Jun 11, 2025

F5 Expands Performance, Multi-Tenancy, and Security Capabilities for Fast-Evolving AI Landscape with NVIDIA

Sesterce validation highlights collaborative innovation between F5 and NVIDIA to help customers...

Sesterce validation highlights collaborative innovation between F5 and NVIDIA to help customers embrace the value of AI-first application delivery

PARIS--(BUSINESS WIRE)-- F5 (NASDAQ: FFIV), the global leader in delivering and securing every app and API, today announced new capabilities for F5 BIG-IP Next for Kubernetes accelerated with NVIDIA BlueField-3 DPUs and the NVIDIA DOCA software framework, underscored by customer Sesterce's validation deployment. Sesterce is a leading European operator specializing in next-generation infrastructures and sovereign AI, designed to meet the needs of accelerated computing and artificial intelligence.

Extending the F5 Application Delivery and Security Platform, BIG-IP Next for Kubernetes running natively on NVIDIA BlueField-3 DPUs delivers high-performance traffic management and security for large-scale AI infrastructure, unlocking greater efficiency, control, and performance for AI applications. In tandem with the compelling performance advantages announced along with general availability earlier this year, Sesterce has successfully completed validation of the F5 and NVIDIA solution across a number of key capabilities, including the following areas:

- **Enhanced performance, multi-tenancy, and security** to meet cloud-grade expectations, initially showing a 20% improvement in GPU utilization.
- **Integration with NVIDIA Dynamo and KV Cache Manager** to reduce latency for the reasoning of large language model (LLM) inference systems and optimization of GPUs and memory resources.
- **Smart LLM routing** on BlueField DPUs, running effectively with NVIDIA NIM microservices for workloads requiring multiple models, providing customers the best of all available models.
- **Scaling and securing Model Context Protocol (MCP)** including reverse proxy capabilities and protections for more scalable and secure LLMs, enabling customers to swiftly and safely utilize the power of MCP servers.
- **Powerful data programmability** with robust F5 iRules capabilities, allowing rapid customization to support AI applications and evolving security requirements.

"Integration between F5 and NVIDIA was enticing even before we conducted any tests," said Youssef El Manssouri, CEO and Co-Founder at Sesterce. "Our results underline the benefits of F5's dynamic load balancing with high-volume Kubernetes ingress and egress in AI environments. This approach empowers us

balancing with high-volume Kubernetes ingress and egress in AI environments. This approach empowers us to more efficiently distribute traffic and optimize the use of our GPUs while allowing us to bring additional and unique value to our customers. We are pleased to see F5's support for a growing number of NVIDIA use cases, including enhanced multi-tenancy, and we look forward to additional innovation between the companies in supporting next-generation AI infrastructure."

Highlights of new solution capabilities include:

LLM Routing and Dynamic Load Balancing with BIG-IP Next for Kubernetes

With this collaborative solution, simple AI-related tasks can be routed to less expensive, lightweight LLMs in supporting generative AI while reserving advanced models for complex queries. This level of customizable intelligence also enables routing functions to leverage domain-specific LLMs, improving output quality and significantly enhancing customer experiences. F5's advanced traffic management ensures queries are sent to the most suitable LLM, lowering latency and improving time to first token.

"Enterprises are increasingly deploying multiple LLMs to power advanced AI experiences—but routing and classifying LLM traffic can be compute-heavy, degrading performance and user experience," said Kunal Anand, Chief Innovation Officer at F5. "By programming routing logic directly on NVIDIA BlueField-3 DPUs, F5 BIG-IP Next for Kubernetes is the most efficient approach for delivering and securing LLM traffic. This is just the beginning. Our platform unlocks new possibilities for AI infrastructure, and we're excited to deepen co-innovation with NVIDIA as enterprise AI continues to scale."

