亚马逊云科技
中国峰会

Powered by 亚马逊云科技：Zilliz 构建企业级 AI 应用的数据底座

Leo Shen

Director of Solution Architect

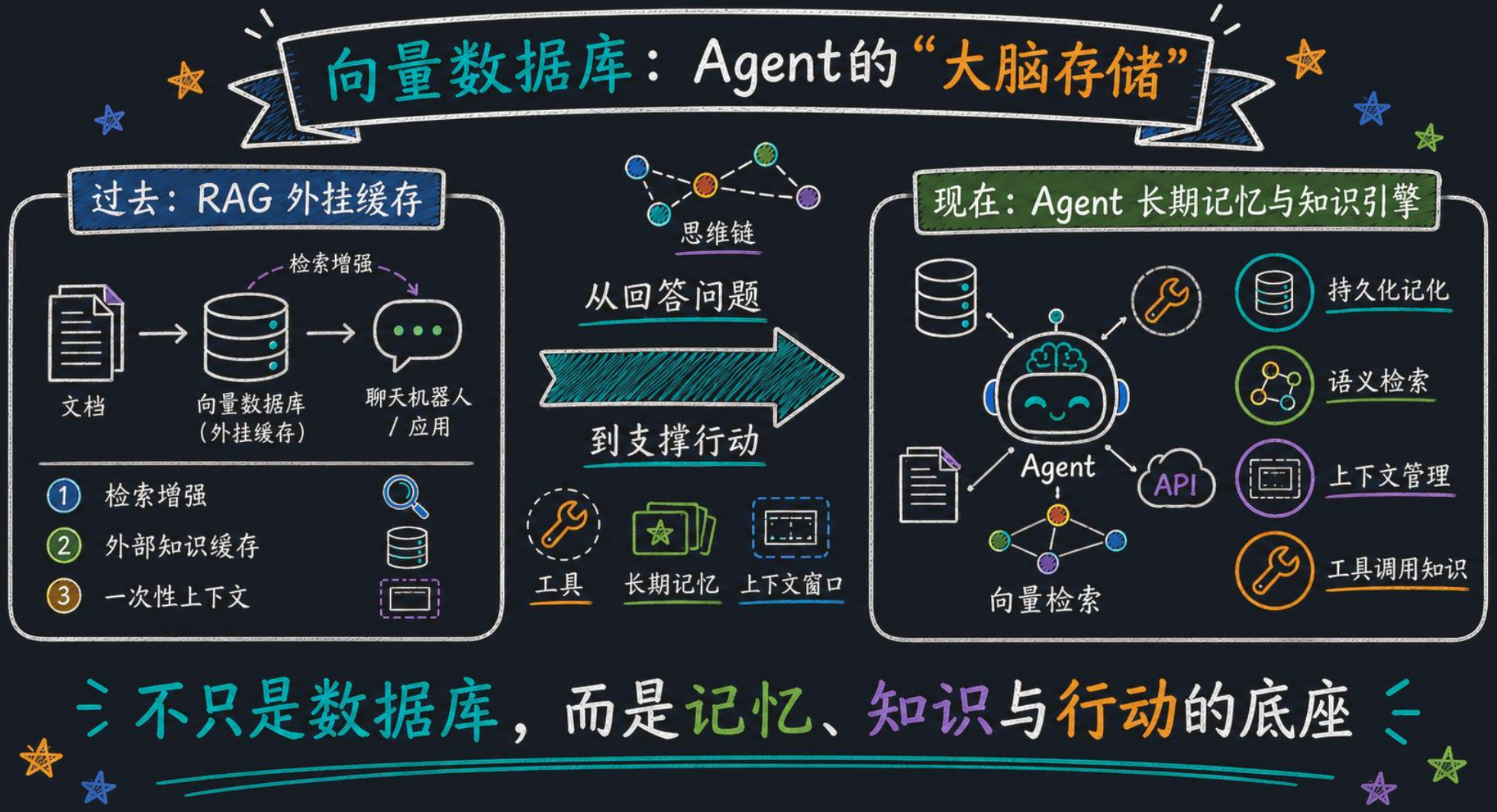
Zilliz

Wang Yi

Solution Architect

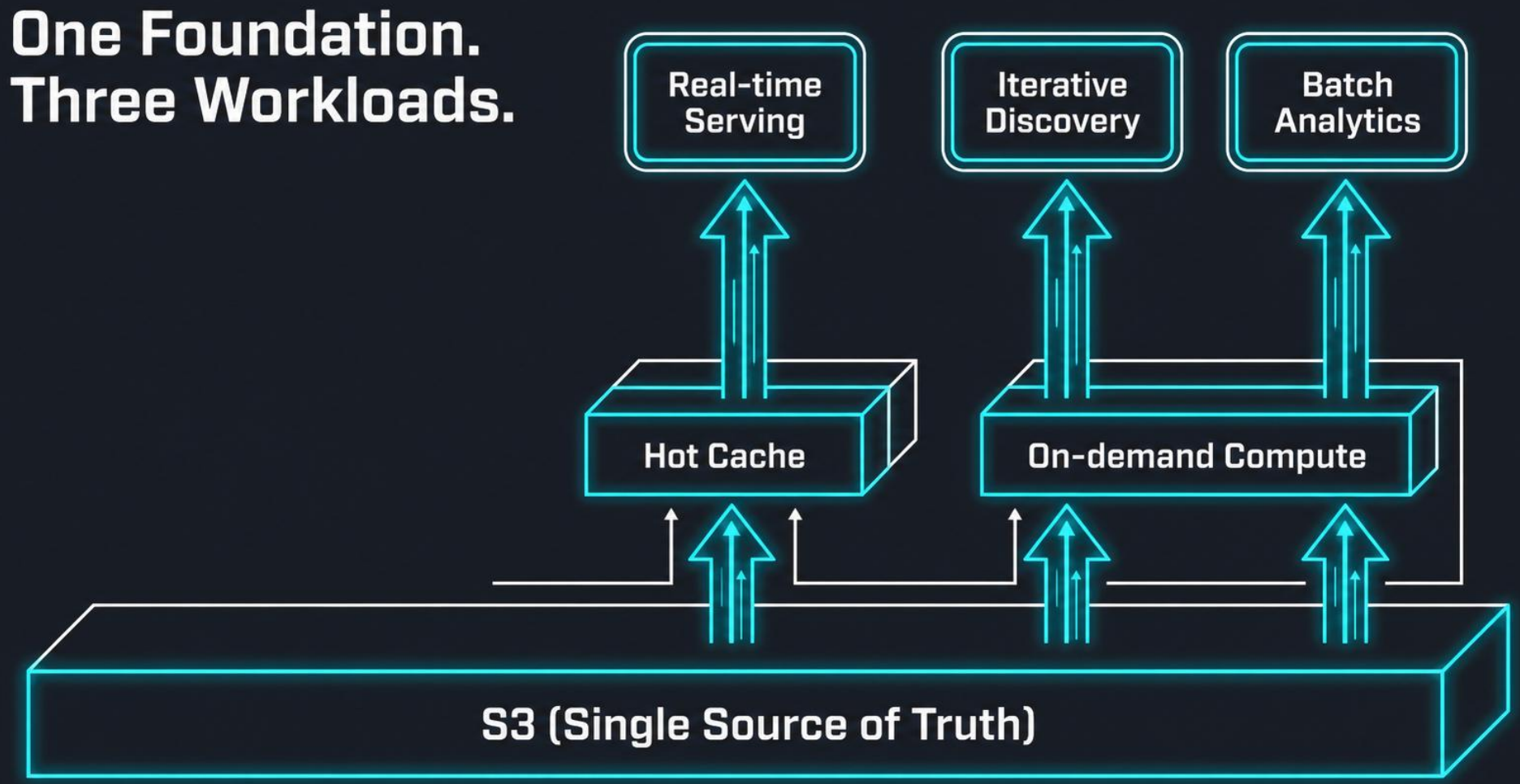
亚马逊云科技

Agentic AI 时代，向量数据库角色进化



*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。

