分布式训练系列

来个总结

ZOMI

BUILDING A BETTER CONNECTED WORLD

Ascend & MindSpore

www.hiascend.com
www.mindspore.cn

About 关于本内容

I. 具体内容

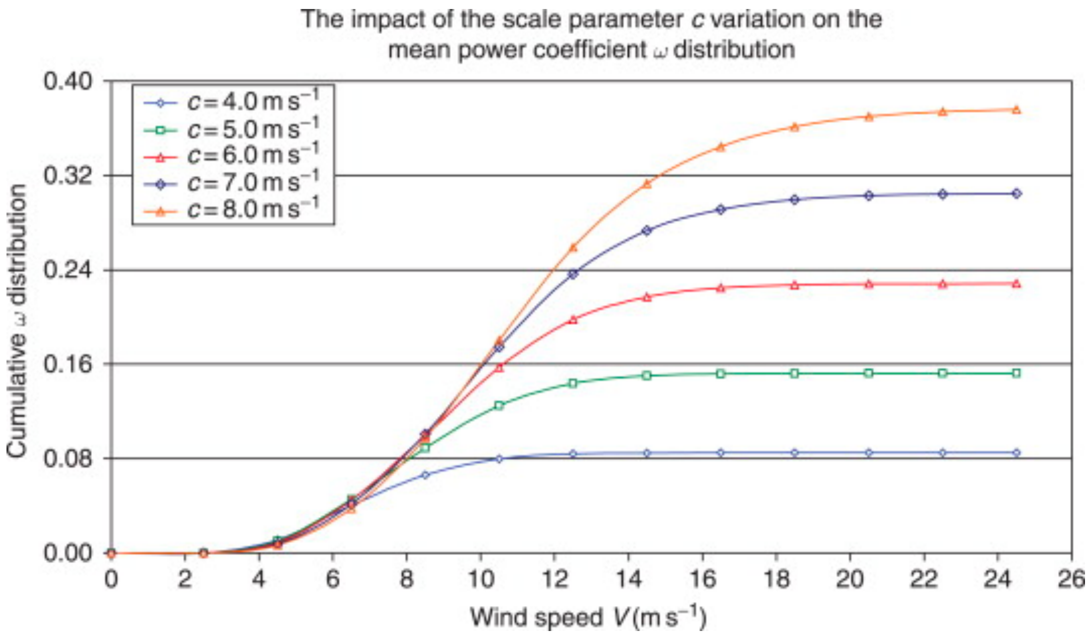
- 大模型训练挑战
- AI框架的分布式
- AI集群架构
- AI集群通信
- 大模型算法
- 分布式并行算法
- 大模型混合并行
- 内存和计算优化

加速比

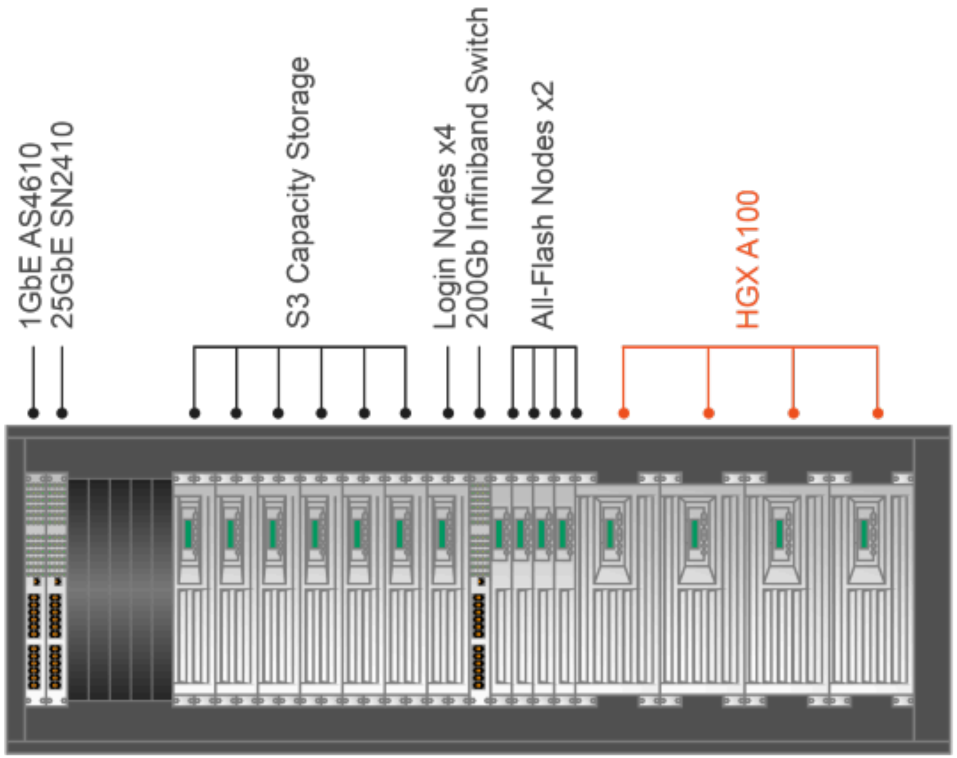
假设单设备吞吐量为T，n个设备系统的吞吐量应为nT，系统实际达到吞吐量为Tn，则加速比为：

