QCon
全球软件开发大会

InfoQ 极客传媒

AI 2.0时代的大模型推理：从模型到硬件的协同优化

曾书霖

上海无问芯穹智能科技有限公司总经理、硬件设计VP

目录

InfoQ 极客传媒

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

以智能革命 引领大模型推理范式变革

以人工智能为代表的第四次工业革命（智能革命）极大提升人类生产力

蒸汽机
解放农业生产力
Cromford Mills
Harland & Wolff

电力
加速工业发展
SIEMENS
Shlnorde Electric

信息技术
提升各产业生产力
intel
Google
Microsoft
Tencent

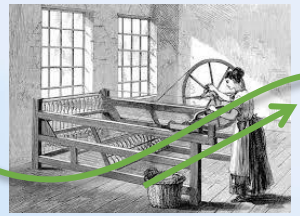
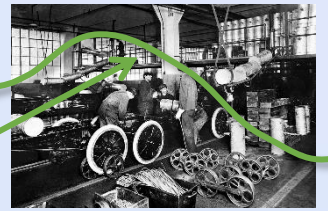
人工智能
智能化应用创造价值
NVIDIA
OpenAI
AMD
阿里云
deepseek
Boston Dynamics

第一次 机械革命
体力密集型生产
生产100米布/人日

第二次 电气革命
体力密集型生产
组装一辆汽车/人日

第三次 数字革命
重复性脑力工作
100阶线性方程组求解
1800人小时→0.05人小时

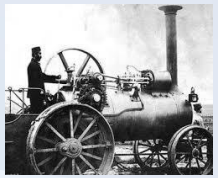
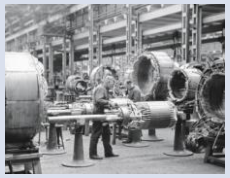
第四次 智能革命
创造性劳动
药物发现
未知时间→月量级

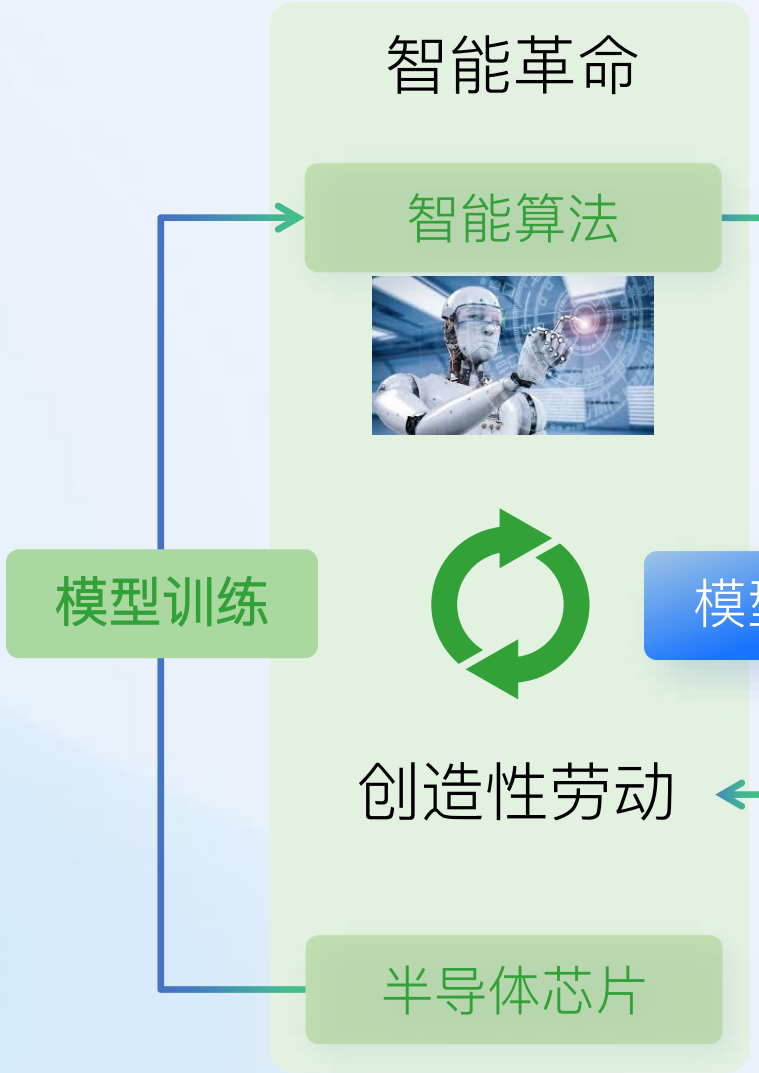
1800 1825 1850 1875 1900 1925 1950 1975 2000 2025 Future

世界GDP增速
世界GDP总量

生产工具与驱动方式的创新

工业革命	机械革命	电气革命	数字革命	智能革命
工具创新	蒸汽机	内燃机	互联网	智能算法
				
人类生产力水平与认知边界不断突破				
替代劳动	体力重复劳动	流程化劳动	知识管理劳动	创造性劳动
驱动方式	蒸汽动力	火电动力	知识信息	半导体芯片

模型推理：技术协同的中枢与产业价值的放大器

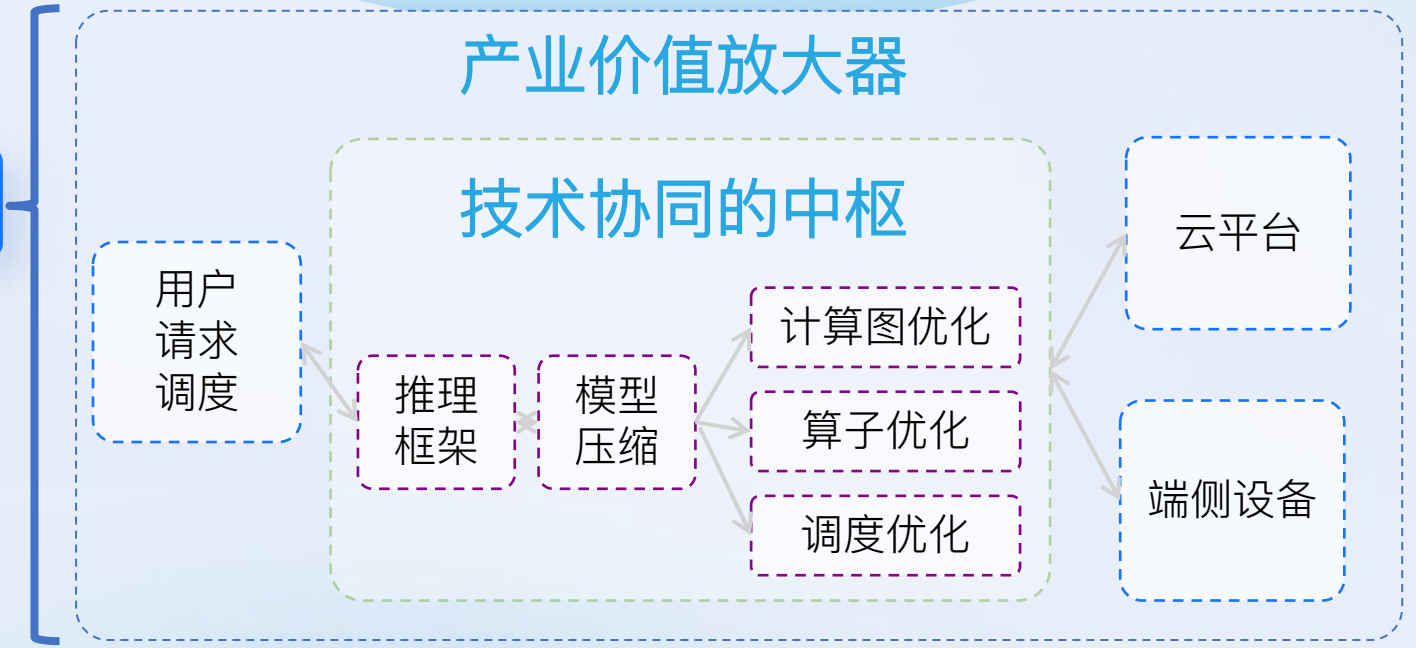



“人工智能+”制造


“人工智能+”金融


“人工智能+”能源


“人工智能+”医疗 …



自2012年以来生成模型发展的关键节点

国外标志性节点

2017年谷歌提出Transformer
Transformer架构奠定了LLM基础，开启大模型时代

2019年谷歌提出T5架构
验证"Text-to-Text"范式在NLP任务中的通用性

2020年OpenAI提出GPT-3
展示LLM强大的的少样本学习能力，引发业界对大模型的研究热潮

2022年OpenAI InstructGPT
提出利用RLHF讲LLM与人类对齐，ChatGPT的基础

2023年OpenAI ChatGPT爆火
引爆全球生成式AI应用，标志AI进入大规模普及阶段

2023年Meta开源Llama
后续逐步成为全球第一大大模型开源模型及生态

2024年OpenAI SORA爆火
火爆全球的视频生成软件，首次实现1分钟长视频的生成，且画面一致性较高

2024年OpenAI提出O1系列模型
将长思维链推理技术带入主流，总结Test Time Scaling，显著提升模型的推理能力

国内标志性节点

2017 2019 2020 2021 2022 2023 2024 2025

2016年商汤、旷世崛起
以计算机视觉技术为核心，推动AI在安防等领域的落地

2021年百度推出ERNIE-3.0
在多项NLP任务中超越GPT-3

智谱推出ChatGLM
第一个国产大模型

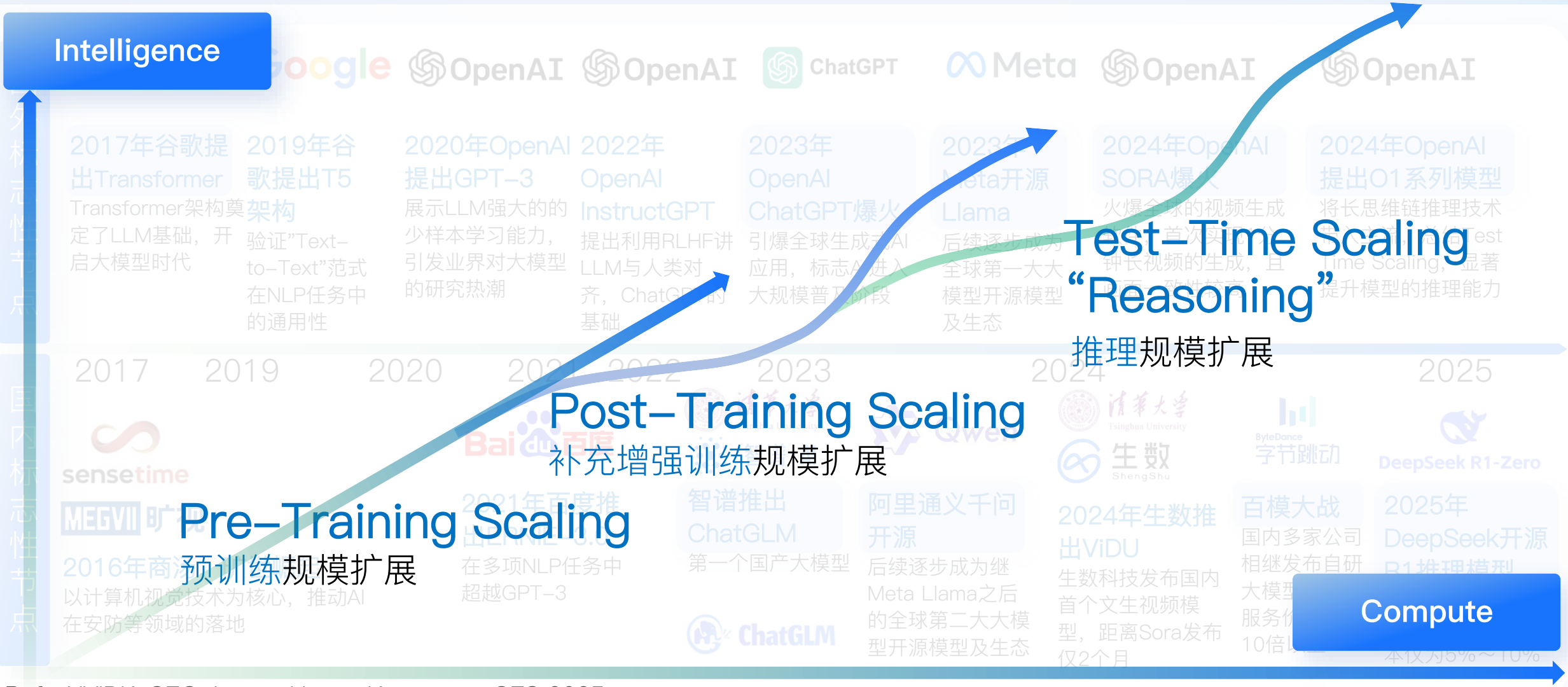
阿里通义千问开源
后续逐步成为继Meta Llama之后的全球第二大大模型开源模型及生态

2024年生数推出ViDU
生数科技发布国内首个文生视频模型，距离Sora发布仅2个月

百模大战
国内多家公司相继发布自研大模型，API服务价格降低10倍以上

2025年DeepSeek开源R1推理模型
比肩OpenAI O1算法性能的同时，成本仅为5%～10%

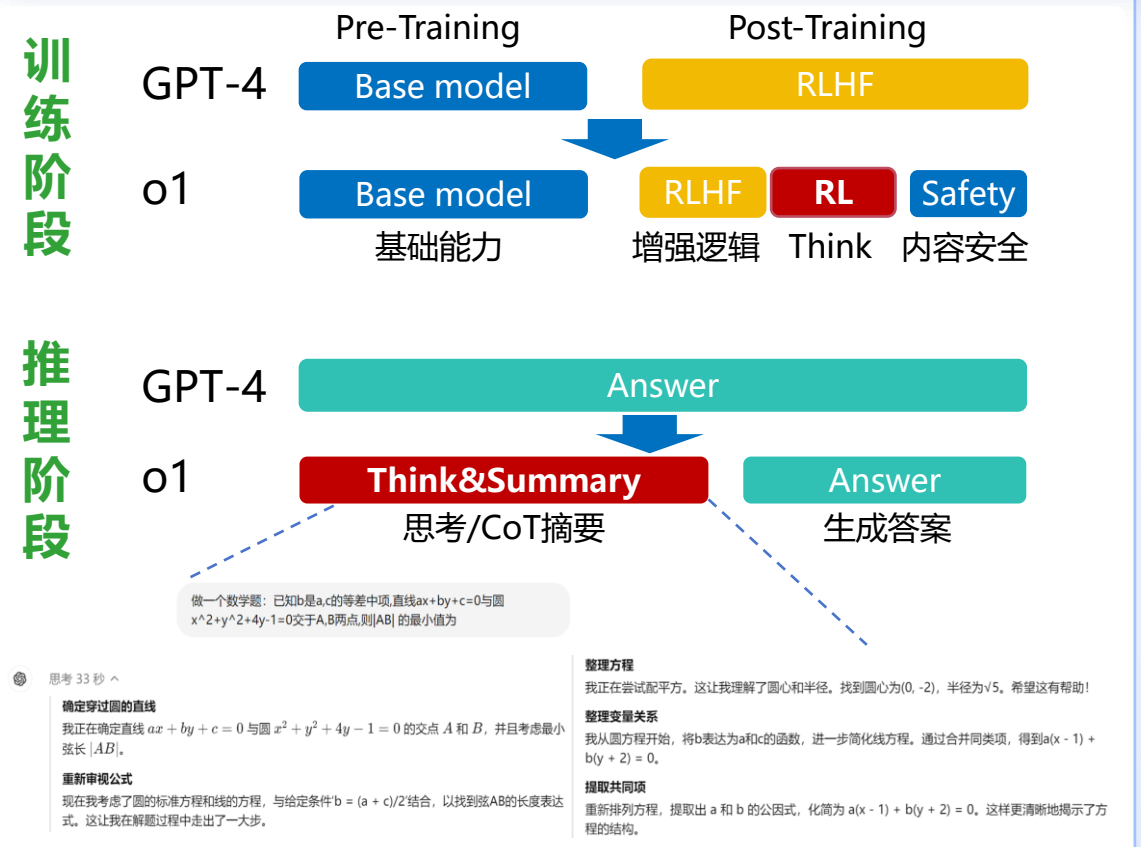
尺度定律(Scaling Law)逐步转向推理规模扩展



Ref：NVIDIA CEO Jensen Huang Keynote at CES 2025

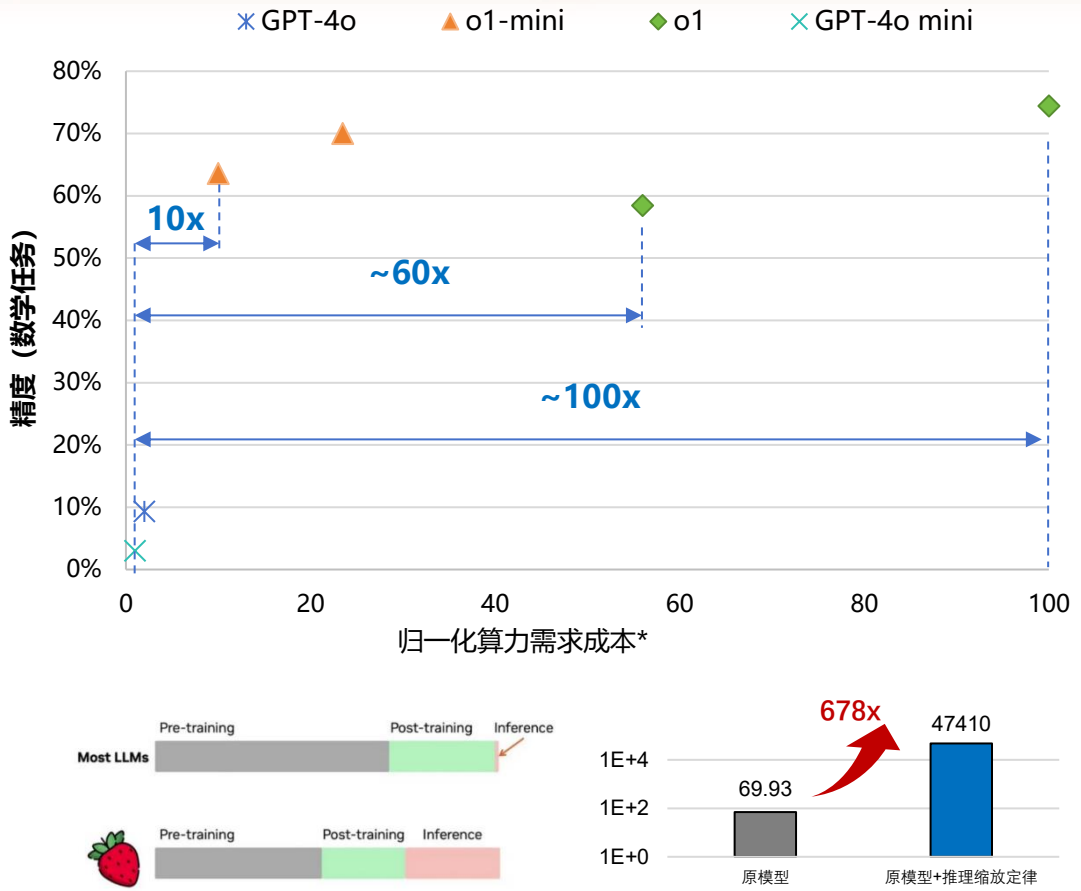
模型推理计算范式转变导致计算需求激增

GPT-o1的计算范式变化：CoT + 强化学习



*ref：[OpenAI o1-mini | OpenAI]

