2025 年 11 月 25 日

GTC25，NVIDIA 发布 Vera Rubin Superchip

——人工智能月度跟踪

行业及产业

电子

强于大市

一年内行业指数与沪深 300 指数对比走势：

资料来源：聚源数据，爱建证券研究所

相关研究

《电子行业周报：数据中心助力 GaN 需求增长》2025-11-17

《电子行业周报：SK hynix 发布存储新路线，重塑 AI 时代新架构》2025-11-10

《电子行业周报：钽电容价格持续上涨》2025-11-04

《电子行业周报：关注半导体新材料的应用进展》2025-10-28

《人工智能月度跟踪：OpenAI 推出新一代音视频工具 Sora 2》2025-10-21

证券分析师

许亮

S0820525010002

0755-83562506

xuliang@ajzq.com

联系人

朱俊宇

S0820125040021

021-32229888-25520

zhujunyu@ajzq.com

投资要点：

- **事件：**2025 年 10 月 29 日，NVIDIA CEO 黄仁勋在美国华盛顿举行的 GTC 大会上发表主题演讲，不仅展示了下一代超级芯片 Vera Rubin 的原型机，更提出了“AI 不是工具，而是会用工具的工人”这一颠覆性观点。

- **作为全球知名芯片设计公司，NVIDIA 长期专注高性能 GPU 及系统芯片研发。**在架构早期阶段，其通过 2010 年前的 Tesla、2010 年的 Fermi、2012 年的 Kepler 及 2014 年的 Maxwell 四大系列，推动 GPU 从“图形加速专用硬件”成功转型为“通用并行计算引擎”，期间持续优化可靠性、能效比与场景适配能力；2016 年后，伴随 AI 与高性能计算需求爆发，NVIDIA 加快迭代节奏，先后推出 2016 年的 Volta（首推 Tensor Core 开启 AI 算力硬件加速）、2022 年的 Hopper（以 Transformer Engine 支撑大模型研发）、2024 年的 Blackwell（全栈协同设计提升 AI 推理与图形渲染性能）等架构，并计划通过下一代 Rubin、Feynman 架构突破算力密度与能效比极限，服务超复杂场景。

- **为突破摩尔定律逼近物理极限、晶体管密度对算力边际贡献下降的瓶颈，NVIDIA 于 2025 年 10 月 29 日推出 Vera Rubin 超级芯片。**该超级芯片并未单纯堆积晶体管，而是通过 CPU 与 GPU 的异构协同、HBM4 高带宽显存的搭配，以及 CUDA 生态的兼容，以架构与系统级创新实现算力跃升。1）从硬件配置来看，Vera Rubin 超级芯片由 Rubin GPU 与 Vera CPU 构成。RubinGPU 被大量电源电路环绕，配备 8 个 HBM4 显存位点，且集成两颗 Reticle 尺寸 GPU 芯片；VeraCPU 则搭载 88 个定制 ARM 核心，总计提供 176 个线程。2）从性能迭代来看，NVIDIA 芯片从 Hopper、Blackwell 演进至 Rubin 架构。Vera Rubin 超级芯片作为该代旗舰，其 VR200、VR300（Ultra）加速器的 FP4 算力分别达 50、100 PFLOPS，搭配 288 GB HBM4 乃至 1025 GB HBM4E 显存，带宽最高 32TB/s，较前代 Blackwell 架构实现数倍提升；同时 CPU 从 Grace 系列升级为 Vera 系列，核心性能与线程数的提升进一步释放了异构协同的算力。

- **伴随 Vera Rubin 超级芯片的持续推进，NVIDIA 计划于 2026 H2、2027 H2 分别推出 Vera Rubin NVL144 平台与更高规格的 Rubin Ultra NVL576 平台。**1）Vera Rubin NVL144 平台采用 Rubin GPU 与 Vera CPU 组合设计。Rubin GPU 包含两颗 Reticle 尺寸核心、具备 50 PFLOPS（FP4 精度）算力及 288 GB HBM4 显存；Vera CPU 提供 88 个定制 Arm 核心、176 线程且 NVLINK-C2C 互联带宽达 1.8 TB/s，该平台性能相较上一代 GB300 NVL72 提升约 3.3 倍，FP4 推理算力 3.6 Exaflops、FP8 训练算力 1.2 Exaflops，系统总显存带宽 13 TB/s、快速存储容量 75 TB；2）而 Rubin Ultra NVL576 作为迭代产品，硬件与性能全面升级。NVL 规模从 144 扩展至 576，CPU 架构不变，GPU 升级为四颗 Reticle 尺寸核心（单颗性能最高 100 PFLOPS FP4 精度、搭载 1 TB HBM4e 显存）。该平台性能较 GB300 NVL72 提升 14 倍，FP4 推理算力 15 Exaflops、FP8 训练算力 5 Exaflops，HBM4 显存带宽 4.6 PB/s、快速存储容量 365 TB，NVLINK 通信能力提升 12 倍（最高 1.5 PB/s），整体算力、存储及连接效率大幅跃升。

