



KubeCon



CloudNativeCon

Europe 2026

#KubeCon #CloudNativeCon

Beyond Nginx Ingress: Higress as the K8s Gateway for the AI Era

Huxing Zhang

Alibaba Cloud



The Retirement of Nginx Ingress



KubeCon



CloudNativeCon

Europe 2026

The Kubernetes SIG Network and Security Response Committee have formally announced the **retirement of Ingress NGINX**.

- Currently in strict **best-effort maintenance** phase only
- Definitively concludes **March 2026**
- Repository will be formally archived — **no further releases, bug fixes, or security patches**

Annotation Snippet Vulnerabilities

To bypass the Ingress API's limitations, Nginx Ingress relied on **annotation-based configuration snippets** — allowing direct injection of arbitrary NGINX config strings. This fundamentally violates least-privilege principles, opening massive vectors for cluster-wide compromise.

<https://kubernetes.io/blog/2026/01/29/ingress-nginx-statement/>

Scale of Impact

An estimated **50% of all global cloud-native environments** currently utilize some form of Nginx Ingress infrastructure.

Continuing beyond March 2026 leaves enterprise networks fundamentally exposed to unpatched zero-day vulnerabilities, violating basic compliance frameworks.

Challenges for Nginx-based ingress controllers



KubeCon



CloudNativeCon

Europe 2026

Traffic Jitter for Long-Lived Connections

gRPC, WebSockets, and Dubbo RPC connections are abruptly terminated or subjected to severe traffic jitter during every reload cycle.

Poor gRPC Load Balancing

Nginx's HTTP/1.1-centric design cannot intelligently balance discrete RPC calls over multiplexed HTTP/2 connections, causing **uneven distribution** and resource saturation.

Scalability issues

High-scale providers encountered critical, unresolvable barriers attempting to scale Nginx-based ingress to support over **thousands of tenants** in a k8s cluster

The evolution of Higress



KubeCon



CloudNativeCon

Europe 2026

Higress 1.0

2020

- ✓ Initial launch with search & recommendation services
- ✓ Integrated with Taobao, Alipay business
- ✓ Expanded to Ele.me, Koubei, Fliggy and other apps
- ✓ **Proven in Double 11** high-traffic scenarios (Nov 2020)

Higress 2.0

2021

- ✓ Standardized Aserver + Ingress solutions
- ✓ Launched as **MSE Cloud-Native Gateway in Alibaba Cloud** (Jul 2021)
- ✓ Combined Traffic Gateway + Microservice Gateway
- ✓ **50% cost reduction** vs self-built gateways

Higress 3.0

2022-

- ✓ **Officially open-sourced** (Nov 2022)
- ✓ Plugin marketplace, WAF protection, Nginx compatibility
- ✓ HTTP to Dubbo protocol conversion
- ✓ **3-in-1 Gateway**: Traffic + Microservice + Security

