

QCon
全球软件开发大会

从云原生到 AI 原生： 模型引发的新一代基础设施构建

余锋（褚霸）

蚂蚁集团
基础设施技术委员会主席
超级计算技术部负责人



目录

- 01 背景：基础设施技术的演进
- 02 新时代的应用：模型成为新的数据库
- 03 模型为中心的基础设施
- 04 AI 中间件：Agent 基础设施
- 05 展望：从“通算+智算”向通智一体的演进

📍 北京

QCon

全球软件开发大会

会议时间：4月10-12日

- 大模型赋能 AIOps
- 越挫越勇的大前端
- 云上业务架构演进
- 多模态大模型及应用
- 海外 AI 应用创新实践

📍 北京

AiCon

全球人工智能开发与应用大会

会议时间：6月27-28日

- 端侧智能
- AI Agent
- 多模态大模型
- 金融+大模型

📍 上海

QCon

全球软件开发大会

会议时间：10月23-25日

- AI Agent
- 大模型训练推理
- 端侧 AI
- 搜推深度融合
- 数智金融

4月

5月

6月

8月

10月

12月

📍 上海

AiCon

全球人工智能开发与应用大会

会议时间：5月23-24日

- 通用大模型
- AI Agent
- 垂直领域应用
- 模型可解释性

📍 深圳

AiCon

全球人工智能开发与应用大会

会议时间：8月22-23日

- 模型效率与部署
- 多模态大模型
- 大模型安全
- 智能硬件

📍 北京

AiCon

全球人工智能开发与应用大会

会议时间：12月19-20日

- 通用大模型
- 智能硬件
- LMOPs
- 具身智能

极客邦科技 2025 年会议规划

促进软件开发及相关领域知识与创新的传播



参会咨询

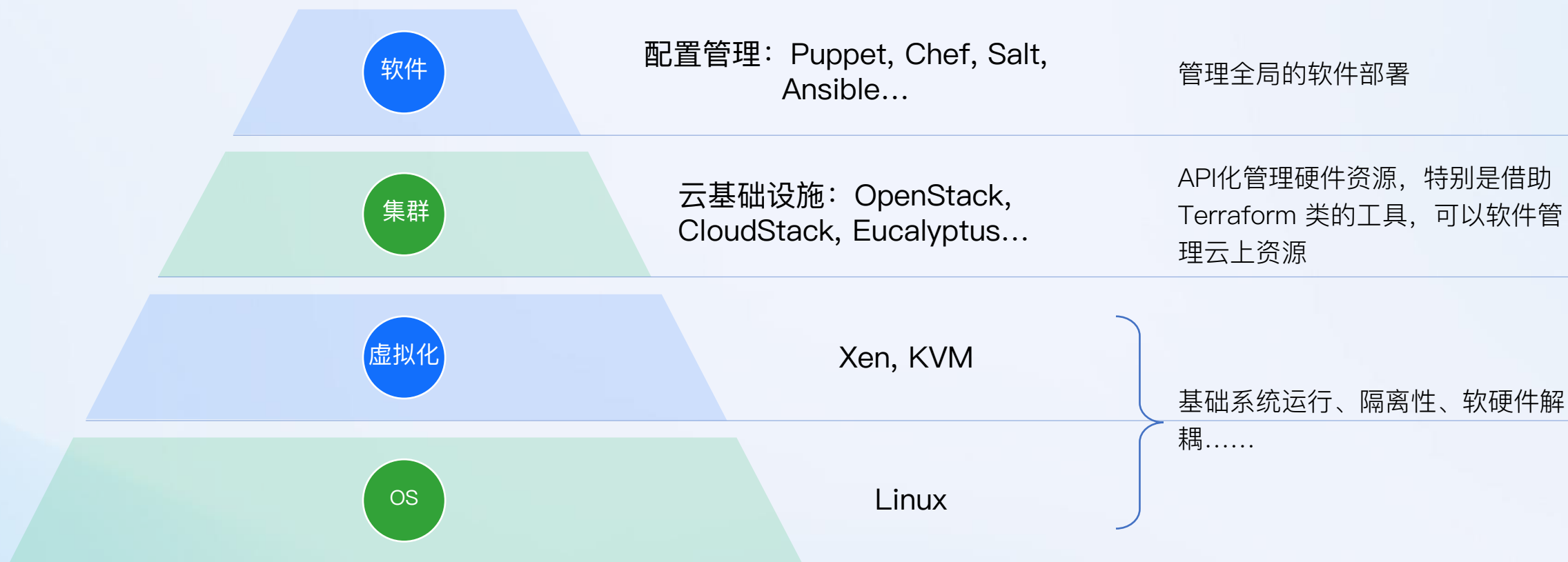


查看会议

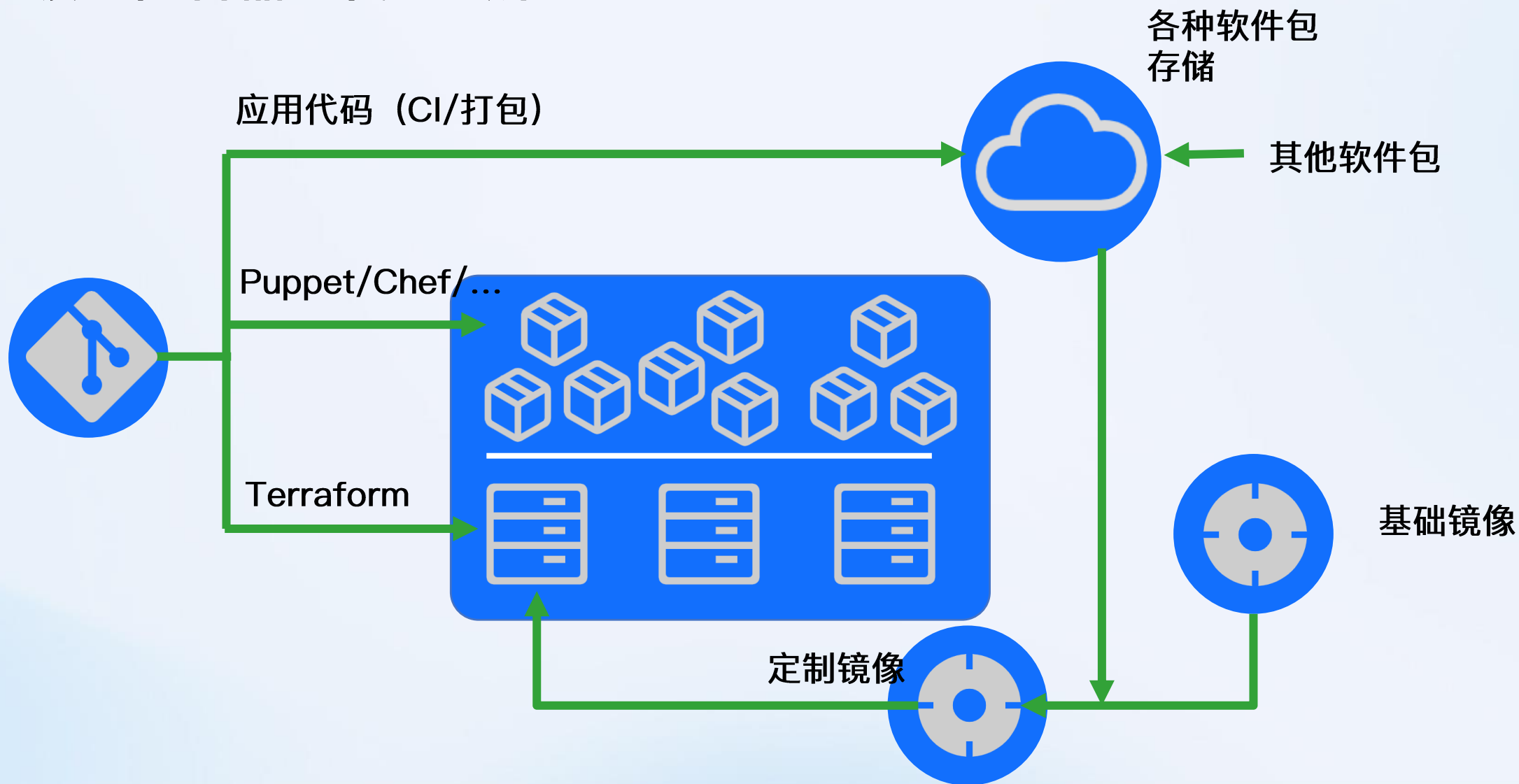
01

基础设施技术的演进

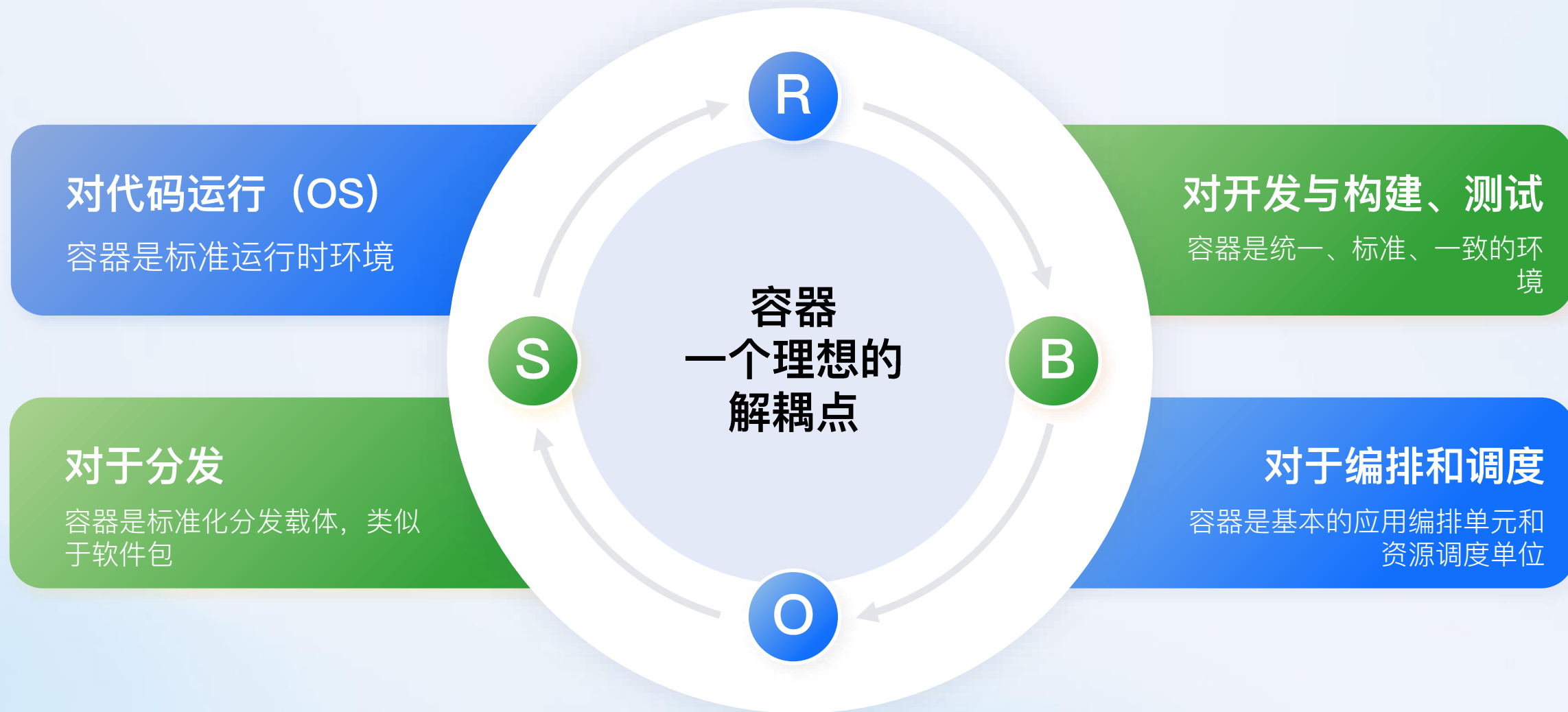
软件定义基础设施之始：虚拟化与云计算



● 过渡：前容器时代的云原生

