

谁是国产英伟达

AIGC行业深度报告(8)

华西计算机团队

2023年6月5日

分析师：刘泽晶

SAC NO: S1120520020002

邮箱：liuzj1@hx168.com.cn

核心逻辑:

- ◆ **全球科技巨头, GPU王者——英伟达。** 英伟达是全球GPU龙头, 2016年至今股价大幅增长近60倍, AI开启全新增长周期。2015年至今, 英伟达持续深耕AI领域, 并发布了多款AI硬件产品。公司GPU产品功能分为用于计算和网络的GPGPU与用于图形处理(Graphics)的GPU; 平台化布局, 打造4条产品线覆盖下游主流应用, 分别是数据中心、游戏、专业可视化、自动驾驶, 其中数据中心和游戏是公司主要业务, 英伟达旗下产品包括云、边、端全面布局。**我们复盘了近10年内英伟达股价走势, 认为英伟达正在开启新一轮成长周期, 我们认为此次AI浪潮不同于元宇宙阶段, 大模型已经产生相关落地应用, 相关大模型的火热势必对算力产生超高需求, 英伟达作为全球算力龙头深度受益。**
- ◆ **为什么是英伟达:** 英伟达作为全球AI算力龙头, **以CUDA架构开启软硬件生态, 形成护城河。** CUDA的本质是“软件定义硬件”, 实现“软件调用硬件”, 可以简单理解, CUDA是英伟达实现软硬件适配的一种“类编译器”, 将软件的代码转换成硬件汇编代码, CUDA是英伟达实现软硬件生态的护城河。此外, CUDA核越多, 计算性能越强, 而GPU的CUDA核数是CPU的上百倍, 因此GPU比CPU更适合于并行计算。此外, 英伟达今年**发布多款AI产品, 助力全球AI生态**, 例如加速库、Grace CPU、DGX超级计算机、全新AI服务平台, AI foundations 云服务, 我们判断英伟达正以AI产品开启第二波成长曲线。
- ◆ **AI硬件自主可控势在必行:** 如果说产品是AI赋能、企业开启第二轮业绩增长曲线的“流量入口”, 那么算力即是大厂开启算力争夺战的“入场券”。我国目前已有较多应用, 大模型短期百家争鸣, “自研大模型热”仍将持续, 国内大模型自研进度明显加速, 势必对算力提出更高要求。近年来, 美国连续发动对我国高科技行业制裁, 执意对我国高科技企业进行制裁, 因此自主可控势在必行, 我国**政策端**持续发力, 加速推动国产自主可控进程, 我国短期发布多条政策助力AI发展, 工作方向主要瞄准推动国产AI芯片突破等。此外, 我国**产业端**积极响应, 智能算力建设正处于持续提速阶段。**我们再次重申观点, 短期算卡为王, 长期自主可控!**
- ◆ **投资建议:** 关注三条投资主线: **1) AI芯片厂商**, 相关受益标的为: **寒武纪、海光信息、景嘉微、龙芯中科**等; **2) AI服务器厂商**, 相关受益标的为: **中科曙光、神州数码、拓维信息、工业富联、浪潮信息**等; **3) AI云厂商**, 相关受益标的为: **首都在线、鸿博股份、青云科技、优刻得、光环新网、新炬网络**等。
- ◆ **风险提示:** 核心技术水平升级不及预期的风险、AI伦理风险、政策推进不及预期的风险、中美贸易摩擦升级的风险。



目录

- 01 什么是英伟达？
- 02 AI硬件自主可控势在必行
- 03 投资建议：梳理算力相关受益厂商
- 04 风险提示

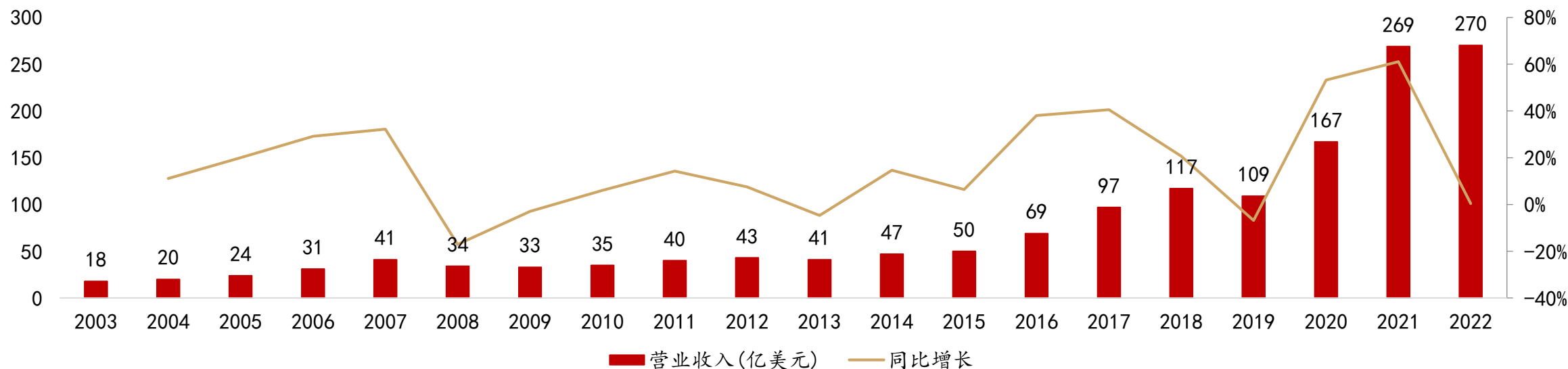


01 为什么是英伟达

1.1 全球科技巨头,GPU王者——英伟达

- ◆ **英伟达(NVIDIA)是全球GPU巨头。**英伟达成立于1993年，总部位于美国加利福尼亚州。公司专注于GPU的研发与制造，2009年发布了费米(Fermi)架构，确立了在游戏领域的主导地位。公司业务包括数据中心、游戏、科学计算和自动驾驶。在人工智能领域，TensorCore作为深度学习的处理单元，为AI提供高效的计算和学习能力。出色的软件研发为公司持续发展提供支持，CUDA平台和深度学习库广泛应用于科研和大数据等领域。
- ◆ **2016年至今股价大幅增长近60倍，AI开启全新增长周期。**英伟达(NVIDIA)截至2023年5月30日的总市值为9631.8亿美元，收盘价为每股389.46美元。根据JonPenddie Reasearch数据，独立显卡市场中，英伟达在全球GPU市场占有84%份额。近20年，英伟达业绩收入大幅增长15倍，业绩持续爆发式增长，我们认为这是英伟达股价持续增高的根本原因。此外，公司在2015年开启布局相应人工智能领域，并于2019年崭露头角并逐渐成为全球AI巨头，如今，随着大模型的爆发，英伟达作为AI硬件龙头，开启第二波成长曲线。

英伟达近20年收入(亿美元)与增速



1.2 30年王者之路，AI硬件巨头崛起

- ◆ **竞争激烈，勇于破冰(1993年-1998年)**：英伟达于1993年进入市场，当时显示芯片行业竞争激烈。1995年，英伟达推出NV1，但成效不明显，财政形势紧张。然而，1997年，公司推出了**全球首款128位3D处理器RIVA 128(加速图形处理芯片)**，仅前四个月就售出超过一百万台，成功逆袭。此后，英伟达在1998年继续发力，发布了两款高性能的3D处理器，RIVA 128ZX和RIVA TNT。
- ◆ **成功上市，高速发展（1999年-2005年）**：1999年1月，英伟达在NASDAQ股票交易所以每股12美元的价格进行了首次公开募资。同年八月，发布了全球首款GPU(**GeForce 256**)，将GPU定义为具备集成变换、照明、三角设置、裁剪和渲染引擎的单片处理器，能够每秒处理至少1000万个多边形。英伟达成为发展最快的半导体公司之一，收入达到**10亿美元**，并被纳入S&P500指数。
- ◆ **CUDA问世，强调生态(2006年-2009年)**：2006年推出了CUDA，一种用于通用GPU计算的革命性架构，使科学家和研究人员能够进行更复杂的计算。2009年发布了首个完整的GPU计算架构（Fermi），其中Quadro 7000代表着一个飞跃，实现了游戏性能和计算性能的双重提升。

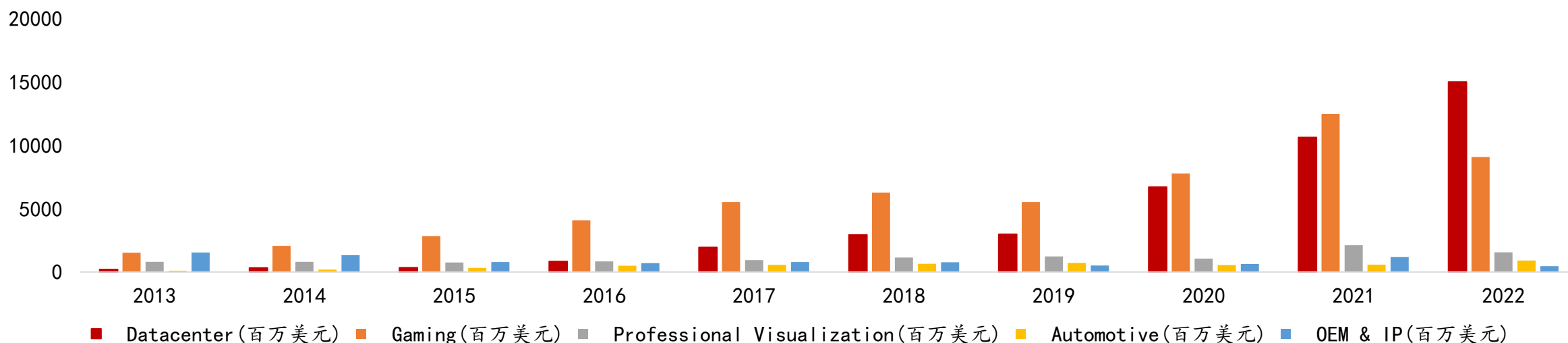
英伟达近30年发展史概括图



1.2 30年王者之路，AI硬件巨头崛起

- ◆ **多领域发展，产品多样(2010年-2014年)：自动驾驶领域：**发布NVIDIA DRIVE，为自动驾驶汽车铺平了道路。超级计算领域：多台超级计算机选用英伟达产品。中国的Tianhe-1A，橡树岭国家实验室Titan。电子产品领域：打造出多款领先市场的平板电脑以及手机2012年。电影领域：为多个名气较高的电影提供技术服务，例如《阿凡达》《星际迷航》和《盗梦空间》等。
- ◆ **深耕AI领域，算力赋能千行百业(2015年-至今)：**2015年，NVIDIA GeForce GTX TITAN X问世，专为训练深度神经网络而打造。2016年，推出NVIDIA® DGX-1™，此为全球首款一体化深度学习超级计算机。公司在 GTC 2019 大会上推出多项创新应用领域，涵盖人工智能，超级计算，自动驾驶，机器人等。并在同年推出NVIDIA® EGX边缘计算平台，将AI引入企业边缘。此外，将AI成功引入多种领域，城市管理、家庭生活、制造/配送/零售、医疗健康(NVIDIA Clara)。2023年，英伟达DGX大会上，持续赋能加速计算AI潮流，推出多款AI产品，例如DGX超级计算机等，并于COMPUTEX大会上推出超级GPU GH200，持续引领AI硬件市场！

2013年-2022财年英伟达各业务营收(百万美元)



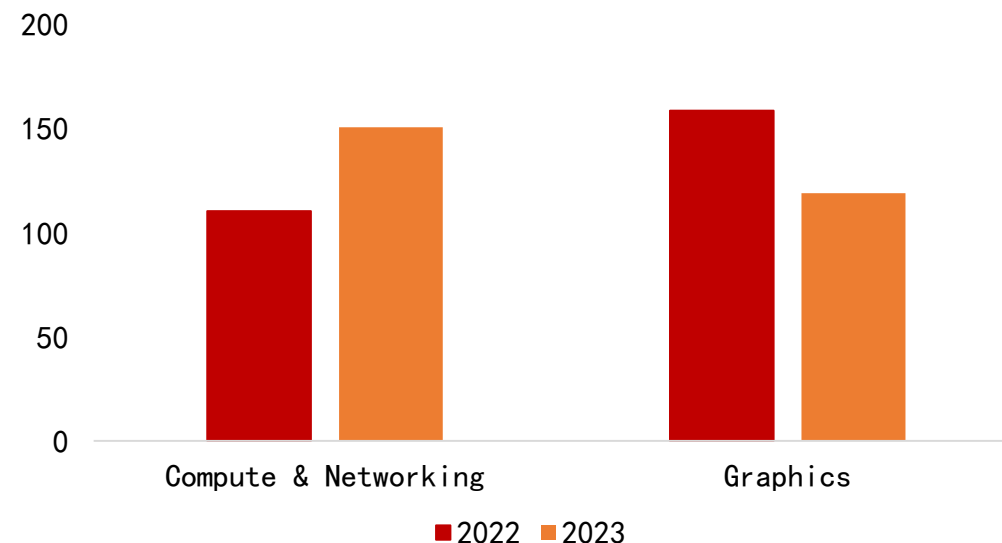
1.3 产品：图形显控+计算中心加速卡双轮驱动

- ◆ 公司GPU产品功能分为1) 用于**计算和网络 (Compute & Networking) 的GPGPU**；2) 用于**图形处理 (Graphics) 的GPU**。
- ✓ **GPGPU (General Purpose GPU)**：通用计算图形处理器。在架构设计中去掉了GPU为图形处理设计的加速硬件单元，保留了SIMT架构和通用计算单元。可以将GPU的并行计算能力应用于科学计算、数据分析、机器学习等领域，提高计算速度和效率。
- ✓ **GPU (Graphics Processing Unit)**：完成图像运算工作的微处理器。作为一个单独的模块，即独立显卡核心或者主板集成显卡核心。主要用于提供高性能的图形渲染能力。
- ◆ 根据2023财年年报，公司计算与网络类GPGPU收入150.68亿美元，图形处理类GPU收入119.06亿美元。其中1) 计算与网络类收入同比+36%，主要应用于数据中心加速计算平台、人工智能驾驶舱、自动驾驶解决方案、电动汽车计算平台、NVIDIA AI企业和其他软件、加密货币挖掘等。2) 图形处理类收入同比-25%，主要应用于游戏和个人电脑的GeForce图形处理器、游戏平台的解决方案、基于云的视觉和虚拟计算的软件、构建和操作3D互联网应用程序的全方位企业软件等。

GPGPU和GPU的区别

	GPGPU	GPU
设计目标	提供高性能的通用计算能力	提供高性能的图形渲染能力
应用场景	用于科学计算、数据分析、机器学习等领域	用于图形渲染和游戏开发
存在形式	通常被集成GPU当中	作为一个单独的模块（独立显卡核心或者主板集成显卡核心）
硬件支持	通常有更多通用计算单元和高速缓存，以支持更广泛的计算任务	通常有专门的硬件支持，如纹理单元、像素处理单元等
软件支持	通常需额外的软件和算法支持，以充分发挥性能优势	通常有丰富的图形API和图形引擎支持
支持计算任务	支持更广泛计算任务，包括浮点计算，整数计算、向量计算等	支持图形渲染相关计算任务，如顶点处理、像素处理、纹理处理等

英伟达Compute & Networking及Graphics类业务收入（亿美元）



1.4 业务: 打造多元产品矩阵，数据中心与游戏为核心

◆ 公司软硬件结合平台化布局，打造4条产品线覆盖下游主流应用：

- ✓ **数据中心**：2023财年收入150.1亿美元，占比56%。基于 GPU、DPU 和 CPU 三种新一代架构构建的 NVIDIA 加速计算平台，推出NVIDIA DGX 人工智能超级计算机，让现代化的数据中心更快速地处理涉及深度学习、机器学习和高性能计算 (HPC) 的工作负载。
- ✓ **游戏**：2023财年收入90.7亿美元，占比34%。产品包括GeForce RTX和GeForce GTX图形处理器，云游戏GeForce NOW，用于流媒体的屏蔽以及芯片系统（SOCs）和游戏机的开发服务。2023财年推出基于Ada Lovelace架构的GeForce RTX 40系列游戏图形处理器。
- ✓ **专业可视化**：2023财年收入15.4亿美元，占比6%。应用于许多领先的3D设计和内容创建，如全宇宙、虚拟现实和增强现实技术。利用 GPU在设计制造、数字内容创建方面提供动力。推出的NVIDIA RTX平台可利用光线跟踪，实时渲染胶片质量、逼真的物体和环境。
- ✓ **自动驾驶**：2023财年收入9.0亿美元，占比3%。包括AV、人工智能驾驶舱、电动汽车计算平台和信息娱乐平台解决方案。根据2023财年年报，公司正与数百名汽车生态伙伴合作，包括汽车产业链制造商、汽车研究机构、地图公司和初创公司，为自动驾驶汽车开发和部署人工智能系统。推出的Drive作为一个人工智能汽车平台，覆盖多种自动驾驶领域。

英伟达四条产品线（2022年）



Data Center

\$15.0 billion revenue
up 41% year on year



Gaming

\$9.1 billion revenue
down 27% year on year



Professional Visualization

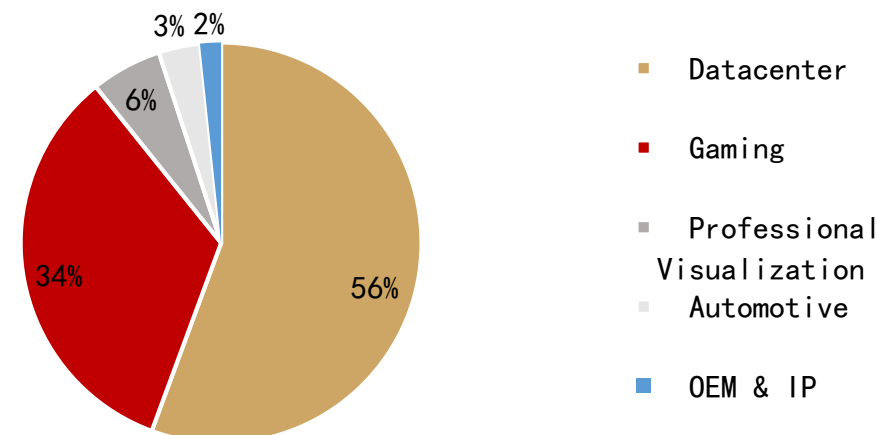
\$1.5 billion revenue
down 27% year on year



Automotive

\$0.9 billion revenue
up 60% year on year

2022年(2023财年)英伟达收入结构



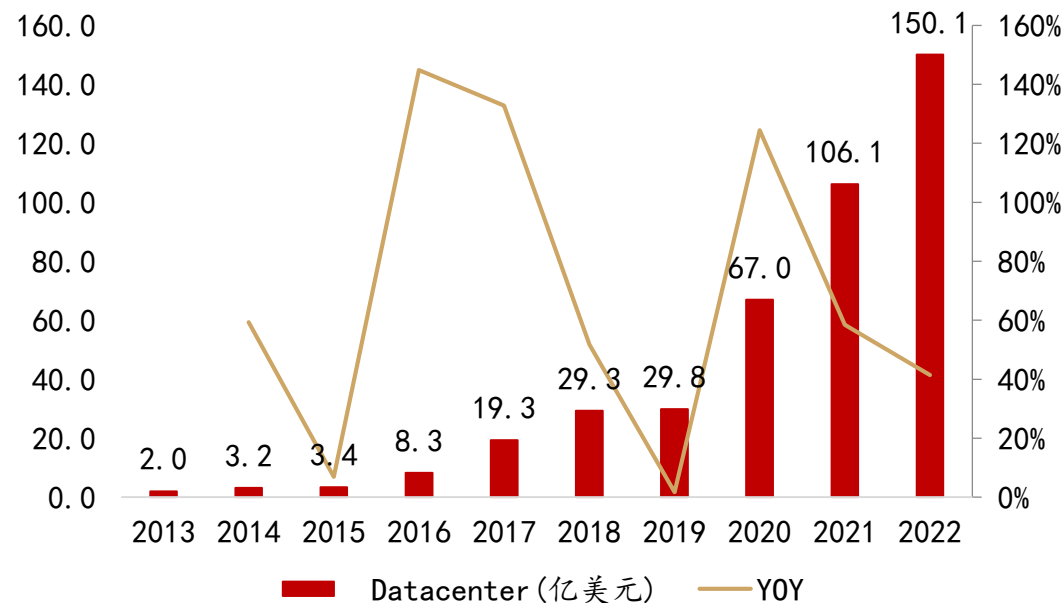
1.4.1 业务: 数据中心GPU , 全球高端领导者

- ◆ **数据中心GPU：全球高端GPU领导者，经数十代产品迭代，技术指标全面升级。**从2011年英伟达推出Tesla M2090数据中心GPU，到2022年H100、L40等型号产品，多项核心技术指标大幅提升。其中，**CUDA作为GPU内部主要的计算单元，从512个升级到超14000个**；芯片工艺尺寸也从40nm降至4nm；单精度浮点算力从1332GFLOPS增至超50TFLOPS。GPU产品性能整体大幅跃升。
- ◆ **数据中心CPU：推出Grace 系列，加速大型 AI、HPC、云和超大规模工作负载。**2022年公司发布首款CPU产品Grace，用于高性能计算和云计算。Grace CPU超级芯片采用NVLink®-C2C 技术，可提供 144 个 Arm®Neoverse V2 核心和 1 TB/s 的内存带宽，**每瓦性能是当今领先 CPU的 2 倍。**此外，公司还推出的Grace Hopper超级芯片将 Grace 和 Hopper 架构相结合，为加速 AI 和高性能计算 (HPC) 应用提供 CPU+GPU 相结合的一致内存模型。

英伟达部分数据中心GPU产品及参数

系列	型号	发布时间	图形处理器	CUDA核心数	内存	线宽度 (位)	工艺尺寸 (nm)	最大功耗 (W)	单精度 FP32
Tesla	M2090	2011年	GF110	512	6GB GDDR5	384	40	250	1332 GFLOPS
Tesla	K40	2013年	GK180	2880	12GB GDDR5	384	28	245	5.046 TFLOPS
Tesla	M40	2015年	GM200	3072	12GB GDDR5	384	28	250	6.832 TFLOPS
Tesla	P100	2016年	GP100	3584	12GB/16GB HBM2	4096	16	250	9.526 TFLOPS
Tesla	V100	2017年	GV100	5120	32GB/16GB HBM2	4096	12	300	14.13 TFLOPS
A100	A100	2020年	GA100	6912	40GB/80GB HBM2e	5120	7	250	19.49 TFLOPS
A2	A2	2021年	GA107	1280	16GB GDDR6	128	8	60	4.531 TFLOPS
L40	L40	2022年	AD102	18176	48GB GDDR6	384	5	300	90.52 TFLOPS
H100	H100	2022年	GH100	14592	80 GB HBM2e	5120	4	350	51.22 TFLOPS

2013-2022年英伟达数据中心业务收入(亿美元)及增速 (%)



