Contribution ID: **100** Type: **Poster**

GPU unified memory in the CMS software and alpaka

The CMS experiment requires massive computational resources to efficiently process the large amounts of detector data. The computational demands are expected to grow as the LHC enters the high-luminosity era. Therefore, GPUs will play a crucial role in accelerating these workloads. The approach currently used in the CMS software (CMSSW) relies on explicit memory management techniques, where data must be manually copied between CPU and GPU memory, leading to increased complexity and requiring careful synchronization to avoid performance bottlenecks.

Unified memory addresses this limitation. It simplifies working with complex data structures by automatically handling memory transfers between CPU and GPU without requiring explicit pointer adjustments. In addition to page fault handling, hardware prefetching, and automatic data migration, the latest generations of NVIDIA and AMD GPUs provide hardware features that further optimise unified memory, like cache-coherent access to the host memory or even a single unified memory pool. As GPUs evolve, unified memory is becoming more significant for writing efficient heterogeneous software.

The impact of unified memory has been studied using the CLUE library used in the CMS software. Based on these results, work is ongoing to implement support for unified memory in alpaka, a performance-portable parallel programming framework used in the CMS software, to ensure efficient tasks submission and scheduling across different hardware architectures.

Significance

References

Experiment context, if any

CMS experiment

Author: CMS COLLABORATION

Presenter: CMS COLLABORATION

Session Classification: Poster session with coffee break

Track Classification: Track 1: Computing Technology for Physics Research