关于几款AI加速器的random thoughts

https://zhuanlan.zhihu.com/p/407119171

None Wed Sep, 08 21:26

所有信息来源于这里，整理一些关于这些AI加速器的random thoughts，用做自娱自乐以及和感兴趣同学的讨论。

1.Cerebras、Graphcore、SambaNova的访存体系的设计几乎都遵循了**large size on-chip SRAM + near device large capacity DDR**的技术路线，而不像NV GPU和Google TPU走的是**片上少量SRAM + stacked HBM**的技术路线。这里的一个argue point是这种on-chip SRAM + DDR的组合，能够获得更划算的性价比以及功耗。更具体一些，通过应用程序控制SRAM的具体用途，将其作为programmable cache，来缓解对DDR的带宽需求压力，从而避免引入HBM这种成本、功耗、封装技术要求都比较高的手段。

	On-chip SRAM	Near-device DDR	Stacked HBM
Cerebras	40GiB(wafer-scale)	4TiB~2.4PiB	NA
Graphcore	304~897MiB	?(expected to be TiB scale?)	NA
SambaNova	>300MiB	1.5TiB	NA
NV A100	40MiB	NA	40~80GiB

这种作法，理论上说是可行的，只要on-chip的SRAM大到一定程度(比如说，一个完整的op计算所需的weight和activation可以被放在on-chip SRAM里，如果超过on-chip SRAM，实际上也可能通过op的拆解来应对，只不过会增加额外的复杂性，这里先不展开)，使得当前op的计算耗时可以hide后继op所需的weight/activation的prefetch耗时。以及，为了节省on-chip SRAM的资源，当前op在完成计算以后，要考虑把输出的activation及时写回到DDR里，以腾挪出SRAM的资源(这种写出的频率可能是per-op的，也可能是per-op block的，相对应的在反向处理的时候，either是更精细的prefetch策略，either是re-compute来用计算换取访存带宽的节省)。

NV之前为什么没有考虑这种on-chip SRAM + DDR的策略呢？我的一个推测是这种作法会要求上层软件对SRAM的使用非常考究，对框架会有比较深的定制要求，NV虽然提供了NGC的框架环境，但毕竟不直接own AI框架，所以通过HBM可以让硬件找到一个访存带宽和内存容量相对满足要求的trade-off点，来简化上层软件人员的负担。对于AI DSA start-up，需要有足够强的差异化，并且HBM的集成对于初创公司还是会引入更多的技术复杂性，所以选择了on-chip SRAM + DDR的策略。

至于Google在TPU里为什么没有类似的考虑，我倒是比较好奇。因为它自己own软硬全栈，完全有机会做这种事情。

2.这几款AI加速器都没有明确提到host跟device交互的overhead。从之前自己曾经有限的评估几款AI新硬件的经验来看，一但涉及到host跟device的交互，性能会有明显的下降，这里有些性能下降是可以通过软件手段来mitigate的，比如异步的kernel发射，有些会更麻烦一些，比如之前讨论过的一些细节。这里有一篇寒武纪和计算所合作的文章探讨了这个问题，并从硬件加速器的角度给出了一个解法。我对这篇文章的工作非常喜欢，这也是目前我看到的公开文献里第一个解决host/device交互开销问题的工作。虽然这个工作还是存在一些局限性，比如

- 对用户模型写法会有侵入性改写的要求（看一下文章里那个NMS的写法就可以理解了）
- 还是不能根解过于Pythonic写法引入的host/device交互开销的问题（毕竟在文章里引入的CPULess的DLPU只是对scalar操作提供了专门的支持，但还没有达到可以支持任意python operation的程度）

但仍然是一个非常有意义的探索。之前和一些朋友讨论，提到过解决host/device的**interaction wall**(文章里的这个提法我觉得很形象，直接拿来引用了)的一种方式是类似于Intel目前的作法，在通用CPU里加入Tensor专用加速单元，这篇文章则提供了另一种作法的参照。

