

AI 系列报告之接口：AI 大算力需求加速新技术渗透

核心观点

- **AI 模型数据量和算力需求大。**ChatGPT3 相较 ChatGPT2 在数据存储端从百 G 提升至 40T，在存储量上有约 100 倍的提升，算力需求同样也呈几何倍增长。据 OpenAI 测算，2012 年以来全球头部 AI 模型训练算力需求 3-4 个月翻一番，每年头部训练模型所需算力增长幅度高达 10 倍，远超摩尔定律的增长速度。美光表示，一个典型的人工智能服务器的 DRAM 容量是普通服务器的 8 倍，NAND 容量是普通服务器的 3 倍。
- **AI 服务器快速增长，内存技术同步升级。**根据 TrendForce，预计 2022 年搭载 GPGPU 的 AI 服务器年出货量占整体服务器比重近 1%，而 2023 年 ChatGPT 相关应用有望再度刺激 AI 相关领域，预计 2023 年出货量增长率可达 8%，2022-2026 年复合成长率将达 10.8%。为提升服务器性能，AI 服务器需搭载多个 CPU 处理器，同时服务器 CPU 性能不断升级，要求内存技术同步升级。DDR5、HBM、CXL、NVLink 等内存技术将加速渗透，有望充分受益于 AI 带来的算力需求增长。
- **DDR5 需要使用更多内存接口芯片。**DDR5 的内存接口缓存芯片 RCD 价格远高于 DDR4，内存接口芯片有望迎来量价齐升。DDR5 芯片需要搭载更多的 SPD、电源管理芯片（PMIC）和温度传感器（TS）等配套芯片。
- **HBM 为大算力芯片提供支撑。**如今搭载于新兴 AI 应用的内存芯片亟待升级，而 HBM 是一种基于 3D 堆叠工艺的 DRAM 内存芯片，被安装在 GPU、网络交换设备、AI 加速器及高效能服务器上。HBM 作为一种带宽远超 DDR/GDDR 的高速内存，将为大算力芯片提供能力支撑，同时生成类模型也会加速 HBM 内存进一步增大容量和增大带宽。TrendForce 预估 2023-2025 年 HBM 市场年复合成长率有望成长至 40-45%以上。
- **CXL 兼容性强、可实现内存一致性。**CXL 为英特尔于 2019 年推出的一种开放性互联协议，能够让 CPU 与 GPU、FPGA 或其他加速器之间实现高速高效的互联，从而满足高性能异构计算的要求。在 AMD、ARM、IBM 以及英特尔等主要 CPU 供应商的支持下，CXL 已经成为领先的行业标准。美光科技在 22 年 5 月与投资者交流时曾预测 CXL 相关产品的市场规模，到 2025 年预计将达到 20 亿美金，到 2030 年可能超过 200 亿美金。
- **NVLink 可最大化提升系统吞吐量。**第四代 NVIDIA® NVLink®技术可为多 GPU 系统配置提供高于以往 1.5 倍的带宽，以及增强的可扩展性。单个 NVIDIA H100 Tensor Core GPU 支持多达 18 个 NVLink 连接，总带宽为 900GB/s，是 PCIe 5.0 带宽的 7 倍。第三代 NVIDIA NVSwitch™基于 NVLink 的高级通信能力构建，可为计算密集型工作负载提供更高带宽和更低延迟。NVIDIA DGX™ H100 等服务器可利用 NVLink 技术来提高可扩展性，进而实现超快速的深度学习训练。

