精选快讯　硬件快讯　2026年11月19日 15:12

忆联携手英特尔发布基于至强6200P-B与UH812a的极简KV Cache存储节点方案

要点

- 忆联联合英特尔发布《基于英特尔® 至强® 6200P-B 系列处理器和忆联 UH812a 企业级 SSD 的极简存储节点 KV Cache G3.5 方案最佳实践》。
- 方案核心采用英特尔至强 6200P-B 系列处理器（代号 Gem Mountain）与忆联企业级 PCIe 5.0 SSD UH812a。
- 在性能指标方面，UH812a 顺序读写速度最高达 14,900/10,500 MB/s，4K 随机读达 3,500K IOPS，顺序读峰值功耗控制在 18W 以内。
- 实测数据显示，在 vLLM + LMCache 环境下，当缓存命中率为 80% 时，首字时延（TTFT）较纯计算减少 57%，系统总吞吐量提升 93%。

忆联联合英特尔发布《基于英特尔® 至强® 6200P-B 系列处理器和忆联 UH812a 企业级 SSD 的极简存储节点 KV Cache G3.5 方案最佳实践》。该方案旨在通过高集成度硬件组合，优化大模型推理场景下的缓存存储性能，提供标准化的部署参考。

方案核心采用英特尔至强 6200P-B 系列处理器（代号 Gem Mountain）与忆联企业级 PCIe 5.0 SSD UH812a。硬件架构由 Gem Mountain 紧凑型 SoC、4 块 UH812a SSD 及 400GbE RDMA 网卡构成独立存储节点。通过 SPDK 与 NVMe-oF 技术优化，该节点实现端到端单 IO 随机读时延约 60μs，随机写时延约 10μs。

在性能指标方面，UH812a 顺序读写速度最高达 14,900/10,500 MB/s，4K 随机读达 3,500K IOPS，顺序读峰值功耗控制在 18W 以内。配合 GMM SoC，单节点 4 盘配置可实现 30TB 以上容量及 11,000,000 IOPS 的吞吐能力。

实测数据显示，在 vLLM + LMCache 环境下，当缓存命中率为 80% 时，首字时延（TTFT）较纯计算减少 57%，系统总吞吐量提升 93%。若实现 100% 缓存命中，时延可降至 0.19ms，吞吐量高达 2.52 token/s，显著提升了推理效率。

报道来源：admin5、中华网

来源网站：https://www.inews.today/