



AI 应用正当时， 详解 AI 应用开发新范式

洛浩

云原生-解决方案架构师



AI 原生、Agent 驱动

AI Agent 正成为下一代应用的核心驱动力

- ✓ **AI 原生应用与Agentic AI 的崛起：** AI 不再仅仅是数据分析的辅助工具，而是逐步成为业务流程和应用的核心驱动力。
- ✓ **成本效益与可持续性：** 随着数据量和AI工作负载的爆炸式增长，成本控制和资源优化将成为企业选择平台的关键因素。**无服务器和按量付费模式将成为主流**，同时，平台将提供更精细的成本管理和优化工具，以确保长期运营的可持续性。
- ✓ **Data+AI 能力的普及化：** 行业致力于**通过自然语言交互、无代码/低代码工具等方式**，将数据洞察和AI 能力赋能给更广泛的非技术业务用户，加速AI在企业内部的普及和应用。
- ✓ **可观测性成为核心竞争力：** 随着现代系统复杂性的不断增加，高效且经济地处理海量日志、指标和追踪数据变得至关重要。可观测性能力正与实时分析紧密结合，成为确保 AI 应用和数据平台稳定运行的关键组成部分。

Contents

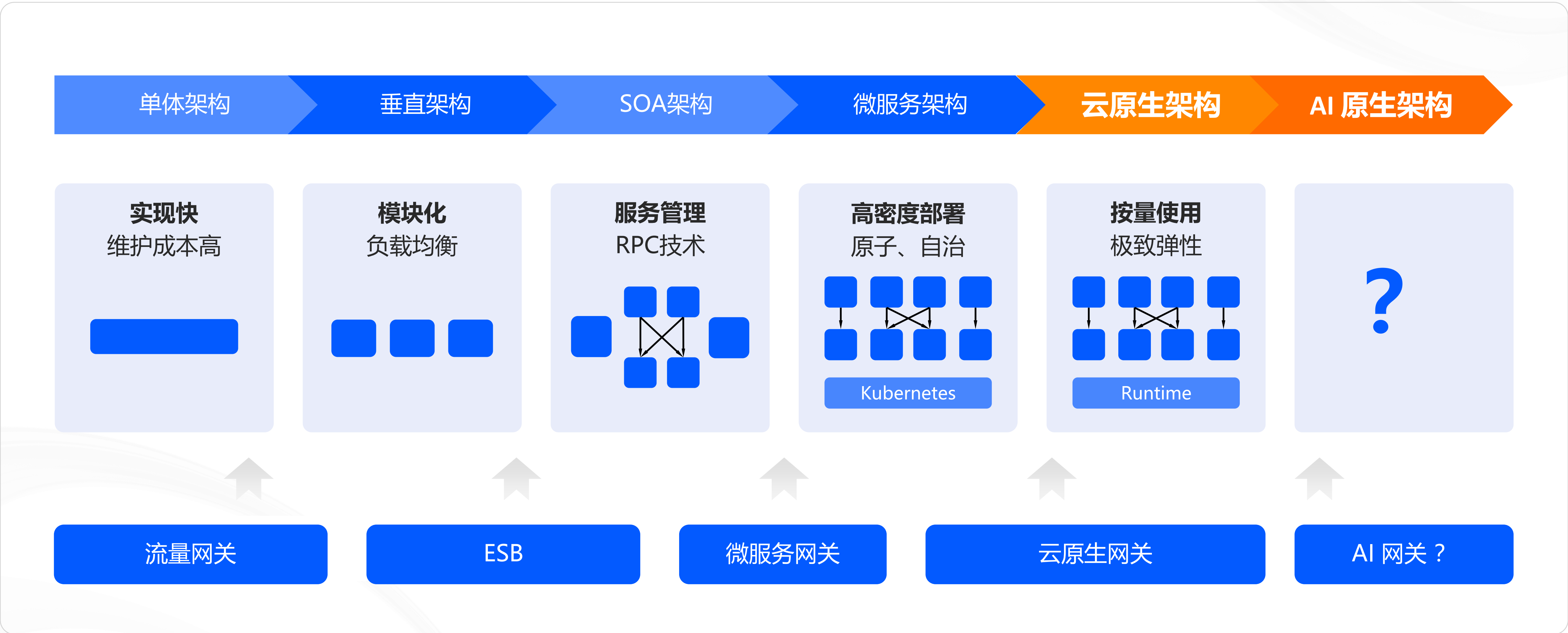
目录

01 从云原生应用到 AI 原生应用

02 MCP 加速推动 AI Agent 应用落地

03 AI Agent Sandbox

应用架构演进



AI 原生应用架构

AI 原生应用架构：AI 原生、Agent 驱动，以数据为中心，整合工具链





阿里云云原生应用平台——从云原生应用到 AI 原生应用

用户业务

云原生应用平台（从云原生应用到 AI 原生应用）

云原生中间件

云原生 API 网关
(API 网关、AI 网关)

微服务引擎 MSE
(注册配置中心、微服务治理、分布式任务调度)

可观测
(日志、应用监控、全链路追踪、云监控、压测)

消息队列
(RocketMQ、Kafka、MQTT、MNS、EventBridge)

开发者

云效

云原生计算运行时

Serverless 应用计算平台
(Serverless 应用引擎 SAE、函数计算 FC、Serverless 应用开发平台 CAP)

容器服务平台
(镜像服务 ACR、K8s 容器服务 ACK、Serverless 容器服务 ACS)

通义灵码

数据库

数据库
(PolarDB/RDS)

AI 大数据/ AI 安全

大数据平台
AI 平台
内容安全
.....

IAAS（计算、存储、网络）

AI 应用架构新范式



Contents 目录

01 从云原生应用到 AI 原生应用

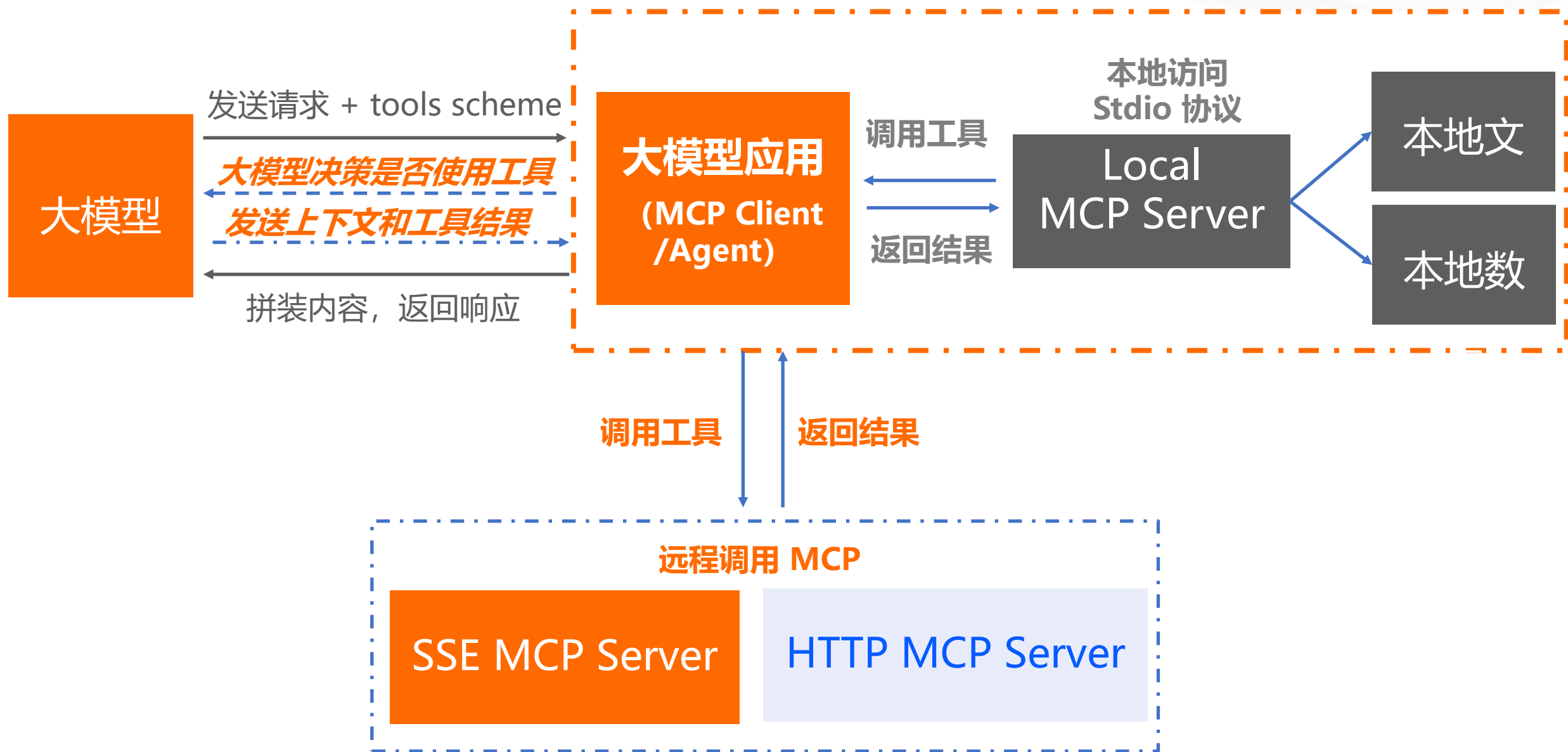
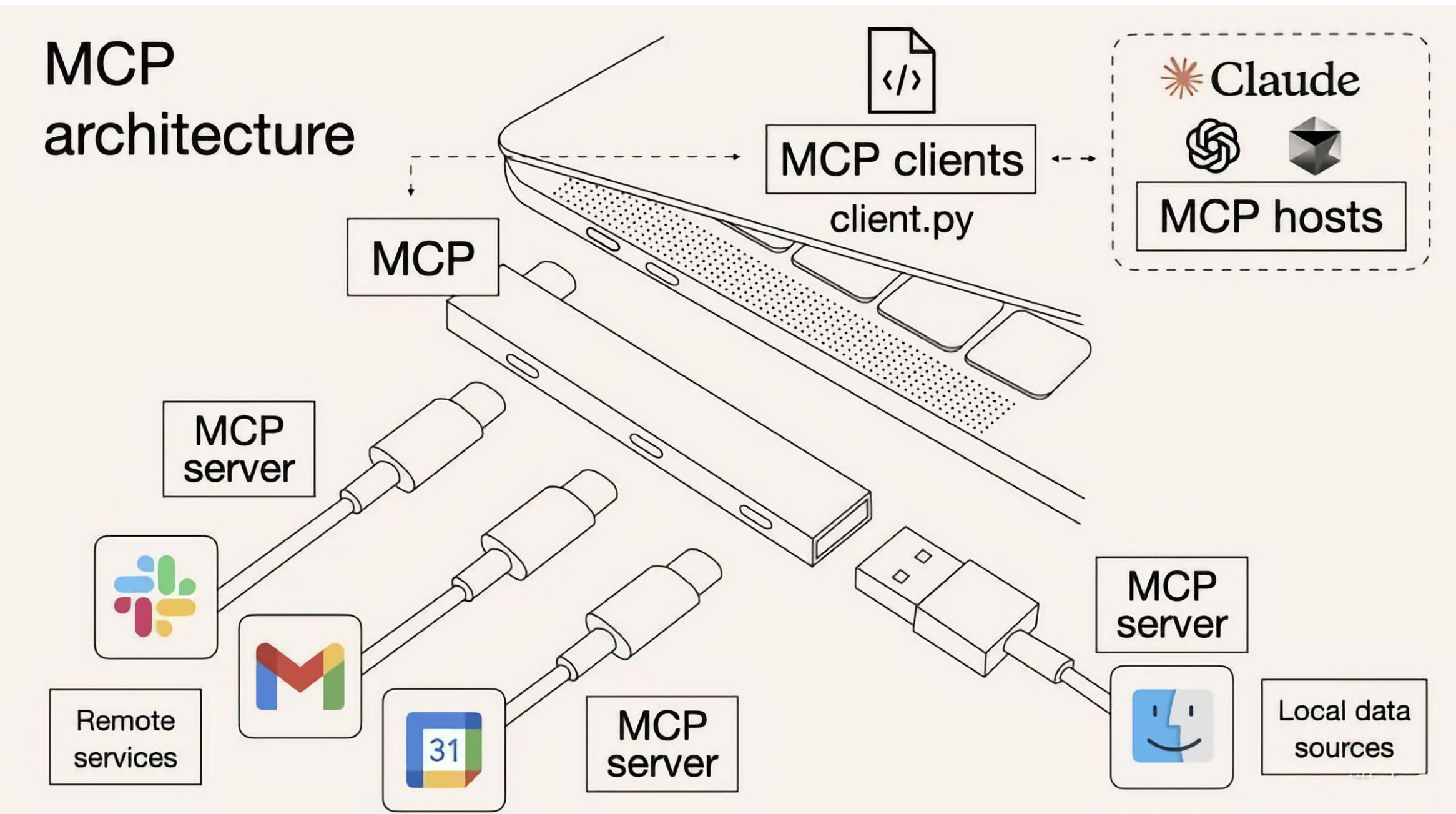
02 MCP 加速推动 AI Agent 应用落地

- MCP 三要素之 Agent 开发
- MCP 三要素之 大模型
- MCP 三要素之 MCP Server 开发
- AI 网关
- AI 应用观测

03 AI Agent Sandbox

MCP 是什么

模型上下文协议（**M**odel **C**ontext **P**rotocol）是一个开源协议，由 Anthropic（Claude开发公司）开发，旨在让大型语言模型（LLM）能够以标准化的方式连接到外部数据源和工具。它就像 AI 应用的通用接口，帮助开发者构建更灵活、更具上下文感知能力的AI应用，而无需为每个AI模型和外部系统组合进行定制集成。



标准化

MCP标准化了LLM访问外部数据的方式，简化了不同数据源和工具的集成。

模块化

MCP促进了模块化设计，允许独立开发和维护不同组件。

可扩展性

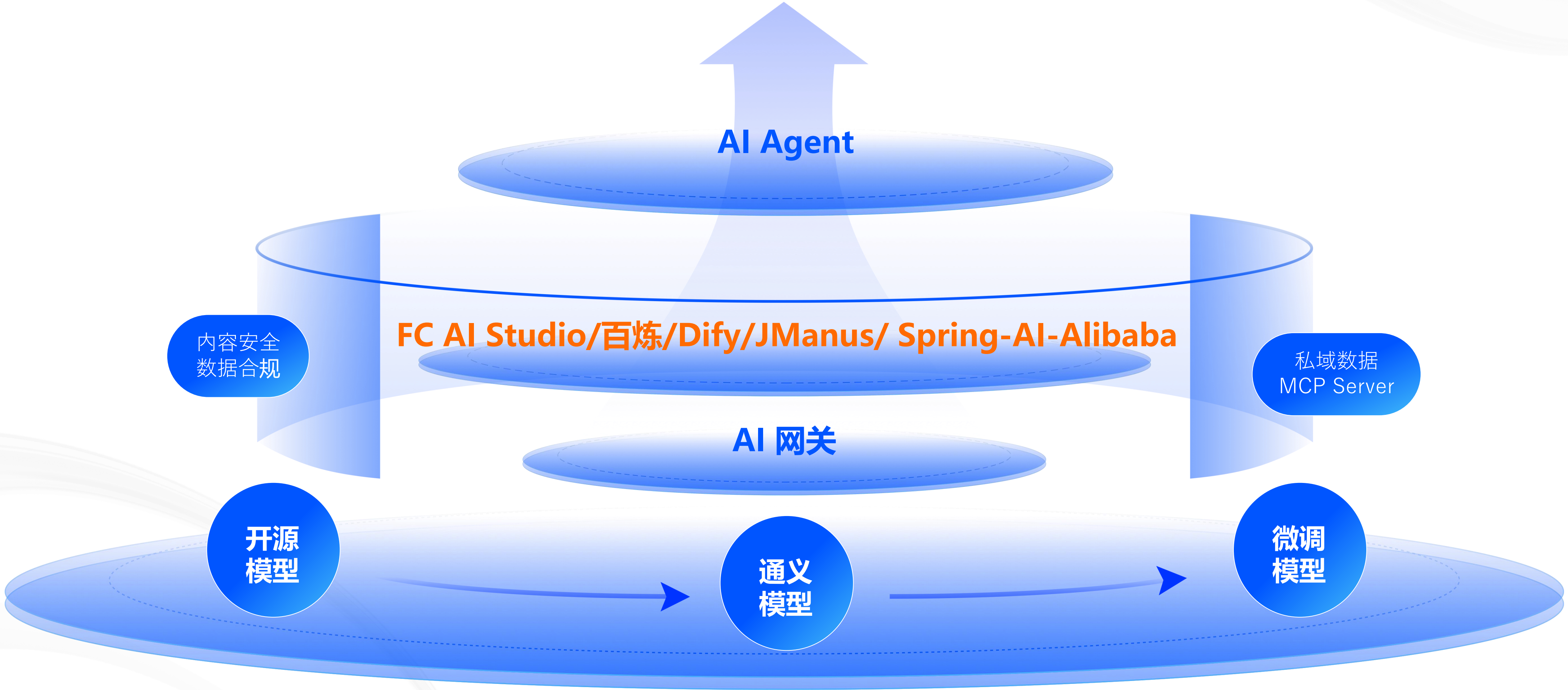
MCP使得添加新数据源或工具变得简单，无需大幅修改现有系统。

安全性

MCP提供结构化的访问模式，内置验证，确保数据交互安全且受控。

AI Agent 实践

打通，挖掘，优化私域数据



Dify Agentic Workflow

可视化构建和编排 AI 应用，让应用快速落地。



支持多模态模型

轻松切换至最优 AI 模型（大语言模型、文本转语音、语音转文本等），完美应对不同任务和业务需求。



精确的记忆管理

确保对话变量有效存储临时数据，在多轮对话中始终维持上下文连贯。



连接、分享和协作

原生支持工作流嵌套，社区共享及团队高效协作。



高容错和高性能兼备

异常处理机制、备用分支和自动重试，保证工作流程持续稳定运行。

Solar Studio TEAM

EXPLORESTUDIOCHATBOT PLAYGROUNDKNOWLEDGETOOLS

PLUGINS

Auto-Saved 21:02:35 · Unpublished

DOC EXTRACTOR

LLM

CODE

LLM

MODEL

GPT-4o

CONTEXT

PROMPT

SYSTEM

USER

MEMORY

Test Run#5

INPUT

RESULT

DETAIL

TRACING

START

DOC EXTRACTOR

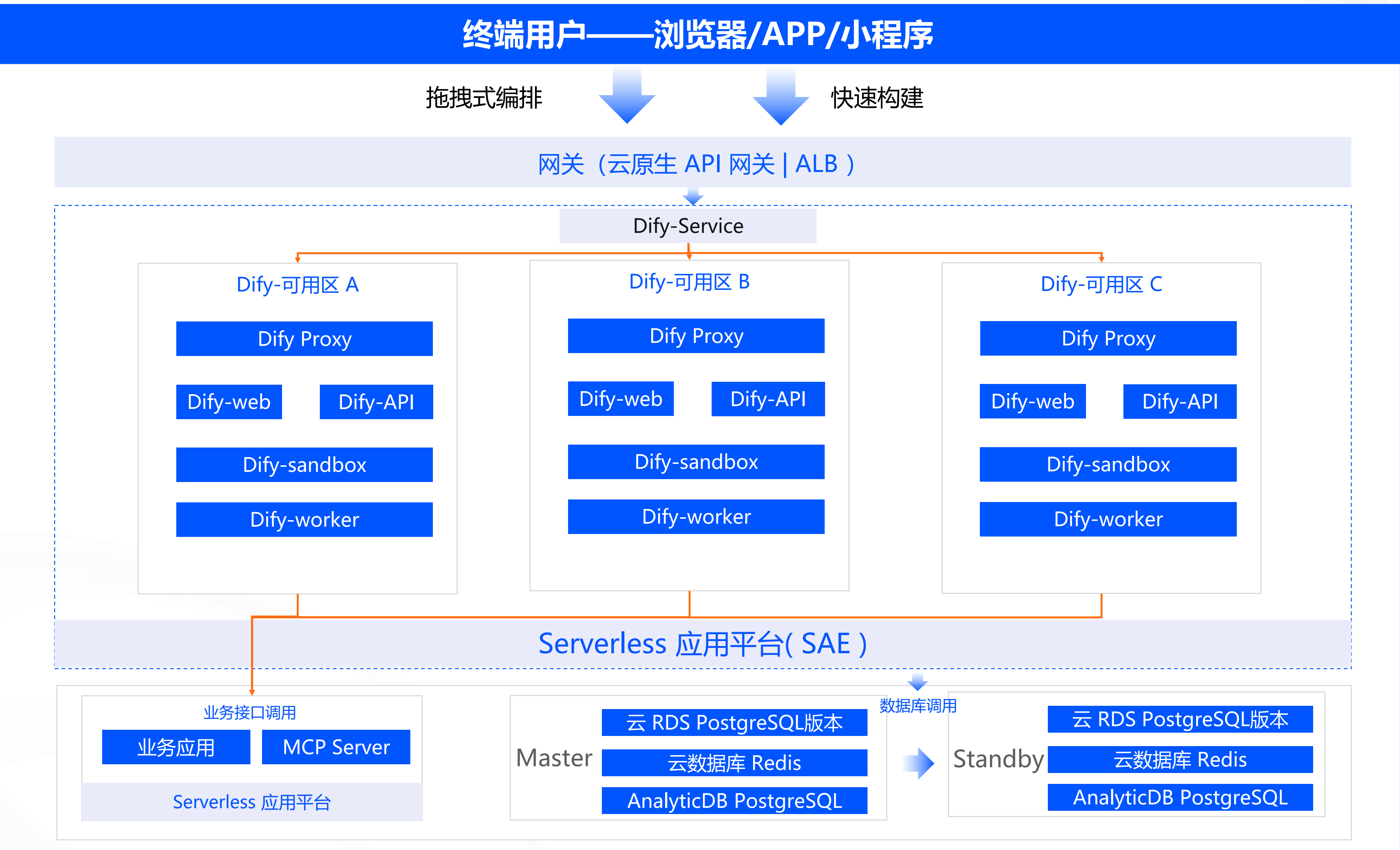
LLM

CODE

INPUT

OUTPUT

一键部署 Dify on SAE —— 支持社区版和商业版



简单易用

- 3 分钟创建 Dify 应用
- 默认集监控日志能力
- 底层资源按需弹缩

高可用

- 默认支持智能化可用区，支持指定三 AZ
- 默认支持负载均衡与健康检查联动，提供无损上下线

成本经济

- 按需按量付费
- 支持多种规格资源，并提供闲时计量模式，夜间成本更低

持续迭代

- Dify 版本更新快，SAE 默认具备灰度发布能力
- AI 应用监控提供 Trace能力，结合日志查询，问题定位更加迅速



百炼智能体和流程编排

百炼智能体和流程编排直接引用阿里云市场积攒多年的成熟 API 服务，助力客户打造 AI 原生应用

智能体/编排应用



工作流应用



插件广场

阿里云百炼
&
阿里云云市场

批量引入
高质量智能体插件

网页阅读

图像与视频

新闻阅读

零售电商

数据搜索

游戏娱乐

金融商业



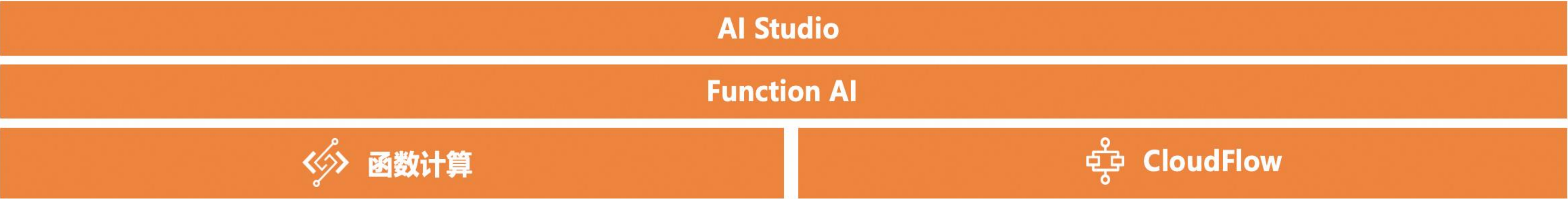
Function AI —— AI Studio 打造高性能低代码 Agent 部署平台

AI Studio 提供类似 Dify 的可视化 Agent 构建平台，同时采用自研工作流执行引擎，10 倍提升 Agent 运行效率

类Dify
可视化流程



自研
流程引擎



Spring AI Alibaba —— 多Agent, 多模态, 多场景

支持 Dify 低代码 DSL 一键转成 SAA 高代码工程方案, 性能提升 5~10 倍

Development Tools

Tools and services to facilitate project initialization and development.

Spring AI Alibaba Studio

Project Initializer

Low-code
workflow

Project
Template

Spring AI Alibaba Studio

RAG
Evaluation

Visualized
Debug

Tracing

Agentic Framework

Framework for creating agent and multi-agent workflows

Spring AI Alibaba Graph

workflow

Agent

ReAct Agent

Multi Agent

.....

Low-level AI Framework

Spring AI based abstractions for developing Chatbot and RAG applications.

