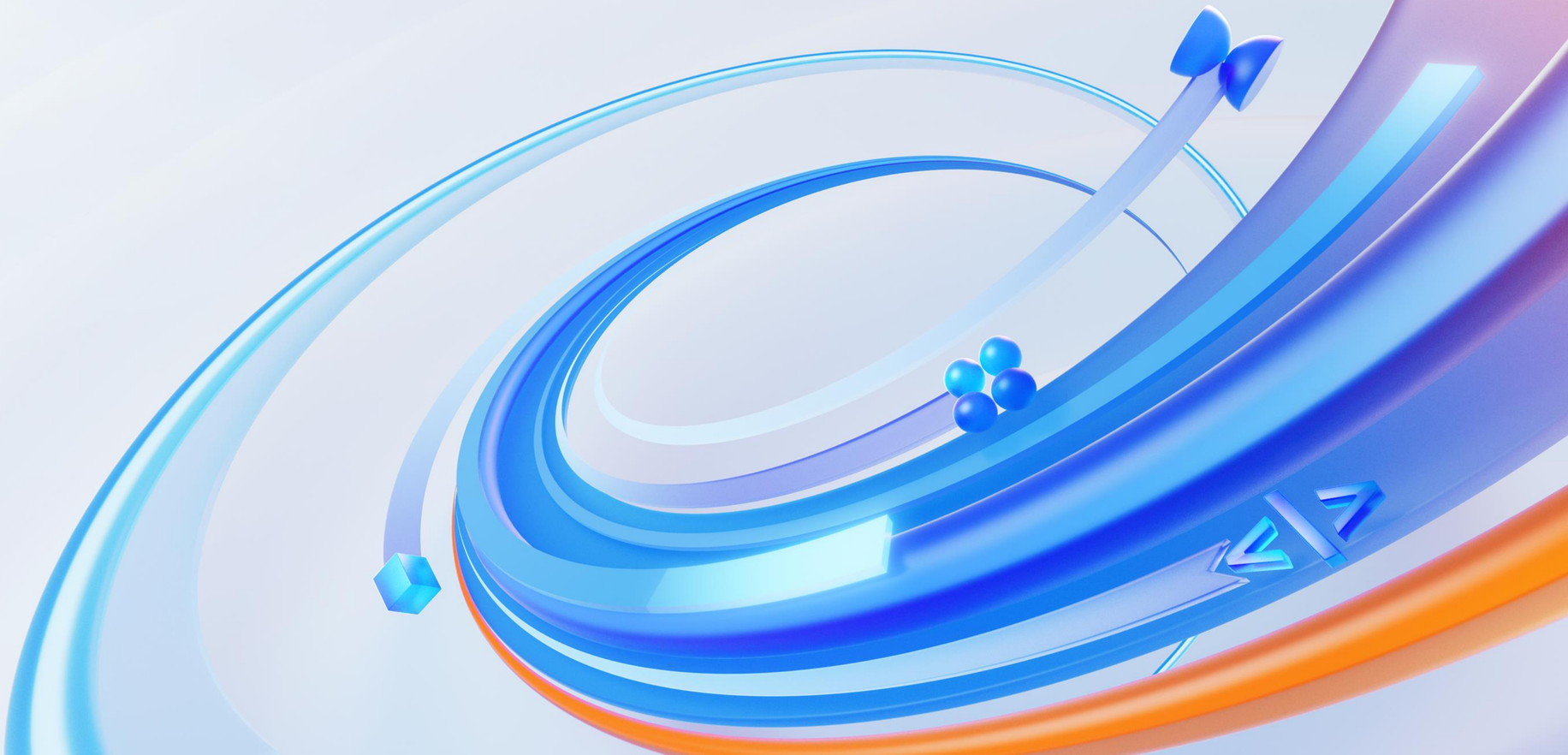
企业级 AI 应用开发：从技术选型到生产落地

黛忻

阿里云 Serverless AI 团队



阿里云

Contents
目录

阿里云

企业级 AI 应用开发运行时选型

01

阿里云

AI 原生范式对基础设施提出全新的要求

构建支持 AI Agent 的高效基础设施

Agent-Centric

基础设施的核心服务对象从“人类用户”转变为“自主 Agent”，以 Agent 而非服务或 API 为中心

以 Agent 为中心

State-First

状态是 Agent 的“记忆”与“人格”载体，基础设施必须原生支持状态的持久化、低延迟访问与跨环境迁移

状态优先

Task-Driven Orchestration

基础设施主动协调 Agent 完成目标，而非被动响应请求，Agent 和 Agent 或者 Agent 和工具之间的协作依靠事件驱动和动态弹性

任务驱动协作

Embrace Uncertainty

承认 LLM 输出的非确定性，通过基础设施能力降低风险，而非追求绝对可控，从“防御性编程”转向“容错自愈”

接受不确定性

AI 时代开发者关注业务创新而非基础设施

Serverless 是 AI 原生架构的最短实现路径

静态资源分配

预留固定资源，利用率低，无弹性或分钟级弹性

状态管理复杂

依赖 Redis/DB 管理，故障恢复复杂，跨节点迁移几乎不可行

API 驱动

依赖网关/MQ实现同步和异步调用或，微服务架构协作复杂度高

运维负担沉重

依赖专业运维团队配置集群、环境，自建监控日志代价高

✕ 从传统架构到 AI 原生架构

✓ 从 Serverless 架构到 AI 原生架构

动态弹性

按请求弹性，毫秒/秒级供给资源，支持低峰缩 0 弹性降本

状态持久化

内置状态持久化或集成工作流，支持快照或会话管理，自带 3AZ 容灾

事件驱动

通过云产品事件驱动自动执行，架构原子化解耦，协作更容易

免运维

安全隔离的多语言运行环境，自带监控日志告警等配套设施

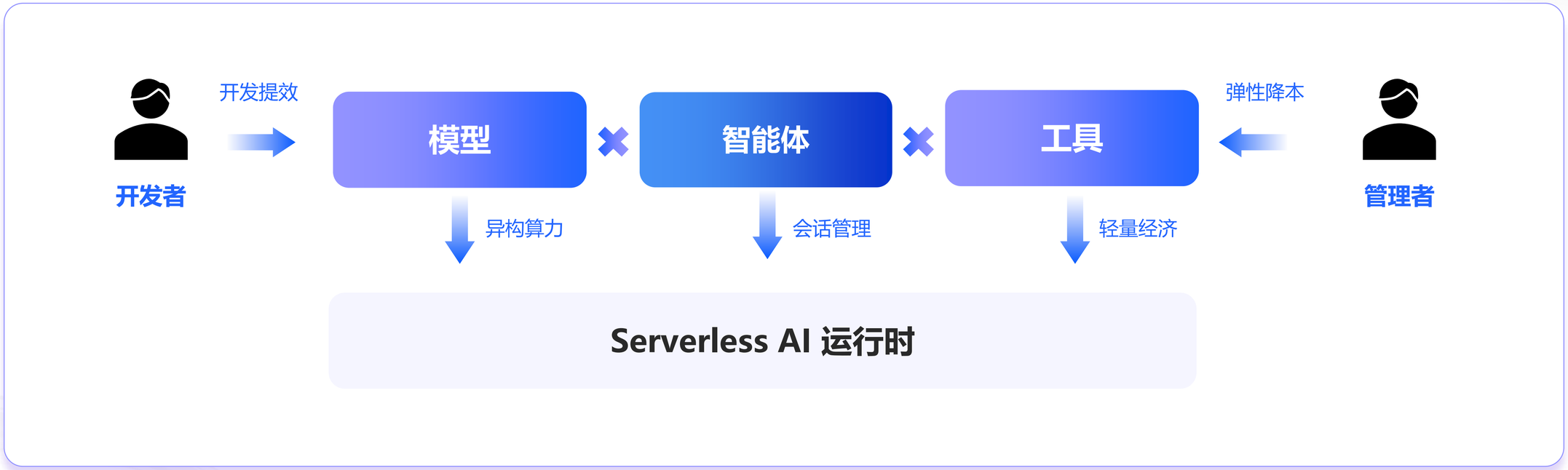
VM/容器运行时

AI 时代需求

Serverless 运行时

从 Serverless 到 Serverless AI

Serverless AI 运行时是 AI 原生应用的最佳选择



Serverless AI 运行时关键技术

02

阿里云

函数计算 FC：Serverless AI 运行时

0 运维、轻量、经济、弹性

异构算力 —— Serverless AI 运行时 —— 生态集成

AI 应用开箱即用

模型服务	图像生成	智能体 Agent	AI 工具
Qwen, DeepSeek, etc.	ComfyUI, SD, etc.	低代码编排、高代码编码	MCP & Sandbox

内置多语言运行环境

Python 3.6~3.11 | Node.js 17~22 | Java 8~21 | PHP/Go/.NET/... | 自定义运行时 | 自定义镜像

函数类型

事件函数 | Web函数 | 任务函数 | GPU 函数

Serverless 平台底座

高密部署，主动调度	安全隔离，自动容灾	快速交付，极致弹性	产品集成，事件驱动
第四代神龙架构	沙箱容器 2.0	资源池化 2.0	云产品 Serverless 化

异构算力

CPU 算力（x86） | GPU 算力（Tesla, Ada, Hopper, etc.） | XPU 算力

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

5倍 规格粒度

最小规格：FC 0.05C128MB，

虚机 1C512MB，容器 0.25C512MB

0 不使用不计费

按请求调度，毫/秒计费，低峰自动缩 0

虚机包月浪费多，容器为集群持续付费

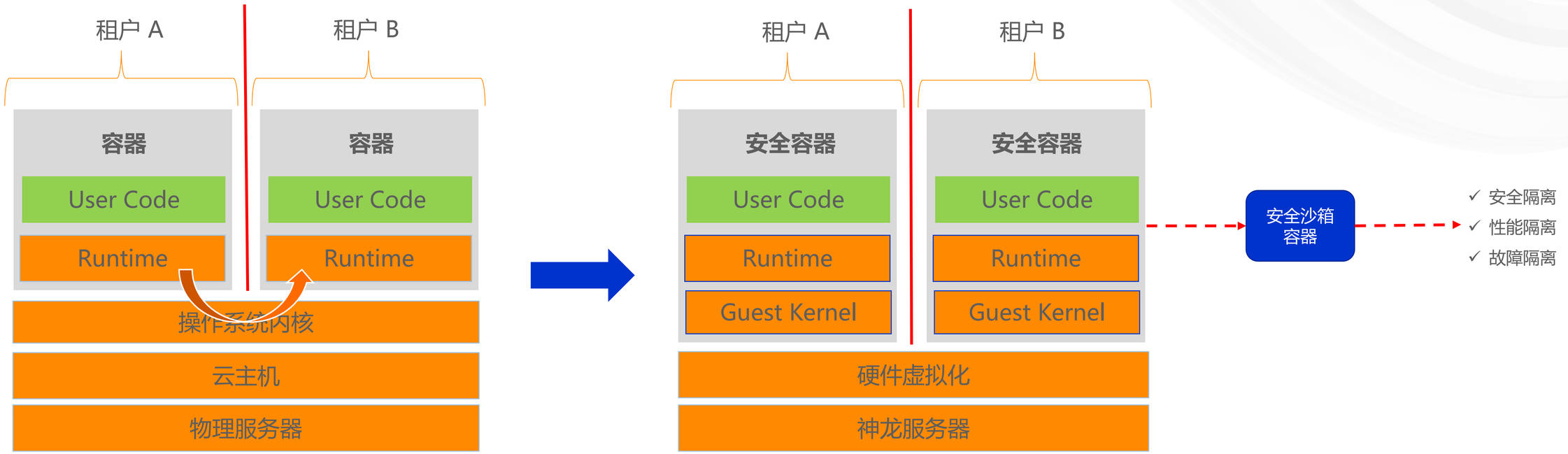
FC 不为 3AZ 容灾额外付费，虚机/容器则需额外付费！

50+ 内置环境

Python/Node/Java/PHP/Go/.NET 等 50+ 内置运行时环境，支持自定义运行时和自定义镜像，方便开发者灵活定制

数据来源：公开数据

Serverless AI 运行时安全 —— 资源强隔离



传统容器技术

普通容器用内核提供的 namespace 和 cgroup 做资源限制和隔离（从机器上圈了一部分资源给容器用），在安全性上存在不足：

- 容器内的进程在宿主机上可以看到
- 容器和宿主机共用内核，可以对宿主机进行破坏

函数计算FC运行时

FC 安全容器安全加固策略（核心是**限制代码破坏范围**）：

- ✓ 安全容器提供基于**虚拟机级别**的隔离
- ✓ 函数调度**尽可能**调度到同一台神龙服务器
- ✓ 加固安全策略：端口封禁、命令行封禁等
- ✓ 组件裁减：精简不必要驱动和内核接口，启动速度更快、资源占用更少
- ✓ 实例回收：销毁重建，避免残留 /tmp目录、日志、环境变量、进程等

模型运行时关键技术

函数计算 Serverless GPU 相对虚拟/容器的核心优势：请求感知调度、毫秒级闲置唤醒、1/N卡切分使用、Serverless 混合调度

请求载感知调度

RT 抖动减少 80%，可缩容到0

独家优势，相比 HPA 抖动率明显下降

用户常驻资源池 +平台弹性资源池 混合调度

Severless 混合调度降本50%

用户成本降低 50% 以上

毫秒级闲置唤醒

百倍启动加速

以 15% 的成本换取冷启动分钟级->毫秒级

1/N卡切分使用

最大降本 93.75%

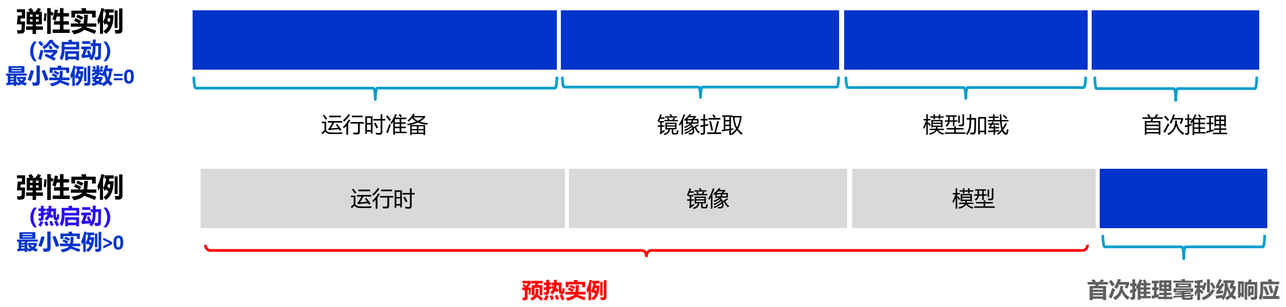
减少 GPU 碎片，优化稀缺资源利用率

powered by 基础软件 sCR，数据来源：测试得出

模型运行时：GPU冷启动优化

函数计算首推 Serverless GPU 启动快照，**实现毫秒级的首次推理响应，**0->1 首包耗时对比 K8s GPU，从分钟级优化至毫秒级

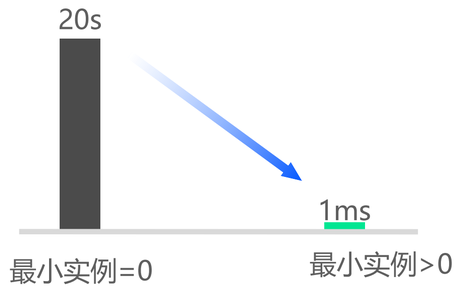
首次推理冷启动耗时分布示意图



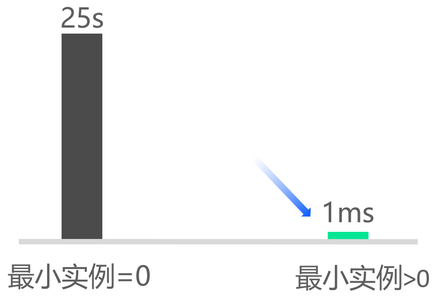
实时/准实时在线推理服务的痛点

1. 低时延：实时/准实时业务时延敏感，一般要求秒级响应，部分场景下需要毫秒级
2. 高并发：高峰期突增的吞吐量可能导致系统性能下降
3. 高成本：低峰期和小规格模型资源浪费，高峰期资源不足，成本优化难
4. 低容错：小流量推理场景单卡容灾能力差，故障率高

SD-v1-5-inpainting（4.27GB） 0->1 TTFI



Qwen-14B-Chat-Int4（9.01GB） 0->1 TTFI



Serverless GPU 的价值

客户案例

Rokid
AR 时代引领者

业务痛点

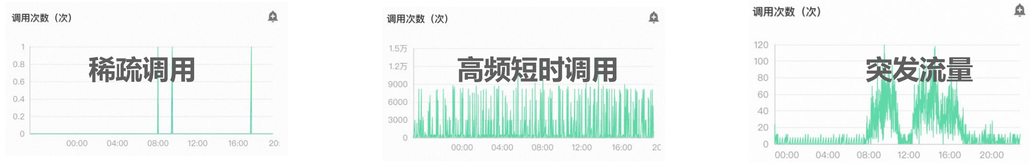
时延极度敏感 高峰期流量大 算力浪费较多

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



模型运行时：模型加载加速

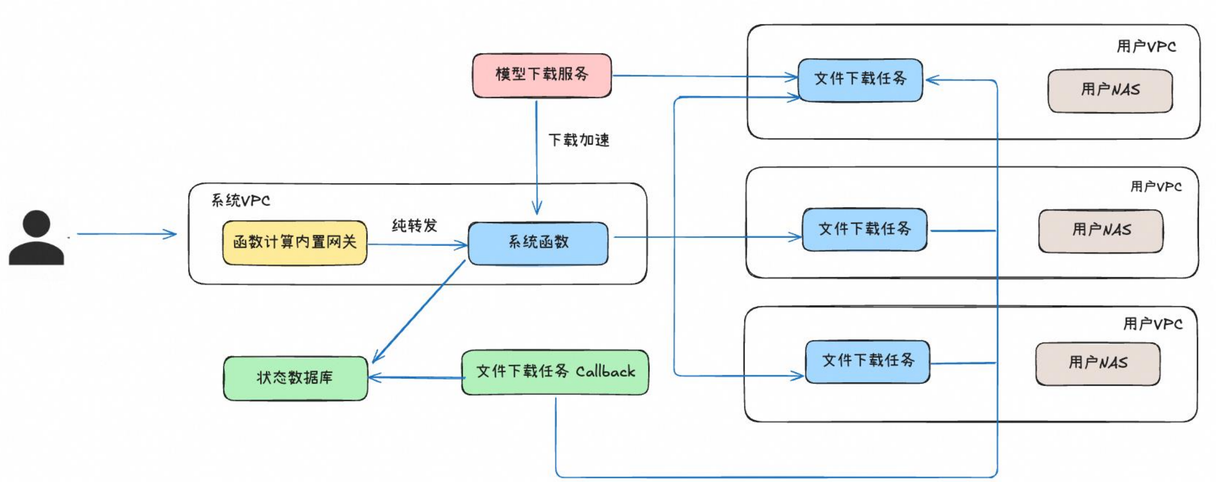
模型随容器镜像分发

适用场景：<1GB 的传统领域模型（CV / TTS），模型变更频率比较低
模型加载加速方案：镜像加速预热 + P2P 镜像分发

模型下载加速

适用场景：模型文件放在 OSS / NAS，应用程序通过挂载点访问。对模型大小没有限制。
- OSS：大量实例并行加载模型、需要本地冗余，或者多地域部署的场景。访问数量较少的大文件。
- NAS： 需要极速的启动性能。

模型加载加速方案：模型下载加速。函数计算用 OSS 缓存常用的模型，下载服务会自动判断系统是否缓存过，已缓存会走 OSS 内网下载。下载本身通过分片下载，多线程/多函数实例下载做了一些优化。



智能体/工具运行时关键技术

函数计算 FC：沙箱即服务、Session 亲和/隔离架构、毫秒级启动与按需付费

标杆客户	ModelScope　Qwen　阿里云百炼　行业头部厂商
沙箱即服务	Code Interpreter　Browser Use　RL Sandbox　Sim Sandbox 服务化 API，支持十万函数百万实例级别的沙箱执行
智能体运行时	Serverless 级 Session 亲和/隔离架构 负载感知调度，按会话弹性伸缩，支持会话亲和/会话隔离
Serverless异构算力	CPU　GPU 零运维，毫秒级启动，最大支持2w实例/分钟极速交付，免费提供 3AZ 自动容灾

业界领先的开箱即用、多语言代码安全执行引擎

沙箱即服务

服务化：提供 Code Interpreter API、Browser API

内置开发环境：Python/Node.js/Java/PHP/ Shell/.NET 等 50+ 多语言环境， 支持 OCI 标准镜像和自定义运行时灵活扩展

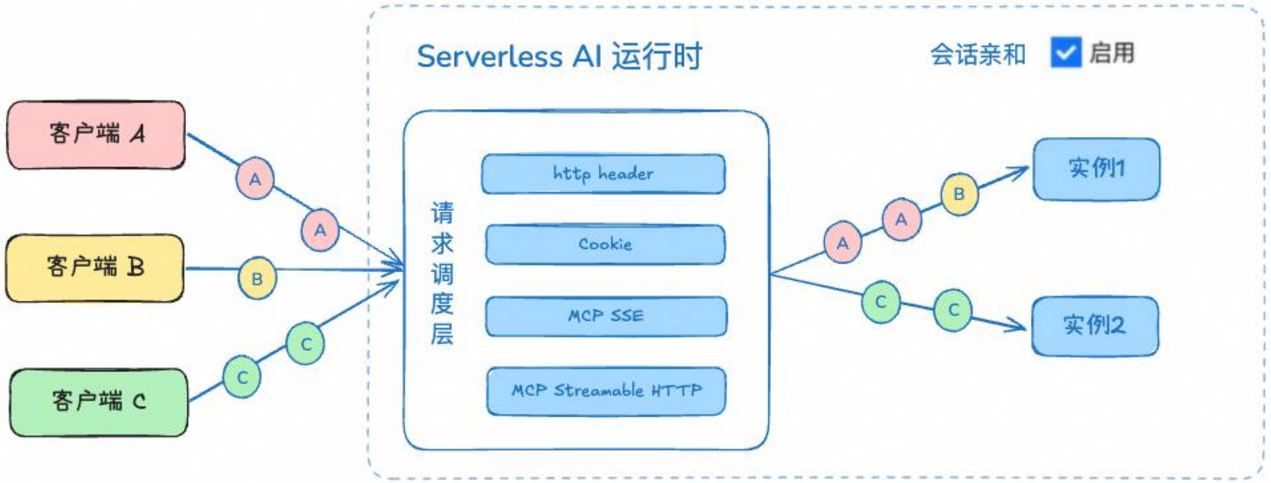
业界首创 Serverless 级 Session 亲和/隔离架构

智能体运行时

开源开放：与 AgentScope、LangChain、LlamaIndex 等主流开发框架集成

毫秒级启动与按需付费：强隔离、突破性上下文保持，启动效率领先传统容器方案 100 倍，按需使用，按量付费，低峰缩 0 成本最优

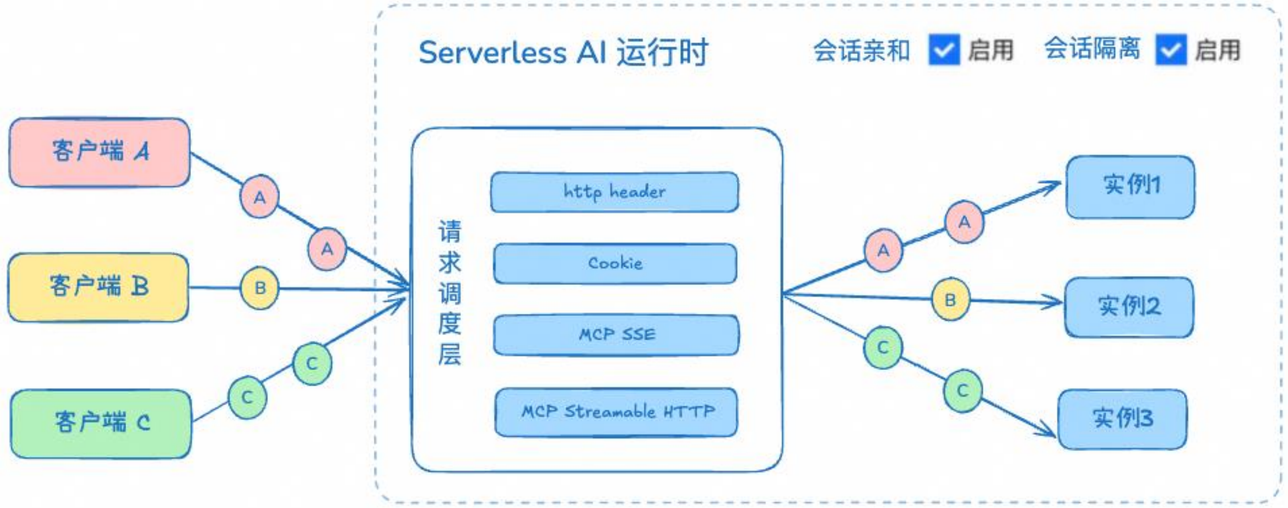
会话管理 — 会话亲和 & 隔离



会话亲和

MCP 场景强诉求。依赖连续会话状态，需确保同一用户请求始终路由至同一个实例，避免因实例切换导致SessionID 丢失和工具调用失败，保障多轮交互连续性与可靠性。

支持 MCP SSE / MCP Streamable HTTP/ Header Field /Cookie 四种亲和类型。

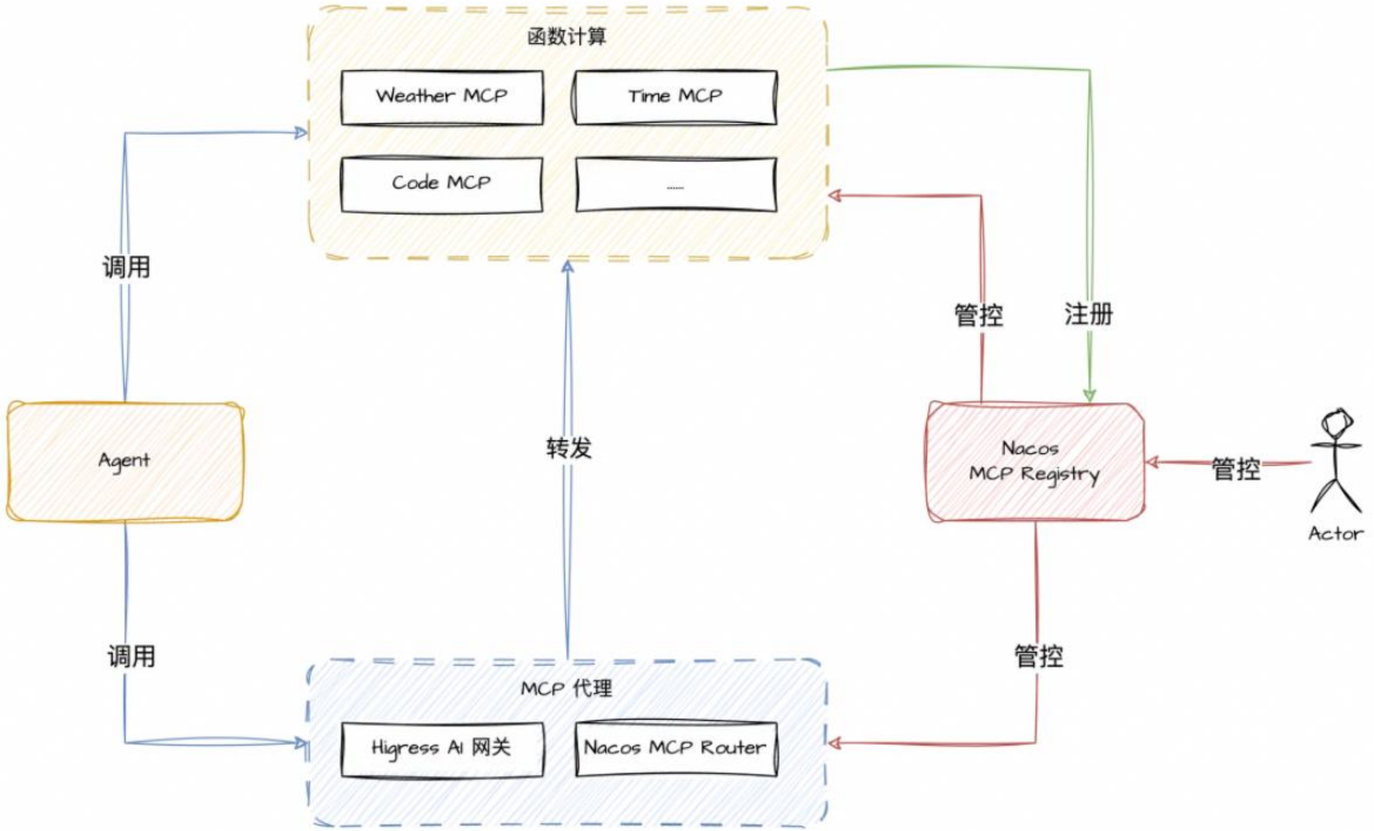


会话隔离

AI Sandbox场景强诉求。一个用户会话（Session）独占并绑定一个函数实例。避免不同请求之间的数据残留和数据泄漏。

函数计算 & MSE Nacos： 轻松部署和管控MCP Server

核心痛点：部署运维成本高，MCP调试迭代必须重启，服务分散难管理，缺乏动态管控能力



函数计算：免运维、低成本、弹性

- 免运维：无需关心基础设施，直接部署。
- 弹性伸缩：按请求自动扩缩，支持0-N。
- 低成本：按负载自动支持忙闲时计费，对稀疏调用非常友好。
- 协议兼容：支持SSE / STDIO / Steamable。

