InfoQ 极客传媒

# 面向多种算力平台的大模型量化推理优化技术

演讲人：汤雄超 博士

清程极智 联合创始人、CEO

AiCon

全球人工智能开发与应用大会

# 演讲人简介

## 汤雄超 博士

本科和博士毕业于清华大学计算机科学与技术系。长期在工业界从事研发工作，目前担任清程极智公司首席执行官。主要研究领域为并行计算的性能优化，例如高性能的大模型训练和推理部署软件等。发表ASPLOS、SC、PPoPP、TPDS等CCF-A类论文十余篇，申请发明专利20余项，获得ACM中国SIGHPC优博奖，ACM戈登贝尔入围奖、深圳市高层次专业人才等荣誉。

## 关于清程极智公司

源自清华大学计算机系，2023年成立，致力于实现高效、普惠的人工智能。业务聚焦于智能算力基础设施，主要产品包括「爱评 AI Ping」大模型服务评测与API调用平台、「赤兔 Chitu」大模型推理部署解决方案、「八卦炉 Bagualu」智能计算软件栈等。持续服务芯片企业、智算中心、大模型企业、AI应用企业等人工智能领域客户。

# 目录

# 极客邦科技 2026 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

## 4月 · 北京 · 1200人

QCon 全球软件开发大会

会议时间：4月16-18日

- Agentic Engineering
- AgentOps
- 下一代模型架构与推理优化
- AI 原生基础设施
- 知识工程实践
- AI 安全

## 6月 · 上海 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：6月26-27日

- AI Infra 系统工程
- 多 Agent 协作与实践
- 多模态融合
- 模型训练与推理创新
- 数据平台与特征服务

## 8月 · 深圳 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：8月21-22日

- Agentic AI
- 轻量化与高效推理
- 多模态应用
- AI + IoT 场景实践
- AI 工业化落地

## 10月 · 上海 · 1200人

QCon 全球软件开发大会

会议时间：10月22-24日

- AI Agent
- Vibe Coding
- 智能可观测
- 推理基建
- 模型攻防
- AI x 创造力

## 12月 · 北京 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：12月18-19日

- 大模型架构创新
- 多模态 AI 产业融合
- 具身智能
- AI for Science
- 大模型安全

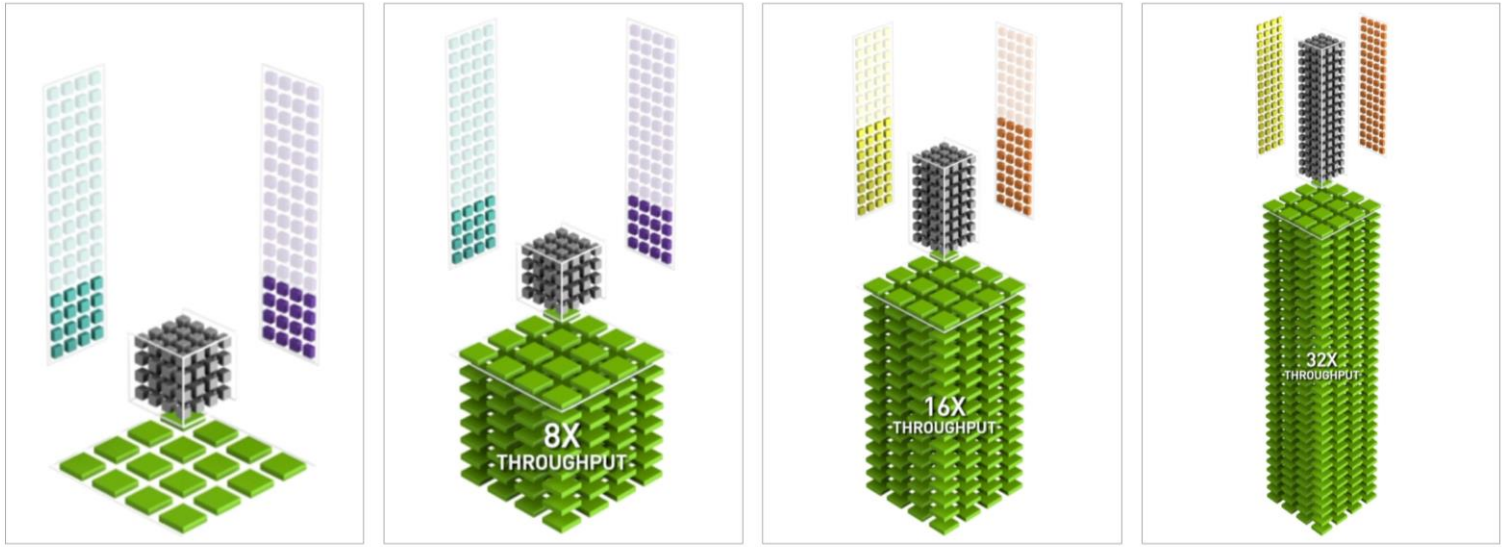
InfoQ 极客传媒

01

# 大部分AI算力都将是低精度算力

# 为什么需要低精度数据类型和低精度算力

随着摩尔定律失速，芯片设计依赖低精度算力提高性能和降低功耗
随着大模型体积增长，模型推理依赖低精度数据类型加速访存和计算，缩短推理时间



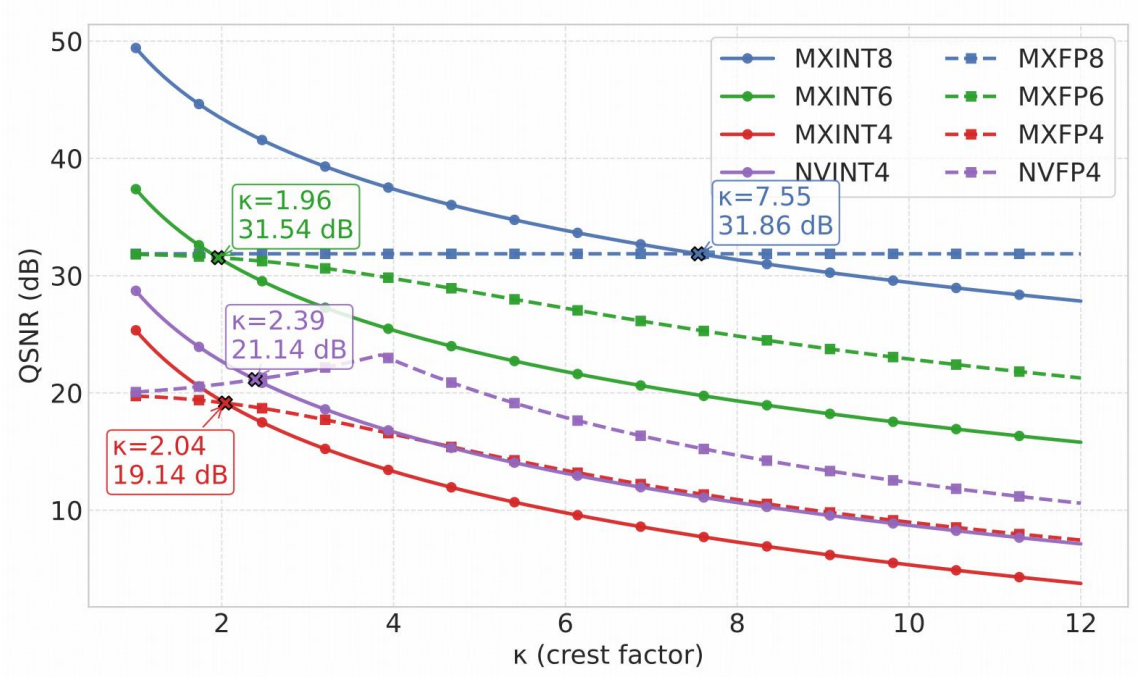

图片来源：NVIDIA网站

# 低精度浮点数与低精度整数的取舍

行业研究显示，低精度浮点数在大多数情况下能够比低精度整数更好地近似大模型的权重参数
与此同时，整数计算比浮点数计算更易实现，功耗也更低

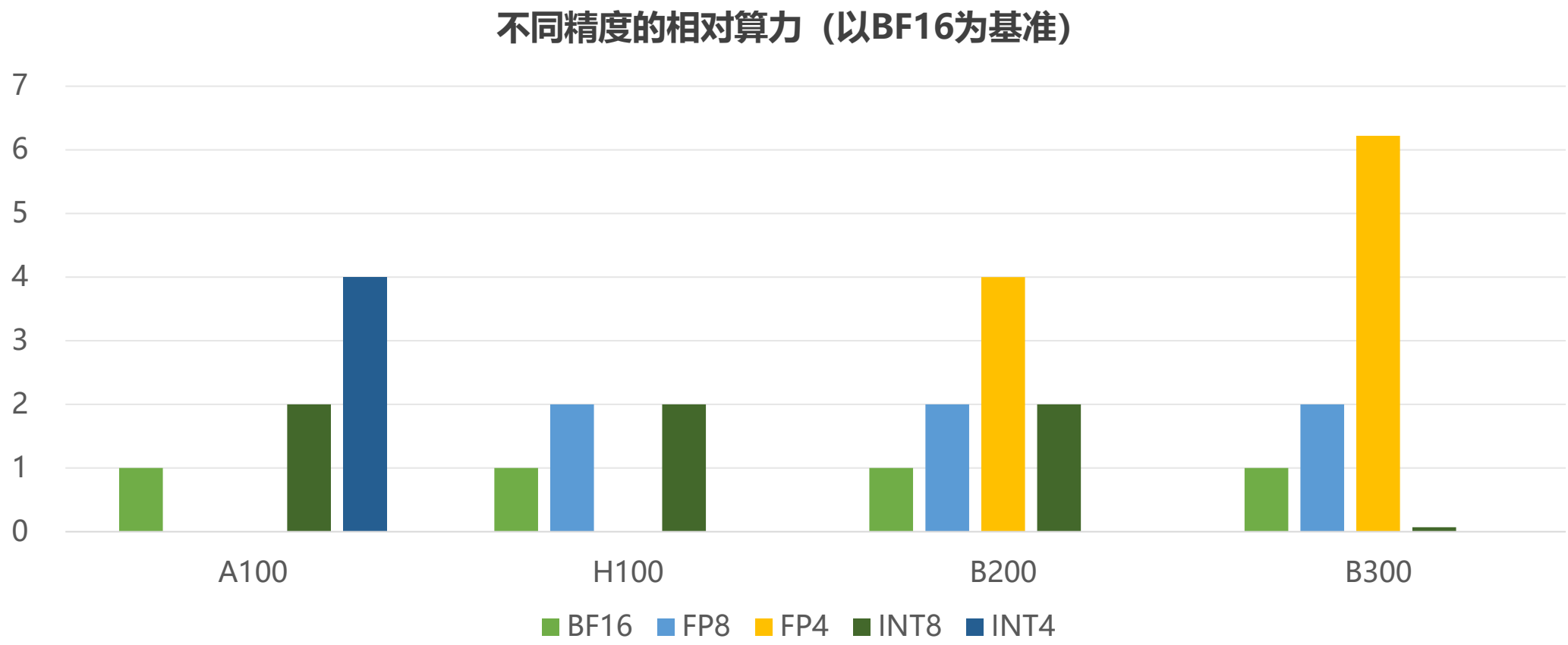
$$\mathrm{QSNR} = -10\log_{10}\left(\frac{\|\mathbf{X}-\mathbf{X}_q\|^2}{\|\mathbf{X}\|^2}\right)$$

