Solution Overview

# Kubernetes Connectivity and Security for AI/ML Workloads

Experience fast, reliable, and secure AI/ML app delivery in Kubernetes with F5 NGINX solutions.

**Key Benefits**

**Simplify operations**
Operationalize AI/ML workloads easily and reliably, reducing complexity through consistency across environments.

**Gain insight**
Improve model serving efficiency, uptime, and SLAs with better visibility into AI/ML app health and performance.

**Improve security**
Protect AI/ML workloads without adding extra complexity and overhead or slowing down release velocity and performance.

# Deploy AI/ML Models in Production at Scale

As new threats emerge, working from a common set of policies allows you to mitigate them far more easily. Security is complex, and while developers should follow secure code practices, they are not security experts—and they really shouldn't be. Instead, they should focus on areas of high exposure and allows IT to address the rest with external controls that protect against both known and unknown threats.

Standardizing on application services that improve application performance can also help ensure good customer experience across applications. User patience with poor response times continues to decline, and developers vary in their ability to code highly performing applications. High turnover may also leave organizations with different developer skill sets. Organizations can compensate for these deficiencies by offering a standard set of performance optimization services for all apps.

AI and machine learning (AI/ML) workloads are revolutionizing how businesses operate and innovate. Kubernetes, the de facto standard for container orchestration and management, is the platform of choice for powering scalable large language model (LLM) workloads and inference models across hybrid, multi-cloud environments.

Out-of-the-box Kubernetes features and capabilities can help:

- Accelerate and simplify the AI/ML application release life cycle
- Enable AI/ML workload portability across different environments
- Improve compute resource utilization efficiency and economics
- Deliver scalability and achieve production readiness
- Optimize the environment to meet business SLAs

At the same time, organizations might face challenges when serving, experimenting with, monitoring, and securing AI/ML models in production at scale. These challenges include:

**AI and machine learning workloads are revolutionizing how businesses operate and innovate.**

- **Difficulty in operating and managing a hybrid, multi-cloud Kubernetes environment**
  Increasing complexity and tool sprawl (caused by disaggregated technologies from multiple vendors, for example) make it difficult for organizations to configure, operate, manage, automate, and troubleshoot hybrid, multi-cloud Kubernetes environments on premises, in the cloud, and at the edge.

- **Poor user experiences in scalable, dynamic environments**
  Application end users often encounter poor experiences because of connection timeouts and errors due to topology changes (such as auto-scaling or pod failures and restarts), extremely high request rates, and service failures due to the dynamic nature of Kubernetes pod orchestration.

- **Insufficient visibility into app health and performance**
  Aggregated reporting and lack of granular real-time and historical metrics in complex Kubernetes environments can cause performance degradation and downtime, and also make troubleshooting harder and slower.

- **Increased risk of exposure to cybersecurity threats across distributed app environments**
  Traditional security models are not designed to protect loosely coupled distributed applications, which creates significant risk of exposure to both external and internal threats in on-premises, hybrid, and multi-cloud Kubernetes environments.

To address these challenges, F5 delivers Kubernetes connectivity and security solutions that leverage F5® NGINX® Ingress Controller and F5 NGINX® App Protect.

Leveraging one tool that combines ingress controller, load balancer, and API gateway capabilities provides better uptime, protection, and visibility at scale no matter where you run Kubernetes—while also reducing complexity and operational cost. This industry-leading layer 7 app protection technology from F5 helps mitigate OWASP Top 10 cyberthreats for LLM applications and defends AI/ML workloads from denial-of-service (DoS) attacks.

**Leveraging one tool that combines ingress controller, load balancer, and API gateway capabilities provides better uptime, protection, and visibility at scale no matter where you run Kubernetes.**

## Simplify and Streamline Model Serving, Experimentation, Monitoring, and Security

NGINX Connectivity Stack for Kubernetes provides fast, reliable, and secure communications between Kubernetes clusters running AI/ML applications and their users—on premises and in the cloud. These benefits include:

- **Scalability, Availability, and Performance**
  Operationalize AI/ML workloads easily and reliably with adaptive load balancing, non-disruptive reconfiguration, A/B testing, and canary deployments, reducing complexity through consistency across environments.

- **Visibility and Insight**
  Improve model serving efficiency, uptime, and SLAs by resolving app connectivity issues quickly with extensive, granular metrics and dashboards using real-time and historical data.

- **Security and Control**
  Protect AI/ML workloads without adding extra complexity and overhead with strong security controls across distributed environments without slowing down release velocity or performance.

