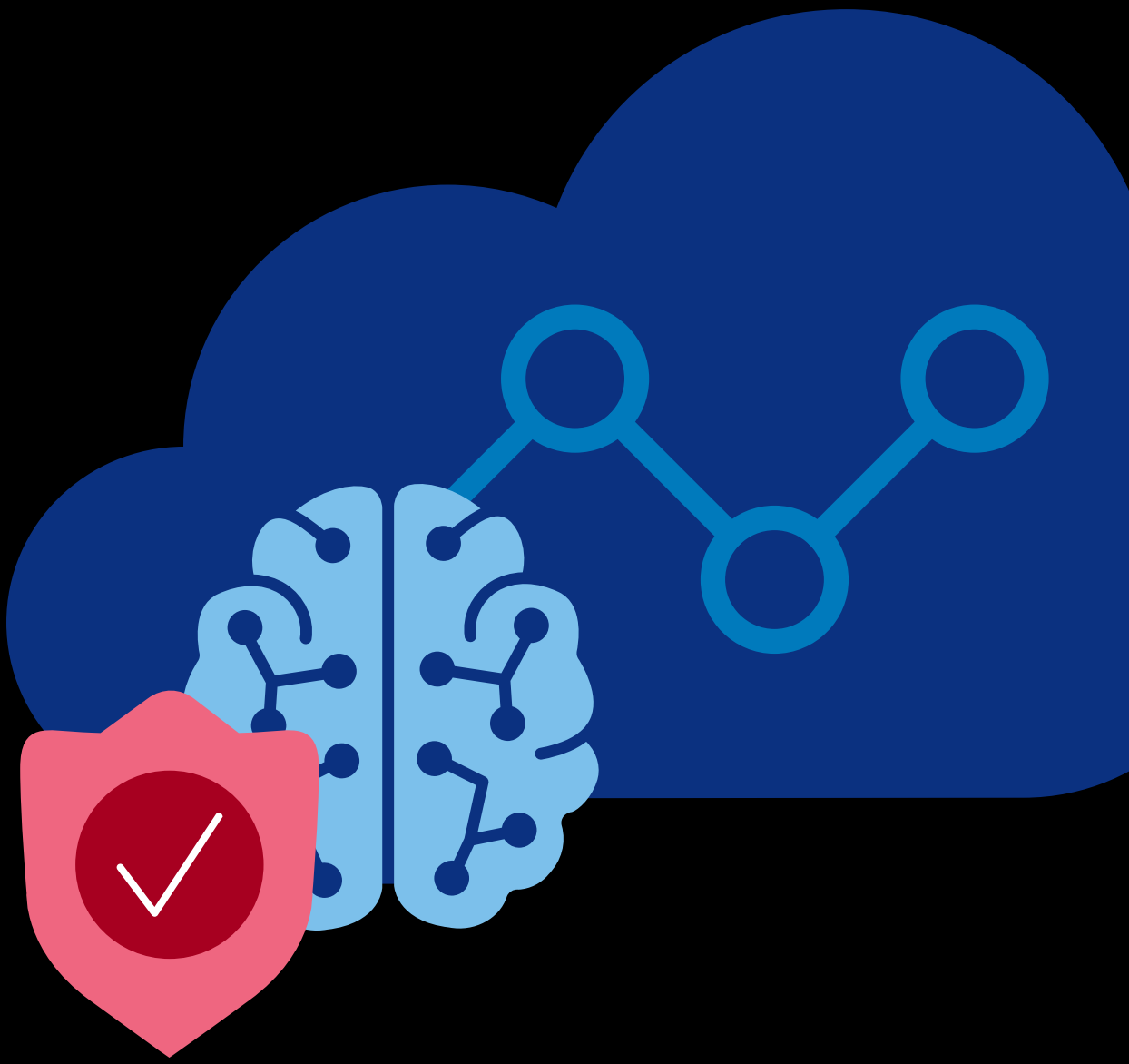
Solution Overview

Distributed Inference: Manage, Secure, and Connect



Generative AI represents a transformative technological shift for the modern enterprise, with artificial intelligence leveraged in new innovative ways using large data sets and natural language learning models. Applications are being deployed to collect data and deliver inference in a variety of locations at the edge—no longer confined to the public cloud. While there are many benefits for organizations adopting AI, there are many challenges related to managing and securing these distributed applications and architecture. Organizations need to address these before they expand their AI investments.

The number of consumer edge-enabled Internet of Things (IoT) devices throughout the world is forecast to grow to almost 6.5 billion by 2030, an increase of more than four billion compared to 2020 (Total global edge-enabled IoT devices 2020-2030, by market published by Thomas Alsop, Dec 16, 2022). The adoption of AI will only exacerbate this figure as new portable application services are deployed across remote edge sites. Also, processing and storing data at the edge reduces latency and bandwidth requirements—but introduces new security challenges.

Expanded Attack Surface with Distributed Sites

Managing and protecting a highly distributed and diverse set of devices and applications, often with limited resources and connectivity, is challenging for many organizations. Devices such as sensors, actuators, and gateways, are often deployed in harsh or remote environments. This makes them vulnerable to physical attacks, tampering, or theft. In addition, these devices often employ a variety of operating systems and protocols, making it difficult to ensure consistent and comprehensive security across the entire network.

