Solution Brief

Intel® Infrastructure Processing Unit

Optimizing OpenVINO™ toolkit Model Deployments by Offloading NGINX Plus to Run on Intel® IPUs.

F5 is a premier provider of application delivery and security services for enterprises worldwide. Through innovative software solutions seamlessly integrated with high-performance hardware, F5 optimizes network infrastructure, enhances security, and accelerates application delivery. The cutting-edge technologies ensure top-notch system performance, enabling businesses to maximize server resources and deliver exceptional online experiences reliably and securely.

Authors

Joel Moses
CTO of Systems and Platforms
F5

Tenille Medley
IPU Product Manager
Intel NEX Cloud Computing Group

Swati Mittal
Cloud Solutions Architect
Intel NEX Cloud Computing Group

The advancement of artificial intelligence (AI) technology has spurred financial growth in organizations through a plethora of applications such as AI-optimized workflows, competitively enabled AI products and services, and cost-saving predictive analytics and decision-making. It is critical for AI model developers and independent software vendors (ISVs) to develop and deploy secure AI models that deliver high performance, scalability, and cost effectiveness. The OpenVINO™ toolkit is an open-source software that provides developers with the tools to optimize and deploy deep learning models.

Intel and F5 have partnered to develop a solution that creates a performance-optimized mechanism for model developers, ISVs, and end users to operate independently and securely. F5 NGINX Plus functions as a reverse proxy agent that runs on the Intel® Infrastructure Processing Unit (Intel® IPU) Adapter E2100 to protect OpenVINO toolkit models and data, creating a security zone between the model server and client connection (Figure 1). Crypto operations are managed through OpenSSL in software. The Intel IPU can provide higher performance and enhanced security compared to NGINX Plus running on the host central processing unit (CPU).

This solution is developed on a Dell PowerEdge R760 Server with Intel® Xeon® processors and the Intel IPU Adapter E2100 delivering performance and versatility for compute-intensive workloads. The server is integrated with the Dell iDRAC integrated management controller, which provides closed-loop thermal control of the Intel IPU.

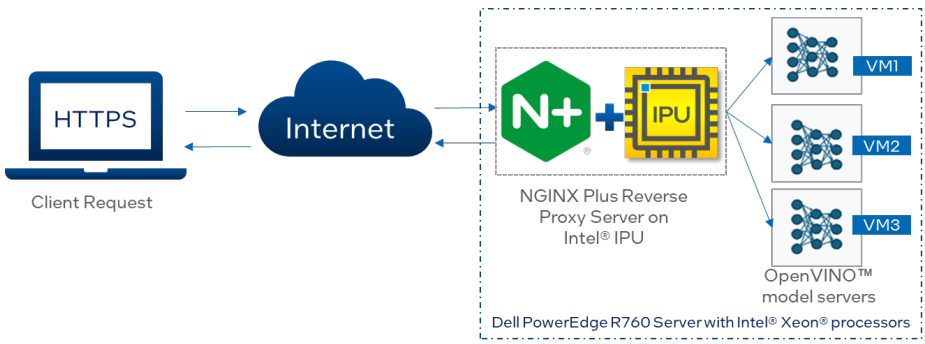

Figure 1. A schematic view of the Intel and F5 joint solution. AI Model Servers deployed on three virtual machines

The OpenVINO toolkit

The OpenVINO toolkit accelerates AI inference with lower latency and higher throughput while maintaining accuracy, reducing model footprint, and optimizing hardware use. It streamlines AI development and integration of deep learning in domains such as computer vision, large language models (LLM), and generative AI.

Secure and Independent Deployments

Secure and independent AI model deployments enable developers to protect sensitive and proprietary data and mitigate security vulnerabilities that can disrupt business continuity. The NGINX Plus reverse proxy server aids in secure and independent AI OpenVINO model server deployments by facilitating traffic routing, SSL/TLS termination, load balancing, security enforcement, performance optimization, and monitoring. By leveraging these capabilities, organizations can ensure the confidentiality, integrity, availability, and performance of AI-driven applications while maintaining independence and scalability.

Security Separation Zone for AI

NGINX Plus is a lightweight, high-performance HTTP and reverse proxy/web server based on a Berkeley Software Distribution (BSD)-like license. NGINX Plus establishes a security separation zone by functioning as an intermediary between external requests (client connection) and AI models (model server). It implements features such as access control and authentication and provides TLS termination to encrypt communication between the user application and the server. These features ultimately safeguard confidentiality, enhance security in AI deployment models, and shield data from potential threats and vulnerabilities.

Secure AI Model Deployment

Secure AI model deployments are achieved through multiple tenets of the OpenVINO toolkit ecosystem. The OpenVINO toolkit, enables developers and enterprises to rapidly optimize and deploy deep learning models and accelerate AI workloads using Intel and third-party hardware. With the joint F5 and Intel IPU solution, the OpenVINO model server is used to bolster security measures and maintain protection against potential vulnerabilities while facilitating and simplifying the deployment of AI models.

