# AI-Based DDoS Attacks

## Judgment day 2024

Ondrej Ciz
F5, Solution Engineer III

# Agenda

AI Evolution And Adoption

AI-Based DDoS Attack Architecture

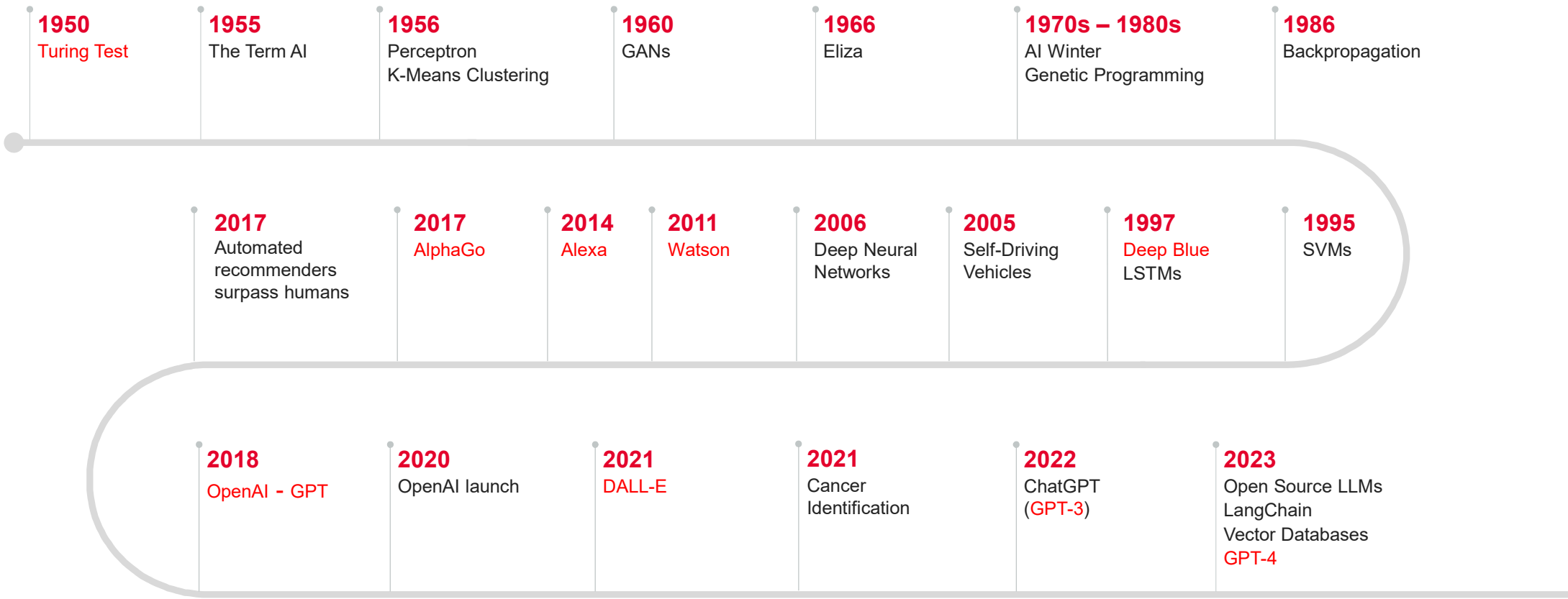
AI-Based DDoS Attack Mitigation

Q&A

# AI is not new

## History of AI



# AI adoption set to define new business models, transform industries, global economies

**80%**

of tech executives will increase investment in AI in the next year[1]

AI will likely contribute an estimated

**$15.7 Trillion**

to the global economy by 2030[3]

The Europe Artificial Intelligence Market Size is Expected to reach

**USD 325.29 Billion**

by 2032[5]

AI will have an estimated

**21%**

net increase on United States GDP in 2030[2]

Asia's Top 1000 Companies to Allocate

**Over 50%**

of their IT Spending on AI Initiatives by 2025 according to IDC[4]

# Top AI Challenges

AI is a business imperative but requires governance programs

Lack of visibility into AI use (Shadow AI)

Complexity of managing increasingly distributed AI workloads

Defending against AI attacks on apps

Safely connecting AI models with your data that lives in disparate environments

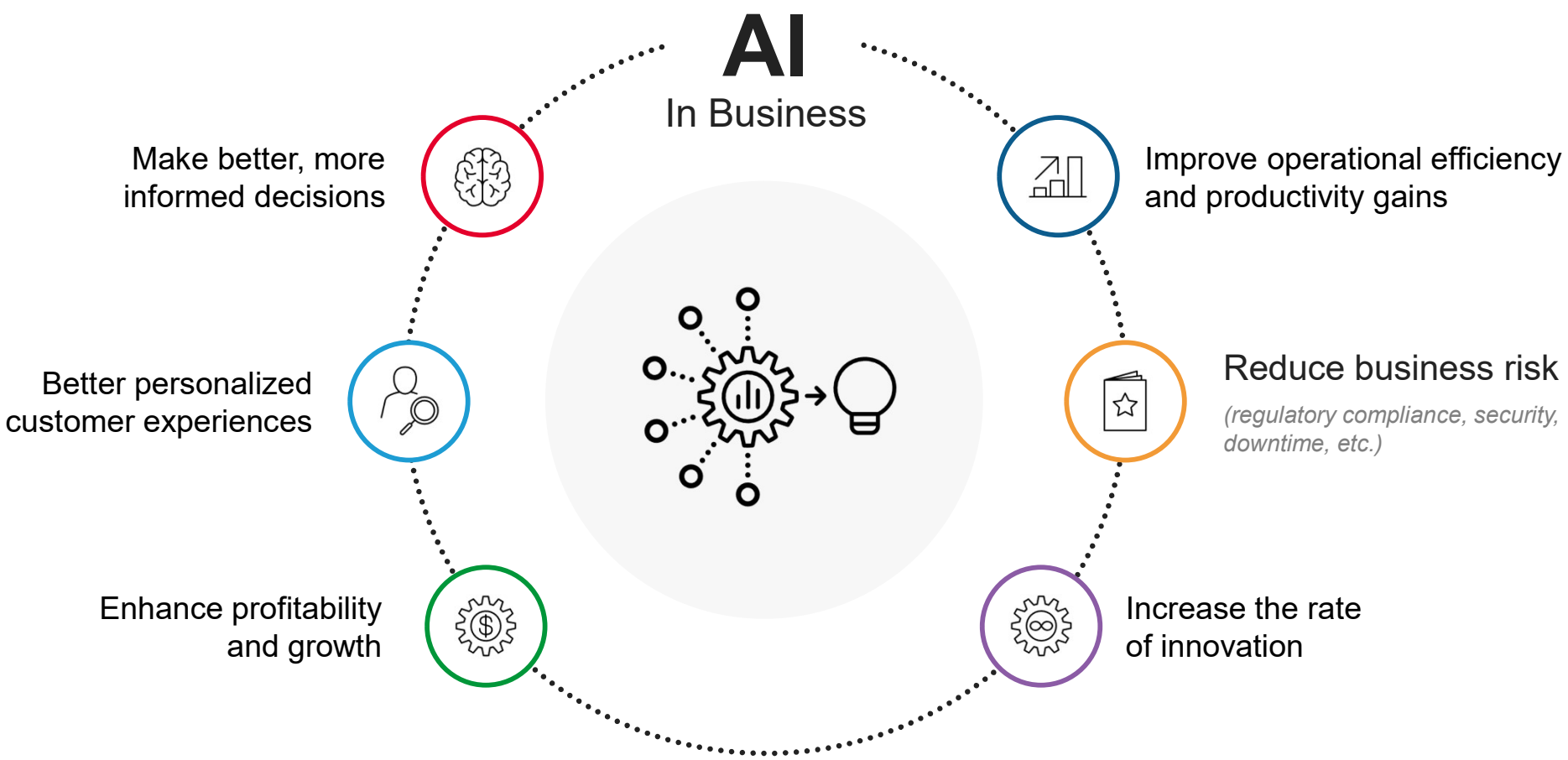
Protecting the explosion new API endpoints required to connect AI/ML models

SecOps' lack of visibility or governance of new API endpoints and middleware dependencies

Managing AI workload complexity without impacting business agility and innovation

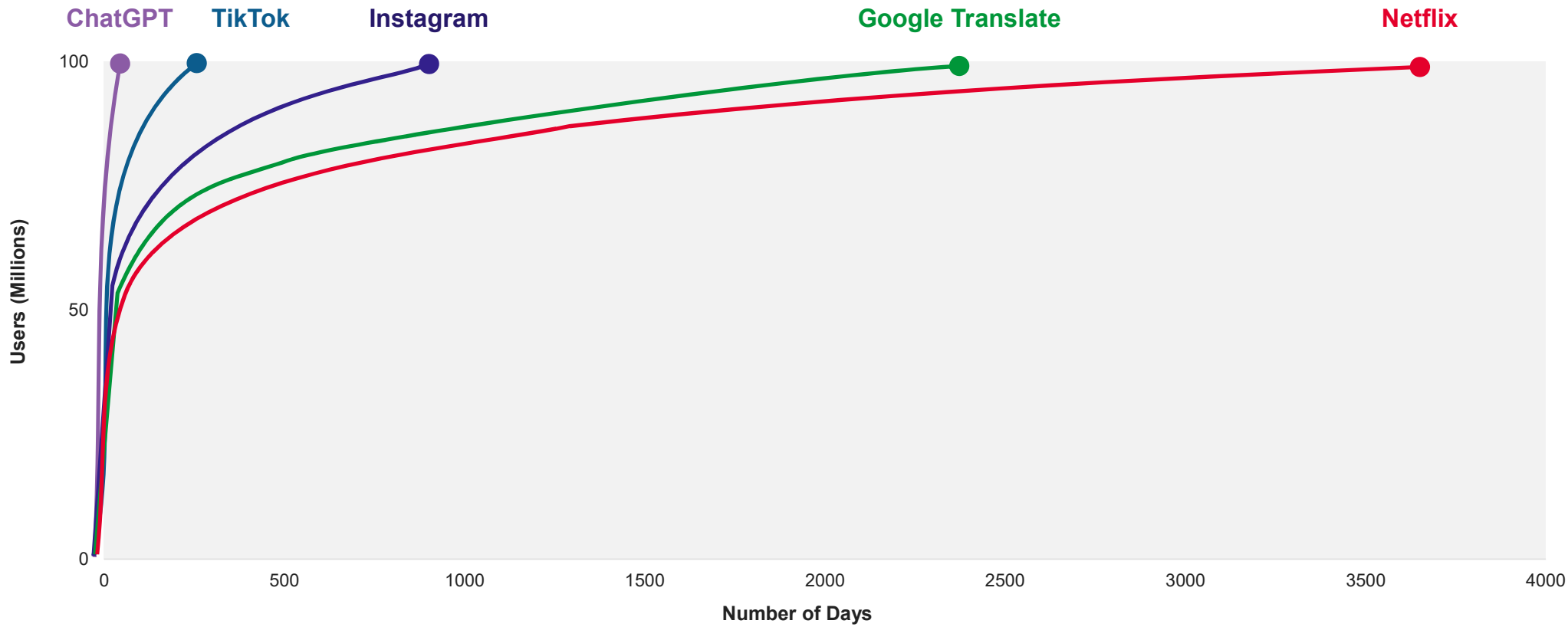
# Top business drivers and considerations for AI adoption



# The Pace of Adoption of Generative AI Has Been Astounding

Time it took companies to reach 100 million users:




Sources: Global X ETFs with info derived from: BBC News. (2018, Jan 23). Netflix's history: From DVD rentals to streaming success; Cerullo, M. (2023, Feb 1). ChatGPT user base is growing faster than TikTok. CBS News.