投资建议与投资标的

- 我们看好服务器带来的接口芯片技术升级趋势，建议关注澜起科技、聚辰股份、裕太微、龙迅股份。

风险提示

- 服务器需求拉动不及预期、新技术渗透率不及预期。

行业评级

看好（维持）

国家/地区

中国

行业

电子行业

报告发布日期

2023 年 04 月 18 日



证券分析师

蒯剑

021-63325888*8514

kuaijian@orientsec.com.cn

执业证书编号：S0860514050005

香港证监会牌照：BPT856

李庭旭

litingxu@orientsec.com.cn

执业证书编号：S0860522090002

联系人

杨宇轩

yangyuxuan@orientsec.com.cn

韩潇锐

hanxiaorui@orientsec.com.cn

张释文

zhangshiwen@orientsec.com.cn

薛宏伟

xuehongwei@orientsec.com.cn

相关报告

大模型增加存储需求，行业景气底部反暖 2023-04-17

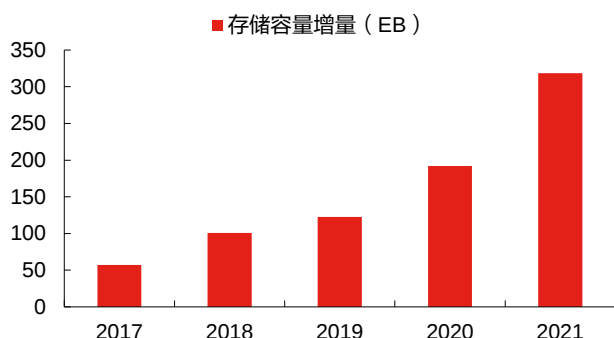
AI 系列报告之 PCB：AI 算力需求增长，服务器 PCB 厂商受益 2023-04-07

算力需求提升，光芯片正扬帆 2023-03-26

AI 大算力需求加速新技术渗透

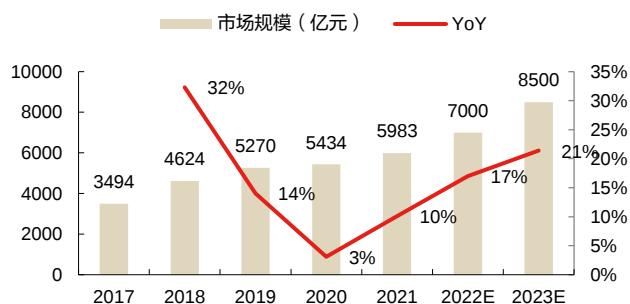
AI 模型数据量和算力需求大。ChatGPT3 相较 ChatGPT2 在数据存储端从百 G 提升至 40T，在存储量上有约 100 倍的提升，算力需求同样也呈几何倍增长。据 Open AI 测算，2012 年以来全球头部 AI 模型训练算力需求 3-4 个月翻一番，每年头部训练模型所需算力增长幅度高达 10 倍，远超摩尔定律的增长速度。随着算力的不断进步，所需存储的数据量在以指数级的增长速度攀升。得益于人工智能、物联网、云计算、边缘计算等新兴技术在中国的快速发展，中国数据圈正在迎来快速增长。据此前 IDC 预测，预计到 2025 年，中国数据圈将增长至 48.6ZB，占全球数据圈的 27.8%，成为全球最大的数据圈。美光首席商务官萨达纳表示，一个典型的人工智能服务器的 DRAM 容量是普通服务器的 8 倍，NAND 容量是普通服务器的 3 倍。