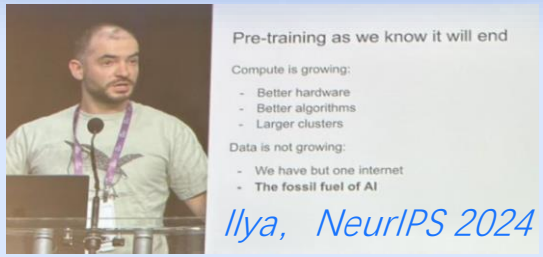
增加模型的推理迭代次数，计算量增加2个量级



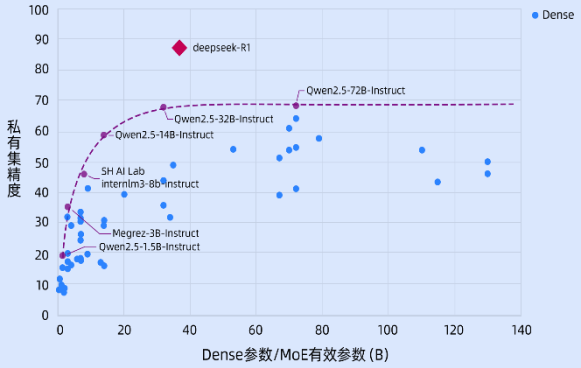
Deepseek等模型推动应用落地加速带来推理需求爆发

未来推理算力需求或是预训练需求的百倍以上

预训练时代即将结束

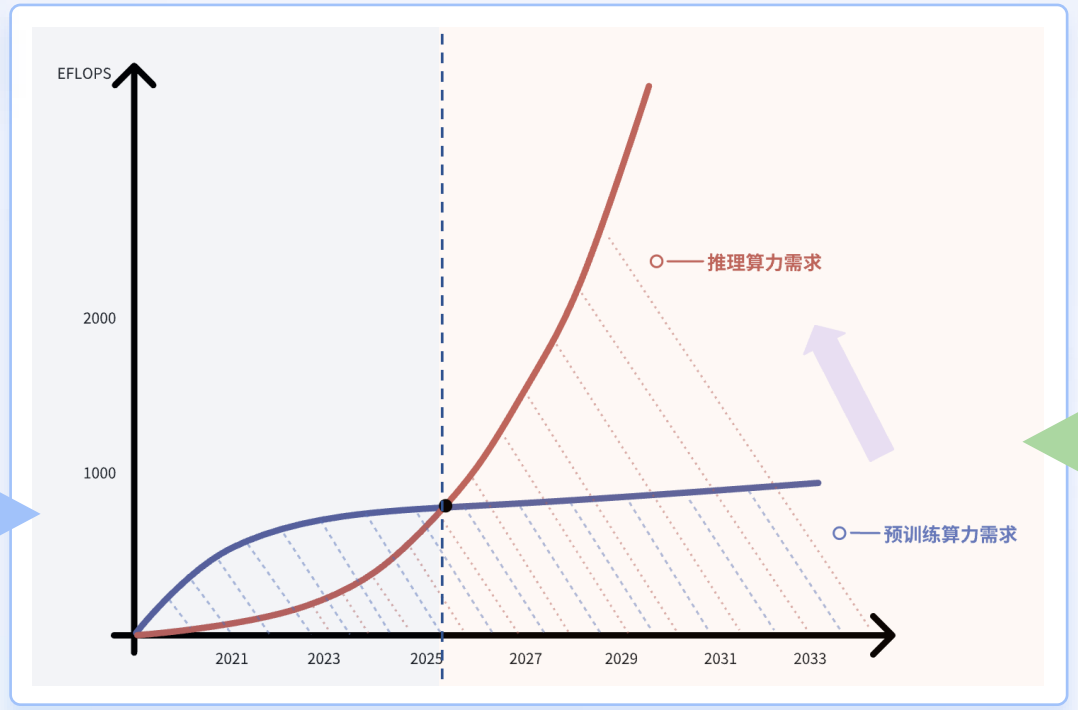


步入2025 年，受限于预训练数据、大模型迭代周期、硬件成本等因素

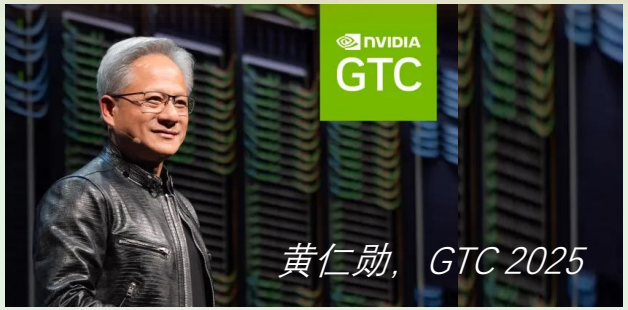


Pre-Training Scaling

收益正在衰减



推理算力需求“喷发式增长”

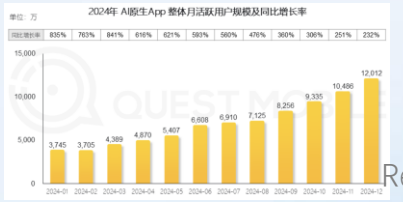


为了保持模型的响应速度，让用户不会因等待而失去耐心，我们现在需要将计算速度提高10倍，黄仁勋表示：

整体计算需求

轻松增长至100倍

推理模型需要在回答之前进行多轮内部推理，因此计算需求大得多，响应时间也更长。



已有**70%**中国亿级用户APP进行了“AI转型”

共962个传统APP备案了深度合成算法

Ref: QuestMobile A产业洞察研究院 2025年1月

大模型推理系统形态

手机	AIPC	一体机	云服务
			
应用需求：100–1000 token/s*		理论值：25–30k token/s/节点+	
现有实现：5–10 token/s	现有实现：20–25 token/s	现有实现：~8000 token/s	现有实现：~14.8k token/s/节点
模型：Qwen2.5–7B 硬件设备：骁龙8 Gen1 推理框架：llama.cpp+OpenCL	模型：Qwen2–7B 硬件设备：AMD Krackan 推理框架：llama.cpp	模型：DeepSeek–R1 硬件设备：8xH20 推理框架：SGLang	模型：DeepSeek–R1 硬件设备：22台H800 推理框架：DeepSeek

端侧推理系统　　云侧推理系统

*前沿应用——机器人智能的基本需求，参考自Deray, Jeremie, Joan Sola, and Juan Andrade–Cetto, “Joint on–manifold self–calibration of odometry model and sensor extrinsics using pre–integration.” 2019 European Conference on Mobile Robots (ECMR). IEEE, 2019.

+理论估计上界可达性，假设集群配置为计算瓶颈，考虑计算量和网络延迟计算上界

大模型推理系统形态

端侧推理系统

优化目标：降低延时

核心特点：单用户、少请求

应用需求：100–1000 token/s*

请求

用户 → 端侧设备

现有实现：5–10 token/s

模型：Qwen2.5–7B
硬件设备：骁龙8 Gen1
推理框架：llama.cpp+OpenCL

现有实现：20–25 token/s

模型：Qwen2–7B
硬件设备：AMD Krackan
推理框架：llama.cpp

端侧推理关键问题：资源受限

端侧推理系统

云侧推理系统

一体机

理论值：25–30k token/s/节点+

现有实现：~8000 token/s

模型：DeepSeek–R1
硬件设备：8xH20
推理框架：SGLang

云服务

现有实现：~14.8k token/s/节点

模型：DeepSeek–R1
硬件设备：22台H800
推理框架：DeepSeek

*前沿应用——机器人智能的基本需求，参考自Deray, Jeremie, Joan Sola, and Juan Andrade–Cetto, “Joint on–manifold self–calibration of odometry model and sensor extrinsics using pre–integration.” 2019 European Conference on Mobile Robots (ECMR). IEEE, 2019.

+理论估计上界可达性，假设集群配置为计算瓶颈，考虑计算量和网络延迟计算上界

大模型推理系统形态

端侧推理系统

优化目标：降低延时

核心特点：单用户、少请求

请求

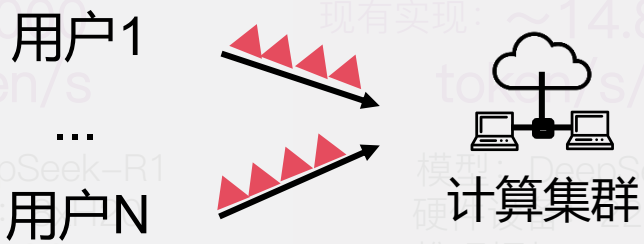
端侧推理关键问题：资源受限

云侧推理系统

优化目标：满足用户延时要求，提升吞吐率

核心特点：多用户、多请求

请求



云侧推理关键问题：资源调度

*前沿应用——机器人智能的基本需求，参考自Deray, Jeremie, Joan Sola, and Juan Andrade-Cetto, “Joint on-manifold self-calibration of odometry model and sensor extrinsics using pre-integration.” 2019 European Conference on Mobile Robots (ECMR). IEEE, 2019.

+理论估计上界可达性，假设集群配置为计算瓶颈，考虑计算量和网络延迟计算上界

大模型推理系统的效率挑战

端侧推理系统

优化目标：降低延时

核心特点：单用户、少请求

小规模负载计算
(利用率：
50%→90%)

重叠传输
延时开销

协同挑战
多处理器硬件资源
协同处理

计算挑战
提升生成式模型的
计算利用率

大规模负载并行计算

存储挑战
同时存储大量请求
的KV cache

存储需求
(存储量
100GB→10GB)

提升显存利用效率
(利用率
70%→90%)

调度挑战
不同用户间资源竞
争及延时需求

满足用户目标延时

云侧推理系统

优化目标：满足用户延时要求，提升吞吐率

核心特点：多用户、多请求

02

以弹性算力集群 驱动云侧智能升级

挑战：云侧大模型推理

① 计算挑战

Prefill阶段 计算密集型 ⟷ **Decode阶段** 访存密集型

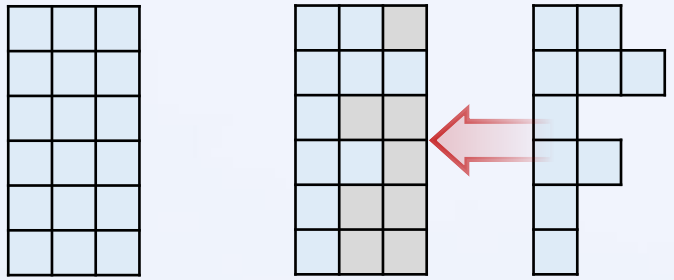
input = ”用python实现冒泡排序”
输入的prompt
↓ Prefill
output = ”def”
第1个token
↓ Decode
output = ”def bubble”
第2个token
↓ Decode
output = ”def bubble_”
第3个token
.....

算子	Prefill计算访存比	Decode计算访存比
Q/K/V/O	$\frac{ld}{l+d}$	$\frac{d}{d+1}$
QK^T	l $<1!$	$\frac{l}{l+1}$
AttenV	l	$\frac{l}{l+1}$
FFN	$\frac{ld_{FFN}}{l+d_{FFN}}$	$\frac{d_{FFN}}{d_{FFN}+1}$

Prefill和Decode两阶段计算特性不同
需分别针对性优化计算

② 存储挑战

云侧集群 存储容量大 ⟷ **推理服务** 存储利用率低



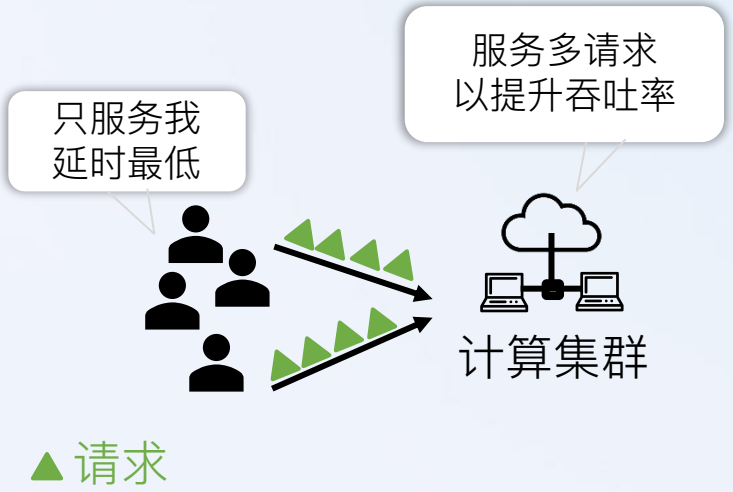
集群存储
TB量级

请求不等长
存储碎片造成利用率低（~70%）

请求动态生成引入存储碎片
导致系统存储利用率低

③ 调度挑战

用户需求 低延时 ⟷ **系统需求** 高吞吐率



用户间存在资源竞争
在满足用户目标前提下提升系统吞吐率

背景：云侧大模型核心技术

云侧推理服务：多用户、多请求

计算挑战
提升生成式模型的计算利用率
→ 算子和计算图优化

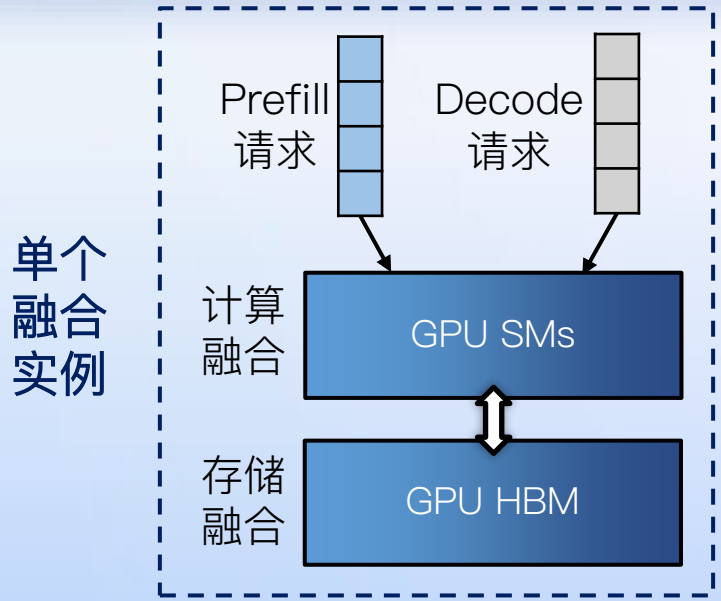
存储挑战
同时存储大量请求的KV cache
→ 分页式存储 前缀缓存

调度挑战
不同用户间资源竞争及延时需求
→ 批处理、请求顺序、阶段间并行优化

云侧推理服务技术Milestone

背景：融合式与分离式实例

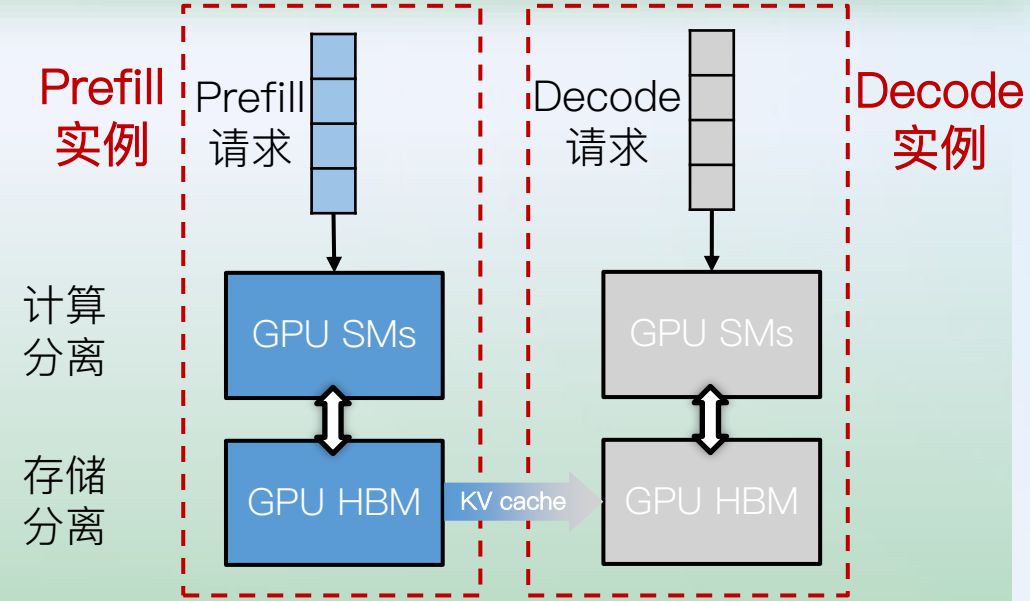
融合式实例：P/D计算和存储共享资源



请求的Prefill和Decode阶段均在相同实例进行计算和存储

代表框架 vLLM SGL

分离式实例：P/D计算和存储资源分离



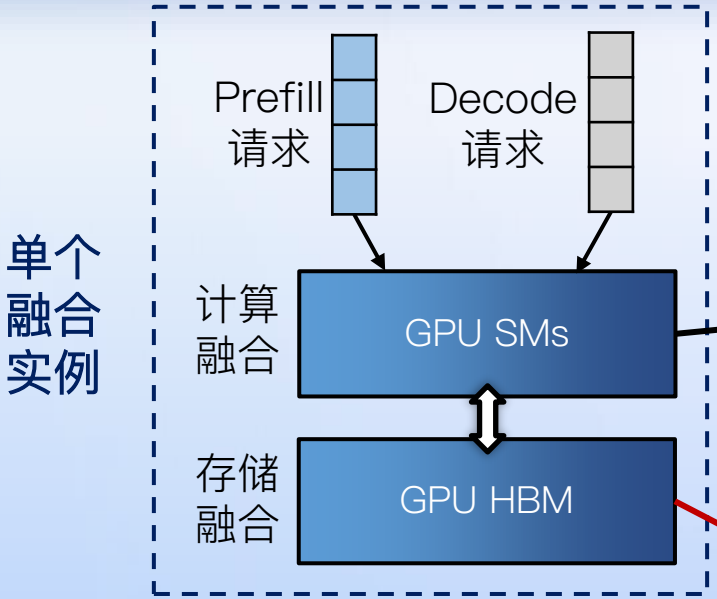
请求在Prefill实例上完成Prefill阶段后，传输KV cache至Decode实例进行计算

代表框架 Mooncake deepseek

[1] W. Kwon, et al. ” Efficient Memory Management for Large Language Model Serving with PagedAttention.”, SOSP, 2023.
[2] L. Zheng, et al. “SGLang: Efficient Execution of Structured Language Model Programs”, NeurIPS, 2024.
[3] DeepSeek-AI. “DeepSeek-V3 Technique Report”, arXiv, 2025.
[4] R. Qin, et al. “Mooncake: Trading More Storage for Less Computation”, FAST, 2025.

