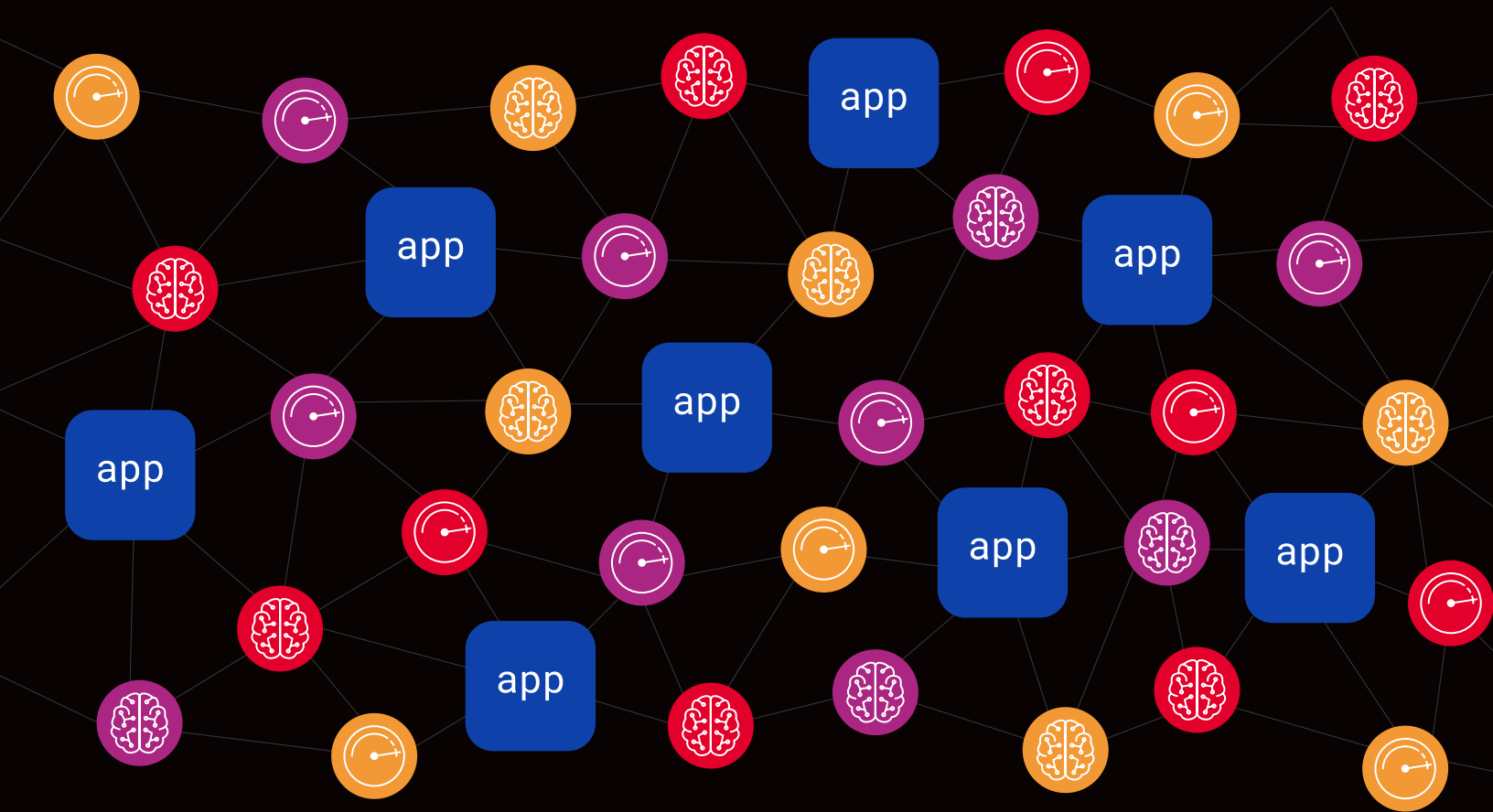# Securely Power Intelligent Apps with AI Inference Using F5 and Intel

Make AI more accessible through lightweight, optimized models. Add the security and availability your Intel® OpenVINO™ model servers need with F5® NGINX® Plus.

Key Benefits

Optimize performance
SSL termination and intelligent load balancing keep AI model servers at peak performance.

Protect against misuse
With NGINX Plus, you can use https and mTLS certificates to protect access to your AI model servers.

Ensure availability
Active health checks and high availability mean inference is ready whenever needed.

Release to production faster
NGINX Plus is lightweight and easy to use during development, so AI models and apps are secure from the start.

Integrate easily
Backed by award-winning support, NGINX documentation is available for deploying with OpenVINO model servers.

Usage of AI Models Is Growing Rapidly

AI adoption in the enterprise is expanding, especially for use in intelligent applications. Gartner predicts that by 2026, 30% of new applications will use AI to drive personalized adaptive user interfaces, up from less than 5% today.[1]

The AI models used to power these applications require high performance and low latency, ideally close to the point of input, to make inferences and deliver responses quickly. Tools are also being developed to make developing and deploying these AI-powered solutions easier, with relatively few lines of code.

Improve AI Accessibility for Developers

Intel® OpenVINO™ is an open-source software toolkit for optimizing and deploying deep learning models for inference tasks. It makes AI more accessible via a smaller footprint, hardware optimization, and a write-once, deploy-anywhere approach.

