DOI：10.19816/j.cnki.10-1594/tn.2020.02.036

GPU的发展历程、未来趋势及研制实践

熊庭刚
（中国船舶重工集团公司第七〇九研究所 武汉 430205）

摘　要：凭借GPU强大的计算能力，超级计算机在数据处理、物理模拟、天气预测、现代制药、基因测序、先进制造、人工智能、密码分析等方面都有着广泛的应用。在2020年的新冠肺炎疫情中，更是为医疗卫生科研人员提供了巨大的帮助，为抗疫斗争赢得了宝贵的时间。从GPU在新冠肺炎疫情中的实际应用情况出发，回顾了GPU诞生至今40余年的主要发展历程，分析其发展趋势，提出了未来可能的若干发展方向，并结合本单位GPU研制实践，阐述了国产GPU研制的特点及现状，列举了多项关键技术，对国产GPU的未来发展提出了展望。

关键词：图形处理器；光线追踪；科学计算；国产化

中图分类号：TP303　　**文献标识码：**A　　**国家标准学科分类代码：**510

History, future and practice of GPU

XIONG Tinggang

(The 709th Research Institute of China Shipbuilding Industry Corporation, Wuhan 430205, China)

Abstract: With the powerful computing power of GPUs, supercomputers are widely used in data processing, physical simulation, weather prediction, modern pharmaceuticals, gene sequencing, advanced manufacturing, artificial intelligence, and cryptographic analysis. In the 2020 outbreak of new coronapneumonia, supercomputers have provided tremendous help for medical and health researchers, and won valuable time for the fight against epidemics. Based on the practical application of GPU in the new coronapneumonia outbreak, this paper reviews the main development course of GPU since its birth for more than 40 years, analyzes its development trend, puts forward some possible development direction in the future, expounds the characteristics and status of Chinese GPU development based on the practice of my team, lists several of key technologies, and puts forward a prospect for the future development of Chinese GPU.

Keywords: GPU; ray tracing; scientific computing; localization

0 引 言

2020年伊始，全人类都在为抗击新冠肺炎而努力，在这场与病毒生死时速的竞赛中，速度至关重要，而以GPU集群为计算核心的超级计算机成为了抗疫竞速的加速器。近期，一篇利用超级计算机进行病毒研究的论文引发关注，美国橡树岭国家实验室的研究人员利用IBM的Summit超级计算机寻找可以对抗新冠肺炎的最有效的现有药物[1]。Summit是目前世界上最强大的超级计算机，它由4 608个计算节点组成，每个节点包含2个22核Power9 CPU和6个Nvidia Volta V100 GPU计算卡，单节点双精度浮点运算能力42TFLOPS，整个集群峰值运算能力超过200 PFLOPS。研究人员从8 000多种化合物入手，借

熊庭刚，研究员，主要研究方向为计算机体系结构与GPU芯片设计。E-mail：xtg_hb@aliyun.com

助Summit的算法寻找可以与蛋白质结合并阻止病毒发挥作用的药物，已筛选出77种化合物，这一筛选过程，如果手动进行，需要数年才能完成，如果在低速计算平台上进行需要数月，在超级计算机上，时间缩短到以天计，而这主要得益于GPU的超级计算能力[2]。GPU历经40多年的发展，已经从单纯的图形显示加速功能发展到如今日益丰富的广泛使用场景，本文综述GPU的发展历程及未来趋势，并介绍国内GPU的一些研制实践。

1 GPU发展历程

1999年，NVIDIA公司在发布其标志性产品GeForce256时，首次提出了GPU的概念，尽管如此，追溯GPU的历史，要从图形显示控制器说起。世界上第一台个人电脑IBM5150于1981年由IBM公司发布，这台PC搭载了黑白显示适配器（monochrome display adapter，MDA）和彩色图形适配器（color graphics adapter，CGA），这便是最早的图形显示控制器[3]。后来，IBM又推出EGA（enhanced graphics adapter），并于1987年提出了VGA（video graphics array）标准，它是IBM为PS/2系统中的Model 50、60和80机型所内建的显示系统，VGA在文字模式下可支持720×400分辨率，绘图模式下可支持640×480×16色和320×200×256色输出，VGA标准一直沿用至今。为了保证兼容性，当今的显卡依然会遵循VGA标准。

