证券研究报告/行业研究

报告日期：2024 年 03 月 26 日

Kimi 智能助手热度高涨，国产大模型加速发展

——人工智能专题研究系列五

投资要点

➢ Kimi智能助手爆火，AI大模型商业化可期

2024年3月18日月之暗面公告Kimi智能助手的上下文窗口已从20万汉字拓展至200万汉字，更长的上下文窗口意味着大模型可以更精准细微的捕捉用户需求。我们认为长上下文窗口将有助于大模型类智能产品的商业化落地，目前Kimi已在智能搜索、高效阅读、资料整理、辅助创作和编程助力等方面具有国际先进表现，数据显示2024年2月Kimi月访问量居于国内第三，最新周访问量达225万次，国产大模型产品商业化加速推进。

➢ 百度/阿里发布大模型新品，国产大模型进展加速

百度智能云基于千帆大模型平台发布多款新品，包括文心大模型、轻量大模型和垂直场景模型等多个系列。3月22日阿里通义千问宣布免费开放1000万字的长文档处理功能。我们认为Kimi热度出圈及文心一言和通义千问的新品迭代有望激发国内大模型行业的竞争意识，加快国内其他大模型产品的更新迭代速度，长期看将利好国内AI产业发展。从中短期来看，算力仍是发展AI产业的首要基础，建议关注AI算力产业链的相关机遇。

➢ AI行业加速发展，关注算力产业链机会

算力是大模型行业发展的基础，大模型的训练参数量可高达千亿级，背后是大规模数据中心的支持，发展人工智能行业需要先发展算力基础设施。服务器、光模块和芯片是算力基础设施中的核心器件，预计算力基础设施的建设将极大拉动上述器件的需求，其中光模块和服务器国内公司的份额较高，在产品迭代基础上量价齐升逻辑明确。而芯片目前国产化率仍较低，具备较大国产化空间，卡脖子风险下自主可控紧迫度提升，相关制造产业链值得关注。

➢ 投资建议

1）服务器：算力缺口是明确发展机遇，建议关注：浪潮信息、中科曙光等；

2）光模块：AI产业拉动高速光模块需求，建议关注：中际旭创、新易盛等；

3）光芯片：光芯片决定光通信效率和稳定性，25G及以上光芯片国产化率仍有较大提升空间，建议关注：源杰科技等。

➢ 风险提示

AI技术发展不及预期，AI商业化推进不及预期，中美贸易摩擦加剧，国产化及市场开拓不及预期。

投资评级：看好

分析师：吴起涤

执业登记编号：A0190523020001

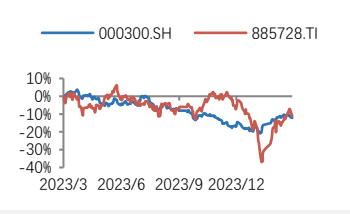
wuqidi@yd.com.cn

研究助理：程治

执业登记编号：A0190123070008

chengzhi@yd.com.cn

人工智能指数与沪深 300 指数走势对比



资料来源：同花顺 iFinD，源达信息证券研究所

相关报告：

1.《人工智能专题研究系列一：大模型推动各行业AI应用渗透》2023.08.02

2.《人工智能专题研究系列二：AI大模型开展算力竞赛，推动AI基础设施建设》2023.08.03

3.《人工智能专题研究系列三：Gemini 1.0有望拉动新一轮AI产业革新，算力产业链受益确定性强》2023.12.12

4.《人工智能专题研究系列四：OpenAI发布Sora文生视频模型，AI行业持续高速发展》2024.02.19

目录

图表目录

一、事件：Kimi 智能助手爆火，AI 产业商业化有望加快

月之暗面（Moonshot AI)创立于 2023 年 3 月，团队核心成员曾参与 Google Gemini、Google Bard、盘古 NLP 和悟道等多个大模型的研发，多项核心技术被 Google PaLM、Meta LLaMa、Stable Diffusion 等主流产品采用。**2023 年 10 月月之暗面推出全球首个支持输入 20 万汉字的智能助手产品 Kimi。**

图 1：月之暗面推出智能助手产品 Kimi

资料来源：月之暗面官网，源达信息证券研究所

根据月之暗面近期公告，Kimi 智能助手的上下文窗口已从 20 万汉字拓展至 200 万汉字。上下文窗口即界面输入窗口和输出窗口的长度总和，更长窗口意味着大模型可以更精准细微的捕捉用户需求，提供更好服务。我们认为长上下文窗口将有助于大模型类智能产品的落地，目前 Kimi 已在智能搜索、高效阅读、资料整理、辅助创作和编程助手等方面具有国际先进表现，为用户提供高效服务。**Kimi 智能助手有望推动国产大模型产品的商业化进程。**