图 1：2017-2021 我国存储容量增量（EB）



数据来源：中国信通院、东方证券研究所

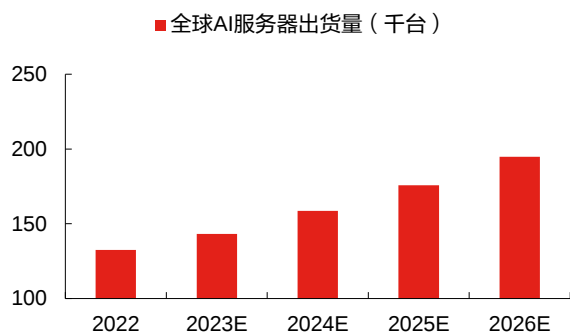
图 2：2017-2023E 中国数据存储市场规模及增速



数据来源：观研报告网、东方证券研究所

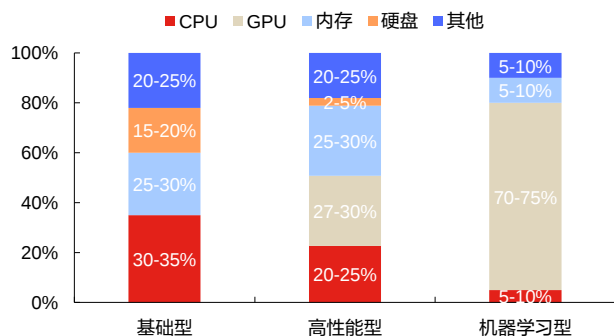
AI 服务器快速增长，内存技术同步升级。根据 TrendForce，在自动驾驶汽车、AIoT 与边缘运算等新兴应用的带领下，自 2018 年起诸多大型云端业者开始大量投入 AI 相关的设备建设，预计 2022 年搭载 GPGPU 的 AI 服务器年出货量占整体服务器比重近 1%，而 2023 年 ChatGPT 相关应用有望再度刺激 AI 相关领域，预计 2023 年出货量增长率可达 8%，2022-2026 年复合成长率将达 10.8%。内存接口芯片是高端服务器中 CPU 与内存之间进行数据传输的核心部件，其主要作用是提升高速内存数据访问的速度及稳定性，满足服务器 CPU 对内存模组日益增长的高性能及大容量需求。为提升服务器性能，AI 服务器需搭载多个 CPU 和 GPU 处理器，同时服务器 CPU 性能不断升级，要求内存技术同步升级。DDR5、HBM、CXL、NVLink 等内存技术将加速渗透，有望充分受益于 AI 带来的算力需求增长。

图 3：2022-2023E 全球 AI 服务器出货量（千台）



数据来源：TrendForce、东方证券研究所

图 4：不同类型服务器成本构成



数据来源：行研查研究中心、东方证券研究所

表 1：服务器型号对比

	浪潮普通服务器	浪潮 AI 服务器	英伟达 AI 服务器	英伟达 AI 服务器
服务器型号	NF5280M6	NF5688M6	DGX A100	DGX H100
CPU 型号	英特尔第三代 Xeon 处理器	英特尔第三代 Xeon 处理器	AMD 第二代 Epyc 处理器 Rome 7742	英特尔第四代 Xeon 处理器 Sapphire Rapids
CPU 数量	1-2	2	2	2
CPU DDR	最大支持 32 根内存条，支持 RDIMM 类型	支持 32 条 DDR4 RDIMM/LRDIMM 内存	2TB	2TB
GPU 型号	/	NVIDIA A800	A100 Tensor Core	H100 Tensor Core
GPU 数量	/	8	8	8
GPU 内存	/	每个对应 80GB HBM2e	每个对应 80GB HBM2e	每个对应 80GB HBM2e
硬盘 NAND	最大支持 20 块 3.5 英寸硬盘或 39 块 2.5 英寸硬盘	8 块 2.5 英寸 NVMe SSD/16 块 2.5 英寸 SATA/SAS SSD	操作系统：2 个 1.92TB NVMe M.2 驱动器 内部存储：8 个 3.84TB NVMe U.2 驱动器	操作系统：2 块 1.9TB NVMe M.2 硬盘 内部存储：8 块 3.84TB NVMe U.2 硬盘

数据来源：浪潮信息官网、英伟达官网、东方证券研究所

DDR5 对比 DDR4 性能全面提升，需要使用更多内存接口芯片。1) 性能：DDR5 内存频率从 3200MHz 起跳（一般都会在 4800MHz 以上）、最高可达 8400MHz；2) 核心容量密度：SK 海力士给出 8Gb、16Gb、24Gb、32Gb、64Gb 丰富的选择，DDR5 单条内存最大可到 128GB；3) 功耗：DDR5 支持 1.1V 电压，进一步降低了功耗。另外，DDR5 搭载 DQS 间隔振荡器，具备错误检查与清理能力，具备更强的抵抗环境变化的能力和更强的可靠性。DDR5 性能全面提升，要求新一代内存接口芯片支持高速、低功耗、系统扩展等要求。DDR5 模组需要搭载 SPD、电源管理芯片（PMIC）和温度传感器（TS）等配套芯片。此外，在 LRDIMM 模组中，相比 DDR3 只采用 1 颗寄存缓存芯片、DDR4 最多采用“1+9”个内存芯片，DDR5 对内存接口的需求进一步提升，最多达到“1+10”个（1RCD+10DB），内存接口芯片有望迎来量价齐升。

表 2：内存接口芯片随着内存升级使用量增加

接口协议标准	内存模组	内存接口芯片	内存模组配套芯片
DDR2	FBDIMM	AMB	
DDR3	RDIMM	1 RCD	1 TS
	LRDIMM	1 MB	1 TS
DDR4	RDIMM	1 RCD	1 TS
	LRDIMM	9DB、1RCD	1 TS
DDR5	RDIMM	1 RCD	1 SPD、2 TS、1 PMIC
	LRDIMM	10 DB、1 RCD	1 SPD、2 TS、1 PMIC

