分布式训练系列

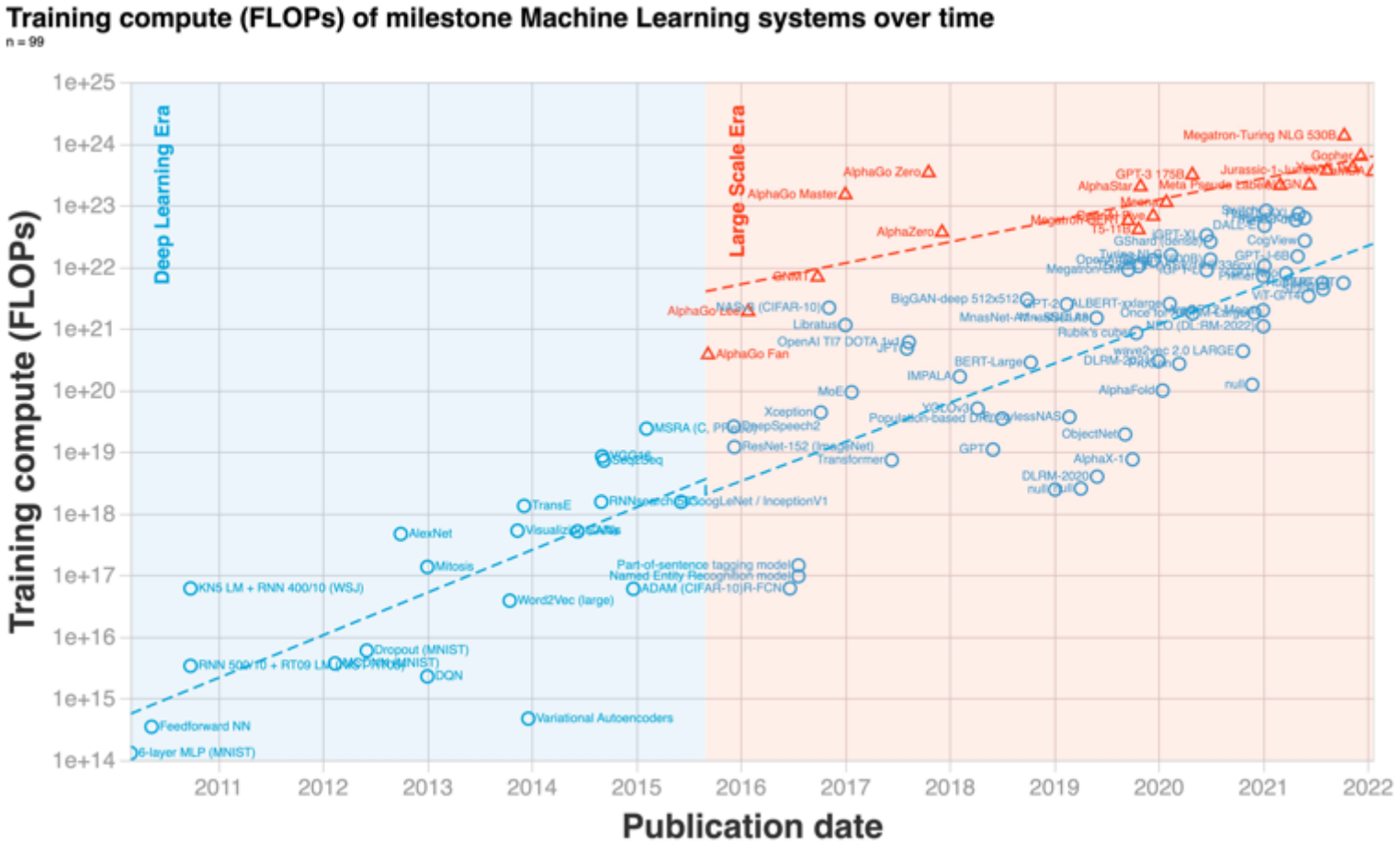
混合并行

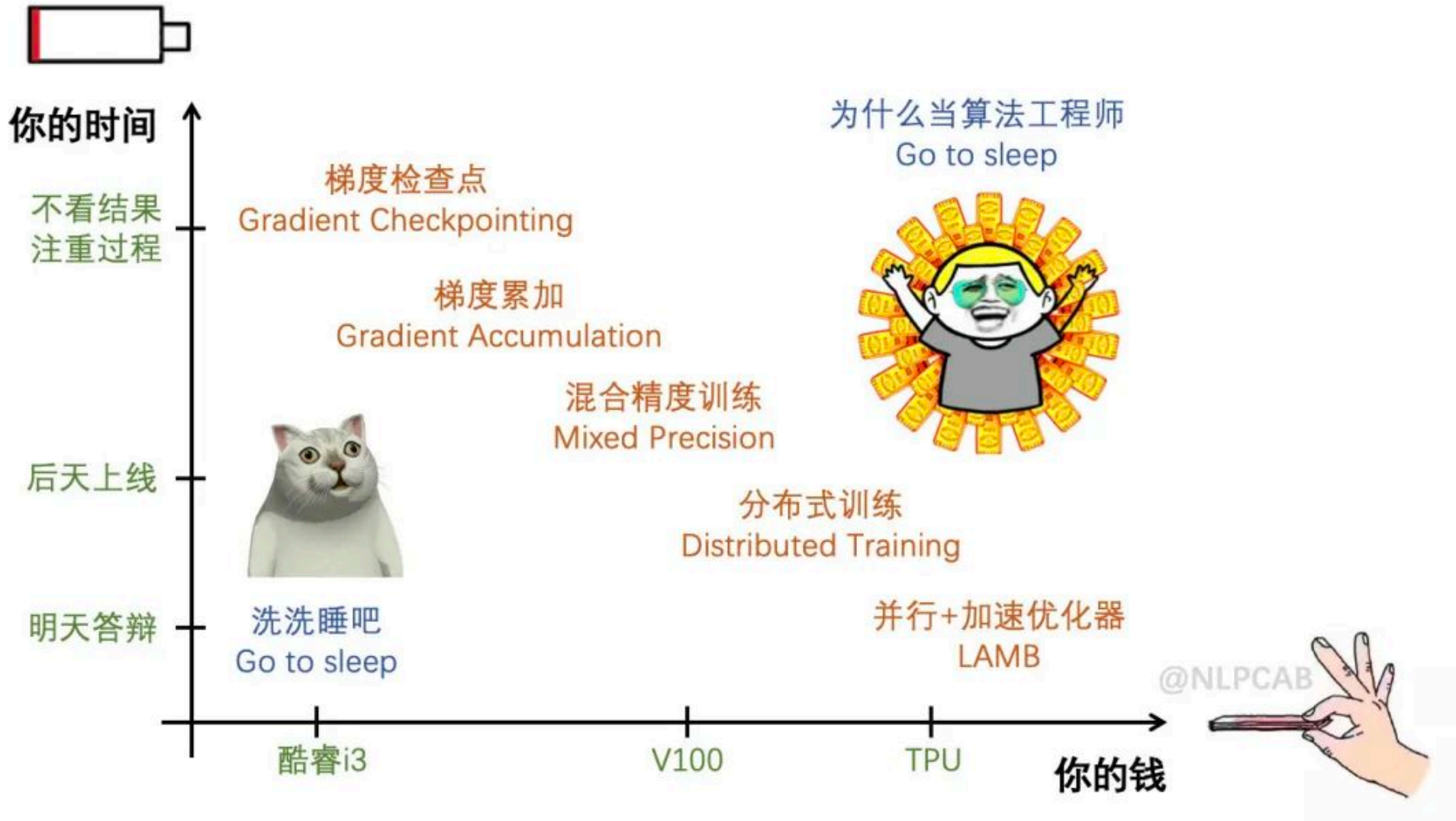
ZOMI

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. 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
3. 解决模型碎片化，提供预训练方案

