

## Rubin 按计划 Q3 出货，CPU 带来 2000 亿美元新增市场

## 英伟达 (NVDA.O) FY27Q1 跟踪报告

## 事件:

英伟达发布 FY27Q1 季报，本季营收 816 亿美元，同比+85%/环比+20%，创历史新高；non-GAAP 毛利率为 75%，同比+14.2pcts/环比-0.1pct。综合财报及交流会议信息，总结要点如下：

## 评论:

## 1、FY27Q1 营收高增再超指引，公司当前总储备规模达 1450 亿美元。

FY27Q1 营收 816 亿美元，同比+85%/环比+20%，超出此前预期（780±2% 亿美元），本季营收、营业利润和自由现金流均再创历史新高；non-GAAP 毛利率为 75%，符合此前指引（74.5%-75.5%），同比+14.2pcts/环比-0.1pctsc。公司面向超大规模云厂商、大模型研发企业、AI 云服务商及各国政企客户等全品类终端市场，持续规模化落地 Blackwell 全系算力产品。一季度，公司总储备规模（库存、预付采购订单等）达 1450 亿美元。

## 2、NV AI 基础设施价值不断提升，消费产品需求受内存价格影响小幅回落。

1) 数据中心：营收 750 亿美元创历史新高，同比+92%/环比+21%，细分来看，面向前沿大模型领域的数据中心算力业务营收 600 亿美元，同比+77%，数据中心网络业务营收 150 亿美元，同比增长近两倍。按市场划分，①超大规模算力市场：营收 380 亿美元，占数据中心总营收约五成，环比+12%。公司 AI 基础设施的价值在不断提升：H100 租赁价格年初至今上涨 20%，A100 云端定价上涨近 15%；②ACIE (AI Clouds、Industrial 和 Enterprise)：营收 370 亿美元，环比+31%，其中 AI 云业务营收同比增长超两倍；2) 边缘算力：涵盖 agentic 与 physical AI 相关的设备，营收 64 亿美元，同比+29%/环比+10%，Blackwell 工作站的强劲需求是增长的重要贡献，但受内存及整机硬件成本上涨影响，消费级终端产品需求小幅回落。公司 Physical AI 业务发展势头强劲，近十二个月累计营收已突破 90 亿美元。

## 3、下季营收指引同比高增，2030 年预计全球 AI 资本开支达 3-4 万亿美元。

1) FY27Q2 业绩指引：营收指引中值为 910 亿美元（±2%），同比+95%/环比+12%，业绩展望未纳入来自中国市场的任何收入；Non-GAAP 毛利率预计为 75%（±0.5pct），环比持平。全年毛利率仍预期处于 mid-70s 区间；2) CSP 资本开支：2027 年全球头部云厂商资本开支总额将突破 1 万亿美元，叠加 Agent AI 全面渗透各行各业，到 2030 年末，全球 AI 基础设施投入规模有望达到 3-4 万亿美元。

## 4、Vera 将带来 2000 亿美元新增市场，预计 Rubin 整体处于供应紧缺状态。

1) CPU：Vera CPU 成功为英伟达打开规模达 2000 亿美元的全新蓝海市场，公司有望成为全球领先的 CPU 供应商。今年公司 CPU 业务整体营收预期规模接近 200 亿美元。按计划 Vera-Rubin 系列将于第三季度量产出货。2) Vera 四种使用方式：第一种 Vera Rubin，公司会卖出数百万颗 Rubin，每两颗 Rubin 配一颗 Vera。第二种 Vera 独立 CPU。第三种 Vera+CX9+存储相关软件栈。第四种 Vera+CX9+用于安全、计算隔离和机密计算的软件栈。预计在

## 推荐（维持）

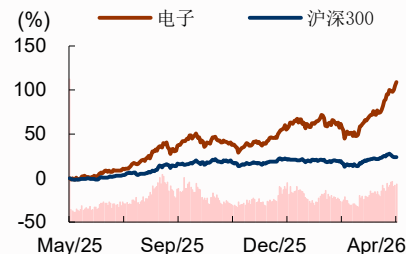
TMT 及中小盘/电子

## 行业规模

|          |         | 占比%  |
|----------|---------|------|
| 股票家数（只）  | 529     | 10.2 |
| 总市值（十亿元） | 20476.8 | 17.2 |
| 流通市值（十亿） | 17065.7 | 16.0 |

## 行业指数

| %    | 1m   | 6m   | 12m   |
|------|------|------|-------|
| 绝对表现 | 22.7 | 54.3 | 107.1 |
| 相对表现 | 20.7 | 48.0 | 82.7  |



资料来源：公司数据、招商证券

## 相关报告

- 1、《英伟达 GTC 2026 跟踪报告—25-27 年 DC 收入超 1 万亿美元，Kyber 将使用铜光等多种互连形式》2026-03-18
- 2、《英伟达 (NVDA.O) FY26Q4 跟踪报告—本季营收与指引均高增，战略备货以满足未来市场需求》2026-02-27
- 3、《英伟达 CES 2026 跟踪报告—Vera Rubin 已正式量产，展示全新 Agentic 和 Physical AI 平台》2026-01-06

