亚马逊云科技

中国峰会

2026年6月23日-24日 上海 · 世博中心

亚马逊云科技
中国峰会

亚马逊云科技
中国峰会

SESSION 304

Token 经济时代，算力的新战场

大规模 AI 推理基础设施的工程实践

唐安波

硅基流动

解决方案总监

汪其香

亚马逊云科技解决方案架构师

推理的供需挑战

01 / INFERENCE USERS

8M → 800M

两年内 100× 增长,推理用户从千万级冲向十亿级

Source: NVIDIA Jensen Huang · GTC 2025

02 / COMPUTE DEMAND

100× of last-gen

推理模型的算力需求,是上一代生成式 AI 模型的 100 倍

Source: NVIDIA · GTC 2025 Keynote

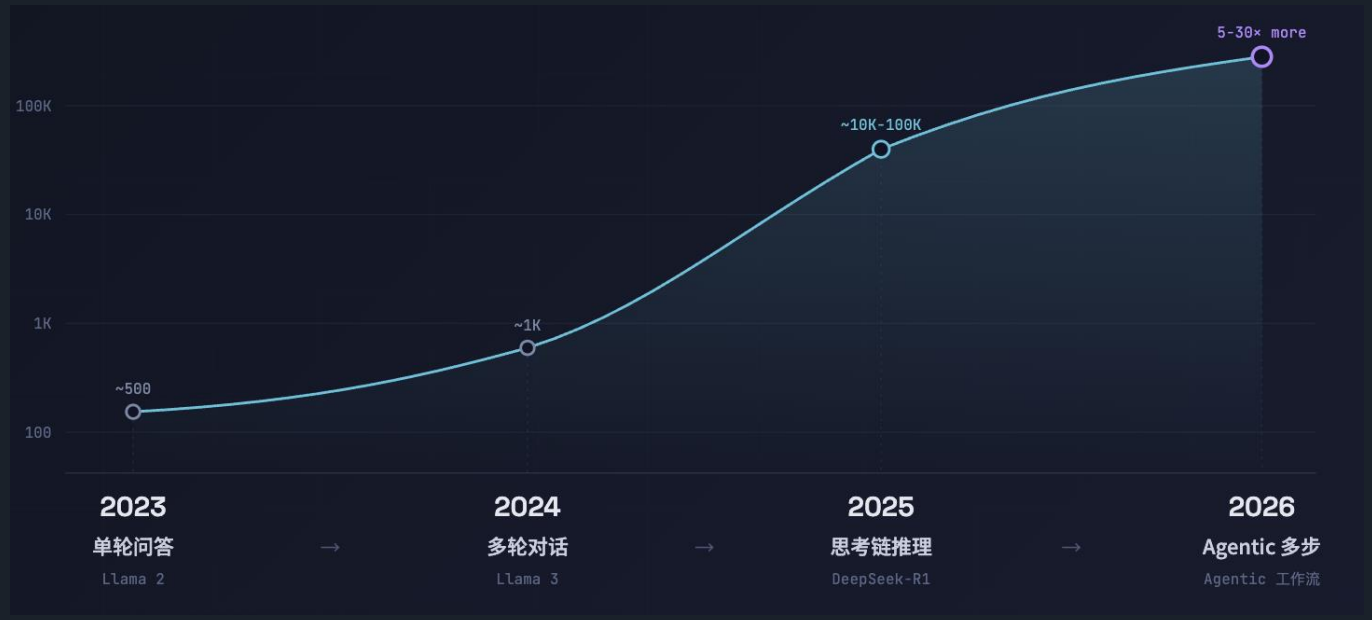
*前述特定亚马逊云科技生成式人工智能相关的服务目前在亚马逊云科技海外区域可用。亚马逊云科技中国区域相关云服务由西云数据和光环新网运营，具体信息以中国区域官网为准。

推理需求，正在快速增长

THE SURGE OF INFERENCE DEMAND

PER-QUERY TOKEN CONSUMPTION (log scale)

单次回答的 Token 消耗 · 指数级跃升



每一步 AI 交互形态的演化都会带来 Token 消耗的指数级上升

01 / SCALE

模型上线即持续产出 Token，
消耗随用户与频率攀升

每一次问答、每一次 Agent 调用,都是新增的算力账单

02 / TEST-TIME COMPUTE

思考链推理模型，
单次推理 ~100× 算力

Source: NVIDIA GTC 2025

03 / AGENTIC

Agent 单任务消耗
是普通 Chatbot 5-30×

Source: Gartner 2026 · 多步 fan-out 调用

MARKET AI 推理市场**$106B (2025)** → **$255B (2030)** CAGR 19.2%

*前述特定亚马逊云科技生成式人工智能相关的服务目前在亚马逊云科技海外区域可用。亚马逊云科技中国区域相关云服务由西云数据和光环新网运营，具体信息以中国区域官网为准。