# DDoS attacks are on the rise significantly YoY 2022/2023

**DDoS impacts include**:

- Loss of revenue
- Loss of productivity (IT Ops / Security)
- Hiring specialized consultants
- Credits to consumers
- Legal and compliance fees
- Public relations

**Indirect costs often include**:

- Damage to the brand
- Theft of vital data
- Loss of valuable customers and
- Opportunity loss

55 seconds between attacks in 2022
14 seconds between attacks in 2023

*https://www.link11.com/en/download/european-cyber-report-2023/

# For Cyber Security AI Is a Double-Edged Sword

## Creates new attack surfaces and new attack capabilities

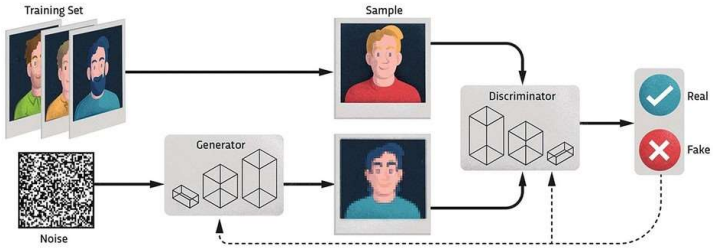
Moving at alarming speed, no reservations

## Enhances cyber security

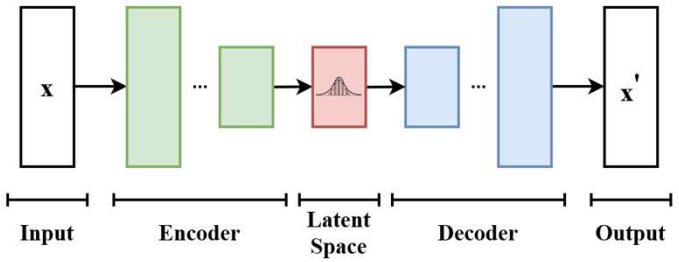
Exercising caution with deferment to human expertise

# AI-Based DDoS Attack Architectures

- **Machine Learning-Enhanced Attackers: Training Models**
  - Understand and mimic normal network behavior, difficult to detect malicious activity.
- **Adaptive Attack Strategies:**
  - ML algorithms can adapt in real-time, allowing attackers to modify their strategies based on the effectiveness of the ongoing attack and the defensive measures in place.
- **Neural Networks (GAN, VAE):Traffic Generation:**
  - GANs can be employed to generate realistic-looking traffic that can mimic legitimate user behavior, making it challenging for traditional rule-based defenses to distinguish between malicious and legitimate traffic.




**Generative Adversarial Networks**



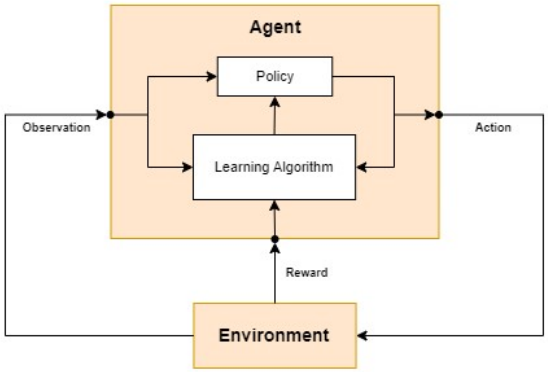

**Variational autoencoder**

*https://www.linkedin.com/pulse/future-ai-based-ddos-attacks-proactive-defense-pratik-barahatte-dahuf

# AI-Based DDoS Attack Architectures

- **Reinforcement Learning (RL) to optimize Attack Strategies**
  - Attackers might use RL algorithms to optimize the parameters of their attack in real-time, making it more difficult for defenders to predict and mitigate the attack.

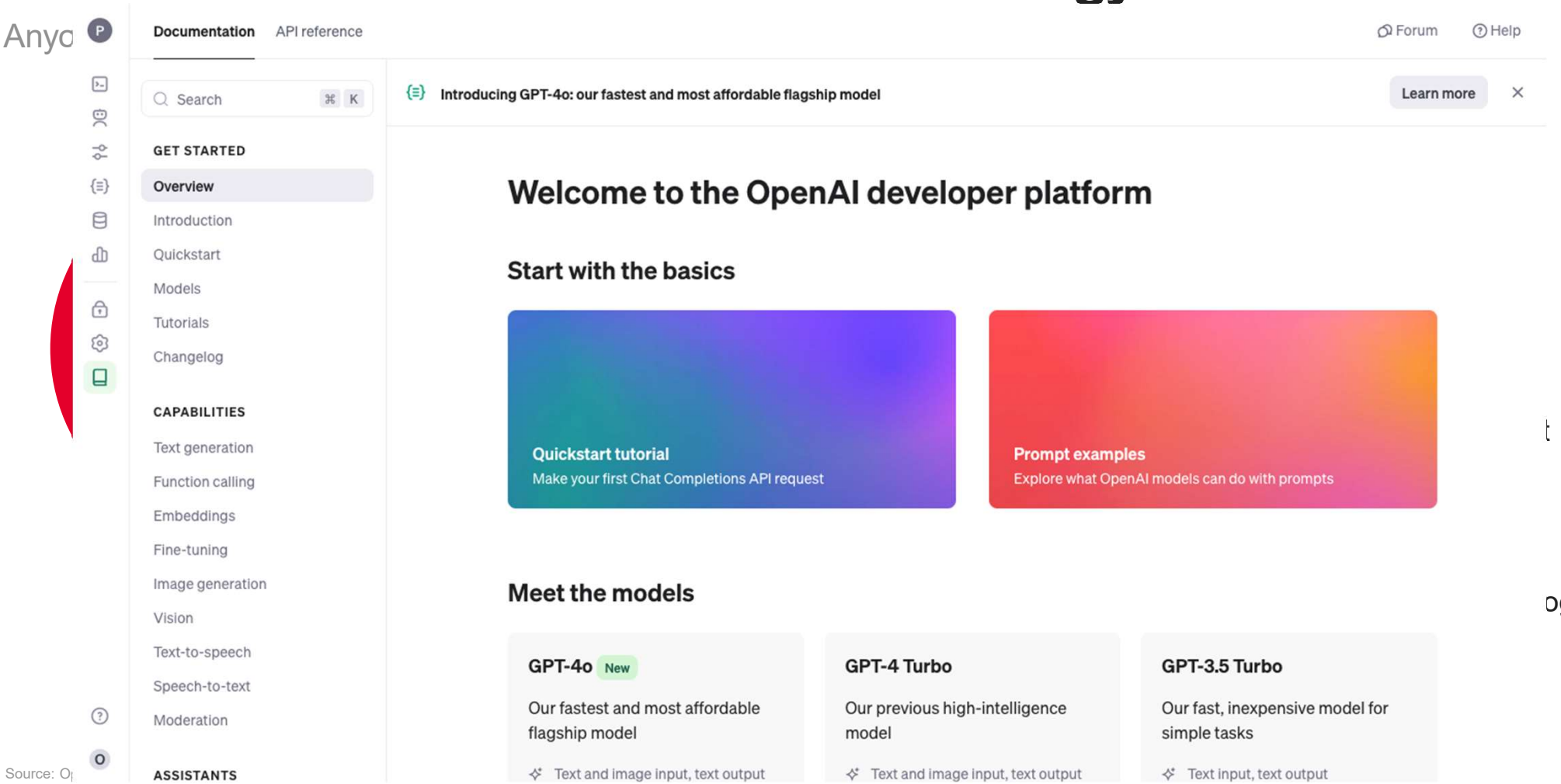



**Reinforcement Learning**

- **IoT botnets**
  - IoT devices are often insufficiently protected and are therefore easily compromised. Compromised devices can be used to expand powerful botnet networks for DDoS attacks (e.g. Mirai botnet attacks of 2016). AI algorithms can be used to coordinate these botnets more efficiently and generate more sophisticated attack patterns.

*https://www.linkedin.com/pulse/future-ai-based-ddos-attacks-proactive-defense-pratik-barahatte-dahuf

# Generative AI Has Democratized Technology



# DDoS attack mitigation

# F5 has been an AI pioneer for decades – and continues to evolve to meet market needs

**2006**
**Policy Builder** to automatically create WAF security policy

**2011**
**Machine Learning (ML)** to optimize BIG-IP network stack

**2012**
**ML-based learning mode** to improve WAF efficacy

**2014**
**Behavioral anomaly analysis** to improve WAF efficacy

**2015**
**ML in iHealth** to identity customer BIG-IP issues and misconfigurations

**2016**
**Behavioral DoS** to mitigate attacks using dynamic signatures

**2019**
**Stage 2 retrospective analysis** to deter bots and automated attacks

**2020**
**Shape acquisition** AI powered Anti- Bot & Fraud

**2021**
**Dynamic API Discovery** to mitigate risk from rogue APIs

**2021**
**False Positive Suppression** to reduce burden on SecOps teams

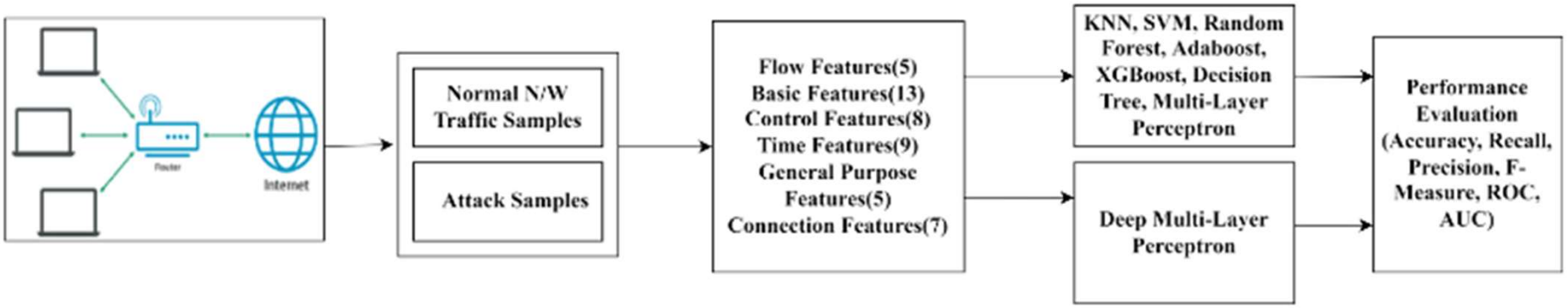
**2022**
**Malicious User Detection** to automatically mitigate bad actors

**2023**
**API Endpoints Risk Score** to measure risk of discovered APIs

# Mitigation Strategies

Detection of Network Attacks using Machine Learning and Deep Learning Models

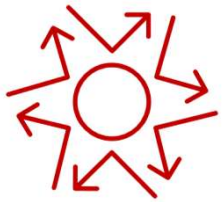



*https://pdf.sciencedirectassets.com

# Mitigation Strategies

- Traffic Analysis and Anomaly Detection
- Behavioral Analysis
- Rate Limiting and Traffic Filtering
- Web Application Firewalls (WAF)
- Intrusion Detection and Prevention Systems (IDPS)
- Challenge-Response Mechanisms
- Cloud-Based DDoS Protection Services
- Regularly Update and Patch Systems
- Collaborate and Share Threat Intelligence
- Incident Response Planning



