Contribution ID: **1** Type: **not specified**

Harnessing Integrated CPU-GPU System Memory for HPC: a first look into Grace Hopper

Wednesday, 9 October 2024 15:30 (1 hour)

Abstract:

Memory management across discrete CPU and GPU physical memory is traditionally achieved through explicit GPU allocations and data copy or unified virtual memory. The Grace Hopper Superchip, for the first time, supports an integrated CPU-GPU system page table, hardware-level addressing of system allocated memory, and cache-coherent NVLink-C2C interconnect, bringing an alternative solution for enabling a Unified Memory system. In this presentation, we provide an in-depth analysis of the memory system of the Grace Hopper Superchip. We further detail how this hardware can be leveraged by programmers through various programming interfaces, comparing system-allocated memory and managed memory. We evaluate the performance of a suite of six representative applications, including the Qiskit quantum computing simulator, using system memory and managed memory. Using our memory utilization profiler and hardware counters, we quantify and characterize the impact of the integrated CPU-GPU system page table on GPU applications. Our study focuses on first-touch policy, page table entry initialization, page sizes, and page migration. Our results show that as a new solution for unified memory, the system-allocated memory can benefit most use cases with minimal porting efforts.

The talk will be held in English | Join Zoom Meeting
https://fz-juelich-de.zoom.us/j/63570550943?pwd=S1rwqKaB6ayfMKW8kkaQK3unkK4iKl.1

Presenter: SCHIEFFER, Gabin (KTH)

Session Classification: Presentations