3.最近似乎因为Tesla AI day的原因，对Cerebras和Dojo的工作关注的同行比较多，对SambaNova的工作关注的比较少。我自己是对SambaNova这种可配置计算的作法会更感兴趣，因为更灵活，更有可能结合上层workload的变化完成domain specific的适配。在几年前明星公司Wave曾经想推出类似的产品，但未能成功deliver，一度让行业觉得这种作法并不可行，SambaNova把这个非常激进的技术想法变成了实际的产品并完成落地，蛮值得关注。SambaNova的show case里的一个案例里把预处理，SQL处理和模型训练通过重配置放在了一组RDU计算设备里，这在一定程度上体现了SambaNova这种架构的优越性。目前公开可查的资源，还不容易确定SambaNova这个架构的可配置的边界在哪里（这里的边界是指在确保decent性能的边界），因为我总是觉得不会存在系统设计的silver bullet，SambaNova在拿到这种灵活性优势的同时，应该会有一些代价吧。

4.这几款硬件理论峰值算力都不算非常高(Cerebras的我没有查到明确的数字)，相同工艺水准下和晶体管数目下，绝对峰值算力相比A100都没有明显优势（甚至有些是存在gap的）。但是这几家的访存系统以及全栈系统设计和NV都存在明显的差异。也就是说，这几家在通过全栈系统优化追求更高的有效算力，而不是一昧飙理论峰值算力。

	工艺	晶体管数量	峰值算力(TFlops)
Cerebras	7nm	2.6T(wafer-scale)	?
Graphcore	7nm	59B	250 FP16
SambaNova	7nm	40B	>300 BF16
NV A100	7nm	54B	312 BF16

5.这几家新硬件的最高性能表现还是集中在BF16/FP16低精度格式上，而NV A100在V100时代完成了FP16的业务试水之后，在A100上为TF32也加入了TensorCore支持(156TFlops)，并已经将TF32变成AI框架的default mode，这种精度的de facto标准，可能会给新硬件带来压力。毕竟，建模同学是不喜欢引入训练过程的过多不确定性的。

6.研发一款7nm的硬件大约会消耗多少钱呢？基于这里和这里的一些公开信息，范围在**1.2亿**美金到**4.2亿**美金。考虑到这些信息略有些老，稍微拍脑袋打点折，可以勉强normalize成**1.5亿美金**好了。A100的八卡整机价格是**20万美金**左右。一块A100显卡的公开报价是**1万美金**左右。所以一款7nm的研件研发成本可以用于购买**15000张**A100显卡(假设整机不购买A100，可以通过其他方式装配到更便宜的服务器整机上)，**750台**A100 DGX高配整机(合计**6000张**A100)。基于前面的假设性推演，研发一款7nm AI DSA新硬件，至少需要卖到这个规模才可能勉强cover研发成本。如果能够通过行业的技术进步(可能是RISC-V，可能是CIRCT，可能是MLIR，也可能是其他的什么东东)，把AI DSA的研发成本进一步拉低，对于DSA更多涌现高效解决vertical的问题会有帮助。当然这也意味着NV的压力可能就更大了：）

7.上面讨论的还都是算力的供给，AI新硬件公司和NV、Intel、AMD都主要扮演的是硬件算力提供者的角色。而AI算力能够被有效消耗，还依赖于AI能够更加普及被应用到行业里。从这个角度来说，仅有算力是远远不够的，还需要一些什么呢？可能会需要一些AI落地工具化，方法论的东西(感谢 @mackler 的讨论，这些观点的核心在他的一些文章里能够找到一些线索):

- 经典的AI建模的抽象原语（比如Conv、Attention、Residual、BN等等)
- 配套工具，便于对模型进行debug和profiling(这里的profiling不是常规意义的性能profiling，而是模型效果的profiling)，最简单的例子，比如方便地看梯度在训练过程中的分布变化，各个layer feature map的分布变化趋势，以及不同layer的模型权重和label的相关系数变化趋势等等。
- 方法论，用于指导根据建模实验的反馈，下一步应该做些什么可以把模型结果变得离期望更近。
- 高效的算力和AI系统，让建模过程中trial and error的cycle变短，无需太关心模型运行过程中的底层软硬件细节，可以专注于模型调试profiling本身。
- 流畅的MLOps工具的支持，让训练好的模型能够顺滑地完成布署、上线以及更新迭代。

我个人的观察，目前只有最后两项(算力和MLOps工具)相对成熟一些，而上面三点的成熟度会决定最后两项的影响力上界，因为再hard core的技术，也最终需要落实解决终端具体问题，才能完成商业闭环。AI作为一种通用目的性技术发挥更广泛的影响，还有很长的路要走。

编辑于 09-06

文章被以下专栏收录