Spring AI Alibaba Core

ChatClient

ChatModel

DocumentReader

Embedding

Nacos

LLM Model

Function

TextSplitter

Retriever

Higress

Memory

Prompt

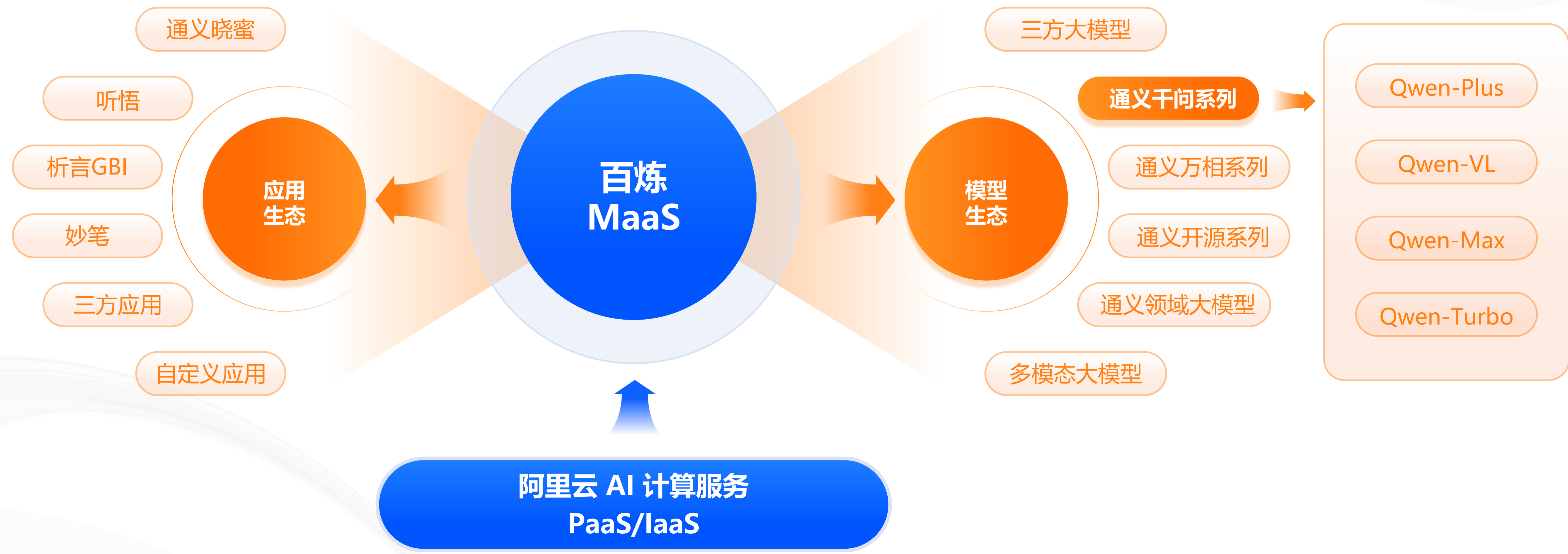
VectorStore

Retriever

Observability

阿里云百炼 —— LLMs API

基于通义大模型，面向企业和开发者客户，打造一站式大模型服务和大模型应用构建平台



函数计算 GPU 极速模式 —— 优化弹性算力成本

业界独家，通过启动 GPU 快照锁定稀缺 GPU 资源并实现毫秒级到秒级首次推理启动，弹性效率比虚拟机/容器快 5~10 倍

函数
计算
FC

FC GPU 聚焦推理场景：

- **开源 LLM 托管**：通义千问系列 0.6B~235B、Deep Seek 系列 1.5B~671B、其他开源 LLM 模型
- **自训模型托管**：OCR、ASR、TTS、图像识别、视频转码、内容审核、3D建模
- **AIGC 生图模型托管** **重磅升级**：托管 SD、ComfyUI 等业界主流生图模型并进行生产级改造，支持前后端分离与 Serverless API，解决开源托管的性能与稳定性问题

FC GPU 弹性模式：

- 标准模式（按请求自动弹性 + 冷启）：秒级冷启动
- **极速模式（按请求自动弹性 + 基于快照启动）** **推荐**：毫秒级快照热启(< 1ms) + 秒级快照温启 (~2s)，**可实现持续锁卡**

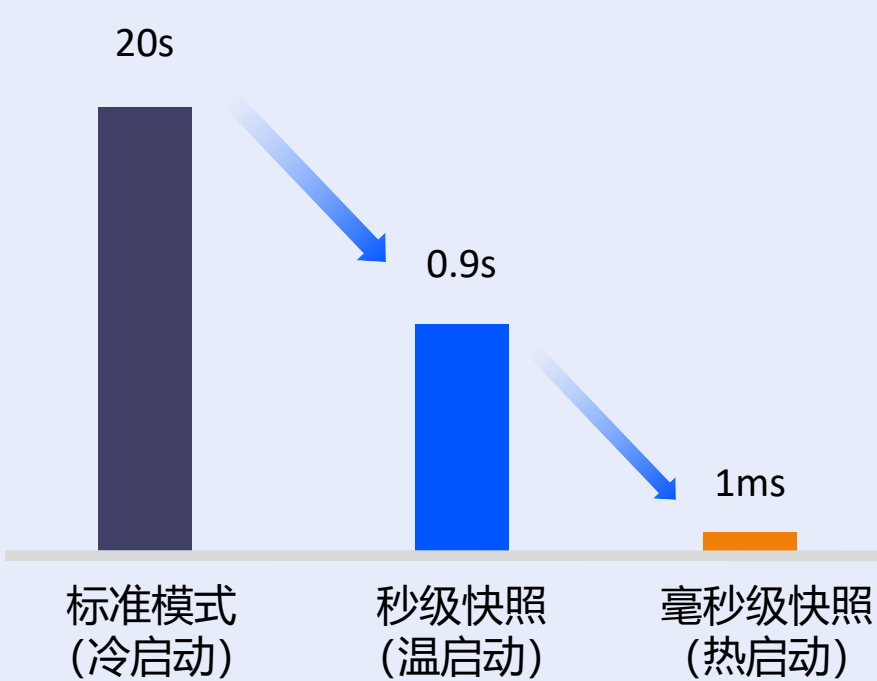
FC GPU 的技术优势

1. **弹得最快**：按请求弹性，通过启动快照可加速至秒级/毫秒级
2. **成本最省**：规格灵活，支持切卡使用；快照技术既可锁卡，又能降本
3. **使用简单**：标准镜像部署，内置调度、负载均衡、高可用、监控、日志等
4. **安全容灾**：3AZ，宕机自动恢复

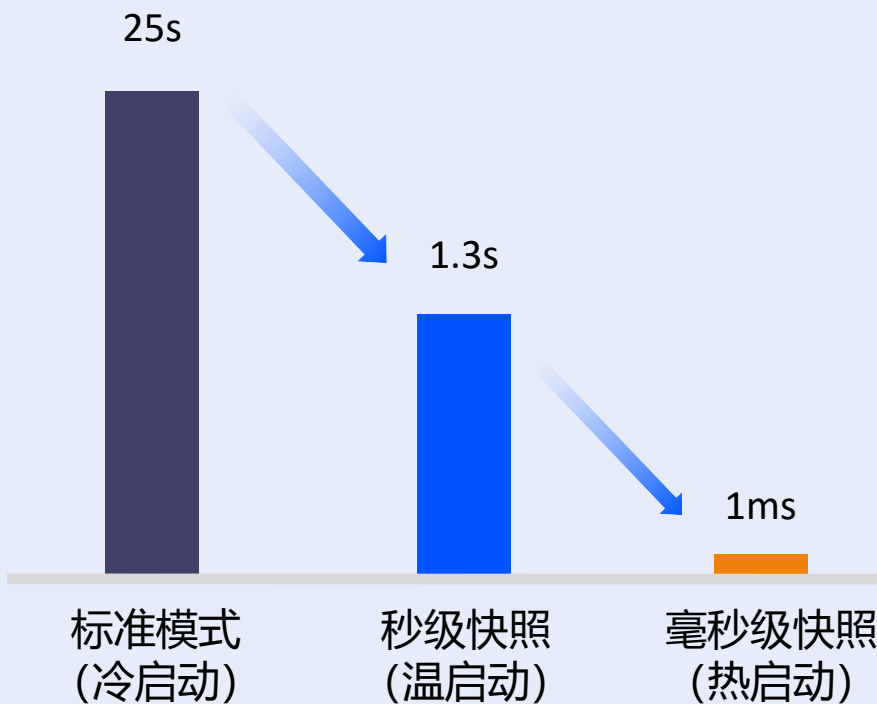
首次推理冷启动耗时分布示意图



SD-v1-5-inpainting (4.27GB) 0->1 TTFI



SD-v1-5-inpainting (4.27GB) 0->1 TTFI





Function AI —— 为 AI 应用开发提效

Function AI 聚焦 AI 应用场景：**模型服务**（魔搭社区/HuggingFace 热门模型托管）；
图像生成（SD、ComfyUI Serverless API）；**MCP & Agent**（企业专属 MCP 市场）

Function AI 一键创建应用

Function AI / 探索

返回函数计算控制台 钉钉答疑群 帮助

Serverless + AI，让应用开发更简单！

Function AI 是基于函数计算构建的 Serverless AI 应用开发平台，内置极致的弹性算力、探索丰富的 AI 应用模板、开箱即用的 AI 工程能力与应用的全生命周期管理。让技术不再是束缚，让AI化作创新的画笔，即刻体验，开启您的 Serverless AI 之旅！

筛选应用模板

☐ 热门模版

☐ 人工智能

☐ MCP Server

☐ 音视频处理

☐ Web 开发框架

☐ 文件处理

☐ Web 应用

☐ 游戏

☐ 流程式开发

搜索应用模板

基于 Qwen-QwQ 推理模型构建 AI 聊天助手

官方 AI Ollama Open-WebUI

使用 Qwen-QwQ 系列模型，部署 Ollama 与 Open-WebUI 到阿里云云应用开发平台

2025.03.06 更新 272

基于 DeepSeek-R1 构建 AI 聊天助手

社区 大模型 AI Ollama Open-WebUI

部署 Ollama 与 Open-WebUI 到阿里云云应用开发平台, 使用 DeepSeek-R1-Distill-Qwen 系列模型

2025.03.04 更新 1055

基于 DeepSeek-R1 构建 RAG 应用

社区 大模型 AI 检索增强生成 模板

CAP rag with deeepseekr1

2025.03.16 更新 173

AI 剧本生成与动画创作

官方 AI 大模型 阿里云技术内容

AI 剧本生成与动画创作示例工程

2025.03.04 更新 707

组装式开发，弹性开放，按需选择

模型服务
(Qwen, DeepSeek, etc.)

图像生成
(SD、ComfyUI)

MCP 服务
(STDIO、SSE、Streamable HTTP)

Agent 服务
(Dify、RAG、Bolt.diy, Open Manus, etc.)

无缝升级 AI 应用开发范式

应用开发

应用市场

AI 运行时内置 vLLM, SGLang, Ollama, HuggingFace 推理框架

海量 Serverless AI 应用模板，开箱即用，支持二次开发

MCP 运行时内置 MCP 开发框架, STDIO 零改造升级 SSE

魔搭社区、百炼官方集成，开源模型/智能体应用增强

AI 网关 —— AI 应用流量入口与集成利器

AI 网关支持多模型协议代理、丰富的插件策略、Rest to MCP 一键转换、完善的可观测服务等



函数计算 MCP 运行时 —— 加速 AI 应用落地最后一公里

函数计算联合阿里云百炼发布**业界首个全生命周期 MCP 服务市场**，函数计算联合魔搭社区发布**最大 MCP 中文社区**

Function AI 一键创建 AI 应用

痛点

Promote: 调用不稳定
Function Calling: 重复开发
Plugin: 仅支持 HTTP 远程调用

MCP

AI 领域的
“USB-C接口”

优势

稳定调用: 弱化提示词依赖
开放标准: 无需适配, 直接使用
调用灵活: Studio/SSE/
Streamable HTTP

函数计算 FC MCP 运行时

内置开发框架

一键部署免运维

自动弹性低成本



百炼 MCP 市场

全生命周期 MCP 服务:

- MCP 广场: 官方 MCP 服务
- MCP 管理: 自定义 MCP 服务
- MCP 使用: Agent/工作流调用



魔搭社区 MCP 广场

全生命周期 MCP 服务:

- MCP 广场: 1500 种 MCP 服务
- MCP 实验场: 提供工具和调试环境
- 降低门槛: 云上托管/本地部署/三方集成



函数计算 Function AI

Function AI MCP 市场

开源托管 (STDIO 零改造升级 SSE) , 开放使用 (注册到百炼或自建模型/Agent调用) :

- 主流云产品官方 MCP 托管: ECS, OSS, RDS, PolarDB, OTS, Hologress, DataWorks 等
- 企业特性增强: 加持 FC 弹性降本优势, 不使用自动缩 0, 原子化易管理, 省 token

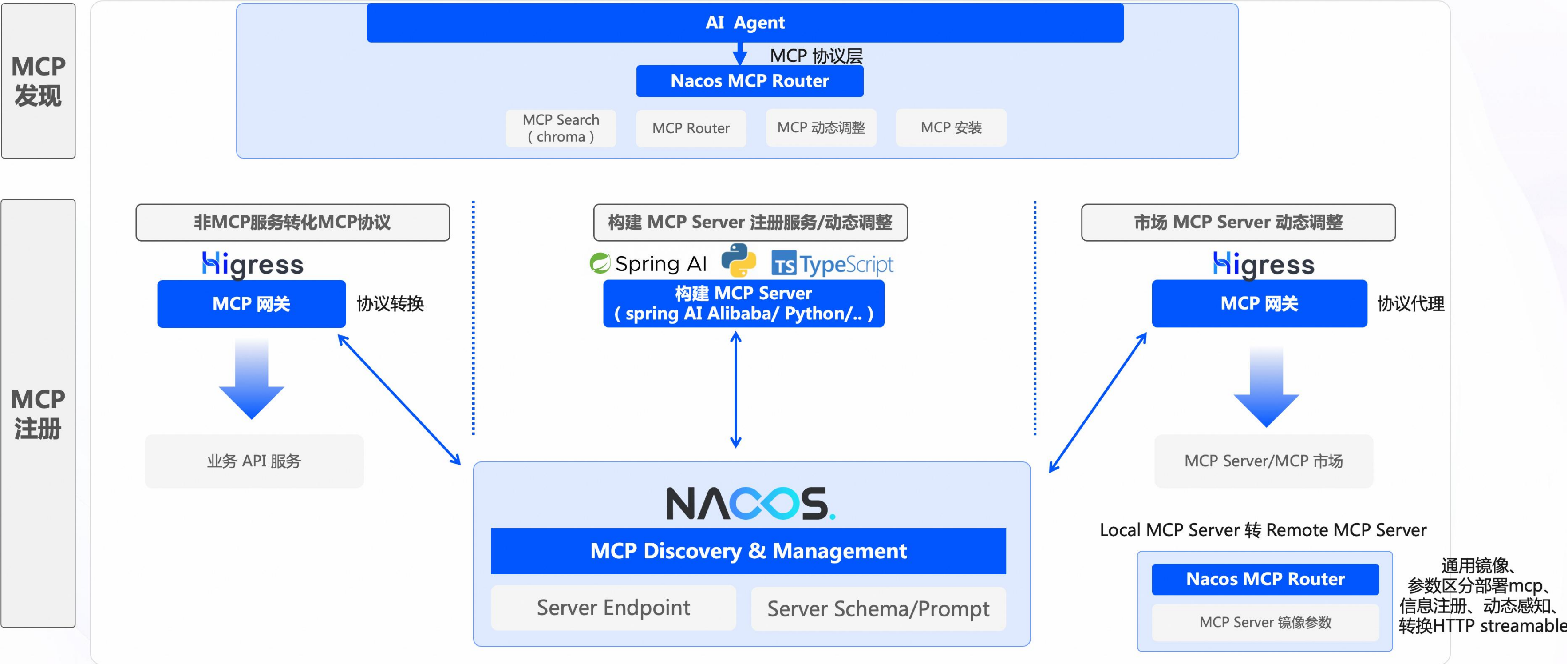


MSE Nacos 企业版 —— 构建企业私有 MCP Registry

- ✓ **MCP Registry:** 搭配 AI 网关，支持 Rest API 直接转换成 MCP Server
- ✓ **MCP Router:** 支持过滤和筛选所需的 MCP 服务避免大范围传入 MCP 服务列表，减少Token消耗。
- ✓ **MCP 服务自动发现:** 配合 Spring AI Alibaba 等应用框架自动注册发现 MCP/Agent 服务

MCP 服务管理中心，适配MCP协议，接入AI模型调用

Nacos 3.0架构作为MCP 服务管理中心，进行信息筛选帮用户安全管理大规模MCP Server



AI 应用可观测体系

为 GenAI 应用可观测而生

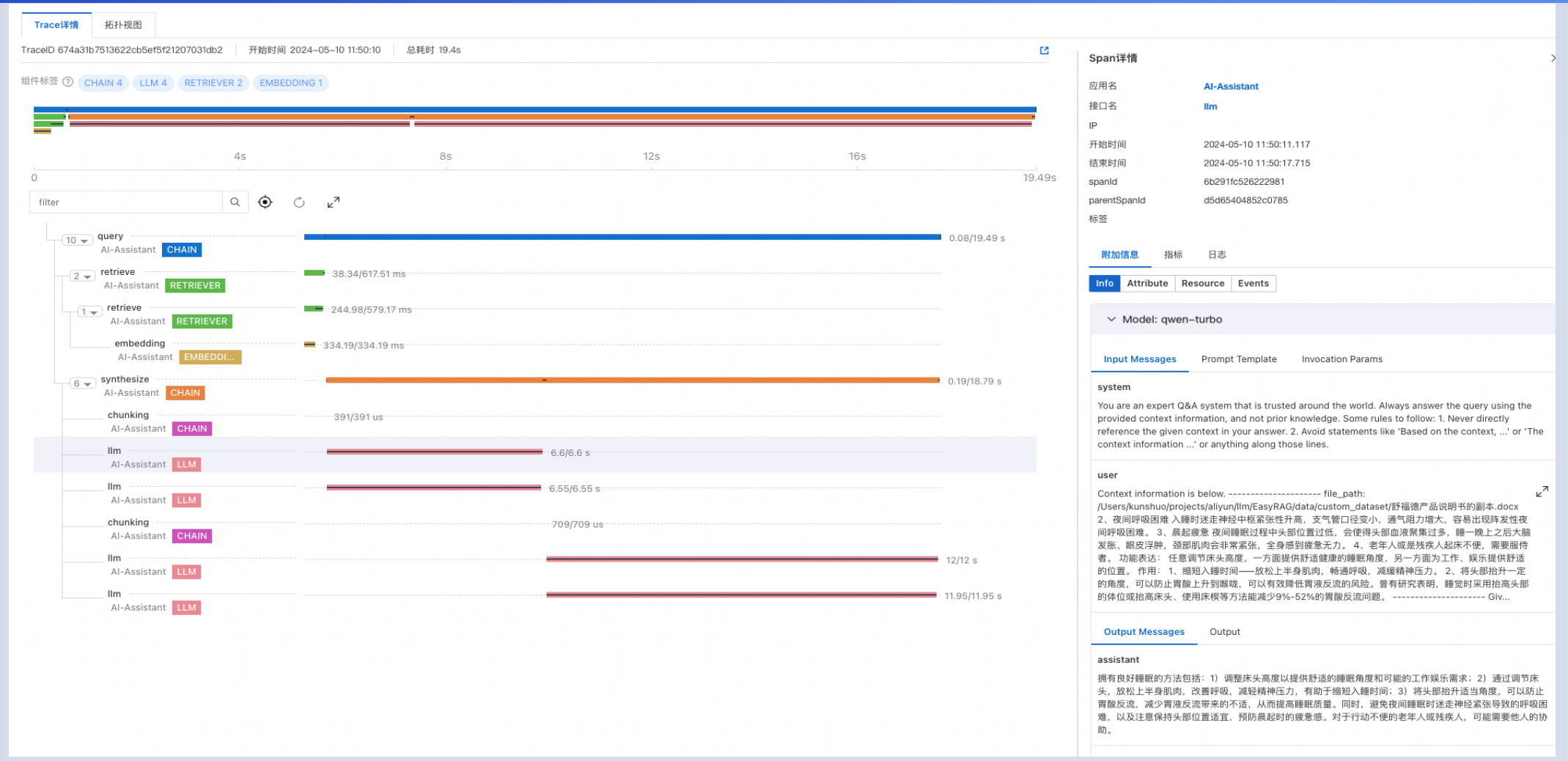
阿里云 ARMS	可观测链路追踪 Open Telemetry 版
 LlamaIndex  LangChain	 Dify  Spring AI Alibaba
 Open AI	 通义千问

Open Telemetry GenAI 语义约定

持续剖析	稳定性	LLM SDK
阿里云 OTel Python 发行版	阿里云 OTel Java 发行版	阿里云 Go 探针

- 遵循最新 OpenTelemetry 社区 GenAI 语义约定。
- 支持常见的AI框架和AI模型，包括 Spring AI Alibaba / LlamaIndex / Langchain / 通义千问 / OpenAI / PromptFlow等。
- 相比社区规范提供更加精细化的埋点和属性。
- 支持在不同的调用链中传播会话信息。

大模型应用专属分析视图



- RAG 过程观测
- 提示词输入、输出观测
- Token 消耗观测

一个典型的 AI 原生应用全链路可观测

AI 全栈统一监控

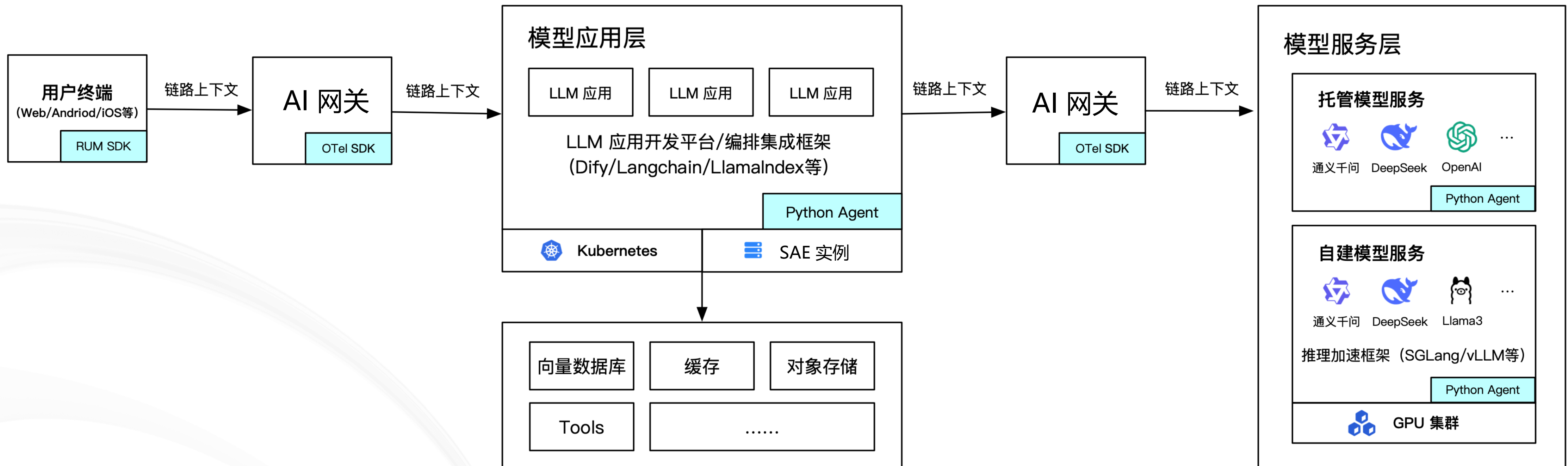
- 基于 Prometheus 实现前后端应用、网关、模型服务与基础设施等全栈统一监控
- 构建 LLM 领域专属视图，包括模型性能分析、Token 成本分析、GPU资源异动分析等

模型调用端到端链路追踪

- 用户终端 (Web/App/小程序等)
- 网关 (API 网关、AI 网关等)
- 后端应用 (模型应用/编排框架等)
- 依赖组件 (缓存、数据库、Tools等)
- 模型服务 (开源自建、三方服务等)

