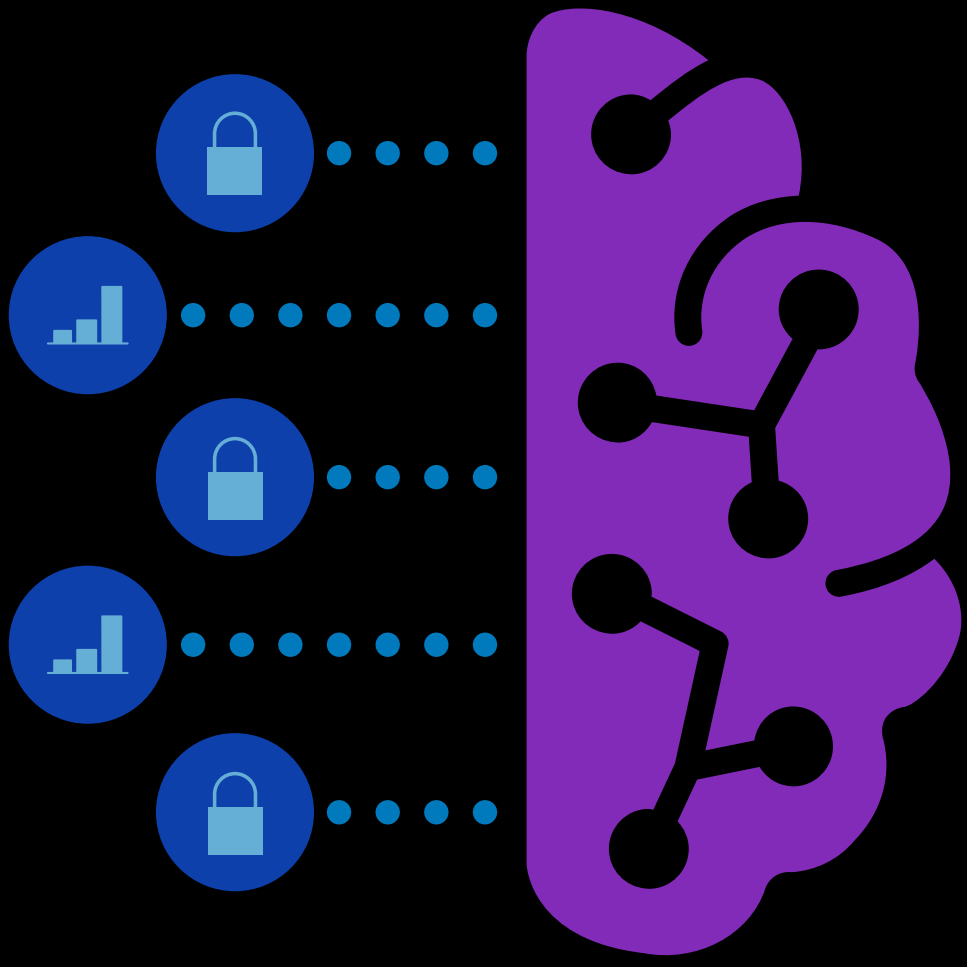
Solution Overview

Driving AI Business Outcomes with Intelligence and Security at Scale

Leverage advanced intelligence and enhanced security to unlock the full business potential of AI infrastructure with F5® BIG-IP Next for Kubernetes deployed on NVIDIA BlueField-3 data processing units (DPUs).



NVIDIA

Key Benefits

Enhanced AI Performance
Accelerate AI workloads with optimized traffic management, faster inference, and efficient resource utilization.

Improved User Experience
Deliver low-latency, high-performance AI services that scale seamlessly with demand.

Streamlined Operations
Simplify AI infrastructure management with centralized traffic control and integration with existing systems.

Future-Ready Scalability
Adapt to evolving AI workflows and workloads with flexible, high-performance infrastructure built for growth.

The Growing Complexity of Cloud-Scale AI Infrastructures

Today's AI-driven world relies on increasingly complex and resource-intensive infrastructures to support large-scale AI applications. AI factories—massive storage, networking, and computing investments built to meet high-volume, high-performance training and inference requirements—are pushing the boundaries of conventional data center capabilities. These environments are tasked with delivering low-latency performance, handling exponential data growth, and scaling to support demanding workloads such as large language models (LLMs), generative AI systems, retrieval-augmented generation, and agentic AI.

However, maximizing business outcomes from AI infrastructure investments comes with significant challenges. Distributed AI systems require precise traffic management to route data efficiently between GPUs and compute nodes, often straining traditional networking solutions. High-performance inferencing, especially for LLMs, demands not only rapid data ingestion but also the ability to dynamically allocate resources across multiple users and workloads. At the same time, securing AI infrastructures has become a critical concern—as open protocols like Model Context Protocol (MCP) gain adoption, data flow and application-layer vulnerabilities create new opportunities for cyber threats to target these environments.