**NGINX Connectivity Stack for Kubernetes helps simplify and streamline model serving, experimentation, monitoring, and security across any Kubernetes environment.**

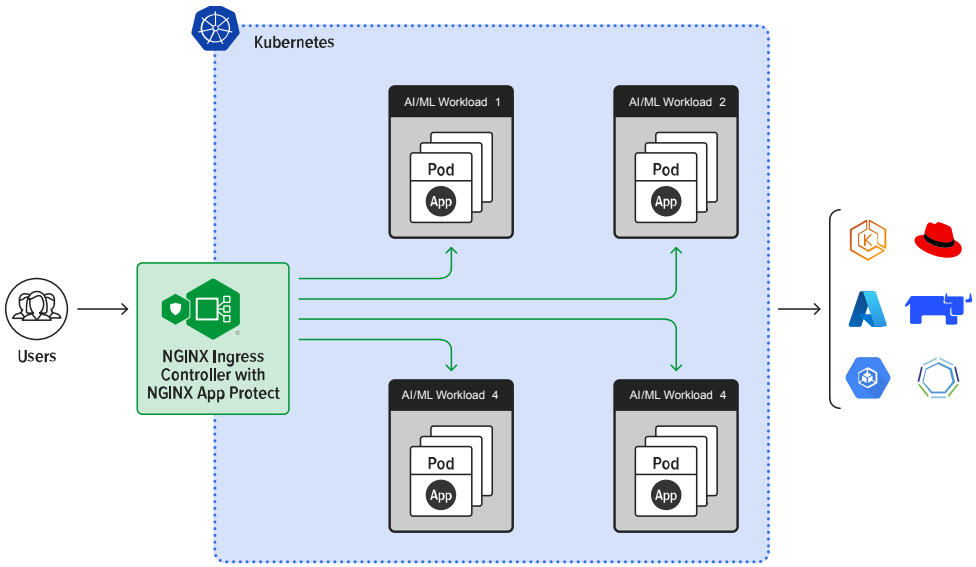


**Figure 1:** NGINX Connectivity Stack for Kubernetes

NGINX Connectivity Stack for Kubernetes helps simplify and streamline model serving, experimentation, monitoring, and security across any Kubernetes environment, enhancing capabilities of cloud provider and pre-packaged Kubernetes offerings with higher degrees of protection, availability, and observability at scale.

### Simplify operations

- Experience better uptime, protection, and visibility with a universal Kubernetes-native implementation that includes an API gateway, load balancer, and ingress controller in one tool.
- Reduce complexity with data and control plane consistency across any hybrid, multi-cloud environment.

### Deliver AI/ML apps without disruption

- Optimize workload distribution with least-time load balancing and active health checks.
- Implement blue-green and canary deployments to avoid downtime when rolling out a new version of a model and A/B testing for online experimentation with the model.
- Use dynamic reconfiguration, rate limiting, circuit breaking, and request buffering to prevent connection timeouts and errors due to topology changes (such as auto-scaling, pod failures, and restarts), extremely high request rates, and service failures.

**Key Features**

**Increase uptime**
Roll out new versions of and experiment with the models without disruption.

**Improve visibility**
Collect, monitor, and analyze health and performance metrics for the model.

**Strengthen security**
Ensure holistic model protection with strong and consistent security controls.

**Mitigate cybersecurity threats**

- Manage user and service identities and their authorized access and actions with JSON Web Tokens (JWTs), OpenID Connect (OIDC), and role-based access control (RBAC).
- Secure incoming and outgoing communications through end-to-end encryption (SSL/TLS passthrough, TLS termination).
- Protect apps from the majority of OWASP Top 10 for LLMs and layer 7 DoS attacks with industry-leading F5 technology.
- Shield precious GPU resources needed for model serving in a Kubernetes cluster from misuse or depletion.

**Gain better insight**

- Collect, monitor, and analyze your model metrics through prebuilt integrations with your favorite ecosystem tools, including Grafana, Prometheus, and Jaeger.
- Reduce outages and downtime by discovering problems before they impact your customers.
- Simplify troubleshooting by quickly finding the root cause of app issues.

**Release AI/ML apps faster**

- Focus on implementing core model expertise and functionality within the app.
- Offload security and other non-functional requirements to the platform layer.
- Enable self-service governance across multi-tenant MLOps teams.

# Conclusion

NGINX Connectivity Stack for Kubernetes delivers built-in security, availability, and visibility for AI/ML workloads across hybrid and multi-cloud Kubernetes environments, reducing complexity, improving uptime, and gaining better insight into app health and performance at scale

# Next Steps

- See how NGINX Connectivity Stack for Kubernetes works with a free trial, start today.
- Contact F5 to find out how F5 products and solutions can enable you to achieve your goals.