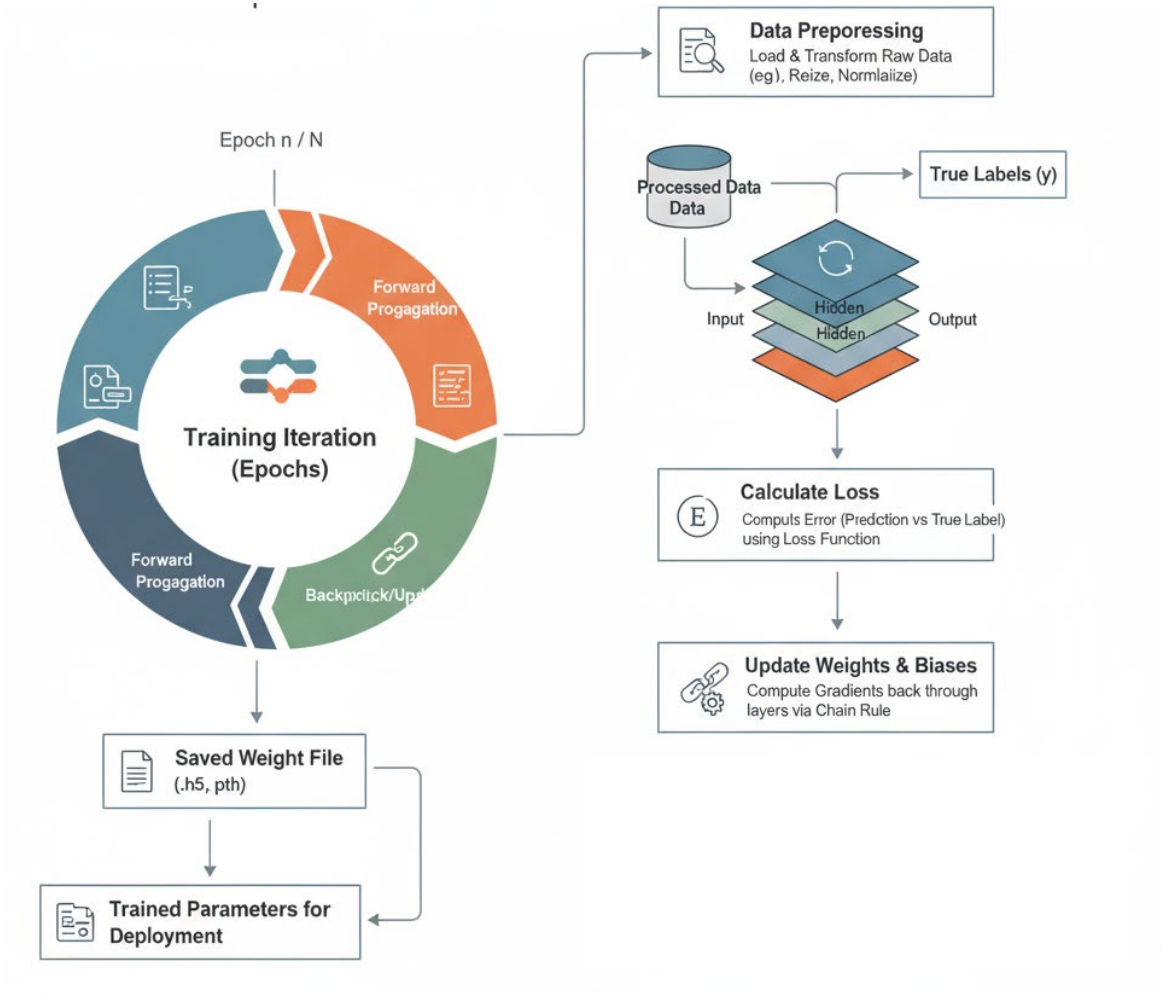