从MDA到VGA，图形图像的运算都由CPU来完成，图形卡的作用主要是将其显示出来。1991年，S3 Graphics推出的“S3 86C911”，正式开启2D图形硬件加速时代，它能进行字符、基本2D图元和矩形的绘制。到了1995年，几乎所有的显卡都具备2D加速功能，2D图形接口GDI、DirectFB等也都相继出现，并延续至今。

1994年，3DLabs发布的Glint300SX是第一颗用于PC的3D图形加速芯片，它支持高氏着色、深度缓冲、抗锯齿、Alpha混合等特性，开启了显卡的3D加速时代，然而这个阶段的显卡大多没有执行统一的标准，加速功能也不尽相同，直到NVIDIA推出GeForce256，它整合了硬件变换和光照（transform and lighting，T&L）、立方环境材质贴图和顶点混合、纹理压缩和凹凸映射贴图、双重纹理四像素256位渲染引擎等，并且兼容DirectX和OpenGL[4]，被称为世界上第一款GPU。硬件T&L的引入，极大减轻了CPU的负担，是这一时代GPU的标志。2001年微软发布DirectX 8，提出了渲染单元模式（shader model）的概念，根据操作对象的不同引入了2种Shader，分别是顶点着色器（vertex shader）和像素着色器（pixel shader），从此，硬件T&L被抛弃，进入shader时代，此时的GPU架构是固定管线[5]。

固定管线架构持续多年，直到微软推出DirectX 10。shader不再扮演固定的角色，每一个shader都可以处理顶点和像素，这就是统一渲染着色器（unified shader），它的出现避免了固定管线中顶点着色器和像素着色器资源分配不合理的现象发生，使得GPU的利用率更高。第一款采用统一渲染架构的GPU是ATI在2005年与微软合作的游戏主机XBOX 360上采用的Xenos，它是ATI第一代统一渲染架构，而真正具有影响力的，是NVIDIA在2006年发布的GeForce 8800 GTX（核心代号G80），它是第一款采用统一渲染架构的桌面GPU，其架构影响了日后的数代产品，是一款极具划时代意义的GPU[6]。

与G80一同发布的，还有著名的CUDA（compute unified device architecture），它能利用NVIDIA GPU的运算能力进行并行计算，拓展了GPU的应用领域，然而这时的CUDA只能算是GPU的副业。2011年TESLA GPU计算卡发布，标志着NVIDIA将正式用于计算的GPU产品线独立出来，凭借着架构上的优势，GPU在通用计算及超级计算机领域，逐渐取代CPU成为主角[7]。

GPU的发展历程如表1所示。

2 GPU未来趋势

GPU的未来方向，可以从NVIDIA 2019年的中国GTC（GPU technology conference）大会窥见一斑。GTC会议是NVIDIA近年来每年举办的一场GPU技术大会，汇集全球顶级的GPU专家，提供GPU领域颇具热门话题的相关培训和演讲。在这个大会上展示的是全球GPU研究人员的最新的研究和应用方

向,通过GTC会议可以窥见GPU的未来。2019年中国GTC大会设置了两大主题,分别是AI和图形,两个大主题之下各自又有一些小主题[8],如表2所示。

从表2不难看出,GPU的未来趋势无外乎3个:大规模扩展计算能力的高性能计算(GPGPU)、人工智能计算(AIGPU)、更加逼真的图形展现(光线追踪Ray Tracing GPU)。虽然GPU的最基本功能-显示技术在大会主题中没有“显式”的提及,但是众多应用方向均与之密切相关,譬如:智慧医疗和生命科学、游戏、虚拟现实/增强现实、工业设计与工程、自动驾驶与交通等,因此支持更加清晰和动感的高清显示是无需强调的未来趋势。此外,由于GPU越来越广泛地应用到手机、终端、边缘计算节点等嵌入式设备,所以高效能也是一个永恒的追求。

