大模型AI训练的数据加速

肖文聪

2024.4.13　　Qcon 北京站

InfoQ 极客传媒

已结束 北京站 2000 人

QCon 全球软件开发大会 暨智能软件开发生态展

架构演进 | AIGC
研发效能 | 向量数据库

6月14-15日 深圳站 1000 人

ArchSummit 全球架构师峰会 暨智能软件开发生态展

业务架构 | 高可用
AI 架构 | 架构师成长

10月18-19日 上海站 1500 人

QCon 全球软件开发大会 暨智能软件开发生态展

AIGC | AI 原生应用
软件工程 | 金融科技

4月 5月 6月 8月 10月 12月

5月17-18日 北京站 2000 人

AiCon 全球人工智能开发与应用大会 暨大模型应用生态展

通用大模型 | AI Agent
LLMOps | AI Infra

8月16-17日 上海站 1000 人

FCon 全球金融科技大会 暨智能软件开发生态展

金融大模型 | 核心系统
数字化转型 | 数据安全

12月20-21日 北京站 1000 人

ArchSummit 全球架构师峰会 暨智能软件开发生态展

未来架构 | 高可用高并发
AI 应用 | 架构师思维

极客邦科技 2024 年会议规划

促进软件开发及相关领域知识与创新的传播

访问大会官网

参会咨询

肖文聪

- 阿里云-PAI机器学习　高级技术专家
- 负责PAI灵骏GPU集群管理、容错和稳定性、AI数据加速、LLM推理等方向
- 在OSDI/NSDI/ATC等系统顶会上发表论文30余篇，引用2000+