模型日志存储与评估分析

- 模型调用日志存储
- 模型调用结果评估 (质量、安全等)
- 统一日志分析平台



Contents

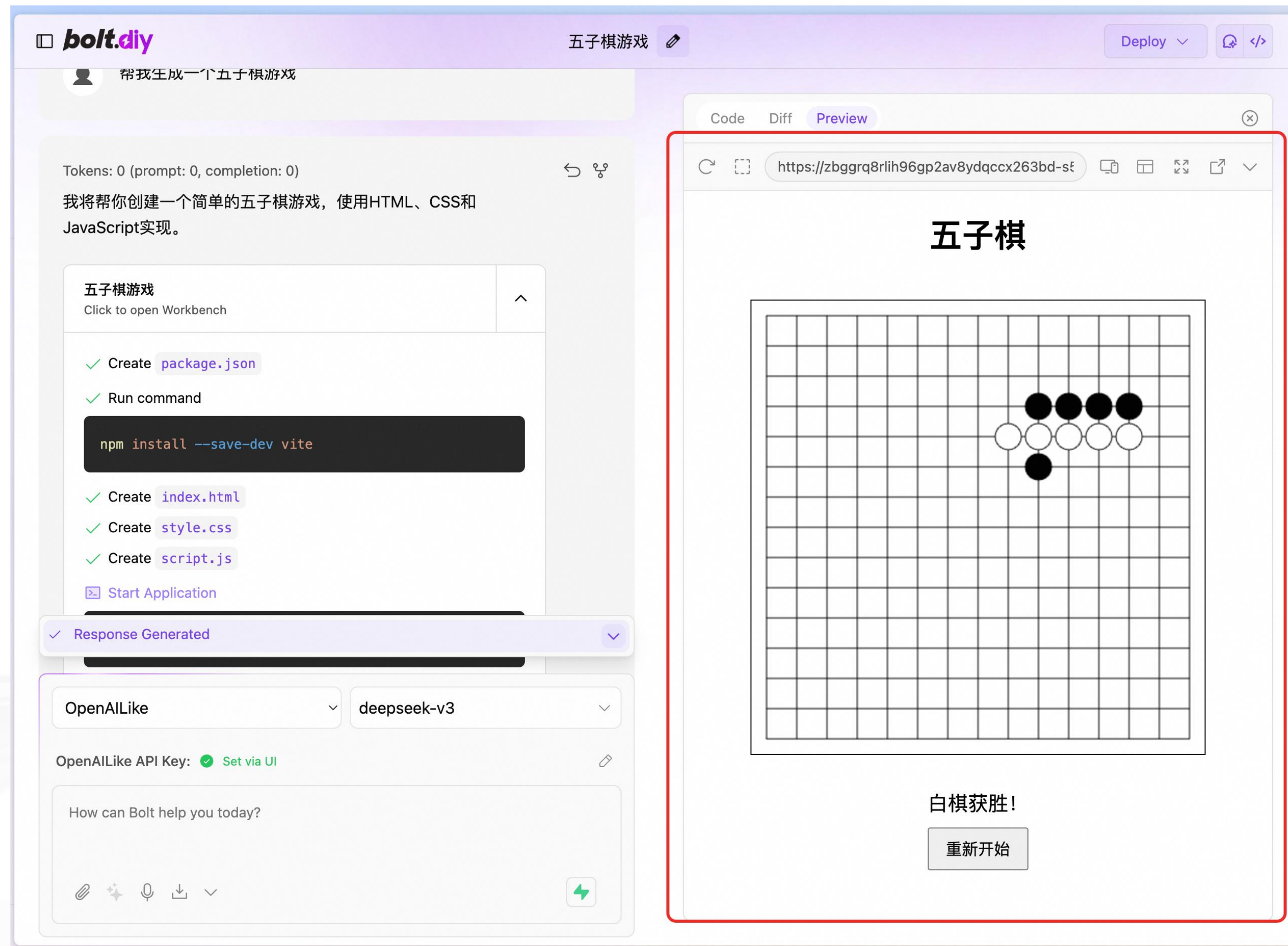
目录

01 从云原生应用到 AI 原生应用

02 MCP 加速推动 AI Agent 应用落地

03 AI Agent Sandbox

函数计算为自研 AI Agent 提供安全高效的 Sandbox 运行环境



SAE

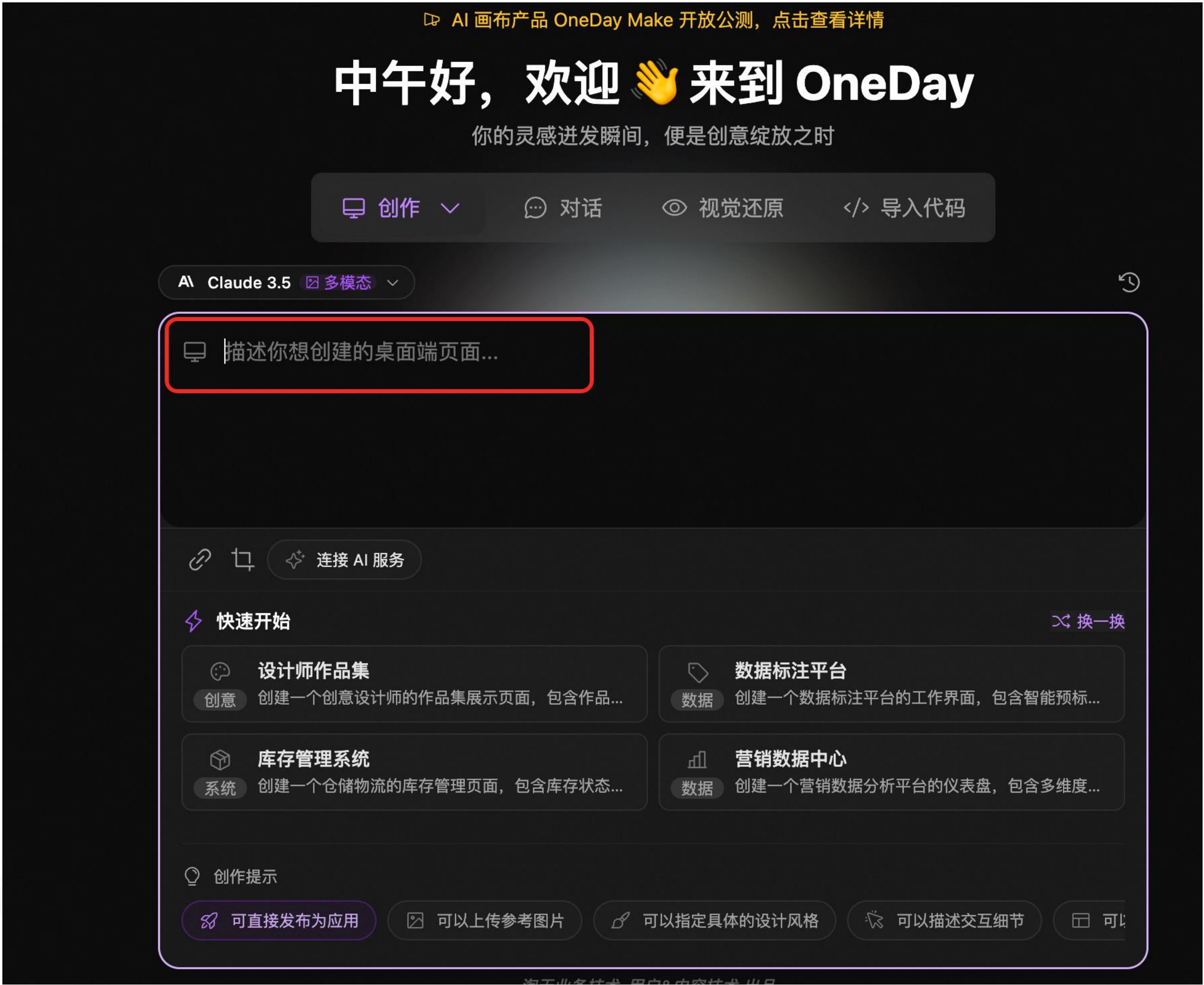
- 自研 Agent 的 long run 运行环境
- 支持运行开源 Agent 框架，如 Spring AI Alibaba、LangChain、LlamaIndex 等



FC

- **Agent Sandbox Runtime:** 如 Code Sandbox、Browser Use Sandbox、仿真训练 Sandbox、RL Sandbox
- **多租强隔离:** 安全容器+动态存储挂载，为每个启动的 Agent 提供干净独占的运行环境
- **成本最优:** 按请求扩缩，提升 AI Agent 资源利用率，降低资使用源成本
- **启动速度快:** CPU 百毫秒弹性伸缩
- **内置完善的监控日志能力**

