利用高性能计算加速深度学习算法

张广勇(QQ: 331526010, Email: zhang03_11@163.com)

2015 年 10 月底

1. 深度学习

深度学习是机器学习研究中的一个新的领域，其动机在于建立、模拟人脑进行分析学习的神经网络，它模仿人脑的机制来解释数据，例如图像，声音和文本。深度学习典型应用为图像识别和语音识别。（由于本人不是深度学习专业人士，对深度学习理论知识不多介绍，说多了就班门弄斧了，后面主要介绍下这些深度学习算法如何进行并行化设计和优化）

2. CPU+GPU 异构协同计算简介

近年来，计算机图形处理器（GPU，Graphics Process Unit）正在以大大超过摩尔定律的速度高速发展（大约每隔半年 GPU 的性能增加一倍），远远超过了 CPU 的发展速度。

CPU+GPU 异构协同计算模式(图 1)，利用 CPU 进行复杂逻辑和事务处理等串行计算，利用 GPU 完成大规模并行计算，即可以各尽其能，充分发挥计算系统的处理能力。

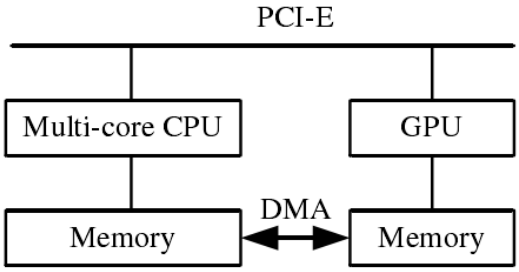

图 1 CPU+GPU 异构体系结构

目前，主流的 GPU 具有强大的计算能力和内存带宽，如图 2 所示，无论性能还是内存带宽，均远大于同代的 CPU。对于 GPU， Gflop/$和 Gflops/w 均高于 CPU。

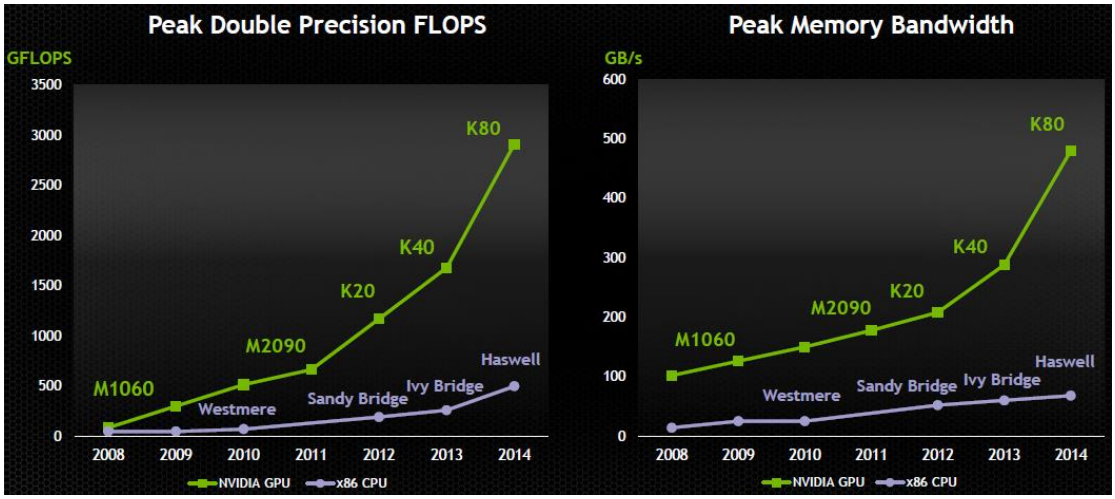

图 2 GPU 计算能力

3. 深度学习中的 CPU+GPU 集群架构

CPU+GPU 集群工作模式（图 3），每个节点内采用 CPU+GPU 异构模式，并且每个节点可以配置多块 GPU 卡。节点间采用高速 InfiniBand 网络互连，速度可以达到双向 56Gb/s,实测双向 5GB/s。后端采用并行文件系统。采用数据划分、任务划分的方式对应用进行并行化，适用于大规模数据并行计算。

