DIANA Fellowship Proposal: Optimizing the AMD back-end for Allen

Allen is a fully GPU-based implementaion of the first level trigger (data-ingestion and reduction system) designed for the upgrade of CERN's LHCb experiment that will start to take data in ealy 2022. It is designed to process the 40 Tbit/s data rate produced by the detector's electronics and perform a wide variety of pattern recoginition tasks. These include reconstructing charged particle trajectories, finding proton-proton collision points, distinguishing between hadrons and muons, and selecting a small subset of the data to be persistsed for processing by a second level trigger. The framework supports C++, CUDA, and HIP backends for CPUs, nVidia GPUs, and AMD GPUs, respectively. The CUDA implementation already meets all performance specifications. The HIP implementation produces the same physics results as the CUDA implementation, but its throughput is substantially lower.

The goal of this project is to improve the performance of the HIP backend. The incumbent will work under the supervision of lead developers and other experts at CERN and Maastricht University to identify which elements of Allen consume disproportionately more time in the HIP implementation than in the CUDA implementation and will then identify and implement strategies to improve the HIP backend performance. Close collaboration with engineers from AMD will be provided via CERN. Candidates for this position should have a good working knowledge of C++ and use of `git`. Experience with multi-threaded applications, especially CUDA or HIP, is preferred. The anticipated duration of the project is the three month period May - July, 2021, although there is some flexibility related to the exact start and finish dates.

The physics goals and performance of Allen are discussed in *Allen: A High-Level Trigger on GPUs for LHCb*. For an overview of the software itself, see the Allen homepage in GitHub.

Niko Neufeld (CERN) and Daniel Campora (NIKHEF) will supervise the student. A timeline, describing the work plan and deliverables, is provided on the next page.

Timeline

weeks 1-2 learn how to run Allen with AMD and nVidia GPU back-ends and interact with the code development system; learn the existing continuous integration/continuous development setup;

weeks 3-4 working with AMD engineers (an other experts), learn to run profiling tools to identify hot-spots, identify contention areas, etc.

weeks 5-6 working with experts, define a strategy for systematically identifying problem areas and improving performance;

weeks 7-8 begin to execute the systematic studies defined in weeks 5-6 and mitigate the most egregious problems.

weeks 9-11 carefully document prior work; continune to identify and mitigate problems;

weeks 12-13 define a road-map for future work; complete documentation.

At the end of the project, the student will present his/her work at an LHCb RTA (Real Time Analysis) meeting and also at an IRIS-HEP topical meeting.