GPU三巨头开启竞争新时代

https://mp.weixin.qq.com/s/m1efcK7voh_4VhU5GAJ7VQ

畅秋 Mon Aug 16 21:05

随着GPU在数据中心当中的应用越来越普及，原本用于CPU的MCM（多芯片封装模块）开始向GPU领域渗透，特别是在高性能计算领域，在业界受到了越来越多的关注。

近日，在AMD第二季度财报中，该公司确认了其具有第二代 CDNA 架构的下一代 Instinct MI200加速器的初始出货量。据悉，MI200配置了Aldebaran GPU，这是一个双芯片处理器，采用了MCM封装。该GPU开始出货，标志着AMD成为第一家向客户交付采用MCM技术的GPU公司，领先于竞争对手英特尔和英伟达（NVIDIA）。

何为MCM?

MCM是由同一器件中的多个Die（两个或更多）组成的电子封装系统。它安装在基板上，衬底上的管芯由导线连接。

与传统架构（如用于GPU的SLI和CrossFire）相比，MCM可提供更高的性能，并减小组件的尺寸。通过MCM封装系统，器件或模块可以克服重量和尺寸限制，并提供超过30%的效率。

MCM的优点可以概括如下：更高的可靠性；更灵活地集成不同的半导体技术；通过减少阵列之间互连的长度来提高性能；尺寸更小；产品可快速上市；降低了复杂性并简化了设计。

通常，MCM模块有3种类型，具体取决于基板技术：MCM层压（MCM-L）、沉积（MCM-D）和陶瓷（MCM-C）。

以前，MCM主要用于CPU，最近逐步进入GPU领域。

MCM用于GPU的优势和挑战

MCM GPU与传统带有多个GPU的显卡之间的最大区别在于，前者是一个单独的封装，后者是一个PCB板卡，前者的板载桥接器取代了两个独立显卡之间的Crossfire或SLI桥接器。

传统的SLI 和 CrossFire需要 PCIe 总线来交换数据、纹理、同步等。由于GPU之间的渲染时间会产生同步问题，因此在许多情况下，传统的双GPU显卡，即单个PCB上的两个芯片由它互连，每个芯片都有自己的VRAM。SLI或CrossFire的能耗很大，冷却也是一个挑战，这些在很长一段时间内都困扰着工程师。

MCM GPU就是为了解决以上问题而出现的。不过，MCM GPU并不完美，它同样面对着诸多技术挑战和难题。

在进行MCM GPU设计之前，需要解决封装和互连方面的软件问题，因为两个或更多GPU，无论多么紧密地连接在一起，要想在一起协同工作，并不是一件容易的事。MCM作为能够用于并行处理的组件，其GPU之间使用不同的内存访问，设计的复杂性会成倍增加。这需要开发人员在软件方面进行大量“修补”。在消费级的PC应用方面，很少有游戏玩家实际运行多GPU设置，因为其回报很少，因此没有人愿意做这么多的软件工作。不过，如果应用于数据中心和云计算，情况就不同了，这样的高性能计算应用对GPU提出了更高的要求。虽然多芯片GPU系统还是新生事物，许多图形工作负载不能很好地扩展（有些甚至根本不能扩展），但每台服务器有多个GPU，由于具有超级计算和数据中心的并行化性质，这就可以很好地扩展工作量。

而如果能解决MCM GPU的瓶颈问题，回报将是诱人的。这也正是MCM GPU首先出现在数据中心应用领域的主要原因，今后，随着技术的不断成熟，以及PC应用性能的提升，其在消费电子领域的应用也将会出现。

三强争霸

在企业界，最早应用MCM技术的是IBM，那是在上世纪70年代和80年代之间，主要用于该公司的POWER架构CPU。而将MCM发扬光大的是英特尔，自然也是用于CPU。2013年，该公司的22nm制程处理器Haswell就用到了该技术。2014年，14nm制程的Broadwell 架构问世，这是一个SoC平台，它使用了“堆叠”基板架构，也就是MCM，将多个阵列垂直堆叠在了一起。