- **投资建议：建议关注 NVIDIA 服务器国产供应链的持续成长机遇。**

- **风险提示：1）国际贸易摩擦加剧 2）下游需求不及预期 3）技术升级进度滞后**

事件：GTC25，NVIDIA 发布 Vera Rubin 超级芯片

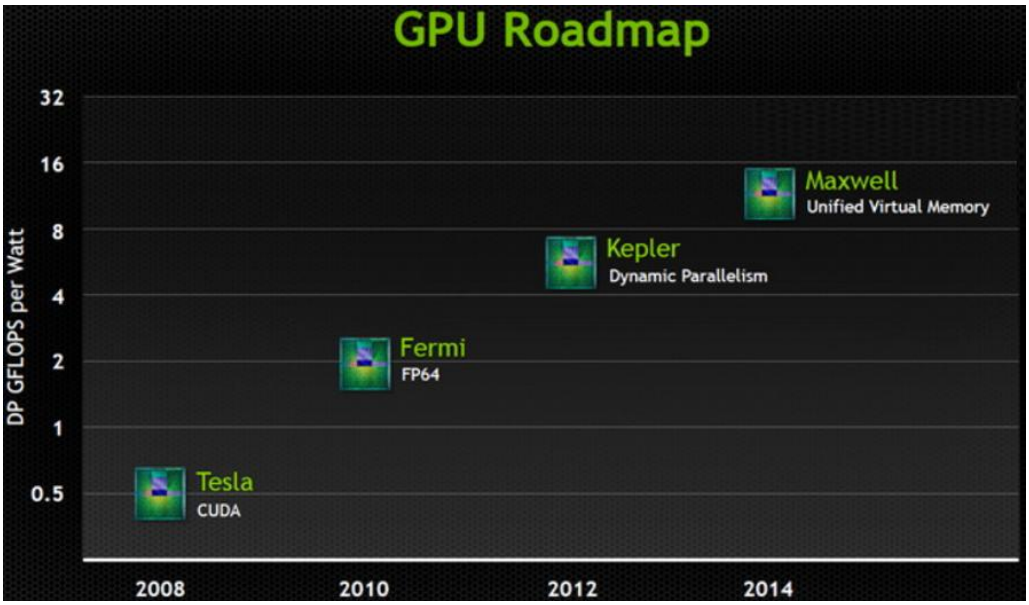
2025 年 10 月 29 日，NVIDIA CEO 黄仁勋在美国华盛顿举行的 GTC 大会上发表主题演讲，不仅展示了下一代超级芯片 Vera Rubin 的原型机，更提出了“AI 不是工具，而是会用工具的工人”这一颠覆性观点。

1. NVIDIA 从 Tesla、Fermi 向 Vera Rubin 迭代升级

NVIDIA 作为全球知名的芯片设计公司，专注于高性能图形处理器（GPU）及系统芯片的研发与升级。

在 GPU 架构的早期演进中，NVIDIA 通过 Tesla、Fermi、Kepler、Maxwell 四大系列，实现了 GPU 从“图形加速专用硬件”到“通用并行计算引擎”的关键技术跃迁。2010 年前，NVIDIA 推出的 Tesla 架构，正式开启了 GPU 从图形加速向通用计算的跨越式转型；2010 年，Fermi 架构聚焦可靠性与通用性升级，不仅首次引入 ECC 错误校验码内存以保障数据计算准确性，还对 CUDA 核心进行优化，使其能支持更多编程语言；2012 年，Kepler 架构以“能效比革命”为核心，推出 SMX 流式多处理器以提升并行效率，同时首次支持 GPUDirect 技术，实现了 GPU 间及 GPU 与存储设备的直接数据传输；到了 2014 年，NVIDIA 又推出 Maxwell 架构，该架构采用台积电 28nm 工艺制程。彼时移动设备兴起，市场对低功耗、高性能 GPU 的需求大增，NVIDIA 也针对性优化了该架构在不同应用场景的适配能力。

