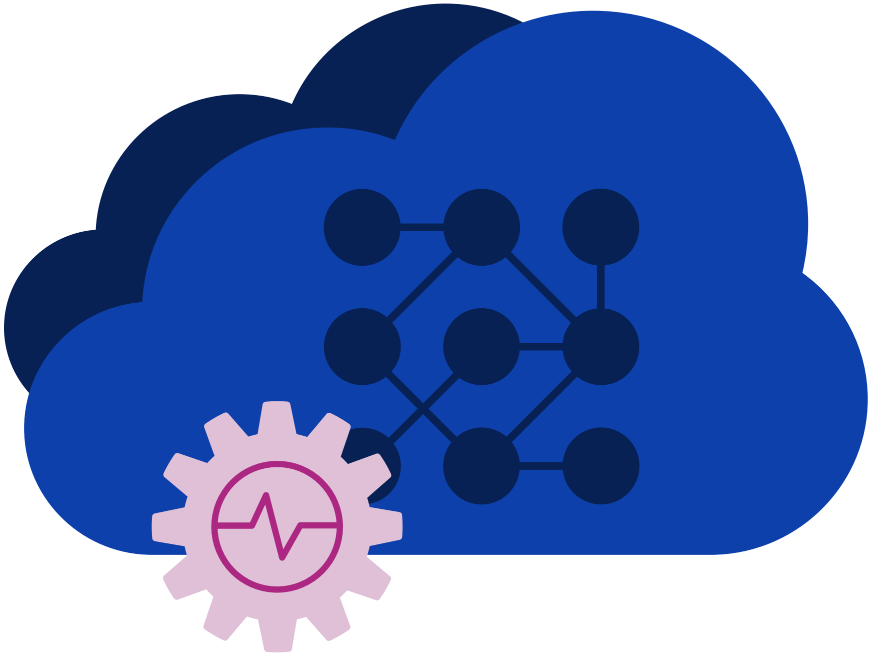
Securely connect and deliver AI apps across clouds

Build resilient AI infrastructure that seamlessly integrates data sources while maintaining security and performance

Contents

Multicloud strategies unlock the full potential of AI applications

Apps today are everywhere. Nearly 90% of surveyed organizations deploy apps in multiple environments,[1] and AI workloads are expected to follow this hybrid approach, too. All workloads are increasingly distributed across hybrid and multicloud infrastructures, driven by cost considerations—and in the case of AI—the strategic advantages of specialized AI services from various providers. Organizations are carefully constructing multicloud environments to provide the precise combination of performance, data proximity, security, and AI tools needed to meet their business objectives. These purpose-built digital foundations maximize AI capabilities while maintaining compliance with evolving regulations.



Why organizations are choosing multicloud for AI

Flexibility

Organizations can strategically distribute AI workloads across multiple platforms, ensuring each task is handled by the most suitable infrastructure and location, reducing the risk of downtime.

Scalability

Multicloud environments facilitate real-time scaling of computational power for data analysis and model training to manage varying workload demands without overprovisioning resources.

Tool diversity

Different environments may offer specialized AI services, allowing businesses to leverage specific AI functionalities that are more advanced or provide unique features that are only available on certain platforms.

Iterative development

With access to a wider array of development environments, multicloud supports more agile AI development processes to test and refine AI models across different cloud infrastructures.

AI applications face unique challenges in multicloud environments

While multicloud strategies offer significant advantages for AI workloads, they also introduce a complex set of challenges to navigate for successful outcomes.

Enhanced security concerns

AI apps attract different threats, creating security considerations that go beyond traditional apps. The distributed nature of multicloud AI compounds these concerns, making it difficult to enforce consistent security.

- **Model protection:** Preventing model theft and unauthorized access to proprietary AI assets becomes more difficult across environments.
- **New attack vectors:** Techniques like prompt injection require specialized detection and prevention mechanisms.
- **Inconsistent controls:** Each environment has different security capabilities, causing management overhead and visibility gaps.