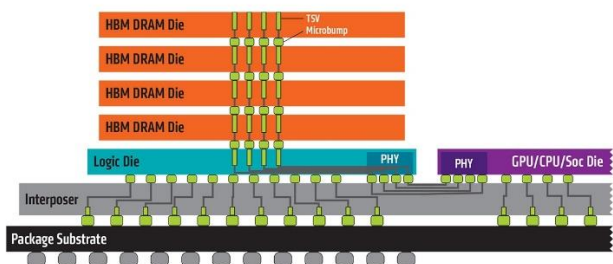
数据来源：澜起科技官网、IDT、东方证券研究所

HBM 为大算力芯片提供支撑。ChatGPT 依赖更高算力，且更多是趋向于矩阵类和卷积类的计算，其对内存的需求也体现在训练和推理 AI 芯片，或是加速模块里的内存带宽，而 HBM 是一种基于 3D 堆叠工艺的 DRAM 内存芯片，被安装在 GPU、网络交换设备、AI 加速器及高效能服务器上。

有关分析师的申明，见本报告最后部分。其他重要信息披露见分析师申明之后部分，或请与您的投资代表联系。并阅读本证券研究报告最后一页的免责声明。

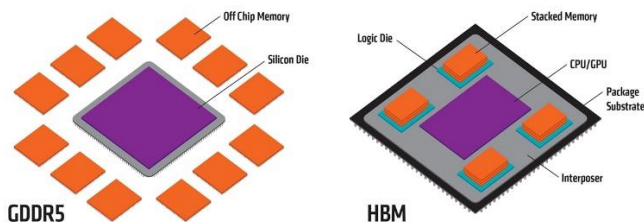
HBM 能大幅提高数据处理速度，每瓦带宽比 GDDR5 高出 3 倍还多，且 HBM 比 GDDR5 节省了 94% 的面积。HBM 作为一种带宽远超 DDR/GDDR 的高速内存，将为大算力芯片提供能力支撑，同时生成类模型也会加速 HBM 内存进一步增大容量和增大带宽。TrendForce 预估 2023-2025 年 HBM 市场年复合成长率有望成长至 40-45% 以上。

图 5：HBM 的立体结构



数据来源：AMD、东方证券研究所

图 6：HBM 与 GDDR5 对比



数据来源：AMD、东方证券研究所

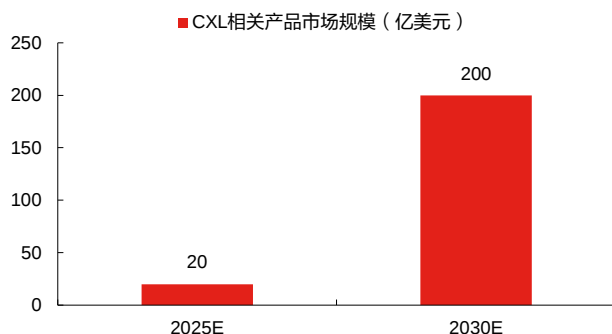
CXL 兼容性强、可实现内存一致性。CXL 全称为 Compute Express Link，是由英特尔于 2019 年 3 月在 Interconnect Day 2019 上推出的一种开放性互联协议，能够让 CPU 与 GPU、FPGA 或其他加速器之间实现高速高效的互联，从而满足高性能异构计算的要求。在 AMD、ARM、IBM 以及英特尔四个主要 CPU 相关供应商的支持下，CXL 已经成为领先的行业标准。CXL 构建于 PCI-e 逻辑和物理层级之上，所以其兼容性很高，更容易被现有支持 PCI-e 端口的处理器（绝大部分的通用 CPU、GPU 和 FPGA）所接纳，同时 CXL 可在 CPU，以及 GPU、FPGA 等之间建立高速且低延迟的互连，维护 CPU 内存空间和连接设备上的内存之间的内存一致性，允许 CPU 与 GPU 之间绕过 PCIe 协议，用 CXL 协议来共享、互取对方的内存资源，从而大幅提升数据运算效率。CXL 有望在未来的服务器中以有效和可扩展的方式增加内存、容量和带宽，成为 ChatGPT 等 AI 应用的重要配合性技术。美光科技在 22 年 5 月预测 CXL 相关产品的市场规模，到 2025 年预计将达到 20 亿美金，到 2030 年可能超过 200 亿美金。

