华金证券
Huajin Securities
华 发 集 团 旗 下 企 业

证券研究报告
通信行业/行业专题报告
评级：领先大市（维持）
2023年9月17日

算力底座：算力承载与网络中枢

通信行业专题报告

分析师：李宏涛 S0910523030003

核心观点

- 算力进展：大模型：算力驱动迭代，应用在向通用型场景和垂直行业型场景落地；服务器：作为算力的承载，随着大模型不断涌现，训练需求爆发，AI服务器的价值凸显；交换机：算力网络中枢，在智算组网下对高速率交换机升级迫切，400G交换机将拉动整体市场增长，具备翻倍增长空间。
- 算力测算与算力租赁：据测算，2022年训练用算力为95Eflops，至2025年训练用算力749E，复合增长43%；在算力总需求高增长背景下，智算租赁景气度提升，目前该市场百花齐放，将成为公司第二增长曲线。
- 算力商业机会演进的路线选择：我们认为，算力方向将演变为以国内驱动为主，国内驱动与海外驱动并行态势；国内驱动逻辑：1、基础设施：大模型出现带动AI服务器需求，数据中心不断扩容升级；2、内延：短距离传输、高速背板连接需求大涨；3、算力资源：全国智算中心建设加速，得卡者得生产力；
- 建议关注标的：紫光股份、菲菱科思、浪潮信息、共进股份、中贝通信、润泽科技等。
- 风险提示：模型算法及应用市场拓展不及预期；芯片成本过高、订单周期过长影响落地进度；智算中心建设进度不及预期。

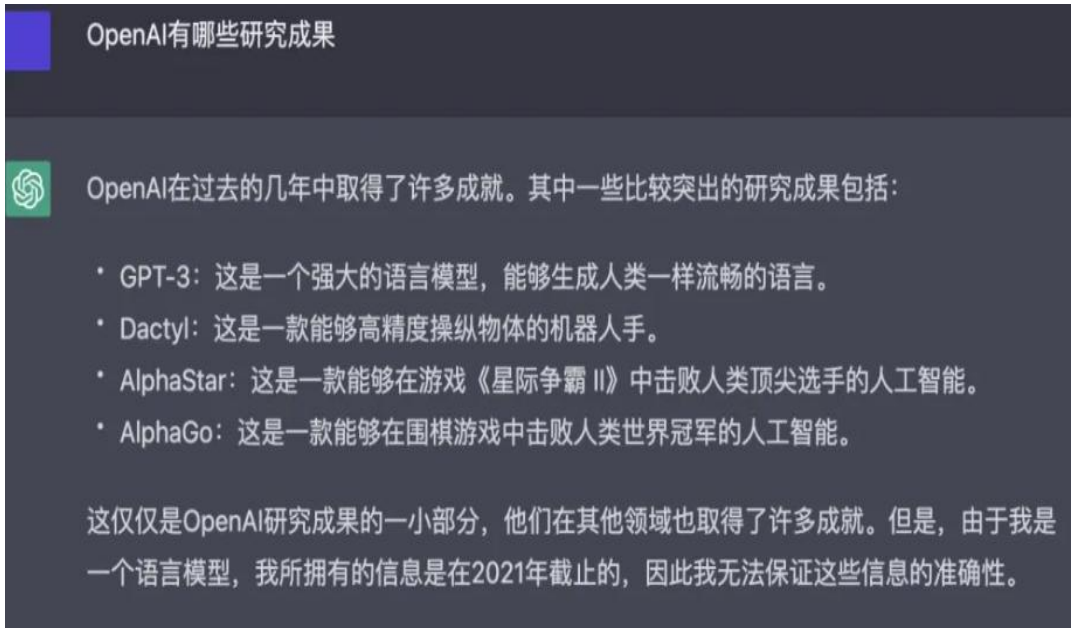
目录

趋势一：大模型发布持续超预期发展

- **ChatGPT模型的推出持续超预期发展。**1）ChatGPT注册用户激增，5天突破百万；2）GPT模型不断迭代：GPT3于20年6月推出，在2022年11月推出的GPT3.5基础上再次更新，23年6月推出GPT 3.5 turbo 16K。3）AI渗透千行百业：基于大模型的多场景应用也不断拓展。拉动智能算力规模高速增长。根据OpenAI的研究，AI训练所需算力指数增长，且增长速度超越硬件的摩尔定律。

图表1 ChatGPT用户比较图

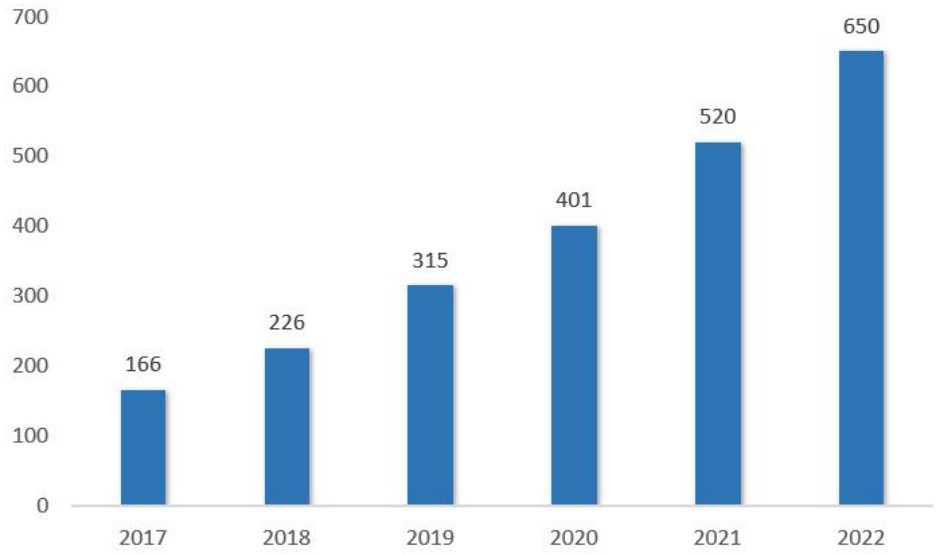
图表2 中国移动交换机招标情况对比



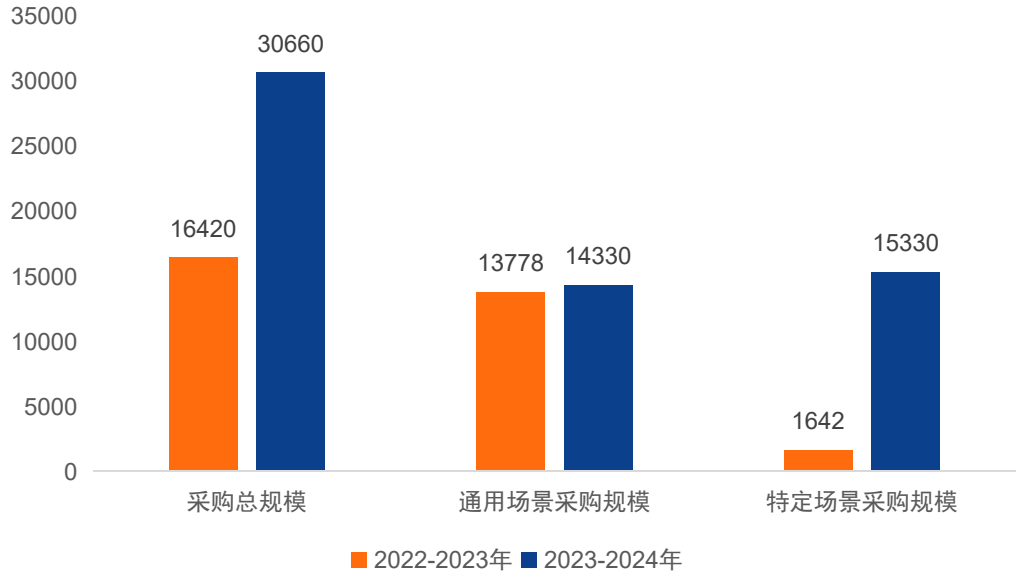
趋势二：数字基础设施建设加速，景气度提升

- **1、我国数据中心总体建设进度加速。**据《数字中国发展报告》，我国数据中心机架总规模2022年达到650万机架，比去年增长130万架，近5年年均增速超过30%。
- **2、算力基础设施采购量明显提升。**以移动为代表的公司，采购交换机数量加大，移动公司2023-2024年采购交换机30660台，其中特定场景交换机需求放量增长，采购15330台。

图表3 2017年—2022年我国在用数据中心机架规模（单位：万架）



图表4 中国移动交换机招标情况对比（单位：台）



趋势三：模型百花齐放，推动AI在各领域广泛应用

- **1、国内模型百花齐放：**众多公司推出大模型，比如阿里推出通义千问、腾讯推出混元，百度的文心一言大模型，科大讯飞的星火大模型
- **2、面向垂直行业的大模型应用效果显著：**紫天科技旗下河马游戏推出《大侦探智斗小AI》，下载量进入榜单TOP10；医联所推出的MedGPT具备全流程智能化诊疗能力，与三甲专家诊断一致性超96%。“天擎”美亚公共安全大模型，具备警务意图识别、警务情报分析、案情推理等推理能力，可实现全流程闭环进化。将重构两大应用场景：电子数据取证和智慧警务系统。

图表5 科大讯飞星火认知大模型

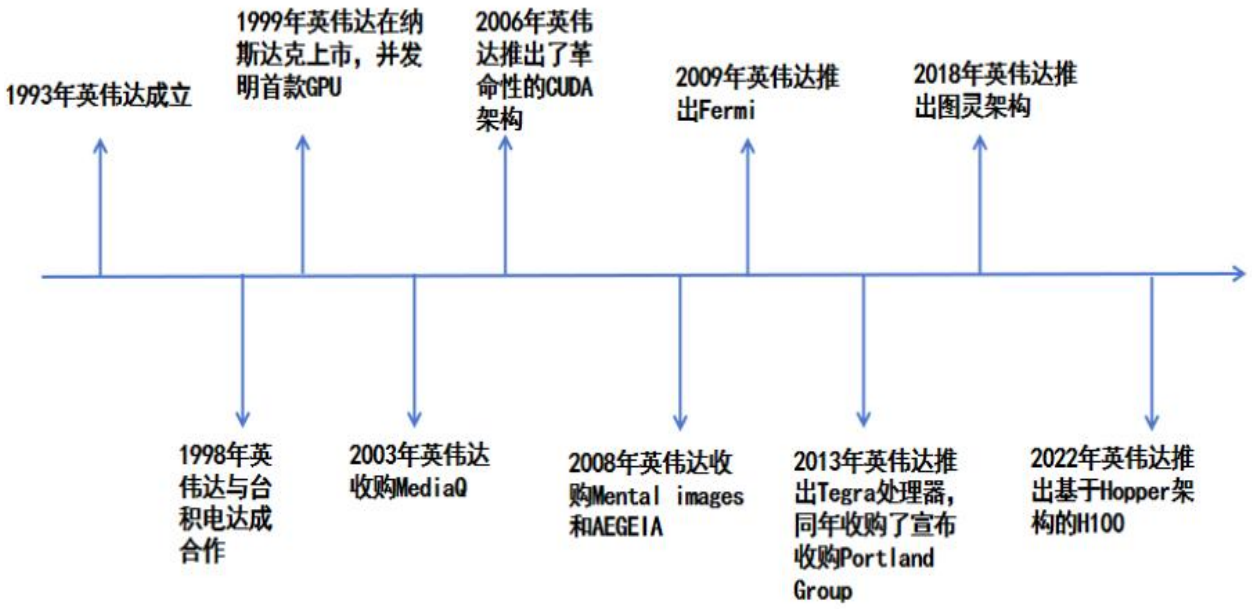
图表6 医联MedGPT

目录

AI芯片是大模型迭代的基础

- **英伟达为代表的芯片厂商性能持续进步。**英伟达2017年上市V100芯片，2020年上市A100芯片，2022年上市H100芯片，H100较V100计算性能提升3倍以上。并且为了迎合AI大模型的潮流，H100配有Transformer引擎，可更好的支撑大模型的架构。
- **AI芯片成为大模型不断更新的算力基础。**以GPT-3为例，模型包含1750亿参数，训练成本达1200万美元。而谷歌发布的PaLM-E包含5620亿参数，GPT-4包含数万亿级别参数。所以大模型的更新迭代必须以先进的AI芯片作为算力基础。

图表7 英伟达发展大事记



图表8 英伟达主流芯片比较

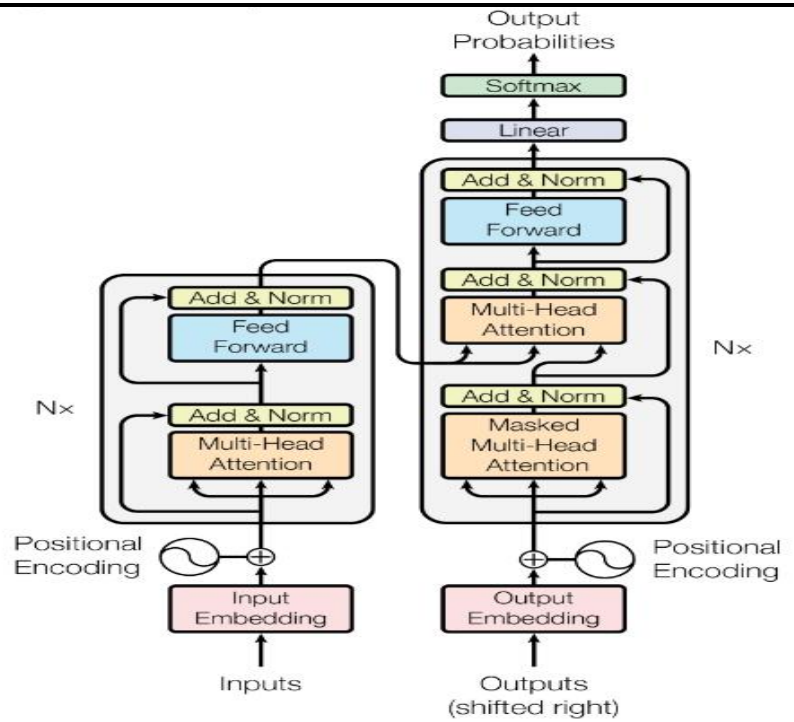
芯片型号	V100	A100	H100
FP64	7TFLOPS	9.7TFLOPS	26TFLOPS
FP32	14TFLOPS	19.5TFLOPS	51TFLOPS
GPU显存	32/16GB	80GB	80GB
GPU显存带宽	900GB/s	1935GB/s	2TB/s
最大设计功耗	250W	300W	300-350W
Interconnect	NVLink:300GB/s Pcle:32GB/s	NVLink:600GB/s Pcle:64GB/s	NVLink:300GB/s Pcle:128GB/s

资料来源：英伟达官网、新浪VR、澎湃新闻、搜狐、知乎、华金证券研究

Transformer架构是AI大模型与传统模型不同的核心

- **Transformer模型是AIGC大模型与传统模型不同的核心。**AIGC大模型起源于NLP，并基于Attention机制构建Transformer模型。传统模型多基于CNN 和 RNN 结构。
- **Transformer模型架构是现代大语言模型所采用的基础架构。**Transformer模型是一种非串行的神经网络架构，最初被用于执行基于上下文的机器翻译任务。Transformer模型以Encoder-Decoder架构为基础，引入“注意机制”（Attention），具有能够并行运算、关注上下文信息、表达能力强等优势。

图表9 Transformer模型架构图



图表10 Transformer、CNN与RNN优缺点对比

RNN架构缺点	①语言的长距离信息会被弱化。 ②串行处理机制所带来的计算效率低。
CNN架构缺点	①生物学基础支持不足，没有记忆功能。 ②全连接模式过于冗余而低效。 ③胜在特征检测，但特征理解不足。
Transformer架构优点	①可以实现完全并行的计算 ②自注意力机制可以更好地捕捉长距离的依赖关系。 ③可计算全局的依赖关系，进而捕捉全局信息。 ④更容易地解释模型的预测结果 ⑤可处理多模态数据，提高模型表现。

底层架构各异导致训练参数不同

- **底层架构的不同导致训练参数的要求不同。**以Transformer为架构的大模型一般可达百亿、千亿、万亿级别，而以CNN或RNN为底层架构的传统模型则是亿级别及更少级别。
- **AI大模型参数级别庞大，需要强大的算力和硬件支撑。**以ChatGPT3.0为例进行拆解，训练一次的成本约为140万美元。对于一些规模更大的模型来说，训练成本介于200万美元-1200万美元之间。

图表11 不同架构模型的参数差别

基于架构	模型名	模型参数量
RNN	LSTM	几百万-几千万
CNN	LeNet5	数万
	AlexNet	6000万
	VGG16	1亿
Transformer	INTERN	100亿
	盘古大模型	1000亿
	MT-NLG	530亿
	Switch Transformer	16000亿

图表12 ChatGPT3.0拆解

模型参数	1750亿
所需算力	3650 PFLOPS/day
所需内存	175GB
所需带宽	896GB/s
所需硬件（A100 GPU）	18953张

资料来源：知乎、CSDN、电子发烧友、数据猿、佐思汽车研究、华金证券研究所

AI大模型带动并行计算，训练消耗更多算力

- **训练和推理是大模型运行的重要环节。**训练环节是大模型的学习过程，可提高模型在各种任务上的性能。推理环节是大模型的判断过程，利用已有训练效果对新的输入进行预测和决策。
- **训练和推理需要大量算力支撑，其中训练消耗更多。**援引Open AI测算，自2012年起，全球头部AI模型训练算力需求每3-4个月翻一番，每年头部训练模型所需算力增长幅度高达10倍。而推理阶段则根据模型上线后的搜索量来计算，新输入数据的量级相较于训练环节的大规模数据量级较低。

图表13 推理和训练阶段所用服务器

阶段	训练阶段	推理阶段
服务器	英伟达A100服务器	华为Atlas 800服务器
芯片数量	8张A100	7张Atlas 300I
提供算力	5PFLOPS	616TFOPS

图表14 英伟达服务器提供算力与内存

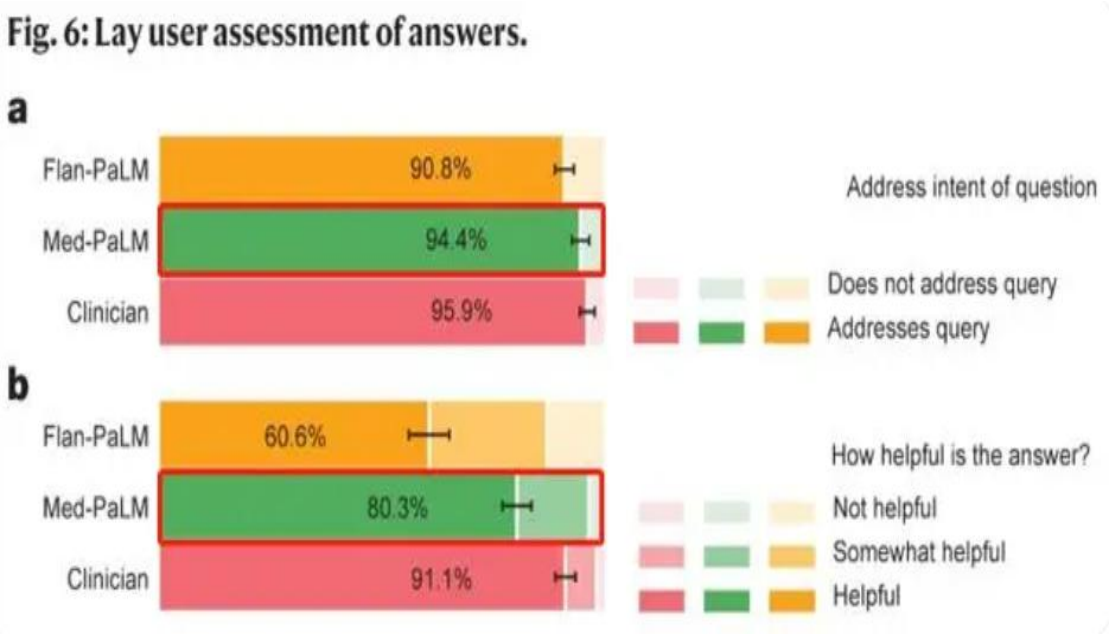
服务器型号	服务器算力	服务器内存
A100	5 PFLOPS	640GB
H100	32 PFLOPS	640GB
GH200	1 exaFLOPS	144TB

资料来源：英伟达官网、广州日报、华金证券研究所

进展：国外大模型—通用模型领先，商业化首落地

- **国外大模型主要以美国高科技公司为主。**Google、Meta、微软等美国公司处于了世界大模型发展的领先地位，现在几乎所有AI大模型训练时采用的Transformer网络结构，Transformer的提出让大模型训练成为可能。
- **微软产品Copilot实现商业化落地。**Copilot产品基于GPT-4，将生成式AI能力全面应用于各大办公套件，可作为办公场景下智能写手。7月18日，Copilot提供订阅收费服务，每名用户每月的价格从12.5美元到57美元不等。产品的商业化落地体现出企业对人工智能的未来前景持续看好。

图表15 谷歌医疗模型效果比较



图表16 Copilot产品

进展：国内大模型—通用和垂直两条路演绎

- **国内AI大模型数量呈现爆发式增长。**国内公开发布的大模型已达80多个。研究大模型的公司有百度、商汤、科大讯飞、华为、阿里、京东、第四范式等公司。同时科研机构也在积极入局，有清华大学、复旦大学、中科院等高校。
- **国内大模型分为通用和垂直应用。**文心一言、通义千问等打造跨行业通用化人工智能能力平台，其应用正从办公、生活、娱乐向医疗、工业、教育等加速渗透。与此同时，一批针对生物制药、遥感、气象等垂直领域的专业类大模型，提供针对特定业务场景的专业化解决方案。

图表17 国内通用领域大模型进展

公司	模型	模型规模	特点
腾讯	混元大模型	千亿级	五大权威数据集榜单登顶，实现跨模态领域大满贯
百度	文心一言	2600亿	首个中国版ChatGPT
京东	K-PLUG	10亿	与电商场景紧密结合，从电商领域中学习特定知识
阿里巴巴	M6	10万亿	当时最大规模的中文多模态预训练大模型
华为云	盘古NLP	千亿级	业界首个千亿参数的中文预训练大模型
中科院	紫东太初三模态	千亿级	全球首个三模态大模型

图表18 国内垂直领域大模型进展

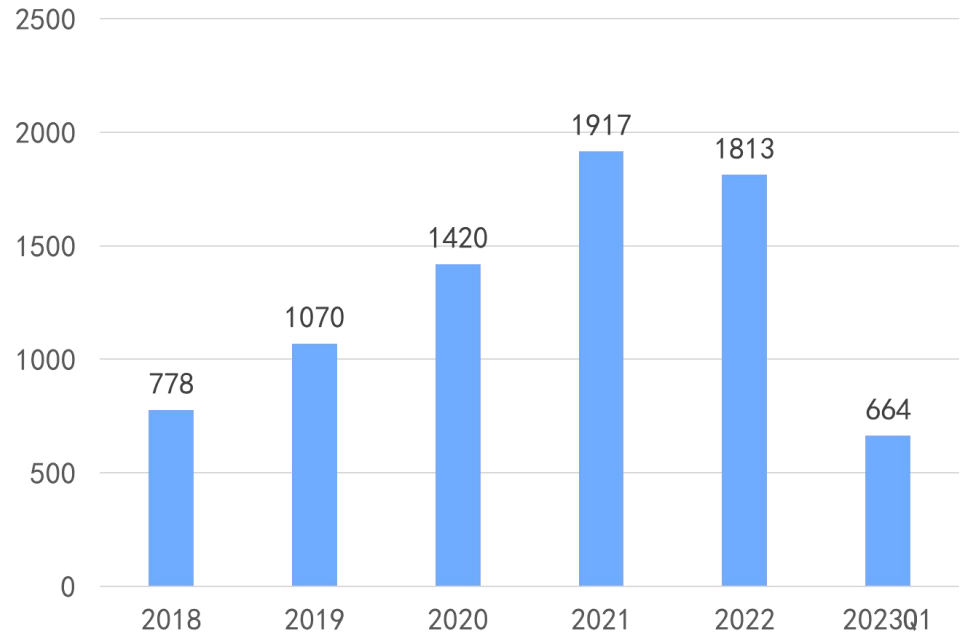
公司	模型	垂直领域	模型介绍
声智科技	壹元模型 AzeroGPT	医联网	融合声学和人工智能技术，面向医联网的多技能与多模态的对话式基础开发框架，提供运行健康检测技能集，疫苗技能集，流调技能集，影像技能集等功能
达观数据	“曹植”大模型	长文本报告	支持多种语言长文本的自动化写作和多语种翻译等功能，能够针对不同行业、领域的文案需求进行深度优化和个性化定制。
蜜度信息	蜜度蜜巢 蜜度文修	政务媒体	通过对公文、新闻等内容的深度学习，创作内容更符合中国文化价值观和文化元素。对中文常见易混淆词语进行增强学习训练，能够辨析词语的细微语义差异。
星环科技	星环无涯	金融量化	基于上百万高质量金融语料、上百种事件类型、20余万事件实例的二次预训练，星环无涯可实现基本面、技术面、消息面在内的金融通识领域的准确理解能力。
金山办公	WPS AI	办公文档	涵盖了包括文字内容生成、PPT生成及美化、文档阅读理解、表格操作等方面的功能，同时可进行文档内容问答，总结摘要。
澜舟科技	孟子预训练模型	轻量化	以十亿参数刷新了此前百亿、千亿级别参数模型轮番霸榜的中文语言理解权威评测基准 CLUE 榜单。可促进自然语言处理技术在更广泛实际场景中的应用

