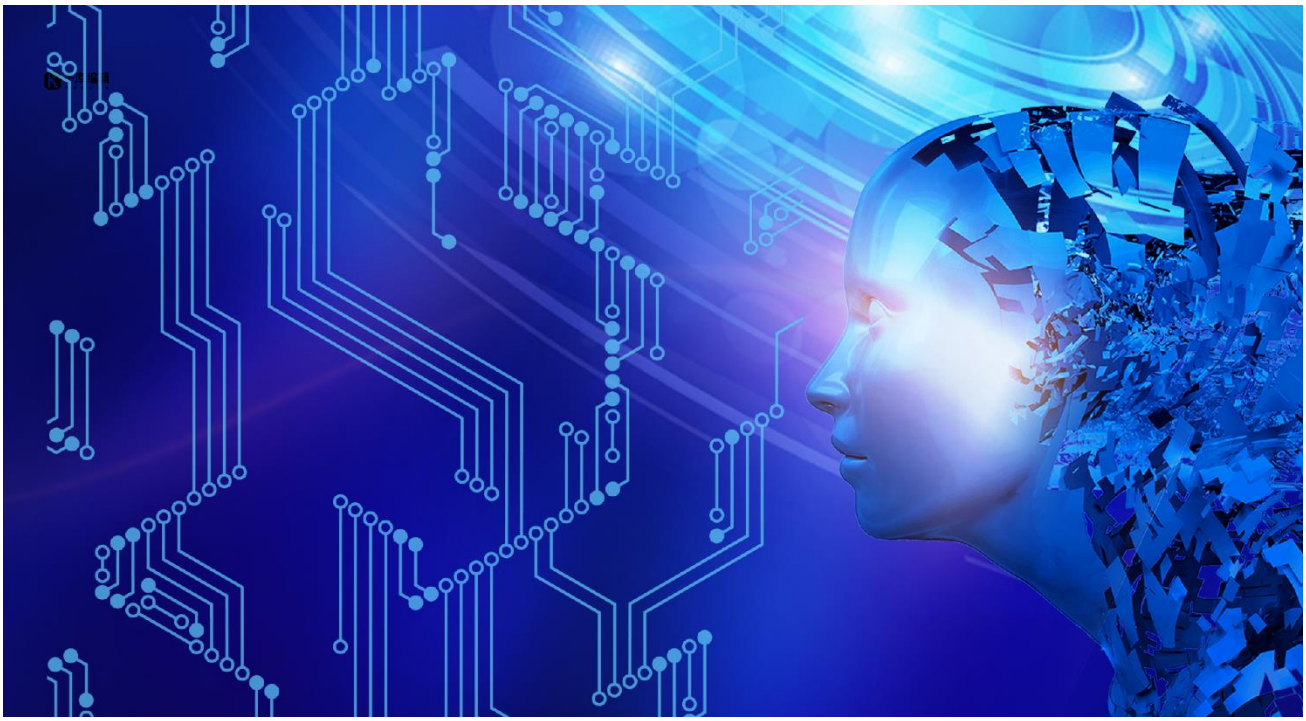
芯跑科技
iRUN VENTURE CAPITAL
立德家园使命感 点燃产业满天星

