GLake：大模型时代显存+传输 管理与优化

蚂蚁集团-基础智能-AI Infra

赵军平

2024.5

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

查看会议

参会咨询

QCon 北京
全球软件开发大会
暨智能软件开发生态展
会议时间：已结束

ArchSummit 深圳
全球架构师峰会
暨智能软件开发生态展
会议时间：6月14–15日

AiCon 上海
全球人工智能开发与应用大会
暨大模型应用生态展
会议时间：8月18–19日

ArchSummit 北京
全球架构师峰会
暨智能软件开发生态展
会议时间：12月20–21日

4月 5月 6月 8月 8月 10月 12月

AiCon 北京
全球人工智能开发与应用大会
暨大模型应用生态展
会议时间：5月17–18日

FCon 上海
全球金融科技大会
暨智能软件开发生态展
会议时间：8月16–17日

QCon 上海
全球软件开发大会
暨智能软件开发生态展
会议时间：10月18–19日

提纲

蚂蚁集团 ANT GROUP

- 背景与技术挑战
  - 显存墙、传输墙
- GLake介绍：显存与传输优化
  - 训练场景
  - 推理场景
  - 其他：混部、serverless
- 总结：开源

自我介绍

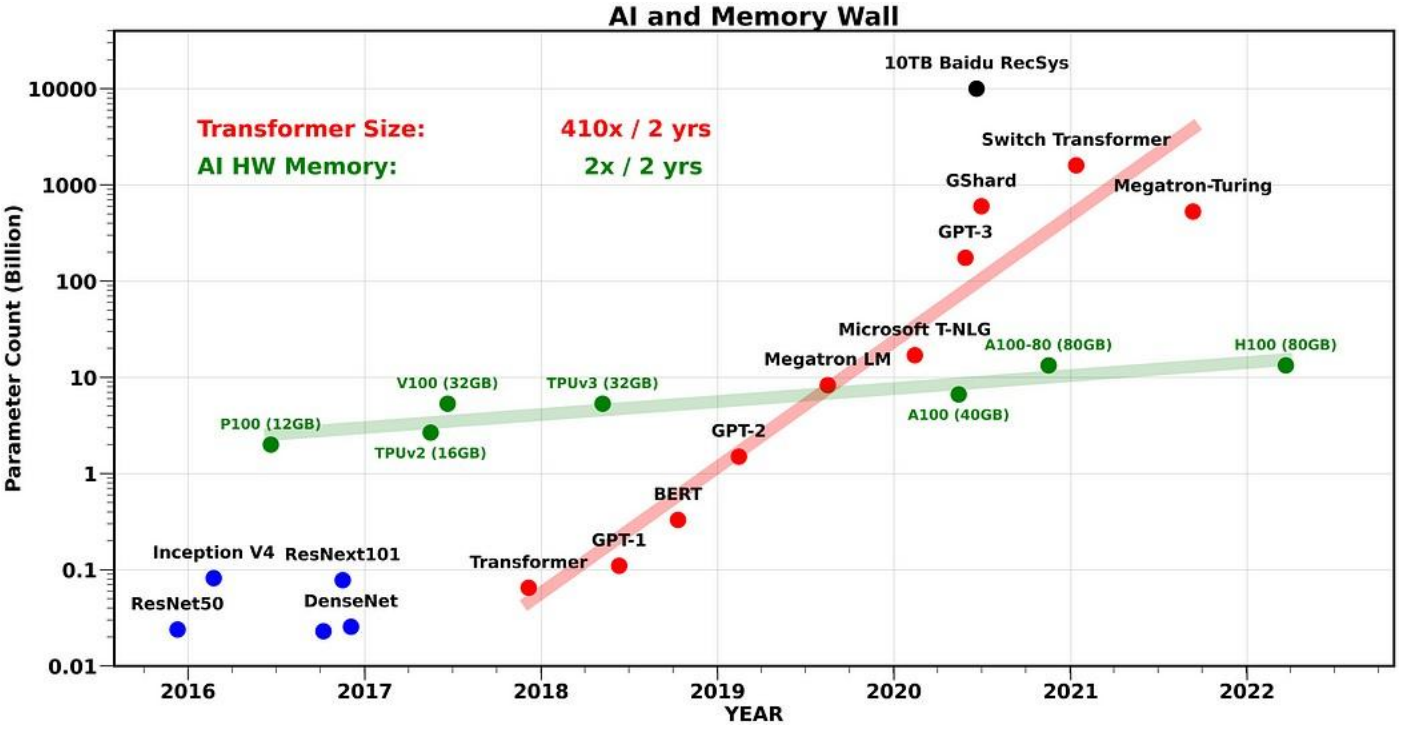
蚂蚁集团 ANT GROUP

- **赵军平，蚂蚁异构计算与推理引擎**负责人

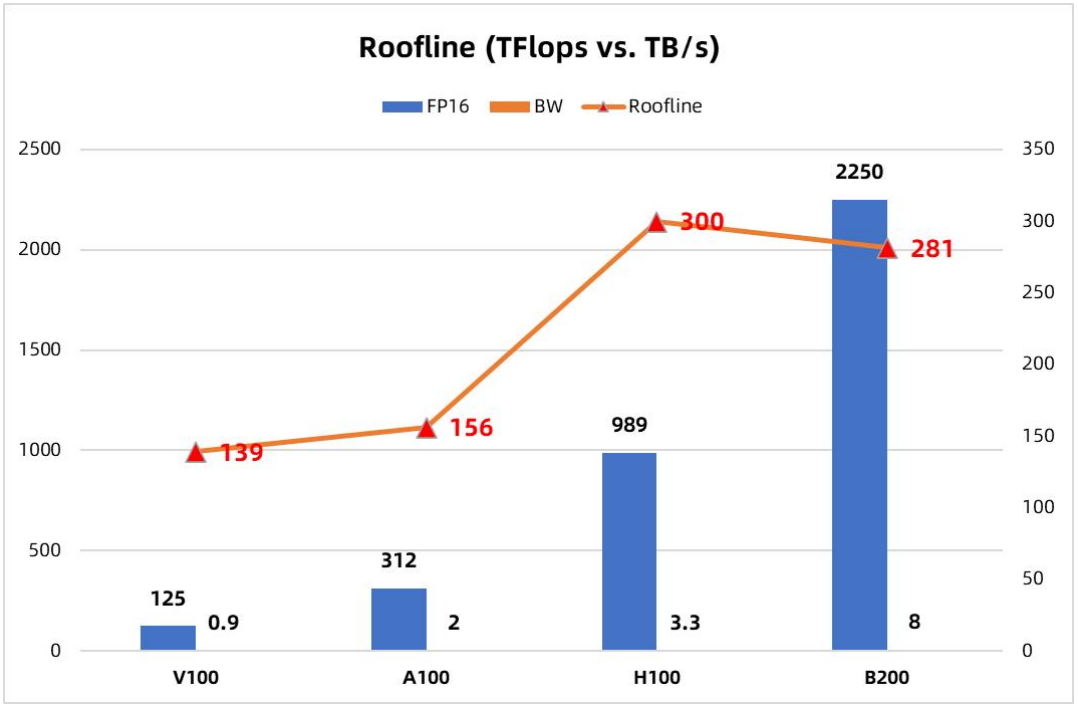
- CCF HPC和存储专委委员，~200 中/ 美技术专利
  - 异构加速，虚拟化，K8S，推理优化，文件系统，企业级存储-保护等
  - “ 数据密集型应用系统设计” 译者
  - 2007~2019：EMC、Dell EMC，Start-up

