

证券代码：301396

证券简称：宏景科技

宏景科技股份有限公司投资者关系活动记录表

编号：2026-004

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 （请文字说明其他活动内容）
参与单位名称	广发证券、浙商证券、华泰证券、长江证券、华创证券、广发基金、摩根基金、华安基金、长江养老、民生加银基金、华夏基金、华商基金、西部利得基金、长盛基金、鹏华基金、兴证全球基金、交银施罗德基金、博时基金、中银基金、泉果基金、东方阿尔法、国联基金、前海人寿保险、上海南土资产、中邮基金
时间	2026 年 9 月 3 日、9 月 8 日、9 月 9 日
地点	上海、公司会议室
上市公司接待人员姓名	董事会秘书 张铁舰
投资者关系活动主要内容介绍	投资者提出的问题及公司回复情况 公司就投资者在本次说明会中提出的问题进行了回复： 1、近期市场的需求有什么变化？市场中的需求结构是否有变化？ 今年以来，行业边际变化较为明显。一方面，单一客户的需求体量有所增加；另一方面，整个行业的客户数量有所增多，客户群体进一步多元化。在需求结构上，从过去以大模型预训练的大规模训练集群需求为主，推理类算力需求占比正在持续快速抬升，AI Agent、MoE 架构模型带动 Token（词元）消耗

激增，部分客户不再一味追求超大的集群规模，更加看重弹性调度、混部能力、时延指标等，对性价比的关注度明显提升，客户会把每 token 成本（per token cost）作为核心考核指标，与海外 Anthropic 依靠“年化运行收入”ARR（Annualized Run-Rate）拉动算力采购的逻辑相通，客户的 API 业务收入直接与 Token 调用量挂钩，每 token 成本直接决定自身模型业务的毛利水平。对于已经实现商业化 API 输出的客户，不再单纯看硬件，会把集群调优、推理优化、混部调度、存储网络优化等全部折算到单 token 成本，倒逼算力服务商提供整体解决方案，而不是简单售卖硬件资源。并且，MoE、Agent 场景下 Token 消耗量放大，客户对单 token 成本的敏感度也会进一步放大。在需求主体上，具备实际 B 端商业化 API 落地、有真实付费场景的厂商订单诉求在提升；并且，垂直行业的算力需求在逐渐崛起，工业制造、医疗、医药研发、政务等行业大模型、本地私有化部署项目的咨询和项目需求有所增多，不再只有云厂商、AI 模型公司采购算力。

2、当前国内外主流模型的不断更新迭代，会怎样改变算力需求结构？在算法的不断优化下，算力需求是否会有过剩的情况？

目前，MoE 架构、AI Agent 似乎已经成为行业迭代主线，最大变化是推理算力需求爆发，单 Agent 任务的 Token 消耗量是普通对话的数十倍以上，推理需求规模已经显著超过训练算力。训练算力追求超大集群互联；推理算力更看重弹性调度、多实例混部、低延迟、性价比、异构硬件兼容。

关于算法优化是否带来算力过剩的问题，市场确实存在不同声音的讨论。随着模型量化、MoE 路由算法、推理优化框架持续迭代，完成同等效果任务的单位 Token 算力消耗会下降，理论上会带来单位任务算力需求的降低；但从产业现实看，算法优化并不直接等同于整体算力全面过剩，这主要是因为：