超大模型对推理基础设施提出新要求

THE FOUR WALLS OF INFERENCE AT SCALE

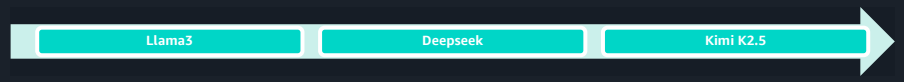
WALL 01 / LATENCY

模型规模持续增长

模型参数规模从 70B，400B 到 1T，持续演进

集群推理优化提高 token 性价比

超大集群的成为推理基础设施的核心竞争力



WALL 02 / UNIT ECONOMICS

单位经济模型恶化

Token 越来越便宜,企业 AI 账单却越来越贵 —— 用量爆炸性增长
长文本，高 Cache 命中率成为推理负载新模式

↓280× ↑320%

两年 Token 单价 / 同期企业 AI 支出 Source: Epoch AI · Introl 2025

WALL 03 / UTILIZATION

GPU 利用率不足与资源浪费

显存 / 算力单元 / 带宽，三者难以同时拉满

GPU 空闲率高达 ~70%

KV-Cache 碎片、Prefill/Decode 错配、批处理粒度不匹配 Source: Hedgehog 2025

WALL 04 / ENTERPRISE SLA

企业级稳定性不足

多租户隔离 · SLA 保障 · 灰度发布 · 可观测性

可用性 99.9%+硬要求

传统在线服务的标配，与推理引擎尽力而为模式存在天然冲突 Source: Flexential 2025

于是，我们需要一座 **Token工厂** —— 把推理变成可工业化、可规模化、可经济化的生产线

亚马逊云科技
中国峰会

硅基流动
SiliconFlow

加速 AGI 普惠人类

硅基流动：打造标准化、超高效能的生成式 AI Infra 产品

- 将生成式 AI Infra 产品化
  降低 AI 应用开发和使用门槛
- 极致性能优化
  降低规模化部署成本
- 攻坚 AI Infra 技术难点
  形成标准化产品

*前述特定亚马逊云科技生成式人工智能相关的服务目前在亚马逊云科技海外区域可用。亚马逊云科技中国区域相关云服务由西云数据和光环新网运营，具体信息以中国区域官网为准。

Token：AGI 时代的核心基础资源

自然语言处理的最小单元 — Token

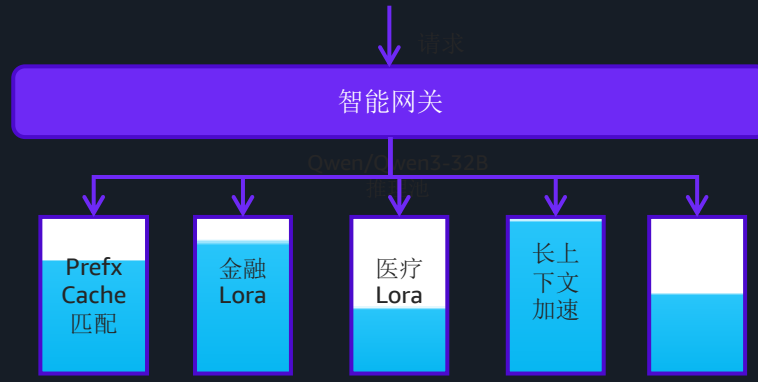
Token 是大模型处理信息的最小计量单位，是 AI 模型理解与生成语言的基 础粒子。当前我国日均词元调用量已超过 140 万亿，相比 2024 年初的 1000 亿增长了 1000 多倍，相比 2025 年底的 100 万 亿，三个月时间又增长了 40% 多。

AI 智能时代的“算力货币”

Token 具有可计量、可定价、可交易的核心属性，正在成为连接技术供给 与商业需求的核心“结算单位”，是衡量 AI 模型活跃度与产业价值的关键指 标，如同数字经济的新型“水电资源”。

大模型网关智能路由分发+大模型框架层推理加速+算力层动态伸缩 = 大幅提升集群吞吐

大模型网关智能路由分发

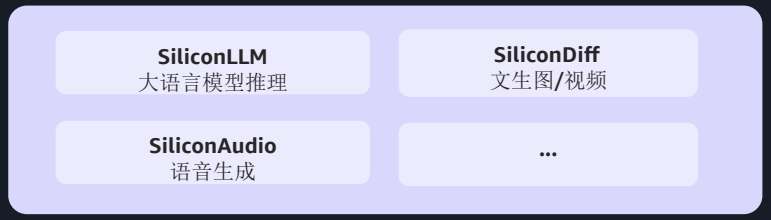


- 上下文长度感知
- Prefix cache 感知
- Lora 感知
- 负载感知
- ……

大模型推理服务平台

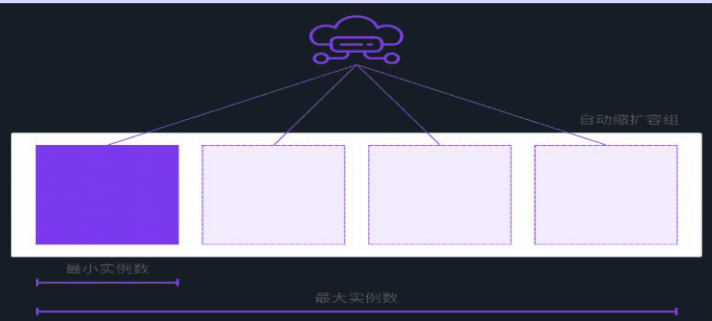
- 模型快速适配
- 内置大模型最佳实践
- 大模型高效微调
- 一键部署快速启动
- ……

