

中国平安 PINGAN

金融·科技

半导体系列报告（四）
AI计算芯片市场展望

Brief Outlook for AI Computing Chip Market

证券研究报告

GPU将成为主流，国产化曙光初现

证券分析师

朱 琨 投资咨询资格编号：S1060518010003

2021年5月31日

通信行业评级：中性（维持）

请务必阅读正文后免责条款

报告摘要

1、算力将成为数字经济引擎和智能社会基石：数字技术在数字经济当中的应用，需要海量数据处理作为支撑。数据处理则需要算力的支撑。算力作为基础设施，支撑大数据和智能化应用，为商品流通和交易、企业管理提供有效支撑，是数字经济引擎和智能社会基石。



2、异构芯片组合提供海量算力，GPU将成为主流：未来的AI计算，将形成以CPU为控制中心，GPU、FPGA、ASIC为特定场景加速卡的异构计算格局。从架构部署的灵活性、效率性以及人工智能算法的本质特性来看，GPU将成为AI计算需求量最大的芯片，预计2025年需求占比将达57%。

3、中国市场将高速增长，GPU国产化曙光初现：预计到2024年，中国人工智能技术市场规模将达到172亿美元；全球占比将从2020年12.5%上升到15.6%，是全球市场增长的主要驱动力；在AI计算的训练和推理两个领域，已经有不少初创公司发布了GPU产品；若以10%-15%的国产化率来估算，国产芯片市场规模大约在37-56亿元。



目录CONTENTS

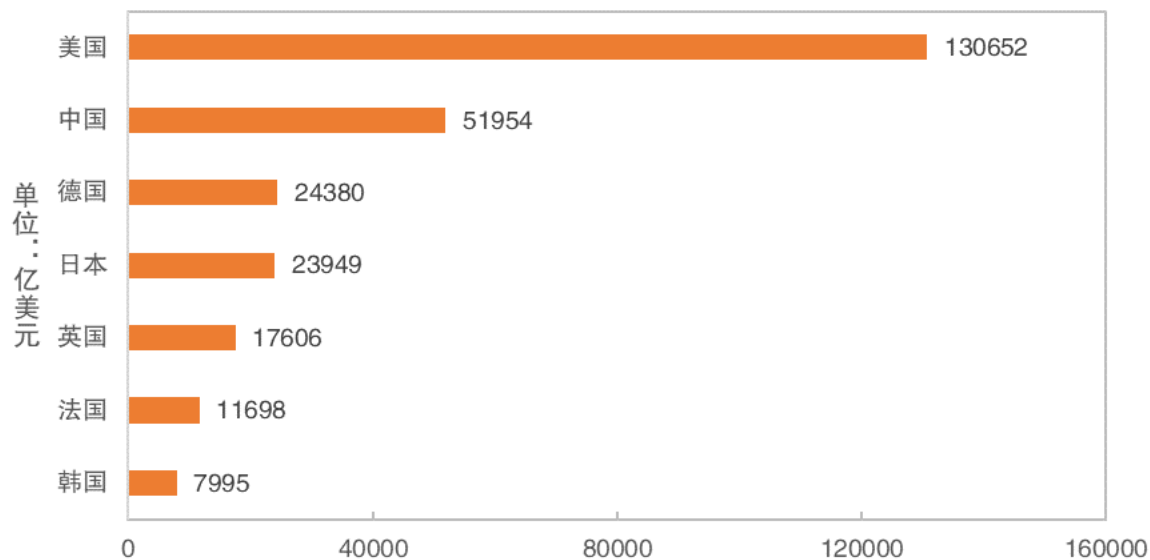
- ◎ 算力：数字经济引擎，智能社会基石
- ◎ 芯片：异构是趋势，GPU将成为主流
- ◎ 展望：中国市场将高速增长，GPU国产化任重道远
- ◎ 投资建议及风险提示

数字经济已经成为全球经济增长主要驱动力

■全球数字经济规模再上新台阶：2019年，可测算47个国家数字经济增加值规模达到了31.8万亿美元；高收入国家数字经济规模全球占比约77%，是中等和低收入国家规模的2.8倍；排名前5的经济体，数字经济规模全球占比达到了78.1%。

■全球数字经济在国民经济中地位持续提升：2019年，可测算47个国家数字经济的GDP占比达到了42%；高收入国家的占比超过了全球平均水平，达到了48%；德国、英国和美国数字经济占比超过了60%，中国占比略低，只有36%。

主要经济体数字经济规模



数据来源：CAICT，平安证券研究所

■全球数字经济实现逆势上扬：近年来，国际经济环境日趋复杂严峻，整体的经济下行压力也在持续增大，但是全球数字经济仍然保持了较快增长。数字经济的各领域稳步推进，新兴产业快速发展，传统产业数字化进程快速推进。2019年，全球数字经济平均名义增速为5.4%，高于全球GDP名义增速。

算力是数字经济发展的关键驱动力

数字经济四化框架



■数据是数字经济最有价值的资源：作为数字经济全新的、关键的生产要素，贯穿于数字经济发展的全部流程，将引发生产要素多领域、多维度、系统性的突破。

■算力是数字经济发展的关键驱动力：数字技术在数字经济当中的应用，需要海量数据处理作为支撑。数据处理则需要算力的支撑。算力工具作为基础设施，支撑大数据和智能化应用，为商品流通和交易、企业管理提供有效支撑。

数据来源：IDC, CAICT, 平安证券研究所

人类社会向智能社会的演进，离不开算力的支撑



数据来源：华为公司，平安证券研究所

2030年人工智能所需算力需求将达到16206EFLOPS

人工智能在以下场景
拥有较大的应用潜力

无人驾驶：借助AI替
代司机

AI医疗：在成像、诊
断、预测分析和
管理领域实现突破

IDC预测，2025年，
人工智能领域的市场
规模预计将达到2081
亿美元

人工智能发展离不开
算力的支撑

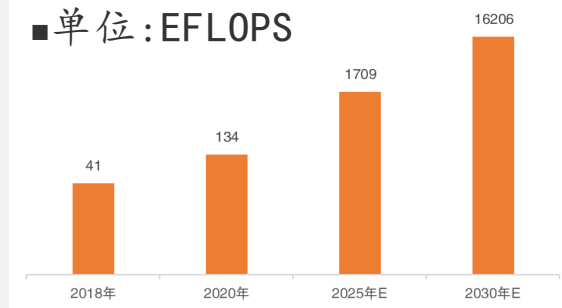
■单场景芯片算力需
求迅速增长

■需要在多种环境随
时提供算力，覆盖
范围要求大，传输
时延要求低

■人工智能在各个场
景渗透率持续提升

人工智能对算力的需
求将快速提升

■单位：EFLOPS



■2030年，人工智能算
力需求相当于1600亿
颗高通骁龙855的算力



目录CONTENTS

◎ 算力：智能社会基石，数字经济引擎

◎ 芯片：异构是趋势，GPU将成为主流

◎ 展望：中国市场将高速增长，GPU国产化任重道远

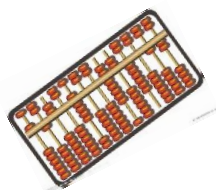
◎ 投资建议及风险提示

算力的定义以及载体演进



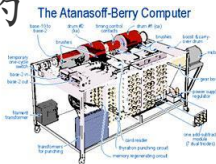
算力是设备根据内部状态的改变，每秒可处理的信息数据量

--2018年诺贝尔奖获得者William D. Nordhaus



算盘：基于人力的算力，简便的计算工具

1937年



电气化算力：
阿塔纳索夫-贝瑞发明的电子数字计算机

1993年



移动化算力：
IBM公司推出第一款智能手机 Simon

多样化算力
(1993年至今)

云端数据中心



边缘设备



智能手机和穿戴设备



1642年

机械式计算器：法国科学家Blasie Pascal发明的第一部机械式计算器



1947年

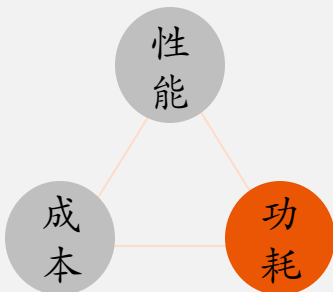
集成电路化算力：
-1947年，贝尔实验室发明晶体管
-1958年，集成电路问世