$$\kappa := \frac{\max(|\mathbf{X}|)}{\sigma}$$




图片来源：https://arxiv.org/pdf/2510.25602，Mengzhao Chen等

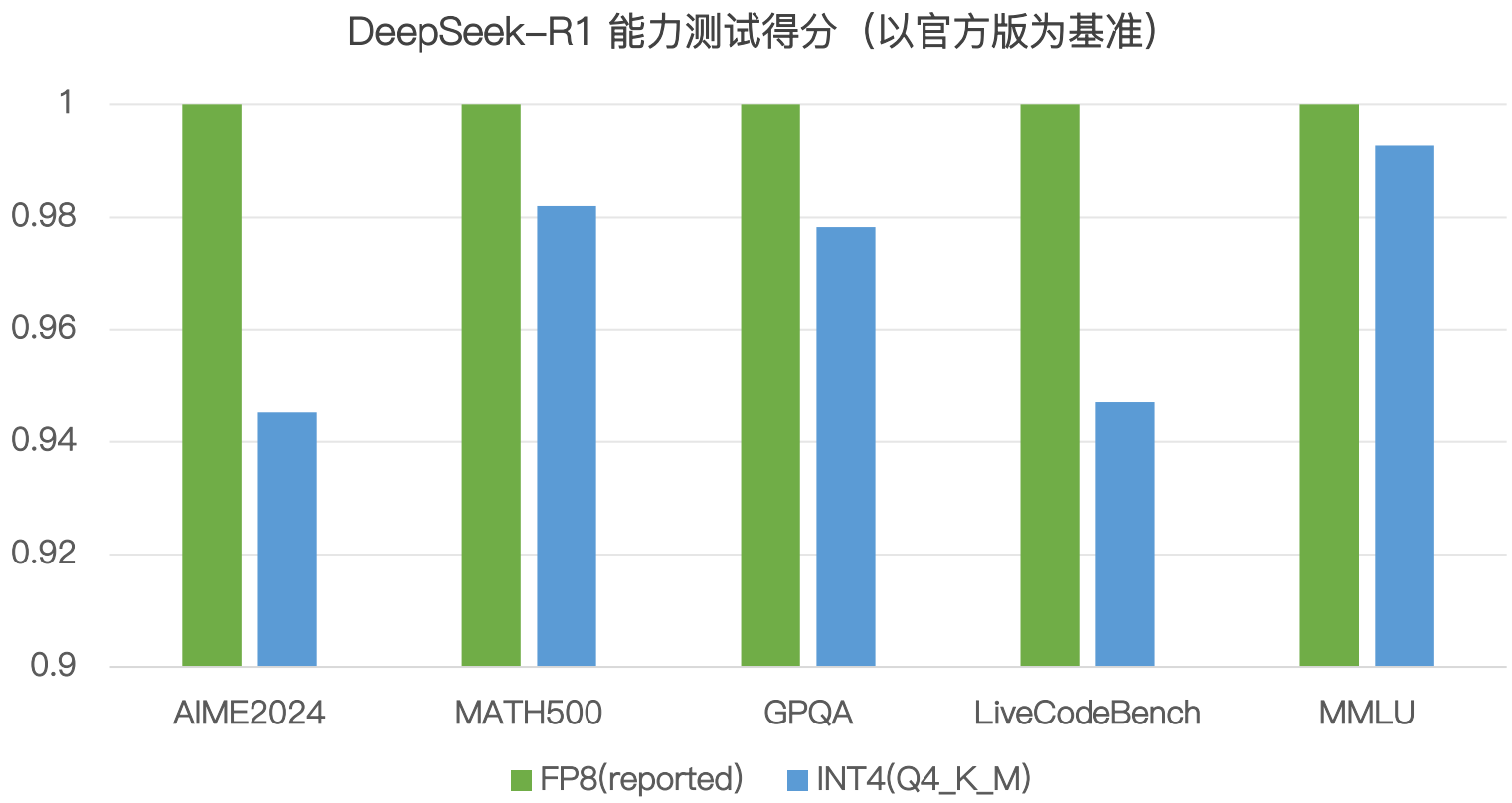
# 低精度浮点数算力正在获得主导地位



02

# 好的量化算法可以保持模型能力

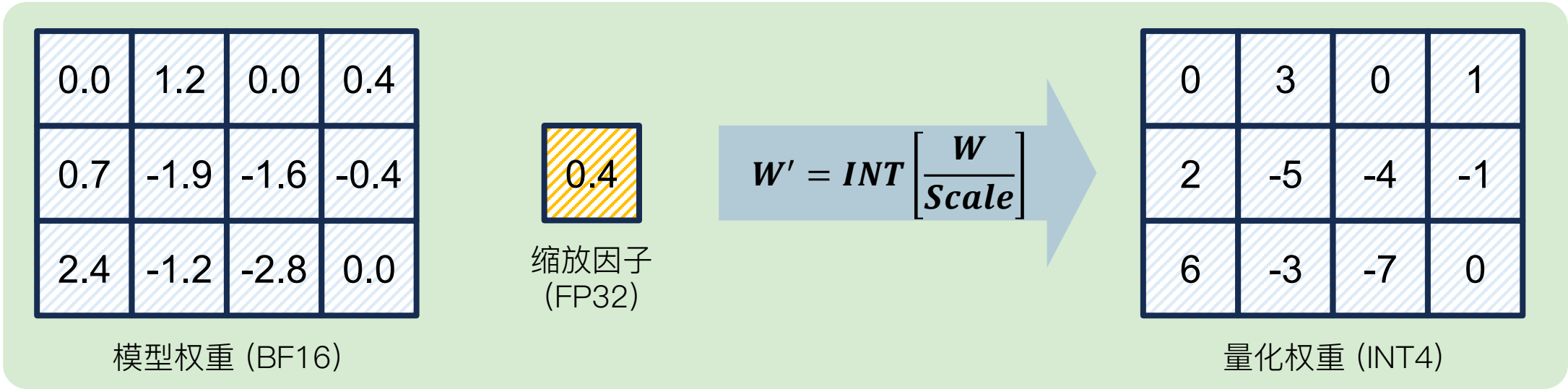
# 低精度类型算得快，但可能损害模型能力



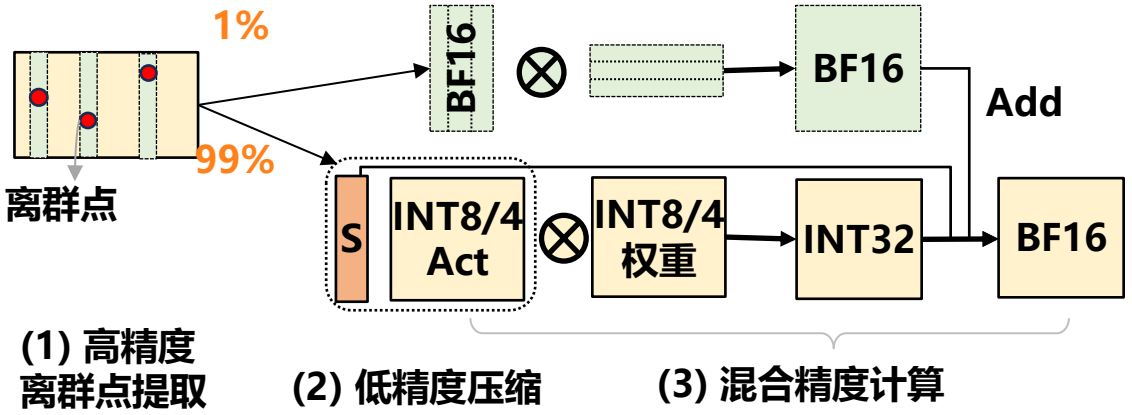

资料来源：https://arxiv.org/html/2505.02390v1, Enbo Zhao等

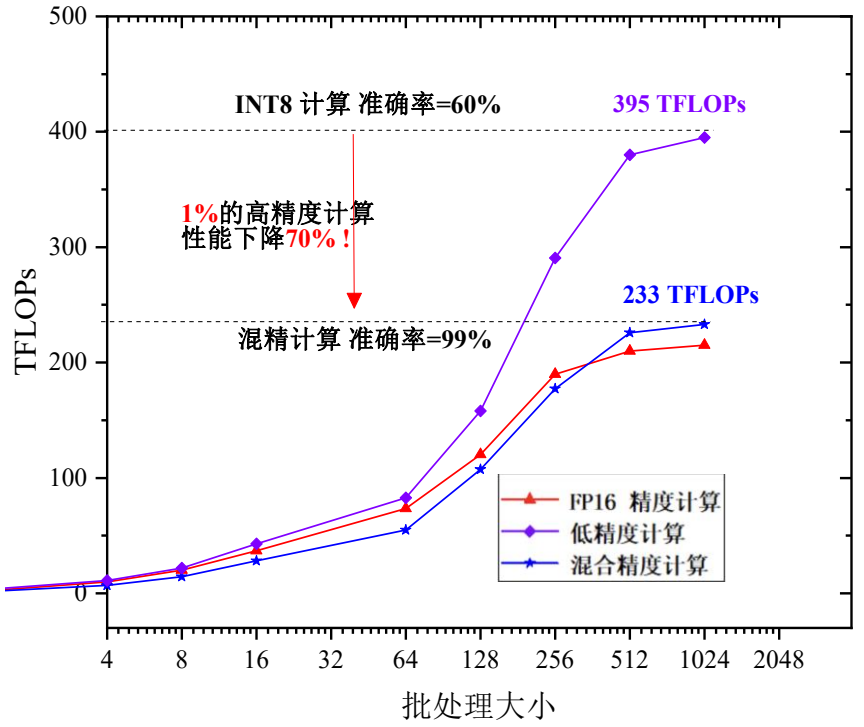
# 退一步海阔天空？WxAy 的纠结

- 模型量化将张量从高精度数值表示转换为低精度数值表示，减少比特位数
- 模型量化分类
  - **仅权重量化**（W4A16、W4A8）：减少模型体积，使用高精度计算单元**（访存加速、保持模型能力）**
