CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH zürich

Reductions on GPUs

CSCS Summer School 2025

Andreas Jocksch, Radim Janalik, Prashanth Kanduri, & Ben Cumming

Example reduction on the GPU

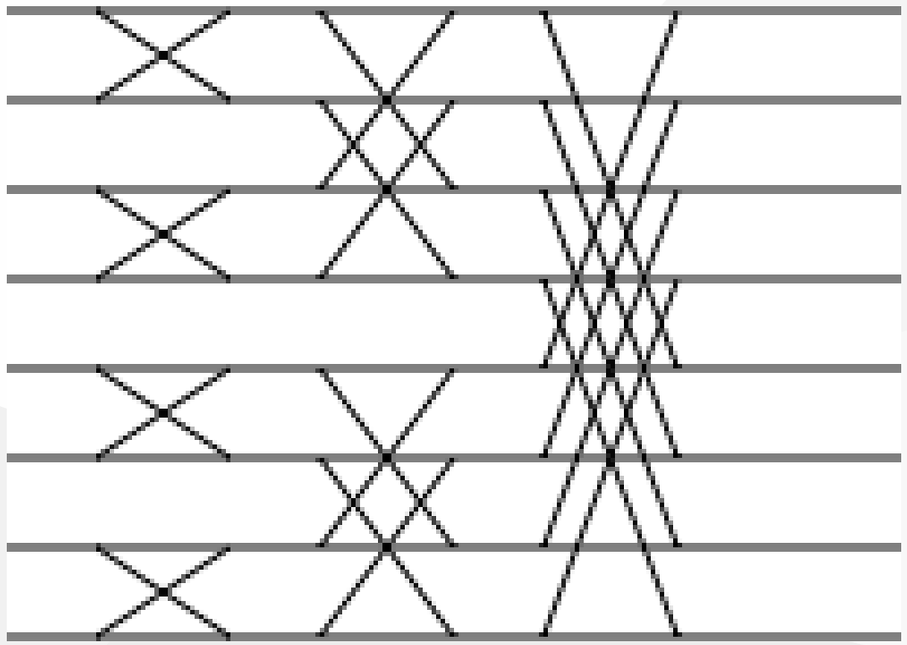
Overview

Reductions

1. Similar to our dot product example.
2. Race condition needs to be avoided.
3. Different levels of parallelism are exploited.

Warp level

- Thread cooperation with shuffle operations. (compute capability 3.x or higher)
  - `__shfl_xor_sync`
  - Butterfly algorithm



Block level

- Thread cooperation with shared memory

Kernel level

- Thread cooperation with atomic operations

Most recent hardware (compute capability 8.x or higher)

- NVIDIA Grace-Hopper cards provide warp reduction operations in hardware

reduction.cu