Page Contents

Palo Alto Networks: Boosting cybersecurity in the age of generative AI

Get started for free

GOOGLE CLOUD RESULTS

- ✓ Reduces machine learning latency to help clients defend against data loss
- ✓ Provides a more cost-effective environment than previous cloud service provider
- ✓ Simplifies cloud platform management, streamlining workflows, and accelerating troubleshooting
- ✓ Boosts performance of applications that benefit from parallel processing and complex calculations
- ✓ Scales fast to match the needs of an expanding Google Kubernetes Engine environment

By minimizing machine learning latency with Google Cloud and NVIDIA Triton Inference Server, Palo Alto Networks enables its customers to detect and repel cyberattacks in real time.

Defending against data loss by reducing machine learning latency

As technology advances, so do the efforts of cybercriminals to exploit new vulnerabilities. Generative Artificial Intelligence (gen AI), with its ability to quickly create high volumes of realistic content, is no exception.

Palo Alto Networks is a leading provider of cybersecurity solutions, helping businesses protect their systems in the age of gen AI. The California-based multinational company offers a platform that integrates advanced firewalls with cloud-based solutions for comprehensive threat detection and mitigation. In cybersecurity, the speed of threat identification is crucial. Any delay between a suspicious action being detected and a defensive response can allow cybercriminals to overwhelm systems and execute malicious actions before security measures are deployed.

> “Running NVIDIA Triton Inference Servers on Google Cloud has been a game changer for us. Previously, we experimented with other cloud providers and on-prem infrastructure. Going forward, we plan to migrate all our operations to Google Cloud for more streamlined, efficient workflows.”
>
> **Ashwin Kannan**
> Machine Learning Engineer, Palo Alto

Networks

To address this need for real-time threat detection, Palo Alto Networks leverages NVIDIA Triton Inference Server, part of the NVIDIA AI Enterprise software platform, to deploy and execute AI models on NVIDIA GPUs in Google Cloud. These models analyze network traffic, user behavior, and other data sources for signs of malicious activity. By minimizing inference latency — the time it takes for the AI models to process data and detect a threat — Palo Alto Networks aim to significantly reduce the window of opportunity for cyberattacks. This approach enables near real-time threat detection and response, enhancing overall security posture in an increasingly complex digital landscape. The need to process vast amounts of data, detect multiple threats simultaneously, and make split-second decisions across distributed networks necessitates the use of accelerated computing and highly scalable infrastructure. Ashwin Kannan, Machine Learning Engineer, Palo Alto Networks, says, "Having low latency with high throughput is the most critical feature when it comes to building any solution to prevent data loss."

Like many companies looking to deploy ML models, these resources inevitably impact the company's bottom line. Palo Alto Networks constantly looks for more efficient architectures to keep costs down, and found just that in a combination of Google Cloud and the NVIDIA AI platform.

Locking systems down as data traffic shoots up

Before migrating its cybersecurity solution to Google Cloud, Palo Alto Networks used a combination of on-prem servers and a cloud service provider to run its security platform. But as the volume of AI data traffic skyrocketed, scaling this architecture to minimize

“We need to be able to scale at speed and accelerate new machine learning models equipped to latency became increasingly complex and expensive.

Switching to Google Cloud delivered an immediate performance lift. NVIDIA GPU instances on Google Cloud Compute Engine substantially reduced inference latency, enabling Palo Alto Networks to meet the stringent real-time response requirements of its data loss prevention solution. Kannan also calls out the importance of NVIDIA accelerated computing for deep learning and other computationally intensive tasks. "When combined with Google Cloud, the full-stack NVIDIA AI platform, which includes the Triton Inference Server and NVIDIA GPUs create a highly efficient, scalable platform for deploying and managing AI models," he says.

tackle the latest security threats. Google Cloud allows us to add and remove capacity cost effectively as we experience peaks in processing demand.”

Ashwin Kannan
Machine Learning Engineer, Palo Alto Networks

The ability to choose and attach GPUs to instances gave Palo Alto Networks more options compared to other cloud service providers. This flexibility allowed them to customize their infrastructure for specific workloads, optimizing performance and resource utilization. NVIDIA GPUs on Google Cloud, integrated with the AI Hypercomputer supercomputing architecture, enhance productivity for high-intensity AI training, tuning, and service workloads.

As well as increased performance, Google Cloud is more cost-effective than the previous cloud service provider thanks to optimized pricing models and efficient resource allocation.

Kannan and his team also appreciate intuitive and easy-to-use features in Google Cloud that reduce troubleshooting and strengthen workflows. One example is integrating a robust data preparation pipeline into real time inferencing for their Data Loss Prevention platform. The holistic offering of Google Cloud reduces the effort required to manage cybersecurity infrastructure. Further, Palo Alto Networks uses BigQuery to store internal model versioning and analyze insensitive data points.

Kannan is also looking to build, deploy, and scale machine learning models faster with fully managed tools on Google Vertex AI that can be applied to many business uses. "Vertex AI provides a powerful environment for developing and deploying cutting-edge cybersecurity solutions," he says. "Its scalability, integration with Google Cloud services, and access to advanced machine learning tools keep us ahead of evolving threats."

Based on the success of the initial migration, Palo Alto Networks plans to move its remaining workloads, including Google Kubernetes Engine (GKE) production deployments, to Google Cloud in the near future. This will enable the organization to benefit from improved scalability, resource availability, and centralized management.

Kannan says, "With Google Cloud, we can build and adapt quickly to an always-changing cybersecurity landscape, especially the risk of data loss when using foundational models. It's great news for our clients, their employees, and their customers."

Palo Alto Networks is a global cybersecurity company that offers a comprehensive suite of network security products and services to protect organizations against cyber threats.

Industry: Technology

Location: United States

Products: BigQuery, Compute Engine, Vertex AI, Google Kubernetes Engine, AI Hypercomputer

About Google Cloud partner- NVIDIA

Since its founding in 1993, NVIDIA has been a pioneer in accelerated computing. The company's invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined computer graphics, ignited the era of modern AI, and is fueling industrial digitalization across markets. NVIDIA is now a full-stack computing company with data-center-scale offerings that are reshaping industries.

GOOGLE CLOUD PARTNERS

Why Google

- Choosing Google Cloud
- Trust and security
- Modern Infrastructure Cloud
- Multicloud
- Global infrastructure
- Customers and case studies
- Analyst reports
- Whitepapers
- Blog

Products and pricing

- Google Cloud pricing
- Google Workspace pricing
- See all products

Solutions

- Infrastructure modernization
- Databases
- Application modernization
- Smart analytics
- Artificial Intelligence
- Security
- Productivity & work transformation
- Industry solutions
- DevOps solutions
- Small business solutions
- See all solutions

Resources

- Google Cloud Affiliate Program
- Google Cloud documentation
- Google Cloud quickstarts
- Google Cloud Marketplace
- Learn about cloud computing
- Support
- Code samples
- Cloud Architecture Center
- Training
- Certifications
- Google for Developers
- Google Cloud for Startups
- System status
- Release Notes

Engage

- Contact sales
- Find a Partner
- Become a Partner
- Events
- Podcasts
- Developer Center
- Press Corner
- Google Cloud on YouTube
- Google Cloud Tech on YouTube
- Follow on X
- Join User Research
- We're hiring. Join Google Cloud!
- Google Cloud Community

About Google | Privacy | Site terms | Google Cloud terms

Our third decade of climate action: join us

Sign up for the Google Cloud newsletter Subscribe

English ▾