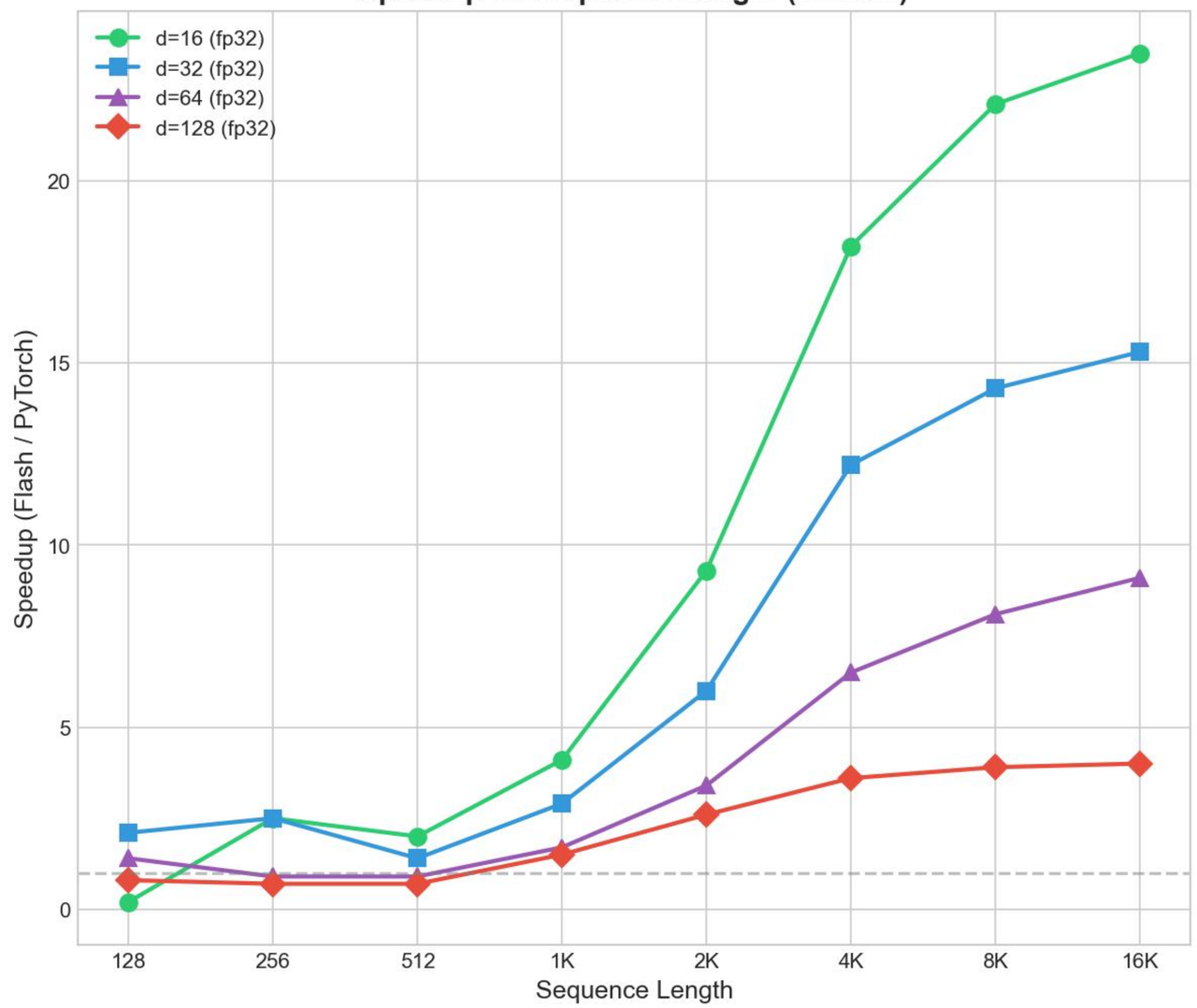
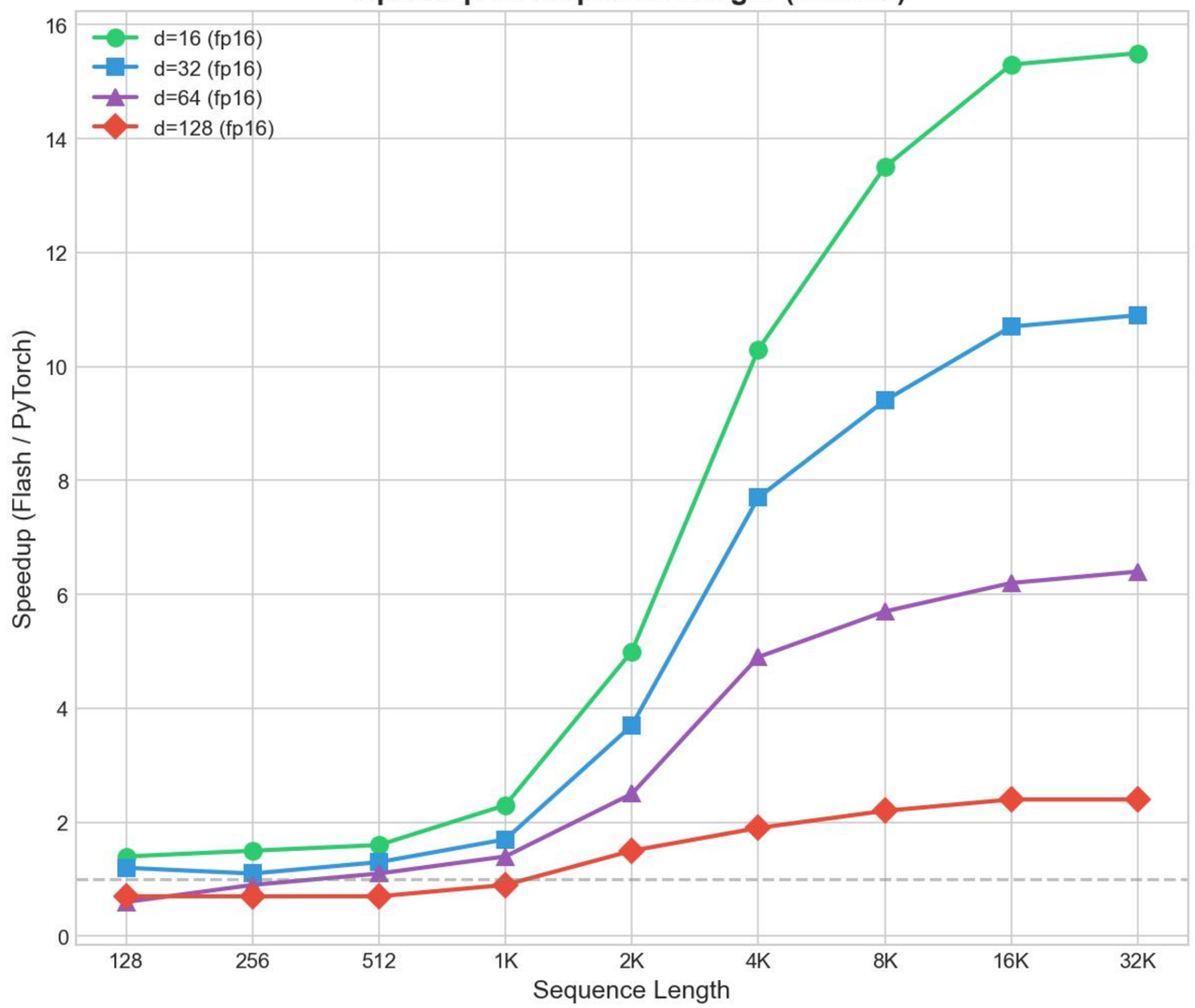
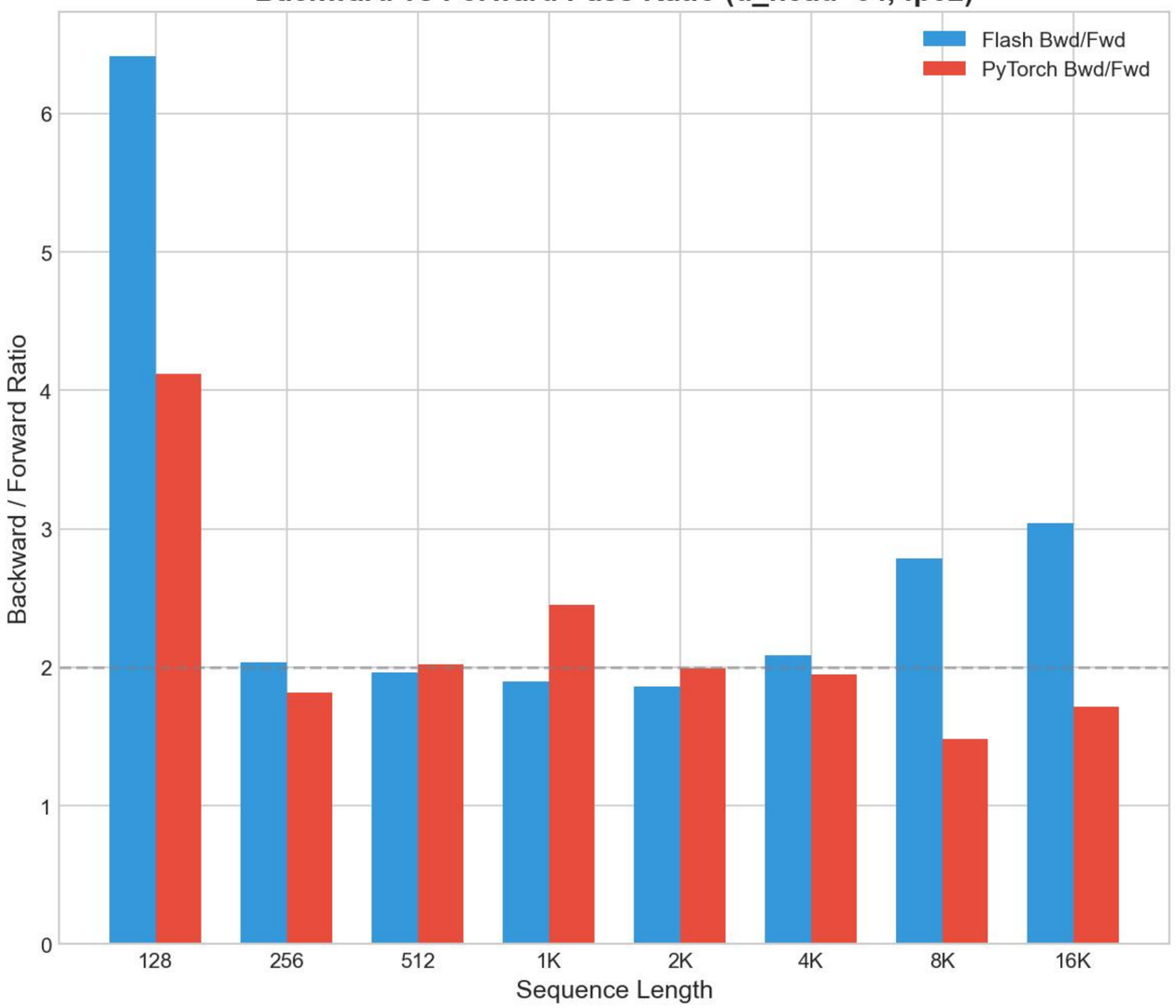