Distributed data sources

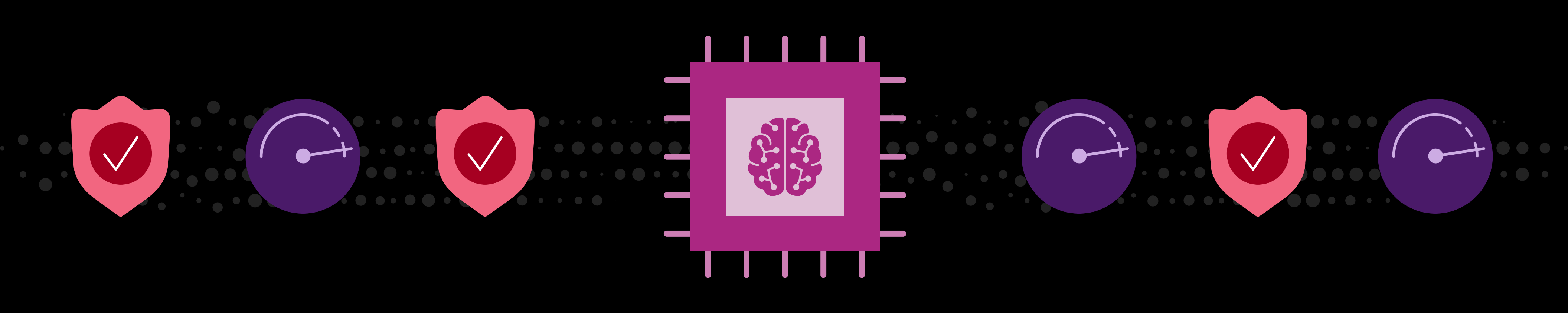
AI applications depend on data that often spans multiple environments. Data used for training or retrieval-augmented generation (RAG) may reside in on-premises data centers due to privacy requirements, while inference services run in public clouds to leverage scale.

- **Inefficient data movement**: Moving large datasets between environments introduces latency and bandwidth constraints.
- **Consistency challenges:** Maintaining coherent data across environments becomes difficult as data evolves.
- **Compliance complexities:** Different regional regulations may require specific data handling practices based on geographic location.

Network performance and reliability issues

Fast, reliable connectivity between AI components is essential for both model training and inference. The massive data volumes and computational requirements of AI workloads amplify the impact of network limitations.

- **Latency sensitivity:** AI inference services require consistently low latency to maintain responsiveness.
- **Bandwidth requirements:** Training processes depend on efficient data transfer between storage and compute resources.
- **Network security complexity:** Each connection between environments represents a potential security boundary that must be protected.



Secure AI connectivity requires specialized architecture components

Overcoming the challenges of multicloud AI requires infrastructure components that work together to create secure, high-performance pathways for AI workloads while maintaining visibility and control across distributed resources.

Private connectivity backbone

- AI workloads need secure, reliable communication pathways that minimize exposure while maintaining high performance.
- Low-latency connections via a distributed network that's closer to users with optimized routing
- A software-defined network with simple, one-click deployment
- Network points of presence that span regions for global distribution and compliance
- Private network segments to protect sensitive data

Unified security framework

Maintaining consistent security across environments is essential for AI workloads that span multiple clouds and on-premises resources.

- Single control plane for uniform policy definition and enforcement
- Specialized detection for AI-related threats like prompt injection and model abuse
- Authentication mechanisms working consistently across cloud boundaries
- Encrypted communication channels protecting sensitive data

Intelligent traffic management

AI applications generate complex traffic patterns that require sophisticated routing and optimization.

- Traffic distribution based on application needs and resource availability
- Centralized API discovery and management with authentication and rate limiting
- Performance tuning for specific AI communication patterns
- Intelligent caching of AI responses to reduce redundant processing

End-to-end observability

Managing AI deployments across environments requires detailed insights into performance, security, and resource utilization.