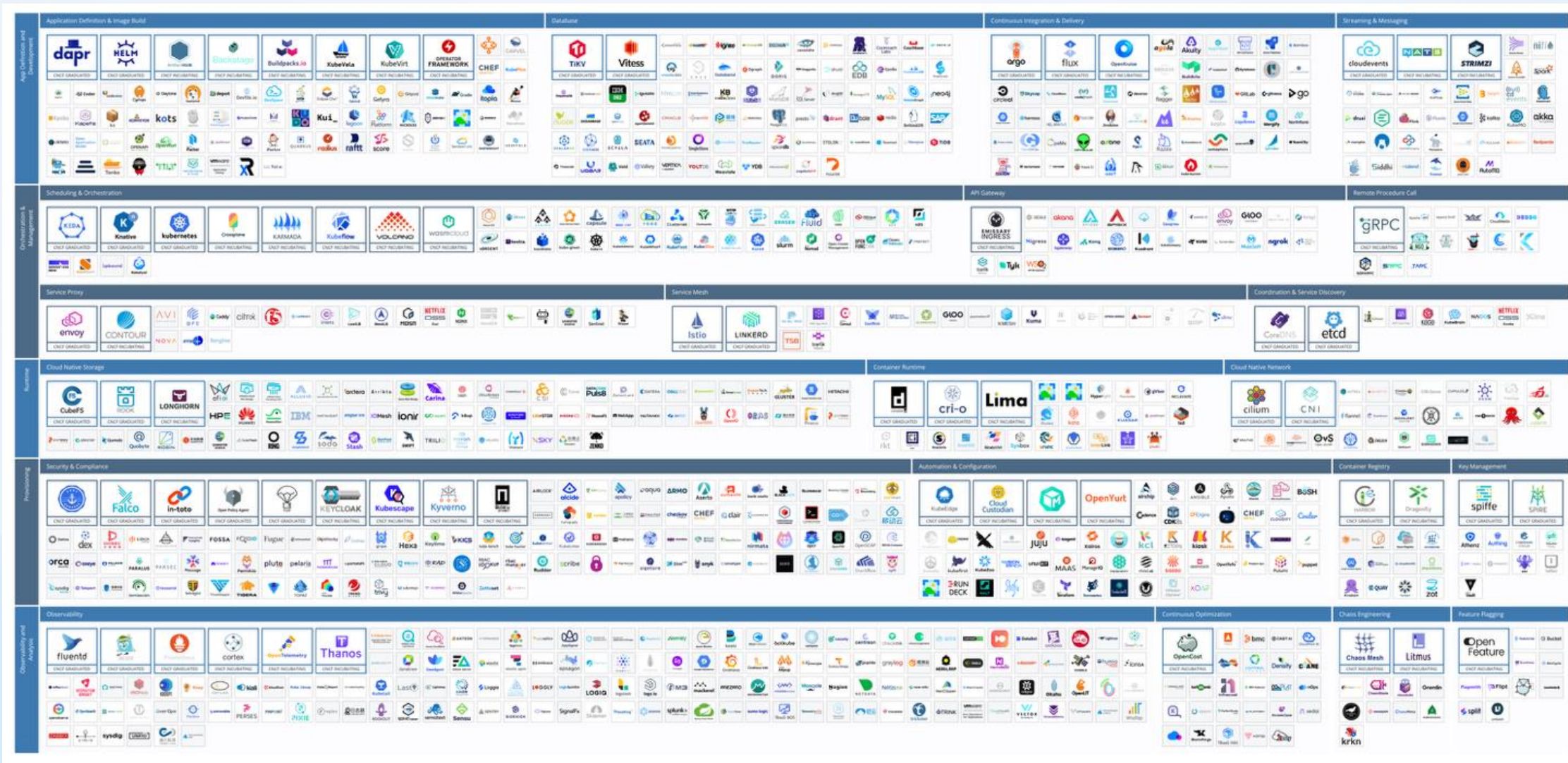


容器与云原生：应用为中心





云原生：良好架构带来的生态繁荣



小结：基础设施的演化方向



性能与效率



适应负载变化



应对环境的不稳定性和风险

降低运维复杂度 → 直至免运维

02

GenAI 时代的应用

模型成为新的数据库

● GenAI 之前的（互联网）应用 Pattern

LAMP 的演进之路

数据库是应用系统的核心

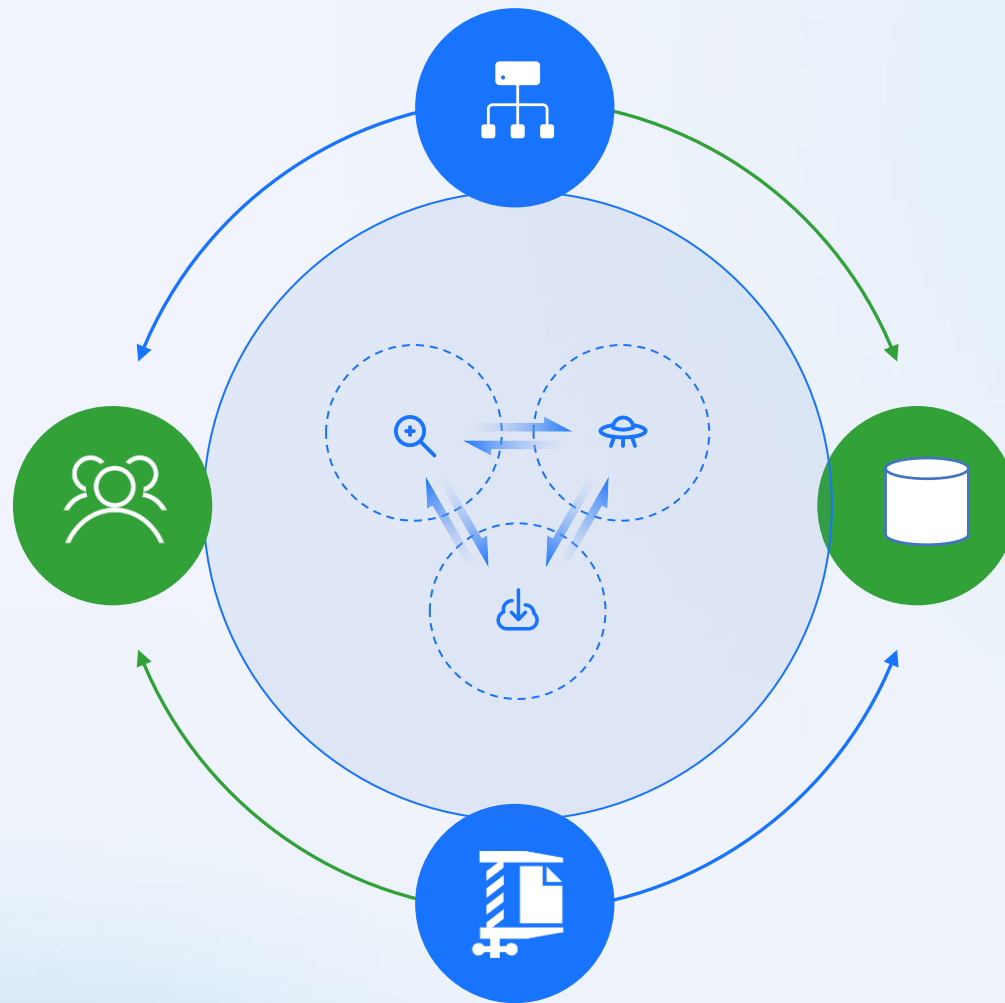
大规模分布式数据库、数据中间件等，帮助构建高性能大规模应用。数据库是核心的应用基础设施。

面向互联网规模的复杂应用

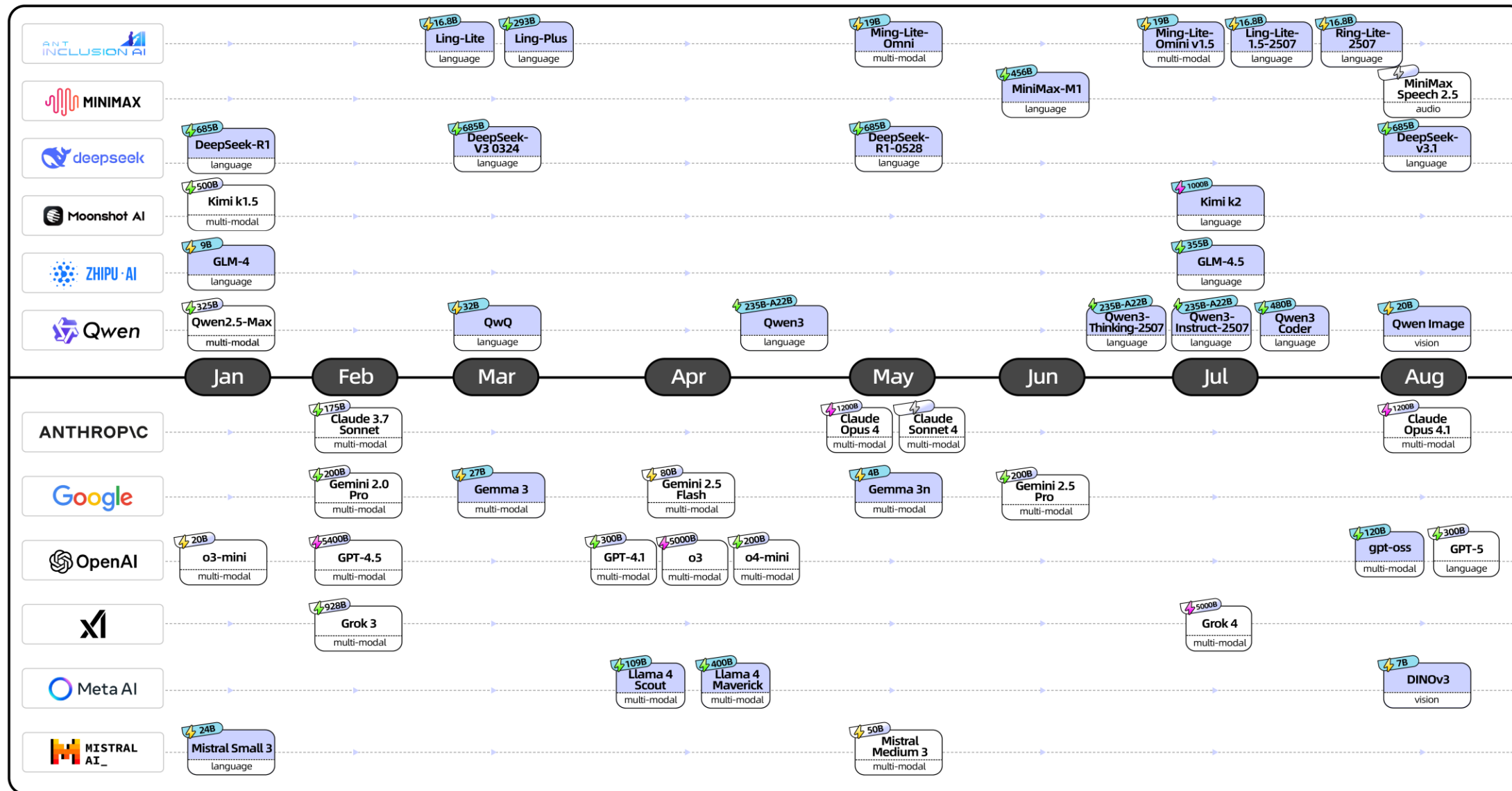
消息队列、服务发现、RPC、应用框架等（Spring），用于构建大规模并行应用；云原生时代，引入了 ServiceMesh（istio）、AppRuntime（dapr），将应用和应用基础设施解耦。

