

Higress是什么？

Higress 简介

- 什么是 AI 网关
- 核心优势
- 使用场景

Higress 简介

Higress 是一款云原生 API 网关，内核基于 Istio 和 Envoy，可以用 Go/Rust/JS 等编写 Wasm 插件，提供了数十个现成的通用插件，以及开箱即用的控制台（demo 点这里）

Higress 在阿里内部为解决 Tengine reload 对长连接业务有损，以及 gRPC/Dubbo 负载均衡能力不足而诞生。

阿里云基于 Higress 构建了云原生 API 网关产品，为大量企业客户提供 99.99% 的网关高可用保障服务能力。

Higress 基于 AI 网关能力，支撑了通义千问 APP、百炼大模型 API、机器学习 PAI 平台等 AI 业务。同时服务国内头部的 AIGC 企业（如零一万物），以及 AI 产品（如 FastGPT）

APSARA 云栖大会

典型客户 – 阿里云内部实践

通义App

Higress

业务 Server 业务Server

Higress作为流量网关，白屏操作

✓ 支持长连接 SSE/WebSocket，热更新对长连接流量无损

✓ 支持流式传输，满足AI大带宽/高延时特性诉求

✓ 支持多种安全认证与限流防护

AI 应用

DashScope
模型服务灵积

Higress

业务网关 业务网关

Higress作为流量网关，白屏操作

✓ 支持长连接 SSE/WebSocket，热更新对长连接流量无损

✓ 支持流式传输，满足AI大带宽/高延时特性诉求

AI 模型服务平台

PAI

Higress

Model Model

使用 Ingress 配置路由，通过Ingress集成PAI的管控

✓ 支持1W+ 超大路由/域名规模场景，多租共享集群模式，切换到Higress后路由配置生效 RT 从原10分钟降到30秒内

✓ 构建完善可观测体系

AI 模型

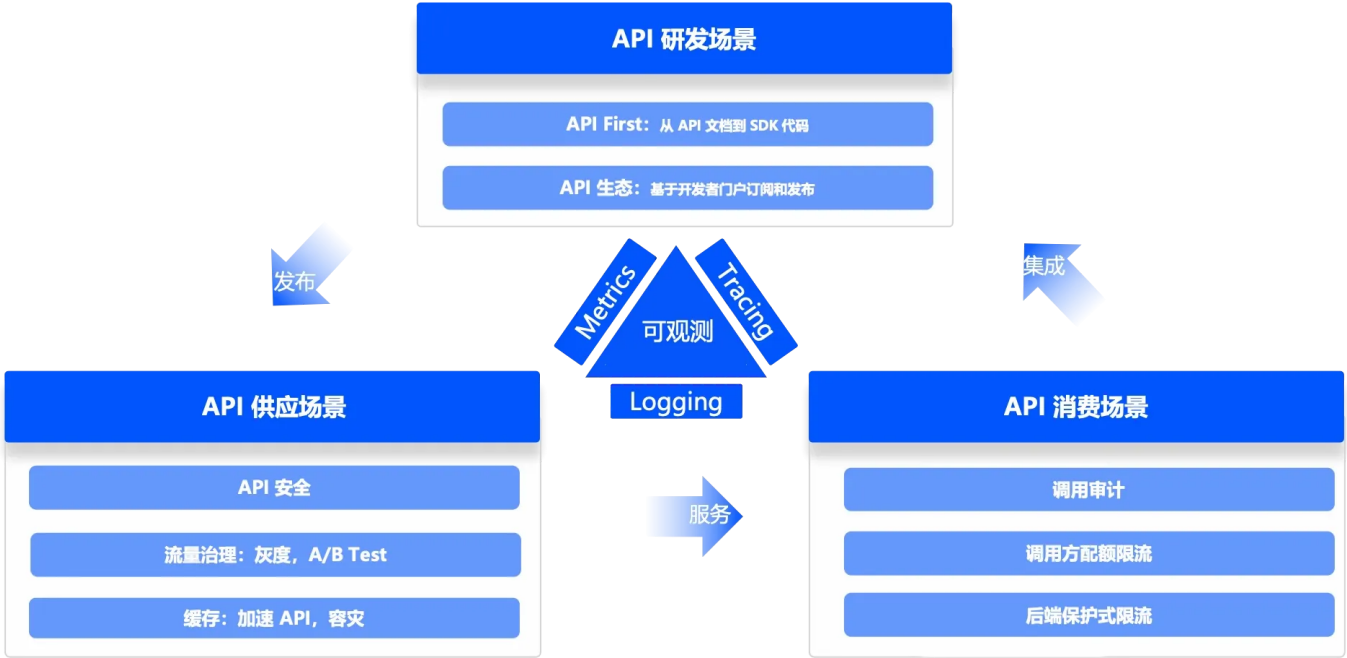
1

什么是 AI 网关

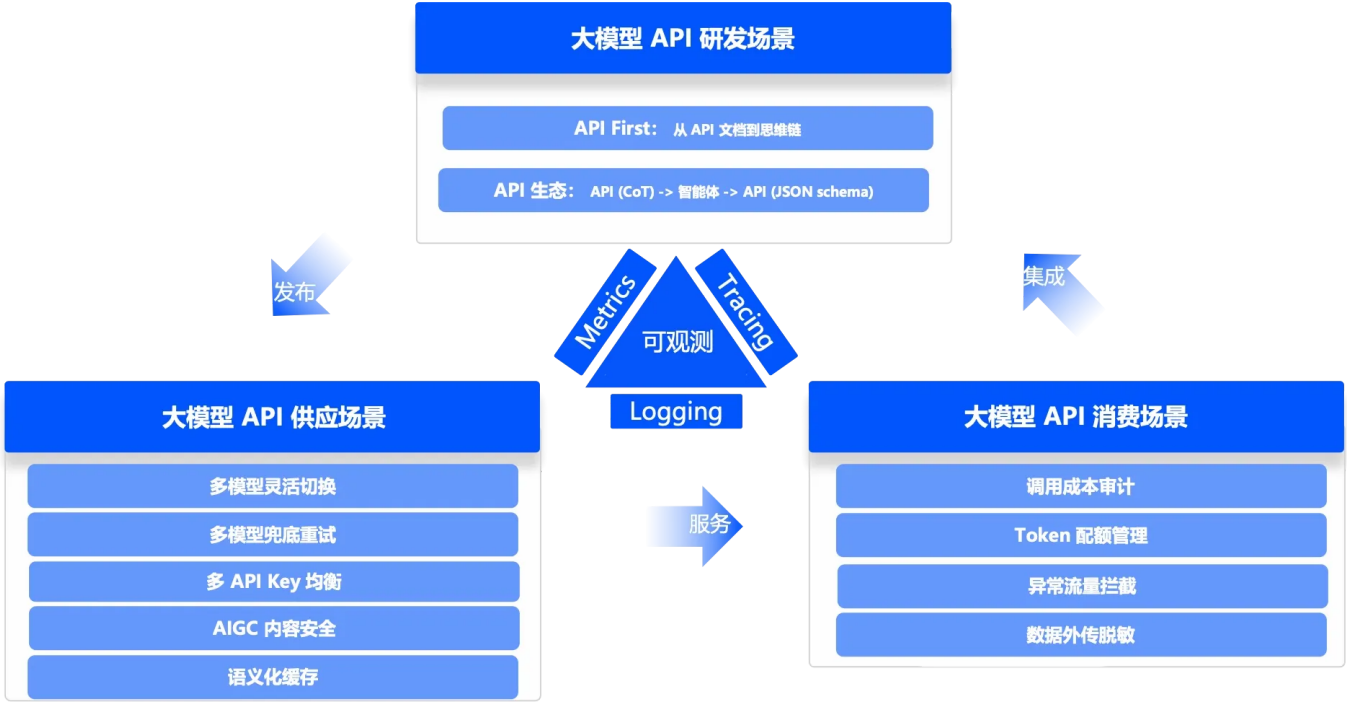
AI Gateway = AI Native API Gateway

AI 网关的本质依然是 API 网关，AI 原生的意义在于，在这样的 API 网关里，AI 是一等公民。API 研发，API 供应，API 消费，以及 API 观测都基于 AI 场景下的需求，演进出全新的能力。