- **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





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：**推荐大模型 – LLM大模型
- 内存和计算优化
- 分布式并行总结

DLRM 推荐大模型

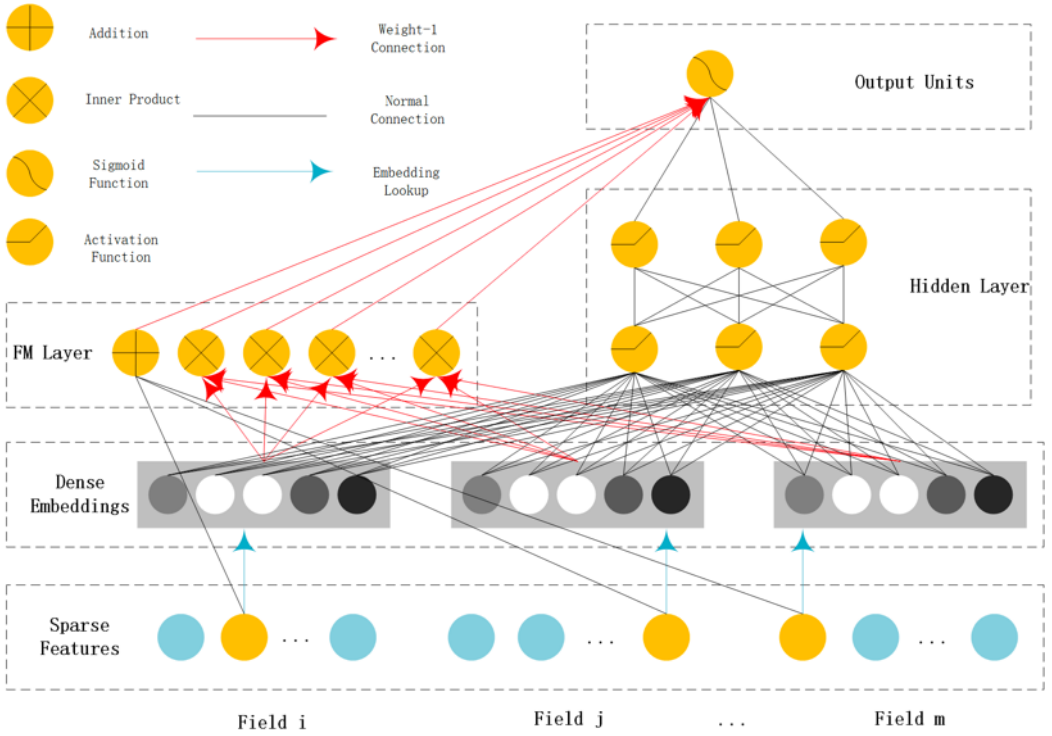
- CTR，Click-Through-Rate，点击率预估模型

Continuous features:

1. 用户本身特征：用户的年龄、性别等；
2. 用户行为特征：点击/购买过的物品等；
3. 上下文特征：用户登录设备、当前时间等；

Categorical features 待排序物品特征：

1. 物品 ID；
2. 物品商品信息；
3. 物品被点击次数；
4. 物品点击率等；



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

DLRM 推荐大模型

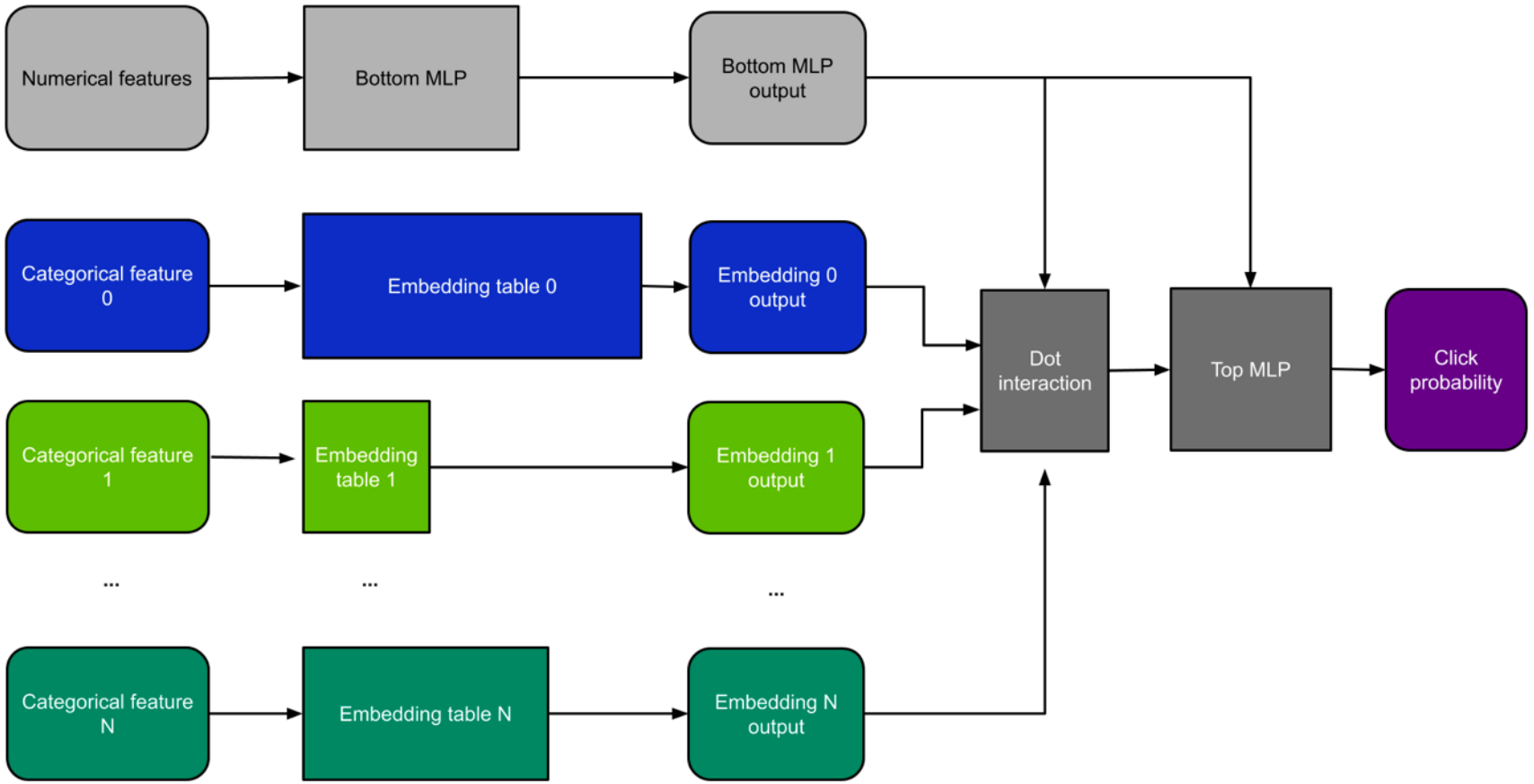
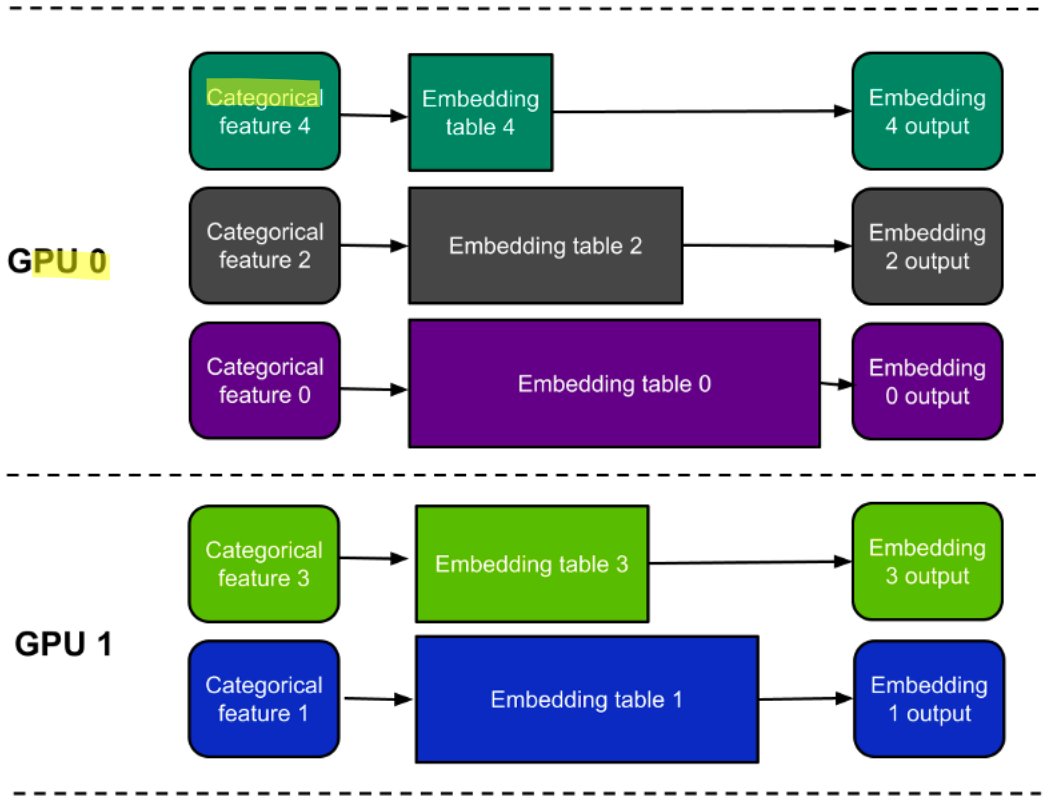


