IDC

SPOTLIGHT

Sponsored by: F5

Flexible solutions for AI initiatives — both in functionality and deployment form factors — are essential for streamlining deployments and ensuring robust security in a rapidly evolving landscape.

AI Training and Inference Infrastructure for Applications and Data Requires Optimized Security and Performance

November 2025

Written by: Paul Nicholson, Research VP, Cloud and Datacenter Networks

Introduction

The adoption of AI is rapidly evolving, with organizations moving from the initial training phases to generative AI (GenAI) deployment in production environments. Organizations are increasingly recognizing this as a strategic initiative, which CIOs often lead. According to IDC research, while many are still taking an opportunistic approach to AI, industry leaders are shifting toward a more structured methodology. Such a strategy is becoming essential as the complexity of AI solutions grows and as GenAI becomes embedded across business units, requiring organizations to be agile when responding to new demands.

Looking ahead to 2026 and 2027, the focus of AI adoption will shift decisively from ad hoc and siloed deployments to repeatable, scalable practices. This transition enables organizations to embed key attributes such as performance and security into their AI initiatives, reducing the need for reactive add-ons and ensuring robust, efficient, scalable, and secure operations from the outset. The need to support a growing number of GenAI applications — which rely on data sourced from multiple locations, including external partners, to deliver specialized insights — will drive this evolution. The expansion of interconnection and the need for flexible, reusable architectures will be critical for meeting diverse use cases and enabling organizations to pivot as business needs change while consistently ensuring reliability, security, and compliance.

Successful GenAI deployments depend on coordinated efforts across traditionally siloed teams, including network, infrastructure, storage, cloud, and data science. IDC research shows that enterprises are planning to deploy a vast array of new AI applications or augment existing ones, often in hybrid environments that span public and private clouds as well as on-premises datacenters and the network edge. Connections to external partner data sources are also becoming integral to the AI ecosystem.

AT A GLANCE

WHAT'S IMPORTANT

AI application deployments are expanding, even as IDC research indicates that the initial scramble of ad hoc deployments has resulted in varying levels of success.

KEY TAKEAWAY

Organizations seek a repeatable approach to align AI initiatives and achieve greater success in meeting business objectives. The AI application landscape is complex, rapidly evolving, and highly distributed. To support AI transformation goals, robust and repeatable infrastructure for application deployment, data delivery, and interconnection is essential.

Recent IDC data highlights the scale and speed of this transformation. According to IDC's 4Q24 AI in Networking Special Report, organizations expect that, within two years, 54% of the corporate data they use for GenAI training and inferencing will reside within private on-premises systems rather than public cloud platforms. This prediction underscores the importance of robust data delivery pipelines, such as retrieval-augmented generation (RAG), to ensure complete and accurate inference results.

In addition, survey data from IDC's 2025 IaaS Network Services: Requirements, Adoption, and Impact Special Report reveals strong momentum toward hybrid and multicloud architectures in general. Over three-quarters (75.8%) of respondents already operate a hybrid cloud, while 17.6% plan to adopt one in the next year. Similarly, almost two-thirds (65.6%) currently operate a multicloud environment as 31.4% plan to do so soon. These trends reflect a clear preference for advanced multicloud use cases.

The race to deliver GenAI applications is accelerating. IDC's AI in Networking Special Report found that only 1% of organizations have no plans to roll out GenAI in the next year, while a remarkable 74.4% plan to integrate GenAI into at least 11 applications (with some organizations planning for over 30 applications). This unprecedented wave of rollouts demands that networking and IT departments ensure performance, reliability, and security. Uniform AI infrastructure and AI/MLOps practices will be essential for consistently and quickly bringing these applications from proofs of concept (POCs) to production successfully.

To meet these challenges, organizations must prepare their data environments and coordinate across IT silos to implement standardized, performant, and secure solutions, while also avoiding unnecessary budget expenditures. To ensure compliance and data protection, they must also address unique requirements for sovereign datacenters and healthcare and financial applications with proprietary data.

