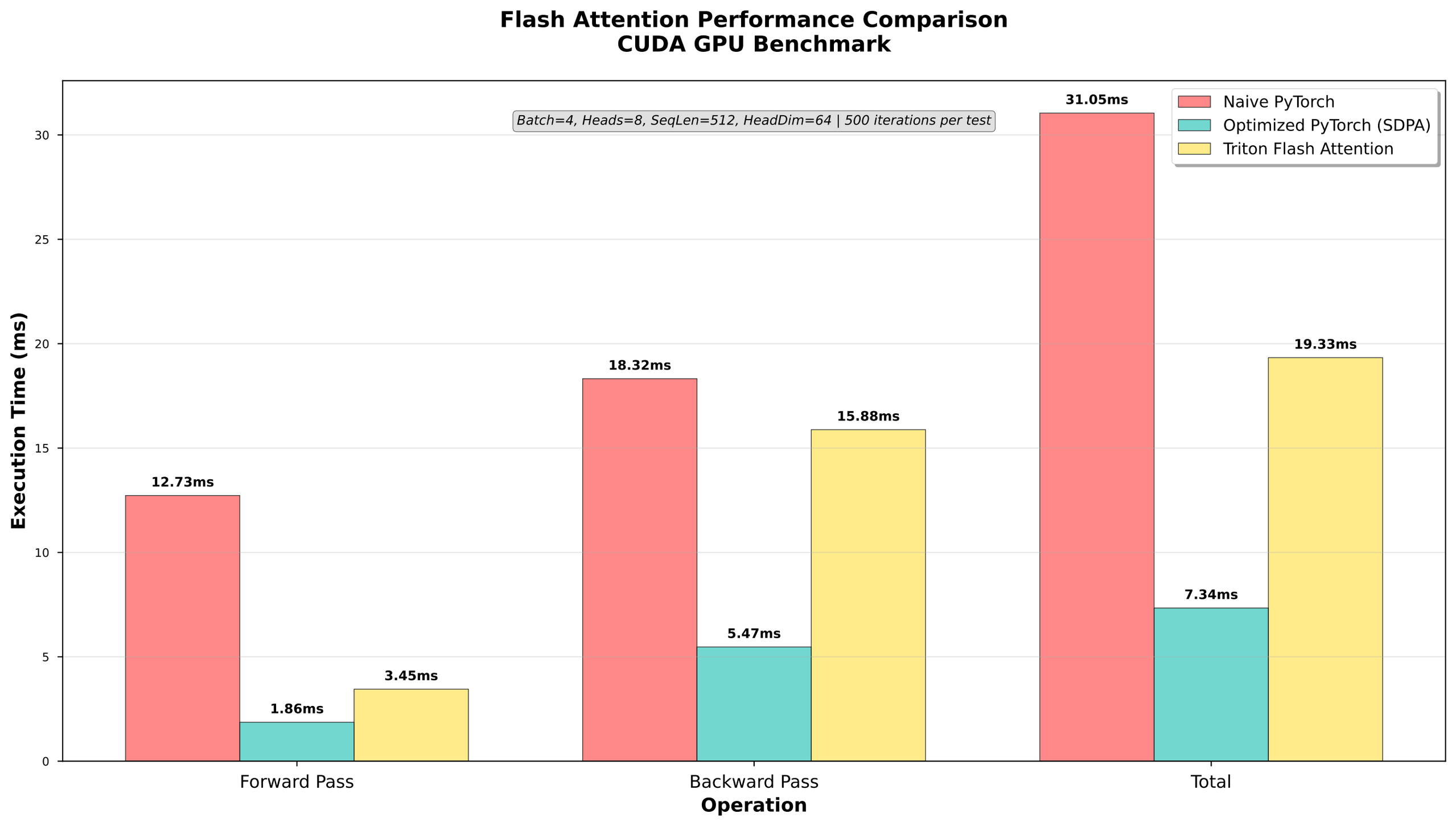
Flash Attention Performance Comparison
CUDA GPU Benchmark
Batch=4, Heads=8, SeqLen=512, HeadDim=64 | 500 iterations per test
Naive PyTorch
Optimized PyTorch (SDPA)
Triton Flash Attention
Execution Time (ms)
0
5
10
15
20
25
30
12.73ms
1.86ms
3.45ms
18.32ms
5.47ms
15.88ms
31.05ms
7.34ms
19.33ms
Forward Pass
Backward Pass
Total
Operation