Contribution ID: **15** Type: **Short talk**

Memory-Aware Scheduling of Tasks Sharing Data on Multiple GPUs with Dynamic Runtime Systems

Thursday 23 March 2023 14:10 (10 minutes)

The use of accelerators such as GPUs has become mainstream to achieve high performance on modern computing systems. GPUs come with their own (limited) memory and are connected to the main memory of the machine through a bus (with limited bandwidth). When a computation is started on a GPU, the corresponding data needs to be transferred to the GPU before the computation starts. Such data movements may become a bottleneck for performance, especially when several GPUs have to share the communication bus.

Task-based runtime schedulers have emerged as a convenient and efficient way to use such heterogeneous platforms. When processing an application, the scheduler has the knowledge of all tasks available for processing on a GPU, as well as their input data dependencies. Hence, it is possible to produce a tasks processing order aiming at reducing the total processing time through three objectives: minimizing data transfers, overlapping transfers and computation and optimizing the eviction of previously-loaded data. We focus on this problem of partitioning and ordering tasks that share some of their input data on multiple GPUs. We present a novel dynamic strategy based on data selection to efficiently allocate tasks to GPUs and a custom eviction policy, and compare them to existing strategies using either a well-known graph partitioner or standard scheduling techniques in runtime systems.
We present their performance on tasks from tiled 2D, 3D matrix products and Cholesky factorization, as well as a sparse matrix product.

All strategies have been implemented on top of the StarPU runtime, and we show that our dynamic strategy achieves better performance when scheduling tasks on multiple GPUs with limited memory.

JLESC topic

Primary authors: Mr MARCHAL, Loris (CNRS); GONTHIER, Maxime (INRIA Bordeaux); Mr THIBAULT, Samuel (Université de Bordeaux)

Presenter: GONTHIER, Maxime (INRIA Bordeaux)

Session Classification: Short Talks on Tasking

Track Classification: Programming languages and runtimes