Solution Overview

Maintain Connectivity and Security for AI-Based Apps

Connect, secure, and scale your AI workloads on one platform.

Key Benefits

Enable service connectivity across environments
Connect AI inference apps to LLMs wherever they are deployed.

Establish consistent security
Enable a consistently high level of security for all AI workloads, regardless of environment.

Enhance scalability
Rapidly deploy new inference apps to new regions and environments to meet demand.

AI Is a Powerful Modern Application, Not a Magic Wand

Generative AI is a transformative technology, driving radical shifts in the ways organizations around the world operate. Built on large language models (LLMs), AI applications are deployed to collect data, train other deep learning models, and deliver inference across multiple environments, including on-premises data centers, at the edge, and across public clouds. While AI and machine learning have been widely used in a variety of forms for two decades, the current wave of generative AI solutions makes training based on large sets of data far more accessible and leverages a host of modern applications that deliver predictable inference to applications across distributed environments.

Industry excitement regarding the potential of AI to catalyze business transformation has fueled a rapid proliferation of myths about AI's capabilities. According to Gartner, organizations and industries are dramatically rethinking key business and human resource approaches in the context of AI, pushing generative AI to the "Peak of Inflated Expectations" level on the Gartner 2023 Hype Cycle for Artificial Intelligence. In addition, a March 2023 KPMG survey of U.S. executives found that almost two thirds (65%) believe "generative AI will have a high or extremely high impact on their organization in the next three to five years," but "fewer than half of respondents say they have the right technology, talent, and governance to implement generative AI successfully."

Successful implementation of generative AI will require organizations to move beyond the hype and recognize AI as a powerful modern application, not a magic wand. At its core, generative AI infrastructure consists of highly distributed modern applications. AI solutions are built upon modern Kubernetes-based architectures (or, in the case of generative AI, modern apps built on LLMs), so any solution for successfully deploying AI will require the same fundamental components as any other modern app—connectivity, security, and scalability. Organizations must address these challenges in order to effectively deploy, manage, and secure AI solutions and expand their AI investments.

Successful implementation of generative AI will require organizations to move beyond the hype and recognize AI as a powerful modern application, not a magic wand.

Connect AI Workloads across Distributed Environments

The first step in successfully deploying generative AI is managing access and connecting to data. LLMs require vast amounts of data on which to train. Centralized data repositories, which can reside in an on-premises data center or in the cloud, provide these resources. Inference apps leverage the LLM, connecting to that LLM from wherever the app is deployed. To maximize performance, inference apps should be deployed as close to users as possible so that they can provide information with the highest levels of responsiveness.

If AI models need to be deployed quickly around the world to respond to demand, having a platform-agnostic set of security policies that can be implemented just as quickly to defend against new threats is imperative.

As global demand for AI continues to accelerate, the demand for secure, low-latency connectivity from inference apps—accessible to users and the LLMs they reference—will also increase. Managing connectivity in a distributed environment without creating additional complexity requires a single pane of glass view to maintain visibility across hybrid and multi-cloud environments.

Secure AI Workloads against Emerging Threats

After addressing data connectivity, the next step is to ensure that those connections and the training data are protected, and each interface is secured.

AI workloads are subject to a wide spectrum of attacks, many of which leverage their own adversarial AI models to find new vectors. The OWASP Top 10 for LLM Applications highlights several potential threats to AI apps, including Model Denial of Service. It is critical that organizations implement a robust app and API security solution to manage these new risks and establish a consistently high level of security across any environment. Every cloud provider and on-premises data center have unique security tooling and controls. If AI models need to be deployed quickly around the world to respond to demand, having a platform-agnostic set of security policies that can be implemented just as quickly to defend against new threats is imperative.

Another step toward secure AI workloads is to ensure that protection is localized with the support of a huge threat intelligence library. For workloads to be properly secured, protection must be available in real time, and be as up to date as possible so that no potential threat can slip through. Point protection must live on each environment where an AI workload will operate, with links back to any centralized resources like registries, threat intelligence libraries, user authentication tools, and all other security tools. Additionally, telemetry that leads to actionable insights, in conjunction with automated protections based on machine learning, will be the only way security can keep pace with generative AI.

