

算力大时代，AI 算力产业链全景梳理

TMT

核心观点

生成式 AI 取得突破，我们对生成式 AI 带来的算力需求做了上下游梳理，并做了交叉验证，可以看到以 ChatGPT 为代表的大模型训练和推理端均需要强大的算力支撑，产业链共振明显，产业链放量顺序为：先进制程制造->以 Chiplet 为代表的 2.5D/3D 封装、HBM->AI 芯片->板卡组装->交换机->光模块->液冷->AI 服务器->IDC 出租运维。综合来看，大模型仍处于混战阶段，应用处于渗透率早期，AI 板块中算力需求增长的确 定性较高，在未来两年时间内，算力板块都将处于高景气度阶段，重点推荐 AI 算力产业链各环节相关公司。

摘要

生成式 AI 取得突破，实现了从 0 到 1 的跨越，以 ChatGPT 为代表的人工智能大模型训练和推理需要强大的算力支撑。自 2022 年底 OpenAI 正式推出 ChatGPT 后，用户量大幅增长，围绕 ChatGPT 相关的应用层出不穷，其通用性能力帮助人类在文字等工作上节省了大量时间。同时在 Transformer 新架构下，多模态大模型也取得新的突破，文生图、文生视频等功能不断完善，并在广告、游戏等领域取得不错的进展。生成式 AI 将是未来几年最重要的生产力工具，并深刻改变各个产业环节，围绕生成式 AI，无论是训练还是推理端，算力需求都将有望爆发式增长。

训练和推理端 AI 算力需求或几何倍数增长。首先是训练侧，参考 OpenAI 论文，大模型训练侧算力需求=训练所需要的 token 数量*6*大模型参数量。可以看到从 GPT3.5 到 GPT4，模型效果越来越好，模型也越来越大，训练所需要的 token 数量和参数量均大幅增长，相应的训练算力需求也大幅增长。并且，与 GPT4 相关的公开论文也比较少，各家巨头向 GPT4 迈进的时候，需要更多方向上的探索，也将带来更多的训练侧算力需求。根据我们的推算，2023 年-2027 年，全球大模型训练端峰值算力需求量的年复合增长率有望达到 78.0%，2023 年全球大模型训练端所需全部算力换算成的 A100 芯片总量可能超过 200 万张。其次是推理侧，单个 token 的推理过程整体运算量为 2*大模型参数量，因此大模型推理侧每日算力需求=每日调用大模型次数*每人平均查询 Token 数量*2*大模型参数量，仅以 Google 搜索引擎为例，每

维持

强于大市

武超则

wuchaoze@csc.com.cn

010-85156318

SAC 编号:S1440513090003

SFC 编号:BEM208

阎贵成

yanguicheng@csc.com.cn

010-85159231

SAC 编号:S1440518040002

SFC 编号:BNS315

刘双锋

liushuangfeng@csc.com.cn

SAC 编号:S1440520070002

SFC 编号:BNU539

金戈

jinge@csc.com.cn

010-85159348

SAC 编号:S1440517110001

SFC 编号:BPD352

于芳博

yufangbo@csc.com.cn

010-86451607

SAC 编号:S1440522030001

崔世峰

cuishifeng@csc.com.cn

SAC 编号:S1440521100004

刘永旭

liuyongxu@csc.com.cn

010-86451440

SAC 编号:S1440520070014

杨伟松

yangweisong@csc.com.cn

SAC 编号:S1440522120003

范彬泰

fanbintai@csc.com.cn

SAC 编号:S1440521120001

发布日期：2023 年 06 月 14 日

年调用次数至少超过 2 万亿，一旦和大模型结合，其 AI 算力需求将十分可观。随着越来越多的应用和大模型结合，推理侧算力需求也有望呈现爆发增长势头。根据我们的推算，2023 年-2027 年，全球大模型云端推理的峰值算力需求量的年复合增长率有望高达 113%。

算力产业链价值放量顺序如下：先进制程制造->以 Chiplet 为代表的 2.5D/3D 封装、HBM->AI 芯片->板卡组装->交换机->光模块->液冷->AI 服务器->IDC 出租运维。

先进封装、HBM：为了解决先进制程成本快速提升和“内存墙”等问题，Chiplet 设计+异构先进封装成为性能与成本平衡的最佳方案，台积电开发的 CoWoS 封装技术可以实现计算核心与 HBM 通过 2.5D 封装互连，因此英伟达 A100、H100 等 AI 芯片纷纷采用台积电 CoWoS 封装，并分别配备 40GB HBM2E、80GB 的 HBM3 内存。全球晶圆代工龙头台积电打造全球 2.5D/3D 先进封装工艺标杆，未来几年封装市场增长主要受益于先进封装的扩产。先进封装市场的快速增长，有望成为国内晶圆代工厂商（中芯国际）与封测厂商（长电科技、通富微电、甬矽电子和深科技）的新一轮成长驱动力。

AI 芯片/板卡封装：以英伟达为代表，今年二季度开始释放业绩。模型训练需要规模化的算力芯片部署于智能服务器，CPU 不可或缺，但性能提升遭遇瓶颈，CPU+GPU 异构方案成为大算力场景标配。其中 GPU 并行计算优势明显，CPU+GPU 成为目前最流行的异构计算系统，而 NPU 在特定场景下的性能、效率优势明显，推理端应用潜力巨大，随着大模型多模态发展，硬件需求有望从 GPU 扩展至周边编解码硬件。AI 加速芯片市场上，英伟达凭借其硬件产品性能的先进性和生态构建的完善性处于市场领导地位，在训练、推理端均占据领先地位。根据 Liftr Insights 数据，2022 年数据中心 AI 加速市场中，英伟达份额达 82%。因此 AI 芯片需求爆发，英伟达最为受益，其 Q2 收入指引 110 亿美金，预计其数据中心芯片业务收入接近翻倍。国内厂商虽然在硬件产品性能和产业链生态架构方面与前者有所差距，但正在逐步完善产品布局和生态构建，不断缩小与行业龙头厂商的差距，并且英伟达、AMD 对华供应高端 GPU 芯片受限，国产算力芯片迎来国产替代窗口期。当前已经涌现出一大批国产算力芯片厂商：1）寒武纪：国内人工智能芯片领军者，持续强化核心竞争力；2）海光信息：深算系列 GPGPU 提供高性能算力，升级迭代稳步推进；3）龙芯中科：自主架构 CPU 行业先行者，新品频发加速驱动成长；4）芯原股份：国内半导体 IP 龙头，技术储备丰富驱动成长；5）工业富联：提供 GPU 芯片板卡组装服务。

交换机：与传统数据中心的网络架构相比，AI 数据网络架构会带来更多的交换机端口的需求。交换机具备技术壁垒，中国市场格局稳定，华为与新华三（紫光股份）两强争霸，锐捷网络展现追赶势头，建议重点关注。

光模块：AI 算力带动数据中心内部数据流量较大，光模块速率及数量均有显著提升。训练侧光模块需求与 GPU 出货量强相关，推理侧光模块需求与数据流量强相关，伴随应用加速渗透，未来推理所需的算力和流量实际上可能远大于训练。目前，训练侧英伟达的 A100 GPU 主要对应 200G 光模块和 400G 光模块，H100 GPU 可以对应 400G 或 800G 光模块。根据我们的测算，训练端 A100 和 200G 光模块的比例是 1:7，H100 和 800G 光模块的比例是 1:3.5。800G 光模块 2022 年底开始小批量出货，2023 年需求主要来自于英伟达和谷歌。在 2023 年这个时间点，市场下一代高速率光模块均指向 800G 光模块，叠加 AIGC 带来的算力和模型竞赛，我们预计北美各大云厂商和相关科技巨头均有望在 2024 年大量采购 800G 光模块，同时 2023 年也可能提前采购。建议关注中际旭创、天孚通信、新易盛、华工科技、源杰科技、太辰光、光迅科技、光库科技、中瓷电子、剑桥科技、博创科技、联特科技、德科立、仕佳光子等。

光模块上游——光芯片：以 AWG、PLC 等为代表的无源光芯片，国内厂商市占率全球领先。以 EEL、VCSEL、DFB 等激光器芯片、探测器芯片和调制器芯片为代表的有源光芯片是现代光学技术的重要基石，是有源光器件的重要组成部分。以源杰科技、光库科技为代表的国内光芯片厂商不断攻城拔寨，在多个细分产品领域取得了

较大进展，国产替代化加速推进，市场空间广阔。

液冷：AI 大模型训练和推理所用的 GPU 服务器功率密度将大幅提升，以英伟达 DGX A100 服务器为例，其单机最大功率约可达到 6.5kW，大幅超过单台普通 CPU 服务器 500w 左右的功率水平。根据《冷板式液冷服务器可靠性白皮书》数据显示，自然风冷的数据中心单柜密度一般只支持 8kW-10kW，通常液冷数据中心单机柜可支持 30kW 以上的散热能力，并能较好演进到 100kW 以上，相较而言液冷的散热能力和经济性均有明显优势。同时“东数西算”明确 PUE（数据中心总能耗/IT 设备能耗）要求，枢纽节点 PUE 要求更高，同时考虑到整体规划布局，未来新增机柜更多将在枢纽节点内，风冷方案在某些地区可能无法严格满足要求，液冷方案渗透率有望加速提升。目前在 AI 算力需求的推动下，如浪潮信息、中兴通讯等服务器厂商已经开始大力布局液冷服务器产品。在液冷方案加速渗透过程中，数据中心温控厂商、液冷板制造厂商等有望受益，建议关注：英维克、高澜股份、网宿科技、曙光数创等。

AI 服务器：预计今年 Q2-Q3 开始逐步释放业绩。具体来看，训练型 AI 服务器成本中，约 7 成以上由 GPU 构成，其余 CPU、存储、内存等占比相对较小，均价常达到百万元以上。对于推理型服务器，其 GPU 成本约为 2-3 成，整体成本构成与高性能型相近，价格常在 20-30 万。根据 IDC 数据，2022 年全球 AI 服务器市场规模 202 亿美元，同比增长 29.8%，占服务器市场规模的比例为 16.4%，同比提升 1.2pct。我们认为全球 AI 服务器市场规模未来 3 年内将保持高速增长，市场规模分别为 395/890/1601 亿美元，对应增速 96%/125%/80%。根据 IDC 数据，2022 年中国 AI 服务器市场规模 67 亿美元，同比增长 24%。我们预计，2023-2025 年，结合对于全球 AI 服务器市场规模的预判，以及对于我国份额占比持续提升的假设，我国 AI 服务器市场规模有望达到 134/307/561 亿美元，同比增长 101%/128%/83%。竞争格局方面，考虑到 AI 服务器研发和投入上需要更充足的资金及技术支持，国内市场的竞争格局预计将继续向头部集中，保持一超多强的竞争格局。重点推荐：1）浪潮信息：全球服务器行业龙头厂商，其 AI 服务器多次位列全球市占率第一；2）工业富联：为英伟达提供 H100 等芯片组装，以及 AI 服务器生产；3）紫光股份：子公司新华三 AI 服务器在手订单饱满，同时可以提供交换机、路由器等；4）中科曙光：高性能计算及国产化服务器龙头；5）中兴通讯：服务器业务快速增长；6）拓维信息：华为昇腾+鲲鹏核心合作伙伴；7）联想集团：全球领先的 ICT 设备企业。

IDC：在数字中国和人工智能推动云计算市场回暖的背景下，IDC 作为云基础设施产业链的关键环节，也有望进入需求释放阶段。在过去两年半，受多重因素影响下，云计算需求景气度下行，但 IDC 建设与供给未出现明显放缓，2021 年和 2022 年分别新增机柜数量 120 万架和 150 万架，因此短期内出现供需失衡情况（核心区域供需状况相对良好），部分地区上电率情况一般。所以 IDC 公司 2022 年业绩普遍承压。当前，我们认为国内 IDC 行业有望边际向好。随着宏观经济向好，平台经济发展恢复，AI 等拉动，IDC 需求有望逐步释放，叠加 2023 新增供给量有望较 2022 年减少（例如三大运营商 2022 年新增 IDC 机柜 15.6 万架，2023 年计划新增 11.4 万架）。展望未来，电信运营商在云计算业务方面仍将实现快速增长，百度、字节跳动等互联网公司在 AIGC 领域有望实现突破性进展，都将对包括 IDC 在内的云基础设施产生较大新增需求，相关 IDC 厂商有望获益，建议关注润泽科技、宝信软件、奥飞数据、数据港、光环新网等。

目 录

一、AI 有望明显拉动算力基础设施投资.....	1
1.1 ChatGPT 爆红引发了人们对于人工智能发展的高度关注.....	1
1.2 人工智能需要强大算力支撑	2
1.3 AI 算力产业链涉及环节较多，行业需求有望全面提升	3
二、AI 芯片需求爆发式增长.....	5
2.1 AI 大规模落地应用对 AI 芯片性能、数量提出全方位要求	5
2.2 英伟达龙头地位稳固，国内厂商正逐步追赶.....	23
2.3 先进封装成为高性价比替代方案，存算一体应用潜力巨大.....	30
三、AI 服务器渗透率快速提升.....	40
3.1 AI 服务器是算力基础设施最主要的硬件，训练型主要成本来自于 GPU 芯片	40
3.2 AI 服务器市场规模有望保持高速增长，当前订单饱满.....	43
3.3 AI 服务器市场集中度有望提升，国内厂商呈现一超多强格局.....	45
3.4 全球服务器市场规模预计保持平稳	47
3.5 标的推荐	47
四、AI 正在推动高速率光模块需求放量.....	49
五、AI 将会拉动交换机市场需求.....	59
六、AI 提升大功率 IDC 机柜需求，液冷渗透率随之提升	62
6.1 “东数西算”统筹全国算力网络建设，云计算需求可能将回暖	62
6.2 AI 大算力服务器需要高功率机柜，液冷或成必选项.....	64
6.3 人工智能算力需求有望推动海底数据中心规模化发展.....	68
七、海外大模型进展	74
7.1 谷歌	74
7.2 微软	79
7.3 Meta.....	82
八、投资建议	86

图表目录

图表 1: AIGC 发展历程	1
图表 2: 国内外公司 AIGC 相关产品	2
图表 3: GPT 模型示意图	2
图表 4: NVIDIA DGX A100 AI 服务器	2
图表 5: 全球算力规模及增速	3
图表 6: 我国算力规模及增速	3
图表 7: 全球 AI 服务器市场规模测算	4
图表 8: 中国 AI 服务器市场规模测算	4
图表 9: 光模块和交换机速率演进示意图	5
图表 10: CPU+AI 芯片的异构计算	6
图表 11: 2021 年中国 AI 芯片市场规模占比	6
图表 12: CPU 与 GPU 架构对比	6
图表 13: NVIDIA GPU 主要产品线	7
图表 14: NVIDIA Fermi 架构至 Hopper 架构的变化	7
图表 15: 低精度比特位宽为 AI 计算带来的好处	8
图表 16: 不同精度计算消耗的能量和硅片面积	8
图表 17: NVIDIA 数据中心 GPU 支持的比特位宽变化	8
图表 18: V100 中 FP32 硬件单元和 FP64 硬件单元的数量关系	8
图表 19: 专门的硬件单元 Tensor Core 加速矩阵乘加计算	9
图表 20: A100 与 H100 的 FP16 Tensor Core 吞吐量对比	9
图表 21: FP16 Tensor Core 与 FP8 Tensor Core 吞吐量对比	9
图表 22: FP16 Tensor 算力快速增长	10
图表 23: FP16 Tensor 每单位核心的算力明显优于 FP16	10
图表 24: AI 训练服务器需要更高的内存容量	10
图表 25: NLP 负载中存储和计算的能耗占比	10
图表 26: GDDR 与 HBM 差异	11
图表 27: 语言模型的参数数量呈指数级增长	11
图表 28: GPU 之间通过 PCIe 连接	12
图表 29: GPU 之间通过 NVLink 连接	12
图表 30: NVLink 1.0—NVLink 4.0	12
图表 31: NVSwitch 连接多颗 GPU	13
图表 32: NVSwitch 支撑的 GPU 计算集群	13
图表 33: NPU 典型架构	14
图表 34: 麒麟 970 NPU 加速图像识别	14
图表 35: 脉动阵列运行矩阵乘法的示意图	14
图表 36: 谷歌 TPU 架构及其内部的脉动阵列	15
图表 37: 谷歌 TPU	15
图表 38: Tesla FSD 搭载 NPU 模块	15
图表 39: AI 训练与 AI 推理对比	16

图表 40:	云端推理占比逐步提升	16
图表 41:	AIGC 引发内容生成范式革命	16
图表 42:	NVIDIA 云端训练 GPU 与推理 GPU 参数对比	17
图表 43:	不同规模大模型所需的显存容量估计	17
图表 44:	边缘端 AI 推理芯片及其算力案例	18
图表 45:	大模型参数量及训练所需 Tokens	18
图表 46:	神经网络的前向传播过程	19
图表 47:	神经网络的反向传播过程	19
图表 48:	不同大模型训练过程中的算力利用率	19
图表 49:	全球大模型训练所需算力/AI 芯片数量测算	19
图表 50:	大模型云端推理所需算力/AI 芯片数量测算（算力角度）	21
图表 51:	大模型云端推理所需算力/AI 芯片数量测算（显存角度）	22
图表 52:	AI 芯片市场竞争格局	23
图表 53:	2022 年 AI 加速芯片市场份额	23
图表 54:	全球独显 GPU 市场份额	24
图表 55:	国内外主流图形渲染 GPU 产品性能对比	24
图表 56:	2022 年人工智能加速芯片在云上部署情况	25
图表 57:	英伟达芯片在 AI 学术论文中的出现频次	25
图表 58:	国内外主流 GPGPU 产品性能对比	26
图表 59:	谷歌 TPU v4 与英伟达 A100 性能指标对比	27
图表 60:	TPU v4 与英伟达 A100 在不同模型中的表现	27
图表 61:	国内外主流 ASIC 产品性能对比	28
图表 62:	CUDA 构建强大生态支持所有主流深度学习框架	29
图表 63:	CUDA 生态和 ROCm 生态对照	29
图表 64:	昇腾计算产业生态示意图	30
图表 65:	寒武纪软件开发平台	30
图表 66:	每百万门晶体管的成本在 28nm 后开始上升	30
图表 67:	先进制程芯片的研发费用大幅上升	30
图表 68:	Chiplet 有利于提升良率	31
图表 69:	用 Chiplet 技术的 7nm+14nm 的造价 vs 7nm	31
图表 70:	先进封装的层次	31
图表 71:	先进封装依据互连密度和性能排名	31
图表 101:	通用服务器与 AI 服务器的不同	41
图表 102:	GPU 与 CPU 产品特点	41
图表 103:	GPU 与 CPU 内部结构	41
图表 104:	AI 服务器训练及推理区别	42
图表 105:	AI 服务器产业链概览	42
图表 106:	各类型服务器成本结构拆分	43
图表 107:	浪潮 AI 服务器售价及 GPU 成本占比估算	43
图表 108:	全球 AI 服务器市场规模测算	44
图表 109:	中国 AI 服务器市场规模测算	44

图表 110:	2022 年上半年全球 AI 服务器市场份额	45
图表 111:	2022 年中国 AI 服务器市场份额	46
图表 112:	浪潮信息服务器产品体系	48
图表 113:	拓维信息研发体系	49
图表 114:	传统三层网络架构	50
图表 115:	叶脊网络架构	50
图表 116:	英伟达 DGX A100 SuperPOD 采用胖树网络三层架构示意图	50
图表 117:	英伟达 DGX A100 SuperPOD 系统示意图	51
图表 118:	Mellanox HDR 200Gb/s Infiniband 网卡示意图	51
图表 119:	DGX H100 服务器背板连接图	51
图表 120:	NVLink 不同代际的升级 Roadmap	52
图表 121:	PCIe 不同代际的性能参数表	52
图表 122:	A100 和 H100 POD 采用 IB 和 NVLink 网络的示意图	53
图表 123:	GH200 的网络连接示意图	53
图表 124:	GH200 的网络连接示意图	54
图表 125:	Intel 的 100G 硅光模块示意图	55
图表 126:	硅光、InP、体材料铌酸锂和薄膜铌酸锂调制器的对比示意图	55
图表 127:	交换机发展示意图	56
图表 128:	LPO 方案的优势	56
图表 129:	光模块厂商目前拥有的 800G 光模块产品	57
图表 130:	北美云厂商资本开支（百万美元）	58
图表 131:	中际旭创股价复盘	58
图表 132:	微软 Azure 的 DGX H100 AI 超级计算机系统	59
图表 133:	不同网络架构的对比	60
图表 134:	2022 年全球前五大以太网交换机厂商	60
图表 135:	2021 年中国交换机市场份额	60
图表 136:	交换机发展示意图	61
图表 137:	交换机内部 SerDes 功耗占比大幅提升	61
图表 138:	网络部分的功耗在数据中心中占比大幅提升	61
图表 139:	CPO 可以降低功耗	62
图表 140:	CPO 所降低的功耗拆分示意图	62
图表 141:	“东数西算”工程设立 8 个节点	63
图表 142:	“东数西算”工程设立 10 个集群	63
图表 143:	中国 IDC 标准机架规模	64
图表 144:	IDC 机房的各类消耗	64
图表 145:	我国数据中心能耗分布	64
图表 146:	液冷数据中心制冷架构示意图	65
图表 147:	各类制冷方式情况梳理	65
图表 148:	浪潮信息液冷服务器产品	66
图表 149:	中兴通讯全液冷数据中心项目获奖	66
图表 150:	华北地区某数据中心节能改造示意图	67

图表 151:	数据港 Capex 支出构成	68
图表 152:	数据港 OPEX 支出构成	68
图表 153:	水下数据中心示例图	69
图表 154:	中国海上风电装机量 (GW)	70
图表 155:	海上风电经济性指标测算	70
图表 156:	建设在海边的水下数据中心	71
图表 157:	IDC 机房的各类消耗	71
图表 158:	我国数据中心能耗分布	71
图表 159:	水下数据中心与传统陆上 IDC 部分指标对比	72
图表 160:	微软 Natick 项目测试指标	73
图表 161:	微软 Natick 项目第二阶段——水下数据中心	73
图表 162:	微软 Natick 项目第二阶段位置图	74
图表 163:	不同大语言模型的预训练数据集结构 (%)	75
图表 164:	Google 在分布式集群计算资源利用率方面处于相对领先地位	75
图表 165:	TPUv4 在多个下游场景中表现优于 A100	75
图表 166:	TPU v4 在 BERT 上表现优于 A100	76
图表 167:	TPU v4 在 ResNet 上表现优于 A100	76
图表 168:	目前学界/业界提升模型计算效率的策略分类	76
图表 169:	OPT-175B survived 143K steps	77
图表 170:	Fine-tuning performance of the T5 Base, Large, and 11B on the GLUE dev set	77
图表 171:	SAM 提升了模型对标签噪声的稳健性, 并优化了模型训练效率	78
图表 172:	当模型性能超越一般人时, Alignment 成为挑战	79
图表 173:	ZeRO 优化下 POS 实现显存占用优化至基准方法的 26.2%	80
图表 174:	ZeRO-Offload 对 GPU/CPU 计算的切分	80
图表 175:	PipeDream 结合模型并行、数据并行和流水并行降低通信成本	81
图表 176:	不同并行化策略下计算资源利用率情况 (%)	81
图表 177:	LoRA 只调试低秩的 A、B, 预训练权重保持不变	82
图表 178:	LoRA 调试下 GPT-2 模型实现训练参数压缩, 同时性能优化	82
图表 179:	LoRA 调试策略下训练参数大幅减少, 同时性能与 Fine-tune 持平或更好	82
图表 180:	通过调整学习率, ResNet-50 mini-batch 训练可实现 8K 内性能不损失	83
图表 181:	对于 AlexNet 网络, 不同层的权值和其梯度的范数的比值差异很大	83
图表 182:	LARS 优化器主要根据范数的比值来调节每一层的学习率	83
图表 183:	W/O LARS 时 AlexNet-BN 8K 训练存在性能损失	84
图表 184:	W/ LARS 时 AlexNet-BN 8K 训练不存在性能损失	84
图表 185:	LARS 优化器将 ResNet 50 无损训练批量提升至 32K	84
图表 186:	LARS 与 LAMB 算法对比	85
图表 187:	LAMB 优化器训练下 BERT 模型的训练批量可扩展至 32K	85
图表 188:	GEM 算法	85
图表 189:	FSDP workflow	86

一、AI 有望明显拉动算力基础设施投资

1.1 ChatGPT 爆红引发了人们对于人工智能发展的高度关注

人工智能（AI）是指由机器展示的智能，即计算机基于大数据模拟人脑的各项功能，例如推理、视觉识别、语义理解、学习能力及规划与决策能力等。人工智能生成内容（AIGC）是指利用人工智能技术来生成内容，包括绘画、作曲、剪辑、写作等。AIGC 的萌芽可追溯到上世纪 50 年代，90 年代从实验性向实用性逐渐转变，但受限于算法瓶颈，无法直接生成内容，从 21 世纪 10 年代开始，随着以生成对抗网络（GAN）为代表的深度学习算法的提出和迭代，AIGC 迎来了快速发展阶段。

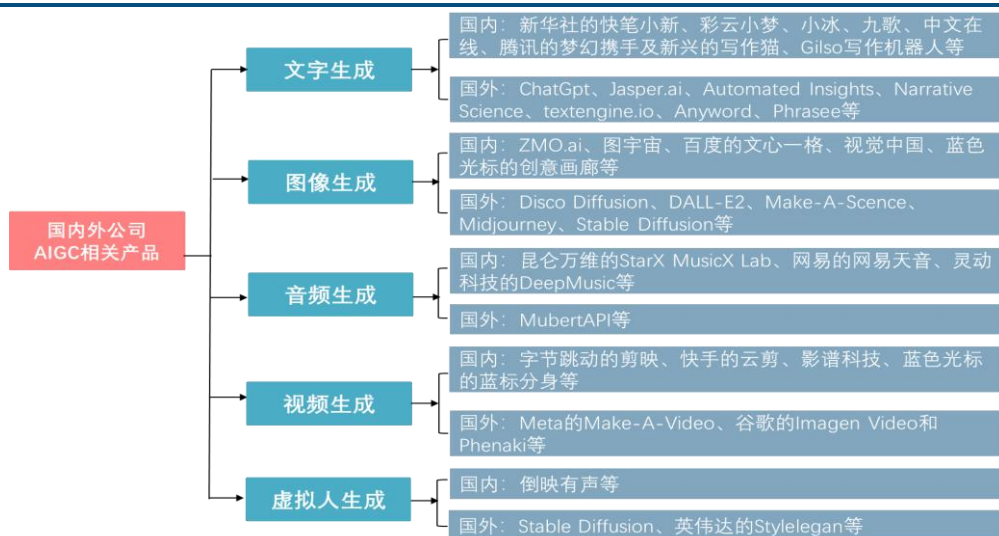
图表1： AIGC 发展历程



数据来源：《人工智能生成内容白皮书 2022》，中信建投

市场需求推动 AIGC 技术加速落地。1) 降低人力和时间成本：AIGC 可以帮助人们完成许多繁琐工作，从而节省人力资本和工作时间，并可以在相同的时间内产出更多内容。2) 改善内容质量。AIGC 被认为是继专业生产内容（PGC）、用户生产内容（UGC）之后的新型内容生产方式。尽管 PGC 和 UGC 的内容更具多元化、个性化，但受限于激励措施和创作者自身因素影响，市场存在供给不足的现象。3) 促进产业数字化，助力数字经济发展。产业数字化是数字经济的融合部分，是传统产业应用数字技术所带来的生产数量和效率提升，其新增产出构成数字经济的重要组成部分，AIGC 为数字经济提供了重要的数据要素。

图表2：国内外公司 AIGC 相关产品



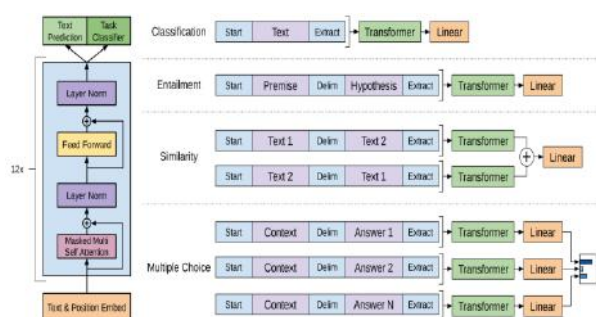
数据来源：《人工智能生成内容（AIGC）的演进历程及其图书馆智慧服务应用场景》，中信建投

ChatGPT 的爆红引发了人们对于人工智能发展的高度关注。2022 年 11 月 30 日，OpenAI 发布语言模型 ChatGPT。该模型采用对话的形式与人进行交互，可以回答后续问题、承认错误、挑战不正确的前提、拒绝不适当的请求。ChatGPT 不仅在日常对话、专业问题回答、信息检索、内容续写、文学创作、音乐创作等方面展现出强大的能力，还具有生成代码、调试代码、为代码生成注释的能力。

1.2 人工智能需要强大算力支撑

以 ChatGPT 为代表的人工智能应用在运行背后需要强大的算力支撑。OpenAI 在 2018 年推出的 GPT 参数量为 1.17 亿，预训练数据量约 5GB，而 GPT-3 参数量达 1750 亿，预训练数据量达 45TB。在模型训练阶段，ChatGPT 的总算力消耗约为 3640PF-days，总训练成本为 1200 万美元，在服务访问阶段则会有更大消耗。

图表3：GPT 模型示意图



数据来源：OpenAI，中信建投

图表4：NVIDIA DGX A100 AI 服务器



数据来源：NVIDIA，中信建投

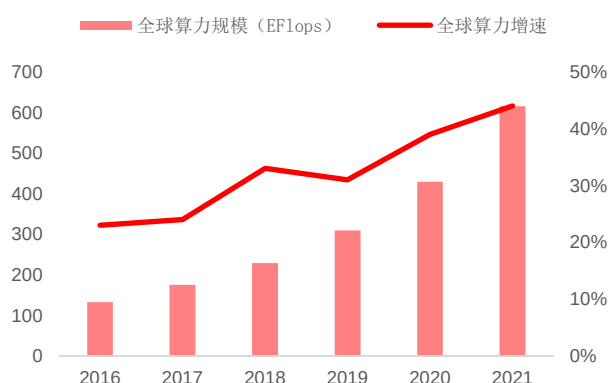
IDC 数据显示：2021 年全球人工智能 IT 投资额为 929.5 亿美元，预计 2026 年将增至 3014.3 亿美元，复合年增长率约 26.5%。2026 年中国市场 AI 投资预计将达 266.9 亿美元，约占全球投资 8.9%，居世界第二位，复合年增长率约 21.7%。未来五年，硬件将成为中国人工智能最大的细分市场，占人工智能总投资的 50% 以上。IDC 预测，2026 年，中国在人工智能硬件市场的 IT 投资将超过 150 亿美元，接近美国人工智能硬件的市场规模，五

年复合年增长率 16.5%。服务器作为硬件市场的主要组成部分，预计将占总投入的 80%以上。

人工智能的发展将对算力提出更高要求，算力网络基础设施需求有望持续提升。根据中国信通院数据，2021 年全球计算设备算力总规模达到 615EFlops(每秒浮点运算次数)，同比增长 44%，其中基础算力规模为 369EFlops，智能算力规模为 232EFlops，超算算力规模为 14EFlops，预计 2030 年全球算力规模将达到 562EFlops，平均年均增长 65%。

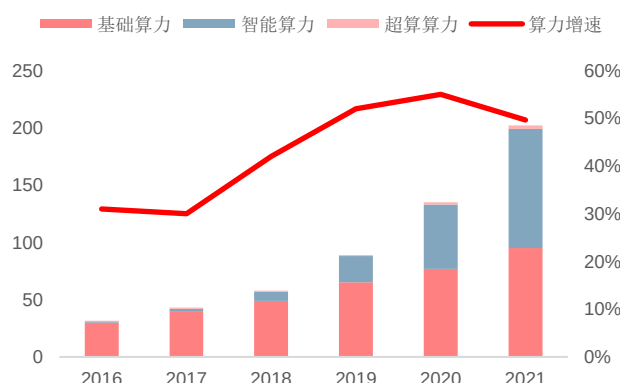
我国智能算力规模持续高速增长，2021 年智能算力规模已经超过通用算力。根据中国信通院数据，我国计算设备算力总规模达到 202EFlops，全球占比约为 33%，保持 50%以上的高速增长态势，增速高于全球，其中智能算力增长迅速，增速为 85%，在我国算力中的占比超过 50%。

图表5： 全球算力规模及增速



数据来源：中国信通院，中信建投

图表6： 我国算力规模及增速



数据来源：中国信通院，中信建投

1.3AI 算力产业链涉及环节较多，行业需求有望全面提升

AI 算力产业链涉及环节较多，按照算力基础设施构成来看，包括 AI 芯片及服务器、交换机及光模块、IDC 机房及上游产业链等。其中，随着训练和推理需求提升，AI 芯片及服务器需求将率先放量；AI 算力对数据中心内部数据流量较大，光模块速率及数量均有显著提升，交换机的端口数及端口速率也有相应的增长；IDC 也有望进入需求释放阶段，预计液冷温控渗透率将快速提升，海底数据中心也可能将迎来产业化的关键节点。

1、AI 芯片和服务器需求将率先放量

根据测算，2023 年-2027 年全球大模型训练端峰值算力需求量的年复合增长率为 78.0%。2023 年全球大模型训练端所需全部算力换算成的 A100 总量超过 200 万张。从云端推理所需算力角度测算，2023 年-2027 年，全球大模型云端推理的峰值算力需求量的年复合增长率为 113%，如果考虑边缘端 AI 推理的应用，推理端算力规模将进一步扩大。

根据 IDC 数据，2022 年全球 AI 服务器市场规模 202 亿美元，同比增长 29.8%，占服务器市场规模的比例为 16.4%，同比提升 1.2pct。我们认为全球 AI 服务器市场规模未来 3 年内将保持高速增长，市场规模分别为 395/890/1601 亿美元，对应增速 96%/125%/80%。根据 IDC 数据，2022 年中国 AI 服务器市场规模 67 亿美元，同比增长 24%。我们预计，2023-2025 年，结合对于全球 AI 服务器市场规模的预判，以及对于我国份额占比持

续提升的假设，我国 AI 服务器市场规模有望达到 134/307/561 亿美元，同比增长 101%/128%/83%。

图表7： 全球 AI 服务器市场规模测算

	2021	2022	2023E	2024E	2025E
大模型带动 GPU 存量空间（亿美元）	-	-	276.6	622.7	1120.9
GPU 占 AI 服务器成本比例（%）	-	-	70.0	70.0	70.0
GPU 芯片升级/算法效率提升比例测算（%）	-	-	100.0	120.0	150.0
AI 服务器存量规模（亿美元）	156.0	202.0	395.2	889.6	1601.3
AI 服务器增量规模（亿美元）	-	46.0	193.2	494.4	711.7
市场增速（%）	39.1	29.8	95.6	125.1	80.0

资料来源: OpenAI, IDC, Nvidia, 中信建投

图表8： 中国 AI 服务器市场规模测算

	2021	2022	2023E	2024E	2025E
全球市场规模（亿美元）	156.0	202.0	395.2	889.6	1601.3
中国市场占全球市场比重（%）	34.6	33.2	34.0	34.5	35.0
市场增速（%）	68.2	24.0	100.5	128.4	82.6
市场规模（亿美元）	54.0	67.0	134.4	306.9	560.5

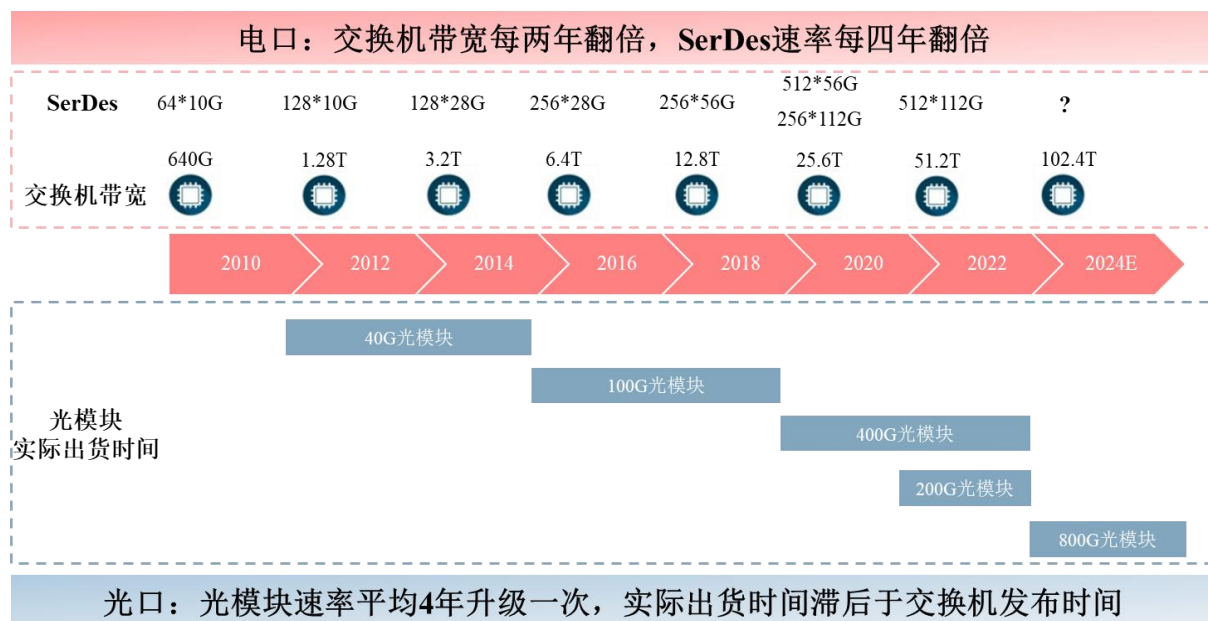
资料来源: OpenAI, IDC, Nvidia, 中信建投

2、AI 算力改变数据中心内部网络架构，光模块和交换机速率及需求提升

AI 数据中心中，由于内部数据流量较大，因此无阻塞的胖树网络架构成了重要需求之一，光模块速率及数量均有显著提升，交换机的端口数及端口速率也有相应的增长。

800G 光模块 2022 年底开始小批量出货，2023 年需求主要来自于英伟达和谷歌，2024 年有望大规模出货，并存在时间前移的可能。从交换机的电口来看，SerDes 通道的速率每四年翻倍，数量每两年翻倍，交换机的带宽每两年翻倍；从光口来看，光模块每 4 年升级一次，实际出货时间是晚于电口 SerDes 及交换机芯片新版发布的时间。2019 年作为 100G 光模块升级的时间点，市场分成了 200G 和 400G 两条升级路径。但是在 2023 年这个时间点，市场下一代高速率光模块均指向 800G 光模块，叠加 AIGC 带来的算力和模型竞赛，我们预计北美各大云厂商和相关科技巨头均有望在 2024 年大量采购 800G 光模块，同时 2023 年也可能提前采购。

图表9：光模块和交换机速率演进示意图



数据来源：思科，中信建投证券

3、IDC 需求有望释放，AI 服务器高功率密度或将推升液冷渗透率

IDC 作为算力基础设施产业链的关键环节，也有望进入需求释放阶段。在过去两年半，受多重因素影响下，云计算需求景气度下行，但 IDC 建设与供给未出现明显放缓，2021 年和 2022 年分别新增机柜数量 120 万架和 150 万架，因此短期内出现供需失衡情况（核心区域供需状况相对良好），部分地区上电率情况一般。所以 IDC 公司 2022 年业绩普遍承压。随着平台经济发展恢复以及 AI 等拉动，IDC 需求有望逐步释放，叠加 2023 新增供给量有望较 2022 年减少（例如三大运营商 2022 年新增 IDC 机柜 15.6 万架，2023 年计划新增 11.4 万架）。

人工智能大模型训练和推理运算所用的 GPU 服务器的功率密度将大幅提升，以英伟达 DGX A100 服务器为例，其单机最大功率约可以达到 6.5kW，大幅超过单台普通 CPU 服务器 500w 左右的功率水平。在此情况下，一方面需要新建超大功率的机柜，另一方面为降低 PUE，预计液冷温控渗透率将快速提升，海底数据中心也可能将迎来产业化的关键节点。

二、AI 芯片需求爆发式增长

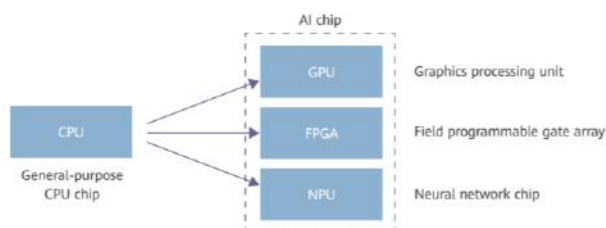
2.1 AI 大规模落地应用对 AI 芯片性能、数量提出全方位要求

从广义上讲，能运行 AI 算法的芯片都叫 AI 芯片。CPU、GPU、FPGA、NPU、ASIC 都能执行 AI 算法，但在执行效率层面上有巨大的差异。CPU 可以快速执行复杂的数学计算，但同时执行多项任务时，CPU 性能开始下降，目前行业内基本确认 CPU 不适用于 AI 计算。

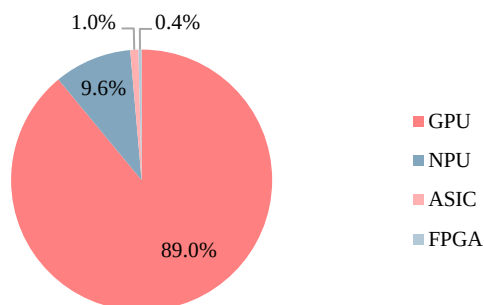
CPU+xPU 的异构方案成为大算力场景标配，GPU 为应用最广泛的 AI 芯片。目前业内广泛认同的 AI 芯片类型包括 GPU、FPGA、NPU 等。由于 CPU 负责对计算机的硬件资源进行控制调配，也要负责操作系统的运行，

在现代计算系统中仍是不可或缺的。GPU、FPGA 等芯片都是作为 CPU 的加速器而存在，因此目前主流的 AI 计算系统均为 CPU+xPU 的异构并行。CPU+GPU 是目前最流行的异构计算系统，在 HPC、图形图像处理以及 AI 训练/推理等场景为主流选择。IDC 数据显示，2021 年中国 AI 芯片市场中，GPU 市占率为 89%。

图表10：CPU+AI 芯片的异构计算



图表11：2021 年中国 AI 芯片市场规模占比



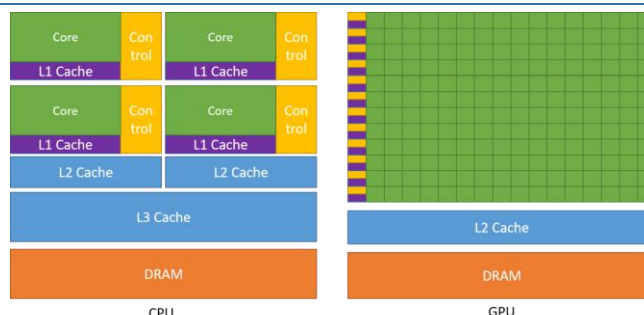
资料来源：华为，中信建投

资料来源：IDC，中信建投

2.1.1 GPU 性能、功能经历长期迭代升级，成为 AI 芯片中应用最广泛的选择

GPU 能够进行并行计算，设计初衷是加速图形渲染。NVIDIA 在 1999 年发布 GeForce 256 图形处理芯片时首先提出 GPU（Graphic Processing Unit）的概念，并将其定义为“具有集成转换、照明、三角形设置/裁剪和渲染引擎的单芯片处理器，能够每秒处理至少 1000 万个多边形”。从计算资源占比角度看，CPU 包含大量的控制单元和缓存单元，实际运算单元占比较小。GPU 则使用大量的运算单元，少量的控制单元和缓存单元。GPU 的架构使其能够进行规模化并行计算，尤其适合逻辑简单，运算量大的任务。GPU 通过从 CPU 承担一些计算密集型功能（例如渲染）来提高计算机性能，加快应用程序的处理速度，这也是 GPU 早期的功能定位。

图表12：CPU 与 GPU 架构对比



资料来源：NVIDIA，中信建投

CUDA 将 GPU 的计算能力扩展至图形处理之外，成为更通用的计算设备。在 GPU 问世以后，NVIDIA 及其竞争对手 ATI（被 AMD 收购）一直在为他们的显卡包装更多的功能。2006 年 NVIDIA 发布了 CUDA 开发环境，这是最早被广泛应用的 GPU 计算编程模型。CUDA 将 GPU 的能力向科学计算等领域开放，标志着 GPU 成为一种更通用的计算设备 GPGPU（General Purpose GPU）。NVIDIA 也在之后推出了面向数据中心的 GPU 产品线。

图表13: NVIDIA GPU 主要产品线

产品线	定位	应用场景	代表型号
GeForce	计算机的图形处理和游戏运行	消费者应用的中高端 PC 市场	GeForce RTX 4090 GeForce RTX 4080
NVIDIA RTX/Quadro	专业视觉计算平台	建筑设计、媒体与娱乐等行业专业用户的 PC、工作站	NVIDIA RTX A6000 Quadro GV100
Data Center	数据中心加速计算平台	AI、数据分析、高性能计算(HPC)	NVIDIA H100 NVIDIA A100

资料来源: NVIDIA, 中信建投

GPU 性能提升与功能丰富逐步满足 AI 运算需要。2010 年 NVIDIA 提出的 Fermi 架构是首个完整的 GPU 计算架构,其中提出的许多新概念沿用至今。Kepler 架构在硬件上拥有了双精度计算单元(FP64),并提出 GPU Direct 技术,绕过 CPU/System Memory,与其他 GPU 直接进行数据交互。Pascal 架构应用了第一代 NVLink。Volta 架构开始应用 Tensor Core,对 AI 计算加速具有重要意义。简要回顾 NVIDIA GPU 硬件变革历程,工艺、计算核心数增加等基础特性的升级持续推动性能提升,同时每一代架构所包含的功能特性也在不断丰富,逐渐更好地适配 AI 运算的需要。

