中国科学技术大学
University of Science and Technology of China

综述

LLM as OS, Agents as Apps:
Envisioning AIOS, Agents and the AIOS-Agent Ecosystem

University Of Science And Technology Of China

汇报人：马斌

2025.7.12

目 录

第一部分

Abstraction

- 论文摘要
- 主要创新

论文摘要

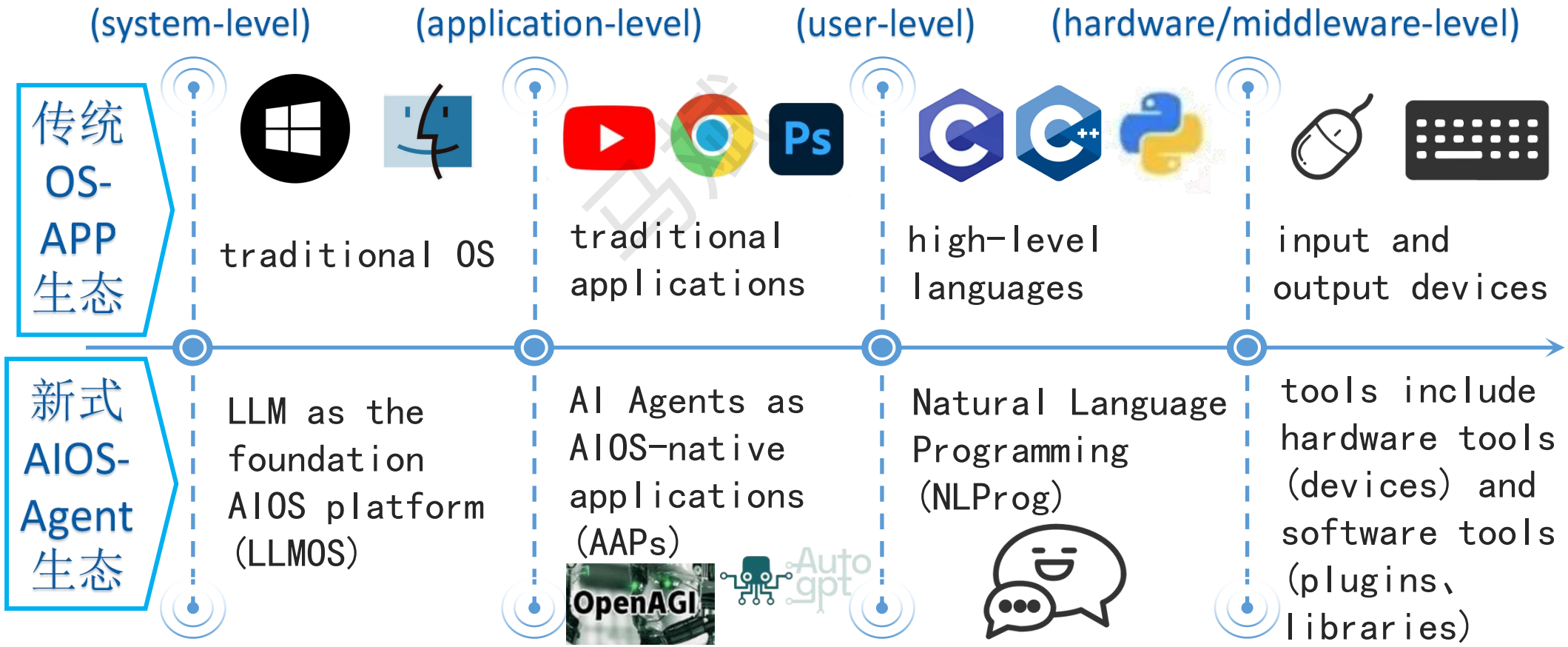
Figure 1: OS-APP ecosystem vs. AIOS-Agent ecosystem.

概述： 本文借鉴传统生态发展路径，构想了一个革命性的 AIOS-Agent 生态系统。

其以大语言模型（LLM）作为智能操作系统（AIOS），其上开发多种基于 LLM 的 AI 代理应用（AAPs），允许用户和开发者用自然语言开发代理应用，标志着从传统 OS-APP 生态的范式转变。

主要创新

创造性提出AIOS-Agent生态及其具体概念框架，可能成为一个跨层级革新路径或对传统(user-level)OS-APP生态发展进行演进。



第二部分

LLM & OS & AIOS

- 问题引入
- 传统OS/APP与AIOS/Agent的联系
- AIOS的具体架构

问题引入

- LLM本身是否具备构建AIOS的能力？
- 是。

1. 卓越的语言理解能力以及解决复杂任务的推理/规划能力；
2. 可处理各类自然语言，为构建 SDK 及应用提供高灵活性平台；
3. 支持自然语言交互，降低技术使用门槛；
4. 可通过学习交互过程，依据用户偏好和历史记录实现个性化体验。

传统OS/APP与AIOS/Agent的联系

架构角度

CPU 管理；Memory 管理；
Storage 管理；Device 管理；
SDK 和编程库

系统调用；
命令行界面(CLI);
图形用户界面(GUI).