$$scale\ factor = \frac{T_n}{nT}$$

边际效应受限

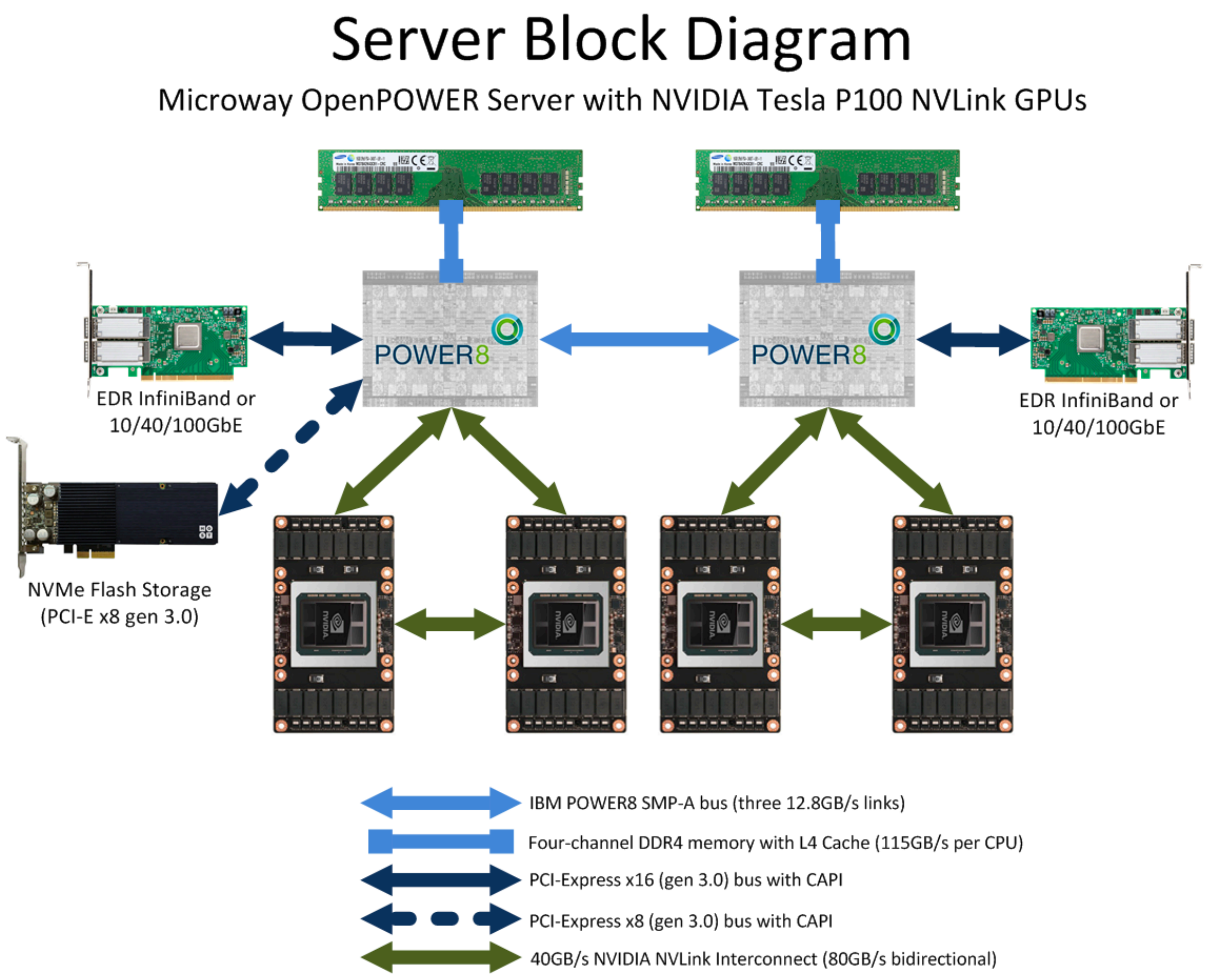


通讯硬件



- 机器内通信
  - 共享内存
  - PCIe
  - NVLink（直连模式）
- 机器间通信
  - TCP/IP网络
  - RDMA网络（直连模式）

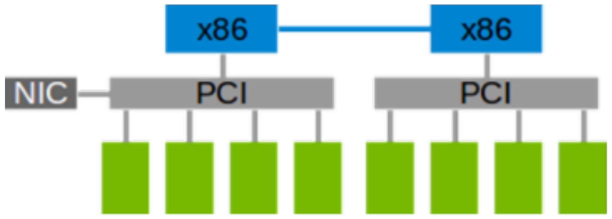
通讯硬件



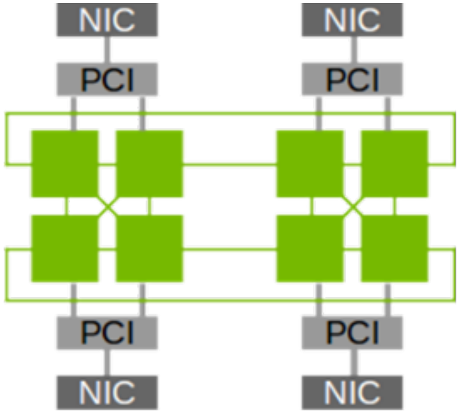
- 机器内通信
  - 共享内存
  - PCIe
  - NVLink（直连模式）
- 机器间通信
  - TCP/IP网络
  - RDMA网络（直连模式）

通信软件：提供集合通信

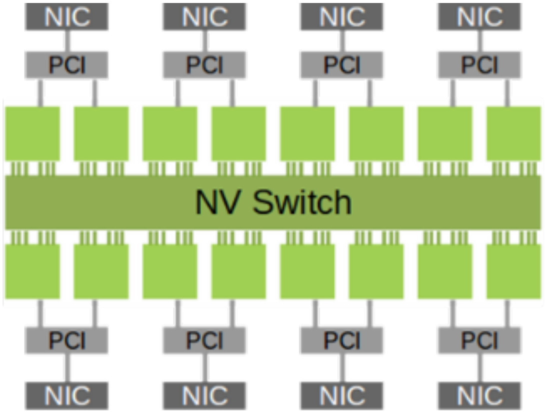
- **MPI**
  - 通用接口，可调用 Open-MPI, MVAPICH2, Intel MPI, etc.
- **NCCL / HCCL**
  - GPU通信优化，仅支持集中式通信