图表14: NVIDIA Fermi 架构至 Hopper 架构的变化

架构	Fermi	Kepler	Maxwell	Pascal	Volta	Turing	Ampere	Hopper
发布时间	2010	2012	2014	2016	2017	2018	2020	2022
工艺	40/28nm	28nm	28nm	16nm	12nm	12nm	8/7nm	4nm
SMTs	16	15	16	60	80	92	108	132
Cuda Cores	512	1536	3072	3584	5120	2560	6912	16896
Tensor Core	/	/	/	/	640	320	432	528
特点	首个完整 GPU 计算架构			NVLink 1.0	NVLink 2.0 Tensor Core 1.0	Tensor Core 2.0 RT Core 1.0	Tensor Core 3.0, NVLink 3.0	Tensor Core 4.0, NVLink 4.0

资料来源: NVIDIA, 中信建投

AI 的数据来源广泛, GPU 逐渐实现对各类数据类型的支持。依照精度差异,算力可从 INT8(整数类型)、FP16(半精度)、FP32(单精度)、FP64(双精度)等不同维度对比。AI 应用处理的数据包括文字、图片或视频,数据精度类型差异大。对于数据表征来讲,精度越高,准确性越高;但降低精度可以节省运算时间,减少成本。总体来看,精度的选择需要在准确度、成本、时间之间取得平衡。目前许多 AI 模型中运行半精度甚至整形计算即可完成符合准确度的推理和训练。随着架构的迭代, NVIDIA GPU 能支持的数据类型持续丰富,例如 Turing 架构 T4 开始支持 INT8, Ampere 架构 A100 的 Tensor Core 开始支持 TF32。

图表15：低精度比特位宽为 AI 计算带来的好处



资料来源：Intel，中信建投

图表16：不同精度计算消耗的能量和硅片面积

计算精度及操作	能量消耗相对值	面积消耗相对值
8b Add	1	1
16b Add	2	2
32b Add	3	4
16b FP Add	13	38
32b FP Add	30	116
8b Mult	7	8
32b Mult	103	97
16b FP Mult	37	46
32b FP Mult	123	214
32b SRAM Read (8KB)	167	-
32b DRAM Read	21333	-

资料来源：《Efficient Method and Hardware for Deep Learning》，中信建投

图表17：NVIDIA 数据中心 GPU 支持的比特位宽变化

	Suppoted CUDA Core Precisions										Suppoted Tensor Core Precisions									
	FP8	FP16	FP32	FP64	INT1	INT4	INT8	TF32	BF16	FP8	FP16	FP32	FP64	INT1	INT4	INT8	TF32	BF16		
NVIDIA Tesla P4	NO	NO	YES	YES	NO	NO	YES	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO		
NVIDIA P100	NO	YES	YES	YES	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO		
NVIDIA Volta	NO	YES	YES	YES	NO	NO	YES	NO	NO	NO	YES	NO	NO	NO	NO	NO	NO	NO		
NVIDIA Turing	NO	YES	YES	YES	NO	NO	YES	NO	NO	NO	YES	NO	NO	YES	YES	YES	NO	NO		
NVIDIA A100	NO	YES	NO	YES	NO	NO	YES	YES	YES	NO	YES	NO	YES	YES	YES	YES	YES	YES		
NVIDIA H100	NO	YES	NO	YES	NO	NO	YES	YES	YES	YES	YES	NO	YES	NO	NO	YES	YES	YES		

资料来源：NVIDIA，中信建投

均衡分配资源的前提下，处理低精度的硬件单元数量更多，表现更高的算力性能。GPU 作为加速器得到广泛应用一定程度上得益于它的通用性，为了在不同精度的数据类型上具有良好的性能，以兼顾 AI、科学计算等不同场景的需要，英伟达在分配处理不同数据类型的硬件单元时大体上保持均衡。因为低精度数据类型的计算占用更少的硬件资源，同一款 GPU 中的处理低精度数据类型的硬件单元的数量较多，对应计算能力也较强。以 V100 为例，每个 SM 中 FP32 单元的数量都为 FP64 单元的两倍，最终 V100 的 FP32 算力（15.7 TFLOPS）也近似为 FP64（7.8 TFLOPS）的两倍，类似的规律也可以在各代架构旗舰 P100、A100 和 H100 中看到。

图表18：V100 中 FP32 硬件单元和 FP64 硬件单元的数量关系

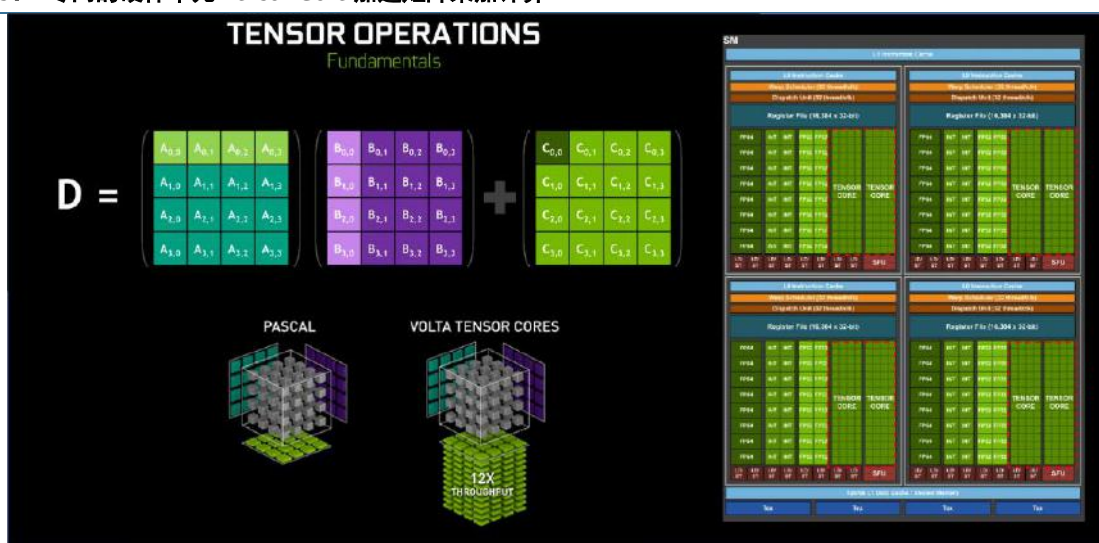


资料来源：NVIDIA，中信建投

GPU 引入特殊硬件单元加速 AI 的核心运算环节。矩阵-矩阵乘法（GEMM）运算是神经网络训练和推理的

核心，本质是在网络互连层中将大矩阵输入数据和权重相乘。矩阵乘积的求解过程需要大量的乘积累加操作，而 FMA (Fused Multiply-accumulate operation, 融合乘加) 可以消耗更少的时钟周期来完成这一过程。传统 CUDA Core 执行 FMA 指令，硬件层面需要将数据按寄存器->ALU->寄存器->ALU->寄存器的方式来回搬运。2017 年发布的 Volta 架构首度引入了 Tensor Core (张量核心)，是由 NVIDIA 研发的新型处理核心。根据 NVIDIA 数据，Volta Tensor Core 可以在一个 GPU 时钟周期内执行 $4 \times 4 \times 4 = 64$ 次 FMA 操作，吞吐量是 Pascal 架构下 CUDA Core 的 12 倍。

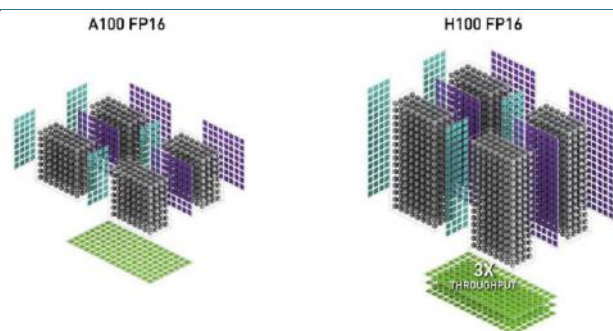
图表19：专门的硬件单元 Tensor Core 加速矩阵乘加计算



资料来源：NVIDIA，中信建投

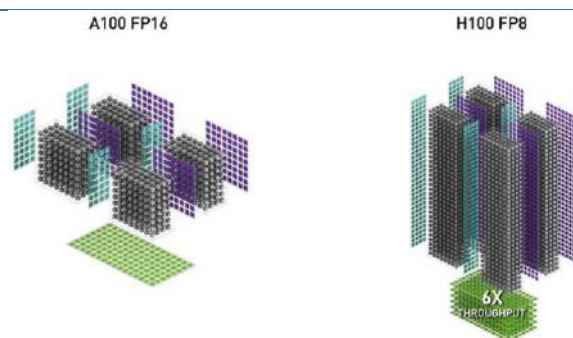
Tensor Core 持续迭代提升其加速能力。Volta 架构引入 Tensor Core 的改动使 GPU 的 AI 算力有了明显提升，后续在每一代的架构升级中，Tensor Core 都有比较大的改进，支持的数据类型也逐渐增多。以 A100 到 H100 为例，Tensor Core 由 3.0 迭代至 4.0，H100 在 FP16 Tensor Core 的峰值吞吐量提升至 A100 的 3 倍。同时，H100 Tensor Core 支持新的数据类型 FP8，H100 FP8 Tensor Core 的吞吐量是 A100 FP16 Tensor Core 的 6 倍。

图表20：A100 与 H100 的 FP16 Tensor Core 吞吐量对比



资料来源：NVIDIA，中信建投

图表21：FP16 Tensor Core 与 FP8 Tensor Core 吞吐量对比

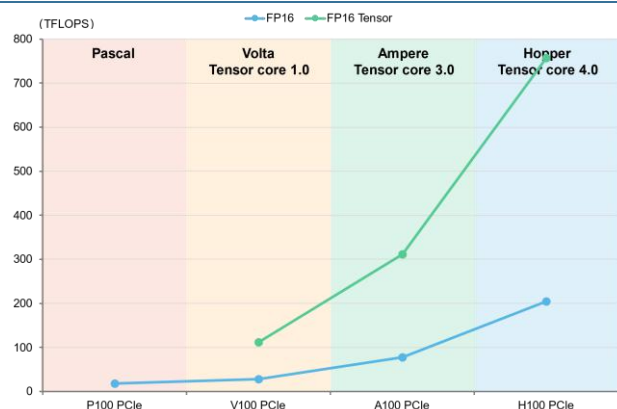


资料来源：NVIDIA，中信建投

Tensor Core 加速下，低精度比特位宽的算力爆发式增长，契合 AI 计算需要。Tensor Core 的应用使算力快速、高效增长，选取 Pascal 至 Hopper 架构时期每一代的旗舰数据中心显卡，对比经 Tensor Core 加速前后的 FP16 算力指标可以得到：（1）经 Tensor Core 加速的 FP16 算力明显高于加速之前。（2）每单位 Tensor core 支持的算

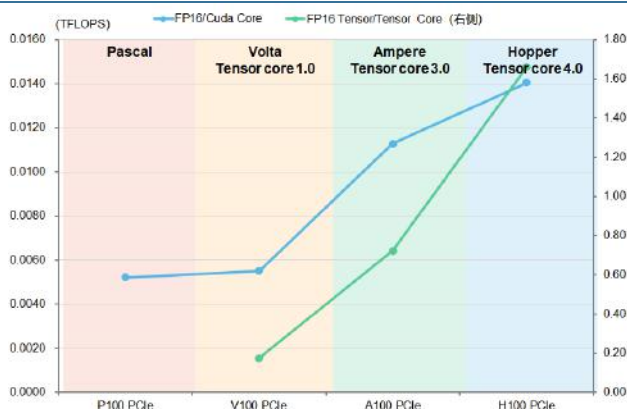
力明显高于每单位 Cuda Core 支持的算力。同时，Tensor Core 从 2017 年推出以来首先完善了对低精度数据类型的支持，顺应了 AI 发展的需要。

图表22： FP16 Tensor 算力快速增长



资料来源：NVIDIA, techpowerup, 中信建投

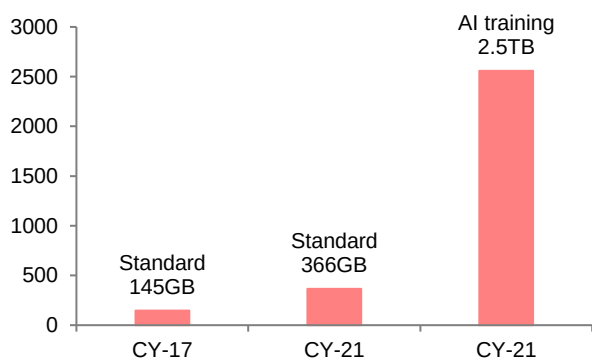
图表23： FP16 Tensor 每单位核心的算力明显优于 FP16



资料来源：NVIDIA, techpowerup, 中信建投

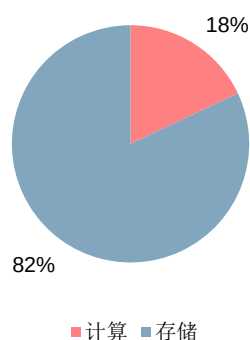
数据访问支配着计算能力利用率。AI 运算涉及到大量数据的存储与处理，根据 Cadence 数据，与一般工作负载相比，每台 AI 训练服务器需要 6 倍的内存容量。而在过去几十年中，处理器的运行速度随着摩尔定律高速提升，而 DRAM 的性能提升速度远远慢于处理器速度。目前 DRAM 的性能已经成为了整体计算机性能的一个重要瓶颈，即所谓阻碍性能提升的“内存墙”。除了性能之外，内存对于能效比的限制也成为一个瓶颈，Cadence 数据显示，在自然语言类 AI 负载中，存储消耗的能量占比达到 82%。

图表24： AI 训练服务器需要更高的内存容量



资料来源：Cadence, 中信建投

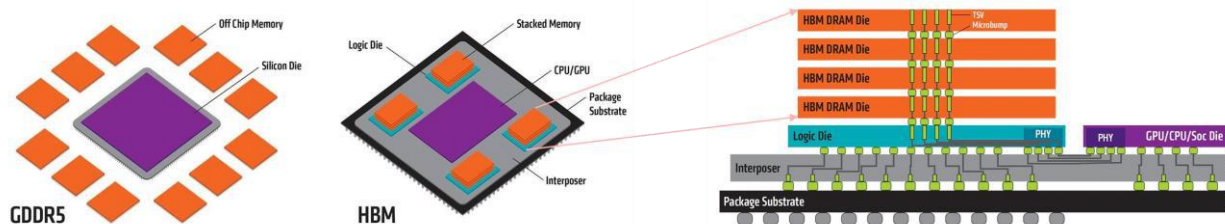
图表25： NLP 负载中存储和计算的能耗占比



资料来源：Cadence, 中信建投

GPU 采用高带宽 HBM 降低“内存墙”影响。为防止占用系统内存并提供较高的带宽和较低的延时，GPU 均配备有独立的内存。常规的 GDDR 焊接在 GPU 芯片周边的 PCB 板上，与处理器之间的数据传输速率慢，并且存储容量小，成为运算速度提升的瓶颈。HBM 裸片通过 TSV 进行堆叠，然后 HBM 整体与 GPU 核心通过中介层互连，因此 HBM 获得了极高的带宽，并节省了 PCB 面积。目前，GDDR 显存仍是消费级 GPU 的行业标准，HBM 则成为数据中心 GPU 的主流选择。

图表26: GDDR 与 HBM 差异



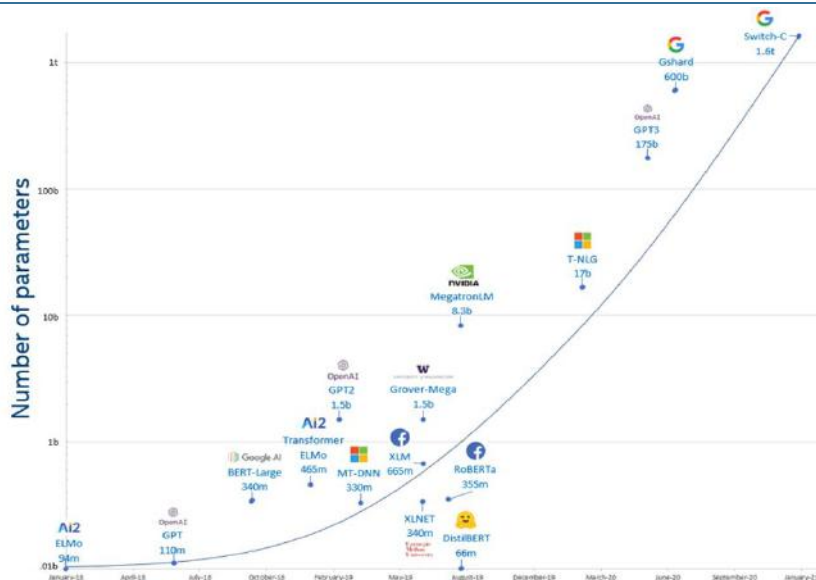
资料来源: NVIDIA, 中信建投

硬件单元的改进与显存升级增强了单张 GPU 算力的释放, 然而, 随着 Transformer 模型的大规模发展和应用, 模型参数量呈爆炸式增长, GPT-3 参数量达到了 1750 亿, 相比 GPT 增长了近 1500 倍, 预训练数据量更是从 5GB 提升到了 45TB。大模型参数数量的指数级增长带来的诸多问题使 GPU 集群化运算成为必须:

(1) 即使最先进的 GPU, 也不再可能将模型参数拟合到主内存中。

(2) 即使模型可以安装在单个 GPU 中 (例如, 通过在主机和设备内存之间交换参数), 所需的大量计算操作也可能导致在没有并行化的情况下不切实际地延长训练时间。根据 NVIDIA 数据, 在 8 个 V100 GPU 上训练一个具有 1750 亿个参数的 GPT-3 模型需要 36 年, 而在 512 个 V100 GPU 上训练需要 7 个月。

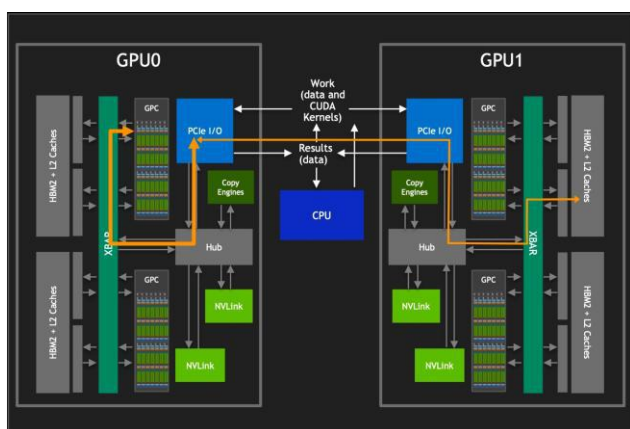
图表27: 语言模型的参数量呈指数级增长



资料来源: HEITS.DIGITAL, 中信建投

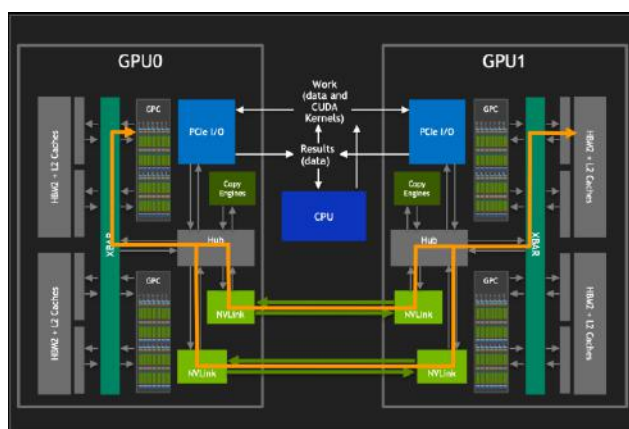
NVIDIA 开发 NVLink 技术解决 GPU 集群通信。在硬件端, GPU 之间稳定、高速的通信是实现集群运算所必须的条件。传统 x86 服务器的互连通道 PCIe 的互连带宽由其代际与结构决定, 例如 x16 PCIe 4.0 双向带宽仅为 64GB/s。除此之外, GPU 之间通过 PCIe 交互还会与总线上的 CPU 操作竞争, 甚至进一步占用可用带宽。NVIDIA 为突破 PCIe 互连的带宽限制, 在 P100 上搭载了首项高速 GPU 互连技术 NVLink (一种总线及通讯协议), GPU 之间无需再通过 PCIe 进行交互。

图表28： GPU 之间通过 PCIe 连接



资料来源：HEITS.DIGITAL，中信建投

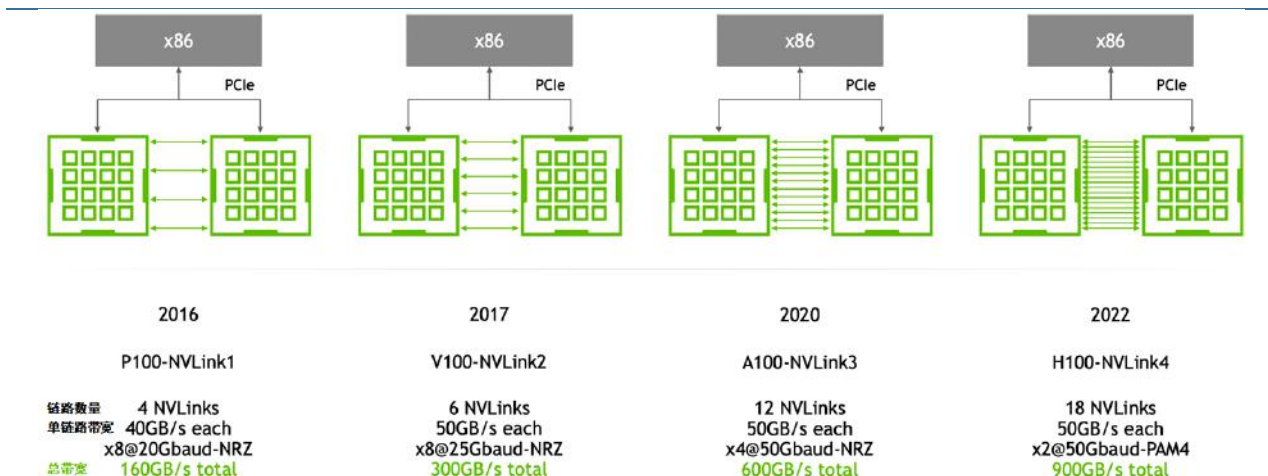
图表29： GPU 之间通过 NVLink 连接



资料来源：NVIDIA，中信建投

NVLink 继续与 NVIDIA GPU 架构同步发展，每一种新架构都伴随着新一代 NVLink。第四代 NVLink 为每个 GPU 提供 900 GB/s 的双向带宽，比上一代高 1.5 倍，比第一代 NVLink 高 5.6 倍。

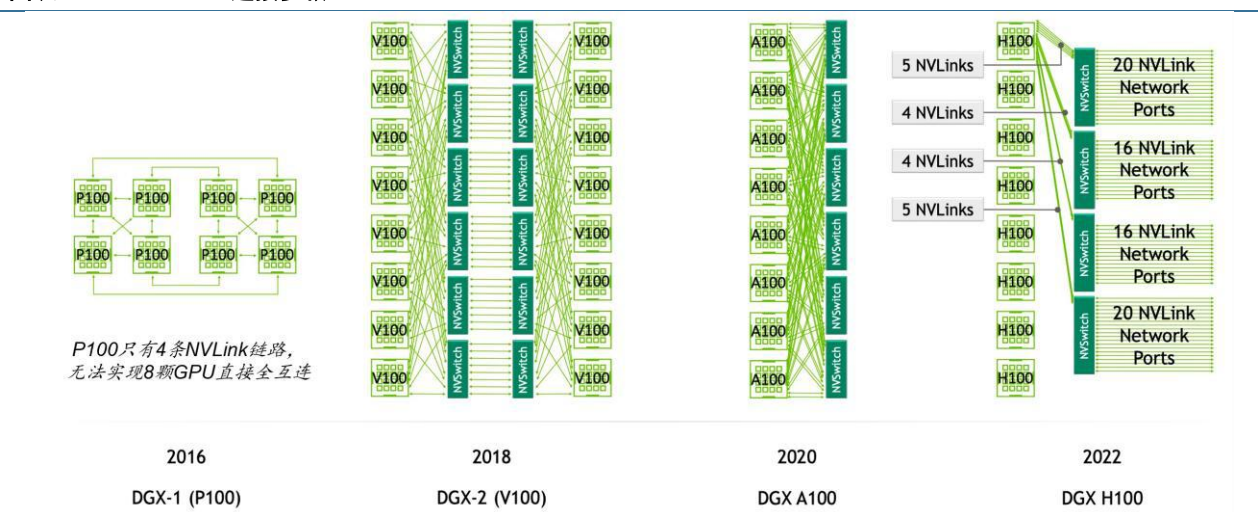
图表30： NVLink 1.0—NVLink 4.0



资料来源：NVIDIA，中信建投

NVIDIA 开发基于 NVLink 的芯片 NVSwitch，作为 GPU 集群数据通信的“枢纽”。NVLink 1.0 技术使用时，一台服务器中的 8 个 GPU 无法全部实现直接互连。同时，当 GPU 数量增加时，仅依靠 NVLink 技术，需要众多数量的总线。为解决上述问题，NVIDIA 在 NVLink 2.0 时期发布了 NVSwitch，实现了 NVLink 的全连接。NVSwitch 是一款 GPU 桥接芯片，可提供所需的 NVLink 交叉网络，在 GPU 之间的通信中发挥“枢纽”作用。借助于 NVswitch，每颗 GPU 都能以相同的延迟和速度访问其它的 GPU。就程序来看，16 个 GPU 都被视为一个 GPU，系统效率得到了最大化，大大降低了多 GPU 系统的优化难度。

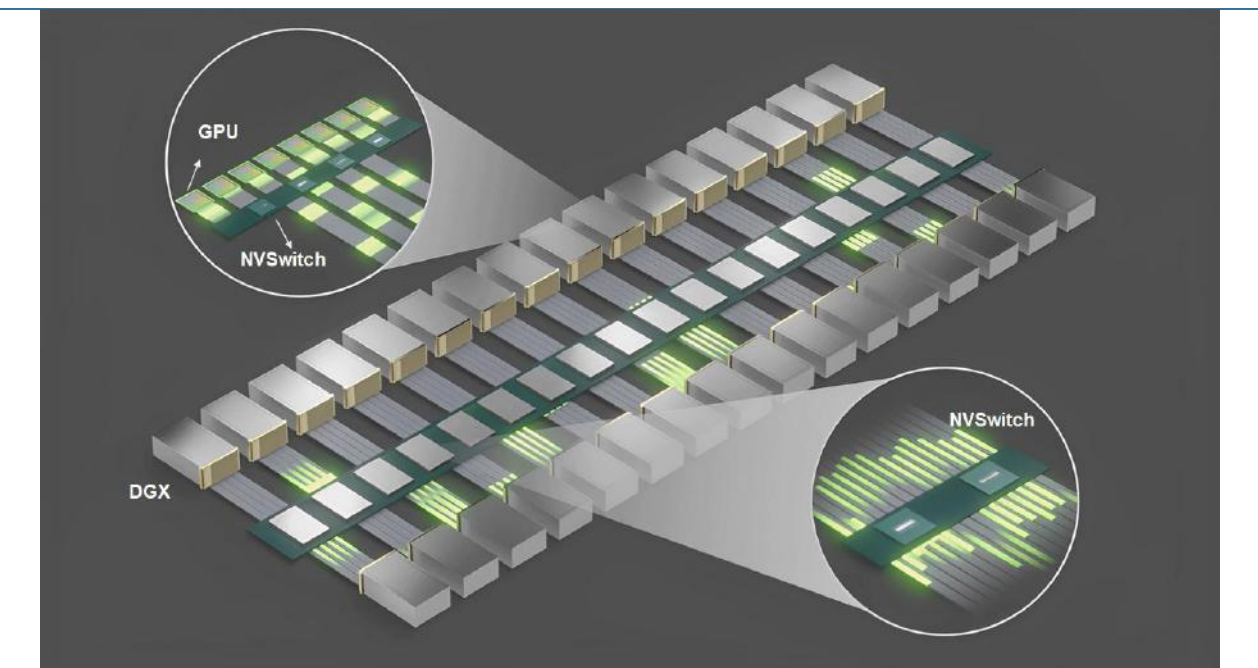
图表31: NVSwitch 连接多颗 GPU



资料来源: NVIDIA, 中信建投

通过添加更多 NVSwitch 来支持更多 GPU, 集群分布式运算得以实现。当训练大型语言模型时, NVLink 网络也可以提供显著的提升。NVSwitch 已成为高性能计算(HPC)和 AI 训练应用中不可或缺的一部分。

图表32: NVSwitch 支撑的 GPU 计算集群



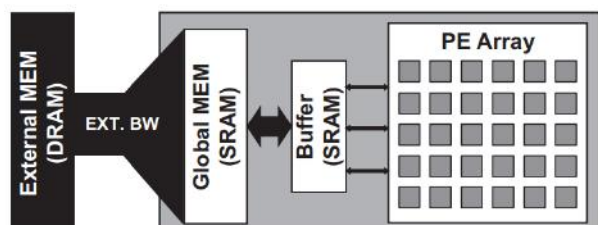
资料来源: NVIDIA, 中信建投

2.1.2 NPU 通过特殊架构设计对 AI 运算起到加速作用

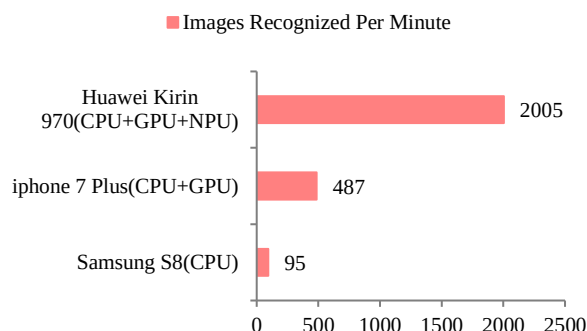
NPU 在人工智能算法上具有较高的运行效率。为了适应某个特定领域中的常见的应用和算法而设计, 通常称之为“特定域架构 (Domain Specific Architecture, DSA)”芯片, NPU (神经网络处理器) 属于其中一种, 常被设计用于神经网络运算的加速。以华为手机 SoC 麒麟 970 为例, NPU 对图像识别神经网络的运算起到了显著

加速效果，使其图像识别速度明显优于同代竞品的表现。

图表33： NPU 典型架构



图表34： 麒麟 970 NPU 加速图像识别



资料来源：《Architecture of neural processing unit for deep neural networks》，中信建投

资料来源：THE TECH REVOLUTIONIST，中信建投

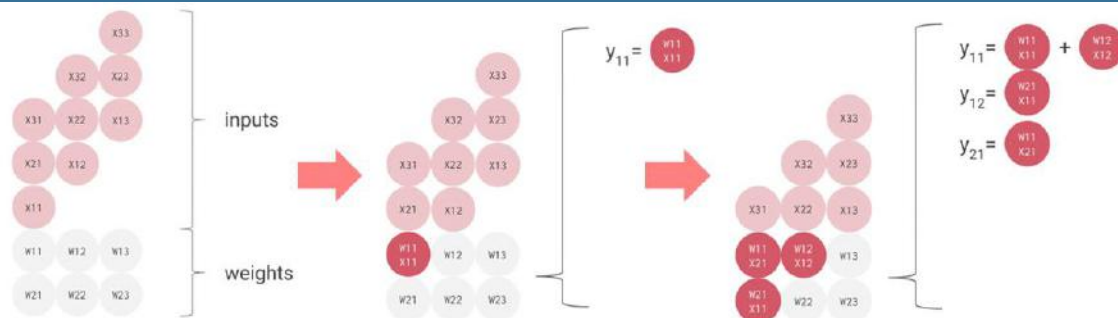
目前已量产的 NPU 或搭载 NPU 模块的芯片众多，其他知名的芯片包括谷歌 TPU、华为昇腾、特斯拉 FSD、特斯拉 Dojo 等。各家厂商在计算核心的设计上有其差异，例如谷歌 TPU 的脉动阵列，华为昇腾的达芬奇架构。

以谷歌 TPU 及计算核心结构脉动阵列为例，对比其相较于 CPU、GPU 的区别：

CPU 和 GPU 均具有通用性，但以频繁的内存访问导致资源消耗为代价。CPU 和 GPU 都是通用处理器，可以支持数百万种不同的应用程序和软件。对于 ALU 中的每一次计算，CPU、GPU 都需要访问寄存器或缓存来读取和存储中间计算结果。由于数据存取的速度往往大大低于数据处理的速度，频繁的内存访问，限制了总吞吐量并消耗大量能源。

谷歌 TPU 并非通用处理器，而是将其设计为专门用于神经网络工作负载的矩阵处理器。TPU 不能运行文字处理器、控制火箭引擎或执行银行交易，但它们可以处理神经网络的大量乘法和加法，速度极快，同时消耗更少的能量，占用更小的物理空间。TPU 内部设计了由乘法器和加法器构成的脉动阵列。在计算时，TPU 将内存中的参数加载到乘法器和加法器矩阵中，每次乘法执行时，结果将传递给下一个乘法器，同时进行求和。所以输出将是数据和参数之间所有乘法结果的总和。在整个海量计算和数据传递过程中，完全不需要访问内存。这就是为什么 TPU 可以在神经网络计算上以低得多的功耗和更小的占用空间实现高计算吞吐量。

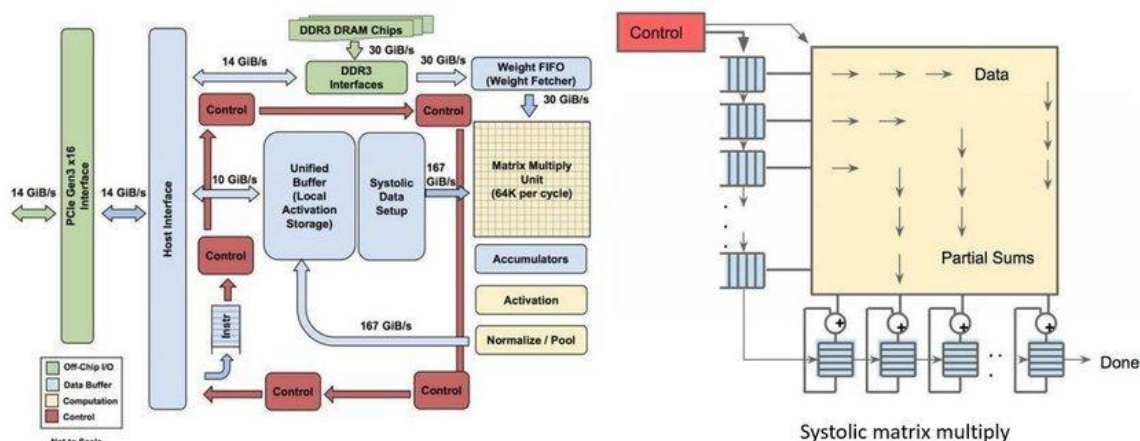
图表35： 脉动阵列运行矩阵乘法的示意图



资料来源：谷歌，中信建投

脉动阵列本质上是在硬件层面多次重用输入数据，在消耗较小的内存带宽的情况下实现较高的运算吞吐率。脉动阵列结构简单，实现成本低，但它灵活性较差，只适合特定运算。然而，AI 神经网络需要大量卷积运算，卷积运算又通过矩阵乘加实现，正是脉动阵列所适合的特定运算类型。脉动阵列理论最早在 1982 年提出，自谷歌 2017 年首次将其应用于 AI 芯片 TPU 中，这项沉寂多年的技术重回大众视野，多家公司也加入了脉动阵列行列，在自家加速硬件中集成了脉动阵列单元。

图表36： 谷歌 TPU 架构及其内部的脉动阵列



资料来源：谷歌，中信建投

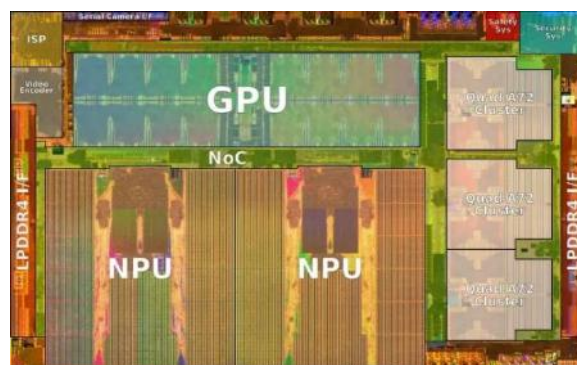
NPU 已经在 AI 运算加速领域获得了广泛应用。在数据中心获得大规模应用的 NPU 案例即 TPU，已被谷歌用于构建数据中心的超级计算机，执行特定神经网络的训练任务。在用户端，手机、汽车、智能安防摄像头等设备开始搭载 AI 计算功能，通常是利用训练好的神经网络模型执行图像处理等工作，此时 NPU 通用性差的劣势被缩小，高算力、高能耗比的优势被放大，因而得到了广泛的应用。在终端设备中，NPU 常以模块的形式包含在 SoC 内部，对 AI 运算进行加速，例如特斯拉自动驾驶芯片 FSD 均包含 NPU。

图表37： 谷歌 TPU



资料来源：谷歌，中信建投

图表38： Tesla FSD 搭载 NPU 模块

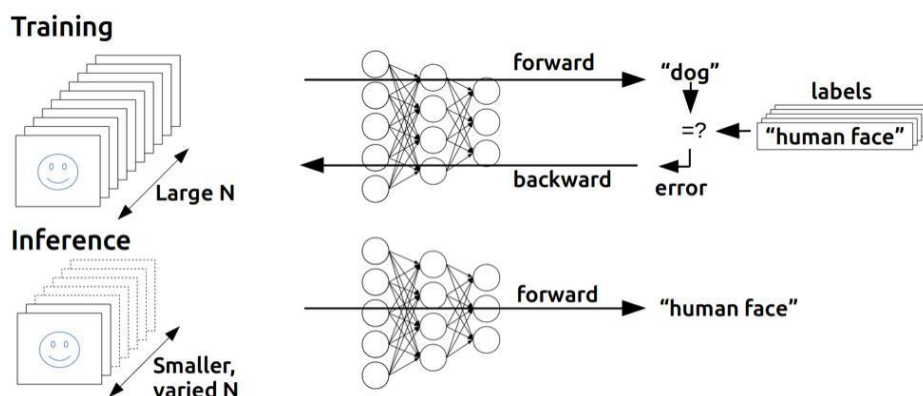


资料来源：Wikichip，中信建投

2.1.3 训练/推理、云/边分别对 AI 芯片提出不同要求，未来推理端的算力需求将远超训练端

AI 技术在实际应用中包括两个环节：训练(Training)和推理(Inference)。训练是指通过大数据训练出一个复杂的神经网络模型，使其能够适应特定的功能。训练需要较高的计算性能、能够处理海量数据、具有一定的通用性。推理是指利用训练好的神经网络模型进行运算，利用输入的新数据来一次性获得正确结论的过程。

图表39: AI 训练与 AI 推理对比



资料来源: NVIDIA, 中信建投

根据所承担任务的不同, AI 芯片可以分为训练 AI 芯片和推理 AI 芯片:

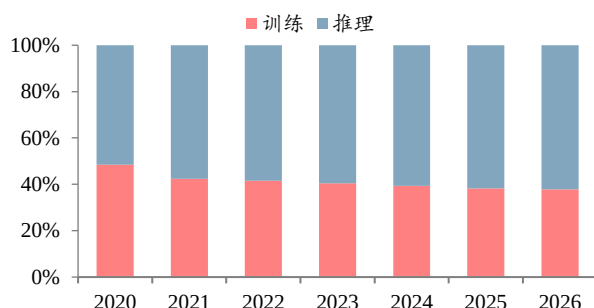
- (1) 训练芯片: 用于构建神经网络模型, 需要高算力和一定的通用性。
- (2) 推理芯片: 利用神经网络模型进行推理预测, 注重综合指标, 单位能耗算力、时延、成本等都要考虑。

根据 AI 芯片部署的位置, 可以分为云端 AI 芯片和边缘端 AI 芯片:

- (1) 云端: 即数据中心, 关注算力、扩展能力、兼容性。云端部署的 AI 芯片包括训练芯片和推理芯片。
- (2) 边缘端: 即手机、安防摄像头等领域, 关注综合性能, 要求低功耗、低延时、低成本。边缘端部署的 AI 芯片以实现推理功能为主。

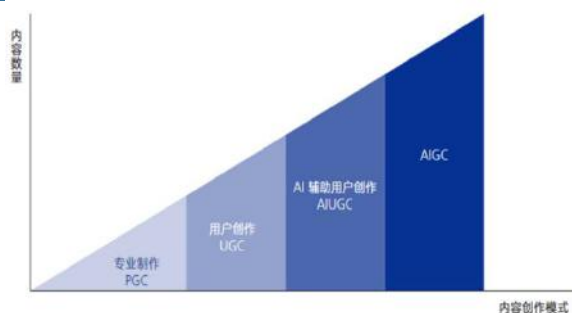
云端推理占比逐步提升, AI 落地应用数量增加。根据 IDC 数据, 随着人工智能进入大规模落地应用的关键时期, 2022 年在云端部署的算力里, 推理占算力已经达到了 58.5%, 训练占算力只有 41.5%, 预计到 2026 年, 推理占到 62.2%, 训练占 37.8%。云端推理占比逐步提升说明, AI 落地应用数量正在不断增加, 人工智能模型将逐步进入广泛投产模式。

图表40: 云端推理占比逐步提升



资料来源: IDC, 中信建投

图表41: AIGC 引发内容生成范式革命



资料来源: 腾讯研究院, 中信建投

目前 GPU 为云端 AI 训练应用的首选, 也有专门面向推理需求设计的 GPU。在云端训练场景, GPU 兼顾通用性和高算力, 同时具有完善的软件生态便于开发, 目前占据主导。云端训练 GPU 常用的型号例如 V100、

A100、H100，上述型号在多种比特位宽具有高算力表现，互连带宽性能也能满足集群分布式训练的需要。在云端推理场景，A100、H100 等型号亦可应用，英伟达也设计了面向推理市场的 T4、A10 等 GPU，这一类型号的性能相比同代旗舰有所下降，但仍具有良好的低精度比特位宽满足 AI 推理的需要，可以满足客户对能耗、成本的综合考虑。

图表42： NVIDIA 云端训练 GPU 与推理 GPU 参数对比

市场定位	训练/推理			推理		
型号	V100 SXM	A100 SXM	H100 SXM	T4 PCIe	A10 PCIe	A30 PCIe
发布时间	2017	2020	2022	2018	2021	2021
制程	12nm	7nm	4nm	12nm	8nm	7nm
FP64	7.8 TFLOPS	9.7 TFLOPS	34 TFLOPS	0.25 TFLOPS	0.97 TFLOPS	5.2 TFLOPS
FP32	15.7 TFLOPS	19.5 TFLOPS	67 TFLOPS	8.1 TFLOPS	31.2 TFLOPS	10.3 TFLOPS
FP16	31.3 TFLOPS	78 TFLOPS	267.6 TFLOPS	-	-	10.3 TFLOPS
FP64 Tensor	-	9.75 TFLOPS	67 TFLOPS	-	-	10.3 TFLOPS
TF32 Tensor	-	156 TFLOPS	495 TFLOPS	-	62.5 TFLOPS	82 TFLOPS
BF16 Tensor	-	312 TFLOPS	990 TFLOPS	-	125 TFLOPS	165 TFLOPS
FP16 Tensor	125 TFLOPS	312 TFLOPS	990 TFLOPS	65 TFLOPS	125 TFLOPS	165 TFLOPS
INT8 Tensor	-	624 TOPS	1979 TOPS	130 TOPS	250 TOPS	330 TOPS
显存类型	HBM2	HBM2e	HBM3	GDDR6	GDDR6	HBM2
显存容量	16/32 GB	40/80 GB	80 GB	16 GB	24 GB	24 GB
显存带宽	900 GB/s	1.56/2.04 TB/s	3.35 TB/s	200 GB/s	600 GB/s	933 GB/s
NVLink	Gen2: 300GB/s	Gen3: 600GB/s	Gen4: 900GB/s	-	-	Gen3:200 GB/s
PCIe	Gen3: 32 GB/s	Gen4: 64 GB/s	Gen5:128 GB/s	Gen3:32 GB/s	Gen4:64 GB/s	Gen4:64 GB/s
TDP	300 W	400 W	700 W	70 W	150 W	165 W

资料来源：NVIDIA，techpowerup，中信建投

带宽、互连速率的限制，使云端超大规模模型推理选择 A100、H100 更优，而非 T4、A10 等推理卡。以 GPT-3 为例，OpenAI 数据显示 GPT-3 模型 1750 亿参数对应超过 350GB 的 GPU 显存需求。假设参数规模与所需显存呈线性关系，且推理的中间参数量按 1 倍估算，则 1 万亿参数规模的大模型推理需要约 4000GB 显存，则需要 50 张 A100（80GB）或者 167 张 A10（24GB）。集群中的 GPU 数量越多意味着更复杂的互连要求，而且 A10 无法应用 NVLink 和 NVSwitch 技术，大量 A10 组成的集群仅依靠 PCIe 通信，互连带宽相比 A100 等显卡的劣势明显，进而可能导致模型推理的时效性不佳。

图表43： 不同规模大模型所需的显存容量估计

参数量（亿）	模型显存需求（E）	推理显存需求（E）	不同型号 GPU 的需求量（E）		
			A100 80GB	A800 80GB	A10 24GB
1750	350GB	700GB	9	9	30
10000	2000GB	4000GB	50	50	167
互连性能备注			NVLink 600GB/s	NVLink 400GB/s	PCIe : 64GB/s

资料来源：OpenAI，中信建投

边缘端靠近数据源头，需求复杂致使 AI 芯片种类丰富多样。边缘端 AI 以推理任务为主，边缘 AI 芯片的特点是靠近数据源头，就近为终端设备提供 AI 算力，减少了网络通信延迟，并不代表算力需求一定弱。边缘 AI 芯片通常要求更为多样化，要求保证具体应用场景的高能效、低延迟、低成本等要求，复杂的需求场景导致边缘 AI 芯片的种类丰富多样。目前边缘端的模型小到使用 CPU 做 AI 运算即可，或大到借助 AI 芯片进行运算加速，INT8 算力从几 TOPS 到几百 TOPS 不等。边缘端 AI 推理芯片依然遵循 CPU+GPU 的异构方案，并由于空间制约多以 SoC 的形式出现，GPU、FPGA、NPU、ASIC 则作为加速模块布置于其中。例如英伟达 Jetson Xavier 内含 Volta 架构 GPU，苹果 M2 配备 NPU 模块。在边缘端的小算力场景，GPU 的功耗较大，NPU 具有较强的竞争力。

