星环科技 20260907

星环科技：GPU 原生数据库实现百倍性能突破，2027 年有望放量

摘要

- 星环科技 2026 年上半年毛利率提升 1.5pct 至 47%，金融与政府行业为核心收入来源，占比合计超 60%。
- 公司实现 GPU 原生数据库全链路执行，TPC-DS 1TB 测试性能比双路 CPU 快 100 倍，并发测试快 70 倍。
- 对比 Databricks 同规格产品，公司 GPU 方案性能领先 16-120 倍，通过缩短处理时间显著降低客户总成本。
- 产品集成关系型、向量及图数据库，单卡向量构建提速 40 倍，查询提速 20-200 倍，适配海光 DCU 实现信创替代。
- 海外市场采用 Serverless 云服务模式，联合 AWS 与英伟达推广，目标切入东南亚及中东 10% 的存量客户群。
- 2026Q4 预计签约大型制造标杆客户，2026 年重点为标杆树立与试用，2027 年有望迎来业务放量增长。
- 针对 AI Agent 循环推理需求，公司数据库解决“内存墙”瓶颈，支持 KVCache 等多模态记忆存储与毫秒级响应。

Q&A

请介绍一下星环科技 2026 年上半年的整体经营情况和财务表现？

2026 年上半年，公司在收入、盈利、现金流及新研发技术方面均表现积极。财务方面，上半年营收达到 1.74 亿元，同比增长近 14%。净亏损持续收窄至近 1 亿元，同比缩减近 20%。从收入结构看，金融行业仍是第一大行业，占比约 32%；政府行业位居第二，占比近 29%；交通、能源、制造等行业占比均在 5%上下，显示了客户群体的多样性和健康性。公司的毛利率持续提升，增加了 1.5 个百分点，达到近 47%。在费用管控方面，研发投入为 1.16 亿元，同时销售费用和管理费用同比分别下降近 10%和 15%，体现了组织效率的持续优化。

公司在上半年取得了哪些关键的技术落地和生态扩展成果？

在业务层面，2026 年上半年的核心是技术落地与生态扩展。公司的 GPU 原生认知数据库实现了全链路的 GPU 执行，能够在 GPU 上完整运行 TPC-DS 基准测试中的全部 99 个查询，无需回退至 CPU，这是对传统数据库架构的重要突破。公司的关系型数据库、向量数据库和图数据库均完成了与英伟达 GPU 的适配。其中，向量数据库还完成了与海光 DCU 的适配，相比传统 CPU 方案性能有大幅提升，为 AI 推理和实时分析场景的规模化部署奠定了基础。在生态与市场方面，公司与合作伙伴联合打造了 SAP HANA 和 CDH 的替换方案，并已开始在 BI 分析领域向客户推广。此外，ArgoDB 在信创领域通过了国家安全可靠测评认证，为国产化替代以及数据与人工智能融合应用提供了技术支撑，也为下半年及 2027 年的信创业务增长储备了动能。

为什么需要发展 GPU 原生数据库，以及星环科技在该领域取得了哪些技术突破和性能优势？

未来数据库主要服务于 AI，而 AI 调用数据时面临两大瓶颈。第一是“内存墙”问题，GPU 算力虽强，但数据加载速度受限于显存、内存和存储，成为性能瓶颈。第二是 AI Agent 在执行循环任务（分解任务、调用数据库、分析数据、调用工具）时，传统数据库处理效率低下。例如，使用传统方案处理 100GB 数据进行 TPC-DS 基准测试，最快需要 5 分钟，若循环 100 次则耗时 500 分钟，导致 AI Agent 运行缓慢且 GPU 利用率极低。尽管提高 CPU 与 GPU 的配比（如从 1:8 提升至 1:1）能有所缓解，但 CPU 核心数量（通常几十至二百余核）远不及单张 GPU（1 至 2 万个核心），性能差距巨大。将数据库迁移至 GPU 运行，可利用其数万核心实现数百倍的性能提升，如同步行与飞机的速度差异。同时，GPU 显存带宽（如 HBM3 的 8TB/s）远超内存（200-400GB/s），也能显著加速数据处理。基于此判断，并预见到未来计算机体系架构将围绕 GPU 重构，公司从 2025 年开始重构数据库。2026 年初，实现了单卡突破，数据库在单张 GPU 上运行，性能比两颗 CPU 快约 20 倍。到 2026 年五六月份，实现了多卡集群突破，可在一台机器或一个机柜（如 72 张卡）内完整运行数据库，完全不依赖 CPU 和内存。在 TPC-DS 1TB 基准测试中，性能比两颗 CPU 快 100 倍；在包含增删查改的完整并发测试中，性能快 70 多倍，实现了 1 到 2 个数量级的性能提升。

与 DataBricks 和 Snowflake 等行业领先者相比，星环科技的 GPU 原生数据库在性能上具体有何差异？

在性能对比上，使用两张 A100 GPU 卡运行 100GB 的 TPC-DS 基准测试，星环科技的方案耗时约 6 到 7 秒。相比之下，DataBricks 使用其 XL 规格的数据仓库（由 32 台虚拟机组成）运行相同测试，最快需要 5 分钟，首次运行则需 10 分钟。这显示出几十倍的显著加速比。尽管美国公司在 GPU 数据库领域已探索多年，但此前行业普遍认为 GPU 加速数据处理最多能快 5 倍，而星环科技实现了远超这一水