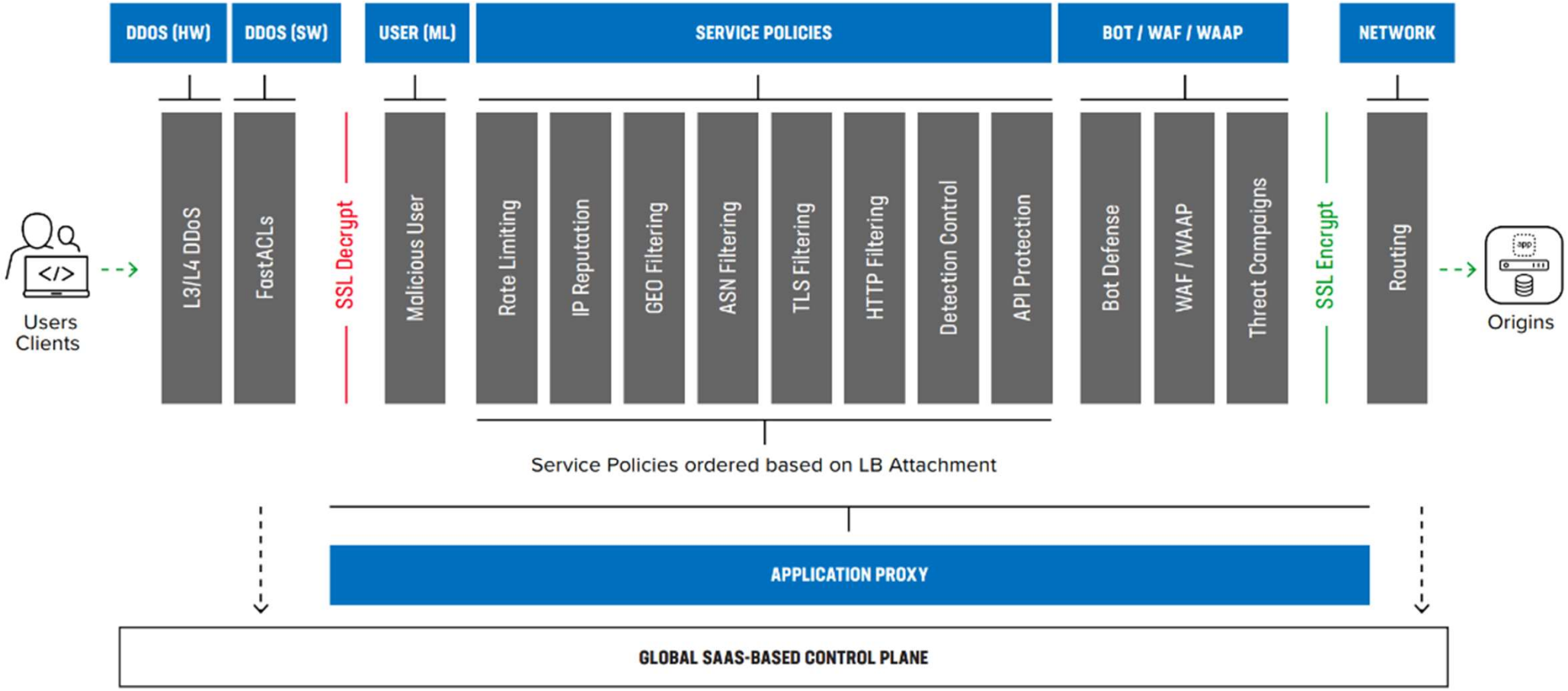
Volumetric DDoS Attacks

Protocol DDoS Attacks


Application DDoS attacks

# The FLOW of Security [Logical Flow]



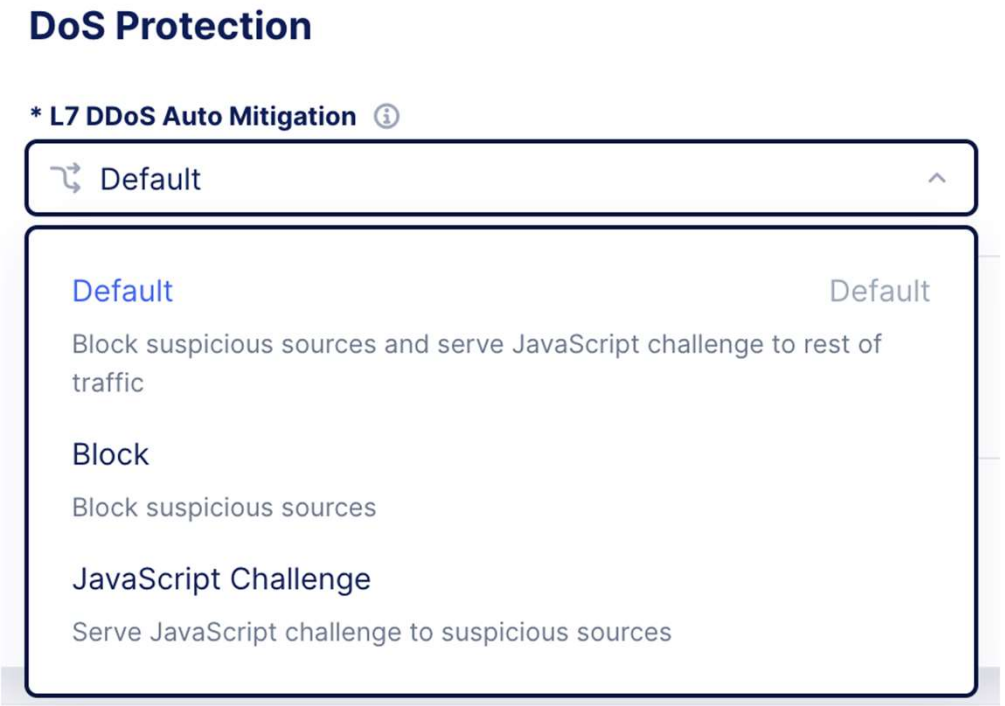

 

Note: All aspects of flow not represented

# Auto DoS Protection

**Enabled by default since Jan 2024 release**

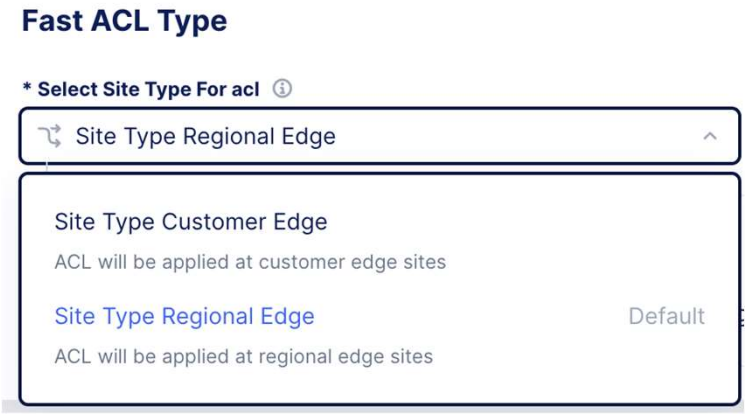
- Included as part of the base package.
- Detects anomalies for request rate, error rate, latency and throughput.
- Dynamically programming
  - Fast ACL to drop attacker at L3 layer
  - L7 ACL in envoy to drop connection based on fingerprint, geo or path
- Auto mitigation rules are created for a duration of 20 minutes, then auto deleted.
- Ability to create trusted client rules to bypass IP addresses from the mitigation.

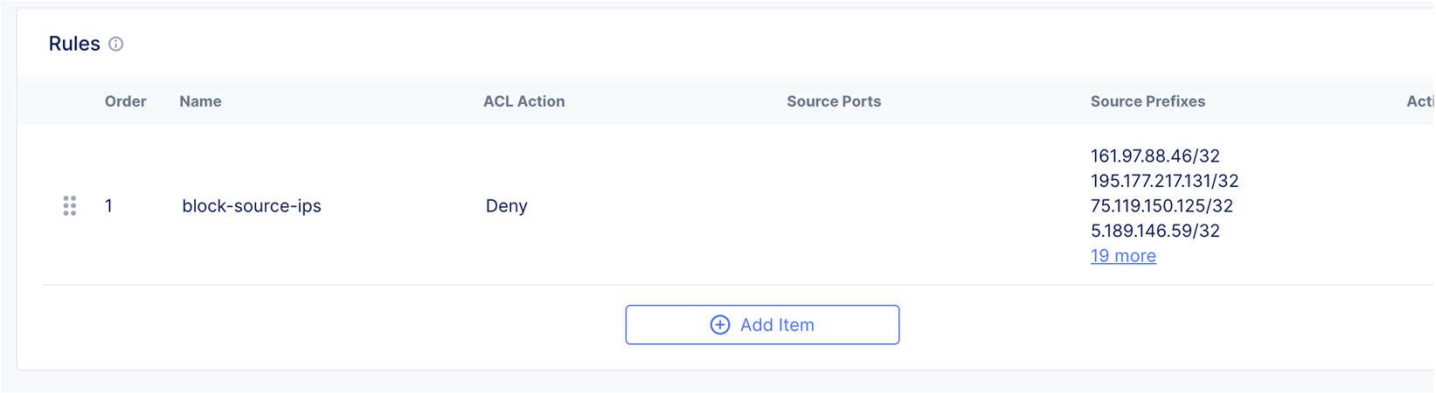


# Fast ACL rules

- Specified in terms of five tuple of the packet (dst ip, dst port, src ip, src port, protocol).
- Very efficient, but hard to do manually
- Block source IP attacker on L3
- Auto DoS Mitigation create them automatically

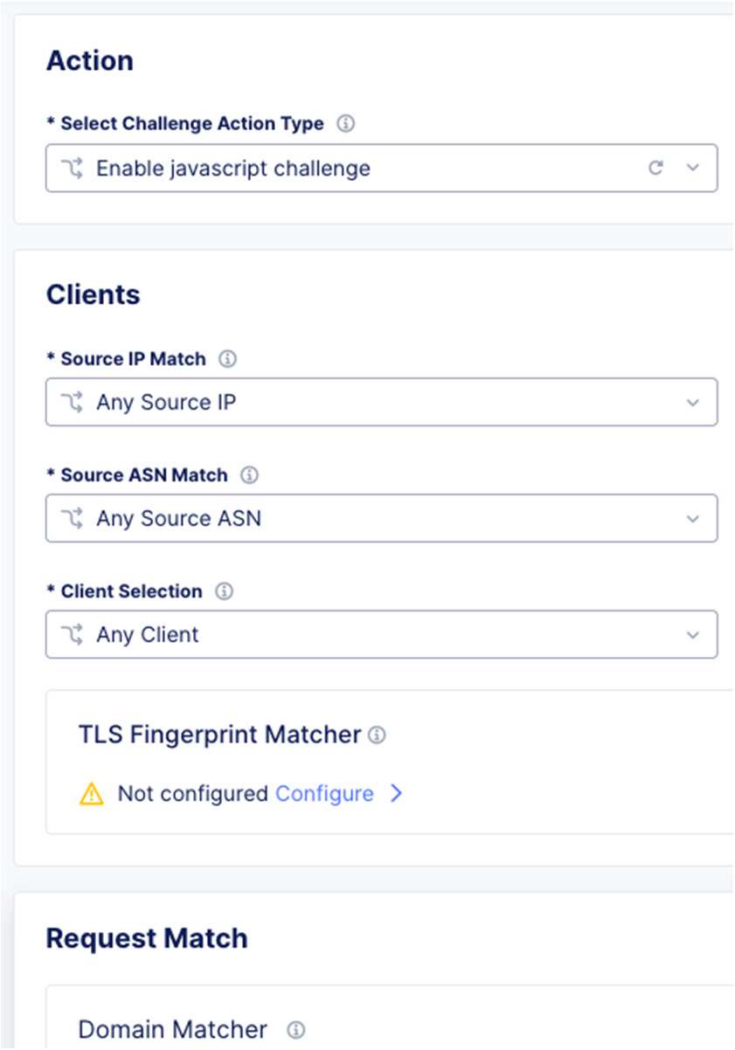


# JavaScript Challenge

- Validate that the request is coming from a browser that is capable of running JavaScript
- Slows down a potential DoS attacks by forcing the browser to run a complex operation that requires many CPU cycles
- The load balancer tags response header with a cookie to avoid JavaScript challenge for subsequent requests.
- Can be configured via Policy Based Challenge rules for advance traffic matching criteria

