AI 大底座，大模型时代的答卷

作者：百度智能云开发者中心
2023.05.09 18:35 浏览量：94

简介：2023 年 3 月，文心一言在这个高性能集群上诞生，并不断迭代出新的能力。

1. 文心一言的诞生

“文心一言就是在这个全国 AI 领域规模最大的高性能 GPU 集群上完成训练的。”

早在 2021 年 6 月，为了满足未来的大模型训练任务，百度智能云开始规划全新的高性能 GPU 集群的建设，联合 NVIDIA 共同完成了可以容纳万卡以上规模的 IB 网络架构设计，集群中节点间的每张 GPU 卡都通过 IB 网络连接， 并在 2022 年 4 月将集群建设完成，提供单集群 EFLOPS 级别的算力。

2023 年 3 月，文心一言在这个高性能集群上诞生，并不断迭代出新的能力。目前，这个集群的规模还在不断扩大。

NVIDIA 中国区解决方案与工程总经理赖俊杰博士：高速 IB 网络互联的 GPU 集群是大模型时代的关键基础设施。NVIDIA 和百度智能云共同建成的这个国内云计算市场最大规模的高性能 GPU/IB 集群，将加速百度在大模型领域获得更大突破。

2. 高性能集群设计

高性能集群并不是算力的简单堆积，还需要经过专门的设计和优化，才能发挥出集群的整体算力。

在分布式训练中 GPU 会在机间和机内不断地进行通信。在利用 IB 、RoCE 等高性能网络为机间通信提供高吞吐、低时延的服务同时，还需要对服务器的内部网络连接，以及集群网络中的通信拓扑进行专门设计，满足大模型训练对通信的要求。

做到极致的设计优化，需要对 AI 任务中的各项操作都对基础设施意味着什么有深刻理解。分布式训练中不同的并行策略，即模型、数据、参数如何进行拆分，会产生不同的数据通信需求，比如数据并行和模型并行会分别引入大量的机内和机间 Allreduce 操作，专家并行会产生机间 All2All 操作，4D 混合并行则会将各类并行策略产生的通信操作都引入。

为此，百度智能云从单机服务器和集群网络两个方面优化设计，构建高性能 GPU 集群。

在单机服务器方面，百度智能云的超级 AI 计算机 X-MAN，目前已经进化到第 4 代。X-MAN 4.0 为 GPU 建立起了高性能的卡间通信，提供单机内部 134 GB/s 的 Allreduce 带宽。这是目前百度定制化程度最高，专用物料最多的服务器产品。在 MLCommons 1.1 榜单中，X-MAN 4.0 在同配置单机硬件性能名列 TOP2。

在集群网络方面，专门设计了面向大模型训练优化过的三层 Clos 架构，确保在大规模训练时集群的性能和加速比。和传统方式相比，该架构经过八导轨的优化，让任一同号卡在不同机器中的通信中的跳步数尽可能少，为 AI 训练中网络流量占比最大的同号卡 Allreduce 操作提供高吞吐和低延时的网络服务。

该网络架构可以最大能支持到 16000 卡的超大规模集群，这个规模是现阶段全 IB 网络盒式组网的最大规模。该集群的网络性能稳定一致性能做到了 98% 的水平，接近一直在稳定通信的状态。经大模型算法团队验证，在此超大规模集群上提交千亿模型训练作业，同等机器规模下整体训练效率是上一代集群的 3.87 倍。

但是，建设大规模高性能异构集群，只是大模型成功落地的第一步。确保 AI 大模型训练任务的顺利完成，还需要更多系统性软硬一体的优化。

3. 大模型训练的挑战

过去几年，大模型的参数规模将达到每年增长 10 倍的速度。2020 年左右，亿级别参数才是大模型，2022 年，已经是需要千亿参数规模才能叫大模型了。

在大模型之前，一个 AI 模型的训练，通常单机单卡、或者单机多卡就可以满足，训练周期在小时到数天之间。现在，为了完成千亿参数大模型的训练，几百台服务器、数千张 GPU/XPU 卡的大集群分布式训练成为必选项，训练周期也扩展到以月为单位。