Speedup vs Sequence Length (Float32)



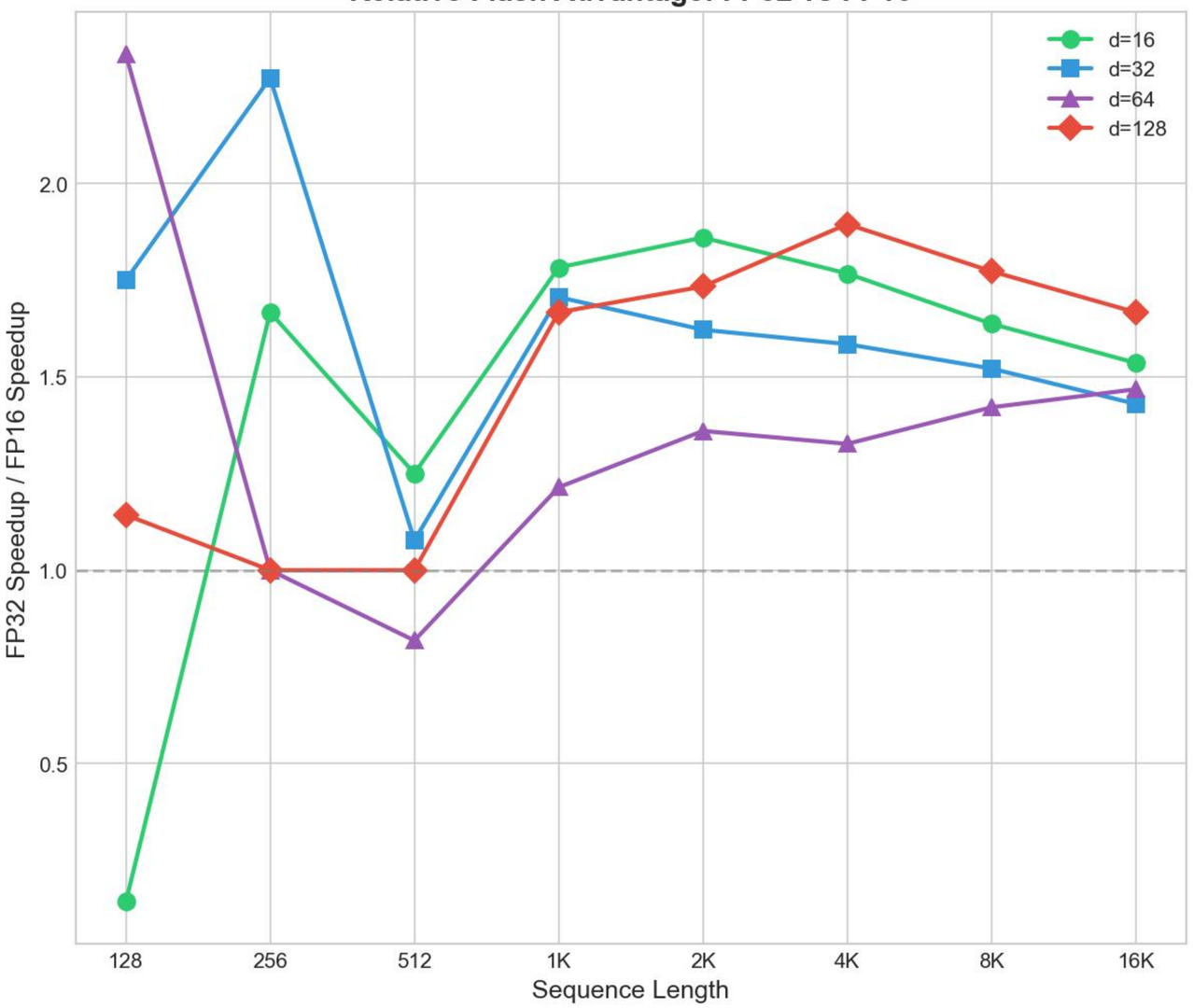
Speedup vs Sequence Length (Float16)