鄢凡 S1090511060002

✉ yanfan@cmschina.com.cn

程鑫 S1090523070013

✉ chengxin2@cmschina.com.cn

湛薇 S1090524070008

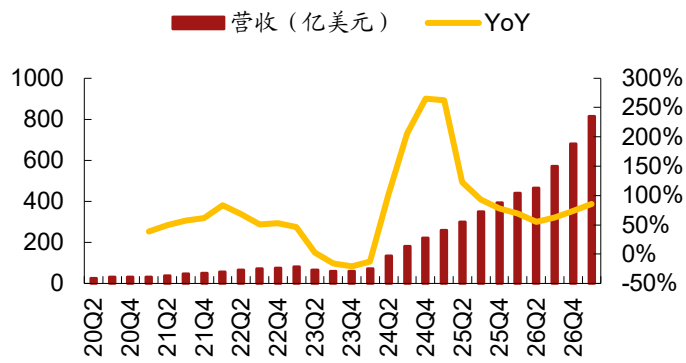
✉ shenwei3@cmschina.com.cn

Vera Rubin 整个生命周期内，公司都将处于供应紧张状态。**3) 股东回报：**公司宣布将季度每股股息从 0.01 美元上调至 0.25 美元，一季度向股东回馈分红总额达 200 亿美元。同时新增 800 亿美元股票回购授权额度，叠加此前回购计划中尚未执行的 390 亿美元额度，回购储备进一步充裕。

**投资建议：**建议关注国际 GPU 龙头英伟达及其产业链标的，关注服务器硬件层面所涉及到的系统组装、GPU、CPU、存储、高速连接器和电光连接、PCB/IC 载板、散热、电源、各类辅助芯片等零部件的投资机会，并关注国产算力厂商和华为昇腾等自主算力产业链相关公司，以及先进制造/封装和 HBM 等产业链机会。①PCB/CCL：胜宏科技、生益科技、景旺电子、沪电股份、方正科技、鹏鼎控股、超颖电子、东山精密、南亚新材等；②PCB 设备：大族数控/大族激光、芯基微装，鼎泰高科（机械&有色联合覆盖）等；③高速互联：立讯精密、汇聚科技、东山精密、致尚科技、沃尔核材等；④液冷/电源：立讯精密、比亚迪电子、领益智造、欧陆通、奥海科技、奕东电子等；⑤组装：工业富联。

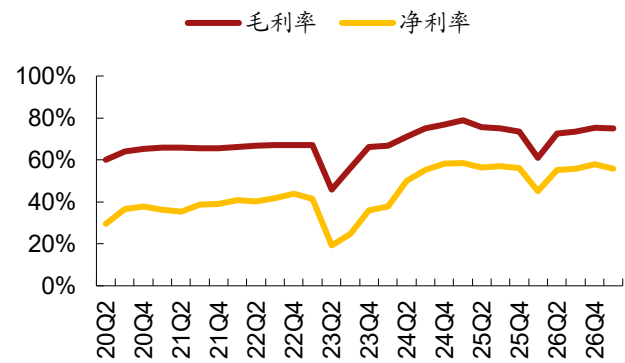
**风险提示：**竞争加剧风险、政策与贸易摩擦风险、技术更新与产品周期风险、宏观经济与行业景气度风险。

图 1：英伟达各财季营收及 YOY



资料来源：英伟达，招商证券

图 2：英伟达各财季毛利率及净利率



资料来源：英伟达，招商证券

图 3：英伟达 FY27Q2 业绩预测

## Q2 FY27 Outlook – Non-GAAP

|                    | Q2 FY27         |               |
|--------------------|-----------------|---------------|
| Revenue            | \$91.0 billion* | + / - 2%      |
| Gross Margin %     | 75.0%           | + / - 50 bps  |
| Operating Expenses | \$8.3 billion   | Approximately |

资料来源：英伟达，招商证券

(后附 FY2027Q1 业绩说明会纪要全文)

## 附录：英伟达 FY2027Q1 业绩说明会纪要

时间：2026 年 5 月 21 日

出席：Colette Kress-EVP & CFO

Jensen Huang-CEO & President

Toshiya Hari-IR

**业绩概述：**本季度业绩再创历史，营收、营业利润、自由现金流均突破此前纪录。总营收 820 亿美元，同比增长 85%，环比增长 20%，连续三个季度同比加速、连续 14 个季度环比增长。

### AI 数据中心业务：

数据中心营收 750 亿美元，同比增 92%，环比增 21%，核心驱动力为 Blackwell 架构持续强劲。GB300 与 VL72 需求尤为旺盛，头部模型厂商与超大规模云厂商已累计部署数十万片 Blackwell GPU，创下公司史上最快产品上量速度。Grace Blackwell 是当前最快训练系统，同时推理侧的单 Token 生成成本最低。

我们专为 AI 打造的端到端以太网平台 Spectrum X，规模已超过所有其他以太网网络厂商总和。InfiniBand 本季度表现强劲，同比增长超 4 倍，得益于下一代 XDR 技术部署。其中，数据中心计算营收 600 亿美元，同比增 77%；数据中心网络营收 150 亿美元，同比增长近乎三倍。

为了更贴合当前与未来增长引擎，公司采用了新的报告框架，公司拥有两大市场平台：数据中心、边缘计算，数据中心进一步划分为两个子市场，Hyperscale 以及 ACIE，Hyperscale 包括来自公有云以及全球最大互联网公司的收入，ACIE 则反映我们在各行业、各国家 AI 专用数据中心与 AI 工厂中的增长机会。边缘计算则突出了用于智能体 AI 和实体 AI 的设备，包括 PC、游戏机、工作站、基于 AI 基础设施的工作站、机器人和汽车。过去九个季度按新框架拆分的营收数据已发布在官网。

