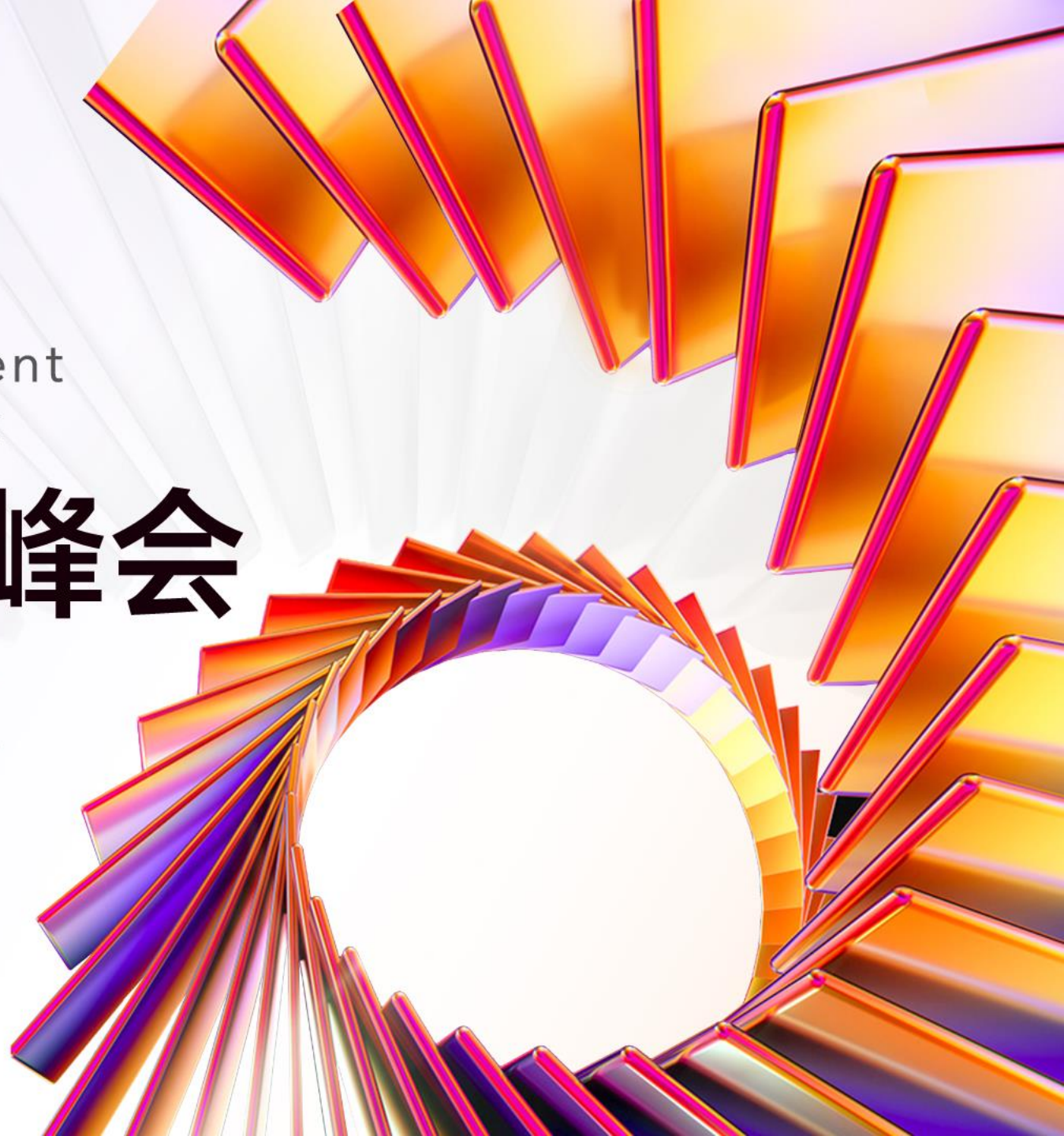
AiDD 8th 2025

第8届 AI+ Development Digital Summit AI+研发数字峰会

拥抱AI 重塑研发

11月14-15日 | 深圳



AiDD × K

2026 AI+ PRODUCT INNOVATION SUMMIT

01.16-17 · ShangHai

AI+产品创新峰会

Track 1: AI 产品战略与创新设计

从0到1的AI原生产品构建

论坛1: AI时代的用户洞家与需求发现

论坛2: AI原生产品战路与商业模式重构

论坛3: AgenticAI产品创新与交互设计

Track 2: AI 产品开发与工程实践

从1到10的工程化落地实践

论坛1: 面向Agent智能体的产品开发

论坛2: 具身智能与AI硬件产品

论坛3: AI产品出海与本地化开发

Track 3: AI 产品运营与智能演化

从10到100的AI产品运营

论坛1: AI赋能产品运营与增长黑客

论坛2: AI产品的数据飞轮与智能演化

论坛3: 行业爆款AI产品案例拆解

2-hour Speech: 回归本质

用户洞察的第一性

——2小时思维与方法论工作坊

在数字爆炸、AI迅速发展的时代，仍然考验“看见”的“同理心”

Panel 1: 出海前瞻

“出海避坑地图”圆桌对话

——不止于翻译:

AI时代的出海新范式

Panel 2: 失败复盘

为什么很多AI产品“叫好不叫座”?

——从伪需求到真价值:

AI产品商业化落地的关键挑战

智能重构产品 数据驱动增长

Reinventing Products with Intelligence, Driven by Data

80+ 业界大咖 汇聚产品大咖的真知灼见

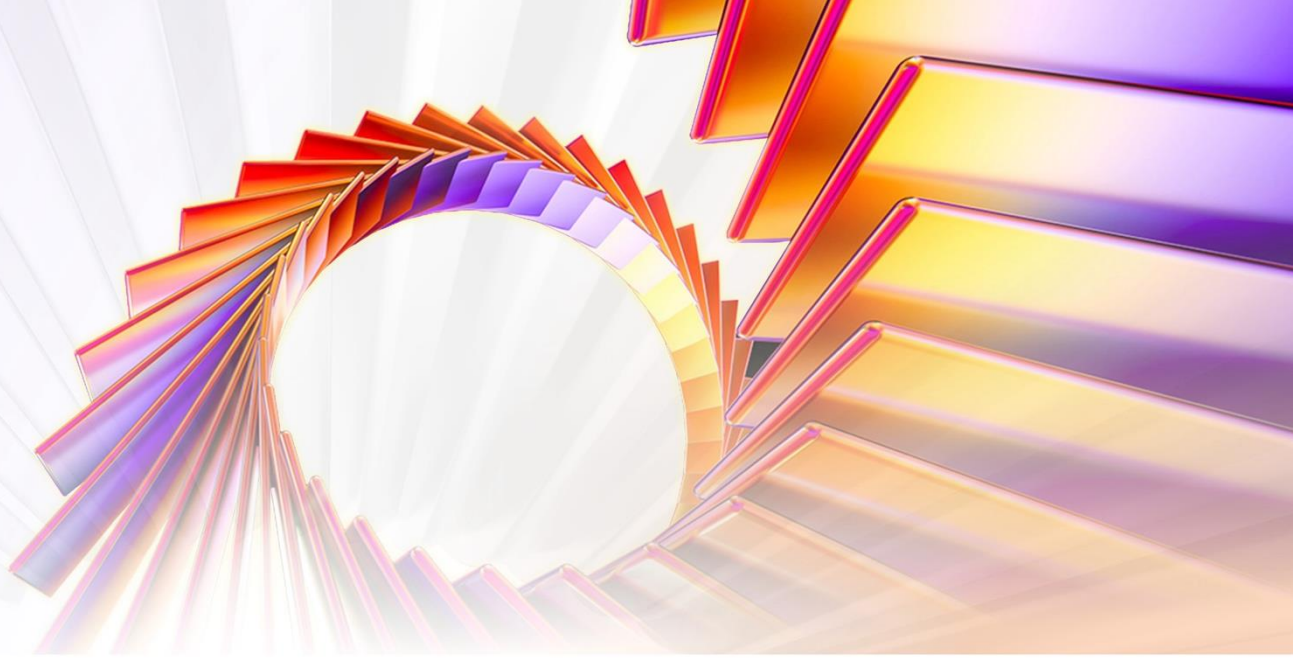
40+ 创新案例 开拓全新AI+产品视角

10+ 分论坛 涵盖产品到商业增长的全链路

800+ 听众规模 共同探索产品的智能重构

扫码查看会议详情

AiDD 8th 2025 | 11月14-15日 | 深圳
第8届 AI+ Development Digital Summit
AI+研发数字峰会
拥抱AI 重塑研发