Higress Architecture

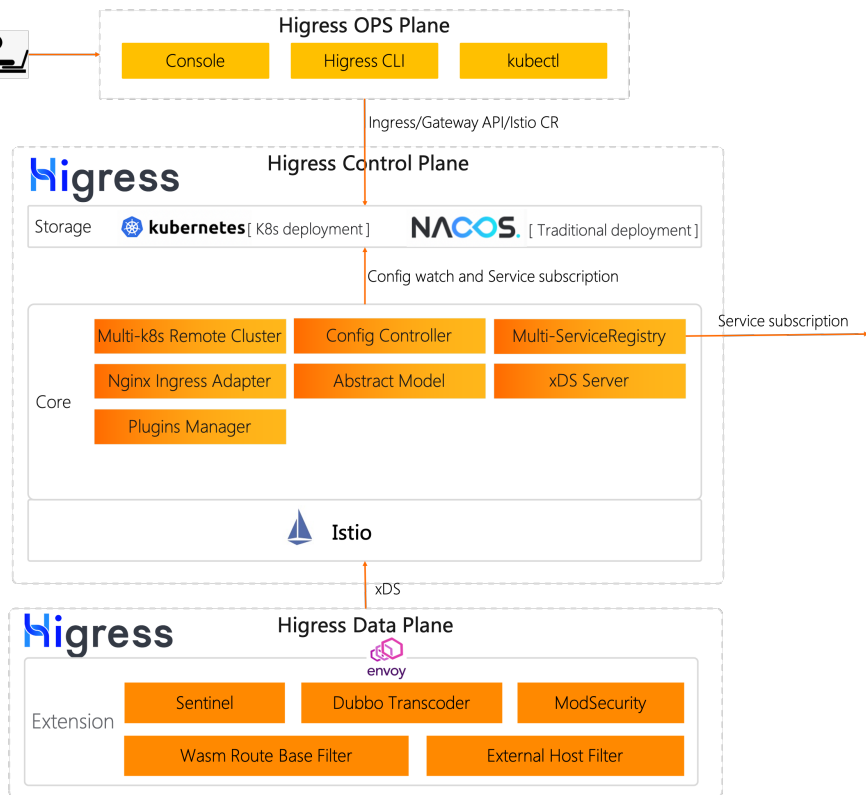


KubeCon



CloudNativeCon

Europe 2026



Control Plane enhancement

- Dynamic certificate management
- Service discovery for Consul, Nacos, Zookeeper...
- Dual stack support Nginx Ingress and Gateway API
- Custom CRDs(WasmPlugin, MCPbridge, Http2rpc)

Data Plane enhancement

- Protocol translation: HTTP to Apache Dubbo
- WASM plugin support
- Golang HTTP filter
- Multi-dimension structured telemetry

Higress Core Features



KubeCon



CloudNativeCon

Europe 2026

Zero downtime configuration reload

- True dynamic configuration eliminates traffic jitter and connection termination historically caused by proxy reloads.

Wasm Extensibility & Memory-Safe Sandboxing

- Hot reload
- Strict Sandbox Isolation
- Polyglot Development (Go/Rust/C++/JS)
- Distribution: OCI image/HTTP/File

Advanced Security Posture

- Integrated WAF, bot detection, replay protection, IP restriction..
- Rich Auth Strategies(JWT/API-KEY/HMAC/OIDC...)
- Guardrails, PII detection..

Migrating from Nginx Ingress to Higress

- Seamless Migration: Zero Modification Required



KubeCon



CloudNativeCon

Europe 2026

Higress is compatible with 90%+ common Nginx Ingress annotations.

Higress natively parses and executes the legacy `nginx.ingress.kubernetes.io` annotation prefix as fully equivalent to its own `higress.io` prefix — eliminating costly refactoring of thousands of YAML manifests.

Feature	Compatibility	Key Annotations Supported
Canary / Traffic Splitting	Fully Compatible	canary, canary-by-header, canary-by-header-pattern (regex), canary-by-cookie, canary-weight
Rewrite & Redirect	Fully Compatible	rewrite-target (Group Capture), upstream-vhost, ssl-redirect, permanent-redirect, app-root
CORS	Fully Compatible	enable-cors, cors-allow-origin, cors-allow-methods, cors-allow-credentials, cors-max-age
Disaster Recovery	Fully Compatible	default-backend, custom-http-errors (auto path rewrite on fallback)
Retry Policies	Fully Compatible	proxy-next-upstream-tries, proxy-next-upstream-timeout, proxy-next-upstream
Rate Limiting	Higress Extension	higress.io/route-limit-rpm, route-limit-rps, route-limit-burst-multiplier (token-bucket)
Security / Access Control	Partially Compatible	whitelist-source-range (CIDR), ssl-cipher. Legacy Basic Auth superseded by Wasm Auth plugins.

Bridging Ingress API to Gateway API

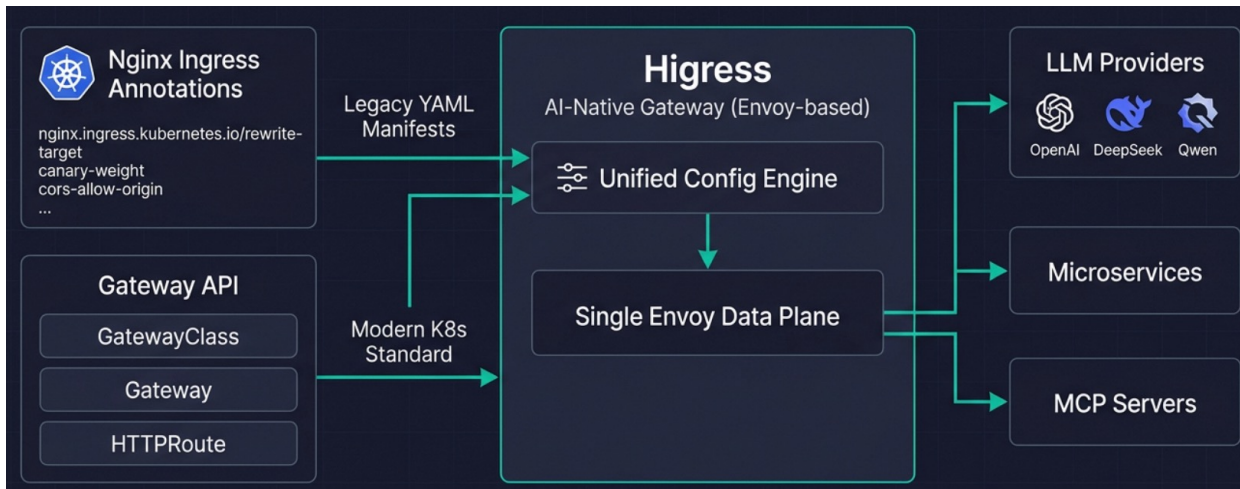


KubeCon



CloudNativeCon

Europe 2026



Higress Dual-Stack Advantage

Higress simultaneously ingests **both legacy Ingress objects and modern Gateway API objects**, applying both to the same Envoy data plane.

