构建AGI时代的推理基础设施

单一舟 华为云

科技生态圈峰会 + 深度研习

——1000 + 技术团队的共同选择

KEYLINK ing

K+峰会

上海站
K+全球软件研发行业创新峰会
时间：2024.06.21-22

敦煌站
K+思考周®研习社
时间：2024.10.17-19

香港站
K+思考周®研习社
时间：2024.11.10-12

K+峰会详情

AiDD峰会

上海站
AI+研发数字峰会
时间：2024.05.17-18

北京站
AI+研发数字峰会
时间：2024.08.16-17

深圳站
AI+研发数字峰会
时间：2024.11.08-09

AiDD峰会详情

演讲嘉宾

单一舟

华为云架构与技术创新部架构师

华为云架设部架构师，博士毕业于加州大学圣地亚哥分校. 研究方向围绕提升数据中心基础设施性价比，包括大模型推理系统，分离式内存，大规模分布式存储系统等. 在华为云主导云存储硬件卸载项目和Serverless大模型推理项目. 在顶级学术会议发表论文20+，研究曾获得OSDI 2018, SYSTOR 2019, FPAG 2024 Runner Up最佳论文.

关于通用人工智能的两个”事实”

“1000个哈姆雷特”

Our mission is to ensure that artificial general intelligence—AI systems that are generally smarter than humans—benefits all of humanity.

Andrew Ng ✓
@AndrewYNg

The definition of AGI I use is "AI that can perform any intellectual task that a human can."

Yann LeCun: "This notion of Artificial General Intelligence (AGI) is complete nonsense because human intelligence is very, very specialized." This is GOLD!!

“快了”

r/singularity • 9 mo. ago
lost_in_trepidation

Andrew Ng still thinks AGI is decades (30-50+ years) away

certainty," that AGI is somewhere between 5 and 20 years from now. His previous prediction was 30-50 years.

