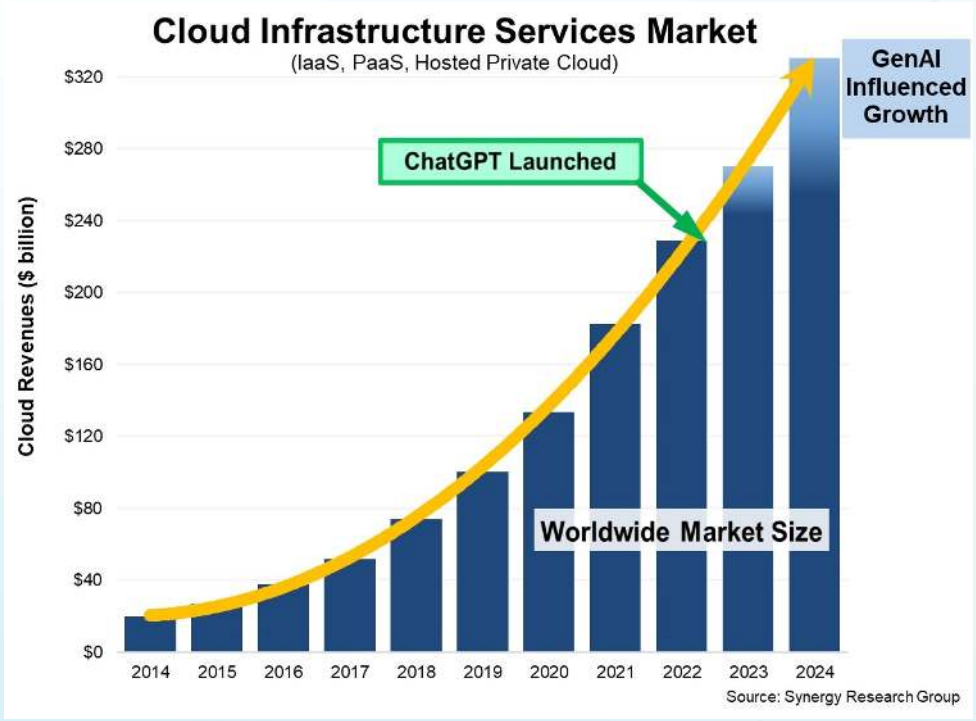
KUBERNETES COMMUNITY DAYS BEIJING 2025

面向GenAI时代的LOKAI基础设施挑战与实践

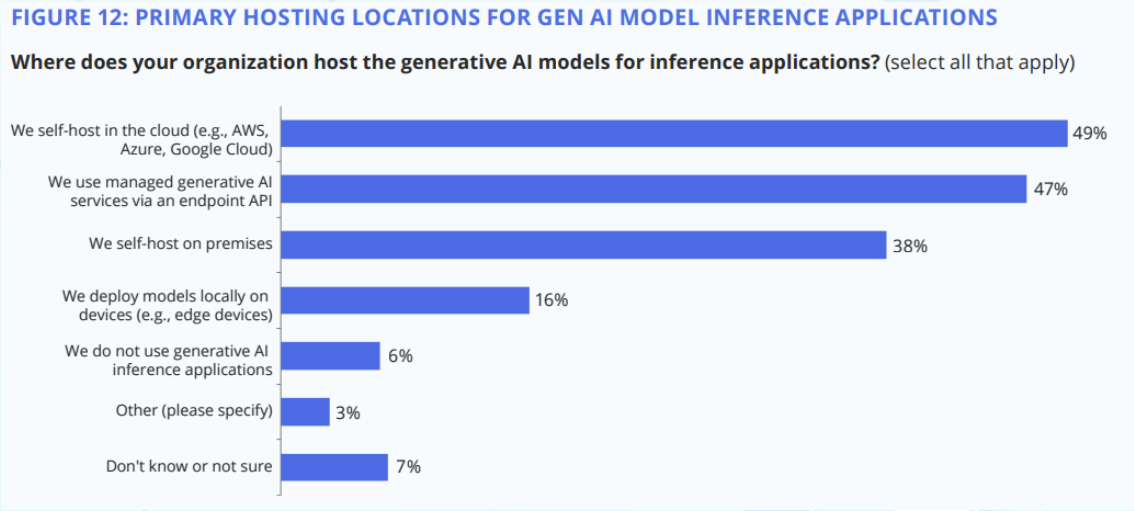
郑振宇 OpenAtom openEuler

GenAI 时代云基础设施的机遇与挑战



- 2024全年云基础设施花费达到$330 Billion,较2023年上涨22%，较2022年上涨44.7%
- ChatGPT于2022年底发布，与基础设施业务爆发式增长节奏相匹配
- 调查表明，超过半数基础设施增长源于GenAI业务

