

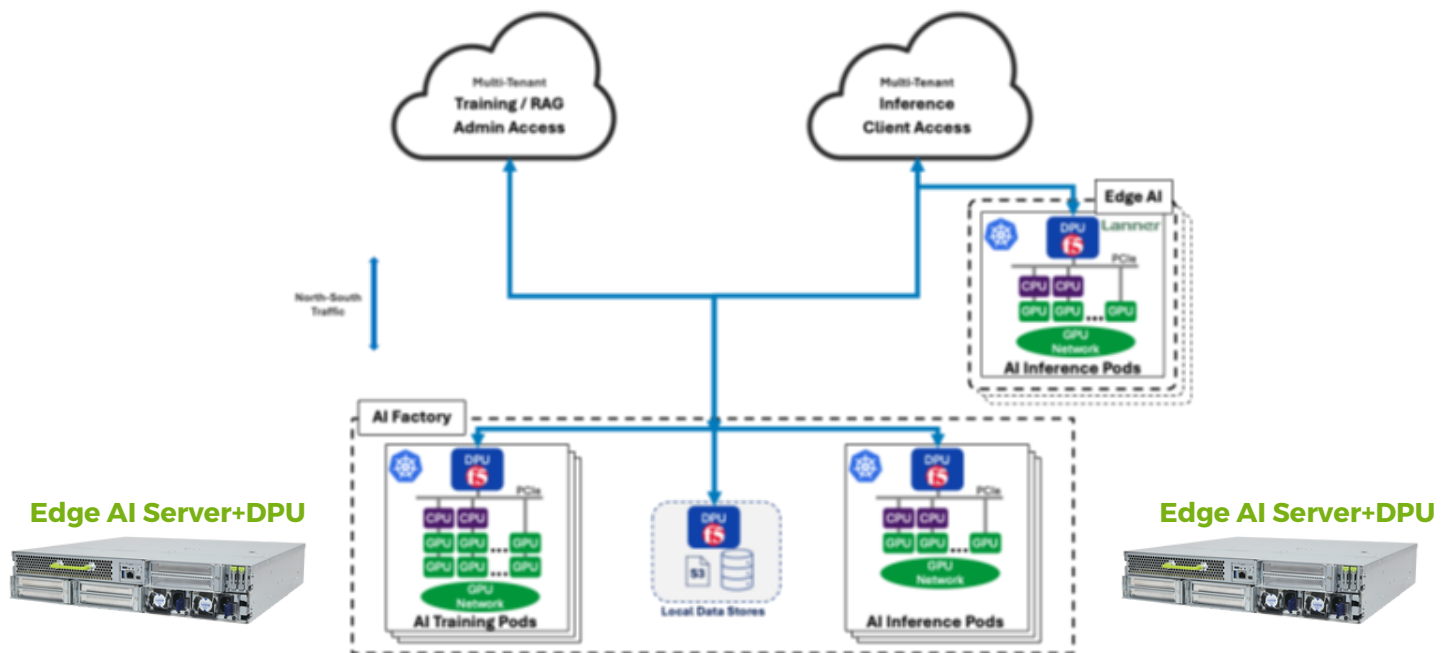
Lanner

OPTIMIZING NETWORKING & SECURITY FOR MODERN AI INFRASTRUCTURE

SOLUTION BRIEF



Lanner collaborates with F5 and NVIDIA to offer a high-performance NVIDIA MGX AI server integrated with BlueField-3 DPUs, designed to enhance data center traffic efficiency for large-scale AI deployments. By providing an integrated approach to networking, traffic management, and security, this solution enables customers to maximize data center resource utilization while freeing up valuable CPU cycles for optimizing AI applications. This offloading not only boosts infrastructure efficiency but also ensures faster, more responsive AI inferencing, ultimately delivering an enhanced AI-driven customer experience.



ECA-6051A-F5

- Intel® Xeon®6 Processor up to 144 Cores
- 8x DDR5 6400MHz RDIMM, Max. 1024GB System Memory
- 3x PCIe*16 slots, with NVIDIA L40S GPU or Bluefield-3 DPU support
- 1x GbE RJ45, 1x RJ45 Console, 1x USB 3.0
- 2x M.2 NVMe (PCIe), 2x E1.S SSD
- 6x Smart Fans, 2x 1600W AC CRPS PSU

Lanner



www.lannerinc.com



contact@lannerinc.com

© Lanner Electronics Inc. All rights reserved.



Edge AI is transforming business. Lanner Edge AI servers enable real-time data processing, deep learning and AI-driven analytics across various applications, from 5G and smart cities to industrial automation. F5 BIG-IP Next for Kubernetes solves the complexity of application delivery without burdening the performance and scale necessary for the demanding edge applications. The integration of F5 and Lanner provides a complete solution to deliver scalable scalability, low latency and performance at the edge.

For more details about F5 BIG-IP Next for Kubernetes on Nvidia DPUs, please go to <https://www.f5.com/partners/technology-alliances/nvidia>

About F5 BIG-IP Next for Kubernetes

BIG-IP Next for Kubernetes delivers high-performance traffic management and security for large-scale AI infrastructure, unlocking greater efficiency, control, and performance for AI applications.

BIG-IP Next for Kubernetes runs natively on NVIDIA BlueField-3 DPUs. This provides enterprises and service providers with a single control point to maximize AI infrastructure usage and accelerate AI traffic, for data ingestion, model training, inference, and retrieval-augmented generation (RAG). Increasingly AI clusters are moving to the edge, to meet the demanding requirements of AI-powered user experiences.

F5 BIG-IP Next for Kubernetes Benefits:

Optimizing AI cluster networks

Providing required cluster network ingress services for AI access consumes 20-30% of valuable AI cluster compute. Offloading networking processes to DPUs frees cluster compute, thus optimizing GPU utilization.

Gaining security for AI clusters

Integrating required cluster networking with SecOps required firewalling removes unnecessary expensive inter-process service chaining. Completely offloading both required networking and firewalling services to the DPU reserves host resources for AI workloads.

Application delivery for AI clusters

Default AI cluster networking lacks multi-protocol support to intelligently deliver application requests to AI services. BIG-IP provides the widest array of protocol support for applications accessing AI.

Multi-tenant AI clusters

Valuable GPU resources introduce the complexity of multi-tenant AI clusters. BIG-IP provides the required mapping of internal cluster segmentation and data center network segmentation required for data center operations.

Observability for AI clusters

BIG-IP provides key logging and telemetry of AI requests. BIG-IP observability data allows for the correlation of application requests to AI cluster and data center metrics. BIG-IP logging and telemetry are based on the industry standard OpenTelemetry protocols.

Lanner MGX AI Server: ECA-6051A-F5