Zilliz Cloud 产品架构概览



*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。

Zilliz Cloud 产品形态

自运维

Milvus

广受欢迎的开源向量数据库

milvus lite　docker　kubernetes

全托管服务

Zilliz Cloud

AI 驱动的高性能、高扩展向量检索

亚马逊云科技

BYOC

Zilliz Cloud BYOC

专为私有化场景打造

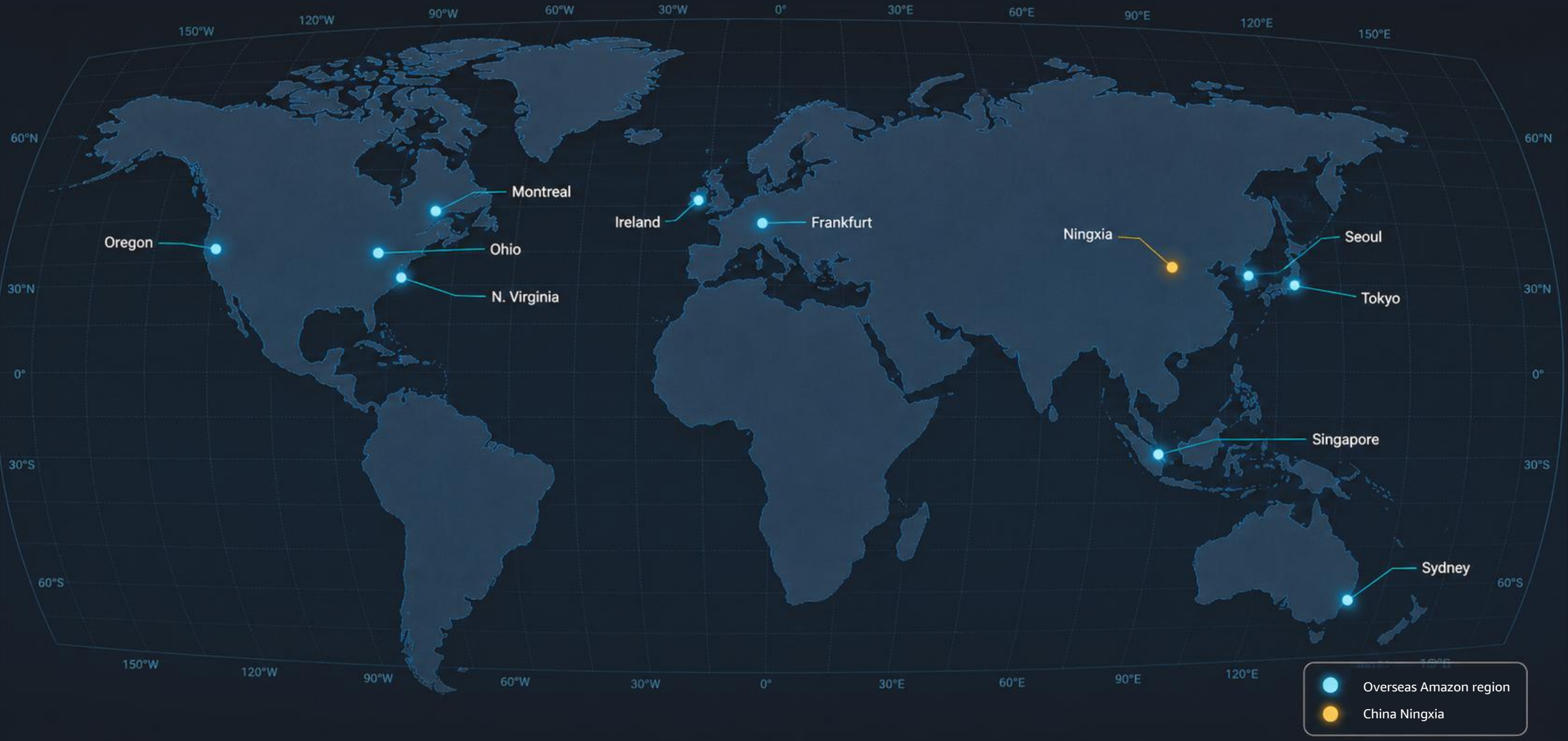
