解析英伟达成长的核心战略：研发为底、生态为径、AI 为翼

行业名称 半导体　　证券研究报告/行业深度报告　　2023 年 05 月 24 日

评级：增持（ 维持 ）

分析师：王芳

执业证书编号：S0740521120002

Email：wangfang02@zts.com.cn

分析师：杨旭

执业证书编号：S0740521120001

Email：yangxu01@zts.com.cn

分析师：李雪峰

执业证书编号：S0740522080004

Email：lixf05@zts.com.cn

基本状况

上市公司数	311
行业总市值(百万元)	4,333,4
行业流通市值(百万元)	2,263,5

行业-市场走势对比

相关报告

【中泰电子】AI 系列 1：从 ChatGPT 看 AI 芯片产业链投资机会

【中泰电子】AI 系列 2：Chiplet：提质增效，助力国产半导体弯道超车

【中泰电子】AI 系列 3：AI 服务器拆解，产业链核心受益梳理

重点公司基本状况

简称	股价	EPS				PE				PEG	评级
	(元)	2020	2021	2022E	2023E	2020	2021	2022E	2023E		
寒武纪	189	-2.0	-3.2	-1.9	-1.2	-95	-59	-101	-163	2.4	NA
工业富联	15	1.0	1.0	1.2	1.3	15	15	13	11	0.7	NA
沪电股份	19	0.6	0.7	0.9	1.1	33	26	21	17	1.0	买入
奥士康	33	1.5	1.0	2.2	2.8	22	35	15	12	0.1	买入
海康威视	36	1.8	1.4	1.8	2.1	20	25	20	17	0.9	NA
大华股份	22	1.0	0.7	1.0	1.2	21	32	21	17	0.4	买入

备注：股价为 2023 年 5 月 22 日收盘价。

报告摘要

- **英伟达是全球领先的 GPU 芯片制造商之一。**公司是全球 GPU 龙头，市场份额遥遥领先。根据 Jon Peddie Research 发布的 GPU 市场数据统计报告，英伟达 2022 年全年 PC GPU 出货量高达 3034 万块，是 AMD 的近 4.5 倍；截至 2022 年四季度，在独立 GPU 市场，英伟达占据 84%的市场份额，远超同业竞争公司。我们认为通过研究英伟达的发展路径和战略，能够帮助国内企业更好地了解 GPU 的应用和未来趋势，为国内企业提供宝贵的借鉴和启示。在本篇研究报告中，我们还原了英伟达所处不同发展阶段的行业背景，深入分析了三个阶段中英伟达通过实施何种战略超越了竞争者，形成了竞争优势。

- **研发为底、生态为径、AI 为翼：研发实力**是一家芯片设计公司的核心竞争力，英伟达从发展初期即重视研发生产力，以高投入换去高回报不断提升产品竞争力。2005 年，AMD 的研发费用为 11 亿美元，是英伟达的 3.2 倍左右，而到了 2022 年，英伟达的研发费用达到了 73.4 亿美元，是 AMD 的 1.47 倍。随着研发投入的不断增长，英伟达通过技术进步降低成本和产品价格，不断推出新的产品吸引更多消费者，优势逐渐凸显；**生态方面，**英伟达推出 CUDA 平台，使得利用 GPU 来训练神经网络等高算力模型的难度大大降低，将 GPU 的应用从 3D 游戏和图像处理拓展到科学计算、大数据处理、机器学习等领域，这一生态系统的建立让很多开发者依赖于 CUDA，进一步增加了英伟达的竞争壁垒；**AI 方面，**人工智能的发展为 GPU 带来更大增长空间，英伟达抓住下游发展新机遇，推出 AI 加速卡，伴随以 ChatGPT 为代表的生成式 AI 大模型发展进入快速增长通道。

- **算力是 AI 芯片底层土壤，未来算力需求将呈爆发式增长。**根据 IDC 数据，未来 5 年我国智能算力规模 CAGR 将达 52.3%。AI 芯片中，GPU 占据主要市场规模。根据 IDC 数据，2022 年国内人工智能芯片市场中，GPU 芯片所占市场份额达 89.0%。GPU 作为市场上 AI 计算最成熟、应用最广泛的通用型芯片，应用潜力较大，其并行计算架构相较于其他 AI 芯片更加适合于复杂数学计算场景，支持高度并行的工作负载。

- **国产厂商加速布局，看好 AI 发展推动国产替代进程提速。**在 ChatGPT 等概念影响下，AIGC 关注度火热。未来 AI 应用的落地离不开庞大算力的支撑，也将推动算力产业链快速增长。据 IDC，2021 年中国 AI 投资规模超 100 亿美元，2026 年将有望达到 267 亿美元，全球占比约 8.9%，排名第二，其中 AI 底层硬件市场占比将超过 AI 总投资规模的半数。看好国产 AI 供应商在产业创新趋势以及国产替代背景下进入快速增长通道。

- **建议关注：（1）AI 算力芯片：**寒武纪、海光信息、景嘉微；**（2）服务器产业链：**工业富联、沪电股份、奥士康；**（3）AI 应用：**大华股份、海康威视；**（4）Chiplet：**通富微电、长电科技、华海清科、长川科技、兴森科技。**（5）C 端 AI 应用：**国光电器、漫步者；瑞芯微、晶晨股份、乐鑫科技、中科蓝讯。

- **风险提示事件：**需求不及预期、产能瓶颈的束缚、大陆厂商技术进步不及预期、中美贸易摩擦加剧、研报使用的信息更新不及时。

内容目录

1、英伟达：算力芯片巨头领跑 AI 时代 - 6 -
1.1 公司简介：全球领先的 GPU 龙头厂商 - 6 -
1.2 公司产品：多元化产品矩阵助力公司长期增长 - 9 -
1.3 公司财务：财务状况良好，反哺研发投入上升 - 14 -
2、英伟达发展历程三部曲 - 17 -
2.1 1993-2000：初具规模，提升研发效率战胜对手 - 17 -
2.2 2001-2006：寡头垄断，逐步成为独显市场霸主 - 19 -
2.3 2007-2023：重“芯”开始，引领人工智能计算 - 20 -
3、英伟达发展历程总结，借鉴意义 - 23 -
3.1 深耕 GPU 算力领域，研发为导向不断提升产品竞争力 - 23 -
3.2 CUDA 自成体系：从单一产业到生态链，构建强护城河 - 26 -
3.3 抓住人工智能发展浪潮，顺利转型切入算力芯片领域 - 32 -
4、算力是 AI 底层土壤，从英伟达看国产发展机遇 - 36 -
4.1 ChatGPT 激起 AI 浪潮，大模型升级推动算力提升 - 36 -
4.2 算力芯片快速增长，GPU 占据 AI 芯片主流地位 - 40 -
4.3 AI 芯片领域，国产芯片迅速崛起 - 43 -
4.4 国产算力公司梳理 - 45 -
5、投资建议及风险提示 - 51 -

图表目录

图 1：引领 GPU 市场的巨头，英伟达的崛起和发展历程 - 6 -
图 2：英伟达时间线 - 7 -
图 3：“CPU+GPU+DPU”业务布局 - 7 -
图 4：英伟达三芯布局产品线 - 8 -
图 5：架构发展历程 - 8 -
图 6：产品及相应架构 - 9 -
图 7：英伟达产品线总览 - 10 -
图 8：英伟达游戏显卡重要时间点 - 10 -
图 9：英伟达游戏显卡详细参数 - 11 -
图 10：英伟达数据中心 GPU 发展历程 - 11 -
图 11：英伟达数据中心 GPU 及其参数 - 12 -
图 12：英伟达 Grace 与 x86+Hopper 对比 - 12 -
图 13：英伟达自动驾驶芯片时间轴 - 13 -
图 14：英伟达几代汽车芯片对比 - 13 -
图 15：英伟达专业可视化产品重要时间点 - 14 -

图 16：英伟达专业可视化产品及其参数 - 14 -
图表 17：英伟达历年营业收入及 yoy（百万美元） - 15 -
图表 18：英伟达历年净利润及 yoy（百万美元） - 15 -
图表 19：英伟达 2003 财年分地区收入占比 - 15 -
图表 20：英伟达 2023 财年分地区收入占比 - 15 -
图表 21：英伟达 2019-2023 财年分业务收入增速 - 16 -
图表 22：英伟达 2019-2023 财年分业务收入（百万美元） - 16 -
图表 23：英伟达 2000-2023 财年净资产收益率 - 16 -
图表 24：英伟达 2000-2023 财年净利率和毛利率 - 16 -
图表 25：2000-2005 年英伟达、ATI 研发费用 - 17 -
图表 26：2005-2022 年英伟达、AMD 研发费用 - 17 -
图表 27：1999-2023 财年英伟达研发费用率 - 17 -
图表 28：2001-2023 财年英伟达研发人员数量 - 17 -
图表 29 ：3D 图像市场竞争图 - 18 -
图表 30 ：Riva 128 与 i740 对比 - 19 -
图表 31：2002-2013 年英伟达收购汇总 - 20 -
图表 32：英伟达产品线 - 20 -
图表 33：2014 年 3G/4G 市场份额 - 21 -
图表 34：英伟达 2007-2023 财年净利润 - 22 -
图表 35：英伟达 2007-2023 财年研发费用 - 22 -
（百万美元） - 22 -
图表 36：英伟达终端用户收入情况（百万美元） - 22 -
图表 37：英伟达终端用户收入年增长率 - 22 -
图表 38：1998 年操作系统占比 - 23 -
图表 39：英伟达追随 DirectX 升级开发产品 - 23 -
图表 40：英伟达商业布局 - 24 -
图表 41：2016-2025 年自动驾驶规模（十亿美元） 错误!未定义书签。
图表 42：英伟达显卡合作伙伴多于 AMD - 25 -
图表 43：AMD 的研发费用被英伟达反超 - 25 -
图表 44：英伟达在天梯图覆盖面广，高端产品领先 AMD - 26 -
图表 45：CUDA 加速计算解决方案 - 27 -
图表 46：CUDA 软件层 - 27 -
图表 47：CUDA 11.0 主要特点 - 28 -
图表 48：大学教授 CUDA 数量（所） - 29 -
图表 49：CUDA 成为英伟达生态基础 - 29 -

图表 50：GPU 编程平台发展历史 - 30 -
图表 51：英伟达人工智能生态系统 - 31 -
图表 52：CUDA 对应 GPU 架构发展 - 31 -
图表 53：CUDA 通过并行架构的改进 - 31 -
图表 54：全球数据总量（ZB） - 32 -
图表 55：Intel 测算的数据流 - 32 -
图表 56：全球 ADAS 市场规模扩大 - 33 -
图表 57：全球自动驾驶功能市场规模扩大 - 33 -
图表 58：智能驾驶层级越高所需传感器越多 - 33 -
图表 59：2018-2025 年 AI 硬件市场收入（十亿美元） - 34 -
图表 60：GPU 打破摩尔定律 - 34 -
图表 61：AI 芯片产业链 - 35 -
图表 62：P4 传输速度大于 FPGA 架构芯片 - 35 -
图表 63：FPGA 和 GPU 对比 - 36 -
图表 64：AI 人工智能与半导体计算芯片发展历程 - 37 -
图表 65：GPT 模型迭代过程 - 38 -
图表 66：大语言模型（LLM）举例 - 39 -
图表 67：Transformer 架构示意图 - 40 -
图表 68：国外部分 AIGC 预训练模型一览 - 40 -
图表 69：中国 AI 算力规模及预测 - 41 -
图表 70：全球 AI 芯片市场规模及预测 - 41 -
图表 71：AI 芯片特点及具体参数对比 - 42 -
图表 72：CPU 的基本结构 - 42 -
图表 73：GPU 的基本结构 - 42 -
图表 74：CPU 与 GPU 对比 - 43 -
图表 75：国内外 AI 芯片产品对比（1）——图形渲染 GPU - 44 -
图表 76：国内外 AI 芯片产品对比（2）——GPGPU - 44 -
图表 77：国内外 AI 芯片产品对比（3）——FPGA/ASIC - 45 -
图表 78：龙架构 LoongArch - 46 -
图表 79：海光 CPU 与 Intel 产品性能对比 - 47 -
图表 80：深算一号与国际同类型产品性能对比 - 47 -
图表 81：高性能通用图形处理器芯片及系统研发项目情况及进程安排 - 48 -
图表 82：公司发展历程时间表 - 48 -
图表 83：FPGA 芯片产品线 - 49 -
图 84：公司发展历程 - 50 -

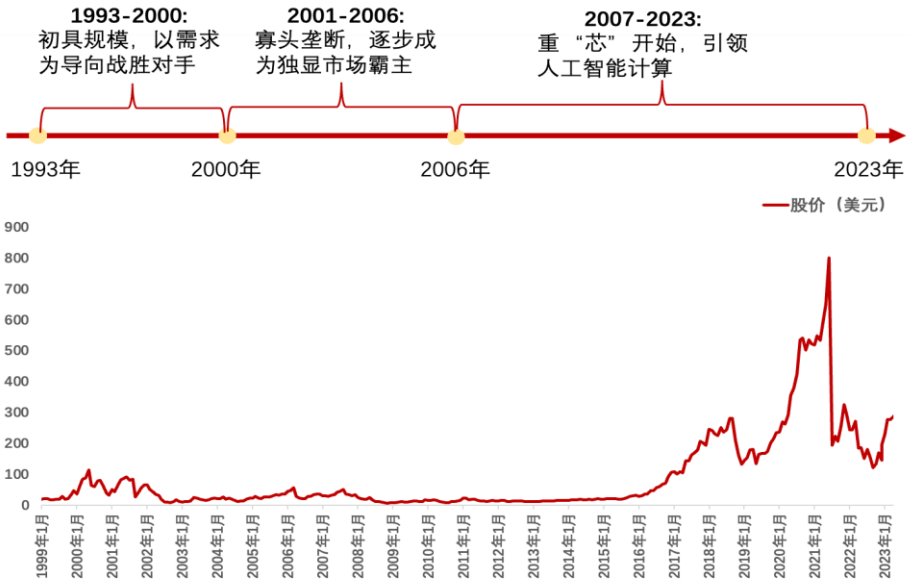
图 85：公司 FPGA 芯片产品线..... - 50 -

1、英伟达：算力芯片巨头领跑 AI 时代

1.1 公司简介：全球领先的 GPU 龙头厂商

- **公司是全球 GPU 龙头，市场份额遥遥领先。**英伟达（NVIDIA）是一家全球知名的技术公司，成立于 1993 年，最初以图形处理器（GPU）起家，通过不断的创新和发展，逐渐成为了高性能计算领域的领导者。根据 Jon Peddie Research 发布的 GPU 市场数据统计报告，英伟达 2022 年全年 PC GPU 出货量高达 3034 万块，是 AMD 的近 4.5 倍；截至 2022 年四季度，在独立 GPU 市场，英伟达占据 84%的市场份额，远超同业竞争公司。

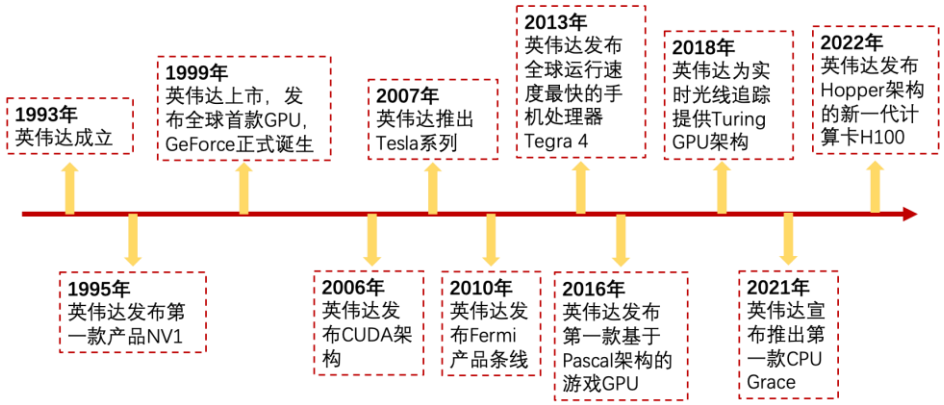
图 1：引领 GPU 市场的巨头，英伟达的崛起和发展历程



来源：WRDS，中泰证券研究所整理

- **英伟达的产品创新和迭代从未止步。**自英伟达成立以来，其经历了多个重要的发展时间点。其中包括 1999 年推出全球第一款 GPU、2006 年发布 Fermi 架构、2012 年发布 Kepler 架构、2016 年推出 AI 加速器 Tesla P100 和 Volta 架构、以及 2020 年发布 Ampere 架构等。这些重要时间点的创新和进步，为英伟达在高性能计算、人工智能、虚拟现实等领域的发展奠定了坚实的基础。

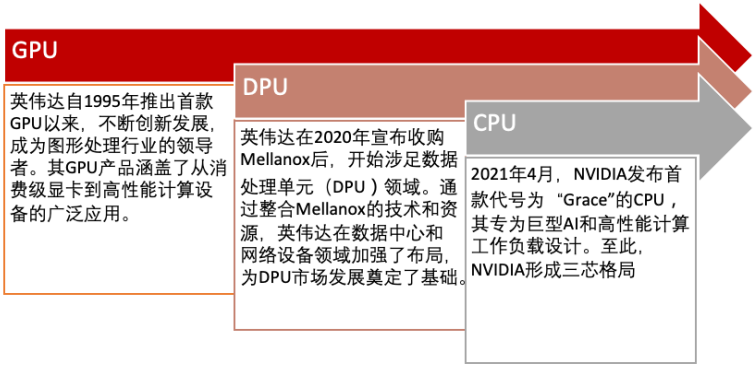
图 2：英伟达时间线



来源：英伟达官网，中泰证券研究所整理

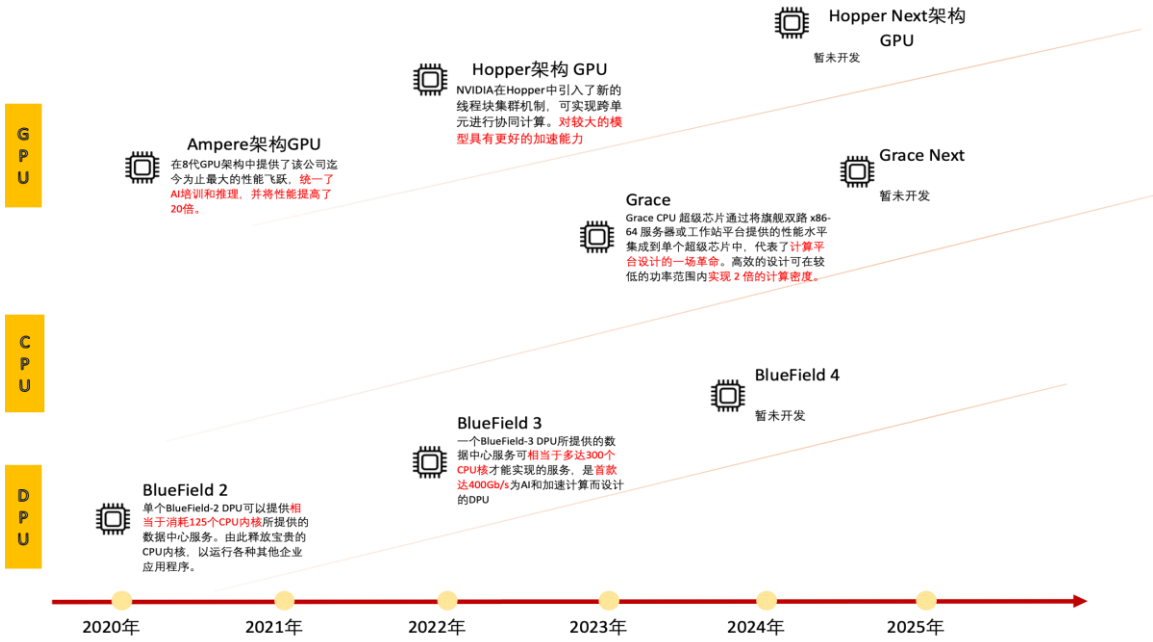
- **横向拓展丰富业务产品线，实现“CPU+GPU+DPU”三芯布局。**英伟达的三芯战略侧重于在数据中心市场实现 CPU、GPU 和 DPU 三类硬件的布局，旨在全面提升竞争力，满足云计算、人工智能及机器学习等高端应用领域的需求。CPU 的加入使英伟达能够更好地应对各种计算任务，尤其是那些需要快速逻辑判断和高度并行处理能力的应用。而 DPU 则针对数据中心和网络设备的需求，具有高效处理数据包和协议的能力，为英伟达的产品线增添了新的价值。通过将 CPU、GPU 和 DPU 集成到同一平台上，英伟达可以为客户提供更加全面、高效的计算解决方案。目前 CPU+GPU 的产品组合获得超级计算中心的采用并即将广泛部署于大型服务器，三芯战略初显成效。