PCI only



DGX-1

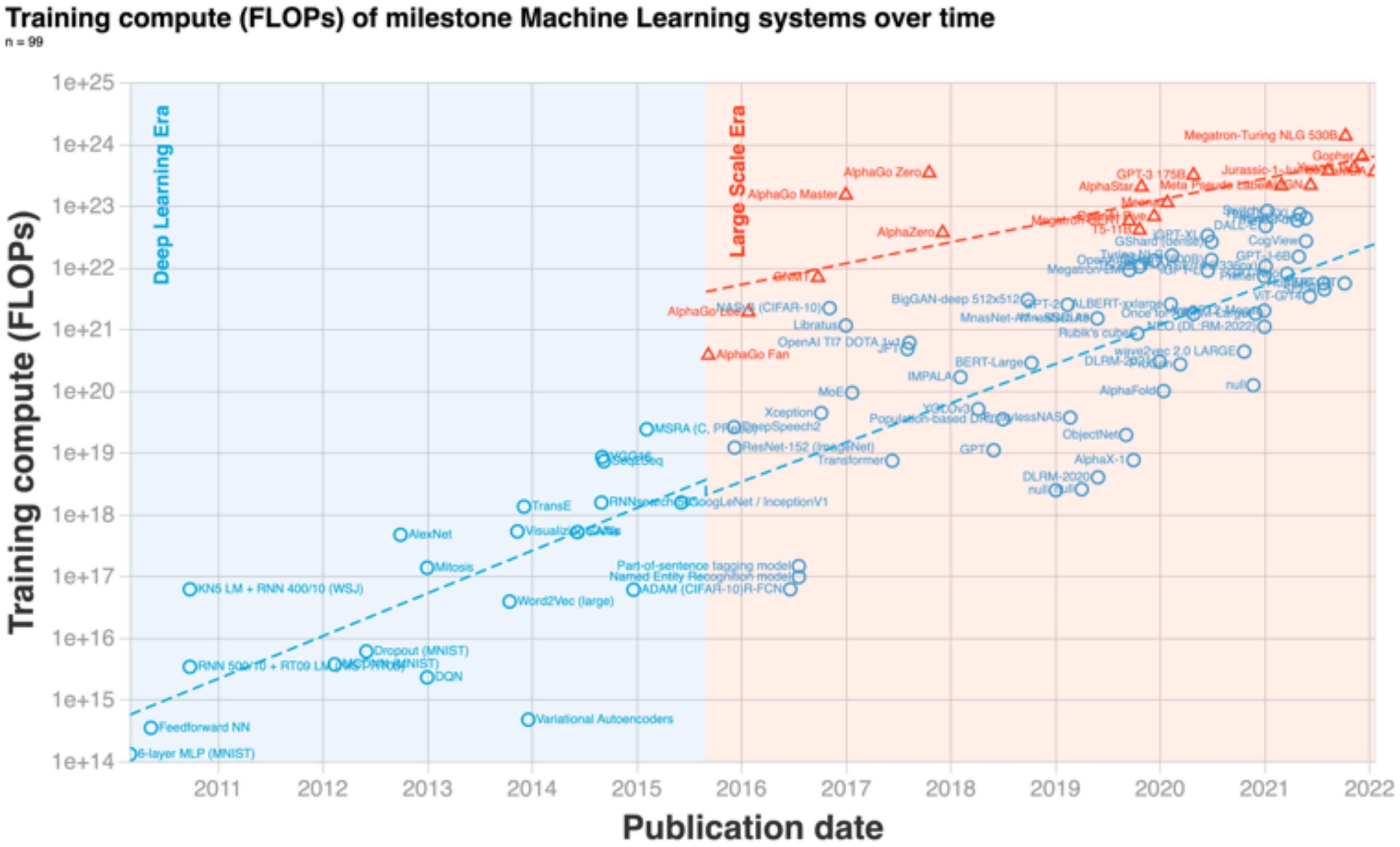


DGX-2

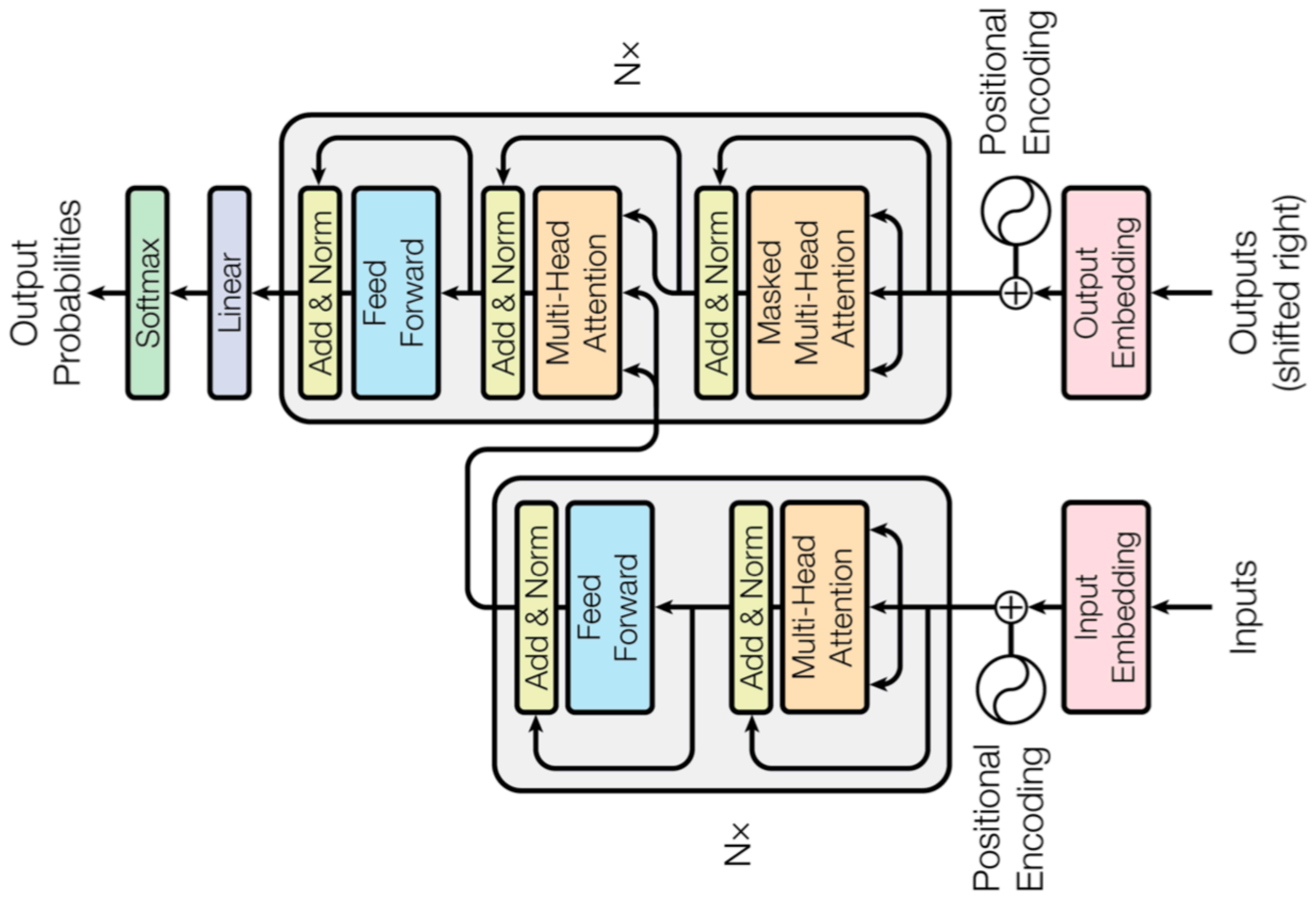
深度学习迎来大模型（Foundation Models）

1. 自监督学习方法，可以减少数据标注，降低训练研发成本
2. 解决模型碎片化，提供预训练方案
3. 模型参数规模越大，有望进一步突破现有模型结构的精度局限

- e.g. 语言模型 GPT-3
  - 8 张 V100，训练时长 36 年
