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	☐特定对象调研　☐分析师会议 ☐媒体采访　☐业绩说明会 ☐新闻发布会　☑路演活动 ☐现场参观　☐电话会议 ☐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	**一、公司最新产品研发进展介绍** 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%；归母净亏损大幅收窄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。 ν 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。 ν 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。 ν 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。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。 **问题 5：公司海外市场拓展计划是什么？** **答：**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。 **问题 6：HBF 出现后，公司认知数据库的产品迭代节奏如何？** **答：**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。 **问题 7：后面公司会进行榜单测试吗？** **答：**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。
附件清单	无
日期	2026 年 4 月 1 日