芯跑科技丨月报资讯

2023 年 12 月，第 048 期：AI 主题研究（二）GPU 发展和国产 GPU 厂商梳理



AI 主题研究（二）GPU 发展和国产 GPU 厂商梳理

芯跑科技研究部 2023.12

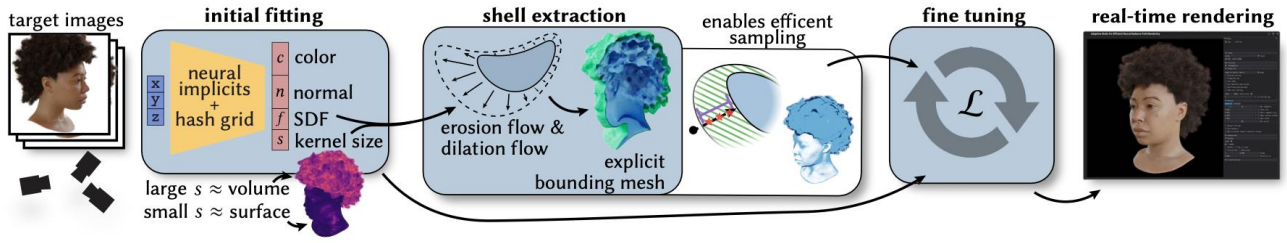
2023 年 11 月，Nvidia 的 Zian Wang 等人发表了《Adaptive Shells for Efficient Neural Radiance Field Rendering》，这是神经辐射场（Neural Radiance Field，NeRF）研究的最新进展。NeRF 出现已经有了数年，但是直到今年 Nvidia 的研究发表，才真正实现了 GPU 上高帧率（100-200fps）高质量 NeRF 渲染，因此今年有望是 NeRF 渲染真正进入实用的开始。NeRF 渲染是高性能 3D 图像渲染的一个发展方向，业界另一个研究方向是法国 INRIA 和德国马克思普朗克研究所完成的 3D Gaussian Splatting（3D GS）。高性能 3D 图像渲染和 3D 图像学一直是推动 GPU 芯片发展的主要动力，对于真实场景和物体的高精度建模/渲染一直是整个学界孜孜不倦追求的目标。

自 GPU 被广泛用于 AI 算法训练，推理和特定算法加速之后，国产 GPU 厂商也逐渐向三个方向发展：传统 GPU，AI GPU 和特定应用方向的 DSA 芯片。

一、 传统 GPU

如常见的桌面 GPU，移动 GPU 等，主要应用是渲染，图像/图形处理，同时也兼顾一定 GPGPU 的运算能力。在这个方向上国产 GPU 厂商通过自研，兼并和授权的方式布局较早、研发积累较长，已经逐步形成了一系列产品，特别是在重点行业的自主可控上已经具有一定的基础。

而 NeRF 和 3D GS 都是 3D 图像学的最新发展，也会逐步影响到传统 GPU 的研发。在过去几十年中，3D 场景和物体建模的主流方式是基于多边形（ploygon mesh）的建模，即把一个 3D 建模的物体表面近似为由大量多边形组成，而多边形数量越多，则 3D 建模和真实物体/场景越接近。在具体 3D 模型创建方面，则是由 3D 建模人员创建模型。在这样的情况下，多边形数量一方面限制了渲染的性能（即主流设备难以渲染多边形足以表现全部真实场景细节的模型），另一方面从建模方面，3D 建模人员也难以完成含有真实物体或者场景所有细节的 3D 模型。因此目前看到的主流 3D 模型都和现实的真实场景有显著区别。



NeRF 渲染流程示意图，参考文献 1

人工智能图像学对于 GPU 提出了新的需求。首先，在基本的 NeRF 或者 3D GS 的渲染中，传统的 GPU 中的多边形渲染流水线已经无法高效支持，因为 NeRF 和 3D GS 的渲染需要一些重要的新计算。对于 NeRF 来说，其场景建模信息都包含在训练过的神经网络中，神经网络的输入就是用户当前的视角，输出则是场景在视角下的 2D 图像。因此，其渲染过程其实就是根据用户的视角来完成神经网络的推理计算。而在 3D GS 中，具体的渲染过程则是把整个场景分成多个块（tile），每个块中根据当前视角首先排序选出对于视觉影响

最大的 N 个 GS，之后再仅仅针对这些 GS 做渲染，从而可以实现高效率。可以看到这些都和当前的多边形渲染流水线有较大不同，为了能高效支持这些 3D 图像学的新范式，GPU 需要能高效支持这些新计算。另外，在新的 3D 图像学是由人工智能驱动的这一潮流下，势必会看到 3D 图像渲染和人工智能的进一步结合，例如在 NeRF 和 3D GS 的场景建模中加入基于神经网络计算的动画或者编辑（光影变化等），这些又进一步说明目前的 GPU 上的多边形渲染流水线对着这类新图像渲染范式已经无法高效支持。

这些新的超高精度 3D 图像学会推动新的 GPU 架构发展，对国产 GPU 厂商既是挑战也是机遇。一方面新架构的演进是技术发展的必然结果，不管是 NeRF 还是 3D GS，抑或是未来新发展出 3D 技术，GPU 架构演进的规律是推陈出新。这势必影响到目前已经日趋成熟的国产 GPU 厂商的研发。另一方面，GPU 架构的选择跟 3D 图像学算法的研究强相关，随着国内 3D 图像学的研究深入，新的 GPU 架构也是国产 GPU 厂商参与新架构定义和竞争的重要方式。

