Contribution ID: **39** Type: **Oral**

Heterogeneous data-processing optimization with CLARA's adaptive workload orchestration

Tuesday 5 November 2019 11:30 (15 minutes)

The hardware landscape used in HEP and NP is changing from homogeneous multi-core systems towards heterogeneous systems with many different computing units, each with their own characteristics. To achieve data processing maximum performance the main challenge is to place the right computing on the right hardware.
In this paper we discuss CLAS12 charge particle tracking workload partitioning that allowed us to utilize both CPU and GPU to improve the performance. The tracking application algorithm was decomposed into micro-services that are deployed on CPU and GPU processing units, where the best features of both are intelligently combined to achieve maximum performance. In this heterogeneous environment CLARA aims to match the requirements of each micro-service to the strength of a CPU or a GPU architecture. In addition, CLARA performs load balancing to minimize idle time for both processing units. However predefined execution of a micro-service on a CPU or a GPU may not be the most optimal solution due to the streaming data-quantum size and the data-quantum transfer latency between CPU and GPU. So, we trained the CLARA workflow orchestrator to dynamically assign micro-service execution to a CPU or a GPU, based on the benchmark results analyzed for a period of the real-time data-processing.

Consider for promotion

Yes

Authors: GYURJYAN, Vardan (Jefferson Lab); HEYES, Graham (Jefferson Lab); ABBOTT, David (Jefferson Lab); TIMMER, Carl (Jefferson Lab); BENKEL, Bruno (FSMTU); MANCILLA, Sebastian (Departamento de Fisica-Univ. Tecnica Federico Santa Maria (UTFSM); ZIEGLER, Veronique (JLAB)

Presenter: GYURJYAN, Vardan (Jefferson Lab)

Session Classification: Track 5 –Software Development

Track Classification: Track 5 –Software Development