Definitions

- **Application delivery controller (ADC)** is a network device or software that manages, secures, and optimizes traffic between applications and users.
- **S3 protocol** is an API standard for storing, retrieving, and managing data in object storage systems. Amazon Simple Storage Service (Amazon S3) originated in 2006. However, multiple vendors have widely adopted the S3 protocol.
- **Data ingestion** is the process of collecting and transferring data from various sources into a system for processing, analysis, or storage. This process is used for the following AI use cases:
  - **RAG** is an AI technique where a model retrieves external data sources to improve the quality and relevance of its responses.
  - **Training data** is a data set that teaches an ML model to recognize patterns and make predictions.
  - **Fine-tuning** is the process of adapting a pretrained model by training it further on domain-specific data to improve accuracy for a particular task.

Benefits

GenAI-ready infrastructure, which enables repeatable and highly flexible operationalization, increases AI deployment success by optimizing performance, application awareness, security, and multi-environment data access — thus ensuring

uptime, privacy, compliance, and adaptability for evolving enterprise needs. A robust, GenAI-ready infrastructure improves the likelihood of successful AI deployments by delivering repeatable, flexible, and secure operational capabilities across diverse environments. Such an infrastructure offers benefits in the following areas:

- **Performance, traffic distribution, and scale:**
  - Optimized flows by host, user, or object ensure efficient resource allocation and scalable traffic management, customized to an organization's application and infrastructure requirements.
- **Application awareness:**
  - This includes HTTP and AI-specific protocols, such as S3 for data delivery and Model Context Protocol (MCP).
  - Abstraction and loose coupling enable seamless maintenance, vendor switching, and smooth transitions from pilot to production.
  - Application awareness and programmability allow dynamic traffic adjustments and bridge compatibility gaps between environments.
  - Flexible architecture supports S3 object storage from multiple vendors, enabling organizations to pivot without disrupting business goals or timelines.
- **Multi-environment data and inference:**
  - This takes the form of structured and accessible data across all compute and networking environments.
  - Proactive network and application-level health checks, leveraging application awareness for rapid issue resolution, achieve higher uptime.
  - Optimized global and inter-datacenter resources offer intelligent redirection based on location, user type, or organization.
  - Robust DDoS protection supports the reliable availability of critical resources.
- **Reduced liability and guaranteed data privacy for sensitive data:**
  - These address risks such as data poisoning or leakage.
  - End-to-end encryption for data in motion protects sensitive information throughout its life cycle.
  - Robust authentication and authorization controls ensure that critical data access is subject to approval.
  - Comprehensive accounting and observability enable effective monitoring, auditing, and compliance.
- **Off-the-shelf and DIY solutions**
  - Off-the-shelf solutions favor repeatability, which can provide predictable costs, reliable support, and consistent features. DIY approaches offer flexibility and adaptability to specific needs, allowing for tailored solutions. But this customization may come with higher integration costs and the increased risk of support and feature limitations.

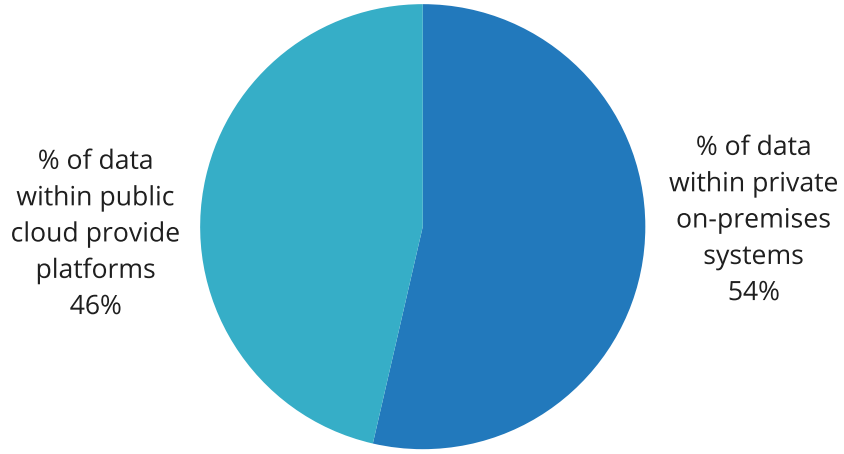
Trends in AI Deployment Show Many Applications in Multiple Locations

