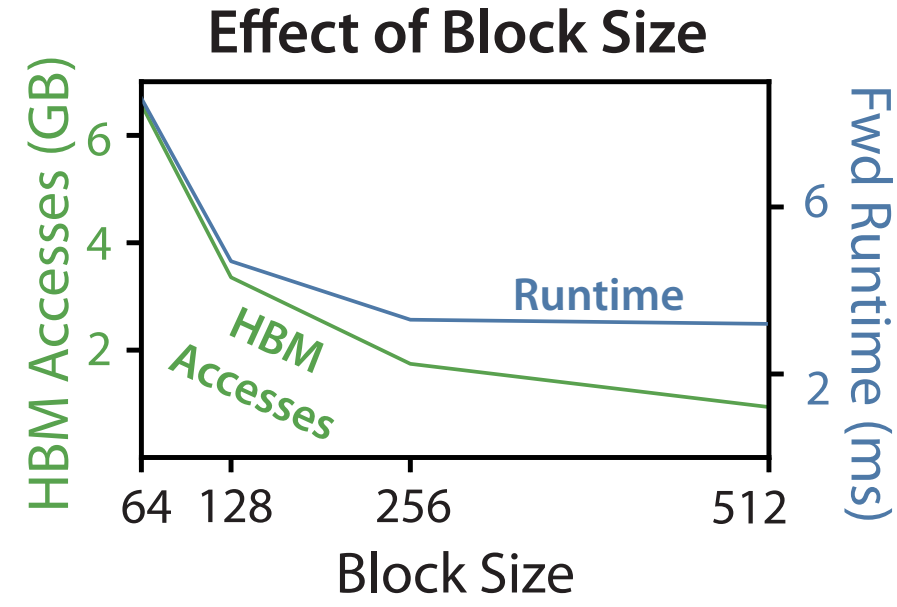
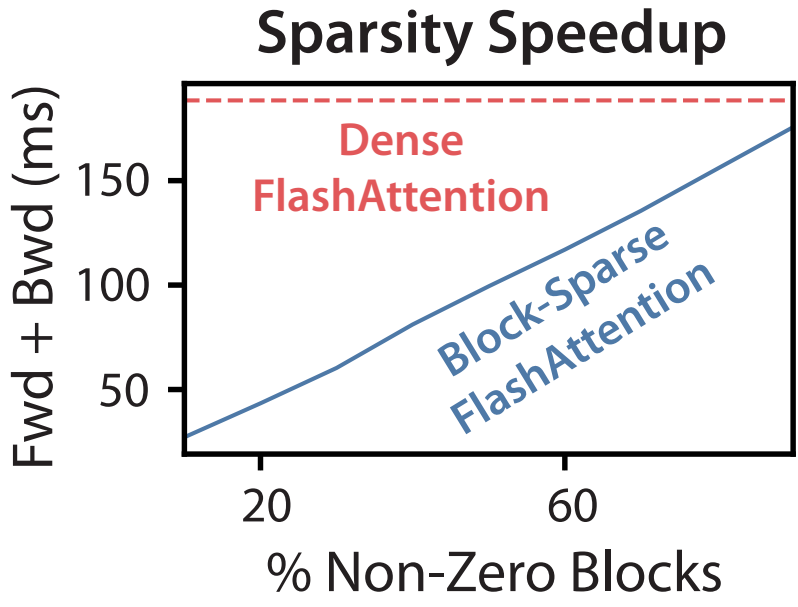

Effect of Block Size
HBM Accesses (GB)
6
4
2
Fwd Runtime (ms)
6
2
Runtime
HBM
Accesses
64
128
256
512
Block Size


Sparsity Speedup
Fwd + Bwd (ms)
150
100
50
Dense
FlashAttention
Block-Sparse
FlashAttention
20
60
% Non-Zero Blocks