资料来源：车百智库、搜狐、每日经济新闻、懂财帝、新浪财经、阿里云、华金证券研究所

资本支出：支出加速，运营商有望成为国内主力军

- **全球云服务CAPEX持续增长。**受益于大模型的持续发展，2023Q1全球云基础设施服务支出增长19%，前三大云厂商AWS、Azure和谷歌云共同增长22%。云服务成为IT市场中增长最快的部分之一。
- **三大运营商成为国内网络建设主力军。**三大运营商中国移动、中国电信、中国联通不断 加大算力投入，优化算力网络布局。根据三大运营商数据，预计2023年算力投入分别为 452/195/149 亿元，分别同比+35%/+40%/+20%。

图表19 2018-2023Q1全球云服务支出规模(亿美元)



图表20 三大运营商2022年CAPEX表现及未来计划

运营商	中国移动	中国电信	中国联通
2022年底 数据中心机架数	46.7万架	51.3万架	36.3万架
算力规模总量	8.0EFLOPS	3.8EFLOPS	-
2023年计划 算力方面投资	452亿元	195亿元	149亿元
2023年 计划新增机架数	4万架	4.7万架	2.7万架
2022年 云业务收入	503亿元	579亿元	361亿元
云业务同比增长	108%	108%	121%

资料来源：东方财富网、中商情报网、巴中在线、华金证券研究所

建设计划：政策—管理办法落地，监管走上正轨

- **国家网信办等七部门联合公布《生成式人工智能服务管理暂行办法》。**上述《办法》自2023年8月15日起施行，国家网信办有关负责人表示，出台《办法》，旨在促进生成式人工智能健康发展和规范应用，维护国家安全和社会公共利益，保护公民、法人和其他组织的合法权益。同时《办法》也指出，应该促进算力资源协同共享，提升算力资源利用效能。

图表21 《生成式人工智能服务管理暂行办法》内容

	关键要点
①	鼓励生成式人工智能技术在各行业、各领域的创新应用，生成积极健康、向上向善的优质内容，探索优化应用场景，构建应用生态体系。
②	鼓励生成式人工智能算法、框架、芯片及配套软件平台等基础技术的自主创新，平等互利开展国际交流与合作，参与生成式人工智能相关国际规则制定。
③	采取有效措施鼓励生成式人工智能创新发展，对生成式人工智能服务实行包容审慎和分类分级监管。
④	支持行业组织、企业、教育和科研机构、公共文化机构、有关专业机构等在生成式人工智能技术创新、数据资源建设、转化应用、风险防范等方面开展协作。
⑤	推动生成式人工智能基础设施和公共训练数据资源平台建设。促进算力资源协同共享，提升算力资源利用效能。
⑥	推动公共数据分类分级有序开放，扩展高质量的公共训练数据资源。鼓励采用安全可信的芯片、软件、工具、算力和数据资源。

图表22 AIGC内容合规建设体系

建设计划：政府—智算中心建设开启，国内需求逐步释放

- **国内智算中心建设明确，后续需求有望逐步释放**。多地政府出台算力规划、三大运营商加速筹备智算中心建设以及互联网厂商多地智算中心建设规划。智能计算中心的建设，整合数据资源结构的同时，也带来更为完善且健全的算力、算法基础设施，为人工智能技术的创新及应用提供强有力的支撑。

图表23 国内智算中心建设情况

类别	项目	算力规模(PFLOPS)	项目进度
企业自建	阿里云张北超级智算中心	12000	投入运营
企业自建	商汤科技AI计算中心	3740	投入运营
企业自建	阿里云乌兰察布智算中心	3000	建设中
政府主导	北京数字经济算力中心	1000以上	规划阶段
政府主导	天津AI计算中心	300	一期完工

图表24 中国银行总行金融科技中心合肥园区项目

投资金额	29亿元
机架数量	一期拟建设5700架机柜 最终规模不少于1万架机柜
服务器数量	10万台服务器
占地规模	占地106亩土地 总建筑面积约20万平方米
项目特点	定位为中行全球化科技研发和运营中心，是中行分布式架构体系重要生产基地，承载数据基础设施、创新中心、新技术研发和科技培训等多元功能。

建设计划：企业—模型推进计划明确，垂直行业延伸

- **科大讯飞提出大模型的建设进度计划**。针对大模型普遍存在的问题，科大讯飞明确规划了一系列的迭代里程碑。计划在不同时间节点进行不同方面的升级，力争超越ChatGPT。目前国内大模型厂商中，科大讯飞制定了明确的追赶GPT-3.5的时间表，表明了他们在技术研发上的整体规划。
- **科大讯飞星火大模型落地多应用场景实现闭环，赋能开放平台共建生态**。公司发布教育、消费者、医疗、政法等多种应用场景的实现，配合公司相关产品使用，进一步提高公司市场渗透率。

图表25 科大讯飞大模型关键时间节点

图表26 科大讯飞星火代码大模型能力

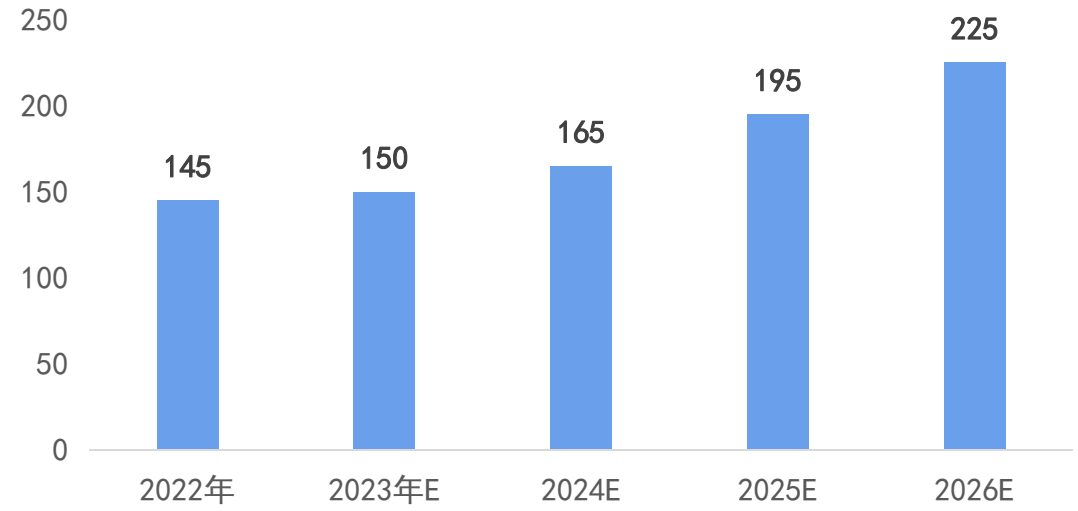
目录

规模：算力增长助力AI服务器规模化布局

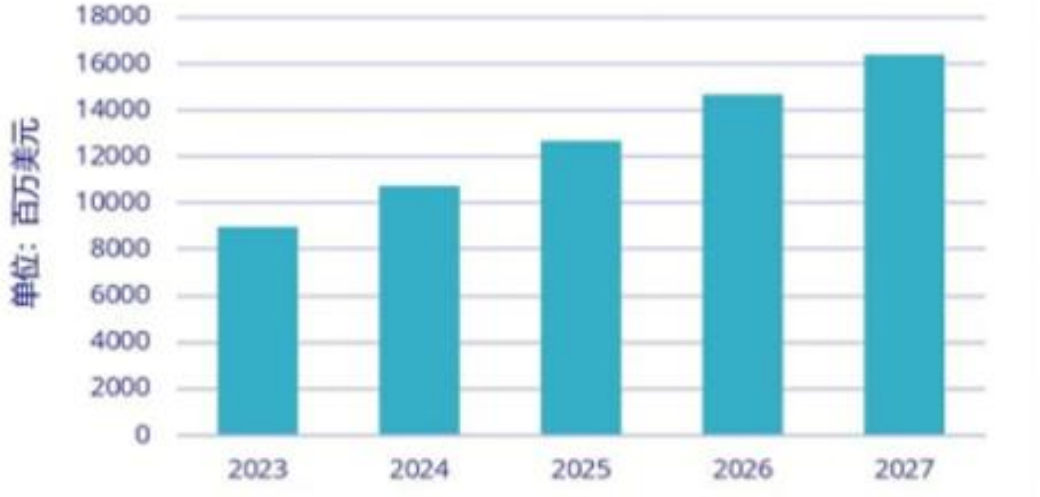
华金证券 Huajin Securities
华发集团旗下企业

- **算力持续增长，AI服务器应用场景更加广泛。**Trendforce预估，预估2022年全球搭载GPGPU的AI服务器年出货量占整体服务器比重近1%，即约14万台。预计2023年其出货量年成长可达8%，到2026年预计全球搭载GPGPU的AI服务器出货量将达到22.5万台左右，2022~2026年CAGR将达10.8%。
- **中国AI算力市场高速发展，将成为第二增长极。**根据IDC预测，到2027年中国加速服务器市场规模将达到164亿美元。其中非GPU服务器市场规模将超过13%。智慧城市、智能机器人、智能家居、工业领域将成为主要应用领域。

图表27 全球AI服务器出货量预测（千台）



图表28 中国加速服务器市场预测



资料来源：观研报告网、IDC中国、华金证券研究所

机架服务器：传统数据中心标配，大型机架渐成主流

- 我国服务器机架数量快速增长，大型机架逐渐成为主流。2021年我国服务器机架总数达429万架，大型规模以上机架361万架，占比达84.1%。
- 企业规模不同，配置的服务器不同。服务器按照可支持的CPU数量可分为单路、双路、四路及多路服务器，单路即为1个CPU。大中小型企业通常配置4-8路、2-4路、1-2路服务器。
- 机架式服务器的通常功耗不超过30W。

图表29 浪潮信息机架服务器 NF5280G7

图表30 NF5280G7服务器性能

处理器×2：最多支持60核；最高睿频4.2GHz;四条UPI互连链路，单条链路最高速率16GT/s;最大热设计功率350W
GPU×4：四条UPI互连链路，单条链路最高速率16GT/s
CPU×2：支持32条DIMM
内存×32：最多支持32条DDR5 4800MT/s内存

刀片服务器：为云而生，高可用、高密度是优势

- 刀片式服务器是指在标准高度的机架式机箱内可插装多个卡式的服务器单元，实现高可用和高密度。每一块"刀片"实际上就是一块系统主板，每块"刀片"都是热插拔。
- 比较：1）刀片式服务器更省空间，6机架42个1U的服务器和1机架的刀片服务器性能相当；2）刀片式服务器布线更简单、更省电；3）机架式比刀片式服务器更加灵活、维护更简单、维护成本更低；4）刀片式服务器对制冷的要求更高；5）刀片式服务器对调度系统的要求更多

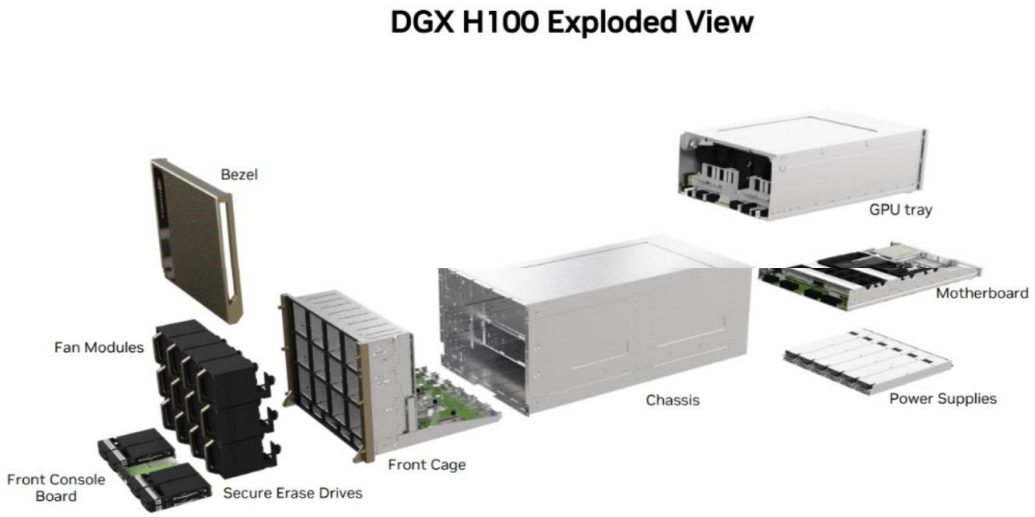
图表31 刀片服务器

图表32 机柜中的刀片服务器

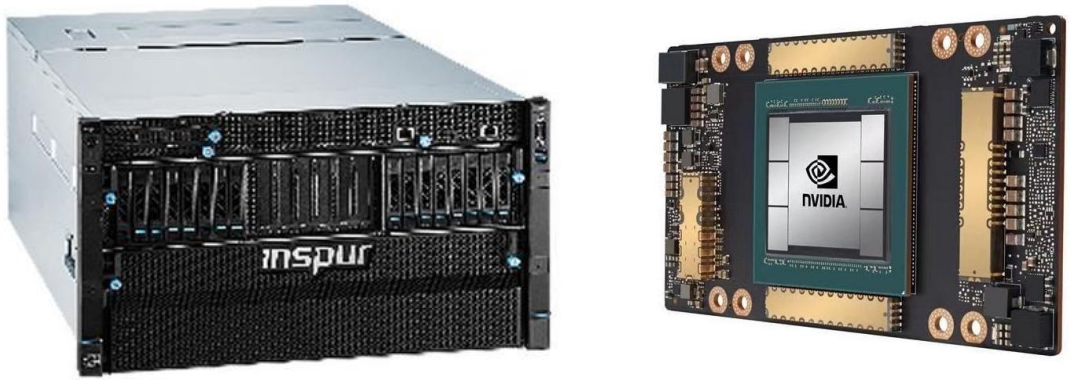
AI服务器：为计算而生，性能全方面升级

- 按照GPU数量分类：AI服务器主要采用加速卡为主导的异构形式，更擅长并行计算。与通用服务器按照CPU数量分类不同，AI服务器一般仅搭载1-2块CPU,GPU数量显著占优。按GPU数量，可分为四路、八路和十六路服务器，其中搭载8块GPU的八路AI服务器最常见。
- AI服务器相较于普通服务器存在各方面升级。AI服务器增加了GPU的使用数量，相配套的在带宽、散热、内存、存储等方面也会有相应升级。

图表33 NVIDIA H100服务器



图表34 浪潮信息服务器



AI服务器：主流AI芯片V100/A100/H100

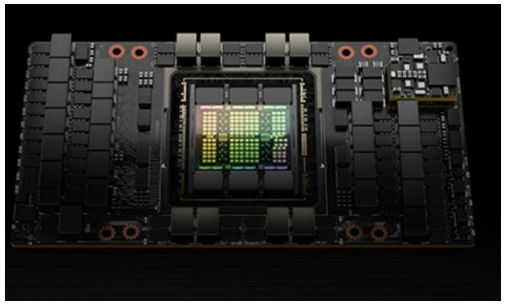
- V100：NVIDIA于2017年5月发布V100 Tensor Core GPU。采用NVIDIA Volta架构，可在单个GPU中提供近100个CPU的性能，减少优化内存使用率的时间。
- A100：于2020年5月发布，采用全新Ampere安培架构的超大核心GA100,7nm工艺，542亿晶体管，826平方毫米面积，6912个核心，搭载5120-bit 40/80GB HBM2显存，带宽近1.6TB/s,功耗400W。可针对AI、数据分析和 HPC 应用场景，在不同规模下实现出色的加速。
- A800：于2022年11月推出,用于替代A100。传输速率为每秒400GB,低于A100的每秒600GB,代表了数据中心的性能明显下降。A800支持内存带宽最高达2TB/s,其他参数变化不大。
- H100：于2022年4月发布，由800亿个晶体管组成，采用全新Transformer引擎和NVIDIA NVLink互连技术，以加速最大规模的AI模型，如大型语言模型，并推动对话式AI和药物发现等领域的创新。

图表35 NVIDIA V100

图表36 NVIDIA A100

图表37 NVIDIA A800

图表38 NVIDIA H100



AI服务器：算力摩尔定律延续

图表39 英伟达主要芯片比较

性能参数	V100 PCIe	A100 80GB PCIe	A800 80GB PCIe	H100 80GB PCIe
FP64	7TFLOPS	9.7TFLOPS		26 TFLOPS
FP32	14TFLOPS	19.5TFLOPS		51 TFLOPS
FP16 Tensor Core		312TFLOPS		756.5 TFLOPS
GPU显存	32/16GB HBM2	80GB HBM2e		80GB
GPU显存带宽	900 GB/s	1935GB/s		2TB/s
最大热设计功耗(TDP)	250瓦	300W		300-350W
多实例GPU			最多7个MIG 每个10GB	
外形规格		PCIe双插槽风冷式 或单插槽液冷式		PCIe双插槽风冷式
互连技术	NVLink:300 GB/s PCIE:32 GB/s	搭载2个GPU的 NVIDIA"NVLink" 桥接器：600GB/s PCIe 4.0:64GB/s	搭载2个GPU的 NVIDIA"NVLink" 桥接器：400GB/s PCIe 4.0:64GB/s	NVLink:600GB/s PCIe 5.0:128GB/s
服务器选项	搭载1至8个GPU的合作伙伴认证系统和NVIDIA认证系统			

资料来源：CSDN、英伟达产品数据表、华金证券研究所

AI服务器：算力硬件成本更高，单台价值突破百万

◆ 以AI服务器NF5688M6为例，该服务器采用2颗第三代Intel Xeon可扩展处理器+8颗英伟达A800 GPU的组合，每颗A800售价104000元，故该服务器芯片成本约96万元，相比于通用服务器NF5280M6，贵出近8倍。

图表40 浪潮信息 NF5688F6

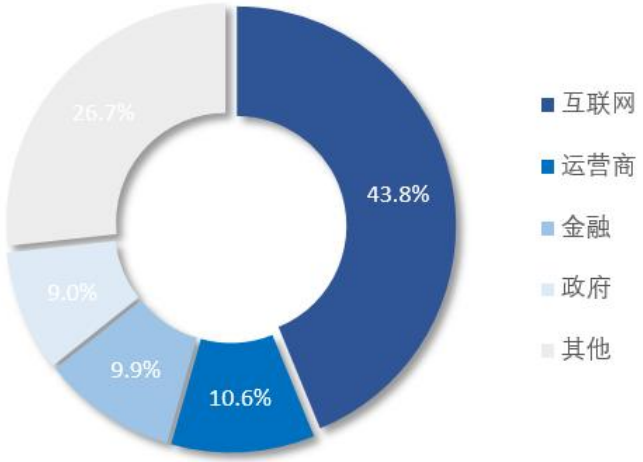
图表41 AI服务器主要结构

处理器×2：第三代英特尔至强可扩展处理器（Ice Lake），单颗处理器高达40核、最高睿频频率 3.7GHz，最多 3 条 UPI 互连链路，速率最高可达11.2GT/s。
GPU×8：任意两个 GPU 之间可以直接进行数据 P2P 交互，GPU 间 P2P 通信速率为400GB/s
CPU×2：配合领先的 UPI 总线互联设计，可为深度学习业务场景提供顶级 AI 计算性能
内存×32：内存支持LRDIMM/RDIMM/BPS，可提供优异的速度、高可用性及最多 4TB 的内存容量
硬盘×8/16：支持多种灵活的硬盘配置方案，提供了弹性的、可扩展的存储容量空间，满足不同存储容量的需求和升级要求。
网卡：可选配1张PCIe 4.0 x16 OCP 3.0网卡，速率支持10G/25G/100G

传统服务器需求：下游需求多点开花，金融行业投资加速

- **服务器下游需求多点开花。**下游需求主体为互联网企业、云计算企业、政府部门、金融机构、电信运营商等。其中互联网行业占比43.8%，位列第一。在互联网行业中服务器需求主要集中在快手、百度等大型企业。运营商、金融、政府占比分别为10.6%、9.9%、9%。
- **各行业服务器投资支出不断提升：**IDC统计数据显示，2022年全球云基础设施支出达到了750亿美元，比2021年的650亿美元提高了15.4%，全球金融行业对服务器的投资增长了20%，达到180亿美元。

图表42 国内服务器下游应用结构分布

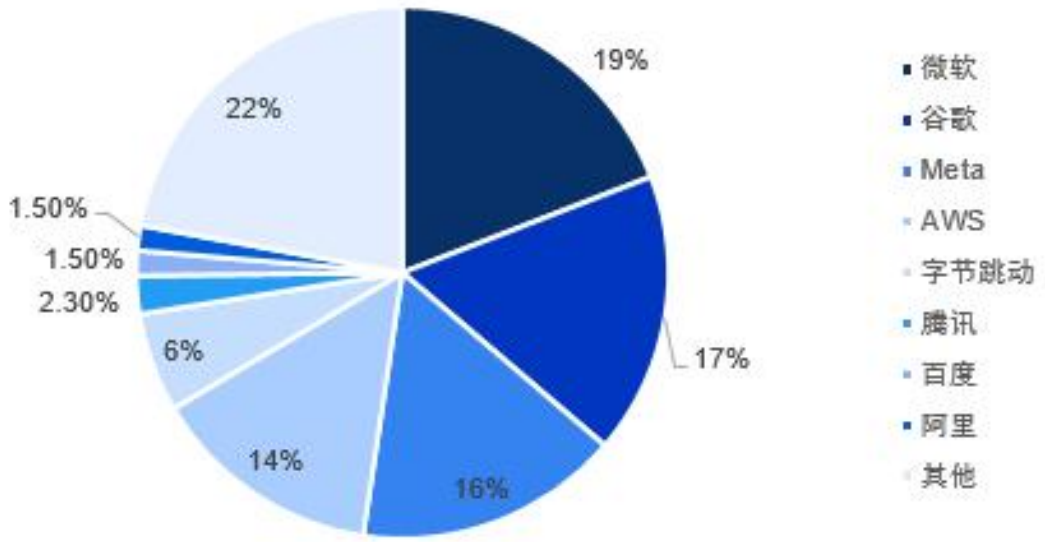


图表43 云计算&互联网厂商资本开支增速与服务器市场增速显著正相关

AI服务器需求：大模型不断涌现，训练需求爆发

- AI服务器的采购群体较为集中。主要集中在互联网厂商，政府、高校的需求量相对较少。其中高校主要需要训练模型。
- 华为盘古 NLP 大模型，该模型采用了深度学习和自然语言处理技术，并使用了大量的中文语料库进行训练。该模型拥有超过 1 千亿个参数，可以支持多种自然语言处理任务，包括文本生成、文本分类、问答系统等。

图表44 2022年AI服务器采购厂商占比

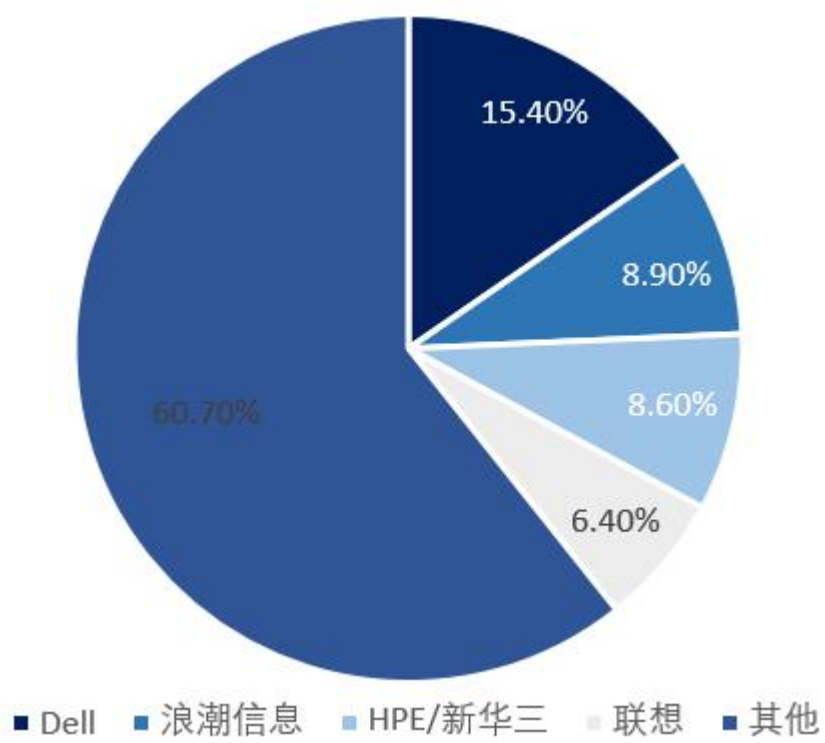


图表45 华为盘古大模型

传统服务器供给：市场增长稳定，产能自给自足

- 全球2021年服务器出货量1353.9万台;中国2021年服务器出货量391.1万台。
- 据IDC《2022年第四季度中国服务器市场跟踪报告Prelim》统计，浪潮信息份额28.1%、新华三份额17.2%、超聚变份额10.1%；浪潮信息服务器22年出货158万台服务器。

图表46 2021年全球服务器厂商份额占比



图表47 全球和中国服务器市场规模测算

	测算出货量	2022E	假设单价	市场空间测算（亿元）
全球	全球2021年服务器出货量1353.9万台	1380万台	2021年全球服务器平均单价高达7328美元/台(约合人民币4.9万元)	1380*4.9=6762
中国	我国2021年服务器出货量391.1万台	400万台	单台双路服务器价格为4-5万(中国市场平均价格约为4.3万元）	400*4.3=1720

资料来源：知乎、C114通信网、财闻网、观研天下、华金证券研究所

AI服务器供给：国内芯片性能及产能受限

- 英伟达芯片产能不足，限制服务器量产：AI服务器需要配备4-8颗GPU，英伟达在GPU市场占有绝对优势，占据全球八成市场份额。其产能有限，同时受到美国禁售影响，影响服务器产能。
- 国产替代化正当时，但仍存在差距：国产寒武纪、昇腾也研发出AI芯片，寒武纪的思元芯片目前多应用在推理场景，通用性较弱，性能不足。