Gateway API with highest priority.

User Case study

- From Minutes to Seconds: How Sealos Conquered the 20,000-Domain Gateway Challenge



KubeCon



CloudNativeCon

Europe 2026



Sealos a cloud provider serving nearly 200,000+ users.



Scale: Successfully handles **20,000+ Ingresses** where Nginx failed at >10,000.



Production Impact: Reduced peak domain activation delay **from 30+ minutes down to <5 seconds**.



Performance: **10x** improvement, e.g., gateway memory usage per node decreased from 3GB to 300MB.



Resilience: Performance is nearly consistent, eliminating the exponential degradation seen with Nginx at scale.

 Alibaba Cloud



Kuaishou



PayPal



Sealos

The AI-Native Paradigm Shift in Enterprise Gateway Architecture



KubeCon



CloudNativeCon

Europe 2026

	Traditional API Gateway	AI-Native Gateway
Traffic Pattern	Stateless, short-lived HTTP request-response cycles	Long-running SSE streams, WebSocket, asynchronous agentic loops
Payload Size	Small, predictable payloads (KB-level)	Massive payloads with variable token-heavy responses (MB-level)
Connection Lifecycle	Milliseconds; open, respond, close	Minutes to hours; persistent connections for streaming inference

- **Higress as a LLM Gateway:** Multi-model routing, failover, token rate limiting, semantic caching, and observability
- **Higress as a MCP Gateway:** Zero-code transformation of legacy APIs into mcp servers.

Higress as a LLM Gateway: core features



KubeCon



CloudNativeCon

Europe 2026

Model fall back

1. Monitor Time-To-First-Token (TTFT) on all backend nodes
2. Trigger failure if first packet exceeds 200ms threshold
3. Eject node when failure rate hits 50% / 1-second window
4. Reroute to cloud fallback (Qwen / Dashscope...)
5. Restore primary node after 30-second ejection period

Token aware rate limiting

Higress uses dedicated WASM plugins (`ai-token-ratelimit` and `ai-quota`) to enforce strict economic boundaries based on **actual token consumption**.

Semantic caching

The `ai-cache` plugin uses **GJSON path syntax** to extract semantic content from nested request/response bodies and cache against a Redis cluster.

Observability

- Metrics: `reasoning_tokens/cached_tokens/TTFT/TPOT..`
- Tracks metrics across **four dimensions**: gateway, route, service, and model.
- Tracing: Zipkin/SkyWalking/OpenTelemetry
- Dynamically configured AI logs (`session_id`, `reasoning`, `tool calls...`)

