英伟达与AMD的巅峰之战

https://mp.weixin.qq.com/s/8V8709gtVv_KJ0X7f5z6Tg

techspot Mon Dec, 14 07:15

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

对于GPU爱好者来说，这是一个漫长的等待。英伟达将Turing产品线维持了两年，然后在2020年9月用Ampere取代了它。AMD更友善一点，他们的新设计间隔了15个月，但大多数人对此并不感兴趣。

他们希望看到的是AMD推出一款高端机型，与英伟达(Nvidia)最优秀的产品展开正面竞争。他们做到了，现在我们已经看到了结果，在花钱买最好的图形卡时，PC游戏玩家现在（在理论上）有了很多选择。

但是驱动它们的芯片呢?其中一个从根本上来说比另一个好吗?继续读下去，看看Ampere和RDNA 2是如何决一死战的!

Nvidia衰退，AMD成长

节点和die尺寸

多年来，高端GPU一直比CPU大得多，而且它们的尺寸一直在稳步增长。AMD最新推出的Navi芯片面积约为520mm^2，是之前Navi芯片的两倍多。不过，这并不是他们最大的——这项荣誉颁给了他们的Instinct MI100加速器（约750 mm^2）中的GPU。

上一次AMD制造的接近Navi 21大小的游戏处理器是为Radeon R9 Fury和Nano显卡设计的，这两款产品在Fiji芯片上采用了GCN 3.0架构。它的裸片面积为596 mm^2，但它是在台积电的28HP工艺节点上生产的。

自2018年以来，AMD一直在使用台积电更小的N7工艺，该生产线生产的最大芯片是Vega 20 (Radeon VII)，面积为331mm^2。他们所有的Navi gpu都是在略微升级的N7P处理器上制作的，所以可以比较这些产品。

Radeon R9 Nano：微型卡，大型GPU

但说到纯粹的die尺寸，英伟达拿下了王冠，并不是说这一定是件好事。最新的基于Ampere的芯片，GA102，是628mm^2。这实际上比它的前身TU102小了17%——GPU面积达到惊人的754mm^2。

与Nvidia巨大的GA100芯片(用于AI和数据中心)相比，这两款芯片的尺寸都相形见绌，其GPU为826 mm^2，采用的是台积电的N7芯片。虽然它从来没有被设计用来驱动桌面显卡，但它确实显示了GPU制造的可能规模。

把它们放在一起突出了Nvidia最大的GPU有多大。Navi 21看起来相当苗条，尽管处理器的功能不仅仅是芯片区。GA102封装了283亿个晶体管，而AMD的新芯片减少了5%，达到268亿个。

我们不知道每个GPU构建多少层，因此我们所能比较的是晶体管与die面积的比率，通常称为die密度。Navi 21的晶体管约为每平方毫米5150万个晶体管，但GA102明显低于41.1，这可能是Nvidia的芯片堆叠程度比AMD的略高，但它更可能表示工艺节点。

如前所述，Navi 21是由台积电生产的，采用N7P生产方法，性能比N7略有提高;但在新产品GA102上，英伟达求助于三星来完成生产任务。这家韩国半导体巨头正在使用他们所谓的8nm节点(标记为8N或8NN)的改良版本，专门为Nvidia设计。

这些节点值，7和8，与芯片组件的实际尺寸没有多大关系:它们只是市场营销术语，用于区分不同的生产技术。也就是说，即使GA102比Navi 21有更多的层，die尺寸确实有一个特殊的影响。

一台300毫米(12英寸)的晶圆片正在台积电的制造工厂进行测试。

微处理器和其他芯片是由高度精炼的硅和其他材料制成的大圆盘，称为晶圆。台积电和三星为AMD和Nvidia使用的是300毫米晶圆，相对于更大的die，使用更小的die，每块晶圆将产生更多的芯片。

这种差异不可能很大，但是在降低制造成本方面，当每片晶圆的生产成本达到数千美元时，AMD相对于Nvidia而言优势较小。 当然，这是假设三星或台积电没有与AMD / Nvidia进行某种财务交易。

如果芯片本身不能很好地完成设计工作，那么所有这些die尺寸和晶体管数量都将是徒劳的。 因此，让我们深入研究每个新GPU的布局，看看它们背后的东西。

剖析die

Ampere GA102和RDNA 2 Navi 21的总体架构

我们从分析Ampere GA102和RDNA 2 Navi 21 GPU的总体架构开始我们对架构的探索——这些图表不一定向我们展示所有的物理布局，但它们给出了处理器有多少组件的明确指示。