亚马逊云科技

milvus

所有部署形态 API 统一，业务逻辑代码灵活复用

*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。

遍布全球亚马逊云科技节点的 Zilliz Cloud



为什么选择亚马逊云科技作为全球化底座

全球覆盖

30+ Region

数据本地化合规

区域先到，业务才能先到

一致体验

Graviton/EKS/S3

全球一致可用

平台能力可复制，交付才可规模化

生态集成

Bedrock/PrivateLink/

Marketplace

更短集成路径，更低企业采用门槛

ISV 合作

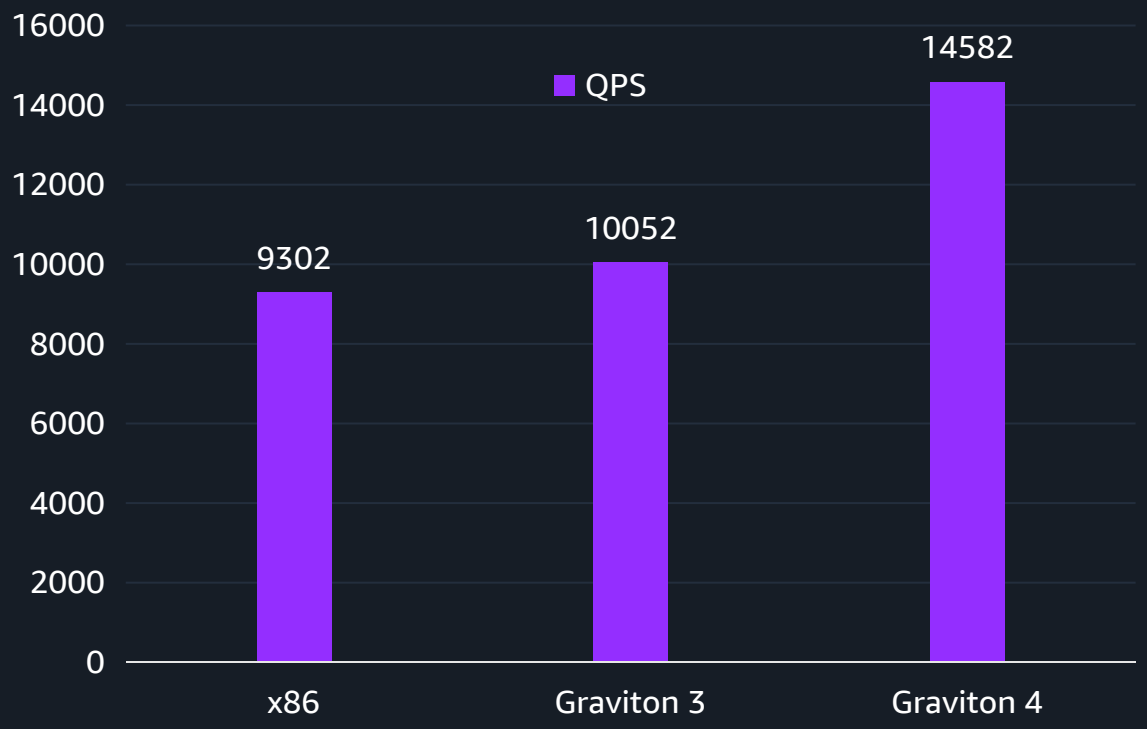
Partner Network

联合销售机制

从技术合作走向商业共赢

*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技术实践 —— 计算与存储



数据集: Cohere 1M 768dim
基准测试工具: https://github.com/zilliztech/VectorDBBench
主机: Zilliz Cloud 8CU-perf

