浙商证券股份有限公司
ZHESHANG SECURITIES CO.LTD

大模型算力需求驱动AI服务器行业高景气

—— AI服务器行业报告

行业评级：看好

2022年4月5日

分析师 刘雯蜀
证书编号 S1230523020002

分析师 李佩京
证书编号 S1230522060001

摘要

1、算力场景向AI升级，CPU+GPU是核心

- 服务器随着场景需求经历通用服务器-云服务器-边缘服务器-AI服务器四种模式，AI服务器采用GPU增强其并行计算能力；
- AI服务器按应用场景可分为训练和推理，训练对芯片算力的要求更高，根据IDC，随着大模型的应用，2025年推理算力需求占比有望提升至60.8%；
- AI服务器按芯片类型可分为CPU+GPU、CPU+FPGA、CPU+ASIC等组合形式，CPU+GPU是目前国内的主要选择（占比91.9%）；
- AI服务器的成本主要来自CPU、GPU等芯片，占比25%-70%不等，对于训练型服务器其80%以上的成本来源于CPU和GPU。

2、 ChatGPT等大模型训练和推理需求激增驱动AI服务器市场高速增长

- 据ARK Invest预测，Chat GPT-4参数量最高达15000亿个，由于参数量与算力需求间存在正比关系，所以可推算GPT-4算力需求最高达到31271 PFlop/s-day。随着国内外厂商加速布局千亿级参数量的大模型，**训练需求有望**进一步增长，叠加大模型落地应用带动**推理需求**高速增长，共同驱动算力革命并助推AI服务器市场及出货量高速增长。

3、 国产芯片推理接近国际一流水平，国产AI服务器有望受到下游需求拉动

- 美国对中国禁售英伟达高性能芯片A100和H100，英伟达特供中国的削弱互联带宽的版本A800或为当前可替代方案；
- 以海光信息、壁仞科技等为代表的国产GPU部分单卡指标接近英伟达，在推理场景中具有一定竞争力；
- 国产AI服务器厂商全球份额超35%，浪潮信息位列榜首；国产AI服务器厂商各具优势，有望受到下游需求拉动；

4、建议关注：浪潮信息、中科曙光、拓维信息、工业富联、神州数码、紫光股份

风险提示

1、国外制裁范围扩大的风险；
2、国内厂商发展不及预期的风险；
3、下游市场需求不及预期的风险；
4、版权、伦理和监管风险等；

1、算力场景向AI演进，
CPU+GPU是AI服务器的核心部件

1.1 AI服务器：算力场景转向AI，加入GPU的AI服务器应运而生

浙商证券股份有限公司
ZHESHANG SECURITIES CO.LTD

	介绍	优势
通用服务器	提供计算服务的硬件设备，位置相对固定；	——
云服务器	将规模级的底层服务器通过集约化、虚拟化来构建云端资源池，并从资源池中调配计算资源组建而成；	支持资源的弹性调配，具有扩展性和灵活性，采用分布式存储；
边缘服务器	将终端采集的数据，交由本地的网关设备直接进行处理后，再上传到云端，使应用程序、数据和计算能力更接近用户；	减少网络传输资源的消耗，满足对实时性的要求，解决安全和隐私问题；
AI服务器	采用异构形式，通过加入GPU采用并行计算模式，解决CPU提供算力时核心数已接近极限的问题，可应用处理密集型运算；	采用GPU并行计算提高算力；

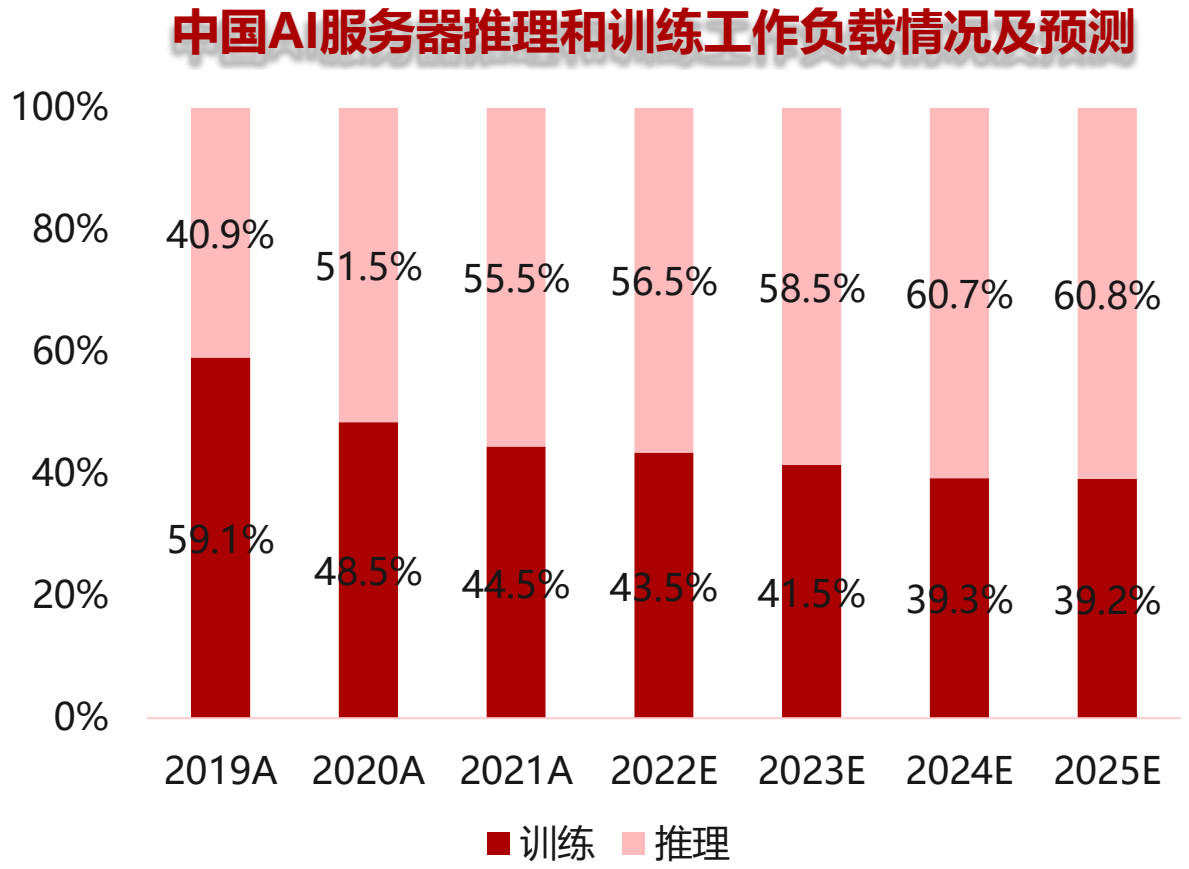
PC单机 → 云架构 → 加入边缘层 → GPU增强并行计算能力

资料来源：立尔讯、陆零网络、亿速云、聚英电子、蓝卡科技、浙商证券研究所

1.2 分类：按应用场景可分为训练和推理，训练对算力要求更高

- AI服务器按应用场景可分为训练和推理两种，2021年中国AI服务器推理负载占比约55.5%，未来有望持续提高；
- 训练对芯片算力要求更高，推理对算力的要求偏低；

	训练	推理
概念	指借助已有的大量数据样本进行学习，获得诸如更准确的识别和分类等能力的过程；	对于新的数据，使用经过训练的算法完成特定任务
算力要求	要求训练芯片应具有强大的单芯片计算能力	对算力的要求较低
部署位置	训练芯片大多部署于云端	推理芯片大多会部署于云端和边缘侧