	一方面，算法优化会持续压缩“单位任务算力消耗”，同样业务效果下的硬件需求会变少，低效、未做深度调优的算力会出现闲置过剩；这也是现在客户把每 token 成本作为核心指标的原因之一，算法进步倒逼算力服务商不能只拼硬件，必须叠加调优、混部、调度等软件服务能力。另一方面，算法优化本身又会打开新的应用边界，正如 Anthropic 的 ARR 增长所反映的，模型效率提升之后，下游会把模型应用到更多 Agent 自动化任务、复杂工作流，应用场景和总 Token 消耗量会扩张，从而抵消掉单任务算力下降的影响。当前，国内日均 Token 调用量仍然维持高速增长。 不过，如果未来 AI 应用落地不及预期，即便算法持续优化，行业依然会出现阶段性算力供给过剩的风险，相关行业风险在公司半年度报告中“公司面临的风险和应对措施”章节已有提示，请各位投资者充分关注。 3、大模型的训练和推理在算力的集群架构、网络拓扑结构等方面有什么差异？ 在集群结构和网络拓扑结构上，训练集群更追求大规模同构集群，大量同规格服务器协同完成分布式训练，集群节点规模大，通常“千卡”、“万卡”级别；计算资源需要高度集中，配套高带宽显存，存储侧以高速并行文件系统为核心，需要支撑海量数据集反复读写；集群整体偏向稳态高负载运行，算力资源尽量满负荷，混部业务极少，一般不穿插其他杂项业务。并以高带宽高速网络为主，节点之间带宽要高，降低集合通信延迟，训练过程中梯度同步对网络丢包、延迟非常敏感。 推理集群更偏向异构弹性架构，同时要应对流量波峰波谷，业务负载波动极大；会做算力分片、实例隔离，兼顾多客户、多模型同时运行；存储侧重点是 KV- Cache 高速缓存，对数据集批量读写要求低，但要求高频小数据访问性能。并以网络的低时延、稳定吞吐为主，对单节点间裸带宽要求低于训
--	--

练集群，通信压力小于训练的梯度同步，拓扑架构更偏向分层接入，前端负载均衡网关协同后端算力池，流量是大量短报文，而非训练场景下超大块数据同步。未来，公司会持续打磨算力集群调优能力，兼顾各类算力硬件的适配，以算力资源池的弹性编排满足客户推理业务的各种特征，公司会持续跟进模型技术迭代情况，依据客户实际落地反馈不断迭代解决方案。但未来仍存在诸多难以预测的因素，相关业务的推进节奏可能存在不确定性。

4、公司在推理侧上的竞争壁垒是什么？推理优化会面临哪些现实挑战？

从表面看，推理集群的硬件门槛似乎低于训练大集群，但实际难点在集群调优、混部调度、存储 IO、网络时延等问题。针对当前行业主流的推理优化路线，公司更多尝试以 KV- cache（键值缓存）复用、PD 分离（Prefill- Decode 预填充- 解码分离）相关的架构方案。针对多轮对话、Agent 场景大量重复 prompt（提示词），通过 KV- cache 前缀缓存复用，减少重复 prefill 计算，降低首 token 时延；同时结合 PD 分离架构思路，将计算密集型的预填充任务与访存密集型的解码任务做资源解耦，区分算力池与缓存资源池，提升集群整体有效吞吐，帮助客户压低每 token 推理成本。公司不做底层大模型算法研发，但会在集群部署层面做好适配、调优，把这些优化能力打包进我们的算力服务中。

但鉴于 KV- cache 复用、PD 分离等技术环节难度较高，存在工程落地的复杂度及技术发展迭代的不确定性等不可预测的因素，未来可能存在公司算力优化、技术迭代、方案研发能力的大幅下降，丧失差异化技术竞争优势，无法适配行业技术迭代节奏的情形。公司提醒广大投资者充分关注并理性评估相关业务和技术的潜在风险，审慎作出投资决策。

	5、从半年报数据看，公司资产负债率处于较高水平，请问公司针对高资产负债率有哪些具体应对措施？ 算力业务属于资本密集型业务，算力设备采购、集群建设等带来固定资产、使用权资产的增加，相应推升了公司的负债规模与资产负债率。一方面，公司将持续推进算力项目的交付验收，加强全流程回款管理，将回款考核落实到业务端，改善经营活动现金流，以经营现金流覆盖利息及到期债务；另一方面，优化客户结构，优先选择信用资质好、付款条款清晰、支付能力强的客户，压降应收账款回收风险。 因算力行业资本投入大、硬件迭代快，如果下游算力需求不及预期、算力资产利用率下滑，会对公司偿债、盈利水平带来压力，上述各项措施实施效果受宏观环境、行业景气度、客户订单落地情况影响，存在不确定性。具体财务指标请以公司定期报告的相关数据为准，公司不做相关业绩预测。
附件清单(如有)	无
日期	2026 年 9 月 9 日