图 7：CXL 联盟董事会成员



数据来源：CXL 官网、东方证券研究所

图 8：CXL 相关产品预测市场规模

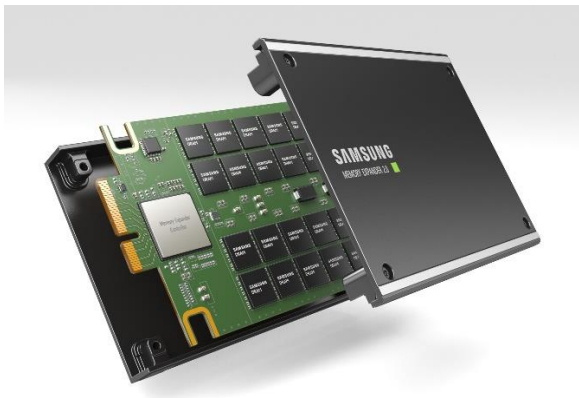


数据来源：澜起科技、美光科技、东方证券研究所

澜起科技在 22 年 5 月与三星电子合作推出 CXL 内存产品，澜起的内存扩展控制器芯片（MXC）用于三星的首款 512GB CXL DRAM 内存，该产品将为服务器系统提供容量和延时性能显著提升的内存解决方案。许多全球顶级的云计算厂商、内存模组厂商等都在进行公司产品的导入测试。

图 9：三星采用澜起科技的 CXL 内存扩展控制器芯片研发的 CXL DRAM 产品

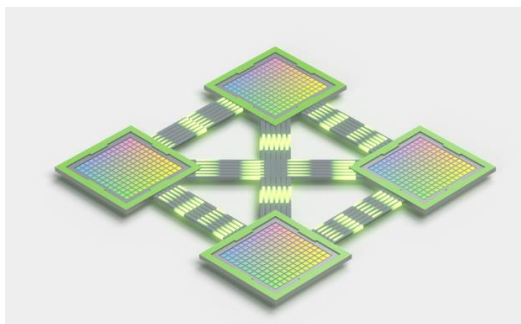
有关分析师的申明，见本报告最后部分。其他重要信息披露见分析师申明之后部分，或请与您的投资代表联系。并阅读本证券研究报告最后一页的免责声明。



数据来源：澜起科技、三星电子、东方证券研究所

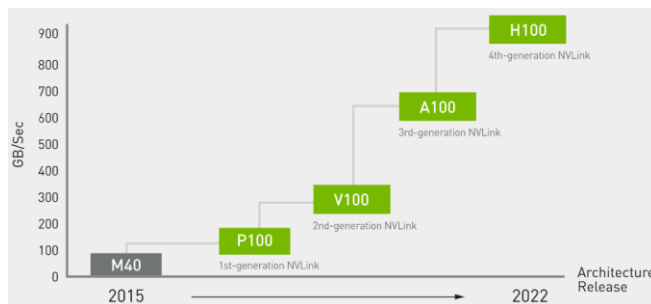
NVLink 可最大化提升系统吞吐量。AI 和 HPC（包括新兴的万亿参数模型）领域的计算需求不断增长，在这一趋势的推动下，对于能够在每个 GPU 之间实现无缝高速通信的多节点、多 GPU 系统的需求也在与日俱增。要打造功能强大且能够满足业务速度需求的端到端计算平台，可扩展的快速互连必不可少。第四代 NVIDIA® NVLink® 技术可为多 GPU 系统配置提供高于以往 1.5 倍的带宽，以及增强的可扩展性。单个 NVIDIA H100 Tensor Core GPU 支持多达 18 个 NVLink 连接，总带宽为 900GB/s，是 PCIe 5.0 带宽的 7 倍。第三代 NVIDIA NVSwitch™ 基于 NVLink 的高级通信能力构建，可为计算密集型工作负载提供更高带宽和更低延迟。为了支持高速集合运算，每个 NVSwitch 都有 64 个 NVLink 端口，并配有 NVIDIA SHARP™ 引擎，可用于网络内归约和组播加速。通过在服务器外部添加第二层 NVSwitch，NVLink 网络可以连接多达 256 个 GPU，并提供 57.6 TB/s 的惊人多对多带宽，从而快速完成大型 AI 作业。NVIDIA DGX™ H100 等服务器可利用 NVLink 技术来提高可扩展性，进而实现超快速的深度学习训练。