  - **权重–激活联合量化**（W8A8、W4A4）：充分利用高吞吐的低精度计算单元**（访存和计算都加速、模型能力受损）**



# 混合精度量化，简单的想法与复杂的现实




**优势**

- 几乎没有精度损失
- 节省显存用量（模型大小减少两倍）

**不足**

- 实现高效的混合精度算子面临挑战

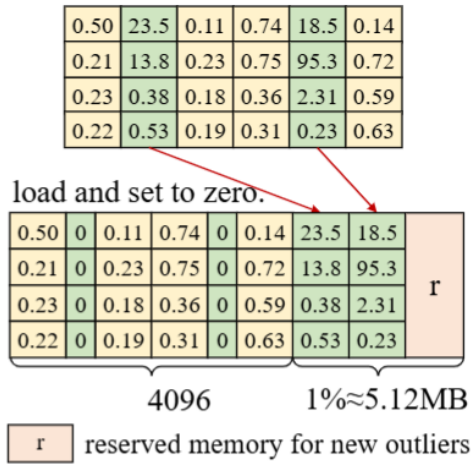
混合精度推理库：Bitsandbytes

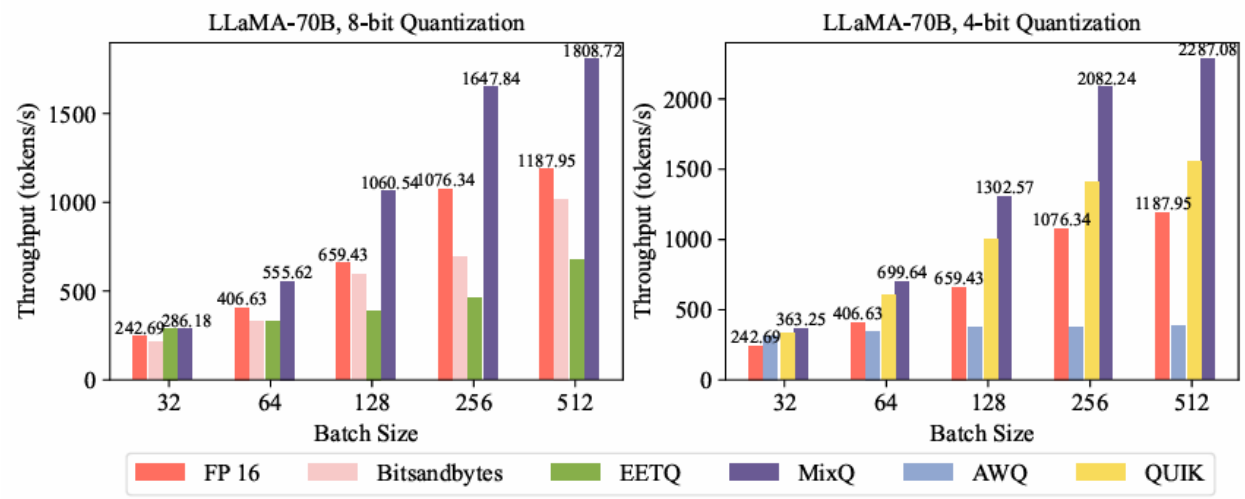
**出现 1% 高精度计算会导致 70% 的性能下降**

资料来源：[SC'24] MixQ: Taming dynamic outliers in mixed-precision quantization by online prediction. Yidong Chen等

# MixQ 的尝试与「以算换存」策略

离群点的数量占比很少，可以复制一份，通过冗余的计算避免昂贵的数据重排操作




资料来源：[SC'24] MixQ: Taming dynamic outliers in mixed-precision quantization by online prediction. Yidong Chen等