发挥 Graviton 的带宽优势：

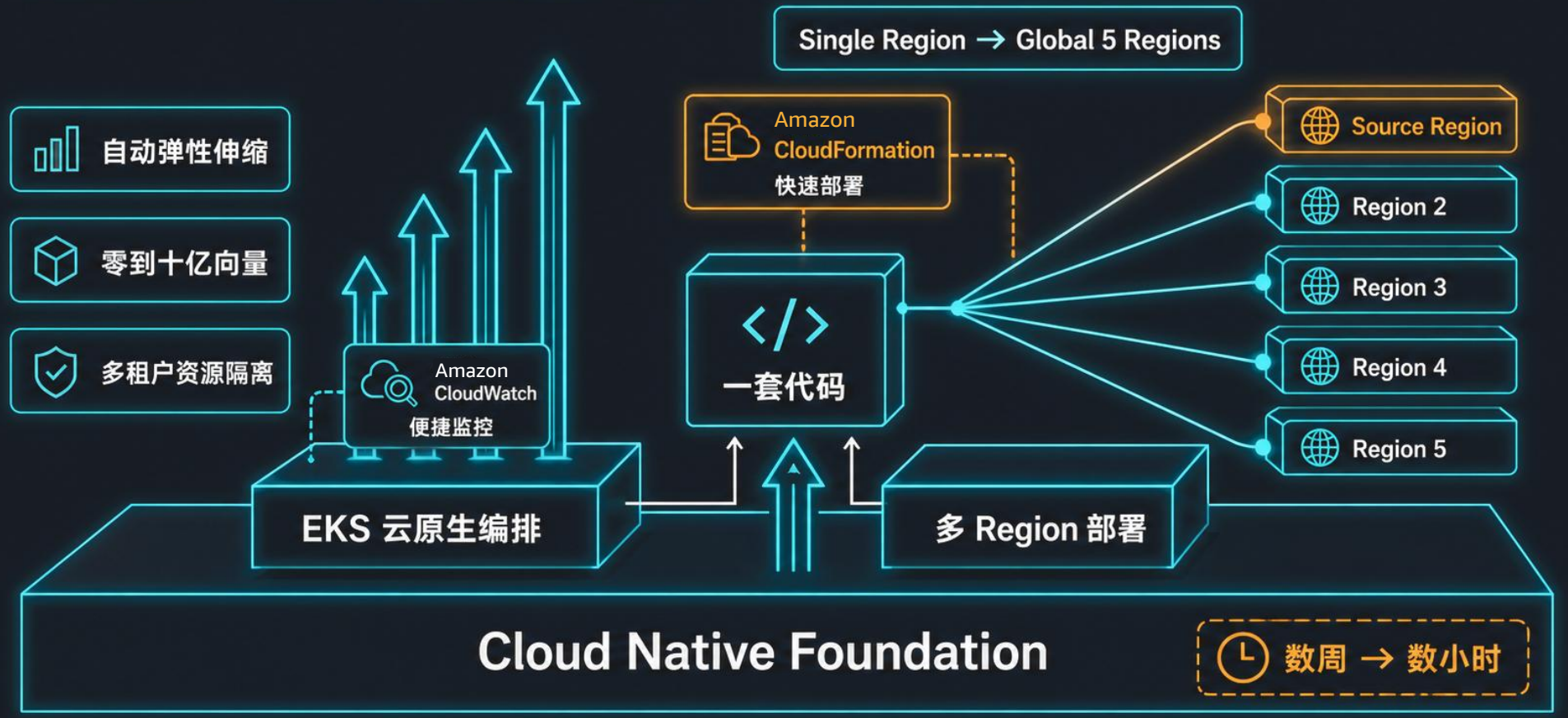
相比传统的 x86 架构，Graviton 芯片提供更高的内存带宽。Zilliz Cloud 相应地调整算法设计，以降低带宽瓶颈，实现更优的计算性能。

