

内容目录

第一章 前言 2

第二章 2023-2028 年 GPU 市场前景及趋势预测 3

 第一节 国产 GPU 发展窗口期已至，生态构筑核心优势 3

 第二节 CUDA：GPU 生态先驱，AI 时代基础设施 4

 第三节 并行计算软件库覆盖度：以 ROCm 为例 5

 一、ROCm 前为何没有能挑战 CUDA 者？ 6

 二、软件不足：OpenCL 对 HPC 支持有限，产品节奏长期滞后 6

 三、硬件不足：早期 MI 产品性能待提高，对数据中心用户吸引力有限 6

 第四节 AMD 系算力走向实用化的转折点：重建基础平台，从 OpenCL 转向 HIP..... 8

 第五节 ROCm 如何融入 CUDA 生态：HIP 通用前端代码+Hipify 转换工具 8

 第六节 ROCm 可以多大程度上替代 CUDA：支持日渐完整，体验仍需进步 10

 一、软件库支持：支持范围持续扩大，基本支持 AI 与 HPC 生态 10

 二、基础软硬件支持：Windows 支持或将面世，CDNA 与 RDNA 架构支持提升 10

 第七节 性能表现与易用性：ROCm 在 AI 框架领域性能无短板，易用性尚需提升 11

 第八节 总结：张量计算待提升，基础软硬件与软件库支持需完善 12

 一、ROCm 的未来：开源是后发者的合理选择，或可参照 Python 与 MATLAB..... 13

 二、AI 框架支持：以昇腾为例，算法库+配套软件基建 14

 第九节 国内厂商生态进展：或复刻 CUDA，或兼容框架 15

第三章 GPU 企业短视频内容制作策略建议 17

 第一节 如何做出优质短视频？ 17

 一、账号定位 18

 二、短视频封面 18

 三、标题 18

 四、短视频内容 19

 第二节 怎么制作高质量短视频？ 19

 一、拍摄前的准备工作 20

 二、拍摄需要注意什么？ 21

 三、如何提升拍摄内容质量？ 21

四、运镜技巧 22

五、剪辑短视频需要注意什么？ 23

六、剪辑短视频遵循以下规则 23

第三节 如何去做高质量短视频 24

一、打造新奇玩法 24

二、拍摄自身企业的内涵段子 24

三、企业服务特色 24

四、邀请知名博主或者主播来访 24

第四节 制作什么样的视频更容易吸引用户？ 25

一、高颜值，帅哥和美女 25

二、分享 25

三、技能类教学 25

四、创意类	25
五、内容选定完成之后还有细节上的处理	25
第四章 GPU 企业《短视频内容制作策略》制定手册	26
第一节 动员与组织	26
一、动员	26
二、组织	27
第二节 学习与研究	27
一、学习方案	27
二、研究方案	28
第三节 制定前准备	29
一、制定原则	29
二、注意事项	30
三、有效战略的关键点	31
第四节 战略组成与制定流程	33
一、战略结构组成	33
二、战略制定流程	34
第五节 具体方案制定	35
一、具体方案制定	35
二、配套方案制定	37
第五章 GPU 企业《短视频内容制作策略》实施手册	38
第一节 培训与实施准备	38
第二节 试运行与正式实施	38
一、试运行与正式实施	38
二、实施方案	39
第三节 构建执行与推进体系	39
第四节 增强实施保障能力	40
第五节 动态管理与完善	41
第六节 战略评估、考核与审计	42
第六章 总结：商业自是有胜算.....	42

第一章 前言

在互联网短视频时代，视频内容天然有很强的吸引力，同时其通过创造诱惑力的特点能更快促进转化。

那么制作什么样的视频更容易吸引用户？

下面，我们先从 GPU 行业市场进行分析，然后分析了短视频时代什么样的效果广告好？短视频营销相较于传统营销的优势，以及最重要的：怎么做好短视频的效果广告？如何去做高质量短视频？

相信通过本文全面深入的研究和解答，您对这些信息的了解与把控，将上升到一个新的台阶。这为您经营管理、战略部署、成功投资提供有力的决策参考价值，也为您抢占市场先机提供有力的保证。

