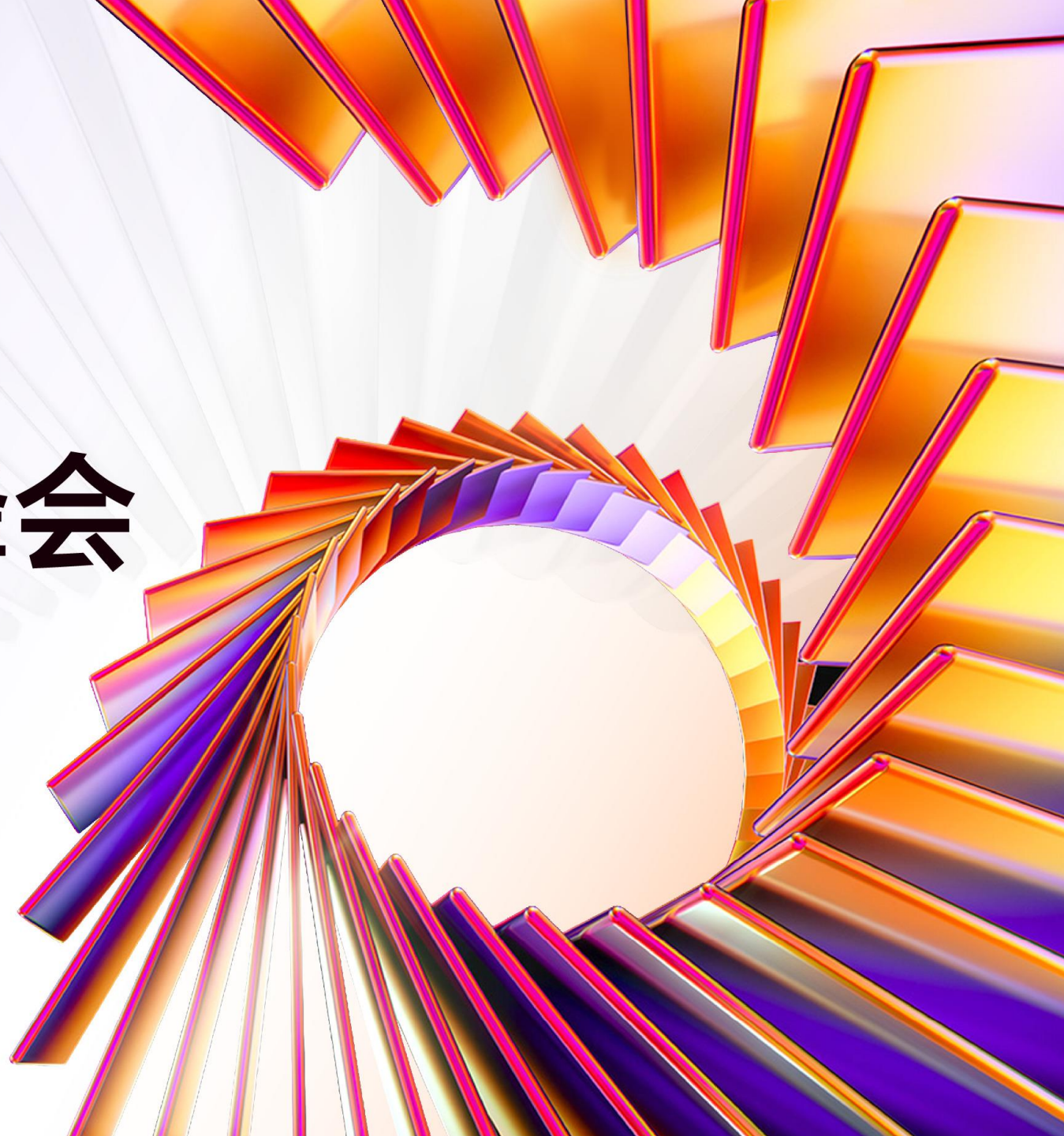
AiDD 6th 2025

2025 AI+ Development Digital Summit

AI+研发数字峰会

拥抱AI 重塑研发

05/23-24 | 上海站



AiDD 7th 2025

2025 AI+研发数字峰会

拥抱AI 重塑研发

AI+ Development Digital Summit

查看会议详情

下一站预告 08/08-09 | 北京站 11/14-15 | 深圳站

北京站论坛设置

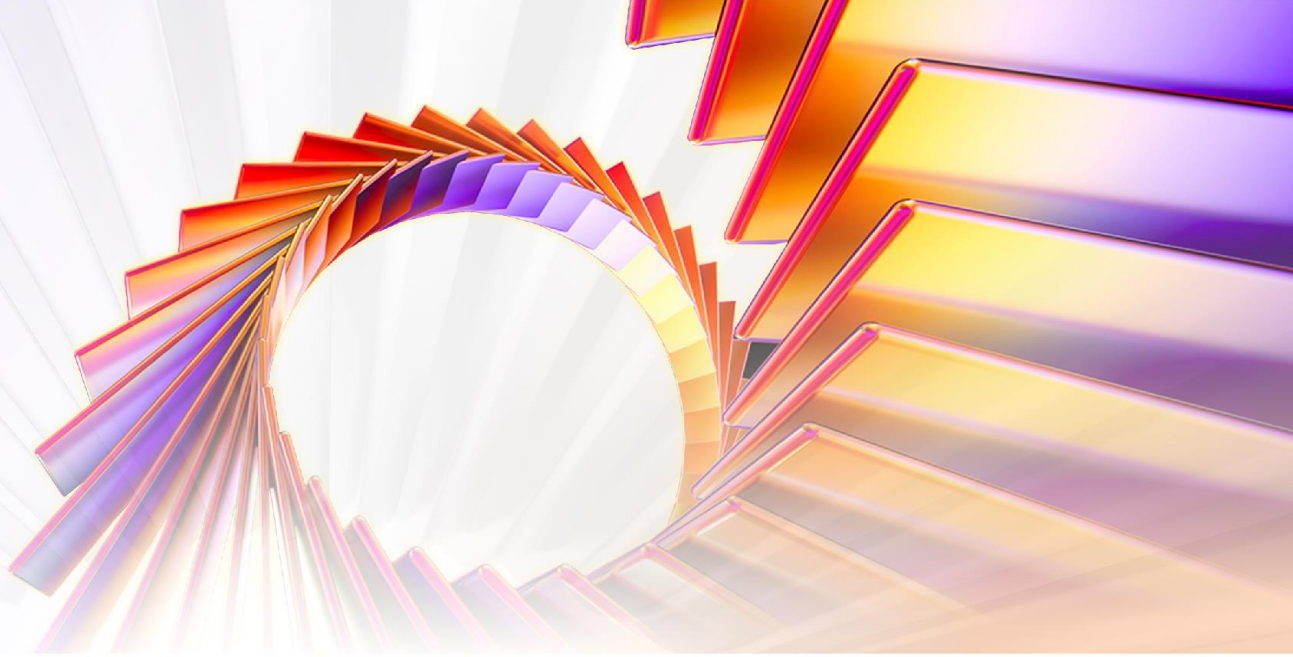
- AI+ 金融业务创新
- 大模型安全与对齐
- 具身智能与机器人
- 大模型和 AI 应用评测
- 智能需求工程
- 大模型应用开发框架与实践
- 代码生成及其改进
- 智能存储与检索技术
- 智能体与研发效率工具
- 智能体经济（Agentic Economy）
- AI+ 新能源汽车
- 下一代知识工程
- AI 产品运营与出海策略
- 智能测试工具的开发与应用
- AI 前沿技术探索与实践

AiDD 6th 2025 | 05/23-24 | 上海站

2025 AI+ Development Digital Summit

AI+研发数字峰会

拥抱AI 重塑研发



基于开源技术栈构建智能弹性大模型推理服务的架构实践

车漾 | 阿里云

车漾

阿里云 高级技术专家

阿里巴巴云原生应用平台高级技术专家，从事 Kubernetes 和容器相关产品的开发，重点探索利用容器技术加速异构计算、深度学习、边缘计算等广泛场景方案的交付与落地，同时是对于开源社区的积极参与者。他是CNCF旗下开源项目Fluid的创始人之一，也是核心维护者。也是业界第一个 GPU 共享调度的主要作者和维护者。他还是Alluxio开源项目的管理委员会成员（PMC Member），Kubernetes，Docker和Kubeflow等社区的积极贡献者。

目录
CONTENTS

Kubernetes 正成为数字化、智能化应用的云原生基础设施

Web/mobile applications
– Stateless
– Idempotent
– Horizontal scalable

Zoo-Keeper | Redis | MySQL | Spark | Flink | Elastic-search | Kafka | Tensor-Flow | PyTorch

