第 3 章

图形处理器

3.1 图形处理器的定义

GPU(graphic processing unit,图形处理器)通常称为显卡,就像大脑对于一个人的重要性一样,GPU 是显卡的大脑,是显卡中最为重要的部件,它决定了显卡的档次和大部分的性能。GPU 拥有上百颗甚至上万颗运算核心,比如 RTX2080Ti 的 GPU 的核心数量多达 4352 个,相比于 CPU 只有个位数的核心数,GPU 的运算能力极强。GPU 类似于流水线,几千个工人干着无差别的劳动,他们不需要统一调配、互相牵制合作,而是每个人只做自己手头上的工作,每个运算核心计算着自己的工作,不用担心计算结果的调配、计算结果的转移等复杂操作。所以,GPU 的主要功能是并行运算,帮助我们做一些计算量大、重复量大的工作,例如图形渲染等。GPU 的目的是协助 CPU 完成高密度的复杂任务。

在图 3-1 中可以看到,对于 GPU 来说,它的控制单元 Control(黄色部分)与缓存单元 Cache(红色部分)比例较小,而运算单元 ALU(绿色)部分比例最多。GPU 采用了流式并行计算的模式,对每个数据在 ALU 中进行独立的并行计算,不依赖于其他同类型数据。举一个简单的例子,在一张图的渲染中,不同的 ALU 负责不同像素点的计算,即有的 ALU 计算图片的左上角部分,有的 ALU 计算图片的右下角部分,不同部分的数据互不影响,最终将生成的整张完整的渲染图片呈现于计算机屏幕上,这便是 GPU 的特点,即不断做一些重复量大、运算量大且关联性不大的工作。

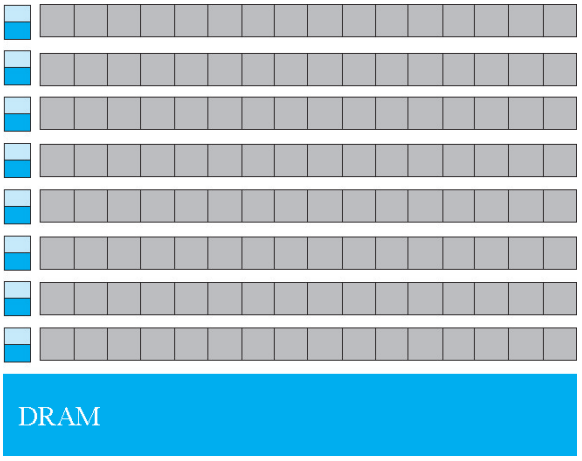


图 3-1　GPU 结构图

彩图 3-1

3.2　图形处理器的发展史

3.2.1　GPU 的前世

1962 年，麻省理工学院的博士伊凡·苏泽兰发表的论文以及他的画板程序奠定了计算机图形学的基础。这促进了 GPU 雏形的形成。20 世纪 70 年代末到 80 年代初，GPU 概念首次被提出，使用单片机集成电路作为图形芯片，用于视频游戏和动画方面，仅能很快地对几张图片进行合成。如图 3-2 所示，1977 年发布的 Atari 2600 是史上第一部真正意义上的家用游戏主机系统。1984 年，SGI 公司推出面向专业领域的高端图形工作站，才有了图形加速器。它们开发的图形系统引入了许多经典的概念，比如顶点变换和纹理映射。在随后的 10 年里，SGI 又不断研发出一系列性能更好的图形工作站。不过，其因为价格较为高昂，在消费级市场的普及度不高，只拥有很小的用户群。这段时期，在消费级领域，还没有专门的图形处理硬件推出，只有一些 2D 加速卡。1991 年，S3 Graphics 公司研究出第一个单芯片 2D 加速器，在 1995 年，3DFX 公司发布了消费级领域史上第一款 3D 图形加速卡 Voodoo，

图 3-2　1977 年发布的 Atari 2600

这也是第一款真正意义上的消费级 3D 显卡。随后几年，AMD 公司发布了 TNT 系列显卡，ATI 公司发布了 Rage 系列显卡。无论是 TNT 系列显卡还是 Rage 系列显卡，这些显卡在硬件上实现 Z 缓存以及双缓存，可以进行光栅化之类的操作，同时也实现了 DirectX 6 的特征集。从此之后，CPU 开始从繁重的像素填充任务中脱离出来，将更多的精力放在其他任务上。

3.2.2　GPU 的今生

1998 年，由 NVIDIA 公司研发的 modern GPU 宣告成功，这一时刻成为 GPU 研发史上的划时代的时刻，宣告着 GPU 研发成为现实。当时研制的 GPU 是图形芯片 Geforce 256，如图 3-3 所示。对于图形芯片领域来说，这是一款史无前例的产品，是第一款提出了 GPU 概念的新兴产品。一般将 20 世纪 70 年代末到 20 世纪 90 年代末这段时间称为 pre-GPU 时期，而将 Geforce 256 诞生后，即 1998 年之后的 GPU 称为 modern GPU。在 pre-GPU 时期，一些图形厂商如 Evans 与 Sutherland，也都在研发自己的 GPU，这些 GPU 现在还没有被淘汰。

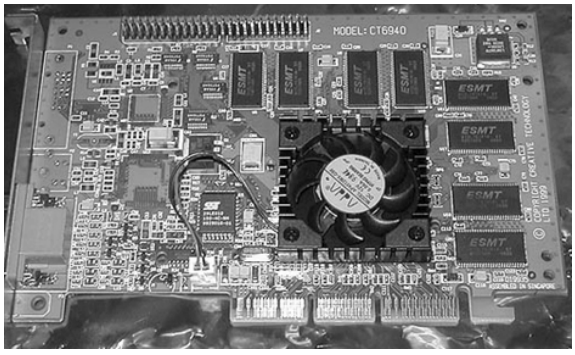


图 3-3　Geforce 256 芯片

2001 年，微软公司研发的 DirectX 8 宣告成功，在这一版本中包含了 Shader Model（优化渲染引擎模式）的 1.0 标准。遵循其标准的 GPU 具备顶点和像素的可编程性，从这时候开始，微软便引领图形硬件的标准。同一时间，NVIDIA 和 ATI 也各自发布了新的 GPU，对于他们发布的这两款 GPU，其都可以支持顶点编程，可以通过应用程序指定指令序列来处理顶点。但不足的是，这一时期的 GPU 还不支持像素编程，对于像素编程工作只能提供简单的配置功能。

2002 年年底，微软发布了 DirectX 9.0b，Shader Model 更新到 2.0 版本，让 Shader 成为其标准配置。2003 年发布的 OpenGL 1.4 中开始正式提供对于 GPU 的编程接口规范。也正是从 2003 年开始，无论是 NVIDIA 发布的产品还是 ATI 发布的产品，都开始同时具备可编程顶点处理和可编程像素处理器的特点，有良好的可编程性。从这时开始，程序员终于可以根据自己的需要灵活控制 GPU 的渲染过程，无须关注其他硬件特性。也正是从这时开始，可编程成为 GPU 的一个特性。

2006 年，Shader Model 4.0 发布。Shader Model 4.0 不同于以往的版本，它采用了一种

统一渲染架构，使用了统一的流处理器。流处理器可以很大程度地提高工作效率，GPU 从单纯的渲染转向通用计算领域，并且扩充了几何编程这一概念，主要用于快速处理一些几何图元和创造新的多边形。

