Solution Overview

Fast and Scalable AI Data Ingest with F5 and NetApp

Maximize your GPU investment and keep traffic moving at the speed of AI with F5 BIG-IP and NetApp StorageGRID.

NetApp

Key Benefits

Improved GPU utilization
Intelligently route traffic at very high speeds and avoid blockages to help maximize GPU utilization.

Protected traffic speeds
Eliminate network bottlenecks to ensure traffic flows rapidly in and out of your AI factory or private data center.

Reduced risk
Thwart threats hidden in encrypted traffic before they have a chance to impact AI operations.

With AI taking the business world by storm, modernizing and optimizing the underlying infrastructure for AI has become a priority.

Amazon Simple Storage Service (Amazon S3) API-compatible storage has quickly become a standard for modern AI applications and model training because of its superior performance over traditional protocols such as NFS or CIFS. Most of the modern applications are also relying on Amazon S3 API-compatible storage. NetApp StorageGRID offers native support for Amazon S3 API, delivering industry-leading storage that supports the massive data scale needed for AI model training, ingestion, and other use cases.

Meanwhile, most GPUs go underutilized even during peak AI model training times.[1] This is a troubling fact given the high expense of these advanced computing resources. Ensuring intelligent traffic management and load balancing in front of your AI services can help maximize the return on your AI infrastructure investments.

F5 BIG-IP for High Duty Cycle GPU

The robust chipsets required to train and fine tune large language (LLM) models and run AI software operate at an incredibly high bandwidth, up to hundreds of gigabits per second.

F5® BIG-IP® app delivery and security service family is the best option for hyperscale load balancing to support your busiest, most complex AI workloads. It can handle the needed speed, concurrency, and volume while removing network bottlenecks to help keep expensive GPUs utilized.

[1] AI Infrastructure Alliance, "The State of AI Infrastructure at Scale 2024," March 2024

Key Features

Hyperscale load balancing
Intelligent load balancing and built-in security coordinate ideal HTTP server utilization.

Accelerators
BIG-IP offers numerous features, including pre-configured profiles, to fine-tune traffic management and optimize infrastructure.

Proprietary hardware
Application delivery controller hardware tuned for BIG-IP enables extreme performance and scale for high-bandwidth conditions.

Encrypted traffic inspection
Threats hidden in encrypted traffic are detected and dealt with at speed, protecting AI asset productivity.

High availability
BIG-IP configured with high availability delivers resilience and uptime needed for AI factory operations.

Load Balancing with BIG-IP

NetApp StorageGRID allows a large-scale set of Amazon S3 API targets using HTTPS to ingest and provide objects. BIG-IP can provide load balancing services for StorageGRID with full security applied in-flight and the needed extreme performance levels to meet enterprise-capacity requirements. Automatic backend synchronization allows any node to be offered up as a target by BIG-IP, allowing storage utilization to be optimized across the node set and performance scaled to reach the highest Amazon S3 API bandwidth levels—all while offering high availability to Amazon S3 API consumers.

BIG-IP maintains the integrity of the NetApp nodes with frequent HTTP-based health checks. If an unhealthy node is detected, it will be dropped from the list of active pool members. When content is written via the Amazon 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.