在这两种情况下，布局都是非常熟悉的，因为它们基本上都是其前身的扩展版本。在处理指令中添加更多的单元将始终提高GPU的性能，因为在最新的3D大片中，在高分辨率下，渲染工作量涉及大量的并行计算。

这样的图表是有用的，但是对于这个特定的分析来说，更有趣的是看看各个组件在GPU中的位置。在设计大型处理器时，您通常希望共享资源（如控制器和缓存）位于中心位置，以确保每个组件都具有相同的路径。

接口系统，如本地内存控制器或视频输出，应该安装在芯片的边缘，以便更容易地将它们连接到连接GPU和显卡其余部分的数千根单独的电线上。

以下是AMD的Navi 21和Nvidia的GA102 die的伪彩色图像。 它们实际上只显示了芯片中的一层；但它们确实给我们提供了一个现代GPU内部的极好视图。

两种设计之间最明显的区别在于，Nvidia在芯片布局上没有遵循集中化的方法——所有的系统控制器和主缓存都在底部，逻辑单元以长列形式运行。他们过去也这样做过，但只针对中低端机型。

例如，Pascal GP106（用于GeForce GTX 1060等）实际上是GP104（来自GeForce GTX 1070）的一半。 后者是较大的芯片，其缓存和控制器位于中间。 这些都移到了它的兄弟姐妹那一边，但这只是因为设计已经被拆分了。

Pascal GP104和GP106 资料来源：Fritzchens Fritz

对于之前所有的高端GPU布局，Nvidia都使用了经典的集中式结构。为什么这里会有变化呢?这不可能是由于接口的原因，因为内存控制器和PCI Express系统都运行在die的边缘。

这也不是出于热学原因，因为即使die 的缓存/控制器部分比逻辑部分的温度更高，您仍然希望在其中间具有更多的硅以帮助吸收和散发热量 。尽管我们不能完全确定更改的原因，但我们怀疑这与Nvidia对芯片中ROP（渲染输出）单元实施的更改有关。

我们将在后面更详细地讨论它们，但是现在让我们说，虽然布局的改变看起来很奇怪，但它不会对性能产生显著的影响。这是因为3D渲染充斥着许多长时间的延迟，通常是由于必须等待数据。因此，由于一些逻辑单元比其他逻辑单元离缓存更远而增加的纳秒数，都被隐藏在了整个系统中。

在我们继续之前，值得注意的是AMD在Navi 21布局中实施的工程改变，与驱动类似Radeon rx5700 XT的Navi 10相比。尽管新芯片在面积和晶体管数量上都比之前的芯片大了一倍，但设计者还设法在不显著增加功耗的情况下提高了时钟速度。

例如，Radeon RX 6800 XT运动的基时钟和升压时钟分别为1825和2250mhz, TDP为300 W;Radeon RX 5700 XT的相同性能为1605 MHz、1905 MHz和225 W。英伟达也通过Ampere提高了时钟速度，但部分原因是使用了更小、更高效的进程节点。

我们对Ampere和RDNA 2显卡的每瓦特性能检查显示，两家供应商在这方面都取得了显著的改进，但AMD和台积电取得了一些相当显著的成就——比较上图中Radeon RX 6800和Radeon VII之间的差异。

后者是他们首次使用N7节点进行GPU合作，并且在不到两年的时间内，他们将每瓦性能提高了64%。的确，如果英伟达继续与台积电合作，那Ampere GA102的性能会好得多。

管理GPU工厂

芯片内部的一切组织方式

当涉及到指令处理和数据传输管理时，Ampere和RDNA2都遵循类似的模式来组织芯片内部的一切。游戏开发人员使用图形API编写标题，以制作所有图像；它可能是Direct3D、OpenGL或Vulkan。这些基本上是软件库，充满了规则、结构和简化指令的“书籍”。

AMD和Nvidia为他们的芯片创建的驱动程序本质上起着翻译的作用:将通过API发布的例程转换为GPU能够理解的操作序列。在那之后，就完全由硬件来管理了，比如什么指令首先执行，芯片的哪个部分执行这些指令，等等。

指令管理的初始阶段由合理地集中在芯片中的一组单元处理。在RDNA 2中，图形和计算着色器通过单独的管线进行路由，这些管线将指令调度并分派到芯片的其余部分。前者称为图形命令处理器，后者是异步计算引擎（ACE）。

Nvidia只是用一个名字来描述他们的一组管理单元，即GigaThread Engine，在Ampere中它执行与RDNA 2相同的任务，尽管Nvidia并未过多说明其实际管理方式。总之，这些命令处理器的功能类似于工厂的生产经理。