数据计算链路

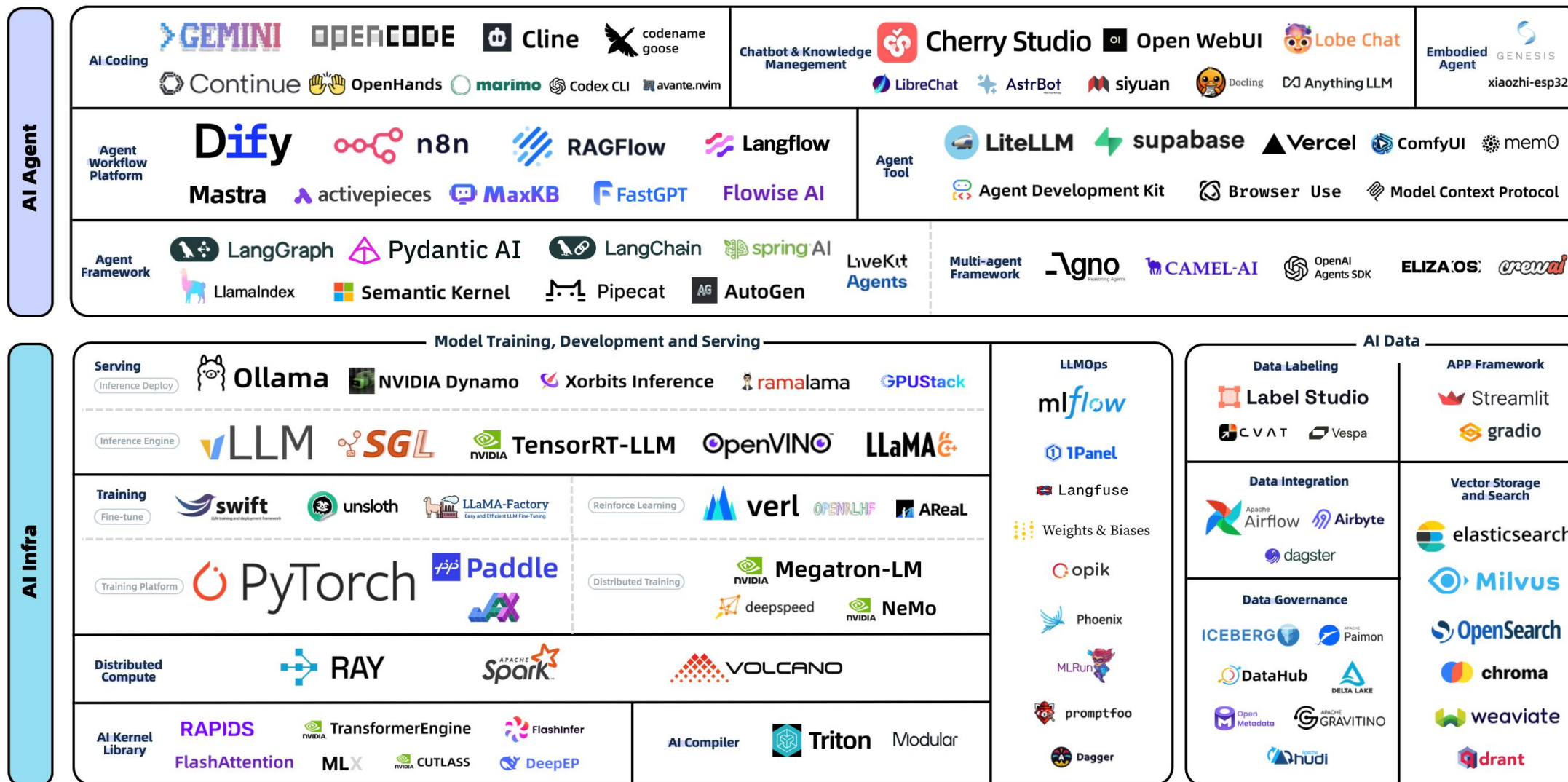
在线服务之外，服务过程产生的数据，会进入数据计算链路，进行离线或实时计算，构成数据驱动的应用。



Large Models Landscape 2025

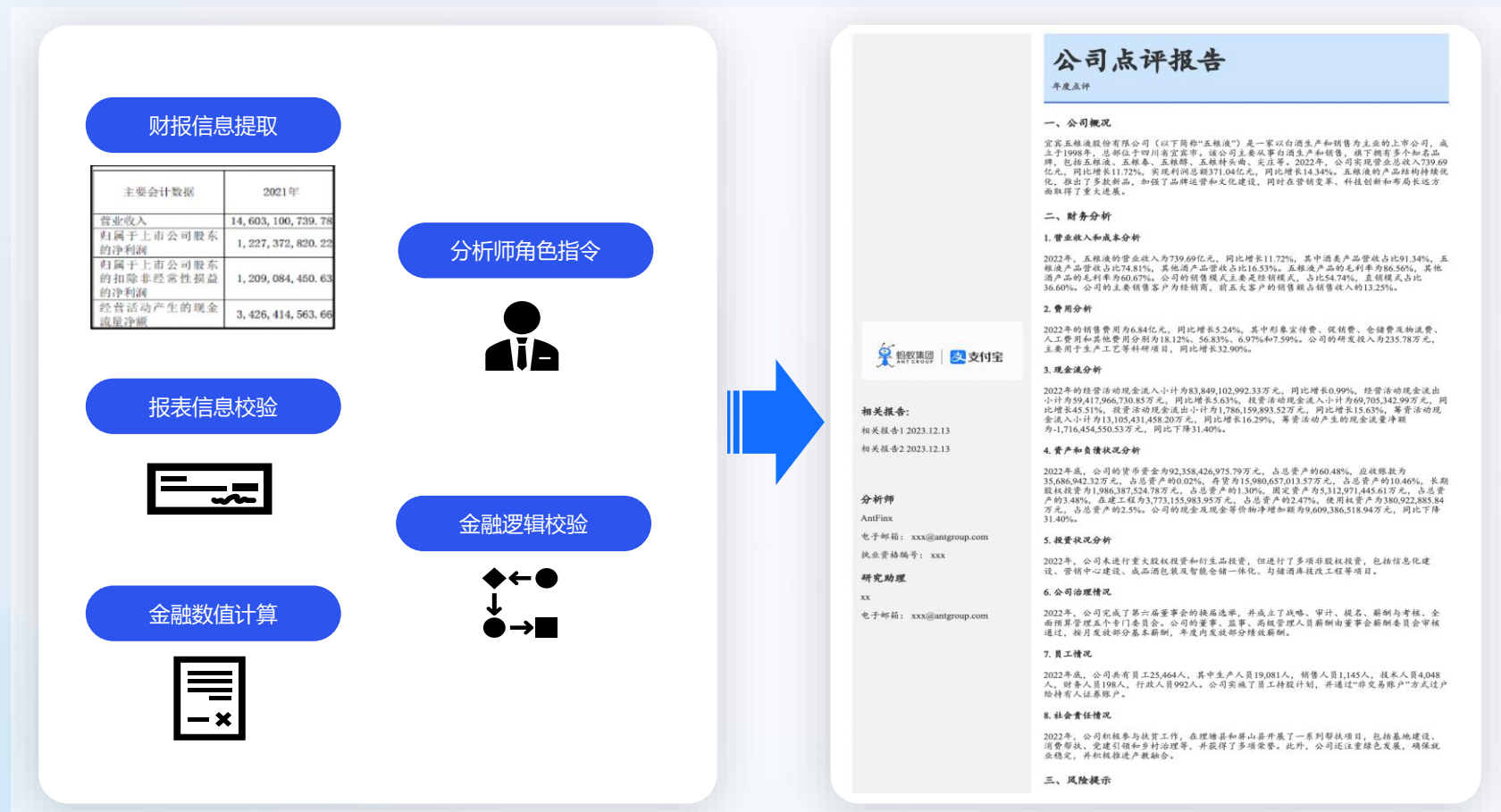


Open Source LLM Development Landscape



<https://github.com/antgroup/llm-oss-landscape>

生态位的变化：模型在成为新的数据库



👉 一个基于多 Agent 的财报解读应用（例中应用基于 AgentUniverse 框架构建）

越来越多的应用，使用模型生成的内容来服务用户，而不是基于数据库来生成内容。GenAI 时代的大模型正在充当 Web 2.0 时代的数据的生态位，这会对基础设施需要满足的需求，和它的架构带来很多变化。