大模型训练和推理中的前沿优化技术

张闰清 | 清华大学

张闰清 博士

清华大学高性能所

清华大学计算机系高性能所博士生，导师为翟季冬教授。研究领域为大模型推理系统优化。本科期间曾获得世界大学生超算竞赛ISC24现场赛总冠军。

目录
CONTENTS

PART 01

并行训练与推理的背景

为什么需要并行训练与推理-显存压力

- 并行推理：使用**多个GPU**同时进行训练/推理
- 大语言模型规模增长迅速：
  - 65B -> 671B
- 显存内需要保留：
  - 参数、Activation、Optimizer（训练）、KVCache（推理）
  - 上下文增加 -> KVCache空间变大
- GPU显存容量有限
  - 32GB, 80GB, 96GB

模型名称	发布时间	参数量
Llama	2023 年 2 月	65B
Llama2	2023 年 7 月	70B
Llama3.1	2024 年 7 月	405B
Deepseek-R1	2024 年 12 月	671B

GPU	显存容量	可容纳参数(fp8)
5090	32GB	32B
H800	80GB	80B
H20	96GB	96B

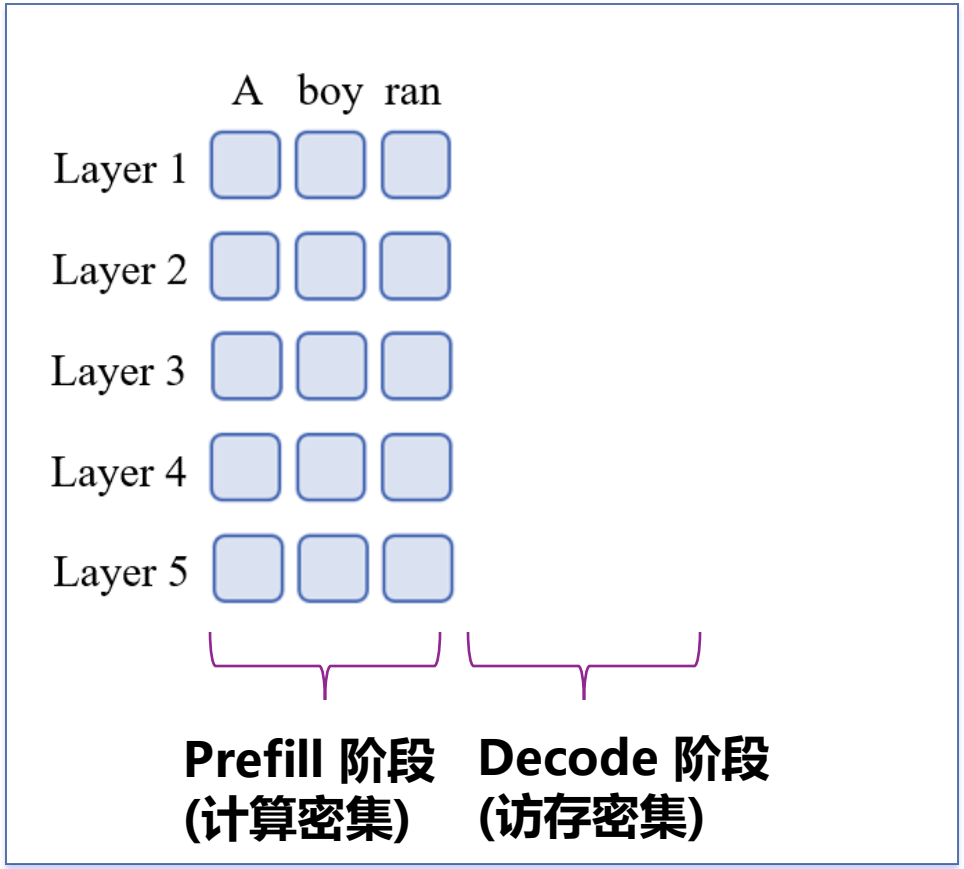
为什么需要并行推理-SLO/高并发

- 真实场景有成千上万用户并发请求。
- 用户期望对话/搜索系统在百毫秒级响应。

- 单个/少量GPU的算力/显存带宽无法满足要求

大语言模型推理过程

- 自回归大模型：
  - Prefill 阶段：
    - 预处理请求中的所有 token
    - 为每个 token 生成相应的 KV Cache
    - 计算密集型
  - Decode 阶段：
    - 基于新生成的 token 和已有 KV cache 计算下一个 token
    - 访存密集型



PART 02

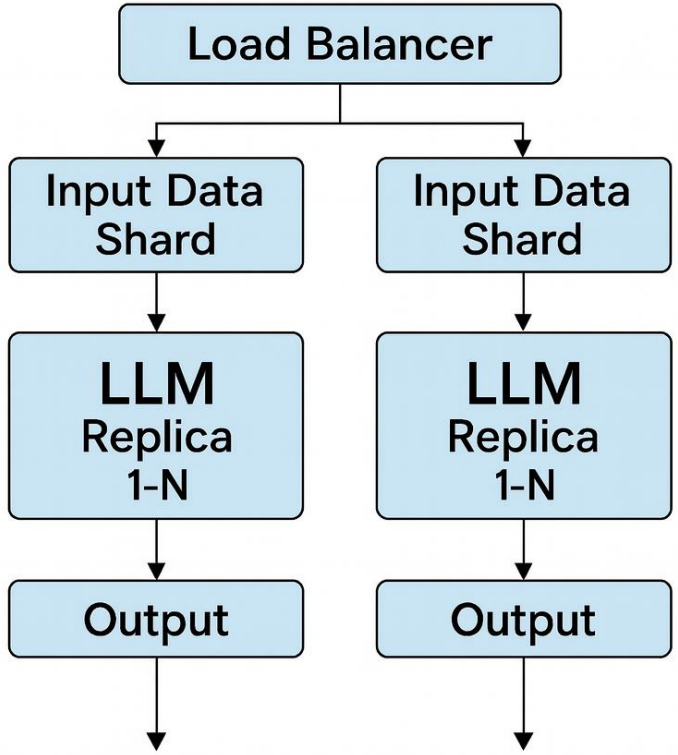
并行方法简介

基本并行方法


数据并行
Data Parallel/DP
张量并行
Tensor Parallel/TP
流水线并行
Pipeline Parallel/PP
专家并行
Expert Parallel/EP

