QCon
全球软件开发大会

InfoQ 极客传媒

飞桨大模型推理实践

——从集中式部署到分离式部署架构演进

蒋佳军

飞桨大模型部署技术负责人
百度 深度学习技术平台部

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

飞桨大模型推理背景

大模型推理需求激增

Google 2025年第二季度月处理Token量

980万亿

推理Token调用量

思考模型RL训练中推理耗时占比

80%+

大模型RL训练

IDC预测推理工作负载占比

73%

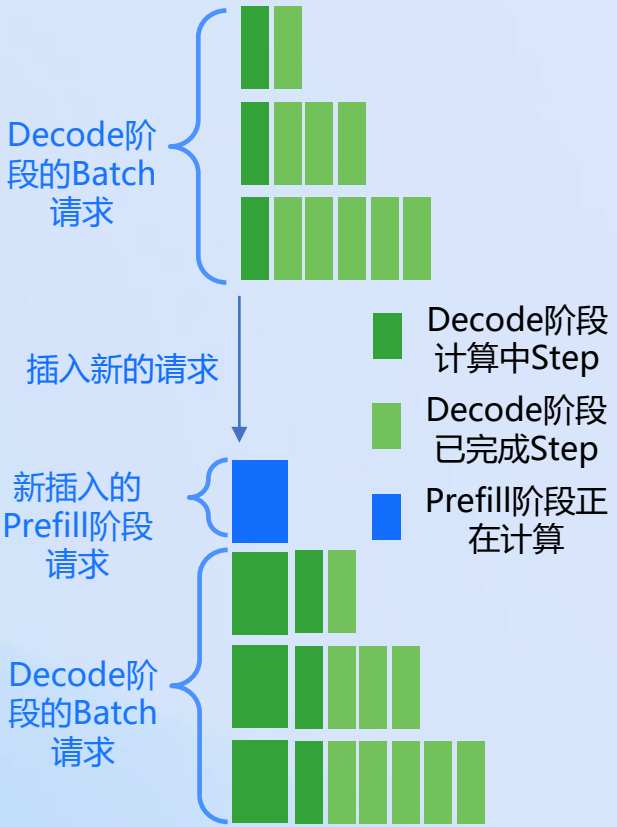
推理服务器需求

数据来源：Google《Q2 earnings call: CEO's remarks》

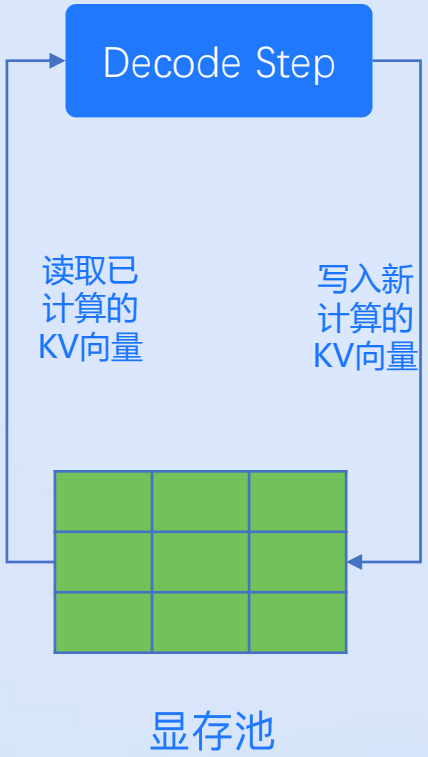
数据来源：IDC浪潮信息 《2025年 中国人工智能计算力发展评估报告》

大模型推理基础流程

集中式部署

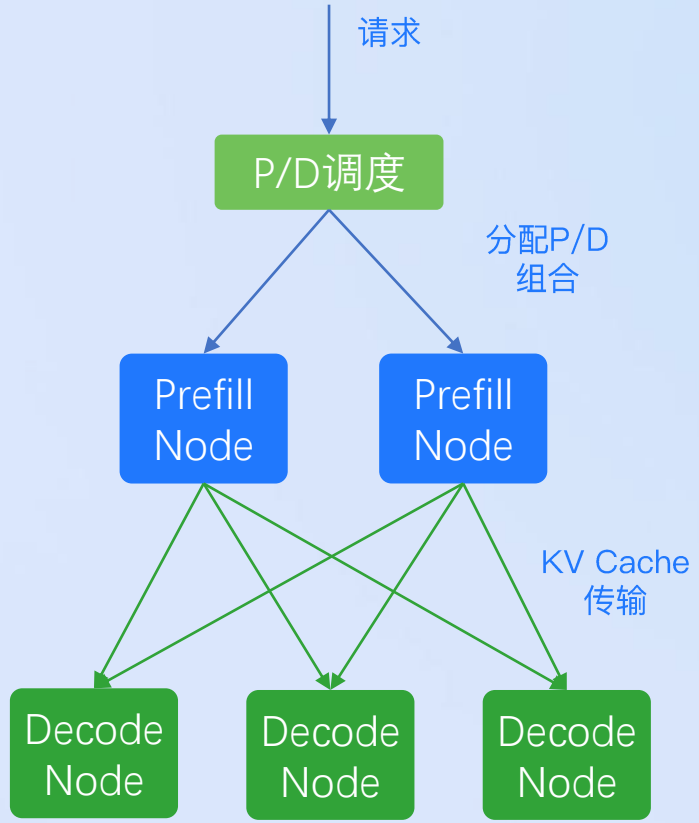


Prefill与Decode混合Batch: 动态插入提升吞吐



存储换计算：KV向量缓存加速Decode计算

分离式部署



PD分离：模型推理拆为Prefill、KVCache传输和Decode三个独立阶段

FastDeploy：飞桨高效大模型推理工具

- 生态兼容的统一接口设计
- 飞桨CINN编译器结合图优化
- 高性能低比特量化推理
- 低时延高吞吐投机解码框架
- 大规模多机PD分离架构
- 多国产化硬件后端支持

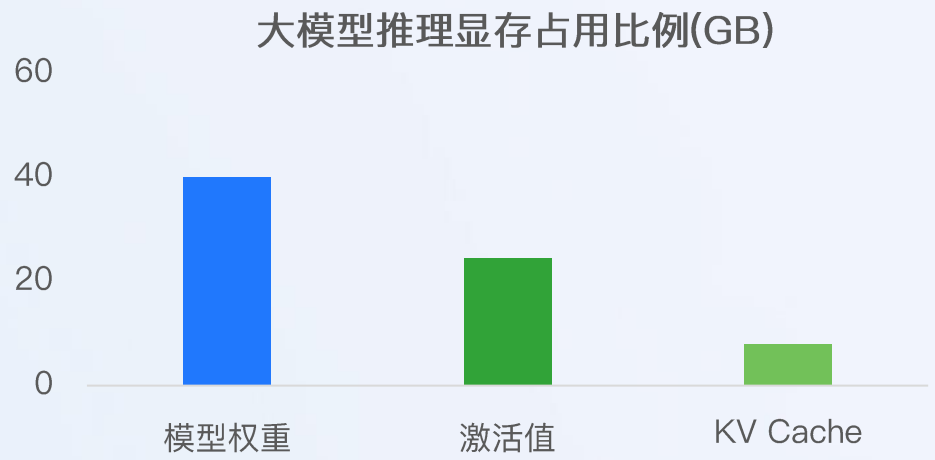
02

集中式部署架构优化

集中式部署架构优化


量化压缩
Weight Only INT8/INT4
Weight Only NT2 CCQ量化
KVCache量化
W4A8 MEPC量化
注意力量化
分布式上下文缓存
分块预填充
图优化与CUDA Graph
投机解码
多头解码
稀疏注意力 PLAS
国产芯片P800
混合精度计算
INT8高性能优化
查表计算优化

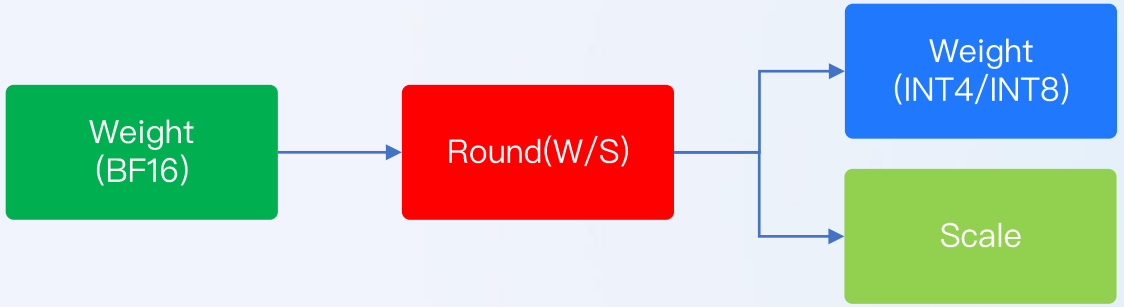
量化压缩：WINT2压缩算法CCQ (Convolutional Code Quantization)