With lightweight containers, OpenVINO can be used to deploy AI models almost anywhere they're needed: in data centers, clouds, or at the edge to reduce latency. It also ensures computational resources are used efficiently to maintain maximum performance.

OpenVINO model servers require many of the same needs as enterprise apps for use in production environments, such as security and traffic management. F5® NGINX® is an enterprise-ready option to secure OpenVINO model servers.

Make Your AI Models Production Ready

F5 NGINX Plus brings enterprise-grade capabilities to deep learning inference. As a reverse proxy server, it can efficiently route requests, perform SSL termination, and secure traffic distribution. With the potential for hundreds, or even thousands, of upstream containers hosting AI models for inference tasks, NGINX Plus serves as the front end to support availability, linking user applications with model servers.

Enterprise-Grade Security and Performance with NGINX Plus

NGINX Plus offers resiliency and scale for your OpenVINO model servers through content caching, session persistence, and multiple algorithms. Active health checks and intelligent routing ensure that all requests land at an operational server. Convenient dashboards quickly show the status of upstream servers to help you get ahead of issues.

Ensure AI inference is always available by configuring NGINX Plus as a high availability cluster. In active-passive configuration, the primary server actively processes traffic, while the backup monitors health to take over in the event of an issue.
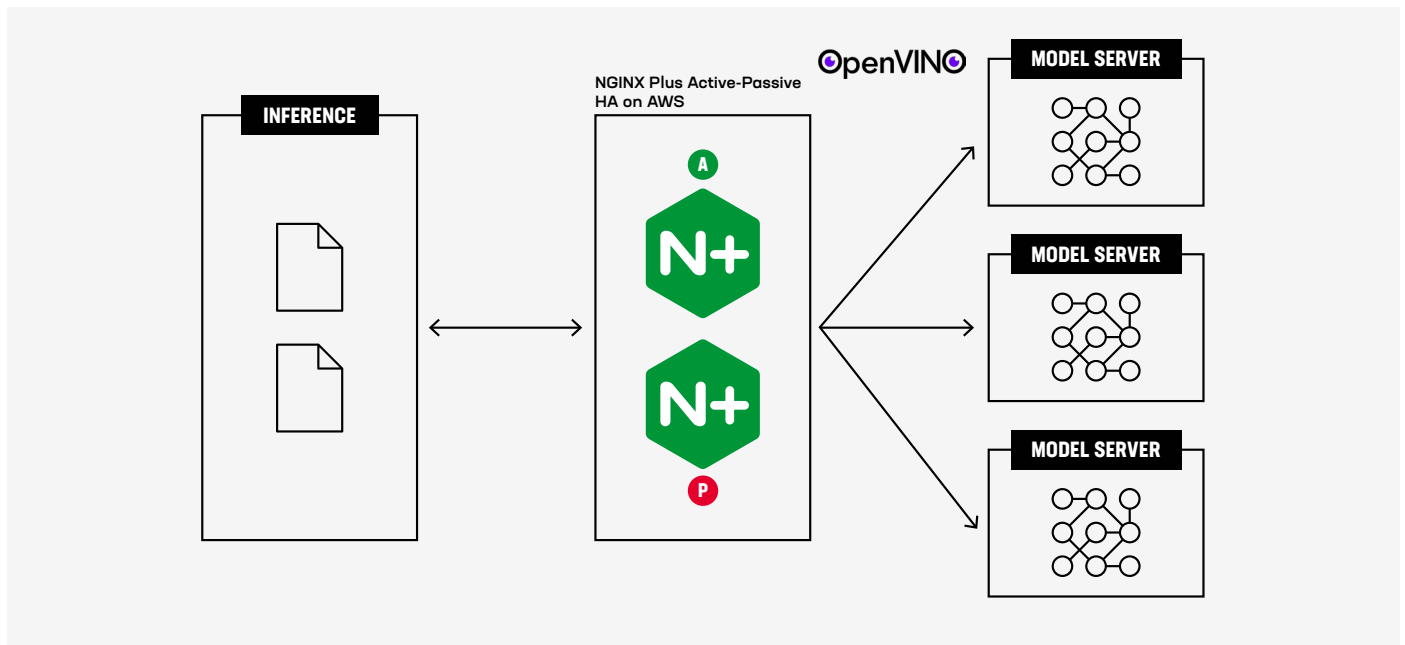
Figure 1: Deploy NGINX Plus in an active-passive pair to offer high availability along with load balancing, active health checks, SSL termination, and content caching.

Performance is also critical for fast AI inference. While OpenVINO offers hardware optimization, NGINX Plus can further increase performance by handling SSL termination. This allows you to use Mutual Transport Layer Security (mTLS)encryption for security without slowing down inference.

NGINX Plus can also help you get your AI models and apps into production sooner. Using a self-service approach, developers can incorporate the lightweight NGINX Plus early in the process to avoid security delays later on. Award-winning NGINX support also assists with installation, deployment, debugging, and more to ensure the best outcomes.

Secure Intelligent Apps with F5 NGINX and Intel OpenVINO

Together, F5 and Intel help you more easily deploy and protect AI models to power intelligent apps in the data center, cloud, or edge. Optimize performance and maintain high availability for reliable operation of your AI models.

Learn more about the F5 and Intel partnership at f5.com/intel.

[1] Gartner, Demand Grows for Intelligent Applications Powered by AI, Sep 2023