Speaker:

Allen Aboytes (aaboytes@ucsc.edu)

Talk Title:

Application-Specified Memory Management

Abstract:

Modern data analytics applications are memory-intensive. For example, vector database applications work with datasets that can take terabytes of memory, spanning beyond the boundaries of a single machine. Far memory technologies (i.e., CXL) potentially solve the capacity problem. However, efficient use of CXL memory devices requires changes to the system stack. Recent work either uses high overhead page-based swapping or ad hoc systems that serve a single use case. An alternative approach is to use a library and runtime system with general memory management APIs.

Although object-based far memory runtimes perform well, they have high metadata overheads and ignore the relationships between objects. Many systems must update the references between objects when moving data between memory devices. Ignoring the semantics of memory objects can mean many expensive trips to far memory. A better approach would be to use semantic information to influence prefetching and allocation policies to increase the locality of reference for related data objects.

To address the issues above, we propose to manage memory with movable objects and capture semantic information at the memory allocation layer. To achieve this, we use object-based allocation APIs supported by a runtime co-designed with the operating system. Data objects enable interoperability between hosts and devices with logical pointers, creating a shared memory abstraction.