第二章 2023-2028 年 GPU 市场前景及趋势预测

第一节 国产 GPU 发展窗口期已至，生态构筑核心优势

AI 发展已经带动 GPU 行业高速发展，国产 AI 也为国产 GPU 的发展提供契机。在国际供应链不确定性背景下，我们认为，未来三年可能是国产 GPU 发展的关键窗口期，这一段时期，国产 GPU 有望取得长足发展，有可能是较好的投资时期。目前，国内 GPU 领域公司数量较多，从性能指标方面来看，已有部分公司能够在理论硬件性能方面接近国际主流水平，理论算力指标较高。

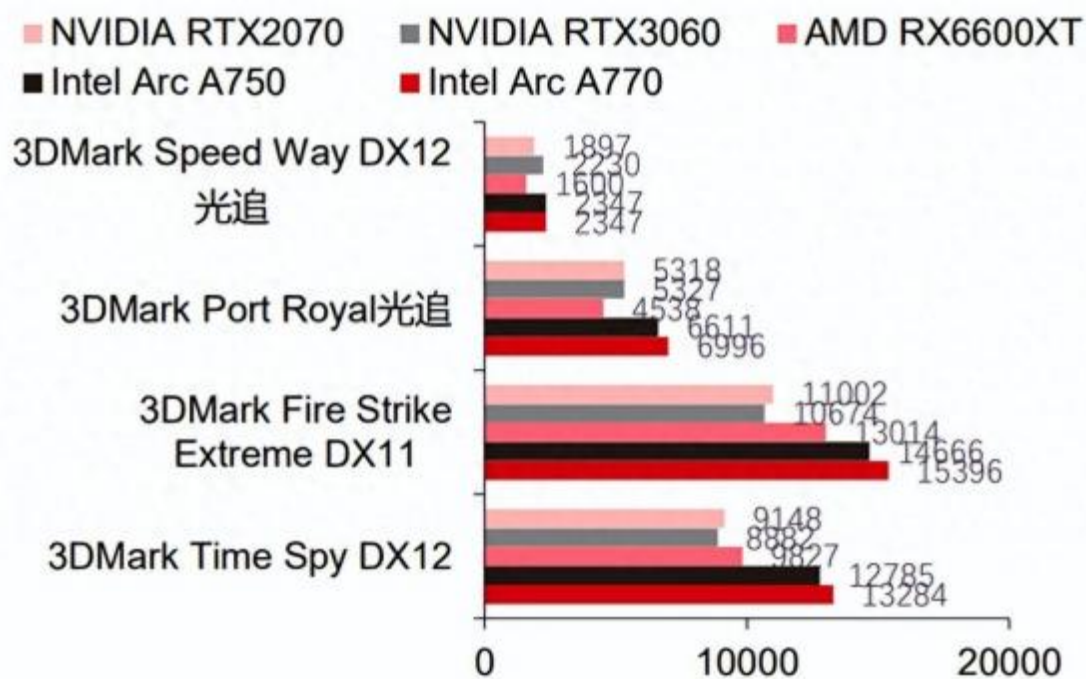
同时，在芯片对外带宽方面，英伟达也并非完全高不可攀。例如，根据 AMD 官网，其 MI 系列产品的带宽就经历了快速进化，从 MI50 到 MI100 再到 MI250，其 Infinity Fabric 通道就从 2 个提升到 8 个，对外带宽从 184GB/s 提升到 800GB/s。国内产品在通信方面也进行了尝试，例如壁仞科技的 BR100 芯片，其 BLink 技术就实现了 8 通道，合计 512GB/s 的对外带宽，技术定位类似于英伟达的 NVLink。

但尽管国产芯片的理论算力和理论带宽均能够做到较为良好的水平，理论性能却往往面临软件生态的限制，即便集成 GPU 领域的领先企业 Intel 也无法摆脱这一规律。Intel 新发布的独显 Arc 系列在使用了更多晶体管的前提下，具备更高的理论性能，但实际使用性能相比友商的基础型显卡有一定差距。

我们从软件测试中可以看见，3Dmark 测试所代表的 GPU 理论性能测试中，Intel 的 Arc 系列均有亮眼表现，而到了实际使用场景中，Intel Arc 系列产品的实际帧率相比同样的友商产品则有一定差距，重要原因在于驱动程序，而驱动程序正是生态的核心组成部分之一。

