Contribution ID: **332** Type: **Poster**

Updates on the Low-Level Abstraction of Memory Access

Tuesday 25 October 2022 16:10 (30 minutes)

Choosing the best memory layout for each hardware architecture is increasingly important as more and more programs become memory bound. For portable codes that run across heterogeneous hardware architectures, the choice of the memory layout for data structures is ideally decoupled from the rest of a program.
The low-level abstraction of memory access (LLAMA) is a C++ library that provides a zero-runtime-overhead abstraction layer, underneath which memory layouts can be freely exchanged, focusing on multidimensional arrays of nested, structured data.
It provides a framework for defining and switching custom memory mappings at compile time to define data layouts, data access and access instrumentation, making LLAMA an ideal tool to tackle memory-related optimization challenges in heterogeneous computing.
After its scientific debut, several improvements and extensions have been added to LLAMA. This includes compile-time array extents for zero memory overhead, support for computations during memory access, new mappings (e.g. int/float bit-packing or byte-swapping) and more. This contribution provides an overview of the LLAMA library, its recent development and an outlook of future activities.

Experiment context, if any

References

Journal paper: https://doi.org/10.1002/spe.3077

Significance

LLAMA provides a general C++ library solution for memory layout and memory access abstractions which are crucial to exploit the heterogeneous landscape of modern hardware architectures, including CPUs, GPUs and FPGAs. LLAMA is an orthogonal framework to compute kernel abstractions such as alpaka and Kokkos and of similar importance.

Primary author: GRUBER, Bernhard Manfred (Technische Universitaet Dresden (DE))

Presenter: GRUBER, Bernhard Manfred (Technische Universitaet Dresden (DE))

Session Classification: Poster session with coffee break