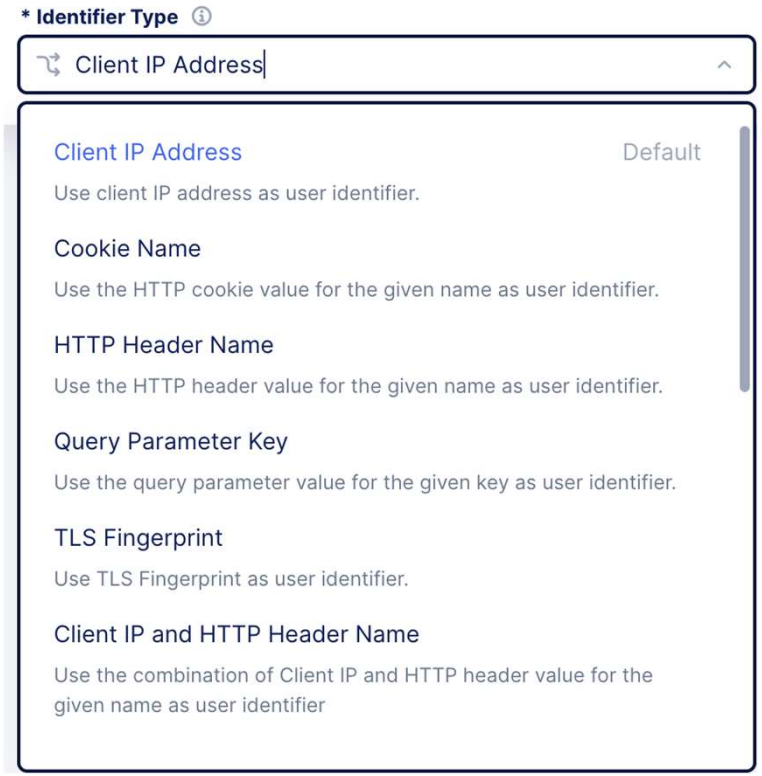
0

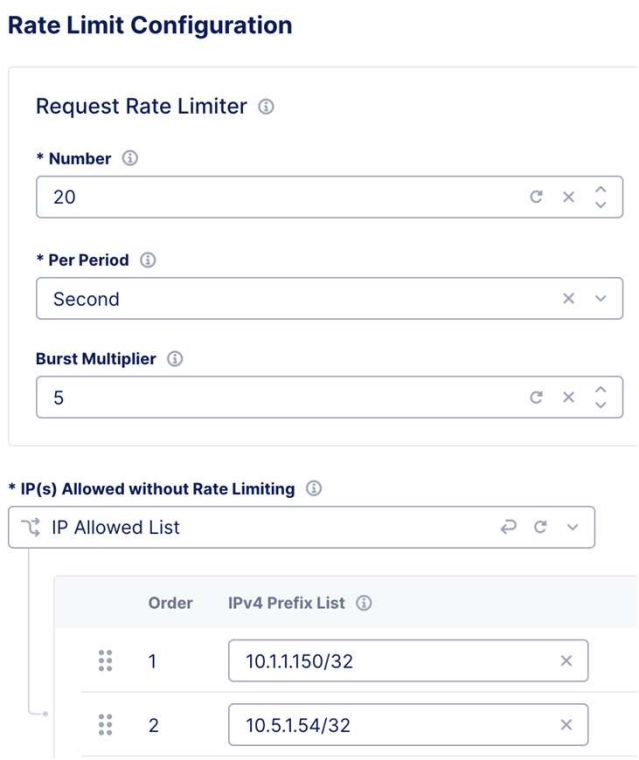
Be extremely careful with the JS challenge in XC.  We turned it on and it caused a considerable amount of downtime for some customers.  If Bot Protection is employed, do not use the JS challenge in XC.  Also, never turn this on without conferring with the customer first.  Just a few guidelines.

Alex Barajas; 2024-02-05T20:04:45.132

# Rate Limiting

- Apply on IP Addresses by default, but can be configured for other identification criteria
- Optionally exempt rate limiting for requests from specified IP prefixes
- Trigger 429 codes in the logs
- The rate limit is always evaluated before any configured network security policy sets.



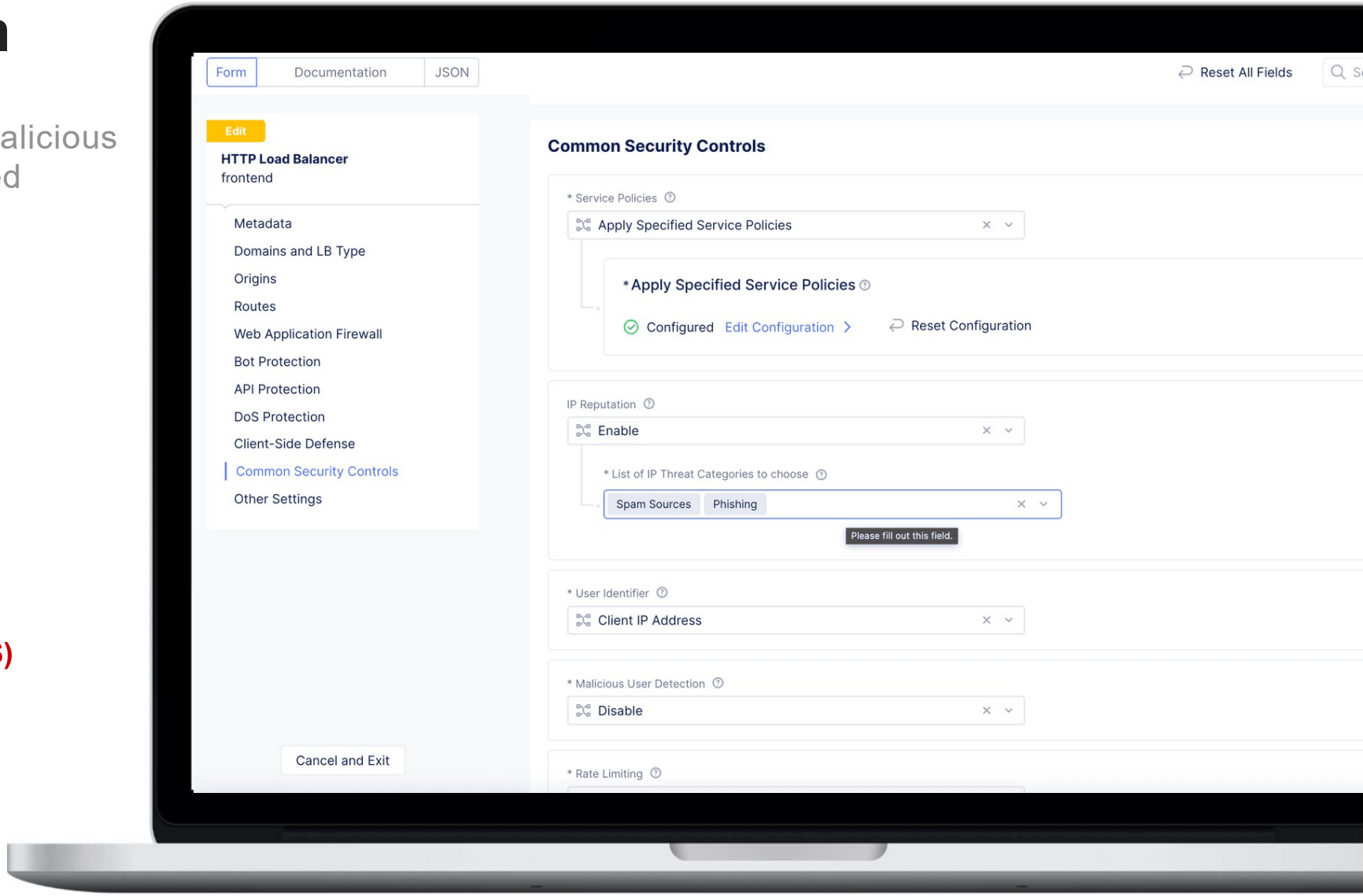



# IP Reputation

Database of known malicious IP addresses classified by threat categories

- Spam sources
- Mobile threats
- Windows exploits
- Web attacks
- **Botnets**
- Scanners
- **Denial of Service (DoS)**
- Phishing



# F5 Distributed Cloud DDoS Mitigation Network

**Responds to DDoS attacks** in < 2 minutes on average.*
Top BGP peering

Modern Service Provider Global Backbone

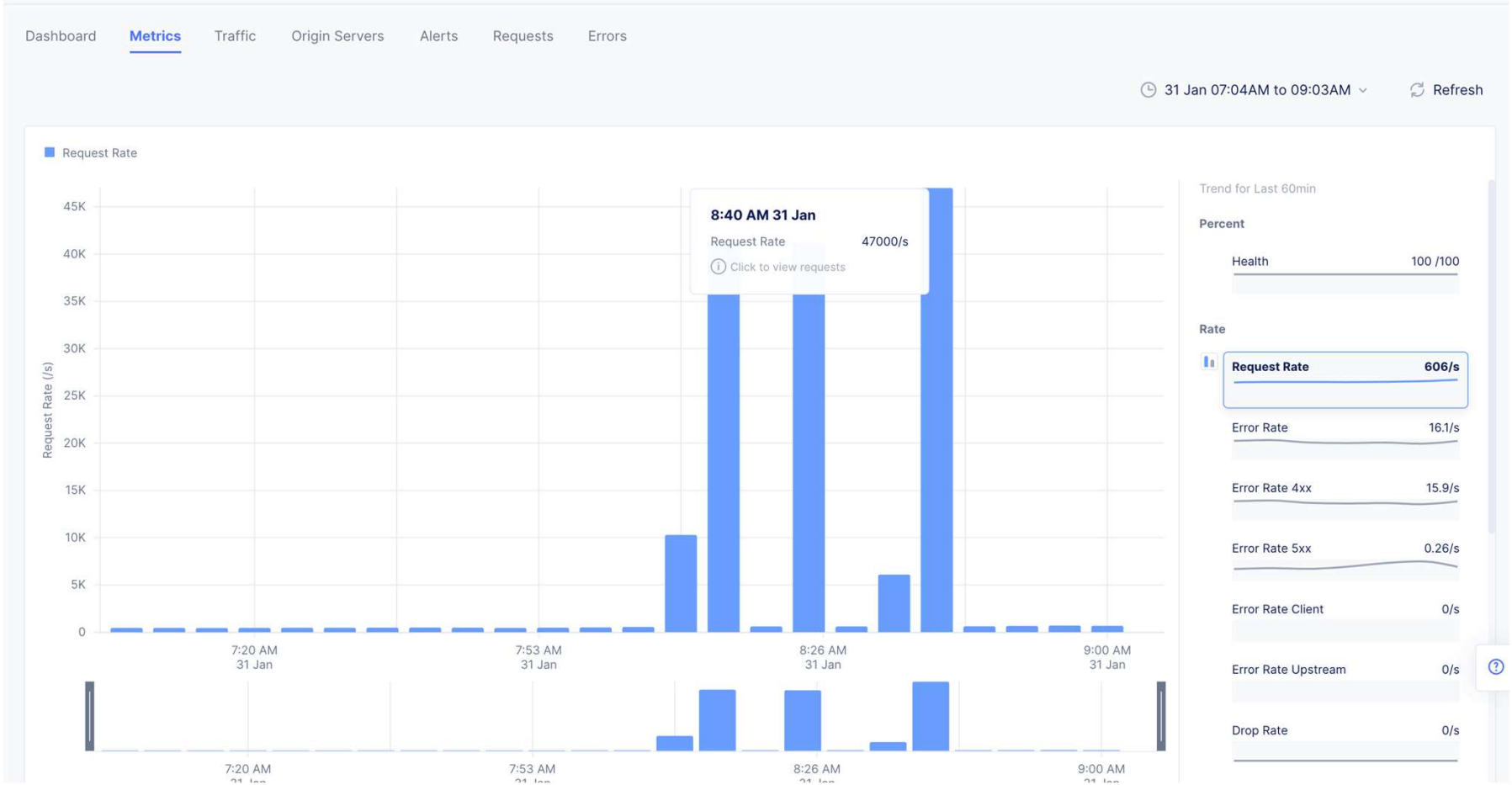
**Flexible Service Options** including Always Available or Always On deployments

**Connect how and where you need** with BGP or Proxy-based traffic redirection and direct connections, peering or GRE tunnels for clean traffic return.


