

- 分布式训练
  - 概述
  - 分布式训练并行策略
  - 分布式训练的集群架构
  - DeepSpeed

# 分布式训练

## 概述

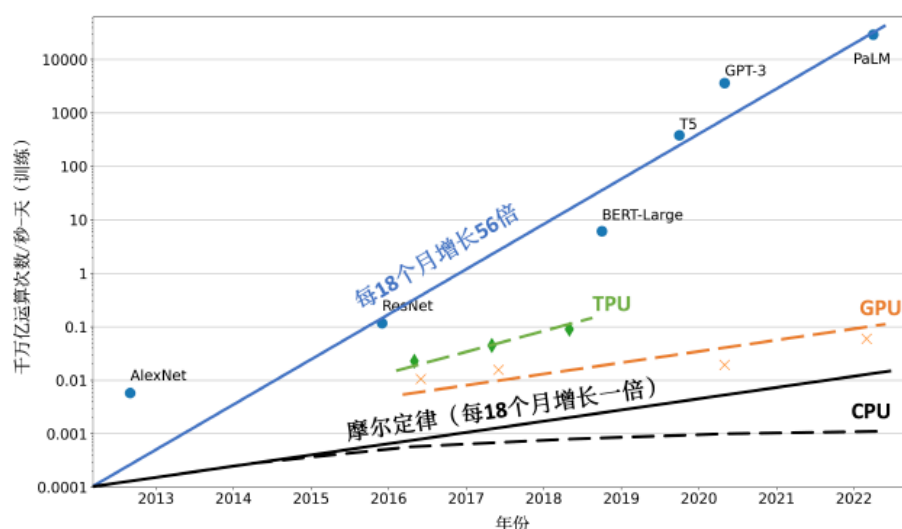



图 4.2 机器学习模型参数量增长和计算硬件的算力增长对比<sup>[128]</sup>

- 总训练速度  $\propto$  单设备计算速度  $\times$  计算设备总量  $\times$  多设备加速比
- 计算墙：单个计算设备所能提供的计算能力与大语言模型所需的总计算量之间存在巨大差异。2022 年3 月发布的NVIDIA H100 SXM 的单卡FP16 算力也只有2000 TFLOPs，而GPT-3则需要314 ZFLOPs 的总算力，两者相差了8 个数量级。
- 显存墙：单个计算设备无法完整存储一个大语言模型的参数。GPT-3 包含1750 亿参数，如果采用FP16 格式进行存储，需要700GB 的计算设备内存空间，而NVIDIA H100 GPU 只有80 GB 显存。
- 通信墙：分布式训练系统中各计算设备之间需要频繁地进行参数传输和同步。由于通信的延迟和带宽限制，这可能成为训练过程的瓶颈。GPT-3 训练过程中，如果分布式系统中存在128个模型副本，那么在每次迭代过程中至少需要传输89.6TB 的梯度数据。而截止2023 年8 月，单个InfiniBand 链路仅能够提供不超过800Gb/s 带宽。

# 分布式训练并行策略

- 首先可以对数据进行切分（**Partition**），并将同一个模型复制到多个设备上，并行执行不同的数据分片，这种方式通常被称为**数据并行（Data Parallelism, DP）**
- 还可以对模型进行划分，将模型中的算子分发到多个设备分别完成，这种方式通常被称为**模型并行（Model Parallelism, MP）**1705152398614
  - 把模型的层切分到不同设备-流水线并行-设备平均使用率降低

