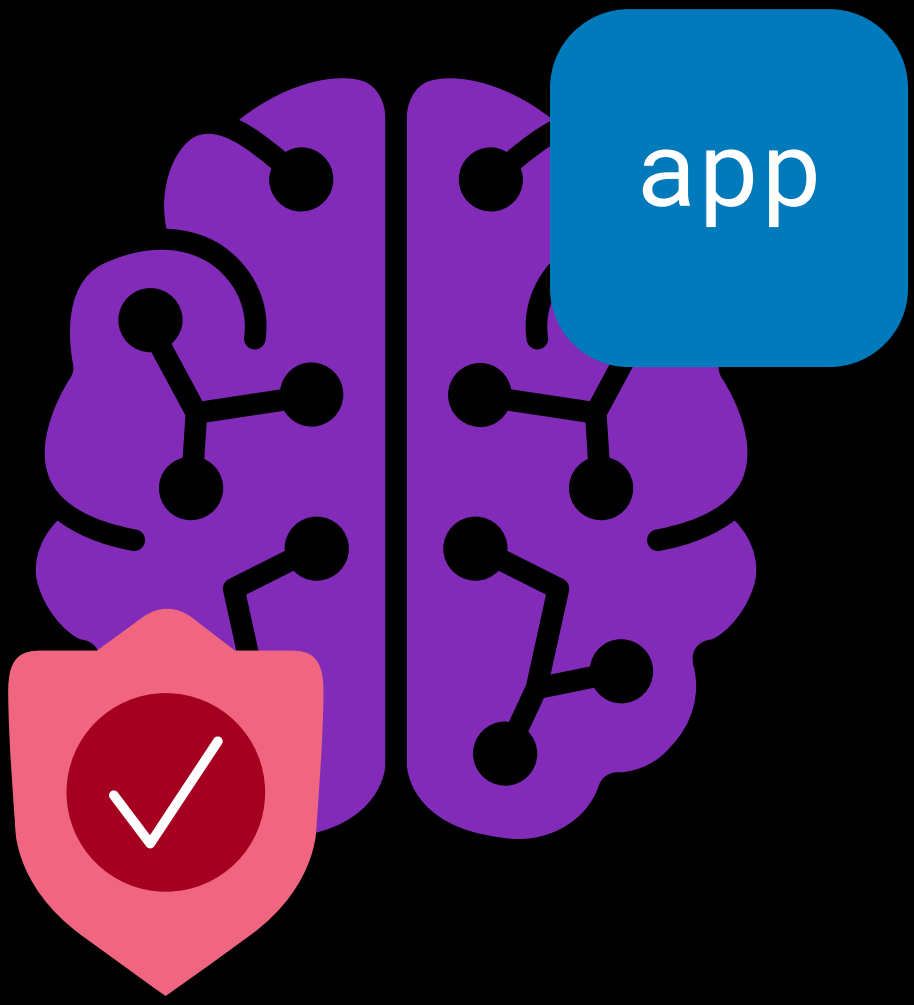
Google Cloud

SOLUTION OVERVIEW

Secure, scale, and optimize your AI applications with F5 and Google Cloud

Overcome the unique challenges of AI adoption with integrated technologies that secure model inputs and outputs, control costs, unify observability, and simplify retrieval-augmented generation across any environment.



Key benefits

Protect AI investments
Safeguard your intellectual property and sensitive data from theft or compromise with protection designed specifically for AI models and apps.

Optimize performance and costs
Control AI expenses and improve response times through resource management and specialized infrastructure.

Ensure consistent deployment anywhere
Deploy AI applications confidently across hybrid multicloud environments with consistent security policies and performance optimization capabilities.

Simplify hybrid multicloud complexity
Reduce operational complexity with unified management and observability tools that provide comprehensive visibility across your entire AI application ecosystem.

Accelerate time to value
Leverage Google Cloud's AI platform, pre-trained models, and F5's application delivery capabilities to rapidly develop, secure, and scale AI applications without deep machine learning expertise.

Navigating the unique challenges of enterprise AI applications

As your organization adopts AI technologies, you face significant obstacles to security, performance, and integration. More than 70% of organizations are currently working on AI applications,[1] which rely heavily on APIs and are increasingly deployed at the edge to improve latency and privacy.[2] Multiple data sources, deployment locations, and connectors expand the attack surface in AI-powered hybrid multicloud environments. .

This complexity requires a new infrastructure approach to deliver and secure AI applications. Traditional security measures aren't designed to address AI-specific threats like prompt injection and model poisoning, while the substantial computational resources required by large language models (LLMs) make optimization critical to control costs and ensure high performance.

Comprehensive AI protection and optimization with F5 and Google Cloud

F5 and Google Cloud have partnered to address these challenges and safely accelerate enterprise AI adoption. By combining Google Cloud's AI-optimized infrastructure and AI models with F5's security and application delivery services, you get an integrated solution for layered protection and performance optimization.