高性能推理框架



- 适配多种算力芯片
- 算子优化、通信优化、调度优化
- ……

算力智能调度



- 弹性扩缩容
- 基于请求队列、性能等多维度指标

*前述特定亚马逊云科技生成式人工智能相关的服务目前在亚马逊云科技海外区域可用。亚马逊云科技中国区域相关云服务由西云数据和光环新网运营，具体信息以中国区域官网为准。

全链路大模型推理优化，一键部署最佳实践，体系化推理加速，端到端运维观测

一键部署

模型权重＋框架镜像分发

数十秒

千亿参数模型下载和预热

数分钟

模型推理框架加载

推理优化

算子加速	PD集群编排	智能网关	分级 KVCache
2-4x 加速	3x 效率提升	60% 降低	10x 提速
自研推理框架	PD 实例自动扩缩容	智能路由 多 Lora 加载	长文本及多轮对话 KVCache 命中率

智能网关

大 EP＋PD 分离分布式推理引擎

Prefi ll → RoCE → Decode

KVCache 分级缓存

运维观测

推理业务全链路可观测

0 侵入

透明埋点

0 侵入

推理业务可观测

全链路观测

最佳实践模版

已适配模型权重文件 100+

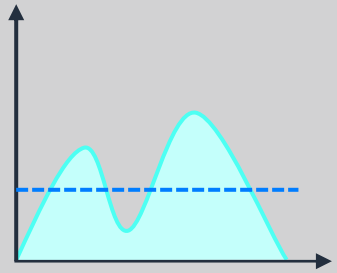
推理加速框架镜像

一键部署服务

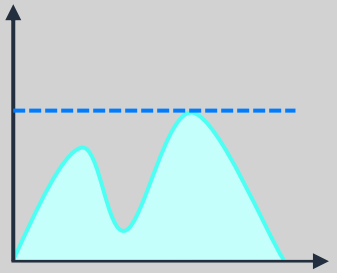
灵活的算力调度：推出 SiliconFlow FaaS，为灵活可调度的 AI 负载设计的新一代算力引擎

算力资源管理需兼顾负载与利用效率

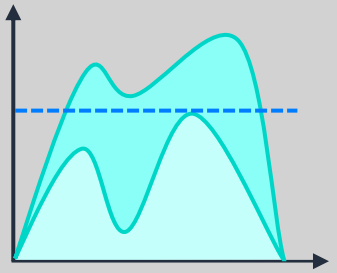
低算力储备造成资源过载



高算力储备导致利用率低



多租户下算力调度重要性更高



客户A业务负载　客户B业务负载　算力资源

弹性灵活的算力调度是降低推理成本、提升算力利用效率的关键要素

硅基流动 SiliconFlow

Function as a Service

新一代 Serverless GPU 调度平台

帮助用户 灵活、便捷、可靠 的将推理服务部署到云端算力之上

算力资源弹性伸缩

算力自由调度，0资源浪费

异构算力无感融合

多集群异构算力统一纳管

极致的成本管控

支持命中缓存计费

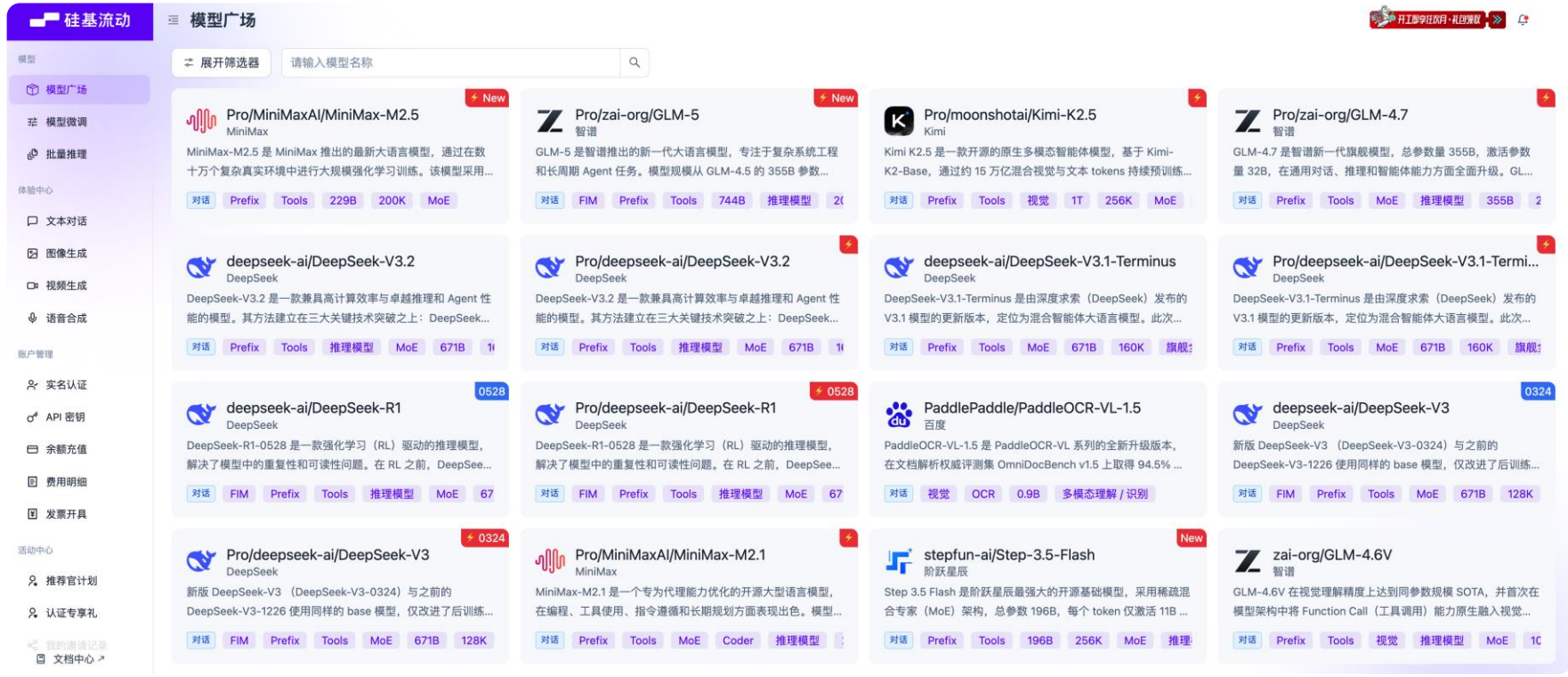
*前述特定亚马逊云科技生成式人工智能相关的服务目前在亚马逊云科技海外区域可用。亚马逊云科技中国区域相关云服务由西云数据和光环新网运营，具体信息以中国区域官网为准。