开发 — OS SDK & 库;
发布&部署 — 平台.

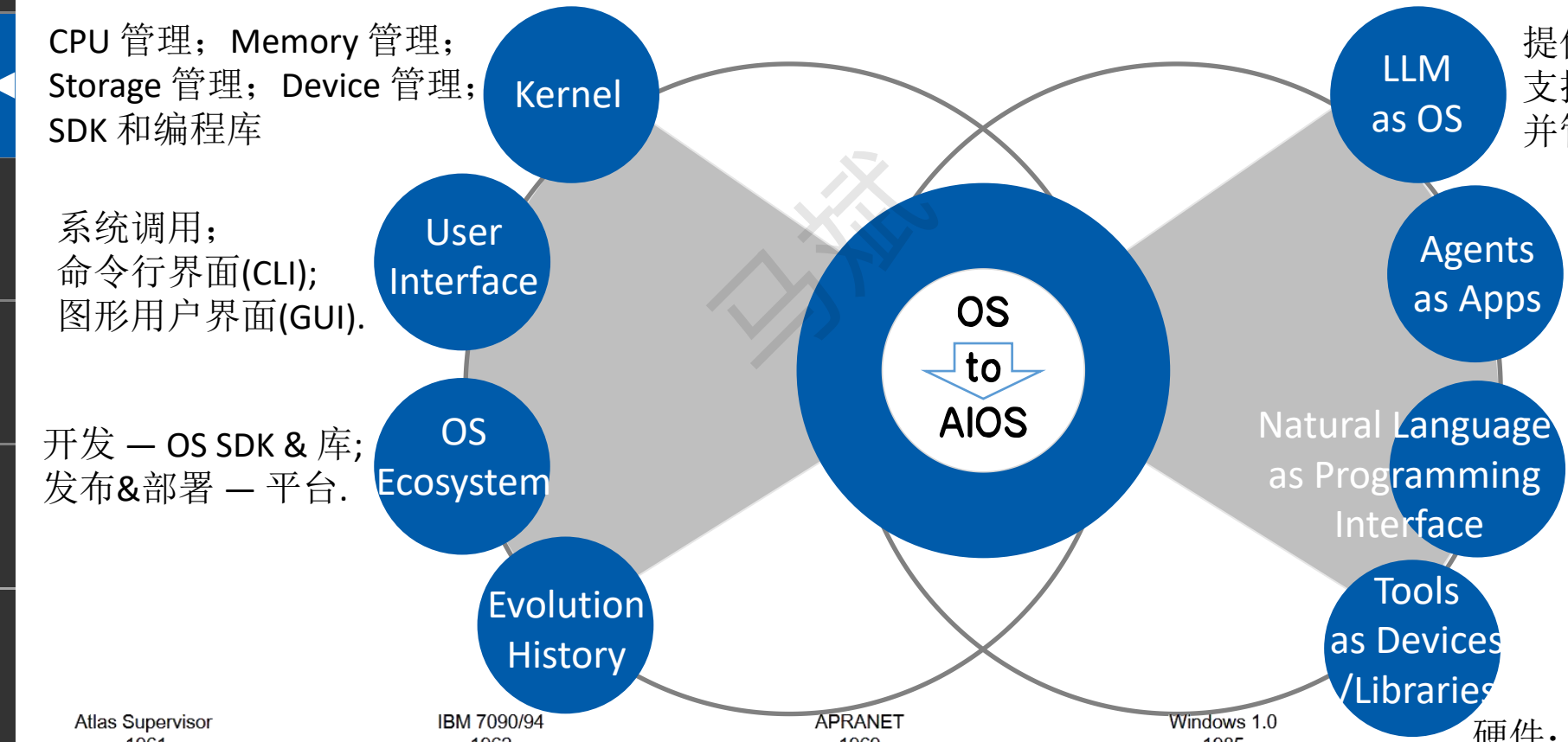
类比角度

提供计算、API 和服务；
支持各种应用程序
并管理各种计算资源

AI Agent
-(AIOS SDK)- >
AIOS-native app

二进制代码
->汇编语言
->高级语言
->自然语言

硬件：Devices->Tool-Drivers
软件：Libraries->Tool-APIs



AIOS的具体架构

LLM (as AIOS Kernel)

NEED:
支撑 AAPs.

AI:
Reasoning and Planning
(planning:Single-path,Multi-path,);
Self-Improving
(from Feedback, Examples)

Context Window (as Memory)

NEED:
定义范围;
可增加内存量

AI:
Efficient attention
(Sparse,Linear,QKV);
Position encoding
(RoPE等)

External Storage (as Files)

NEED:
提高长期内存保留率

AI:
不同格式存储
(Embedding Vectors, Plain-Text Documents, Structured Data);
检索数据
(检索和排名;严重依赖于现有IR方法)

Tools (as Devices/ Libraries)

NEED:
深入领域知识;与外部世界互动

Tools:
Software
(专家模型或 API);
Hardware;
Self-made(LLM设计)

