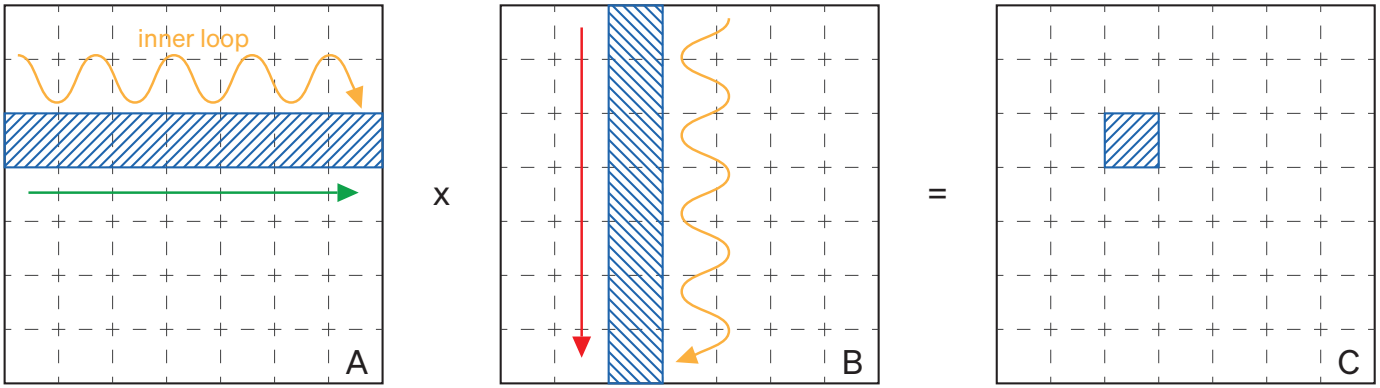
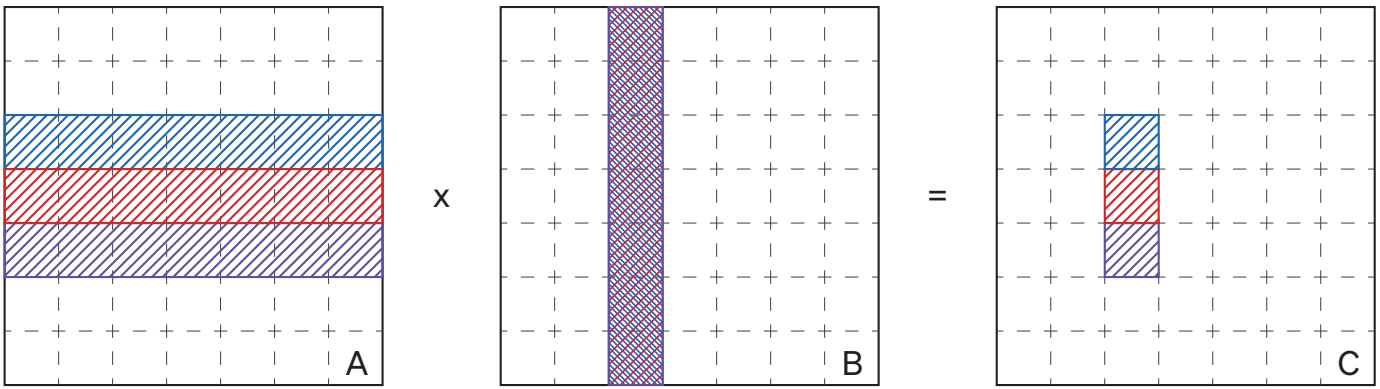
Matrix memory layout:



consecutive values
in memory

non consecutive values
in memory

Naive kernel:

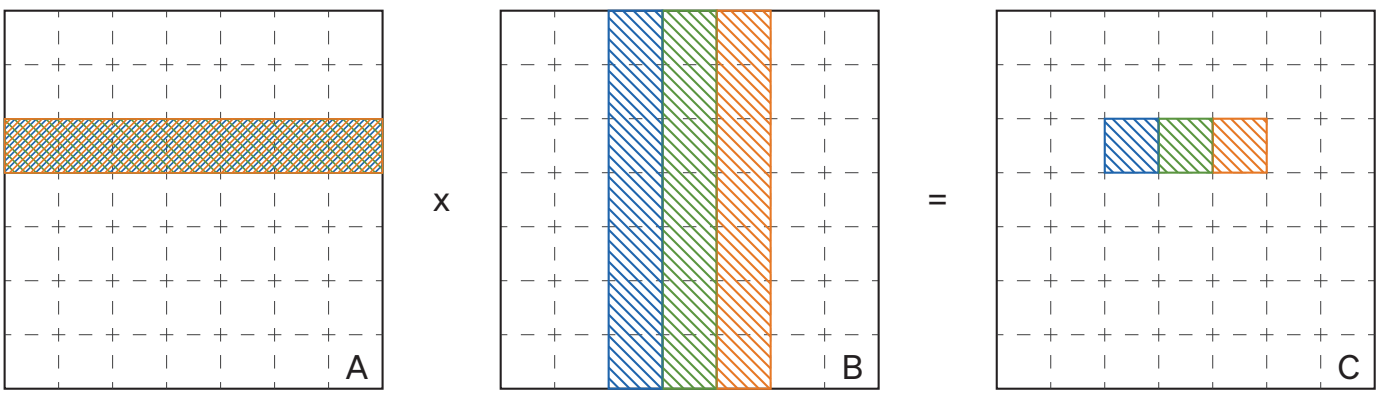


threads access non-consecutive
values => cannot coalesce

all threads access same
values => within-warp broadcast

no benefit in putting these
threads in same warp

Coalescing kernel:



all threads access same
values => within-ward broadcast

threads access consecutive
values => can coalesce

make sure these threads end up
in same warp to exploit coalescing