# 奇妙的数学，RoMeo 对 MixQ 的扩展

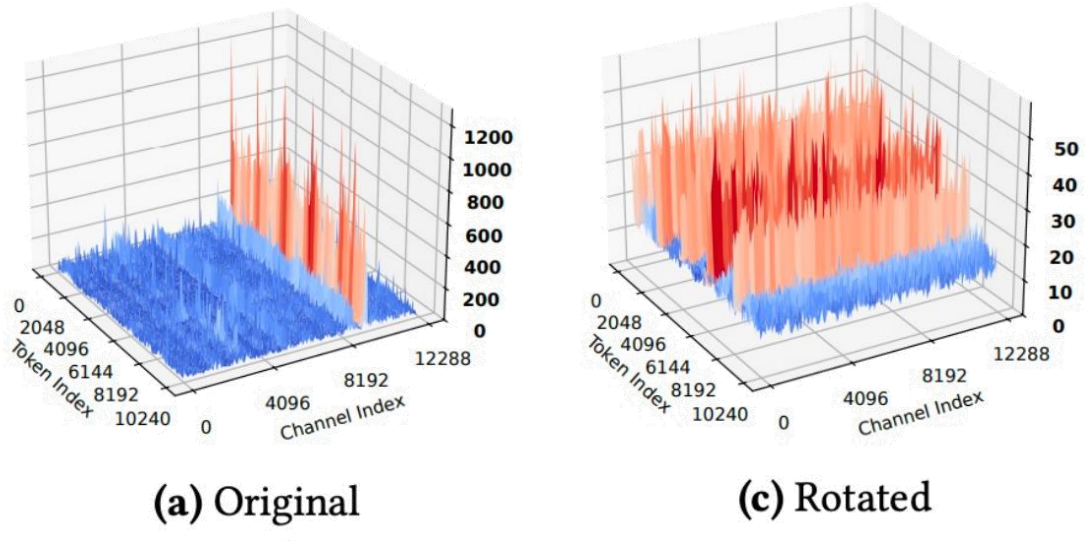
$$H_4 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$$

$$Q[A] \times Q[W]$$

$$\Downarrow$$

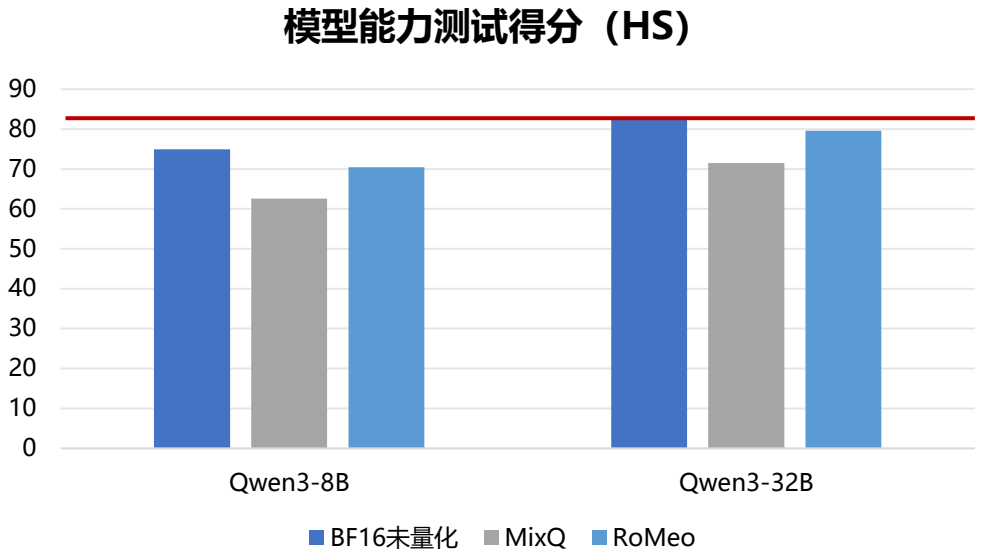
$$\underline{Q[AH]} \times \underline{Q[H^T W]} \quad \text{Offline}$$

Online FWT　Rotated & Quantized




(a) Original　　(c) Rotated

4/8比特混合量化后，32B模型的体积和运行开销与未量化的8B模型相近，但能力更强




资料来源：[PPoPP'26 Accepted] RoMeo: Mitigating Dual-dimensional Outliers with Rotated Token-wise Mixed Precision Quantization, Qihao Zhang等