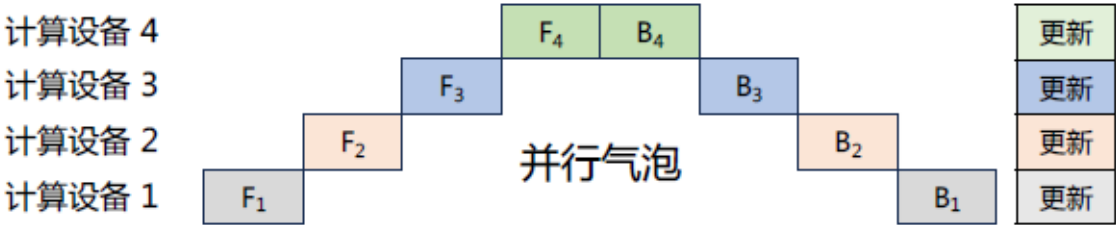


图 4.6 流水线并行样例

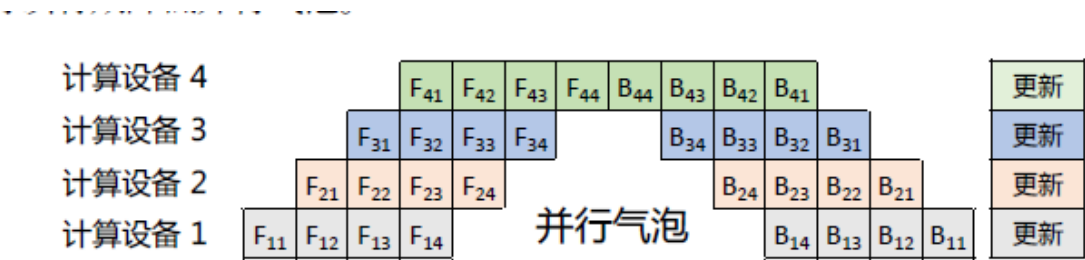


图 4.7 GPipe 策略流水线并行样例<sup>[131]</sup>

- 把计算图层内的参数切分到不同设备-张量并行

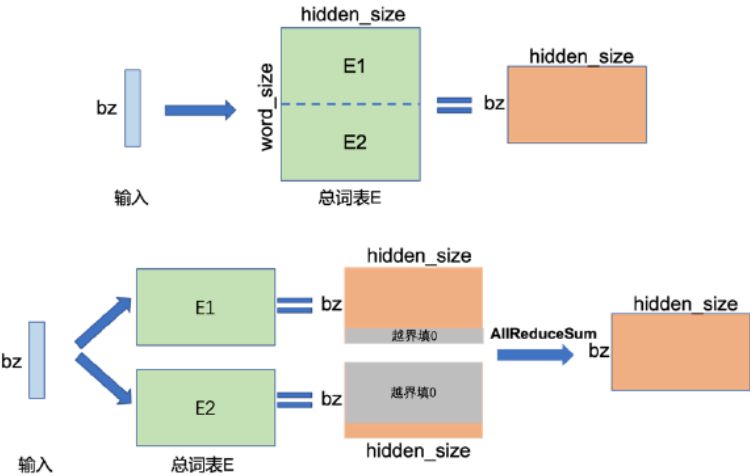


图 4.9 两节点 Embedding 算子张量并行示例

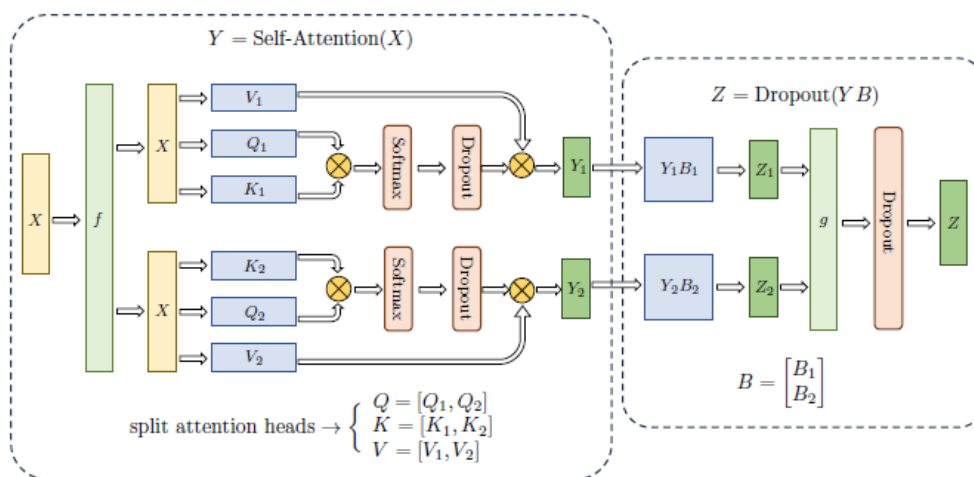


图 4.13 多头自注意力机制张量并行示意图<sup>[130]</sup>

- 当训练超大规模语言模型时，往往需要同时对数据和模型进行切分，从而实现更高层次的并行，这种方式通常被称为混合并行（Hybrid Parallelism, HP）

## 分布式训练的集群架构

在由高速网络组成的高性能计算上构建分布式训练系统，主要有两种常见架构：参数服务器架构（Parameter Server, PS）和去中心化架构（Decentralized Network）

