

seq16384-hd128-16h-32b | Q=16384, K=16384, 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
<b>flash_attention_triton</b>	<b>7850.03</b>	<b>1.00x</b>