

计算机行业

报告日期：2023 年 12 月 04 日

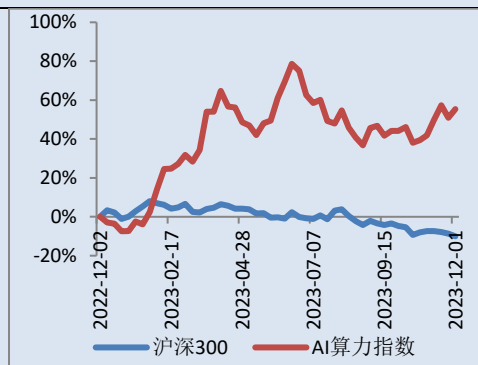
智算供给格局分化，国产化进程有望加速

——AI 算力行业深度研究报告

华龙证券研究所

投资评级：推荐（首次覆盖）

最近一年走势



研究员：孙伯文

执业证书编号：S0230523080004

邮箱：sunbw@hlzqgs.com

相关阅读

请认真阅读文后免责声明

摘要：

- 国产大模型发展方兴未艾。大模型规模、数据量和数量的全面增长将持续拉动 AI 算力需求。国内千亿级参数规模大模型持续落地。2023 年 10 月，百度发布万亿参数级大模型文心一言 4.0，正式宣布对标 GPT-4。近两年国产大模型数量呈爆发式增长态势，仅 2023 年 1 到 7 月，国内大模型就新增 64 个，短期内增速有望维持。
- 美国再收紧对华高端芯片出口标准，AI 芯片国产化替代成为趋势。近年来，受国际局势影响，国内厂商加快自研芯片速度。华为昇腾 AI 算力已经为至少 30 个国产大模型训练提供算力支持，具备大规模商用能力。短期关注华为昇腾和昇思 MindSpore 产业链，长期关注国产芯片研发的突破性进展。
- 算力租赁业务模式逐渐清晰，预期强需求下定价将大幅上涨。随着美国对华高端芯片禁令进一步收紧，较早布局 AI 算力租赁的厂商容易形成资源优势，估算行业成本回收周期约为 17 个月。算力租赁业务的商业化路径逐渐明晰，加之近期行业频释放提价信号，算力租赁利润空间将进一步增长。
- 长期关注芯片性能和算力资源调配效率提升。一方面，Chiplet 技术有望打破芯片的物理极限，延续、甚至提高摩尔定律中提出的芯片性能增长速度。另一方面，AI 算力资源上云有望从资源配置灵活性方面提升算力供给效率。
- 建议关注：1) 开展算力服务的厂商：中贝通信、莲花健康、恒润股份、汇纳科技、鸿博股份；2) 具备国产芯片 IP 设计能力的厂商：寒武纪、海光信息、芯原股份；3) 具备云服务能力的厂商：中科曙光、浪潮信息、紫光股份、神州数码；4) 芯片封测厂商：长电科技、通富微电、华天科技。
- 风险提示：
国际紧张局势加剧；国产大模型发展不及预期；国产芯片技术突破速度不及预期；台积电产能恢复不及预期；数据估算风险。

并购家 免费行业报告网

IPOIPO.CN 每日更新 不用注册会员 无广告 永久免费

内容目录

1 大模型浪潮推动作用下，算力需求缺口将持续扩大	1
1.1 大模型发展对算力需求的推动作用	1
1.1.1 国外大模型的发展	1
1.1.2 国内大模型的发展	3
1.2 国产大模型 AI 算力需求测算	5
1.2.1 通用大模型 AI 算力需求测算	6
1.2.2 行业大模型 AI 算力需求测算	7
1.3 AI 算力供给方面：高端芯片进口受限，国产替代为大势所趋	10
1.3.1 国际形势：美国进一步收紧芯片对华出口标准	10
1.3.2 AI 芯片国产化替代	11
1.3.3 AI 芯片产能：台积电产能复苏伴随订单激增，供不应求情况仍将持续	12
2 AI 算力租赁行业的内在价值	12
2.1 AI 算力租赁对下游公司：带来成本和时间优势	12
2.2 对具备算力资源的公司：算力租赁可为公司带来第二业务增长线	13
2.2.1 AI 算力租赁成本回收周期测算	13
3 AI 算力未来发展方向——增质提效	14
3.1 以云网融合为前提，算力调度成为提高资源配置效率的核心	14
3.2 芯片数据传输效率：关注 Chiplet 技术和芯片互联技术	15
3.2.1 Chiplet 技术	15
3.2.2 芯片互联技术	16
4 重点关注公司	16
4.1 寒武纪-U (688256.SH)	16
4.2 中贝通信 (603220.SH)	17
5 风险提示	19

图目录

图 1: 2018 年海外大模型发布数量.....	1
图 2: 2017-2023 年海外大模型参数量演进情况	2
图 3: 国内大模型发布数量.....	3
图 4: 国产通用大模型与行业大模型分布.....	4
图 5: 国内部分垂类大模型应用.....	4
图 6: 国内通用与垂类大模型比例.....	4
图 7: 华为盘古大模型层级分布.....	5
图 8: 基于华为昇思 MindSpore 的医药垂类大模型-鹏程·神农	8
图 9: Megatron 框架下大模型训练规模和算力利用率	9
图 10: 美国 2023 年 10 月 18 日高端芯片出口禁令标准.....	10
图 11: 寒武纪思元 370 芯片.....	16
图 12: 2020-2025 年寒武纪营业收入及增长率	17
图 13: 2020-2025 年寒武纪归母净利润及增长率	17
图 14: 2020-2025 年寒武纪 ROE(摊薄).....	17
图 15: 2020-2025 年中贝通信营业收入及增长率	19
图 16: 2020-2025 年中贝通信营业归母净利润及增长率	19
图 17: 2020-2025 年中贝通信 ROE(摊薄).....	19
图 18: 2020-2025 年中贝通信 PE.....	19

表目录

表 1: 部分国内外公开数据的通用大模型训练计算量和规模对比.....	6
表 2: 2023 年国产通用大模型训练侧算力需求测算.....	7
表 3: 2023 年国产通用大模型推理侧算力需求测算.....	7
表 4: 2023 年国产垂类大模型算力需求测算.....	9
表 5: 国内主流 AI 芯片性能参数对比	11
表 6: 算力租赁行业成本回收周期测算（以 H800 服务器为租赁产品）	14
表 7: 中贝通信算力租赁业务的成本回收期估算.....	18

1 大模型浪潮推动作用下,算力需求缺口将持续扩大

1.1 大模型发展对算力需求的推动作用

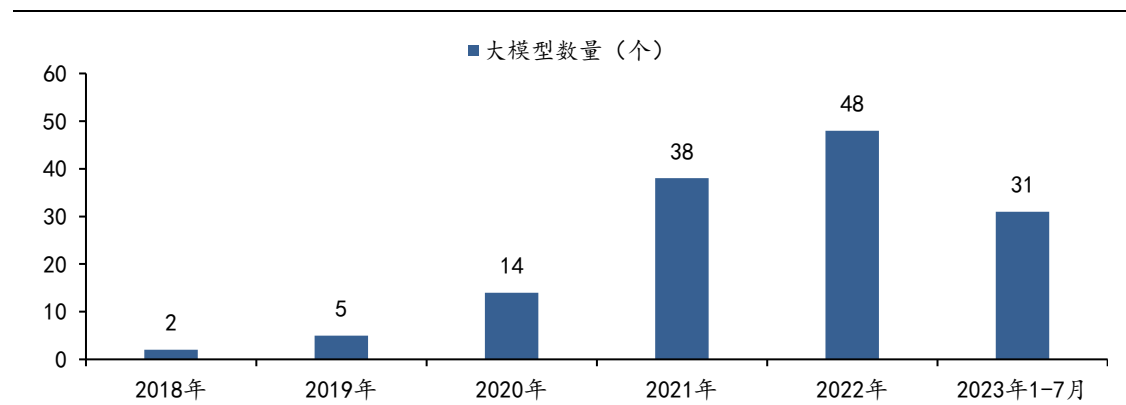
大模型的训练效果、成本和时间与算力资源有密切的关系。大模型发展浪潮有望进一步增加 AI 行业对智算算力的需求规模。