阿里内部 Agent OneDay 落地效果



THANKS



函数计算 FC – AI 时代最佳运行时

墨颀

函数计算产品经理



Contents

目录

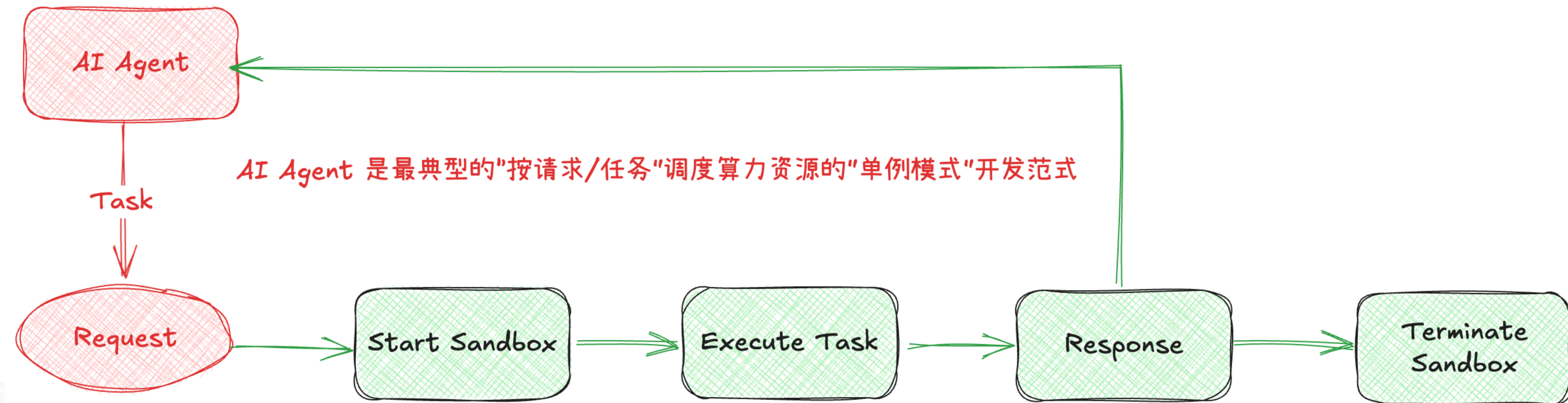
01 AI 时代运行时关键需求

02 AI 时代运行时关键技术

03 AI 时代运行时最佳实践

AI Sandbox Runtime

Deploy AI Agents Fast with **Secure Cloud Sandboxes**



AI 时代运行时关键需求 - 核心场景

AI Sandbox Runtime 是典型的 AI 时代运行时，开发者可以是人类，也可以是 AI，两者场景几乎一致

生成代码

生成代码、运行代码、分析数据、生成UI等



执行命令

执行任务、依赖管理、文件操作、分析诊断等

控制电脑

浏览网页、表单填写、操作软件、内容提取等

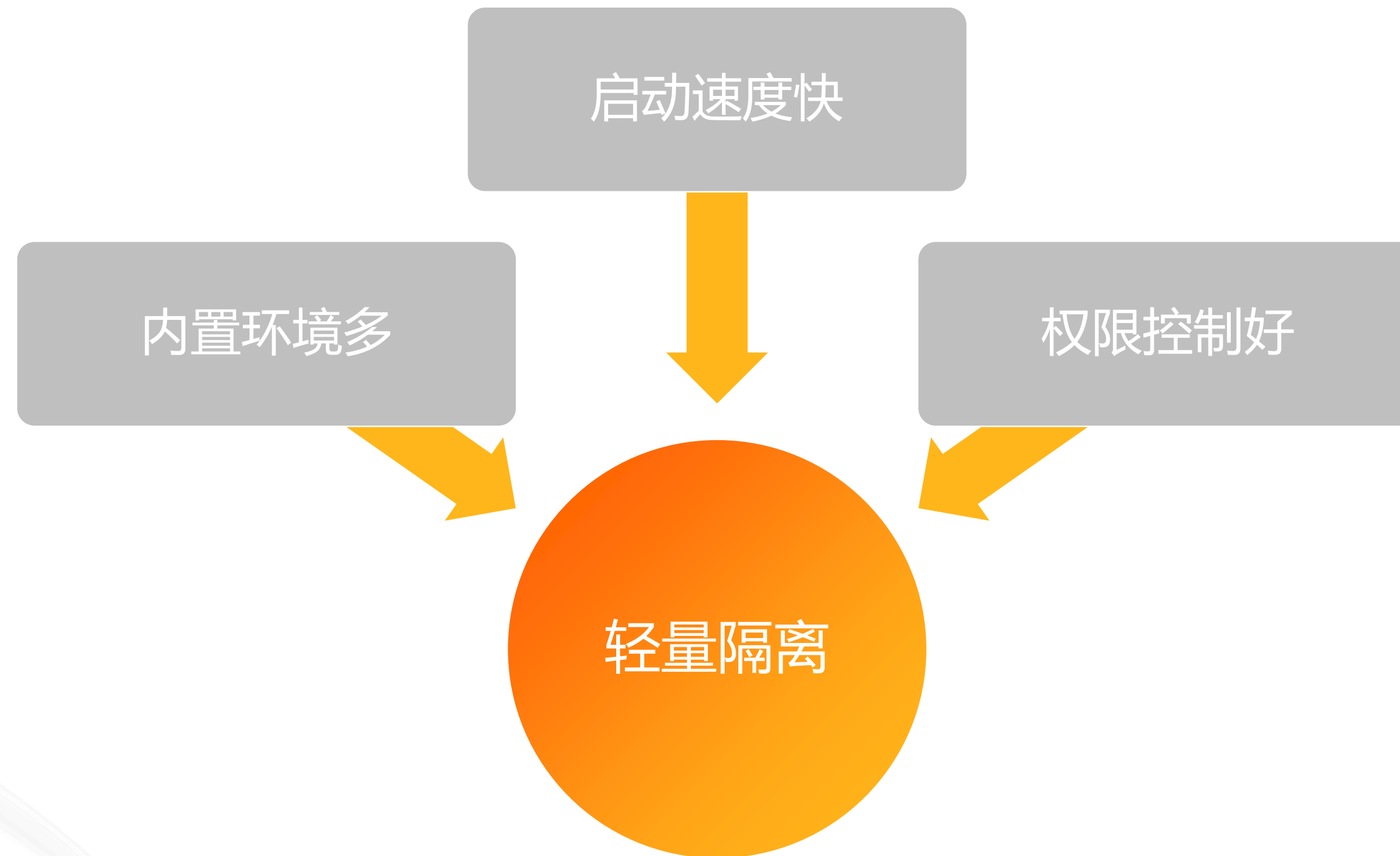


强化学习

有交互性、目标导向、根据环境反馈调整行为等

AI 时代运行时关键需求 – 需求分析

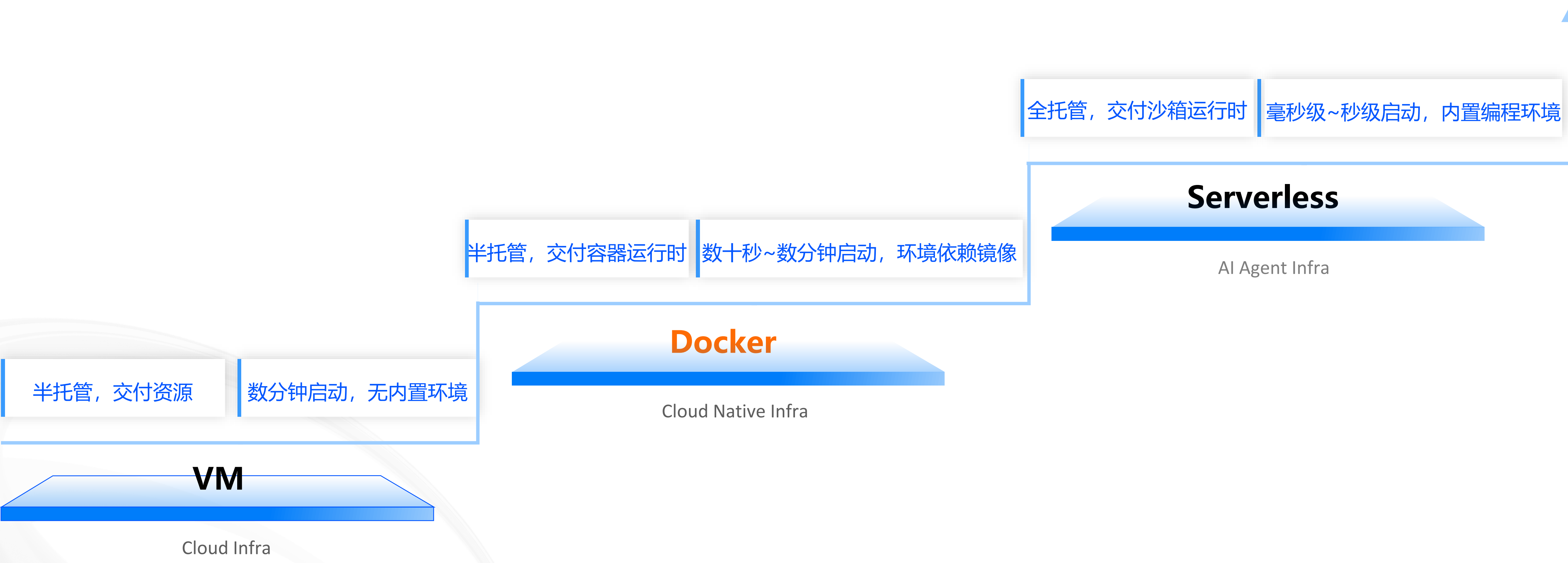
AI Agent 痛点：行为不可预测、易受注入攻击、权限边界模糊





AI 时代运行时关键需求 – 技术选型

开发环境正在从本地迁移到云端，AI 时代运行时首先得是 Serverless 运行时





AI 时代运行时关键需求 – 功能矩阵

函数计算 FC 为 AI 时代运行时提供**弹性算力**和**应用开发**工具套件，助力 Serverless + AI 开发更简单



X 个业务场景

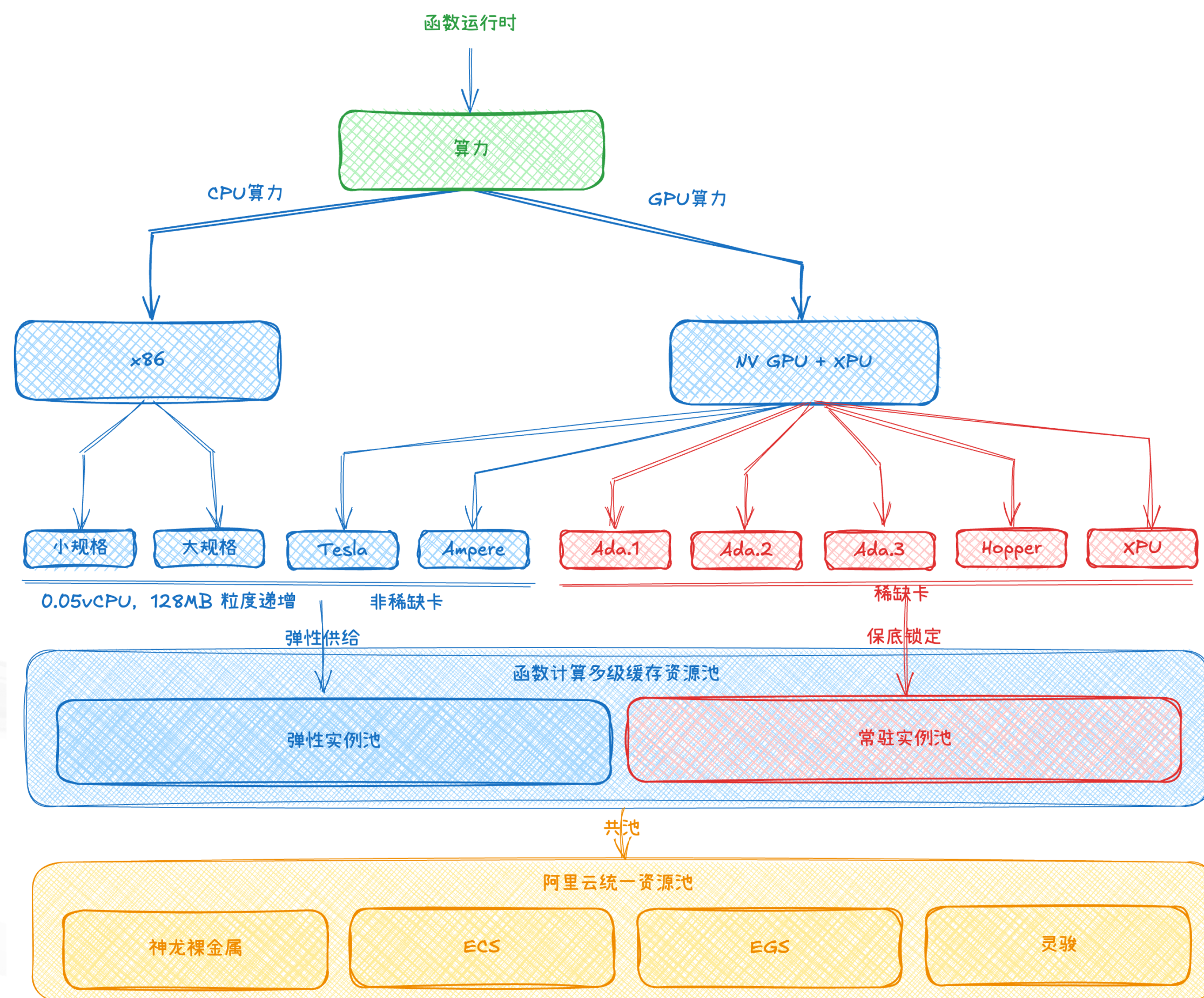
50+ 运行环境/100+产品集成

4 大底座核心能力

N 种异构算力资源池

AI 时代运行时关键技术 - 算力类型

函数计算 FC 为 AI 时代运行时提供**云上最小规格、最丰富推理卡型、最细计费粒度**的全托管 Serverless 异构算力



5倍

规格粒度

最小规格：FC 0.05C128MB，
虚机 1C512MB，容器 0.25C512MB

100倍

启动效率

冷启动速度：FC 毫秒~秒级，
虚机数分钟，容器 30+秒~数分钟

0

不使用不计费

FC 不删除元数据，供给算力即服
务，释放算力即不再计费

元数据：指代码和相关配置

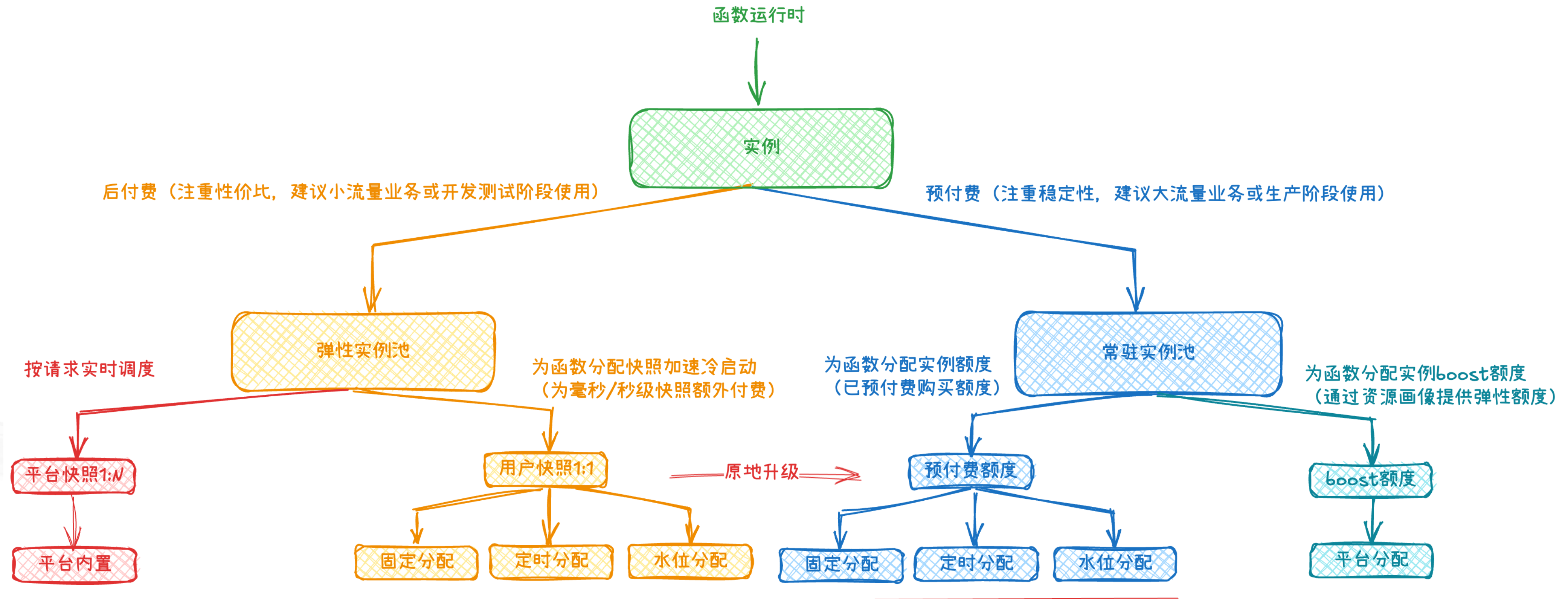
100%

利用率

主动调度，1毫秒计费（GPU 1秒计
费），云上唯一 100% 利用率的计算
类产品（业务进程启动后再计费）

AI 时代运行时关键技术 - 调度策略

函数计算 FC 提供弹性实例（**按请求实时调度**、**按快照加速冷启动**）、常驻实例（**按预付费额度分配实例**）2种调度策略



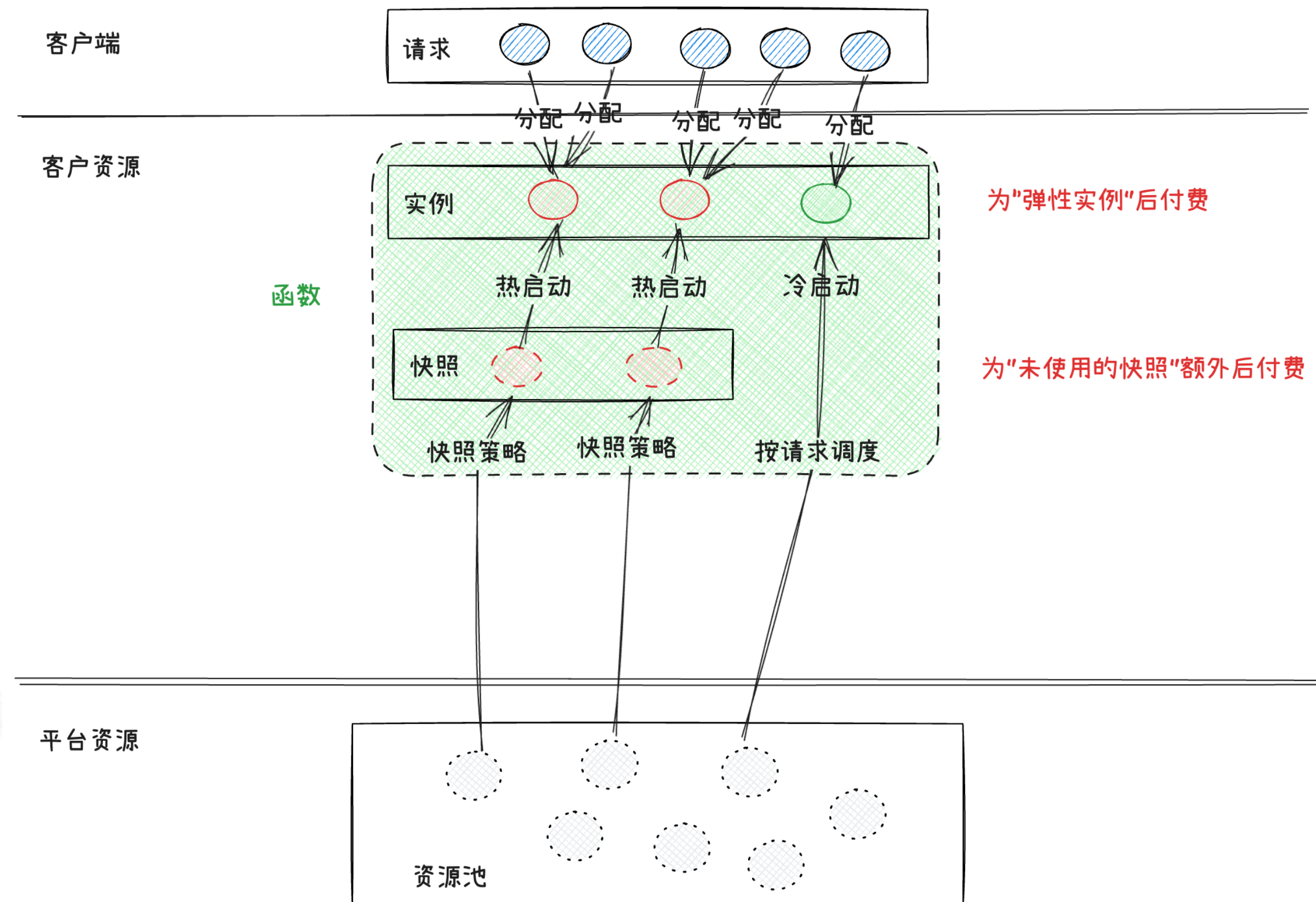
快照分配策略可原地升级为实例分配策略，从后付费无缝切到预付费

AI 时代运行时关键技术 – 弹性实例

函数计算 FC 按请求调度实例，实例**冷启动**（100%利用率）或通过**快照热启动**（额外付费加速启动），为实例+快照付费

弹性实例（按请求实时调度、按快照加速冷启动）

示例：某业务**时延敏感**且关注**性价比**，因此配置了2个快照+单实例2并发，前4个并发优先调度到**基于快照加速启动的实例**（启动时间秒级/毫秒级 - 秒级快照/毫秒级快照），第5个并发因没有可用的快照加速启动，因此要**通过冷启动弹性供给的实例**。



请求



实例



快照 (可选, 加速冷启动)



资源池

AI 时代运行时关键技术 – 快照

函数计算 FC 业界独家技术优势，通过快照**加速冷启动**的同时，**大幅降低资源闲置/锁定成本**

免费

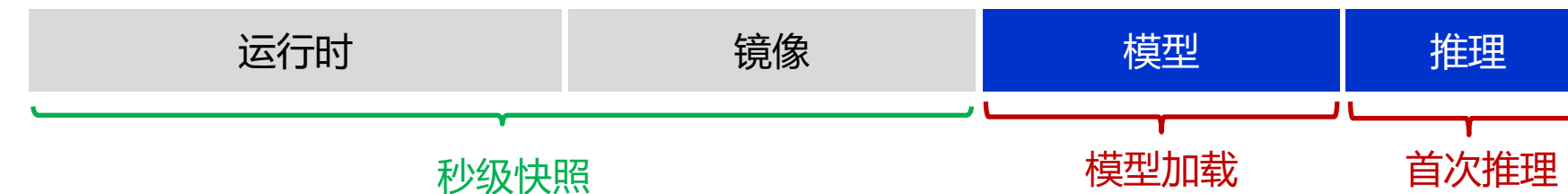
收费

首次推理冷启动耗时分布示意图

平台快照
(冷启动)



秒级快照
(温启动)



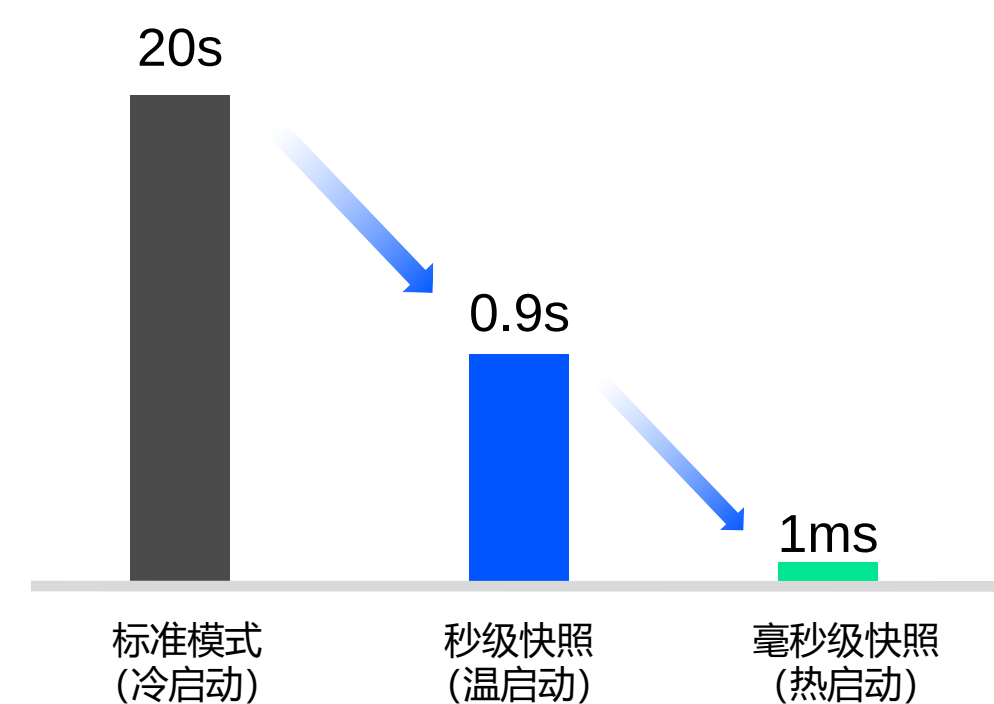
毫秒级快照
(热启动)



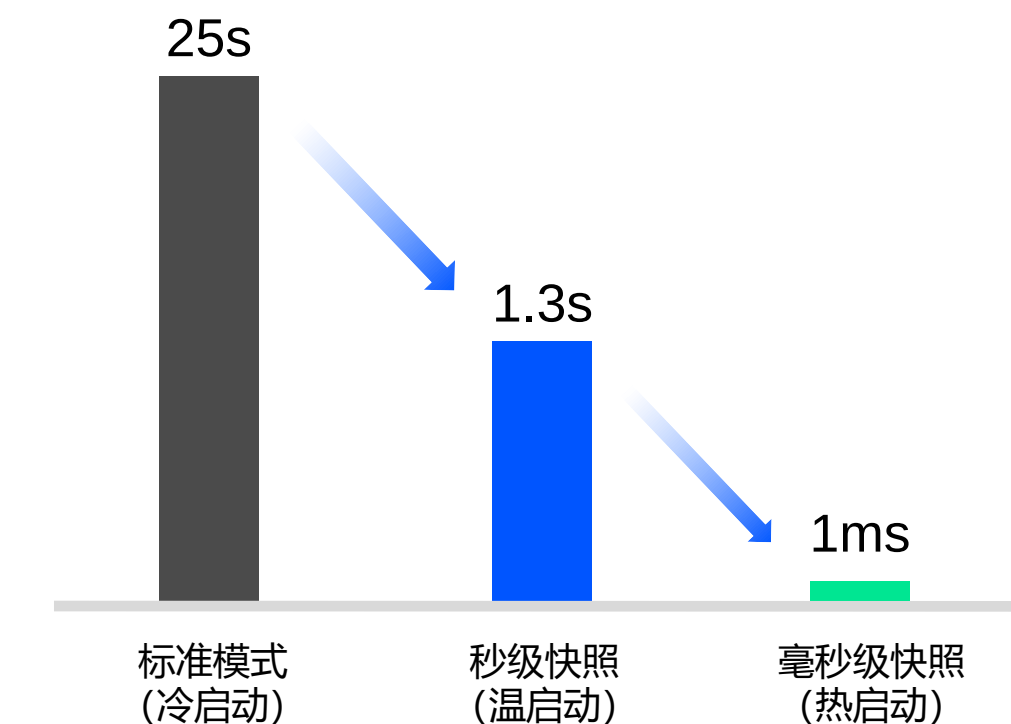
快照分类和收费策略

1. **平台快照**：如 Python Runtime，快照内置 Debian+Python 编程环境，开箱即用
2. **秒级快照**：如 Qwen，秒级快照内置 Debian+CUDA+Qwen，模型未加载到显存
3. **毫秒级快照**：如 Qwen，毫秒级快照内置 Debian+CUDA+Qwen，模型已加载到显存
4. **快照收费**：平台快照 1:N 不计费，**秒级/毫秒级快照 1:1 按标准价格 15%/20% 计费**

SD-v1-5-inpainting (4.27GB) 0->1 TTFI



Qwen-14B-Chat-Int4 (9.01GB) 0->1 TTFI



FC GPU 的客户价值

客户案例

Rokid
AR 时代引领者

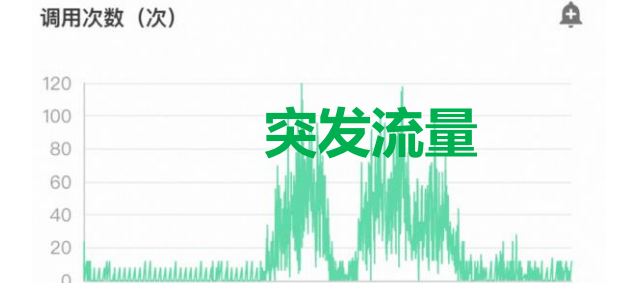
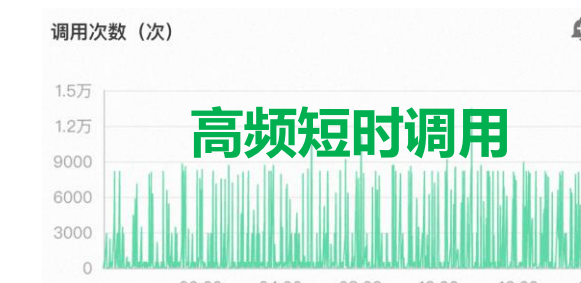
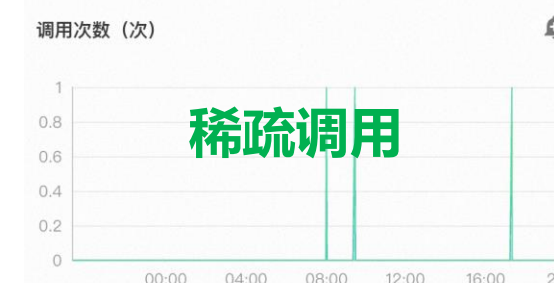
业务痛点

时延极度敏感 高峰期流量大 算力浪费较多

使用价值

+50% 建模效率
-40% GPU 算力成本

FC GPU 适用于三类典型场景



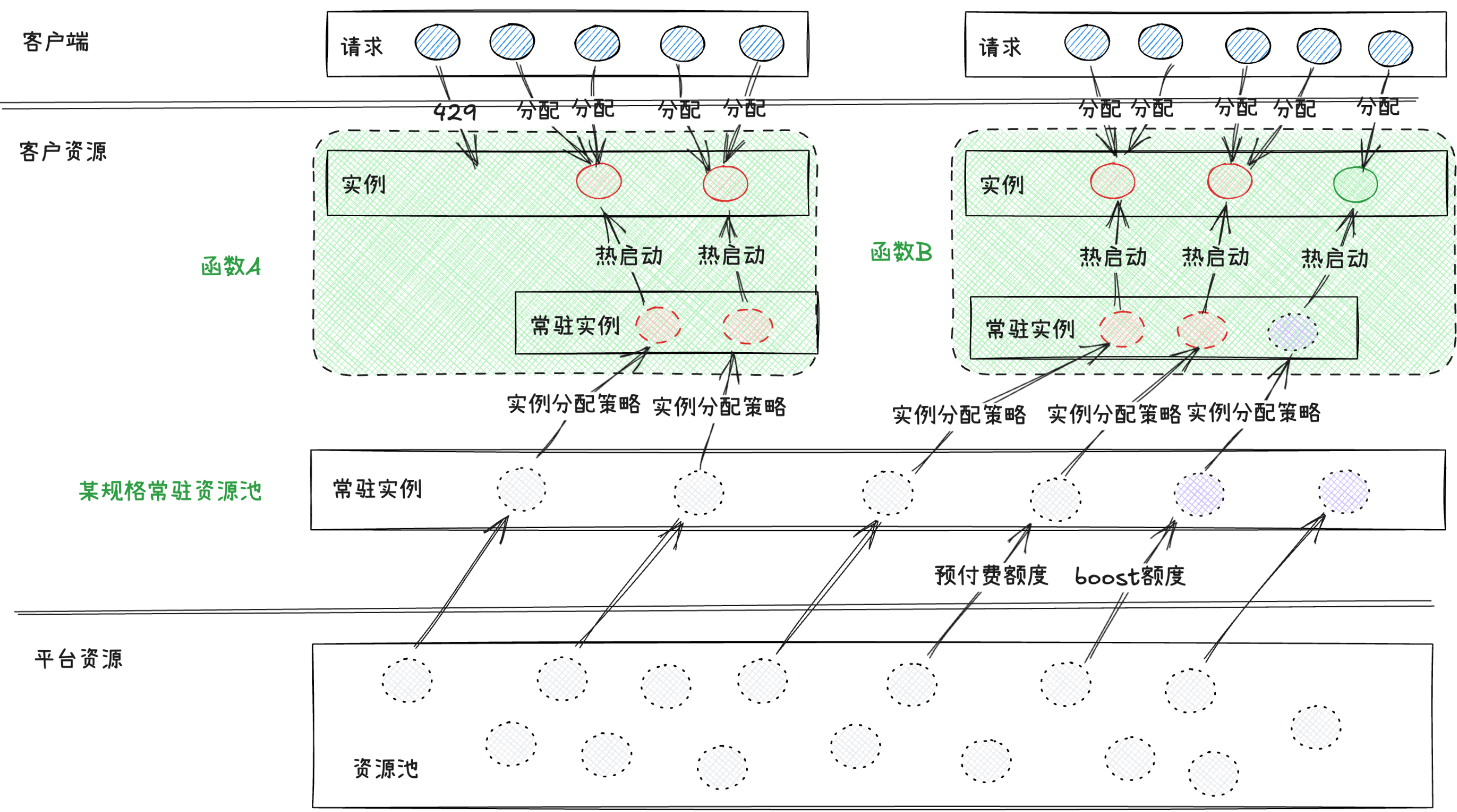


AI 时代运行时关键技术 – 常驻实例

函数计算 FC 按请求调度实例，实例分配自**预付费常驻资源池**，为资源池付费，且以 FC 功能特性最大化使用资源池

常驻实例（为函数分配预付费额度）

示例：某业务期望**固定预算**业务流量极高，因此预购 4 卡 Ada.1（根据资源画像，该用户某些时段有 2 卡 boost 额度），**分配实例额度**：函数 A 2 卡额度，函数 B 3 卡额度，单实例 2 并发，则函数 A 和函数 B 的前 4 个并发均调度给常驻实例，函数 A 的第 5 个并发将 429（**超出额度自动触发限流**），而函数 B 的第 5 个并发则正常响应（**boost 额度满足了突发流量弹性需求**）。



请求

实例

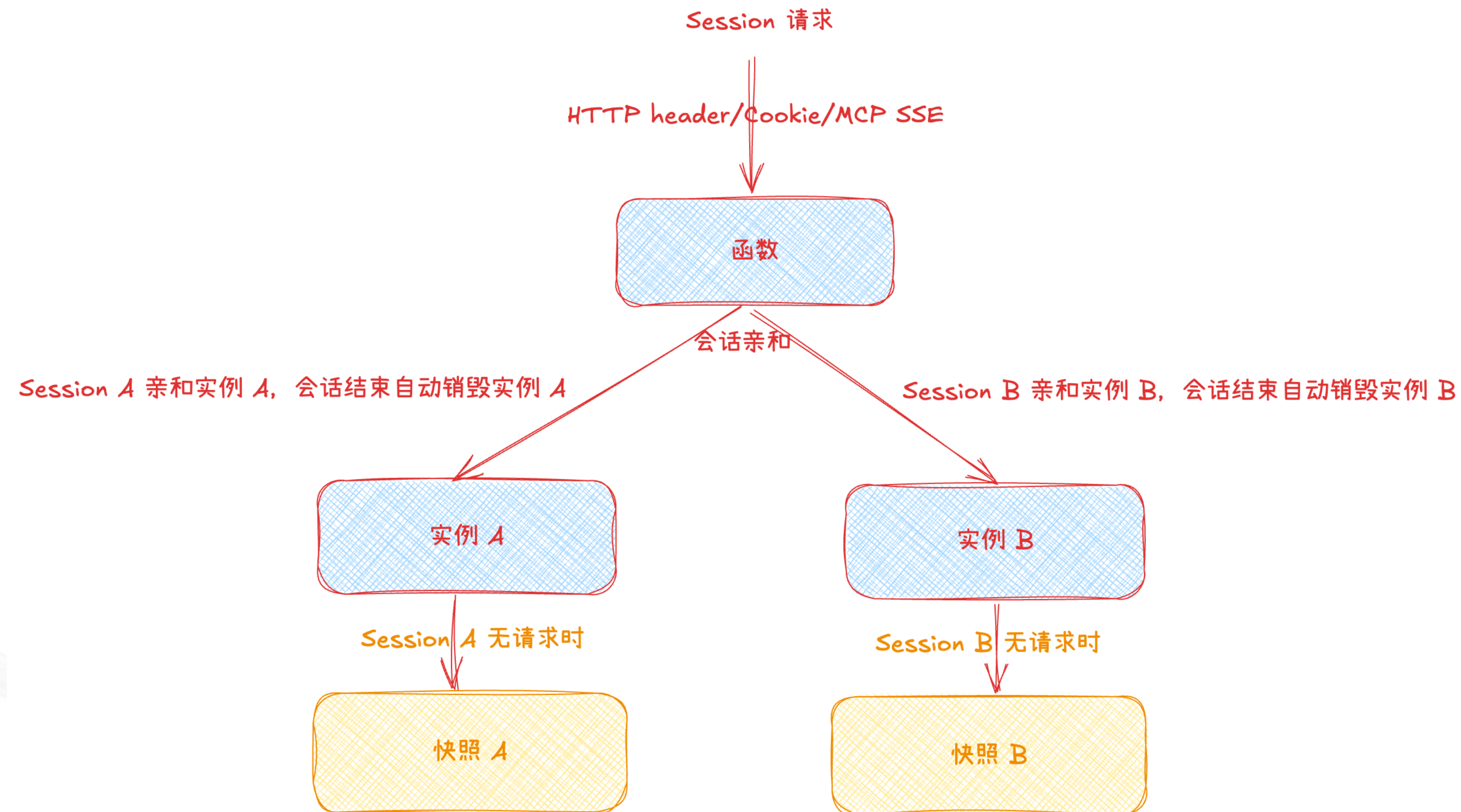
为“常驻资源池”预付费（包月），根据资源画像提供一定boost额度

常驻资源池（可选，包月预付费）

资源池

AI 时代运行时关键技术 – 安全隔离

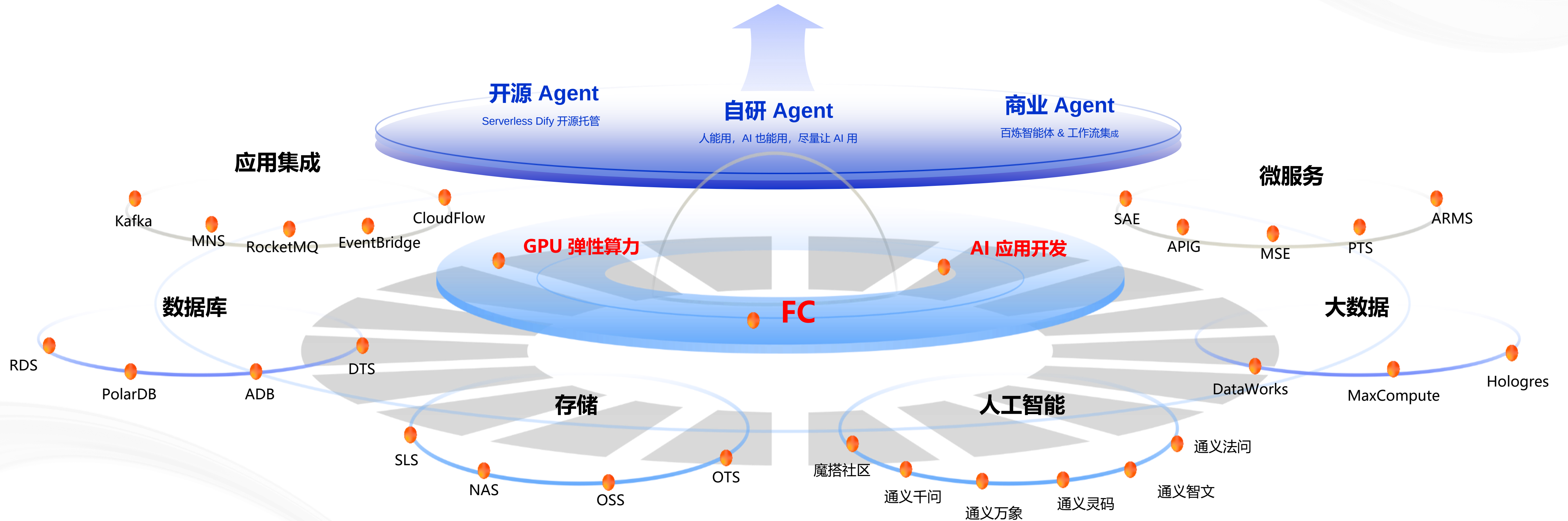
函数计算 FC 业内独家支持会话亲和、长连场景的 Sandbox 安全隔离，AI Agent 上下文不残留