图 2：Kimi 具备论文阅读和总结能力

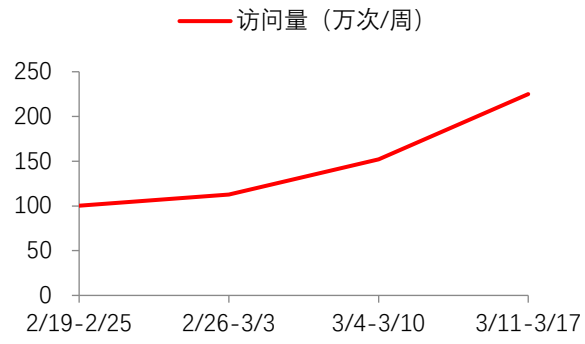
资料来源：Kimi 官网，源达信息证券研究所

图 3：Kimi 可以完成代码问题的解答

资料来源：Kimi 官网，源达信息证券研究所

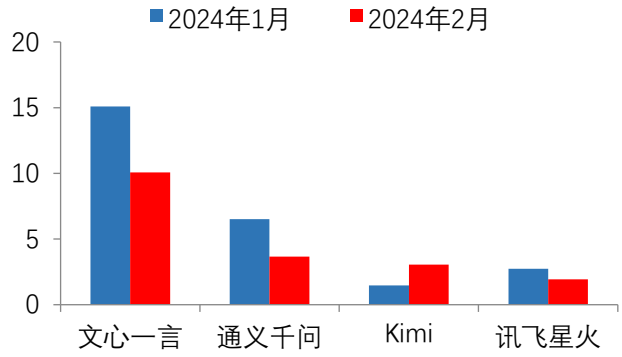
Kimi 自 2023 年 10 月发布后，经过多次迭代及优化，可用性持续增强。根据 Similarweb 数据，2024 年 2 月以来 Kimi 周访问量持续攀升，截至 2024 年 3 月 11 日-17 日，周访问量已达 225 万次。根据 AI 产品榜，2024 年 2 月 Kimi 月访问量已在国内大模型中位居第三，仅次于文心一言和通义千问。

图 4：Kimi 访问量逐周攀升



资料来源：Similarweb，源达信息证券研究所

图 5：国内大模型产品月访问量情况（百万次）



资料来源：AI 产品榜，源达信息证券研究所

百度智能云基于千帆大模型平台发布多款新品。产品包括适用于通用复杂场景的文心大模型：ERNIE3.5、ERNIE4.0；轻量大模型：ERNIE Speed、ERNIE Lite 和 ERNIE Tiny；垂直场景模型：ERNIE Character、ERNIE Functions。

3 月 22 日阿里通义千问大模型宣布向所有人免费开放 1000 万字的长文档处理功能，成为目前全球文档处理容量最大的 AI 应用。

图 6：百度智能云推出多款大模型新品

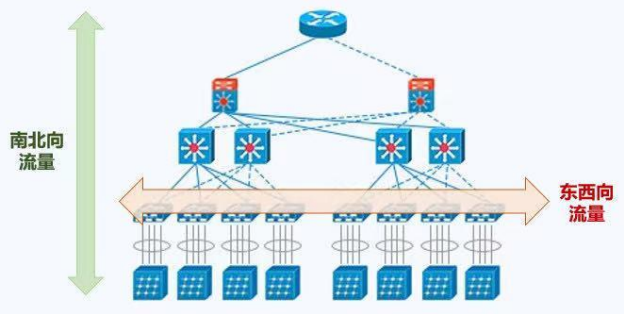
资料来源：百度智能云官网，源达信息证券研究所

Kimi 智能助手的爆火有望推动国内大模型行业发展。Kimi 智能助手的爆火及在上下文长度下的突破，有望推动大模型产品的商业化落地及国内 AI 产业发展。我们认为 Kimi 有望激发国内大模型行业的竞争意识，加快文心一言和通义千问等大模型产品的更新迭代速度，长期看将利好国内 AI 产业发展。从中短期来看，算力仍是发展 AI 产业的首要基础，建议关注 AI 算力产业链的相关机遇。