图表44：边缘端 AI 推理芯片及其算力案例

	瑞芯微 RK3588M	苹果 M2	三星 Exynos 2100	NVIDIA Xavier	Tesla HW 3.0
AI 芯片/模块类型	NPU	NPU	NPU	GPU	NPU
AI 算力	6 TOPS	15.8 TOPS	26 TOPS	32 TOPS	2 * 36.86 TOPS
应用	智能座舱及 ADAS	平板、PC	手机	自动驾驶、机器人	自动驾驶

资料来源：各公司官网，中信建投

经测算，AI 大模型在训练端和推理端都将产生巨量的算力/AI 芯片需求。如果未来大模型广泛商用落地，推理端的算力/AI 芯片的需求量将明显高于训练端。

大模型云端训练对算力的需求测算：

测算原理：从模型的（1）参数规模入手，根据（2）训练大模型所需的 Token 数量和（3）每 Token 训练成本与模型参数量的关系估算总算力需求，再考虑（4）单张 GPU 算力和（5）GPU 集群的算力利用率推导出 GPU 总需求。

（1）参数规模：过去几年，大模型的参数量呈指数上升，GPT-3 模型参数量已达到 1750 亿。GPT-4 具有多模态能力，其参数量相比 GPT-3 会更大。我们在测算中假设 2023 年多模态大模型的平均参数量达到 10000 亿个，之后每年保持 20%的增速；普通大模型的平均参数量达到 2000 亿个，之后每年保持 20%的增速。

（2）训练大模型所需的 Token 数量：参数规模在千亿量级的自然语言大模型 GPT-3、Jurassic-1、Gopher、MT-NLG，训练所需的 Token 数量在千亿量级，而一些多模态大模型在训练过程中所需 Token 数据量也跟随参数量增长而增长，我们在测算中假设多模态大模型训练所需 Token 数量达到万亿级别，并且 Token 数量与模型参数规模保持线性增长关系。

图表45：大模型参数量及训练所需 Tokens

	年份	参数量	训练 Tokens
GPT-3	2020	1750 亿	3000 亿
Jurassic-1	2021	1780 亿	3000 亿
Gopher	2022	2800 亿	3000 亿
MT-NLG	2022	5300 亿	2700 亿

资料来源：《Training Compute-Optimal Language Models》，中信建投

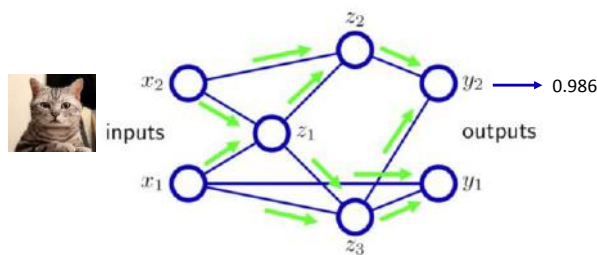
（3）每 Token 训练成本与模型参数量的关系：参考 OpenAI 发布的论文《Scaling Laws for Neural Language Models》中的分析，每个 token 的训练成本通常约为 $6N$ ，其中 N 是 LLM 的参数数量，我们在测算中遵循这一关系。具体原理如下，神经网络的训练过程包括前向传播和反向传播两个过程，其中大致包括四个步骤：

1. 做一个单次的推理操作，得到输出 y ，例如输入猫的图片得到输出 0.986。
2. 求到输出 y 与真实的目标输出 Y （假定设置的目标输出 $Y=1$ ）之间的差值 σ ，例如得到输出与目标真实值的差值为 0.014。
3. 将输出差值回溯，计算差值关于每个参数的梯度关系。
4. 根据输出差值和梯度修正每个神经元的参数，实现神经网络的参数更新，促使输出逼近目标真实值。

因而在一个参数量为 N 的神经网络中，一次输入带来训练过程的整体运算量大致为 $6N$ ，其中 $2N$ 为前向传

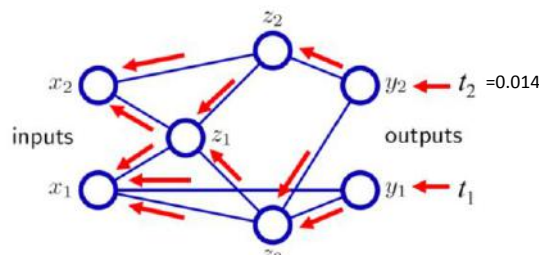
播过程，4N 为反向传播过程。

图表46：神经网络的前向传播过程



资料来源: jameskle, 中信建投

图表47：神经网络的反向传播过程



资料来源: jameskle, 中信建投

(4) **单张 GPU 算力**: 因为在训练大模型时, 主要依赖可实现的混合精度 FP16/FP32 FLOPS, 即 FP16 Tensor Core 的算力, 我们在测算中选取 A100 SXM 和 H100 SXM 对应的算力 312 TFLOPS 和 990 TFLOPS 作为参数。

(5) **GPU 集群的算力利用率**: 参考 Google Research 发布的论文《PaLM: Scaling Language Modeling with Pathways》中的分析, 我们在测算中假设算力利用率约为 30%。

图表48：不同大模型训练过程中的算力利用率

	年份	参数量	加速芯片	算力利用率
GPT-3	2020	1750 亿	10000×NVIDIA V100	21.3%
Gopher	2022	2800 亿	4096×Google TPU v3	32.5%
MT-NLG	2022	5300 亿	2240×NVIDIA A100	30.2%
PaLM	2022	5400 亿	6144 Google TPU v4	46.2%

资料来源: 《PaLM: Scaling Language Modeling with Pathways》, 中信建投

其他基本假设包括多模态研发厂商个数、普通大模型研发厂商个数等。根据所有假设及可以得到, 2023 年-2027 年, 全球大模型训练端峰值算力需求量的年复合增长率为 78.0%。2023 年全球大模型训练端所需全部算力换算成的 A100 总量超过 200 万张。

图表49：全球大模型训练所需算力/AI 芯片数量测算

	2023E	2024E	2025E	2026E	2027E
多模态大模型研发厂商个数	5	8	10	13	15
同时训练模型数目	3	3	3	3	3
多模态大模型平均参数数量 (亿个, N)	15000	18000	21600	25920	31104
YoY		20.00%	20.00%	20.00%	20.00%
训练 Tokens 数量 (亿个)	10000	12000	14400	17280	20736
单个模型单 Token 训练所需运算次数 (TFLOPS, 6N)	9.00	10.80	12.96	15.55	18.66
单模型所需算力 (PFLOPS)	9.0×10^9	1.3×10^{10}	1.9×10^{10}	2.7×10^{10}	3.9×10^{10}
假设单次训练所需时间 (天)	7	7	7	7	7
训练端峰值算力需求 (PFLOPs, 单模型)	223214.29	514285.71	925714.29	1732937.14	2879341.71

请参阅最后一页的重要声明

所需算力×模型数量/（单次训练时间）					
普通大模型研发厂商个数	15	20	25	30	35
同时训练模型数目	3	3	3	3	3
普通大模型平均参数数量（亿个，N）	2000	2400	2880	3456	4147
YoY		20.00%	20.00%	20.00%	20.00%
训练 Tokens 数量（亿个）	4000	4800	5760	6912	8294
单个模型单 Token 训练所需运算次数（TFLOPS-s，6N）	1.20	1.44	1.73	2.07	2.49
单模型所需算力（PFLOPS）	480000000	691200000	995328000	1433272320	2063912141
假设单次训练所需时间（天）	7	7	7	7	7
训练端峰值算力需求（PFLOPS，单模型所需算力×模型数量/（单次训练时间））	35714.29	68571.43	123428.57	213284.57	358318.08
硬件算力效率	30%	30%	30%	30%	30%
H100 SXM FP16 Tensor （TFLOPS）	990	990	990	990	990
A100 SXM FP16 Tensor （TFLOPS）	312	312	312	312	312
H100 需求总量（万张） （只考虑 H100 的情况下）	87.18	196.25	353.25	655.29	1090.12
A100 需求总量（万张） （只考虑 A100 的情况下）	276.63	622.71	1120.88	2079.30	3459.04
H100 需求增量（万张） （只考虑 H100 的情况下）		109.07	157.00	302.05	434.83
A100 需求增量（万张） （只考虑 A100 的情况下）		346.08	498.17	958.42	1379.74

资料来源：OpenAI，Google Research，NVIDIA，中信建投

大模型云端推理对算力的需求测算：在云端推理场景下，我们分别从云端推理所需算力和云端模型部署所需显存两个维度分别进行测算。

算力角度的测算原理：基于前文对参数规模、模型数量等数据的假设，根据（1）大模型日活用户人数、（2）每人平均查询 Token 数量、（3）每 Token 推理成本与模型参数数量的关系估算推理端总算力需求，再考虑（4）单张 GPU 算力和 GPU 集群的算力利用率推导出 GPU 总需求。

（1）大模型日活用户人数：根据 Similarweb 统计数据，2023 年 1 月 ChatGPT 的日活用户数达到 1300 万。我们在测算中假设 2023 年多模态大模型的平均日活量达到 2000 万，普通大模型的平均日活量达到 1000 万，之后每年保持快速增长。

（2）每人平均查询 Token 数量：根据 OpenAI 数据，平均每 1000 个 Token 对应 750 个单词，我们在测算中假设每位用户平均查询的 Token 数量维持在 1000 个。

（3）每 Token 推理成本与模型参数数量的关系：参考 OpenAI 发布的论文《Scaling Laws for Neural Language

Models》中的分析，每个 token 的推理成本通常约为 $2N$ ，其中 N 是 LLM 的参数数量，我们在测算中遵循这一关系。

（4）单张 GPU 算力：由于测算中的大模型参数量级分别在千亿量级和万亿量级，考虑带宽容量和集群计算中的带宽限制，我们在测算中假设采用 H100 或 A100 作为云端推理卡。

图表50： 大模型云端推理所需算力/AI 芯片数量测算（算力角度）

	2023E	2024E	2025E	2026E	2027E
多模态大模型平均参数数量（亿个，N）	15000	18000	21600	25920	31104
YoY		20.00%	20.00%	20.00%	20.00%
多模态大模型日活用户人数（亿人）	0.2	0.5	1	2	4
YoY		150.00%	100.00%	100.00%	100.00%
每人平均每天查询次数（次）	20	20	20	20	20
每人平均每次查询 Tokens 数量（个）	1000	1000	1000	1000	1000
单 Tokens 所需计算次数（TFLOPs-s, $2N$ ）	3.00	3.60	4.32	5.18	6.22
每人每次查询所需计算次数（TFLOPs-s, $2N \times \text{Tokens 数量}$ ）	3000	3600	4320	5184	6220.8
全天计算次数合计（EFLOPs-s, 每人每次查询所需计算次数 $\times$ 查询次数 $\times$ 日活人数）	1200000	3600000	8640000	20736000	49766400
平均每 s 所需峰值算力（EFLOPs）	13.89	41.67	100.00	240.00	576.00
最大并发峰值算力乘数	5	5	5	5	5
最大并发峰值算力（EFLOPs）	69.44	208.33	500.00	1200.00	2880.00

普通大模型平均参数数量（亿个，N）	2000	2400	2880	3456	4147
YoY		20.00%	20.00%	20.00%	20.00%
普通大模型日活用户人数（亿人）	0.2	1	2	4	8
YoY		400.00%	100.00%	100.00%	100.00%
每人平均每天查询次数（次）	10	10	10	10	10
每人平均每次查询 Tokens 数量（个）	1000	1000	1000	1000	1000
单 Tokens 所需计算次数（TFLOPs-s, $2N$ ）	0.40	0.48	0.58	0.69	0.83
每人每次查询所需计算次数（TFLOPs-s, $2N \times \text{Tokens 数量}$ ）	400	480	576	691.2	829.44
全天计算次数合计（EFLOPs-s, 每人每次查询所需计算次数 $\times$ 查询次数 $\times$ 日活人数）	80000	480000	1152000	2764800	6635520
平均每 s 所需峰值算力（EFLOPs）	0.93	5.56	13.33	32.00	76.80
最大并发峰值算力乘数	5	5	5	5	5
最大并发峰值算力（EFLOPs）	4.63	27.78	66.67	160.00	384.00

峰值算力总量（PFLOPs）	246914	787037	1888889	4533333	10880000
----------------	--------	--------	---------	---------	----------

算力效率	30.00%	30.00%	30.00%	30.00%	30.00%
H100 SXM FP16 Tensor (TFLOPs)	990	990	990	990	990
A100 SXM FP16 Tensor (TFLOPs)	312	312	312	312	312
H100 需求量 (万张) (只考虑 H100 的情况下)	83	265	636	1526	3663
A100 需求量 (万张) (只考虑 A100 的情况下)	264	841	2018	4843	11624
H100 需求增量 (万张) (只考虑 A100 的情况下)	-	182	371	890	2137
A100 需求增量 (万张) (只考虑 A100 的情况下)	-	577	1177	2825	6781

资料来源: NVIDIA, OpenAI, 中信建投

根据所有假设及可以得到, 从云端推理所需算力角度测算, 2023 年-2027 年, 全球大模型云端推理的峰值算力需求量的年复合增长率为 113%。

显存角度测算原理: 首先, 目前 SK Hynix 已开发出业界首款 12 层 24GB HBM3, 考虑到一张 GPU 板卡面积有限, 限制了计算核心周围可布置的 HBM 数量, 因此未来一段时间内, GPU 显存容量的提升空间较小。其次, 推理最主要的需求是时效性, 为了满足时效性, 模型所需要的存储空间需要放到显存内。综合 GPU 板卡 HBM 容量有限和推理端模型需放置在 GPU 显存中这两个条件, 我们从模型推理端运行所需显存入手, 先预估推理端运行一个大模型所需显存容量 (1), 再假设业务场景中模型的峰值访问量, 并以此得到总体的显存需求 (2), 最终得到算力/AI 芯片的需求。

(1) 运行一个模型所需显存: 以 1750 亿参数的 GPT-3 模型为例, OpenAI 数据显示参数存储需要 350GB 空间。假设推理计算中间产生的参数按照一倍计算, 因此推理至少需要 700GB 显存空间, 即部署一个模型需要 9 张 80GB 显存版本的 A100。

(2) 业务场景部署模型量及所需显存: 假设该模型能够同时处理的并发任务数量为 100, 即 9 张 A100 80GB 处理 100 用户同时并发访问。业务场景部署以搜索引擎为例, 假设最高并发访问人数为 2000 万, 则需要 2000 万/100*9=180 万张 A100 80GB。

图表51: 大模型云端推理所需算力/AI 芯片数量测算 (显存角度)

GPU 型号	A100 80GB	
模型	GPT-3	
参数 (亿)	1750 亿	10000 亿
FP16 推理精度显存预估 (GB)	350 GB	2000GB
推理中间参数量倍数预估	x1	x1
推理显存需求 (GB)	700 GB	4000 GB
显卡需求 (张)	9	50
业务部署场景假设	搜索引擎	
最高并发访问量 (万)	2000	
模型能同时处理的并发量假设	100	

模型部署量（万）	20	
显卡需求（万）	180	1000

资料来源：OpenAI，NVIDIA，中信建投

根据上述测算可以得到，云端推理的算力需求潜力巨大。在 AI 大模型规模化落地应用的情况下，云端推理所需的算力/AI 芯片将明显超过云端训练。如果考虑边缘端 AI 推理的应用，推理端算力规模将进一步扩大。

2.2 英伟达龙头地位稳固，国内厂商正逐步追赶

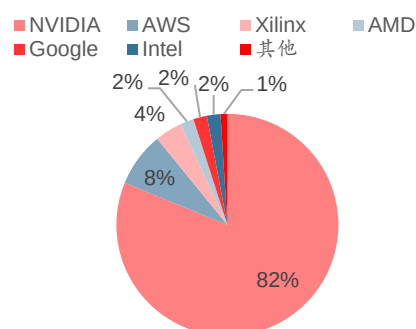
海外龙头厂商占据垄断地位，AI 加速芯片市场呈现“一超多强”态势。数据中心 CPU 市场上，英特尔份额有所下降但仍保持较大领先优势，AMD 持续抢占份额势头正盛。AI 加速芯片市场上，英伟达凭借硬件优势和软件生态一家独大，在训练、推理端均占据领先地位。根据 Liftr Insights 数据，2022 年数据中心 AI 加速市场中，英伟达份额达 82%，其余海外厂商如 AWS 和 Xilinx 分别占比 8%、4%，AMD、Intel、Google 均占比 2%。国内厂商起步较晚正逐步发力，部分加速芯片领域已经涌现出一批破局企业，但目前多为初创企业规模较小，技术能力和生态建设仍不完备，在高端 AI 芯片领域与海外厂商仍存在较大差距。未来，随着美国持续加大对中国高端芯片的出口限制，AI 芯片国产化进程有望加快。

图表52： AI 芯片市场竞争格局

CPU	AMD, Intel, Apple, Qualcomm, Samsung	兆芯, 龙芯中科, HYCON, 飞腾, C-SPARC
GPU	NVIDIA, AMD, Intel	景嘉微, 兆芯, MISTEX, 龙芯中科, HYCON, 壁仞科技
FPGA	XILINX, ADETA, LATTICE, Microsemi	紫光同创, 华恒电子, 安路信息, 复旦微电子, 高云半导体
ASIC	Google, Intel	寒武纪, 地平线, 黑芝麻, 瑞芯微, 全志科技, 晶晨股份

资料来源：各公司官网，，中信建投

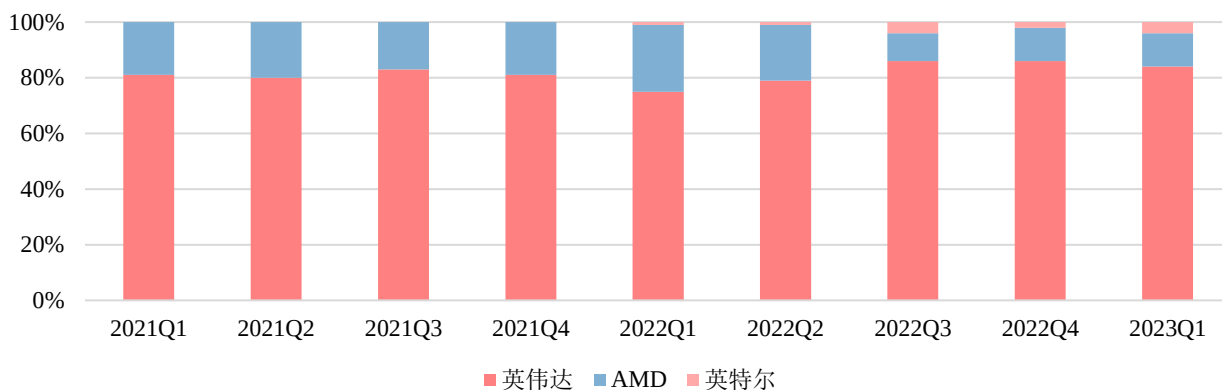
图表53： 2022 年 AI 加速芯片市场份额



资料来源：LIFTR INSIGHTS，中信建投

GPU 市场方面，海外龙头占据垄断地位，国产厂商加速追赶。当前英伟达、AMD、英特尔三巨头霸占全球 GPU 芯片市场的主导地位。集成 GPU 芯片一般在台式机和笔记本电脑中使用，性能和功耗较低，主要厂商包括英特尔和 AMD；独立显卡常用于服务器中，性能更高、功耗更大，主要厂商包括英伟达和 AMD。分应用场景来看，应用在人工智能、科学计算、视频编解码等场景的服务器 GPU 市场中，英伟达和 AMD 占据主要份额。根据 JPR，2023 年 Q1 英伟达的独立显卡（包括 AIB 合作伙伴显卡）的市场份额达 84%，AMD 和 Intel 则分别占比 12%、4%。

图表54： 全球独显 GPU 市场份额



资料来源：JPR，中信建投

图形渲染 GPU：英伟达引领行业数十年，持续技术迭代和生态构建实现长期领先。2006 年起，英伟达 GPU 架构保持约每两年更新一次的节奏，各代际产品性能提升显著，生态构建完整，Geforce 系列产品市占率长期保持市场首位，最新代际 GeForce RTX 40 系列代表了目前显卡的性能巅峰，采用全新的 Ada Lovelace 架构，台积电 5nm 级别工艺，拥有 760 亿晶体管和 18000 个 CUDA 核心，与 Ampere 相比架构核心数量增加约 70%，能耗比提升近两倍，可驱动 DLSS 3.0 技术。性能远超上代产品。AMD 独立 GPU 在 RDNA 架构迭代路径清晰，RDNA 3 架构采用 5nm 工艺和 chiplet 设计，比 RDNA 2 架构有 54% 每瓦性能提升，预计 2024 年前 RDNA 4 架构可正式发布，将采用更为先进的工艺制造。目前国内厂商在图形渲染 GPU 方面与国外龙头厂商差距不断缩小。芯动科技的“风华 2 号”GPU 像素填充率 48GPixel/s，FP32 单精度浮点性能 1.5TFLOPS，AI 运算(INT8)性能 12.5TOPS，实测功耗 4~15W，支持 OpenGL4.3、DX11、Vulkan 等 API，实现国产图形渲染 GPU 突破。景嘉微在工艺制程、核心频率、浮点性能等方面虽落后于英伟达同代产品，但差距正逐渐缩小。2023 年顺利发布 JM9 系列图形处理芯片，支持 OpenGL 4.0、HDMI 2.0 等接口，以及 H.265/4K 60-fps 视频解码，核心频率至少为 1.5GHz，配备 8GB 显存，浮点性能约 1.5TFlops，与英伟达 GeForce GTX1050 性能相近，有望对标 GeForce GTX1080。

图表55： 国内外主流图形渲染 GPU 产品性能对比

厂商	英伟达	英伟达	景嘉微	芯动科技	芯动科技	摩尔线程
型号	GeForce RTX 4090	GTX1080	JM9 系列	风华一号	风华二号	MTT S80
制程	4nm	16nm	14nm	12nm	NA	NA
核心数目	16384	2560	NA	NA	NA	4096 个 MUSA
时钟频率	2.23-2.52GHz	1.61-1.73GHz	1.5GHz	NA	NA	1.8GHz
显存容量	24GB	8GB	8GB	4GB/8GB/16GB	2/4/8GB	16GB
显存类型	GDDR6X	GDDR5X	NA	GDDR6/GDDR6X	NA	GDDR6
FP32 运算性能	82.58 TFLOPS	8.873 TFLOPS	1.5 TFLOPS	5TFLOPS/10 TFLOPS	1.5 TFLOPS	14.4 TFLOPS

请参阅最后一页的重要声明

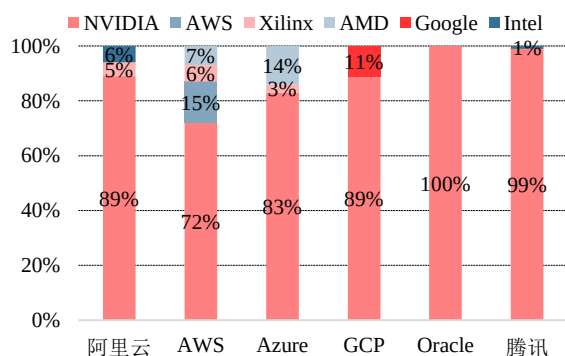
FP16 运算性能 82.58TFLOPS

Int8 运算性能	25TOPS		25TOPS		12.5TOPS	
总线接口	PCIe 4.0 x16	PCIe 3.0 X16	PCIe 4.0 X8	PCIe 4.0 x16	PCIe 3.0 x8	PCIe Gen5 x16

资料来源：各公司官网，中信建投

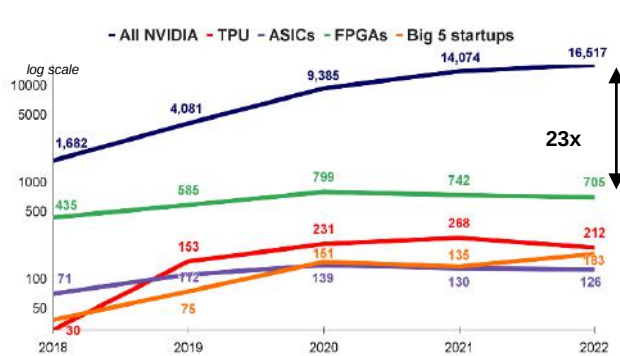
GPGPU：英伟达和 AMD 是目前全球 GPGPU 的领军企业。英伟达的通用计算芯片具备优秀的硬件设计，通过 CUDA 架构等全栈式软件布局，实现了 GPU 并行计算的通用化，深度挖掘芯片硬件的性能极限，在各类下游应用领域中，均推出了高性能的软硬件组合，逐步成为全球 AI 芯片领域的主导者。根据 stateof.AI 2022 报告，英伟达芯片在 AI 学术论文中的出现频次远超其他类型的 AI 芯片，是学术界最常用的人工智能加速芯片。在 Oracle 以及腾讯云中，也几乎全部采用英伟达的 GPU 作为计算加速芯片。AMD 2018 年发布用于数据中心的 Radeon Instinct GPU 加速芯片，Instinct 系列基于 CDNA 架构，如 MI250X 采用 CDNA2 架构，在通用计算领域实现计算能力和互联能力的显著提升，此外还推出了对标英伟达 CUDA 生态的 AMD ROCm 开源软件开发平台。英伟达的 H100 及 A100、AMD 的 MI100、MI200 系列等是当前最为主流的 GPGPU 产品型号。

图表56： 2022 年人工智能加速芯片在云上部署情况



资料来源：LIFTR INSIGHTS，中信建投

图表57： 英伟达芯片在 AI 学术论文中的出现频次



资料来源：stateof.AI，中信建投

国内 GPGPU 厂商正逐步缩小与英伟达、AMD 的差距。英伟达凭借其硬件产品性能的先进性和生态构建的完善性处于市场领导地位，国内厂商虽然在硬件产品性能和产业链生态架构方面与前者有所差距，但正在逐步完善产品布局和生态构建，不断缩小与行业龙头厂商的差距。国内主要 GPGPU 厂商及产品如下：

1) 海光信息：公司第一代 DCU 产品深算一号已于 2021 年实现商业化应用，采用 7nm 制程，基于大规模并行计算微结构进行设计，能支持 FP64 双精度浮点运算，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，且该产品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，高速并行数据处理能力强大，在典型应用场景下，主要性能指标可对标 AMD MI100、英伟达 P100，接近英伟达 A100；第二代 DCU 产品深算二号处于研发阶段，进展顺利。DCU 系列产品全面兼容“类 CUDA”环境，因此能够较好地适配、适应国际主流商业计算软件和人工智能软件，公司积极参与开源软件项目，加快了公司产品的推广速度，并实现与 GPGPU 主流开发平台的兼容。未来有望广泛应用于大数据处理、人工智能、商业计算等领域。

2) 天数智芯：2021 年 11 月宣布量产国内首款云端 7nm GPGPU 产品卡“天垓 100”，采用业界领先的台积电

电 7nm FinFET 制造工艺、2.5D CoWoS 封装技术，搭配台积电 65nm 工艺的自研 Interposer(中介层)，集成多达 240 亿个晶体管，整合 32GB HBM2 内存、存储带宽达 1.2TB，支持 FP32、FP/BF16、INT32/16/8 等多精度数据混合训练，系统接口 PCIe 4.0 x16。支持国内外主流 GPGPU 生态和多种主流深度学习框架。

3) 壁仞科技：2022 年 9 月针对人工智能训练、推理，及科学计算等更广泛的通用计算场景推出 BR100 系列通用 GPU 芯片，目前主要包括 BR100、BR104 两款芯片，基于壁仞科技原创芯片架构研发，采用 7nm 制程，可容纳 770 亿颗晶体管，并在国内率先采用 Chiplet 技术，新一代主机接口 PCIe 5.0，支持 CXL 互连协议，双向带宽最高达 128GB/s，具有高算力、高通用性、高能效三大优势。创下全球算力纪录，16 位浮点算力达到 1000T 以上、8 位定点算力达到 2000T 以上，单芯片峰值算力达到 PFLOPS 级别，达到国际厂商在售旗舰产品 3 倍以上，创下国内互连带宽纪录。

4) 摩尔线程：2022 年基于自研第二代 MUSA 架构处理器“春晓”GPU 推出针对数据中心的全功能 MTT S2000/S3000。MTT S3000 具有 PCIe Gen5 接口，FP32 算力为 15.2 TFLOPS，核心频率 1.9 GHz，显存容量 32 GB，支持 MUSA 安全引擎 1.0 以及 GPU 弹性切分技术，支持在云端的虚拟化和容器化。此外，摩尔线程推出了完备的 MUSA 软件栈，可帮助 MUSA 开发者快速基于摩尔线程全功能 GPU 开发各种不同的应用软件，并可通过 CUDA ON MUSA 兼容 CUDA 语言开发。

5) 沐曦：沐曦首款异构 GPGPU 产品 MXN100 采用 7nm 制程，已于 2022 年 8 月回片点亮，主要应用于推理侧；应用于 AI 训练及通用计算的产品 MXC500 已于 2022 年 12 月交付流片，公司计划 2024 年全面量产。2023 年发布首款 AI 推理 GPU 加速卡——曦思 N100 及解决方案在安防领域的应用。曦思 N100 是一款面向云端数据中心应用的 AI 推理 GPU 加速卡，内置异构的 GPGPU 通用处理器核心“MXN100”，同时集成了 HBM2E 高带宽内存，单卡的 INT8 整数算力达 160TOPS，FP16 浮点算力则达 80TFLOPS，具备高带宽、低延时特性。支持 128 路编码和 96 路解码的高清视频处理能力，兼容 HEVC(H.265)、H.264、AV1、AVS2 等多种视频格式，最高支持 8K 分辨率。

图表58：国内外主流 GPGPU 产品性能对比

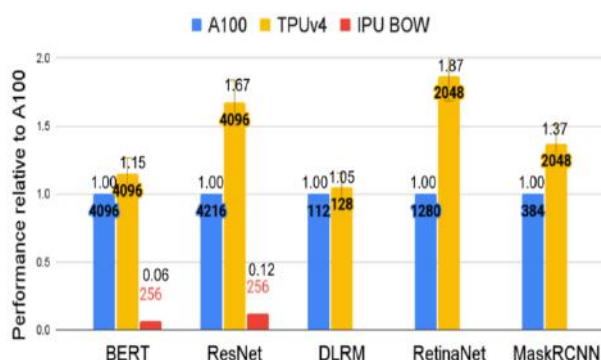
厂商	英伟达	英伟达	英伟达	海光信息	摩尔线程	壁仞科技	天数智芯	沐曦
型号	H100 SXM	A100 SXM	A800 (40G PCIE)	深算一号	MTT S3000	壁砺 100P	天垓 100	MXN100
制程	4nm	7nm	7nm	7nm FinFET		7nm	7nm	7nm
核心数目	15872	6912	6912	4096	4096			
时钟频率	1.07-1.83GHz	0.77-1.41GHz	0.475GHz	1.5-1.7GHz	1.9GHz			
显存容量	80GB	40GB/80GB	40GB	32GB	32GB	64GB	32GB	
显存类型	HBM3	HBM2E	HBM2	HBM2	GDDR6	HBM2E	DRAM HBM2	
FP32 运算性能	67TFLOPS	19.5TFLOPS	19.5 TFLOPS		15.2 TFLOPS	2456TFLOPS	37 TFLOPS	
FP16 运算性能	267.6TFLOPS	78TFLOPS				(BF16) 1024 TFLOPS		80TFLOPS
Int8 运算性能	1979 TOPS	624TOPS				2048 TOPS		160TOPS
互联接口	PCIe 5.0x16, NVLink Gen4: 900GB/s	PCIe 4.0 x16, NVLink Gen3: 600GB/s	PCIe 4.0 x16, NVLink Gen3: 400GB/s	PCIe Gen4 x 16	PCIe Gen5 x16	PCIe 5.0 X16	PCIe Gen4.0 x 16	
TDP	700W	400W	250W	350W	≤35W	450-550W	250W	

资料来源：各公司官网，中信建投

ASIC 市场方面，由于其一定的定制化属性，市场格局较为分散。在人工智能领域，ASIC 也占据一席之地。其中谷歌处于相对前沿的技术地位，自 2016 年以来，就推出了专为机器学习定制的 ASIC，即张量处理器 (Tensor Processing Unit, TPU)，近期，谷歌首次公布了其用于训练人工智能模型的 AI 芯片 TPU v4 的详细信息，其采用低精度计算，在几乎不影响深度学习处理效果的前提下大幅降低了功耗、加快运算速度，同时使用了脉动阵列等设计来优化矩阵乘法与卷积运算，对大规模矩阵的乘法可以最大化数据复用，减少访存次数，大幅提升 Transformer 模型的训练速度，同时节约训练成本。谷歌称在同等规模系统下基于 TPU 的谷歌超级计算机比基于英伟达 A100 芯片的系统最高快 1.7 倍，节能效率提高 1.9 倍。谷歌 TPU 属于定制化 ASIC 芯片，是专门为神经网络和 TensorFlow 学习框架等量身打造的集成芯片，需要在这类特定框架下才能发挥出最高运行效率。

图表59：谷歌 TPU v4 与英伟达 A100 性能指标对比

	NVIDIA A100	Google TPUv4
Production deployment	2020	2020
Peak TFLOPS	312 (bf16), 624 (int8)	275 (bf16 or int8)
Clock Rate Base/Boost (MHz)	1095 / 1410	1050
Tech. node, Die size	7 nm, 826 mm ²	7 nm
Transistor count	54 Billion	22 Billion
Chips per CPU host	4	4
TDP (W)	400	N/A
Inter Chip Interconnect	12 links @ 25 Gb/s	6 links @ 50 Gb/s
Largest scale MLPerf2.8 configuration (chips)	4216	4096
Processor Style	Single Instruction Multiple Threads	Single Instruction 2D Data
Processors / Chip	108	2
Threads / Core	32	1
On Chip Memory (MB)	40	128 (CMEM)+32 (VMEM)+10 (spMEM)
Register File Size (MB)	27	0.25
HBM2 capacity, BW	80 GB, 2039 GB/s	32 GB, 1200 GB/s

图表60：TPU v4 与英伟达 A100 在不同模型中的表现


资料来源：An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embedding，中信建投

资料来源：An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embedding，中信建投

国产厂商快速发展，寒武纪等异军突起。通过产品对比发现，目前寒武纪、海思昇腾、遂原科技等国产厂商正通过技术创新和设计优化，持续提升产品的性能、能效和易用性，推动产品竞争力不断提升，未来国产厂商有望在 ASIC 领域持续发力，突破国外厂商在 AI 芯片的垄断格局。国内主要 AI 用 ASIC 厂商及产品如下：

1) 寒武纪：云端人工智能领域，推出思元系列产品。其中，MLU100 芯片是中国首款高价值云端智能芯片。MLU290 芯片是寒武纪首款云端训练智能芯片，采用了 7nm 工艺，性能功耗上接近英伟达 A100，理论峰值性能分别高达 1024TOPS (INT4)、512TOPS (INT8)。思元 370 (MLU370) 芯片是寒武纪首款采用 Chiplet (芯粒) 技术的人工智能芯片，是寒武纪第二代云端推理产品思元 270 算力的 2 倍。MLU370-X8 与 MLU370-M8 是寒武纪基于思元 370 云端智能芯片打造的两款不同形态的人工智能加速卡。MLU370-X8 采用双芯思元 370 配置，为双槽位 250w 全尺寸智能加速卡，提供 24TFLOPS(FP32)训练算力和 256TOPS(INT8)推理算力；MLU370-M8 是寒武纪面向数据中心场景打造的 OAM 形态智能加速卡，可提供 32TFLOPS(FP32)训练算力和 340 TOPS(INT8)推理算力。两款加速卡均支持寒武纪 MLU-Link 芯片间互联，可满足多样化人工智能模型的训练和推理需求。此外，公司正在开展新一代思元 590 的开发，将采用 MLUarch05 全新架构，能够提供更大的内存容量和更高的内存带宽，其 I/O 和片间互联接口也较上代实现大幅升级。

2) 华为海思：推出昇腾系列产品。其中昇腾 310 在典型配置下可以输出 16TOPS@INT8, 8TOPS@FP16，功耗仅为 8W，采用自研华为达芬奇架构，集成丰富的计算单元，提高 AI 计算完备度和效率，进而扩展该芯片的适用性，全 AI 业务流程加速,大幅提高 AI 全系统的性能，有效降低部署成本。昇腾 910 是业界算力最强的 AI 处理器，基于自研华为达芬奇架构 3D Cube 技术，半精度 (FP16) 算力达到 320 TFLOPS，整数精度 (INT8) 算力达到 640 TOPS，功耗 310W，可支持云边端全栈全场景应用。表现性能上，昇腾 910 芯片性能接近英伟达

A100，但华为是基于自研的深度学习框架 MindSpore 与算力芯片进行相互优化，与 Tensorflow/Pytorch 两大主流深度学习训练框架的融合度不足，未来仍需要一定的时间进行生态建设。

3) 燧原科技：2019 年 12 月首发云端 AI 训练加速芯片邃思 1.0 及训练加速卡产品，2020 年推出推理加速卡，2021 年 7 月推出的第二代云端 AI 训练加速芯片邃思 2.0，单精度 FP32 峰值算力达到 40TFLOPS，单精度张量 TF32 峰值算力达到 160TFLOPS。同时搭载了 4 颗 HBM2E 片上存储芯片，高配支持 64GB 内存，带宽达 1.8TB/s。

4) 昆仑芯：昆仑芯 1 代 AI 芯片于 2020 年量产，在百度搜索引擎、小度等业务中部署数万片，是国内唯一一款经历过互联网大规模核心算法考验的云端 AI 芯片。昆仑芯 2 代 AI 芯片于 2021 年 8 月量产，是国内首款采用 GDDR6 显存的通用 AI 芯片，相比昆仑芯 1 代 AI 芯片性能提升 2-3 倍，且在通用性、易用性方面也有显著增强。昆仑芯 3 代有望在 2024 年规模上市，或将采用了 Huawei Da Vinci（达芬奇）架构，峰值性能为 256 TeraFLOPS，支持更多的运算和深度学习技术，例如 ONNX、TensorFlow 和 PyTorch。

图表61： 国内外主流 ASIC 产品性能对比

厂商	谷歌	寒武纪	寒武纪	海思	燧原	昆仑芯
型号	谷歌 TPUv4	寒武纪 MLU370-X8	寒武纪 MLU590	海思昇腾 910	燧原科技 T20	昆仑芯 2
发布时间	2020	2022	2022	2018	2021	2021
工艺制程	7nm	7nm		7nm	12nm	7nm
浮点算力	BF16 275TFLOPS	FP32 24TFLOPS		BF16 320TFLOPS	BF16 128TFLOPS	FP16 128TFLOPS
INT8 算力	275TOPS	256 TOPS		640TOPS	256TOPS	256 TOPS
互联带宽	1000GB/s	200GB/s			300GB/s	512GB / s
显存	32GB	48GB			32GB	
功耗		250W		350W	300W	120W
生态	TensorFlow XLA	Cambricon Neuware		MindSpore		Ascend 910

资料来源：各公司官网，中信建投

生态体系决定用户体验，是算力芯片厂商最深的护城河。虽然英伟达 GPU 本身硬件平台的算力卓越，但其强大的 CUDA 软件生态才是推升其 GPU 计算生态普及的关键力量。从技术角度来讲，GPU 硬件的性能门槛并不高，通过产品迭代可以接龙头领先水平，但下游客户更在意能不能用、好不好用的生态问题。CUDA 推出之前 GPU 编程需要机器码深入到显卡内核才能完成任务，而推出之后相当于把复杂的显卡编程包装成为一个简单的接口，造福开发人员，迄今为止已成为最发达、最广泛的生态系统，是目前最适合深度学习、AI 训练的 GPU 架构。英伟达在 2007 年推出后不断改善更新，衍生出各种工具包、软件环境，构筑了完整的生态，并与众多客户合作构建细分领域加速库与 AI 训练模型，已经积累 300 个加速库和 400 个 AI 模型。尤其在深度学习成为主流之后，英伟达通过有针对性地优化来以最佳的效率提升性能，例如支持混合精度训练和推理，在 GPU 中加入 Tensor Core 来提升卷积计算能力，以及最新的在 H100 GPU 中加入 Transformer Engine 来提升相关模型的性能。这些投入包括了软件和芯片架构上的协同设计，使得英伟达能使用最小的代价来保持性能的领先。而即便是英伟达最大的竞争对手 AMD 的 ROCm 平台在用户生态和性能优化上还存在差距。CUDA 作为完整的 GPU 解决方案，提供了硬件的直接访问接口，开发门槛大幅降低，而这套易用且能充分调动芯片架构潜力的软件生

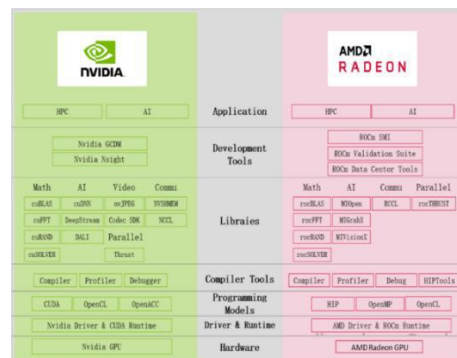
生态让英伟达在大模型社区拥有巨大的影响力。正因 CUDA 拥有成熟且性能良好的底层软件架构，故几乎所有的深度学习训练和推理框架都把对于英伟达 GPU 的支持和优化作为必备的目标，帮助英伟达处于持续处于领先地位。

图表62： CUDA 构建强大生态支持所有主流深度学习框架



资料来源：宽泛科技，中信建投

图表63： CUDA 生态和 ROCm 生态对照



资料来源：英伟达，AMD，中信建投

英伟达领先地位稳固。英伟达凭借良好的硬件性能和完善的 CUDA 生态将持续处于领先地位，但起步较晚的挑战者也在奋起直追，未来有望出现一超多强的多元化竞争格局。训练市场方面，英伟达高算力 GPU 是当前 AI 训练主流选择，谷歌 TPU 面临着通用性的局限，AMD 存在生态构建差距，但在二者的冲击及云厂商自研芯片的竞争下，AI 训练市场也或将出现格局的变动。推理市场方面，GPU 具有较好的生态延续性仍占主流，如英伟达针对推理市场的产品 Tesla T4 上的芯片包含了 2560 个 CUDA 内核，性能达到了 FP64 0.25 TFLOPS、FP32 8.1TFLOPS、INT8 达 130 TOPS，可提供多精度推理性能，以及优于 CPU 40 倍的低延时高吞吐量，可以实时满足更多的请求。但其他解决方案在成本、功耗具有优势，特定市场竞争格局相对激烈，工作负载不同对应的芯片性能需求不同，T4 PCIe，有望出现各类芯片共存的局面。

国内算力芯片厂商具备较好的入局机会。国产算力芯片市场需求巨大，国内人工智能生态环境较好，在 AI 应用领域的步伐处于全球前列，国产 GPU 厂商具孵化和发展的沃土，国内厂商供应链多元化的需求带来了国内 AI 芯片厂商适配窗口期，尤其是当前大模型发展早期是适配的黄金窗口期。其中，寒武纪、华为等兼容 CUDA 和自建生态是国产厂商发展的两大趋势，具备很大的竞争力潜力。短期来看，国内厂商兼容英伟达 CUDA，可以减轻开发和迁移难度，进而快速实现客户端导入。同时需要避开英伟达绝对优势领域，在芯片设计结构上形成差异化竞争；长期来看，国产 GPU 如果完全依赖 CUDA 生态，硬件更新将不得不绑定英伟达的开发进程，应借鉴 AMD、Google 构建自身生态体系，开展软硬件结合的平台化布局，并打造不同领域快速落地垂直解决方案的能力，铸造自己的生态圈核心壁垒。预计硬件性能高效以及能够构建符合下游需求的生态体系的国产厂商有望脱颖而出。

图表64： 昇腾计算产业生态示意图



资料来源：昇腾社区，中信建投

图表65： 寒武纪软件开发平台



资料来源：寒武纪开发者社区，中信建投

2.3 先进封装成为高性价比替代方案，存算一体应用潜力巨大

2.3.1 先进封装：后摩尔定律时代的创新方向，先进制程的高性价比替代方案

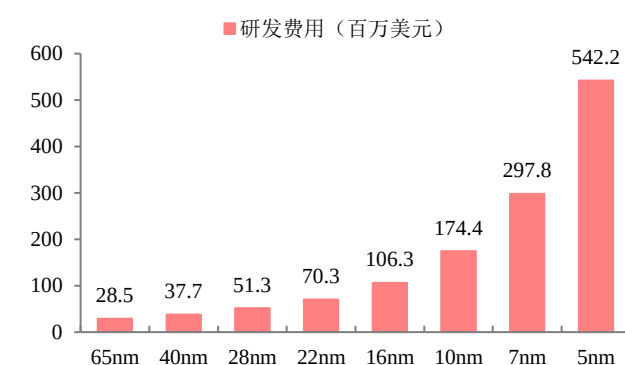
大算力芯片要求性能持续提升，后摩尔时代急需高性价比解决方案。随着大模型参数增加，AI 大模型对于算力需求大幅提升，GPU 等大算力芯片的性能提升遭遇两大瓶颈：一方面，进入 28nm 以后摩尔定律逐渐失效，先进制程的成本快速提升。根据 IBS 统计在达到 28nm 制程节点以后，如果继续缩小制程节点数，每百万门晶体管的制造成本不降反升，摩尔定律开始失效。而且应用先进制程的芯片研发费用大幅增长，5nm 制程的芯片研发费用增至 5.42 亿美元，几乎是 28nm 芯片研发费用的 10.6 倍，高额的研发门槛进一步减少了先进制程的应用范围。另一方面，内存带宽增长缓慢，限制处理器性能。在传统 PCB 封装中，走线密度和信号传输速率难以提升，因而内存带宽缓慢增长，导致来自存储带宽的开发速度远远低于处理器逻辑电路的速度，带来“内存墙”的问题。