多样化模型选择：基于对推理加速的深刻理解和超强的推理引擎架构，最快适配主流开源模型

SiliconFlow 推理引擎具备极强的抽象能力，基于扎实的通用优化框架与前瞻性策略迭代更新，将不同模型的适配与加速沉淀为可复用的标准化路径

140+ 主流模型适配 — *最全*

95%+ 发布当天适配 — *最新*



*前述特定亚马逊云科技生成式人工智能相关的服务目前在亚马逊云科技海外区域可用。亚马逊云科技中国区域相关云服务由西云数据和光环新网运营，具体信息以中国区域官网为准。

全栈自研产品架构：基于高性能推理能力，满足多层次的 AI 应用部署需求

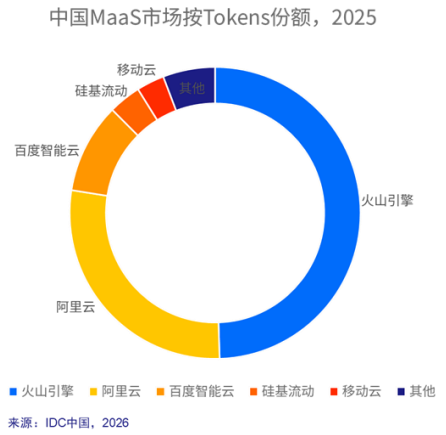
*前述特定亚马逊云科技生成式人工智能相关的服务目前在亚马逊云科技海外区域可用。亚马逊云科技中国区域相关云服务由西云数据和光环新网运营，具体信息以中国区域官网为准。

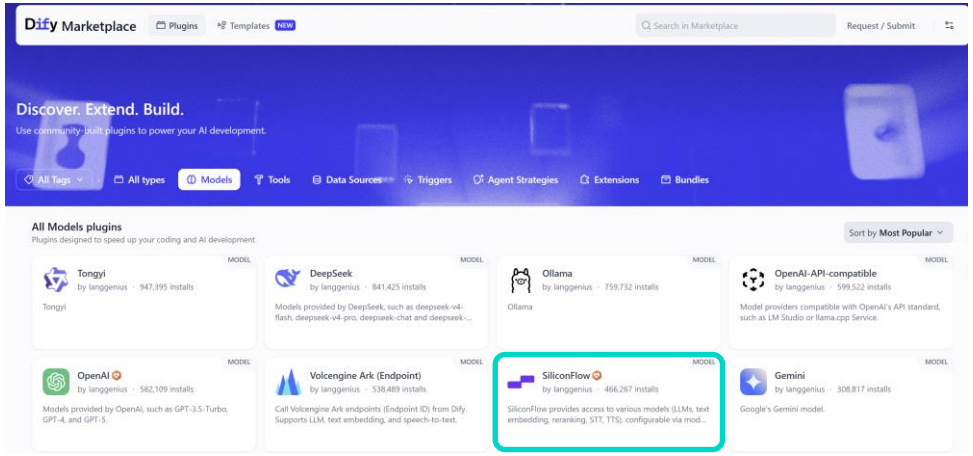
用户心智：最受开发者欢迎的 API 供应商之一，主流 Marketplace 安装量领先