# QAT/QAD 与原生低精度推理

## PTQ 训练后量化

最常见的量化方式

直接对已有模型的权重进行量化

## QAT 量化感知训练

在训练时，用量化方式前向计算

反向传播使用原始精度

## QAD 量化感知蒸馏

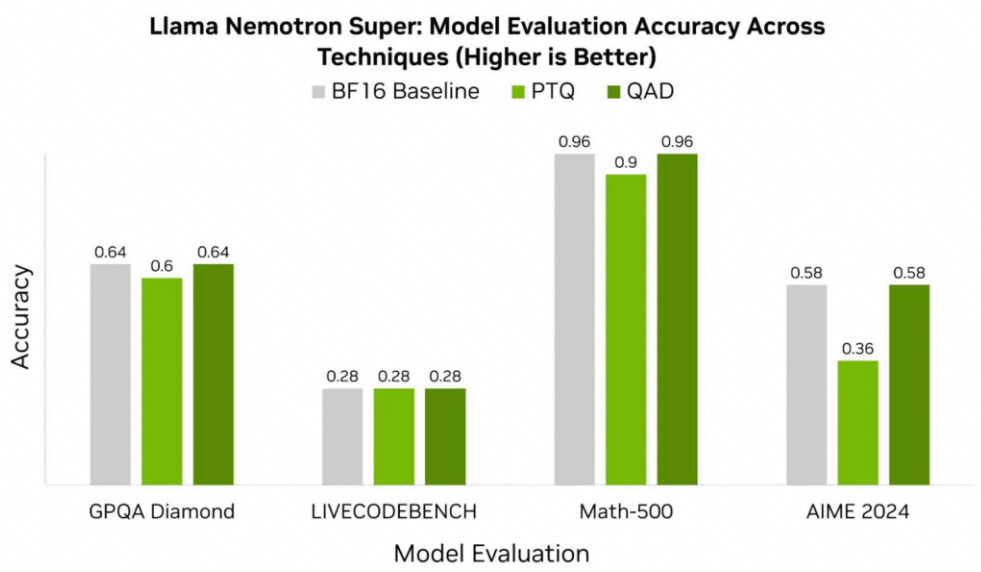
在蒸馏时一并完成量化

用原始精度模型作为教师模型

**Qwen3: BF16 Native**

**DeepSeek-V3: FP8 Native**

**Kimi-K2-Thinking: INT4 Native**




资料来源：NVIDIA网站、HuggingFace等

03

# 高速量化推理需要贴合硬件架构

# 各平台对量化推理的支持不尽相同

## 精度支持不同

是否支持FP8
是否支持FP4
是否支持INT8
是否支持INT4
是否支持MXFP8
是否支持MXFP4
是否支持MXINT8
……

## 算力分配不同

张量、向量、标量算力的比例
高精度、低精度算力的比例
张量和向量是否可以同时计算
张量和向量是否共享高速存储
……

# 量化算法的组合爆炸

**量化分组粒度**

以16/32/256个元素为一组

block or channel or tensor

**累加精度要求**

矩阵乘加运算中的累加

精度是INT32还是FP32


**缩放因子类型**

一层量化 or 两层量化

FP32 or BF16 or FP8

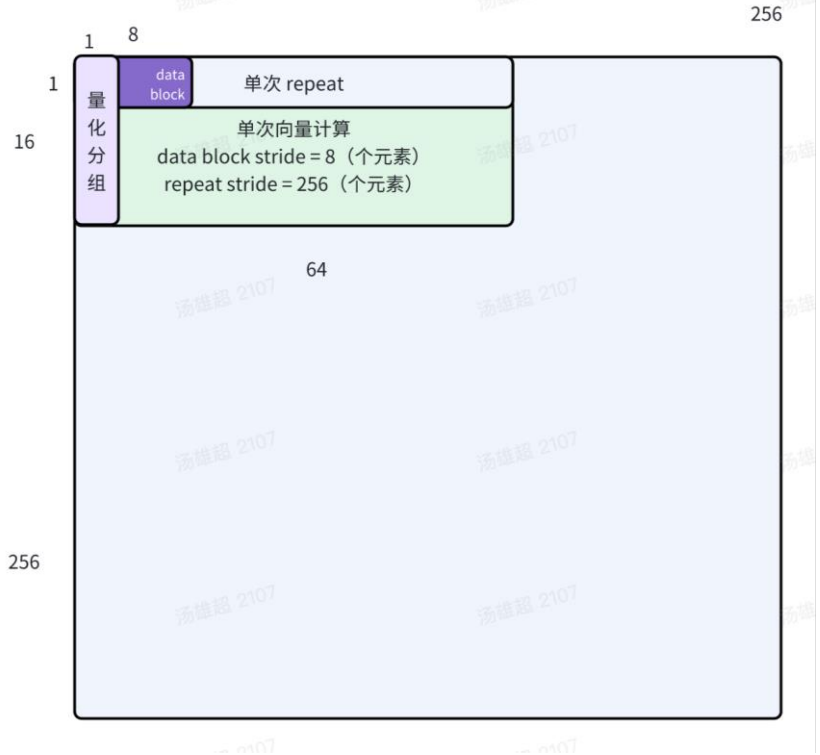
**是否有偏移项**

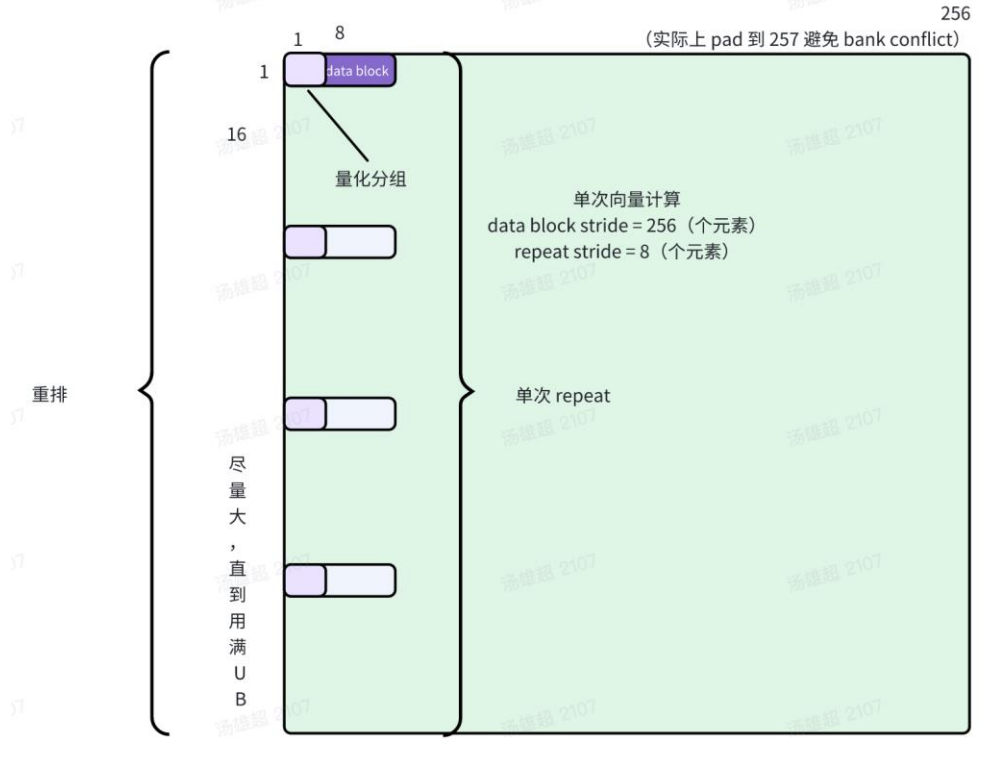
对称量化 or 非对称量化

矩阵乘加后是否有修正

# 面向国产算力重排张量内存布局

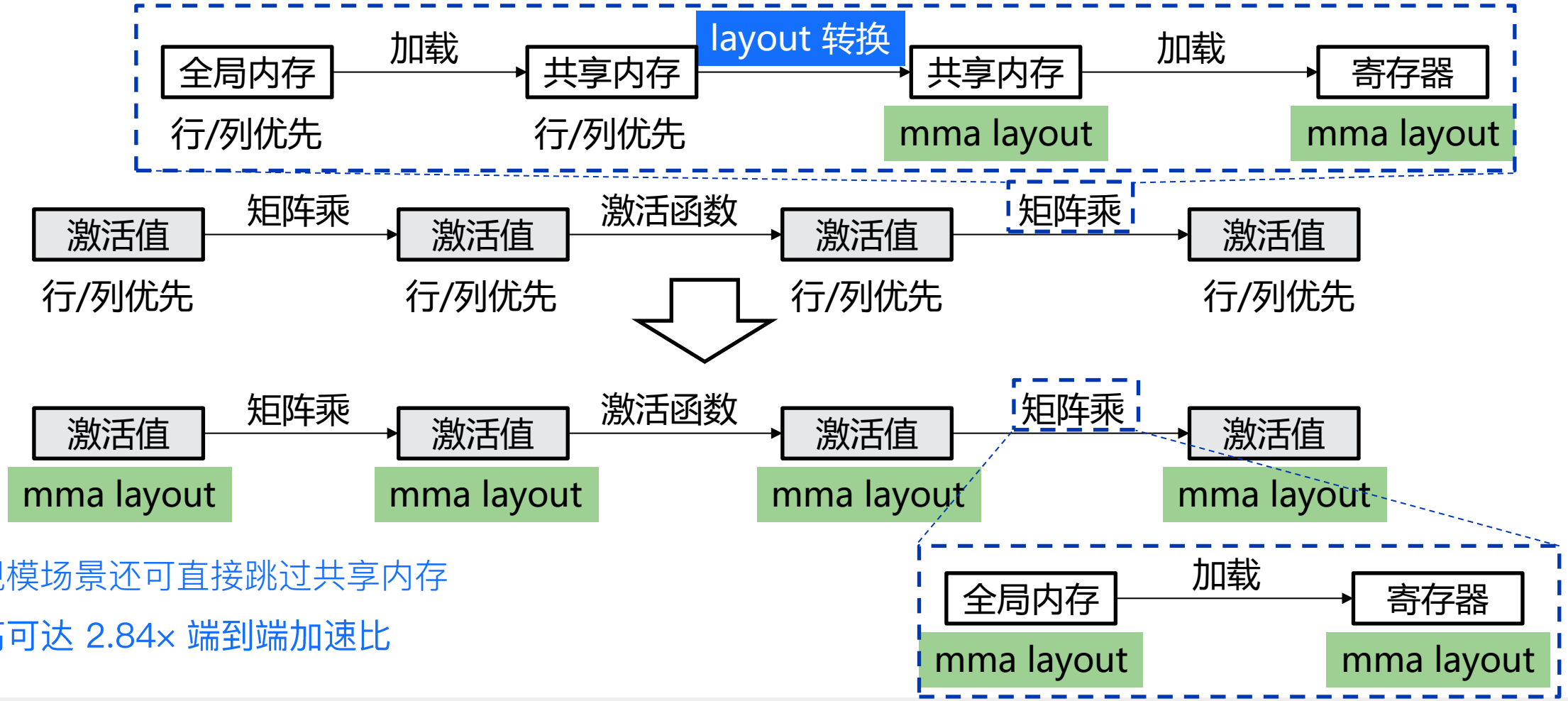
- 昇腾向量指令一次性处理 shape=(r, 8, 8)、stride=(x, y, 1) 的高维数据区域
- 通过预处理重排数据内存布局扩大每次处理的范围
- 推理引擎+算子协同优化，引擎全局重排数据，算子直接调用



