Contribution ID: **119** Type: **Short Talk**

# Performance of CUDA Unified Memory in CMS Heterogeneous Pixel Reconstruction

*Tuesday, 18 May 2021 15:26 (13 minutes)*

The management of separate memory spaces of CPUs and GPUs brings an additional burden to the development of software for GPUs. To help with this, CUDA unified memory provides a single address space that can be accessed from both CPU and GPU. The automatic data transfer mechanism is based on page faults generated by the memory accesses. This mechanism has a performance cost, that can be with explicit memory prefetch requests. Various hints on the inteded usage of the memory regions can also be given to further improve the performance. The overall effect of unified memory compared to an explicit memory management can depend heavily on the application. In this paper we evaluate the performance impact of CUDA unified memory using the heterogeneous pixel reconstruction code from the CMS experiment as a realistic use case of a GPU-targeting HEP reconstruction software. We also compare the programming model using CUDA unified memory to the explicit management of separate CPU and GPU memory spaces.

**Primary authors:** KORTELAINEN, Matti (Fermi National Accelerator Lab. (US)); KWOK, Ka Hei Martin (Fermi National Accelerator Lab. (US))

**Presenter:** KWOK, Ka Hei Martin (Fermi National Accelerator Lab. (US))

**Session Classification:** Accelerators

**Track Classification:** Offline Computing