GPU通过并行执行所有操作来获得性能，因此在整个芯片上复制了下一个组织层次。坚持工厂的类比，这类似于一家拥有中央办公室但在多个地点生产商品的企业。

AMD使用标签着色器引擎（SE），而Nvidia则称其为图形处理集群（GPC）-不同的名称，相同的角色。
对芯片进行这种分区的原因很简单：命令处理单元不能处理所有事情，因为它最终会变得过于庞大和复杂。因此，将一些日程安排和组织职责进一步向下推进是有意义的。这也意味着每个分离分区可以完全独立于其他分区执行某些操作，因此一个分区可以处理大量的图形着色器，而其他分区则在处理长而复杂的计算着色器。
在RDNA 2的例子中，每个SE都有自己一套固定的功能单元:被设计用来完成一项特定任务的电路，程序员通常无法对其进行大量调整。

- mitive Setup unit——获取顶点，准备好进行处理，同时生成更多的顶点(essellation)并将其剔除
- Rasterizer——将三角形的3D世界转换为像素的2D网格
- Render Outputs(ROPs)——读取、写入和混合像素

原始的设置单元以每个时钟周期1个三角形的速率运行。这听起来可能不是很多，但是不要忘记这些芯片运行在1.8到2.2 GHz之间，所以原始的设置不应该成为GPU的瓶颈。对Ampere来说，原始单位是在组织的下一层找到的，我们很快就会讲到。
AMD和Nvidia都没有过多提及他们的光栅化器。后者称为光栅引擎，我们知道它们每个时钟周期处理一个三角形，并输出若干像素，但没有进一步的信息，例如它们的亚像素精度。
Navi 21芯片中的每个SE都有4组8个ROP，总共产生128个渲染输出单元；Nvidia的GA102每GPC包含2组8个ROP，因此整个芯片可运动112个ROP。这看起来AMD在这方面有优势，因为更多的ROP意味着每个时钟可以处理更多的像素。但是这样的单元需要对缓存和本地内存的良好访问，我们将在本文后面详细介绍。现在，让我们继续研究SE/GPC分区是如何进一步划分的。
AMD的着色引擎被划分为双计算单元（DCU），Navi 21芯片本身就有10个DCU——请注意，在一些文档中，它们也被归类为工作组处理器（WGP）。在Ampere和GA102的例子中，它们被称为纹理处理簇（TPC），每个GPU包含6个tpc。Nvidia设计的每一个集群都有一个叫做“变形引擎”的东西——本质上是Ampere的原始设置单元。
Nvidia也以每时钟1个三角形的速度运行，尽管Nvidia的GPU比AMD的低，但他们的TPC数量比Navi 21的SE要多得多。因此，对于相同的时钟速度，GA102应该有一个显著的优势，因为完整的芯片拥有42个原始设置单元，而AMD的新RDNA 2只有4个。但由于每个光栅引擎有6个TPC，GA102实际上有7个完整的原始系统，而Navi 21有4个。由于后者的时钟并没有比前者高75%，当涉及到几何处理(尽管没有游戏可能在这方面受到限制)时，似乎英伟达在这方面具有明显的领先优势。
芯片组织的最后一层是RDNA 2中的计算单元（CU）和Ampere中的流式多处理器（SM），这是我们GPU工厂的生产线。

这些是图形处理器馅饼中的肉和蔬菜，因为它们拥有所有用于处理图形、计算和现在的光线追踪着色器的高度可编程单元。正如你在上图中看到的，每一个芯片都只占整个芯片空间的很小一部分，但是它们仍然是非常复杂的，并且对芯片的整体性能非常重要。

到目前为止，在两个GPU的布局和组织方式方面，还没有什么真正的突破性协议。术语全都不同，但是它们的功能却大同小异。而且由于它们所做的很多事情都受可编程性和灵活性的限制，因此一个相对于另一个所具有的任何优势，都只能归结为规模感，即哪个拥有最大的特色。

但是对于CU和SM，AMD和Nvidia采取了不同的方式来处理着色器。在某些领域，它们有很多共同点，但在其他许多领域则并非如此。

计数核心是Nvidia的方式

由于安培（Ampere）在RDNA 2之前就冒险进入野外，我们首先来看看Nvidia的SM。现在没有必要查看裸片本身的图像，因为它们无法准确告诉我们其中的内容，因此让我们使用组织图。这些不应该代表芯片中各种组件的物理排列方式，而只是每种类型中存在多少种。