  - 512 张 V100，训练近 7 个月

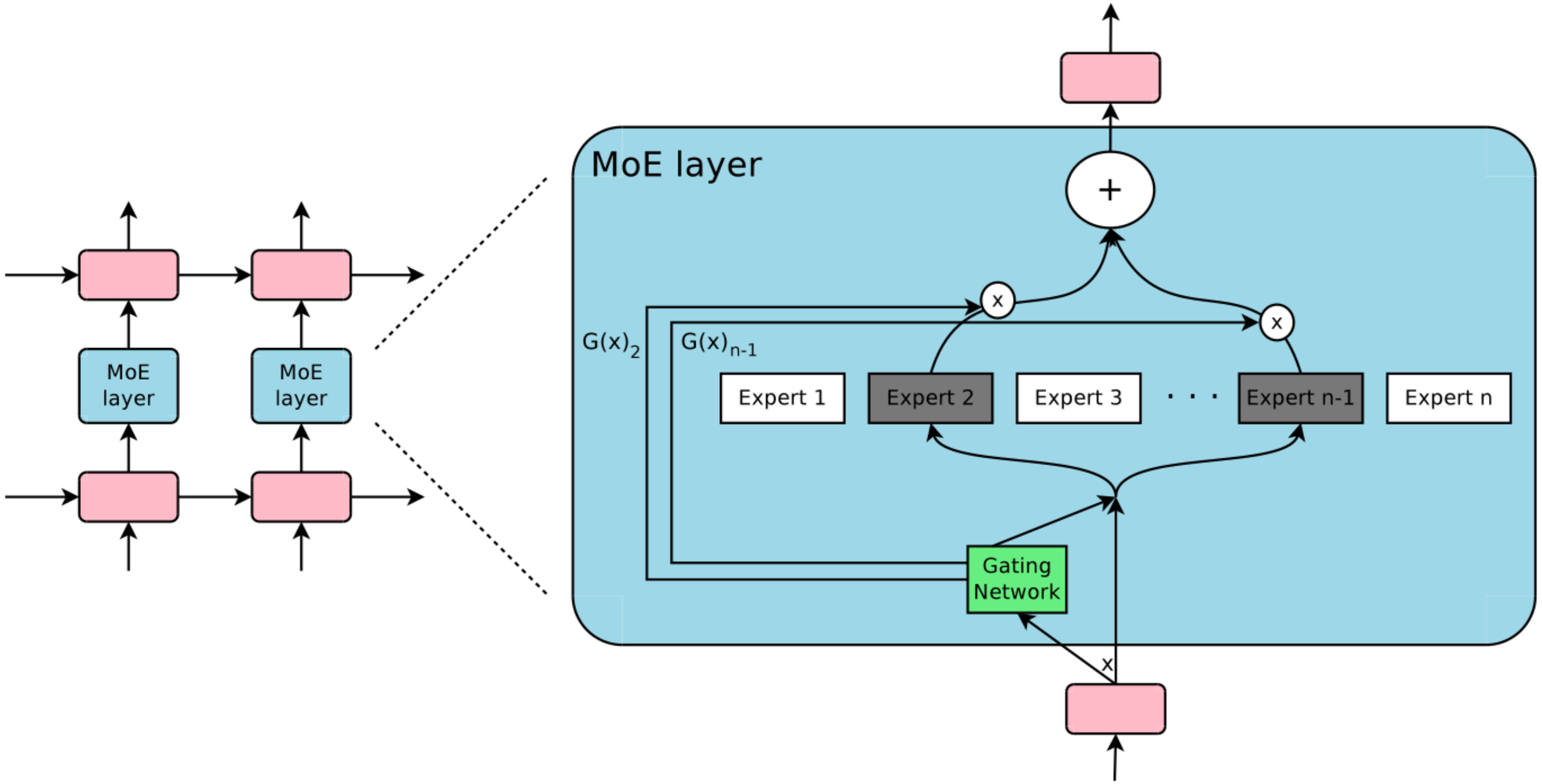


Transformer 取代RNN、CNN进入大模型时代



MoE 稀疏混合专家结构模型参数量进一步突破

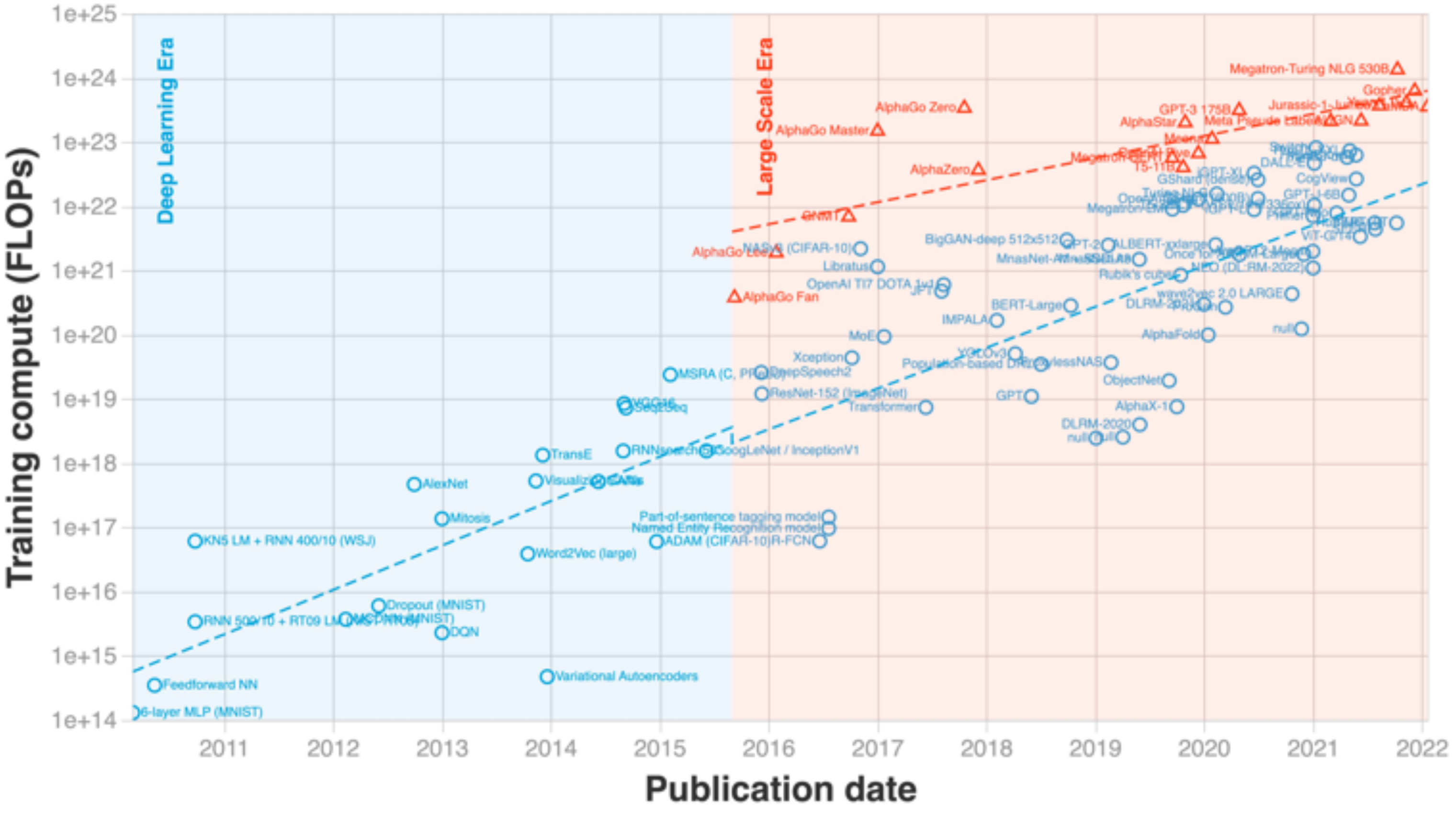
稀疏门控专家混合模型（Sparsely-Gated MoE）：旨在实现条件计算，即神经网络的某些部分以每个样本为基础进行激活，作为一种显著增加模型容量和能力而不必成比例增加计算量的方法。