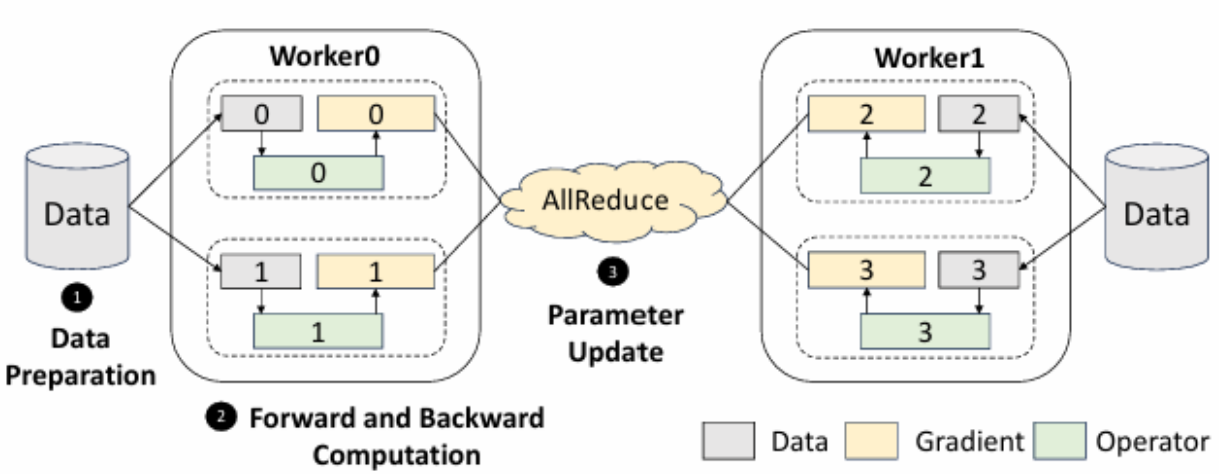
大模型训练优化基础

