大模型（LLMs）分布式训练面

1. 理论篇

1.1 想要训练1个LLM，如果只想用1张显卡，那么对显卡的要求是什么?

显卡显存足够大，nB模型微调一般最好准备**20nGB以上的显存**。

1.2 如果有N张显存足够大的显卡，怎么加速训练?

数据并行（DP），充分利用多张显卡的算力。

1.3 如果显卡的显存不够装下一个完整的模型呢?

最直观想法，需要**分层加载，把不同的层加载到不同的GPU上**（accelerate的device_map）

也就是常见的PP，流水线并行。

1.4 PP推理时，是一个串行的过程，1个GPU计算，其他空闲，有没有其他方式?

1. **横向切分：流水线并行（PP）**，也就是分层加载到不同的显卡上。
2. **纵向切分：张量并行（TP）**，在 DeepSpeed 世界里叫模型并行（MP）

1.5 3种并行方式可以叠加吗?

是可以的，**DP+TP+PP**，这就是3D并行。**如果真有1个超大模型需要预训练，3D并行那是必不可少的。**毕竟显卡进化的比较慢，最大显存的也就是A100 80g。

单卡80g，可以完整加载小于40B的模型，但是训练时+梯度+优化器状态，5B模型就是上限了，更别说activation的参数也要占显存，batch size还得大。而现在100亿以下（10B以下）的LLM只能叫small LLM。

1.6 Colossal-AI 有1D/2D/2.5D/3D，是什么情况?

Colossal-AI 的nD是针对张量并行，指的是TP的切分，对于矩阵各种切，和3D并行不是一回事。

1.7 除了3D并行有没有其他方式大规模训练?

可以使用更优化的数据并行算法FSDP（类似ZeRO3）或者直接使用 DeepSpeed ZeRO 。

1.8 有了ZeRO系列，为什么还需要3D并行?

根据ZeRO论文，尽管张量并行的显存更省一点，张量并行的通信量实在太高，只能限于节点内（有NVLINK）。如果节点间张量并行，显卡的利用率会低到5%

但是，根据Megatron-LM2的论文，当显卡数量增加到千量级，ZeRO3是明显不如3D并行的。

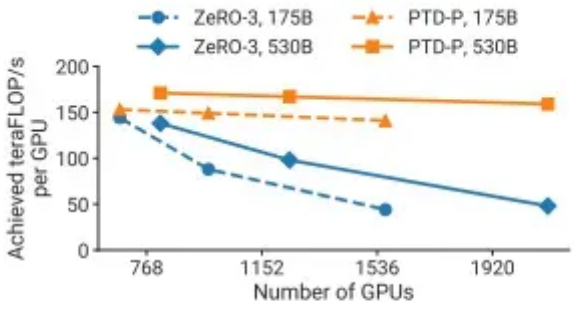

Figure 10: Throughput per GPU of PTD-P and ZeRO-3 for two different GPT models (the 175B GPT-3 model is shown with dotted lines, and the 530B model is shown with solid lines). Global batch sizes are fixed and ZeRO-3 is used without any model parallelism.

1.9 平民适不适合玩3D并行?

不适合。

3D并行的基础是，节点内显卡间NVLINK超高速连接才能上TP。有没有NVLINK都是个问题。

而且，节点间特殊的网络通常有400Gb/s？远超普通IDC内的万兆网络10Gb/s。

1.10 平民适不适合直接上多机多卡的ZeRO3（万兆网）?

不适合。

想象一下，当65B模型用Zero3，每一个step的每一张卡上需要的通信量是195GB（3倍参数量），也就是1560Gb。万兆网下每步也要156s的通信时间，这画面太美。

2. 实践篇

2.1 假如有超多的8卡A100节点（DGX A100），如何应用3D并行策略?

1. 首先，**张量并行**。3种并行方式里，张量并行（TP）对于GPU之间的通信要求最高，而节点内有NVLINK通信速度可以达到600GB/s。
2. 其次，**流水线并行**，每个节点负责一部分层，每35个节点组成一路完整的流水线，也就是一个完整的模型副本，这里一个模型副本需280卡。

3. 最后，**数据并行**，官方也做了8路，10路，12路的并行实验，分别使用280个节点，350个节点和420个节点。

参考 Megatron-Turing NLG 530B

集群规模越大，单个GPU利用率越低。

2.2 如果想构这样一个大规模并行训练系统，训练框架如何选?

可以参考Megatron-Turing NLG 530B，NVIDIA Megatron-LM + Microsoft DeepSpeed

BLOOM **则是PP+DP用DeepSpeed，TP用Megatron-LM**

当然还有一些其他的训练框架，在超大规模下或许也能work。

2.3 训练框架如何选?

