Speaker Name:

LOKESH NAGAPPA JALIMINCHE

Talk Title:

Towards QEMU-Based CXL-SSD Emulation with Hint-driven In-device Data Placement

Modern applications, such as large language models (LLMs) and deep learning recommendation models (DLRMs), demand substantial memory resources, hitting the memory capacity wall where traditional DRAM solutions can no longer scale efficiently. This growing challenge has driven the need for scalable, cost-effective alternatives that can handle large volumes of data without compromising performance.

To address this issue, researchers are exploring the use of hybrid devices like CXL-SSDs, which offer promising density scaling advantages over DRAM. However, accessing data from NAND Flash can still lead to significant performance degradation, highlighting the need for further research on optimizing the use of the device's DRAM cache and minimizing long NAND Flash access latencies. However, since these devices are still in their early stages, there is a lack of platforms to facilitate research on such strategies.

We introduce a QEMU-based CXL-SSD emulator that implements hint-driven in-device data placement, enabling more efficient management of DRAM and Flash within the device. Initial experiments revealed that traditional data placement techniques, like prefetching based on physical addresses, were ineffective due to non-linear virtual-to-physical address mappings. Our approach overcomes this limitation by allowing applications to guide data placement based on access patterns based on virtual addresses. Early results show improved prefetching and caching efficiency when data placement is driven by the host or application, demonstrating the effectiveness of our approach. We believe this emulation platform lays a solid foundation for future research, not only on CXL-SSD but also on hybrid memory systems and effective data management across diverse storage media.