为了训练 1750 亿参数的 GPT-3 （3000 亿 token 数据），1 块 A100 按半精度峰值计算性能折算需要 32 年， 1024 块 A100 按资源利用率 45% 计算需要 34 天时间。当然，即使不考虑时间问题，1 块 A100 也是无法训练千亿参数规模的模型的，因为模型参数已经超过单卡显存容量。

在分布式训练的环境下进行大模型训练，训练周期从单卡几十年缩短到几十天，需要突破计算墙、显存墙、通信墙等各种挑战，使得集群内的所有资源都能被充分利用，加速训练过程，缩短训练周期。

计算墙，指的是单卡算力和模型总算力之间的巨大差异。A100 的单卡算力只有 312 TFLOPS，而 GPT-3 则需要 314 ZFLOPs 的总算力，两者相差了 9 个数量级。

显存墙，指的是单卡无法完整存储一个大模型的参数。GPT-3 的 1750 亿参数本身就需要 700 GB 的显存空间（每个参数按照 4 个字节计算），而 NVIDIA A100 GPU 只有 80 GB 显存。

关于作者

百度智能云开发者中心

5659 被阅读数
5 被赞数
3 被收藏数

关 注

文章标签

论坛公告 开发者运营 人工智能 科技前沿 AI开发平台 了不起的开发者

最热文章

- AI 大底座，大模型时代的答卷
- 云智一体领跑大模型产业发展
- 以太网交换技术
- 【地图开发者专属活动】轻松三步即可开启地图服务能力

计算墙和显存墙的本质是有限的单卡能力和模型的巨大的存储、计算需求之间的矛盾。这可以通过分布式训练的方法解决，但分布式训练之后又会遇到通信墙的问题。

通信墙，主要是分布式训练下集群各计算单元需要频繁参数同步，通信性能将影响整体计算速度。如果通信墙如果处理的不好，很可能导致集群规模越大，训练效率反而会降低。成功的突破通信墙，体现为集群有较强的扩展能力，即集群的多卡加速能力和规模是匹配的。多卡的线性加速比就是评估集群多卡加速能力的指标，数值越高越好。

这几堵墙在多机多卡的训练中开始出现。随着大模型的参数越来越大，对应的集群规模也越来越大，这三堵墙也越来越高。同时，在大集群长时间训练过程中，还会出现设备故障，有可能会影响或者中断训练进程。

4. 大模型训练的过程

一般来说，从基础设施视角看大模型训练，整个过程可以大致分成以下两个阶段：

阶段一：并行策略和训练优化

在提交待训练的大模型后，AI 框架会综合考虑大模型的结构等信息、以及训练集群的能力，为本次训练任务制定出一个并行训练策略，并完成 AI 任务放置。这个过程就是拆开模型、放置任务，即大模型应该被如何拆解，被拆开的各个部分如何放置到集群的各个 GPU/XPU 中。

针对放置在 GPU/XPU 中运行的 AI 任务，AI 框架会联合训练集群在单卡运行时和集群通信层面进行全链路优化，加速大模型训练过程中各个 AI 任务的运行效率，包括数据加载，算子计算、通信策略等。比如将 AI 任务中运行的普通算子替换为经过优化的高性能算子，提供适配当前并行策略和训练集群网络能力的通信策略等。

阶段二：资源管理和任务调度

大模型训练任务按照上面制定的并行策略开始运行，训练集群为 AI 任务提供各类高性能的资源。比如 AI 任务运行在什么环境中，如何为 AI 任务提供资源对接，AI 任务通过什么存储方式读取和保存数据，GPU/XPU 通过什么类型网络设施通信等。

同时，在运行过程中，训练集群会联合 AI 框架通过弹性容错等方式，为大模型的长时间训练提供可靠的环境。比如如何观测和感知集群中各类资源和 AI 任务的运行状态等，如何在集群变化时能够对资源和 AI 任务进行调度等。