*Standard DDoS Service offering MSA specifies a 15 Minute Response SLA.*

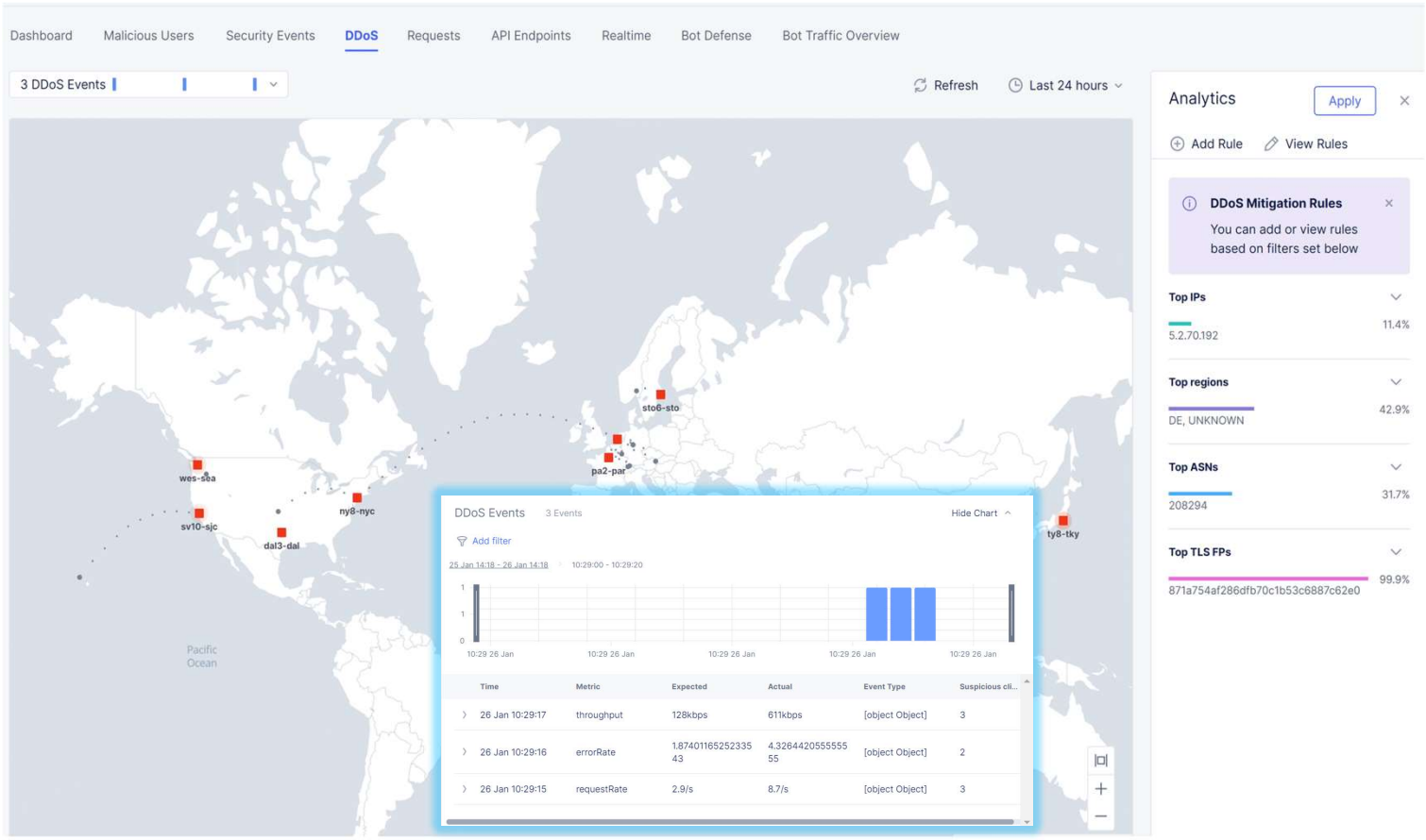
# Metrics request rate & spikes



# DDoS Events map

- Shows details on DDoS events occurring over selected time interval.
- Display DDoS trend including start time, end time, and number of events represented.



# Why F5 for API Security?

## Discovery + Enforcement

Holistic app and API security, with AI/ML powered continuous discovery to uncover unknown, forgotten and shadow APIs - paired with critical protection capabilities (WAF, DoS mitigation, Bot protection etc.)

## Flexibility and Architectural freedom

Combination of hardware, software and SaaS with flexible, lightweight, agnostic tools customers can deploy across cloud, on-premises, and edge environments.


## Goes Beyond Detection and Blocking

A portfolio that can solve the problem more holistically giving Platform Ops and IT teams the tools to enable developers with self-service capabilities to discover, use, publish and manage APIs and security teams consistent oversight and fine-grained controls over all apps and API endpoints.

# Q&A


App
World
2024

# EXTRA SLIDES

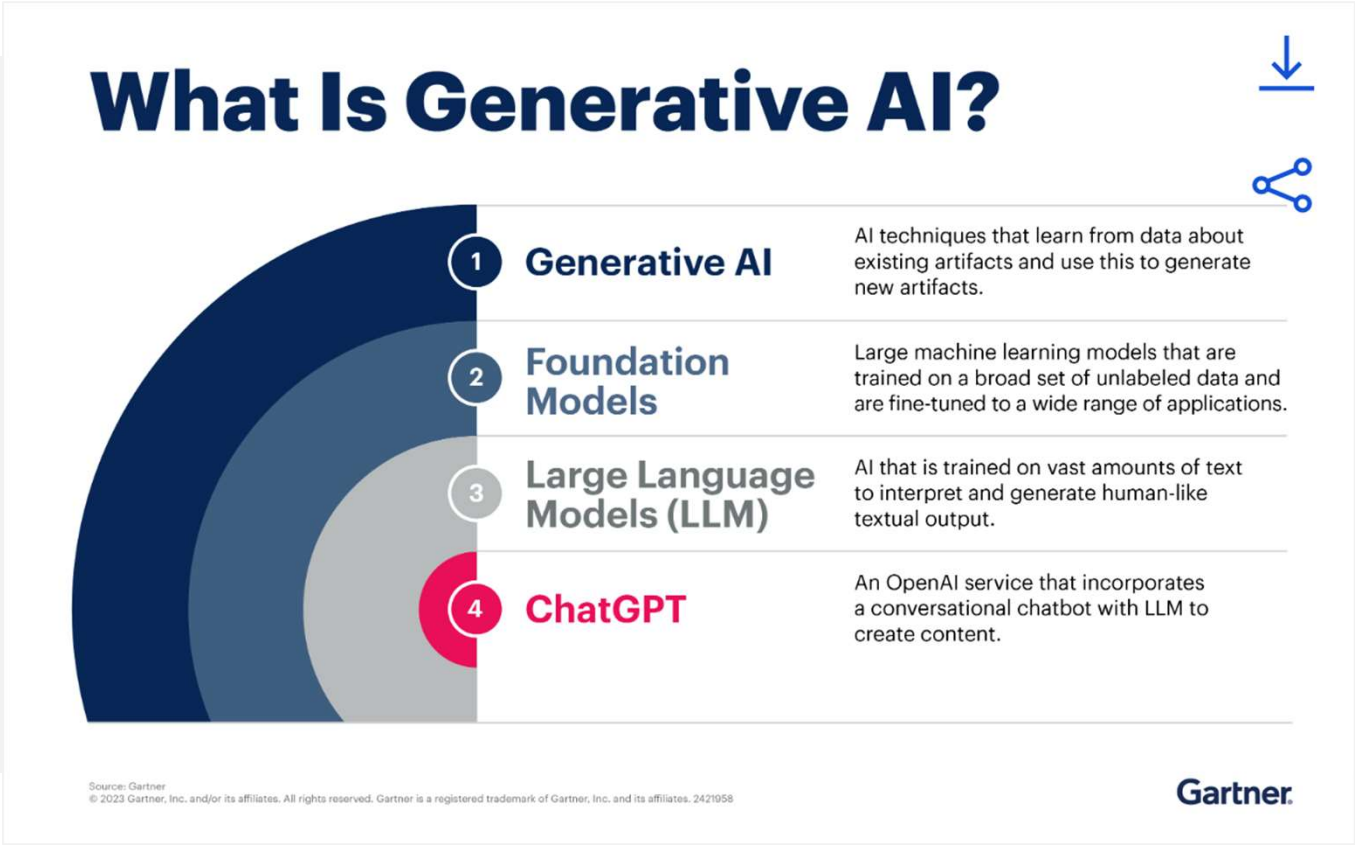
 

# Generative AI: The Democratization of AI

From tool for a few, to resource for many

# Generative AI takes the main stage



# GenAI APIs/GenAI-enabled apps in use already with widely accelerated growth expected over the next few years


More than half of tech executives whose companies are experimenting with generative AI (56%) are doing so for economic savings[1]


Nearly all business leaders say their company is prioritizing at least one initiative related to AI systems in the near term, while 54% of companies have already implemented GenAI in some areas of their business[2]

Gartner

**Press Release**
**STAMFORD, CT, October 11, 2023**

**By 2026, more than 80% of enterprises** will have used generative artificial intelligence (GenAI) application programming interfaces (APIs) or models, and/or deployed GenAI-enabled applications in production environments, **up from less than 5% in 2023.**[3]

# Generative AI Has Engendered both Excitement and Caution

"AI will be the biggest thing this decade…it has potential to teach math and offer medical advice to people with limited access to resources."

**Bill Gates,** Cofounder of Microsoft

"Humans must be unambiguously, unquestionably in charge of powerful AI models to prevent them from going out of control."

**Satya Nadella,** CEO of Microsoft

"We see enormous potential in this space to affect virtually everything we do... It will affect every product and every service that we have."

**Tim Cook,** CEO at Apple

"AI can have huge problems, in terms of a democracy and how it reacts to this… biggest value is in replacing human labor."

