Solution Overview

Powering GPUaaS and AI Inferencing Services with F5 and NVIDIA

F5 and NVIDIA optimize AI infrastructure by enhancing performance, efficiency, and security for GPUaaS and AI inferencing services. Overcome the challenges of traditional data centers using advanced networking and traffic management with F5® BIG-IP Next® for Kubernetes deployed on NVIDIA® BlueField™-3 DPUs.

Key Benefits

Maximize AI Infrastructure Efficiency
Achieve increased AI computing efficiency required for scaling AI factories and cloud data centers.

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 agentic AI.

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

The Growing Demand for Robust and Secure AI Infrastructure

As enterprises and service providers increasingly adopt artificial intelligence (AI) to drive innovation and improve operational efficiency, the need for robust and scalable AI infrastructure has become paramount. The sheer volume of data generated and the complexity of AI models place significant demands on data center and AI factory resources. Traditional network infrastructures often struggle to handle the high data throughput and low-latency requirements essential for AI workloads. This necessitates the adoption of advanced infrastructure solutions that can meet these stringent demands.

Scaling AI deployments in conventional data centers often results in inefficiencies that delay returns on infrastructure investments. These inefficiencies stem from high latency, performance bottlenecks, and the inability to optimize massive parallel computations required for AI model training and inferencing. Additionally, integrating AI workloads into existing IT environments can amplify operational complexities and resource constraints, further reducing the speed at which organizations can monetize their AI infrastructure. Addressing these challenges is crucial to maximizing GPU utilization, enhancing efficiency, and ultimately accelerating ROI for AI initiatives.

Additionally, security is a critical concern for AI infrastructure. AI applications are prime targets for cyber threats due to the sensitive data they process and the critical business functions they support. Ensuring robust security measures, such as network isolation, data encryption, and threat mitigation, is essential to protect AI workloads from potential vulnerabilities. Addressing these challenges requires a comprehensive approach that combines high-performance computing, efficient data management, and stringent security protocols.

Key Benefits

Maximize 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 NVIDIA BlueField-3 DPUs.

Greater Observability
Provides a central point for collecting valuable metrics on ingress and egress traffic, log collection with seamless integration of log ingestion solutions, and provides administrators and engineers the ability to do real time traffic captures, visibility to resource utilizations and statistics.

Enabling Scalable GPUaaS and Efficient Inferencing

F5 BIG-IP Next for Kubernetes deployed on NVIDIA BlueField-3 DPUs offers a cutting-edge solution to the challenges faced by large-scale AI infrastructures. This integration provides high-performance networking, security, and traffic management capabilities specifically designed to optimize AI workloads. By leveraging the NVIDIA BlueField-3 DPUs, F5's solution enhances the efficiency and performance of AI applications, enabling organizations to maximize their infrastructure investments.

The combined solution of F5 BIG-IP Next for Kubernetes and NVIDIA BlueField-3 DPUs significantly improves data center resource management. It accelerates data traffic flow in and out of AI infrastructures, ensuring GPUs are utilized to their full potential. This enables enhanced data ingestion performance and optimized server utilization during AI model training, fine-tuning, and inferencing. The solution also supports critical security features such as zero-trust architectures, distributed denial-of-service (DDoS) mitigation, and API protection, ensuring that AI applications are secure from potential threats.

A significant feature of this integrated solution is support for GPU-as-a-Service (GPUaaS) and inferencing services. GPUaaS provides on-demand access to GPU resources. Inferencing services optimize the deployment of trained AI models for real-time applications, such as predictive analytics and natural language processing. The granular multi-tenancy and dynamic resource allocation capabilities of BIG-IP Next for Kubernetes ensure that these services are delivered efficiently and securely, catering to the diverse needs of multiple users and workloads.

Delivering Secure, Responsive AI Applications

The integration of F5 BIG-IP Next for Kubernetes with NVIDIA BlueField-3 DPUs offers a powerful solution for organizations looking to optimize their AI infrastructures. By addressing the critical challenges of performance, efficiency, and security, this solution enables organizations to fully leverage the potential of AI technologies. The high-performance, accelerated networking and traffic management capabilities ensure AI workloads are processed efficiently, while the robust security features protect sensitive data and applications from potential threats.

The support for GPUaaS and inferencing services further allow organizations to deploy access to GPU resources on demand, scale their AI capabilities seamlessly, and deploy AI models for real-time applications with minimal latency. This not only reduces operational costs but also improves overall user experiences, making AI applications more responsive and effective.

As AI continues to evolve and drive innovation across various industries, having a robust and scalable infrastructure solution is essential. F5 BIG-IP Next for Kubernetes deployed on NVIDIA BlueField-3 DPUs provides the performance, security, and flexibility needed to support the growing demands of AI workloads. By adopting this solution, organizations can stay ahead of the curve, achieve their AI objectives, and deliver exceptional value to their customers.

Next Steps

Contact F5

Deploying NVIDIA Accelerated Computing? Find out how F5 works with 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, which provide application delivery and security for AI models and apps to scale accelerated computing. Explore the collaboration