图表 1：NVIDIA 早期 GPU 技术蓝图

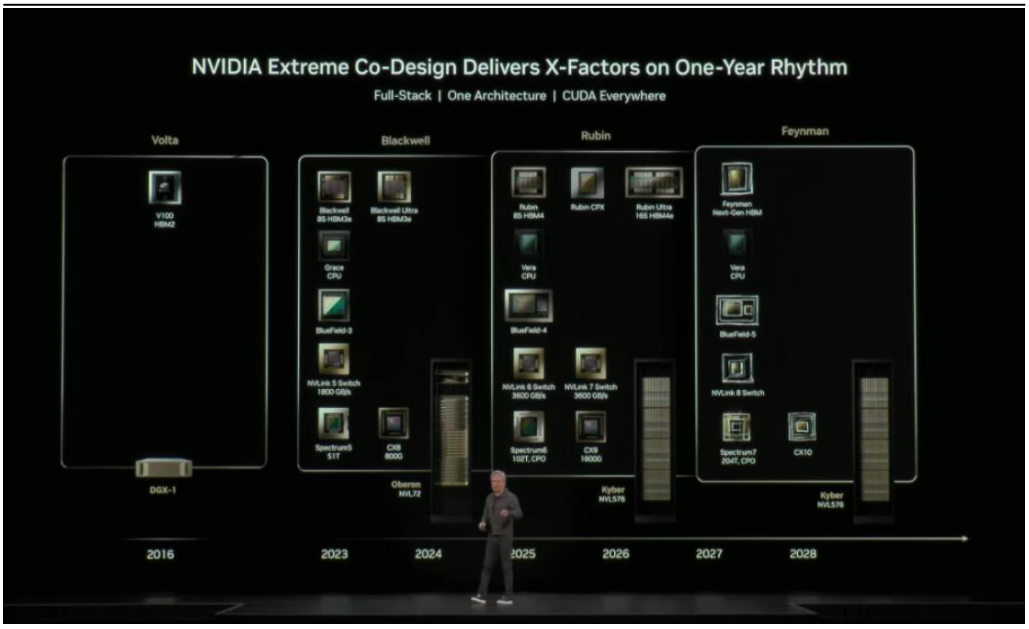


资料来源：EXR Review，爱建证券研究所

随着人工智能与高性能计算需求的持续爆发，NVIDIA 自 2016 年后进一步加快架构迭代节奏，先后推出 Volta、Hopper、Blackwell 等一系列突破性架构，并计划通过下一代架构持续拓展算力边界。2016 年，Volta 架构凭借首创的 TensorCore 技术，开启 AI 算力硬件加速的商业化时代，为深度学习大规模落地提供关键算力支撑；2022

年，Hopper 架构聚焦大模型需求，以 Transformer Engine 夯实千亿参数模型研发基础；2024 年发布的 Blackwell 架构，则通过“芯片-系统-软件”全栈协同设计，在 AI 推理效率与图形渲染性能上实现双重跃升。