计算芯片的技术因素制约驱动算力布局向泛在演进

限制算力的核心因素

核心结论

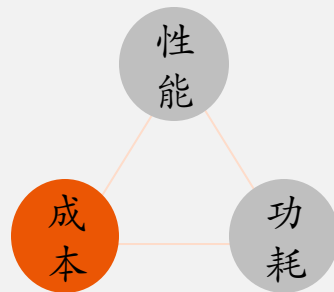
1、单核阶段



- 性能：芯片制程难以持续提高
- 成本：3nm以下制程成本难以被接受

- 硅基芯片单核制程将在3nm达到极限

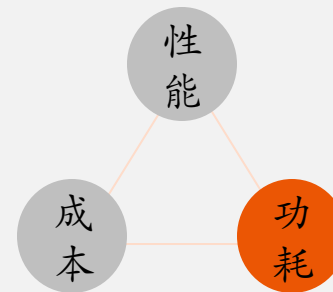
2、多核阶段



- 性能：核数增加放大架构间的不匹配
- 功耗：单位算力功耗将显著增加

- 处理器内核将在128核达到极限

3、网络化阶段

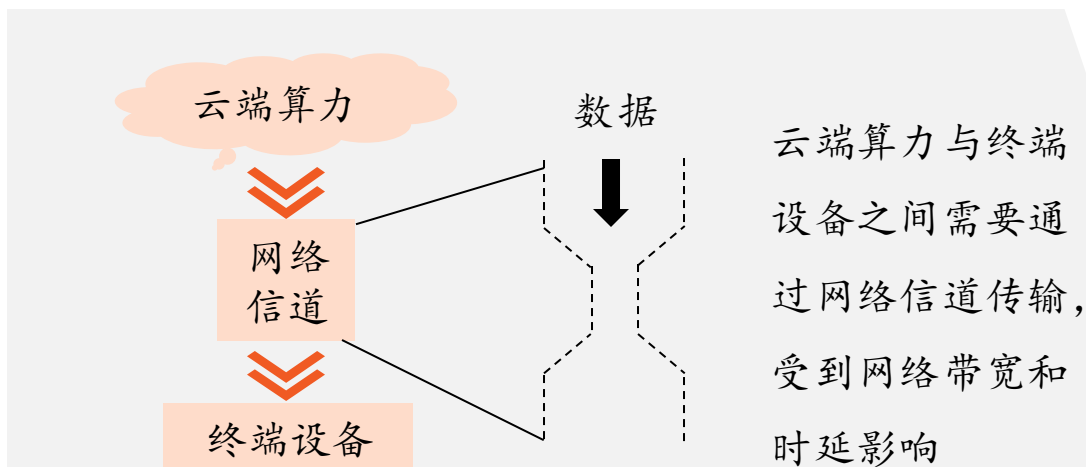


- 性能：网络带宽和时延性能受限
- 成本：带宽成本问题导致供需错配

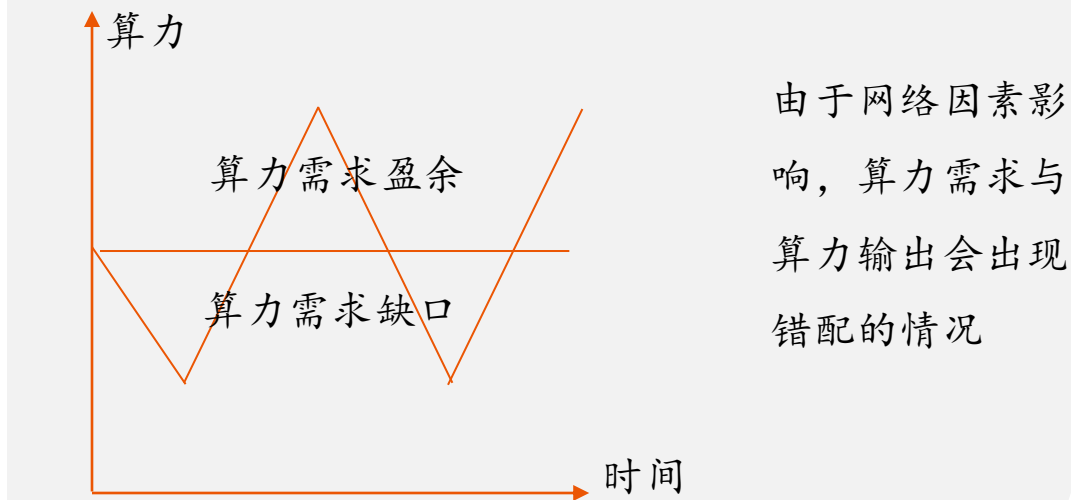
- 算力分布从集中向分布式演进

网络传输的制约驱动算力布局向设备侧靠近

网络限制



算力潮汐效应



云端算力



终端设备

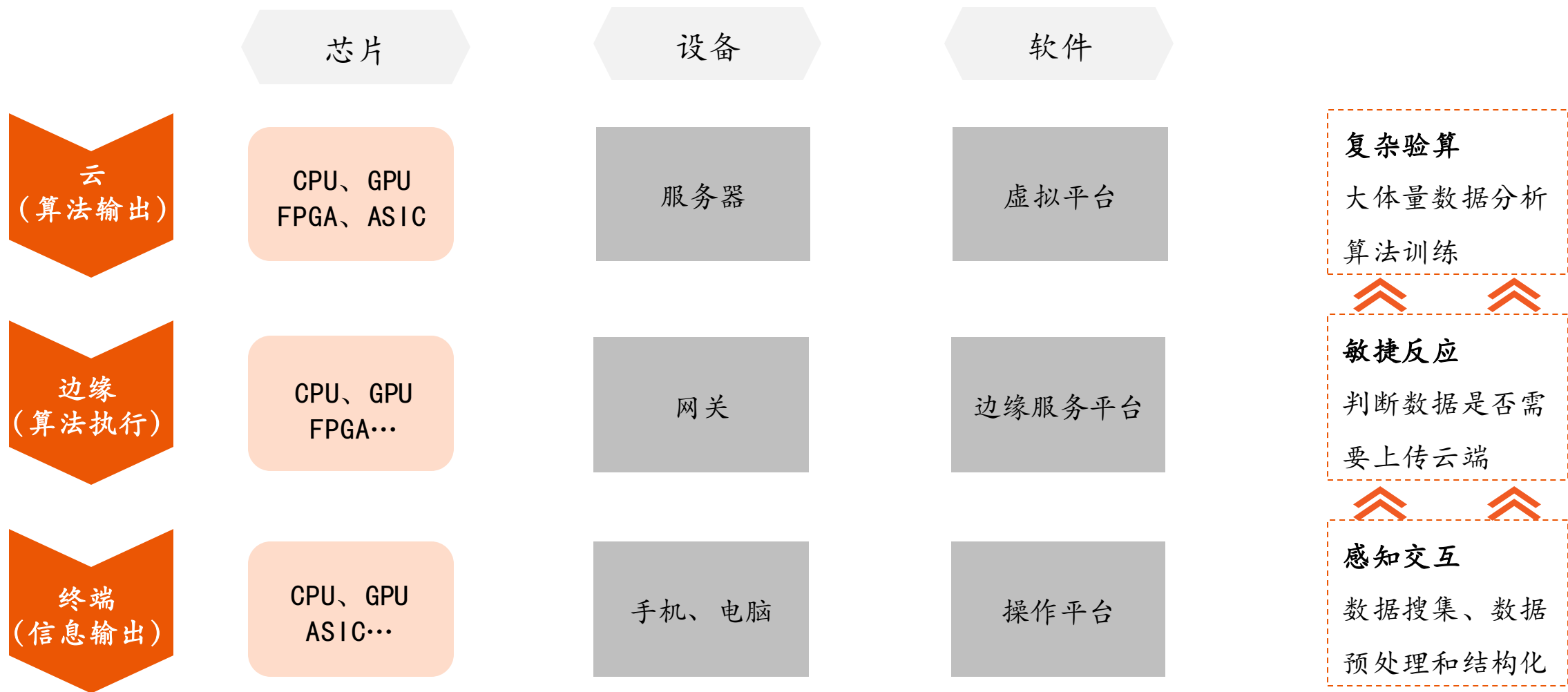


边缘设备

边缘算力的补充具有以下好处

- 1、更灵活：带宽高、时延低、可满足不同量级的算力需求。
- 2、成本低：边缘设备与终端物理距离近，传输成本低。

泛在算力所需要的计算芯片



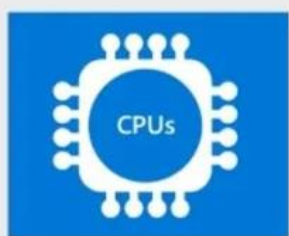
数据来源：华为公司，平安证券研究所

异构计算是灵活性和效率综合平衡的结果

- 训练指模拟人类接收、学习并理解外界信息能力的AI技术；推理指模拟人类通过学习、判断、分析等心理活动获取信息内含逻辑的AI技术。

训练：主要使用CPU、GPU，部分FPGA和ASIC

推理：主要使用CPU、FPGA、ASIC，部分GPU

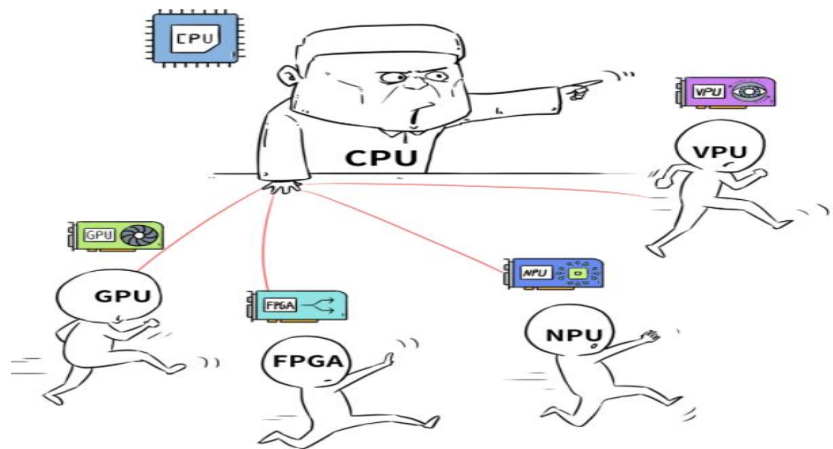


- 从部署的灵活性来看，CPU最为灵活，GPU次之，FPGA和ASIC分列最后两位。

- 从计算的效率来看，ASIC效率最高，FPGA次之，GPU和CPU则分列最后两位。

- 异构计算是一个平衡的结果：考虑到部署的灵活性和计算效率，异构计算是一个平衡的结果，CPU + GPU or FPGA or ASIC是趋势。

异构计算将成为算力基础设施的主要架构



■未来的AI计算，将形成以CPU为控制中心，GPU、FPGA、ASIC（NPU、VPU...）为特定场景加速卡的异构计算格局。

■异构计算是指不同类型的指令集和体系架构的计算单元组成的系统的计算方式，目前“CPU+GPU”以及“CPU+FPGA”都是受关注的异构计算平台。

■异构计算最大的优点是具有比传统CPU并行计算更高效率和低延迟的计算性能，尤其是在业界对计算性能需求水涨船高的情况下，异构计算变得愈发重要。

芯片种类	通用性	并行处理能力	处理速度	功耗	研发成本	量产成本	交付周期
CPU	高	低	低	高	低	中等	短
GPU	高	中等	中等	高	低	中等	中等
FPGA	中等	高	中等	中等	中等	高	中等
ASIC	低	高	高	低	高	低	长