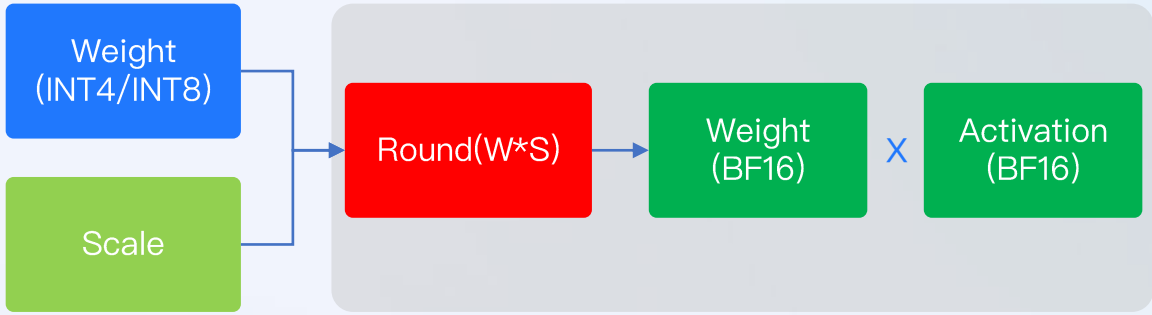
ERNIE-4.5-21B-A3B模型A800单卡BF16部署权重显存占用超过50%

模型	权重
ERNIE-4.5-21B-A3B	42GB
ERNIE-4.5-VL-28B-A3B	55GB
ERNIE-4.5-300B-A47B	562GB
ERNIE-4.5-VL-424B-A47B	789GB

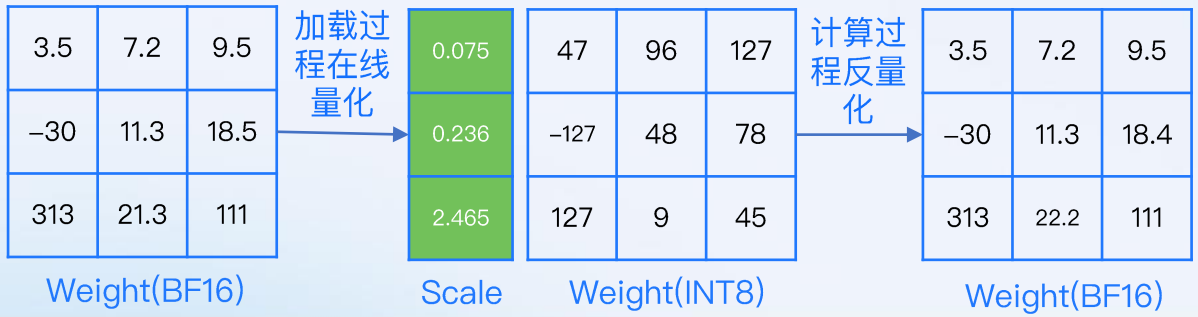
文心系列模型权重大小对比



加载时量化为INT4/INT8精度降低权重显存占用



Kernel读取Weight反量化为BF16计算保障精度



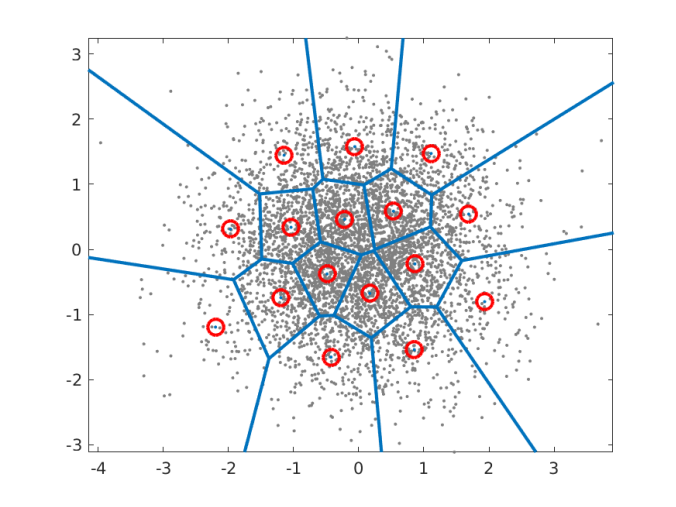
量化压缩：WINT2压缩算法CCQ (Convolutional Code Quantization)

解决标量低比特压缩**精度损失**和向量量化的**推理效率低**问题

标量量化

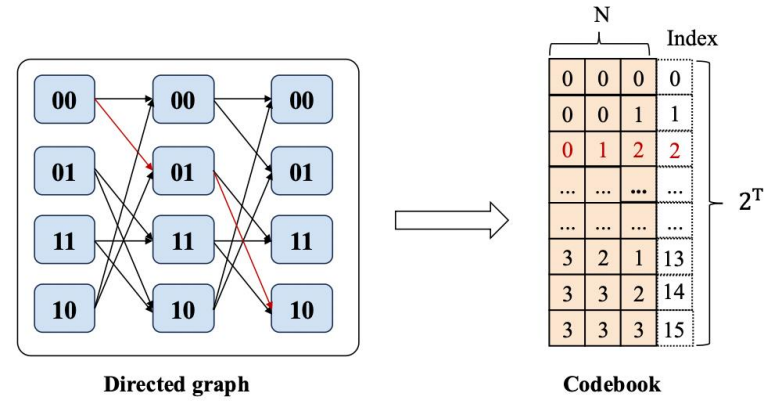
最常用且直观的量化方式，但在低比特如2 bit下精度失严重

向量量化



将向量分组映射，得到“索引表”和“码本”，保留向量数据相关性，解决**标量压缩的精度损失**。但同时也带来**推理查表性能**问题

卷积码



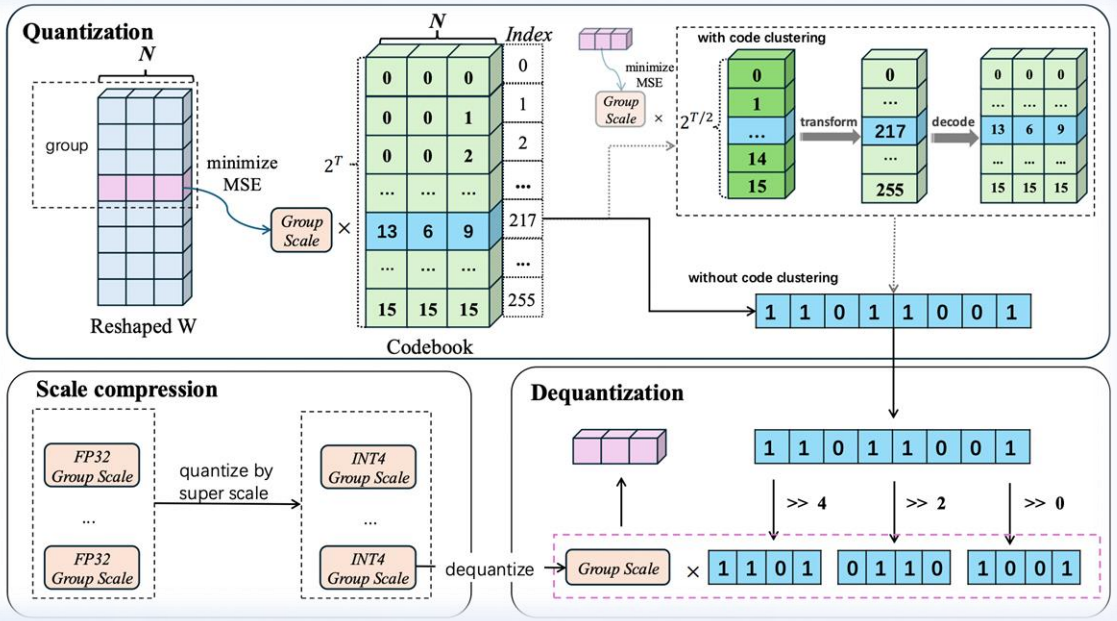
通过卷积码构建“**码本**”，消除反量化过程的查表，仅需**移位**和**掩码反量化**，优化推理性能