Diagram of DLRM architecture.

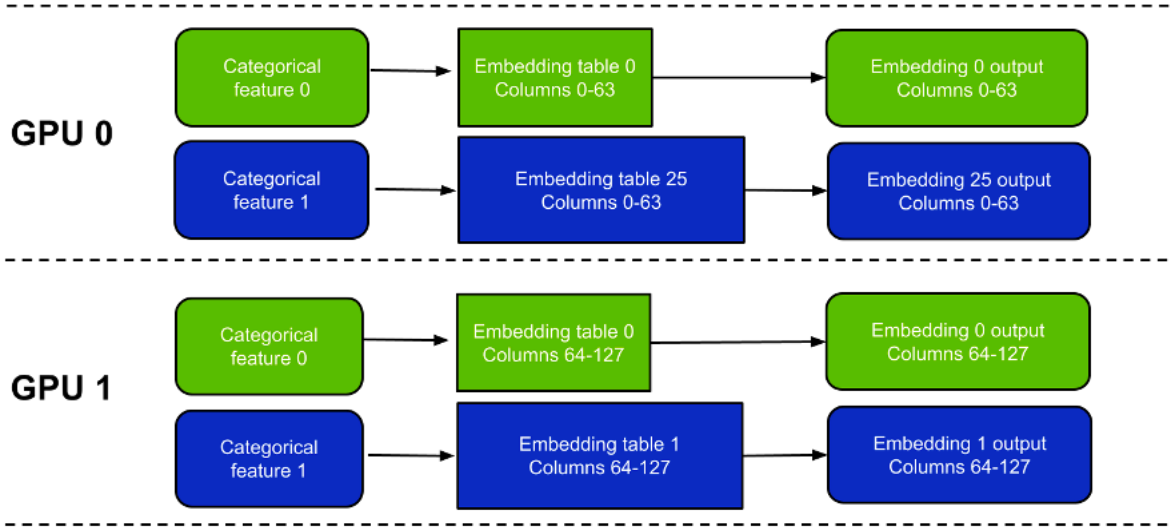
Naumov, Maxim, et al. "Deep learning recommendation model for personalization and recommendation systems." arXiv preprint arXiv:1906.00091 (2019).

DLRM 推荐大模型

Table-wise split mode is when each GPU stores a subset of all the embedding tables

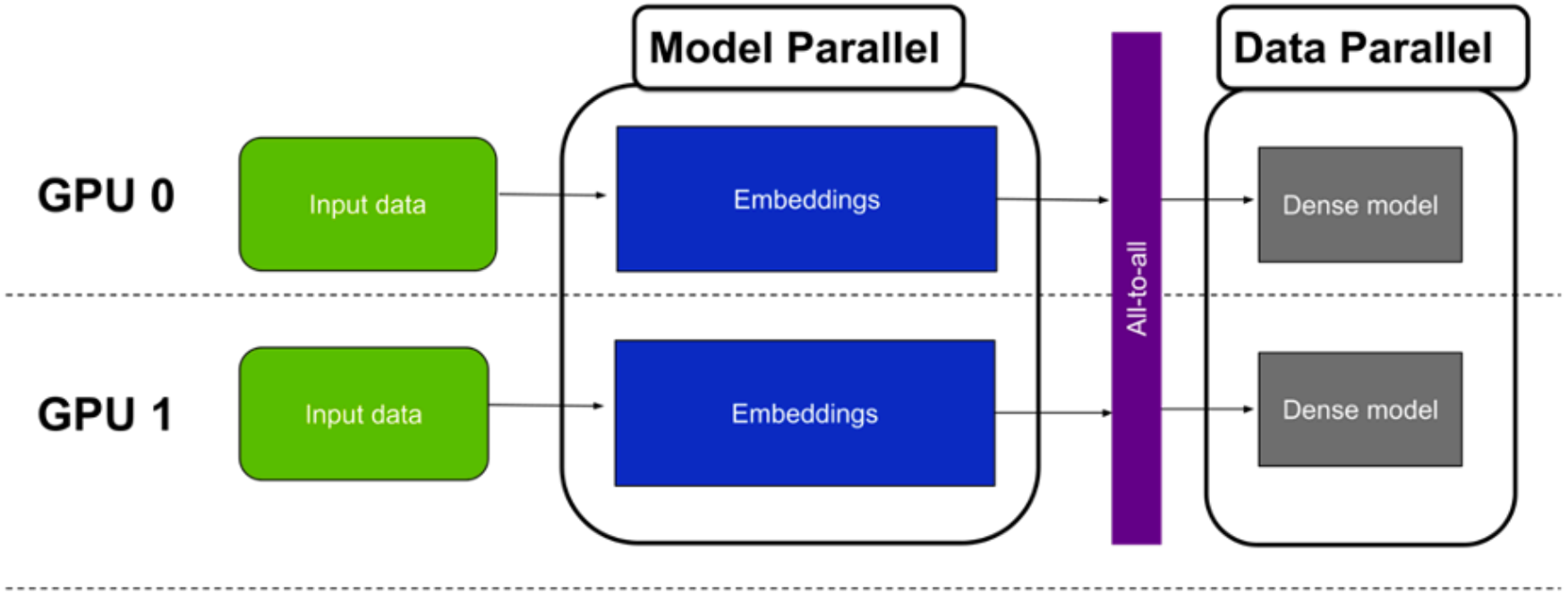


Column-wise split mode is when each device stores a subset of columns from every embedding table