图表48 AI服务器代工情况

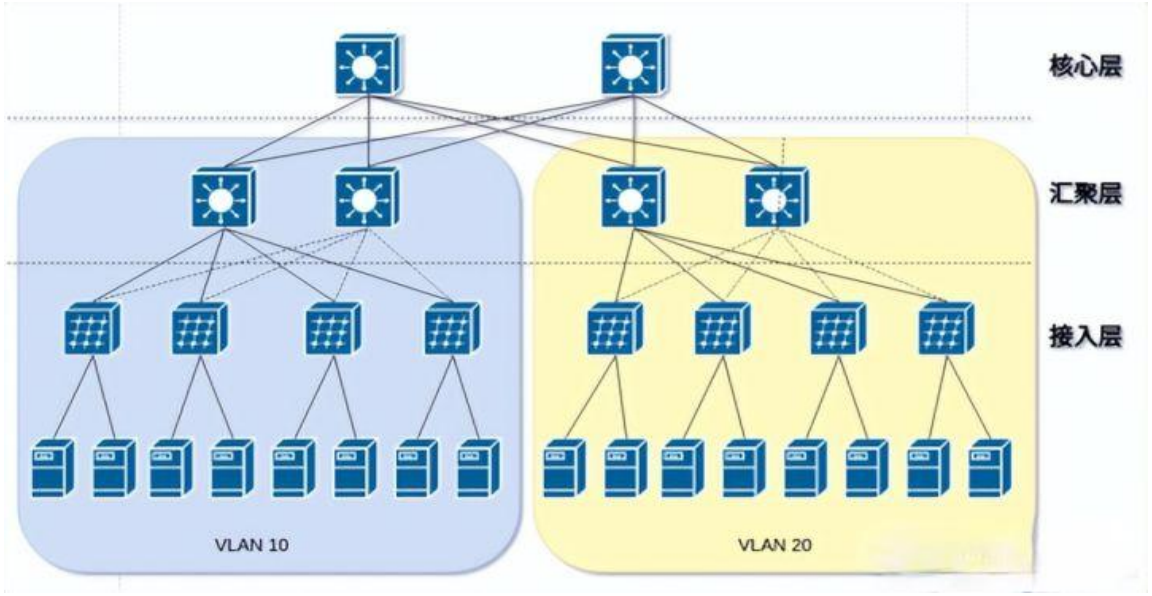
公司	布局	预计业绩增长情况
鸿海	通过旗下工业富联（FII）与鸿佰科技冲刺AI服务器业务，已着手开发下一代AI服务器产品	预计鸿海今年AI服务器占比将超过30%，较去年大增五成，稳居全球最大AI服务器厂。
广达	领先与英伟达合作，通过旗下云达推出英伟达全新HPC-AI服务器。将持续扩充AI服务器产能。	服务器产品营收占比已达三成，预期全年出货可望成长个位数百分比。
仁宝	冲刺发展AI产品多元应用，力拼实质提升市场需求	出货量今年估计可年增双位数百分比；预估2025年5G、服务器等新事业营收占比可望达5%，2026年力拼翻倍至10%。
纬创	通过旗下纬颖布局云端服务器领域，包括提供系统解决方案服务，设计一系列巨型资料中心等各种云端服务。	目前营收占比约20%，今年已到手新订单中，超过五成为AI服务器。
英业达		往年旗下服务器营收占比仅约5%，预估英业达明年AI服务器业务表现年增双位数百分比。

目录

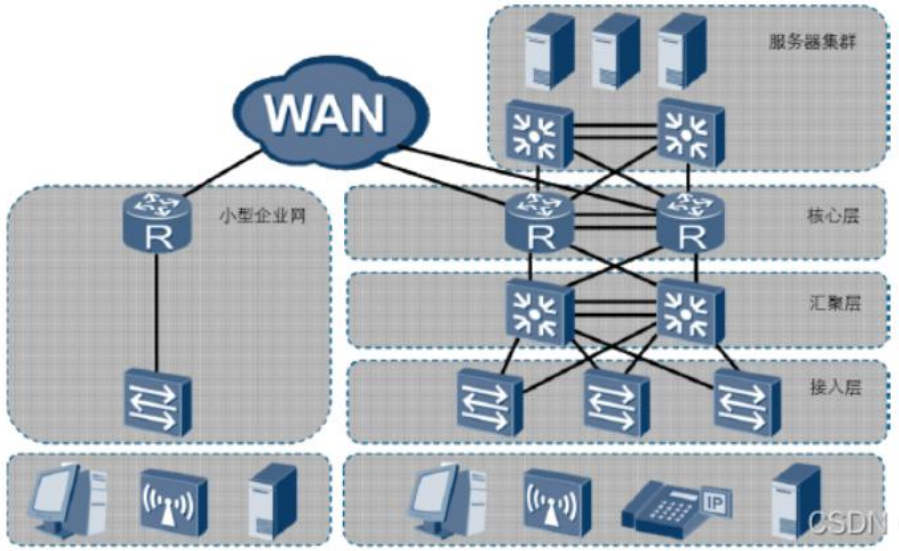
网络架构：传统三层网络——管理网络的经典稳定架构

- 核心层(Core Layer):核心交换机为进出数据中心的包提供高速的转发，为多个汇聚层提供连接性，核心交换机通常为整个网络提供一个弹性的L3路由网络。
- 汇聚层(Aggregation Layer):汇聚交换机连接接入交换机，同时提供其他的服务，例如防火墙，SSL offload，入侵检测，网络分析等。它可以是二层交换机也可以是三层交换机。
- 接入层(Access Layer):接入交换机通常位于机架顶部，所以它们也被称为ToR(Top of Rack)交换机，它们物理连接服务器。通常情况下，汇聚交换机是 L2 和 L3 网络的分界点：汇聚交换机以下的是 L2 网络，以上是 L3 网络。每组汇聚交换机管理一个POD(Point Of Delivery)，每个POD内都是独立的 VLAN 网络。

图表49 传统三层网络架构



图表50 企业网络三层网络架构

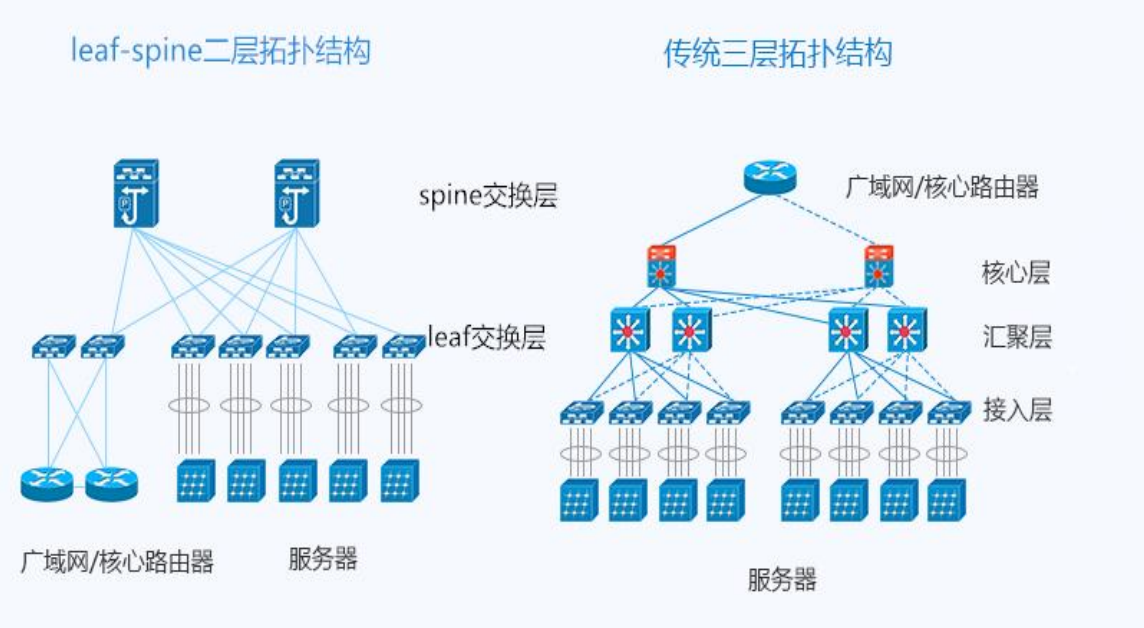


网络架构：叶脊网络——扁平化高速新型架构

华金证券 Huajin Securities
华 发 集 团 旗 下 企 业

- 随着数据中心内流量的快速增长和数据中心规模的不断扩大，传统的三层网络拓扑结构不能满足的数据中心内部高速互连的需求，慢慢演进为leaf-spine叶脊架构，叶脊架构分为两层：Leaf(叶)交换层 、 Spine(脊)交换层。
- Spine(脊)交换层:网络核心节点，提供高速 IP 转发能力，通过高速接口连接各个功能 Leaf 节点。
- Leaf(叶)交换层:网络功能接入节点，提供各种网络设备接入功能。

图表51 叶脊架构与传统架构对比



图表52 Leaf(叶)交换层各节点名称及功能

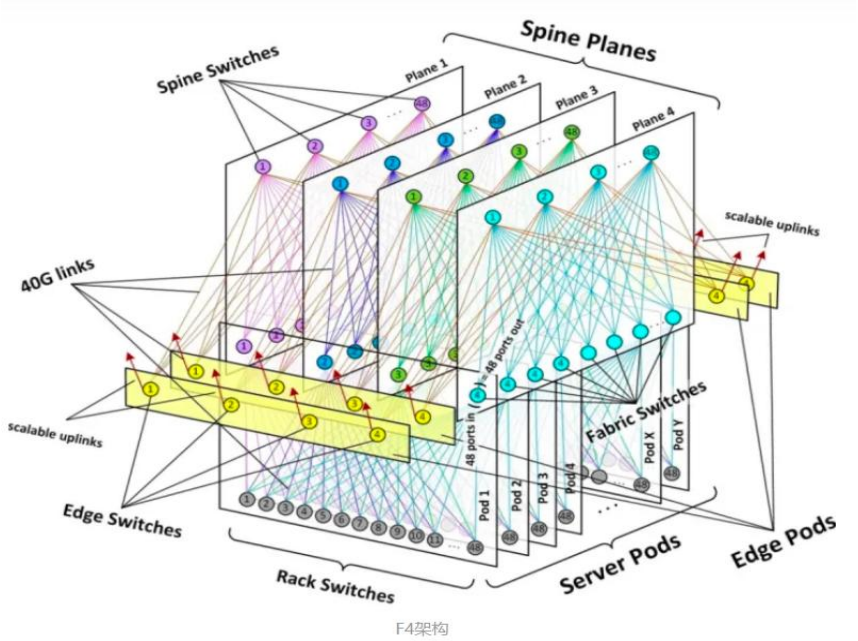
Leaf节点名称	实现功能
Server Leaf	提供虚拟化服务器、物理机等计算资源接入功能
Service Leaf	提供防火墙、负载均衡等增值服务接入功能
Border Leaf	连接外部路由器，提供数据中心外部流量接入功能

图表53 叶脊架构优势

网络架构：基于叶脊的升级——流量模块化架构

- Facebook从14年开始对原有的数据中心网络架构进行改造，原因就在于面对网络流量2-4倍的未来扩张，现有的三层网络架构难以胜任。
- Facebook提出了自己的下一代数据中心网络——data center fabric网络架构（也称为F4网络），在原始叶脊网络基础上进行模块化组网，能够承载数据中心内部的大量东西流量的转发，并且保证了足够的扩展性。
- Facebook于19年提出了F16架构，原因在于F4架构在网络流量持续上涨、硬件更新换代的现在已经难以满足数据中心的需求。

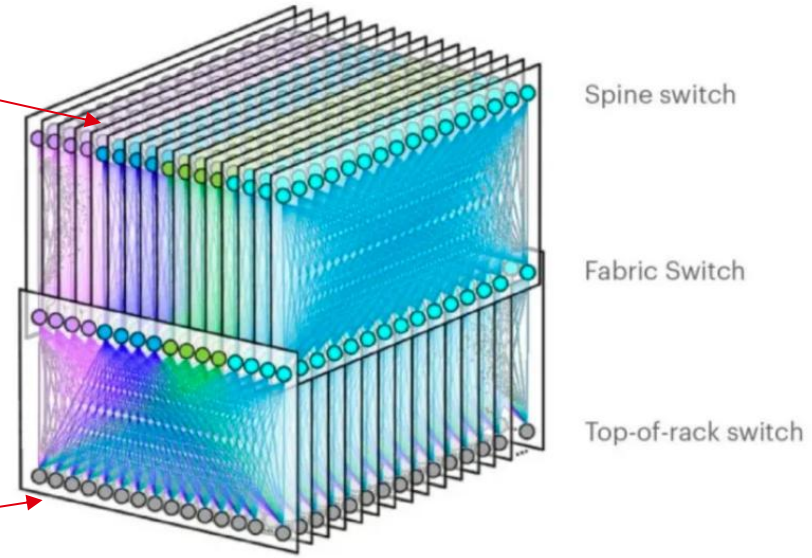
图表54 Facebook F4网络架构



F4架构

- Spine平面增加为16个。单芯片处理能力提升为12.8TBps，减少体积，仅需跨越5个芯片即可通过路径。
- 机架采用Wedge 100s，拥有1.6T上行带宽和1.6T下行带宽，流量收敛比达到1：1；机架上方的平面包含16个128端口的100G光纤交换机。

图表55 Facebook F16网络架构

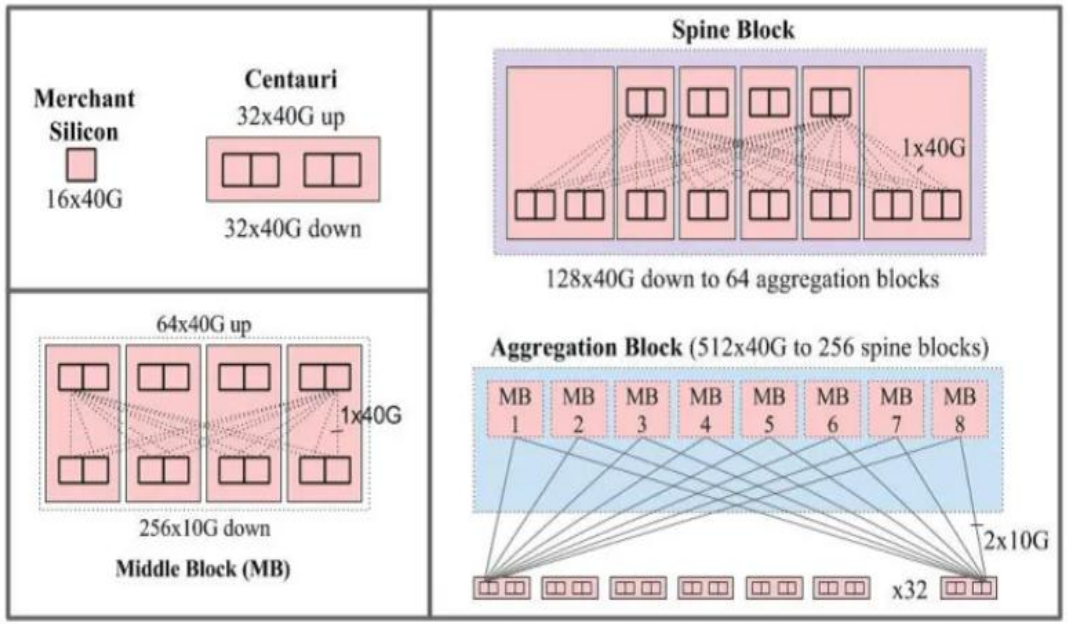


F16架构

网络架构：基于叶脊的升级——动态分布式网络架构

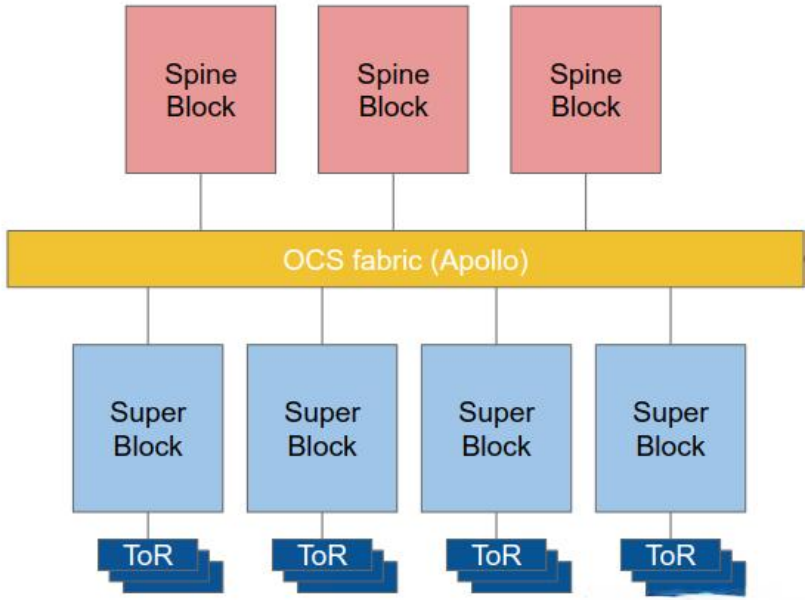
- 谷歌数据中心第五代架构Jupiter Network Fabrics。Jupiter可以视为一个三层Clos。Leaf交换机作为ToR(top-of-rack架顶式交换机)，向北连接到名为Middle Block的Spine交换机，再向北还有一层Super Spine名为Spine Block。
- Apollo结构。Super spine方案的交换机数量可以一直增长，新增交换机需要和原有的Pod全部互联。为解决互联问题，谷歌在Spine层和Pod层之间加入Apollo Fabric。此结构解除了Spine Block和Super Block的直连，但又能够动态地调整连接关系，高效实现了全互联，动态地调整网络流量的分布。

图表56 Google Jupiter网络架构



Jupiter拓扑

图表57 Google Apollo结构



数据中心交换机：算力网络的底座

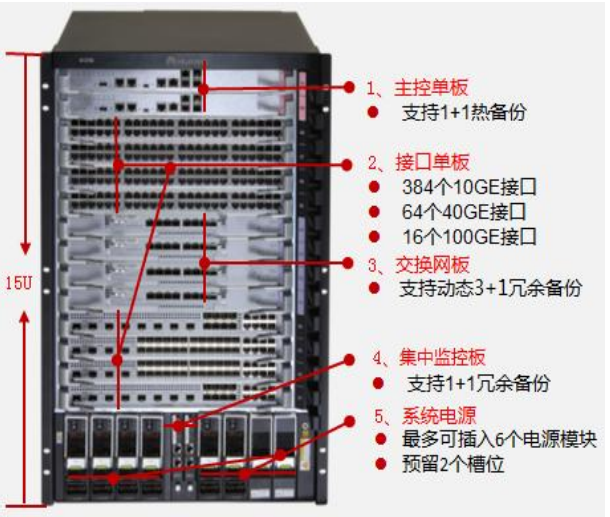
- 交换机相当于一台特殊的计算机，由硬件和软件组成，负责构建网络进而实现所有设备的互联互通，是算力网络底座的主要组成部分。
- 数据中心交换机的功能：1、网络数据流量分发；2、处理大带宽数据。
- 数据中心交换机按照位置分类可分为接入层交换机、汇聚层交换机和核心层交换机。
- 数据中心交换机按照形态分类可分为盒式交换机（多位于接入层和汇聚层）和机框式交换机（多位于核心层）。

图表58 按位置分类交换机特点

接入层	·端口数量较少 ·支持高速连接 ·数据传输低延迟
汇聚层	·承载大量的流量 ·连接方式灵活 ·支持上行、下行链路连接
核心层	·高性能和高可用性 ·大容量的交换能力 ·高速的转发能力

图表59 按形态分类交换机示意图

盒式交换机

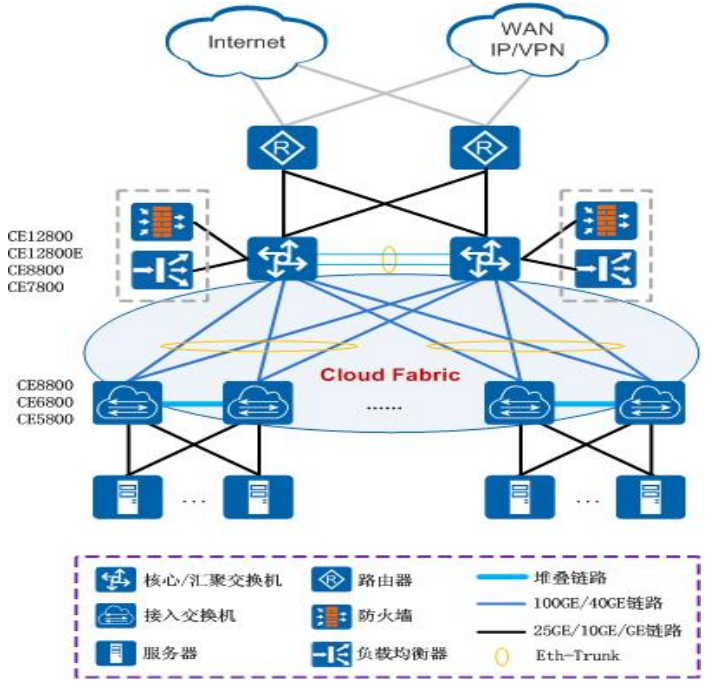


框式交换机

数据中心交换机：背板带宽和包转发率是两大核心

- 接入层、核心层、汇聚层交换机的重要参数是交换机的背板带宽和包转发率。
- 背板带宽是交换机接口处理器或接口卡和数据总线间所能吞吐的最大数据量，标志了交换机总的数据交换能力。
- 包转发率指交换机每秒可以转发多少百万个数据包（Mpps），即交换机能同时转发的数据包的数量，以数据包为单位体现了交换机的交换能力。
- 决定包转发率的重要指标是背板带宽，背板带宽越高，所能处理数据的能力就越强，包转发率也就越高。

图表60 华为交换机产品位置图



图表61 交换机选择标准（以包转发率和背板带宽为主要决定参数）

交换机层级	选择标准（以1000路摄像机为例）
接入层	一般项目中使用百兆交换机较多，一个百兆交换机带机量不要超过8个，8路以上摄像机需采用千兆上联交换机。
汇聚层	汇聚层交换机需要支持同时转发500M以上的交换容量，多选用千兆交换机。交换机需满足背板带宽至少32Gbps，包转发率至少23.44Mpps。
核心层	核心层交换机需要支持4000Mbps的带宽，可选择高带宽千兆交换机或万兆交换机，上联端口应为万兆，交换机需满足背板带宽至少64Gbps，包转发率至少47.56Mpps。

数据中心交换机：主流100/200/400/800G交换机

- 数据中心交换机根据接口速率不同，可分为100/200/400/800G交换机；不同速率交换机的单板能力(每个插槽的端口数量或者带宽)越来越强，所有端口升级一次其传输容量将直接翻倍。
- 我国主要厂商突破技术瓶颈，实现快速发展。新华三旗下S10500、S12500等系列交换机，多采用CLOS架构，提供Seerblade高性能AI计算模块，目前已形成了完整的数据中心产品体系。华为主推CloudEngine系列数据中心交换机，系列基于华为新一代的VRP8操作系统，2023Q1该系列以34.6%的份额排中国数据中心交换机市场第一，其中16800-X为全球首款800GE数据中心核心交换机。

图表62锐捷网络100G交换机

- 产品支持48*100G DSFP端口+8*400G QSFP-DD端口，2*电源模块，6*风扇模块。包转发速率可达5350Mpps。

图表63 华为200G交换机

- 产品采用华为iLossless智能无损算法，支持32*200G端口或8*400G端口。交换容量可达19.2Tbps，包转发速率7200Mpps。

图表64 新华三400G交换机


S10500X-G系列太网核心交换机

- S10500X-G采用CLOS无中板交换架构，交换容量384-1024Tbps，包转发速率72000-1920000Mpps。

图表65 华为800G交换机


CloudEngine 16800-X

- CloudEngine 16800-X支持288*800GE端口，3.5微秒跨板转发时延，支持三网融合，总运营成本可降低36%。

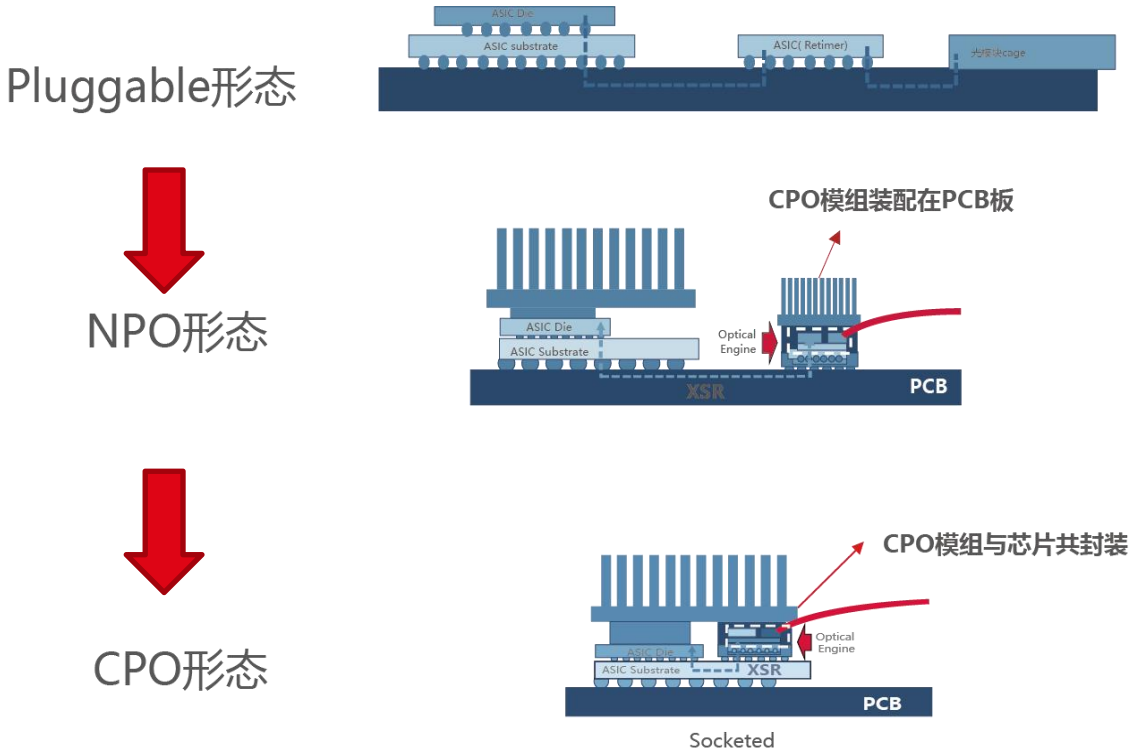
NPO交换机：传统可拔插向CPO过渡（封装角度）

- NPO交换机，应用NPO（Near Packaged Optics）近封装光学技术，属于传统可拔插(Pluggable)向CPO技术的过渡阶段。
- NPO交换机可以在CPO技术生态完备之前，在最短时间内享受到低成本、低功耗的收益。

