

seq4096-hd64-12h-8b | Q=4096, K=4096, head_dim=64, heads=12, batch=8

Kernel	Avg Forward Time (ms)	Relative to Triton Flash
scaled_dot_prod_attention	ERROR	N/A
vectorized_torch	ERROR	N/A
vectorized_torch_compiled	67.40	1.80x slower
flash_attention_triton	37.47	1.00x