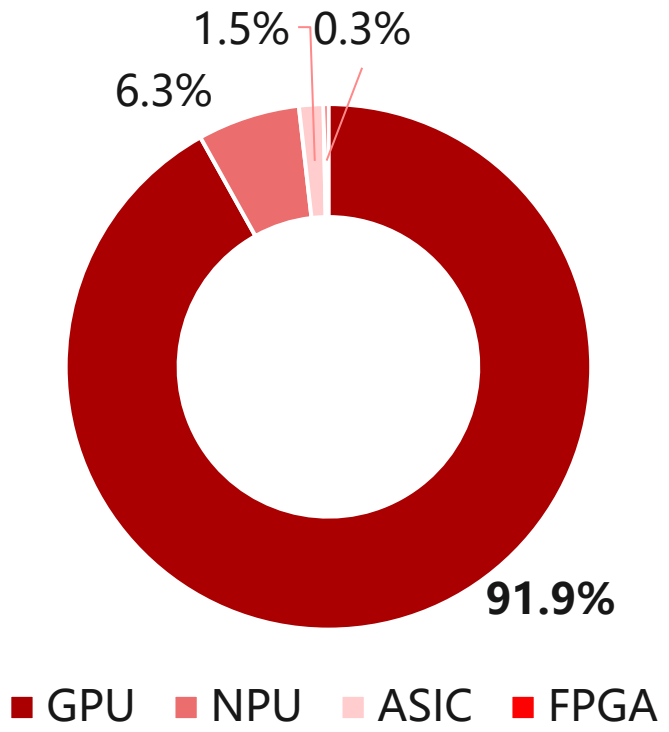


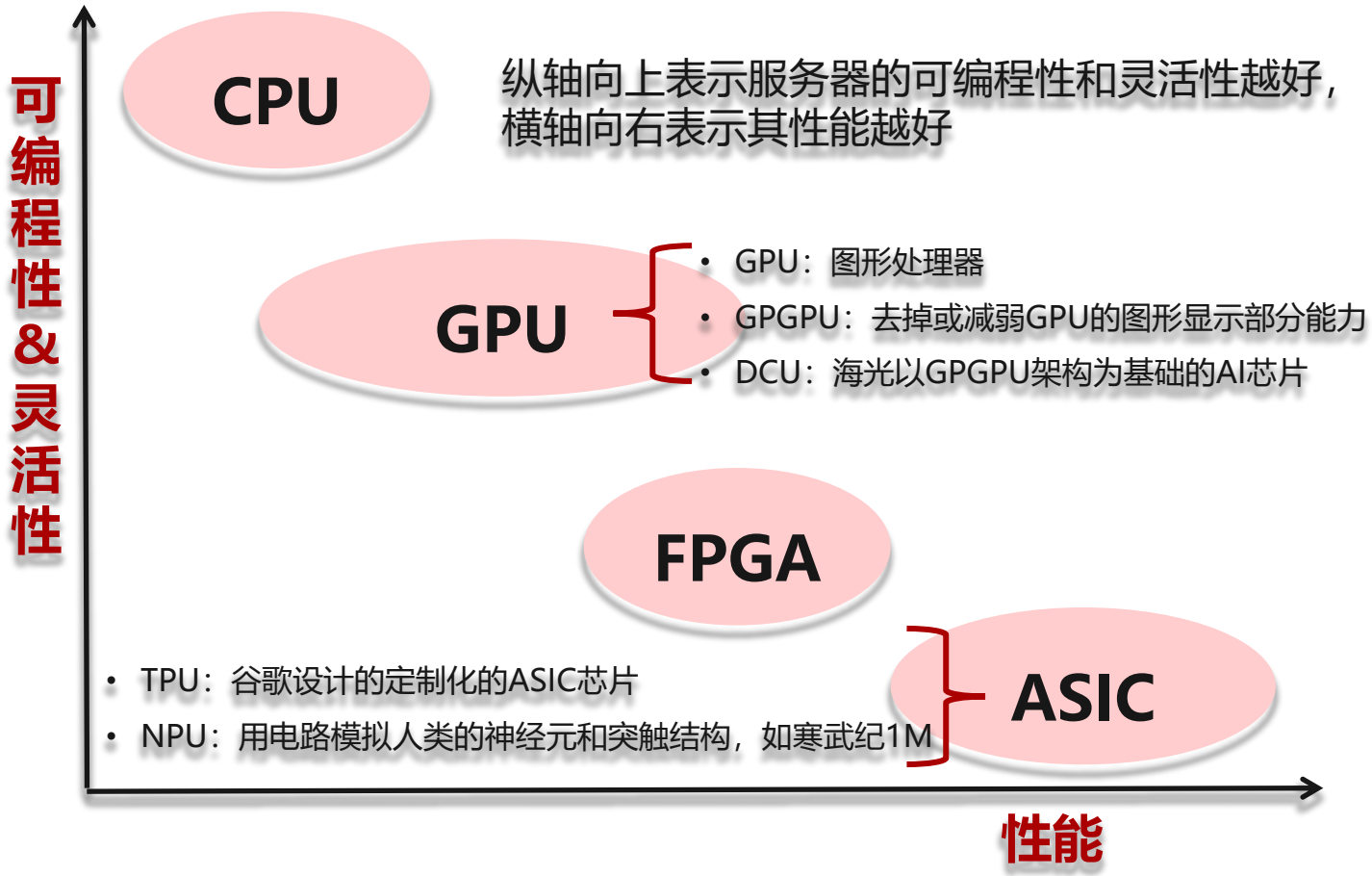
资料来源：浪潮信息官网、IDC、电子发烧友、浙商证券研究所

1.2 分类：按芯片类型可分为GPU、FPGA、ASIC等

- AI服务器采用异构形式，按芯片类型可分为CPU+GPU、CPU+FPGA、CPU+ASIC等组合；
- 目前GPU依然是实现数据中心加速的首选，其他非GPU芯片应用逐渐增多，IDC预计到2025年其他非GPU芯片占比超过20%；
- 一般来说，ASIC的性能最好，但是可编程性和灵活性较弱；在训练或者通用情况下，GPU则是更好的选择。

中国AI服务器按加速卡类型拆分（2021）





资料来源：浪潮信息官网、宽泛科技、CSDN、51CTO、浙商证券研究所

1.3 AI服务器成本主要来自CPU、GPU等芯片，占比50%以上

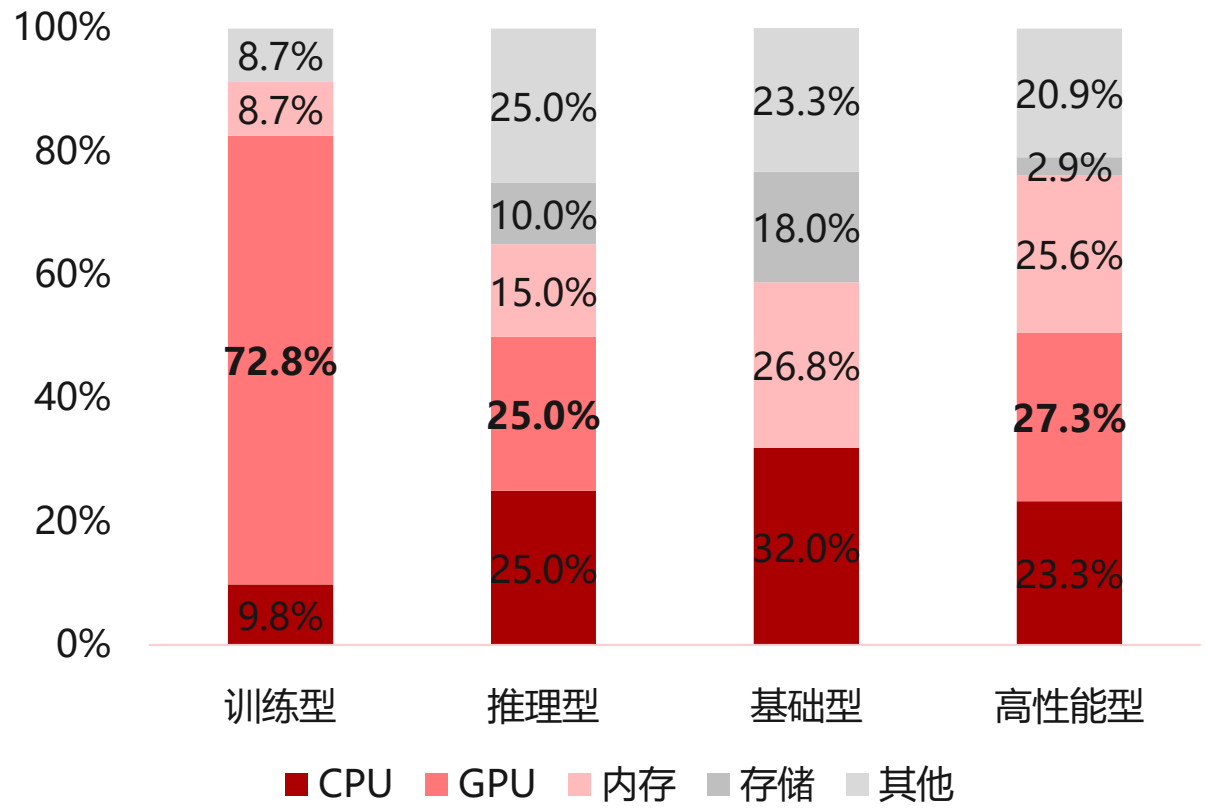
浙商证券股份有限公司 ZHESHANG SECURITIES CO.LTD

- 服务器由电源、CPU、内存、硬盘、风扇、光驱等几部分构成，芯片成本（CPU、GPU等）占比较高，在25%-70%不等；
- 以AI服务器浪潮NF5688M6为例，京东售价约105万人民币，包括2颗Intel Ice Lake处理器（根据cnBeta，约5.3万元/颗）和8颗NVIDIA A800 GPU（根据ZOL，约10.4万元/颗），CPU和GPU的价值量占比分别为10.10%和79.24%；

浪潮NF5688M6技术规格

机型	NF5688M6	
高度	6U	
GPU计算模块	8颗NVIDIA A800 GPU	
处理器	2颗第三代Intel® Xeon® 可扩展处理器(Ice Lake)，TDP 270W，支持3条UPI互联	
芯片组	Intel® C621A Series Chipset (Lewisburg-R)	
内存	支持32条DDR4 RDIMM/LRDIMM内存，速率最高支持3200MT/s	
存储	8块2.5英寸NVMe SSD	16块2.5英寸SATA/SAS SSD
M.2	板载2块 SATA M.2	
PCIe扩展	10个PCIe 4.0 x16插槽，2个PCIe 4.0 x16插槽(PCIe 4.0 x8速率)，1个OCP3.0插槽	6个PCIe 4.0 x16插槽，1个OCP3.0插槽
RAID支持	可选配支持RAID0/1/10/5/50/6/60等，支持Cache超级电容保护提供RAID状态迁移、RAID配置记忆	
网络	可选配1张PCIe 4.0 x16 OCP 3.0网卡，速率支持10G/25G/100G	
前置I/O	1个USB 3.0端口，1个USB 2.0端口，1个VGA端口，1个RJ45管理口	
后置I/O	1个USB 3.0端口，1个VGA端口	
远程管理	内置BMC远程管理模块，支持Redfish/IPMI/SOL/KVM等	
操作系统	Red Hat Enterprise 7.8 64bit、CentOS 7.8、Ubuntu 18.04或更高版本	
散热	N+1冗余热插拔风扇	
电源	6块3000W 80Plus铂金电源，支持3+3冗余	
机箱尺寸	宽447mm，高263.9mm，深850mm	
工作温度	10°C~35°C/50°F~95°F	
满配重量	≤88kg	