Kubernetes（容器服务 ACK / ACS）

异构算力
ECS, GPU, FPGA, ECI

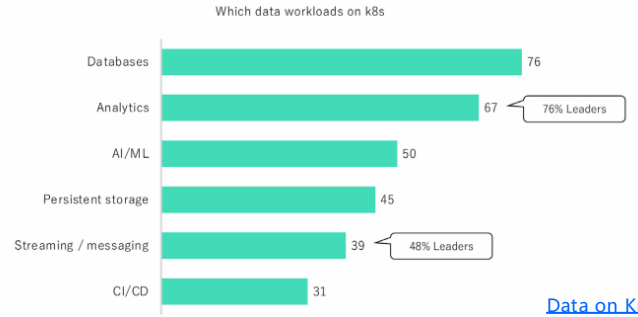
高性能网络
VPC, ENI, RDMA, SLB, DNS

高性能存储
EBS, NAS, CPFS, OSS

Public Cloud | Edge Computing | Private Cloud

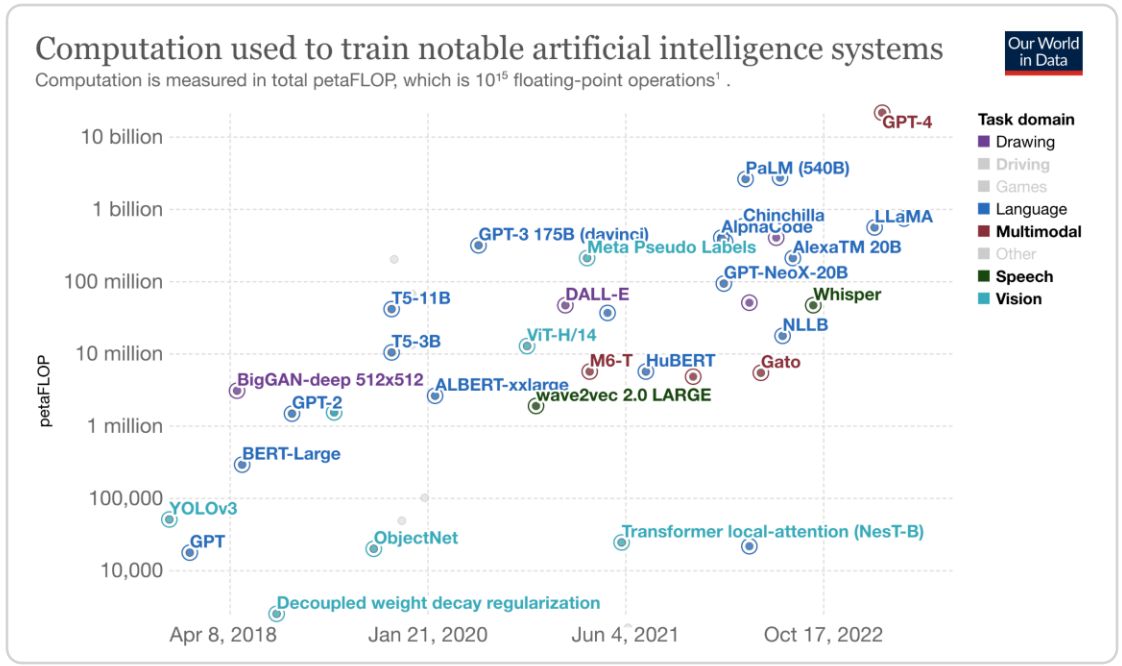
The Data on Kubernetes Community 2022调查报告，90%的受访者认为Kubernetes已经可以很好支持有状态应用，其中70%的受访者已经将其运行在生产环境。

IDC预测：到2025年，几乎50%的用于性能密集型计算（如AI、HPC和大数据分析）的加速基础设施将迁移至云端



大模型推理对基础设施服务带来新的挑战

- 大模型对基础设施服务能力的挑战是阶跃式的。
- 对“规模、性能、效率”的要求，成为LLM/AIGC快速落地的高门槛。



GPT3：175B 参数，单次训练使用45TB数据，近千卡 A100/1个月，成本数百万美元。

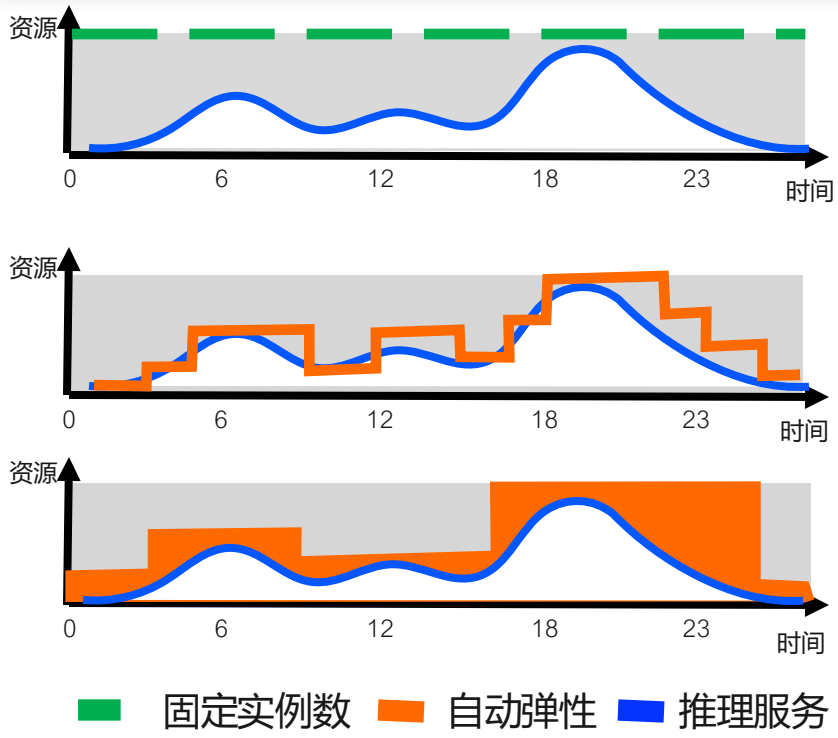
- 资源成本：如何充分利用有限计算资源
- 运维成本：降低复杂度

- 算力：千卡GPU任务，万卡集群
- 数据：PB级存储，TB级吞吐
- 网络：800Gbps ~ 3.2Tbps RDMA

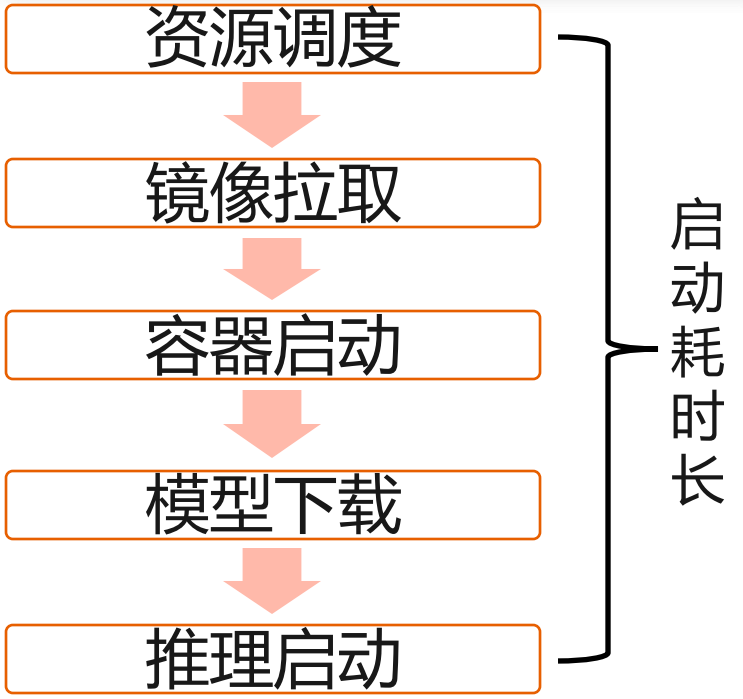
- 训练：分布式，混合并行
- 推理：模型优化、服务QoS

大模型推理对基础设施服务带来新的挑战

大模型弹性面临的问题



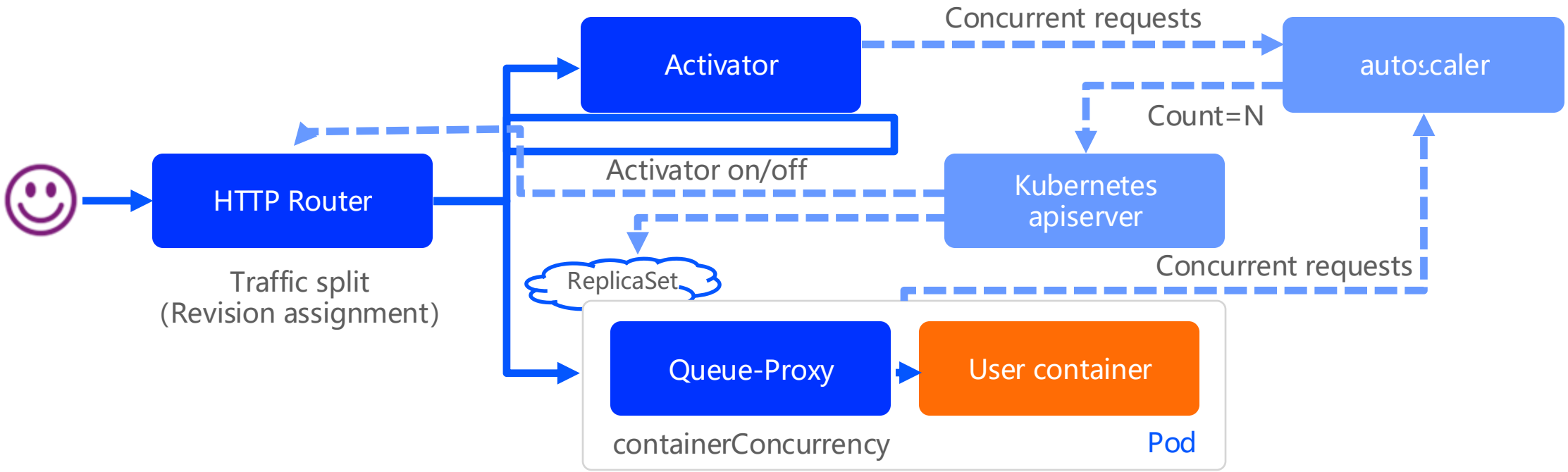
大模型启动冷启动问题



核心的挑战：在提升资源使用率的同时，保障大模型服务的稳定性和用户体验

大模型推理需要基于请求数的自动弹性策略

基于 GPU的弹性，并不能完全反映业务的真实使用情况，而基于并发数或者每秒处理请求（QPS/RPS）,对于推理服务来说更能直接反映服务性能，Knative Serving提供了基于请求的自动弹性能力