数据来源：阿里云，平安证券研究所

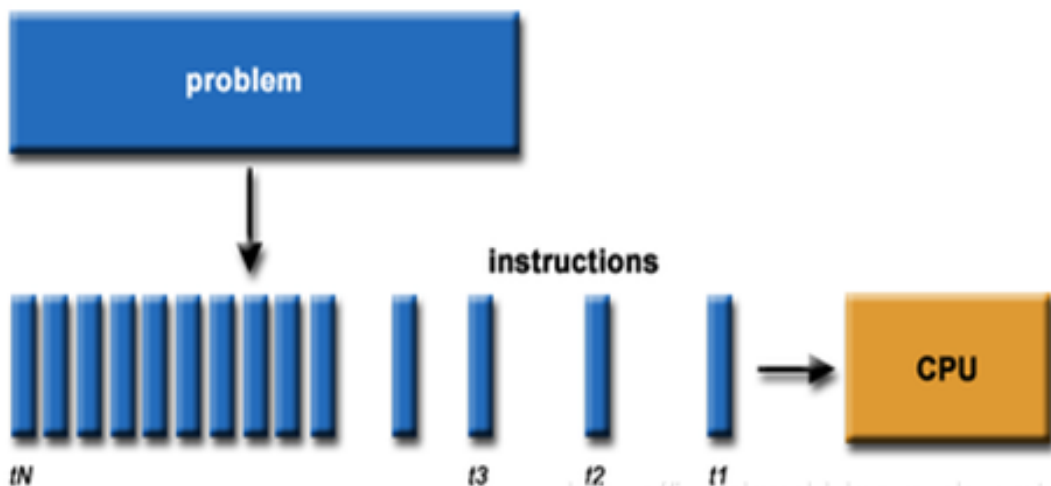
CPU擅长调度，计算能力一般

■ 有强大的 ALU（逻辑运算部件），时钟频率很高；有容量较大的 Cache，这些 Cache 占据相当一部分的片上空间；有复杂的控制逻辑；上述设计使得真正执行运算的 ALU 单元只占据了很小一部分的 CPU 片上空间。

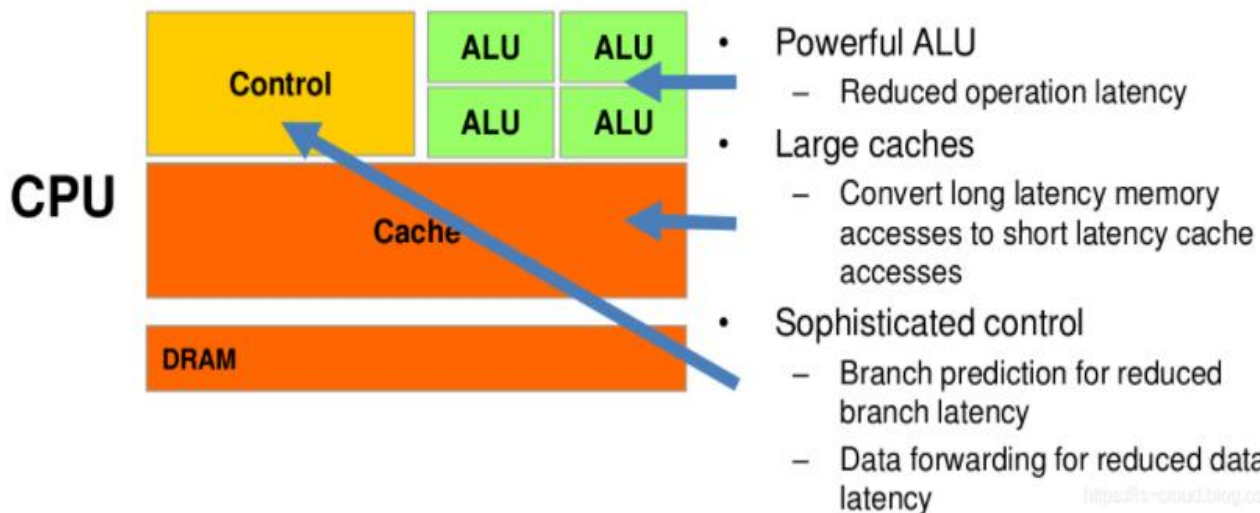


■ 具有多种功能的优秀领导者，优点在于调度、管理、协调能力强，但计算能力一般。

CPU采用串行计算架构



CPU的低延迟设计架构



数据来源：CSDN，平安证券研究所

注释：CPU，中央处理器，Central Processing Unit

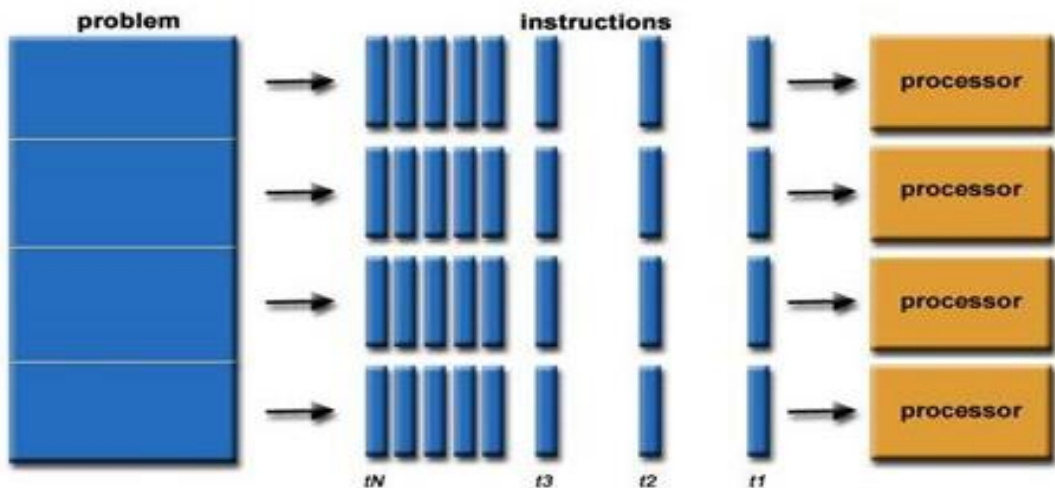
GPU擅长浮点计算，并行处理能力强

■有大量的ALU，Cache很小，缓存的目的不是保存后面需要访问的数据，而是为线程提高服务效率；没有复杂控制逻辑，没有分支预测等组件；上述设计，使得 GPU 擅长大规模并行计算任务。



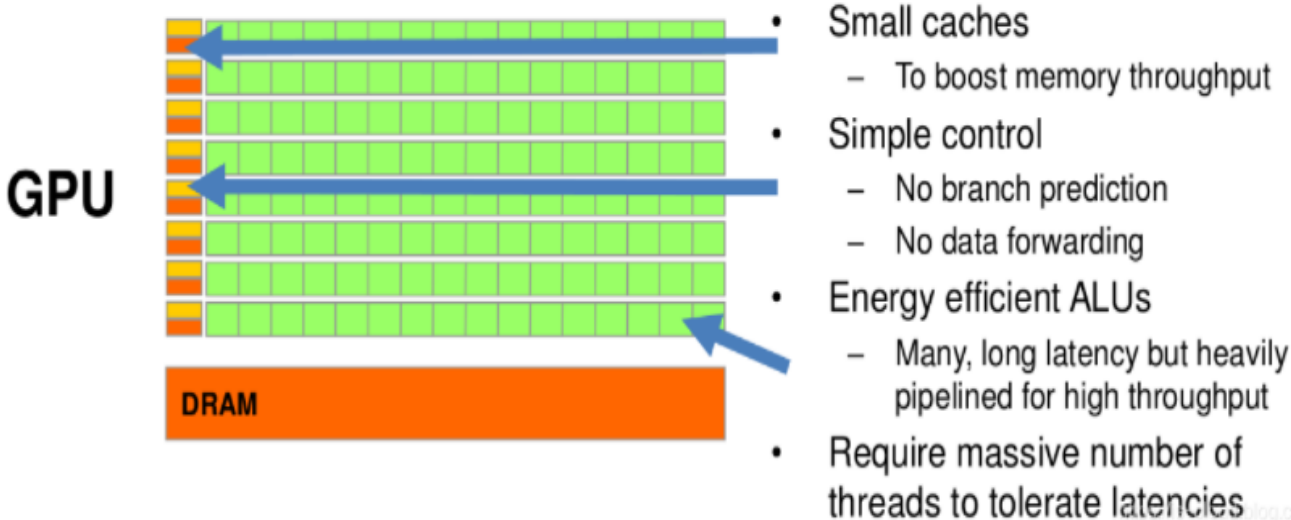
■擅长浮点计算，并行处理能力强，能做到几千核高并发，适合图形处理、机器学习训练等。

GPU采用并行计算架构



数据来源：CSDN，平安证券研究所

GPU的高吞吐设计架构



注释：GPU，图形处理器，Graphics Processing Unit

FPGA可灵活编程，擅长固定业务的高速处理

■可以实现比 GPU 更高的并发处理，在密集处理和高并发能力上占优，而且功耗比 CPU和GPU低。

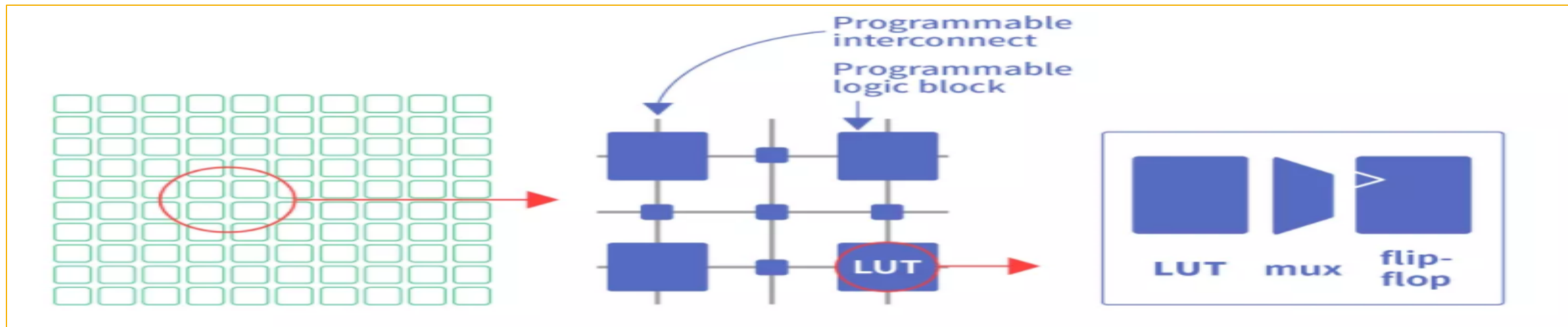


■内部有大量极细粒度的基本单元，但是每个单元的计算能力都远远低于 CPU 和 GPU 中的 ALU模块；速度和功耗相对ASIC仍然存在不小差距；成本要高于ASIC。