目录

- 大模型存储需求背景
- 通用存储架构下的挑战
- 改进的系统设计
- 阿里云DatasetAcc产品

快速增长的大模型规模

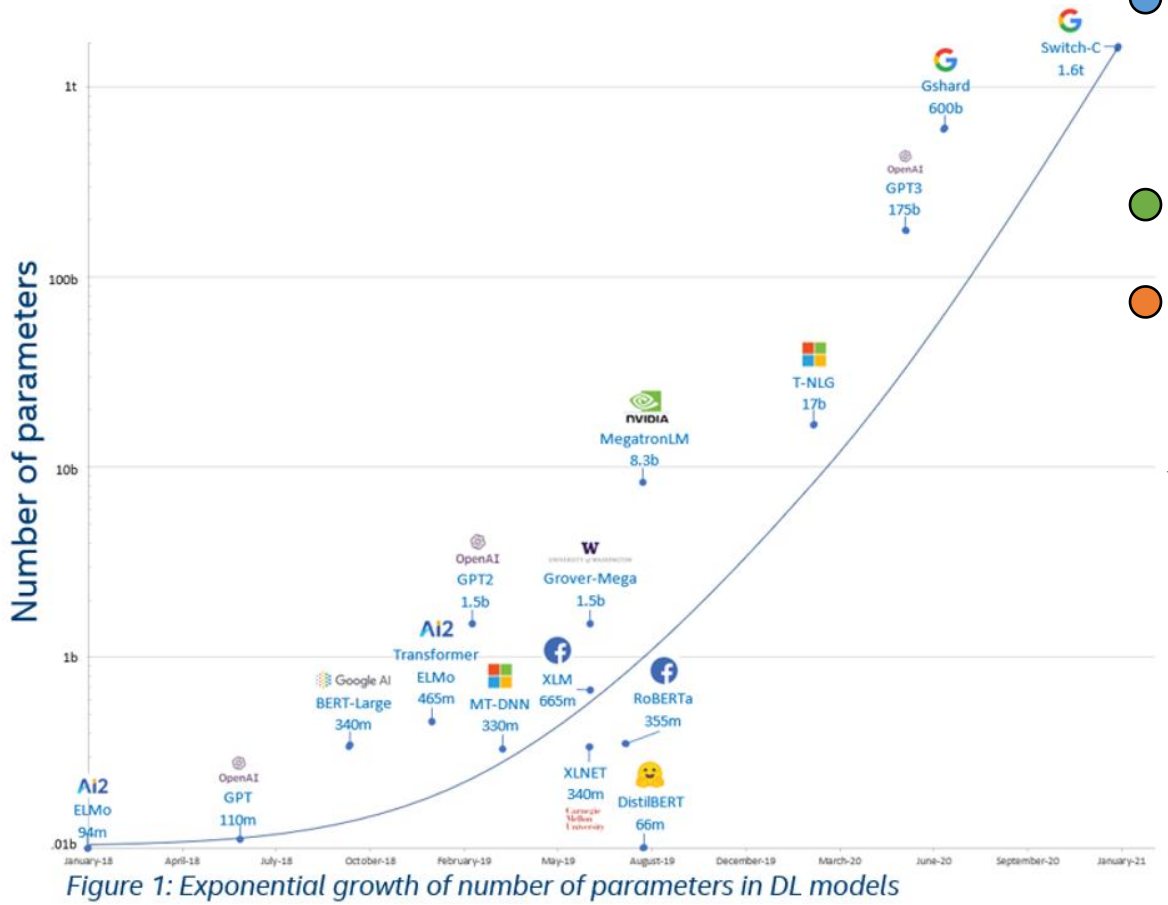


Figure 1: Exponential growth of number of parameters in DL models

GPT5 ??T

GPT4 1.8T （猜测）

Mistral 8*22B

Qwen 72B

至今国内开源模型仍无法追平2020年的GPT3

快速增长的大模型规模

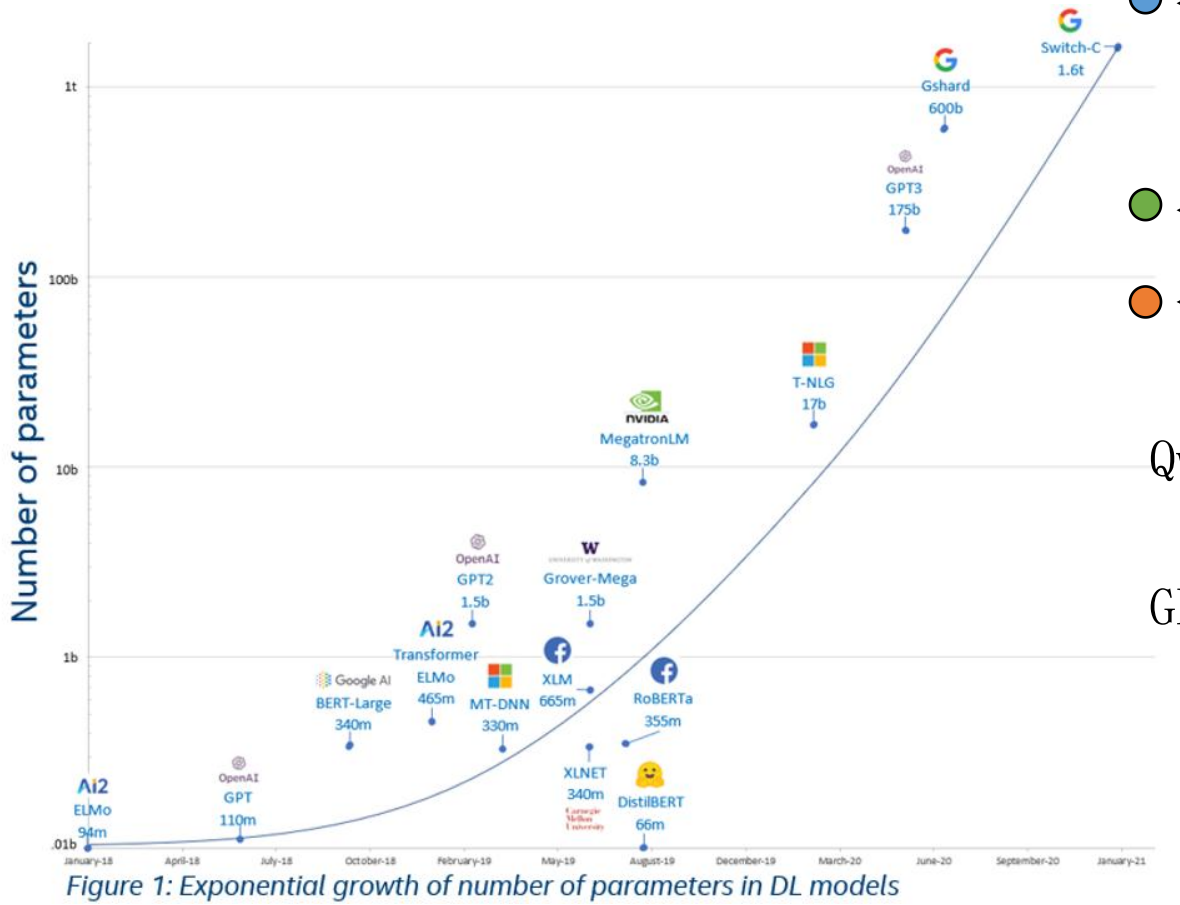


Figure 1: Exponential growth of number of parameters in DL models

GPT5 ??T

GPT4 1.8T （猜测）

Mistral 8*22B

Qwen 72B

Qwen72B：一个模型Checkpoint是150GB

GPT4：Checkpoint大小预计是**3.75TB**

快速增长的大模型规模

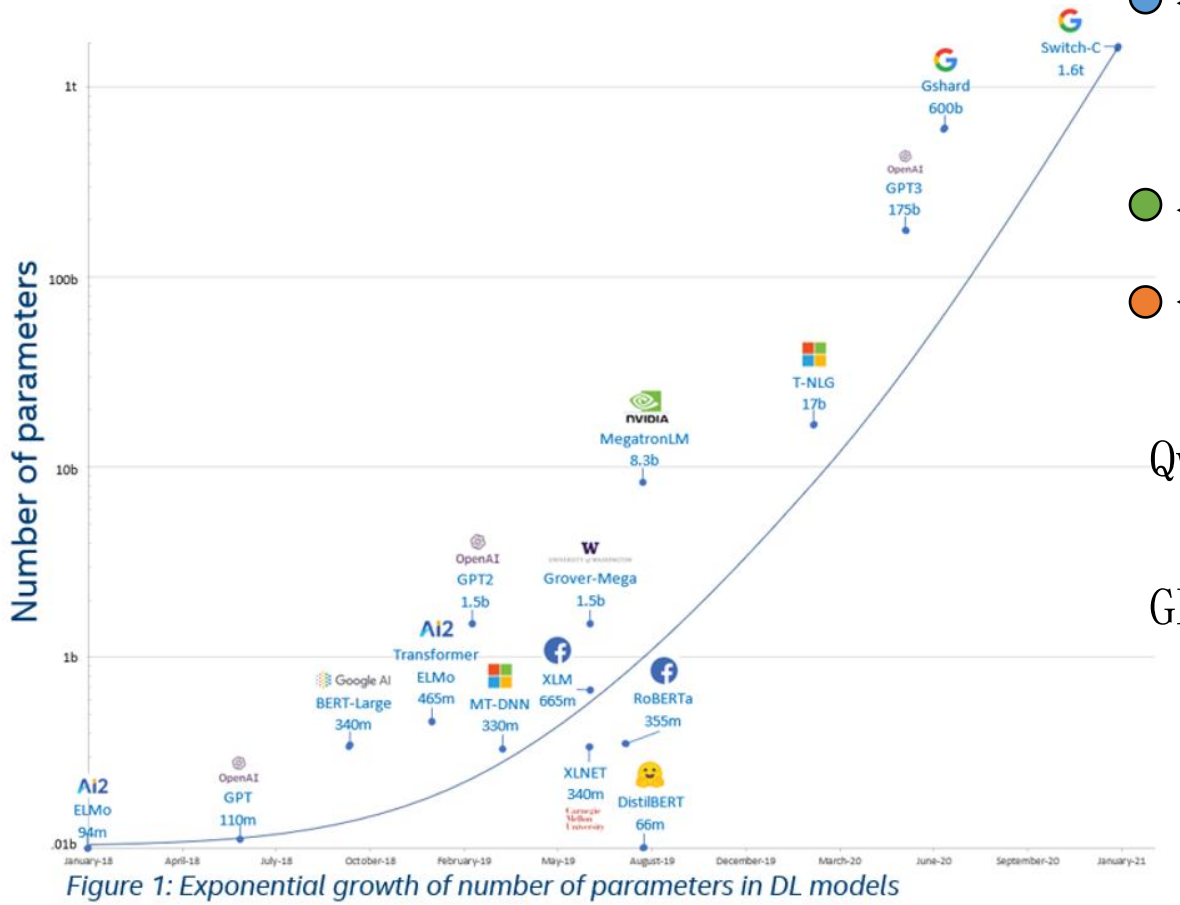


Figure 1: Exponential growth of number of parameters in DL models

GPT5 ??T

GPT4 1.8T （猜测）

Mistral 8*22B

Qwen 72B

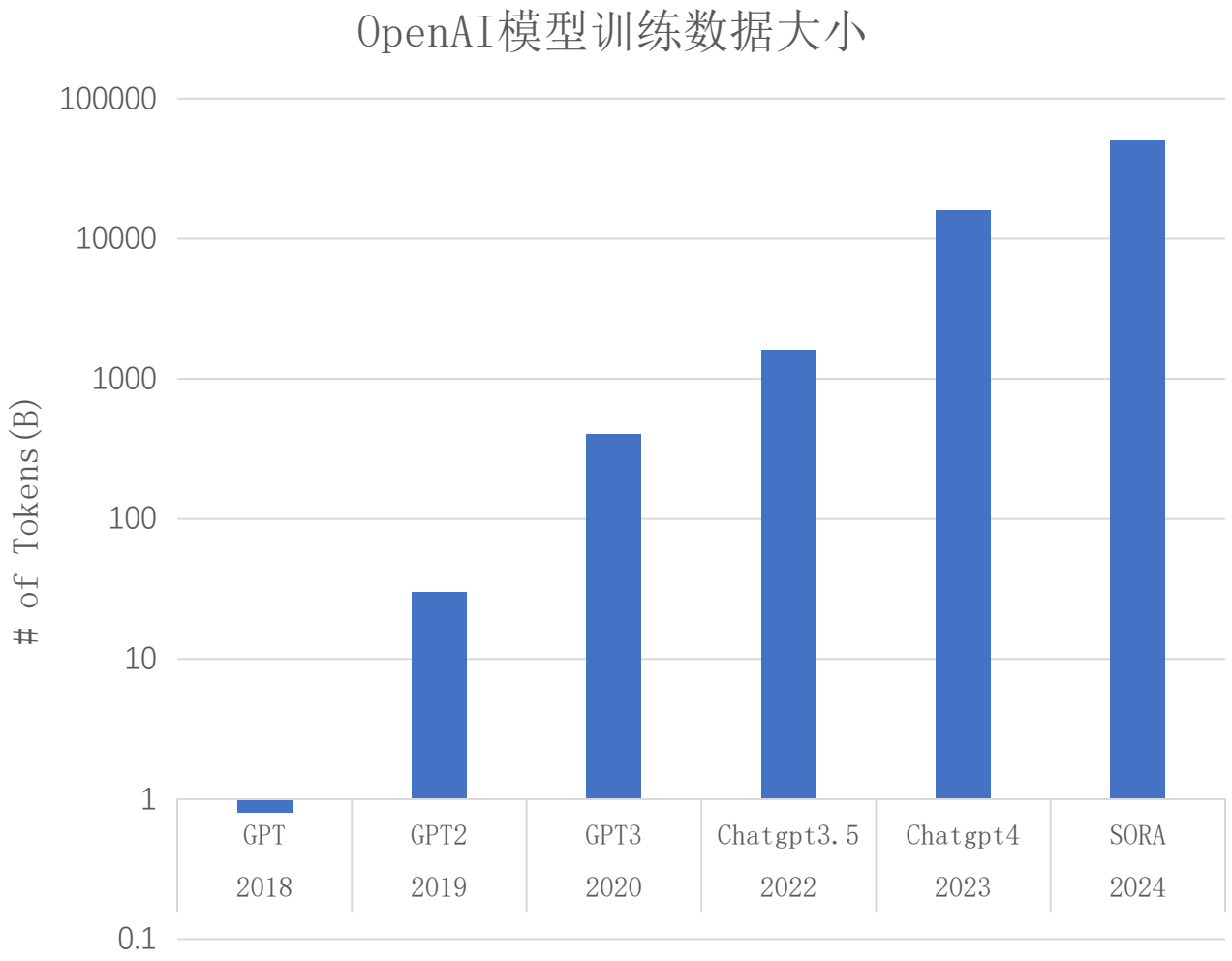
Qwen72B：一个模型Checkpoint是150GB

GPT4：Checkpoint大小预计是**3.75TB**

“秒级服务扩容？！”

激增的AI训练数据

- GPT3约570GB
- GPT4预估20TB
- SORA预估100TB
  - 多模态数据
    - 文本
    - 图片
    - 视频
  - 合成数据



*Chatgpt3.5, Chatgpt4, SORA均引用互联网公开讨论猜测数据规模

Scaling Law

- 算力、数据、模型越大，效果越好！
  - 算力：千卡 -->万卡
  - 数据：3000～50000B tokens
  - 模型：7B-->32B-->200B…