S3 持久化存储：

- 向量索引 + 元数据的分层存储策略
- 生命周期管理实现成本精细化
- 实际存储成本节降约30%

基于亚马逊云科技的技术实践 —— 弹性与全球交付

云原生 + 全球化工程实践



*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。

出海合规与企业级集成

*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。

客户案例 —— 北美法律 AI 独角兽 Filevine

FILEVINE × zilliz

Filevine四亿刀融资跻身法律AI独角兽，背后的infra怎么搭

Original 和你一起学习的 Zilliz 2025年9月24日 18:07

好消息，**Zilliz Cloud用户——法律AI独角兽 Filevine** 今日宣布完成 **4 亿美元融资。**

公开资料显示，Filevine 员工总数仅在500–1000人，但在法律领域，**Filevine 拥有近6000家客户，服务超过10万名法律专业人士**，日均文件上传量超过 2,000 万页，总处理文档超十亿份。成为全球法律AI领域绝对头部玩家。

与此同时，与多数C端AI产品不同之处在于，Filevine 的产品有着超高的用户粘性，核心产品保持着**96%以上的留存率，美元净留存率超过120%**。

那么Filevine 如何用这么小的企业规模，撑起如此庞大的业务量的？

我们**独家采访了Filevine 团队**，在他们看来，选择合适的AI与向量数据库，是公司快速在传统行业突围、加速成长的重要助力。

*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。

亚马逊云科技视角 —— 性能、成本与全球化

Amazon PERSPECTIVE

从向量数据库延伸到完整