**1) 超大规模算力市场：**FY27Q1 超大规模营收 380 亿美元，约占数据中心 50%，环比增 12%；

**2) ACIE：**营收 370 亿美元，环比增 31%，其中 AI 云营收同比增长超三倍，其中 AI 云营收同比增长超三倍，公司的客户已经实现 AI 算力容量的快速部署。

合作伙伴中，功率超 10 兆瓦的数据中心数量一年内近乎翻倍，目前已经超过 80 个站点。主权国家算力营收同比增超 80%。英伟达 AI 基础设施已部署在近 40 个国家，覆盖 GDP 总量 50 万亿美元。我们的客户群体多元且持续扩张，依托庞大生态、存量装机量、丰富的 CUDA 加速应用，以及最低 Token 成本的优势，我们的 AI 计算平台市场空间远超其他竞品。

AI 基础设施需求以前所未有的速度扩张，AI 工厂建设加速。英伟达 AI 基础设施价值持续提升：H100 云租价年初至今上涨 20%，A100 上涨近 15%。得益于平台通用性与软件栈持续性能优化，客户的 GPU 在折旧周期外仍能产生盈利收入。英伟达算力形成庞大可信市场，成为生态为 AI 基础设施投入数百亿资金

的关键基础。

推动 AI 基础设施建设加速的原因主要有两个。第一，从搜索与广告，到推荐系统，再到内容理解，最大的 hyperscale 工作负载正在持续从 CPU 转向 GPU 加速计算。第二，原生 AI 产品与服务的采用正在进入加速拐点。自 ChatGPT 出现以来，我们已经看到主流 AI 从一次性推理，演进到思考推理，再到如今的智能体。AI 已经不再是“可有可无”。AI 现在已经成为提升各行业、各岗位生产力的必要工具。这推动 AI 全产业链收入加速增长，包括能源、芯片、基础设施、大模型、应用。模型层增长惊人，Anthropic、OpenAI 增速持续加快，GPT-5.5 发布后 OpenAI Codex 实现爆发式增长。

分析师预测，2027 年超大规模云厂商资本开支将突破 1 万亿美元，智能体 AI 向全行业渗透，到本年代末，AI 基础设施年支出将达 3 万亿—4 万亿美元。

Blackwell 全面落地，Blackwell 架构已被所有头部超大规模云厂商、云服务商、模型厂商采用部署，上月，OpenAI 发布 GPT-5.5，该模型与 Blackwell 联合设计、基于 Blackwell 训练、并运行在 Blackwell 之上，目前位居 Artificial Analysis 评测榜单榜首。微软 Fairwater—全球算力最强的 AI 数据中心，已提前投入使用，搭载数十万片 Blackwell GPU。从今年开始，AWS 将新增部署超 100 万片 Blackwell 与 Rubin GPU，并与我们合作开展 Spectrum 网络相关建设。在 Google，Blackwell 将作为云服务提供给客户，包括保密计算功能，为安全高性能 AI 建立新的基础。

我们在前沿 AI 算力中的份额持续提升，与 Anthropic 深化合作，成为其战略伙伴，通过 AWS、Azure、CoreWeave、xAI 等为其扩容算力。叠加 OpenAI、Gemini、xAI、Meta 等头部实验室均基于英伟达构建，我们在前沿大模型领域的份额将大幅增长。

当前数据中心是创收型 AI 工厂，受电力与资金约束，运营商必须选择最优架构。我们通过极致协同设计，实现行业最低 Token 成本、最高 Token 吞吐量、最高投资回报率。MLPerf 推理测试结果出炉，Blackwell Ultra 再次横扫所有基准测试，在全模型、全部署场景下吞吐量第一。GB300 相较半年前，吞吐量提升 2.7 倍，单 Token 成本下降 60%，全栈创新功不可没。

英伟达算力不只是最高性能 AI 基础设施，更是最经济、最易融资的方案。客户买的不是 GPU，而是建 AI 工厂；核心指标不是 GPU 采购价，而是 AI 工厂全生命周期成本：Token / 瓦、Token / 美元、可用性、利用率、上线速度、软件耐用性、资产寿命——英伟达在所有维度均领先。

智能体 AI 与强化学习为 CPU 带来全新增长机遇。继 Grace CPU 成功后，Vera CPU 恰逢其时推出，基于定制 Arm 内核，与 Rubin GPU、NVLink 端到端协同设计：单核性能提升 1.5 倍、每瓦性能提升 2 倍、单机柜密度提升 4 倍（对比 x86）

Vera 为英伟达打开 2000 亿美元全新市场，是我们此前从未涉足的领域。所有头部超大规模厂商与系统厂商均已合作部署。我们预计今年 CPU 总营收近 200 亿美元，有望成为全球领先 CPU 供应商。产品迭代速度行业无人能及，是我们市场地位的核心支柱。

Vera Rubin 计划今年下半年（Q3 起）量产发货。整套方案整合 5 个加速机架、7 颗专用芯片，推理吞吐量较 Blackwell 最高提升 35 倍，AI 工厂收入最高提升 10 倍。

Google 的 A5X 裸金属实例已率先采用，可跨机房支持最高 96 万片 Rubin GPU，助力客户运行最大规模 AI 工作负载。

美国政府已批准 H200 向中国客户出口许可，但我们尚未产生任何收入，且不确定最终能否获准进口。因此，与上季度一致，我们的业绩展望未计入中国数据中心计算业务收入。

#### 边缘计算业务：

边缘计算平台营收 64 亿美元，环比增 10%，同比增 29%。Blackwell 工作站需求强劲，消费端因内存与系统价格上涨小幅下滑。

Physical AI 持续增长，过去 12 个月收入超 90 亿美元。与优步合作，2028 年前将为四大洲近 30 座城市的自动驾驶出租车车队提供算力支持。机器人领域，工业、手术、人形机器人等头部企业均基于英伟达技术规模化开发部署。