周宇航

南京大学

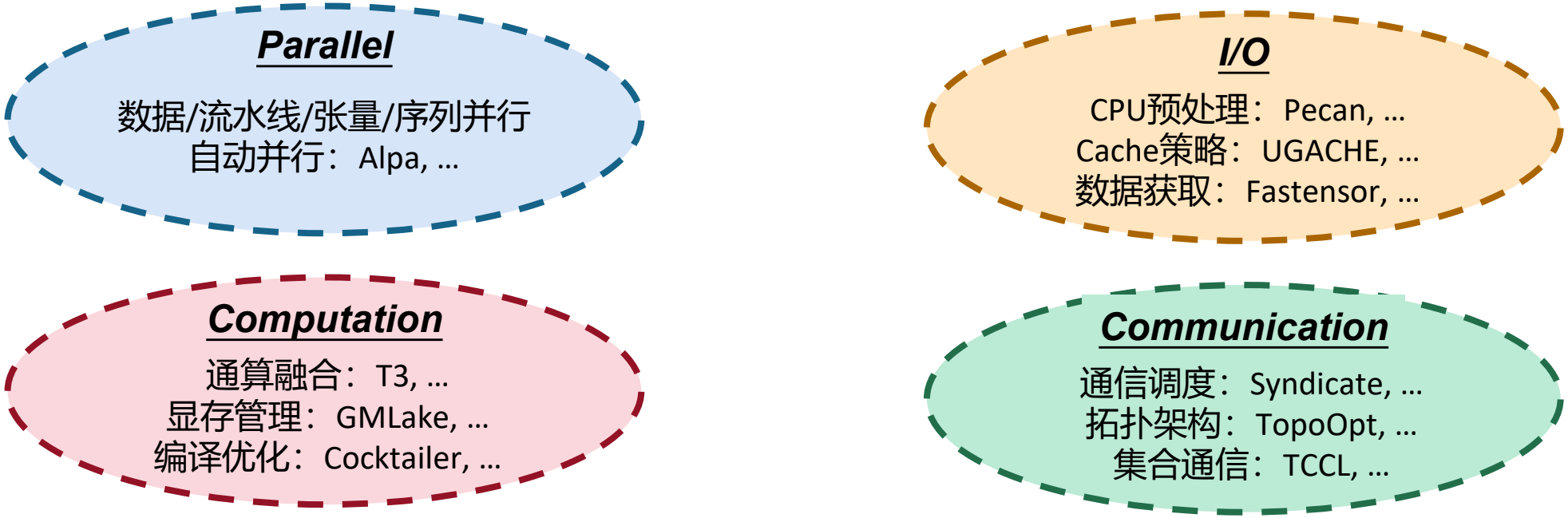
模型训练


Epoch n / N
Training Iteration
(Epochs)
Forward
Progagation
Forward
Progagation
Data Preporessing
Load & Transform Raw Data
(eg), Reize, Normlaiize)
Processed Data
Data
True Labels (y)
Input
Hidden
Hidden
Output
Calculate Loss
Computs Error (Predction vs True Label)
using Loss Function
Update Weights & Biases
Compute Gradients back through
layers via Chain Rule
Saved Weight File
(.h5, pth)
Trained Parameters for
Deployment


Worker0
Worker1
Data
AllReduce
Data
0
0
0
1
1
1
2
2
2
3
3
3
1 Data
Preparation
2 Forward and Backward
Computation
3
Parameter
Update
Data
Gradient
Operator

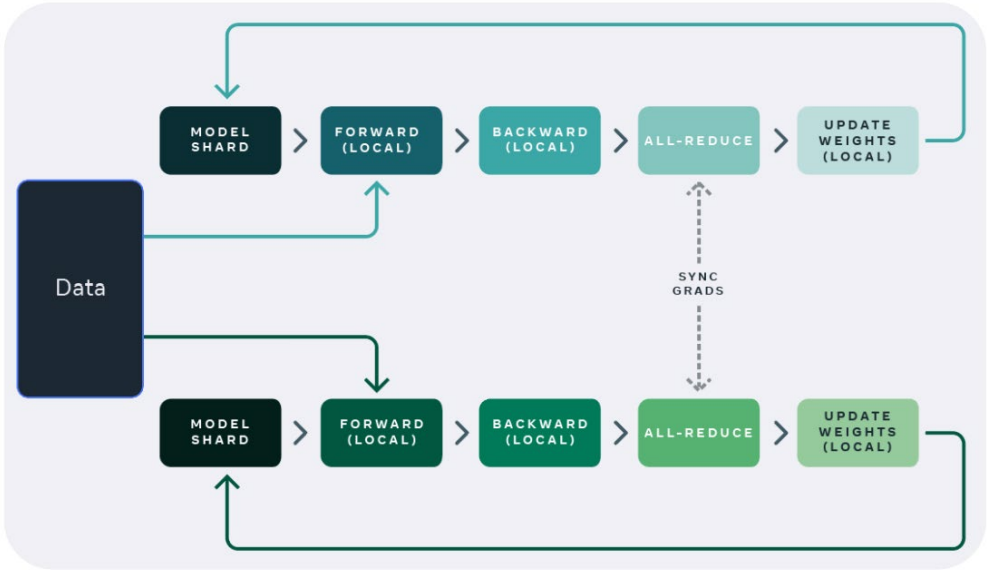
Question: 大模型训练有哪些方面的性能问题?