Scaling Laws for Neural Language Models

Jared Kaplan *
Johns Hopkins University, OpenAI
jaredk@jhu.edu

Sam McCandlish*
OpenAI
sam@openai.com

Tom Henighan
OpenAI
henighan@openai.com

Tom B. Brown
OpenAI
tom@openai.com

Benjamin Chess
OpenAI
bchess@openai.com

Rewon Child
OpenAI
rewon@openai.com

Scott Gray

Alec Radford

Jeffrey Wu

Dario Amodei

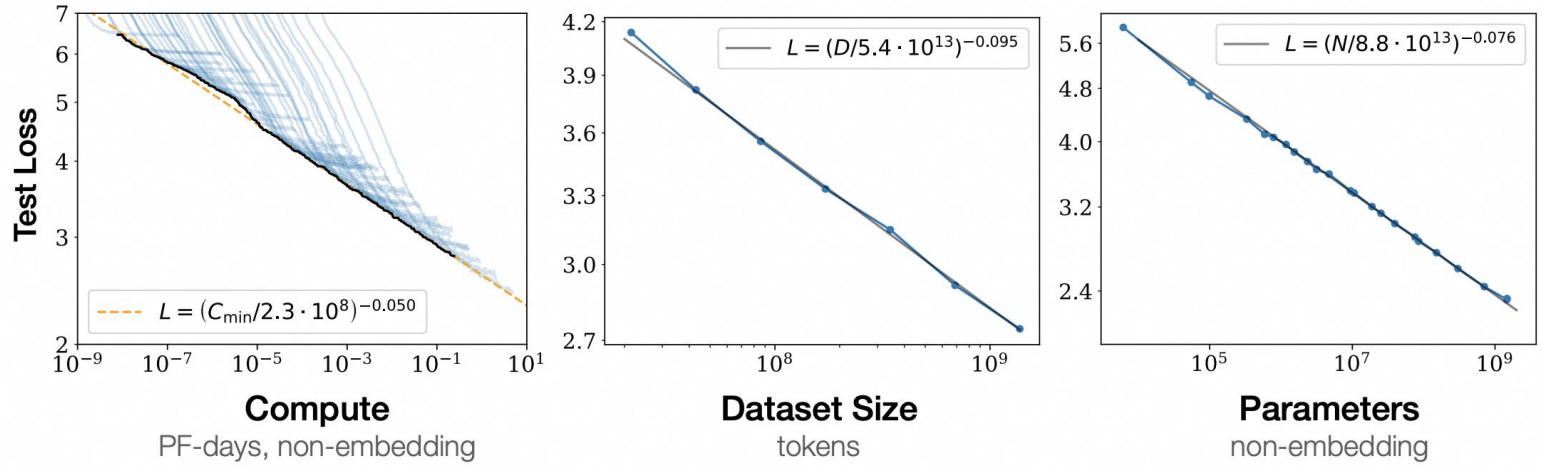
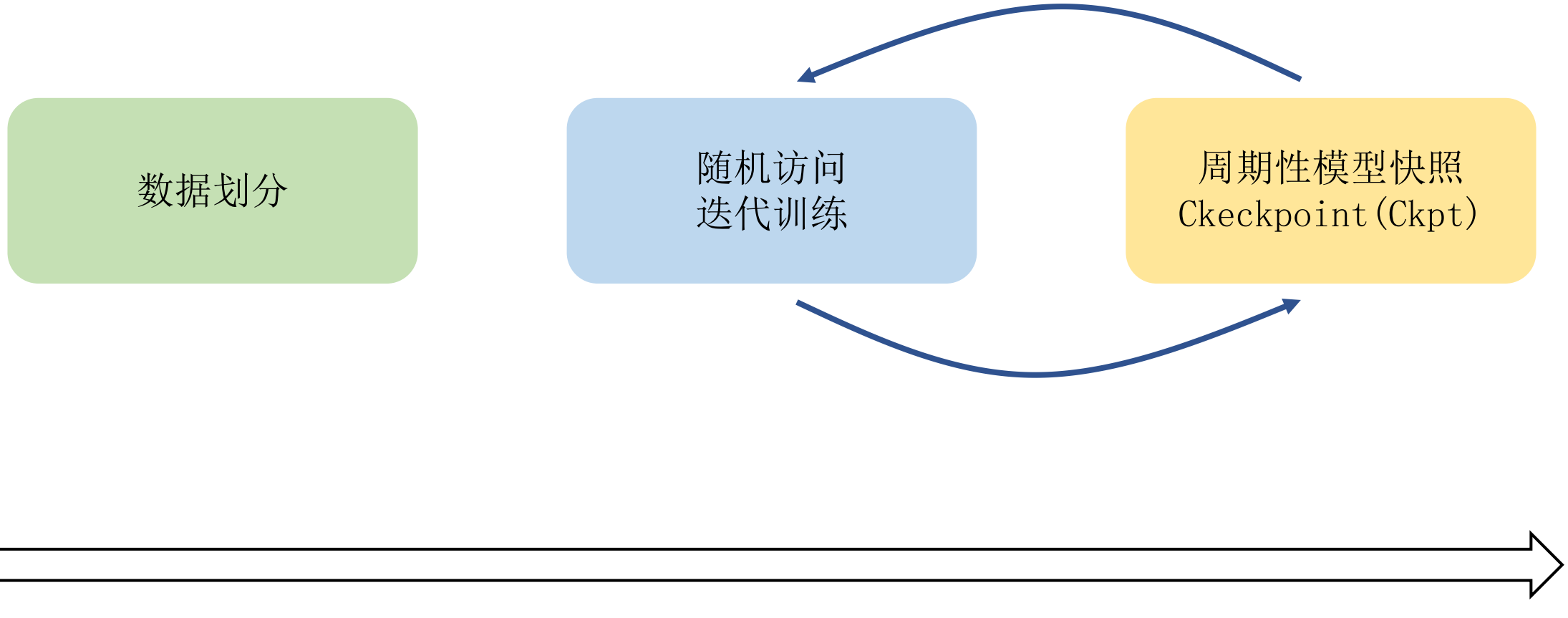


