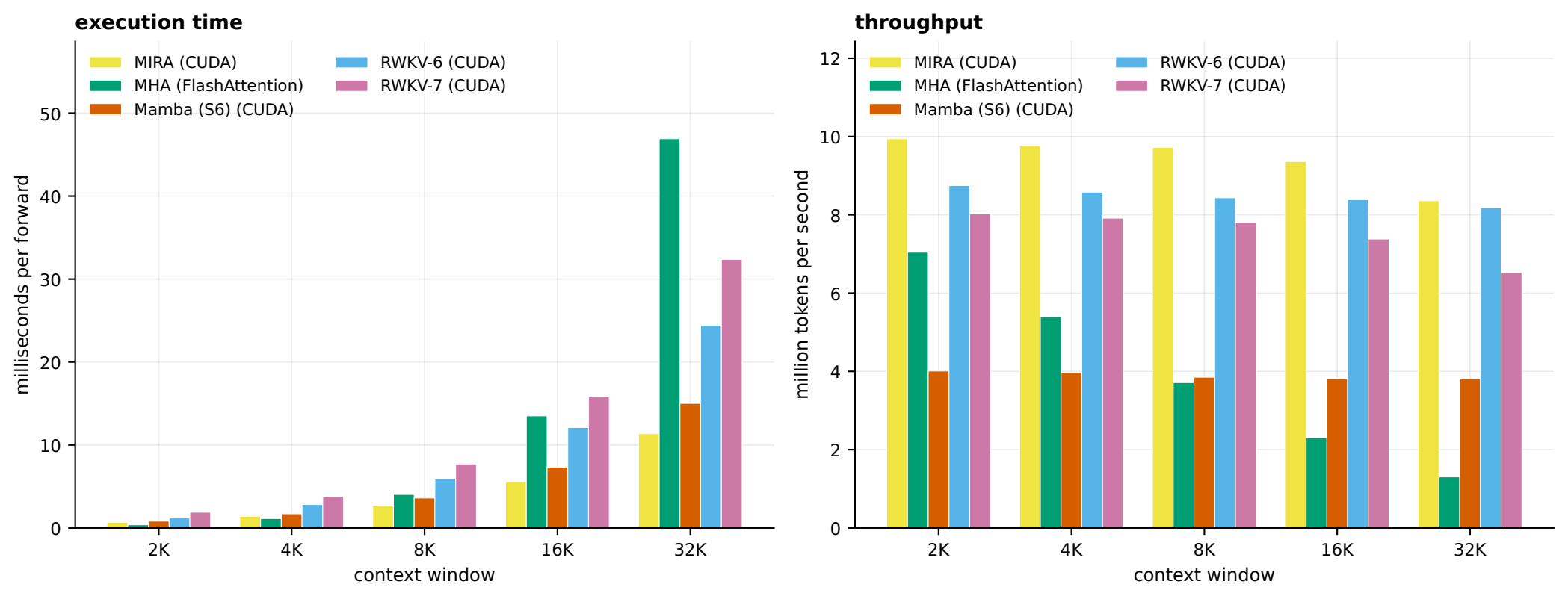

execution time
MIRA (CUDA)
MHA (FlashAttention)
Mamba (S6) (CUDA)
RWKV-6 (CUDA)
RWKV-7 (CUDA)
milliseconds per forward
0
10
20
30
40
50
2K
4K
8K
16K
32K
context window
throughput
MIRA (CUDA)
MHA (FlashAttention)
Mamba (S6) (CUDA)
RWKV-6 (CUDA)
RWKV-7 (CUDA)
million tokens per second
0
2
4
6
8
10
12
2K
4K
8K
16K
32K
context window