二、算力产业链：高速光模块加快放量，国产厂商有望充分受益

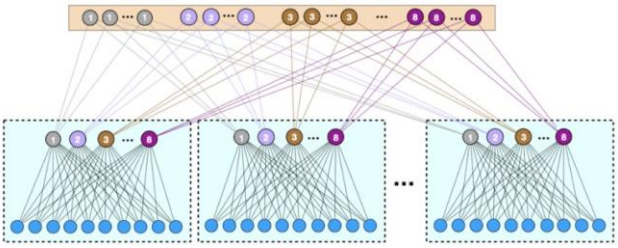
高算力需要与高效传输架构相匹配。AI 大模型通常由多个服务器作为节点，并通过高速网络架构组成集群合作完成模型训练。因此在模型中东西向流量（数据中心服务器间的传输流量）大幅增加，而模型训练过程中南北向流量（客户端与服务器间的传输流量）较少，由于叶脊网络架构相较传统三层架构更适用于东西向流量传输，成为现代数据中心主流网络架构。

图 7：AI 大模型中东西向流量显著增加



资料来源：华为云，源达信息证券研究所

图 8：叶脊网络架构适用于东西向流量传输



资料来源：鹅厂网事，源达信息证券研究所

叶脊网络架构大幅增加对光模块数量需求。由于叶脊网络架构中东西向流量大，因此服务器与交换机相连均需使用光模块，从而大幅增加对光模块数量需求。同时 AI 大模型的高流量对带宽提出更高要求，800G 光模块相较 200G/400G 光模块具有高带宽、功耗低等优点，有望在 AI 大模型网络架构中渗透率提升。

表 1：叶脊网络架构对光模块数量需求大幅提升

架构类型	传统三层架构	改进等三层架构	叶脊网络架构
光模块相对于机柜倍数	8.8	9.2	44/48

资料来源：中际旭创定向增发募集说明书，源达信息证券研究所

我们以 Nvidia DGX H100 网络架构为例。该架构适配 Nvidia H100 GPU，采用叶脊网络架构，分为 1-4 个 SU 单元类型（8 个 GPU 组成一个 H100 服务器节点，32 个服务器节点组成一个 SU 单元）。其中 4-SU 单元架构由 127 个服务器节点组成（其中一个节点用于安装 UFM 网络遥测装置），具有 1016 个 H100 GPU、32 个叶交换机、16 个脊交换机。

表 2：Nvidia DGX H100 架构所需 GPU、交换机数量

SU Count	Cluster Size # Nodes	Cluster Size # GPUs	Leaf Switch Count	Spine Switch Count	Compute + UFM Node Cable Count	Spine-Leaf Cable Count
1	31[1]	248	8	4	252	256
2	63	504	16	8	508	512
3	95	760	24	16	764	768
4	127	1016	32	16	1020	1024

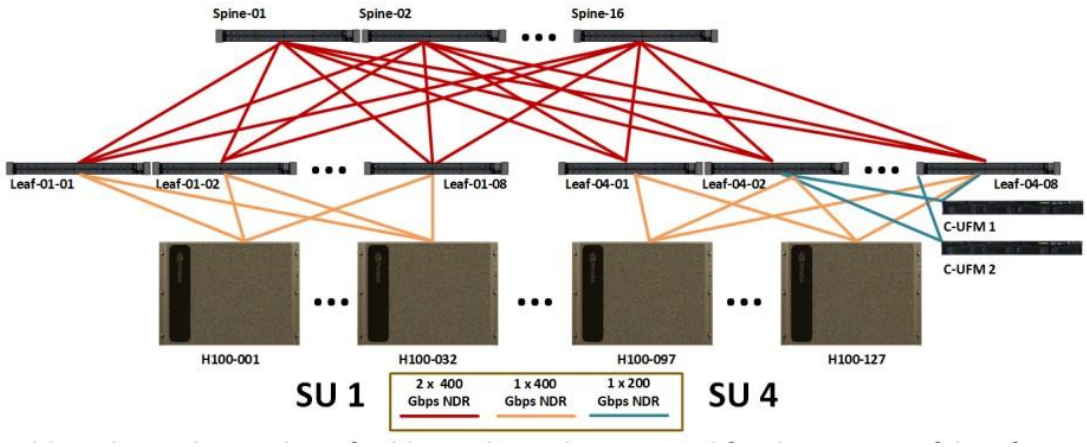
1. This is a 32 node per SU design, however a DGX Node must be removed to accommodate for UFM connectivity.