3.2.3 GPU 的摩尔定律

摩尔定律是由英特尔创始人之一戈登·摩尔提出来的。该定律指出，当价格不变时，集成电路上可容纳的元器件数目，约每隔 18～24 月增加一倍，性能也将提高一倍，相同价格下能买到的计算机性能，将每隔 18～24 月翻一倍以上。GPU 的发展最开始的时候也遵循摩尔定律，在一段时间之内性能会翻倍，处理速度和计算能力也会呈指数上升。

在 GPU 发展的历史中，一开始发展速度非常快，基本上 GPU 的计算能力和运算能力呈翻倍增长，21 世纪初为飞速发展阶段。在 CPU 发展陷入瓶颈的现在，GPU 性能提升的潜力是非常巨大的，按照 NVIDIA 公司的预测，从 2005 年开始到 2025 年，GPU 的计算性能将会提升 1000 倍。总的来说，GPU 发展到现在的阶段，十分遵循摩尔定律的规律，在未来的发展阶段将具有非常大的潜力。

3.3 图形处理器的分类

GPU 的分类如图 3-4 所示，具有两种不同的分类方法：一种是根据与 CPU 的关系，可以分为独立 GPU 和集成 GPU，其依据是是否具有自己的独立板卡；另一种是根据应用终端的不同，GPU 可以分为 PC GPU、服务器 GPU 以及移动 GPU。PC GPU 既有独立的 GPU，也有集成的 GPU，服务器 GPU 则是专门为了计算加速和深度学习领域的独立 GPU，而移动 GPU 一般来说是集成 GPU。

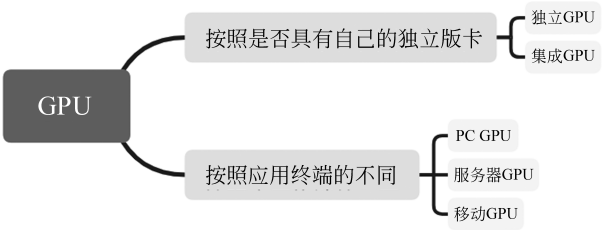


图 3-4 GPU 的分类

第一种分类方法，GPU 分为独立 GPU 和集成 GPU 两部分，在现在的消费市场上，很多计算机供应商会以独立 GPU 为噱头吸引用户购买，使得市场上普遍形成一种风气——独立 GPU 一定是好的，其实这两种类型的 GPU 各有特点，应该根据不同的需求选购不同的 GPU。

首先是集成 GPU。由集成 GPU 的名字可以看出，它是将 GPU 集成，那么集成在哪呢？集成在 3.2 节介绍的 CPU 上。集成 GPU 具有价格低、供电好、兼容好、升级成本低的特点，价格低主要因为 GPU 集成在 CPU 上，节省了空间上的位置，如图 3-5 所示。而且，集成 GPU 没有显存，因此占用计算机其中一部分内存进行运行，这个特点大大减少了制作难度，使得我们制造集成 GPU 的成本要比独立 GPU 低。供电好的特点是因为独立 GPU 需要单

独供电且其运算能力更强大，消耗也更大，所以会造成计算机上的负担。集成 GPU 只要 CPU 供电，就可以满足所需的电量要求，这也使得集成 GPU 更加节约能耗。对于兼容性好、升级成本低的特点，是因为很多计算机厂家在生成计算机时就在 CPU 上安装了集成 GPU，集成 GPU 的升级只更换主板便可，使得升级更简单。而且，因为集成 GPU 和 CPU 配套，所以很少碰到兼容性的问题。但因为其占用 CPU 的空间和内存，没有自己独立的空间和内存，所以一方面集成 GPU 会影响计算机性能，另一方面会出现计算能力不足，渲染某些画面吃力等现象。集成 GPU ARM 公司占据大部分份额，而国产 GPU 在集成 GPU 上也逐渐占据一席之地。

独立 GPU 如图 3-6 所示，是专门用来处理图像的硬件，通过 PCI-Express 扩展插槽与主板连接。正好相反的是，集成 GPU 的优点是独立 GPU 的缺点。独立 GPU 价格昂贵，功耗大。举一个例子来说，一块华硕 GTX1080Ti 尊享版价格达到万元，而相对应的，独立 GPU 具有非常巨大的优点，总结来说就是性能高、画质强、大型游戏运行流畅，具有非常强大的图像处理功能，一般支持高性能游戏的计算机都采用独立 GPU，应用于 VR/AR、人工智能等领域的高性能 GPU 也是独立 GPU。在独立 GPU 领域，英伟达和 AMD 占据大部分市场份额，而在 2020 年，国产 GPU 开始正式向独立 GPU 领域发起进攻的号角。集成 GPU 和独立 GPU 的区别见表 3-1。

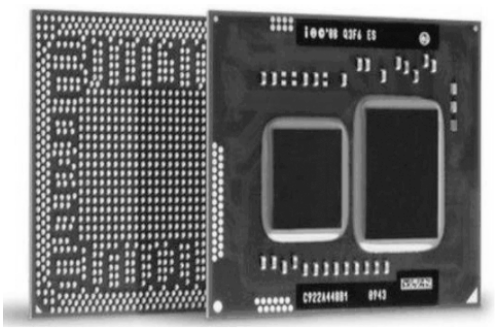


图 3-5　Intel 首款集成 GPU

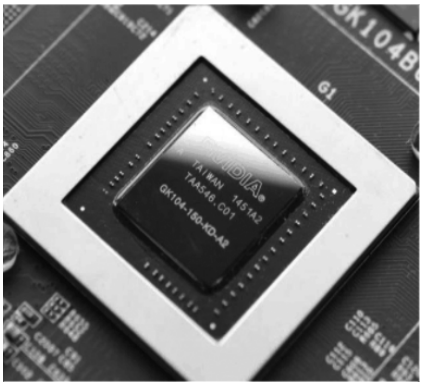


图 3-6　NVIDIA 的独立 GPU

表 3-1　集成 GPU 和独立 GPU 的区别

区　　别	集成 GPU	独立 GPU
价格	低	高
兼容性	强	弱
性能	较差	较好
升级成本	低	高
耗能	低	高
是否占用内存	是	否
应用领域	移动计算市场，例如笔记本和手机	高性能游戏计算机、人工智能、VR/AR

所以，在选择 GPU 时，并不一定选择独立 GPU，而应该根据需求选择。配置高和轻薄是难以双全的两个条件，若想配置高，去做一些图像处理、游戏开发工作，则需要选购独立 GPU，这也导致计算机体积增大、耗电量变大、质量增大等不方便携带的问题。若不需要那么强的图像处理功能时，例如只用于办公和视频浏览等用途，则可以选择集成 GPU，一方面机身更加轻薄，容易携带，另一方面能耗降低，不会造成选购独立而浪费资源的问题。

对于独立 GPU 和集成 GPU 来说，评价指标有所不同，但其运算能力和功耗还是评价这两种 GPU 的两大重要指标。其中，独立 GPU 的厂商将 GPU 芯片、显存、散热器、显卡接口等包装成一个独立的 GPU。其中运算能力、数据传输能力和数据存储能力共同决定了独立 GPU 的运算性能。而功耗和散热可以从散热设计功耗和散热设计两方面考察。而集成 GPU 的评价在独立 GPU 的基础上还要额外考虑内存带宽。集成 GPU 一般用在移动端，不配置专门的独立显存，而是和 CPU 共用内存，因此内存带宽代替显存带宽成为其的一个重要指标。评价独立 GPU 的方法见表 3-2。