03

模型为中心的基础设施

● 模型带来的架构演进：

后云原生时代的超级计算



并行计算再次伟大

大模型的训练和推理都高计算量都超越了节点边界，不论是训练，还是大参数量模型的推理，都需要跨越传统节点的边界



不再以CPU为唯一中心

GPU的算力密度更高，访问显存速度更快，更大量的计算和计算成本都在 GPU 上，GPU 的效率成为设计的核心考量



新的参数面网络

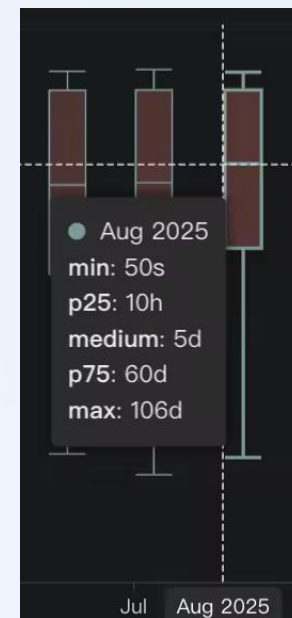
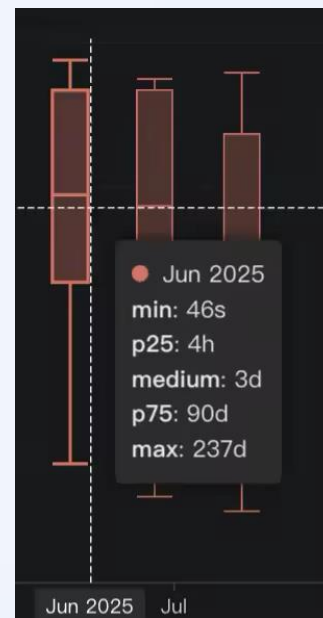
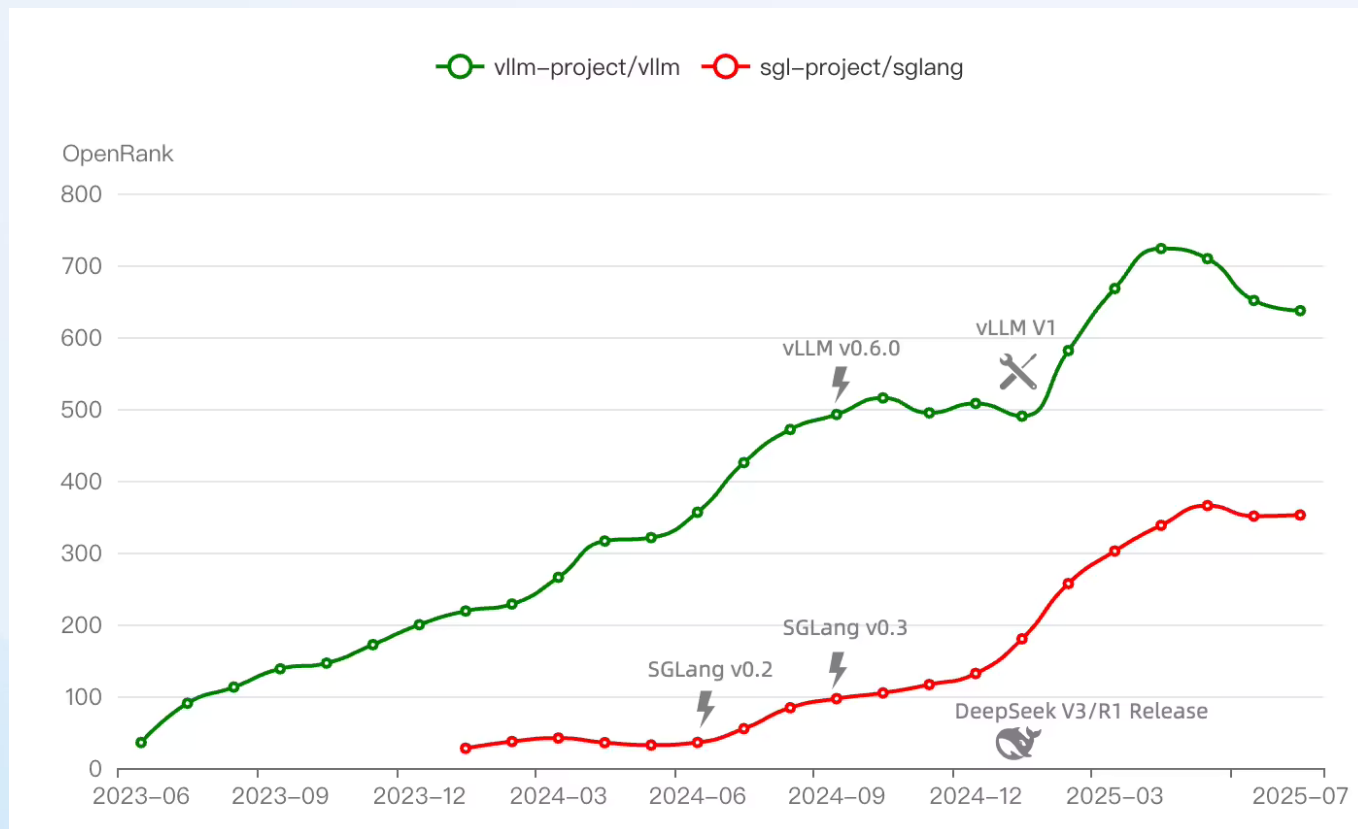
在计算、存储、管理网之外，引入了新的高速参数面网络，相应地，围绕参数面网络和 GPU 的优化成为了调度的重要约束