1.1.1 国外大模型的发展

大模型数量加速增长,算力成为模型竞赛底座。自 2018 年以来,海外云厂商巨头接连发布 NLP 大模型。据赛迪顾问 2023 年 7 月发布的数据显示,海外大模型发布数量逐年上升,年发布数量在五年中由 2 个增长至 48 个。且仅 2023 年 1-7 月就发布了 31 个大模型。

自 2021 年起,海外大模型数量呈现加速增长的趋势,结合 2023 年 1-7 月的情况,该趋势有望延续。

图 1: 2018 年海外大模型发布数量



资料来源: 赛迪顾问, 华龙证券研究所

2017-2023 年,从各公司发布的公开信息来看,大模型在 7 年的时间里实现参数量从千万到万亿级的指数型增长。

2017 年,谷歌团队提出 Transformer 架构,奠定

了当前大模型领域主流的算法架构基础。

2018 年 6 月，OpenAI 发布了 Transformer 模型——GPT-1，训练参数量 1.2 亿。同年 10 月，谷歌发布了大规模预训练语言模型 BERT，参数量超过 3 亿。

2019 年，OpenAI 推出 15 亿参数的 GPT-2。2019 年 9 月，英伟达推出了 83 亿参数的 Megatron-LM。同年，谷歌推出了 110 亿参数的 T5，微软推出了 170 亿参数的图灵 Turing-NLG。

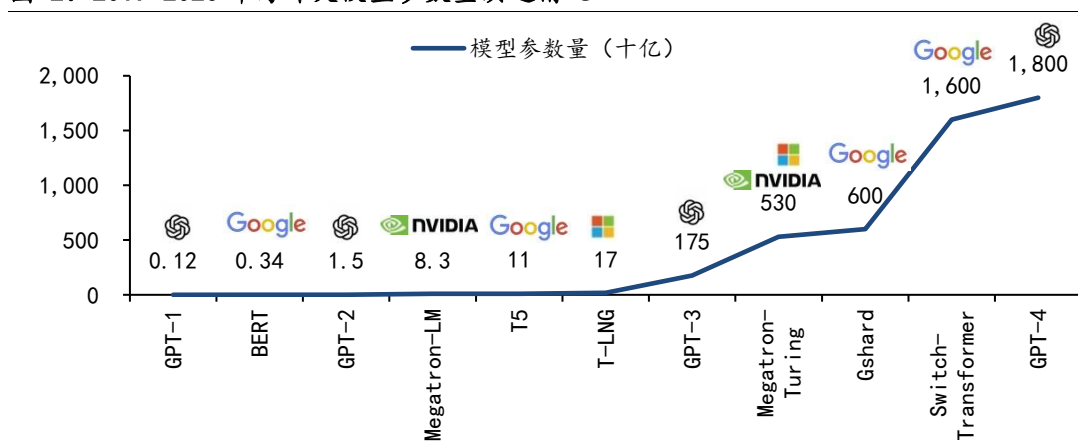
2020 年，OpenAI 推出了大语言训练模型 GPT-3，参数达到 1750 亿。微软和英伟达在同年 10 月联合发布了 5300 亿参数的 Megatron-Turing 大模型。

2021 年 1 月，谷歌推出 Switch Transformer 模型，参数量达到 1.6 万亿，大模型参数量首次突破万亿。

2022 年，OpenAI 推出基于 GPT-3.5 大模型的 ChatGPT，宣告了 GPT-3.5 版本的存在。

2023 年，OpenAI 推出 GPT-4，估计参数规模达到 1.8 万亿。

图 2：2017-2023 年海外大模型参数量演进情况



资料来源：《Attention Is All You Need》等论文，中国信通院，华龙证券研究所

GPU 数量与不同量级大模型所需的算力之间的线性关系。根据 2021 年 8 月 Deepak Narayanan 等人发布的论文，随着模型参数增加，大模型训练需要的总浮点数与 GPU 数量呈现正相关的线性关系。175B 参数量级的大模型所需的 A100 级别芯片数量为 1024 片（Token 数为 300B, 训练 34 天情况下）。当参数增长到 1T 时，大模型训练所需的 A100 芯片数量为 3072 片（Token 数为 450B，训练 84 天情况下）。

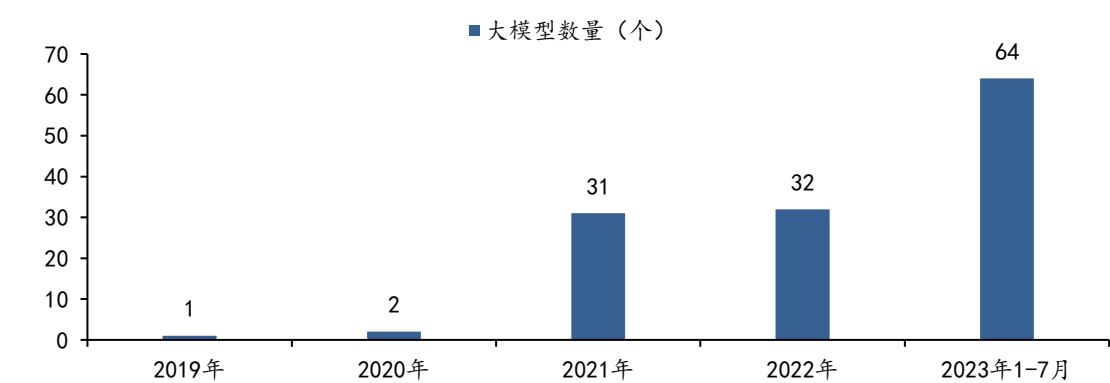
加大成本投入，海外大模型训练周期有望进一步缩

短。从 2020 年 6 月 OpenAI 发布首个千亿参数量级大模型 GPT-3 到 2021 年 1 月谷歌推出首个万亿参数量级的 Switch Transformer 模型，大模型实现参数量级从千亿到万亿的跨越只用了不到一年。随着海外大厂商加大对大模型训练的成本投入，预计大模型发布周期将进一步缩短。在商业逻辑上，大模型发布数量指数型增长，意味着市场竞争越来越激烈。厂商更愿意通过使用高性能的芯片缩短大模型训练时间，使大模型更早投入应用为公司带来业务增长。因此，芯片性能的提高并不会削弱厂商对芯片数量的需求意愿。

1.1.2 国内大模型的发展

数量增长情况与海外类似，短期内呈现密集发布的特点。自 2019 年至 2023 年 7 月底，国内累计发布 130 个大模型，2023 年 1-7 月国内共有 64 个大模型发布，大模型发布数量呈现加速增长趋势。数量增长趋势与海外情况一致，我国大模型研发起步较晚，随着在大模型领域布局的厂商数量快速增加，大模型发布周期逐步缩短，预期未来两到三年内国产大模型数量将呈现爆发式增长局面。