会话亲和、长连接等场景将自动启用快照实现降本

AI 时代运行时关键技术 – 集成生态

函数计算 FC 联动阿里云产品生态，打通 AI Agent + 计算 + 存储 + 网络 + 数据库 + 大数据完整上下游闭环
智能体世界



事件驱动

100+云产品事件驱动，同步/异步数据处理自动化

互联互通

集成到 OSS、NAS、MQ、百炼等作为官方功能，VPC互通更安全

流程编排

流程式编程，可视化编排，易于沉淀业务资产和复用公共模块



AI 时代运行时最佳实践 – 模型服务

函数计算 Function AI 提供 SD/ComfyUI、DeepSeek/Qwen 等 AI 应用样板间，深度集成通义产品系列、魔搭社区等

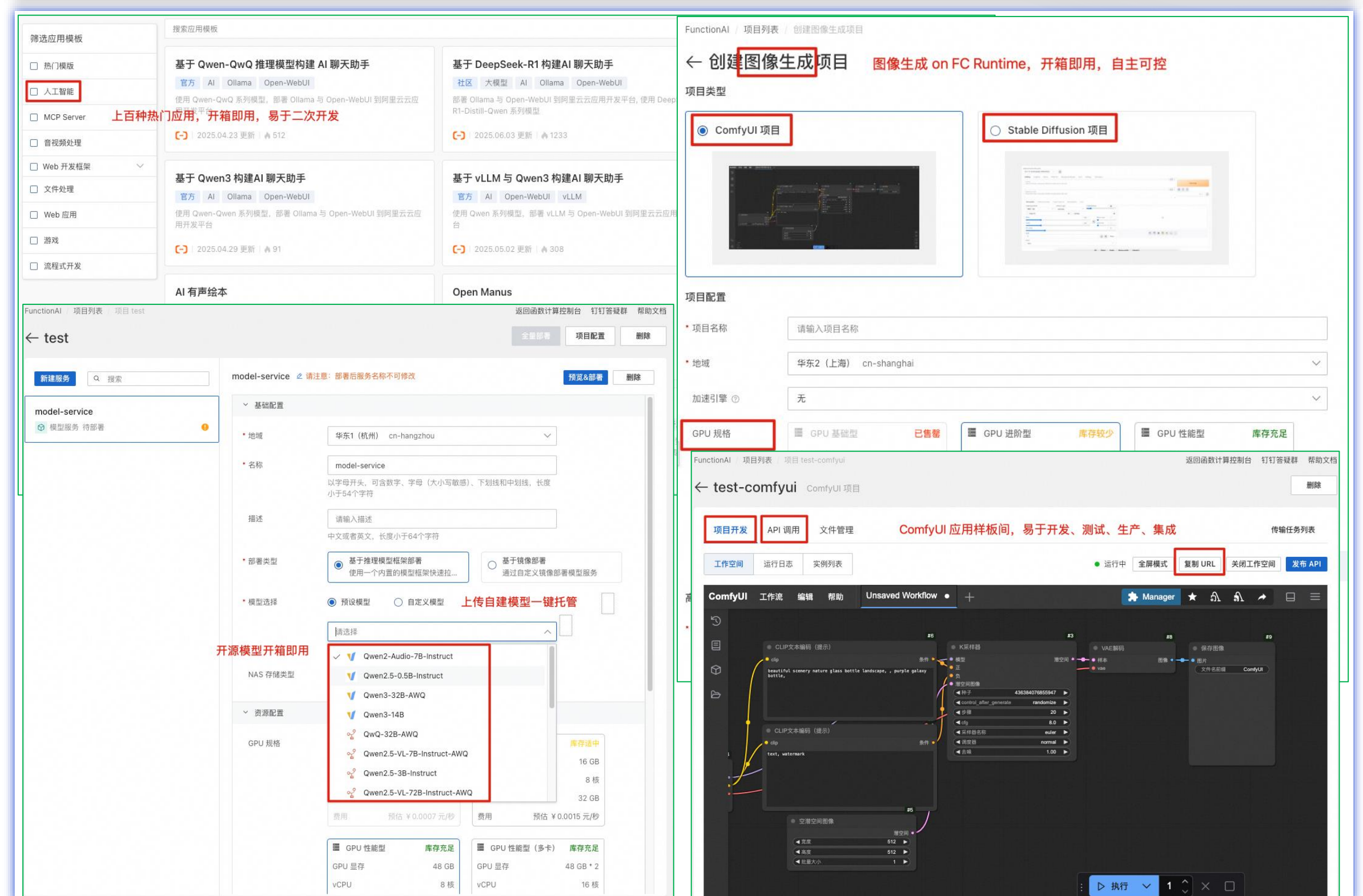
热门模型一键部署



魔搭社区

6000+热门大模型支持一键部署至用户指定函数计算，快速体验和托管主流模型

专属开发空间



模板/模型服务

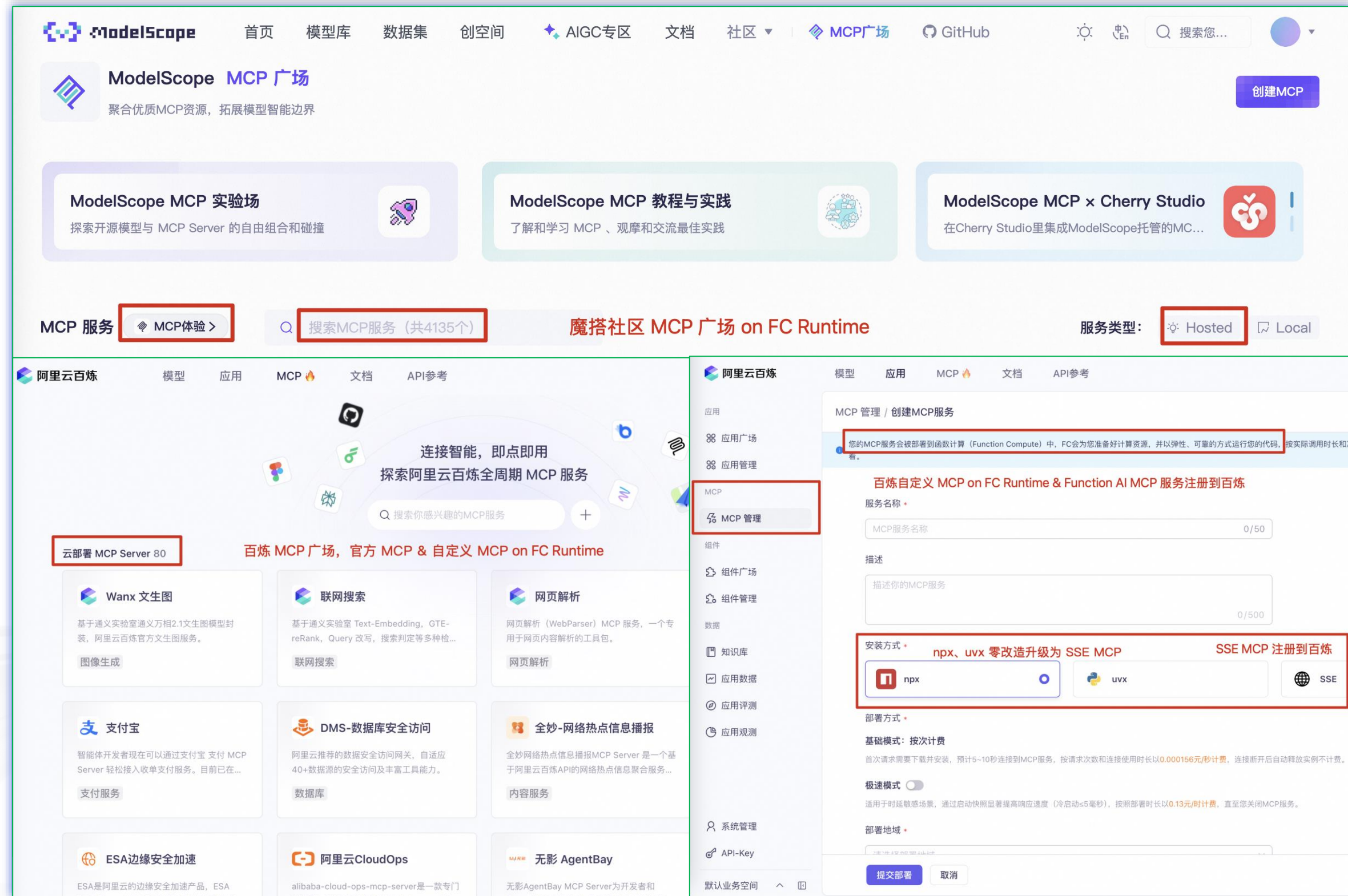
100+热门 AI 应用模板如 SD、ComfyUI、Flux 等支持一键部署，快速托管热门模型如 Qwen、DeepSeek 等



AI 时代运行时最佳实践 – MCP 服务

函数计算 Function AI 联合百炼发布业界首个全生命周期 MCP 服务市场，联合魔搭社区发布最大 MCP 中文社区

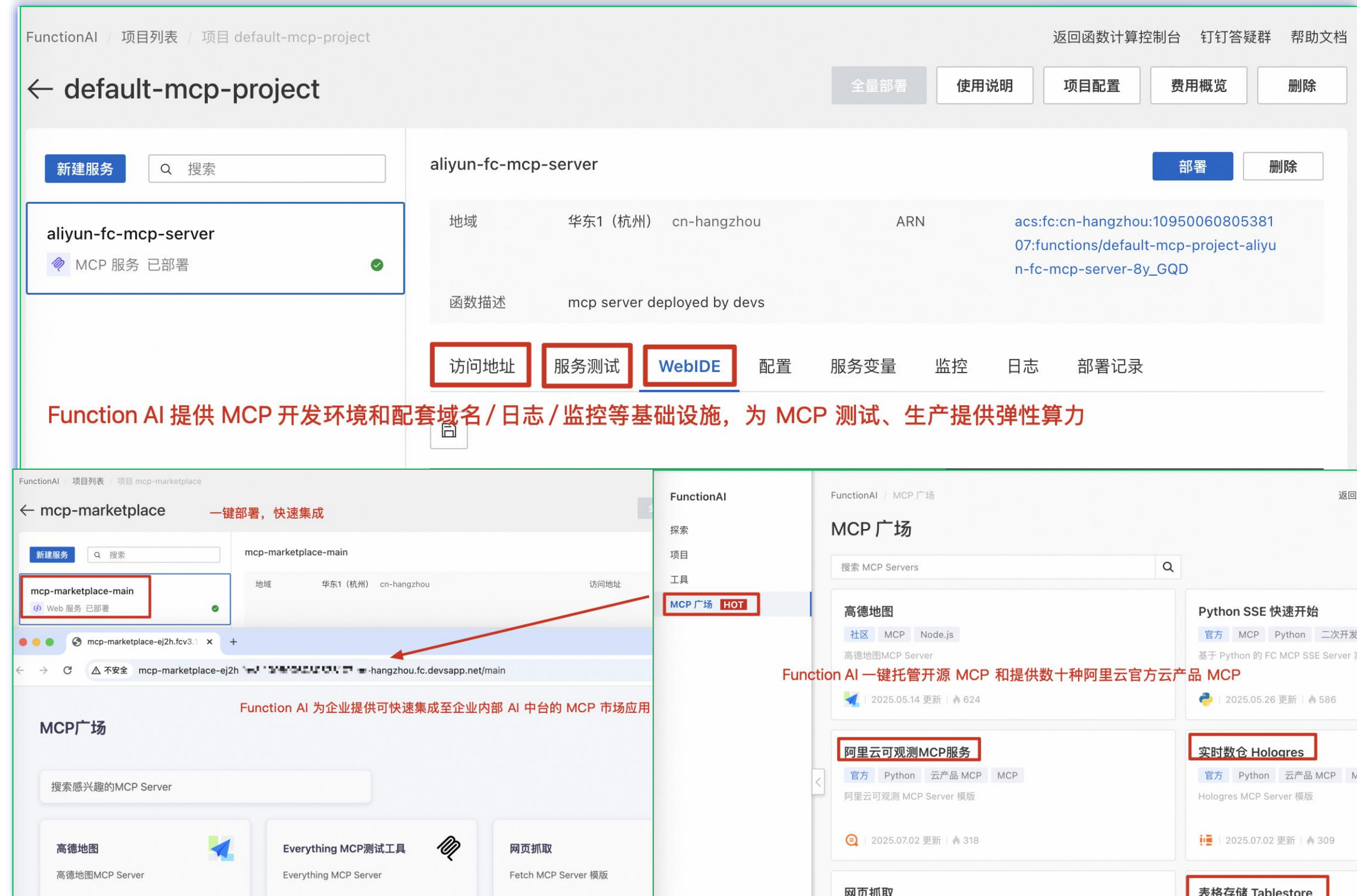
FC MCP Runtime 经过大规模流量验证



百炼、魔搭官方 MCP 市场

百炼、魔搭使用 FC MCP Runtime 为数万开发者的海量访问提供大规模 MCP 服务

助力企业构建专属 MCP 市场，丰富 AI 工具生态



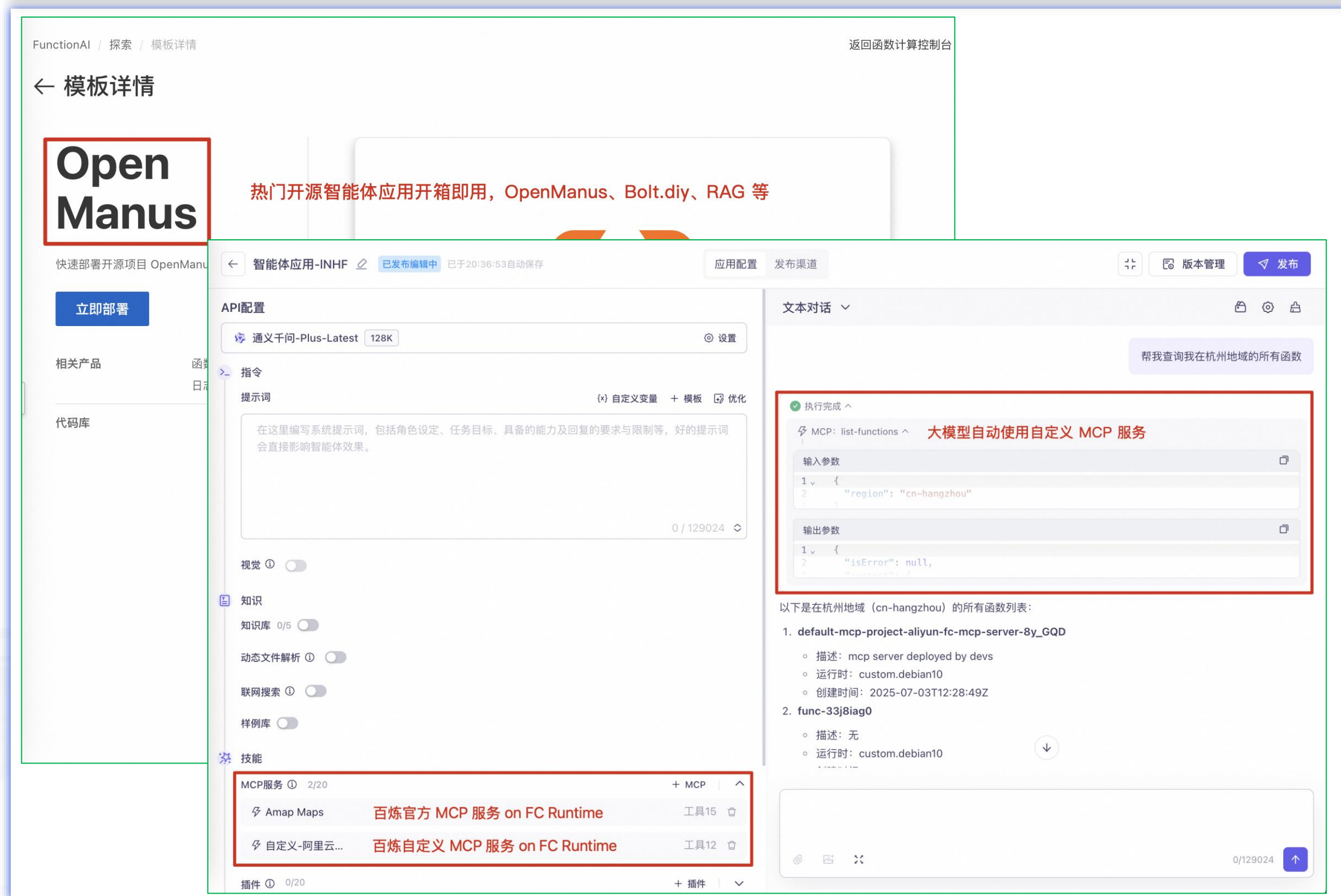
MCP 开发环境 & MCP 市场应用

开发者可在 Function AI 零改造将开源 MCP 升级为 SSE/ Streamable HTTP，快速构建企业专属 MCP 市场

AI 时代运行时最佳实践 – 智能体/工作流

函数计算 Function AI 作为智能体/工作流/ MCP 工具等 AI 运行时底座，深度集成百炼等大模型应用平台

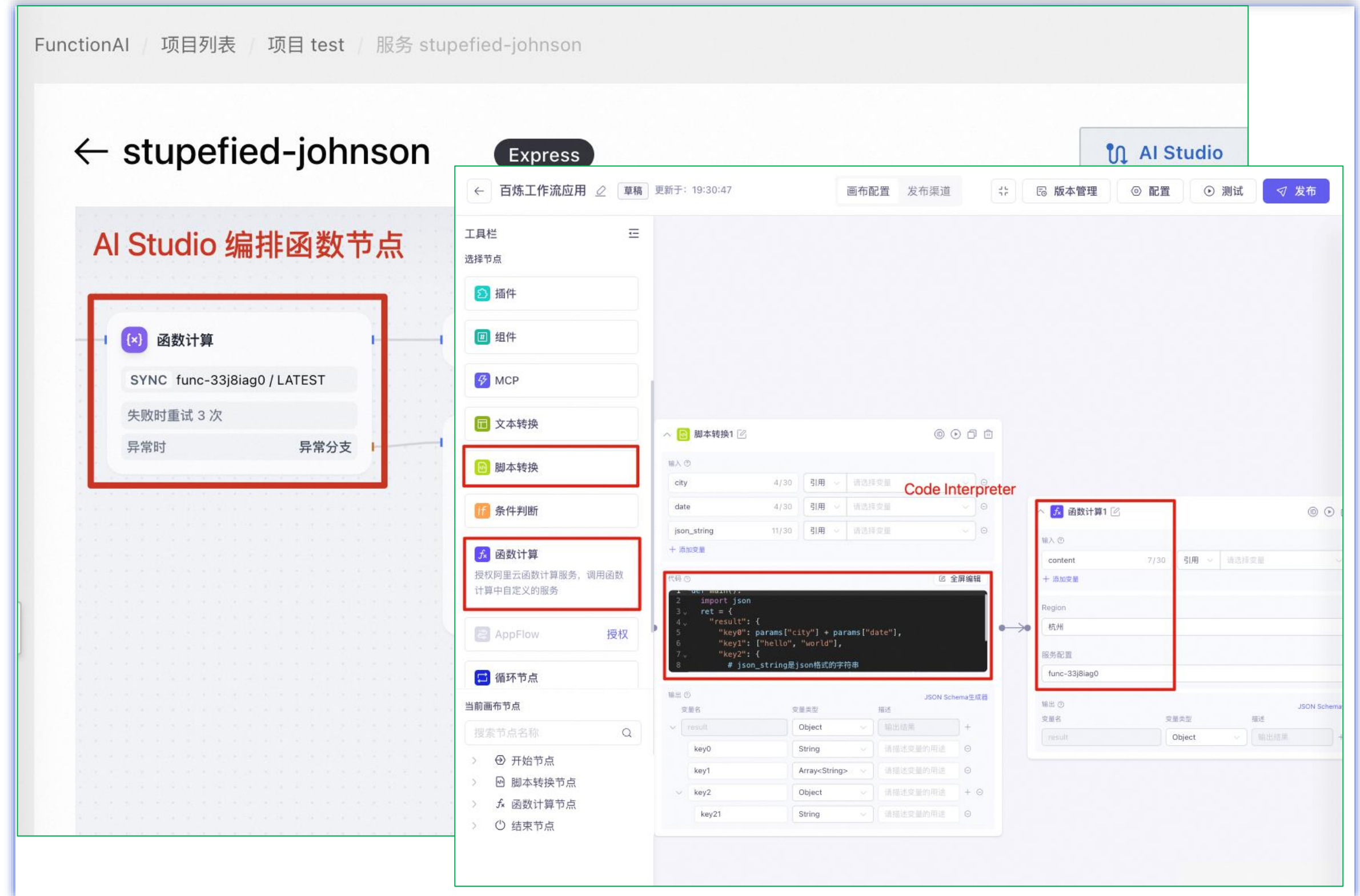
智能体应用工具底座



智能体应用

Function AI 热门开源智能体项目开箱即用，百炼智能体应用集成 FC Runtime 作为 MCP 和插件等工具底座

工作流节点



工作流应用

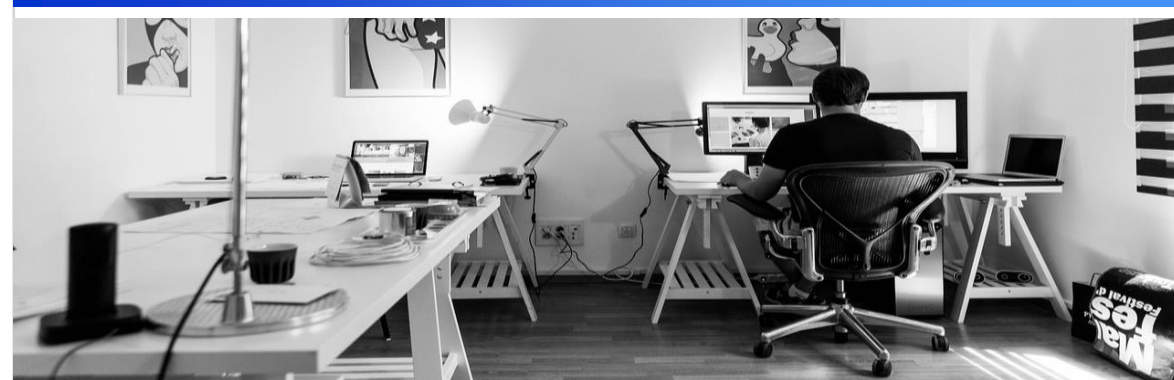
Function AI AI Studio 函数节点，百炼工作流应用集成 FC Runtime 作为 Code Interpreter

AI 时代运行时最佳实践 – 客户故事

Serverless x AI，函数计算 FC 帮助企业跨越技术鸿沟，让 2 人的研发团队享受 2000 人技术团队的红利

设计师

定制个人工作助手



客户原声

场景：某4A景区设计师蒋鹏，希望用AI绘画平台为游客照进行风格化助力，增加趣味性和景区收入

痛点：不会编程、GPU 价格昂贵，SaaS 化平台无法满足个性化定制需求

Function AI + ComfyUI

降本	出图速度
78%	1图/10分钟 -> 1分钟/12图

开发者

开发有趣的 Web 项目



客户原声

场景：wolai 是一款开发者打破“边界”的在线文档产品，其独特的功能和舒适的交互吸引着众多极客用户尽享 Web 体验

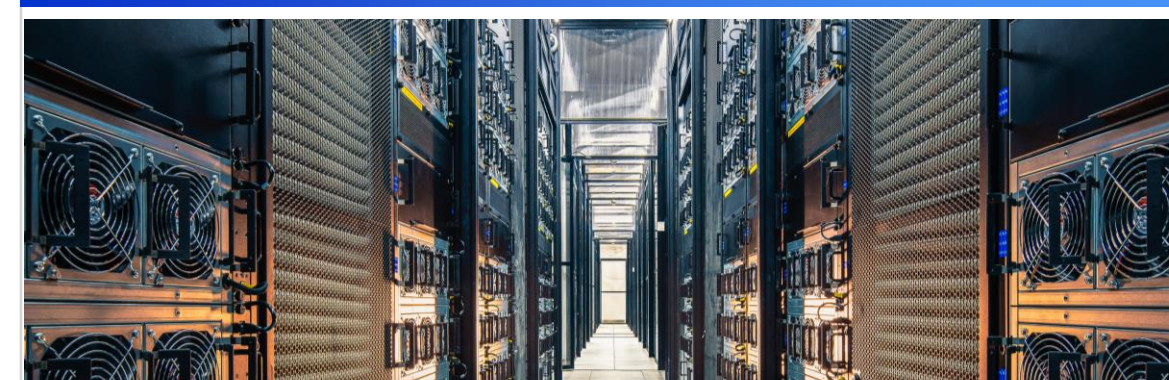
痛点：“idea”要快速验证，访问体验要极度丝滑，增量发布要稳定常态

FC + Web 全栈

降本	开发提效
100万/年 -> 50万/年	前端开发 -> 全栈开发

初创公司

自媒体浪潮中的潮流引领



客户原声

场景：伊对 App 是一家专注音视频恋爱社交的高新企业，高度自动化的直播截帧送审技术是其业务创新的核心推动力之一

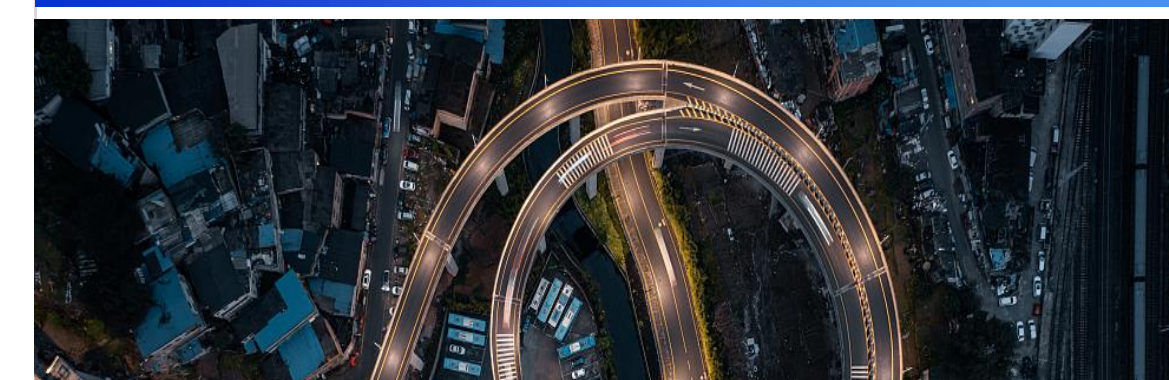
痛点：大规模突发流量，保障业务爆发式发展的同时追求稳定、高效、降本

FC + 任务调度

降本	高度自动化
20%	0 运维

大型企业

灵活可定制，加速业务 AI 创新



客户原声

场景：服装企业森马，希望借助 AI 技术实现服装行业从产品设计、生产制造、营销推广的多场景提效。

痛点：GPU 成本高、开发周期长、适配多种开源大语言模型，定制化要求高

Function AI + Stable Diffusion

降本	迭代效率
33%	3天 -> 30分

THANKS



基于 Function AI 分钟级构建生产级 AI 应用

—— Serverless + AI 让应用开发更简单

世如 (史明伟)