■擅长根据特定需求进行灵活编程，可以高速处理一些相对固定的业务逻辑，比如用于深度学习的检测阶段。

FPGA的内部架构



ASIC可根据需求定制，只能针对特定场景

■与通用集成电路相比具有体积更小、功耗更低、可靠性高、性能高、保密性强、成本低等优点。

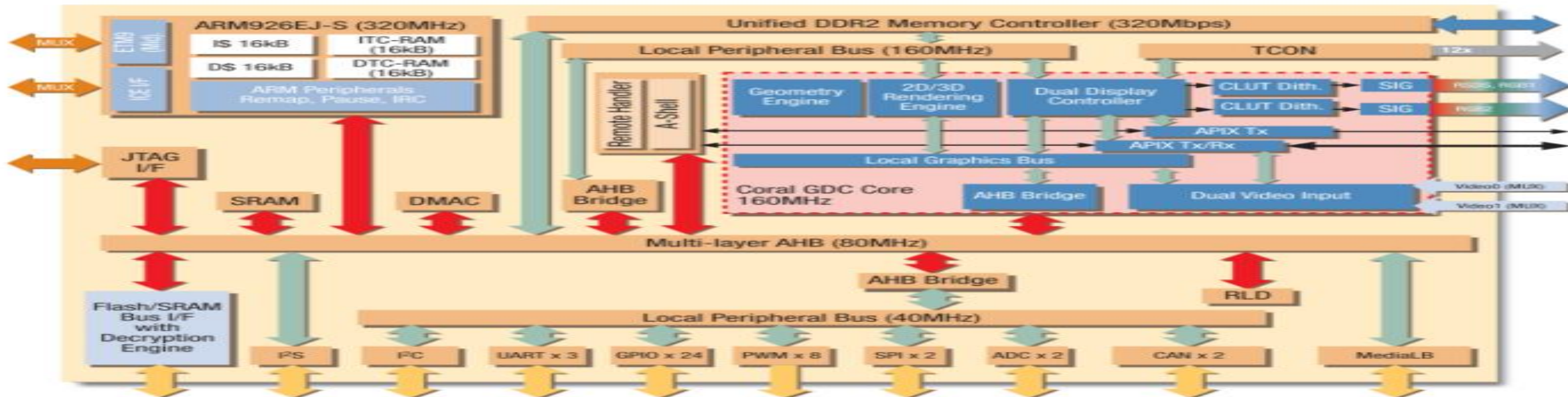


■基于ASIC开发人工智能芯片更像是电路设计，需要反复优化，需要经历较长的流片周期，故开发周期较长。



■计算能力和计算效率都可以根据算法需要进行定制，只能针对特定的某几个应用场景。

某ASIC内部架构示意



数据来源: CSDN, 平安证券研究所

注释: ASIC, Application Specific Integrated Circuit, 专用集成电路

从技术趋势来看，GPU将成为主流AI芯片

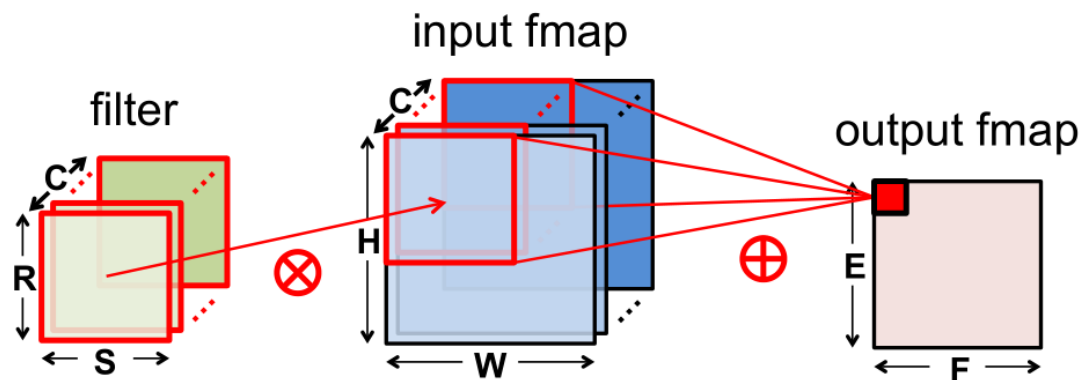
在图形图像处理领域，GPU最终全面取代图形加速卡



通用标准高性能卡是未来云端高性能计算发展趋势



典型的卷积运算



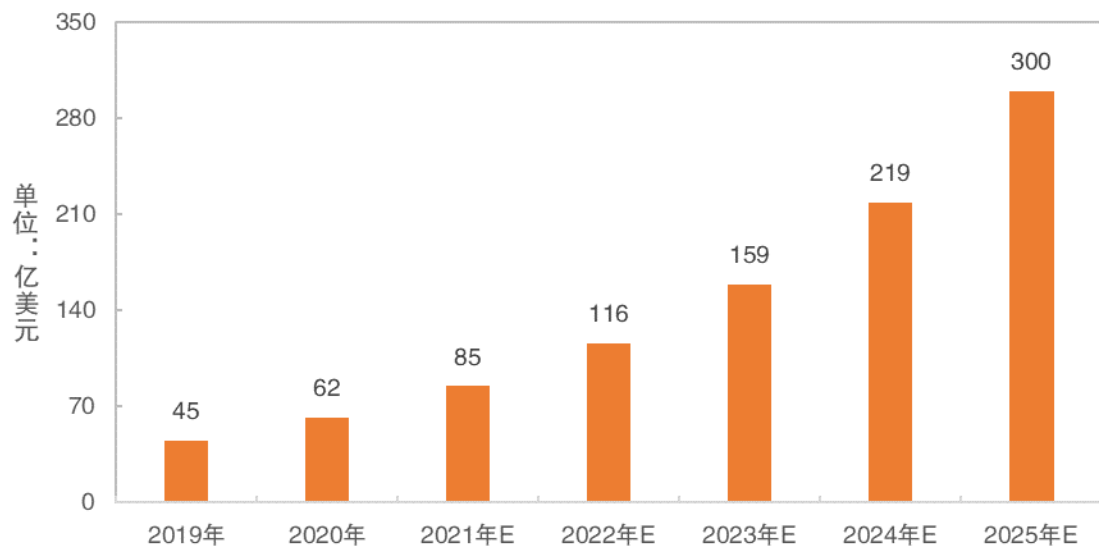
- **GPU最适合AI算法:** AI算法要用到大量卷积运算，对大量不同的数据进行同样的乘加MAC运算，这种运算和GPU Shader执行的SIMD运算高度相似。
- 使用特定算法加速的芯片只能用在某一些特定场景，通用型计算芯片最终将会胜出。

全球AI芯片市场将高速增长，GPU份额有望超过50%

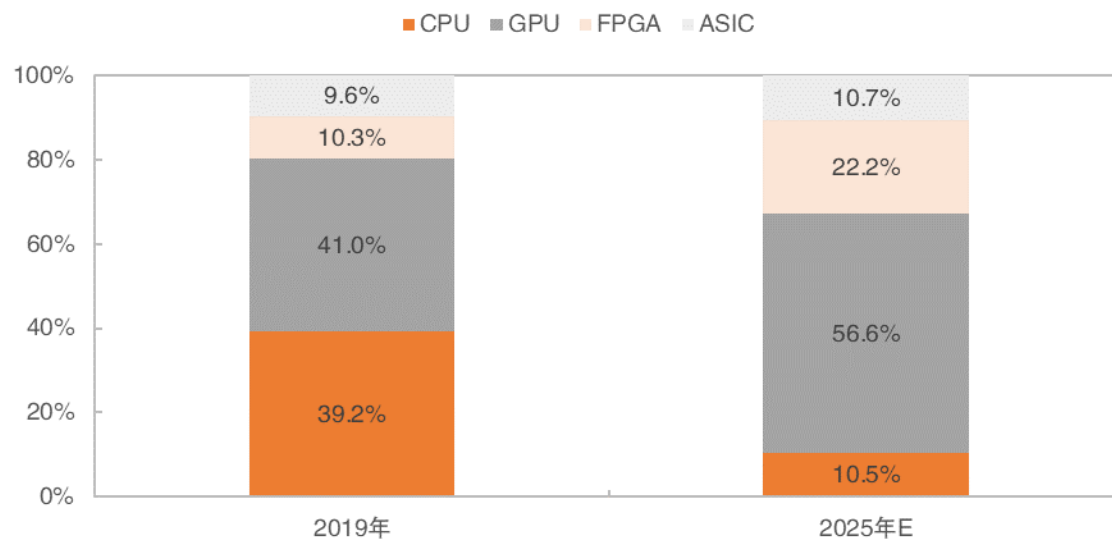
■全球AI芯片市场规模将高速增长：预计到2025年，全球市场规模有望达到300亿美元，2019年-2025年年均复合增速约37%。