图 3：“CPU+GPU+DPU”业务布局



来源：英伟达官网，中泰证券研究所整理

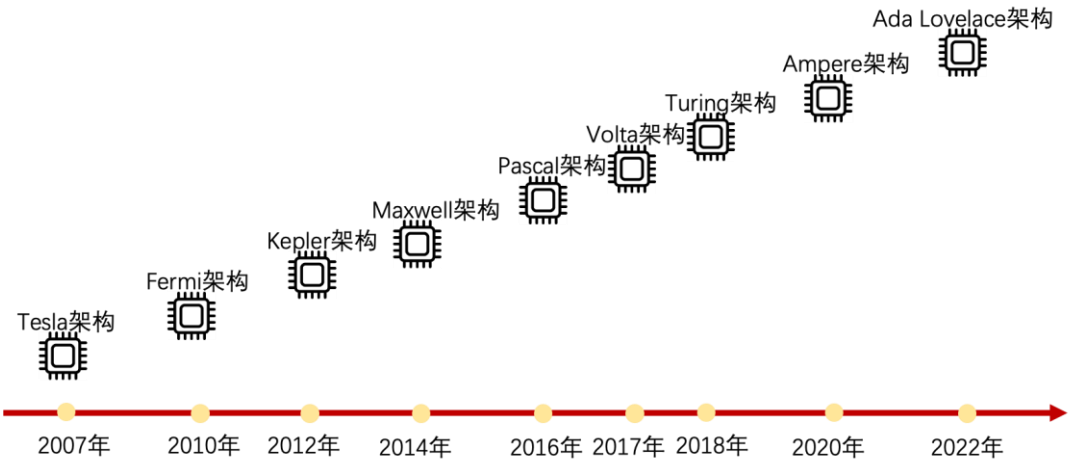
图 4：英伟达三芯布局产品线



来源：英伟达官网，中泰证券研究所整理

- **芯片架构是英伟达的技术核心，快速迭代的新架构为产品带来不断的创新与升级。**自英伟达 GPU 问世以来，其架构经历了多个重要发展阶段。2006 年，Fermi 架构在 GPU 计算领域实现了重大突破，Kepler 架构进一步提高了能效比和 GPU 性能，并引入了动态并行处理技术。随后，Maxwell 架构实现了更加节能和高效的设计，Pascal 架构则引入了深度学习计算中的 Tensor Core 和 NVLink 技术，以及更多的 AI 加速功能。Volta 架构则实现了更高的计算能力和存储带宽，并引入了深度学习加速器 Tensor Cores V100。Turing 架构则进一步提高了光线追踪和图形渲染性能，而 Ampere 架构则在 AI 加速、性能和能效方面实现了重要进展。每一代架构的创新和进步，都为 GPU 技术在高性能计算、人工智能、虚拟现实等领域的应用奠定了坚实的基础。

图 5：架构发展历程



来源：英伟达官网，中泰证券研究所整理

图 6：产品及相应架构

架构名称	系列产品
Tesla	GeForce 8800 GTX/Ultra, Quadro FX 5600/5800
Fermi	GeForce 400/500系列, Tesla M2050/M2070/M2090
Kepler	GeForce 600/700系列, Quadro K/M系列
Maxwell	GeForce 900系列, Quadro M系列
Pascal	GeForce 10系列, Quadro P系列
Volta	Titan V, Quadro GV100
Turing	GeForce 16/20系列, Quadro RTX系列
Ampere	GeForce 30系列, A100, A40, A30, A10
Lovelace	GeForce RTX40系列

来源：英伟达官网，中泰证券研究所整理

1.2 公司产品：多元化产品矩阵助力公司长期增长

- **英伟达产业布局多元化，解决客户不同需求。**GPU 产品为英伟达主要收入来源，收入占比稳定在 80%以上。相比较于 CPU，GPU 在机器学习算法有天生的优势。英伟达一直专注于 GPU 的设计，同时由于 GPU 的并行计算能力，可以通过数千个计算核心进行深度学习，英伟达开始将服务和系统、软硬件和可编程算法结合在一起，提出 CUDA 架构。从下游应用来看，英伟达产品主要集中于游戏、专业可视化、数据中心以及自动驾驶领域：

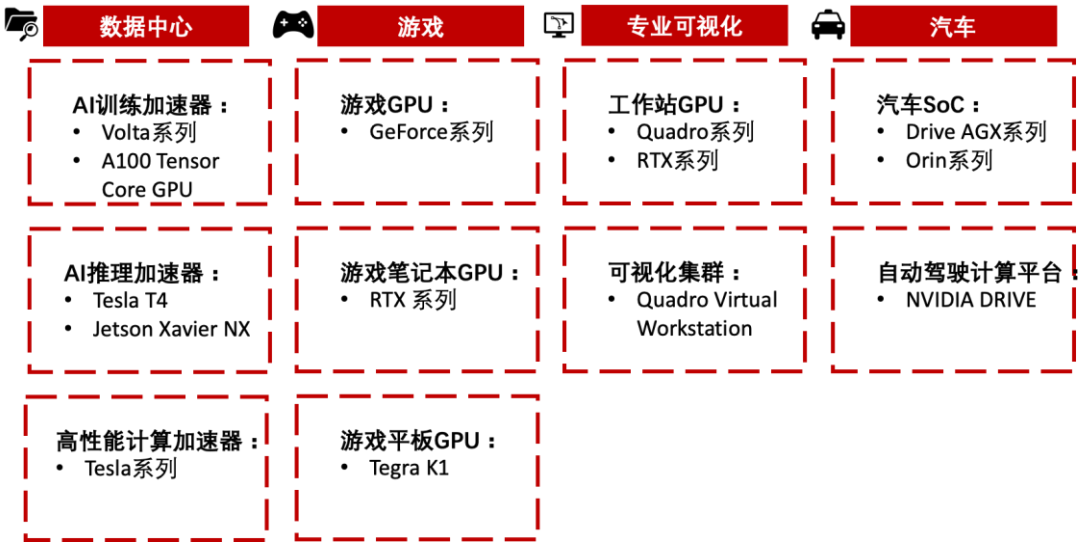
  1）游戏市场：英伟达提供的产品包括 PC 游戏的 GeForce RTX 和 GeForce GTX，用于游戏和流媒体的 SHIELD 设备，用于云端游戏的 GeForce NOW，以及用于专门控制台游戏设备的平台和开发服务；

  2）专业可视化市场：英伟达除了加速 GPU 计算解决方案，同时也为汽车、娱乐、建筑工程、石油和天然气、医疗等行业引入新的解决方案；

  3）数据中心市场：英伟达使用 NVlink 技术将多个 GPU 结合在一起，加速神经网络训练和推理。同时开发出 DGX 超级计算机，进行科学计算、深度学习和机器学习；

  4）自动驾驶市场：英伟达 Drive 作为一个人工智能汽车平台，涵盖了从交通拥堵到机器人出租车自动驾驶的所有领域。2018 年有超过 370 家自动驾驶汽车公司开始使用 Drive，共同开发自动驾驶的人工智能系统。

图 7：英伟达产品线总览



来源：英伟达官网，中泰证券研究所整理

■ **游戏业务：是英伟达主要产品线，作为基本盘见证了其里程碑式的革新。**英伟达在游戏业务领域持续不断的技术升级，以应对玩家日益增长的画质需求。游戏业务一直是英伟达的核心领域，每年都以引人注目的新产品展现其持续的创新力。与前一代产品相比，每一代新显卡都带来了显著的性能提升。从核心数量来看，英伟达显卡产品的 CUDA 核心数量已从最初的 640 颗增长到现在的高达 16384 颗，技术上不断突破，包括实时光线追踪技术等。另外，英伟达在游戏显卡市场上有着广泛的布局，从入门级到专业级，都提供了相应的产品。这一策略允许英伟达满足从独立游戏玩家到专业电竞选手的多元需求。

图 8：英伟达游戏显卡重要时间点

来源：英伟达官网，中泰证券研究所整理

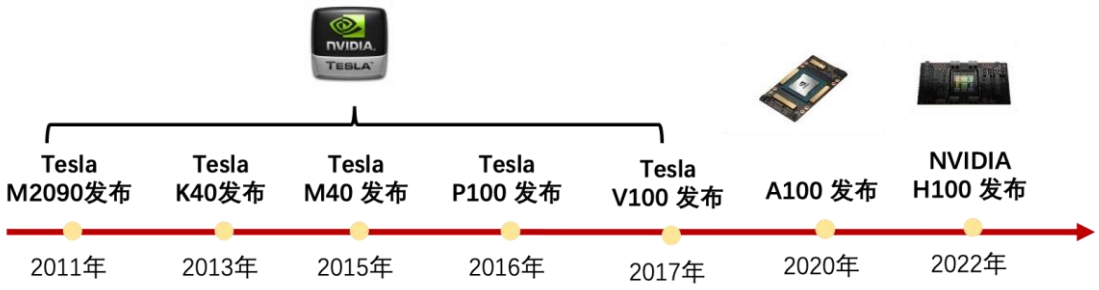
图 9：英伟达游戏显卡详细参数

系列名称	产品型号	发布时间	CUDA 核心数量	加速频率 (GHz)	基础频率 (GHz)	显存容量	制程工艺	功耗（W）
GeForce RTX 40系列	RTX 4090	2022	16384	2.52	2.23	24GB	6nm	450
	RTX 4080	2022	9728	2.51	2.21	16GB	6nm	320
	RTX 4070 Ti	2023	7680	2.61	2.31	12GB	6nm	285
	RTX 4070	2023	5888	2.48	1.92	12GB	6nm	200
GeForce RTX 30系列	GeForce RTX 3090 Ti	2022	10752	1.89	1.56	24GB	8nm	850
	GeForce RTX 3090	2020	10496	1.7	1.4	24GB	8nm	350
	GeForce RTX 3080 Ti	2021	10240	1.67	1.37	12GB	8nm	350
	GeForce RTX 3080	2020	8960/9704	1.71	1.26/1.44	12GB/10GB	8nm	320
	GeForce RTX 3070 Ti	2021	6144	1.77	1.58	8GB	8nm	250
	GeForce RTX 3070	2020	5888	1.73	1.5	8GB	8nm	220
	GeForce RTX 3060 Ti	2020	4864	1.67	1.41	9GB	8nm	200
	GeForce RTX 3060	2021	3584	1.78	1.32	10GB	8nm	170
	GeForce RTX 3050	2022	2560/2304	1.78/1.76	1.55/1.51	11GB	8nm	130
GeForce RTX 20系列	GeForce RTX 2080 Ti	2018	4352	1.64	1.35	11GB	12nm	260
	GeForce RTX 2080 Super	2019	3072	1.82	1.65	8GB	12nm	250
	GeForce RTX 2080	2018	2944	1.8	1.52	8GB	12nm	225
	GeForce RTX 2070 Super	2019	2560	1.77	1.61	8GB	12nm	215
	GeForce RTX 2070	2018	2304	1.71	1.41	8GB	12nm	185
	GeForce RTX 2060 Super	2019	2176	1.65	1.47	8GB	12nm	175
	GeForce RTX 2060	2019	2176/1920	1.65/1.68	1.47/1.37	12GB/6GB	12nm	185/160
GeForce GTX 16系列	GeForce GTX 1660 Ti	2019	1536	1.77	1.5	6GB	12nm	120
	GeForce GTX 1660 Super	2019	1408	1.785	1.53	6GB	12nm	125
	GeForce GTX 1660	2019	1408	1.785	1.53	6GB	12nm	120
	GeForce GTX 1650 Super	2019	1280	1.725	1.53	4GB	12nm	100
	GeForce GTX 1650（G5）	2019	896	1.665	1.485	4GB	12nm	75
	GeForce GTX 1650（G6）	2019	896	1.59	1.41	4GB	12nm	75
	GeForce GTX 1630	2022	512	1.785	1.74	4GB	12nm	75
GeForce GTX 10系列	GeForce GTX 1080 Ti	2017	3584	1.58	1.48	11 GB	16 nm	250
	GeForce GTX 1080	2016	2560	1.73	1.61	8 GB	16 nm	180
	GeForce GTX 1070 Ti	2017	2432	1.68	1.61	8 GB	16 nm	180
	GeForce GTX 1070	2016	1920	1.68	1.51	8 GB	16 nm	150
	GeForce GTX 1060	2016	1280	1.71	1.51	6 GB	16 nm	120
	GeForce GTX 1050 Ti	2016	768	1.39	1.29	4 GB	14 nm	75
	GeForce GTX 1050	2016	640	1.46	1.35	2GB	14 nm	75

来源：英伟达官网，中泰证券研究所整理

- **数据中心：持续发力，高市占率源自于英伟达持续不断的研发与创新。**英伟达长期占据高端 GPU 市场的领导地位，截至目前英伟达占据全球算力芯片 90%的市场份额。高端芯片领域的霸主地位主要源自于公司不断的技术提升所形成强大的技术壁垒。从 2017 到 2022 这五年间，公司先后推出了 Volta、Ampere、Hopper 等针对高性能计算和 AI 训练的架构，以此为基础发布了 V100、A100、H100 等高端 GPU。通过不断的技术革新，英伟达 GPU 产品向量双精度浮点算力已从 7.8 TFLOPS 增至 30 TFLOPS。

图 10：英伟达数据中心 GPU 发展历程



来源：英伟达官网，中泰证券研究所整理

- **英伟达数据中心 GPU 在 11 年间从制程工艺到核心数量，各参数全方位提升。**从 2011 年的 Tesla M2090 开始英伟达不断更新迭代数据中心产品，到

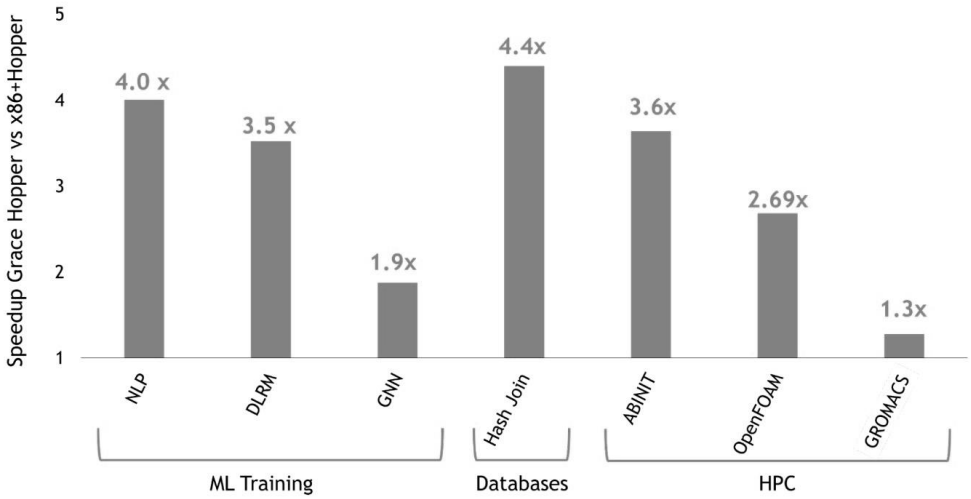
了 2022 年发布的英伟达全新 GPU 产品 NVIDIA H100，性能上已经出现了质的飞跃。此外，英伟达在数据中心的布局不仅仅停留在 GPU，在 CPU 方面英伟达也全面发力，在 2022 年发布了首款 CPU 产品 Grace。Grace 内置下一代 Arm Neoverse 内核，采用第四代 NVIDIA NVLink，从 CPU 到 GPU 连接速度超过 900GB/s，相当于目前服务器 14 倍的带宽速度；从 CPU 到 CPU 的速度超过 600GB/s。并且 Grace 拥有最高的内存带宽，采用的新内存 LPDDR5x 技术，带宽是 LPDDR4 的 2 倍，能源效率提高了 10 倍，能提供更多计算能力。

图 11：英伟达数据中心 GPU 及其参数

系列名称	产品型号	发布时间	CUDA核心数量	加速频率(GHz)	基础频率(GHz)	显存容量	显存位宽（位）	制程工艺	接口	功耗(W)
Tesla	M2090	2011	512	1.3	-	6GB GDDR5	384	40nm	PCIe 2.0	250
Tesla	K40	2013	2880	0.875	-	12GB GDDR5	384	28nm	PCIe 3.0	235
Tesla	M40	2015	3072	1.1	-	12GB/24GB GDDR5	384	28nm	PCIe 3.0	250
Tesla	P100	2016	3584	1.48	-	16GB/12GB HBM2	4096	16nm	PCIe 3.0, NVLink	300
Tesla	V100	2017	5120	1.53	-	16GB/32GB HBM2	4096	12nm	PCIe 3.0, NVLink	250-300
A100	-	2020	6912	1.41	-	40GB/80GB HBM2	5120	7nm	PCIe 4.0, NVLink	400
NVIDIA H100	NVIDIA H101	2022	7296	1.83	-	80GB GDDR6	5120	4nm	-	700

来源：英伟达官网，中泰证券研究所整理

图 12：英伟达 Grace 与 x86+Hopper 对比



来源：英伟达官网，中泰证券研究所整理

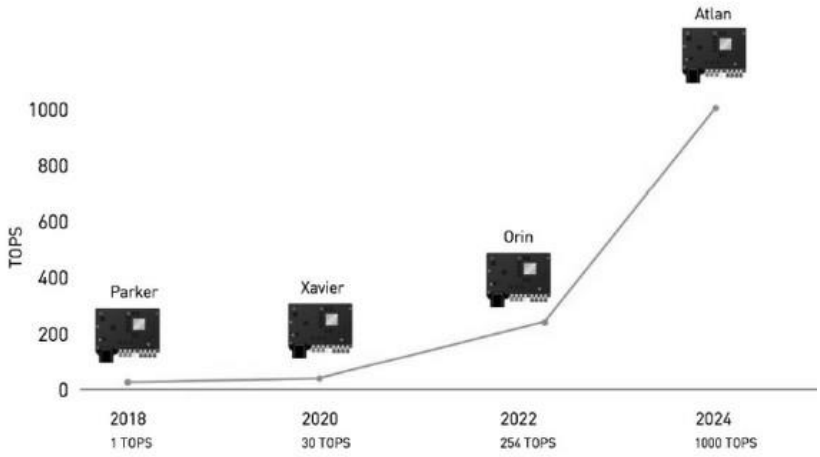
- **自动驾驶业务：为英伟达提供中长期增长曲线。**英伟达的自动驾驶 SoC 产品线以其高性能、高能效和创新技术而著称，致力于满足不断增长的计算需求。英伟达推出的自动驾驶 SoC 产品包括先进的 Atlan 和 Orin 芯片，它们集成了安培架构 GPU 核心、基于 Arm 的 Grace CPU 核心、深度学习和计算机视觉加速器单元以及 BlueField DPU 核心，以实现卓越的算力和性能。英伟达的 SoC 产品线不断创新，为客户提供卓越的性能和可靠性，帮助推动未来智能驾驶和高度互联的汽车发展。

图 13：英伟达自动驾驶芯片时间轴

来源：英伟达官网，中泰证券研究所整理

- **最新款 Atlan SoC 算力获得指数级提升，为自动驾驶提供充足算力。**2021 年，英伟达推出了自动驾驶 SoC Atlan，其单颗算力高达 1000TOPS，是上一代 Orin SoC（254TOPS）的近四倍。Altan 还支持 400Gbs（40 万兆）网络和安全网关，可以满足高速通信需求。同时，Atlan 可与为上一代芯片组编写的软件堆栈（如 Orin 或 Xavier）兼容，使得汽车制造商和 AV 开发人员不需要重新设计软件就能利用新 SoC 的性能提升，大大提升使用的便捷程度。

图 14：英伟达几代汽车芯片对比



来源：英伟达官网，中泰证券研究所整理

- **可视化业务：技术革新助力卓越视觉与计算体验。**在过去几年，英伟达专业可视化业务持续推出了一系列的技术革新，包括新的 GPU 架构（如 Pascal、Volta、Ampere、Ada Lovelace），更高效的显存技术（如 GDDR6X），以及更加智能化的软件工具（如 RTX Studio）。这些创新大幅提升了英伟达专业显卡在高性能计算、人工智能、虚拟现实等领域的性能和可靠性，为专业用户提供了更加卓越的视觉体验和计算能力。

图 15：英伟达专业可视化产品重要时间点

来源：英伟达官网，中泰证券研究所整理

- **英伟达专业显卡技术不断进步，性能显著提升。**随着英伟达专业可视化显卡的不断升级，计算能力和相关性能得到了显著提升。从最初的几百万个 CUDA 核心、数百 GB/s 的显存带宽，到现在的数千万个 CUDA 核心、TB/s 级别的显存带宽，英伟达专业显卡已经成为高性能计算、人工智能、虚拟现实等领域不可或缺的重要组成部分，为专业用户提供了更加卓越的视觉体验和计算能力。

图 16：英伟达专业可视化产品及其参数

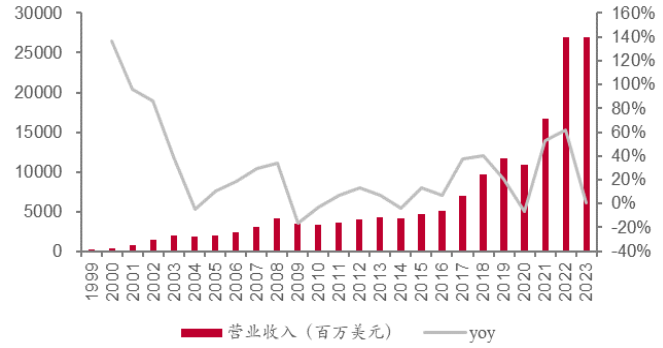
系列名称	产品型号	发布时间	CUDA核心数量	加速频率(GHz)	基础频率(GHz)	显存容量	显存位宽（位）	制程工艺	接口	功耗(W)
Quadro RTX	8000	2018	4608	1.77	1.35	48GB GDDR6	384	12nm	PCIe 3.0 x16	295
	6000	2018	4608	1.77	1.35	24GB GDDR6	384	12nm	PCIe 3.0 x16	295
	5000	2018	3072	1.77	1.65	16GB GDDR6	256	12nm	PCIe 3.0 x16	230
	4000	2018	2304	1.62	1.1	8GB GDDR6	256	12nm	PCIe 3.0 x16	160
Quadro P Series	P6000	2016	3840	1.8	1.48	24GB GDDR5X	384	16nm	PCIe 3.0 x16	250
	P5000	2016	2560	1.8	1.6	16GB GDDR5X	256	16nm	PCIe 3.0 x16	180
	P4000	2017	1792	1.77	1.47	8GB GDDR5	256	16nm	PCIe 3.0 x16	105
	P2000	2017	1024	1.48	1.17	5GB GDDR5	160	16nm	PCIe 3.0 x16	75
	P1000	2017	640	1.85	1.4	4GB GDDR5	128	16nm	PCIe 3.0 x16	47
	P620	2018	512	1.8	1.14	2GB GDDR5	128	14nm	PCIe 3.0 x16	40
	P400	2017	256	0.98	0.64	2GB GDDR5	64	16nm	PCIe 3.0 x16	30
RTX A6000	RTX A6000	2020	10752	1.8	1.41	48GB GDDR6	384	8nm	PCIe 4.0 x16	300
RTX A40	RTX A40	2020	10752	1.74	1.3	48GB GDDR6	384	8nm	PCIe 4.0 x16	300
RTX 6000	RTX 6000	2022	18176	1.77	1.44	48GB GDDR6	384	7nm	PCIe 3.0 x16	300