思路：融合式与分离式实例分析

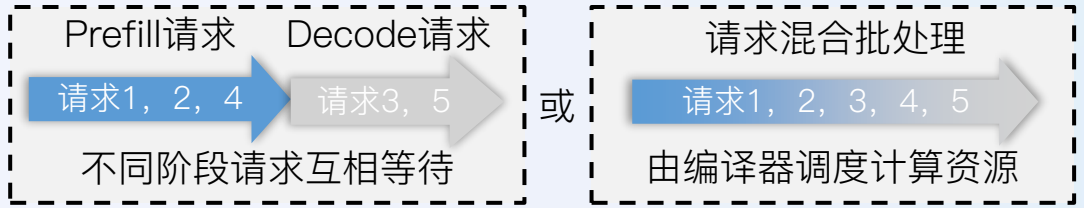
融合式实例：P/D计算和存储共享资源



请求的Prefill和Decode阶段均在相同实例进行计算和存储

代表框架 vLLM SGL

计算劣势



P/D间请求延时干扰严重

P/D计算资源配比不可控

存储优势

P/D计算之间无需传递KV cache

KV cache能使用所有存储资源

[1] W. Kwon, et al. ” Efficient Memory Management for Large Language Model Serving with PagedAttention.”, SOSP, 2023.
[2] L. Zheng, et al. “SGLang: Efficient Execution of Structured Language Model Programs”, NeurIPS, 2024.

思路：融合式与分离式实例分析

计算优势

P/D隔离计算无延时干扰

可以通过GPU数配比计算资源

存储不均衡

Prefill 实例
KV cache生成后即传输

Decode 实例
长期存储KV cache

实例切换需转移存储

Decode 实例1 —切换→ Prefill 实例1

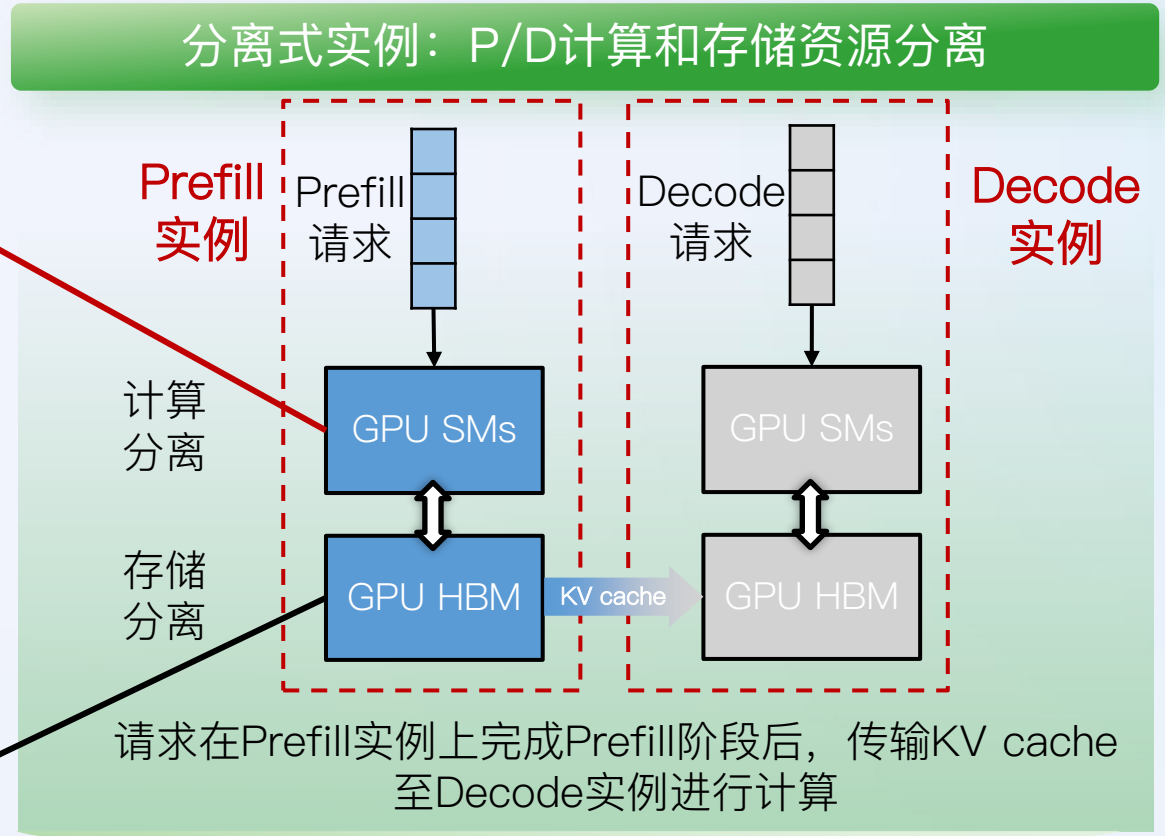
KV cache

Decode 实例2

存储劣势

P/D实例间KV cache存储不均衡
（例，P: 23% v.s. D: 72%*）

资源调整涉及存储搬移开销



代表框架 Mooncake deepseek

*Llama3-70B, Prefill实例TP=4, Decode实例TP=4, ShareGPT数据集

[3] DeepSeek-AI. “DeepSeek-V3 Technique Report”, arXiv, 2025.
[4] R. Qin, et al. “Mooncake: Trading More Storage for Less Computation”, FAST, 2025.

思路：融合式与分离式实例分析

融合式实例：P/D计算和存储共享资源

计算劣势

P/D间请求延时干扰严重

P/D计算资源配比不可控

存储优势

P/D计算之间无需传递KV cache ✓

KV cache能使用所有存储资源

分离式实例：P/D计算和存储资源分离

计算优势

P/D隔离计算无延时干扰 ✓

可以通过GPU数配比计算资源

存储劣势

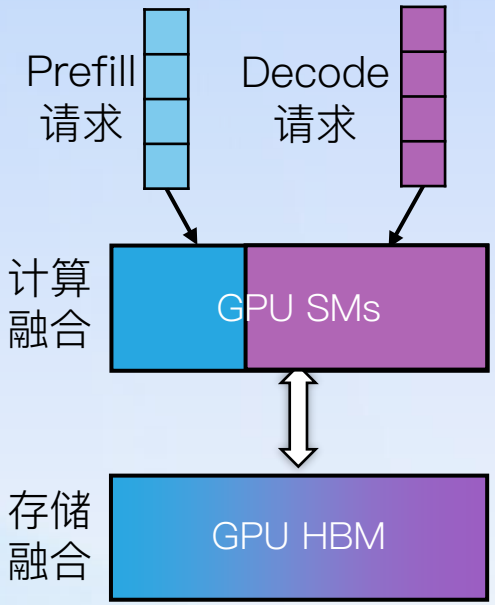
P/D实例间KV cache存储不均衡
(例，P: 23% v.s. D: 72%*)

资源调整涉及存储搬移开销

*Llama3-70B, Prefill实例TP=4, Decode实例TP=4, ShareGPT数据集

方案：semi-PD实例

计算分离 & 存储融合：P/D计算资源隔离，但是共享存储资源



单个semi-PD实例

P/D计算资源隔离且分为不同数据流

但是共享相同的存储资源

计算优势

P/D隔离计算无延时干扰

可以通过GPU数配比计算资源

存储优势

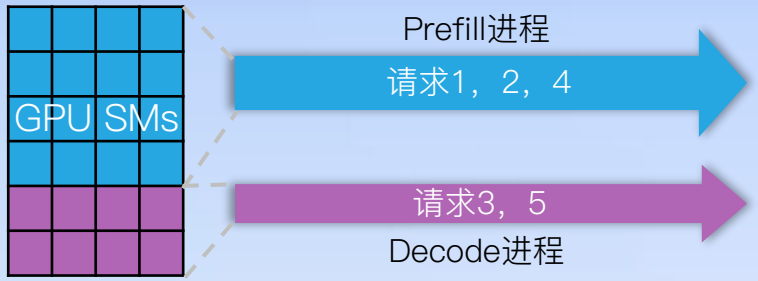
P/D计算之间无需传递KV cache

KV cache能使用所有存储资源

方案：semi-PD技术细节

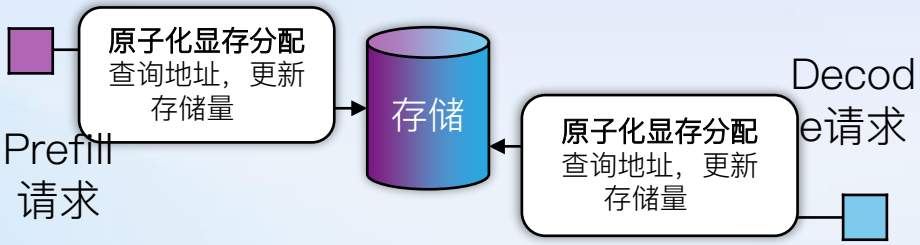
技术细节：计算分离 & 存储融合

计算分离



P/D阶段的计算分别由不同的进程（抽象为worker）完成
进程间实现SM粒度的计算资源划分

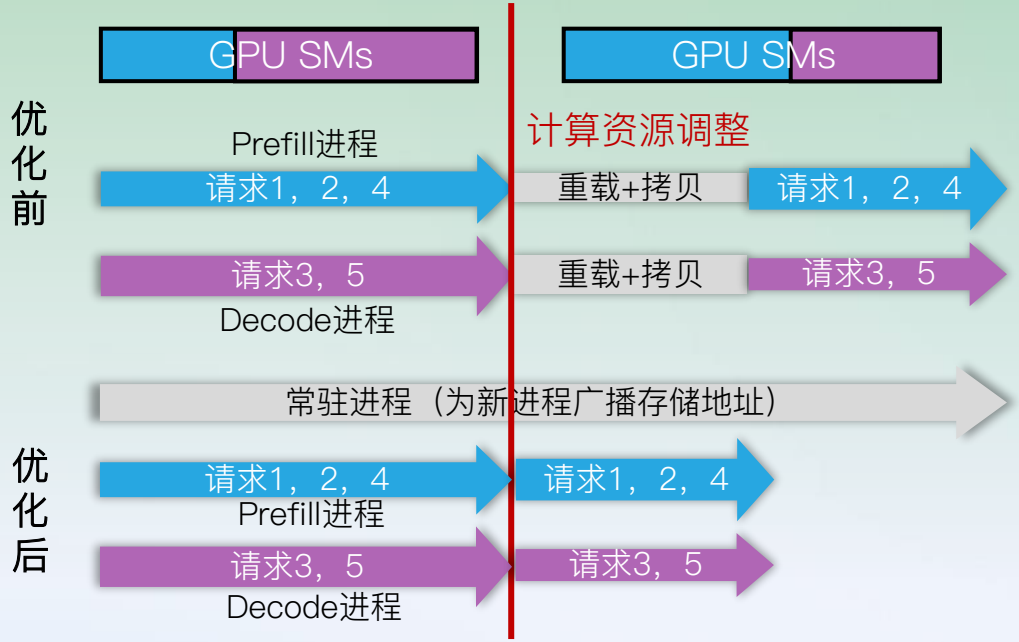
存储融合



通过原子化的显存分配避免P/D worker显存管理的冲突

技术细节：低开销资源调整机制

挑战 P/D间资源调整模型权重重载、KV cache拷贝开销，造成请求阻塞

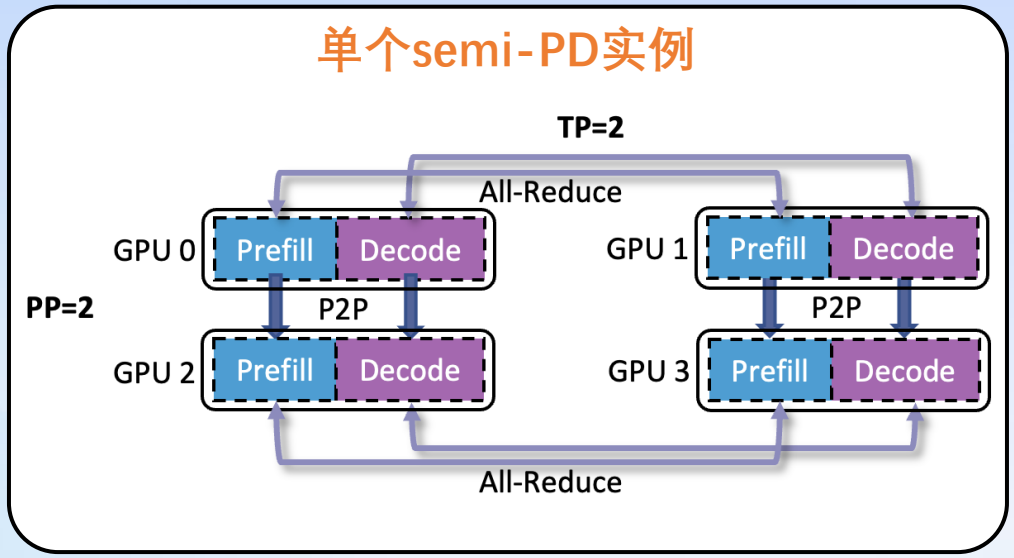


方案 引入常驻进程管理模型权重和KV cache存储，无需重载和拷贝

方案：semi-PD应用适配

实例推理场景

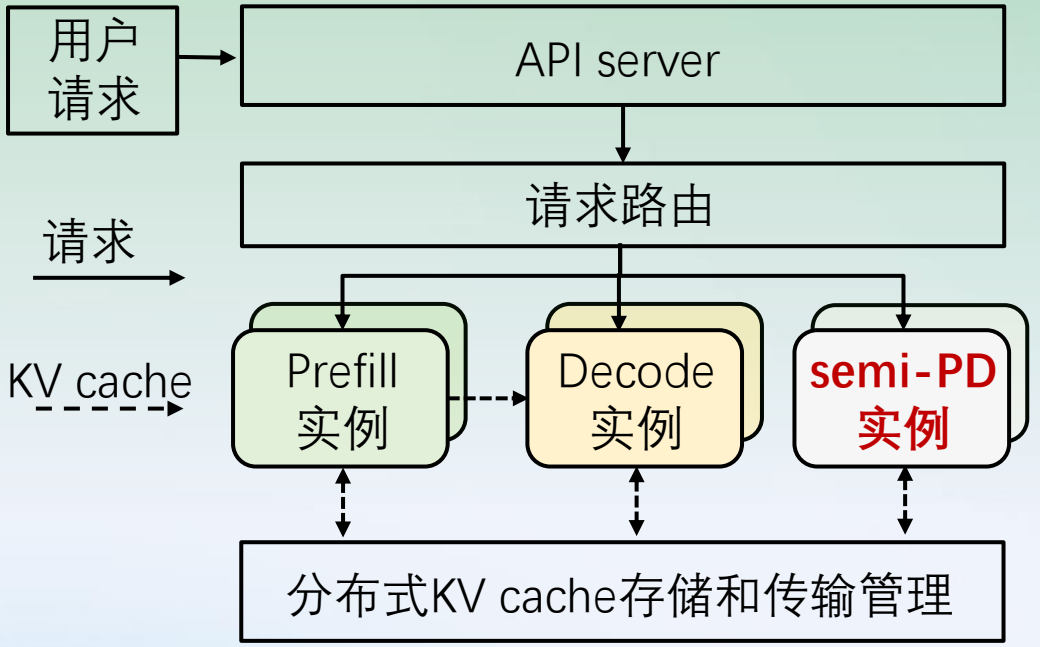
每张GPU上的**P/D worker独立、异步**地与并行组进行**通信**，通信总量不变