# 推理引擎支持的全局内存重排优化

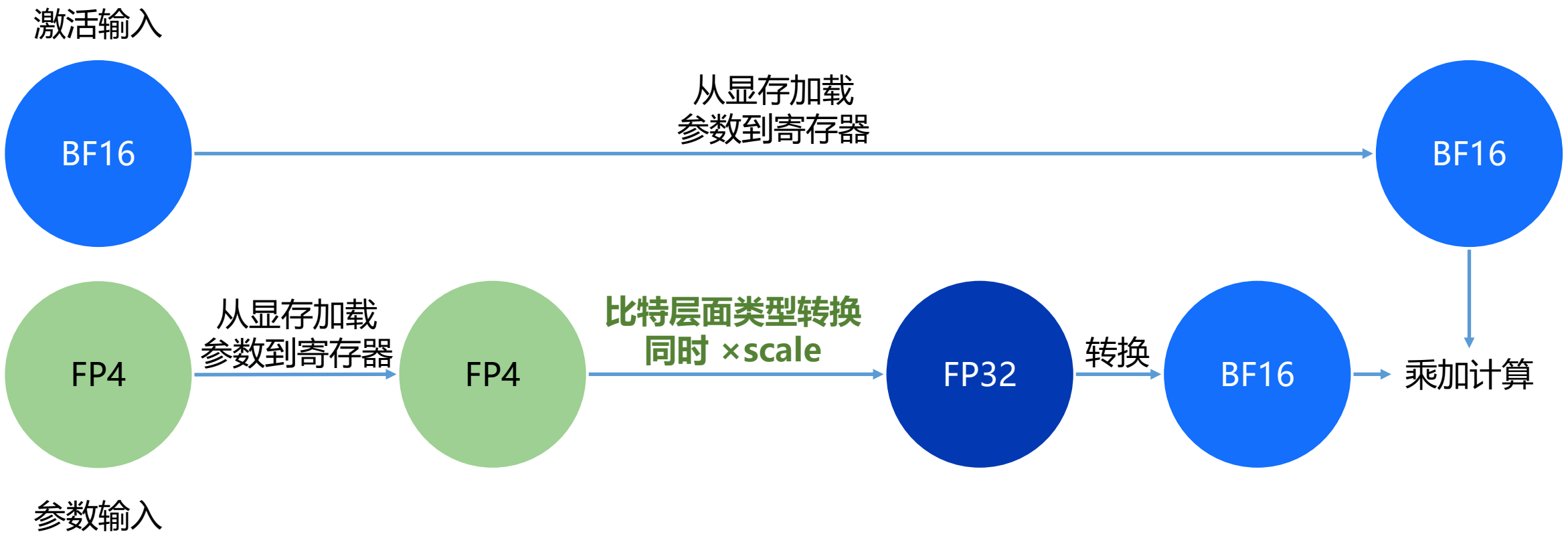
算子–推理引擎联合优化：全局存储 mma layout




小规模场景还可直接跳过共享内存

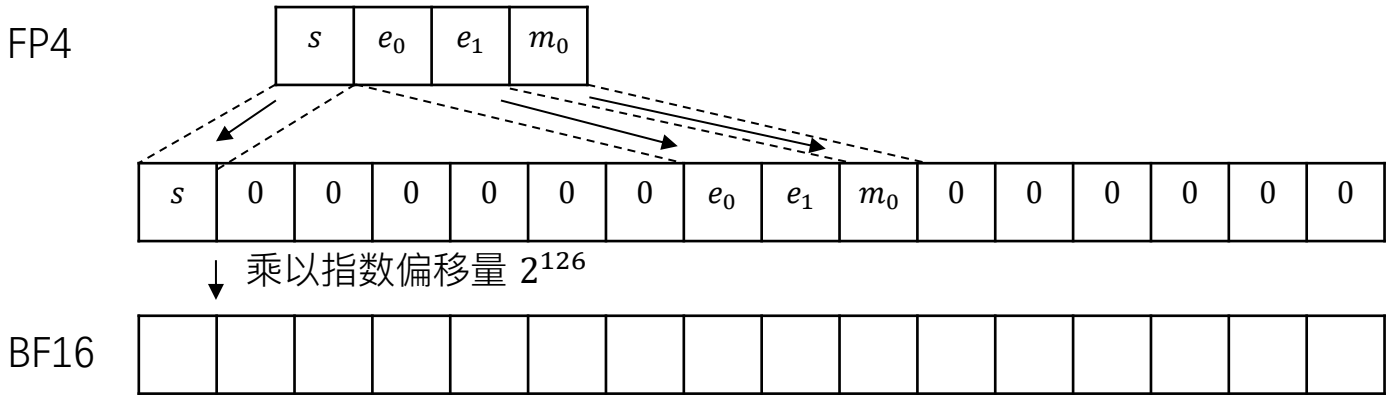
最高可达 2.84× 端到端加速比

# 旧算力设备如何支持新量化算法.



# 旧算力设备如何支持新量化算法..

## 软件实现 FP4 转 BF16 的位运算流程




**核心优势: 纯整数位运算，无需规格数判断**

利用浮点数乘法校正指数位偏差，在权重加载时瞬间完成转换，极大降低延迟。

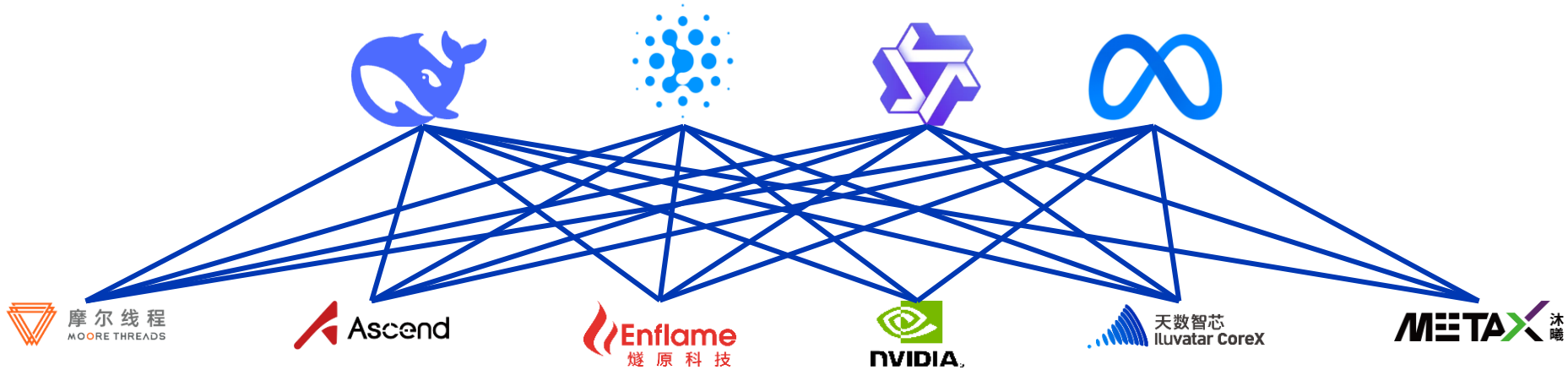
04

# 异构算力平台量化推理解决方案

# 多种模型和多种算力的 MxN 适配问题

m 种模型

n 种加速卡

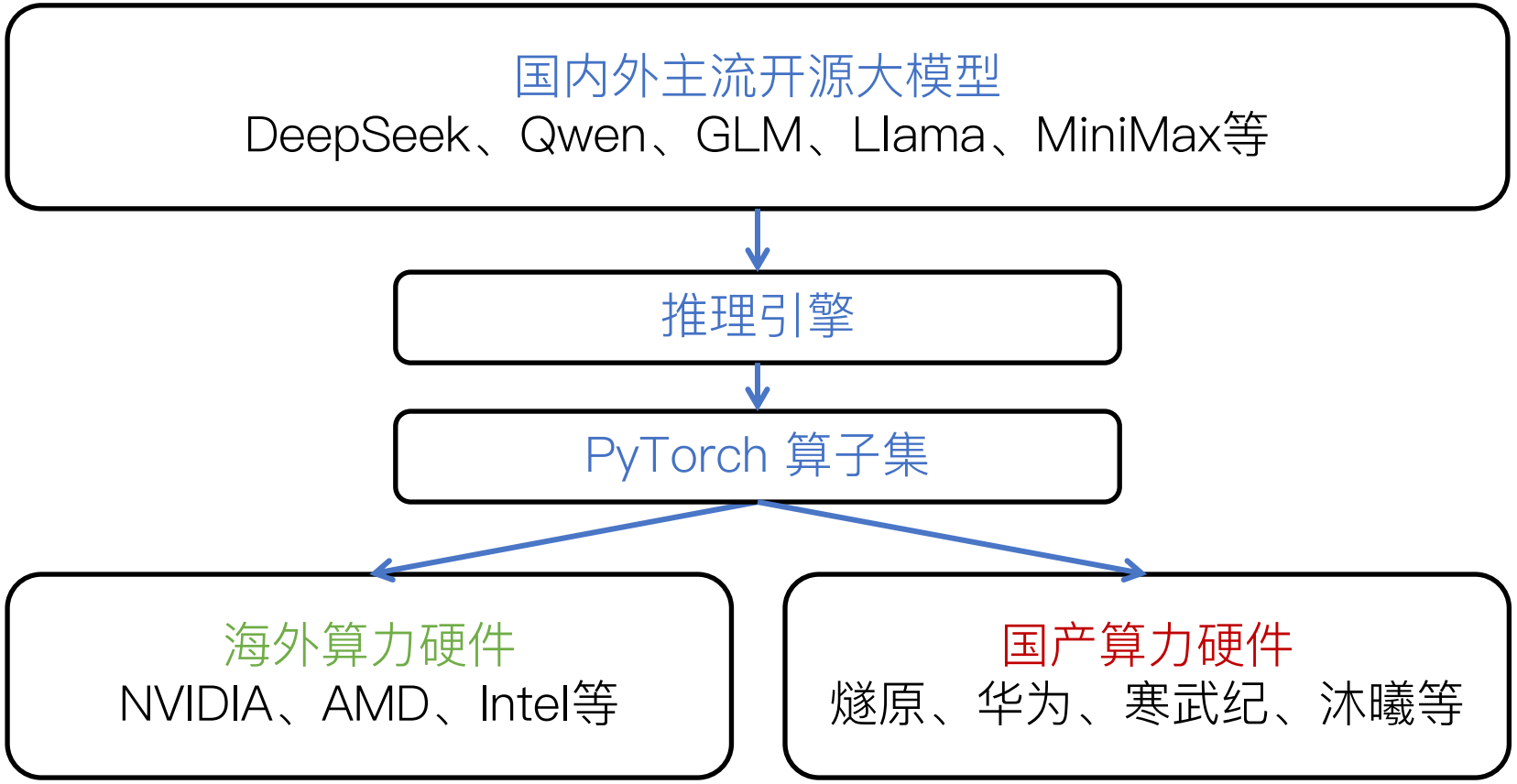



算法+推理引擎+算子 联合优化

➔ m 种模型面向 n 种加速卡的优化

➔ 在 n 种加速卡上优化 m 种模型需要 n × m 种实现，工作量组合爆炸?

# 站在巨人的肩膀上：PyTorch 算子集



# 超越算子集：推理引擎+算子 协同设计优化

## 联合优化已成为高性能推理所必须

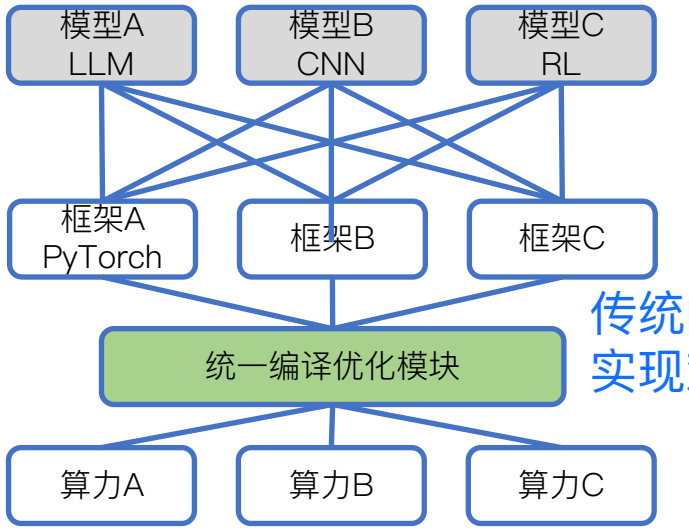