AI 应用基础设施

Graviton 降本增效：
芯片架构 + 容器化实践

*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。

Graviton 广泛适用于容器生态

编排器


Amazon ECS


Amazon EKS


Docker Swarm


Kubernetes

镜像注册表


Amazon ECR


Docker Hub

容器优化Linux发行版


Bottlerocket

无服务器


Fargate


Lambda

*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。

DevOps 生态系统中的广泛 Graviton 支持

完全托管

Amazon CodeBuild

circleci

Cirrus CI

Travis CI

GitHub

混合（托管/自建）

GitHub

GitLab

自管理

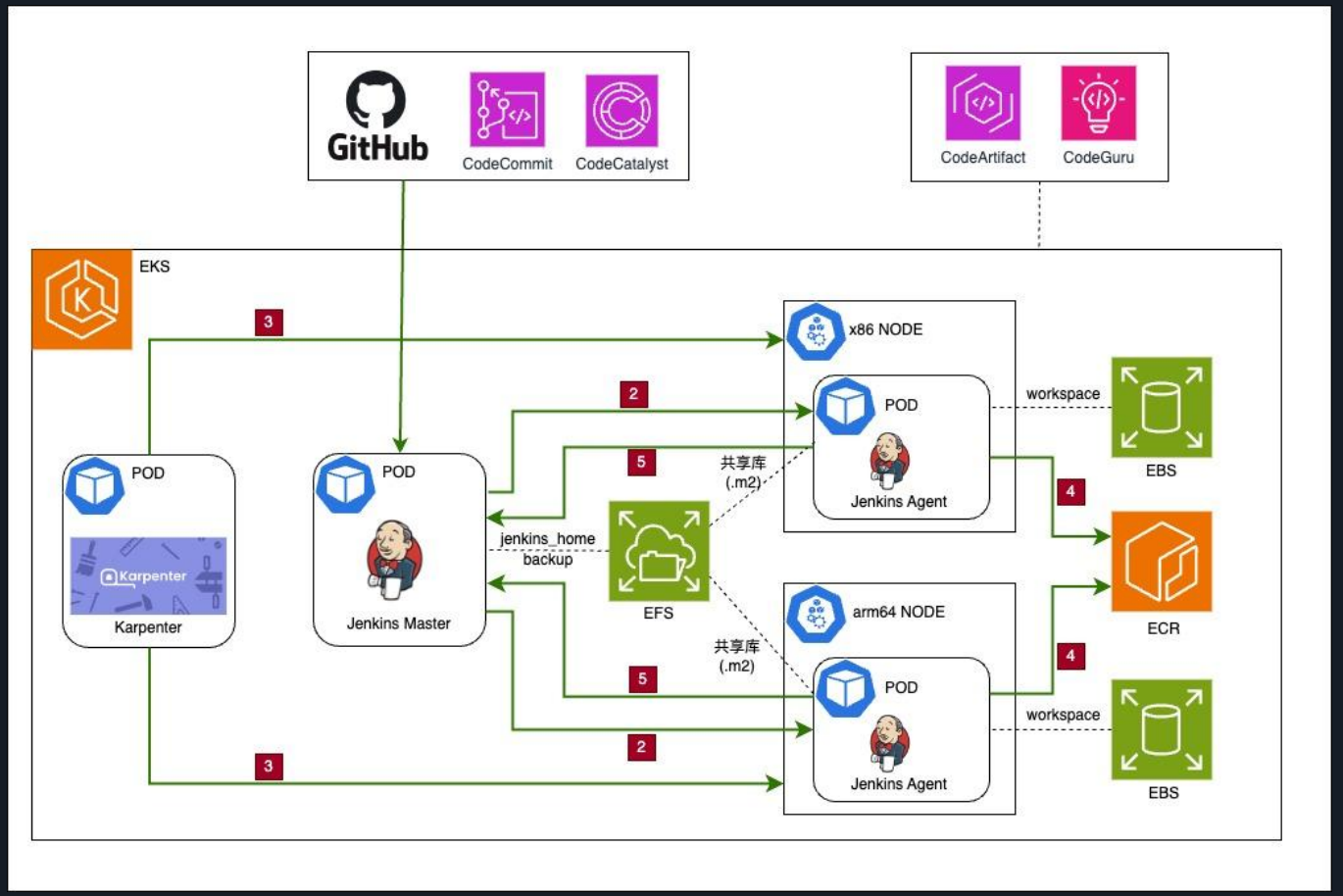
Jenkins

argo

*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。

DevOps 架构：Jenkins in EKS with Graviton

- Graviton提升了Jenkins40%的构建性能，Maven/Gradle构建速度显著加快
- ARM 架构的高效能核心设计，支持Jenkins Agent并发执行
- **Jenkins Master** 运行在稳定的x86节点，**Jenkins Agent** 动态调度到Graviton Spot节点，执行实际构建任务
- **Karpenter** 实时监控作业队列，30秒内自动provision合适规格的Graviton实例, 并提供**混合架构支持**