*以TP=2, PP=2为例

集群推理场景

semi-PD实例作为**可选实例之一**，与P/D实例共同参与请求路由

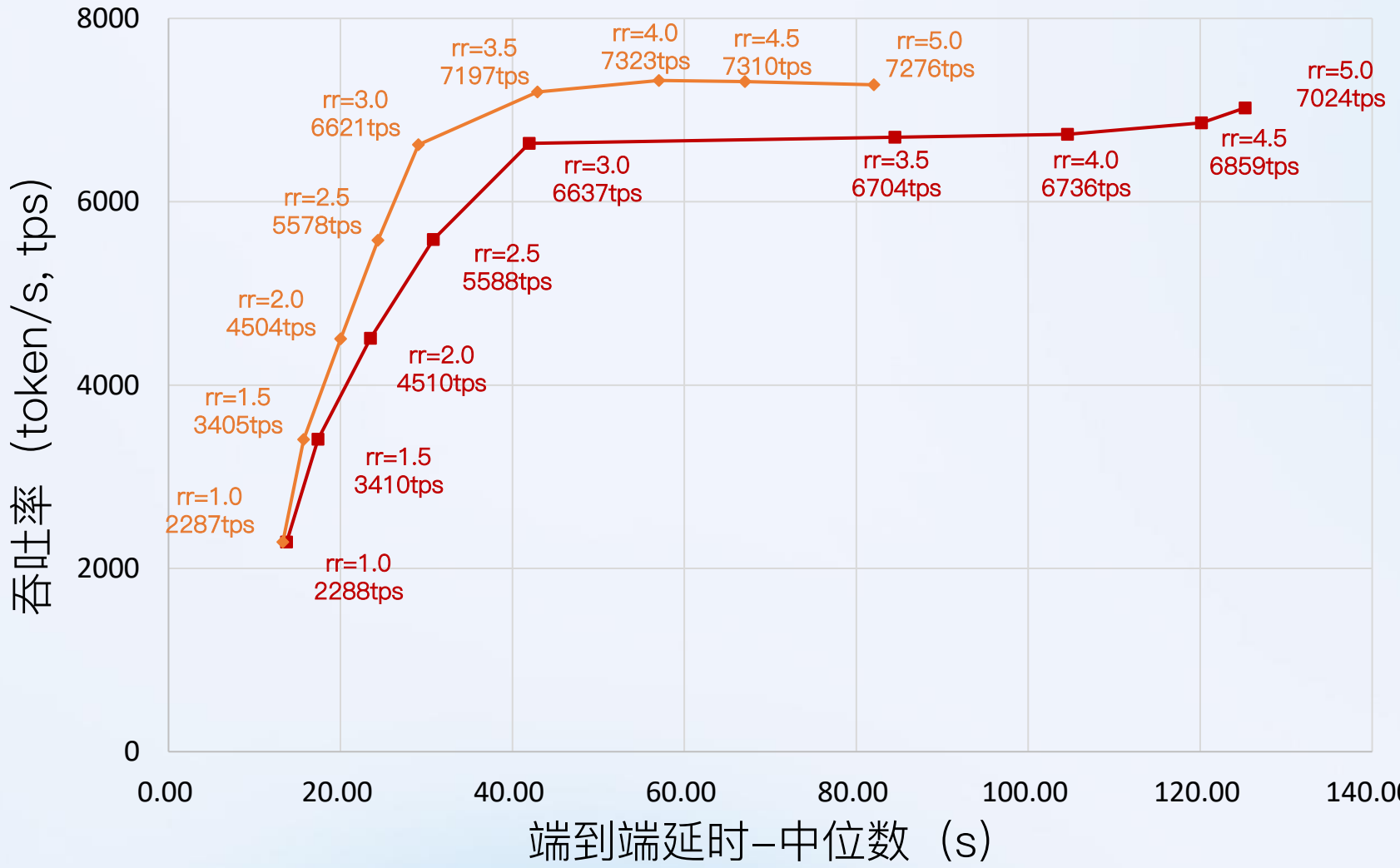


效果：更低的延时 更高的吞吐率

semi-PD+SGLang结果
SGLang结果

测试配置
DeepSeek-V3模型
FP8, TP8, H200
输入均值2k, 输出均值256

同时取得10%的吞吐率提升
和2倍的延时降低



效果：首Token延时和Token平均延时对比

semi-PD+SGLang结果

SGLang结果

测试配置

DeepSeek-V3模型

FP8, TP8, H200

输入均值2k, 输出均值256

P99 TTFT

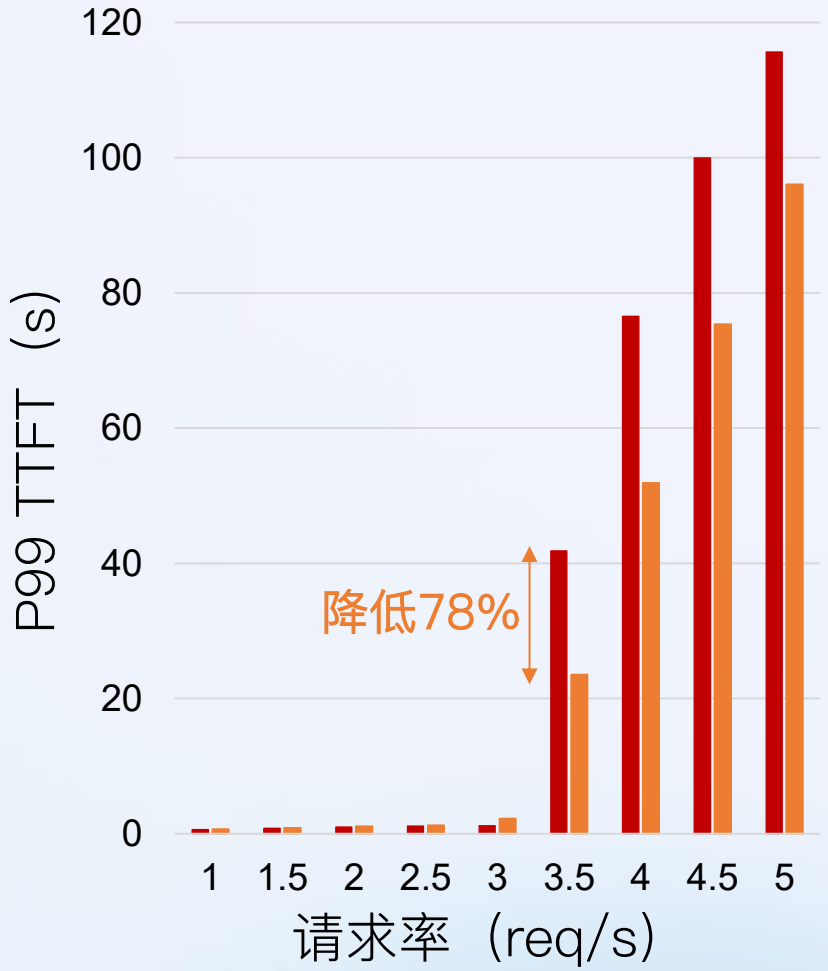
系统中99%请求的

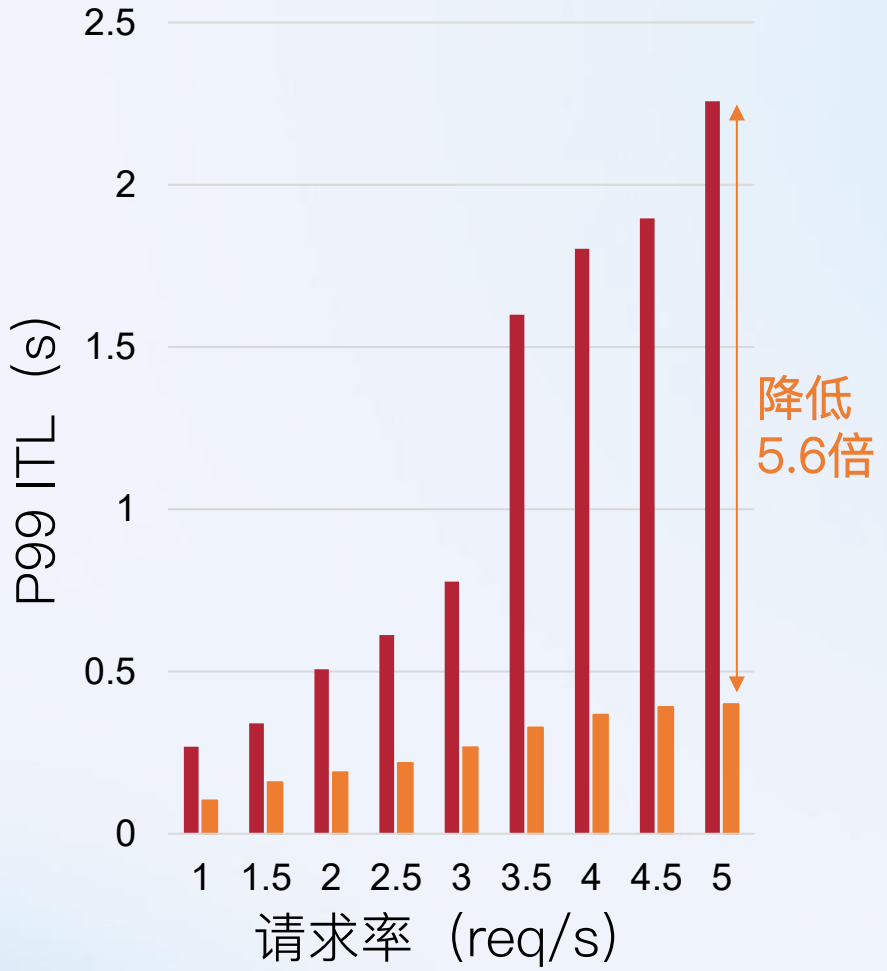
首Token延时不超过的值

P99 ITL

系统中99%请求的

Token间平均延时不超过的值





效果：请求完成率实时对比

semi-PD+SGLang结果

SGLang结果

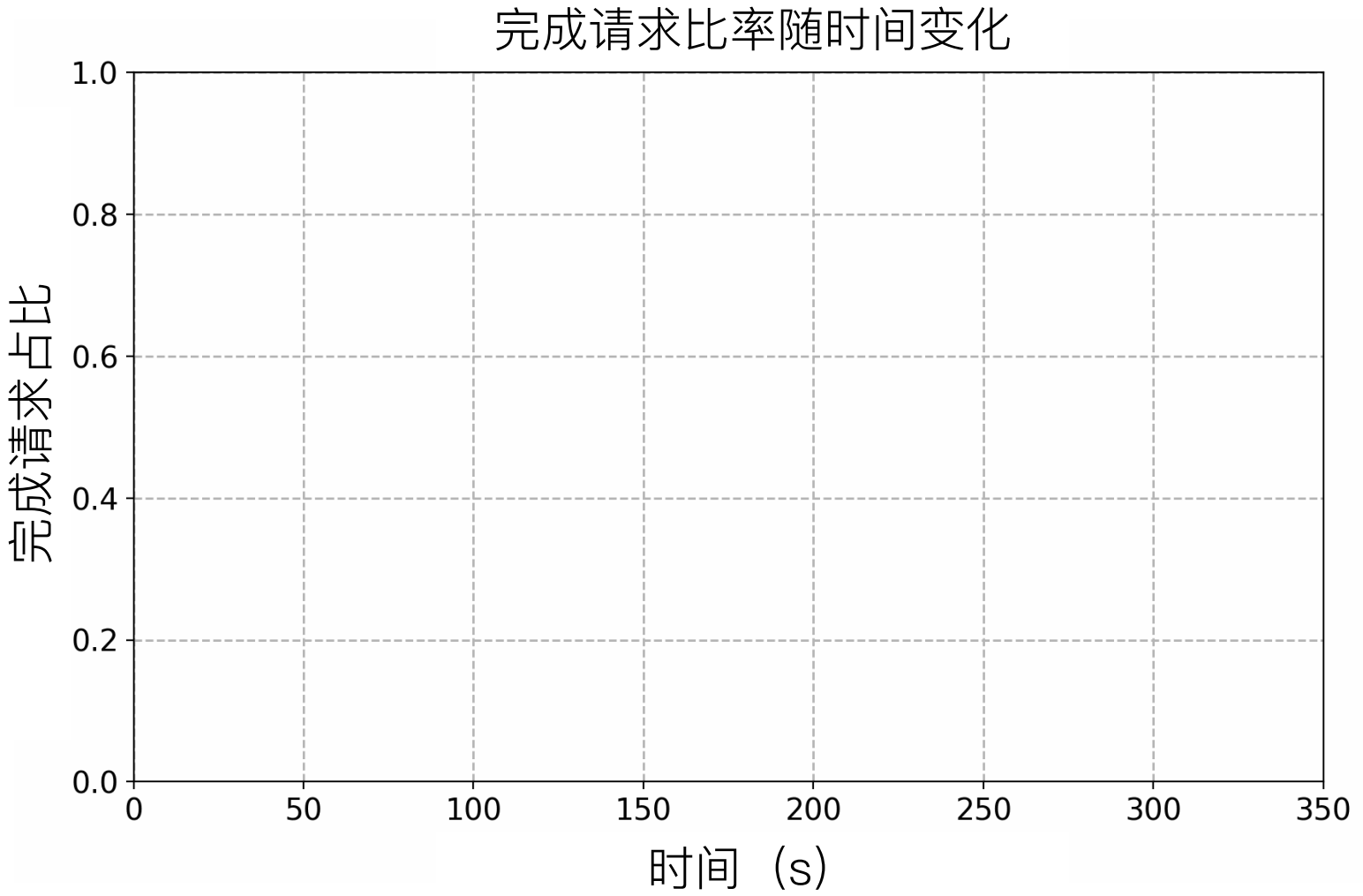
测试配置

DeepSeek-V3模型

FP8, TP8, H20

输入均值2k, 输出均值256

输入请求率=4.0 request/s

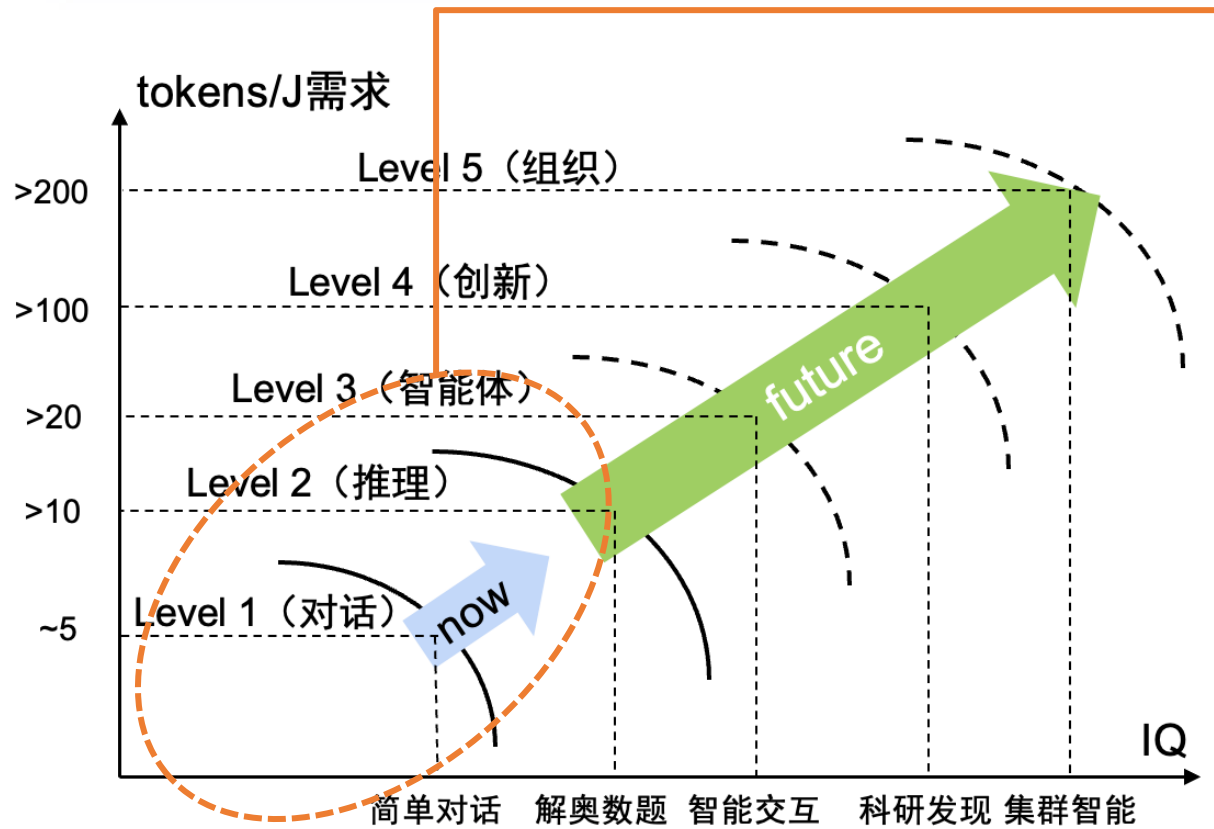


03

面向华为昇腾的推理优化部署实践

推理部署趋势的关键词

关键词：超节点、大EP、长文本



智能变化	推理部署
① 模型规模↑ Llama3系列：405B DeepSeek-V3：671B Kimi-K2：1T…	“超节点” 多卡间通信瓶颈 需要超节点架构
② MoE专家数目↑ Mixtral-8x7B：8个专家 DeepSeek-V3：256个专家 Kimi-K2：384个专家…	“大EP” 专家拆分至不同卡实现高并发计算
③ 上下文长度↑ 对话：<1k Test-time scaling[1]：~8k Agent[2]：>50k	“长文本” 多卡间通信瓶颈 需要超节点架构

[1] DeepSeek team, “DeepSeek Technical Report”, *arXiv preprint arXiv. 2412.19437* (2024).
[2] Q. Wu et al, “AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation”. *arXiv preprint arXiv. 2308.08155* (2023).

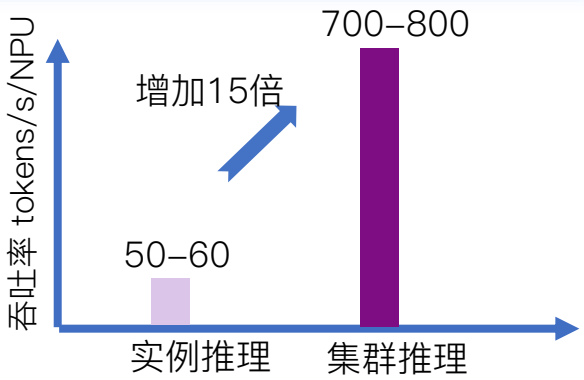
昇腾怎么从能用到好用?