来源：英伟达官网，中泰证券研究所整理

1.3 公司财务：财务状况良好，反哺研发投入上升

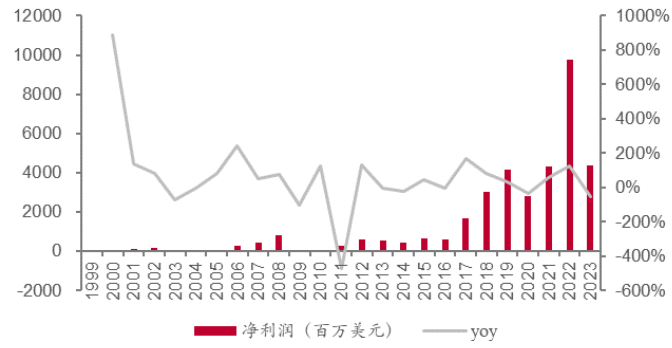
- **公司营业收入高速增长，1999-2023 财年 CAGR 24%。**1996 年英伟达的营业收入仅 391 万美元，净利润亏损超过 300 万美元。此后，英伟达的体量快速增长，到 2023 财年营收和净利润分别达 270 亿美元和 44 亿美元，1999-2023 财年营收 CAGR 24%，净利润 CAGR 34%。

图表 17：英伟达历年营业收入及 yoy（百万美元）



来源：英伟达，中泰证券研究所

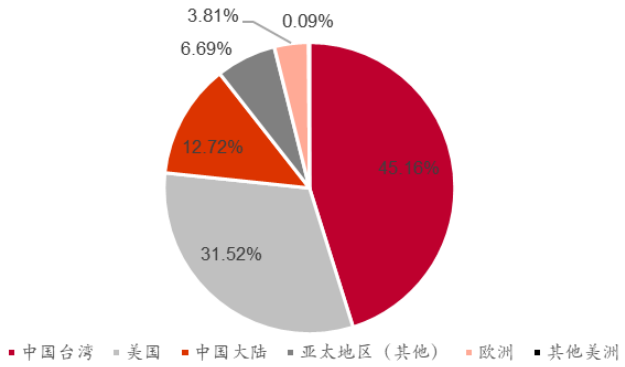
图表 18：英伟达历年净利润及 yoy（百万美元）



来源：英伟达，中泰证券研究所

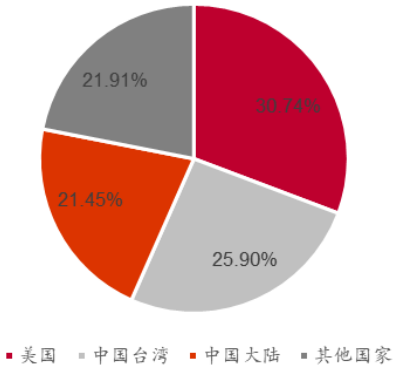
■ **新兴市场成为英伟达主要收入来源地。**分地区看，中国大陆在 2023 财年营业收入达到 58 亿美元，占总收入的 21%，而在 2003 财年中国大陆营业收入只有 2.4 亿美元，占总收入的比例仅为 13%。和中国大陆市场一样，亚太其他地区以及美洲其他地区都出现了较大的增长幅度。相反，中国台湾市场出现了较大的衰退。2003 财年，中国台湾市场占总营收的比例为 45%，到 2023 财年下降到 32%，而美国市场收入则保持稳定在 31%左右。

图表 19：英伟达 2003 财年分地区收入占比



来源：英伟达，中泰证券研究所

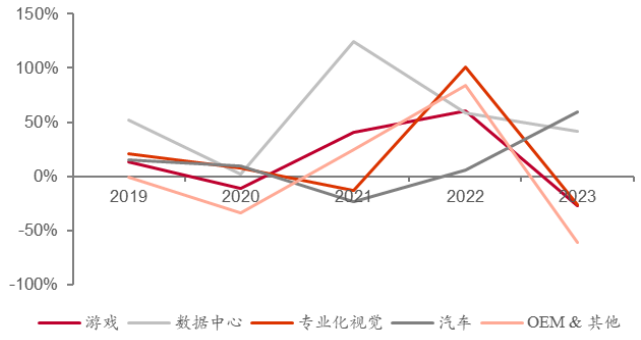
图表 20：英伟达 2023 财年分地区收入占比



来源：英伟达，中泰证券研究所

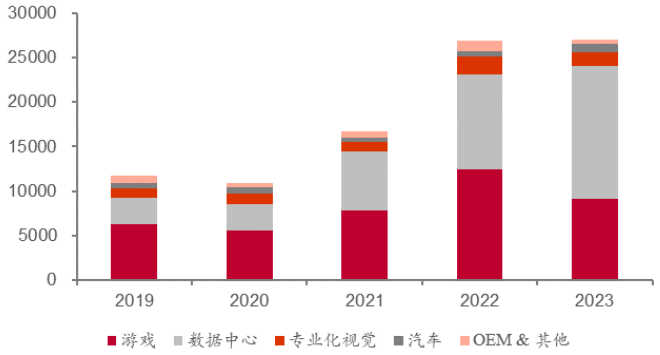
■ **随人工智能发展，数据中心业务收入增速最高，逐步成为公司最大营收占比。**从业务板块看，英伟达下游应用包括游戏、数据中心、专业化视觉、汽车、OEM 及其他。其中，数据中心业务收入在 2023 财年达到 150 亿美元，占据英伟达营业总收入的 56%，数据中心业务收入同比增长 41%，主要增长来源于 AI 发展及美国云服务提供商的推动。游戏业务收入 90.7 亿美元，占总营业收入的 34%，受全球游戏行业需求下行影响同比下降 27%。汽车、代工以及专业可视化业务都保持着低速增长。

图表 21：英伟达 2019-2023 财年分业务收入增速



来源：英伟达，中泰证券研究所

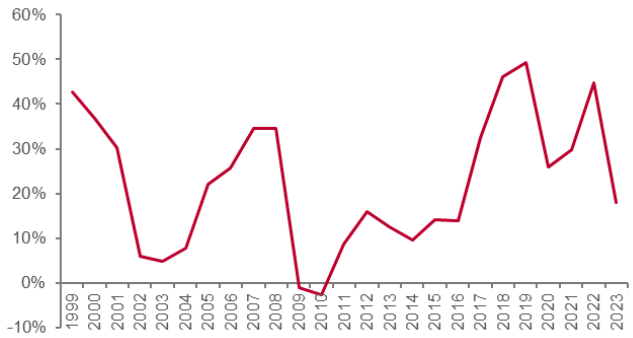
图表 22：英伟达 2019-2023 财年分业务收入（百万美元）



来源：英伟达，中泰证券研究所

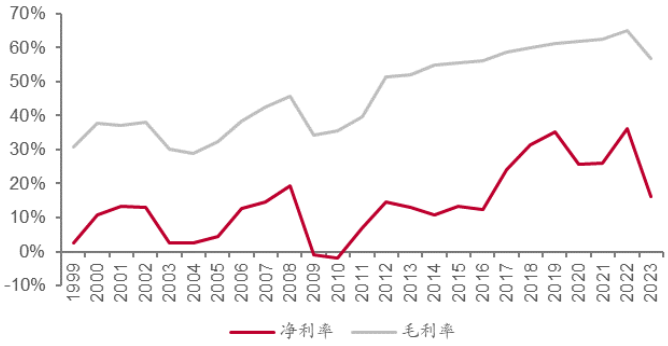
■ **英伟达净资产收益率周期波动，毛利率和净利率总体呈现上升趋势。**2000 财年之后，英伟达开始负责 Xbox（微软公司开发并于 2001 年发售的一款家用电视游戏机）的芯片设计工作，因为 Xbox 相比较于英伟达其他产品有着相对较少的利润率，所以 ROE 和净利率都呈现下降趋势。在此之后英伟达依靠新产品的开发，使得 ROE 重新上升到 34%。2008 财年英伟达已成长为全球图像处理器行业龙头，但因全球经济危机影响，公司 ROE 和净利率创新低，至 2010 财年分别达-2.69%和-2.04%。2010 年后，全球经济复苏，游戏市场在新兴市场蓬勃发展，英伟达游戏部门业务及图形处理器收入平稳上升。2017 财年，英伟达迎来了新一轮的增长期，产品全面发力，GeForce、Tesla、GRID 和 Quadro 销售收入相较于 2016 年都出现大幅度增长。

图表 23：英伟达 2000-2023 财年净资产收益率



来源：英伟达，iFind，中泰证券研究所

图表 24：英伟达 2000-2023 财年净利率和毛利率

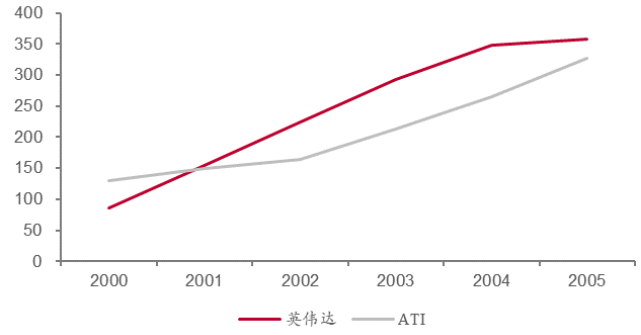


来源：英伟达，iFind，中泰证券研究所

■ **良好的营收状况是公司增加研发投入的基本，研发投入也保障了公司营收的持续健康成长。**相比较于竞争对手 ATI 和 AMD，英伟达在竞争初期都处于下风。随着研发投入的不断增长，英伟达通过技术进步降低成本和产品价格，不断推出新的产品吸引更多消费者，优势逐渐凸显。在与 ATI 竞争的周期中，英伟达的研发费用从 1999 财年的 2507 万美元，以年均 55%的增长率赶上 ATI 的研发费用，在 2005 财年达到 3.6 亿美元。ATI 被 AMD 收购后，英伟达在独立显卡的竞争对手就变为了 AMD。2005 年，AMD 的研发费用为 11 亿美元，是英伟达的 3.2 倍左右，而到了 2022

年，英伟达的研发费用达到了 73.4 亿美元（对应 2023 财年），是 AMD 的 1.47 倍。

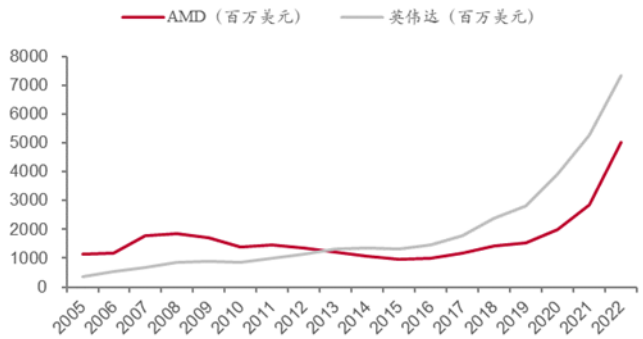
图表 25：2000-2005 年英伟达、ATI 研发费用（百万美元）



来源：英伟达，ATI，中泰证券研究所

注释：各家公司取统一自然年数据对比。

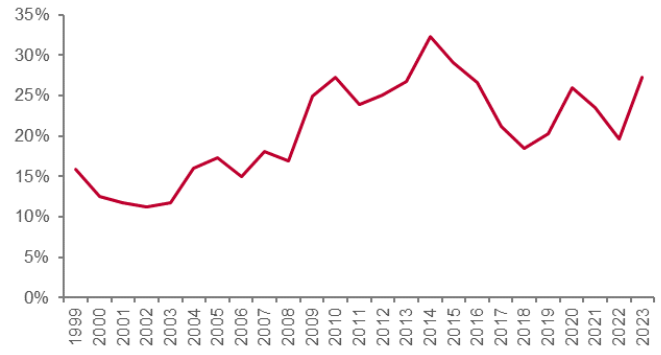
图表 26：2005-2022 年英伟达、AMD 研发费用（百万美元）



来源：英伟达，AMD，中泰证券研究所

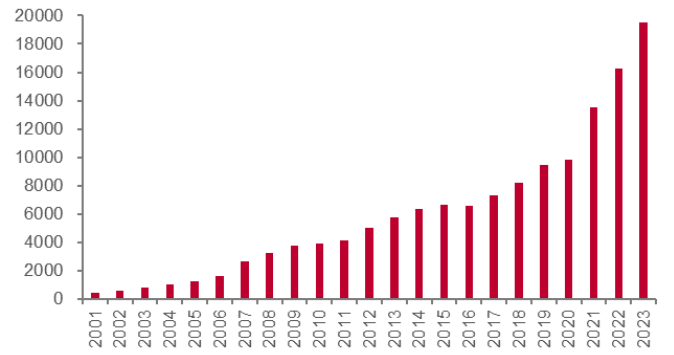
■ **研发费用率保持高位，不断吸引优秀人才加入。**从早期的“三团队-两季度”研发迭代模式开始，英伟达的研发目标就一直走在市场的前端。英伟达研发团队分为软件工程、硬件工程、超大规模集成电路工程、工艺工程、架构和算法团队，负责研究开发统一的硬件和软件架构，提供领先市场的图像加速技术。英伟达研发人员数量持续增长，截至 2023 财年达到 19532 人。

图表 27：1999-2023 财年英伟达研发费用率



来源：英伟达，中泰证券研究所

图表 28：2001-2023 财年英伟达研发人员数量



来源：英伟达，中泰证券研究所

2、英伟达发展历程三部曲

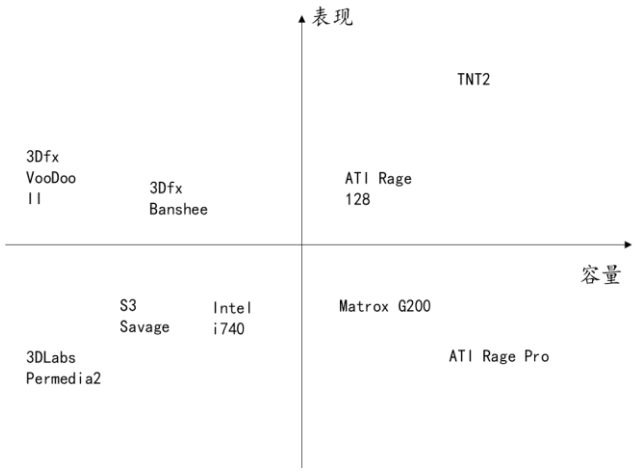
2.1 1993-2000：初具规模，提升研发效率战胜对手

■ 1993 年黄仁勋、克里斯·马拉科夫斯基和柯蒂斯·普利姆在美国加州创立了英伟达。在创建之初，公司设想着个人电脑将会成为游戏、多媒体的主流消费设备。90 年代初，高性能图像被使用在工作站和视频游戏机上，在此之后，3 件独立事件改变了这样的情况，推动了 3D 图像市场的发展：

（1）微软推出的 Windows 95 包括了视频、音频功能，刺激了多媒体市场发展。3D 图形逐渐增加的重要使得个人电脑制造商的差异性更加明显；
（2）电脑仿真渲染动画出现，3D 动作游戏登陆 PC 平台；
（3）在摩尔定律的推动下，IC 的集成度不断提高，能够将大量 3D 图形处理器放在一个芯片上。半导体设计和制造的不断进步，使得以前只能在工作站级别获得的高性能 3D 图形技术，现在能够以合理的价格获得。而图形处理器的不断发展也带动英伟达整体规模不断增长。

- **推进“三团队-两季度”研发模式，新品不断迭代满足下游需求。**一般图形市场产品有两个开发周期：6-9 个月和 12-18 个月，英伟达执行了“三团队-两季度”的运营模式，具体方式包括三个并行开发团队，专注于三个独立的分阶段产品开发。一个在第一年秋季，一个是在第二年春季以及第二年秋季。这样的运行方式允许公司每 6 个月推出一次新产品，与图形市场产品周期一致，并且领先市场 1-2 个研发周期，从而满足下游需求变化。
- **不断丰富产品矩阵满足下游客户不同需求。**GeForce 系列是英伟达为台式机提供图像处理的芯片。2002 年 11 月，英伟达推出为个人电脑市场消费者开发的产品线 GeForce FX 系列。GeForce 系列的其他产品，比如 GeForce2、GeForce3、GeForce4 都能够为不同价位的主流产品提供最高的性能。
- **高研发带来技术水平不断升级，英伟达在 1996 年后接连推出旗舰产品，击败行业竞争者。**英伟达在 1996 年，推出 NV3 系列的 Riva 128 芯片，在性能方面具有优势，并且芯片尺寸更小，因此结构成本更低，通过数据对比，RIVA128 甚至优于下一年 Intel 推出的 i740，而且 i740 不支持任何 OpenGL 驱动程序。在英伟达推出 RIVA TNT 时已经没有产品能够和其匹敌。1999 年，Intel 宣布完全退出独立显卡芯片组业务。而在 2000 年英伟达推出 GeForce 256，全面超过当时行业最大竞争者 3dfx，最终 3dfx 宣布破产并且被英伟达收购。

图表 29 ：3D 图像市场竞争图



来源：英伟达，H&Q，中泰证券研究所

图表 30 ：Riva 128 与 i740 对比

	RIVA 128ZX	i740
填充率（百万像素/秒）	100	66
多边形率	1.5	0.5
晶片大小	8.2mm	11.4mm
最大电容	3.5W	5.8W（散热片需求）
DAC	230MHz	205MHz
储存器接口	128-bit	64-bit
帧缓冲总线带宽	1.6GB/s	800MB/s
纹理缓存	8192byte	256byte

来源：英伟达，Intel，中泰证券研究所

- **总结：英伟达在成立初期，面对着技术不成熟、行业竞争激烈等难题，依靠“三团队-两季度”的研发模式和以 Direct X、OpenGL 为代表的 API 出现，不断进行技术更新、降低产品价格，挤压同业竞争对手的生存空间，从而在早期的 GPU 市场上存活下来。**

2.2 2001-2006：寡头垄断，逐步成为独显市场霸主

- **英伟达在游戏市场率先取得突破。**自 1999 年 Geforce 系列推出以来，它一直在游戏性能的创新和提升方面保持领先地位。GeForce 系列显卡被广大游戏爱好者和电子竞技玩家所推崇，因为它们能提供极高的图形渲染能力和实时光线追踪技术，以实现更加真实的游戏体验。与此同时，Xbox 为代表的游戏主机兴起助力了英伟达在游戏 GPU 行业的发展。英伟达为 Xbox 视频游戏系统设计的处理器利用双处理架构推动了其优秀的图形、音频和网络功能，确立了英伟达在游戏机市场的稳固地位。虽然后续英伟达未能持续成为 Xbox GPU 供应商，但是早期在 Xbox 上的成功已经为英伟达在游戏市场的发展奠定基础。
- **通过收购，技术开发以及广纳人才，英伟达进一步开拓市场，增强自身实力，保持市场领先地位。**英伟达预测未来能够实现通话和多媒体功能的手机半导体将会大放异彩，因此积极通过收购移动端公司来布局移动端图像芯片产业，并紧密融合 Direct3D 和 OpenGL 以最大程度地支持第三方软件。Direct3D 和 OpenGL 作为应用程序编程接口，使软件开发人员能够在不需要深入了解硬件特性的情况下编写应用程序，从而在 3D 图形、视频媒体通信以及超低功耗方面保持其技术的领先地位。为了维护市场的领导地位，英伟达积极地招募业界经验丰富的 3D 图形和通信工程师，并持续开发新一代的 GPU、MCP 以及 UMP。

图表 31：2002-2013 年英伟达收购汇总

年份	收购公司	收购目的
2002	Exluna	提供设计人才，推动 CG 语言进入电影行业
2003	MediaQ	打开快速增长的移动和手持市场领域
2004	iReady	获得用于支持超高性能以太网络的传输技术
2005	ULI Electronics	ULI 为 ATI 提供南桥部件
2006	Hybrid Graphics	打开手持设备领域，开发图像解决方案
2006	PortalPlayer	将 GPU 和 PortalPlayer 应用处理器结合，完善手持产品战略分布
2008	Ageia	将 PhysX 物理引擎和 GPU 集成
2011	Icera	帮助代工厂商缩短产品上线时间，满足下一代移动计算需求
2013	PGI	为 HPC 系统提供关键组件

来源：各公司年报，中泰证券研究所

- **英伟达全面完善产品线，产品覆盖高中低端下游各应用市场。**经过了不断的发展，英伟达的产品线逐渐丰富，覆盖了多种不同的下游应用。首当其冲的是 GeForce 系列显卡，主要针对的是个人电脑的游戏领域。同时，为了满足科研和企业市场的需求，英伟达推出了 Tesla 和 Quadro 系列的 GPU，这些产品被广泛应用于机器学习、数据科学、计算机视觉等领域。此外，英伟达还在汽车自动驾驶等前沿领域推出了专门的解决方案，如 Jetson 和 DRIVE 系列。

