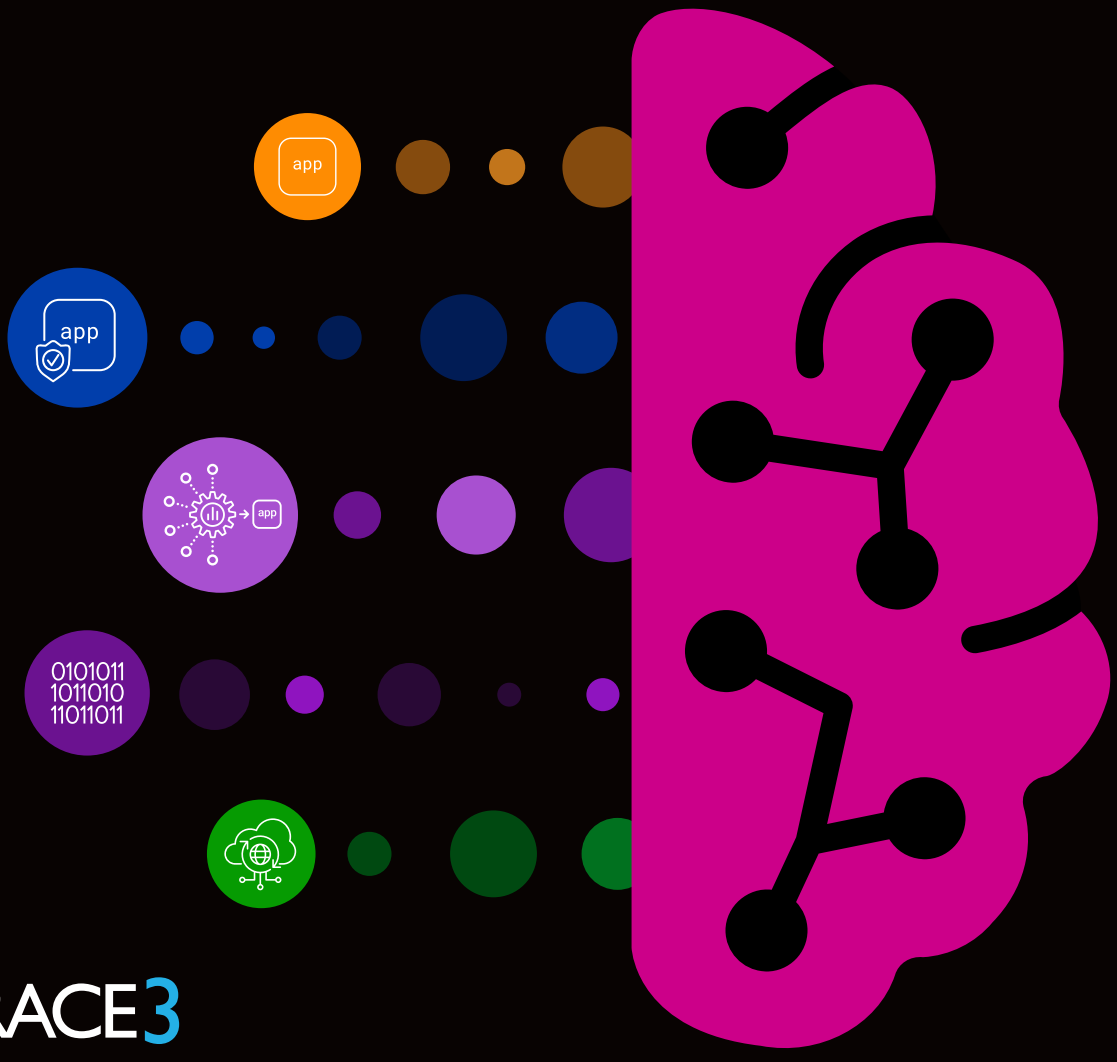
Solution Overview

Accelerating and Optimizing Networking for AI Infrastructure

Streamline the delivery of AI workloads and maximize the efficiency of your AI networking infrastructure with F5 BIG-IP Next for Kubernetes deployed on NVIDIA BlueField-3 DPUs.



TRACE3

Key Benefits

Maximize AI Infrastructure Efficiency
Achieve increased AI computing efficiency required for scaling AI factories.

Optimize AI Infrastructure Investment
Accelerate ROI by supporting multiple customers and use cases on your AI infrastructure.

High-Performance Networking
Meet the performance demands required by large-scale AI container networking infrastructure for training, inference, and retrieval-augmented generation (RAG).

Centralized Security
Provides a single point for network security control for data flowing in and out of your AI infrastructure.

Maximizing AI Infrastructure Investments Starts with the Network

Accelerated computing is at the heart of numerous new AI use cases, including AI factories, AI SaaS, edge AI, and sovereign AI. These modern AI workloads have unique and high-volume demands on the underlying infrastructure. Massively parallel graphics processing unit (GPU)-accelerated AI computing requires high-volume, low-latency networking between users, GPUs, memory, and data storage for training and inferencing. Accelerated networking solutions play a pivotal role in achieving peak AI workload efficiency. The network supporting AI infrastructure must intelligently map traffic to AI resources while being highly performant, low latent, and secure. The absence of any of these capabilities could lead to poorly performing AI models and AI apps or to security breaches. It could also lead to the underutilization of GPUs, resulting in lower ROI for organizations deploying AI infrastructure.