图表66： 每百万门晶体管的成本在 28nm 后开始上升



资料来源：IBS，中信建投

图表67： 先进制程芯片的研发费用大幅上升

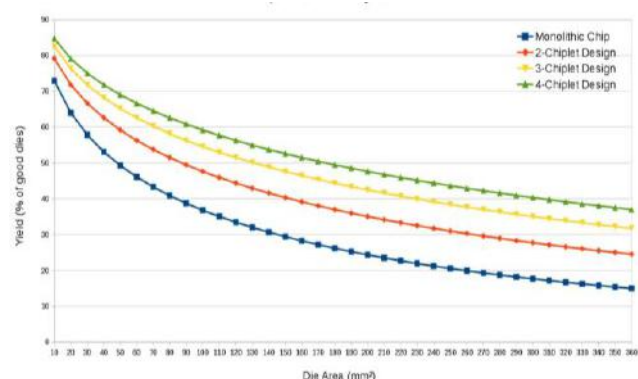


资料来源：IBS，中信建投

Chiplet 设计+异构先进封装提供了性能与成本平衡的最佳方案。Chiplet 即“小芯片”，是指预先制造好、具有特定功能、可组合集成的晶片（Die）。Chiplet 技术背景下，可以将大型单片芯片划分为多个相同或者不同的小芯片，这些小芯片可以使用相同或者不同的工艺节点制造，再通过跨芯片互联和先进封装技术进行封装级别集成，主要优势包括：1）可以突破光罩尺寸对单芯片面积的限制；2）可以充分发挥旧工艺节点的性价比优

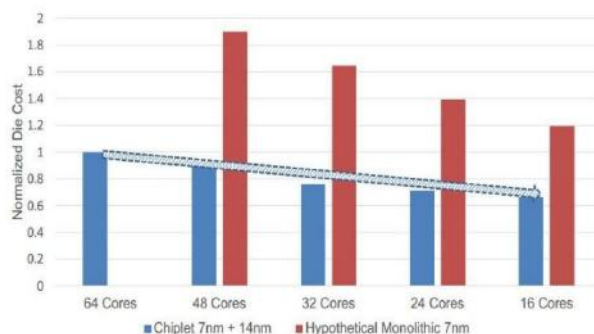
势，有效提升产品的良率，降低成本；3）通过集成不同工艺的芯粒，可以形成更加灵活的产品策略；4）先进封装的走线密度短，信号传输速率有很大的提升空间，同时能大大提高互连密度，成为解决内存墙问题的主要方法之一。

图表68： Chiplet 有利于提升良率



资料来源：Wikichip，中信建投

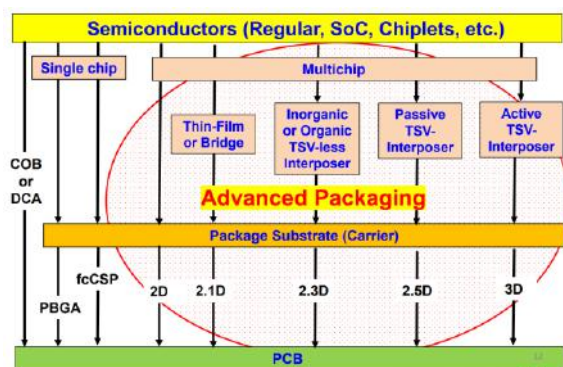
图表69： 用 Chiplet 技术的 7nm+14nm 的造价 vs 7nm



资料来源：AMD，中信建投

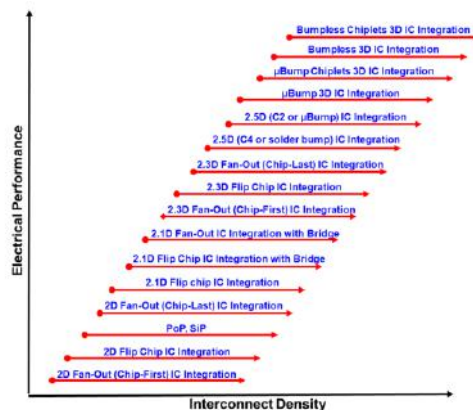
为了使异构集成的 Chiplet 封装实现，需要借助到 2D/2.1D/2.3D/2.5D/3D 等一系列先进封装工艺。先进封装的不同层次主要依据多颗芯片堆叠的物理结构和电气连接方式划分，例如 2D 封装中的芯片直接连接到基板，其他封装则以不同形式的中介层完成互连。其中，2.5D 封装常用于计算核心与 HBM 的封装互连，3D 封装常用于 HBM 显存的多层堆叠，并有望用于不同 IC 的异构集成。

图表70： 先进封装的层次



资料来源：《Recent Advances and Trends in Advanced Packaging》，中信建投

图表71： 先进封装依据互连密度和性能排名

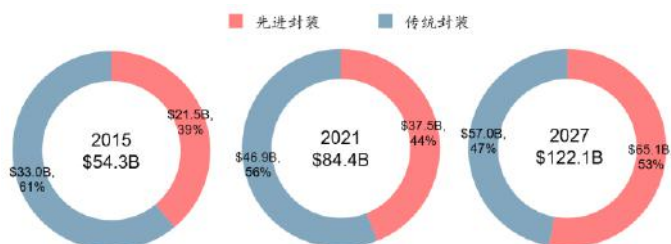


资料来源：《Recent Advances and Trends in Advanced Packaging》，中信建投

先进封装市场快速成长，相对高阶的封装形式将呈现更快增速。预计 2027 年先进封装市场规模增至 651 亿美元，2021-2027 年 CAGR 达到 9.6%。根据 Yole 数据，全球封装市场中，先进封装占比已由 2015 年的 39% 提升至 2021 年的 44%。预计到 2027 年，先进封装市场占比将增至 53%，规模约为 651 亿美元，2021-2027 年 CAGR 约为 9.6%，高于传统封装市场的 3.3% 和市场整体的 6.3%。倒装稳占先进封装最大份额，2.5D /3D、嵌入式芯片和扇出成为增长最快的先进封装平台。根据 Yole 数据，先进封装内份额最大的板块为倒装（包括 FCBGA、FCCSP、FC-SiP），2021 年市场规模约 262.7 亿美元，占比 70%。从增速角度来看，相对高阶的封装

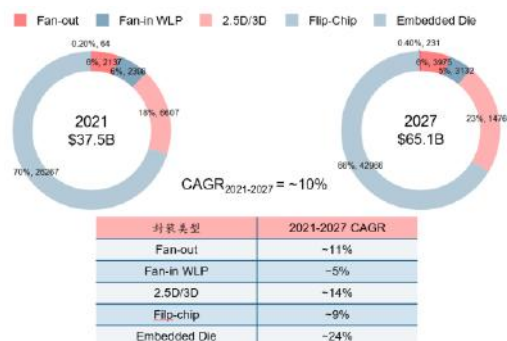
形式 Fan-Out、2.5D /3D、Embedded Die 在智能手机、HPC、自动驾驶等领域需求的推动下，保持高于先进封装整体市场的复合增速。

图表72：全球封装市场规模及结构预测



资料来源: Yole, 中信建投

图表73：先进封装市场规模及结构预测

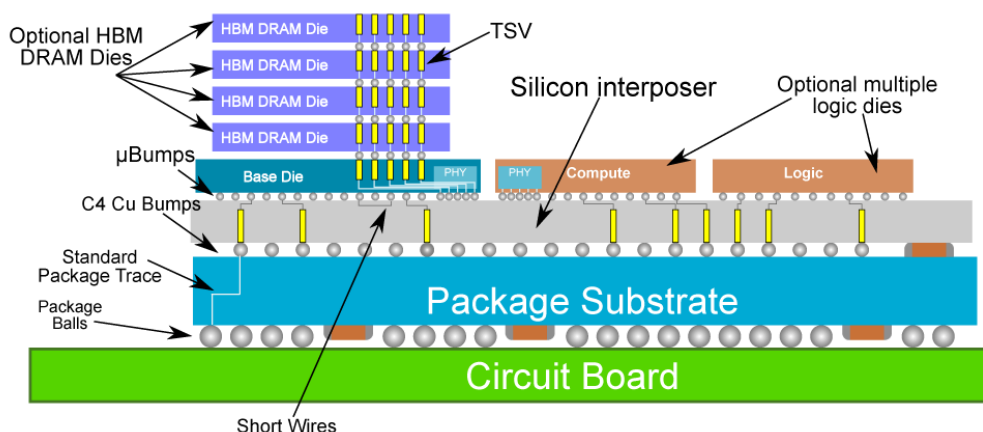


资料来源: Yole, 中信建投

1) CoWoS: 2.5D 封装重要解决方案，实现计算核心与 HBM 封装互连

计算核心与 HBM 通过 2.5D 封装互连，台积电开发的 CoWoS 封装技术为广泛使用的解决方案。台积电早在 2011 年推出 CoWoS 技术，并在 2012 年首先应用于 Xilinx 的 FPGA 上。此后，华为海思、英伟达、谷歌等厂商的芯片均采用了 CoWoS，例如 GP100（P100 显卡核心），TPU 2.0。如今 CoWoS 已成为 HPC 和 AI 计算领域广泛应用的 2.5D 封装技术，绝大多数使用 HBM 的高性能芯片，包括大部分创企的 AI 训练芯片都应用了 CoWoS 技术。

图表74：台积电 CoWoS 技术平台

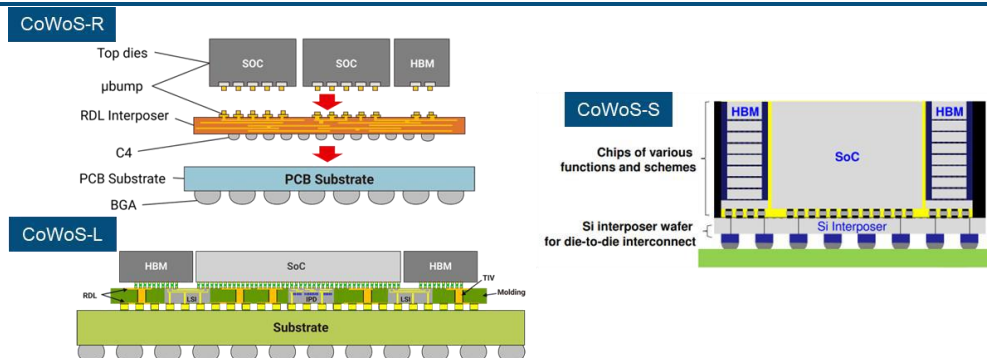


资料来源: 台积电, 中信建投

TSV（Through Silicon Via，硅通孔）是 CoMoS 封装的关键技术。TSV 在芯片和芯片之间、晶圆和晶圆之间制作垂直导通，通过铜、钨、多晶硅等导电物质的填充，实现硅通孔的垂直电气互连，是目前唯一的垂直电互联技术。根据中介层的不同，CoWoS 可以分为 CoWoS-S、CoWoS-R 和 CoWoS-L 三种：1) CoWoS-S 基于硅中介层为先进 SoC 和 HBM 提供系统集成；2) CoWoS-R 更强调小芯片间的互连，利用 RDL 实现最小 4 μm 的布线；3) CoWoS-L 则是最新的 CoWoS 技术，结合了 CoWoS-S 和 InFO 两种技术的优点，使用 RDL 与 LSI（本地硅互连）进行互连，具有最灵活的集成性。硅中介层中的 TSV 采用后通孔工艺（via last）形成，可以由封测厂商完成。CoWoS-S 对最大光罩掩膜版的尺寸有要求，当芯片封装规模大于掩膜版尺寸后将出现一张掩膜

版无法满足芯片完整曝光的需求，多次曝光拼接将带来良率问题，但是 HBM 接口对硅互联有着迫切的需求，因此 CoWoS-L 的 LSI 是目前的发展趋势。

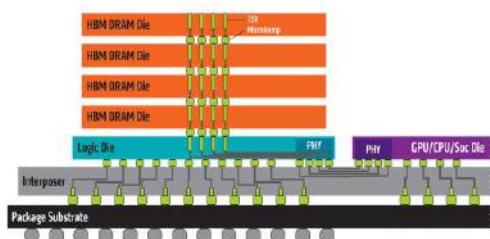
图表75：台积电三种 CoWoS 技术类型



资料来源：台积电，中信建投

CoWoS-S 基于硅中介层（Si-interposer）为先进 SoC 和 HBM 提供系统集成，在 GPU 等算力芯片的封装中应用广泛。CoWoS-S 的特点是混合了宽带内存模块 HBM（High Bandwidth Memory）和大规模 SoC 的高性能子系统，通过 Si 中介层连接 HBM 和 SoC，实现了宽带内存访问。CoWoS-S 最早于 2011 年开发，经历 5 代发展。最初，安装在中介层上的硅芯片是多个逻辑芯片，采用该技术的赛灵思高端 FPGA “7V2000T” 在 CoWoS-S 中配备了四个 FPGA 逻辑芯片。第 3 代开始支持逻辑和内存的混合加载。第 5 代 CoWoS-S 技术使用了全新的 TSV 解决方案，更厚的铜连接线，晶体管数量是第 3 代的 20 倍，硅中介层扩大到 2500mm²，相当于 3 倍光罩面积，拥有 8 个 HBM2E 堆栈的空间，容量高达 128 GB。第 6 代技术有望于 2023 年推出，将会在基板上封装 2 颗运算核心，同时可以板载多达 12 颗 HBM 缓存芯片。

图表76：GPU 与 HBM 封装结构示意图



资料来源：AMD，中信建投

图表77：台积电 CoWoS-S 发展历程



资料来源：台积电，中信建投

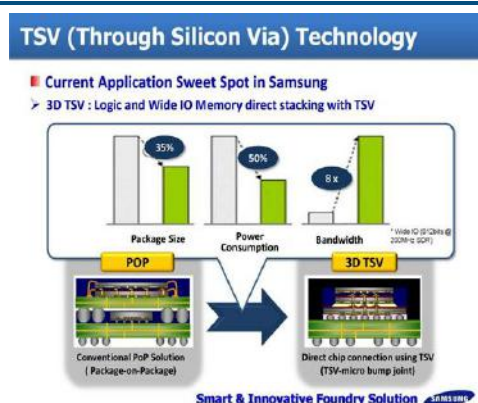
CoWoS 帮助台积电取得英伟达、AMD 等高性能计算芯片订单。根据 DIGITIMES 报道，微软已与台积电及其生态系统合作伙伴接洽，商讨将 CoWoS 封装用于其自己的 AI 芯片。英伟达高端 GPU 都采用 CoWoS 封装技术将 GPU 芯片和 HBM 集合在一起。Tesla P100 通过加入采用 HBM2 的 CoWoS 第三代技术，将计算性能和数据紧密集成在同一个程序包内，提供的内存性能是 NVIDIA Maxwell 架构的三倍以上。V100、A100、等高端 GPU，均采用台积电 CoWoS 封装，分别配备 32 GB HBM2、40GB HBM2E 内存，全新 Hopper 架构的 H100 GPU 也采用 CoWoS 封装，具有 80GB 的 HBM3 内存和超高的 3.2TB/s 内存带宽。**AMD 也将重新采用 CoWoS 封装。**根据 DIGITIMES 报道，AMD MI 200 原本由日月光集团与旗下矽品提供，应用 FO-EB 先进封装（扇出嵌入式桥接），而新 MI 系列数据中心加速器芯片将重新采用台积电先进封装 CoWoS。基于 Aldebaran GPU 的 MI250 或采用第五代 CoWoS 封装技术，可实现 128GB HBM2E 内存等超高性能配置。

请参阅最后一页的重要声明

2) HBM：3D 封装打造多层堆叠内存，突破容量与带宽瓶颈

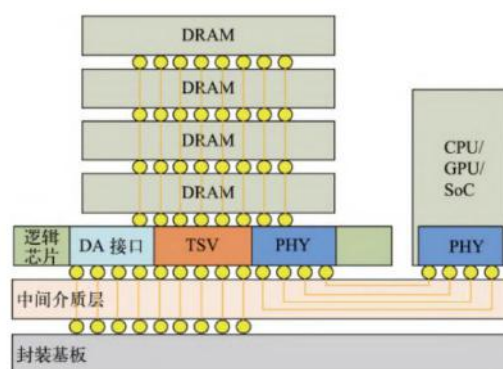
HBM 采用 3D 封装，通过 TSV 将多个 DRAM die 垂直堆叠。在后摩尔时代，存储带宽制约了计算系统的有效带宽，导致芯片算力性能提升受到限制，HBM 应运而生，与传统 DRAM 不同，HBM 是 3D 结构，它使用 TSV 技术将数个 DRAM 裸片堆叠起来，形成立方体结构，即 DRAM 芯片上搭上数千个细微孔并通过垂直贯通的电极连接上下芯片；DRAM 下面是 DRAM 逻辑控制单元，对 DRAM 进行控制。从技术角度看，HBM 促使 DRAM 从传统 2D 加速走向立体 3D，充分利用空间、缩小面积，契合半导体行业小型化、集成化的发展趋势。HBM 和硅互联技术突破了内存容量与带宽瓶颈，被视为新一代 DRAM 解决方案。而相较传统封装方式，TSV 技术能够缩减 30% 体积，并降低 50% 能耗。

图表78：3D TSV 封装技术



资料来源：三星，中信建投

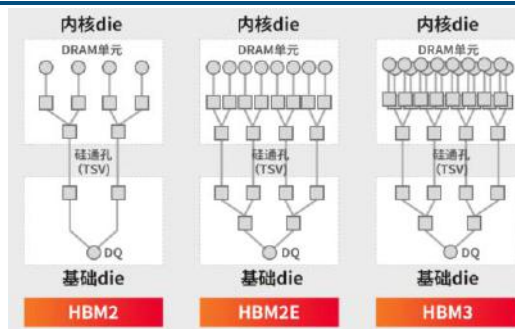
图表79：HBM 堆叠结构



资料来源：电子与封装，中信建投

HBM 相对传统内存数据传输线路的数量大幅提升。存储器带宽指单位时间内可以传输的数据量，要想增加带宽，最简单的方法是增加数据传输线路的数量。在典型的 DRAM 中，每个芯片有八个 DQ 引脚²，也就是数据输入/输出引脚。在组成 DIMM3 模块单元之后，共有 64 个 DQ 引脚。然而，随着系统对 DRAM 和处理速度等方面的要求有所提高，数据传输量也在增加。因此，DQ 引脚的数量（D 站的出入口数量）已无法保证数据能够顺利通过。HBM 由于采用了系统级封装（SIP）⁴和硅通孔（TSV）技术，拥有高达 1024 个 DQ 引脚，但其外形尺寸（指物理面积）却比标准 DRAM 小 10 倍以上。由于传统 DRAM 需要大量空间与 CPU 和 GPU 等处理器通信，而且它们需要通过引线键合⁵或 PCB 迹线⁶进行连接，因此 DRAM 不可能对海量数据进行并行处理。相比之下，HBM 产品可以在极短距离内进行通信，增加了 DQ 路径，显著加快了信号在堆叠 DRAM 之间的传输速度，实现了低功耗、高速的数据传输。

图表80：HBM 数据传输线路的数量大幅提升



资料来源：海力士，中信建投

目前 HBM 产品带宽增加了七倍，已接近 1TB/秒的里程碑节点。显存带宽=显存等效频率×显存位宽/8，因此频率和带宽决定显存性能。HBM 显存可以提供 1024bit 起跳的显存位宽，4 颗粒堆叠式的显存可达到 128GB/s 的带宽。HBM 能大幅提高数据处理速度，每瓦带宽比 GDDR5 高出 3 倍多，且 HBM2 比 GDDR5 节省了 94% 的面积，减少 20%+ 的功耗。2021 年，SK 海力士和 Rambus 先后发布最高数据传输速率 6.4Gbps 和 8.4Gbps 的 HBM3 产品，每个堆栈将提供超过 819GB/s 和 1075GB/s 的传输速率，支持 16-Hi 堆栈，堆栈容量达到 64GB。HBM3 带宽达 819GB/s，相对初代增加了 7 倍，是 LPDDR5 的近 100 倍，较 DDR5、GDDR6 高出 10 倍以上。与传统内存相比，HBM 的存储密度更大、功耗更低、带宽更高，多用于与数据中心 GPGPU 配合工作，可以取代传统的 GDDR，HBM 优势在于高位宽，但是频率相对偏低。

图表81：DDR 与 HBM 技术指标对比

Item	GDDR6	GDDR6X	HBM2E	HBM3
DRAM density	2GB(per chip)	2GB(per chip)	24GB(per stack)	24GB(per stack)
#Channels/DRAM package	2 channels	2 channels	8 channels	16 channels
#Bits in a channel	16 bits	16 bits	128 bits	64 bits
Speed	16 Gbps	21 Gbps	3.6 Gbps	6.4 Gbps
Overall bandwidth	64GB/s	84 GB/s	460GB/s	819GB/s
Power efficiency			Better than GDDR6/6X	
Cost	Lower cost than HBM2E/3			
Packaging process	PCB	PCB	2.5D/3D	2.5D/3D

资料来源：奎芯科技，中信建投

图表82：历代 HBM 性能持续提升



资料来源：海力士，中信建投

HBM 正在成为 AI 服务器 GPU 的标配。AI 服务器需要在短时间内处理大量数据，对带宽提出了更高的要求，HBM 成为了重要的解决方案。AI 服务器 GPU 市场以 NVIDIA H100、A100、A800 以及 AMD MI250、MI250X 系列为主，基本都配备了 HBM。HBM 方案目前已演进为较为主流的高性能计算领域扩展高带宽的方案。SK 海力士 HBM3 显存的样品已通过 NVIDIA 的性能评估工作，在 2022 年 6 月向 NVIDIA 正式供货，2023 GTC 大会发布的 ChatGPT 专用最新 H100 NVL GPU，也配置了 188GB HBM3e 内存；Rambus HBM3 或将在 2023 年流片，实际应用于数据中心、AI、HPC 等领域。IDC 数据显示，2019 年中国 AI 加速服务器单机 GPGPU 搭载量最多达到 20 颗，加权平均数约为 8 颗/台。单颗 GPU 配套的 HBM 显存存储容量达到 80GB，对应价值量约为 800 美元。

图表83：目前推出的搭载 HBM 和 GDDR 的 GPU 产品

Graphics Card Name	Memory Technology	Memory Speed	Memory Bus	Memory Bandwidth	Release
AMD Radeon R9 Fury X	HBM1	1.0 Gbps	4096-bit	512 GB/s	2015
NVIDIA GTX 1080	GDDR5X	10.0 Gbps	256-bit	320 GB/s	2016
NVIDIA Tesla P100	HBM2	1.4 Gbps	4096-bit	720 GB/s	2016
NVIDIA Titan XP	GDDR5X	11.4 Gbps	384-bit	547 GB/s	2017
AMD RX Vega 64	HBM2	1.9 Gbps	2048-bit	483 GB/s	2017
NVIDIA Titan V	HBM2	1.7 Gbps	3072-bit	652 GB/s	2017
NVIDIA Tesla V100	HBM2	1.7 Gbps	4096-bit	901 GB/s	2017
NVIDIA RTX 2080 Ti	GDDR6	14.0 Gbps	384-bit	672 GB/s	2018
AMD Instinct MI100	HBM2	2.4 Gbps	4096-bit	1229 GB/s	2020
NVIDIA A100 80GB	HBM2e	3.2 Gbps	5120-bit	2039 GB/s	2020
NVIDIA RTX 3090	GDDR6X	19.5 Gbps	384-bit	936 GB/s	2020
AMD Instinct MI200	HBM2e	3.2 Gbps	8192-bit	3200 GB/s	2021
NVIDIA RTX 3090 Ti	GDDR6X	21.0 Gbps	384-bit	1008 GB/s	2022
AMD Instinct MI300	HBM3	满血6.4 Gbps	8192-bit	>5000 GB/s	E2023
NVIDIA H100	HBM3	满血6.4 Gbps	8192-bit	>5000 GB/s	E2023

资料来源：奎芯科技，中信建投

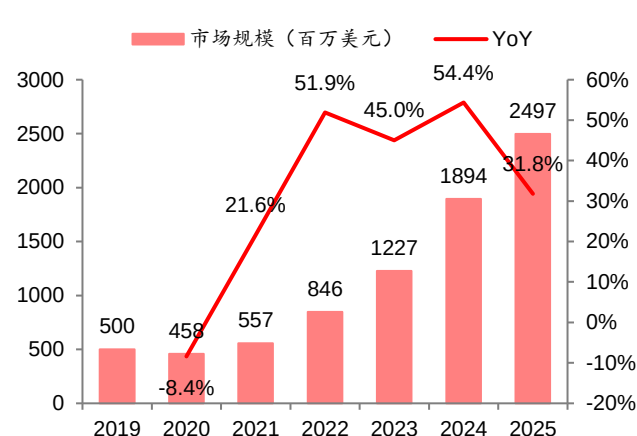
SK 海力士是 HBM 开发的先行者，并在技术开发和市场份额上占据领先地位。2014 年，SK 海力士与 AMD 联合开发了全球首款 HBM 产品。SK 海力士的 HBM3 发布 7 个月实现了量产，将搭载于 NVIDIA H100 之上。根据 BussinessKorea 的报道，SK 海力士在 HBM 市场已获得 60%-70% 的市场份额。SK 海力士之后，三星、美光推出了各自的 HBM 产品，分别迭代至 HBM3 和 HBM2E。晶圆代工厂商包括台积电、格芯等也在发力 HBM 相关的封装技术。

随着 HBM3 的性能提升，未来市场空间广阔。以位元计算，目前 HBM 占整个 DRAM 市场比重仅约 1.5%，渗透率提升空间较大。在将 GPU 等 AI 芯片推向高峰的同时，也极大带动了市场对新一代内存芯片 HBM（高带宽内存）的需求，据悉，2023 年开年以来，三星、SK 海力士的 HBM 订单就快速增加，价格也水涨船高。根据 TrendForce 咨询，2023-2025 年 HBM 市场 CAGR 有望成长至 40-45% 以上，至 2025 年市场规模有望快速增至 25 亿美元。

图表84：SK 海力士的 HBM 产品迭代

	HBM1	HBM2 Gen1	HBM2 Gen2	HBM2E	HBM3
Operating Frequency (Mbps)	~1600	1600	2000-4000	3200-3600	4400-6400
VDD	1.2V	1.2V	1.2V	1.2V	1.1V
Die Density (Stack)	2GB (4-Hi)	8GB (4Hi)	8GB (4Hi/8Hi)	16GB (4Hi/8Hi)	16-24GB (4/8/12Hi)
Release Year	2016	2017	2018	2020	2022

资料来源：SK 海力士，中信建投

图表85：全球 HBM 市场规模预测


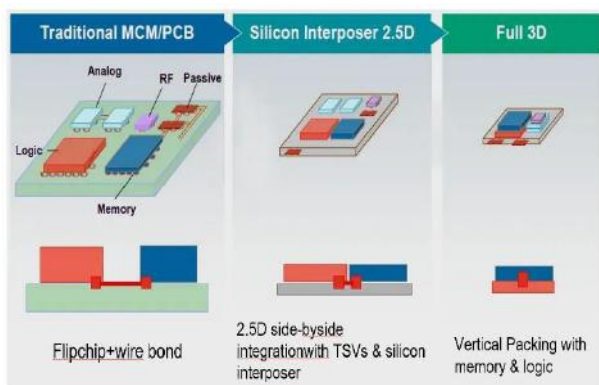
资料来源：TrendForce，Omdia，中信建投

3) 3D IC：多芯片垂直堆叠增强互联带宽，未来发展潜力巨大

3D IC 是指使用 FAB 工艺在单个芯片上堆叠多个器件层，包括多 Logic 芯片间的堆叠。与 2.5D 封装相比，

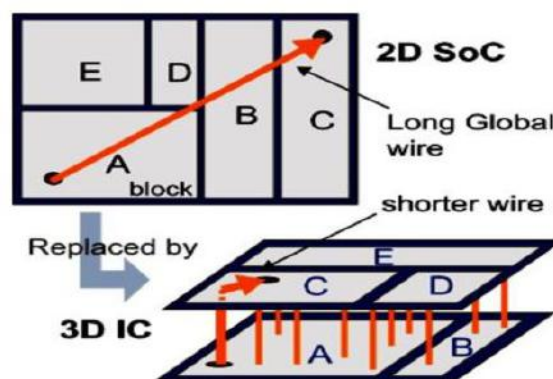
3D IC 封装在互连方式有所不同。2.5D 封装是通过 TSV 转换板连接芯片，而 3D IC 封装是将多个芯片垂直堆叠在一起，并通过直接键合技术实现芯片间的互连。在 2.5D 结构中，两个或多个有源半导体芯片并排放置在硅中介层上，以实现极高的芯片到芯片互连密度。在 3D 结构中，有源芯片通过芯片堆叠集成，以实现最短的互连和最小的封装尺寸。另一方面，2.5D 封装和 3D IC 封装的制造工艺也有所不同，2.5D 封装需要制造硅基中介层，并且需要进行微影技术等复杂的工艺步骤；而 3D IC 封装需要进行直接键合技术等高难度的制造工艺步骤。当前 3D IC 封装主流产品包括台积电 SoIC 技术、英特尔 Foveros 技术和三星 X-Cube 技术。

图表86：普通封装、2.5D 封装、3D IC 的区别



资料来源: einfochips, 中信建投

图表87：2D SoC 和 3D IC 互连线长度模型

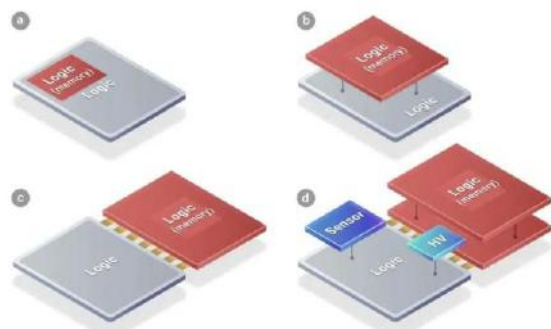


资料来源: 2cm, 中信建投

台积电 SoIC 是 3D 异构集成的技术平台，采用 wafer-on-wafer 键合技术。SoIC 技术采用 TSV 技术，可以实现非凸点键合结构，将许多不同性质的相邻芯片集成在一起。SoIC 技术将同构和异构小芯片集成到单个类似 SoC 的芯片中，该芯片具有更小的占用空间和更薄的外形，可以整体集成到 CoWoS 和 InFO 中。从外观上看，新集成的芯片就像一个普通的 SoC 芯片，但嵌入了所需的异构集成功能。

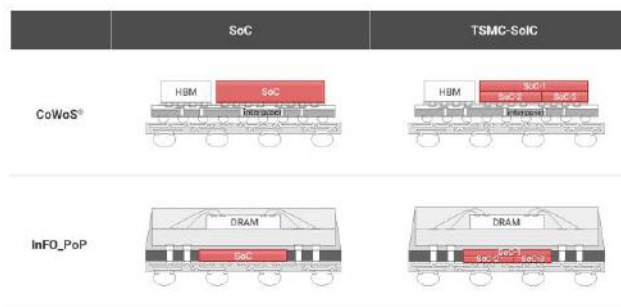
SoIC 主要分为 SoIC_CoW (Chip on Wafer) 和 SoIC_WoW (Wafer on Wafer)。1) **SoIC_CoW 技术**将不同尺寸、功能、节点的晶粒进行异质整合。2) **SoIC_WoW 技术**通过晶圆堆叠工艺实现异构和同质 3D 硅集成。紧密的键合间距和薄的 TSV 可实现最小的寄生以实现更好的性能、更低的功耗和延迟以及更小的外形尺寸。WoW 适用于高良率节点和相同裸片尺寸的应用或设计，甚至支持与第 3 方晶圆的集成。台积电在 CoW 方面正在开发 N7-on-N7 和 N5-on-N5 等；WoW 方面，台积电则在开发 Logic-on-DTC (Deep Trench Capacitor)。

图表88：台积电 SoIC 技术



资料来源: 台积电, 中信建投

图表89：SoIC 与 InFO_PoP、CoWoS 联合应用



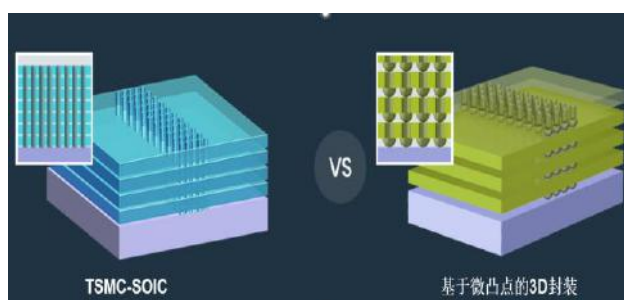
资料来源: 台积电, 中信建投

基于微凸块的 3D 封装借助微凸点连接芯片，在连接密度、性能等方面受限。传统 3D 封装在后端工艺中借

助微凸点 (Pump) 连接堆叠的芯片, 但微凸点的尺寸很难缩小到 $10\ \mu\text{m}$ 以下, 限制了堆叠芯片的 I/O 针脚计数。此外, 按比例排列的微凸点增加了寄生电容、电阻和电感, 降低了其性能和功率。

台积电 SoIC 3D 封装技术使芯片连接紧密, 并在互联带宽和散热上表现优异。台积电 SoIC 的键合技术在前端工艺完成, 接合间距更小, 使芯片更紧密地连接在一起, 提供超过 $10\text{K}/\text{mm}^2$ 的垂直互连密度, 用于超高带宽互连。在热性能方面, 台积电 SoIC 键合的热阻比微凸点下降低 35%。

图表90: SoIC 与基于微凸点的 3D 封装对比



图表91: SoIC 具有更优异的热性能表现



资料来源: 台积电, 中信建投

资料来源: 台积电, 中信建投

台积电公布了其 SoIC 研发进度, CoW 和 WoW 的研发进度基本一致, 为 N7/N6 工艺, 预计 2023 年将会实现基于 N5 工艺, 并预计将于 2035 年前实现 $1\ \mu\text{m}$ 以内的 SoIC 互连。3D IC 未来有望迎来快速发展和商用化进程。

图表92: 台积电 SoIC 研发进度规划



资料来源: 台积电, 中信建投

2.3.2 存算一体: 解决传统冯诺依曼架构“存储墙”, 能效比提升潜力巨大

存算一体有望解决传统冯诺依曼架构下的“存储墙”。由于处理器的设计以提升计算速度为主, 存储则更注重容量提升和成本优化, “存”“算”之间性能失配, 从而导致了访存带宽低、时延长、功耗高等问题, 即通常所说的“存储墙”和“功耗墙”。访存愈密集, “墙”的问题愈严重, 算力提升愈困难。随着以人工智能计算单元为代表的访存密集型应用快速崛起, 访存时延和功耗开销无法忽视, 计算架构的变革显得尤为迫切。存算

一体作为一种新型算力，指计算单元与存储单元融合，在完成数据存储功能的同时可以直接进行计算，有望解决传统冯诺依曼架构下的“存储墙”、“功耗墙”问题，以其巨大的能效比提升潜力，有望成为人工智能时代的先进应用技术。

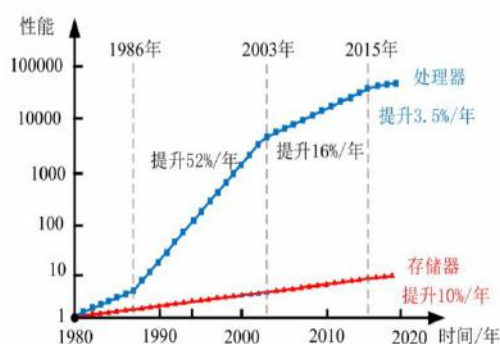
存储墙：数据搬运慢、搬运能耗大等问题是高速计算的关键瓶颈。从处理单元外的存储器提取数据，搬运时间往往是运算时间的成百上千倍，整个过程的无用能耗大概在 60%-90% 之间，能效非常低。

图表93：不同存储墙的带宽、功耗对比

	数据带宽	位宽	数据搬运能耗
片外 HBM	960GB/s	1024-bit	10nj
片外 DDR4	40GB/s	64bit	10nj
片内 SRAM	10-100TB/s	8bit、16bit、 32bit	50pj
计算功耗	-		5pj

资料来源：电子发烧友，中信建投

图表94：CPU 与存储器发展趋势

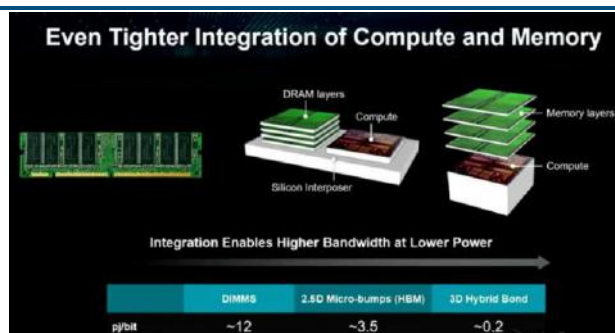


资料来源：半导体产业纵横，中信建投

根据存储与计算的距离远近，将广义存算一体的技术方案分为三大类，分别是近存计算（Processing Near Memory, PNM）、存内处理（Processing In Memory, PIM）和存内计算（Computing in Memory, CIM）。

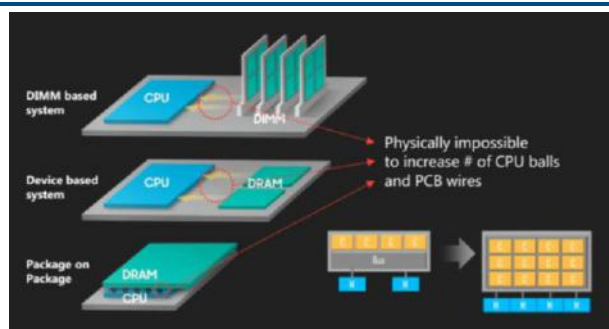
PNM：即 HBM 与 CPU 一体，用于高性能计算芯片，采用 HBM 堆叠，2.5D 封装，硅中介层（Interposer）内联在基板上。通过中介层紧凑而快速地连接后，HBM 具备的特性几乎和芯片集成的 RAM 一样。

图表95：近存计算大幅减少功耗



资料来源：Planet，中信建投

图表96：近存计算可以克服计算与存储之间的瓶颈

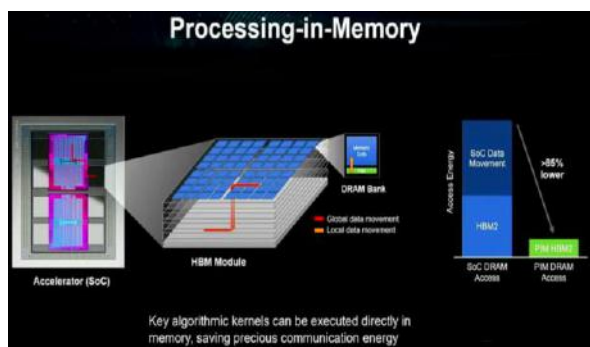


资料来源：Planet，中信建投

PIM：用硅通孔（Through Silicon Via, TSV，2010 年实现）技术将计算单元塞进内存上下 bank 之间。

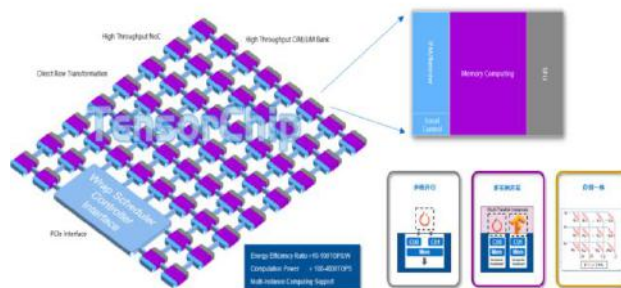
CIM：计算操作由位于存储芯片/区域内部的独立计算单元完成，存储和计算可以是模拟的也可以是数字的。这种路线一般用于算法固定的场景算法计算。目前主要路线是基于 NOR flash，多数情况下存储容量较小，这使得 NOR flash 单片算力达到 1TOPS 以上器件代价较大，通常业内大算力一般是 20-100TOPS 以上。而其他存储器，包括 SRAM、RRAM 等，可以用来做到大算力的存算一体。

图表97: PIM 原理(实例: Xilinx 的 Alveo U280)



资料来源: 三星, 中信建投

图表98: CIM 原理



资料来源: 电子发烧友, 中信建投

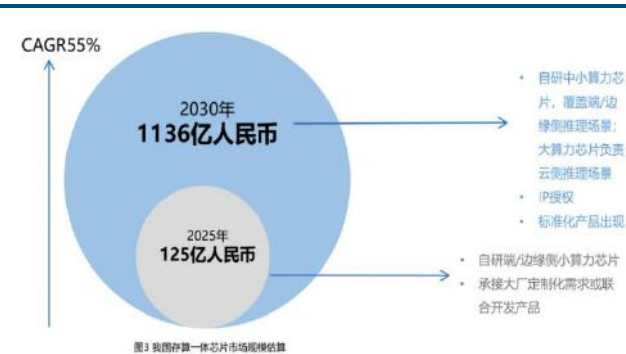
科研院所与龙头厂商积极布局, 未来市场潜力较大。2011 年, 存算一体芯片开始受到学界关注, 2016-2017 年成为学界热议话题, 随之而来学术大佬与业界领军厂商纷纷开启其商业化探索。科研院所方面, 加州大学圣芭芭拉分校谢源教授团队致力于在新型存储器件 ReRAM (阻变存储) 里面实现计算的功能研究, 即 PRIME 架构。清华大学刘勇攀教授团队和汪玉教授团队均参与了 PRIME 架构的研发, 目前已实现在 150nm 工艺下流片, 在阻变存储阵列里实现了计算存储一体化的神经网络, 功耗降低 20 倍, 速度提高 50 倍。此外, 清华大学与 SK 海力士联合成立智能存储计算芯片联合研究中心, 未来五年, 中心将致力于研发存算一体与近存储处理技术。在产业应用方面, 英特尔、博世、美光、Lam Research、应用材料、微软、亚马逊、软银都投资了 NOR 闪存存算一体芯片。其中, 英特尔发布的傲腾固态硬盘采用片外存储技术, 实现 CPU 与硬盘之间数据高速搬运, 从而平衡高级分析和人工智能等大规模内存工作负载的性价比。SK 海力士在今天的 ISSCC 发表存内计算的开发成果-基于 GDDR 接口的 DRAM 存内计算, 并展示了其首款基于存内计算技术产品-GDDR6-AiM 的样本。根据量子位智库预计, 2030 年基于存算一体的大算力芯片将实现规模量产, 应用场景覆盖大数据检索、蛋白质/基因分析、数据加密、图像处理等。2030 年, 基于存算一体技术的中小算力芯片市场规模约为 1069 亿人民币, 基于存算一体技术的大算力芯片市场规模约为 67 亿人民币, 总市场规模约为 1136 亿人民币。

图表99: 存算一体市场发展趋势



资料来源: 量子位, 中信建投

图表100: 我国存算一体市场规模估算



资料来源: 量子位, 中信建投

三、AI 服务器渗透率快速提升

3.1 AI 服务器是算力基础设施最主要的硬件, 训练型主要成本来自于 GPU 芯片

3.1.1 AI 服务器采用异构架构，主流结构为 CPU+多颗 GPU

与普通服务器的绝大多数空间分配给 CPU 相比，AI 服务器是采用异构形式的服务器，在异构方式上可以根据应用的范围采用不同的组合方式，一般采取 CPU+多颗 GPU 的架构，也有 CPU+TPU、CPU+其他的加速卡等组合。相较普通服务器，AI 服务器更擅长并行运算，具有高带宽、性能优越、能耗低等优点。

在大模型的预训练中，一方面侧重对文本上下文的理解，另一方面算法上存在大量的向量、矩阵计算，这让并行计算的 AI 服务器更擅长处理大模型的预训练任务。人工智能与通用大模型作为数字经济中的新兴行业，带动了大量的算力需求，也成为国内算力基础设施建设中最主要的硬件之一。

图表101：通用服务器与 AI 服务器的不同

	通用服务器	AI 服务器
硬件架构	2CPU	2CPU+8GPU、CPU+TPU 等
计算性能	擅长串行计算	擅长并行计算
代表供应商	Dell、HPE	浪潮
应用场景	传统金融、通信等行业	深度学习、大模型训练

资料来源：IDC，中信建投

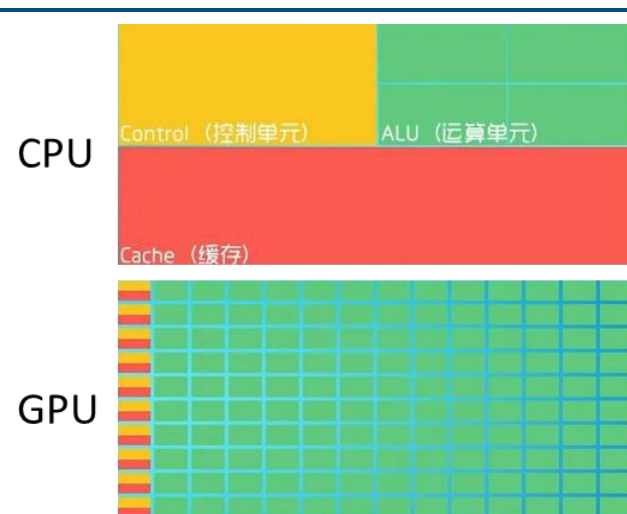
以 GPU 为核心的异构服务器未来将成为主流。对比 CPU 和 GPU 的内部架构，CPU 采用整块的 ALU(运算单元)，且大量空间用于控制单元和缓存，串行计算能力强；而 GPU 采用分立的大量 ALU，很少空间分配给控制单元和缓存，并行计算能力强。而由于图像识别、视觉效果处理、虚拟现实、大模型训练等任务都包含大量的简单重复计算、矩阵计算等，更适合用搭载 GPU 更多的异构型 AI 服务器进行处理，而随着企业的智能化变革和通用大模型的兴起，以 GPU 为核心的异构型 AI 服务器将在算力基础设施建设中占据愈发重要的地位。

图表102：GPU 与 CPU 产品特点

	GPU	CPU
核心数量	数千个加速核心(双卡 M40 高达 6144 个加速核心)	几十个核心
产品特点	1. 高效众多的运算单元 (ALU) 支持并行处理 2. 多线程已达到超大并行吞吐量	1. 复杂的逻辑控制单元 2. 强力的运算单元
适用场景	计算密集、易于并行的程序	逻辑复杂、串行计算的程序

数据来源：IDC，中信建投

图表103：GPU 与 CPU 内部结构



数据来源：IDC，中信建投

AI 服务器按应用场景又可分为训练和推理两种，其中训练对芯片算力需求更高，推理对算力需求相对较低。从部署方面来看训练服务器往往部署于云端，推理服务器则根据需求会部署于云端及边缘侧。