The dynamic nature of these environments challenges maintaining a reliable and up-to-date inventory of assets and vulnerabilities. Edge-based devices also typically have limited computing power and battery life, making it harder to leverage sophisticated security measures such as encryption, authentication, and access control—and if those are not deployed, the entire network may be at risk.

Maintaining app-to-app connectivity across multiple environments and locations

Ensuring network connectivity at remote sites is dependent on the availability of reliable network bandwidth in these areas. Remote sites in rural or isolated areas either lack access to traditional network services, or the available options may be prohibitively expensive or unreliable. In such cases, organizations may need to consider alternative connectivity

options, such as satellite, cellular, or mesh networks. They can be more expensive, but they provide greater coverage and flexibility. When AI is deployed, it will often be more efficient to deploy inference models locally in these sites, to ensure faster responses to dynamic conditions where immediate actions are required.

Maintaining Inference Apps at Scale

Maintaining distributed AI inference across hundreds of sites poses significant challenges. Each site requires consistent updates to ensure the latest large language models (LLMs) are in use, which can be complex and time-consuming. Coordinating these updates is difficult, especially when the deployment spans various environments, each with its own infrastructure and compatibility issues. Ensuring seamless integration and consistent performance across diverse platforms complicates deployment further. Moreover, managing the network latency, data synchronization, and security across these distributed systems adds to the intricacies—making the process demanding and resource intensive.

Simplify inference app lifecycle management and security

F5® Distributed Cloud Services is a SaaS-based platform that connects and secures applications deployed across distributed environments. The platform extends serviceability to remote edge sites and branches, enabling localized delivery of services and streamlining application lifecycle management.

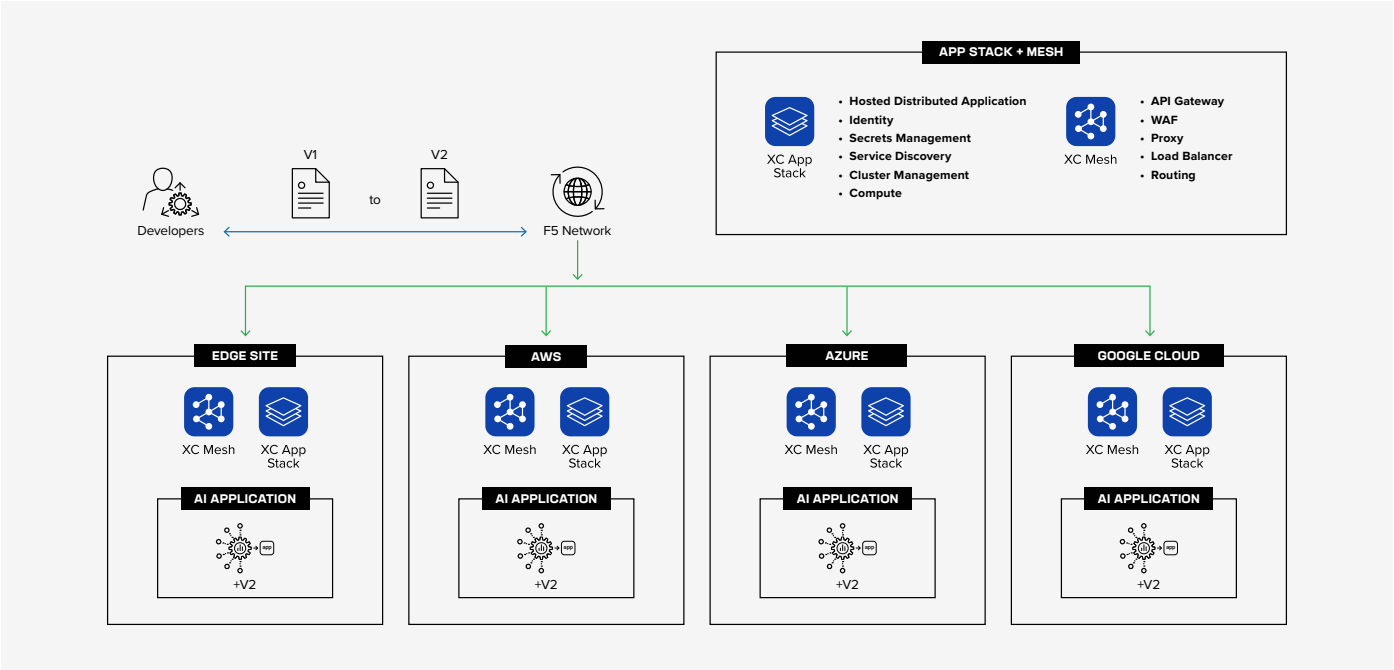


Figure 1: Application Lifecycle Management

The F5® Distributed Cloud Platform provides secure connectivity and reduces the maintenance burden for your AI application deployments.