云原生-Serverless- 函数计算&AI 研发负责人



Contents

目录

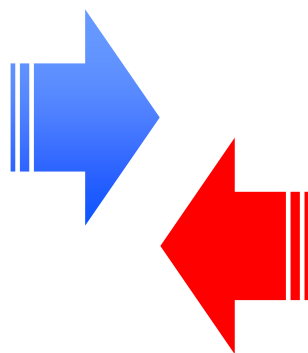
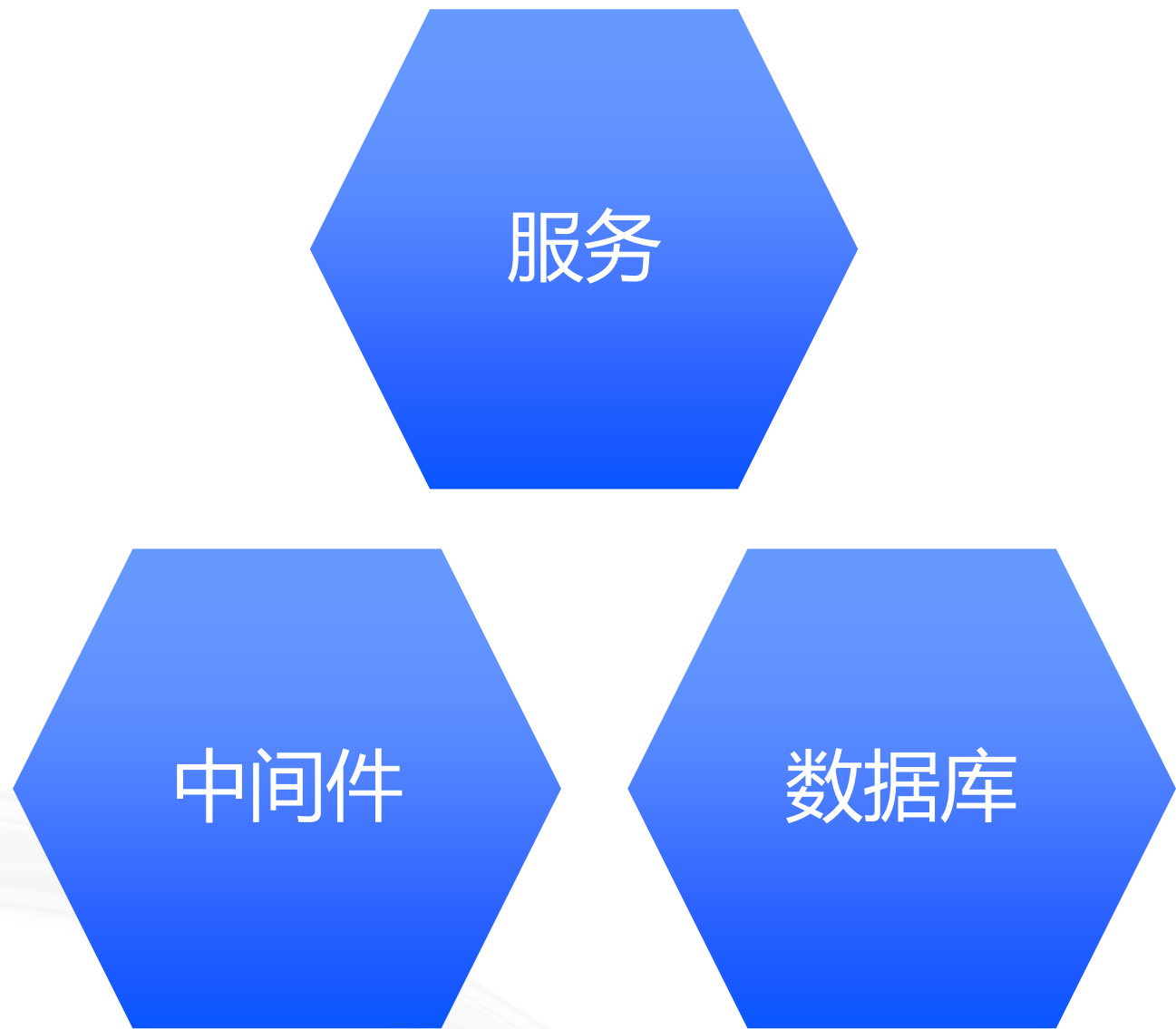
01 分钟级构建 AI 应用的挑战

02 Function AI 开发平台介绍

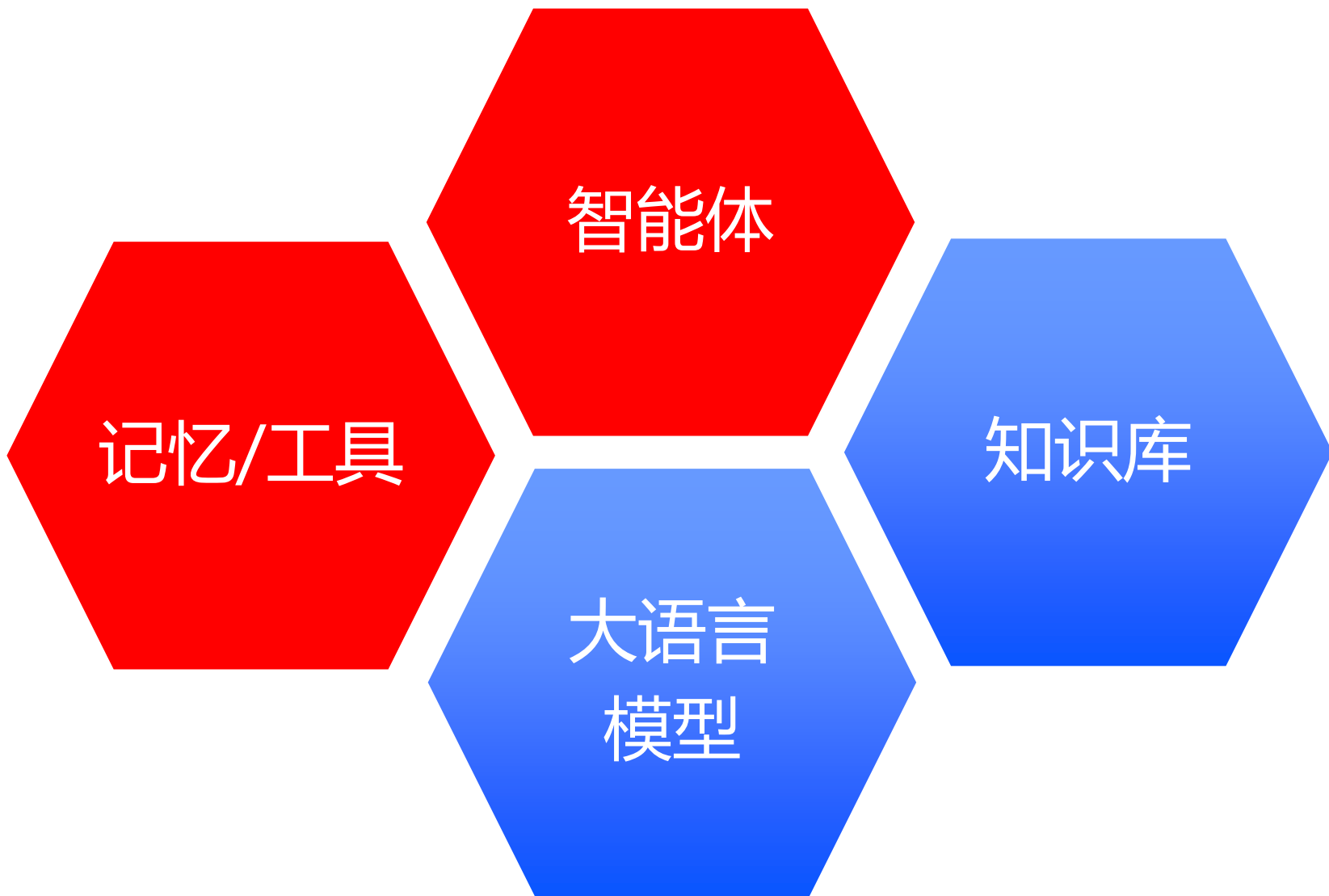
03 分钟级构建生产级 AI 应用

应用抽象发生变化

传统应用开发



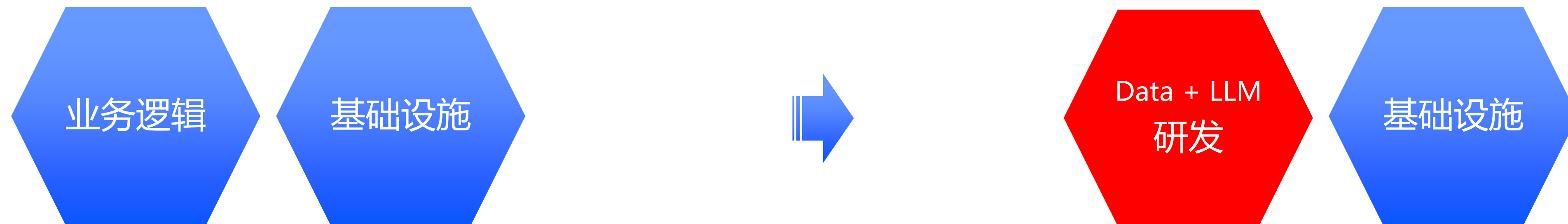
AI 应用开发



研发的关注点发生变化

传统应用开发：如何确保业务逻辑正确稳定运行

AI 应用开发：如何最大可靠的发挥LLM价值



AI 应用基本要素



分钟级构建 AI 应用的挑战

克服LLM幻觉

缺少解决方案

开发部署效率

快速测试访问

业务灵活定制

业务成本最优

运行时安全隔离

数据安全访问

基础设施稳定性

基础设施弹性与扩展性

Serverless 在解决什么问题？

Serverless



目标和策略

成本目标

- 按需付费，降低用户成本，提供产品竞争力

效率目标

- 简单，易用，减少发布/扩容时间，提升发布/扩容的效率
- 通过不断的优化资源供给能力：降低用户保有资源的成本，提高资源利用率，降低资源使用成本；
- 通过不断的加强和云产品及周边生态的集成，降低用户业务构建的门槛，减少业务发布和扩容运维时间，提升业务效率；

- **平台及体验能力升级**：云服务集成，事件驱动，函数编排，应用模版，计费优化，观测能力集成

Contents

目录

01 分钟级构建 AI 应用的挑战

02 Function AI 开发平台介绍

03 分钟级构建生产级 AI 应用

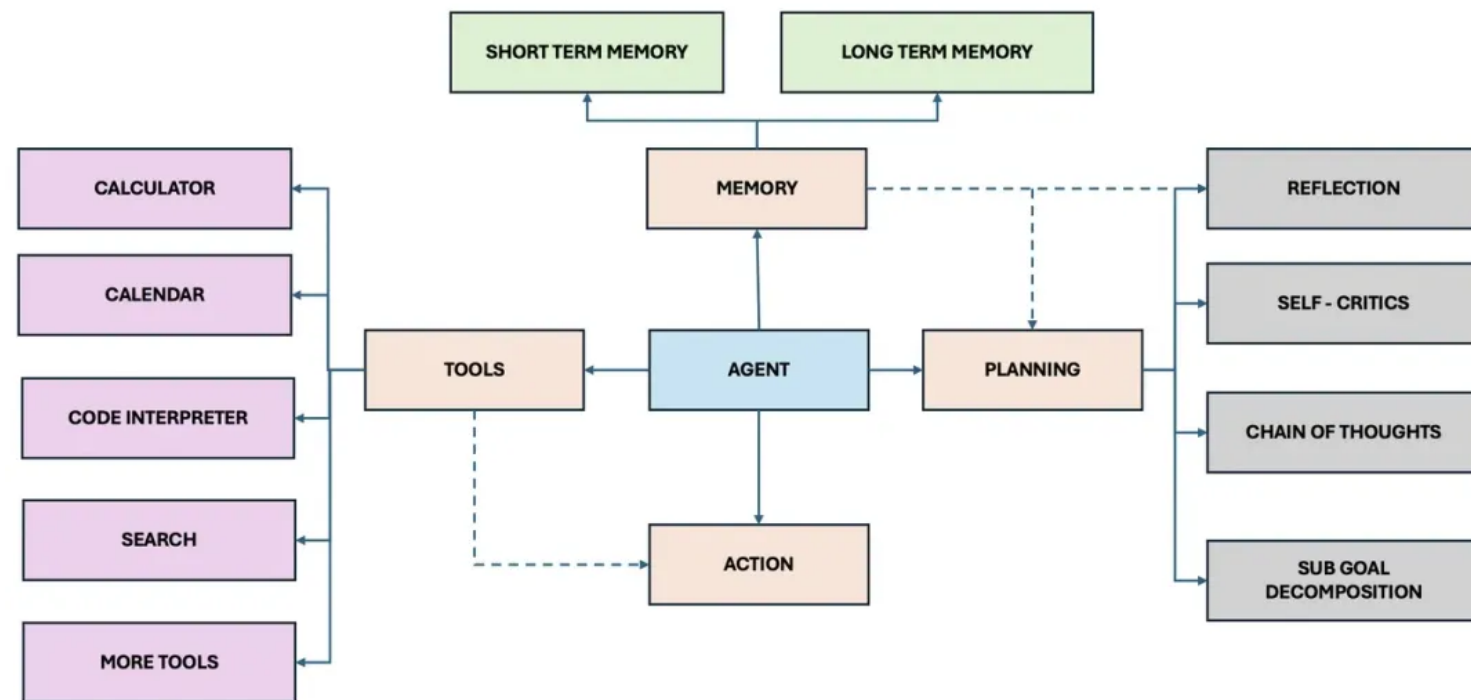
Function AI: Serverless + AI 让应用开发更简单



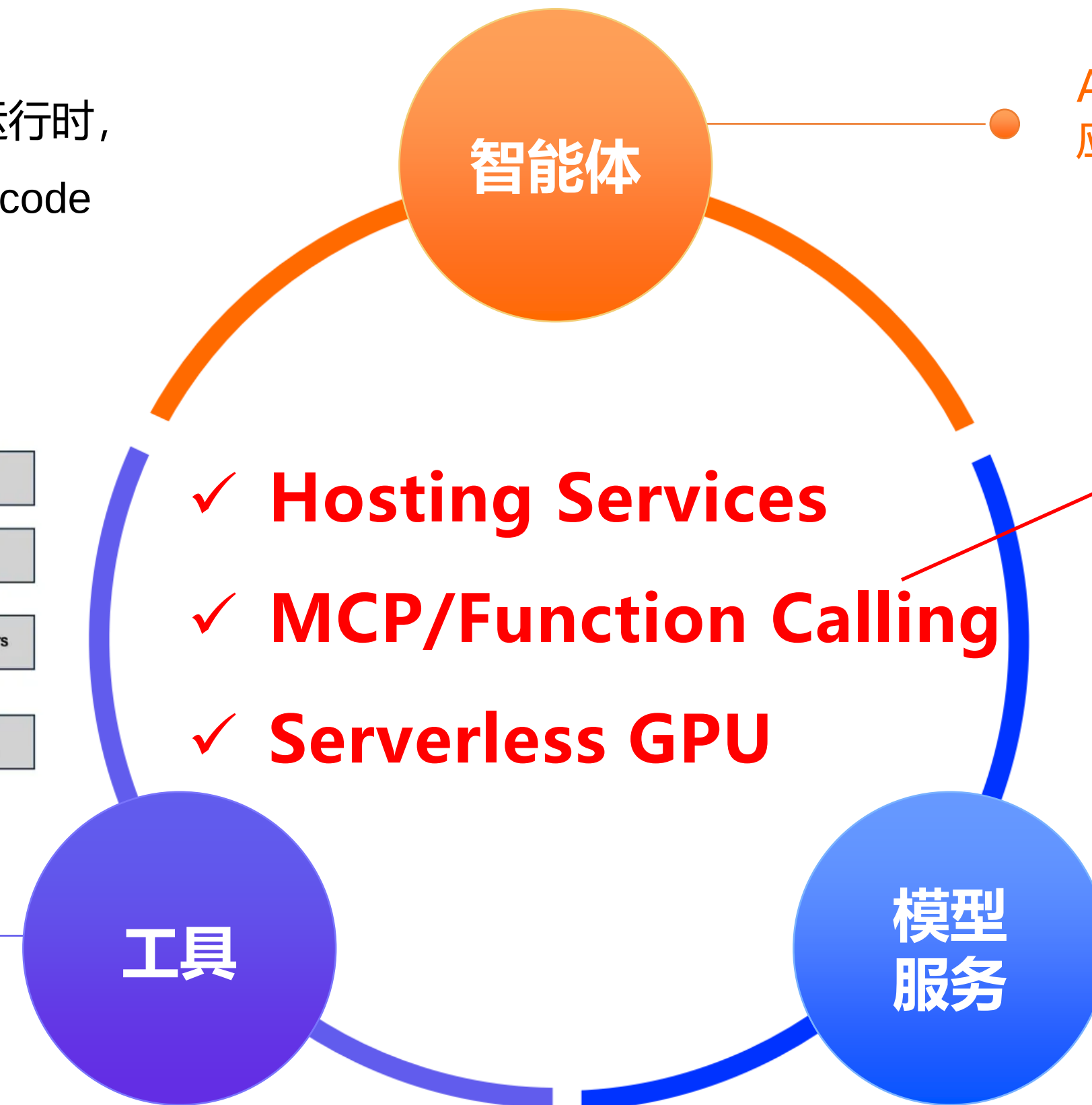
Function AI: 聚焦 AI 应用运行时

LLM 和现实世界的结合

- 包含感知、决策、行动
- 行动：碎片化代码，需要灵活、快速、轻量的运行时，函数是最合适的形态（MCP, function calling, code interpreter,）



托管生态及 Serverless 算力，扩展及运维价值



AI 领域应用的主体
应用集成及应用运维价值

Serverless 是 Agent 的
最佳运行时承载

Serverless GPU 支撑自
定义领域模型托管



Function AI：为 AI 应用开发提效

Function AI 聚焦 AI 应用场景：**模型服务**（魔搭社区/HuggingFace 热门模型托管）；
图像生成（SD、ComfyUI Serverless API）；**MCP & Agent**（企业专属 MCP 市场）

Function AI 一键创建应用

Function AI / 探索

返回函数计算控制台 钉钉答疑群 帮助

Serverless + AI，让应用开发更简单！

Function AI 是基于函数计算构建的 Serverless AI 应用开发平台，内置极致的弹性算力、探索丰富的 AI 应用模板、开箱即用的 AI 工程能力与应用的全生命周期管理。让技术不再是束缚，让AI化作创新的画笔，即刻体验，开启您的 Serverless AI 之旅！

筛选应用模板

☐ 热门模版

☐ 人工智能

☐ MCP Server

☐ 音视频处理

☐ Web 开发框架

☐ 文件处理

☐ Web 应用

☐ 游戏

☐ 流程式开发

搜索应用模板

基于 Qwen-QwQ 推理模型构建 AI 聊天助手

官方 AI Ollama Open-WebUI

使用 Qwen-QwQ 系列模型，部署 Ollama 与 Open-WebUI 到阿里云云应用开发平台

2025.03.06 更新 272

基于 DeepSeek-R1 构建AI 聊天助手

社区 大模型 AI Ollama Open-WebUI

部署 Ollama 与 Open-WebUI 到阿里云云应用开发平台, 使用 DeepSeek-R1-Distill-Qwen 系列模型

2025.03.04 更新 1055

基于 DeepSeek-R1 构建 RAG 应用

社区 大模型 AI 检索增强生成 模板

CAP rag with deeepseekr1

2025.03.16 更新 173

AI 剧本生成与动画创作

官方 AI 大模型 阿里云技术内容

AI 剧本生成与动画创作示例工程

2025.03.04 更新 707

组装式开发，弹性开放，按需选择

模型托管

图像生成

MCP 服务

Agent 服务

无缝升级 AI 应用开发范式

应用开发

应用市场

AI 运行时内置 vLLM, SGLang, Ollama, HuggingFace 推理框架

海量 Serverless AI 应用模板，开箱即用，支持二次开发

MCP 运行时内置 MCP 开发框架, STUDIO 零改造升级 SSE

魔搭社区、百炼官方集成, 开源模型/智能体应用增强

模型服务：让模型使用更简单，推理性能更卓越

产品价值	价格低 弹性模式 冷启动长 价格低 闲置模式 无冷启动 价格高 预留模式 无冷启动			• 低延时：规避GPU按量场景下的冷启动 • 降成本：对比K8S自建，大幅降低AI落地成本 • 保交付：GPU紧张背景，确定资源交付，保证业务永续		
应用场景	传统的在线推理场景 延时高度变化，负载高度不确定，偏C端的应用形态，日均资源利用率普遍较低；例如：传统的CV类模型（OCR）、NLP模型（机器翻译），CosyVoice，SenseVoice		新兴的大模型推理场景 AIGC浪潮下的新兴推理应用，LLM文生文、Stable Diffusion文生图、FunASR文生音频等			
模型来源	 ModelScope	 Hugging Face	 文件存储 NAS	 对象存储 OSS	 人工智能 PAI	 自定义模型镜像
推理加速框架	 vLLM 为大型语言模型(LLM)设计的高性能推理框架 获取支持模型列表 🔗		 SGLang 面向大语言模型的高性能推理框架 获取支持模型列表 🔗		 Ollama 轻量级推理框架，专注于量化模型部署及各种开源 LLM 部署 获取支持模型列表 🔗	 Hugging Face Transformers 由 Hugging Face 提供的模型推理框架 获取支持模型列表 🔗

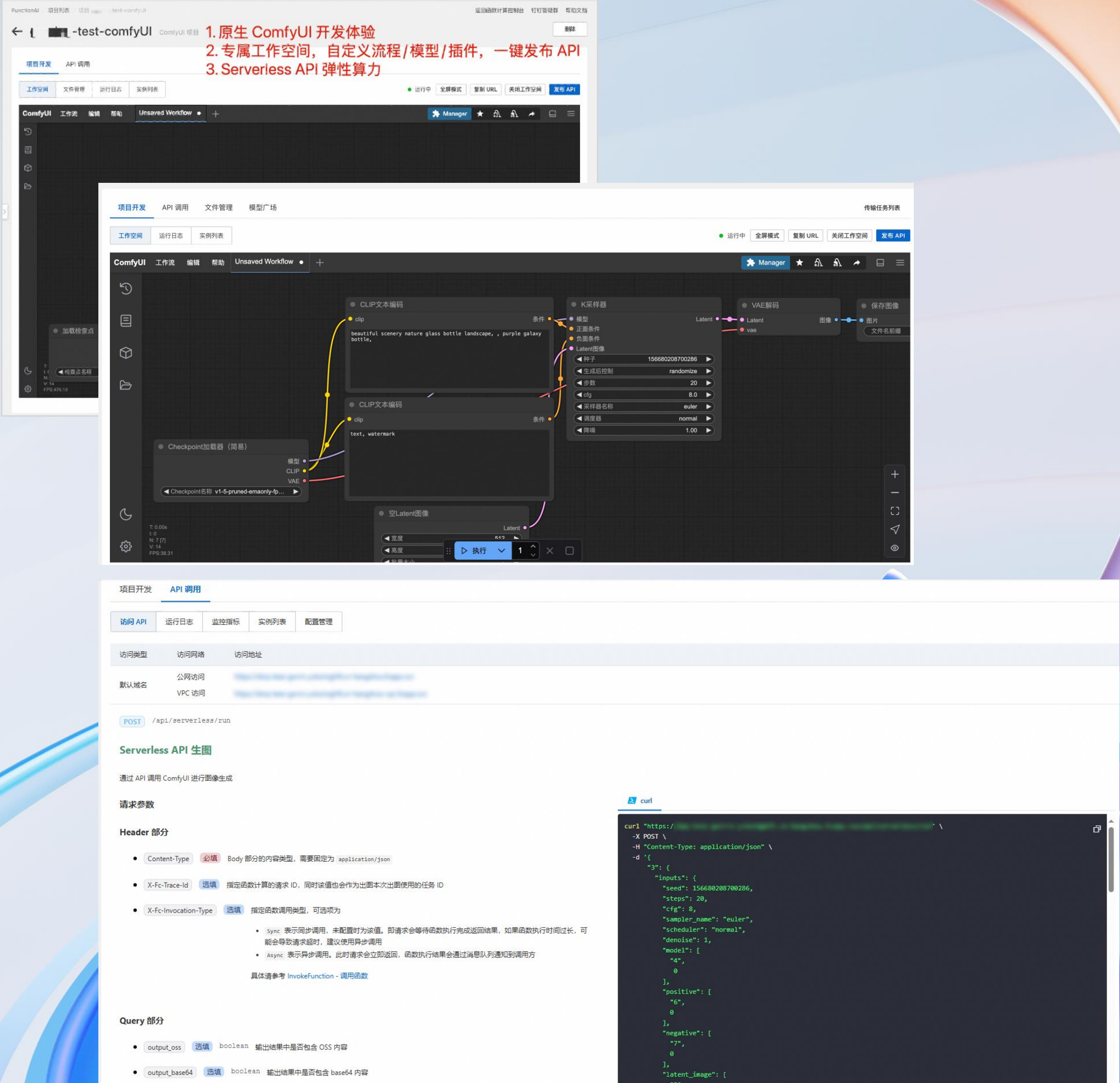
图像生成： 领域模型加速落地

行业痛点

- SaaS 昂贵且无法自主可控
- 虚拟机/容器自建弹性效率弱、利用率低、成本高
- 文生图 WebUI/API 开发门槛高，模型/插件下载国内网络不可用

Serverless图像生成平台价值

- 深度集成**Serverless GPU**，兑现弹性降本优势、快速出图
- **专属工作空间**：激活、销毁快，简单易用
- **企业级特性增强**： Serverless API，自带流量灰度，多实例并发出图加速出图效率
- 文件管理：方便模型，插件上传，支持本地上传
- 模型广场：快速找到符合业务需要的模型
- 工作流广场：提供流程和流程依赖的整体部署