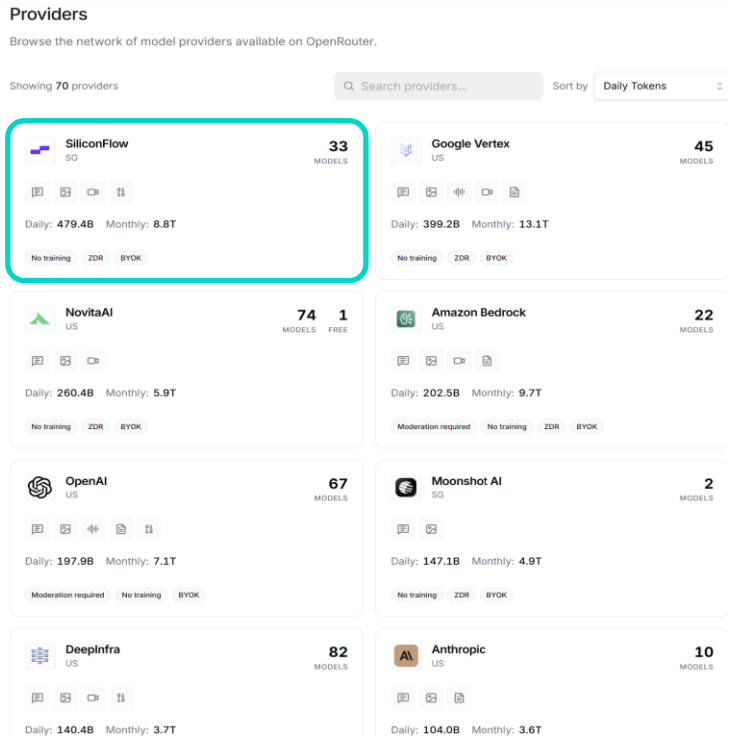
硅基流动 SiliconFlow

在主流开发者 Marketplace 和第三方统计报告 上的使用量处于第一梯队，整体领先主流云厂商与头部模型公司，已与同领域创业公司拉开约 300 倍差距，已在开发者群体中建立极强的用户心智

IDC 2026年5月发布，硅基流动在中国 MaaS 市场 Token 使用量排名第四，唯一进入前五名的创业公司







*前述特定亚马逊云科技生成式人工智能相关的服务目前在亚马逊云科技海外区域可用。亚马逊云科技中国区域相关云服务由西云数据和光环新网运营，具体信息以中国区域官网为准。

硅基流动在 Token 供应链各环节领跑全行业，打造 Token 经济中的“超级供应商”

硅基流动 SiliconFlow

全球领先的**新一代 AI Native Cloud**，不做简单的 API 转售，而是针对多种模型和算力底层重构的推理引擎。开发者可以**针对特定场景选择最佳的“模型 x 算力”组合**，彻底解决多模型组合及性价比问题，这背后是扎实的**“AI 供应链基础设施”**

最丰富的算力方案

NVIDIA　AMD

...

全面适配多种算力芯片

满足国内外多元化的客户需求

最灵活的算力调度

- 算力资源弹性伸缩
- 异构算力无感融合
- 算力成本极致管控

经过检验的云服务架构能力

多 SKU“模型 x 算力”云端秒级响应

最全的模型选择

deepseek　Qwen

智谱·AI

Moonshot AI　MINIMAX

...

全面适配加速 140+ 款主流开源模型

绝大多数模型发布当天完成适配

*前述特定亚马逊云科技生成式人工智能相关的服务目前在亚马逊云科技海外区域可用。亚马逊云科技中国区域相关云服务由西云数据和光环新网运营，具体信息以中国区域官网为准。

携手亚马逊云科技打造全球 Token 工厂

算力资源丰富

- 全球 30+ 区域提供多种高性能计算实例
- SiliconFlow 可按需获取大规模异构算力，快速匹配不同模型对算力的差异化需求

弹性与规模

- 应对流量波峰，支撑大规模 Token 生产
- 确保 Token 生产的"工厂产能"始终匹配实际需求，同时最大化资源利用率、控制成本

全球化服务

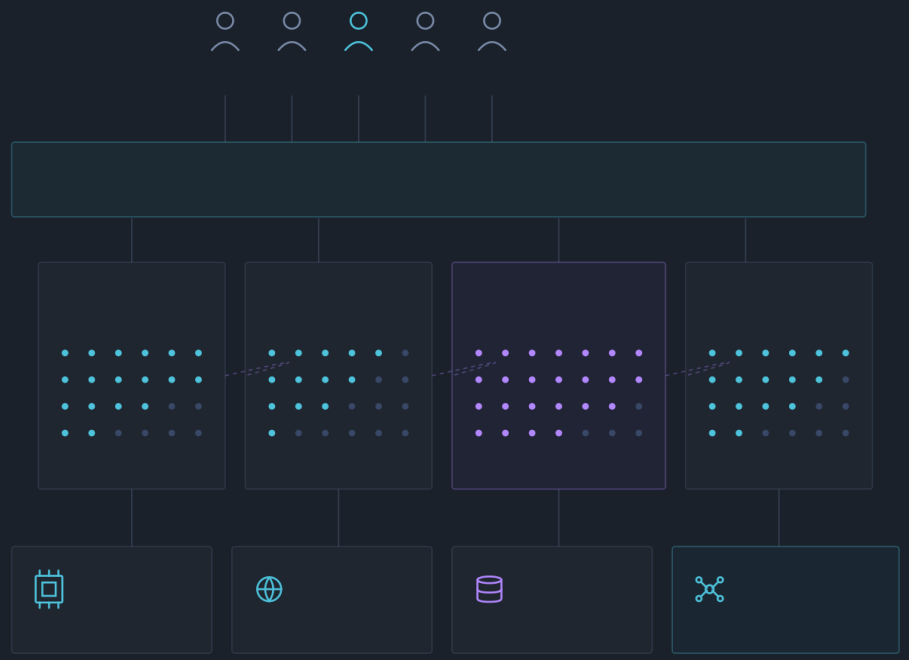
- 遍布全球的基础设施与网络加速能力
- SiliconFlow 为海外开发者与企业客户就近提供推理服务，实现低延迟、高可用的全球化 Token 交付