图 3：国内大模型发布数量



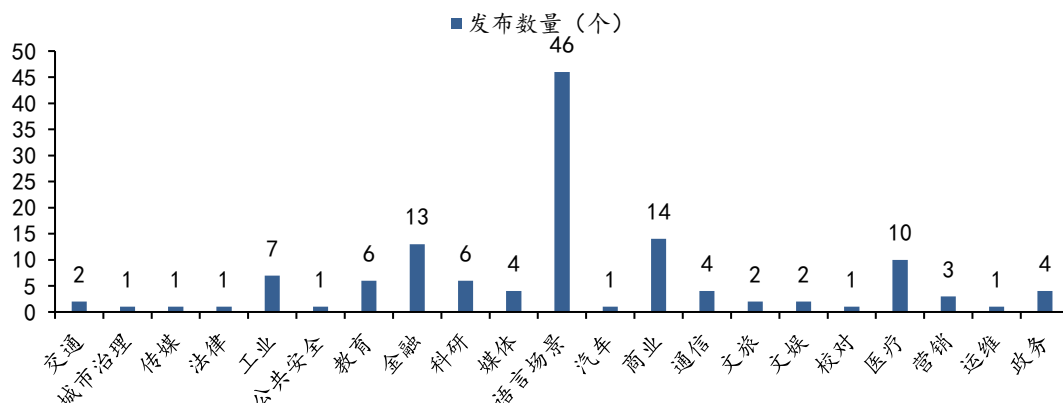
资料来源：赛迪顾问，华龙证券研究所

巨头引领，千亿级参数规模大模型陆续落地。2023 年 3 月，百度发布文心一言 1.0；同年 4 月，阿里发布通义千问大模型、商汤科技发布日日新大模型体系；同年 5 月，科大讯飞发布星火大模型；同年 7 月，华为发布面向行业的盘古大模型 3.0，千亿级参数规模大模型密集发布。2023 年 10 月，随着百度发布万亿级参数大模型文心一言 4.0，国产大模型或将具备对标 GPT-4 性能的能力。

按类型划分，大模型分为行业大模型和通用大模型。

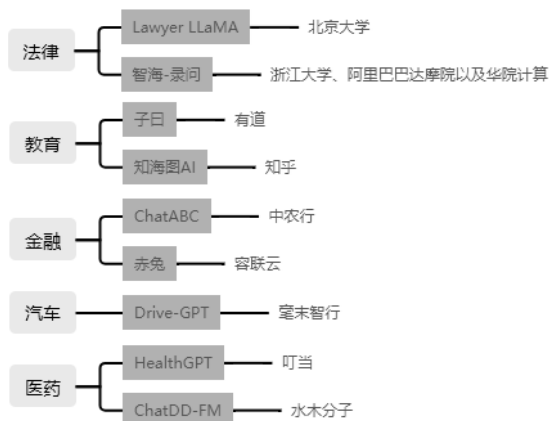
据赛迪顾问于 2023 年 7 月统计的数据显示，我国通用大模型和行业大模型占比分别为 40%和 60%。行业大模型分布较多的领域为商业（14 个）、金融（13 个）、医疗（10 个）、工业（7 个）、教育（6 个）和科研（6 个）。研究显示，通用大模型在行业领域及行业细分场景的表现一般。但行业模型可以在通用模型的基础上通过行业数据库进一步训练出来。

图 4：国产通用大模型与行业大模型分布



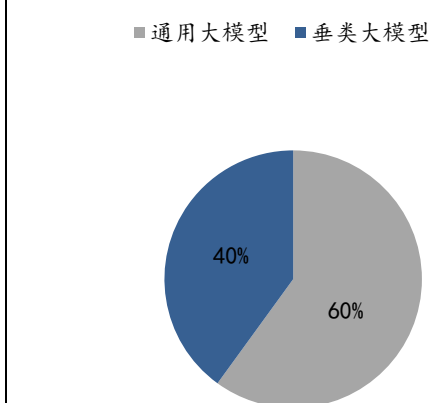
资料来源：赛迪顾问，华龙证券研究所

图 5：国内部分垂类大模型应用



资料来源：中国日报，京报网，中国网科技，各公司官网，华龙证券研究所

图 6：国内通用与垂类大模型比例



资料来源：赛迪顾问，华龙证券研究所

大模型应用向细分场景下沉。华为发布的盘古大模型实际分为 L0（基础大模型）L1（行业大模型）L2（场景模型）三个层级。采取 5+N+X 模式，即 5 个基础大模型、N 个行业大模型和 X 个细分场景应用模型。目前行业模型主要应用于矿山、政务、气象、汽车、医学、数字人和研发共七大领域，覆盖 14 个细分场景。这种

通过基础大模型+行业大模型实现大模型商业化落地的模式已经逐渐得到验证，未来行业大模型有望带动大模型本地化部署热潮，在解决行业长尾问题上将发挥更大优势并成为打通大模型“最后一公里”的桥梁。

图 7：华为盘古大模型层级分布

5	L0基础大模型	盘古自然语言大模型	盘古多模态大模型	盘古视觉大模型	盘古预测大模型	盘古科学计算大模型		
N	L1行业大模型	盘古矿山大模型	盘古政务大模型	盘古气象大模型	盘古汽车大模型	盘古医学大模型	盘古数字人大模型	盘古研发大模型
X	L2场景大模型	传送带异物检测	政务热线	台风路径预测	自动驾驶开发	报告解读	数字人直播	智能测试
		重介选煤洗选	城市事件处理	降水预测	车辆辅助设计	辅助医疗	智能问答	智能运输

资料来源：华为官网，华龙证券研究所

1.2 国产大模型 AI 算力需求测算

大模型算力需求测算方法：

根据 2023 年 8 月腾讯公布的大模型算力评估通用方法，在大模型训练过程中，训练侧算力需求可量化表达为：

训练所需浮点运算量（FLOPs）=6×参数量×Training Tokens

若训练中使用激活重计算技术，则对应算力需求可量化表达为：

训练所需浮点运算量（FLOPs）=8×参数量×Training Tokens

同时，在大模型推理过程中的算力需求可量化表达为：

推理侧所需浮点运算量（FLOPs）=2×参数量×Prompt Tokens

由于激活重计算技术是可选的，因此假设在训练中没有选择使用激活重计算技术，按照以上计算方法可得：

训练 GPT-3 量级的大模型算力需求估算为 3.15E+23 FLOPs；

训练 GPT-4 量级的大模型算力估算为 2.15E+25 FLOPs，由于 GPT-4 采用了混合专家（MoE）模型，实际训练调用参数量按约 2770 亿计算。

表 1：部分国内外公开数据的通用大模型训练计算量和规模对比

名称	发布时间	发布方	模型规模 (参数量)	估计训练数据量 (TOKENS)	估计训练计算量 (FLOPs)
Baichuan2	2023 年 10 月	百川智能	7B、13B	2600B	1.092E+23、2.028E+23
Llama 2	2023 年 7 月	Meta	7B、13B、34B、70B	2000B	8.4E+22-8.4E+23
书生浦语	2023 年 6 月	商汤等	104B	1600B	9.98E+23
GPT-4	2023 年 3 月	OpenAI	1800B	1300B	2.15E+25
GLM-130B	2022 年 10 月	清华大学、智谱 AI	130B	400B	3.12E+23
GPT-3	2020 年 6 月	OpenAI	175B	300B	3.15E+23

