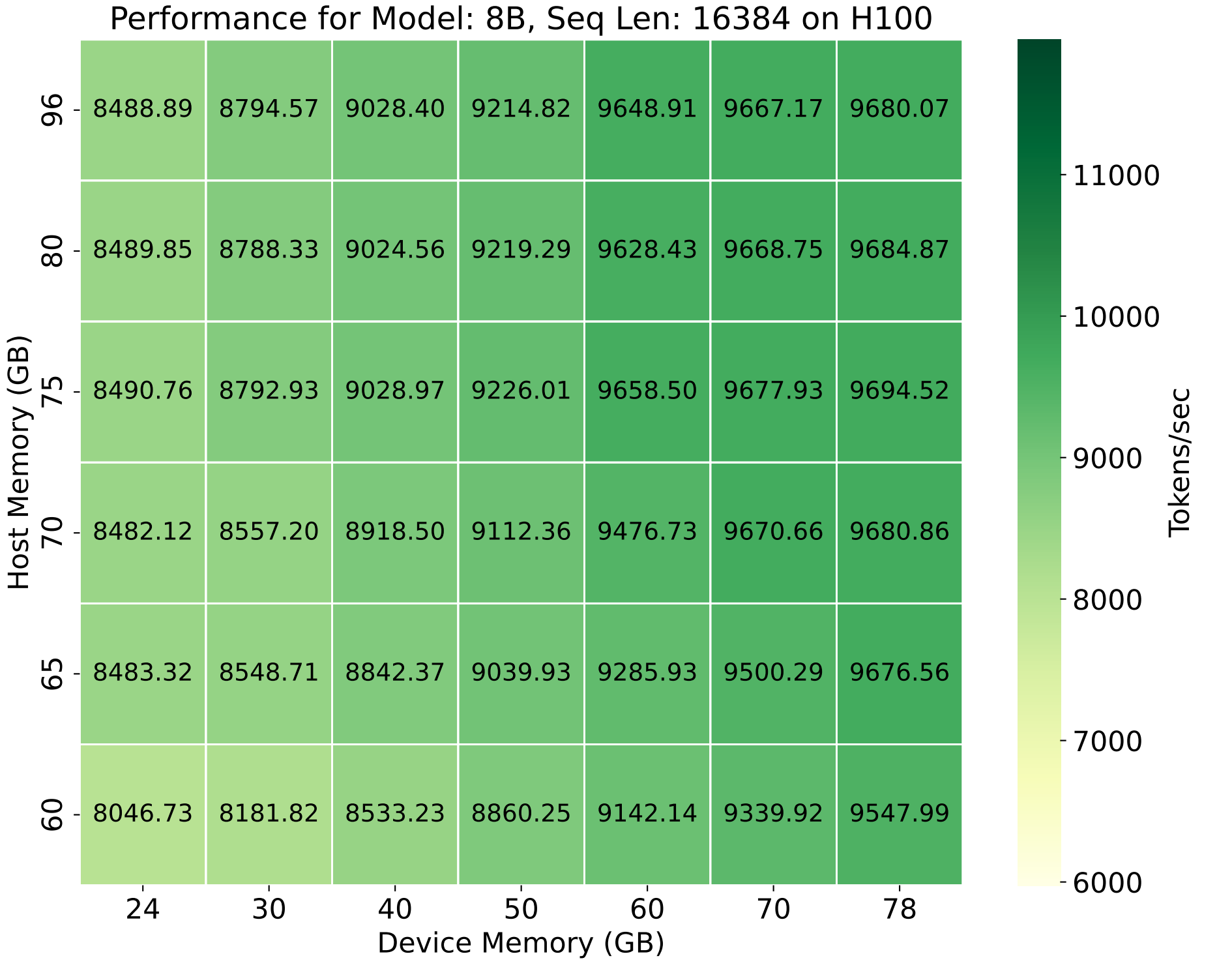
Performance for Model: 8B, Seq Len: 16384 on H100
Host Memory (GB)
96
80
75
70
65
60
8488.89 8794.57 9028.40 9214.82 9648.91 9667.17 9680.07
8489.85 8788.33 9024.56 9219.29 9628.43 9668.75 9684.87
8490.76 8792.93 9028.97 9226.01 9658.50 9677.93 9694.52
8482.12 8557.20 8918.50 9112.36 9476.73 9670.66 9680.86
8483.32 8548.71 8842.37 9039.93 9285.93 9500.29 9676.56
8046.73 8181.82 8533.23 8860.25 9142.14 9339.92 9547.99
24
30
40
50
60
70
78
Device Memory (GB)
Tokens/sec
11000
10000
9000
8000
7000
6000