernel	5.11.0-1026-gcp	5.11.0-1026-gcp
Microcode	0xffffffff	0xffffffff
IRQ Balance	Disabled	Disabled
CPU Model	2nd Gen Intel® Xeon® Scalable processor @ 2.80GHz	3rd Gen Intel® Xeon® Scalable processor @ 2.60GHz
Base Frequency	2.8GHz	2.6GHz
CPU(s)	8	8
Thread(s) per Core	2	2
Core(s) per Socket	4	4
Socket(s)	1	1
NUMA Node(s)	1	1
Turbo	Disabled	Disabled
Installed	32 GB	32 GB
Automatic NUMA Balancing	Disabled	Disabled

Appendix B - Software Configuration

Software	
Operating System	Ubuntu 20.04.3 LTS
Kernel	5.11.0-1026-gcp
Workload & version	PAN MLC2 model
AI Framework	TensorFlow 2.7.0
Libraries	N/A
WL Specific Details	
	Mean Inference time under the following 1, Saved_model under TensorFlow 2.7 2, FP32 frozen model under TensorFlow 2.7 without Intel oneDNN 3, FP32 frozen model under TensorFlow 2.7 with Intel oneDNN enable 4, INT8 frozen model under TensorFlow 2.7 The WL performance will take advantage of Intel oneDNN and Intel® Neural Compressor which converts FP32 model to INT8 model