这是传统 API 网关的功能范畴，AI 场景下仍然有其通用价值：



在 AI 场景下，基于 Higress 可以将 API 网关的功能范畴进一步扩展：



核心优势

- 生产等级

脱胎于阿里巴巴多年生产验证的内部产品，支持每秒请求量达数十万级的大规模场景。

彻底摆脱 Nginx reload 引起的流量抖动，配置变更毫秒级生效且业务无感。对 AI 业务等长连接场景特别友好。

- 便于扩展

提供丰富的官方插件库，涵盖 AI、流量管理、安全防护等常用功能，满足90%以上的业务场景需求。

主打 Wasm 插件扩展，通过沙箱隔离确保内存安全，支持多种编程语言，允许插件版本独立升级，实现流量无损热更新网关逻辑。

- 安全易用

基于 Ingress API 和 Gateway API 标准，提供开箱即用的 UI 控制台，WAF 防护插件、IP/Cookie CC 防护插件开箱即用。

支持对接 Let's Encrypt 自动签发和续签免费证书，并且可以脱离 K8s 部署，一行 Docker 命令即可启动，方便个人开发者使用。

- 流式处理

支持真正的完全流式处理请求/响应 Body，Wasm 插件很方便地自定义处理 SSE（Server-Sent Events）等流式协议的报文。

在 AI 业务等大带宽场景下，可以显著降低内存开销。

使用场景

- AI 网关:

Higress 能够用统一的协议对接国内外所有 LLM 模型厂商，同时具备丰富的 AI 可观测、多模型负载均衡/fallback、AI token 流控、AI 缓存等能力：



- **Kubernetes Ingress 网关:**

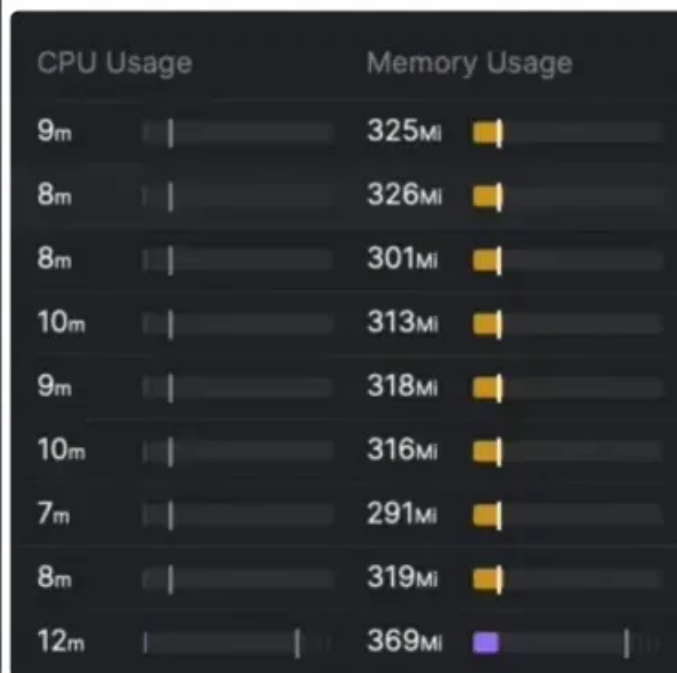
Higress 可以作为 K8s 集群的 Ingress 入口网关, 并且兼容了大量 K8s Nginx Ingress 的注解, 可以从 K8s Nginx Ingress 快速平滑迁移到 Higress。

支持 Gateway API 标准, 支持用户从 Ingress API 平滑迁移到 Gateway API。

相比 ingress-nginx, 资源开销大幅下降, 路由变更生效速度有十倍提升:



Ingress-nginx



Higress

```

$./test.sh new-ingress.yaml nginx
Applying YAML...
ingress.networking.k8s.io/new-ingress created
YAML applied
Ingress YAML took 34138 milliseconds to take effect.
Deleting YAML...
ingress.networking.k8s.io "new-ingress" deleted
YAML deleted
Ingress YAML took 36061 milliseconds to be removed.
  
```

创建生效需 34 秒
Nginx
移除生效需 36 秒

```

$./test.sh new-ingress.yaml higress
Applying YAML...
ingress.networking.k8s.io/new-ingress created
YAML applied
Ingress YAML took 3065 milliseconds to take effect.
Deleting YAML...
ingress.networking.k8s.io "new-ingress" deleted
YAML deleted
Ingress YAML took 1763 milliseconds to be removed.
  