下面这个图是bloom的一个实验，DP/TP/PP都能降显存，核心是要降到单卡峰值80g以下。

真大模型就是要TP=8，充分利用NVLINK，然后优先PP，最后DP。

GPUs	Size	DP	TP	PP	MBS	Mem	TFLOPs	Notes
8	20B	1	8	1	1	68GB	107.48	02-17
80	200B	1	8	10	1	75GB	97.82	02-17
160	200B	2	8	10	1	53GB	96.19	02-17

然而假大模型（7B）比如LLaMA-7B，可以不用3D并行，直接用DeepSpeed ZeRO更方便，参考open-llama项目。

3. 并行化策略选择篇

3.1 单GPU

1. 显存够用： 直接用
2. 显存不够：上offload，用cpu

3.2 单节点多卡

1. 显存够用（模型能装进单卡）：DDP或ZeRO
2. 显存不够：TP或者ZeRO或者PP

重点：没有NVLINK或者NVSwitch，也就是穷人模式，要用PP

3.3 多节点多卡

如果节点间通信速度快（穷人的万兆网肯定不算）

ZeRO或者3D并行，其中3D并行通信量少但是对模型改动大。

如果节点间通信慢，但显存又少。

DP+PP+TP+ZeRO-1

4. 问题篇

4.1 推理速度验证

ChatGML在V100单卡的推理耗时大约高出A800单卡推理的40%。

ChatGML推理耗时和问题输出答案的字数关系比较大，答案字数500字以内，A800上大概是每100字，耗时1秒，V100上大概是每100字，耗时1.4秒。

1. ChatGML在A800单卡推理耗时统计

问题	运行次数	平均答案长度	平均耗时
给我介绍一下苹果公司,50个字	5	146.6	1.9458990097045898
给我介绍一下微软公司,50个字	5	104.6	1.2978283882141113
给我介绍一下苹果公司,100个字	5	165.4	2.0990972995758055
给我介绍一下微软公司,100个字	5	154.4	1.9092102527618409
给我介绍一下苹果公司,200个字	5	168.4	2.1302066326141356
给我介绍一下微软公司,200个字	5	208.8	2.6089860916137697
给我介绍一下苹果公司,300个字	5	443.4	5.4034130573272705
给我介绍一下微软公司,300个字	5	484.2	5.980589342
给我介绍一下苹果公司,500个字	5	525.4	6.246328496932984
给我介绍一下微软公司,500个字	5	591.6	6.961390399932862

1. ChatGML在V100单卡推理耗时统计

问题	运行次数	平均答案长度	平均耗时
给我介绍一下苹果公司,50个字	5	146.6	1.9458990097045898
给我介绍一下微软公司,50个字	5	104.6	1.2978283882141113
给我介绍一下苹果公司,100个字	5	165.4	2.0990972995758055
给我介绍一下微软公司,100个字	5	154.4	1.9092102527618409
给我介绍一下苹果公司,200个字	5	168.4	2.1302066326141356
给我介绍一下微软公司,200个字	5	208.8	2.6089860916137697
给我介绍一下苹果公司,300个字	5	443.4	5.4034130573272705
给我介绍一下微软公司,300个字	5	484.2	5.980589342
给我介绍一下苹果公司,500个字	5	525.4	6.246328496932984
给我介绍一下微软公司,500个字	5	591.6	6.961390399932862

1. 结论:
2. 训练效率方面: 多机多卡训练，增加训练机器可以线性缩短训练时间。
3. 推理性能方面:
4. ChatGML在V100单卡的推理耗时大约高出A800单卡推理的40%。
5. ChatGML推理耗时和问题输出答案的字数关系比较大，答案字数500字以内，A800上大概是每100字，耗时1秒，V100上大概是每100字，耗时1.4秒。

4.2 并行化训练加速

可采用deepspeed进行训练加速，目前行业开源的大模型很多都是采用的基于deepspeed框架加速来进行模型训练的。如何进行deepspeed训练，可以参考基于deepspeed构建大模型分布式训练平台。

deepspeed在深度学习模型软件体系架构中所处的位置:

DL model—>train opitimization(deepspeed)—>train framework —> train instruction (cloud)—>GPU device

当然需要对比验证deepspeed 的不同参数，选择合适的参数。分别对比stage 2,3进行验证，在GPU显存够的情况下，最终使用stage 2。

4.3 deepspeed 训练过程，报找不主机

解决方法：deepspeed的关联的多机的配置文件，Hostfile 配置中使用ip，不使用hostname

4.4 为什么 多机训练效率不如单机?

多机训练可以跑起来，但是在多机上模型训练的速度比单机上还慢。

通过查看服务器相关监控，发现是网络带宽打满，上不去了，其他系统监控基本正常。原理初始的多机之间的网络带宽是64Gps，后面把多机之间的网络带宽调整为800Gps，问题解决。

实验验证，多机训练的效率，和使用的机器数成线性关系，每台机器的配置一样，如一台GPU机器跑一个epoch需要2小时，4台GPU机器跑一个epoch需要半小时。除了训练速度符合需求，多机训练模型的loss下降趋势和单机模型训练的趋势基本一致，也符合预期。

4.5 多机训练不通，DeepSPeed配置问题

多机间NCCL 不能打通

1. 解决方法：

新建 .deepspeed_env 文件，写入如下内容

NCCL_IB_DISABLE=1

NCCL_DEBUG=INFO

NCCL_SOCKET_IFNAME=eth0

NCCL_P2P_DISABLE=1