深度学习迎来大模型（Foundation Models）

Training compute (FLOPs) of milestone Machine Learning systems over time

n = 99



分布式训练系统

定义：能够分布式地执行深度学习的训练的系统

- **分布式用户接口**
  - 用户通过接口，实现模型的分布化
- **执行单节点训练**
  - 产生本地执行的逻辑
- **通信协调**
  - 实现多节点之间的通信协调

意义：提供易于使用，高效率的分布式训练

分布式训练系统

大模型算法
分布式并行与同步策略
硬件内存优化和计算优化
分布式训练系统
分布式训练 通信与协调
分布式并行架构

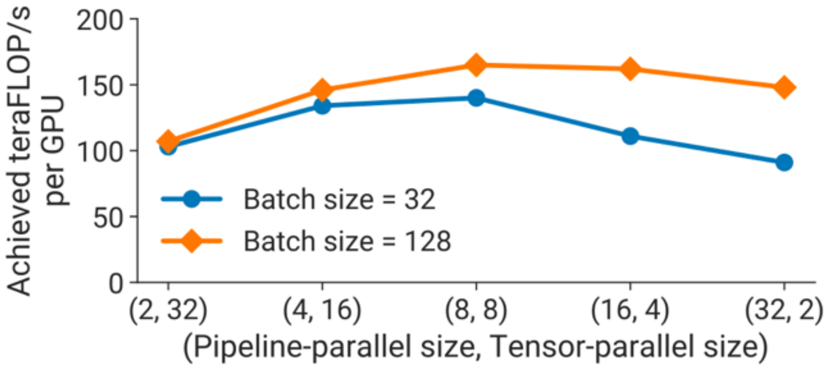


Figure 13: Throughput per GPU of various parallel configurations that combine pipeline and tensor model parallelism using a GPT model with 162.2 billion parameters and 64 A100 GPUs.

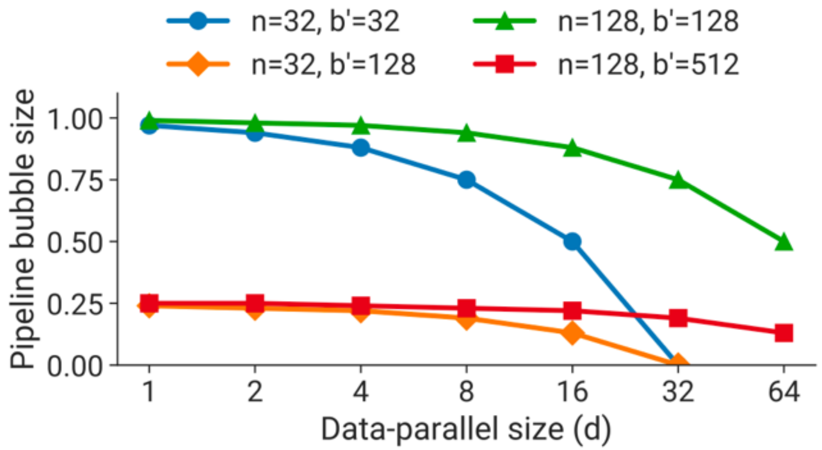


Figure 6: Fraction of time spent idling due to pipeline flush (pipeline bubble size) versus data-parallel size (d), for different numbers of GPUs (n) and ratio of batch size to microbatch size ($b' = B/b$).

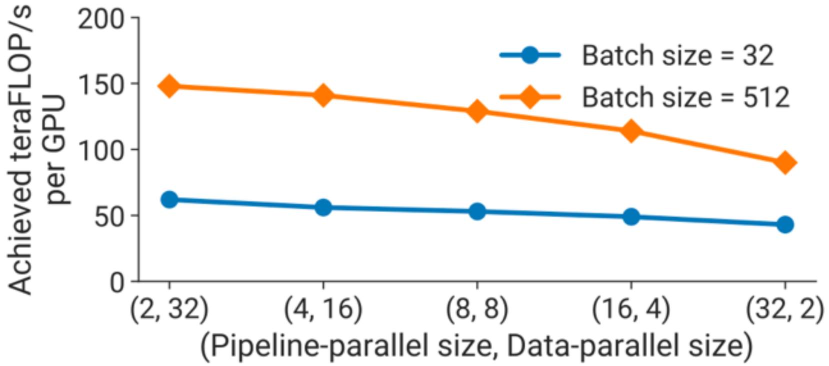


Figure 14: Throughput per GPU of various parallel configurations that combine data and pipeline model parallelism using a GPT model with 5.9 billion parameters, three different batch sizes, microbatch size of 1, and 64 A100 GPUs.