**Warren Buffet,** Business magnate, investor, and philanthropist

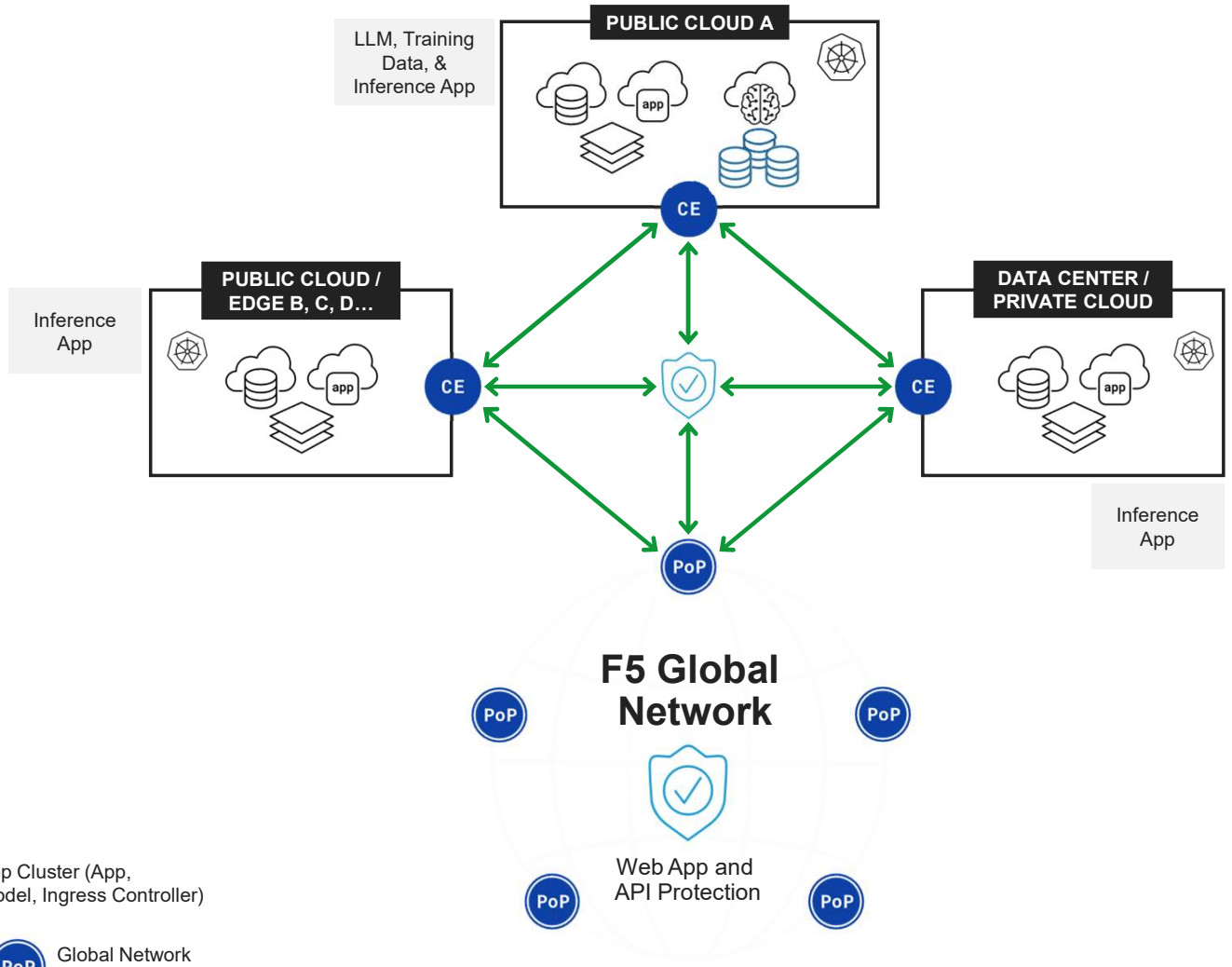
# Generative AI Multicloud Networking Challenges

Connecting Customer Edge nodes directly to each other for internal apps with secure service connectivity, or via the Regional Edge for public-facing apps that need a front door service.

Connecting and sharing data between inference apps, the central LLM, and training data through secure network connectivity.

Enablng inference app deployment and maintenance at scale with an ingress controller in each app cluster.

Enabling load balancing and security at each node, with centralized DDoS, Bot, and API protection, plus distributed web app firewalls locally.




**Key**

LLM

Training Data

Inference App Cluster (App, Inference Model, Ingress Controller)

Web App & API Protection

Secure Network & Service Connectivity

CE Customer Edge

PoP Global Network Regional Edge

# Challenges with Running AI/ML Workloads in Kubernetes

Across hybrid, multi-cloud environments with disaggregated technologies

- Connection timeouts and errors → Poor user experiences
- Insufficient visibility into app health and performance → Troubleshooting difficulties and downtime
- Difficulties with securing distributed app environments → Increased risk of exposure to cybersecurity threats
- Increasing complexity and tool sprawl → Hard to operate, manage, and troubleshoot

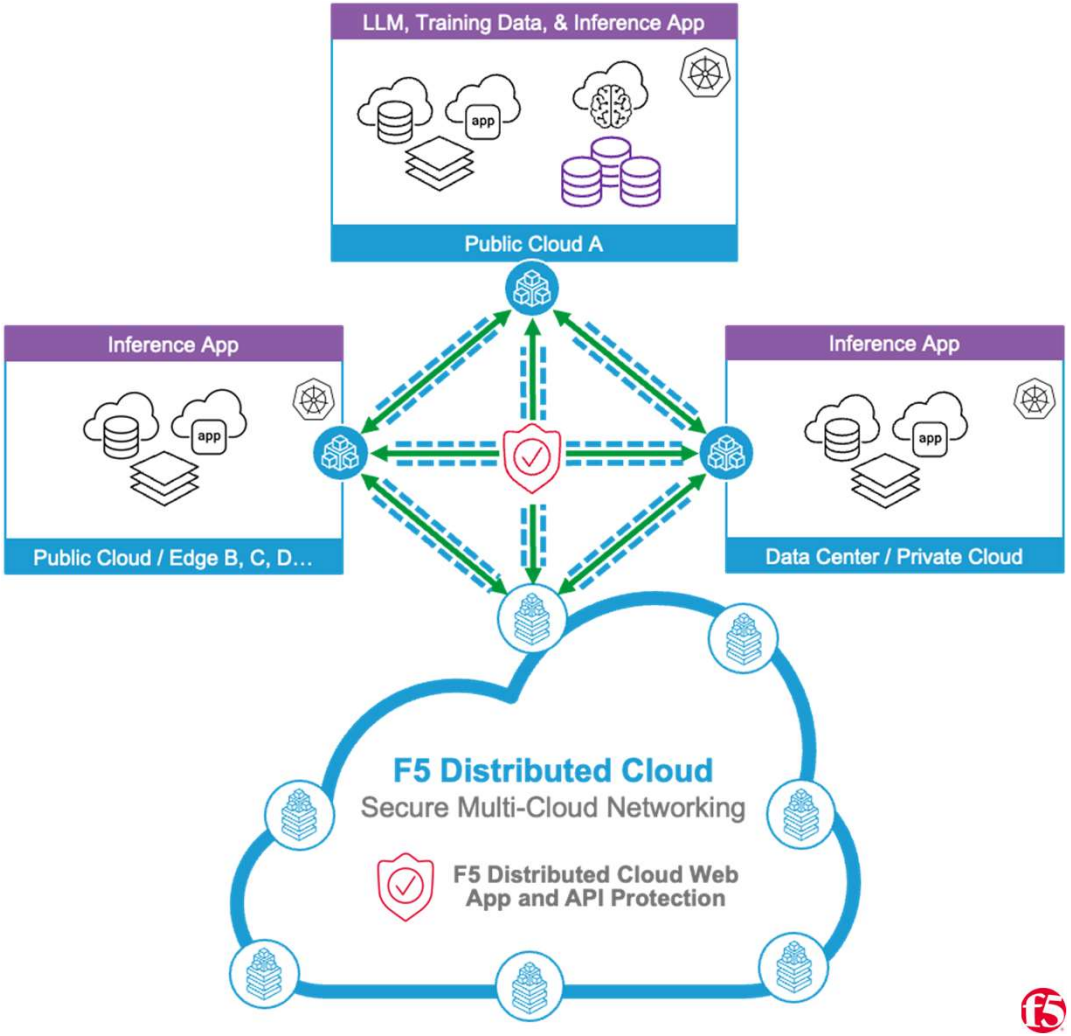
 

# The challenge of connecting AI inference apps

Traditionally, LLMs, training data, and inference apps are all deployed centrally, and any user accessing them is routed to a single, centralized location.

Now, those inference apps are deployed in distributed environments, and are decoupled from the LLM.

However, **secure connectivity is still required** to protect the LLM, training data, and apps.



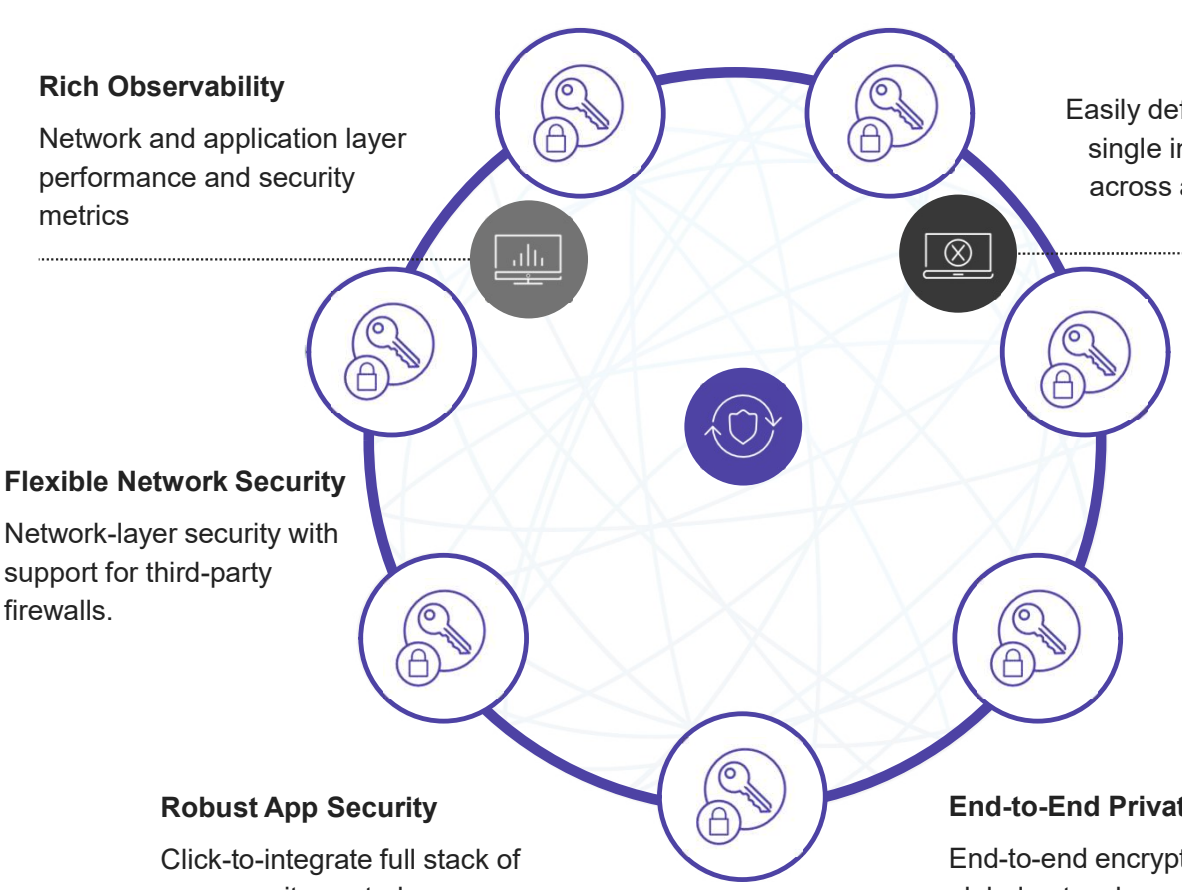


# Enabling connectivity across AI inference apps

**Secure Multi-Cloud Networking** ensures that inference apps can be deployed quickly to any cloud, edge, or on-prem site.