CCQ

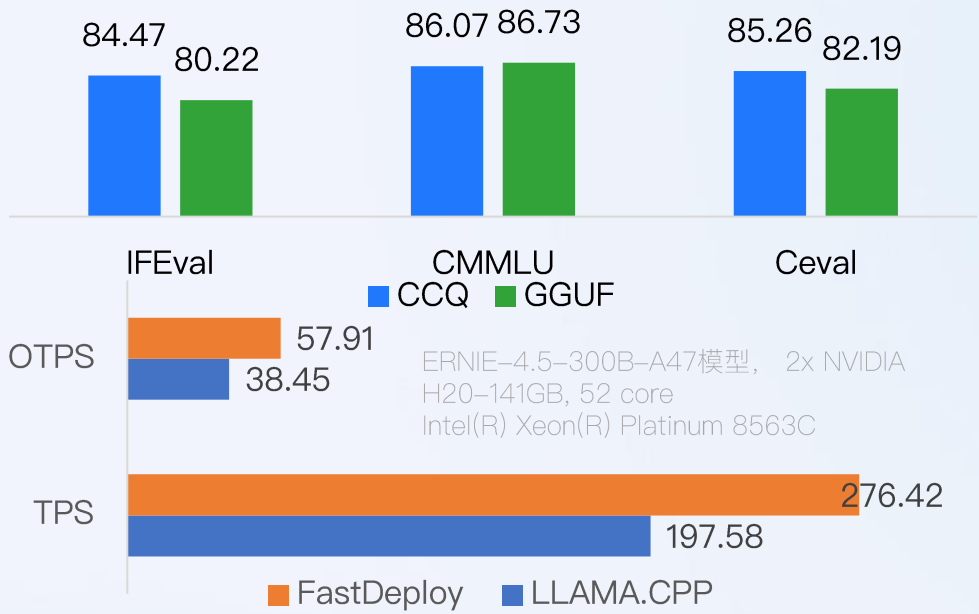
飞桨自研高性能低比特大模型WINT2压缩算法

图例来源：https://speechprocessingbook.aalto.fi/Modelling/Vector_quantization_VQ.html

量化压缩： WINT2压缩CCQ算法



CCQ算法量化与反量化过程



CCQ与GGUF量化对比

H20单卡即可实现文心4.5 300B大模型部署

ERNIE-4.5-300B-A47B	权重大小	GSM8K	DROP	MMLU
RTN(WINT8)	281GB	96.62	91.13	86.52
RTN(WINT4)	149GB	96.21	91.17	86.16
CCQ(WINT2)	89GB	95.30	88.34	82.31

量化压缩：MoE模型的W4A8多专家并行协同量化

MoE模型W4A8量化挑战

量化效率

- 专家多粒度小，计算资源利用不充分
- 激活所有专家，需要大量训练数据

量化效果

- 张量并行下异常点多节点分布问题
- 节点内的异常值影响

MEPC算法
(Multi-Expert Parallel Collaboration)

专家并行协同量化

- 多专家拼接并行校准，提升计算资源利用率
- 激活专家推算未激活专家系数，降低训练数据量

跨节点异常点转移算法

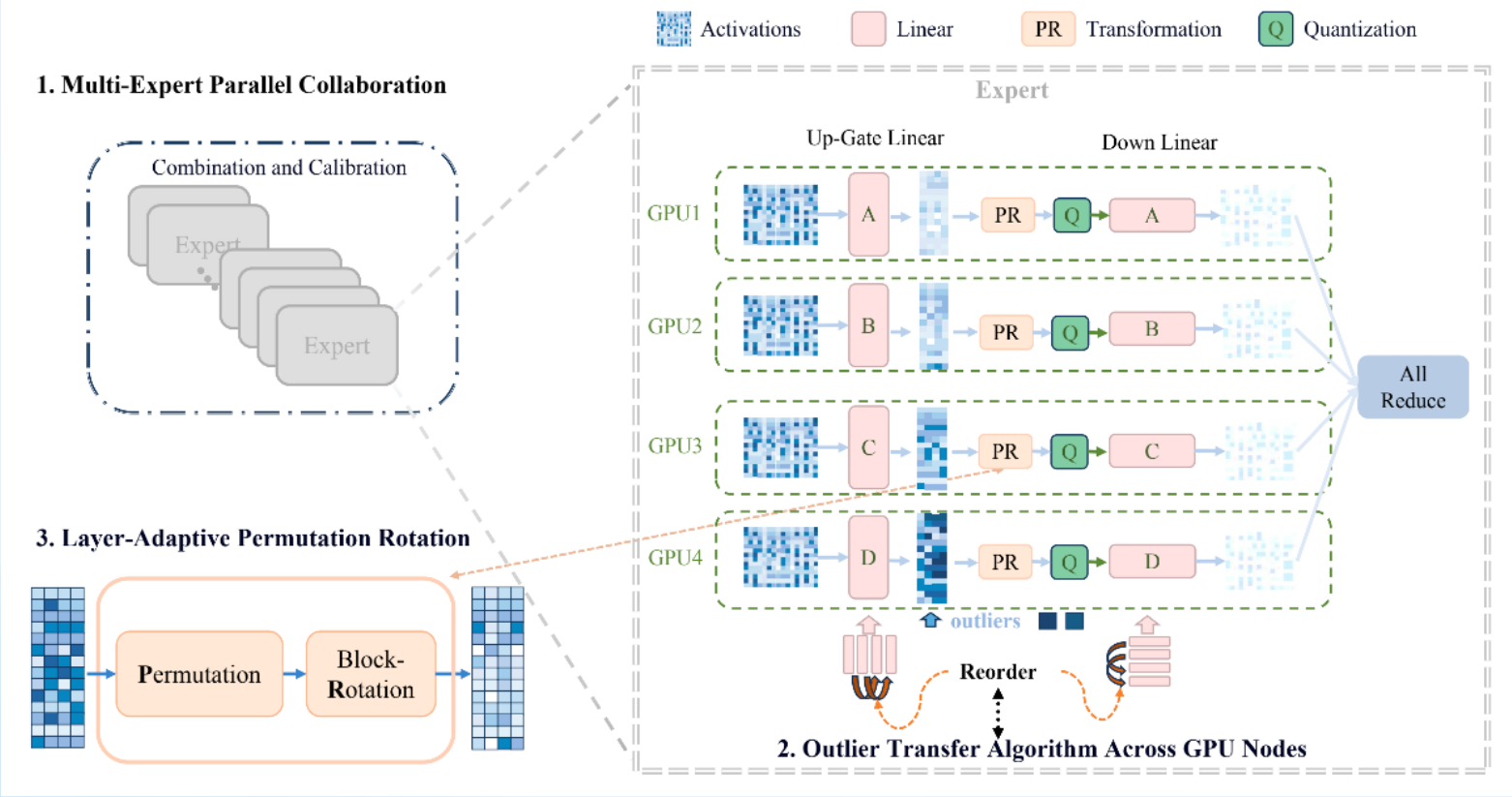
- 通过权重重排操作，所奖有的outliers迁移，集中到一张卡，减少整体的量化损失

层自适应的旋转转置

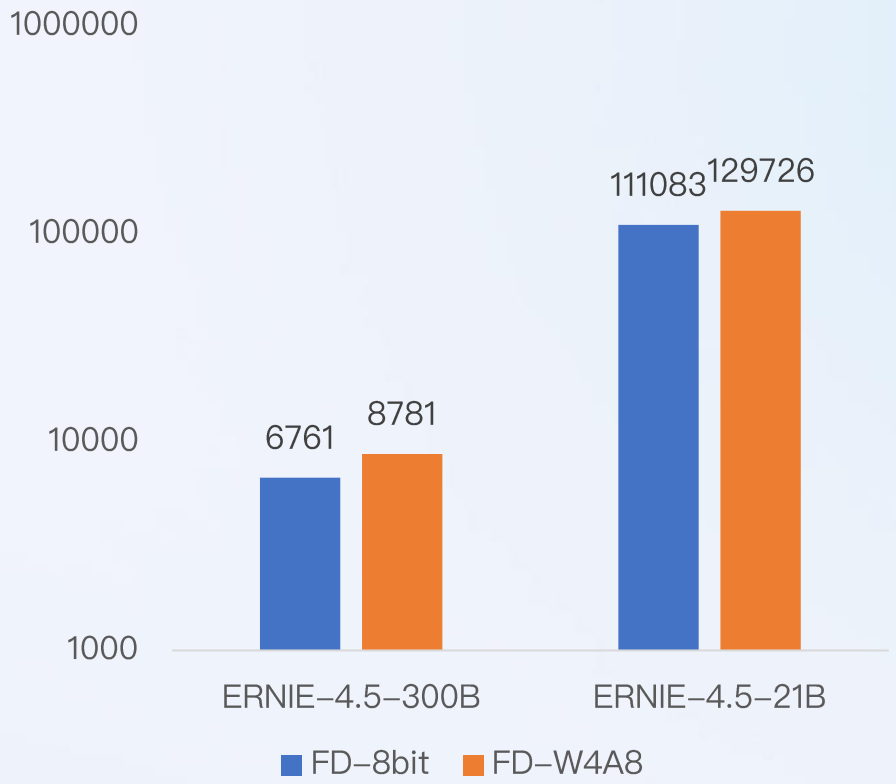
- 分块旋转与排列结合，兼顾推理效率和异常值平滑
- 根据各层特点自适应选择优先处理激活或处理权重