函数计算： FunctionAI 图像生成

- ✓ SD WebUI/CompyUI 一键部署，提供专属工作空间和网络加速能力，支持上传、下载自定义模型及插件
- ✓ 开发调试一键发布Serverless API，隔离测试和生产环境，通过API调用实现同步/异步/流式/WebSocket出图

插件广场：加速智能体应用集成

MCP 为LLM连接数据源和工具提供了标准化的方法

Serverless Hosting MCP server 的优势

- MCP 稀疏调用、非稳态调用、脉冲型调用特征明显，算力规格要求小，天然适合弹性
- Nodejs、Python主流，启动快，天然适合FaaS运行时
- 多租隔离，满足安全要求
- 内置负载均衡以及安全防护，支持集成网关实现更高阶流量管理

Serverless Hosting MCP server 的劣势

- 大部分MCP server都是STDIO模式，不支持多并发，不具备鉴权能力
- SSE 会话亲和性特性，不适合Serverless按量弹性不同实例
- SSE 协议长连接特性，造成函数实例长时间活跃，持续产生费用
- 本地调试困难，问题难定位，缺少开发及部署MCP的最佳实践

FunctionAI MCP Runtime & MCP 广场

- 内置MCP Proxy，存量STDIO无改造直接生成SSE/Streamable Http服务，且具备单实例多并发能力
- 发布MCP 会话亲和性调度，保证相同sessionID会话绑定到相同实例，具备会话配额控制保证单实例会话公平、MCP会话优雅升级
- 发布长连接闲置计费，最高降低87%成本
- 提供模板一键部署、自定义MCP服务，满足开源托管以及二次开发需求
- 提供隐私变量管理、服务鉴权、内网访问能力，本地STDIO协议无需改造即可提供安全的企业级服务



AI Studio：打造稳定高性能低代码 Agent 部署平台

AI Studio 提供类似 Dify 的可视化 Agent 构建平台，同时采用自研工作流执行引擎，10 倍提升 Agent 运行效率

AI 智能生成工作流框架

和错误处理也是通过FC实现的逻辑，确保系统的稳定性和用户体验。

3. 注册新用户：对于新用户注册请求，同样通过FC处理，检查用户名是否已被占用，然后创建新账户。

4. 简单的错误与成功反馈：无论是登录还是注册，流程的最后都会通过一个简单的Pass状态来反馈操作结果，成功或失败，以及可能的错误信息。

继续优化 使用该方案 →

创建新用户并返回注册成功的消息。同时，此步骤中加入限流逻辑，如果FC端因为并发请求过多触发限流，将捕获这一错误并返回给用户相应的错误信息。

3. 响应处理：根据FC的处理结果，HTTP任务会构造响应给前端，告知用户登录或注册的结果。

注意：实际应用中，限流的错误处理应该更加细致，可能涉及到重试策略、熔断机制等，但在此简化示例中，我们主要展示如何在CloudFlow中表达基本的流程和简单的错误处理。

继续优化 使用该方案 →

灵感 生成注释 细化流程 优先函数计算 优先自建网关

生成一个用户登陆注册业务流程，细化流程，优先函数计算，同时需要能够对于并发请求出现限流增加错误处理。

生成工作流

← luohao

请输入搜索内容

操作 流程 模式

常用

InvokeFunction

FC:InvokeFunction

StartExecution

FNF:StartExecution

SendMessage

MNS:SendMessage

Request

HTTP:Request

PutObject

OSS:PutObject

PutEvents

EventBridge:PutEvents

弹性计算

函数计算 FC

函数工作流 FNF

云服务器 ECS

类Dify
可视化流程

自研
流程引擎

AI Studio

Function AI

函数计算

CloudFlow

Contents

目录

01 分钟级构建 AI 应用的挑战

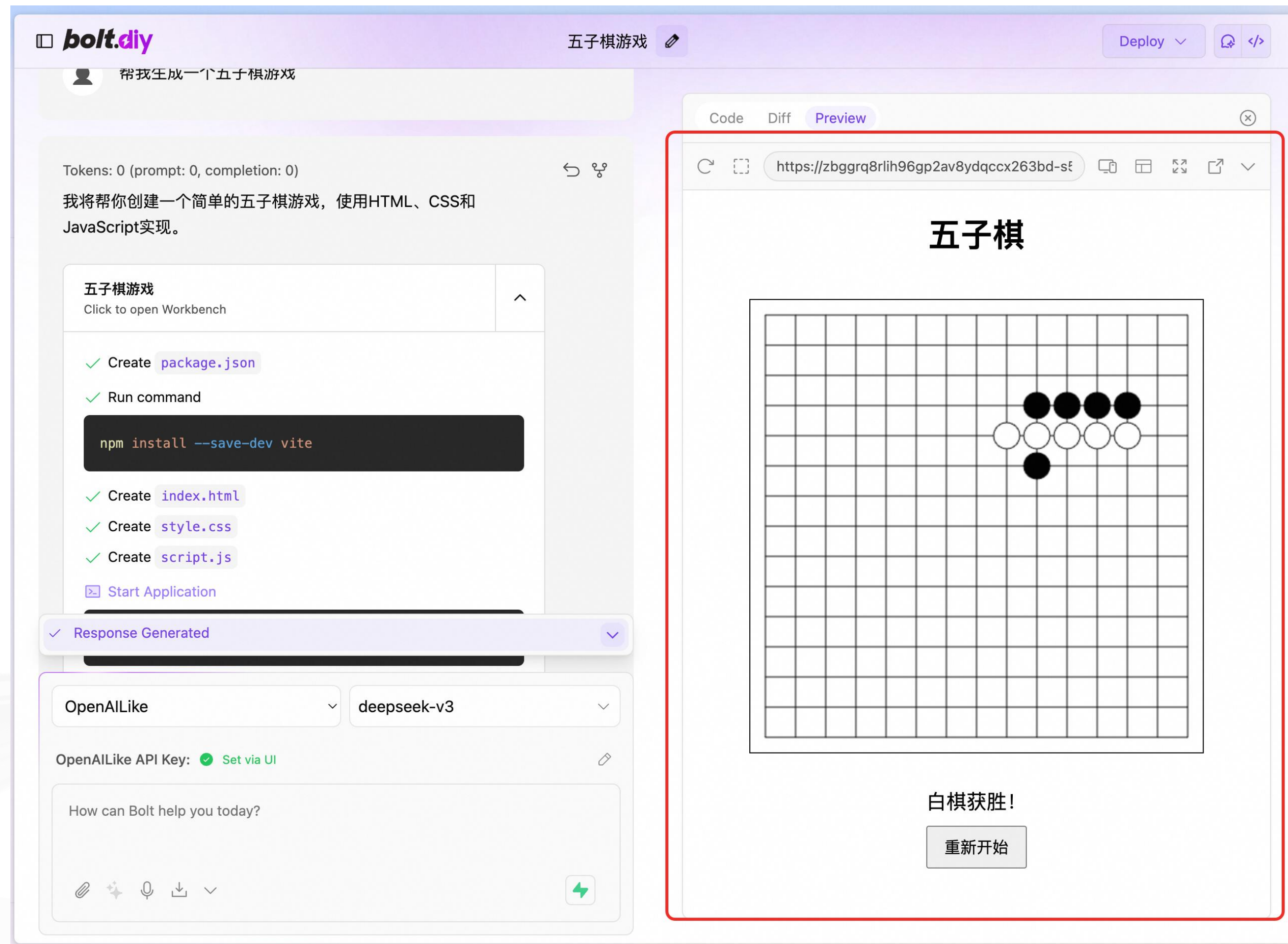
02 Function AI 开发平台介绍

03 分钟级构建生产级 AI 应用

Function AI：分钟级落地AI应用架构新范式



Function AI: AI Agent 提供安全高效的 Sandbox 运行环境



- **Agent Sandbox Runtime:** 如 Code Sandbox、Browser Use Sandbox、仿真训练 Sandbox、RL Sandbox
- **成本最优:** 按请求扩缩，提升 AI Agent 资源利用率，降低资使用源成本
- **多租强隔离:** 安全容器+动态存储挂载，为每个启动的 Agent 提供干净独占的运行环境
- **启动速度快:** CPU 百毫秒弹性伸缩
- **内置完善的监控日志能力:** OpenTelemetry Trace 支持
- **会话亲和调度:** 保持业务运行时及业务上下文
- **业务隔离:** 运行时支持按请求隔离，会话隔离

THANKS



AI 时代的“智能流量中枢”， AI 网关搭建与落地

赵世振

阿里云云原生产品架构师



Contents

目录

01 AI 应用发展趋势及挑战

02 AI 网关产品核心能力

03 AI 网关落地实践

04 AI 网关应用场景及客户案例

AI 应用发展趋势及挑战

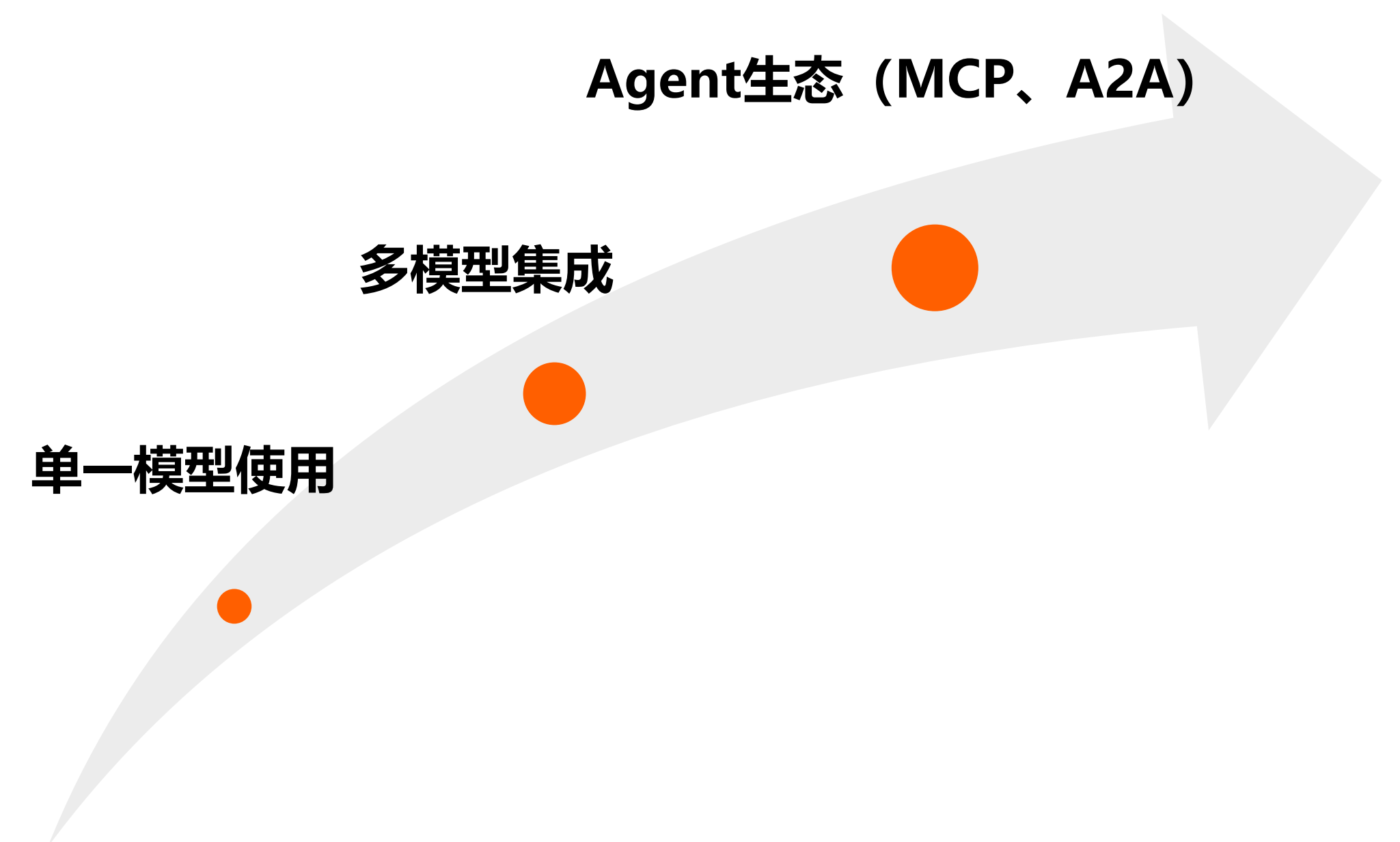
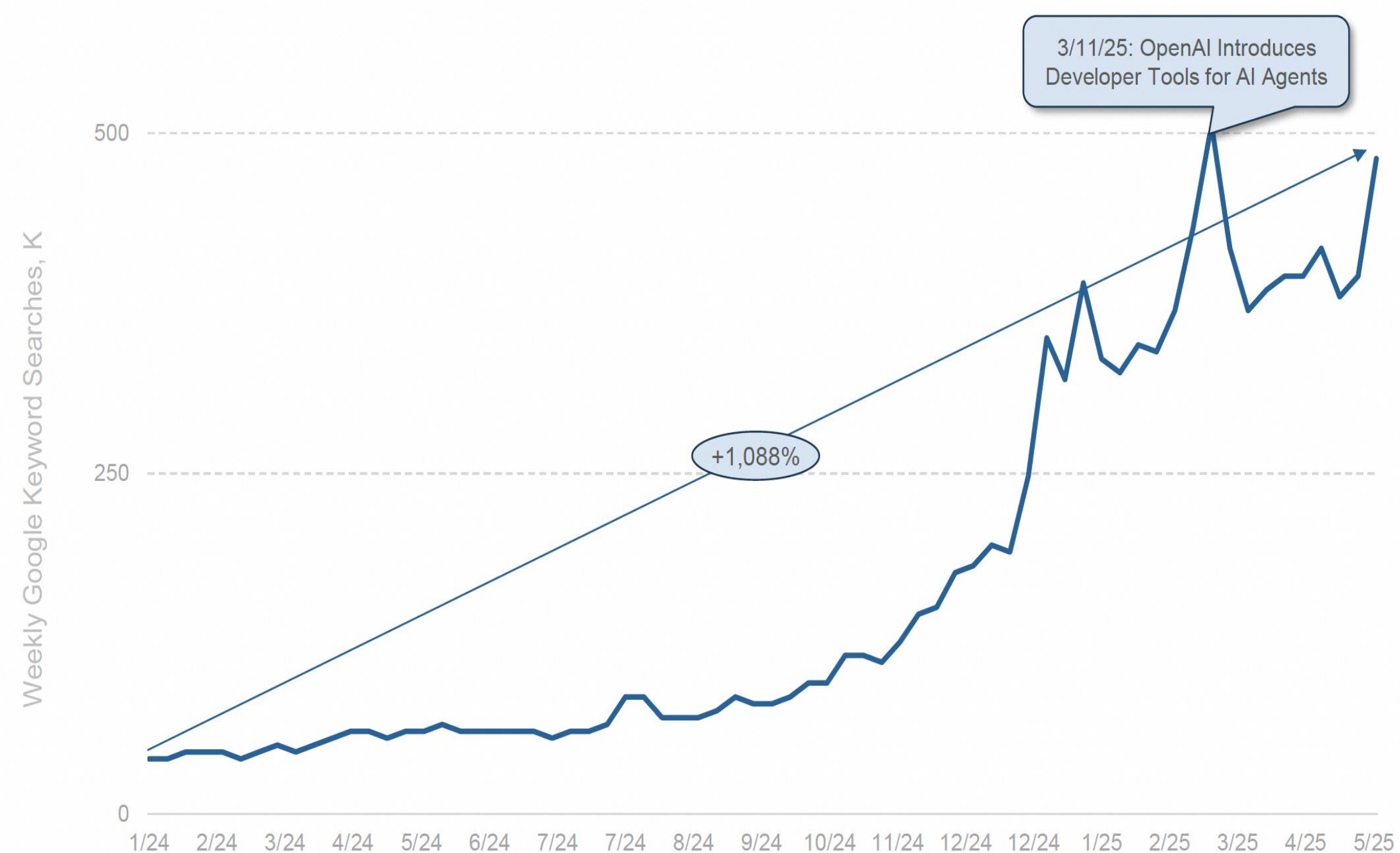
New Future on Cloud New Future on Cloud New Future on Cloud

01

AI 应用发展趋势

AI Agent Interest (Google Searches) =
+1,088% Over Sixteen Months

Global Google Searches for 'AI Agent' (K) – 1/24-5/25, per Google Trends



Source: Google Trends via Glimpse (5/15/24), OpenAI (3/25)

AI 应用落地过程中遇到的核心挑战

安全与合规风险

- API-KEY泄漏风险
- 敏感数据泄露
- 生成内容不可控

高可用保障

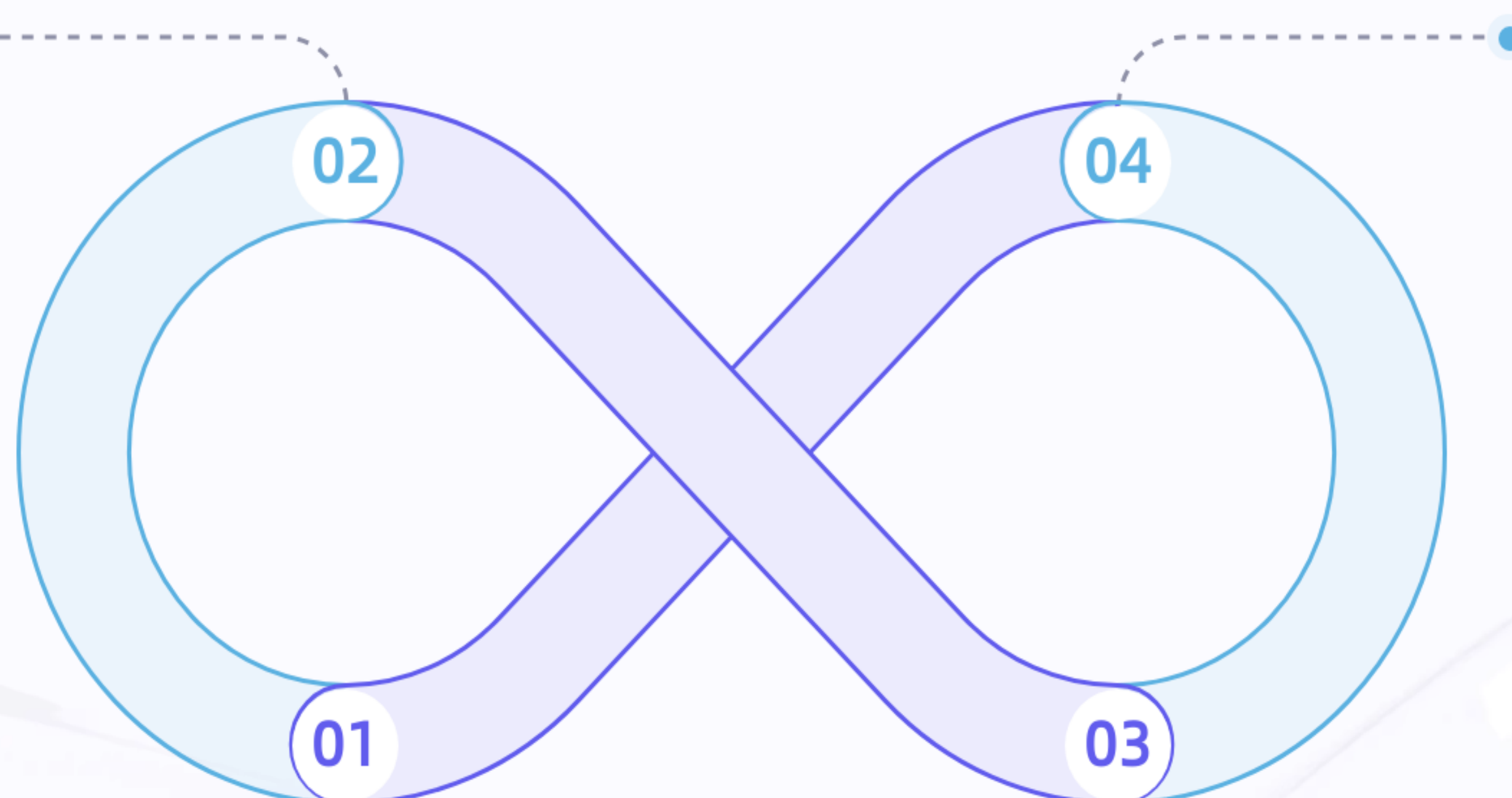
- 模型服务中断
- RT和成功率波动影响用户体验
- 监控告警缺失

集成复杂度

- 统一多模型代理
- MCP 服务、agent集成
- 权限管理复杂

成本失控

- 重复问答浪费
- 无限流机制



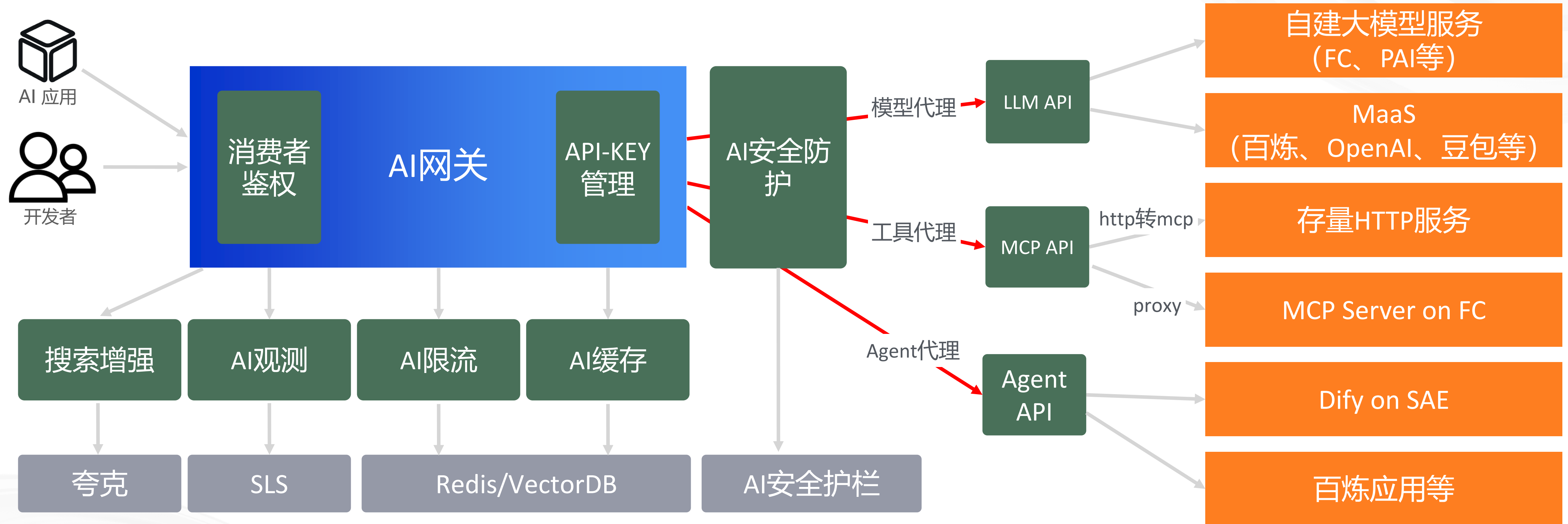
AI 网关产品核心能力

New Future on Cloud New Future on Cloud New Future on Cloud

02

AI 网关产品核心能力

企业访问大模型、MCP工具和Agent的统一入口，提供全方位的治理和优化能力，构建企业级 AI 智能流量中枢



统一接入代理

- 多模型统一代理
- MCP协议适配
- Agent API封装

安全鉴权管理

- API-KEY托管、二次分发
- 消费者鉴权
- AI安全防护

高可用保障

- Fallback机制
- AI限流
- 联网搜索优化模型幻觉

成本优化

- AI缓存
- AI观测

AI 网关落地实践

New Future on Cloud New Future on Cloud New Future on Cloud

03

AI网关落地实践--解决同一域名/API访问不同模型的问题

核心问题1： 公司GPU资源有限，部署了满血版DeepSeek R1，还有其他一些小模型以及使用百炼的模型服务，现在域名都不统一，分发、管理、集成的成本都很高，如何使用同一个域名来访问不同的模型？

解法：

- 满血DS R1和其他模型或者闭源LLM API服务共存，保持同一个API接口，不同业务通过请求中的模型名称，切换不同的模型。
- 满血DS R1和其他模型或者闭源LLM API服务共存，保持同一个API接口，不同业务通过请求中（Header，Cookie等）携带的业务标识，匹配到不同的模型。