图表 32：英伟达产品线

来源：英伟达，中泰证券研究所

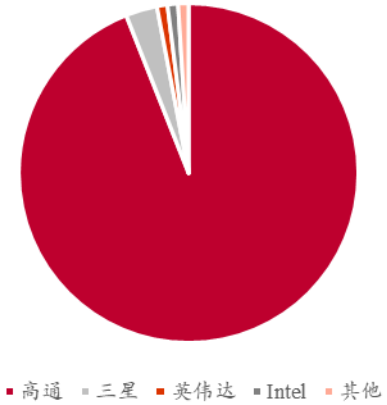
- **总结：在经历了行业发展初期洗牌之后，英伟达在独立显卡市场上的主要竞争对手只剩下 ATI，整个独立显卡行业逐步向寡头垄断转变。在这六年时间里，公司曾因产品定位和市场需求贴合度不够而落后，但通过坚持投入研发，完善产品线，竞争力持续提升。2006 年 AMD 收购 ATI 后，英伟达终成行业霸主。**

2.3 2007-2023：重“芯”开始，引领人工智能计算

- **智能手机浪潮来临，但是由于时机和定位上的失误，英伟达错失机遇。**自苹果系列产品推出后，智能手机成为一大热点，引领时代风潮。在这

样的背景下，Intel 推出了 Atom，英伟达推出了 Tegra。然而手机芯片市场并没有像 PC 市场一样被这两个大场占领，相反高通依靠着基带技术的垄断成为了移动端市场的主流。Tegra 系列在最初是依靠英伟达在图像处理的优势为平板和游戏机研发的。当英伟达推出 Tegra 2 系列时，3G/4G 技术开始成为移动端市场追逐的目标。但由于 Tegra 3 没有能够整合基带技术从而失去了占领市场的必要条件，而 Tegra 4 迟迟没有发布以及低性价比也失去了市场的青睐。

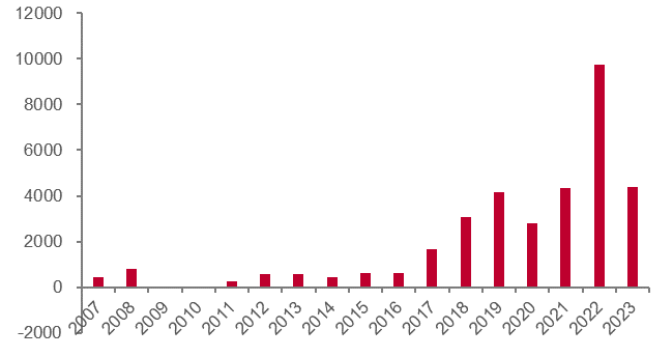
图表 33：2014 年 3G/4G 市场份额



来源：Strategy Analytics，中泰证券研究所

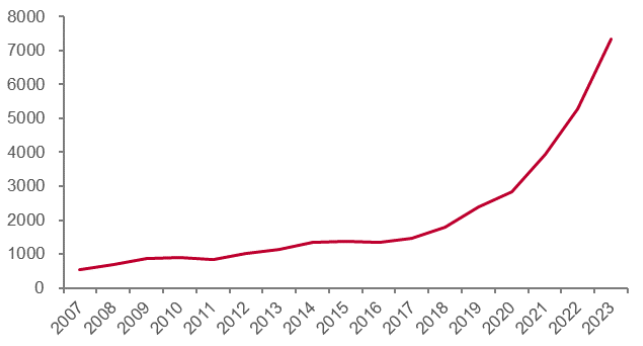
- **英伟达退出手机市场，转向汽车、人工智能市场，调整竞争策略。**在经历了手机市场的挑战后，英伟达进行了战略调整，从手机市场退出，并将其研发重心转向了汽车和人工智能市场。这一转变对于英伟达来说，不仅是其业务发展的一次机遇，更是对于行业趋势的敏锐洞察。在汽车市场中，英伟达通过自己的技术优势，推出了一系列高效能的自动驾驶处理器，逐渐在此领域确立了自己的领导地位。而在人工智能市场，英伟达的 GPU 产品凭借其超强的并行计算能力，成为了支撑深度学习和机器学习应用的核心设备，展现出强大的市场竞争力。
- **事实证明英伟达的转型抓住了市场需求的改变，英伟达 2015-2023 年营收增速可观。**在 2008 年全球经济危机爆发之后，英伟达的业务收入也受到了经济危机的影响，在 2009 年和 2010 年财报中净利润呈现负值，亏损达到三千万和六千万美元，但在之后的几年中，英伟达依靠着在游戏行业中的基础，继续拓宽在可视化计算、人工智能业务，并且借助于比特币和区块链对于显卡芯片的高增长需求，在 2016 年之后，保持着高增长的营业收入增长趋势。高增长的净利润得益于英伟达每年研发费用的投入，使得英伟达的产品领先同行业的竞争对手，更快地拓展新业务，更早地形成进入壁垒。而净利润的增长又会使得英伟达有更多的资金进行新产品的研发，从而达成良性循环，占据市场领先的地位。

图表 34：英伟达 2007-2023 财年净利润（百万美元）



来源：英伟达，中泰证券研究所

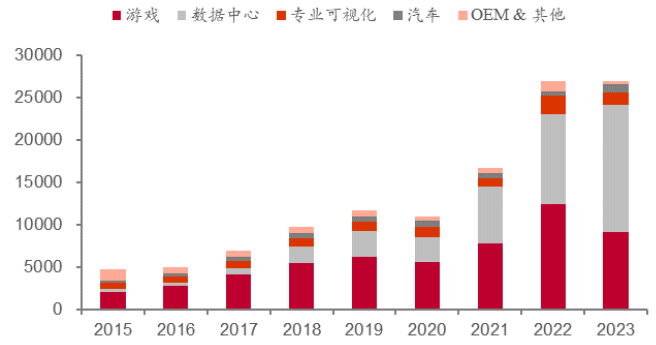
图表 35：英伟达 2007-2023 财年研发费用（百万美元）



来源：英伟达，中泰证券研究所

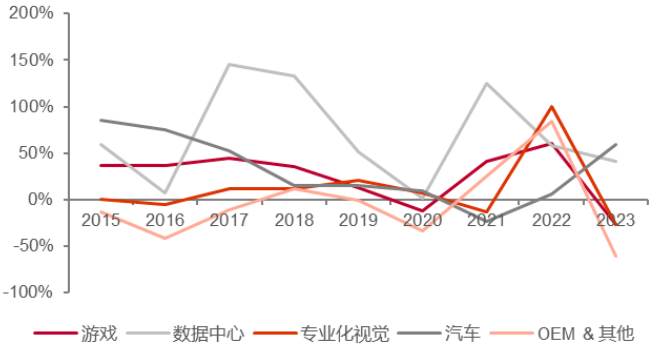
- **具体从各终端来看，英伟达各方面业务保持收入增长，全面发展。**从英伟达终端用户划分来看，各终端产品收入都保持着稳定增长。数据中心发展加速，游戏终端依旧是英伟达重要的业务收入基础。

图表 36：英伟达终端用户收入情况（百万美元）



来源：英伟达，中泰证券研究所

图表 37：英伟达终端用户收入年增长率



来源：英伟达，中泰证券研究所

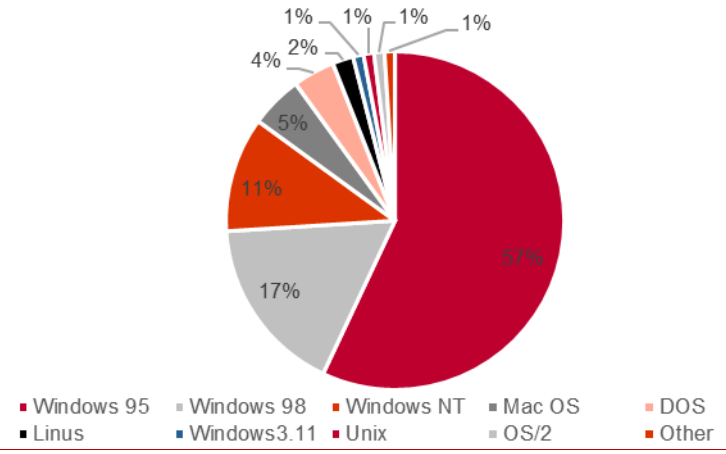
- **英伟达不断更新产品技术， AI 市场成为主要目标，给英伟达带来新的增长。**随着 AI 市场的蓬勃发展，英伟达敏锐地将其定位为公司的主要发展目标。英伟达开发了一系列专门针对 AI 应用的 GPU，如 Tesla、Titan 以及 Quadro 系列。这些产品能够高效处理深度学习和机器学习的大规模并行计算，极大地推动了 AI 的发展。2020 年在 SC20 超级计算大会上，NVIDIA 发布了新一代 DGX Station A100 以及 NVIDIA A100 80GB GPU 支持诸如 BERT Large 推理等复杂的对话式 AI 模型。此后在 2022 年 3 月，NVIDIA 又宣布推出第四代 NVIDIA® DGX™系统，是全球首个基于全新 NVIDIA H100 Tensor Core GPU 的 AI 平台，彻底占据 AI 市场领先地位。
- **总结：当 iPhone 出现后，全球智能手机市场的帷幕被拉开。移动端 GPU 市场逐渐成为了大家的焦点。但英伟达并没有能够在手机 GPU 市场取得较打的成功，但英伟达将手机 GPU 芯片 Tegra 用在了其他应用领域，为公司打开了新的业务市场。**

3、英伟达发展历程总结，借鉴意义

3.1 深耕 GPU 算力领域，研发为导向不断提升产品竞争力

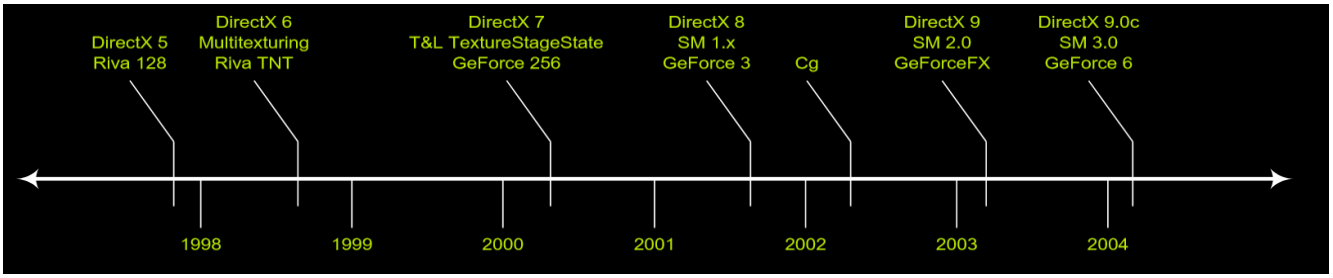
- **采用主流 API，借助微软推广产品。**从英伟达创立时，公司就以市场需求为导向。通过匹配主流 API，不断技术更新逐渐减低产品价格，达到消费者需求，以此来达到一家初创公司占领市场的目的。英伟达在设计 NV2 后的产品时，都将微软推出的 DirectX 作为匹配的 API。凭借着微软 Windows 系列在操作系统市场占有大量份额，同时对 DirectX 和 OpenGL 加速优化，使得英伟达的产品广受欢迎。

图表 38：1998 年操作系统占比



来源：IDC，中泰证券研究所

图表 39：英伟达追随 DirectX 升级开发产品



来源：各公司年报、中泰证券研究所

- **压缩开发周期领先市场，为下游厂商提供更好的产品。**英伟达把握住了从 2D 到 3D 过渡的风潮，通过成熟的研发体系，用速度甩开 2D 图形厂商。英伟达图形业务的快速产品周期得益于其运营模式。一般图形市场产品有两个开发周期：6-9 个月和 12-18 个月，英伟达执行了“三团队-两季度”的运营模式，具体方式包括三个并行开发团队，专注于三个独立的分阶段产品开发。一个在第一年秋季，一个是在第二年春季以及第二年秋季。这样的运行方式允许公司每 6 个月推出一次新产品，与图形市场产品周期一致，并且领先市场 1-2 个研发周期。此外为解决芯片硬件开发比软件开发慢的问题（软件可以快速测试并经过调试，通常是每

天或每周一次；相比之下，芯片硬件就必须构建掩模并在进行电子测试之前完成初步制造），英伟达大力投资了仿真技术，从而提升效率。

- **在产品布局多元化初期，用产品交叉服务市场。**英伟达在经历了手机端芯片市场开拓的失败之后，并没有停止 Tegra 处理器的研发，而是改变产品定位，将 Tegra 处理器运用在智能汽车、智慧城市和云端服务上。于是英伟达初步奠定了“两产品条线-四市场”的商业模式。两产品条线包括了英伟达传统产品 GPU 和 Tegra 处理器，而四市场则包括了游戏、企业级、移动端、云端。

图表 40：英伟达商业布局

GPU	GeForce	游戏
	Quadro	专业工作站
	Tesla	超级计算机、工作站
	GRID 图像模块	服务器工作站
	GRID 系统	可视化计算
Tegra	Tegra	平板、手机、游戏设备
	Icera	基带处理器
	集成芯片解决方案	集成 Tegra 和 Icera
	Shield 项目	游戏、云端处理
	Tegra VCM	智能汽车
	嵌入计算平台	数字消费设备

来源：英伟达、中泰证券研究所

- **英伟达的商业模式战略很好的应对了图像处理器市场的发展趋势。**当时的图像处理器市场产品细化，主要分布在游戏玩家、企业级、平板电脑和移动端用户，不同客户的需求差异化明显，针对不同下游英伟达推出了对应的产品方案：

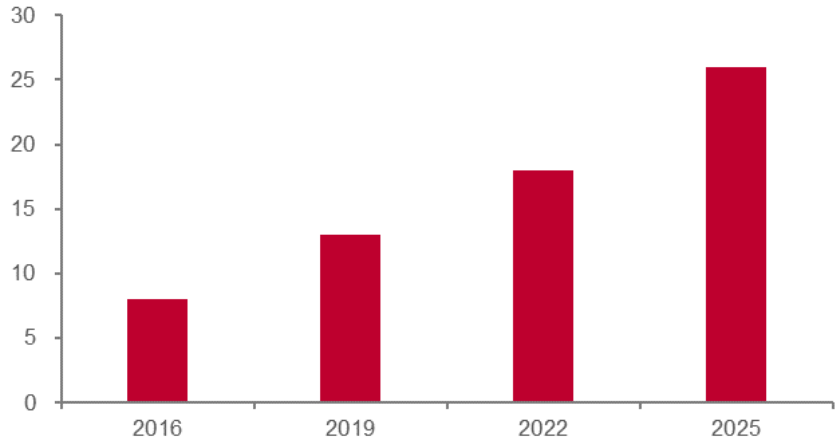
（1）**游戏市场：**玩家希望能够在不同的平台无缝的进行游戏体验，英伟达为此推出了端到端的服务：游戏能够在云端运行，不需要玩家拥有足够高性能的电脑。大大提高了玩家碎片时间的利用率和娱乐的灵活性。

（2）**企业级：**产品则是为汽车、电影、天然气等行业提供可视化解决方案，目的是提高行业生产力。英伟达面向企业市场的产品包括用于工作站的 Quadro，用于高性能计算服务器的 Tesla 和用于企业 VDI 应用程序的 GRID。

（3）**移动端：**英伟达不再将移动端客户拘泥于手机端用户，而是将移动端扩展到移动智能设备市场，比如智能汽车、智能家居行业。英伟达的移动战略转变为了将 Tegra 应用到需要视觉设计的设备中。

（4）**云端服务：**伴随着计算机行业的发展也成为了可视化计算服务的重要一环。凭借云端技术，英伟达将 GPU 的应用从 PC 端拓展到服务器和数据中心，使得更多的用户可以使用。英伟达开发的 GRID 使 Adobe Photoshop 远程运行，并与应用程序交互。

图表 41：2016-2025 年自动驾驶规模（十亿美元）



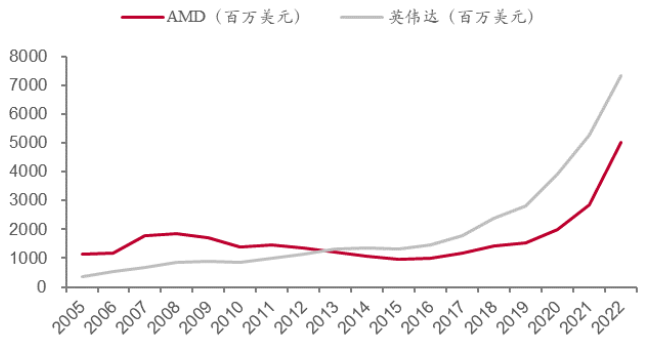
来源：Bain & Company，中泰证券研究所

- **英伟达 AMD 双寡头垄断显卡市场。**在 2009 年 Intel 取消 Larrabee 图形显卡项目之后，独立显卡市场逐渐成为了双寡头市场。相较于行业的潜在新进入者，英伟达和 AMD 拥有更长的经营历史、更大的客户基础、更全面的知识产权和专利保护，以及更多的融资、销售、营销和分销资源，二者共同构筑了行业新进入者无法逾越的天堑。研发方面，2005 年，AMD 的研发费用为 11 亿美元，是英伟达的 3.2 倍左右，而到了 2022 年，英伟达的研发费用达到了 73.4 亿美元（对应 2023 财年），是 AMD 的 1.47 倍。由于性能、构建、价格的不同，二者逐步产生差异化，形成了错位竞争。至 2022 年第三季度，英伟达基本占据 88%市场份额，AMD 则降低至 8%。

图表 42：英伟达显卡合作伙伴多于 AMD

来源：AMD，英伟达，中泰证券研究所

图表 43：AMD 的研发费用被英伟达反超



来源：AMD，英伟达，中泰证券研究所

图表 44：英伟达在天梯图覆盖面广，高端产品领先 AMD

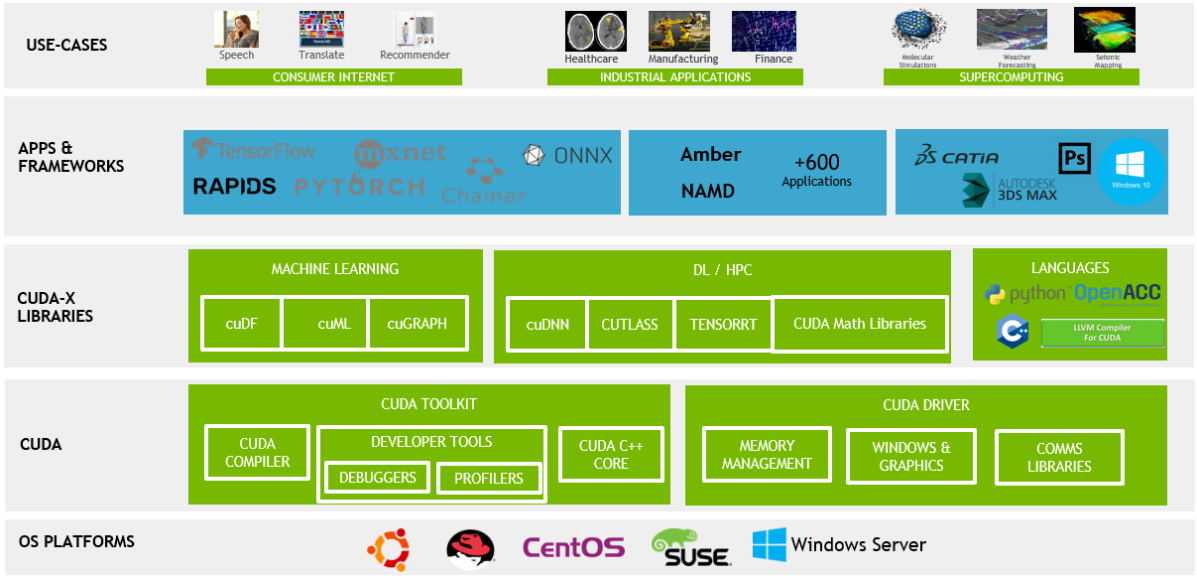
来源：各公司年报，中泰证券研究所

- **研发投入带给英伟达高回报，在主流游戏和显卡天梯测评上，AMD 落后于英伟达。**英伟达在 2018 年推出的 Titan RTX 和 RTX 2080 Ti 全面超过当时的 Radeon VII，其采取的策略是推出比 AMD 稍高的性能和价格。即使技术比 AMD 领先，也会等到 AMD 推出更高性能的产品之后，才会推出，以此来获得比 AMD 更高的收益。
- **专利数量方面英伟达逐步反超 AMD。**AMD 此前在专利数量上一直多于英伟达，但申请的数量呈现下降趋势。英伟达在 2011 年之后申请专利数量开始爆发，主要因其在 2007 年之后开始研发移动端 GPU 和深度学习领域，最终给 GPU 市场带来了新的框架和更高性能的芯片。

3.2 CUDA 自成体系：从单一产业到生态链，构建强护城河

- **CUDA 助力英伟达成长为 AI 产业龙头，构建强大生态护城河壁垒。**CUDA 是英伟达基于其生产的 GPUs 的一个并行计算平台和编程模型，目的是便于更多的技术人员参与开发。开发人员可以通过 C/C++、Fortran 等高级语言来调用 CUDA 的 API，来进行并行编程，达到高性能计算目的。CUDA 平台的出现使得利用 GPU 来训练神经网络等高算力模型的难度大大降低，将 GPU 的应用从 3D 游戏和图像处理拓展到科学计算、大数据处理、机器学习等领域。这种生态系统的建立让很多开发者依赖于 CUDA，进一步增加了英伟达的竞争优势。