Use Cases

The NGINX Plus reverse proxy server and OpenVINO model server architecture enable applications to leverage their infrastructure efficiently across diverse environments such as edge computing, content delivery networks and microservices architecture.

By implementing NGINX Plus routing and security capabilities alongside OpenVINO toolkit-optimized AI algorithms, organizations can capitalize on their infrastructure's intelligence, efficiency, and resilience while delivering innovative AI-driven experiences to their users.

IPUs provide an isolated execution environment separate from the host's CPU execution environment by offloading the infrastructure processing done by NGINX Plus from the host to the IPU. A combined IPU/Host solution makes a strong case for deployment of AI edge applications such as the following:

- EV charging stations serving as a data aggregation and analysis points-of-presence for the copious amounts of data generated by electric car sensors (especially cameras) while charging is performed.
- Medical diagnosis via high-definition video and audio using AI models to determine whether critical care patients are experiencing observable symptoms and alerting critical care providers, enabling a larger number of patients to be supported by fewer nurses.
- Computer vision models to automate and improve quality control on high-speed assembly lines.
- Instantaneous collection and analysis of sensor data from commercial airliners local to the arrival airports to reduce the time it takes to turn a flight around.

Architecture without Intel IPU

The NGINX Plus reverse proxy server functions as a gateway to streamline the deployment and management of OpenVINO toolkit AI applications, enhance performance, and fortify security.

Figure 2 illustrates the NGINX Plus reverse proxy server running on the same host server as the OpenVINO model servers, which run on three separate virtual machines. In this implementation, incoming client requests are reverse proxied by NGINX Plus to different OpenVINO model servers on the host. Additionally, all NGINX Plus-related crypto operations are executed on the host CPU.

While this configuration enables efficient and scalable traffic routing, there are limitations with running NGINX and OpenVINO model servers together on the host.

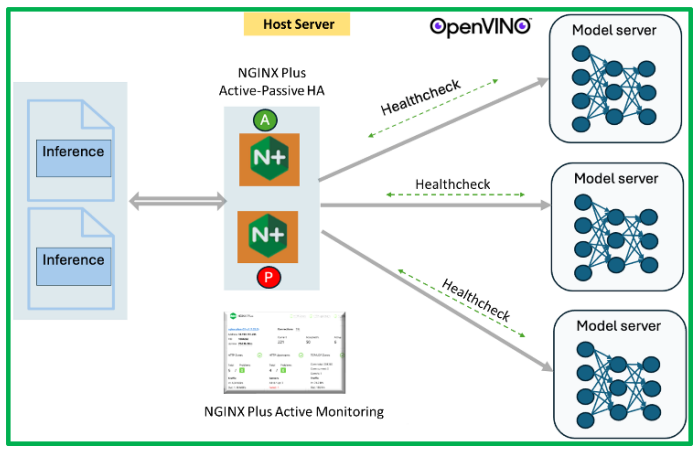

Figure 2. OpenVINO toolkit and NGINX Operating on Host Server

Firstly, running NGINX Plus alongside OpenVINO model

servers on the host CPU, increases CPU overhead by forcing NGINX Plus and OpenVINO model servers to compete for CPU resources, network bandwidth, and memory. This can degrade performance, slow responsiveness, and increase latency, especially for compute-intensive tasks.

Secondly, the host CPU may have limited processing capacity, restricting the scalability of NGINX Plus and OpenVINO toolkit deployments. Adding more virtual machines to accommodate increased demand may incur additional overhead and complexity, potentially limiting scalability and agility.

Thirdly, hosting NGINX Plus and OpenVINO model servers on the same virtual machines exposes them to shared vulnerabilities, leading to data breaches, unauthorized access, or service disruptions.

These limitations highlight the importance of considering alternative deployment strategies such as offloading NGINX Plus operations to a dedicated IPU to optimize system performance, security, and efficiency.

Architecture with Intel IPU

integrating NGINX Plus on the Intel IPU creates a new, secure, and highly capable location for the insertion of application and infrastructure services. Previously, provisioning services of this type required implementation of a physical appliance form-factor, a separate commercial off-the-shelf (COTS) host server – forcing an architecture that either dramatically increases power, environmental, and space requirements, or contends on the host for the same CPU capacity that the hosted application requires.

The F5 and Intel joint solution proposes to create a performance-optimized mechanism for developers to operate independently and securely by integrating NGINX Plus on the Intel IPU in a COTS server configuration as seen in Figure 3. In this implementation, the NGINX Plus reverse proxy server operates on the IPU while the OpenVINO model server runs on the host CPU. All the crypto operations are managed with OpenSSL in software. This creates a security "air gap" between the model server, where the protected IP of the AI model resides, and the incoming client connection.