1

维护多个模型服务

- 无论是FC上部署的，IDC部署的，还是闭源LLM API，都可以作为模型服务被维护在AI网关。

2

AI API代理多个模型服务

- 使用多模型服务类型创建AI API，在一个AI API中可以添加多个模型服务。
- 模型名称通过Glob语法进行匹配。

3

同一个API请求不同模型

- 同一个API，不同业务传入不同的model name，即可实现模型切换。



建议

- 优先推荐使用模型名称匹配切换的模式，更遵循OpenAI协议。

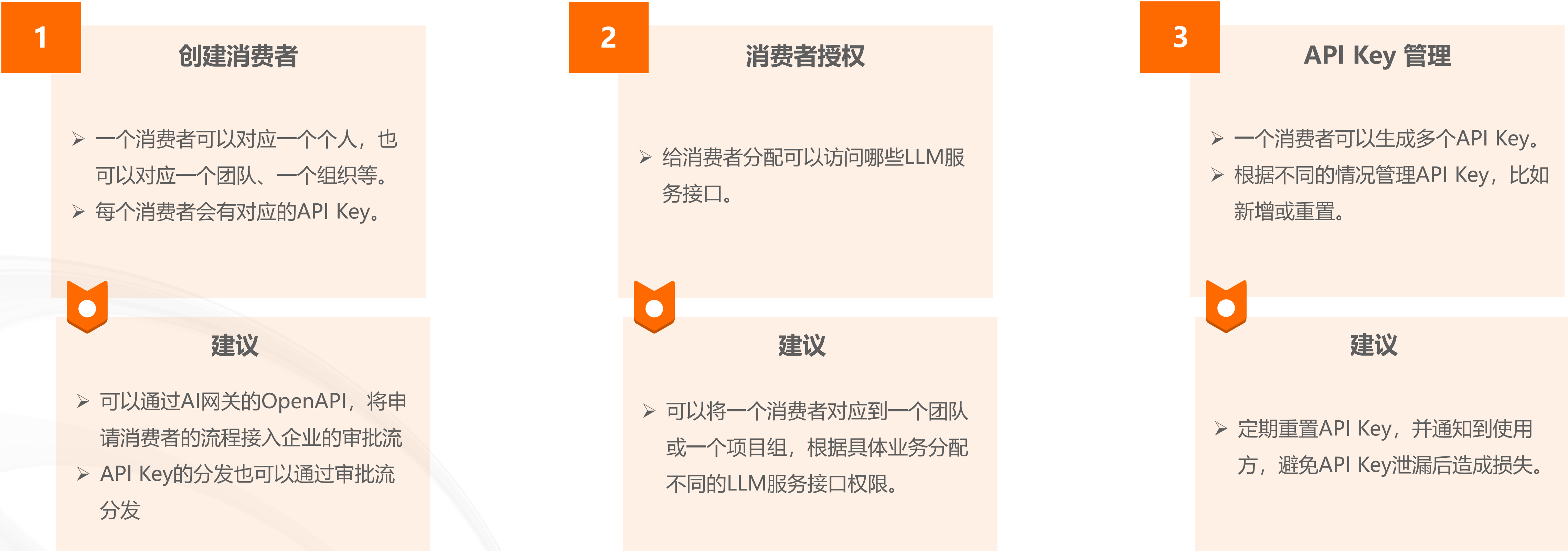
AI网关落地实践--解决用户管理失控问题

核心问题1：我以什么样的方式将LLM服务和能力暴露给大家呢？

解法：OpenAI API的协议基本已经是标准协议，目前市场面上几乎所有的LLM都支持OpenAI API协议。所以提供遵循OpenAI API协议的HTTP接口就可以让企业员工通过各种方式使用LLM服务和能力。

核心问题2：企业内部部署DeepSeek R1 满血版，公司好几千人，但GPU资源有限，如何限制用户？

解法：AI 接口一旦暴露出去，基本上不可能只让一小部分人知道，所以需要对访问LLM服务的用户做以限制，只让能访问的人访问，不能访问的人即便知道了接口也无法访问。



AI网关落地实践--解决百炼DS R1 QPM/Token限制的问题

核心问题： 我们使用百炼平台上提供的DS R1 671B 模型的API，但是有15000 QPM和1200000 TPM的配额限制，不能满足业务需求，但是每次升配很麻烦。

解法：

- 不只是百炼，目前所有的模型托管平台都有QPM和TPM的限制，并且有些平台是很难升配这个限制的，所以大多数用户都会选择申请多个帐号（API Key），变相的撑大这个配额限制，但缺点是在业务里管理多个API Key是一件很麻烦的事。
- 对输入/输出内容做缓存，减少对模型服务的请求次数以及Token消耗，从而提升业务侧的请求性能。

1

模型服务支持多

API Key

- AI网关，每个模型服务都可以配置多个API Key（也可以是自建模型的API Key）。
- 每次请求会轮询拿API Key，对模型服务做请求。

2

API Key可实时维护

- 当监控到API Key配额水位较高时，可以实时动态添加模型服务的API Key。



建议

- 通过AI网关OpenAPI将添加API Key的行为集成到客户自己的自动化平台中。

3

AI API维度结果缓存

- AI API维度支持将输入和输出缓存到Redis，只需要配置Redis地址即可
- 支持精确匹配
- 支持向量化检索匹配



建议

- 在非常垂直类的应用场景下适合开启结果缓存，但建议开向量化检索匹配
- 在非常垂直类，问题和答案非常固定的应用场景下可以开精确匹配
- 在泛业务场景下开启结果缓存可能会降低推理精度或准确性，需要结合业务判断和考量

AI网关落地实践--解决模型服务高可用的问题

核心问题：公司的主力模型是PAI上部署的DS R1 671B，但GPU资源并不是基于流量峰值储备的，所以当高峰期时，DS服务会请求失败，有什么办法可以保证业务健壮性？

解法：有两种做法，并且可以搭配使用：

- 可以构建多个个兜底模型服务，如果要想保证模型一致，可以主力使用PAI上部署的，兜底使用百炼平台提供的。实现当PAI上部署的DS服务请求失败时，Fallback到百炼平台托管的DS R1 服务。从而保证业务的连续性和健壮性。
- 通过基于Tokens的限流策略，解决Burst流量，保护后端模型服务。

1

维护多个模型服务

➤ 无论是FC部署的，IDC部署的，还是百炼LLM API服务，都可以作为模型服务被维护在AI网关。

2

开启AI API限流策略

➤ AI API限流策略需要配合Redis实现，但是只需要开通Redis和在AI网关侧配置即可。

➤ 支持多种限流判断条件：
Header，Query参数，Cookie，消费者，客户端IP

3

开启AI API
Fallback策略

➤ AI API一键开启Fallback策略。

➤ 当主LLM服务出现异常后Fallback到指定的其他LLM服务。

➤ 支持配置多个Fallback模型服务。

AI 网关落地实践--解决安全合规的问题

核心问题：模型托管平台自带好几层内容安全审核机制，但是我们在IDC部署或者在FC部署的，如何能方便的接入内容安全审核服务？

解法：AI网关中的AI API集成了阿里云的内容安全防护和AI安全护栏服务，可以一键开启。安全防护的规则还是要在内容安全服务侧配置。

- 支持请求内容检测。
- 支持响应内容检测。

API详情

策略与插件

消费者认证

使用指南

统计

日志

路由列表

预置策略与插件

内容安全防护

* 防护服务

https://cn-beijing.aliyuncs.com

默认填充内容安全的URL，您可以[参考文档](#)

防护服务

内容安全

AI 安全护栏

检查请求

开启

关闭

检查响应

开启

关闭

检查响应用于检查大模型的回答内容是否合规，生效时会使流式响应变为非流式

防护等级

中

保存

取消

AI 网关落地实践—联网搜索解决大模型不满血的问题

核心问题：公司部署了DeepSeek R1 671B的模型，但推理的结果和DS官网推理的结果有差距，似乎不满血？

解法：推理的结果和DS官网推理的结果有差距大概率是因为DS官网开启了联网搜索。DeepSeek R1 671B的模型推理能力是很强，但训练的数据也是有限的，所以要解决幻觉还需是要在推理前先搜索和处理出比较确切的信息后，再由DS R1推理，所以联网搜索是非常关键的。目前模型托管平台提供的DS R1 API和自己部署的DS R1都需要自己实现联网搜索。

联网搜索

* 搜索引擎

夸克

服务状态

试用中 正式开通流程说明

搜索配置

* API Key

显示 申请 API Key

返回结果数量

10

超时时间

3000

ms

查询时间范围

1周内

行业 (可选)

请选择

结果渲染

默认语言

中文

输出引用来源

否

内容类型

正文

引用格式

资料引用:
%s

填写示例

提示模版

Here are the search results for your question:

{search_results}

Your question was: {question}

填写示例

1

支持夸克联网搜索

- AI网关在AI API维度集成了夸克和必应的联网搜索能力
- AI API策略中一键开启，快速配置

2

搜索结果自动融合

- 搜索策略有多种配置项。
- 搜索结果自动融合进输入的Prompt，无需用户额外处理。

3

问题意图识别

- 默认使用小模型对用户的问题做意图识别，避免无效的联网搜索



AI 网关落地实践- MCP server的统一代理和鉴权

核心问题：建设企业内部的 MCP Marketplace，能够将企业自有的工具以 MCP server 的形式提供给大模型调用，同时能够将企业管理的外部 API（如地图 API 等）以 MCP server 的方式提供给内部 AI Agent 使用，且能支持对外部工具鉴权的统一收口和二次签发。

解法：通过 AI 网关（with MCP），为 AI Agent 调用工具，提供统一的接入点和管理界面，支持对接各类内外部 API 和运行时，能够在网关上统一对 MCP 调用进行观测和治理，从而降低 AI Agent 使用工具的门槛，提升企业 AI 应用的安全性和稳定性。

包含 3 种接入形式：

- 直接代理 MCP servers
- 接入 MCP servers 运行时（如 FC）
- adapt 用户 API 或服务为 MCP servers

AI 网关 / 实例 / aigw-demo / MCP 管理

← aigw-demo

概览

LLM API

MCP 管理

Agent API

服务

插件

安全管理

观测分析

参数配置

创建 MCP 服务

同步 Nacos MCP 服务

搜索MCP服务名称

发布状态 全部

全部服务 8

native-mcp-get-time

已发布

直接代理

get-express-info

已发布

HTTP 转 MCP

kuaidi-mcp-demo

已发布

HTTP 转 MCP

http2mcp

已发布

HTTP 转 MCP

express

已发布

HTTP 转 MCP

stock

已发布

直接代理

amap

已发布

直接代理

doc

已发布

HTTP 转 MCP

编辑工具

编辑方式

自定义YAML

手动创建 MCP 工具

基于Swagger3

通过上传 Swag

MCP 工具列表

后端服务认证

认证类型

凭证

添加

```
1 server:
2   name: get-express-info
3 tools:
4   - name: logistics
5     description: 获取物流信息
6     args:
7       - name: orderId
8         description: 订单ID, 订单号
9         type: string
10        required: true
11    requestTemplate:
12      url: http://express-svppmj rxyt.cn-hangzhou.fcapp.run:80/logistics
13      method: GET
14      argsToUrlParam: true
15    responseTemplate:
16      prependBody: |
17        以下是该订单的详细物流信息，以JSON格式返回。字段说明:
18        - **logisticsId**: 物流ID
19        - **orderId**: 订单ID
20        - **status**: 物流状态
21        - **history**: 物流跟踪信息
22        - **location**: 地点
23        - **time**: 时间
24        - **description**: 物流详细信息
25
```



AI 网关落地实践- AI agent的鉴权、观测

核心问题：公司当前已经在使用百炼应用、dify等平台构建 AI Agent，不支持二次 API-KEY 签发和更细力度的限流等策略，也不支持 openAI 的兼容协议。希望能够统一代理和管理 AI agent平台，包括托管 API-KEY、限流等

解法：通过 AI 网关的 Agent API，统一代理 Agent，实现 Agent 对外的二次鉴权，并完成观测、限流、fallback 和安全防护等治理动作，提升 Agent 的稳定性和安全性。

阿里云

工作台

华北2（北京）

搜索

费用ICP备案企业支持工单购物车通知帮助简中

API网关 / 实例管理 / gateway_0425

gateway_0425

概览

LLM API

MCP Server

Agent API

服务

插件

安全管理

观测分析

业务监控

资源监控

日志中心

事件中心

链路追踪

AI 诊断

参数配置

创建 Agent API

请输入要检索的Agent API名称

API名称	协议
test-agentAPI-0601	百炼
test-agentAPI-0602	Coze
test-agentAPI-0603	Dify
test-agentAPI-0604	百炼

创建 Agent API

基本信息

* API 名称

test-agentAPI-0605

名称唯一，支持英文、数字、下划线“_”，不超过63个字符

* 域名

test-aliyun.com × test-alibaba.com ×

创建域名

* Base Path

/base

* 协议

阿里云百炼

AI请求观测

开启

开启后可查看AI请求的Metrics、Logging、Tracing信息 Logigng和Tracing依赖SLS日志投递服务，请前往 开启日志投递

当您启用AI请求观测功能时，系统网关将自动记录所有AI请求的内容（包括请求体body）至访问日志中。请您务必确保所使用的SLS（日志服务）配置恰当，并采取必要的安全措施以保护日志内容的安全性，防止任何潜在的信息泄露或其他形式的信息安全风险。

高级配置

* 资源组

resource_0089dhudu12

创建资源组

消费者认证

消费者认证

开启

认证方式

API Key

JWT

确定

取消

AI 网关应用场景及案例

New Future on Cloud New Future on Cloud New Future on Cloud

04

AI 网关落地场景及案例

场景一：在FC/PAI/ECS上自建大模型服务，通过AI网关提升自建大模型服务的稳定性

客户画像

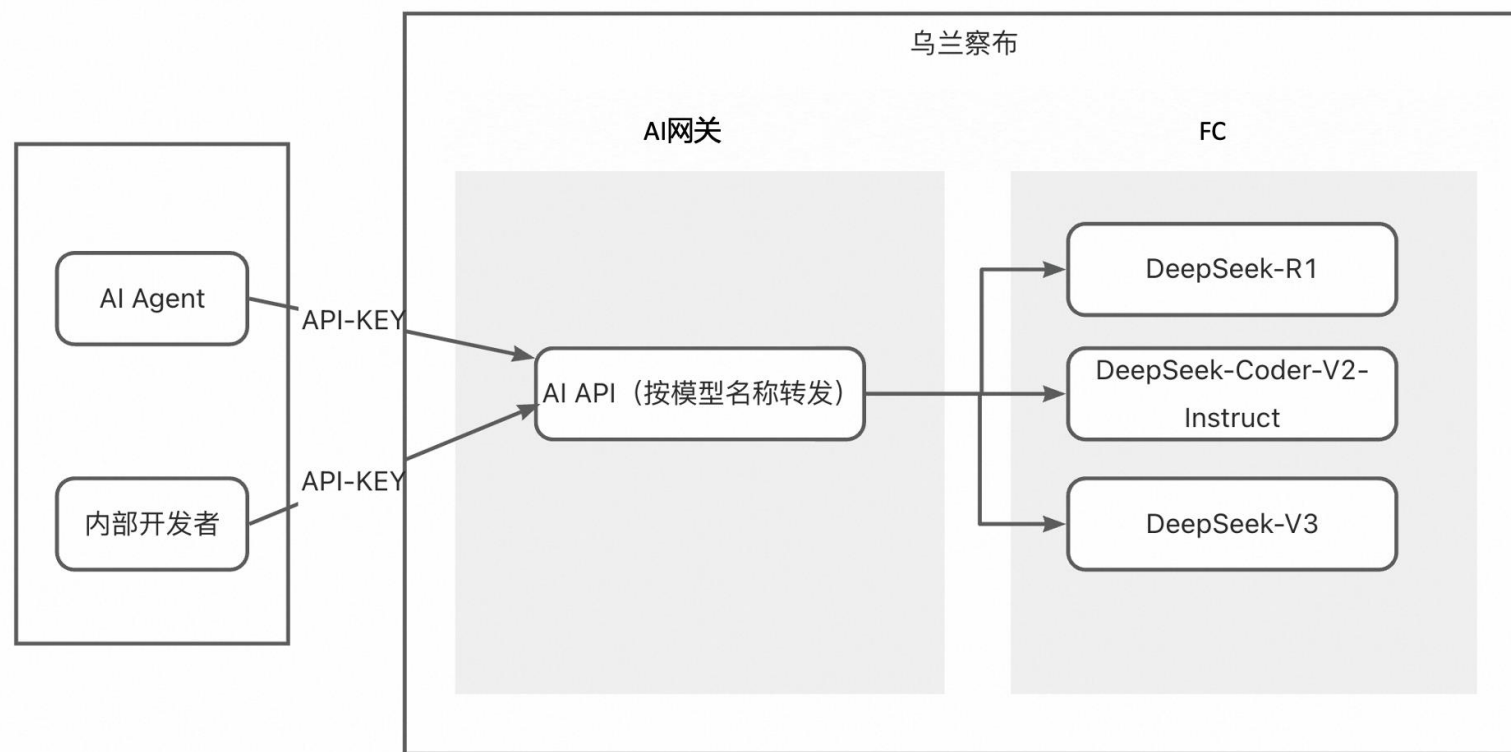
在阿里云FC/PAI/ECS等自建大模型服务的客户

客户问题

客户在阿里云上自建的大模型服务经常因为业务方大量调用而导致不可用，同时难以定位异常调用方。

解决方案

客户在阿里云上通过AI网关接入自建模型，在网关侧对业务方进行API-KEY授权、token限流和统一观测，在自建模型不可用时自动fallback到百炼等服务上。



代表客户：地平线

场景二：使用包括百炼在内的多家大模型服务，通过AI网关实现按openai的API标准的统一接入和统一管控

客户画像

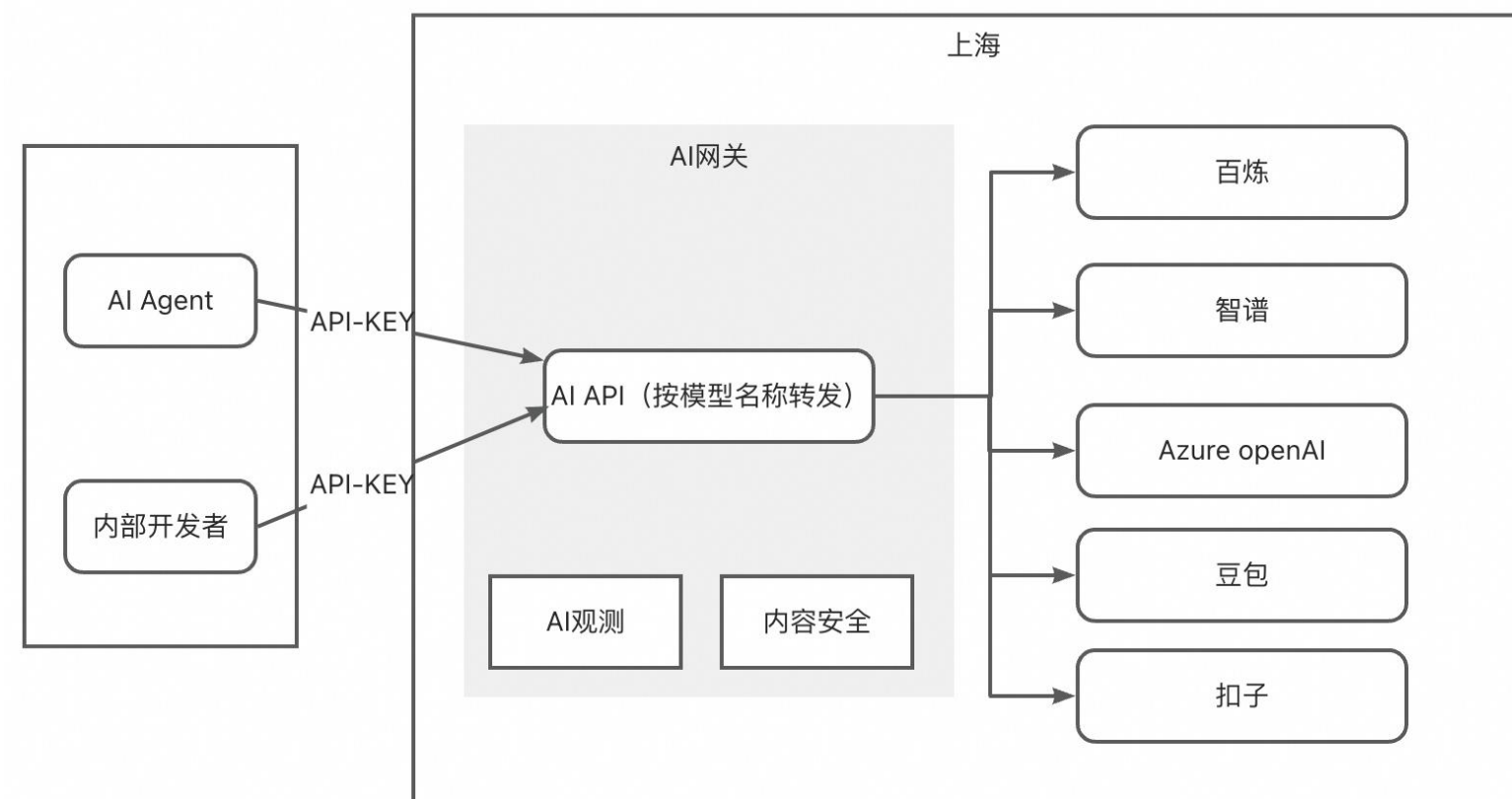
同时使用多个供应商的大模型服务，需要满足统一接入、管控和安全诉求的客户。

客户问题

不同供应商的模型的调用方式不同，需要业务侧适配，切换成本高；API-KEY管理不当会造成异常费用；安全和合规风险较高。

解决方案

客户在阿里云上通过AI网关统一接入各个大模型供应商，通过网关统一为openai标准。在网关侧对业务方进行授权、限流、安全检测并记录访问日志。



代表客户：国泰产险

场景三：通过AI网关建设企业自有MCP Server、agent市场，并对MCP Server、agent进行统一管控

客户画像

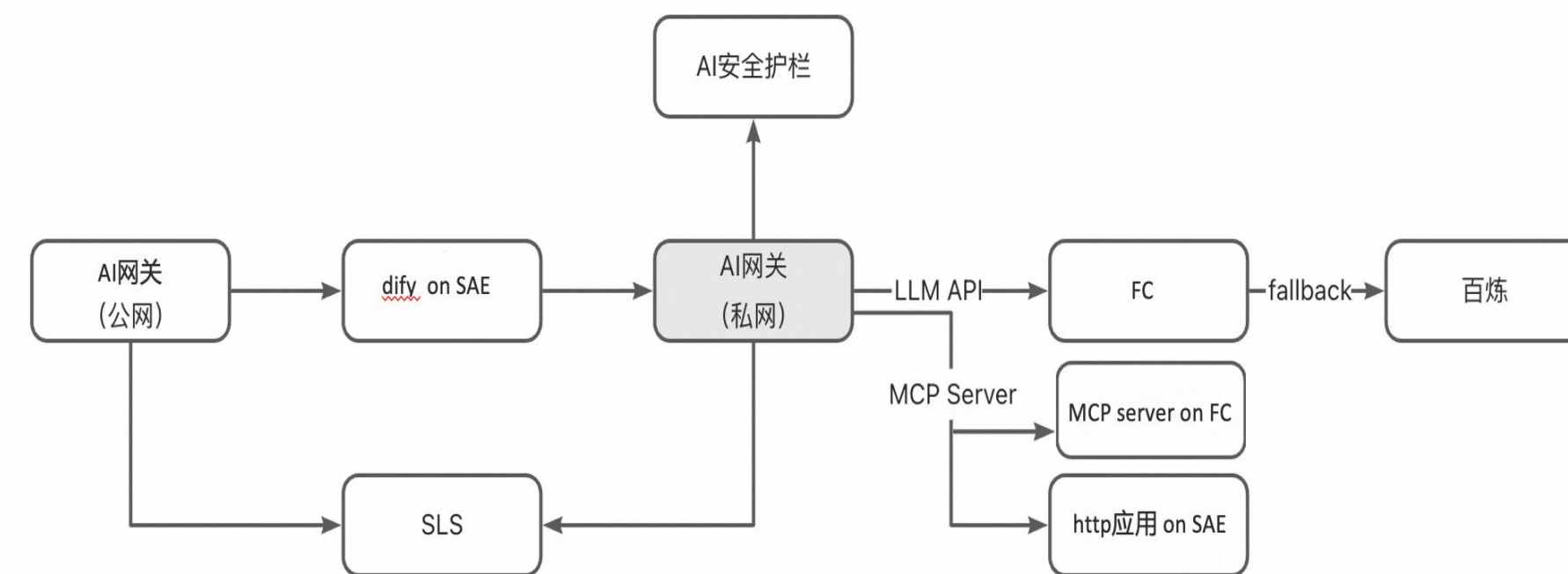
通过dify等自建AI应用平台，通过AI应用进行业务创新，同时有访问大模型和访问工具的诉求。

客户问题

存量的服务无法被AI应用方便地使用，缺乏对AI应用访问工具和大模型的统一管控。

解决方案

客户在阿里云上通过AI网关统一管理MCP Server和大模型调用。在网关侧自动将存量应用转换为MCP Server，对AI应用访问大模型和MCP Server进行授权和观测，并根据观测数据对各个业务方进行计量和ROI评估。



代表客户：某大型物流公司

THANKS