1.4.2 游戏: 长期布局多规格产品

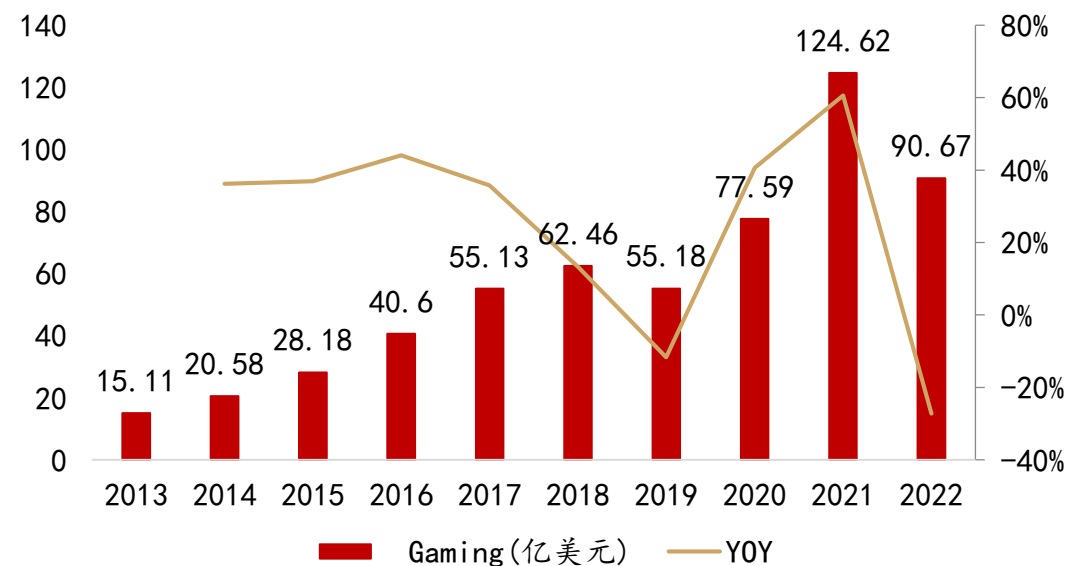
- ◆ **游戏业务：长期布局多规格产品，重磅推出ACE 游戏开发版。**游戏业务是英伟达第二大产品线，每一代新品推出带来显著的性能提升。CUDA核心数从512个增长到超16000个，目前芯片尺寸最低达到5nm。
- **NVIDIA RTX：**具有光线追踪技术，可以模拟光的物理行为，从而在计算机生成的场景实现更逼真的效果。还提供深度学习技术NVIDIA DLSS，可以提高帧率，为游戏生成美丽清晰的图像。去年推出GeForce RTX 40系列，采用第三代RTX技术，提供上一代4倍的性能。
- **GeForce NOW：**用于云端，链接数字在线PC游戏平台让玩家串联游戏库。
- **ACE 游戏开发版**（Avatar Cloud Engine for Games）：据东方财富网消息，23年5月29日公司推出全新定制AI模型代工服务ACE 游戏开发版，利用AI驱动的自然语言交互技术，为游戏中的非玩家角色（NPC）带来智能。中间件、工具及游戏开发者可以使用“ACE 游戏开发版”在他们的游戏和应用中建立和部署定制的语音、对话及动画AI模型。

英伟达部分游戏显卡及参数

系列	型号	发布时间	CUDA核心数	加速频率 (GHz)	基础频率 (GHz)	标准显存配置	工艺尺寸 (nm)	显存位宽	显卡功率 (W)
GeForce GTX 16	GTX 1630	2022年	512	1.785	1.74	4GB GDDR6	12	64	75
	GTX 1650 (G6)	2019年	896	1.59	1.41	4GB GDDR6	12	128	75
	GTX 1650 (G5)	2019年	896	1.665	1.485	4GB GDDR5	12	128	75
	GTX 1650 Super	2019年	1280	1.665	1.53	4GB GDDR6	12	128	100
	GTX 1660	2019年	1408	1.785	1.53	6GB GDDR5	12	192	120
	GTX 1660 Super	2019年	1408	1.785	1.53	6GB GDDR6	12	192	125
	GTX 1660 Ti	2019年	1536	1.770	1.5	6GB GDDR6	12	192	120
GeForce RTX 20	RTX 2060	2019年	2176/1920	1.65/1.68	1.47/1.37	12GB/6GB GDDR6	12	192	185 / 160
	RTX 2060 SUPER	2019年	2176	1.65	1.47	8GB GDDR6	12	256	175
	RTX 2070	2018年	2304	1.71	1.41	8GB GDDR6	12	256	185
	RTX 2070 SUPER	2019年	2560	1.77	1.41	8GB GDDR6	12	256	215
	RTX 2080	2018年	2944	1.8	1.52	8GB GDDR6	12	256	225
	RTX 2080 SUPER	2019年	3072	1.82	1.65	8GB GDDR6	12	256	250
	RTX 2080 Ti	2018年	4352	1.64	1.35	11GB GDDR6	12	352	260
GeForce RTX 30	RTX 3050	2022年	2560/2304	1.78/1.76	1.55/1.51	8GB GDDR6	12	128	130
	RTX 3060	2021年	3584	1.78	1.32	12GB/8GB GDDR6	8	192/128	170
	RTX 3060 Ti	2020年	4864	1.67	1.41	8GB GDDR6/GDDR6X	8	256	200
	RTX 3070	2020年	5888	1.73	1.5	8GB GDDR6	8	256	220
	RTX 3070 Ti	2021年	6144	1.77	1.58	8GB GDDR6	8	256	290
	RTX 3080	2020年	8960/8704	1.71	1.26/1.44	12GB/10GB GDDR6X	8	384/320	350/320
	RTX 3080 Ti	2021年	10240	1.67	1.37	12GB GDDR6X	8	384	350
GeForce RTX 40	RTX 3090	2020年	10496	1.70	1.4	24GB GDDR6X	8	384	350
	RTX 3090 Ti	2022年	10752	1.86	1.56	24GB GDDR6X	8	384	450
	RTX 4060	2023年	3072	2.46	1.83	8GB GDDR6	5	128	115
	RTX 4060 Ti	2023年	4352	2.54	2.31	16GB/8GB GDDR6	5	128	165.16
	RTX 4070	2023年	5888	2.48	1.92	12GB GDDR6X	6	192	200
	RTX 4070 Ti	2023年	7680	2.61	2.31	12GB GDDR6X	6	192	285
	RTX 4080	2022年	9728	2.51	2.21	16GB GDDR6X	6	256	320
	RTX 4090	2022年	16384	2.52	2.23	24GB GDDR6X	6	384	450

资料来源：techpowerup, Bloomberg, 英

2013-2022年英伟达数据中心业务收入(亿美元)及增速 (%)



1.4.3 专业可视化、自动驾驶打造多元产品矩阵

- ◆ **专业可视化业务：着力提升视觉体验，扩大应用场景。**从2017年推出的Quadro P400到当前RTX A6000，显卡专业性能实现质的飞跃，CUDA核心数从256个增至超18000个，工艺尺寸达7nm。根据2023财年年报，元宇宙、虚拟现实或增强技术正被纳入越来越多的企业应用程序中，公司的GPU正为虚拟汽车展厅、外科培训、设计和制造、数字内容创建等多种场景提供动力。同时Pascal、Volta、Ampere等新架构，以及RTX Studio等软件工具，深度赋能英伟达提供解决方案的能力。
- ◆ **自动驾驶业务：业务稳健增长，提供澎湃算力。**公司推出的NVIDIA DRIVE嵌入式超级计算平台，可处理来自摄像头、普通雷达和激光雷达传感器的数据，以感知周围环境、在地图上确定汽车的位置，然后规划并执行安全的行车路线。此外，包括Atlan和Orin芯片在内的SoC产品，是智能车辆的中央计算机，为自动驾驶功能、置信视图、数字集群以及 AI 驾驶舱提供动力支持。根据官网新闻，于2021年推出的最新款Atlan可提供每秒超过1000万亿次（TOPS）运算次数，将用于多家汽车制造商的2025年车型上。

专业可视化及自动驾驶业务主要产品概览

专业可视化

- **工作站GPU：**
RTX系列、Quadro系列
- **虚拟工作站：**
Quadro vWS、RTX vWS

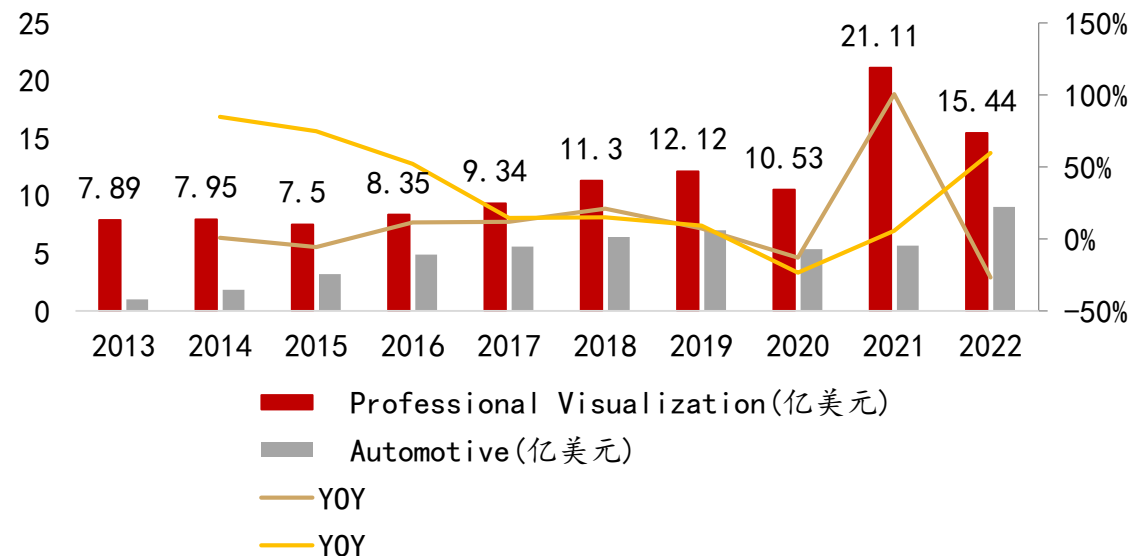


自动驾驶

- **系统级芯片：**
Orin SoC、Altan SoC
- **平台：**
NVIDIA DRIVE Hyperion、Thor等



2014-2023财年英伟达专业可视化及自动驾驶业务收入（亿美元）



1.5 产品形态多元化，实现云、边、端全面布局

◆ 云、边、端全面布局，推出平台化解决方案：

- ✓ **云端**：提供云端解决方案，目前所有高级云平台均支持NVIDIA GPU 加速解决方案，为世界各地创新者提供巨大算力。此外，IaaS 产品Omniverse Cloud 可连接在云端、边缘设备等，让客户在任何位置设计、发布和体验交互式 3D 互联网应用。Omniverse Cloud支持自部署云容器以及托管服务（软件即服务）。
- ✓ **边缘计算**：推出适用于自主机器和其他嵌入式应用平台Jetson，包括 Jetson 模组(高性能计算机)、用于加速软件的 NVIDIA JetPack SDK，以及包含传感器、SDK、服务和产品的生态系统，从而加快开发速度。每个NVIDIA Jetson 都是一个完整的系统模组，包括GPU、CPU、内存、电源管理和高速接口等。Jetson 生态系统合作伙伴提供软件、硬件设计服务以及涵盖载板到完整系统的现成兼容产品，客户可以借助 AI 嵌入式边缘设备更快地打入市场。
- ✓ **终端**：在游戏、可视化、智能驾驶等多领域提供包括驱动器、显示器、智能软件服务系统等终端解决方案及产品，不断深入各应用场景。

产品实现云、边、端全面布局

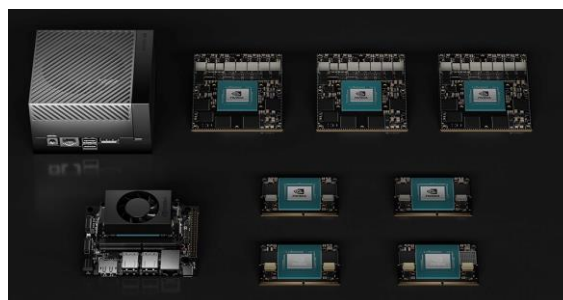
云端

- GPU加速云计算（在云端完成计算）
- Omniverse Cloud：自部署云容器、托管服务



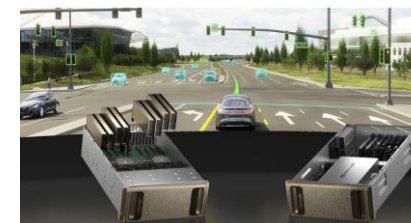
边缘计算

- Jetson嵌入式系统：Orin系列、Xavier系列、TX2系列、Nano（在数据源或数据源附近完成计算）



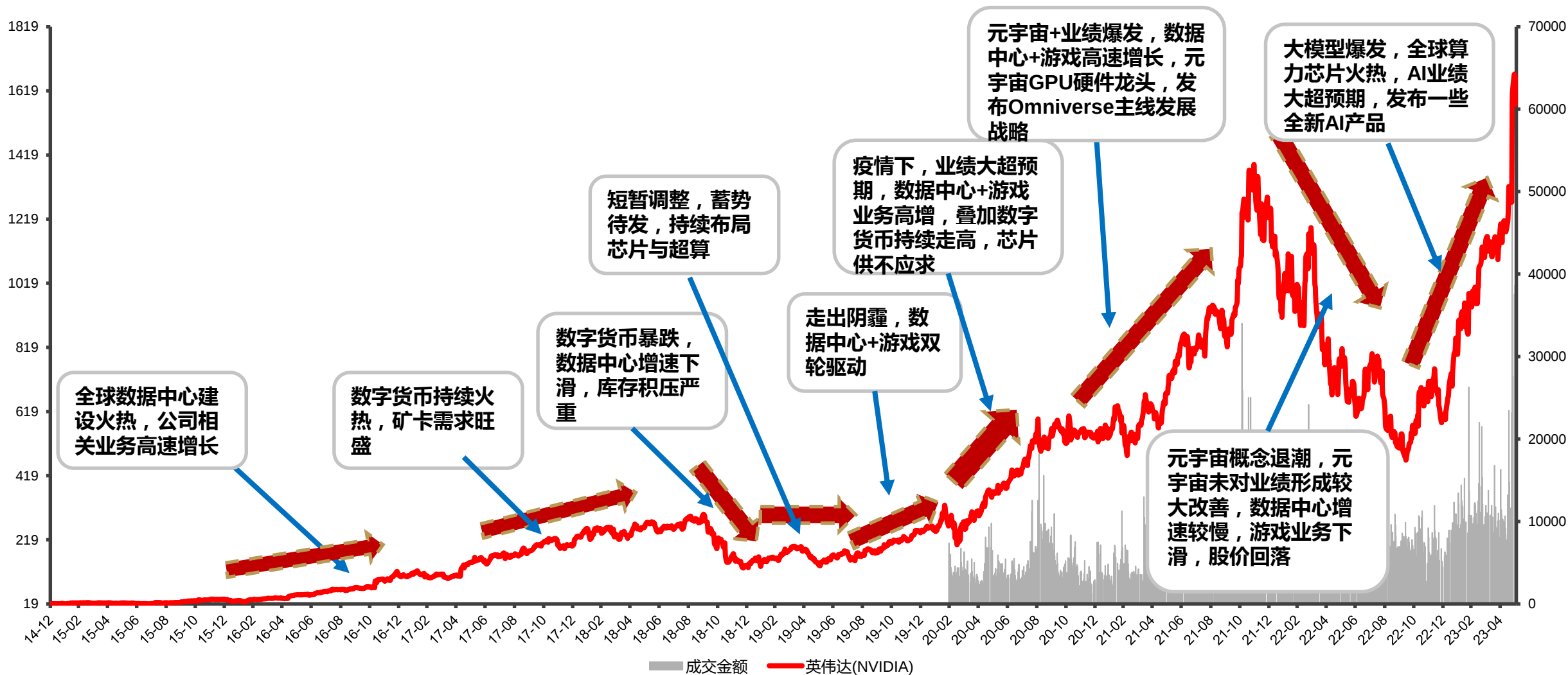
终端

- **游戏**：驱动器、Reflex、G-SYNC 显示器
- **可视化**：虚拟工作站、NVIDIA RTXDI光线追踪等
- **智能驾驶**：舱内智能服务软件、地图软件、辅助驾驶平台等



1.6 复盘英伟达十年成长曲线

2015年至今英伟达走势复盘图



资料来源：， 华西证券研究所

更多免费行业报告

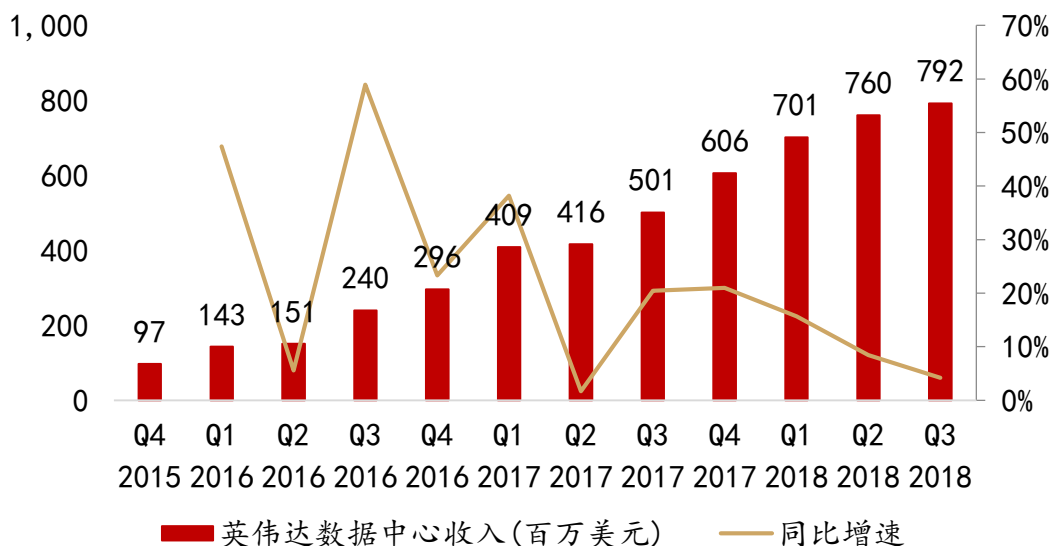
并购家

www.ipoipo.cn

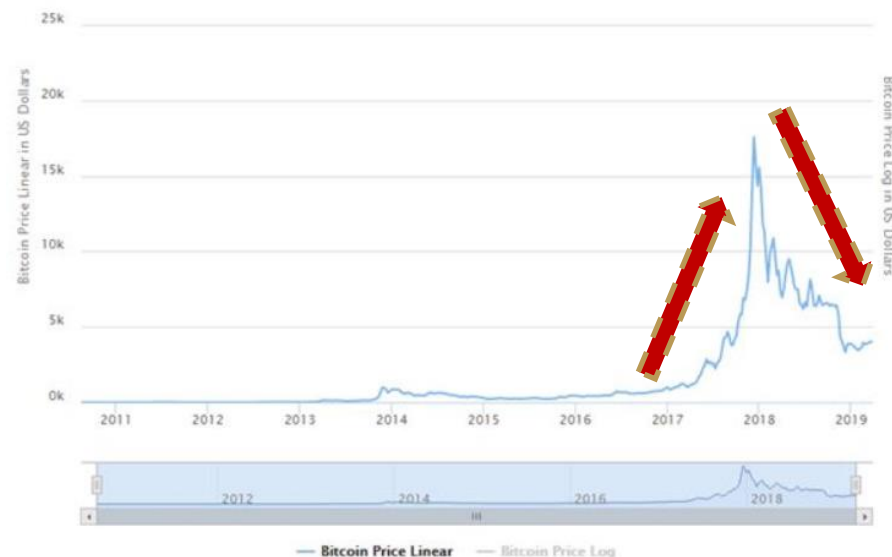
1.6.1 2016初-2018年中，数据中心业务高速增长+数字货币持续火热

- ◆ **2016年-2017年底，得益于数据中心营收高速增长：**受益于全球云计算和大数据业务高速发展，2015年底开始，全球开启大规模的数据中心建设任务，英伟达数据中心业务迎来快速发展期，根据Bloomberg数据，英伟达2016年Q1、Q2、Q3、Q4、2017年Q1、Q2、Q3、Q4数据中心收入规模分别为143、151、240、296、409、416、501、606亿元，同比增速分别为47%、6%、59%、23%、38%、2%、20%、21%，可以看到英伟达数据中心业务短期呈现爆发趋势。
- ◆ **2017年-2018年年中，数字货币大幅增长，带动相关矿卡需求：**以比特币价格为例，比特币于2017年前后开始大幅增长，于2018年年初达到顶峰，极大的增加了对于大数据处理的需求，2016年开始，英伟达推出**GeForce 10系列显卡**，产品GPU并行计算能力远超当时所有产品，很快成为了众多投资者的必选项。数字货币的高速发展，同时造就了英伟达股价的大幅增长。

英伟达2015年至2018年数据中心收入规模及增速



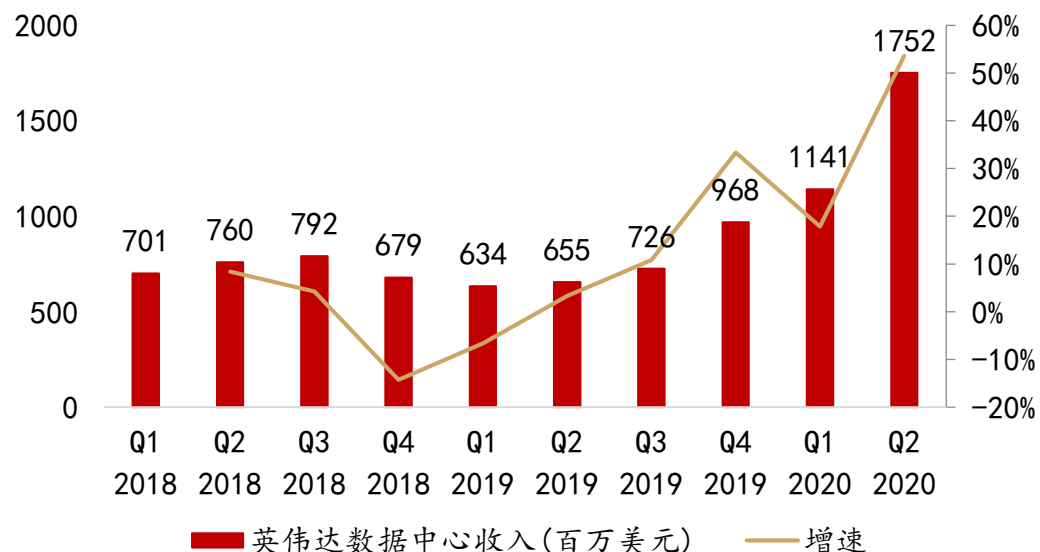
2015年-2019年比特币价格走势



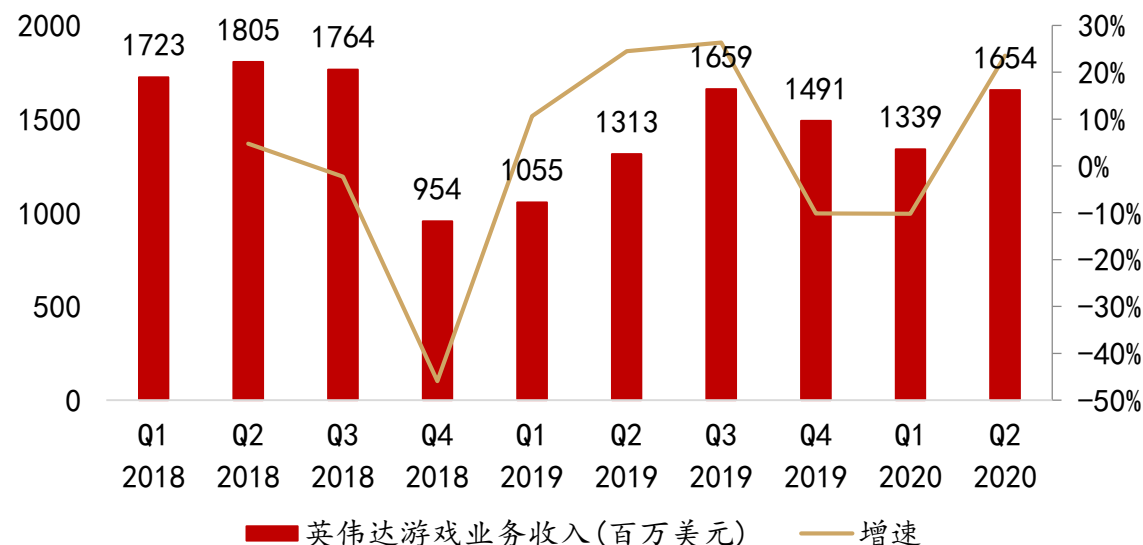
1.6.2 2018年下至2019Q3上，短暂调整，蓄势待发

- ◆ **2018Q3-Q4，数字货币暴跌+数据中心增速下滑，股价下跌:** 英伟达2018Q3、Q4数据中心营业收入为792、679亿元，同比4.21%、-14.27%，数据中心业务陷入瓶颈期。同时数字货币的价格下跌，导致矿机芯片需求量锐减，供过于求使库存严重积压。英伟达股价从原来的72.5美元直线下跌至30.88美元。
- ◆ **2019Q1-Q3，短暂调整，蓄势待发:** 2019年Q1-Q3，无论是游戏业务还是数据中心业务进入明显的业绩低迷期。然而英伟达进入快速调整阶段，年初即停止了大规模的游戏支出，3月11日，收购以色列芯片设计公司Mellanox，希望摆脱对加密货币以及游戏市场的依赖，从而专心将业务中心放在芯片技术上。公司逐步从游戏渠道库存过多的“困境”中走出，公司持续布局芯片与超算，6月，英伟达推出了全球排名第22的超级计算机DGX SuperPOD，短期来看，虽然增速较低，但已逐步“恢复元气”。