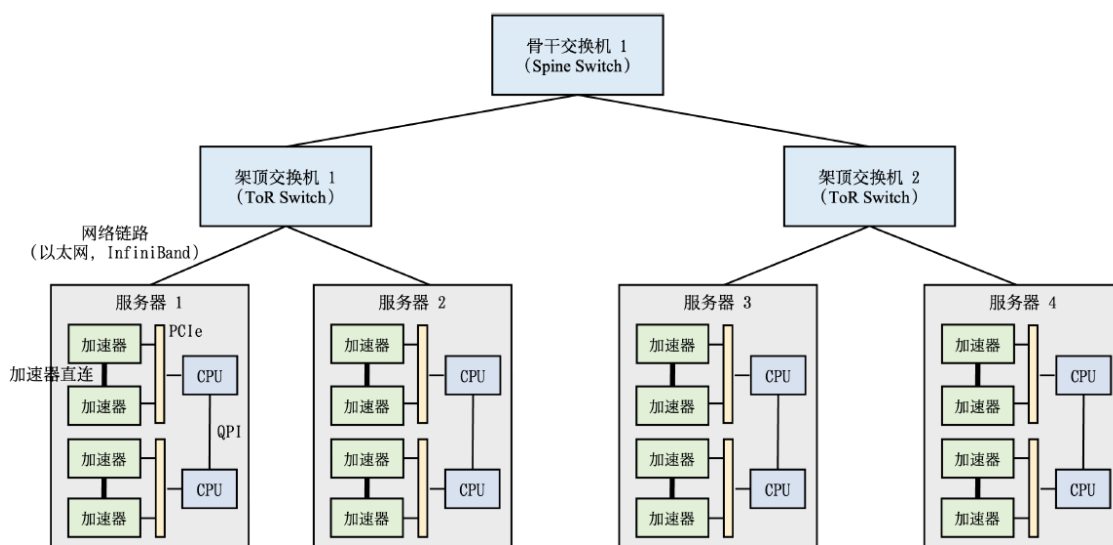


图 4.18 典型用于分布式训练的计算集群硬件组成<sup>[128]</sup>

# DeepSpeed

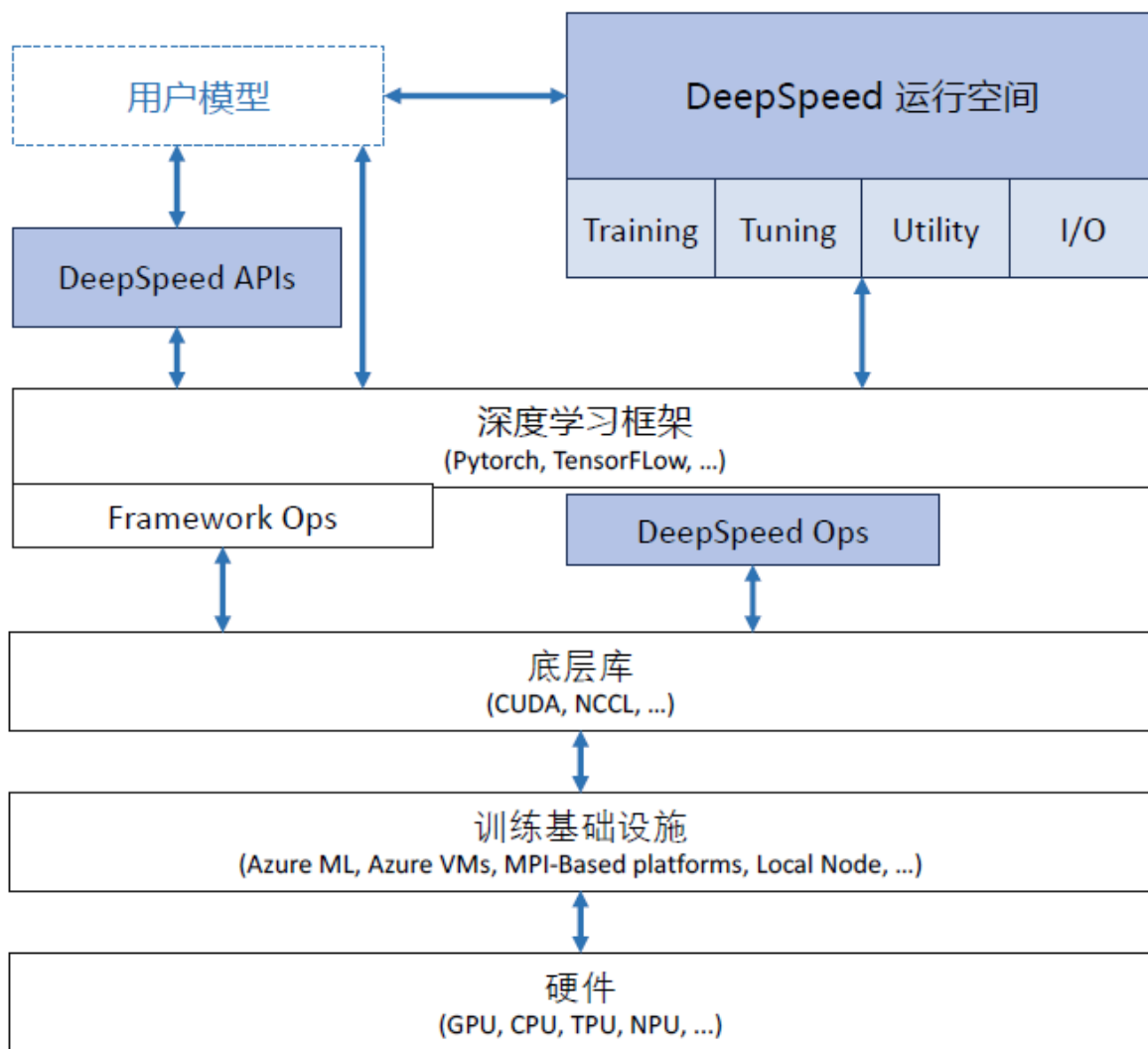


