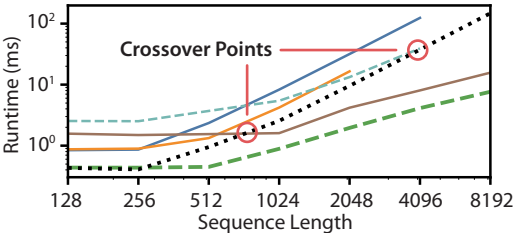
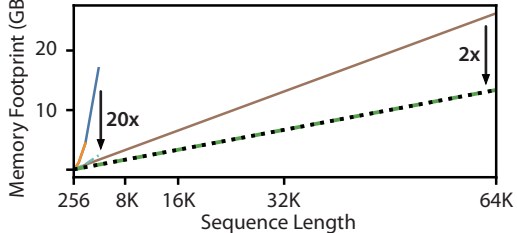


### Attention Runtime (Fwd Pass + Bwd Pass)



### Attention Memory Usage



FlashAttention

Block-Sparse FlashAttention

PyTorch Attention

Megatron Attention

Linformer Attention

OpenAI Sparse Attention