量化压缩：MoE模型的W4A8多专家并行协同量化

文心4.5模型精度接近无损，相比FP8吞吐提升17%~30%



基于MEPC算法解决MoE模型的量化效率与效果问题



W4A8与FP8推理吞吐对比

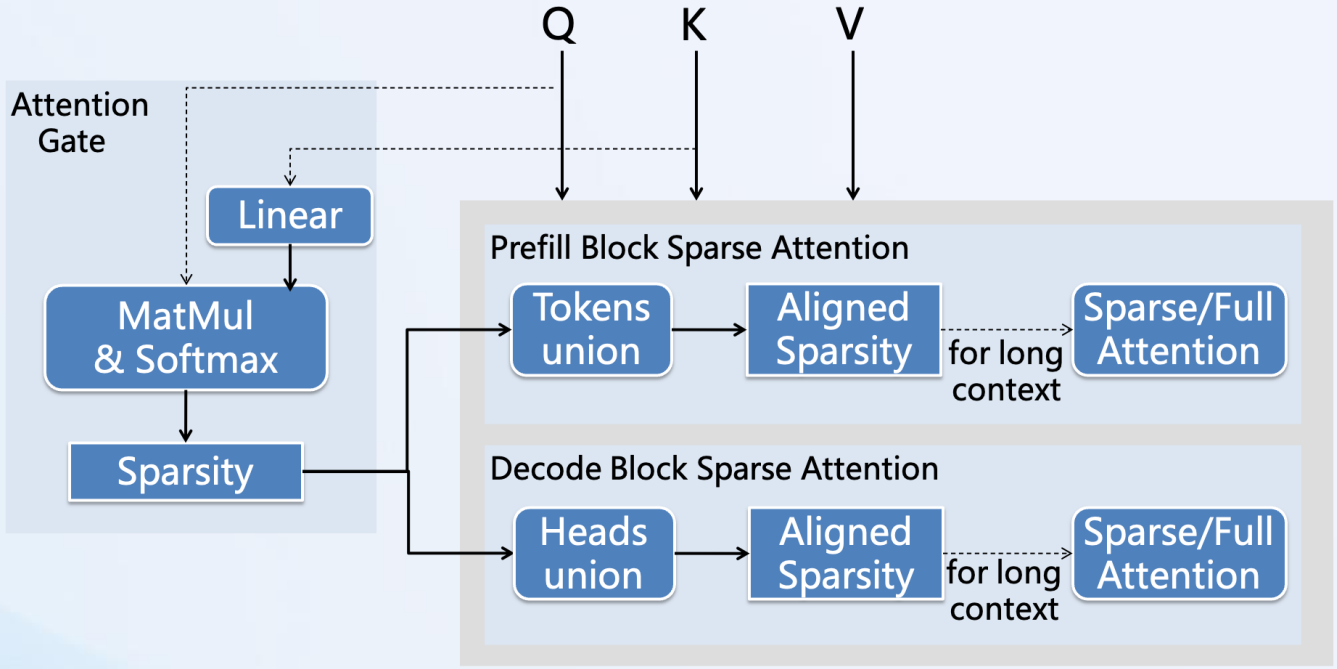
8x NVIDIA H800-80GB，CUDA 12.8
48 core Intel(R) Xeon(R) Platinum 8468V

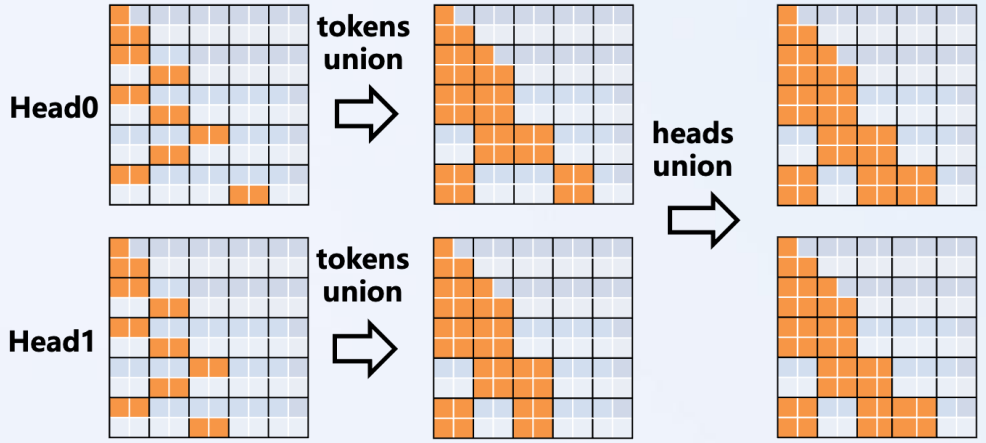
突破长文性能瓶颈：可插拔式稀疏注意力PLAS

低成本蒸馏微调注意力门控选择重要性块

根据输入长度自适应选择 Full Attention 或 Sparse Attention

Head及Token维度对齐实现高性能推理同时加速P/D 性能

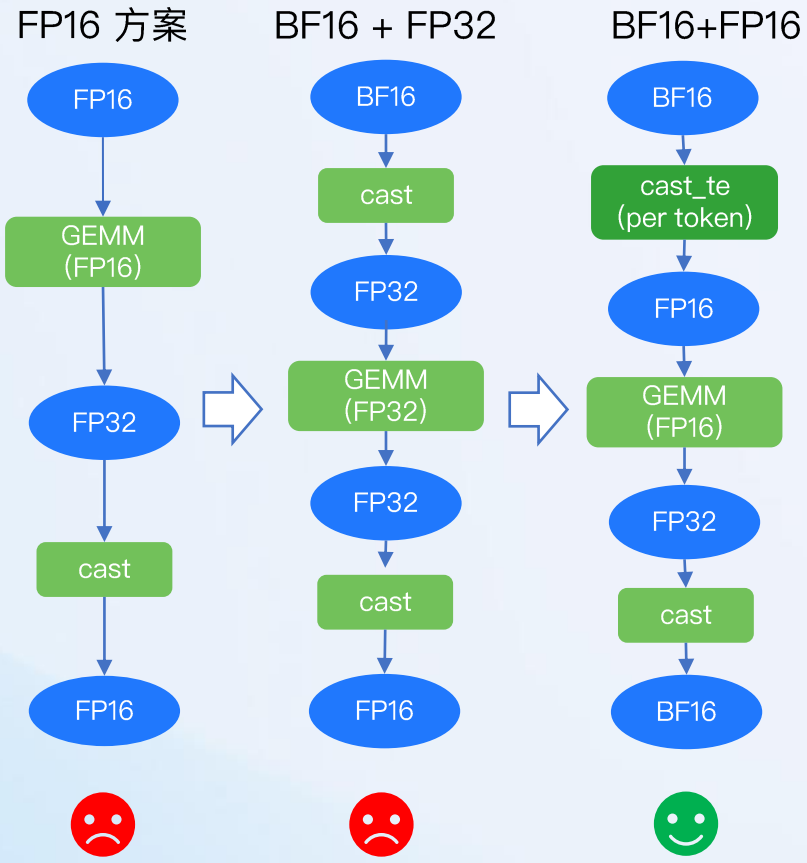




模型	精度（LongBenchV2）		性能		
	FullAttention	PLAS	TTFT	TPOT	QPS
ERNIE-4.5-300B-A47B-128K	40.02	40.05(+0.03%)	-30%	-34%	+23%
ERNIE-4.5-21B-A3B-128K	30.82	30.41(-0.41%)	-48%	-31%	+48%