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

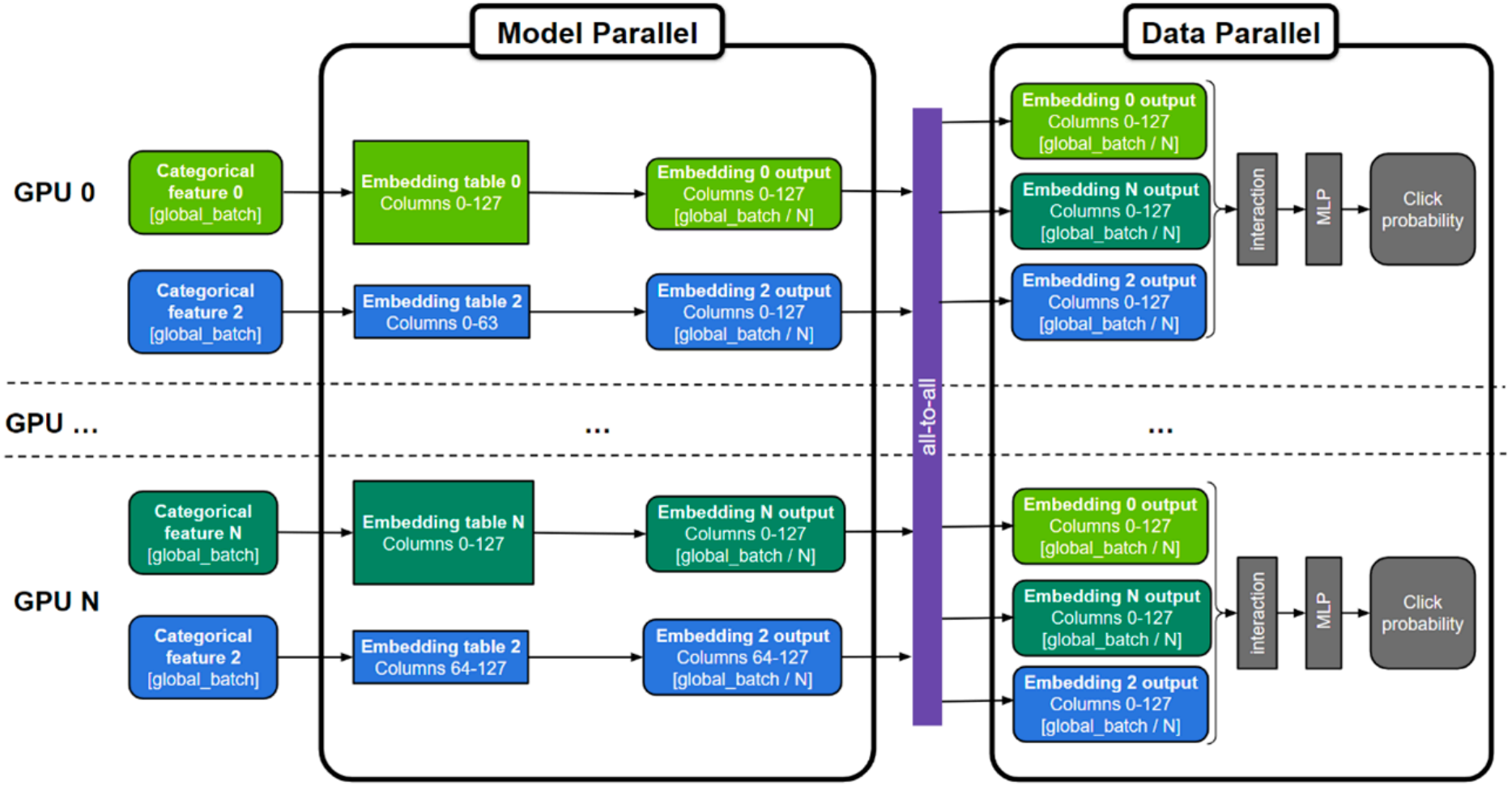
DLRM 推荐大模型



General hybrid-parallel approach for training large recommender systems

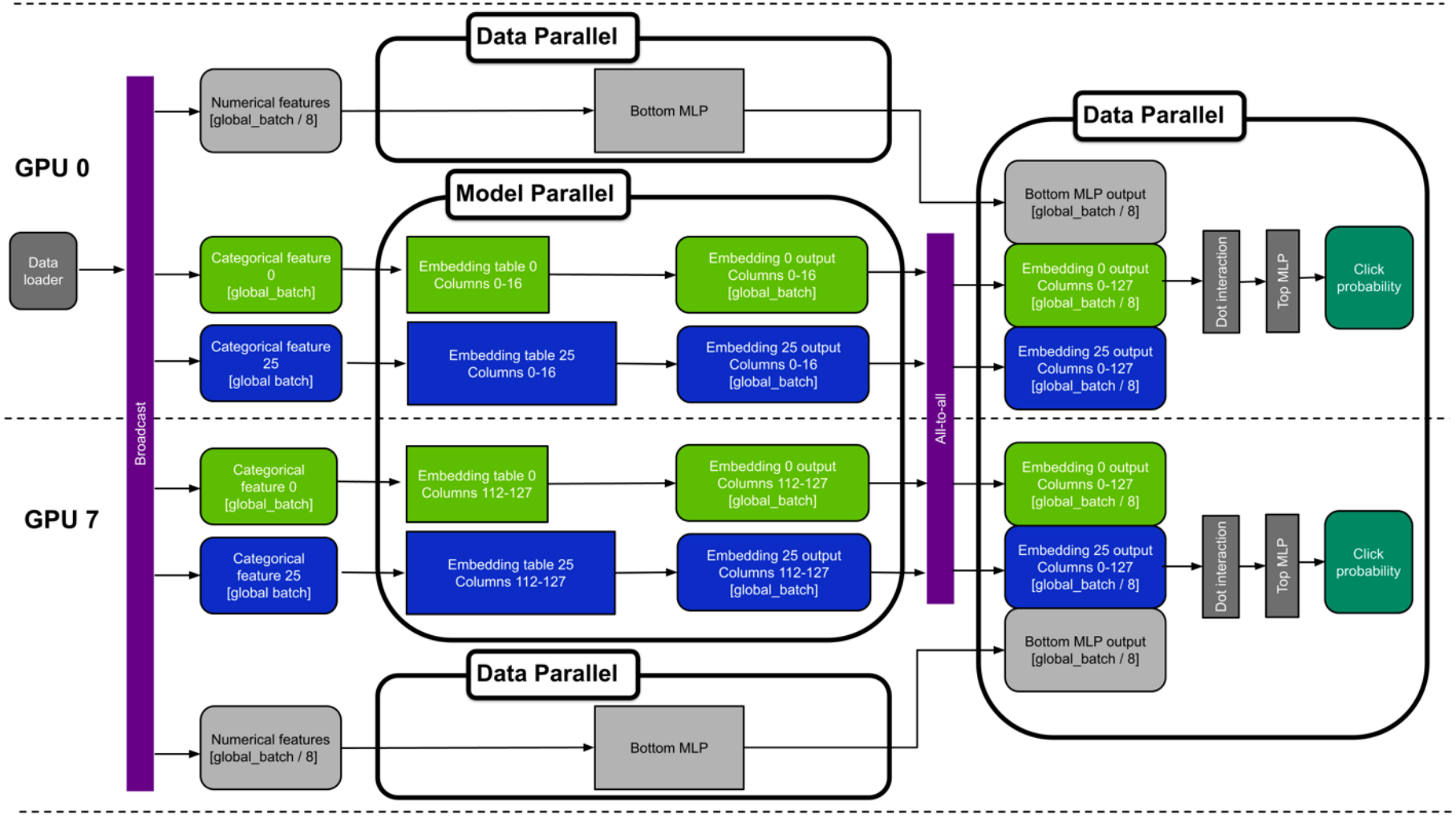
Naumov, Maxim, et al. "Deep learning recommendation model for personalization and recommendation systems." arXiv preprint arXiv:1906.00091 (2019).