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

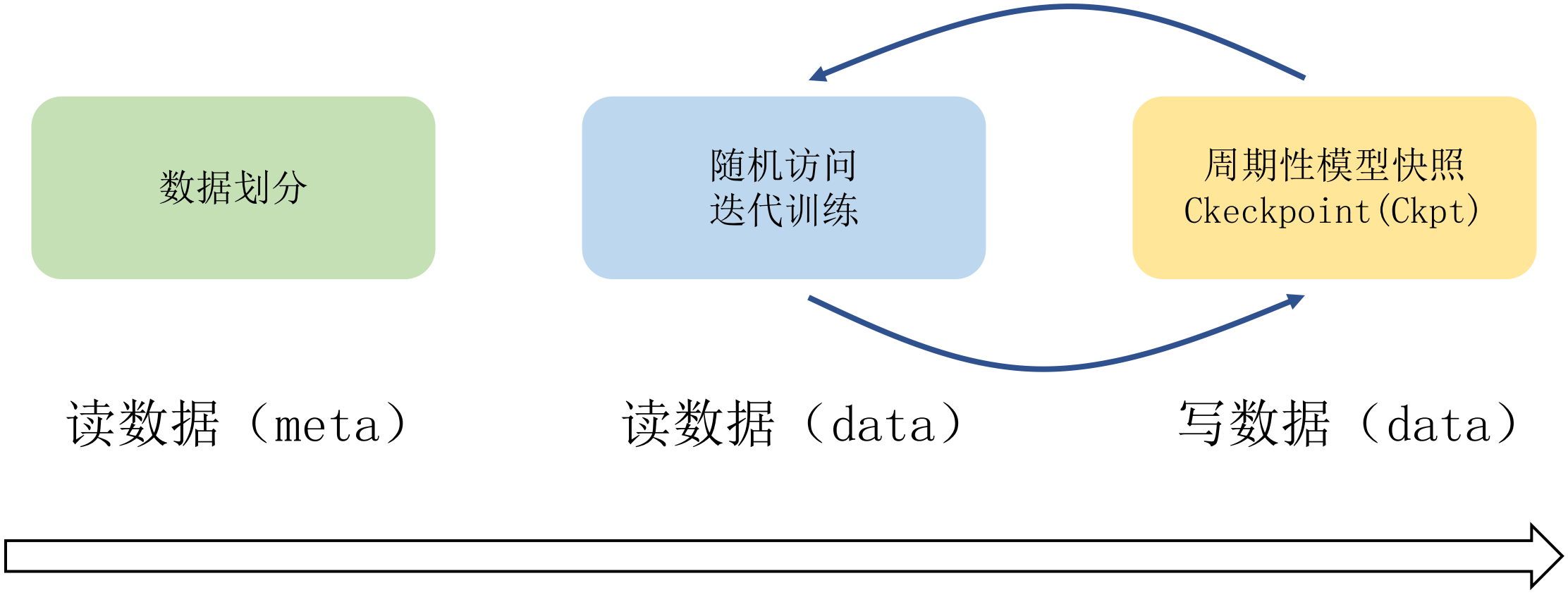
大模型带来的规模化数据

- 模型规模增大（Ckpt在～TB级别）
- 训练数据的增长（RawData在～100TB级别）
  - 多模态融合
  - 合成数据
  - 数据增强

回顾AI大模型训练流程



回顾AI大模型训练流程



挑战1：读数据

- 通用存储系统
  - 强一致Meta管理
  - 高可用的三副本

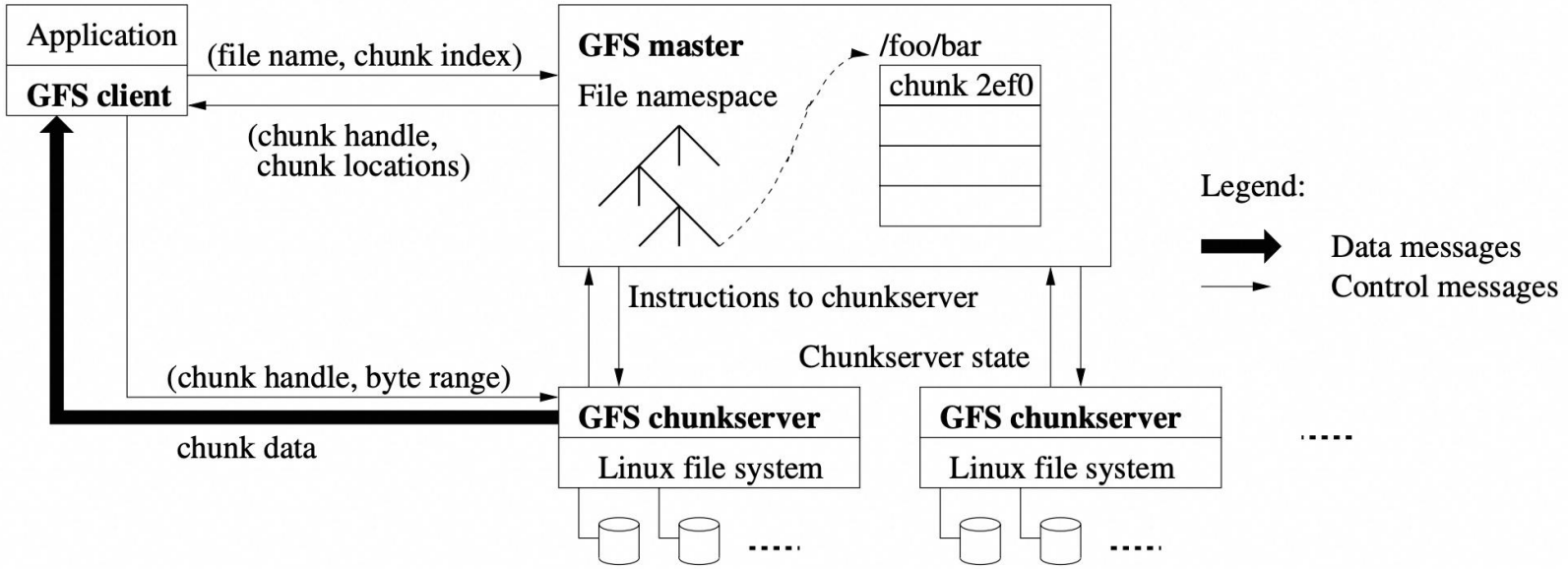


Figure 1: GFS Architecture

云存储：OSS/NAS等均沿用类似于GFS的架构设计

挑战1：读数据

- 通用存储系统
- AI小文件随机访问
  - Meta访问压力
  - 有限的IOPS*
    - OSS：10K
    - NAS：100K

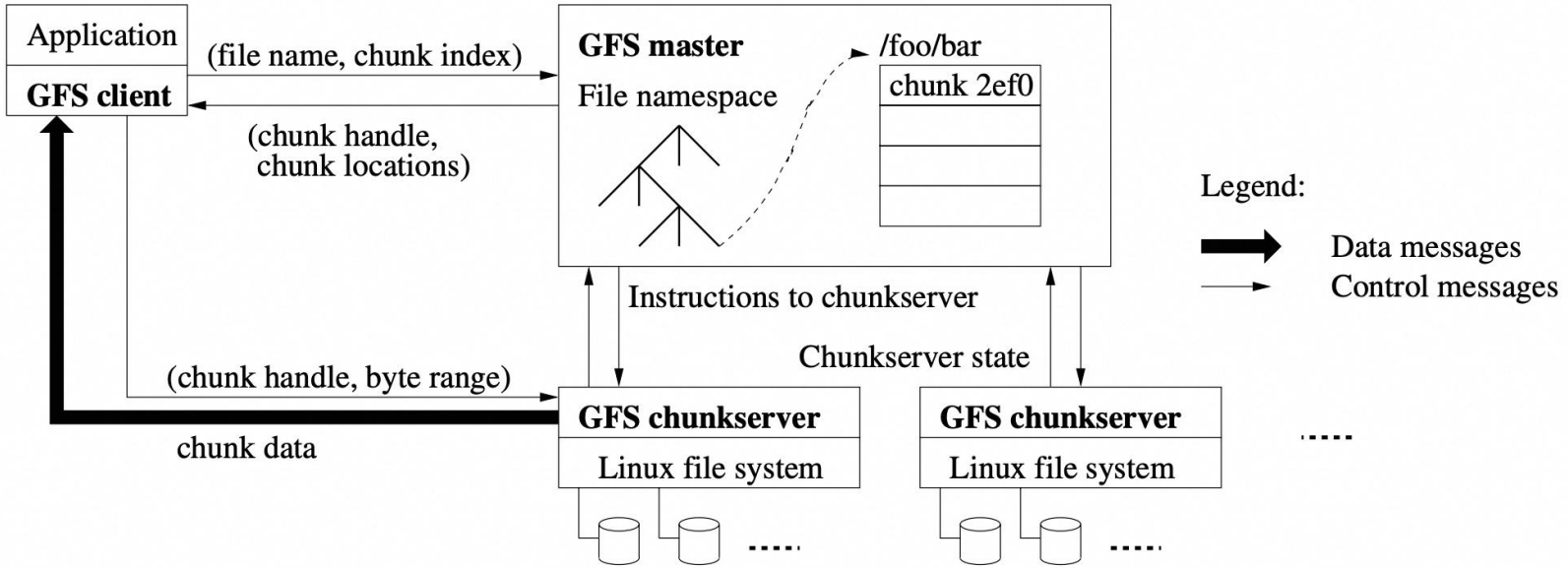


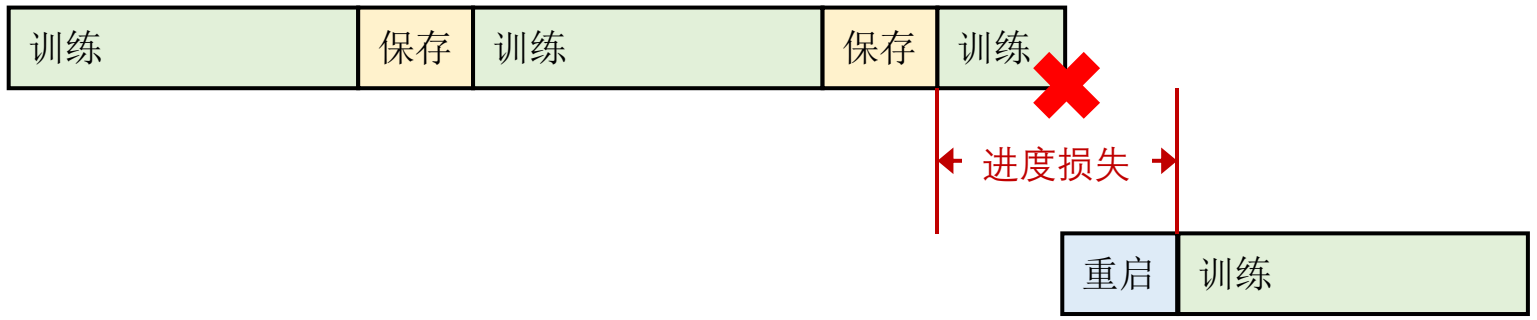
Figure 1: GFS Architecture

云存储：OSS/NAS等均沿用类似于GFS的架构设计

*单台GPU机器训练ResNet50需求约10K image/s

挑战2：写数据

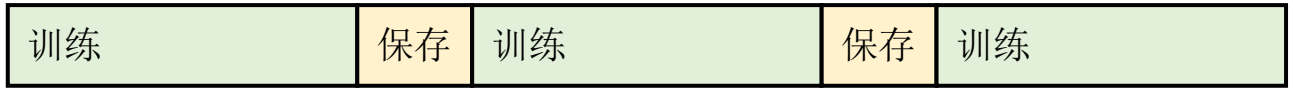
- 大规模分布式AI训练（e.g., 千卡规模）硬件故障不可避免
  - 通常采用周期性Ckpt进行容错



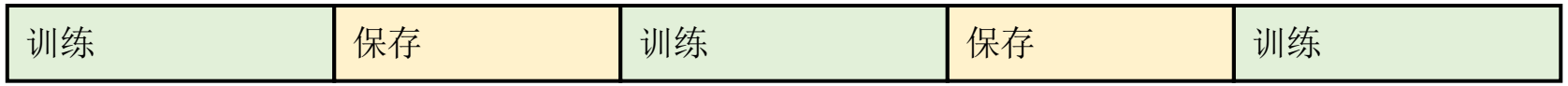
- 典型场景30min保存一次全量模型Ckpt
  - E.g., 150GB for Qwen 72B 模型