图灵对其台式机前身Pascal进行了实质性更改（去掉了一堆FP64单元和寄存器，但是增加了张量核和光线跟踪），而Ampere实际上是一个相当温和的更新-至少从表面上看。不过，就Nvidia的市场部门而言，新设计使每个SM中CUDA内核的数量增加了一倍以上。

在图灵中，流多处理器包含四个分区（有时称为处理块），每个分区中容纳16个INT32和16x FP32逻辑单元。这些电路旨在对32位数据值执行非常具体的数学运算：INT单位处理整数，而FP单位处理浮点数（即十进制）。

英伟达表示，一个Ampere SM总共有128个CUDA内核，但严格来说，这是不正确的-或者，如果我们必须坚持这一点，那么图灵（Turing）也是如此。该芯片中的INT32单元实际上可以处理浮点值，但只能以非常少量的简单操作进行。对于Ampere，Nvidia已开放了它们支持的浮点数学运算范围，以匹配其他FP32单元。这意味着每个SM的CUDA内核总数并没有真正改变。只是其中的一半现在拥有更多功能。

每个SM分区中的所有内核都可以随时处理同一条指令，但是由于INT / FP单元可以独立运行，因此Ampere SM每个周期最多可以处理128x FP32计算，或一起处理64x FP32和64x INT32操作。而图灵只是后者。

因此，新的GPU可能使FP32的输出量比其上一代产品大一倍。对于计算工作负载，尤其是在专业应用程序中，这是向前迈出的一大步。但是对于游戏而言，优势却远远没有达到预期。当我们首次测试GeForce RTX 3080时，这一点很明显，它使用启用了68个SM的GA102芯片。

尽管FP32的峰值吞吐量比GeForce 2080 Ti高出121％，但平均帧速率仅提高了31％。那么，为什么所有这些计算能力都会浪费掉呢？一个简单的答案是，游戏并非一直在运行FP32指令。

当Nvidia在2018年发布Turing时，他们指出， GPU处理的指令平均约有36％涉及INT32例程。这些计算通常用于计算内存地址，两个值之间的比较以及逻辑流/控制。

因此，对于这些操作，双速率FP32功能不起作用，因为具有两个数据路径的单元只能执行整数或浮点运算。而且，只有在当时由它处理的所有32个线程都排队处理相同的FP32操作时，SM分区才会切换到此模式。在所有其他情况下，安培中的分区与图灵中的分区一样运行。

这意味着在INT + FP模式下运行时，GeForce RTX 3080之类的FP32仅比2080 Ti具有11％的FP32优势。这就是为什么在游戏中看到的实际性能提升没有原始数据所预期的那么高的原因。
至于其他改进。每个SM分区的Tensor Core更少，但每个都比Turing中的功能强大得多。这些电路执行非常具体的计算（例如将两个FP16值相乘并用另一个FP16编号累加答案），每个内核现在每个周期执行32次这些操作。
它们还支持一种名为“细粒度结构稀疏性”的新特性，在不涉及所有细节的情况下，这意味着通过剔除那些对答案没有影响的数据，计算率可以翻倍。同样，这对于从事神经网络和人工智能工作的专业人员来说是个好消息，但目前对游戏开发者来说并没有什么明显的好处。
光线跟踪核心也已进行了调整：它们现在可以独立于CUDA核心工作，因此，在进行BVH遍历或光线原始相交数学时，SM的其余部分仍可以处理着色器。处理射线是否与原语相交测试的RT核心的部分性能也增加了一倍。
RT内核还具有附加的硬件，可帮助将光线跟踪应用于运动模糊，但是此功能目前仅通过Nvidia专有的Optix API公开。
还有其他一些调整，但是整体方法是明智但稳定的演进之一，而不是主要的新设计。但是考虑到图灵的原始功能并没有什么特别的错误，因此看到这一点不足为奇。
那么AMD怎么办-他们对RDNA 2中的计算单元做了什么？