图表 2：NVIDIA 2016-2026 年 GPU 技术路线持续迭代



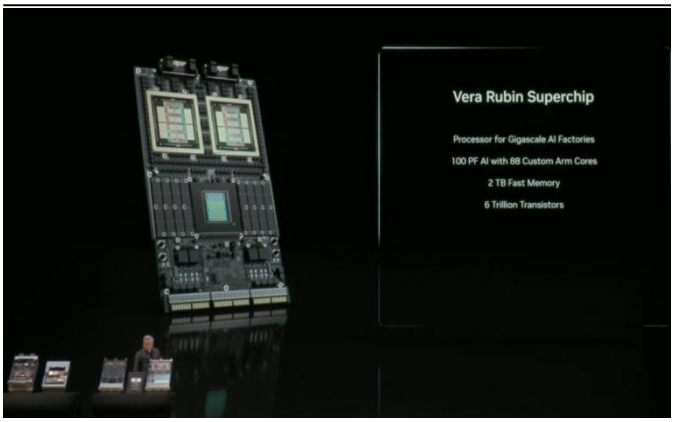
资料来源：NVIDIA，IT 之家，爱建证券研究所

而 NVIDIA 计划推出的下一代架构，将进一步突破算力密度与能效比的极限，为超大规模 AI 集群、量子计算协同模拟等复杂场景提供底层技术支撑。

2. NVIDIA Vera Rubin 超级芯片以异构协同与架构创新实现算力跃升

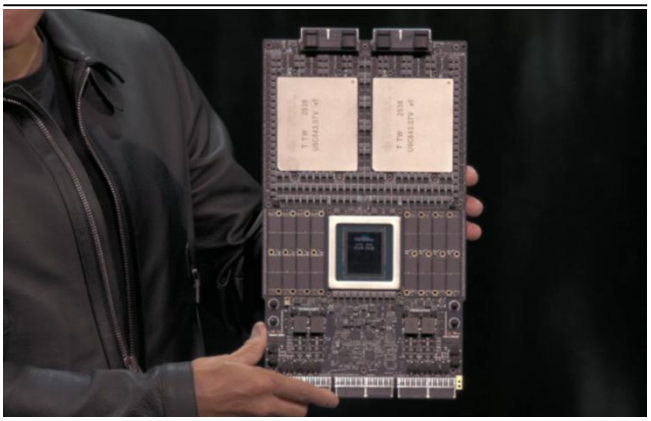
伴随“摩尔定律”逐步逼近物理极限，晶体管密度提升对算力的边际贡献持续走低。为突破这一瓶颈，NVIDIA 于 2025 年 10 月 29 日推出 Vera Rubin 超级芯片。该超级芯片并未单纯堆积晶体管，而是通过 CPU 与 GPU 的异构协同、HBM4 高带宽显存的搭配，以及 CUDA 生态的兼容，以架构与系统级创新实现算力跃升。

图表 3：Vera Rubin 超级芯片构成图



资料来源：NVIDIA，IT 之家，爱建证券研究所

图表 4：芯片主板由 Vera CPU 与 Rubin GPU 构成



资料来源：NVIDIA，IT 之家，爱建证券研究所

每块 Rubin GPU 被大量电源电路环绕，配备 8 个 HBM4（HBM4 高宽带显存）显存

位点，集成两颗 Reticle 尺寸（半导体光刻机掩模版的最大制造尺寸）GPU 芯片；Vera CPU 搭载 88 个定制 ARM 核心，总计提供 176 个线程。

从性能迭代维度看，在 NVIDIA 芯片从 Hopper、Blackwell 到 Rubin 的架构演进中，Vera Rubin 超级芯片作为旗舰产品，其 VR200、VR300（Ultra）加速器的 FP4 算力分别达 50、100 PFLOPS，搭配 288 GB HBM4 乃至 1025 GB HBM4E 显存，带宽最高达 32TB/s，较前代 Blackwell 架构实现数倍跃升；同时，Nvidia CPU 从 Grace 系列升级为 Vera 系列，核心性能与线程数的提升进一步支撑了异构协同的算力释放。

图表 5: NVIDIA 芯片从 Hopper、Blackwell、Rubin 架构演进

架构	Hopper		Blackwell			Rubin	
时间	2022	2023	2024	2025		2026	2027
Accelerator	H100	H200	B200/GB200	GB300(Ultra)	GB300(B300A)	VR200	VR300 (Ultra)
GPT TDP (w)	700	700	700/1200	1400	600	1800	3600
Foundry Node	4N	4N	4NP	4NP	4NP	**3NP**	**3NP**
FP4 PFLOPS	4	4	10	**15**	**4.6**	**50**	**100**
HBM	80 GB HBM3	141GB HBM3E	192GB HBM3E	288GB HBM3E	144GB HBM3E	**288GB HBM4**	**1025GB HBM4E**
HBM Stacks	**5**	6	8	8	4	8	**16**
HBM BandWidth(TB/s)	**3.35**	4.8	8	8	4	13	**32**
Packaging	CoWoS-S	CoWoS-S	CoWoS-L	CoWoS-L	CoWoS-L	CoWoS-L	CoWoS-L
SerDes speed(G)	112	112	224	224	224	**224**	**448**
Nvidia CPU	Grace					**Vera**	