- Continuous batching
  - 推理引擎调度策略 + 变长 attention 算子
- Paged KV cache
  - 推理引擎 page 管理 + paged attention 算子
- 分组量化
  - 量化算法 + 分组 scale 低位宽矩阵乘算子
- 专家并行
  - 推理引擎并行策略 + AllToAllV 等通信算子

## 面向NV算力的联合优化方案<br>大部分无法在国产算力上使用

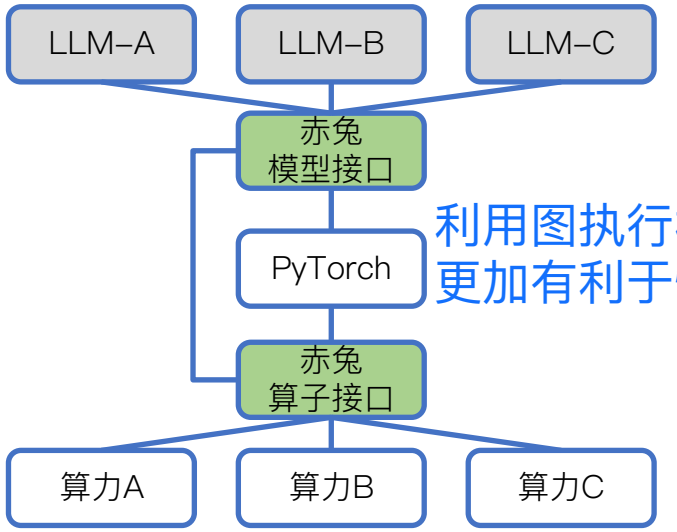
- 列优先 weight
  - 假设矩阵乘算子的特定设计
- 变长 attention 算子
  - 假设与序列长度无关的算子内分块策略
- 分组量化
  - 假设向量单元与张量单元的高效交互
- 新兴量化数据类型
  - 假设硬件对数据类型的支持

# 超越算子集：推理引擎+算子 协同设计优化

1. LLM 作为特定类型的神经网络，具有更多的共性
    1. 基于 pytorch 运行
    2. 通过记录 kernel 指针的 graph 机制加速
2. 推理引擎重点与 Python 变量、函数、类的交互，无需成为完整 op set 上的编译器