Each app has network connectivity already provisioned, with consistent L7 security, and visibility into traffic flows and threats across the network.



**Rich Observability**

Network and application layer performance and security metrics

**Intent-based Policy Enforcement**

Easily define, deploy, and enforce a single intent-based security policy across all apps and environments

**Flexible Network Security**

Network-layer security with support for third-party firewalls.

**Robust App Security**

Click-to-integrate full stack of app security controls

**End-to-End Private Network**

End-to-end encryption across the F5 global network

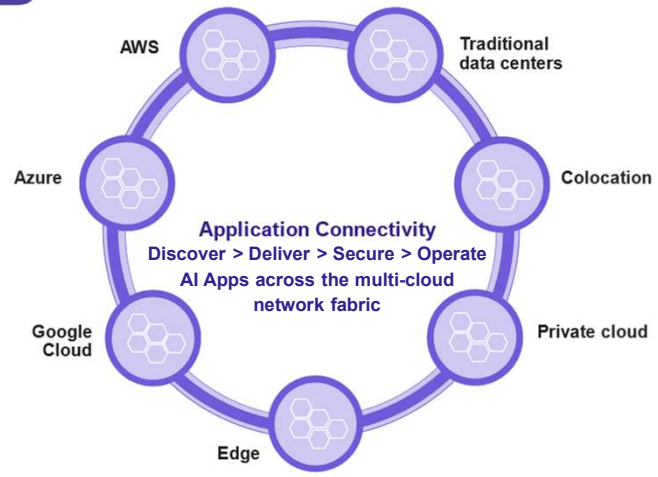
 

# F5 Distributed Cloud Secure Multi-Cloud Networking App Connect

Discover > Deliver > Secure > Operate **Any Inference App** across **Any Environment**

**Connect AI apps to speed up deployment at the edge**

### Distributed Cloud App Connect




### Key Capabilities for AI Apps

Secure, flexible deployment of inference apps, with connectivity to LLMs and data in any public cloud, on-premises, or edge site. App Connect offers:

- Service discovery
- App segmentation
- Service networking
- Multi-cluster ingress, egress, and deployment
- End-to-end encryption
- End-to-end observability
- Standardized app security integration across sites (e.g. WAAP, Bot Defense)

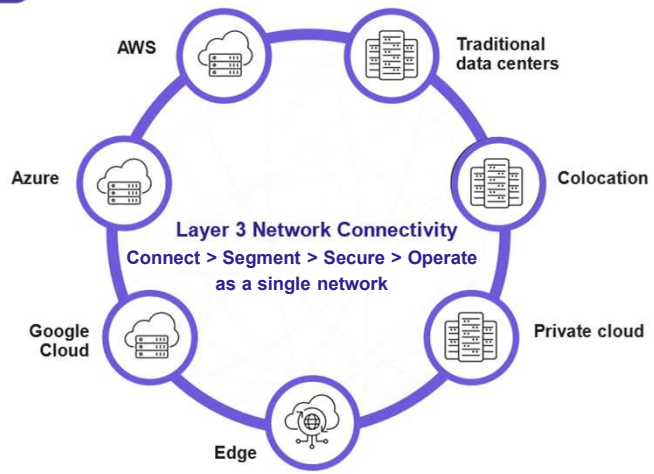
0 Same comment as slide 30 (this looks like key capabilities).
Byron McNaught; 2024-01-12T19:55:59.799

# F5 Distributed Cloud Secure Multi-cloud Networking Network Connect

Connect > Segment > Secure > Operate **Networks** across **Any Environment**

**Connect Networks to secure AI Apps**

### Distributed Cloud Network Connect




### Key Capabilities for AI Apps

Easily and securely network across public clouds, hybrid clouds, and edge sites, wherever your AI and LLM workloads are running. Network Connect offers:

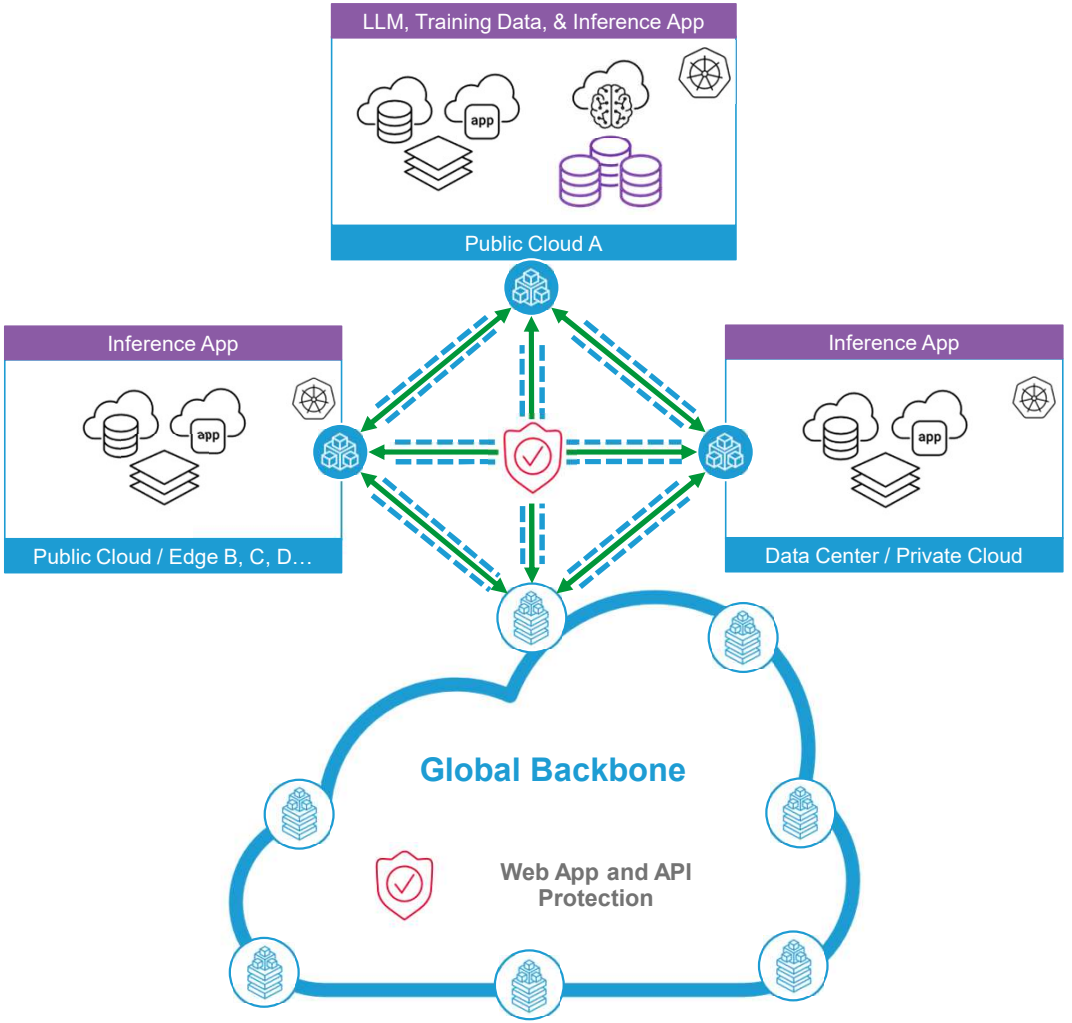
- Cloud orchestration and private connectivity
- Centralized resource access for identity control and shared data
- Global network segmentation
- Embedded network security and firewall service insertion at CEs
- End-to-end observability
- App networking integration

0 Should this be “key capabilities for AI Apps”?
Byron McNaught; 2024-01-12T19:55:38.902

0 0 Yeah, that works
Colin Clauset; 2024-01-18T16:14:45.549

# Generative AI Multi-Cloud Networking Challenges

- Connecting Customer Edge nodes directly to each other for internal apps with secure service connectivity, or via the Regional Edge for public-facing apps that need a front door service.
- Connecting and sharing data between inference apps, the central LLM, and training data through secure network connectivity.
- Enabling inference app deployment and maintenance at scale with an ingress controller in each app cluster.
- Enabling load balancing and security at each node, with centralized DDoS, Bot, and API protection, plus distributed web app firewalls locally.




LLM

Training Data

Inference App Cluster (App, Inference Model, Ingress Controller)

Secure Network & Service Connectivity

Customer Edge

Global Network Regional Edge

Web App & API Protection

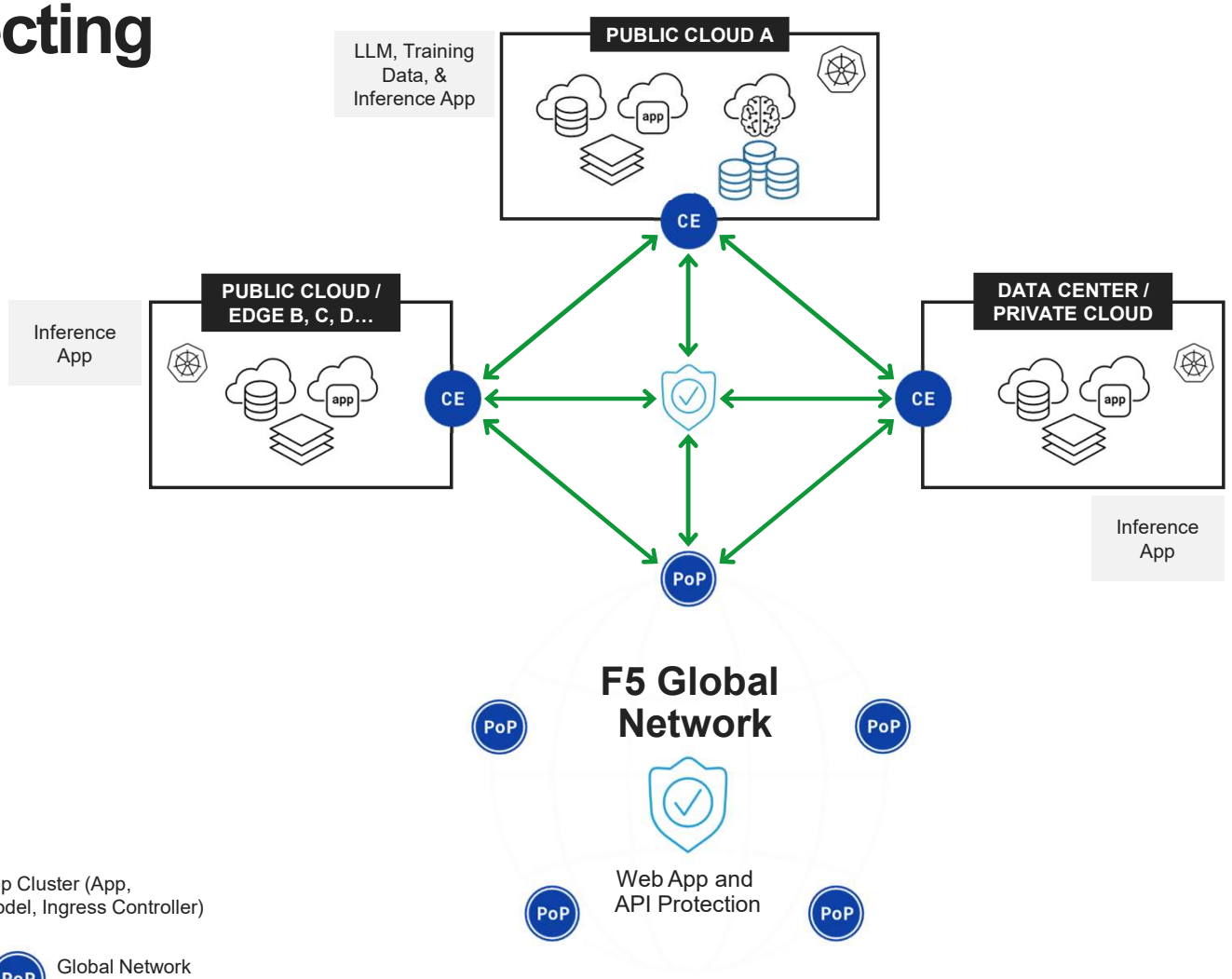
 

# The challenge of connecting AI inference apps

Traditionally, LLMs, training data, and inference apps are all deployed centrally, and any user accessing them is routed to a single, centralized location.