MSE Nacos：统一管控，可维护

- 自动服务注册发现：自动注册到指定MSE Nacos实例，统一服务发现，便于Agent 客户端/AI网关动态寻址与调用。
- 动态元信息管理：支持对工具描述、参数定义等元信息进行运行时更新，变更实时生效，无需重启服务。
- Tools 动态开关：对特定Tool启用禁用，实现故障隔离/灰度控制。
- 全链路集成：和 AI 网关、Nacos MCP router无缝对接。

AI 时代计费演进 —— 从请求驱动到价值驱动

阶段一：按资源租用计费

虚机 / 容器 计费模式

问题：为实例的持续运行付费。无请求时不能缩0仍计费，资源空转成本高。

阶段二：按请求计费

传统 FaaS 计费模式

问题：为代码运行时刻付费，无请求时 0 成本。但长连接保活场景（如 MCP Server/WebSocket）因低负载存活仍计费，成本高。

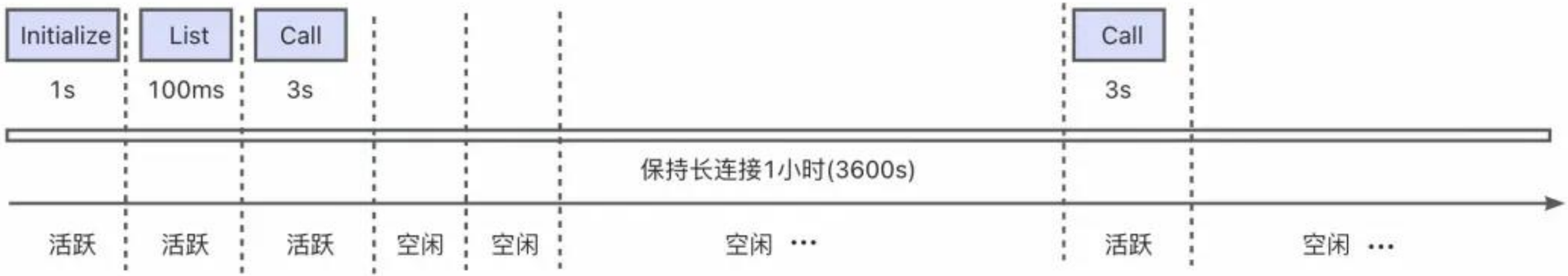
阶段三：按实际资源消耗计费

Serverless AI 计费模式

价值：按实际资源消耗，精准区分忙闲时计费。

消除长会话/低负载保活冗余成本，无缝支持 AI 强交互场景。

MCP Server 基于Serverless AI 的计费方案，长连接闲置计费最高降低 87%



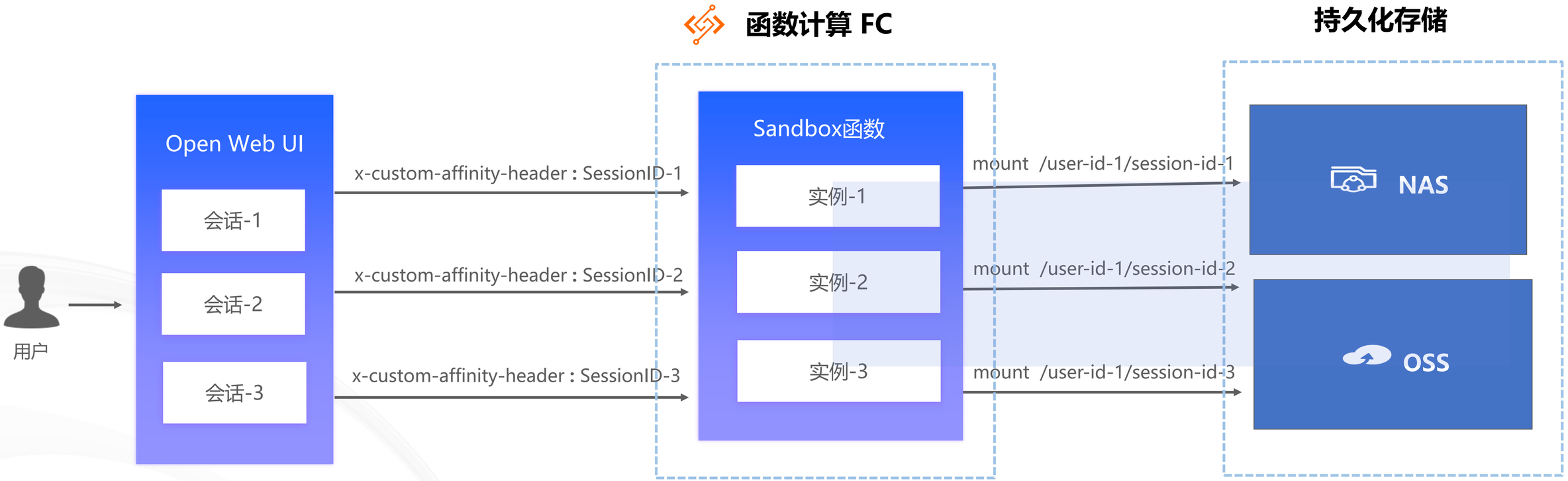
Sandbox 实例动态挂载 — 从计算隔离到存储隔离延伸

传统共享存储问题（虚机/容器/FaaS架构）

Agent Code Sandbox 多租户数据共享，有安全问题。无法满足同一个函数的每个实例路径不同的需求。挂载存储路径是变化不确定的。

Serverless AI 解决方案

- 引入会话粒度度存储粘性，将会话和一个持久化的，归属特定租户的存储子目录进行强绑定。
- 平台基于POSIX 标准多租存储安全实践框架，落地层次化纵深防御体系



阿里云

客户案例-
Serverelss + AI 让应用开发更简单

03

阿里内部案例 — 智能体/工具运行时最佳实践

魔搭社区、Qwen、百炼，大规模使用函数计算 FC 提供的 Serverless 运行时构建模型、智能体和 AI 工具

ModelScope × 函数计算 FC

模型

- 领域模型
- 大语言模型

2万+热门模型一键托管到云上环境

- ✓ 开箱即用，DevPod 二次开发，弹性交付 GPU，低峰缩 0
- ✓ 内置 vLLM/SGLang/Ollama/LMDeploy 推理框架
- ✓ 最快 30s 将开源模型转化为生产级 OpenAI 兼容 API

Serverless 模型运行时

Qwen × 函数计算 FC

智能体

- 高代码 Agent
- RL Sandbox

助力 Qwen3-Coder 成为领先的的代码模型

- ✓ 50+内置多语言运行环境，代码执行延迟 <100ms
- ✓ 会话亲和/隔离架构，满足 Agent/Sandbox 上下文保持需求
- ✓ 零运维、毫秒级弹性、按量付费的企业级安全沙箱，兼容 E2B

Serverless 智能体运行时

阿里云百炼 × 函数计算 FC

工具

- MCP Server
- Code Interpreter

业界首个全生命周期 MCP 服务市场

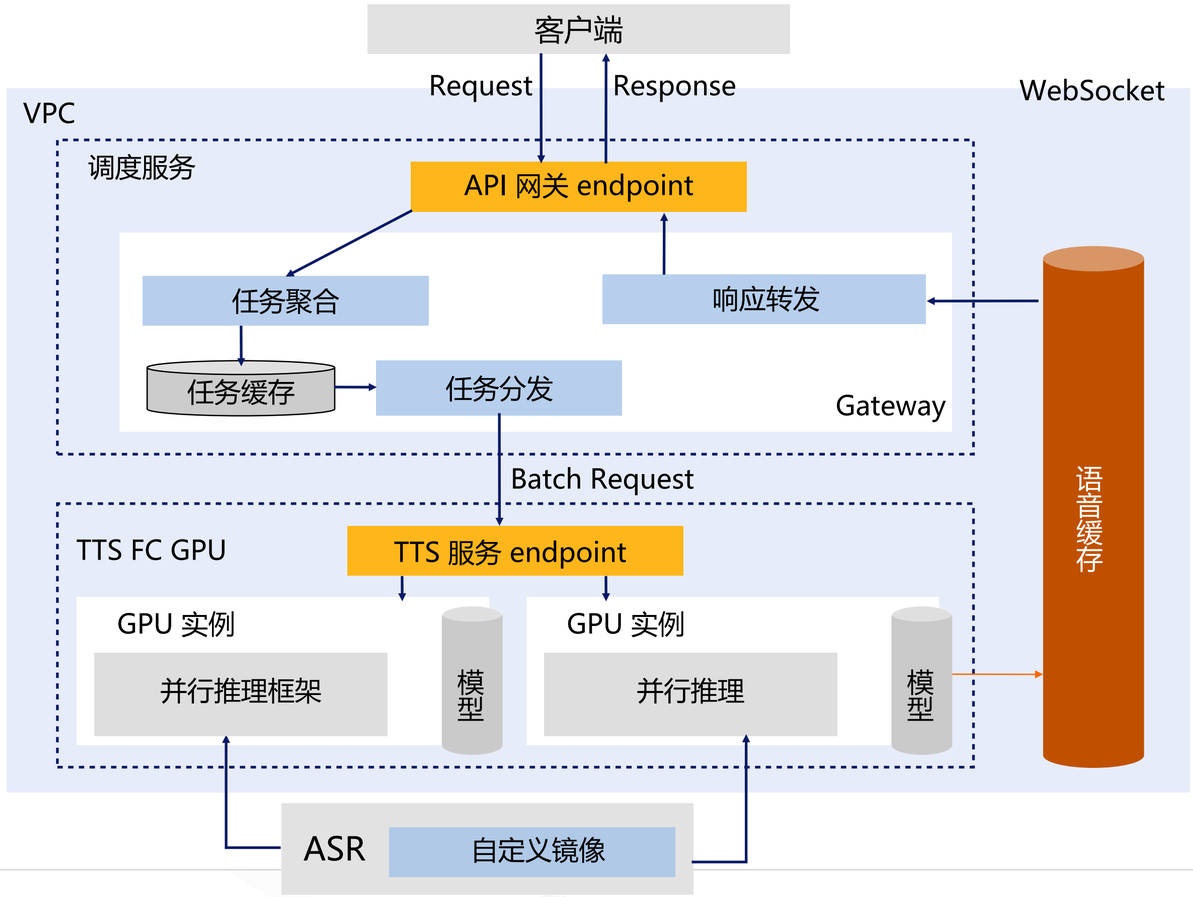
- ✓ 零改造将开源 STDIO MCP 升级为 SSE/Streamable HTTP
- ✓ 原子化解耦，MCP Server 冷热分离，弹性伸缩 RT 抖动少
- ✓ MCP SSE 亲和调度，闲置计费，低峰自动缩 0 成本最优

Serverless 工具运行时

Serverless 运行时已经成为阿里云 AI 原生应用的核心载体

实时/准实时推理场景— Serverless GPU 解决方案

函数计算给吉利 AI 座舱的交互和娱乐功能提供大规模推理服务，共同打造**大规模、高可用、高性能**的推理引擎。
场景覆盖：意图解析、文生图、情感TTS 等。



痛点 & 挑战

- 高性能：冷启动低延迟 & 模型预热、推理请求批量执行等。
- 低成本：提升 GPU 资源利用率
- 高可用：模型请求高可靠接入、推理服务高可靠、故障恢复。
- 故障恢复策略：快速定位和恢复。

解决方案

基于 FC+ 网关+可观测能力打造的 AI原生架构，具备端到端高可用、高性能、低时延、性价比高的算力供给。
预热实例+忙闲时自动计费、一站式集成、模型优雅上下线、多AZ等

客户价值

- 算力成本：相比之前 IaaS GPU 资源供给，FC 成本优化33%。
- 算力集群简化维护：降低运维压力，让业务专注模型和业务本身。
- 端到端高可用SLA 99.99%（持续建设中）：AI 推理引擎的端到端可观测、Trace追踪、灰度发布，SRE保障机制等。

模型服务 – 生产实践

Serverless x AI，函数计算 FC 帮助企业跨越技术鸿沟，让模型服务兼具 **SaaS 体验** 和 **自建优势**

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，希望开箱即用，无需操心软硬件维护，费用可控。

ComfyUI

降本	出图速度
78%	1图/10分钟 -> 1分钟/12图

开发者

少儿阅读 App 创新功能

客户原声

场景： 某朵科技是一家儿童阅读解决方案提者，引入2.5w+绘本童书，1.5w+音频内容，其利用生成式语音大模型开发的 AI 互动阅读解锁阅读新方式。

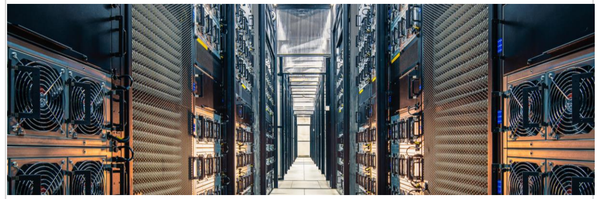
痛点： “idea”要快速验证，开发效率要高，特别是要避免繁琐的环境和资源维护

CosyVoice

降本	开发提效
1万/年 -> 0.5万/年	客户端开发 -> 全栈开发

初创公司

智能家居领域的潮流引领



客户原声

场景：某睿科技是一家聚焦于一站式智能家居视频解决方案的高新技术企业，通过 AI 的增强大大扩宽场景、提升体验。

痛点：视频业务有明显的峰谷特性，深受 SaaS 限流、昂贵且无法定制的困扰

Qwen/DeepSeek

降本	灵活定制
40%弹性不限流	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。

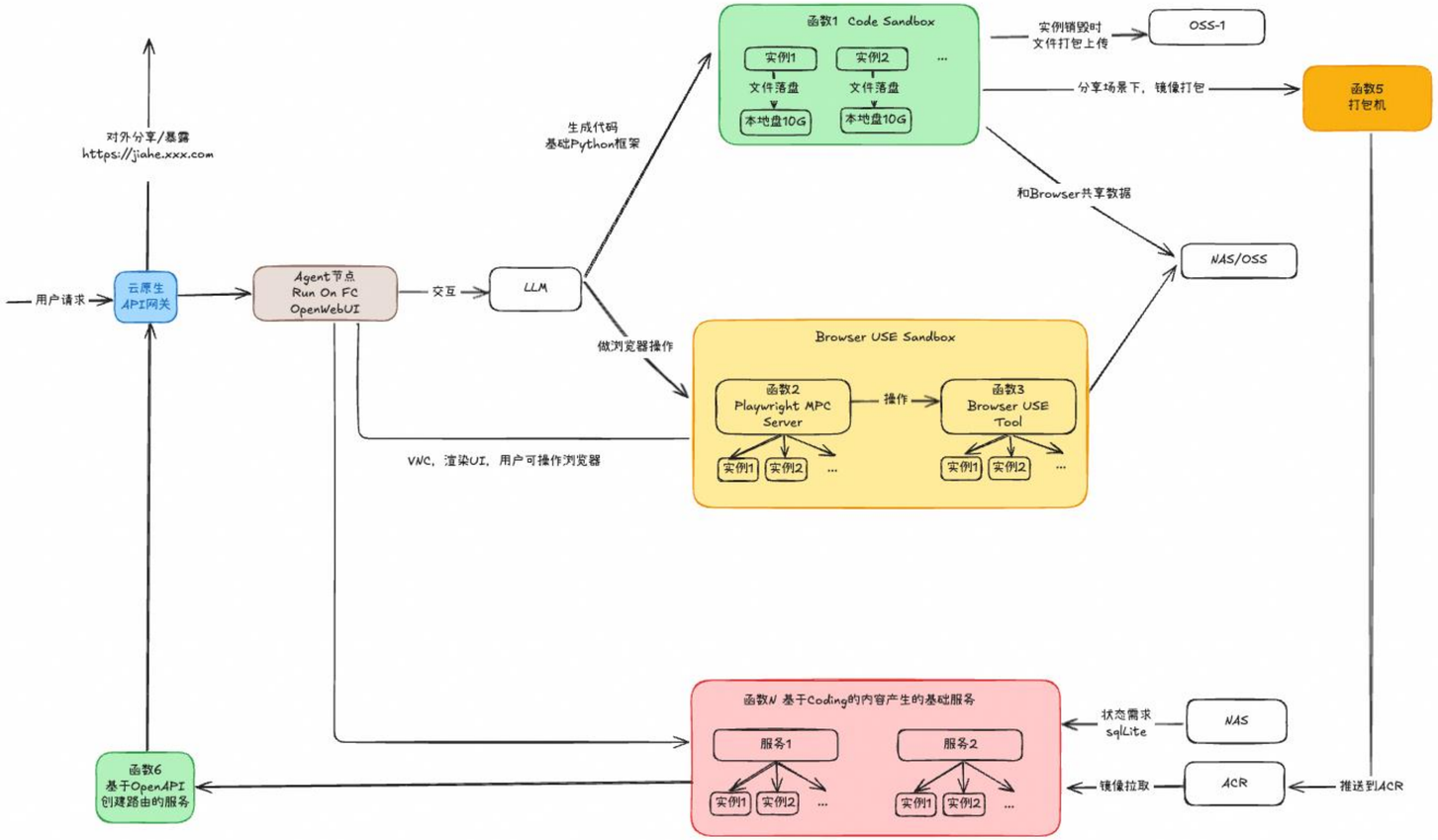
痛点： 适配多种开源大模型， 定制化要求高、开发周期长、GPU 持有浪费多

Stable Diffusion

降本	迭代效率
33%	3天 -> 30分

基于函数计算 FC Sandbox 构建智能体开发平台

千问强化学习　阿里云百炼　行业头部厂商



客户痛点

- 安全风险：AI 代码风险未知，需强隔离
- 成本压力：流量不可预期，资源常备成本高昂
- 弹性瓶颈：突发流量扩容不及时，错失爆款良机

方案价值

- **轻量级安全沙箱：**提供轻量、快速、虚拟机级别强隔离的执行环境。
- **Serverless降本：**缩容到 0 以及根据请求/CPU 真实负载的忙闲时智能计费，完美匹配 AI 应用的脉冲式调用，整体成本优化 60% 左右
- **极致弹性：**基于内存快照，1 毫秒内从挂起状态极速恢复会话
- **海量并发承载：**数十万级函数轻量化管理支持项目级别的分享，支持 C 端用户爆点项目的可扩展；百万级 Session 超高并发，满足 Agent 轻量灵活与极速弹性

阿里云

THANKS

Function AI：生成式 AI 的落地实践与案例分享

—— Serverless 到 Serverless AI 让 AI 应用开发更简单

世如(史明伟)

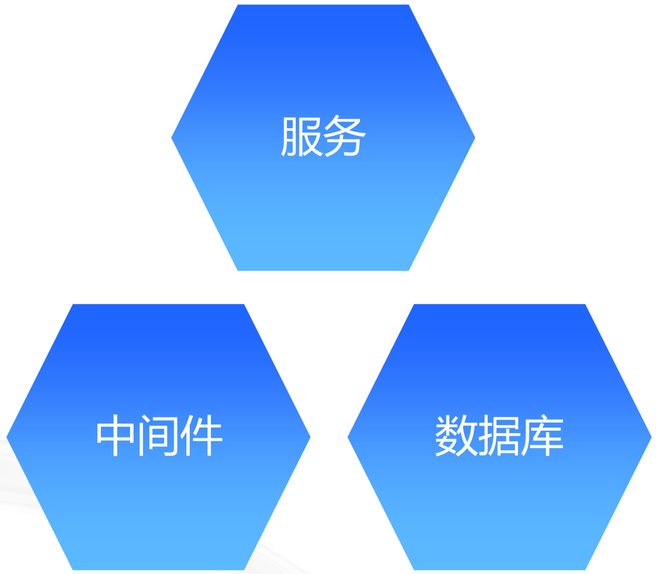
阿里云云原生—Serverless—函数计算&AI 研发负责人

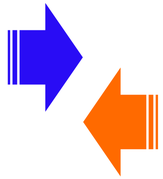
阿里云

Contents
目录

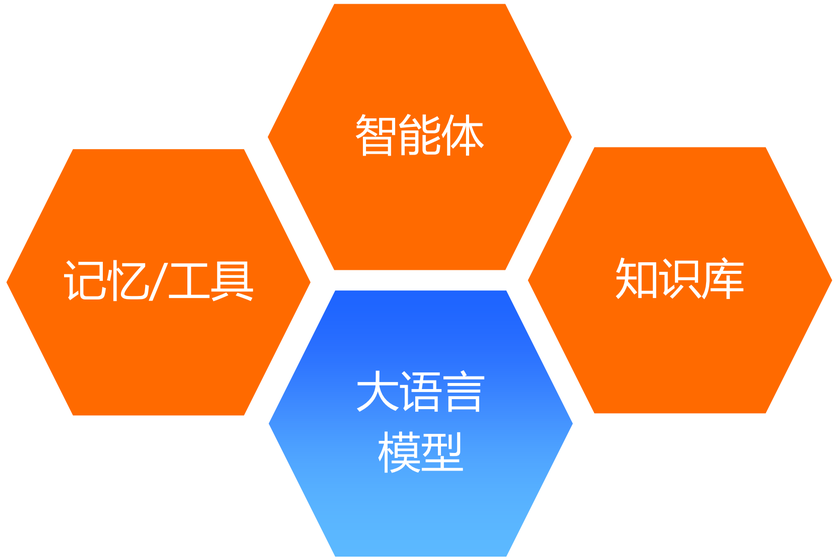
AI 应用组成核心抽象发生变化

传统应用开发





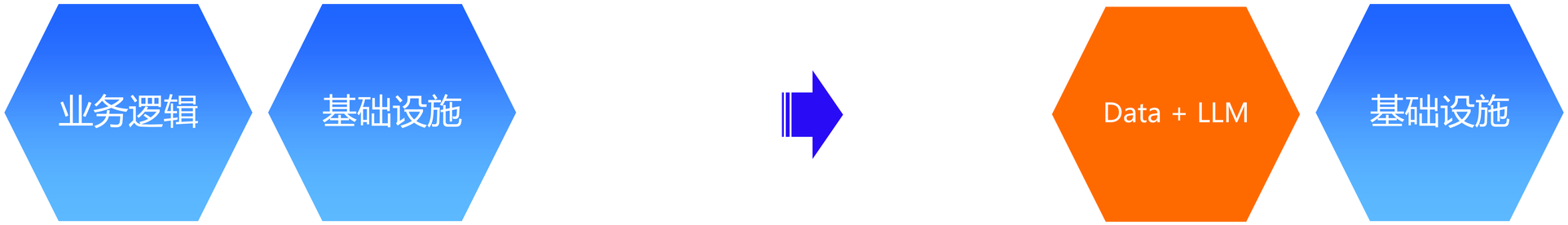
AI 应用开发



AI 应用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 业务优先要求基础设施更加敏捷高效

阿里云

Serverless AI 原生架构新范式

Building a New Paradigm of AI-Native Architecture on Serverless

AI应用可观测：云监控2.0针对AI应用提供全栈智能可观测能力

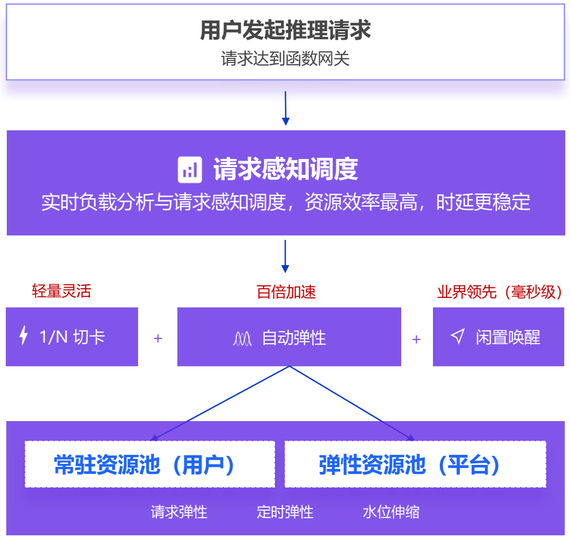
全栈Serverless **全栈高可用** **双层安全** **简单易用**

Function AI：从算力到应用，AI 全栈升级

魔搭社区、Qwen、百炼、PAI、Qoder，大规模使用函数计算 FC 提供的 Serverless AI 运行时构建模型、智能体和工具

FC GPU

1. 开箱即用的 Serverless GPU 服务，请求感知调度，缩容到 0，3AZ容灾
2. 整卡/切卡，细粒度资源灵活售卖；毫秒级闲置唤醒技术，冷启动快百倍；
3. 常驻资源池 + 弹性资源混池，FC 提供开箱即用的 Serverless 混合调度能力，更省成本，业界领先

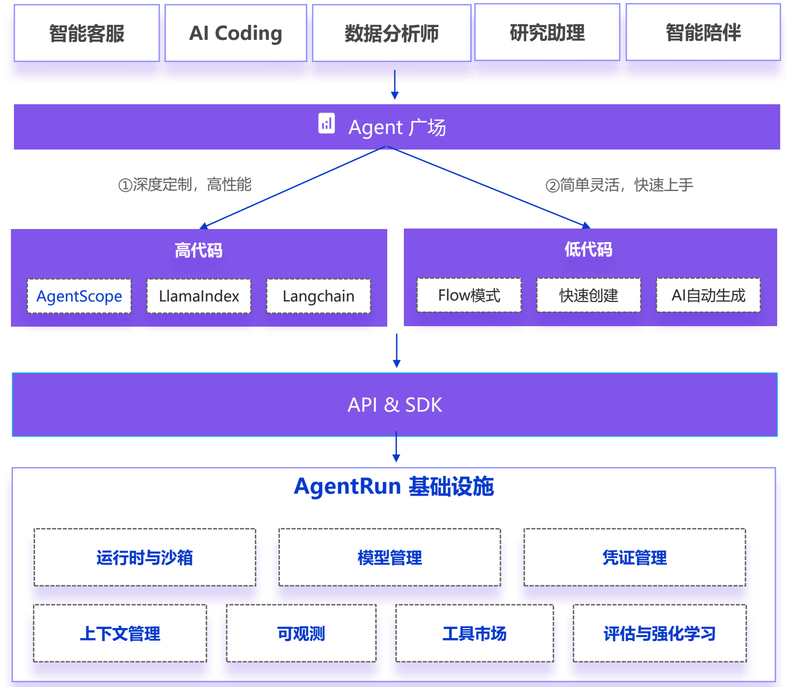


模型服务

1. 丰富的模型广场开箱即用，支持魔搭、HugginFace 以及自训练开源模型一键托管，Serverless GPU 成本更优
2. 内置 vLLM/SGLang 等框架，支持 DeepGPU/Nunchaku 推理加速
3. DevPod 模型开发/自动打镜像，灵活定制兼具 SaaS 体验和自建优势

AI 应用

1. Agentic AI 应用开发平台，低代码兼容 dify 性能提升 10 倍，高代码默认集成 AgentScope、Langchain、LammaIndex、CrewAI 等主流开发框架
2. 开箱即用、安全隔离、极致弹性的 Serverless AI 运行时与沙箱，平均 TCO 降 60%
3. 集成健康检查、日志收集、监控告警，减少 90% 的人工运维工作量