资料来源：SemiAnalysis，爱建证券研究所

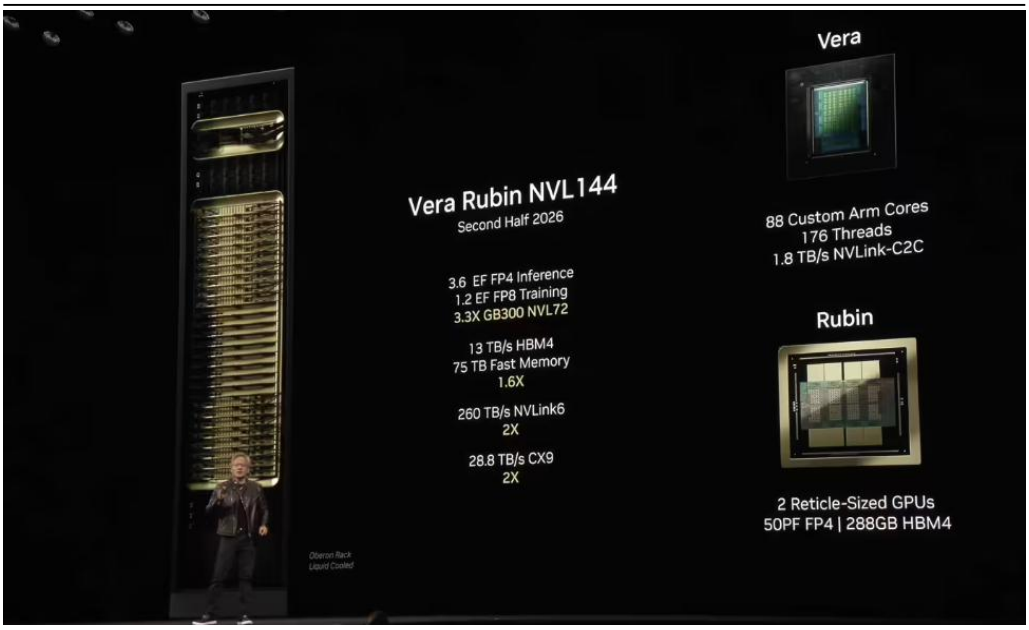
3. NVIDIA 2026-2027 H2 计划推出 Vera Rubin 平台

针对新一代计算平台，黄仁勋宣布，公司预计于 2026 年 H2 推出 Vera Rubin NVL144 平台，并计划在 2027 年 H2 进一步推出 Rubin Ultra NVL576 平台。

Vera Rubin NVL144 平台的核心硬件采用 Rubin GPU 与 Vera CPU 的组合设计。其中，Rubin GPU 由两颗 Reticle 尺寸核心构成，不仅具备 50 PFLOPS（FP4 精度）的算力，还配备 288 GB HBM4 显存；Vera CPU 则提供 88 个定制 Arm 核心与 176 线程，其 NVLINK-C2C 互联带宽可达到 1.8 TB/s。

性能层面，Vera Rubin NVL144 平台相较上一代 GB300 NVL72 实现持续提升。其中，该平台的 FP4 推理算力达 3.6 Exaflops、FP8 训练算力达 1.2Exaflops，较 GB300 NVL72 提升约 3.3 倍；系统总显存带宽为 13 TB/s，快速存储容量为 75 TB，两项指标较上一代分别提升 60%。此外，该平台的 NVLINK 与 CX9 通信能力也实现双倍提升，最高速率分别可达 260 TB/s 与 28.8 TB/s。