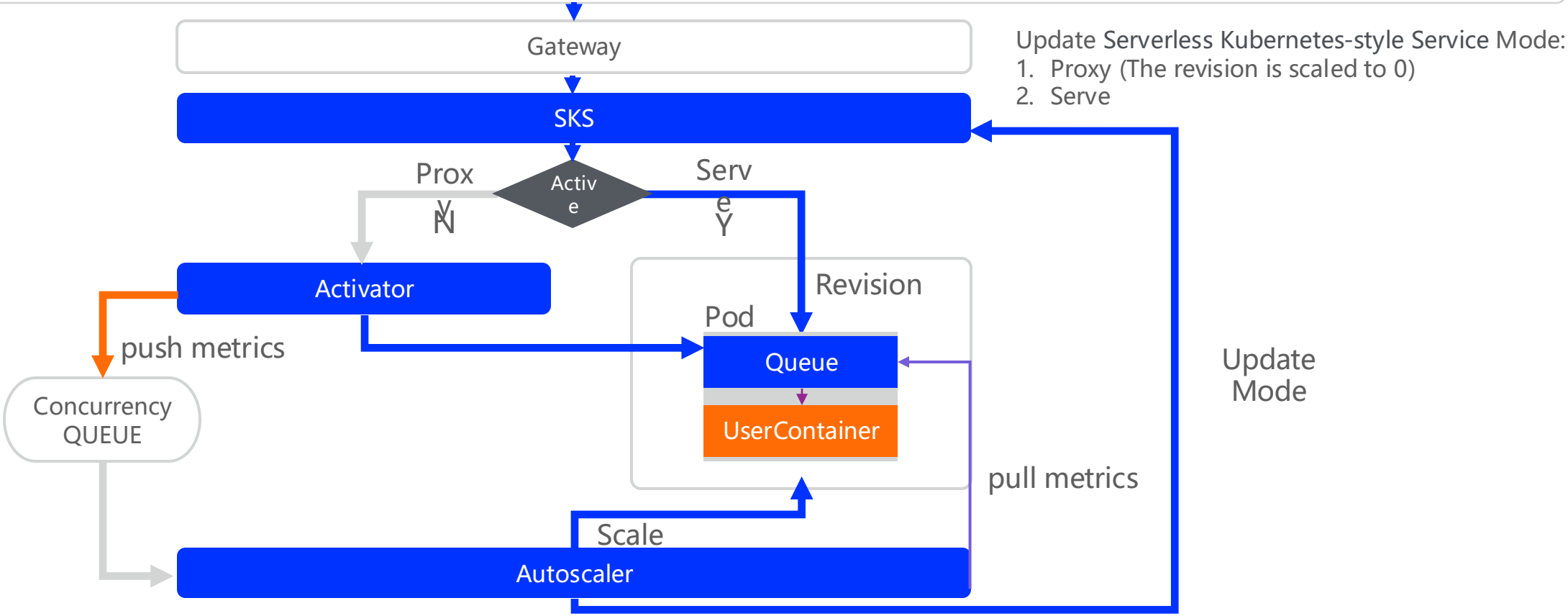


Pod数=并发请求总数/(Pod最大并发数*目标使用率)

特定离线推理场景需要缩容到0

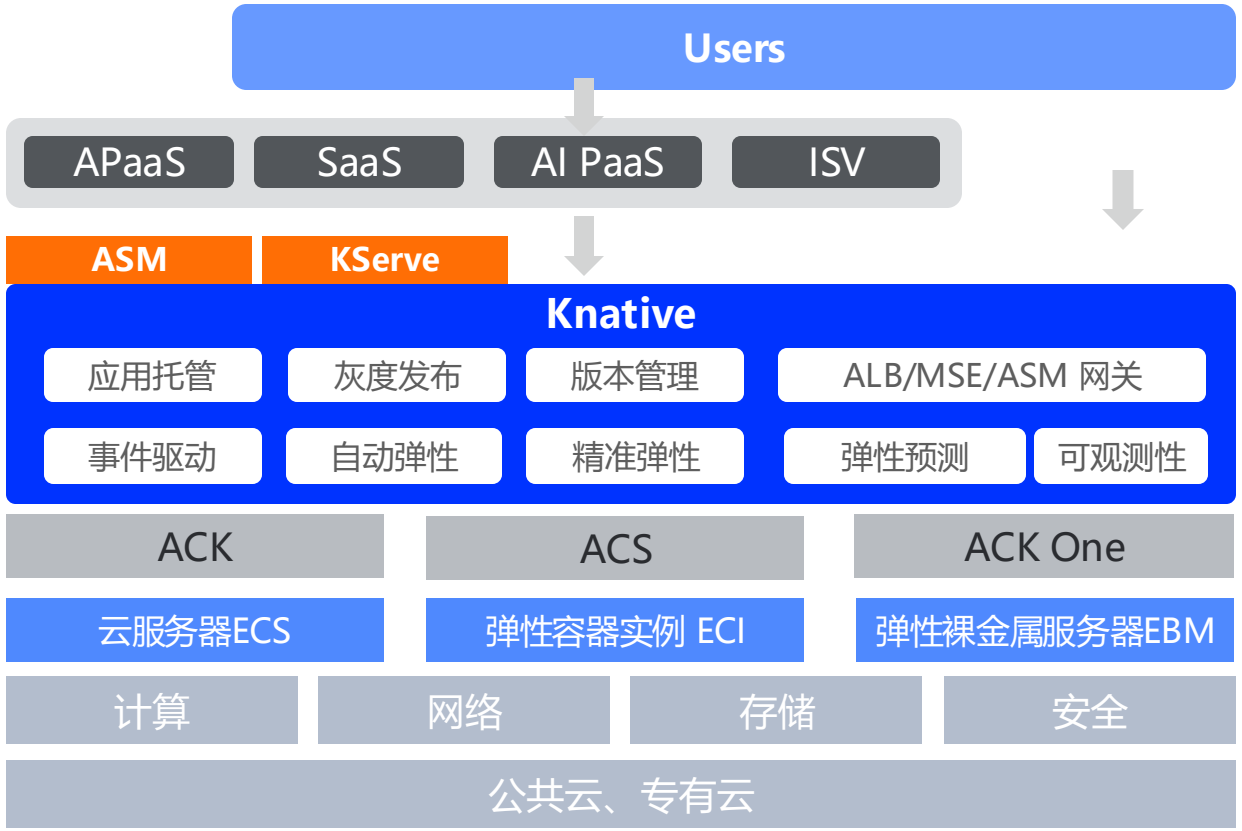
Knative Serving 中定义了 2 种请求访问模式：Proxy和Serve。

- Proxy 顾名思义，代理模式，也就是请求会通过 activator 组件进行代理转发
- Serve模式是请求直达模式，从网关直接请求到Pod， 不经过 activator 代理



阿里云容器服务 Knative 支持更多计算资源

- 构建容器 Serverless Framework 解决方案，基于K8s开放标准，被用户平台集成, 同时可插拔
- 丰富阿里云 Serverless 产品家族，同时打造更面向应用的 Serverless 容器平台