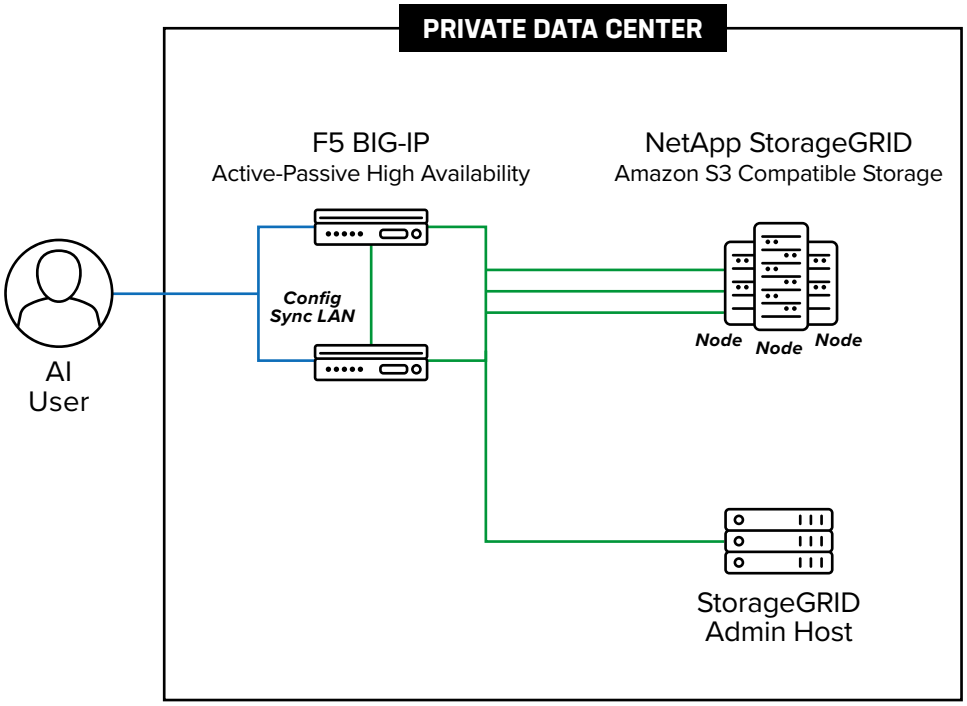

Figure 1: BIG-IP load balancing with NetApp StorageGRID Amazon S3-compatible storage for AI applications running in AI factories and private data centers

Tune for High Performance Amazon S3-Compatible Storage Service Delivery

BIG-IP offers profiles it can run via F5 BIG-IP Local Traffic Manager™ (LTM), which is the heart of the server load balancing function. Different profiles can be selected to reach the right tradeoff between the absolute capacity of stateful load-balanced traffic versus rich layer 7 functions, including F5 iRules™ or authentication.

FastL4 Profile with Hardware Acceleration

The FastL4 protocol profile can increase virtual server performance and throughput for supported platforms by leveraging the embedded Packet Velocity Acceleration (ePVA) chip to accelerate traffic. FastL4 delivers high-performance layer 4 throughput by offloading traffic processing to the hardware acceleration chip.

- IP Intelligence hardware
  The ePVA is also used to process and implement IP Intelligence (IPI) rules to block malicious actors. if denial-of-service (DoS) sweep protection detects a bad actor or group of actors, it can set an auto-deny list. It can also signal the ePVA to drop the offending IP addresses in hardware on some BIG-IP platforms so they are not sent on for further processing.

- DoS protection hardware
  Many attack vectors such as bad headers, floods, and fragmented packets are processed in hardware and mitigated using the ePVA chip. Addressing these attacks with hardware rather than software improves performance of the BIG-IP solution.

OneConnect Profile

F5 OneConnect™ works with HTTP Keep-Alive, which allows BIG-IP to minimize the number of server-side TCP connections by making them persistent. This reduces excessive TCP three-way handshakes and mitigates unnecessary congestion from new TCP session starts.

Fast HTTP Profile

The Fast HTTP profile is designed to speed up certain types of HTTP connections and strives to reduce the number of connections opened to back-end HTTP servers. This is accomplished by combining features from the TCP, HTTP, and OneConnect profiles into a single profile optimized for network performance.