阿里云

Contents

目录

FunctionAI：让 AI 应用开发更简单

FunctionAI 一键创建应用

组装式开发，弹性开放，按需选择

模型托管	图像生成	Agent 开发	Agent 低代码
FunModel	FunArt	AgentRun	AIStudio

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

阿里云

模型服务：AI 模型一键转化为 Serverless API

Model Service: One-Click AI Model to Serverless API Transformation

开源模型一键部署，AI模型一键 Serverless 化，云端模型开发部署零门槛

模型类型

- 大语言模型
- 垂类模型

模型市场

- Modelscope
- HuggingFace

模型框架

- vLLM/SGLang/Ollama
- Transformers/PyTorch

自然语言处理 | 语音 | 计算机视觉 | 计算科学 | 多模态

模型服务

享受PaaS自由度的同时获得SaaS便捷性，按调用量付费，模型部署成本降低90%

100+热门模型在线体验
200+模型一键部署

6+模型框架一键托管
与Agent服务一键集成

Devpod 模型在线开发/构建
按量付费 CPU+GPU 集群

函数计算 – Serverless GPU

Serverless API

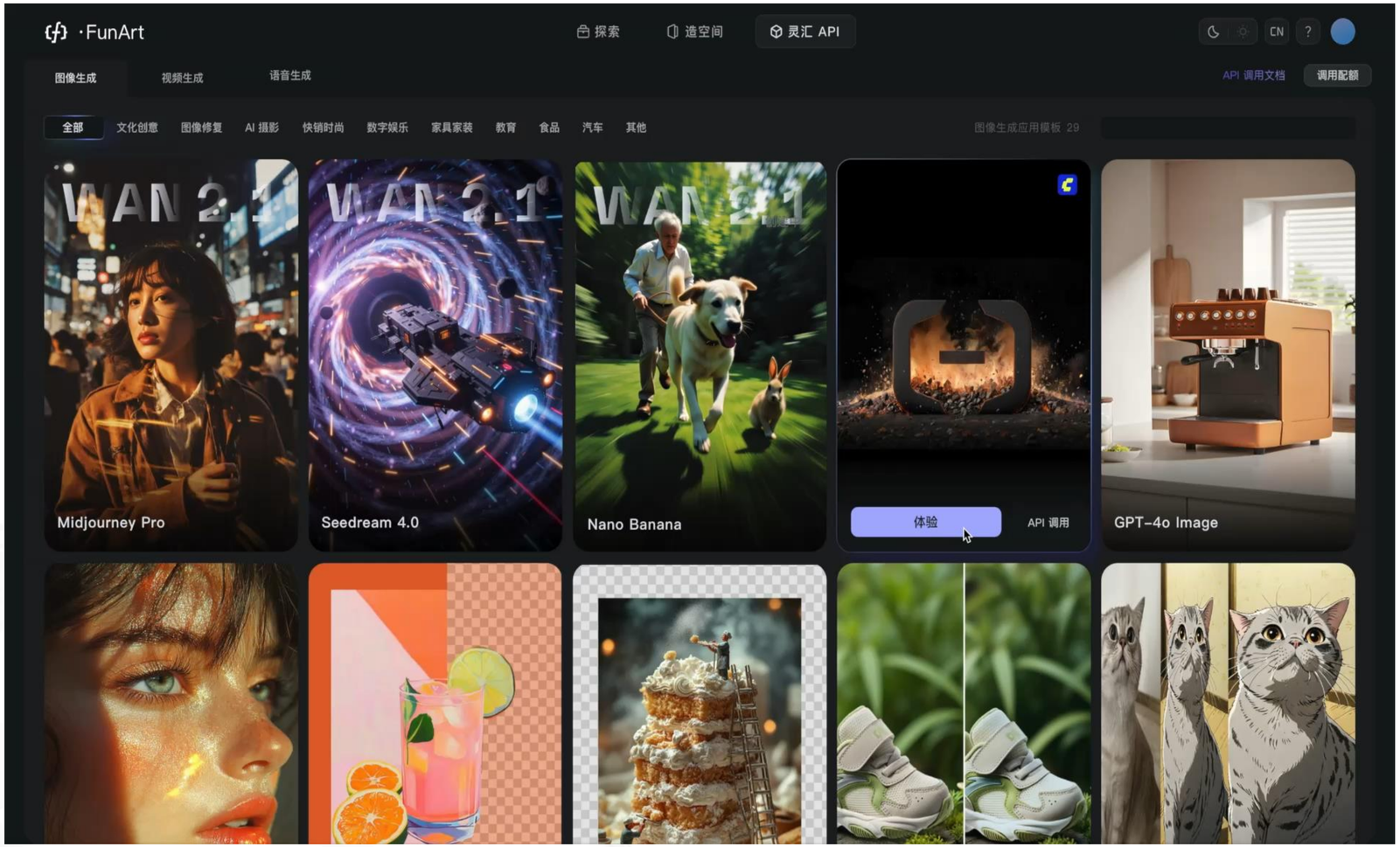
- 按量付费，降低成本
- 免运维/低运维
- 毫秒级弹性*
- 百万级实例规模*
- 与 ComfyUI/SD
  Dify 等生态一键联动

* 毫秒级弹性和百万规模集群紧针对部分模型的测试结果，并不代表全部模型都可以具备该能力

文生图/声音/视频：多模态AI创作平台

Text-to-Image/Audio/Video: Multimodal AI Creative Platform

多模态内容生成引擎，AI创作零门槛



文生图生图 / 文生视频

提供开箱即用的云端 ComfyUI，Stable Diffusion 工具，解决部署安装复杂，及本地环境显存限制问题。提供模型广场，插件广场，工作流广场加速业务开发。

CPU/GPU 分离（即将发布）

将开发和运行分离，开发态不再消耗 GPU 资源，由CPU资源负责用户的流程开发，只有调试执行生图操作时才调度GPU资源完成内容渲染。

LoRA 模型训练

提供 Muse LoRa，Kohya Lora 等多种 LoRA 训练器，解决从数据集打标到模型训练的全过程。

企业级 Serverless API 调用

高性价比的 GPU 弹性算力；支持一键将 ComfyUI 工作流、Stable Diffusion 发布为API，实现资源独享，自动弹性扩容，内置队列，支持异步调用

阿里云

AgentRun: Agentic AI 应用基础设施

AgentRun: Agentic AI Application Infrastructure

函数计算/Function AI 赋能企业 AI 应用高效开发与稳定运行

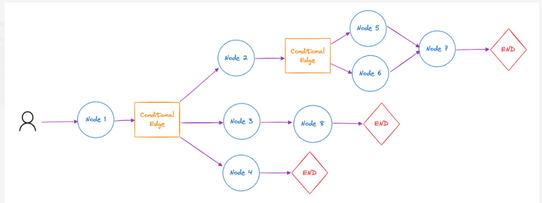
开发 | 部署 | 运维

开发

高代码深度定制，低代码快速搭建，全面提升 AI 应用开发效率

高代码

低代码



快速创建 | 自然语言 AI 生成 | 流程编排

百炼 | CloudFlow | Dify | ……

部署

轻量、安全隔离、极致弹性的 Serverless AI 运行时

开箱即用

Runtime			Sandbox			
Agent Runtime	Model Runtime	Tools Runtime	Code Interpreter	Browser Use	Compute Use	mobile Use

Serverless

会话亲和 | 安全隔离（请求/会话/函数） | 内置多语言执行引擎

轻量化函数 | 细粒度资源 | 毫秒级弹性 | GPU 算力解耦 & 1/N切分

异构算力

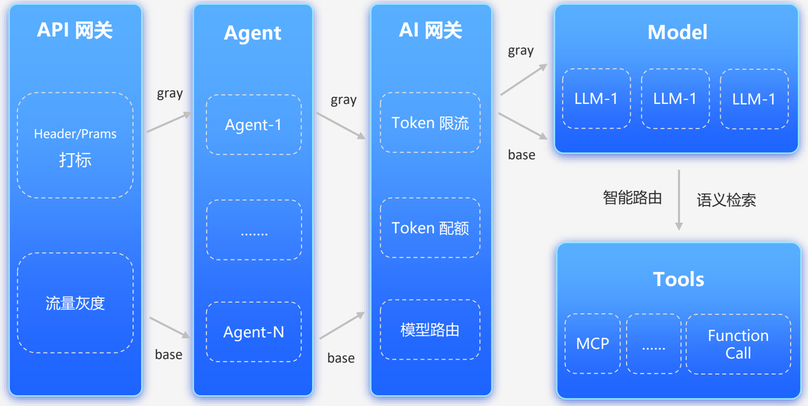
CPU 算力 | GPU 算力 | xPU 算力

函数计算 FC

连续三次入围《Forrester Wave™: Serverless Development Platforms》领导者象限
亚太区第一，全球前三

运维

服务治理与可观测，为 AI 应用保驾护航



全栈监控

智能体框架无侵入埋点和链路跟踪，全链路 Agent、LLM、网关、运行时指标和日志

AI/MI平台预集成

与 FC、容器，PAI，百炼默认集成，一键接入可观测数据采集与可观测大盘

成本管理及质量评估

使用评测模型对输入输出进行评估，涵盖语义、质量、安全风险

AI 网关 | ARMS | SLS | 云监控

AIStudio: 高性能低代码 Agent 开发平台

触发调度

- 定时触发
- 消息队列
- 日志
- HTTP
- ...

交互式开发

画布&视图

- AIStudio
- CloudStudio

工作流引擎

- 标准工作流（Standard）
- 快速工作流（Express）

基础组件

- LLM/Agent
- 知识检索
- 条件分支/并行/迭代
- 更多...

拓展

- MCP Markplace
- Model Registry

自定义

- 代码执行
- 函数计算

可观测/可运维/可靠交付

- 版本管理
- 日志/监控/链路追踪
- 别名流量灰度
- 沙箱隔离/弹性扩缩

阿里云

Contents

目录

与光同尘 AIGC 案例：业务特点及挑战

业务流量在 平峰 与 高峰期 有十几倍的 Gap

成本难以控制

平常资源消耗相对平稳，传统包月方案资源配置不够灵活，资源闲置率较高

+

弹性能力要求极高

AIGC 课程高峰期需要几分钟内弹出几百张卡的 GPU 资源，弹性能力要求很高

+

运维管理投入大

业务发展迅速，人员需要聚焦在业务本身，传统方案运维程序繁琐，大大增加运维的工作量

技术选型

Technology Solution Comparison

一般有 2 种技术选型：模型服务商（如 OpenAI、百炼等）、开源自建（Qwen、DeepSeek、ComfyUI、SD 等）

维度	SaaS（模型服务商）	PaaS（函数计算 Function Compute）	IaaS （VM/容器自建）
安全	✗ **数据风险高：** 数据在第三方，合规不可控	✓ **数据风险低：** 数据在客户私网，厂商基础安全保障	✓ **数据风险低** 数据在客户私网，厂商基础安全保障
效率	✓ 开发效率最高： 开箱即用，零配置 ✗ 几乎不可定制： 无法修改底层框架，可选模型少	✓ 开发效率高： 开箱即用，无需管服务器及其环境依赖 ✓ 定制效率高： 框架/模型自由，开源选择多	✗ 开发效率低： 需手动配置集群、网络、依赖 ✗ 定制效率低： 小规模效率高，大规模效率低
可靠	✓ **可靠性最高：** 服务商提供 SLA，自动容灾 ✗ 完全黑盒： 故障依赖服务商修复	✓ **可靠性高：** 3AZ 高可用，自动容灾 ✓ 黑盒+白盒： 自带监控日志链路追踪等工具	✗ **可靠性低：** 自主实现高可用架构 ✓ 完全白盒： 需要自建监控日志，代价高
弹性	✓ **有限弹性：** 按请求弹性，配额受限则不可再弹 ✗ **成本不可控：** 按请求单价高，突发流量费用激增	✓ **极致弹性：** 按请求弹性，毫秒/秒级供给资源 ✓ **成本可控：** 按资源单价低，利用率高浪费少	✗ **普通弹性：** 手动/自动扩展 VM 或 Pod，分钟级 ✗ **成本可控：** 按资源单价低，利用率低浪费多

基于上页提到的挑战，我们可以发现IaaS基于成本、效率、稳定性、弹性上均不满足我们要求。

与光作为创企，在 PoC 阶段快速验证效果，SaaS / PaaS 的开箱即用无疑是最简单的选择。
随着规模扩大，从百卡 GPU 需求增长到千卡级别，核心业务的自建和创新PaaS无疑是 ROI 最高的选择。
等到业务成熟，有专业团队，PaaS / IaaS 自建是追求自主可控的必然选择。

函数计算作为 PaaS 产品，从安全、成本、告警监控、问题排查、异步调度等多个方面能平滑满足各阶段的需求

技术架构

Technical Architec0ture

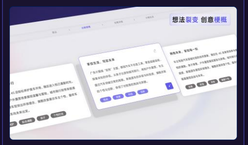
应用层

商业影片生成

短剧创作

AIGC教育培训

想法裂变

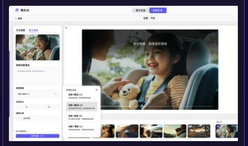


脚本扩写

智能分镜生成

图片编辑

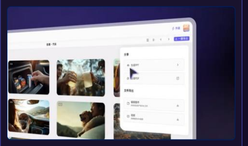
快速生成视频



轻松拖拽时间线

协作管理

一键导出提案



平台层

资产管理

工作流广场 | 模型广场

数据资产 | 函数管理

模型训练与评测

模型盲评 | 模型训练

数据采集与处理 | 模型设计

平台运维

实例登陆 | 账号管理 | 弹性 API

监控报警 | 应用运维 | 异步任务

统一资源调度

基础设施层

函数计算 FC

数据库 RDS

对象存储 OSS

容器服务 SAE

网络

消息队列 RocketMQ

模型 API

万相系列

Qwen 系列

第三方模型

效率革命与成本重构

Efficiency Transformation and Cost Restructuring

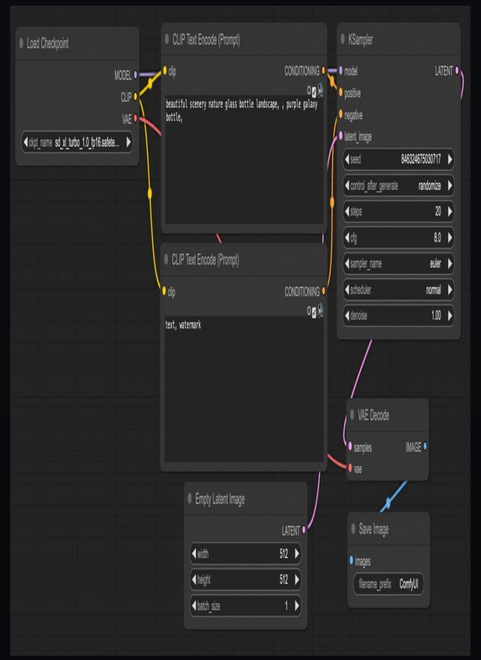
一键部署 + 模型管理 UI 界面	工作空间 + 生产 API 独立计费	ComfyUI + Flux 提速 2 倍	多种生产工作流可用

模型

- SD 模型
- FLUX 模型
- WANX 模型
- Qwen-Image
- 其他模型... ...

+

ComfyUI



图片编辑　分镜生成　文/图生视频

视觉生成 API → 集成 → 与光AI

50% 运维效率提升

60% 发布效率提升

80%+ 推理速度提升

30% 模型加载速度提升

服务发布上线速度 **分钟级**

系统 SLA 达到 **4 个 9**

AI 让生活更美好

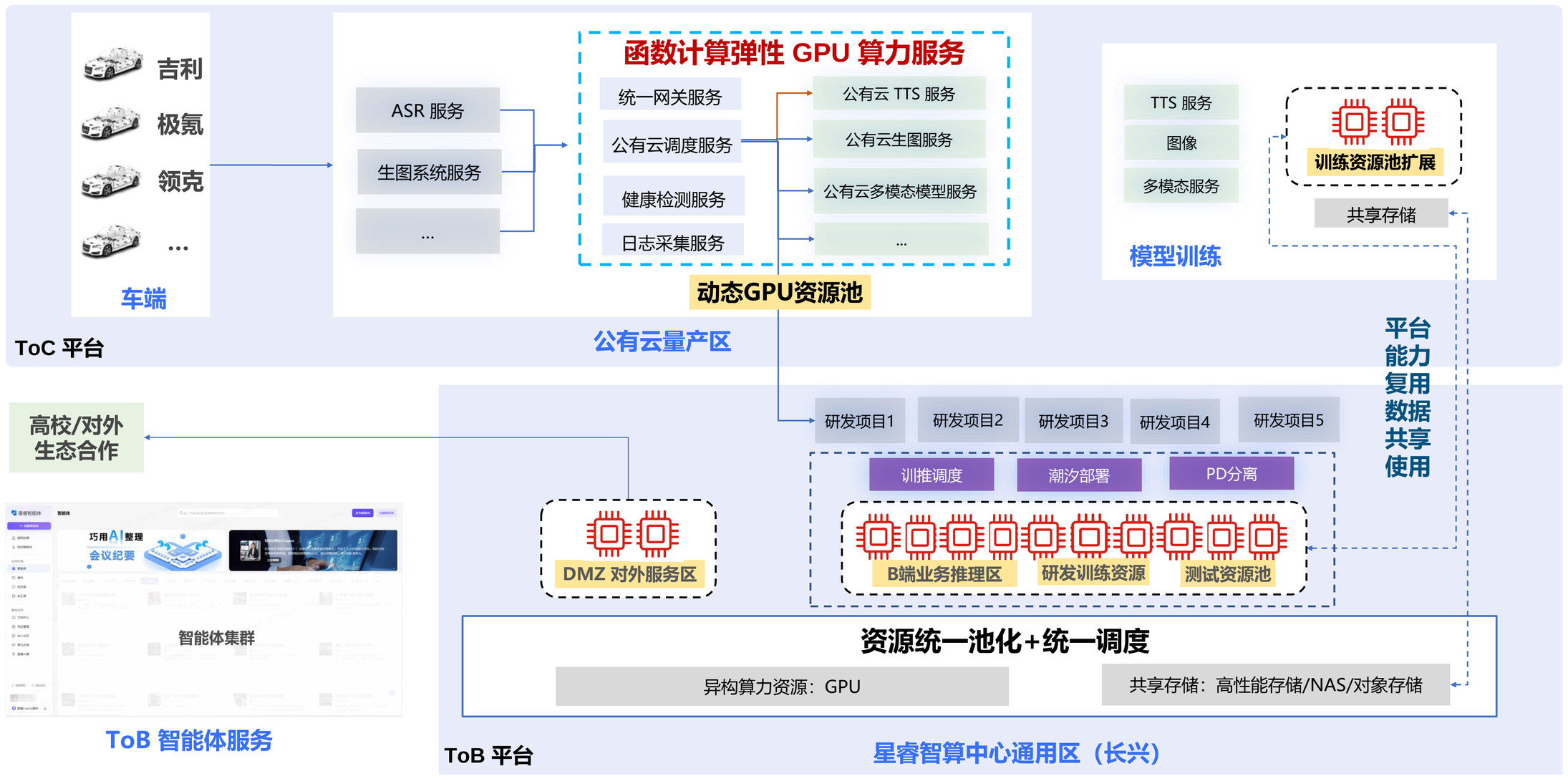
AI Makes Life Better

AI 上更阔气的家/AI 上更百变的家/AI 上更智慧的家/AI 上更强大的家/AI 上更轻松的家/AI 上更安心的家

- 随着 AI 浪潮的发展，吉利汽车在今年推出新一代 AI智能化服务，成为核心竞争力之一，在吉利银河 M9 车型深度搭载。
- 基于阿里云函数计算的 Serverless GPU 算力集群为 AI 座舱的交互和娱乐功能提供大模型推理服务，共同打造大规模、高可用、高性能的推理引擎。
- 场景涵盖意图解析、文生图、情感 TTS 等。尤其在语音交互方面，吉利自研超拟人 TTS，采用大量微调和训练，打造吉利汽车专属的逼真、富有情感、自然生动又稳定一致的语音交互体验。

吉利汽车 AI 业务部署架构

Geely Automobile AI Application Deployment Architecture



AIStudio 场景案例： 社区内容审查

挑战

客户 Z 作为中国最大的问答社区之一，每天面临海量内容审核需求，传统人工审核方式效率低下，难以应对突发流量。

解决方案

基于AIStudio工作流引擎构建LLM内容审查系统，采用多级审核流程：

- 第一级：基于规则的快速过滤
- 第二级：LLM语义理解和上下文分析
- 第三级：人工确认和反馈学习

技术亮点

- 利用AIStudio高吞吐特性，支持1000QPS的内容审核需求
- 沙箱环境确保敏感内容处理安全性
- 弹性伸缩应对突发流量，降低资源成本

95% 审核准确率	92% 召回率
70% 人工介入减少	3.4倍 处理速度提升

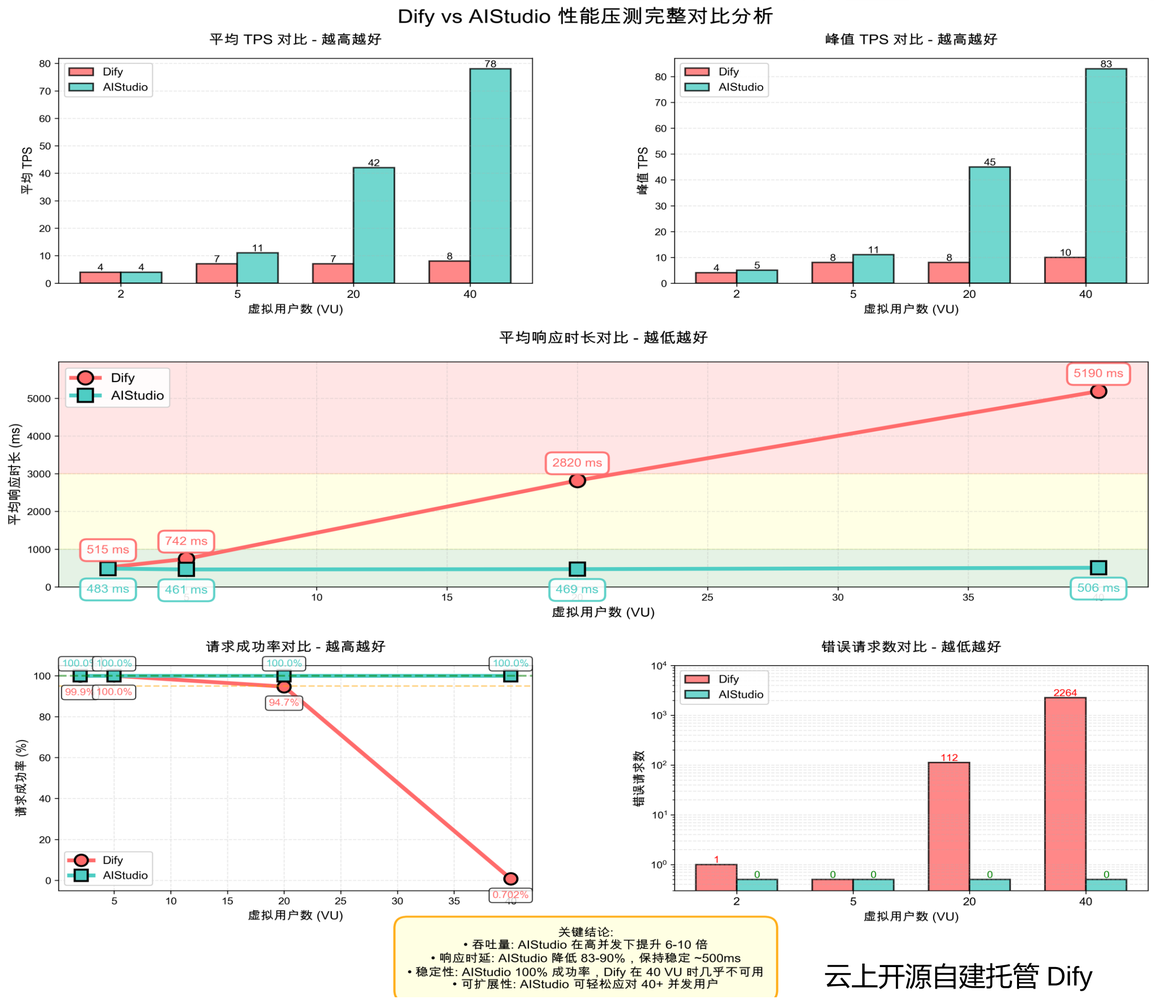
AIStudio 场景案例： 智能家电交互

挑战

客户 S 作为中国家电市场著名品牌，需要为客户提供智能化家电控制能力，通过手机，语音对话进行家电交互。对于Agent的响应效率和并发处理性能提出了严苛的要求。

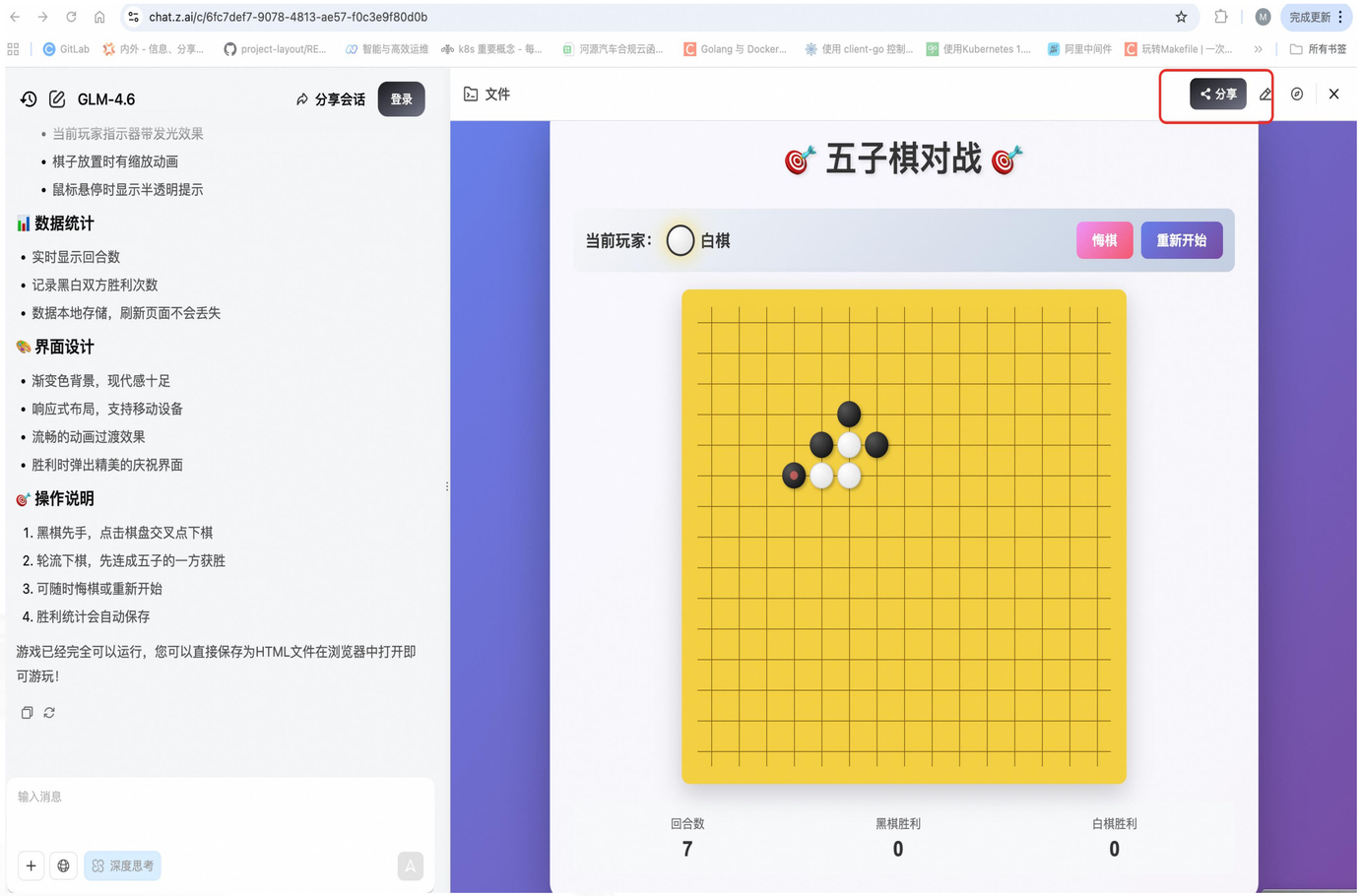
技术亮点

- 利用 AIStudio 高吞吐，高并发满足并发要求，高性能满足响应要求
- 代码处理节点，沙箱环境确保敏感内容处理安全性
- 弹性伸缩应对突发流量，降低资源成本，应对家电用户动态流量需求



云上开源自建托管 Dify

智谱 Sandbox 场景案例： 为 Chat Coding 提供安全的验证环境

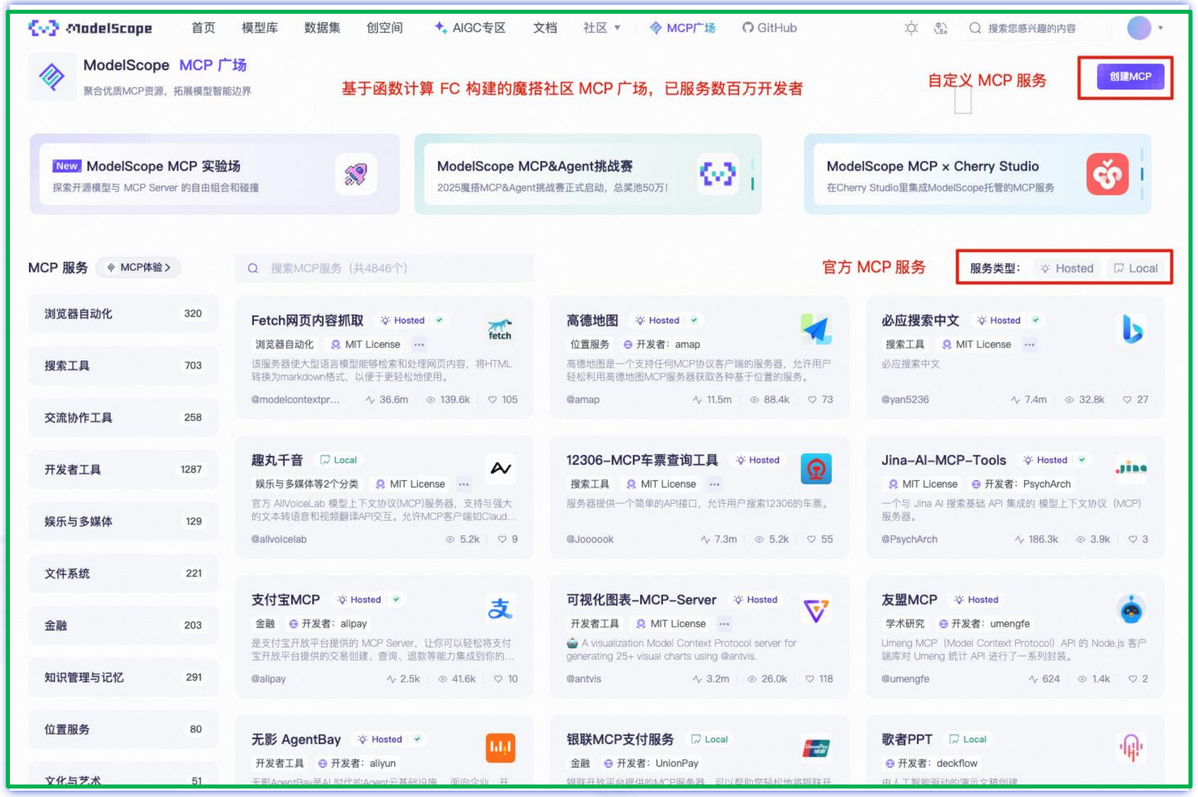


- **Agent Sandbox Runtime：**如 Code Sandbox、Browser Use Sandbox、仿真训练 Sandbox、RL Sandbox
- **成本最优：**按请求扩缩，提升 AI Agent 资源利用率，降低资使用源成本
- **多租强隔离：**安全容器+动态存储挂载，为每个启动的 Agent 提供干净独占的运行环境
- **启动速度快：**CPU 百毫秒弹性伸缩
- **内置完善的监控日志能力：**OpenTelemetry Trace 支持
- **会话亲和调度：**保持业务运行时及业务上下文
- **业务隔离：**运行时支持按请求隔离，会话隔离