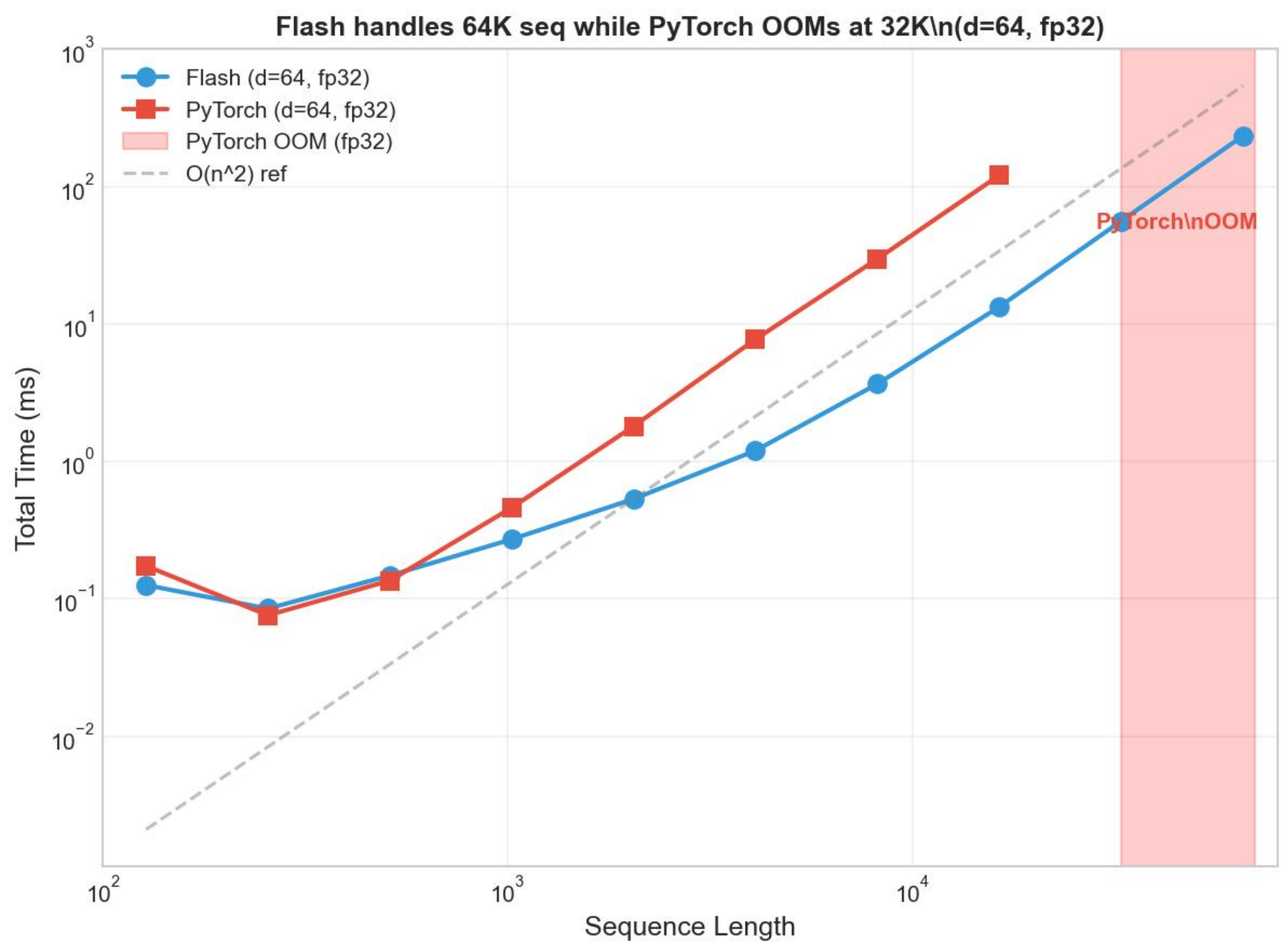


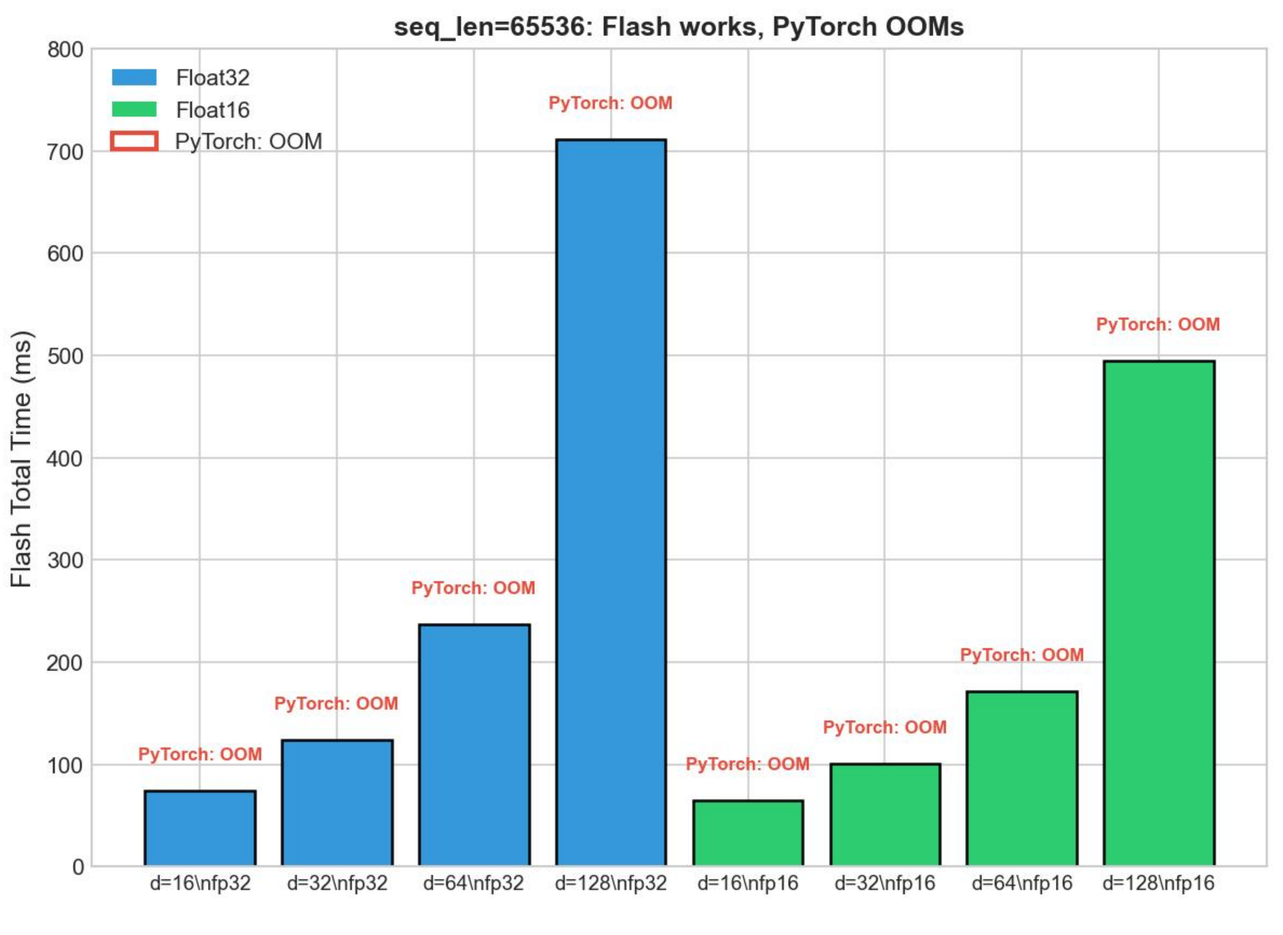
Backward vs Forward Pass Ratio (d_head=64, fp32)



Relative Flash Advantage: FP32 vs FP16




Flash handles 64K seq while PyTorch OOMs at 32K\n(d=64, fp32)
Flash (d=64, fp32)
PyTorch (d=64, fp32)
PyTorch OOM (fp32)
O(n^2) ref
PyTorch\nOOM
Total Time (ms)
Sequence Length


seq_len=65536: Flash works, PyTorch OOMs
Float32
Float16
PyTorch: OOM
Flash Total Time (ms)
d=16\nfp32
d=32\nfp32
d=64\nfp32
d=128\nfp32
d=16\nfp16
d=32\nfp16
d=64\nfp16
d=128\nfp16