2018年全球不同类型服务器成本结构拆分



资料来源：浪潮官网、京东、IDC、产业信息网、智研咨询、ZOL、浙商证券研究所

2、ChatGPT等大模型训练和推理需求激增驱动AI服务器市场高速增长

2.1 国内外厂商布局大模型，千亿级参数量推动算力需求增长

浙商证券股份有限公司 ZHESHANG SECURITIES CO.LTD

- 参数量与算力需求呈正比，据ARK Invest预测，GPT-4参数量最高达15000亿个，则GPT-4算力需求最高可达31271 PFlop/s-day；
- 与此同时，国内外厂商加速布局大模型，其参数量均达到千亿级别，同步带动算力需求爆发式增长；

参数量与算力需求呈正比关系（以GPT为例）

模型名称	参数量	算力需求
GPT 3 Small	1.25 亿个	2.6 PFlop/s-day
GPT 3 175B	1746 亿个	3640 PFlop/s-day
GPT 4	最高 15000 亿个	最高 31271 PFlop/s-day

其他国内外厂商加速布局大模型

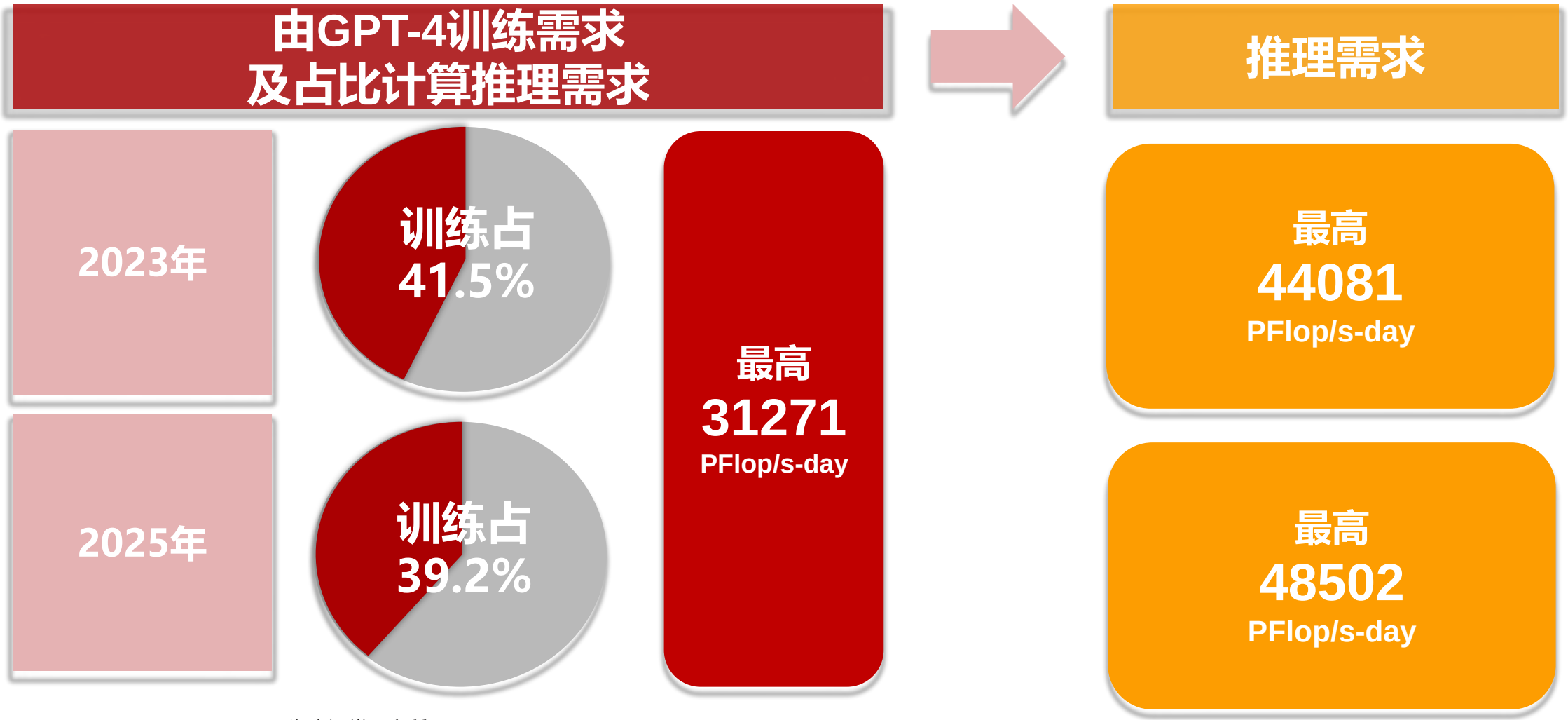
	厂商	模型名称	参数量 亿个	算力需求 PFlops-day
国外厂商	Google	LaMDA	1370	2850
		PaLM-E	5620	11690
	Hugging Face	Bloom	1750	3640
国内厂商	百度	ERNIE 3.0 Titan	2600	5408
	阿里	M6-OFA	100000	208000
	华为云	盘古 NLP	2000	4160
	腾讯	混元 AI	>1000	>2080

资料来源：IDC、ARK Invest、Open AI、澎湃新闻、金融界，《LaMDA: Language Models for Dialog Applications》，Google，机器之心，至顶智库，阿里巴巴达摩院等，浙商证券研究所

2.2 随着训练需求占比下降，推理需求有望进一步提升

浙商证券股份有限公司
ZHESHANG SECURITIES CO.LTD

- 据IDC预测，2023年AI服务器训练需求占比达41.5%，随着大模型的应用，该比例在2025年将降低至39.2%；
- 将GPT-4的推算结果作为训练需求，进一步推算2023/2025年推理需求最高达44081/48502 PFlop/s-day；



资料来源：IDC，Open AI，ARK Invest，浙商证券研究所

2.3 以GPT-4为例，为满足算力需要近千台浪潮NF5688M6服务器

浙商证券股份有限公司 ZHESHANG SECURITIES CO.LTD

训练需求
最高 31271
PFlop/s-day

推理需求（2023）
最高 44081
PFlop/s-day

浪潮 NF5688M6
2 CPU + 8 GPU

根据ARK Invest预计GPT-4最高参数量15000亿推算算力需求

算力需求	计算速度	完成时间	采购数量
最高 75352 PFlop/s-day	5 PFlop/s-day	5 天	3015 台
		15 天	1005 台
		20 天	754 台