In addition to these security and performance concerns, organizations face difficulties in optimizing resource utilization. GPUs, as one of the most costly and valuable components of AI infrastructure, are central to AI workloads but rely on optimized resource allocation and seamless traffic flows to achieve their full potential. Challenges in managing network and memory resources can limit the scalability of critical applications, delaying ROI and impacting end-user experiences. Addressing these interconnected challenges is essential for organizations seeking to build robust AI factories capable of supporting next-generation AI deployments.

Key Features

Advanced Security Integration
Protect sensitive AI workloads with robust protocol security, zero trust architecture, and comprehensive traffic control.

Optimized GPU Utilization
Maximize infrastructure ROI by dynamically balancing resources and avoiding computational bottlenecks.

Scalable Multi-Tenancy
Confidently support multiple users, workloads, and applications with secure, isolated environments.

Intelligent LLM Routing
Optimize query handling by dynamically assigning tasks to the most suitable AI models for faster, smarter results.

BIG-IP Next for Kubernetes provides a reverse proxy capability that protects MCP servers from threats such as denial-of-service attacks while maintaining seamless traffic flow.

Optimizing Performance, Scalability, and Security for AI Workloads

The growing complexity of large-scale AI workloads requires infrastructure that can intelligently manage traffic, optimize GPU utilization, and safeguard emerging AI protocols. F5® BIG-IP Next for Kubernetes deployed on NVIDIA BlueField-3 DPUs addresses these critical challenges with a suite of advanced capabilities tailored for AI environments. By enhancing performance, scalability, and security, this combined solution enables organizations to scale AI infrastructures with efficiency, confidence, and control.

Smarter LLM Routing with Dynamic Load Balancing

Dynamic LLM routing ensures that AI-related queries are intelligently assigned to the most appropriate language model based on task complexity and resource availability. Lightweight tasks are routed to simpler, cost-efficient LLMs, while advanced models handle more demanding and complex queries. Additionally, domain-specific models can be prioritized to elevate the accuracy and quality of results. By programming routing rules directly onto the DPU, the solution minimizes latency, optimizes time to first byte, and efficiently distributes workloads to reduce computational strain.

Optimizing GPUs for Distributed AI Inference

For large-scale distributed AI frameworks, the integration of BIG-IP Next for Kubernetes with NVIDIA Dynamo introduces significant performance optimizations. The NVIDIA Dynamo KV Cache Manager intelligently routes requests and leverages key-value caching to accelerate generative AI and reasoning models. By offloading these operations from GPUs and CPUs to DPUs, organizations achieve faster inference operations, reduced computational overhead, and greater cost efficiency. These improvements also maximize throughput, enabling AI infrastructure to scale efficiently while delivering faster response times for resource-intensive inference workloads.

Enhanced Security for MCP Servers

As AI protocols such as MCP standardize the delivery of context for LLMs, security and resilience become critical requirements. BIG-IP Next for Kubernetes provides a reverse proxy capability that protects MCP servers from threats such as denial-of-service attacks while maintaining seamless traffic flow. The solution's programmability through F5 iRules allows organizations to quickly adapt to evolving AI protocol requirements, ensuring robust defenses for sensitive workloads. Together, these capabilities secure AI deployments while supporting the scalable and high-performance use of MCP and similar protocols for evolving AI use cases.

Unlocking the Future of Scalable and Secure AI Infrastructure

The demands of AI factories and cloud-scale infrastructure deployments are unprecedented, requiring organizations to address critical challenges in traffic management, resource optimization, and security. BIG-IP Next for Kubernetes deployed on NVIDIA BlueField-3 DPUs delivers a transformative solution tailored to these complex needs. By enabling smarter LLM routing, accelerating distributed AI inference, and fortifying security for emerging protocols like MCP, this integrated solution empowers enterprises to fully capitalize on their AI investments.

As AI workloads grow in complexity, the ability to dynamically scale infrastructure, optimize GPU utilization, and secure sensitive applications becomes essential to maintaining competitive performance and reducing operational costs. With capabilities that ensure high-performance data ingestion, low-latency response times, and robust security, this solution is purpose-built to support the next generation of enterprise AI applications. The collaboration between F5 and NVIDIA allows organizations to redefine AI operations, delivering greater ROI and ensuring scalability for future growth.

Next Steps

Multi-cloud is here to stay. If you want to maximize the benefits and minimize the risks, then putting together the people, processes, and technology to deliver consistent, high-quality application performance and security services in all your private and public cloud locations is a smart move. Partnering with F5 to do it is an even smarter one.

Contact F5

Deploying NVIDIA Accelerated Computing? Find out more about how F5 works with NVIDIA BlueField-3 DPUs and enables you to achieve greater efficiency, performance, and security for AI workloads. **Contact us**.

What Is an AI Factory?

Amidst the AI technological evolution, the concept of an AI factory has emerged as an analogy for how AI models and services are created, refined, and deployed. **Read the article**.

F5 Solutions, Accelerated by NVIDIA

F5 taps into NVIDIA technologies to create AI infrastructure solutions that provide application delivery and security for AI models and apps to help scale accelerated computing. **Explore the collaboration**.