AGI this Decade: Demis Hassabis

Technology

Tesla's Musk predicts AI will be smarter than the smartest human next year

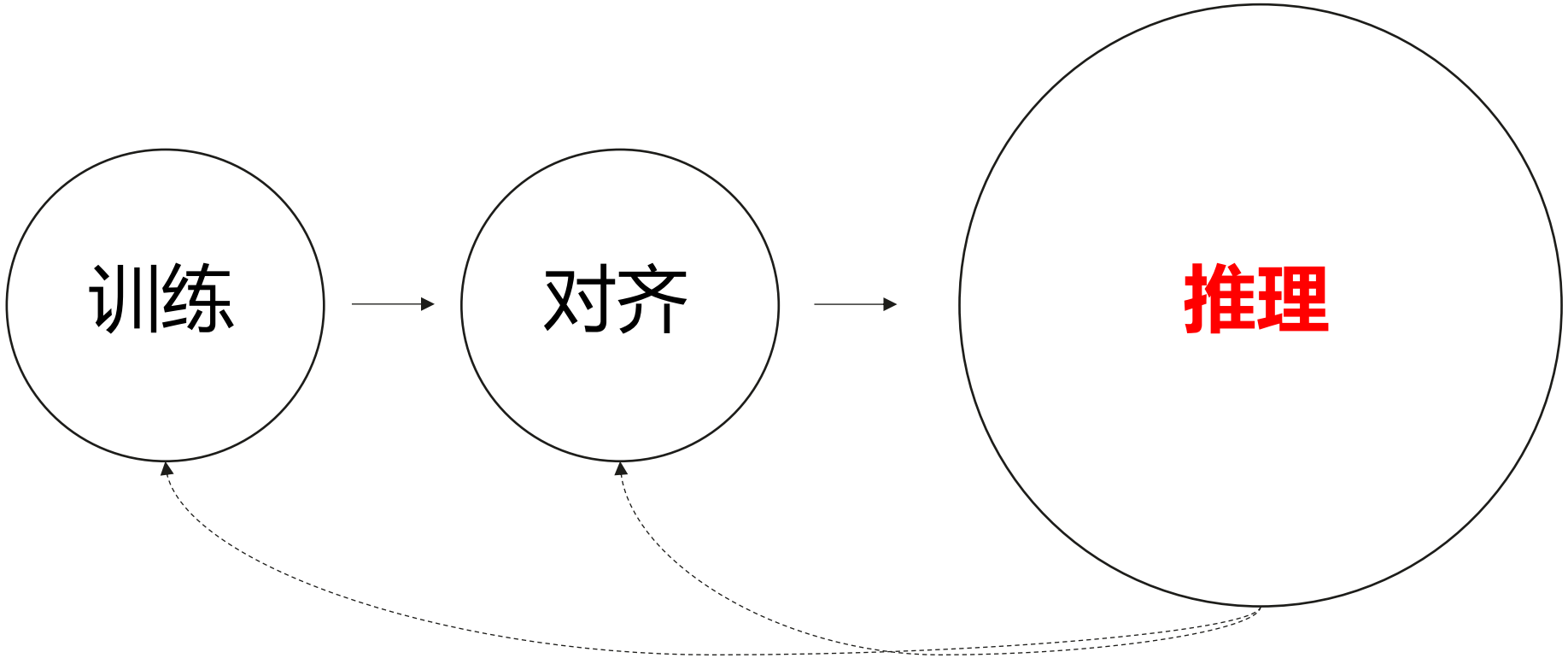
By Reuters

April 8, 2024 8:09 PM UTC · Updated ago

走向AGI的三个简单步骤

训练出高质量模型 是开始

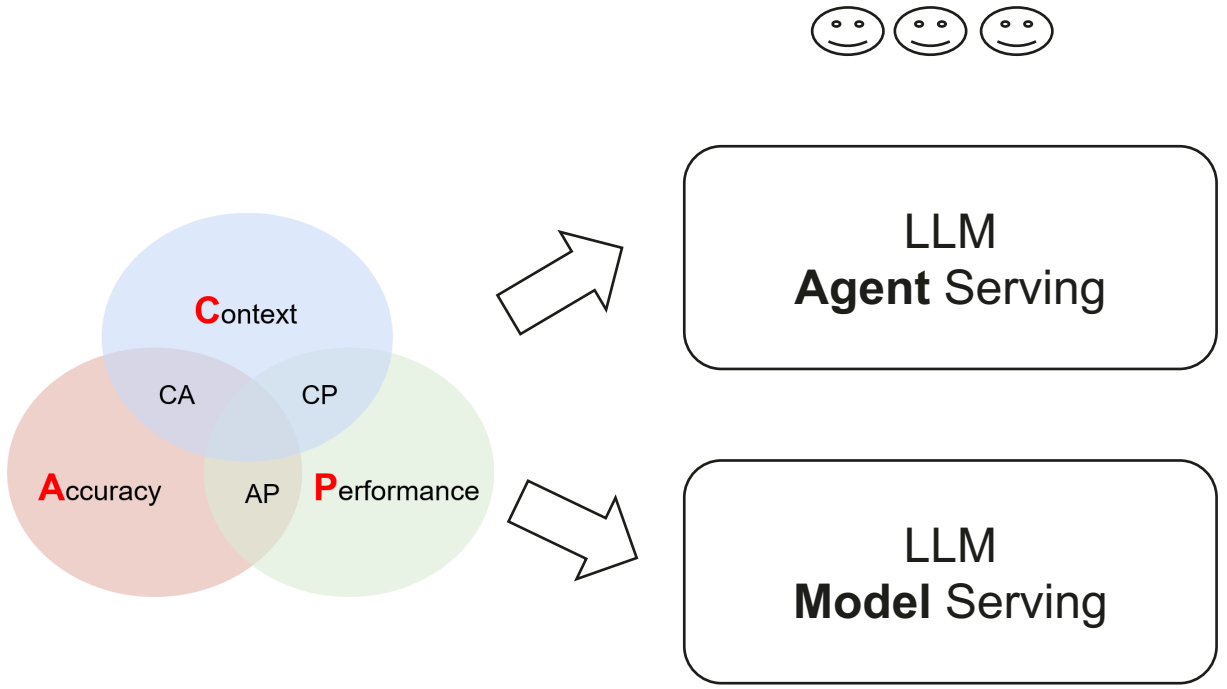
如何大规模低成本高性能推理是 关键



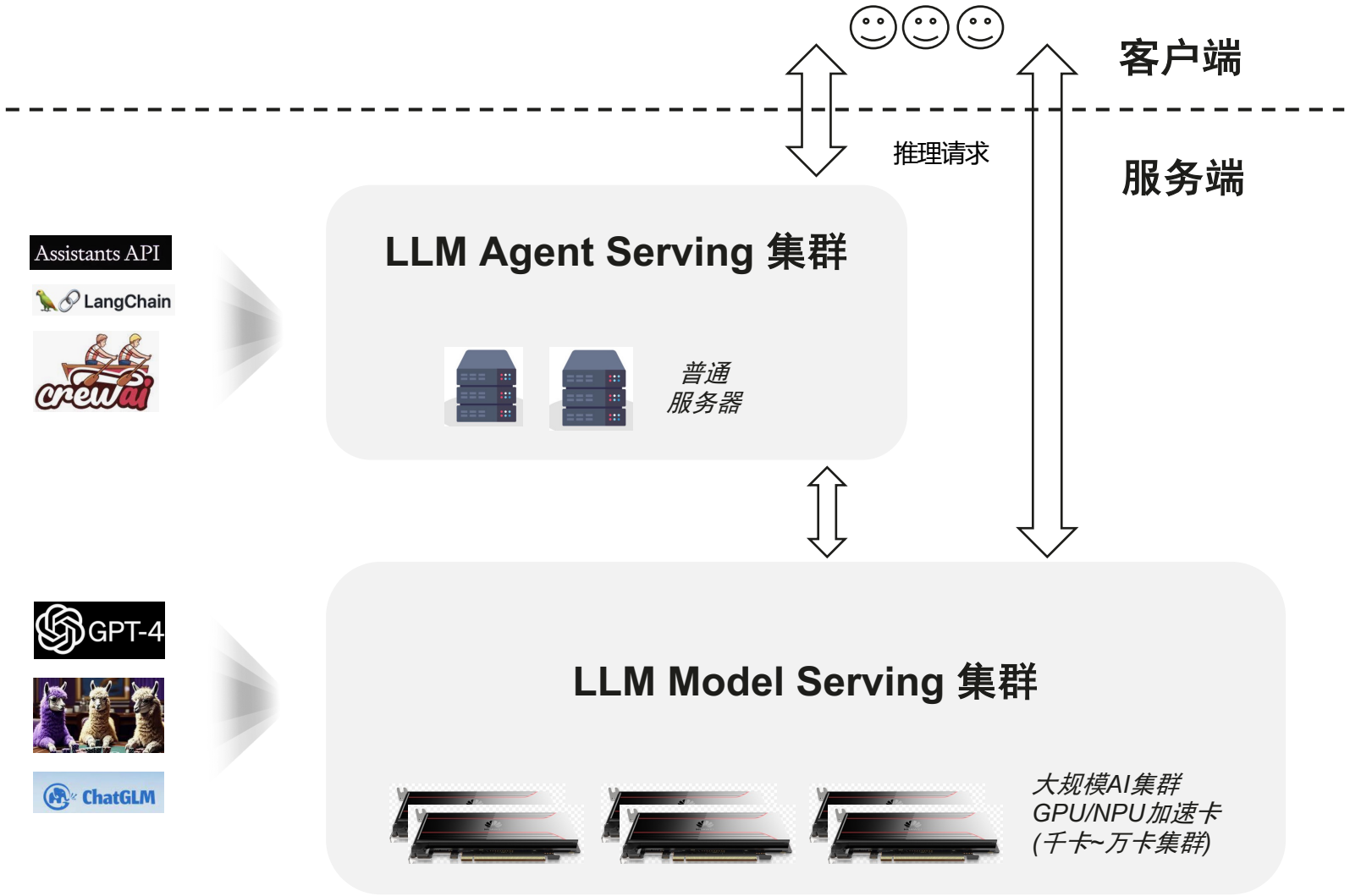
本次Talk重点

Outline

1. 整体介绍 / CAP理论
2. LLM Model Serving
3. LLM Agent Serving
4. 总结



数据中心如何部署推理系统?

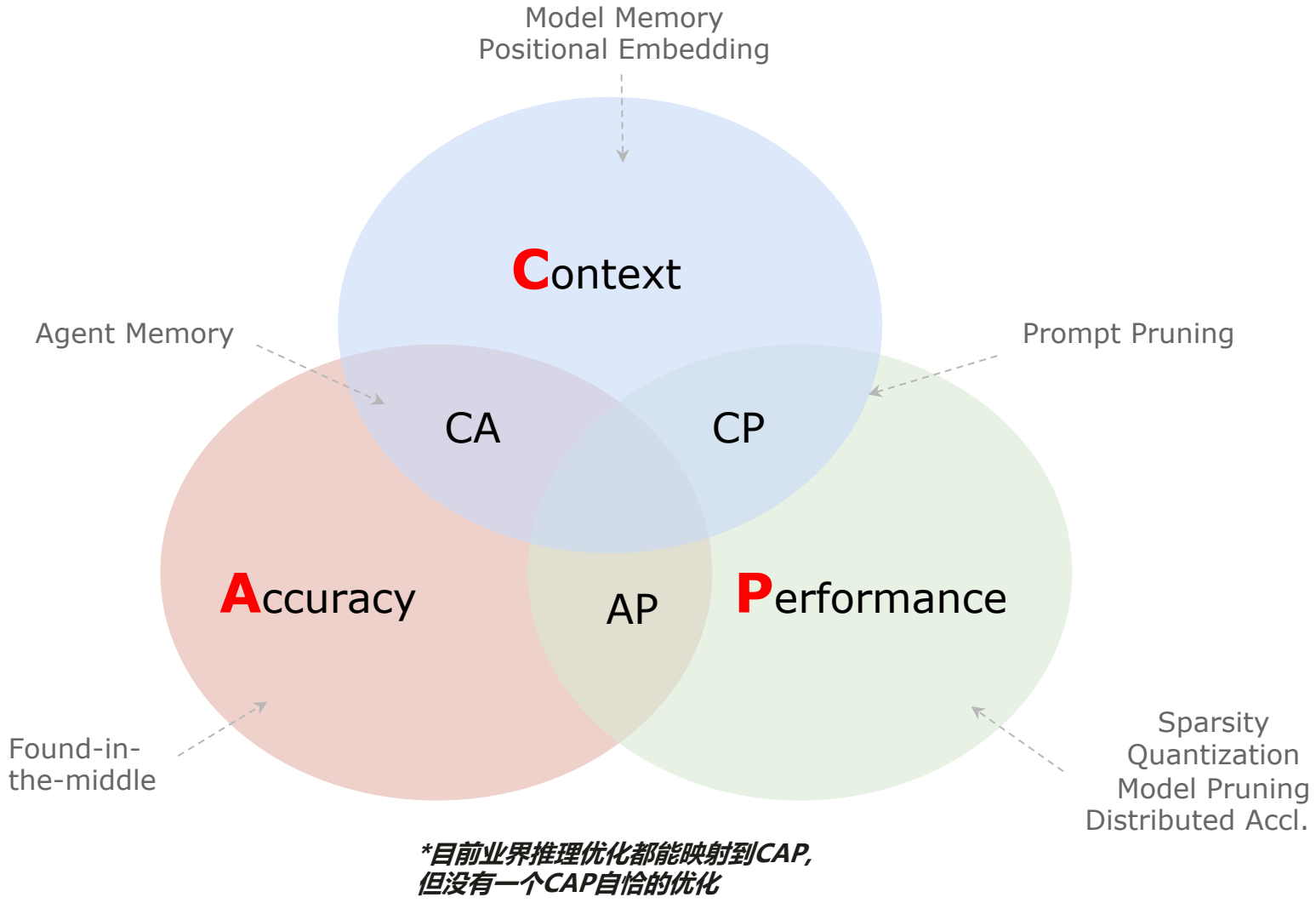


推理系统有什么挑战？我们总结为 CAP Principle

最大的挑战: 推理系统要优化三个目标,

但这三个目标冲突不自恰!

长度 (Context)
精度 (Accuracy)
性能 (Perf)