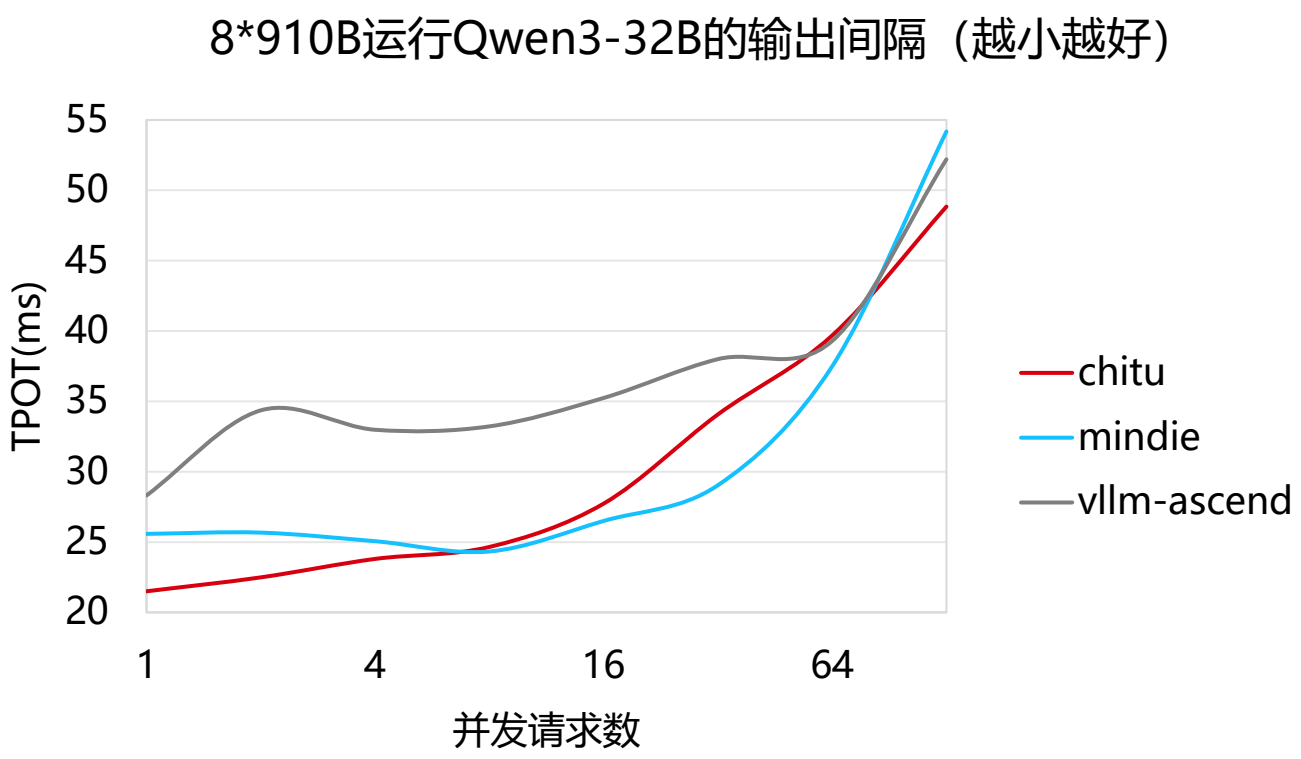

传统的图编译优化方式
实现难度高，工程量大



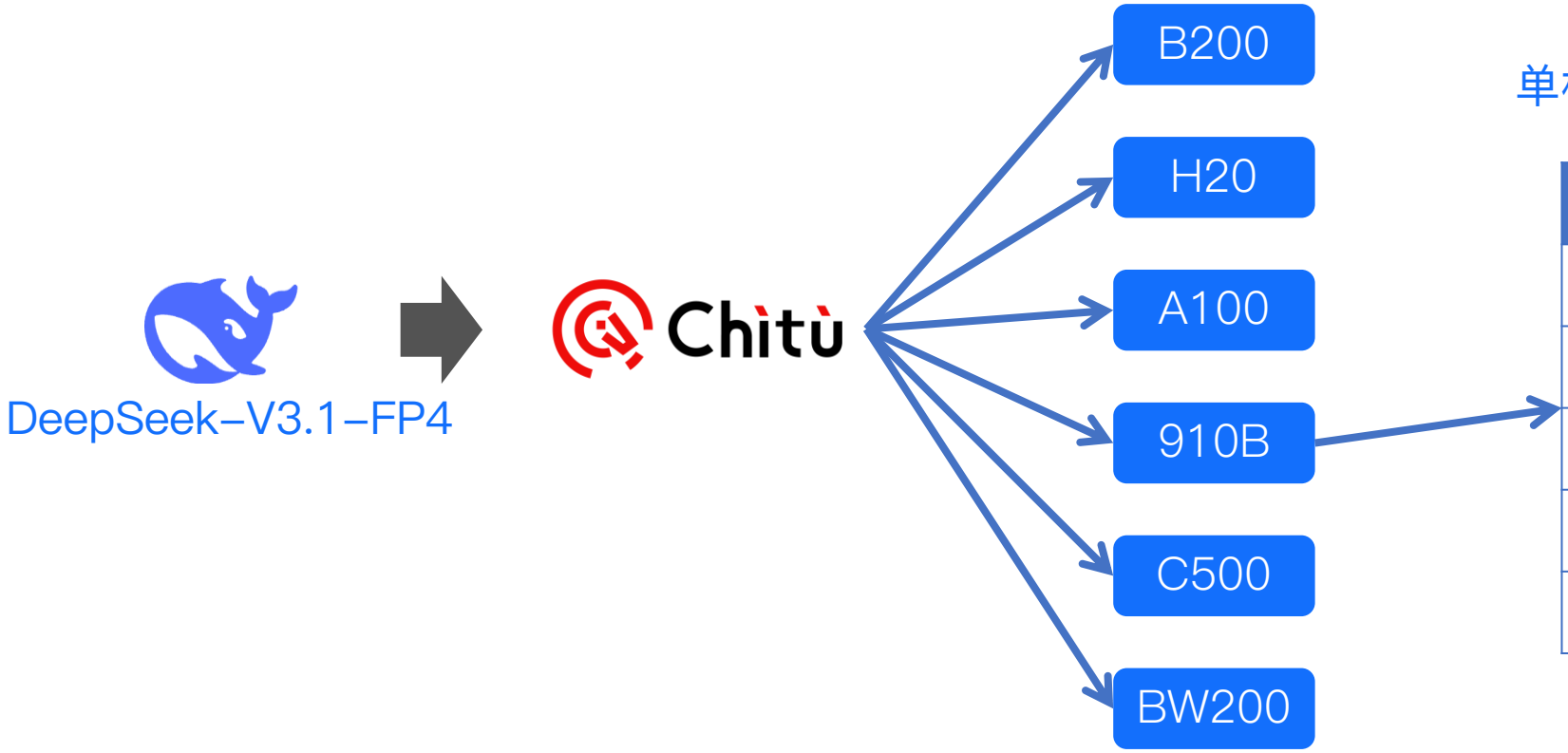

利用图执行机制和语言特性，
更加有利于快速实现

# 案例：面向昇腾算力的高效实现

- 在推理引擎的框架层面实现张量内存重排
- 重排后的张量有助于实现更高性能的Attention算子
- 若无框架层面协同优化，需要在每个算子内部重排，得不偿失



# 案例：多种算力对 FP4 量化推理的统一支持




单机八卡910B运行DS满血版性能

| Batch Size | Token/s |
|---|---|
| 1 | 17.76 |
| 2 | 26.95 |
| 4 | 55.97 |
| 8 | 87.00 |
| 16 | 114.40 |

# Take Away Points

### 低精度算力

未来的绝大部分AI算力都是低精度算力

### FP4成为主流

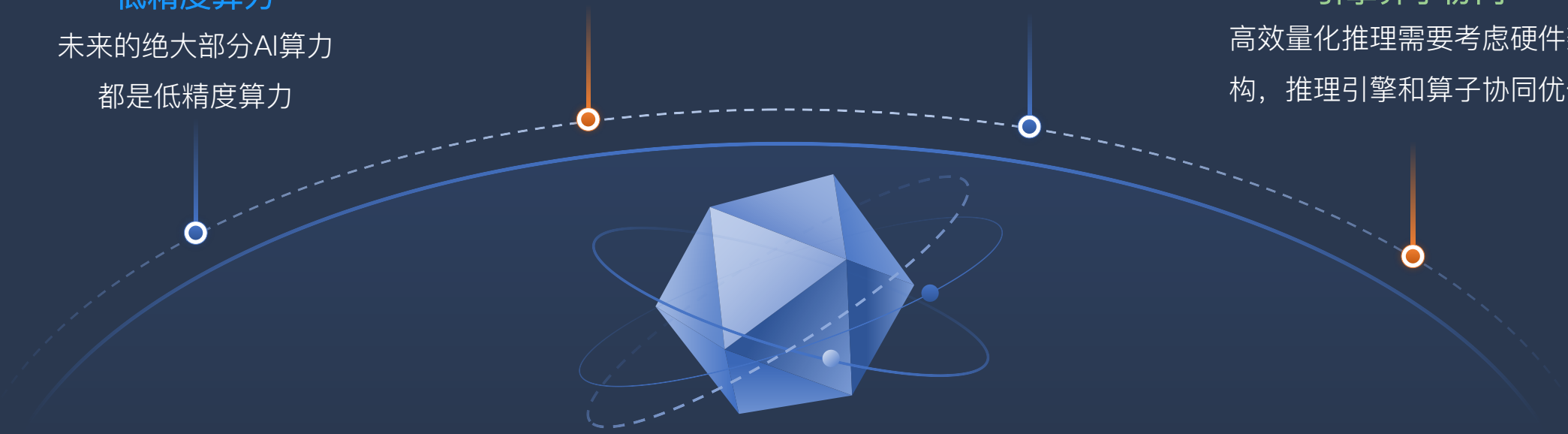
低精度浮点数和低精度整数各有利弊，目前FP4正在成为主流算力

### 控制量化损失

可以在利用低精度算力加速推理的同时，保持模型能力几乎不变

### 引擎算子协同

高效量化推理需要考虑硬件架构，推理引擎和算子协同优化



# 极客邦科技 2026 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

## 4月 · 北京 · 1200人

QCon 全球软件开发大会

会议时间：4月16-18日

- Agentic Engineering
- AgentOps
- 下一代模型架构与推理优化
- AI 原生基础设施
- 知识工程实践
- AI 安全

## 6月 · 上海 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：6月26-27日

- AI Infra 系统工程
- 多 Agent 协作与实践
- 多模态融合
- 模型训练与推理创新
- 数据平台与特征服务

## 8月 · 深圳 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：8月21-22日

- Agentic AI
- 轻量化与高效推理
- 多模态应用
- AI + IoT 场景实践
- AI 工业化落地

## 10月 · 上海 · 1200人

QCon 全球软件开发大会

会议时间：10月22-24日

- AI Agent
- Vibe Coding
- 智能可观测
- 推理基建
- 模型攻防
- AI x 创造力

## 12月 · 北京 · 1000人

AiCon 全球人工智能开发与应用大会

会议时间：12月18-19日

- 大模型架构创新
- 多模态 AI 产业融合
- 具身智能
- AI for Science
- 大模型安全

InfoQ 极客传媒

InfoQ 极客传媒

# THANKS

## 探索 AI 应用边界

Explore the limits of AI applications

AiCon

全球人工智能开发与应用大会