- Unified dashboards for visibility into apps and networks across all environments
- Predictive analytics to identify potential issues earlier
- Comprehensive logging and analysis for improved threat detection
- Visibility into resource consumption to control spending

By implementing these essential components, you can build the foundation for secure, high-performance AI applications that operate seamlessly across hybrid and multicloud environments.

F5 and AWS deliver secure AI connectivity

The partnership between F5 and AWS provides organizations with integrated solutions that address the unique challenges of AI applications in multicloud environments. With over a decade of collaboration, this alliance enables you to build, connect, and secure distributed AI workloads.

Security and application delivery solutions

F5 provides the essential connectivity and security foundation for AI applications, spanning from the public cloud to remote edge locations.

- **Global network infrastructure:** The F5® Global Network delivers high-performance, private connectivity between environments with built-in security.
- **Distributed deployment models:** Flexible deployment options support AI workloads across public and private clouds, private data centers, and edge locations.
- **AI-specific security:** F5® AI Gateway protects AI models and apps against abuse.
- **Centralized management:** A single console for F5® Distributed Cloud Services provides consistent visibility and control across all environments and deployment locations.

AI services and infrastructure

AWS offers specialized services for building and operating AI applications at scale:

- **Managed AI models:** Amazon Bedrock is a fully managed service to access foundation models through a unified API without infrastructure management.
- **End-to-end ML platform:** Amazon SageMaker enables building, training, and deploying AI models.
- **Compute options:** Specialized instance types are optimized for AI training and inference workloads.
- **Global infrastructure:** Extensive regional presence supports AI deployment with local data residency requirements.
- **Edge solutions:** Services like AWS Local Zones and AWS Outposts extend AI infrastructure to edge locations with low-latency requirements.

NetApp data management integration

F5 and AWS work with NetApp storage solutions, including NetApp BlueXP and NetApp ONTAP, to unify data management and secure data connections for AI workloads.

Secure connections between AI models and data sources accelerate RAG implementation

RAG represents one of the most powerful use cases for AI in the enterprise, allowing you to enhance foundation models with your organization's proprietary data for more valuable and accurate responses. Implementing RAG effectively requires secure, high-performance connections between AI models and distributed data sources across your hybrid and multicloud environments.

Secure multicloud networking in Distributed Cloud Services unifies data for RAG through:

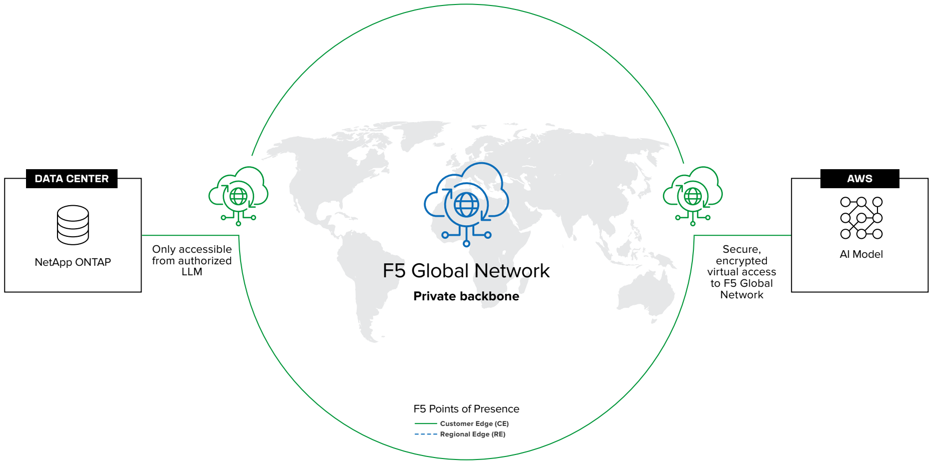
- **One-click deployment:** Create connections between cloud and on-premises environments quickly and easily
- **Layer 3 network connectivity:** Establish secure pathways between model inferencing and data volumes with F5® Distributed Cloud Network Connect
- **TCP and HTTPS load balancing:** Prevent networking issues like overlapping IP ranges and protect data connections with support for Amazon S3 protocols and robust security controls
- **Private network transit:** Leverage the F5 Global Network to securely move data between environments without public internet exposure

NetApp storage integration for RAG

NetApp and F5 work together to unify data management for RAG implementations. Connect NetApp storage volumes on AWS, other clouds, or on premises directly to AI models without migration or copying, maintaining the integrity of your data. This integration helps you maintain control of sensitive information while making it available for AI processing, balancing AI value with security requirements.