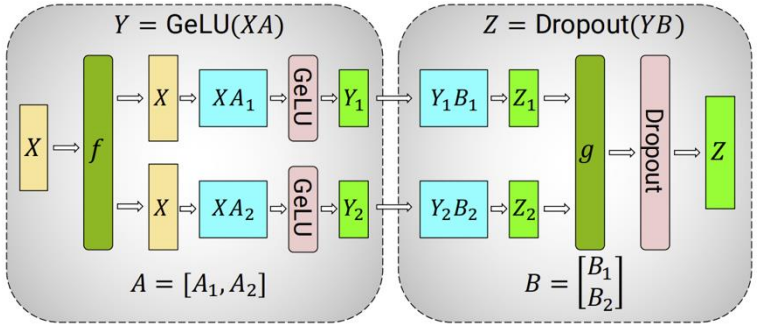
数据并行

- 数据并行会在不同的GPU上复制整个模型
  - 数据并行无法缓解显存压力
- 开销很小
  - 只有输入数据分发需要通信
- 多种实现范式
  - 多个推理引擎实例+HTTP负载均衡
  - 单个推理引擎实例内数据并行

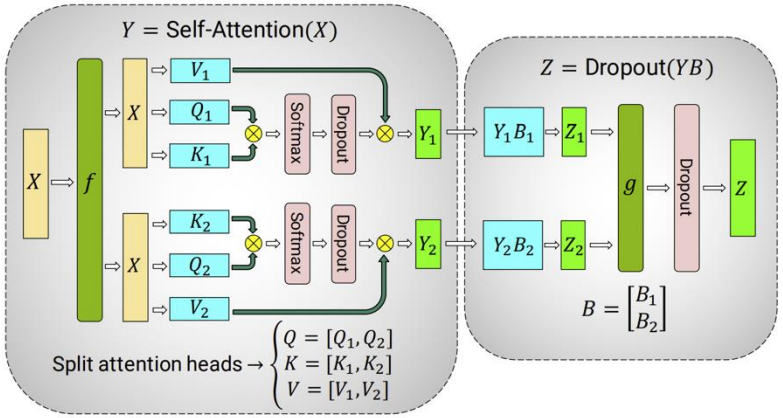


张量并行

- 将模型中的张量（矩阵）在某些维度上切开
  - 切开的结果分布到不同的GPU上
  - 可以缓解显存压力、计算压力
- 通信开销较大
  - 引入Allreduce将各个部分的结果合并
  - 通过规划切分维度减少通信次数
- 扩展性较差
  - 通信量较大
  - 一般不进行跨机张量并行



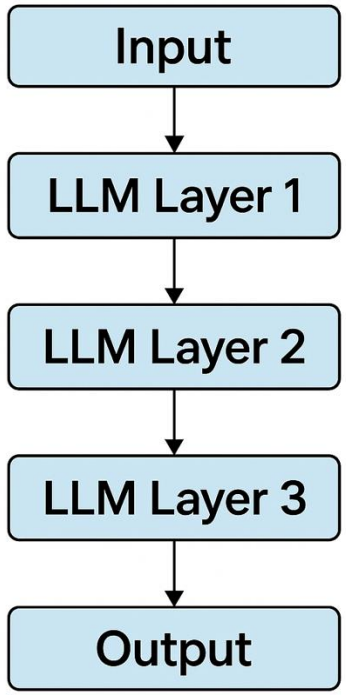
(a) MLP.

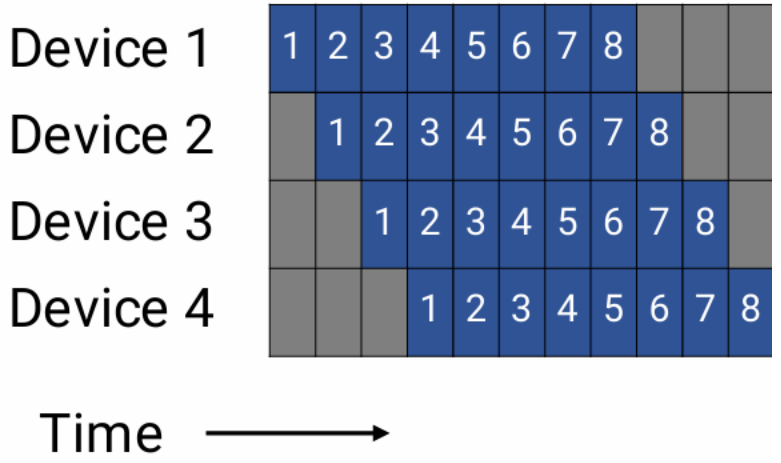


(b) Self-Attention.

流水线并行

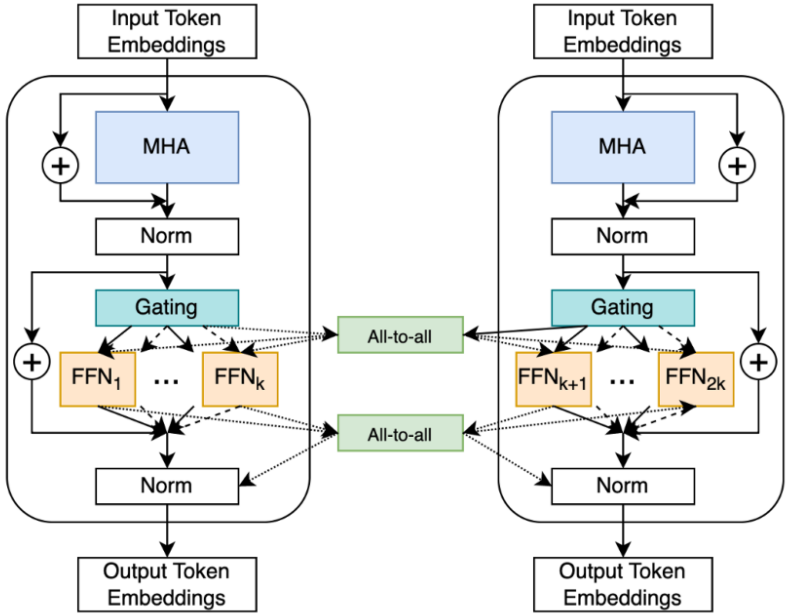
- 流水线并行将模型在Layer维度上进行切分
  - 切开的结果分布到不同的GPU上
  - 可以缓解显存压力、计算压力
  - 在运行时，依次在GPU上进行运算
- 通信开销较小
  - 只需要在切换流水线阶段时点对点发送Activation
- 问题
  - 流水线气泡
  - 延迟较高

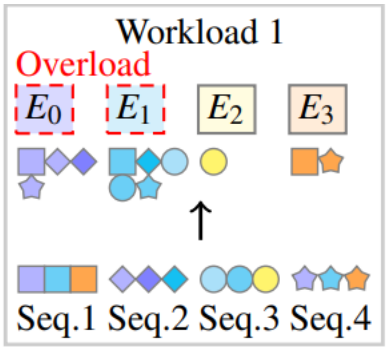




专家并行

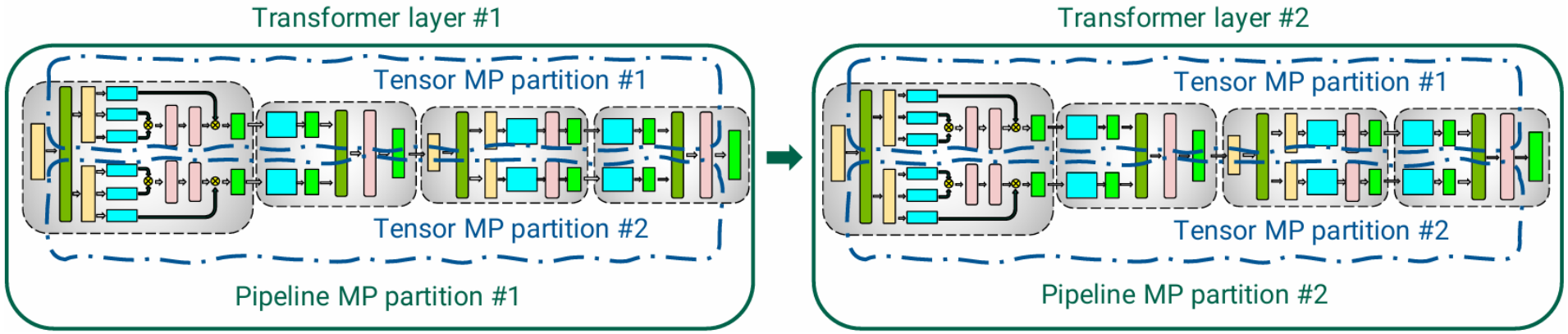
- 在MoE模型中，可以将模型在专家维度上分割
  - 不同GPU存储不同专家
  - 缓解显存压力、计算压力
- 引入All-to-All通信
  - 通信模式复杂
- 引入负载不均问题
  - 需要进行负载均衡
- 扩展性较好