表 3-2　评价独立 GPU 的方法

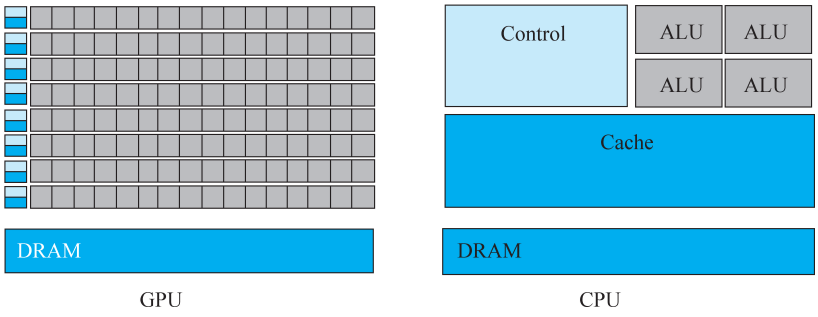
独立 GPU	具体能力	性能指标	指 标 意 义
性能方面	计算能力	核心数目	数目越多，单位时间内可以执行的运算越多
		时钟频率	时钟频率越高，运算的速度越快
	数据传输能力	显存位宽	指的是显存在一个时钟周期内所能传输的数据的位数，目前主流的有 64 位、128 位和 256 位
		显存带宽	指 GPU 与显存之间的数据传输的速率，是影响深度学习性能的最重要的因素
	数据存储能力	显存容量	类似于计算机内存，决定着 GPU 处理过的或者即将提取的数据量的大小
功耗方面	功耗和散热	散热设计功耗	即 TDP，是芯片在真实运行时所散发的最大热量，TDP 越大，越费电
		散热设计	有风冷和水冷散热两种类型。水冷散热效果好、噪声低、价格贵。而风冷散热有普通风扇和涡轮风扇两类

最后，这两种 GPU 都具有非常巨大的市场和发展潜力，国产 GPU 将在这两种不同类型的 GPU 领域齐头并进，不断发展，抢占市场，形成自己的 GPU 配置的生态圈。

3.4　图形处理器与中央处理器的对比

对比 GPU 与 CPU，它们在结构上有非常大的差异[36]。图 3-7 展示了 GPU 和 CPU 的结构对比。从图 3-7 可以看出，CPU 的大部分面积为控制器（黄色部分）和寄存器（红色部分 Cache），与之不同的，GPU 则拥有更多的 ALU 用于数据处理，对于 GPU 来说，它可以处理更加复杂的运算数据，所以 GPU 逐渐应用于机器学习等一系列需要大量计算的应用领域。

相比于 CPU，GPU 并没有特别良好的控制体系，所以无法处理一些复杂多样化的问题，类似于一些需要知道数据相关性的算法，在 GPU 上很难实现。



彩图 3-7

图 3-7　GPU 和 CPU 的结构对比

如果举一个例子进行对比的话，CPU 类似于公务员体系中的一个部门，不同职位间的工作人员互相帮忙，互相协作，做自己不同的工作，不断地开会、总结，最终形成一套完整的解决方案，然后每个部门负责不同的工作，不同部分之间可以进行沟通，开会讨论，最终解决一个复杂的问题。而 GPU 类似于加工厂，每个人做着自己的工作，不用担心其他人做什么工作、工作进度如何，只专注于自己手头上的工作，甚至大多数人做着一样的工作。他们不需要开会，不需要总结，很多时刻很多人只干着同一份工作。最终，各个厂房的人做出来的东西被集中在一起，工作进度就算完结。尽管两方的工作模式差异特别大，但我们可以知道的是，在社会上，公务员团队和加工厂团队都是不可或缺的，只有两个团队好好配合，才能完成社会上各种各样的工作，这也是计算机运行的一个准则。总的来说，GPU 与 CPU 各方面数据的比较见表 3-3。

表 3-3　GPU 与 CPU 各方面数据的比较

对　比　项	处理器	
	GPU	CPU
高速缓冲存储器	多	少
线程数	少	多
寄存器	多	少
单指令多数据流	快	慢
控制器	少	多

可以看出，由于 GPU 中的控制器少于 CPU，因此在控制流方面 GPU 弱于 CPU。控制器的主要功能是取指令，并指出下一条指令的位置，协助计算机各部分有条不紊地工作。但因为 GPU 具有更多的寄存器计算单元等，所以 GPU 运算的速度显著大于 CPU，适合对大量数据进行计算。GPU 上的数据不需要依赖其他数据进行计算，例如渲染一张图片，CPU 需要不断通过循环语句遍历像素，所以代码的量和考虑的逻辑语句较多。而在 GPU 上，只通过一条代码即可完成，不需要过多地考虑其他像素点的属性。

当认识了 GPU 和 CPU 之后，我们会产生一个疑问，什么程序适合在 GPU 上运行呢？根据 GPU 独特的结构特征，可以得出计算密集型的程序与易于并行的程序更适合在 GPU 上运行。

3.5 图形处理器的应用领域

GPU 应用最广泛的领域是图形渲染，包含 PC 端的图形处理和移动端的图形处理。最开始 GPU 的应用领域是 PC 端的图形处理，虽然近年来 GPU 在 PC 端的整体销量下滑，但仍有游戏硬件市场这个增长的亮点。而移动 GPU 作为提升智能手机性能的核心部件，是过去几年 GPU 的主要增长点，未来几年仍可能是 GPU 增长的主力。

对于 PC GPU 来说，受到全球 PC 出货量降低的影响，游戏领域成为其增长的主力。最开始，GPU 伴随着 20 世纪 90 年代 PC 的大范围普及而迎来其发展的黄金时期。但随着 2012 年开始 PC 的出货量持续下滑，PC GPU 的出货量也持续下滑。2018 年一季度 PC 端 GPU 出货量同比增加 3.4%，环比下滑 10.4%。图 3-8 反映了近几年市场需求的平淡。而与之相比的是，游戏持续火爆，带动了 PC 游戏硬件市场的飞速增长。截至 2018 年年底，我国游戏产业用户的规模已经达到 5.83 亿人，对比 2009 年增长了将近 8 倍。2016 年，全球 PC 游戏硬件市场首次超过 300 亿美元，这是一个非常巨大的突破，在原本的预测中，该规模只有到 2018 年才有可能突破。而游戏需求的上涨推动着独立 GPU 的增长。2015 年，PC GPU 比重为 27%，此后份额将持续增长[37]。

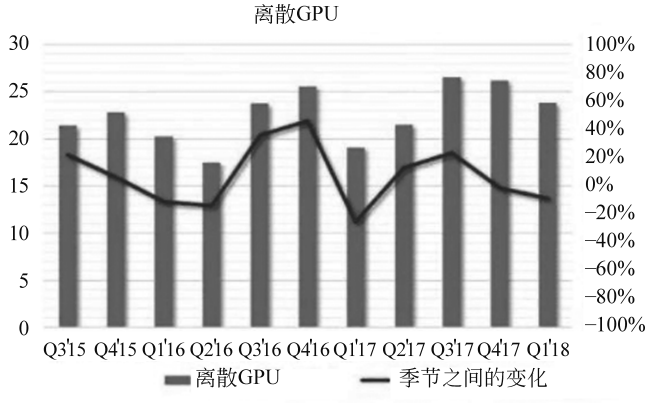


图 3-8 PC GPU 出货量及其环比增长

对于移动 GPU 来说，其崛起于智能手机的浪潮，持续受益于存量替换。移动 GPU 是提升手机性能的一个核心部件，其能决定设备的流畅程度、游戏的顺畅程度等参数。其应用领域的火热程度取决于智能手机行业的景气度。根据 IDC 统计，2017 年全球智能手机出货量为 14.72 亿台，增速首次转为负值，在手机市场需求较为饱和的情形下，未来移动端 GPU 市场将主要受产品迭代的存量替换需求驱动[37]。