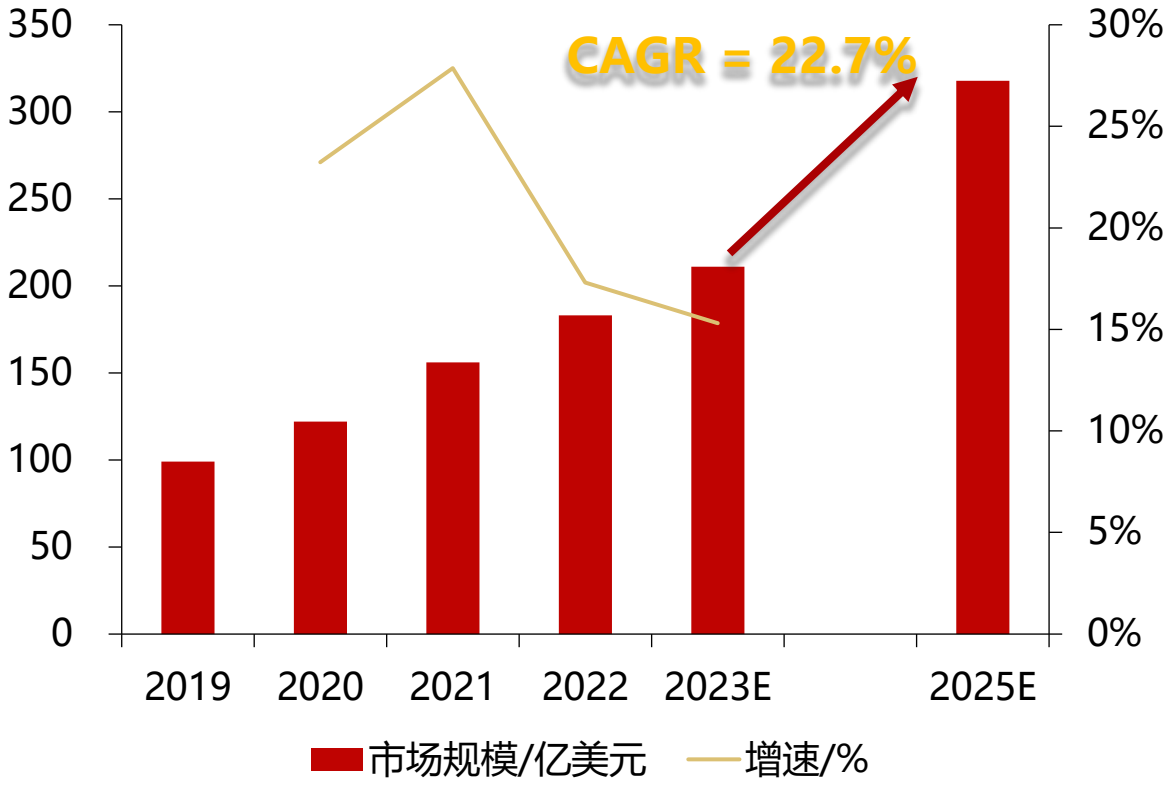
资料来源：IDC，ARK Invest，浪潮信息，浙商证券研究所

2.4 全球AI服务器出货量和市场规模保持快速增长

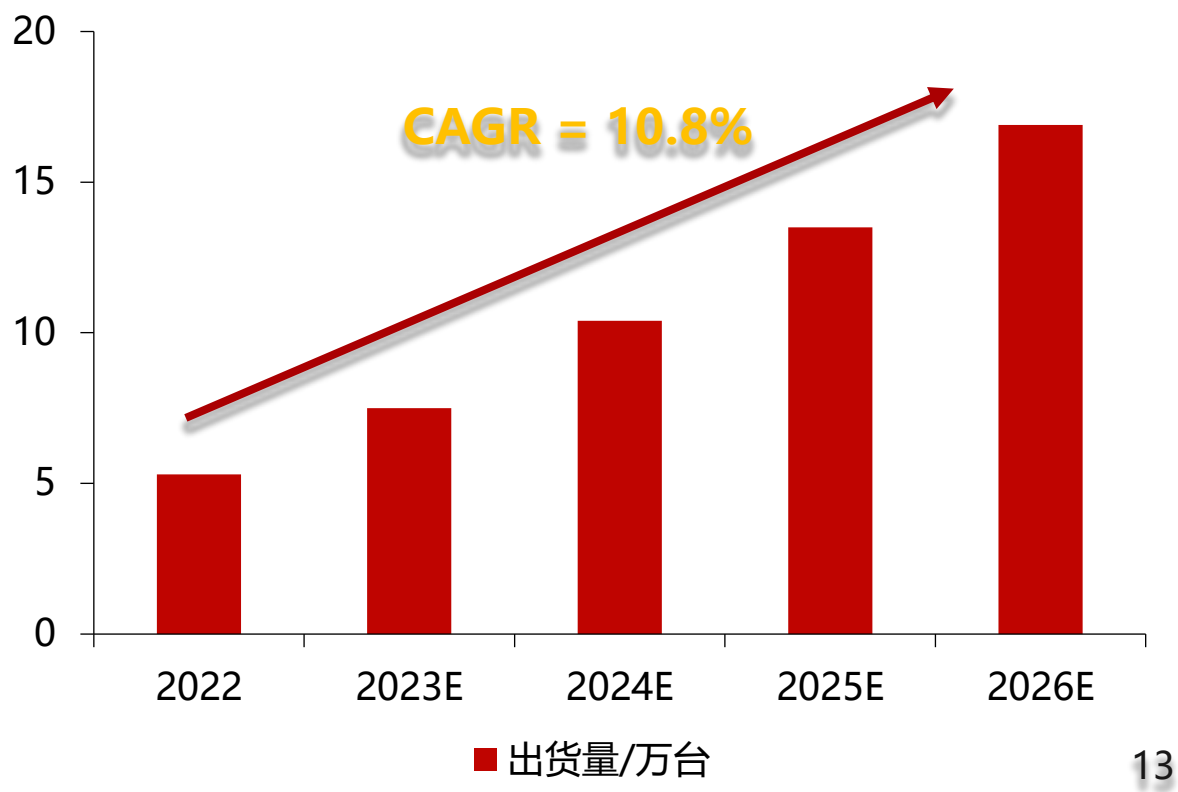
浙商证券股份有限公司 ZHESHANG SECURITIES CO.LTD

- 据IDC预测，2023年全球AI服务器市场规模为211亿美元，预计2025年达317.9亿美元，2023-2025年CAGR为22.7%；
- 据Trend Force预测，2026年全球AI服务器出货量将进一步提升，2022-2026年CAGR达到10.8%；

2019-2025年全球AI服务器市场规模及预测



2022-2026年全球AI服务器出货量

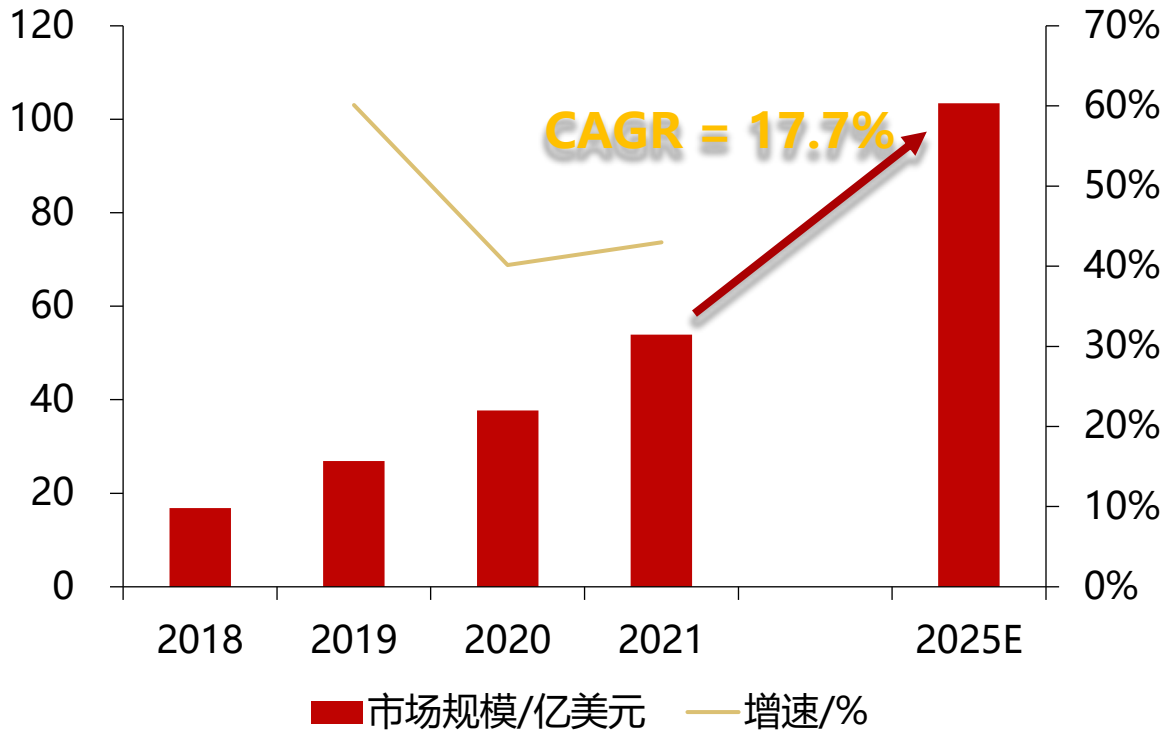


资料来源：IDC，中商产业研究院，Trend Force，浙商证券研究所

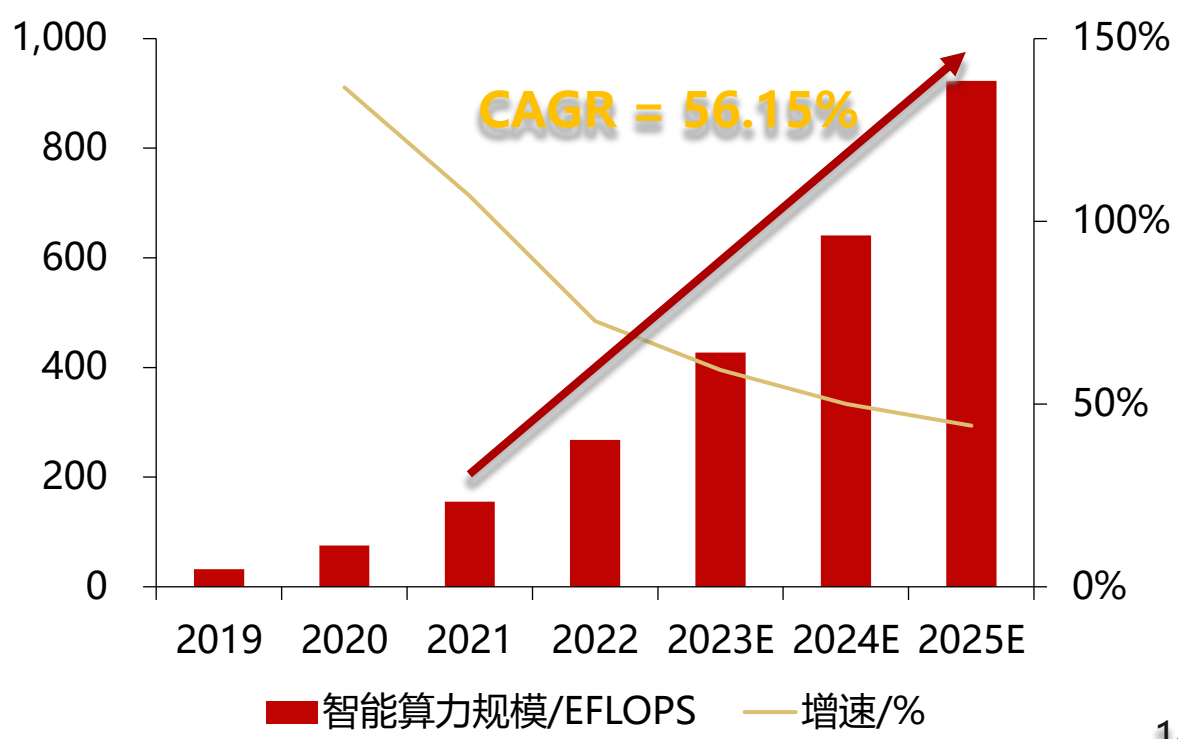
2.5 中国AI服务器出货量和市场规模保持快速增长

- 据IDC统计，2021年中国AI服务器市场规模为53.9亿美元，预计2025年达到103.4亿美元，2021-2025年CAGR达17.7%
  2021年中国智能算力规模为155.2 EFLOPS，预计2025年达922.8 EFLOPS，2021-2025年CAGR达56.15%