图 10：支持 NVLink GPU 之间连接的 NVIDIA H100



数据来源：NVIDIA、东方证券研究所

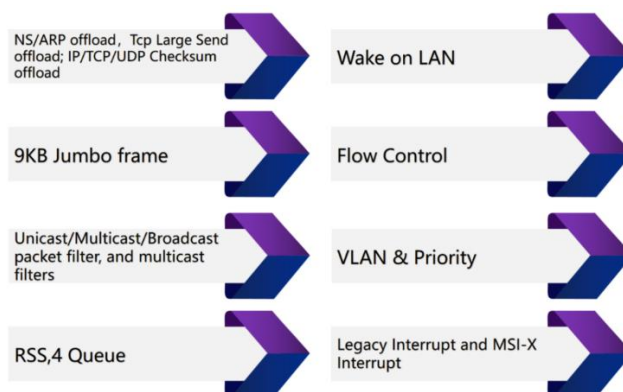
图 11：NVLink 性能



数据来源：NVIDIA、东方证券研究所

裕太微自研 PCIE 接口千兆以太网卡芯片。裕太微电子已推出首款自主研发千兆以太网卡芯片。该系列产品采用 PCIE 接口，是目前国内为数不多拥有完全自主知识产权的千兆以太网卡芯片。该系列产品支持 10/100/1000Mbps 以太网速率，采用 PCIE1.1 接口，可与 Windows、Linux 等多种操作系统适配。据测试结果显示，该款芯片以太网物理层接口在 CAT5E 线缆上的连接距离超过 130 米，PCIE 接口眼图性能优异。双向打流带宽超过 1.5Gbps，居于国际先进水平，将为 PC 机、服务器客户带来更大的经济价值。

图 12：裕太微千兆以太网卡芯片



数据来源：裕太微官网、东方证券研究所

龙讯股份重点研发 SERDES 技术和 PCIe Hub/Switch 系列芯片。为了支持总交换机芯片带宽的增加，SERDES 的速度、数量和功率也在随之增加，其速度从 10 Gbps 增加到 112 Gbps，芯片周围的 SERDES 数量从 64 通道增加到 51.2 Tbps 一代的 512 通道。基于在高速信号传输领域的技术积累，龙讯股份对 5G 高速信号高质量传输技术进行研究开发，其基于单通道 12.5Gbps SERDES 技术研发的通用高速信号延长芯片可在 5G 通信领域实现国产化应用，另外还在高速数据传输和视频传输接口技术基础上，研究 PCIe Hub/Switch 系列芯片，可用于终端系统的数据通信、接口扩展和超高清显示。

图 13：光迅科技 800G/400G 高速光模块基于 112G SERDES



数据来源：CIOE 中国光博会、东方证券研究所

投资建议

我们看好服务器带来的接口芯片技术升级趋势，建议关注澜起科技(688008，买入)、聚辰股份(688123，未评级)、裕太微-U(688515，未评级)、龙讯股份(688486，未评级)

风险提示

服务器需求拉动不及预期、新技术渗透率不及预期。

分析师申明

每位负责撰写本研究报告全部或部分内容的研究分析师在此作以下声明：

分析师在本报告中对所提及的证券或发行人发表的任何建议和观点均准确地反映了其个人对该证券或发行人的看法和判断；分析师薪酬的任何组成部分无论是在过去、现在及将来，均与其在本研究报告中所表述的具体建议或观点无任何直接或间接的关系。

投资评级和相关定义

报告发布日后的 12 个月内的公司的涨跌幅相对同期的上证指数/深证成指的涨跌幅为基准；

公司投资评级的量化标准

买入：相对强于市场基准指数收益率 15%以上；
增持：相对强于市场基准指数收益率 5% ~ 15%；
中性：相对于市场基准指数收益率在-5% ~ +5%之间波动；
减持：相对弱于市场基准指数收益率在-5%以下。

