# DIANA Fellowship Proposal: Profiling the CUDA 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 back-ends for CPUs, nVidia GPUs, and AMD GPUs, respectively. The CUDA implementation already meets all performance specifications, but its performance does not scale linearly with the number of CUDA cores on a GPU board.

The goal of this project is to build a toolkit to carefully profile the performance of the CUDA back-end and identify bottlenecks wherever they occur, including use of processor units, registers, and memory access. The incumbent will work under the supervision of a lead developer and another expert from Maastricht University and the University of Cincinnati to build this toolkit and use it to study Allen's performance on a variety of nVidia GPUs. Close collaboration with engineers from nVidia 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 is preferred. Prior knowledge of GPU architectures will be useful.

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.

Daniel Campora (Maastricht) and Tom Beottcher (Cincinnati) 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 CPU and nVidia GPU back-ends and interact with the code development system;

weeks 3-4 learn how to use profiling tools by hand;

weeks 5-6 begin to identify bottlenecks in the CUDA back-end; run targeted profiling (e.g., Roofline models) of specifidc algorithms;

weeks 7-8 integrate profile runs into the Allen continuous integration; document work to date;

weeks 9-11 test application-wide settings (such as L1 cache preferences or launch bounds for specific kernels) and measure their impact on throughput and profiling;

weeks 12-13 continue profiling studies, finish integration of profile runs into the Allen continuous integration, and fully document work done.

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.