2019-2025年我国AI服务器市场规模及预测



中国智能算力规模及预测



资料来源：IDC，产经网，浙商证券研究所

3、国产推理芯片逐步接近一流水平，AI服务器厂商有望各自获益

3.1 美国限制向中国出口先进GPU，可购买削弱带宽的A800

浙商证券股份有限公司 ZHESHANG SECURITIES CO.LTD

- 英伟达推出的三代GPU芯片V100、A100和H100可用于AI模型训练和推理，最新一代的H100较A100计算速度快约3倍（67/19.5）；
- 2022年8月，美国要求英伟达停止向中国企业出售A100和H100两款GPU计算芯片，目前中国企业仅能购买特供的A800芯片，该芯片较A100在互联带宽方面被削弱1/3，成为当前可行的替代方案。暂时未被禁售的V100工艺为12nm，难以满足目前计算需求。

	V100			A100		H100		A800		
	NVLink V100	PCIe V100	PCIe V100S	80GB PCIe	80GB SXM	SXM	PCIe	**40GB PCIe**	**80GB PCIe**	**80GB SXM**
FP64（双精度）	7.8 TFLOPS	7 TFLOPS	8.2 TFLOPS	9.7 TFLOPS		34 TFLOPS	26 TFLOPS	**9.7 TFLOPS**		
FP64 Tensor Core				19.5 TFLOPS		67 TFLOPS	51 TFLOPS	**19.5 TFLOPS**		
FP32（单精度）	**15.7 TFLOPS**	**14 TFLOPS**	**16.4 TFLOPS**	**19.5 TFLOPS**		**67 TFLOPS**	**51 TFLOPS**	**19.5 TFLOPS**		
Tensor Float 32 (TF32)				156 TFLOPS/312 TFLOPS		989 TFLOPS	756TFLOPS	**156 TFLOPS/312 TFLOPS**		
BFLOAT16 Tensor Core				312 TFLOPS/624 TFLOPS		1979 TFLOPS	1513 TFLOPS	**312 TFLOPS/624 TFLOPS**		
FP16 Tensor Core				312 TFLOPS/624 TFLOPS		1979 TFLOPS	1513 TFLOPS	**312 TFLOPS/624 TFLOPS**		
INT8 Tensor Core				624 TOPS/1248 TOPS		3958 TOPS	3026 TOPS	**624 TOPS/1248 TOPS**		
GPU 显存	32/16GB HBM2	32/16GB HBM2	30GB HBM2	80GB HBM2	80GB HBM2e	80GB	80GB	**40GB HBM2**	**80GB HBM2e**	**80GB HBM2e**
GPU 显存带宽	900 GB/s	900 GB/s	1134 GB/s	1935 GB/s	2039 GB/s	3.35TB/s	2TB/s	**1555 GB/s**	**1935 GB/s**	**2039 GB/s**
最大热设计功耗 (TDP)	300W	250W	250W	300W	400W	700W	300-350W	**250W**	**300W**	**400W**
多实例 GPU				最大为7MIG@5GB	最大为 7MIG@10GB	最多 7 MIG @ 10GB		**最大为7MIG@ 5GB**	**最大为 7 MIG @ 10 GB**	
互连	NVLINK 300GB/s	PCIe 32GB/s	PCIe 32GB/s	NVLink:600 GB/s PCIe 4.0:64 GB/s	NVLink:600 GB/s PCIe 4.0:64 GB/s	NVLink:900GB/s PCIe 5.0:128 GB/s	NVLink:600GB/s PCIe 5.0:128GB/s	**NVLink:400 GB/s PCIe 4.0:64 GB/s**	**NVLink:400 GB/s PCIe 4.0:64 GB/s**	**NVLink:400 GB/s PCIe 4.0:64 GB/s**
价格	**约1万美元**			**约1.3-2.7万美元**		**约3.6万美元**		**约1.4万美元**		

资料来源：英伟达官网、videocardz、Zaker、36氪、tomshardware、viperatech、microway、浙商证券研究所

3.1 国产GPU单卡指标接近英伟达，推理应用更具竞争力

浙商证券股份有限公司
ZHESHANG SECURITIES CO.LTD

- 国产算力GPU的主要厂商包括海光信息、寒武纪、平头哥、华为昇腾、天数智芯、燧原科技、摩尔线程、壁仞科技、沐曦等公司，部分产品的单卡指标和参数已经与英伟达产品接近或持平。
- 目前国产算力GPU芯片在推理场景应用较多且具备一定竞争力，如含光800、思元370、MTT S3000等等。

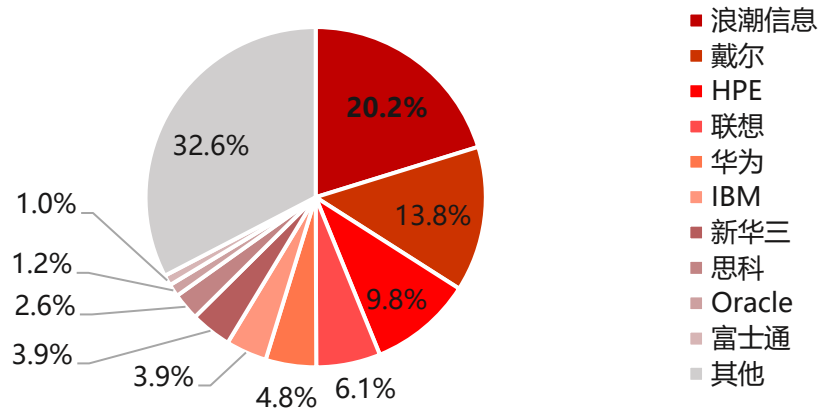
		寒武纪			平头哥	华为昇腾		天数智芯	燧原科技	摩尔线程	壁仞科技		海光信息
		思元370	思元290	思元270	含光800	昇腾310	昇腾910	天垓100	云燧T20/T21	MTT S3000	壁砺100P	壁砺104P	DCU
算力指标	FP64												10.8TFLOPS
	FP32	**24TFLOPS**						**37/18.5TFLOPS**	**32TFLOPS**	**15.2TFLOPS**	**240TFLOPS**		
	TF32								128TFLOPS		480TFLOPS	256TFLOPS	
	FP16	96TFLOPS					320TFLOPS	147/37TFLOPS	128TFLOPS				
	BF16	96TFLOPS							128TFLOPS		960TFLOPS	512TFLOPS	
	INT16	128TOPS	256 TOPS	64TOPS	205TOPS	8TOPS	640TOPS						
	INT8	256TOPS	512 TOPS	128 TOPS	825TOPS	16TOPS		295TOPS	256TOPS		1920TOPS	1024TOPS	
内存容量		24GB LPDDR5	32GB HBM2	16GB DDR4				32GB HBM2	32GB HBM2E	32GB GDDR6	64GB HBM2E	32GB HBM2E	32GB HBM2
内存带宽		307.2 GB/s	1228 GB/s	102 GB/s					1.6TB/s	448GB/s	1.64TB/s	819GB/s	1TB/s
功耗		150W	350W	70w	276W	8W	310W	250W	300W	250W	450-550W	300W	260-350W

