SICM: Simplified Interface to Complex Memory

Exascale systems will have complex, heterogenous memories that need to be effectively managed either directly by the programmer or by the runtime in order to achieve high performance. Natively supporting each memory technology is challenging, as each has its own separate programming interface. The SICM project addresses the emerging complexity of exascale memory hierarchies by providing a portable, simplified interface to complex memory that application programmers and library developers can use to achieve their performance goals.

The SICM project is creating a universal interface for discovering and managing complex memory hierarchies and sharing resources within them. Memory technologies to be supported include, for example, high-bandwidth memory associated with accelerators, nonvolatile memory, 3D stacked memory, and phase-change memory. The result of the SICM project will be a portable, simplified memory interface and software library that will allow application programmers, library developers, and vendors to use these emerging memory technologies without having to program to each technology-specific programming interface.

The SICM team will provide a unified, two-tier node-level complex memory interface. The low-level interface will allow full control of what memory types are being used and is meant for expert developers. This interface will provide support for discovery, allocation/de-allocation, partitioning, and configuration of the memory hierarchy and information on the properties of each specific memory, such as capacity, latency, bandwidth, volatility, and power usage.

The high-level interface will enable developers to define coarser-level constraints on the types of memories needed and leave out the details of the memory management. This interface would potentially be a catalyst for more research as intelligent allocators, migrators, and profiling tools are developed. The high-level interfaces will leverage the low-level interface and library to further decouple applications and libraries from hardware configurations. Specifically, it will emphasize ease of use by developers with a policy/intent–driven syntax enabled through runtime intelligence and system support. Developers will specify which attributes are a priority for each allocation, and the interface will provide the most appropriate configuration.

The impact of the SICM project will be immediate and wide reaching, as developers in all areas are struggling to add support for new memory technologies, and the simplified interface to complex memory can alleviate these challenges.

Progress to date

- The SICM team prototyped and delivered an initial low-level library that supports high-bandwidth memory associated with accelerators and nonvolatile memory.

- The SICM team has provided a high-level interface that supports graph-type allocations on block-based nonvolatile memory devices and produced tools for performance analysis of data structure to inform developers of the potential benefits of migrating those data structures to high-bandwidth memory.

- The SICM team has been working with OpenMP to support the generation of SICM library calls from OpenMP directives to ease adoption of SICM and provide access to heterogenous memory with minimal code changes.

PI: Mike Lang, Los Alamos National Laboratory

Collaborators: Los Alamos National Laboratory, Lawrence Livermore National Laboratory, Oak Ridge National Laboratory, Sandia National Laboratories