16bit×4bit (group=128) mul with 72k×18k matrix on NVIDIA A10 - LOCKED BASE CLOCK CUTLASS) torch-nightly bitsandbytes Ideal dense **AWQ** Marlin exllamav2 16 128 32 64 Batchsize

Speedup