AI use cases, such as AI SaaS and sovereign AI, require infrastructure to support multiple customers and users to realize maximum return on their GPU investment. Multi-tenancy support enables the delivery and scale of AI use cases.

Networks must provide isolation between customers, users, and workloads while delivering flexible and high-performance connections. Securing AI infrastructure and AI workloads will be top of mind for security operations teams as AI models and apps will attract bad actors and new security threats. Networks must integrate critical security features and zero trust architecture, including a network firewall, DDoS mitigation, web application firewall (WAF), API protection, intrusion prevention, encryption, and certificate management to provide AI models and apps isolation from threats.

As AI evolves, purpose-built networks will play a more prominent role and impact the successes or failures of AI for businesses. Networking for AI requires high-performance traffic management, security, and multi-tenancy to meet the demands of AI use cases.

Key Features

Shift CPU Resources
Offload network traffic management, load balancing, and security features onto NVIDIA BlueField-3 DPUs, freeing up valuable CPU resources.

Multi-Tenancy Support
Confidently host multiple users and AI use cases, providing network and customer isolation for AI applications, enabling efficient deployment across an AI infrastructure.

DPU-Driven Zero Trust
Manage critical security features and establish zero-trust architecture, including firewall, DDoS mitigation, API protection, intrusion prevention, encryption, and certificate management on the programmable BlueField-3 DPU.

Simplified Operations
Provides a central point for managing networking and security for AI infrastructure, greatly simplifying operations.

F5 Networking for AI Workloads

AI cloud infrastructure primarily leverages Kubernetes within deployed clusters to support varying workloads, multiple tenants, and automation. Kubernetes performs well for simple web apps but, at the core, are not designed for complex network requirements (e.g., multiple tenants, non-HTML protocols). Infrastructure teams today may attempt to piece together solutions to solve limitations within Kubernetes networking. While this may perform to satisfy proofs of concept, deploying at scale results in operational inefficiencies.

F5® BIG-IP® Next™ for Kubernetes is purpose-built software resolving high-volume Kubernetes networking challenges posed by accelerated compute workloads. BIG-IP Next for Kubernetes deployed on NVIDIA® BlueField®-3 data processing units (DPUs) now adds powerful new functionality needed for AI workloads and brings high-performance traffic management and security to large-scale AI infrastructure. When deployed, enterprises gain greater data ingestion performance, GPU utilization during model training, and better user experience from performance during inferencing with RAG. BIG-IP Next for Kubernetes seamlessly integrates into existing data center networking and is deployed on NVIDIA BlueField-3 DPUs, freeing CPU cycles for revenue-generating applications.

For cloud-native AI infrastructure and AI factories, BIG-IP Next for Kubernetes running on NVIDIA BlueFIeld-3 DPUs scales networking and security services as a central control point, including advanced traffic management, routing, and security (e.g., firewall, DDoS, and API protection).

Ensure Your AI Infrastructure Is Running at Optimal Performance

Building AI infrastructure is costly, and enterprises investing in AI must maximize their investments. However, traditional networking deployed in modern data centers does not satisfy the demands of new and upcoming AI workloads, potentially leading to inefficient usage of valuable GPU servers. Networks supporting AI infrastructure must be optimal to reduce underutilized GPUs. BIG-IP Next for Kubernetes integrates with NVIDIA BlueField-3 DPUs to deliver technology advancements and ensure AI infrastructures operate at optimal performance. BIG-IP Next for Kubernetes delivers high-performance networking and security for AI Infrastructure, addressing the need for higher efficiency and simplifying operations through multi-tenancy support and a central point of control for networking and security.

Reducing Hardware Footprint and Energy Consumption

Intelligent load balancing, routing, and security services are necessary for improving compute efficiency and securing AI workloads. Physical hardware appliances could be used to deliver these services but may not meet all requirements, like maximizing GPU utilization, and their collective power consumption could be greater than BIG-IP Next for Kubernetes deployed on NVIDIA BlueField-3 DPUs. For example, a hardware appliance delivering 200Gbps of throughput for load balancing, routing, and security functions consumes between approximately 1,000 and 1,200 watts.[1] With BIG-IP Next for Kubernetes deployed on NVIDIA BlueField-3 DPUs, separate appliances are no longer required, and power consumption for the same functions is reduced to approximately 160 watts.[2] This decrease in energy consumption translates to substantial energy savings for large-scale AI infrastructure deployments for these networking services.

Additional Resources Available Online

Deploying NVIDIA GPUs at Scale? Contact F5

Find out how F5 works with Bluefield-3 DPUs and enables you to achieve greater efficiency, performance, and security for AI workloads.

What Is an AI Factory?

Amid 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.

F5 Solutions, Powered by NVIDIA

F5 taps into NVIDIA technologies to create AI infrastructure solutions, which provide application delivery and security for AI models and apps to scale accelerated computing.

To learn more about Trace3 and F5, please visit trace3.com.

[1, 2] Energy consumption numbers and estimates are based on F5 internal testing with NVIDIA BlueField-3 DPUs and F5 rSeries hardware appliances in 2024.