MCP 行业案例： 百炼、魔搭社区 MCP 服务市场

联合魔搭社区发布最大 MCP 中文社区，联合百炼发布业界首个全生命周期 MCP 服务市场，将 MCP 沉淀为平台资产

STDIO 零改造升级为 MCP



魔搭社区 MCP 市场

快速开发和部署 MCP 服务

百炼全周期 MCP 服务

FunctionAI 客户交流群

https://functionai.console.aliyun.com/welcome

阿里云

THANKS

阿里云

Serverless 应用引擎SAE：从传统应用到AI应用的一站式托管

赋能企业安全高效构建部署 AI 智能体应用

卢令

Serverless 应用引擎（Serverless AI|App Engine）技术负责人

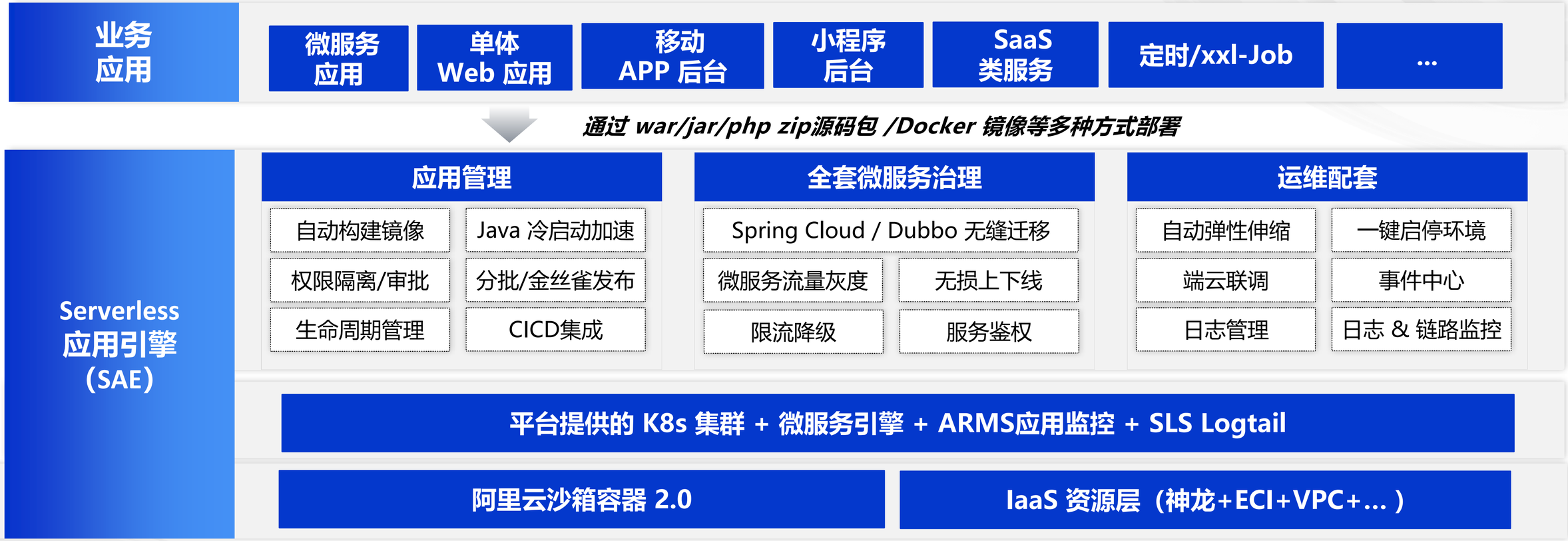
阿里云

Contents
目录

传统应用运维的‘简、稳、省’优化之道

01

简：一站式容器托管平台，让运维更简单



兼容容器生态

- 无缝容器镜像部署模式，解决无厂商绑定
- 支持 K8s API 接口，无缝支持 ArgoCD，Rancher 等CICD 平台
- 支持 Sidecar 模式，可以自定义安装 Agent 能力
- 支持 Kubectl Yaml 部署，保留容器黑屏操作能力

屏蔽 K8s 系统复杂性

- 产品化支持多种灰度策略：蓝绿发布，滚动升级，全链路灰度，标签路由等。
- 以应用维度暴露给业务使用方，以优化的界面暴露给业务研发。
- 免费提供全链路的无侵入式监控，和微服务治理能力
- 企业级特性的加强：符合企业习惯的权限管理，优化的批量操作，巡检等。

稳：从架构到防护，全程保障线上稳定运行

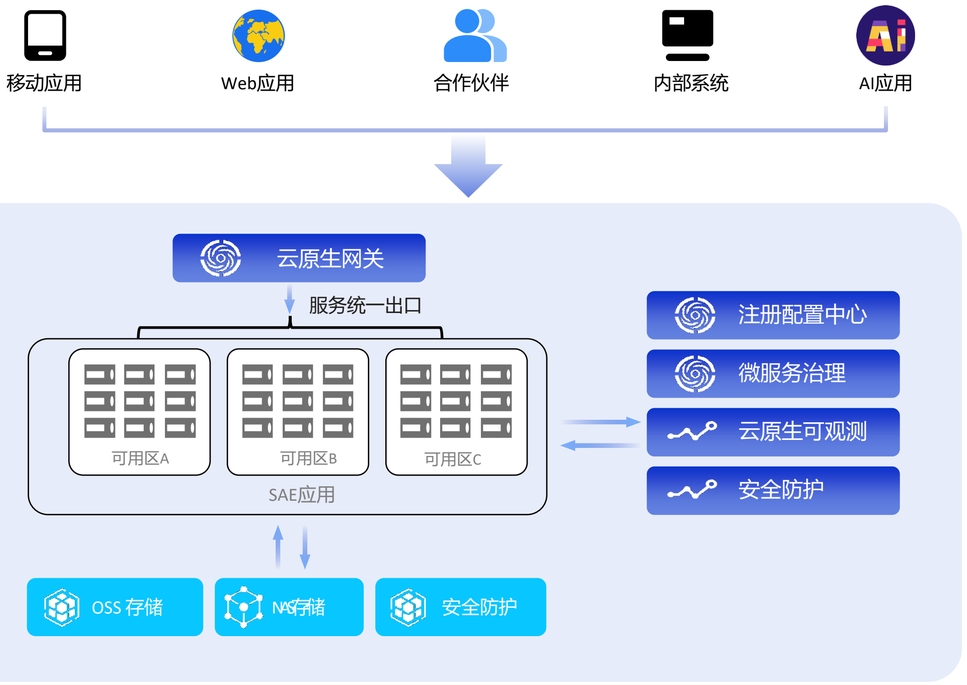
多可用区容灾

多可用区容灾对于SAE是默认的能力：

一键开启

SAE默认应用实例分散部署在多个可用区，实现跨机房容灾。单个可用区故障时，流量自动切换至其他可用区，保障业务连续性。

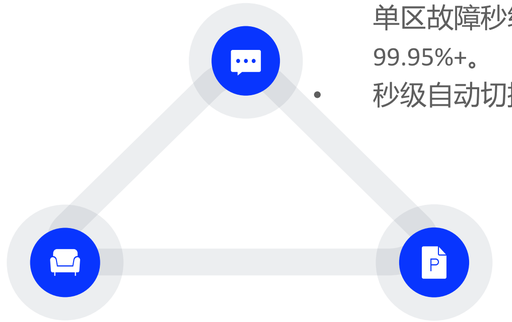
SAE内部架构示意图



多可用区优势

跨可用区容灾（Multi-AZ）

- 应用实例自动分发至多可用区（AZ），单区故障秒级流量切换，可用性达99.95%+。
- 秒级自动切换，RTO≈0，RPO≈0



智能流量调度

- 同可用区优先路由：优先访问同 AZ 实例，跨区延迟降低 80%（1ms→0.2ms）。
- 全局负载均衡（SLB）：故障时自动切换至健康 AZ，业务零中断。

全托管运维

- 无需维护资源，自动维护多 AZ 资源池，无需手动配置，运维成本降低 70%。
- 按需跨 AZ 弹性伸缩，资源利用率提升 50%

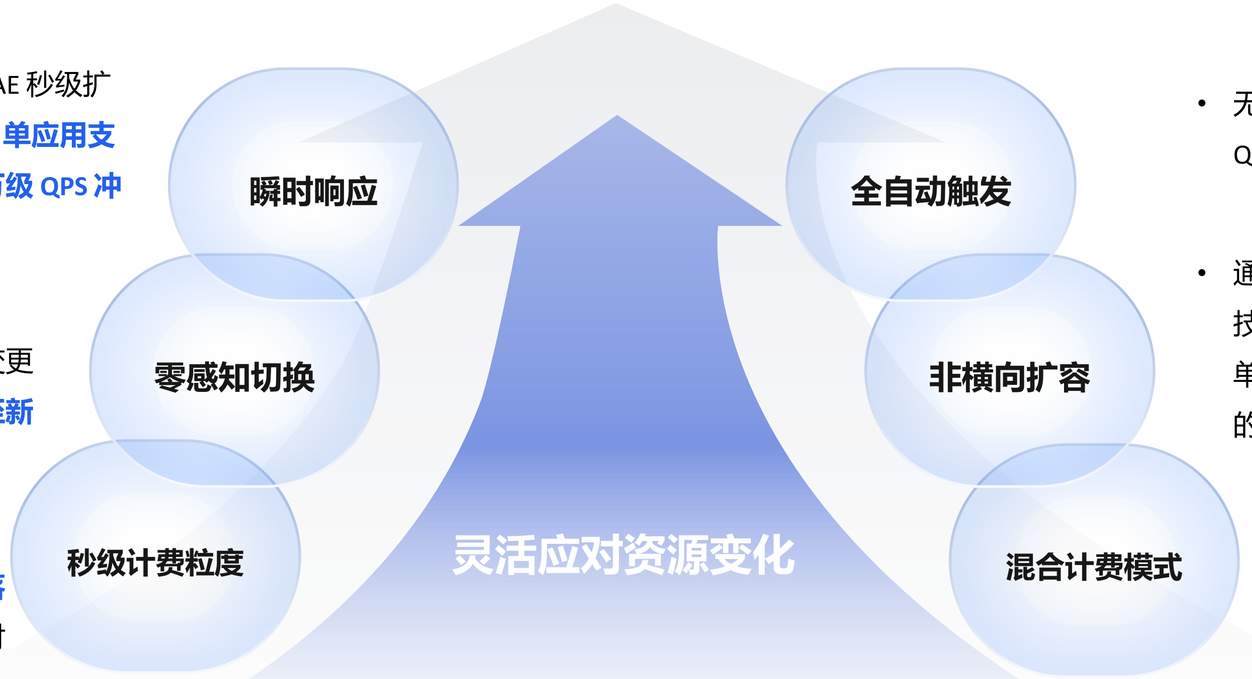
省：极致资源利用率：按需付费 + 零冗余 + 超卖让利

核心竞争力

- 当流量突发超过预设阈值时，SAE 秒级扩展资源至 集群规模 100% 上限，**单应用支持万级实例瞬间拉起，承载百万级 QPS 冲击**，保障业务零降级、零中断。

- 扩容过程不断开现有连接、不变更服务 IP，**业务流量无感知切换至新实例，服务连续性 100% 保障**。

- **突发时按秒计费扩容，流量回落后自动缩容至基线实例数**，闲时成本直降 50%。



- 无需人工干预，系统实时监控流量（如 QPS、CPU）并触发扩容。

- 通过资源超分配（Over-Subscription）技术，在不增加实例数的条件下提升单实例处理能力，避免传统横向扩容的复杂度。

- 基线 + 弹性组合都Serverless 付费。

极速扩容（秒级拉起实例）、高并发承载能力

加速 AI 创新：从快速探索到高效落地

02

规模化落地 AI 应用的痛点

开源平台性能差

- 各组件（如：Worker、Plugin、数据库等）参数非最优配置
- 管控面与数据链路耦合，高并发无法保证稳定性
- 数据源存储格式单一，推理服务需要大量的计算资源，资源分配不均会导致性能瓶颈

运维复杂度高

- 本地部署复杂且维护成本高，需要频繁升级版本
- 需要自己管理应用的版本发布
- 周边配套不完善：没有配套的治理、可观测体系，事前事后无法及时发现并定位问题

成本不可控

- 资源错配，要么业务低峰期闲置烧钱，要么业务高峰期瞬间被打满，影响业务
- 人力维护投入大

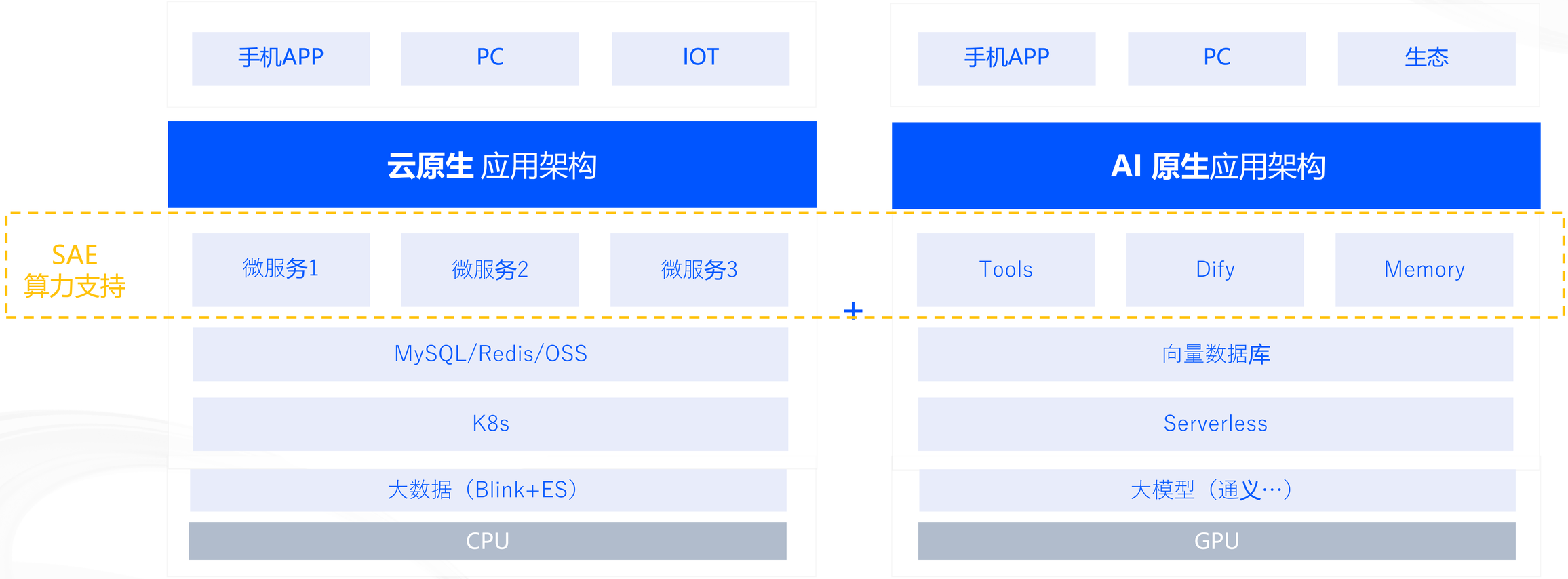
安全合规风险

- 流量防护弱，很容易被穿透
- 数据隐私与合规性管理困难

企业真正需要的是：开箱即用的开发体验 + 生产级的性能、稳定性及安全保障

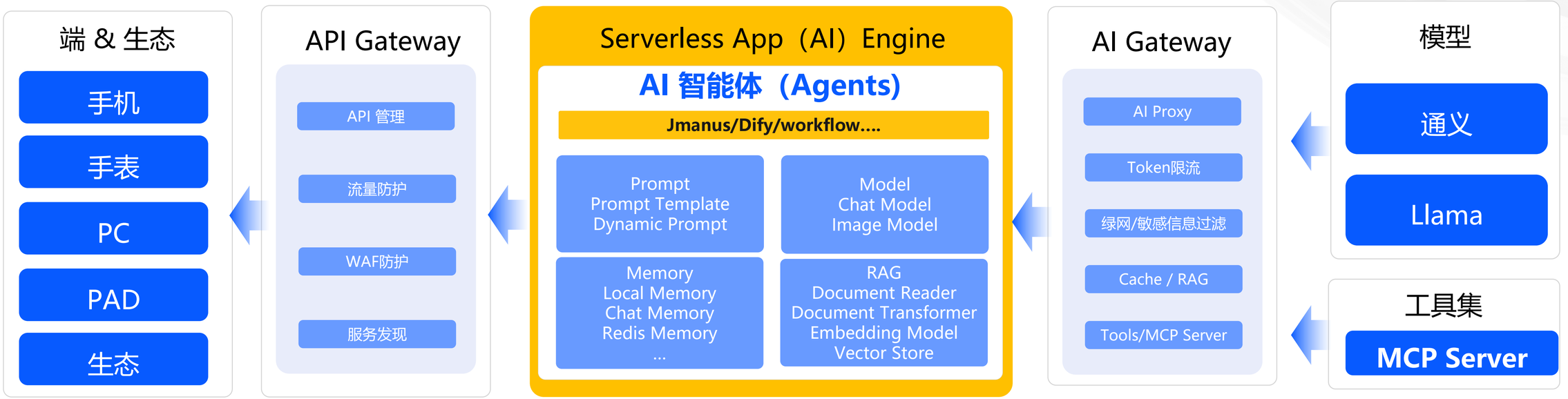
SAE 在 AI 原生应用领域的定位

不做开发平台的替代者，而是做它们的“护航舰”



SAE 致力于托管主流开源 AI 智能体应用开发平台（深度适配+全局赋能）

SAE 全托管 AI 智能体解决方案



Serverless （AI）应用引擎托管 AI Agents 方案优势

简单易用

- 一分钟创建 AI 应用，无需任何额外配置
- 默认集成全链路监控，保证系统稳定性
- 无需关系底层资源，按需弹缩资源

稳定高可用

- 配置化，支持三 AZ 部署，默认支持智能化可用区，实例粒度的自动化迁移
- 默认支持负载均衡与健康检查联动保证无损上下线

低成本

- 按需按量付费，潮汐流量弹性使用，无需冗余保证资源
- 支持多种规格资源，并提供闲时计量资源类型，提供更低成本的算力

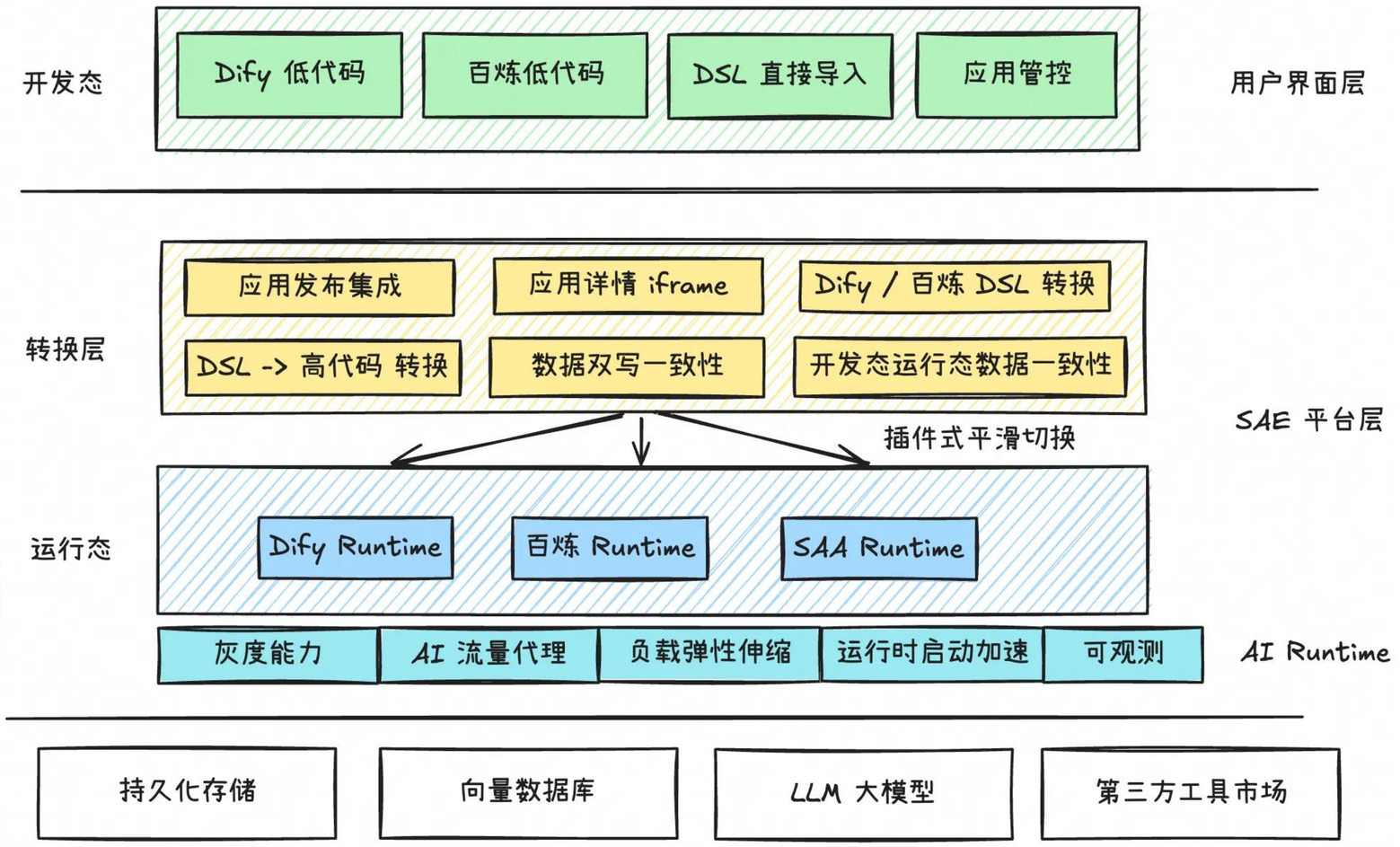
安全保障

- 全链路提供防护策略：Ddos 防护，Web防护墙，流量防护，云安全中心。
- VPC 内独立部署，数据不出安全域，保证数据绝对安全

持续迭代

- SAE 默认具备灰度发布，分批发布，镜像加速，Pod 粒度监控，保证 Dify 进行安全二次开发
- Dify 版本更新快，通过 SAE 可安全兼容升级。

降本增效 - 低成本托管 AI 应用套件



低成本

- 推出 Besteffort，轻量版算力，让业务享受到云算力的让利。
- 无需扩缩，就可以用闲置能力支持业务的低峰，成本降低 80%。
- 专业版免费提供全链路的无侵入式监控和微服务治理能力。

灵活+性能

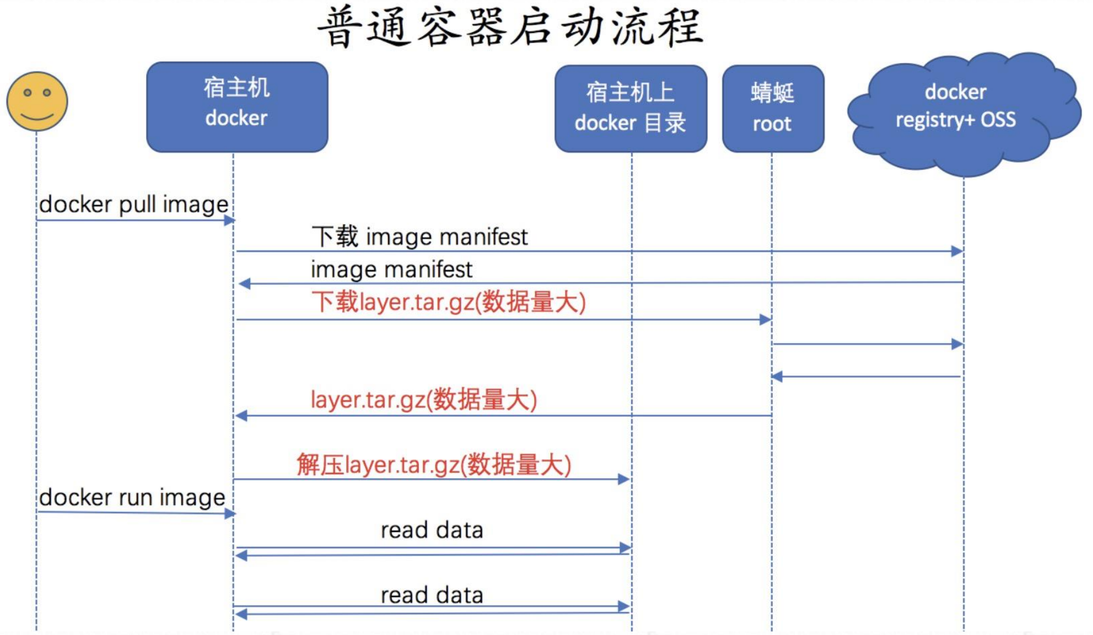
- 一键部署及版本升级：Dify，Jmanus，OpenManus，Airflow 等成熟框架
- 支持 K8s yaml，kubectl 快速部署任何的开源 AI 框架
- 性能优化：参数调优、控制面与数据面分离
- 无厂商绑定