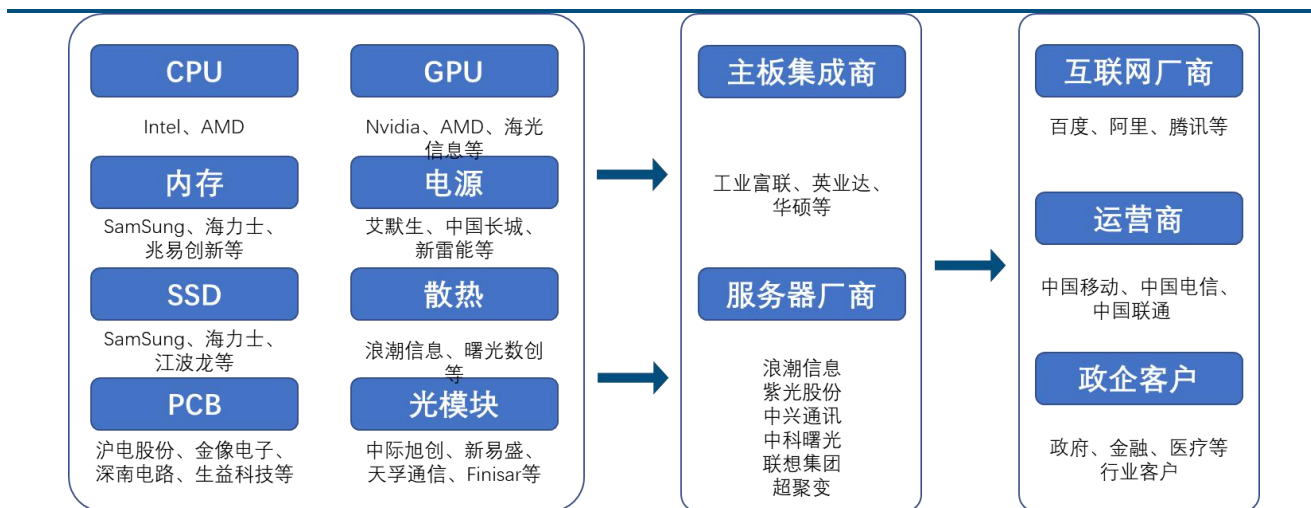
图表104： AI 服务器训练及推理区别

类别	训练	推理
定义	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能力的过程	对新的数据，使用经过训练的算法完成特定任务
算力要求	训练芯片应当具备强大的单芯片计算能力	推理芯片算力需求相对较低
部署位置	云端为主	云端和边缘侧
搭载 GPU 芯片	英伟达 A100、A800 等	英伟达 T4 等

资料来源：英伟达，华经产业研究院，中信建投

3.1.2 AI 服务器产业链上下游&成本结构拆解

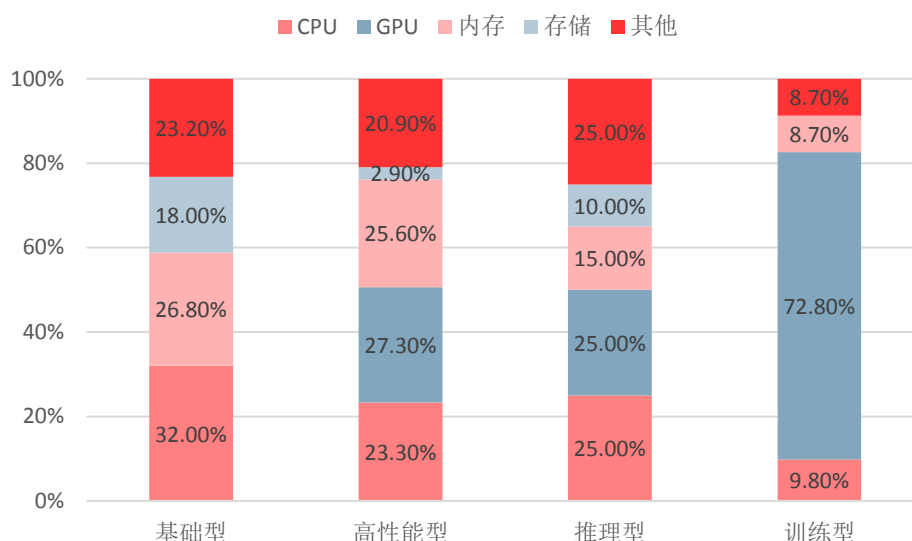
AI 服务器产业链上游主要由服务器元器件生产商组成，其中 CPU、GPU 作为核心组件，主要由 Intel、AMD、Nvidia 供应，国产供应商占比较少，其他部件包括内存、SSD、PCB、光模块、电源等存在更多的国产供应商；产业链中游包括主板集成商和服务器厂商，先由主板集成商将众多芯片集成，再交由服务器厂商装配成整机销售。目前国内企业在服务器厂商中占据重要地位；产业链下游主要包括以 BAT 为首的互联网厂商，移动、电信、联通三大运营商和众多政企客户（主要集中在政府、金融、医疗三大行业，因其最需要 AI 客服等相关产品）。

图表105： AI 服务器产业链概览


数据来源：， 中信建投

通用服务器成本主要由 CPU、存储、内存及其他部分构成，而 AI 服务器由于采用了多颗 GPU 芯片组成异构架构，其成本构成也会发生变化。具体来看，训练型 AI 服务器由于需要处理大量数据，具备更强的计算能力，训练芯片价格显著高于推理芯片。训练型 AI 服务器成本中，约 7 成以上由 GPU 构成，其余 CPU、存储、内存等占比相对较小。对于推理型服务器，其 GPU 成本约为 2-3 成，整体成本构成与高性能型相近。

图表106：各类型服务器成本结构拆分



数据来源：华经产业研究院，中信建投

以浪潮 AI 服务器旗舰型号 NF5468M6 为例，在其他配置均相同的情况下，训练型和服务型价格如下。以一块 Tesla A800-80G 显卡价格约为 10 万元计算，对于 8 块 A800 显卡的型号，GPU 成本占总成本的 72%；同样，按一块 Tesla T4-16G 显卡价格约为 1 万元计算，对于 8 块 T4 显卡的型号，GPU 成本占总成本的 28%。以上数据进一步印证了，训练型 AI 服务器的成本主要来自于 GPU 成本，而推理型 AI 服务器的 GPU 成本占比约为 25%左右，与高性能服务器基本相当。

图表107：浪潮 AI 服务器售价及 GPU 成本占比估算

服务器类型	配置	价格（人民币）	GPU 成本/总成本（估算）
训练	8 块 Tesla A800-80G 显卡	1106660	72%
	6 块 Tesla A800-80G 显卡	878760	68%
推理	8 块 Tesla T4-16G 显卡	290160	28%
	4 块 Tesla T4-16G 显卡	232060	17%

数据来源：浪潮信息，京东，中信建投

3.2 AI 服务器市场规模有望保持高速增长，当前订单饱满

3.2.1 全球 AI 服务器近三年将保持高速增长

根据 IDC 数据，2022 年全球 AI 服务器市场规模 202 亿美元，同比增长 29.8%，占服务器市场规模的比例为 16.4%，同比提升 1.2pct。我们认为随着数据量的持续提升，大模型参与玩家和单个模型参数量提升，以及数字化转型推进等多因素影响，AI 服务器市场规模将继续保持较快增长。

结合 2.1.3 节图表 45 我们对于大语言模型带来 AI 芯片的增量需求测算，我们认为 2023-2025 年全球 AI 服

服务器有望实现高速增长。以目前企业对于 AI 服务器的实际需求来看，虽然推理端需求更为旺盛，但从采购角度更倾向于搭载 A100/A800GPU 的训练/推理一体服务器。因此我们结合 3.1.2 节对于训练型、推理型 AI 服务器的成本拆解测算，预估 2023-2025 年增量的 GPU 需求约占 AI 服务器成本比重为 70%。此外，随着包括 H100/H800 等新一代芯片的推出、算法迭代升级均有望带来整体效率提升，AI 服务器增量市场空间可能略低于大模型需求预期。结合上述假设，我们认为全球 AI 服务器市场规模未来 3 年内将保持高速增长，市场规模分别为 395/890/1601 亿美元，对应增速 96%/125%/80%。由于互联网厂商等主要下游客户倾向于为未来潜在需求提前备货，因此 2023 年市场增速可能高于预测值，同时 2024、2025 年市场增速可能略低于预测值。

图表108： 全球 AI 服务器市场规模测算

	2021	2022	2023E	2024E	2025E
大模型带动 GPU 存量空间（亿美元）	-	-	276.6	622.7	1120.9
GPU 占 AI 服务器成本比例（%）	-	-	70.0	70.0	70.0
GPU 芯片升级/算法效率提升比例测算（%）	-	-	100.0	120.0	150.0
AI 服务器存量规模（亿美元）	156.0	202.0	395.2	889.6	1601.3
AI 服务器增量规模（亿美元）	-	46.0	193.2	494.4	711.7
市场增速（%）	39.1	29.8	95.6	125.1	80.0

资料来源：OpenAI, IDC, Nvidia, 中信建投

3.2.2 中国 AI 服务器近三年将保持高速增长

根据 IDC 数据，2022 年中国 AI 服务器市场规模 67 亿美元，同比增长 24%。其中 GPU 服务器占据主导地位，市场份额为 89%至 60 亿美元。同时，NPU、ASIC 和 FPGA 等非 GPU 加速服务器以同比 12%的增速占有了 11%的市场份额，达到 7 亿美元。在大模型浪潮到来前，由数字经济和“东数西算”等政策影响下，中国 AI 算力在 2021 年实现了 68.2%的同比高速增长。据浪潮信息、国际数据公司(IDC)和清华大学联合推出的《2021-2022 全球算力指数评估报告》显示，中国 AI 算力发展领跑全球，AI 服务器支出规模位列全球第一。我们认为，在大模型浪潮下，叠加数字经济、东数西算带动的数据中心、智算中心建设，AI 服务器市场中我国的份额在当前约全球 1/3 比例上有望进一步提升。我们预计，2023-2025 年，结合对于全球 AI 服务器市场规模的预判，以及对于我国份额占比持续提升的假设，我国 AI 服务器市场规模有望达到 134/307/561 亿美元，同比增长 101%/128%/83%。由于互联网厂商等主要下游客户倾向于为未来潜在需求提前备货，因此 2023 年市场增速可能高于预测值，同时 2024、2025 年市场增速可能略低于预测值。

图表109： 中国 AI 服务器市场规模测算

	2021	2022	2023E	2024E	2025E
全球市场规模（亿美元）	156.0	202.0	395.2	889.6	1601.3
中国市场占全球市场比重（%）	34.6	33.2	34.0	34.5	35.0
市场增速（%）	68.2	24.0	100.5	128.4	82.6
市场规模（亿美元）	54.0	67.0	134.4	306.9	560.5

资料来源：OpenAI, IDC, Nvidia, 中信建投

3.2.3 当前 AI 服务器厂商在手订单充分，AI 服务器市场高增长确定性较强

自去年 ChatGPT 带动的大模型浪潮以来，国内外头部互联网厂商纷纷加入 AI 算力的军备竞赛，加大对于 AI 算力侧的资源投入。AI 算力的高景气带动 AI 服务器需求端爆发式增长，并体现在 AI 服务器厂商订单端。

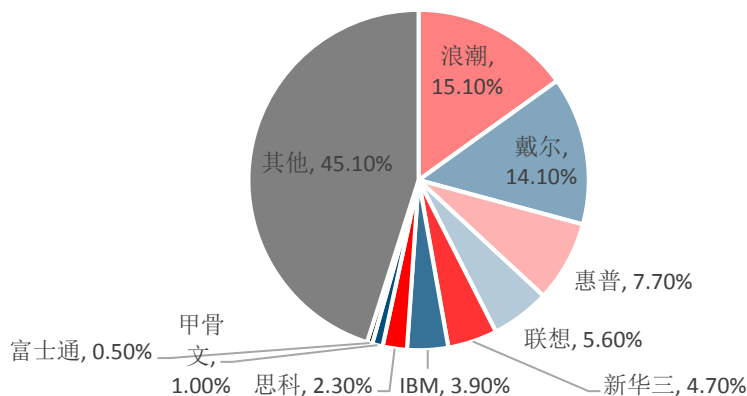
全球 AI 服务器出货金额排名第一位的龙头厂商浪潮信息，提到一季度以来 AI 服务器市场迎来明显增长，客户关注点由价格转向能否及时满足自身需求。此外，据紫光股份于投资者互动平台的回复，其 AI 服务器订单今年一季度有很大提升，产能满足市场需求不存在问题，针对 GPT 场景优化的 GPU 服务器已经完成开发，预计今年二季度全面上市。作为全球 ICT 设备龙头企业的联想集团，根据其最新公布的财报数据，ISG（基础设施解决方案业务集团）在 2023 年 1-3 月实现营收同比增长 56.2%，全财年营收同比增长 36.6%，主要受益于海外 AI 服务器需求爆发以及存储业务的高速增长，公司预期新财年 AI 服务器收入增速将显著快于通用服务器，带动 ISG 部门营收增长超市场平均水平 20% 以上。中科曙光深度布局算力领域，包括上游芯片、中游服务器解决方案、液冷技术、以及下游算力调度等业务，公司于投资者互动平台多次回复，会根据用户需求提供通用算力和智能算力产品及服务，随着我国算力需求的增长，各类产品销售均呈现增长态势，伴随我国人工智能技术和产业的发展，预计智能计算产品需求将逐步提升。

3.3 AI 服务器市场集中度有望提升，国内厂商呈现一超多强格局

3.3.1 全球 AI 服务器竞争格局

据 IDC 数据，2022 年上半年全球 AI 服务器市场中，浪潮信息、戴尔、惠普、联想、新华三分别以 15.1%、14.1%、7.7%、5.6%、4.7% 的市场份额位居前五位。市场格局相对分散，龙头厂商份额较为接近。此外，由于以北美云厂商为主的需求方偏向于采用 ODM 模式，因此非品牌商份额占比较高，接近 50%。

图表110： 2022 年上半年全球 AI 服务器市场份额

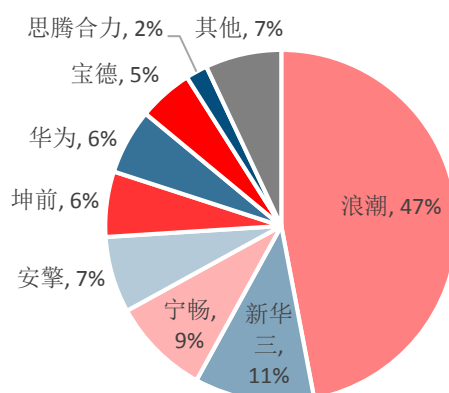


数据来源：IDC，中信建投

3.3.2 中国 AI 服务器竞争格局

据 IDC 数据，2022 年我国 AI 服务器市场按销售额统计市场份额中，浪潮信息、新华三、宁畅位居前三位，市场份额分别为 47%、11%、9%。市场格局呈现一超多强局面，除浪潮外其与厂商份额相对接近。由于国内头部厂商采用类 ODM 模式服务互联网客户，因此 ODM 厂商份额占比偏低。

图表111： 2022 年中国 AI 服务器市场份额



数据来源：IDC，中信建投

3.3.3 AI 服务器竞争格局未来演进趋势

从 AI 服务器的研发与交付考虑，品牌商和代工厂的模式及时间线略有不同，品牌商研发周期更长但交付更快，代工厂研发周期略短但交付产品时间略长。5 月 29 日，英伟达 CEO 在台北国际电脑展 COMPUTEX 2023 大会带来主题演讲，演讲中发布了目前台系 ODM 厂商针对客户需求做出的 AI 服务器雏形，并将进一步根据客户需求做定制化开发，由定制化开发到产品交付客户预计需要数月时间。对于 OEM 厂商来说，包括浪潮、联想、新华三等厂商的研发周期相对较长，需要接近一年的时间进行验证，并根据不同客户做不同配置规格进行进一步验证。OEM 厂商验证完成后的成熟产品在交付中相比 ODM 厂商可以实现更快交付。

从全球维度来看，考虑品牌商及代工厂两种商业模式，当前 AI 服务器市场份额中代工厂（ODM）厂商份额会高于品牌商（OEM），且 ODM 份额有逐步提升趋势。目前全球对于 AI 服务器的爆发式需求更多来自于大型互联网公司，在除中国外的全球市场，尤其以北美为主的市场中，大型互联网公司往往通过 ODM 满足其需求。从合作厂商来看，海外服务器出货排名靠前的戴尔、惠普、联想等厂商，其中戴尔和惠普均不参与 ODM 业务，联想在海外则是承接一部分 ODM 业务。除联想外，ODM 市场主要被英业达、纬创、富士康、广达等台系 ODM 传统厂商占据，上述厂商也在积极扩充墨西哥、东南亚现有产能，以更好满足北美云厂商的 AI 服务器强劲需求。预计 2023 年 ODM 海外产能比重将升至 50%，海外 AI 服务器市场 ODM 份额比重有望进一步提升。从品牌商的竞争格局来看，AI 服务器相比通用服务器复杂程度更高，研发周期更长，前期投入规模更大，需要企业具备一定的规模、资金储备、技术能力等，预计 AI 服务器市场中 OEM 份额将更进一步向龙头企业集中。

从国内维度看，OEM 份额占据主导，市场呈现一超多强的竞争格局。作为 OEM 厂商的浪潮、联想、新华三、宁畅等会以类 ODM 模式服务以互联网厂商为主的主要需求方，例如浪潮采用 JDM（Joint Design Manufacture 联合开发模式）模式，联想采用 ODM+ 模式。相比于海外 AI 服务器市场 ODM 厂商占据主流，国内市场的绝大多数份额由 OEM 厂商所占据，2022 年国内市场份额中，ODM 厂商份额不足两成，而浪潮、新华三、宁畅前三大 OEM 厂商占据 67% 市场份额。考虑到 AI 服务器研发和投入上需要更充足的资金及技术支持，OEM 的竞争格局预计将继续向头部集中。展望未来，国内 AI 服务器需求除互联网厂商外，还包括政府端的智算中心等建设，预计 OEM 的份额占比有望进一步提升。

3.4 全球服务器市场规模预计保持平稳

3.4.1 通用服务器仍处库存去化阶段，全球市场规模预计将出现下滑

根据研究机构 TrendForce 5 月 17 日发布的报告，2023 年服务器市场需求展望不佳，再次下调今年全球服务器整机出货量预测至 1383.5 万台，同比减少 2.85%。TrendForce 称，美国谷歌、微软、Meta、亚马逊四大互联网公司陆续下调服务器采购量；同时戴尔、HPE 等 OEM 厂商也在 2~4 月间下调全年出货量预估，同比分别减少 15%、12%；此外，受国际形势以及经济因素等多种因素导致全年服务器需求展望不佳。2023 年 Q1 受淡季效应以及终端库存修正的影响，全球服务器出货量环比减少了 15.9%。TrendForce 对于二季度产业回暖信心偏低，产业旺季并未如期发生，环比增长预估仅为 9.23%。此外，ESG 方面的讨论使得美国四大互联网公司延长服务器的使用年限，进而降低采购量，控制资本支出，这也是影响服务器市场的因素之一。预计库存去化完成将在今年下半年或明年上半年到来，若库存去化进度不及预期，全年服务器市场规模预测可能会进一步下调。

根据研究机构 IDC 报告，2022 年我国服务器市场规模为 273.4 亿美元，预计 2023 年将达 308 亿美元。2023 年一季度，国内服务器市场走势与全球水平较为相似，出现较大幅度下滑。展望未来，从通用服务器来看，下半年互联网企业库存消耗结束，有望释放订单；此外，包括政府、运营商等行业的采购下半年预计逐步展开，有望带来进一步需求。下半年需求回暖有望弥补上半年同比下滑的市场规模，国内通用服务器市场全年实现持平或温和增长。

3.4.2 AI 服务器出货量占比进一步提升，对全球服务器市场整体出货量贡献有限

去年底以来，ChatGPT 等人工智能应用的火热带动了 AI 服务器需求暴增，英伟达芯片出现供不应求情况。包括微软、谷歌、Meta、腾讯、百度等国内外云服务提供商纷纷积极加大 AI 算力投入。根据 TrendForce 预估，2023 年 AI 服务器出货量将同比实现 10% 增长，但由于从台数来看 AI 服务器占比不足 10%，对于整个市场影响相对有限，预计全球全年服务器出货量整体呈现持平或小幅下滑趋势。

从国内市场来看，互联网厂商及智算中心建设推动 AI 服务器需求暴涨，一季度相关厂商新增订单同比超 4 成，全年预计出货金额将保持高速增长。考虑到通用服务器市场下半年需求有望回暖，全年市场规模有望持平或小幅增长，叠加 AI 服务器的快速增长，根据 IDC 预测，预计全年服务器市场规模有望实现超 10% 的增长。

3.5 标的推荐

3.5.1 全球服务器行业龙头厂商——浪潮信息

浪潮信息是全球领先的 IT 基础设施产品、方案和服务提供商，为客户提供更先进的云计算、大数据、人工智能、边缘计算等各类创新产品和解决方案，并积极参与开放计算技术创新，加快全球计算生态的开放融合进程。通用服务器、边缘计算服务器和 AI 服务器为公司核心产品，具有很强的国际竞争力。

根据 IDC 数据，从 2019 年至 2022 年，浪潮信息始终保持服务器出货量、营业收入在国内服务器厂商中均排名第一；AI 服务器方面，浪潮信息多次位列全球市占率第一，且在中国 AI 加速计算市场市占率连续多年接近或超过 50%。浪潮信息推动 AI 领域开放计算的发展，参与制定了 OCP 社区的 OAM 规范以及 ODCC 社区的 GPU 服务器规范，为不同的 AI 技术提供统一的技术标准。JDM（Joint Design Manufacture，联合设计制造）模式，是浪潮提出的区别于传统 OEM、ODM 的一种供应链模式，即让客户参与到服务器产品的设计、研发和交付

的流程中，实现全运营链定制化，开启了服务器产业从大规模标准化到需求驱动的大规模定制化时代。

通用服务器方面，代表产品有 NF5280M6 机架式服务器、ORS3000S 数据中心液冷整机柜服务器；AI 服务器方面，浪潮代表产品有 NF5688M6（拥有高达 5 PFLOPS 的强大 AI 计算性能）、NF5468M6-行业专属。

图表112： 浪潮信息服务器产品体系

产品样图				
产品型号	浪潮NF5280M6	浪潮ORS3000S	浪潮NF5688M6	浪潮NF5468M6-行业专属
核心架构	2个英特尔®至强系列第三代可扩展处理器	-	Intel Ice Lake CPU*2 NVIDIA A800 GPU*8	支持2 CPU+8 GPU；支持CPU+业界各种AI加速卡
应用场景	数据分析处理、深度学习、分布式存储	数据中心	超大规模数据中心	互联网AI公有云、企业级AI云平台、视频编解码

数据来源：浪潮信息，中信建投

3.5.2 高性能计算及国产化服务器龙头——中科曙光

中科曙光作为我国核心信息基础设施领军企业，为中国及全球用户提供创新、高效、可靠的 IT 产品、解决方案及服务。作为中科院产业化联盟的一员，中科曙光始终坚持自主创新，努力实现服务器、芯片国产化。普通服务器方面，中科曙光基于海光、AMD 合作生产的 EPYC 处理器，推出了天阔 A620-G30 机架式服务器；基于龙芯中科生产的龙芯 3B4000 处理器，推出了 L620-G35 机架式服务器。在信创产业进一步发展的未来，基于国产处理器的曙光服务器能在国产化进程中占据有利的竞争地位。在 AI 服务器方面，中科曙光推出了兼备训练与推理功能的全能型 GPU 服务器 X785-G40。

中科曙光参股国产 CPU 稀缺标的海光信息，积极布局服务器上游行业。多家国产服务器厂商采用了海光芯片，在电信行业信创招标中，搭载海光芯片的国产服务器占比逐渐升；海光信息于科创板上市后，为中科曙光的服务器业务扩张进一步助力。公司子公司曙光数创拥有浸没式相变液冷核心技术，预计将广泛应用于 AI 服务器散热领域。

此外，中科曙光还积极布局全栈云相关业务，覆盖 IaaS、PaaS 等方面，结合不同行业需要为国内企业上云提供技术支持，根据 IDC 报告，曙光政务云位列 2020 年中国政务云服务运营厂商市场第一阵营。在未来数字经济繁荣发展的大背景下，中科曙光的云业务也有望迎来较快增长。

3.5.3 华为昇腾+鲲鹏核心合作伙伴——拓维信息

拓维信息是中国领先的软硬一体化产品及解决方案提供商。公司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 多个国家，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。

拓维信息是鸿蒙生态重要建设者、鲲鹏战略合作伙伴、昇腾战略合作伙伴、华为云同舟共济战略合作伙伴，以基石研究院为中心，借助华为鸿蒙、昇腾、鲲鹏等核心技术，促进公司云计算、AI、物联网、大数据等核心能力的提升，产品包括咨询服务、运营服务、解决方案和 PaaS 产品。

图表113： 拓维信息研发体系



数据来源：拓维信息，中信建投

3.5.4 全球领先的 ICT 设备企业——联想集团

联想集团作为全球领先 ICT 科技企业，秉承“智能，为每一个可能”的理念，为用户与全行业提供整合了应用、服务和最佳体验的智能终端，以及强大的云基础设施与行业智能解决方案。联想与众多国际芯片厂商合作密切，更是与英伟达建立了战略合作伙伴关系，为联想 AI 服务器芯片的长期稳定供应打下了坚实基础。

联想集团具备全球供应能力。联想曾收购 IBM 的 X86 服务器业务，借助 IBM 的成熟市场完成了联想的全球供应链布局；同时，联想的 ODM+模式在海外优势明显，相比于传统 ODM 企业可以更快触达客户，且相比于戴尔与惠普两大全球服务器巨头具备更低的供应链成本，能为客户提供更具性价比的选择。

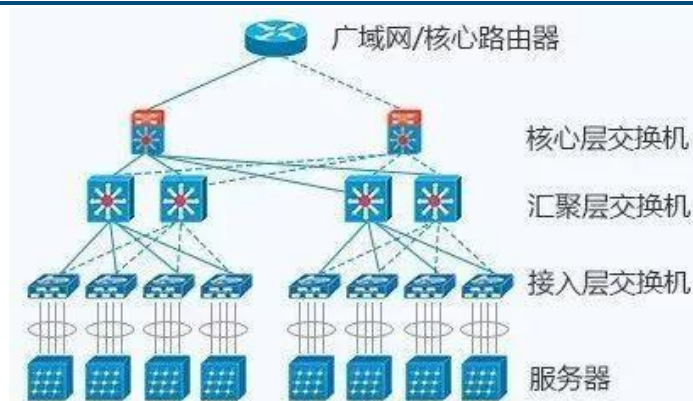
联想在全球个人 PC 市场深耕近 20 年，积累大量的 PC 成熟客户和完善的经销渠道；同时，在产业数字化的背景下，PC 客户大都有服务器的购买需求，这些都将成为联想服务器的潜在客户，这也使得联想集团在全球服务器市场竞争中占得了先机。公司主要客户微软、甲骨文、字节跳动等当前 AI 服务器订单充分，预期今年将拉动 ISG 部门实现超市场平均增速 20%的水平。

四、AI 正在推动高速率光模块需求放量

在传统的数据中心中，网络侧主要包括传统树形三层架构和叶脊架构。早期的数据中心一般采用传统的三层结构，包括接入层、汇聚层和核心层，其中接入层用于连接计算节点与机柜交换机，汇聚层用于接入层的互联，核心层用于汇聚层的互联且实现与外部网络连接。随着数据中心内部东西向流量的快速提升，三层网络架构的核心层和汇聚层任务加重，性能提升需求高，设备成本将大幅提升。因此，适用于东西向流量的扁平化的叶脊网络架构应运而生，叶交换机直接与计算节点相连，脊交换机相当于核心交换机，通过 ECMP 动态选择多条路径。叶脊网络架构具备带宽利用率高、扩展性好、网络延迟可预测和安全性高等优势，在数据中心中实现

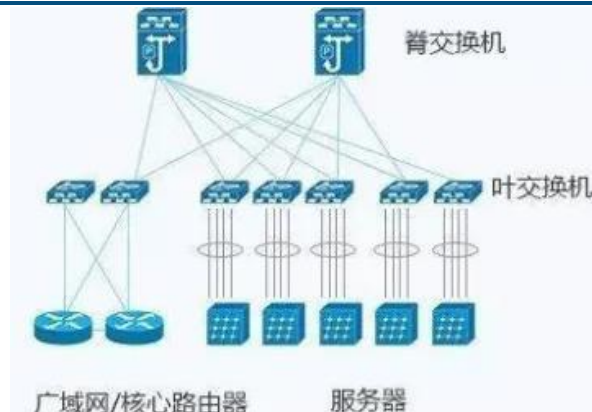
广泛的应用。

图表114：传统三层网络架构



数据来源：鲜枣课堂，中信建投证券

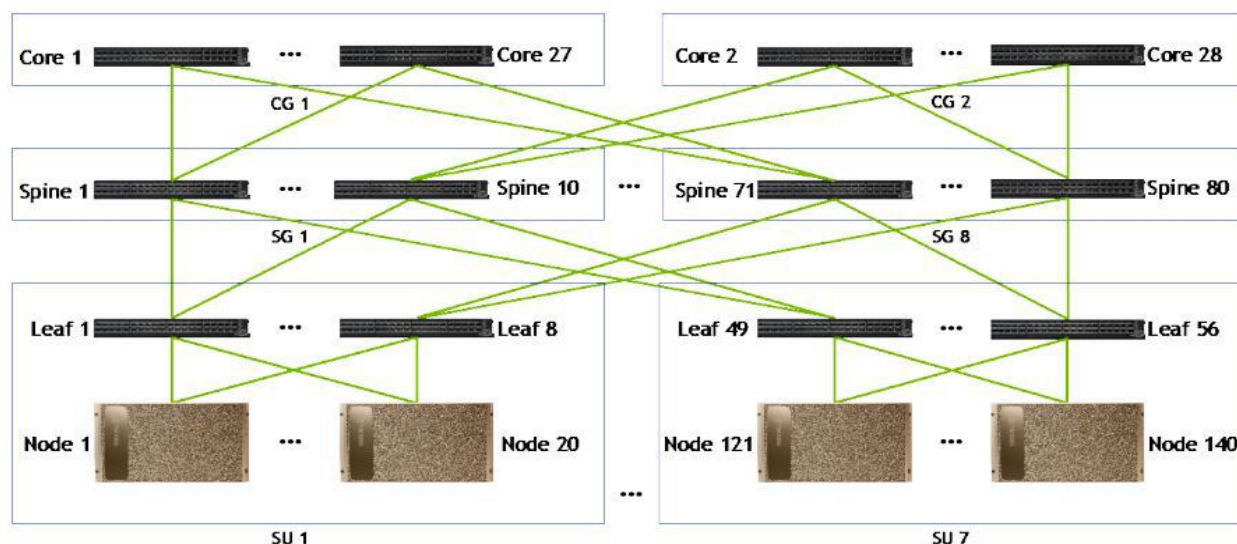
图表115：叶脊网络架构



数据来源：鲜枣课堂，中信建投证券

AI 数据中心中，由于内部数据流量较大，因此无阻塞的胖树网络架构成了重要需求之一。英伟达的 AI 数据中心中，采用了胖树（fat-tree）的网络架构来实现无阻塞的功能。胖树的网络架构基本理念为：使用大量低性能的交换机，构建出大规模的无阻塞网络，对于任意的通信模式，总有路径让他们的通信带宽达到网卡带宽，架构中用到的所有交换机都是相同的。胖树网络架构一般用于网络要求较高的数据中心中，如超算中心和 AI 数据中心等。

图表116：英伟达 DGX A100 SuperPOD 采用胖树网络三层架构示意图



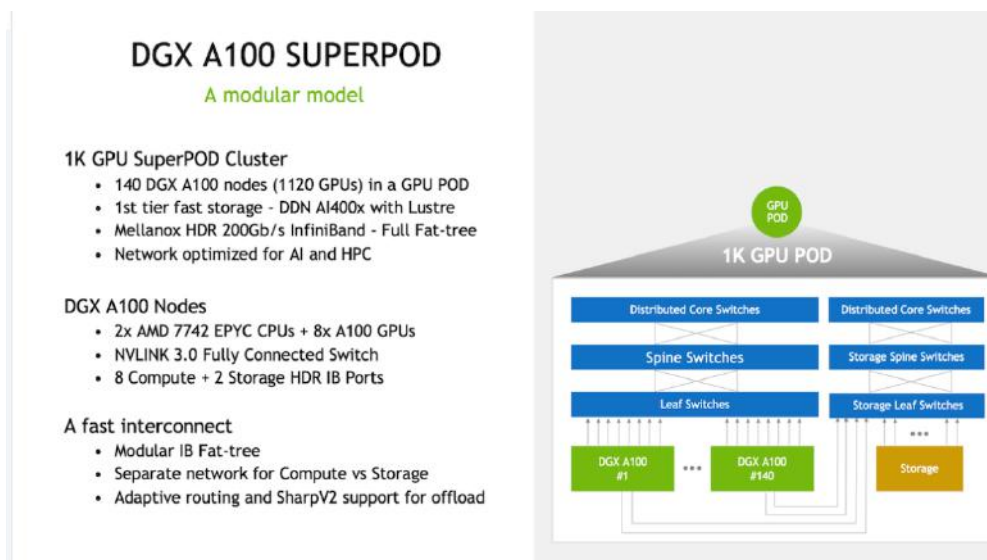
数据来源：英伟达，中信建投证券

在英伟达 DGX A100 SuperPOD 的 AI 数据中心系统中，三层交换机全部为 Nvidia Quantum QM8790 的 40 端口交换机。第一层交换机与 1120 张 Mellanox HDR 200G Infiniband 网卡连接；第二层交换机下传端口与第一层相连，上传端口与第三层互联；第三层交换机只有下传端口，与第二层相连。此外，存储侧独立组网，与计算侧网络架构分开，也需要一定数量的交换机和光模块。因此，相比较传统数据中心，AI 数据中心中的交换机及光模块数量大幅提升。根据我们的测算，训练端 A100 和 200G 光模块的比例是 1:7，H100 和 800G 光模块的比例

请参阅最后一页的重要声明

是 1:3.5。

图表117： 英伟达 DGX A100 SuperPOD 系统示意图



数据来源：英伟达，中信建投证券

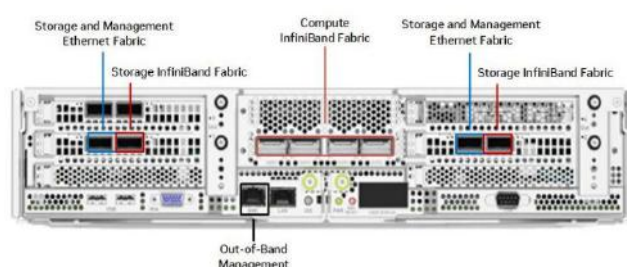
英伟达的 A100 GPU 主要对应 200G 光模块，H100 GPU 可以对应 400G 或 800G 光模块。每个 A100 GPU 配一张 Mellanox HDR 200Gb/s Infiniband 网卡，每个 H100 GPU 配一张 Mellanox NDR 400Gb/s Infiniband 网卡。英伟达在 H100 SuperPOD 的设计中，采用了 800G 的光模块，在光口采用 1 个 800G 光模块可以替代 2 个 400G 光模块，在电口也可以将 8 个 SerDes 通道进行整合，与光口的 8 个 100G 通道一一对应。因此这种设计下，交换机的通道密度提高，物理尺寸显著降低。

图表118： Mellanox HDR 200Gb/s Infiniband 网卡示意图



数据来源：英伟达，中信建投证券

图表119： DGX H100 服务器背板连接图



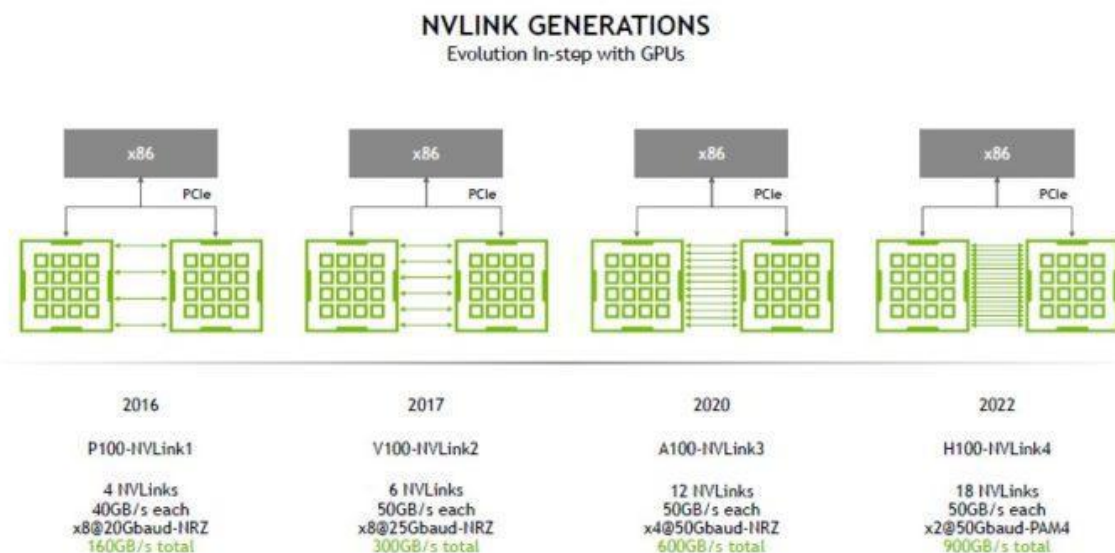
数据来源：英伟达，中信建投证券

光模块速率由网卡决定，网卡的速率受限于 PCIe 通道速率。英伟达 A100 的 DGX 服务器内部通过 NVLink3 连接，单向带宽为 300GB/s，但是 A100 GPU 连接 ConnectX-6 网卡是通过 16 个 PCIe 4.0 通道，带宽总和为 200G 左右，因此网卡带宽为 200G，需要连接 200G 的光模块或者 DAC 电缆。H100 的 DGX 服务器内部通过 NVLink4 连接，单向带宽为 450GB/s，但是 H100 GPU 连接 ConnectX-7 网卡是通过 16 个 PCIe 5.0 通道，带宽总和为 400G 左右，因此单个网卡带宽为 400G。可以看出，光模块速率是由于网卡与 GPU 之间的 PCIe 带宽所决定。假设 A100

请参阅最后一页的重要声明

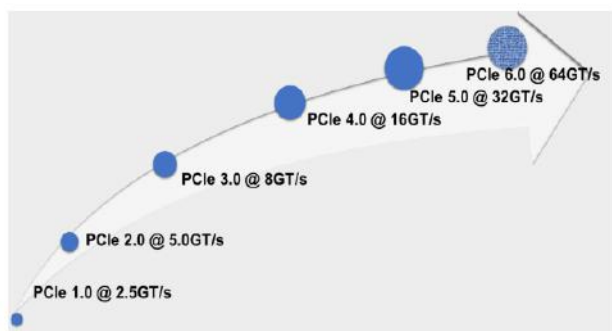
和 H100 的 DGX 服务器内部所用 PCIe 通道速率达到 800G（即 PCIe 6.0），那么也可以采用 800G 带宽的网卡，即也可以采用 800G 光模块，大大提升系统计算效率。

图表120： NVLink 不同代际的升级 Roadmap



数据来源：STH 网站，英伟达，中信建投证券

图表121： PCIe 不同代际的性能参数表



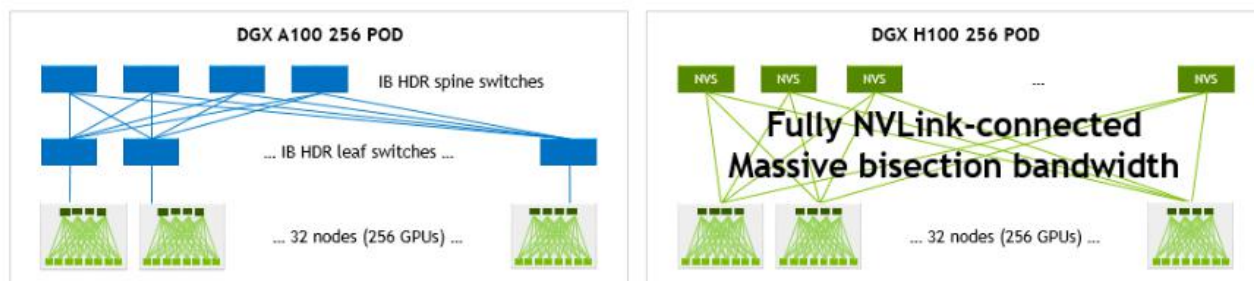
PCIe Specification	Data Rate(Gb/s) (Encoding)	x16 B/W per dirn*	Year
1.0	2.5 (8b/10b)	32 Gb/s	2003
2.0	5.0 (8b/10b)	64 Gb/s	2007
3.0	8.0 (128b/130b)	128 Gb/s	2010
4.0	16.0 (128b/130b)	252 Gb/s	2017
5.0	32.0 (128b/130b)	504 Gb/s	2019
6.0	64.0 (PAM-4, Flit)	1024 Gb/s (~1Tb/s)	2021

* - bandwidth after encoding overhead

数据来源：PCI-SIG，中信建投证券

NVLink 带宽远大于网卡侧的 PCIe 带宽，因此若将 NVLink 从服务器内部 GPU 互连拓宽至不同服务器之间的 GPU 的互连，将显著提升系统的带宽。若要实现不同服务器之间按照 NVLink 协议的 GPU 互连，除了需要采用 NVSwitch 芯片的物理交换机，还需要物理器件来实现交换机和服务器之间的连接，那么光模块也成为了重要的组成部分，从而也会大幅增长 800G 光模块的需求。近日，英伟达创始人兼 CEO 黄仁勋在 NVIDIA Computex 2023 演讲中宣布，生成式 AI 引擎 NVIDIA DGX GH200 现已投入量产。GH200 通过 NV Link4 的 900GB/s 超大网络带宽能力来提升算力，服务器内部可能采用铜线方案，但服务器之间我们认为可能会用光纤连接。对于单个 256 GH200 芯片的集群，计算侧 1 个 GH200 对应 9 个 800G 光模块；对于多个 256 的 GH200 集群，计算侧 1 个 GH200 对应 12 个 800G 光模块。

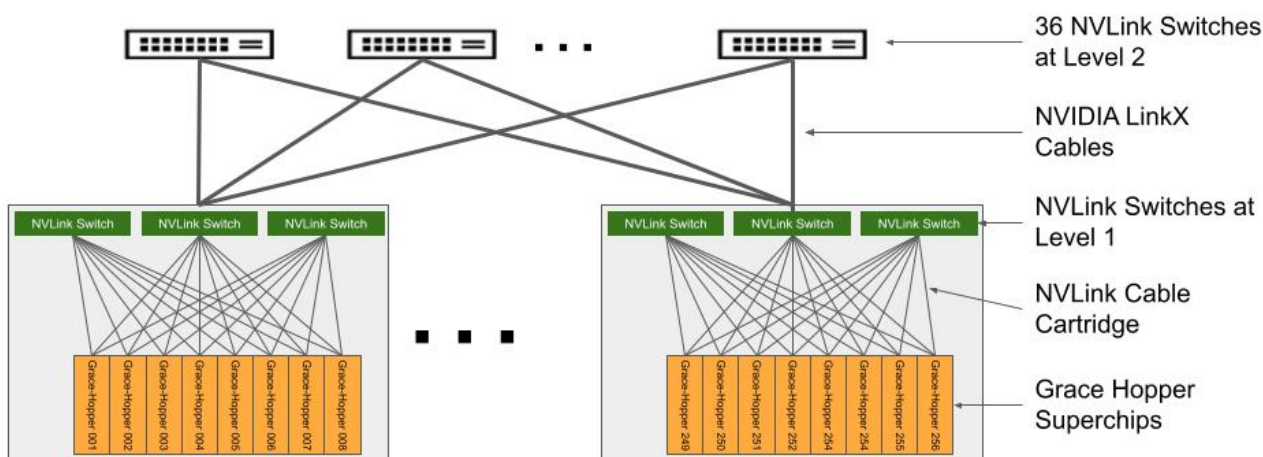
图表122: A100 和 H100 POD 采用 IB 和 NVLink 网络的示意图



数据来源: 英伟达, 中信建投证券

图表123: GH200 的网络连接示意图

Fully Connected NVLink across 256 GPUs



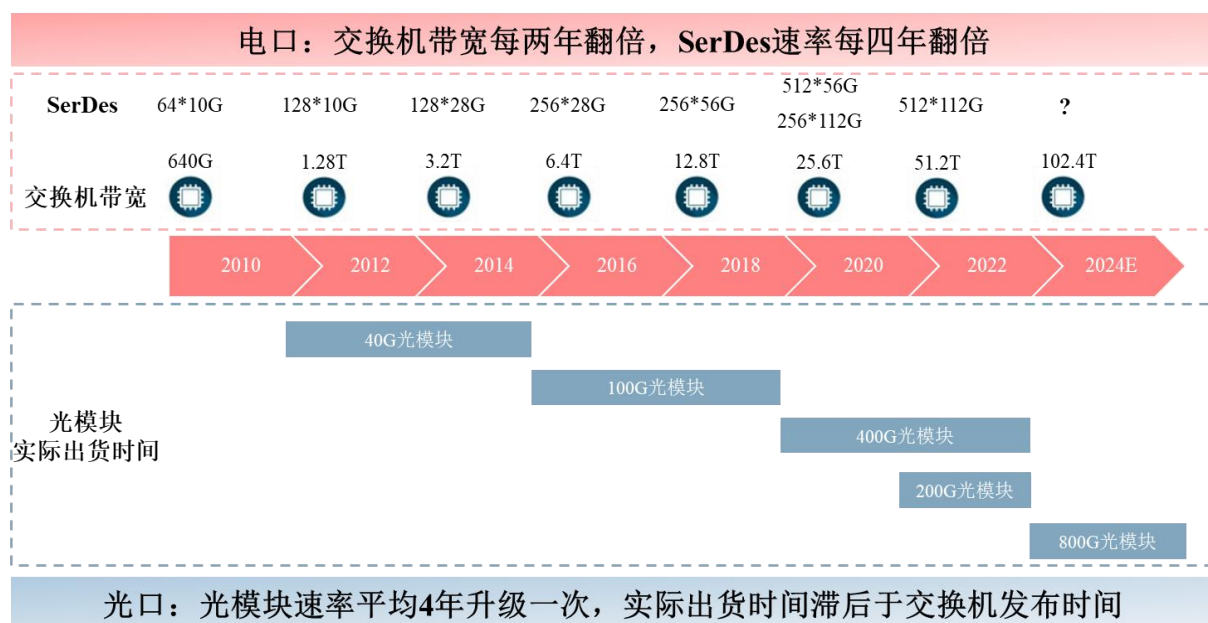
数据来源: 英伟达, 中信建投证券

训练侧光模块需求与 GPU 出货量强相关, 推理侧光模块需求与数据流量强相关。AI 对光模块需求的拉升主要分为两个阶段, 训练和推理。其中, 训练侧的网络架构以胖树架构为主, 因为在大模型训练过程中, 对于网络性能的要求很高, 网络无阻塞是重要的需求之一, 比如腾讯用于大模型训练的星脉网络采用了胖树架构。同时, 我们认为大部分厂商会采用 Infiniband 协议的网络, 时延远低于以太网, 可以提升计算效率, 缩短模型训练时间。训练侧光模块的需求与所用 GPU 显卡的数量强相关, 根据胖树架构中 GPU 和光模块的比例关系可以得到所需光模块的数量, A100 对应 200G 光模块, H100 对应 400G 或者 800G 光模块。推理侧面向用户侧, 网络架构更接近于传统云计算数据中心的叶脊架构, 主要用于承载 AI 应用带来的数据流量增量。传统云计算主要是 ToB 市场, 用户数量不多, 若未来出现图片或视频相关的爆款 AI 应用, 一方面用户数量有望大幅提升, 另一方面单个用户产生的数据流量可能会显著增长, 因此数据总流量将暴增, 所以推理所需的算力和流量实际上可能远大于训练, 因此对于包括光模块在内的网络设备需求将起到有力的支撑和提振。