Scale AI Workloads to Meet Demand

Once an AI workload is deployed and protected, the final step is to ensure it can grow as demands increase. During an era in which the use of generative AI is growing exponentially, waiting for connectivity and security to be provisioned can drive customers away from less responsive workloads to a faster competitor. This can lead to shadow AI and inadvertent exposure of personally identifiable information (PII) or company secrets.

Having a solution in place that automates connectivity and security helps maintain deployment velocity, removes complexity from the deployment process, and allows development teams to push new functionality to existing AI workloads. Additionally, changes in the broader AI landscape can lead to new deployment and regulatory requirements. A platform that supports all architectures and environments within vast AI ecosystems can manage these changes and ensure predictable inference and security across distributed environments, helping to keep pace with a highly competitive and dynamic market.

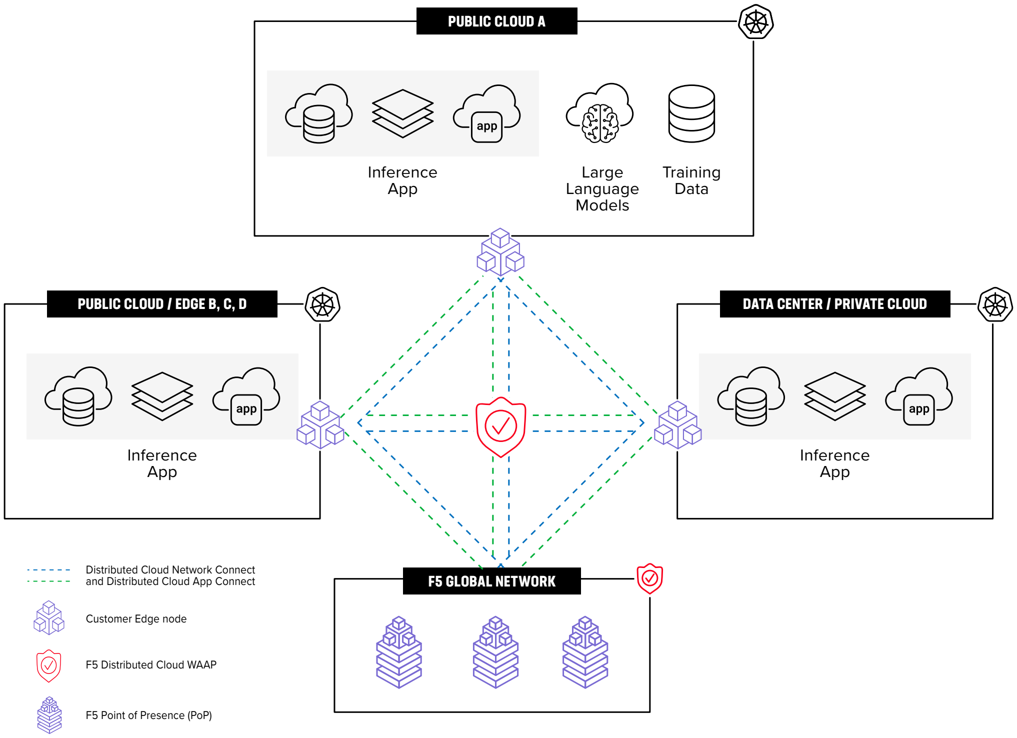
PUBLIC CLOUD A
app
Inference App
Large Language Models
Training Data
PUBLIC CLOUD / EDGE B, C, D
app
Inference App
DATA CENTER / PRIVATE CLOUD
app
Inference App
F5 GLOBAL NETWORK
Distributed Cloud Network Connect and Distributed Cloud App Connect
Customer Edge node
F5 Distributed Cloud WAAP
F5 Point of Presence (PoP)

Figure 1: How F5 Distributed Cloud secure multi-cloud networking solutions can be deployed to support AI workloads anywhere.