■GPU市场份额有望达50%：AI的应用需要用到大量卷积算法，正是GPU擅长的领域；预计2025年，市场份额将达到约57%。

全球AI计算芯片市场规模



全球AI计算芯片市场份额





目录CONTENTS

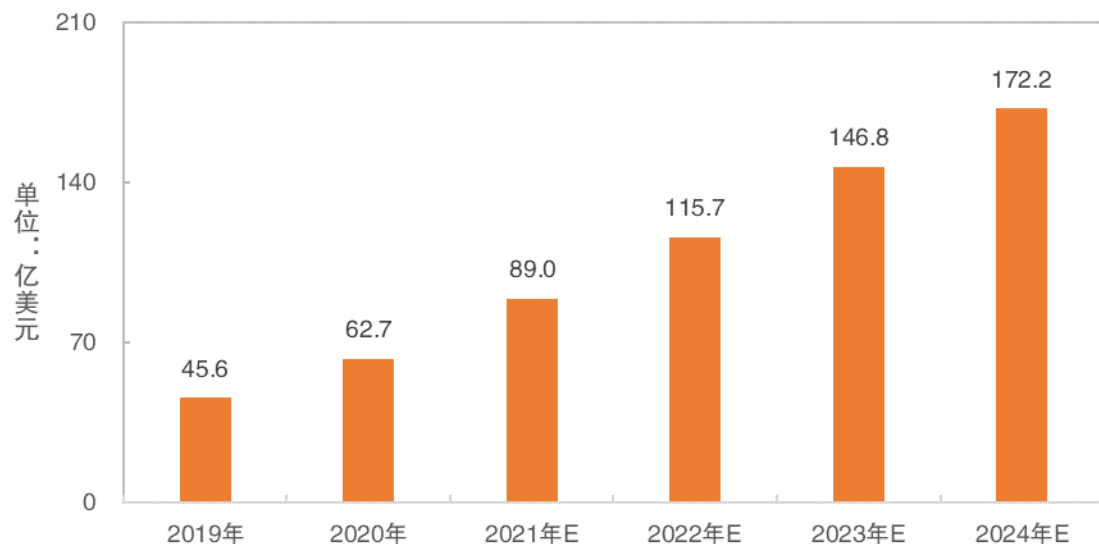
- ◎ 算力：智能社会基石，数字经济引擎
- ◎ 芯片：异构是趋势，GPU将成为主流
- ◎ 展望：中国市场将高速增长，GPU国产化任重道远
- ◎ 投资建议及风险提示

中国人工智能市场将高速增长，服务器是市场主体

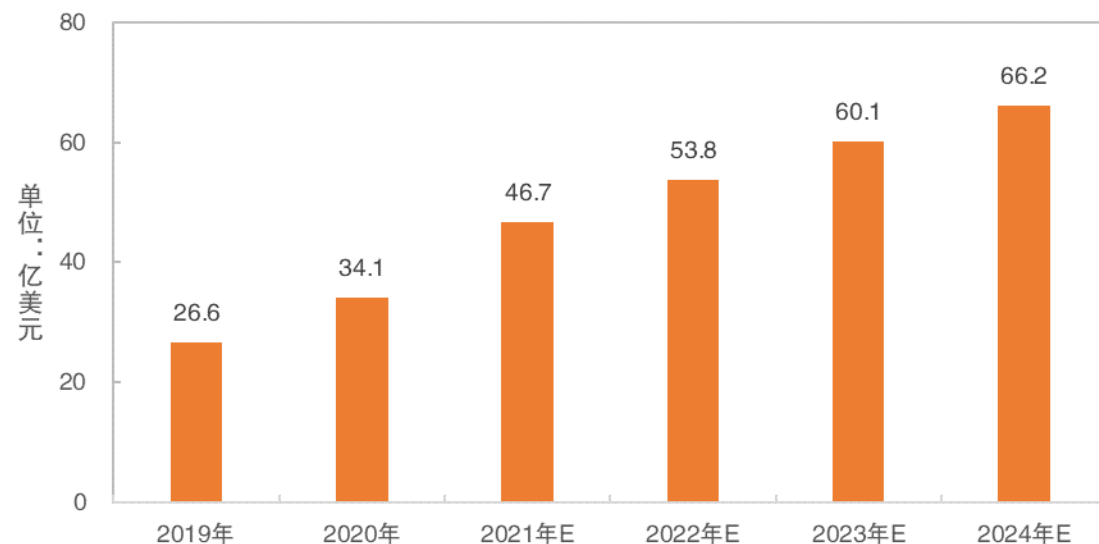
■中国人工智能技术市场将高速增长：预计到2024年，中国人工智能技术市场规模将达到172亿美元；全球占比将从2020年12.5%上升到15.6%，是全球市场增长的主要驱动力。

■服务器在整体市场中将保持主体地位：预计到2024年，中国人工智能服务器市场规模将达到66亿美元，在整体市场中占比约38%；相比2019年，下降约20个百分点。

中国人工智能技术市场规模



中国人工智能服务器市场规模



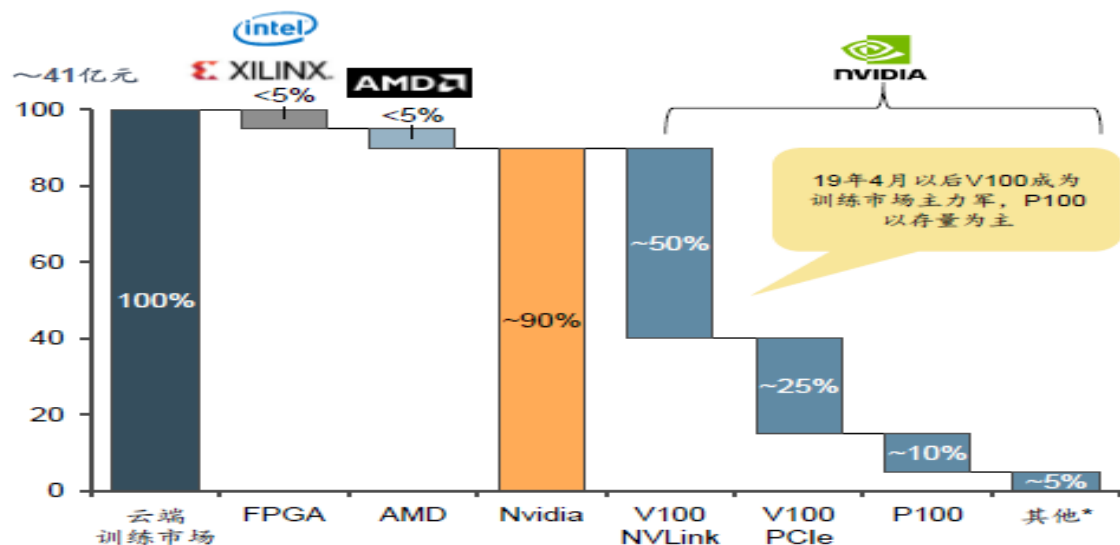
数据来源：IDC，平安证券研究所

训练芯片市场以GPU为主，国内公司有待突破

■训练芯片市场以GPU为主：2019年，中国训练芯片市场规模约41亿元，英伟达公司凭借V100系列等产品占据了90%的市场份额；考虑到AMD公司的产品也是GPU，GPU占据了训练芯片市场95%的份额。

■国内公司有待突破：虽然华为公司推出了Atlas系列产品，并且在性能、功耗等主要指标上均不弱于英伟达V100；但是由于软件生态等因素，尚未实现大规模商用。

2019年中国训练芯片市场格局



2019年中国训练芯片市场主要产品

	英伟达 V100	AMD M150	华为 Atlas300-9000
工艺	12nm	7nm	7nm+
软件生态	CUDA、OpenCL 2.0	ROCm 2.0、OpenCL 2.0	自研Mindspore
性能	<u>7.8TFLOPS@FP64</u>	<u>6.6TFLOPS@FP64</u>	<u>256TFLOPS@FP16</u>
功耗	300W	300W	310W
应用场景	训练、推理、超算	训练、推理、超算	训练、推理、超算

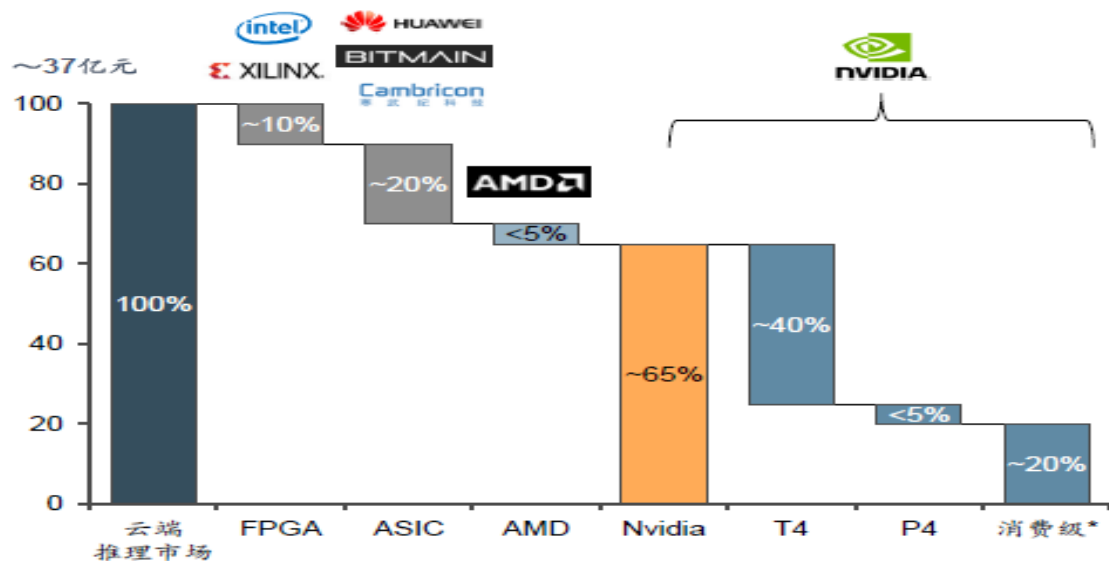
数据来源：天数智芯，平安证券研究所

推理芯片市场呈现多元化趋势，国内公司开始展露头角

■推理芯片市场多元化趋势较为明显：2019年中国推理芯片市场规模约37亿元，相比于训练市场，FPGA、ASIC份额均有所提升，AMD的GPU产品份额亦有所提升；GPU产品不再是一家独大。

■国内公司开始展露头角：以华为公司、寒武纪和比特大陆为代表的中国公司，凭借ASIC产品，通过合适的价格和解决方案，占据了大约20%的市场份额。

2019年中国推理芯片市场格局



2019年中国推理芯片市场主要产品

	英伟达 T4	英伟达 P4	英伟达 2080Ti	AMD MI8
工艺	12nm	16nm	12nm	28nm
生态	CUDA、OpenCL 2.0	CUDA、OpenCL 2.0	CUDA、OpenCL 2.0	ROCm 2.0、OpenCL 2.0
性能	8.1TFLOPS@FP32	5.5TFLOPS@FP32	13.4TFLOPS@FP32	8.19TFLOPS@FP32
功耗	70W	75W	250W	175W
应用场景	训练、推理、超算	训练、推理、超算	训练、推理、超算	训练、推理、超算