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).

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).

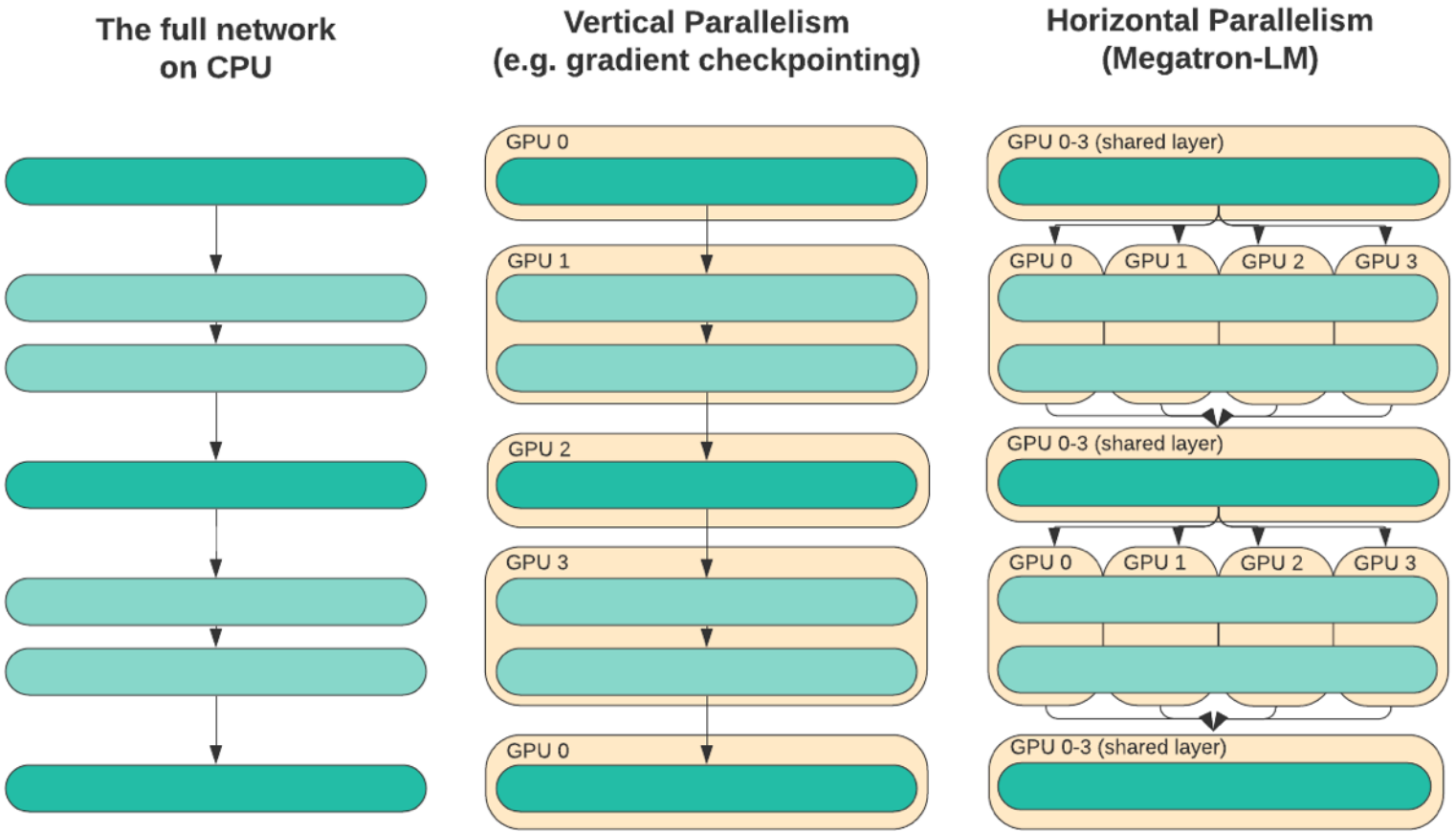
DLRM 推荐大模型

- Comparison of CPU and GPU training throughput for a 113-billion parameter Deep Learning Recommendation Model (DLRM).

Hardware	Throughput [samples/second]	Speedup over CPU
2xAMD EPYC 7742	17.7k	1x
A100-80GB + 2xAMD EPYC 7742(large embeddings on CPU)	768k	43x
DGX A100 (8xA100-80GB) (hybrid parallel)	11.9M	672x

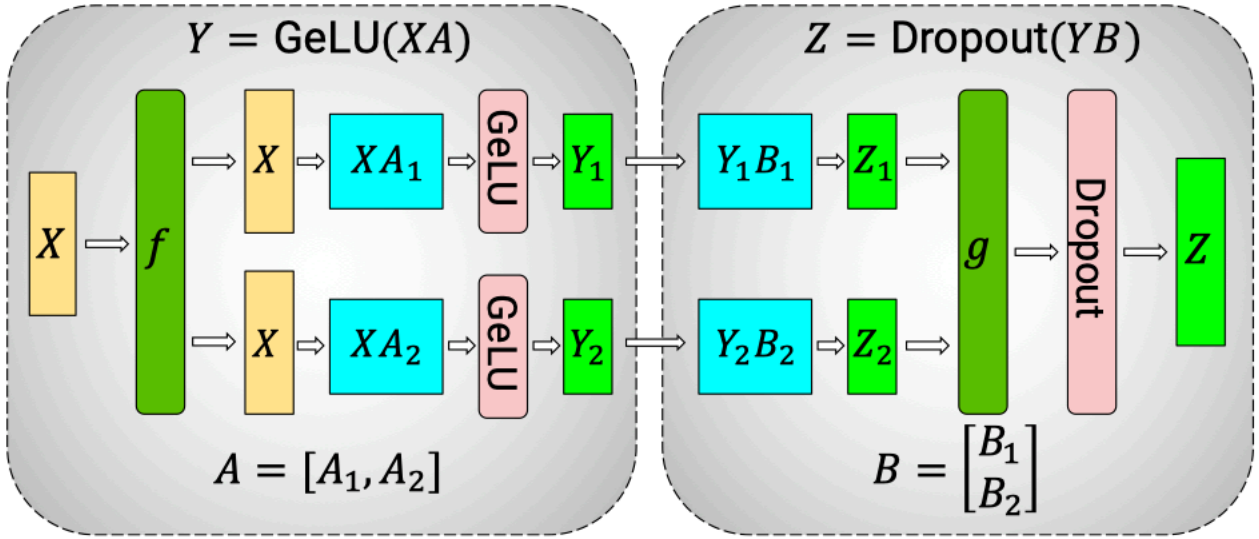
Naumov, Maxim, et al. "Deep learning recommendation model for personalization and recommendation systems." arXiv preprint arXiv:1906.00091 (2019).