IDC's *AI in Networking Special Report* highlights an acceleration in GenAI application deployment (see Figure 1). It is apparent that the trends in application and datacenter architecture must undergo upgrades to meet the rising AI-related demands. In detail:

- When IDC asked organizations about their plans for integrating GenAI model results in 2025, 1% reported no planned rollouts, 24 .6% anticipated deploying GenAI in 1–10 applications, and 74.4% expected integration across at least 11 applications (with some organizations planning for more than 30 applications). As organizations increasingly augment multiple business and IT applications with AI, the demand for adaptable, multipurpose infrastructure that can scale rapidly becomes critical to support this surge in rollouts.
- In the same report, organizations project that, within two years, 54% of the corporate data they use for GenAI training and inferencing will reside within private on-premises systems rather than public cloud platforms. This can be further influenced not just by cost and compliance/privacy but also large data sets that are difficult to move (i.e., the concept of data gravity). In such cases, it is deemed more beneficial to bring the application to the data. This shift underscores the growing need for distributed data environments and robust interconnection strategies. For GenAI inference, secure and reliable access to diverse data sources — both internal and external — will be essential to ensure performance, accuracy, and compliance.

FIGURE 1: ***Planned Data Placement for GenAI Training and Inferencing in Two Years***

Q ***In two years, what percentage of your organization's corporate data that you use for GenAI training and inferencing will be located within private on-premises systems versus public cloud provider platforms?***



n = 1,209

Source: IDC's Worldwide AI in Networking Special Report, 2024

Considering F5 to Augment AI Application and Data Delivery

Founded in 1996, F5 has evolved into a leading provider of application delivery and security solutions.

F5 offers the BIG-IP suite as a core component of its Application Delivery and Security Platform (ADSP). The platform integrates essential services to provide consistent security, high availability, and intelligent orchestration for applications, APIs, and components across edge, cloud, and on-premises environments. Designed to support demanding workloads, BIG-IP enables organizations to deliver and protect their digital assets with reliability and flexibility.

In wide use today, ADC technology can provide advanced solutions in the delivery of AI applications that may be absent in more basic implementations. For example, "load balancing" does not denote the advanced capabilities that an ADC provides, even though an ADC typically has full load-balancing capabilities. Advanced application capabilities can be important when rolling out AI applications. In addition to essential networking and security functionalities, they also provide features that can overcome deployment issues both foreseen and unforeseen. Thus they can be critical to ensuring successful and timely rollouts, a process that has proven to be challenging in early and current AI deployments.

Examples of advanced application delivery include:

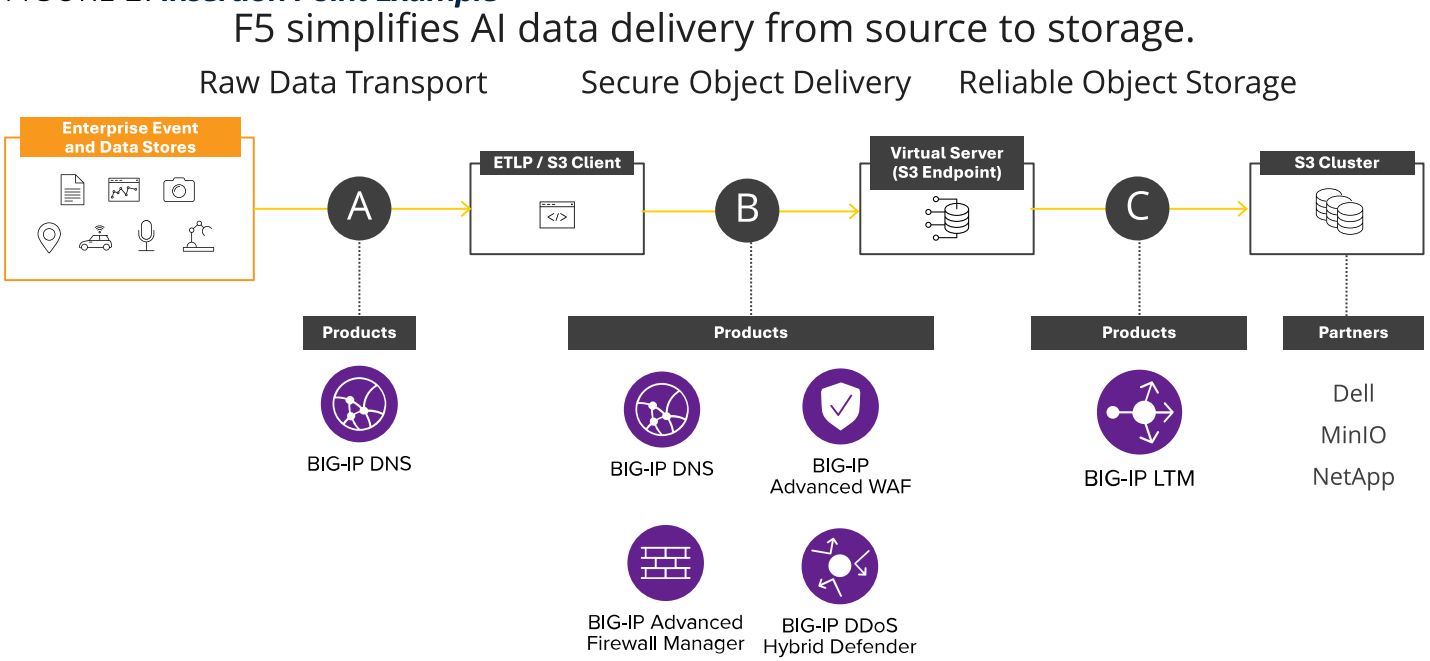
- **Application awareness** enables deeper visibility and control over application traffic as compared with traditional layer 4 solutions, allowing for more granular management of protocols, user sessions, and application-specific requirements.
- **Global server load balancing** provides intelligent distribution of traffic across multiple geographically dispersed datacenters, improving application availability, performance, and disaster recovery capabilities.
- **Programmability/iRules flexibility** supports advanced customization through scripting and iRules, empowering organizations to tailor traffic management, security policies, and application behavior to meet specific business needs and respond quickly to changing requirements.
- **Optimization** enhances application performance by ensuring that S3 nodes operate optimally, avoiding overloads and hotspots, implementing throttling and QoS to prevent resource starvation, and supporting features such as caching and compression for efficient resource utilization and a better end-user experience.
- **Security** supports TLS encryption, strong authentication mechanisms, and integrated web application firewall capabilities to protect applications from threats and ensure compliance with security standards.

Data organization and data transport are key to successful rollouts and diverse delivery use cases, applicable to any data ingestion scenario including training, fine-tuning, or RAG. Further:

- **Raw data transport** facilitates the efficient movement of unprocessed data between sources and destinations, ensuring that foundational data sets are available for AI workflows and analytics without unnecessary transformation delays.
- **Secure object delivery** ensures that data objects are transmitted with robust security measures such as encryption and access controls, protecting sensitive information during transfer and meeting compliance requirements for privacy and data protection.
- **Reliability for object storage** provides dependable access to stored data objects, leveraging redundancy, health checks, and failover mechanisms to guarantee data availability and integrity for critical AI processes and business operations.

F5 capabilities can facilitate multiple use cases and have multiple different insertion points. Figure 2 shows data ingestion with these solutions inserted in different stages:

FIGURE 2: ***Insertion Point Example***



Source: F5, 2025

F5 BIG-IP enables a wide range of application delivery and security solutions, offering deployment flexibility across hardware, software, container network functions, and consumables via public cloud marketplaces. This versatility enables organizations to select the optimal form factor to meet their evolving infrastructure needs, as they may require a combination. In detail:

- **BIG-IP Local Traffic Manager:** Full proxy architecture enforces network protocols and protects endpoints by controlling and optimizing application traffic flows. Application layer functionality, including advanced awareness (e.g., S3) and application data scripting, increases the number of achievable use cases as compared with what is possible with more basic capabilities.
- **BIG-IP DNS:** High-performance, scalable DNS and global server load balancing services are delivered for fast, reliable application access and optimal site redirection/selection for application processing.
- **BIG-IP Advanced Firewall Manager:** Robust network-layer security is provided to block malicious traffic and safeguard infrastructure.
- **BIG-IP Advanced Web Application Firewall:** Web applications are protected from advanced threats and vulnerabilities with intelligent, adaptive security.
- **BIG-IP DDoS Hybrid Defender:** It defends against volumetric and targeted DDoS attacks, ensuring continuous application availability and traffic regulation.