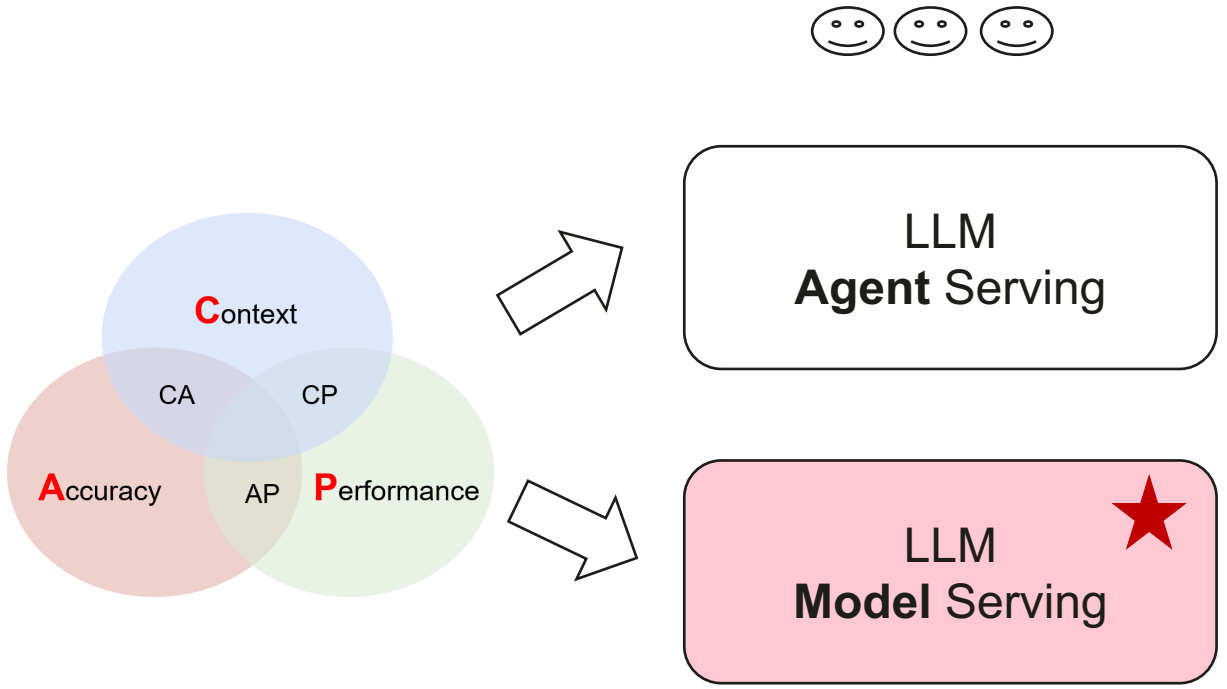


*受经典的分布式系统CAP原则启发

***目前业界推理优化都能映射到CAP,
但没有一个CAP自恰的优化***

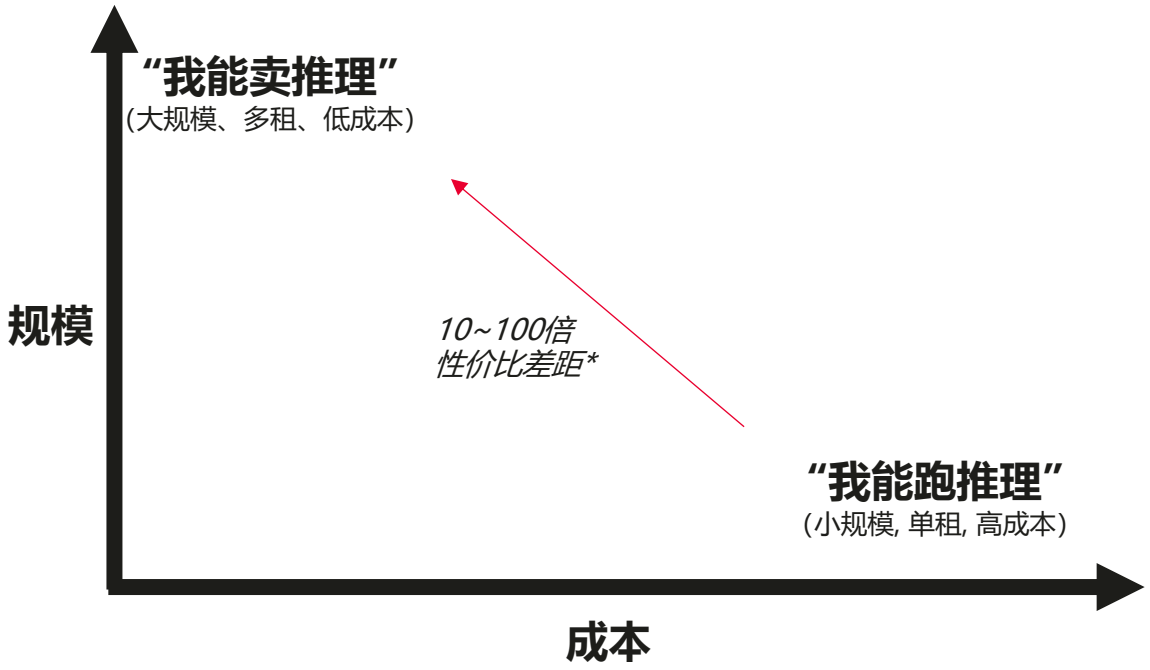
Outline

1. 介绍
2. **LLM Model Serving**
3. **LLM Agent Serving**
4. **未来**



LLM Model Serving 核心指标是什么? 性价比

从能跑到能卖有GAP



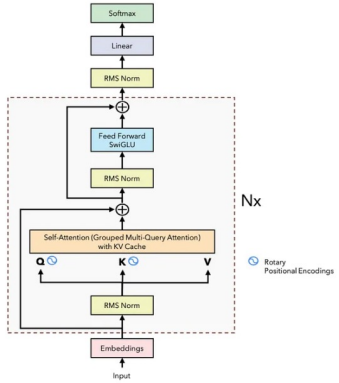
性价比是推理的核心竞争力

https://www.deepseek.com/

GPT-4	GPT-4o	GPT-3.5
$60 /MToken	$15 /MToken	$1.5 /MToken

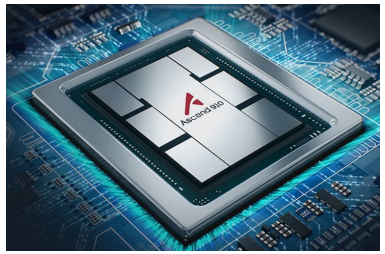
提升Model Serving性价比挑战是什么？模型架构与芯片架构的冲突

Transformer大模型



+

AI 芯片

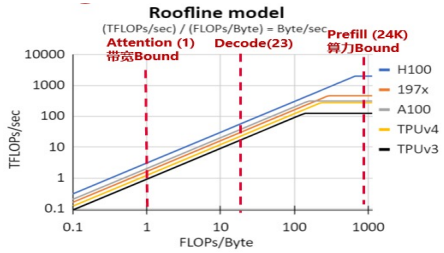


大模型
全量计算
+
芯片存算比
↓
算力和带宽不匹配

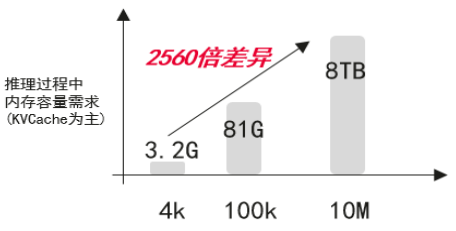
大模型
增量计算
+
芯片容算比
↓
算力和容量不匹配

大模型
组合计算
+
芯片连算比
↓
算力和网络不匹配

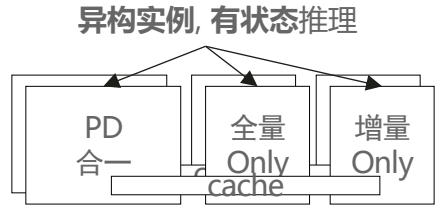
1 算力瓶颈



2 内存瓶颈



3 调度瓶颈



以存代算，缓解大模型全量推理的算力瓶颈，降低首字时延，提升用户体验

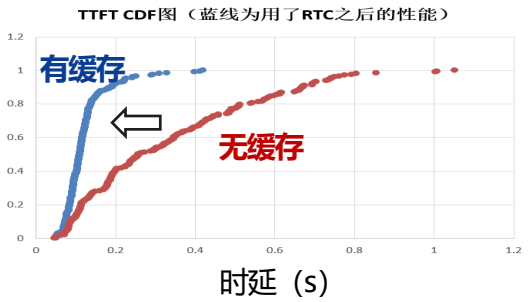
介绍

解决的问题: 在许多应用场景, 用户Prompt存在大量重复前缀, 因此大模型全量计算时会存在大量的重复计算, 导致首字时延高(time-to-firs-token).

解决的思路: 将推理过程中产生的临时数据 (KV Cache)缓存起来, 在用户下一次调用推理时检索是否有缓存的数据, 若有, 则无需计算直接使用已有的缓存, 加速TTFT.

实现和效果: 实现的实体是Cache Manager (CM), CM内部有一颗索引树, 管理HBM和DRAM空间, 等功能

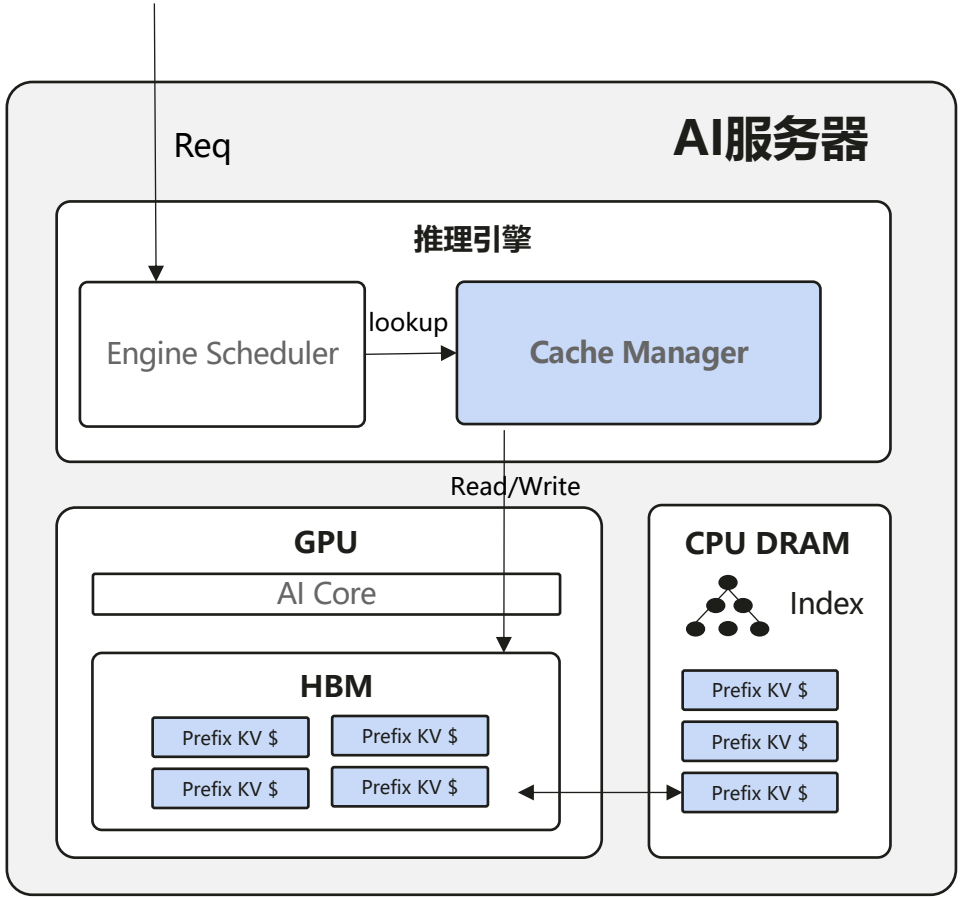
某语音助手推理场景
降低首字时延 >60%



相关工作: SGLang, Pensieve

架构

多轮对话/RAG/Multi-Agent等大模型应用



分离式内存弹性伸缩，解决大模型推理的内存墙，降低推理集群成本

介绍

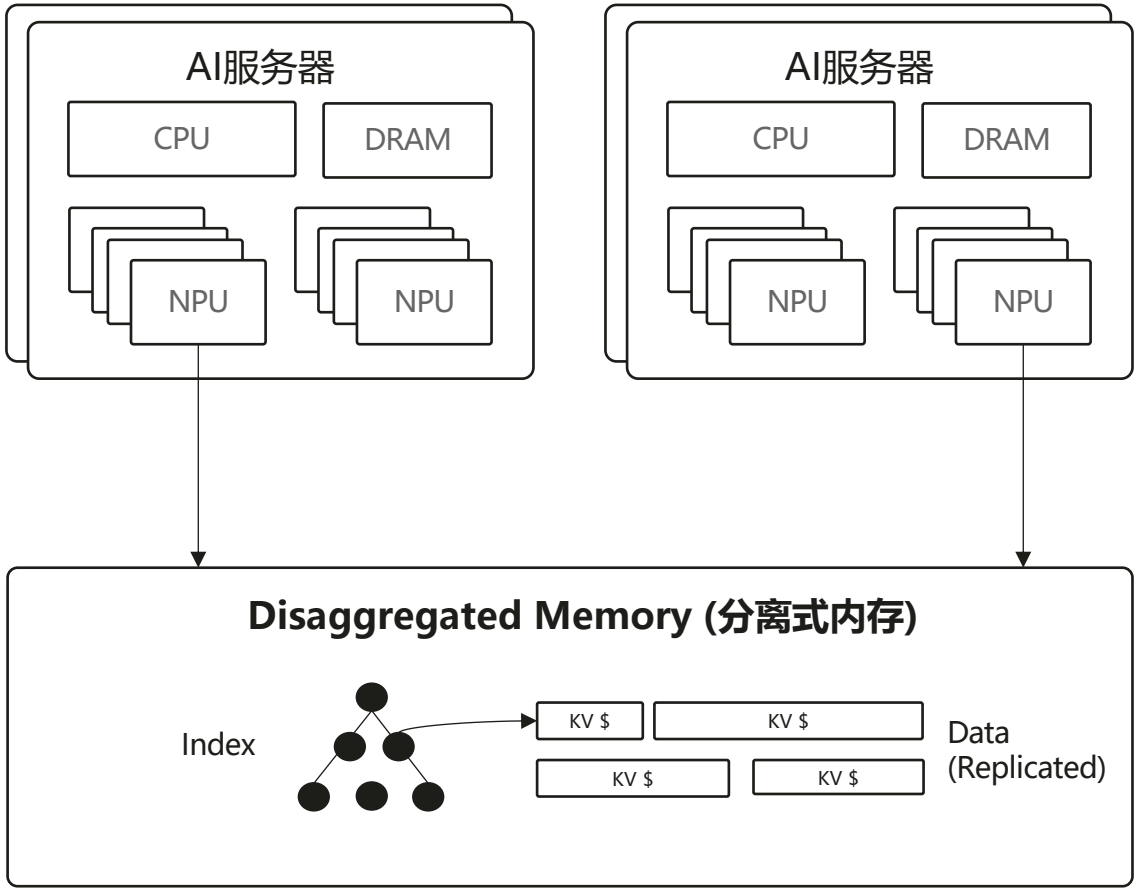
解决的问题: (1) 大语言模型在推理过程中会产生大量的缓存数据; (2) 推理服务器会配置大量内存, 但内存利用率不高, 资源浪费; (3) 数据无容错, 节点故障需重启推理.