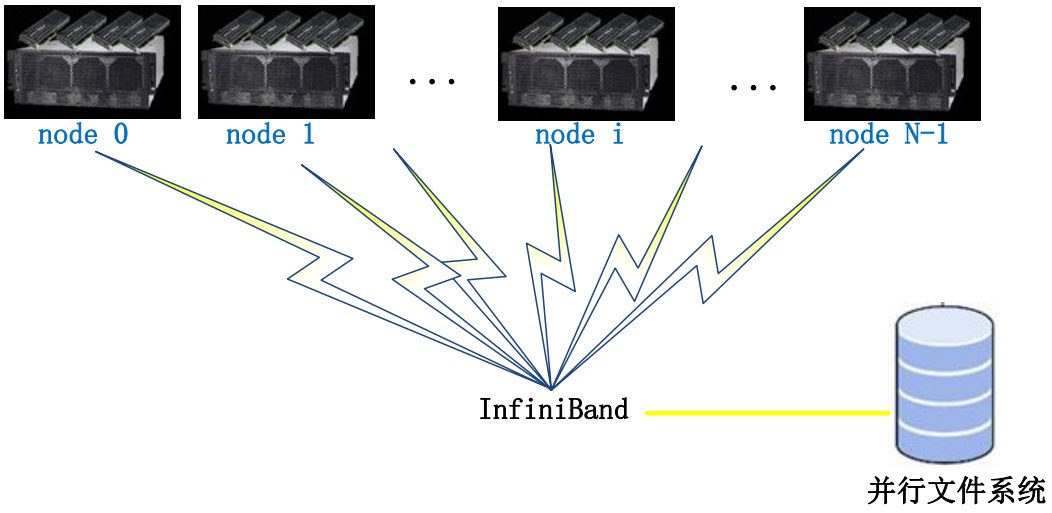

图 3 CPU+GPU 集群架构

4. 利用 GPU 加速深度学习算法

4.1. 单 GPU 并行

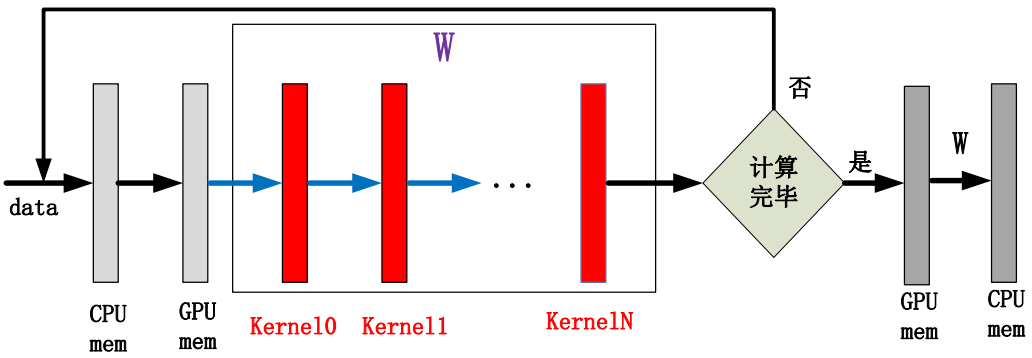

图 4 单 GPU 计算流程

针对每次训练数据，模型内计算通过多次 GPU 内核的调用完成计算。权重 W 值一直存在 GPU 内存中，直到所有训练数据计算完毕之后回传到系统内存中。Data 为图像或语音数据。

4.2. 多 GPU 卡并行

多 GPU 并行计算时，各 GPU 有自己独立的内存，卡之间的并行属于分布式计算模式。针对深度学习算法，采用多 GPU 卡计算可以采用两种并行方法：数据并行和模型并行。

4.2.1. 数据并行

数据并行是指不同的 GPU 计算不同的训练数据，即把训练数据划分给不同的 GPU 进行分别计算，由于训练是逐步训练的，后一个训练数据的计算需要前一个训练数据更新的 W，数据并行改变了这个计算顺序，多 GPU 计算需要进行 W 的互相通信，满足训练的特点，使训练可以收敛。

数据并行如图 5 所示，多 GPU 训练不同的数据，每训练一次需要同步 W，使得后面的训练始终为最新的 W。

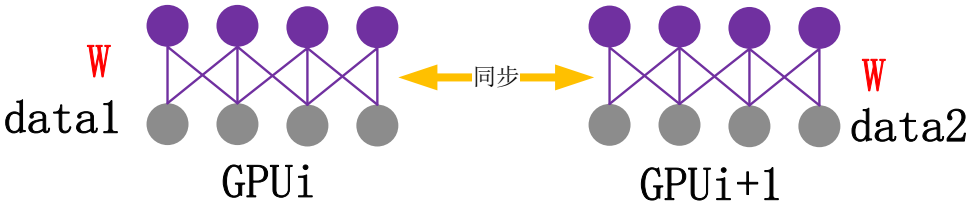

图 5 数据并行

数据并行的特点：

1) 优点
   a) 实现比较容易，也比较容易扩展
   b) 只需要进行 W 的通信，模型内的数据不需要通信
2) 缺点
   a) 当模型较大时，GPU 内存无法满足存储要求，无法完成计算