混合并行

- 适应不同的GPU数量、网络拓扑
- 张量并行、流水线并行混合

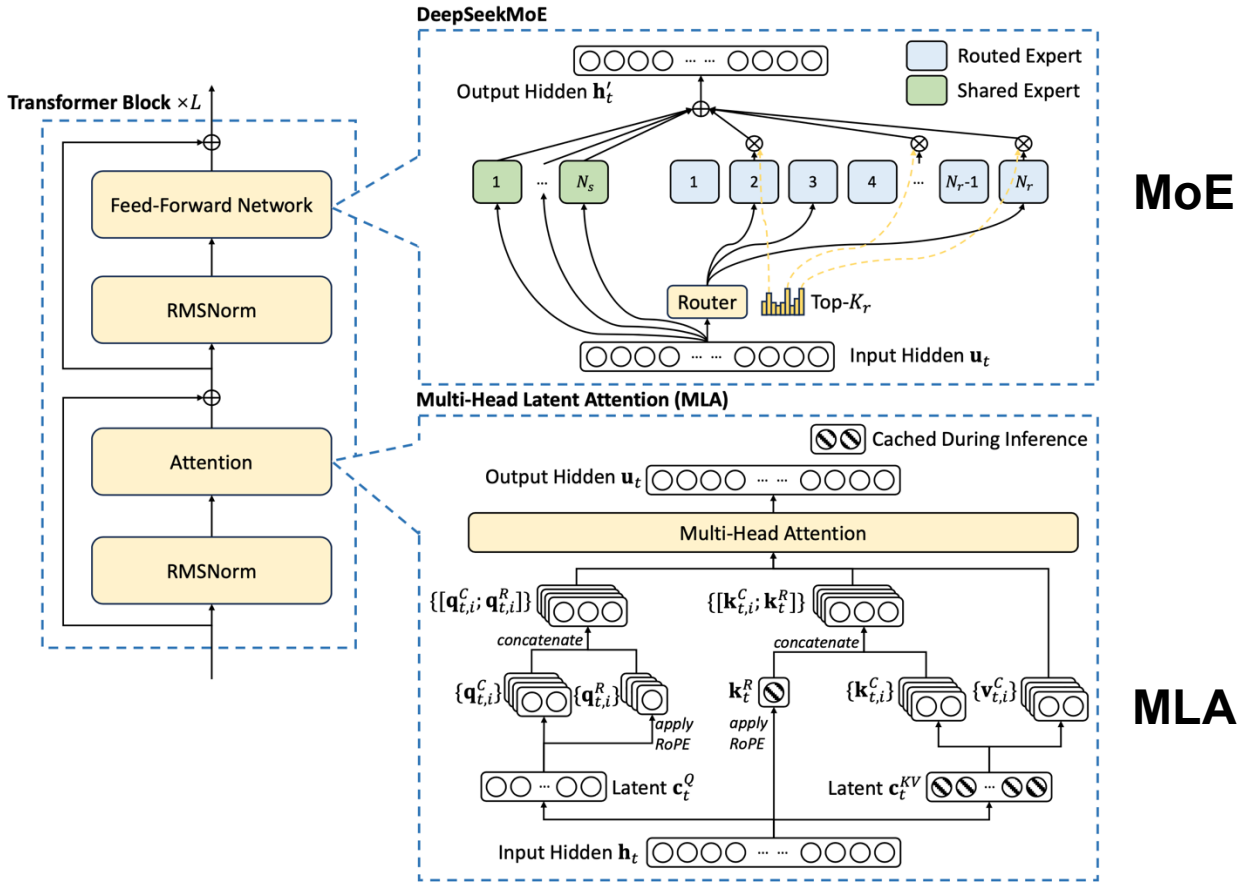


PART 03

Case Study DeepSeek-V3

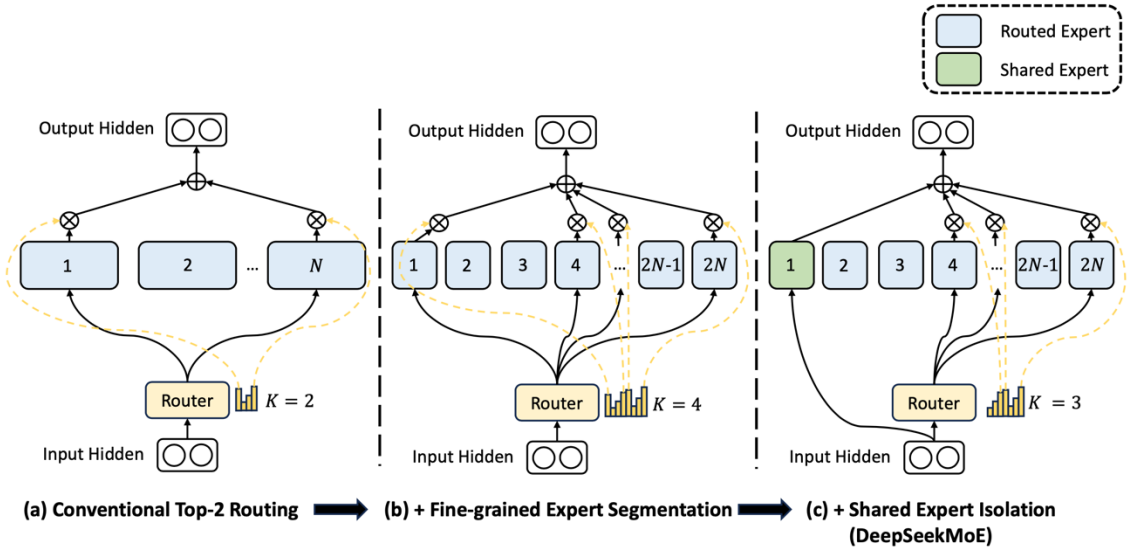
Case Study – DeepSeekV3

- DeepSeek模型架构：MoE+MLA

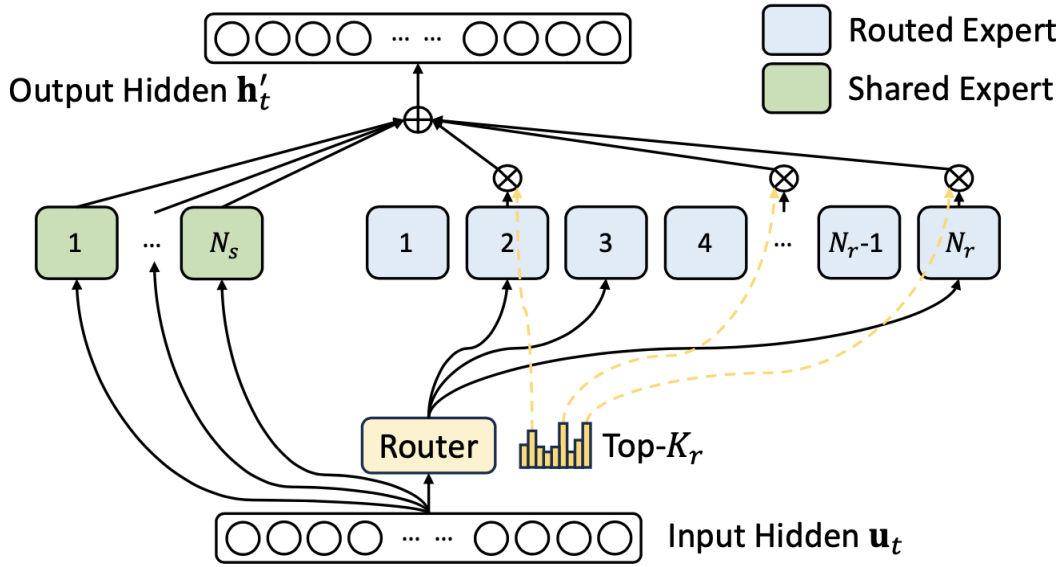


DeepSeekV3 - MoE

- 核心思想：共享专家+大量细粒度路由专家
- 共享专家： 捕获通用知识、降低知识冗余
- 路由专家：量大、细粒度、灵活组合、方便知识表达
- V3：1共享专家+256路由专家、每token激活8个路由专家



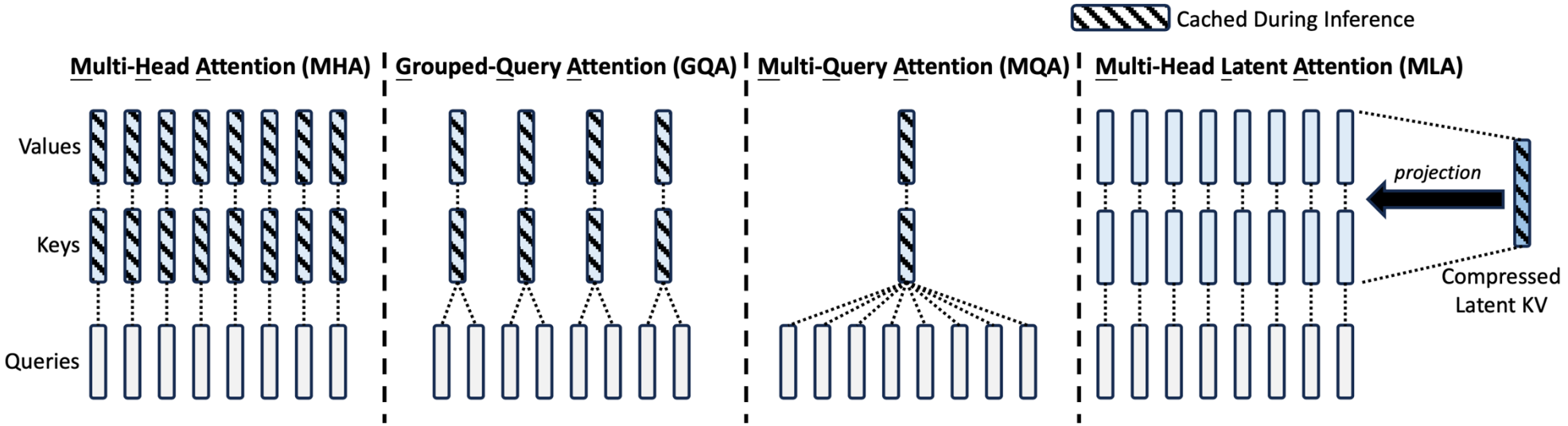
不同MoE模型架构



共享专家+路由专家MoE架构

DeekSeekV3 - MLA

- 为了降低大模型推理成本，提出 MLA 架构
- 核心思想：通过低秩压缩 KV，显著降低推理时 KV cache 的存储空间需求
- MLA 存储需求降低，同时可以更好地保持模型精度
- MLA的KVCache无法通过张量并行分割



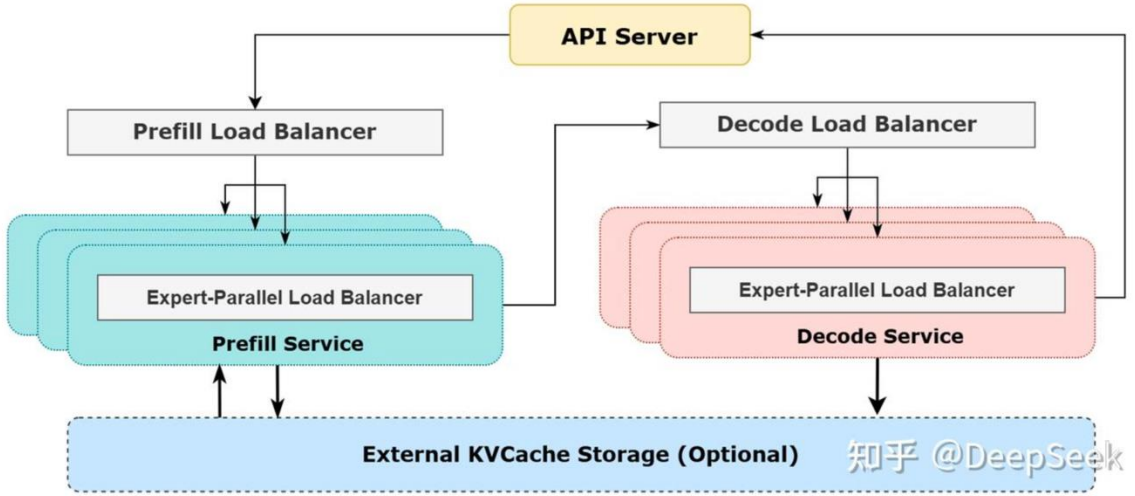
MQA：PaLM、Gemini 等模型采用