挑战2：写数据

- 大文件写入带宽Bound
  - 高带宽：



  - 低带宽：

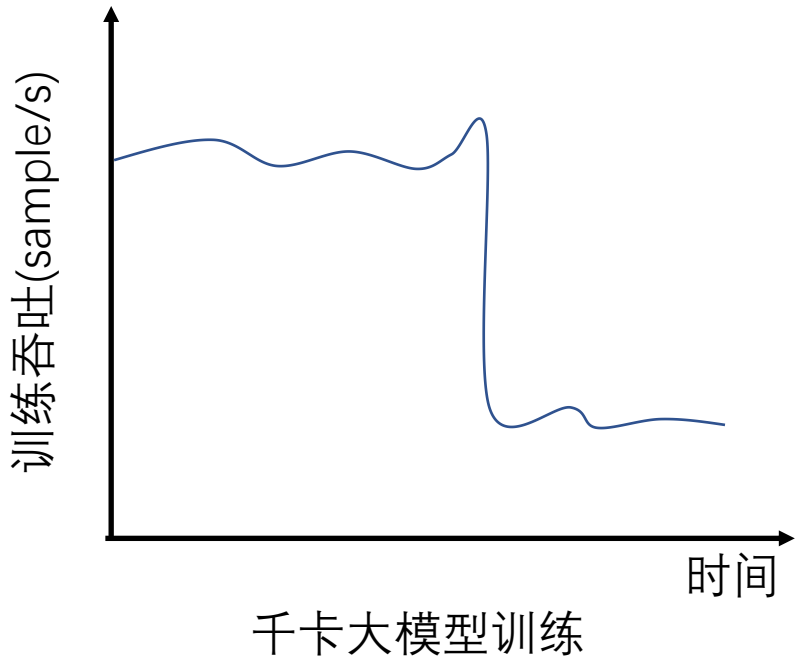


- 带宽 vs 成本
  - 高带宽全闪存储带来昂贵的成本

	带宽(GB/s)	价格(元/GB月)
OSS（对象存储）	1.25	0.033
CPFS（全闪）	10	1.6
倍数	8	48.48

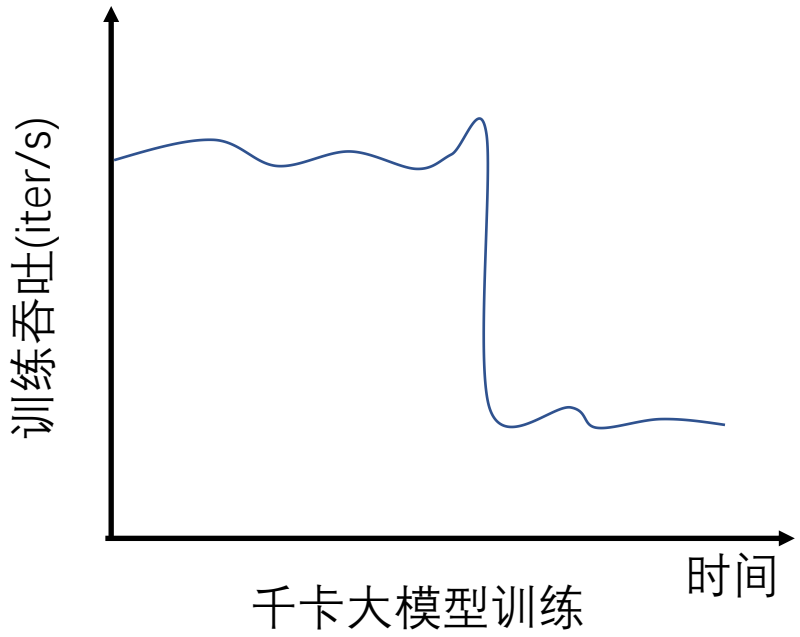
*以阿里云乌兰察布为例

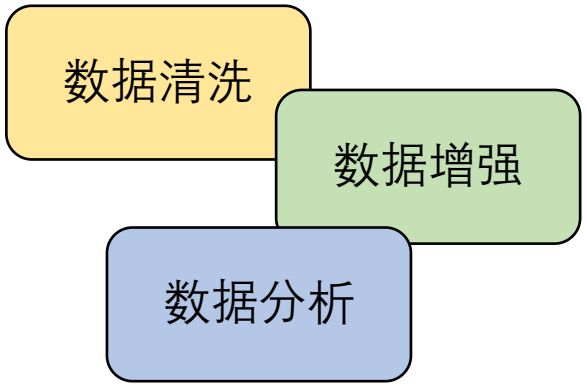
挑战3：性能隔离

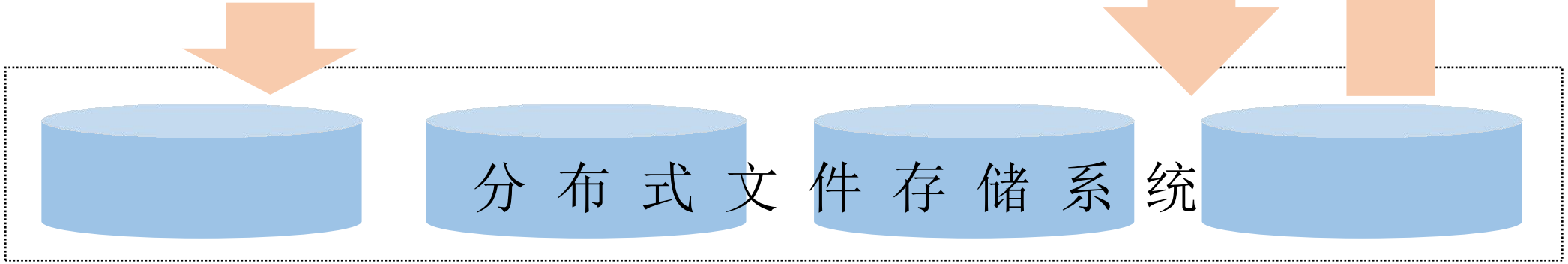


- 作业异常
  - 迭代时间变长
  - GPU利用率下降
  - 性能抖动

挑战3：性能隔离



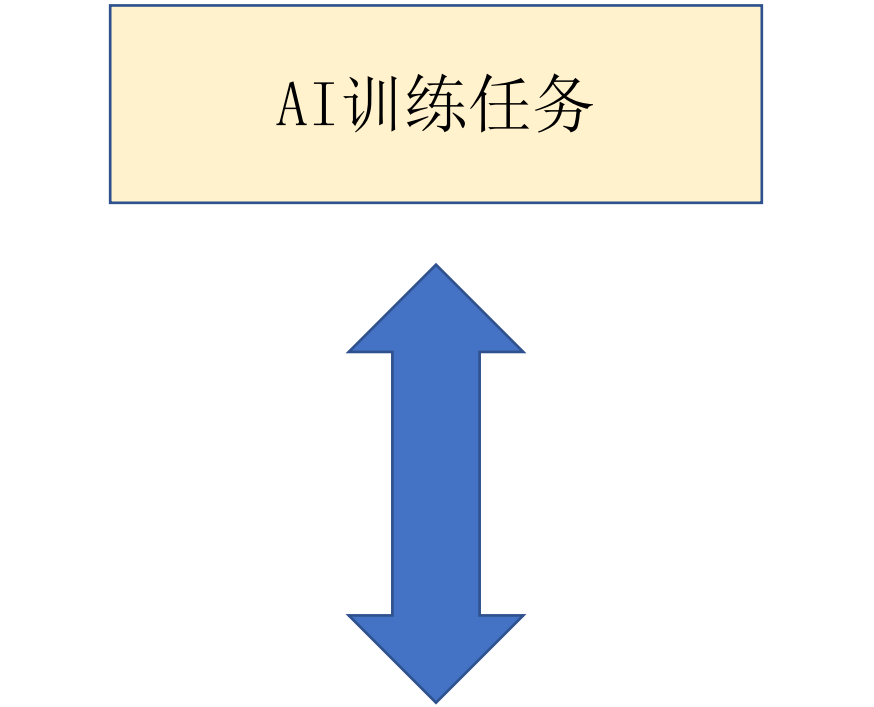




根因：文件存储系统并非为AI任务设计

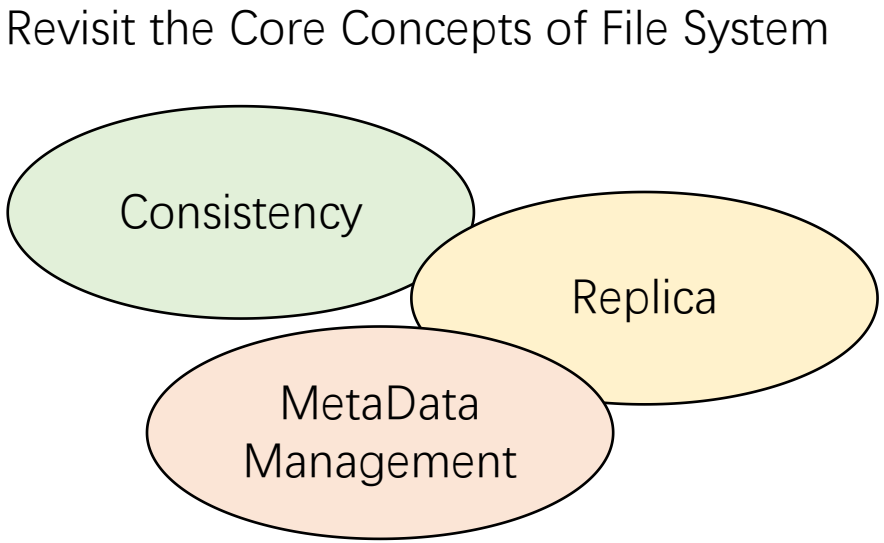
- AI训练要求“随机”访问数据对通用存储系统不友好
- AI训练带来大量爆发式Ckpt流量
- AI任务是大规模同步训练易受影响
- AI任务的特性没有被文件系统很好的利用