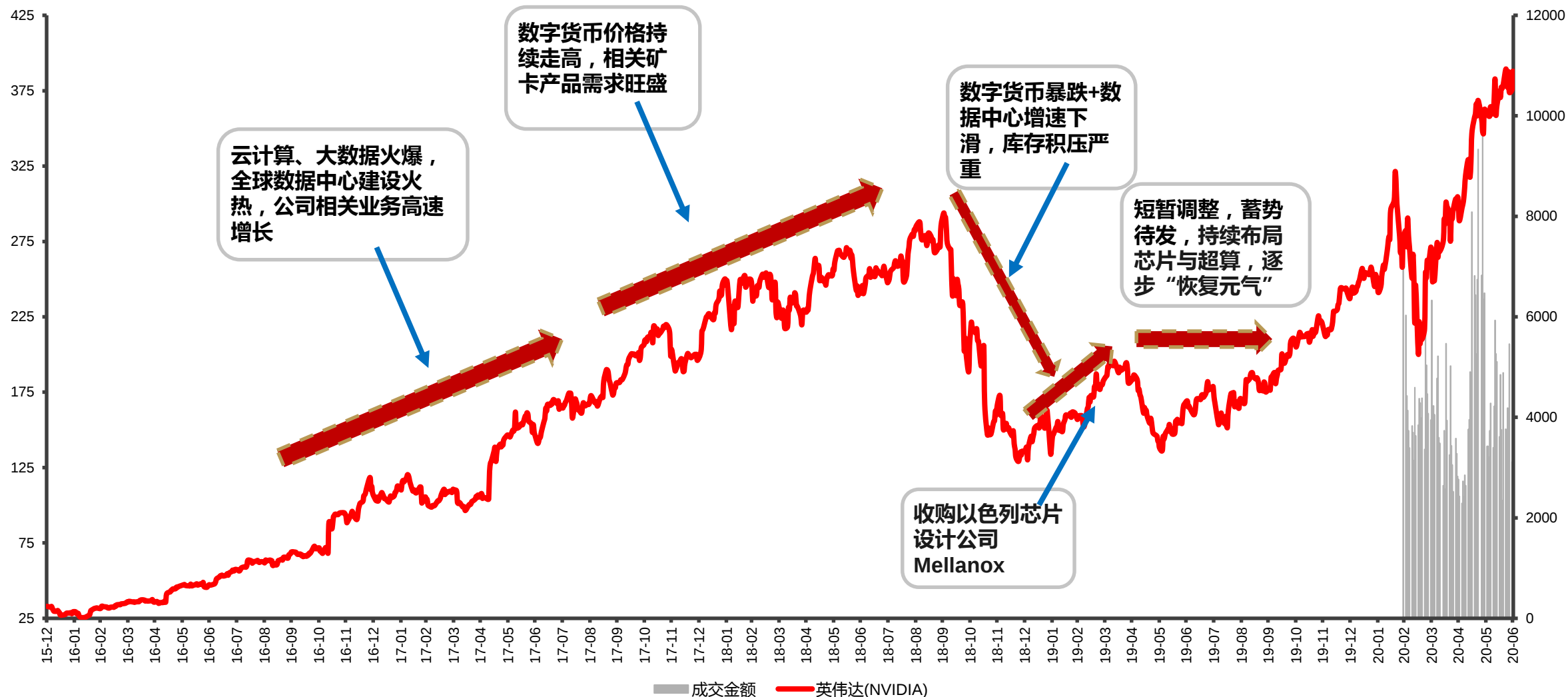
英伟达2018年Q1-2020年Q2数据中心业务收入及增速



英伟达2018年Q1-2020年Q2游戏业务收入及增速



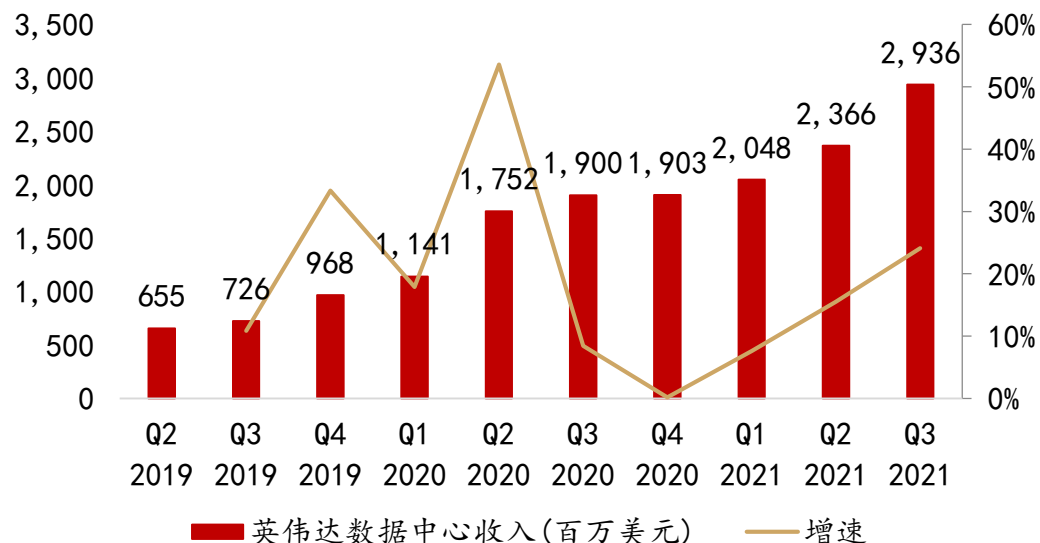
1.6.3 2016年Q1至2019年Q3复盘



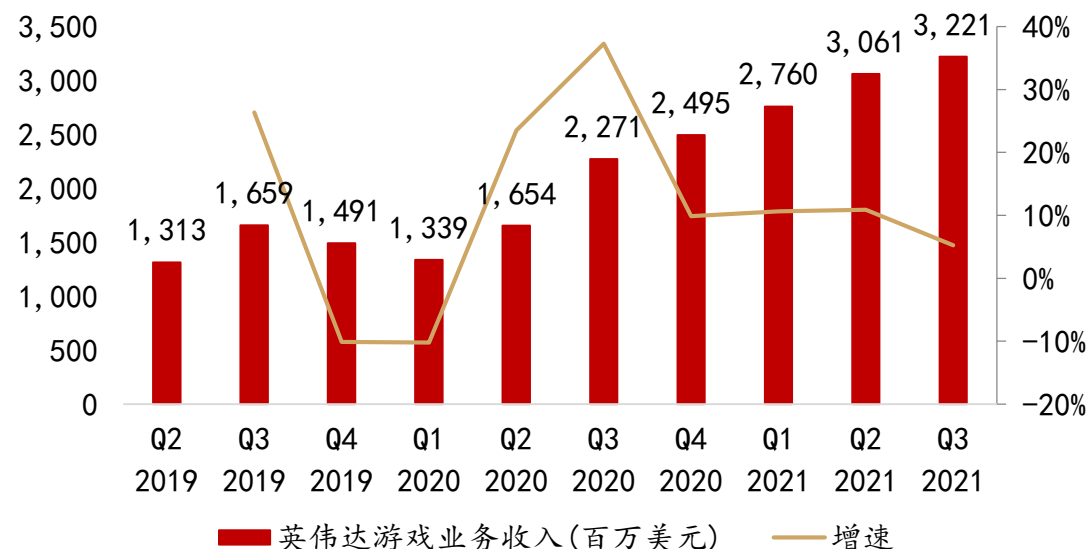
1.6.4 19年Q3至20Q4，王者归来，业绩高速增长

- ◆ **2019Q3-Q4，数据中心+游戏双轮驱动，业绩高速增长：**在经历长达一年的战略调整，英伟达已经走出阴霾，业绩重回高速增长，以数据中心业务和游戏业务为主，2019年Q3数据中心和游戏的收入分别为726亿美元、1659亿美元，同比增速分别为11%，26%，开启新一轮成长曲线。
- ◆ **2020Q1，疫情爆发，美国科技股全面下挫**由于疫情原因，美国科技股超预期地全面崩盘，NVDA股价也随之从原来的75.77美元开始下跌。
- ◆ **2020Q1-2020Q3，业绩超预期：**由于疫情原因，全球人们被迫居家，视频、游戏等成为主要娱乐活动，公司于2020年游戏业务持续超预期，公司2020年Q1、Q2、Q3游戏业务收入分别为1339、1654、2271亿元，同比增长分别为-10%、24%、37%，同时，英伟达数据中心业务同样高速增长，此外，数字货币持续走高，同样带动相关矿卡需求，英伟达芯片“供不应求”。

英伟达2019年Q1-2021年Q3数据中心业务增速



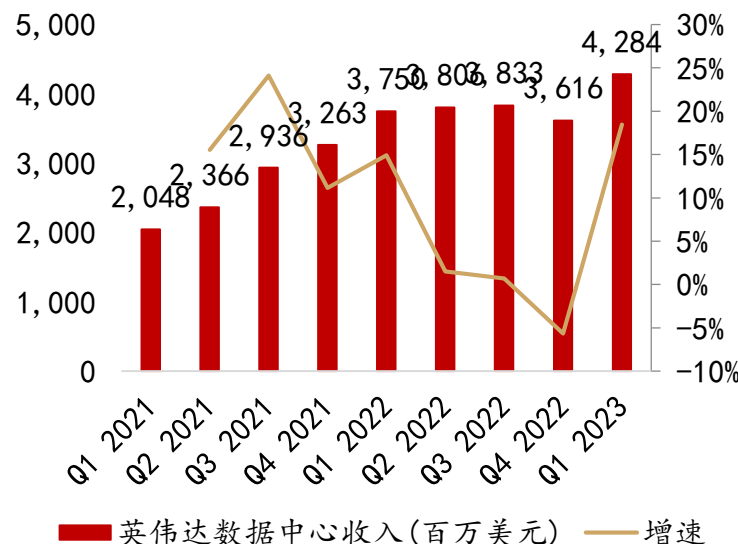
英伟达2019年Q1-2021年Q3游戏业务增速



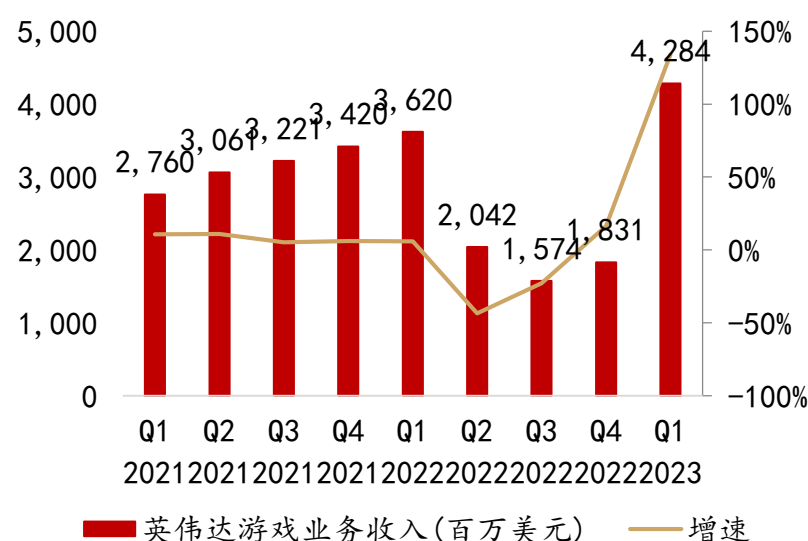
1.6.5 21Q1至22Q3，元宇宙超级巨星，Omniverse开启成长曲线

- ◆ **2021年，游戏与数据中心双轮驱动，元宇宙超级巨星出世：**2021年，英伟达数据中心板块与游戏板块持续高速增长，同时2021年，元宇宙概念兴起，元宇宙可以笼统地理解为一个平行于现实世界的虚拟世界，现实中人们可以做的事，都可以在元宇宙中实现，相关GPU硬件是构成未来元宇宙的硬件核心底座，因此英伟达作为全球GPU龙头深度受益。
- ◆ **同年，英伟达发布以Omniverse为主线发展的发展战略：**英伟达CEO黄仁勋在Computex 2021时，描绘了英伟达元宇宙业务策略。并于在2021年11月发布了Omniverse Avatar。Omniverse Avatar是基于语音、机器视觉、自然语言处理等技术形成的交互式AI产品。Omniverse是为虚拟世界建设者打造的平台，在上面可以运行逼真的虚拟世界，并与其他数字平台相连。英伟达希望未来科学家和企业可以借此大量建造其他虚拟世界，使虚拟世界如同今天的互联网那样不断涌现，创造数字双胞胎和工业元宇宙。
- ◆ **2022年，概念“退潮”，英伟达股价回落：**然而元宇宙由于硬件等相关条件制约仍停留在“梦想”阶段，并未在2022年对英伟达业务形成较大改善，因此，2022年，公司股价回落。

英伟达2020年Q1-2022年Q4数据中心业务增速



英伟达2020年Q1-2022年Q4游戏业务增速



2021年-2022年元宇宙和NFT搜索频率



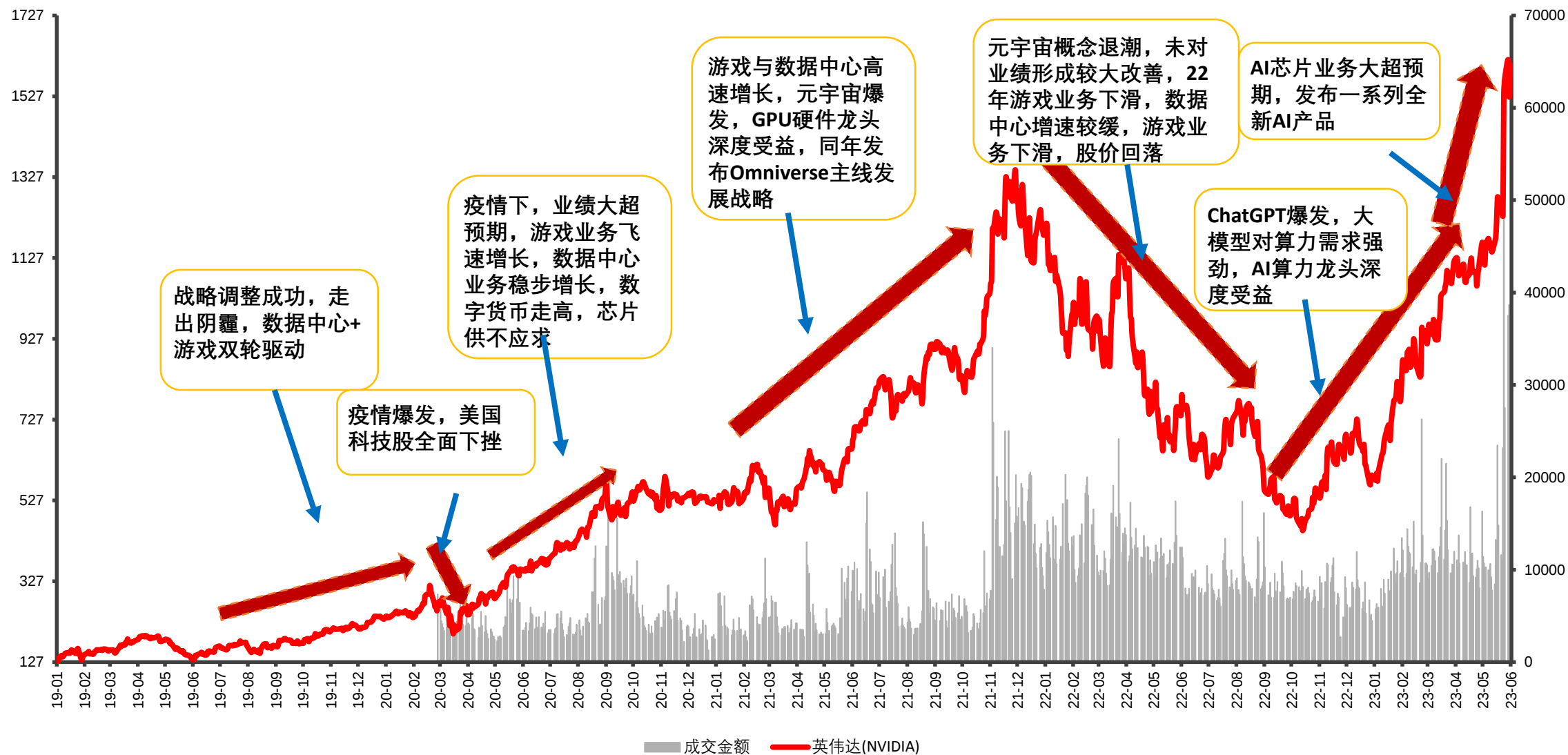
1.6.5 22Q4-至今，全球AI算力王者归来

- ◆ **AI大模型时代拉开帷幕，对AI芯片需求量明显增多。** AI从2012年发展至今，随着模型参数量越来越大，对于计算量的要求也逐步增高。2022年11月份OpenAI发布ChatGPT，以ChatGPT为代表的大模型持续引爆市场。早在2019年，微软斥资几亿美元为Open AI的训练打造一台超级服务器，其中包括上万张英伟达A100，旨意为ChatGPT和New Bing提供算力基础，而相关大模型的火爆，AI硬件竞争持续升温，芯片遭“哄抢”导致价格大涨，根据界面新闻，英伟达AI旗舰芯片H100已明显提价。
- ◆ **AI芯片业务大超预期：**根据公司2024财年发布会，公司 AI 芯片所在数据中心业务收入达 42.8 亿美元(环比+18%，同比+14%)，创历史新高,二季度营收预期约 110 亿美元，同比增近 33%；毛利率超 64%。业绩会上英伟达称，众多云公司竞相部署 AI 芯片，其锁定了数据中心芯片的大幅增长，计划下半年大幅增加供应。
- ◆ **英伟达持续发布全新产品，助力全球算力:** 英伟达今年陆续发布AI Foundations，DGX超级计算机，全新RTX4070，以及DGX GH200 AI超级计算机等全新AI产品。我们认为此次AI浪潮不同于元宇宙阶段，大模型已经产生相关落地应用，相关大模型的火热势必对算力产生超高需求，英伟达作为全球算力龙头深度受益。

英伟达2023年相关AI产品



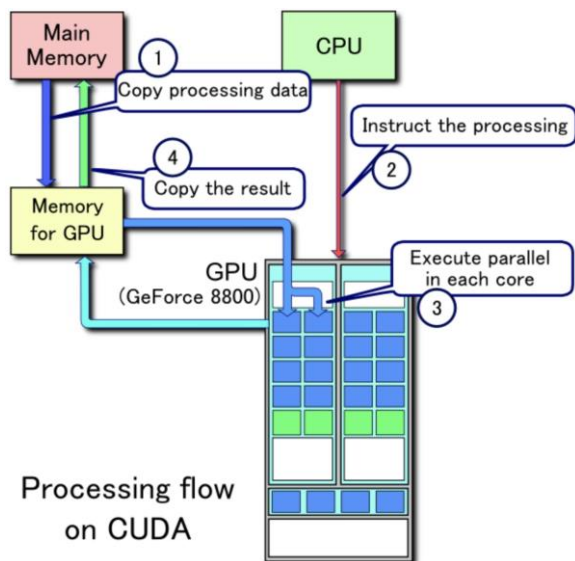
1.6.6 2019Q4-至今复盘



1.7 CUDA开启软硬件生态，形成护城河

- ◆ **GPU适用于处理大数据集，CUDA核是本质原因。** 最开始，GPU(图形处理单元)作为一种专用计算机处理器，可以满足实施高分辨率3D图形计算密集型任务的需求。到2012年，由于GPU已经发展成为高度并行的多核系统，让它具备了处理大量数据的能力。简而言之，CPU做的专注线性计算，GPU做的是并行计算(数据之间没有直接关系)，而本质的原因是CUDA核的不同，**CUDA核越多，计算性能越强**，而GPU的CUDA核数是CPU的上百倍，如AMD EPYC 7003系列7763核心数为64个，而英伟达A100 40GB核心数为6912个。
- ◆ **CUDA的本质是“软件定义硬件”，实现“软件调用硬件”。** CUDA是一种并行计算平台和应用程序编程接口(API)，允许软件使用特定类型的图形处理单元(GPU)进行通用目的的处理，称为通用图形处理单元计算(GPGPU)。CUDA提供了直接访问GPU虚拟指令集和并行计算元素的软件层，用于执行计算内核。CUDA支持的GPU还可以使用编程框架，通过将代码编译为CUDA来使用HIP。CUDA将从前多种不同的代码整合成了一气呵成的代码，这样极大的加快了开发模型的训练速度。可以简单理解，CUDA是英伟达实现软硬件适配的一种“类编译器”，将软件的代码转换成硬件汇编代码，CUDA是英伟达实现软硬件生态的护城河。

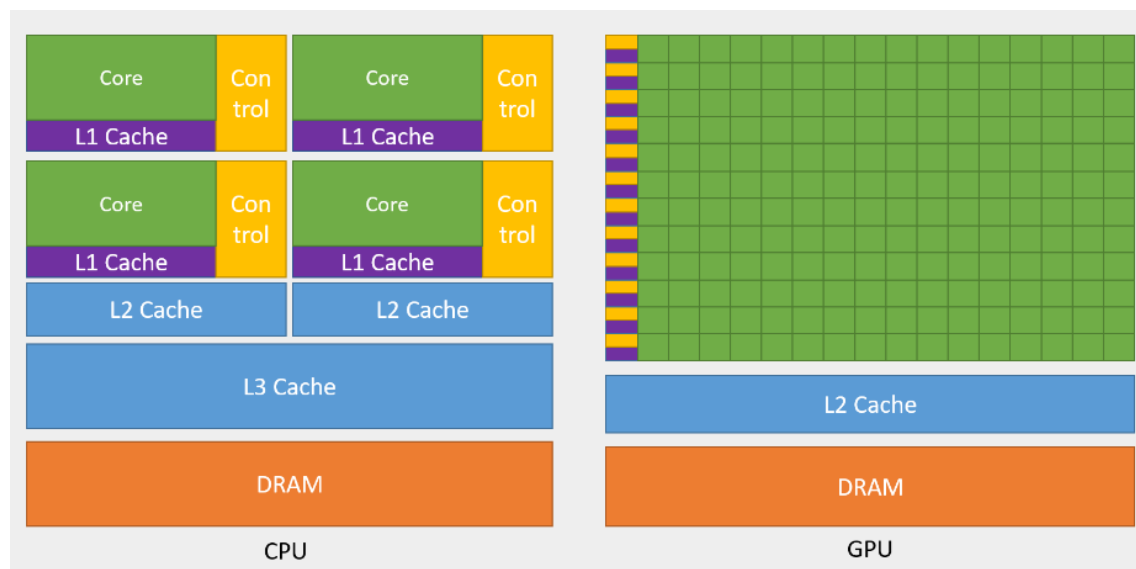
CUDA处理流程



CUDA处理流程：

- 1.将数据从驻内存复制到GPU内存
- 2.CPU启动GPU计算内核
- 3.GPU的CUDA内核并行执行计算
- 4.将生成的数据从GPU内存输送到内存

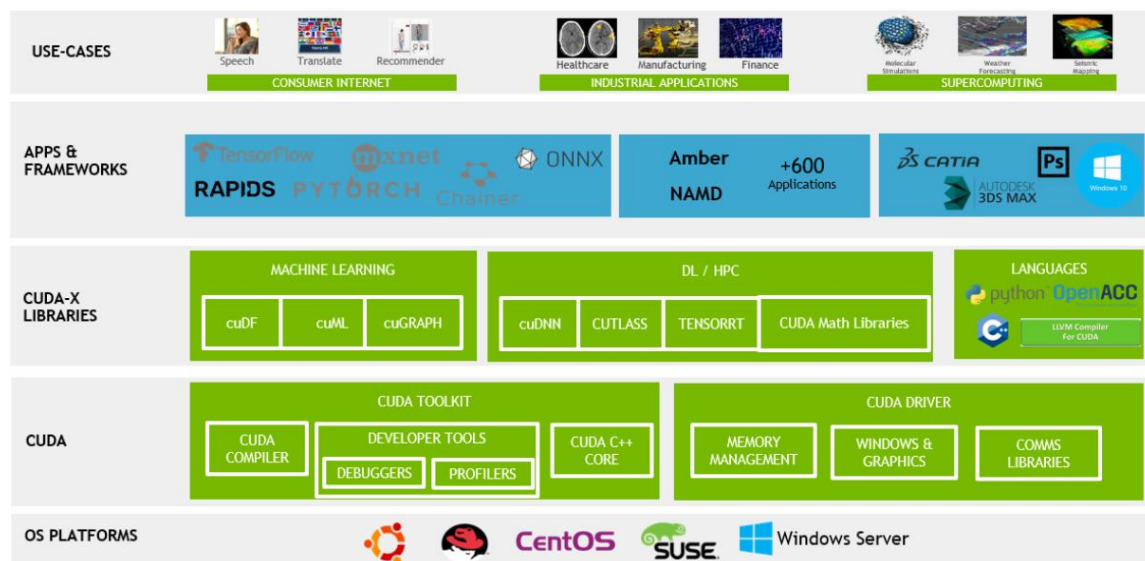
CPU和GPU计算资源差异



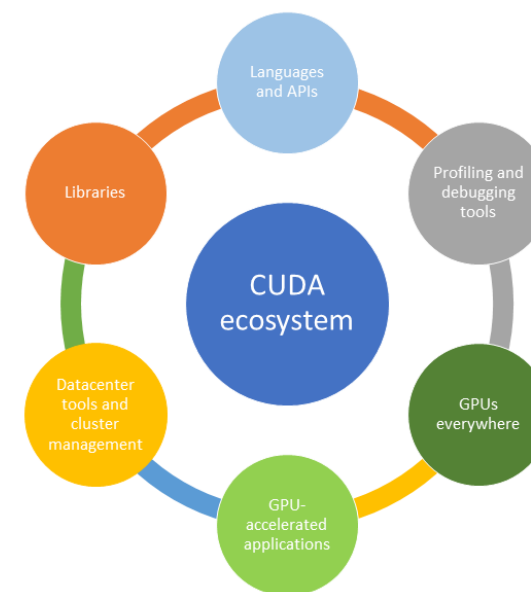
1.7 CUDA开启软硬件生态，形成护城河

- ◆ **CUDA是英伟达GPU产品线核心技术，构建高竞争壁垒。** CUDA是 NVIDIA 专为图形处理单元 (GPU) 上的通用计算开发的并行计算平台和编程模型，开发者能凭借CUDA利用 GPU 的强大性能显著加速计算应用。使用 CUDA 时，开发者可用主流语言（如 C、C++、Fortran、Python 和 MATLAB）进行编程，并通过扩展程序以几个基本关键字的形式来表示并行性。CUDA使得GPU可以用于更加广泛的航空航天、生物科学研究、机械和流体模拟及能源探索等领域的应用，**我们认为其拓宽了高算力模型的应用前沿，为公司构建起生态链的护城河。**
- ◆ **CUDA 生态系统发展迅速，已涵盖多种服务及解决方案。** NVIDIA 于 2006 年发布 CUDA，即首款用于 GPU 通用计算的解决方案。CUDA 充当 NVIDIA 各 GPU 系列的通用平台，因此客户可以跨 GPU 配置部署并扩展应用。目前CUDA 生态系统已涵盖软件开发工具、多种服务以及基于合作伙伴的解决方案，通过CUDA开发的数千个应用已部署到嵌入式系统、工作站、数据中心和云中的 GPU。