让每个人都能成为 行业专家、管理专家、营销专家、战略专家、成功企业家……

图 2：理论性能表征——3Dmark 跑分对比



另外生态在数据中心与开发者场景中起到更重要的作用，能够重塑整个数据科学/AI 工作流程，加强开发者黏性，形成正反馈。在开发者的日常流程中，首个环节是数据管理，包括数据从数据来源的提取（Extraction）、变形（Transform）、加载到应用端（Load），合并称为 ETL，随后还有数据的存储等。其次是数据训练、验证（可视化）、部署（推理）等多个环节。足够良好的 GPU 生态能够极大影响上述工作流，首先可以将模型训练环节迁移到 GPU，随后还可以将 ETL 环节（传统上由 CPU 完成）也迁移到 GPU 上，让整个数据科学/AI 工作流几乎全部在 GPU 上完成。

通过发达的 GPU 软件支持实现工作流迁移，能够极大提高效率，从而重塑开发者的软件使用习惯，让开发者的黏性极大增加，形成正反馈，持续提高软件生态的壁垒。

形成上述正反馈的前提是在每个环节中都必须有足够丰富、可靠、易于使用的工具供开发者使用，且最好是开发者熟悉、拥有成熟社区的，由此可见生态的重要性。

第二节 CUDA：GPU 生态先驱，AI 时代基础设施

让每个人都能成为

行业专家、管理专家、营销专家、战略专家、成功企业家……

构建在 CUDA 的基础之上。CUDA 也是软件生态的标杆，从软件库的覆盖面、AI 框架和算子库的支持程度两方面来讲，都是目前最完善的。CUDA 的诞生：实现从 GPU 到 GPGPU 的转变。在

CUDA 问世之前，想要调用 GPU 的计算能力必须编写大量的底层语言代码或借用图形 API，对使用高级语言为主的程序员十分不便。英伟达公司的首席技术官 David Kirk 以及 CEO 黄仁勋主导推出了 CUDA。2006 年 CUDA 发布，2007 年正式推出 CUDA1.0 公测版本。此后程序员无需再通过图形 API 来调用 GPU，而是可以直接采用类似 C 语言的方式直接操控 GPU。2008-2010 年，

CUDA 平台进一步发展，拓展了新局域的同步指令、扩充全速常量内存并且支持递归，NVIDIA 向各软件厂商免费提供开发工具，使得 CUDA 生态初具规模。

发展：十年苦坐冷板凳，与 AI 生态深度绑定。在 CUDA 系统诞生之初，英伟达大量投入研发对 CUDA 进行不断更新与维护，但并没有立即得到市场的认可，CUDA 用户不多但大量研发却直接影响利润，英伟达市值长期不振，直到 2016 年市值才超过 2007 年的水平。但 CUDA 对科学计算领域的渗透一直在进行，2012 年深度学习革命开始以来更是一直在渗透 AI 领域，2012 年 ImageNet 挑战赛冠军使用的就是英伟达 GTX580 GPU。此后 CUDA 持续迭代，2023 年 CUDA 已经推出最新的 12.0 版本 API。如今整个 HPC 与 AI 生态都已经与 CUDA 形成了深度绑定，CUDA 已经成功塑造了用户习惯，成为了 AI 时代开发者最熟悉的工具箱。

CUDA 生态：组件庞杂、取代难度高。CUDA 所包含的生态组分众多，包含编程语言和 API、开发库、分析和调试工具、数据中心和集群管理工具，以及 GPU 硬件等多个大类。每一大类中都包含了大量的组件，是英伟达以及开源生态开发者们在二十年间日积月累所形成，如今要将其取代难度较大，需要巨量的时间和资源投入。CUDA 生态虽已成为 AI 和 HPC 几乎不可分割的一部分，但英伟达占据了绝大部分市场份额，对下游用户来说，其使用成本也不低（根据英伟达财报，近 5 年来其毛利率基本维持在 60%上下），用户有意愿寻求低成本替代。且在当前国际局势下，中国本土企业面临一定的供应链不确定性。因此，对于诸多企业而言，在 CUDA 之外寻找一个第二选项，是对于企业经营可行的选择，要做到这一点，就需要尽可能跨越 CUDA 的护城河。

