Contribution ID: **26** Type: **Oral**

# Evaluating Application Characteristics for GPU Portability Layer Selection

*Wednesday 13 March 2024 14:30 (20 minutes)*

GPUs have become the dominant source of computing power for HPCs and are increasingly being used across the High Energy Physics computing landscape for a wide variety of tasks. Though NVIDIA is currently the main provider of GPUs, AMD and Intel are rapidly increasing their market share. As a result, programming using a vendor-specific language such as CUDA can significantly reduce deployment choices. There are a number of portability layers such as Kokkos, Alpaka, SYCL, OpenMP and std::par that permit execution on a broad range of GPU and CPU architectures, significantly increasing the flexibility of application programmers. However, each of these portability layers has its own characteristics, performing better at some tasks and worse at others, or placing limitations on aspects of the application. In this presentation, we report on a study of application and kernel characteristics that can influence the choice of a portability layer and show how each layer handles these characteristics. We have analyzed representative heterogeneous applications from CMS (patatrack and p2r), DUNE (Wire-Cell Toolkit), and ATLAS (FastCaloSim) to identify key application characteristics that have different behaviors for the various portability technologies. Using these results, developers can make more informed decisions on which GPU portability technology is best suited to their application.

## Significance

Flexibly porting code originally written for CPUs to diverse heterogeneous architectures is currently an unsolved problem in the HEP community. While some experiments have ported some code bases to a single or a small number of platforms as they have already purchased their selected hardware backends, there has not been a systematic study of the problem addressing all currently available heterogeneous architectures. Some experiments have selected technologies such as Alpaka or HIP, simply because it functioned for their code bases. This does not help other experiments make a portability layer selection, as their use cases are likely different. This study is cross-cutting in nature, identifying application characteristics that result in different performance for the various layers. By using this information, application developers can more easily select a portability technology without having to try each one.

## References

## Experiment context, if any

**Primary author:** Dr LEGGETT, Charles (Lawrence Berkeley National Lab (US))

**Co-authors:** VIREN, Brett (Brookhaven National Laboratory); MOHAMMAD ATIF, FNU (Brookhaven National Laboratory); YU, Haiwang; ESSEIVA, Julien (Lawrence Berkeley National Lab. (US)); KWOK, Ka Hei Martin (Fermi National Accelerator Lab. (US)); DEWING, Mark; KORTELAINEN, Matti (Fermi National Accelerator Lab. (US)); BHATTACHARYA, Meghna (Fermilab); LIN, Meifeng; STRELCHENKO, Oleksii (Fermi National Accelerator Lab. (US)); GUTSCHE, Oliver (Fermi National Accelerator Lab. (US)); WANG, Tianle (Brookhaven National Lab); TSULAIA, Vakho (Lawrence Berkeley National Lab. (US)); DONG, Zhihua

**Presenter:** Dr LEGGETT, Charles (Lawrence Berkeley National Lab (US))

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research