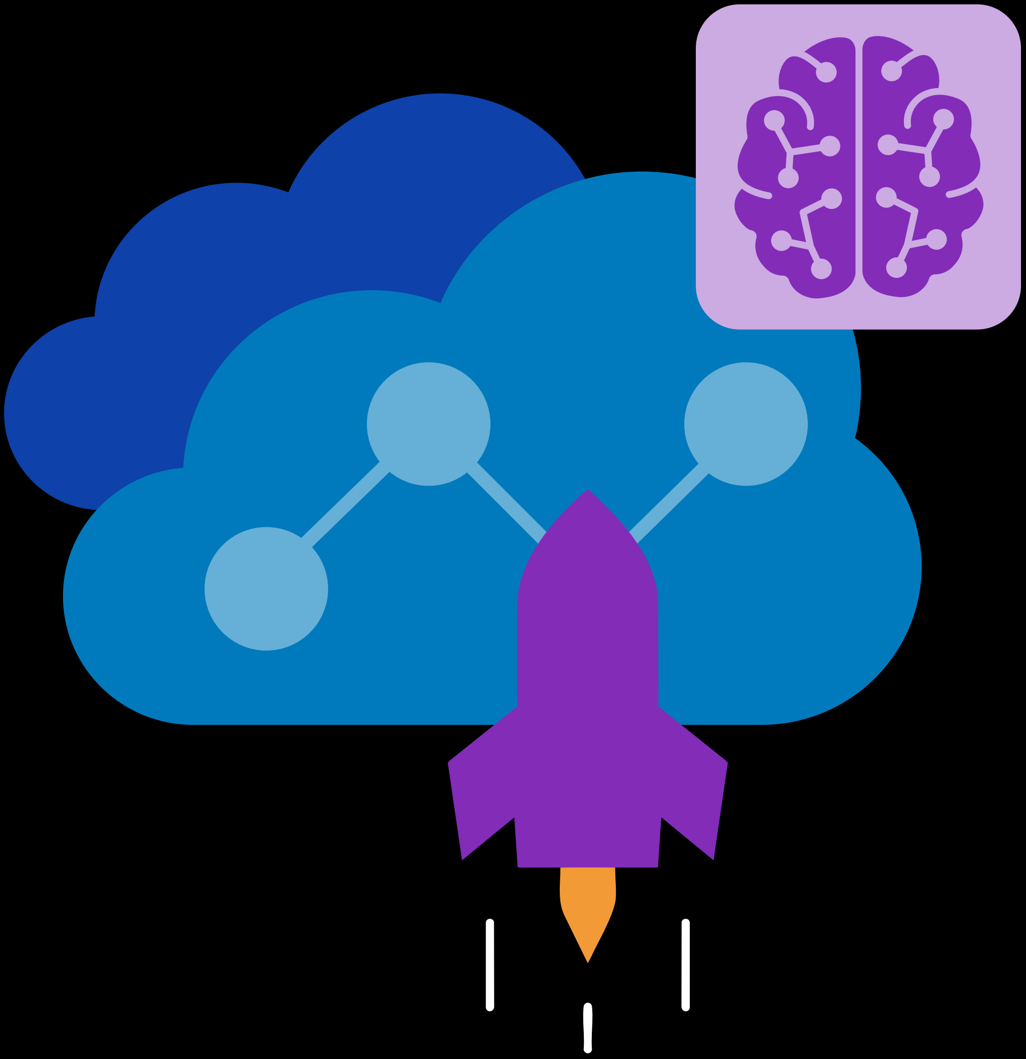
SOLUTION OVERVIEW

High-performance AI data delivery for S3 storage with F5 and NetApp

Deliver AI data at scale with F5® BIG-IP® load balancing S3 workloads for NetApp StorageGRID.



Key benefits

Shape, prioritize, and load balance large dataset transfers to maximize throughput and protect object stores from being overwhelmed by surging AI workloads.

Improve resiliency of storage architectures to serve high transfer rates and large datasets.