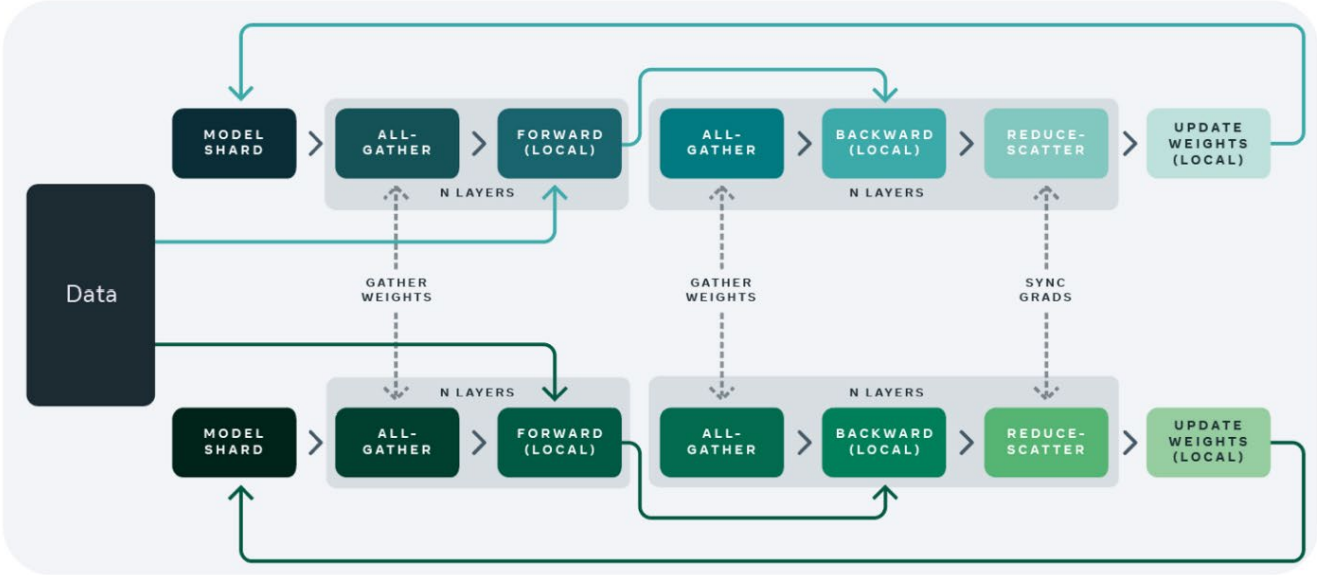
优化分类


Parallel
数据/流水线/张量/序列并行
自动并行：Alpa, …
I/O
CPU预处理：Pecan, …
Cache策略：UGACHE, …
数据获取：Fastensor, …
Computation
通算融合：T3, …
显存管理：GMLake, …
编译优化：Cocktailer, …
Communication
通信调度：Syndicate, …
拓扑架构：TopoOpt, …
集合通信：TCCL, …

数据并行

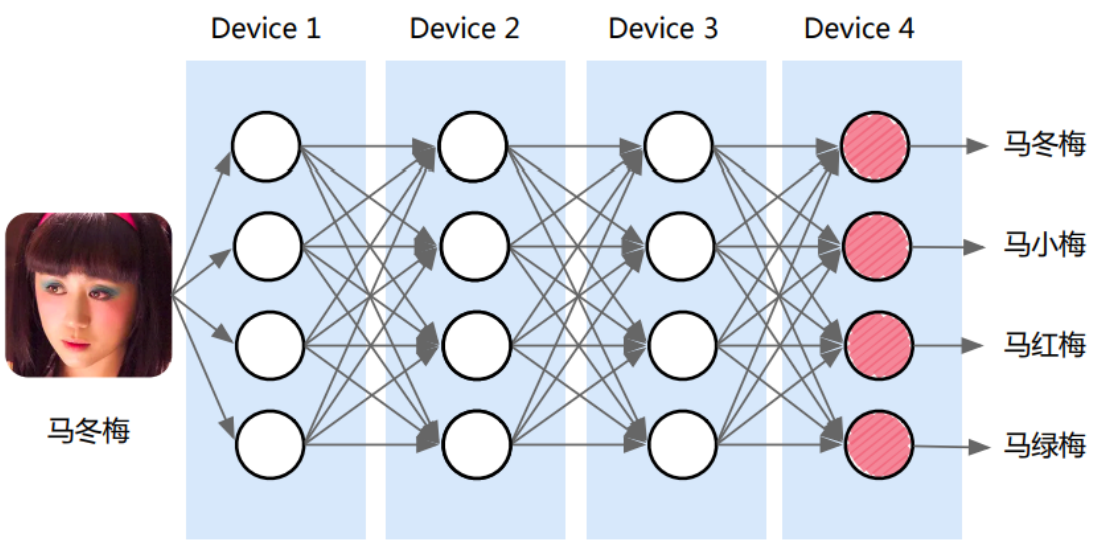
- 数据并行会自动拆分训练数据，并将模型作业发送到多个 GPU。每个模型完成后，数据并行会累积梯度。
- FSDP 会将模型的所有参数、梯度和优化器状态分片到多个GPU上，并可以选择将分片后的模型参数卸载到 CPU 上。