最近几年，英特尔开始研发独立的GPU，也就是其Xe架构产品，为了顺应技术发展和应用需求，该公司开始将MCM应用于其最新的GPU产品，据悉是基于Xe HPC架构的Ponte Vecchio加速器，但具体问世时间还未确定。

AMD则快人一步。2020年，该公司把游戏卡与专业卡的GPU架构分家了，游戏卡的架构是RDNA，而专业卡的架构叫做CDNA，首款产品是Instinct MI100系列。今年6月，AMD首席执行官苏姿丰博士提到了CDNA 2架构及其产品，表示会在年内推出，不久前发布的Q2财报则确认CDNA 2 GPU已经向客户发货了。CDNA 2基于CDNA架构，是专为数据中心设计的。

近日，AMD更新了CDNA 2的说明，其GPU核心代号是Aldebaran，它会成为AMD第一款采用MCM多芯片封装的产品，也就是Instinct MI200。Aldebaran是AMD的第一款MCM GPU，但它是为数据中心准备的。在PC方面，2022年引入下一代RDNA 3架构后，基于MCM的消费级Radeon GPU也会出现。

据悉，采用MCM封装的CDNA 2内部将整合两个Die，每个芯片上有128组CU单元，如果每组CU还是128个流处理器的话，预计会拥有16384个流处理器，预计还会搭载128GB的HBM2e显存，而目前的Instinct MI100只有7680个流处理器，搭载32GB的HBM2显存。

制造多芯片计算 GPU 类似于制造多核 MCM CPU，例如Ryzen 5000或Threadripper处理器。首先，将芯片靠得更近可以提高计算效率。AMD 的 Infinity 架构确保了高性能互连，有望使两个芯片的效率接近一个的。其次，使用先进的工艺技术批量生产多个小芯片比大芯片更容易，因为小芯片通常缺陷较少，因此比大芯片的产量更好。

AMD 的合作伙伴HPE证实，即将推出的 Frontier 超级计算机将使用 AMD 代号为 Trento CPU（最有可能是具有额外缓存或其他增强功能的 Milan 版本）和 Instinct MI200 加速器，成为世界上最快的超级计算机，峰值性能为 1.5 ExaFLOPS。

除了AMD和英特尔，另一大GPU厂商英伟达也在摩拳擦掌，很可能紧随AMD之后推出其首款MCM GPU产品Hopper。

据悉，Hopper GPU架构是为数据中心应用专门设计的，与英伟达的Ampere架构产品不同，后者同时服务于 GPGPU（数据中心/工作站）和游戏市场。

早期的爆料称，Hopper由两个称为GPM 的GPU 模组构成，每个模组有144 个SM 单元，同时 Hopper 由于是专为运算所规划的架构，相较Ampere 应该会取消用于光线追踪加速的RT Core ，并强化包括FP64 、 FP16 与Tensor Core 等运算与AI 技术会使用到的单元。

据悉，Hopper GPU 将采用台积电的5nm制程工艺，性能比 Ampere 提高 3 倍。这是一个很大的提升，具体情况如何，还要看今后爆出的更多关于Hopper的信息。

有报道称，Hopper GPU很快就会流片。

据悉，推出Hopper GPU之后，英伟达还将推出Ampere Next 和 Ampere Next Next，它们将采用MCM封装。Ampere Next GPU 预计在 2022 年推出，而 Ampere Next Next 将在 2024 年推出。

结语

MCM的自身特点使其在高性能计算领域如鱼得水，不只是是CPU，如今在GPU领域也得到了拓展，而随着数据中心、边缘云、物联网的发展，以及CPU、GPU、DPU等产品形态的日益增多和复杂，留给MCM的发展空间可能会越来越大。

新技术产品和应用的发展给以MCM为代表的芯片封装、整合技术提供了更多的想象空间。

今天是《半导体行业观察》为您分享的第2769内容，欢迎关注。

推荐阅读

★5G R17带来的芯片机会

★SiC迎来“上车”时刻？

★芯片制造利润窜升，各家争先恐后再添一把火

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！