APS Scientific Computation Seminar Series

Speaker:
Steven E. Hahn, Software Scientist, Computer Science and Mathematics Division, Oak Ridge National Laboratory

Title: Integrating ORNL's HPC and Neutron Facilities with a Performance-Portable CPU/GPU Ecosystem

Date: May 19, 2025

Time: 1:00 p.m. (Central Time)

Location:
Join ZoomGov Meeting https://argonne.zoomgov.com/j/1601444470?pwd=N1phbHZVdCtmcVR5cGh0c1Zhc0orZz09
Meeting ID: 160 144 4470
Passcode: 937918
One tap mobile
+16692545252,,1601444470# US (San Jose)
+16468287666,,1601444470# US (New York)
Dial by your location
+1 669 254 5252 US (San Jose)
+1 646 828 7666 US (New York)
+1 646 964 1167 US (US Spanish Line)
+1 669 216 1590 US (San Jose)
+1 415 449 4000 US (US Spanish Line)
+1 551 285 1373 US
Meeting ID: 160 144 4470
Find your local number: https://argonne.zoomgov.com/u/af2crdvQy

Hosts: Mathew Cherukara and Nicholas Schwarz

Abstract:
We explore the development of a performance-portable CPU/GPU ecosystem to integrate the Oak Ridge Leadership Computing Facility (OLCF) and the Spallation Neutron Source (SNS), both of which are housed at Oak Ridge National Laboratory. We select a data reduction workflow use-case to obtain the differential scattering cross-section from data collected by SNS's CORELLI and TOPAZ instruments. Inconvenient execution times from complex data processing techniques and large volumes of data discourages users from utilizing all experimental data. The current CPU-only production implementation using the Garnet Python multiprocess package based on the C++ Mantid framework is compared against our proposed CPU/GPU implementation that uses the LLVM-based Julia scientific language and the JACC.jl performance portability package. To understand and address performance challenges, two proxy apps were developed: (i) an app for extracting relevant Mantid kernels (MDNorm) in C++ and (ii) the Julia MiniVATES.jl miniapp. We introduce algorithmic improvements: (i) parallelization strategies across multiple GPU nodes, (ii) reformulating key calculations into a tall-skinny matrix multiplication kernel, and (iii) exploring intermediate data layout representations for optimizing reading and processing times. Performance results from NVIDIA A100, H200 and AMD MI100 GPUs and AMD EPYC and NVIDIA Grace CPUs will be presented. These computational experiments show 10x+ speed ups and provide insights for future generations of data reduction software that can take advantage of developments in productivity and performance portability, making high-performance computing more accessible for an integrated research infrastructure across DOE's experimental and computational facilities.