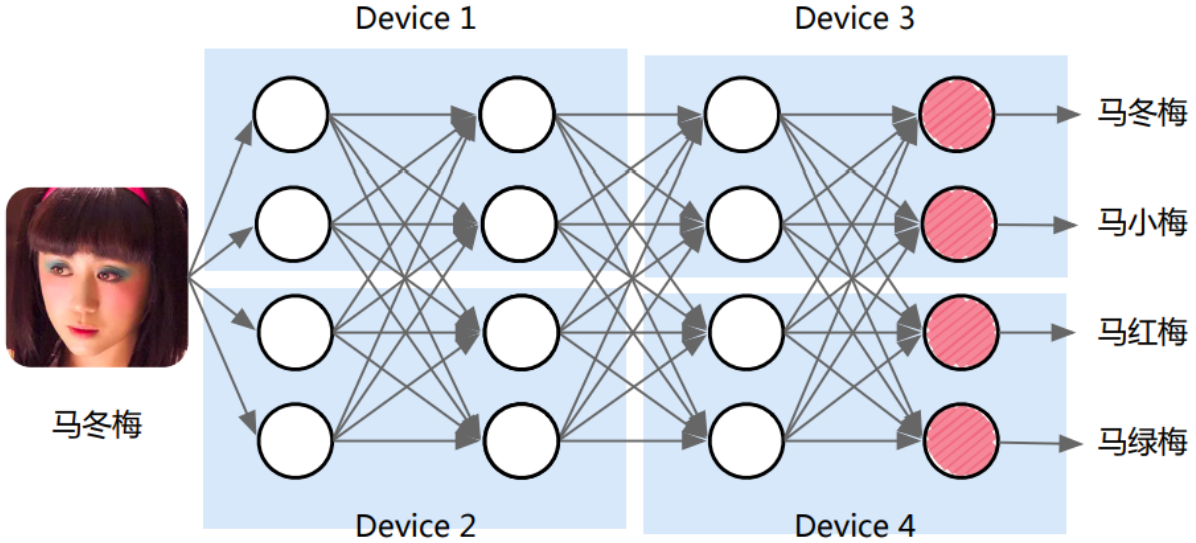


模型并行

- 流水线并行：按模型layer层切分到不同设备，即层间并行
- 张量并行：将计算图中的层内的参数切分到不同设备，即层内并行



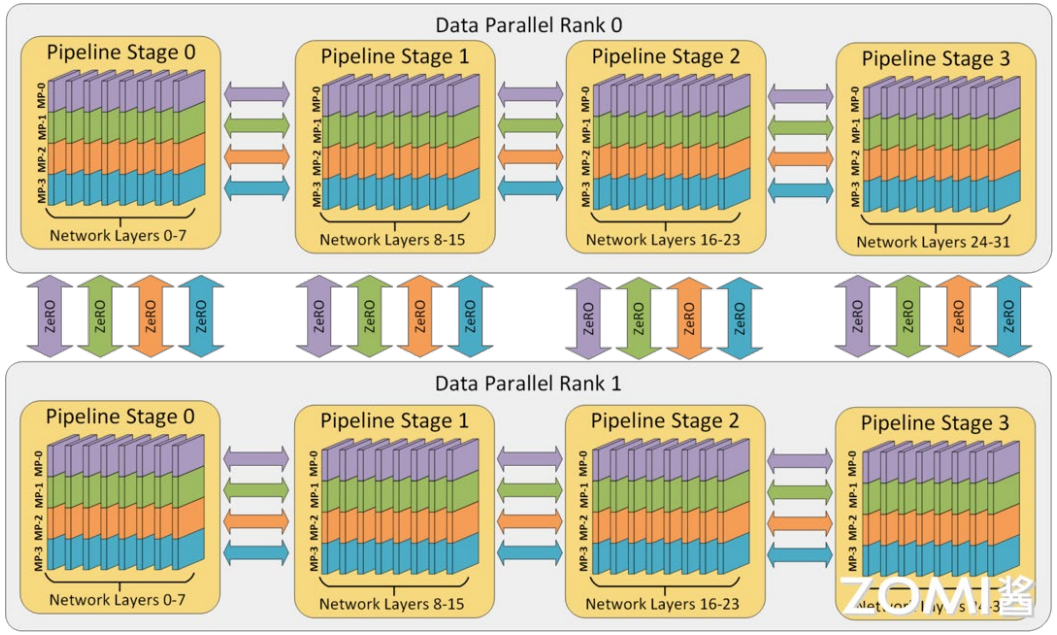
流水线并行（层间并行）