Source: Synergy data and analysis



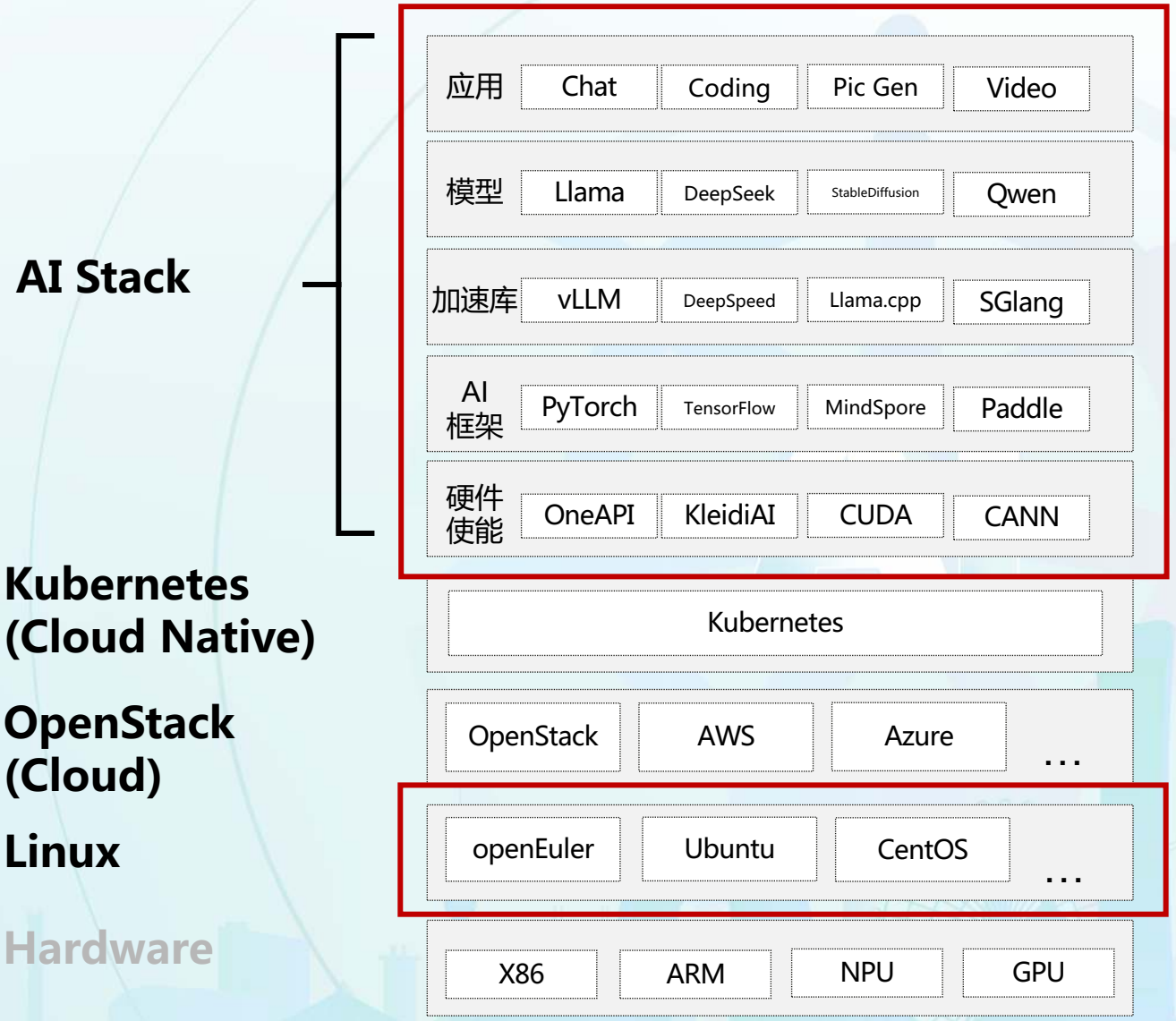
- 近半数GenAI业务自部署AI基础设施软件栈

Source: Linux Foundation Research

GenAI 时代云基础设施的机遇与挑战

核心挑战：

- **算力支持**：硬件种类繁多，GenAI时代大规模集群与异构融合成为刚需；
- **算力释放**：加速库与训、推工具链全量支持，硬件使能层释放多样性算力；
- **软件生态**：丰富的AI软件生态，加速库与训、推工具链全量支持，释放多样性算力；
- **集群能力**：集成云与云原生能力，打造坚实、灵活的算力底座；
- **快速响应**：AI软件栈日新月异，快速使能、快速集成、快速响应；



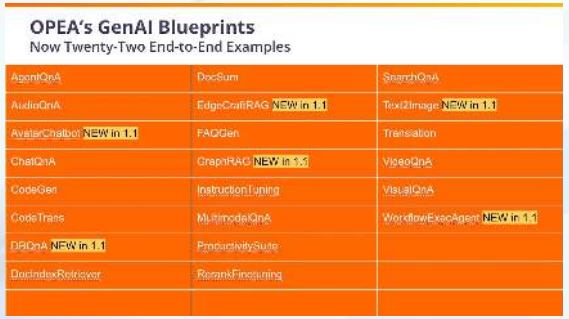
OPEA: AI应用E2E快速部署

- Open Platform for Enterprise AI
- LF AI&DATA 旗下开源项目
- 为用户提供端到端企业级GenAI应用部署方案：
  - 用于构建生成式人工智能解决方案的组件，包括检索增强：
  - 生成式人工智能模型 - 大型语言模型（LLM），大型视觉模型（LVM）等。
  - 系统组件 - 例如，嵌入模型；向量数据库；排序，提示处理等。
  - 用于构建AI代理和创建完整端到端生成式人工智能流程的组合能力
  - 用于微调、定制和优化的工具，包括数据中心/本地设置
  - 各种经过验证、准备就绪的端到端参考流程
- v1.2 Released in 2025.1.27
  - 22个企业级GenAI应用

https://opea.dev/

openEuler：多样性算力支持、易用稳定的Linux发行版

- **算力支持**：Arm/x86/RISC-V/GPU/NPU/DPU/Power/...算力全量支持；
- **算力释放**：主流加速库与工具链全栈使能核心加速库全量验证，支撑算力释放；
- **软件生态**：主流框架全部支持，海量模型开箱即用，使能应用开发；
- **集群能力**：使能异构算力融合，集群资源动态调整，节点间异构设备协同，推理场景性能相对提升20% ；
- **社区响应**：极具活力的开源操作系统社区；