公司积极保障供应链，一季度将总供应能力（含库存、预付采购承诺）提升至 1450 亿美元。虽面临供应链挑战，但我们有信心支撑未来增长。核心供应商的长期合作与规模优势持续支撑业务。

#### FY27Q1 财务情况：

**毛利率：**本季度公认会计原则（GAAP）毛利率为 74.9%，非公认会计原则（Non-GAAP）毛利率为 75%，环比基本持平，主要受益于 Blackwell 架构出货量持续提升，高端产品占比稳定。

**费用：**公认会计原则（GAAP）营业费用环比增长 12% 非公认会计原则（Non-GAAP）营业费用环比增长 12% 增长主要源于研发投入、员工薪酬及算力 / 基础设施成本上升。

**税率：**第一季度非公认会计原则（Non-GAAP）有效税率为 16%，低于此前预期区间（17%-19%），主要受公司业务地域结构优化影响。

**库存：**公司战略性增加库存与采购承诺，已锁定产能以满足未来多个季度的市场需求。本季度将总供应能力（含库存、预付采购承诺）提升至 1450 亿美元，储备时间跨度超出常规水平，体现了对市场需求的长期清晰判断。尽管先进架构产品供应紧张状况仍将持续，但凭借公司规模优势、完善的供应链体系及长期稳定的合作伙伴关系，我们有信心把握未来的增长机遇。

**自由现金流：**本季度公司产生自由现金流 490 亿美元，创历史新高。

**股东回报：**本季度公司向股东返还了创纪录的 200 亿美元 资金；全年计划将 50% 的自由现金流返还股东。具体措施包括：季度股息从 0.01 美元上调至 0.25 美元，新增 800 亿美元股票回购授权，叠加当前剩余 390 亿美元额度，公司将持续投资技术研发和生态建设，培育市场发展，推动长期增长，最终为股东创造超越市场和同行的总回报。重要的是，我们将继续秉持战略性和纪律性的投资原则，始终致力于向股东返还资本。

#### FY27Q2 展望

**营收：**预计总营收为 910 亿美元，上下浮动 2%，环比增长主要由数据中心业

务驱动。与上一季度一致，我们的业绩展望未纳入来自中国市场的任何数据中心计算业务营收。

**毛利率：**GAAP 毛利率预计为 74.9%，Non-GAAP 毛利率预计为 75%，均上下浮动 50 个基点。全年来看，我们预计毛利率将维持在 75% 左右。随着 Vera Rubin 平台的转型推进，我们将及时向市场更新进展。

**费用：**GAAP 营业费用预计约为 85 亿美元，Non-GAAP 营业费用预计约为 83 亿美元，2027 财年全年，Non-GAAP 营业费用预计同比增长 40% 高位，因我们将持续投资于不断扩大的市场机遇。**税率：**排除一次性项目和税务环境的重大变化，2027 财年 GAAP 和 Non-GAAP 税率预计在 7% 至 19% 之间。

## 总结

这是非凡的一季度，需求呈抛物线式爆发，智能体 AI 已经到来。AI 能创造真实价值，Token 开始盈利，模型厂商竞相扩产。

AI 时代，算力就是收入、就是利润，英伟达是这个时代的核心平台。我的五大核心观点：

英伟达是唯一运行所有前沿 AI 模型的平台，叠加 Anthropic 合作，前沿 AI 份额持续增长覆盖所有超大规模云，支撑其核心数据处理、机器学习、内部 AI 服务与公有云需求全栈 AI 工厂方案+全球生态，独家覆盖 AI 原生云、主权算力、企业 / 工业本地新市场，CUDA 生态延伸至边缘：机器人、自动驾驶、医疗仪器、AI 基站，实体 AI 是下一波浪潮。

Vera CPU：全球首款智能体 AI 专用 CPU，打开 2000 亿美元全新 TAM，所有头部厂商均已合作。

世界正为智能体 AI 与实体机器人 AI 重构计算体系，英伟达站在这一转型的中心。我们用三十年打造统一架构、庞大生态、芯片-系统-网络-软件极致协同设计，为这一刻做好准备。

## Q&A

**Q：贵公司调整业务分部口径的考虑是什么？两个细分市场在竞争格局上有何差异？提到的 200 亿美元 CPU 数字应如何理解？两类客户中的分布情况如何？**

**A：**首先纠正一下，是将季度股息从每股 0.01 美元上调至 0.25 美元，这额外的 5 美分对大股东来说应该很重要。关于业务分部的调整，我们希望帮助大家更好地理解我们的业务。AI 和计算本质上具有高度的多样性，体现在以下几个维度：

第一，AI 本身的多样性。语言模型是其一，但根据不同行业，AI 可以呈现为制造业和工业机器人领域的 3D 图形、生命科学领域的蛋白质研究、生命科学或材料科学领域的小分子化合物、物理科学领域的物理学问题（无论是能源领域、科学实验室还是高等教育领域）等等。AI 的形态本身极其多样。

第二，应用场景的多样性。AI 可能应用于企业级场景、能源行业、制造业等。

第三，部署位置的多样性。可以在超大规模云中运行，也可以在 AI 原生云（全

球各地涌现的 AI 原生企业网络) 中运行，还可以在企业本地部署、工厂车间的工业现场、超级计算中心，以及边缘端运行。边缘端包括大家所熟知的自动驾驶汽车和机器人，也包括制造业工厂内部日益庞大的计算机集群——无论是芯片制造厂、封装厂还是计算机制造厂等各种类型的制造工厂。展望未来，每一个基站、每一个无线网络都将成为 AI 驱动的无线网络。这就是“在哪里运行”的多样性。