资料来源：《Llama 2: Open Foundation and Fine-Tuned Chat Models》等论文，商汤科技，中国信通院，华龙证券研究所

1.2.1 通用大模型 AI 算力需求测算

训练侧：

2023 年 10 月，百度发布文心一言 4.0 大模型。据百度公开的信息，该大模型在综合水平上可以对标 GPT-4。

乐观预期下，2023 年内，国内头部互联网厂商中，百度能够训练出 GPT-4 量级的大模型。且假设阿里、腾讯、字节跳动、商汤、科大讯飞、浪潮、华为这 7 家厂商能够训练出 GPT-3 量级的大模型。

参考 2023 年上半年国产大模型发布数量情况，预估到 2023 年年底，国内大模型发布数量可达约 200 个，年内新增约 134 个，其中通用大模型新增约 80 个。除头部大厂商外，其他厂商和科研机构发布的通用大模型数量估计为 72 个，参数在百亿至千亿之间，保守估计平均参数量级为 500 亿。

由此计算，2023 年年内国产通用大模型训练侧算力需求为 3.03E+25 FLOPs。

表 2：2023 年国产通用大模型训练侧算力需求测算

大模型量级	单模型所需 算力 (FLOPs)	数量 (个)	各量级大模型总算力需求 (FLOPs)
其他量级	9E+22	72	6.48E+24
GPT-3 量级	3.15E+23	7	2.2E+24
GPT-4 量级	2.16E+25	1	2.16E+25
训练侧总算力需求 (FLOPs)			3.03E+25

资料来源：华龙证券研究所

推理侧：

推理侧的算力需求需要在用户的访问峰值情境下计算。通常情况下，访问时间中，80%的访问量都集中在 20%的访问时间里。

以 GPT-4 为基准，按日访问量 2 亿（次），每个用户占用 Tokens 数 80 计算，单模型推理算力需求为 4.10E+17 FLOPs。

以 GPT-3 为基准，按日访问量 1 亿（次），每个用户占用 Tokens 数 80 计算，单模型推理算力需求为 1.30E+17 FLOPs。

以 500 亿参数大模型为基准，按日访问量 5000 万（次），每个用户占用 Tokens 数 80 计算，单模型推理算力需求为 1.85E+16 FLOPs。

结合各参数基准下的大模型数量，2023 年国产通用大模型推理侧算力需求预估为 2.65E+18 FLOPs。

表 3：2023 年国产通用大模型推理侧算力需求测算

大模型量级	单模型算力需求 (FLOPs)	数量 (个)	各量级大模型算力需求 (FLOPs)
GPT-4 量级	4.10E+17	1	4.10E+17
GPT-3 量级	1.30E+17	7	9.07E+17
其他量级	1.85E+16	72	1.33E+18
推理侧总算力需求 (FLOPs)			2.65E+18

资料来源：华龙证券研究所

由此，2023 年国产通用大模型训练和推理侧的算力需求总和为 3.03E+25 FLOPs。

1.2.2 行业大模型 AI 算力需求测算

我国垂类大模型主要分布在遥感、生物制药、气象、轨道交通、代码生成/编辑、金融等领域。未来垂类大模型数量有望随着其在各行业细分场景的渗透上升而加速增长。华为已经在算力和软硬件方面，为多个国产垂类大模型的训练提供支持。在医疗方面，华为和医渡科技于 2023 年 9 月在华为全联接大会上联合发布医疗垂类领域大模型训推一体机。该一体机由昇腾 AI 提供算力支持，内置医渡科技研发的医疗垂类大模型，目标是帮助医院、机构等医疗场所实现大模型私有化。在遥感方面，2022 年 8 月，中科院推出了“空天·灵眸”遥感预训练大模型。该大模型基于华为昇腾 AI 澎湃算力和 MindSpore 训练而成，有望在中科星图的线下业务中，通过 AI 赋能公司的数字化产品。

总结近年来国内大模型商业化落地的过程和效果可以得出，商业化的一般路径为：厂商基于通用大模型训练行业垂类大模型，再通过定制化服务为企业提供所处行业的细分场景 AI 解决方案。

从垂类大模型数量上看，截至 2023 年上半年，垂类大模型占国产大模型的 40%，预计 2023 年新增量为 54 个。

图 8：基于华为昇思 MindSpore 的医药垂类大模型-鹏程·神农



资料来源：MindSpore 官网，华龙证券研究所

垂类大模型训练侧算力需求测算：

2023 年国内发布的垂类大模型参数量在百亿-千亿量级范围内，按平均 500 亿参数和 5,000 亿 Tokens 估算，训练侧总算力需求为 8.1E+24 FLOPs。

垂类大模型推理侧算力需求测算：

按日访问量 3000 万（次），每个用户占用 Tokens 数 80 计算，2023 年国产垂类大模型推理侧算力需求为 6E+17 FLOPs。

由此,2023 年国内发布的垂类大模型训练侧和推理侧总算力需求为 8.1E+24 FLOPs。

表 4: 2023 年国产垂类大模型算力需求测算

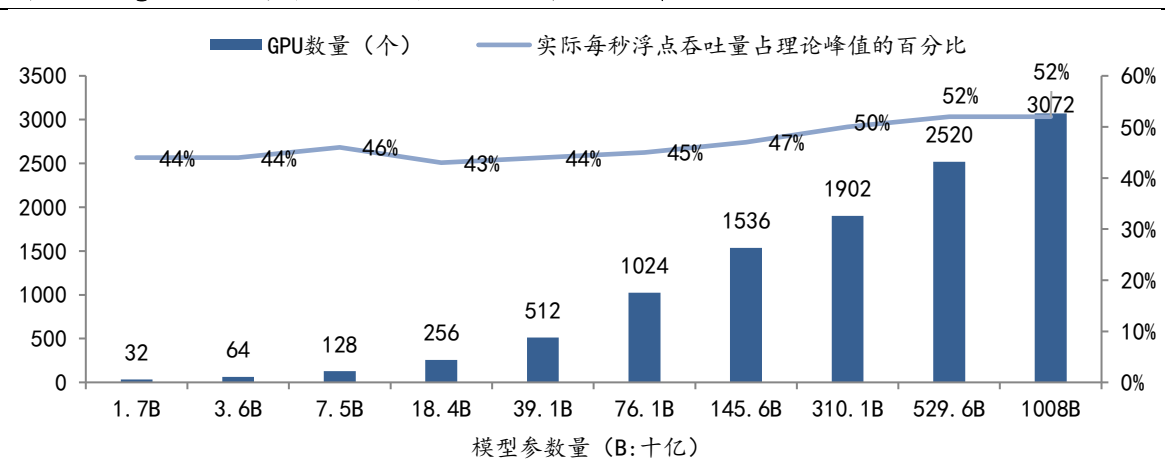
训练侧算力需求 (FLOPs)		推理侧算力需求 (FLOPs)	
单模算力需求	1.5E+23	单模算力需求	1.11E+16
数量	54	数量	54
训练总算力需求	8.1E+24	推理总算力需求	6E+17
垂类大模型总算力需求 (FLOPs)		8.1E+24	