根据多 GPU 卡之间的 W 通信，下面介绍两种通信方法：主从模式和令牌环模式。

1) 主从模式

主从模式：选择一个进程或线程作为主进程或线程，各个 GPU 把每次训练得到的 ΔW 发给主进程或线程，主进程或线程进行 W 更新，然后再发送给 GPU，GPU 再进行下一个数据的训练。如图 6 所示。

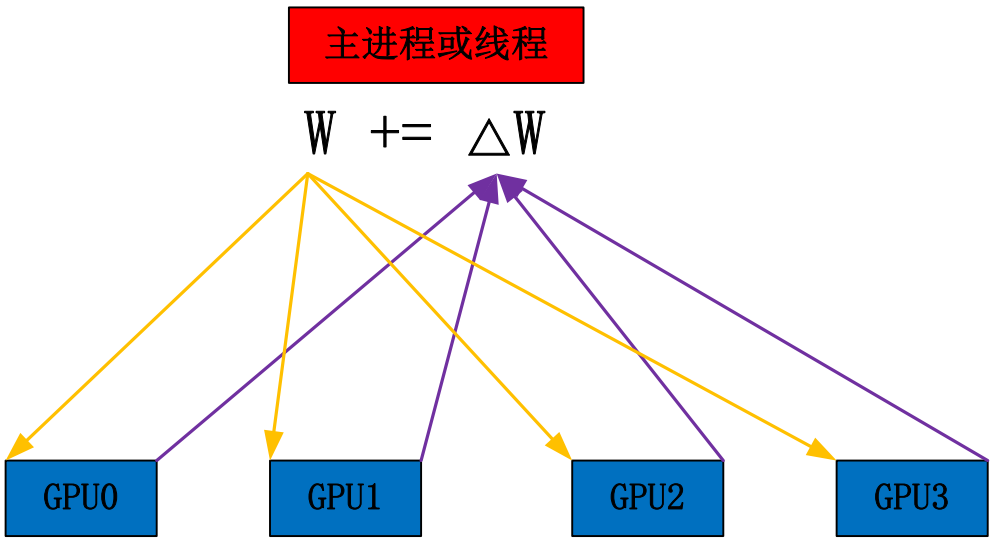

图 6 主从模式

2) 令牌环模式

令牌环模式：每个 GPU 把自己训练得到的 ΔW 更新到 W 上，并且发送到下一个 GPU，保证令牌环上的 W 始终为最新的 W。如图 7 所示。

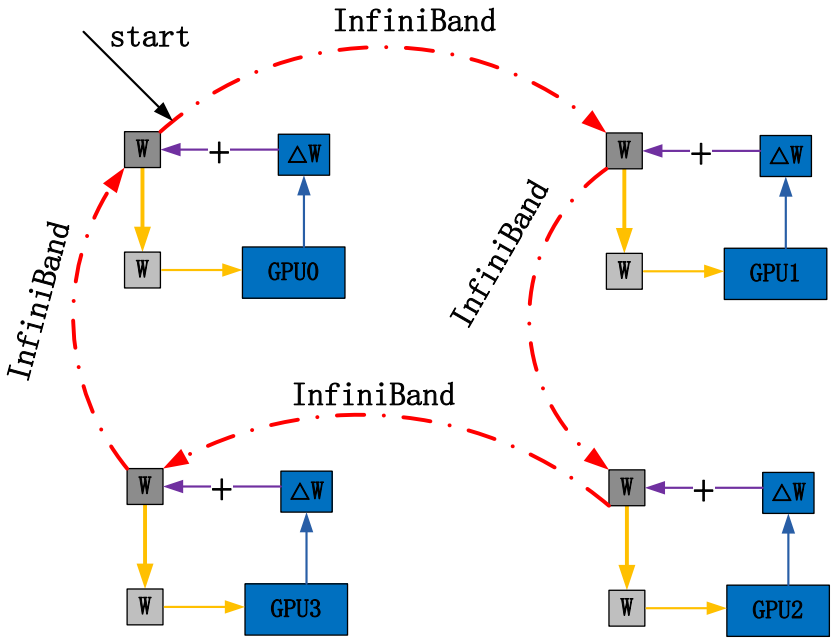

图 7 令牌环模式

两种模式对比如表 1

表 1 主从模式和令牌环模式对比

模式	优点	缺点
主从模式	收敛速度更快	GPU 计算需要等待，影响 GPU 计算；主进程或线程压力较大
令牌环模式	GPU 计算不需要等待通信，性能更好	通信速度影响收敛的速度