二、 AI GPU

AI GPU 不再以图像处理为主要任务，Nvidia 主导的 GPGPU 在人工智能浪潮的前几年（2012-2017）是 Nvidia 能够占据人工智能霸主地位的核心，因为 GPGPU 的开放接口可以让 GPU 去做人工智能计算。在这之后，随着人工智能应用进入主流地位，Nvidia 开始给人工智能做专用优化，引入了包括 Tensor Core 等重要新架构，人工智能在 Nvidia 的 GPU 上已经不再主要依赖其 GPGPU 思路，而是更多依赖 Nvidia 的人工智能架构设计。

Nvidia GPU 也成为人工智能三大基石中算力的代表。这带来的直接结果就是如 A100、H100 的热销，Nvidia 数据中心业务的增长和股价的飙升。特别是 ChatGPT 引爆全球对人工智能新的期许以后，OpenAI 甚至上演了对人工智能未来方向的斗争。国内各大厂在大模型上的布局也激起国产数据中心厂商的算力储备，进而也传导到 AI GPU 需求上。国产 GPU 厂商布局 AI GPU 产品集中在 2015 年之后，业界比较出名的初创 AI GPU 厂商基本都成立在这段时间，而早期 GPU 厂商也大多在近几年开始推出 AI 加速芯片。

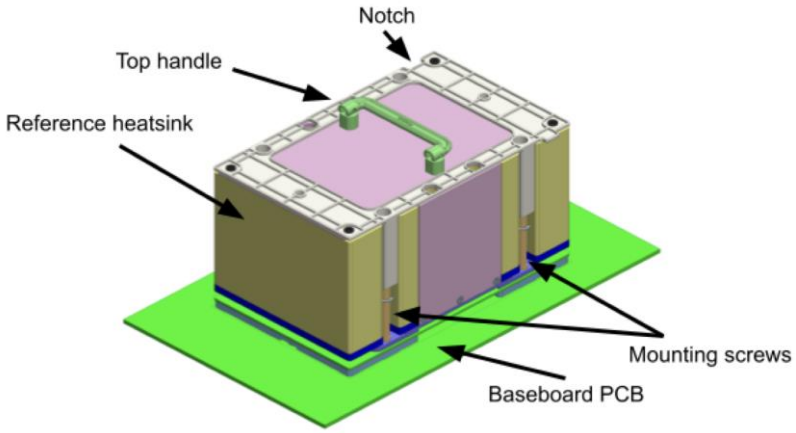
随着中美贸易战的开展和升级，如 A100，H100 的高端 AI GPU 上了美国限制出口的名单，尽管 Nvidia 迅速推出了可以绕过监管的改进型号产品 A800 和 H800，美国政府的监管政策也进一步收紧到对总性能和算力密度的管控，至此不但高端 AI GPU 上榜，连 GP 算力较强的 RTX 4090 也被禁售了。与此同时，国产 AI GPU 公司如壁仞科技，摩尔线程等也受到影响，从流片，EDA 工具到 IP 授权都受到了限制。

Nvidia 在 AI GPU 上布局多年，国产 GPU 厂商在 AI GPU 的航道上主要有三方面的难关需要攻克。

首先在硬件角度，高端 AI GPU 严重依赖先进制程工艺，目前国产 AI GPU 主要集中在 7nm 和 12nm 工艺。Nvidia 现有产品 H100 芯片采用 TSMC 4nm 工艺，而下一代产品 B100 芯片计划在 2024 年 4 季度推出，将采用 TSMC 3nm 工艺。在当前美国加严技术管控的前提下，国产 AI GPU 公司可能在先进制程流片角度受到较大影响。如果不能使用先进制程，AI GPU 的散热和性能就会在一定程度上打折扣，这会影响到国产 AI GPU 的部署和应用场景。而美国一旦将国产 GPU 公司放入管控的实体清单，除了流片，还会影响到相应公司的 EDA 软件使用和 IP 授权，进而更深层次的影响到芯片设计和开发。

软件角度，CUDA（Compute Unified Device Architecture）是英伟达自 2006 年推出的一种通用并行计算平台和编程模型，可以让开发者利用 GPU 的强大计算能力来加速各种类型的应用计算。CUDA 提供了一套完整的软件工具和库，支持多种编程语言和操作系统，简化了 GPU 编程的难度和复杂度。CUDA 还提供了一套丰富的生态系统，包括各种深度学习框架（如 TensorFlow、PyTorch、MXNet 等）、机器学习库（如 cuDNN、

cuML、RAPIDS 等)、科学计算库（如 cuBLAS、cuFFT、cuSPARSE 等）等，可以帮助开发者快速构建和部署 AI 应用程序。时至今日，CUDA 已成为全球 AI 基础设施，主流的 AI 框架、库、工具都以 CUDA 为基础进行开发。国产 AI GPU 在推广过程中普遍采用兼容 CUDA 的方式，以降低客户跨平台算法部署的难度。