今后，人工智能将为 GPU 的发展提供新兴的需求。GPU 在人工智能领域的应用包括云端以及终端。云端有云平台和超级计算机，而终端有机器人和智能汽车。这些领域是近

期 GPU 发展最为迅猛的领域，也将是 GPU 发展的一大前景。

对于云平台来说，GPU 已经成为最主流的芯片，技术服务的形式非常多元化。GPU 非常适配云端深度学习的过程，其良好的可编程性、成熟的技术体系能使得开发更加便捷、开发周期更加短暂。目前，GPU 已经成为云端最主流的 AI 芯片，亚马逊、微软、腾讯、百度、华为等 IT 巨头的云平台均采用 GPU 进行云平台上的深度学习计算。

对于超级计算机来说，我们可以使其搭载 GPU 从而成为人工智能超级计算机。以前没有搭载 GPU 的超级计算机的优势主要是强大的计算能力，而现在搭载 GPU 的人工智能超级计算机的优势主要是可以应用在人工智能领域。所以，深度学习就成为超级计算机和人工智能最为紧密的结合点。一方面，神经网络的训练需要大量的超级运算性能；另一方面，超级计算机借助其强大的计算能力训练模型，让神经模型具备更好、更快、更加准确的识别和推理的能力。所以，如图 3-9 所示，搭载了 GPU 的人工智能超级计算机具有非常巨大的发展前景和发展优势，其可以缩短数据处理时间，加速深度学习框架，并设计出更加复杂的神经网络，获得更快速的速度、更巨大的模型，以及更精准的结构，是未来超级计算机的一个重要的发展方向。

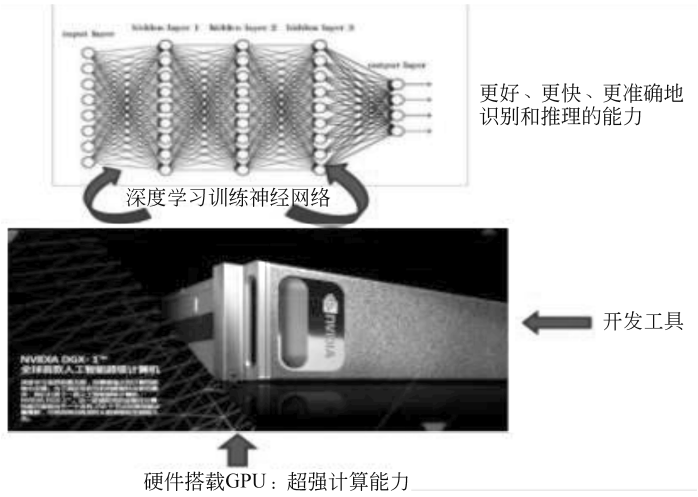


图 3-9　人工智能超级计算机示意图

对于机器人的终端应用，2017 年 GPU 巨头公司 NVIDIA 发布了高性能、低功耗的计算机平台 NVIDIA Jetson TK2，成为机器人、无人机、智能摄像头等计算密集型终端的一个理想的深度学习平台。在该平台上，用户可以通过其提供的开发套件和教程构建属于自己的平台投入生产，该系统应用于机器人凌云的 Fellow、快递机器人领域的 Marble 等国外企业以及海康威视、京东等国内企业。

对于智能汽车的应用，GPU 主要应用于 AI 自动驾驶领域。如图 3-10 所示，智能驾驶汽车是通过从摄像头、超声波等收集的信息感知道路的情况，自动规划汽车行车路线，并控制车辆安全。对于自动驾驶来说，车辆必须快速处理传感器传输的海量数据，这便需要高效快速的 GPU 的支持。NVIDIA 推出了自动驾驶车载计算机 Drive PX 2，其单精度计算能力等同于 150 台苹果 MacBook Pro，是上一代速度的 4 倍，而深度学习速度可以达到每秒 24

万亿次操作，是上一代的 10 倍之多。随着自动驾驶概念的不断普及，人工智能 GPU 需求有望实现持续的增加。

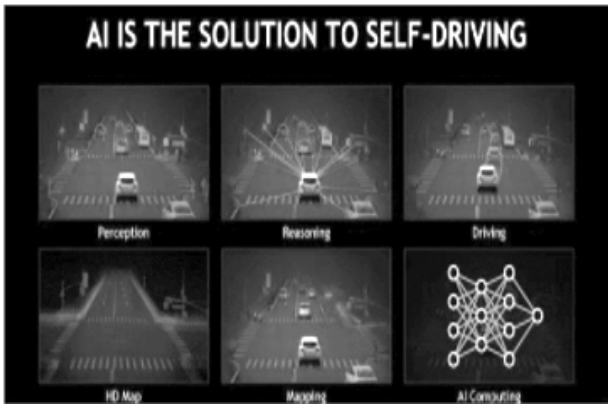


图 3-10　人工智能技术应用于自动驾驶

3.6　经典图形处理器

3.6.1　超高性价比的 Radeon 9550

在 Radeon 9550 出现之前，Radeon 9700 的口碑在市场上极其成功，但其还是很难在市场层面上击败竞争对手，因为其属于少数人的旗舰架构。所以，这时需要研发一款适合大众的可以攻占市场份额的中端产品，可以有效占据市场大量份额，这时候 Radeon 9550 芯片如图 3-11 所示，横空出世，强势攻占市场。

图 3-11　Radeon 9550 芯片

Radeon 9550 不仅继承了 Radeon 9700 的全部优点，而且其处于中端，甚至中低端的售价也非常受消费者欢迎[38]。但这些都不是其最大优点，其最大的优点是它不可思议的可超频能力以及超频后性能。在当时普遍 CPU 超频都不超过 50％的时代，Radeon 9550 具有 250MHz 的“超高默认核心频率”，而且每一块甚至都能达到 400MHz，这种超过 60％的超频

幅度非常不可思议，超频之后的芯片具备了接触旗舰级边缘的实力。所以，对于这种价格与定位不相称、“超高性价比”的超频，Radeon 9550 占据了大量的市场，在当时基本上与“性价比显卡”这个词等价。许多显卡爱好者人生中的第一块 GPU 都是 Radeon 9550，这足以说明其在显卡历史中的地位。

3.6.2　SweetSpot 的前身——GeForce 9600GT

GeForce 9600GT 芯片如图 3-12 所示，对于现代显卡来说具有非常重要的意义，在当时市场上，它是一款风靡全球的经典终端游戏显卡，占据市场很大份额，赢得了消费者的欢迎[38]。而且对后面 GPU 的影响也是非常巨大的，其是炙手可热的 SweetSpot 级显卡的雏形，对后面显卡的发展影响巨大。