4.2.2. 模型并行

模型并行是指多个 GPU 同时计算同一个训练数据，多个 GPU 对模型内的数据进行划分，如图 8 所示。Kernel 计算和通信流程如图 9 所示，在一次训练数据多层计算过程中，每个 GPU 内核计算之后需要互相交换数据才能进行下一次的计算。

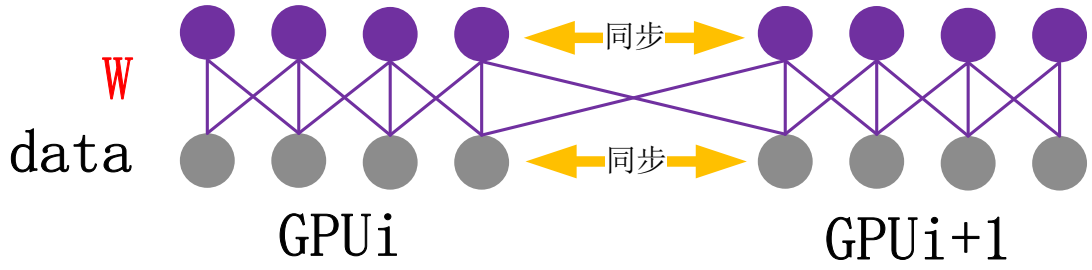

图 8 模型并行

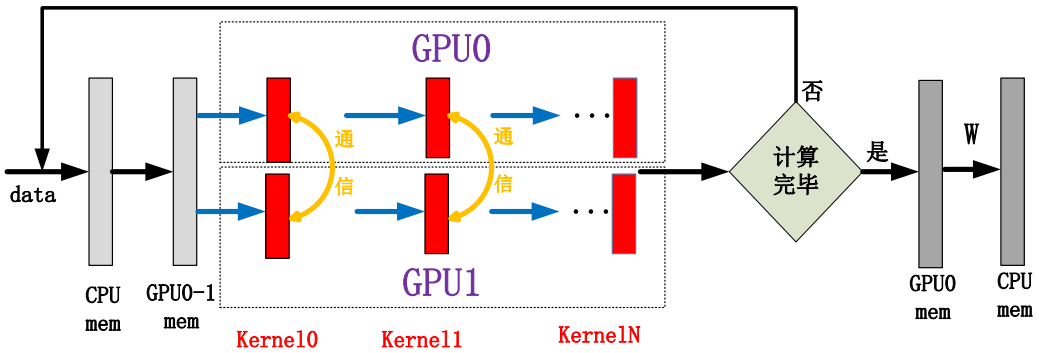

图 9 模型并行：多 GPU 计算内核和通信示意图

模型并行特点：

1) 优点
   a) 可以处理大模型的情况
2) 缺点
   a) 需要更频繁的通信，增加通信压力
   b) 实现难度较大

4.3. GPU 集群并行

GPU 集群并行模式即为多 GPU 并行中各种并行模式的扩展，如表 2 所示。节点间采用 InfiniBand 通信，节点间的 GPU 通过 RMDA 通信，节点内多 GPU 之间采用 P2P 通信。

表 2 GPU 集群并行模式

模式	节点间	节点内	特点
模式 1	令牌环		单一模式的缺点放大
模式 2	主从		
模式 3	模型并行		
模式 4	令牌环	主从	结合各种模式的有点，避免某一模式的缺点放大
模式 5	主从	令牌环	