差异性

标准化

- 兼容社区Knative
- 提供k8s 标准 API
- 无需担心厂商绑定

产品化

- 提供UI 控制台
- 智能弹性AHPA
- 基于ACK/ACS产品底座

高集成

- EventBridge
- 云效
- SLS
- 可观测监控

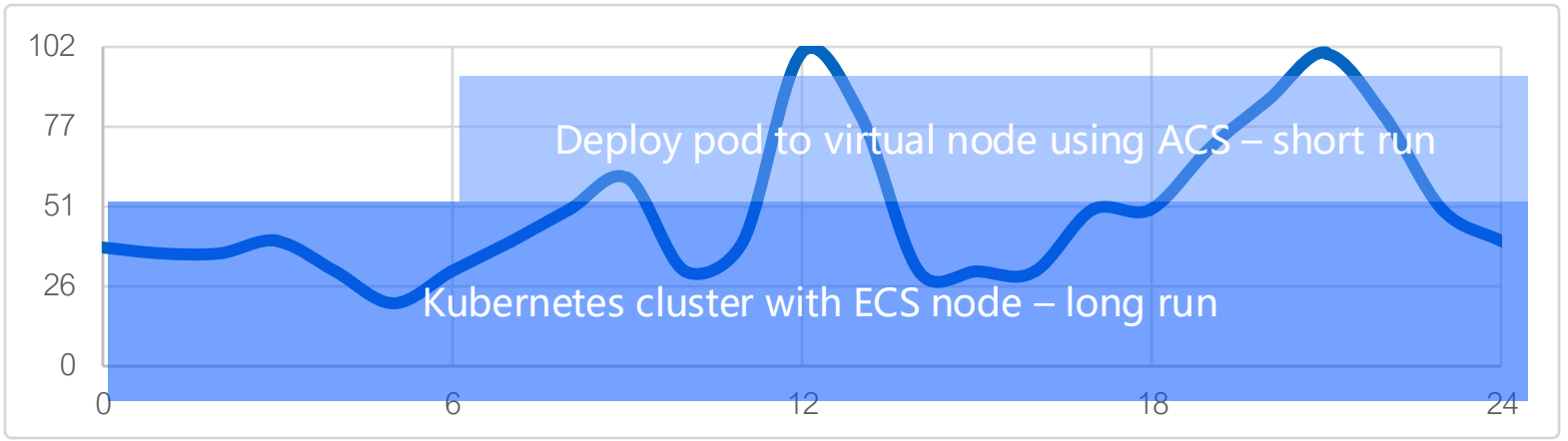
高可用

- 组件全托管、高可用
- 云产品网关：ALB、ASM、MSE

通过保留资源池降低资源使用成本

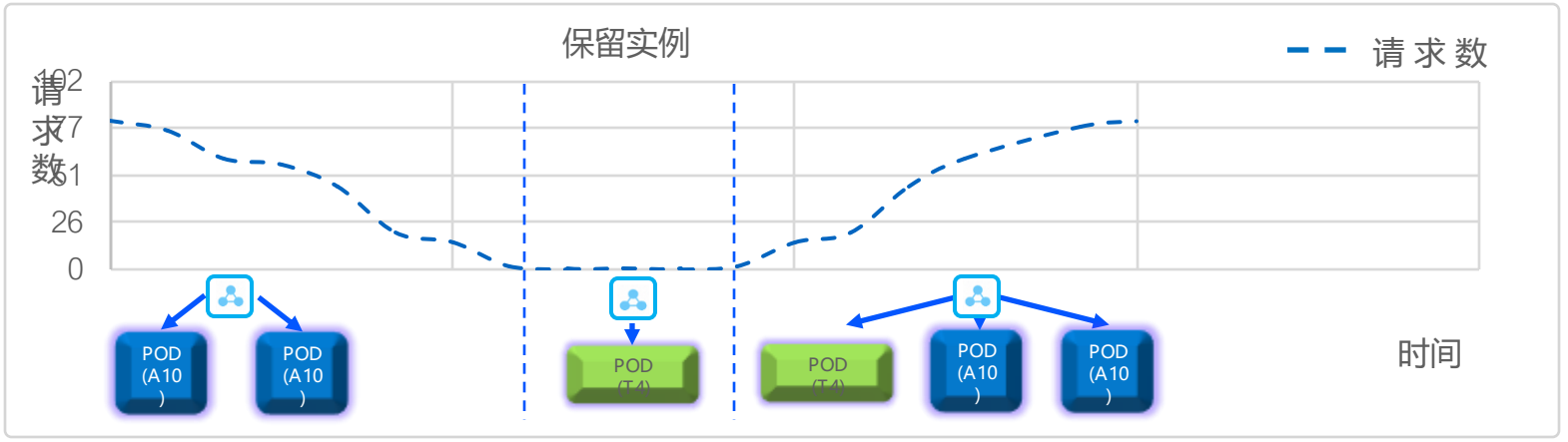
ECS 和 ACS 混合使用

常态情况下使用ECS 资源，突发流量使用ACS 资源

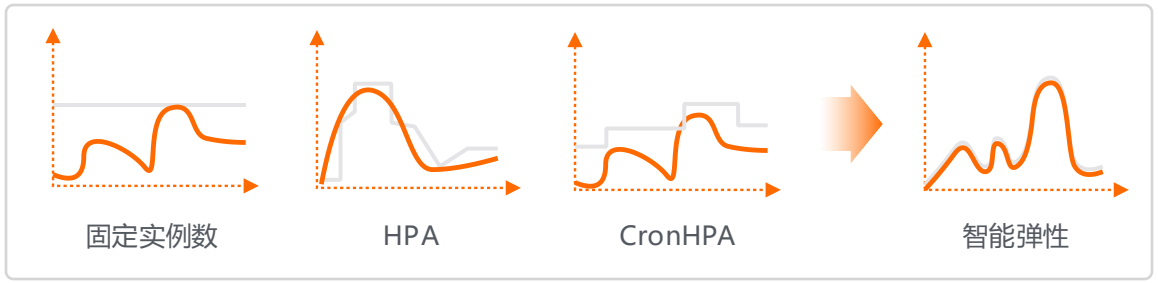


资源预热

完全使用ACS的场景，也可以通过保留资源池实现资源预热



AHPA优化大模型的智能弹性



智能弹性

- 资源提前预热，实时调整容量
- 无需人工干预，自动弹性规划
- 弹性降级保护，快速兜底容错

资源提前预热

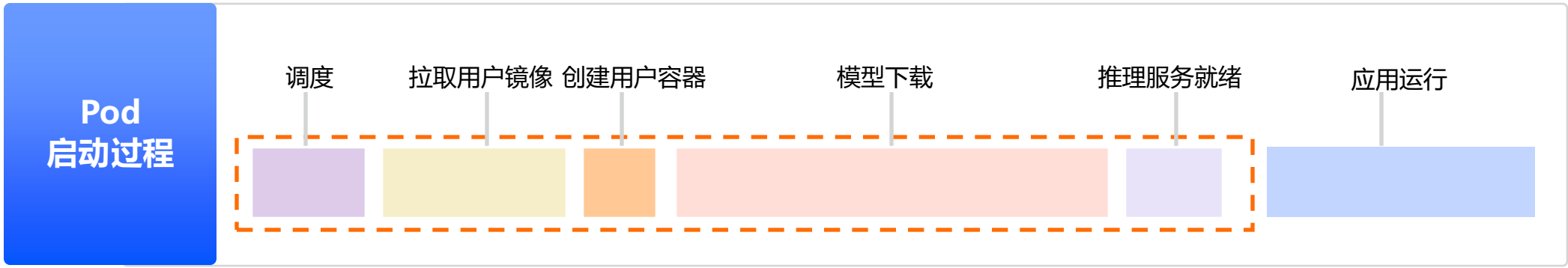
解决客户弹性滞后冷启动的问题，通过弹性预测，提前预热资源。

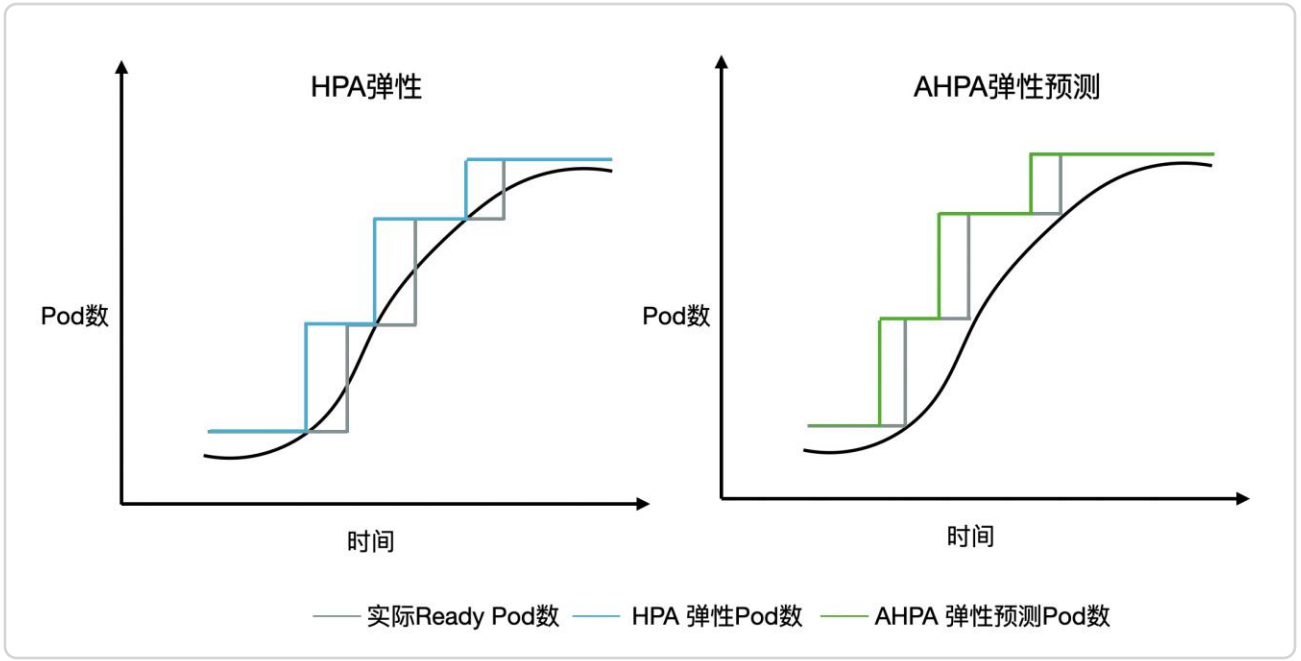
自动弹性规划

根据业务趋势，自动进行弹性策略规划，避免人工规划导致预估不准（过高导致资源浪费，过低导致业务不稳定）