资料来源：各公司官网、海光信息招股书、浙商证券研究所

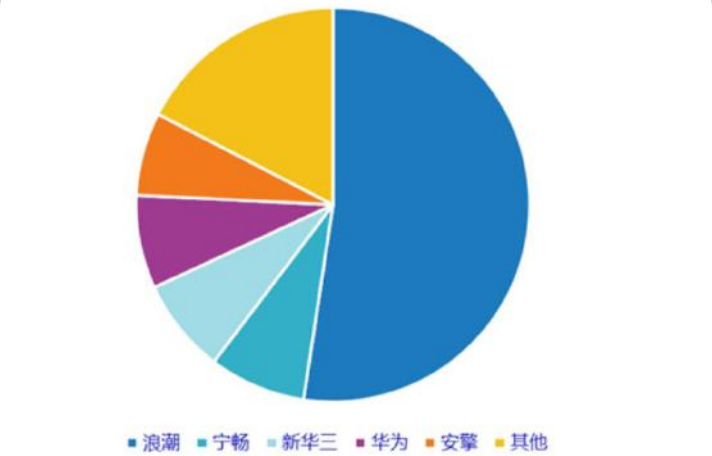
3.2 国产厂商全球市场份额占比超35%，浪潮信息位列国内外榜首

浙商证券股份有限公司 ZHESHANG SECURITIES CO.LTD

2021H1全球AI服务器市场竞争格局



2021年中国加速计算服务器厂商市场份额



AI 服务器产业链相关上市公司

AI芯片

海光信息	龙芯中科	中国长城	CPU
		景嘉微	GPU
安路科技	复旦微电	紫光国微	FPGA
寒武纪	澜起科技		ASIC

服务器集成

浪潮信息	采购国外芯片
中科曙光	海光、寒武纪芯片
拓维信息	华为昇腾芯片
神州数码	华为昇腾芯片
工业富联	供货微软
紫光股份	新华三

资料来源：IDC，浙商证券研究所

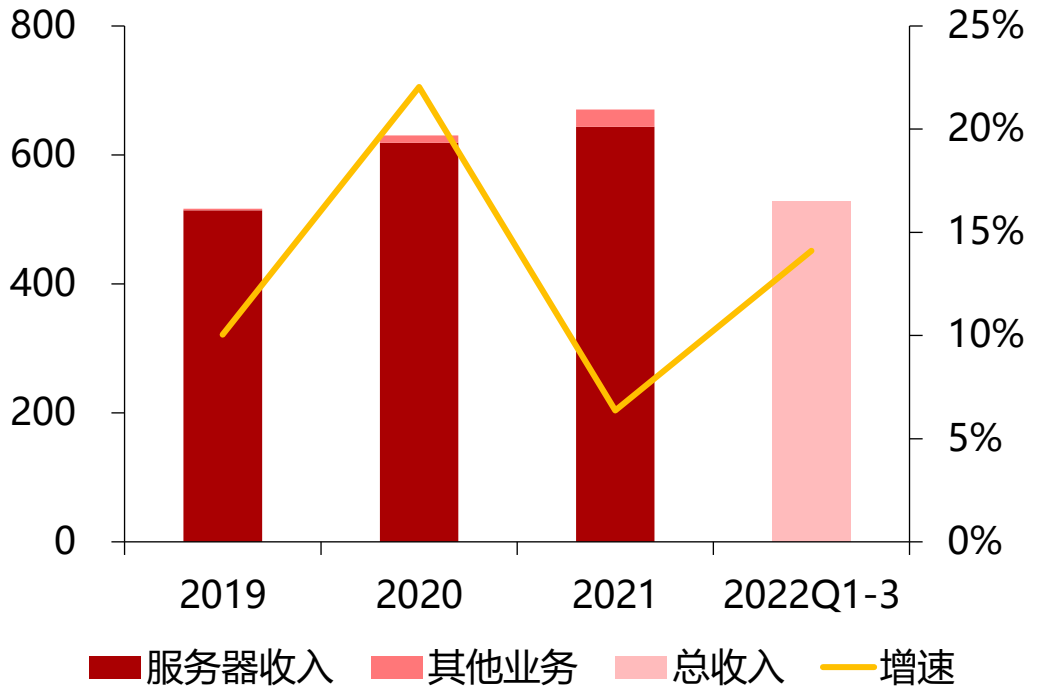
3.3 浪潮信息：市场份额多年第一，AI服务器产品优势显著

浙商证券股份有限公司 ZHESHANG SECURITIES CO.LTD

- 浪潮信息是全球领先的服务器生产商，2021年全球AI服务器市场份额第一；拥有全面的 AI 计算产品阵列以及性能领先的Transformer训练服务器，具备从芯片、板卡、整机到平台软件的全栈AI计算方案提供能力；2022年MLPerf基准评测中浪潮AI服务器获超半数赛道的冠军；
- 浪潮NF5688M6是目前公司算力最强的AI服务器之一，主要用于超大规模数据中心；

型号	高度	处理器	GPU
NF5688M6	6U	2颗Intel Ice Lake处理器	8颗英伟达A800
NF5488A5	4U	2颗AMD EPYC	8颗英伟达A800
NF5468A5	4U	2颗AMD EPYC	支持8个A800、A40等
NF5448A6	4U	2颗AMD EPYC 7003	4颗英伟达A800
NF5468M6	4U	2颗Intel Ice Lake处理器	支持8颗A800、A30等
NF5280M5	2U	2颗第二代Intel Xeon Scalable系列处理器	支持4颗V100、P100等

浪潮信息营业收入增长及拆分（亿元、%）



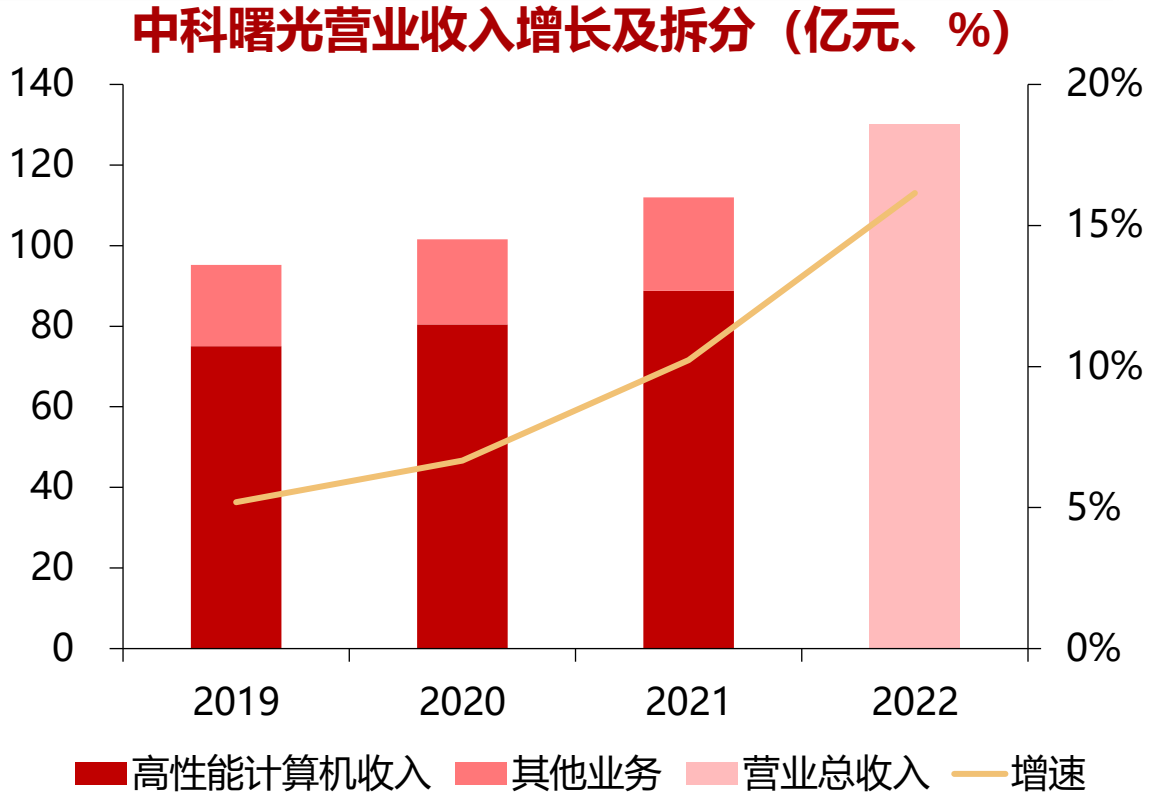
资料来源：浪潮信息官网、凤凰网、Wind、美通社、浙商证券研究所

3.3 中科曙光：国内高性能计算领域领军企业，布局超算和智算中心建设

浙商证券股份有限公司 ZHESHANG SECURITIES CO.LTD

- 中科曙光是国内高性能计算领域的领军企业，亚洲第一大高性能计算机厂商，大力发展云计算、大数据、人工智能、边缘计算等先进计算业务，构建了完整的AI计算服务体系，参建国内多家超算中心和智算中心；
- 中科曙光目前AI服务器分为训练和推理两种，主要采用海光/寒武纪芯片，目前已和百度“文心一言”展开合作，为其产业化应用提供算力支持；

型号	X785-G30	X785-G40
类型	深度学习训练	智能应用推理
处理器	支持英特尔至强可扩展处理器，高速UPI互连总线，大容量三级缓存	2颗第三代智能英特尔至强可扩展处理器
内存	16个插槽，支持DDR4 RDIMM/LRDIMM ECC内存	32个DDR4，支持内存ECC
GPU扩展	支持8块FHFL双宽GPU和16块HHHL单宽GPU	支持英伟达A100、A40、A10、T4



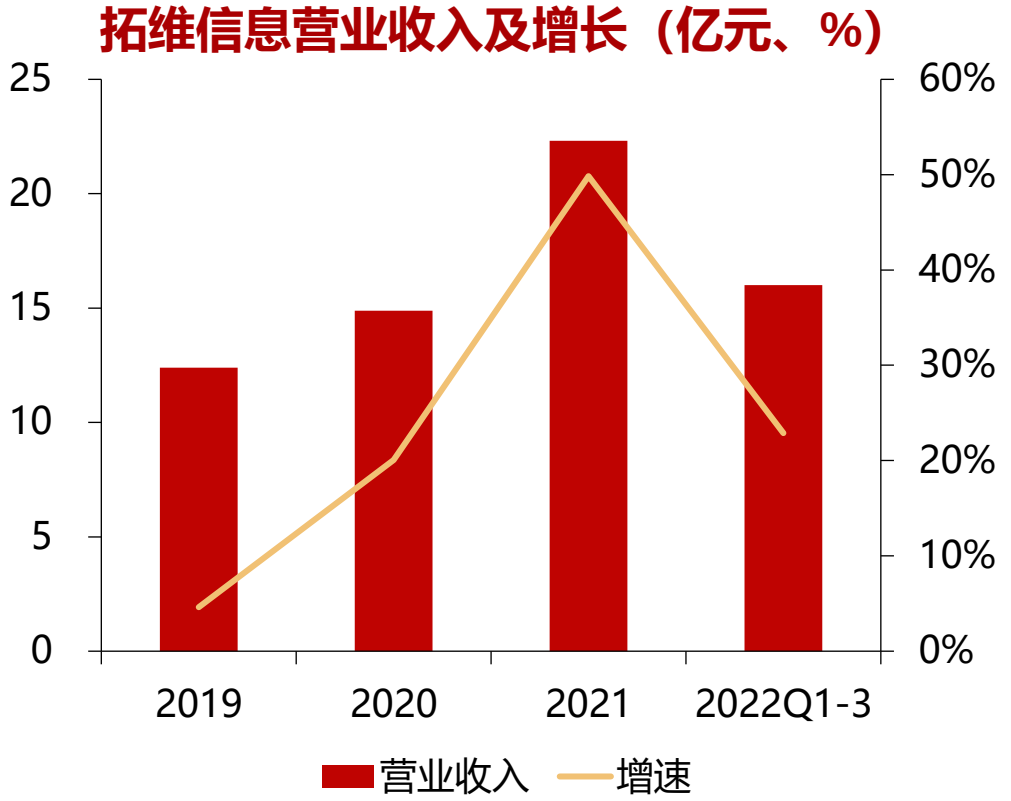
资料来源：中科曙光官网、财联社、Wind、浙商证券研究所

3.3 拓维信息：与华为昇腾AI深度合作，支持云边端全栈全场景应用

浙商证券股份有限公司 ZHESHANG SECURITIES CO.LTD

- 拓维信息是华为昇腾AI战略合作伙伴、华为全方位战略合作伙伴，在“鸿蒙+鲲鹏+昇腾”基础软硬件领域和华为构建全面合作；其中兆瀚系列AI服务器是基于华为达芬奇架构3D Cube技术的昇腾AI处理器；
- 兆瀚AI服务器涵盖训练、推理等领域，目前广泛应用于智慧城市、运营商等行业的数据中心、算力中心相关场景；