第四，治理方式的多样性。有些由公有云运营，但有些因工业监管原因不能部署在受监管的云环境中，有些受限于机密计算的要求，有些则出于国家安全的考虑。不同类型的数据中心必须以不同的方式建设。

英伟达的独特价值在于：我们是唯一一家构建所有技术组件的公司。我们以极致的协同设计方式、端到端全栈的方式进行开发，但同时我们将平台开放，使其能够集成到各种不同的环境中。然而，有些环境（例如企业级用户）需要的恰恰是一家公司提供所有已协同运作的技术组件，这样它们就无需自行构建，而是直接购买并运营即可。因此，数据中心市场存在许多不同的细分市场，在这些市场中，英伟达的全套解决方案——全栈集成但仍保持开放性的产品交付方式——显得尤为重要。

基于此，我们将所有上述维度归纳为三大板块（以求用最简单的方式划分）：

第一大类——超大规模云（Hyperscale）。在这个板块中，我们的运营方式有三种：其一，帮助超大规模云商加速其数据处理和机器学习工作负载；其二，加速和支持其内部的 AI 处理；其三，将大量英伟达生态系统的业务引入其公有云平台。

第二大类——AI 原生云、企业本地部署、工业本地部署和主权 AI（ACIE）。这个板块正在以极快的速度增长，因为每个行业、每个国家、每家公司都需要 AI。然而，每个主体都希望以不同的方式建设 AI。我们提供完整解决方案的能力，大大降低了建设门槛，甚至使得原本不可能的事情变为可能。

第三大类——机器人边缘（Robotic Edge）。过去的计算主要是关于个人计算，而未来的计算将是关于个人 AI。个人 AI 的其中一个例子就是自动驾驶汽车——它本质上是一辆作为你个人 AI 的机器人系统。未来还将出现各种不同类型的机器人系统，甚至包括我之前提到的基站无线网络，本质上也会成为一种机器人系统。

这就是我们如此划分业务的原因，这是理解我们业务的最简洁方式。每一类业务在技术栈、操作系统和运营方式上都有很大不同，我们的市场进入策略也截然不同。超大规模云商仅五六家，市场进入最简单；而其余市场代表着全球约 25 万家公司，市场进入极为复杂和多样化，对 AI 的理解也必须极其广泛和深入。正如各位所知，英伟达拥有全球最庞大的加速库套件——从计算光刻、流体力学、粒子物理到分子动力学等等。所有这些库对于我们覆盖第二类和第三类所代表的垂直行业至关重要。总而言之，随着我们的业务已经演进并增长到如此庞大的规模，通过这样的分部方式来帮助大家更好地理解我们的业务运作是很有必要的。

关于 CPU 数字的补充，200 亿美元为独立 CPU 销售收入。Vera 共有四种使用方式：配套 Vera Rubin、独立 CPU、搭配 CX9 的存储方案、搭配 CX9 的安全和机密计算方案，预计整个生命周期内都将供应紧张。

**Q: 本季数据中心业务（除中国外）增长约 120%，而下季指引约 100%。许多分析师预测今年超大规模云商 CapEx 将增长 90%-100%。您此前谈到数据中心支出到本十年末仍有望达到 3 万亿-4 万亿美元。公司是否认同收入增速将快于超大规模云商 CapEx 的观点？您是否仍认为超大规模云商 CapEx 在今年之后仍将保持高速增长？**

**A:** 首先，我们的增速应当快于超大规模云商的 CapEx 增速。原因正可由我刚才描述的业务分部来说明。我们的数据中心业务有两大组成部分（实际不止两类，但为便于理解，简化为两类）。

第一类就是超大规模云商，也就是你刚才谈到的超大规模 CapEx 的对应方。今年这一规模已达 1 万亿美元，我完全有理由预期它将继续增长，这背后有根本性的良好原因——这就是未来计算的运行方式。如果他们不拥有算力，就无法产生营收。道理非常清楚：算力即营收，算力即利润。世界正在改变。过去的软件、SaaS 并不需要如此庞大的算力，但 AI 需要巨大的算力，同时也带来了远超以往的变现能力。这就是为什么我们看到 Anthropic 和 OpenAI 等前沿 AI 公司正以惊人的速度增长。它们在一个月实现的的增长，某些 SaaS 公司过去可能需要十年才能完成，这本身就说明了很多问题。因此，第一类市场——超大规模云的 CapEx 当前为 1 万亿美元，并正朝着 3 万亿-4 万亿美元的方向增长。

第二类是所有的 AI 原生云。它们遍布各地、呈区域性分布，全球各地的初创企业为其提供支持。还有企业级市场——全球约 25 万家企业，其中许多将不得不或希望为自己建设并运营 AI 工厂。许多工业公司别无选择，只能将计算机部署在业务现场、在行动发生的地方。你不可能把那种场景放到云上去运行。它必须每一次都可靠、快速地做出响应。无法想象一座芯片制造工厂连接到云服务商去运行，这完全讲不通。主权 AI 云同理。因此，存在一整个类别的数据中心市场，半定制芯片在此根本不适用——因为这些数据中心的业主需要的是购买系统并自行运营，而非自行设计或组建。因此，第二类市场极其多样化，不是由五六家或七家公司贡献我们第一类市场的收入，而是由成百上千、未来将达数十万家公司组成，每家的部署规模相对较小。这类市场将继续以惊人的速度增长。当我谈论物理 AI，谈论全球 100 万亿美元、过去 30 年未被 IT 深刻影响的经济体即将被 AI 改造时，我指的正是这个市场。第二类市场正在极快增长。