资料来源：Nvidia，源达信息证券研究所

我们以 Nvidia DGX H100 架构为例测算 GPU 与光模块的对应数量。在 4-SU 的 Nvidia DGX H100 架构中，每 32 台服务器节点组成一个 SU 单元，并与 8 台叶交换机相连，因此服务器节点与叶交换机之间共有 1024 个连接（32×8×4）；32 台叶交换机需分别与 16 台脊交换机相连，因此叶交换机与脊交换机之间共有 512 个连接（32×16）；

在 Nvidia DGX H100 的目前方案中，脊-叶连接采用 800G 光模块，需要 1024 个 800G 光模块；叶-服务器连接中，每个服务器节点通过一个 800G 光模块与两台叶交换机向上连接，需要 512 个 800G 光模块（128×4），同时每台叶交换机通过一个 400G 光模块与一个服务器节点连接，需要 1024 个 400G 光模块（128×8）。综上计算得一个 4-SU 单元的 DGX H100 架构需要 1016 个 GPU、1536 个 800G 光模块、1024 个 400G 光模块，**GPU：800G 光模块：400G 光模块的比例约等于 1：1.5：1。**

图 9：Nvidia DGX H100 架构示意图



资料来源：Nvidia，源达信息证券研究所

国产光模块厂商在 2022 年全球光模块企业 TOP10 排名中占据 7 席。TOP10 中国内企业为中际旭创（Innolight）、华为（Huawei）、光迅科技（Accelink）、海信（Hisense）、

新易盛（Eoptolink）、华工正源（HGG）、索尔思光电（已被华西股份收购）。而在高端光模块领域，中际旭创已在 2022 年实现 800G 光模块批量出货。

表 3：中际旭创在 2022 年全球光模块企业排名中位居第一

2018	2021	2022
Finisar Innolight	Ⅱ-Ⅳ&Innolight	Innolight&Coherent
Hisense	Huawei（HiSilicon）	Cisco(Acacia)
Accelink	Cisco（Acacia）	Huawei（HiSilicon）
FOIT（Avago）	Hisense	Accelink
Lumentum/Oclaro	Broadcom（Avago）	Hisense
Acacia	Eoptolink	Eoptolink
Intel	Accelink	HGG
Aoi	Molex	Intel
Sumitomo	Intel	Source Photonics

资料来源：Light counting，源达信息证券研究所

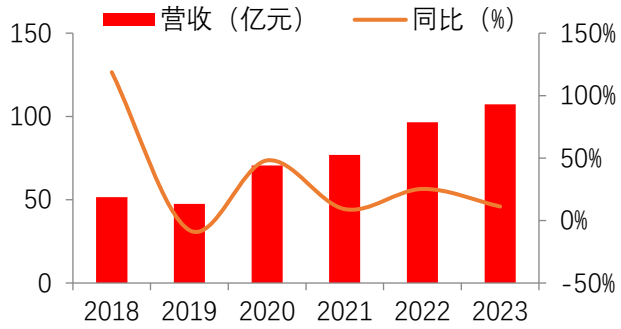
三、建议关注

1.中际旭创

中际旭创是全球光模块行业的领军企业，并持续拓展高速光模块产品。公司于 2020 年推出首个 800G 光模块产品，并积极布局 CPO 和 LPO 等新技术。在 AI 算力建设加速情况下，公司已成为北美重点客户的高速光模块的长期供货商。

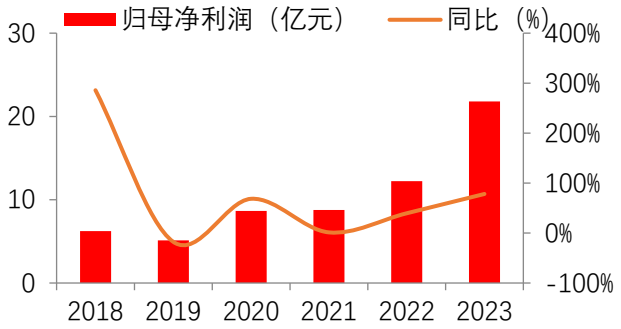
2024 年 2 月 29 日公司发布 2023 年业绩快报，2023 年全年营收在 107.25 亿元，同比增长 11.23%；归母净利润在 21.81 亿元，同比增长 78.19%。算力需求激增带动 800G 等高速光模块需求增长，公司受益 800G 等高端产品出货增长及产品结构优化，业绩保持高速增长，盈利能力持续提升。