英伟达CUDA加速计算解决方案



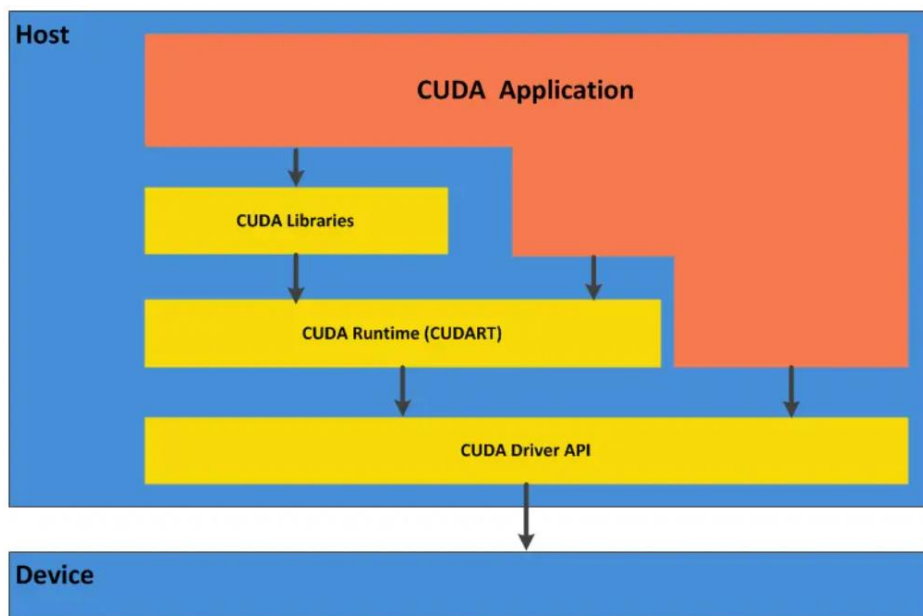
英伟达CUDA生态系统



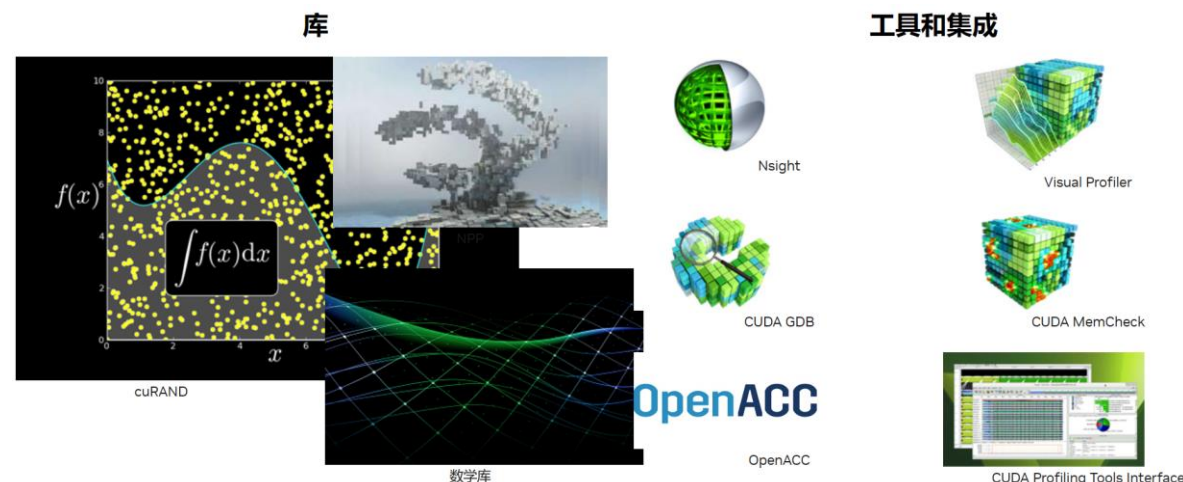
1.7 CUDA开启软硬件生态，形成护城河

- ◆ **CUDA助力加速计算及深度学习**：GPU通过图形应用程序的算法存在算法密集、高度并行、控制简单、分多个阶段执行等特征，英伟达引入的CUDA使GPU超越了图形领域。同时，CUDA的框架和库可以充分发挥GPU的并行计算能力，提供高效的矩阵运算、卷积运算等计算任务的实现，大大**简化深度学习的编程工作**，提高开发效率和代码质量。在经GPU加速的应用中，工作负载的串行部分在CPU上运行，而应用的计算密集型部分则以并行方式在数千个GPU 核心上运行，能够**大幅提升计算效率**。目前NVIDIA H100 GPU的CUDA数已达到14592个，远超AMD EPYC Genoa-X CPU的96个核心。
- ◆ **CUDA生态合作者规模翻倍增长**。根据英伟达2023财年年报，**目前有400万名开发者正在与CUDA合作**，而且规模还在不断扩大。英伟达通过12年的时间达到200万名开发者，在过去的两年半里该数字翻了一番。**目前CUDA的下载量已经超过了4000万次**。

CUDA软件架构



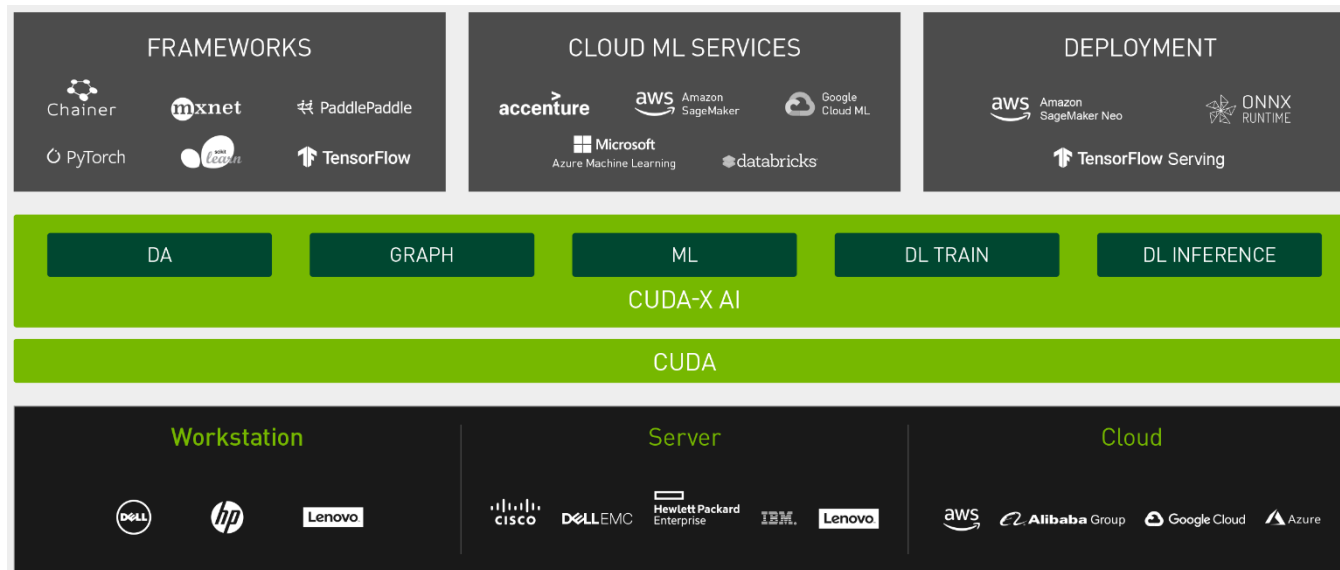
英伟达CUDA工具包



1.7 CUDA开启软硬件生态，形成护城河

- ◆ **CUDA X-AI加速库**：CUDA-X AI 是软件加速库的集合，这些库建立在 CUDA之上，提供对于深度学习、机器学习和高性能计算必不可少的优化功能，是针对数据科学加速的端到端平台。这些库与 NVIDIA Tensor Core GPU 配合工作，能够将机器学习和数据科学工作负载加速至高达50倍。CUDA-X AI 的软件加速库集成到所有深度学习框架和常用的数据科学软件中，且可以部署到多种设备内的 NVIDIA GPU 上，其中包括台式机、工作站、服务器、云计算和物联网 (IoT) 设备。CUDA-X AI 让开发人员能够提高工作效率，加速开发基于 AI应用程序的多步骤处理计算。
- ◆ **CUDA-X作为CUDA平台上集合层，开发人员可快速部署多领域常用库，加强CUDA软件计算平台性能，将应用层和算力层更好的适配。** CUDA-X AI已得到渣打银行、微软、PayPal、SAS和沃尔玛等顶尖公司所采用，已集成至主流深度学习框架中，如TensorFlow、PyTorch和MXNet。全球主要云服务提供商均在使用CUDA-X AI来加速自身云服务。全球八大计算机制造商宣布其数据科学工作站和服务器经优化后能够运行NVIDIA的CUDA-X AI库。

CUDA-X AI 生态系统图



1.8 英伟达今年发布多款AI产品，助力全球AI生态

- ◆ **2023年3月23日GTC会议，英伟达全新AI相关产品助力全球AI生态。** 1) 基础软件：推出全新加速库；2) 芯片方面：推出数据中心 Grace CPU，具备高能效、高运行速度等优势；3) 服务器：推出DGX超级计算机；4) **全新AI服务平台（DGX云与生成式AI服务）**，AI的“iPhone”时刻已经来临，**AI foundations 云服务**能够构建、改进和操作定制的大型语言模型和生成式 AI 模型，助力初创企业具备拥有生成式AI的能力，且已经具备多种生成式AI模型和相应案例。
- ◆ **平台实为模型和算力之间的“桥梁”，是AIGC或大模型生成的必备要素**，不论是数据库还是编译器，都需要通过平台来实现资源的合理分配以达到软硬件的最优组合，从而大幅提升模型效率。平台通过调用数据包来适配软硬件之间的结构，来达到模型的最优组合，从而提升模型乃至整个虚拟机的效率。

英伟达加速计算库



英伟达DGX H100



英伟达AI Foundations能力示意图



1.8 英伟达今年发布多款AI产品，助力全球AI生态

- ◆ **2023年5月30日COMPUTEX主题演讲，英伟达展示多款新系统、软件及服务，让生成式AI触手可及，革新了广告、制造、电信等行业**
- ✓ **DGX GH200 AI超级计算机**：由NVIDIA GH200 Grace Hopper超级芯片和NVIDIA NVLink Switch System驱动，相比上一代将NVLink带宽提升了48倍以上。据腾讯网消息，谷歌云、Meta和微软是首批有望接入DGX GH200探索其用于生成式AI工作负载能力的公司。英伟达还计划将DGX GH200设计作为蓝图提供给云服务提供商等。
- ✓ **模块化参考架构NVIDIA MGX**：能够灵活地兼容多代NVIDIA产品，制造商可以使用通用架构和模块化组件构建CPU和加速服务器。
- ✓ **NVIDIA Spectrum-X网络平台**：是全球首款专为AI网络打造的51Tb/s以太网交换机，提高了基于以太网AI云的性能与效率，与现有以太网的堆栈实现互通。单台交换机即可实现突破性的256个200Gb/s端口的连接，以支持AI云的增长和扩展。
- ✓ **ACE游戏开发版**：在Omniverse的基础上，“ACE游戏开发版”为语音、对话和角色动画提供优化的AI基础模型。其代工服务协助开发者微调游戏模型，然后通过 NVIDIA DGX Cloud，GeForce RTX PC 或现场加以部署，以实时进行推理。在客户端侧，NVIDIA和微软合作，将加强1亿台配备RTX GPU的PC性能，这些GPU中的Tensor Core可大幅提升400多个AI加速应用和游戏的性能。
- ✓ **NVIDIA Omniverse的虚实融合**：在数字广告领域，全球最大营销服务机构WPP正与NVIDIA一起在Omniverse Cloud上构建首个生成式AI内容引擎；在工业制造领域，企业可通过Omniverse和生成式AI 的API接口，构建工厂数字孪生。其他已运用相关产品的企业包括和硕、富士康、Techman Robot等

英伟达DGX GH200 AI超级计算机



英伟达“ACE游戏开发版”构建NPC场景



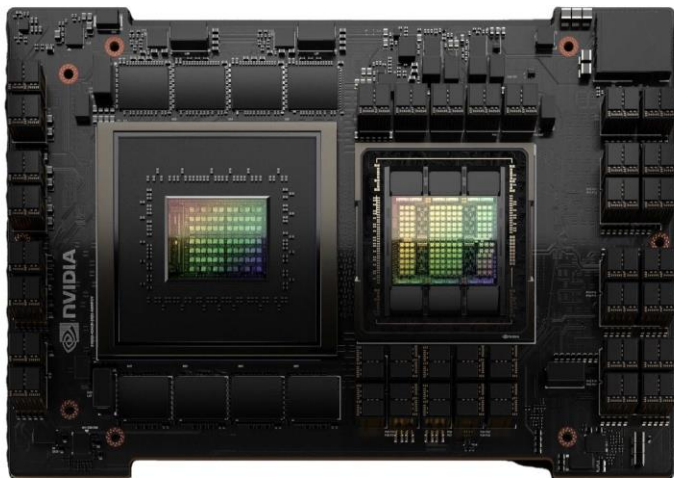
英伟达最新组件Metropolis for Factories



1.9 英伟达提供算力的方式: AI芯片、AI服务器、AI云

- ◆ **芯片方向**：发布应用于大规模人工智能和高性能计算应用程序的突破性加速CPU+GPU—NVIDIA Grace Hopper超级芯片，该超级芯片可以为运行TB级数据的应用程序提供高达10倍的性能提升，使科学家和研究人员有能力解决更为复杂的问题。
- ◆ **显卡方向**：NVIDIA TITIAN RTX显卡，它由Turing（图灵）架构提供支持，为用户的PC带来130 Tensor TFLOPs的性能，576个Tensor核心和24GB超快GDDR6内存，并且加入了加速AI和光线追踪的最新Tensor Core和RT Core技术。
- ◆ **服务器方向**：DGX超级计算机，配有 8 个 H100 GPU 模组，H100 配有 Transformer 引擎，旨在处理类似令人惊叹的 ChatGPT 模型，8 个 H100模组通过 NVLINK Switch 彼此相连，以实现全面无阻塞通信。
- ◆ **云服务方向**：NVIDIA DGX平台为企业AI而构建，将NVIDIA软件，基础设施和专业知识的精华结合在一个从云端到本地数据中心的现代统一AI开发解决方案中。DRX平台具有（1）**先进的人工智能开发平台**：除常规应用，包含用于训练模型、优化框架和加速数据科学软件库的NVIDIA AI Enterprise 软件（2）**注入NVIDIA AI 专业知识**：可直接访问NVIDIA DGXperts帮助优化企业的AI工作负责，以获得更高的投资回报率。（3）**前所未有的性能**：为AI基础设施提供清晰的可预测的成本模型。

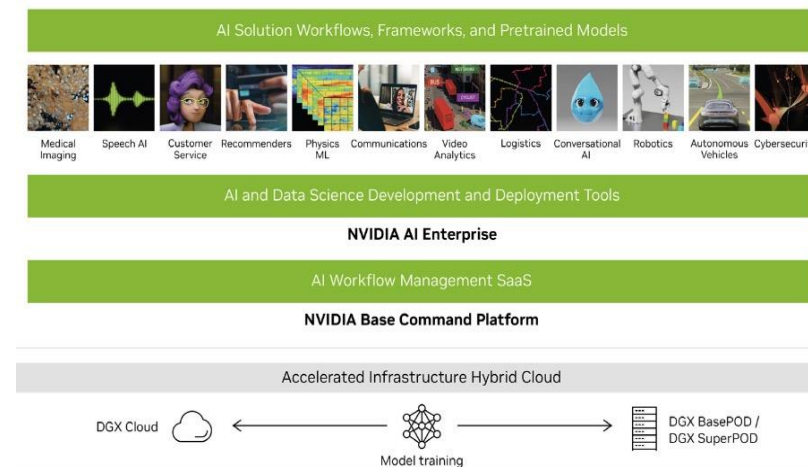
NVIDIA Grace Hopper超级芯片



NVIDIA TITIAN RTX 显卡



NVIDIA DGX云平台概念图



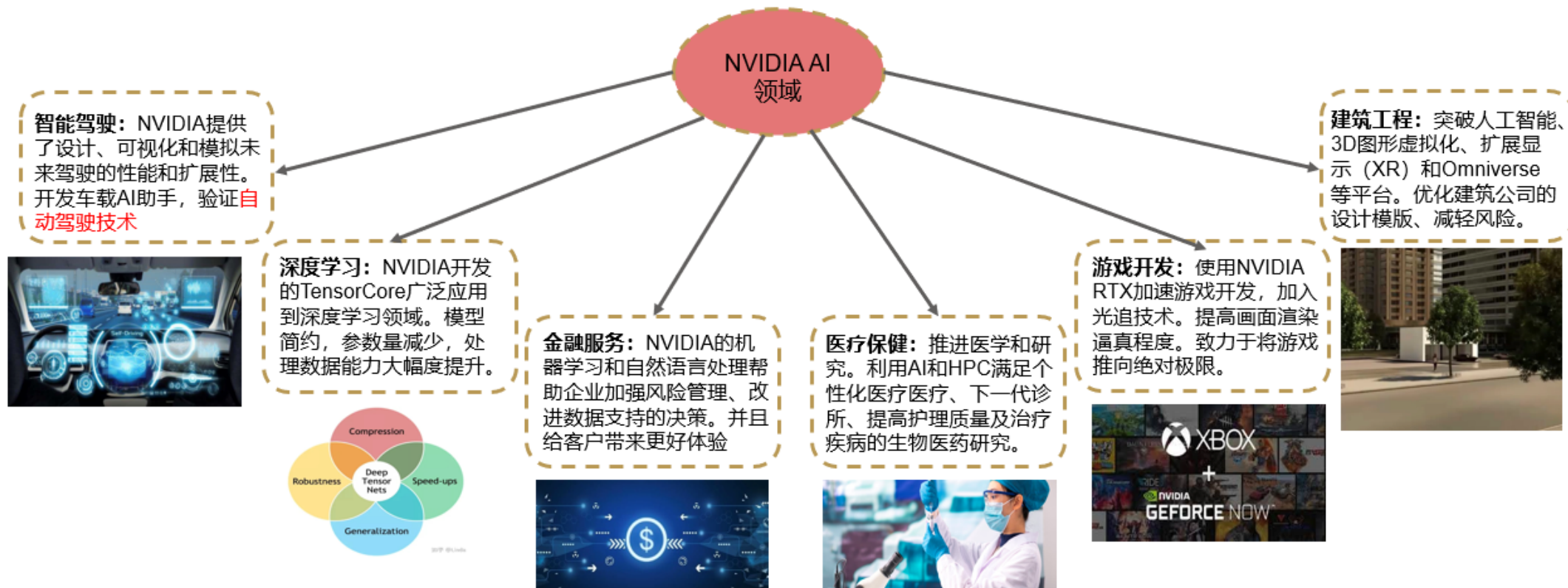


02 AI硬件自主可控势在必行

2.1 英伟达成功转型成全球AI硬件龙头

- ◆ **英伟达已经成功的从一家图形处理器公司转型为引爆人工智能的综合性硬件公司。** GPU技术在AI应用中发挥着重要的作用，其并行计算的能力处于领先地位。AI平台—NVIDIA DGX系统作为一套高性能计算解决方案，集成了英伟达的GPU和软件工具，让开发者更加便捷的构建、训练和部署AI。软件方面推出NVIDIA CUDA和NVIDIA cuDNN，提供了快速的神经网络训练和推断功能。英伟达的技术在AI领域被广泛应用于智能驾驶、深度学习、金融服务、医疗保健、游戏开发和建筑工程等各个领域。