解决的思路: Consolidate over Static-Provisioning, 把内存池化拉远, 形成分离式内存, 解决资源的弹性/利用率/容错等问题.

实现和效果: 构建一个分离式内存池, AI服务器与分离式内存池直通, 通过高速网络 (700~900GB/s带宽)互联. 分离式内存保存索引, 数据以及其副本. 预定在5%性能下降的前提下节省40%内存 , 提升性价比1.6x

相关工作: 无系统工作, 硬件趋势见GH200

架构



资源感知调度，缓解大模型推理的调度墙，提升推理集群利用率

介绍

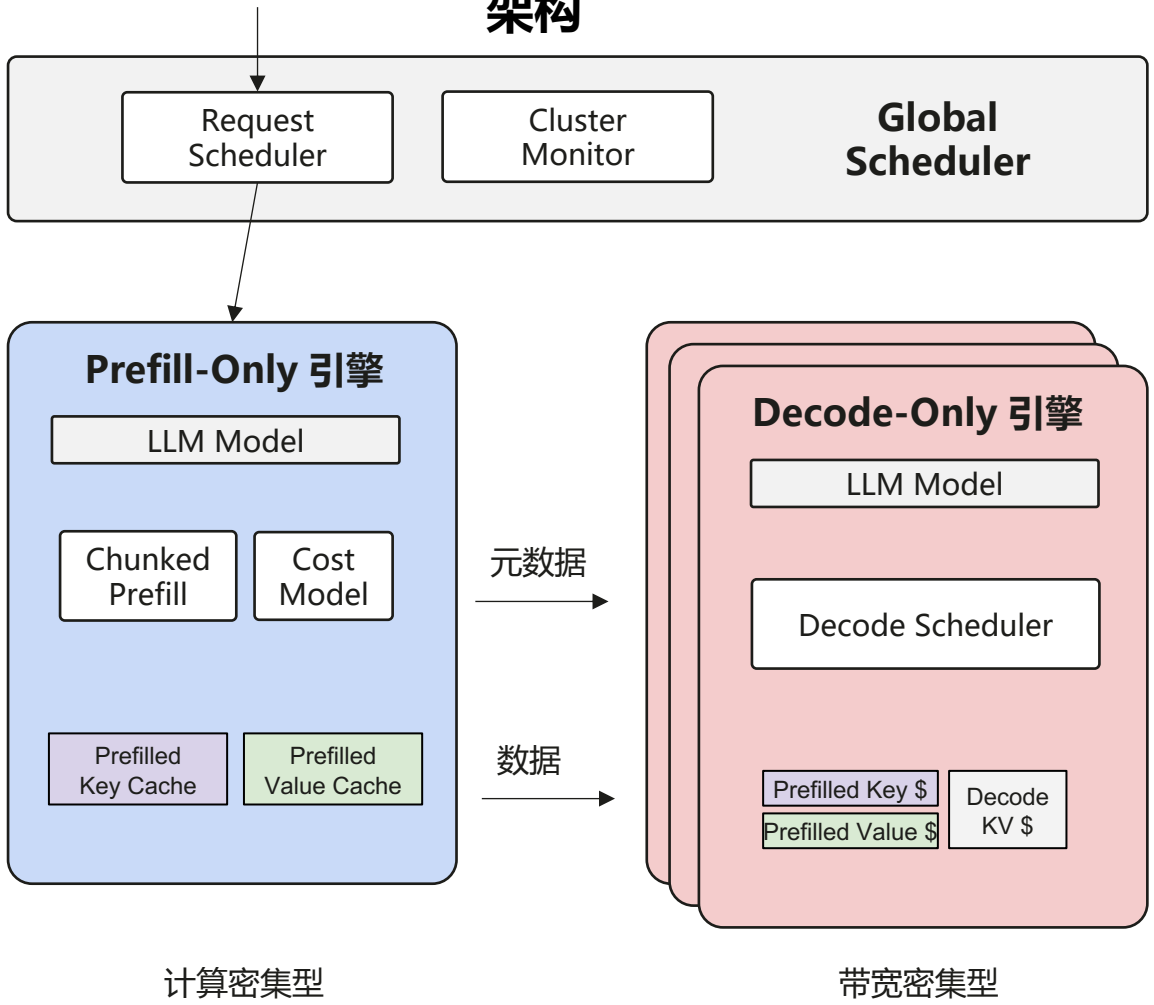
解决的问题: 大模型推理的全量和增量阶段呈现迥然的资源需求, 一个是计算密集一个是带宽密集, 因此对内存和算力的需求不同.

解决的思路: 全量和增量分离, 叠加上分布式调度, 集群二级调度, 提升集群的资源利用率.

实现和效果: 全量和增量分别部署, 增量可根据负载伸缩. 从实测看, PD分离架构适用于Prompt短Decode长的负载, 能提升约2x性价比.

相关工作: TetriServe, DistServe, Splitwise等.