重新定义节点边界

通过参数面网络访问其他卡上的显存的时延，已经低于通过 PCIE 访问主存，什么才是本地，以及节点的边界，这个架构概念已经发生了动摇

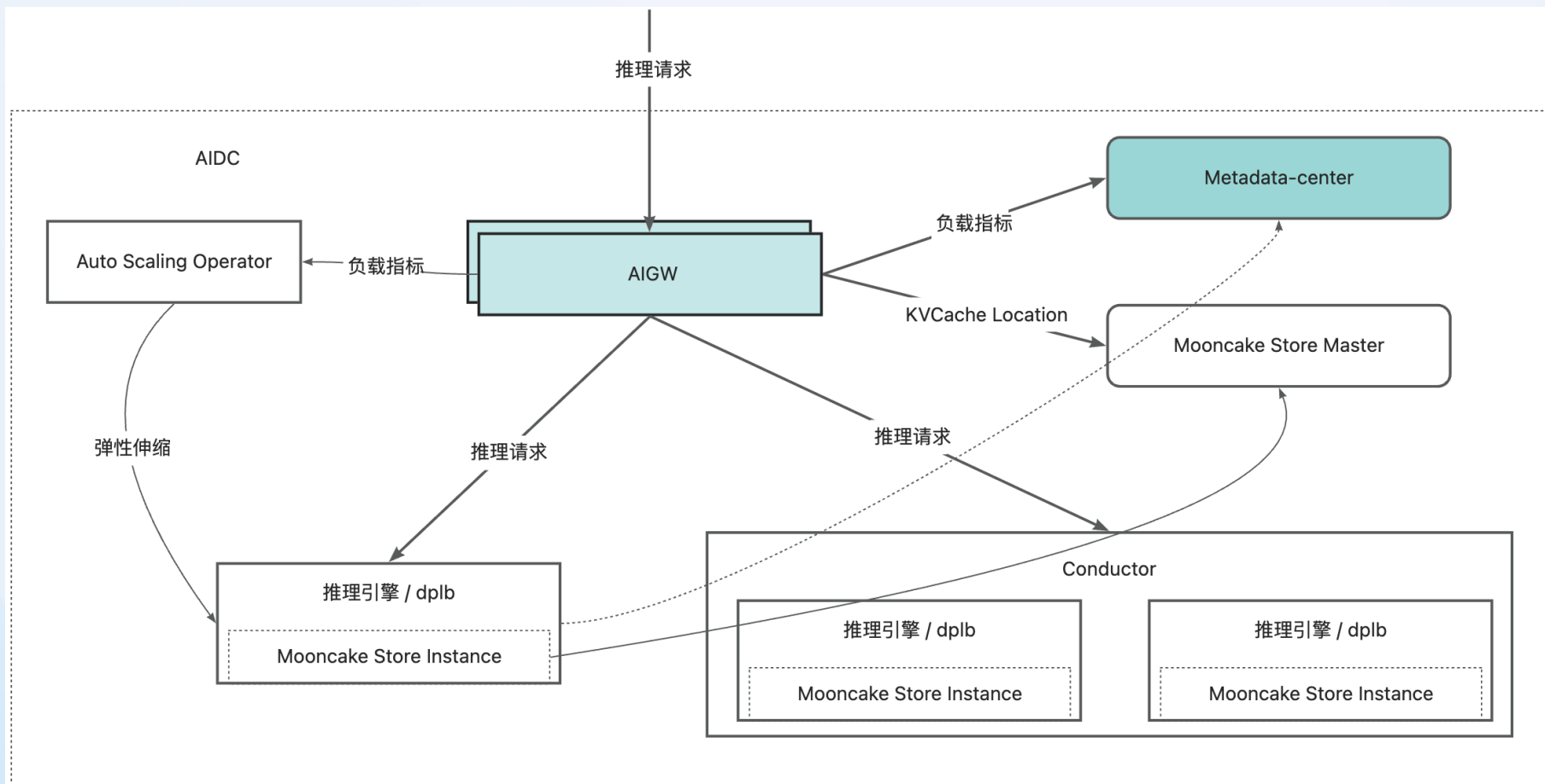
开源模型推理引擎 vLLM 和 SGLang 快速崛起



两个项目成长迅速，不仅性能竞赛一直在持续，而且都有上千开发者，开发速度都很快，大量的 Issue 会在几个小时内就得到响应。任何单一一家公司都很难跟上社区的脚步。