图表66 锐捷NPO交换机

- 锐捷51.2T硅光NPO冷板式硅光交换机在1RU的空间内，实现了64口800G的超高密度端口设计，升级了外置激光源的设计方案。采用独立迷你交换板设计，大幅降低了高速率PCB的成本。

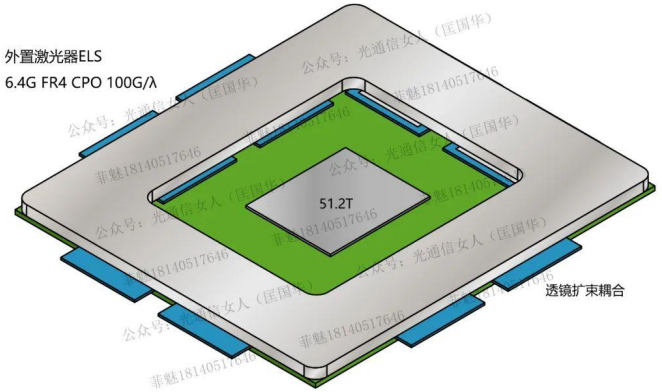
图表67 硅光技术形态概览



CPO交换机：硅光技术的最终形态（封装角度）

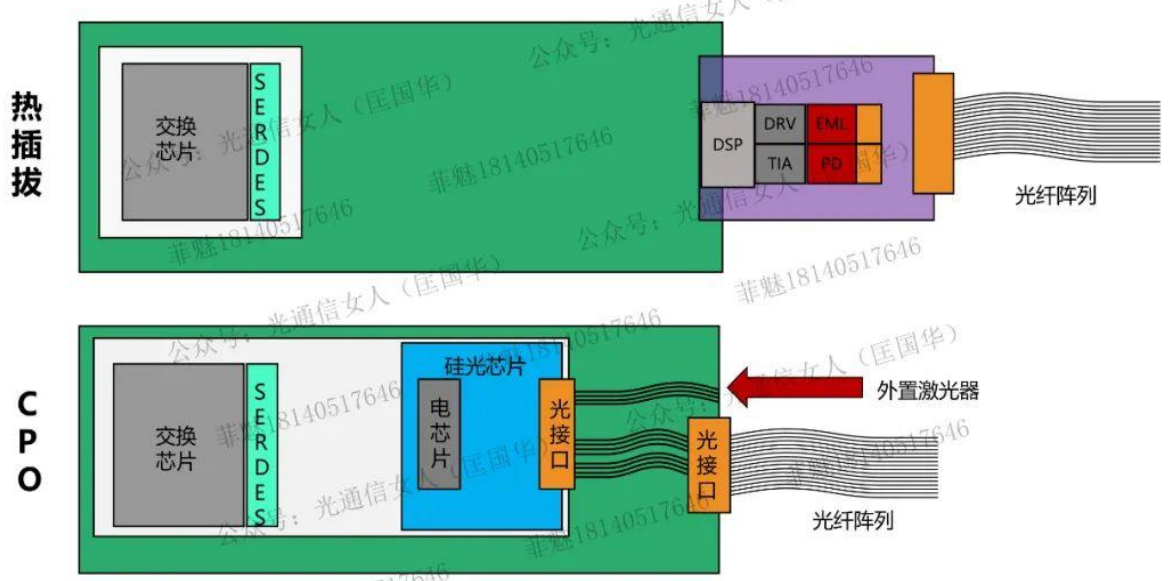
- CPO交换机，应用CPO（Co-Packaged Optics）光电共封装技术，将网络交换芯片和光模块共同装配在同一个插槽上，形成芯片和模组的共封装，可以缩短交换芯片和光引擎之间的距离，以帮助电信号在芯片和引擎之间更快地传输。
- CPO具备高密度、低成本、小体积、低功耗的特点，成为AI高算力情况下降低能耗从而降低成本的可选方法。
- 海外市场，高速率板块对CPO需求更为迫切，国内上量节奏紧随其后。

图表68 OFC第二代51.2T CPO



- 第二代CPO容量倍增，单波100Gbps，一个引擎64个通道，采用外置激光器。光模块部分实现了低功耗，可达7pJ/bit。

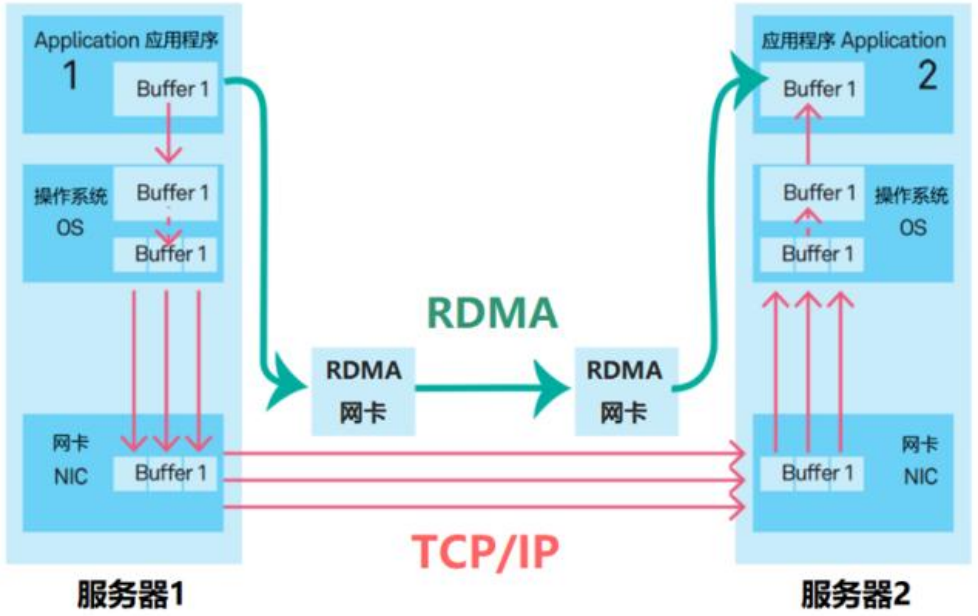
图表69 CPO技术框架图



IB交换机：适用于IB网络的交换机（网络协议角度）

- IB 交换机全称InfiniBand 交换机，IB属于RDMA技术，与TCP/IP技术在机理上不同。在市场上，Mellanox InfiniBand 交换机、Intel 和 Oracle InfiniteBand 交换机是三大名牌领先的 IB 交换机。其中Mellanox IB交换机 QM8790交换机可提供多达提供多达 40 个 200Gb/s 端口。

图表70 RDMA工作原理



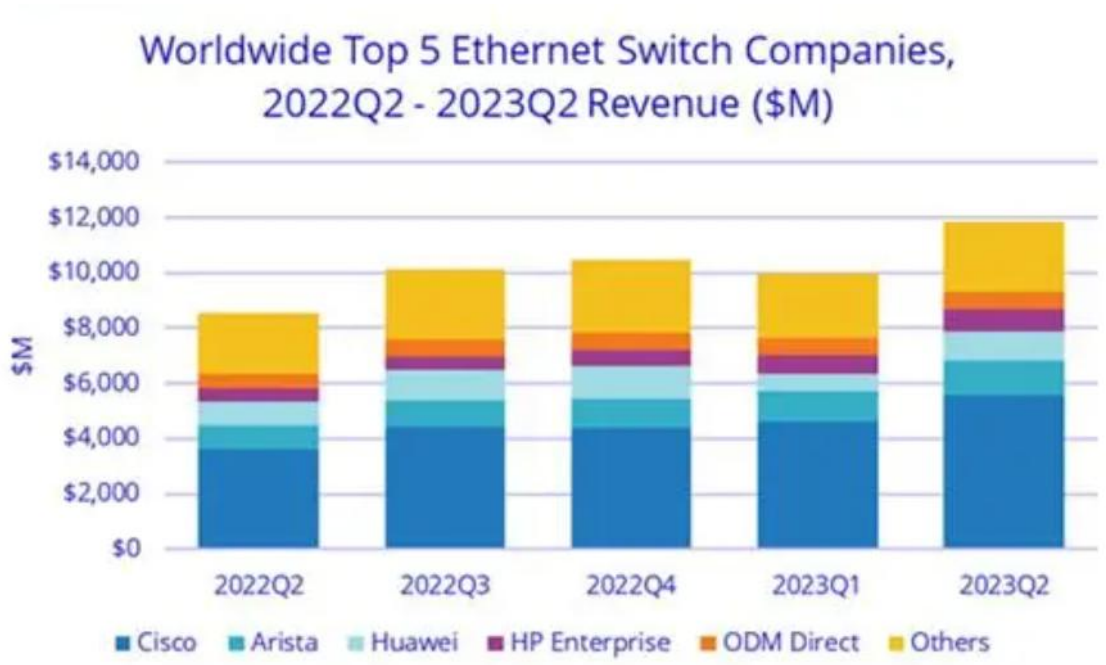
图表71 IB交换机和以太网交换机区别

区别	IB交换机	以太网交换机
速率不同	InfiniBand目前可实现400Gb/s网速的单端口传输带宽。	千兆交换机支持1000Mbps，万兆交换最大支持100G。
应用场景	IB交换机通常用于高性能计算领域，如超级计算机、数据中心等。	以太网交换机则更常用于企业网络、服务器机房等。
包转发率	IB交换机通常使用硬件转发方式，具有更高的包转发率。	以太网交换机通常使用软件转发方式，因此包转发率相对较低
协议与标准	IB交换机使用InfiniBand协议和技术标准。	以太网交换机使用以太网协议和IEEE 802.3技术标准。

全球：200/400G拉动整体市场增长

- 云计算业务和云流量快速发展，全球数据中心交换机市场继续保持强劲增长。2022年，全球交换机市场规模为365亿美元，同比增长18.7%。全球数据中心用以太网交换机收入占比达47%，同比增长20.8%，相应的端口出货量同比增长3.7%；
- 200/400Gb交换机拉动整体市场增长。2022年200/400GbE交换机的市场收入增长超过300%，第四季度收入环比增长为0.7%。2022年第四季度，100GbE收入同比增长18.7%，全年增长22.0%，第四季度25/50GbE收入同比增长30.1%，全年增长29.8%。

图表72 2023年第二季度全球以太网交换机市场



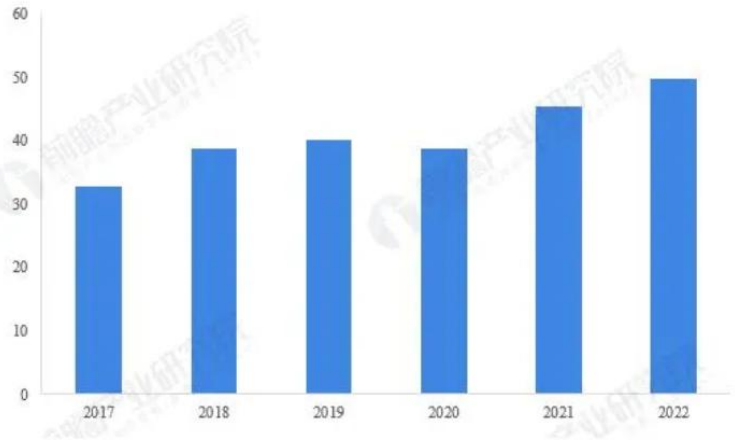
图表73 各速率交换机市场同比变化情况

速率	2022Q4同比	2022全年同比
1GbE	增长21.4%	增长12.6%
2.5/5GbE	增长122.1%	增长53.1%
10GbE	增长1.0%	增长0.4%
25/50GbE	增长30.1%	增长29.8%
100GbE	增长18.7%	增长22.0%
200/400GbE	增长0.7%	增长超过300%

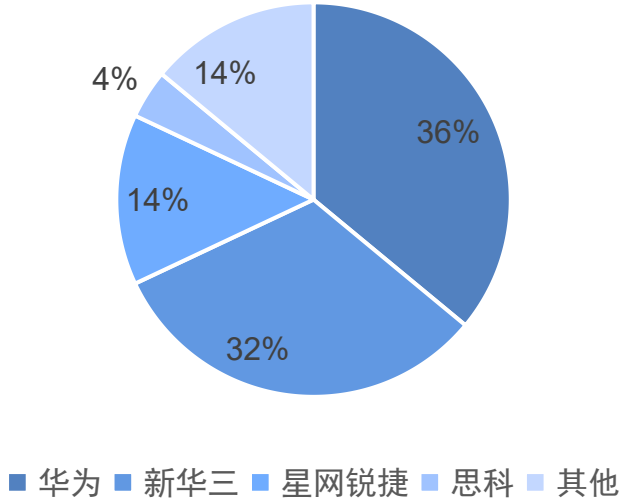
国内：国产厂商快速崛起，海外仍有较大扩张空间

- 受益于强劲的AI需求，中国交换机市场规模持续扩张。2021年中国交换机行业市场规模为501亿元，同比增长17.33%，预计2023年市场规模达到685亿元。
- 国产设备商快速崛起。2021年中国市场数据中心交换机前四厂商为华为、新华三、星网锐捷和思科。从份额占比来看当前国内交换机市场双龙头格局特征显著，华为占比36%、新华三占比32%。
- 中国厂家在全球市场仍有较大扩张空间。2022年全球交换机市场份额，上榜公司有华为、新华三，共占比15.7%。主要市场仍为思科占领，为43.3%。

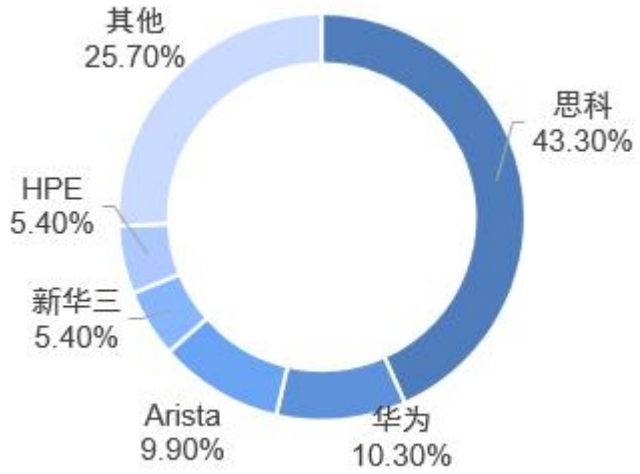
图表74 17-22年中国交换机市场规模变化趋势（单位：亿美元）



图表75 2021年中国交换机各公司市场份额



图表76 2022年全球交换机各公司市场份额

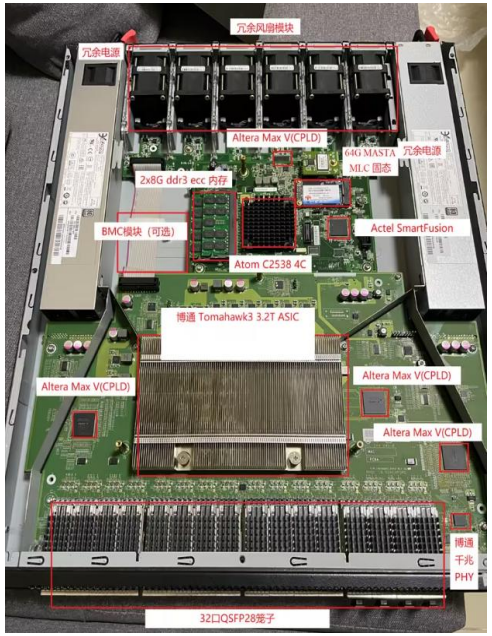


数据中心交换机：硬件国产化进度仍需加快

华金证券 Huajin Securities
华发集团旗下企业

- 交换机的主要硬件可分为接口模块、控制模块、交换模块、供电模块几个部分，可分为黑盒交换机和白盒交换机，白盒交换机将网络中的物理硬件和操作系统（NOS）进行解耦，让标准化的硬件配置与不同的软件协议进行匹配
- 交换机芯片市场国产程度较低，主要由国外三大厂商掌控。从2022年国内商用以太网交换芯片市场份额占比来看，前三大厂商共占比97.8%。中国大陆厂商盛科通信销售额排名第四，占比1.6%，在国内厂商中排名第一。

图表77 Edge-core 白盒交换机



图表78 各硬件部位功能拆解

硬件名称	硬件功能
业务接口（Port）	分为普通接入接口和上行汇聚端口。接口数量的多少直接决定了交换机的接入性能，接口类型则直接决定了交换机在网络中的应用位置。
主板（背板）	提供各业务接口和数据转发单元的联系通道。背板吞吐量（bps）也称背板带宽.
主控单板（引擎）	提供设备的管理和控制功能以及数据平面的协议处理功能，负责处理各种通信协议；
交换网板	负责跨接口单板卡之间的数据转发交换，负责各接口板之间报文的交换、分发、调度、控制等功能。
接口单板	提供业务传输的外部物理接口，完成报文接收和发送。承担部分协议处理和交换/路由功能。一般来说就是提供交换机端口的模块。

图表79 交换机芯片厂商

公司	主营芯片	芯片介绍
BROADCOM	StrataXGS产品	StrataXGS有两种系列芯片，Trident系列适用于数据中心、园区及无线交换等场景；Tomahawk系列是基于超大规模云网络，存储网络及HPC等场景考虑的。
	StrataDNX产品	StrataDNX产品线提供了业界商用硅交换机最大的可扩展性和可扩展性，可通过DNX架构元素连接在一起，构建业界领先的具有数百TB带宽的交换机。
CISCO	Silicon One系列	系列芯片有Q100L/200L/201L/202L等，主要应用于盒式和框式交换机，Q100 是业界首款仅单个 ASIC 即可突破 10 Tbps 性能的芯片
	Cloud Scale系列	系列芯片主要应用于数据中心交换机，可以为数据中心定制的3.6TB和6.4TB单芯片交换机提供更好的性能，降低功耗并实现更精细的网络遥测。
centec	CTC系列	系列芯片有CTC8180、7132、2118、5118等，其中CTC8180是盛科面向5G和云计算，推出的第七代网络交换核心芯片。支持2.4T I/O带宽，提供从1000M到400G的全速率端口能力。

传统数据中心架构中，400G交换机两年翻倍增长

◆ 传统数据中心中，据测算，2022年高功率机架约为4.2万个（对应1.6万400G交换机），预计2025年达8万个（对应3.2万个400G交换机）。

依据1、400G交换机应用前置条件是高算力和高功率；目前人工智能的功率范围在20-40KW。

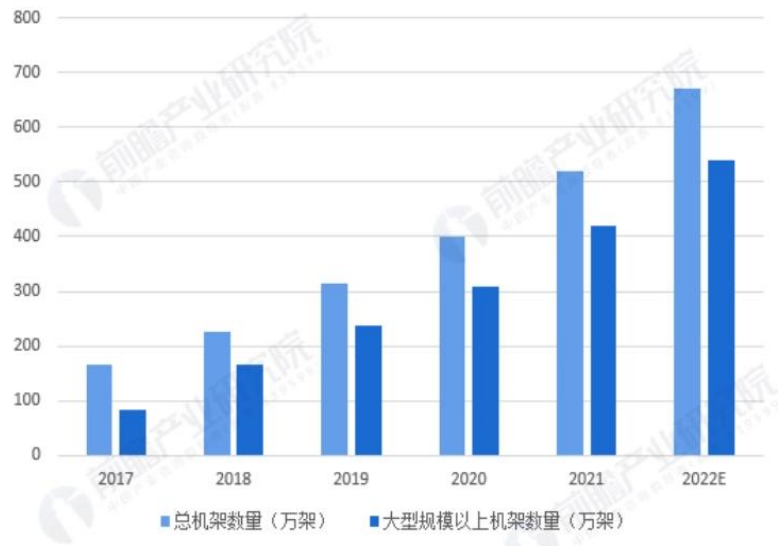
依据2、目前30KW以上占比为1.9%；15-30KW以上占比3.3%；我们预估2022年，20KW-30KW总占比为3.5%。

依据3、2022年我国大型规模以上数据中心新增120万台，年增长率约为30%。

图表80 2021年单机柜功率增长速度

分类	单机柜功率	2020年(亿元)	比例	2021年(亿元)	比例	增长率
低密度数据中心	4KW以下	26.55	10.50%	27.89	10.00%	5.00%
中、低密度数据中心	4<X<6kW	95.07	37.60%	105.42	37.80%	10.90%
	6<X<8kW	65.49	25.90%	71.11	25.50%	8.60%
中、高密度数据中心	8<X<10kW	31.35	12.40%	34.3	12.30%	9.40%
	10<X<15kW	22.5	8.90%	25.66	9.20%	14.00%
高密度数据中心	15<X<30kW	7.84	3.10%	9.2	3.30%	17.40%
	30kW以上	4.05	1.60%	5.3	1.90%	31.00%
总计		252.84	100%	278.88	100%	10.30%

图表81 17-22年中国IDC机架规模（单位：万架）



目录

架构比较：传统数据中心-三层架构基础要素配比

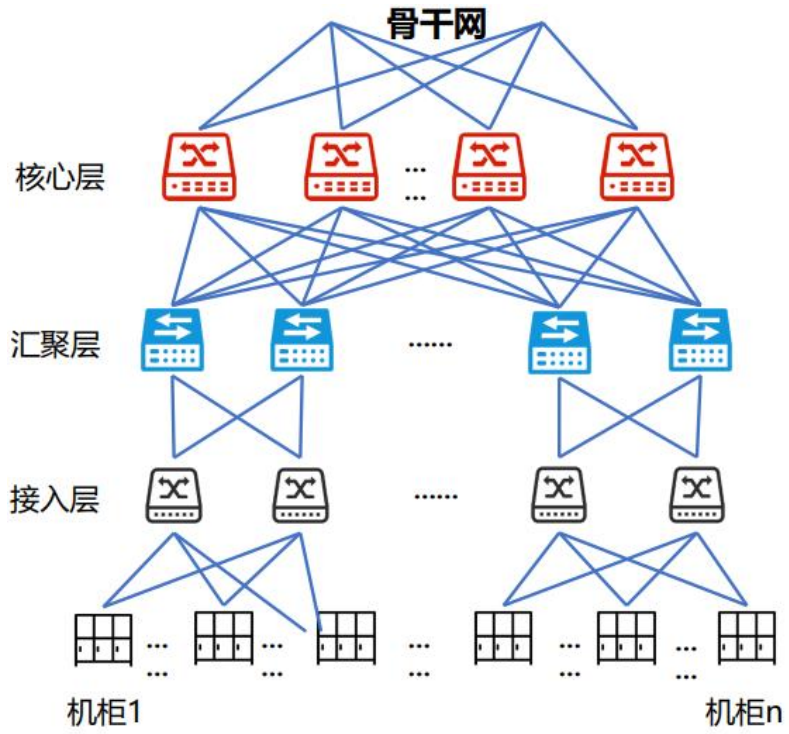
- **机柜数：服务器数：交换机数=1:20:2**（机柜数：250台；服务器数量5000台，交换机数量500台左右）

依据1、传统数据中心，每台机柜的占地面积3-5㎡，此处取均值4进行计算。

依据2、超大型数据中心面积大于2000㎡，大型数据中心为800-2000㎡，此处取1000㎡进行计算。

依据3、按照42U高的标准机柜来存放2U/550W的服务器，可存放20台左右服务器，此处取20进行计算。

图表82 传统三层架构下各层级交换机配比



2台核心交换机

- 每台核心交换机有2路输出。下行与汇聚层交换机相连，每台核心交换机与所有汇聚交换机两两互联。

5台汇聚层交换机

- 每台汇聚层交换机上行与所有核心交换机交叉互联。

500台接入层交换机

- 一般每个机柜顶部放置2台接入层交换机，每台接入交换机上行都有2台汇聚层交换机互联，下行都与每台服务器直连。

架构比较：传统数据中心-新型叶脊架构基础要素配比

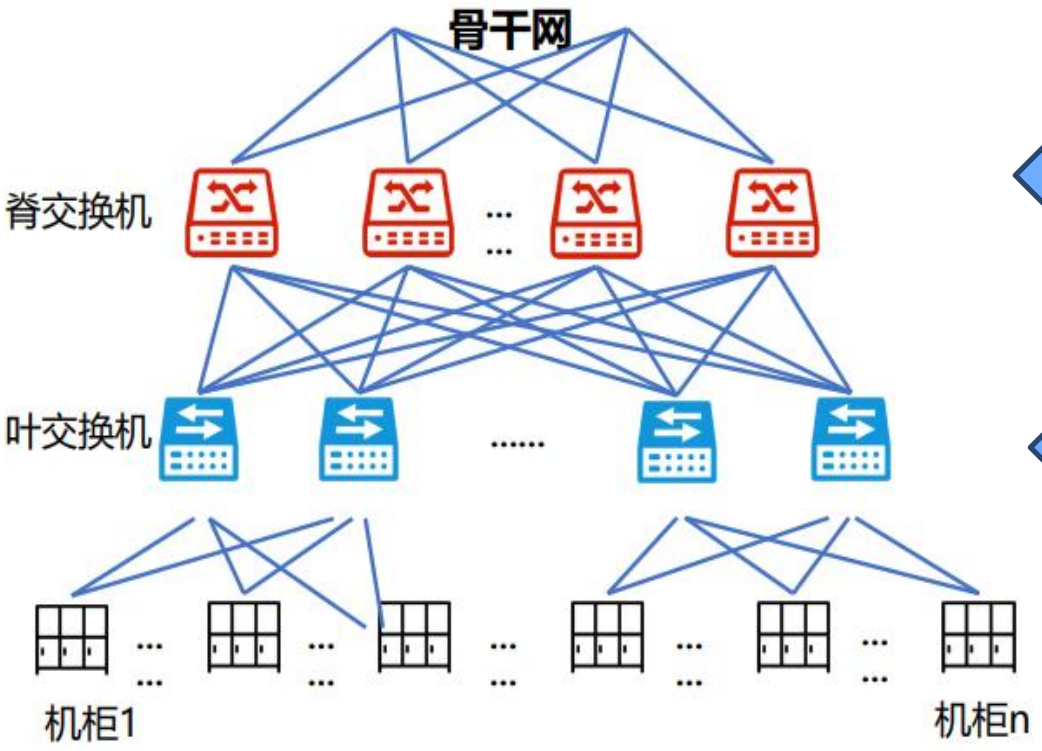
- **机柜数：服务器数：交换机数=1:20:0.4**（机柜数：250台；服务器数量5000台，交换机数量100台左右）