国产昆仑P800芯片性能优化

优化1：整网BF16存储+FP16计算方案

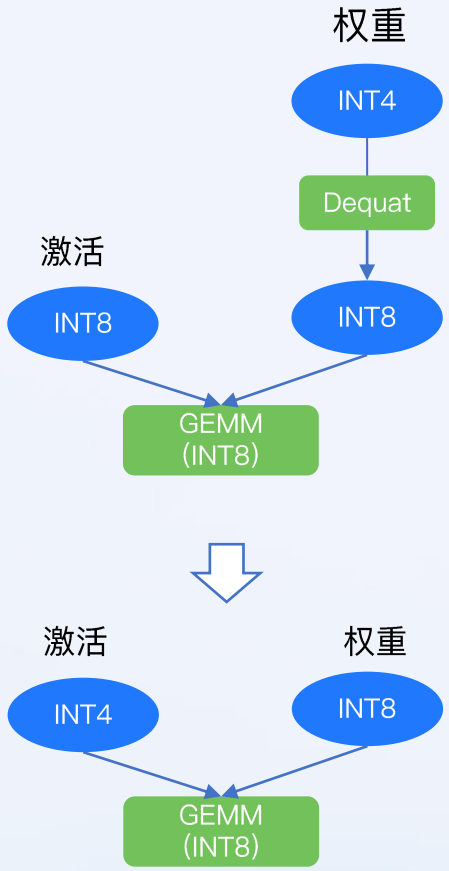


模型效果无法达标

性能劣化30%

效果性能达标

优化2：INT4*INT8直接计算

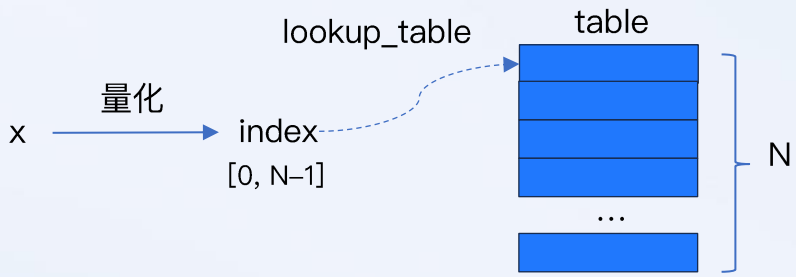


利用硬件特性，发挥算力优势，**性能提升10%！**

优化3：基于查表优化复杂函数计算

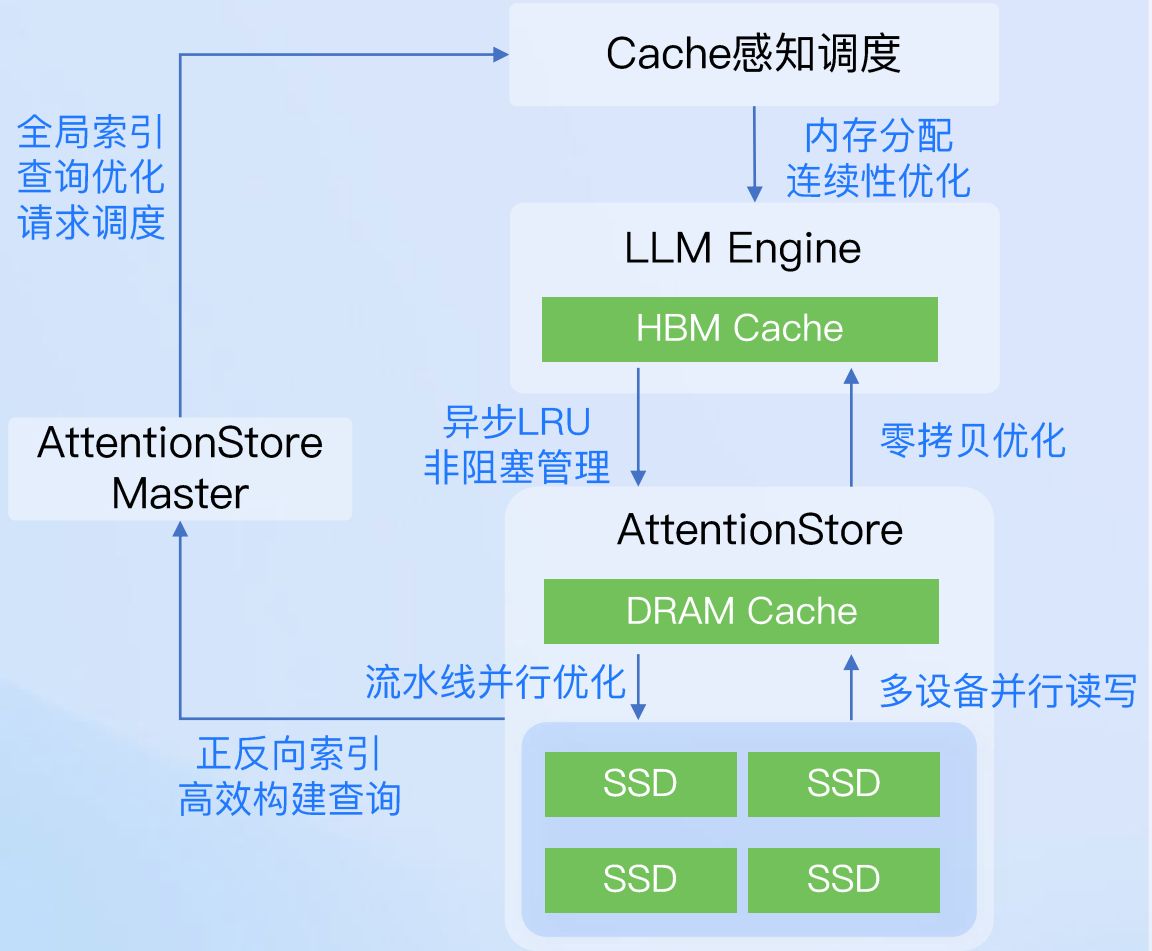
$$Swish_\beta(x) = \frac{x}{1+e^{-\beta x}}$$

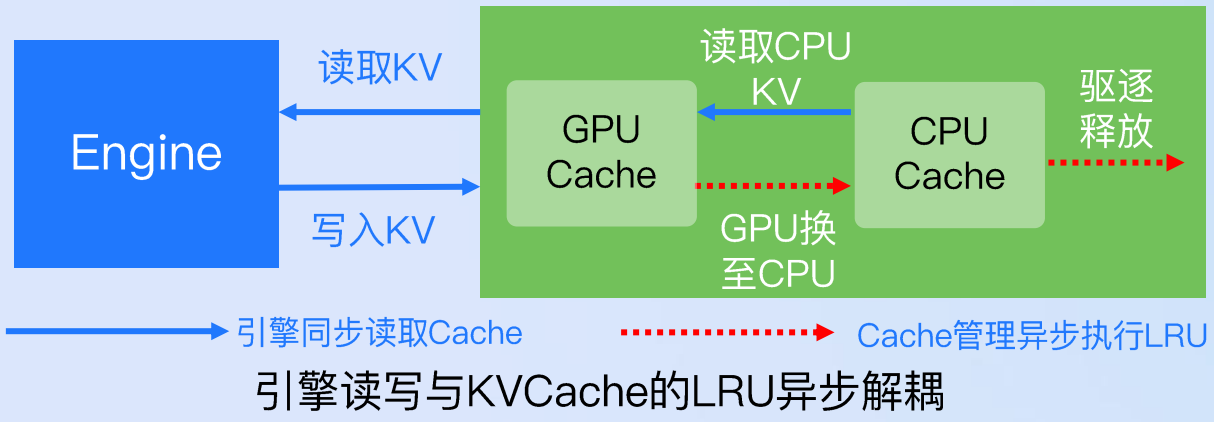
$$Softmax(x) = \frac{e^{x_i}}{\sum_j e^{x_j}}$$