Benefits

- Enhanced value from AI investments
- Effortless data migration across zones and regions that maintains performance
- Enterprise data protection at rest an in transit
- Granular positive security controls for proprietary data
- Data access restricted to only authorized AI models



Bring AI workloads to the edge with secure connectivity between cloud and remote locations

AI workloads are increasingly moving beyond centralized data centers to edge locations, where they can process data closer to its source and deliver low-latency insights. This distributed approach creates new opportunities for innovation, such as industrial IoT monitoring and video analysis. However, edge environments often lack the robust network infrastructure and security controls available in traditional data centers. Limited computing resources, intermittent connectivity, and inconsistent security all complicate AI deployments at the edge.

Extending cloud AI capabilities to the edge

With Distributed Cloud Services, you can securely connect edge locations to your cloud-based AI infrastructure. The platform extends networking and security services to remote edge sites, creating consistency regardless of location. Even on-premises edge sites can run Distributed Cloud Services by using F5® Distributed Cloud Customer Edge deployed locally to run and secure applications. This approach allows you to deploy AI inferencing at the edge while maintaining secure connections to centralized models and data sources.

You can also run AI Gateway at the edge to be closer to your AI apps. Its Kubernetes-based architecture means it can be readily deployed wherever needed to provide load balancing and security designed for AI.

Use F5 solutions with AWS edge services like AWS Local Zones and AWS Outposts, which bring AWS infrastructure and services closer to your applications. By combining F5's security and connectivity solutions with AWS edge infrastructure, you can create a seamless experience that spans from cloud to edge.

BENEFITS

- Faster response times for critical AI applications
- Reduced data transfer costs by processing locally
- Improved customer experiences through real-time insights
- Enhanced operational efficiency in remote locations
- Greater business continuity despite connectivity issues
- Minimized security risk across distributed systems

F5 and AWS are addressing key AI challenges together

F5 and AWS deliver solutions to build, secure, and optimize AI-powered applications. Backed by a robust portfolio and over a decade of joint engineering and innovations, F5 and AWS provide an easier path to innovation while protecting both your sensitive data and the user experience through:

Security consistency: Enable uniform protection policies across all environments where AI workloads operate

Performance optimization: Combine F5 traffic management with AWS infrastructure for responsive AI experiences

Data accessibility: Unify secure access to distributed data sources

Operational simplicity: Reduce the complexity of maintaining AI apps across environments

Learn more about F5 solutions for AWS at f5.com/aws.

THE F5 AND AWS PARTNERSHIP

10+

years of collaboration

Over 1K

joint customers

Competencies for containers, networking, and security

Service validations for AWS WAF, Amazon CloudFront, AWS Outposts, and Linux

Appendix

[1] F5, 2024 State of Application Strategy Report, May 2024

ABOUT F5

BRINGING A BETTER DIGITAL WORLD TO LIFE

F5, Inc. (NASDAQ: FFIV) is the global leader that delivers and secures every app. Backed by three decades of expertise, F5 has built the industry's premier platform—F5 Application Delivery and Security Platform (ADSP)—to deliver and secure every app, every API, anywhere: on-premises, in the cloud, at the edge, and across hybrid, multicloud environments. F5 is committed to innovating and partnering with the world's largest and most advanced organizations to deliver fast, available, and secure digital experiences. Together, we help each other thrive and bring a better digital world to life.

For more information, go to f5.com.

Learn more about F5 solutions for AWS at f5.com/aws.