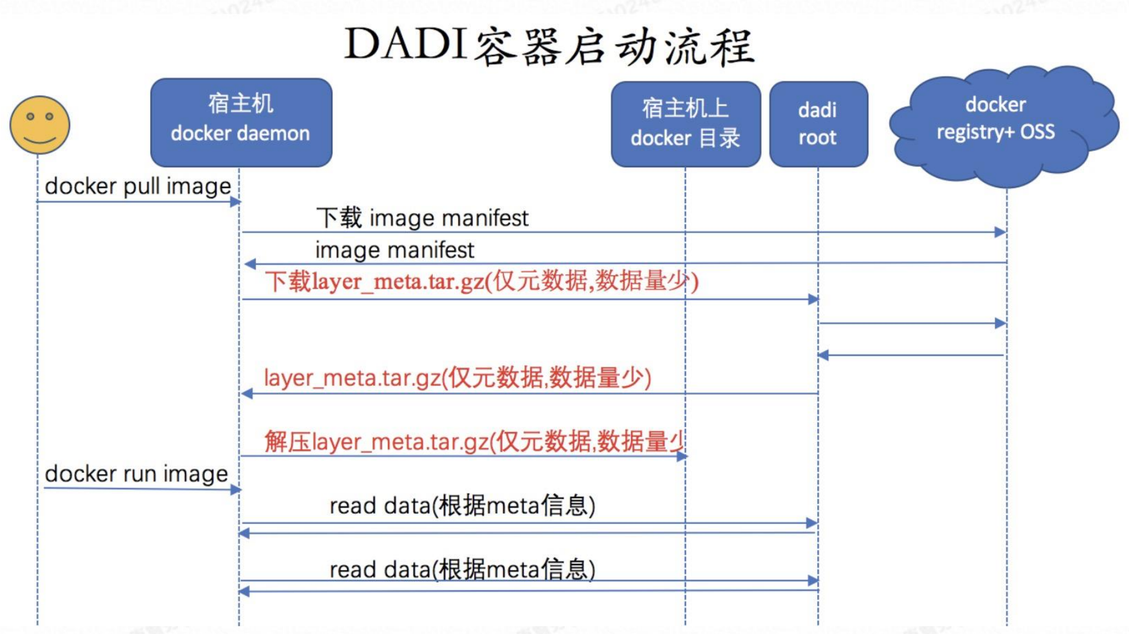
Agent 启动加速 - 默认具备镜像加速

普通容器启动慢根因：1.容器启动之前，OCIV1标准的镜像下载和解压耗时大　2.应用启动仅依赖6.4%的镜像数据

DADI镜像加速方案：**转化镜像格式、OnDemand read（按需加载启动容器）**

加速原理





运维配套 – 自定义弹性伸缩

作为 SAE 的核心竞争力，相对传统 ECS 的弹性，SAE 更精准更降本；相对 K8s 弹性，SAE 的指标和策略更丰富，上手门槛更低。

1. 指标弹性（CPU、 Mem、 QPS 、 RT等）

适用于有突发流量、典型脉冲的应用场景，多用于互娱 / 游戏 / 社交平台 / 电商等行业。

优势：比开源K8s HPA 指标丰富，且可以自定义指标。

2.定时弹性

适用于资源画像存在周期性的应用场景，多用于餐饮 / 出行 / 证券 / 医疗政府等行业 。

优势：操作简单，易用。

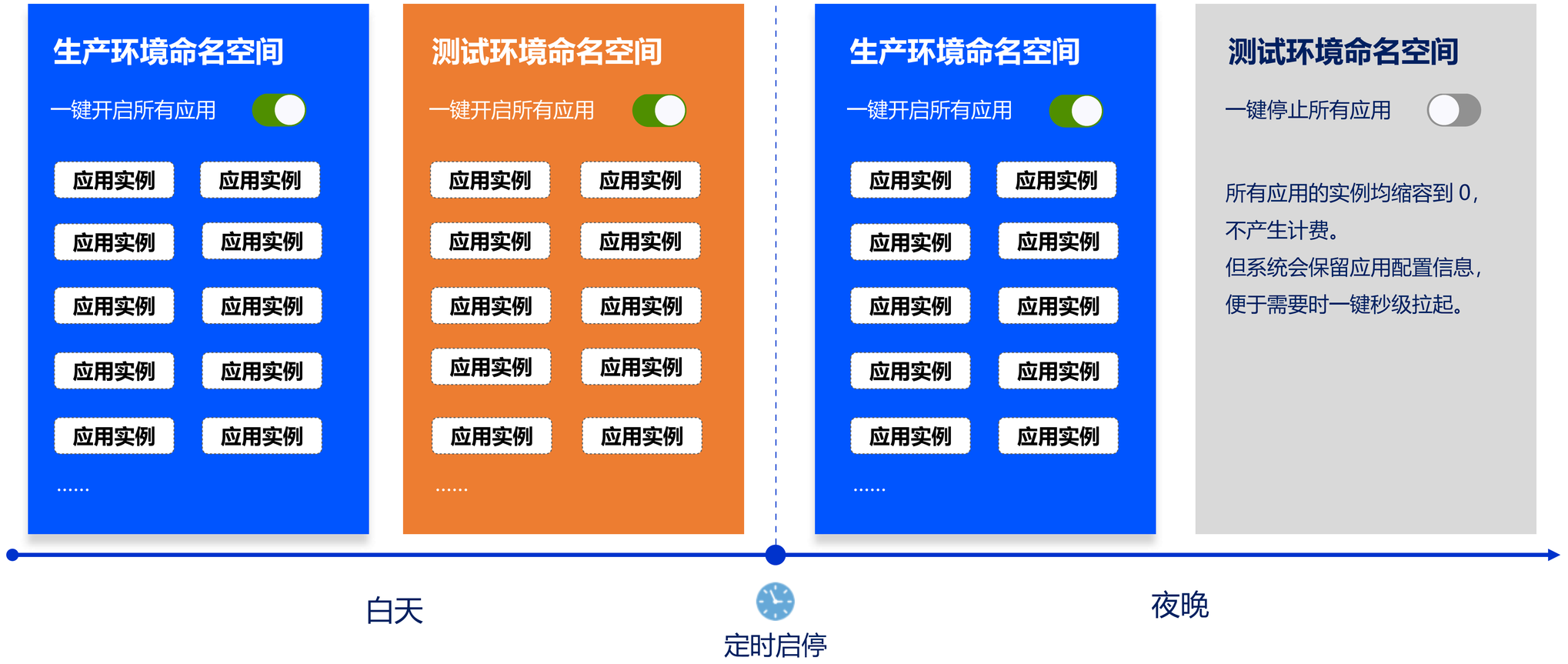
3. 混合弹性（定时弹性 & 指标弹性混用）

适用于固定时段内有突发流量、典型脉冲，常稳时段内流量波动不均的应用场景，多用于 媒体报社 / 在线教育 / 语音识别合成等行业。

优势：比开源K8s HPA 指标丰富，且可以自定义指标。

一键起停开发测试环境

中大型企业多套环境，内部环境长期资源浪费，使用 SAE 一键启停，可以节省一部分资源成本。



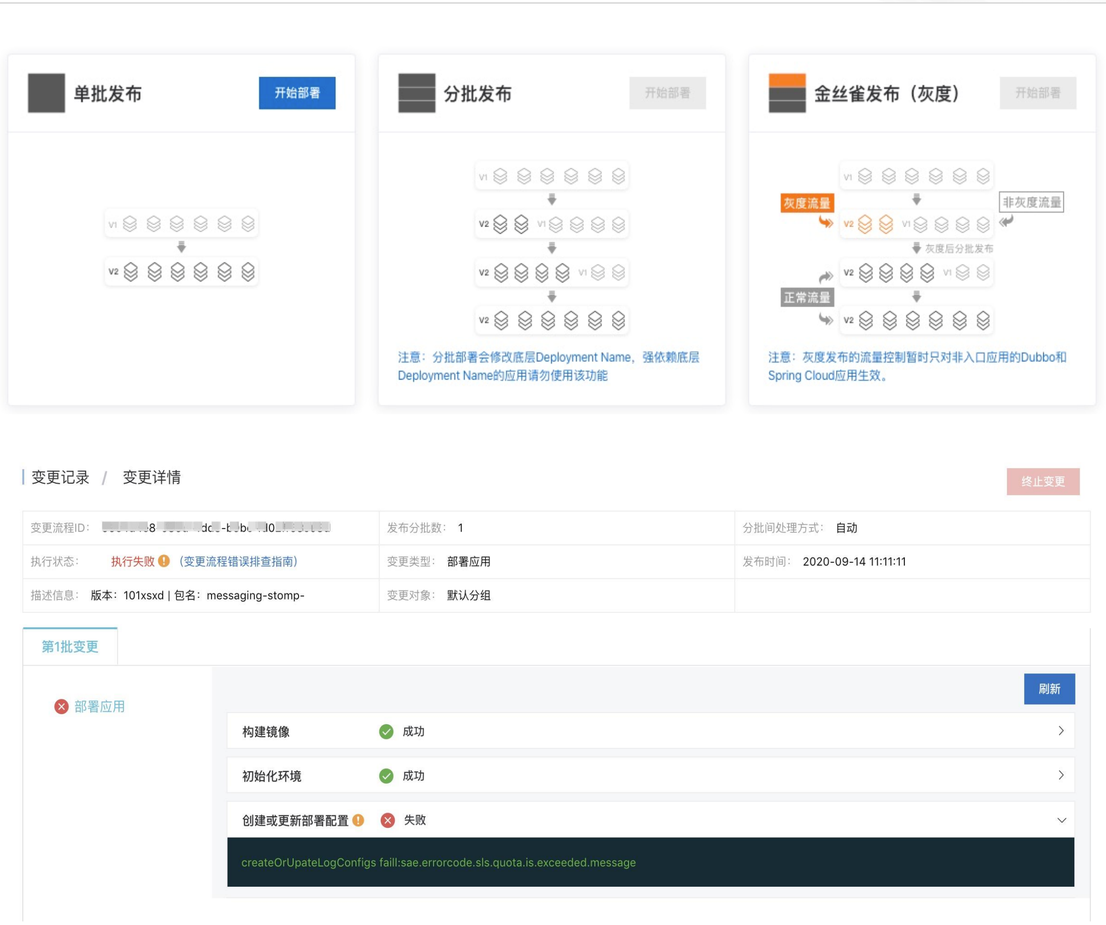
AI 应用全栈统一监控


针对大模型应用提供四大核心可观测能力
端到端诊断
对常见的智能体和智能应用框架进行无侵入埋点和链路跟踪
全栈智算监控
覆盖模型应用、模型服务、网关、容器、智算节点等各层的关键指标和日志
成本管理优化
提供 Token 消耗和 GPU 使用分析，结合网关提供灵活的路由、限额限流
质量安全评估
使用评测模型对输入输出进行评估，涵盖语义、质量、安全风险等
大模型服务 (MaaS)
应用
RAG、MCP 应用运行监控
AI App 监控
模型调用监控
安全诊断
模型
模型监控
模型日志
模型链路
AI 平台 (PaaS)
模型推理
服务与组件监控
推理引擎监控
服务组件
向量数据库监控
AI 网关监控
模型训练
训练任务监控
数据流监控
容器 Kubernetes
容器调度
Workload 可观测、控制面可观测、GPU 监控、Ingress 监控，事件、日志审计，AI 套件可观测
智算服务器 (IaaS)
AI 基础设施
灵骏计算节点监控
网络 RDMA 监控
高速存储 CPFS 监控
GPU 云服务器监控

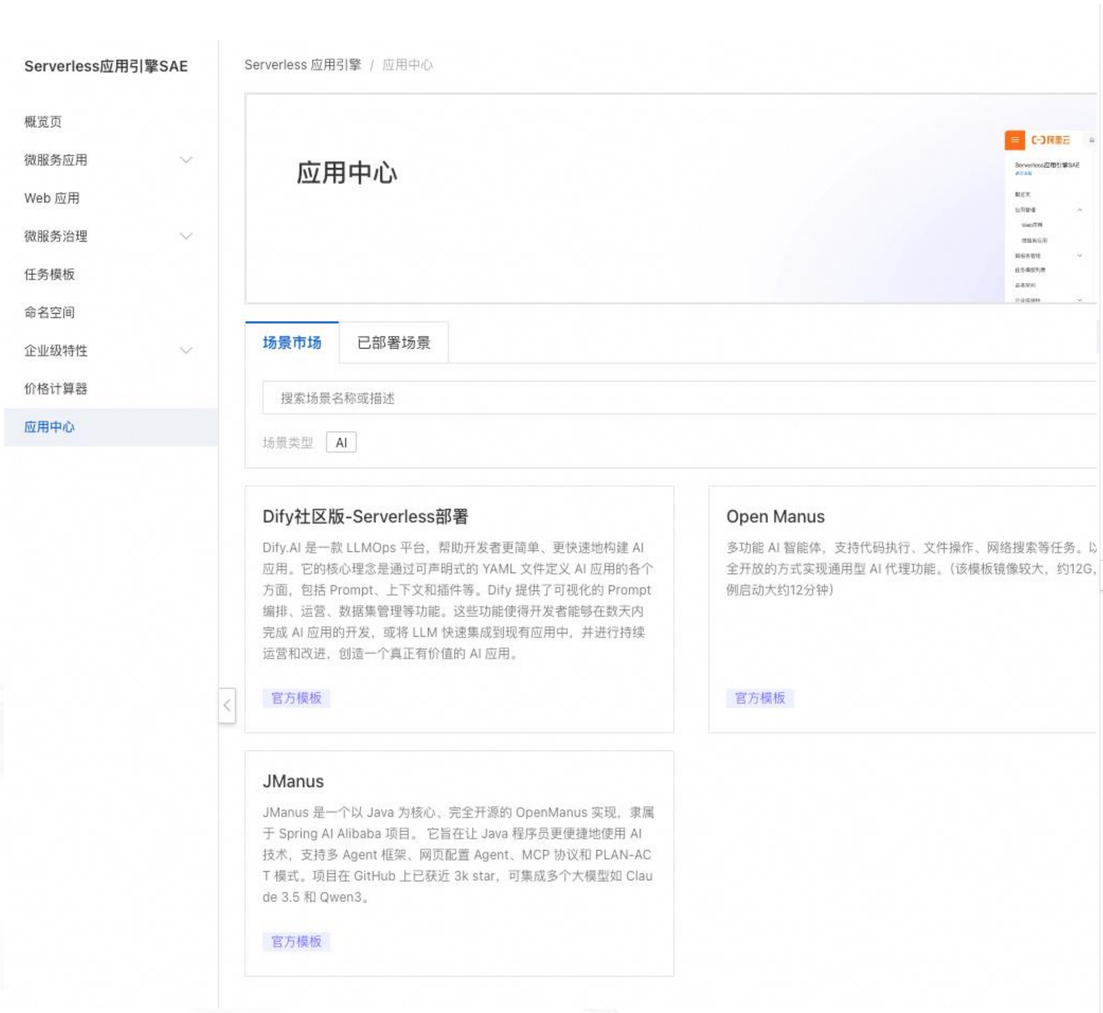
运维配套 – 可灰度、可观测、可回滚

将阿里巴巴沉淀多年的安全生产实践产品化，应用生命周期管理纯白屏化操作，应用发布可灰度、可观测、可回滚。

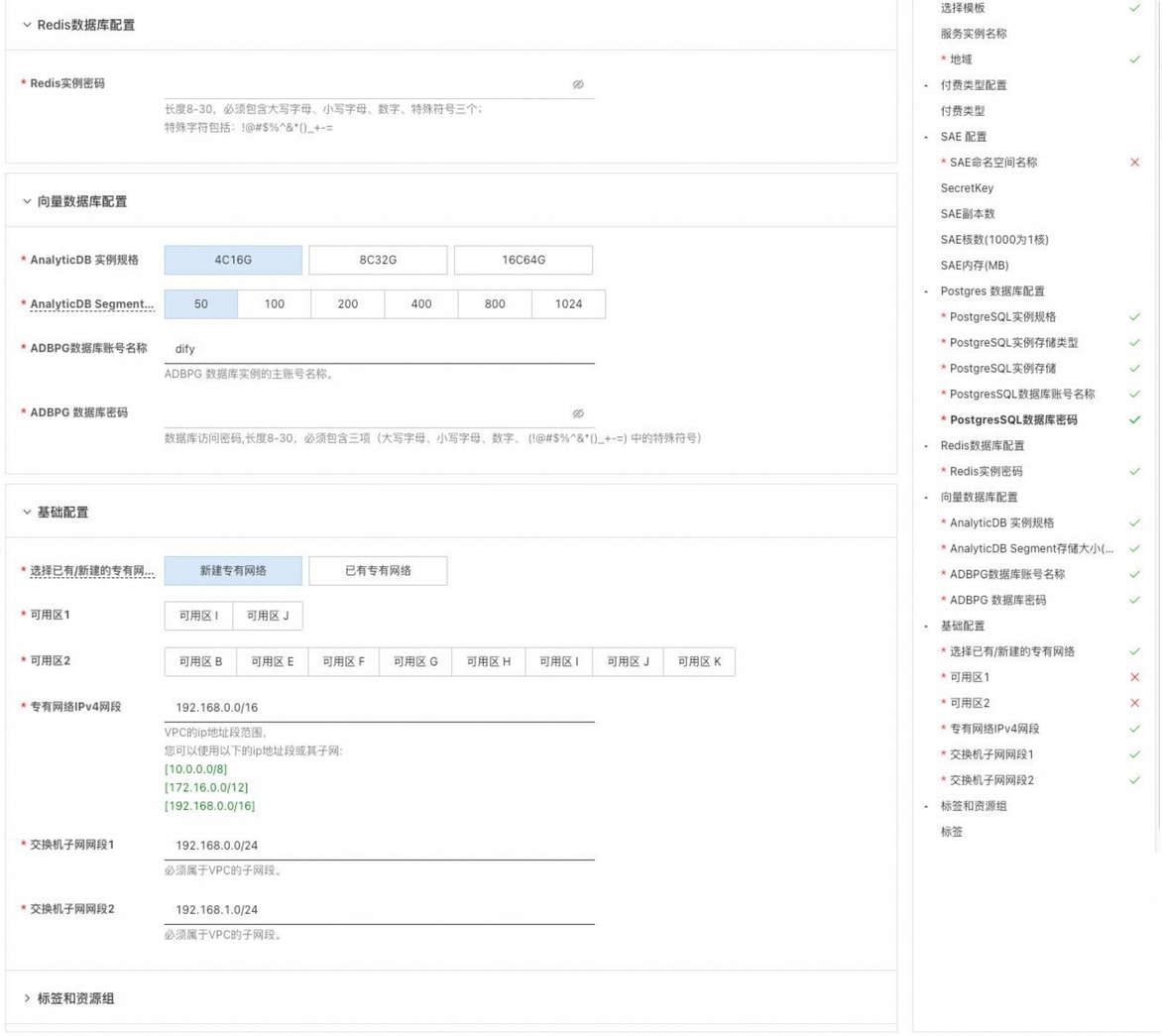
- 支持单批、分批、金丝雀等发布策略。支持按流量灰度、按内容灰度。批次间自动/手动任选
- 发布过程可监控，白屏化实时查看发布日志和结果，及时定位，
- 允许人工介入控制发布流程，手动确认通过、异常中止、一键回滚



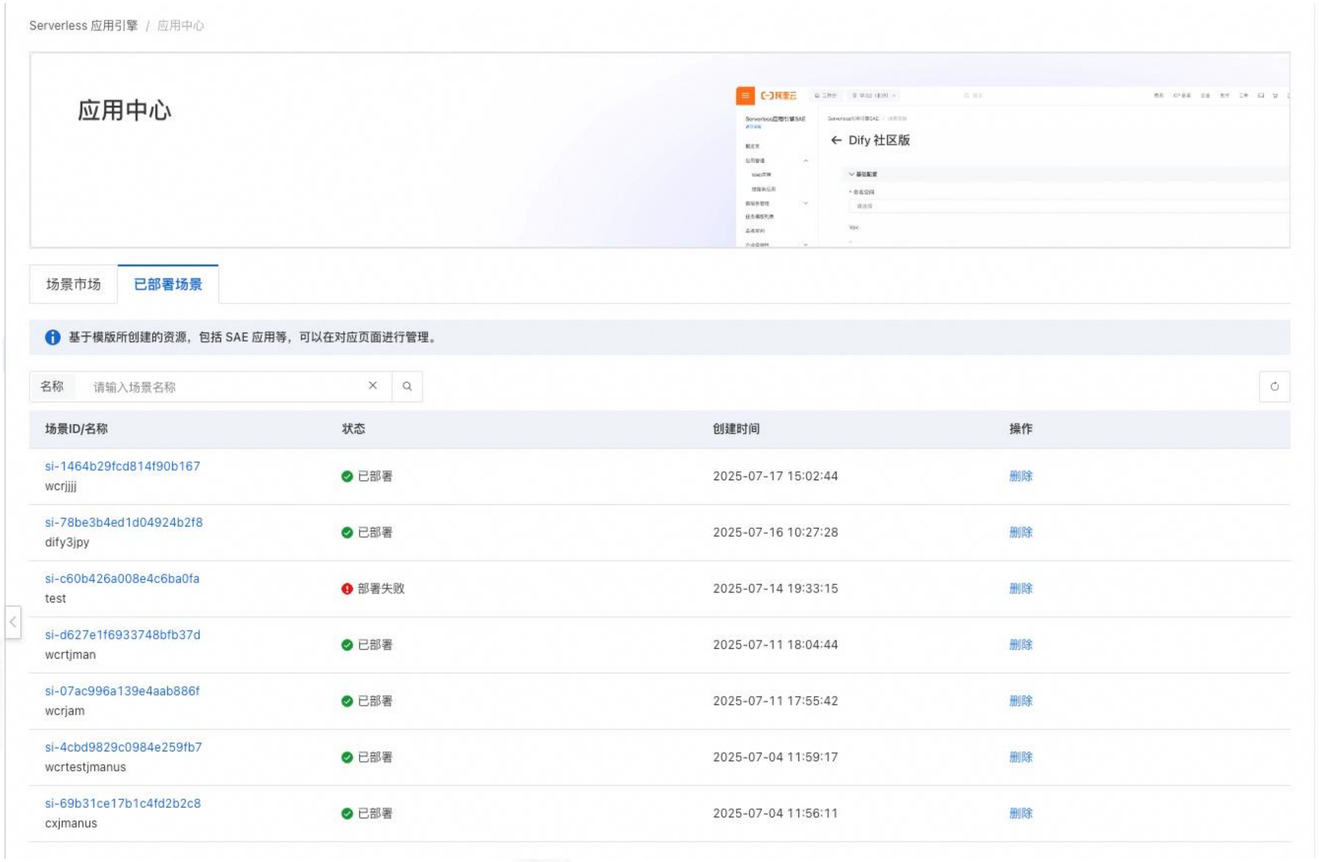
SAE AI 场景模板市场

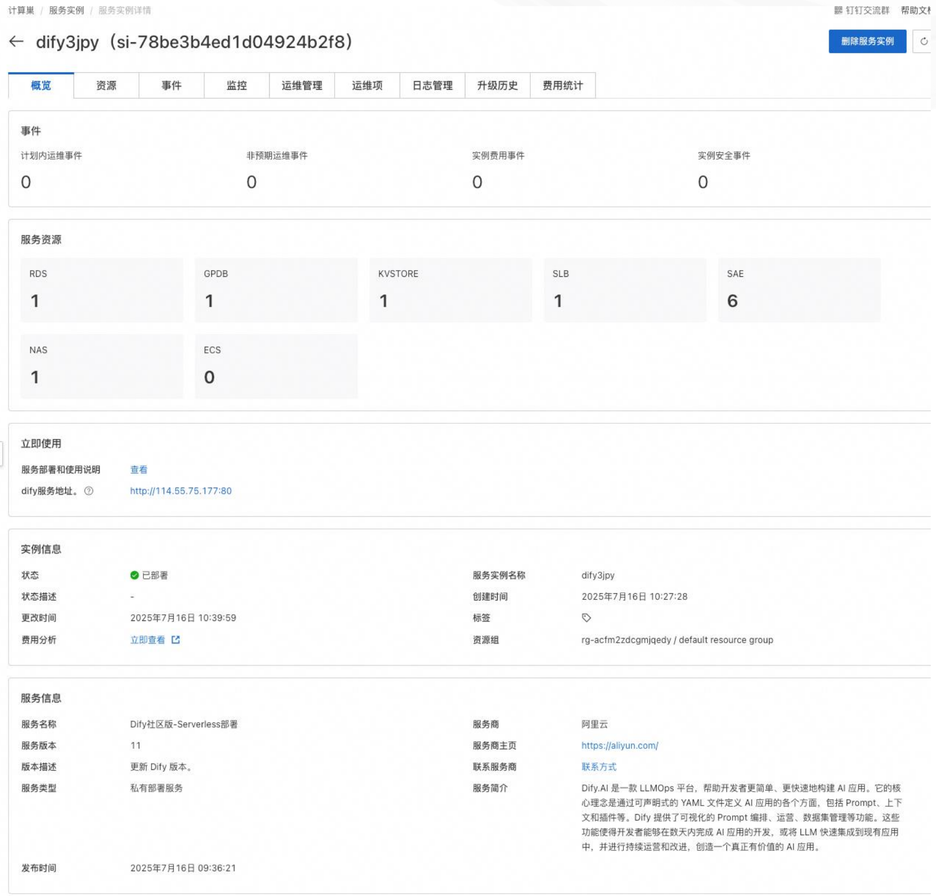


SAE AI 场景模板市场 - Dify 创建



SAE AI 场景模板市场 - Dify 创建





阿里云

THANKS

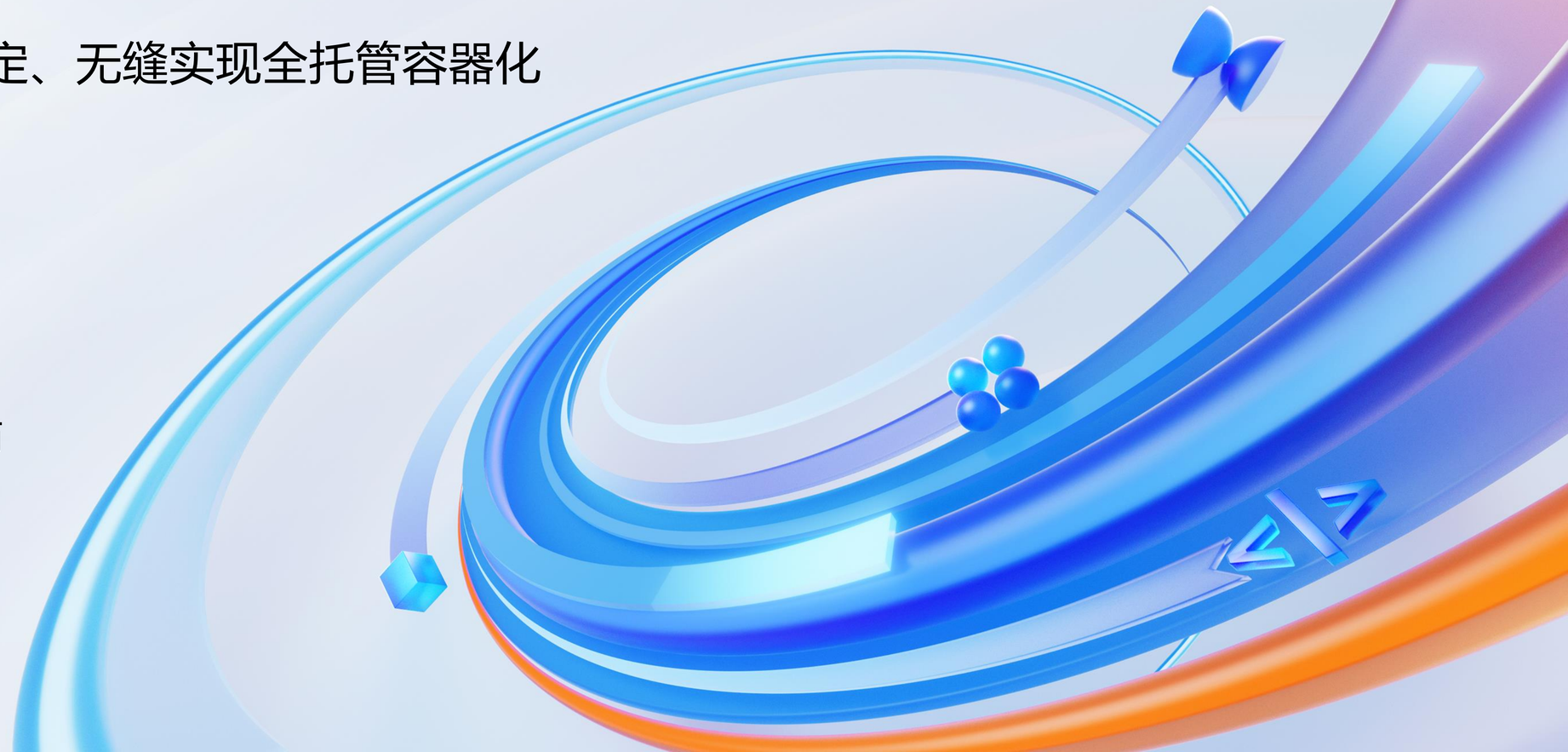
阿里云

传统架构 Serverless 容器化迁移实践

SAE赋能企业高效、稳定、无缝实现全托管容器化

杨跃山

阿里云云原生高级架构师



企业云原生上云的挑战

技术

- 部署在IDC或运营商云IaaS
- 技术架构滞后
- 缺少体系化的PaaS服务构建
- 稳定性与高可用性保障
- 安全性面向挑战大
- 架构转型难度
- toC及新业务建设和舆情压力大

管理

- 运维管理复杂度高
- 多级组织
- 研发依赖外包、缺少统一的技术规范、技术栈多样
- 运维管理复杂度高、风险大
- 建设周期长

成本

- 成本压力大
- 历史包袱大，需要考虑遗留系统，系统考虑迁移路径
- 系统集成复杂度高
- 如何快速、低成本融合先进的云原生技术：容器化 + 微服务 +可观测 + Serverless

阿里云

Serverless容器化架构的演讲趋势

01

阿里云

All On Serverless 让应用开发更简单

千行百业 百万开发者

互娱/教育	游戏	新零售/电商	传媒/医疗	汽车出行	制造/文旅
新浪微博 weibo.com 新东方 XDF.CN 爱奇艺·体育	心动 沃娱互动	世纪联华 CenturyMart Semir	Focus Media 分众传媒 PONY 谱尼测试 Pony Testing International Group	ZEEKR 高德地图	JL·TOUR 捷旅假期 SKG