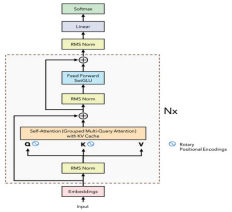
架构



Put it together

大模型推理 **负载和硬件间的冲突** **整系统解决方案**

Transformer



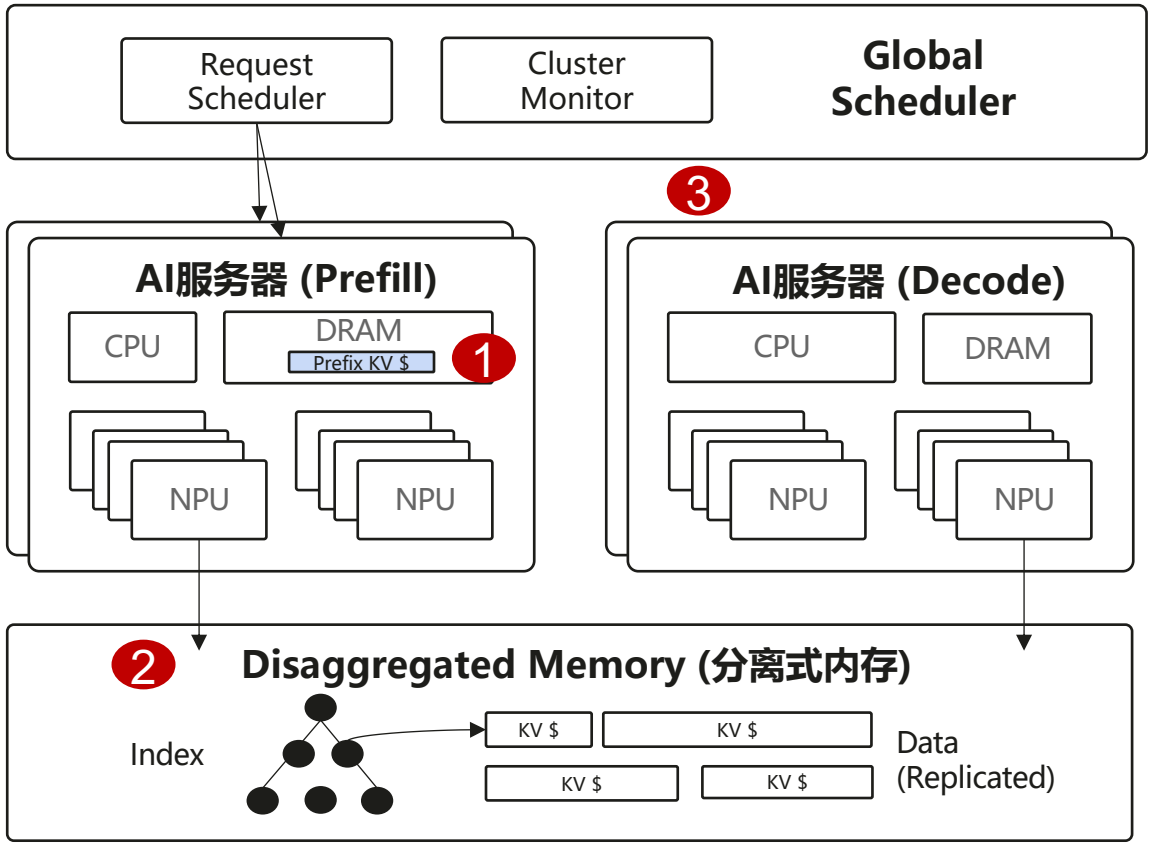
\+

AI 芯片

1 算力瓶颈

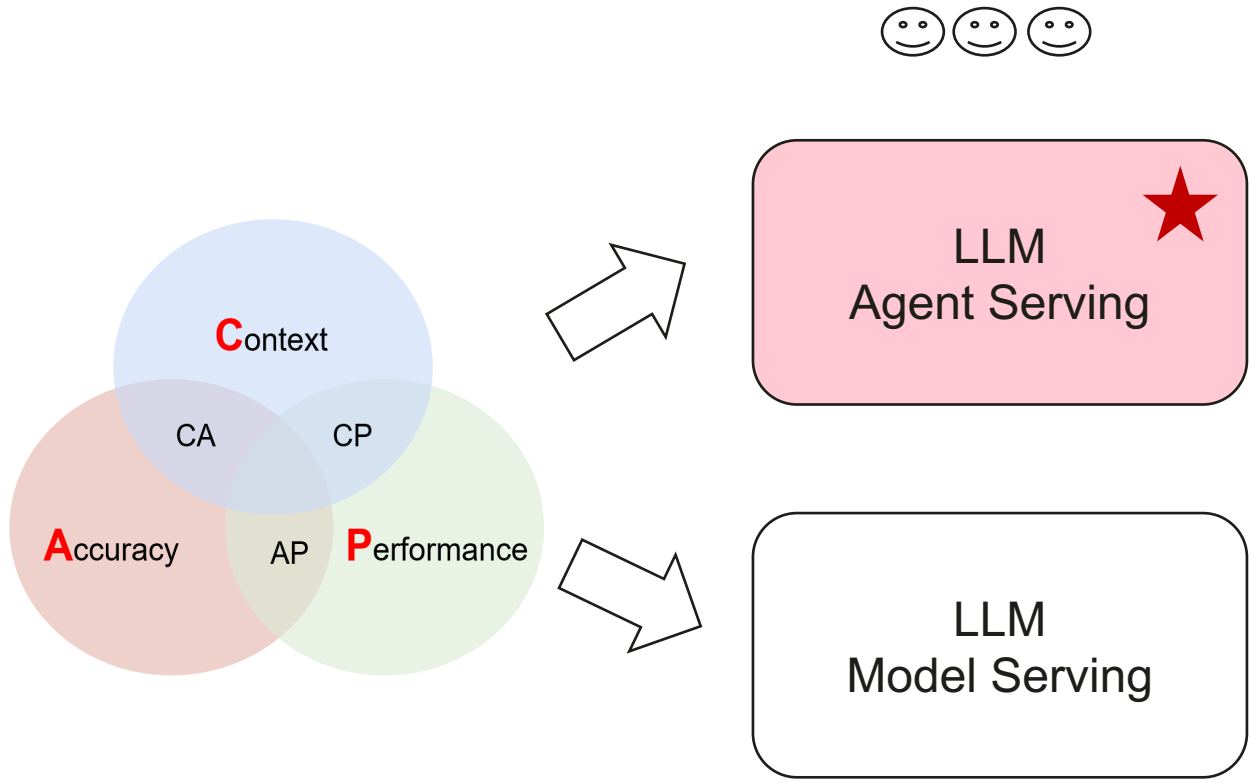
2 内存瓶颈

3 调度瓶颈

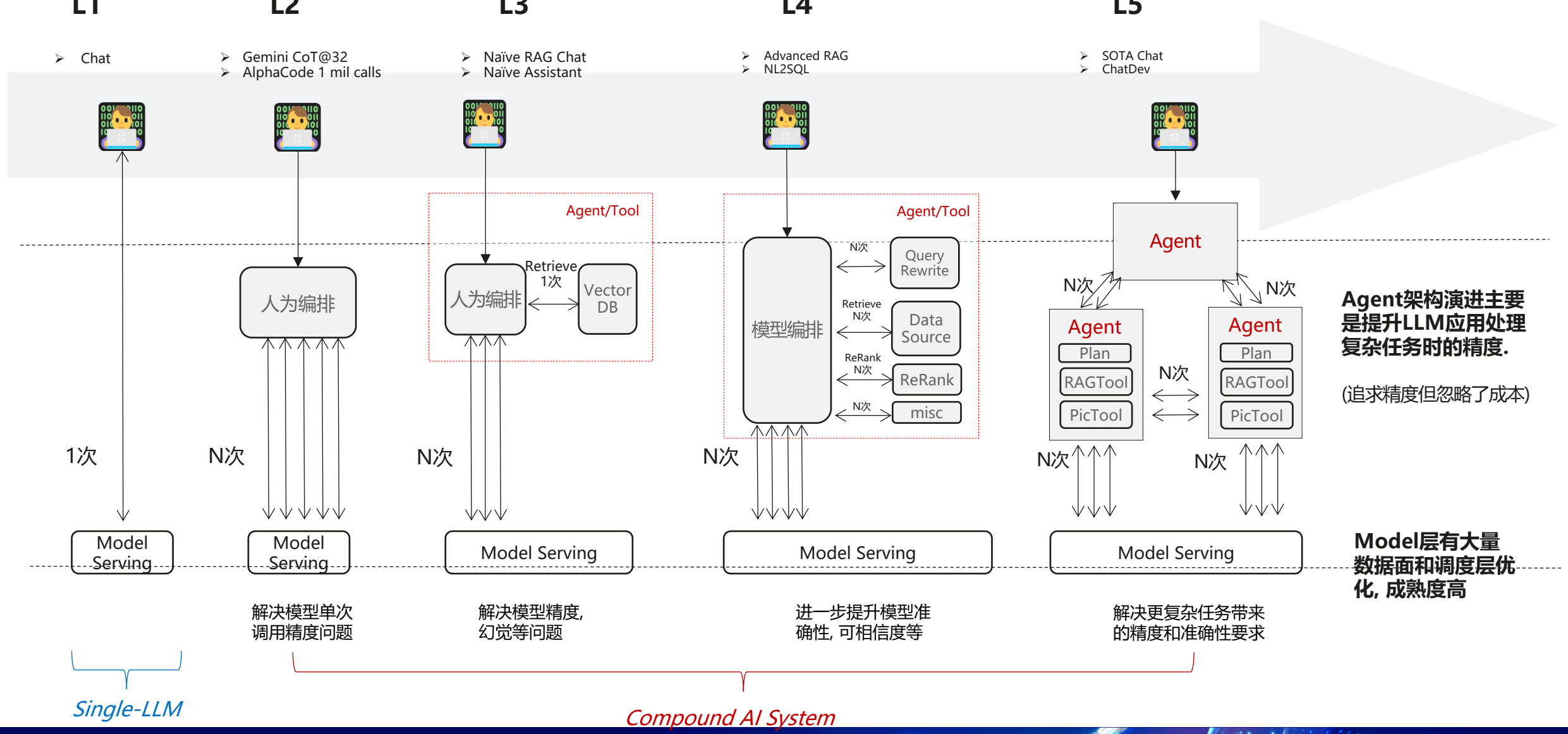


Outline

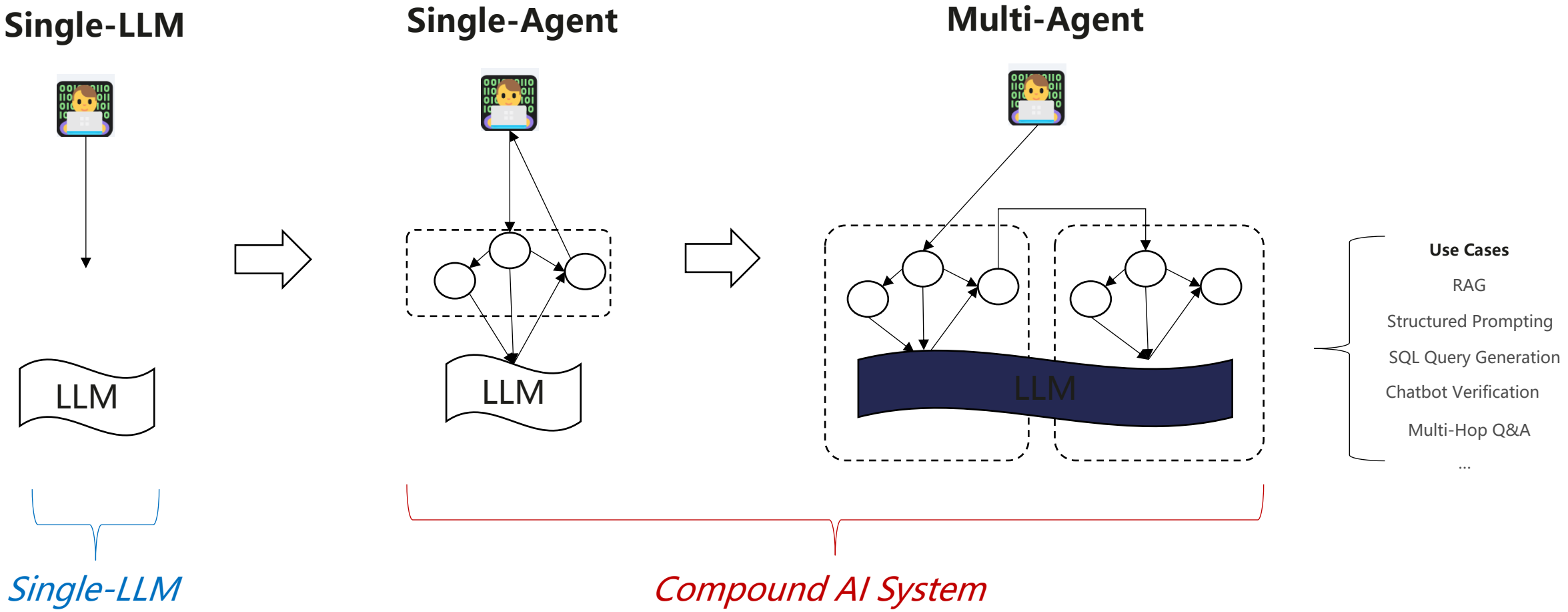
1. 介绍
2. LLM Model Serving
3. **LLM Agent Serving**
4. **未来**