资料来源: 华龙证券研究所

综上,根据我们对国产通用大模型和垂类大模型的算力需求测算,预计 2023 年国产大模型总算力需求为 3.84E+25 FLOPs。

AI 芯片需求——大模型算力需求具象表现。据英伟达 2023 年 5 月的研究数据所示,训练 GPT-3 的 GPU 数量随着模型规模的增长而增加,同时 GPU 的利用效率从 44%提升到了 52%,说明 GPU 的利用率存在较大的限制。因此在大模型算力需求细化到 GPU 数量需求上时,需考虑 GPU 在模型训练时的实际每秒浮点吞吐量。按 44.8%的 GPU 利用率来计算(GPT-3 训练用 A100 的实际利用率),A100 在 FP16 精度下的算力约为 140TFLOPS。

图 9: Megatron 框架下大模型训练规模和算力利用率



资料来源: NVIDIA, 华龙证券研究所

随着模型参数量增加, GPU 实际利用率会相应有所提升,因此以大模型训练周期 60-90 天、GPU 效率 50% 计算,2023 年国内大模型训练和推理一共约需要 31,648-47,472 块 A100 级别芯片。

根据以上结论,结合芯片性能和深度学习时代的算

力需求增速情况,2025 年大模型带来的算力需求估算如下:

芯片性能方面,按摩尔定律所述,芯片算力每 18 个月性能会提升一倍。

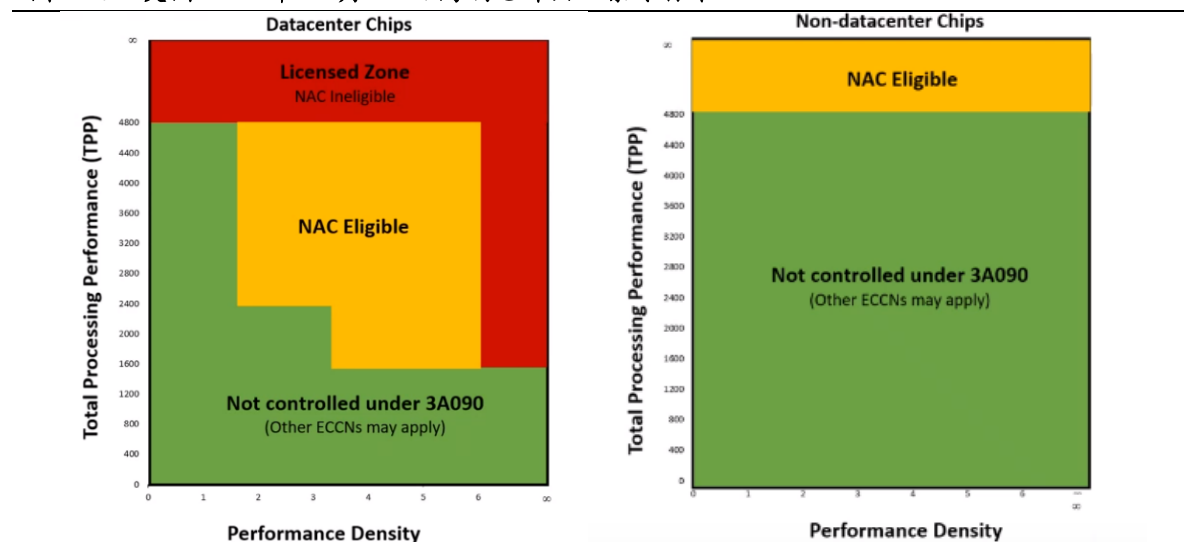
根据 OpenAI 的测算,在深度学习快速发展的 2012 年之后,训练大模型的算力需求约每 3.4 个月翻一倍。近年来,从 2020 年 6 月 GPT-3 发布到 2023 年 3 月 GPT-4 发布,大模型计算量增长约 7 倍。未来大模型计算量增速可能受限于成本和硬件效率,因此估计未来两年,即到 2025 年,训练大模型所需的算力需求增速范围约为 5 倍到 133 倍。对比摩尔定律中芯片算力的增速,训练大模型带来的算力需求增长速度预估远大于算力性能的增长速度。

1.3 AI 算力供给方面: 高端芯片进口受限, 国产替代为大势所趋

1.3.1 国际形势: 美国进一步收紧芯片对华出口标准

2023 年 10 月 18 日,美国发布新禁令提出对高端芯片出口限制标准,从原来对单芯片算力(TTP)的关注向“性能密度阈值”(PD)转移,首次提出对小型高性能芯片的出口限制。意在防范 Chiplet 技术对芯片性能利用率的提升效果。

图 10: 美国 2023 年 10 月 18 日高端芯片出口禁令标准



资料来源: BIS 官网, 华龙证券研究所

1.3.2 AI 芯片国产化替代

智算规模计划稳步提升，以长足发展为目标。根据10月8日，工业和信息化部等六部门联合印发的《算力基础设施高质量发展行动计划》，我国算力发展的主要目标是：到2023年，智算规模达到5.5 E+19 FLOPS。到2025年，算力规模超过300 EFLOPS，智能算力占比达到35%，将超过1.05 E+20 FLOPS。按照计划指标，国内智算供给与实际需求差距较大。

国产AI芯片短期看好华为，长期关注各厂商研发进度。按本次美国禁售芯片的性能标准，市面上主流国产芯片中只有少数能够对标美国禁售的A100/A800等芯片。国产芯片替代化道路还处于起步阶段，距离在大模型训练中大规模使用仍有一段距离。一方面，在芯片IP设计之后，厂商需要根据芯片在大规模生态中应用的实际效果对算子做出调整，不断做出优化以使芯片达到实际应用级别。另一方面，在芯片量产的过程中还需考虑芯片代工厂商的制造工艺、交货周期和定价等等。目前华为发布的昇腾910与英伟达A100/A800性能较为接近，且经过大模型自用和调整，已经具备了大规模商业化应用的条件。截至2023年5月，基于昇腾算力的华为昇腾AI基础软硬件平台已孵化和适配了30多个主流大模型。随着更多国内厂商宣布进入芯片自研领域和发布自研芯片，如百度、腾讯等，未来将持续看好国产芯片领域。

表 5：国内主流 AI 芯片性能参数对比

公司	产品型号	精度	算力	制程	生产商	理论对标英伟达产品
华为	昇腾 910	-	-	-	-	英伟达 A100/A800
	昇腾 310	FP16	8TOPS	12nm	-	
壁仞	BR-100	FP16	>1000TFLOPS	7nm	台积电	
寒武纪	思元 590	-	-	-	-	英伟达 A100/A800
	思元 370	INT8	256TOPS	7nm	-	
百度	昆仑芯 2 代	FP16	128TFLOPS	7nm	-	
	昆仑芯 3 代	-	-	4nm(计划)	-	
阿里	含光 800	INT8	825TOPS	12nm	台积电	
海光	深算二号	-	-	-	-	

资料来源：各公司官网、财联社、中国网、华龙证券研究所

1.3.3 AI 芯片产能：台积电产能复苏伴随订单激增，供不应求情况仍将持续

短期内台积电的芯片制造工艺难有替代。目前台积电依靠 2.5D、3D 等适用于高端芯片的先进封装技术，在芯片制造行业仍然处于垄断地位。其他代工厂商，如三星、格芯等，所占市场份额较少。为提高良率、降低成本、提高芯片制造的精度，目前各芯片 IP 厂商在芯片量产环节广泛依赖台积电。

台积电产能复苏，同时英伟达等大客户订单激增。2023 年 10 月，台积电的产能利用率释放回暖信号，目前 7/6nm 产线利用率从 40%恢复到 60%，到年底预估可以达到 70%。另外，5/4nm 产线利用率为 75-80%。预计台积电明年的 CoWoS（即 2.5D、3D 封装技术）月产能将同比增长 120%。

