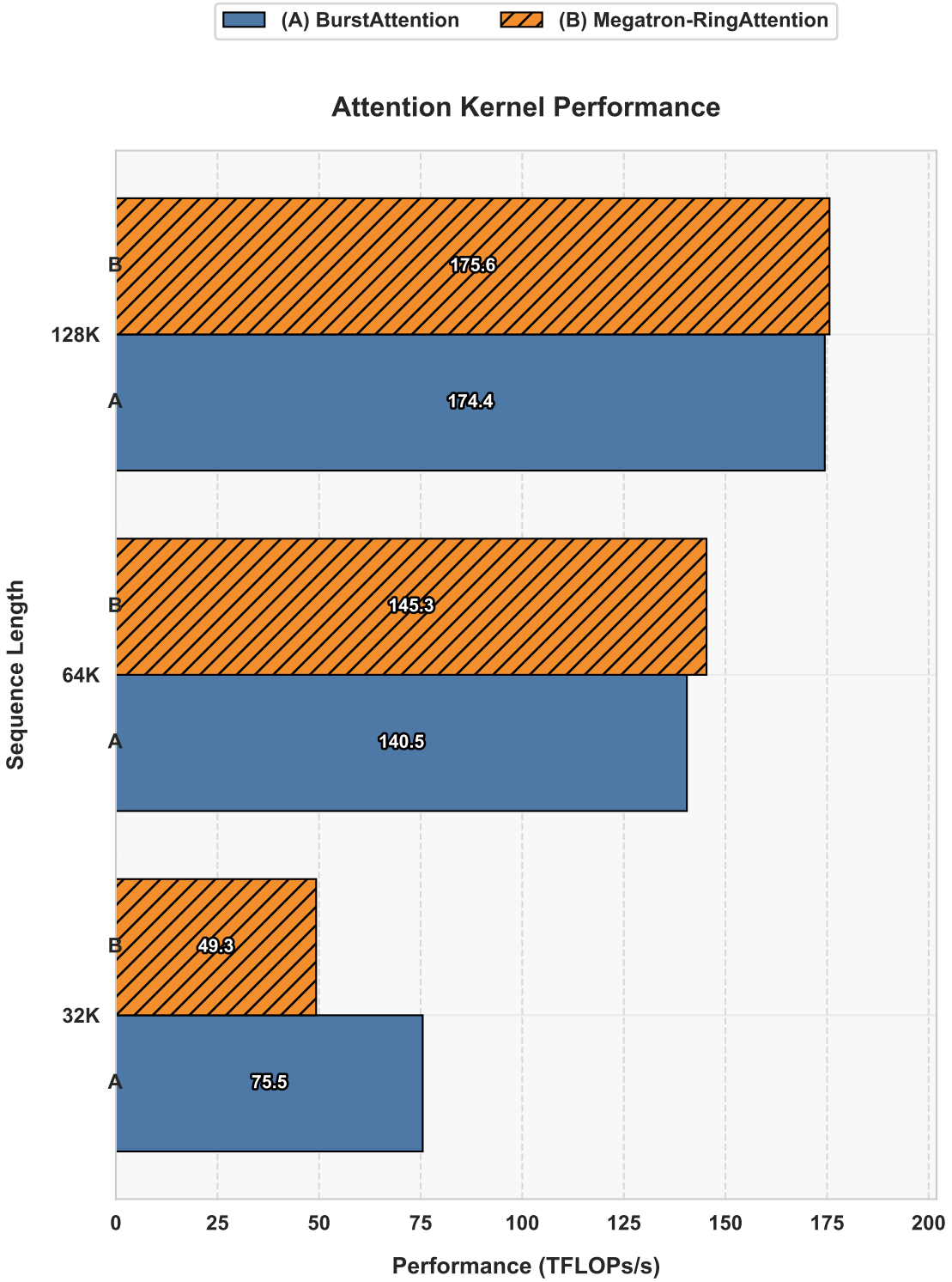
(A) BurstAttention
(B) Megatron-RingAttention
Attention Kernel Performance
Sequence Length
128K
64K
32K
B
A
175.6
174.4
145.3
140.5
49.3
75.5
0
25
50
75
100
125
150
175
200
Performance (TFLOPs/s)