第三部分

AIOS-Agent Ecosystem

该生态的创新点

该生态的创新点

Agents as Applications — More tailored

Natural Language Programming for Agents — More accessible

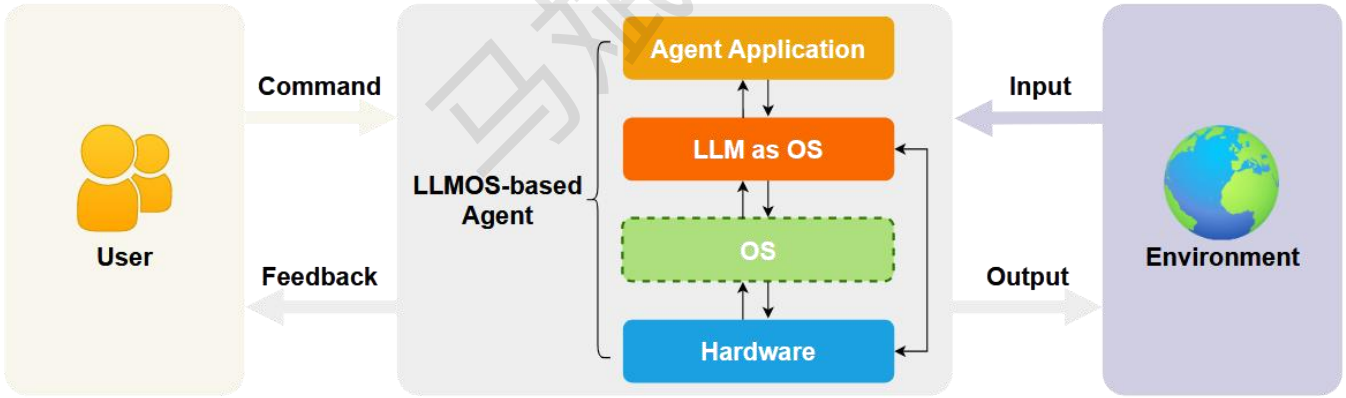


Figure 4: An illustration of LLMOS-based AI Agent.

总结：多Agent协作；可扩展性；自学习性

中国科学技术大学
University of Science and Technology of China

第四部分

LLMOS in Practice: AI Agents

- LLMOS现状及应用场景

LLLMOS现状及应用场景

现阶段基于 LLMOS 的代理，其实际部署场景多体现在Agent Apps。

Single Agent Apps

Physical: Scientific Research,Robotics,Autonomous Driving
Virtual/Digital: Coding,Web Service,Games,Recommendation

Multi-Agent Apps

Collaborative(合作), Adversarial(对抗),
此外，还有来自(H. Luo, et al., 2025.7.1)的Ensemble Integration(集成关系:分别输出，最后按权组合)

Human-Agent Apps

允许对代理进行更大的自定义和微调，促进人类用户和代理之间有意义交互。

中国科学技术大学
University of Science and Technology of China

第五部分

Inspiration and Conclusion

- 历史经验及未来发展方向
- 论文总结

历史经验及未来发展方向

01

Resource Management

内存管理：
上下文窗口扩展的限制；
工具管理：
系统社区，版本控制系统

02

Communication

域特定语言 (DSL)：
平衡操作系统和用户，增强multi-AIOS通信

03

Security

系统漏洞后果：
系统崩溃，诈骗，资源窃取
检测和捕获：
静态分析，模糊测试

论文总结

本文提出 AIOS-Agent 生态系统的未来新愿景，借鉴于又区别于传统 OS-APP 生态系统。
该生态系统是技术交互方式的根本性变革，将 LLM 置于系统层、智能体（Agents）作为应用、工具（Tools）作为设备 / 库、自然语言作为编程接口。
这一范式转变有望推动软件开发和使用的民主化，实现NLProg，打破传统生态中专业编程技能的限制。

Comment：全面，大胆，浅显。

近期进度

论文阅读：

[1] LUO H, LIU Y, ZHANG R, 等. Toward Edge General Intelligence with Multiple-Large Language Model (Multi-LLM): Architecture, Trust, and Orchestration[EB/OL]. arXiv, 2025[2025-07-10]. http://arxiv.org/abs/2507.00672.
[2] PAN J, LI G. A Survey of LLM Inference Systems[EB/OL]. arXiv, 2025[2025-07-10]. http://arxiv.org/abs/2506.21901.
[3] GE Y, REN Y, HUA W, 等. LLM as OS, Agents as Apps: Envisioning AIOS, Agents and the AIOS-Agent Ecosystem[EB/OL]. arXiv, 2023[2025-07-11]. http://arxiv.org/abs/2312.03815.

知识点补充：

1. GPT系列、Llama系列的基本结构及其原始文章的abstract
2. Transformer的整体架构初探
3. 概念理解：QKV 矩阵，API调用，RoPE 旋转位置编码

题外话：发现在不同论文中，正文部分对于参考文献的引用话术有一定的规律，可以考虑利用编程设计一个辅助器，实现仅输入“文献”和“概要”，输出“正文内容”，使论文书写过程中作者仅需关注论文的想法本身。

中国科学技术大学
University of Science and Technology of China

Thanks

感 谢 老 师 的 观 看

University Of Science And Technology Of China

汇报人：马斌

2025. 7. 12