我们在其中的份额当然是非常高的。我们在服务这一市场方面具有相当的独特性。我们的平台在构建时采用垂直集成的方式，确保一切协同运作，但随后我们又可将其解耦，使人们能够按照自己想要的配置进行采购和组装。然而，这第二类市场目前被资本市场理解得相当不充分，因为其中包含太多小型公司，每家公司的部署规模相对一家超大规模云商而言显得很小。

因此，当你审视这一业务分部及各部分的规模时，可以看到：一方面，我们在超大规模云市场中的份额正在增长，因为我们获得了 Anthropic 这一新合作伙伴的大力支持，并正在帮助他们在未来几年大幅扩大算力容量。另一方面，在第二类市场，由于我们拥有的平台解决方案，极少有公司能够触及这一领域，而我们的敞口极为独特。

**Q: 随着 Vera Rubin 即将量产，公司对前沿模型更新以及优化多样化 AI 工作负载的新技术有着深刻洞察。鉴于投资者高度关注贵公司在推理市场的份额，请问您如何看待 Vera Rubin 及贵公司的极致协同设计对 2026 年底至 2027 年推理市场份额的影响？**

A: 我们正在推理市场中获得份额，而且增长速度非常非常快。原因如下：

今年前沿模型公司的数量有所增加。出现了 Cursor、Perplexity，以及一些新的模型公司如 TML、Reflection 等等。前沿模型厂商的阵容在扩大，同时我们今年将 Anthropic 纳入了合作伙伴关系。他们正在以极快的速度扩张。我们已与他们合作，在 Azure、AWS、CoreWeave 等多个平台为其锁定算力容量。还有其他一系列我们已宣布的平台正在为他们上线。今年和明年我们将为 Anthropic 上线的算力容量将相当可观、非常巨大。需要指出的是，在此之前我们在 Anthropic 的覆盖率几乎为零，直到最近才发生改变。因此，我们在推理市场的份额正以极快的速度增长。

Vera Rubin 的起步势头将比 Grace Blackwell 更加成功。我想不出有哪一家前沿模型公司不会从首发阶段就全力拥抱 Vera Rubin——每一个前沿模型公司都会从最初就采用 Vera Rubin。而在 Blackwell 时代，情况并非如此。因此，Vera Rubin 的开局极为强劲，它必将比 Grace Blackwell 更加成功。

此前关于推理问题的所有解释，主要是聚焦在超大规模市场领域。别忘了，还有一个整个第二类的 AI 数据中心市场，我们几乎是以近乎独有的方式在服务这个市场。这个板块非常碎片化，需要一个高度集成的平台解决方案以及非常庞大的市场覆盖体系。在这个市场板块中，所有的推理、几乎绝大多数的推理算力都是由英伟达提供的。此外，在物理 AI 领域，英伟达目前也几乎是唯一一家服务该市场的公司，我们在物理 AI 领域已经深耕了很长时间，这一板块同样在增长。因此，我们在推理市场的份额正在快速攀升。

Q: 想请教一下贵公司在 CPX、LPX 等定制化产品方面的市场接受度。您此前曾提到 Groq 可能占据约 20% 的市场份额，因此我推测 LPX 应该也取得了不错的进展。能否谈谈这方面的进展，以及 LPX 如何融入公司更广泛的平台战略？

A: LPX 的设计定位是低延迟和高 token 速率，但其吞吐量较低、模型规模容量有限，并且其上下文处理能力——即吸收大量上下文的能力（例如用于软件编程、Agentic 工作负载所需的能力）——也相对较弱。因此，其核心挑战在于应用场景并不广泛。

LPX 面向的是那些拥有较大规模且不同类型 token 服务组合的客户。对于那些 token 速率极高的服务，或许这些服务相当高端、客户数量不算庞大，但 token 并发速率极高。这与我们此前的表述完全一致，我也仍然维持这一判断。因此，我预计 LPX 及其他基于 SRAM、专注于解码端高 token 速率生成的加速器，在未来相当一段时间内仍将是利基产品。

正如各位所知，Grace Blackwell 和 Vera Rubin 支持 AI 的完整生命周期——从数据处理（为训练做准备）、预训练、后训练强化学习，一直到推理。Grace Blackwell 是全球完成这一切的最佳平台。而在某些特定场景下，只要客户（即服务提供商）已经拥有可提供的高 token 速率服务，我们就可以为其叠加一个 LPX 模块，使其能更好地交付该服务。

至于份额是 20% 还是 10%，这取决于我们处于 AI 发展的哪个阶段。我认为今天这一比例远低于 20%。未来随着高端 token 服务的发展，这一比例有可能达到 20%。我们已做好准备与各类服务提供商合作来赋能这一能力，我对此充满期待。

**Q:** 当前市场对 Agentic 应用中的 CPU 充满期待，有很多关于 CPU 数量实际上将超过 GPU 的讨论。首先，这是否是增量工作负载，还是会蚕食 GPU 本应完成的工作？其次，您给出的 200 亿美元数字是指独立 Vera CPU，还是已包含在 Vera Rubin 配套的 CPU 中？能否请您阐释一下 CPU 与 GPU 的角色定位——是替代关系还是增量关系？以及如何将 200 亿美元这一数字与公司通常的销售模式（即 CPU 作为 GPU 的一部分）联系起来理解？

**A:** 200 亿美元是指独立 CPU 的销售收入。需要说明的是，Vera 一共有四种使用方式。让我从大家最熟悉的一种开始：

第一种是 Vera Rubin 平台。我们将销售数百万颗 Rubin GPU，每两颗 Rubin 搭配一颗 Vera CPU，当然我们对这部分 CPU 也进行了恰当的定价。这是第一种使用场景。