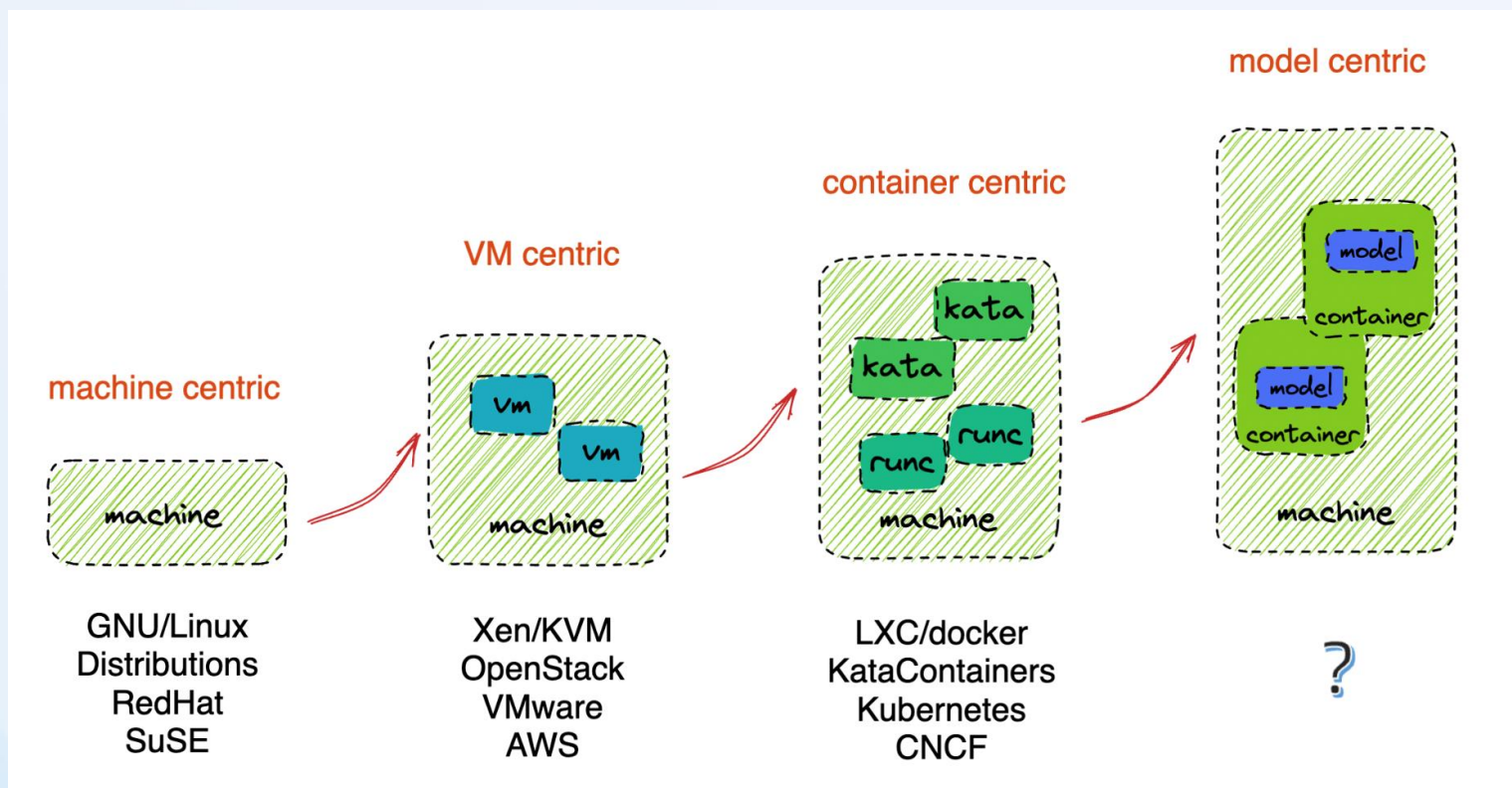
模型的优化部署与服务



模型的规模化服务，以开源的蚂蚁 AIGW + Mooncake 为例

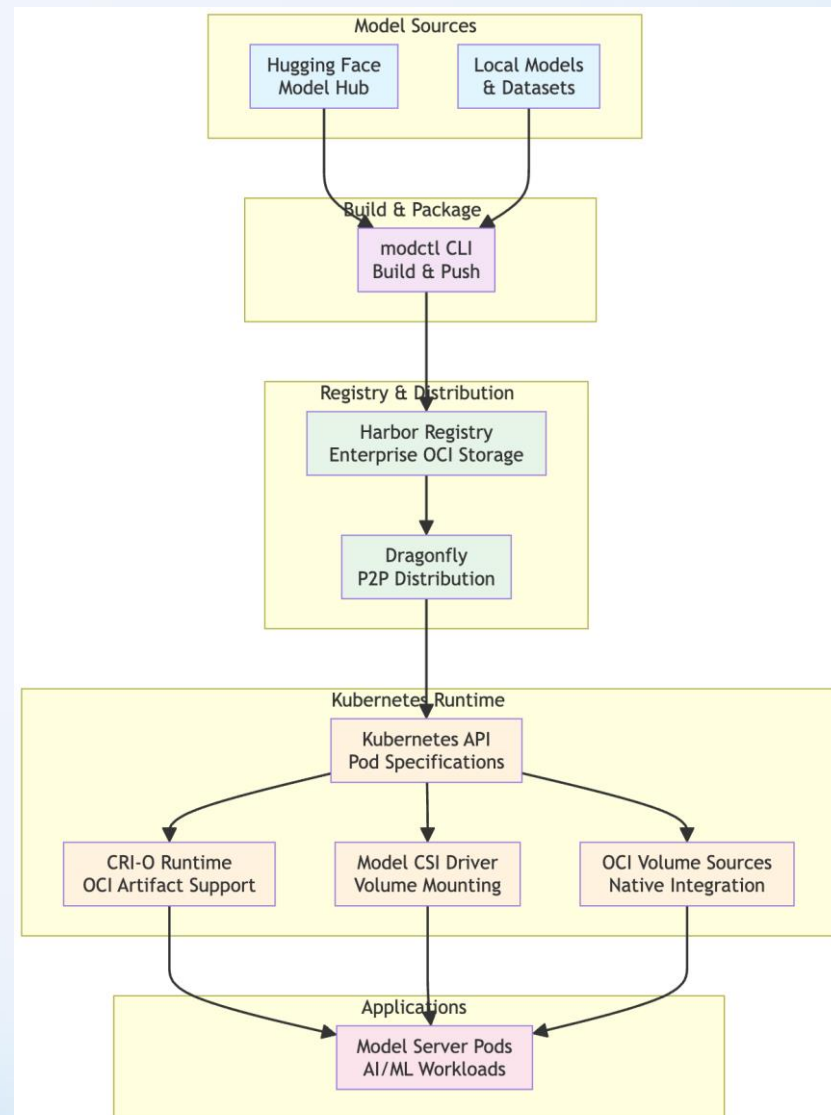
大模型应用的请求与数据库应用的模式不同，每个请求的计算消耗量大，且请求之间差异很大，需要整个基础设施的系统化适应和改进，而不仅是推理引擎本身的优化。LLM-d 等项目同样是在处理这个系统问题。

更进一步，从云原生到 AI 原生，从应用为中心到模型为中心



From CNCF Sandbox Project ModelPack (<https://github.com/modelpack/model-spec>)

借助已有的云原生生态，将开放接口、开放标准、不可变基础设施的理念，与大模型的生命周期管理与优化部署结合，帮助构建更高效的模型开发与服务生态

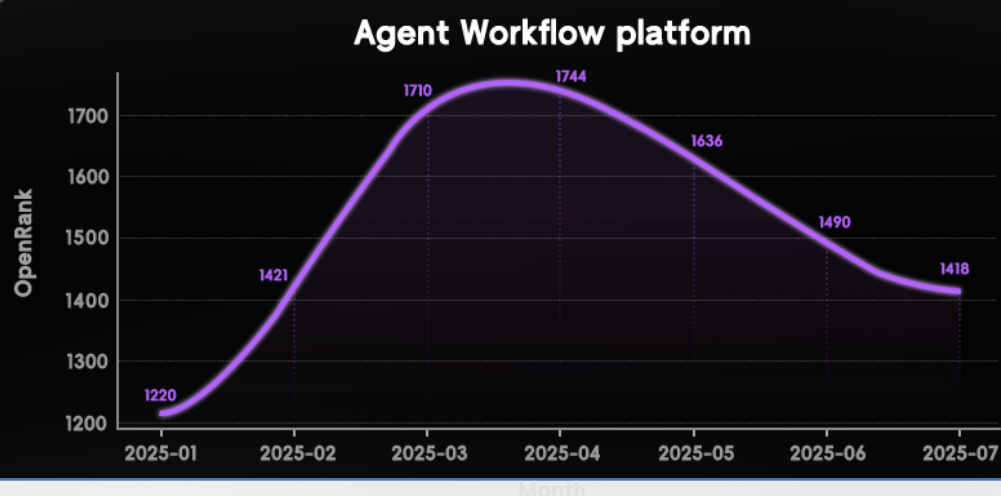
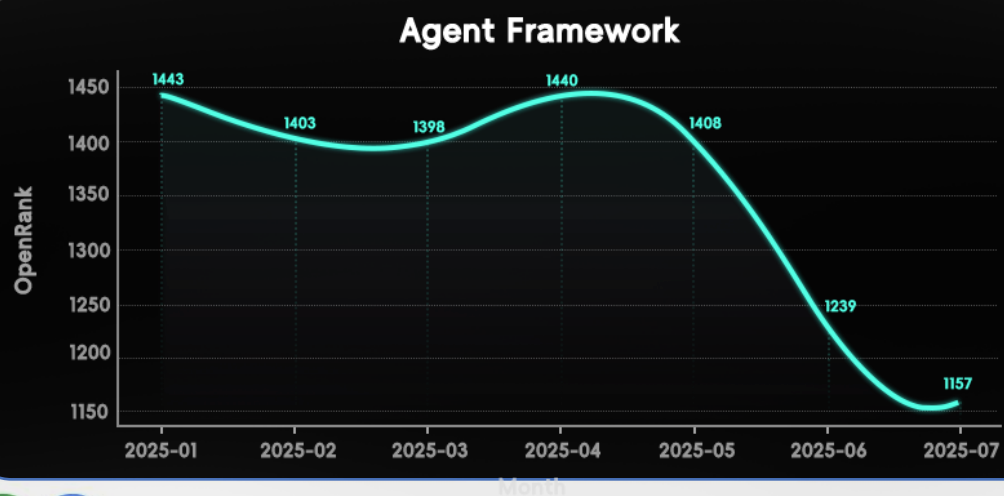
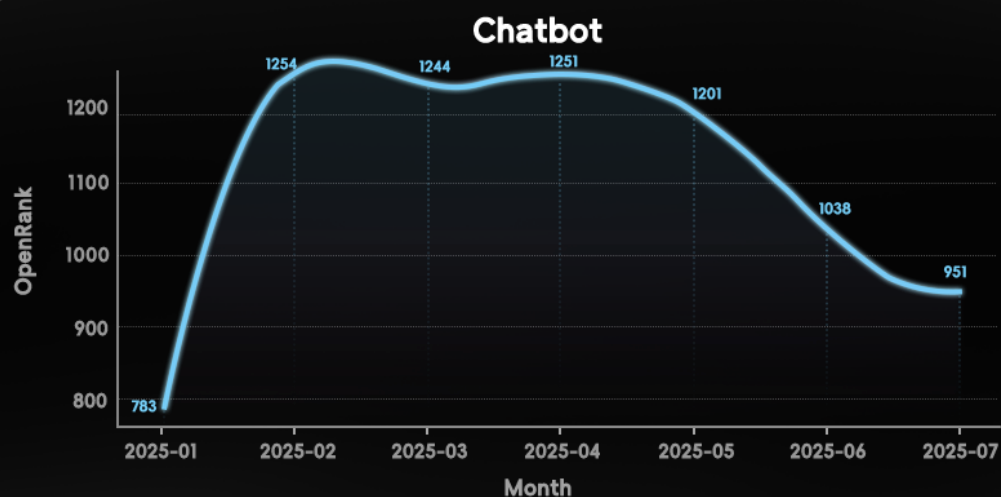
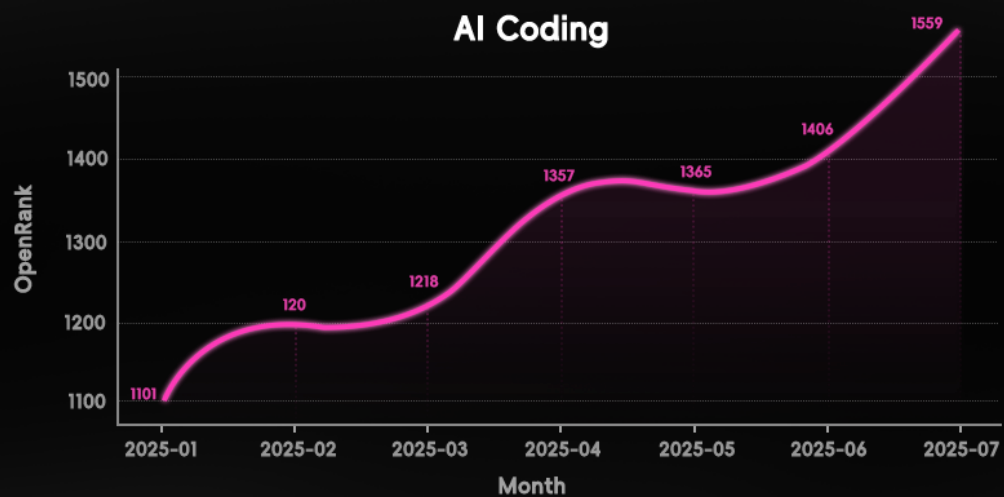