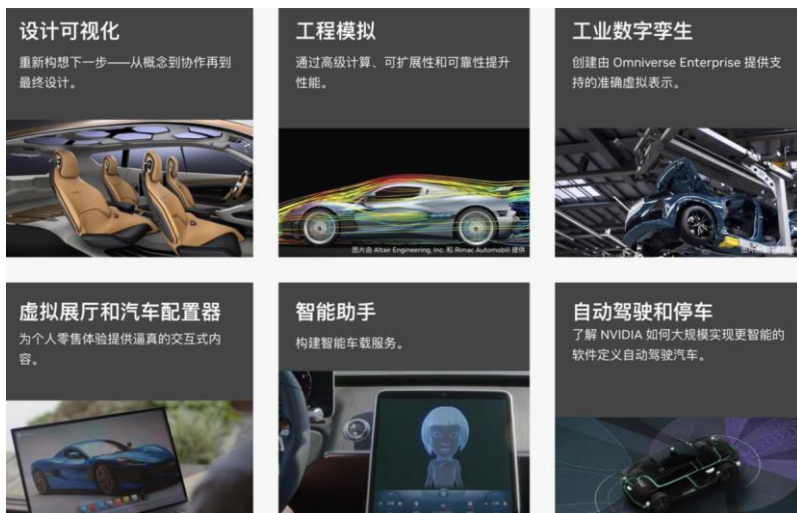
英伟达部分AI领域应用



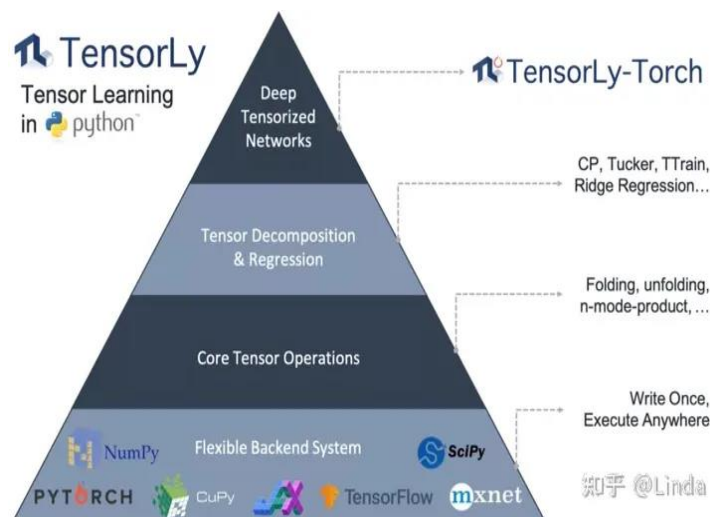
2.1 英伟达成功转型成全球AI硬件龙头

- ◆ **智能驾驶**：为设计者提供便利，借助基于NVIDIA Hopper和Ada Lovelace架构，让设计者拥有身临其境、实时、物理上准确的可视化效果。构建智能车载服务，DRIVE Concierge核心建立在DRIVE平台上，提供信息娱乐和游戏，充当乘车人的数字助理。自动驾驶和停车，Infrastructure用于数据社区管理和训练，DRIVE AGX平台使自动驾驶系统能够处理大量传感数据作出实时驾驶决策。
- ◆ **深度学习**：Tensor将矩阵推广到二维以上，广泛应用于深度神经网络特征到大数据处理。Tensor强大的性能，拥有简约模式、大幅减少的参数数量和更好的归纳方法。Tensor优秀的生态系统，Tensor具有高级的API，可以选择并无缝集成计算后端，例如Numpy、PyTorch、MXNet和Tensorflow等。
- ◆ **金融服务**：为银行业带来更智能安全的服务，GPU驱动的AI减少人为工作时间、加速风险计算、减少数字支付所带来的欺诈以及更为准确的推荐系统增强客户服务。带领保险公司超越传统的索赔管理，通过拥抱数字机会并采用先进的分析方法，AI能更快更准确的处理案件的索赔，很大程度上减少人为处理案件的错误率。

NVIDIA 智能驾驶部分应用方向



NVIDIA 深度学习部分应用方向



NVIDIA 金融服务部分应用方向



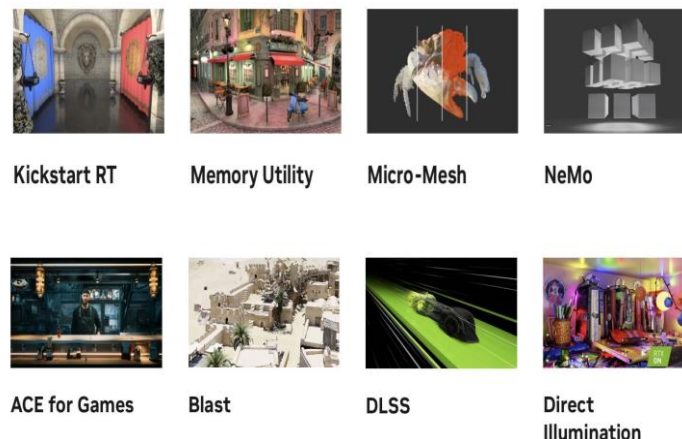
2.1 英伟达成功转型成全球AI硬件龙头

- ◆ **医疗保健**：加快药物的研发速度，NVIDIA推出NVIDIA Clara Discovery的AI加速计算平台，可支持化学信息学研究、蛋白质结构预测、药物筛选和分子动力学等研究。推出大语言模型BioNeMo可用于训练和部署大型生物分子语言模型。更新医疗设备，运用NVIDIA强大的GPU的分析和成像技术，打造用于影像诊断、数字手术和病人监护等多种医疗设备，大大提高治疗精度和效率。
- ◆ **游戏开发**：为游戏开发提供NVIDIA游戏技术，例如，有生成式AI提供支持的智能游戏角色铸造平台—NVIDIA ACE、AI神经图形技术—NVIDIA DLSS、可扩展的多GPU实时推理开发平台（用于3D方针和设计）—NVIDIA Omniverse平台等。这些平台帮助游戏开发这一创纪录的速度构建逼真且精确的游戏。
- ◆ **建筑工程**：加速设计流程，提高工作效率，NVIDIA RTX驱动的工作站通过实时光线追踪、虚拟现实、工程模拟和支持AI的应用程序增强建筑和基础设施设计工作流程。Omniverse平台克服了设计软件之间操作冲突的问题。NVIDIA Quadro GPU可快速分析复杂的流体场景，使用NVLink技术加速，可以大大缩短设计、建模、模拟和检查等过程的时间。

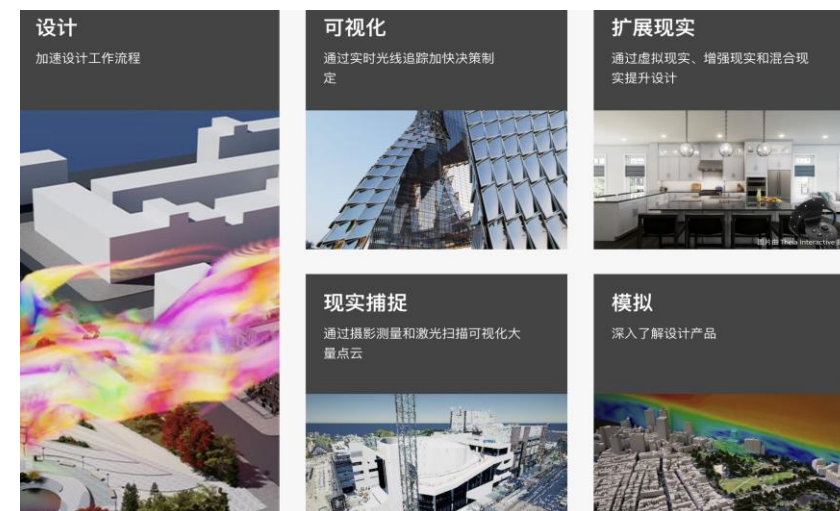
NVIDIA 医疗保健部分应用方向



NVIDIA 游戏开发部分应用方向



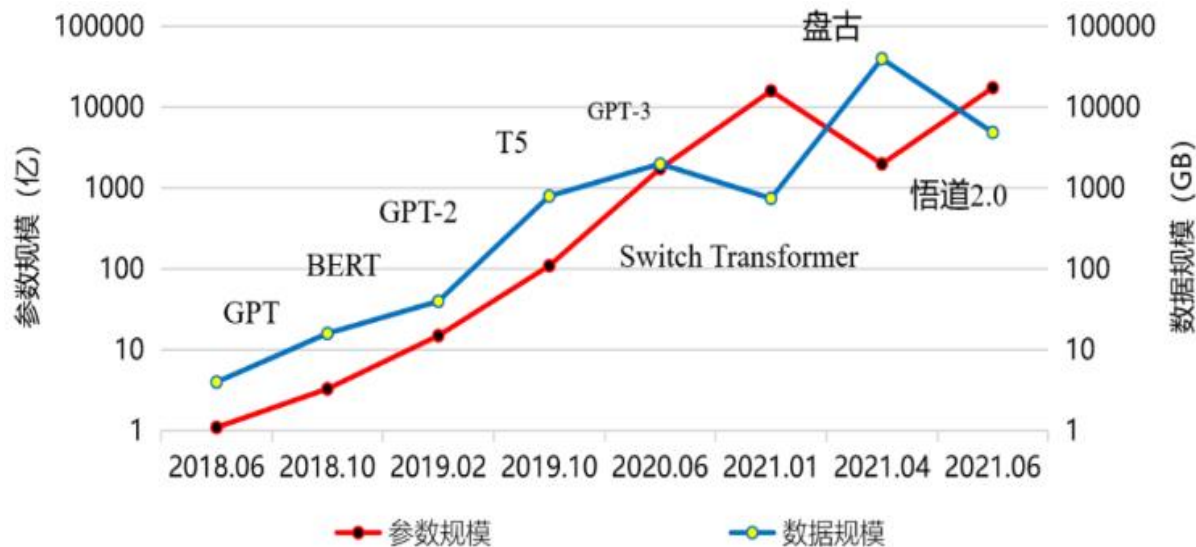
NVIDIA 建筑工程部分应用方向



2.2.1 再三强调，大模型背景下算力势必迎来爆发

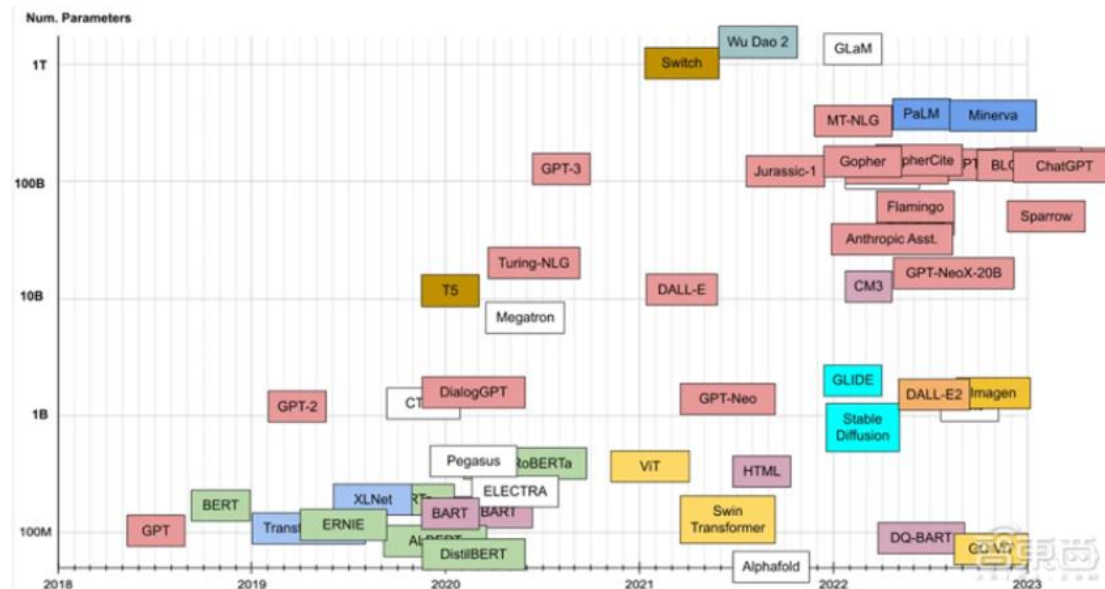
- ◆ **ChatGPT开启算力军备赛：**我们已经在《ChatGPT：百度文心一言畅想》中提到数据、平台、算力是打造大模型生态的必备基础，且算力是训练大模型的底层动力源泉，一个优秀的算力底座在大模型(AI算法)的训练和推理具备效率优势；同时，我们在《ChatGPT打响AI算力“军备战”》中提及算力是AI技术角逐“入场券”，其中AI服务器、AI芯片等为核心产品；此外，我们还在《ChatGPT，英伟达DGX引爆AI“核聚变”》中提到以英伟达为代表的科技公司正在快速补足全球AI算力需求，为大模型增添必备“燃料”。
- ◆ **大模型参数呈现指数规模，引爆海量算力需求：**根据财联社和OpenAI数据，ChatGPT浪潮下算力缺口巨大，根据OpenAI数据，模型计算量增长速度远超人工智能硬件算力增长速度，存在万倍差距。运算规模的增长，带动了对AI训练芯片单点算力提升的需求，并对数据传输速度提出了更高的要求。根据智东西数据，过去五年，大模型发展呈现指数级别，部分大模型已达万亿级别，因此对算力需求也随之攀升。

大模型参数数量和训练数据规模快速增长



资料来源：新浪，智东西，可创办日报，华

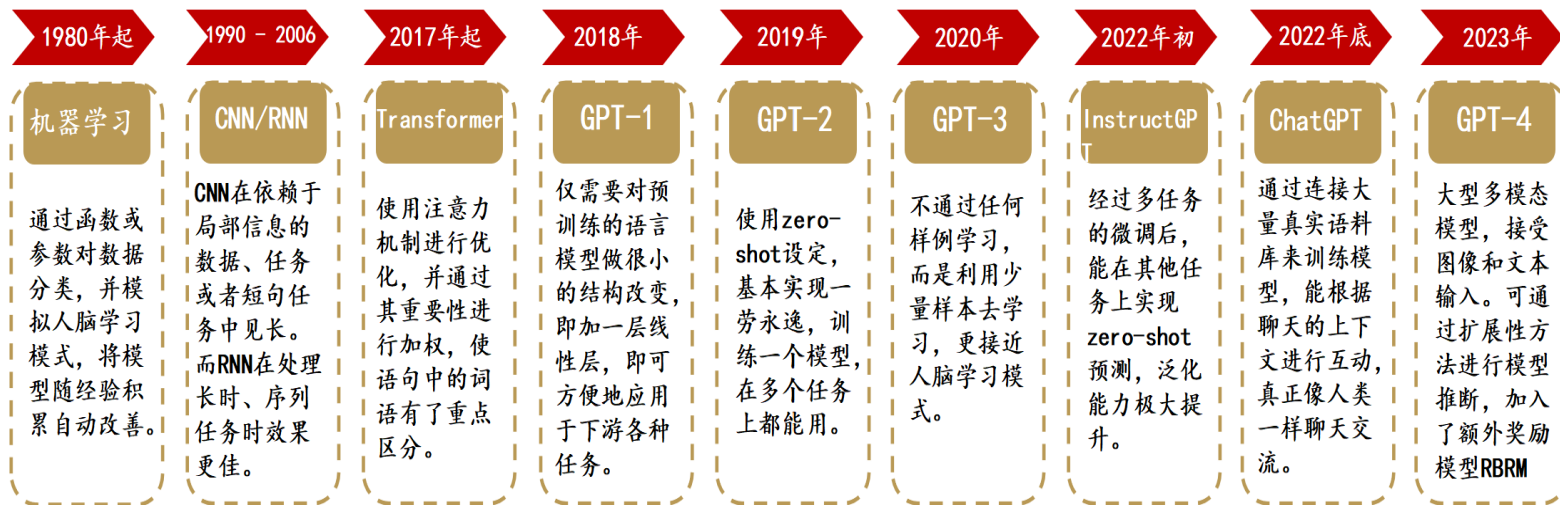
近年大模型的参数规模增长趋势



2.2.2 GPT-4、多模态正逐级点燃AI领域

- ◆ **GPT-4，多模态算法，AI大模型又一跨越里程碑式的巨作。** GPT-4于北京时间2023年3月15日横空出世，为多模态大型语言模型，支持图像和文本输入，以文本形式输出扩写能力增强，能处理超过 25000 个单词的文本；更具创造力，并且能够处理更细微的指令。
- ◆ **多模态，AI的旷世之作：**过去辅助式AI只注重于以重模态，例如图像、视频、语言等，多模态旨再通过机器学习方法处理和理解多源模态信息的能力，例如图像、视频、音频、语义相结合。
- ◆ **DALL·E2，文生图震撼发布：**DALL·E2是OpenAI旗下产品，可以根据文字描述创建原创、逼真的图像和艺术作品。它可以组合概念、属性和样式，我们认为文生图功能对于传统图型生成工具具有颠覆性。
- ◆ **Runway，AI生成视频，多模态的下一站：**Gen-2模型震撼发布，多模态人工智能技术实现了从AI文生图到AI文生视频的跨越，实为解放生产力的双手，我们认为其功能颠覆摄影、传媒、电影制作等行业。

GPT算法的发展历程



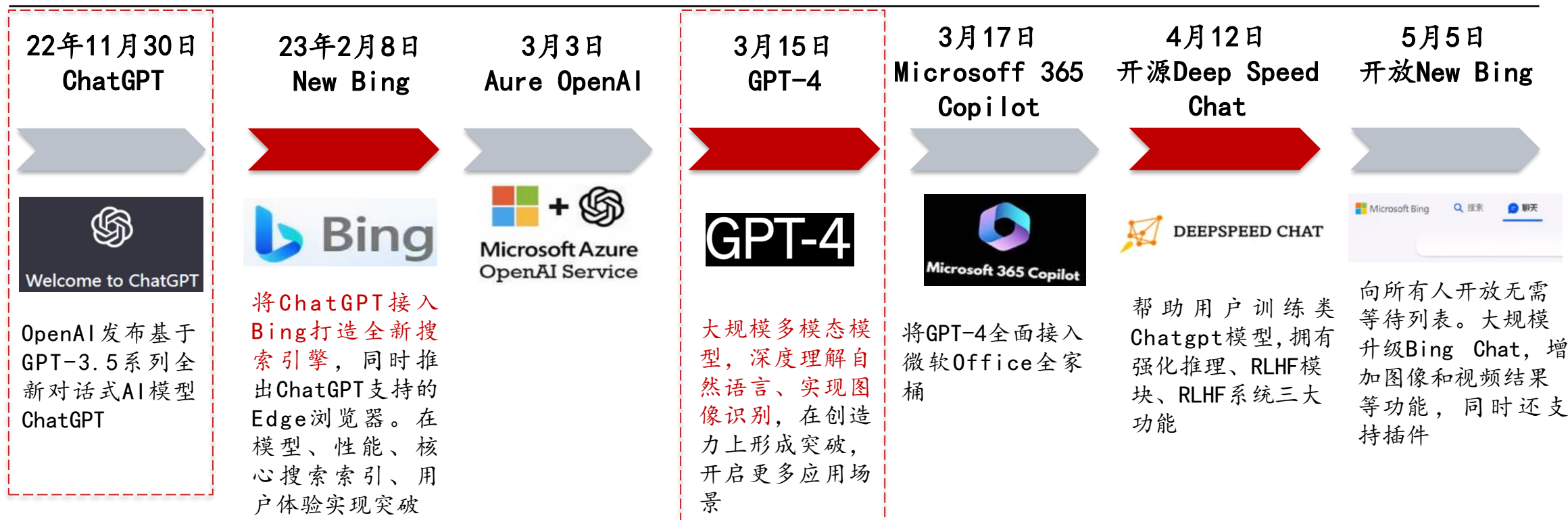
Gen-2文生视频示意图



2.2.3 大模型赋能千行百业，巨头指引——微软

- ◆ **引领AI浪潮，大模型技术里程碑：**22年11月OpenAI发布基于GPT-3.5系列全新对话式AI模型ChatGPT，具备人工智能算法的迭代升级跨时代意义；今年2月将ChatGPT接入Bing，重新定义搜索引擎；3月多模态大型语言模型GPT-4震撼发布，在“理解+创造”上展现的能力是AI算法历史的里程碑。
- ◆ **微软已将大模型能力赋能自身产品：**例如个人计算New Bing，Xbox等、应用软件Office365 Copilot、Dynamics 365 Copilot等一系列办公软件与工业软件中，势必成为解放生产力的双手，如果说产品是AI赋能、企业开启第二轮业绩增长曲线的“流量入口”，那么算力即是大厂开启算力争夺战的“入场券”。

微软与大模型及相关产品时间概述图



2.2.4 大模型赋能千行百业，巨头指引——谷歌

- ◆ **紧追OpenAI，AI竞速升级**：面对OpenAI陆续推出的GPT系列火爆全球，谷歌步步紧追，今年2月和3月分别推出对标ChatGPT的Bard和史上最大多模态具身视觉语言模型PaLM-E；并于5月11日正式打响“反击战”，发布大语言模型PaLM 2直指GPT-4痛点，同时在25+款应用上接入AI。我们预计谷歌和微软的竞速将持续升级，有望推动AI技术指数级发展。
- ◆ **AI大战全面升级，更加突出算力的重要性**：以微软和谷歌为代表，全球科技巨头已经进入AI“争夺战”的重要时间点，以大模型赋能自身产品开启新一轮成长周期，而大模型短期呈现高需求的模式势必对底层算力提出更严峻的需求。

谷歌与大模型及相关产品时间概述图



2.2.5 国内AI应用已有雏形，下游应用领域不断扩大

- ◆ **AI赋能千行百业，下游应用领域不断扩大。**我们认为国内AI产品赋能已有雏形，目前“AI+”的应用方面，在教育、办公软件、金融、电商和互联网传媒等行业已有一些企业推出产品和落地解决方案，而大模型应用领域的增多，势必对算力提出更高需求。

国内大模型应用场景



2.2.5 国内大模型百家争鸣，自研进度加速，技术迭代超预期

- ◆ **大模型短期百家争鸣，“自研大模型热”仍将持续：**国产大模型于今年3、4月份密集发布，国产自研AI大模型进入“百花齐放”阶段。原因如下
 - ✓ **1、AI变革中机遇挑战并存：**对于早布局AI的中小厂商，大模型打开了弯道超车的机会窗口；头部厂商若不加紧布局或将在未来面临较大竞争压力
 - ✓ **2、未来AIOS会成为重要流量入口：**长期技术层面下各家大模型趋同，差异化将体现于渠道禀赋、商业化能力等方面。对于头部公司，立足已有禀赋布局AI大模型，将是未来重要增收手段。
- ◆ **目前，国内大模型自研进度明显加速，技术迭代超预期：**我们认为，大模型仍然处于百家争鸣的状态，国内的科技巨头兼在研发通用或垂类的大模型，通用类大模型例如百度、阿里、腾讯、华为，垂类大模型例如航天宏图、三六零等企业，且自研速度明显加速，例如讯飞星火大模型“三步走”实现技术升级，10月对标ChatGPT。而大模型短期的爆发势必会带来算力的高需求。

国内大模型发布时间

公司	最新发布时间	大模型	参数规模	模型特点
网易	-	伏羲	110亿	NLP大模型、多模态大模型
云从科技	研发	行业精灵	百亿-千亿	行业大模型
京东	待发布	ChatJD	千亿级	新一代产业大模型
字节跳动	预计今年	自研大模型		在语言和图像两种模态上发力
科大讯飞	2023. 5. 6	1+N认知智能大模型		
知乎	2023. 4. 13	知海图AI	10亿级	NLP大模型
阿里巴巴	2023. 4. 11	通义千问	超10万亿	大语言模型
	2021. 3	M6	10万亿	多模态大模型
昆仑万维	2023. 4. 10	天工3. 5	百亿级	多模态大模型、代码大模型
商汤科技	2023. 4. 10	日日新	1300亿	商量：NLP大模型、代码大模型
			超10亿	AIGC文生图、数字人物生成等
360	2023. 4. 9	360智脑		
华为	2023. 4. 8	盘古CV	超30亿	NLP 大模型、CV 大模型、科学计算大模型（气象）
		盘古NLP	千亿级	
百度	2023. 3. 16	文心大模型	2600亿	NLP大模型、CV大模型、跨模态大模型、生物计算大模型
澜舟科技	2023. 3. 14	孟子Mchat	百亿-千亿	NLP大模型
复旦大学	2023. 2. 20	MOSS	175亿	NLP大模型
腾讯	2022. 4	混元AI 大模型	万亿级	NLP大模型、CV大模型、多模态大模型
浪潮信息	2021. 9	源1. 0	2457亿	NLP大模型

2.3.1 美国连续发动对我国高科技行业制裁，自主可控势在必行

- ◆ 自2018年来，美国通过多种制裁手段，严重限制我国高科技领域发展。
- ✓ 根据美国提出的《国家量子倡议法》（2018）、《美国人工智能发展倡议》（2019）以及《出口管制改革法案》（2018）等相关法案和计划，美国已对我国在14类新兴和基础技术领域，包括AI技术、人工智能芯片、机器人、量子计算、脑机接口和先进材料等方面实施出口和技术合作限制措施。
- ✓ 2022年8月，拜登正式签署《芯片与科学法案》，其中提到禁止接受联邦奖励资金的企业，在中国扩建或新建先进半导体的新产能；同年10月，美国政府进一步紧缩半导体产品对华出口的政策，主要包括限制英伟达、AMD等公司向中国出售高算力人工智能芯片；限制应用材料、泛林、科磊等美国设备厂商向任何中国公司出售半导体设备；将31家中国公司、研究机构及其他团体列入所谓“未经核实的名单”（UVL清单），限制它们获得某些受监管的美国半导体技术能力

美国制裁、限制事件汇总

时间	事件
2018/11/1	美国商务部发布涉及人工智能和机器学习技术、先进计算技术、数据分析技术等14项新兴和前沿技术的对华出口管制框架
2019/5/1	“布拉格5G安全大会”召开：联合发布了“布拉格提案”，该提案从政策、安全、技术、经济四个方面探讨如何排除中国5G技术产品。
2020/1/1	特朗普政府发布限制人工智能软件出口新规，应用于智能化传感器、无人机和卫星的目标识别软件都在限制范围之内。
2020/2/1	推动42个加入《瓦森纳协定》的国家扩大半导体对华出口管制范围，旨在加强防备相关技术外流到中国。
2020/2/1	美国商务部更新《出口管制条例》，将“用于自动分析地理空间图像的软件”列入对华管制清单中，应用于智能化传感器、无人机、卫星和其他自动化设备的目标识别软件。
2020/5/1	美国宣布将加入七国集团“人工智能全球合作伙伴组织”，力图以霸权力量主导构成不利于中国的全球人工智能管理规则，限制中国人工智能技术发展。
2020/5/1	发起七国集团（G7）加澳大利亚、韩国和印度的“D10俱乐部”（D10 Club），以减少对中国电信技术的依赖。
2020/10/1	美国家人工智能安全委员会提出通过多边合作、数字联盟等形式与北约、印度等建立国际联盟，推广美国标准和规则，形成对我人工智能的封锁围堵之势。
2022/7/1	美国半导体设备制造商收到美商务部的通知，拟要求禁止向中国大陆供应用于14nm或以下芯片制造的设备。
2022/8/1	美国总统拜登正式签署《芯片与科学法案》，以补贴美国的半导体产业。关于补贴资助对象资格的内容里，明确写到，禁止接受联邦奖励资金的企业，在中国等对美国国家安全构成威胁的特定国家扩建或新建某些先进半导体的新产能，期限为10年，违反禁令或未能修正违规状况的公司，可能需要全额退还联邦补助款。
2022/9/1	美国两大芯片制造巨头英伟达（NVIDIA）与AMD同时发布公告，声称均已接到美国拜登政府下达的最新命令，要求停止向中国出口用于人工智能的最先进芯片。制裁主要针对的两个芯片是Nvidia A100和H100图形处理单元以及AMD的MI250 人工智能芯片。
2022/10/1	BIS修订《出口管理条例》：美国从多方面加强对出口到中国的半导体的管制措施。新的管控措施主要涉和先进计算及半导体制造业以及超级计算机和半导体最终用途。

2.3.2 美国限制高端芯片流入中国，严重干扰国内大模型发展生态

- ◆ **美国政府禁止英伟达、AMD向中国出口用于人工智能的顶级计算芯片。**
- ✓ 根据钛媒体，2022年9月，美国商务部宣布限制英伟达（NVIDIA）和AMD等美国公司向中国出口先进计算机图像处理器（GPU），该禁令主要限制了英伟达的A100、H100高端芯片以及AMD的MI250出口中国，目的是瞄准国内先进计算进行遏制，影响国内人工智能领域发展。
- ✓ **浪潮集团被加入“实体清单”**：根据钛媒体，2023年3月美国商务部发布浪潮被加入被加入美国“实体清单”，限制了美国科技公司对浪潮的技术、产品支持。浪潮的服务器业务在CPU、GPU等关键芯片技术商高度依赖外国厂商，此外，截至2022年末，浪潮服务器及部件占总营收99.17%，若此次制裁被严格落实，其服务器业务将严重停滞。

英伟达H100、A100芯片



DGX H100

AI supercomputer optimized for large generative AI and other transformer-based workloads.



