Mar 17, 2026

F5 and NVIDIA Advance AI Factory Economics With New Capabilities for Accelerated AI Inference

F5 BIG-IP Next for Kubernetes accelerated with BlueField DPUs improves token throughput, reduces...
F5 BIG-IP Next for Kubernetes accelerated with BlueField DPUs improves token throughput, reduces cost per token, and enables secure multi-tenant AI infrastructure, transforming AI factories for the agentic era

SEATTLE--(BUSINESS WIRE)-- F5 (NASDAQ: FFIV), the global leader in delivering and securing every app and API, today announced expanded capabilities in its ongoing collaboration with NVIDIA to accelerate and optimize AI inference infrastructures.

The expanded integration combines F5 BIG-IP Next for Kubernetes with NVIDIA BlueField-3 DPUs, creating an intelligent, telemetry-aware infrastructure layer that increases token throughput with better GPU utilization, reduces latency, and enables secure multi-tenant AI platforms at scale.

In AI systems, tokens represent the measurable unit of AI output—the words, symbols, or data fragments generated and processed during inference. The volume and velocity of token production ultimately determine user experience, infrastructure efficiency, and revenue per accelerator.

As enterprises and GPUaaS providers race to monetize AI and move from AI experimentation to revenue-generating services, infrastructure efficiency has become a defining metric. Success is increasingly measured not simply by deployed GPU capacity, but by token economics, sustained token throughput, time to first token (TTFT), cost per token, and revenue per GPU accelerator. The F5 and NVIDIA joint solution is designed to directly address these metrics.

Optimizing tokenomics through intelligent AI infrastructure

The shift from application-centric inference to agent-driven AI workflows demands new architectural approaches to optimize token throughput and reduce costs. BIG-IP Next for Kubernetes now leverages NVIDIA NIM statistics, Dynamo runtime signals, and GPU telemetry to make inference-aware routing decisions before execution. By matching workloads to the most appropriate accelerators in real time, the solution increases sustained utilization while reducing latency and re-compute.

“AI infrastructure is no longer just about access to GPU or scaling their deployments. It has evolved into maximizing economic output per accelerator,” said Kunal Anand, Chief Product Officer, F5. “Together with NVIDIA, we are enabling AI factories to treat token production as a measurable business metric. BIG-IP Next for Kubernetes provides the intelligence and governance required to increase GPU yield, reduce cost per token, and scale shared AI platforms confidently.”

Validated infrastructure efficiency: A structural uplift

The performance numbers speak for themselves. In testing validated by The Tolly Group, BIG-IP Next for

The performance numbers speak for themselves. In testing validated by The Tolly Group, BIG-IP Next for Kubernetes, accelerated by NVIDIA BlueField-3 DPUs, delivered up to a **40% increase in token throughput**, a **61% faster time to first token (TTFT)**, and a **34% reduction in overall request latency**.

These are not incremental gains. By offloading networking, TLS/encryption, AI-aware load balancing, and traffic management to NVIDIA BlueField-3 DPUs, BIG-IP Next for Kubernetes preserves host CPU capacity and frees GPUs to do what they were built for: sustained, high-throughput inference at scale. The result is improved GPU utilization, reduced queuing delays, and increased token yield—enabling lower cost per token within a fixed infrastructure footprint. Critically, no model modifications were required, making these gains immediately deployable across existing AI factory infrastructure. For enterprises and NeoCloud providers competing on token economics, this is the difference between infrastructure that constrains AI output and infrastructure that accelerates it.

“NVIDIA’s accelerated computing infrastructure coupled with F5’s AI-aware Application Delivery and Security Platform unlocks superior AI factory tokenomics—delivering scalable and cost-effective inference without making any changes to the models,” said Kevin Deierling, SVP, Networking, NVIDIA. “Together, F5 and NVIDIA are empowering enterprises to scale AI factory inference efficiently and economically.”

Built for agent-driven AI and multi-tenant AI platforms

Modern AI workloads are increasingly agent-driven, persistent, and context-aware. They demand intelligent traffic control that traditional load balancing cannot provide. The enhanced BIG-IP Next for Kubernetes solution can now support:

Inference-aware routing for agentic AI workflows
Integration with NVIDIA DOCA Platform Framework (DPF) to simplify NVIDIA BlueField DPU deployment and lifecycle management
EVPN-VXLAN with dynamic VRFs for secure network-level multi-tenancy
Integrated security, token governance, and observability within Kubernetes AI environments

These capabilities enable enterprises and NeoCloud providers to securely share GPU infrastructure across business units or external customers while preserving performance isolation and predictable service levels.

A control plane for AI factory economics

F5 and NVIDIA provide enterprises with validated tools and best practices to optimize inference architecture. With these advancements, BIG-IP Next for Kubernetes is positioned to become a strategic control plane for AI factory economics, governing token consumption, optimizing traffic flows, and maximizing infrastructure return on investment.

Rather than overprovisioning to compensate for inefficiencies, organizations can now extract greater economic value from every GPU already in production. The result is improved revenue per GPU, lower operational overhead, and scalable AI services built for sustained growth. By combining NVIDIA’s infrastructure telemetry and DPU acceleration with F5’s traffic intelligence and security capabilities, the companies are helping enterprises transform AI factories into efficient, monetizable platforms ready for the agentic era.

Supporting materials

Blog: AI factories need intelligent infrastructure. New results from The Tolly Group show why.

Report: Independent testing by Tolly: F5 BIG-IP Next for Kubernetes

About F5

F5, Inc. (NASDAQ: FFIV) is the global leader that delivers and secures every app. Backed by three decades of expertise, F5 has built the industry's premier platform—F5 Application Delivery and Security Platform (ADSP)—to deliver and secure every app, every API, anywhere: on-premises, in the cloud, at the edge, and across hybrid, multicloud environments. F5 is committed to innovating and partnering with the world's largest and most advanced organizations to deliver fast, available, and secure digital experiences. Together, we help each other thrive and bring a better digital world to life.

For more information visit f5.com

Explore F5 Labs threat research at f5.com/labs

Follow to learn more about F5, our partners, and technologies: Blog | LinkedIn | X | YouTube | Instagram | Facebook

Forward-looking statements

This press release contains forward-looking statements regarding the expected performance, efficiency, and economic benefits of F5's integration with NVIDIA, which are subject to risks and uncertainties.

Source: F5, Inc.

View source version on businesswire.com: https://www.businesswire.com/news/home/20260317977850/en/

Dan Sorensen
F5
(650) 228-4842
d.sorensen@f5.com

Holly Lancaster
We. Communications
(415) 547-7054
hlancaster@wecommunications.com

Source: F5, Inc.