2.1 高性能计算

NVIDIA最新发布的Tesla V100s高性能计算GPU,集成5 120个CUDA Core,640个Tensor Core,采用32 GB HBM2显存,显存带宽达1 134 GB/S,单精度浮点计算能力达16.4 TFLOPS。

GPGPU在图形GPU的基础上进行了优化设计,使之更适合高性能并行计算,加上CUDA多年来建立的完整生态系统,其在性能、易用性和通用性上比图形GPU更加强大。基于这种特性,GPGPU将应用领域扩展到了图形之外,在自动驾驶、智慧医疗、生命科学、深度学习、云计算、数据处理、金融等方面均得到广泛应用,关于它的科研成果和新应用模式也层出不穷。

相比CUDA,OpenCL具有更好跨平台性和通用性,得到更多GPU硬件厂家的支持,但由于其对开发者的友好程度不高,直接应用反而不多。

2.2 人工智能计算

GPU的并行处理结构非常适合人工智能计算,但传统的基于流处理器的GPU,其流处理器一般只能处理FP32/FP64等精度的运算,而AI计算的精度要求往往不高,INT4/INT8/FP16往往可满足绝大部分AI计算应用。针对AI应用,NVIDIA设计了专用的Tensor Core用于AI计算,支持INT4/INT8/FP16等不同精度计算,RTX 2080集成了544个Tensor Core,INT4计算能力可达455 TOPS。

基于NVIDIA GPU的AI应用绝大多数情况下应用在服务器端、云端,基于GPU的AI计算往往具有更好的灵活性和通用性,在数据中心、云端等环境下具有更广泛的适用性。与之相对应的,在分布式应用领域AI计算更倾向于独立的面向特定应用领域的专用芯片,而不依赖于GPU,如手机、平板等移动端SOC都集成了专用的NPU IP。

2.3 光线追踪-更加逼真的图形展现

传统的图形GPU都使用光栅化技术显示3D物体,对物体进行3D建模,将其分割成若干三角形,三角形的细粒度很大程度上决定最后的成像质量,然后将三角形转换为2D屏幕上的像素点并分配初始颜色值,接下来进行像素处理,基于场景修改像素颜

表1　GPU发展历程

Table 1　GPU development history

时间	80年代	80年代末	90年代初	90年代后期	2004 ~ 2010	2011 ~ 至今
类型	图形显示	2D加速	部分3D加速	固定管线	统一渲染	通用计算
相关标准	CGA,VGA	GDI,DirectFB	OpenGL(1.1 ~ 4.1), DirectX(6.0 ~ 11)			CUDA,OpenCL1.2 ~ 2.0
代表产品	IBM 5150	86C911	Glint300SX	GeForce256	G80	TESLA
基本特征	光栅生成器	2D图元加速	硬件T&L	shader功能固定	多功能shader	完成与图形处理无关的科学计算

表2　2019中国GTC大会主题

Table 2　2019 China GTC subject

AI			图形
AI框架及应用	AI开发及工具	科学计算中的AI	光线追踪在游戏中的应用
数据中心,云计算和图形虚拟化		自主机器:机器人和物联网	工业设计与工程
边缘计算/AI+5G赋能行业		自动驾驶与交通	产品渲染和光线追踪
加速数据科学	智慧金融	智慧医疗和生命科学	AI　加速图形应用

色,并将纹理应用于像素,从而生成像素的最终颜色[9]。

光线追踪与光栅化的实现原理不同,它最早由IBM的Arthur Appel于1969年在“Some Techniques for Shading Machine Renderings of Solids”[10]中提出,光线追踪通过从观察点对每一个像素发射一条光线并找到在世界场景中阻挡光线路径的最近物体来渲染场景,光线有两种,第一种是视者发射的光线,来寻找场景中的交点,另一种是从交点发到灯光的阴影射线,看自身是否是处于阴影当中,光线追踪的一个显著优点是能够处理不平整的表面和固体。