Megatron-LM 语言大模型



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.

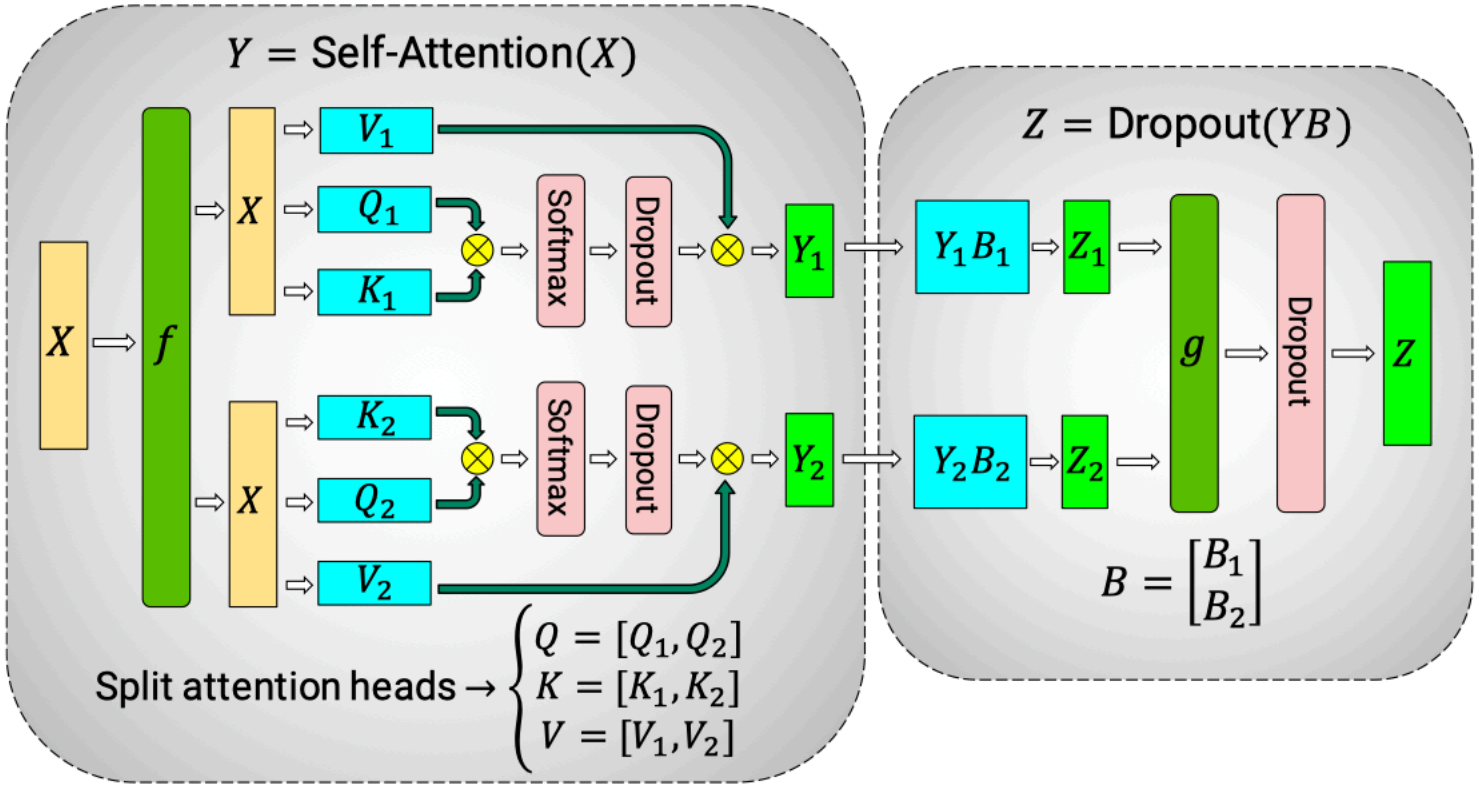
Megatron-LM 语言大模型



(a) MLP.

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.

Megatron-LM 语言大模型



(b) Self-Attention.

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.

