Speaker:

Allen Aboytes (aaboytes@ucsc.edu)

Talk Title:

Semantic Data Tiering for CXL Memory Systems

Abstract:

Vector processing applications, such as vector databases, place intense pressure on memory due to their growing working sets that span beyond the boundaries of a single machine. Far memory technologies (e.g., CXL) potentially solve the memory capacity problem by adding terabytes to system configurations. However, efficient use of CXL memory requires changes to the system stack. Recent work uses page-based memory management, memory object-based far memory runtimes, or systems that serve a single use case to widen the memory hierarchy. Despite these advances, each approach has its downsides; they either have high metadata overheads, cannot scale to serve large memory deployments or ignore the semantics between memory objects, wasting precious memory capacity and bandwidth. To address the issues, we propose a tiered memory system, Mnemonic Memory Tiers (M2T), that captures semantic information at the memory allocation layer to place related objects within the same memory region called SemSlabs. M2T tracks and migrates entire SemSlabs, composed of multiple pages, between local and CXL memory and enables transparent access to far memory through application libraries that use our proposed application-aware memory allocator.