With Google Cloud Vertex AI, you can leverage end-to-end machine learning model development and deployment without requiring deep machine learning expertise. Your developers can use pre-trained first- and third-party models as starting points, deploy them through Google Kubernetes Engine (GKE), and run them on secure, purpose-built AI infrastructure. F5 complements Google Cloud with security and performance capabilities engineered to help you scale your AI inititaitves with confidence. Regardless of where your AI applications or models are deployed, F5 provides consistent security policies and optimization capabilities, giving you deep visibility into AI operations while controlling resource utilization.

Key features

AI-specific threat prevention
Defend against unique AI attack vectors with specialized protection that monitors prompt interactions, validates model responses, and prevents data loss.

Intelligent traffic management
Distribute AI requests based on real-time performance metrics, resource availability, and cost considerations.

Comprehensive API security
Discover and protect API endpoints connecting your AI apps with models and data sources across all environments.

Centralized observability
Monitor key metrics, including request volumes, token usage, and response times, across all AI apps in a single view using a flexible open standard.

Edge AI deployment
Deploy AI workloads closer to data sources or users while maintaining consistent security and performance.

Secure data connectivity
Establish encrypted connections between AI models and data sources across environments without exposing sensitive information to the public Internet.

Secure AI model inputs and outputs

Your customized AI models are valuable intellectual property, often with access to sensitive corporate data, making them attractive targets for attackers. F5 works with Google Cloud to operationalize the Secure AI Framework (SAIF) across your entire AI application stack, delivering security that goes beyond traditional defense-in-depth approaches.

F5® AI Gateway and Google Cloud Model Armor provide adaptive controls recommended by SAIF to keep up with evolving threats. They monitor prompt inputs to prevent attacks, block malicious requests, and prevent legitimate users from accidentally providing confidential data that could pose a security or compliance risk. Along with Google Cloud Sensitive Data Protection, these solutions automatically detect and redact PII or other confidential data from prompts and responses.

Extend detection and response to protect the APIs that connect your AI apps with models and data sources with F5® Distributed Cloud API Security. By discovering shadow APIs and preventing evolving AI risks like unbounded consumption, you can reduce compute costs and AI token usage.

Deploying F5 solutions alongside Google Cloud Vertex AI and Model Armor creates layered defenses for your AI workloads with consistent protection and unified management across GKE, Google Distributed Cloud, other cloud platforms, data centers, and edge sites.

Optimize AI application performance and costs

Each AI model interaction consumes costly computing resources and potentially incurs per-token fees. The combination of high-performance hardware like Google's Cloud Tensor Processing Units (TPUs) and AI Gateway helps accelerate AI operations while giving you tigher control over costs. F5's intelligent load balancing distributes AI requests across a diverse model ecosystem based on real-time metrics to avoid bottlenecks and use the most appropriate models based on complexity, cost, or compliance requirements.

Rate limiting sets usage thresholds to prevent resource overload, while semantic caching saves costs by reusing results from similar prompts to reduce AI token and compute usage. For deployments requiring low latency, Google Distributed Cloud enables you to run AI workloads closer to data sources or users while F5 optimizes traffic to ensure consistent performance across all locations, maximizing efficiency without sacrificing user experience.

Unify observability for AI applications

It can be hard to keep track of everything in widely distributed AI ecosystems, and that lack of visibility leads to issues with security, performance, and model degradation. F5 and Google Cloud provide unified observability for your AI applications to support MLOps, enabling your teams to monitor, troubleshoot, and optimize more easily. AI Gateway integrates with Google Cloud's observability suite to track key metrics like request volumes, token usage, and response times across all AI applications in centralized dashboards.

Native OpenTelemetry support in AI Gateway lets you export standardized metrics, logs, and traces that can be consumed by Google Cloud's observability platform, third-party SIEM, or SOAR tools for maximum flexibility. Unified visibility extends across AI applications running in GKE, Google Distributed Cloud, or elsewhere in your hybrid multicloud environment to eliminate blind spots with customizable dashboards.

Simplify retrieval-augmented generation

Retrieval-augmented generation (RAG) adds data from enterprise sources for more accurate AI responses, but if your enterprise data is spread across multiple environments, connecting it with AI models can be challenging. F5, Google Cloud, and NetApp offer an integrated solution to create secure, efficient data flows for your RAG implementations by using F5® Distributed Cloud Services to establish encrypted connections between Google Cloud Vertex AI and data sources, including NetApp storage systems in data centers or Cloud Volumes ONTAP for Google Cloud.

Automated provisioning eliminates the complexity of building and maintaining connections across your hybrid multicloud environment, while the private F5® Global Network keeps sensitive data from traversing the public Internet. Consistent security controls protect your connections and data in transit through a single platform with unified policy enforcement.

Secure, efficient, and cost-effective AI for your business

By integrating F5's application delivery and security capabilities with Google Cloud infrastructure and AI models, you can build and operate AI apps confidently. Centralized management simplifies protection and performance no matter where your AI models and applications run. Together, F5 and Google Cloud help you innovate safely with AI while ensuring your applications remain secure, efficient, and cost-effective.

Learn more about F5 and Google Cloud at f5.com/gcp

[1] Applause, The State of Digital Quality in AI in 2025, Mar 2025 [2] Google Cloud, 2024 State of Edge Computing, Jul 2024
[3] Google Cloud, 2025 State of AI infrastructure, Apr 2025