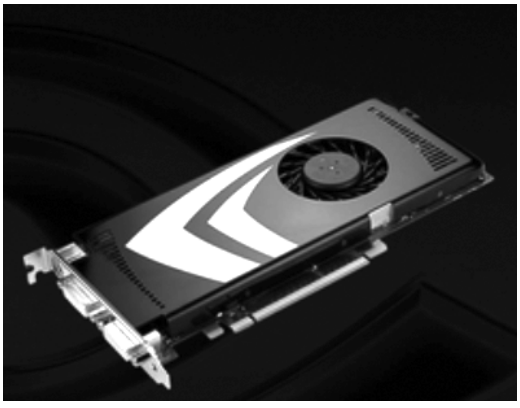


图 3-12　GeForce 9600GT 芯片

在 DirectX 末期时，当时人们对显卡的架构发展争论了很长一段时间，对于负责直接图元操纵的后端与负责 Shader Program 运算的 ALU 谁更为重要陷入了争执，两派不同意见的人争论不休。当时，不同的公司在这两个方向上分别进行尝试，GeForce 9600GT 是其中最为经典的一个。

GeForce 9600GT 基于 G94 的架构，拥有 64 个流处理器，16 组 ROP 单元以及 256bit 显存位宽。对比旗舰版的 G80/92，其 ALU 削减了一半，但 GeForce 9600GT 的后端规模却与其相同水平。其后端规模为其带来优秀的带宽和处理能力，因此使得 GeForce 9600GT 具有非常优秀的性能，超越了当时市场上的很多中端显卡。

GeForce 9600GT 拥有更小的芯片面积、更低的功耗以及发热表现，使得其可控性成本更优，大大提高了游戏性能和使用体验，对当时争论不休的问题给出了自己的答案——后端对于游戏显卡非常重要。它所取得的结论性经验不仅影响了 NVIDIA 后续中高端显卡的研发，还一定程度上改变了 NVIDIA 公司的发展轨道和发展方向，而且也成为后来众多 SweetSpot 显卡借鉴的一个对象。

3.6.3　亘古未有的经典——Radeon HD 4850

对于 AMD 公司来说，R600 的失败使得其在显卡领域大受打击，AMD 公司需要一款性

能上具有独特优势的显卡作为主打产品来重新占据市场、树立信心，于是在当时发布了RV770。出乎意料的是，作为RV770开路先锋的Radeon HD 4850居然成为一块名留史册的显卡。

RV770在当时取得了巨大的成功，它不仅弥补了之前R600失败带来的损失，还通过其惊人的性能和极高的性价比在市场上大受欢迎，重新使得市场恢复了对AMD图形部门的信心。后来，研究显卡历史的人说，“整个AMD都要感谢Radeon HD 4850”。

Radeon HD 4850是真正意义上的前无古人，它具有与旗舰RV770相同的800个流处理器，40个TMU单元以及16组ROP阵列，尽在显卡端采用了带宽较低的GDDR3颗粒，使得其成功拥有铁剑Radeon HD 4870的性能，并且凭借其超高的性价比成为了显卡历史上最具“良心”的芯片之一。不仅如此，其在通用计算领域也开始了初步的尝试，为后来的RV970的追赶打下了基础。其芯片面积远低于NVIDIA即将到来的架构，这让其背负了很大的压力。其载入史册的最大原因是它把小芯片策略、多芯片互联手段以及成功的市场切入点和运作结合在一起，最终在市场竞争中取得了巨大的成绩。其通过优秀的架构设计带来优秀的成本控制，将价格控制在市场和公司都能接受的位置，填补了该市场位置的空白。Radeon HD 4850的横空出世，使得AMD公司在短时间内夺回了市场的主动权。Radeon HD 4850是人类历史上第一块将性能、架构、工艺、特性支持、市场策略和运作手段等显卡的全部有关属性都发挥到极致的显卡。其经典程度前无古人。所以，当我们回顾GPU的历史时，Radeon HD 4850(见图3-13)无疑是最耀眼的一颗明星，是一个亘古未有的经典。

图3-13　Radeon HD 4850芯片

3.6.4　幸运时代的幸运产物——GeForce GTX 680

对于如图3-14的GeForce GTX 680芯片，它有一个史无前例的评价——具备7个“更”的显卡：更小、更短、更轻、更凉、更省电、更便宜，同时也更快。与前面的GPU不同，该显卡是一款旗舰级显卡，但其又不具备旗舰显卡的一些特点，它仅达到Tahiti架构80%的芯片面积、75%的单元规模和75%的显存带宽。

它在历史上的幸运大多是竞争对手赋予的。其由NVIDIA生产，而且凭借其一款产品对抗了AMD的整条产品链，因为当时AMD架构研发的积弊而背负了太多不该背负的负担，使其定价以及推广策略都非常糟糕，从而使得GeForce GTX 680在当时风靡一时，获得了极高的市场占有率，并在当时为NVIDIA获得了极为丰厚的利润回报。

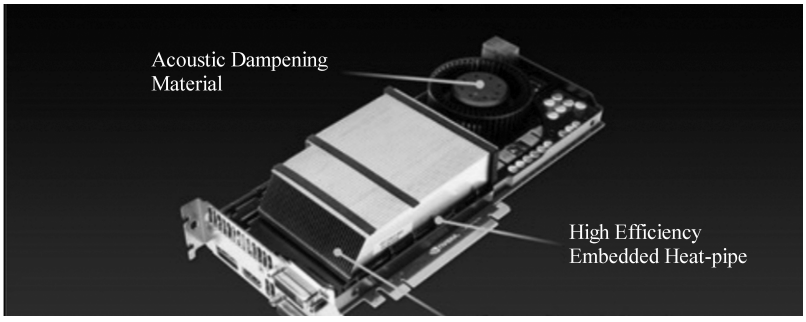


图 3-14　GeForce GTX 680 芯片

GeForce GTX 680 身上有许多令人瞩目的优点——极高的性能功耗比、超高性价比、非常优秀的显卡性能、运转时良好的温度，以及低分贝的噪声等[38]。但对于该显卡来说，在显卡历史上，其最特别的无法复刻的特点便是它的“幸运”，它的幸运成为显卡历史上一道有趣的风景线，成为一种不可复刻的经典。

3.7　国产图形处理器

从前面可以看出，GPU 对于计算机来说是非常主要且必不可少的一部分。在当今社会各国贸易间不断摩擦，竞争逐渐加剧，贸易壁垒不断提升的情况下，发展国产 GPU 是刻不容缓的一件事。只有自己掌握核心技术，把握科技发展，才能应对复杂多变的国际形势，才能更好地发展自我国家的实力，建立属于自己的国产计算生态圈，在未来发展领域拥有自己的一席之地。

在国外封锁技术且不断打压的情况下，我们国家还是涌现出一批优秀的国产 GPU 制造商，他们突破重围，投入大量资金，研制属于我们国家自己的 GPU。其中业内公司可以分为 3 种类型。

1. 自主研发系

自主研发系的内涵指企业从 GPU 的架构和算法等底层的 GPU 领域着手，采取自主研发的产品进行一些产品开发。这些公司能对自主开发的 GPU 进行升级以及迭代的操作，这些能力保证其产品有一个完整的循环链，有能力进入军事或者民用市场与国外 GPU 进行竞争。其主要包括景嘉微和中船系。景嘉微发布的首款 GPU JM5400 有力填补了国内高端 GPU 市场空白，它的出现打破了国外芯片在军用领域的垄断。中船系下属 2 个研究所——中船重工 709 所以及中船 716 所则各自研发了一款国产 GPU，有力填补了市场需求，为我国国产计算生态平台的构建注入不可缺少的动力。

2. 学术课题系

该系列以西邮微电为代表，其特点是以学术研发作为主导，多诞生于高校实验室。西邮微电子科技有限公司脱胎于西安邮电大学 GPU 团队。其研制的“萤火虫 1 号”历经 5 年开发，于 2015 年通过陕西省支持的成果鉴定，成为自主研发的 GPU 的雏形芯片。