GQA：LLAMA3、ChatGLM3、DeepSeek V1等模型采用

DeepSeekV3的并行推理部署

- DeepSeekV3的参数量为671B
  - 使用fp8推理时，至少需要671GB显存
    - 16张80G显存GPU/8张96G显存GPU
- 可行的并行方案
  - 单机内做完全张量并行（TP），机间做流水线并行（PP)
    - TP跨机扩展性不好
  - MoE部分做专家并行（EP），Attention部分做数据并行（DP）
    - 通过DP Attention减少KVCache冗余
    - 目前主流的部署方法
    - 可适用于大规模并行推理

DeepSeekV3 论文中的实践

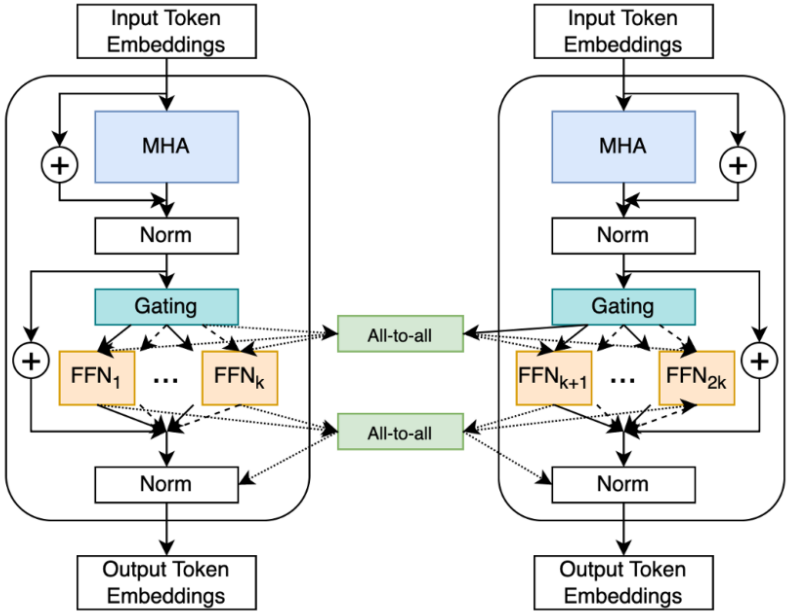
- DeepSeek 推理配置
- PD 分离策略
  - Prefill 阶段以 32 GPU 为1个部署单元
  - Decode 阶段以 144 GPU 为1个部署单元
- Prefill 阶段并行策略
  - Attention 部分使用 4路 张量并行，8路 数据并行
  - MoE 部分使用 32路 专家并行
- Decode 阶段并行策略
  - Attention部分使用 4路 张量并行，36路 数据并行
  - MoE部分使用 144路 专家并行
  - 每个GPU计算2个专家，并有冗余的GPU计算热门专家



