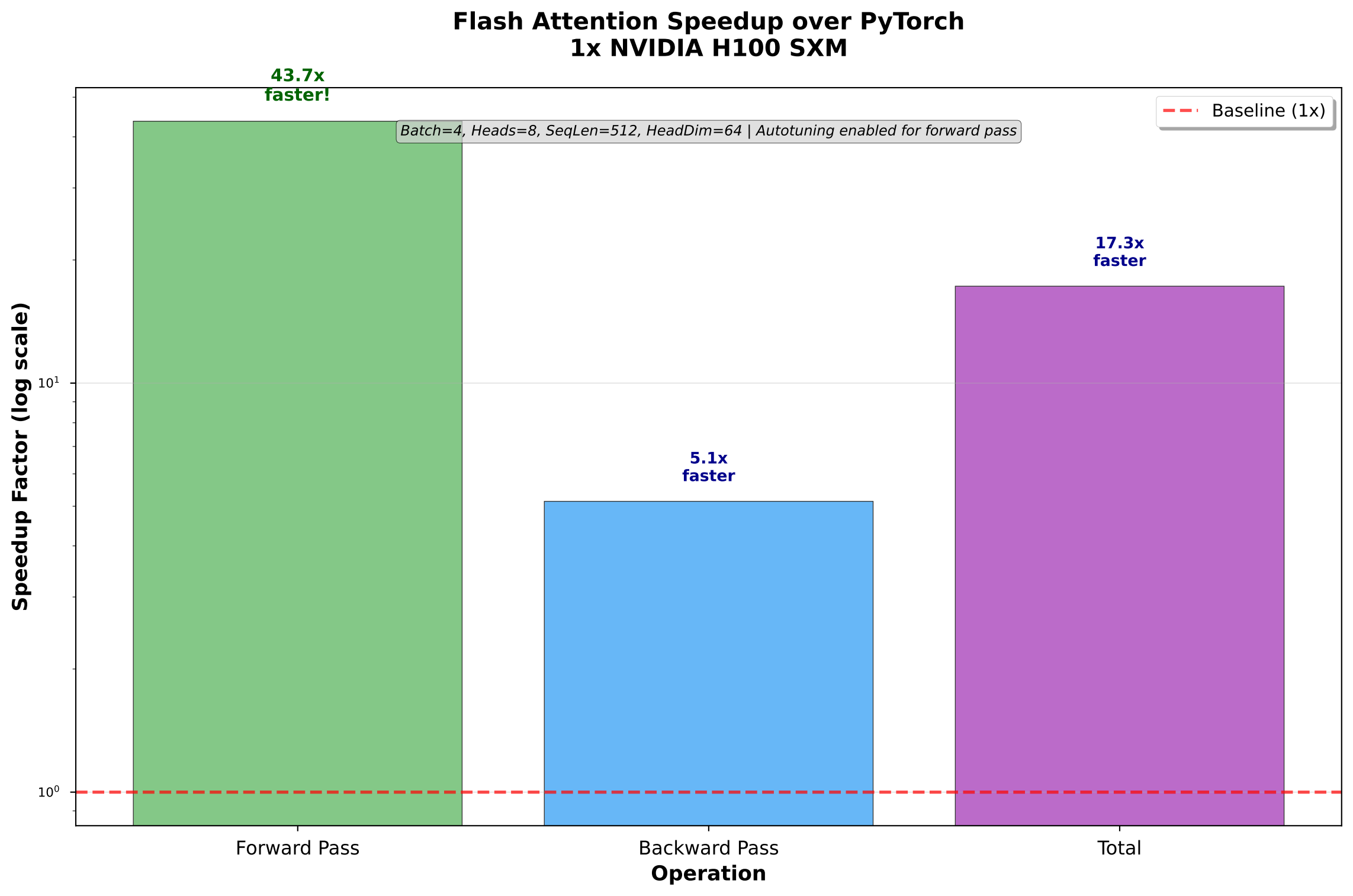
Flash Attention Speedup over PyTorch
1x NVIDIA H100 SXM
43.7x
faster!
Batch=4, Heads=8, SeqLen=512, HeadDim=64 | Autotuning enabled for forward pass
Baseline (1x)
17.3x
faster
5.1x
faster
Speedup Factor (log scale)
$10^1$
$10^0$
Forward Pass
Backward Pass
Total
Operation