请参阅最后一页的重要声明

800G 光模块 2022 年底开始小批量出货，2023 年需求主要来自于英伟达和谷歌，2024 年有望大规模出货，并存在时间前移的可能。从交换机的电口来看，SerDes 通道的速率每四年翻倍，数量每两年翻倍，交换机的带宽每两年翻倍；从光口来看，光模块每 4 年升级一次，实际出货时间是晚于电口 SerDes 及交换机芯片新版发布的时间。2019 年作为 100G 光模块升级的时间点，市场分成了 200G 和 400G 两条升级路径。但是在 2023 年这个时间点，市场下一代高速率光模块均指向 800G 光模块，叠加 AIGC 带来的算力和模型竞赛，我们预计北美各大云厂商和相关科技巨头均有望在 2024 年大量采购 800G 光模块，同时 2023 年也可能提前采购。

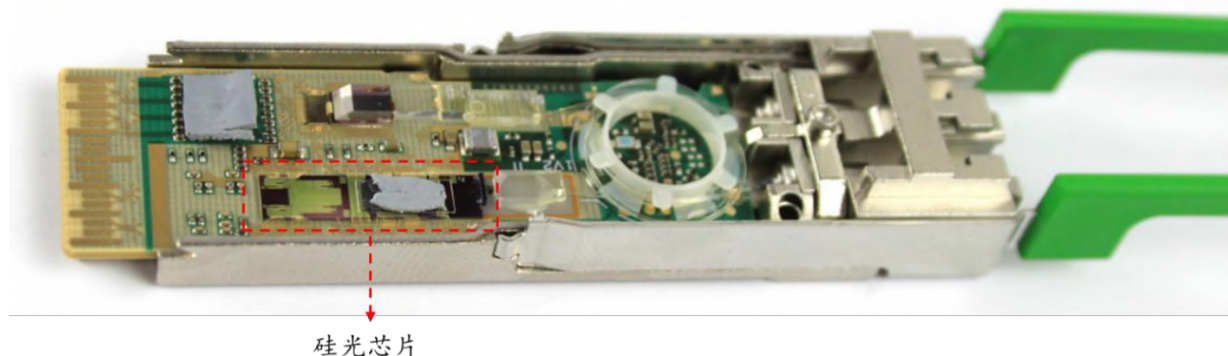
图表124： GH200 的网络连接示意图



数据来源：思科，中信建投证券

硅光子技术是以硅或硅基材料（Si, SiO₂, SiGe）作为衬底材料，利用与集成电路兼容的 CMOS 工艺制造对应的光子器件和光电器件，以实现光的激发，调制，响应等，广泛应用于光通信，光传感，高性能计算等。数通领域的硅光模块同样实现了大规模商用，未来份额有望不断提升。随着数据中心的快速发展，对于光模块的需求爆发式增长，多家厂商开始大力研发用于数据中心的硅光模块。初期是 40G 硅光数通光模块小规模应用，Intel 和 Luxtera 的 100G 硅光模块大规模应用，目前 400G 的硅光模块已经实现量产，800G 亦在验证中。目前国内的硅光模块厂商具备较强的竞争力，包括中际旭创、新易盛、华工科技等公司有自研的硅光芯片，博创科技等公司与海外硅光芯片巨头厂商深度合作，有望在 800G 光模块市场取得突破。

图表125： Intel 的 100G 硅光模块示意图



数据来源：SystemPlus，中信建投证券

铌酸锂材料的优势在调制器上体现，目前主要应用在电信领域。LiNbO₃ 具有电光系数大、本征调制带宽大、波导传输损耗小、稳定性好等优点，同时也有偏振敏感、尺寸大、调制电压高的缺点。LiNbO₃ 调制器是目前发展较成熟的调制器，其利用线性电光效应实现电信号对光信号的调制，通过外加电场改变光在晶体中传播的折射率，进而改变光的相位和偏振。目前 LiNbO₃ 调制器的应用场景主要在长距离的相干光通信领域以及军事及航天的陀螺仪等产品中，未来有望应用到 800G 等更高速率的数通光模块中。

图表126： 硅光、InP、体材料铌酸锂和薄膜铌酸锂调制器的对比示意图

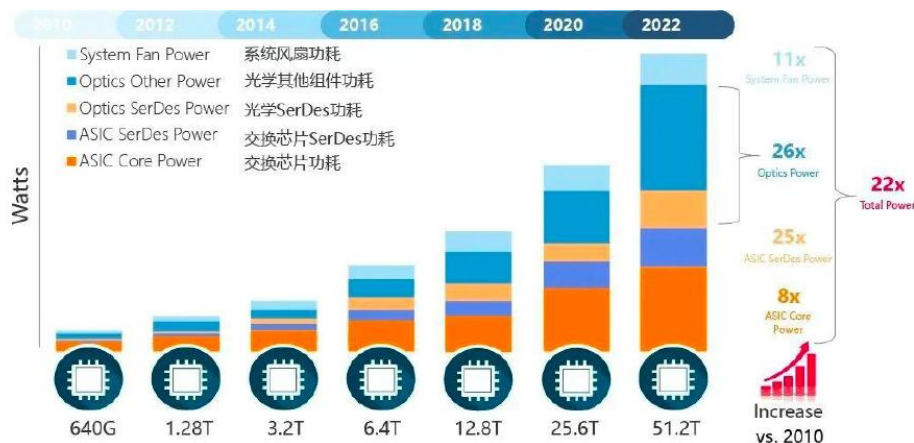
调制器种类	优点	缺点	应用场景	主要供应商
硅光	- 兼容CMOS工艺 - 低成本 - 低功耗	- 耦合损耗大 - 光源集成难度高 - 偏振敏感	- 相干光模块 - 数通光模块 - CPO交换机中	Intel、Cisco等
磷化铟	- 调制效率高 - 驱动电压小 - 与光源易于集成	- 工艺要求高 - 成本高 - 材料要求高	中短距离传输场景	Lumentum、II-VI和新飞通等
体材料铌酸锂	- 调制带宽大 - 传输损耗小 - 稳定性好	- 尺寸较大 - 偏振敏感 - 调制电压高	- 相干光通信 - 军事及航天	富士通、住友、光库（Lumentum）
薄膜铌酸锂	- 尺寸小 - 易于集成 - 功耗低	- 工艺难度大 - 带宽降低 - 技术尚不成熟	- 相干光通信 - 军事及航天 - 数通光模块	富士通、住友、铌奥、光库

数据来源：光库科技，中信建投证券

Co-packaged Optics，即共封装光学，光学引擎 PIC 与电学引擎 EIC 合封在一起的封装技术。CPO 交换机主要分为交换机芯片、SerDes 和光学部分，过去 10 年交换机带宽增长了 80 倍。交换机芯片的带宽每两年提升一倍；电接口的 SerDes 数量和速率也在提升，速率从 10G/s 提升到 112G/s，数量从 64 个通道提升到 51.2T 时代的 512 个通道。交换机带宽从 640G 提升到 51.2T，交换机芯片功耗提升 7.4 倍，每个 Serdes 通道的功耗提升 2.84 倍，结合 Serdes 通道数的增加，总功耗增加 22.7 倍。而 CPO 可以降低功耗（核心优势）、降低成本和减小尺寸。CPO 参与公司主要包括云服务厂商、设备商和芯片厂商等。目前，CPO 仍有很多技术难题，例如光源的功耗问题，光源作为核心的部件之一，虽然外部光源在配置上更加灵活，但是激光器在高温下效率较低，因此给多个

通道同时提供光源时，高功率带来低效率，其功耗反而会更高。而且，光引擎紧密排布在交换机芯片的周围，巨大的发热量如何进行有效地散热，光引擎失效后如何进行灵活地更换，新的光学连接器如何定义等这些技术难题都需要更加有效的解决方案。此外，CPO 产品是将光模块和交换机集成在一起，因此将对光模块和交换机行业产生较大的影响，在制定好相关产品标准之后如何使得两个产业链更好的协同，也将是一个重要的挑战。

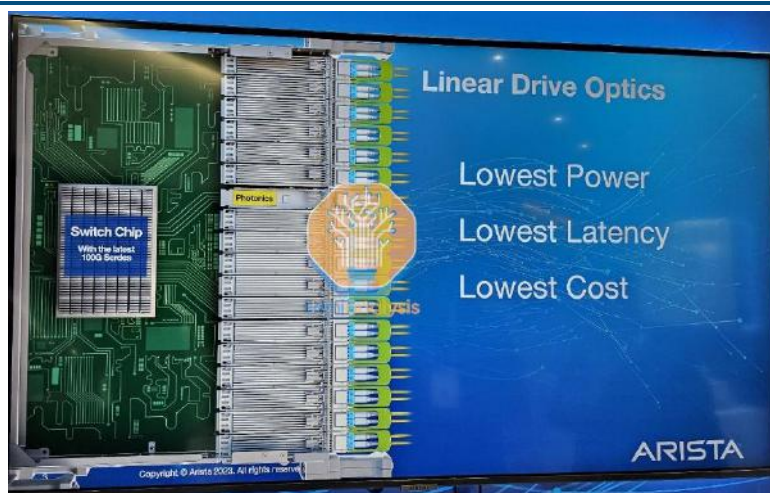
图表127： 交换机发展示意图



数据来源：思科，中信建投证券

LPO，即 linear drive pluggable optics，线性直驱可插拔光模块。主要方式为，在光模块中不再采用 DSP，只留下 driver 和 TIA 等电芯片，而将 DSP 的功能集成到交换芯片中。LPO 光模块中的 driver 和 TIA 需要分别集成 CTLE 和 Equalization 功能，用于对高速信号进行一定程度的补偿。Driver 的主要功能是线性放大，输出电压也是线性变化的。从交换机中发出的信号，不再需要通过 CDR 恢复产生，而是直接传递给 driver，进行线性调制。相比较传统的带 DSP 的光模块，LPO 光模块可以降低功耗、延迟和成本。我们认为，LPO 光模块的核心在于交换机芯片，如交换机芯片进展顺利，性能优异，也将直接推动 LPO 光模块的进展。

图表128： LPO 方案的优势



数据来源：Arista，中信建投证券

多家光模块厂商具备 800G 光模块能力，国内多家厂商具备较强的竞争力。在 2023 年的 OFC 光博会上，各

家光模块公司均推出了自己的 800G 光模块产品，涵盖不同封装方式、材料和传输距离等种类。值得一提的是，国内厂商在 100G 和 400G 光模块时代已经取得了显著的进展，跻身全球先进水平。在数通 800G 光模块时代，以中际旭创和新易盛为代表的国内厂商已经在海外云厂商的供应链体系中，确定性较强。而华工科技、剑桥科技、联特科技、博创科技、光迅科技和德科立等公司也有望取得突破，同样值得重视。

图表129：光模块厂商目前拥有的 800G 光模块产品

光模块厂商	800G 光模块产品
中际旭创	20230FC 推出了其基于 5nm DSP 和先进硅光子技术的第二代 800G 模块，同时拥有功耗低于 14W 的 800G OSFP DR8+和 2xFR4 光通信模块。公司具备 800G 全系列光模块产品，包括不同封装和传输距离，竞争力保持全球领先。
新易盛	20230FC 现场演示基于薄膜铌酸锂（TFLN）调制器技术的 800G OSFP DR8 光模块产品，搭配 5nm DSP 芯片，功耗 11.2W。同时推出 LPO 光模块，包含 EML、TFLN 和 SiPh 三种方案。
华工科技	公司的 800G SR8 已经在国内市场头部厂商送样测试，DR8 和 FR8 产品在微软和英伟达同样，预计 23 年下半年出货量将快速增长。
剑桥科技	公司的 800G 光模块基于传统 EML 和硅光两种方案，对于薄膜铌酸锂方案也在积极关注。公司也推出了线性驱动的 800G 光模块产品。
光迅科技	20230FC 上展示了 800G QSFP-DD800 的 SR8 光模块产品，同时公司拥有 800G QSFP-DD 2x400G FR4 和 DR8 光模块。
光模块厂商	800G 光模块产品
博创科技	公司具备 800G 硅光模块的产品能力，同时也在研发 CPO 相关产品。
源杰科技	公司具备 10G EML 和 25G DFB 激光器芯片能力，预计今年发布 100G PAM4 EML 激光器芯片，主要用于 400G 和 800G 光模块。
Coherent	公司目前拥有 800G OSFP DR8 和 QSFP-DD800 2xFR4 等产品。
Intel	公司具备 800G OSFP DR8 硅光模块、2x400G FR4 硅光模块等产品。

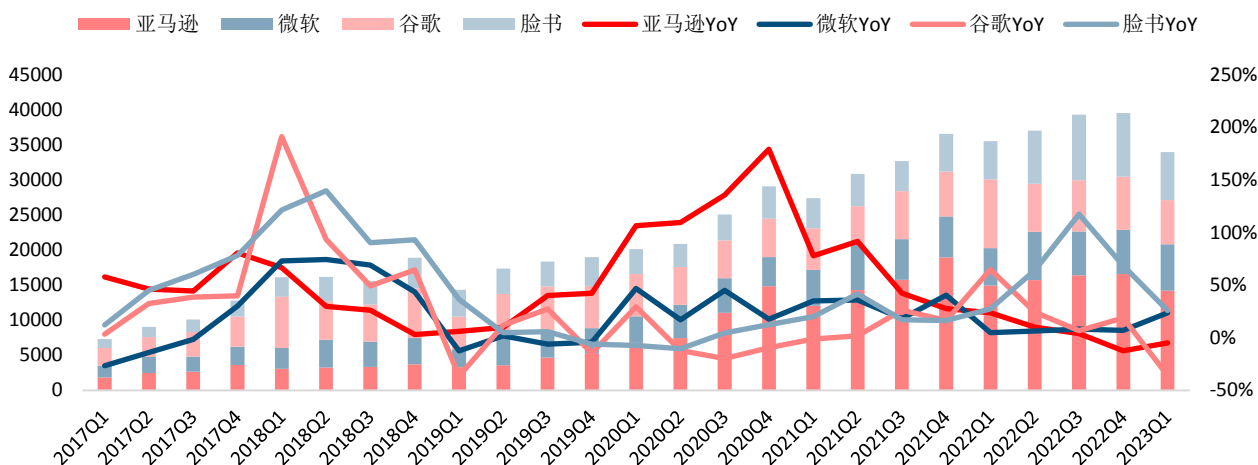
资料来源：中际旭创，新易盛，华工科技，剑桥科技，光迅科技，源杰科技，中信建投证券

我们认为，本轮光模块板块行情可以参考 2016-2018H1 与 2019H2-2020H1。

数通光模块行业在 2016-2018H1 处于景气周期，中际旭创期间股价表现较好，2018H2-2019H1 全球云计算及互联网巨头资本开支迎来调整，期间股价也下行。北美 FAAM（Facebook、Amazon、Alphabet、Microsoft）2016-2018 Capex 增速为 29.65%、27.94%、62.74%，虽然 2018 年全年增速强劲，但 2018Q3 起增速显著放缓。经过近 3 年（2016-2018H1）的景气周期，云厂商基础设施如服务器、光网络等利用率不够饱满，相当于计算、存储、网络能力有一定的“库存”，叠加宏观经济及中美摩擦导致的不确定性，企业信息化投入收缩，企业上云放缓，互联网巨头面临增长压力，因此资本开支增速明显放缓，直至 2019Q1 资本开支负增长。

2018H2-2019H1 北美云基础设施需求放缓，只是云厂商的“库存”调整。2019Q2 之后北美云厂商的资本开支同比出现增长，整体延续回暖态势，其中亚马逊、谷歌较为显著，亚马逊 2019Q3 Capex 同比增长 40.13%，中际旭创股价在 2019 年下半年开始反应市场预期。2020-2022H1 年北美四家云厂商资本开支高增近 3 年，2022 年下半年资本开支明显降速。2022Q4，北美四家云厂商资本开支 395.04 亿美元，同比增长 8.07%，明显降速（2022 年前三季度单季同比增速基本都在 20%或以上）。亚马逊作为资本开支大户，2022Q4 出现 2015Q4 以来第一次单季度负增长，下降 12.37%。因此，2022 年虽然光模块公司业绩普遍表现较好，但股价与估值不断下跌。

图表130： 北美云厂商资本开支（百万美元）



数据来源: Bloomberg, 中信建投

图表131: 中际旭创股价复盘



数据来源: wind, 中信建投

从历史估值来看，中际旭创 2019 年 PE-TTM 高点时超过 70 倍，2020 年高点超过 100 倍，过去 5 年平均 PE-TTM 为 47.28 倍，2019 年、2020 年基于当年业绩与当年市值的 PE 为 72.49x、41.91x。

我们认为，在 AI 带动下，叠加宏观经济企稳，数字经济发展，国内外云厂商资本开支有望在今年企稳，在 2024 年或将出现显著提升，因此本轮光模块行情走势建议参考 2019Q1-2020Q2，同时我们认为是板块性行情，因此**建议重点关注中际旭创、天孚通信、新易盛、华工科技、源杰科技、太辰光、光迅科技、光库科技、中瓷电子、剑桥科技、博创科技、联特科技、德科立、仕佳光子等。**

请参阅最后一页的重要声明

五、AI 将会拉动交换机市场需求

AI 带来数据中心的网络架构变化，光模块速率及数量均有显著提升，因此交换机的端口数及端口速率也有相应的增长。以 ChatGPT 为代表的 AIGC 技术，依靠强大的 AI 模型和海量数据，能够在多个应用场景下产生优质的内容，有望推动人工智能更广泛的应用。算力作为 AIGC 技术的重要支撑之一，是影响 AI 发展与应用的核心因素。算力基础设施成了目前行业亟需布局的资源，除了 CPU/GPU 等算力硬件需求强劲，网络端也催生了更大带宽需求，以匹配日益增长的流量。与传统数据中心的网络架构相比，AI 数据网络架构会带来更多的交换机端口的需求。

图表132： 微软 Azure 的 DGX H100 AI 超级计算机系统



数据来源：英伟达，中信建投证券

训练侧大概率会采用 Infiniband 或者类 IB 的低时延网络协议，推理侧预计会采用以太网协议的交换机。InfiniBand 是一种开放标准的高带宽，低时延，高可靠的网络互联技术，随着人工智能的兴起，也是 GPU 服务器首选的网络互联技术。相比较以太网协议的网络，Infiniband 网络在带宽、时延、网络可靠性、和组网方式上都有一定的优势。当然，以太网的兼容性更好，成本更低，可以应用在各种应用场景中，适配各种不同的设备终端。AI 训练端对时延要求较高，因此训练侧大概率会采用 Infiniband 网络，也可以采用 ROCE 网络，即基于以太网的 RDMA 技术，也能够达到较低的时延。而英伟达 NVLink 技术，其带宽大幅提升，NVLink4 的双向带宽可以达到 900GB/s，在训练侧也将具备较强的优势。在推理侧，我们认为网络协议可以沿用云计算数据中心的以太网。

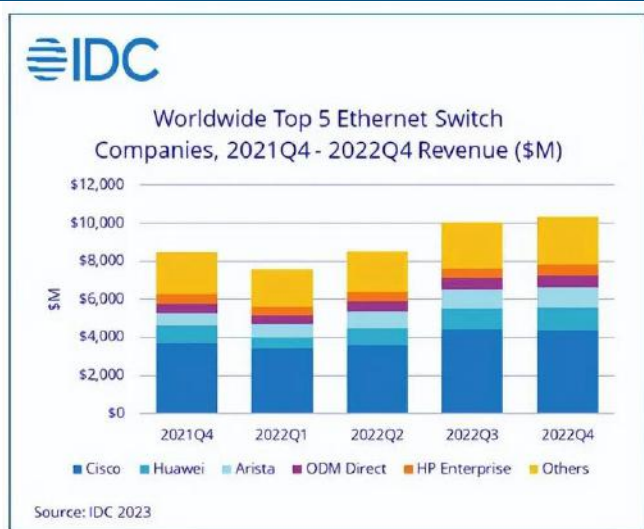
图表133： 不同网络架构的对比



数据来源：英伟达，中信建投证券

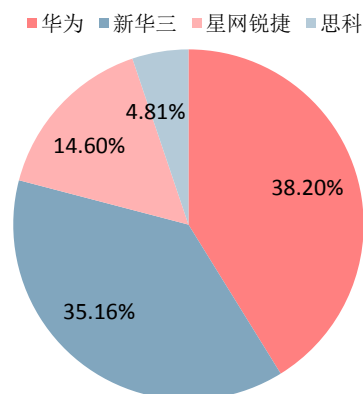
交换机具备技术壁垒，中国市场格局稳定，华为与新华三（紫光股份）两强争霸，锐捷网络展现追赶势头。全球来看，思科一家独大，份额近 50%，但呈现下滑趋势，华为列全球第二（9%）、新华三列第五（4.5%）。华为在数据中心、电信运营商市场均展现出较强竞争力，新华三与锐捷网络目前均以数据中心为主，正突破运营商。其中，锐捷网络近年进一步取得突破，在前期中国移动招标中，系仅有的两家中标者之一，获近 50% 份额。交换机除了应用于数据中心、电信运营商外，还有政企市场。**建议重点关注：紫光股份、锐捷网络等。**

图表134： 2022 年全球前五大以太网交换机厂商



数据来源：IDC，中信建投

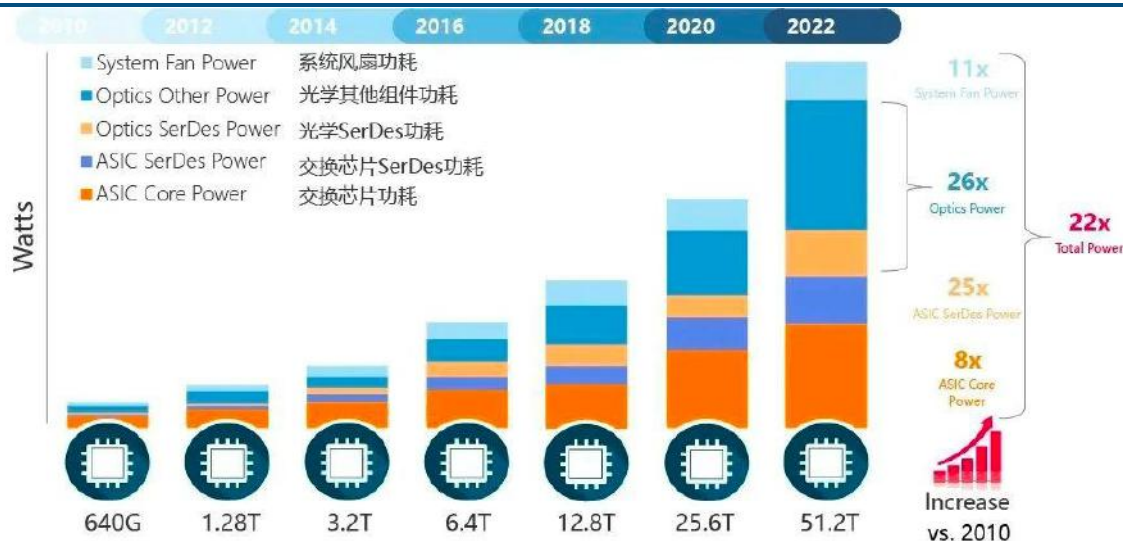
图表135： 2021 年中国交换机市场份额



数据来源：IDC，中信建投

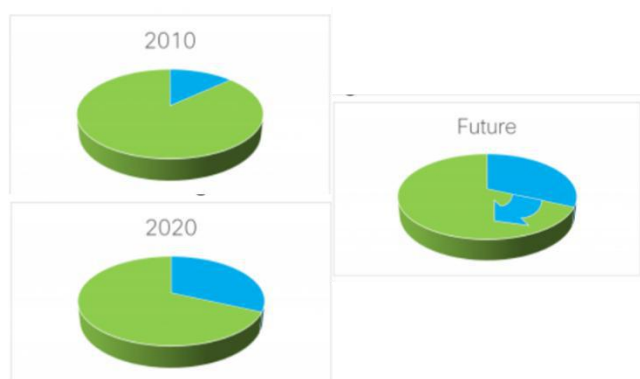
交换机中 SerDes 的功耗大幅提升。随着单个 SerDes 带宽提升带来功耗的提升，同时结合 SerDes 数量的提升，未来 SerDes 的总功耗在交换机中的功耗占比将大幅提升。**网络部分的功耗在数据中心中的功耗大幅提升：**根据 Facebook 的测算，随着数据中心内部流量的大幅提升，网络部分的功耗占比增加明显，到下一代网络部分的功耗占比将从现在的 2% 左右提升到 20% 左右。**传输距离越近，SerDes 功耗越低。**缩短交换机和光模块之间电信号需要传输的距离，可以简化 Serdes 芯片的功能，同时降低电信号的发射功率，从而降低 SerDes 的功耗。

图表136：交换机发展示意图



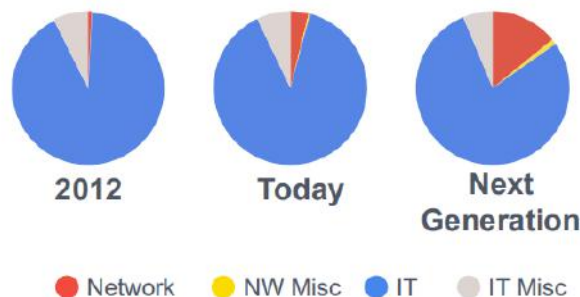
数据来源：Cisco，中信建投

图表137：交换机内部 SerDes 功耗占比大幅提升



数据来源：Cisco，中信建投

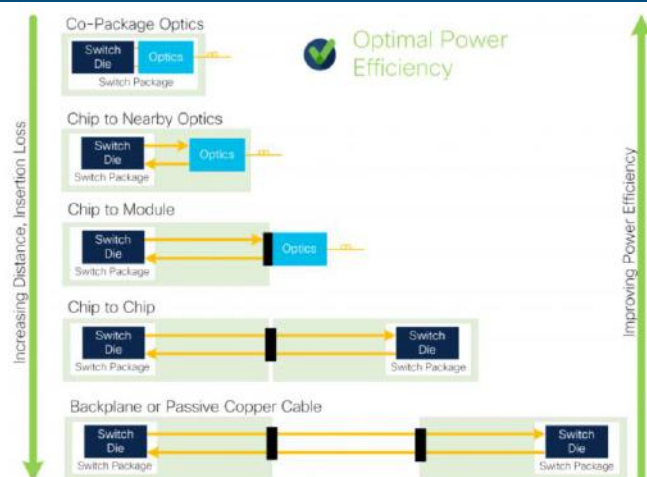
图表138：网络部分的功耗在数据中心中占比大幅提升



数据来源：Meta，中信建投

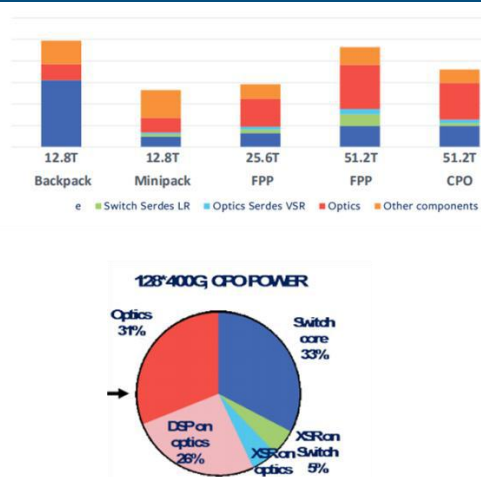
CPO 部署将在很大程度上受到交换演进的推动。交换演进将在 2025 年达到 102.4Tbps。一旦交换达到这个水平，可插拔收发器将逐渐消失，与使用可插拔光学器件相比，CPO 承诺将功耗降低 30%，每比特成本降低 40%。到 2027 年，共封装光学的市场收入将达到 54 亿美元，2025 年全球 CPO 组件市场将超 13 亿美元，到 2028 年将增长到 27 亿美元。根据 LightCounting 的报告，从长远来看，CPO 不局限于硅光、不局限于数据中心，还有更大的前景。在 2027 年，CPO 端口将占总 800G 和 1.6T 端口的近 30%。**CPO 参与公司主要包括云服务厂商、设备商和芯片厂商等。**Meta 在 2022 年的 OFC 会上展示了新一代的基于 51T ASIC 和 NPO 端口的交换机，4RU 的尺寸；Marvell 推出的 NPO，基于自家 Teralynx 交换芯片平台，集成到标准 1RU 32 端口设备中，未来计划发展到支持 51.2T 交换机的 3.2T CPO 平台；Intel 的样机计划于 2024 年上市，此前先后收购了 Optoscribe 和 Tower；博通在 2023 年 OFC 推出了 51.2T 的 CPO 产品；IBM 推出了基于 VCSEL 的 CPO 产品。**国内的紫光股份和锐捷网络等公司也均有布局 CPO 相关技术，有望紧跟行业演进趋势，保持竞争力。**

图表139： CPO 可以降低功耗



数据来源：Cisco，中信建投

图表140： CPO 所降低的功耗拆分示意图



数据来源：Meta，中信建投

六、AI 提升大功率 IDC 机柜需求，液冷渗透率随之提升

6.1 “东数西算”统筹全国算力网络建设，云计算需求可能将回暖

2021 年 5 月，发改委、网信办、工信部、能源局联合印发《全国一体化大数据中心协同创新体系算力枢纽实施方案》，明确提出布局全国算力网络国家枢纽节点，启动实施“东数西算”工程，构建国家算力网络体系。

《全国一体化大数据中心协同创新体系算力枢纽实施方案》围绕国家重大区域发展战略，根据能源结构、产业布局、市场发展、气候环境等，在京津冀、长三角、粤港澳大湾区、成渝以及贵州、内蒙古、甘肃、宁夏等地布局建设全国一体化算力网络国家枢纽节点，引导数据中心集约化、规模化、绿色化发展，构建数据中心集群。国家枢纽节点间将进一步打通网络传输通道，加快实施“东数西算”工程，提升跨区域算力调度水平。

根据《全国一体化大数据中心协同创新体系算力枢纽实施方案》要求，京津冀、长三角、粤港澳大湾区、成渝等节点，用户规模较大、应用需求强烈，要重点统筹好城市内部和周边区域的数据中心布局，优化数据中心供给结构，扩展算力增长空间，满足重大区域发展战略实施需要，城市内部加快对现有数据中心的改造升级，优先满足对实时性要求高的业务需求。贵州、内蒙古、甘肃、宁夏等节点，可再生能源丰富、气候适宜、数据中心绿色发展潜力较大，要重点提升算力服务品质和利用效率，充分发挥资源优势，夯实网络基础保障，积极承接全国范围需后台加工、离线分析、存储备份等非实时算力需求，打造面向全国的非实时性算力保障基地。

为实现全国一体化算力网络国家枢纽节点布局，就要在集群和集群之间建立高速数据中心直联网络，构建形成以数据流为导向的新型算力网络格局，助力实施“东数西算”工程，支撑大规模算力调度。从数据中心网络时延的产生来看，主要分为传输时延（受制于物理距离）和传输节点时延（受制于节点数量和单节点转发时延），减少长距离传输时延的方法主要是路由优化，提供更短的光缆路由。传统上我国通信网络主要围绕人口聚集程度进行建设，网络节点普遍集中于北上广等一线城市，数据中心对网络依赖性强，随之集中于城市部署。推进“东数西算”工程，就要推进网络一体化建设，夯实西部地区的网络基础保障，围绕集群建设数据中心直连网，增大网络带宽，提高传输速度，降低传输费用，推进新型互联网交换中心、互联网骨干直连点建设。

2022 年 2 月，国家发展改革委、中央网信办、工业和信息化部、国家能源局再次联合印发通知，同意在京津冀、长三角、粤港澳大湾区、成渝、内蒙古、贵州、甘肃、宁夏 8 地启动建设国家算力枢纽节点，规划了 10 个国家数据中心集群，标志着全国一体化大数据中心体系完成总体布局设计，“东数西算”工程正式全面启动。

图表141：“东数西算”工程设立 8 个节点



数据来源：发改委，中信建投

图表142：“东数西算”工程设立 10 个集群

东部	枢纽节点	集群	起步区政策
西部	贵州枢纽	贵安数据中心集群	贵安新区贵安电子信息产业园
	内蒙古枢纽	和林格尔数据中心集群	和林格尔新区、集宁大数据产业园
	甘肃枢纽	庆阳数据中心集群	庆阳西峰数字经济产业集聚区
	宁夏枢纽	中卫数据中心集群	中卫工业园区云计算基地
东部	京津冀枢纽	张家口数据中心集群	张家口市怀来县、张北县、宣化区
	长三角枢纽	长三角生态绿色一体化发展示范区数据中心集群	上海市青浦区、江苏省苏州市吴江区、浙江省嘉兴市嘉善县
	粤港澳大湾区枢纽	韶关数据中心集群	韶关南雄市、曲江新区、龙湖区、龙华市
	成渝枢纽	重庆数据中心集群	重庆两江新区、南岸区、巴南区、南岸区
	贵安枢纽	贵安数据中心集群	贵安新区
	中卫枢纽	中卫数据中心集群	中卫市沙坡头区、红墩峡镇、沙坡头区

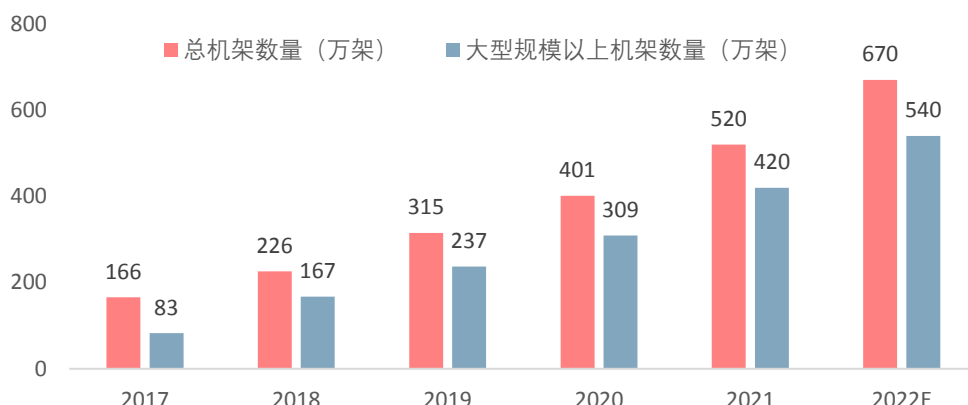
数据来源：发改委，中信建投

根据发改委表态，“东数西算”工程总体思路有三方面：一是推动全国数据中心适度集聚、集约发展；二是促进数据中心由东向西梯次布局、统筹发展；三是实现“东数西算”循序渐进、快速迭代。在当前起步阶段，8 个算力枢纽内规划设立 10 个国家数据中心集群，划定了物理边界，并明确了绿色节能、上架率等发展目标，如集群内数据中心的平均上架率至少要达到 65% 以上，要求张家口、韶关、长三角、芜湖、天府、重庆集群的 PUE 在 1.25 以下，和林格尔、贵安、中卫、庆阳集群的 PUE 在 1.2 以下。我们认为，10 个国家数据中心集群更多属于新建项目，各地方此前已发放的能耗指标及相关 IDC 公司在其它地区的投资规划可能多数会继续实施（现有 IDC 供应商本来在上述 10 个区域的投放安排就少），因此对于 IDC 建设产业链带来利好。

当前阶段的数据中心集群规划有两点值得关注：一是强调平均上架率至少要达到 65% 以上，供给增加要匹配需求增长，避免无序扩张；二是对 PUE 提出严格要求，大部分中小规模的 IDC 服务商在绿色数据中心设计、建设和运维方面的能力较为一般，难以满足 PUE 在 1.25 以下的能耗要求。对此，我们认为：一是在上架率要求的背景下，各数据中心集群的建设有望分期建设，边建设边交付边上架，最终投资金额仍待持续跟踪，假设上架速度较慢，可能存在短期停建可能；二是“东数西算”作为 IDC 供给侧改革的重要举措，预计未来其它区域的 IDC 供给将会进一步被压缩，未来东部核心区域的 IDC 资源将愈发具有稀缺性，因此目前拥有较多热点区域 IDC 资源的公司值得重视；三是降低 PUE 的主要方法是提高温控系统的工作效率，空调机组将从风冷型和水冷型向冷冻水型、双冷源型转化，未来液冷技术有望逐步普及，此外间接蒸发制冷的应用渗透率也有望提升。根据央视新闻报道，“东数西算”工程自启动至今，全国新增投资超过 4000 亿元，整个“十四五”期间，将累计带动各方面投资超过 3 万亿元。“东数西算”工程的 8 个国家算力枢纽节点建设已全部开工，工程从系统布局进入全面建设阶段。在已经开工的 8 个国家算力枢纽中，今年新开工的数据中心项目近 70 个，其中西部新增数据中心的建设规模超过 60 万机架，同比翻倍，至此国家算力网络体系架构初步形成。

在数字中国和人工智能推动云计算市场回暖的背景下，IDC 作为云基础设施产业链的关键环节，也有望进入需求释放阶段。在过去两年半，受多重因素影响下，云计算需求景气度下行，但 IDC 建设与供给未出现明显放缓，2021 年和 2022 年分别新增机柜数量 120 万架和 150 万架，因此短期内出现供需失衡情况（核心区域供需状况相对良好），部分地区上电率情况一般。所以 IDC 公司 2022 年业绩普遍承压。

图表143： 中国 IDC 标准机架规模



数据来源：中国信通院，中信建投

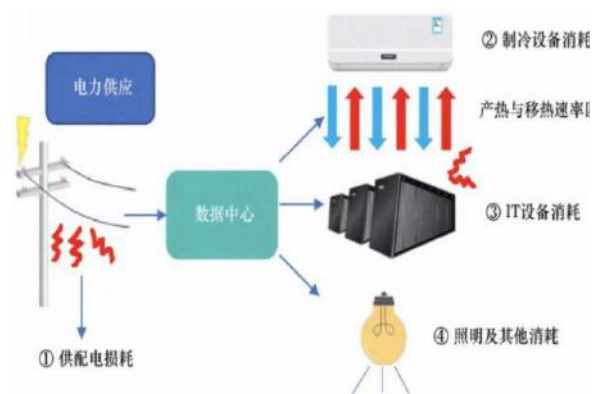
当前，我们认为国内 IDC 行业有望边际向好。随着宏观经济向好，平台经济发展恢复，AI 等拉动，IDC 需求有望逐步释放，叠加 2023 新增供给量有望较 2022 年减少（例如三大运营商 2022 年新增 IDC 机柜 15.6 万架，2023 年计划新增 11.4 万架）。**展望未来，电信运营商在云计算业务方面仍将实现快速增长，百度、字节跳动等互联网公司在 AIGC 领域有望实现突破性进展，都将对包括 IDC 在内的云基础设施产生较大新增需求，相关 IDC 厂商有望获益，建议关注润泽科技、宝信软件、奥飞数据、数据港、光环新网等。**

6.2 AI 大算力服务器需要高功率机柜，液冷或成必选项

人工智能大模型训练和推理运算所用的 GPU 服务器的功率密度将大幅提升，以英伟达 DGX A100 服务器为例，其单机最大功率约可以达到 6.5kW，大幅超过单台普通 CPU 服务器 500w 左右的功率水平。在此情况下，一方面需要新建超大功率的机柜，另一方面为降低 PUE，预计液冷温控渗透率将快速提升。

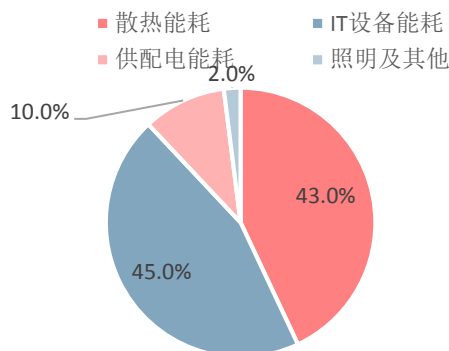
PUE 值是衡量 IDC 能效的重要指标。PUE 的计算方法为数据中心的总耗电量比上 IT 设备的耗电量，数值越接近 1，表明 IDC 的能效越高。根据赛迪顾问的统计数据，2019 年中国数据中心的能耗中约有 43% 是用于 IT 设备的散热，基本与 45% 的 IT 设备自身的能耗持平。因此，设备散热能耗成为降低 PUE 的关键影响因素。

图表144： IDC 机房的各类消耗



数据来源：中国热管理网，中信建投

图表145： 我国数据中心能耗分布

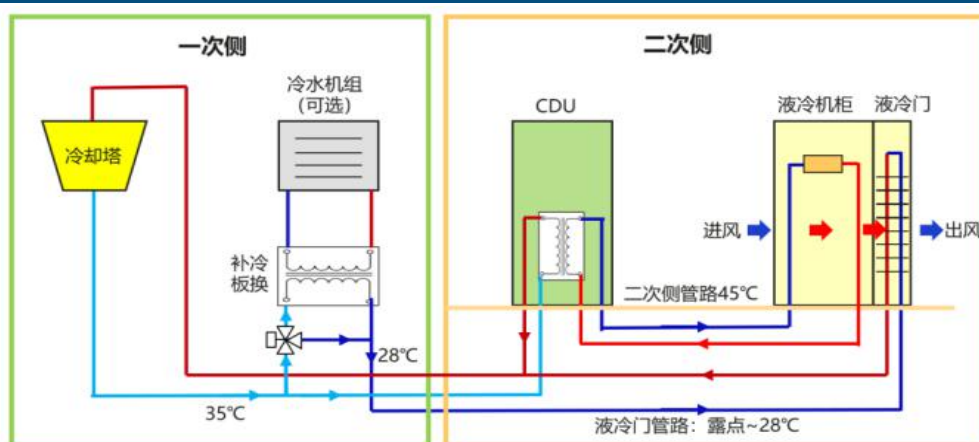


数据来源：赛迪顾问，中信建投

液冷数据中心适合提供高密算力，提升单柜部署密度，提高数据中心单位面积利用率。根据《冷板式液冷服务器可靠性白皮书》数据显示，液体相对空气能够传热更快(相差 20-25 倍)，能够带走更多热量(相差 2000-3000 倍)，给高密部署提供了较好方案。通常液冷数据中心单机柜可以支持 30kW 以上的散热能力，并能较好演进到 100kW 以上。自然风冷的数据中心单柜密度一般只支持 8kW-10kW，冷热风道隔离的微模块加水冷空调水平制冷在 15kW 以上性价比将大幅降低，相比较而言液冷的散热能力和经济性均有明显优势。

由于 AIGC 的发展，大功率 AI 服务器出货量有望快速增长，进而要求单机柜功率要明显提升，业界已经开始规模建设 20kW、30kW 功率的机柜。同时，数据中心降 PUE 也是刚需。在此背景下，由于风冷技术在高功率机柜制冷方面的短板比较明显，因此液冷有望成为 AI 大算力数据中心的主要制冷方案。

图表146： 液冷数据中心制冷架构示意图



数据来源：ODCC，中信建投

数据中心液冷方案主要分为冷板式和浸没式两种技术路径，此外还有喷淋式。根据《中国液冷数据中心发展白皮书》，液冷是指使用液体取代空气作为冷媒，为发热部件进行换热的技术。一般来说，行业将液冷分为直接冷却和间接冷却，其中直接冷却以浸没式液冷技术为主，间接冷却以冷板式液冷技术为主。

图表147： 各类制冷方式情况梳理

方式	PUE	支撑单机柜功率	示意图
传统风冷	1.4 以上	10kW 以下	

冷冻水、间接
蒸发冷却等 1.2-1.4 20kW 以下



冷板式液冷 1.2 以下 20-50kW



浸没式液冷 1.2 以下 50kW 以上



数据来源：英维克，依米康，阿里巴巴，中信建投

无论是冷板式液冷还是浸没式液冷，都需要数据中心温控和 ICT 设备厂商彼此配合，此前市场对于产业链的协作问题存在疑虑。目前在 AI 算力需求的推动下，服务器厂商已经开始大力布局液冷服务器产品，液冷的产业化进度有望加速。2022 年，浪潮信息将“AI in 液冷”纳入公司发展战略，全栈布局液冷，实现通用服务器、高密度服务器、整机柜服务器、AI 服务器四大系列全线产品均支持冷板式液冷，建成年产能 10 万台的亚洲最大液冷数据中心研发生产基地，实现了业界首次冷板式液冷整机柜的大批量交付。2022 年，中兴通讯发布了《中兴通讯液冷技术白皮书》，公司建设的全液冷数据中心项目获得了 2022 年 CDCC 数据中心科技成果奖，近期公司 G5 系列服务器在泰国进行海外市场首发，支持液冷散热技术，采用冷板式液冷散热。