Secure sensitive AI data in transit and protect against OWASP Top 10, denial-of-service attacks, and API threats.

Enable performance-based, QoS-aware traffic steering, leveraging S3 protocol headers.

Surge in demand for high-rate S3 file ingestion

AI model training, fine tuning, and RAG workflows, require scalable, highly performant ingestion of files. Legacy network-attached storage (NAS) file sharing protocols such as SMB or NFS are unable to deliver the scale required, shifting the industry towards S3 object storage which utilizes single API calls over HTTPS. NetApp StorageGRID S3 object storage are large, high-capacity servers designed to store and manage vast amounts of data within a distributed storage cluster—ideal for demanding AI workloads.

Advanced traffic management for S3 storage

NetApp StorageGRID architectures can be augmented with F5® BIG-IP®, an HTTPS load balancer that can read and write to dozens of enterprise nodes through a simplified target address. BIG-IP is a core component of the F5 Application Delivery and Security Platform (ADSP), which extends across your entire technology stack. The F5 ADSP ensures that every app and API benefits from consistent, comprehensive security and performant delivery. BIG-IP also supports TCP profiles designed for high concurrency and massive amounts of parallel HTTPS sessions, applicable to large-scale S3 architectures, to help ensure consistent AI data delivery at scale.

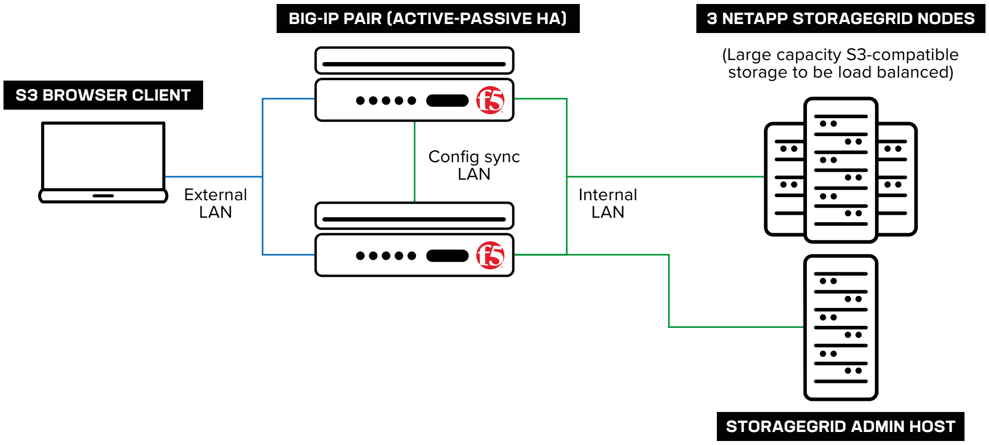

BIG-IP PAIR (ACTIVE-PASSIVE HA)
3 NETAPP STORAGEGRID NODES
(Large capacity S3-compatible storage to be load balanced)
S3 BROWSER CLIENT
Config sync LAN
External LAN
Internal LAN
STORAGEGRID ADMIN HOST

Figure 1: Shape, prioritize, and load balance massive dataset transfers with F5 BIG-IP

S3 high availability

BIG-IP maintains the integrity of the NetApp nodes through frequent HTTP-based health checks. Should an unhealthy node be detected, it will be dropped from the list of active pool members and traffic will continue to be dispersed to the remaining healthy nodes. When content is written via the S3 protocol to any node in the pool, the other members are synchronized to serve up content should they be selected by BIG-IP for future read requests. BIG-IP can also be deployed as a highly available solution.

Key features

Global server load balancing manages traffic across geographically distributed storage nodes, directing clients to the nearest site.

BIG-IP scales extreme, high-performance AI data delivery with FPGA assist, available for demanding NetApp StorageGRID deployments.

Advanced WAF, flexible policies, and response inspection enable data security and ensure delivery to AI systems.

Load balancing supports post-quantum-computing-resistant cryptography, per the NIST FIPS 203 standard, protecting data against future threats enabled by traffic harvesting done today.