LLM技术挑战1：显存墙

- 显存**容量**、访存**带宽**（特别是推理小batch场景）



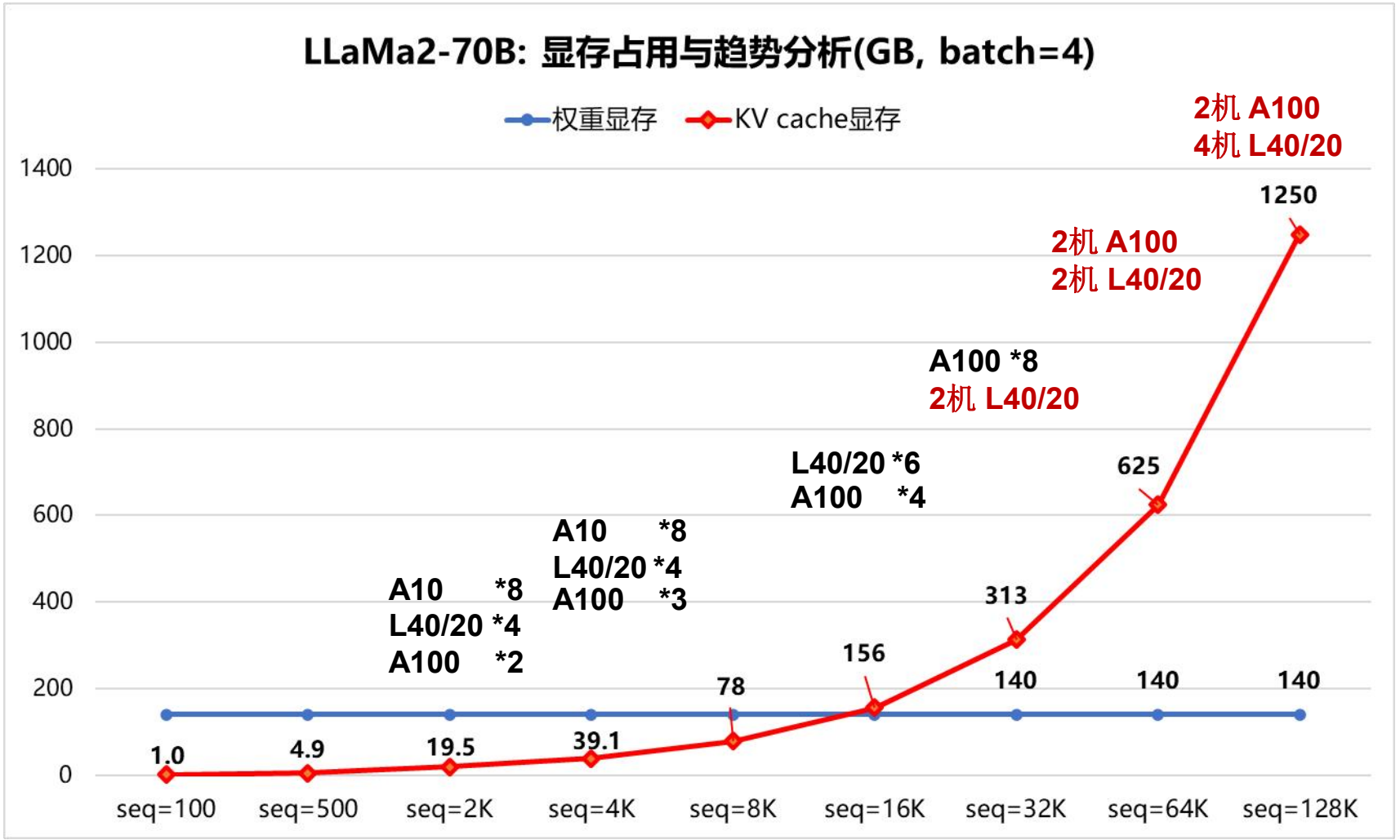
模型参数量 vs. 单卡显存容量



单卡算力 vs. 访存带宽

Llama2-70B推理为例

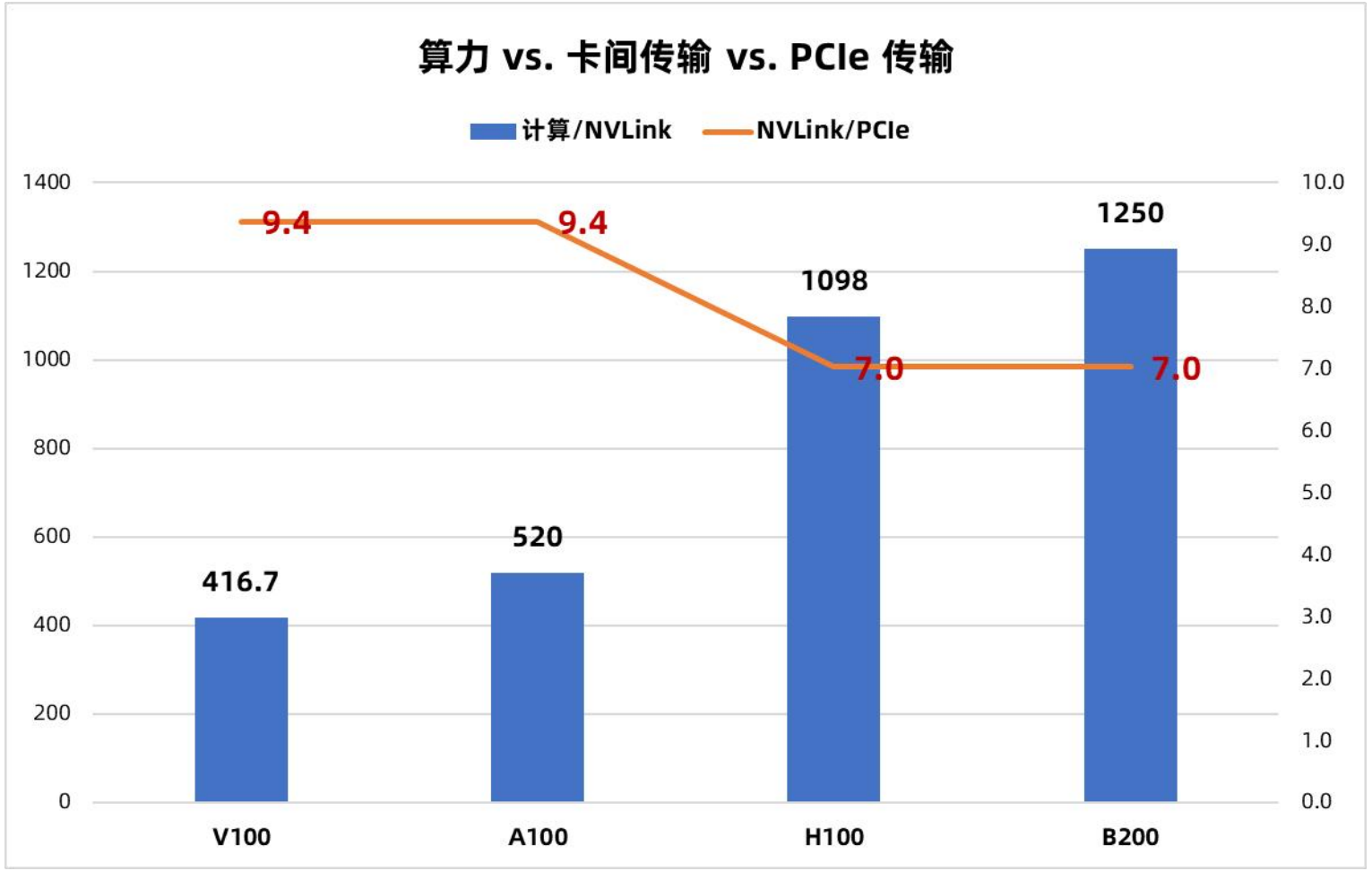
- 权重与KV cache的显存消耗分析 （假定FP16 未优化）



KV cache显存占用 = H# * H_dim * L# * 2 * 2B * seq# * b#

LLM技术挑战2：传输墙

- 访存、卡-卡传输、PCIe传输等发展 << 算力发展



业界方案

蚂蚁集团 ANT GROUP

思路	精度	性能	范围
模型/算法/实现改造 • DeepSeek • YOCO • MoD, MoE	Case by case	Case by case	case by case
梯度累计	depends	中~高	训练过程适配
重计算	无影响	中	通用 计算换显存
offloading/swapping	无影响	低(同步) ~ 高(异步)	通用 包括UMA
混合精度、小型化	可能影响 - 无损压缩不影响	中-高	Dependents QAT, AMP, INT8
多卡并行:PP, TP, EP, SP	-	-	Depends 模型改造
以上组合			