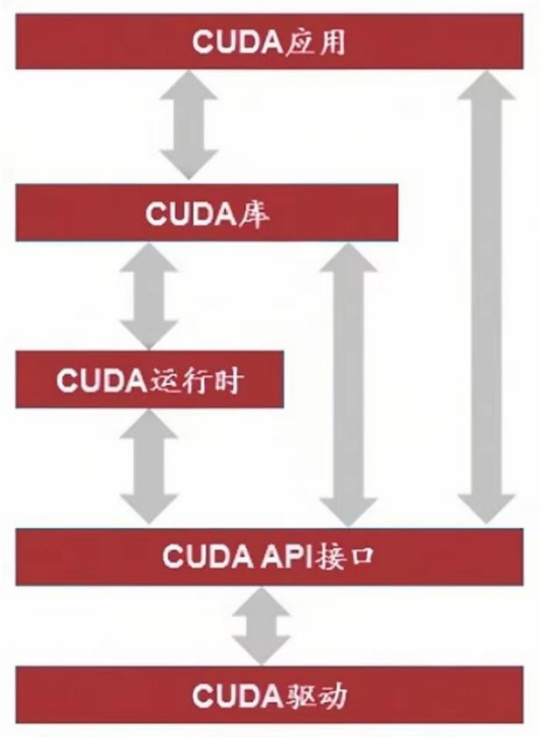
图表 45：CUDA 加速计算解决方案



来源：英伟达官网，中泰证券研究所

- **CUDA 的低成本和兼容性成为其最重要的吸引点之一。**英伟达的 CUDA 是一个免费、强大的并行计算平台和编程模型。安装过程简单且明确，让开发者能够轻松快速地启动并行编程。CUDA 对新手极其友好，特别是对 C 语言、C++和 Fortran 的开发者。同时为支持其他编程语言，如 Java、Python 等，CUDA 还提供第三方包装器进行扩展。为广大开发者提供了极大的便利和高效的编程体验。操作系统方面，CUDA 在多种操作系统上也都有良好的兼容性，包括 Windows、Linux 和 macOS。

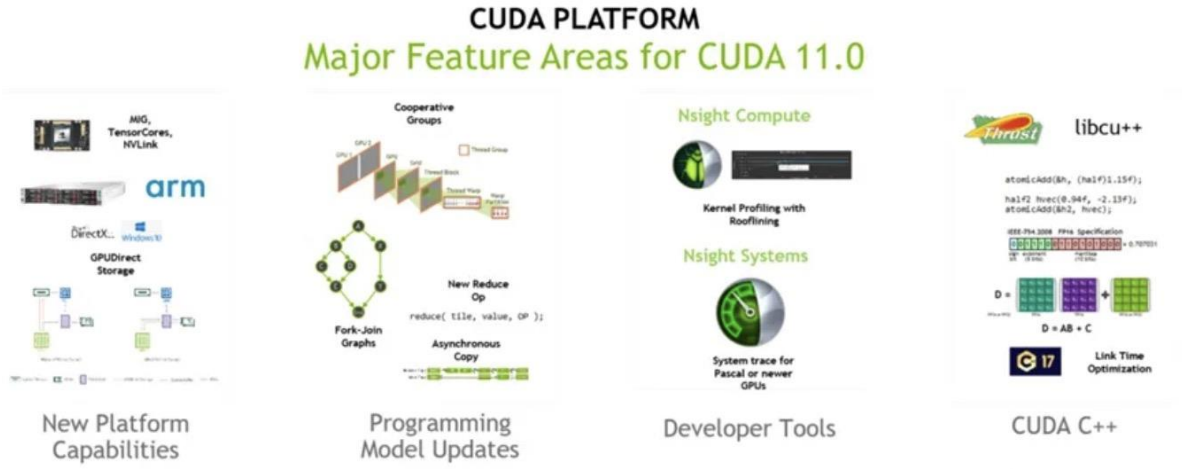
图表 46：CUDA 软件层



来源：The CUDA handbook，中泰证券研究所

- **CUDA 有着丰富的社区资源和代码库，为编程提供良好的支持。**英伟达的 CUDA 享有强大的社区资源，这个社区由专业的开发者和领域专家组成，他们通过分享经验和解答疑难问题，为 CUDA 的学习和应用提供了丰富的支持。另外，CUDA 的代码库资源涵盖各种计算应用，具有极高的参考价值，为开发者在并行计算领域的创新和实践提供了宝贵的资源。这两大特点共同推动了 CUDA 在并行计算领域的领先地位。

图表 47：CUDA 11.0 主要特点

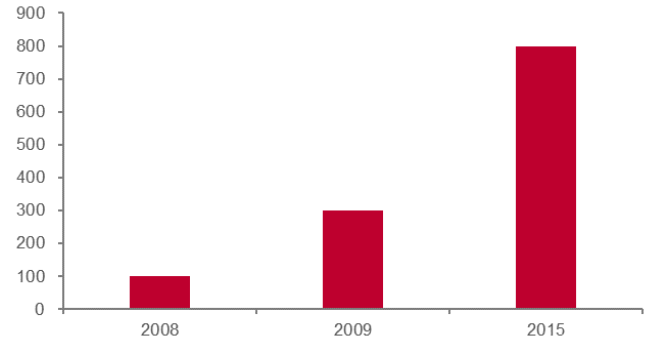


来源：英伟达官网，中泰证券研究所

- **CUDA 借助燕尾服效应，搭配 GeForce 覆盖多元市场。**CUDA 技术最初是为了配合 GeForce 系列芯片而推出的，利用 GeForce 在游戏市场的广泛覆盖率，作为一个技术杠杆，推动 CUDA 的普及和发展。作为一项可以帮助 GeForce 拓展新的市场的重要技术，CUDA 极大地提高了视频和图像应用（如 CyberLink、Motion DSP 和 Nero）的性能，实现了多倍的效率提升。
- **创业公司的大量采用使得 CUDA 应用场景进一步得到拓展，游戏不再是唯一应用领域。**随着时间的推移，超过一百家创业公司开始利用 CUDA 的强大计算能力，使其应用领域得以扩展，不再局限于游戏方面。在视频编码领域，英伟达与 Elemental 公司合作，利用并行计算技术加速了高清视频的压缩、上传和存储速度。这一成功的合作不仅体现了 CUDA 在各种场景下的适用性，也进一步推动了 CUDA 技术的发展。当 Elemental 公司后被亚马逊收购，其基于 CUDA 的视频处理技术也成为 AWS 的服务组成部分，这一过程也让 CUDA 的使用场景得到了进一步的丰富和拓宽。

- **CUDA 形成完整生态链，通过大学普及学习以推广 CUDA。**英伟达将 CUDA 引入了大学的课堂中，从源头上扩大了 CUDA 的使用范围和受众群体。早在 2010 年，已经有关于 CUDA 数千篇论文，超过 350 所大学进行 CUDA 教学课程。在此基础之上，英伟达建立了 CUDA 认证计划、研究中心、教学中心，不断完善 CUDA 的生态链。从结果看：2008 年仅有 100 所大学教学 CUDA 课程，在 2010 年英伟达全球建立了 20 个 CUDA 研发中心后，2015 年已有 800 所大学开放 CUDA 课程。

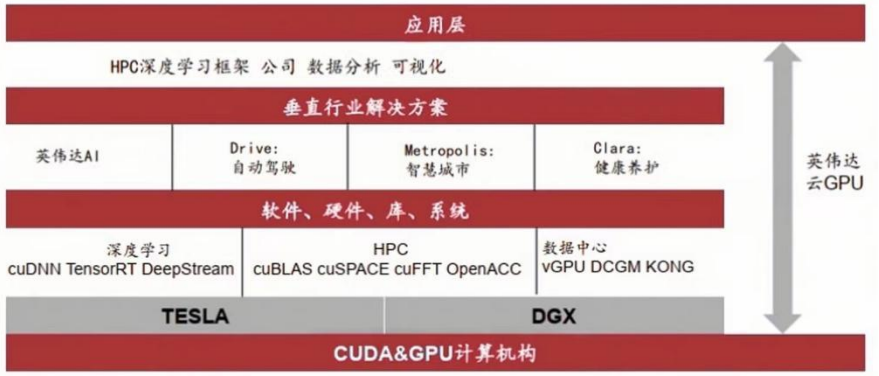
图表 48：大学教授 CUDA 数量（所）



来源：英伟达，中泰证券研究所

- **对比 OpenCL，CUDA 是英伟达 GPU 编程的更优解。**OpenCL 虽然具有更广的兼容性，但 CUDA 由于与英伟达的硬件紧密结合，能更有效地利用其 GPU 的性能。同时，CUDA 的编程模型相比 OpenCL 更加简洁，易用，并提供完整的开发工具链。此外，CUDA 的社区资源丰富，代码库多样，使得在科学计算、深度学习等领域的应用更为便捷。因此，对英伟达 GPU 的开发者来说，CUDA 往往是更优的选择。

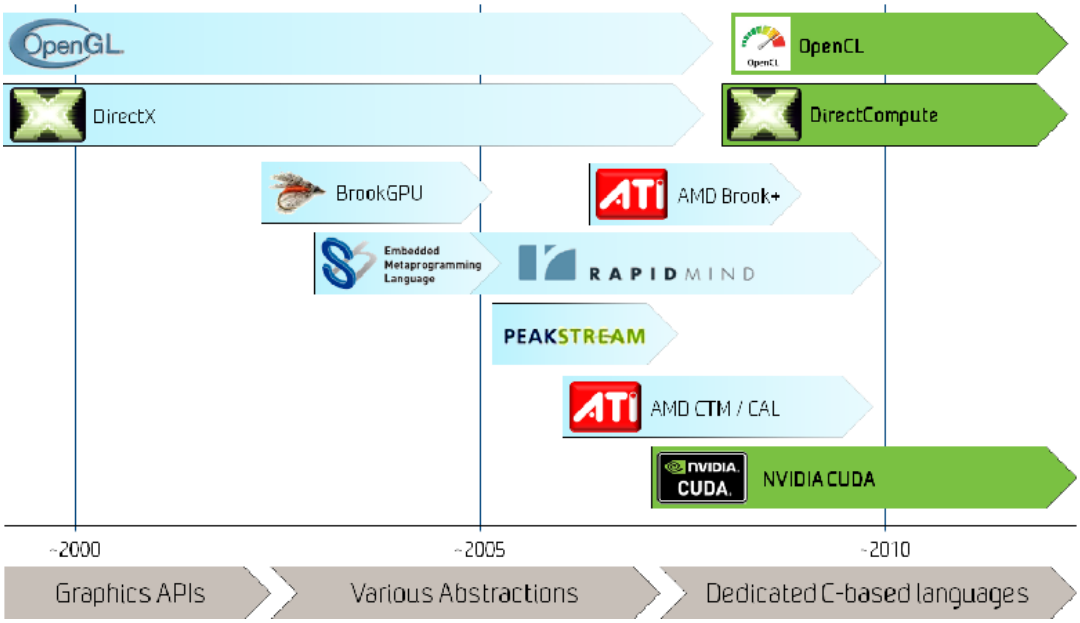
图表 49：CUDA 成为英伟达生态基础



来源：英伟达，中泰证券研究所

- **对比 ADM 的 CTM 编程模型，CUDA 拥有更广泛的应用和更高的操作性。**从操作性来说，由于 CTM 更接近硬件，因此开发者需要有更深入的硬件知识才能进行开发，但是这也意味着 CTM 能够提供更精细的控制和优化。对比之下，CUDA 提供了一套完整的开发工具链，包括编译器、调试器和性能分析工具，以及丰富的库函数，为开发者提供了极大的便利。从应用来说，CUDA 已经在各种领域获得了广泛的应用，尤其是在科学计算和深度学习等领域，CUDA 拥有大量的优化库和开发工具。而 CTM 的应用相对较少，但是由于它提供了对硬件的低级别控制，因此在一些特定的应用场景中会具有优势。

图表 50：GPU 编程平台发展历史



来源：Semantic Scholar，中泰证券研究所

- **对比微软的 DirectCompute，CUDA 胜在配套设施的支持。**与 DirectCompute 相比，CUDA 由于其丰富的功能库，完善的开发工具和广泛的应用支持，尤其在科学计算和深度学习领域，具有明显优势。CUDA 在英伟达 GPU 上的性能优化也更为出色。而 DirectCompute 作为跨平台工具，其优势在于与 DirectX 的兼容性以及对多种硬件的支持。但从英伟达 GPU 的应用广泛度来看，使用 CUDA 才是开发者的首选。总的来说，虽然 DirectCompute 的通用性更强，但英伟达的 CUDA 在功能、性能和应用范围上提供了更强大的支持，对于使用英伟达硬件的开发者来说是更优的选择。
- **CUDA 的开发提升了英伟达的品牌竞争力和影响力。**CUDA 的开发使英伟达的 GPU 超越了仅用于图形处理的传统角色，转变为通用的并行计算设备，极大地提升了其在市场上的竞争力。英伟达因此能够满足广泛的高性能计算和人工智能需求，使其产品得以进入新的市场领域。同时，随着 CUDA 在各类高性能计算任务，特别是人工智能领域的广泛应用，英伟达的品牌影响力得到了显著增强。越来越多的人开始认知和使用英伟达的产品，这不仅加强了英伟达的市场地位，也为其未来的发展奠定了坚实的基础。

图表 51：英伟达人工智能生态系统

来源：英伟达，中泰证券研究所

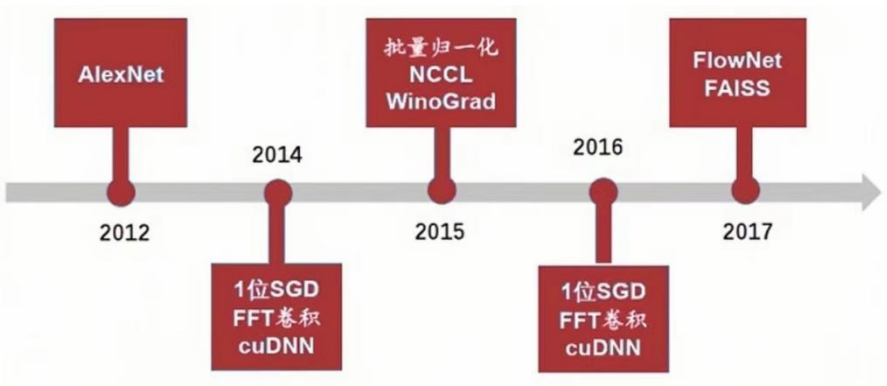
- **CUDA 促进了英伟达的产品创新，激发更多可能性。**CUDA 的开发推动了英伟达在并行计算技术领域的创新，尤其在硬件架构方面。这不仅体现在优化了的 GPU 架构上，例如将流多处理器（SM）配置为处理并行线程的方式，也在软件架构上如 CUDA 自身的持续更新和优化。为了更有效地满足用户对于更高性能和更易用并行计算工具的需求，英伟达不断创新，致力于提高 CUDA 的性能和用户体验。这一切不仅反映出英伟达对创新的重视，也为其在并行计算技术领域的领导地位提供了坚实的技术支撑。

图表 52：CUDA 对应 GPU 架构发展

	Kepler	Maxwell	Pascal	Volta	Turing
推出时间	2013	2014	2016	2017	2018
制程	28nm	28nm	TSMC 16nm/三星 14nm	TSMC 12nm	TSMC 12nm
芯片尺寸(平方毫米)	118	148	471	815	754
晶体管（十亿）	1.3	1.87	11.8	21.1	18.6
CUDA 核心数量	1536	2048	3840	5120	4608
存储容量	12GB	24GB	16GB	16GB	48GB
表现（浮点运算）	5.04	6.8	10.6	15.7	16

来源：英伟达，中泰证券研究所

图表 53：CUDA 通过并行架构的改进



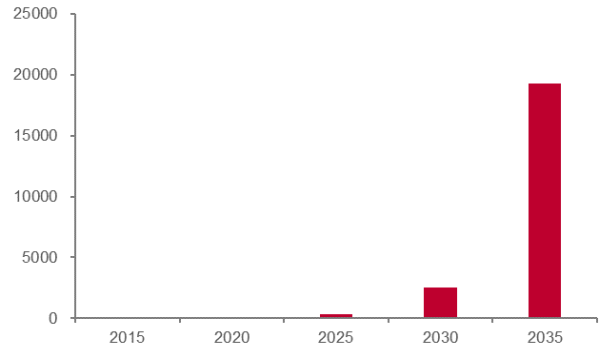
来源：英伟达，中泰证券研究所

- **英伟达的 CUDA 技术凭借其广泛应用和强大合作伙伴网络，巩固了英伟达的领导地位。**CUDA 技术在众多领域均有广泛应用，包括但不限于深度学习、图像和自然语言处理、天气模拟、流体动力学、分子动力学、量子化学以及天体物理模拟。因此，适配 CUDA 的应用程序数量繁多，进而催生了对 CUDA 的广阔需求空间。英伟达与诸如 Dell、HP、联想等知名 OEM 厂商，以及 Netapp、Pure Storage 等渠道合作伙伴和如埃森哲等服务公司展开了深度合作。
- **CUDA 整合英伟达体系，培养了开发者和使用者的用户粘性。**当开发者融入 CUDA 的生态系统，他们往往会被其卓越的计算性能、充裕的库函数和出色的易用性所吸引，因此更倾向于持续利用此技术。另一方面，为 CUDA 优化的代码移植至其他平台通常需要消耗大量的精力和时间，这进一步增强了客户的留存度。此外，英伟达不断推陈出新，发布新的硬件产品及 CUDA 版本，从而维持用户对其技术的关注并持续使用。这种深度使用使得用户在选购硬件产品时倾向于选择对 CUDA 有更好支持的英伟达产品，进而建立起稳固的客户忠诚度。

3.3 抓住人工智能发展浪潮，顺利转型切入算力芯片领域

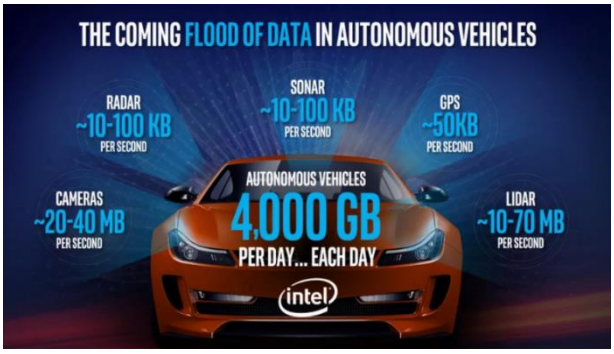
- 根据 IDC 的测算，全球数据总量将以每年 50%的增速不断增长，在 2025 年数据量会增加到 334ZB,在 2035 年则将达到 19267ZB。**随着 5G 落地，应用方案更加具象化，未来随之出现的数据总量和数据分析需求将会持续上升。**增长的数据量主要来源于 IoT、移动互联网、智慧城市、自动驾驶。大数据的应用将会从商业分析向工业、交通、政府管理、医疗、教育等等行业渗透，并且成为产业供应链中不可或缺的一部分。

图表 54：全球数据总量（ZB）



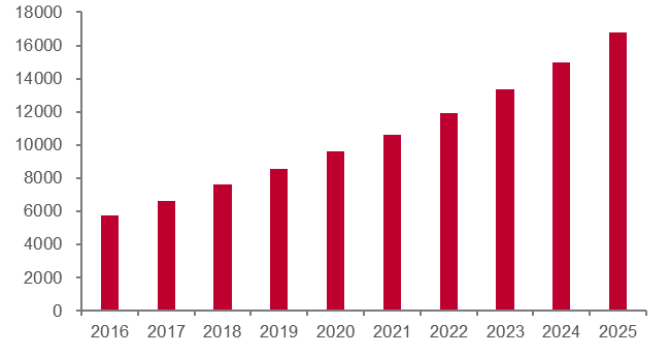
来源：IDC，中泰证券研究所

图表 55：Intel 测算的数据流



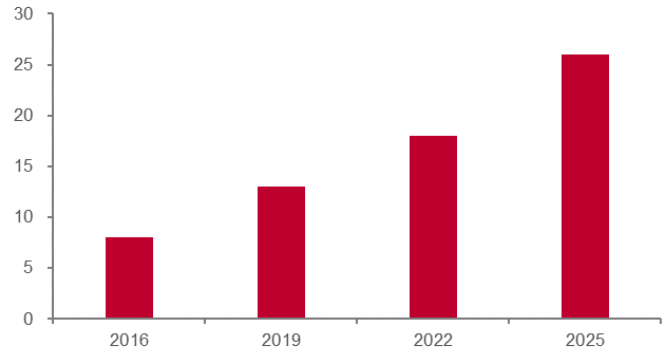
来源：Intel，中泰证券研究所

图表 56：全球 ADAS 市场规模扩大（亿美元）



来源：BIS Research，中泰证券研究所

图表 57：全球自动驾驶功能市场规模扩大（十亿美元）



来源：Bain & Company，中泰证券研究所

- **从目前的测算来看，智能驾驶将是算力要求最高的应用层面。**一方面汽车驾驶对于安全可靠性性要求最高，另一方面 L5 级别的汽车会携带的传感器将达到 32 个，据麦肯锡估算一辆自动驾驶汽车的数据量将达到 4TB/h，Intel 测算出的一天数据量将达到 4000GB。而英伟达的 Xavier 目前只有 1.3TFlops，还达不到处理 L5 的数据能力，自动驾驶和 ADAS 市场在接下来的 10 年之间有望保持较高增长的态势，因此智能驾驶以及 ADAS 存在着巨大的算力缺口。