图表148： 浪潮信息液冷服务器产品



数据来源：浪潮信息，中信建投

图表149： 中兴通讯全液冷数据中心项目获奖



数据来源：中兴通讯，中信建投

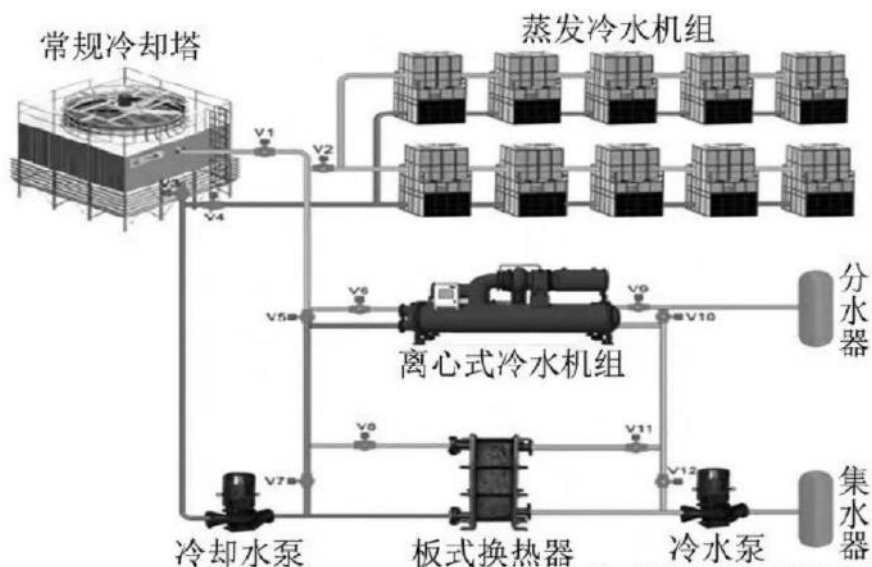
当前我国数据中心年用电量已占全社会用电的 2%-3% 左右，东部核心地区针对数据中心 PUE 已经提出严格要求。为确保实现“碳达峰碳中和”目标，需要在数据中心建设模式、技术、标准、可再生能源利用等方面进一步挖掘节能减排潜力。多地已针对数据中心的绿色低碳发展提出规划方案。北京发布《北京市数据中心统筹发展实施方案（2021-2023 年）》，提出将有序关闭腾退低利用率的数据中心，**新建云数据中心更强调“绿色”，PUE 不高于 1.3**，用于数据存储功能的机柜功率占比不超过 20%；广东省能源局 2021 年 4 月发布《关于明确全省数据中心能耗保障相关要求的通知》，明确提出“利用市场和行政手段，推动绿色低碳发展”，要求加大节能技术改造力度，“十四五”期间 PUE 降至 1.3 以下；上海在 2019 年发布的信息基础设施三年行动计划中提出，**新建数据中心 PUE 限制在 1.3 以下，存量数据中心 PUE 不高于 1.4**。

根据发改委表态，“东数西算”工程总体思路有三方面：一是推动全国数据中心适度集聚、集约发展；二是促进数据中心由东向西梯次布局、统筹发展；三是实现“东数西算”循序渐进、快速迭代。在当前起步阶段，8 个算力枢纽内规划设立 10 个数据中心集群，划定了物理边界，并明确了绿色节能、上架率等发展目标，如集群内数据中心的平均上架率至少要达到 65% 以上，**要求张家口、韶关、长三角、芜湖、天府、重庆集群的 PUE 在 1.25 以下，和林格尔、贵安、中卫、庆阳集群的 PUE 在 1.2 以下**。

我们认为，无论是在原来的东部核心区域，还是“东数西算”工程的枢纽节点内，政策端均对新建数据中心以及存量数据中心的 PUE 提出严格要求，其中枢纽节点内的要求更高，同时考虑到整体规划布局，未来新增机柜更多将在枢纽节点内，因此采用高效的机房温控方案来降低 PUE 是大势所趋。

根据《间接蒸发冷却在华北地区某数据中心的应用》数据显示，华北地区某数据中心一期 IT 备总负荷为 3150kW，采用集中式冷水机组+房间级空调末端的供冷架构，机房年均 PUE 为 1.4，采用新增间接蒸发冷却冷水机组的技术措施对一期机房实施节能改造后，机房年均 PUE 降至 1.28。但值得注意的是，“东数西算”枢纽节点对于 PUE 的要求为 1.25 或 1.2 以下，采用冷冻水或间接蒸发冷却（风冷）方案在某些高温、高湿地区可能无法严格满足上述要求，因此我们预计液冷方案渗透率有望加速提升。

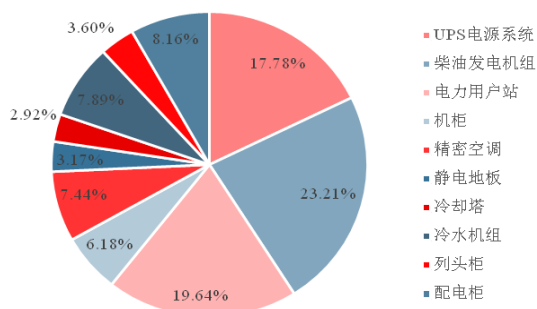
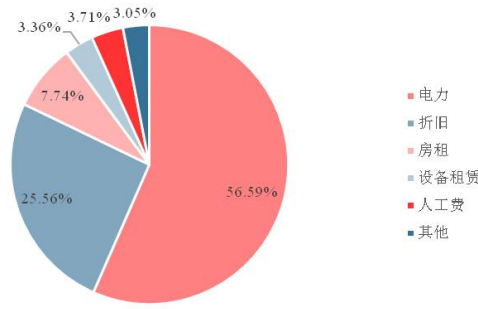
图表150： 华北地区某数据中心节能改造示意图



数据来源：《间接蒸发冷却在华北地区某数据中心的应用》，中信建投

制冷系统约占数据中心建设投资的 20% 左右，短期来看液冷方案的价值量更高。根据数据港的招股说明书

数据显示，制冷系统占到 Capex 的比例约为 20%，按照 IT 负载进行估算，单 kW 对应冷冻水方案（风冷）制冷系统的价值量约为 7000 元，我们预计如果采用液冷方案，投资将提升至 1.5 万元以上/kW。

图表151： 数据港 Capex 支出构成

图表152： 数据港 OPEX 支出构成


数据来源：数据港招股书，中信建投

数据来源：数据港招股书，中信建投

假设国内每年新增机柜数为 100 万架（按照单机柜 2.5kW 计算），新增 IT 负载量为 250 万 kW，若全部采用风冷方案，单 kW 价值量为 7000 元，则对应温控市场规模为 175 亿元；若液冷方案渗透率达到 70%，单 kW 液冷价值量为 1.5 万元，则对应温控市场规模为 315 亿元（+80%），其中液冷温控市场规模为 262.5 亿元。

考虑到 AIGC 的发展，“东数西算”工程 PUE 的要求，以及超算/智算中心的建设需求，我们认为，一是未来 2-3 年将是国内大功率 IDC 新增建设的高峰期，液冷方案的渗透率可能更高，二是存量 IDC 机房为满足现在更严格的 PUE 要求，可能需要进行改造，其中制冷系统将是改造和投资的重点，三是产业链目前的液冷产能规划可能在短期内将呈现供不应求状态（液冷设备系统现有产能较少，叠加储能等新能源需求快速增长），因此在液冷方案加速渗透过程中，数据中心温控厂商、液冷板制造厂商等有望受益，建议关注：网宿科技（全资子公司绿色云图深耕液冷技术多年，传统主业 CDN 发展势头向好，且向边缘计算方向发展良好）、英维克（数据中心与储能制冷方案领先供应商，液冷布局深厚）、科创新源（液冷板产品有望应用液冷数据中心及服务器）、飞荣达（液冷板产品有望应用液冷数据中心及服务器）、依米康（数据中心制冷方案主要供应商）等。

6.3 人工智能算力需求有望推动海底数据中心规模化发展

我们认为，海底数据中心可能将迎来产业化的关键节点。一是中国通信工业协会已于 2022 年 12 月 14 日批准发布标准 T/CA 303—2022《水下数据中心设计规范》。二是中国及全球近两年海上风力发电取得大发展，海底数据中心可就近消纳海上风电。三是东部沿海城市算力及 IDC 需求旺盛，海底数据中心可就近满足需求。四是 AIGC 需要单机柜功耗可能达几十 kW，海底数据中心单机柜功率可达 35kW 左右，利用海水冷却，无压缩机运行，单舱 PUE 可以低于 1.10，且无需冷却塔，可节约大量的水资源。五是全球在海底数据中心布局领先的是微软，2015 年开始启动测试，两次测试都成功，2022 年美国 subsea cloud us 计划推出商用海底数据中心。

6.3.1 国内海底数据中心相关设计规范已经发布

国内第一个水下数据中心标准已发布并已开始实施。中国通信工业协会已于 2022 年 12 月 14 日批准发布标准 T/CA 303—2022《水下数据中心设计规范》，该标准于 2023 年 1 月 1 日起开始实施。该标准遵循开放、公平、透明、协商一致和促进贸易和交流的原则，按照全国团体标准信息平台公布的标准制定程序文件制定，由深圳

海兰云数据中心科技有限公司、中国通信工业协会数据中心委员会、中国长江三峡集团有限公司、海洋石油工程股份有限公司、维谛技术有限公司、清华大学等单位共同起草。

该标准适用于指导和规范新建、改建和扩建部署于海洋的水下数据中心设计工作。《水下数据中心设计规范》基于海底数据中心水下密封、无氧无尘、空间受限、无人值守等特点，规定了水下数据中心的分级与性能要求、选址与系统组成、水下舱体系统设计要求、电气系统设计要求、空调系统设计要求、监控系统设计要求、网络与布线系统设计要求、动力与通讯缆线系统设计要求、消防与安全系统设计要求。部署于湖泊、江水等水下数据中心亦可参照执行。

该标准的发布有利于推进我国水下数据中心的发展。《水下数据中心设计规范》标准的发布，有利于推进我国水下数据中心的发展，保障水下数据中心工程顺利实施，从而可以科学有序地衔接设计、建设、运维工作，确保水下数据中心安全、稳定、可靠运行，做到技术先进、经济合理、节能环保。更好地为用户、营运商和业务主管部门提供水下数据中心设计规范，从而为数据中心“碳中和”的实现贡献力量，为新业态、产业领域的多融合探索、构建树立标杆。

图表153： 水下数据中心示例图



数据来源：Dgtl Infra，中信建投

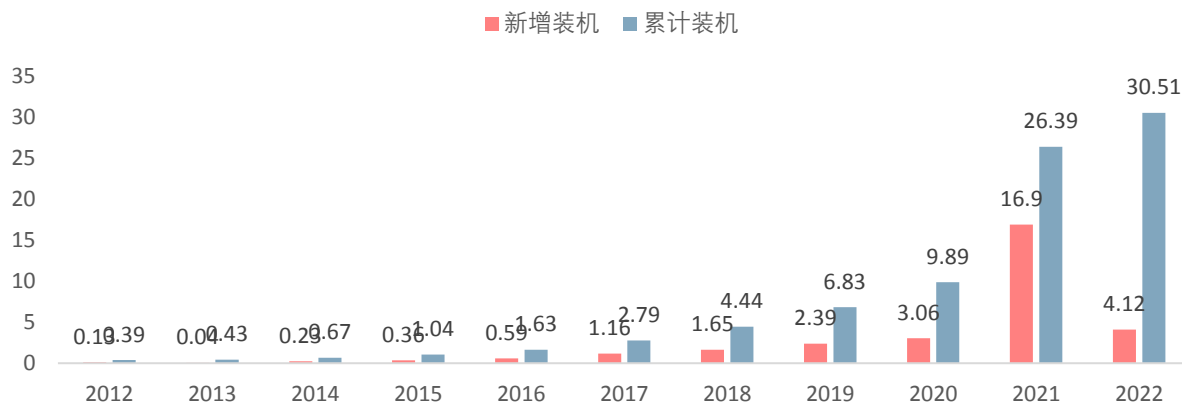
水下数据中心的建设符合我国“双碳”发展目标和新型数据中心发展战略，该标准顺应了数据中心向“零碳”方向发展的趋势。“双碳”目标下各领域节能减碳力度正不断增强，零碳被广泛提及，对于高耗能的数据中心而言，零碳更是早已成为关注焦点。但各产业的算力需求不断提升，数据中心规模不断扩大，碳排总量提高，想要实现零碳挑战巨大。这需要对现有减碳方式进行升级，编制并落地更具指导性的标准、规范，进一步优化相应的技术及解决方案，《水下数据中心设计规范》便应运而生。

6.3.2 海上风电已经实现规模化发展，有望与海底数据中心结合产生新商业模式

经历了 2020-2021 年海风抢装潮，国内海上风电产业链加速成熟。我国海上风电探索起源于 2007 年。是年 11 月 8 日，首座安装有 1 台金风科技 1.5 MW 风电机组的海上风电项目在渤海绥中油田建成发电，经历了十余年的发展，2020 年末，我国海上风电装机量达到了 9.89GW。2019 年 5 月 24 日，国家发改委发布《关于完善风电上网电价政策的通知》，提出将海上风电标杆上网电价改为指导价，新核准海上风电项目全部通过竞争方式确定上网电价；对 2018 年底前已核准的海上风电项目，如在 2021 年底前全部机组完成并网的，执行核准时的上

网电价（约 0.85 元/千瓦时，补贴力度超 0.4 元/千瓦时），极具诱惑力的补贴价格，带来了海上风电的抢装潮，仅 2021 年中国海上风电新增装机量超过 16.9GW，抢装也加速了我国海风产业链的成熟，2010 年我国海上风电的单 GW 造价水平大约在 240 亿左右，目前已经降至 120-130 亿元。截至 2022 年末，中国海上风电装机量达 30.51GW。

图表154： 中国海上风电装机量（GW）



数据来源：国家能源局，中信建投

风机大型化带来发电效率提升叠加产业链降本推进，部分地区或已经实现平价。通过十多年的海上风电场设计建造的经验，以及装备制造水平的提升，根据华东勘测设计院的测算，在福建、广东、海南等风资源较好且标杆煤电价格较高的省份已经基本具备平价上网的条件。

图表155： 海上风电经济性指标测算

省（市）	年平均风速 (m/s)	等效满负荷小 时数 (h)	目前可研概算 水平 (元/kW)	标杆煤电价格 (元/kWh)	最小电价差 (元/kWh)	最小造价差 (元/kW)
辽宁	6.5-8	2750-3200	13000	0.3749	0.083	2600
天津	6.5-8	2750-3300	13000	0.3655	0.086	2740
河北	6.5-8	2750-3300	13500	0.372	0.095	3030
山东	6.5-8	2750-3300	13500	0.3949	0.025	1300
江苏	7.0-8.0	3080-3300	13000	0.391	0.015	1000
上海	7.0-8.0	3080-3450	14000	0.4155	0.006	1190
浙江	7.0-8.0	3080-3450	14500	0.4153	0.039	1670
福建	7.5-10	3300-4100	15000	0.3932	~	~
广东	7.0-9.0	2750-3700	15500	0.453	~	~
广西	6.5-8.0	2420-3200	13000	0.4207	0.045	1400
海南	6.5-8.5	2420-3550	13000	0.4298	~	~

资料来源：华东勘测设计研究院，中信建投。注：最小电价差和最小造价差以资本金 IRR6% 反算。

发展海上风电的省份均为东部发达地区，同时也是对于算力需求较高的省份。IDC 本身对于能耗需求较高，使用海上风电与海底数据中心联合作业方式，既可实现对于海上风电能源的就近消纳，数据中心自身也可以使用绿色能源来实现“零碳”目标，二者结合有望诞生新的商业模式。

6.3.3 海底数据中心节能优势突出，可较好满足沿海地区的旺盛算力需求

海底数据中心 UDC 是水下数据中心的一种。海底数据中心是将服务器等信息基础设施安装在海底密封的压力容器中，利用流动海水进行散热，并利用海底复合缆供电且将数据回传至互联网的新型数据中心。海底数据中心具有节能、节地、低时延、安全可靠等显著的绿色低碳特征和多方面的优点，符合绿色低碳发展趋势。

海底数据中心一般建设在海岸线 10-20 公里之处，可满足沿海地区较高的算力、数据存储及低延迟的要求。水下数据中心为低延迟连接提供了一种解决方案，即减少数据在源和目的地之间传输所需的时间。西部内陆地区的数据中心可以进行一些冷数据的存储和延迟要求较低的计算，但对于延迟要求较高的还是需要在东部沿海地区寻找数据中心资源。东部沿海城市算力需求旺盛，海底数据中心可以利用较近的距离为基数巨大的沿海人口提供低延迟连接，因为世界上超过 50%的人口居住在距离海岸 120 英里（200 公里）的范围内。

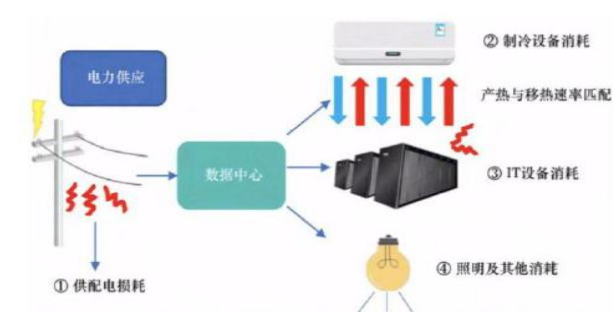
图表156： 建设在海边的水下数据中心



数据来源：Dgtl Infra，中信建投

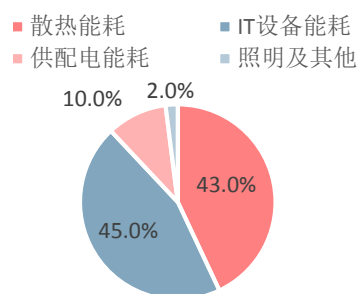
通过将水下数据中心放置在世界大部分人口附近，可以为服务不足的社区提供更快、更流畅的互联网浏览、视频流、游戏和云服务。因此，水下数据中心可能成为包括亚马逊网络服务（AWS）、微软 Azure 和谷歌云在内的云服务提供商的重要边缘计算工具。

图表157： IDC 机房的各类消耗



数据来源：中国热管理网，中信建投

图表158： 我国数据中心能耗分布



数据来源：中国热管理网，中信建投

PUE 值是衡量 IDC 的重要指标。PUE 的计算方法为数据中心的总耗电量比上 IT 设备的耗电量，数值越接近 1，表明 IDC 的能效越高。工信部明确规定 2025 年底，新建数据中心的 PUE 值必须在 1.3 以下。

在单机功率、PUE 等方面，海底数据中心优于陆上 IDC。AIGC 的计算需要单机柜功耗可能达到几十 kW，目前广泛使用的英伟达 DGX A100 服务器单机功率就有 6.5kW；海底数据中心单机柜功率可达 35kW 左右，利用海水冷却，无压缩机运行，单舱 PUE 可低于 1.10，且无需冷却塔，可节约大量的水资源。此外，在冷却效率、延迟、建造时间与成本、可靠性及可持续性方面，以海底数据中心为代表的水下数据中心也表现出了一定优势。

图表159：水下数据中心与传统陆上 IDC 部分指标对比

指标	传统陆上 IDC	水下数据中心	UDC 的特点
冷却效率	微软新建的陆基数据中心的 PUE 约为 1.125。现有海南陆地 IDC 的 PUE 值(1.8-2.0)	微软水下数据中心在海平面下 36 米处，温度比陆基数据中心低约 10 摄氏度。PUE 为 1.07。海兰信实测结果单舱 PUE 值为 1.076，公司预计海南示范项目实际 PUE 可以成功控制在 1.10 左右	海洋提供持续的寒冷环境，降低冷却成本
延迟	路基 IDC 对地理环境要求高，在中国多部署在离大城市较远的中西部地区，有时延	世界上超过 50%的人口居住在距离海岸 120 英里（200 公里）的范围内，水下数据中心分布在沿海发达城市 10-20 公里海域内，距离用户更近，可能成为云服务提供商的重要边缘计算工具。	海洋数据中心可以为沿海人口提供低延迟连接
建造时间和成本	在陆地上，数据中心的“建设”需要许可和适应各种物理环境。一般而言，建造数据中心的成本在每平方英尺 600 至 1,100 美元之间，或每兆瓦 IT 负载 7 万至 12 万美元。（具体见下文表格）传统数据中心的完整建设周期在 400 天左右，模块化数据中心可以缩短建设周期到 2-3 个月。	水下数据中心更多地涉及“制造”过程，旨在大规模生产模块，以便在非常相似的海洋条件下部署。相比路基 IDC，单千瓦 TCO（建设成本+运营成本）节约 15-20%左右，土地占用仅 1/5。	水下数据中心是作为预制和标准化模块构建的，这样可以快速构建和交付时间
可靠性	陆基数据中心平均寿命为 10-15 年（来自华为的数据）	海洋数据中心可以在现场无人且无需维护的情况下运行长达 5 年。生命周期在 20 年，每 5 年进行一次重新加载服务器和部署。	水下数据中心具有高度的可靠性和更可预测的数据中心性能，因为这些预制模块是在受控的工厂环境中以精确的规格构建的
可持续性	陆基数据中心 2021 年行业平均 WUE 是每千瓦时 1.8 升水	水下数据中心的可再生能源包括海上风能、太阳能、潮汐能和波浪能。通过不连接到电网，这些海洋数据中心可以减轻当地电网的压力。此外，水下数据中心用水效率 WUE 为 0，不消耗水。	水下数据中心可以使用可再生能源，满足可持续性要求

资料来源：Dgtl Infra，中信建投

此外，海底数据中心在资源节约与互补、安全性等方面也存在优势。海底数据中心的岸站占地极少；没有冷却塔，节约大量的水资源（200 立方米/机柜·年，典型规模年省水 60 万立方米）。同时，海底数据中心亦可利用海上风能、太阳能、波浪能和潮汐能等可再生能源实现多能互补。安全性方面，由于海底数据中心可满足

恒温、恒湿、恒压、无氧、无尘的条件，其可充分保障数据的物理安全。并且水下数据中心可预制、作为标准化模块构建，因而可以快速构建和交付，实现工业化部署与模块化生产。**海底数据中心的建设与使用在综合利用海洋资源的同时，陆海统筹、生态用海、集约用海，也响应了高效利用海洋的国家战略。**

6.3.4 全球海底数据中心建设案例——微软 Natick 项目

全球首个海底数据中心于 2015 年由美国微软公司研制，微软在海洋中建立水下数据中心和放置服务器的研究实验——Natick 项目目前已完成了为期 4 个月的水下概念验证测试与为期两年的水下数据中心测试。该项目第一阶段的目的是有效地测试水下数据中心的冷却系统。第二阶段的目的则在于确定全尺寸水下数据中心模块的制造可行性以及在 90 天内部署它们的经济可行性。此外，在两年的时间里，微软还能够测试和监控水下数据中心服务器的性能和可靠性。

图表160： 微软 Natick 项目测试指标

Project Natick	Phase 1	Phase 2	Phase 3
Launched	August 2015	June 2018	Future
Duration	105 days	2 years	5 years
Location	California	Scotland	TBD
Racks	1	12	144
Servers	24	864	10,368
Length	10ft(3m)	40ft(12.2m)	<300ft(<91.5m)
Depth	Shallow	117ft(36 m)	>131ft(>40m)

资料来源：微软，中信建投

Natick 项目的第一阶段是一个概念验证原型水下数据中心，于 2015 年 8 月启动。Natick 项目的第一阶段被放置在平静的浅水中的海底，距离美国加利福尼亚州圣路易斯奥比斯波附近的阿维拉海滩太平洋海岸约 0.6 英里（1 公里）。这个海洋数据中心的规格为 10 英尺（3 米）x 7 英尺（2.1 米）、38,000 磅重，装有 1 个标准 42U 机架，包含 24 台服务器，服务器占据机架空间的 1/3，其他 2/3 服务器装满“负载托盘”以产生热量，目的为有效地测试水下数据中心的冷却系统。

图表161： 微软 Natick 项目第二阶段——水下数据中心

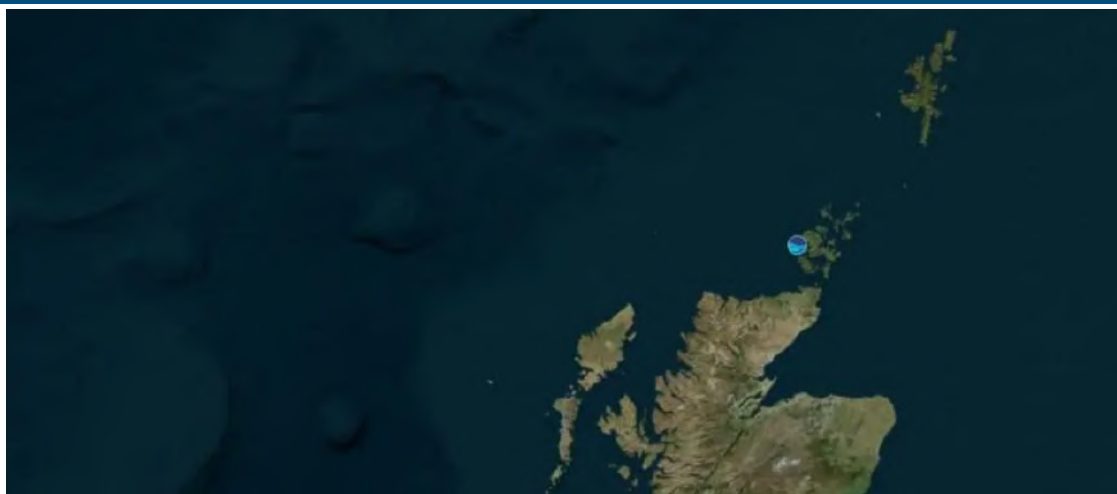


数据来源：微软，中信建投

Natick 项目的第二阶段是一个水下数据中心，部署时间为 2018 年 6 月到 2020 年 7 月，时长达两年。在微

软最初的概念验证测试之后，海洋数据中心的规模不断扩大，项目的第二阶段是一个集装箱大小的数据中心，承载 12 个机架，包含 864 台服务器。第二阶段被放置在北部群岛的海底——117 英尺（36 米）深的岩石板海底，具体位于英国苏格兰奥克尼群岛的欧洲海洋能源中心（EMEC）。该设施包括一个装有 12 个机架的水下数据中心，其中包含 864 台具有 FPGA 加速功能的标准服务器。864 台服务器中的每一台都有 32TB 的磁盘，相当于 27.6PB 的总磁盘。在电力消耗方面，微软 Natick 项目的第二阶段需要 240 千瓦（kW），这意味着在满负荷运行时，功率不到四分之一兆瓦。这种电力来自 100% 当地生产的可再生电力，包括陆上风能和太阳能，以及海上潮汐能和波浪能。

图表162： 微软 Natick 项目第二阶段位置图



数据来源：微软，中信建投

微软 Natick 项目未来的第三阶段被描述为“试点”。具体来说，微软将为 Natick 项目的第 3 阶段建立一个“更大规模”的水下数据中心，该数据中心“可能是多艘船”，并且“可能是与第二阶段不同的部署技术”。微软 Natick 项目的第 3 阶段将被放置在大于 117 英尺（36 米）的深度。

微软通过 Natick 项目探索了海底数据中心发展的潜力。Natick 项目第二阶段测试结果显示，海底数据中心的 PUE 为 1.07，故障率是地面数据中心故障率的八分之一。同时，微软通过 Natick 项目发现，水下数据中心可实现快速部署，并可密封在类似潜艇的管道内，在海床上运行多年，而无需人工进行任何现场维护。初步分析表明，服务器在水下具有卓越性能的主要原因是避免了湿气和氧气的腐蚀。

但需要注意的是，目前海底数据中心也存在发展瓶颈。一是海底数据中心需要高额的建设成本，包括购买数据舱、服务器、布线、配电系统、通信系统等。二是海底数据中心的技术难度大，需要具备海洋环境下的建设、抗潮汛、抗海浪、抗噪声等技术。三是海底数据中心的运维工作复杂，由于海底环境条件复杂多变，需要特殊的技术和设备才能完成运维工作。

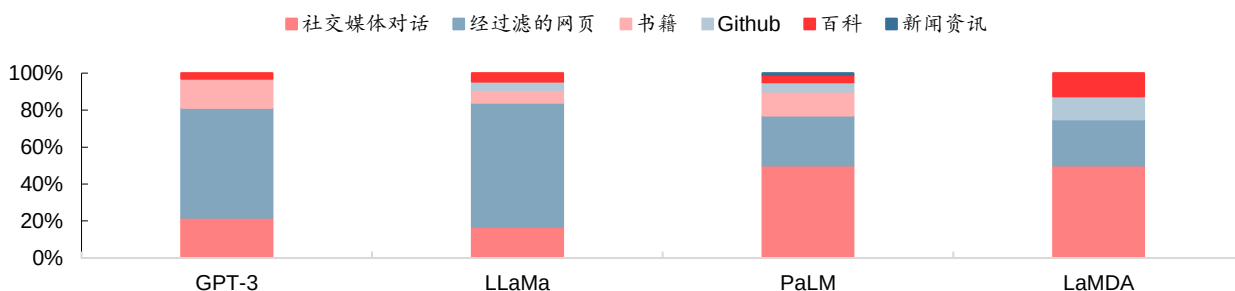
七、海外大模型进展

7.1 谷歌

谷歌在训练集方面不具备明显优势。在训练数据集方面，现有的大模型主要采用书籍、文章、网页等

文本数据，这些数据能够帮助大模型积累语法、文本知识、文本理解、上下文连贯逻辑等能力，而在前文“综述”部分我们提到代码对语言模型的逻辑推理能力具备帮助，因此训练数据集的多样性较为重要，确保大模型积累多样化的能力以便后续激活，这里的问题主要是，例如逻辑推理的培养需要一定比例的高质量代码数据，1) 如何定义高质量的数据，怎么对原始数据进行清理、去重¹、标注等？2) 多大比例的数据能够积累能力？就我们的知识范围，目前学术界/业界尚未有较为公开且权威的研究能够回答上述问题，但总体而言，数据质量上论文/书籍/百科≥代码/文章≥对话≥网页。从这一角度看，Google 在数据源方面不存在明显的优势。

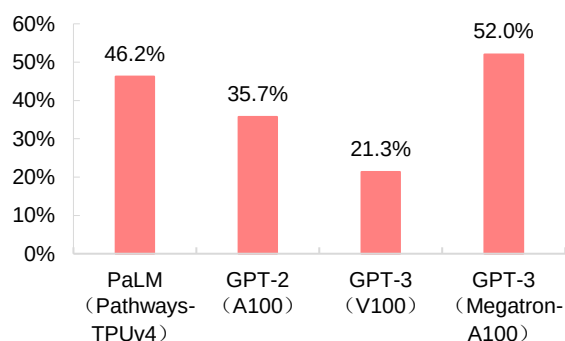
图表163： 不同大语言模型的预训练数据集结构（%）



数据来源: OpenAI, Google, Meta, 中信建投

谷歌在 AI 架构、芯片方面处于行业领先地位。Google 在《Pathways: Asynchronous Distributed Dataflow for ML》提出了 Pathways 作为新一代 AI 架构，其特点是多任务，多通道，稀疏激活。在《PaLM: Scaling Language Modeling with Pathways》中，Google 提到 Pathway 下 MFU (Model Flops Utilization) 达到 46.2%，高于 GPT-2/3 在 A100/V100 集群上的利用率 35.7%/21.3%，但低于 GPT-3 基于英伟达 Megatron-A100 集群实现的利用率 52%。TPU 方面，TPU 在 MLPerf 部分场景的性能测试中表现优于 A100，其中 TPU v4 在 4096 块芯片，应用 BERT 场景下性能是 A100 的 1.15 倍左右；ResNet 场景下 TPU v4 则是 A100 性能的 1.67 倍。

图表164： Google 在分布式集群计算资源利用率方面处于相对领先地位



数据来源: Nvidia, 《Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM》, 《PaLM: Scaling Language Modeling with Pathways》, 中信建投

图表165： TPUv4 在多个下游场景中表现优于 A100

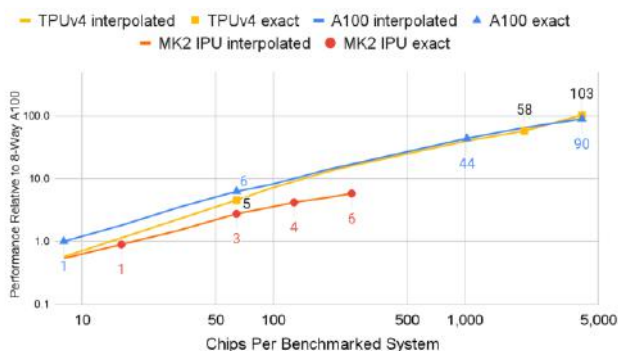


数据来源: 《TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings》, 中信建投

¹ 2022 年 5 月，Anthropic 团队在《Scaling Laws and Interpretability of Learning from Repeated Data》指出重复数据对 LLM 的损失产生较大影响。

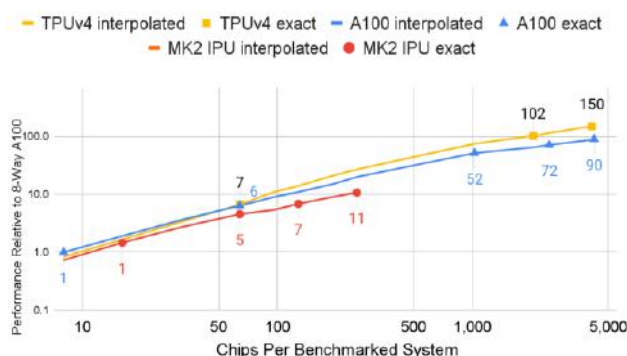
注：Megatron-A100 的利用率 52% 是在 529.6B-1008B 的参数规模实现的。

图表166： TPU v4 在 BERT 上表现优于 A100



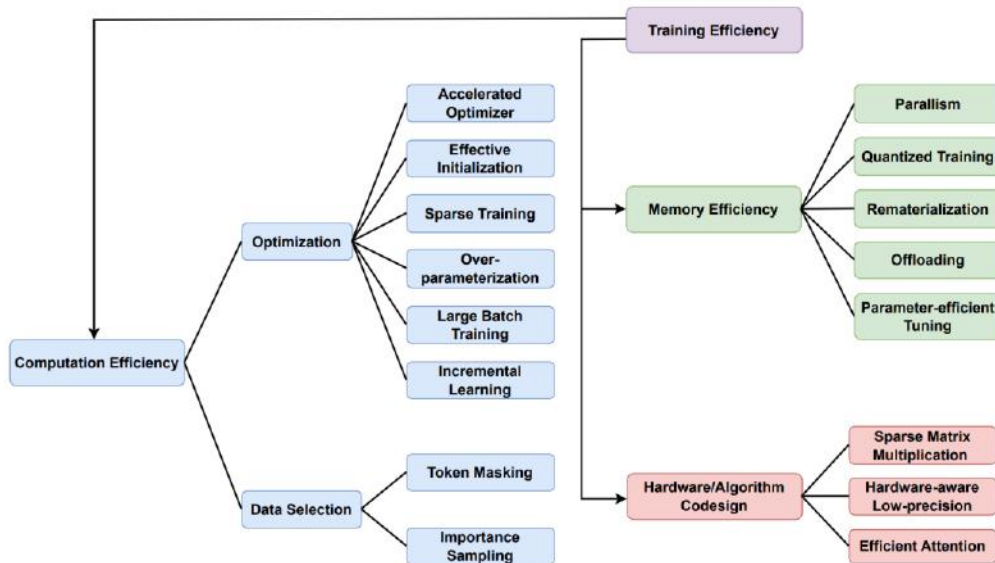
数据来源：《TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings》，中信建投

图表167： TPU v4 在 ResNet 上表现优于 A100



数据来源：《TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings》，中信建投

图表168： 目前学界/业界提升模型计算效率的策略分类

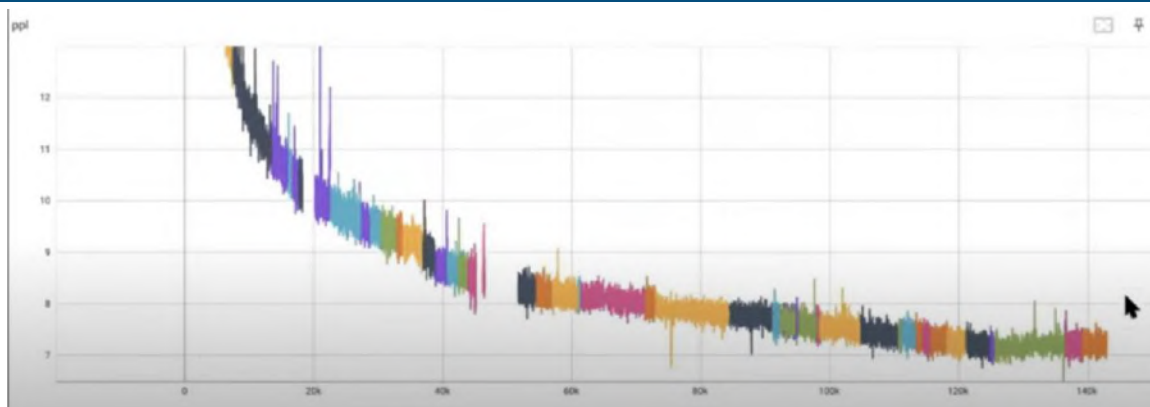


数据来源：《A Survey on Efficient Training of Transformers》，中信建投

大模型的训练稳定性是过去研究涉及较少的。由于小模型训练时长较短，涉及的软硬件协同面较窄，扩展至大模型下集群出现异常或错误的概率大幅提升，相应带来模型训练的不稳定性（Training instability），以及资源的额外耗费（一般需要回到 checkpoint 重新训练）。在训练策略上，Google 团队在 PaLM 论文中提到模型训练过程中多次出现损失函数的突变（we observed spikes in the loss roughly 20 times during training²），而 Susan Zhang 在 Stanford 分享 OPT 模型训练过程中展示了模型训练中也出现了多次波动。

² 《PaLM: Scaling Language Modeling with Pathways》。

图表169： OPT-175B survived 143K steps



数据来源：Stanford，中信建投

谷歌在模型训练方面具有较好积累。Diederik P. Kingma 和 Jimmy Lei Ba 2014 年发表《Adam: A method for stochastic optimization》，Adam 是一种可以替代传统随机梯度下降过程的一阶优化算法，它能基于训练数据迭代地更新神经网络权重。Diederik P. Kingma 于 2015 年与其他合伙人共同创立 OpenAI，并于 2018 年加入 Google Brain。而 Google 团队于 2023 年 2 月提出了 Lion 优化器³，此前流行的 AdamW 等自适应优化器需要同时保存一阶和二阶矩相比，Lion 只需要动量，这意味着内存占用降低，且在训练大型模型和大 Batch size 时效果显著。

图表170： Fine-tuning performance of the T5 Base, Large, and 11B on the GLUE dev set

Model	Optimizer	CoLA	SST-2	MRPC	STS-B	QQP	MNLI -m	MNLI -mm	QNLI	RTE	Avg
Base	AdamW	60.87	95.18	92.39 / 89.22	90.70 / 90.51	89.23 / 92.00	86.77	86.91	93.70	81.59	87.42
	Lion	61.07	95.18	92.52 / 89.46	90.61 / 90.40	89.52 / 92.20	87.27	87.25	93.85	85.56	87.91
Large	AdamW	63.89	96.10	93.50 / 90.93	91.69 / 91.56	90.08 / 92.57	89.69	89.92	94.45	89.17	89.46
	Lion	65.12	96.22	94.06 / 91.67	91.79 / 91.60	90.23 / 92.67	89.85	89.94	94.89	90.25	89.86
11B	AdamW	69.50	97.02	93.75 / 91.18	92.57 / 92.61	90.45 / 92.85	92.17	91.99	96.41	92.42	91.08
	Lion	71.31	97.13	94.58 / 92.65	93.04 / 93.04	90.57 / 92.95	91.88	91.65	96.56	93.86	91.60

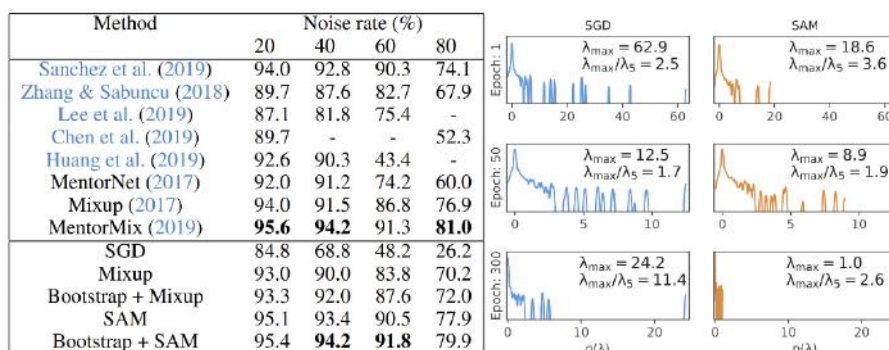
数据来源：《Symbolic Discovery of Optimization Algorithms》，中信建投

此外，Google 研究团队在 2021 年《Sharpness-aware minimization for efficiently improving generalization》提出 Sharpness-aware minimization (SAM)方法,除了提升模型的泛化表现，SAM 可以提高模型对标签噪声的稳健性（robustness）。后续 NUS⁴和字节跳动⁵的研究团队进一步优化了 SAM 方法。

³ 《Symbolic Discovery of Optimization Algorithms》。

⁴ 《Efficient sharpness-aware minimization for improved training of neural networks》。

⁵ 《Sharpness-aware training for free》。

图表171： SAM 提升了模型对标签噪声的稳健性，并优化了模型训练效率


数据来源：《Sharpness-aware minimization for efficiently improving generalization》，中信建投

模型初始化策略方面，MIT 和 Google⁶2019 年提出 Fixup 策略，避免梯度爆炸或消失，并可以应用于超过 1 万层的神经网络。后续 UCSD⁷和 Google⁸进一步在此基础上提出 Rezero 和 SkipInit，具体到每一层执行操作，实现进一步优化。

Google 团队在模型调试和 Prompt engineering 方面积累领先行业。在前文综述部分，我们提到谷歌团队开创了 CoT 研究，其论文《Chain-of-Thought Prompting Elicits Reasoning in Large Language Models》引入 CoT Prompt，并通过对比实验探测出模型能力涌现的界限大约是 62B 和 175B。Google 团队在 2022 年 12 月比较了不同参数规模下直接 prompt 以及 CoT 下的表现，得出以下结论：对于所有小于 62B 的模型，直接用提示词都好于思维链。结合 GPT-3 模型规模，至少需要大于 175B，思维链的效果才能大于 Fine-tuned 小模型的效果。东京大学和 Google 团队《Large Language Models are Zero-Shot Reasoners》更进一步提出 Zero-Shot Prompting，即加入“Let’s think step by step”可以显著的提升模型性能。

对齐调优方面，OpenAI 及 Anthropic 相对领先。OpenAI 团队⁹提出通过递归法能够实现对长难文本的归纳总结，并指出这类方法可以进一步泛化至其他类型的任务上，实现与人类的对齐。此外，论文指出 RL 比 SL 更有效地帮助模型对比。具体细节方面，John Schulman 在《Reinforcement Learning from Human Feedback: Progress and Challenges》¹⁰提到，SFT 与其等价的 BC 存在固有缺陷，即训练越充分越容易出现欺骗（即模型并不明确自己知识的边界），RLHF 则是让模型明确有些问题自己不知道。原理上是因为 SL 训练时只有正反馈，而且对偏离样本的惩罚较重，RL 多样性更好，因而在面对不知道的问题时，SL 训练充分的模型倾向于回答（胡编乱造），而非反馈不知道¹¹。需要指出的是，OpenAI 提出 alignment tax，即模型牺牲部分性能实现与人的对齐。

⁶ 《Fixup initialization: Residual learning without normalization》。

⁷ 《ReZero is All You Need: Fast Convergence at Large Depth》。

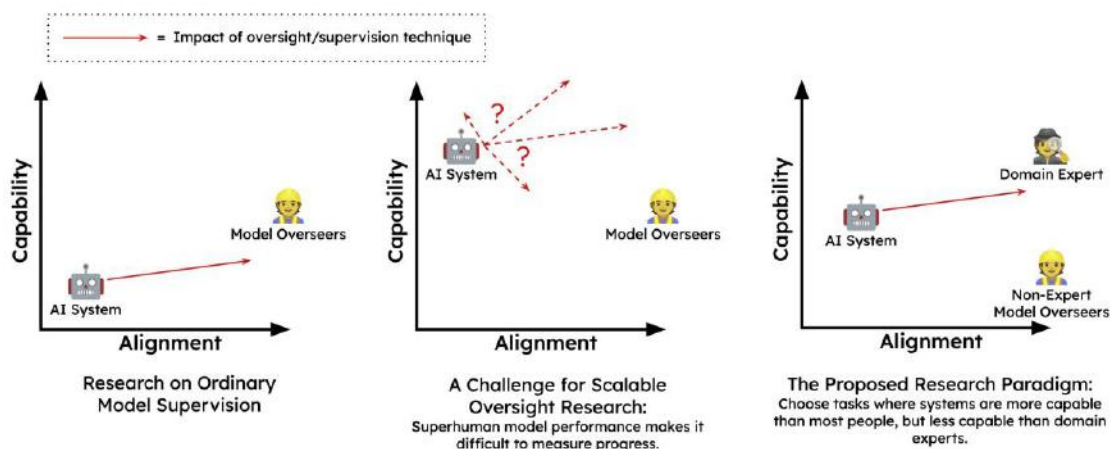
⁸ 《Batch Normalization Biases Residual Blocks Towards the Identity Function in Deep Networks》。