张量并行（层内并行）

混合并行

- DP+PP+TP



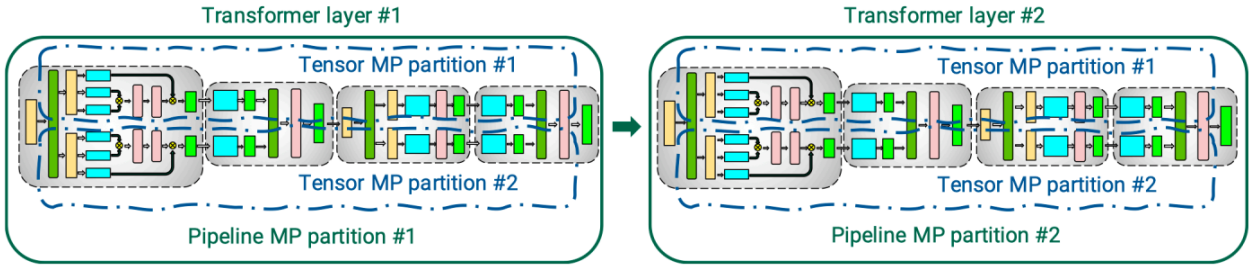
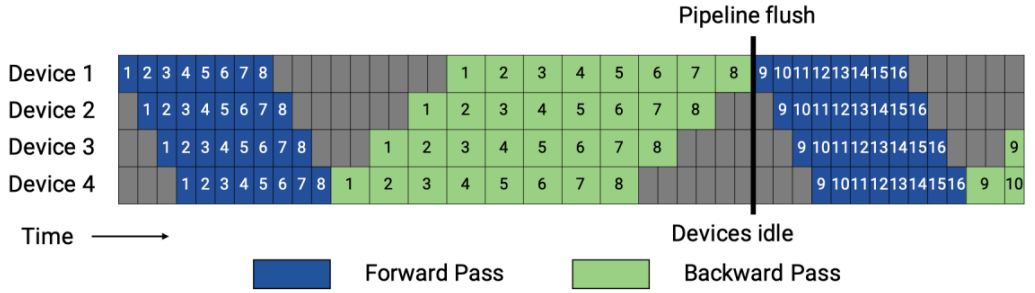
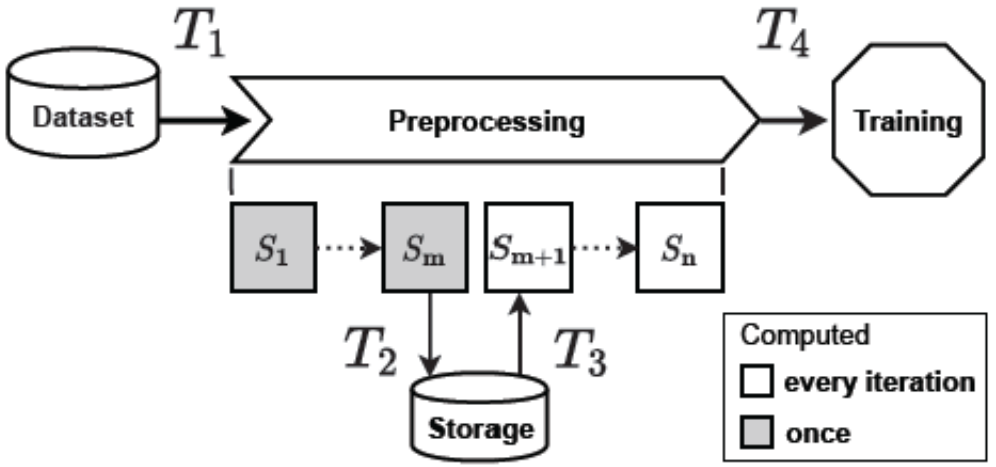