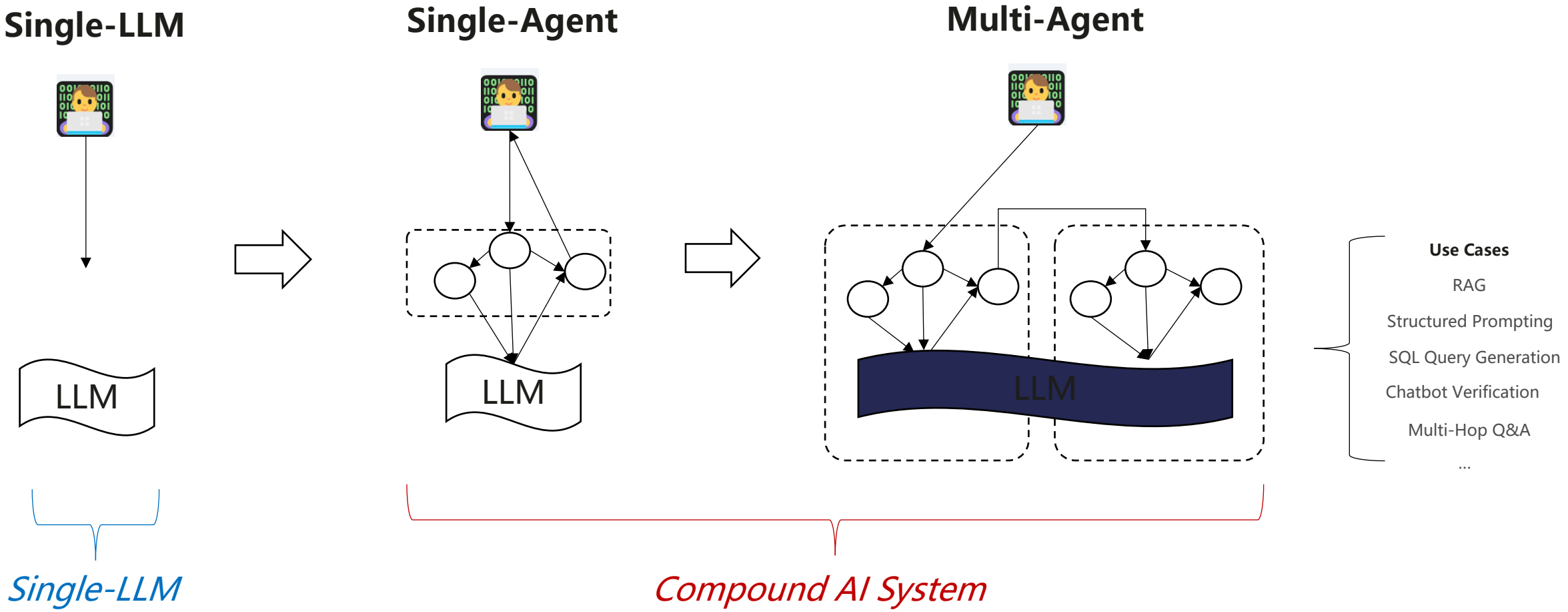
AI应用体系架构趋势: 从Single-LLM走向了Compound AI System

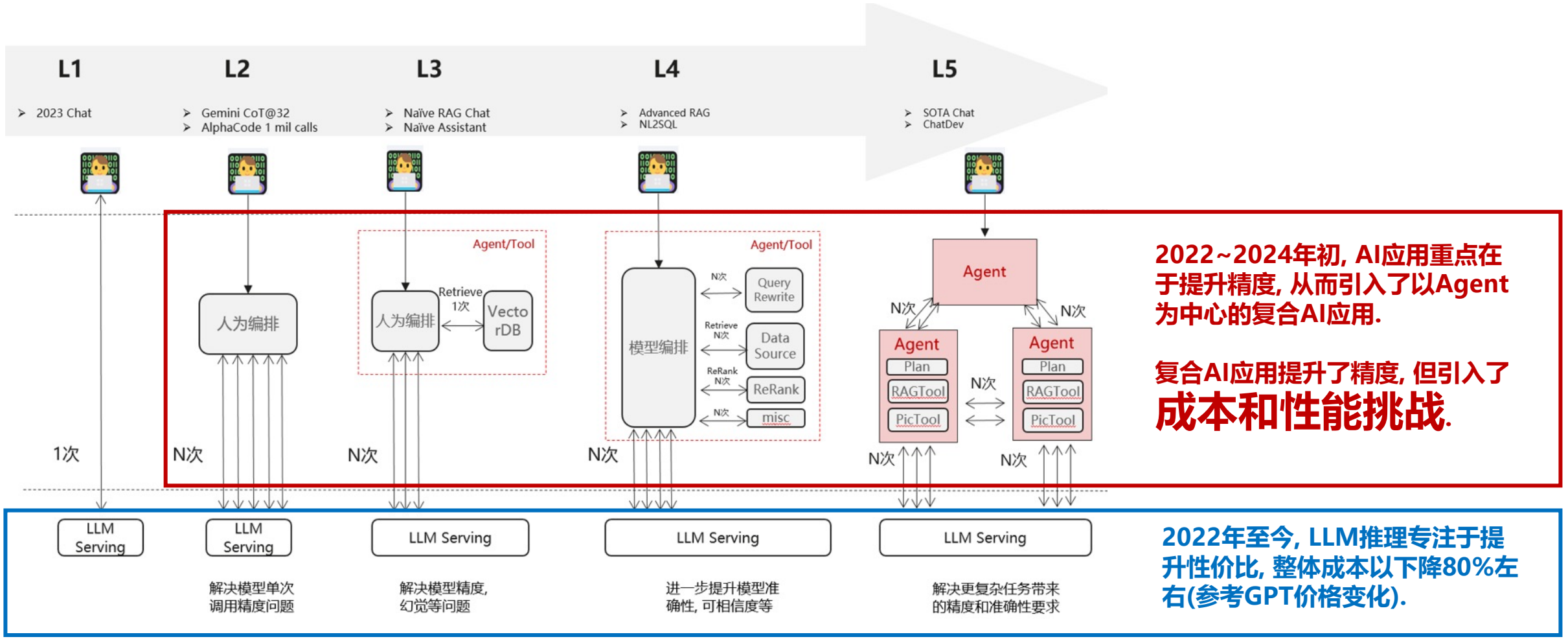


AI应用体系架构趋势:
从Single-LLM走向了Compound AI System



AI应用体系架构趋势:
从Single-LLM走向了Compound AI System



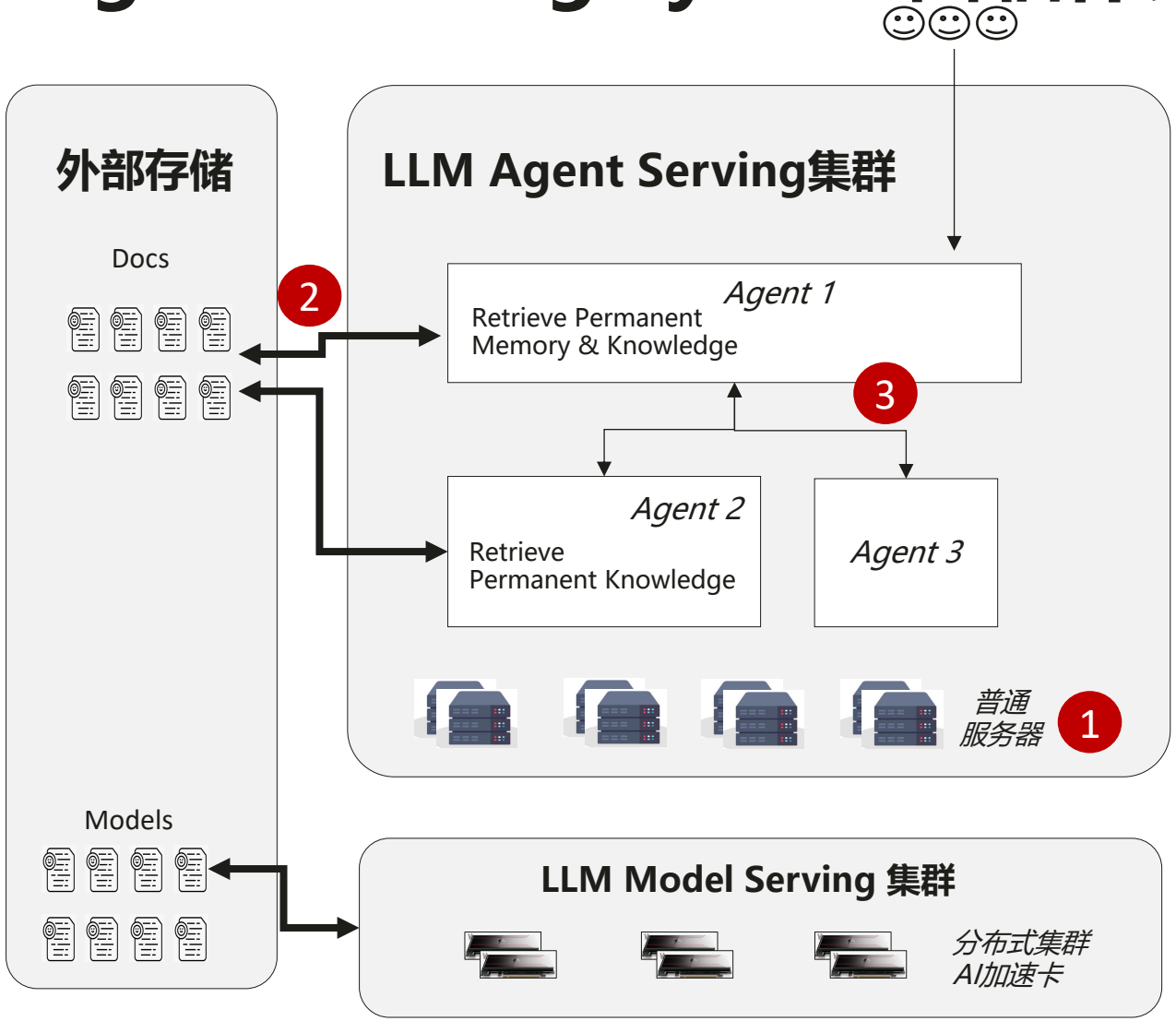


2022~2024年初, AI应用重点在于提升精度, 从而引入了以Agent为中心的复合AI应用.

复合AI应用提升了精度, 但引入了成本和性能挑战.

2022年至今, LLM推理专注于提升性价比, 整体成本以下降80%左右(参考GPT价格变化).

Agent Serving System面临什么挑战?