追寻美妙的光线

从表面上看，AMD在计算单元方面并没有太大变化-它们仍然包含两组SIMD32向量单元，一个SISD标量单元，纹理单元以及各种缓存堆栈。关于它们可以执行的数据类型和相关的数学运算，已经发生了一些变化，我们稍后将详细介绍。对于普通消费者而言，最明显的变化是AMD现在为光线跟踪中的特定例程提供了硬件加速。
CU的这部分执行ray-box或ray-triangle交叉检查——与安培中的RT内核相同。然而，后者也加速了BVH遍历算法，而在RDNA 2，这是通过使用SIMD 32单元计算着色器来完成的。
不管一个着色器内核有多少个，或者它们的时钟速率有多高，使用设计为仅完成一项工作的定制电路总是比通用方法更好。这就是为什么首先发明GPU的原因：渲染世界中的所有事物都可以使用CPU来完成，但是它们的通用性使其不适合于此。
RA单元紧邻纹理处理器，因为它们实际上是同一结构的一部分。早在2019年7月，我们就报道了AMD申请的一项专利的内容，该专利使用``混合"方法详细处理了光线追踪中的关键算法...
尽管该系统确实提供了更大的灵活性，并且消除了当存在光线追踪工作量时裸片的一部分不做任何事情的需求，但AMD的第一个实现确实有一些缺点。最值得注意的是，纹理处理器在任何时候都只能处理涉及纹理或ray-primitive交点的操作。
鉴于Nvidia的RT核心现在完全独立于SM的其余部分而运行，与RNDA 2相比，在通过光线跟踪所需的加速结构和交叉测试进行磨削时，这似乎给Ampere带来了明显的领先优势。
尽管我们仅简要检查了AMD最新图形卡中的光线追踪性能，但到目前为止，我们确实发现使用光线追踪的影响很大程度上取决于所玩的游戏。

例如，在Gears 5中，Radeon RX 6800（使用Navi 21 GPU的60 CU变体）仅降低了17％的帧速率，而在《古墓丽影》的阴影中，平均损失达到52％。相比之下，英伟达的RTX 3080（使用68 SM GA102）在这两款游戏中的平均帧率损失分别为23％和40％。

需要对射线追踪进行更详细的分析来说明AMD的实现，但是作为该技术的第一个迭代，它看起来很有竞争力，但对应用程序正在进行的射线追踪很敏感。

如前所述，RDNA 2中的计算单元现在支持更多数据类型。最值得注意的是低精度数据类型，例如INT4和INT8。它们用于机器学习算法中的张量运算，而AMD具有用于AI和数据中心的单独架构（CDNA），但此更新适用于DirectML。

该API是Microsoft DirectX 12家族的最新成员，硬件和软件的组合将为光线跟踪和时间放大算法中的降噪提供更好的加速。对于后者，Nvidia当然拥有自己的名称，称为DLSS。他们的系统使用SM中的Tensor核心执行部分计算，但是鉴于可以通过DirectML构建类似的过程，因此这些单元似乎有些多余。但是，在Turing和Ampere中，Tensor核心还可以处理所有涉及FP16数据格式的数学运算。

对于RDNA 2，此类计算是使用着色器单元，使用打包数据格式完成的，即每个32位向量寄存器都包含两个16位寄存器。那么哪种方法更好呢？AMD将其SIMD32单元标记为矢量处理器，因为它们能针对多个数据值发出一条指令。

每个向量单元包含32个流处理器，由于每个流处理器只处理单个数据片段，因此实际操作本身是标量的。这本质上与安培中的SM分区相同，其中每个处理块还针对32个数据值携带一条指令。

但是，在Nvidia设计中的整个SM每个周期最多可以处理128个FP32 FMA计算（融合乘加）时，单个RDNA 2计算单元只能处理64个。在执行标准FP16数学时，使用FP16可以将其提高到每个周期128 FMA，这与Ampere的Tensor核心是一样的。

Nvidia的SM可以处理指令时可以同时处理整数和浮点值（例如64 FP32和64 INT32），并且具有用于FP16操作，张量数学和光线跟踪例程的独立单元。尽管AMD CU具有独立的支持简单整数数学的标量单元，但它们在SIMD32单元上承担了大部分工作量。

因此，安培似乎在这方面有优势:GA102比Navi 21拥有更多的CU，而且在峰值吞吐量、灵活性和提供的功能方面，它们的表现更出色。但是AMD有一个相当不错的锦囊妙计。

内存系统，多层缓存

拥有成千上万个逻辑单元的GPU，在复杂的数学运算中一路高歌猛进，这一切都很好，但是但如果他们不能足够快地提供所需的指令和数据，他们将在海量的数据中挣扎。这两种设计都拥有丰富的多级缓存，拥有巨大的带宽。

让我们来看看安培的。总体而言，内部发生了一些显着变化。2级缓存的数量增加了50％（图灵TU102的运动速度分别为4096 kB），并且每个SM中的1级缓存的大小都增加了一倍。

与以前一样，就可以为数据，纹理或一般计算用途分配多少缓存空间而言，Ampere的L1缓存是可配置的。但是，对于图形着色器（例如顶点，像素）和异步计算，缓存实际上设置为:

- 64 kB用于数据和纹理
- 48 kB用于共享通用内存

- 16 kB保留用于特定操作

只有在完全计算模式下运行时，L1才可以完全配置。从好的方面来说，可用带宽的数量也增加了一倍，因为缓存现在可以每个时钟读取/写入128个字节（尽管没有关于延迟是否得到改善的消息）。
内部存储器系统的其余部分在Ampere中保持不变，但是当我们仅移至GPU外部时，对于我们来说，这是一个很好的惊喜。Nvidia与DRAM制造商美光合作，将GDDR6的修改版本用于其本地内存需求。从本质上讲，它仍然是GDDR6，但数据总线已被完全替换。GDDR6X使用四个电压，而不是使用传统的每个引脚设置1位（信号只是在两个电压（又名PAM）之间快速反弹）设置的方式：
进行此更改后，GDDR6X每个周期每个引脚有效传输2位数据-因此，对于相同的时钟速度和引脚数，带宽增加了一倍。GeForce RTX 3090具有24个GDDR6X模块，它们以单通道模式运行，额定速率为19 Gbps，提供的峰值传输带宽为936 GB / s。
与GeForce RTX 2080 Ti相比，这增加了52％，并且不能轻易忽视。过去仅通过使用类似HBM2的方式获得了这样的带宽数字，与GDDR6相比，HBM2的实现成本很高。
但是，只有Micron可以制造这种存储器，而PAM4的使用为生产过程增加了额外的复杂性，对信号的公差要严格得多。AMD走了一条不同的道路-他们没有寻求外部机构的帮助，而是利用其CPU部门为桌面带来了新的东西。与它的前代产品相比，RDNA 2中的整个内存系统没有太大变化，只有两个主要变化。
每个着色器引擎现在都有两组Level 1缓存，但是由于它们现在具有两组Dual Compute Unit（RDNA拥有一组），因此这种改变是可以预期的。但是将128 MB的3级缓存缓存到GPU中吗？这让很多人感到惊讶。利用其EPYC系列Zen 2服务器芯片中的L3高速缓存的SRAM设计，AMD在该芯片中嵌入了两组64 MB高密度高速缓存。数据事务由16组接口处理，每个接口每个时钟周期移位64个字节。
所谓的无限缓存具有其自己的时钟域，并且可以在1.94 GHz上运行，从而提供1986.6 GB / s的内部峰值传输带宽。而且由于它不是外部DRAM，因此涉及的延迟非常低。这种高速缓存非常适合存储光线跟踪加速结构，并且由于BVH遍历涉及大量数据检查，因此Infinity高速缓存应特别为此提供帮助。
两个64 MB的Infinity缓存条和Infinity Fabric系统
目前，尚不清楚RDNA 2中的3级缓存是否以与Zen 2 CPU中相同的方式运行：即作为2级victim缓存。通常，当需要清除最后一级的高速缓存以为新数据腾出空间时，对该信息的任何新请求都必须发送到DRAM。
victim缓存存储的数据已经被标记为要从下一层内存移除，并且有128MB的数据可用，Infinity缓存可能存储32个完整的L2缓存集。该系统的结果是，在GDDR6控制器和DRAM上放置的需求更少。
AMD的较旧GPU设计一直在缺乏内部带宽的情况下苦苦挣扎，尤其是当它们的时钟速度提高后，但是额外的缓存将使该问题逐渐消失。

那么，哪种设计更好呢？GDDR6X的使用为GA102提供了到本地内存的巨大带宽，并且更大的缓存将有助于减少缓存未命中（这会使线程的处理停滞）的影响。Navi 21的大型3级缓存意味着DRAM不必经常被窃听，并利用了以更高的时钟速度运行GPU的能力，而不会出现数据匮乏的情况。

AMD决定坚持使用GDDR6的决定意味着第三方供应商可以使用更多的内存，同时任何制造GeForce RTX 3080或3090的公司都必须使用美光。尽管GDDR6有多种模块密度，但GDDR6X当前限于8 Gb。

RDNA 2中的缓存系统可以说是比Ampere中使用的缓存系统更好的方法，因为与外部DRAM无关，使用多个级别的片上SRAM始终比外部DRAM提供更低的延迟和更好的性能（在给定的功率范围内）。

GPU的来龙去脉

渲染管线

这两种架构都对渲染管线的前端和后端进行了大量更新。DirectX12 Ultimate中的Ampere和RDNA 2完全具有运动型网格着色器和可变速率着色器，但Nvidia的芯片具有更出色的几何性能，而这要归功于Nvidia用了更多的任务处理器。