Metrics 指标收集、Pod生命周期 → 周期检测、资源需求预测 → 配置保护、安全降级 → 实时生效

AHPA优化大模型的智能弹性

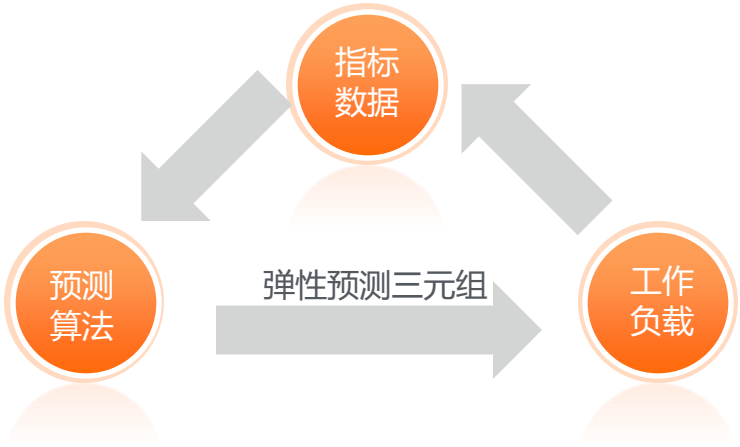




提前扩容的因素

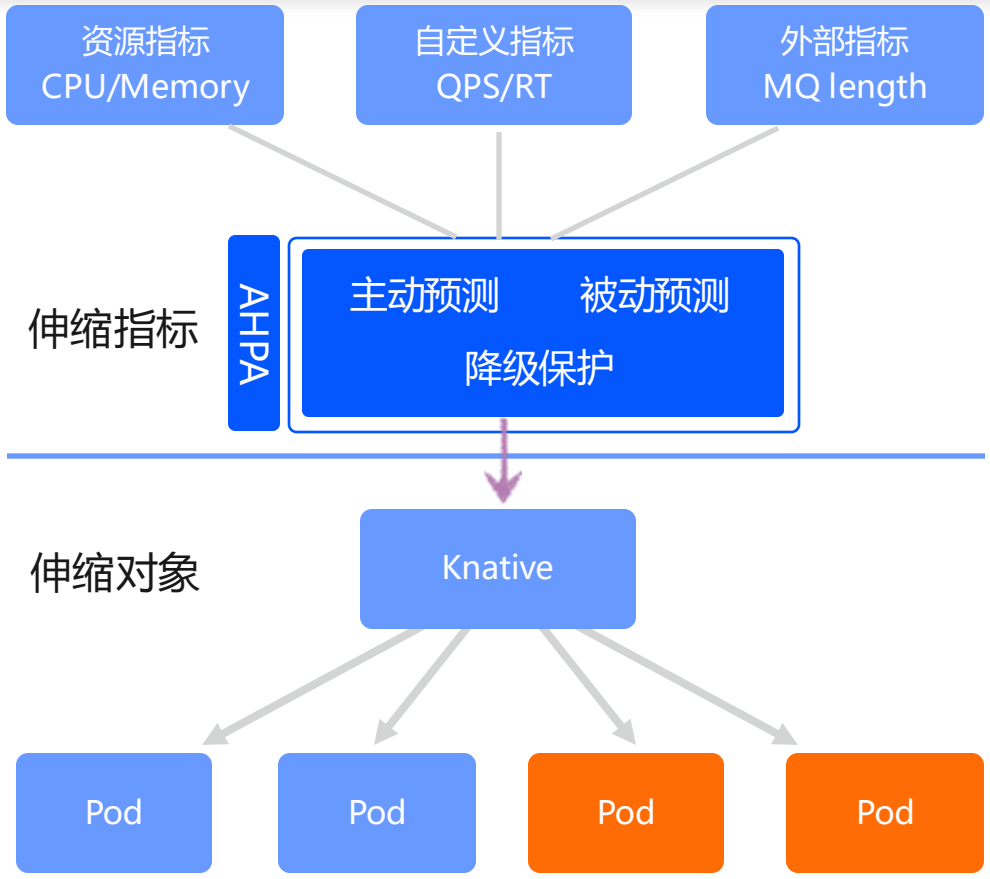
目标GPU使用率/RT/QPS等

根据POD生命周期计算 POD 冷启动时间



AHPA优化大模型的智能弹性

智能弹性：主动预测+被动预测



适合模型推理场景

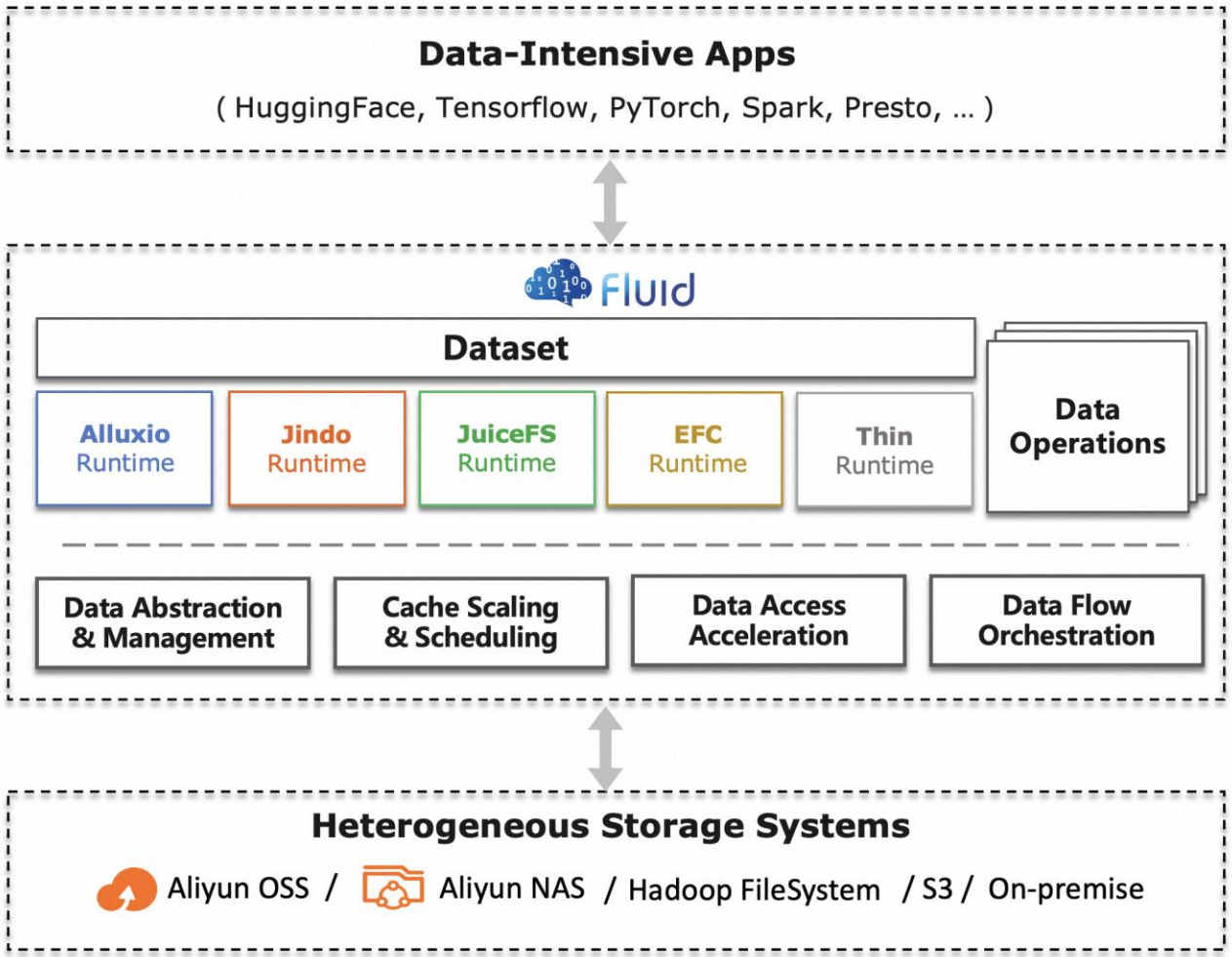
1. 有明显周期性。
2. 固定实例数+弹性兜底。
3. 需要推荐实例数配置

更快 更准 更稳

```
apiVersion: v1
kind: ConfigMap
metadata:
  name: application-intelligence
  namespace: kube-system
data:
  prometheusUrl: "https://cn-
shanghai.arms.aliyuncs.com:9443/api/v1/prometheus/da9d7dece9f1db4c9fc7
f5b9c40e93e/1581204543170042/417d182c6d430fb062ec364e6dfb49/cn-
shanghai"
```