2018年NVIDIA发布的RTX 2080 GPU,采用Turing架构,在GPU中集成了68个独立的RT(ray tracing) Core,用于光线追踪,光线处理能力达到了10 Giga/S,1 080P@60Hz需要处理的光线约为6 Giga/S,实测基于光线追踪的应用其帧率大致在50 FPS左右,基于RTX 2080的光线追踪达到了可用的程度,光线追踪对于反射和阴影有着更逼真的处理效果,尽管目前仍然是采用光线追踪和传统光栅图形处理相结合的方式来进行图形渲染,但其效果已经远超传统光栅图形处理,对于游戏、电影等追求逼真光影效果的应用,光线追踪能提供电影级画质的实时渲染,带来视觉效果上质的飞跃。

除了游戏、电影方面的应用,产品设计师和建筑师也可以享受到光线追踪带来的好处,借助光线追踪工具和高性能GPU,可以实时生成逼真的产品模型,提高设计迭代速度。

NVIDIA的下一代图形GPU,采用Ampere架构,计划于今年发布,相信在光线追踪方面带来新的提升。

2.4 高清显示

2.4.1 高刷新率

目前主流屏幕的刷新率为60 Hz,就是一秒能刷新60张画面,但近年来用户要求不断提高,游戏、电影都提出了90 Hz、120 Hz、144 Hz刷新率的要求,VR基于良好的用户体验也提出了120 Hz刷新率的要求,高刷新率能带来更加流畅连贯的画面显示效果,提供更好的感官体验,目前市场上已经推出了280 Hz刷新率的显示器,可见的未来显示刷新率会不断提高。

高刷新率对GPU带来了两个挑战,一方面需要每秒输出更多的像素数据,另一方面需要解决GPU与显示器刷新率不匹配造成的画面撕裂问题。对于第一个挑战,目前GPU采用了更快的显存如GDDR6/HBM以及提升GPU自身的处理能力以提升刷新率,同时在接口方面采用了PSR/PSR2等技术,即只对变化的像素点进行更新,以降低显示接口输出的压力。对于第二个挑战,AMD/NVIDIA使用Free Sync/G-Sync等技术,在显示器内安置一枚可与GPU直接通信的芯片,以协调显示器与GPU显示输出之间的数据同步,使显示器根据GPU的实际输出进行刷新率动态调节,以解决刷新率不匹配造成的画面撕裂问题。

2.4.2 高分辨率

目前2 K显示已经成为主流,但无论桌面端还是移动端4 K显示的硬件基础已具备,随着片源问题的逐步解决,未来4 K、甚至8 K显示必然会逐步普及,而VR则要求16 K乃至32K的分辨率以期给用户带来更好的沉浸感,最新推出的HDMI 2.1支持10 K显示,而DP 2.0显示接口已经能够支持16 K显示。

高分辨率给GPU的显示接口以及处理能力提出了更高要求,16 K模式下每一帧图像像素点达到了1.32亿,考虑同时存在的高刷新率需求,高分辨率对GPU的像素处理能力要求极高。随着消费端对沉浸式高分辨率显示的不断追求,GPU厂商需要进一步提升GPU图形处理能力以及显示接口的传输速率。

2.5 高效能

GPU擅长处理计算密集型任务,但大部分应用场景都需要在满足计算或者图形处理性能的条件下尽量降低功耗。传统的GPU架构将存储和计算分离,会遇到很多瓶颈:增加核心数量来达到高性能的方式,有芯片面积、功耗和可靠性的限制;纹理和顶点数据移动的功耗远多于图形计算的功耗。目前存内计算的方式,已经成功应用于人工智能领域中,来提升深度学习芯片的能效比。因此,本文提出了开展基于新型存储器的存算一体图形处理架构的研究,以提升图形处理器的能效比,对GPU的发展有积极意义。

