DATA CENTER PRIVATE MAAS

数据中心私有化 MaaS 完整技术架构

01

平台分层总图

一套中央运营平台，连接三类客户入口，多个数据中心与三种服务资源池。

CONTROL PLANE

◆ REFERENCE STATE

01 · DEMAND
企业客户
金融、制造、能源、政务及集团企业

02 · SUPPLY
算力资产方
国企智算中心、上市公司、IDC 与园区

03 · ECOSYSTEM
生态伙伴
模型商、集成商、行业软件与定付伙伴

04 · OPERATIONS
联合运营团队
商务、交付、SRE、模型工程与客户成功

业务入口统一整合

L1

接入与体验层

EXPERIENCE

运营控制台
资产、客户、合同、运营驾驶舱

企业门户
模型申请、部署、API Key、用量与 SLA

伙伴门户
模型上架、替代方案、项目协作与交付

统一 API / SDK
OpenAI 兼容 · 多模态 · 异步任务与管理 API

认证 · 授权 · ENTITLEMENT

L2

商业运营控制面

FOREGROUND CORE · SELF-BUILT

A 客户与组织
组织、项目、角色、环境、合作伙伴

B 服务目录
模型商品、资源规格、共享/专属套餐

C 合同与 SLA
额度、折扣、期限、承诺、超售与续费

D 资产商品化
服务器、GPU、机房、资源池与可售库存

E 部署编排
运行时选择、Placement、发布、升级与回滚

F 计量与权益
Token、GPU 小时、套餐额度、降级及限额

G 运营与工单
巡检、故障、变更、交付和客户成功

H 经营驾驶舱
利润率、收入能力、单位成本、SLA 与预测

产品化资源池 ▼

◆ 对外 / 内部 / 计量事件

L3

平台公共服务

SHARED CAPTIVES

API Gateway
Envoy Gateway · WAF · 限流 · 路由

身份与密钥
Keycloak · OIDC/SAML · Vault/KMS

模型供应链
许可证归档 · 安全扫描 · Registry · 镜像

事件与数据
Kafka/NATS · PostgreSQL · Redis

用量聚合
原始计量流水 · OpenMeter · 对账

可观测性
Prometheus · OTel · Loki · Langfuse

FLEET API · CLUSTER AGENT · GITOPS

L4

多集群与安全域

FLEET MANAGEMENT

中央 Fleet Hub
集群登录、Placement、策略、服务和资源视图

Cluster Agent
资源发现、心跳、策略实施、离线补偿

OCM
集群注册、治理、策略和工作负载投放

Karmada · 位置
跨集群调度、灾备和多云资源传播

安全边界 中央控制面不保存 Prompt、文档、图片，敏感规则编译后推送至本地；完全隔离区只拉取资源指标与变名用量事件。

共享 API 流量

企业专属流量

资源 / 成员事件

POOL A

共享 LLM MaaS 池

高利用率 · 按量服务

AIBrix
Gateway · KV/Die-aware Routing · Autoscaling · LoRA

vLLM	SGLang	TRT-LLM
默认引擎	高吞吐场景	延迟敏感场景

标准大模型、Embedding 与高并发 API，通过成本和时延表现选择引擎。

POOL B

企业专属 / 多模态池

隔离部署 · 统一生命周期

KServe
Router · Scale · Canary · Health · Runtime

BentoML	Triton	TEI / ASR / TTS
业务开发	低时延推理	专用引擎

专属 LLM、图片、视频、语音、OCR、Embedding 与客户自定义模型。

POOL C

训练与批处理池

异步调度 · 资源复用

Volcano + Kueue
Gang · Priority · GP Aware · Hang Scheduling

Ray / KubeRay	MLflow	Batch Inference
分布式计算	实验与登记	异步任务

数据、评测、批量推理、数据处理及复杂多模型 Pipeline。

KUBERNETES RESOURCE API · DEVICE PLUGIN · SCHEDULER

L5

集群基础设施

PER CLUSTER

Kubernetes
Namespace · RBAC · NetworkPolicy · CNI

GPU / NPU 管理
GPU Operator · DCGM · 厂商 Device Plugin

异构资源共享
HAMi · MIG · vGPU · 软件模拟调度

数据与供应链
S3 对象存储 · 预热仓库 · 模型缓存 · PVC

安全与审计
镜像签名 · Secret · 审计日志 · 准入策略

电信 · 国有 · 网络 · 交通 · 电力与环境监测

L6

异构算力与部署位置

HETEROGENEOUS INFRASTRUCTURE

ACCELERATOR

NVIDIA
H100 / H200 / B 系列 / L 系列

ACCELERATOR

国产 GPU / NPU
华为 · 寒武 · 摩尔线 · 壁仞等厂商

LOCATION

合作智算中心
共享池、专属池、弹性高密区

LOCATION

客户本地机房
电信 · 国有企 · 政务不出域

LOCATION

海外区域
出海业务、数据跨境与合规交付