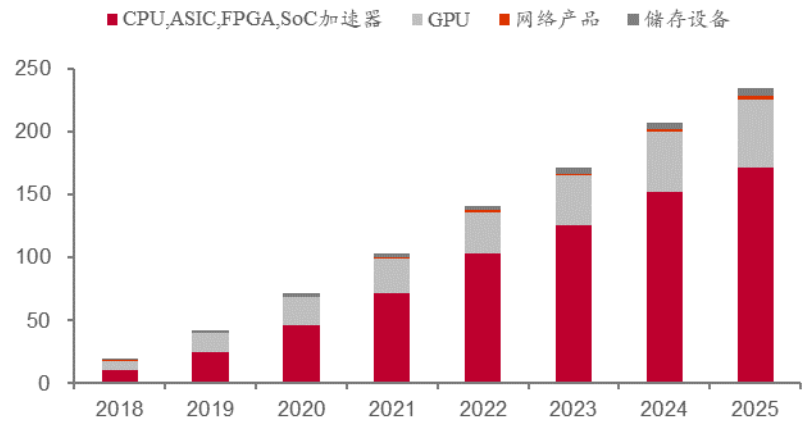
图表 58：智能驾驶层级越高所需传感器越多

	L1	L2	L3	L4	L5
软件应用	主动巡航控制 车道偏离警告系统	停车辅助 车道保持辅助	自动紧急制动 驾驶员监控 交通堵塞辅助	传感器融合 高速无人驾驶辅助	随时随地无人驾驶辅助
硬件需求		超声波传感器 4 个 长距雷达传感器 1 个 环视摄像头 1 个	超声波传感器 4 个 长距雷达传感器 1 个 短距雷达传感器 4 个 环视摄像头 1 个	超声波传感器 10 个 长距雷达传感器 2 个 短距雷达传感器 6 个 环视摄像头 5 个 长距离摄像头 2 个 立体摄像机 1 个 Ubolo1 个 激光雷达 1 个 航位推算 1 个	超声波传感器 10 个 长距雷达传感器 2 个 短距雷达传感器 6 个 环视摄像头 5 个 长距离摄像头 4 个 立体摄像机 2 个 Ubolo1 个 激光雷达 1 个 航位推算 1 个

来源：Deloitte，中泰证券研究所

- **人工智能产业将拉动 GPU 行业发展。**根据 Tractica 的数据，2018 年全球 AI 硬件市场的收入为 196 亿美元，其中 GPU 的收入占 36.2%为 71 亿美元。而在 2025 年将达到 2349 亿美元，其中 GPU 的收入占 23.2%为 545 亿美元。尽管 GPU 市场占比出现下滑，但是全球 AI 硬件市场在不断上升，将会给 GPU 市场带来更多的增长空间。

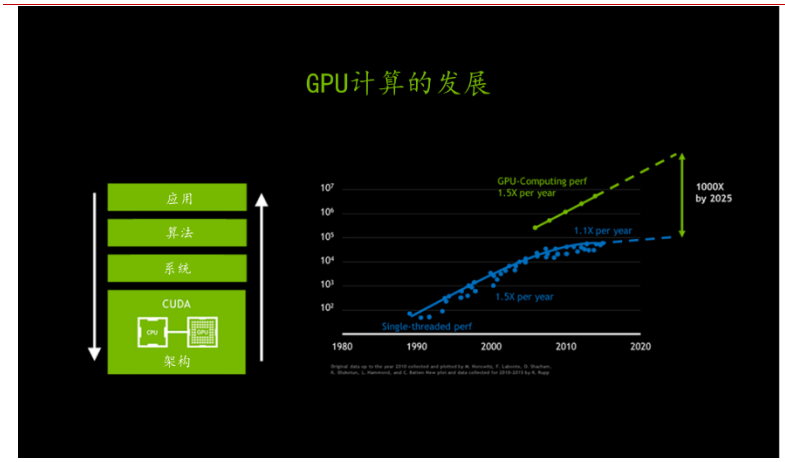
图表 59：2018-2025 年 AI 硬件市场收入（十亿美元）



来源：Tractica，中泰证券研究所

■ **CPU 受到摩尔定律约束，应用性能增幅下降。**人工智能的到来没有因为摩尔定律的放缓而停止到来。登纳德定律是通过缩小晶体管的尺寸和电压让设计师在保持功率密度时提高晶体管的密度和速度。但目前受到物理条件的限制，CPU 架构师需要大量增加电路和能量，获得有限的 ILP(指令级并行)。因此，在后摩尔定律时代，CPU 晶体管需要消耗更多的性能导致应用性能的小幅提高。最近几年 CPU 的性能仅以每年 10%的速度增长，而过去是每年 50%。

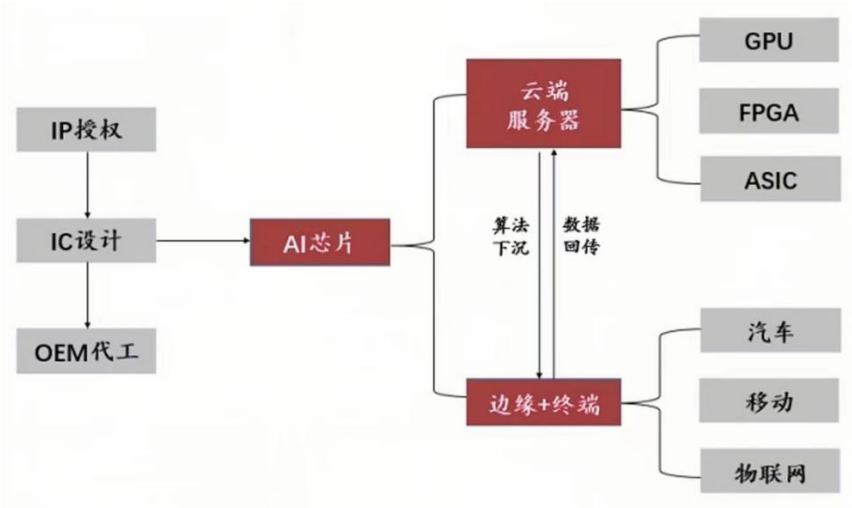
图表 60：GPU 打破摩尔定律



来源：英伟达，中泰证券研究所

■ **GPU 凭借超高运算速度成为高性能计算的宠儿，相比 CPU 提高了数倍的计算效率。**在人工智能领域，图形处理器（GPU）凭借其卓越的并行计算能力和大规模处理单元，成为大数据运算的主力。这在深度学习模型的训练和推断等任务中表现得尤为明显。尽管中央处理器（CPU）在顺序化任务处理方面具有优势，但其并行计算性能远不如 GPU。自 2006 年亚马逊在卷积神经网络（CNN）中首次使用 GPU，其效率就已显著优于 CPU。如今，随着技术不断发展和更新，GPU 在运算效率上进一步超越 CPU，坚定地确立了其在 AI 时代的算力核心地位。

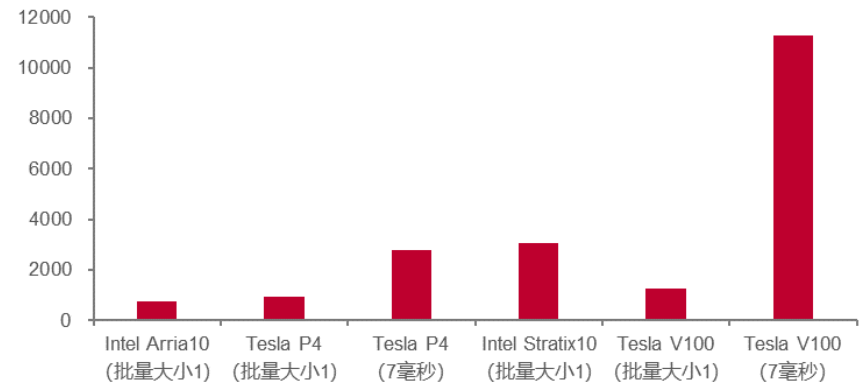
图表 61：AI 芯片产业链



来源：中泰证券研究所

- **尽管 FPGA 具有编程灵活性，但 GPU 的处理效率却更胜一筹。**FPGA 是专用集成电路（ASIC）中的一种半定制电路，被称为现场可编程门阵列，其性能介于定制电路和可编程器件之间。可以通过硬件描述语言（HDL）按照特定任务或者应用程序的需求来搭配相应的 HDL。相比较于 ASIC，FPGA 能够使用 OpenCL 快速编程，更加具有成本效益。微软预测基于 FPGA 的 BrainWave 推理平台可以在达到每秒约传输 500 张图像。但是对比这两种硬件，GPU 从效率上更显优势，例如英伟达的 Tesla P4 就能够在 75w 的能耗上一秒传输 1676 张图像。

图表 62：P4 传输速度大于 FPGA 架构芯片（秒）



来源：BIS Research，中泰证券研究所

图表 63：FPGA 和 GPU 对比

	比较	优势方
浮点计算	GPU 每秒的浮点运算总数要高于 FPGA	GPU
延时性	FPGA 中实现的算法提供了确定性定时，延迟比 GPU 小一个数量级。	FPGA
功耗	FPGA 的 GFPLOS/watt 比 GPU 高 3-4 倍	FPGA
接口	通过 PCI-E 接口，FPGA 可以通过几乎任何接口连接到其他设备	FPGA
通用性	为旧的 GPU 设计的软件在新的设备上还可以使用，而 FPGA 需要做出调整	GPU
大小	FPGA 较低的功耗对于散热的需求低，因此尺寸较小	GPU
开发	FPGA 开发昂贵并且困难	GPU

来源：Bain & Company，中泰证券研究所

- **英伟达以其创新力和高市场份额，在全球 GPU 市场中居于领导地位。**英伟达 GPU 领域享有显著的市场份额，在 2022 年 Q3 英伟达市场占有率达到了历史新高 88%。这一优势体现了其在行业中的龙头地位。其持续的技术创新和强大的研发实力，尤其是 CUDA 并行计算平台的推出，进一步巩固了其在市场中的优越地位。在人工智能和机器学习的加速应用下，英伟达的 GPU 已成为行业内的首选解决方案，广泛应用于各个领域，从游戏和专业视觉应用到数据中心和自动驾驶汽车。

- **回顾英伟达的发展历程，其成功的经验在于以下几点：**
（1）英伟达持续十几年深耕 GPU 高性能计算潜力，产品矩阵丰富，抓住下游人工智能和 5G 浪潮，推动 GPU 市场从游戏显卡转变为 AI 计算加速处理器；
（2）搭建 CUDA 生态，提高自身产品附加值，构建强大的软件护城河壁垒。通过大学、研究院加快 CUDA 开发，吸引人工智能行业人员加入 CUDA，将其打造成英伟达 GPU 核心竞争力；
（3）加大研发投入，强大的研发能力使英伟达能够实施创新技术，不断更新 GPU 架构拓展业务范围，扩大 GPU 市场，提高营收和利润率，达成产业链的良性循环。
- **解析复盘英伟达，借鉴算力龙头发展路径。**伴随生成式 AI 火热，大模型持续发展对算力提出更高要求，英伟达进入快速发展时期。我们认为复盘英伟达能够学习海外龙头的成长经验，为国产厂商提供发展思路。此外，随国内算力行业国产替代进程持续推进，看好国产供应链厂商发展机遇。

4、算力是 AI 底层土壤，从英伟达看国产发展机遇

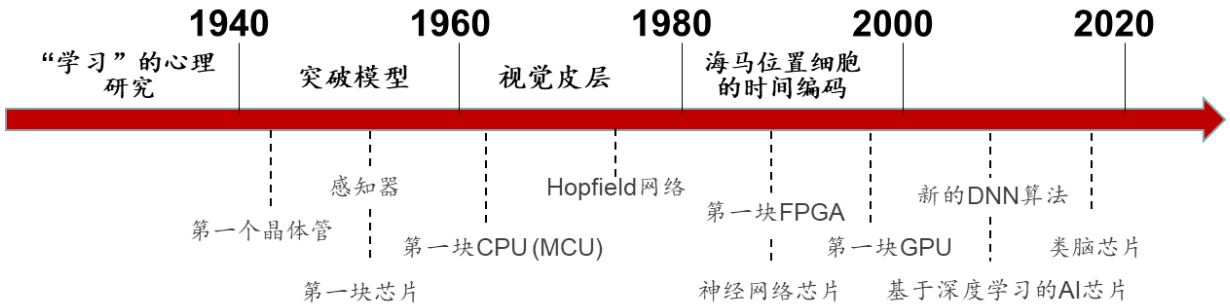
4.1 ChatGPT 激起 AI 浪潮，大模型升级推动算力提升

- **AI 人工智能的发展主要依赖两个领域的创新和演进：一是模仿人脑建立起来的数学模型和算法，其次是半导体集成电路 AI 芯片。**AI 的发展一直伴随着半导体芯片的演进过程，20 世纪 90 年代，贝尔实验室的杨立

昆（Yann LeCun）等人一起开发了可以通过训练来识别手写邮政编码的神经网络，但在那个时期，训练一个深度学习卷积神经网络（Convolutional Neural Network，CNN）需要 3 天的时间，因此无法实际使用，而硬件计算能力的不足，也导致了当时 AI 科技泡沫的破灭。

- **AI 芯片是 AI 发展的底层基石。**英伟达早在 1999 年就发明出 GPU，但直到 2009 年才由斯坦福大学发表论文介绍了如何利用现代 GPU 远超过多核 CPU 的计算能力（超过 70 倍），把 AI 训练时间从几周缩短到了几小时。算力、模型、数据一直是 AI 发展的三大要素，而 AI 芯片所代表的算力则是人工智能的底层基石。

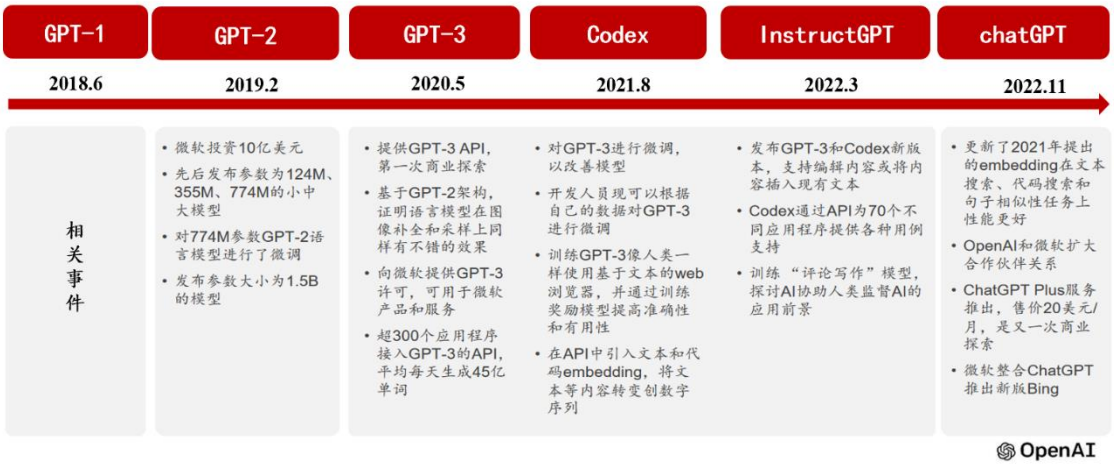
图表 64：AI 人工智能与半导体计算芯片发展历程



来源：《AI 芯片：前沿技术与创新未来》，中泰证券研究所

- **ChatGPT 爆火的背后是人工智能大模型的迭代升级。**ChatGPT 是基于 GPT-3.5 微调得到的新版本模型，能够借助人类反馈的强化学习（RLHF）技术来指导模型训练，实现模型输出与人类预期的需求，使对话内容更加人性化和富有逻辑性。从 2008 年第一代生成式预训练模型 GPT-1 诞生以来，GPT 系列模型几乎按照每年一代的速度进行迭代升级，未来随着大语言模型（LLM）技术的不断突破，AI 相关应用有望加速落地，AI 产业或将迎来新一轮发展机遇。
- **大模型发展将不断降低 AI 应用门槛，助力下游降本增效推动渗透率提升。**预训练大模型在海量数据的学习训练后具有良好的通用性和泛化性，用户基于大模型通过零样本、小样本学习即可获得领先的效果，能够显著降低 AI 应用的门槛，目前主流的大模型包括 Transformer、GAN、CNN 以及 RNN 等架构。

图表 65：GPT 模型迭代过程



来源：OpenAI 官网，World of Engineering，中泰证券研究所

- **大语言模型（LLM）代表着 AI 领域的重大进步，并有望通过习得的知识改变该领域。**根据英伟达，在过去几年中，LLM 的规模每年增加 10 倍，而且随着这些模型的复杂程度和规模的增加，其性能也在不断发展。大语言模型是一种基于深度学习的自然语言处理技术，其主要目的是通过学习大量文本数据从而自动生成符合语言规则的语句、段落甚至文章。基于 Transformer 架构，大语言模型的核心思想是利用深度神经网络来学习自然语言的语法、语义等特征，从而能够预测下一个词汇的出现概率，并根据这些概率生成新的语句。Transformer 架构在海量数据集上并行处理数据排序的计算能力是大语言模型背后的最大驱动力。
- **基于 Transformer 架构，GPT 采用预训练的方法来学习语言的概率分布模型，经过微调后可以解决各种自然语言处理任务、生成自然流畅的文本。**其工作原理如下：

1）数据收集和监督训练。模型用示例提示进行训练，由人类向模型演示所需的输出。然后由人类对模型进行监督和微调，直到它能够进行满足所需性能水平的输出。

2）可比数据集手机和模型的奖励训练。对于相同的示例，向模型演示多个输出并从最好到最差排序。已经通过监督训练的模型将生成尽可能接近预期结果的输出。用多组数据输出来训练模型以生成排名最高的输出。

3）使用强化学习来优化模型。当该模型产生一个排名最高的输出时，它会得到奖励以强化这种积极的结果。训练过的模型生成输出后，奖励模型计算奖励，如果新输出排名较高则模型策略会自动更新。

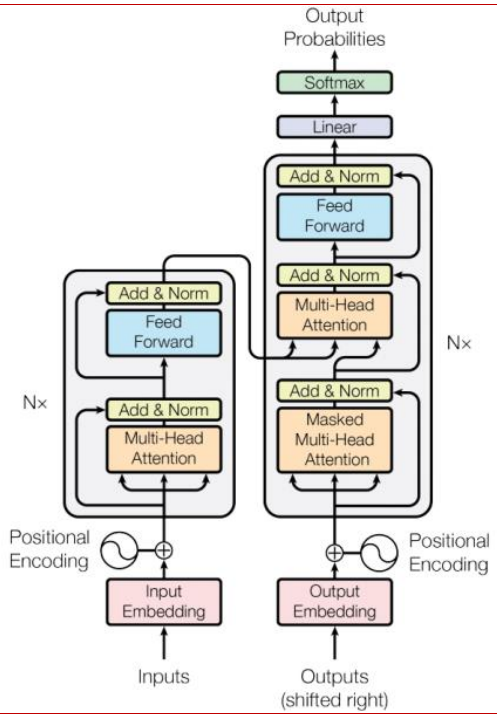
图表 66：大语言模型（LLM）举例

类型	描述	备注
GPT-3/4 (Generative Pretrained Transformer 3/4)	由 OpenAI 研发的 GPT-3/4，能够进行自然语言生成、文本分类和问答等任务	GPT3 参数达 1750 亿个，GPT4 参数达 3.5 万亿个。
BERT (Bidirectional Encoder Representation from Transformers)	由谷歌研发的双向 Transformer 的 Encoder，因为 Decoder 是不能获要预测的信息的	拥有 1.1 亿个参数
XLNet (eXtreme Language understanding NETwork)	由 CMU 和 Google 共同开发的预训练语言模型，能够进行自然语言处理任务，如文本分类、问答等	拥有 1.5 亿个参数
RoBERTa (Robustly Optimized BERT Pretraining Approach)	由 Facebook AI 研究院开发的预训练语言模型，能够进行自然语言处理任务如文本分类、问答等	拥有 1.25 亿个参数
ERNIE Bot（文心一言）	由百度开发的全新一代知识增强大语言模型	
T5 (Text-to-Text Transfer Transformer)	由谷歌开发的预训练语言模型，能够进行多种自然语言处理任务，如文本生成、问答、摘要等	拥有 11 亿个参数

来源：公开资料整理，中泰证券研究所

- **与传统人工智能不同，生成式 AI 具有创造性能力。**传统的人工智能模型建立在具有预测性的判别统计模型上，其侧重于从现有数据中识别模式。而生成式模型则可以基于一组底层数据输入来生成新的数据实例。生成式人工智能从底层数据集生成、创建新内容，以文本、图像、音频、视频、代码等形式生成原始想法，超越了传统的模式检测和扭曲数据分析。生成式 AI 能够打破人与机器之间的通信障碍，使得人类用自然语言而不是编程语言与计算机进行通信。
- **Transformer 架构已经成为神经网络学习中最重要的架构之一。**传统上，自然语言处理领域中使用的大多数模型都基于循环神经网络（RNN），这些模型存在计算复杂度高、难以并行计算等众多局限性。而 Transformer 架构基于注意力机制，比传统的 RNN 和 CNN 更快、更稳定，并且具有更高的准确率，更容易并行化。该技术大大减少了训练模型的时间和对结构化数据集的依赖，提高了人工智能的自主学习能力。