- **Connectivity.** The platform enables secure connectivity of AI services from the public cloud to far-reaching edge sites.
- **Security.** The F5 global network provides a natively secure and private backbone for all AI applications with additional app security services such as web app firewall, bot detection, DDoS mitigation, and API security.
- **Life cycle management.** Easily deploy software updates from the console to your applications across all distributed sites, ensuring all app infrastructure is available and secure.

F5® Distributed Cloud App Stack is an integral part of the platform that is built to support AI inference models at the edge. It enables hybrid deployment models and consistent app infrastructure to be maintained across VMs 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 many individually managed Kubernetes clusters and allows customers to automate application deployment, scaling, security, and operations across their entire deployment as a “unified cloud”.

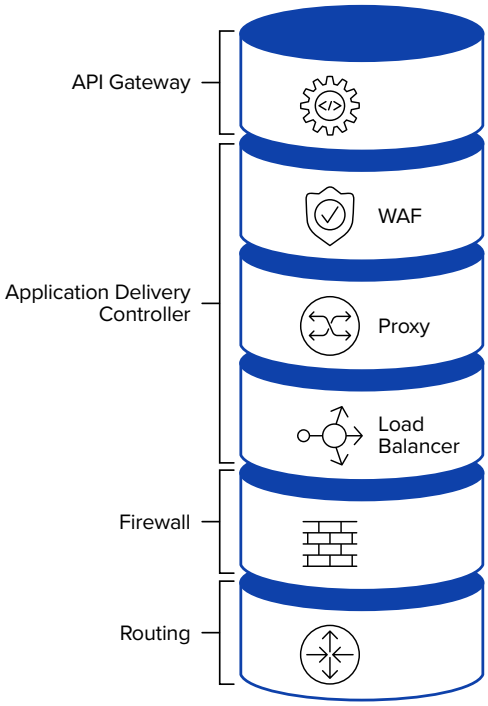


Figure 2: Mesh; Integrated High Performance Networking Stack [L3-L7]

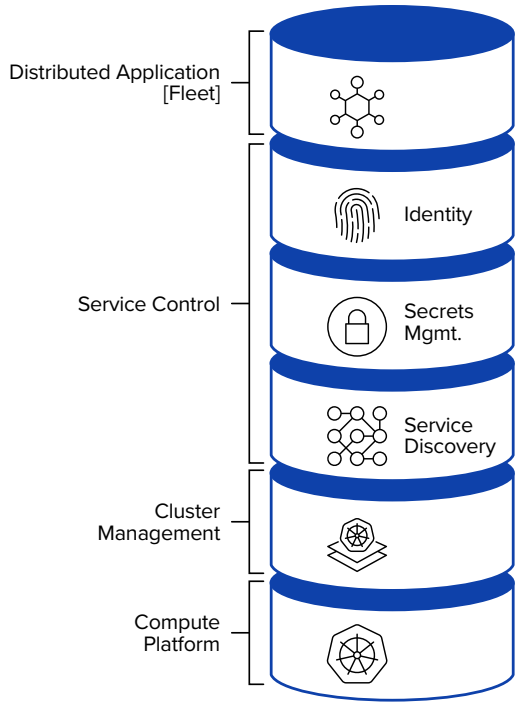


Figure 3: App Stack; Simplified Application Infrastructure Stack

Key Benefits of Distributed Cloud App Stack:

- Improved AI app performance and connectivity
- Enhanced application security
- Edge operational efficiency and simplification
- Scalability to bring on more sites quickly

Key Capabilities:

- Efficiently manage the complete lifecycle of applications and infrastructure services across multiple distributed sites, incorporating intuitive policies and configurations based on user intent.
- A single control plane across F5's global infrastructure, capable of scaling seamlessly to accommodate large numbers of application clusters.
- Seamlessly integrate web and API protection services within the platform, offering comprehensive security features such as Layer 7 security, API discovery, and robust defense against DDoS attacks and malicious bots.

- Enhance the performance of latency-sensitive applications by strategically deploying them, or specific components, to edge sites or the network edge. Harness the power of AI/ML applications at the edge and leverage GPU as a Service within the network for optimal performance.
- Specifically crafted to streamline Kubernetes operations from Day 0 to Day 100, providing the simplest methods for creating, managing, upgrading, and maintaining clusters.
- Distributed Cloud App Stack can be deployed anywhere to support your distributed applications—on F5 infrastructure at the regional edge or in the public or private cloud. It can be deployed on commodity hardware and support applications built using either containers or virtual machines.

Distributed Cloud App Stack deployment options:

Public clouds	Customer Edge	F5 Regional Edge	Bring your own Kubernetes
Manage and protect application workloads hosted across clouds, including AWS, Azure, and GCP.	Manage and protect applications at your data center and edge sites.	Manage and protect application workloads from any of locations of presence on the F5 global network.	Leverage existing managed container platform deployments and protect and connect them to the F5 global network.

Distributed Cloud App Stack stands out for its scalability and flexibility in deployment options, making it adaptable to any environment. The Distributed Cloud Platform, coupled with Distributed Cloud App Stack, is a powerful solution for organizations that are navigating the complexities of deploying new generative AI infrastructure—ensuring efficiency, security, and scalability across highly distributed, heterogeneous environments.

To learn more, contact your F5 representative, or visit F5.com.