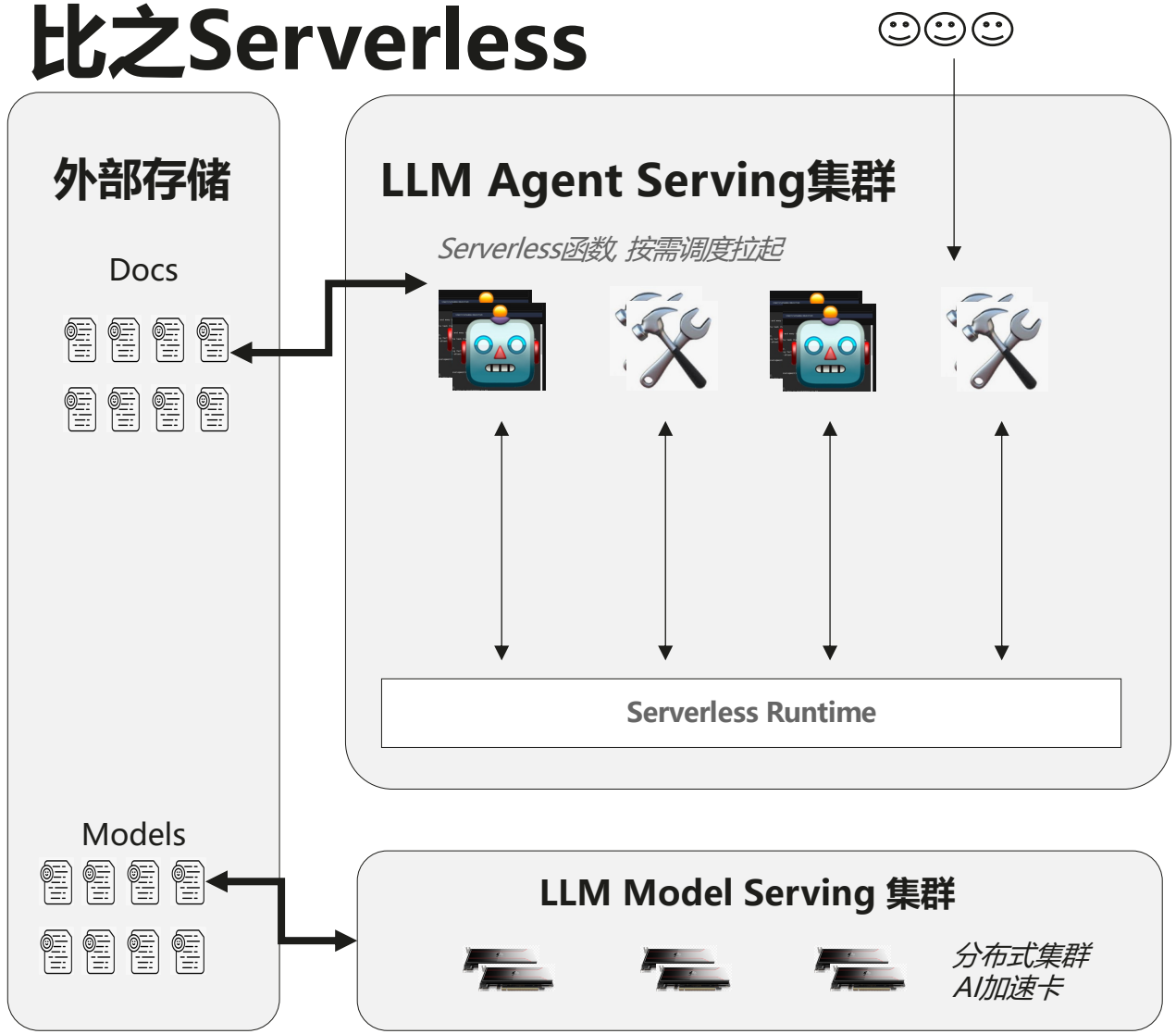


核心: 提升精度的同时, 引入了很多系统组件和流程, 以及大量的数据流转, 从而引入成本/性能挑战.

SSS挑战

	挑战	描述
1	**Server** **静态资源配置挑战**	使用静态资源运行AI Agents, 会Over-Provision, 带来成本问题.
2	**Storage** **持久化数据访问挑战**	Agents多次访问持久化存储的外部知识和Memory, 存在时延和带宽挑战.
3	**Scheduling** **数据流转挑战**	Multi-Agent协同有大量数据传递, 存在数据流转开销挑战.

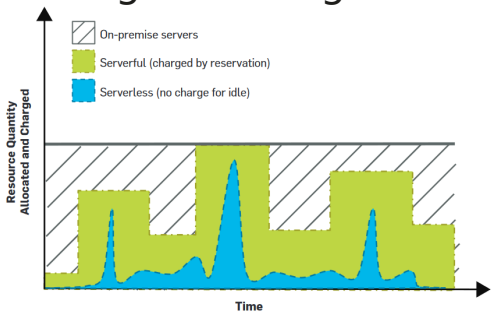
如何提升Agent Serving System的性价比? 数据中心提性价比之Serverless



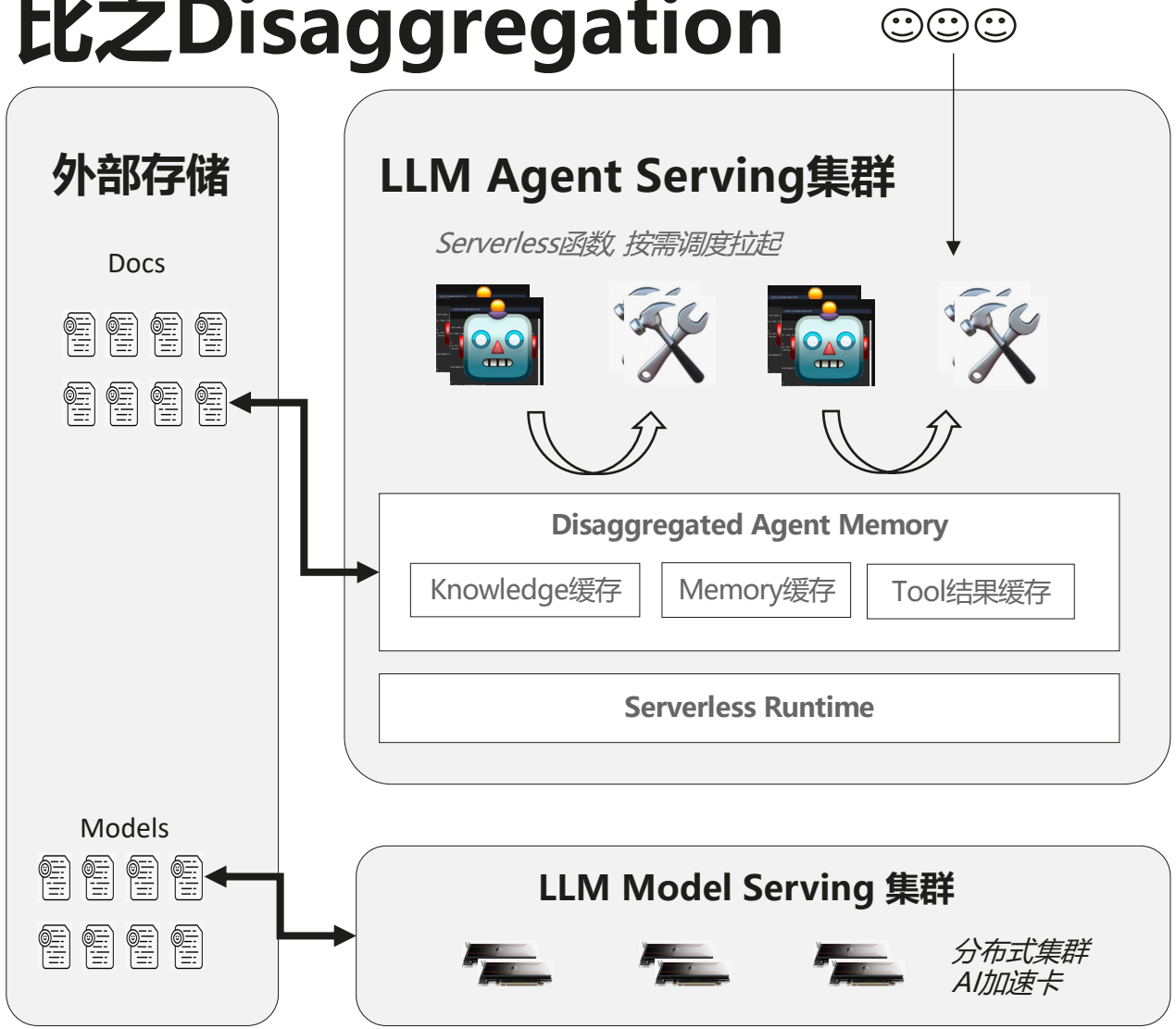
应对SSS挑战 - Serverless

	挑战	解决思路
1	**Server** **静态资源配置挑战**	Agent/Tool执行**Serverless**化
2	**Storage** **持久化数据访问挑战**	
3	**Scheduling** **数据流转挑战**	

将Agent/Tool等Serverless化,
降低Agent Serving成本



如何提升Agent Serving System的性价比? 数据中心提性价比之Disaggregation



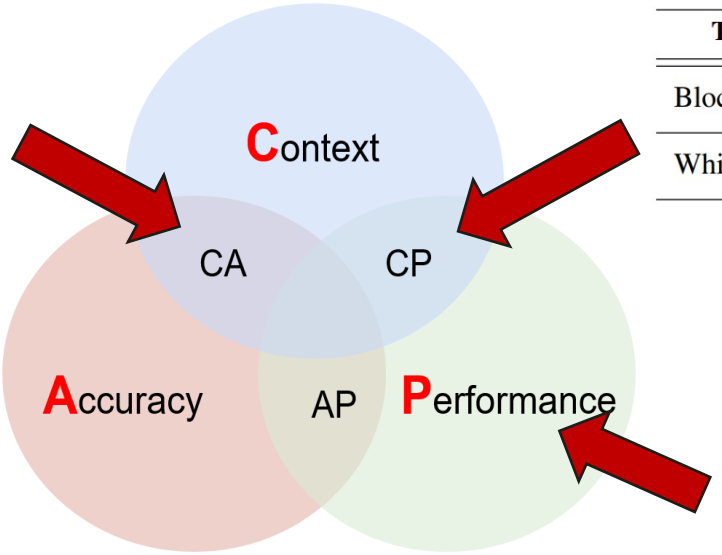
应对SSS挑战 - Disaggregation

	挑战	解决思路
1	**Server** **静态资源配置挑战**	Agent/Tool执行**Serverless**化
2	**Storage** **持久化数据访问挑战**	**分离式Agent Memory缓存**
3	**Scheduling** **数据流转挑战**	