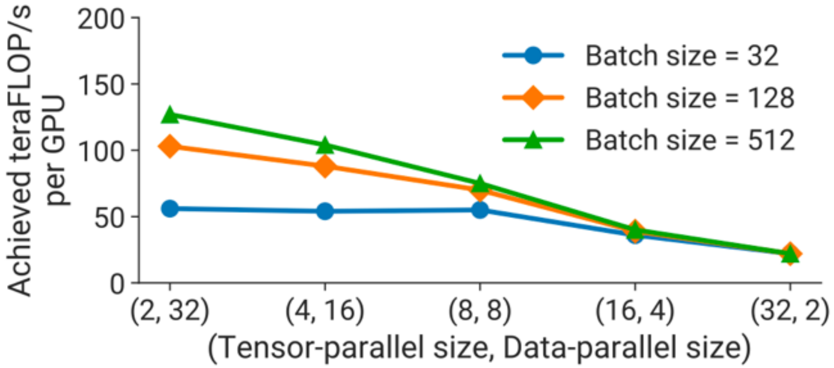


Figure 15: Throughput per GPU of various parallel configurations that combine data and tensor model parallelism using a GPT model with 5.9 billion parameters, three different batch sizes, microbatch size of 1, and 64 A100 GPUs.

Narayanan, Deepak, et al. "Efficient large-scale language model training on gpu clusters using megatron-lm." Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. 2021.

大模型训练挑战

成功路上总有绊脚石！

大模型训练挑战

内存墙

200B参数，参数内存占用754GB内存，训练过程需要3500GB+内存（权重+激活+优化器状态），一个模型需要100多张卡才能存放下

性能墙

大规模训练技术中，不仅要求AI芯片的计算性能足够强悍，同时也依赖于AI框架的大规模分布式训练的运行和调度效率，以及在分布式并行等各种优化手段的权衡。

通讯墙

通讯过程，需要综合考虑数据参数量、计算量、计算类型、数据样本量、集群带宽拓扑和通讯策略等不同的因素，才能设计出性能较优的切分策略，最大化利用通讯效率，提高通讯比。

调优墙

在数千节点的集群上，要保证计算的正确性/性能/可用性，手工分布式难免全面兼顾

大模型的分布式训练

考验的是算法、数据、框架、资源调度等

全栈和全流程的综合能力

大模型训练挑战

- **深度学习训练耗时：**

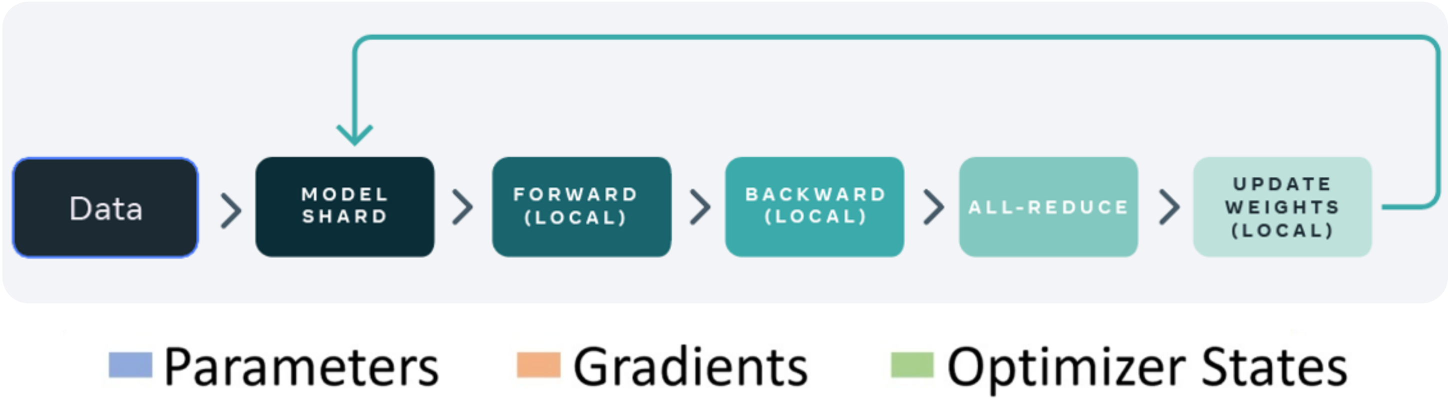
$$\text{训练耗时} = \underbrace{\text{训练数据规模} \times \text{单步计算量}}_{\text{模型相关}} \,/\, \underbrace{\text{计算速率}}_{\text{可变因素}}$$

- **计算速率：**

$$\text{计算速率} = \underbrace{\text{单设备计算速率}}_{\substack{\text{混合精度}\\ \text{算子融合}\\ \text{激活重计算}\\ \text{加速优化器}}} \times \underbrace{\text{设备数}}_{\substack{\text{服务器架构}\\ \text{通信拓扑优化}}} \times \underbrace{\text{多设备并行效率（加速比）}}_{\substack{\text{数据并行}\\ \text{模型并行}\\ \text{流水并行}}}$$

