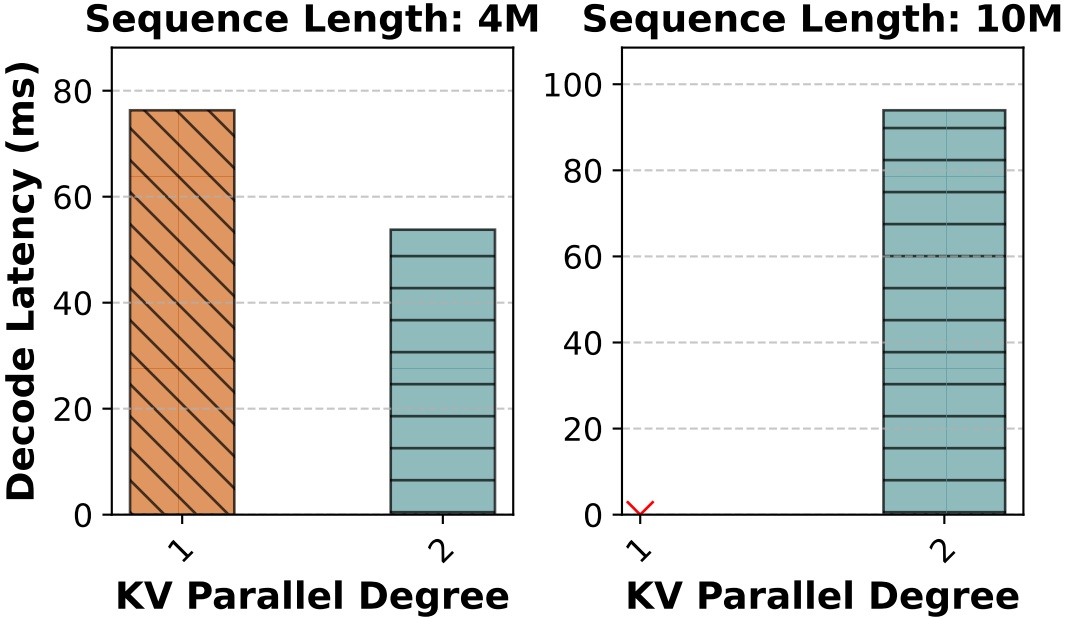
Sequence Length: 4M
Sequence Length: 10M
Decode Latency (ms)
0
20
40
60
80
100
1
2
KV Parallel Degree
KV Parallel Degree