Figure 2: Combination of tensor and pipeline model parallelism (MP) used in this work for transformer-based models.



I/O优化



问题来源

- 随机访存
- CPU预处理
- 多线程并行效率低

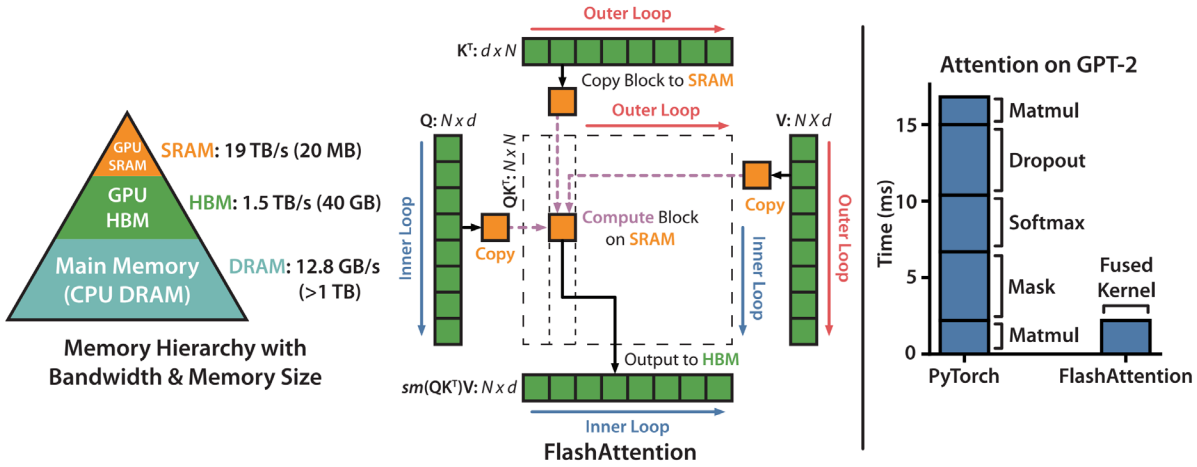
解决方案

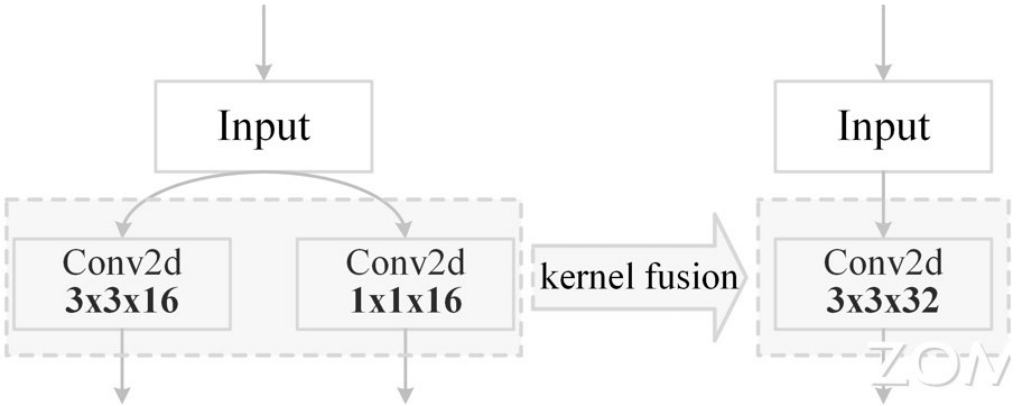
- 文件合并、数据压缩
- 缓存策略
- 多线程并行