DeepSeek推理架构图

DeepEP

- DeepEP 是专为MoE架构与专家并行而设计的通信库。
- 针对大规模分布式训练与推理场景中，专家（Expert）之间的 **all-to-all** 通信成为瓶颈，DeepEP 提供优化的解决方案。
- 关键特性
  - 高吞吐率 & 低延迟的 GPU all-to-all 内核（Combine&Dispatch）
  - 支持低精度运算，包括 FP8 dispatching + BF16 combining。
  - 针对异构域（如 NVLink 域 → RDMA 域）通信做了专门优化。
  - 提供低延迟专用内核（使用 RDMA 路径+NvLink）适合推理解码环节。
  - 支持流 / 计算重叠机制（hook based overlap）以减小 SM 占用。



DeepGEMM

- DeepGEMM 是一个专为通用矩阵乘法（GEMM: General Matrix Multiplication）设计的高性能库。
- 支持 FP8 精度（即 8-位浮点）以及正在进行中的 BF16 支持，用于常规 GEMM 和专家混合 (Mixture-of-Experts, MoE) 分组场景。
- 用 CUDA 实现，无需内核编译安装时编译，使用轻量级 JIT 模块在运行时生成内核。
- 关键特性
  - 支持 FP8 精度矩阵乘法，细粒度缩放 (fine-grained scaling) 以提升运算效率。
  - 支持 MoE 分组 GEMM（即专家并行/混合专家系统中的 grouped GEMM 场景）。
  - 纯 CUDA 实现，运行时 JIT 内核，无需用户手动编译。
  - 面向推理场景，当前主要用于 DGRAD、Inference。

EPLB-Expert Parallelism Load Balancer

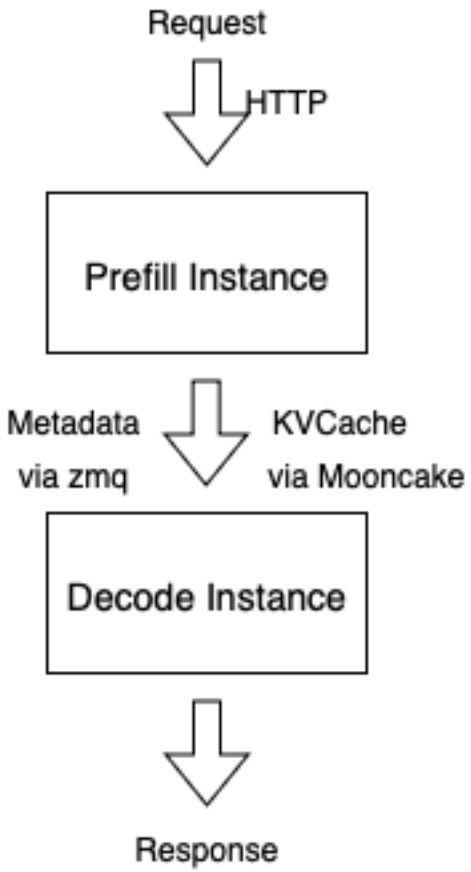
- EPLB 全称 “Expert Parallelism Load Balancer”，由 DeepSeek-AI 开源。
- 在专家并行 架构中，不同专家分布于不同 GPU / 节点，但各专家的实际负载常常不均衡，导致部分 GPU 空闲、部分 GPU 过载，从而拉低整体利用率。
- EPLB 的目标是：基于估算的每个专家的负载，计算一个复制与放置专家的方案，以在 GPU / 节点间实现更均衡的负载分配。
- 关键特性
  - 支持 **层级负载均衡策略（Hierarchical Load Balancing）**：当服务器节点数能整除专家组数时，先将专家组均匀分配给节点，再在节点内复制专家、最后分配至 GPU。
  - 支持 **全局负载均衡策略（Global Load Balancing）**：在节点数与专家组数不整除时，进行全局规律复制 / 放置专家至各 GPU。
  - 提供简单接口（例如 rebalance_experts）用于计算专家的复制与 GPU 映射方案。

PART 04

大模型训练和推理中的前沿优化技术

PD分离

- PD分离：Prefill、Decode分离
  - 在空间上分离：Prefill、Decode运行在不同的GPU上
- Why?
  - Prefill与Decode互相干扰
  - GPU资源与并行策略耦合
  - Continuous Batching、Chunked Prefill存在不足
- 在主流推理引擎中都得到了支持



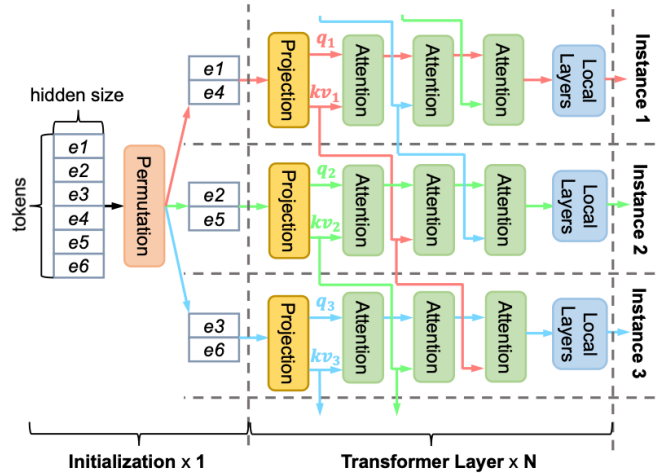
PD分离

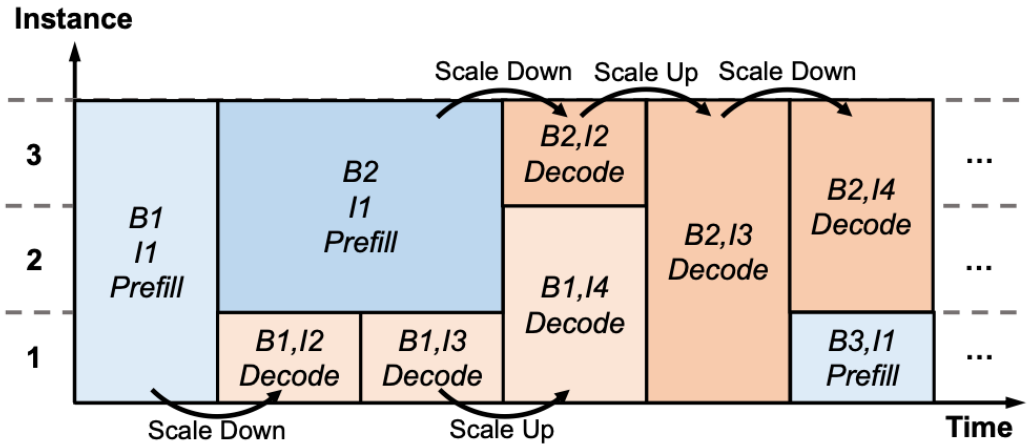
- Pros
  - 消除Prefill、Decode干扰
  - P、D并行方案解耦
  - 提高在满足SLO要求下的吞吐