Serverless 工具链

Serverless 应用中心

Serverless Devs + AI Studio

All On Serverless 解决方案

REST API
Restful 风格 极简编程

EDA 架构
云产品深度集成，构建事件驱动

AI 推理
GPU 虚拟化+ AIGC

媒体处理
音视频转码/直播录制/图片加工

ETL 数据加工
Kafka、MQ 等上下游联动计算

游戏/汽车等
战斗结算/游戏打包/移动车联等

微服务
微服务架构和容器平台的深度融合，集成微服务治理、应用监控等

WEB 应用
面向单体应用的全托管平台，集成流量网关，支持实例缩 0、闲置计费等

Job/批任务
极简批量任务运行平台，支持定时、XXL-Job 等开源调度框架

BAAS 服务联动

存储服务
OSS、NAS、OTS、SLS

数据库服务
RDS、PolarDB、MongoDB

消息/微服务中间件
EventBridge、RocketMQ、Kafka、MNS、MSE

网络服务
API GW、CLB、ALB

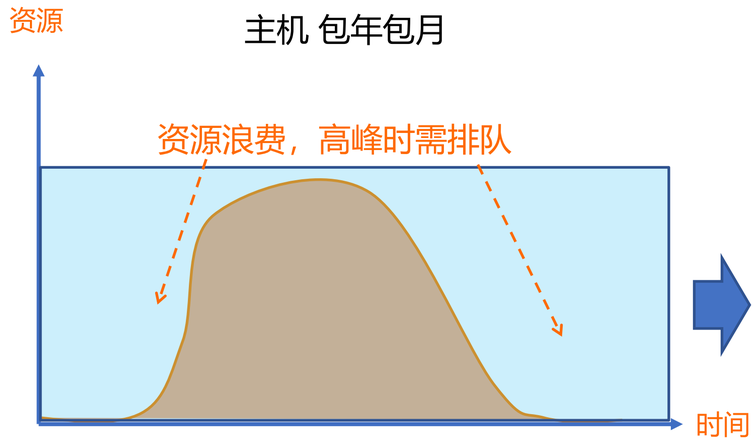
Serverless 计算

函数计算 FC
事件驱动型无服务器计算服务，丰富的云产品集成触发

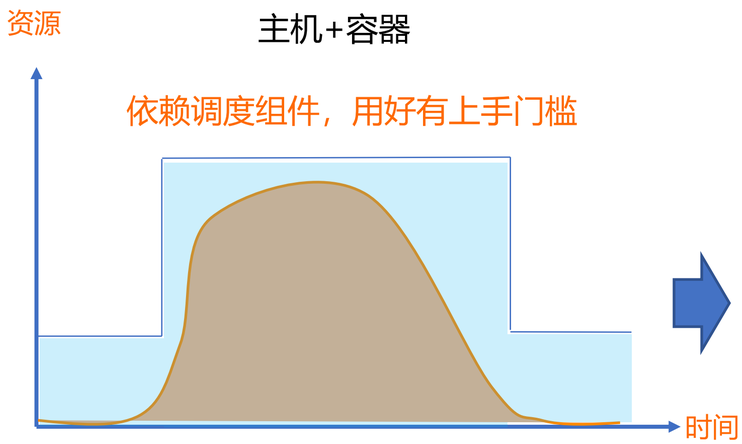
Serverless 应用引擎 SAE
面向应用的全托管、简化运维、高可用 K8s 平台

Serverless先进生产力，让「应用开发运维」更简单

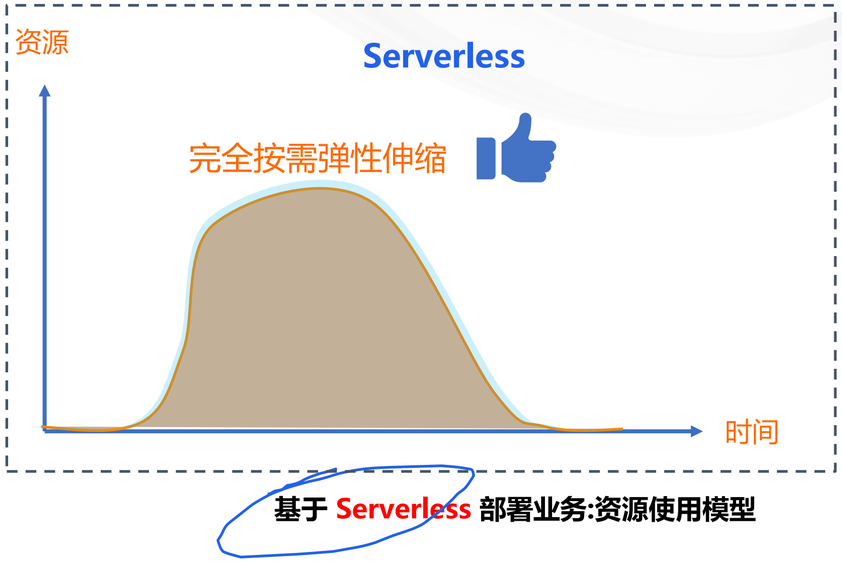
Serverless 会极大促进企业技术架构升级，提升效率：资源效率、运维效率、研发效率



基于云主机方案：资源使用模型



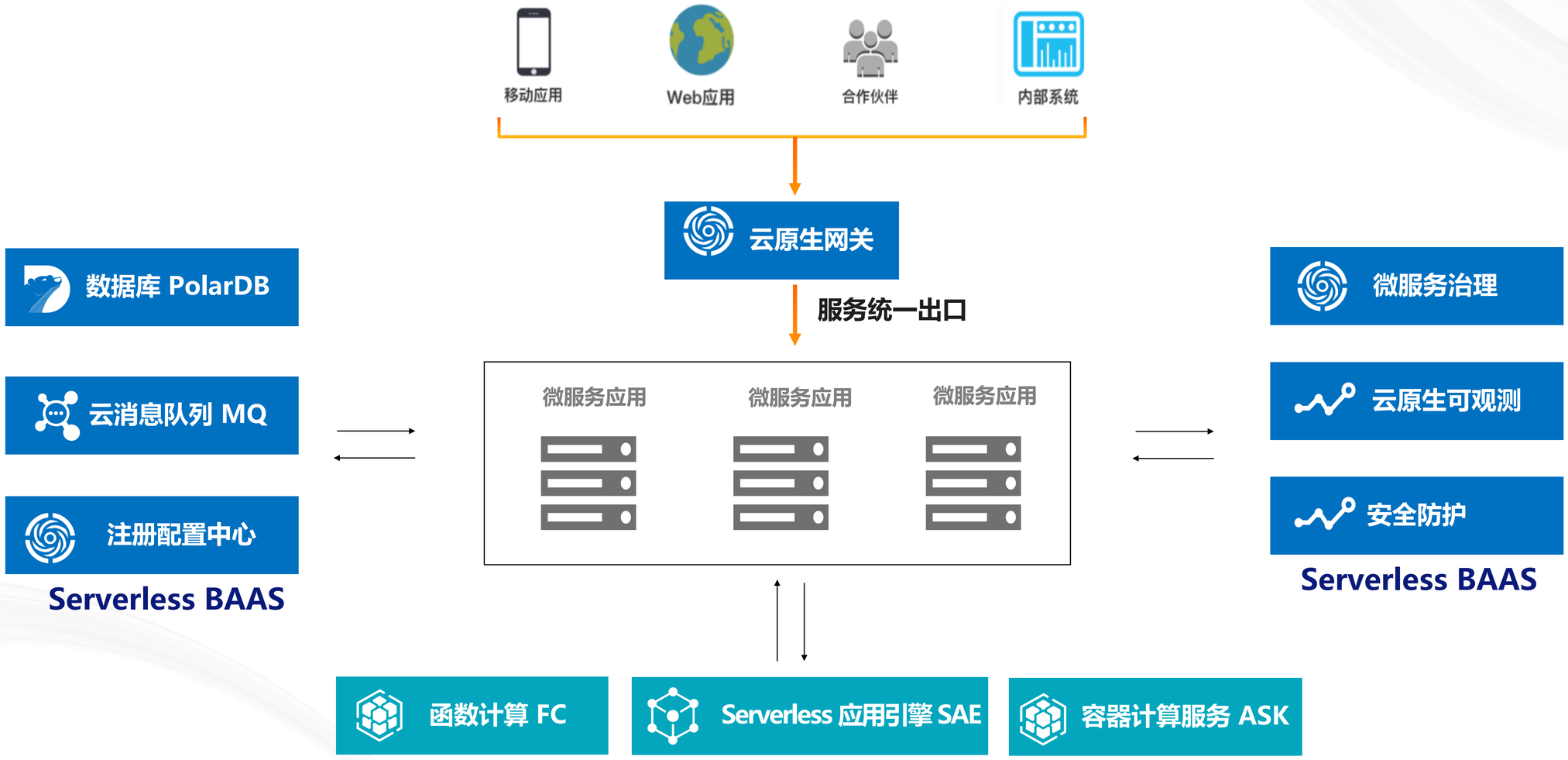
基于K8S 方案：资源使用模型



基于 Serverless 部署业务:资源使用模型

资源管理革命	从人工运维	到云平台工具运维	到 Serverless 免运维	资源免维护
资源利用率革命	从预算采购低利用率	到有限弹性高利用率	到 Serverless 100%资源利用率	业务高弹性
资源成本革命	从固定成本支出	到根据资源策略伸缩	到 Serverless 根据业务策略适配	平台低成本

基于 Serverless 架构的分布式应用



SAE 适用场景：IDC 上云/云上架构升级/EDAS 升级方案

IDC 转公有云
平滑迁移，快速上云

1. 对标虚拟机迁移上云：
   - 可免资源购买、维护
   - 可具备自动弹性
   - 微服务免改造
   - 单体需存算分离
2. 对标容器化迁移上云：
   - 镜像平滑迁移，CD 对接
   - 简化 K8s 管理维护
   - 提升系统弹性能力
   - 提升系统容灾能力
   - 全套微服务治理、应用可观测

阿里云架构升级
降本/弹性/简化运维

1. 升级 ECS 上部署的微服务：
   - 平滑迁移，多种灰度策略
   - 可提升资源利用率 + 弹性效率，简化维护
2. 升级 ECS 部署的单体应用：
   - SAE = ECS + LB + NAT网关 + 应用监控 + 应用管理 + 不停机灰度策略 + 秒级弹性
3. 升级自建 K8s + 微服务：
   - 平滑迁移，简化运维
   - 屏蔽 K8s 复杂度

阿里公共云 EDAS 升级
降本/弹性/简化运维

1. 公共云 EDAS 新用户会推荐升级到 SAE，存量 EDAS 客户会引导迁移到 SAE
2. 升级 EDAS 到 SAE：
   - SAE = EDAS + ECS/K8s
   - 运维更简单，更易用
3. 迁移方案：
   - 平滑迁移，支持不停服
   - 迁移工具辅助迁移

SAE 适用的软件架构：微服务，单体，Job（任务）

兼容的部署平台：容器、K8s、ECS

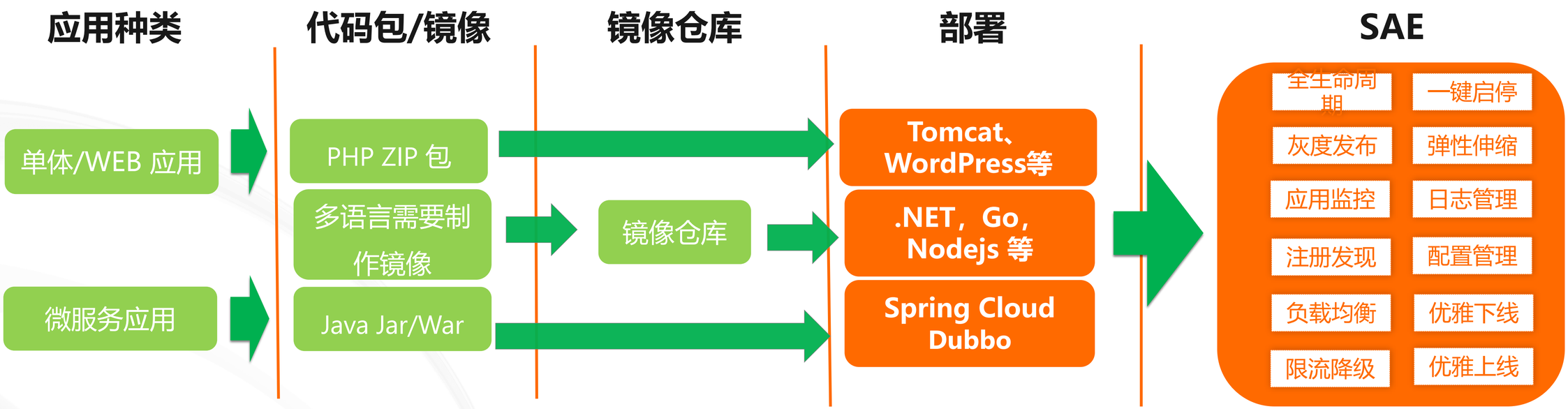
从传统架构到 SAE 的平滑路径

02

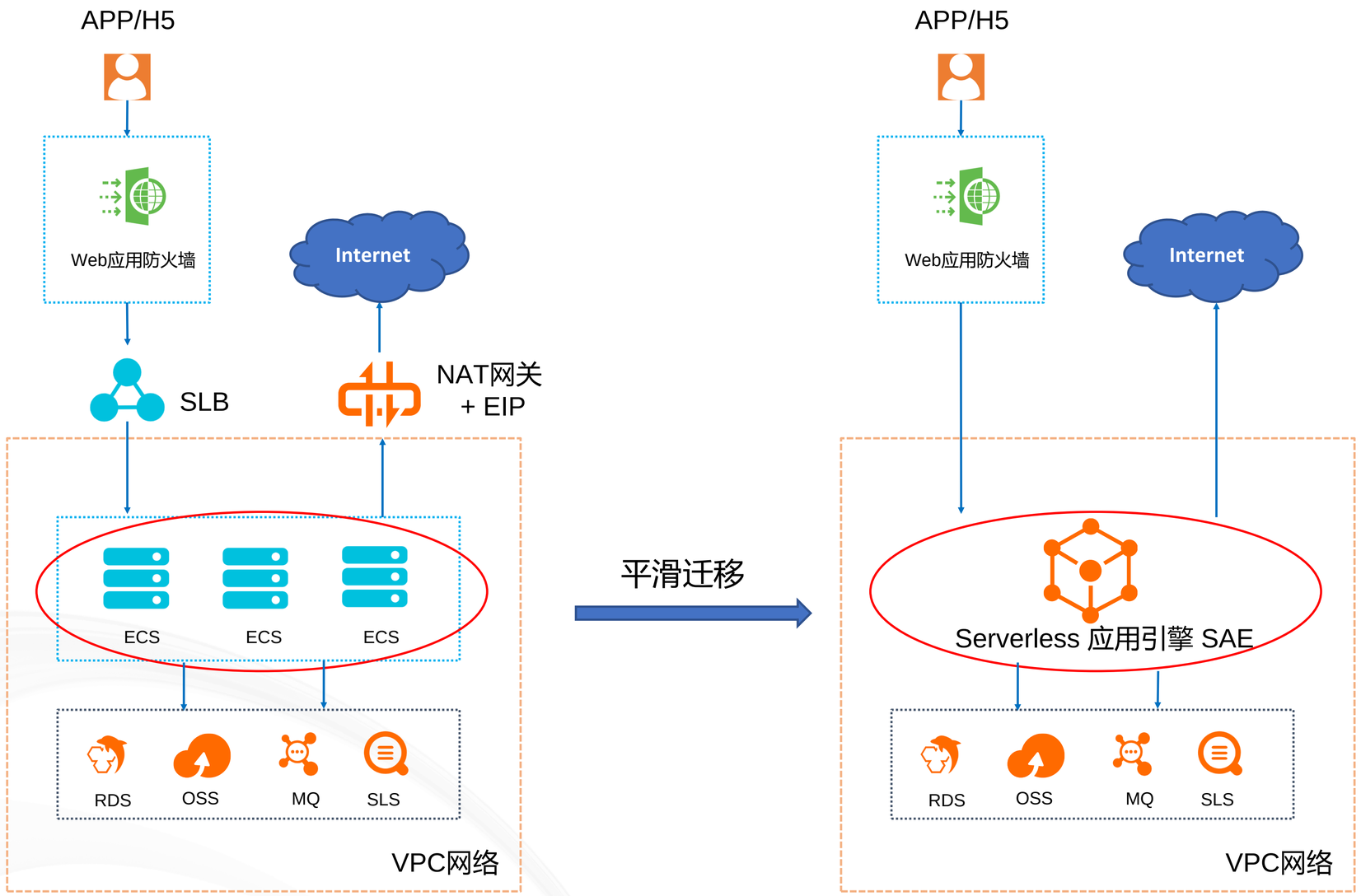
微服务/单体应用代码平滑迁移

资源使用方式	资源准备阶段	研发&部署阶段	运维阶段
云主机自建	购买ECS 创建集群 集群初始化	部署件业务代码到 ECS 部署监控日志组件到 ECS	运维 ECS（高可用，弹性） 运维监控日志组件
使用SAE	免购买 免创建集群	代码平滑迁移 Jar/ZIP/镜像	免运维

代码部署方式



场景一：单体应用（已经存算分离）架构迁移



使用问题：

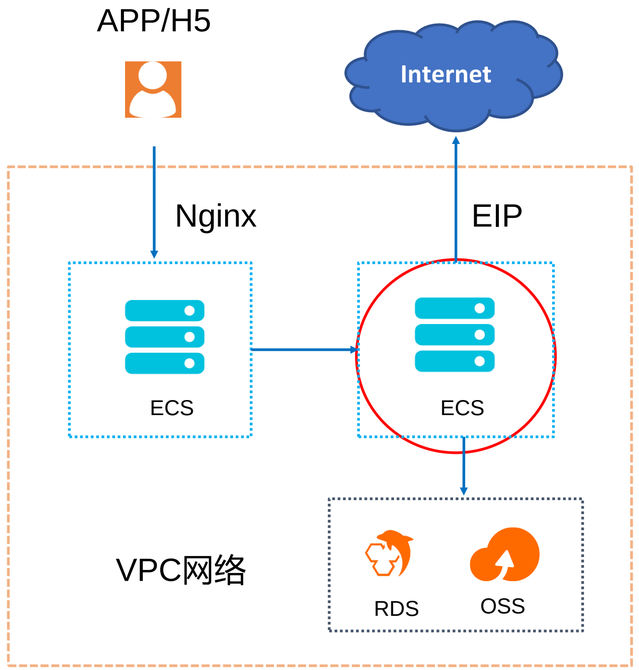
1. 原单体服务运维复杂：
   a. 弹性扩缩复杂
   b. 灰度发布复杂
   c. 缺少应用级监控
2. 存在资源浪费

迁移价值：

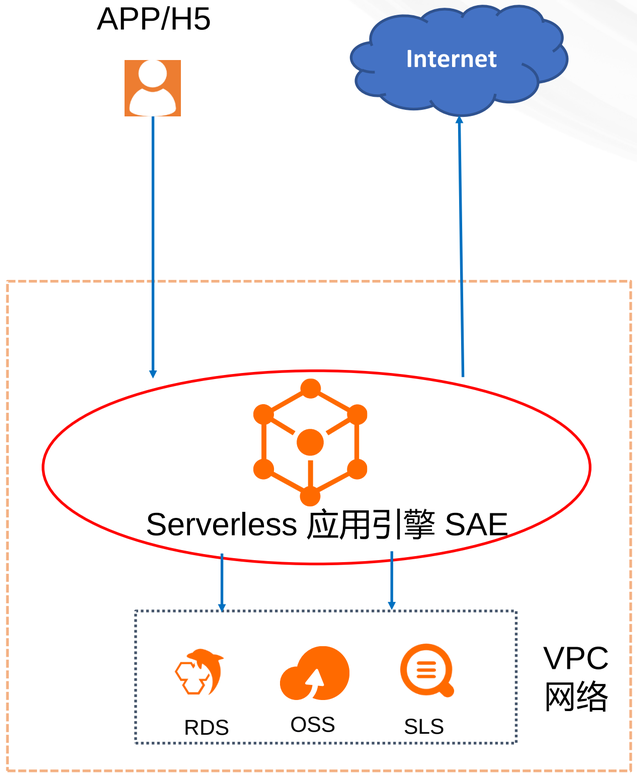
1. 简化运维：
   a. 扩缩容简单，秒级弹性
   b. 支持跨可用区容灾
   c. 支持精细化的版本发布
   d. 支持请求级别的监控能力
2. 成本优化：
   a. 仅在有请求时分配 CPU
   b. 降本幅度高达 30%（有波峰波谷）

场景二：单体应用迁移到 SAE

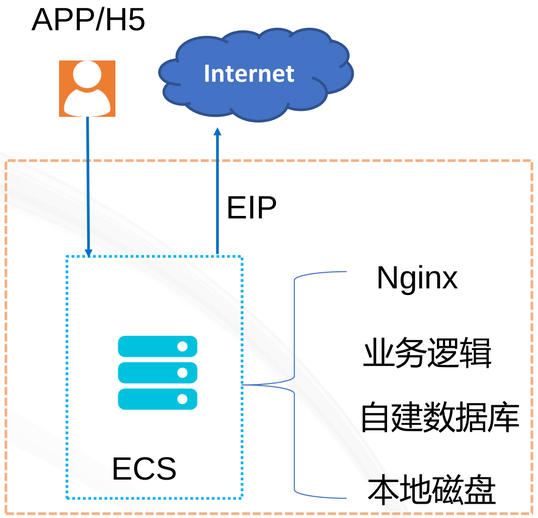
部署方式一
Nginx +
已经存算分离



推荐迁移方案



部署方式二
大单体

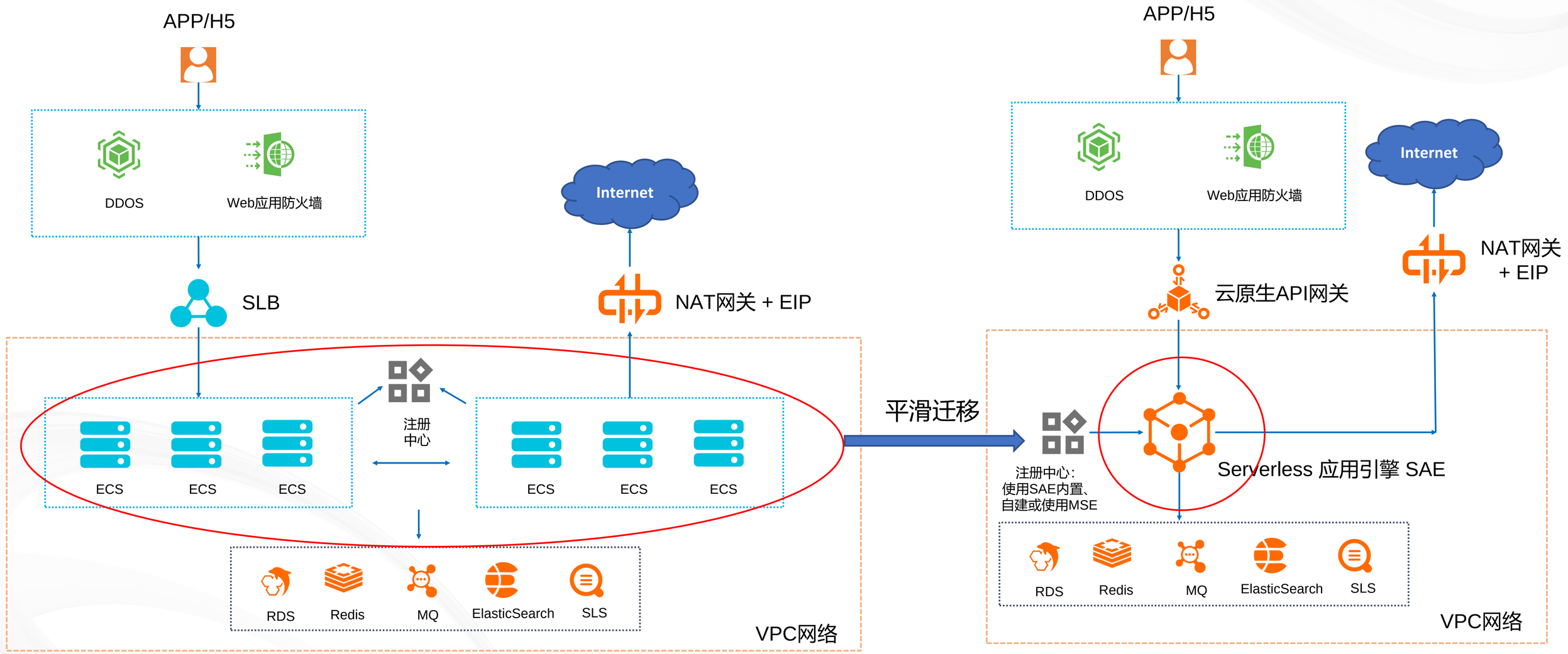


推荐迁移方案

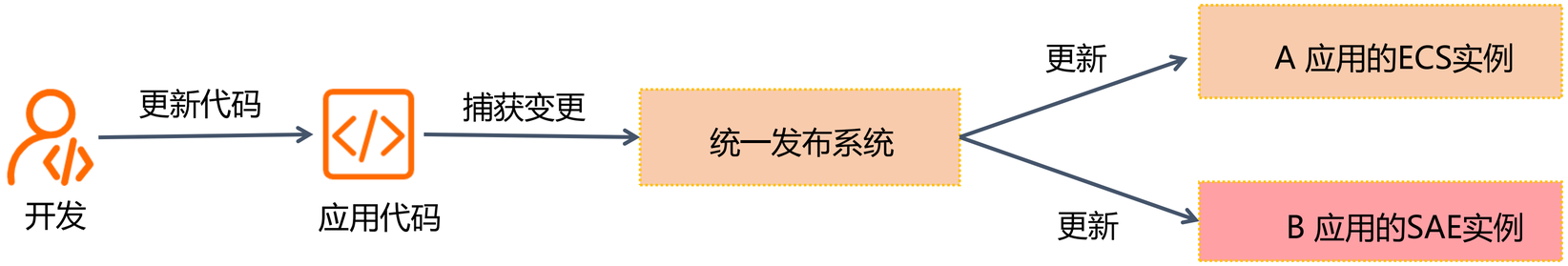
1. Nginx 可以直接省掉，如果配置过于复杂，可把 Nginx 也部署到 SAE 上
2. NAT网关+EIP模式可以直接用 SAE 的默认出公网能力，如果需要固定公网出口IP，可以保留 NAT网关+EIP
3. 大单体需要拆分成存算分离架构，SAE 代替其中业务逻辑部分：易扩展、稳定性好

微服务架构“0”改造上云/升级到 SAE

只用把代码从 IDC/ECS 虚拟机以代码包或者镜像的方式部署到 SAE，中间件、数据库等可复用，支持存量业务混部、互访

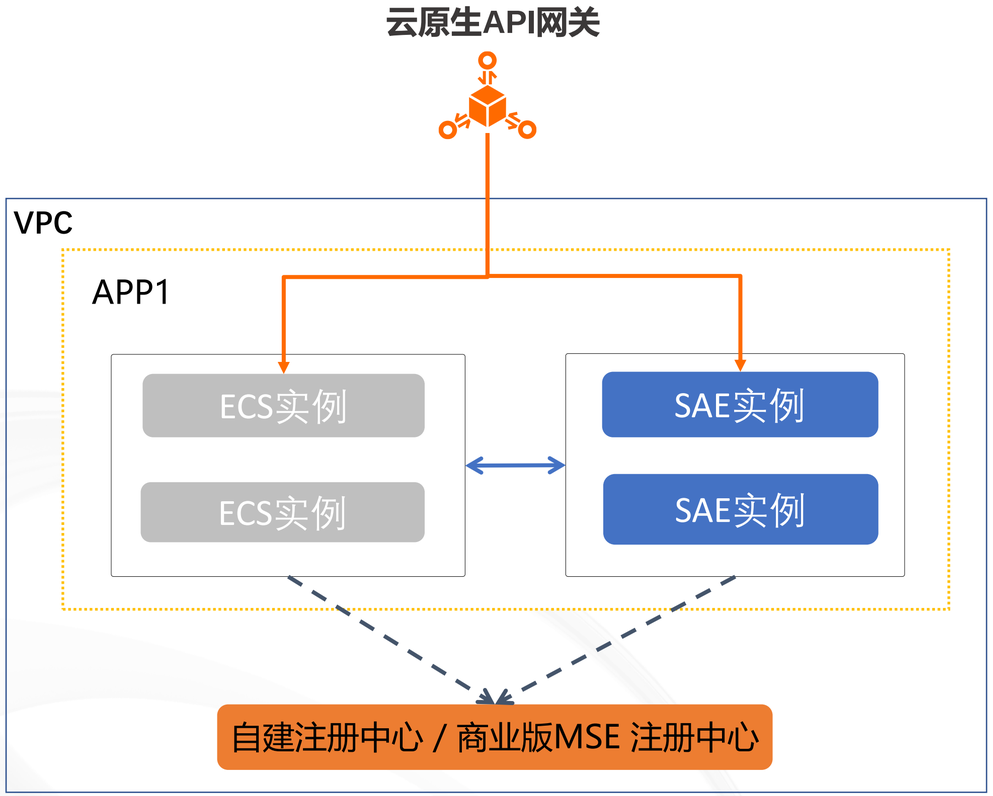


从自建逐步切换到 SAE 部署模式：以应用粒度迁移



迁移步骤：

- 发布系统需要同时对接 SAE 和ECS；
- 选取独立应用部署到 SAE；
- SAE 和 ECS 混跑，逐步迁移。



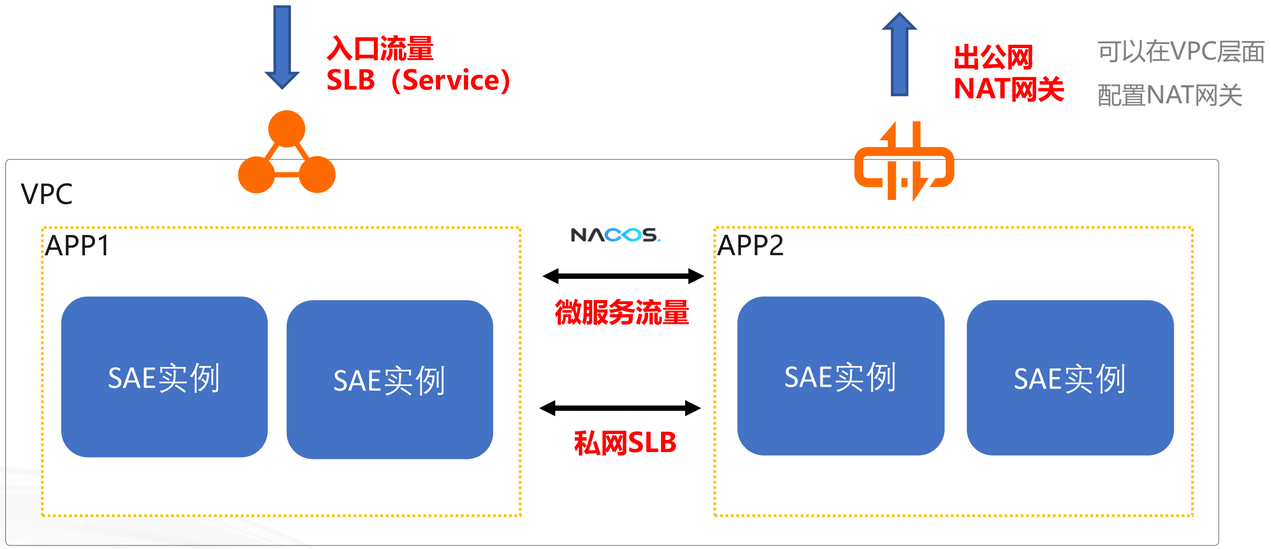
方案说明：

1. 统一注册中心：SAE 上部署的应用和 ECS 上部署的应用统一都注册到自建注册中心（ZK/Nacos/Eureka）或者商业版 MSE 提供的注册中心。
2. 应用间网络通信：需要保证 SAE 上的应用和 ECS、注册中心都在一个VPC 内部。
3. **云原生API网关：关联证书和域名，通过配置网关路由把流量转发到 SAE。**
4. 日志服务：统一投递到 SLS。
5. 监控数据：可以统一上报到云监控、ARMS 或者 Prometheus。