图表 67：Transformer 架构示意图



来源：壁仞科技研究院，中泰证券研究所

- **生成式 AI 主要依赖于人工智能大模型，具有参数多、包含数据量大等特点。**这些模型通常包含数十亿至数万亿个参数，需要庞大的数据集进行训练，根据《AIGC 发展报告 2023》数据，国外主要 AIGC 预训练模型参数规模在 6.4 亿至 5400 亿之间，平均参数量高达 1541 亿。未来大模型的训练数据不仅限于文字，还可以包括图像、视频等多种形式。与自然语言处理模型相比,多模态模型训练数据为图像、视频等，规模远大于语言类模型，因此需要更多的计算资源和算力来支持模型的训练和推理。

图表 68：国外部分 AIGC 预训练模型一览

厂商	预训练模型	应用	参数量	领域
谷歌	BERT	语言理解与生成	4810亿	NLP
	PaLM	语言理解与生成、推理、代码生成	5400亿	NLP
	Imagen	语言理解与图像生成	110亿	多模态
	Parti	语言理解与图像生成	200亿	多模态
微软	Florence	视觉识别	6.4亿	CV
	Turing-NLG	语言理解、生成	170亿	NLP
Deep Mind	Gato	多面手的智能体	12亿	多模态
	Gopher	语言理解与生成	2800亿	NLP
	AlphaCode	代码生成	414亿	NLP
Open AI	GPT3	语言理解与生成、推理等	1750亿	NLP
	CLIP&DALL-E	图像生成、跨模态检索	120亿	多模态
	Codex	代码生成	120亿	NLP
	ChatGPT	语言理解与生成、推理等	13-1750亿	NLP

来源：腾讯《AIGC 发展报告 2023》，中泰证券研究所

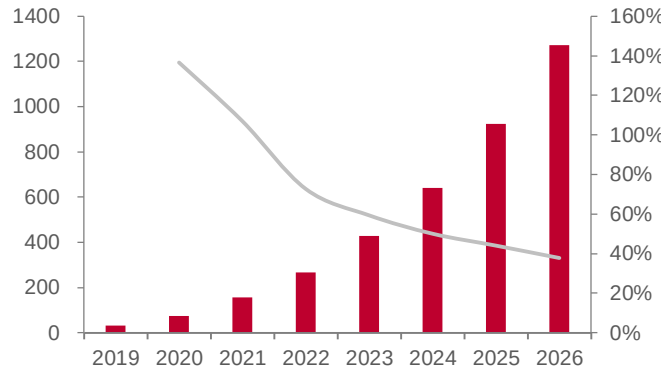
4.2 算力芯片快速增长，GPU 占据 AI 芯片主流地位

- **算力需求激增，AI 应用发展仍需跨越当前算力鸿沟。**根据 OpenAI 数据，

2012-2018 年期间，人工智能训练任务中使用的算力正呈指数级增长，速度为每3.5个月翻一倍，人们对于算力的需求增长了超过300,000倍。相比之下，摩尔定律是每 18 个月翻倍，如果是以摩尔定律的速度，这期间只会有 12 倍的增长。根据 IDC 数据，中国 AI 算力规模将保持高速增长，预计到 2026 年将达 1271.4EFLOPS，CAGRA（2022-2026 年）达 52.3%。

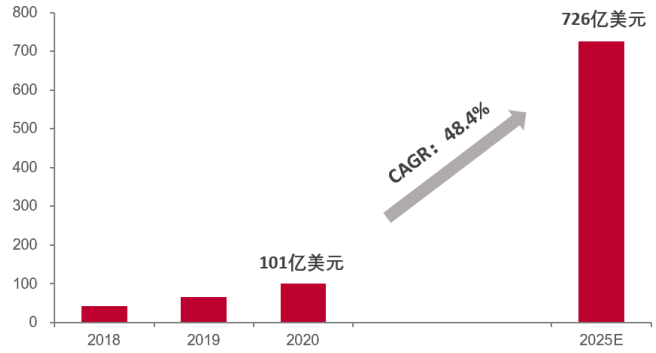
- **算力需求的快速增长与芯片计算能力的增长形成剪刀差，推动 AI 芯片市场规模不断发展。**当前模型计算量的增长远超人工智能硬件算力的增长，二者性能增长之间的不匹配，剪刀差的扩大将带来对算力基础设施供给需求的不断增长，以及算力硬件供给需求的快速增长。根据 Gartner 数据，2025 年人工智能芯片市场规模将从 2020 年的 101 亿美元增长至 726 亿美元，CAGR（2020-2025）为 48.4%。

图表 69：中国 AI 算力规模（百亿亿次浮点运算/秒）



来源：《2022-2023 中国人工智能计算力发展评估报告》，中泰证券研究所

图表 70：全球 AI 芯片市场规模及预测



来源：IDC，Gartner，OpenAI，中泰证券研究所

- **AI 芯片是 AI 算力的核心，需求有望率先扩张。AI 芯片是用于加速人工智能训练和推理任务的专用硬件，主要包括 GPU、FPGA、ASIC 等，具有高度并行性和能够实现低功耗高效计算的特点。**CPU 是 AI 计算的基础，负责控制和协调所有的计算操作。在 AI 计算过程中，CPU 用于读取和准备数据，并将数据来传输到 GPU 等协处理器进行计算，最后输出计算结果，是整个计算过程的控制核心。根据 IDC 数据，CPU 在基础型、高性能型、推理型、训练型服务器中成本占比分别为 32%、23.3%、25%、9.8%，是各类服务器处理计算任务的基础硬件。
- **GPU、FPGA、ASIC 是 AI 计算的核心，作为加速芯片处理大规模并行计算。**具体来看，GPU 通用性较强，适合大规模并行计算，且设计及制造工艺较成熟，目前占据 AI 芯片市场的主要份额；FPGA 具有开发周期短、上市速度快、可配置性等特点，目前被大量应用于线上数据处理中心和军工单位；ASIC 根据特定需求进行设计，在性能、能效、成本均极大的超越了标准芯片，非常适合 AI 计算场景，是当前大部分 AI 初创公司开发的目标产品。

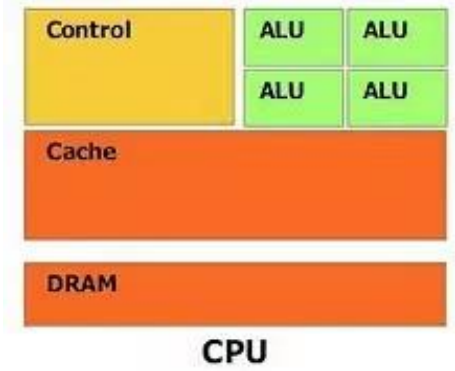
图表 71：AI 芯片特点及具体参数对比

特点	CPU	GPU	FPGA	ASIC
基本架构	60%逻辑单元 40%计算单元	60%-70%计算单元 30%逻辑控制单元	门电路资源	固化的门电路资源
架构图				
定制化程度	通用型	通用型	半定制化	定制化
延迟	高	较高	低（约为GPU的1/10）	低（约为GPU的1/10）
优势	复杂逻辑运算能力强，擅长逻辑控制	擅长并行计算，浮点数据计算能力强，软硬件体系一致	可进行数据并行和流水线并行，可编程，灵活度高	AI运算效率高，功耗低，体积小
劣势	核数少，不擅长处理并行任务	面积大，功耗高，由于通用性要求难以专一面对某一模型深度优化	开发周期长，复杂算法开发难度大	灵活性差，算法支持有限，算法迭代后需重新开发
AI训练效果	效果较差	唯一量产可用于训练的硬件	效率不高	可能是用于训练的最佳芯片，但目前没有量产产品
应用场景	主要用于推断场景	在云端和边缘端均占据主导地位，云端训练份额最高	主要用于推断场景	主要应用于推断场景
具体芯片对比	**E5-2699 V3**	**Tesla K80**	**Virtex7-690T**	**Google TPU**
计算单元个数（个）	18（256bit）	7804（32bit）	3600（32bit）	65536（8bit）
峰值运算能力（TOPS）	1.33（单精度浮点）	8.74（单精度浮点）	1.8（单精度浮点）	92（8bit整点）
功耗（W）	145	300	30	40
能耗比（GFLOPS/W）	9	29	60	2300

来源：中泰证券研究所

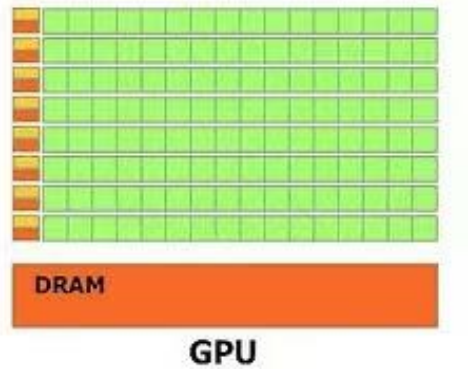
- **相比于少核心串行结构的 CPU，多核心的并行结构 GPU 更适合处理图形图像（矩阵结构）信息。**CPU 通常有 4 个、8 个或 16 个强力 ALU 核心（arithmetic logic unit，算术逻辑单元），适合做复杂的通用串行任务。GPU 是图形计算的重要元件，主要用来处理与图形图像相关的数据。与 CPU 不同的是，GPU 有数百甚至数千个简单 ALU 核心，单个 ALU 处理能力相比 CPU 的更弱，但能够实现多个 ALU 并行计算，适合做简单特定的并行任务。因此，对于复杂的单个计算任务来说，CPU 的执行效率更高，通用性更强；而对于图形图像这种矩阵式多像素点的简单计算，更适合用 GPU 来处理，但通用性较弱。

图表 72：CPU 的基本结构



来源：Imagination，中泰证券研究所

图表 73：GPU 的基本结构



来源：Imagination，中泰证券研究所

- **GPU 的并行计算能力强于 CPU。**并行计算一般分为两种类型：一是基于任务的并行处理，这种模式将计算任务拆分成若干个小的但不同的任务，不同任务相接组成一道流水线，由此完成整体的任务要求；二是基于数据的并行处理，这种模式将数据分解为多个部分，让多个运算单元分别去计算小块的数据，最后再将其汇总起来满足任务要求。一般而言，

CPU 的多线程编程偏向于第一种并行模式，GPU 并行编程模式则偏向于第二种，其对于数据的计算能力更加强大。

- **具有并行计算架构的 GPU 是 AI 算力的重要支撑，相较 CPU 在 AI 研究和开发中具有更高的效率。**21 世纪初期，研究人员意识到，由于机器学习算法通常具有与图形处理算法相同类型的计算，因此 GPU 可以为机器学习提供基于 CPU 计算的更有效的替代方案。GPU 能够提供卓越的并行性能，由此可以有效加速 AI 计算，满足不断发展的算力需求。GPU 提供的计算效率不仅仅能够简化了分析过程，还能促进更广泛的模型训练以获得更高的准确性，扩大了模型搜索过程的范围以防止替代规范，使以前无法实现的某些模型变得可行，并允许对替代数据集增加额外的敏感性以确保其稳健性。

图表 74：CPU 与 GPU 对比

	CPU	GPU	原因
设计目标	通用性高，需要处理不同数据类型	主要用于处理计算密集型的程序和易于并行的程序并加快图像处理速度	
核数	个位数	众核，个数可达三位数	
控制器	复杂	简单	CPU 面对分支程序时，通过预测分支结果来降低延时
ALU	较少	较多	CPU 需要在短时间内完成复杂计算，时钟频率高达达到 1.532~3 千兆赫
Cache	较多，并分级	较少	CPU 利用大缓存来降低延迟
寄存器	较少	较多	GPU 依靠较多的寄存器来支持多线程
线程数	较少	较多	GPU 不依靠多级 Cache 来降低访问内存的延迟，利用多线程来应对大吞吐量
功耗	一般	较高	

来源：英伟达，公开资料整理，中泰证券研究所

- **GPU 在 AI 研究和开发中的重要性不断增加。**AI 芯片中，GPU 占据主要市场规模。根据 IDC 数据，2022 年国内人工智能芯片市场中，GPU 芯片所占市场份额达 89.0%。GPU 作为市场上 AI 计算最成熟、应用最广泛的通用型芯片，应用潜力较大，其并行计算架构相较于其他 AI 芯片更加适合于复杂数学计算场景，支持高度并行的工作负载。
- **英伟达是全球领先的 GPU 和 AI 芯片制造商之一。**在国内 GPU 市场，英伟达占据了主要份额。IDC 数据显示，2020 年英伟达在国内 GPU 服务器市场几乎占据 95%左右的市场份额。通过研究英伟达的发展路径和战略，能够帮助国内企业更好地了解 GPU 的应用和未来趋势，为国内企业提供宝贵的借鉴和启示。

4.3 AI 芯片领域，国产芯片迅速崛起

- **全球 GPU 芯片市场主要由海外厂商占据垄断地位，国产厂商加速布局。**全球 GPU 市场被英伟达、英特尔和 AMD 三强垄断，英伟达凭借其自身 CUDA 生态在 AI 及高性能计算占据绝对主导地位；国内市场中，景嘉微在图形渲染 GPU 领域持续深耕，另外天数智芯、壁仞科技、登临科技等一批主打 AI 及高性能计算的 GPGPU 初创企业正加速涌入。
- **图形渲染 GPU：目前国内厂商在图形渲染 GPU 方面与国外龙头厂商差**

距不断缩小。芯动科技的“风华 2 号”GPU 采用 5nm 工艺制程，与 Nvidia 最新一代产品 RTX40 系列持平，实现国产图形渲染 GPU 破局。景嘉微在工艺制程、核心频率、浮点性能等方面虽落后于 Nvidia 同代产品，但差距正逐渐缩小。

图表 75：国内外 AI 芯片产品对比（1）——图形渲染 GPU

产品类型	厂商	产品型号	发布年份	制作工艺	显存类型	显存位宽	显存容量	显存带宽	核心频率	像素填充速率	浮点性能（FP32）	总线接口
图形渲染GPU	Nvidia	RTX4090	2022	5nm	GDDR6X	384bit	24GB	——	2230MHz	——	——	PCIe 4.0 x 16
		GTX1050	2016	14nm	GDDR5	128bit	2GB	112GB/s	1354MHz	43.3G Pixel/s	1.8TFLops	PCIe 3.0 x 16
	芯动科技	风华1号	2021	12nm	GDDR6X	128bit	16GB	304GB/s	——	160G Pixels/s	5TFLops	PCIe 4.0 x 8
		风华2号	2022	5nm	LPDDR5X	——	8GB	102.4GB/s	——	48G Pixels/s	1.5TFLops	PCIe 3.0 x 8
	景嘉微	JM7200	2018	28nm	DDR3	64bit	4GB	17GB/s	1300MHz	5.2G Pixels/s	0.5TFLops	PCIe2.0 x16
		JM9231	2021	14nm	——	——	8GB	256GB/s	1500MHz	32G Pixels/s	2TFLops	PCIe 3.0 x 16

来源：各公司官网，中泰证券研究所

- **在 GPGPU 方面，目前国内厂商与 Nvidia 在 GPGPU 上仍存在较大差距。**制程方面，目前 Nvidia 已率先到达 4nm，国内厂商多集中在 7nm；算力方面，国内厂商大多不支持双精度（FP64）计算，在单精度（FP32）及定点计算（INT8）方面与国外中端产品持平，天数智芯、壁仞科技的 AI 芯片产品在单精度性能上超过 NVIDIA A100；接口方面，壁仞科技与 Nvidia 率先使用 PCIe5.0，其余厂商多集中在 PCIe4.0；生态方面，国内企业多采用 OpenCL 进行自主生态建设，与 Nvidia CUDA 的成熟生态相比，差距较为明显。

图表 76：国内外 AI 芯片产品对比（2）——GPGPU

产品类型	厂商	产品型号	发布时间	工艺制程	浮点算力-Tflops			INT8定点算力（TOPS）	互联带宽	显存	接口	功耗	生态
					FP64	FP32	BF16						
GPGPU	NVIDIA	H100 SXM5	2022	4nm Hopper	30	500	1000	2000	900GB/s	80GB	SXM5	700W	CUDA
		H100 PCIe	2022	4nm Hopper	24	48	800	1600	900GB/s	80GB	PCIe5.0	350W	CUDA
		A100 PCIe	2020	7nm Ampere	9.7	19.5	312	624	600GB/s	80GB	PCIe4.0	400W	CUDA
		Tesla V100	2017	12nm Volta	7.8	15.7	125	62	150GB/s	32GB	PCIe4.0	300W	CUDA
	AMD	Instinct MI250X	2021	6nm CNDA 2	47.9	47.9	383	383	——	128GB	PCIe 4.0	560W	AMD ROCm
		Instinct MI250	2021	6nm CNDA 2	47.9	45.3	362	362	——	128GB	PCIe 4.0	560W	AMD ROCm
		Instinct MI100	2020	7nm CNDA 1	11.5	23.1	92.3	184.6	——	32GB	PCIe 4.0	350W	AMD ROCm
	天数智芯	天垓100	2021	7nm	×	37	147	295	64GB/s	32GB	PCIe 4.0	250W	SIMT
	壁仞科技	壁砺 100P	2022	7nm	×	240	960	1920	448 GB/s	64GB	PCIe 5.0	550W	BIRENSUPA
		壁砺104P	2022	7nm	×	112	448	896	192GB/s	32GB	PCIe5.0	300W	BIRENSUPA
	登临科技	Goldwasser-L	2020	12nm	×	×	×	512	——	64GB	PCIe 3.0	120W	——

来源：各公司官网，中泰证券研究所

- **FPGA 全球市场呈现“两大两小”格局，Altera 与 Xilinx 市占率共计超 80%，Lattice 和 Microsemi 市占率共计超 10%；**整体来看，安路科技、紫光同创等厂商处于国际中端水平，仍需进一步突破。工艺制程方面，当前国产厂商先进制程集中在 28nm，落后于国际 16nm 水平；在等效 LUT 数量上，国产厂商旗舰产品处于 200K 水平，仅为 XILINX 高端产

品的 25%左右。

- **ASIC 不同于 CPU、GPU、FPGA，目前全球 ASIC 市场并未形成明显的头部厂商，国产厂商快速发展；**通过产品对比发现，目前国产厂商集中采用 7nm 工艺制程，与国外 ASIC 厂商相同；算力方面，海思的昇腾 910 在 BF16 浮点算力和 INT8 定点算力方面超越 Googel 最新一代产品 TPUv4，遂原科技和寒武纪的产品在整体性能上也与 Googel 比肩。未来国产厂商有望在 ASIC 领域继续保持技术优势，突破国外厂商在 AI 芯片的垄断格局。

图表 77：国内外 AI 芯片产品对比（3）——FPGA/ASIC

产品类型	厂商	产品型号	发布时间	工艺制程	浮点算力（Tflops）			INT8定点算力（TOPS）	互联带宽	显存	接口	功耗	生态
					FP64	FP32	BF16						
ASIC	Googel	TPUv4i	2020	7nm	×	×	138	138	400GB/s	8GB	——	——	TensorFlow XLA
		TPUv4	2021	7nm	×	×	275	275	1000GB/s	32GB	——	——	TensorFlow XLA
	海思	昇腾910	2018	7nm	×	×	320	640	——	——	PCIe 4.0	350W	MindSpore
	遂原科技	T20(32GB)	2021	12nm	×	32	128	256	300 GB/s	32GB	PCIe 4.0	300W	——
	寒武纪	MLU370-X4	2021	7nm	×	24	96	256	200GB/s	24GB	PCIe 4.0	150W	Cambricon Neuware
		MLU370-S4	2021	7nm	×	18	72	192	200GB/s	24GB	PCIe 4.0	75W	Cambricon Neuware

产品类型	厂商名称	产品型号	工艺制程	系统逻辑单元	等效LUT数量	分布式RAM	DSP数量	User IO
FPGA	XILINX	Artix UltraScale+ AU25P	16nm	308K	141K	4.7Mb	1200	304
		Kintex UltraScale+ XCKU19P	16nm	1843K	842K	11.6Mb	1080	540
	紫光同创	PG2L200H	28nm	——	239.7K	2.528Mb	——	500
		PG2T160H	——	160K	——	2.188Mb	——	400
	安路科技	PH1A180SFG676	28nm	——	210.24K	3.277MB	600	396
		EF3LA0CG642	55nm	——	11.776K	94K	——	475

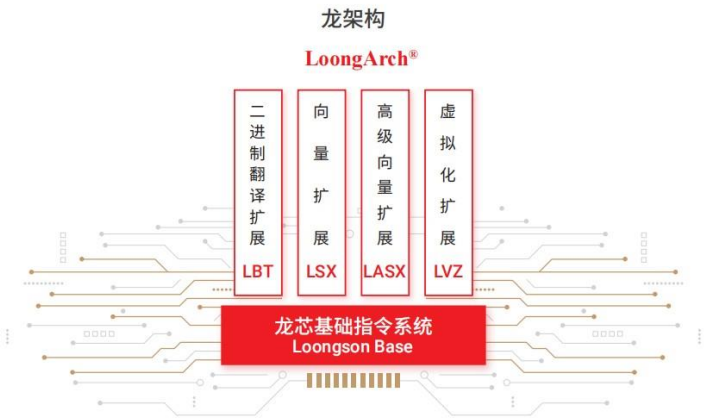
来源：各公司官网，中泰证券研究所

4.4 国产算力公司梳理

龙芯中科：国产 CPU 设计标杆，自主研发 GPGPU

- **公司主要从事处理器（CPU）及配套芯片的研制、销售及服务。**主要产品包括龙芯 1 号、龙芯 2 号、龙芯 3 号三大系列处理器芯片及桥片等配套芯片，系列产品在电子政务、能源、交通、金融、电信、教育等行业领域已获得广泛运用。
- **龙芯成功构建独立信息技术体系的 CPU，不断推出基于 LoongArch 架构的芯片。**龙芯基于自主指令系统，决心构建独立于 Wintel 和 AA 体系的开放信息技术体系的 CPU。龙芯技术上的持续积累使其成功地建立了自己的指令系统架构 LoongArch。在 2021 年和 2022 年，公司相继推出了多款基于 LA 架构的芯片产品，目前已经拥有 9 颗基于 LA 架构的芯片产品。