*前述特定亚马逊云科技生成式人工智能相关的服务目前在亚马逊云科技海外区域可用。亚马逊云科技中国区域相关云服务由西云数据和光环新网运营，具体信息以中国区域官网为准。

基于亚马逊云科技的全球化部署

GLOBAL INFERENCE, POWERED BY AMAZON WEB SERVICES

SYSTEM ARCHITECTURE



01 / GLOBAL RESOURCE POOL

全球算力资源池

覆盖多大洲多 Region 的 GPU 资源,就近接入、就近推理，从基础设施层面降低端到端延迟

02 / ELASTIC SCHEDULING

GPU 弹性调度

按流量动态伸缩、跨 Region 资源融通，峰谷错峰提升整体利用率

03 / FOUNDATION STACK

算力 · 网络 · 存储协同优化

GPU 算力集群 + 高带宽低延迟网络 + 高吞吐对象存储，三层协同支撑大规模推理

04 / HIGH AVAILABILITY

企业级稳定性与容灾

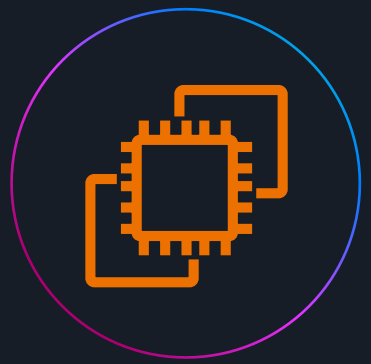
多可用区部署、故障自动转移、全链路可观测

全球用户的每一次 **Token 请求**，都能就近、稳定、低延迟地完成

*前述特定亚马逊云科技生成式人工智能相关的服务目前在亚马逊云科技海外区域可用。亚马逊云科技中国区域相关云服务由西云数据和光环新网运营，具体信息以中国区域官网为准。

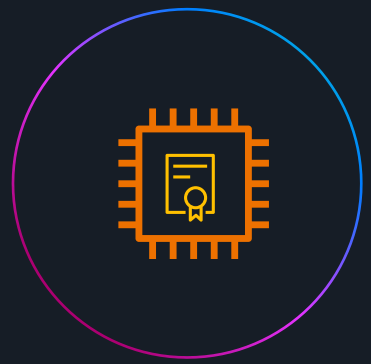
亚马逊云 GPU 使用方式

按需使用



按需预置
可变用量
灵活性高

容量预留



预留容量
稳态运行
更高可用性

容量块

加速计算实例
有时间限制
低延迟集群

EC2 竞价实例

可能会中断
价格低

* 最新超大规模实例：以 Capacity Blocks 为主（按时段预留集群）

*前述特定亚马逊云科技生成式人工智能相关的服务目前在亚马逊云科技海外区域可用。亚马逊云科技中国区域相关云服务由西云数据和光环新网运营，具体信息以中国区域官网为准。