DGX A100

AI supercomputer delivering world-class performance for mainstream AI workloads.

浪潮AI服务器



2.3.3 英伟达应对制裁，提出特供版A800芯片

- ◆ **A800是A100的下位替代版。**
 - ✓ 根据快科技，在美国限制英伟达向中国出售高算力芯片A100、H100后，英伟达发布公告，确认发布新款中国特供版A800 GPU芯片来替换A100，以满足制裁政策。A800完全符合美国政府有关出口管制的测试。
- ◆ **A800在带宽性能方面劣于A100。**
 - ✓ A800芯片的数据传输速率为400GB/s，低于A100芯片的600GB/s，而其他参数变化不大。这也说明A800相比于A100在整体通信带宽性能上低了33%左右，影响了多卡互联性能，但是单卡性能保持不变。在一定程度上，这种削弱会导致在AI大模型训练上消耗更长的时间。
- ◆ **英伟达表示会继续推出H100的替代版H800。**
 - ✓ 根据快科技，2023年3月21日，英伟达在GTC 2023春季图形大会上，NVIDIA近日宣布为中国市场开发了第二个特供版H800，该产品是在已有的H100基础上进行了调整，以符合美国政府的规定。

A100、A800对比

NVIDIA A100 TENSOR CORE GPU SPECIFICATIONS
(SXM4 AND PCIE FORM FACTORS)

	A100 80GB PCIe	A100 80GB SXM
FP64	9.7 TFLOPS	
FP64 Tensor Core	19.5 TFLOPS	
FP32	19.5 TFLOPS	
Tensor Float 32 (TF32)	156 TFLOPS 312 TFLOPS*	
BFLOAT16 Tensor Core	312 TFLOPS 624 TFLOPS*	
FP16 Tensor Core	312 TFLOPS 624 TFLOPS*	
INT8 Tensor Core	624 TOPS 1248 TOPS*	
GPU Memory	80GB HBM2e	80GB HBM2e
GPU Memory Bandwidth	1,935GB/s	2,039GB/s
Max Thermal Design Power (TDP)	300W	400W***
Multi-Instance GPU	Up to 7 MIGs @ 10GB	Up to 7 MIGs @ 10GB
Form Factor	PCIe dual-slot air cooled or single-slot liquid cooled	SXM
Interconnect	NVIDIA® NVLink® Bridge for 2 GPUs: 600GB/s ** PCIe Gen4: 64GB/s	NVLink: 600GB/s PCIe Gen4: 64GB/s
Server Options	Partner and NVIDIA-Certified Systems™ with 1-8 GPUs	NVIDIA HGX™ A100-Partner and NVIDIA-Certified Systems with 4, 8, or 16 GPUs NVIDIA DGX™ A100 with 8 GPUs

NVIDIA A800 TENSOR CORE GPU SPECIFICATIONS
(SXM4 AND PCIE FORM FACTORS)

	A800 40GB PCIe	A800 80GB PCIe	A800 80GB SXM
FP64	9.7 TFLOPS		
FP64 Tensor Core	19.5 TFLOPS		
FP32	19.5 TFLOPS		
Tensor Float 32 (TF32)	156 TFLOPS 312 TFLOPS*		
BFLOAT16 Tensor Core	312 TFLOPS 624 TFLOPS*		
FP16 Tensor Core	312 TFLOPS 624 TFLOPS*		
INT8 Tensor Core	624 TOPS 1248 TOPS*		
GPU Memory	40GB HBM2	80GB HBM2e	80GB HBM2e
GPU Memory Bandwidth	1,555GB/s	1,935GB/s	2,039GB/s
Max Thermal Design Power (TDP)	250W	300W	400W***
Multi-Instance GPU	Up to 7 MIGs @ 5GB	Up to 7 MIGs @ 10GB	Up to 7 MIGs @ 10GB
Form Factor	PCIe (dual-slot air cooled or single-slot liquid cooled)		SXM
Interconnect	NVIDIA® NVLink® Bridge for 2 GPUs: 400GB/s ** PCIe Gen4: 64GB/s		NVLink: 400GB/s PCIe Gen4: 64GB/s
Server Options	Partner and NVIDIA-Certified Systems™ with 1-8 GPUs		NVIDIA HGX™ A800-Partner and NVIDIA-Certified Systems with 4 or 8 GPUs

2.4.1 政策端持续发力，加速推动国产自主可控进程

- ◆ **政府聚焦人工智能产业，发布多条政策助力AI发展。**
 - ✓ 2023年5月30日，北京市发布《北京市加快建设具有全球影响力的人工智能创新策源地实施方案（2023-2025年）》和《北京市促进通用人工智能创新发展的若干措施》。
 - ✓ **工作方向主要瞄准：**突破基础理论，引领关键核心技术创新，强化可信人工智能技术；推动国产AI芯片突破，研发通用高算力训练芯片、低功耗边缘端芯片和创新架构；加强自主开源深度学习框架研发，实现软硬件深度协同；提升算力供给能力，建设公共算力中心，实施算力伙伴计划；加强公共数据开放共享，推动数据融合创新；构建高效协同的大模型技术产业生态，加强人工智能企业梯度培育，强化企业多维服务。构建人工智能生态系统。

北京市人民政府官网最新政策界面



北京政府关于AI政策概述

时间	部门	政策	主要内容
2023. 5. 30	北京市人民政府	《北京市加快建设具有全球影响力的人工智能创新策源地实施方案（2023-2025年）》	1. 夯实人工智能底层基础，聚焦突破人工智能关键技术。2. 积极引导国产大模型研发应用国产人工智能芯片，加速提升算力供给的国产化率。3. 深化国产芯片部署应用，推动自主可控软硬件算力生态建设
2023. 5. 30	北京市人民政府	《北京市促进通用人工智能创新发展的若干措施》	1. 推动算力基础建设，针对算力需求，打造多云算力调度平台。2. 优化相关数据要素质量，助力大模型训练。3. 构筑大模型等通用人工智能技术体系，推动下游AI创新场景应用。

2.4.1 政策端持续发力，加速推动国产自主可控进程

- ◆ 加快推动人工智能高质量发展、创新应用场景。
- ◆ 2023年5月31日，深圳市政府发布《深圳市加快推动人工智能高质量发展高水平应用行动方案（2023—2024年）》。
- ✓ **工作方向主要瞄准：**建设城市级智能算力平台，打造大湾区智能算力枢纽，建设企业级智能算力平台；加强科技研发攻关，支持创新产品研发；规划建设产业集聚区，大力培育企业梯队，依托鹏城云脑搭建城市级人工智能生态孵化平台，为中小企业提供低成本智能算力资源；推进人工智能产业发展和应用，包括搭建供需对接平台、推进公共服务和城市治理的人工智能应用，培育企业梯队，建设产业集聚区，以及推动各行业的人工智能应用和创新发展；优化数据提供和高技术人员构成；加强组织领导，并成立对应的工作专班。
- ◆ 北深两地政策发布有望加速我国人工智能发展，推动算力供给环节国产化替代进程。
- ✓ 1) 算力是AI发展的基础，决定处理数据的能力和模型的性能，但是在多个环节仍然受制于海外制裁、存在技术“卡脖子”等问题。2) 关键技术自主可控仍然是国内发展AI的重中之重，在当前国际形势下，只有大力发展自主可控才可以保障数据安全和占据主动权。

深圳市人民政府官网最新政策界面



5月31日，深圳正式印发《深圳市加快推动人工智能高质量发展高水平应用行动方案（2023—2024年）》（以下简称《行动方案》），同步发布首批“城市+AI”应用场景清单，统筹设立规模1000亿元的人工智能基金群，以最充足的算力、最大的政策支持、最优的产业生态、最好的人才环境、最丰富的场景应用，积极打造国家新一代人工智能创新发展试验

深圳政府关于AI政策概述

时间	部门	政策	内容
2023年5月30日	中共深圳市委办公厅、深圳市人民政府办公厅	《深圳市加快推动人工智能高质量发展高水平应用行动方案（2023—2024年）》	1. 加强智能算力集群供给。2. 聚焦核心技术和产品创新能力。3. 增强产业聚集。4. 优化下游应用。5. 培育高质量数据要素市场，汇集高水平研发人员。

2.4.1 政策端持续发力，加速推动国产自主可控进程

◆ 针对算力产业，我国激励政策陆续出台。

- ✓ 当前我国已进入《新型数据中心发展三年行动计划（2021-2023年）》落地见效的关键年。《行动计划》主要目标为用3年时间，基本形成布局合理、技术先进、绿色低碳、算力规模与数字经济增长相适应的新型数据中心发展格局。到2023年底，全国数据中心机架规模年均增速保持在20%左右，平均利用率力争提升到60%以上，总算力超过200 EFLOPS，高性能算力占比达到10%。

◆ 各地全力保障数字基础设施建设，积极带动关联产业集聚发展。

- ✓ 在第七届世界智能大会上，中国电子董事长曾毅表示没有强大的算力，新一代人工智能将是无本之木。此外，5月12日，《北京市促进通用人工智能创新发展的若干措施（2023-2025年）（征求意见稿）》，在“加强算力资源统筹供给能力”等5个方面提出21项具体措施推动人工智能创新落地。同时，在5月19日，北京市启动通用人工智能产业创新伙伴计划推动大模型产业加速落地，该计划提出八大任务作为支撑，分别为加快满足近期迫切算力需求、提升中长期算力供给能力、推出一批高质量训练数据、谋划建设国家级数据训练基地、大模型应用创新标杆试点工程、推动大模型赋能千行百业等。

工信部印发相关计划



北京市通用人工智能产业创新伙伴计划成员名单（第一批）



2.4.2 产业端积极响应，智能算力建设持续提速

- ◆ 2023年年4月17日国家超算互联网联合体成立，算力建设持续提速。
- ◆ 科技部高新司2023年4月17日在天津组织召开国家超算互联网工作启动会，会议发起成立了国家超算互联网联合体。超算互联网是用互联网思维运营超算，将全国众多超算中心通过算力网络连接起来，构建一体化算力服务平台，解决当前亟待突破的现有单体超算中心运营模式，以应对算力设施分布不均衡、接口不统一、应用软件自主研发和推广不足等问题。
- ◆ “超算”向“智算”跨越，“AI+”时代步入“+AI”时代。
- ✓ 区别于超算中心，智算中心立足于赋能产业，可为大规模AI算法和模型研究形成条件支撑，主要支持人工智能与传统行业的融合应用。由于利用超算系统完成人工智能计算任务的成本高、效率低，我们认为从通用算力建设过渡到专用算力是大势所趋。

国家超算互联网正式启动



资料来源：中华人民共和国网，国务院，国家

中国智算相关政策及产业部署

时间	政策/产业部署	发布主体	内容
2023年1月	《智能计算中心创新发展指南》	国家信息中心	提出构建智算中心的“四化”技术路线：以算力基建化为主体，使得AI算力成为城市的公共基础资源，供政府、企业、公众按需使用；以算法基建化为引领，服务模式从提供算力为主向提供“算法+算力”转变；以算法基建化为引领，以低代码甚至无代码开发的模式，为用户提供使用便捷的智能算力；以设施绿色化为支撑：通过采用液冷技术等节能降碳技术
2020年11月	《智能计算中心规划建设指南》	国家信息中心	提升AI算力生产供应，智算中心基于新型硬件架构和人工智能算法模型，保证规划建设的技术领先性；促进数据开放共享汇聚各行业领域数据资源，全面提升AI算法训练数据质量；培育区域智能生态：推动AI产业创新聚集推动AI产业创新聚集，加速AI应用场景落地
2017年7月	《新一代人工智能发展规划》	国务院	首次提及智算中心概念。强调建设布局人工智能创新平台，重点突破人机协同的感知与执行一体化模型等核心技术，建立人工智能超级计算中心、大规模超级智能计算支撑环境、在线智能教育平台等

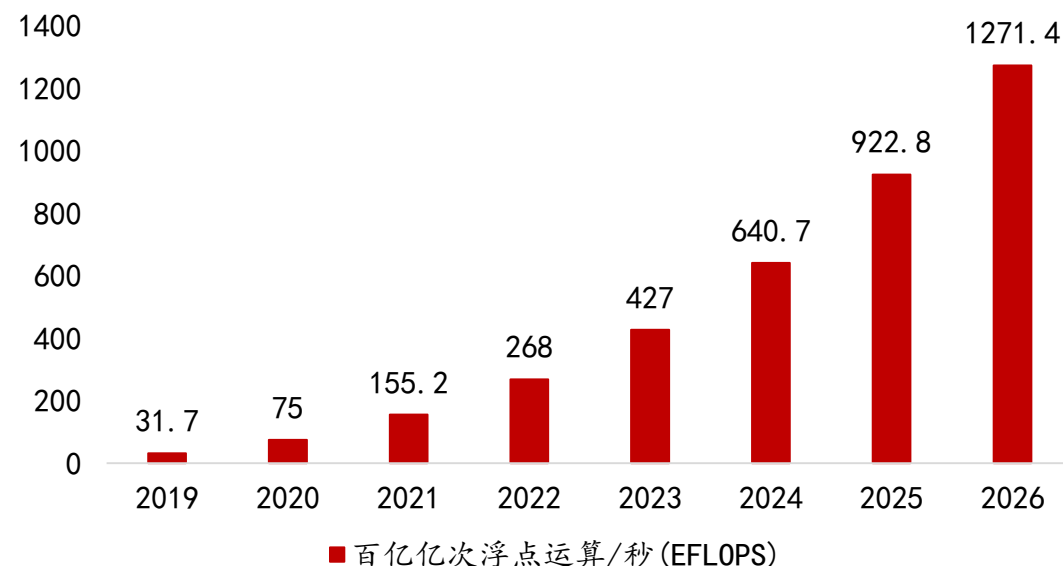
2.4.2 产业端积极响应，智能算力建设持续提速

- ◆ **智算算力增速超通用算力，26城抢建智算中心。**
 - ✓ 根据人民网，目前国家在8地启动建设国家算力枢纽节点，并规划了10个国家数据中心集群，协调区域平衡化发展。根据智东西数据，截止2022年2月，全国共有至少26个城市在推动或刚完成当地智算中心建设，其中合肥、庆阳、大连、沈阳、深圳、长沙等至少6个城市已经宣布开工建设。
 - ✓ 中国智能算力规模持续高速增长，据IDC预计，到2026年智算规模将达1271.4EFLOPS，未来CAGR达52.3%，同期通用算力规模CAGR为18.5%。

8大算力枢纽智算中心建设进度（21.1-22.2）

国家算力枢纽节点	建成或正在建设的智算中心
甘肃枢纽	庆阳智算中心
京津冀枢纽	中国电信京津冀大数据智能算力中心
	河北人工智能计算中心
长三角枢纽	商汤科技人工智能计算中心
	南京智能计算中心
	昆山智算中心
	杭州人工智能计算中心
	腾讯智慧产业长三角（合肥）智算中心
	合肥先进计算中心
粤港澳大湾区枢纽	广州人工智能公共算力中心
	深圳市人工智能公共算力中心
成渝枢纽	成都智算中心

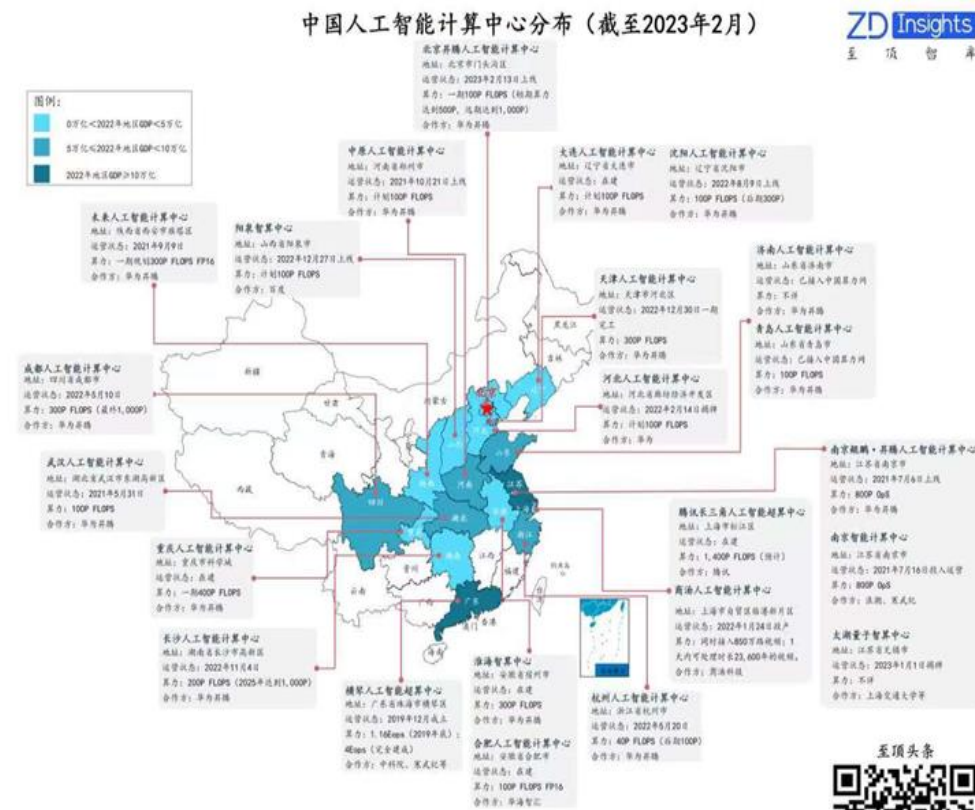
中国智算算力规模预测（EFLOPS）



2.4.2 产业端积极响应，智能算力建设持续提速

- ◆ **北京昇腾人工智能计算中心正式点亮**：北京昇腾人工智能计算中心正式点亮，将推动北京人工智能产业高质量发展。该智能计算中心采用昇腾AI基础软硬件，充分释放硬件算力，加速人工智能企业创新应用和模型孵化。
- ◆ **贵州省大数据局印发《面向全国的算力保障基地建设规划》**：总体目标是到2025年，面向全国的算力保障基地建设任务全面完成，贵州超大规模数据中心集群的地位更加巩固，存算比更加合理，优化基础设施布局、结构、功能和系统集成，数据中心实现集约化、规模化、绿色化发展，网络互联互通、能源安全可靠提高到新的水平，打造具有国际竞争力的数字产业集群。
- ◆ **上海市经济信息化委印发《上海市推进算力资源统一调度指导意见》**：主要目标为到2023年底，依托本市人工智能公共算力服务平台，接入并调度4个以上算力基础设施，可调度智能算力达到1,000 PFLOPS（FP16）以上；到2025年，市人工智能公共算力服务平台能级跃升，完善算力交易机制，实现跨地域算力智能调度，通过高效算力调度，推动算力供需均衡，带动产业发展作用显著增强。
- ◆ **惠州首个超大规模数据及算力中心力争年内投产**：2023年年初，作为大数据及关联产业发展的重要支撑点的粤港澳大湾区（惠州）数据产业园建设取得明显成效。落户该园区的润泽（惠州）国际信息港一期项目试运行工作进展顺利，预计年内正式投产。其目标是构建具有国际领先技术水平的算力基础设施，带动数据服务及硬件研发制造等关联产业集聚发展。

中国人工智能计算中心分布图（截至2023年2月）



2.4.2 产业端积极响应，智能算力建设持续提速

- ◆ **山东首个人工智能计算中心上线运行，竞逐人工智能赛道**：2023年3月17日青岛市人工智能产业园正式开园，同步上线的青岛人工智能计算中心，成为山东首个上线运行的人工智能计算中心。中心首期具备100P算力，相当于5万台高性能PC的算力，将面向青岛乃至胶东地区的企业、高校和科研机构提供普惠公共算力服务。
- ◆ **河南省数字化转型战略工作方案出炉，推进郑州、洛阳构建超大型绿色数据中心集群**：2023年3月30日河南省制造强省建设领导小组办公室印发《2023年河南省数字化转型战略工作方案》，目标今年电子信息制造业营业收入力争突破8000亿元，先进计算、软件产业规模均超过500亿元。
- ◆ **天津市人工智能计算中心揭牌，加快打造天津数字经济发展新动能**：2023年3月18日，天津市人工智能计算中心正式揭牌上线，助力人工智能产业创新发展。人工智能中心不仅提供基础算力服务，还提供应用创新服务、产业孵化服务等，把算力、算法、数据、应用场景和人才进行5要素的聚集，帮助企业在人工智能科研创新上降本增效。

2023年河南省数字化转型战略目标任务分解

序号	城市	智能工厂/智能车间	贯标升级版/对标升级版（家）	数字化转型项目（个）	企业上云
1	郑州市	22	36/360	116	10740
2	开封市	7	15/150	48	1120
3	洛阳市	12	24/240	80	2080
4	平顶山市	7	15/150	52	1540
5	安阳市	7	15/150	52	460
6	鹤壁市	6	12/120	40	290
7	新乡市	12	21/210	72	1840
8	焦作市	8	18/180	60	440
9	濮阳市	5	12/120	40	570
10	许昌市	9	18/180	60	1610
11	漯河市	6	12/120	40	730
12	三门峡市	5	12/120	40	350
13	南阳市	12	21/210	72	2730
14	商丘市	9	18/180	56	1710
15	信阳市	4	15/150	52	1510
16	周口市	8	15/150	52	1360
17	驻马店市	7	15/150	48	770
18	济源示范区	4	6/150	20	150
合计		150	300/3000	1000	30000

