证券代码：001339　　证券简称：智微智能

深圳市智微智能科技股份有限公司
投资者活动记录表

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 ☑其他（线上会议）
参与单位名称及人员姓名	国信计算机艾宪、国信计算机熊莉、信达澳亚基金陈泽昆、浦银安盛基金聂乔、工银瑞信基金胡志利、国信证券侯睿、长城基金程书峰、中邮人寿保险张雪峰、诺安基金王浩然、华金证券王臣复、平安基金薛冀颖、国泰基金张阳、宝盈基金容志能、景顺长城基金刘煜、鹏华基金杨飞、东北证券吴源恒、北京人寿祁朝瑞、中国人寿资产翟晓刚、招商基金宋达、鹏扬基金程阳、永赢基金王嘉玮、华安基金宁柯瑜、华宝基金易镜明、大成基金黄浩峻等 191 家机构参会人员。
时 间	2026/3/5 19:00
地 点	线上会议
上市公司接待人员姓名	杭州元川微科技有限公司创始人 杨滨 智微智能董事会秘书 张新媛 智微智能IRD 汪伟杰
投资者关系活动主要内容介绍	**一、关于参股公司杭州元川微科技有限公司的基本情况介绍** 深圳市智微智能科技股份有限公司近年围绕 AI 进行了一系列布局，2024 年公司布局智算业务，瞄准训练市场，订单快速增长；2025 年在具身智能领域开发大小脑控制器，并围绕具身智能对触觉传感、电机、机器人大脑模型等标的进行投资；2026 年，公司继续秉承全面拥抱 AI 的战略，积极把握 AI 从训练到推理范式转变的重大产业机遇，近日智微通过曜腾投资参股了杭州元川微科技有限公司，深度布局 AI 推理芯片领域。 元川微专注于 AI 推理算力创新，通过回归 AI 推理的第一性原理，聚焦边端智能场景，是国内领先的基于 LPU 架

	构的算力芯片科技公司；依托自研的硬数据流架构与全资源编译器等核心技术，推出了面向大模型、多模态和端侧应用场景的 Mountain（算力）、River（Agent）两大系列 LPU+产品，大幅降低推理应用的部署复杂度和总拥有成本 TCO，精准满足推理应用对确定性、超低时延、高算力、高能效与低成本的核心需求。 通过对元川微进行投资，智微智能旨在通过与上游芯片原厂的深度绑定，强化自身从训练端到推理端的优势卡位，增强在 AI 服务器、具身智能、边缘及端侧领域的产品能力，以期实现双方在技术、生态、资源、市场上的全面协同。 **二、互动交流** **1、Q：请介绍一下什么是 LPU？** A：作为专为推理设计的芯片（为单一任务量身定做的芯片），LPU（Language Processing Unit，语言处理单元）与GPU存在根本性差异。GPU源于图形渲染需求，凭借强大的并行计算能力被拓展至AI领域，支撑模型训练与推理；而LPU则聚焦语言处理场景，针对文本数据的特性深度优化，在自然语言理解、文本生成等任务中实现更高效的处理——如同为“文本引擎”量身定制的专用加速器，在语义解析、对话交互等垂直场景中，展现出比通用计算单元更精准的能效比与响应速度，重新定义了语言智能的硬件实现路径。 LPU采用大容量片上SRAM架构，数据直接集成于芯片，访问延迟远低于传统GPU的“仓库-生产线”分离模式，实现“生产线旁即仓库”的极速响应；其确定性执行架构通过“静态时序”规划，将计算与通信步骤精确到时钟周期，保障稳定高吞吐量。 更关键的是，LPU抛弃了传统“存算分离”的冯诺依曼架构包袱，如同专为推理定制的“F1赛车”，在低时延、高吞吐、低成本、高能效四大维度形成综合优势，成为大模型推理的“性能引擎”。 **2、Q：在2月份英伟达大会上，英伟达一直在强调，Token即收入，单位Token的成本，关注最佳每瓦性能，能否分享**

