

证券代码：688031

证券简称：星环科技

## 星环信息科技（上海）股份有限公司

### 投资者关系活动记录表

编号：2026-002

投资者关系活动类别	<input type="checkbox"/> 特定对象调研 <input type="checkbox"/> 分析师会议 <input type="checkbox"/> 媒体采访 <input type="checkbox"/> 业绩说明会 <input type="checkbox"/> 新闻发布会 <input checked="" type="checkbox"/> 路演活动 <input type="checkbox"/> 现场参观 <input type="checkbox"/> 电话会议 <input type="checkbox"/> 其他（请文字说明其他活动内容）
参与单位名称	长江养老、南方基金、广发基金、鹏华基金、中信证券、华创、银河基金、诺安基金、国泰海通、国金证券、怀信基金、姚河泾、安联投资、天襄资本、中财资本、怀信基金、东方证券、东吴证券、广发证券、睿郡资产、大成基金、浙商证券、中金资管、运舟资本、九方智投、汇证资产、时间资本、国金证券、上银基金、南土资产、目汐资本、卓盈投资、浙商资管、华永信资本、贝溢投资、深圳展博投资、上海老渔民投资、景顺长城、巨鼎私募、翊安投资、骐楷资产、国投瑞银、惠升基金、渤海证券
时间	2026年4月1日 10:00-12:00
地点	公司11楼
上市公司接待人员姓名	董事长、总经理：孙元浩先生 董事、董事会秘书、财务总监：李一多女士 投融资总监、投资者关系负责人：赵梦笛女士 证券事务代表：王诗瑶女士
投资者关系活动主要内容介绍	<b>一、公司最新产品研发进展介绍</b>  公司从“以数据为中心”到“AI 基础设施软件”，到当前拟开辟以 Agent 为中心的增量市场，这个市场是基于新硬件架构变革重构软件产品支撑的。过去的硬件架构以 CPU 为主，现在所有的硬件创新都围绕着 GPU 来做，包括 HBM、HBF、高速 SSD 等。AI 推理过程中面临存储墙（memory wall）的问题，数据从内存装载到 GPU 里面受限于带宽影响，让 GPU 在等待。现在 agents 都跑在 GPU 上，通过 GPU 数据库可以较好解决 memory wall 的问题，加速推理过程。数据处理也可以充分利用 GPU 的多核的特性，意味着需要根据新的硬件架构重构数据库软件。

公司尝试在 GPU 上加速数据处理效率，获得了让人振奋的性能进展：其中在 SQL 上，在关系型数据库的标准 TPC-DS（全 99 个场景）150GB 数据规模的测试中，性能较 128vCore 传统 CPU 性能提升约 26 倍，相关数据在英伟达 GTC 大会数据加速专场中被公开提及。在向量检索上，我们也用 GPU 来加速。构建索引的时间上，在一张 GB300 上的构建性能相比在 96vcore 的 CPU 上面提速约 40 倍。在向量查询任务中，基于 GPU 加速的向量检索在不同的数据索引项下提升约 20-200 倍不等，这个加速的前提是把数据库索引放在显存里。目前我们也做到了仅留 5%的向量数据库索引放在显存里面，剩下的 offload 至内存和 SSD 上。在这种情况下用 GPU 做向量检索，也比传统的 CPU 处理方式快 12 倍左右。也就是说同样的任务，只需要原先 HBM 需求的 5%。谷歌 TurboQuant 也可以达到类似的压缩效果。通过将数据库索引 offload 至内存或 SSD 上，我们可以有更大的容量来存放索引。

我们拟发布的认知数据库，它的定位是完全基于 GPU 全新架构的数据库，主要为 AI 服务（未来数据库由 Agent 使用，而非人），满足 AI Agent 的短期/长期记忆存储、高效交互（与 AI Agent 同 GPU 部署降低延迟）、数据内容理解（提供自然语言接口）。功能构成上，包含了知识库（例如向量库、图数据库）+记忆系统（将所有 agent 和环境交互的内容都存下来）+关系型数据库（做数据分析和轻量级的交易）。这款高性能数据库的适用场景，主要针对做 AI 推理加速（agent 会多次调数据库，或者多个 agent 会调数据库，等待的时间太长。前端的数据处理工作我们来做，嵌在 AI 的推理管道里面）、时延敏感性应用（如量化交易、银行实时反欺诈）、探索性分析（如药物研发、能源勘探）、高复杂度高数据量的计算等。我们计划先向市场推出单机多卡的版本，单机可实现约 1TB 数据的分析，性能较传统架构提升几十倍。我们拟采用云服务和私有化部署两种方式。公司现有产品作为历史数据的存储、加工，认知数据库作为一个加速层，支持 AI Agent。

私有化部署模式下，我们已经有了数据工程、数据治理、知识工程等工具，这部分产品我们也在同步做 AI 化。我们的认知数据库还是以 agent 为中心，可以支持各种 agent、OpenClaw 等，我们作为底层数据支撑。现在企业在逐步部署 GPU，在 GPU 上可以直接部署我们的认知数据库。Agent 部署日益普及，需要高性能数据库来配合。

