特斯拉Dojo芯片，领先竞争对手一个数量级

https://mp.weixin.qq.com/s/EPgSUdG_3_t3XpeXVQne0A

None Sat Aug 21 09:05

来源：内容由半导体行业观察（ID:icbank）编译自「semianalysis」，谢谢。

特斯拉刚刚召开了他们的人工智能日，并透露了其软件和硬件基础设施的内部工作原理。此次公开的部分一是之前公开的Dojo AI训练芯片。特斯拉称他们的 D1 Dojo 芯片具有 GPU 级别的计算能力、CPU 级别的灵活性，以及网络交换机IO。

几周前，我们推测该系统的封装是 TSMC 晶圆集成扇出系统 (InFO_SoW)。我们解释了这种类型的封装的好处以及这种大规模扩展训练芯片所涉及的冷却和功耗。此外，我们估计该软件包的性能将优于 Nvidia 系统。所有这些似乎都是有根据的推测。今天，我们将深入的挖掘更多的半导体细节。

在我们深入讨论硬件之前，让我们先谈谈评估基础架构。特斯拉不断地重新训练和改进他们的神经网络。他们评估任何代码更改以查看是否有改进。在汽车和服务器中部署了数以千计的相同芯片。他们每周进行数百万次评估。

多年来，特斯拉一直在扩大其GPU集群的规模。如果特斯拉停止所有实际工作负载，运行Linpack，并将其提交到Top500榜单，他们目前的训练集群将成为第五大超级计算机。然而，对于特斯拉和其雄心来说，这种性能的扩展是不够的，所以他们在几年前就开始开发自己的Dojo芯片项目。特斯拉需要更高的性能，以节能且经济高效的方式实现更大、更复杂的神经网络。

特斯拉的架构解决方案是分布式计算架构。当我们听他们的细节时，这个架构看起来很像Cerberus。我们分析了Cerebras晶圆级引擎及其架构。每个AI训练架构都是以这种方式布置的，但计算元素、网络和结构的细节差别很大。这些类型的网络最大的问题是扩大带宽和保持低延迟。为了扩大网络规模，特斯拉特别关注于此，这影响了他们设计的每个部分，从芯片材料到封装。

该功能单元被设计为可通过1个时钟周期，但足够大，以至于同步开销和软件不是主要问题。因此，他们得出的设计几乎和Cerebras一模一样。由高速结构连接的单个单元的网状结构，该结构在一个时钟内的功能是单元之间的路由通信。每个单元都有一个大的1.25MB SRAM暂存板和多个具有SIMD功能的超标量CPU核，以及支持所有常见数据类型的矩阵乘法单元。此外，他们还引入了一种名为CFP8的新数据类型，可配置浮点8。每个单元可在BF16 / CFP8上支持1TFlop，FP31上64GFlops，每个方向512GB/s的带宽。

CPU也是毫不逊色，它在矢量管道上是4宽2宽。每个核心可以承载4个线程，以最大限度的提高利用率。遗憾的是，特斯拉使用了定制的ISA，而不是像 RISC V 这样的顶级开源 ISA。这个定制的ISA引入了转置，收集，广播，和链接遍历的指令。

这354个功能单元的全芯片可达到BF16或CFP8的362 TFlops和FP32的22.6 TFlops。它总共有645mm²和500亿个晶体管。每个芯片都有惊人的400W TDP，这意味着功率密度高于大多数配置

的Nvidia A100 GPU。有趣的是，特斯拉实现了每mm²上集成7750万个晶体管的有效晶体管密度，仅次于移动芯片和苹果M1，比其他所有高性能芯片都要高。

基本功能单元的另一个有趣的方面是NOC路由器。它与Tenstorrent有非常类似的在芯片内部和芯片间的扩展方式。毫不意外，特斯拉正在采用与其他备受推崇的人工智能初创公司类似的架构。Tenstorrent 非常适合扩展培训，而特斯拉非常关注这方面。

在芯片上，特斯拉拥有惊人的10TBps定向带宽，但这个数字在实际工作负载中没有多大意义。特斯拉相对于Tenstorrent的一个巨大优势是芯片之间的带宽要高得多。他们在 112GTs 上有 576 个 SerDes。这产生了总共 64Tb/s 或 8TB/s 的带宽。

我们不确定特斯拉每条边的4TB/s是从哪里得到的，更有可能是X轴和Y轴上的数字。先不说这张令人困惑的幻灯片，这个芯片的带宽是疯狂的。目前已知的最高外部带宽芯片是32Tb/s网络交换机芯片。特斯拉能够通过大量的SerDes和先进的封装将这个数字翻倍。

