打造『国模国芯』生态样板

云天励飞千卡AI推理集群落地

深圳侨报记者 尹萌 通讯员 卢凯平 文/图

3月12日，深圳欧美同学会副会长陈宁所在企业深圳云天励飞技术股份有限公司（简称“云天励飞”）中标湛江市AI渗透支撑新质生产力基础设施建设项目，中标金额4.2亿元。项目将基于云天励飞自研的国产AI推理加速卡，建设国产AI推理千卡集群。

该集群将搭载DeepSeek等国产大模型，为政务、产业及各类应用场景提供更加便捷、低成本的AI能力，探索打造“国模国芯”的AI生态样板。

AI算力从“训练优先”走向“推理优先”

智算集群是人工智能时代的基础设施。如果说电力支撑了工业时代，互联网支撑了信息时代，那么智算正在成为支撑AI时代的重要底座。

在AI算力体系中，算力大体可以分为训练算力与推理算力。训练算力决定模型如何完成“从0到1”的能力构建，而推理算力则直接支撑AI应用落地。无论是春节期间大热的SeeDance，近期广泛讨论的“小龙虾”，还是各行业不断上线的AI Agent应用，背后都离不开推理算力的支撑。根据Gartner预测，到2026年，约55%的AI专用云基础设施支出将用于推理工作负载。

过去，国内许多智算中心普遍采用“训推一体”的建设模式。而此次在湛江建设的集群，则定位为专注推理任务的AI推理集群，主要面向各类行业应用场景，为传统产业的AI化提供直接支撑。

此次云天励飞建设的AI推理集群，也将与DeepSeek等国产模型进行深度适配，为更多行业应用提供算力支撑。

面向推理时代的千卡集群架构

在大模型应用场景中，推理系统通常需要同时满足高并发、高吞吐与低延迟三项要求。为提升整体效率，当前业界普遍采用“Prefill－Decode分离”的推理架构，通过对不同阶段进行资源优化，实现系统性能的整体提升。

其中，Prefill阶段主要负责对长上下文进行理解和计算，计算量大、带宽需求高；而Decode阶段则负责持续生成Token，对系统延迟更加敏感。如何在两个阶段之间进行合理的资源配置，成为推理系统架构设计的重要问题。

与此同时，随着大模型上下文长度不断增加，大量中间状态需要以KV Cache的形式存储。业内普遍认为，未来推理系统的性能瓶颈将越来越多来自数据访问效率，而不仅仅是计算能力。

在这一背景下，算力、存储与网络之间的协同设计，正逐渐成为AI基础设施的重要竞争力。

此次在湛江落地的千卡推理集群，正是围绕这一思路进行构建。

该集群采用云天励飞自主研发的AI推理芯片，并在系统架构上确立了“优先优化Prefill、兼顾Decode”的技术路线。通过在芯片设计中对计算资源与存储带宽进行针对性配置，使系统在长上下文推理场景下依然能够保持较高的吞吐效率。

在网络互联方面，系统采用统一高速互联架构，通过400G光网络构建集群物理层网络，实现节点之间的高带宽、低延迟通信。与传统在节点内和节点间分别采用不同协议构建网络的方式相比，这种同构互联架构减少了协议转换带来的额外开销，也简化了系统部署。

在部署能力上，该架构既可以支持单节点数十卡规模扩展，也能够平滑扩展至千卡级集群规模，从而适配不同规模的AI应用需求。

此外，针对大模型推理中KV Cache访问带来的压力，系统在计算互联与存储互联层面进行了协同优化。通过计算网络与存储网络的联合调度，可以显著提升数据读取效率，使模型在长上下文推理场景下依然保持稳定性能。

通过芯片架构、网络互联以及系统调度等多层优化，这一推理集群在整体效率与成本控制方面形成了明显优势，为AI规模化应用提供了更加经济的算力方案。

自研芯片构建低成本推理能力

据悉，本次AI推理集群将分三期建设，并全部采用云天励飞自研的国产AI推理加速卡。

其中，一期项目将部署云天励飞 X6000 推理加速卡；二、三期建设将率先搭载公司最新一代芯片产品。

根据公司规划，未来三年云天励飞将推出三代AI推理芯片产品。

第一阶段，将推出面向长上下文场景优化的Prefill芯片，通过提升计算效率与内存访问能力，为OpenClaw、各类AI Agent提供基础算力支撑。

第二阶段，将研发专注于Decode阶段低延迟优化的芯片产品，进一步提升实时推理能力。

第三阶段，则通过系统级协同优化，实现Prefill与Decode性能的整体提升，向毫秒级推理时延目标迈进。

其中，首款Prefill芯片DeepVerse100预计将在年内完成流片，并计划在湛江集群中率先部署。

在更长期的规划中，云天励飞提出“1001计划”，即以“百亿Token一分钱”为长期目标，通过芯片与系统协同优化持续降低大模型推理成本。

过去几年，AI算力建设往往以“堆算力”为主要路径——通过不断扩大GPU规模来获得更高性能。但随着大模型逐渐进入应用阶段，产业关注点正从“算力峰值”转向“单位成本效率”。

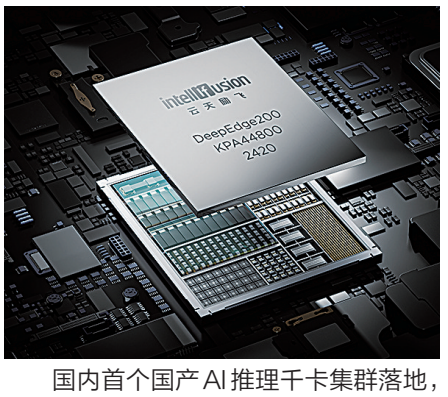
换句话说，未来AI产业竞争的重要维度，不仅在于模型能力本身，还在于谁能够以更低成本提供稳定的大规模推理能力。

湛江项目的落地，也为这一目标提供了重要的实践场景。千卡级推理集群不仅能够满足当前AI应用需求，同时也为更大规模算力系统提供技术部署平台。

在典型架构下，一个千卡级集群通常由多级扩展结构组成：从单节点8卡、32卡，到64卡甚至百卡级超节点，再到跨节点的大规模集群。通过这一规模系统的实际运行，可以充分验证卡间互联、节点通信和负载均衡等关键技术，为未来更大规模AI算力系统建设积累经验。

随着大模型逐步进入产业应用阶段，AI基础设施的发展逻辑也正在发生变化——从单纯追求算力规模，转向更加注重效率与成本。

在业内看来，推理算力将成为决定AI应用规模化落地的关键基础设施。谁能够以更高效率、更低成本提供稳定的大规模推理能力，谁就有机会在新一轮人工智能产业竞争中占据先机。



国内首个国产AI推理千卡集群落地，采用云天励飞全自研AI推理芯片。