	下LPU和GPU相比，我们的Token输出速度能快多少，以及成本能下降多少，能耗能下降多少？ A：根据 Groq CEO 在 2024 年 ISSCC 国际固态电路大会公布的实测数据：LPU 的 Token 生成速度达到英伟达 H100 的 6 倍，单 Token 成本降至 H100 的 1/4，推理能耗降至 H100 的 1/3。元川微自身架构验证数据与 Grok 接近，且通过进一步架构优化，有望在能效和成本上表现更优。 **3、Q：能否分享一下 LPU 速度快的原因是什么？** A：LPU速度领先的核心原因，主要来自三大技术特性：①硬流水体系结构：LPU采用纯硬件流水线架构，天然规避冯·诺依曼体系中系统调度、仲裁、多级缓存等额外开销，时延极短。 ②片上大容量SRAM及高带宽：大容量片上SRAM使模型处理长上下文时无需将数据卸载至片外存储，显著降低访存时延；高带宽提升单次Token并行处理能力，并增强算子融合效率——原需3个算子完成的计算可融合为2至2.5个算子执行。 ③ 静态编译调度机制：所有调度工作在编译期静态完成，运行时无需动态仲裁。类比高铁运行图，调度预先确定，拥塞概率极低；传统GPU动态调度犹如高速公路自由行驶，个体随机性在数学上必然导致系统性拥塞。 **4、Q：根据 Groq 的设计，LPU 跨过了内存墙，使用了大量 SRAM，一方面 SRAM 的价格比较昂贵，成本问题如何解决？另一方面 SRAM 的容量较低，能否承载超大模型推理？** A：关于 LPU 采用大容量 SRAM 带来的成本与超大模型承载问题，我们从架构本质和产品迭代两个层面给出清晰解答。成本方面，芯片制造成本的核心决定因素是芯片裸片面积与光罩层数，而非片上组件类型，以相同工艺、相同裸片面积、相同光罩层数的芯片为例，晶圆厂的制造成本并无差异，虽然 SRAM 单比特面积大于 DRAM，但 LPU 凭借专用架构优势，实现同等算力所需的芯片面积远小于传统

	GPU，节省出来的面积可直接用于部署大容量 SRAM，在总芯片面积不变的前提下，SRAM 的扩容由计算单元的面积红利支撑，不会带来额外制造成本，而 GPU 受限于通用架构，即便叠加片上 SRAM 也无法复制 LPU 的系统性优势，二者存在代际级的技术差异。容量方面，Groq 第一代 LPU 所搭载的 SRAM 容量，是适配当时主流 AI 模型的产物，已无法满足当前大语言模型、具身智能与智能体场景的部署需求，这属于产品代际的定位差异，并非 LPU 架构的固有缺陷，针对超大模型推理，新一代 LPU 可实现单芯片加载超大规模模型，大幅减少部署所需芯片数量，后续还能通过按需叠加算力卡的方式灵活扩展，支撑按需付费的商业化模式，显著降低客户的前期投入与部署成本。 **5、Q：展望 2027-2028 年，LPU 与 GPU 的市场占比终局关系如何？** A：未来推理算力占比将达 90%（训练占 10%），LPU 在推理领域能效比、性价比、时延等方面具有显著竞争力，将主导推理市场，而 GPU 可能主要聚焦训练及部分推理场景。 **6、Q：若 GPU、NPU 集成 3D SRAM 吸收 LPU 优点，后续竞争格局如何？** A：GPU、NPU 属于冯诺依曼结构，仅微架构优化，底层依赖多级存储和 Memory 机制，算力密度受限；LPU 核心优势在于非冯诺依曼硬数据流架构及全资源编译器，SRAM 仅是架构环节之一，GPU、NPU 集成 SRAM 无法复制其底层优势。
附件清单（如有）	
日期	2026年3月5日