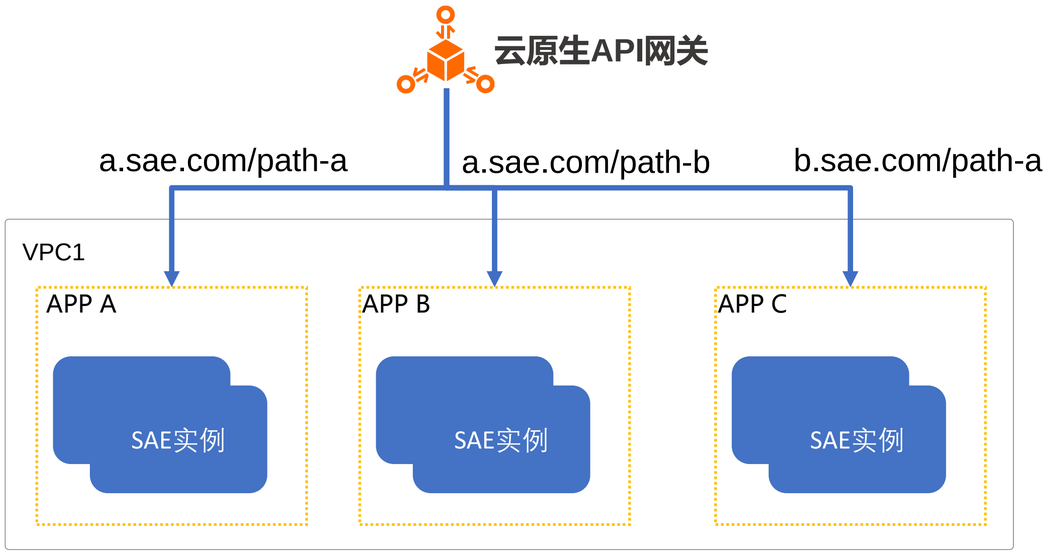
基于 SAE 的流量管理

南北流量

- SLB 为公网流量入口
- 如果 SAE 实例需要主动出公网，需要借助 NAT 网关 + EIP



- 网关路由：一个 网关 + 域名，通过一个端口，根据不同的 Path 路由到不同的应用

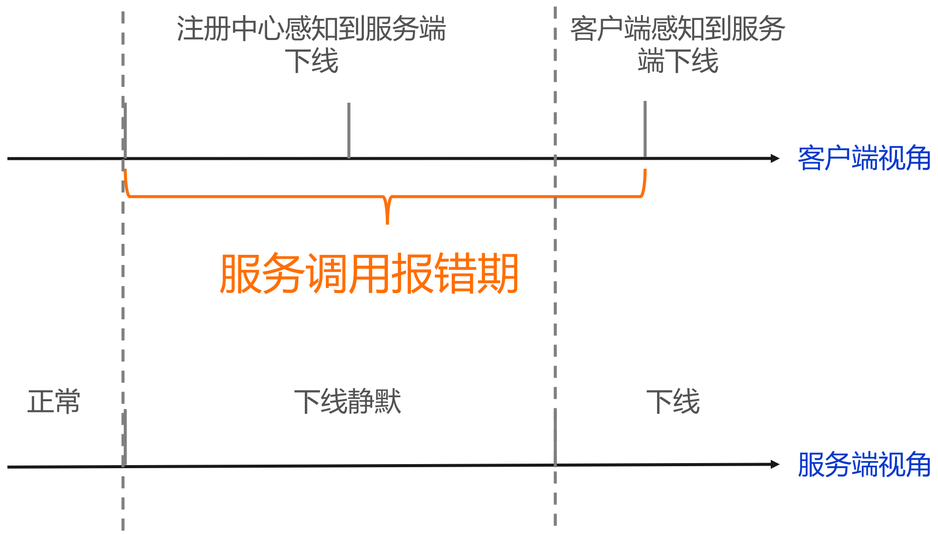


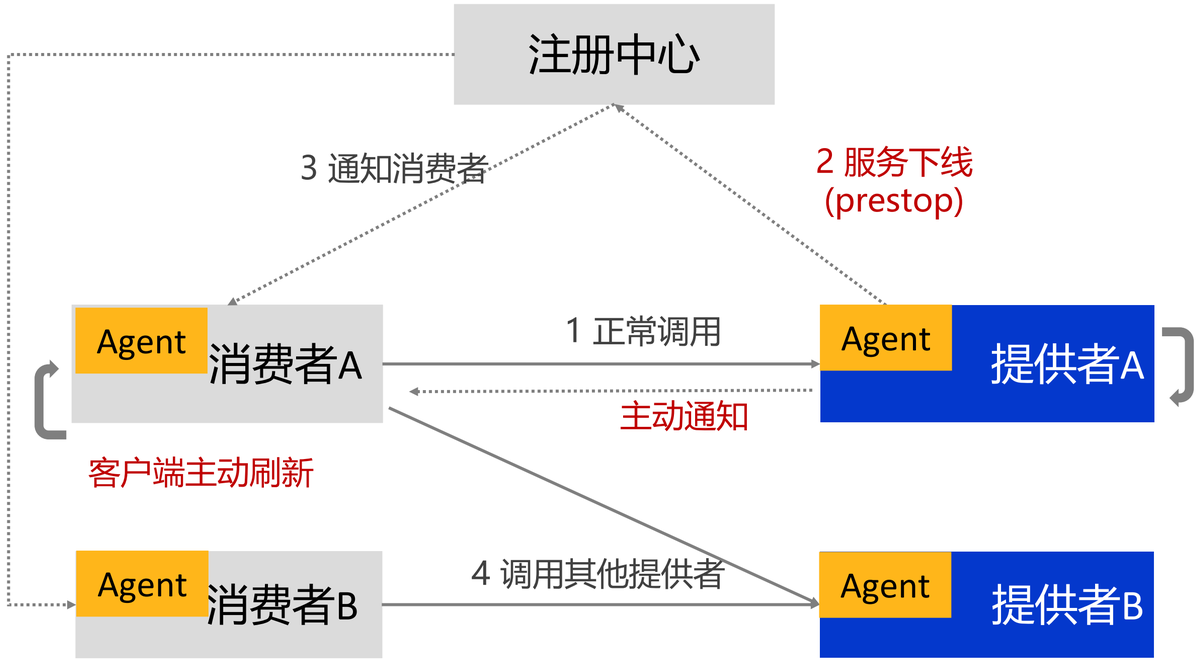
东西流量（微服务）

针对微服务应用的东西向流量，SAE 托管了注册中心，
也可以使用自建注册中心，以及众多微服务治理功能

全套微服务治理-无损下线

通过无损下线，有效避免服务端升级时客户端调用报错，实现在白天大流量发布也能丝般润滑

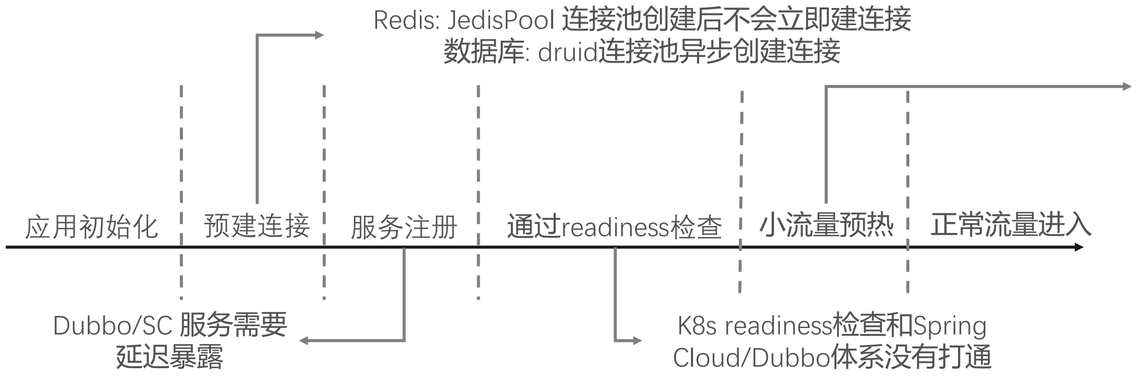




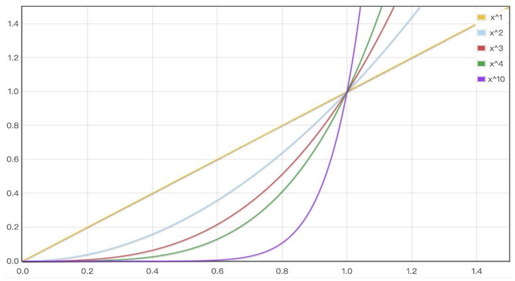
SAE方案 VS 开源自研的优势： 1.延时更短 2. 稳定性兜底 3. 无需投入研发

全套微服务治理-无损上线

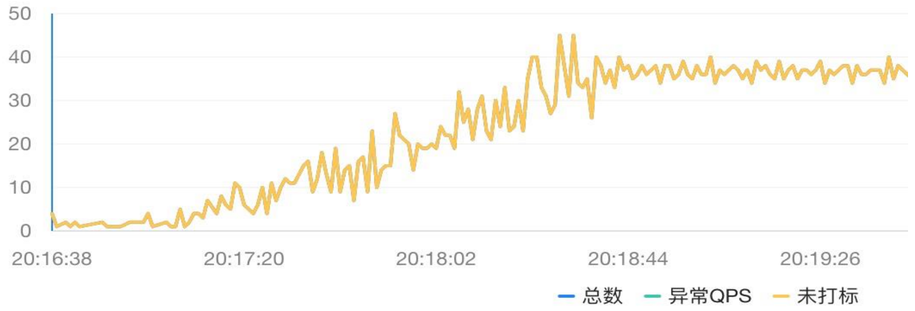
通过无损上线，避免应用尚未启动完成就承接正常流量导致实例被打挂，通过延迟注册和小流量预热实现扩容时流量无损。



- 流量需要缓慢增加，但Dubbo 2.7.4.1以下版本存在服务预热不生效问题
- Fastjson/Jetty低版本没有开启并行类加载
- JVM JIT 编译问题引起cpu飙高
- 日志异步化



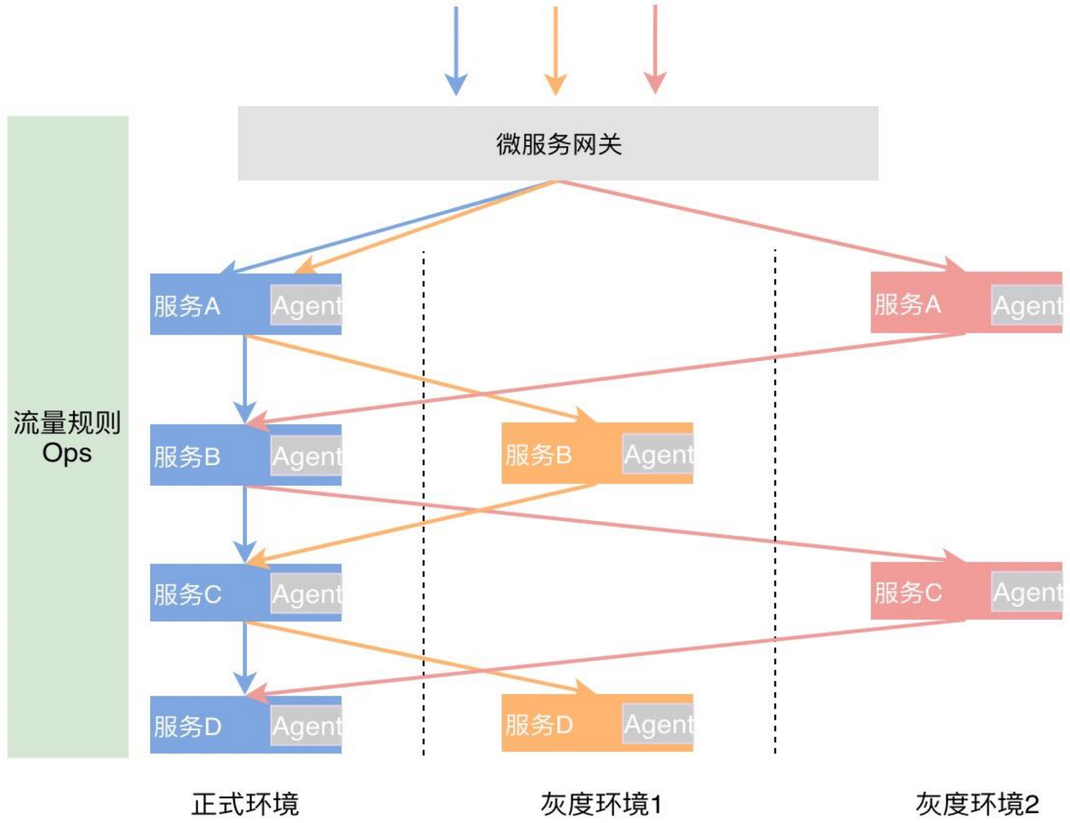
预热曲线动态调整



SAE方案 VS 开源自研的优势： 1. 无需投入研发 2. 稳定性有保障 2. 白屏化使用，体验更简单 3. 完整的可观测能力

全套微服务治理-端到端的全链路灰度发布

通过全链路灰度发布，白屏化配置规则，实现精准灰度能力，比传统多环境部署的方案大幅降低了运维和机器成本。



适用场景

- 调用链中因个别/部分应用新上线，需要精准用户灰度验证

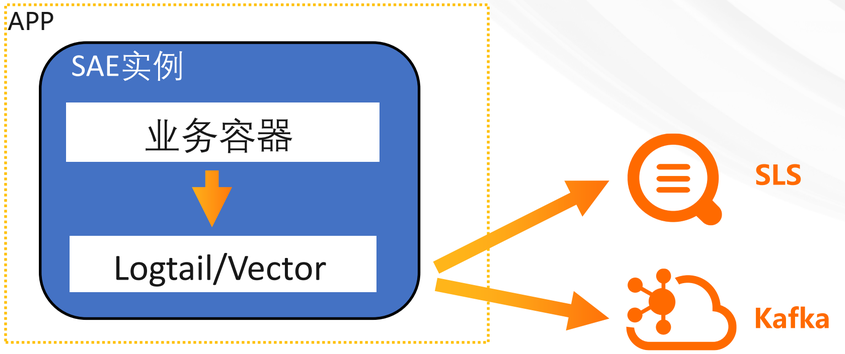
价值

- 控制最小爆炸半径
- 无需重新搭建一整套新环境，降低部署运维和硬件成本
- 能覆盖七层流量到四层微服务接口、方法的灰度

日志管理

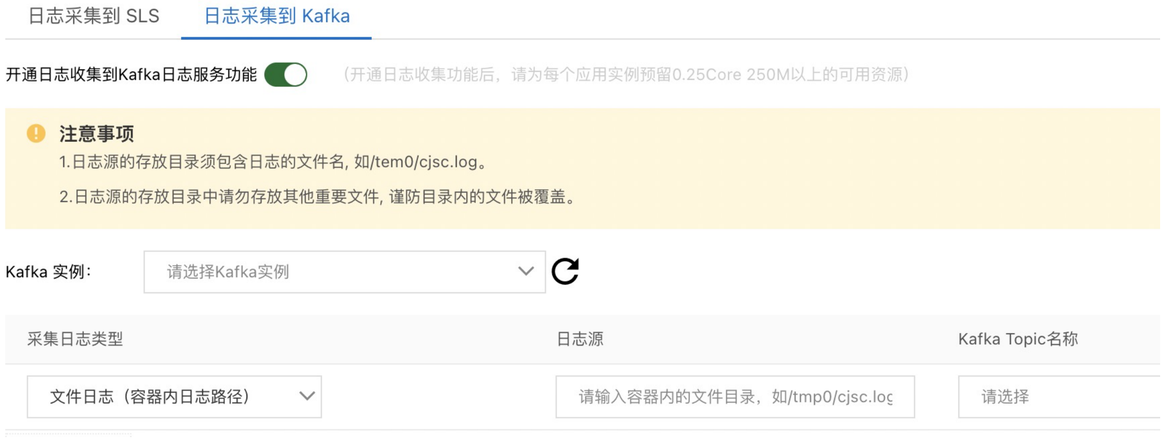
➢ 最佳实践：设置日志收集至 SLS

- SAE无缝集成了SLS，在 SAE 中部署的应用建议将日志输出到文件，并配置SLS 日志采集，并在 SLS 侧做进一步的分析、可视化、告警等。
- 同时也可以配置将 SAE 运行的代码，打印的 stdout 标准日志重定向输出到 SLS。
- SAE 默认将标准输出的内容收集并展示在 SAE 的实时日志中，支持查看最近500行日志信息，方便快速定位问题。



➢ 最佳实践：设置日志收集至 Kafka（ELK、日志分析等场景）

- SAE 无缝集成了 Kafka，在 SAE 中支持将应用实例的标准输出（stdout），以及将应用实例指定路径的日志采集到消息队列 Kafka版 中。
- 适用于需要将 Kafka 的数据投递到例如 Elasticsearch 等其他持久化库中，便于集中管理和分析日志。



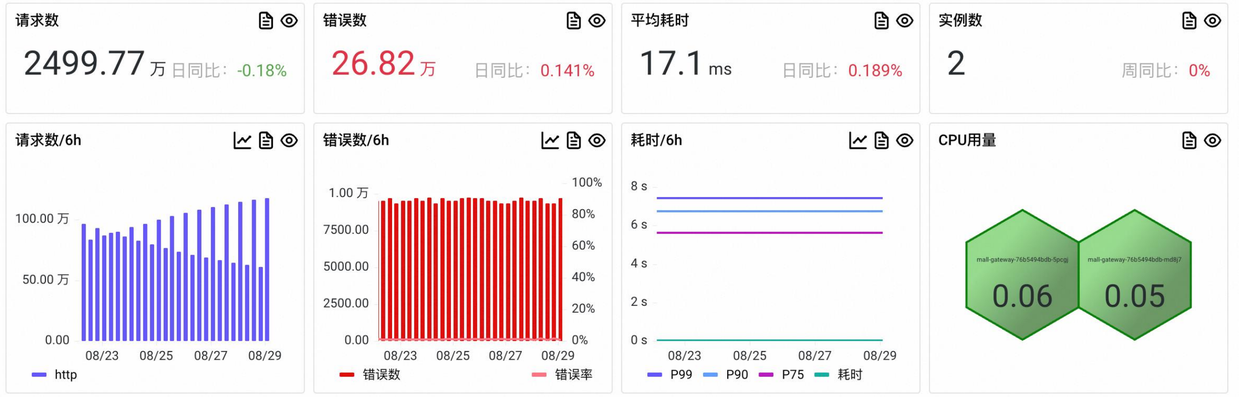
➢ 不推荐：将日志写入到 NAS 盘

- 文件存储 NAS 为多客户端提供了统一名字空间的文件共享读写能力，但在多进程/多客户端并发写同一个文件的场景中会出现文件覆盖、传行等问题，解决方案：
- （推荐）不同进程/客户端写入同一文件系统的不同文件中，后续分析处理时再进行归并，这个方案能够很好地解决并发写入导致的问题，同时无需使用文件锁，不会对性能造成影响。
- 对于并发追加写同一个文件（如日志）的场景，可以使用文件锁 + seek 机制来保证写入的原子性和一致性。但是文件锁 + seek 是一个比较耗时的操作，可能会对性能产生显著的影响。

应用监控

➢ 基础监控：云监控

- 应用迁移到 SAE 后，基于云监控可以查看实例的运行信息， SAE 会对应用所运行设备的CPU、负载、内存、网络和磁盘进行数据采集与分析。
- 在云监控试图下，可以统一查看所有云服务的监控大盘。
- 注意事项：由于从数据采集到分析存在一定延时，时延约2分钟。如果托管在SAE上的应用在升级或扩缩容导致 Pod 发生变更，那么将会导致监控数据产生断点，不能保持连续。
- 支持调用云监控的 API 接口来拉取基础监控数据，上报到自建监控系统。

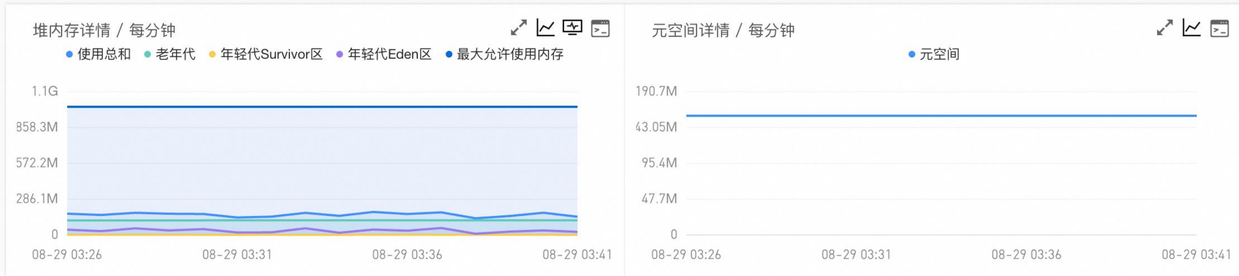


➢ 应用监控：SAE 集成了 ARMS 无侵入提供应用级监控

- SAE 目前提供的应用监控主要是 JAVA 语言和 PHP（5.2～7.3）；其他语言的应用监控还在补充。
- 应用监控：支持选定时间内的总请求量、平均响应时间、实时实例数、Full GC次数、慢SQL次数，以及这些指标和上周、上一天的同比升降幅度等。

➢ 应用监控：支持对接 Prometheus，构建统一监控

- 在云上开通 **Prometheus for VPC 版本**，或者自建 **Prometheus** 。
- 采用服务框架自带的上报组件进行应用监控数据的投递。
- 注意：产品化的 Prometheus 投递能力还在对接开发中。

