Kokkos: State of Experimental Backends

Rahul Gayatri, NERSC, LBNL
Seyong Lee, ORNL

Kokkos User Group Meeting 2023

December 14, 2023

- OpenMPTarget and OpenACC are the current experimental backends supported.
- Not all features are implemented.
- Not all supported compilers and versions are tested and supported.
- Latest compiler versions might lead to feature or performance regressions. This might be less frequent for non-experimental backends.

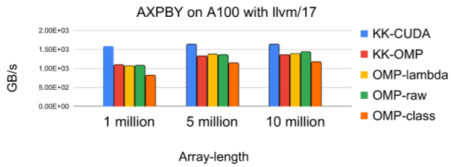
- Uses *target* directives from OpenMP5.0 and above to offload Kokkos parallel patterns onto GPUs.
- Backend is supported on NVIDIA, AMD and Intel architectures
- Supports multiple compilers on a single architecture whenever possible
- Vendor compilers and *clang* on NVIDIA and AMD architecture
- *Intel* compiler on Intel architectures.

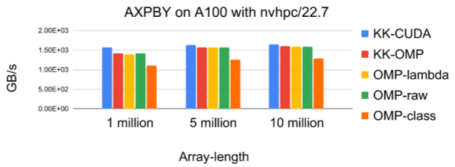
- No 3 level parallelism
- No SIMD support inside *target* region

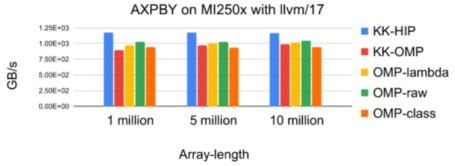
- No true L0 scratch support.
- llvm extensions that allow requesting scratch memory is now in *develop*
- No abort from inside kernel
- Minimum team size of 32 threads

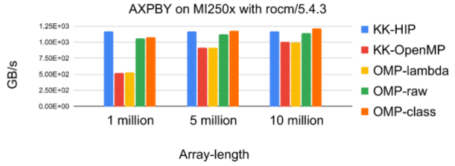
- Actively working with llvm extensions to make OpenMP performant inside a target region.
- PR for using L0 on NVIDIA and AMD GPUs with OpenMPTarget is out
- Soon to be extend this to Intel GPUs.

AXPBY Results (Higher is Better)

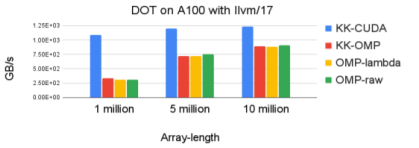


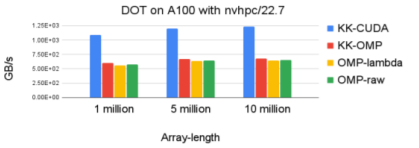


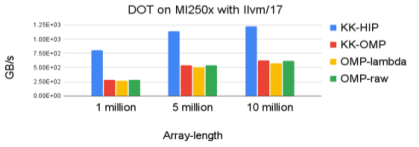


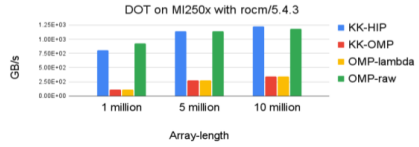


DOT Results (Higher is Better)



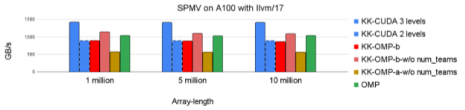


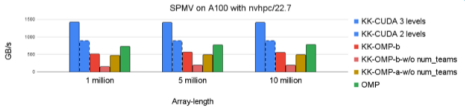


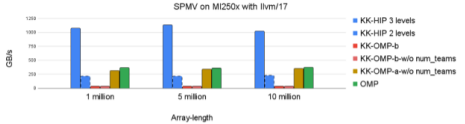


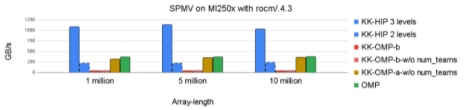
13

SPMV Results









- Uses OpenACC *parallel* directives to offload Kokkos parallel patterns onto GPUs.
- Backend is supported on NVIDIA GPUs, AMD GPUs, and Intel/AMD/IBM CPUs.
- NVIDIA NVHPC compiler (*nvc++*) and *CLACC* compiler[1] for NVIDIA GPUs
- *CLACC* compiler for AMD GPUs
- NVIDIA NVHPC compiler (*nvc++*) compiler for INTEL/AMD/IBM CPUs

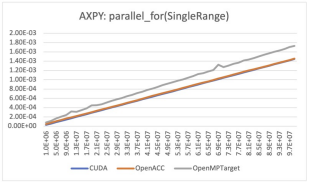
[1] Open-source OpenACC compiler built on LLVM/OpenMP; available in the LLVM-DOE fork (https://github.com/llvm-doe-org/llvm-project/wiki)

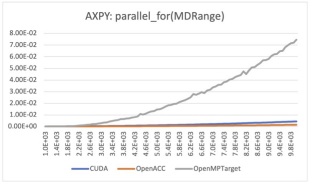
- No atomic operation support (available in the *develop* branch; will be added in the next release (V4.3))
- No scratch pad memory support
- Not-supported team-level APIs: *team_barrier()*, *team_broadcast()*, *team_reduce()*, *team_scan()*, etc.
- No abort from inside kernel
- No custom reduction support
- No SIMD support inside *parallel* region

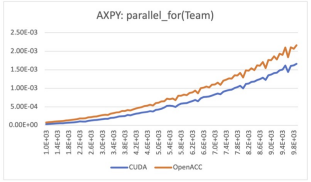
- Add atomic operation support
- Add custom reduction support
- Fully support hierarchial parallelism using CLACC as an OpenACC backend compiler

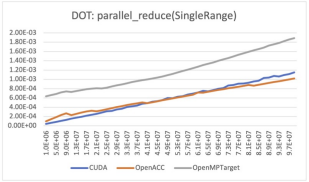
Performance Evaluation on GPUs

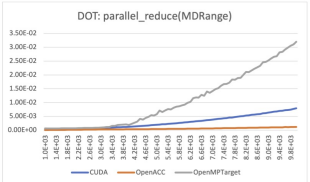
ORNL SUMMIT
- 1x NVIDIA Volta V100 GPU (16 GB)
- OpenMP Target backend (LLVM 17.0.0)
- CUDA backend (CUDA 11.5.2)
- OpenACC backend (NVHPC 22.11)

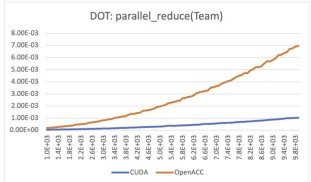












Performance Evaluation on CPUs

- Intel, AMD, and IBM CPUs
- Mini-benchmarks (AXPY and DOT)