Higress as a MCP Gateway: core features



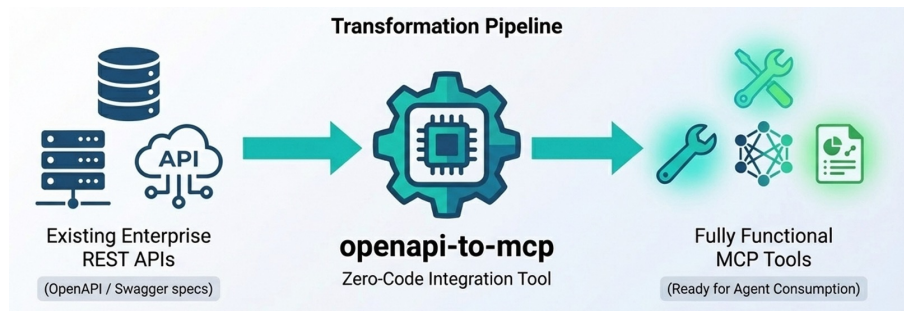
KubeCon



CloudNativeCon

Europe 2026

OpenAPI-to-mcp Conversion



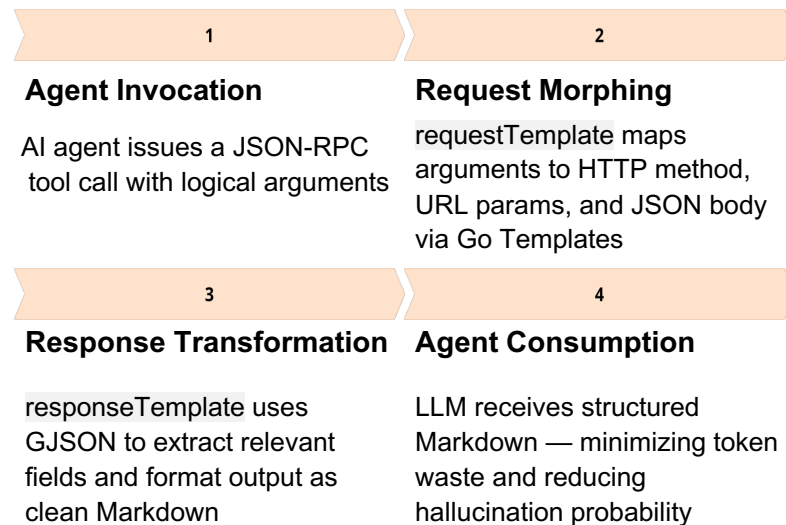
- **Parameter Positioning**

Maps OpenAPI schema locations (in: path, in: query, etc.) to MCP tool arguments — agents call logical functions, not HTTP endpoints.

- **Schema Flattening**

Parses complex allOf object references into simplified representations, preventing LLM hallucination from fragmented API definitions.

Advanced Payload Morphing



Case study: Ctrip and Ant Group



KubeCon



CloudNativeCon

Europe 2026



Ctrip (Trip.com) AI Gateway

Ctrip built its internal AI gateway on Higress due to its battle-tested Istio/Envoy architecture and extensible Wasm framework. The gateway serves as the sole ingress for all internal AI traffic, standardizing disparate LLM interfaces and managing MCP services via SSE and Streamable HTTP.

Key outcomes:

- Bearer Token authentication
- cross-business-unit token cost allocation
- automated zero-downtime failovers



Ant Group(Alipay) SOFA AI Gateway

The SOFA AI Gateway leverages the open-source Higress kernel to manage massive enterprise topologies. Focused on **intelligent traffic routing**, **unified protocol access**, and **advanced rate-limiting**, SOFA implementation offloads complex traffic management, security, and model scheduling to the gateway layer — drastically **reducing integration complexity** and **significantly accelerating AI agent deployment velocity** across the enterprise.



Sealos

Higress Joins CNCF Sandbox!



KubeCon



CloudNativeCon

Europe 2026

Higress just passed the CNCF sandbox vote!^[1]

Website: <https://higress.io/en>

Github: <https://github.com/alibaba/higress>

Contributions are welcome!

[1] <https://github.com/cncf/sandbox/issues/445>



git-vote bot 3 days ago – with [Git-Vote](#)

Vote closed

The vote **passed!** 🎉

90.91% of the users with binding vote were in favor and 0.00% were against (pas

Summary

In favor	Against	Abstain	Not voted
10	0	0	1

Binding votes (10)

User	Vote	Timestamp
@TheFoxAtWork	In favor	2026-03-10 15:47:32.0 +00:00:00
@angellk	In favor	2026-03-10 16:08:25.0 +00:00:00
@chira001	In favor	2026-03-11 14:52:43.0 +00:00:00
@dims	In favor	2026-03-10 16:36:12.0 +00:00:00
@jeremyrickard	In favor	2026-03-10 20:13:29.0 +00:00:00
@kevin-wangzefeng	In favor	2026-03-10 18:31:25.0 +00:00:00
@kfaseela	In favor	2026-03-10 15:38:37.0 +00:00:00
@kgamanji	In favor	2026-03-11 15:40:50.0 +00:00:00
@linsun	In favor	2026-03-11 15:57:05.0 +00:00:00
@rochaporto	In favor	2026-03-10 16:26:05.0 +00:00:00

Roadmap



KubeCon



CloudNativeCon

Europe 2026

NOW

NEXT (v2.2.0)

FUTURE



**Mature AI & Ingress
Gateway**



**Embracing Standards:
Gateway API Inference
Extension**

**Enabling Real-Time AI:
WebRTC Support**

Deepening Kubernetes Integration (v2.2.0): Initial support for the Gateway API Inference Extension is complete.

Next-Gen Protocols (In Development):
Proactively building WebRTC support for the next wave of real-time(e.g. OpenAI RealTime API), bidirectional AI applications.

Higress is a vital, mature, and forward-looking project providing critical infrastructure for the entire CNCF ecosystem, stabilizing the present while building the future.



KubeCon



CloudNativeCon

Europe 2026

Q & A

higress.io |
github.com/alibaba/higress