CUDA 的两大生态护城河：软件库覆盖率、AI 框架支持度。对于后来者而言，关键是采取切实可行的措施向 CUDA 学习靠拢，进行自身生态建设。我们认为，CUDA 的第一个壁垒是软件库覆盖

率，既包含了基础的并行计算软件库，也包含了细分行业软件库，还有配套的辅助软件等；第二个壁垒则是针对 AI 领域而言的 AI 框架支持率。对于我国芯片公司而言，中短期内可以尽可能投入资源实现的是并行计算软件+辅助软件库开发，以及对 AI 框架的完善支持，而对于细分行业的支持，则需要较长的时间和积累。

第三节 并行计算软件库覆盖度：以 ROCm 为例

目前，有望替代 CUDA 的选项已经存在，即 ROCm。ROCm 是 AMD 开发的对标 CUDA 的软件平

让每个人都能成为

行业专家、管理专家、营销专家、战略专家、成功企业家……

台。ROCm 全称为 Radeon Open Computing platform，是基于 AMD 系列 GPU 设计的开源计算生态，同时也能够相当大程度上兼容 CUDA、支持 NVIDIA GPU，其目标是建立有望替代 NVIDIA CUDA 生态的平台。以下我们将以 ROCm 为例，分析生态的第一个核心要素——并行计算软件库覆盖度，包括软件支持覆盖范围的变化及其影响，未来的演变趋势等。

一、ROCm 前为何没有能挑战 CUDA 者？

OpenCL 自身不足，AMD 旧产品性能待提升

ROCm 出现之前，AMD 也意识到 GPU 在 AI 计算领域的潜力。根据五二电子工作室网站信息，AMD 曾经尝试利用 OpenCL 打入深度计算领域，先是尝试移植 cuda-convnet，后转向 Caffe，最后的成果是 OpenCL Caffe，虽然是有益的尝试，但这项努力并没有引起很多的关注，其成果也局限在单一工具，并未构建起有效的平台。此前尝试失败的一个重要原因在于 AMD 此前一直使用 OpenCL，而 OpenCL 本身具有不足之处：语言支持少（2015 年以前只支持 C），更多面向底层软件而非上层软件。另外硬件层面，主导 ROCm 生态的 AMD 此前在计算卡系列仍然采用 GCN 架构，产品性能相对不足。

二、软件不足：OpenCL 对 HPC 支持有限，产品节奏长期滞后

OpenCL 最早提出是为了应对单核性能极限与多核并行计算的发展，最初由苹果开发，并与 AMD、Intel、英伟达、高通、IBM 的团队合作完成 API 方案，各家公司在 2008 年合作成立 Khronos 计算工作组，并于当年发布 OpenCL 标准，各公司采用统一的 API，但各自完成实现方式。需要注意的是，这时 GPU 的算力价值还没有被广泛认知，Khronos 成员除了英伟达外均以 CPU 为主业，OpenCL 也更多考虑了多核 CPU 计算，导致 OpenCL 更多考虑底层硬件驱动/操作系统支持，对高层的科学计算库等支持则相对有限。这一点从支持的语言也可以看出，OpenCL 长期以来仅支持 C 语言（底层驱动、操作系统等多使用 C），而科学计算库常用的 C++支持直到 2015 年的 OpenCL2.1 版本才得以推出。与之相比，英伟达早在 2010 年的 CUDA 3.0 版本就已经添加 C++和 Fortran 支持，2011 年的 4.0 版本更通过 GPU Direct 2.0 等技术在 HPC 领域站稳脚跟，2013 年的 GTC13 上，CUDA 添加了 Python 支持，易学易用水平再度提升，到 OpenCL 发布 C++支持的 2015 年，CUDA 已经迭代到 7.0 版本，形成了相对稳定的护城河。可见二者在科学计算领域的统治力完全不可同日而语。根据 AnandTech，OpenCL 始终没有在 HPC 领域获得大规模应用，遑论后续的 AI。

三、硬件不足：早期 MI 产品性能待提高，对数据中心用户吸引力有限

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/585004321311012021>