XKBlas: under the hood

Thierry GAUTIER

Inria, France

Abstract

XKBlas is a BLAS Level3 library with a MUMPS-supported C interface to leverage multi-GPUs. XKBlas is ported to NVidia GPUs through CUDA/CUBLAS APIs as well as AMD GPUs via ROCM/HIP/HIPBLAS APIs through the portability layer called XKaapi. In this presentation, I will focus on how the XKBlas C API was designed to promote asynchrony to enable overlap between communication and compute operations. Performance on top of multi-GPUs will be presented to illustrate the impact of some internal design decisions.