未评级 —— 由于在报告发出之时该股票不在本公司研究覆盖范围内，分析师基于当时对该股票的研究状况，未给予投资评级相关信息。

暂停评级 —— 根据监管制度及本公司相关规定，研究报告发布之时该投资对象可能与本公司存在潜在的利益冲突情形；亦或是研究报告发布当时该股票的价值和价格分析存在重大不确定性，缺乏足够的研究依据支持分析师给出明确投资评级；分析师在上述情况下暂停对该股票给予投资评级等信息，投资者需要注意在此报告发布之前曾给予该股票的投资评级、盈利预测及目标价格等信息不再有效。

行业投资评级的量化标准：

看好：相对强于市场基准指数收益率 5%以上；
中性：相对于市场基准指数收益率在-5% ~ +5%之间波动；
看淡：相对于市场基准指数收益率在-5%以下。

未评级：由于在报告发出之时该行业不在本公司研究覆盖范围内，分析师基于当时对该行业的研究状况，未给予投资评级等相关信息。

暂停评级：由于研究报告发布当时该行业的投资价值分析存在重大不确定性，缺乏足够的研究依据支持分析师给出明确行业投资评级；分析师在上述情况下暂停对该行业给予投资评级信息，投资者需要注意在此报告发布之前曾给予该行业的投资评级信息不再有效。

免责声明

本证券研究报告（以下简称“本报告”）由东方证券股份有限公司（以下简称“本公司”）制作及发布。

本报告仅供本公司的客户使用。本公司不会因接收人收到本报告而视其为本公司的当然客户。本报告的全体接收人应当采取必要措施防止本报告被转发给他人。

本报告是基于本公司认为可靠的且目前已公开的信息撰写，本公司力求但不保证该信息的准确性和完整性，客户也不应该认为该信息是准确和完整的。同时，本公司不保证文中观点或陈述不会发生任何变更，在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的证券研究报告。本公司会适时更新我们的研究，但可能会因某些规定而无法做到。除了一些定期出版的证券研究报告之外，绝大多数证券研究报告是在分析师认为适当的时候不定期地发布。

在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议，也没有考虑到个别客户特殊的投资目标、财务状况或需求。客户应考虑本报告中的任何意见或建议是否符合其特定状况，若有必要应寻求专家意见。本报告所载的资料、工具、意见及推测只提供给客户作参考之用，并非作为或被视为出售或购买证券或其他投资标的的邀请或向人作出邀请。

本报告中提及的投资价格和价值以及这些投资带来的收入可能会波动。过去的表现并不代表未来的表现，未来的回报也无法保证，投资者可能会损失本金。外汇汇率波动有可能对某些投资的价值或价格或来自这一投资的收入产生不良影响。那些涉及期货、期权及其它衍生工具的交易，因其包括重大的市场风险，因此并不适合所有投资者。

在任何情况下，本公司不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任，投资者自主作出投资决策并自行承担投资风险，任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。

本报告主要以电子版形式分发，间或也会辅以印刷品形式分发，所有报告版权均归本公司所有。未经本公司事先书面协议授权，任何机构或个人不得以任何形式复制、转发或公开传播本报告的全部或部分内容。不得将报告内容作为诉讼、仲裁、传媒所引用之证明或依据，不得用于营利或用于未经允许的其它用途。

经本公司事先书面协议授权刊载或转发的，被授权机构承担相关刊载或者转发责任。不得对本报告进行任何有悖原意的引用、删节和修改。

提示客户及公众投资者慎重使用未经授权刊载或者转发的本公司证券研究报告，慎重使用公众媒体刊载的证券研究报告。

东方证券研究所

地址：上海市中山南路 318 号东方国际金融广场 26 楼

电话：021-63325888

传真：021-63326786

网址：www.dfzq.com.cn

东方证券股份有限公司经相关主管机关核准具备证券投资咨询业务资格，据此开展发布证券研究报告业务。

东方证券股份有限公司及其关联机构在法律许可的范围内正在或将要与本研究报告所分析的企业发展业务关系。因此，投资者应当考虑到本公司可能存在对报告的客观性产生影响的利益冲突，不应视本证券研究报告为作出投资决策的唯一因素。