特斯拉通过PCIe 4.0将Dojo芯片的计算平面连接到连接主机系统的接口处理器上。这些接口处理器还支持更高的基数网络连接，以补充现有的计算平面网格。

25个D1芯片被封装成“扇出晶圆工艺（fan out wafer process）”。特斯拉并没有像我们几周前猜测的那样确认这个封装是台积电的集成晶圆扇形系统(InFO_SoW)，但考虑到疯狂的芯片间带宽和他们特别提到的扇出晶圆，这看起来很有可能。

特斯拉开发了一种专有的高带宽连接器，可以保留这些芯片之间的芯片外带宽。每个芯片都有令人印象深刻的9PFlops BF16/CFP8和36tb /s的off-tile带宽。这远远超过了Cerebras的晶圆外带宽，使特斯拉系统的横向扩展能力甚至比横向扩展设计（例如 Tenstorrent 架构）还要好。

电源传输是独一无二的，定制的，也非常令人印象深刻。由于具有如此大的带宽和超过 10KW 的功耗，特斯拉在电力传输方面进行了创新，并垂直供电。定制稳压器调制器直接回流到扇出晶片上。功率、热量和机械都直接与芯片连接。

即使芯片本身的总功率只有10KW，但芯片的总功率仍然是15KW。电力传输、IO和晶圆线也在消耗大量的电力。能量从底部进来，热量从顶部出来。特斯拉的规模单位不是芯片，而是25块芯片。这个贴图远远超过了Nvidia, Graphcore, Cerebras, Groq, Tenstorrent, SambaNova，或任何其他AI训练项目的单位性能和扩展能力。

所有这些似乎都是非常遥远的技术，但特斯拉声称，他们已经在实验室的真实人工智能网络上以2GHz的频率运行了芯片。

扩展到数千个芯片的下一步是服务器级别。Dojo可扩展为2 x 3的tile配置，在一个服务器中有两个这样的配置。对于那些在家计数的人来说，每个服务器总共有12个tile，每个服务器总共有108个PFlops，超过100,000个功能单元，400,000个定制核和132GB SRAM是令人震惊的数字。

特斯拉不断扩大其网格中的机柜级别。芯片之间没有带宽中断，它是一个具有惊人带宽的同质芯片网格。他们计划扩大到10个机柜、1.1 Exaflops、1,062,000个功能单元、4,248,000个核心和1.33TB的SRAM。

软件方面很有趣，但我们今天不会太深入讨论。他们声称他们可以对其进行虚拟细分。他们说不管集群的大小如何，软件都可以在Dojo处理单元(DPU)之间无缝扩展。Dojo 编译器可以处理

硬件计算平面的细粒度并行和映射网络。它可以通过数据模型图并行性来实现这一点，但也可以进行优化以减少内存占用。

模型并行性可以跨芯片边界扩展，甚至不需要大批量的轻松解锁具有数万亿参数甚至更多参数的下一级AI模型。他们不需要依赖手写的代码来在这个庞大的集群上运行模型。

总的来说，与英伟达的GPU相比，成本相当，但特斯拉声称他们可以实现4倍的性能，每瓦性能提高1.3倍，减少5倍的面积。特斯拉的TCO优势几乎比英伟达的AI解决方案好一个数量级。如果他们的说法是真的，特斯拉已经超越了人工智能硬件和软件领域的所有人。我对此表示怀疑，但这也是硬件极客的美梦。

我们都要试图冷静下来，等一等，看看它什么时候会实际部署到生产环境中。

*免责声明：本文由作者原创。文章内容系作者个人观点，半导体行业观察转载仅为了传达一种不同的观点，不代表半导体行业观察对该观点赞同或支持，如果有任何异议，欢迎联系半导体行业观察。

今天是《半导体行业观察》为您分享的第2773内容，欢迎关注。

推荐阅读

★半导体设备公司，赚大发了！

★台积电再上顶峰

★热闹的WiFi 6芯片赛道

半导体行业观察

『半导体第一垂直媒体』

实时 专业 原创 深度

识别二维码，回复下方关键词，阅读更多

晶圆｜集成电路｜设备｜汽车芯片｜存储｜MLCC｜英伟达｜模拟芯片

回复 **投稿**，看《如何成为“半导体行业观察”的一员 》

回复 **搜索**，还能轻松找到其他你感兴趣的文章！