与此同时，国外大厂商也在大量追加订单。在英伟达 10 月份确定扩大下单后，苹果、超威、博通、迈威尔等重量级客户近期也开始向台积电追单。加上国内四大厂商到 2024 年共计 50 亿美元的芯片订单，台积电大客户订单量全面激增。虽然在台积电在明年计划将 7nm 以下芯片代工定价提高 3-6%，但英伟达、微软等大客户对定价接受度比较高，侧面表现出大厂商对台积电代工的依赖程度比较大。因此短期内订单量仍有保持增速的趋势，且考虑到台积电出货周期拉长以及产能恢复周期的情况，短期内可能出现大量订单积压的现象。

2 AI 算力租赁行业的内在价值

具有大模型训练需求的厂商，按算力付费方式，可分为自建算力的重资产模式和租赁算力的轻资产模式。

2.1 AI 算力租赁对下游公司：带来成本和时间优势

对下游公司来说，在大模型训练方面，自建算力成本过高，且自建算力对设备的运维能力要求很高。这就意味着自建算力的公司除了支付购买硬件设备的高额成本之外还需要支付运维成本，组建运维团队以及付出额外的时间成本。目前市场上购买硬件设备及安装调试的时间过长（1-2 年），过长的设备等待时间会导致大模型训练速度和数量落后于行业整体水平。另一方面，

能够通过自建算力形成规模效应的大厂商较少，小型厂商既有算力需求又难以通过自建算力的方式形成规模效应。这类小型厂商的算力需求使拥有算力资源的公司从业务中分化出算力租赁这一新的业务模式。

2.2 对具备算力资源的公司：算力租赁可为公司带来第二业务增长线

2.2.1 AI 算力租赁成本回收周期测算

算力租赁业务模式近年来在国外头部云厂商中已经得到验证。

国内具备 AI 算力资源的公司可以分为三类。一类是传统云计算服务提供商，如三大运营商、阿里、腾讯等。一类是具备 IDC 建设运营能力的企业，如云赛智能、中科曙光（海光信息）、中贝通信等。以及跨界厂商，如恒润股份、莲花健康等。

国内算力租赁目前的定价方式分为两种，一种按单台设备定价、一种按每 P 算力定价。

国内市场上用于算力租赁的服务器主要有 A100/A800/H800 等型号。本报告中以英伟达 H800 服务器（8 卡）为例测算短期算力租赁成本回收周期。

按恒润股份披露的采购公告，英伟达 H800 服务器的采购价格为 228 万元每台。每台服务器搭载 8 块 GPU，算力约为 16P。按鸿博股份 2023 年披露数据显示，该公司 H 系列服务器刊例价格为 29.9 万/月/台，与其他同行业同类型服务的租金基本持平，或价差浮动不超过 5%。参考阿里云年租定价折扣（年租约 5 折），同时考虑到租金浮动因素，因此市场面上的租赁均价按 12 万元/P/年计算。

国内厂商平均算力租赁成本回收周期按以下公式计算（以 H800 为主要租赁服务器）：

年租赁收入（万元）=服务器算力（P）×年租金（万元/P）

年费用支出（万元）=人工/管理/销售费用（万元）+修理费用（万元）+机柜租赁费用（万元）+宽带费用（万元）

总设备投入成本（万元）=服务器购买合同价格（万元）

算力租赁成本回收周期(月)

$$= 12 \times \frac{\text{总设备投入成本(万元)}}{\text{年租赁收入(万元)} - \text{年费用支出(万元)}}$$

假设厂商出租 H800 服务器的算力租赁利用率为 100%，未来市场上租金价格较为稳定且 H800 服务器的折旧年限为 5 年（无残值）。

则当前以 H800 服务器作为租赁产品的厂商成本回收周期约为 17 个月。实际上，中贝通信算力租赁价格已有上涨趋势。2023 年 11 月，汇纳科技也释放出大幅提价信号，拟将 A100 服务器的算力服务价格提高 100%。预期国内厂商未来的算力租赁定价受强需求影响将有普遍上涨的趋势，整体算力租赁成本回收周期将进一步缩短。

表 6：算力租赁行业成本回收周期测算（以 H800 服务器为租赁产品）

租赁现金流入/流出项	H800
市场价（万元/台）	228
租赁设备投资额（万元）	228
算力（P）	16
租赁价格（万/p/年）	12
租赁收入（万元）	192
人工/管理/销售费用（万元）	22.80
机柜租赁费用（万元）	2.98
宽带费用（万元）	1.20
修理费用（万元）	1.1
总费用（万元）	38.38
成本回收周期（月）	16.70

注：

人工/管理/销售费用：按 10%的投资额计算

机柜租赁费用：按一个机柜（平均 6KW）功率，月租费用为 9,545 元

宽带费用：按约 1000 台服务器所用费用均摊计算

修理费用：按 0.5%的投资额计算

资料来源：华龙证券研究所

3 AI 算力未来发展方向——增质提效

3.1 以云网融合为前提，算力调度成为提高资源配置效率的核心

算力资源紧缺使算力供给加速进入到共享时代。对有算力需求的公司，租用云端算力可以大幅度降低硬件成本，提高对成本的控制能力。另一方面，算力上云也为算力调度奠定了基础。有效的算力调度能够使算力资源能够得到精准管理，分时复用，从而解决算力闲置问题。近两年，我国已着力在算力资源调度方面布局。2022年2月，我国“东数西算”工程正式启动，旨在通过构建数据中心、云计算、大数据一体化的新型算力网络体系，解决东西部算力供需不匹配的问题。2023年6月，全国一体化算力算网调度平台发布，该平台是我国首个实现多元异构算力调度的全国性平台，有望成为“东数西算”项目的有力助益。国内多家上市公司，如中兴通信、浪潮信息、中科曙光、商汤科技、思特奇等以不同的形式参与了该平台的建设。

3.2 芯片数据传输效率：关注 Chiplet 技术和芯片互联技术

3.2.1 Chiplet 技术

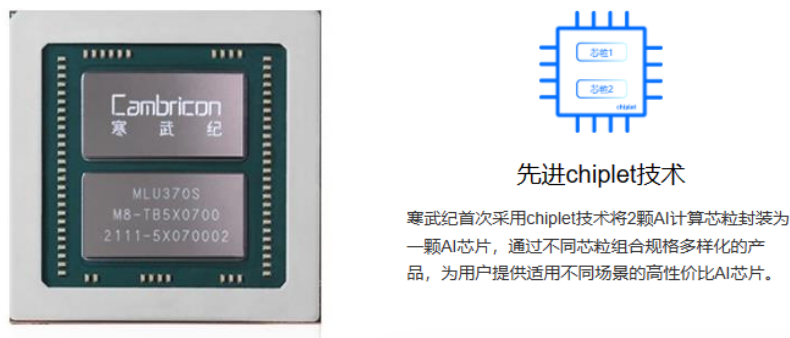
芯片性能提升存在物理极限，摩尔定律或将失效。摩尔定律的核心内容为：集成电路上可以容纳的晶体管数目在大约每经过 18 个月到 24 个月便会增加一倍。换言之，处理器的性能大约每两年翻一倍，同时价格下降为之前的一半。但单集成电路的面积存在物理极限，不能无限容纳晶体管。近年来，摩尔定律中所预言的增速已经明显减慢。随着单芯片面积的缩小，芯片制造的难度和成本也大大提高。根据 AMD 在 IEDM 会议上的资料，若将生产 250 平方毫米的 45nm 芯片的生产成本定为基准 1，14/16nm 芯片的成本将达到 2，而生产 7nm 芯片的成本更将翻倍达到 4。