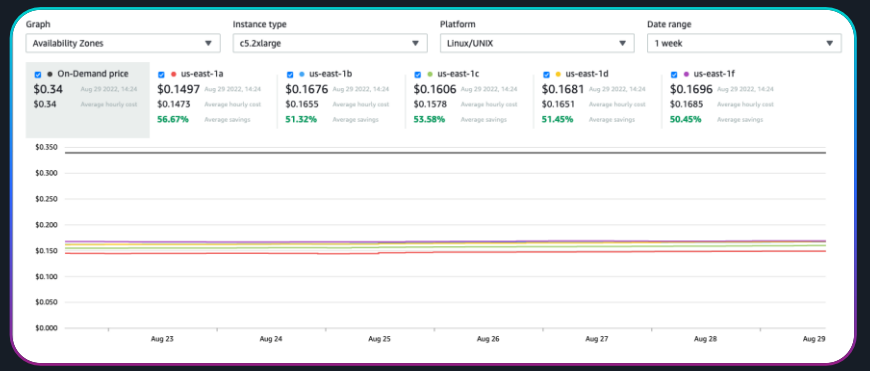
GPU 竞价实例

解决成本问题

Spot 实例是空闲 EC2 容量，可作为购买选项提供给客户，并享受大幅折扣

空闲 EC2 容量

与按需部署相同的基础架构



高达 90% 的按需价格折扣

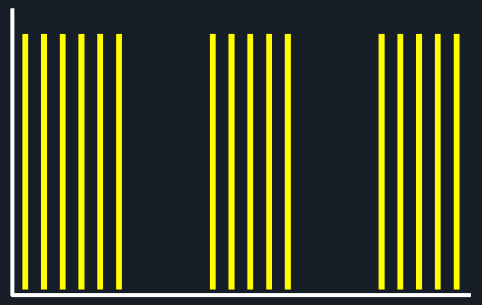
Spot 价格基于 EC2 实例的长期供需趋势（无竞价）

可中断

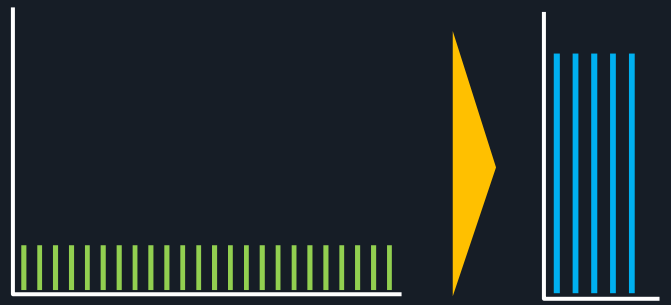
如果 EC2 特定容量池需要恢复容量，Spot 实例可能会中断，并在中断前 2 分钟发出警告

*前述特定亚马逊云科技生成式人工智能相关的服务目前在亚马逊云科技海外区域可用。亚马逊云科技中国区域相关云服务由西云数据和光环新网运营，具体信息以中国区域官网为准。

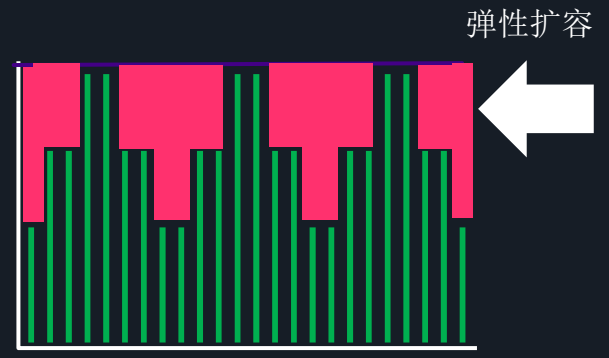
Spot GPU 的使用模式



微调、基准测试工作负载



实验、开发 — 低负载持续使用，偶尔短时爆发



突发实时灵活推理 — 用 Spot 实例作为弹性扩容，补充峰值推理容量

*前述特定亚马逊云科技生成式人工智能相关的服务目前在亚马逊云科技海外区域可用。亚马逊云科技中国区域相关云服务由西云数据和光环新网运营，具体信息以中国区域官网为准。

Spot 中断最佳实践 – 灵活性，中断处理

实例灵活

地区灵活

使用尽可能多且深的容量池

- 不同实例类型
- 不同实例大小
- 不同可用区

时间灵活

价格灵活

时间灵活和/或价格灵活可以进一步降低中断率，
提高应用程序正常运行的时间

EC2 实例再平衡信号
(主动式)

- 当您的 Spot 实例面临较高的中断风险时的通知
- 内置支持与 EC2 Auto Scaling、EKS 托管节点组等服务的集成

Spot 实例终止通知
(被动)

- 在 Spot 实例被中断前 2 分钟做出响应
- 内置支持相同的亚马逊云科技服务
- 中断处理（亚马逊云科技提供了推荐用例的最佳实践方案）

*前述特定亚马逊云科技生成式人工智能相关的服务目前在亚马逊云科技海外区域可用。亚马逊云科技中国区域相关云服务由西云数据和光环新网运营，具体信息以中国区域官网为准。

Spot Placement Score (SPS)

- 能够指示在给定时间点，哪些实例类型和区域最有可能成功启动 Spot 实例
- P 系列特殊情况 — 支持单一实例类型请求
- SPS 需要实例多样性（至少 3 种类型）才能给出可靠分数
- 策略：
  - 根据性能需求（模型大小、GPU 性能、驱动程序等）混合使用不同加速器
  - 使用 SPS 识别最适合用例的区域和时间

```
“ aws ec2 get-spot-placement-scores
 --region us-east-1
--instance-types p5en.48xlarge
--target-capacity 1920
--target-capacity-unit-type vcpu
 --region-names us-east-2 us-west-1”
```