指标源配置

Fluid：弹性数据集编排和加速



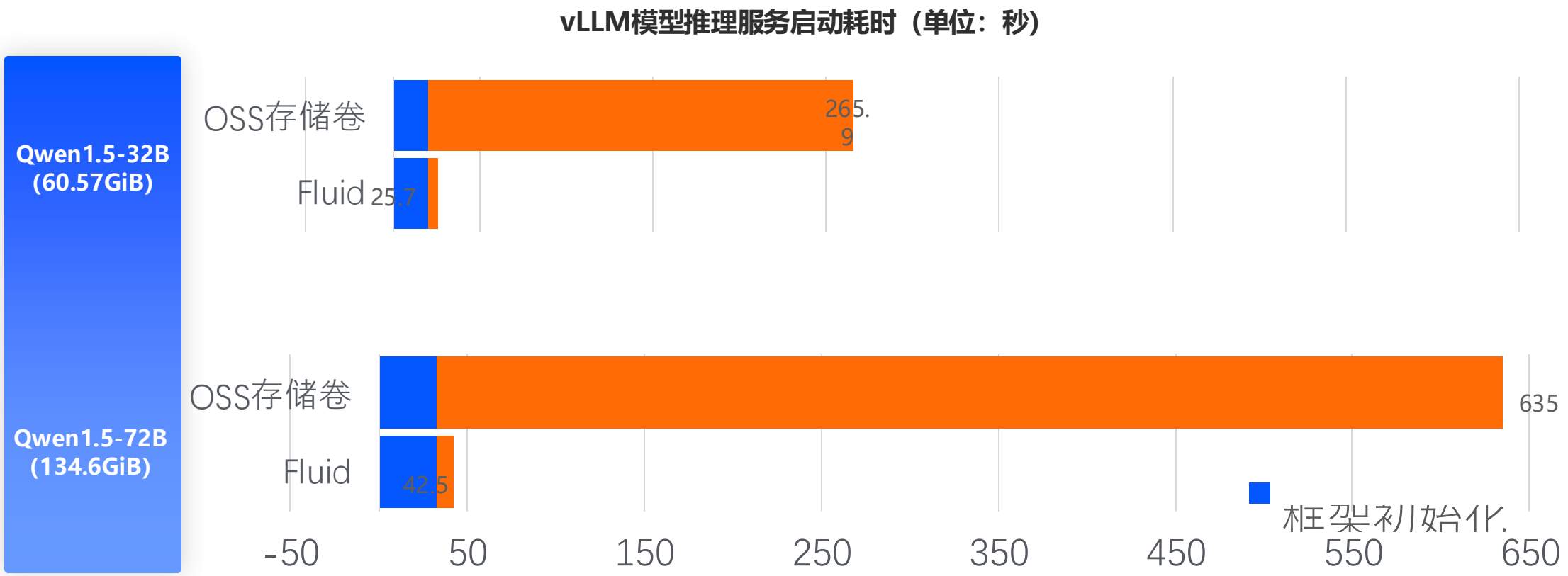
CLOUD NATIVE SANDBOX

核心功能：

- **标准化**：用于数据访问和分布式缓存管理的API。
- **自动化**：针对数据的操作，如数据处理、预取、迁移和缓存扩缩容的操作。
- **加速**：通过弹性分布式缓存和亲和度调度提高数据访问效率。
- **屏蔽差异**：支持不同 Kubernetes 环境的CSI和边车模式。
- **数据流**：自动编排数据操作和任务调度。

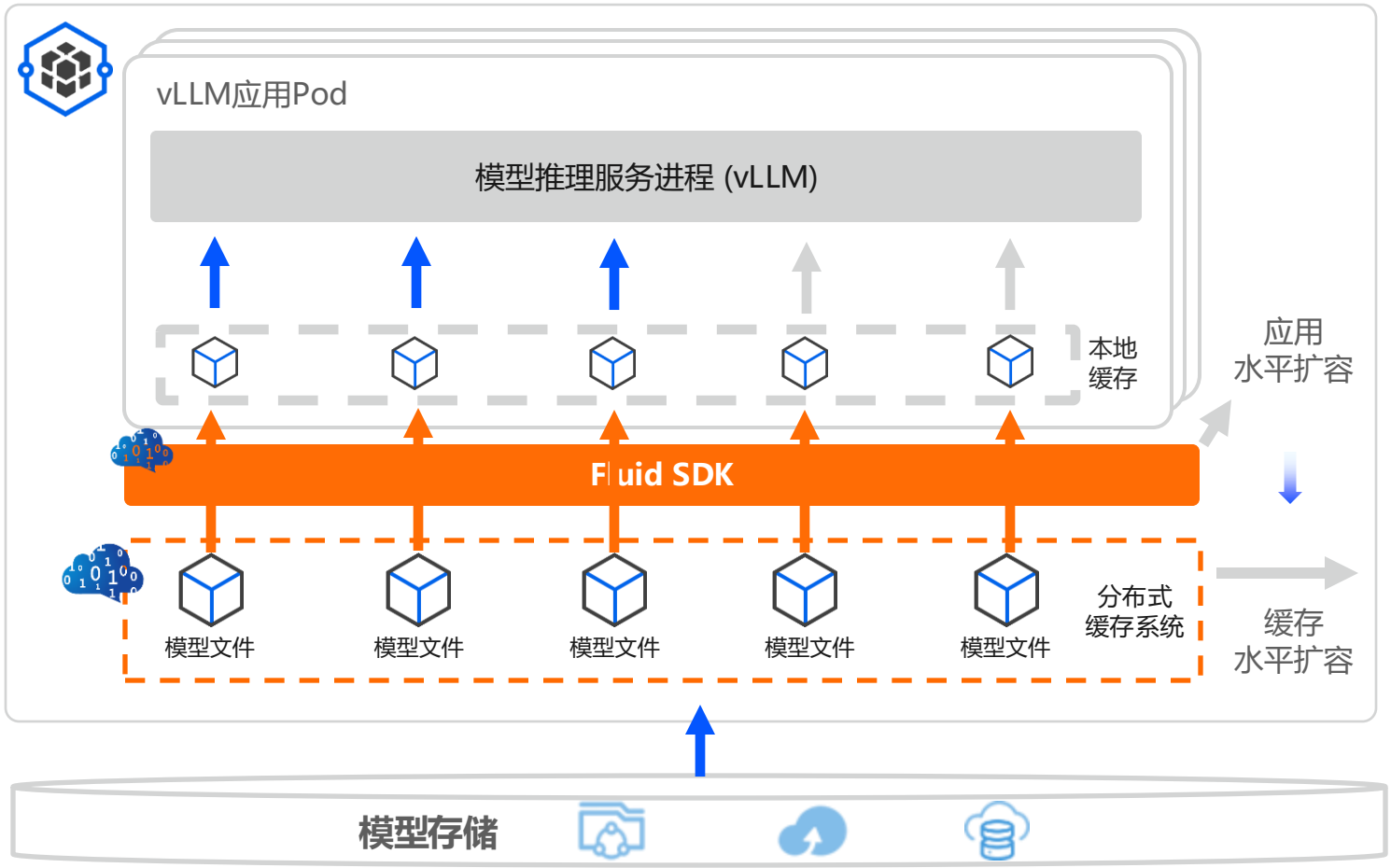
Fluid对于推理服务启动速度的优化效果

Fluid使vLLM + Qwen系列开源LLM模型推理服务启动耗时缩短10.3倍(32B)、14.9倍(72B)



[测试机型：1 x ecs.ebmgn7ix.32xlarge + 3 x ecs.g7ne.12xlarge（Fluid缓存）；OSS带宽上限：100 Gbps；vLLM框架版本：v0.4.3；测试时间：2024.05]

优化模型加载的I/O过程



AI应用侧模型文件预读

1 Fluid SDK在AI应用Pod中多线程并发预读对LLM模型文件，**最大化利用单个AI应用Pod的可用带宽，加速模型加载过程**

弹性伸缩的分布式缓存

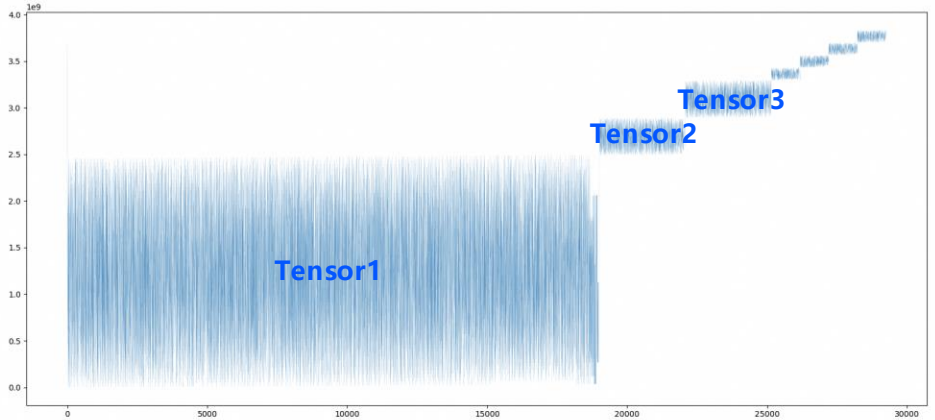
2 计算侧分布式缓存系统可根据每个AI应用Pod的带宽需求灵活扩缩容，**为多个AI应用Pod分配更大的可用带宽，支撑模型加载过程的高带宽需求**