第二种是 Vera 独立 CPU。

第三种是 Vera 搭配 CX9 及存储软件栈。

第四种是 Vera 搭配 CX9 及安全、计算隔离与机密计算软件栈。

以上每一种使用场景都基于 Vera 构建。我的判断是，在 Vera Rubin 的整个生命周期内，我们都将面临供应紧张的局面。无论如何，回到你的问题——200 亿美元是独立 CPU 的收入。

关于 CPU 的使用：Agent 本质上就是人们所说的“执行框架”（harness）。这个框架可以是 OpenClaw、Hermes 等形式。Claude Code 本质上就是围绕 Claude 模型（Opus 模型）的一个框架，OpenAI 的 Codex 则是围绕 GPT-5.5 模型的一个框架。这些框架提供 I/O、编排、内存管理、工具调用等功能——连接到浏览器、C 编译器、Python 编译器等各种工具。这些框架运行在 CPU 上，工具调用也运行在 CPU 上。例如，如果 AI 要进行搜索或使用浏览器，这些操作都会在 CPU 上运行。

全球目前有 10 亿人类用户。我的判断是，未来全球将有数十亿个 AI Agent——不是今天，而是逐步增长至此。这数十亿个 Agent 都将使用工具，这些工具可以类比为 PC，就像今天人类使用 PC 一样。未来，你会有一个 Agent 在使用 PC。因此，如果你试着沿着这个思路去想——当前 Agent 的数量可能只有几十万个，但未来将达到数十亿——我可以想象它们每一个都拥有一台可供使用的“PC”。更重要的是，每一个 Agent 都将生成子 Agent（sub-agents），每一次生成都需要进行推理——那是“思考”发生的地方。

所有的思考都在 GPU 上完成。所有的编排则本质上运行在 CPU 上。子 Agent 被生成出来后，当它们进行思考时，使用的也是 GPU。当 Agent 使用仿真器时，那些可以在 CPU 或 GPU 上运行——这也是为什么我们与 Cadence 和 Synopsys 紧密合作，致力于加速全球所有工具的原因。我们正在加速全球所有的工具、数据处理引擎和数据库引擎，因为 Agent 会使用这些工具，而它们对等待的容忍度远低于人类，它们希望事情能快速完成。因此，我们正在将全球所有工具加速运行在 CUDA 上。各位可以看到我们正在做这件事——我们与 Cadence、Synopsys、Siemens 以及 Adobe 等公司的合作，正是因为我们希望让全球所有工具都运行在 GPU 上，因为客户已经拥有 GPU，而且在 GPU 上运行要快得多。所以，我们将需要更多的 CPU。

Vera 正是为 Agentic 工作负载而设计的 CPU。过去的 CPU 设计理念是拥有众多核心，以便于租用——人们租用的是“核心”。但 Agent 不会租用核心，它们只

希望工作能快速完成。过去的经济模式是“每核心多少美元”，那是过去云计算的经济学。未来 AI 的经济学是“每美元多少 token”或“每 token 多少成本”。因此，未来我们需要做的是尽可能快速地生成和处理 token，而这正是 Vera 所擅长并表现卓越的领域。因此，我们预期 Vera 将取得巨大成功。

归根结底，我们正在为 AI 建设基础设施。AI 需要极其强大的存储——这就是我们打造 STX 的原因；需要极其出色的网络——这就是我们打造 Spectrum-X 的原因；当然还需要卓越的 GPU 和推理能力——这就是 NVLink 72 的意义所在；需要极其强大的安全和机密计算能力——这就是为什么 Vera Rubin 是全球首个端到端机密计算平台；还需要强大的 CPU。我们已经全面覆盖了这一切。

**Q: 我想回到业务分部的问题。首先，AI 原生云（Neo clouds）在两个细分市场中如何归类？是属于超大规模云还是 AI 云？我倾向于认为是后者，但不太确定。其次，从规模上看，两个板块现在规模已基本相当。您似乎暗示 AI 云市场未来的增速可能比超大规模云更快，这是您想表达的意思吗？还是说您认为两个板块将保持相似的增速？**

**A:** 首先，你的判断是正确的——AI 原生云确实属于第二类市场（ACIE）。原因在于：AI 原生云不制造芯片，也不设计自己的芯片。他们不想、也做不到将互不相关的部件自行组装成一个 AI 工厂。他们对“从部署到产出首个 token 的时间”容忍度极低，同时对架构的要求极高——需要一个能够运行所有模型、承接来自任何地方的客户的架构，即架构需要具备极强的“可租赁性”。

这就是为什么英伟达的架构对他们而言如此完美。我们提供每一个组件，我们未提供的部分由合作伙伴生态系统补足，且所有组件都完全集成、协同运作。能够从 AI 原生云处租赁算力的客户数量极其庞大——基本上涵盖了全球每一位 AI 开发者、每一家 AI 原生初创公司、SaaS 公司、企业公司和工业公司。因此，我们的计算架构是全球可租赁性最强的计算平台：性能最强、组装最易、可租赁性最高、总体拥有成本（TCO）最优、融资也最容易。所有这些属性都极其契合 AI 原生云的独特需求，这与 OEM 厂商及大型企业等需求非常相似。因此，我们将 AI 原生云归入第二类市场。