- 22个E2E GenAI应用
- https://opea.dev/

OpenEuler

docker hub

- 100+官方容器镜像
- Docker Official Supported OSS
- hub.docker.com/u/openeuler

Azure aws

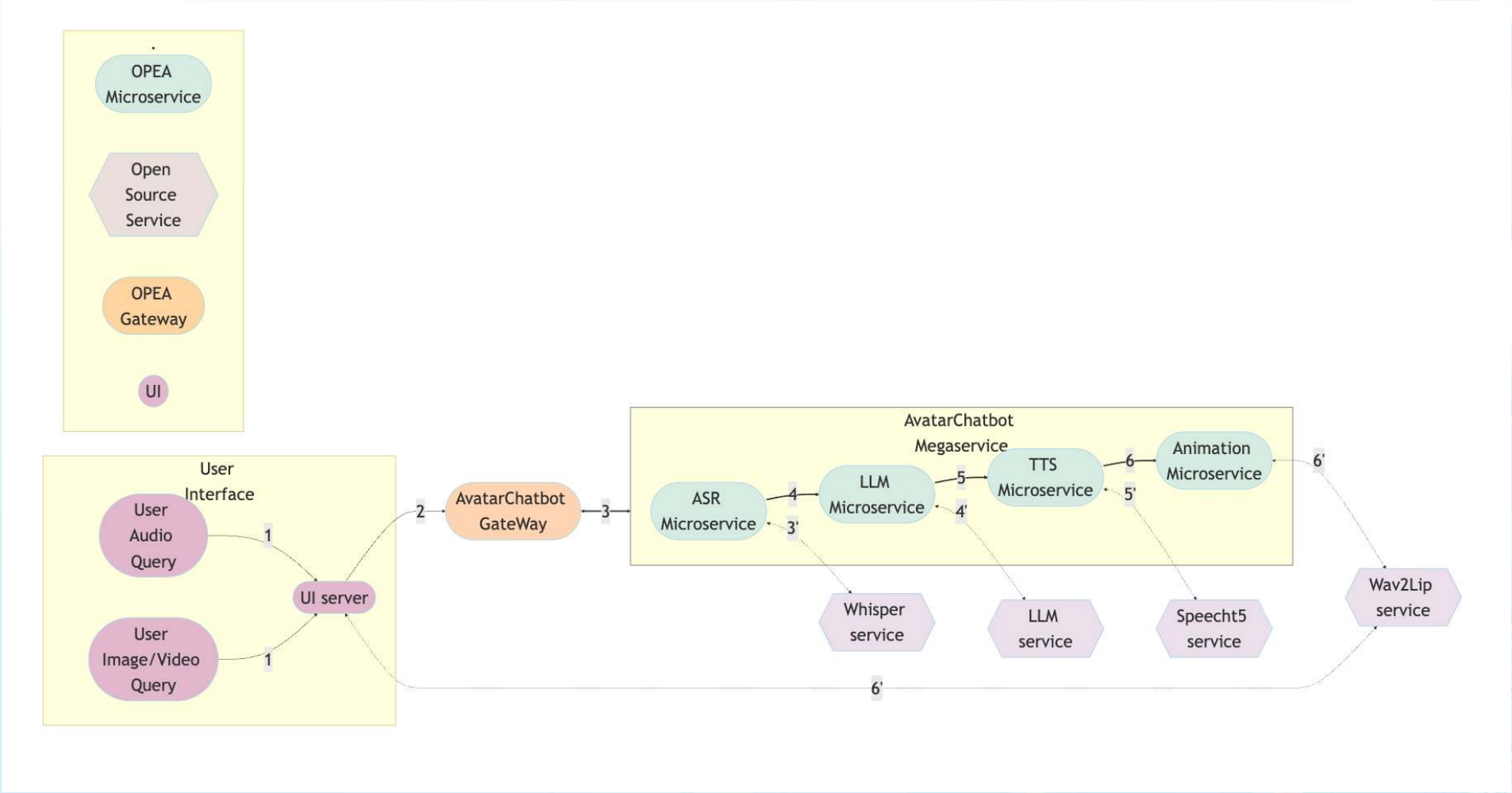
HUAWEI 腾讯云 阿里云

全球主流公有云镜像上线

Talk is cheap, show me the DEMO

AvatarChatbot

https://opea-project.github.io/latest/GenAIExamples/AvatarChatbot/README.html



Open Platform for Enterprise AI

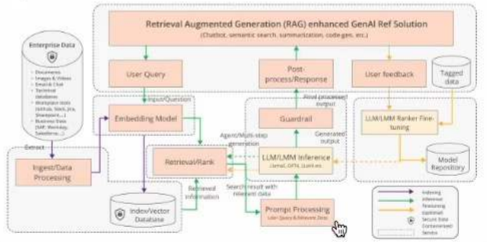
Sign up for our mailing list to stay up to date on OPEA!

Simplify development, production, & adoption of GenAI apps for enterprise

The OPEA platform includes:

- Detailed framework of composable building blocks for state-of-the-art generative AI systems including LLMs, data stores, and prompt engines
- Architectural blueprints of retrieval-augmented generative AI component stack structure and end-to-end workflows
- A four-step assessment for grading generative AI systems around performance, features, trustworthiness and enterprise-grade readiness

Pipeline Blueprint - RAG Flow



Efficient
Harnesses existing infrastructure, the AI accelerator or other hardware of your choosing.

Seamless
Integrates with enterprise software, with heterogeneous support and stability across system & network.

Open
Brings together best of breed innovations and is free from proprietary vendor lock-in.

Ubiquitous
Runs everywhere through a flexible architecture built for cloud, data center, edge and PC.

Trusted
Features a secure enterprise-ready pipeline and tools for responsibility, transparency, and traceability.

Scalable
Access to a vibrant ecosystem of partners to help build and scale your solution.

了解更多关于openEuler的内容

Thanks.