AI训练真的需要通用存储吗？

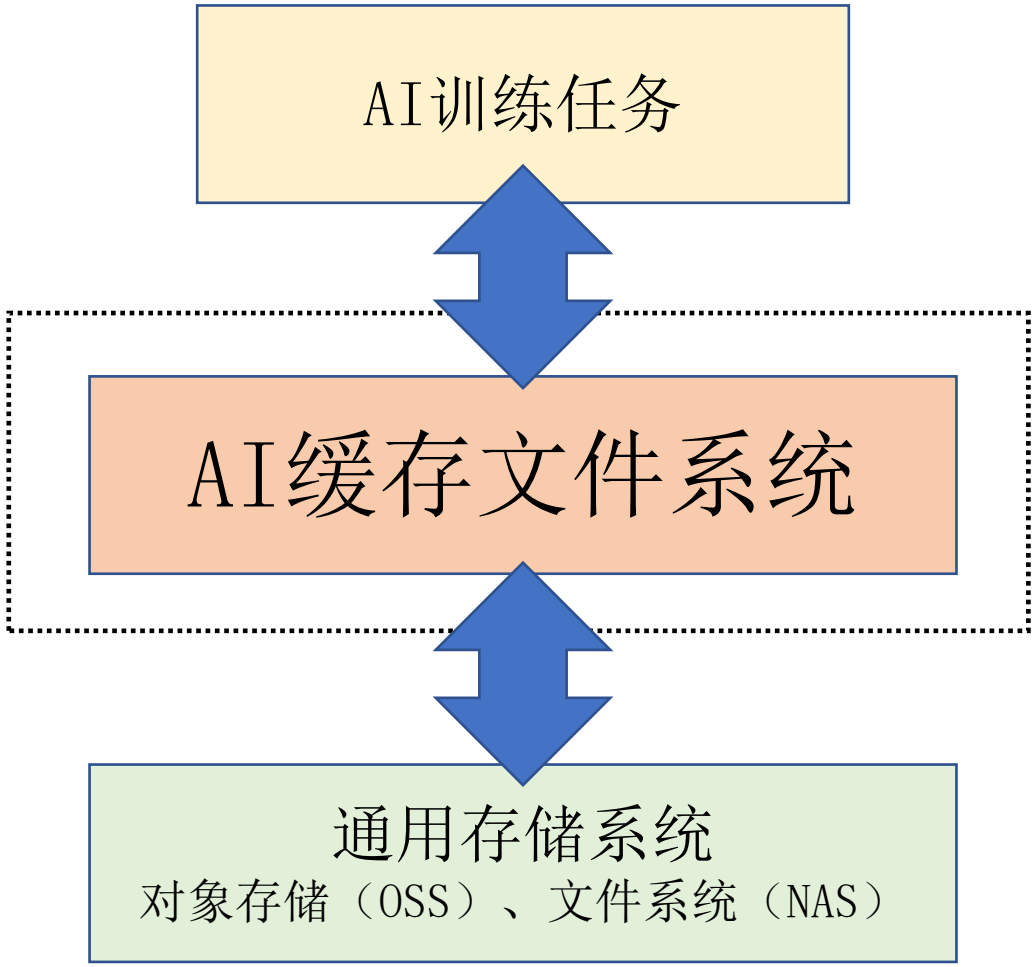


AI训练真的需要通用存储吗？

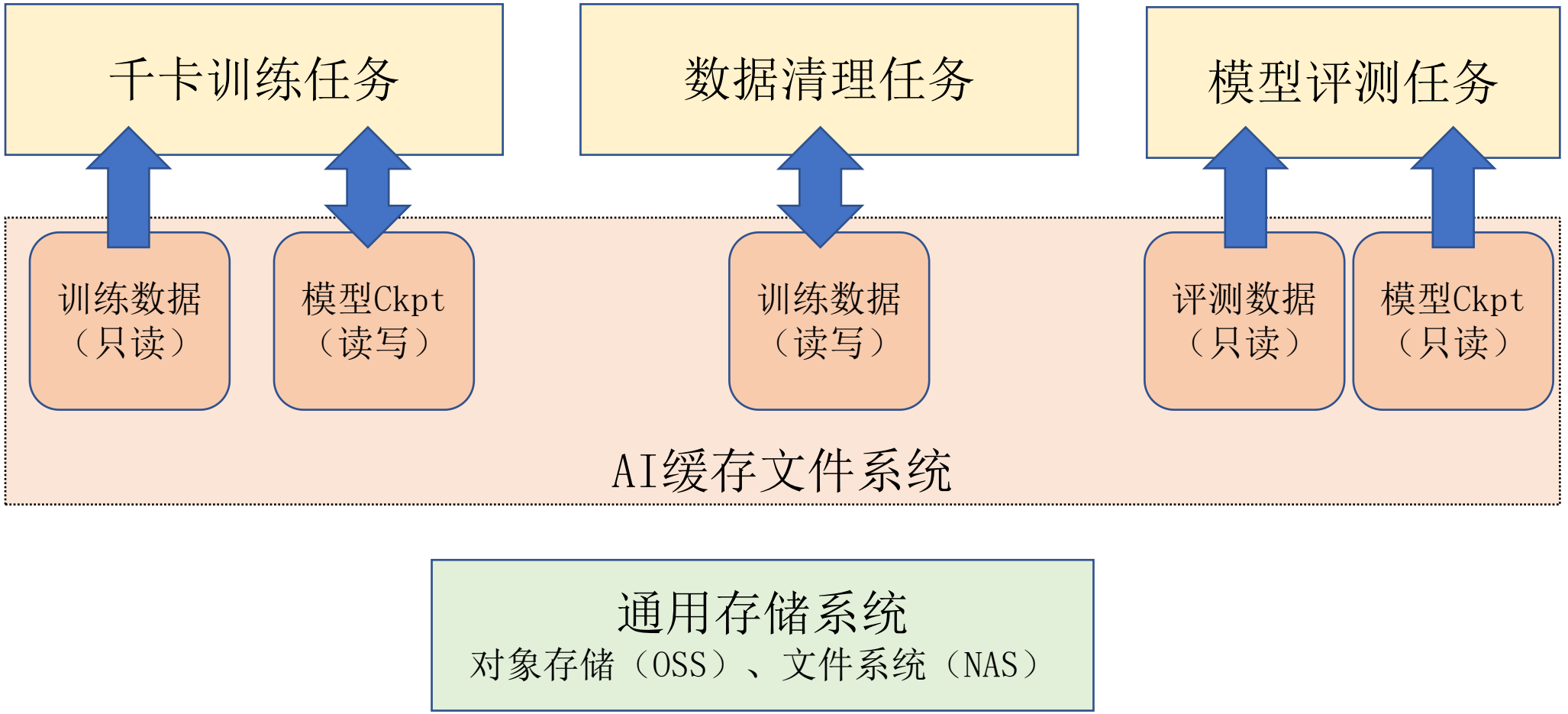
- 通用文件系统的弊端
  - 强一致性限制了架构的可扩展性
  - 多副本限制性能抬高成本
  - 读写混合潜藏着干扰
  - 缺乏任务间隔离能力



探讨：一个可能的缓存系统架构

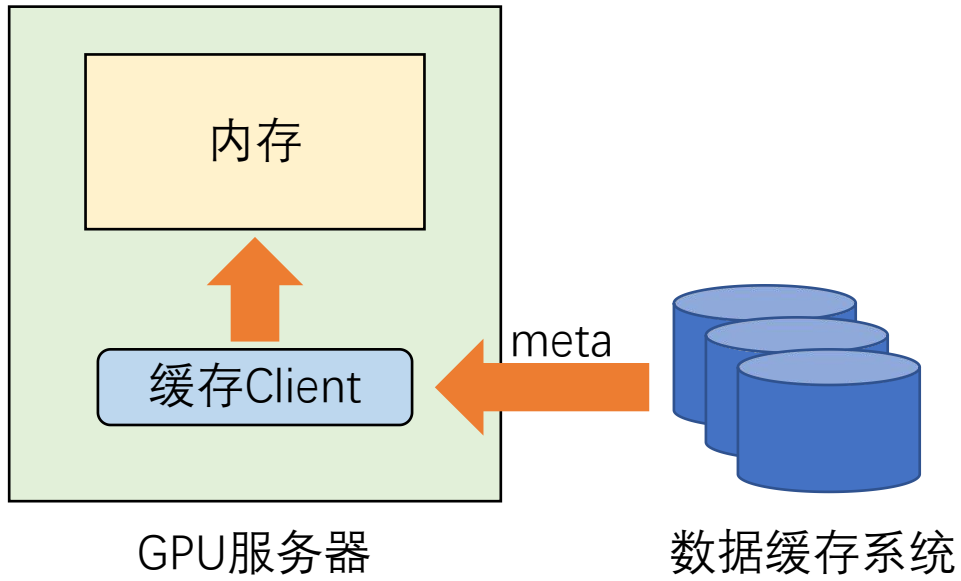


核心设计：读写分离缓存提供性能隔离



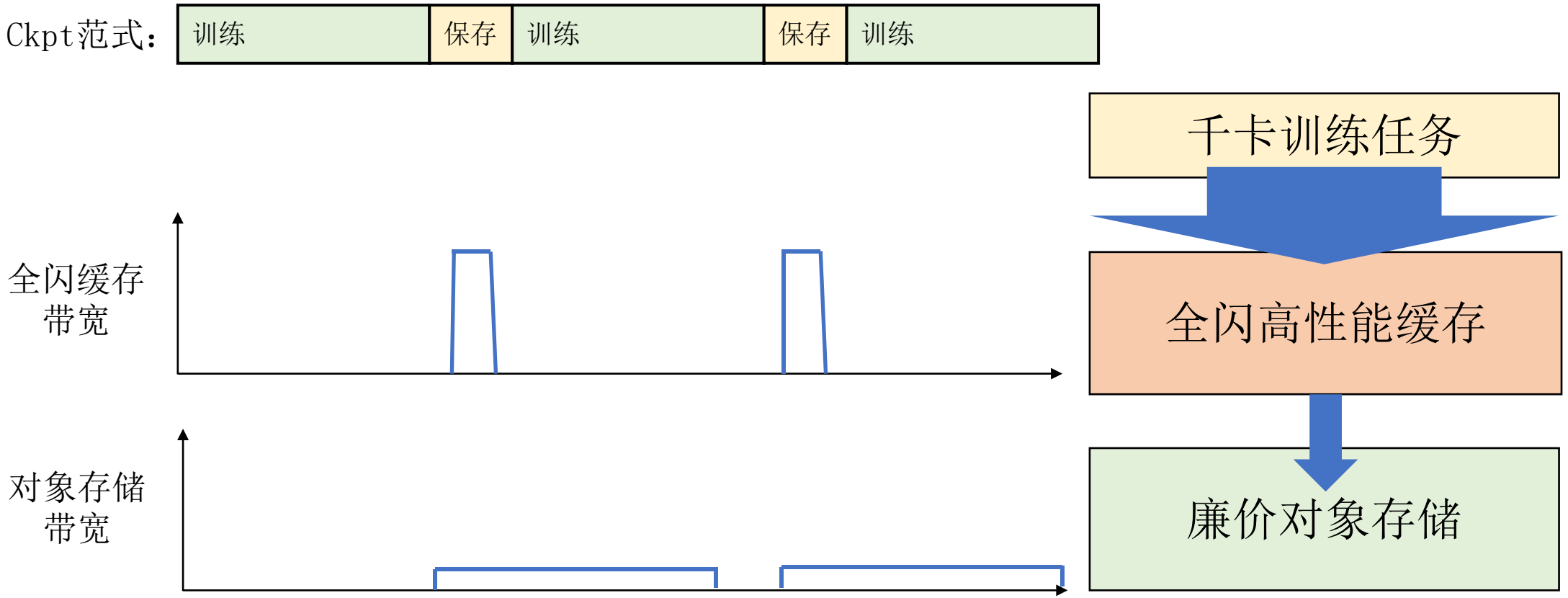
核心设计：Meta裁剪+主动近端Cache

- 并不是所有的Meta都有用
  - E.g.，修改时间，创建时间…
- 优先PyTorch DataLoader的需求
  - 文件名、大小
- 典型的H100服务器有2TB内存
  - 优先缓存Meta信息，可支持～100TB数据集