能用 → 好用

首个支持大EP、PD分离功能的国产芯片

910B


华为昇腾服务器 DeepSeek V3/R1 推理部署最佳实践



910C


Serving Large Language Models on Huawei CloudMatrix384

Method	Batch Size	KV Cache Length	TPOT (ms)	Throughput (tokens/s)
DeepSeek (Blog) on H800	N/A	4,989	~50.0	1,850
DeepSeek (Profile) on H800	128	4,096	~50.2	**2,325**
SGLang (Simu. MTP) on H100	128	4,000	~55.6	2,172
CloudMatrix-Infer	96	4,096	49.4	**1,943**

1943 tokens/s/NPU

离好用仍有距离

1. 长文本支持仍有诸多挑战

- Prefill阶段注意力计算量平方增长
- Decode阶段Vector单元计算压力大
- Attention和MoE对计算访存分歧呈更大剪刀差

2. 超节点高性能、低灵活性

- 大量定制优化，社区和学术界**难以**基于良好的软件基础**自发**开展创造
- CloudMatrix 384**超节点规模**和**模型规模**难以灵活匹配

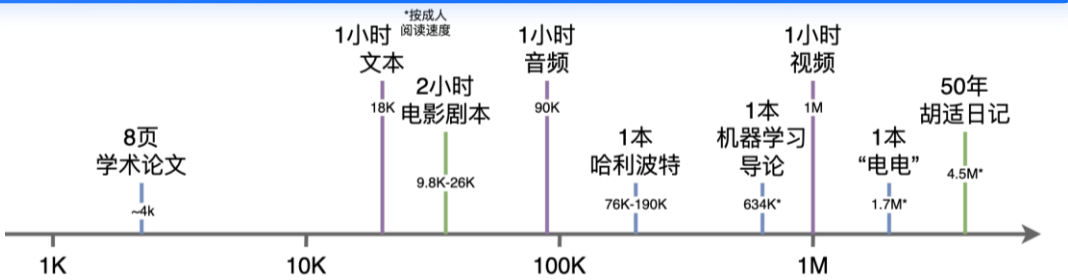
能用：性价比昇腾 vs 英伟达

1）性价比比肩英伟达；2）910B性价比不低于910C

Method	Batch Size	KV Length	TFLOPS	TPOT	Throughput	Throughput per TPLOPS	Max Power per Node (KW)	Max Power per GPU(KW)	Tokens per J
DeepSeek (Blog) on H800	N/A	4,989	1979 (FP8)	~50.0	1,850	0.93	10–13	1.25–1.63	1.48
DeepSeek (Profile) on H800	128	4,096	1979 (FP8)	~50.2	2,325	1.17	10–13	1.25–1.63	1.86
SGLang (Simu. MTP) on H100	128	4,000	1979 (FP8)	~55.6	2,172	1.10	10–13	1.25–1.63	1.74
CloudMatrix–Infer	96	4,096	1504 (INT8)	49.4	1,943	1.29	15–16	1.87–2	1.03
Atlas 800I A2 (910B)	–	–	626 (INT8)	100.0	766	1.22	5–6	0.63–0.75	1.21

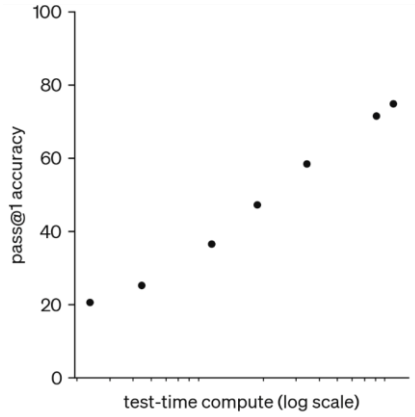
好用：长文本推理仍有诸多挑战

长文本的需求：数据、应用与能力



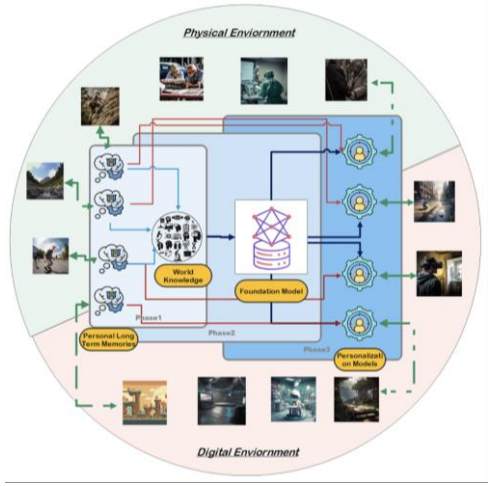
推理能力

回应用户前产生**很长的内部思维链**



随着思考时间（思维链总长度）增加，
强推理模型表现会**持续提高**[1]

具身智能

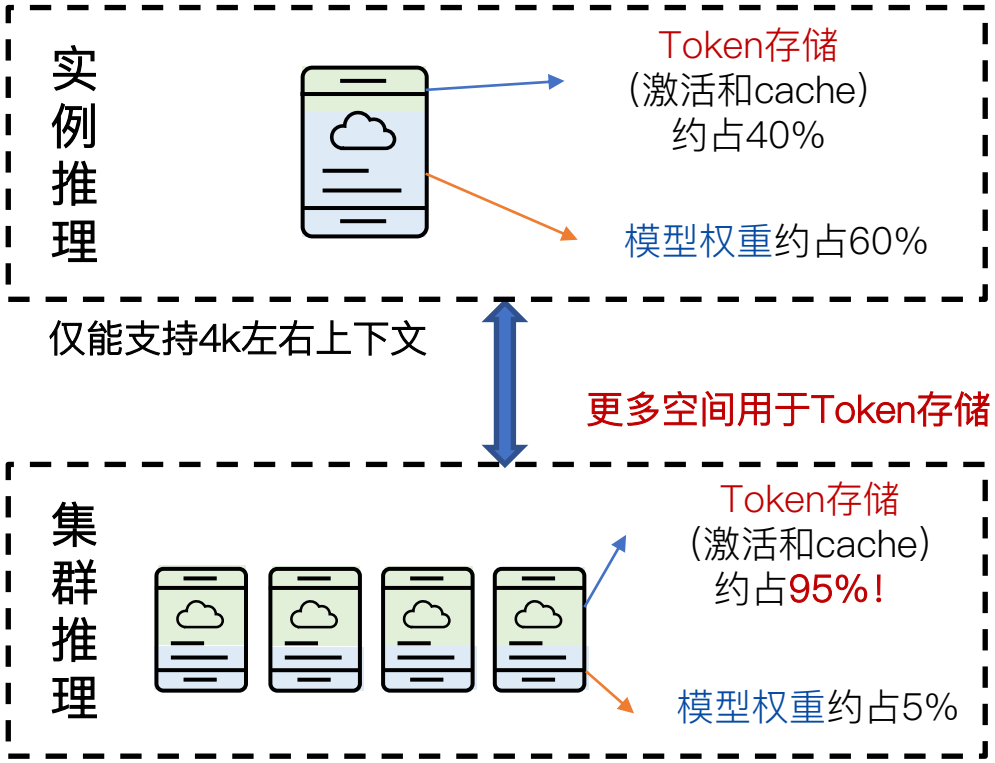


有**长程记忆**的模型通过与环境的交互，
逐渐发展出**新的认知能力** [2]

集群使高效长文本推理成为可能

单Token的内存占用：约1MB

*DeepSeek-V3，考虑激活存储复用



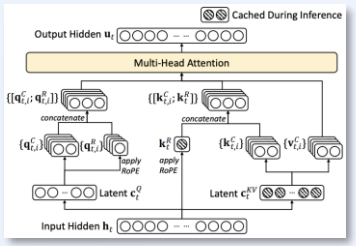
[1] OpenAI. ”Learning to Reason with LLMs.” OpenAI, www.openai.com/index/learning-to-reason-with-llms/. Released 12 Sept. 2024.
[2] Jiang, Xun et al. “Long Term Memory: The Foundation of AI Self-Evolution.” (2024).

长文本场景带来的问题跃迁

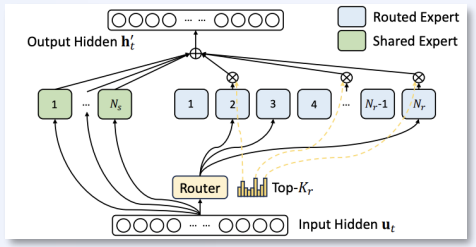
华为昇腾服务器推理部署最佳实践 ☺

长文本场景的新挑战 ☹

计算



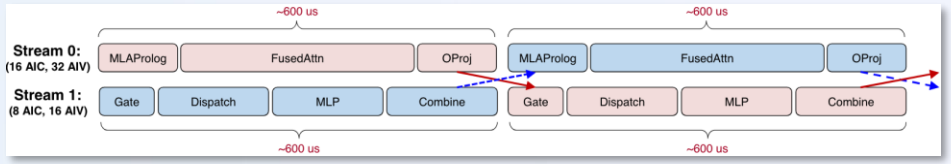
MLA算子优化



MoE算子优化

Prefill阶段注意力计算：呈二次方增长
思路：集群的多卡协作优势

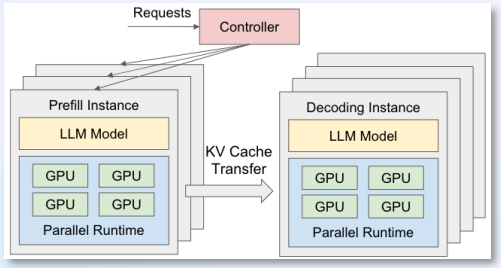
通信



计算-通信重叠

序列并行引入权重和KV Cache通信
思路：基于超节点优化并行和任务编排策略

框架



P/D分离式系统

模型规模和集群规模匹配失衡、
Attention和MoE对计算访存需求失衡
思路：利用UB搭建的微服务架构

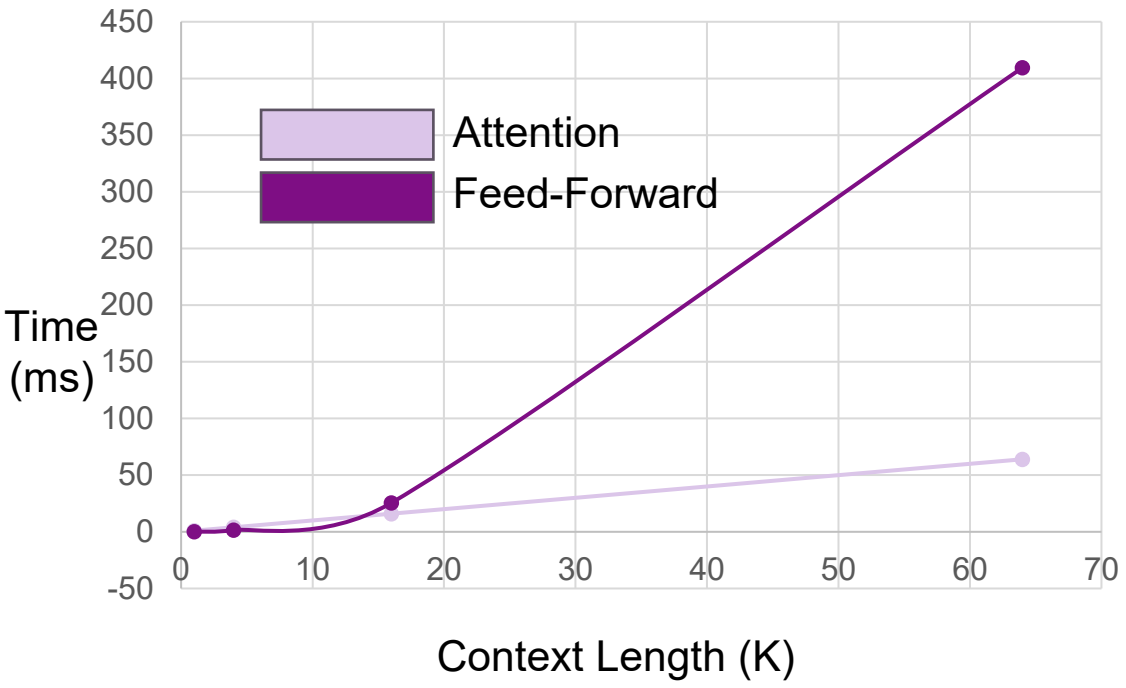
上下文长度：2k~4k → **>100k**

面向长文本场景的集群推理

问题1

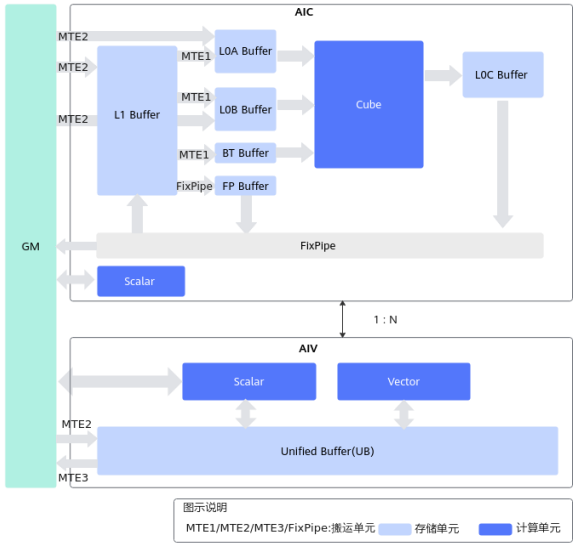
长文本下的计算问题：Attention成为瓶颈对标量核压力增加

Prefill阶段Attention计算和上下文呈二次方



向量计算在昇腾芯片可能成为新瓶颈

随长度增加，标量计算对算子需求迅速增长，昇腾AIC、AIV算力悬殊问题凸显



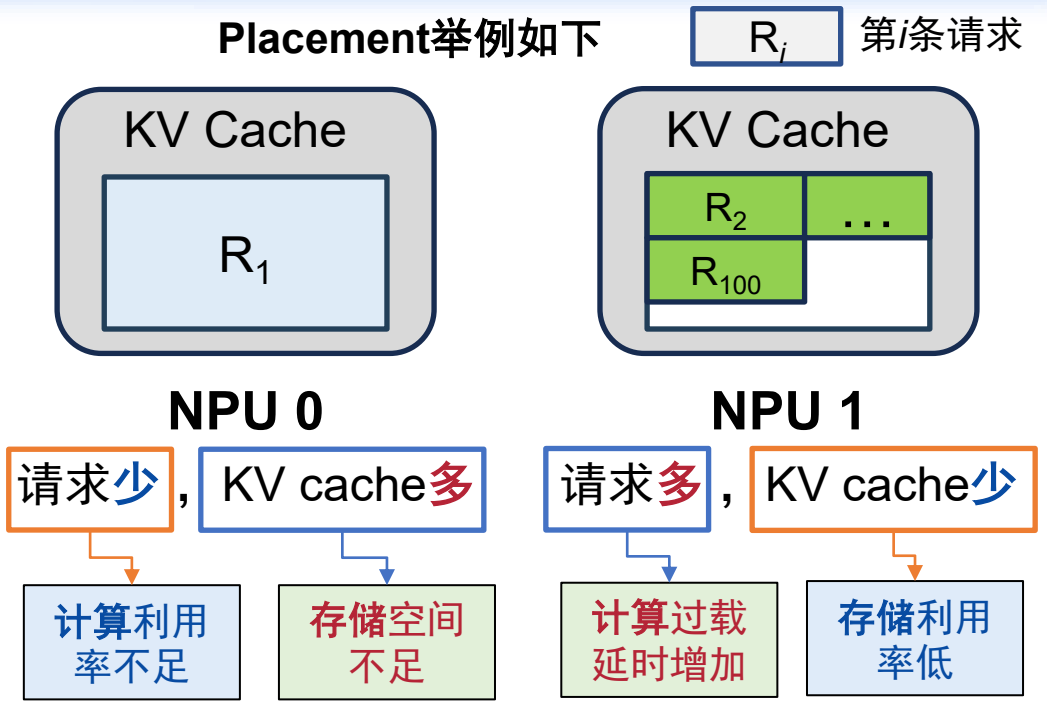
AIV算力紧张，如何协同AIC和AIV成为新问题

面向长文本场景的集群推理

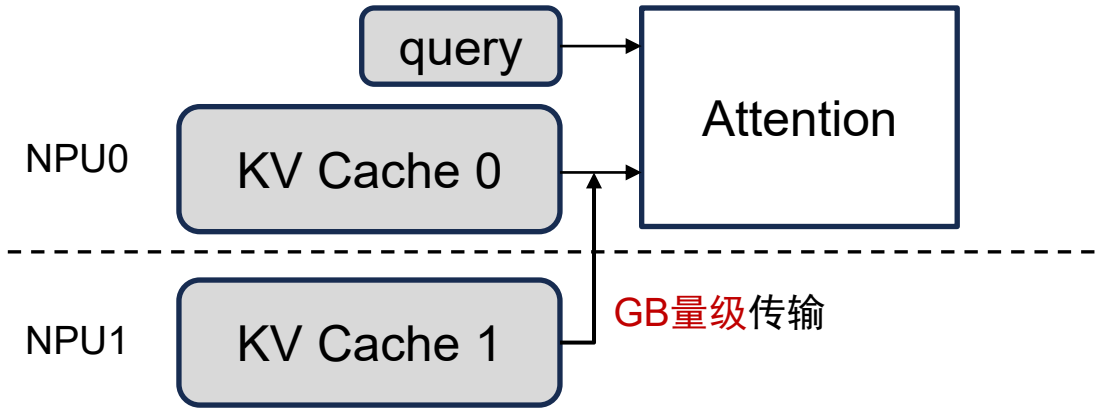
问题2

长文本加剧KV Cache存储均衡问题，传输KV Cache成为新瓶颈

不同卡KV Cache的Placement导致计算和存储均衡问题



多卡间KV Cache传输通信量大，如何掩盖KV Cache传输延时



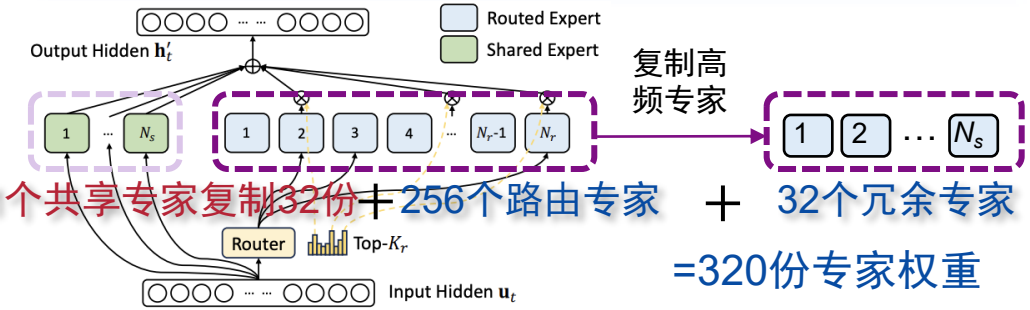
长文本KV Cache拆分到多卡存储
注意力计算时，若传KV cache则**数据传输量大**

