

seq8192-hd128-16h-32b | Q=8192, K=8192, head_dim=128, heads=16, batch=32

Kernel	Avg Forward Time (ms)	Relative to Triton Flash
scaled_dot_prod_attention	ERROR	N/A
vectorized_torch	ERROR	N/A
vectorized_torch_compiled	ERROR	N/A
flash_attention_triton	2568.95	1.00x