```

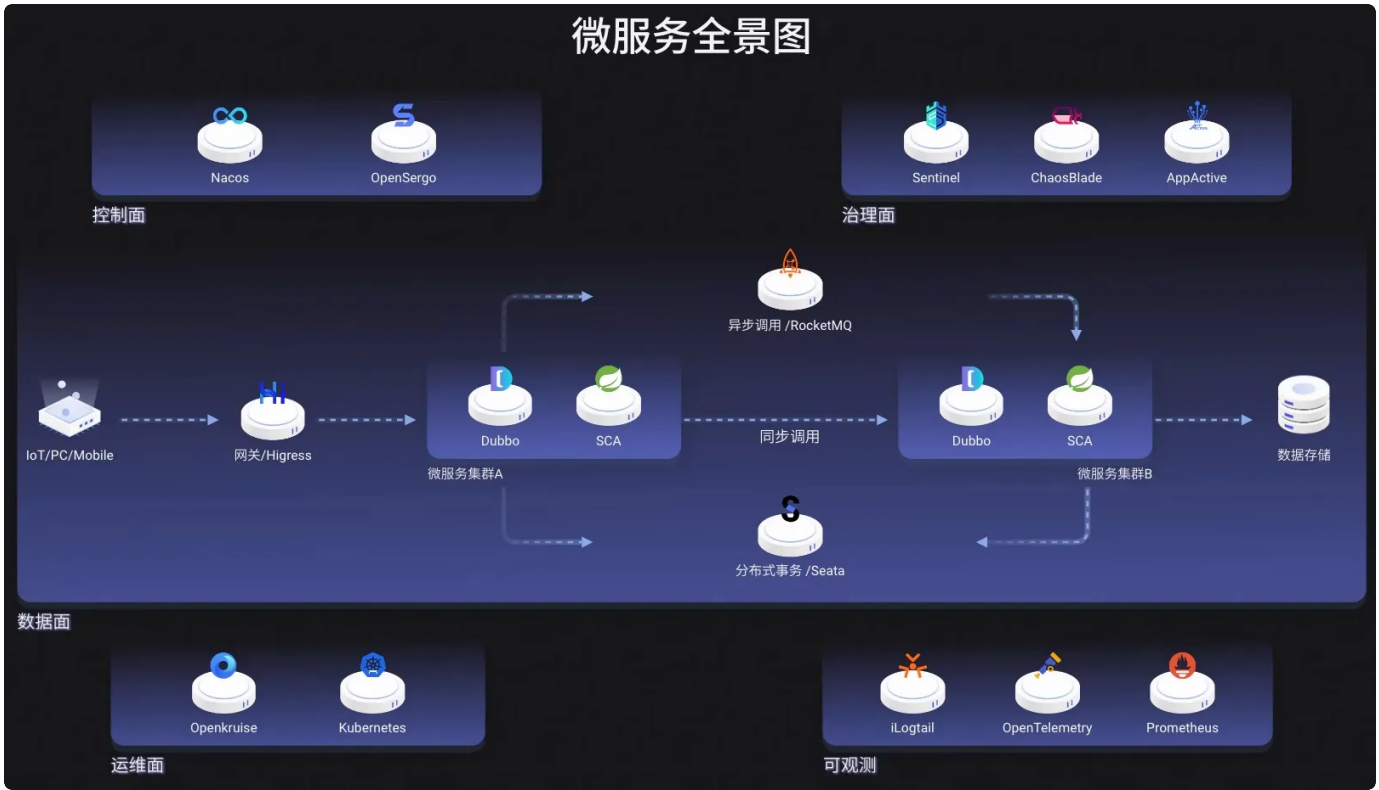
创建生效需 3 秒
Higress
移除生效需 2 秒

10000 个路由下变更一条路由的耗时对比

- 微服务网关:

Higress 可以作为微服务网关, 能够对接多种类型的注册中心发现服务配置路由, 例如 Nacos, ZooKeeper, Consul, Eureka 等。

并且深度集成了 Dubbo, Nacos, Sentinel 等微服务技术栈, 基于 Envoy C++ 网关内核的出色性能, 相比传统 Java 类微服务网关, 可以显著降低资源使用率, 减少成本。



• 安全防护网关:

Higress 可以作为安全防护网关, 提供 WAF 的能力, 并且支持多种认证鉴权策略, 例如 key-auth, hmac-auth, jwt-auth, basic-auth, oidc 等。