天津市人工智能计算中心内的算力服务器



2.5.1 重申强调，算力在大模型的背景下势必迎来爆发

- ◆ **ChatGPT用户数量暴增，同样侧面证明了AI产业革命下，对于算力基础的高度需求。**根据SimilarWeb的数据，2023年2月，ChatGPT访问数量为10亿次/每月，而2023年4月，ChatGPT的访问量增加至17.56亿次/每月。
- ◆ **根据我们的测算，目前ChatsGPT产品运营需英伟达A100 GPU约71296片，预计投入算力成本达17.73亿美元。**据SimilarWeb数据，2023年5月（至5月24日）ChatGPT官网（chat.openai.com）总访问量为14.08亿次。据环球零碳研究中心数据，每次用户与ChatGPT互动，ChatGPT的每个响应词在A100 GPU上需要350毫秒。英伟达DGXA100服务器单机售价约为19.9万美元/台，每台大约可搭载8片A100 GPU。根据测算结果，从今年2月至今，对英伟达A100 GPU需求持续增长，我们预计后续该趋势有望保持。
- ◆ **短期算卡为王，长期自主可控：**重申强调，算力在大模型的背景下势必迎来爆发，而算卡作为算力的心脏其重要性不言而喻。**长期来看，美国连续发动对我国高科技行业制裁，其目的是阻碍我国高科技及AI的科技发展，因此发展自主可控算力芯片势在必行。**

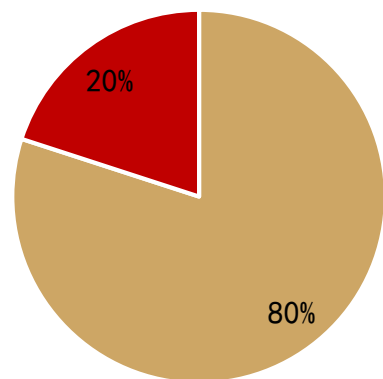
ChatGPT运行的算力需求及成本测算

	单位	2023年2月	2023年3月	2023年4月	2023年5月 (至5.24)
ChatGPT访问量	亿次/月	10.00	15.59	17.56	14.08
月均天数	天	28	31	30	24
日均访问量	百万次/天	35.71	50.29	58.53	58.67
咨询量	字/次/天	300	300	300	300
总咨询量	亿字/天	107.14	150.87	175.60	176.00
A100 GPU算力耗量	ms/字	350	350	350	350
A100需求量	片/天	43,403	61,117	71,134	71,296
A100 GPU售价	万美元/片	1.5	1.5	1.5	1.5
DGXA100服务器搭载A100数	片/台	8	8	8	8
DGXA100服务器需求量	台	5,425	7,640	8,892	8,912
DGXA100系统售价	万美元/台	19.9	19.9	19.9	19.9
初始算力成本	亿美元	10.80	15.20	17.69	17.73

2.5.2 国产AI算力芯片自主可控势在必行

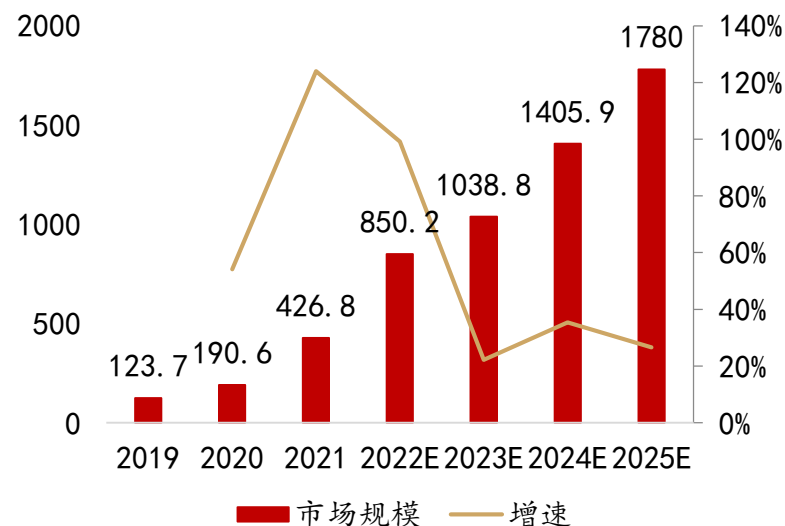
- ◆ **我国AI芯片方面仍处于“垄断局面”，高端 AI芯片仍需自主可控，我国相关企业已初具竞争实力：**根据 IDC 数据，2021 年，中国加速卡数量出货超过 80 万片，其中 Nvidia 占据超过 80% 市场份额。此外还包括 AMD、百度、寒武纪、燧原科技、新华三、华为、Intel 和赛灵思等。
- ◆ **人工智能逐渐成为主流的发展趋势，中国人工智能市场投资规模呈上升趋势。**在中国市场，IDC 预测，2026 年中国人工智能投资有望达到 266.9 亿美元，约占全球投资的 8.9%，在其他国家中排名世界第二。
- ◆ **AI 算力规模的快速增长将催生更大的 AI 芯片需求：**根据亿欧智库的数据，预计 2023 年中国 AI 芯片市场规模将达到 1039 亿元，2025 年中国 AI 芯片市场规模将达到 1780 亿元，三年 GAGR 为 19.66%。

2021 年我国服务器加速卡市场份额

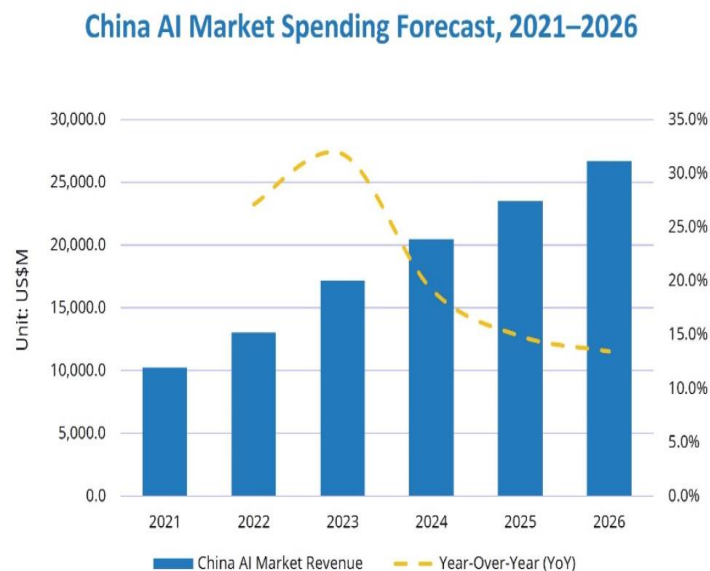


■ 英伟达 ■ 其他

中国 AI 芯片市场规模及其预测（亿元）



中国人工智能市场投资情况及其预测



2.5.2 国产AI算力芯片自主可控势在必行

◆ **发展国产AI芯片势在必行：**我国AI芯片已经呈现百舸争流的情况，国产化AI芯片势在必行，相关厂商积极加速推进AI芯片布局，促进AI芯片市场发展。

中国AI芯片主要厂商及其产品情况

公司	产品型号	应用（训练或推理）	算力	频率	功耗	制程
寒武纪	思元370	训练+推理	256TOPS (INT8)	/	/	7nm
	思元290	训练	512TOPS (INT8)			
	思元270	推理	128TOPS (INT8)			
燧原科技	T20	训练	256TOPS (INT8)	1.5GHz	300w	/
	T21	训练	256TOPS (INT8)		300w	
	i20	推理	256TOPS (INT8)		150w	
昆仑芯	昆仑芯2代AI芯片	训练+推理	256TOPS (INT8)		/	7nm
平头哥	含光800	推理	820TOPS	/	/	12nm
沐曦	MAN100	训练+推理		/	/	7nm
华为海思	HUAWEI Ascend310	边缘计算AI	16TOPS (INT8)	/	8w	12nm
	HUAWEI Ascend910		640TOPS (INT8)		310w	N7+
紫光展锐	SC9863A	边缘计算AI	/	1.6GHz	/	/
后摩智能	™H30	感存算AI	156TP0S	/	35w	/
云天励飞	DeepEye 2000	安防/人脸识别AI	/	/	/	22nm
地平线	征程®5	自动驾驶AI	128TP0S	/	/	16nm
景嘉微	JM9	图形处理	1.5TFLOPS (FP32)	1.5GHz	/	14nm
龙芯中科	2K1000LA	可信+边缘计算AI	/	/	5w	/
海光信息	深算一号	计算AI	/	2GHz	350w	7nm
清微智能	TX8	训练+推理	/	/	/	/

资料来源：公司官网、华西证券研究所整理

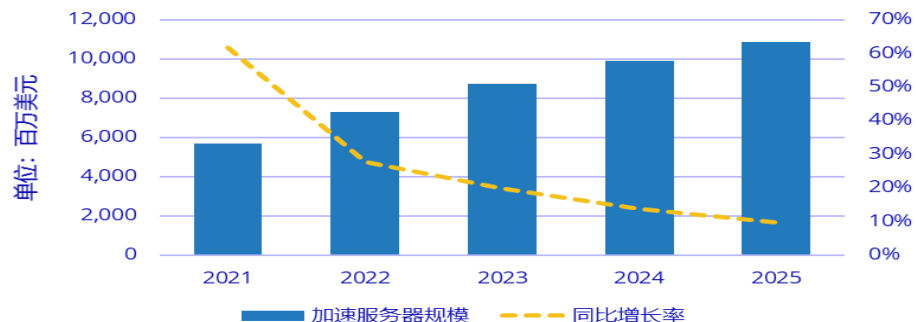
2.5.3 国产加速计算服务器时代到来

- ◆ **人工智能应用场景下的加速计算服务器是中国服务器的核心驱动力:** AI服务器作为AI芯片的载体景气度上行，大模型的出现带动AI服务器呈现加速状态，根据IDC的数据，在2021年的统计，预计到2025年中国加速服务器市场规模将达到108.6亿美元，且2023年仍处于中高速增长期，增长率约为20%。
- ◆ **AI服务器作为算力载体为数字经济时代提供广阔动力源泉:** AI服务器更专精于海量数据处理和运算方面，我们认为其可以为人工智能、深度学习、神经网络、大模型等场景提供广阔的动力源泉，并广泛应用于医学、材料、金融、科技等千行百业。

2021-2025年中国服务器市场规模及增速(亿美元)



中国半年度加速计算市场预测，2021-2025



资料来源：IDC，华西证券研究所

公司名称	名称/型号	处理器	内存支持	AI加速卡/AI处理器	AI算力
浪潮科技	NF5468M6	2颗第三代Intel® Xeon®可扩展处理器 (Ice Lake), TDP 270W, 支持3条UPI互联	支持32条DDR4 RDIMM/LRDIMM内存, 速率最高支持3200MT/s	/	/
神州数码	KunTai A222	1*鲲鹏920处理器, 24核, 主频2.6GHz	4个DDR4 RDIMM, 最高速率3200MT/s内存保护支持ECC、SEC/DED、SDDC、Patrol scrubbing功能; 单根内存条容量支持16GB/32GB/64GB/128GB	最大支持3张Atlas 300V 视频解析卡或Atlas 300I Pro 推理卡或Atlas 300V Pro 视频解析卡	最大420 TOPS INT8
	KunTai A722	2*鲲鹏920处理器, 支持32、48、64核可选, 主频2.6GHz	16个或32个DDR4 RDIMM, 最高速率2933MT/s内存保护支持ECC、SEC/DED、SDDC、Patrol scrubbing功能; 单根内存条容量支持16GB/32GB/64GB/128GB	最大支持8张, Atlas 300V 视频解析卡或Atlas 300I Pro 推理卡或Atlas 300V Pro 视频解析卡	最大1120 TOPS INT8
	KunTai A924	4*鲲鹏920处理器, 支持48核, 主频2.6GHz	支持32个DDR4内存插槽, 速率最高2933MT/s内存保护支持ECC、SEC/DED、SDDC、Patrol scrubbing功能; 单根内存条容量支持32GB/64GB/128GB	8*昇腾910, 支持直出100G RoCE网络接口	最大512Tops Int8或256Tops FP16
拓维信息	兆瀚RA2300-A	支持两颗华为鲲鹏920处理器, CPU主频2.6GHz。单CPU最多64个内核, 最大功率180w。	最多支持32个DDR4内存DIMM插槽, 最高速率2933MT/s	支持Atlas 300I Pro推理卡和Atlas 300V Pro视频解析卡	最大1.12 POPS INT8; 最大560 TFLOPS PF16
	兆瀚SA300	支持一颗华为鲲鹏920处理器, CPU主频2.6GHz。单CPU最多64个内核, 最大功率181w。	最多支持4个DDR4内存DIMM插槽, 最高速率2934MT/s	支持Atlas 300I Pro推理卡/Atlas 300V Pro 视频解析卡	最大420 TOPS INT8或384路1080P 30 FPS视频解析 (硬件解码能力)
	兆瀚RA5900-A	支持四颗华为鲲鹏920处理器, CPU主频2.6GHz。单CPU最多64个内核, 最大功率182w。	最多32个DDR4内存插槽, 支持RDIMM。单根内存条容量支持32 GB/64GB	8*昇腾910	/
	兆瀚RA2302-B	2*64核青松处理器	32个DDR4内存插槽, 最高3200 MT/s, 支持ECC	最大支持4个Atlas 300I/V Pro	最大560 TOPS INT8
龙芯中科	KU 2208-L2	支持2颗国产龙芯3B4000处理器; 8核心 主频1.8GHz-2.0GHz, 单颗功耗达50W	板载8个DDR4 DIMM扩展插槽, 支持DDR4 ECC RDIMM; 内存单根内存容量: 8GB、16GB, 内存可达: 128GB	/	/
	KU 2208-L3	支持2颗国产龙芯3C5000L处理器, 高达32核心; 主频2.2GHz, 功耗高达130W	板载8个DDR4 DIMM插槽, 支持双通道DDR4 ECC RDIMM, 支持内存单根内存容量: 8GB、16GB、32GB、64GB, 支持至: 512GB内存	/	/

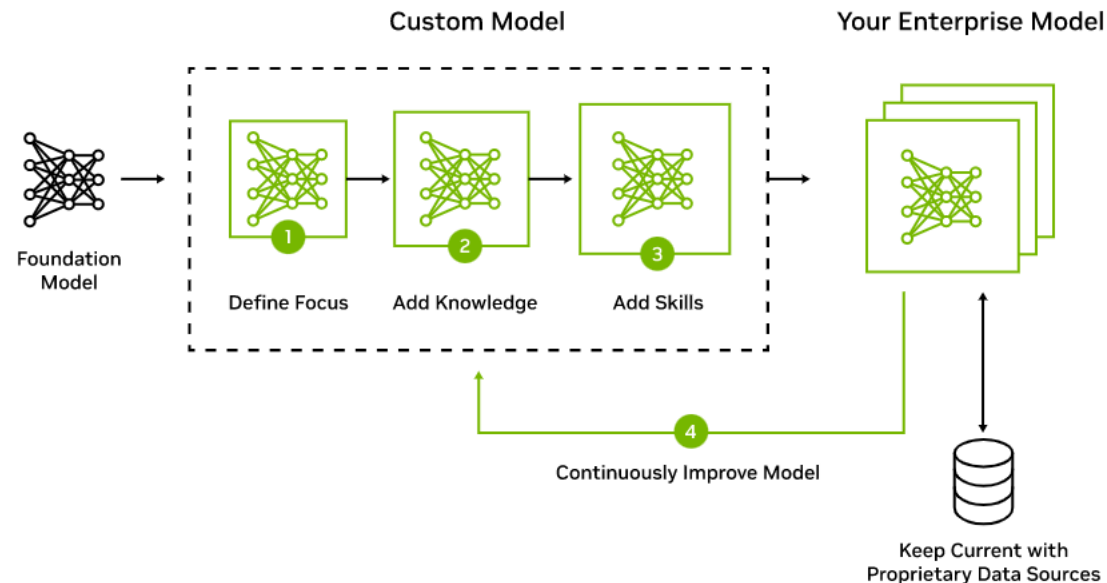
2.5.4 “共享AI能力与算力”，AI云需求高增

- ◆ **部署生成式 AI 应用难度较高，AI云提供平台定制化能力。**随着大模型带来的人工智能产业崛起，AI应用如文本生成、自动客服、自动驾驶等领域快速扩张。对于大多数企业自己部署这样的能力是非常困难的，我们认为一是由于目前的算力缺口及训练的边际成本较高，二是由于全链部署应用需要深厚的软硬件结合生态技术。而AI云可将如英伟达等专业供应商的AI能力整合到云上，让企业能够直接接入应用或从基础层进行预训练，进而形成自己的模型和应用。
- ◆ **AI云需求快速增长，云算力革命开启。**在企业对大模型训练、人工智能应用部署等AI能力需求持续上行的态势下，AI云产品受到市场的青睐。以阿里、腾讯为代表的平台型公司在云端市场布局上，更多地关注的是通用云的打造；而华为、曙光更多地是从硬件的角度着手加入云市场布局。我们认为软硬结合的AI云玩家（英伟达、首都在线）符合市场的需求。同时，AI云需求高增也意味着未来掌握智算卡的企业将继续占领市场高地，万变不离其宗，我们坚持认为拥有算力的企业在下一阶段进行应用和平台化竞争的过程中具有先发优势。

首都在线云游戏解决方案



英伟达NeMo云服务





03 投资建议：梳理AIGC相关受益厂商

3.1 投资建议: 梳理AIGC的受益厂商

- ◆ 我们认为大模型有望赋能千行百业，算力作为“底层燃料”其重要性不言而喻，以英伟达为首的科技巨头有望借助算力开启新一轮成长曲线，再次强调我们的观点，**短期算卡为王，长期自主可控**。拥有算卡的厂商有望开启新一轮成长曲线，而长期自主可控为大势所趋，积极的推荐以下三条投资主线:
- ◆ **1) AI芯片厂商**，相关受益标的为: **寒武纪、海光信息、景嘉微、龙芯中科**等；
- ◆ **2) AI服务器厂商**，相关受益标的为: **中科曙光、神州数码、拓维信息、工业富联、浪潮信息**等；
- ◆ **3) AI云厂商**，相关受益标的为: **首都在线、鸿博股份、青云科技、优刻得、光环新网、新炬网络**等。

AIGC的A股受益标的

公司名称	股票代码	收盘价	市值(亿元)	EPS (元)			PE (倍)		
		2023/6/5	2023/6/5	2022	2023E	2024E	2022	2023E	2024E
赛武纪	688256. SH	230. 32	958. 23	-3. 14	-1. 95	-1. 34	-	-	-
海光信息	688041. SH	83. 50	1940. 82	0. 38	0. 56	0. 81	219. 7	149. 1	102. 5
景嘉微	300474. SZ	97. 35	443. 11	0. 64	0. 90	1. 26	152. 1	108. 4	77. 1
龙芯中科	688047. SH	150. 49	603. 46	0. 14	0. 52	0. 98	1074. 9	291. 7	153. 3
中科曙光	603019. SH	51. 76	757. 77	1. 06	1. 38	1. 78	48. 8	37. 6	29. 0
神州数码	000034. SZ	27. 55	184. 37	1. 57	1. 84	2. 20	17. 6	14. 9	12. 5
拓维信息	002261. SZ	14. 99	188. 25	-0. 82	0. 10	0. 13	-	155. 7	112. 0
工业富联	601138. SH	18. 93	3760. 05	1. 02	1. 19	1. 34	18. 6	15. 8	14. 1
首都在线	300846. SZ	17. 90	83. 56	-0. 41	0. 16	0. 37	-	112. 9	48. 0
鸿博股份	002229. SZ	35. 86	178. 71	-0. 15	-0. 03	0. 12	-	-	288. 3
光环新网	300383. SZ	11. 91	214. 09	-0. 49	0. 36	0. 45	-	32. 6	26. 2
新炬网络	605398. SH	38. 35	31. 94	0. 68	0. 85	1. 03	56. 4	45. 0	37. 1
优刻得	688158. SH	20. 97	95. 01	-0. 92	-0. 53	-0. 33	-	-	-

注：均来自wind一致预测

资料来源：， 华西证券研究所

3.2.1 海光信息：支持全精度，GPU实现规模量产

- ◆ **海光信息主要从事高端处理器、加速器等计算芯片产品和系统的研究、开发，主要产品包括海光CPU和海光DCU:**2018年10月，公司启动深算一号DCU产品设计，海光8100采用先进的FinFET工艺，典型应用场景下性能指标可以达到国际同类型高端产品的同期水平。2020年1月，公司启动DCU深算二号的产品研发。
- ◆ **海光DCU性能强大:**海光DCU基于大规模并行计算微结构进行设计，不但具备强大的双精度浮点计算能力，同时在单精度、半精度、整型计算方面表现同样优异，是一款计算性能强大、能效比较高的通用协处理器。海光DCU集成片上高带宽内存芯片，可以在大规模数据计算过程中提供优异的数据处理能力。

海光信息主要产品



系列	7000系列CPU	5000系列CPU	3000系列CPU
核心规格	最大32个物理核心	最大16个物理核心	最大8个物理核心
应用领域	高端通用服务器、先进计算系统	通用服务器	个人工作站、工控设备等终端产品

系列	8000系列DCU
核心规格	60-64个深度计算单元
应用领域	先进计算系统、人工智能

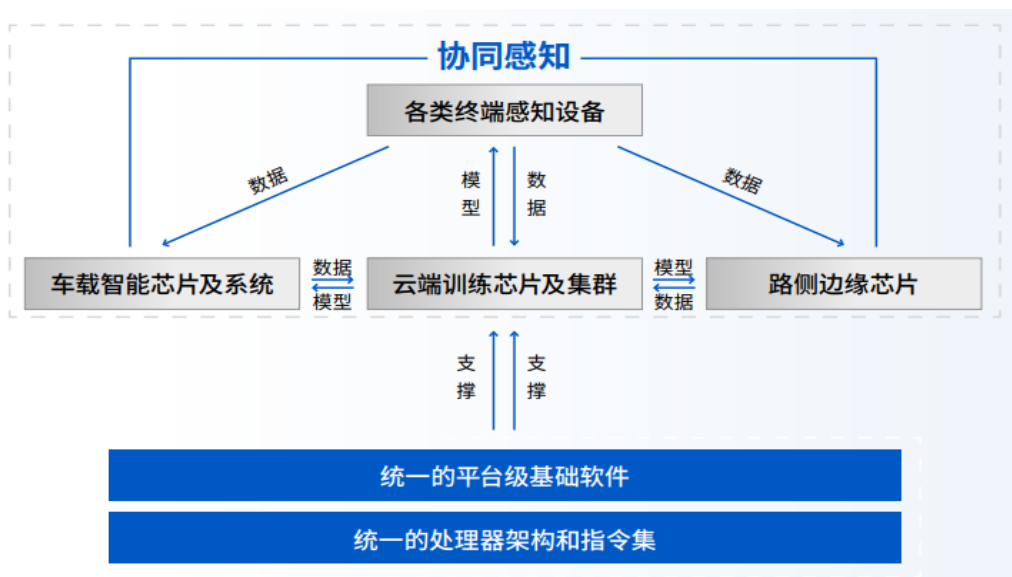
海光深算一号性能达到国际同类产品水平

项目	海光	NVIDIA	AMD
品牌	深算一号	Ampere 100	MI100
生产工艺	7nm FinFET	7nm FinFET	7nm FinFET
核心数量	4096 (64CUs)	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
显存位宽	4096 bit	5120 bit	4096bit
显存频率	2.0 GHz	3.2 GHz	2.4 GHz
显存带宽	1024 GB/s	2039 GB/s	1228 GB/s
TDP	350 W	400 W	300 W
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

3.2.2 寒武纪：少数全面掌握AI芯片技术的企业之一

- ◆ **寒武纪是目前国际上少数几家全面系统掌握了通用型智能芯片及其基础系统软件研发和产品化核心技术的企业之一：**寒武纪主营业务是应用于各类云服务器、边缘计算设备、终端设备中人工智能核心芯片的研发和销售。公司的主要产品包括终端智能处理器IP、云端智能芯片及加速卡、边缘智能芯片及加速卡以及与上述产品配套的基础系统软件平台。
- ◆ **公司AI技术积累浓厚：**能提供云边端一体、软硬件协同、训练推理融合、具备统一生态的系列化智能芯片产品和平台化基础系统软件。2022年3月，寒武纪正式发布了新款训练加速卡“MLU370-X8”，搭载双芯片四芯粒封装的思元370，集成寒武纪MLU-Link多芯互联技术，主要面向AI训练任务。