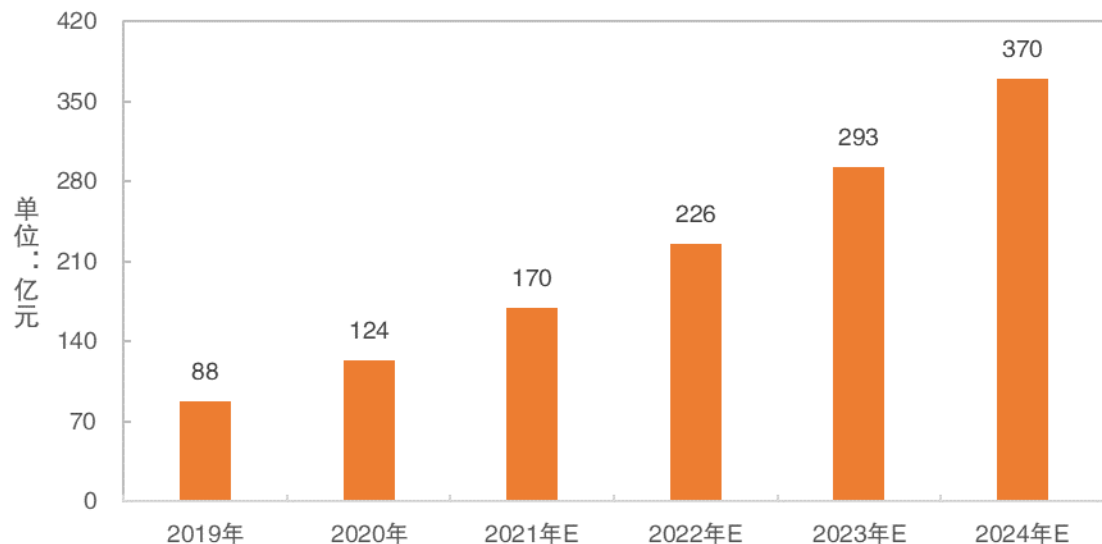
数据来源：天数智芯，平安证券研究所

中国GPU芯片板卡市场将高速增长，互联网和安防/政府是主要领域

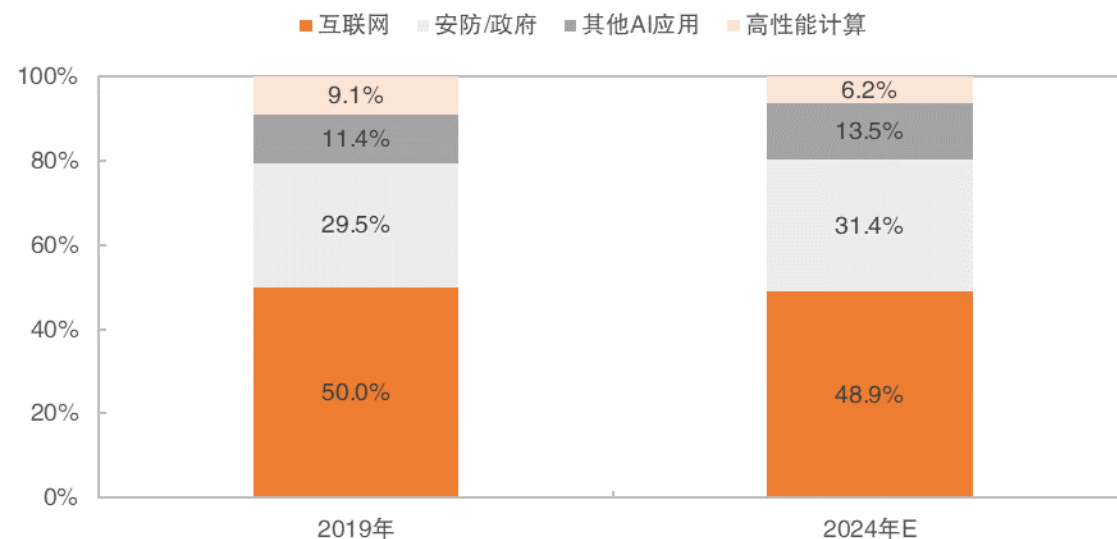
■中国GPU芯片板卡市场将高速增长：预计2024年，中国GPU芯片板卡市场规模将达到370亿元，年均复合增速约30%；训练市场规模占比约36%，推理市场占比约58%，高性能计算市场约6%。

■互联网和安防/政府是主要领域：预计2024年，互联网和安防/政府行业市场份额将分别达到48.9%和31.4%，依然占据市场的主要地位；主要还是因为这两个行业是AI应用的重要需求方。

中国GPU芯片板卡市场规模预测



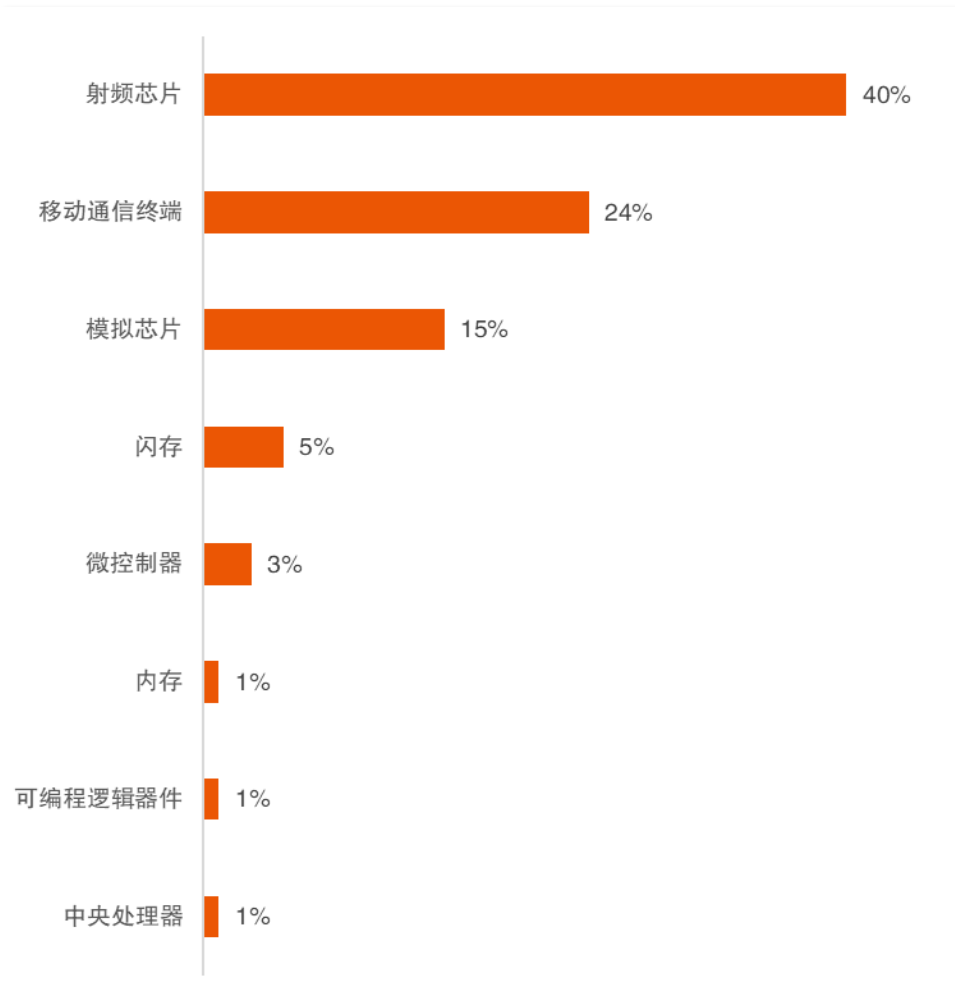
中国GPU芯片板卡市场份额（应用领域）



数据来源：公司网站，平安证券研究所

谨慎估计GPU芯片的国产化规模约37-56亿元

主要芯片国产化率情况



各类芯片代表性公司

- 昂瑞微、卓胜微
- 华为海思、紫光展锐
- 圣邦微、思瑞浦
- 兆易创新、长江存储
- 兆易创新、华大半导体
- 长鑫、晋华
- 紫光同创、国微电子
- 海光、龙芯

■手机产业链芯片国产化率超过15%：由于中国手机厂商占据了全球主要份额，带动了国内产业链的快速发展；与此同时，芯片与软件应用耦合紧密度相对较低，因此国产化率大于15%。

■服务器产业链国产化率低于5%：由于CPU指令集复杂程度较高，产业链主要环节均在海外，国产化率低于5%。

■谨慎估计GPU芯片的国产化规模约37亿元：由于GPU指令集复杂程度相对较低，国内有一定基础，国产化率有望达到10%-15%，对应37-56亿元的市场规模。

数据来源：平安证券研究所



目录CONTENTS

- ◎ 算力：智能社会基石，数字经济引擎
- ◎ 芯片：异构是趋势，GPU将成为主流
- ◎ 展望：中国市场将高速增长，GPU国产化任重道远
- ◎ 投资建议及风险提示

投资建议

芯片设计

软件生态

芯片制造

产业
瓶颈

■无论是GPU，还是FPGA，两种芯片的国产化率都很低

■GPU和FPGA两类芯片的设计和应用都需要有自主软件生态支持

■GPU和FPGA的制程都需要用到16nm甚至7nm和5nm的制程



投资
建议

■建议关注国内FPGA龙头紫光国微和GPU已有应用场景的景嘉微

■建议关注具有完全自主知识产权的FPGA公司紫光国微

■建议关注国内具有晶圆先进制程升级潜力的中芯国际公司

建议关注公司汇总

代码	名称	市值 (亿元)	收盘价 (元)	EPS (元)		PE		评级
				2021年E	2022年E	2021年E	2022年E	
002049.SZ	紫光国微	808	133.20	2.27	3.04	59	44	推荐
688981.SH	中芯国际	1081	55.78	0.44	0.54	127	103	未评级
300474.SZ	景嘉微	245	81.36	1.16	1.67	70	49	未评级