Inspection of both headers and payload can enable priority routing based on metadata as well as real-time security alerting.

Built-in security for S3 traffic

BIG-IP's core ADC (application delivery controller) security features are fully applied to S3 high-performance traffic loads with minimal impact to performance.

- **Advanced WAF**
  Application layer security modules, including Data Guard, protect against behavioral threats, automated attacks, and sensitive data disclosure including a redaction feature for sensitive information such as PII.
- **Highly customizable policies**
  iRules provide powerful, flexible, and real-time customization of network and application traffic, allowing administrators to implement custom business logic, security policies, and application-specific features beyond the standard BIG-IP interface.
- **F5 AI Assistant**
  Engage through a natural language generative AI interface for BIG-IP iRules to accelerate the creation of new iRules and simplify management, making traffic management and security more accessible and efficient than ever before.
- **Modern, real-time security alerting**
  Use HTTP or syslog to feed security alerts to SIEM or observability platforms.

Tune BIG-IP for high performance with profiles

F5 BIG-IP Local Traffic Manager™ (LTM) offers different profiles to tune traffic management behavior. The FastL4 profile can substantially increase virtual server performance by exclusive focus on layer 4, TCP characteristics. Supported platforms using the embedded Packet Velocity Acceleration (ePVA) chip can further accelerate traffic. The ePVA chip is a hardware-acceleration field-programmable gate array (FPGA) that delivers high-performance layer 4 throughput by offloading traffic processing to the hardware acceleration chip. BIG-IP makes flow acceleration decisions in software and then offloads eligible flows to the ePVA chip for that acceleration. For platforms that do not contain the ePVA chip, the system performs acceleration actions in software.

Additional BIG-IP profiles include OneConnect for layer 7 HTTP-based load balancing. This profile minimizes the number of server-side TCP connections required for a NetApp StorageGRID cluster, reducing overall memory consumption on the storage nodes.

Global server load balancing for highly resilient storage systems

F5 BIG-IP DNS enables global server load balancing for geographically distributed NetApp StorageGRID sites for a robust, resilient, highly available global service. BIG-IP can inspect transactions requesting S3 data and route requests to the closest location for the best performance. In addition, you can keep a data center on standby to utilize for disaster recovery. In the event of a failure, BIG-IP can route traffic to the secondary site to avoid disruption or downtime.

Improved visibility into S3 traffic for priority routing and insights

Route priority traffic

S3 metadata in traffic headers may be configured to indicate if its data is high QoS, i.e. designated as critical and therefore needs to be prioritized for performance. BIG-IP can act on this information using its data plane programmability, powered by iRules, and route the traffic to a higher-performing storage pool such as a set of full SSD nodes. Subsequently, it can route lower-priority traffic to less-performant hybrid or spinning hard disk drives.

S3 traffic insights

The BIG-IP Application Visibility and Reporting (AVR) module for onboard analytics provides details on the nuances of the S3 traffic being proxied. With AVR, you can see the URL values used by S3 transactions incorporating the object names as URLs.

Share traffic with third-party NetOps tooling

Richly detailed, per-transaction visibility is achievable with F5 BIG-IP SSL Orchestrator® where copies of bi-directional S3 traffic decrypted within the load balancer can be sent to packet loggers, analytics tools, or protocol analyzers. Active in-line Intrusion Protection System (IPS) can be plumbed into S3 traffic flows to add further capabilities to modern security postures.

AI data delivery at scale

Utilizing BIG-IP with NetApp StorageGRID enables AI data delivery at scale, with intelligent local and global load balancing, advanced WAF, and granular control over policies to address complex requirements.

To learn more about F5 and NetApp solutions, visit f5.com/netapp.

©2025 F5, Inc. All rights reserved. F5, and the F5 logo are trademarks of F5, Inc. in the U.S. and in certain other countries. Other F5 trademarks are identified at f5.com.
Any other products, services, or company names referenced herein may be trademarks of their respective owners with no endorsement or affiliation, expressed or implied, claimed by F5, Inc.

DC 09.2025 | JOB-CODE-1751501505