3. 技术引进系

以中科曙光为代表，其通过与国际上知名的 GPU 生产商（如 AMD）进行合作引进技术用于生产 GPU。例如，中科曙光与 AMD 公司进行合作，并且其还收购了 Imagination 的凯桥资本以及美国图芯的芯原。

下面主要介绍自主研发系以及学术课题系，这两大系列都研制了属于我们自己的 GPU，从 0 到 1 打破了国外 GPU 对国内 GPU 市场的垄断，帮助有力地建立属于自己的国产生态平台。“中国心，中国芯”，这些国产 GPU 必将焕发更加巨大的活力，构建属于我们自己的计算生态平台。

3.7.1 景嘉微

景嘉微公司（以下简称景嘉微）是国内率先成功自主研发国产化 GPU 并且产业化的企业，于 2016 年成立。其产品主要涉及图形显控、小型专用化雷达和芯片等领域[39]。在 2006 年成立的时候，刚好是中国军用飞机显控系统慢慢转向 GPU 升级的时刻，景嘉微把握时机开始投入军用飞机图形显控领域的研究。2010 年，公司“图形加速器技术研究”项目荣获“核高基”项目立项，从此走上了自主知识产权图形处理芯片的研发道路。2014 年，第一代 GPU 研制成功，其性能优于军工电子显控领域主流的一些进口芯片。2018 年，第二代 GPU 研制成功，它在第一代的基础上作了重大改进，有了非常巨大的进步。目前，第一代 GPU 已经运用于各种军用显控系统中，而第二代 GPU 已经获得党政计算机的意向订单，有望进一步推动党政机关实现自主的国产计算机的全面生态化。

2018 年，国家集成电路基金作为国有法人入股景嘉微，成为公司的第二大股东[40]。国有法人股东的加入体现出产品技术水平、发展前景、发展潜力得到了国家的认可，也体现出国家对国产 GPU 的重视和肯定，这一次入股也进一步提升了景嘉微的市场地位，体现了其作为国产 GPU 的领军企业的身份。

景嘉微具备高层次科技人才的核心团队，推动了军用图形显控国产化的进程。景嘉微具备高端人才的储备力量，招揽了各大学校的高技术人才。景嘉微的核心团队基本来自中国人民解放军国防科技大学，且都是在各自领域具备丰富研发经验的资深研发专家。由于拥有一批军校科技人才作为核心人员，因此景嘉微更容易把握国内顶尖的科技产品研发的技术方向，更能把握用户的真实需求和未来发展的趋势，更能与国防军用产品进行技术对接，这是其蝉联于军用图形显控产品市场龙头地位的重要支撑。

1. JM5400

JM5400 芯片如图 3-15 所示，是景嘉微第一代 GPU，是一个具有重大意义的划时代产品。它的出现打破了国外芯片在军用领域的垄断，关乎国家命脉的军用领域不再受制于他人，也不再担心国家军事的消息遭到泄露，从 0 到 1 实现了国产化代替，这款产品于 2014 年发布，是国内首款具有完全知识产权的图形处理芯片，也是一款高性能、低功耗的优质 GPU 芯片。JM5400 芯片取代了中国军用飞机传统的、常用的多款海外芯片，类似于 ATIM9、M54、M72、

图 3-15 JM5400 芯片

M96 等。相比于海外芯片，JM5400 的性能更高，工作温度范围更宽，并且功耗更低。JM5400 主要指标见表 3-4。

表 3-4　JM5400 主要指标

参　数	具 体 数 值
工艺	65nm CMOS
时钟频率	内核时钟频率最大为 550MHz，存储器时钟频率最大为 800MHz，软件可配置
主机接口	PCI 2.3 规范，33/66MHz
存储器	片上封装两组 DDR3 存储器，每组位宽为 32 位，共 1GB 容量
渲染能力	含 4 条渲染流水线，像素填充率为 2.2G Pixel/s
工作温度	－55～125℃
存储温度	－65～150℃
功耗	功耗不超过 6W，内部各功能模块可独立关闭，可进一步减少功耗
封装	FCBGA 1331 脚，MCM 封装
尺寸	37.5mm×37.5mm

下面根据表 3-4 介绍芯片的各项指标。首先，其 65nm CMOS 指的是在最初栅极上留下 65nm 宽度的光刻胶，是一种极为精细的工艺，通常该尺寸越小，晶体管密度越高，GPU 的性能越好。时钟频率则表现 GPU 的处理速度，时钟频率越高，则 GPU 速度越快。存储器表现 GPU 内存大小，表示其可以存储数据的多少。像素填充率是指图形处理单元在每秒内所渲染的像素数量，JM5400 芯片每秒可以处理 22 亿像素。工作温度和存储温度表示其工作于存储环境的限制。功耗则是其工作时每秒消耗的能量。封装和尺寸表示其外型参数，GPU 基本上朝着越小越薄的方向前进。

2. JM7200

JM7200 芯片如图 3-16 所示，是景嘉微的第二代 GPU，其于 2018 年 9 月完成流片、封装阶段工作。其制程工艺 28nm，核心频率 1.2GHz，搭配 4GB DDR3 显存，性能跟 NVIDIA 的 GT 640 显卡相近。发展国产 GPU 有一个得天独厚的优势，即可以与国内的 CPU 操作系统等形成自己的国产生态圈，形成属于我们自己的国产计算平台，不再受制于其他国家的技术限制，而 JM7200 便完成了适配构建这一步，并且有希望进行大规模推广。根据目前的适配测试，JM7200 的产品性能已经能够适配台式计算机，满足国内计算机基础使用的需求，满足计算机市场推广的条件。

图 3-16　JM7200 芯片

而且，JM7200 芯片获党政市场意向订单，景嘉微有望占据民用计算机市场。2019 年，景嘉微签署了《战略合作协议》，为了打造战略合作伙伴关系，为政企用户提供基于 JM7200 芯片的国产图形显卡的以及其一系列的解决方案，湖南长沙在 2020 年购进 10 万套基于 JM7200 的图形显卡，这笔订单标志着该公司正式开拓民用计算机市场，并且在民用计算机市场具有非常巨大的前景和潜力。JM5400 与

JM7200 的对比见表 3-5。

表 3-5　JM5400 与 JM7200 的对比

型　号	JM5400	JM7200
发布时间	2014	2018
工艺	65nm	28nm
外存类型	DDR3	DDR3
内核频率	550MHz	1.2GHz
等效运算频率	160GFlops	500GFlops
存储器带宽	12.8GB/s	16GB/s
存储器容量	1GB	2GB/1GB
OpenGL 支持	OpenGL1.3	OpenGL1.5
研发周期	8 年	4 年

从表 3-5 可以清晰地看到，工艺从 65nm 飞跃到 28nm，是一个很大的跨度。晶体管的宽度从 65nm 下降到 28nm，不仅使得一块芯片上可以搭载的晶体管数量大幅增加，也使得每一个晶体管的运算速度呈几何倍增长，芯片的性能突破了一个档次，中国 GPU 开始向国际领先技术靠拢。内核频率的上升表示其一秒能进行更多次运算，在相同时间内，能处理的数据量更加巨大，对于用户来说，简单明了的感觉便是画面更加清晰、流畅，因为其对每一个画面的计算更加迅速、准确。等效运算频率也是相同的道理。存储器带宽增加也表示其存储速度提升，单位时间内从存储器读出和写入的数据更多，与存储器的交互更加迅速，与存储器的交流更加紧密。而存储器容量也表示其能存储更多的数据，方便进行一些大数据的计算和存储。而且，其能支持更高版本的 OpenGL，代表该 GPU 具有更强大的处理能力，具有更加强大的图形渲染功能，能完成更加复杂的图形处理渲染。最后，从研发周期成倍缩短可以看出我国 GPU 研发技术的不断蓬勃发展，更新速度更加迅速代表技术更加娴熟、平台更加完善、政策更加扶持。JM7200 是一个划时代的产品，比起上一代 GPU 有飞跃式的发展，是中国芯片与国际接轨的重要一步。