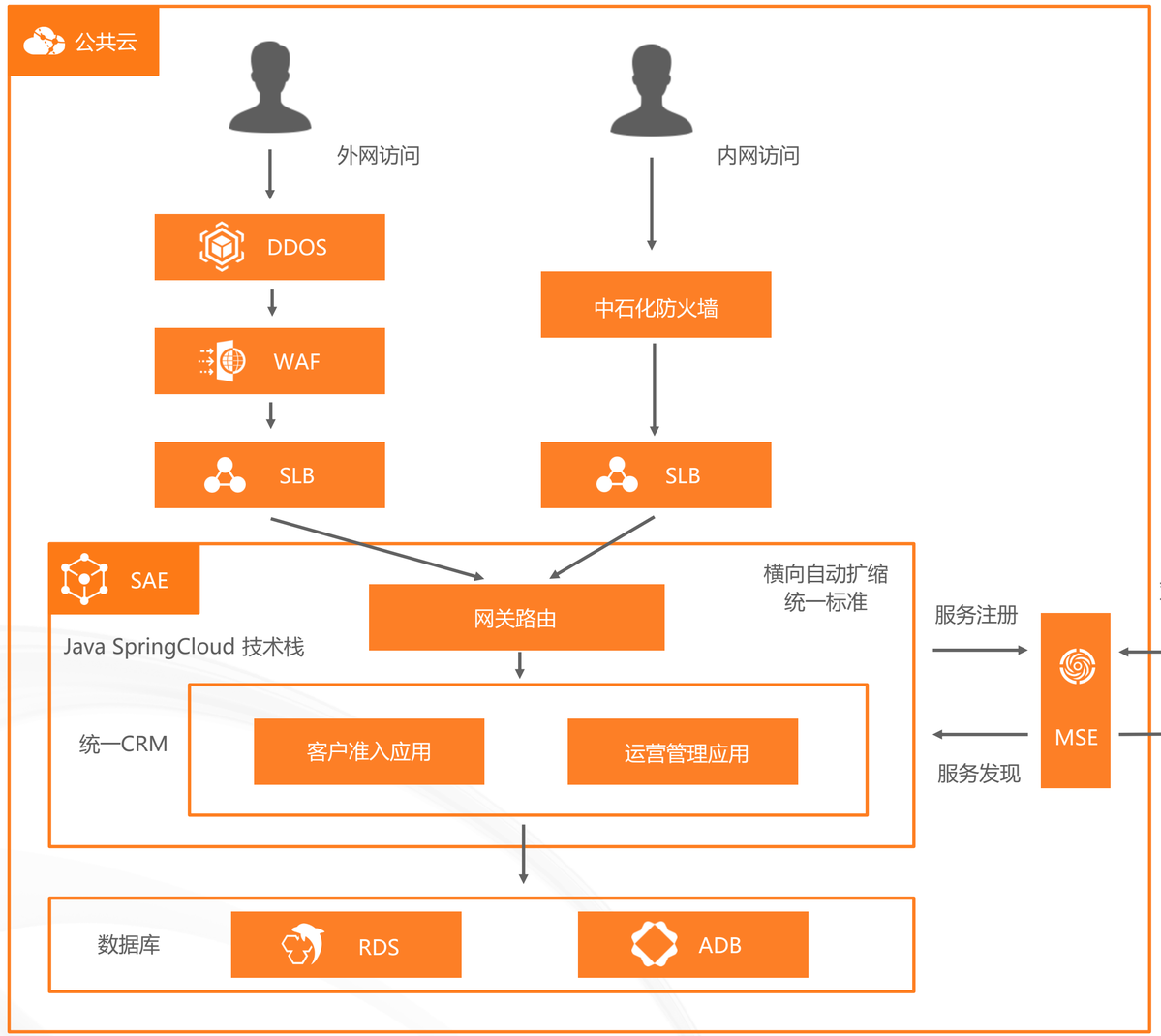


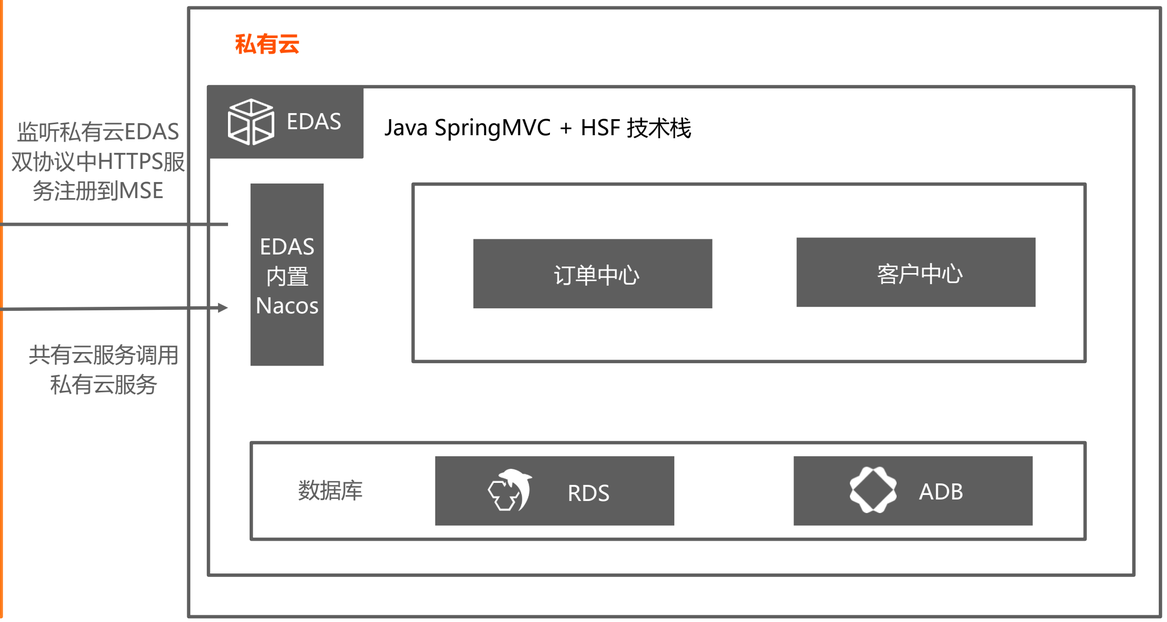
企业落地 SAE 的案例

03

传统能源行业 Serverless 新突破 – 中国石化

以 Microservice on Serverless 方案为核心指导，以构建基于 SAE 的应用层横向扩缩标准为核心抓手，逐步推进新、老系统转型



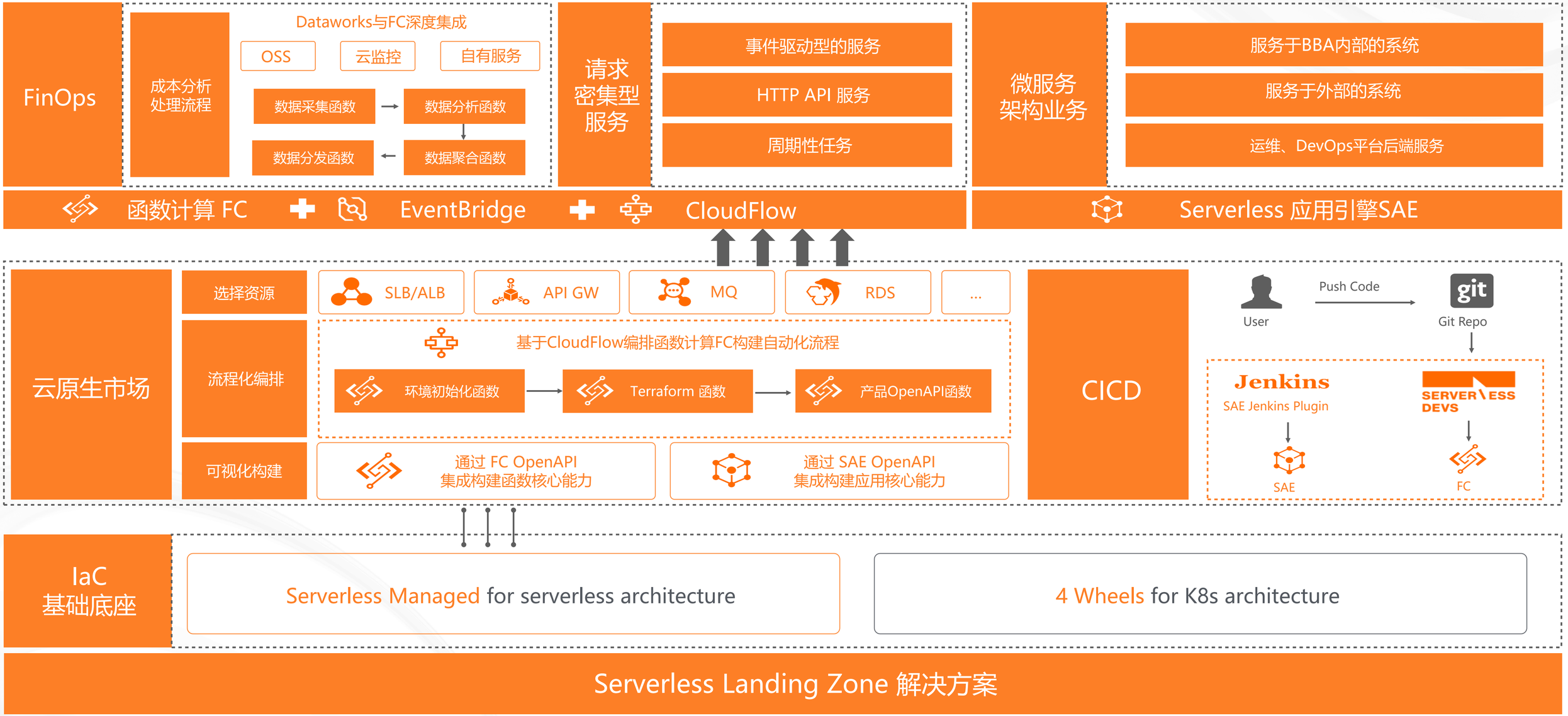


多个系统完成 Serverless 转型

- 中石化危化品物流系统-运单管理模块，企业地图模块，任务系统
- 中石化 CRM 系统
- 中石化电子钱包系统

构建企业级 Serverless 转型标准 – 中国宝马

以 Serverless Landing Zone 方案为核心，构建 Serverless 环境 IaC 标准，构建统一标准，逐步推进上层业务转型



茶饮头部企业- 全栈云原生助力业务高速发展

门店1000→3000家，运维团队持续一个人，三年增长6倍，周年活动7天筹备，0故障支持120万QPS，业务遍布马来、泰国、新加坡、美国、日本、韩国等

客户痛点

- **业务快速发展要求**：业务增长数倍，线上线下快速发展、全渠道融合，IDC中部署、SaaS托管等，不能满足业务的快速发展需求，维护复杂度和成本高，容易出现稳定性问题。
- **大促活动频繁业务峰值变化大**：大促、秒杀、521、周年庆、代言人、健康大使等各种活动频繁、活动常态化，在IDC中需要按照峰值做整体规划，长期保持较多资源，使用率低，成本高。
- **自建可观测体系维护成本高**：开发人员和运维人员少，增长迅速，客户可观测体系缺失或自建，维护复杂度和成本都越来越高，部分快速定位和根因排查

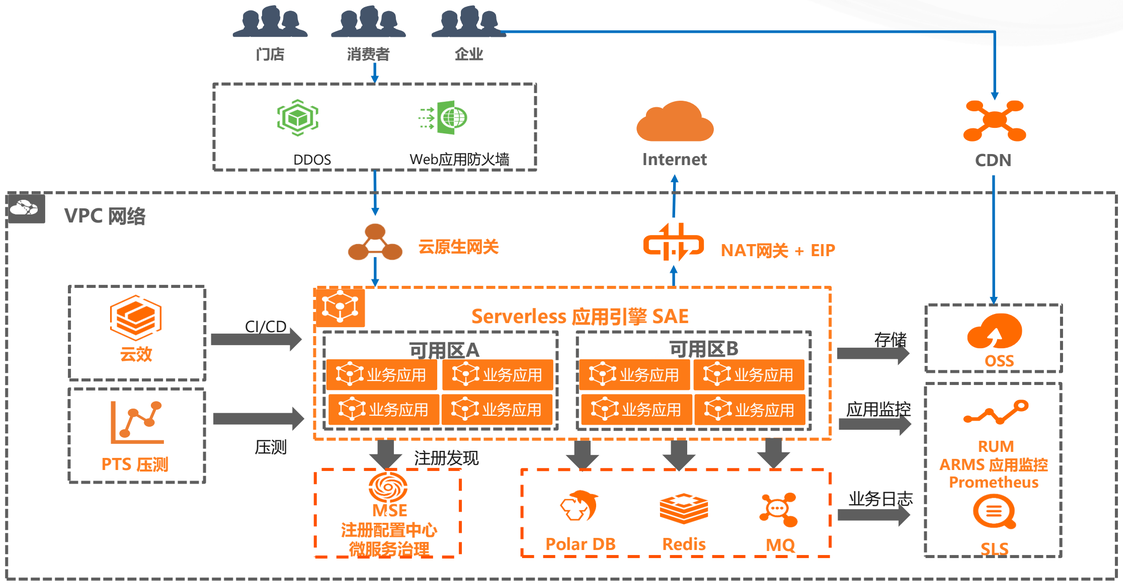
方案亮点

经历过IDC、SaaS托管，在上云过程中，全栈采用云原生的方案，包括应用构建基于SAE 应用引擎、MSE微服引擎、完整的可观测体系，消息服务等：

- **基于SAE构建全部应用，免运维、高弹性、全托管**： 无论是toB还是toC业务，都基于SAE快速部署，大促活动期间及业务高峰等，配置策略，按照业务流量自动弹性扩容，运维人员少，大大减轻了运维工作量，降低了资源和运维成本。
- **提升系统稳定性**：云原生网关、MSE微服务注册中心、所有应用接入服务治理，提供无损上下线能力，前端灰度，全链路灰度等，构建业务高可用。网关支持业务活动的限流，构建了支撑百万QPS流量洪峰的网关及后端系统处理能力。
- **电商业务异步解偶**：大促活动基于rocket MQ构建了海量优惠券发券，订单、配送、客户服务等异步系统解耦，提供可以支持核心业务的高可靠的消息处理。
- **构建完整的可观测体系**：ARMS应用监控，RUM前端用户体验监控，rum-网关-arms-SLS完成的链路分析，Prometheus接入所有云服务，构建完整的统一智能告警体系，SLS构建分布式系统的统一日志服务，Grafana构建面向领导层、研发、运维等不同角色的监控大盘。
- **业务监控构建**：Rum自定义指标、事件支持大促活动的活动转化效果实时掌控，SLS的业务日志构建基于业务流程的稳定性大盘和业务大盘。

相关产品：

- SAE Serverless应用引擎
- MSE微服务引擎
- Prometheus 服务
- Rocket MQ消息服务
- 应用实时监控服务 ARMS
- 用户体验监控 RUM
- 日志服务SLS
- Kafka消息服务



建设成果

全栈基于云原生，构建了完整的云原生应用平台，满足业务的高速发展、上市，业务出海，全球布局。全栈云原生支撑业务快速扩张、业务创新，自动提供业务高可用保障，**整体系统设计为可以支撑百万QPS活动**，经历了多个节日、秒杀、周年庆等日常促销和大促活动。SAE帮助客户大大节约了运维成本，**构建的完整的可观测体系**，提供端到端的全链路监控，智能告警，智能洞察，根因洞察等。基于微服务引擎，助力客户快速构建业务活动和业务创新，构建多套全链路灰度环境进行验证及流量管理，满足客户业务的快速迭代。

客户证言：B端和C端系统全部跑在阿里云Serverless应用引擎 SAE上，日常活动比较多，大大减轻了运维压力，不需要平时保有大量机器，运维效率提升了50%，业务可靠性达到了99.99%。

AI网关落地实践

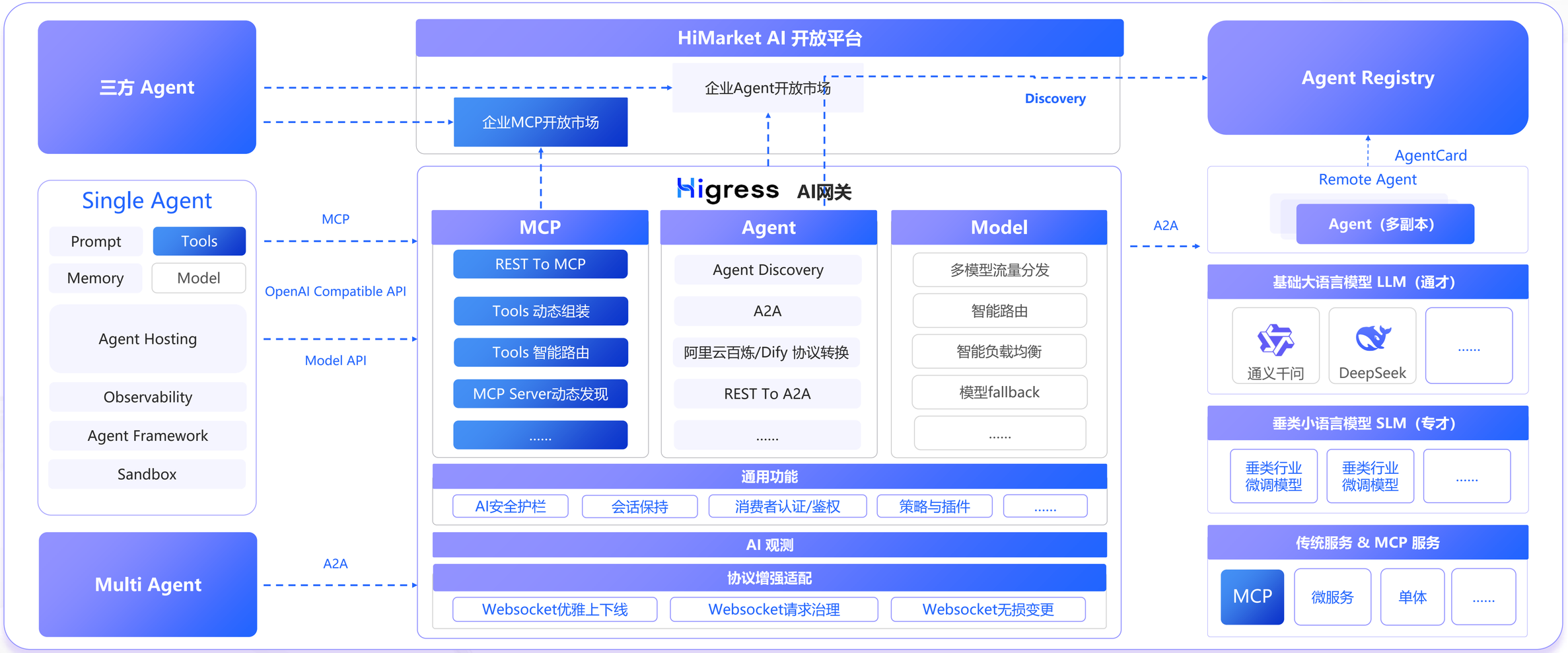
04

阿里云

AI 网关在 AI 原生架构中的位置

The Position of the AI Gateway in the AI-Native Architecture

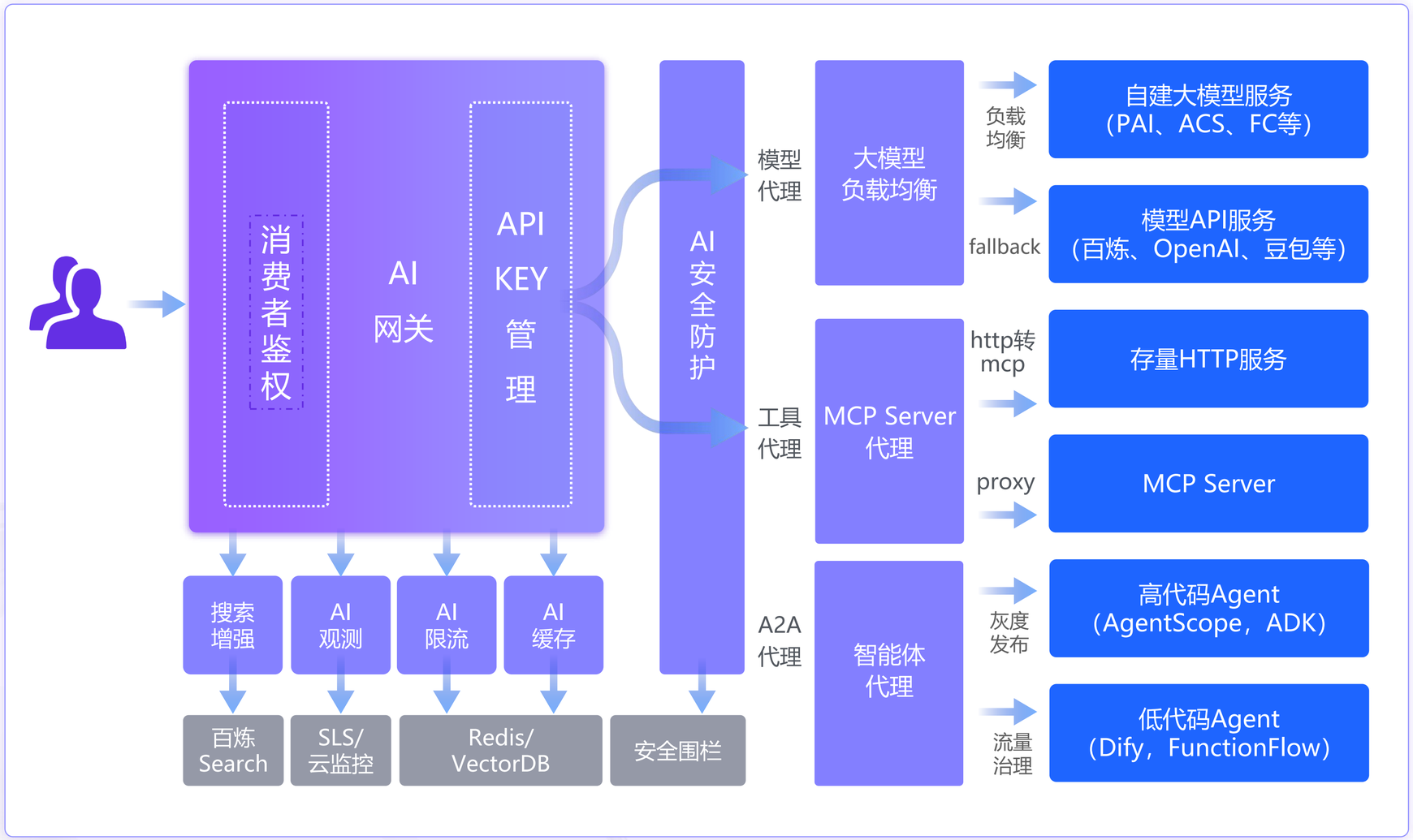
AI 网关是 AI 原生架构的关键基础设施



AI 网关提升 AI 应用企业级能力

AI Gateway Improves Enterprise-Grade Capabilities of AI Applications

企业AI应用访问大模型，工具，多智能体的统一出口，解决企业流量管理，安全，协议转化等一系列企业诉求



面向失败设计提高可用性

通过多模型路由，Failover、灰度发布、并发控制、限流等能力将模型调用可用性提升到 99.9%+

预集成安全能力

对 API-key，消费者凭证进行统一管理，分权管理
预集成WAF防护，安全护栏能力，提升安全水位

多协议支持

集中管理智能体各种调用，包括模型调用，MCP 工具调用，以及A2A调用

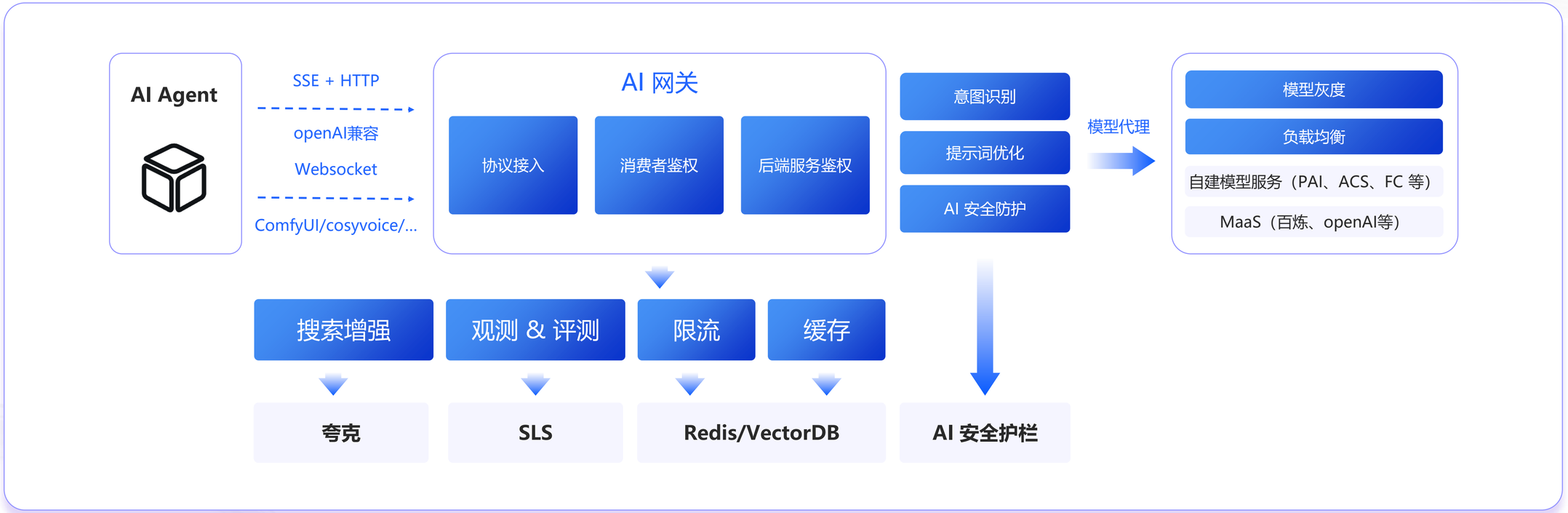
发布 Serverless 版

免容量预估，500QPS 秒级弹性
大幅降低起步成本

模型：从文生文到多模态

Model: from Text-to-Text to Multimodal

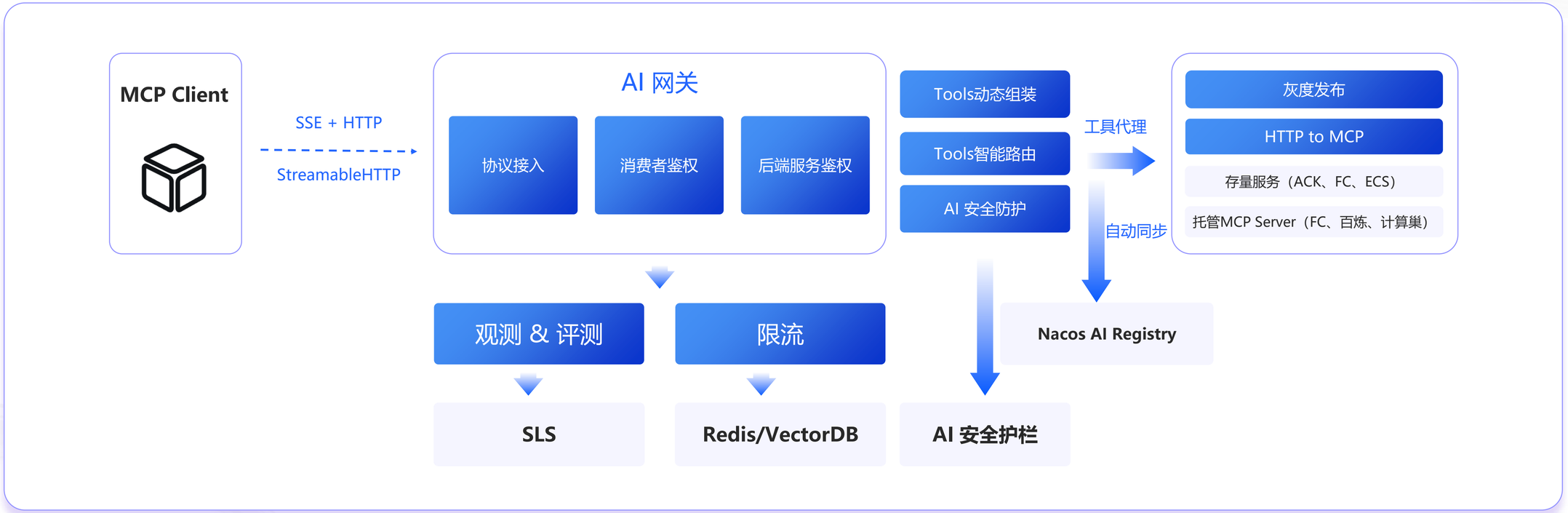
满足多模型、多模态、多场景的诉求，做到高安全、高稳定



从 MCP 代理到工具组装

From MCP Proxy to Tool Assembly

通过手动和智能组装 MCP Server，更好地使用工具



阿里云

云监控为 AI 应用提供全栈可观测能力

Providing Full-Stack Observability for AI Applications

针对大模型应用提供四大核心可观测能力

端到端诊断
无侵入支持主流AI应用框架
与其他AI组件形成全链路诊断能力

全栈监控
覆盖模型应用到智算节点
各层的关键指标和日志

AI/MI平台预集成
与容器服务，PAI，百炼等平台默认集成
一键接入可观测数据采集，默认提供大盘

成本管理及质量评估
基于token/GPU监控
与AI网关限流路由能力

大模型应用 (MaaS)

应用
- RAG、MCP 应用运行监控
- AI App 监控
- 模型调用监控
- 安全诊断

模型
- 模型监控
- 模型日志
- 模型链路

AI 平台 (PaaS)

模型推理
- 服务与组件监控
- 推理引擎监控

服务组件
- 向量数据库监控
- AI 网关监控

模型训练
- 训练任务监控
- 数据流监控

容器 Kubernetes

容器调度
- Workload 可观测、控制面可观测、GPU 监控、Ingress 监控，事件、日志审计，AI 套件可观测

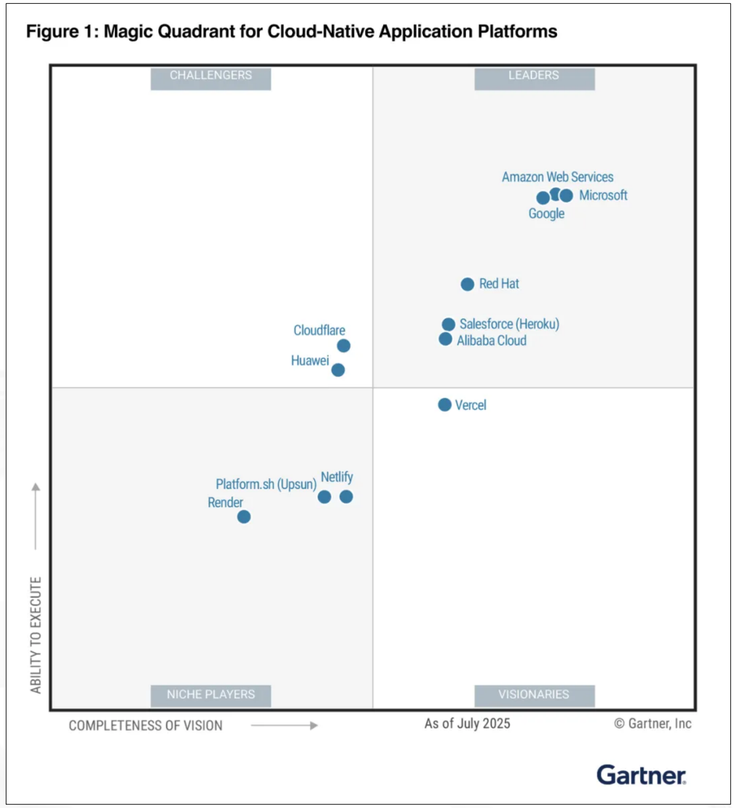
智算服务器 (IaaS)

AI 基础设施
- 灵骏计算节点监控
- 网络 RDMA 监控
- 高速存储 CPFS 监控
- GPU 云服务器监控

Serverless x AI，让应用开发更简单

2025年8月7日，Gartner®发布了2025年度全球《云原生应用平台魔力象限》报告：阿里云凭借Serverless应用引擎SAE（以下简称SAE）和函数计算FC，成为亚太地区唯一入选「领导者象限」的科技公司。

阿里云获「Gartner云原生应用平台」领导者象限

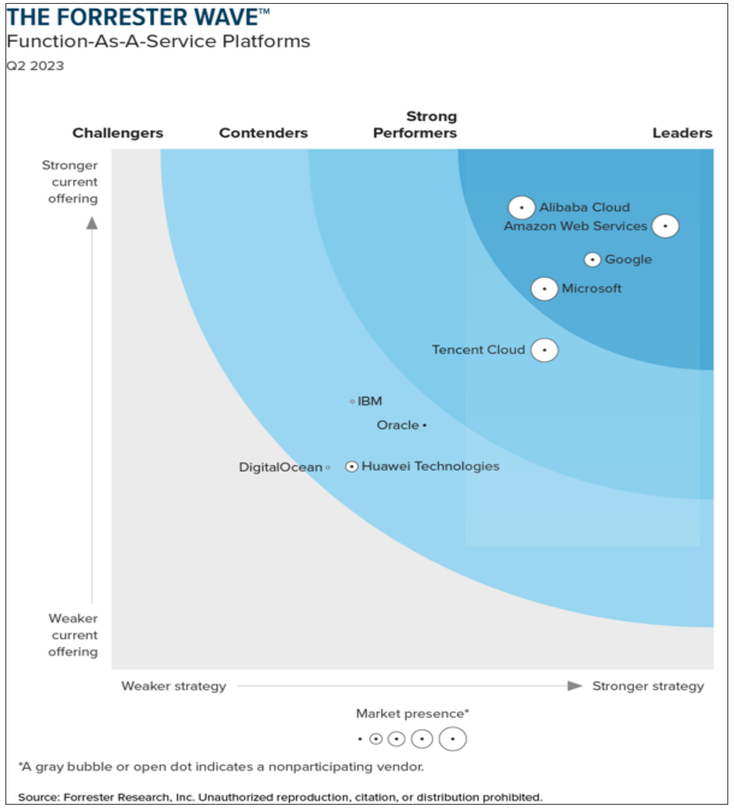


Forrester发布Serverless开发平台评估报告《Forrester Wave™: Serverless Development Platforms, Q2 2025》：在21项测评中，阿里云函数计算FC和Serverless应用引擎SAE产品能力拿到了9项最高分。阿里云成功进入领导者象限，是国内唯一进入该象限的科技公司。

阿里云获得「 Forrester FaaS」 领导者象限

Forrester： 阿里云 Serverless 连续获得（2021年，2023年） FaaS平台领导者象限， 全球领先、国内唯一。2023年，函数计算产品能力在 40 个评分项目中拿下 24 个最高分，蝉联全球领导者象限，中国唯一

阿里云蝉联 Forrester FaaS 领导者象限



阿里云

THANKS