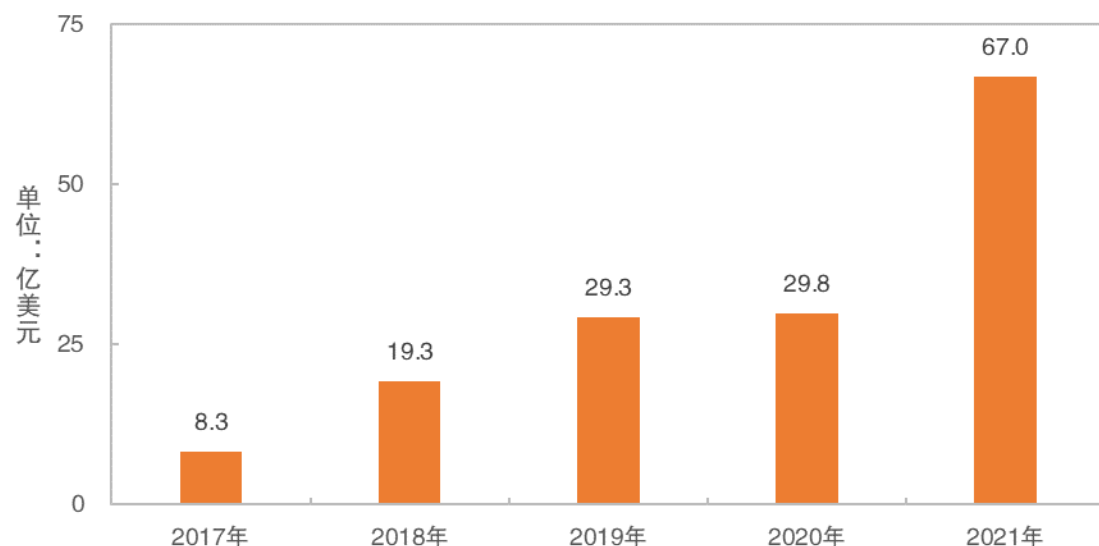
注：收盘价和市值截止2021年5月28日；未评级公司盈利预测为Wind一致预期；中芯国际市值为：A股市值=（A股流通股+A股限售股）*A股收盘价

英伟达：全球领先的GPU供应商

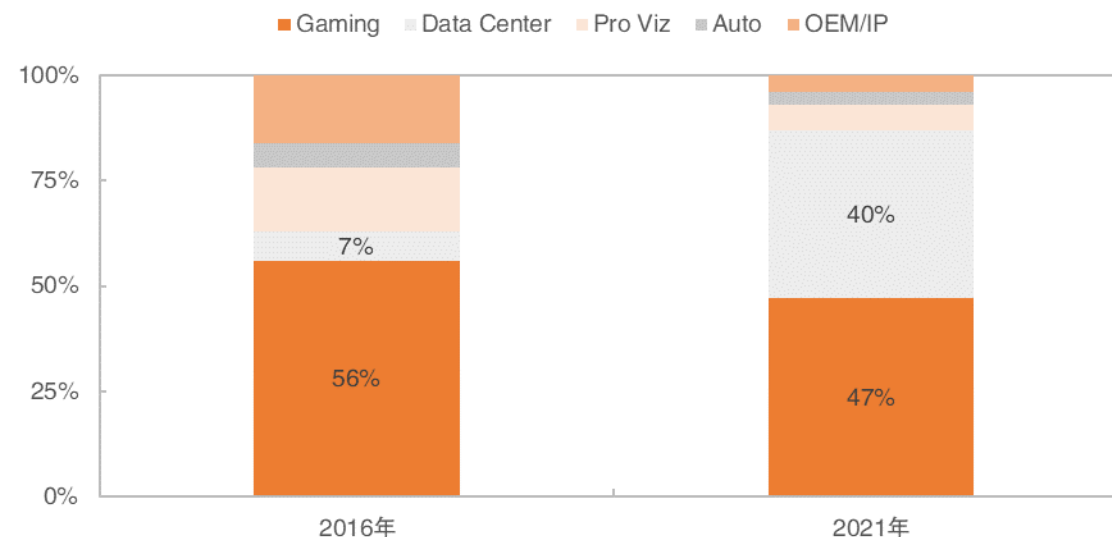
■英伟达是全球领先的GPU供应商：公司是全球GPU市场绝对领先者，2021财年公司营业收入规模达到了约167亿美元；数据中心收入占比从2016年的7%提升到了40%。

■A100产品处于业内绝对领先地位：NVIDIA A100由NVIDIA Ampere架构提供支持，作为NVIDIA数据中心平台的引擎，A100的性能比上一代产品提升高达20倍。

英伟达数据中心GPU营业收入情况



英伟达营业收入行业拆分



数据来源：Wind，平安证券研究所

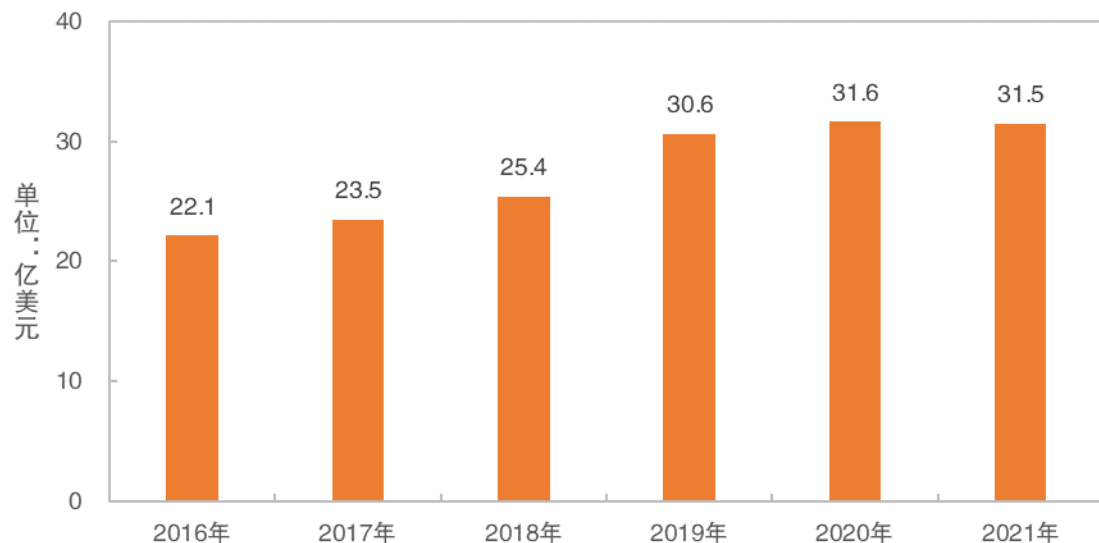
注：英伟达财年为上年1月到当年1月

赛灵思：全球领先的FPGA供应商

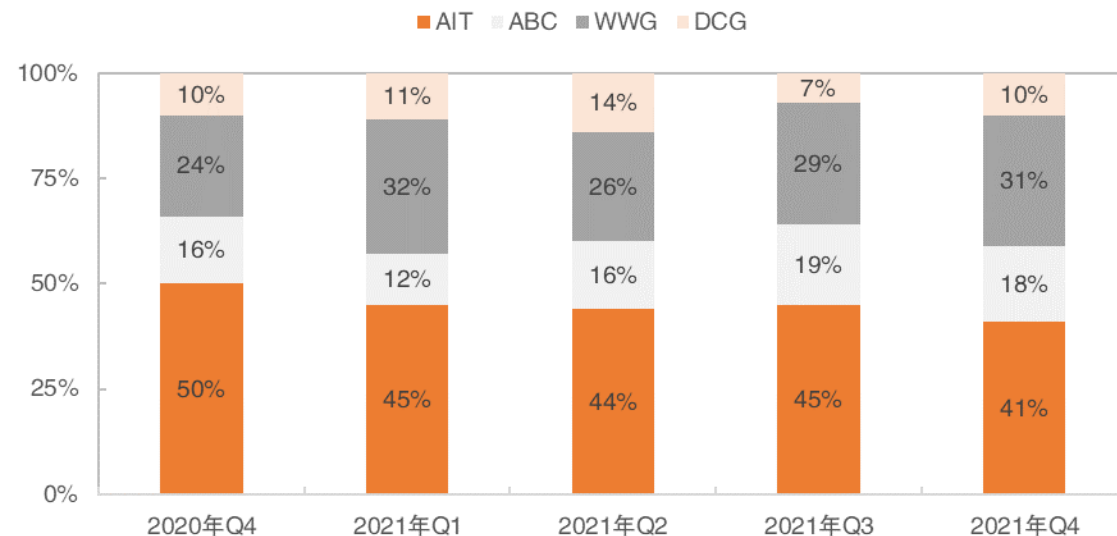
■赛灵思是全球领先的FPGA供应商：公司是FPGA产品的发明者，拥有574项重要的行业专利；2021财年营业收入规模达到了31.5亿美元；2019年市场占比达到了42%，是FPGA行业领先的供应商。

■无线和航天军工是公司收入主要来源：AIT和WWG对应航天军工以及无线通信两个行业，在过去的5个季度，这两个行业为公司贡献了70%以上的收入；DCG（数据中心）行业收入贡献只有10%左右。

