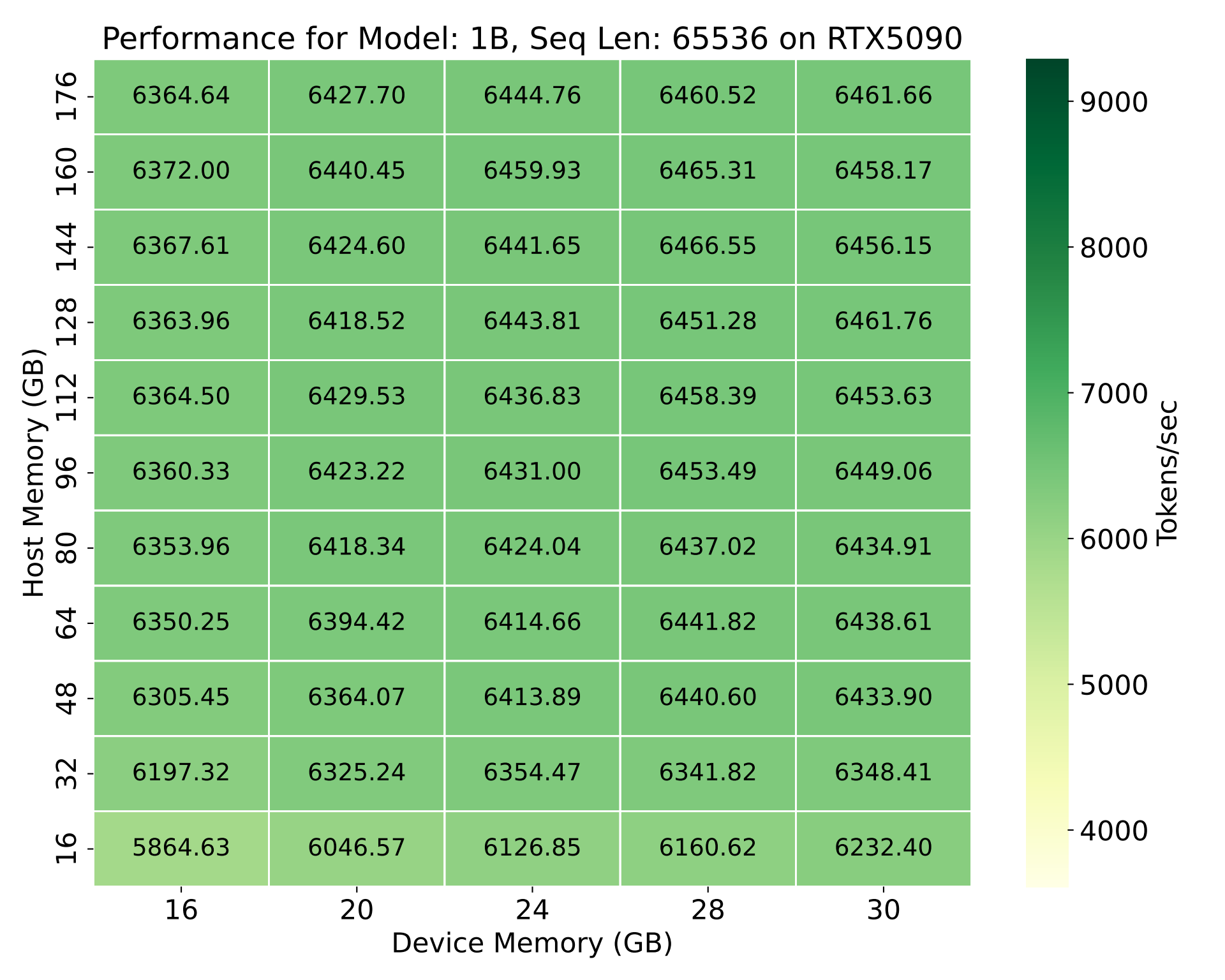
Performance for Model: 1B, Seq Len: 65536 on RTX5090
Host Memory (GB)
176
160
144
128
112
96
80
64
48
32
16
6364.64 6427.70 6444.76 6460.52 6461.66
6372.00 6440.45 6459.93 6465.31 6458.17
6367.61 6424.60 6441.65 6466.55 6456.15
6363.96 6418.52 6443.81 6451.28 6461.76
6364.50 6429.53 6436.83 6458.39 6453.63
6360.33 6423.22 6431.00 6453.49 6449.06
6353.96 6418.34 6424.04 6437.02 6434.91
6350.25 6394.42 6414.66 6441.82 6438.61
6305.45 6364.07 6413.89 6440.60 6433.90
6197.32 6325.24 6354.47 6341.82 6348.41
5864.63 6046.57 6126.85 6160.62 6232.40
16
20
24
28
30
Device Memory (GB)
Tokens/sec
9000
8000
7000
6000
5000
4000