- Cons
  - KVCache通信开销
  - 推理系统复杂性提高
  - GPU消耗量更大，不适用于小规模集群

LoongServe

- 背景
  - 在**超长上下文**(~1M)情境下，单一GPU显存无法存下一个Request的，需要引入序列并行（Sequence Parallel，SP）
  - 上下文长度是一个变动幅度、变动频率都很高的特性
  - 传统的解决方案（如PD分离、静态并行）无法适应
- 解决方案
  - 弹性序列并行
  - 统一KVCache存储池
  - 全局调度器+调度算法





LoongServe

- Scale up、Scale down的Overhead
  - 会涉及到KVCache存储位置改变
  - 通过改进算法规避Overhead

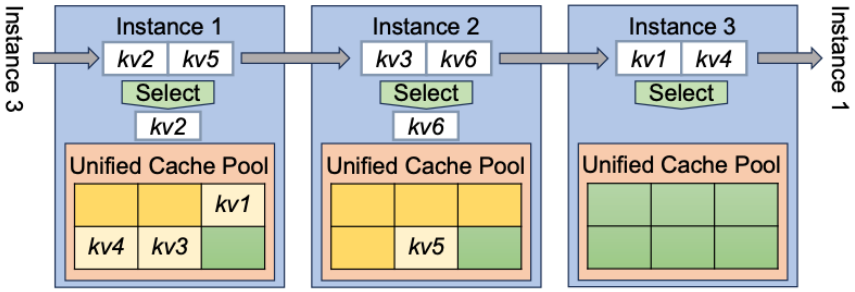


Figure 7. Elastic scale down in the prefill phase.

- Scale down
  - Proactive migration
  - 提前只在scale down后保留的节点存储KV
- Scale up
  - Multi-master distributed decoding
  - 在并行组内的多个GPU都可以作为master进行decode

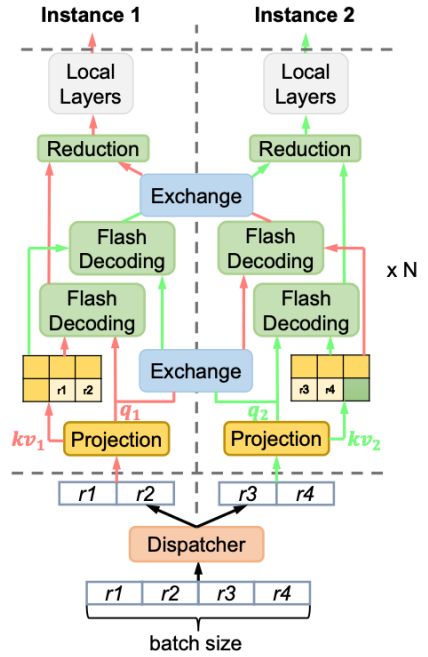
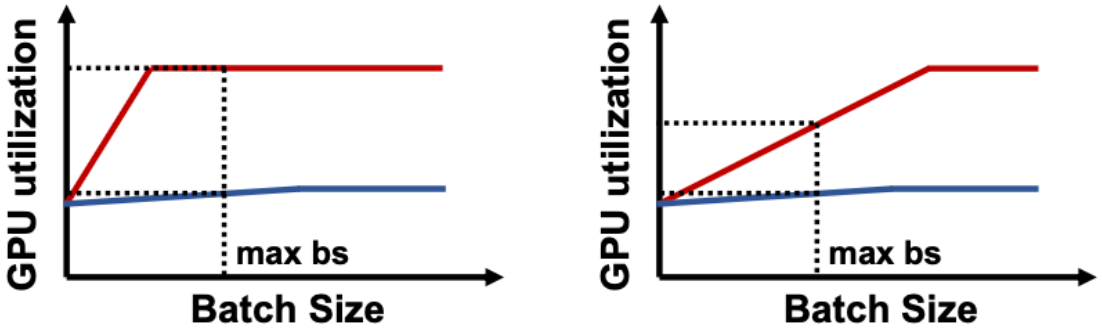
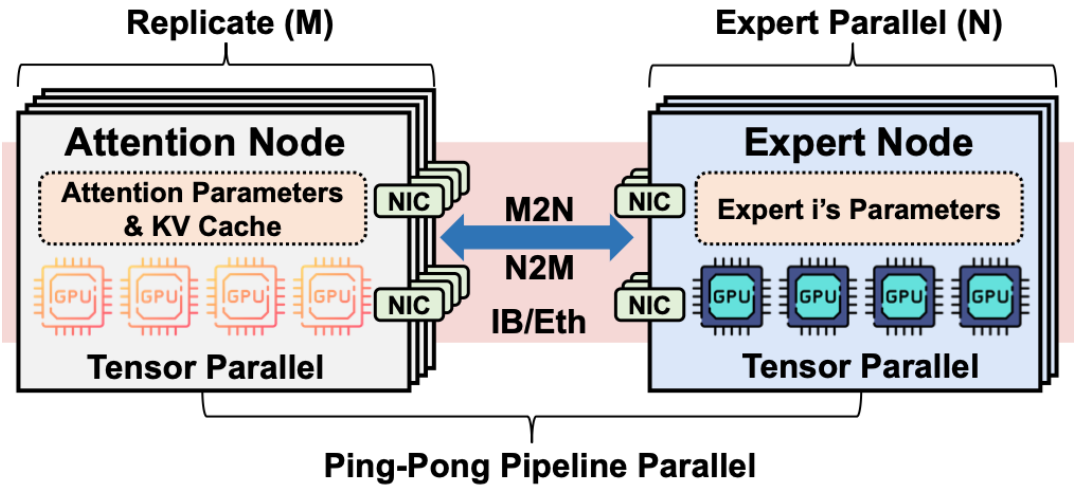


Figure 8. Elastic scale up in the decoding phase.

AF分离

- Attention FFN分离
- 背景
  - Decode阶段中：Attention、FFN具有不同的计算特征
    - Attention：访存密集（KVCache）
    - FFN/MoE：计算密集
  - Attention、FFN对于算力要求不同
- 实现方法
  - Attention、FFN部署在不同GPU/节点
  - 二者通过M2N算子进行通信
  - Ping-Pong流水线并行





AF分离

- M2N通信
  - 将token发送至其激活的专家所在GPU
  - 不同于专家并行中的All2All
  - 一种新的通信范式
- Ping-Pong流水线并行
  - 通过流水线排布，掩盖通信开销
  - 减少AF分离带来的Overhead