从以上两个阶段的拆解中，我们可以发现整个大模型训练的过程，都依赖 AI 框架和训练集群的密切配合，完成对三堵墙的突破，共同确保大模型训练的高效和稳定。

5. 全栈融合，「AI 大底座」加速大模型训练

结合多年在 AI 和大模型领域的技术积累和工程实践，百度在 2022 年底推出了全栈自研的 AI 基础设施「AI 大底座」，包括「芯片 – 框架 – 模型 」三层技术栈，在各个层面都拥有关键自研技术和领先产品，分别对应昆仑芯、飞桨 （PaddlePaddle）、文心大模型。

在这三层技术栈的基础上，百度智能云推出了两大 AI 工程平台，「AI 中台」和「百度百舸· AI 异构计算平台」，分别在开发和资源层面进行提效，完成对三堵墙的突破，加速训练过程。

其中，「AI 中台」依托 AI 框架为大模型训练过程制定并行策略和优化过的环境，覆盖训练的全生命周期。「百度百舸 」实现了高效的芯片使能，提供各类 AI 资源的管理和任务调度的能力。

百度「AI 大底座」对各层的技术栈进行了全栈融合、系统优化，完成了云和智的技术一体化建设，可以实现对大模型训练的端到端优化和加速。

百度集团副总裁侯震宇：大模型训练是一个系统工程，集群规模、训练时间、花费金额，相比过去都提高了很多。如果不是全栈优化，很难保证大模型训练的顺利完成。百度多年来在大模型上的技术投入和工程实践，使得我们建立起了一套完整的软件栈能力，用来加速大模型的训练。

接下来，我们将结合上文提到的大模型训练过程的两阶段，讲述「AI 大底座」的各层技术栈是如何相互融合、系统优化，实现大模型训练的端到端优化和加速。

5.1 并行策略和训练优化5

模型拆分

飞桨可以为大模型训练提供数据并行、模型并行、流水并行、参数分组切片、专家并行等丰富的并行策略。这些并行策略可以满足从十亿到千亿、甚至万亿参数规模大模型的训练，实现对计算墙和显存墙的突破。2021 年 4 月，飞桨在业界第一个提出 4D 混合并行策略，可支持千亿级大模型的训练在月级别完成。

拓扑感知

百度百舸拥有专为大模型训练场景准备的集群拓扑感知能力，包括节点内架构感知、节点间架构感知等，比如每台服务器内部的算力大小、 CPU 和 GPU/XPU、GPU/XPU 和 GPU/XPU 链接方式，以及服务器之间 GPU/XPU 和 GPU/XPU 网络链接方式等信息。

自动并行

在大模型训练任务开始运行前，飞桨可以依据百度百舸平台的拓扑感知能力，对集群形成统一分布式资源图。同时，飞桨根据待训练的大模型形成的统一逻辑计算视图。

综合这两张图，飞桨自动化地为模型搜索出最优的模型切分和硬件组合策略，将模型参数、梯度、优化器状态按照最优策略分配到不同的 GPU/XPU 上，完成 AI 任务的放置以提升训练性能。

比如将模型并行的 AI 任务都放置在同一台服务器的不同 GPU 上，这些 GPU 通过服务器内部的 NVSwitch 链接。将数据并行、流水线并行的 AI 任务放置在不同服务器的同号 GPU 上，这些 GPU 通过 IB 或者 RoCE 链接。通过这种依据 AI 任务的类型进行 AI 任务放置的的方法，使得集群资源能够被高效使用，加速大模型训练。

端到端自适应训练

在训练任务运行过程中，如果集群发生了变化，比如有资源出现了故障，或者集群规模有变化，百度百舸会进行容错的替换或者弹性扩缩容。由于参与计算的节点所在位置发生了变化，它们之间的通信模式也许已经不是最优。飞桨能够依据最新的集群信息，自动调整模型切分和 AI 任务放置策略。同时，百度百舸完成相应的任务和资源的调度。