优化一：AI应用侧模型文件预读

现存问题

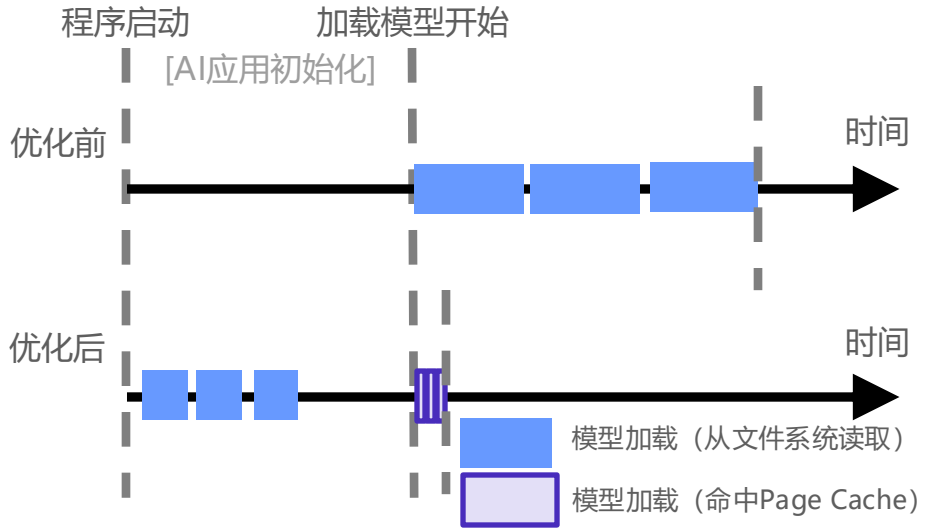
- LLM模型参数大多以Safetensors文件格式分发
- Safetensors文件的加载过程涉及大量跳变的随机读，受存算分离架构下文件系统实现影响，I/O效率低下。
- （机器节点带宽利用率不到20%）

模型参数文件(.safetensors) 读取过程read offset变化



Fluid解决方案

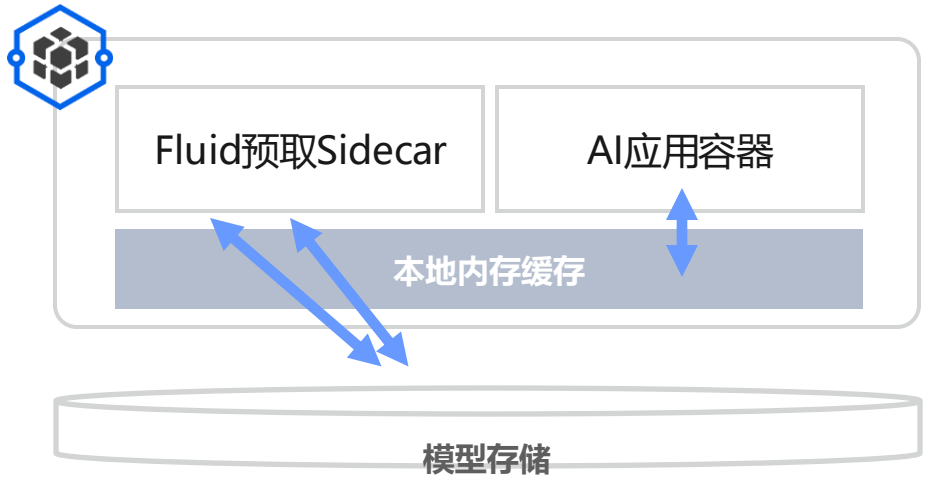
- 将模型参数文件提前预读到随机读友好的本地内存缓存（Page Cache）
- 多线程并行顺序预读，节点带宽利用率提升至80%以上



优化一：AI应用侧模型文件预读

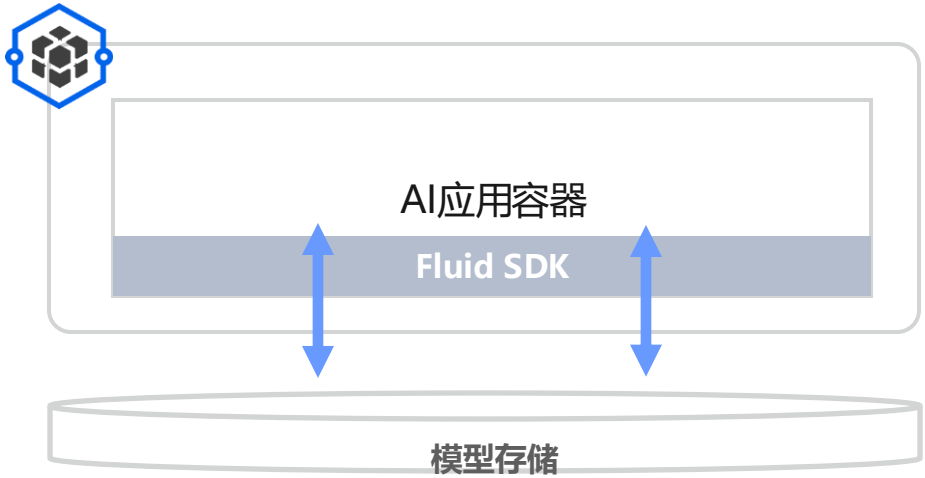
使用方法一：Sidecar预取

- 由Fluid自动对AI应用Pod注入Sidecar容器，由Sidecar发起多线程预取
- 优势：对应用无侵入，适用简单场景



使用方法二：应用容器调用Fluid SDK

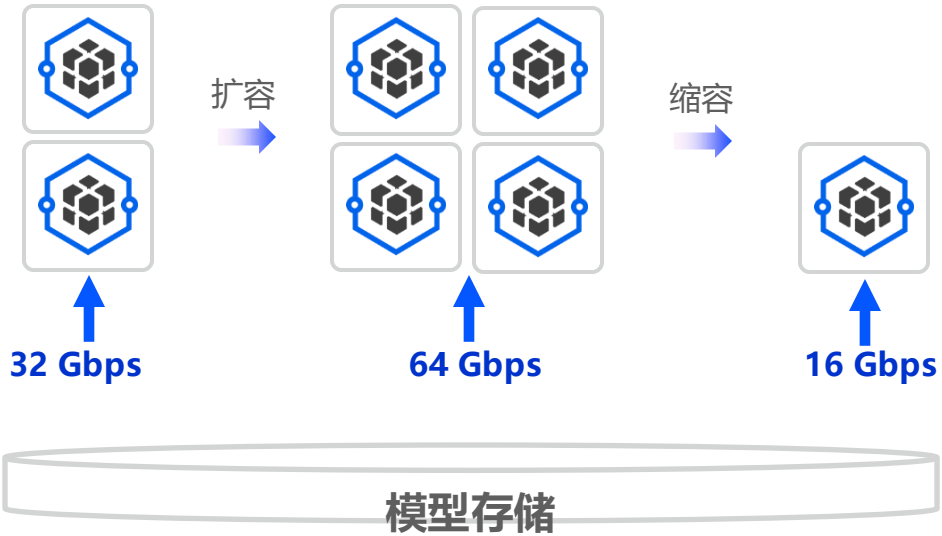
- 改造AI应用，根据业务需求适时调用Fluid SDK发起多线程预取
- 优势：支持运行过程中模型汰换等灵活场景