图表 78：龙架构 LoongArch



来源：龙芯科技官网，中泰证券研究所

- **坚持自主研发指令系统、IP 核等核心技术。**龙芯中科掌握指令系统、处理器核微结构、GPU 以及各种接口 IP 等芯片核心技术，在关键技术上进行自主研发，拥有大量的自主知识产权，已取得专利 400 余项。
- **GPU 产品进展顺利，正研制新一代图形及计算加速 GPGPU 核。**公司在 2022 年上半年完成了第一代龙芯图形处理器架构 LG100 系列，目前正在启动第二代龙芯图形处理器架构 LG200 系列图形处理器核的研制。根据公司在2022年半年度业绩交流会信息，第一代GPU核 (LG100) 已经集成在 7A2000 中，新一代 GPGPU 核(LG200) 的研制也取得了积极进展。

海光信息：国产高端处理器龙头，CPU+DCU 双轮驱动

- **公司主营产品包括海光通用处理器（CPU）和海光协处理器（DCU）。**海光 CPU 主要面向复杂逻辑计算、多任务调度等通用处理器应用场景需求，兼容国际主流 x86 处理器架构和技术路线。从应用场景看，海光 CPU 分为 7000、5000、3000 三个系列，分别定位于高端服务器、中低端服务器和边缘计算服务器。海光 DCU 是公司基于 GPGPU 架构设计的一款协处理器，目前以 8000 系列为主，面向服务器集群或数据中心。海光 DCU 全面兼容 ROCm GPU 计算生态，能够较好地适配国际主流商业计算软件，解决了产品推广过程中的软件生态兼容性问题。
- **CPU 与 DPU 持续迭代，性能比肩国际主流厂商。**CPU 方面，目前海光一号和海光二号已经实现量产，海光三号已经正式发布，海光四号目前进入研发阶段。海光 CPU 的性能在国内处于领先地位，但与国际厂商在高端产品性能上有所差距，接近 Intel 中端产品水平；DCU 方面，深算一号已实现商业化应用，深算二号已于 2020 年 1 月启动研发。在典型应用场景下，公司深算一号指标达到国际上同类型高端产品的水平。
- **高研发力度成为海光产品快速迭代的基石。**从 2019 到 2021 年，海光信息的研发投入从 8.65 亿元增至 15.85 亿元，增长 83.3%。拥有千人级高端处理器研发团队，且 90.2%的员工是研发人员。公司已取得多项处理

器核心技术突破，并拥有 179 项专利、154 项软件著作权和 81 项集成电路布图设计专有权，构建了全面的知识产权布局。

图表 79：海光 CPU 与 Intel 产品性能对比

产品名称		Intel8380HL	Intel8380H	Intel8376HL	海光7285	Intel8360HL	Intel8360H
4路测试结果	Speccpu INT	784	784	765	-	690	688
	Speccpu FP	657	653	641	-	599	597
双路测试结果	Speccpu INT	392	392	383	348	345	344
	Speccpu FP	329	327	321	308	300	299
性能差异(Intel数据/海光数据-1)	Speccpu INT	12.64%	12.64%	9.91%	-	-0.86%	-1.15%
	Speccpu FP	6.66%	6.01%	4.06%	-	-2.76%	-3.08%

来源：海光招股说明书，中泰证券研究所

图表 80：深算一号与国际同类型产品性能对比

项目	海光	NVIDIA	AMD
产品	深算一号	Ampere 100	MI100
生产工艺	7nm FinFET	7nm FinFET	7nm FinFET
核心数量	4096 (64 CUs)	2560 CUDA processors 640 Tensor processors	120CUs
内核频率	Up to 1.5GHz（FP64） Up to 1.7Ghz（FP32）	Up to 1.53Ghz	Up to 1.5GHz（FP64） Up to 1.7Ghz（FP32）
显存容量	32GB HBM2	80GB HBM2e	32GB HBM2
显存频率	2.0 GHz	3.2 GHz	2.4 GHz
TDP	350W	400W	300W
CPU to GPU 互联	PCIe Gen4 x 16	PCIe Gen4 x 16	PCIe GEN4 x 16
GPU to GPU 互联	xGMI x 2, Up to 184 GB/s	NVLink up to 600 GB/s	Infinity Fabric x 3，up to 276 GB/s

来源：海光招股说明书，中泰证券研究所

寒武纪：国产 AI 芯片领先者

- **寒武纪是 AI 芯片领域的独角兽。**公司成立于 2016 年 3 月 15 日，专注于人工智能芯片产品的研发与技术创新，产品广泛应用于消费电子、数据中心、云计算等诸多场景。公司是 AI 芯片领域的独角兽：采用公司终端智能处理器 IP 的终端设备已出货过亿台；云端智能芯片及加速卡也已应用到国内主流服务器厂商的产品中，并已实现量产出货；边缘智能芯片及加速卡的发布标志着公司已形成全面覆盖云端、边缘端和终端场景的系列化智能芯片产品布局。
- **人工智能的各类应用场景，从云端溢出到边缘端，或下沉到终端，都离不开智能芯片的高效支撑。**公司面向云端、边缘端、终端推出了三个系列不同品类的通用型智能芯片与处理器产品，分别为终端智能处理器 IP、云端智能芯片及加速卡、边缘智能芯片及加速卡。
- **寒武纪产品线丰富，应用场景广泛。**公司拥有一套全面的产品线，包括已经发布的面向云端和边缘端的智能芯片、加速卡、训练机、处理器 IP 以及软件，能够满足在云、边、端各个尺度的人工智能计算需求。在 2022 年的 3 月份，公司推出了新的训练加速卡 MLU370-X8。这款加速卡配备了双芯片四核思元 370，并整合了寒武纪 MLU-Link 多核互联技术，主要针对训练任务。在广泛应用于各个领域的 YOLOv3、Transformer 等训练任务中，8 卡计算系统的并行性能平均超过了 350WRTXGPU 的 155%。

图表 81：高性能通用图形处理器芯片及系统研发项目情况及进程安排

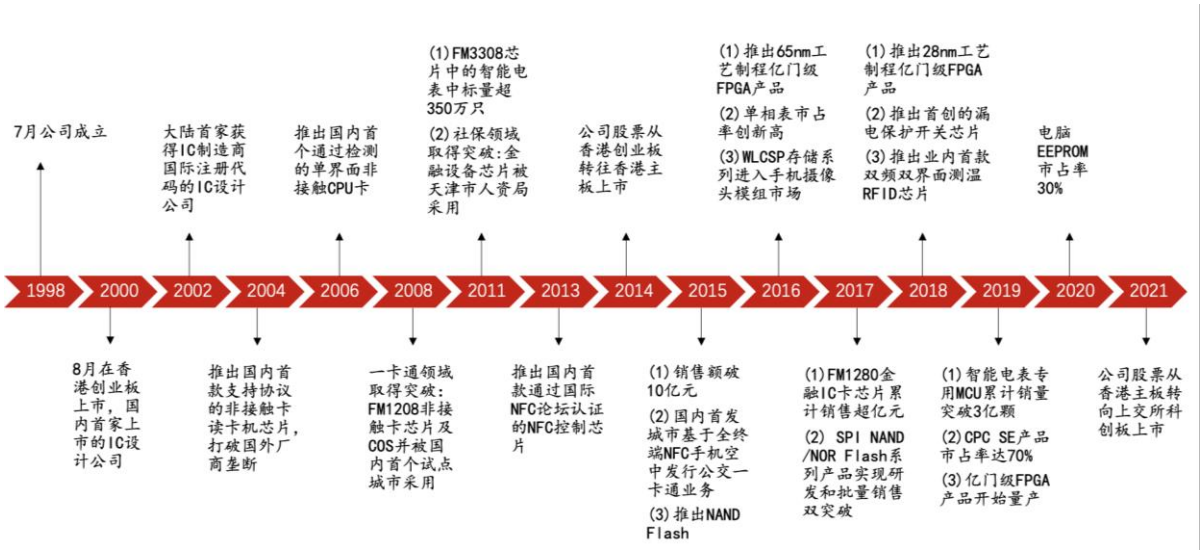
应用场景	芯片需求	典型计算能力	典型功耗	典型应用领域
终端	低功耗、高能效、推理任务为主、成本敏感、硬件产品形态众多	<8TOPS	<5瓦	各类消费类电子、物联网产品等
云端	高性能、高计算密度、兼有推理和训练任务、单价高、硬件产品形态少	>30TOPS	>50瓦	云计算数据中心、企业私有云等
边缘端	对功耗、性能、尺寸的要求常介于终端与云端之间、推理任务为主、多用于插电设备、硬件产品形态相对较少	5TOPS 至 30TOPS	4瓦至15瓦	智能制造、智能家居、智能零售、智慧交通、智慧金融、智慧医疗、智能驾驶等众多应用领域

来源：龙芯中科股说明书，中泰证券研究所

复旦微电：国内 FPGA 领军企业，多元化产品打开增长空间

- **24 年潜心钻研，高端芯片国产化的开创性先锋。**公司成立两年后就成为国内首家上市的 IC 设计公司，二十余年发展过程中又陆续成为首家获得 IC 制造商的国际注册代码，推出国内首个非接触卡等产品，打破国外厂商垄断并在诸多领域取得突破。目前，复旦微的 RFID 芯片、智能卡芯片、EEPROM、智能电表 MCU 等多类产品市占率都位列前茅。此外，复旦微推出的亿门级 FPGA 产品已实现供货，填补了国产高端 FPGA 的空白，具有取代进口 FPGA 产品的巨大潜力。

图表 82：公司发展历程时间表



来源：公司官网，公司招股书，中泰证券研究所整理

- **复旦微电专注于 FPGA 高端产品的开发，有望引领高端 FPGA 国产替代。**FPGA 在 5G 通信、人工智能以及特种集成电路等多个领域拥有巨大发展潜力。作为国内最早推出亿门级 FPGA 产品的企业，复旦微电子在国内 FPGA 芯片设计行业占据了领先地位。至 2022 年半年报，公司已向超过 500 家客户销售相关 FPGA 产品，同时，10 亿门级产品的研

发正在进行中，预计在未来将引领高端 FPGA 的国产替代进程。

- **复旦微电产品应用广泛，获得业内认可。**作为国内顶尖的集成电路设计企业，公司以二十多年的深厚积淀构建了多元化的产品线。其 RFID 芯片、智能卡芯片、EEPROM 以及智能电表 MCU 等多种产品在市场上享有领先地位，获得了三星、LG、VIVO 等海内外知名厂商的高度认可。
- **FPGA 芯片国内技术领军者，国产替代种子选手。**复旦微在国内 FPGA 领域处于领先地位，目前可提供千万门级、亿门级 FPGA 芯片以及嵌入式可编程芯片等系列产品。研发方面，当前一方面基于 14/16nm 工艺制程开发 10 亿门级产品，另一方面丰富 28nm 制程的 FPGA 及 PSoC 芯片种类，继续保持在国产 FPGA 技术的领先地位。

图表 83：FPGA 芯片产品线

产品线	产品介绍	应用	产品或终端图
千万门级 FPGA	采用 65nm CMOS 工艺，是一系列高性能、高性价比 SRAM 型 FPGA 产品	适用于网络通信、信息安全、工业控制、高可靠等高性能、大规模应用	
亿门级 FPGA 芯片	采用 28nm CMOS 工艺，是一系列高性能、大规模的 SRAM 型 FPGA 产品	适用于 5G 通信、人工智能、数据中心、高可靠等高性能、大带宽、超大规模应用	
嵌入式可编程器件 PSoC	采用 28nm CMOS 工艺，是一系列嵌入式可编程片上系统产品	适用于视频、工控、安全、AI、高可靠等应用	

来源：公司招股书、中泰证券研究所整理

安路科技：民用 FPGA 领先厂商，国产替代正当时

- **安路科技是国内领先的 FPGA 芯片设计企业。**安路科技成立于 2011 年，自成立至今，公司一直专注于 FPGA 芯片设计领域，通过多年的技术累积，公司在 FPGA 芯片设计技术、SoC 系统集成技术、FPGA 专用 EDA 软件技术、FPGA 芯片测试技术和 FPGA 应用解决方案等领域均有技术突破。公司主要专注于 FPGA 芯片和专用 EDA 软件的研发、设计和销售。产品的主要下游应用领域主要包括工业控制、网络通信、消费电子和数据中心等。

图 84：公司发展历程

来源：公司公告，中泰证券研究所整理

- **公司主要向客户提供 FPGA 产品，包括 FPGA 芯片和专用 EDA 软件两部分。**基于目前的核心技术体系，公司成功构建了由 ELF 系列、EAGLE 系列和 PHOENIX 系列 FPGA 芯片和 TangDynasty 系列专用 EDA 软件组成的产品矩阵，2021 年，公司 FPSoC 产品新增了面向工业和视频接口的低功耗 SWIFT 系列。公司产品覆盖 28nm-55nm 的工艺制程，形成了多种逻辑规模 FPGA 芯片和软件的全产品线覆盖，并持续致力于高容量、高性能的 FPGA 和 FPSoC 芯片的研发与拓展。公司目前已成为国内领先的 FPGA 芯片供应商，产品已广泛应用于工业控制、网络通信、数据中心、消费电子等产业中。

图 85：公司 FPGA 芯片产品线

系列名称	细分系列	量产时间	制程	逻辑容量	产品介绍	应用领域	产品图片
ELF 系列	ELF1 系列	2015 年	130nm	336~688	ELF1 系列 FPGA 定位低成本、低功耗可编程市场。快速上电启动、支持单电源供电、无需要外部配置器件等特性使得 ELF1 器件非常适用于功能扩展、电源管理等应用场景。	消费电子、工业控制	
	ELF2 系列	2018 年	55nm	1.5K~4.5K	ELF2 系列 FPGA 是 ELF 的第二代产品，定位低功耗可编程市场。无需外部配置器件、低密度逻辑容量、丰富的存储器、高达 1Gbps 的 IO 速率等特性，使得 ELF2 器件非常适用于高速接口扩展与转换、高速总线扩展、高速存储器控制等应用场景。 ELF2 系列中包括一款嵌入 CPU 核的 FPSoC 芯片，已在多家客户获得应用。	消费电子、网络通信	 
	ELF3 系列	2019 年	55nm	1.5K~9.2K	ELF3 系列 FPGA 是 ELF 的第三代产品，定位工业控制、网络通信、数据中心等功能扩展应用市场，最多支持 336 个 IO，满足客户板级功能扩展多样性应用需求。ELF3 器件经过功耗与性能优化，使系统设计师在降低成本和功耗的同时又可满足不断增长的带宽要求。	工业控制、网络通信、数据中心	

EAGLE 系列	AL3 系列	2014 年	65nm	5.8K~11.1K	AL3 系列 FPGA 定位高性价比的逻辑 控制市场。AL3 器件系列建立在一个 优化的工艺基础之上，并通过较低的成本实现较高的功能性，具有合适的 逻辑规模，丰富的存储资源。	工业控制	
	EAGLE4 系列	2016 年	55nm	20K	EAGLE4 系列是 AL3 的升级产品，定位在高性价比逻辑控制和图像处理市场，数量适中的逻辑和乘法器，丰富多样的片内存储器，高达 1Gbps 的 IO 速率，使得 EAGLE4 器件非常适 合于图像预处理，伺服控制和高速图像接口转换等领域。	工业控制、网络通信、数据中心	
FPSoC	EF2M45				EF2M45 是嵌入 ARM 处理器核的 FPSoC 芯片，单颗芯片实现灵活的硬件可编程系统控制功能，已在多家客户获得应用。	工业控制、消费电子	
	SWIFT1 系列				SWIFT1 系列是全新低功耗 FPSoC 产品，芯片集成了逻辑单元、存储单元、视频处理单元、RISC-V 处理器核等资源，定位高带宽的视频数据处理和桥接可编程系统芯片市场，在保持低功耗的前提下，提供高达 17.6Gbps 带宽的 MIPI 数据收发能力。	消费电子、工业控制	
PHOENIX 系列	PHOENIX 系列	2020 年	28nm	127K	PHOENIX1 系列 FPGA 采用 28nm 工艺，产品架构支持 100K~600K 等效逻辑单元、高速运算单元、丰富的存储资源和高达 16Gbps 的 SerDes 接口资源，定位高性能可编程逻辑市场。 针对高带宽应用场景，PHOENIX1 能够提供良好的信号处理和数据的传输功能。PHOENIX1 能够满足工业控制、网络通信、数据中心等市场需求。	工业控制、网络通信、数据中心	
TangDynasty	TangDynasty	2014 年	-	-	TangDynasty 软件为公司所有 FPGA 芯片产品系列提供简洁高效的应用 设计开发环境。该软件会根据 ELF 系列、EAGLE 系列、PHOENIX 芯片系列需要进行算法升级和迭代。	专用 EDA 软件	

来源：招股说明书，中泰证券研究所整理

■ **布局卡位打造软硬件生态体系行业壁垒较高，公司未来成长逻辑清晰。**公司立足于中低端 CPLD 产品起步，产品从几十、几百 K 的高性价比产品到目前 400K 的中高端产品全覆盖，客户积累深厚。公司采用软硬件协同模式，软件配套构建良好生态，其自主开发的 FPGA 专用 EDA 软件拥有较高技术水平，是国内目前拥有最多客户的国产 EDA 厂商，可以立足于广大的客户群体，不断反馈完善自身软件和配套的生态体系，打造自身软硬件护城河。正是因为 FPGA 芯片行业需要厂商同时具备较高的硬件芯片设计能力以及软件开发能力，行业进入壁垒较高。公司作为目前国产 FPGA 芯片行业的领先厂商，立足 FPGA 行业快速增长，拥有广阔国产替代空间，提前布局卡位未来竞争优势明显，稀缺性成长性兼备，伴随中高端产品放量，未来成长逻辑清晰。

5、投资建议及风险提示

- **持续看好 AIGC 发展下算力+应用两大方向投资机遇，建议关注：**
  （1）AI 算力芯片：寒武纪、海光信息、景嘉微；
  （2）服务器产业链：工业富联、沪电股份、奥士康；
  （3）AI 应用：大华股份、海康威视；
  （4）Chiplet：通富微电、长电科技、华海清科、长川科技、兴森科技。
  （5）C 端 AI 应用：国光电器、漫步者；瑞芯微、晶晨股份、乐鑫科技、中科蓝讯；

- **风险提示：**
- **行业需求不及预期的风险：**若包括手机、PC、可穿戴等终端产品需求回暖不及预期，则产业链相关公司的业绩增长可能不及预期。
- **产能瓶颈的束缚：**2021 年缺芯潮带来产业链公司业绩快速增长，产能成关键限制因素，若包括代工厂、封测厂等产能扩张进度不及预期，则可能影响公司业务的增速速度。
- **大陆厂商技术进步不及预期、中美贸易摩擦加剧、研报使用的信息更新不及时的风险。**

投资评级说明：

	评级	说明
股票评级	买入	预期未来 6~12 个月内相对同期基准指数涨幅在 15%以上
	增持	预期未来 6~12 个月内相对同期基准指数涨幅在 5%~15%之间
	持有	预期未来 6~12 个月内相对同期基准指数涨幅在-10%~+5%之间
	减持	预期未来 6~12 个月内相对同期基准指数跌幅在 10%以上
行业评级	增持	预期未来 6~12 个月内对同期基准指数涨幅在 10%以上
	中性	预期未来 6~12 个月内对同期基准指数涨幅在-10%~+10%之间
	减持	预期未来 6~12 个月内对同期基准指数跌幅在 10%以上
备注：评级标准为报告发布日后的 6~12 个月内公司股价（或行业指数）相对同期基准指数的相对市场表现。其中 A 股市场以沪深 300 指数为基准；新三板市场以三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）为基准；香港市场以摩根士丹利中国指数为基准，美股市场以标普 500 指数或纳斯达克综合指数为基准（另有说明的除外）。		

重要声明：

中泰证券股份有限公司（以下简称“本公司”）具有中国证券监督管理委员会许可的证券投资咨询业务资格。本报告仅供本公司的客户使用。本公司不会因接收人收到本报告而视其为客户。

本报告基于本公司及其研究人员认为可信的公开资料或实地调研资料，反映了作者的研究观点，力求独立、客观和公正，结论不受任何第三方的授意或影响。本公司力求但不保证这些信息的准确性和完整性，且本报告中的资料、意见、预测均反映报告初次公开发布时的判断，可能会随时调整。本公司对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。本报告所载的资料、工具、意见、信息及推测只提供给客户作参考之用，不构成任何投资、法律、会计或税务的最终操作建议，本公司不就报告中的内容对最终操作建议做出任何担保。本报告中所指的投资及服务可能不适合个别客户，不构成客户私人咨询建议。

市场有风险，投资需谨慎。在任何情况下，本公司不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。

投资者应注意，在法律允许的情况下，本公司及其本公司的关联机构可能会持有报告中涉及的公司所发行的证券并进行交易，并可能为这些公司正在提供或争取提供投资银行、财务顾问和金融产品等各种金融服务。本公司及其本公司的关联机构或个人可能在本报告公开发布之前已经使用或了解其中的信息。