依据1：叶脊架构将传统三层架构变为两层，扁平化处理提升网络效率。

依据2：配比数据计算仍以1000㎡数据中心，250台单机柜，5000台服务器。

依据3：以设计网络时的收敛比不超过3:1的标准来计算，即上行口与下行口的配比数量不超过3:1。

图表83 叶脊架构下各层级交换机配比



4台脊交换机

- 根据收敛比不超过3:1，叶交换机下行带宽48*10G=480G，上行带宽最大为480G/3=160G，则最多可连接160G/40G=4台脊交换机。

105台叶交换机

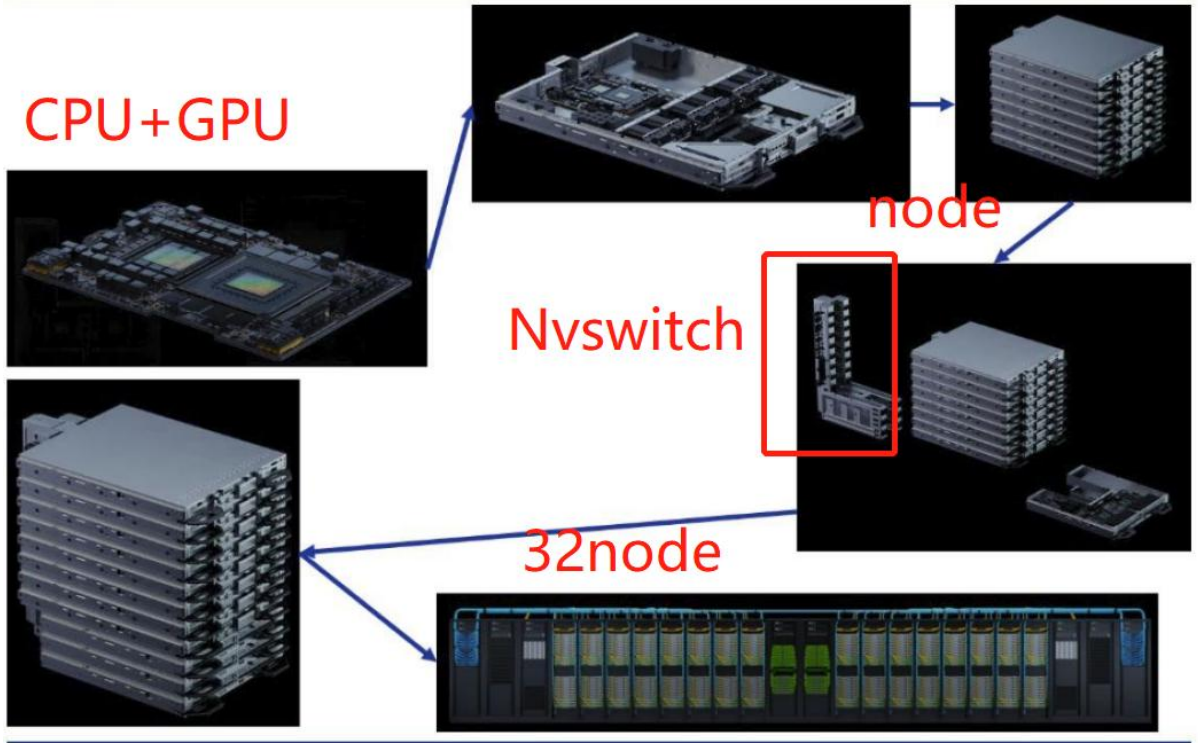
- 以叶交换机48个下行端口计算，最少需要105台叶交换机。其中每台服务器都与两台叶交换机直连。

架构比较： DGX GH200 光模块、服务器、交换机配比

华金证券 Huajin Securities
华发集团旗下企业

- **GH200服务器：**英伟达最新一代AI服务器，集成了256个GH200芯片，拥有144TB共享内存，性能可达1exaFLOPS。
- **光模块：**新架构全光方案刺激高速光模块需求，GPU:800G光模块数量比1:18，400G为1:36。

图表84 DGX GH200 结构图



配比依据：

① 1个node配8个（CPU+GPU）

② 1个Nvswitch配8个node

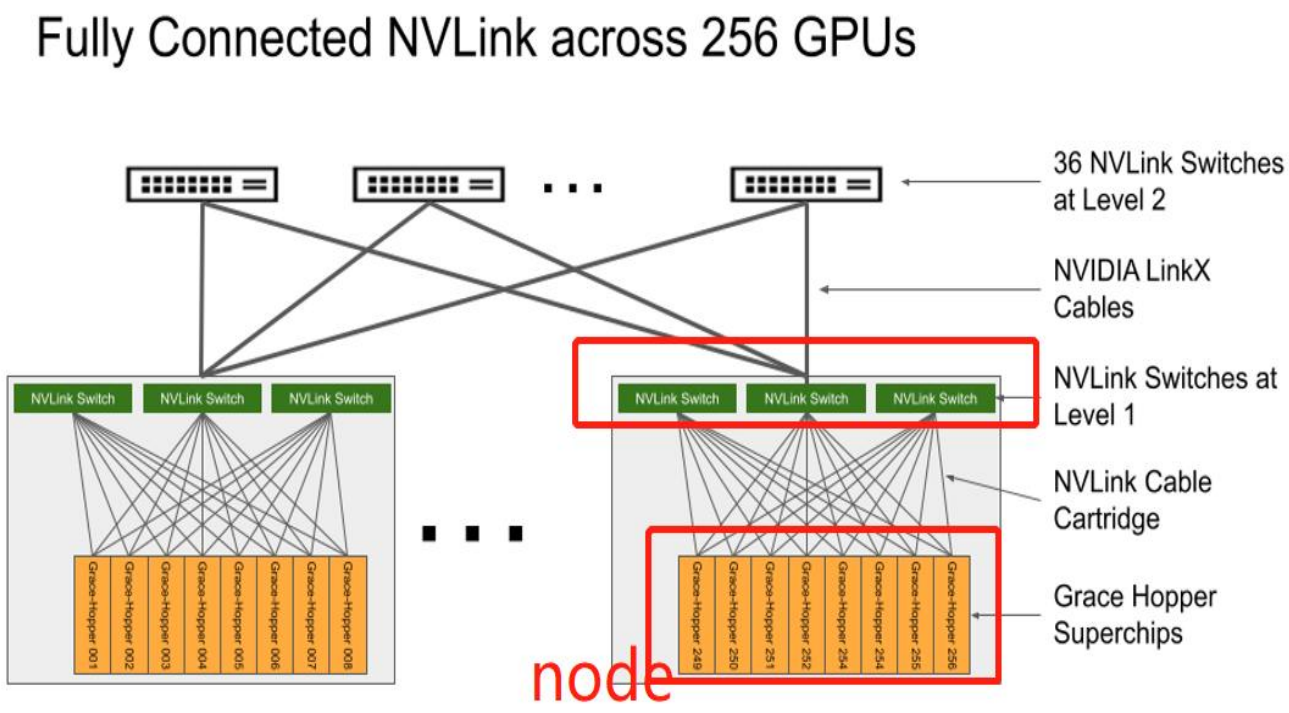
③ 1个GH200配32个node=256个GPU

①	1个GPU与1个Nvswitch传输带宽是900GB/S；单向带宽是450GB/S
②	800Gbps光模块传输速率是100GB/s
③	1个Nvswitch配8个node（8个GPU）=450*8=3.6TB/s
④	3.6TB/s==36个100GB/s==36个800G光模块
⑤	1个Nvswitch共两个上行、两个下行方向==36*4=144个光模块
⑥	全光方案：8个GPU配144个光模块；1个GPU配18个光模块
⑥	256个GPU配2304个光模块

架构比较： DGX GH200 光模块、服务器、交换机配比

- **服务器和交换机：**GH200通过96组L1 NVLink Switch、36组L2 NVLink Switch，互连256个GH200超级芯片，交换机和超级芯片配比为132:256

图表85 DGX GH200网络架构



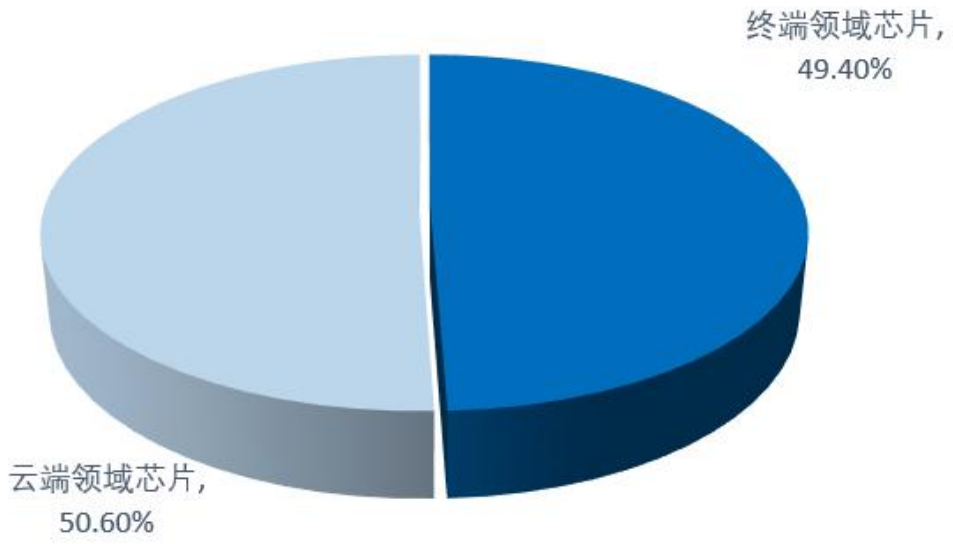
L1层Nvswitch共有96组
① 1个node配8个GPU
② 1个Nvswitch组合配1个node
③ 1个Nvswitch组合中有3个Nvswitch
④ 全组网架构中共有32个Node
⑤ L1层Nvswitch共有96组

L2层Nvswitch共有36组
① 1个Node配1个L2层NVSwitch
② 32个Node配32个L2层NVSwitch
③ 整体结构中4台另有其他功能使用
④ L2层Nvswitch共有36组

算力测算逻辑：2022年训练用算力为95Eflops

- 市场规模：根据测算，2022年中国AI芯片市场中，英伟达GPU模约为150亿；
- 总量：按照10万/片，共计A100芯片15万片A100芯片；
- 算力：按照单台DGX A100算力为5P FLOPS计算，折合1.9万个英伟达A100 AI服务器；可提供95Eflops算力。

图表86 2022年中国AI芯片市场规模细分占比



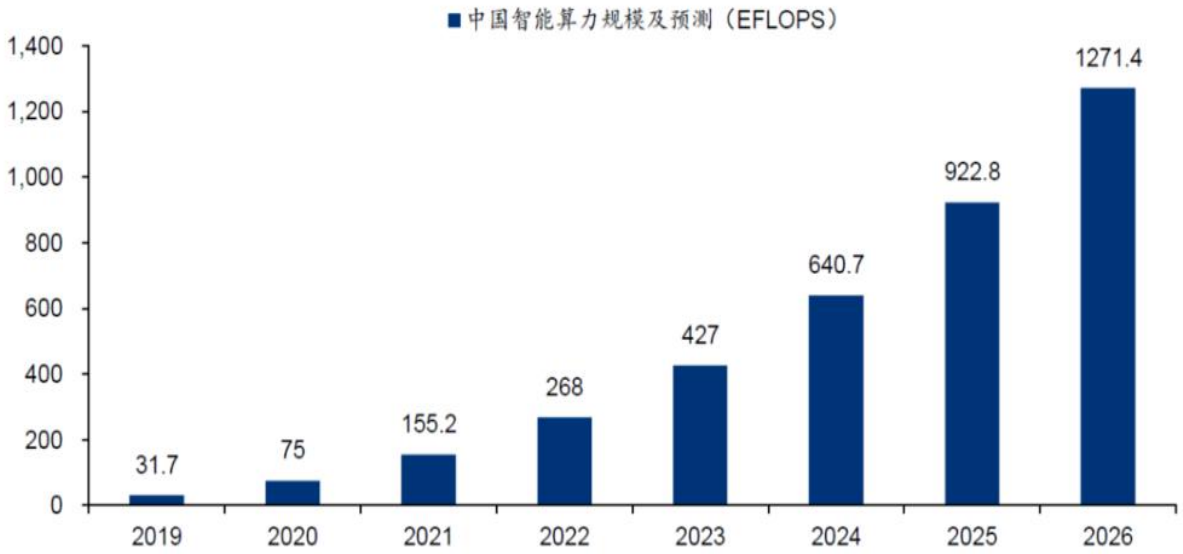
- 依据1：2022年我国AI芯片市场300亿元；从我国AI芯片市场规模细分来看，其中2022年中国云端领域芯片市场规模为152.96亿元，占总市场规模的49.2%；终端领域芯片市场规模为156.54亿元，占总市场规模的50.6%。

- 依据2：假设2022年GPU占比为50%。AI芯片2023年出货量将增长46%。其中英伟达GPU为AI服务器市场搭载主流，市占率约60～70%，其次为云端厂商自主研发的AISC芯片，市占率逾20%。

算力测算逻辑：至2025年训练用算力749E，复合增长43%

- **2023年/2024年/2025年我国训练算力为254E、467.7E、749.8E，每年新增训练算力为159E、213.7E、282.1E**
- 依据假设1：以IDC智算规模为依据；
- 依据假设2：英伟达AI服务器主做训练应用，至2025年，我国新增智算算力均为训练用算力；
- 依据假设3：训练算力需采用智算架构、配备高速率光模块和高速率交换机

图表87 中国智能算力规模及预测



图表88 训练算力测算

	2022	2023E	2024E	2025E
训练算力(EFLOPS)	95	254	467.7	749.8
其中新增训练算力(EFLOPS)	-	159	213.7	282.1
其他用途算力(EFLOPS)	173	173	173	173
合计	268	427	640.7	922.8

资料来源：IDC 、华金证券研究所整理

算力紧缺背景下，算力租赁景气度提升

- 按照《中国算力白皮书(2022年)》的定义，算力主要分为四部分:通用算力、智能算力、超算算力、边缘算力。
- 通用算力以CPU芯片输出的计算能力为主；智能算力以GPU、FPGA、AI芯片等输出的人工智能计算能力为主；超算算力以超级计算机输出的计算能力为主；边缘算力主要是以就近为用户提供实时计算能力为主，是前三种的组合，边缘算力只是解决网络延迟的问题，不算一种新的算力。
- 各地都在加速建设数据中心，打造关键基础设施算力网络，全国的算力规模会保持高速增长。

图表89 算力分类

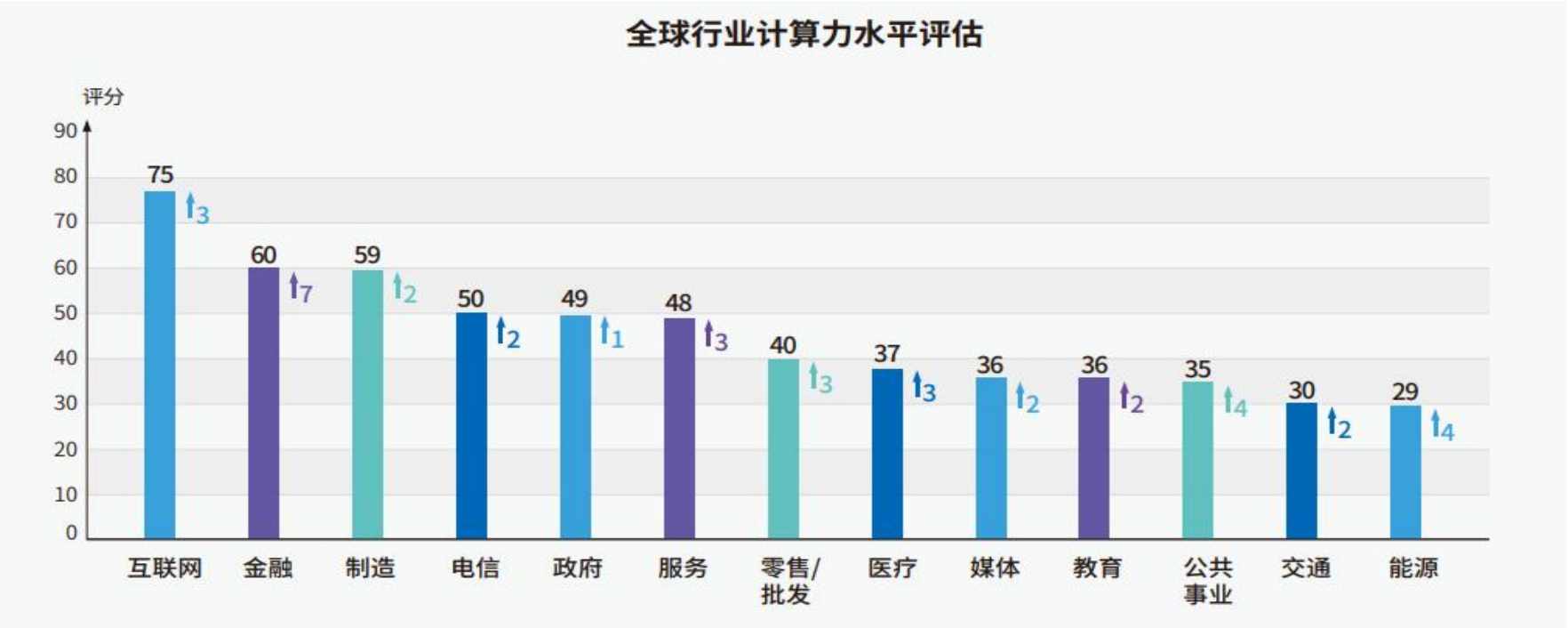
通用算力	基于CPU芯片的服务器所提供的计算能力，主要用于基础通用计算和对精度要求不高的数字化场景。
智能算力	基于AI芯片的加速计算平台，提供人工智能训练和推理的计算能力，主要应用于AI模型的训练和推理计算，比如语音、图像和视力频的处
超算算力	基于超级计算机等高性能计算集群所提供的计算能力，主要用于天体物理、气象研究、航天航空等高精尖科研领域。

图表90 全国主要算力节点

算力紧缺背景下，智能算力租赁景气度提升

- 算力租赁：对算力进行出租，是一种通过云计算服务提供商租用计算资源的模式。用户可以根据自己的需求租赁服务器或虚拟机实现大规模的计算任务，无需拥有自己的计算资源。算力租赁是一种灵活、高效、成本低廉的计算服务，适用于各种大规模计算需求的场景。
- 优势包括：无需投入大量资金购买计算设备、高效稳定的计算服务、灵活的扩容或缩减，更好地满足用户的需求、非常灵活的计费方式，可以根据实际使用情况进行计费。

图表91　全球行业计算水平评估

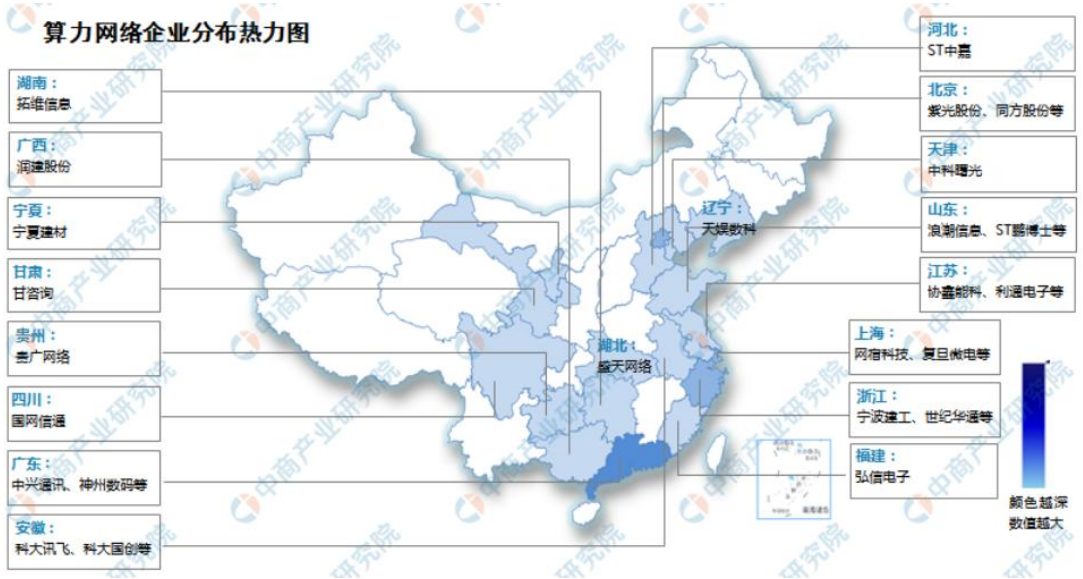


算力租赁市场百花齐放

华金证券 Huajin Securities
华发集团旗下企业

- 算力租赁或成中小企业 AI 算力解决方案。AI 服务器成本每台可达百万元人民币。由于大模型参数量巨大，对应的训练和推理过程都需要消耗大量算力资源，成本高昂，仅有资金实力雄厚、算力资源储备丰富的巨头可以承担。
- 算力租赁对象：科技公司、科研高校、互联网公司
- 算力租赁优势：无需投入大量资金购买计算设备；短时间缩短研发周期和成本；灵活扩容和保障

图表92　算力网络企业分布热力图



图表93 算力租赁参与方

第三方机构	亚康股份、南凌科技、英博数科（鸿博股份全资子公司）、朗源股份、利通电子、宝腾互联（中青宝旗下）、铜牛信息、顺网科技、世纪华通等。
服务器厂商	中科曙光、浪潮信息、新华三（紫光股份旗下）、工业富联、拓维信息等。

已开展算力租赁的公司分析：算力互联

➢ 公司是在中国科学院科技算力基础设施建设运营实践基础上由科技计算产业链相关企业联合发起设立的高新技术企业，主营业务包括以HPC、AI、云原生为主的科技算力基础设施运营与多云互联科技计算云服务、面向科技和产业应用的算力服务通信基础设施运营与算力互联网。

图表94 服务内容

- **技术服务：**计算服务自营+互联的HPC、AI、云原生大规模算力资源池，支持K8S、Slurm双调度系统弹性共享和资源独占。
- **数据服务：**支持文件和对象的高性能在线存储服务和备份存储，互联魔方百TB级存储单元提供离线数据迁移服务。
- **软件工具：**自主研发的面向多云环境统一运维平台，应用软件工具集成和面向三方平台的开放API支持。
- **数据中心：**核心城市与资源型地区A级数据中心标准托管服务，蒸发冷却模块化装配绿色数据中心解决方案PUE1.1。

图表95 算力互联主要租赁节点介绍

资源类别	资源名称	标准单价（元/节点天）
HPC计算节点	I-Sky40c-192	96
	I-Sky36c-256	103.68
	I-Sky48c-256	115.2
	I-Sky56c-384	161.28
	I-Sky56c-768	201.6
	I-Cas32c-256	115.2
AI/图形计算节点	I-H800-80N	3166.67
	I-A800-80N	2933.34
	I-A100-40D	2496
	N-V100-32N	864
Cloud主机	云主机H-1C4G	3
	云主机H-8C16G	20
	容器实例I-2C8G	4
	容器实例I-4C16G	10
	裸金属I-4C16G	12
	裸金属I-20C128G	50
	裸金属I-40C192G	100