优化二：弹性伸缩的分布式缓存

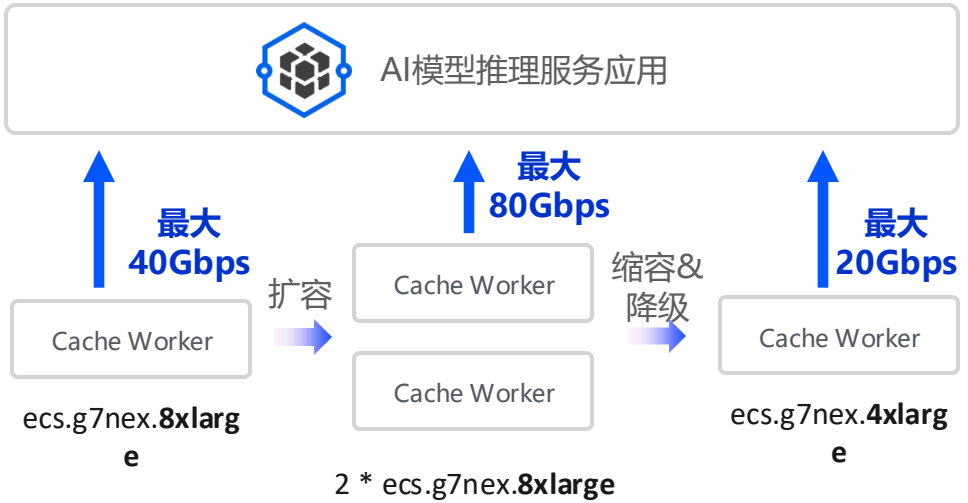
现存问题

- Knative应用的灵活弹性 -> 存储侧供给带宽的灵活弹性
- 存储系统聚焦于数据持久化可靠性、稳定性，往往无法提供弹性带宽的灵活选择。



Fluid解决方案

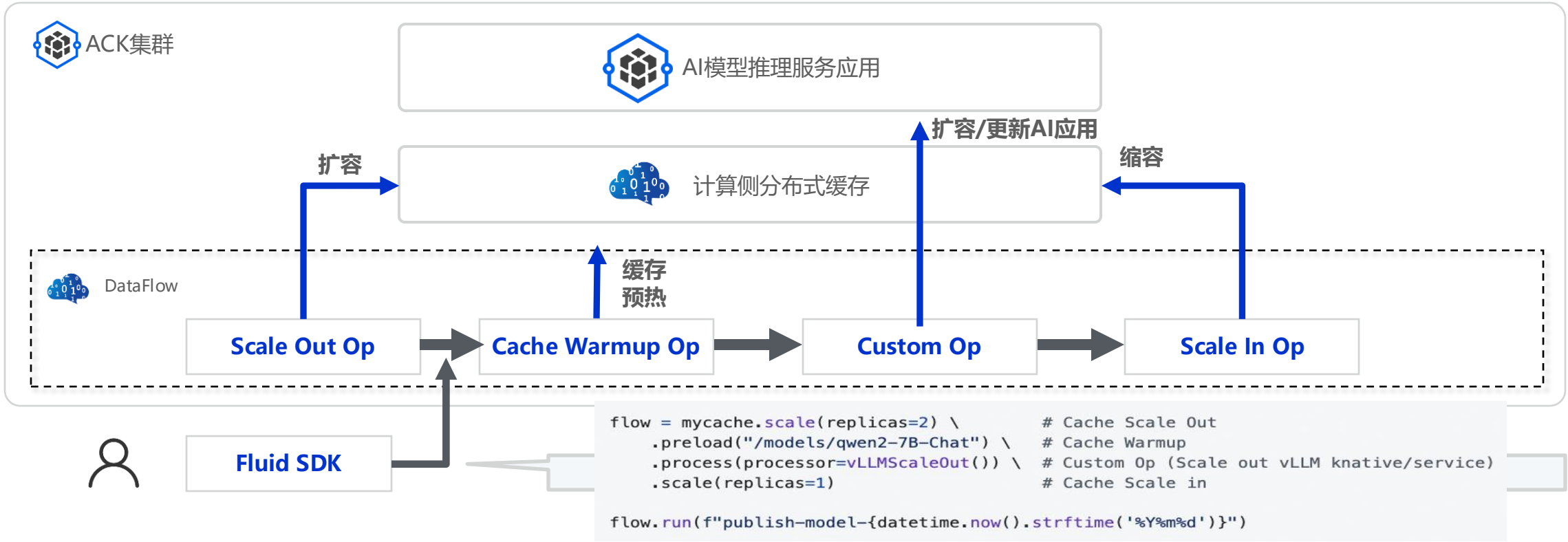
- 在Kubernetes集群内构建可弹性扩缩容的计算侧分布式缓存
- 支持主动扩缩容：根据业务场景主动执行扩缩容操作
- 支持自动扩缩容：HPA、CronHPA、AHPA、...



Fluid DataFlow：串联AI应用更新发布流程

缓存运维带来的复杂度

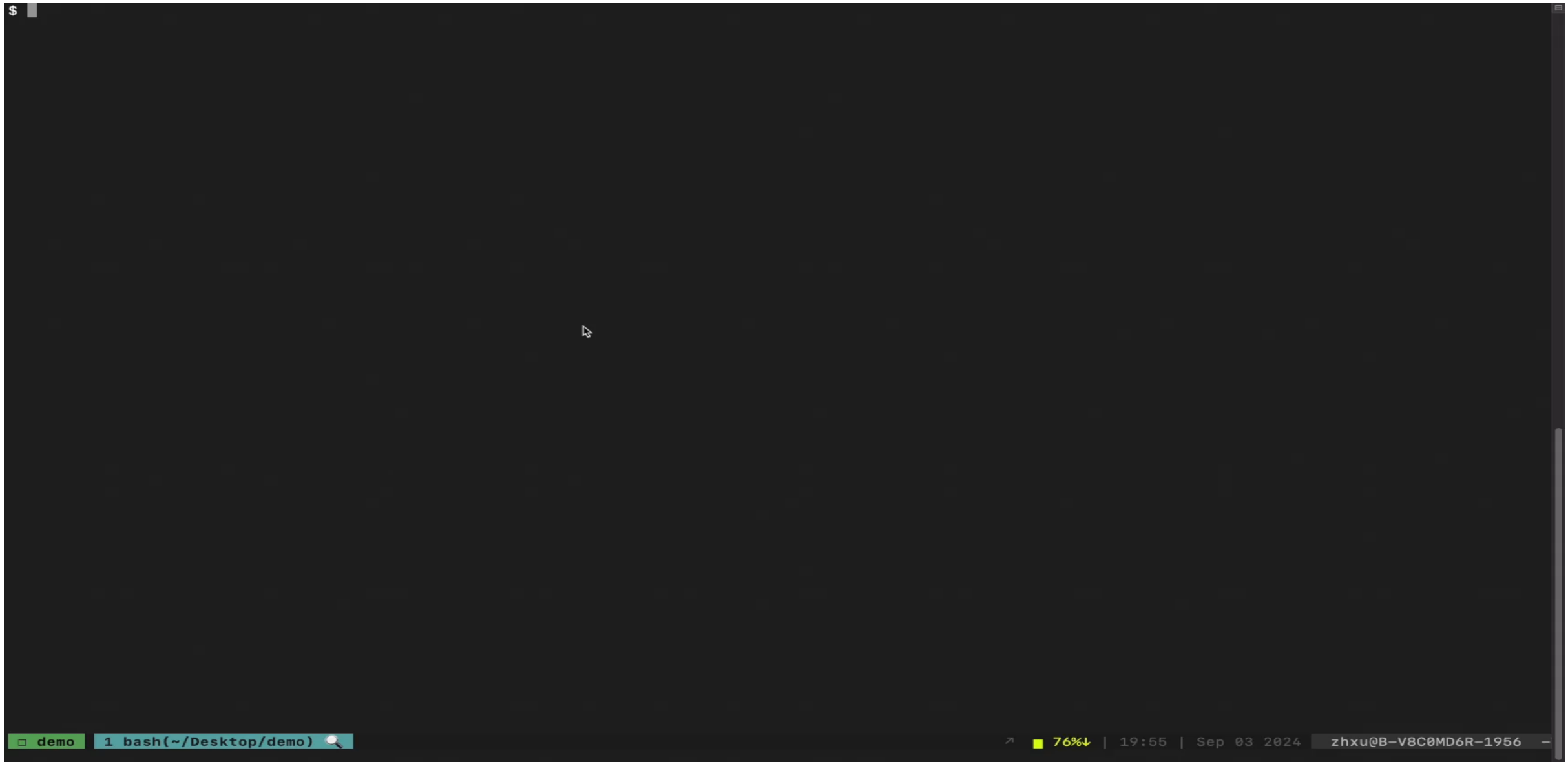
- 一次AI应用扩容/更新的流程：缓存扩容 -> 缓存数据预热 -> 扩容/更新AI应用 -> 等待AI应用就绪 -> 缓存缩容
- **如何简化、自动化AI应用的扩容和更新发布流程?**




```
flow = mycache.scale(replicas=2) \          # Cache Scale Out
    .preload("/models/qwen2-7B-Chat") \     # Cache Warmup
    .process(processor=vLLMScaleOut()) \    # Custom Op (Scale out vLLM knative/service)
    .scale(replicas=1)                      # Cache Scale in

flow.run(f"publish-model-{datetime.now().strftime('%Y%m%d')}")
```

Demo演示：
Fluid加速vLLM + Qwen模型的Knative服务扩容


$
demo
1 bash(~/Desktop/demo)
76%
19:55
Sep 03 2024
zhxu@B-V8C0MD6R-1956

Demo演示：Fluid加速vLLM + Qwen模型的Knative服务扩容

优化效果

相比直接读取OSS，Fluid方案能够使：

- 模型加载耗时下降为原来的约 **1/16**
- 端到端首请求延时下降 **67%**

方案	模型加载耗时	端到端首请求延时
OSS存储卷	33s	48.25s
Fluid	**2s**	**15.79s**

[测试机型：1 x ecs.ebmgn7i.32xlarge + 2 x ecs.g7nex.8xlarge（Fluid缓存）；OSS带宽上限：100 Gbps；测试时间：2024.09]

AiDD 6th 2025

参与调研您将优先获得

AiDD定制版
《AI+软件研发精选案例》

专属学习顾问
1对1需求对接

AiDD会后小调研

AiDD峰会致力于协助企业利用AI技术深化计算机对现实世界的理解，推动研发进入智能化和数字化的新时代。作为峰会的重要共建者，您的真知灼见对我们至关重要。衷心感谢您的参与支持！

扫码参与调研

2025 AI+研发数字峰会

拥抱AI 重塑研发

科技生态圈峰会 + 深度研习

——1000 + 技术团队的共同选择

K+峰会

K+峰会 敦煌站
K+ 思考周®研习社
时间：2025.08.29-30

K+峰会 上海站
K+ 金融专场
时间：2025.09.26-27

K+峰会 香港站
K+ 思考周®研习社
时间：2025.11.17-18

K+峰会详情

AiDD峰会

AiDD峰会 上海站
AI+研发数字峰会
时间：2025.05.23-24

AiDD峰会 北京站
AI+研发数字峰会
时间：2025.08.08-09

AiDD峰会 深圳站
AI+研发数字峰会
时间：2025.11.14-15

AiDD峰会详情

AiDD 6th 2025

2025 AI+研发数字峰会

AI+ Development Digital Summit

感谢聆听！

扫码领取会议PPT资料