公有云服务模式下，我们会在公有云、NeoCloud、私有云上开展服务，面向三种类型的用途：（1）AI 驱动的数据工程、数据治理、知识工程，如果客户原来使用 Snowflake/Databricks 的产品，他们的数据存在 S3 上，我们的认知数据库也能直接进行读取，并实现更高的性价比；（2）推出一个给开发者用的工具集（Sophon LLMops），可以开发各种各样的 Agent、用 AI 来编程；（3）同时也提供 API，作为大模型能够直接访问的数据库，提供容器或者虚拟机实例，嵌入到 Agent 推理的链路中去。

## 二、关于公司 2025 年核心财务情况介绍

公司 2025 年营收 4.47 亿，同比增长 20.47%；综合毛利达历史新高 54.09%；归母净利润大幅收窄 30.6%；经营活动现金流净流出显著改善，由 2024 年 3.3 亿元缩窄至 2025

年度的 1.1 亿元。

1、毛利端，公司 2025 年主营业务毛利率约 54%，较 2024 年增加 3.45 个百分点。主要驱动因素为：（1）项目利润占公司考核的 50%，推动营业收入中高毛利的软件收入和维保收入占比提升；（2）公司软件产品成熟度持续提高，配套服务所需人力投入相应减少，标准化产品通过生态合作伙伴交付，同时项目执行与交付效率进一步提升。

2、公司现金流实现显著改善，主要得益于持续优化合同收款条款：通过设置付款进度达到 80% 方可授予永久授权 license，并按月开展客户信用管理，有效保障销售款项及时回笼、严控坏账风险。同时，公司目前付费客户累计达 1,800 家，维保服务收入稳步增长，且相关维保合同均约定季度预付模式，进一步加快资金周转效率。

3、费用端，公司 2025 年加强各部门费用管控，整体费用结构持续优化：销售费用绝对值较 2024 年下降约 22%，一方面公司超 70% 收入来自老客户复购，获客成本显著降低；另一方面公司采取更具性价比的市场推广形式，加强销售人员的考核频次，细化管理指标的颗粒度，提升销售人员人效。管理费用同比下降约 5%，主要系公司自研 ERP 系统线上应用，实现项目全周期成本、进度及回款一体化管理，大幅提升后台运营效率。研发投入（考虑费用化和资本化的部分）合计 2.5 亿元，相比 2024 年略有下降。公司在研发人员未增加的情况下，通过内部活水调配、扁平化管理及 AI 工具赋能，持续保障研发进度和效率。

#### Q&A

**问题 1：我们看到 GPU 适合并行计算任务，而 Agent 有比较复杂的工具调用，我们的认知数据库是否会涉及 CPU 和 GPU 的分工？**

**答：**技术原理上，公司通过 GPU 直连技术减少 CPU 与 GPU 交互，将绝大部分工作放在 GPU 上，仅让 CPU 做调度，甚至实现 GPU 直连存储不经过 CPU。设计原则是尽量在机器内完成操作，减少跨节点交互。

**问题 2：请问公司后面关于产品的规划、商业化落地的时间线。**

**答：**商业化策略上，预计今年下半年上线云版本，先进入海外市场，以高性价比吸引海外客户；同时，在国内寻找对延迟敏感或 AI 部署较多的客户，争取今年有标杆客户部署落地。

**问题 3：NV 关于 cuDF 和 cuVS 的发布对我们有什么影响？**

**答：**我们是基于 NV 的 cuDF 进行开发，部分索引也基于 cuVS，未来也会基于国产卡。

**问题 4：我们公司和大型模型公司的关系？**

**答：**大模型更偏向静态知识存储，核心是不断强化自身的推理与编程能力。公司提供的是外部动态知识库，用来存放实时经营数据、合同、视频等不断变化的信息，和大型模型内部的静态知识用途不同；同时这套系统还可以永久记住用户偏好，导致记忆体量持续变大。

	<p>因此，外部动态知识库与大模型内部知识是互补关系，我们还专门面向 AI Agent 记忆市场，提供短期、长期、永久记忆的完整解决方案。</p> <p><b>问题 5：公司海外市场拓展计划是什么？</b></p> <p><b>答：</b>我们目前在海外市场同步推进东南亚、中东及美洲区域的拓展，主要依靠本地渠道及其他合作渠道进行推广。业务上依托现有云资源搭建服务，主打两大优势：一是全新的增量市场和技术概念，为 Agent 提供数据处理服务；二是高性价比，在大幅提升性能的同时，有效降低整体使用成本。从各渠道反馈来看，市场反响整体积极。</p> <p><b>问题 6：HBF 出现后，公司认知数据库的产品迭代节奏如何？</b></p> <p><b>答：</b>HBF 与 GPU 之间的连接速度很快，适合存放模型权重，不过 KV cache 目前还是要放在外部。这对 GPU 数据库来说其实是个优势：现在数据放在显存里，后续可以迁移到 HBF，整体容量能进一步提升。产品迭代节奏上，单机版已开发完成，预计近期就会发布；今年下半年将推出多卡版本，计算能力会进一步增强，数据量也能实现线性扩展，数据仍存放在显存或内存中；到年底，随着硬件成熟，我们将采用 AI SSD，把数据和索引都下沉到 SSD 上。</p> <p><b>问题 7：后面公司会进行榜单测试吗？</b></p> <p><b>答：</b>目前公司的认知数据库为单卡版本。后续多卡版本发布后，我们计划完成 TPC-DS 1TB 的全量测试；若届时顺利通过，即代表数据库核心功能已完备。</p>
附件清单	无
日期	2026 年 4 月 1 日