Where Is My Training Bottleneck? Hidden Trade-Offs in Deep Learning Preprocessing Pipelines (SIGMOD 22)

计算优化

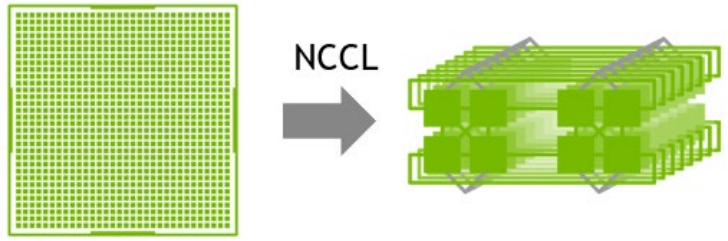
- Flash Attention
- 算子融合

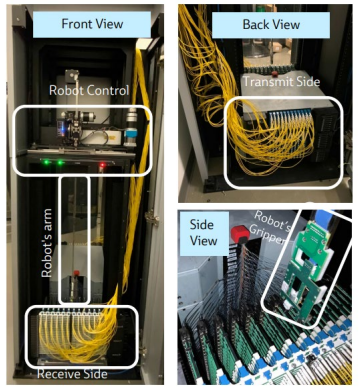


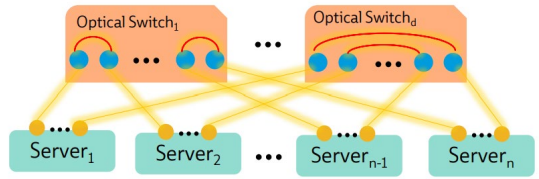


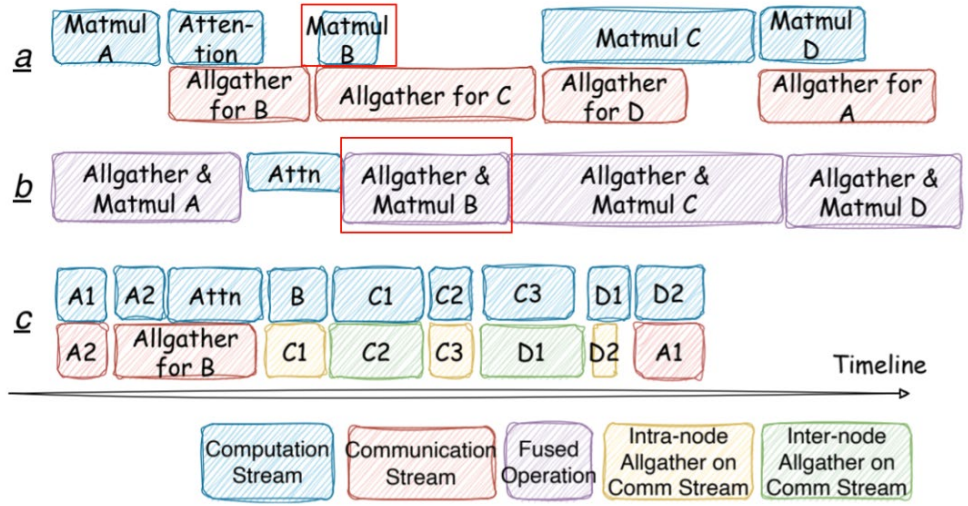
通信优化

- NCCL等通信库
- 通信拓扑
- 计算通信并行









Thanks!

Q&A