

seq8192-hd64-12h-8b | Q=8192, K=8192, head\_dim=64, heads=12, batch=8

<b>Kernel</b>	<b>Avg Forward Time (ms)</b>	<b>Relative to Triton Flash</b>
scaled_dot_prod_attention	ERROR	N/A
vectorized_torch	ERROR	N/A
vectorized_torch_compiled	ERROR	N/A
<b>flash_attention_triton</b>	<b>149.89</b>	<b>1.00x</b>