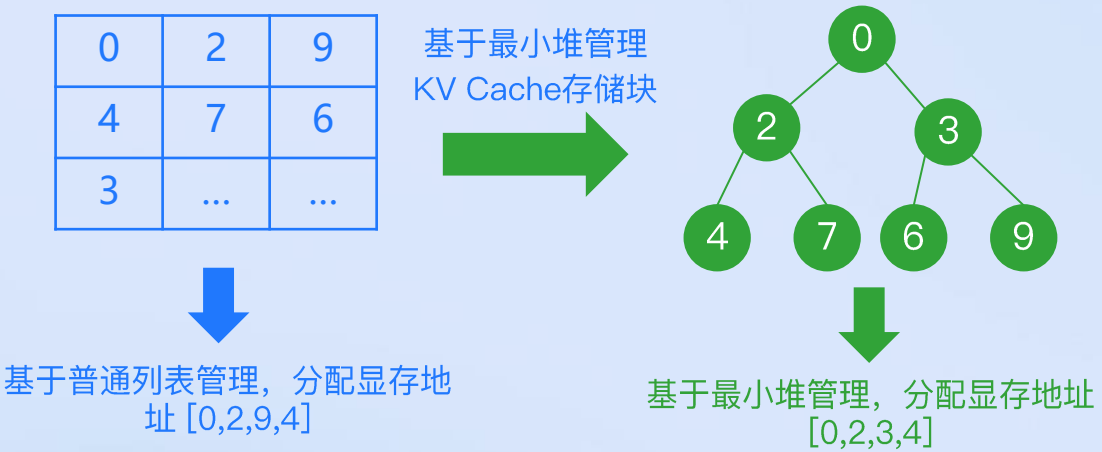
利用硬件查表指令，实现对复杂函数快速计算，**性能提升5%！**

分布式多级上下文缓存高效管理





引擎读写与KVCache的LRU异步解耦



优化后显存分配地址连续性强，更友好跨介质间的数据拷贝

03

分离式部署架构优化

分离式部署需求分析

集中式推理问题

Prefill与Decode所需资源类型不同

Prefill为计算密集型，依赖算力强劲的GPU加速
Decode为存储密集型，依赖更高的内存带宽

Prefill混合并行带来的性能波动

Prefill与Decode混合并行的『木桶效应』：长输入的Prefill任务导致Decode阶段时延急剧上升

资源浪费和成本增加

生产环境中Prefill与Decode的资源需求随输入输出长度动态变化，集中式的部署无法灵活按需配置资源

1

独立优化与异构部署

针对Prefill与Decode分别做推理上的优化，采用不同的硬件满足计算和存储需求

2

更优的性能与更可靠的SLO

摆脱『木桶效应』，更快的请求响应速度，同时解决长输入请求带来的时延上升问题，保障稳定性

3

精细化资源管理

通过系统输入输出长度变化，以及对首Token，Token间时延的要求，精细调整Prefill/Decode资源

分离式部署架构

分离式部署挑战

负载均衡

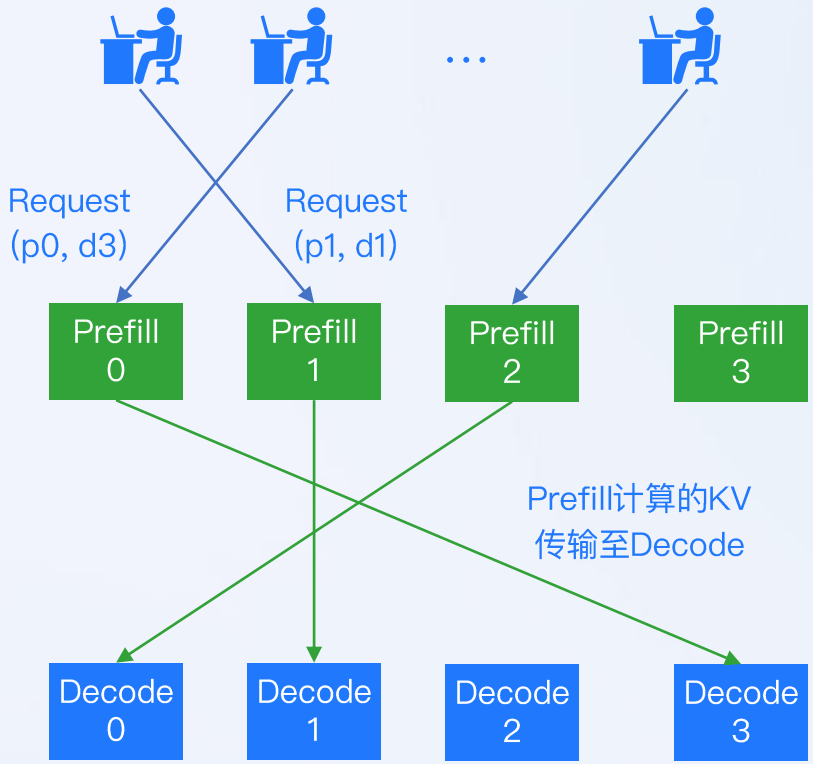
相比集中式部署，需分别平衡Prefill与Decode的负载；针对MoE模型EP并行，需进一步平衡各专家的负载

KV传输与通信

Prefill计算完的KV需要跨进程跨节点传输给Decode计算，额外引入传输耗时。在MoE模型采用EP并行后，进一步引入专家间数据传输

系统运维

分离式部署架构中，各Prefill和Decode不再是独立的实例节点，Prefill需感知所有的Decode实例，同时处理在KV Cache传输或者Decode实例异常问题



PD分离下，请求先在Prefill计算完生成首Token，再在Decode继续生成剩余的Token

分离式部署与分布式上下文缓存加速推理

负载均衡

Prefill负载均衡

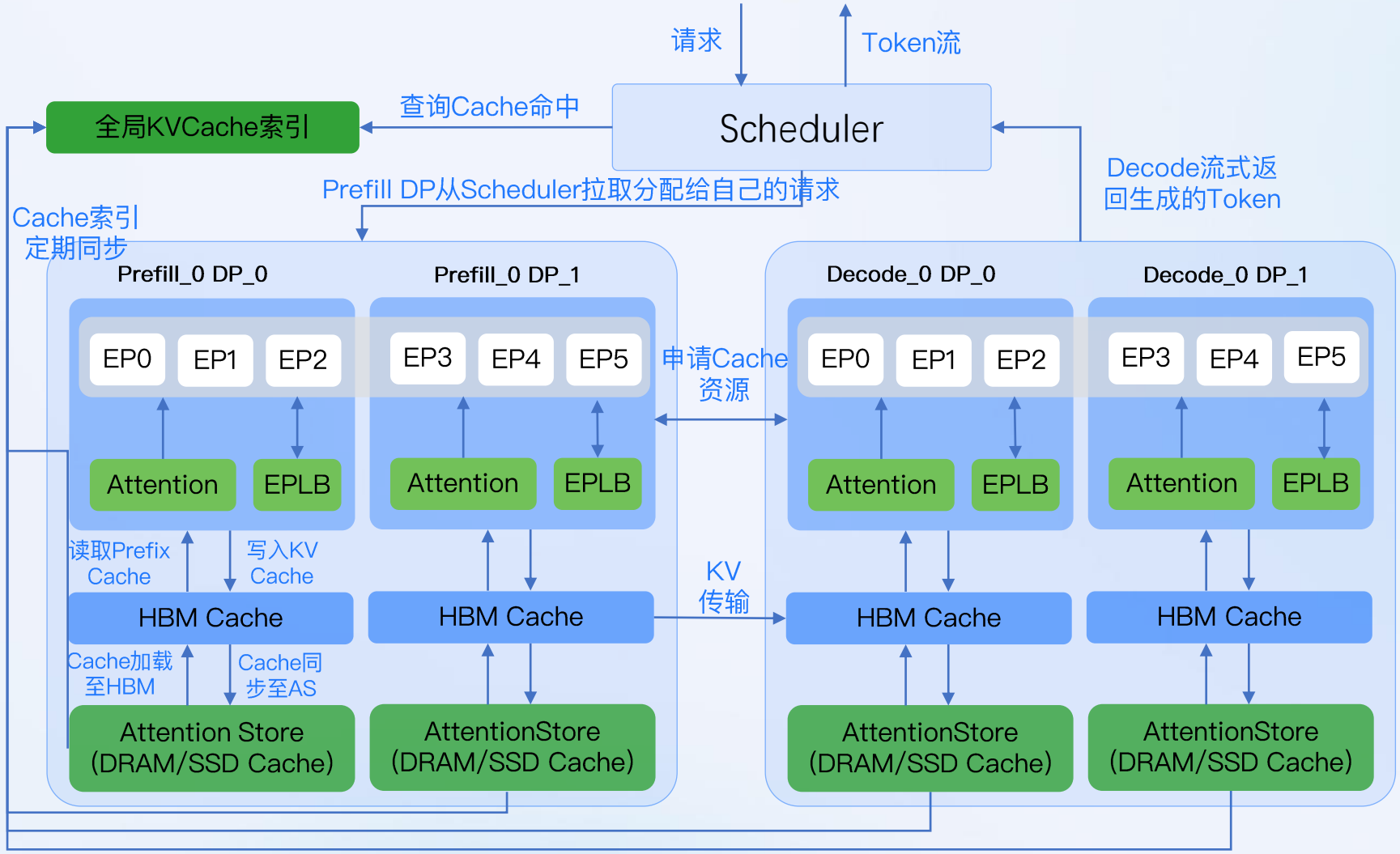
Prefill各DP负载指标为未完成请求Token总数与Cache命中Token数差值，Scheduler以此为均衡条件进行调度