面向长文本场景的集群推理

问题3

长文本场景加剧了集群推理中的资源匹配的难度

如何灵活自适应匹配节点规模和模型规模

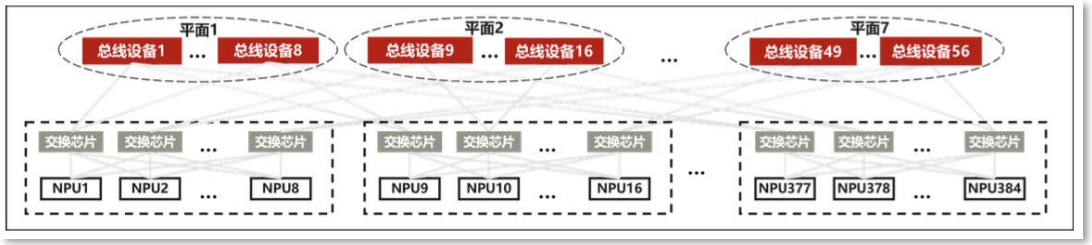


1个共享专家复制32份 + 256个路由专家 + 32个冗余专家

=320份专家权重

专家并行EP=320（解码阶段），即单实例需要160张NPU

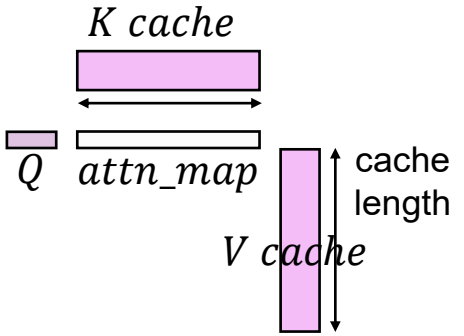
如何匹配？换个模型又该如何



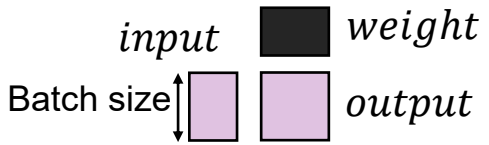
超节点CloudMatrix384，共384张NPU

如何匹配注意力和MoE所需计算资源

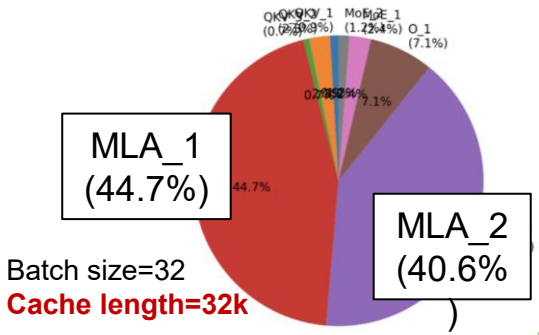
注意力部分对计算需求高、MoE阶段对访存需求高



注意力部分计算富集：计算压力大



MoE部分显存需求富集：存储压力大



Batch size=32

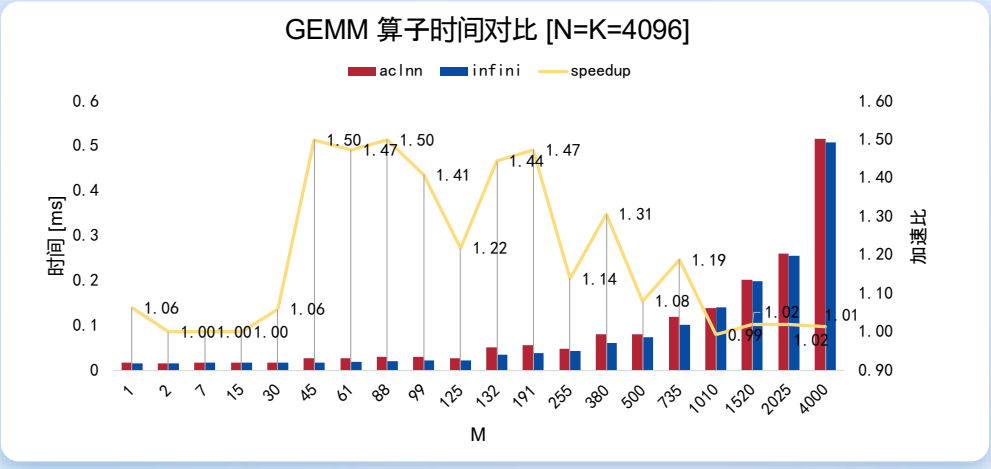
Cache length=32k

我们在昇腾基础软件上的技术积累

计算：GEMM/Group GEMM 算子优化

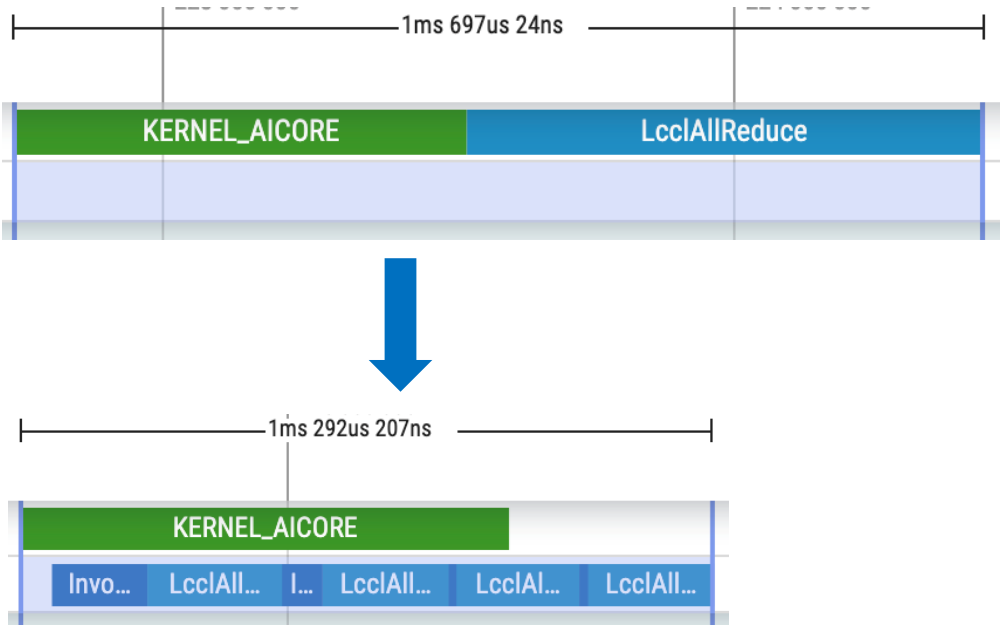
- 基于CCE（底层编程语言）实现更精细的计算、数据搬运和流水编排
- 精细的L2->L1-L0数据搬运和复用策略
- GEMM算子相比aclNN优化性能提升高达1.5倍

Llama3-8B典型矩阵乘法形状相比aclNN提升高达1.5倍

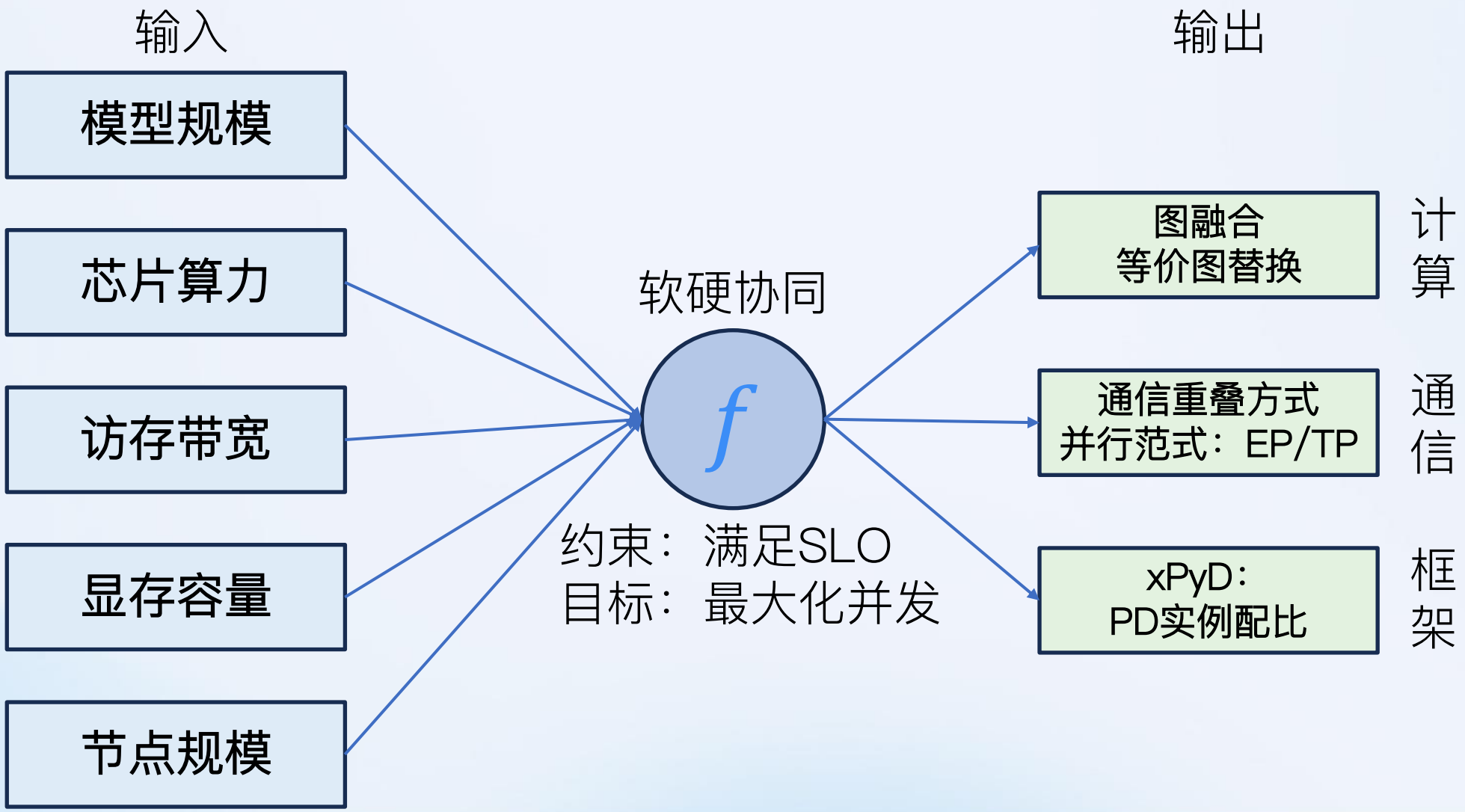


通信：计算通信重叠FlashOverlap

- 基于控制信号实现计算和通信双流并行且解除数据依赖
- 联合昇腾高性能计算库catlass和通信库hccl完成实践
- 相比顺序执行性能提升高达1.35倍



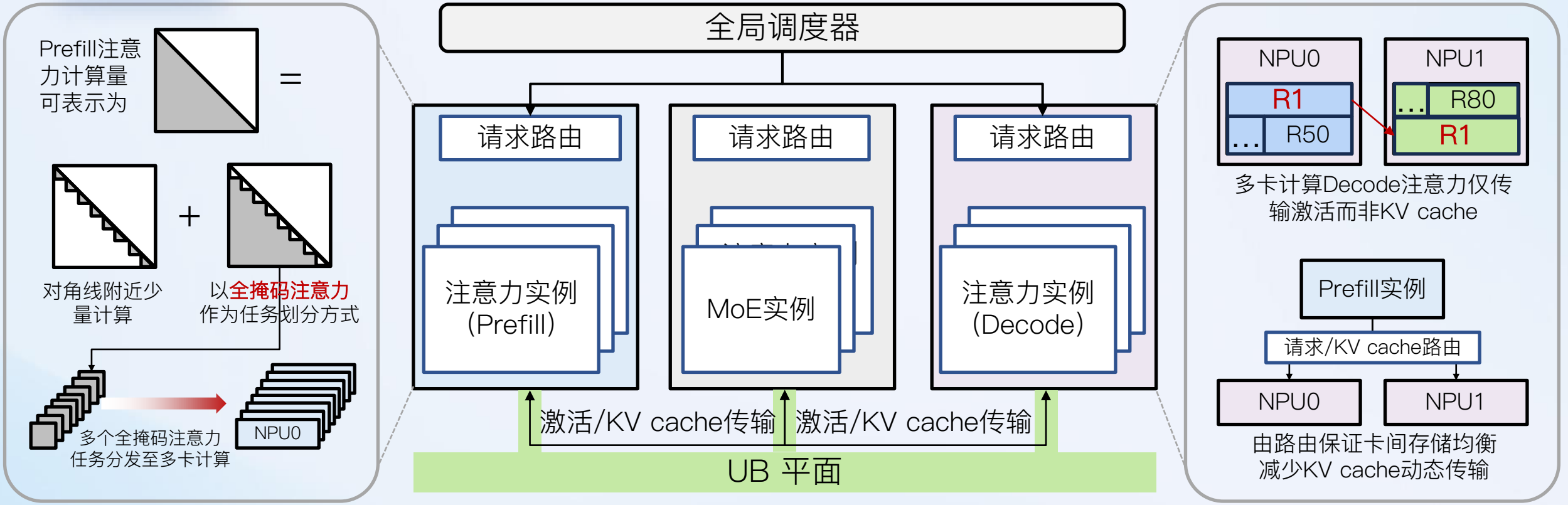
集群推理方法论


输入
模型规模
芯片算力
访存带宽
显存容量
节点规模
软硬协同
f
约束：满足SLO
目标：最大化并发
输出
图融合
等价图替换
计算
通信重叠方式
并行范式：EP/TP
通信
xPyD：
PD实例配比
框架

长文本场景的集群推理整体方案

思路

计算、通信、框架设计实现通信/访存量降低和灵活资源匹配



计算：全掩码注意力任务分发
提升卡内注意力计算效率

框架：以大算子为粒度的微服务架构
实现注意力、MoE的计算需求和超节点规模灵活匹配

通信：KV Cache Placement策略减少KV cache通信

04

以有限算力架构 释放终端应用潜能

端侧应用落地迎接爆发 亟需强推理能力

云端协同释放算力，数十亿终端进入大模型时代

基座模型预训练

模型云端部署

模型微调/推理

算力资源优化

算力协同及数据返回

基座模型及算力调用

端侧协同推理

端侧模型部署

场景增量学习

AI手机年出货
5.52亿台[注2]

AI PC年出货
1.02亿台[注3]

智能汽车年出货
近**5000万台**[注4]

未来泛端侧应用需要>100tokens/s推理能力

泛端侧应用大模型推理性能需求

- 经推测，GPT-4o/o1参数量**超过200B**
- 实际端侧应用需要4o/o1能力的模型实现**100~1000 token/s** 的推理性能

GPT-4o/o1
(~200B+强化学习)

端侧智能机器人

自动驾驶
感知决策
>100 token/s
[Tesla HW3]

无人机
路径规划
>100 token/s
[Nature 2023封面文章]

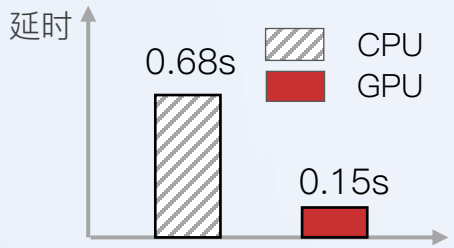
具身智能
动作决策
>1000 token/s
[Jim Fan, NV高级研究科学家]

端侧大模型推理的效率挑战

① 计算挑战

GPU 算力高计算快 ⟷ CPU 算力低计算慢

各部分计算延时占比
(Llama2-7B GPU16层，CPU16层)