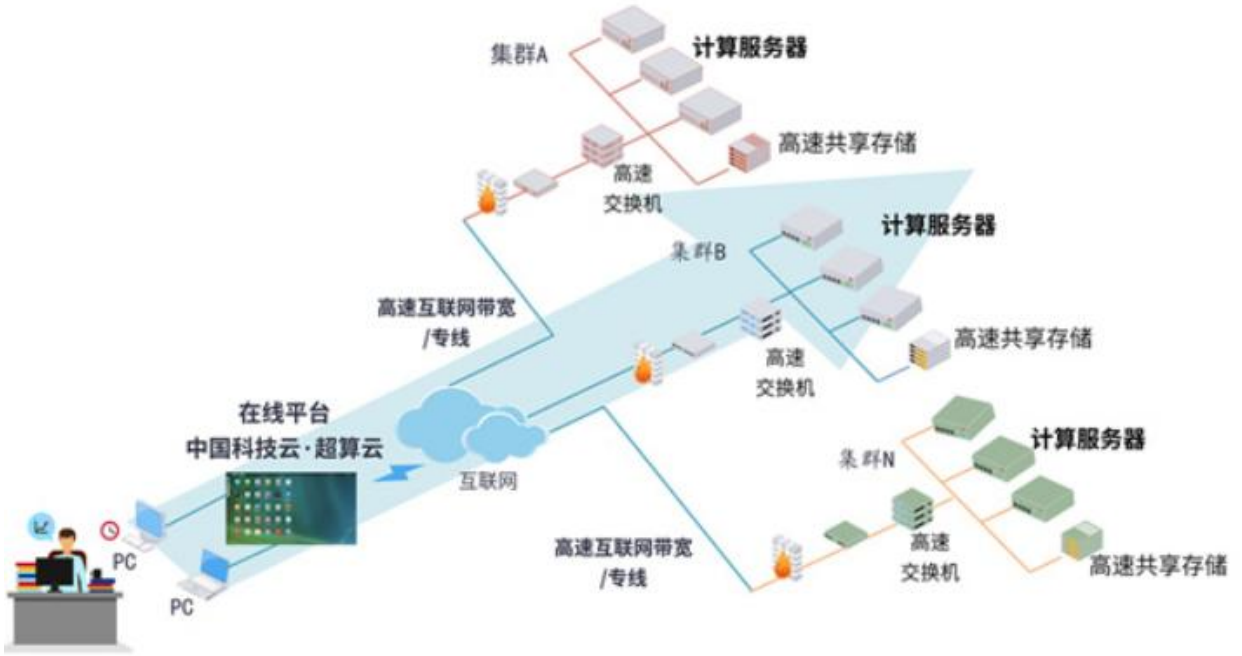
已开展算力租赁的公司分析：北京超级云计算中心

- 公司成立于2011年11月1日，由中国科学院和北京市政府共建，依托中科院计算机网络信息中心建设，运营主体为北京北龙超级云计算有限责任公司。在2020年中国HPC TOP100榜单中，北京超级云计算中心A分区以Linpack测试性能3.74PFlops，荣获2020 HPC TOP100榜单中国超算TOP3，通用CPU算力第一。

- 服务能力

公司总核心数共50万核，服务用户（机构用户）数超过10万+用户。在持续扩容计算资源的同时，上线计算性能达百PFlops的国产服务器资源，满足大规模并行计算需求。

图表96 超级云计算布局图



上市公司智能算力布局百花齐放

华金证券 Huajin Securities
华 发 集 团 旗 下 企 业

- 目前已有多家上市公司布局智能算力业务，包括恒润股份、中贝通信、利通电子、云赛智联、鸿博股份、润建股份、首都在线、真视通等。其中润建股份宣布开发了具有自主知识产权的“‘曲尺’人工智能开放平台”；“曲尺”平台已经在为管维业务的目标检测、行为检测、视频优化、环境风险、物联网设备状态、网规网优、物体识别等多个场景提供 AI 赋能。

图表97　润建股份‘曲尺’平台功能图

图表98　控股子公司收购芜湖六尺智算科技有限公司100%股权

上海六尺核心团队具有丰富的 AI 智算中心(GPU 算力)建设、运营经验和算力市场资源。团队深耕 GPU 算力多年，与上游 GPU 供应厂商英伟达、新华三等有深度合作关系。

基于上海六尺在算力租赁业务的资源优势，公司与上海六尺共同投资设立上海润六尺，公司持有上海润六尺 51%股权，上海六尺持有上海润六尺 49%股权。根据公司与上海六尺于 2023 年 7 月 28 日签署的《江阴市恒润重工股份有限公司与上海六尺科技集团有限公司签署的合资经营合同》(以下简称“《合资经营合同》”)约定，对于上海六尺及其关联方在《合资经营合同》签订之日前，自营或与第三方合作经营的算力租赁业务，将根据上海润六尺的发展需求，双方协商一致按照合理估值择机优先并入上海润六尺。

算力租赁业务将成为公司第二增长曲线

- 出资布局算力市场，合资建设智算中心，部分公司算力租赁已有合作对象：
- 中贝通信与中国联通青海分公司签订3.4亿算力服务框架协议，协议包括服务总算力960P、服务费用12万/P/年等；
- 润建股份规划建设总规模为 20,000 个平均功率 6KW 的机柜的超大型数据中心“五象云谷”云计算中心总投资约 36 亿元。第一阶段投入资金 2亿元采购行业内顶级算力服务器，打造润建股份智能算力中心，提供最高可达 2533Pops（Int8）定点算力

图表99 中贝通信关于签订算力服务框架协议的公告

合同名称：算力服务框架协议

合同金额：34,560万元（含税）

合作双方：中贝通信集团股份有限公司、中国联通有限公司青海省分公司

合作内容：本协议双方以H800设备为基础搭建算力服务平台，提供960P算力服务，服务费按照含税¥12万元/P/年计算，协议服务期限三年，服务费总金额为含税¥34,560万元；三年服务期满后，双方协商确定后续服务事项。

图表100 润建股份算力规划

可承载算力规模

算力类型	参考模型	浮点算力	定点算力	服务器功耗（插8卡）	数据中心最大承载服务器数量（台）	浮点算力总和	定点算力总和
算力	NVIDIA H800 SXM	67 TFLOPS（FP32）	3958 Tops（Int8）	6.5KW	6000	3,216 PFLOPS(FP32)	189984 Pops(Int8)

- 五象云谷云计算中心一期可提供6000个平均功率6KW数据机柜，总设计电荷容量为36000KW；
- 最大可承载6000台搭载NVIDIA H800 SXM卡算力服务器；
- 可承载最大3,216P单精度浮点算力和189,984P Int8 定点算力输出。

目录

算力下一站：政府算力平台建设

- 政府驱动加速算力落地。近期北上广深政府出台一系列促进算力落地的政策，如北京提出了到2025年基本建成具有全球影响力的人工智能创新策源地；上海提出了2025年数据中心超过18000PFLOPS；深圳与科技部共同布局国家超算深圳中心二期项目。
- 我们认为地方国企背景公司有望成为地方智算中心的建设方和运营方，如云赛智联，母公司为仪电集团，是上海国资系统的数据中心运营商和云服务运营商，中标上海超算中心项目；广电运通，背靠广州国资委，搭建省算力中心，承建广州人工智能公共算力中心以及多个政府部门重点算力工程项目。

图表101 北上广深政府算力平台建设规划梳理

地区	规划
北京	《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）》，全面推动人工智能自主技术体系建设及产业生态发展，到2025年基本建成具有全球影响力的人工智能创新策源地。
上海	上海市经信委、上海市发改委发布《关于推进本市数据中心健康有序发展的实施意见》，到2025年，预期上海市数据中心总规模能力达到28万标准机架左右，平均上架率提升至85%以上；数据中心算力超过14000 PFLOPS（FP32）；。
深圳	《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）》提出的首要任务是强化智能算力集群供给通过积极布局算力基础设施，预计2025年建成国家超算深圳中心二期项目。

图表102 部分政府算力平台梳理

公司	布局
云赛智联	上海市人工智能算力的承建方和运营方，建项目包含智能算力平台、算法平台、专用算法模型平台、算力调度平台、配套基础设施建设和机房改造
鸿博股份	子公司博数科已建成的北京AI算力中心100P，与上海、湖南国资共商建设超大规模算力中心
广电运通	承建广州人工智能公共算力中心以及多个政府部门重点算力工程项目

从招标文件看算力对数据中心增量机会：液冷

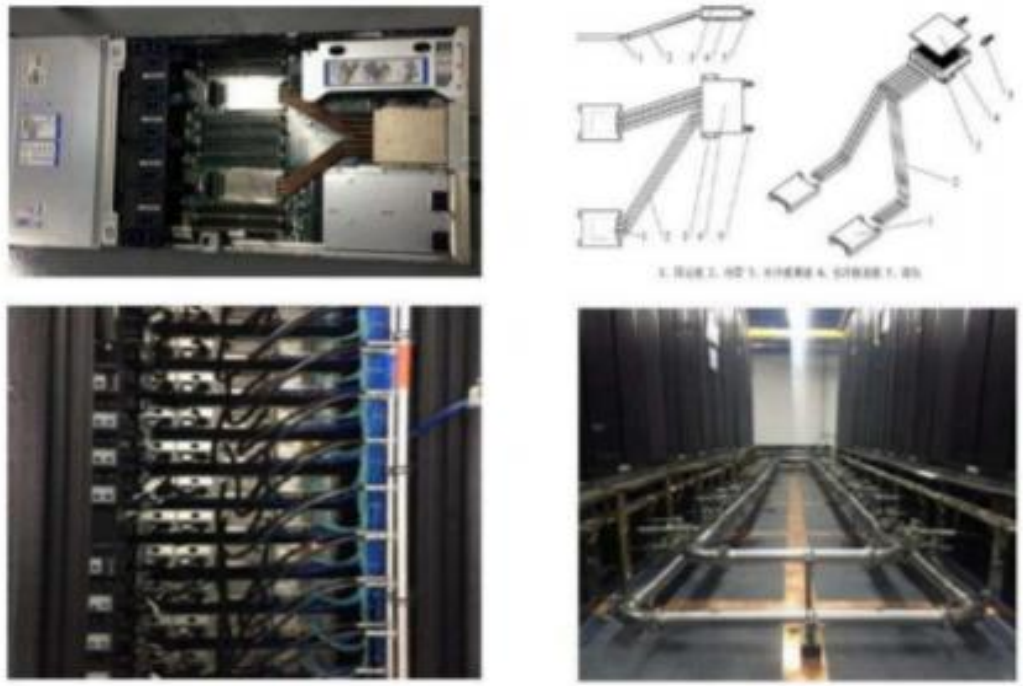
华金证券 Huajin Securities
华发集团旗下企业

- 液冷是指借助高比热容的液体作为热量传输介质满足服务器等IT设备散热需求的一种冷却方式，以云计算为代表的传统数据中心，主要以CPU芯片为主，平均功耗为50-100W；AI芯片普遍功耗高，英伟达A100芯片，功耗达300-400W。
- 目前国内传统数据中心以风冷为主，云计算数据中心、智算中心以冷板式液冷为主，超算中心主要以浸没式液冷为主。冷却力是空气的1,000-3,000倍，相比传统风冷系统约节电30%-50%；2022－2027年，中国液冷服务器市场年复合增长率将达到56.6%，2027年市场规模将达到95亿美元。

图表103 青岛“海之心”人工智能计算中心项目招标文件

项目	技术需求
计算子系统	计算子系统计划提供多种不同类型精度算力支撑，可以根据实际场景需求适配不同的算力类型支持应用，半精度浮点运算峰值总算力≥150PFlops@FP16，单精度浮点运算峰值总算力≥13PFlops@FP32，双精度浮点运算峰值总算力≥2PFlops@FP64。投标人根据智算中心规模自行设计相应规模计算子系统。 计算系统使用先进的绿色节能技术，包括但不限于高效浸没式液冷相变冷却技术或风/液冷混合制冷技术。 计算节点包含国产 X86 处理器、通用人工智能加速卡、国产人工智能加速卡适配。计算子系统方案包含机柜及配套设施。 本项目全部服务器 CPU 全部采用国产 x86 处理器，单颗

图表104 液冷数据中心展示图



资料来源：《电信运营商液冷技术白皮书》、云头条、财联社、CSDN、华金证券研究所

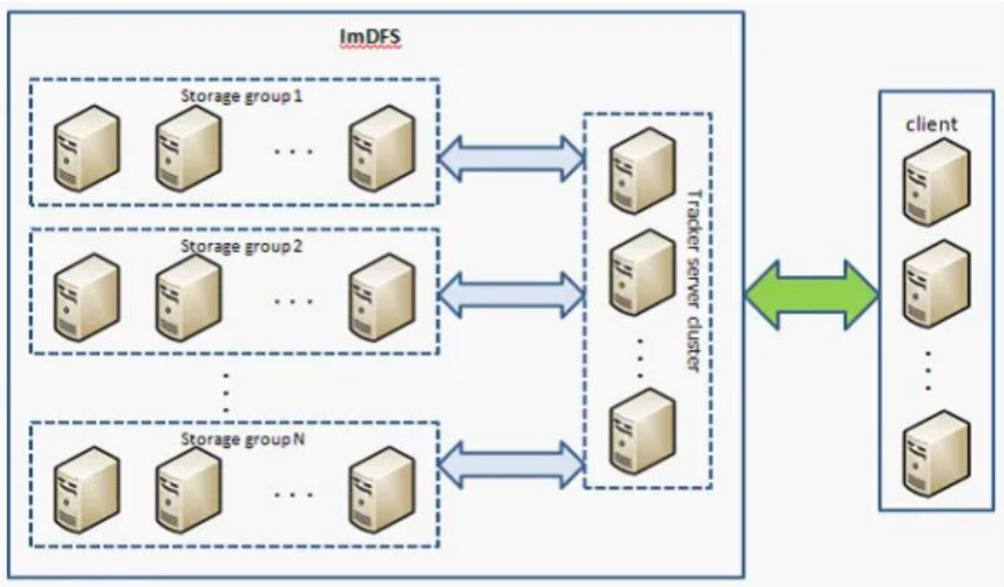
从招标文件看算力对数据中心增量机会：分布式存储

- 新增的超算场景，如气象预报、能源勘探、GIS 等应用的精度带来了数据量的快速增长；其中基因测序等 80% 以上是 PB 级的数据密集型场景，部分业务单文件数量达到 TB 级别，需要超算存储可以提供更大的带宽、更高的 IOPS、支持超大算力的访问能力。
- 分布式存储是指基于分布式架构，通过软硬件协同，依托高效网络连接多个节点来实现存储功能的IT产品和服务。用户提供包括分布式文件、分布式块、分布式对象、混闪和全闪系列集 中式存储等产品及一体化解决方案

图表105 青岛“海之心”人工智能计算中心项目招标文件

2	存储子系统	投标人根据智算中心规模自行设计相应规模存储子系统。存储系统采用分布式并行存储系统，支持多副本、N+M 纠删码等数据保护技术、全冗余设计，支持单一存储命名空间、支持容量海量扩展，性能线性扩展，能够满足智能计算中心海量文件并发读写需求。提供分布式文件/对象/块/备份等存储资源，负责智算、科学/工程计算相关数据的输入、计算结果等的存储与备份。 分布式文件/对象混闪存储系统存储裸容量≥25PB。 分布式块全闪存储系统储裸容量≥1200TB。

图表106 常见分布式存储示意图



从招标文件看算力对数据中心增量机会：智算可视化

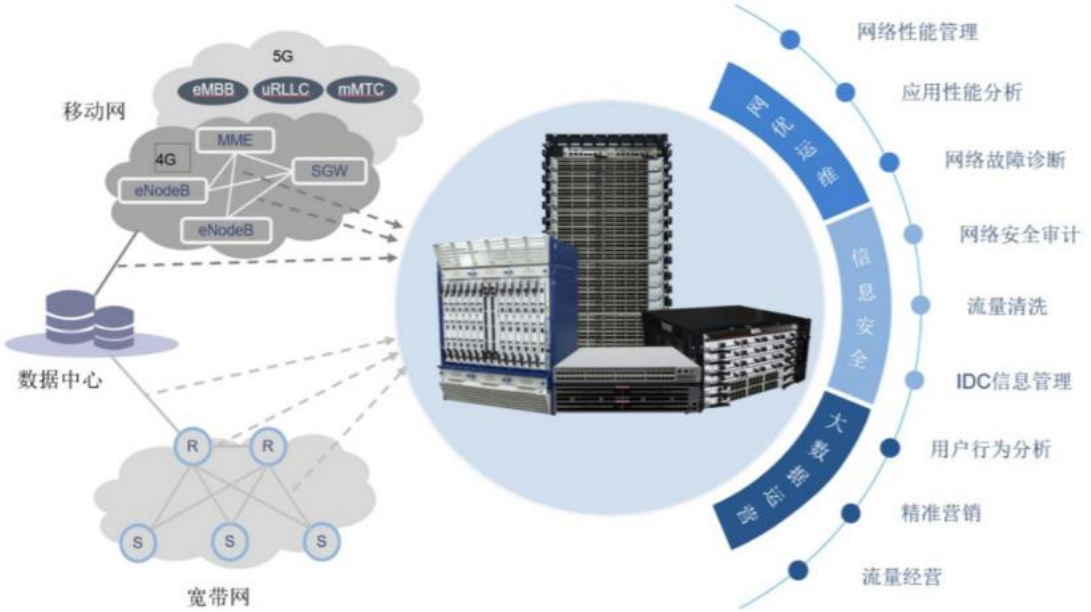
华金证券 Huajin Securities
华发集团旗下企业

- AI应用的升级和优化，需要海量的网络数据，网络可视化是网络数据的源头，网络可视化主要功能是对网络数据进行监测、管理。以网络流量的采集与深度检测为基本手段，实现网络管理、信息安全与商业智能的一类应用系统。
- 恒为科技与中贝通信技术团队已围绕高速光模块在智算可视化上的应用开展深度合作，恒为科技的网络可视化产品主要部署在运营商宽带骨干网、移动网、IDC出口、以及企业和行业内部网络等不同场景，在其主要网络节点通过多种物理链路信号采集技术，进行全流量数据采集。

图表107 青岛“海之心”人工智能计算中心项目招标文件

3	网络子系统	投标人根据智算中心规模自行设计相应规模网络子系统。网络系统主要为节点之间的管理控制信息以及业务数据信息提供高速、安全、可靠的通信通道，主要包含高速训练网络系统、高速推理网络系统、业务网络、存储网络、管理网络和监控网络等。 智算系统训练网络采用 200Gb/s 高速网络，用作并行计算程序的计算网络。 高速推理网络系统采用 25Gb/s 以太高速网络组网。 业务网络采用万兆以太网作为业务网。 AI 集群内部存储网络采用 200Gb/s IB 网络和 25Gb/s 以太高速网络作为存储网络。

图表108 恒为科技网络可视化应用系统



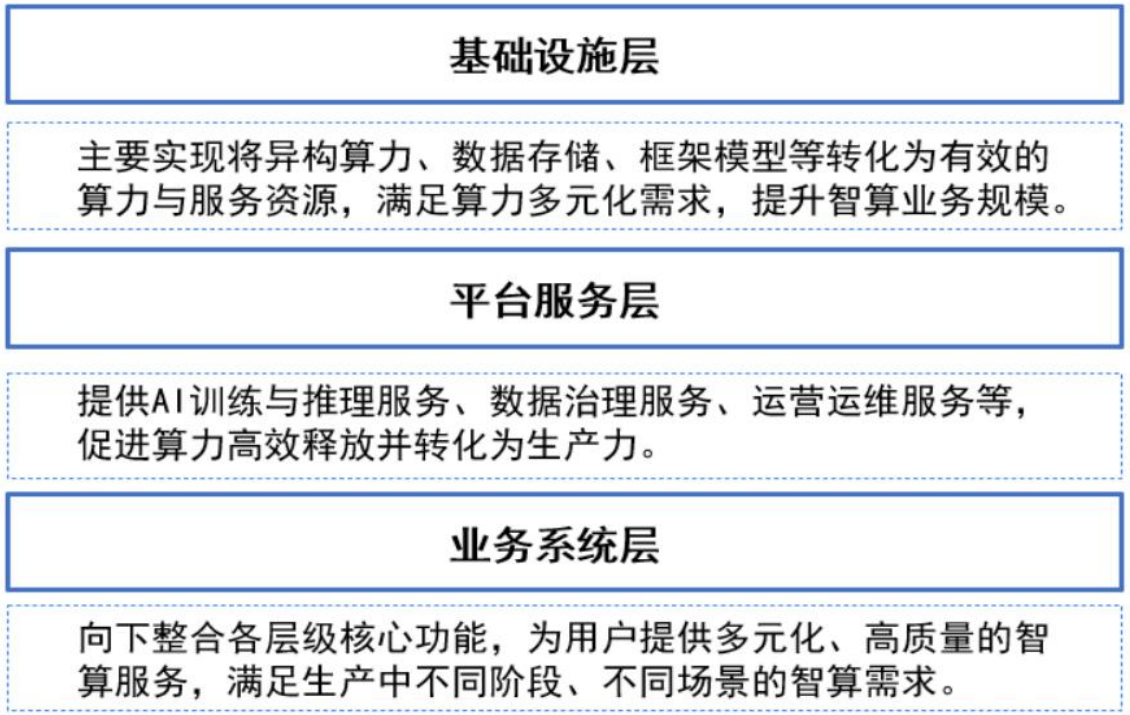
从招标文件看算力对数据中心增量机会：智算OS

- 智算中心操作系统，是以智算服务为对象，对智算中心基础设施资源池进行高效管理和智能调度的产品方案，分为基础设施层、平台服务层、业务系统层，提升资源调度效率，降低算力使用门槛。
- 2020年，浪潮信息推出浪潮云海OS，并成功完成全球最大规模单一集群达1000节点的云数智融合实践。3天完成1000台服务器部署 ，在单一集群中融合支撑了海量大数据处理业务及大规模云原生业务。

图表109 青岛“海之心”人工智能计算中心项目招标文件

人工智能服务平台	投标人根据智算中心规模自行设计相应规模的人工智能服务平添，提供数据集管理、模型管理、模型训练、超参调优、推理等验证、部署上线的全流程人工智能开发应用支持能力。
全栈云平台	投标人根据智算中心规模自行设计相应规模全栈云平台，通过对存储、网络、计算等资源虚拟化，为用户提供动态可扩展、高可靠性、高性价比资源池，实现按需部署、灵活可控的云端计算。
智算产业情报分析系统	投标人根据智算中心规模自行设计相应规模智算产业情报分析系统，提供产业全景分析、产业研判、产业地图监测、产业政策监测及研报，产业舆情监测等功能，打造智慧化、可视化、场景化的产业情报平台，推动优质项目与产业链资源精准匹配。

图表110 智算OS架构



从招标文件看算力对数据中心增量机会：调度平台

- 主要分为算力服务虚拟化和算力服务协同调度。算力服务虚拟化，弱化底层算力芯片供给的技术差异性，为用户提供标准化的算力供给服务。对底层计算能力的GPU、FPGA、ASIC等AI芯片进行统一管理和调度，以PFLOPS、EFLOPS作为计算能力单位向用户提供算力服务；算力服务协同调度，作为算力基础单元，通过云服务方式融入全国算力调度体系中。
- 目前我国首个实现多元异构算力调度的全国性平台——全国一体化算力算网调度平台（1.0版）正式宣发；北京提出建设统一的多云算力调度平台，上海市依托人工智能公共算力服务平台实现算力调度，探索建设算力交易中心。

图表111 青岛“海之心”人工智能计算中心项目招标文件

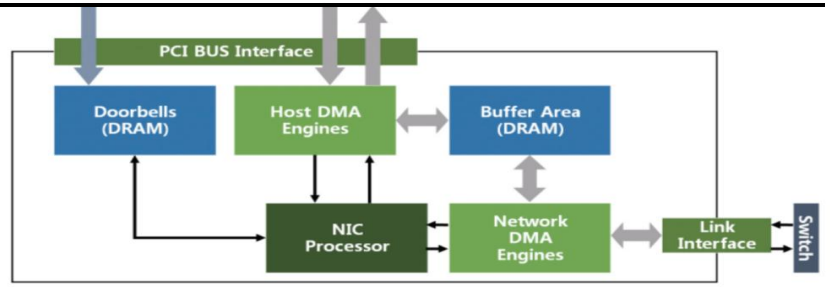
序号	项目	技术需求
1	计算服务平台	投标人根据智算中心规模自行设计相应规模计算服务平台，为整个智算中心提供用户管理、权限管理、计量计费等运营和在线运维能力。
2	集群融合调度系统	投标人根据智算中心规模自行设计相应规模集群融合调度系统，提供面向 HPC 作业、AI 算子的编排和调度等提供统一的调度引擎。
3	集群资源	投标人根据智算中心规模自行设计相应规模的集群资源

图表112 全国一体化算力算网调度平台

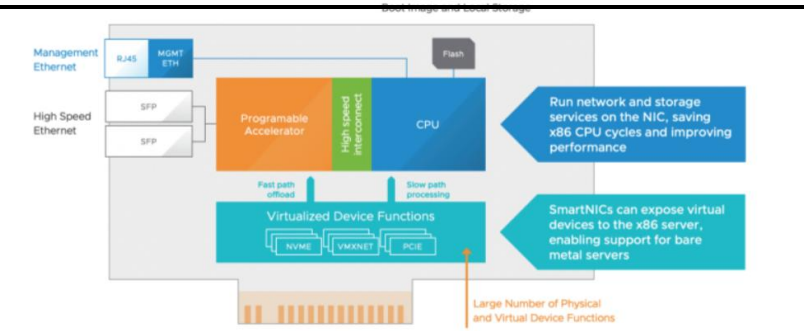
未来趋势一：智能网卡崛起，需求越发迫切

- 基础功能网卡只能提供基本的网络功能，而智能网卡具备计算能力，拥有灵活可编程的功能。
- 智能网卡可处理数据中心节点间、节点内和网络系统三方面问题。智能网卡可以提升服务器数据交换效率，加强数据传输可靠性，提高数据中心模型执行效率、I/O切换效率，完善服务器架构，保障网络系统安全性等作用。
- 近年来数据中心流量以每年25%速度增长。CPU计算能力增速远远低于网络传输速率增速，且差距持续增大。对网络功能卸载到可编程软件的需求，即智能网卡的需求越来越迫切。
- 智能网卡复合年均增长率可达20%。2020年智能网卡市场规模为6亿美元，预计2026年市场规模将达到16亿美元。