在此情况下，Chiplet 技术有望降低芯片成本增速，也能够突破单芯物理极限方面持续发挥作用。Chiplet 技术是指通过把不同芯片的能力模块化，利用新的设计、互连接口、封装等技术，在一个封装的产品中使用来自不同技术、不同制程甚至不同工厂的芯片。其优势在于通过缩小单个计算芯粒的面积，提高良率、降低成本、提高算力性能，也可满足定制化需求。

Chiplet 技术在国内已有显著发展成果。下游应用方面，寒武纪在 2021 年就已推出公司首款采用 Chiplet 技术的芯片——思元 370。芯片封测厂商中，长电科技、

通富微电等已宣布掌握 Chiplet 技术。

图 11：寒武纪思元 370 芯片



资料来源：寒武纪官网，华龙证券研究所

3.2.2 芯片互联技术

芯片互联技术是 Chiplet 技术得以存续和发展的底层技术之一。2022 年 3 月 3 日，英特尔、AMD、Arm、高通、台积电、三星、日月光、Google 云、Meta、微软等十大行业巨头联合成立了 Chiplet 标准联盟，正式推出了通用 Chiplet 高速互联标准“Universal Chiplet Interconnect Express”(通用芯粒互连，简称“UCIe”)，UCIe 是 PCIe 的扩展，不但支持 PCIe、CXL，还支持用户定制的 Raw Mode。国内厂商中，芯片 IP 厂商芯原股份、封测厂商长电科技也加入了 UCIe 产业联盟。

总体来说，建立广泛兼容的芯片互联标准，有利于形成良好的上下游生态，提高 Chiplet 的应用范围和技术提升的底层基础。

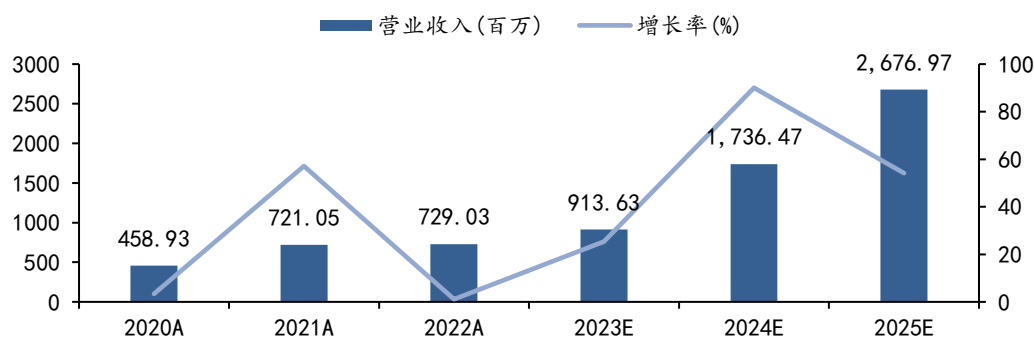
国内芯片互联标准发展进程——构建自有 Chiplet 标准。2023 年 1 月 13 日，中国计算机互连技术联盟（CCITA）发布《小芯片接口总线技术要求》。国内构建符合国内行业状况的自有芯片互联标准有利于加快芯片国产替代化速度，能够有效预防国际技术封锁，也为国产芯片行业从 IP 设计到封测的全产业链提供了价值增长点。

4 重点关注公司

4.1 寒武纪-U（688256.SH）

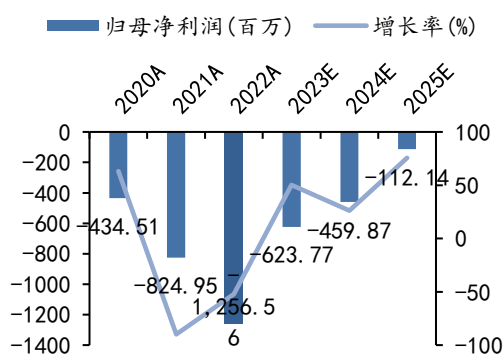
公司推出自研芯片思元系列，有望搭乘 AI 芯片国产替代化的东风。寒武纪现已推出基于思元 290、思元 270 和思元 370 芯片的智能加速卡系列产品，Cambricon-1M 和 Cambricon-1H 终端智能处理器 IP 以及基于思元 220 芯片的智能边缘计算模组。目前产品体系已覆盖云端、边缘端的智能芯片及其加速卡、终端智能处理器 IP，实现云、边、端三大场景的布局。在公司推出的中高端 AI 芯片思元系列中，思元 590 芯片已在百度文心一言大模型中投入使用，在一定程度上对标英伟达 A100 芯片，有望成为除华为昇腾 910/920 芯片外的 A100 国产替代芯片。

图 12：2020-2025 年寒武纪营业收入及增长率



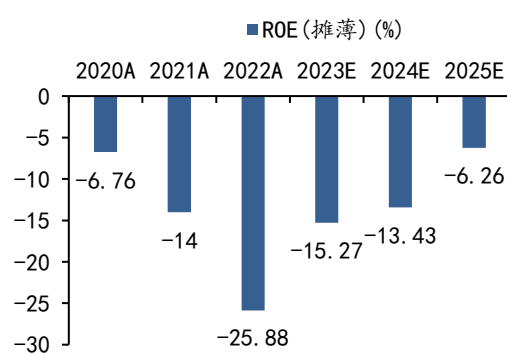
资料来源：一致预测，华龙证券研究所

图 13：2020-2025 年寒武纪归母净利润及增长率



资料来源：一致预测，华龙证券研究所

图 14：2020-2025 年寒武纪 ROE (摊薄)



资料来源：一致预测，华龙证券研究所

4.2 中贝通信 (603220.SH)

算力租赁业务为公司带来第二业务增量，业务利润率有望持续提升。中贝通信在 2023 年 8 月份发布公告，

拟投资建设中贝通信合肥智算中心项目，计划总投资金额约 8.5 亿元。中贝通信合肥智算中心建成后拟提供算力租赁服务，为 AI 企业提供人工智能模型训练和推理等服务。截至 2023 年 11 月 15 日，公司已公告签订的算力服务合同金额总计约合人民币 67,160 万元。中贝通信在算力租赁业务上布局较早，有比较大的先发优势。

2023 年 11 月，中贝通信在关于签订算力服务框架合同的公告中披露，12 个月租期下算力租赁定价为 18 万元/P/年，租赁容量为 1920P，共计提供 128 套高性能算力一体机柜，单机柜提供算力不低于 15P。合同涉及金额共计 34,560 万元。H800 服务器（8 卡 GPU）单台约提供 16P 算力，因此按租赁用算力设备为 H800 服务器测算租赁回收周期。参考 H800 服务器市场价约 228 万元/台，该合同涉及的设备投入总额约为 29,184 万元。

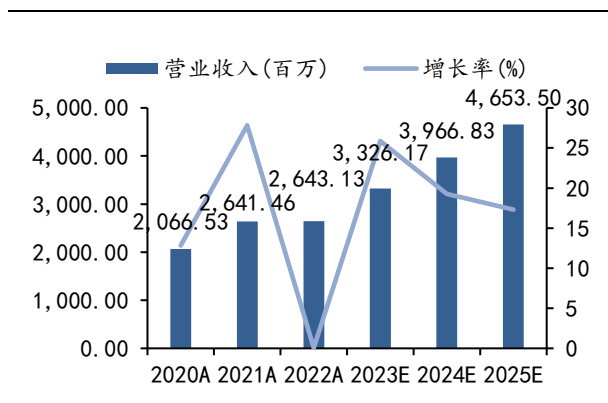
若折旧年限按五年计算，参照前文公式，中贝通信在该合同中的算力租赁业务成本回收周期约为 11 个月，低于按市场定价测算的成本回收周期。且中贝通信的算力租赁业务有涨价趋势，业务利润有进一步增长空间。