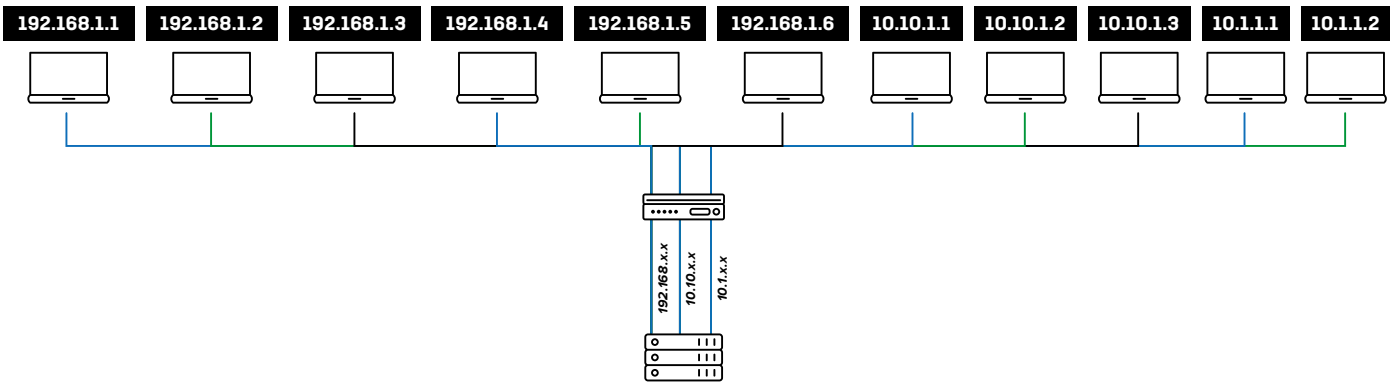

Figure 2: OneConnect connection reuse

Protect AI Assets from Encrypted Threats

Bad actors take advantage of SSL/TLS encryption to hide malicious payloads to outsmart and bypass security controls. F5 BIG-IP SSL Orchestrator® can send copies of bidirectional Amazon S3 traffic, decrypted within the load balancer, to packet loggers, analytics tools, and protocol analyzers. This rich transaction detail can be used to enable security inspection that exposes threats and stops attacks before they happen.

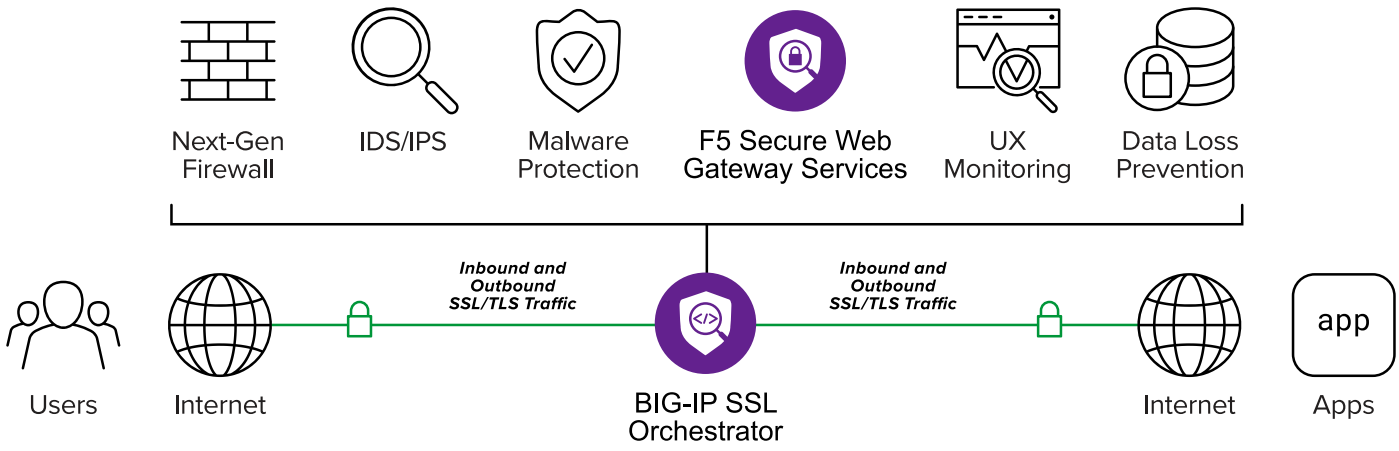

Figure 3: BIG-IP SSL Orchestrator overview

Read the DevCentral article for more information about F5 BIG-IP solution configuration for NetApp StorageGRID.

Learn more at f5.com/products/big-ip-services/local-traffic-manager.