Towards Uniform Tasking on CPUs and GPUs

Laura Morgenstern, Chemnitz University of Technology & Jülich Supercomputing Centre, Forschungszentrum Jülich GmbH

Task parallelism is omnipresent these days; whether in data mining or machine learning, for matrix factorization or even molecular dynamics (MD). Despite the success of task parallelism on CPUs, there is currently no performant way to exploit the task parallelism of synchronization-critical algorithms on GPUs. Due to this shortcoming, we develop a tasking approach for GPU architectures. Our use case is a fast multipole method for MD simulations, which targets strong scaling. Hence, the application tends to be latency- and synchronization-critical. Therefore, offloading as the classical programming model for GPUs is unfeasible. We share our experience with the design and implementation of tasking as alternative programming model for GPUs using CUDA. We describe the tasking approach for GPUs based on the design of our tasking approach for CPUs. Following this, we reveal several pitfalls implementing it.