图 4.30 DeepSpeed 软件架构

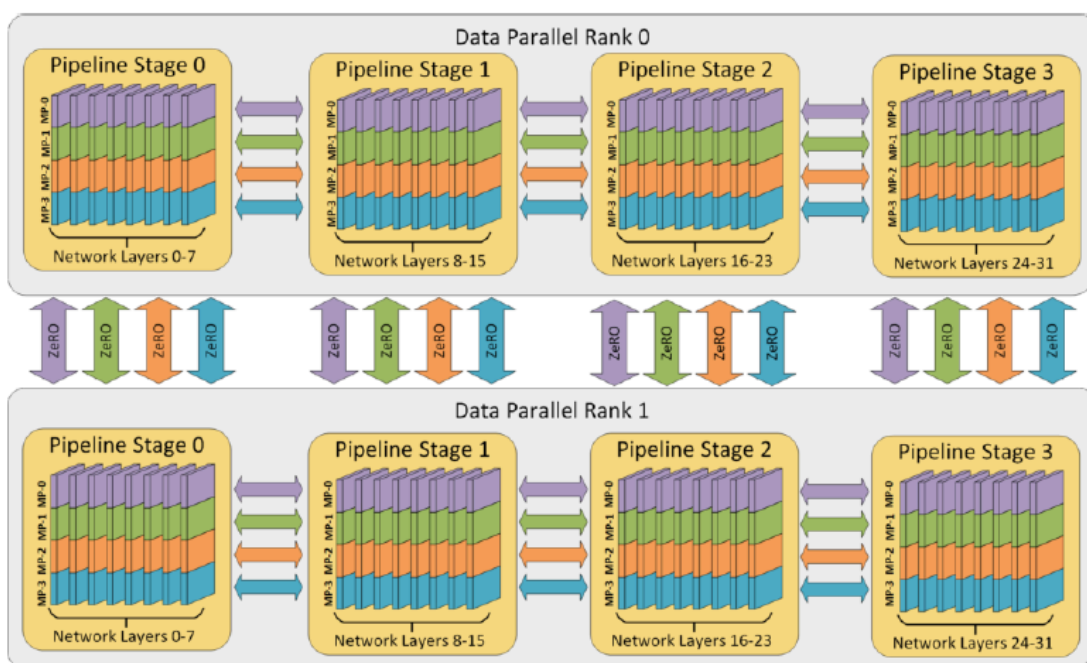


图 4.29 DeepSpeed 3D 并行策略示意图<sup>[138]</sup>

- 由Microsoft 公司开发的开源深度学习优化库, 基于Pytorch构建