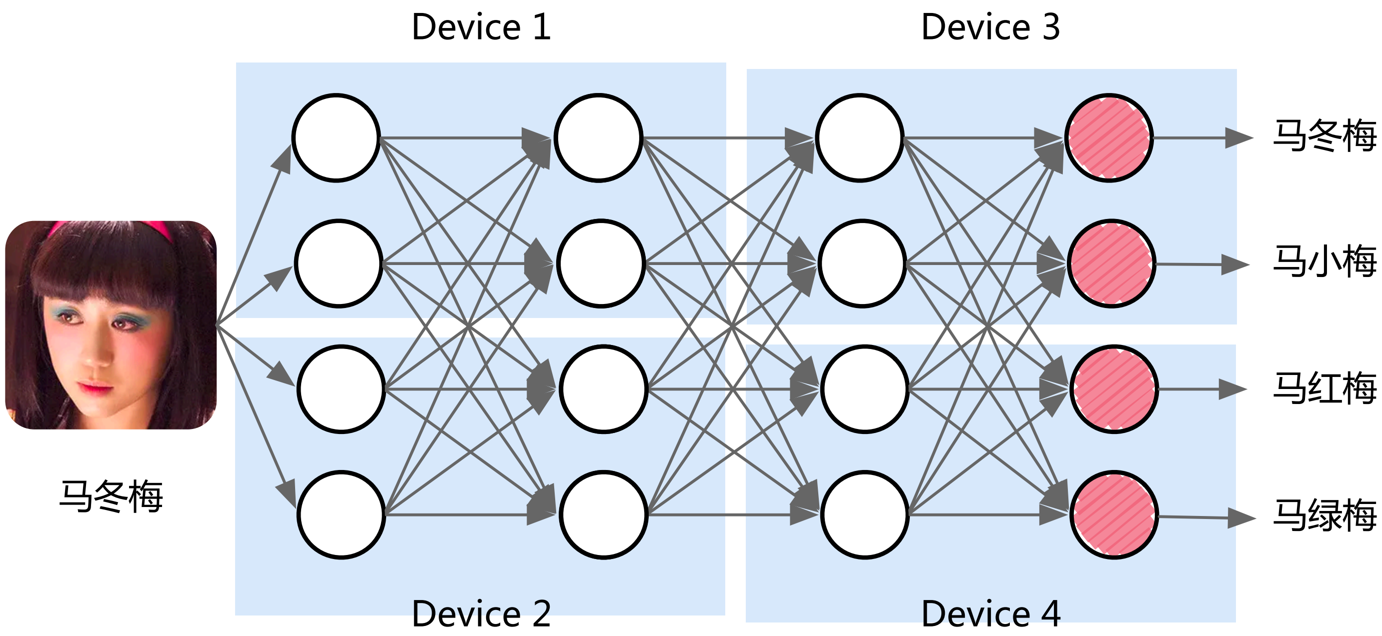
Data parallelism 数据并行

1. Data parallelism, DP
2. Distribution Data Parallel, DDP
3. Fully Sharded Data Parallel, FSDP



MP: Tensor parallelism 张量并行

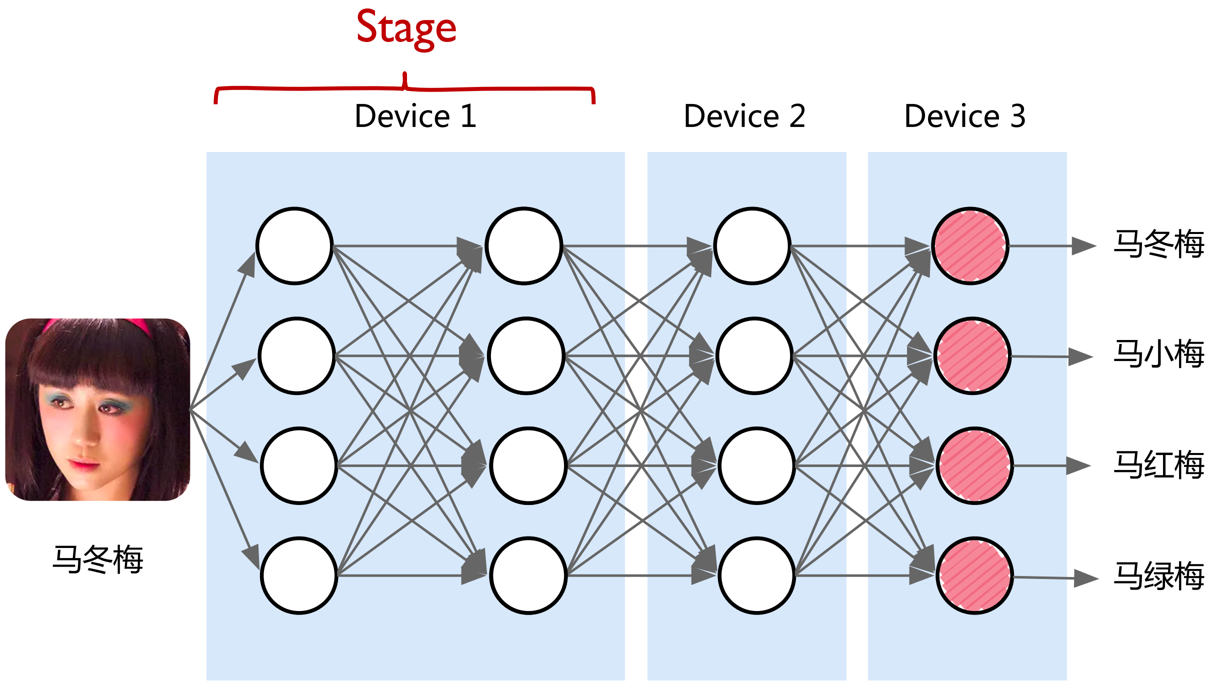
- Divide parameters in the layer into different devices, which we called tensor model parallelism
- 张量并行：将计算图中的层内的参数切分到不同设备，即层内并行