Now, those inference apps are deployed in distributed environments, and are decoupled from the LLM.

However, **secure connectivity is still required** to protect the LLM, training data, and apps.




**Key**

LLM

Training Data

Inference App Cluster (App, Inference Model, Ingress Controller)

Web App & API Protection

Secure Network & Service Connectivity

CE Customer Edge

PoP Global Network Regional Edge

# Enabling connectivity across AI inference apps

**Secure Multicloud Networking** ensures that inference apps can be deployed quickly to any cloud, edge, or on-prem site.

Each app has network connectivity already provisioned, with consistent L7 security, and visibility into traffic flows and threats across the network.

**Robust App Security**
Click-to-integrate full stack of app security controls

**Flexible Network Security**
Network-layer security with support for third-party firewalls

**End-to-End Private Network**
End-to-end encryption across the F5 global network

**Intent-based Policy Enforcement**
Easily define, deploy, and enforce a single intent-based security policy across all apps and environments

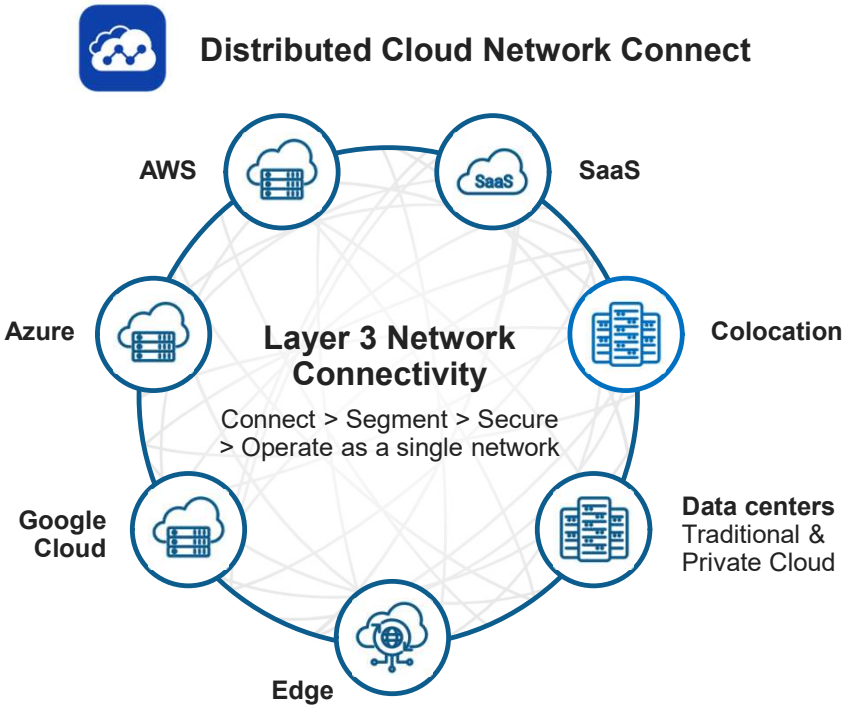
**Rich Observability**
Network and application layer performance and security metrics

# F5 Distributed Cloud Secure Multicloud Networking Network Connect

Connect > Segment > Secure > Operate **Networks** across **Any Environment**

## Connect Networks to secure AI Apps




### Key Capabilities for AI Apps

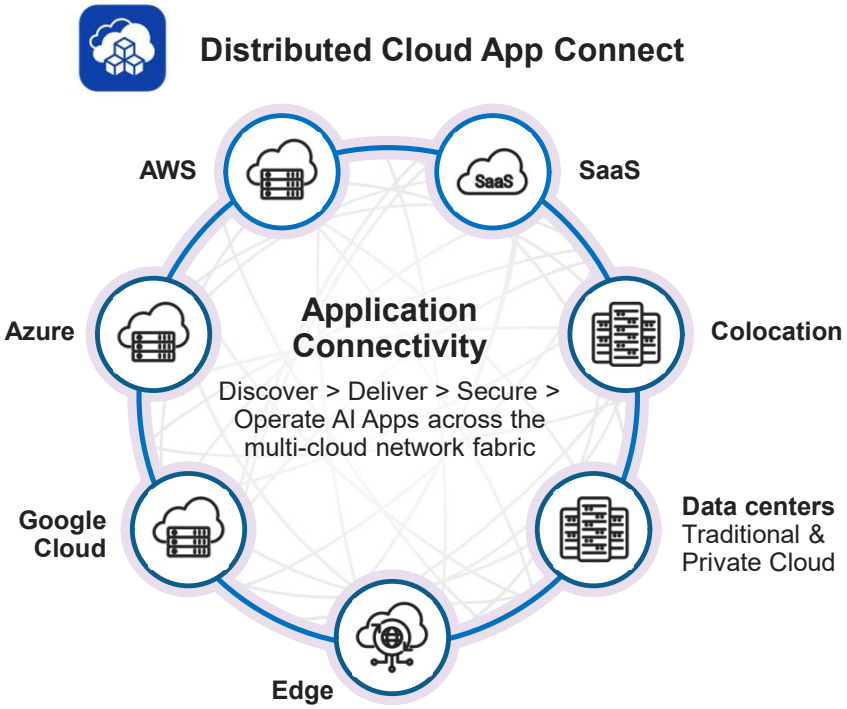
Easily and securely network across public clouds, hybrid clouds, and edge sites, wherever your AI and LLM workloads are running. Network Connect offers:

- Cloud orchestration and private connectivity
- Centralized resource access for identity control and shared data
- Global network segmentation
- Embedded network security and firewall service insertion at CEs
- End-to-end observability
- App networking integration

# F5 Distributed Cloud Secure Multicloud Networking App Connect

Discover > Deliver > Secure > Operate **Any Inference App** across **Any Environment**

**Connect AI apps to speed up deployment at the edge**




## Key Capabilities for AI Apps

Secure, flexible deployment of inference apps, with connectivity to LLMs and data in any public cloud, on-premises, or edge site. App Connect offers:

- Service discovery
- App segmentation
- Service networking
- Multi-cluster ingress, egress, and deployment
- End-to-end encryption
- End-to-end observability
- Standardized app security integration across sites (e.g. WAAP, Bot Defense)

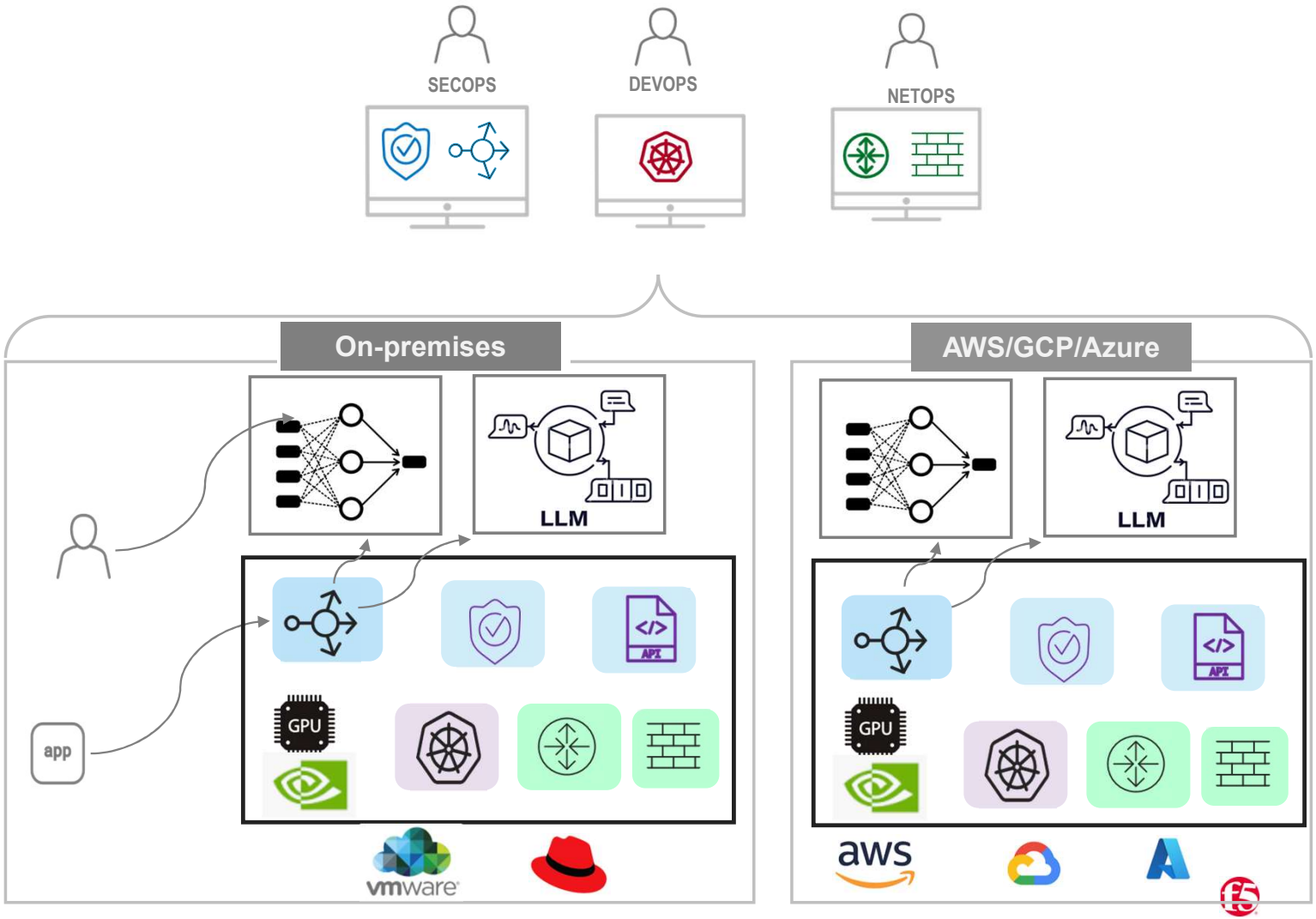
 

# Customer Edge Use-Case : AI/ML-as-a-service Challenges

Inconsistent Deployment & Operational Models

Multiple Point Products Cobbled Together Per Cloud

Slow TTM & AI/ML Transformation



# Customer Edge Use-Case : AI/ML-as-a-service

Consistent Deployment Model

Anywhere – On-prem/Cloud

Fully Integrated Stack

Delivering Inference Apps

Fast TTM

Lower TCO