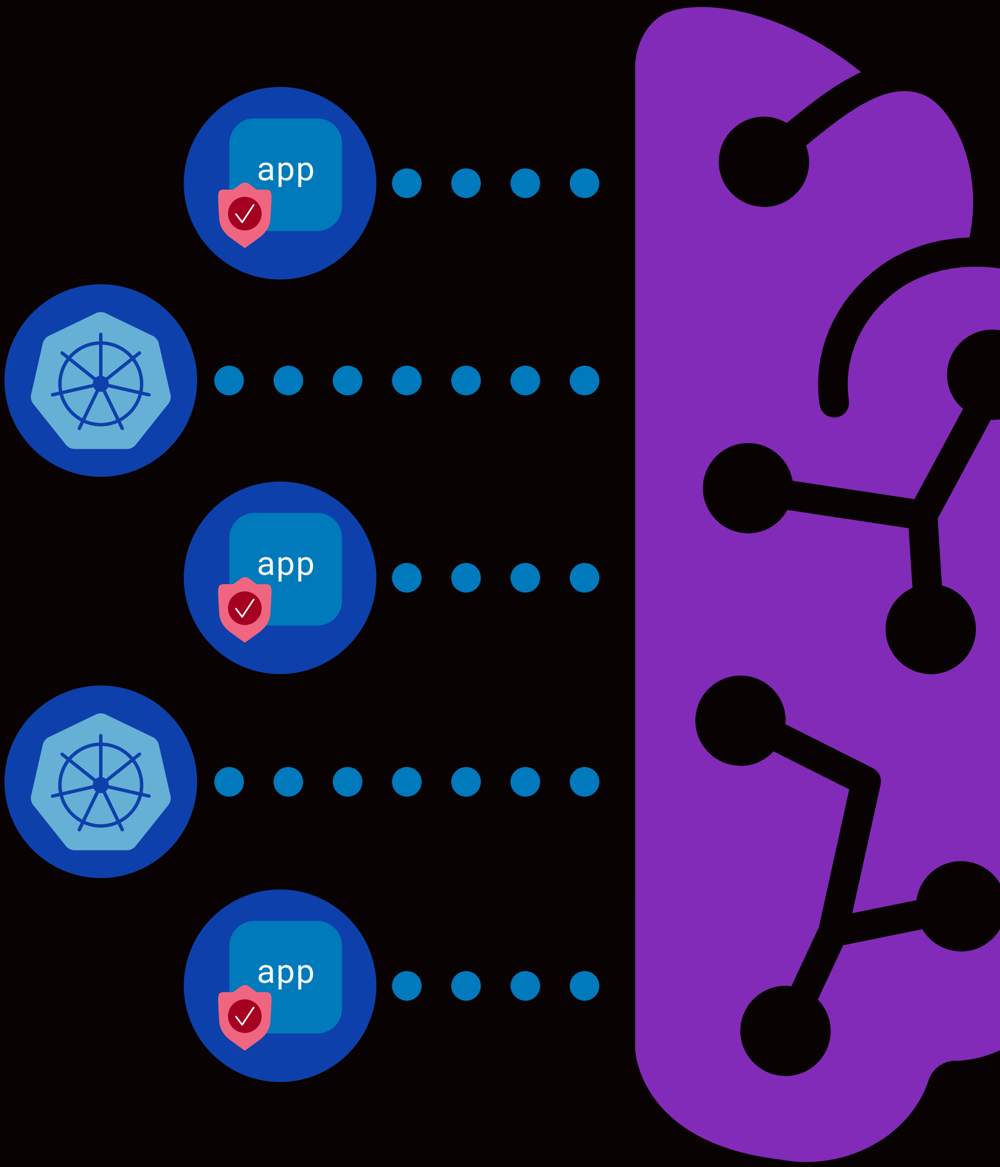
SOLUTION OVERVIEW

# Secure AI application delivery in Kubernetes environments

Overcome the challenges of delivering AI applications in containerized environments with F5 and AWS.



**Key benefits**

**Enable faster, more efficient AI models**
Drive intelligent load balancing in Kubernetes for more responsive AI performance, efficient GPU utilization, and increased deployment flexibility.

**Enhance security controls**
Protect your valuable AI models, sensitive data, and application infrastructure to ensure user trust, brand reputation, and regulatory compliance.

**Optimize performance**
Leverage AI-specific metrics to fuel optimization efforts, identify opportunities for improvement, and respond quickly to changing conditions.

**Quickly scale across distributed environments**
Deliver everything AI needs to scale in Kubernetes environments on AWS, in data centers, and across public and private clouds.

## AI applications have unique delivery requirements in Kubernetes environments

Kubernetes is the platform of choice for modern apps and AI, with 66% of surveyed organizations using it in production.[1] However, AI applications differ from traditional containerized workloads. Unlike conventional microservices with consistent and predictable resource needs, AI applications require specialized handling to address variable processing demands and elevated security risks.

### Container ingress controllers struggle with AI

AI services handle everything from simple text prompts to complex multimedia analyses. These variable workloads create dramatic differences in processing needs that largely depend on GPU resources. However, traditional ingress controllers lack awareness for GPU availability, resulting in uneven workload distribution that leaves some GPUs congested and others underutilized, driving up costs and slowing performance.

### Traditional Kubernetes security wasn't designed for AI-specific threats

AI services typically require numerous API connections to various model and data repositories and other software services, creating complex, difficult-to-manage environments with larger attack surfaces. These environments are attractive to cyber criminals and vulnerable to AI-specific lateral movement and data extraction techniques, such as prompt injection and jailbreaking. To address this added complexity, organizations need additional layers of protection that go beyond standard Kubernetes security.

