# seq1024-hd64-12h-8b | Q=1024, K=1024, head_dim=64, heads=12, batch=8

| **Kernel** | **Avg Forward Time (ms)** | **Relative to Triton Flash** |
|---|---|---|
| scaled_dot_prod_attention | 19.69 | 3.03x slower |
| vectorized_torch | 21.04 | 3.24x slower |
| vectorized_torch_compiled | 7.64 | 1.18x slower |
| **flash_attention_triton** | **6.49** | **1.00x** |