3 国产GPU的研制实践

我国现有的绝大部分计算机中所使用的GPU均为美国芯片巨头(NVIDIA、AMD)所垄断,尽管在民用领域目前看来没有太大问题,但是在党政军办公和国民经济的关键领域,存在严重的信息安全隐患和供货保障问题。因此,亟需开展国产GPU的研制工作,并加速推广应用。

研究团队针对上述GPU的发展现状及未来趋势,多年来展开了一系列关键技术的研究,包括可扩展的科学计算与图形渲染统一架构、多核多线程调度与管理、生态环境建设、国产计算机平台适配与优化等等,研制了多款国产GPU芯片。在显示方面,提出了一种基于图层的高刷新率高分辨率显示技术,满足了比较广泛的高清图像显示应用需求。未来,更多高清3D应用的出现将带来GPU高清图形图像显示需求的持续增长,对GPU处理能力也是一项不小的挑战,持续改进GPU系统架构和设计方法,提高运算能力和综合显示能力,以应对高清显示的发展要求。

4 结论与展望

近些年,国外GPU技术快速发展,已经大大超出了其传统功能的范畴。国内GPU芯片的研制虽然可满足目前大多数图形应用需求,但在科学计算、人工智能及新型的图形渲染技术方面仍然和国外领先水平存在较大差距,未来持续发展国产GPU势在必行。

国产GPU下阶段的发展方向可以主要考虑3个方面:第一,进一步提升图形图像显示水平,提升国产GPU的基本能力;第二,扩展科学计算和人工智能计算能力,增强国产GPU的非传统功能;第三,建设全系统解决方案及生态系统,寻求用户的最优体验。

参考文献

[1] MICHOLAS S, JEREMY C S. Repurposing therapeutics for COVID-19: supercomputer-based docking to the SARS-CoV-2 viral spike protein and viral spike protein-human ACE2 interface[EB/OL]. (2020-03-11)[2020-04-09]. https://doi.org/10.26434/chemrxiv.11871402.v47.

[2] 肖漫. IBM超级计算机筛选出77种抗病毒化合物,成抗疫新力量[EB/OL]. (2020-03-22)[2020-05-15]. https://www.leiphone.com/news/202003/vqpKTghBXrgVB2iA.html.
XIAO M. IBM supercomputer selected 77 kinds of antiviral compounds, new forces into the fight against COVID-19[EB/OL]. (2020-03-22)[2020-05-15]. https://www.leiphone.com/news/202003/vqpKTghBXrgVB2iA.html.

[3] BRIDGES R A, IMAM N, MINTZ T M, et al. Understanding GPU power: a survey of profiling, modeling, and simulation methods[J]. ACM Computing Surveys, 2016, 49(3): 41:1-41:27.

[4] NVIDIA. NVIDIA launches the World's first graphics processing unit: GeForce 256[EB/OL]. (2002-01-11)[2020-04-10]. https://www.nvidia.com/object/IO_20020111_5424.html.

[5] OWENS J D. GPU architecture overview[C]// International Conference on Computer Graphics and Interactive Techniques. ACM, 2007.

[6] MACRI J. AMD's next generation GPU and high bandwidth memory architecture: FURY[C]// Hot Chips Symposium. IEEE, 2015: 1-26.

[7] HU L, CHE X, ZHENG S Q, et al. A closer look at GPGPU[J]. ACM Computing Surveys, 2016, 48(4): 1-20.

[8] BENAMOU J. Big ray tracing[J]. Journal of Computational Physics, 1996, 128(2): 463-474.

[9] APPEL A. Some techniques for shading machine renderings of solids[C]// Fall Joint Computer Conference, 1968: 37-45.

[10] ATWELL C. AMD unveils freesync at CES2014[J]. Design News, 2014, 69(3): 28-29.