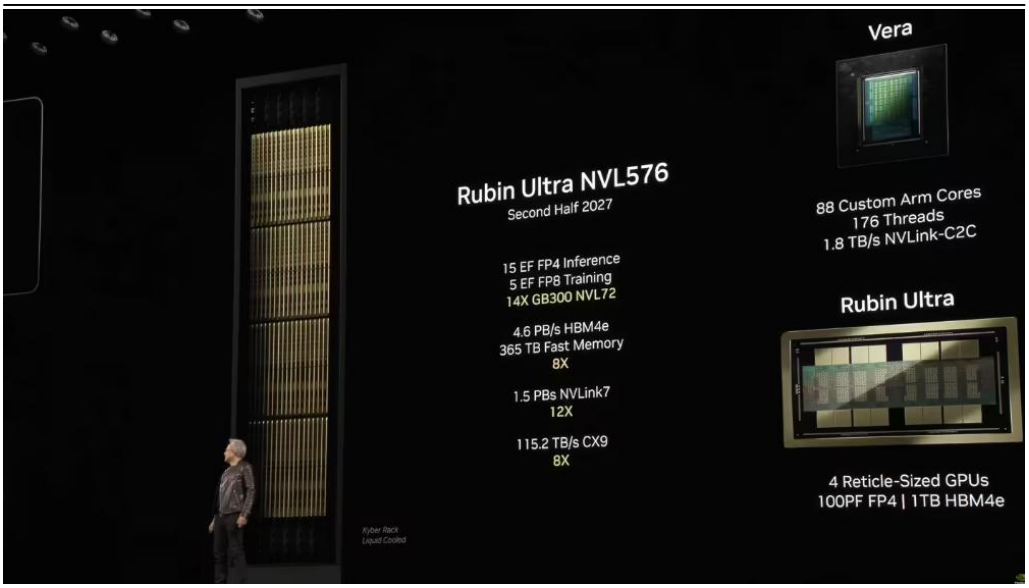
图表 6：Vera Rubin NVL144 平台参数性能



资料来源：NVIDIA，IT 之家，爱建证券研究所

Rubin Ultra NVL576 平台作为更高规格的迭代产品，在硬件与性能上均实现全面升级。1）其 NVL 规模从 144 扩展至 576，CPU 架构保持不变，GPU 则升级为四颗 Reticle 尺寸核心；单颗 GPU 性能最高可达 100 PFLOPS（FP4 精度），并搭载 1TB HBM4e 显存。2）性能层面，该平台可实现 15 Exaflops（FP4 推理）与 5 Exaflops（FP8 训练）算力，相较上一代 GB300 NVL72 提升 14 倍，同时 HBM4 显存带宽达到 4.6 PB/s、快速存储容量达 365 TB；通信能力上，NVLINK 与 CX9 分别提升至 12 倍与 8 倍，最高速率依次达到 1.5 PB/s 与 115.2 TB/s，整体算力、存储与连接效率均大幅跃升。

图表 7：Rubin Ultra NVL576 平台参数性能



资料来源：NVIDIA，IT 之家，爱建证券研究所

4. 风险提示

1）国际贸易摩擦加剧

2）下游需求不及预期

3）技术升级进度滞后

爱建证券
AIJIAN SECURITIES

爱建证券有限责任公司

上海市浦东新区前滩大道 199 弄 5 号
电话：021-32229888　　传真：021-68728700　　服务热线：956021
邮政编码：200124　　邮箱：ajzq@ajzq.com　　网址：http://www.ajzq.com

评级说明

投资建议的评级标准

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 6 个月内的相对市场表现，也即：以报告发布日后的 6 个月内的公司股价（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。其中：A 股市场：沪深 300 指数（000300.SH）；新三板市场：三板成指（899001.CSI）（针对协议转让标的）或三板做市指数（899002.CSI）（针对做市转让标的）；北交所市场：北证 50 指数（899050. BJ）；香港市场：恒生指数（HIS.HI）；美国市场：标普 500 指数（SPX.GI）或纳斯达克指数（IXIC.GI）。

股票评级

买入	相对同期相关证券市场代表性指数涨幅大于 15%
增持	相对同期相关证券市场代表性指数涨幅在 5%～15%之间
持有	相对同期相关证券市场代表性指数涨幅在-5%～5%之间
卖出	相对同期相关证券市场代表性指数涨幅小于-5%

行业评级

强于大市	相对表现优于同期相关证券市场代表性指数
中性	相对表现与同期相关证券市场代表性指数持平
弱于大市	相对表现弱于同期相关证券市场代表性指数

分析师声明

本报告署名分析师在此声明：我们具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，本报告采用信息和数据来自公开、合规渠道，所表述的观点均准确地反映了我们对标的证券和发行人的独立看法。研究报告对所涉及的证券或发行人的评价是分析师本人通过财务分析预测、数量化方法、或行业比较分析所得出的结论，但使用以上信息和分析方法可能存在局限性，请谨慎参考。

法律主体声明

本报告由爱建证券有限责任公司（以下统称为“爱建证券”）证券研究所制作，爱建证券具备中国证监会批复的证券投资咨询业务资格，接受中国证监会监管。

本报告是机密的，仅供我们的签约客户使用，爱建证券不因收件人收到本报告而视其为爱建证券的签约客户。本报告中的信息均来源于我们认为可靠的已公开资料，但爱建证券对这些信息的准确性及完整性不作任何保证。本报告中的信息、意见等均仅供签约客户参考，不构成所述证券买卖的出价或征价邀请或要约。该等信息、意见未考虑到获取本报告人员的具体投资目的、财务状况以及特定需求，在任何时候均不构成对任何人的个人推荐。客户应当对本报告中的信息和意见进行独立评估，并应同时考量各自的投资目的、财务状况和特定需求，必要时就法律、商业、财务、税收等方面咨询专家的意见。对依据或者使用本报告所造成的一切后果，爱建证券及其关联人员均不承担任何法律责任。

本报告所载的意见、评估及预测仅为本报告出具日的观点和判断。该等意见、评估及预测后续可随时更改。过往的表现亦不应作为日后表现的预示和担保。在不同时期，爱建证券可能会发出与本报告所载意见、评估及预测不一致的研究报告。

版权声明