Decode负载均衡

Decode各DP负载指标为未守成请求数，Scheduler以此为均衡条件进行调度

专家负载均衡

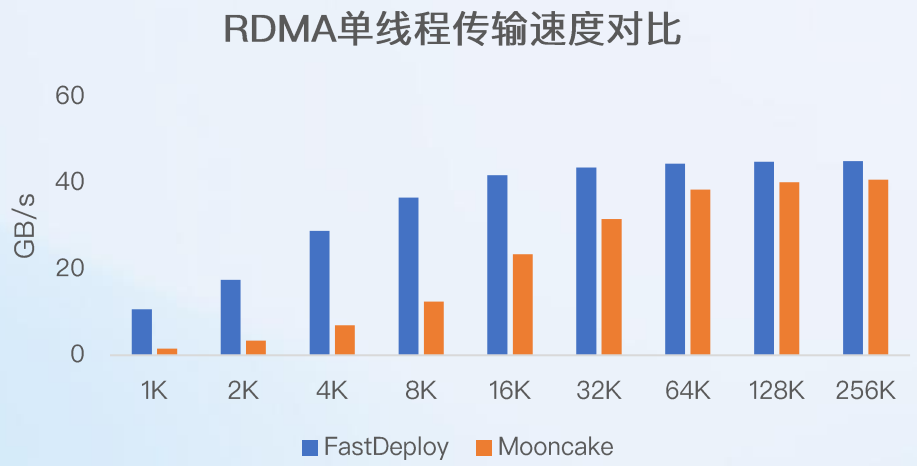
收集Prefill/Decode各DP上专家负载，定期重排各节点及各卡上的专家权重



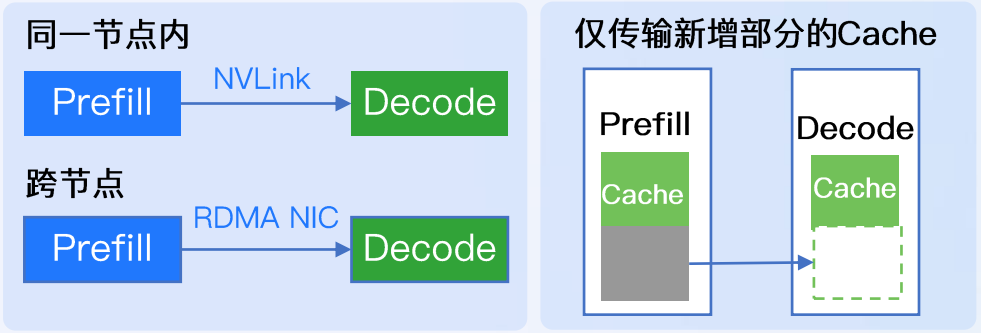
KV传输与通信优化

轻量高效的RDMA传输库

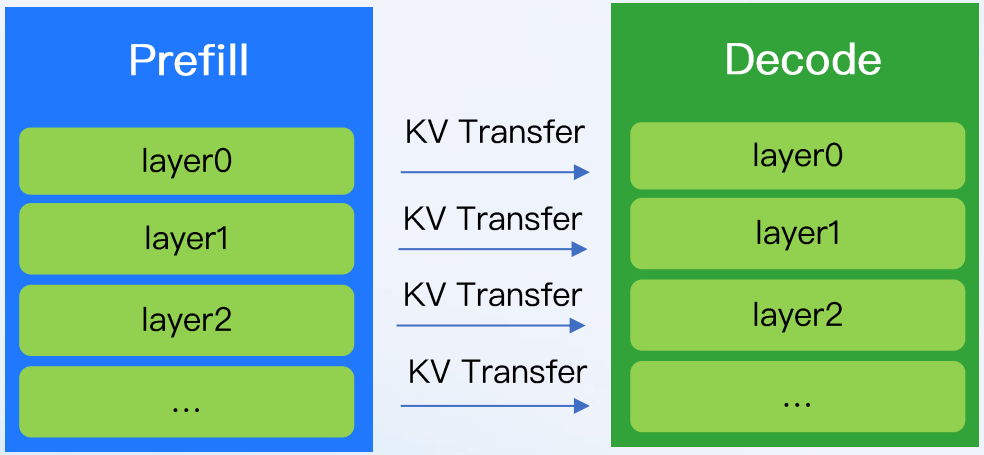
- 轻量： 部署简单轻量，仅需基础的RDMA运行环境即可使用
- 高性能：减少CQE数量和支持PCIe Relaxed Ordering等优化，单线程传输效率更高，多线程带宽打满
- 动态建连：支持PD比例动态扩缩容



传输协议自适应的增量Cache传输



Layerwise方式掩盖传输耗时



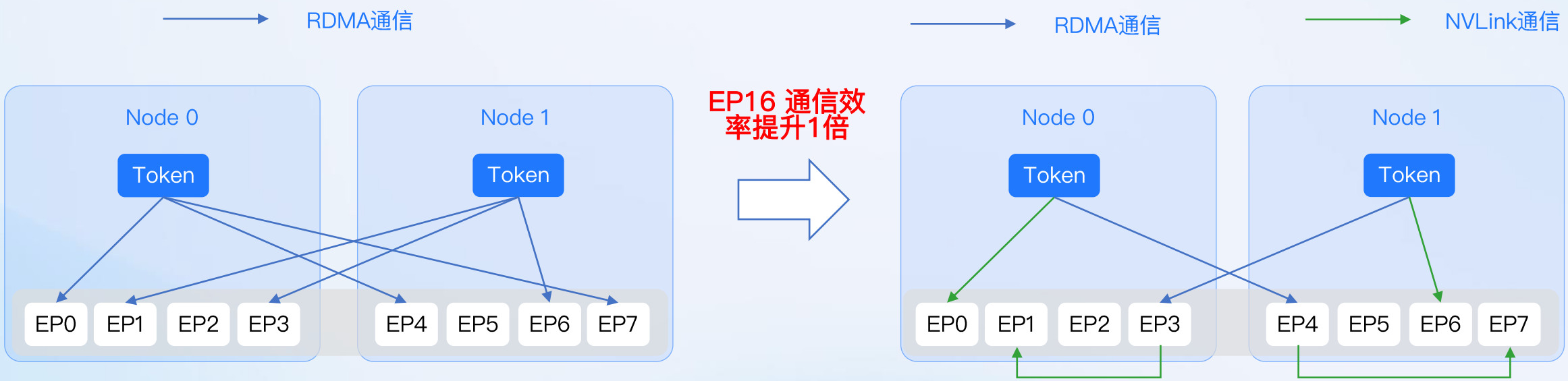
基于DeepEP的自适应双阶段通信优化

纯RDMA通信带宽受限

- EP并行度低的场景下，存在大量节点内通信
- 跨节点通信存在重复使用RDMA进行传输需求

自适应双阶段通信优化

- 节点内通信采用NVLink，节点间通信采用RDMA
- 跨节点通信第一阶段基于RDMA将数据传输至对应节点对应GPU卡，第二阶段在对应节点内基于NVLink将数据传至其它GPU卡