Megatron-LM 语言大模型

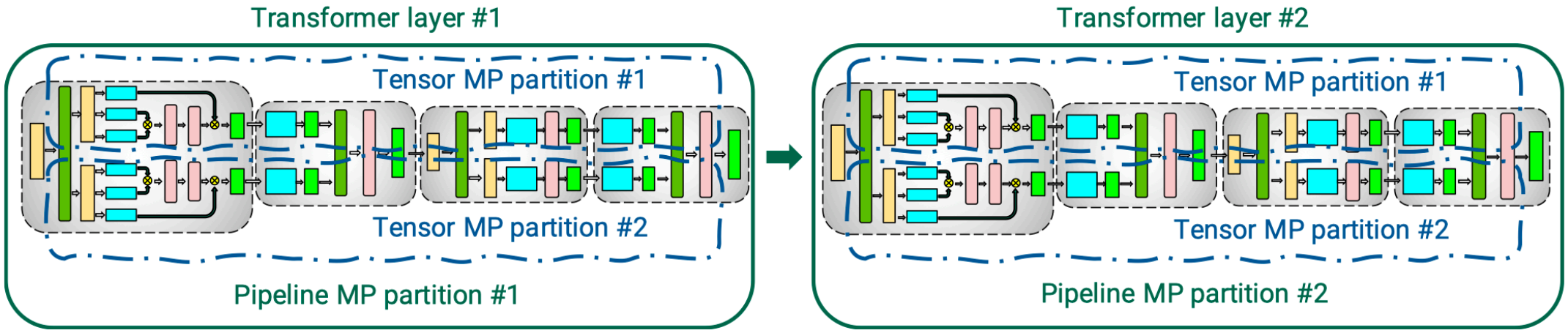
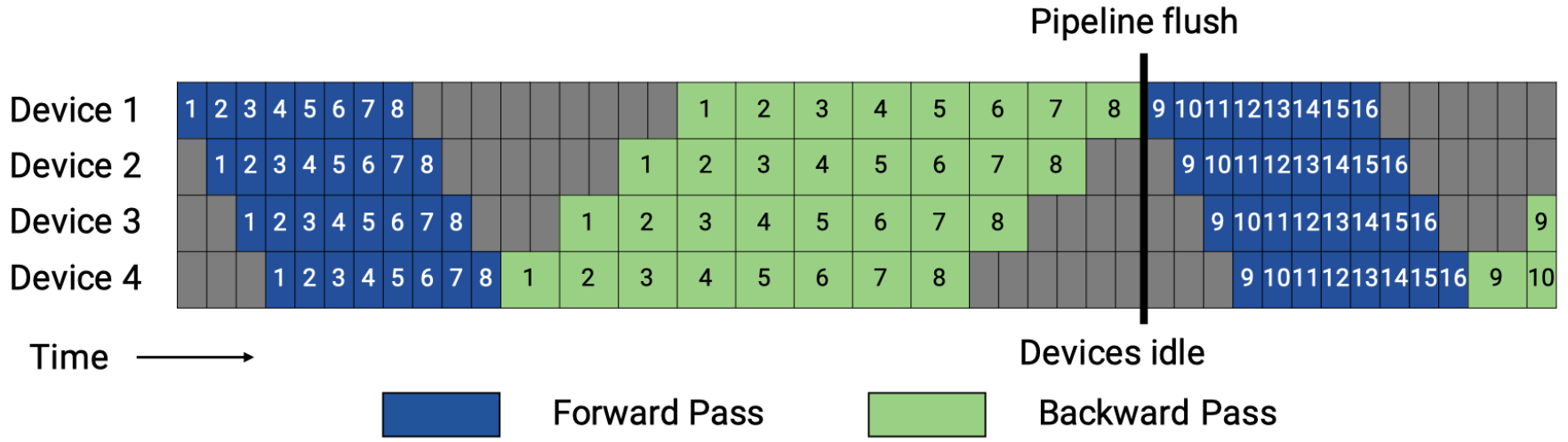
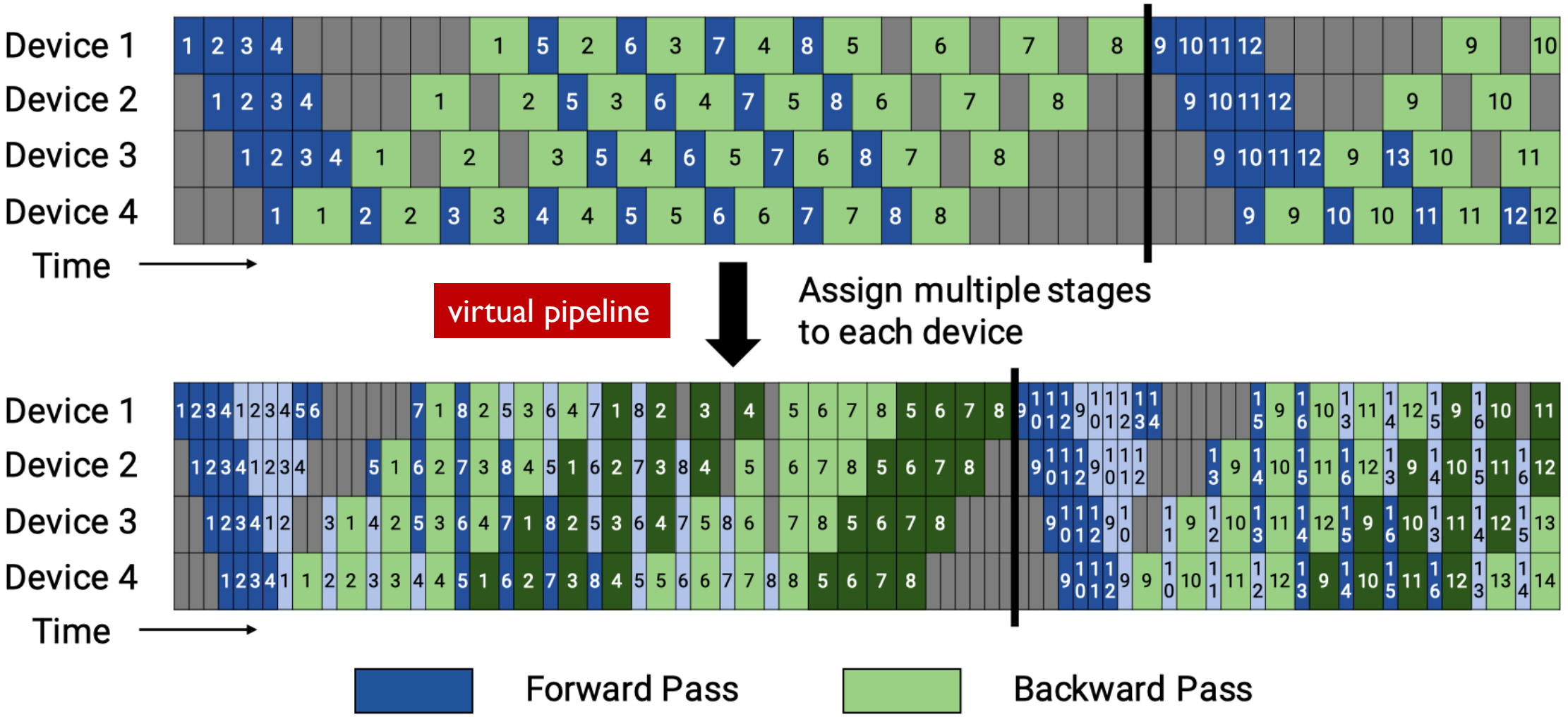


Figure 2: Combination of tensor and pipeline model parallelism (MP) used in this work for transformer-based models.



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.