OCP OAM 示意图，参考文献 2

除了标准的 PCIe 形态 GPU 卡之外，要想获得更高的 GPU 计算性能，就需要使用多 GPU 互联的 OCP OAM 模块。可以说 OCP OAM 是 GPU 服务器在接下来的几年里最重要的技术发展方向之一。而当 2020 年 ChatGPT-3 大模型横空出世的时候，人们意识到大模型千亿级别参数的训练，只有把 GPU 服务器级联起来，构成 GPU Pod，才能有足够的资源在可接受的时间范围内完成 AI 模型训练。NVIDIA 在 GPU 互联方向上的解决方案是 NVLink 接口和 NVSwitch 互联芯片。NVLink 接口最早在 2014 年推出，目前版本是 4.0。国产 GPU 目前还没有互联芯片量产，因为 NVLink 是个私有协议，需要 Nvidia 授权，国产 GPU 可以考虑 PCIe/CXL Switch 的方向。

模型名称	参数大小	MMLU平均分	CEval平均分	AGIEval平均分	GSM8K平均分	发布者
PaLM	5400.0	69.3	/	/	56.5	
PaLM 2	3400.0	78.3	/	/	80.7	
Gemini-ultra	3000.0	90.04	/	/	94.4	
GPT-3	1750.0	53.9	/	/	/	
GPT-3.5	1750.0	70.0	54.4	/	57.1	
GPT-4	1750.0	86.4	68.7	/	92.0	
OPT	1750.0	25.2	25.0	24.2	/	

大模型参数数量对比，参考文献 3

三、 DSA

国产 GPU 厂商的第三个发展方向是针对特定应用领域的 DSA。从长期来看，芯片设计总有一条通路是从通用、高性能，向专用、高性价比发展的。不管是图像处理的应用，还是 AI 算法云端、边缘的加速，一旦算法和应用场景成熟，就会催生出用户和设计公司走向 DSA，甚至 ASIC。

国产 GPU 厂商很早便在视频相关 DSA 上发力，在成熟确定的视频编解码基础上加入可编程的 AI 算法模块。针对体积小、能耗少、性能略低的应用场景，例如集成于工控机、工作站中，支持计算机视觉、视频处理和自然语言处理等领域，发挥 DSA 同时具备高并发视频解码能力，来支持工业视觉、智慧交通、智慧医疗等行业。这方向的国产 GPU 厂商也面对来自视频行业芯片开发企业的竞争，其中关键点在于相应的 AI 算法和加速硬件的掌握程度。

此外 AI DSA 也是国产 GPU 厂商不断拓展的方向，包括 AI 加速卡。AI DSA 不期待取代 CPU 和 GPU，而是在特定任务上实现高得多的效率，是体现在真实 workload 下的实际/W 性能和/$性能，即更高的性价比。相对的，牺牲灵活性和对更多应用场景的覆盖程度。AI DSA 向这个方向发展，最终期待运算模式简单的 AI 模型能够覆盖更多的应用，这样专用的 AI DSA 系统就可能真的覆盖更多的市场。

总体而言，国产 GPU 公司在过去几年中经历了较大发展和成长，已经逐步形成了在传统 GPU，AI GPU 和 DSA 方向的布局。接下来整理了目前主要国产 GPU 厂商名录供参考：