表 7：中贝通信算力租赁业务的成本回收期估算

现金流入流出项	租赁设备：H800 服务器
总出租算力（P）	1920
年租金（万元/P/年）	18
年收入（万元/P/年）	34560
成本（万元）	29184
人工/管理/销售费用（万元/年）	2918
修理费用（万元/年）	146
机柜租赁费用（万元/年）	122.18
宽带费用（万元/年）	154
总费用（万元/年）	3340.10
折旧年限（年）	5
回收周期（月）	11.22

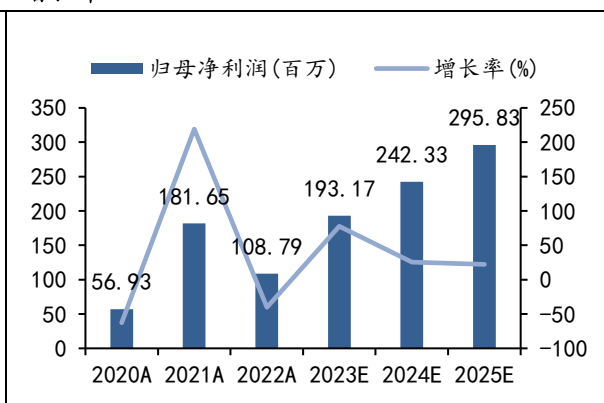
资料来源：，中贝通信公司公告，华龙证券研究所

图 15:2020-2025 年中贝通信营业收入及增长率



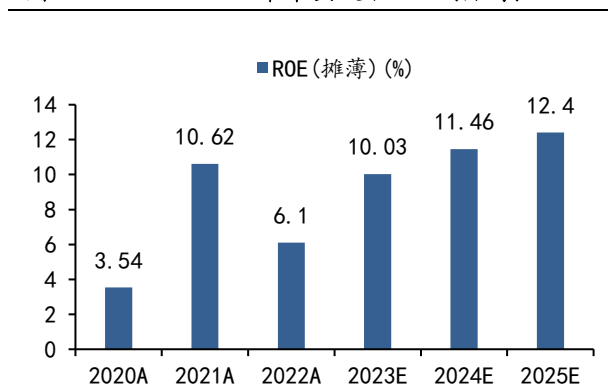
资料来源：一致预测，华龙证券研究所

图 16:2020-2025 年中贝通信营业归母净利润及增长率



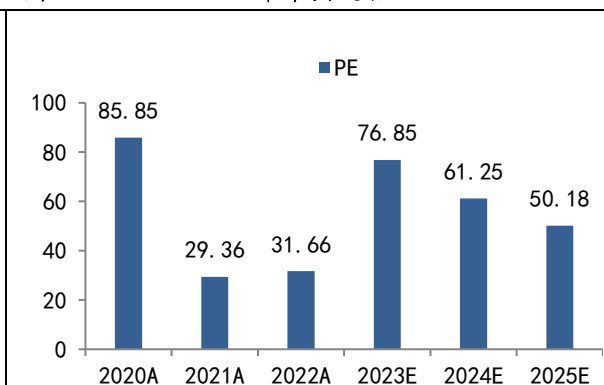
资料来源：一致预测，华龙证券研究所

图 17: 2020-2025 年中贝通信 ROE (摊薄)



资料来源：一致预测，华龙证券研究所

图 18: 2020-2025 年中贝通信 PE



资料来源：一致预测，华龙证券研究所

建议关注：1) 开展算力服务的厂商：中贝通信、莲花健康、恒润股份、汇纳科技、鸿博股份；2) 具备国产芯片 IP 设计能力的厂商：寒武纪、海光信息、芯原股份；3) 具备云服务能力的厂商：中科曙光、浪潮信息、紫光股份、神州数码；4) 芯片封测厂商：长电科技、通富微电、华天科技。

5 风险提示

国际紧张局势加剧（美国可能进一步限制高端芯片对华出口）；

国产大模型发展不及预期(大模型数量和参数规模增长上可能因为技术问题而出现瓶颈期)；

国产芯片技术突破速度不及预期（从芯片 IP 设计

到封测阶段，各项技术突破都需要较长时间）；

台积电产能恢复不及预期（台积电产能有限，芯片供给可能出现空窗期）；

数据估算风险（预测数据可能存在误差的风险）。

免责及评级说明部分

分析师声明：

本人具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，以勤勉尽责的职业态度，独立、客观、公正地出具本报告。不受本公司相关业务部门、证券发行人士、上市公司、基金管理公司、资产管理公司等利益相关者的干涉和影响。本报告清晰准确地反映了本人的研究观点。本人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。据此入市，风险自担。

投资评级说明：

投资建议的评级标准	类别	评级	说明
报告中投资建议所涉及的评级分为股票评级和行业评级（另有说明的除外）。评级标准为报告发布日后的6-12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。其中：A股市场以沪深300指数为基准。	股票评级	买入	股票价格变动相对沪深300指数涨幅在10%以上
		增持	股票价格变动相对沪深300指数涨幅在5%至10%之间
		中性	股票价格变动相对沪深300指数涨跌幅在-5%至5%之间
		减持	股票价格变动相对沪深300指数跌幅在-10%至-5%之间
		卖出	股票价格变动相对沪深300指数跌幅在-10%以上
	行业评级	推荐	基本面向好，行业指数领先沪深300指数
		中性	基本面稳定，行业指数跟随沪深300指数
		回避	基本面向淡，行业指数落后沪深300指数

免责声明：

华龙证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因为任何机构或个人接收到报告而视其为当然客户。

本报告信息均来源于公开资料，本公司对这些信息的准确性和完整性不作任何保证。编制及撰写本报告的所有分析师或研究人员在此保证，本研究报告中任何关于宏观经济、产业行业、上市公司投资价值等研究对象的观点均如实反映研究分析人员的个人观点。报告中的内容和意见仅供参考，并不构成对所述证券买卖的价格的建议或询价。本公司及分析研究人员对使用本报告及其内容所引发的任何直接或间接损失及其他影响概不负责。

在法律许可的情况下，本公司及所属关联机构可能会持有报告中提及的公司所发行的证券并进行证券交易，也可能为这些公司提供或正在争取提供投资银行、财务顾问或金融产品等相关服务，投资者应充分考虑本公司及所属关联机构就报告内容可能存在的利益冲突。

版权声明：

本报告版权归华龙证券股份有限公司所有，未经书面许可任何机构和个人不得以任何形式翻版、复制、刊登。任何人使用本报告，视为同意以上声明。引用本报告必须注明出处“华龙证券”，且不能对本报告作出有悖本意的删除或修改。

华龙证券研究所

北京	兰州	上海
地址：北京市东城区安定门外大街189号天鸿宝景大厦F1层华龙证券	地址：兰州市城关区东岗西路638号甘肃文化大厦21楼	地址：上海市浦东新区浦东大道720号11楼
邮编：100033	邮编：730030	邮编：200000
	电话：0931-4635761	