F5's technology alliance partnerships focusing on S3 extend beyond AWS to encompass leading storage providers such as NetApp, MinIO, and Dell, all of which leverage the S3 protocol for data integration. These collaborations enable organizations to optimize AI data delivery and storage across hybrid multicloud environments, supporting greater flexibility and scalability for modern workloads.

F5 continues to expand its AI-focused solutions with advanced support for MCP in BIG-IP version 21, introducing standardized traffic management that enables load balancing, optimization, and enhanced security for AI applications (and data sources). In addition, F5 has strengthened its AI security portfolio through the acquisition of CalypsoAI in September 2025, integrating AI Guardrails and AI Red Team capabilities. These tools empower enterprises to validate, monitor, and safeguard AI model behavior, enabling robust protection and compliance as AI deployments become more complex and widespread.

Challenges

IT departments are frequently organized in silos, which can make cross-team coordination challenging and can hinder efficient operations. F5 will have to reach the technical decision-makers in these different groups while also identifying the budget holders. It should continue to demonstrate to stakeholders how its unified platform consolidates multiple vendor solutions, enabling organizations to streamline their processes and reduce operational complexity. This integrated approach helps align previously siloed groups with a consistent strategy for application delivery and security, which is especially important for managing AI data delivery.

For traditional ADC users, the concept of S3 support for data delivery may be unfamiliar and require additional explanation. As organizations expand their use of S3 for AI and other data-driven applications, they must educate both existing users and the broader market about the benefits and implementation of S3 support within ADC environments.

While some organizations continue to rely on private connections, basic firewalling, and encryption for data distribution, this approach is becoming less common as distributed data sources and interconnection proliferate. The inability to flexibly accommodate diverse data sources can impede AI application rollouts and leave organizations without the advanced security features necessary to protect sensitive information in increasingly complex environments.

Conclusion

Flexible solutions for AI initiatives — both in functionality and deployment form factors — are essential for streamlining deployments and ensuring robust security in a rapidly evolving landscape with the challenges (both predictable and unpredictable) that emerging and evolving technologies typically face.

Organizations that invest in adaptable platforms are well positioned to enhance their AI initiatives and have a greater chance of success, meeting current needs while remaining agile as requirements and technologies advance.

About the Analyst

Paul Nicholson, *Research VP, Cloud and Datacenter Networks*

Paul Nicholson is IDC's research vice president for cloud and datacenter networks. He provides thought leadership and actionable insights on cloud and datacenter networking markets and technologies. Paul has a deep understanding of the networking market along with its business and application requirements, technologies, product road maps, competitive differentiation, and go-to-market strategies, enabling him to provide informed guidance for vendors, cloud providers, and enterprise IT buyers and practitioners. Paul works closely with IDC's Enterprise Networking, Server, Cloud, and Security research teams on assessing the impact of emerging IT, networking, and interconnect infrastructures.

MESSAGE FROM THE SPONSOR

Enterprises accelerating AI adoption face significant challenges in delivering massive data sets for training, fine-tuning, and retrieval-augmented generation. As workloads scale, data must flow seamlessly across hybrid and multicloud environments while remaining secure, observable, and resilient. Without the right approach, organizations risk GPU underutilization, data poisoning, bottlenecks, and stalled deployments.

The F5 Application Delivery and Security Platform (ADSP) provides a programmable control point for AI data delivery. By unifying traffic management, policy-driven security, and endpoint resilience, the F5 ADSP enables organizations to abstract applications from specific storage back ends while maintaining availability and governance. This ensures data pipelines remain flexible, observable, and performant as infrastructure evolves.

With the F5 ADSP, enterprises can standardize data ingestion and delivery processes that scale efficiently and reliably across environments.

To learn more, visit https://www.f5.com/ai.

IDC Custom Solutions

The content in this paper was adapted from existing IDC research published on www.idc.com.

IDC Research, Inc.
140 Kendrick Street
Building B
Needham, MA 02494, USA
T 508.872.8200
F 508.935.4015
blogs.idc.com
www.idc.com