图 10：中际旭创 2018-2023 年营收情况



资料来源：Wind，源达信息证券研究所

图 11：中际旭创 2018-2023 年归母净利润情况



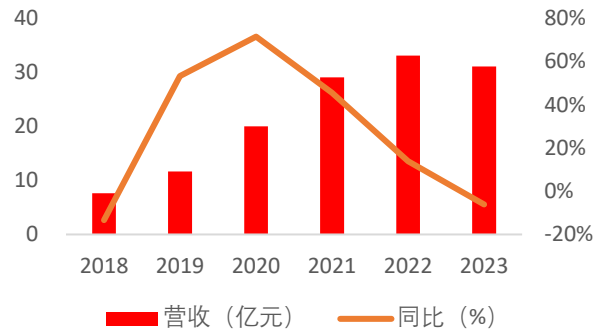
资料来源：Wind，源达信息证券研究所

2.新易盛

公司是全球光模块行业的领先企业，2022 年在全球光模块企业 TOP10 中排名第 7。公司积极布局前沿产品，2019 年即实现 400G 光模块的批量出货，2020 年推出 800G 光模块样品，并已具有基于硅光基数的 400G、800G 光模块和产品和基于 LPO 技术的 800G 光模块产品，有望充分受益算力建设浪潮。

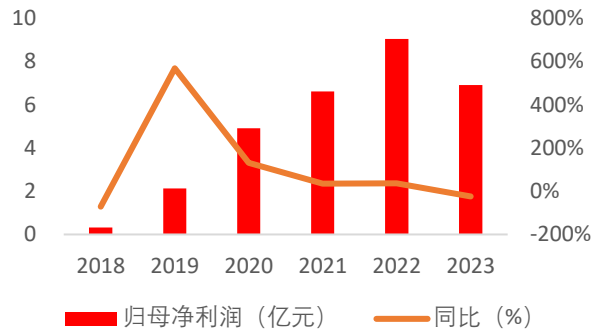
2024 年 2 月 29 日公司发布 2023 年业绩快报，2023 年全年公司实现营收 31.08 亿元，同比下降 6.13%；归母净利润 6.91 亿元，同比下降 23.57%。业绩出现下滑系传统电信行业的需求不佳所致，期待公司后续高速光模块产品的放量及相关订单兑现。

图 12：新易盛 2018-2023 年营收情况



资料来源：Wind，源达信息证券研究所

图 13：新易盛 2018-2023 年归母净利润情况



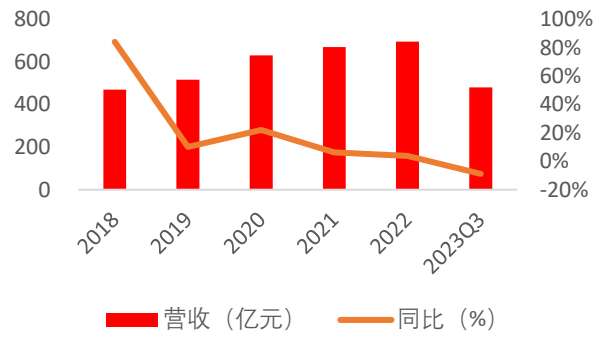
资料来源：Wind，源达信息证券研究所

3.浪潮信息

公司是全球 AI 服务器行业的领军企业。根据 IDC 数据，公司在 2023 年第三季度全球服务器出货量和销量份额中均位居第二，市场份额分别为 10.3%和 9.1%。公司持续加大算力基础设施的研发投入和相关布局。2023 年 1 月 27 日公司发布“源 2.0”基础大模型，并宣布正式开源。该模型作为千亿级基础大模型，并可分为 1026、518、21 亿三种参数规模的版本。

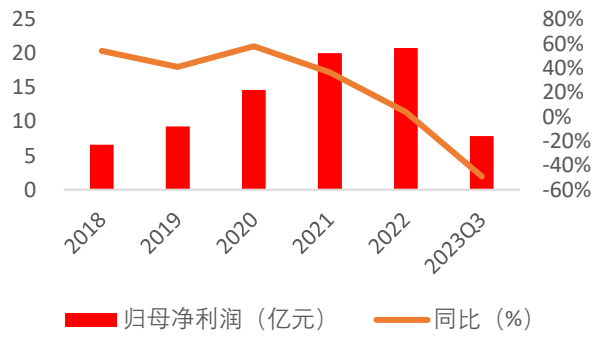
截至 2023 年前三季度，公司实现营收 480.96 亿元，同比下降 8.85%；归母净利润 7.87 亿元，同比下降 49.12%。单三季度公司实现营收 232.99 亿元，同比增长 30.04%，归母净利润 4.61 亿元，同比下降 22.06%。业绩下滑系 2023 年上半年传统服务器行业需求有节奏性放缓，而第三季度公司营收已恢复高速增长，盈利能力下滑系公司前三季度出货中互联网客户占比增加，客户特性导致毛利率有一定降低，后续客户结构优化有望带动公司盈利能力的修复。