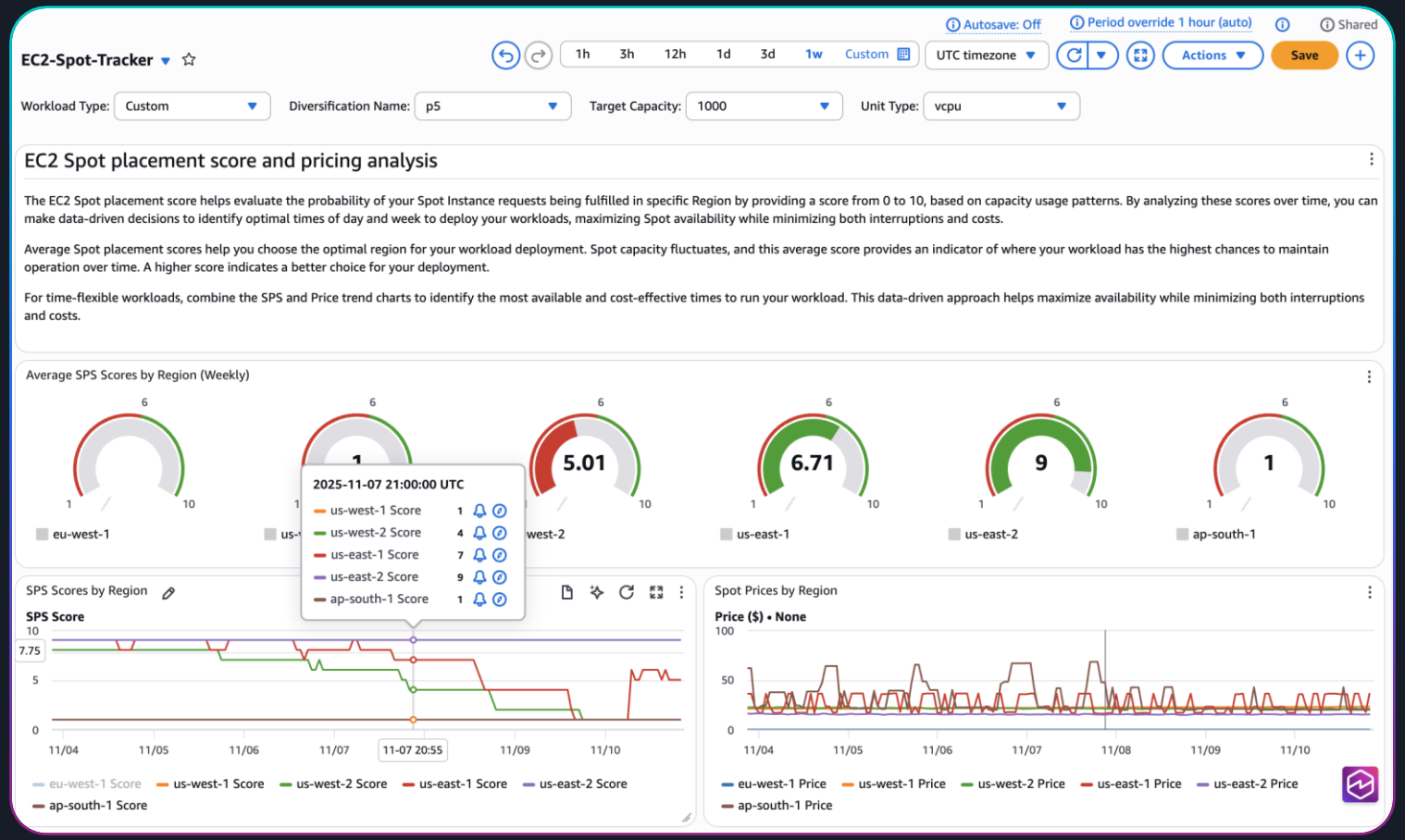
p5en.48xlarge × 10 台 — us-west-2

AZ	评分	判断
usw2-az4	**9**	✅ 非常好，大概率拿到
usw2-az3	**8**	✅ 非常好，大概率拿到
usw2-az2	6	🟡 有概率，但不确定

10 台在美西2表现不错，**usw2-az4 评分 9 分**，非常稳。建议优先选 az4。

SPOT Placement Score Tracker

- SPS Tracker 提供一段时间内的趋势分析，帮助识别多样化策略
- 实时、短期、突发推理 — 识别最优区域
- 时间灵活的研发、实验、模型调优 — 识别最优时间



github.com/aws-solutions-library-samples/guidance-for-ec2-spot-placement-score-tracker

*前述特定亚马逊云科技生成式人工智能相关的服务目前在亚马逊云科技海外区域可用。亚马逊云科技中国区域相关云服务由西云数据和光环新网运营，具体信息以中国区域官网为准。

总结

THREE TAKEAWAYS

01

TOKEN FACTORY

Token 工厂 =
工业化推理生产体系

把推理从「手工作坊」升级为可规模化、可经济化、可工程化的生产线

02

FULL-STACK OPTIMIZATION

全栈优化实现
10× 性能提升

从调度、引擎、KV-Cache 到底层算力，全栈端到端协同优化

03

OPTIMAL TCO

亚马逊云全球算力 +
SiliconFlow 推理引擎

全球资源池 + 工业级推理引擎，性能、稳定性、成本三者兼得

*前述特定亚马逊云科技生成式人工智能相关的服务目前在亚马逊云科技海外区域可用。亚马逊云科技中国区域相关云服务由西云数据和光环新网运营，具体信息以中国区域官网为准。

亚马逊云科技
中国峰会

Thank you

扫码关注“硅基流动”公众号

contact@siliconflow.cn

清华科技园启迪科技大厦 D 座 23 层

亚马逊云科技
中国峰会

Thank you

扫码关注“硅基流动”公众号

contact@siliconflow.cn

清华科技园启迪科技大厦 D 座 23 层