## F5 and AWS optimize AI application delivery for Kubernetes

F5 and AWS work together to address the unique challenges of containerized AI services by enhancing performance, security, and operational efficiency.

### AI model deployment and protection at scale

F5® NGINX® Ingress Controller and NGINX App Protect operate as a single tool that combines ingress, load balancing, and API gateway functions for better uptime, protection, and visibility. The solution provides flexible release strategies, A/B testing for model experimentation, and dynamic reconfiguration for reliable AI delivery even during scaling events or pod failures.

### Key features

**Distribute workloads intelligently**

Deploy AI-aware traffic management that directs requests based on content, payload size, and expected processing requirements rather than simple round-robin distribution.

**Protect AI with container-native security**

Address AI-specific threats such as prompt injection and jailbreaking with API security, distributed denial-of-service (DDoS) protection, input/output validation, and robust identity and access controls.

**Scale workloads based on AI-specific patterns**

Implement scaling capabilities that understand GPU utilization and model loading time based on actual demand patterns rather than generic resource consumption.

**Empower optimization efforts**

Get AI-specific metrics alongside standard container health indicators, including prompt volume, token usage, inference latency, and model performance statistics.

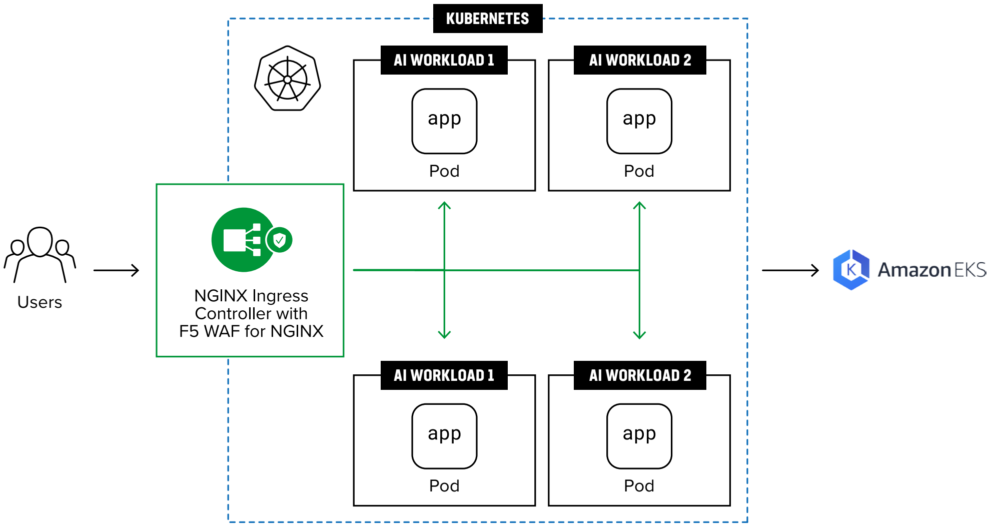


**Figure 1:** F5 provides AI-aware traffic management and protection for Amazon Elastic Kubernetes Service (EKS).

## AI-specific app security and optimization

F5 AI Gateway offers intelligent traffic management and security for AI apps in a flexible, Kubernetes-based format that runs across public clouds, private clouds, and on-premises environments. It supports leading AI platforms including OpenAI, Anthropic, and Ollama, along with HTTP-based language models, providing consistent, location-agnostic protection that addresses the OWASP Top 10 for LLMs.

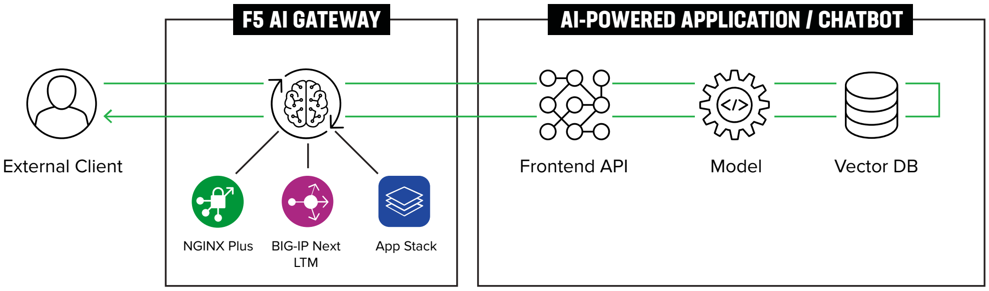


**Figure 2:** AI Gateway simplifies AI delivery across hybrid multicloud environments.

### Managed Kubernetes platform

Amazon EKS reduces operational complexity while maintaining full Kubernetes compatibility. F5 solutions enhance EKS with AI-specific delivery capabilities, creating a comprehensive platform for containerized AI workloads that balances innovation with operational stability.

## Protect and enhance AI with F5 and AWS

Kubernetes is an indispensable platform for containerized workload efficiency and portability. Whether you're deploying on AWS or across hybrid multicloud infrastructure, F5 solutions work consistently to enhance Kubernetes and bridge critical AI performance, security, and scalability gaps, empowering your organization and setting the stage for success in the era of AI.

**Contact F5 to get started or learn more at f5.com/aws.**

[1] Cloud Native Computing Foundation, CNCF 2023 Annual Survey, April 2024.

DC 09.2025 | JOB-CODE-1723245974