关于增速差异：如果你观察第二类市场的发展轨迹，会发现它是在超大规模云的 AI 生态发展起来之后才开始增长的。超大规模云之所以率先发展 AI，有很多原因：它们拥有卓越的计算机科学能力、出色的数据中心运营能力，并且主要聚焦于消费级应用——在这些场景中，即便 AI 不完美也并非致命问题，只要能改善服务就足够了。而对许多其他应用（工业应用、企业应用）而言，在 AI 变得足够强大、能真正完成高生产力的工作并安全可靠地执行之前——即能够真正产生实际影响和收入之前——这些应用并不会真正被采用。因此，可以预期第二类市场的发展会比超大规模云更慢，这一点在数据中也能看到。

然而，长期来看，工业和企业市场显然才是未来经济价值所在，因为它们代表着全球约 50 万亿至 80 万亿美元的经济规模，而且借助 AI，这一规模还将进一步扩大。因此，我预计第二类市场终将成为更大的市场。

短期来看，在未来数年内，两个板块无疑都将以极快的速度增长。我预计第二类市场的增速仍将更快一些，但两者都将高速增长。而在未来五年内，我期待物理 AI 和机器人板块也将实现爆发式增长。

**Q:** 回到此前 GTC 大会上的讨论，您曾提到对 Blackwell 和 Rubin 平台累计营收有 1 万亿美元的可预见度，但我认为这一数字未包含 LPX、Rubin CPX 以及 Vera CPU 机架等产品。能否请您谈谈，Vera CPU 是否会成为超越这 1 万亿美元的最大增量来源？此外，公司是否在考虑包括 CPU 在内的其他产品组合，以获取更大的整体 TAM 份额？

**A:** 在 1 万亿美元之上，增量来源按重要度排序如下：

第一，前沿 AI 模型份额的持续增长。我预期我们将获得更多份额，这部分将持续增长。

第二，独立 Vera CPU，这部分我们完全没有包含在此前 1 万亿美元预期之内。因此，我预计这将是第二大增量来源。Agentic 系统的 TAM 当然非常庞大，我们所有的客户都对 Vera 充满期待，我们将销售大量的 Vera。

第三才是 LPX。正如我此前解释的，LPX 因其 SRAM 架构而具有非常低的延迟和极高交互性的优势，但其吞吐量和上下文处理能力也相对有限，这是基于 SRAM 的系统的固有特性。

关键在于组合：通过 Vera、Vera Rubin 和 LPX 的组合，我们将能够覆盖从预训练、后训练到推理、Agentic 系统的 AI 全谱系。

**Q:** Colette，您在准备好的发言中提到 GB300 是公司历史上爬坡最快的产品。我们应该如何看待 Vera Rubin 与此标杆的对比？它显然在芯片层面采用了全新架构，但机架形态类似。这是否意味着 Vera Rubin 的爬坡斜率将与 GB300 类似，还是因采用全新芯片而更加平缓？

**A:** 是的，我们此前已多次表示将在下半年推出 Vera Rubin。我们将在 Q3 开始初步出货，将各个组件逐步整合在一起。进入 Q4 后，我们预计爬坡将持续加速。目前还很难说哪个产品的爬坡会更快，但需要强调的是，我们的需求已经规划就绪，采购订单已经到手，几乎所有主要客户都已准备就绪。这些系统极其复杂，我们需要将其中的各个系统全部整合到位。因此，我认为关键在于将所有这些不同系统组织生产并推向市场所需的时间，除此之外别无其他制约因素。

现在下结论还为时过早，但可以确定的是，我们将在 Q3 启动并在 Q4 持续加速爬坡，而明年 Q1 无疑也将是一个非常巨大的季度。



## 分析师承诺

负责本研究报告的每一位证券分析师，在此申明，本报告清晰、准确地反映了分析师本人的研究观点。本人薪酬的任何部分过去不曾与、现在不与、未来也将不会与本报告中的具体推荐或观点直接或间接相关。

## 评级说明

报告中所涉及的投资评级采用相对评级体系，基于报告发布日后 6-12 个月内公司股价（或行业指数）相对同期当地市场基准指数的市场表现预期。其中，A 股市场以沪深 300 指数为基准；香港市场以恒生指数为基准；美国市场以标普 500 指数为基准。具体标准如下：

### 股票评级

强烈推荐：预期公司股价涨幅超越基准指数 20%以上

增持：预期公司股价涨幅超越基准指数 5-20%之间

中性：预期公司股价变动幅度相对基准指数介于±5%之间

减持：预期公司股价表现弱于基准指数 5%以上

### 行业评级

推荐：行业基本面向好，预期行业指数超越基准指数

中性：行业基本面稳定，预期行业指数跟随基准指数

回避：行业基本面转弱，预期行业指数弱于基准指数

## 重要声明

本报告由招商证券股份有限公司（以下简称“本公司”）编制。仅对中国境内投资者发布，请自行评估接收相关内容的适当性。本公司不会因收到、阅读相关内容而视为中国境内投资者。本公司具有中国证监会许可的证券投资咨询业务资格。本报告基于合法取得的信息，但本公司对这些信息的准确性和完整性不作任何保证。本报告所包含的分析基于各种假设，不同假设可能导致分析结果出现重大不同。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出价，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。除法律或规则规定必须承担的责任外，本公司及其雇员不对使用本报告及其内容所引发的任何直接或间接损失负任何责任。本公司或关联机构可能会持有报告中所提到的公司所发行的证券头寸并进行交易，还可能为这些公司提供或争取提供投资银行业务服务。客户应当考虑到本公司可能存在可能影响本报告客观性的利益冲突。

本报告版权归本公司所有。本公司保留所有权利。未经本公司事先书面许可，任何机构和个人不得以任何形式翻版、复制、引用或转载，否则，本公司将保留随时追究其法律责任的权利。