寒武纪“云边端车”协同



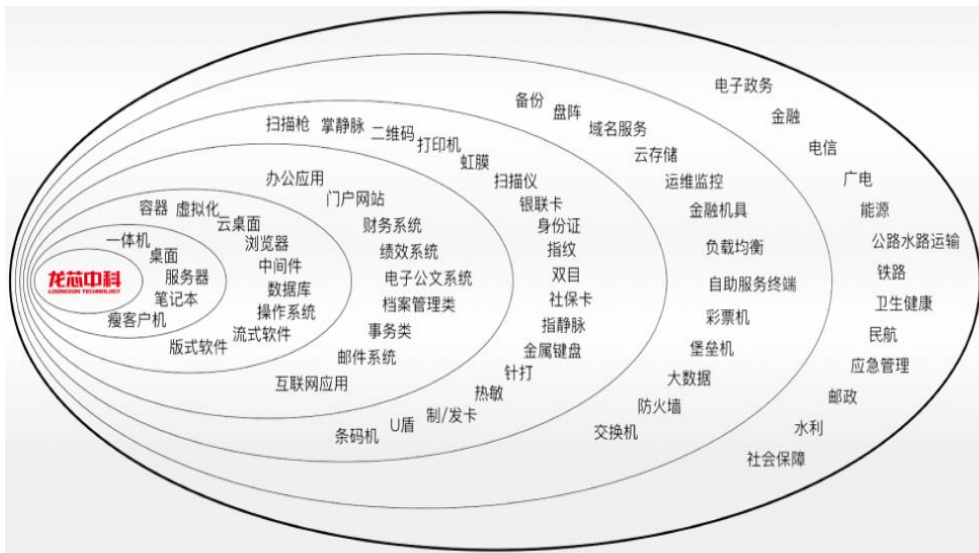
寒武纪产品技术图谱



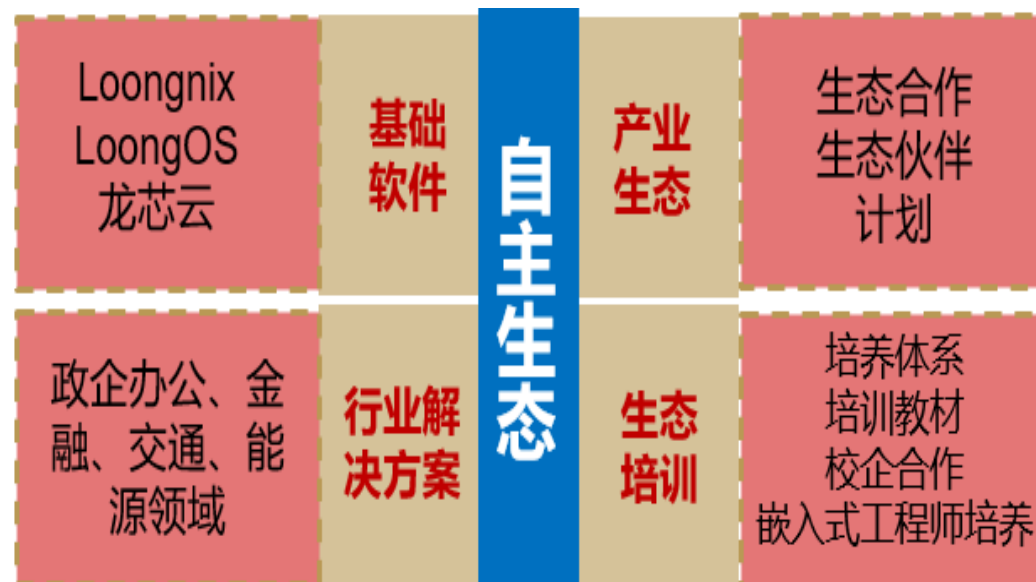
3.2.3 龙芯中科：2K2000系列集成自主GPU

- ◆ **龙芯中科主营业务为处理器及配套芯片的研制、销售及服务：**主要产品与服务包括处理器及配套芯片产品与基础软硬件解决方案业务。公司基于信息系统和工控系统两条主线开展产业生态建设，面向网络安全、办公与业务信息化、工控及物联网等领域与合作伙伴保持全面合作，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应用。
- ◆ **公司自主研发2K200系列GPU：**2022年12月，龙芯2K2000完成了初步功能调试及性能测试，达到其设计目标，2023年将推出试用。龙芯2K2000集成了两个LA364处理器核，典型工作频率为1.5GHz，共享2MB的L2缓存，SPEC2006INT（base）单核定/浮点分值达到13.5/14.9分。龙芯2K2000芯片集成了龙芯自主研发的GPU，并优化了图形算法和性能。

龙芯中科生态合作示意图



龙芯中科自主生态



3.2.4 景嘉微：新一代JM9系列有望打开商用市场

- ◆ **国产GPU龙头企业：**公司成立于2006年，主要从事军用电子产品的研发、生产、销售，目前形成了三大业务板块分别是图形线控模块、小型专用雷达和芯片业务。GPU方面，2014年首推JM5400实现了军用GPU的国产替代；第二款芯片JM7200于2018年研发成功，具备了PC端的功能；日前，公司9系列芯片研发成功，具备高性能计算能力。
- ◆ **新一代JM9系列有望打开商用市场：**日前，公司JM9系列图形处理芯片已顺利发布，应用领域涵盖地理信息系统、媒体处理、CAD辅助设计、游戏、虚拟化等高性能显示和人工智能计算领域。目前，信创市场为公司提供了新的业务增长点，JM9系列图形处理芯片的成功发布将为公司未来进一步拓展通用市场提供强有力的产品支撑。

景嘉微GPU系列产品



景嘉微7系列GPU示意图



3.3.1 中科曙光：我国高性能计算、智能计算领军企业

- ◆ **中科曙光作我国核心信息基础设施领军企业：**在高端计算、存储、安全、数据中心等领域拥有深厚的技术积淀和领先的市场份额，并充分发挥高端计算优势，布局智能计算、云计算、大数据等领域的技术研发，打造计算产业生态，为科研探索创新、行业信息化建设、产业转型升级、数字经济发展提供了坚实可信的支撑。
- ◆ **依托先进计算领域的先发优势和技术细节，中科曙光全面布局智能计算：**完成了包括AI核心组件、人工智能服务器、人工智能管理平台、软件等多项创新，构建了完整的AI计算服务体系。并积极响应时代需求，在智能计算中心建设浪潮下，形成了5A级智能计算中心整体方案。目前，曙光5A智能计算中心已在广东、安徽、浙江等地建成，江苏、湖北、湖南等地已进入建设阶段，其他地区也在紧张筹备和规划中。

中科曙光主要产品

通用服务器 机架式服务器 高密度服务器 刀片服务器 核心应用服务器	智能计算服务器 深度学习训练 智能应用推理	终端&工作站 微型计算机 工作站	高性能计算机 通用高性能计算机 高性能计算机系统组件 高性能计算机的服务支撑
机房冷却设施 微模块产品 液冷基础设施产品	存储产品 分布式统一存储 多控统一存储 高密度存储服务器 备份一体机	网络安全产品 数据中心安全产品 汇聚分流设备 智能加速卡 网络内容识别分析系统 网络态势感知系统	大数据平台软件 大数据智能引擎系列 数据工程服务系列 视频智能分析系列 大数据与人工智能实训平台
云计算平台软件 云计算操作系统 超融合一体机 云桌面 云容灾	计算服务 弹性计算服务 混合计算服务 专有计算服务 API 托管、运营	云计算服务 云服务器 ECS 裸金属 BMS 对象存储 OSS 云容器实例 CCI 人工智能服务 数据开发 DDS 数据治理中心 DGS 数据服务 DSS 数据可视化 DAV 数据集成 Data Integration	城市云 智慧城市 国资云 交通云 医疗云
5A级智算中心			

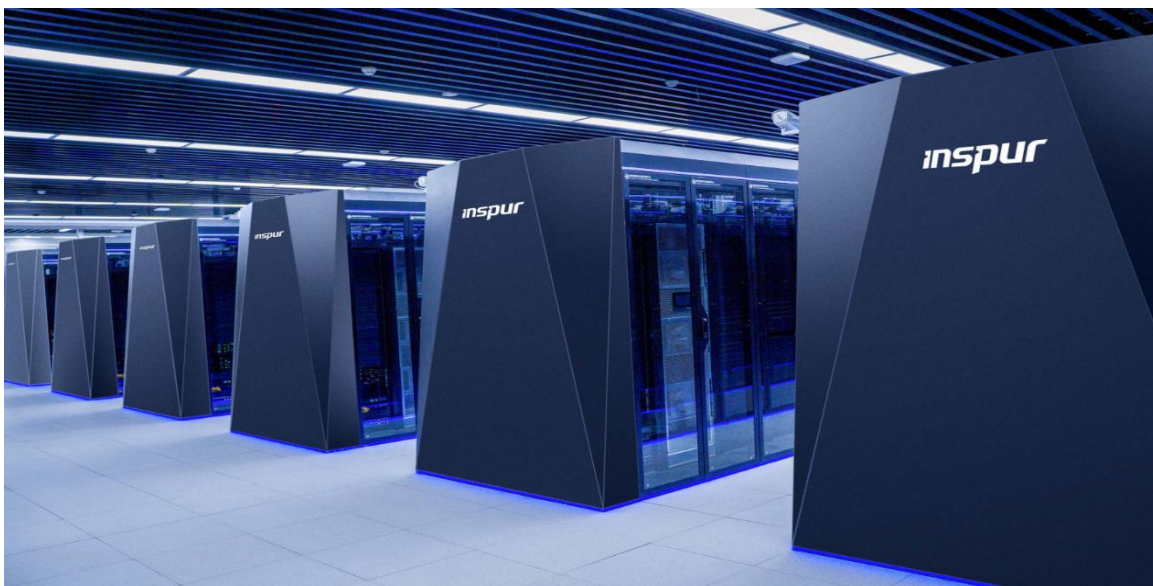
中科曙光硅立方液体相变冷却计算机



3.3.2 浪潮信息：中国服务器/AI服务器市占率稳居榜首

- ◆ **浪潮信息是全球领先的新型IT基础架构产品、方案及服务提供商：**公司是全球领先的 AI 基础设施供应商，拥有业内最全的人工智能计算全堆栈解决方案，涉及训练、推理、边缘等全栈 AI 场景，构建起领先的 AI 算法模型、AI 框架优化、AI 开发管理和应用优化等全栈 AI 能力，为智慧时代提供坚实的基础设施支撑。
- ◆ **公司算力技术壁垒浓厚：**生产算力方面，公司拥有业内最强最全的 AI 计算产品阵列，业界性能最好的Transformer 训练服务器 NF5488、全球首个 AI 开放加速计算系统 MX1、自研 AI 大模型计算框架 LMS。聚合算力层面，公司针对高并发训练推理集群进行架构优化，构建了高性能的NVMe 存储池，深度优化了软件栈，性能提升 3.5 倍以上。调度算力层面，浪潮信息 Aistation 计算资源平台可支持 AI 训练和推理，是业界功能最全的 AI 管理平台；同时，浪潮信息还有自动机器学习平台 AutoML Suite，可实现自动建模，加速产业化应用。

浪潮信息智算中心



浪潮信息智算中心



3.3.3 神州数码: 华为生态核心践行者

- ◆ **神州数码领先的数字化转型：**神州数码围绕企业数字化转型的关键要素，开创性的提出“数云融合”战略和技术体系框架，着力在云原生、数字原生、数云融合关键技术和信创产业上架构产品和服务能力，为处在不同数字化转型阶段的快消零售、汽车、金融、医疗、政企、教育、运营商等行业客户提供泛在的敏捷IT能力和融合的数据驱动能力。
- ◆ **神州数码为华为生态核心践行者：**公司旗下的神州鲲泰基于华为鲲鹏处理器多款不同种类的服务器产品，包括1、单路服务器：R222、R224；2、双路服务器：R522、R524、R722、R724、R2240、R2260、R2280。3、四路服务器：R822。此外，公司基于华为鲲鹏920处理器与昇腾Atlas AI加速卡，神州数码开发了采用ARM架构的一系列AI服务器。

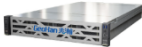
神州数码服务器及相关参数

名称	示意图	形态	处理器	内存支持	AI加速卡/AI处理器	AI算力
KunTai A222		2U单路边缘机架式服务器	1*鲲鹏920处理器，24核，主频2.6GHz	4个DDR4 RDIMM，最高速率3200MT/s内存保护支持ECC、SEC/DED、SDDC、Patrol scrubbing功能；单根内存条容量支持16GB/32GB/64GB/128GB	最大支持3张Atlas 300V 视频解析卡或Atlas 300I Pro 推理卡或Atlas 300V Pro 视频解析卡	最大420 TOPS INT8
KunTai A722		2U 双路推理型 AI 机架式服务器	2*鲲鹏920处理器，支持32、48、64核可选，主频2.6GHz	16个或32个DDR4 RDIMM，最高速率2933MT/s内存保护支持ECC、SEC/DED、SDDC、Patrol scrubbing功能；单根内存条容量支持16GB/32GB/64GB/128GB	最大支持8张，Atlas 300V 视频解析卡或Atlas 300I Pro 推理卡或Atlas 300V Pro 视频解析卡	最大1120 TOPS INT8
KunTai A924		4U四路训练型AI机架式服务器	4*鲲鹏920处理器，支持48核，主频 2.6GHz	支持32个DDR4内存插槽，速率最高2933MT/s内存保护支持ECC、SEC/DED、SDDC、Patrol scrubbing功能；单根内存条容量支持32GB/64GB/128GB	8*昇腾910，支持直出100G RoCE网络接口	最大512Tops Int8或256Tops FP16

3.3.4 拓维信息: 华为生态重要参与者

- ◆ **拓维信息是领先的软硬一体化解决方案提供商:** 公司1996年成立，业务涵盖政企数字化、智能计算、鸿蒙生态，覆盖全国31个省级行政区、海外10+国家，聚焦数字政府、运营商、考试、交通、制造、教育等重点领域和行业，服务超过1500家政企客户，为其提供全栈国产数字化解决方案和一站式全生命周期的综合服务。
- ◆ **拓维信息为华为生态重要参与者:** “兆瀚”系列通用服务器是基于ARM架构，搭载鲲鹏920处理器设计开发的机架式型服务器，拥有高的性能、可靠性、高效环保、兼容性强等特点；“兆瀚”系列AI服务器能够满足当前各类主流AI场景与AI大模型的训练需求，已经在国内多个区域人工智能计算中心、城市人工智能中枢、通用AI服务器场景中得到了应用，已经在国内多家头部互联网企业开展适配测试。

拓维信息旗下“兆瀚”系列服务器产品介绍

种类	名称	示意图	形态	处理器	内存支持	AI加速卡/AI处理器	AI算力
通用服务器	兆瀚RH220系列		2U双路机架	支持两颗华为鲲鹏920处理器，CPU主频2.6GHz。单CPU最多64个内核，最大功率180w。	最多支持32个DDR4内存DIMM插槽，最高速率2933MT/s	/	/
	兆瀚RH520系列		4U机架服务器	支持两颗华为鲲鹏920处理器，CPU主频2.6GHz。单CPU最多64个内核，最大功率180w。	最多支持32个DDR4内存DIMM插槽，最高速率2933MT/s	/	/
AI服务器	兆瀚RA2300-A		2U推理服务器	支持两颗华为鲲鹏920处理器，CPU主频2.6GHz。单CPU最多64个内核，最大功率180w。	最多支持32个DDR4内存DIMM插槽，最高速率2933MT/s	支持Atlas 300I Pro推理卡和Atlas 300V Pro视频解析卡	最大1.12 POPS INT8；最大560 TFLOPS PF16
	兆瀚SA300		2U智能边缘服务器	支持一颗华为鲲鹏920处理器，CPU主频2.6GHz。单CPU最多64个内核，最大功率181w。	最多支持4个DDR4内存DIMM插槽，最高速率2934MT/s	支持Atlas 300I Pro推理卡/Atlas 300V Pro视频解析卡	最大420 TOPS INT8 或 384路1080P 30 FPS视频解析（硬件解码能力）
	兆瀚RA5900-A		4U训练服务器	支持四颗华为鲲鹏920处理器，CPU主频2.6GHz。单CPU最多64个内核，最大功率182w。	最多32个DDR4内存插槽，支持RDIMM。单根内存条容量支持32 GB/64GB	8*昇腾910	/
	兆瀚RA2302-B		2U AI 服务器	2*64核青松处理器	32个DDR4内存插槽，最高3200 MT/s，支持ECC	最大支持4个Atlas 300I/V Pro	最大560 TPOS INT8

资料来源：公司官网，华西证券研究所

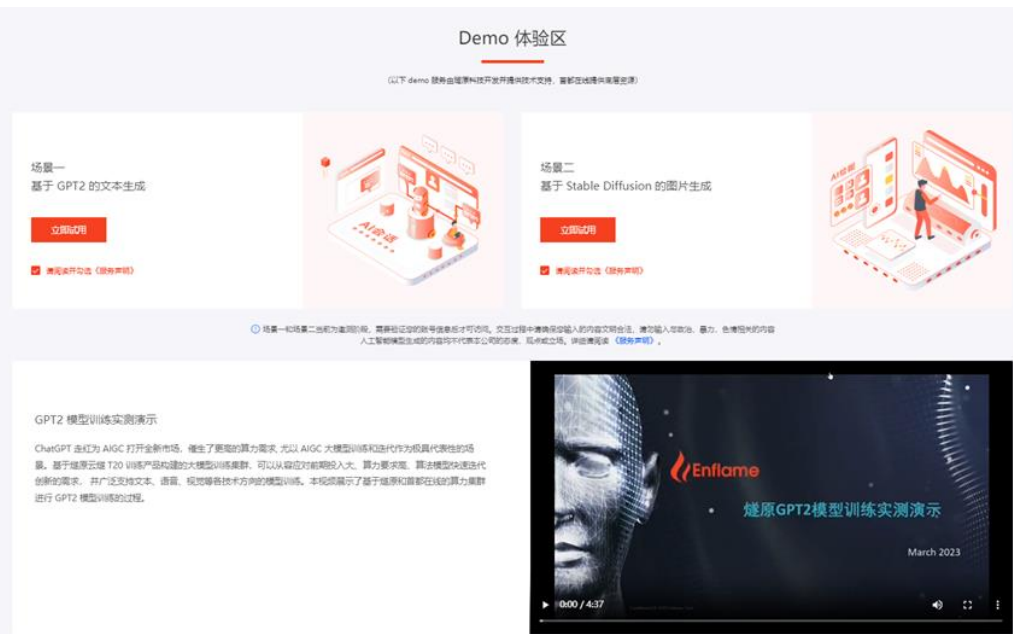
3.4.1 首都在线: AI算力云龙头, AIGC “挖井人”

- ◆ **公司绑定英伟达、燧原, AI云开启第二波成长曲线。**公司已摆脱单一的IaaS 公有云, 重点转向AI算力云转型, 有望借助底层英伟达GPU算力储备, 以AI云为抓手, 开启第二波成长曲线, 我们认为算力网络以及边缘节点是公司AI云的核心壁垒之一。公司首云星图云算力平台已经震撼发布, 深度绑定英伟达, 算力平台采用A100、A40、A5000, 为全球数字世界多场景提供澎湃算力。同时, 公司携手燧原科技, 开启AIGC芯征程, 重点针对大模型 MaaS开展联合攻关, 正式推出云燧i20支撑的AIGC实时推理应用。
- ◆ **海外游戏具备竞争优势, AI算力云赋能千行百业。**我们认为公司AI云平台产品发布与公司底层算力储备密不可分, 借助通过算力、网络、存储等核心能力构建“云-网-数”一体的边缘计算平台, 就近为高算力业务场景如**云游戏、AI、XR、数字人、数字孪生、元宇宙、智能制造**等各领域提供了算力支持。其中云游戏方面, 我们判断云游戏市场处于“技术成熟走向商业可行”与“商业可行走向商业腾飞”的交替阶段, 公司坐拥算力和算力网络双重竞争优势。此外, 公司传统IDC和云业务积极布局海外, 也将会是公司另一个业绩爆发点。

公司星图云底层算例示意图

公司与遂原科技合作示意图

高性能算力供给





04 风险提示

- ◆ **核心技术水平升级不及预期的风险:** AIGC相关产业技术壁垒较高，公司核心技术难以突破，进程低于预期，影响整体进度。
- ◆ **AI伦理风险:** AI可能会生产违反道德、常规、法律等内容。
- ◆ **政策推进不及预期的风险:** 受到宏观经济、财政、疫情影响，政策推进节奏不及预期。
- ◆ **中美贸易摩擦升级的风险:** 供应链存在部分海外提供商，容易受到美国“卡脖子”制裁，导致产品研发不及预期。

分析师与研究助理简介

刘泽晶（首席分析师）2014-2015年新财富计算机行业团队第三、第五名，水晶球第三名，10年证券从业经验。

分析师承诺

作者具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，保证报告所采用的数据均来自合规渠道，分析逻辑基于作者的职业理解，通过合理判断并得出结论，力求客观、公正，结论不受任何第三方的授意、影响，特此声明。

评级说明

公司评级标准	投资评级	说明
以报告发布日后的6个月内公司股价相对上证指数的涨跌幅为基准。	买入	分析师预测在此期间股价相对强于上证指数达到或超过15%
	增持	分析师预测在此期间股价相对强于上证指数在5%—15%之间
	中性	分析师预测在此期间股价相对上证指数在-5%—5%之间
	减持	分析师预测在此期间股价相对弱于上证指数5%—15%之间
	卖出	分析师预测在此期间股价相对弱于上证指数达到或超过15%
行业评级标准		
以报告发布日后的6个月内行业指数的涨跌幅为基准。	推荐	分析师预测在此期间行业指数相对强于上证指数达到或超过10%
	中性	分析师预测在此期间行业指数相对上证指数在-10%—10%之间
	回避	分析师预测在此期间行业指数相对弱于上证指数达到或超过10%

华西证券研究所：

地址：北京市西城区太平桥大街丰汇园11号丰汇时代大厦南座5层

网址：<http://www.hx168.com.cn/hxzq/hxindex.html>

华西证券股份有限公司（以下简称“本公司”）具备证券投资咨询业务资格。本公司签约客户使用。本公司不会因接收人收到或者经由其他渠道转发收到本报告而直接视其为本公司客户。

本报告基于本公司研究所及其研究人员认为的已经公开的资料或者研究人员的实地调研资料，但本公司对该等信息的准确性、完整性或可靠性不作任何保证。本报告所载资料、意见以及推测仅于本报告发布当日的判断，且这种判断受到研究方法、研究依据等多方面的制约。在不同时期，本公司可发出与本报告所载资料、意见及预测不一致的报告。本公司不保证本报告所含信息始终保持在最新状态。同时，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，投资者需自行关注相应更新或修改。

在任何情况下，本报告仅提供给签约客户参考使用，任何信息或所表述的意见绝不构成对任何人的投资建议。市场有风险，投资需谨慎。投资者不应将本报告视为做出投资决策的惟一参考因素，亦不应认为本报告可以取代自己的判断。在任何情况下，本报告均未考虑到个别客户的特殊投资目标、财务状况或需求，不能作为客户进行客户买卖、认购证券或者其他金融工具的保证或邀请。在任何情况下，本公司、本公司员工或者其他关联方均不承诺投资者一定获利，不与投资者分享投资收益，也不对任何人因使用本报告而导致的任何可能损失负有任何责任。投资者因使用本公司研究报告做出的任何投资决策均是独立行为，与本公司、本公司员工及其他关联方无关。

本公司建立起信息隔离墙制度、跨墙制度来规范管理跨部门、跨关联机构之间的信息流动。务请投资者注意，在法律许可的前提下，本公司及其所属关联机构可能会持有报告中提到的公司所发行的证券或期权并进行证券或期权交易，也可能为这些公司提供或者争取提供投资银行、财务顾问或者金融产品等相关服务。在法律许可的前提下，本公司的董事、高级职员或员工可能担任本报告所提到的公司的董事。

所有报告版权均归本公司所有。未经本公司事先书面授权，任何机构或个人不得以任何形式复制、转发或公开传播本报告的全部或部分内容，如需引用、刊发或转载本报告，需注明出处为华西证券研究所，且不得对本报告进行任何有悖原意的引用、删节和修改。

THANKS