平的性能突破。

与 CPU 方案相比，公司基于 GPU 的数据库产品在性能和经济性方面有何具体优势，其产品功能构成是怎样的？

在性能方面，尽管英伟达发布的软件声称其在 TPCH 基准测试中能比 CPU 快最高 5 倍，且与英伟达合作近三年的经验显示平均加速比约为 1.5 倍，IBM 也公开宣布其方案能快 3 倍左右，但考虑到 GPU 相较于 CPU 更高的成本，1.5 至 3 倍的平均加速比并不具备经济性。公司的数据库产品是目前唯一能将数据库功能完整运行在 GPU 上，并实现 20 倍以上加速比的产品。 该数据库产品包含三部分功能：第一，处理结构化数据，支持传统关系型数据库的增删查改操作，性能加速比可达 20 倍至 100 多倍。 第二，集成了向量数据库和图数据库，用于 AI 的模糊检索。在使用单张 GPU 卡的情况下，向量数据库的构建速度可提升约 40 倍，查询速度可提升 20 倍至 200 倍。 第三，将 Agent Memory、Agent 上下文及对话历史等存入数据库。 这三部分功能全部在 GPU 上运行并获得显著加速。在 AI 推理的 OfficeQA 基准测试中，通过将美国财政部自上世纪初至今的所有财报、公报、国债及财政开支等数据导入数据库进行端到端分析，整体处理速度可提升约 6 倍，主要得益于将数据处理环节从 CPU 迁移至 GPU，解决了 GPU 等待 CPU 处理数据的瓶颈。

公司计划在海外市场推出的云服务，其商业模式、定价策略以及与 Databricks 等竞品的对标情况是怎样的？

公司计划在海外市场推出的云服务，其模式与 Databricks 完全一致。我们定义了从 Small、Medium、Large 到 X-Large 等不同规格的数据仓库。在每个规格中，尽管我们处理的工作集数据量与 Databricks 相近，但我们的性能表现更优。Databricks 采用由 1 到 64 个节点不等的 CPU 集群，而我们仅使用一台虚拟机，配置不同型号的 GPU 卡，并支持多卡扩展。在各个规格上，我们的性能比 Databricks 快 16 倍至 120 倍不等。 在定价策略上，我们采用按时收费模式。虽然我们的单位时间价格因使用 GPU 而略高于 Databricks，但由于运算速度极快，总处理时间能缩短至几分之一甚至十分之一，从而为客户显著降低总成本，实现了高性价比和优越的交互体验。

公司在海外市场，特别是东南亚和中东地区，计划如何与 Databricks 和 Snowflake 等公司竞争，核心竞争优势是什么？

随着 AI 应用的加速部署，特别是 Agent 产生的数据量急剧增加，数据基础设施的需求正在放大。Databricks 和 Snowflake 的营收快速增长，部分得益于 AI 数据分析带来的需求。我们认为，继 AI 编程之后，AI 办公和 AI 数据分析将
一个投资回报率被验证为正的领域，这将带动 AI 和数据库部署的新一轮
我们在国际市场的核心竞争优势体现在两点： 第一，我们即将推出的云
用与 Databricks 相似的 Serverless 方案，用户无需关心底层硬件，可在

人在草木间

一分钟内启动 GPU 数据库并快速导入数据进行分析，大部分查询能在数百毫秒内完成。在提供相似用户体验的同时，我们能以快几十倍的性能处理相同数据集，从而加速客户的 AI 负载并降低总成本。 第二，我们通过 AI 赋能数据分析，极大地降低了用户的使用门槛。在东南亚和中东等地区，数据科学家和工程师较为稀缺，我们的方案允许业务人员直接使用 AI 进行数据分析、数据库操作及运维优化，替代了原本需要雇佣数据科学家的成本。 我们的目标是获取东南亚和中东市场现有客户群体的 10%作为切入点。

公司产品的落地节奏如何，是否有明确的标杆客户以及预计的商业化进展？

目前，产品已有私有化部署的客户正在使用。预计在 2026 年第四季度，会有一家大型知名制造业客户签订合同。同时，交通、零售及生物制药等行业的公司也正在试用，并将陆续产生订单。2026 年的主要目标是树立行业标杆，让客户充分体验产品优势。 在渠道方面，我们正与 AWS 合作，将海外云服务构建在 AWS 之上，并由 AWS 协助进行市场推广。同时，我们也在与英伟达开展联合营销，共同寻找标杆客户。目前已有部分客户在试用中。 总体节奏规划是，2026 年主要致力于建立标杆案例，期望在 2027 年实现业务放量增长，届时收入将体现出增长趋势。

针对近期业界关注的循环推理架构（如“阿雀”模型），公司的产品能提供怎样的技术支持和价值？

“阿雀”这类模型的发布，其特点在于将模型的思考过程内部化，并能自动化执行编程、日常办公等复杂工作流，这意味着 AI 需要使用更多工具并产生海量数据。企业的工作流高度依赖数据，因此未来 AI 调用数据库的频次和产生的数据量将迅速提升。 这种趋势对数据库提出了新的要求。首先，数据库需要能处理多样化的数据类型，包括短期记忆（如 KV Cache）、中期记忆（如对话历史、任务中间结果）和长期记忆（如知识库）。这些都将成为 AI Agent 不可或缺的基础设施。其次，数据库性能必须与 GPU 的速度相匹配。传统分析引擎处理 100GB 数据需要数分钟，在 AI 数百万次调用的场景下是无法接受的。 我们的产品能够满足这些新需求，为 AI Agent 提供高性能、多模态的数据处理能力。随着 AI Agent 执行的任务日益复杂，对我们这类高性能数据库的需求将大幅提升，这对我们而言是一个重要的市场机遇。