计算大模型
CPU速度比GPU速度慢4倍

② 存储挑战

云模型 模型参数多 ⟷ 端设备 内存容量小

内存相差
2个数量级

DeepSeek-R1模型约617GB参数

单机内存仅12GB
无法在本地运行

端侧内存容量与云模型参数量
相差2个数量级

③ 协同挑战

单处理器 无调度低延时 ⟷ 多处理器 通信+调度

GPU Laptop

CPU

多处理器协同计算
需要复杂的通信和调度

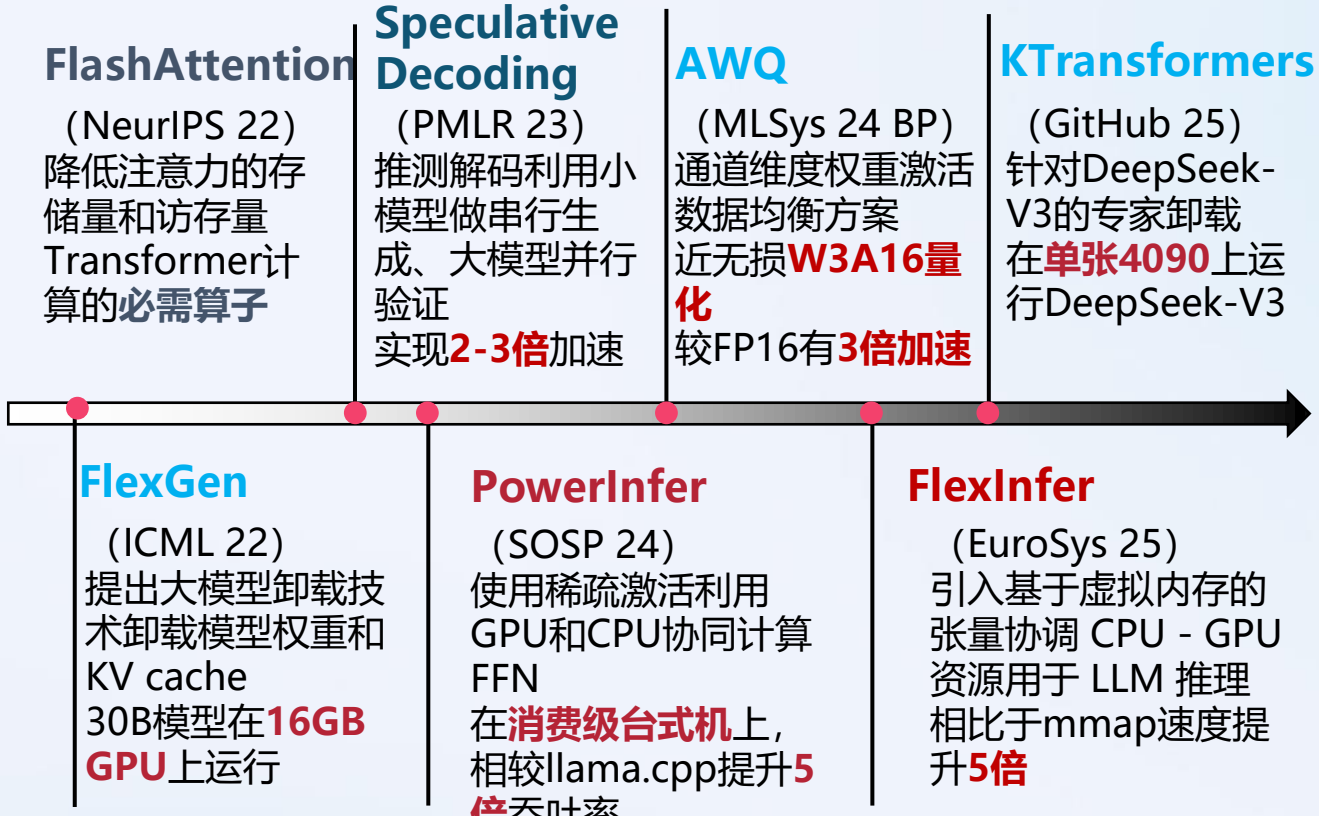
端侧大模型核心技术

端侧应用智能：单用户、少请求

计算挑战
提升生成式模型的计算利用率
→ 投机解码等算法优化 编译调度等系统优化

存储挑战
端侧存储难以容纳参数和中间变量
→ 模型轻量化设计 模型卸载技术

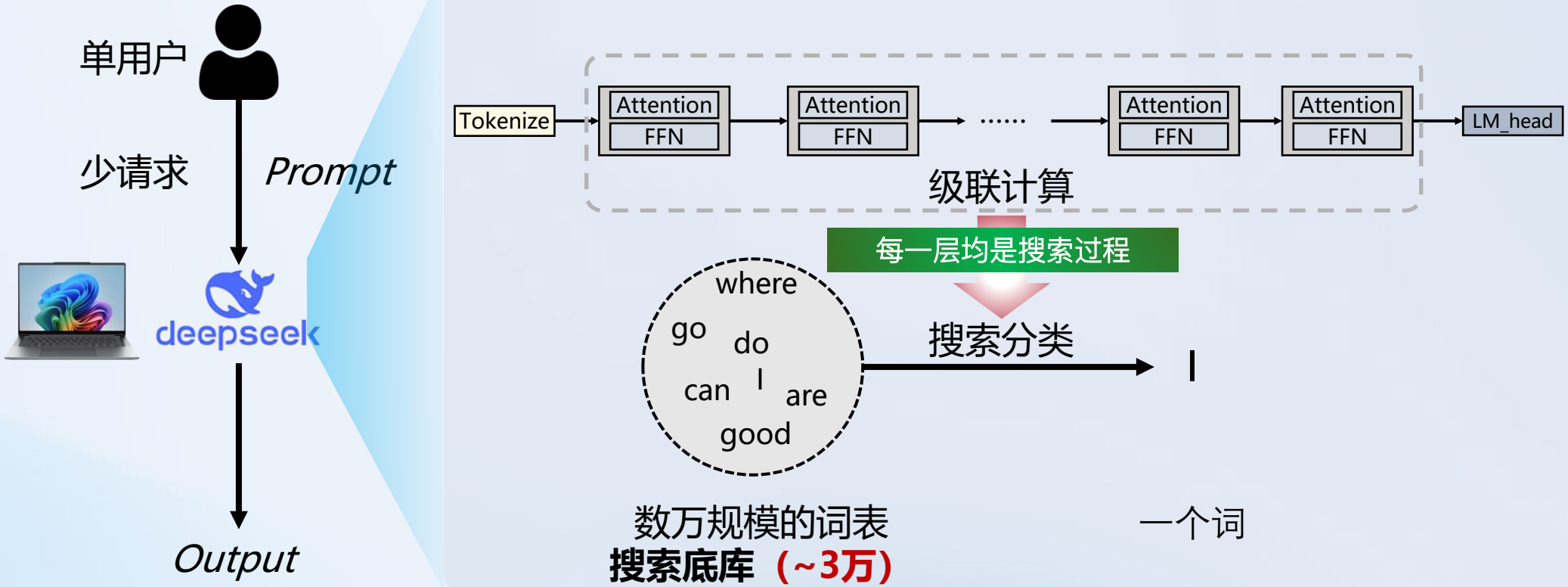
协同挑战
多处理器硬件资源协同处理
→ 任务拆分与并行等异构计算优化



端侧推理应用技术Milestone

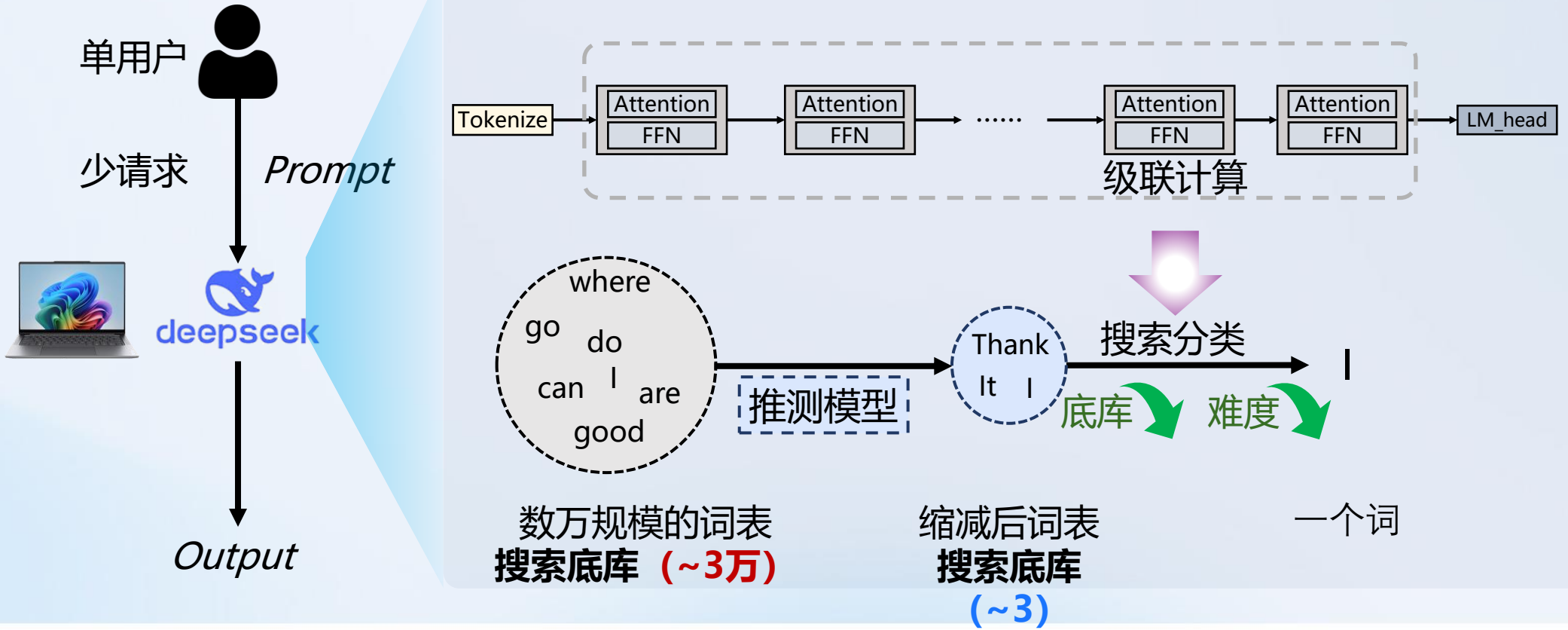
单用户下大模型端侧推理

关键分析：单用户下大模型推理是底库很大的搜索分类问题



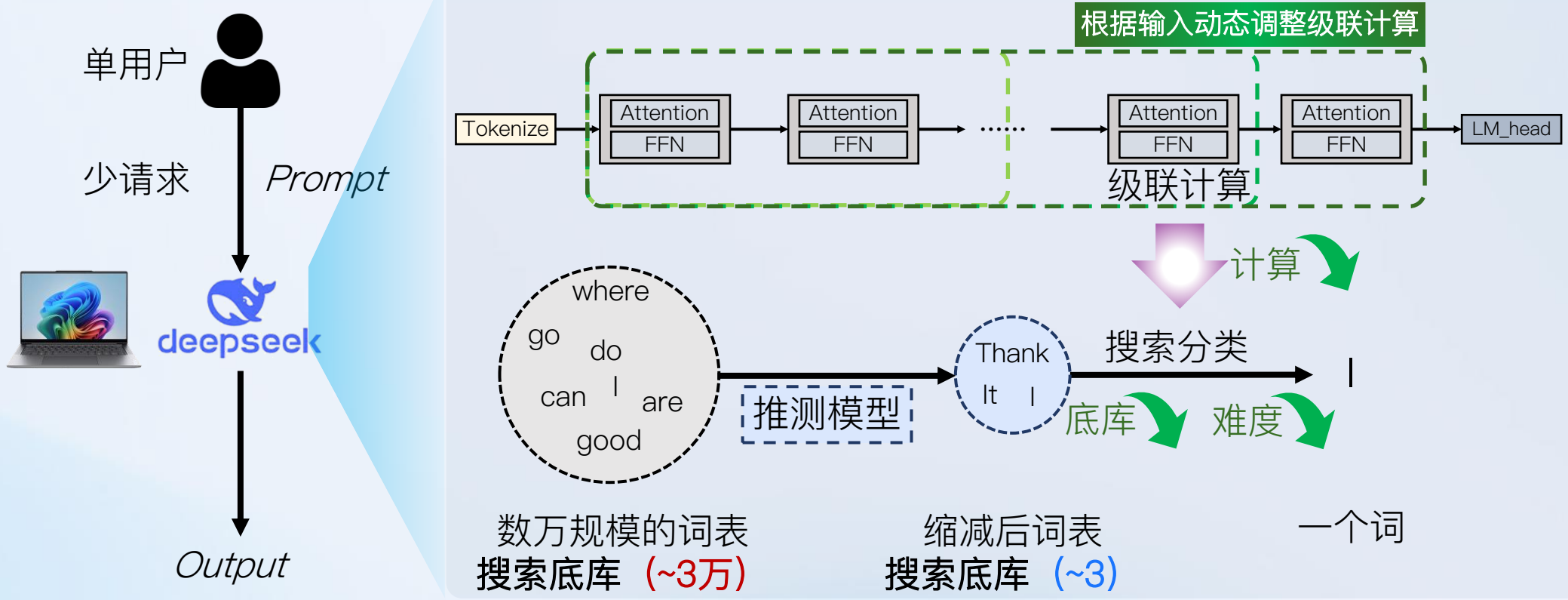
端侧推理引擎：SpecEE

核心思想：利用小参数的LLM作为推测模型缩减搜索底库



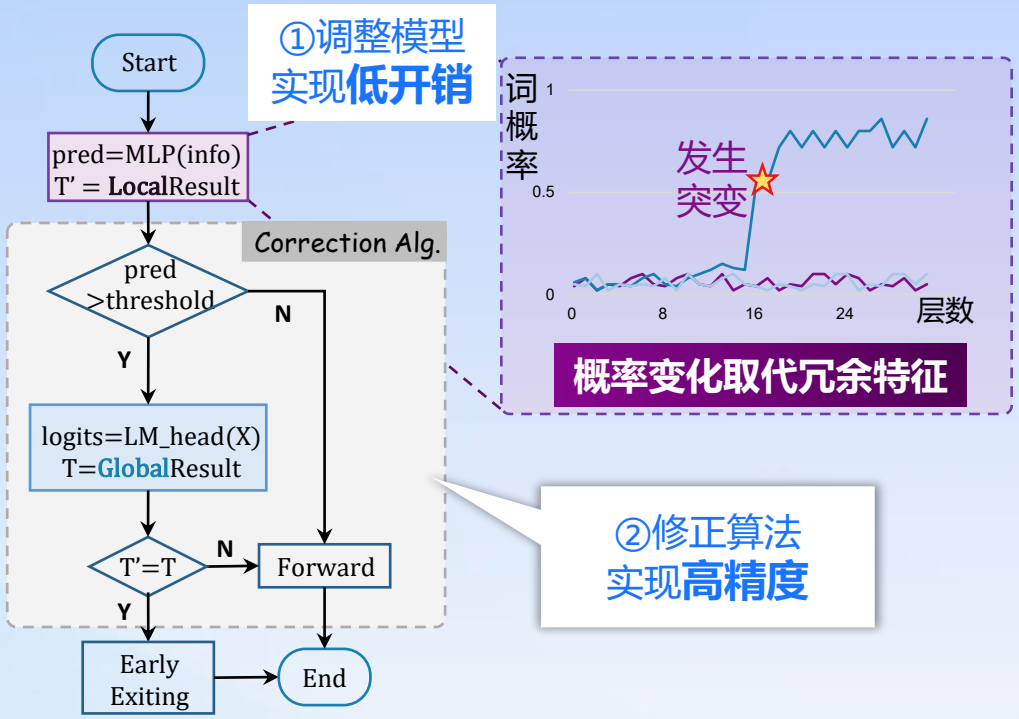
端侧推理引擎：SpecEE

技术核心：基于缩减后搜索底库动态调整单用户下级联计算



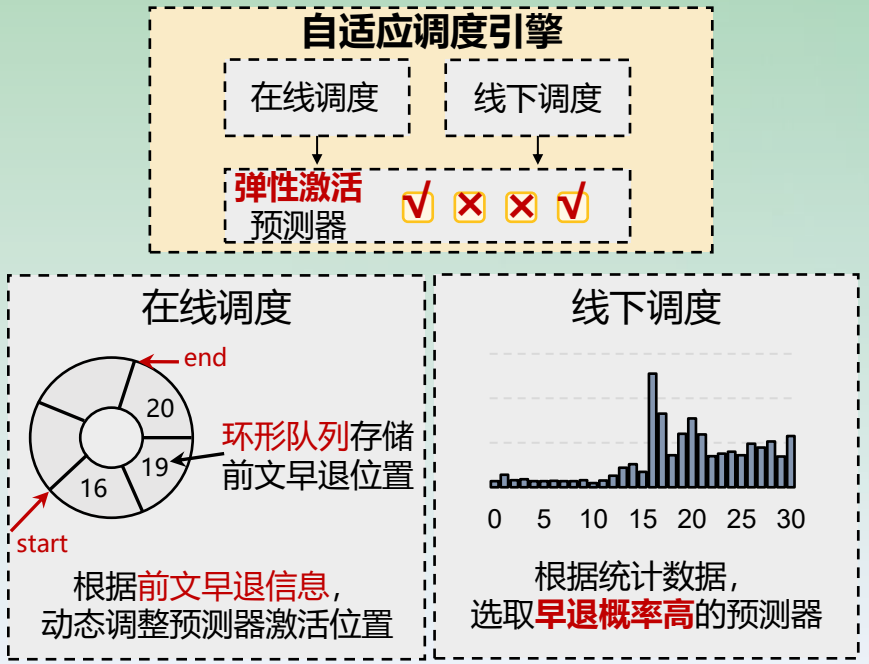
SpecEE技术细节：如何高效动态调整级联计算

算法：低开销高精度调整级联计算



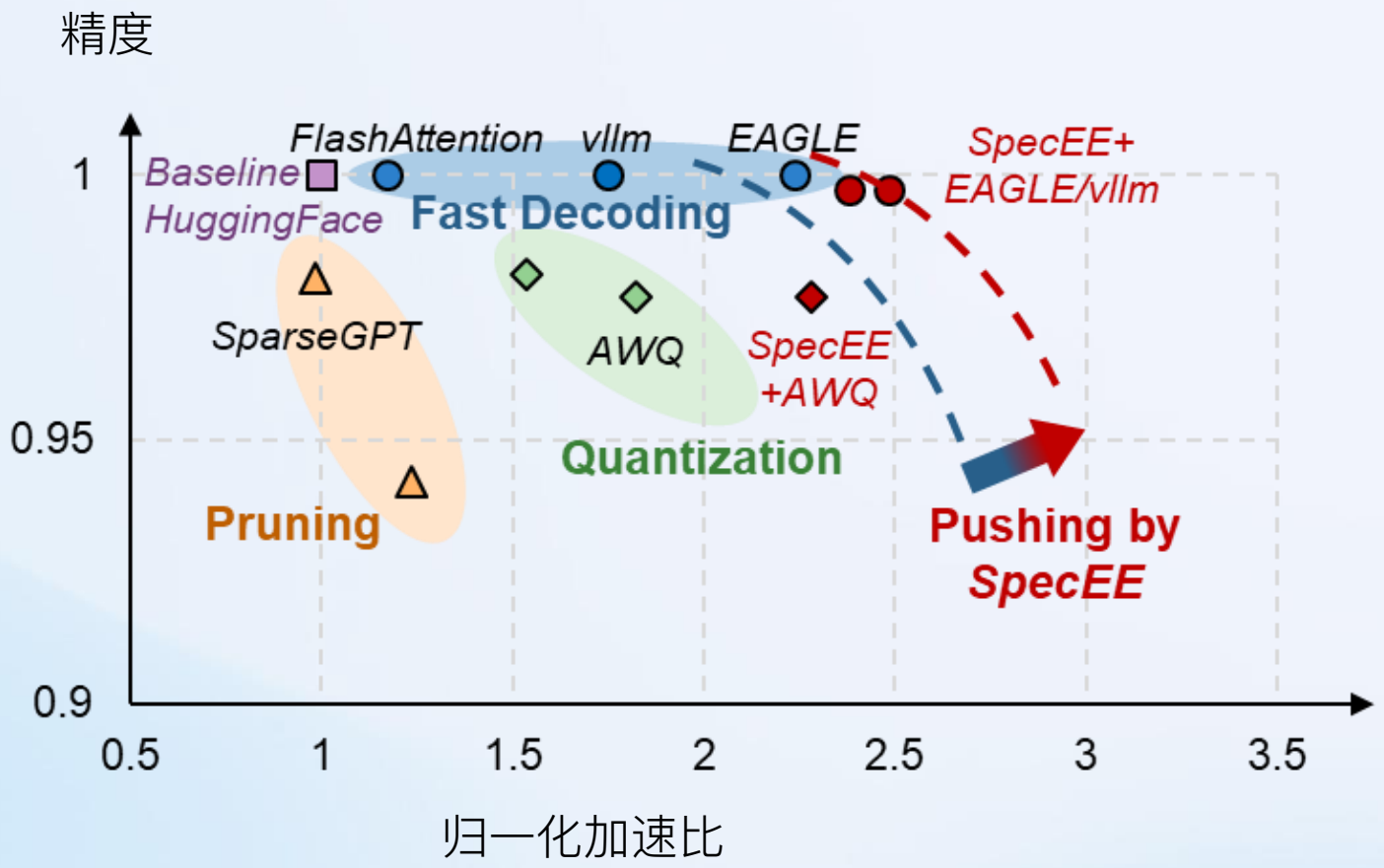
预测模型与修正算法双重保障实现低开销高精度

系统：动态调度引擎适配端侧部署



引入自适应调度引擎实现弹性激活调整机制

SpecEE结果：推进速度与精度的帕累托前沿



SpecEE

Explain key principles in evaluating an argument in analytical writing in details.