3. JM9231 与 JM9271

JM9 系列是公司正在研发的第三代 GPU，该 GPU 较第一代 GPU 和第二代 GPU 的性能有很大的提升，有望进入人工智能的市场。公司此前的第二代 GPU JM7200 虽然已经支持可编程的架构，但是其 GPU 内核与国外的 GPU 公司的产品仍具有一定的性能差异。但是，JM9 系列有望弥补国内 GPU 与国外 GPU 的明显差距，其使用与国际公司通用的做法以及业界主流的统一渲染架构，增加了可编程计算的模块数量，与国际上显卡主流的趋势对接。JM9 系列如果研发成功，将有望追赶上国外主流 GPU 产品 2016—2017 年的性能水平，这将成为国产 GPU 的一个里程碑。国产 GPU 和国外 GPU 差距一直在十年左右，该芯片研制成功将可以追赶上 GPU 的中低端市场水平，将能占据世界 GPU 市场的一席之地，有能力与国外 GPU 公司竞争。而且该 GPU 缩短了国产 GPU 与国外 GPU 的巨大差距，给国产

GPU 的研发带来了巨大的信心。该系列不逊色于 GTX1080，预期可以达到 2017 年国际高端的显卡水平。其核心频率不低于 1.8GHz，支持 PCIe 4.0 x16，采用 16GB HBM 显存，频宽为 512GB/s，浮点性能可达 8 TFlops，未来可以进一步应用于人工智能等高端应用领域。

目前，JM9 系列正处于测试阶段。表 3-6 是 JM9 系列与 NVIDIA 的 GTX 系列产品的对比。

表 3-6　JM9 系列与 NVIDIA 的 GTX 系列产品的对比

指　　标	JM9231	GTX1050	JM9271	GTX1080
API 支援	OpenGL 4.5，OpenGL 1.2	OpenGL 4.6，DX12	OpenGL 4.5，OpenGL 2.0	OpenGL 4.6，DX12
显存时钟频率	1500MHz	1455MHz	1800MHz	1733MHz
PCIe 卡	PCIe 3.0	PCIe 3.0	PCIe 4.0	PCIe 3.0
显存带宽	256GB/s	112GB/s	512GB/s	320GB/s
显存容量	8GB	2GB	16GB	8GB
像素填充率	≥32G Pixel/s	46.56G Pixel/s	≥128G Pixel/s	≥110G Pixel/s
单精度浮点性能	2TFlops	1.862TFlops	≥8TFlops	≥8.873TFlops
影像输出	HDMI 2.0，DisplayPlot 1.3	HDMI 2.0，DisplayPlot 1.4	HDMI 2.0，DisplayPlot 1.3	HDMI 2.0，DisplayPlot 1.4
视频解码	H.265/4K 60FPS	H.265/4K 60FPS	H.265/4K 60FPS	H.265/4K 60FPS
功耗	150W	75W	200W	180W

其中，API 支援表示其应用程序接口版本，兼容版本越高，越能通过应用程序执行更加复杂的操作，可以看到，JMP9 系列与 GTX 系列的 API 版本差距很小，只有 0.1 的版本差距。JM9 系列的视频解码性能已经追赶上 GTX 的指标，而在衡量 GPU 性能的重要参数时钟频率、显存带宽、显存容量以及像素填充率上，JM9 系列普遍比 GTX 快，这是一个质的飞跃，是国产 GPU 追赶上国际先进标准的一个重要信号。国产 GPU 已走上了追赶国际先进标准，甚至超越国际先进标准的道路。而且，值得注意的是，该产品的研究周期为 2～3 年，相比于第一、二代 GPU 的 8 年、4 年的漫长研发周期，第三代 GPU 的研发速度是一个很大的突破。研发周期的不断缩短，有利于产品更快地更新迭代，缩短与国际先进技术水平产品的差距，提升公司产品的竞争力，进一步扩大应用市场空间。而且，缩短研发周期也可以侧面反映出我国研制 GPU 的技术越来越熟练、研发水平越来越精进、研发的流程越来越完善。

4. 三代 GPU 的对比

芯片的研发是一项巨大而复杂的系统工程，景嘉微从数学公式推导开始，在架构设计、算法模型、原理验证、硬件实现、驱动开发等环节全面实现了自主研发，三代 GPU 不断地改进。三代 GPU 的对比见表 3-7。

表 3-7　三代 GPU 的对比

GPU 名称	研发周期	应 用 领 域	国外同类技术水平产品
JM5400	8 年	军用图形显控领域	ATI M96 芯片
JM7200	4 年	主要用于军用市场，拓展至国产化计算机市场	英伟达 GT640
JM9 系列	预 计 2 ～ 3 年	消费电子领域以及人工智能、安防监控、语音识别、深度学习、云计算等高端应用领域	英伟达 GTX1080

首先，显而易见的是 GPU 的研发周期不断缩短，代表研发技术更加成熟、研发流程更加完善。而在应用领域的拓展上，可以看出 GPU 的应用范围更加广阔，功能更加齐全，更能稳定地占据更多的市场，拓宽自我应用领域。军用图形显控领域的应用如图 3-17 所示。而国产 GPU 从单单占据军用市场开始，一步步走向民用市场；从单单只能处理图形显控领域开始，一步步迈向高端应用。这不仅预示着国产 GPU 实力的不断强盛，也代表着国产 GPU 发展的宏伟蓝图。国产 GPU 势必不断追赶上国外 GPU 的顶尖水平，也要不断占据更加广阔的市场，打破国外 GPU 垄断的现状。而从国外同类技术水平产品对比中可以看到，国产 GPU 和国外 GPU 的差距越来越小，从最开始的 10 年左右到现在的 4 年左右的差距，国产 GPU 正迈着坚毅的步伐不断追赶、不断缩短与国外 GPU 顶尖水平的差距。