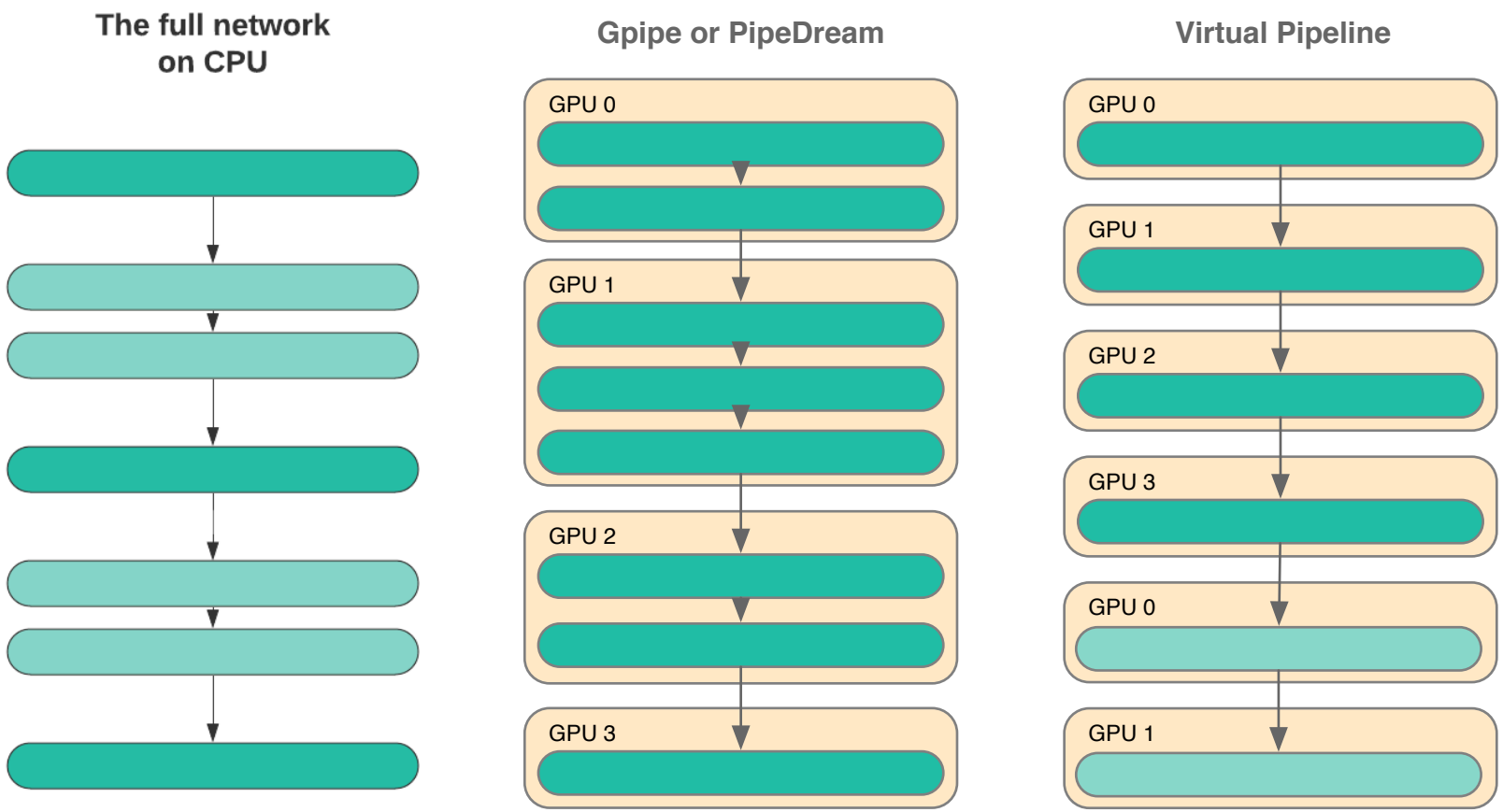
Megatron-LM 语言大模型



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.

Megatron-LM 语言大模型

在 device 数量不变的情况下，分出更多的 pipeline stage，以更多的通信量，换取空泡比率降低



Megatron-LM 语言大模型

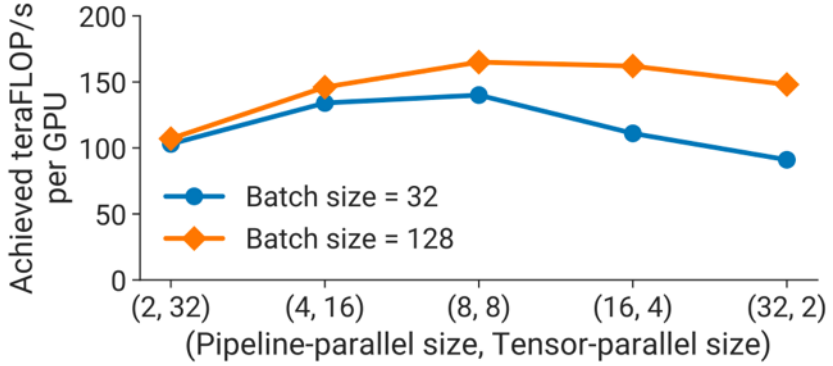


Figure 13: Throughput per GPU of various parallel configurations that combine pipeline and tensor model parallelism using a GPT model with 162.2 billion parameters and 64 A100 GPUs.

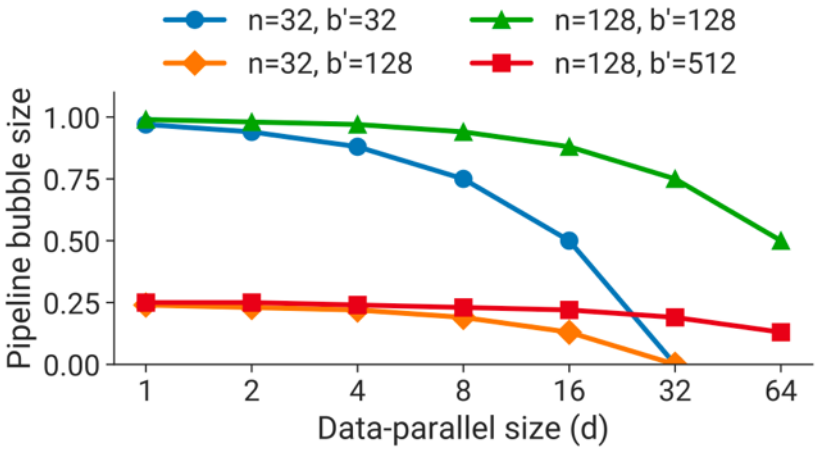


Figure 6: Fraction of time spent idling due to pipeline flush (pipeline bubble size) versus data-parallel size (d), for different numbers of GPUs (n) and ratio of batch size to microbatch size ($b' = B/b$).

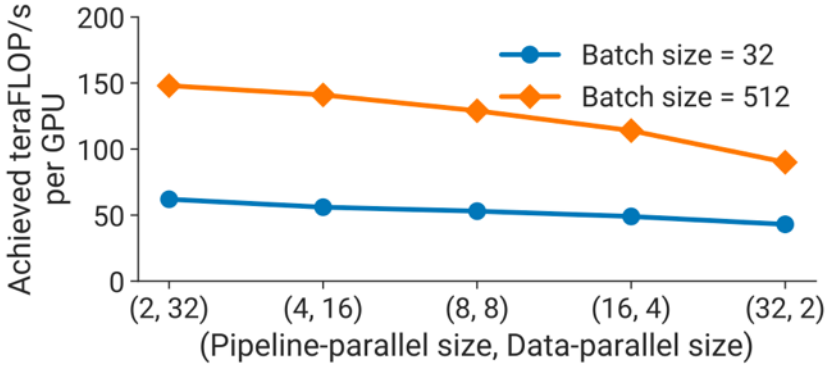


Figure 14: Throughput per GPU of various parallel configurations that combine data and pipeline model parallelism using a GPT model with 5.9 billion parameters, three different batch sizes, microbatch size of 1, and 64 A100 GPUs.

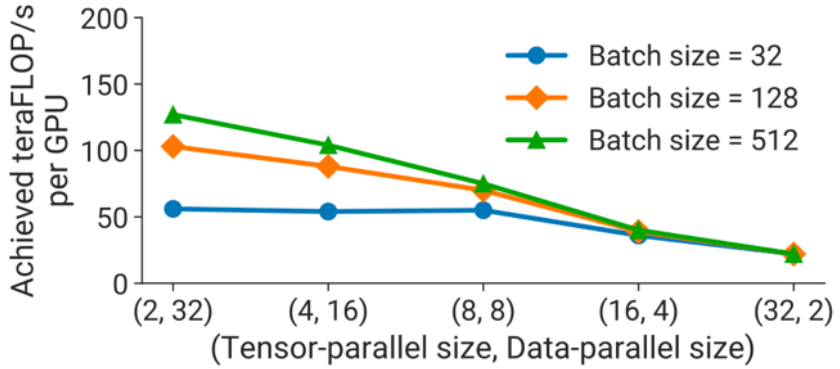


Figure 15: Throughput per GPU of various parallel configurations that combine data and tensor model parallelism using a GPT model with 5.9 billion parameters, three different batch sizes, microbatch size of 1, and 64 A100 GPUs.

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.

Inference

I. https://zhuanlan.zhihu.com/p/450854172 全网最全-超大模型+分布式训练架构和经典论文

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

III. 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.

IV. https://developer.nvidia.com/blog/training-a-recommender-system-on-dgx-a100-with-100b-parameters-in-tensorflow-2/

V. https://developer.nvidia.com/blog/fast-terabyte-scale-recommender-training-made-easy-with-nvidia-merlin-distributed-embeddings/