将Agent状态做池化统一管理,
让Agent/Tool更加弹性

Agent Memory

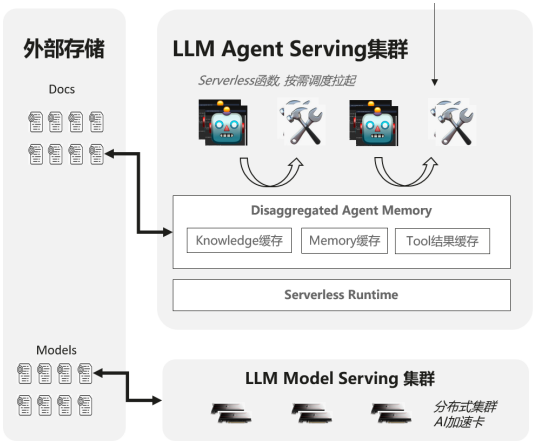
Work	Online Memory Management	Offline Memory Reflection
MemWalker [106]	✓	/
WebGPT [107]	✓	/
MemGPT [105]	✓	/
TheSim [108]	✓	✓
ChatDev [109]	✓	✓
MetaGPT [110]	✓	✓
Self-Refine [111]	✓	✓
Reflexion [112]	✓	✓
MLCopilot [113]	✓	✓



Prompt Compression

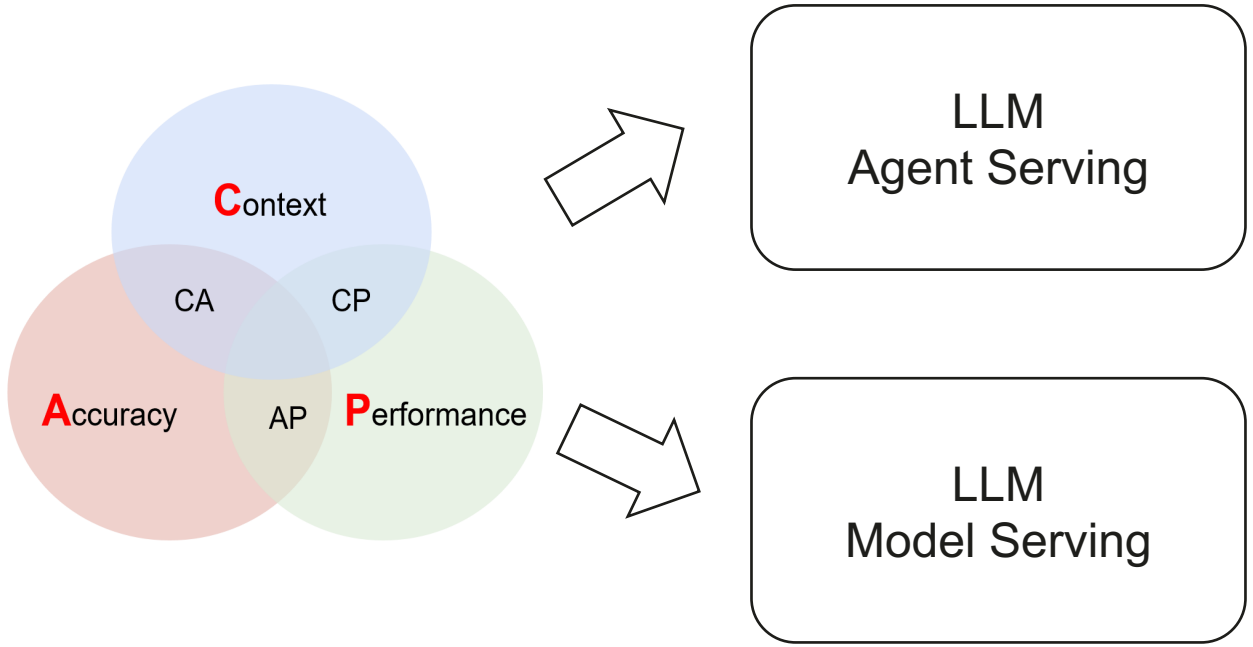
Type	Work
Block-Box	Selective Context [98], LLMLingua [99], LongLLMLingua [100], LLMLingua2 [101]
White-Box	Gist-Token [102], PCCC [103], ICAE [104], AutoCompressor [32]

Serverless + Disaggregation for Agent
属于提升P的工作



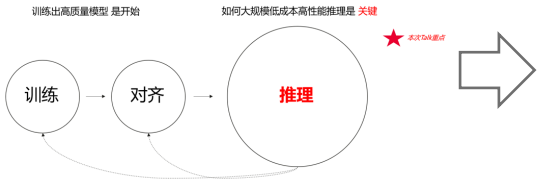
Outline

- 介绍
- LLM Model Serving
- LLM Agent Serving
- **总结**

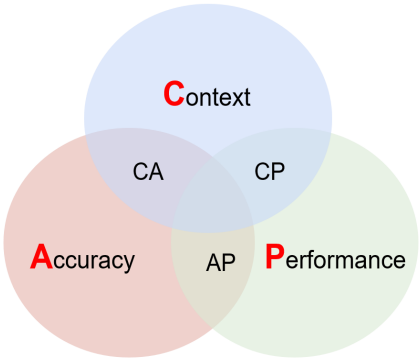


回顾

AGI三步骤
推理很重要



推理有挑战, 平衡
长度/精度/性能



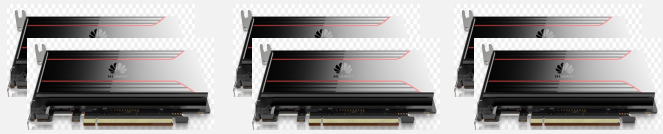
LLM Agent Serving 集群

Serverless Agent
Disaggregate 内存

LLM Model Serving 集群

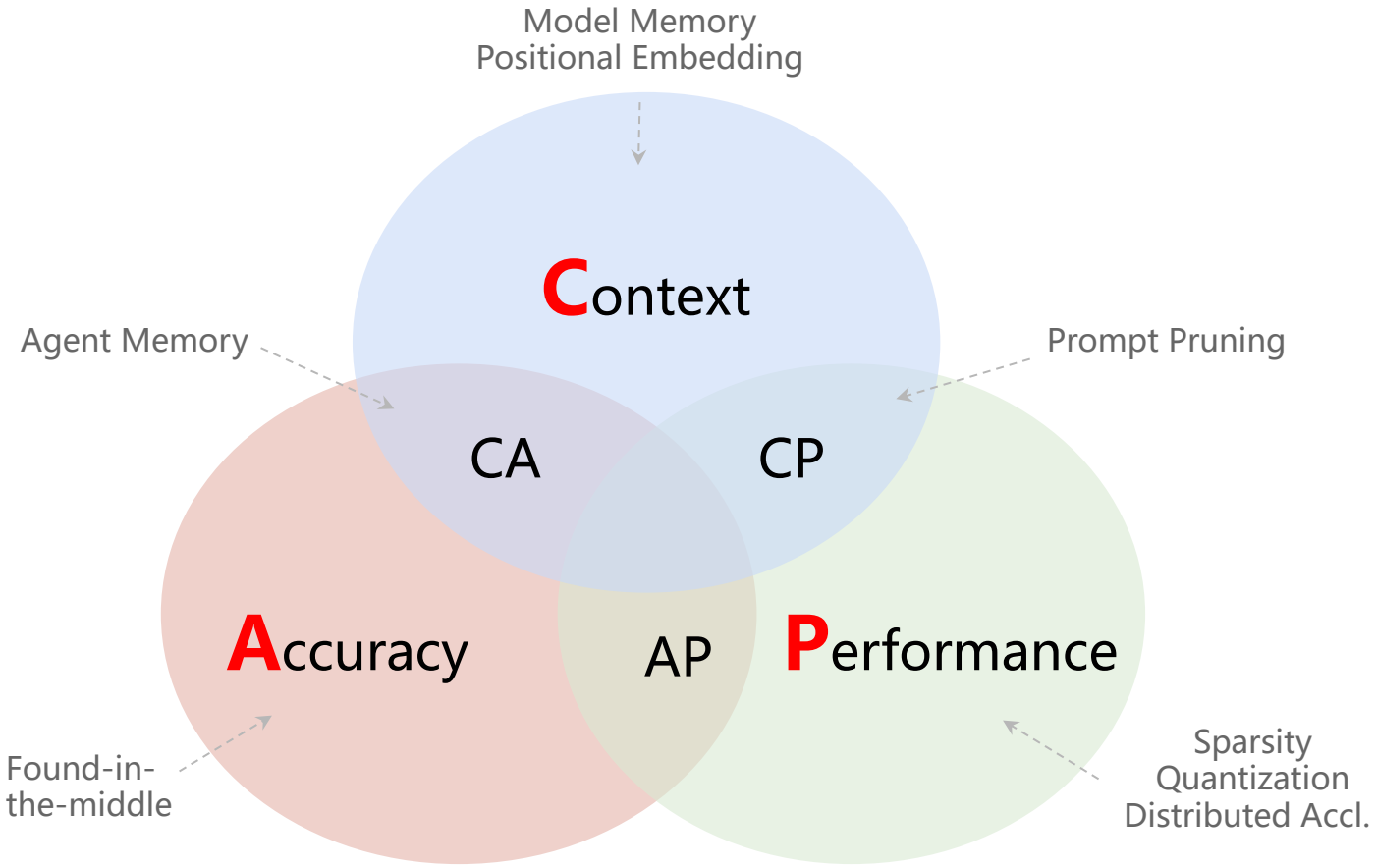
Global Scheduling

Caching ***Disaggregated-Mem & PD***



分布式集群
AI加速卡

以终为始，基于CAP原则，构建产业需要的通用人工智能Serving平台



在大规模场景下，通过Agent系统与模型系统的整合优化，打造最佳性价比的

AI Serving Infra for AGI

科技生态圈峰会 + 深度研习

——1000 + 技术团队的共同选择

KEYLINK ing

K+峰会

K+峰会 上海站
K+全球软件研发行业创新峰会
时间：2024.06.21-22

K+峰会 敦煌站
K+思考周®研习社
时间：2024.10.17-19

K+峰会 香港站
K+思考周®研习社
时间：2024.11.10-12

K+峰会详情

AiDD峰会

AiDD峰会 上海站
AI+研发数字峰会
时间：2024.05.17-18

AiDD峰会 北京站
AI+研发数字峰会
时间：2024.08.16-17

AiDD峰会 深圳站
AI+研发数字峰会
时间：2024.11.08-09

AiDD峰会详情

AiDD

THANKS