Optimizing GPUs for Distributed AI Inference at Scale with NVIDIA Dynamo and KV Cache Integration

Earlier this year, NVIDIA Dynamo was introduced, providing a supplementary framework for deploying generative AI and reasoning models in large-scale distributed environments. NVIDIA Dynamo streamlines the complexity of running AI inference in distributed environments by orchestrating tasks like scheduling, routing, and memory management to ensure seamless operation under dynamic workloads. Offloading specific operations from CPUs to BlueField DPUs is one of the core benefits of the combined F5 and NVIDIA solution. With F5, the Dynamo KV Cache Manager feature can intelligently route requests based on capacity, using Key-Value (KV) caching to accelerate generative AI use cases by speeding up processes based on retaining information from previous operations (rather than requiring resource-intensive recomputation). From an infrastructure perspective, organizations storing and reusing KV cache data can do so at a fraction of the cost of using GPU memory for this purpose.

"BIG-IP Next for Kubernetes accelerated with NVIDIA BlueField-3 DPUs gives enterprises and service providers a single point of control for efficiently routing traffic to AI factories to optimize GPU efficiency and to accelerate AI traffic for data ingestion, model training, inference, RAG, and agentic AI," said Ash Bhalgat, Senior Director of AI Networking and Security Solutions, Ecosystem and Marketing at NVIDIA. "In addition, F5's support for multi-tenancy and enhanced programmability with iRules continue to provide a platform that is well-suited for continued integration and feature additions such as support for NVIDIA Dynamo Distributed KV Cache Manager."

Improved Protection for MCP Servers with F5 and NVIDIA

Model Context Protocol (MCP) is an open protocol developed by Anthropic that standardizes how applications provide context to LLMs. Deploying the combined F5 and NVIDIA solution in front of MCP servers allows F5 technology to serve as a reverse proxy, bolstering security capabilities for MCP solutions and the LLMs they support. In addition, the full data programmability enabled by F5 iRules promotes rapid

and the LLMs they support. In addition, the full data programmability enabled by F5 iRules promotes rapid adaptation and resilience for fast-evolving AI protocol requirements, as well as additional protection against emerging cybersecurity risks.

“Organizations implementing agentic AI are increasingly relying on MCP deployments to improve the security and performance of LLMs,” said Greg Schoeny, SVP, Global Service Provider at World Wide Technology. “By bringing advanced traffic management and security to extensive Kubernetes environments, F5 and NVIDIA are delivering integrated AI feature sets—along with programmability and automation capabilities—that we aren’t seeing elsewhere in the industry right now.”

F5 BIG-IP Next for Kubernetes deployed on NVIDIA BlueField-3 DPUs is generally available now. For additional technology details and deployment benefits, go to www.f5.com and visit the companies at NVIDIA GTC Paris, part of this week’s VivaTech 2025 event. Further details can also be found in a companion blog from F5.

About Sesterce

Founded in 2018, Sesterce is a leading European operator specialized in high-performance computing and artificial intelligence infrastructure. With full control over the value chain, the company delivers flexible, sovereign, and sustainable solutions tailored to the needs of startups, large enterprises, and academic institutions. Sesterce aims to become the European leader in AI infrastructure—empowering innovators to scale while upholding ethical and environmental standards.

In this spirit, Sesterce also offers an “AI-native” service layer on top of its infrastructure: it provides high-level data preparation that ingests and transforms heterogeneous real-time streams, delivers dedicated support for Very Large Language Models (VLLMs) to organizations developing or deploying large-scale models, and supplies modular business intelligence solutions designed for both AI-native startups and established enterprises. Moreover, Sesterce ensures end-to-end privacy and control with private AI and inference environments fully compliant with European sovereignty and confidentiality standards.

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

F5, BIG-IP, BIG-IP Next, and iRules are trademarks, service marks, or tradenames of F5, Inc., in the U.S. and other countries. All other product and company names herein may be trademarks of their respective owners.

Source: F5, Inc.

View source version on businesswire.com: https://www.businesswire.com/news/home/20250611617990/en/

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

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

Source: F5, Inc.