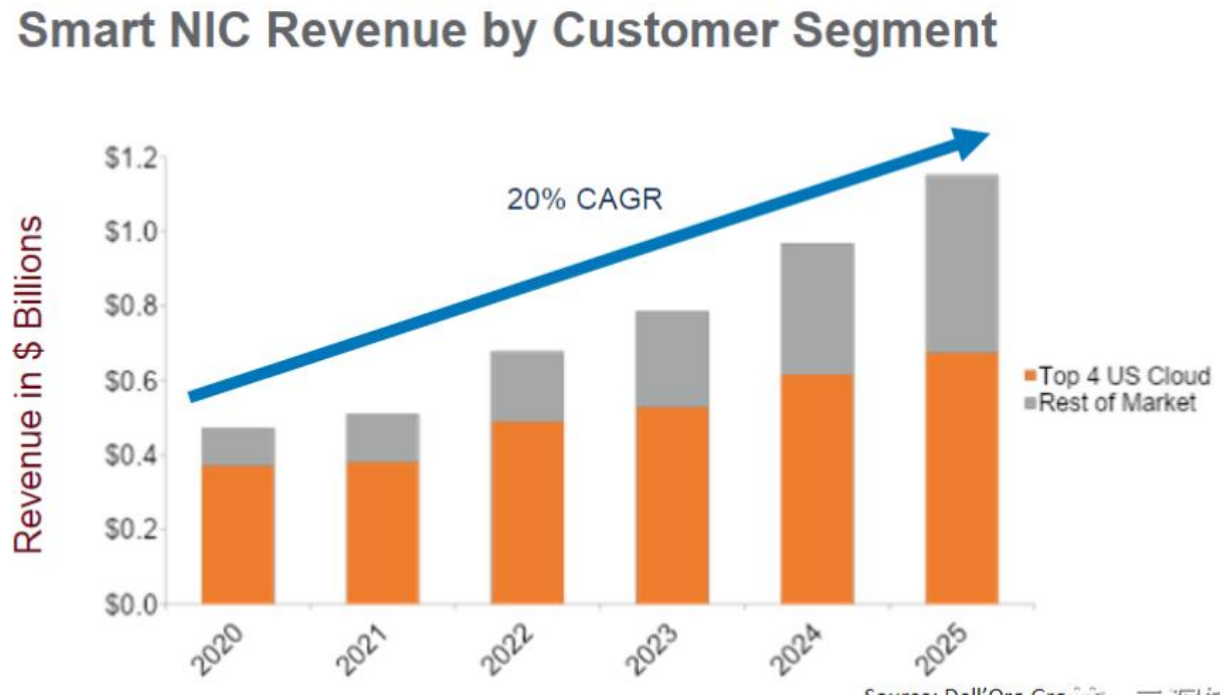
图表113 普通网卡技术架构图



图表114 智能网卡技术架构图



图表115 智能网卡在数据中心的作用



未来趋势一：智能网卡——第二代DPU成为发展新方向

- 目前智能网卡分为第一代SNIC智能网卡和第二代DPU智能网卡，第二代加入了CPU进一步提高性能。DPU智能网卡作为服务器硬件基础核心单元与应用支撑，将作为解决算力性能瓶颈和保障网络安全的重要承载。
- 全球厂家积极布局智能网卡。英伟达ConnectX以太网和IB网系列产品，最高可支持400G网络。Intel坚持ASIC + FPGA实现形式，推出IPU系列产品，支持快速可编程数据包处理引擎和开源开发工具包IPDK。阿里云推出CIPU产品，构建大规模弹性RDMA高性能网络，时延最低可达5us。锐捷网络湛卢系列，通过FPGA+SOC支持转发和控制功能的网络全卸载。

图表116 智能网卡各形式的主要厂家及技术壁垒

SNIC实现形式	NP智能网卡	FPGA智能网卡	ASIC智能网卡
SNIC主要厂家	FUNGIBLE HUAWEI	XILINX	intel BROADCOM NVIDIA
DPU实现形式	CPU	FPGA+CPU	ASIC+CPU
DPU主要厂家	aws MARVELL	Ruijie锐捷 阿里云	NVIDIA
技术壁垒及特点	·性能瓶颈明显 ·开发生态封闭	·成本功耗高 ·自研投入高	·性能优势明显 ·量产成本低 ·特性迭代慢

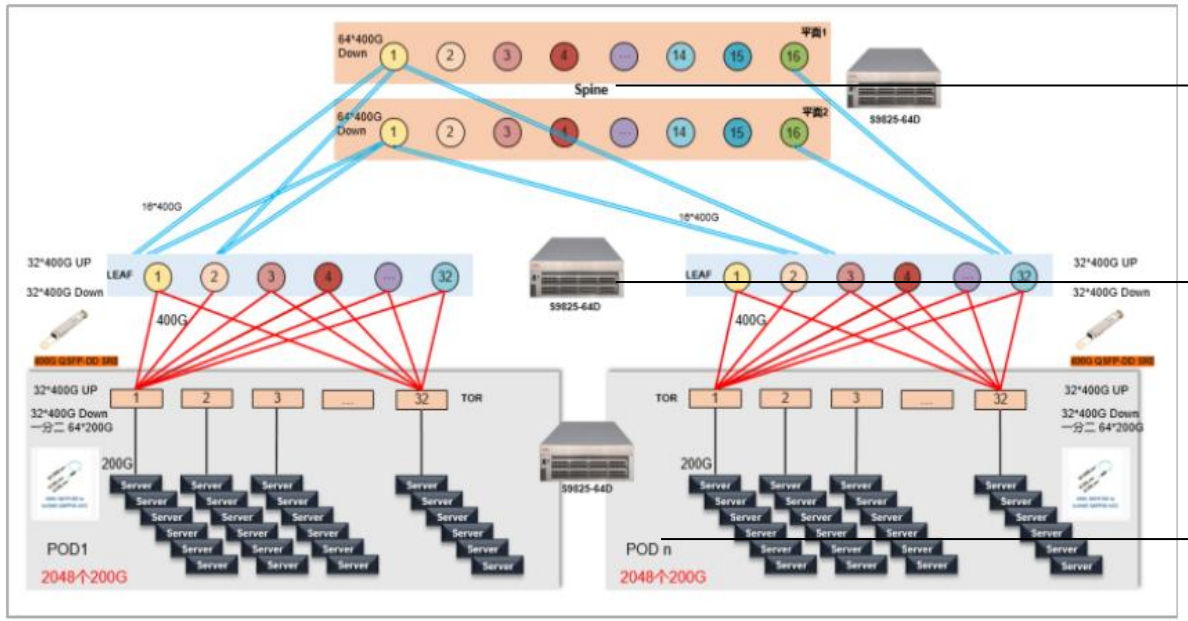
图表117 英伟达ConnectX-7网卡

- 英伟达ConnectX-7 支持超低时延、400Gb/s 吞吐量，支持最高4端口RDMA，拥有基于硬件的拥塞控制，使用NVMe over Fabrics 网络功能卸载。

未来趋势二：无损网络——下一代数据中心网络

- 网络连接诉求加强。ChatGPT的总算力消耗约为3640PF-days，算力500P的数据中心才能支撑运行，GPU之间的通信互联带宽可达400GB/s，0丢包、低时延、高吞吐成为下一代数据中心网络的三个核心诉求。
- 无损网络通过网络协议、分布式架构、上下行非对称设计、流控技术、拥塞控制、流量调度等方面的解决来实现。目前常用的网络协议有RDMA、InfifiniBand，其中IB网络架构封闭，兼容现网差。华为、新华三、腾讯、星网锐捷都推出了无损网络架构。

图表118 新华三智能无损网络盒盒组网框架图

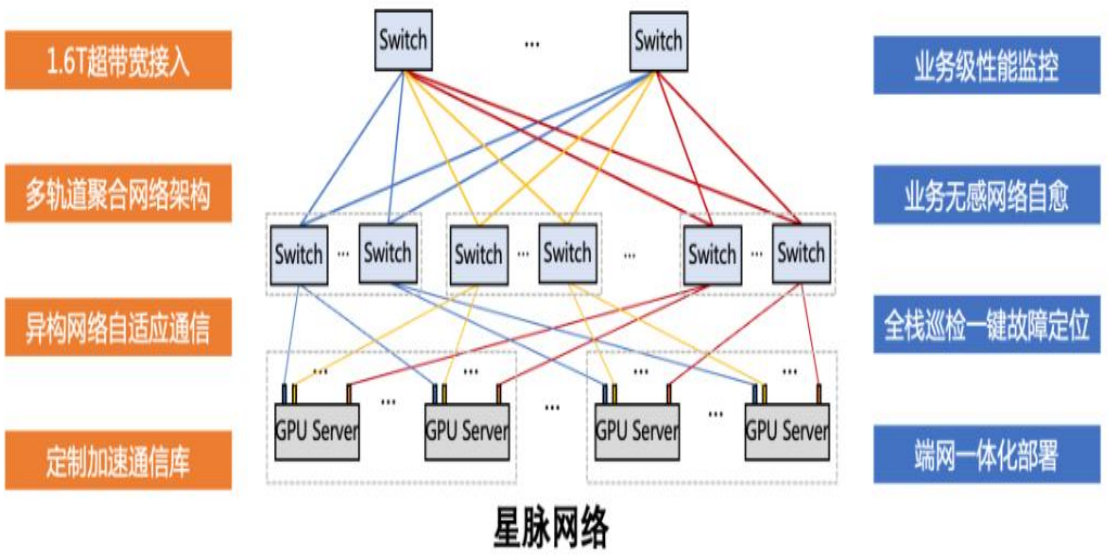


- Spine采用双平面并且保证ToR上下行收敛比满足1：1要求。
- 采用最新的S9825/S9855等系列三层ToR/Leaf/Spine组网架构。
- 服务器接入速率400G下单PoD可以支持1024台服务器，集群最大支持32个PoD理论可满足6.5万台服务器规模接入；若采用100G速率，最大可超过十万台服务器规模接入。

未来趋势二：无损网络——腾讯五位一体新一代计算集群

- 腾讯云发布新一代HCC高性能计算集群。HCC基于自研网络及存储架构，具备TB级吞吐能力和千万级IOPS，可实现3.2T通信带宽，训练时间可缩短至4天。
- 新一代HCC通过三方面解决大模型算力缺乏问题。主要通过计算层面的单机算力提升，网络层面的网络架构优化以及存储层面的存储性能加强。
- 星脉超算网络架构实现网络硬件平台全自研，打造“胖树组网拓扑+3.2T超带宽+异构网络+多轨道流量聚合+定制加速通信库”五位一体化设计。

图表119 腾讯云星脉网络架构图



图表120 腾讯云新一代HCC关键要素

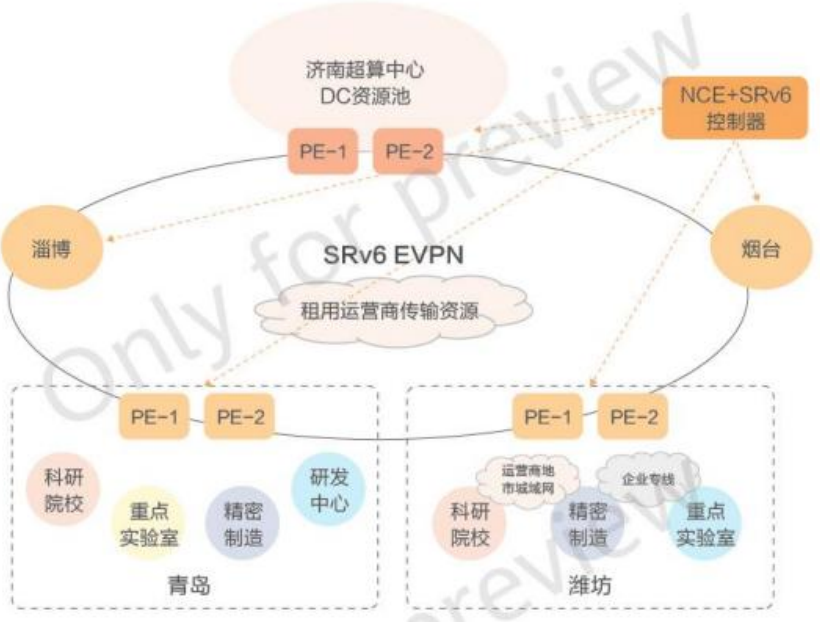
关键要素	解决方案	方案特点
计算层面	自研星星海服务器	星星海具备超高密度+一体化设计提升单点算力性能，搭载英伟达H800 Tensor Core GPU，支持输出最高1979 TFlops的算力。采用6U超高密度设计，每节点支持8块H800，相较行业可支持的上架密度提高30%；
应用场景	星脉高性能计算网络	星脉打造业界最高3.2T RDMA通信带宽。集群整体算力提升20%。自研高性能集合通信库TCCL，能够为大模型训练优化40%负载性能，消除多个网络原因导致的训练中断问题。
存储层面	腾讯云最新自研存储架构	最新自研的存储架构具备TB级吞吐能力和千万级IOPS，COS+GooseFS对象存储方案提供基于对象存储的多层缓存加速，CFS Turbo多级文件存储方案充分满足大模型场景下大数据量、高带宽、低延时的存储要求。

未来趋势三：DCI算力互联——算力增长需求解决方向

- 数据中心互联需求倍增。数据中心建设规模、IP流量及算力呈现出直线增长。随着国内互联网大厂相继布局GPT类大模型，预训练+持续调优+应用推理带来大量算力需求，对全社会算力资源提出更高要求。
- 超算业务面临用户接入难、算力变现难、算力资源使用不均匀现象。国家超级计算中心目前共有9所，分别为：天津中心、广州中心、深圳中心、长沙中心、济南中心、无锡中心、郑州中心、昆山中心和成都中心；算力互联有望进一步统筹计算资源，提高产能利用率，推动算力产业建设进一步加速。

图表121 算力网络组网框架图

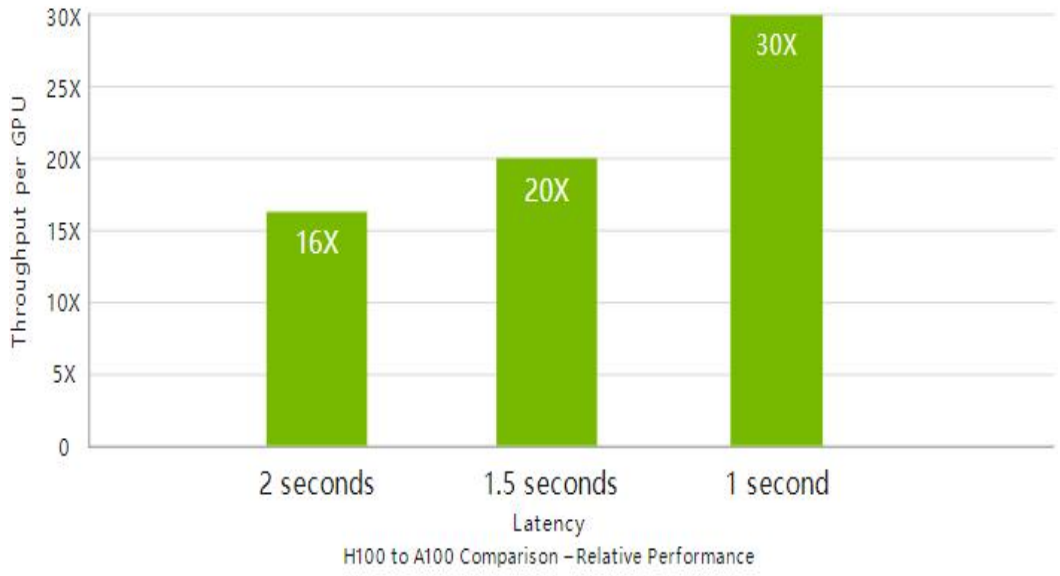
图表122 山东省算力配给网组网图



未来趋势三：DCI算力互联——提供高性能AI算力支持

- 算力互联提供国内顶级的高性能AI算力支持。算力互联启动新一轮自营AI算力资源池大规模集群扩容，包括H800 AI算力服务，采用ACMCube™一体化算力基础设施解决方案，核心计算节点HGX单主板八个H800 80G GPU NVSwitch，使用Nvme高速本地数据盘和全新的PCIE5.0 IB网络，由多种最新技术构建了目前最为强大的AI算力节点。
- 算力互联提供灵活丰富的科技算力服务，全面覆盖下沉市场。公司算力AI/图形计算服务分为标准型和共享型，标准型按每节点每天计算，最低可至28.8元/节点天;共享型按每卡时计算,最低可至0.15元/卡时。

图表123 AI 推理性能提升示意图



图表124 算力互联AI/图形计算节点标准型价格表（部分）

资源名称	节点技术规格	标准单价(元/节点天)
I-H800-80N	8*H800 80G Nvlink/Ice 96c 2.1GHz/2TB DDR5/15.36TB NVME/800Gb	¥ 3,166.67
I-H800-80N	8*H800 80G Nvlink/Ice 80c 2.0GHz/2TB DDR5/15.36TB NVME/800Gb	¥ 3,166.67
I-A800-80N	8*A800 80G Nvlink/Ice 64c 2.6GHz/2TB DDR4/16TB NVME/400Gb	¥ 2,933.34
I-A100-80N	8*A100 80G Nvlink/Ice 72c 2.7GHz/2TB DDR4/16TB NVME/400Gb	¥ 3,168.00
I-A100-40N	8*A100 40G Nvlink/Ice 64c 3.0GHz/1TB DDR4/6.4T NVME/400Gb	¥ 2,496.00
I-A100-40D	8*A100 40G Nvlink DGX/Ice 64c 2.6GHz/1TB DDR4/8T NVME/800Gb	¥ 2,496.00
I-A100-40P	8*A100 40G Clink/Ice 56c 2.6GHz/512GB DDR4/3.84TB NVME/200Gb	¥ 1,382.40
N-V100-32N	8*V100 32G Nvlink/Ice 32c 2.4GHz/512GB DDR4/7.68TB NVME/200Gb	¥ 864.00
N-V100-16	2*Tesla V100 GPU/Dhyana 64c 2.0GHz/192GB DDR4/1TB SATA/10Gb	¥ 144.00

算力方向演变：国内与海外驱动并行

华金证券 Huajin Securities
华发集团旗下企业

上半年

以海外驱动为主；海外驱动大于国内

海外驱动逻辑：
1、算力爆发，景气提升：算力超预期，A100/H100产能紧缺
2、光模块订单：英伟达、谷歌持续超预期下单800G光模块

下半年

以国内驱动为主，国内驱动与海外驱动并行

国内驱动逻辑：
1、基础设施：大模型出现带动AI服务器需求，数据中心不断扩容升级
2、内延：短距离传输、高速背板连接需求大涨
3、算力资源：全国智算中心建设加速，得卡者得生产力

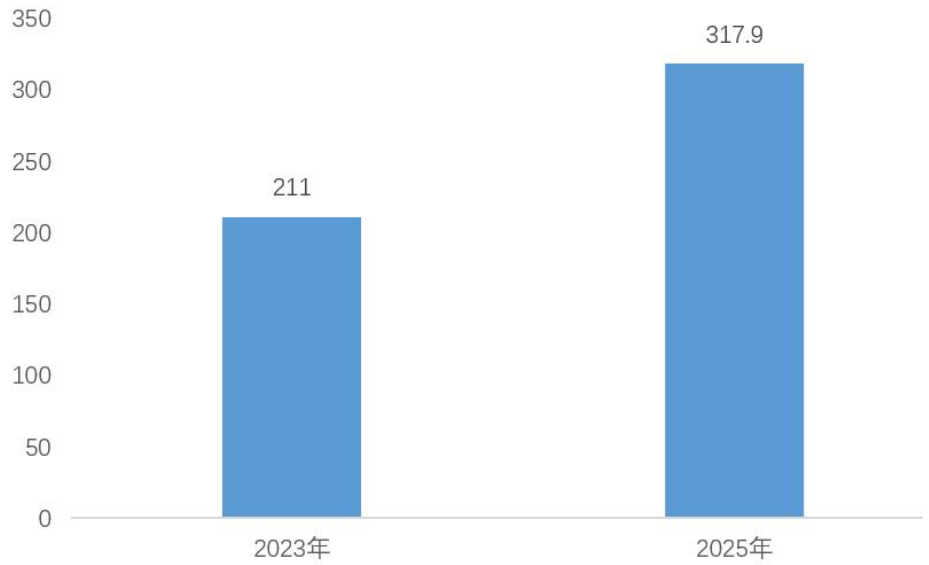
+

海外驱动逻辑：
1、海外厂商：不断订单驱动，仍利好光模块
2、光模块验证：800G光模块验证结果陆续落地

布局方向一：算力基础设施

- AI带动产业增长：据国家信息中心联合浪潮信息日前发布的《智能计算中心创新发展指南》测算，十四五”期间，在智算中心实现80%应用水平的情况下，城市对智算中心的投资可带动人工智能核心产业增长2.9至3.4倍，带动相关产业增长36至42倍。
- 数据中心交换机不断扩容升级：数据流量发展加速，交换机不断扩容升级，以 400GE 为代表的关键技术，不断加速部署，持续放量；国内主流的数据中心交换机端口速率正在由10G/40G向100G/400G升级演进，市场不断推出25.6T、51.2T交换机芯片。

图表125 全球AI服务器市场规模（亿美元）



图表126：新华三800G CPO硅光数据中心交换机

紫光股份——新一代云计算基础设施领先者

- 公司看点：紫光股份是全球新一代云计算基础设施建设和行业智慧应用服务的领先者，提供技术领先的网络、计算、存储、云计算、安全和智能终端等全栈 ICT 基础设施及服务，在国内交换机和服务器领域的市场份额名列前二。
- 业务增速： 2023年上半年营业收入360.45亿元，同比增长4.78%；净利润18.03亿元，同比增长7.14%。公司加快“云-网-算-存-端”布局和数智创新，主动把握新一轮技术发展方向，在“AI in ALL”技术战略指引下，推出多款服务器、交换机等数字基础设施。

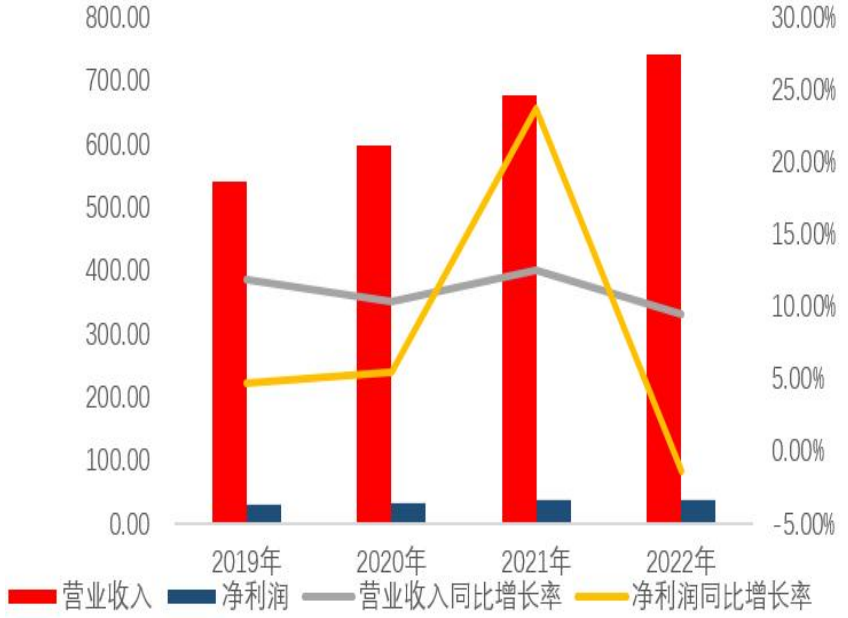
图表127 紫光股份子公司新华三交换机

- 紫光股份旗下交换机主要分为数据中心交换机、园区网交换机、全光网络和工业和安防交换机。

图表128 紫光股份子公司新华三服务器

- 紫光股份旗下服务器主要分为机架式、刀片、塔式、高密度、边缘和关键业务服务器。
- 旗下有H3C(新华三)服务器品牌及HPE(惠普)服务器代理品牌。

图表129 紫光股份2019-2022营业收入、净利润及增速（单位：亿元）



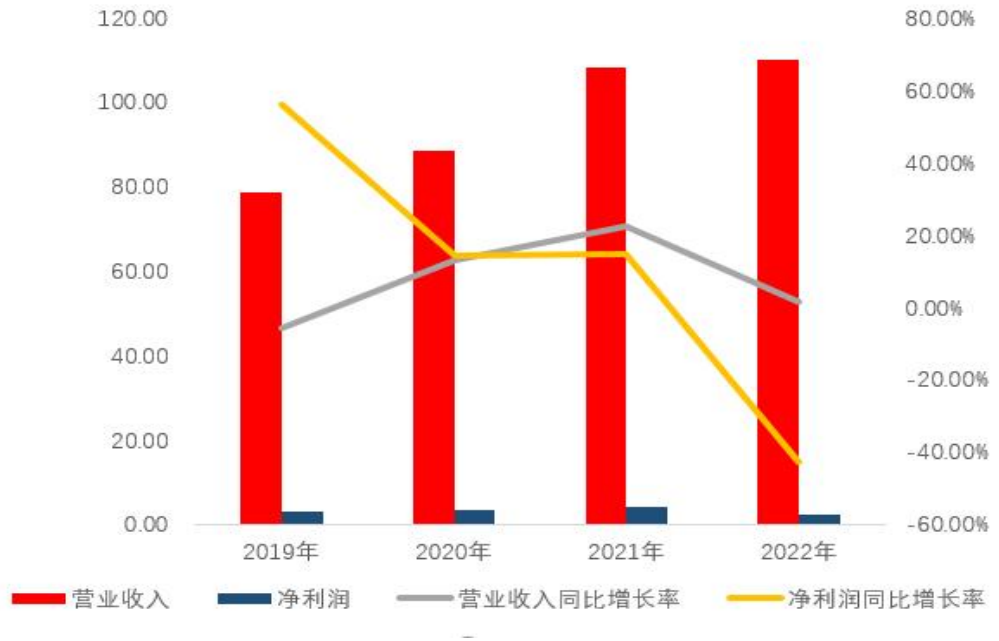
共进股份——深耕通信终端设备领域

- 公司看点：公司聚焦通信终端设备领域，公司牢牢把握数字经济发展机遇，扩大与核心客户在园区交换机、数据中心交换机等领域的合作，目前公司交换机业务主要以国内客户为主，同步覆盖海外优质客户，包括数据中心交换机、园区交换机等，以 100GbE、400GbE 为主。
- 业务增速： 2023年上半年营业收入43.55亿元，同比下降16.17%；净利润1.96亿元，同比增长2.90%。公司智慧通信业务辐射全球，其中就包括交换机和服务器业务。随着北美市场需求强劲，公司将持续加大海外订单的获取力度，扩大核心客户份额比例，探索突破直营模式。