Connect, Secure, and Scale AI Workloads with a Single Platform: F5 Distributed Cloud Services

F5® Distributed Cloud secure multi-cloud networking solutions satisfy the rigorous connectivity, security, and scalability requirements of generative AI, all in one platform.

Distributed Cloud secure multi-cloud networking solution packages are comprised of several different elements, all working together to create a seamless service mesh and network fabric that can secure, deploy, and manage AI apps across hybrid, cloud, and on-premises environments.

Key Features

Enable app layer networking
Leverage proxy-based architecture with granular policy for transparent interconnect and load balancing for TCP, UDP, or HTTP/S, decoupled from the underlying network.

Automate cloud network provisioning
Establish connectivity and security among clouds, data centers, and edge locations with one-click provisioning.

Unify app and API security enforcement
Combine in-line app and API security capabilities with WAF, including a granular L7 policy engine delivering rate limiting, IP reputation, Allow/Deny, and L7 DoS functionality, to control and block API endpoints.

Centralize app and API monitoring and visualization
Identify the most used and attacked endpoints, usage patterns, and sensitive data including PII, plus correlate good and bad actor activity to optimize and tune protection policies for apps and APIs.

Centralize observability and diagnostics
Gain centralized visibility and insights into network, security, apps, and users, eliminating the need to gather data from multiple sources.

- **F5 Distributed Cloud Network Connect** enables layer 3 networking across the entire environment. It includes a virtual router and network firewall with globally orchestrated control for point-and-click connectivity, fully segmented and encrypted in transit, to connect networks across cloud, on-premises, and branch locations. If each environment has a customer edge instance installed, Distributed Cloud Network Connect enables AI workloads to access resources on any of those locations as if it were the same network.
- **F5 Distributed Cloud App Connect** is an integrated stack of layer 7 networking and security services. It includes a distributed load balancer, application firewall, API proxy for app-to-app and cross-cluster API delivery, and API discovery for stealth API security. This provides automatic and scalable orchestration to connect apps among public clouds, data centers, co-location facilities, and edge locations (including retail stores or manufacturing facilities). Distributed Cloud App Connect enables AI workloads to be deployed to any customer edge location without exposing the underlying networking layer, making AI workloads easily deployable and scalable across regions and clouds to meet demand.
- **F5 Distributed Cloud Web App and API Protection (WAAP)** protects and secures any traditional, modern, or hybrid app, including AI. Distributed Cloud WAAP leverages a diverse set of security services with machine learning and globally sourced F5 threat intelligence. These services operate across the F5 global delivery network to enable SaaS-based application protection, including web application firewall (WAF), API discovery and security, bot defense, and distributed denial-of-service (DDoS) mitigation.
- **F5 Distributed Cloud App Stack** is an integral part of the Distributed Cloud platform, built to support AI inference models at the edge. It enables hybrid deployment models, serves as an ingress controller for Kubernetes clusters, and maintains consistent app infrastructure across virtual machine instances and containers. Distributed Cloud App Stack delivers a logically centralized cloud that can be managed using industry-standard Kubernetes APIs. This singular control plane removes the overhead of individually managing Kubernetes clusters and allows customers to automate application deployment, scaling, security, and operations across their entire deployment as a “unified cloud.”

F5 Distributed Cloud secure multi-cloud networking solutions satisfy the rigorous connectivity, security, and scalability requirements of generative AI, all in one platform.

Conclusion

Given the hype and exponential pace of adoption, generative AI can seem incredibly daunting. But in reality, it is just another modern application. F5 Distributed Cloud Services are designed to connect and secure any app, any API, anywhere—from monolithic, legacy enterprise resource planning deployments to modern, lightweight apps built on collections of microservices. Generative AI is one more variation, and one that is naturally supported by Distributed Cloud secure multi-cloud networking solutions from F5.

Next Steps

- Successfully deploy and manage generative AI with robust solutions from F5, start today.
- Scale AI workloads with secure multi-cloud networking solutions from F5. Learn how.