飞桨统一的资源和计算视图以及自动并行能力，再结合百度百舸的弹性调度能力，实现了大模型的端到端自适应分布式训练，可以覆盖集群训练的全生命周期。

这是 AI 框架和 AI 异构算力平台的深入交互，实现了算力、框架、算法三位一体的系统优化，支持大模型自动弹性的进行训练，端到端实测有 2.1 倍的性能提升，保证了大规模训练的高效性。

训练优化

完成模型的拆分和 AI 任务的放置后，在训练过程中为了确保算子在飞桨、Pytorch 等各类主流 AI 框架和各类计算卡上可以加速计算，百度百舸平台中内置了 AI 加速套件。AI 加速套件包括了数据层存储加速、训练和推理加速库 AIAK ，分别从数据加载、模型计算、分布式通信等维度进行了全链路优化。

其中，数据加载和模型计算的优化可以有效提高单卡的运行效率；分布式通信的优化，结合集群的 IB 或者 RoCE 等高性能网络和专门优化的通信拓扑，以及合理的 AI 任务放置策略，共同解决通信墙问题。

百度百舸在千卡规模集群中的多卡加速比达到了 90%，使得集群拥有的整体算力可以被充分释放出来。

在 2022 年 11 月发布的 MLPerf Trainning v2.1 测试结果中，百度使用飞桨加百度百舸提交的模型训练性能结果，位列同等 GPU 配置下世界第一，端到端训练时间和训练吞吐均超越 NGC PyTorch 框架。

5.2 资源管理和任务调度

百度百舸通过容器引擎 CCE 承载所有 AI 任务的运行，并通过相关容器插件的方式提供各类 AI 资源管理、架构感知、弹性容错等能力，在资源效能层面完成计算墙、显存墙、通信墙的突破。

资源管理

百度百舸可以提供各类计算、网络、存储等 AI 资源，包括百度太行·弹性裸金属服务器 BBC、IB 网络、RoCE 网络、并行文件存储 PFS、对象存储 BOS、数据湖存储加速 RapidFS 等各类适合大模型训练的云计算资源。

在任务运行时，可以将这些高性能资源进行合理的组合，进一步提升 AI 作业的效率，全流程实现 AI 任务的计算加速。在 AI 任务开始前可以预热对象存储 BOS 中的训练数据，通过弹性 RDMA 网络将数据加载至数据湖存储加速 RapidFS 中。弹性 RDMA 网络相比传统网络可以降低 2 至 3 倍通信时延，在高性能存储的基础上，加速 AI 任务数据的读取。最后通过高性能的百度太行·弹性裸金属服务器 BBC 或者云服务器 BCC，进行 AI 任务的计算。

弹性容错

AI 任务运行时，不仅需要高性能的资源，还需要确保集群的稳定，最大程度降低资源故障发生率以免打断训练。但是，资源的故障不能绝对避免，AI 框架和训练集群需要联合保证训练任务被打断后能够从最近的状态恢复，从而为大模型的长时间训练提供可靠环境。

百度自研的异构集合通库 ECCL ，支持昆仑芯和其他异构芯片的通信，支持慢节点和故障节点的感知。通过百度百舸的资源弹性和容错策略，将慢节点和故障节点剔除，并将最新的架构拓扑反馈给飞桨，重新进行任务布置，对应训练任务调配至其他 XPU/GPU 上，确保训练的平滑高效运行。

6. 大模型时代的 AI 普惠

大模型是人工智能迈向通用智能的里程碑技术，驾驭好大模型是完成智能化升级路径上的必答题。超大规模的算力、全栈融合的软件优化，是对这道必答题的最好回答。

为了帮助社会和产业快速训练出自己的大模型，抢占时代先机，2022 年底百度智能云发布了阳泉智算中心，搭载百度「AI 大底座」的全栈能力，可以提供 4 EFLOPS 的异构算力。这是目前亚洲单体规模最大、技术最先进的数据中心。

目前，百度智能云已经将「AI 大底座」的全部能力对外开放，实现大模型时代的 AI 普惠，通过各个地域的中心云、边缘云 BEC、本地计算集群 LCC、私有云 ABC Stack 等多种形式进行交付，使得社会和产业可以方便的获得智能服务。