F5 to Supercharge AI Application Delivery for Service Providers and Enterprises with NVIDIA BlueField-3 DPUs

Oct 24, 2024 1:30 AM

F5 BIG-IP Next for Kubernetes, F5's new intelligent proxy, combined with NVIDIA BlueField-3 DPUs, transforms application delivery for AI workloads

SEATTLE--(BUSINESS WIRE)-- F5 (NASDAQ: FFIV) today announced the availability of BIG-IP Next for Kubernetes, an innovative AI application delivery and security solution that equips service providers and large enterprises with a centralized control point to accelerate, secure, and streamline data traffic that flows into and out of large-scale AI infrastructures.

This press release features multimedia. View the full release here: https://www.businesswire.com/news/home/20241023073581/en/

The solution harnesses the power of high-performance NVIDIA BlueField-3 DPUs to enhance the efficiency of data center traffic that is critical to large-scale AI deployments. With an integrated view of networking, traffic management, and security, customers will be able to maximize data center resource utilization while achieving optimal AI application performance. This not only improves infrastructure efficiency but also enables faster, more responsive AI inference, ultimately delivering an enhanced AI-driven customer experience.

F5 BIG-IP Next for Kubernetes is a purpose-built solution for Kubernetes environments that has been proven in large-scale telco cloud and 5G infrastructures. With BIG-IP Next for Kubernetes, this technology is now tailored for leading AI use cases such as inference, retrieval-augmented generation (RAG), and seamless data management and storage. The integration with NVIDIA BlueField-3 DPUs minimizes hardware footprint, enables granular multi-tenancy, and optimizes energy consumption while delivering high-performance networking, security, and traffic management.

The combination of F5 and NVIDIA technologies allows both mobile and fixed-line telco service providers to ease the transition to cloud-native (Kubernetes) infrastructure, addressing the growing demand for vendors to adapt their functions to a cloud-native network functions (CNFs) model. F5 BIG-IP Next for Kubernetes offloads data-heavy tasks to the BlueField-3 DPUs, freeing up CPU resources for revenue-generating applications. The solution is particularly beneficial at the network edge for virtualized RAN (vRAN) or DAA for MSO, and in the core network for 5G, enabling future potential for 6G.

Designed specifically for high-demand service providers and large-scale infrastructures, F5 BIG-IP Next for Kubernetes:

- ***Streamlines delivery of AI services at cloud scale:*** BIG-IP Next for Kubernetes seamlessly integrates with customers' front-end networks, significantly reducing latency while delivering high-performance load balancing to handle the immense data demands of multi-billion-parameter AI models and trillions of operations.
- ***Enhances control of AI deployments:*** The solution offers a centralized integration point into modern AI networks with rich observability and fine-grained information. BIG-IP Next for

Kubernetes supports multiple L7 protocols beyond HTTP, ensuring enhanced ingress and egress control at very high performance.
- ***Protects the new AI landscape:*** Customers can fully automate the discovery and security of AI training and inference endpoints. BIG-IP Next for Kubernetes also isolates AI applications from targeted threats, bolstering data integrity and sovereignty while addressing the encryption capabilities critical for modern AI environments.

Availability for BIG-IP Next for Kubernetes running on NVIDIA BlueField-3 DPUs will begin in November. Additional information can also be found in a companion blog post from F5.

Supporting Quotes:

- "The proliferation of AI has catalyzed an unprecedented demand for advanced semiconductors and technologies. Organizations are building out AI factories, highly optimized environments designed to train large AI models and deliver the requisite processing power for inference scale at an astounding rate, and with minimal latency. The synergy between F5's robust application delivery and security services and NVIDIA's full-stack accelerated computing creates a powerful ecosystem. This integration provides customers with enhanced observability, granular control, and optimized performance for their AI workloads across the entire stack, from the hardware acceleration layer to the application interface." - Kunal Anand, Chief Technology and AI Officer at F5

- "Service providers and enterprises require accelerated computing to deliver high-performance AI applications securely and efficiently at cloud scale. NVIDIA is working with F5 to accelerate AI application delivery, better ensuring peak efficiency and seamless user experiences powered by BlueField-3 DPUs." - Ash Bhalgat, Sr. Director of AI Networking and Security Partnerships at NVIDIA

- "Realizing the potential of AI requires more data processing capabilities than the industry had previously prepared for. For many companies, deploying cutting-edge AI requires massive infrastructure buildouts that tend to be very complex and expensive, making efficient and secure operations more important than ever. F5 BIG-IP Next for Kubernetes addresses performance and security concerns for large-scale AI infrastructure. By delivering optimized traffic management, organizations gain greater data ingestion performance and server utilization during AI model inferencing. This leads to a vastly improved customer experience for AI application users." - Kuba Stolarski, Research Vice President, Computing Systems Research Practice at IDC

- "The explosion of AI workloads has created a new wave of massive demand for scalable, optimized, and enhanced control of Kubernetes ingress and egress. With F5 now delivering the established benefits of BIG-IP Next for Kubernetes directly on NVIDIA BlueField-3 DPUs, this unleashes an already proven technology now deployable at an ideal insertion point for large-scale AI deployments. WWT clients will be able to benefit from greater data ingestion performance and GPU utilization during model training and better user experiences during inference, while gaining a strategic control point for security services. Technology from F5 and NVIDIA—two of our most strategic partnerships—further strengthens our Global Cyber mission to deliver digital security excellence." - Todd Hathaway, Global Practice Manager, AI, App, and API Security Solutions at WWT

Supporting Resources:

- F5 Blog Post
- Technology Collaboration Page

- NVIDIA Blog Post

About F5

F5 is a multicloud application security and delivery company committed to bringing a better digital world to life. F5 partners with the world's largest, most advanced organizations to secure every app—on premises, in the cloud, or at the edge. F5 enables businesses to continuously stay ahead of threats while delivering exceptional, secure digital experiences for their customers. For more information, go to f5.com. (NASDAQ: FFIV)

You can also follow @F5 on X or visit us on LinkedIn and Facebook to learn about F5, its partners, and technologies. F5, BIG-IP, and BIG-IP Next 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. The use of the terms "partner," "partners," "partnership," or "partnering" in this press release does not imply that a joint venture exists between F5 and any other company.

This press release contains forward-looking statements including, among other things, statements regarding potential benefits and availability of F5 BIG-IP Next for Kubernetes running on NVIDIA BlueField-3 DPUs. Each customer's unique environment, objectives, and constraints could impact potential benefits, while surrounding factors could affect availability timing.

Source: F5, Inc.

View source version on businesswire.com: https://www.businesswire.com/news/home/20241023073581/en/

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

Holly Lancaster
WE Communications
(415) 547-7054
hluka@we-worldwide.com

Source: F5, Inc.