⁹ 在 2021 年 9 月发布《Recursively Summarizing Books with Human Feedback》。

¹⁰ https://www.youtube.com/watch?v=hhiLw5Q_UFg

¹¹ <https://gist.github.com/yoavg/6bfff0fec65950898eba1bb321cfbd81>

图表172： 当模型性能超越一般人时，Alignment 成为挑战



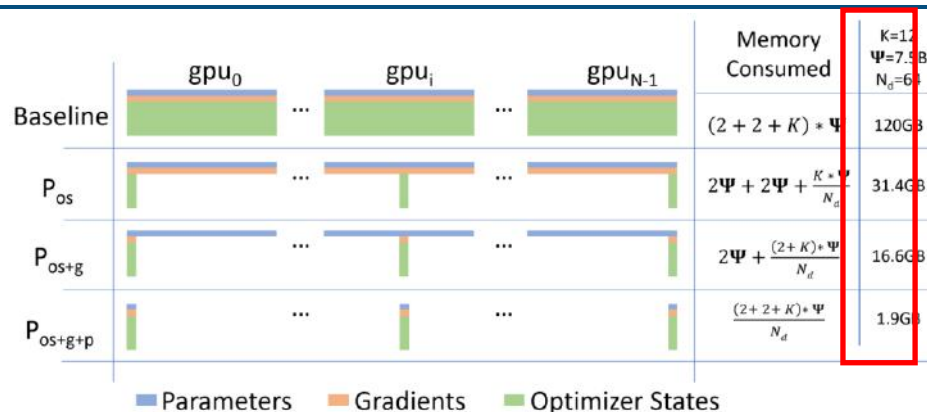
数据来源：《Measuring Progress on Scalable Oversight for Large Language Models》，中信建投

总结来看，谷歌在大模型领域的布局是全方位的，涵盖上游芯片、分布式计算集群、深度学习框架，以及模型训练、调试优化策略，并且在多数环节保持领先地位，OpenAI 的成功则是建立在与微软、英伟达等公司相互合作的基础上，并且是 OpenAI 与微软是通过股权投资绑定利益关系，这意味着其他竞争者模仿的难度较大，而就互联网平台而言，Google 在 AI 领域的积累深厚，整体并不落后于 OpenAI 的情况。

7.2 微软

微软整体上在 LLM 领域的工程化方向有较为深厚的积累。微软研究团队于 2019 年 10 月提出 ZeRO，通过分片（partition）显著优化显存和通信花费。值得一提的是，后续 DeepSpeed 超大规模训练工具正是基于 ZeRO 为代表的一系列工作。在《ZeRO: Memory Optimizations Toward Training Trillion Parameter Models》中，研究团队探讨了内存显存的结构，并将模型训练阶段每张卡中显存内容分为两类：1) 模型状态（model states）；2) 剩余状态（residual states），而模型状态占用了主要显存，因为深度学习中应用比较广泛的 Adam 优化器涵盖了参数梯度、梯度的一阶动量和二阶动量，并且在混合精度训练下需要存储 FP16 的模型参数、梯度，FP32 的 Adam 状态（模型参数、梯度备份、梯度的一阶动量和二阶动量）。定量看，如果单模型参数量为 Ψ ，则实际需要 16Ψ 的存储空间（其中 75% 来自 Adam 优化器）。因此研究团队提出通过分片（partitions）优化 Adam 带来的模型状态显存占用，通过动态通信策略提升通信效率， P_{OS+g} 下模型通信量和标准的数据并行一致。

图表173: ZeRO 优化下 P_{OS} 实现显存占用优化至基准方法的 26.2%

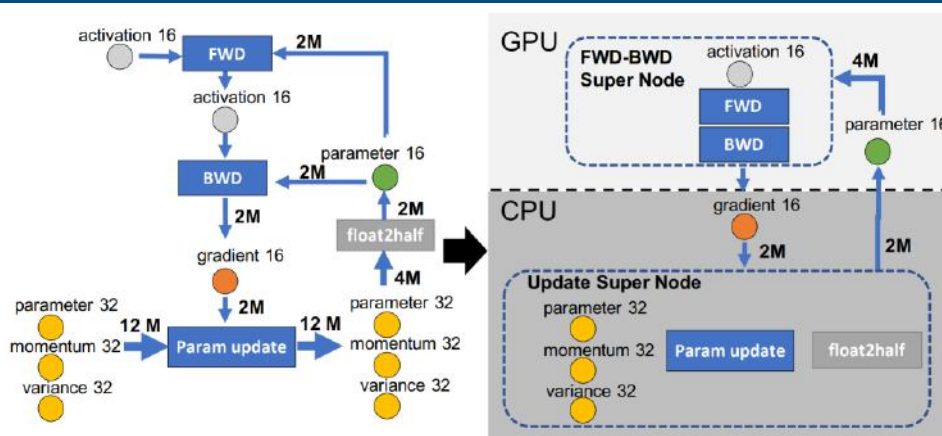


数据来源:《ZeRO: Memory Optimizations Toward Training Trillion Parameter Models》, 中信建投

注: P_{OS} 指对 Adam 优化器进行分片, P_{OS+g} 指对 Adam 优化器及梯度进行分片, P_{OS+g+p} 指对 Adam 优化器、梯度、模型参数进行分片。

微软研究团队后续于 2021 年 1 月、2021 年 4 月发布 ZeRO-Offload 与 ZeRO-Infinity。其中 ZeRO-Offload 就是引入相对 GPU 显存更廉价的 CPU 内存, 但尽可能避免 CPU 通信对整体系统的拖累; ZeRO-Infinity 相较于 Offload 聚焦单卡场景, 更适用于超大规模训练场景 (业界应用), 资源利用率达到 40% 水平。

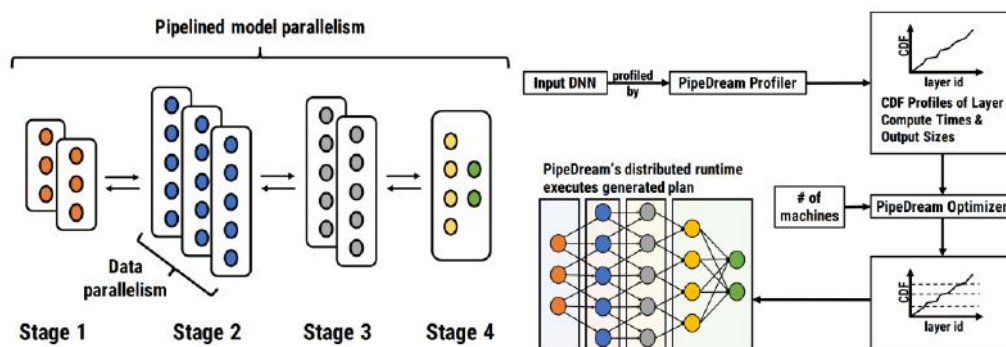
图表174: ZeRO-Offload 对 GPU/CPU 计算的切分



数据来源:《ZeRO-Offload: Democratizing Billion-Scale Model Training》, 中信建投

微软和斯坦福大学、CMU 的研究团队于 2018 年 6 月提出 PipeDream, 通过结合模型、数据、流水并行解决数据并行带来的大量通信成本。

图表175: PipeDream 结合模型并行、数据并行和流水并行降低通信成本



数据来源:《PipeDream: Fast and Efficient Pipeline Parallel DNN Training》, 中信建投

图表176: 不同并行化策略下计算资源利用率情况 (%)



数据来源: Nvidia,《Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM》,《PaLM: Scaling Language Modeling with Pathways》,《ZeRO-infinity: breaking the gpu memory wall for extreme scale deep learning》, 中信建投

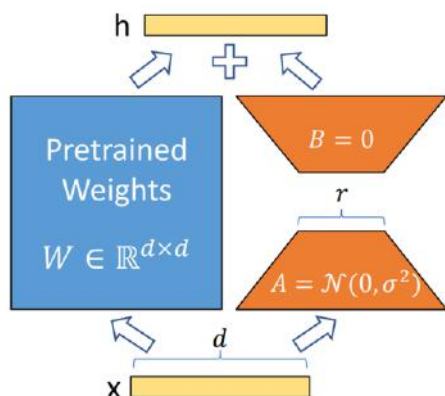
注: Megatron-A100 在不同参数规模上利用率不同,我们选取了与 GPT-3 参数规模接近的情况。

由于 Scaling Law 及 CoT 带来的规模竞赛,当前主流预训练大模型的参数规模普遍在数十亿乃至千亿级别,这导致对所有参数做精调难度较大,因此学术界提出只微调部分参数的思路,但过往的研究一般存在性能损失等问题,即微调后模型性能无法与全参数调试的性能相比。微软和 CMU 的研究团队于 2021 年 6 月提出 LoRA¹²,LoRA 的核心思想是过参数模型存在低内在维度,因此可以通过秩¹³分解矩阵来间接训练神经网络中的一些密集层,同时冻结预训练模型权重,降低了存储占用,同时提升训练速度(减少计算量)。

¹² 《Low-Rank Adaptation of Large Language Models》。

¹³ 矩阵的最大非零子式阶数。

图表177: LoRA 只调试低秩的 A、B, 预训练权重保持不变



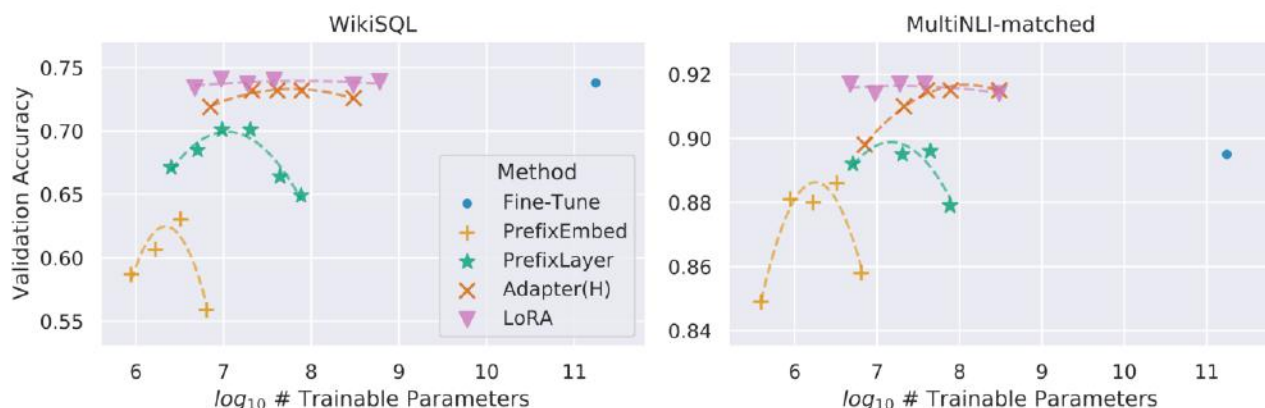
图表178: LoRA 调试下 GPT-2 模型实现训练参数压缩, 同时性能优化

Model & Method	# Trainable Parameters	E2E NLG Challenge				
		BLEU	NIST	MET	ROUGE-L	CIDEr
GPT-2 M (FT)*	354.92M	68.2	8.62	46.2	71.0	2.47
GPT-2 M (Adapter ^L)*	0.37M	66.3	8.41	45.0	69.8	2.40
GPT-2 M (Adapter ^L)*	11.09M	68.9	8.71	46.1	71.3	2.47
GPT-2 M (Adapter ^L)*	11.09M	67.3 ± 0.6	8.50 ± 0.07	46.0 ± 0.2	70.7 ± 0.2	2.44 ± 0.01
GPT-2 M (FT ^{Top2})*	25.19M	68.1	8.59	46.0	70.8	2.41
GPT-2 M (PreLayer)*	0.35M	69.7	8.81	46.1	71.4	2.49
GPT-2 M (LoRA)	0.35M	70.4± 0.1	8.85± 0.02	46.8± 0.2	71.8± 0.1	2.53± 0.02
GPT-2 L (FT)*	774.03M	68.5	8.78	46.0	69.9	2.45
GPT-2 L (Adapter ^L)*	0.88M	69.1 ± 0.1	8.68 ± 0.03	46.3 ± 0.0	71.4 ± 0.2	2.49± 0.0
GPT-2 L (Adapter ^L)*	23.00M	68.9 ± 0.3	8.70 ± 0.04	46.1 ± 0.1	71.3 ± 0.2	2.45 ± 0.02
GPT-2 L (PreLayer)*	0.77M	70.3	8.85	46.2	71.7	2.47
GPT-2 L (LoRA)	0.77M	70.4± 0.1	8.89± 0.02	46.8± 0.2	72.0± 0.2	2.47± 0.02

数据来源: 《Low-Rank Adaptation of Large Language Models》, 中信建投

数据来源: 《Low-Rank Adaptation of Large Language Models》, 中信建投

图表179: LoRA 调试策略下训练参数大幅减少, 同时性能与 Fine-tune 持平或更好



数据来源: 《Low-Rank Adaptation of Large Language Models》, 中信建投

7.3 Meta

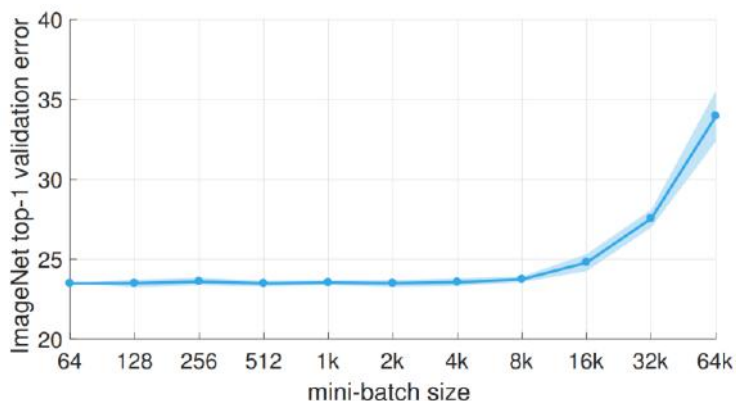
Meta 在预训练、下游调试等环节研究基本处于第一梯队, 在全面性方面落后于 Google, 但其 LLM 选择开源的差异化路线, 并开始探索不同于 Transformer 架构的 Megabyte 路线。

在加速器方面, Meta 研究团队 2017 年 6 月¹⁴则提出通过调整学习率 (learning rate), 以及配合 Warm-up 等操作, 基于 ResNet-50 大批量训练的性能损失能够显著减少, 但后续 UCB、CMU 和英伟达团队 2017 年 8 月的研究¹⁵表明这一方法难以推广至其他模型, 并因此提出基于 SGD 的 LARS 优化器。

¹⁴ 《Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour》。

¹⁵ 《Large Batch Training of Convolutional Networks》。

图表180： 通过调整学习率，ResNet-50 mini-batch 训练可实现 8K 内性能不损失



数据来源：《Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour》，中信建投

注：1K=1024。

图表181： 对于 AlexNet 网络，不同层的权值和其梯度的范数的比值差异很大

Layer	conv1.b	conv1.w	conv2.b	conv2.w	conv3.b	conv3.w	conv4.b	conv4.w
$\ w\ $	1.86	0.098	5.546	0.16	9.40	0.196	8.15	0.196
$\ \nabla L(w)\ $	0.22	0.017	0.165	0.002	0.135	0.0015	0.109	0.0013
$\frac{\ w\ }{\ \nabla L(w)\ }$	8.48	5.76	33.6	83.5	69.9	127	74.6	148
Layer	conv5.b	conv5.w	fc6.b	fc6.w	fc7.b	fc7.w	fc8.b	fc8.w
$\ w\ $	6.65	0.16	30.7	6.4	20.5	6.4	20.2	0.316
$\ \nabla L(w)\ $	0.09	0.0002	0.26	0.005	0.30	0.013	0.22	0.016
$\frac{\ w\ }{\ \nabla L(w)\ }$	73.6	69	117	1345	68	489	93	19

数据来源：《Large Batch Training of Convolutional Networks》，中信建投

注：如果比值差异很大，增大 Batch size，同时提升 learning rate，可能会导致一些层无法更新权重。

图表182： LARS 优化器主要根据范数的比值来调节每一层的学习率

Algorithm 1 SGD with LARS. Example with weight decay, momentum and polynomial LR decay.

Parameters: base LR γ_0 , momentum m , weight decay β , LARS coefficient η , number of steps T

Init: $t = 0, v = 0$. Init weight w_0^l for each layer l

while $t < T$ **for each layer** l **do**

$g_t^l \leftarrow \nabla L(w_t^l)$ (obtain a stochastic gradient for the current mini-batch)

$\gamma_t \leftarrow \gamma_0 * (1 - \frac{t}{T})^2$ (compute the global learning rate)

$\lambda^l \leftarrow \frac{\|w_t^l\|}{\|g_t^l\| + \beta \|w_t^l\|}$ (compute the local LR λ^l)

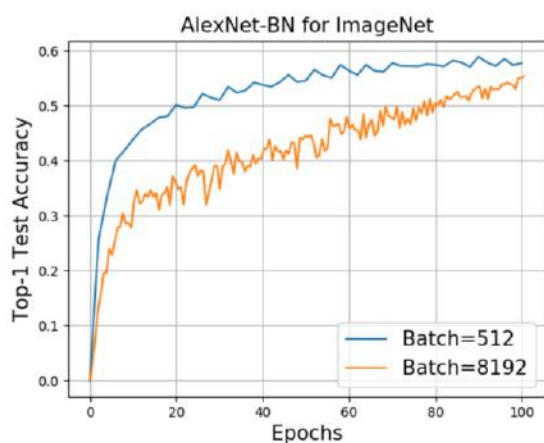
$v_{t+1}^l \leftarrow mv_t^l + \gamma_{t+1} * \lambda^l * (g_t^l + \beta w_t^l)$ (update the momentum)

$w_{t+1}^l \leftarrow w_t^l - v_{t+1}^l$ (update the weights)

end while

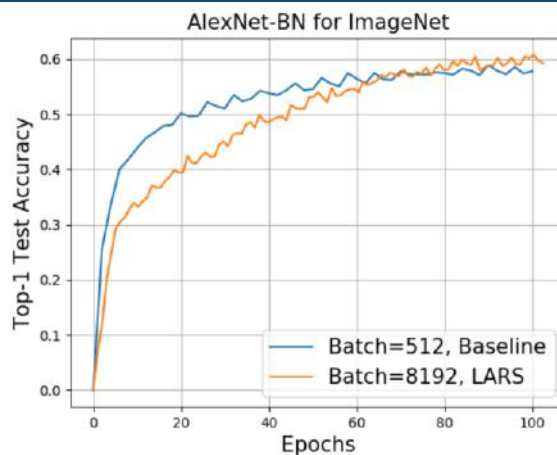
数据来源：《Large Batch Training of Convolutional Networks》，中信建投

图表183: W/O LARS 时 AlexNet-BN 8K 训练存在性能损失



(a) Training without LARS

图表184: W/ LARS 时 AlexNet-BN 8K 训练不存在性能损失

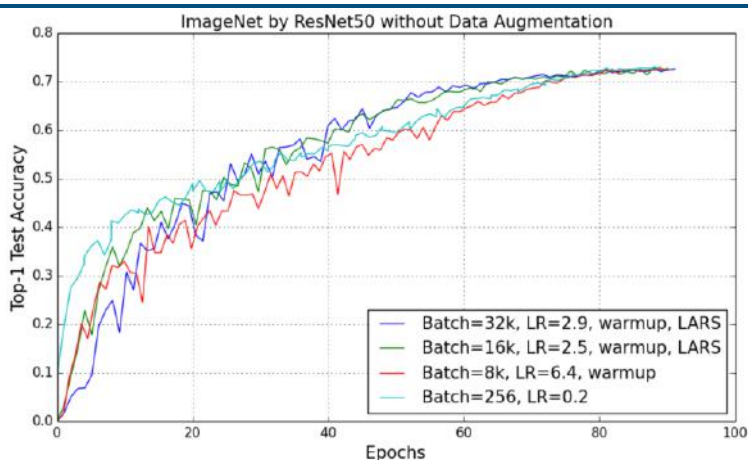


(b) Training with LARS

数据来源:《Large Batch Training of Convolutional Networks》, 中信建投

数据来源:《Large Batch Training of Convolutional Networks》, 中信建投

图表185: LARS 优化器将 ResNet 50 无损训练批量提升至 32K



数据来源:《Large Batch Training of Convolutional Networks》, 中信建投

由于 LARS 优化器在 BERT 等模型应用仍存在缺陷, 谷歌、UCB、UCLA 团队于 2020 年 4 月提出¹⁶基于 Adam 的 LAMB 优化器, 将此前的思路移植到 Attention 机制的相关模型, 例如 BERT, 并实现较好的效果。

¹⁶ 《Large Batch Optimization for Deep Learning: Training BERT in 76 minutes》, 论文一作尤洋也是 LARS 论文一作 (其在英伟达实习期间的工作)。加入谷歌后, 其延续此前工作思路, 提出 LAMB 优化器。

图表186: LARS 与 LAMB 算法对比

Algorithm 1 LARS	Algorithm 2 LAMB
Input: $x_1 \in \mathbb{R}^d$, learning rate $\{\eta_t\}_{t=1}^T$, parameter $0 < \beta_1 < 1$, scaling function ϕ , $\epsilon > 0$ Set $m_0 = 0$ for $t = 1$ to T do Draw b samples S_t from $\mathbb{P}$ Compute $g_t = \frac{1}{ S_t } \sum_{s_t \in S_t} \nabla \ell(x_t, s_t)$ $m_t = \beta_1 m_{t-1} + (1 - \beta_1)(g_t + \lambda x_t)$ $x_{t+1}^{(i)} = x_t^{(i)} - \eta_t \frac{\phi(\ m_t^{(i)}\)}{\ m_t^{(i)}\ } m_t^{(i)}$ for all $i \in [b]$ end for	Input: $x_1 \in \mathbb{R}^d$, learning rate $\{\eta_t\}_{t=1}^T$, parameters $0 < \beta_1, \beta_2 < 1$, scaling function ϕ , $\epsilon > 0$ Set $m_0 = 0, v_0 = 0$ for $t = 1$ to T do Draw b samples S_t from $\mathbb{P}$ Compute $g_t = \frac{1}{ S_t } \sum_{s_t \in S_t} \nabla \ell(x_t, s_t)$ $m_t = \beta_1 m_{t-1} + (1 - \beta_1)g_t$ $v_t = \beta_2 v_{t-1} + (1 - \beta_2)g_t^2$ $m_t = m_t / (1 - \beta_1^t)$ $v_t = v_t / (1 - \beta_2^t)$ Compute ratio $r_t = \frac{m_t}{\sqrt{v_t} + \epsilon}$ $x_{t+1}^{(i)} = x_t^{(i)} - \eta_t \frac{\phi(\ r_t^{(i)}\)}{\ r_t^{(i)}\ + \lambda x_t^{(i)}} (r_t^{(i)} + \lambda x_t^{(i)})$ end for

数据来源:《Large Batch Optimization for Deep Learning: Training BERT in 76 minutes》, 中信建投

图表187: LAMB 优化器训练下 BERT 模型的训练批量可扩展至 32K

Batch Size	512	1K	2K	4K	8K	16K	32K
Learning Rate	$\frac{5}{2^{3.0} \times 10^3}$	$\frac{5}{2^{2.5} \times 10^3}$	$\frac{5}{2^{2.0} \times 10^3}$	$\frac{5}{2^{1.5} \times 10^3}$	$\frac{5}{2^{1.0} \times 10^3}$	$\frac{5}{2^{0.5} \times 10^3}$	$\frac{5}{2^{0.0} \times 10^3}$
Warmup Ratio	$\frac{1}{320}$	$\frac{1}{160}$	$\frac{1}{80}$	$\frac{1}{40}$	$\frac{1}{20}$	$\frac{1}{10}$	$\frac{1}{5}$
F1 score	91.752	91.761	91.946	91.137	91.263	91.345	91.475
Exact Match	85.090	85.260	85.355	84.172	84.901	84.816	84.939

数据来源:《Large Batch Optimization for Deep Learning: Training BERT in 76 minutes》, 中信建投

在增量学习方面, Meta 团队 2017 年 6 月提出 Gradient Episodic Memory¹⁷ (GEM, 梯度片段记忆), 主要逻辑是不更新旧参数, 并且针对新参数更新施加约束, 希望更新后的模型在原有任务的表现不下降。总体来说, 基于回放的增量学习需要额外的计算资源和存储空间用于回忆旧知识, 当任务种类不断增多时, 可能存在计算成本和内存占用增加, 且存储旧知识可能涉及数据安全与隐私保护。

图表188: GEM 算法

Algorithm 1 Training a GEM over an ordered continuum of data	Algorithm 2 EVALUATE
procedure TRAIN(f_θ , Continuum _{train} , Continuum _{test}) $\mathcal{M}_t \leftarrow \{\}$ for all $t = 1, \dots, T$. $R \leftarrow 0 \in \mathbb{R}^{T \times T}$. for $t = 1, \dots, T$ do : for (x, y) in Continuum _{train} (t) do $\mathcal{M}_t \leftarrow \mathcal{M}_t \cup (x, y)$ $g \leftarrow \nabla_\theta \ell(f_\theta(x, t), y)$ $g_k \leftarrow \nabla_\theta \ell(f_\theta, \mathcal{M}_k)$ for all $k < t$ $\tilde{g} \leftarrow \text{PROJECT}(g, g_1, \dots, g_{t-1})$, see (11). $\theta \leftarrow \theta - \alpha \tilde{g}$. end for $R_{t,:} \leftarrow \text{EVALUATE}(f_\theta, \text{Continuum}_{\text{test}})$ end for return f_θ, R end procedure	procedure EVALUATE(f_θ , Continuum) $r \leftarrow 0 \in \mathbb{R}^T$ for $k = 1, \dots, T$ do $r_k \leftarrow 0$ for (x, y) in Continuum(k) do $r_k \leftarrow r_k + \text{accuracy}(f_\theta(x, k), y)$ end for $r_k \leftarrow r_k / \text{len}(\text{Continuum}(k))$ end for return r end procedure

数据来源:《Gradient Episodic Memory for Continual Learning》, 中信建投

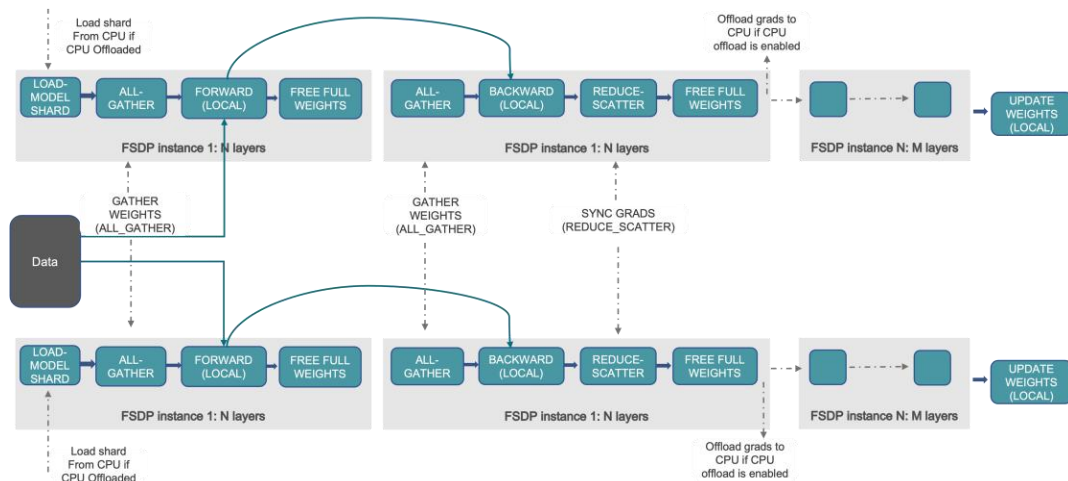
在训练策略优化方面, Meta 研究团队于 2021 年 7 月提出 FSDP¹⁸, 对标 ZeRO。FSDP 在此前 DDP¹⁹ (分

¹⁷ 《Gradient Episodic Memory for Continual Learning》。

¹⁸ <https://engineering.fb.com/2021/07/15/open-source/fsdp/>

布式数据并行）基础上做了优化，结合参数切分（parameter sharding），实现训练效率提升。

图表189： FSDP workflow



数据来源：Pytorch²⁰， 中信建投

八、投资建议

生成式 AI 取得突破，实现了从 0 到 1 的跨越，以 ChatGPT 为代表的人工智能大模型训练和推理需要强大的算力支撑。自 2022 年底 OpenAI 正式推出 ChatGPT 后，用户量大幅增长，围绕 ChatGPT 相关的应用层出不穷，其通用性能力帮助人类在文字等工作上节省了大量时间。同时在 Transformer 新架构下，多模态大模型也取得新的突破，文生图、文生视频等功能不断完善，并在广告、游戏等领域取得不错的进展。生成式 AI 将是未来几年最重要的生产力工具，并深刻改变各个产业环节，围绕生成式 AI，无论是训练还是推理端，算力需求都将有望爆发式增长。

训练和推理端 AI 算力需求或几何倍数增长。首先是训练侧，参考 OpenAI 论文，大模型训练侧算力需求=训练所需要的 token 数量*6*大模型参数量。可以看到从 GPT3.5 到 GPT4，模型效果越来越好，模型也越来越大，训练所需要的 token 数量和参数量均大幅增长，相应的训练算力需求也大幅增长。并且，与 GPT4 相关的公开论文也比较少，各家巨头向 GPT4 迈进的时候，需要更多方向上的探索，也将带来更多的训练侧算力需求。根据我们的推算，2023 年-2027 年，全球大模型训练端峰值算力需求量的年复合增长率有望达到 78.0%，2023 年全球大模型训练端所需全部算力换算成的 A100 芯片总量可能超过 200 万张。其次是推理侧，单个 token 的推理过程整体运算量为 2*大模型参数量，因此大模型推理侧每日算力需求=每日调用大模型次数*每人平均查询 Token 数量*2*大模型参数量，仅以 Google 搜索引擎为例，每年调用次数至少超过 2 万亿，一旦和大模型结合，其 AI 算力需求将十分可观。随着越来越多的应用和大模型结合，推理侧算力需求也有望呈现爆发增长势头。根据我们的推算，2023 年-2027 年，全球大模型云端推理的峰值算力需求量的年复合增长率有望高达 113%。

算力产业链价值放量顺序如下：先进制程制造->以 Chiplet 为代表的 2.5D/3D 封装、HBM->AI 芯片->板卡组装->交换机->光模块->液冷->AI 服务器->IDC 出租运维。

¹⁹ 《PyTorch Distributed: Experiences on Accelerating Data Parallel Training》。

²⁰ <https://pytorch.org/blog/introducing-pytorch-fully-sharded-data-parallel-api/>

先进封装、HBM：为了解决先进制程成本快速提升和“内存墙”等问题，Chiplet 设计+异构先进封装成为性能与成本平衡的最佳方案，台积电开发的 CoWoS 封装技术可以实现计算核心与 HBM 通过 2.5D 封装互连，因此英伟达 A100、H100 等 AI 芯片纷纷采用台积电 CoWoS 封装，并分别配备 40GB HBM2E、80GB 的 HBM3 内存。全球晶圆代工龙头台积电打造全球 2.5D/3D 先进封装工艺标杆，未来几年封装市场增长主要受益于先进封装的扩产。先进封装市场的快速增长，有望成为国内晶圆代工厂商（中芯国际）与封测厂商（长电科技、通富微电、甬矽电子和深科技）的新一轮成长驱动力。

AI 芯片/板卡封装：以英伟达为代表，今年二季度开始释放业绩。模型训练需要规模化的算力芯片部署于智能服务器，CPU 不可或缺，但性能提升遭遇瓶颈，CPU+GPU 异构方案成为大算力场景标配。其中 GPU 并行计算优势明显，CPU+GPU 成为目前最流行的异构计算系统，而 NPU 在特定场景下的性能、效率优势明显，推理端应用潜力巨大，随着大模型多模态发展，硬件需求有望从 GPU 扩展至周边编解码硬件。AI 加速芯片市场上，英伟达凭借其硬件产品性能的先进性和生态构建的完善性处于市场领导地位，在训练、推理端均占据领先地位。根据 Liftr Insights 数据，2022 年数据中心 AI 加速市场中，英伟达份额达 82%。因此 AI 芯片需求爆发，英伟达最为受益，其 Q2 收入指引 110 亿美金，预计其数据中心芯片业务收入接近翻倍。国内厂商虽然在硬件产品性能和产业链生态架构方面与前者有所差距，但正在逐步完善产品布局和生态构建，不断缩小与行业龙头厂商的差距，并且英伟达、AMD 对华供应高端 GPU 芯片受限，国产算力芯片迎来国产替代窗口期。当前已经涌现出一大批国产算力芯片厂商：1）寒武纪：国内人工智能芯片领军者，持续强化核心竞争力；2）海光信息：深算系列 GPGPU 提供高性能算力，升级迭代稳步推进；3）龙芯中科：自主架构 CPU 行业先行者，新品频发加速驱动成长；4）芯原股份：国内半导体 IP 龙头，技术储备丰富驱动成长；5）工业富联：提供 GPU 芯片板卡组装服务。

交换机：与传统数据中心的网络架构相比，AI 数据网络架构会带来更多的交换机端口的需求。交换机具备技术壁垒，中国市场格局稳定，华为与新华三（紫光股份）两强争霸，锐捷网络展现追赶势头，建议重点关注。

光模块：AI 算力带动数据中心内部数据流量较大，光模块速率及数量均有显著提升。训练侧光模块需求与 GPU 出货量强相关，推理侧光模块需求与数据流量强相关，伴随应用加速渗透，未来推理所需的算力和流量实际上可能远大于训练。目前，训练侧英伟达的 A100 GPU 主要对应 200G 光模块和 400G 光模块，H100 GPU 可以对应 400G 或 800G 光模块。根据我们的测算，训练端 A100 和 200G 光模块的比例是 1:7，H100 和 800G 光模块的比例是 1:3.5。800G 光模块 2022 年底开始小批量出货，2023 年需求主要来自于英伟达和谷歌。在 2023 年这个时间点，市场下一代高速率光模块均指向 800G 光模块，叠加 AIGC 带来的算力和模型竞赛，我们预计北美各大云厂商和相关科技巨头均有望在 2024 年大量采购 800G 光模块，同时 2023 年也可能提前采购。建议关注中际旭创、天孚通信、新易盛、华工科技、源杰科技、太辰光、光迅科技、光库科技、中瓷电子、剑桥科技、博创科技、联特科技、德科立、仕佳光子等。

光模块上游——光芯片：以 AWG、PLC 等为代表的无源光芯片，国内厂商市占率全球领先。以 EEL、VCSEL、DFB 等激光器芯片、探测器芯片和调制器芯片为代表的有源光芯片是现代光学技术的重要基石，是有源光器件的重要组成部分。以源杰科技、光库科技为代表的国内光芯片厂商不断攻城拔寨，在多个细分产品领域取得了较大进展，国产替代化加速推进，市场空间广阔。

液冷：AI 大模型训练和推理所用的 GPU 服务器功率密度将大幅提升，以英伟达 DGX A100 服务器为例，其单机最大功率约可达到 6.5kW，大幅超过单台普通 CPU 服务器 500w 左右的功率水平。根据《冷板式液冷服务器可靠性白皮书》数据显示，自然风冷的数据中心单柜密度一般只支持 8kW-10kW，通常液冷数据中心单机柜

可支持 30kW 以上的散热能力,并能较好演进到 100kW 以上,相较而言液冷的散热能力和经济性均有明显优势。同时“东数西算”明确 PUE (数据中心总能耗/IT 设备能耗)要求,枢纽节点 PUE 要求更高,同时考虑到整体规划布局,未来新增机柜更多将在枢纽节点内,风冷方案在某些地区可能无法严格满足要求,液冷方案渗透率有望加速提升。目前在 AI 算力需求的推动下,如浪潮信息、中兴通讯等服务器厂商已经开始大力布局液冷服务器产品。在液冷方案加速渗透过程中,数据中心温控厂商、液冷板制造厂商等有望受益,建议关注:英维克、高澜股份、网宿科技、曙光数创等。

AI 服务器: 预计今年 Q2-Q3 开始逐步释放业绩。具体来看,训练型 AI 服务器成本中,约 7 成以上由 GPU 构成,其余 CPU、存储、内存等占比相对较小,均价常达到百万元以上。对于推理型服务器,其 GPU 成本约为 2-3 成,整体成本构成与高性能型相近,价格常在 20-30 万。根据 IDC 数据,2022 年全球 AI 服务器市场规模 202 亿美元,同比增长 29.8%,占服务器市场规模的比例为 16.4%,同比提升 1.2pct。我们认为全球 AI 服务器市场规模未来 3 年内将保持高速增长,市场规模分别为 395/890/1601 亿美元,对应增速 96%/125%/80%。根据 IDC 数据,2022 年中国 AI 服务器市场规模 67 亿美元,同比增长 24%。我们预计,2023-2025 年,结合对于全球 AI 服务器市场规模的预判,以及对于我国份额占比持续提升的假设,我国 AI 服务器市场规模有望达到 134/307/561 亿美元,同比增长 101%/128%/83%。竞争格局方面,考虑到 AI 服务器研发和投入上需要更充足的资金及技术支持,国内市场的竞争格局预计将继续向头部集中,保持一超多强的竞争格局。重点推荐: 1) 浪潮信息: 全球服务器行业龙头厂商,其 AI 服务器多次位列全球市占率第一; 2) 工业富联: 为英伟达提供 H100 等芯片组装,以及 AI 服务器生产; 3) 紫光股份: 子公司新华三 AI 服务器在手订单饱满,同时可以提供交换机、路由器等; 4) 中科曙光: 高性能计算及国产化服务器龙头; 5) 中兴通讯: 服务器业务快速增长; 6) 拓维信息: 华为昇腾+鲲鹏核心合作伙伴; 7) 联想集团: 全球领先的 ICT 设备企业。

IDC: 在数字中国和人工智能推动云计算市场回暖的背景下, IDC 作为云基础设施产业链的关键环节,也有望进入需求释放阶段。在过去两年半,受多重因素影响下,云计算需求景气度下行,但 IDC 建设与供给未出现明显放缓,2021 年和 2022 年分别新增机柜数量 120 万架和 150 万架,因此短期内出现供需失衡情况(核心区域供需状况相对良好),部分地区上电率情况一般。所以 IDC 公司 2022 年业绩普遍承压。当前,我们认为国内 IDC 行业有望边际向好。随着宏观经济向好,平台经济发展恢复, AI 等拉动, IDC 需求有望逐步释放,叠加 2023 新增供给量有望较 2022 年减少(例如三大运营商 2022 年新增 IDC 机柜 15.6 万架, 2023 年计划新增 11.4 万架)。展望未来,电信运营商在云计算业务方面仍将实现快速增长,百度、字节跳动等互联网公司在 AIGC 领域有望实现突破性进展,都将对包括 IDC 在内的云基础设施产生较大新增需求,相关 IDC 厂商有望获益,建议关注润泽科技、宝信软件、奥飞数据、数据港、光环新网等。

风险提示: 国产替代进程不及预期。GPU 的国产替代过程中面临诸多困难,国产替代进程可能不及预期; AI 技术进展不及预期。当前 AI 技术的快速进步带动了巨大的 AI 算力需求,如果 AI 技术进展不及预期,可能对 GPU 市场的整体需求产生不利影响; 互联网厂商资本开支不及预期。互联网厂商是 AI 算力和 GPGPU 的重要采购方和使用方,如果互联网厂商资本开支不及预期,可能会对 GPGPU 的需求情况产生不利影响; 在 GPU 需求旺盛的背景下,国内外涌现出诸多 GPU 行业的新兴玩家,众多参与厂商可能导致整体竞争格局恶化。

分析师介绍

武超则

中信建投证券研究所所长兼国际业务部负责人，董事总经理，TMT 行业首席分析师。新财富白金分析师，2013-2020 年连续八届新财富最佳分析师通信行业第一名；2014-2020 年连续七届水晶球最佳分析师通信行业第一名。专注于 5G、云计算、物联网等领域研究。中国证券业协会证券分析师、投资顾问与首席经济学家委员会委员。

阎贵成

中信建投证券通信&计算机行业首席分析师，北京大学学士、硕士，专注于云计算、物联网、信息安全、信创与 5G 等领域研究。近 8 年中国移动工作经验，6 年多证券研究经验。系 2019-2021 年《新财富》、《水晶球》通信行业最佳分析师第一名，2017-2018 年《新财富》、《水晶球》通信行业最佳分析师第一名团队核心成员。

刘双锋

中信建投证券电子首席分析师。3 年深南电路，5 年华为工作经验，从事市场洞察、战略规划工作，涉及通信服务、云计算及终端领域，专注于通信服务领域，2018 年加入中信建投通信团队。2018 年 IAMAC 最受欢迎卖方分析师通信行业第一名团队成员，2018《水晶球》最佳分析师通信行业第一名团队成员。

金戈

中信建投证券研究发展部计算机行业联席首席分析师，帝国理工学院工科硕士，擅长云计算、金融科技、人工智能等领域。

于芳博

中信建投人工智能组首席分析师，北京大学空间物理学学士、硕士，2019 年 7 月加入中信建投，主要覆盖人工智能等方向，下游重点包括智能汽车、CPU/GPU/FPGA/ASIC、EDA 和工业软件等方向

崔世峰

海外研究首席分析师，南京大学硕士，6 年买方及卖方复合从业经历，专注于互联网龙头公司研究，所在卖方团队获得 2019-2020 年新财富传媒最佳研究团队第二名。2022 年新财富海外研究最佳研究团队入围。

刘永旭

通信行业分析师，南开大学学士、硕士，曾从事军工行业研究工作，2020 年加入中信建投通信团队，主要研究云计算 IDC、工业互联网、通信新能源、卫星应用、专网通信等方向。2020-2021 年《新财富》、《水晶球》通信行业最佳分析师第一名团队成员。

杨伟松

通信行业分析师，南京大学理学学士，浙江大学工学硕士。6 年光通信行业研发及管理经验，曾就职于光通信头部企业 Coherent。2022 年 2 月加入中信建投通信团队，主要研究光通信、ICT 设备和激光雷达等方向。

范彬泰

中信建投电子行业分析师。电子科技大学工学学士，香港理工大学会计学硕士。2018年5月加入国金证券研究所，担任半导体行业研究员，重点覆盖集成电路和显示面板两大产业链，2021年加入中信建投电子团队。

研究助理**郑寅铭**

zhengyinming@csc.com.cn

何昱灵

heyuling@csc.com.cn

感谢庞佳军、樊文辉对本报告的贡献。

评级说明

投资评级标准		评级	说明
报告中投资建议涉及的评级标准为报告发布日后6个月内的相对市场表现,也即报告发布日后的6个月内公司股价(或行业指数)相对同期相关证券市场代表性指数的涨跌幅作为基准。A股市场以沪深300指数作为基准;新三板市场以三板成指为基准;香港市场以恒生指数作为基准;美国市场以标普500指数为基准。	股票评级	买入	相对涨幅 15%以上
		增持	相对涨幅 5%—15%
		中性	相对涨幅-5%—5%之间
		减持	相对跌幅 5%—15%
		卖出	相对跌幅 15%以上
	行业评级	强于大市	相对涨幅 10%以上
		中性	相对涨幅-10-10%之间
		弱于大市	相对跌幅 10%以上

分析师声明

本报告署名分析师在此声明:(i)以勤勉的职业态度、专业审慎的研究方法,使用合法合规的信息,独立、客观地出具本报告,结论不受任何第三方的授意或影响。(ii)本人不曾因,不因,也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿。

法律主体说明

本报告由中信建投证券股份有限公司及/或其附属机构(以下合称“中信建投”)制作,由中信建投证券股份有限公司在中华人民共和国(仅为本报告目的,不包括香港、澳门、台湾)提供。中信建投证券股份有限公司具有中国证监会许可的投资咨询业务资格,本报告署名分析师所持中国证券业协会授予的证券投资咨询执业资格证书编号已披露在报告首页。

在遵守适用的法律法规情况下,本报告亦可能由中信建投(国际)证券有限公司在香港提供。本报告作者所持香港证监会牌照的中央编号已披露在报告首页。

一般性声明

本报告由中信建投制作。发送本报告不构成任何合同或承诺的基础,不因接收者收到本报告而视其为中信建投客户。

本报告的信息均来源于中信建投认为可靠的公开资料,但中信建投对这些信息的准确性及完整性不作任何保证。本报告所载观点、评估和预测仅反映本报告出具日该分析师的判断,该等观点、评估和预测可能在不发出通知的情况下有所变更,亦有可能因使用不同假设和标准或者采用不同分析方法而与中信建投其他部门、人员口头或书面表达的意见不同或相反。本报告所引证券或其他金融工具的过往业绩不代表其未来表现。报告中所含任何具有预测性质的内容皆基于相应的假设条件,而任何假设条件都可能随时发生变化并影响实际投资收益。中信建投不承诺、不保证本报告所含具有预测性质的内容必然得以实现。

本报告内容的全部或部分均不构成投资建议。本报告所包含的观点、建议并未考虑报告接收人在财务状况、投资目的、风险偏好等方面的具体情况,报告接收者应当独立评估本报告所含信息,基于自身投资目标、需求、市场机会、风险及其他因素自主做出决策并自行承担投资风险。中信建投建议所有投资者应就任何潜在投资向其税务、会计或法律顾问咨询。不论报告接收者是否根据本报告做出投资决策,中信建投都不对该等投资决策提供任何形式的担保,亦不以任何形式分享投资收益或者分担投资损失。中信建投不对使用本报告所产生的任何直接或间接损失承担责任。

在法律法规及监管规定允许的范围内,中信建投可能持有并交易本报告中所提公司的股份或其他财产权益,也可能在过去12个月、目前或者将来为本报中所提公司提供或者争取为其提供投资银行、做市交易、财务顾问或其他金融服务。本报告内容真实、准确、完整地反映了署名分析师的观点,分析师的薪酬无论过去、现在或未来都不会直接或间接与其所撰写报告中的具体观点相联系,分析师亦不会因撰写本报告而获取不当利益。

本报告为中信建投所有。未经中信建投事先书面许可,任何机构和/或个人不得以任何形式转发、翻版、复制、发布或引用本报告全部或部分内容,亦不得从未经中信建投书面授权的任何机构、个人或其运营的媒体平台接收、翻版、复制或引用本报告全部或部分内容。版权所有,违者必究。

中信建投证券研究发展部

北京
东城区朝内大街2号凯恒中心B座12层
电话:(8610) 8513-0588
联系人:李祉瑶
邮箱:lizhiyao@csc.com.cn

上海
上海浦东新区浦东南路528号南塔2103室
电话:(8621) 6882-1600
联系人:翁起帆
邮箱:wengqifan@csc.com.cn

深圳
福田区福中三路与鹏程一路交汇广电金融中心35楼
电话:(86755) 8252-1369
联系人:曹莹
邮箱:caoying@csc.com.cn

中信建投(国际)

香港
中环交易广场2期18楼
电话:(852) 3465-5600
联系人:刘泓麟
邮箱:charleneliu@csci.hk