尽管使用网格着色器可以使开发人员创建更加逼真的环境，但没有一款游戏的性能会完全以来于渲染过程中的这个阶段。因为大部分最难的工作是在像素或光线跟踪阶段。

这就是使用可变率着色器发挥作用的地方——基本上该过程涉及在一个像素块而不是单个像素上应用照明和颜色着色器。这类似于为了提高性能而降低游戏的分辨率，但由于它只能应用于选定的区域，因此视觉质量的损失并不明显。

无论是否使用可变速率着色器，这两种体系结构都已更新了其渲染输出单元（ROP），这将提高在高分辨率下的性能。在所有以前的GPU中，Nvidia都将ROPs与内存控制器和二级缓存绑定在一起。

在Turing中，8个ROP单元(统称为一个分区)会直接链接到一个控制器和一个512kb的高速缓存上。添加更多的ROP会带来问题，因为它需要更多的控制器和缓存，因此对于Ampere而言，现在ROP已完全分配给了GPC。GA102每个GPC拥有12个ROP（每个时钟周期处理1个像素），整个芯片共有112个单位。

AMD采用了与Nvidia的旧方法类似的系统（即与内存控制器和L2缓存片绑定），尽管它们的ROP主要使用1级缓存进行像素读/写和混合。在Navi 21芯片中，已经为它们提供了急需的更新，并且每个ROP分区现在每个周期以32位颜色处理8个像素，并以64位处理4个像素。

Nvidia还为Ampere带来了RTX IO，这是一种数据处理系统，可以让GPU直接访问存储驱动器，复制所需的数据，然后使用CUDA内核解压。但是，目前该系统不能在任何游戏中使用，因为Nvidia正在使用DirectStorage API（另一种DirectX12增强功能）来控制它，并且尚未准备好公开发布。

目前使用的方法包括让CPU管理所有这一切：它接收来自GPU驱动程序的数据请求，将数据从存储驱动器复制到系统内存中，进行解压缩，然后再复制到图形卡的DRAM中。

除了涉及大量浪费的复制之外，这种机制在本质上是串行的——CPU一次只能处理一个请求。Nvidia声称可以达到“100倍的数据吞吐量”和“20倍的CPU利用率低”，但是在系统能够在现实世界中测试之前，并不能证明它能够实现这种效果。

当AMD推出RDNA 2和新的Radeon RX 6000图形卡时，他们推出了称为Smart Access Memory的产品。这不是他们对Nvidia的RTX IO的答案——实际上，它甚至不是真正的新功能。默认情况下，每个单独的访问请求中，CPU中的PCI Express控制器最多可以寻址256 MB的图形卡内存。此值由基址寄存器（BAR）的大小设置，并且早在2008年，PCI Express 2.0规范中就有一项可选功能，可以调整其大小。这样做的好处是，只需处理较少的访问请求即可访问整个卡的DRAM。

该功能需要操作系统，CPU，主板，GPU及其驱动程序的支持。当前，在Windows PC上，系统仅限于Ryzen 5000 CPU，500系列主板和Radeon RX 6000图形卡的特定组合。

测试时，这个简单的功能给出了一些令人吃惊的结果——在4K环境下将性能提升15%是不容小觑的，所以英伟达表示他们将在不久的将来为RTX 3000范围提供这个特性也就不足为奇了。是否可调整大小的BAR支持是否适用于其他平台组合还有待观察，但是它的使用无疑是受欢迎的，即使它不是Ampere / RDNA 2的体系结构功能。

视频取代了广播

多媒体引擎，视频输出

GPU世界通常由核心访问量、TFLOPS、GB/s和一些其他指标为主导，由于YouTube内容创造者和直播游戏流的崛起，使得市场开始注意GPU的显示和多媒体引擎的能力。

随着支持此类功能的显示器价格的下降，对所有分辨率下的超高刷新率的需求也在增长。两年前，一台144 Hz 4K 27”的HDR显示器要花你2000美元;今天，你可以用几乎一半的价格买到类似的东西。

两种架构均通过HDMI 2.1和DisplayPort 1.4a提供显示输出。前者提供了更多的信号带宽，但是在HDR和240 Hz时，它们的额定频率均为4K，在60 Hz时，其额定值为8K。这是通过使用4：2：0色度二次采样或DSC 1.2a实现的。这些是视频信号压缩算法，可显着减少带宽需求，而不会损失太多的视觉质量。如果没有它们，即使HDMI 2.1的峰值带宽为6gb /s，也不足以以6hz的速率传输4K图像。

