精选快讯　硬件快讯　2026年11月12日 14:05

SK海力士发布多服务器共享“内存池”架构概念，称与英特尔合作推进

要点

- SK海力士提出名为“memory pool”的架构概念，目标是让多台服务器共同使用大容量内存，以解决AI系统面临的内存瓶颈。
- 按照展示方案，在GPU服务器机架和CPU服务器机架之外，会增加一个容纳数十至数百TB的独立内存机架。
- 在与合作伙伴进行的多轮AI推理评估中，生成首个token的平均时间从传统重计算方法的80.76秒降至使用远程CXL DRAM的15.45秒，提升约5.2倍。
- 这一架构概念在实现前仍需共同开发内存池与服务器的连接规范及软件生态，并在数据中心规模验证CXL共享内存的延迟和可靠性。

SK海力士提出名为“memory pool”的架构概念，目标是让多台服务器共同使用大容量内存，以解决AI系统面临的内存瓶颈。该架构不采用将内存附着于单个GPU的传统方式，而是把计算资源与存储资源分离。SK海力士副总裁Joo Young-pyo在Intel AI Summit Seoul 2026的主题演讲中表示，进入agentic AI时代，内存容量和性能压力将加剧。Joo Young-pyo称，SK海力士正从系统设计阶段起与客户及合作伙伴讨论，并与英特尔紧密合作推进该架构。

按照展示方案，在GPU服务器机架和CPU服务器机架之外，会增加一个容纳数十至数百TB的独立内存机架。现有服务器本地内存保留，并在此基础上增加共享内存池。机架之间通过RDMA连接，机架内部使用CXL连接。SK海力士将AI系统的三大瓶颈概括为内存容量不足、数据访问延迟和数据管理复杂。

在与合作伙伴进行的多轮AI推理评估中，生成首个token的平均时间从传统重计算方法的80.76秒降至使用远程CXL DRAM的15.45秒，提升约5.2倍。同一实验中，从远程SSD加载缓存耗时33.56秒，表明CXL内存在速度上占优。

这一架构概念在实现前仍需共同开发内存池与服务器的连接规范及软件生态，并在数据中心规模验证CXL共享内存的延迟和可靠性。SK海力士正与客户和合作伙伴继续相关测试。

报道来源：biggo

来源网站：https://www.inews.today/