核心设计：平衡带宽与性价比

- 高速单副本缓存叠加廉价持久化存储



核心设计：平衡带宽与性价比

- 与全闪文件系统相比
  - 性能相似，一样的磁盘写入带宽
  - 价格分析
    - 因仅需单副本，相比3副本高可用文件系统，成本仅1/3
  - 可靠性：Ckpt可以接受极端情况下丢失

PAI-DatasetAcc

全托管、面向机器学习的云原生AI数据集加速服务

云原生

完全基于云原生基础设施
Kubernetes native， 容器化
支持客户自建ACK场景

多样加速策略

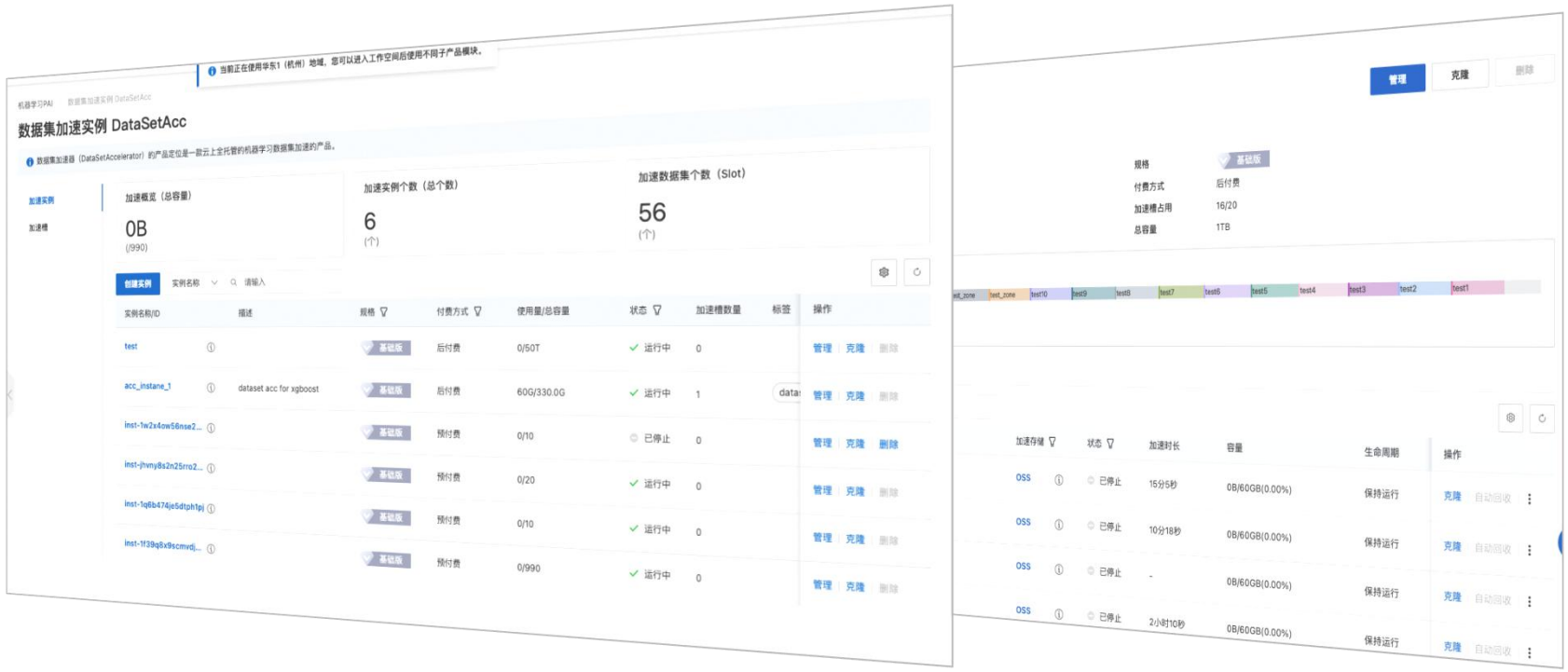
适配多种存储：对象存储、NAS、CPFS、ODPS
多级样本加速，训练效率达到最优
网络、训练框架协同制定最优加速策略