llama.cpp

速度更快

精度更高

05

以大模型推理技术创新 融合人工智能产业创新

破解推理Scaling时代 大模型推理系统的效率挑战

端设备

优化目标：降低延时

核心特点：单用户、少请求

云平台

优化目标：满足用户延时要求，提升吞吐率

核心特点：多用户、多请求

计算挑战
提升生成式模型的计算利用率

小规模负载计算
(利用率：
50%→90%)

大规模负载并行计算

协同挑战
多处理器硬件资源协同处理

重叠传输
延时开销

调度挑战
不同用户间资源竞争及延时需求

满足用户目标延时

存储挑战
同时存储大量请求的KV cache

存储需求
(存储量
100GB→10GB)

提升显存利用效率
(利用率
70%→90%)

推理系统上云 打造云端AI产业生态创新引擎

高效集群推理，加速大模型能力注入千行百业

猎聘 · 招商银行 CHINA MERCHANTS BANK · 极狐 GitLab · 看见概念 · OPPLE 欧普照明 · initi:AI 自由量级 · FUZFL · 中国科学院深圳先进技术研究院 SHENZHEN INSTITUTE OF ADVANCED TECHNOLOGY CHINESE ACADEMY OF SCIENCES

MOOTION · 蜂鸟 AI · FUXI·AI 伏羲智能 · 万能小in

等近百家AI应用/企业排队入驻中

徐汇模速空间算力生态平台

投入运营

北京海淀公共算力服务平台

投入运营

杭州市算力资源服务平台

投入运营

登上央视1套《新闻联播》

云平台

cloud.infini-ai.com 大模型服务平台

semi-PD

全球首创第三代推理集群系统，推理性能行业第一

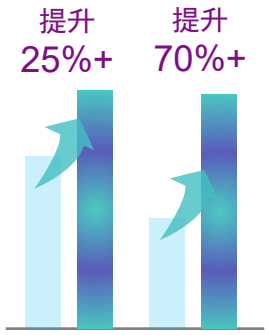
推理系统入端 释放智能终端应用潜能

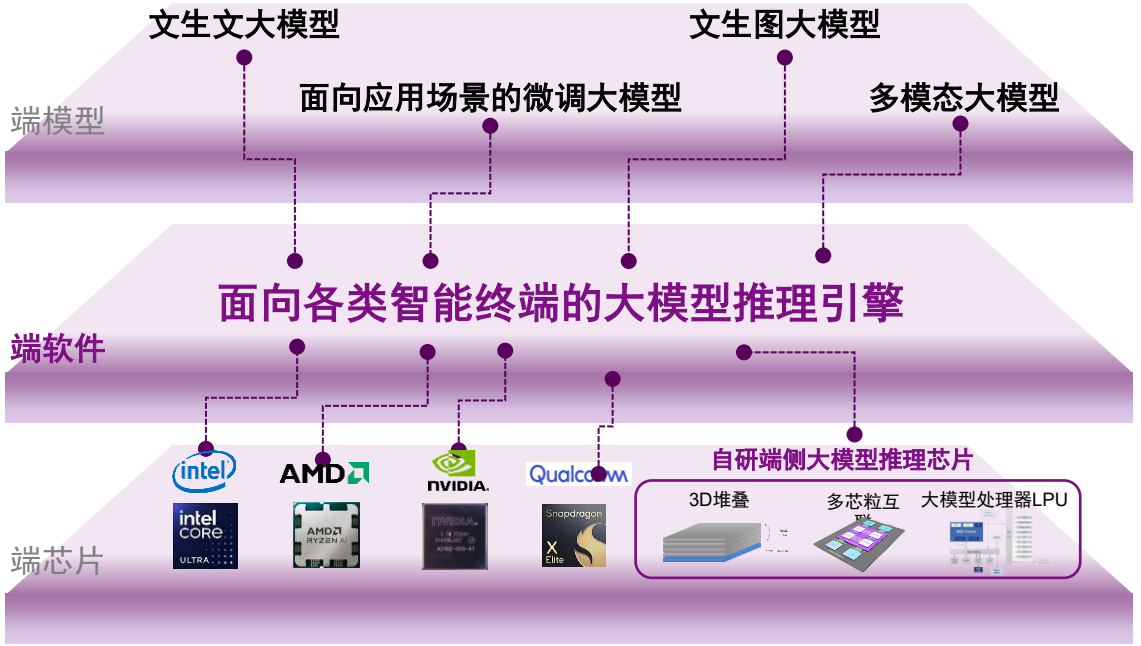
高效端侧推理，软硬协同加速下一代智能终端落地

大模型推理引擎
端到端性能提高 **70%**

YOGA Pro 14 元启版

ThinkBook 14+ 元启版





- 量产软件合作
- 签订战略合作备忘录
- 建立联合实验室
- 联合定义终端场景

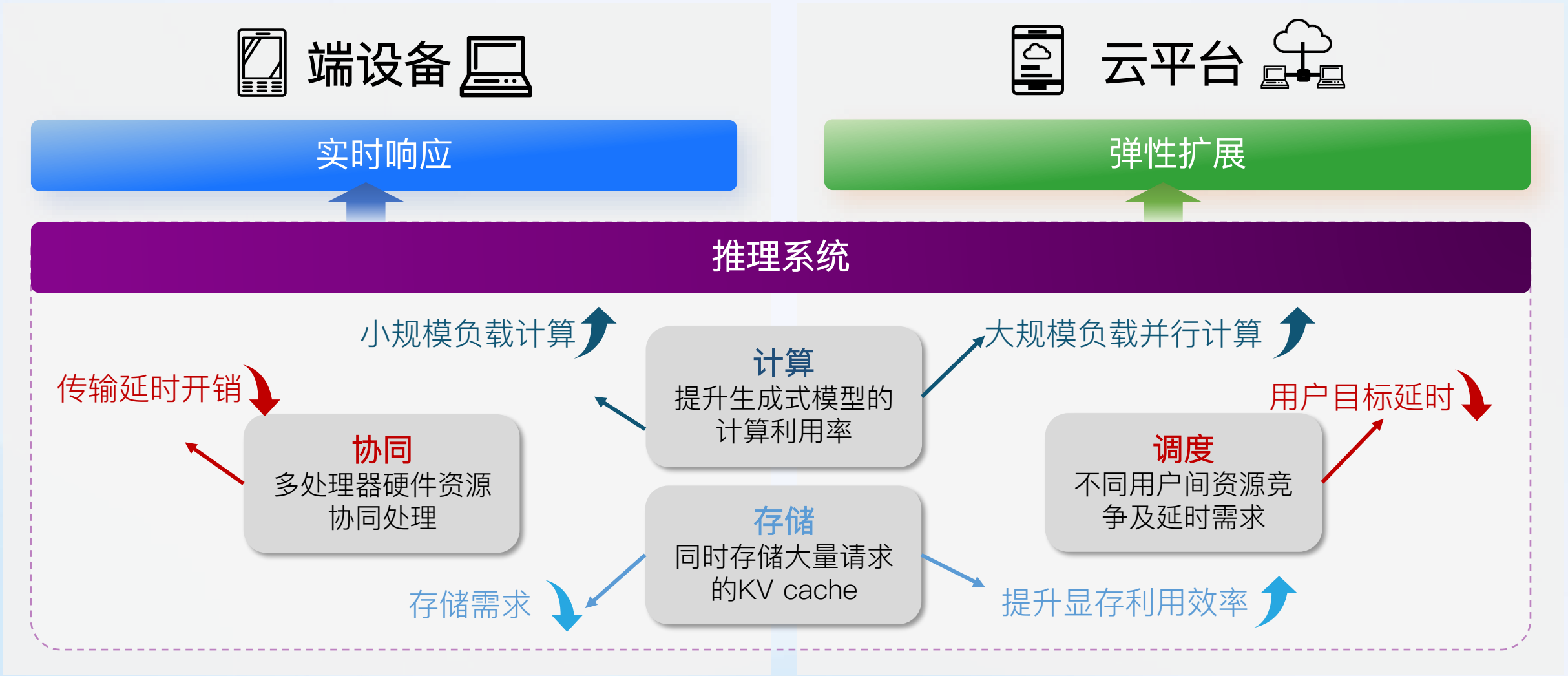
端设备

“端模型+端软件+端IP”智能终端一体化解决方案

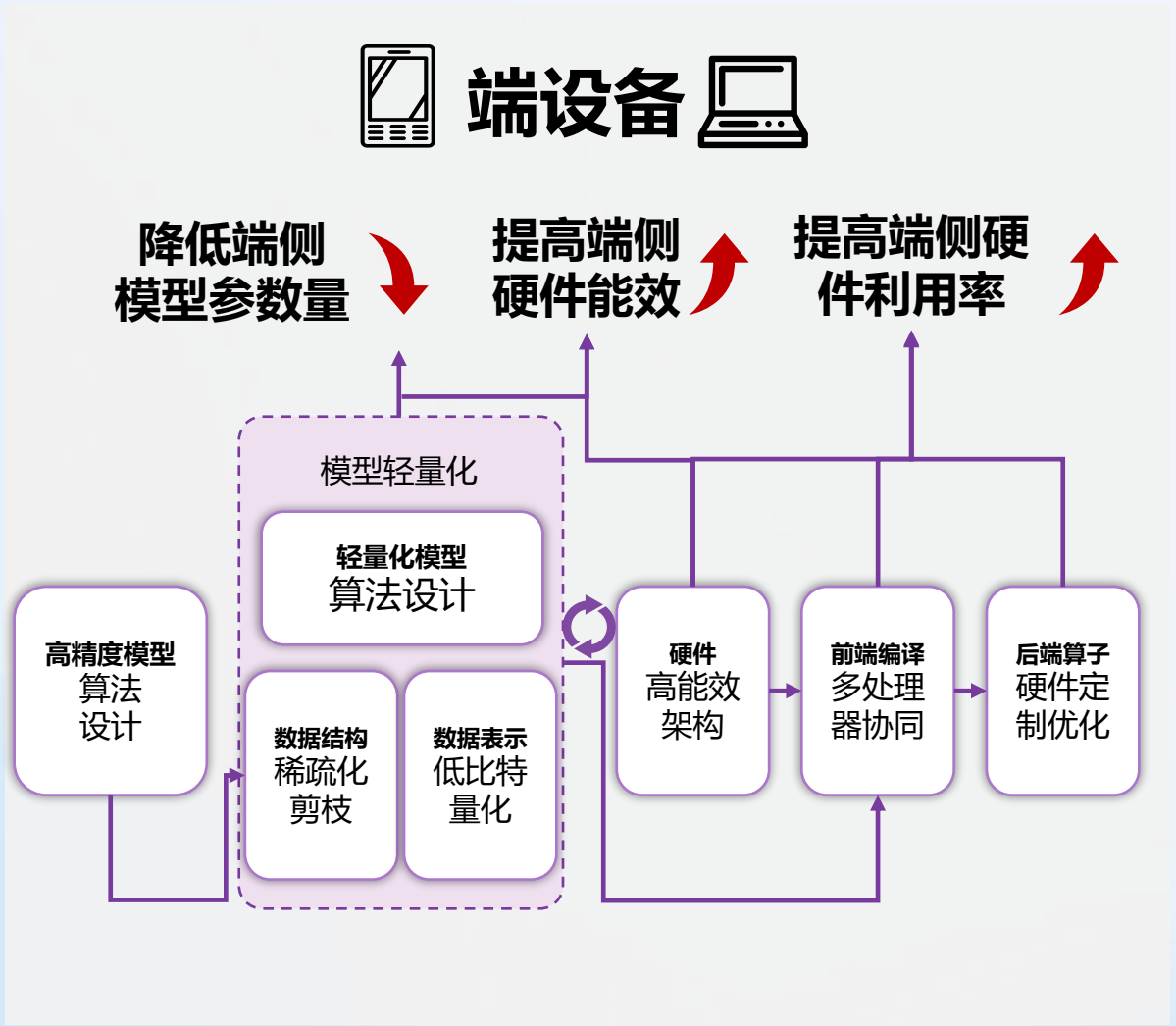
SpecEE

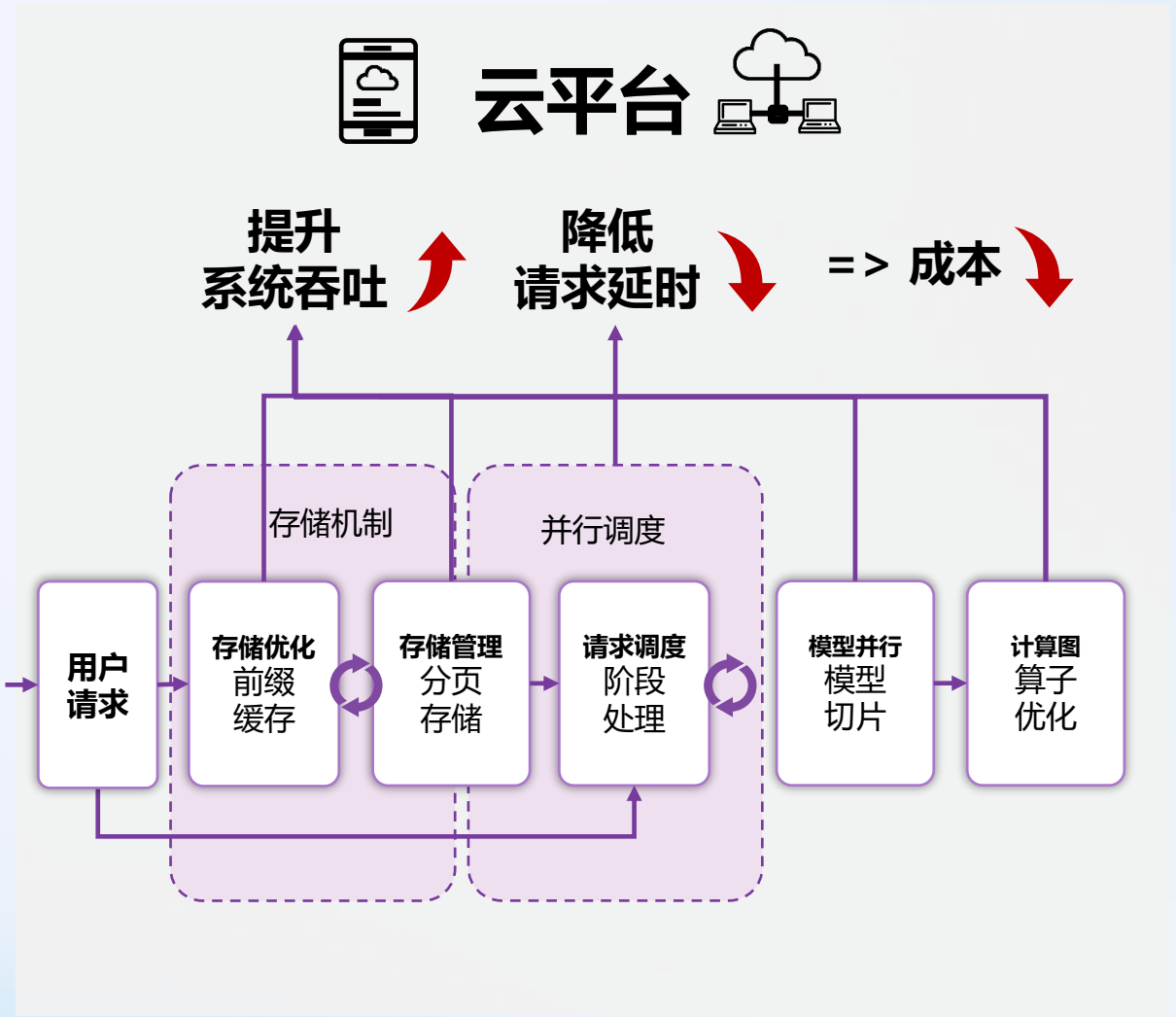
全球首创推测性早退技术，AI PC推理速度行业第一

端侧实时守护 云侧全局智能



端+云推理加速 数量级降低大模型落地成本





基础设施的“魔法”

1995

让民用电走进

10000个县镇

2025

让大模型成本

10000倍下降

InfoQ 极客传媒

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

InfoQ 极客传媒

THANKS

大模型正在重新定义软件

Large Language Model Is Redefining The Software