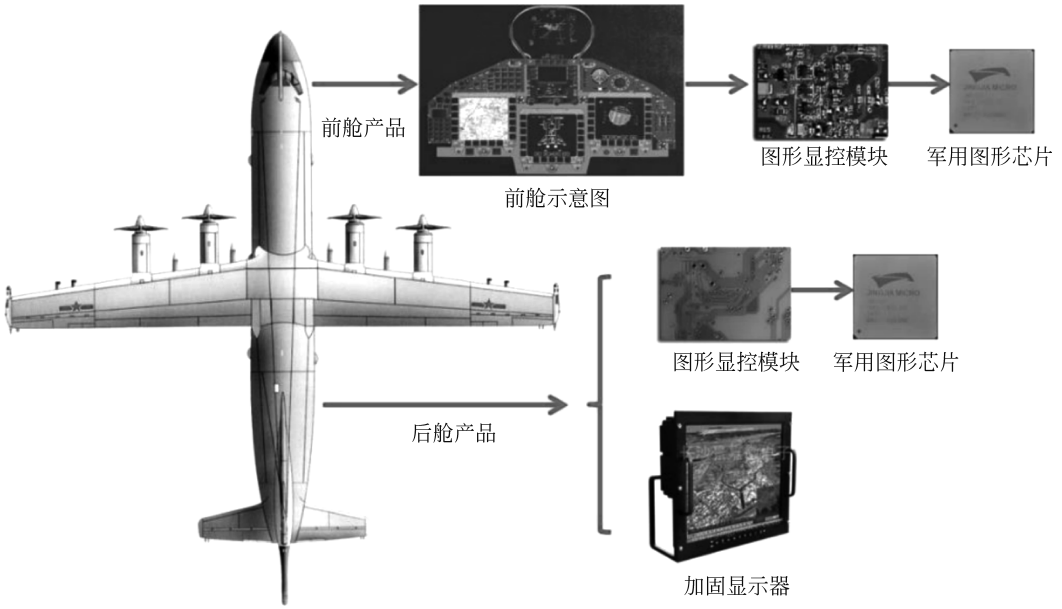


图 3-17　军用图形显控领域的应用

3.7.2　兆芯

兆芯公司（以下简称兆芯）是成立于 2013 年的国资控股公司，总部位于上海张江，在北京、西安、武汉、深圳等地设有研发中心和分支机构。兆芯拥有一大批具备硕士、博士学历的专职研发人员，公司同时掌握了 CPU、GPU、芯片组三大核心技术，具备三大核心芯片及相

关 IP 设计研发能力，并获评“高新技术企业”资质[41]。

在 x86 领域上，兆芯是第三大生产厂家，其能跻身前三的原因是其 20 多年的历史，为什么说 2013 年成立的兆芯有 20 多年的历史呢？因为它的前身是台湾威盛电子公司。威盛公司出技术，上海政府出资金，于 2013 年成立兆芯，其发布了众多 CPU，在 CPU 界具有一定的地位。

而在 GPU 上，兆芯宣布了 GPU 独立显卡发展计划，将帮助其成为国内少数的同时掌握 GPU、CPU、芯片组核心技术的公司，也将使其具有一个较为完善的国产硬件生态圈，帮助该公司更好地跻身国产计算生态平台，建立起属于中国的一个较为完善的计算机生态平台。

兆芯的优势在于，其具有较为完备的硬件生产技术，其生产 CPU 的技术已经较为娴熟。因为 GPU 和 CPU 的部分相似性，该公司生产 GPU 将具有一定的技术基础和生产经验。兆芯对 GPU 的生产也将互补于 CPU 生产线，使得整个公司的计算生态硬件平台更加完善，产品更加全面，可以进一步占据市场。

IDC 预测，到 2023 年中国 GPU 服务器市场规模将达到 43.2 亿美金，未来 5 年整体市场年复合增长率(CAGR)为 27.1%。国内 GPU 市场非常火热，而中美贸易战的前景下我国自主研发的独立显卡将具有非常巨大的发展前景和发展潜力。并且，因为兆芯在 CPU 这块早已在市场上有很高的评价，在消费者人群中有很高的认同感，所以推出该独立显卡将可以有力挺进 GPU 市场，为建设我国自主的计算生态平台添砖加瓦。

兆芯在 2020 年官方视频中宣布了 GPU 独立显卡，其宣传图如图 3-18 所示，表示其最快于 2020 年年底发布，慢则在 2021 年发布，该显卡一经问世，将填补国内 GPU 独立显卡的空白，帮助国产芯片进一步完善属于自己的计算生态的硬件平台。

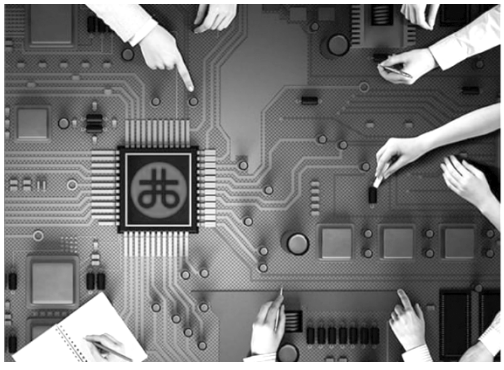


图 3-18　兆芯 GPU 宣传图

其中公布了 GPU 的一些参数，首先，该 GPU 采用相对较低的 70W TDP，功耗相对较低。而在规模上采用了台湾积体电路制造股份有限公司的 28nm 制程[42,43]，采用 28nm 的工艺大多考虑到中美贸易战，美国可能会阻止台湾积体电路制造股份有限公司向中国客户提供 16nm 以及更小的工艺，因为这些较新的工艺涉及一部分美国知识产权。

兆芯现有的 iGPUs 支持 DX11、OpenCL 1.1 和 OpenGL 3.2，并支持硬件加速的视频编码和解码，但细节很少，GPU 监控应用无法获取架构组件的更多细节。集成显卡支持 DisplayPort、eDP、HDMI 和 VGA 接口，可以同时输出到两个 4K 分辨率的屏幕上。

3.7.3 浪潮信息

浪潮电子信息产业股份有限公司(以下简称浪潮)具有非常悠久的历史。浪潮的前身是20世纪50年代成立的山东电子设备厂。1970年发射的人造卫星“东方红一号”就采用了其生产的晶体管。1988年,浪潮信息公司成立,并在2000年上市。2000年,浪潮服务器打破世界纪录,这也是国产服务器首次打破世界纪录。

浪潮拥有三家上市公司,现今向全球超过100个国家和地区提供IT产品和服务。浪潮服务器销售额非常巨大,全球前三,中国第一。由图3-19可知,在2019年上半年,浪潮以50.8%的市场份额占据中国市场第一,是国内GPU服务器的龙头企业。

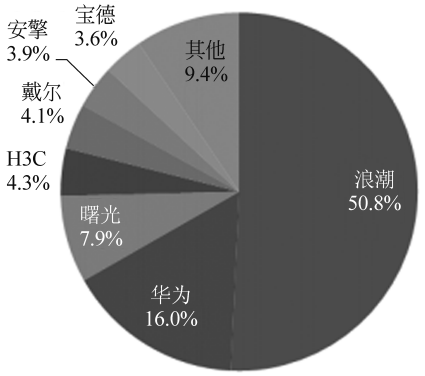


图3-19 2019年中国GPU服务器厂商销售额占比

浪潮推出GPU服务器在市场上广受好评,在2019年服务器市场持续走低的情况下,浪潮持续领涨势头。2019年,浪潮销售额和出货量增速分别为18%及11%,均实现了两位数的快速增长,这归功于公司不断增加的研发投入。一家优秀的研发公司对研发经费的投入非常重要,只有重视研发,掌握核心科技,才能凭借过硬的技术立足市场,这也是浪潮能立足世界服务器领域的一个重要原因。

1. NF5488A5

NF5488A5服务器如图3-20所示,是浪潮信息全新发布的五款支持NVIDIA A100 GPU的AI服务器中的其中一款,该服务器取得了巨大的成绩。在全球首个AI测试标准Mlperf的测试下,性能排行全球第一。NF5488A5提供非常强大的单机训练性能和超高的数据吞吐,对于众多AI应用具有非常好的应用。其具备非常多的优点,如极致的AI训练性能,最高可提供5 petaFlops AI算力[44],而且其具备非常极致的通信速率,对比上一代的带宽翻倍,可极大降低数据延迟,是一个巨大的飞跃。并且其具备非常优秀的硬件设计,在4U空间中,模块化设计,适用于更加广泛的数据中心环境,可以在极大程

图3-20 NF5488A5服务器