04

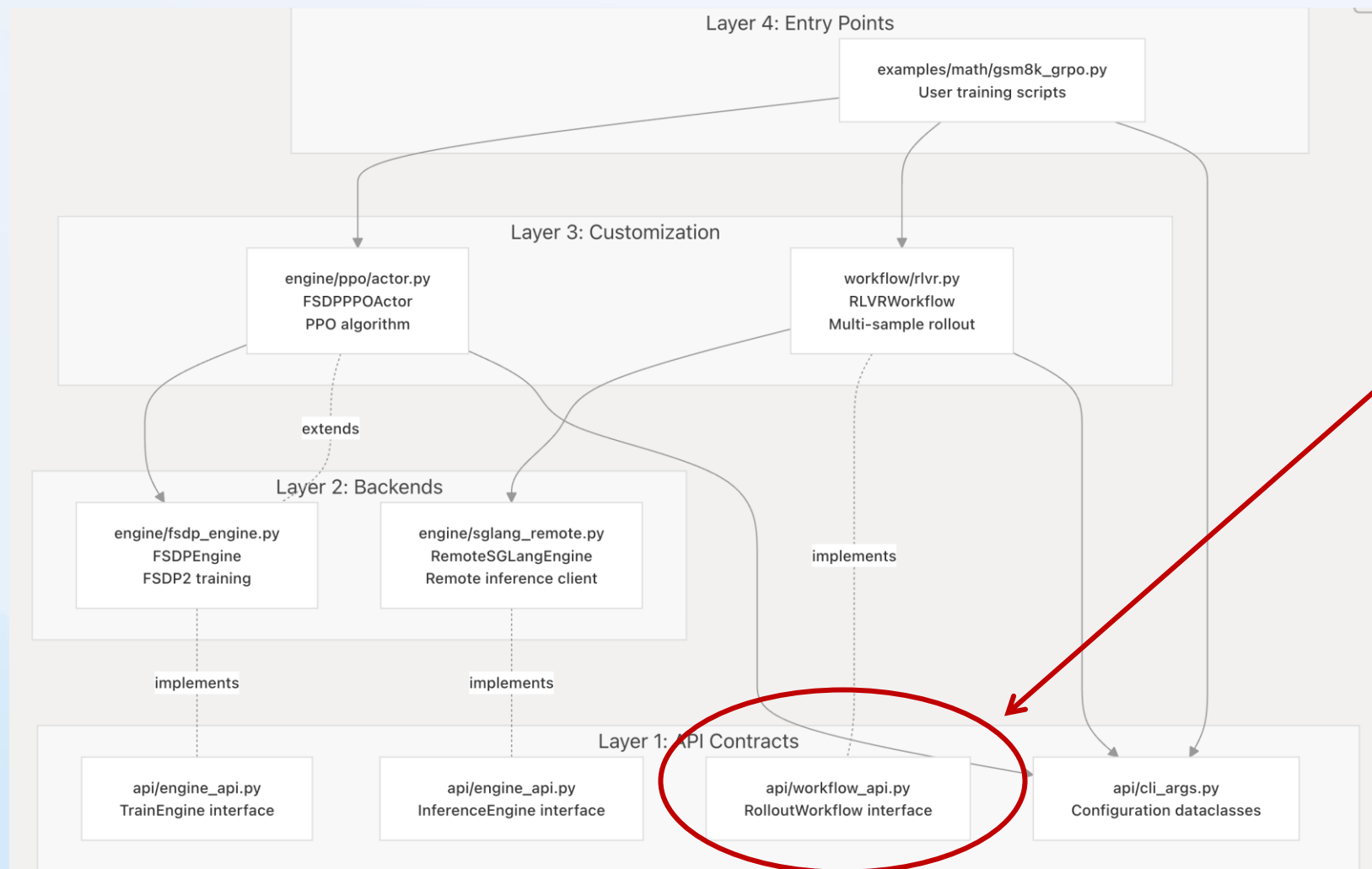
AI 应用的基础设施

或者说，Agent 基础设施

Agent 领域开源项目的活跃度趋势



同样，模型训练也在引入 Agentic RL



以 inclusionAI/AReaL 项目为例，作为一个强化学习框架，在 AReaLite 版本中增强了 Agentic RL，通过 Agent，来提升模型的泛化能力。

这样任意定义/引入的 Agent 代码，在训练过程中被执行，给基础设施也提出了更多需求

Agent Sandbox 与 Agent Tools

一切都是Agent，未来可能需要有海量的 Agent 和运行它们的 Sandbox，给已有的 Sandbox 技术和平台带来了新的挑战 and 机会



功能性

从简单的计算器，到搜索，再到浏览器、Computer Use，乃至执行任意代码，都可能是 Agent 或 Agent Tools 需要完成的功能



性能与并发量

对于 Agentic RL，可能需要同时并发执行大量的 Agent，并要求响应速度很快，镜像和容器加速项目都可能会被用到



Sandbox 的隔离性

因为 Agent 有可能执行任意获取的代码，或者是模型生成的代码，因此需要使用安全容器或其他安全的配置，来保护基础设施，免受意外或恶意的攻击，Kata Containers, Firecracker, gVisor, 都是被大量提到的沙箱技术。

05 展望

从“通算+智算”向通智一体的演进

从“通算+智算”向通智一体的演进

很多前 AI 时代的应用也将向 Agent 演进，Agent 也会做一些“普通”的工作

挑战：

- 「基础设施的稳定性」与「模型技术的快速变化」的矛盾，工程质量的保障需要被更有效地关注；
- 如上提到的 Agent 的任意执行可能，会引发更多的安全和滥用风险，需要谨慎地解决或防御这些问题

应对或机遇：

- 我们有机会在这样的变革来临时，和应用一起进步，构建更好的 CloudNative AI，或者是 AI-Native 架构；
- 我们已经有的沙箱技术可以帮我们应对安全威胁；
- 利用 AI 本身的发展，也可以帮助我们更好地、智能地应对这些风险和挑战。

📍 北京

QCon

全球软件开发大会

会议时间：4月10-12日

- 大模型赋能 AIOps
- 越挫越勇的大前端
- 云上业务架构演进
- 多模态大模型及应用
- 海外 AI 应用创新实践

📍 北京

AiCon

全球人工智能开发与应用大会

会议时间：6月27-28日

- 端侧智能
- AI Agent
- 多模态大模型
- 金融+大模型

📍 上海

QCon

全球软件开发大会

会议时间：10月23-25日

- AI Agent
- 大模型训练推理
- 端侧 AI
- 搜推深度融合
- 数智金融

4月

5月

6月

8月

10月

12月

📍 上海

AiCon

全球人工智能开发与应用大会

会议时间：5月23-24日

- 通用大模型
- AI Agent
- 垂直领域应用
- 模型可解释性

📍 深圳

AiCon

全球人工智能开发与应用大会

会议时间：8月22-23日

- 模型效率与部署
- 多模态大模型
- 大模型安全
- 智能硬件

📍 北京

AiCon

全球人工智能开发与应用大会

会议时间：12月19-20日

- 通用大模型
- 智能硬件
- LMOps
- 具身智能

极客邦科技 2025 年会议规划

促进软件开发及相关领域知识与创新的传播



参会咨询



查看会议

QCon
全球软件开发大会

THANKS

大模型正在重新定义软件（以及硬件）

Large Language Model Is Redefining The Software and More



Backup 一张 ModelPack