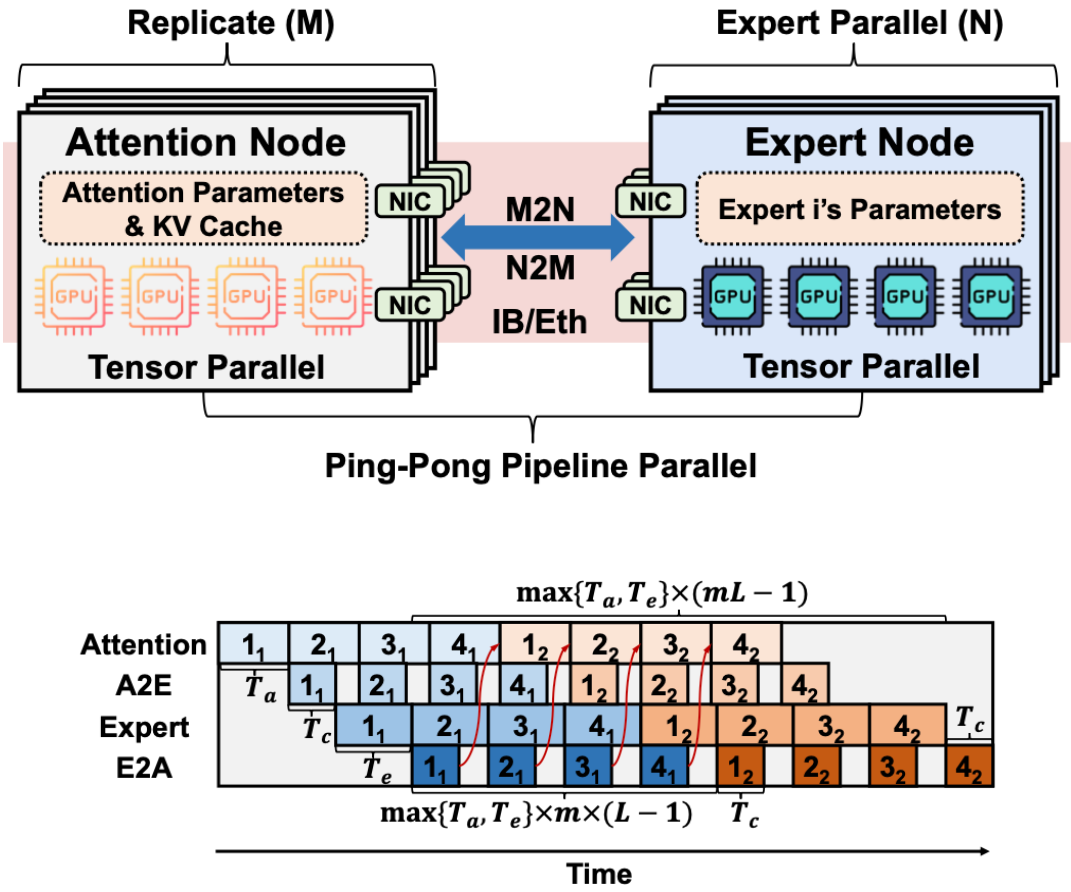
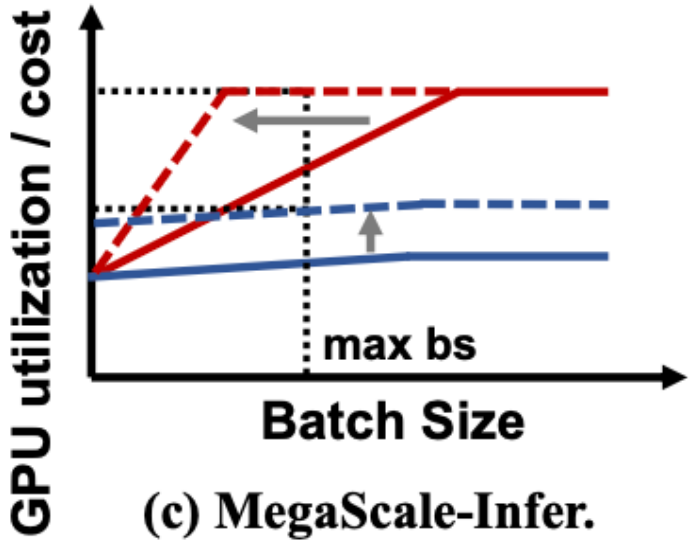


Figure 4 Illustration of ping-pong pipeline parallelism.

AF分离

- 优点
  - Attention、FFN解耦
    - 并行方案可变
    - 独立扩展
    - 针对性优化
  - 异构硬件友好
- 挑战
  - Ping-pong流水线气泡
  - 低延迟通信库
  - 推理系统复杂性提高



(c) MegaScale-Infer.

PART 05

总结与展望

总结

- 并行优化的核心目标：在有限算力与显存条件下最大化吞吐与响应速度。
- 当前主流方法：
  - **数据并行 / 张量并行 / 流水线并行 / 专家并行**（DP/TP/PP/EP） 各有优势与适用场景。
  - 混合并行策略结合不同方法，实现跨机、跨节点扩展。
- 典型案例：DeepSeek-V3
  - 通过 MoE + MLA 架构降低计算与KVCache存储压力。
  - 引入 专家负载均衡（EPLB） 等新型优化策略。
  - 配合 DeepEP、DeepGEMM 等高性能通信与算子库，有效提升推理性能。
- 前沿优化技术
  - PD分离/LoongServe
  - AF分离

未来发展方向&展望

- **更细粒度的计算解耦**
  - PD / AF 分离的理念将进一步扩展到模块级与算子级。
  - 支持更灵活的异构调度（GPU、NPU、CPU混合计算）。
- **通信与算力融合优化**
  - 低延迟通信库（如DeepEP）将持续演进，减少跨域All-to-All与M2N开销。
  - GEMM类内核将更深度结合FP8/混合精度与硬件特性（Hopper、Blackwell）。
- **系统层与算法层协同设计**
  - 从“模型并行”走向“系统共优”，强调推理框架、通信库、硬件栈一体化设计。
  - 未来推理架构将更加模块化、可重构、智能化。

- **展望**
  - **算力边界的突破**：从GPU集群走向异构超级计算平台。
  - **架构创新驱动推理革命**：PD/AF分离、MoE负载均衡等将成为标准组件。
  - **终极目标**：实现 **高吞吐、低延迟、可扩展、智能调度** 的下一代大模型推理系统。

科技生态圈峰会 + 深度研习

——1000 + 技术团队的共同选择

KEYLINK ing

AiDD 峰会

AiDD峰会 上海站
AI+研发数字峰会
时间：2026.05.22-23

AiDD峰会 北京站
AI+研发数字峰会
时间：2026.08.21-22

AiDD峰会 深圳站
AI+研发数字峰会
时间：2026.11.20-21

AiDD峰会详情

AiDD × AI+产品峰会

AiDD × 上海站
AI+产品创新峰会
时间：2026.01.16-17

AiDD × 北京站
AI+产品创新峰会
时间：2026.08.23-24

产品峰会详情

AiDD × K

2026 AI+ PRODUCT INNOVATION SUMMIT

01.16-17 · ShangHai

AI+产品创新峰会

Track 1: AI 产品战略与创新设计

从0到1的AI原生产品构建

论坛1: AI时代的用户洞家与需求发现

论坛2: AI原生产品战路与商业模式重构

论坛3: AgenticAI产品创新与交互设计

Track 2: AI 产品开发与工程实践

从1到10的工程化落地实践

论坛1: 面向Agent智能体的产品开发

论坛2: 具身智能与AI硬件产品

论坛3: AI产品出海与本地化开发

Track 3: AI 产品运营与智能演化

从10到100的AI产品运营

论坛1: AI赋能产品运营与增长黑客

论坛2: AI产品的数据飞轮与智能演化

论坛3: 行业爆款AI产品案例拆解

2-hour Speech: 回归本质

用户洞察的第一性

——2小时思维与方法论工作坊

在数字爆炸、AI迅速发展的时代，仍然考验“看见”的“同理心”

Panel 1: 出海前瞻

“出海避坑地图”圆桌对话

——不止于翻译:

AI时代的出海新范式

Panel 2: 失败复盘

为什么很多AI产品“叫好不叫座”?

——从伪需求到真价值:

AI产品商业化落地的关键挑战

智能重构产品 数据驱动增长

Reinventing Products with Intelligence, Driven by Data

80+ 业界大咖 汇聚产品大咖的真知灼见

40+ 创新案例 开拓全新AI+产品视角

10+ 分论坛 涵盖产品到商业增长的全链路

800+ 听众规模 共同探索产品的智能重构

扫码查看会议详情

第8届AI+研发数字峰会
AI+ Development Digital Summit

感谢聆听！

扫码领取会议PPT资料