48英寸LG CK OLED'显示器'-120 Hz时的4K需要HDMI 2.1

Ampere和RDNA 2还支持可变刷新率系统（用于AMD的FreeSync，用于Nvidia的G-Sync），在视频信号的编码和解码方面，也没有明显的区别。

无论您使用哪种处理器，您都会发现对8K AV1、4K H.264和8K H.265解码的支持，尽管它们在这种情况下的性能究竟如何还没有得到彻底的研究。两家公司都没有详细说明他们的显示和多媒体引擎的内部结构。尽管它们很重要，但GPU的其余部分依旧值得关注。

不同的策略

为计算而生，还是为游戏而生

GPU历史的爱好者将知道AMD和Nvidia过去在架构选择和配置上采用了截然不同的方法。但是，随着3D图形越来越受到计算世界和API的同质化的支配，它们的总体设计也越来越相似。如今的游戏并不是以渲染需求为架构定下基调，GPU产业已经扩展到的市场领域才是引领方向。在撰写本文时，Nvidia有三种使用Ampere 技术的芯片:GA100、GA102和GA104。

GA104可以在GeForce RTX 3060 Ti中找到

最后一个只是GA102的精简版——每个GPC拥有的TPC更少（整体GPU更少），而二级缓存则只有三分之二。其他所有内容都完全相同。而GA100则是与此完全不同的产品。

GA100没有RT核，也没有INT32 + FP32支持的CUDA核。相反，它打包了许多额外的FP64单元，更多的加载/存储系统以及大量的L1 / L2缓存。它也没有显示器或多媒体引擎。这是因为它完全是为用于AI和数据分析的大规模计算集群而设计的。

从应用场景上看，GA102/104需要覆盖Nvidia瞄准市场是:游戏爱好者、专业图形艺术家和工程师，以及小规模的人工智能和计算工作。Ampere则想要成为“万事通”，并且能够精通所有行业，但这可不是一件容易的事。

750平方毫米的Arcturus CDNA

RDNA 2专为PC和游戏机上的游戏而设计，尽管它可以转向与Ampere相同的应用市场。然而，AMD选择继续他们的GCN架构，并更新它，以满足今天的专业客户的需求。

RDNA 2产生了“ Big Navi”，而CDNA产生了“ Big Vega”-Instinct MI100装有Arcturus芯片，这是一个拥有128个计算单元的500亿晶体管GPU。与Nvidia的GA100一样，它也不包含显示器或多媒体引擎。

尽管英伟达凭借Quadro和特斯拉(Tesla)等在专业市场占据了主导地位，但Navi 21之类的产品并非旨在与之抗衡，而是进行了相应的设计。这是否会使RDNA 2成为更好的结构体系; Ampere 适应多种市场的要求是否在限制了它的发展?

当您查看证据时，答案似乎是：不。

AMD将很快发布Radeon RX 6900 XT，它使用了完整的Navi 21(没有禁用CUs)，其性能可能与GeForce RTX 3090或更好。但是那张卡上的GA102也没有被完全释放，所以Nvidia总是可以选择升级到一个“超级”版本，就像他们去年对 Turing所做的那样。

可能有人会说，因为RDNA 2被用在Xbox系列X/S和PlayStation 5中，游戏开发者会倾向于将这种架构用于他们的游戏引擎。但是，您只需要查看在Xbox One和PlayStation 4中使用了GCN的时间，便可以了解这种情况如何发挥作用。

前者在2013年的第一次发布中使用了基于GCN 1.0架构的GPU——这种设计直到第二年才出现在桌面PC图形卡中。2017年发布的Xbox One X使用的是GCN 2.0，这是一个已经使用了3年多的成熟设计。

那么，所有为Xbox One或PS4制作的游戏，只要移植到PC上，就能在AMD显卡上运行得更好吗?实际上，并没有。因此，尽管RDNA 2具有令人印象深刻的功能，但我们不能认为该产品会与RDNA 2有所不同。

但这些最终都无关紧要，因为这两种GPU设计都具有非凡的能力，是半导体制造领域的奇迹。英伟达和AMD带来了不同的工具，因为他们都在试图解决不同的问题;Ampere的目标是面向所有人，RDNA 2主要是关于游戏的。
这一次，战斗陷入了僵局，尽管双方都可以在特定的一两个区域宣告胜利。GPU之战将持续到明年，一个新的竞争者将加入这场战斗:英特尔的Xe系列芯片。在不久的将来，我们就能看到这场战斗的结果了!
★ 点击文末【阅读原文】，可查看本文原链接。

*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2521内容，欢迎关注。

推荐阅读

★IDM：我太难了！

★“逃离”英伟达

★3D闪存，176层了！

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

存储｜晶圆｜华为｜FPGA｜英特尔｜高通｜射频｜封测

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!

点击阅读原文，可查看本文
原文链接!