

Shallow water simulations on FPGAs with a portable SYCL implementation

Author:

Christoph Alt

Co-authors:

Markus Büttner, Tobias Kenter, Harald Köstler, Christian Plessl, Vadym Aizinger

The shallow water equations are widely used in ocean or climate simulations and describe fluid flows where the vertical dimension is much smaller than the horizontal scale. This poster presents the FPGA-specific details of a portable implementation of the shallow water equations, which also runs on CPUs and GPUs. A single-device version was first presented at PASC 2024, with the multi-device extension following at PASC 2026. By using SYCL as a portability layer, the implementation provides a shared numerical code base while supporting architecture-specific memory layouts and communication routines. Unlike CPUs and GPUs, FPGAs lack a cache-based memory hierarchy that mitigates the cost of accessing slow off-chip memory. However, FPGA fabrics incorporate fast RAM blocks, which the shallow water solver uses as static, array-specific caches to store necessary data in fast on-chip memory. When the entire mesh fits within these caches, the FPGA design nearly achieves the ideal throughput of one element per clock cycle. While the communication for CPUs and GPUs relies on MPI, the FPGA communication employs a direct streaming scheme. Combined with the on-chip caches, this enables super-linear scaling in a strong scaling scenario, as the optimal per-FPGA performance is reached when the partition fits entirely into the on-chip caches.

Another strategy to reduce the memory bottleneck is the adoption of block-structured meshes instead of fully unstructured ones. Block-structured meshes consist of locally structured partitions arranged in a globally unstructured topology. This mesh type reduces the complexity of the memory access pattern within a block, promising more efficient bandwidth utilization. This poster presents also an FPGA-Design that utilizes block-structured meshes, currently under development. Initial results show a more efficient global memory bandwidth utilization compared to fully unstructured meshes.