赛灵思营业收入情况



赛灵思营业收入行业拆分



数据来源：Wind，平安证券研究所

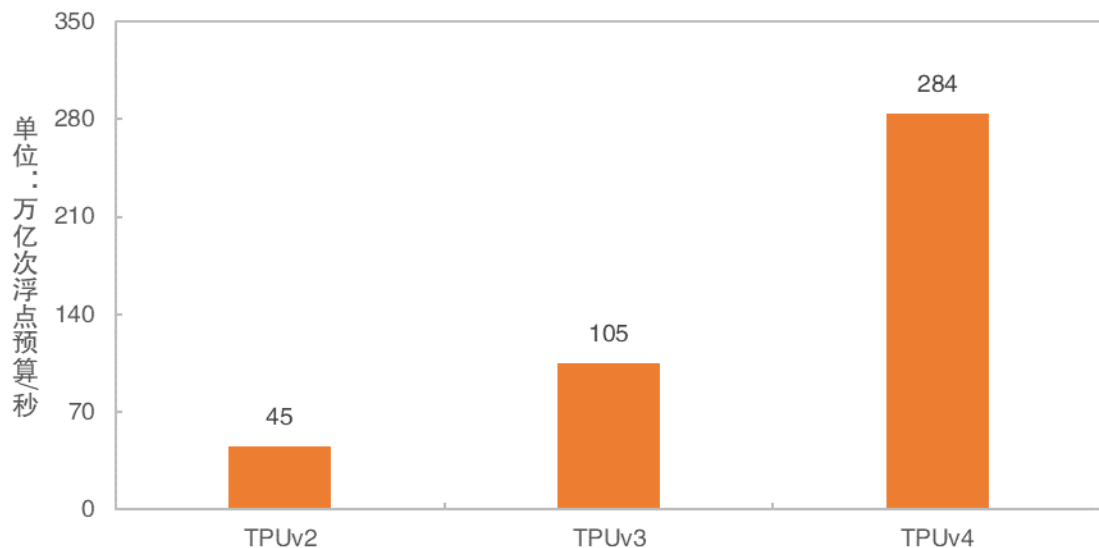
注：赛灵思财年为上年3月到当年3月

Google: 自用AI芯片研发再下一程, TPU4正式面世

■Google公司于近日发布TPUv4: TPU是Google公司自己研发的专门用于人工智能计算的芯片, 可以归为ASIC产品; 近日发布的TPUv4产品, 单片计算性能达到了TPUv3的2.7倍。

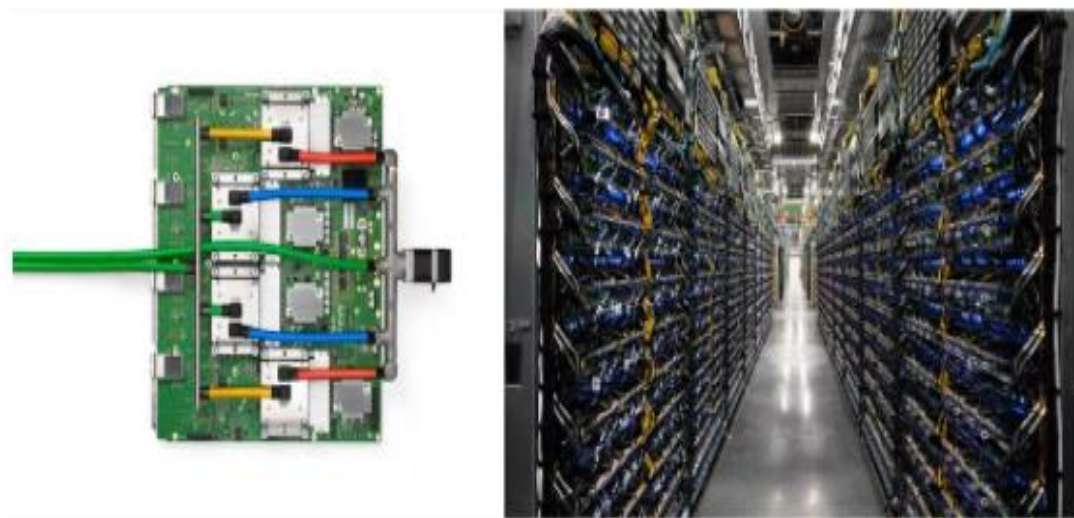
■训练效率显著提升: 根据MLperf发布的数据显示, 针对不同的AI模型, TPUv4 Pod的训练效率是TPUv3 Pod训练效率的2.2-3.7倍; 已经接近真正意义上的E级别超算时代。

Google公司历代TPU产品性能对比



数据来源: 数据中心前沿技术, 平安证券研究所

部署在Google公司数据中心内部的TPUv4

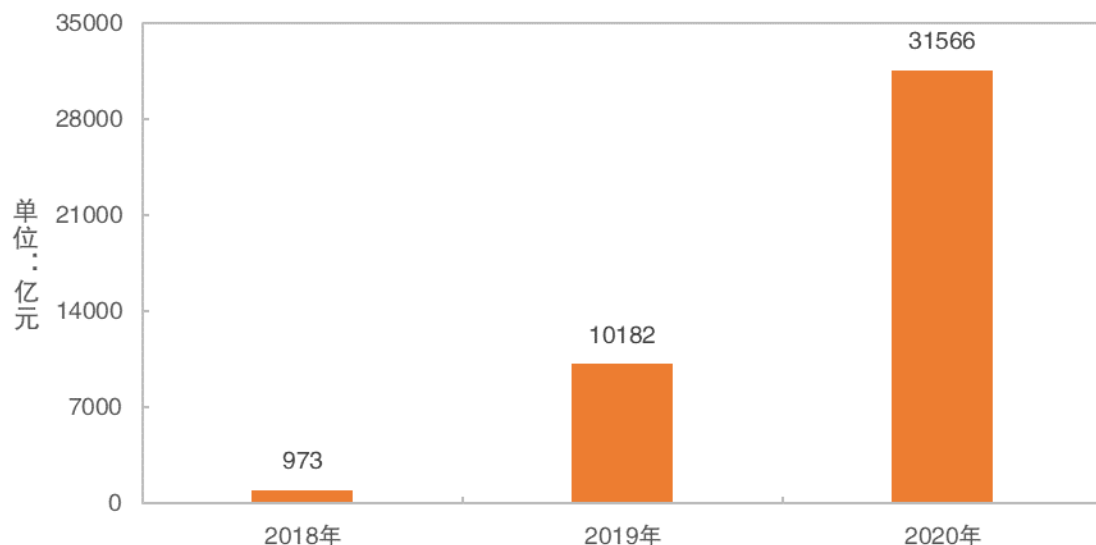


紫光国微：拥有国产自主知识产权的FPGA龙头

■拥有国产FPGA自主知识产权：孙公司紫光同创专门从事FPGA产品研发与生产销售10余年，拥有专利200项，研发出中国第一款国产自主知识产权千万门级高性能FPGA产品-Titan（40nm）。

■紫光同创销售规模快速增长：2020年销售规模达到了31566万元，与2018年973万元的收入相比，实现了几何级的增长；随着公司28nm制程的产品量产，收入规模有望持续增长。

紫光同创营业收入情况



数据来源：Wind，平安证券研究所

紫光同创40nm工艺Titan系列FPGA主要参数

参数名称	Titan系列-PGT180H
等效LUT4	174019
Flip-Flops (个)	217524
分布式RAM (Kbit)	570
PLL (时钟资源)	8
最大用户IO	611

风险提示

1、互联网公司芯片自研带来的风险

从Google的情况来看，互联网公司自研AI芯片会是趋势；对于那些为互联网公司开发定制芯片的公司来说，这将是一个风险点。

2、中美贸易摩擦的风险

部分公司GPU芯片制作需要用到7nm的先进制程，若是中美贸易摩擦升级，有可能会使这些公司无法使用7nm的先进制程来生产芯片。

3、人工智能在安防/政府业务领域应用不及预期的风险

安防/政府行业是人工智能应用渗透率较高的领域；若是安防/政府行业需求不及预期，将使得国产GPU芯片等各类AI计算芯片的出货量不及预期。

电子信息团队

行业	分析师	邮箱	资格类型	编号
通信	朱琨	zhukun368@pingan.com.cn	投资咨询	S1060518010003
智能制造	吴文成	wuwencheng128@pingan.com.cn	投资咨询	S1060519100002
电子	徐勇	xuyong318@pingan.com.cn	投资咨询	S1060519090004
计算机	付强	fuqiang021@pingan.com.cn	投资咨询	S1060520070001
	闫磊	yanlei511@pingan.com.cn	投资咨询	S1060517070006

股票投资评级：

强烈推荐（预计6个月内，股价表现强于沪深300指数20%以上）

推 荐（预计6个月内，股价表现强于沪深300指数10%至20%之间）

中 性（预计6个月内，股价表现相对沪深300指数在±10%之间）

回 避（预计6个月内，股价表现弱于沪深300指数10%以上）

行业投资评级：

强于大市（预计6个月内，行业指数表现强于沪深300指数5%以上）

中 性（预计6个月内，行业指数表现相对沪深300指数在±5%之间）

弱于大市（预计6个月内，行业指数表现弱于沪深300指数5%以上）

公司声明及风险提示：

负责撰写此报告的分析师（一人或多人）就本研究报告确认：本人具有中国证券业协会授予的证券投资咨询执业资格。

本公司研究报告是针对与公司签署服务协议的签约客户的专属研究产品，为该类客户进行投资决策时提供辅助和参考，双方对权利与义务均有严格约定。本公司研究报告仅提供给上述特定客户，并不面向公众发布。未经书面授权刊载或者转发的，本公司将采取维权措施追究其侵权责任。

证券市场是一个风险无时不在的市场。您在进行证券交易时存在赢利的可能，也存在亏损的风险。请您务必对此有清醒的认识，认真考虑是否进行证券交易。市场有风险，投资需谨慎。

免责声明：

此报告旨在发给平安证券股份有限公司（以下简称“平安证券”）的特定客户及其他专业人士。未经平安证券事先书面明文批准，不得更改或以任何方式传送、复印或派发此报告的材料、内容及其复印本予任何其他人。

此报告所载资料的来源及观点的出处皆被平安证券认为可靠，但平安证券不能担保其准确性或完整性，报告中的信息或所表达观点不构成所述证券买卖的出价或询价，报告内容仅供参考。平安证券不对因使用此报告的材料而引致的损失而负上任何责任，除非法律法规有明确规定。客户并不能仅依靠此报告而取代行使独立判断。

平安证券可发出其它与本报告所载资料不一致及有不同结论的报告。本报告及该等报告反映编写分析员的不同设想、见解及分析方法。报告所载资料、意见及推测仅反映分析员于发出此报告日期当日的判断，可随时更改。此报告所指的证券价格、价值及收入可跌可升。为免生疑问，此报告所载观点并不代表平安证券的立场。

平安证券在法律许可的情况下可能参与此报告所提及的发行商的投资银行业务或投资其发行的证券。

平安证券股份有限公司2021版权所有。保留一切权利。