*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。

DevOps 架构：ArgoCD can run on Graviton easily

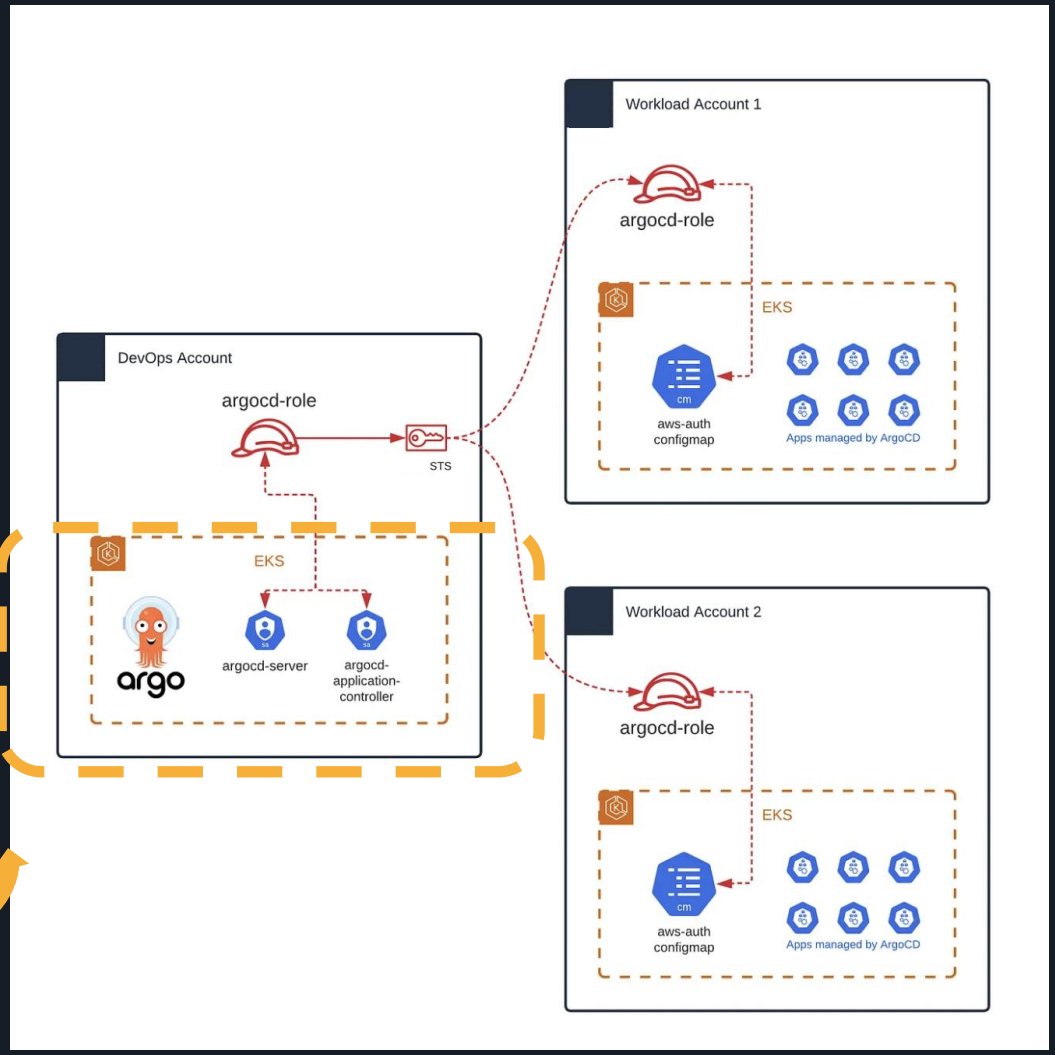
GitHub

argoproj/argo-cd

#6830 Argo-cd support for arm64 based docker images

原生 ARM64 支持

- 官方支持 - ArgoCD 正式支持 ARM64 镜像
- 无缝迁移 - 所有组件完美运行在 Graviton 上
- 性能提升 - 同步和部署速度提升30%



Graviton 迁移复杂度

易用性	工作负载	操作
无缝迁移	Amazon RDS、Amazon Aurora、Amazon ElastiCache、Amazon OpenSearch、Amazon MemoryDB 和 Amazon Neptune	更新至最新版可使用最新功能
	Amazon EMR	一般可直接使用
	基于 Graviton 支持的 ISV（开源/商业版）	根据应用而定，但通常能无缝迁移
简单操作	Amazon Lambda	通常只需结合 Lambda 托管的运行时或基础镜像 ** 检查 Java 本地接口 (JNI)、共享对象或原生模块*
一般难度的操作	Linux – 解释型和即时编译型语言（如 Java、PHP、Node.js）	选择 Arm64 AMI 并安装 **其他操作（若容器化）** ** 检查 Java 本地接口 (JNI)、共享对象或原生模块*
相对复杂的操作	Linux – 编译型语言或依赖（如 C/C++、Python、Go）	选择 Arm64 AMI 并编译 ** 移植任何 intrinsics 函数、汇编或原生模块*
多种操作，高回报	Microsoft Windows – .NET	迁移至 Linux + 基于 Arm64 的 .NET Core

*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