图表130 共进股份交换机产品序列

交换机类型	产品图片	产品简介
数据中心交换机		T&W的数据中心解决方案，采用高速DCN交换芯片，配套BMC模组，给客户提供定制化的解决方案；
非管理交换机		T&W非管理交换机是一种便捷的解决方案，可通过即插即用的方式快速扩展办公网络，可以以千兆速度将有线以太网设备连接在一起。
24GE 4SFP交换机		T&W的二层换机解决方案包括一流的管理功能，同时对于更复杂的网络，Web Smart交换机支持高级管理功能。

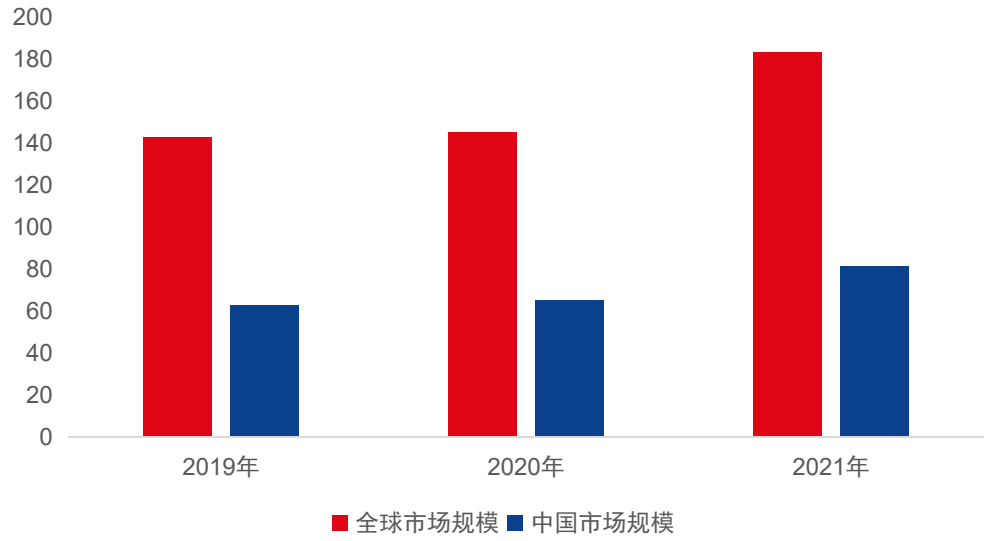
图表131 共进股份2019-2022营业收入、净利润及增速（单位：亿元）



布局方向二：内延挖掘——连接

- 高速背板连接器需求大涨。随着华为5G、6G建设推进，数据中心、交换机、服务器等设备应用愈发广泛，高速背板连接器需求大幅增长，特别是AI服务器。根据Bishop&Associates 的预测，2022 年我国高速背板连接器市场规模将达到6.13亿美元
- 高速背板连接器技术壁垒高：应用于高端服务器，是连接器里面技术壁垒最高的，一块PCB板上多达几十个。在小型服务器中大概有3-5对高速背板连接器。

图表132：全球及中国通讯连接器市场规模（亿美元）



图表133：易天光通信200G AOC

创益通：数据连接景气提升，公司迎来爆发机遇

华金证券 Huajin Securities
华发集团旗下企业

◆**公司看点：**主要从事设计研发与制造精密连接器、连接线、精密结构件等互连产品。产品主要应用于5G通信、新能源汽车、数据存储、医疗、消费电子等领域。

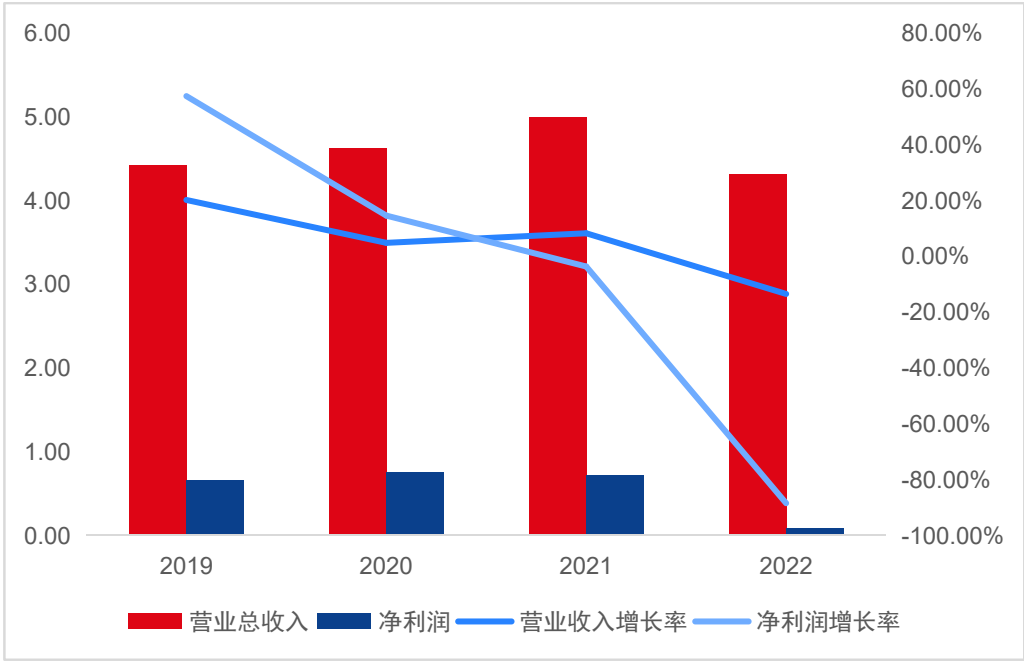
◆**重点客户：**小米及其旗下公司、西部数据（晟碟）、星科金朋、莫仕、安克创新、公牛集团等国内外知名公司。

◆**业务增速：**2023年上半年营业收入2.18亿元，同比增长14.17%；净利润-0.12亿元，同比下降328.56%。随着数据存储行业不断复苏，公司未来盈利空间有望进一步扩大。

图表134：公司主流产品及应用场景

类别	主流产品	产品	产品图	应用场景
数据存储互连产品及组件	高速连接器及组件、高频高速数据线	SATA、DDR4 系列、SD、连接器组件等		高速存储、数据传输、支撑和固定电子产品零部件及电磁屏蔽等
消费电子互连产品及组件	通用连接器、数据线	工业连接器、Apple Lightning 等		为多媒体数据传输、充电、显示、数据传输
通讯互连产品及组件	通讯互连产品及组件	高速背板连接器组件、高速射频连接器组件		能为电子系统中各模块间高速数据信号传输；为超小型盲插射频连接器射频信号传输
精密结构件	新能源	Package 箱体系列、软/硬铜排系列、单体锂电池组件		为锂离子/单体电池包提供安全保护；为串联电路提供大电流传输能力

图表135：2019-2022营业收入、净利润及增速（单位：亿元）



华丰科技：受益国防建设+国产替代，打开连接器成长空间

华金证券 Huajin Securities
华发集团旗下企业

◆**公司看点：**全球光电连接器及互连方案提供商。研制的连接器从低频到高频、从电口到光口，从标准型到特殊定制，能适应深海环境、高压环境、高振动冲击环境、电磁波干扰环境、高低温环境、太空宇航环境等多种特殊环境。为用户提供互连解决方案一站式服务。致力于向全球防务装备、通讯装备、轨道交通装备和新能源汽车提供更专业、更可靠、更智能的互连技术和产品。

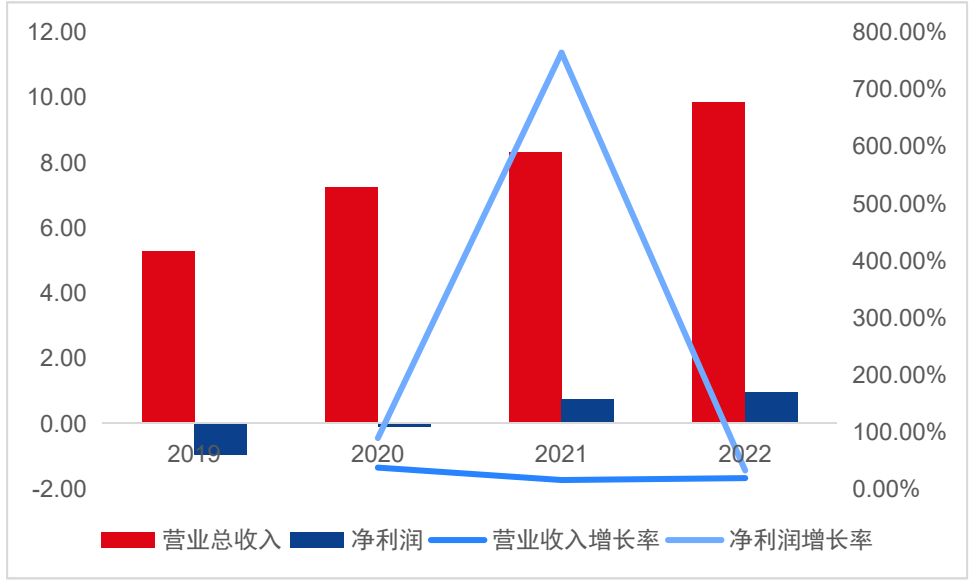
◆**重点客户：**华为、中兴等移动通信设备制造商，航天科工、中国电科、中国兵工等航空航天及防务集团下属单位以及上汽通用五菱、比亚迪等汽车制造厂商。

◆**业绩增速：**2023年上半年营业收入4.15亿元，同比下降14.35%；净利润0.36亿元，同比下降31.04%。在连接器行业加速国产替代化的背景下，公司有望开拓成长空间，实现快速增长。

图表136：公司主流产品及应用场景

类别	主流产品	产品	产品图	应用场景
防务类连接产品	系统互连产品、防务连接器、组件	J29M矩形盲插连接器、CPBS无缆化母板组件、树型拓扑结构线缆组件		应用领域主要为航天、船舶、电子、防务装备等领域的信息系统电子设备与设备间、设备内部、模块与板卡间、印制板间的系统互连。
通讯类连接产品	高速连接器、印制板连接器、光通讯连接器、线缆组件等	P 系列、I/O连接器等、PCB 板上电源连接器、光模块等		主要用于点对点及点对多点传输接口、背板交换应用、相控阵雷达数据通信、以太网、光纤通道、Infiniband QDR 等场合
工业类连接产品	轨道交通类产品 、新能源汽 车类产品	轨道交通连接器、电气车钩总成、BDU/PDU 充配电系统总成等		广泛应用于高速列车、电力机车及地铁轻轨的电气控制与集成布线系统中，为不同设备或功能单元之间的电气或信号提供电连接。

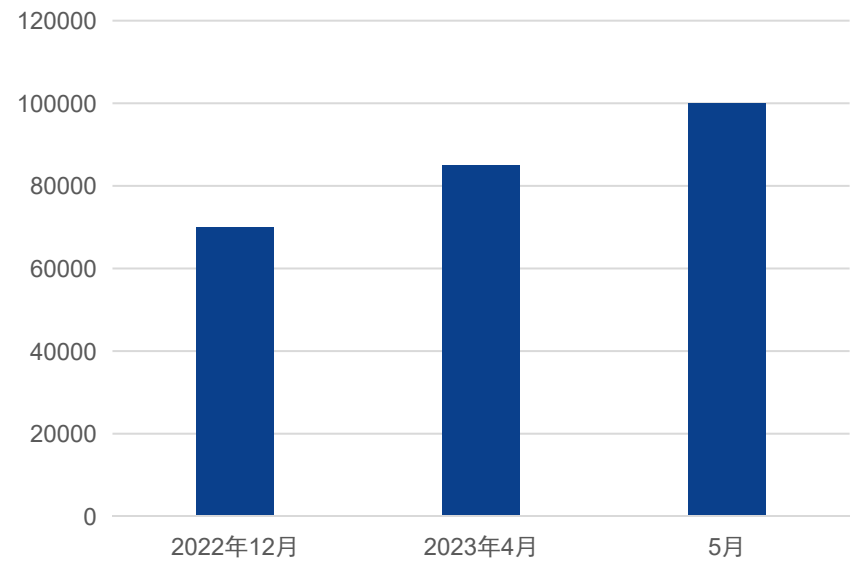
图表137：2019-2022营业收入、净利润及增速（单位：亿元）



布局方向三：国内算力资源池——算力资源+IDC

- 全国智算中心建设加速。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%。在新一轮AI浪潮和大模型催化下，智算中心建设投资将迎来高峰，产业链各核心环节有望迎来爆发机遇。
- 英伟达A800价格飙升，得卡者得生产力。需求起势，供不应求，A800价格不断上涨。在下游互联网企业、云厂商们更加广泛地布局AI大模型的形势下，得GPU者得生产力。

图表138：A800价格走势



图表139：我国综合型云厂商

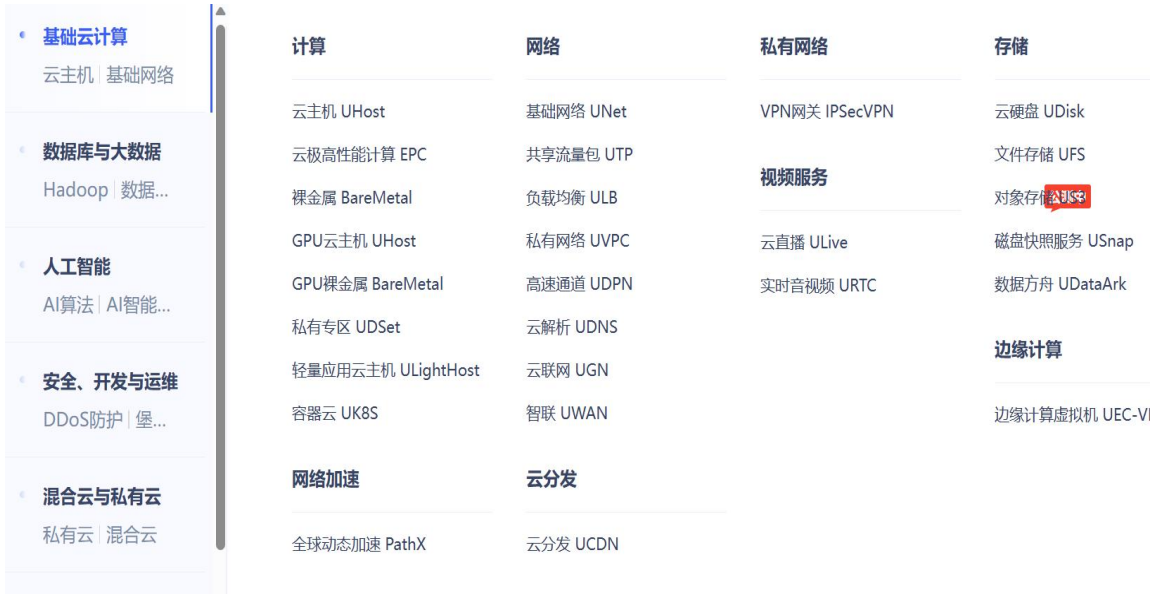
类别	公司名称	概述
综合型云厂商	阿里云	IaaS+PaaS公有云市场第一，且市占率超2-4名之和
	腾讯云	IaaS+PaaS公有云市场第二，游戏类公有云、视频云垂直领域第一
	华为云	IaaS+PaaS公有云市场第三，中国容器云市场第一
	天翼云	IaaS+PaaS公有云市场第四，全球运营商云第一，中国政务云第二
	金山云	与鞍钢集团共建中国制造业首个行业云
	百度云	中国智能媒体解决方案、工业质检解决方案市场第一
	浪潮云	拥有中国最大的分布式云体系，在政务云市场上位列第一。
	京东云	发布行业内首个混合云操作系统‘云舰”

优刻得：持续优化业务布局，技术研发带来竞争力提升

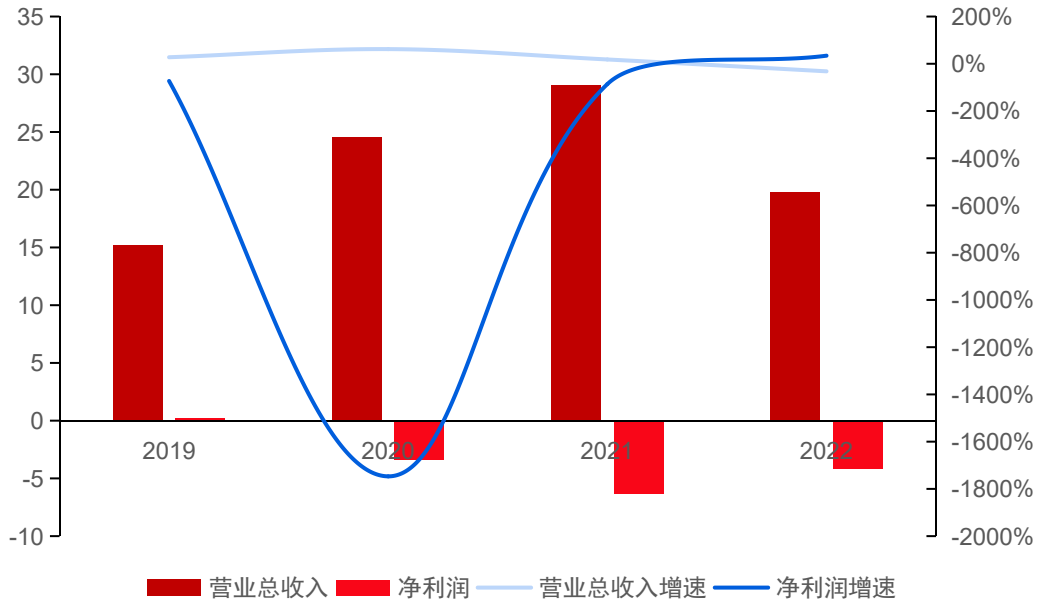
华金证券 Huajin Securities
华发集团旗下企业

- 公司看点：公司自主研发IaaS、PaaS、大数据流通平台、AI服务平台等一系列云计算产品，并深入了解互联网、传统企业在不同场景下的业务需求，提供公有云、混合云、私有云、专有云在内的综合性行业解决方案。
- 重点客户：主要来自互动娱乐、移动互联、企业服务三个领域。
- 业务增速：2023年上半年营业收入7.39亿元，同比下降29.31%；净利润-1.90亿元，同比增长28.28%。公司业务优化调整，造成利润水平下降，但长期看盈利能力提升。

图表140 公司主流业务及应用场景



图表141 2019-2022营业收入、净利润及增速（单位：亿元）



润泽科技：国内IDC龙头，盈利能力领先同行业

- **公司看点**：公司是国内IDC 龙头企业，致力于面向大数据、云计算、物联网、5G 技术等行业应用需求，以算力为基础、数字技术为手段、智慧应用为示范，为各行业提供新一代数字经济产业技术、产品、服务和系统解决方案。
- **重点客户**：中国电信（第一）、中国联通（第二）、字节跳动、华为、京东、快手等。
- **业务增速**：2023年上半年，公司实现营业收入16.83亿元，同比增长32.30%；净利润7.02亿元，同比增长47.91%。

图表142 公司主流业务（单位：万元）

产品	总投资金额	拟使用募集资金
润泽（佛山）国际信息港A2、A3、数据中心项目	169668	169668
润泽（平湖）国际信息港A2数据中心项目	81306	76306
偿还银行借款	243800	209026
中介机构费用及相关发行费用	15000	15000
合计	509774	470000

图表143 长三角·平湖润泽国际信息港B区项目效果图

建议关注标的

华金证券 Huajin Securities
华发集团旗下企业

证券代码	证券简称	总市值（亿元）	EPS-2022A	EPS-2023E	EPS-2024E	PE-2022A	PE-2023E	PE-2024E
300308.SZ	中际旭创	792.23	1.68	2.05	4.00	64.73	47.96	25.77
000988.SZ	华工科技	316.83	0.92	1.11	1.48	34.97	27.87	21.41
600105.SH	永鼎股份	88.49	0.17	0.24	0.32	39.14	26.76	20.55
000938.SZ	紫光股份	704.15	0.78	0.93	1.13	32.63	26.51	21.61
000063.SZ	中兴通讯	1,476.21	1.88	2.09	2.44	19.30	15.44	13.29
000977.SZ	浪潮信息	527.32	0.99	1.61	2.02	25.35	22.61	17.62
301191.SZ	菲菱科思	64.47	2.61	3.28	4.38	33.00	28.21	20.88
600602.SH	云赛智联	158.61	0.13	0.16	0.20	102.88	83.69	68.14
300442.SZ	润泽科技	418.77	1.12	1.06	1.35	34.95	23.05	17.90
300913.SZ	兆龙互连	82.25	0.45	0.58	0.81	63.05	53.06	40.19
603220.SH	中贝通信	78.00	0.41	0.57	0.74	71.70	40.06	31.69
688629.SH	华丰科技	100.50	0.19	0.25	0.34	101.73	88.40	62.04

目录

投资建议

- 我们认为，下半年随着国内订购的英伟达芯片逐渐到位，算力下沉赛道在国内会最先爆发，包括AI服务器放量，随着智算中心建设加速对算力组网所需的400G投资机会以及国内互联网公司、高校、初创公司对训练模型算力的需求延伸出的算力租赁投资机会
- 建议重点关注：紫光股份、菲菱科思、浪潮信息、共进股份、中贝通信、润泽科技等

风险提示

- 模型算法及应用市场拓展不及预期;随着chatGPT带动，生成式算法迭代加速，但主流大模型的模型参数达上千亿甚至万亿，国内大模型迭代速度不及预期。大模型落地需要应用场景加持，目前市场在等待爆款应用的推出。
- 芯片成本过高、订单周期过长影响落地进度。目前英伟达芯片供给虽有改善，但需求增长的速度远大于产能速度，订单执行周期过长，此外芯片价格偏高，导致最终产品售价过高，影响产品的普及性。
- 智算中心建设进度不及预期。智算中心建设需要综合考虑，土建、设备、能耗指标都是硬性条件，此外智算所需的卡以及客户都是市场关心的重点，这些资源要素在推进过程中面临着不确定性。

评级说明

公司评级体系

收益评级：

买入 — 未来6个月的投资收益率领先沪深300指数15%以上；

增持 — 未来6个月的投资收益率领先沪深300指数5%至15%；

中性 — 未来6个月的投资收益率与沪深300指数的变动幅度相差-5%至5%；

减持 — 未来6个月的投资收益率落后沪深300指数5%至15%；

卖出 — 未来6个月的投资收益率落后沪深300指数15%以上。

风险评级：

A — 正常风险，未来6个月投资收益率的波动小于等于沪深300指数波动；

B — 较高风险，未来6个月投资收益率的波动大于沪深300指数波动。

评级说明

行业评级体系

收益评级：

领先大市 — 未来6个月的投资收益率领先沪深300指数10%以上；

同步大市 — 未来6个月的投资收益率与沪深300指数的变动幅度相差-10%至10%；

落后大市 — 未来6个月的投资收益率落后沪深300指数10%以上；

风险评级：

A — 正常风险，未来6个月投资收益率的波动小于等于沪深300指数波动；

B — 较高风险，未来6个月投资收益率的波动大于沪深300指数波动。

法律声明

华金证券 Huajin Securities
华发集团旗下企业

分析师声明

李宏涛声明，本人具有中国证券业协会授予的证券投资咨询执业资格，勤勉尽责、诚实守信。本人对本报告的内容和观点负责，保证信息来源合法合规、研究方法专业审慎、研究观点独立公正、分析结论具有合理依据，特此声明。

本公司具备证券投资咨询业务资格的说明

华金证券股份有限公司（以下简称“本公司”）经中国证券监督管理委员会核准，取得证券投资咨询业务许可。本公司及其投资咨询人员可以为证券投资人或客户提供证券投资分析、预测或者建议等直接或间接的有偿咨询服务。发布证券研究报告，是证券投资咨询业务的一种基本形式，本公司可以对证券及证券相关产品的价值、市场走势或者相关影响因素进行分析，形成证券估值、投资评级等投资分析意见，制作证券研究报告，并向本公司的客户发布。

法律声明

华金证券 Huajin Securities
华发集团旗下企业

- **免责声明：**
- 本报告仅供华金证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因为任何机构或个人接收到本报告而视其为本公司的当然客户。
- 本报告基于已公开的资料或信息撰写，但本公司不保证该等信息及资料的完整性、准确性。本报告所载的信息、资料、建议及推测仅反映本公司于本报告发布当日的判断，本报告中的证券或投资标的价格、价值及投资带来的收入可能会波动。在不同时期，本公司可能撰写并发布与本报告所载资料、建议及推测不一致的报告。本公司不保证本报告所含信息及资料保持在最新状态，本公司将随时补充、更新和修订有关信息及资料，但不保证及时公开发布。同时，本公司有权对本报告所含信息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。任何有关本报告的摘要或节选都不代表本报告正式完整的观点，一切须以本公司向客户发布的本报告完整版本为准。
- 在法律许可的情况下，本公司及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，提请客户充分注意。客户不应将本报告为作出其投资决策的惟一参考因素，亦不应认为本报告可以取代客户自身的投资判断与决策。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，无论是否已经明示或暗示，本报告不能作为道义的、责任的和法律的依据或者凭证。在任何情况下，本公司亦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。
-
- 华金证券股份有限公司对本声明条款具有惟一修改权和最终解释权。

法律声明

华金证券 Huajin Securities
华发集团旗下企业

风险提示:

- 报告中的内容和意见仅供参考，并不构成对所述证券买卖的出价或询价。投资者对其投资行为负完全责任，我公司及其雇员对使用本报告及其内容所引发的任何直接或间接损失概不负责。
-
-
- 华金证券股份有限公司
- 办公地址：
- 上海市浦东新区杨高南路759号陆家嘴世纪金融广场30层
- 北京市朝阳区建国路108号横琴人寿大厦17层
- 深圳市福田区益田路6001号太平金融大厦10楼05单元

- 电话：021-20655588
- 网址： www.huajinsc.cn