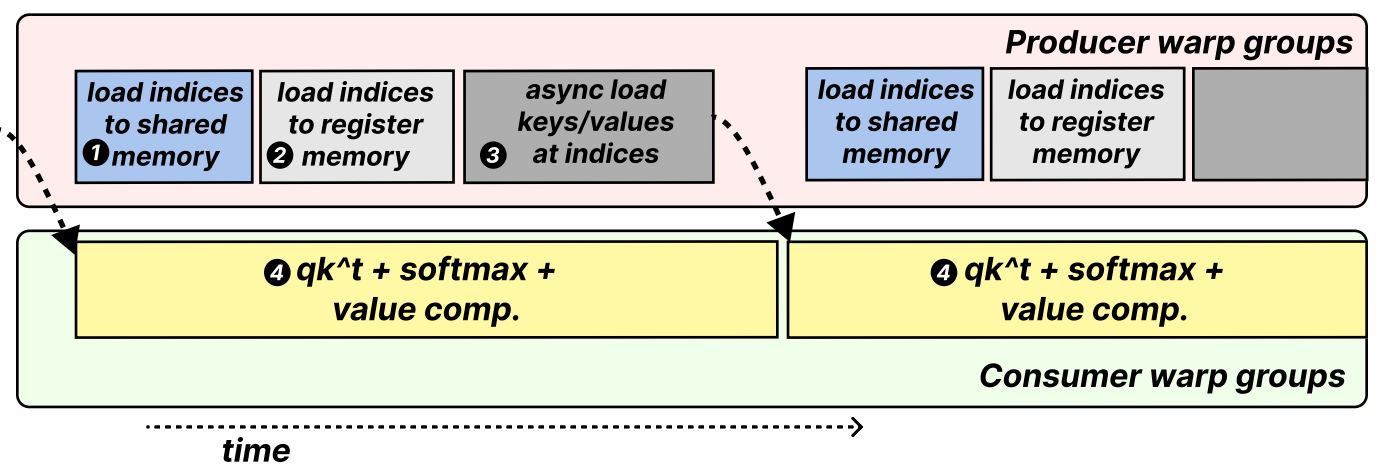
Producer warp groups
load indices to shared memory
❶
load indices to register memory
❷
async load keys/values at indices
❸
load indices to shared memory
load indices to register memory
❹ qk^t + softmax + value comp.
❹ qk^t + softmax + value comp.
Consumer warp groups
time