图 14：浪潮信息 2018-2023 年前三季度营收情况



资料来源：Wind，源达信息证券研究所

图 15：浪潮信息 2018-2023 年前三季度归母净利润情况



资料来源：Wind，源达信息证券研究所

四、投资建议

Kimi 智能助手的爆火及在上下文长度下的突破，有望推动大模型产品的商业化落地及国内 AI 产业发展。我们认为 Kimi 有望激发国内大模型行业的竞争意识，加快文心一言和通义千问等大模型产品的更新迭代速度，长期看将利好国内 AI 产业发展。从中短期来看，算力仍是发展 AI 产业的首要基础，建议关注 AI 算力产业链的相关机遇：

1）服务器：服务器是算力基础设施，AIGC 行业的快速发展将产生持续大额的算力缺口，拉动服务器需求，建议关注：浪潮信息、中科曙光等；

2）光通信模块：AI 时代的网络架构对 400G/800G 等高端光模块用量有望大幅提升，目前国产光模块企业在全球市场中已占据一定市场地位，建议关注：中际旭创、新易盛等；

3）光芯片：光芯片的性能决定光通信效率和稳定性，25G 及以上光芯片国产化率仍有较大提升空间，出于保供考虑光芯片国产化率有望提升，建议关注：源杰科技等。

五、风险提示

AI 技术发展不及预期；

AI 商业化推进不及预期；

中美贸易摩擦加剧；

国产化及市场开拓不及预期。

投资评级说明

行业评级	以报告日后的 6 个月内，证券相对于沪深 300 指数的涨跌幅为标准，投资建议的评级标准为：		
	看	好：	行业指数相对于沪深 300 指数表现＋10%以上
	中	性：	行业指数相对于沪深 300 指数表现 - 10%~＋10%以上
	看	淡：	行业指数相对于沪深 300 指数表现 - 10%以下
公司评级	以报告日后的 6 个月内，行业指数相对于沪深 300 指数的涨跌幅为标准，投资建议的评级标准为：		
	买	入：	相对于恒生沪深 300 指数表现＋20%以上
	增	持：	相对于沪深 300 指数表现＋10%~＋20%
	中	性：	相对于沪深 300 指数表现 - 10%~＋10%之间波动
	减	持：	相对于沪深 300 指数表现 - 10%以下

办公地址

石家庄

河北省石家庄市长安区跃进路 167 号源达办公楼

上海

上海市浦东新区民生路 1199 弄证大五道口广场 1 号楼 2306C 室

分析师声明

作者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，以勤勉的职业态度，独立、客观地出具本报告。分析逻辑基于作者的职业理解，本报告清晰准确地反映了作者的研究观点。作者所得报酬的任何部分不曾与，不与，也不将与本报告中的具体推荐意见或观点而有直接或间接联系，特此声明。

重要声明

河北源达信息技术股份有限公司具有证券投资咨询业务资格，经营证券业务许可证编号：911301001043661976。

本报告仅限中国大陆地区发行，仅供河北源达信息技术股份有限公司（以下简称：本公司）的客户使用。本公司不会因接收人收到本报告而视其为客户。本报告的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，也不保证所包含信息和建议不发生任何变更。本公司已力求报告内容的客观、公正，但文中的观点、结论和建议仅供参考，不包含作者对证券价格涨跌或市场走势的确定性判断。本报告中的信息或所表述的意见均不构成对任何人的投资建议，投资者应当对本报告中的信息和意见进行独立评估。

本报告仅反映本公司于发布报告当日的判断，在不同时期，本公司可以发出其他与本报告所载信息不一致及有不同结论的报告；本报告所反映研究人员的不同观点、见解及分析方法，并不代表本公司或其他附属机构的立场。同时，本公司对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

本公司及作者在自身所知情范围内，与本报告中所评价或推荐的证券不存在法律法规要求披露或采取限制、静默措施的利益冲突。