分离式系统稳定性保障

稳定性保障

重调度与实例更新

- 重调度接口支持Prefill/Decode即时通知调度处理异常请求
- 专家负载均衡定期滚动更新，实例异步权重加载

续推Agent

- 针对实例软硬件问题，即时将异常请求与已生成结果拼接进行续推，避免PV Lost同时，优化计算资源

抢占Agent

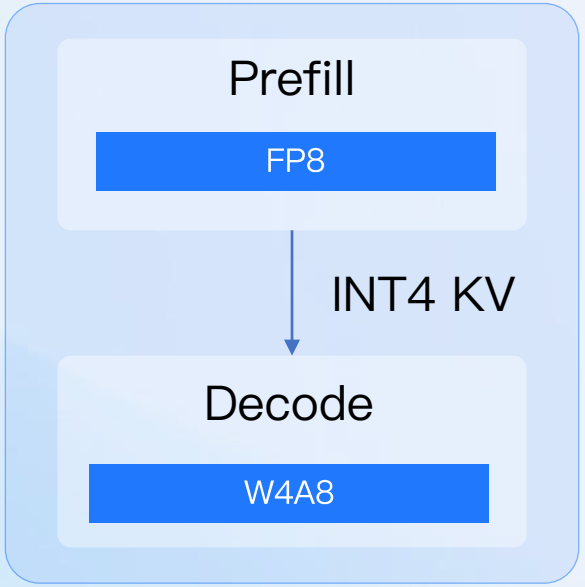
- 针对集群资源动态调整需求，结合续推Agent，实现推理实例实时退出

文心大模型的分离式部署优化

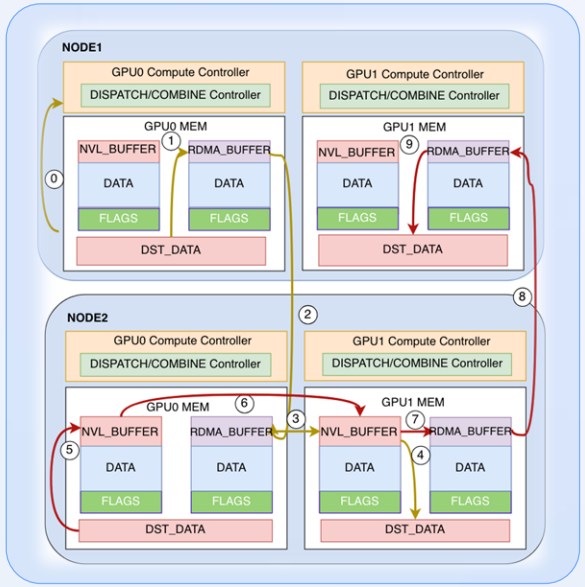
PD分离分布式架构，文心性能再提升

ITPS 56K，OTPS 21K，TPOT 50ms

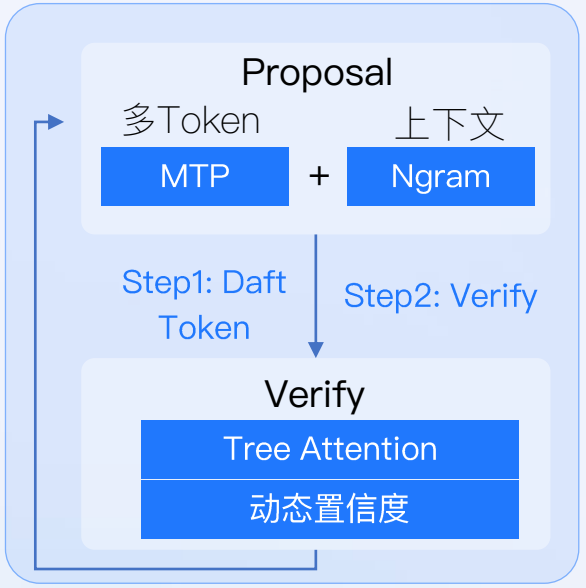
ERNIE-4.5-300B-A47B



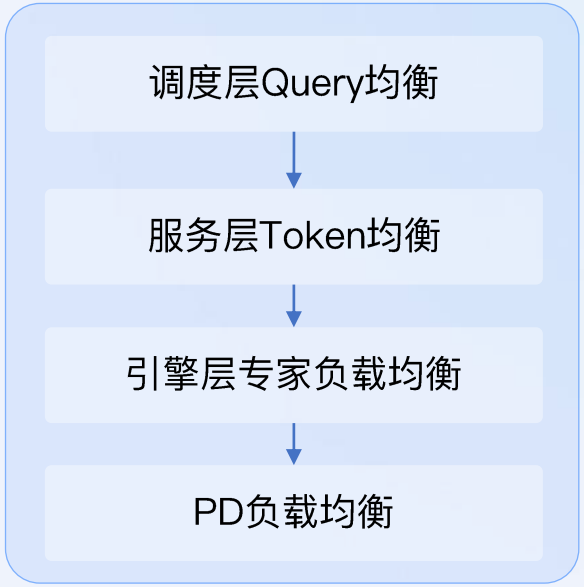
Prefill/Decode/KV传输优化



自适应双阶段通信优化



投机解码加速优化



多级负载均衡优化

04

总结与展望

文心飞桨推理架构的演进史

未来计划

四级缓存

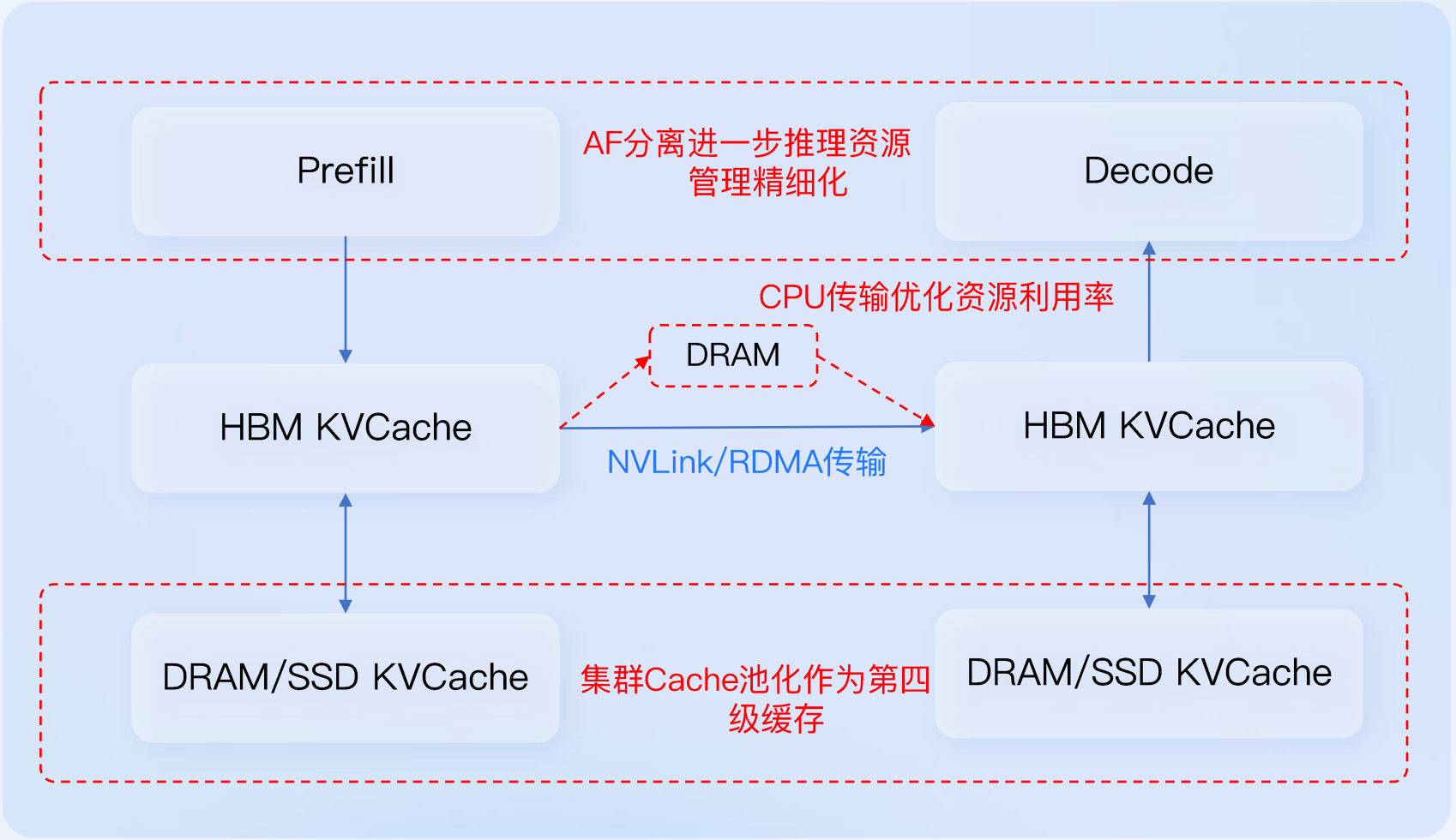
通过集群KVCache池化支持推理实例从其它推理实例获取KVCache加速Prefill计算

KV传输

GPU直传依赖提前占用Decode显存资源导致资源利用不充分，新增CPU传输优化吞吐

AF分离

针对Attention与FFN任务的差异进一步精细化部署的资源管理，降低推理成本



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