型号	兆瀚 RA2300-A	兆瀚SA300	兆瀚 RA5900-A	兆瀚 RA2302-B
形态	2U推理服务器	2U智能边缘服务器	4U训练服务器	2U AI 服务器
处理器	2*鲲鹏920处理器	1*鲲鹏920	4*鲲鹏920	2*64核青松处理器
CPU内存	32个DDR4内存插槽	4个DDR4内存插槽	32个DDR4内存插槽	32个DDR4内存插槽
AI加速卡/处理器	Atlas 300I Pro推理卡/Atlas 300V Pro视频解析卡	Atlas 300I Pro推理卡/Atlas 300V Pro 视频解析卡	8*昇腾910	4个Atlas 300I/V Pro



资料来源：湘江鲲鹏官网、拓维信息官网、公司公告、九方智投、wind、浙商证券研究所

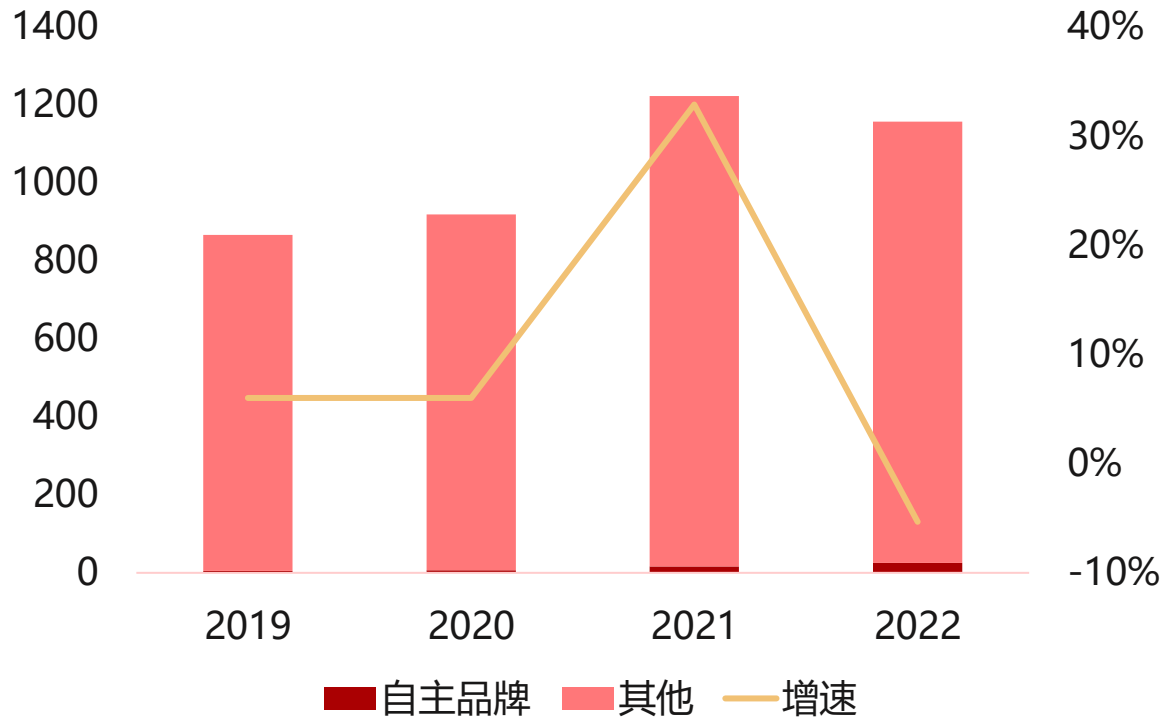
3.3 神州数码：云计算产业国内领先，华为鲲鹏生态重要实践者

浙商证券股份有限公司
ZHESHANG SECURITIES CO.LTD

- 神州数码拥有国内领先的云计算产业相关服务技术优势和国内最大的To B销售网络渠道，云管理服务已经覆盖全球五大公有云，形成了全面的云和数字化生态整合能力；并率先提出“数云融合”理念，并将“数云融合”作为企业数字化转型的方法框架；
- 神州鲲泰系列AI服务器覆盖训练、推理和边缘三种类型，均采用鲲鹏920处理器，成为鲲鹏产业生态的重要实践者，积极布局基于“鲲鹏+昇腾”的自有品牌体系；

型号	KunTai A222	KunTai A722	KunTai A924
类型	2U单路边缘型	2U双路推理型	4U四路训练型
处理器	1*鲲鹏920处理器	2*鲲鹏920处理器	4*鲲鹏920处理器
AI加速卡	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	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	8*昇腾910
内存	4个DDR4 RDIMM	16个或32个DDR4 RDIMM	32个DDR4内存插槽
AI算力	420 TOPS INT8	1120 TOPS INT8	512Tops Int8或256Tops FP16