模式 6	令牌环	模型并行	
模式 7	主从	模型并行	

4.4. 性能分享

4.4.1. 基于 GPU 集群的 Caffe 并行加速

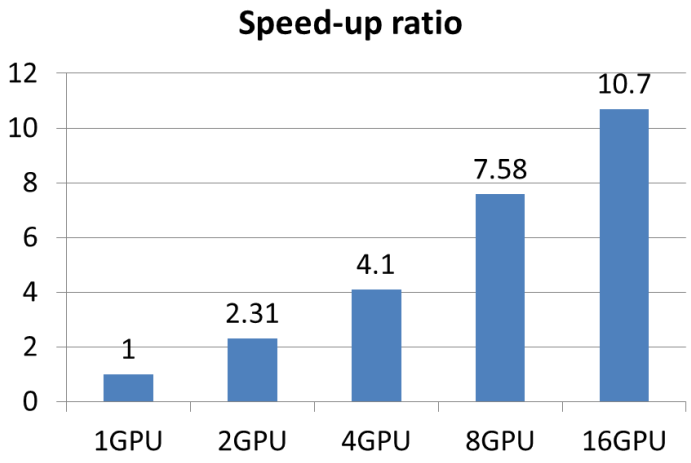

图 10 Caffe 性能

8 节点 GPU 服务器，2 K20m GPU/节点，56Gb/s InfiniBand 网络，Lustre 并行文件系统

4.4.2. 基于 GPU 集群的 DNN 并行加速

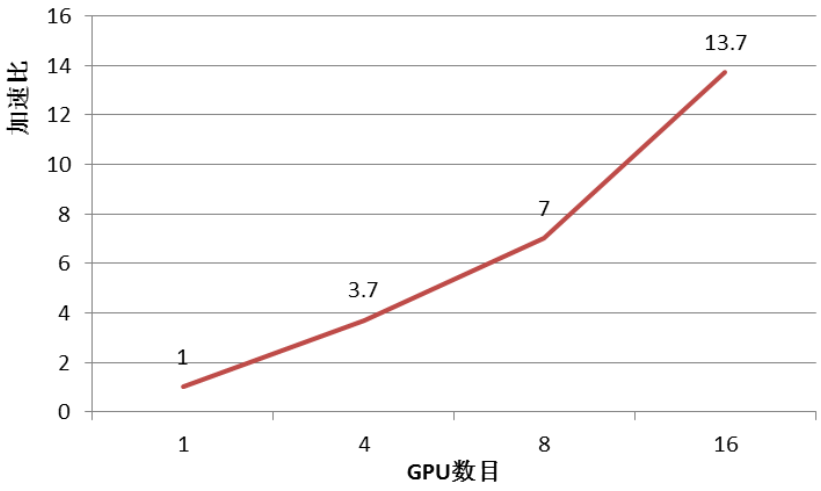

图 11 DNN 测试性能

4 节点 GPU 服务器，4 K20m GPU/节点，56Gb/s InfiniBand 网络

5. CPU+FPGA 协同计算加速线上计算

相对于训练计算，线上识别的计算是小而众多的任务计算，每次请求的计算比较小，但请求的任务数十分庞大，GPU 计算虽然获得很好的性能，但功耗仍然是个严峻的问题。

目前主流的 FPGA 卡功耗只有主流 GPU 的十分之一，性能相差 2-3 倍，FPGA 相对于 GPU 具有更高的 GFlops/W。

利用 FPGA 解决线上识别计算可以采用分布式+FPGA 计算的模式，如图 12 所示，上层可以采用 Hadoop 或 Spark 调度，底层利用 FPGA 进行计算。

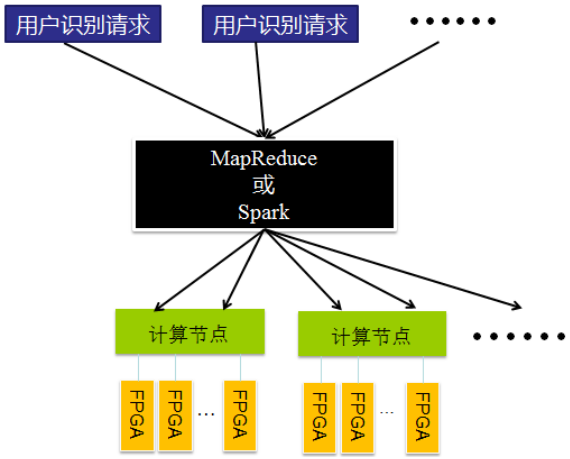

图 12 分布式计算+FPGA 计算

目前，FPGA 已开始支持高级语言，如 Altera FPGA 支持 OpenCL，Xilinx FPGA 支持 HLS，这对程序员利用 FPGA 开发减低了难度。这些新平台的支持还有很多问题，也许后面会支持的越来越好。

注：由于对深度学习算法了解比较肤浅，以上内容难免无误，请大家理解并提出修改意见。