云上全托管

弹性资源，动态伸缩
可运维，多级监控和管理

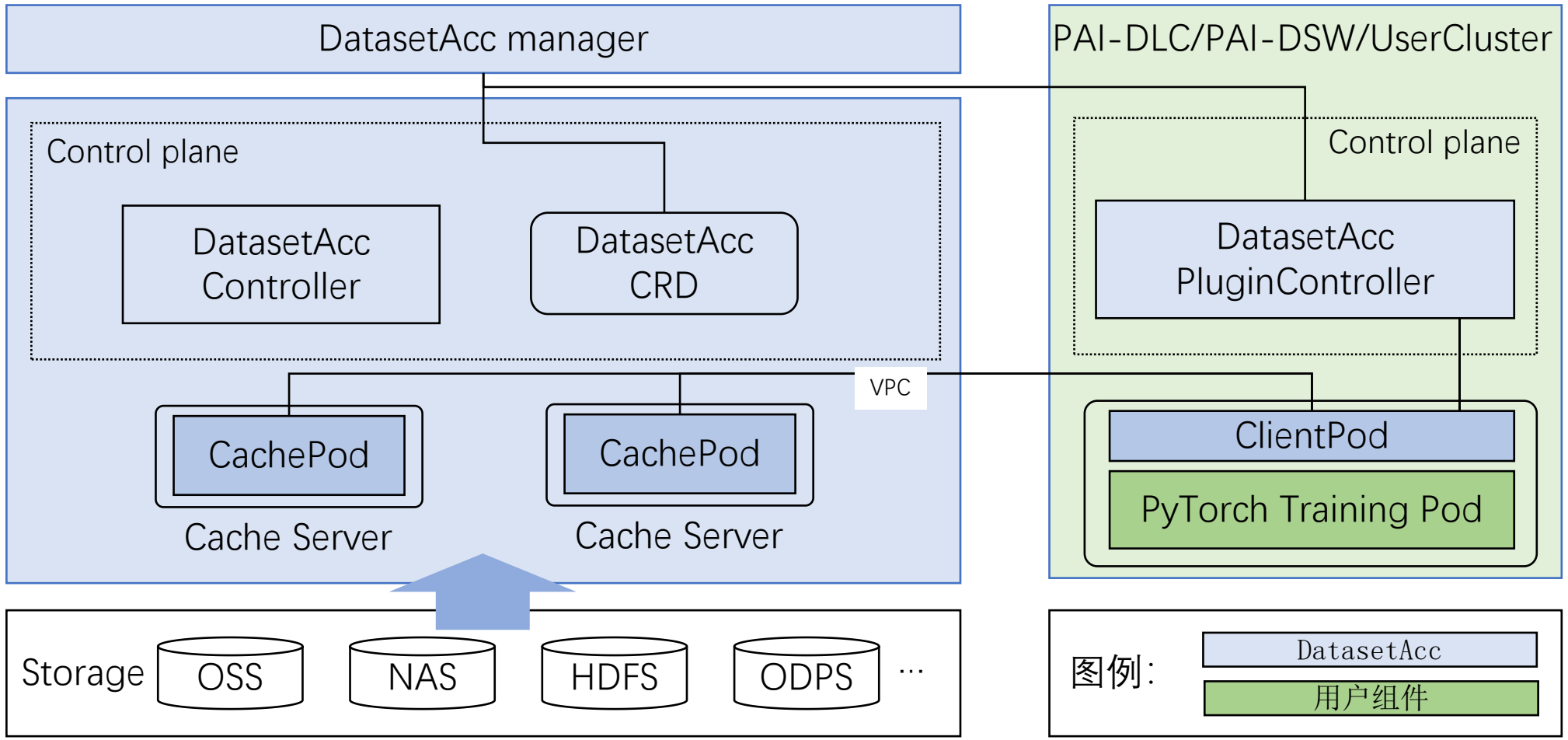
更易用

PAI-DSW、PAI-DLC 深度适配
训练代码无侵入
标准 OpenAPI，易于被集成



PAI-DatasetAcc

全托管、面向机器学习的云原生AI数据集加速服务



加速槽性能隔离

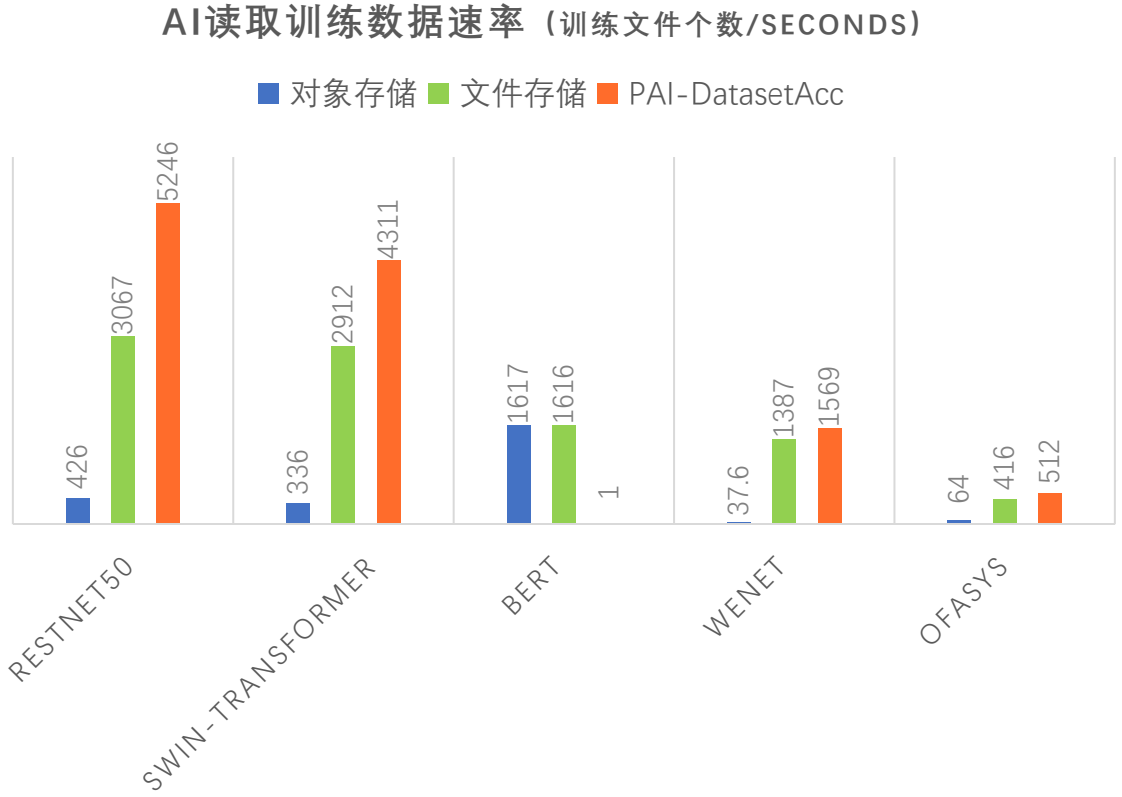
- 按需按量加速，文件夹级性能隔离

- E.g.， 10T加速实例
  - 1T只读加速训练数据
  - 5T为任务A加速Ckpt
  - 4T为任务B加速Ckpt

- 显式区分“只读”“读写”，极致优化性能

加速IO瓶颈的AI训练任务

- 更优的AI数据集读取性能



模型类型	业务模型	相比于云上对象存储（OSS）的加速比	
		文件存储（极速型）	PAI-DatasetAcc
图像分类	RetNet50	7.20X	12.31X
	SwinTransformer	8.67X	12.83X
多模态	OfaSys	6.5X	8.0X
语音识别	Wenet	36.89X	41.73X
NLP	Bert	1X	1x

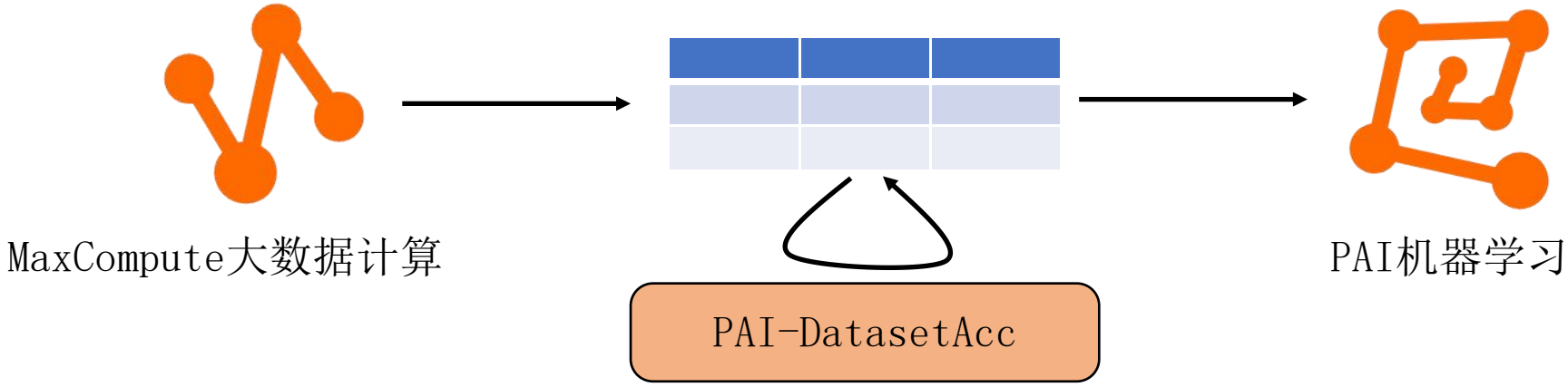
视觉任务　　文本任务　语音任务 多模态任务

*测试环境：单机8卡A100

高性价比加速模型Checkpoint

- Qwen 72B训练
  - 快照耗时：～分钟
  - DatasetAcc+OSS价格是全闪存储的70%

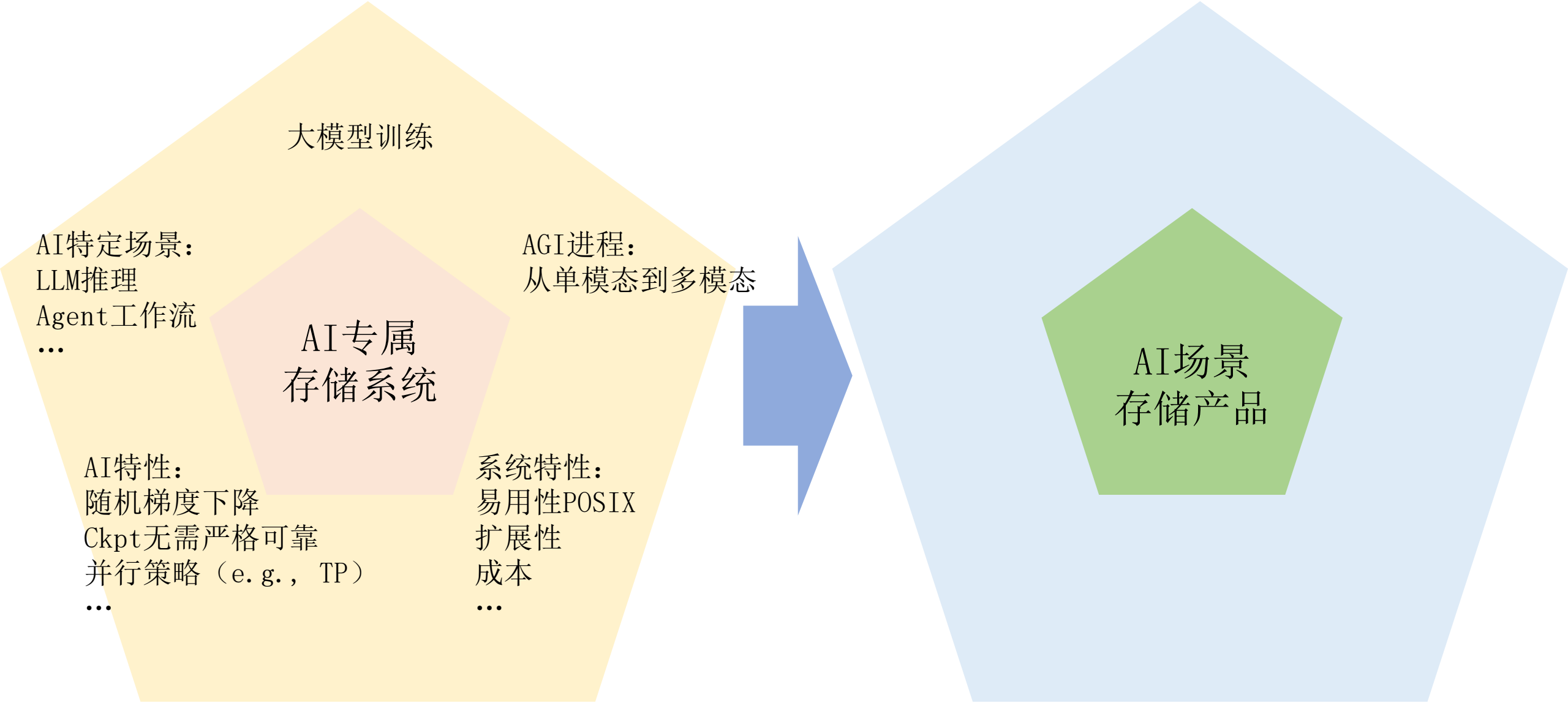
加速大数据结构化数据



- 500w条生产数据端到端性能测试
  - 2~3x性能提升

测试编号	BatchSize	使用DatasetAcc 加速读取（单位 秒）	使用原生MaxCompute耗时（单位秒）
1	512	14.71841049194336	39.86837840080261
2	1024	15.056357860565186	44.40856122970581
3	2048	15.917996883392334	37.80163049697876

总结：是时候为AI构建专属存储系统了！



InfoQ 极客传媒

已结束 北京站 2000 人

QCon 全球软件开发大会 暨智能软件开发生态展

架构演进 | AIGC
研发效能 | 向量数据库

6月14-15日 深圳站 1000 人

ArchSummit 全球架构师峰会 暨智能软件开发生态展

业务架构 | 高可用
AI 架构 | 架构师成长

10月18-19日 上海站 1500 人

QCon 全球软件开发大会 暨智能软件开发生态展

AIGC | AI 原生应用
软件工程 | 金融科技

4月 5月 6月 8月 10月 12月

5月17-18日 北京站 2000 人

AiCon 全球人工智能开发与应用大会 暨大模型应用生态展

通用大模型 | AI Agent
LLMOps | AI Infra

8月16-17日 上海站 1000 人

FCon 全球金融科技大会 暨智能软件开发生态展

金融大模型 | 核心系统
数字化转型 | 数据安全

12月20-21日 北京站 1000 人

ArchSummit 全球架构师峰会 暨智能软件开发生态展

未来架构 | 高可用高并发
AI 应用 | 架构师思维

极客邦科技 2024 年会议规划

促进软件开发及相关领域知识与创新的传播

访问大会官网

参会咨询

谢谢 & QA