Asynchronous batched methods on GPUs

Pratik Nayak[1] Hartwig Anzt[2]

Batched methods leverage the embarrassing parallelism available to make maximal use of the resources of a GPU. While there has been extensive research on batched dense and batched direct methods, there has been a lack of research and implementation for batched sparse and iterative methods. In this talk, we will explore implementing asynchronous batched Krylov methods on GPUs. Many applications require solution of multiple independent small linear systems, which are very suitable for GPUs due to the massive parallelism available in them. To make maximal use of these GPUs though, it is necessary to maximize usage of the cache hierarchy, reduce the kernel launch latency by fusing the kernels and optimize the occupancy of the GPU. In this talk, we will explore a few applications, namely a combustion application and a fusion application and showcase results comparing the state of the art batched methods with our batched iterative methods and show that we can obtain significant performance benefits using these batched iterative methods. Finally, we will explore some ideas needed to use these methods as block-preconditioners and the benefits these can bring to the solution of larger monolithic linear systems.

[1]Karlsruhe Institute of Technology
pratik.nayak@kit.edu

[2]Karlsruhe Institute of Technology
hartwig.anzt@kit.edu