公司	成立时间	当前轮次	产品	应用方向
登临科技	2017 年 11 月	B 轮	Goldwasser-UL(MXM)，Goldwasser-UL，Goldwasser-L	自动驾驶、智慧城市、生物医疗、AI 计算等领域
壁仞科技	2019 年 09 月	B 轮	BR100GPU	人工智能、云计算、图形渲染、大数据等领域
海光信息	2014 年 10 月	688041.SH	深算一号 DCU 产品	人工智能训练等领域
寒武纪	2016 年 03 月	688256.SH	MLU370-X8	人工智能训练等领域
海飞科	2020 年 08 月	B+	Compass C10	云计算，计算机视觉，自然语言，语音处理，科学计算
沐曦集成	2020 年 09 月	Pre-B	MXN（推理）MXC（GPGPU）MXG（渲染）	深度学习、数据分析、物理仿真、云游戏等领域
摩尔线程	2020 年 06 月	B+	MTTS60. S70. S80.MTTS2000. MTTS3000	物理仿真、AI 计算、游戏娱乐、自动驾驶等领域
芯动科技	2007 年 10 月	A 轮	风华 1 号、风华 2 号	办公上网、娱乐游戏、工程制图、GIS 等领域
上海兆芯	2013 年 04 月	C 轮	KX-6000 中的 C-960GPU	高性能桌面、便携终端、嵌入式等领域

智绘微电子	2018 年 12 月	Pre-A	IDM919	图形图像显示、科学计算、人工智能等领域
芯瞳半导体	2019 年 12 月	A 轮	GenBu01GPU，GB2062 GPU	信创、能源、交通、金融、人工智能等领域
深流微智能	2021 年 05 月	A 轮	XST-G01. XST-E01.XST-C01	视觉计算、人工智能、高性能计算等领域
砺算科技	2021 年 08 月	Pre-A	TrueGPU 图形芯片	端、云、边、车等领域
芯原股份	2001 年 08 月	688521.SH	Arcturus GC8800.GC8400. GC8200.Gc8000, GPU Nano IP	小型物联网 MCU、人工智能等领域
景嘉微	2006 年 04 月	300474.SZ	JM5400. JM7201.JM92 系列	军用工业、人工智能、云计算金融、教育等领域
龙芯中科	2008 年 03 月	688047.SH	7A2000 桥片的集成 GPU	金融、政务办公、网安、教育、通信、医疗等领域
凌久电子	2007 年 08 月	/	GP101	桌面显示，适配国产 CPU
中微电	2009 年 04 月	C 轮	南风一号，二号，三号	桌面显示，PC 显卡，满足党政，金融和安防需求
杭锦科技	1997 年 09 月	000818.SZ	SG****0、SG****1	面向特殊领域市场和通用领域的市场需求
格兰菲	2020 年 12 月	A 轮	Arise-GT10C0	桌面、商业显示以及通用计算等
飞腾	2014 年 08 月	B+	X100 套片	桌面显示
红山微电子	2019 年 04 月	B 轮	Sc3 系列	汽车、互联网、云计算、机器人、医药、零售、工业
瀚博半导体	2018 年 12 月	B 轮	SV100， VA1 视频加速卡	视频领域
象帝先	2020 年 09 月	A+	天钧一号，天钧二号	AIOT 市场，适用于移动、桌面、边缘计算、云数据中心等
速显微	2015 年 12 月	C 轮	天都一号	智能座舱，家居家电
西安翔腾	2004 年 07 月	A 轮	HKM9000	座舱显控

GPU 做为算力的代名词，在当今人工智能呼之欲出的场景下，承担着突破封锁，承前启后的作用。中国

AI 算法研究，AI 生态建立，AI 产业的蓬勃发展，很大程度上也与国产 GPU 的发展息息相关。国产 GPU 通过多年的发展已经在桌面、移动 GPU 上取得了一定布局。但是在 AI GPU 方面，国产 GPU 还面对三个主要的困难要克服，特别是制程问题，背后实际是我国在先进制程方向遇到各种问题的缩影。美国实体清单管控的加码、断供的升级，既影响到国产 GPU 的发展，也迫使我们不断推陈出新，调整国产 GPU 的发展策略。人工智能的发展也进一步催生出 DSA 在端侧的应用机会，未来在端侧释放出大量算力，反过来也会促进 AI 算法对端侧的覆盖和迭代。相信随着先进制程工艺技术的突破，国产 GPU 的发展也会强力促进我国人工智能在数据中心和端侧的应用推广。

参考文献：

1. 《Adaptive Shells for Efficient Neural Radiance Field Rendering》, Zian Wang 等, Nvidia
2. OCP OAM 设计标准 v1.5, COP 官网
3. 大模型综合能力评测对比表，数据学习
4. A100, H100, NVLINK and NVSwitch, Nvidia 中国官网
5. 《GPU，巨变前夜》，半导体行业观察
6. CUDA 专区，Nvidia Developer 频道
7. 《AI DSA 芯片的发展方向》，唐杉，知乎专栏
8. CXL™ 3.1 Specification, CXL association