神州数码营业收入增长及拆分（亿元、%）



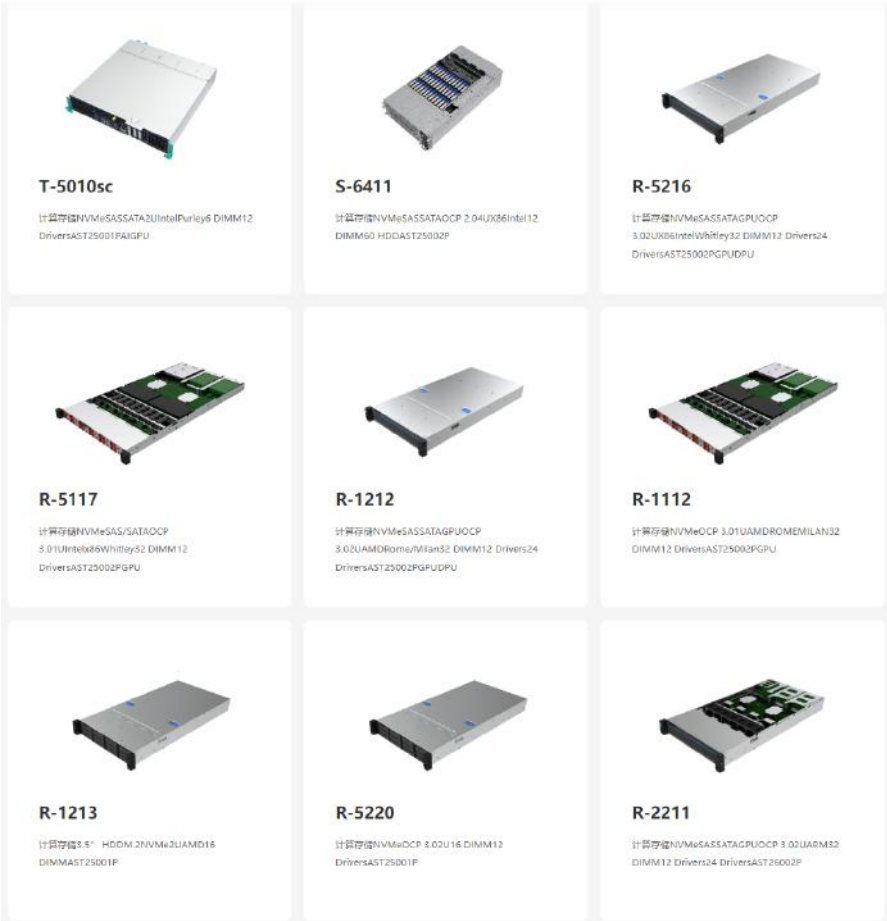
资料来源：神州数码官网、神州鲲泰官网、公司公告、Wind、浙商证券研究所

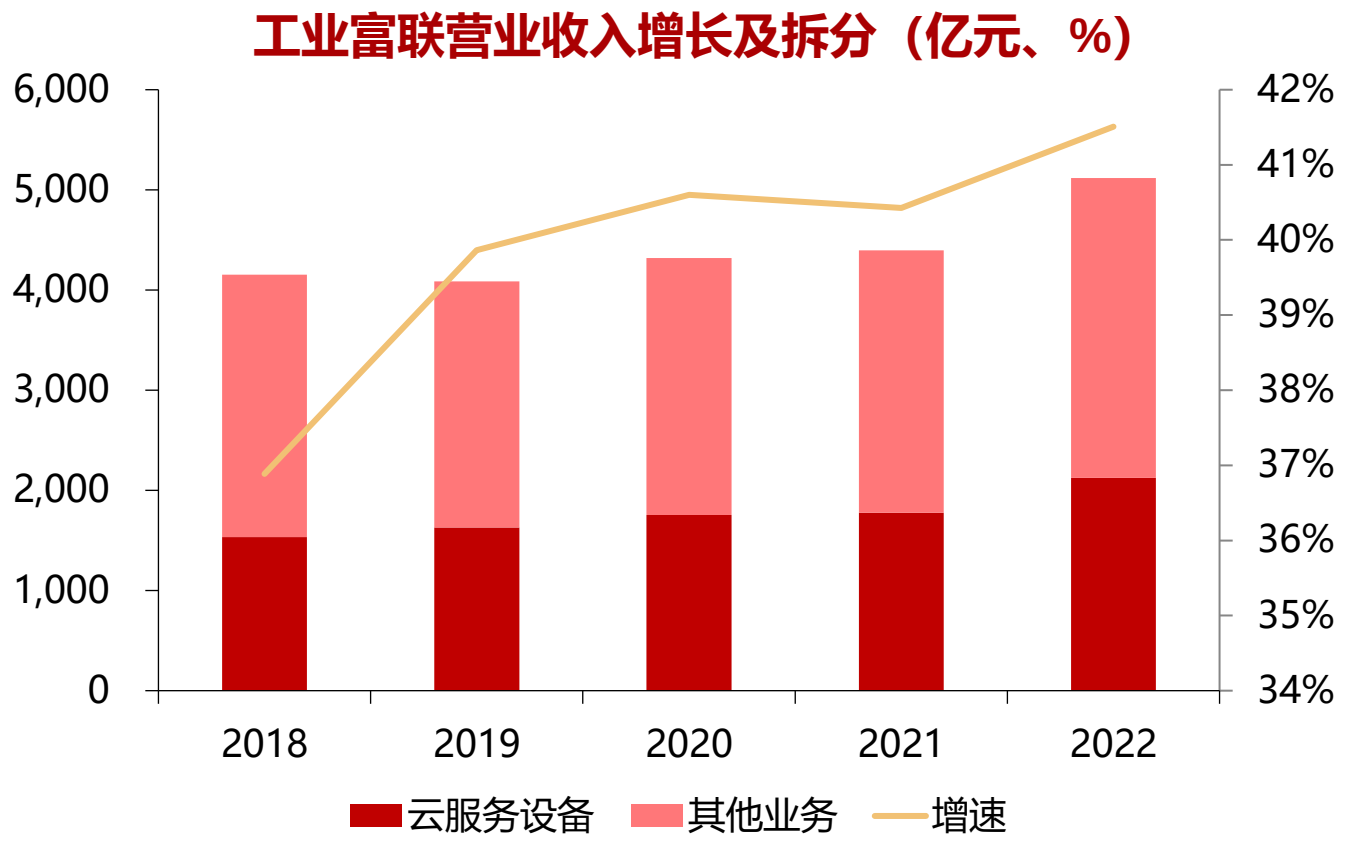
3.3 工业富联：云计算服务器出货量全球第一，有望受益于海外AI服务器需求

浙商证券股份有限公司 ZHESHANG SECURITIES CO.LTD

- 工业富联在云计算服务器出货量持续全球第一，并推出新一代云计算基础设施解决方案（模块化服务器、高效运算HPC）以解决AIGC算力井喷需求；其客户布及全球，覆盖微软、谷歌、英伟达、英特尔等海外头部大厂；
- 工业富联云计算产品涉及云服务器、高性能服务器、AI服务器、边缘服务器及云储存设备等；

工业富联主要高性能服务器产品





资料来源：工业富联官网、金融界、Wind、浙商证券研究所

3.3 紫光股份：国内ICT行业龙头，推进收购新华三加码智慧计算

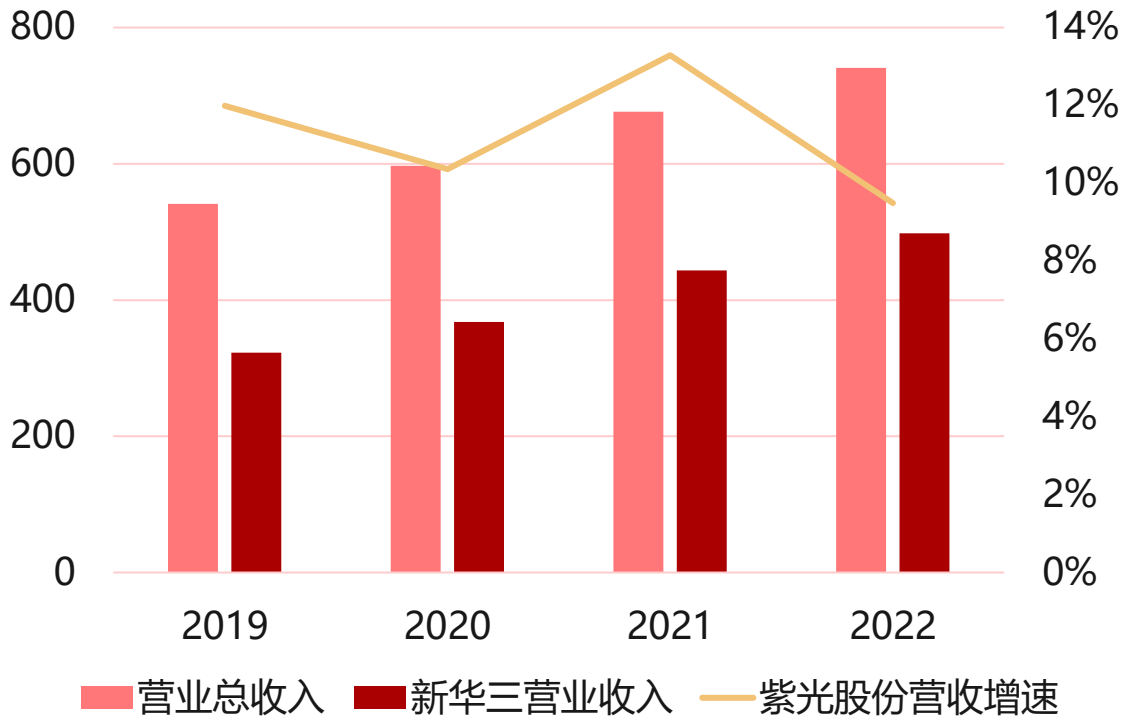
浙商证券股份有限公司
ZHESHANG SECURITIES CO.LTD

- 紫光股份致力于打造“云—网”产业链，向云计算、移动互联网和大数据处理等信息技术的行业应用领域全面深入，目前核心业务基本覆盖IT服务的重要领域；紫光股份对子公司新华三剩余49%的股权收购有望能在今年内完成，正持续推进；
- 新华三是目前国内前三、全球前十的AI服务器厂商，其相关产品包括UniServer AI一体机和一系列AI服务器等，新华三AI服务器在MLPerf训练及推理测试中共斩获86项世界第一，具备深厚技术积淀和领先实力。

型号	GPU支持	处理器	内存
UniServer R5350 G6	10张双宽GPU	2颗AMD EPYC 9004	24个DDR5
UniServer R5300 G5	NVIDIA HGX A800 4-GPU,8个双宽GPU,20个单宽GPU	2颗英特尔至强第三代可扩展家族处理器或澜起津逮处理器	32个DDR4,16个英特尔傲腾持久内存PMem 200
UniServer R5500 G5	NVIDIA HGX A800 8-GPU	2颗AMDEPYC Rome/Milan/Milan-X,2颗英特尔至强第三代可扩展家族处理器或澜起津逮处理器	32个DDR4,16个英特尔傲腾持久内存PMem 200
UniServer R5300 G3	8个双宽GPU,20个单宽GPU	2颗英特尔至强可扩展家族处理器或澜起津逮处理器	24个DDR4,12个英特尔傲腾数据中心级持久内存

资料来源：新华三官网、紫光股份官网、紫光股份公司公告、环球网、Wind、浙商证券研究所

紫光股份和新华三营业收入情况（亿元、%）



风险提示

1、国外制裁范围扩大的风险：美国对中国制裁范围可能继续扩大，导致国内服务器厂商无法购买到A800等芯片而经营受到影响；

2、国内厂商发展不及预期的风险：国内算力芯片厂商的研发和推进效果不及预期所带来的风险；

3、下游市场需求不及预期的风险：目前大模型发展仍不成熟且存在一定缺陷，发展可能出现瓶颈，导致下游市场需求不及预期；

4、版权、伦理和监管风险等： AIGC 生成的内容依赖现有版权素材，另外不当使用或模型自身问题可能导致不良后果；

行业评级与免责声明

行业的投资评级

以报告日后的6个月内，行业指数相对于沪深300指数的涨跌幅为标准，定义如下：

1、看好 ：行业指数相对于沪深300指数表现＋10%以上；

2、中性 ：行业指数相对于沪深300指数表现－10%～＋10%以上；

3、看淡 ：行业指数相对于沪深300指数表现－10%以下。

我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重。

建议：投资者买入或者卖出证券的决定取决于个人的实际情况，比如当前的持仓结构以及其他需要考虑的因素。投资者不应仅仅依靠投资评级来推断结论

行业评级与免责声明

法律声明及风险提示

本报告由浙商证券股份有限公司（已具备中国证监会批复的证券投资咨询业务资格，经营许可证编号为：Z39833000）制作。本报告中的信息均来源于我们认为可靠的已公开资料，但浙商证券股份有限公司及其关联机构（以下统称“本公司”）对这些信息的真实性、准确性及完整性不作任何保证，也不保证所包含的信息和建议不发生任何变更。本公司没有将变更的信息和建议向报告所有接收者进行更新的义务。

本报告仅供本公司的客户作参考之用。本公司不会因接收人收到本报告而视其为本公司的当然客户。

本报告仅反映报告作者的出具日的观点和判断，在任何情况下，本报告中的信息或所表述的意见均不构成对任何人的投资建议，投资者应当对本报告中的信息和意见进行独立评估，并应同时考量各自的投资目的、财务状况和特定需求。对依据或者使用本报告所造成的一切后果，本公司及/或其关联人员均不承担任何法律责任。

本公司的交易人员以及其他专业人士可能会依据不同假设和标准、采用不同的分析方法而口头或书面发表与本报告意见及建议不一致的市场评论和/或交易观点。本公司没有将此意见及建议向报告所有接收者进行更新的义务。本公司的资产管理公司、自营部门以及其他投资业务部门可能独立做出与本报告中的意见或建议不一致的投资决策。

联系方式

浙商证券研究所

上海总部地址：杨高南路729号陆家嘴世纪金融广场1号楼25层

北京地址：北京市东城区朝阳门北大街8号富华大厦E座4层

深圳地址：广东省深圳市福田区广电金融中心33层

邮政编码：200127

电话：（8621）80108518

传真：（8621）80106010

浙商证券研究所：http://research.stocke.com.cn