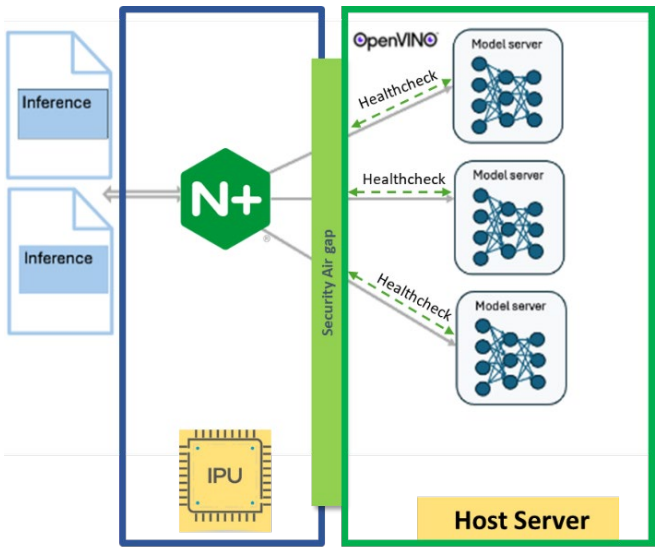

Figure 3. NGINX on Intel® IPU; OpenVINO toolkit on Host Server

The Intel® IPU Adapter E2100 is equipped with hardware accelerators that offer significant performance, security, and scalability benefits to NGINX Plus reverse proxy servers. Because the NGINX Plus reverse proxy server runs on the IPU instead of the host CPU, this solves multiple problems that exist in current AI application deployments:

- Restores the separation of duties between DevOps and NetSecOps and solves an acute organizational issue of domain-of-control for many enterprises.
- Prevents unexpected variations in the compute requirements of the applications from interfering with the infrastructure services and vice-versa while greatly simplifying node size estimation when provisioning and allocating nodes at scale.

The Intel® Infrastructure Processing Unit Adapter E2100

The Intel® IPU Adapter E2100 supports multiple hardened accelerators that deliver high performance, low latency, and better efficiency required in the new generation of data centers. It drives workload acceleration by leverages the IPU's ARM Compute Complex (ACC) to offload NGINX Plus workloads from the host CPU and performs crypto operations using OpenSSL in software. This frees up cycles and provide availability of hardware resources on the host CPU.

The adapter's compute complex has 16 Arm Neoverse N1 cores. These 16 high-frequency cores are backed by a large 32 MB system-level cache and 3x dual-mode LPDDR4x/DDR4 controllers for improved memory bandwidth. These features give the IPU the bandwidth and horsepower to take on large infrastructure workloads.

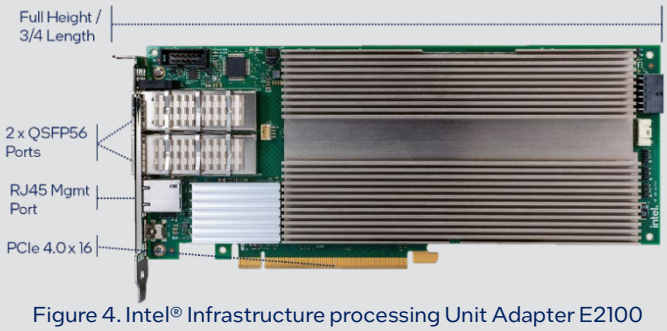

Figure 4. Intel® Infrastructure processing Unit Adapter E2100

Overall, Intel IPUs provide vital benefits such as enhanced security through isolation, accelerated AI workloads, rapid packet inspection, infrastructure optimization, and virtual storage enablement for maximum data center flexibility. By utilizing a combination of acceleration hardware and software running in the compute complex, Intel IPUs enables the rapid innovation necessary for the modern data center.

Summary

The integration of the Intel® IPU Adapter E2100 with NGINX Plus empowers organizations to harness the full potential of OpenVINO toolkit AI model deployments. While NGINX Plus enhances security in AI deployments by serving as a reverse proxy agent, integrating NGINX Plus on the Intel IPU in a COTS server provides an isolated execution environment separate from the host CPU. This solution sets a new standard for secure, scalable, and high-performance AI deployments by addressing limitations with host CPU-only architectures.

The IPU represents a democratization of optimized compute power that can be used anywhere a data inferencing task exists. Combined with the right software stack, AI applications can be distributed safely to the edge, while still being able to offer the same efficiencies and benefits we've come to expect from large-scale datacenter deployments of AI learning systems. To demonstrate how powerful this approach can be, F5 has developed an IPU-deployed example involving NGINX and OpenVINO toolkit for strong AI model protection.

This solution is fully validated on the Dell PowerEdge R760 Server with Intel Xeon processors and can be ordered with the Intel-IPU Adapter E2100 preinstalled. Contact your local Dell representative to learn more.

Contributors

Tony Vo (Intel), Eric Vallone (Intel), Sai Pracheetha (Intel), Paul Pindell (F5), Cyrus Rafii (F5), Sanjay Shitole(F5)

Resources

- Intel® IPU Adapter E2100
- Intel OpenVINO™ toolkit
- F5 NGINX Plus