张量并行（层内并行）

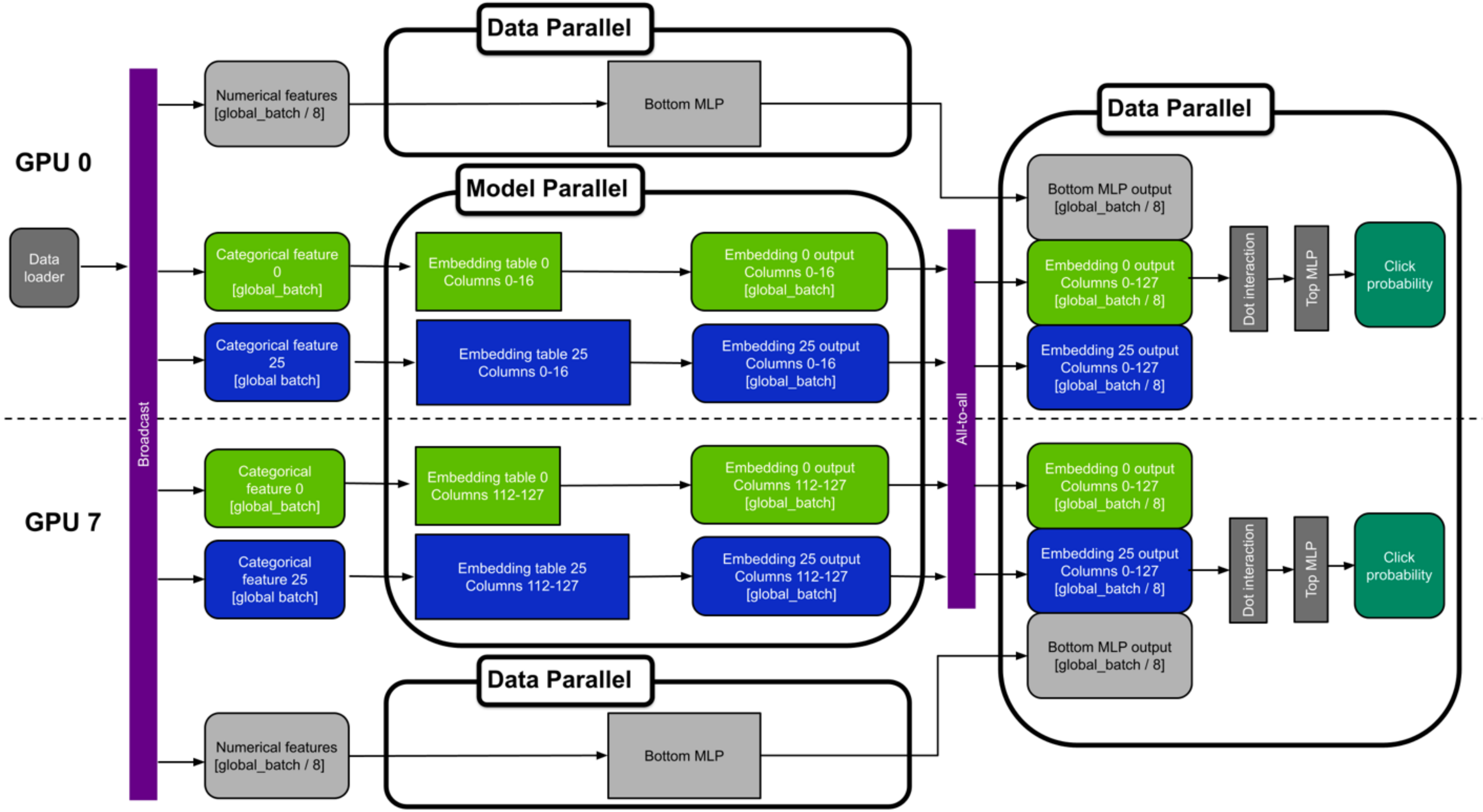
MP: Pipeline parallelism 流水线并行

- Model divided layers into different devices, which we called pipeline parallelism
- 流水线并行：按模型layer层切分到不同设备，即层间并行



流水线并行（层间并行）

混合并行：DLRM 推荐大模型



Naumov, Maxim, et al. "Deep learning recommendation model for personalization and recommendation systems." arXiv preprint arXiv:1906.00091 (2019).

混合并行：Megatron-LM 语言大模型

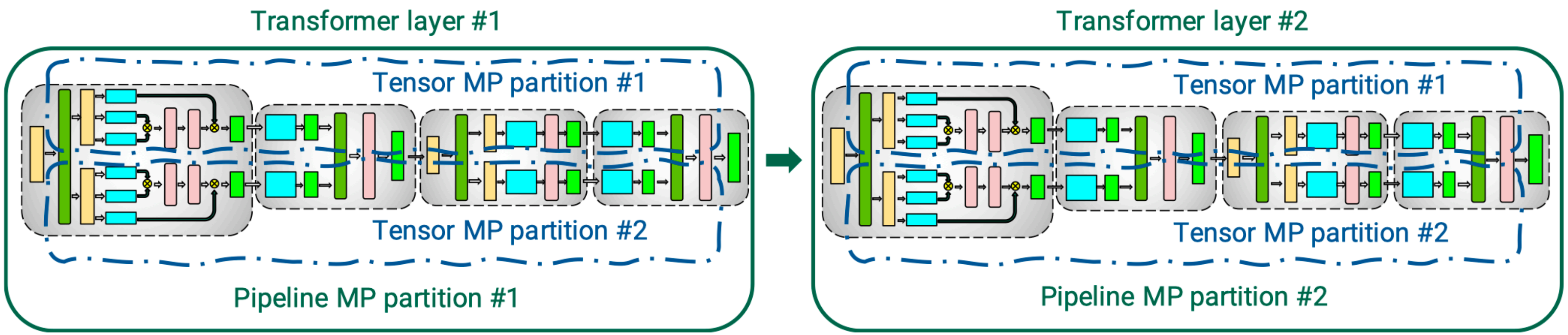
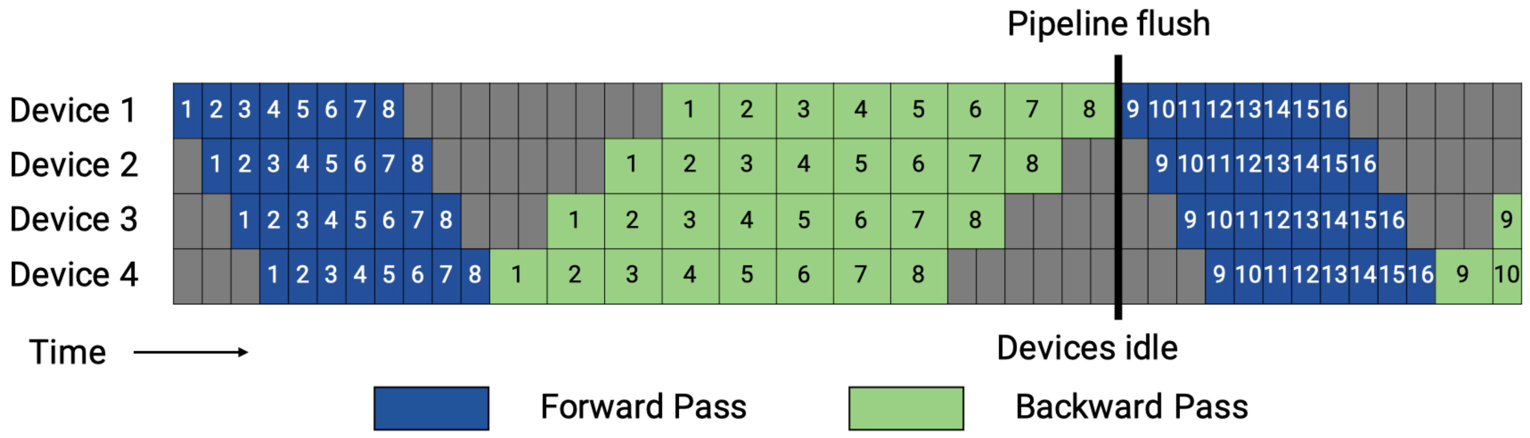


Figure 2: Combination of tensor and pipeline model parallelism (MP) used in this work for transformer-based models.



Narayanan, Deepak, et al. "Efficient large-scale language model training on gpu clusters using megatron-lm." Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. 2021.

BUILDING A BETTER CONNECTED WORLD

THANK YOU