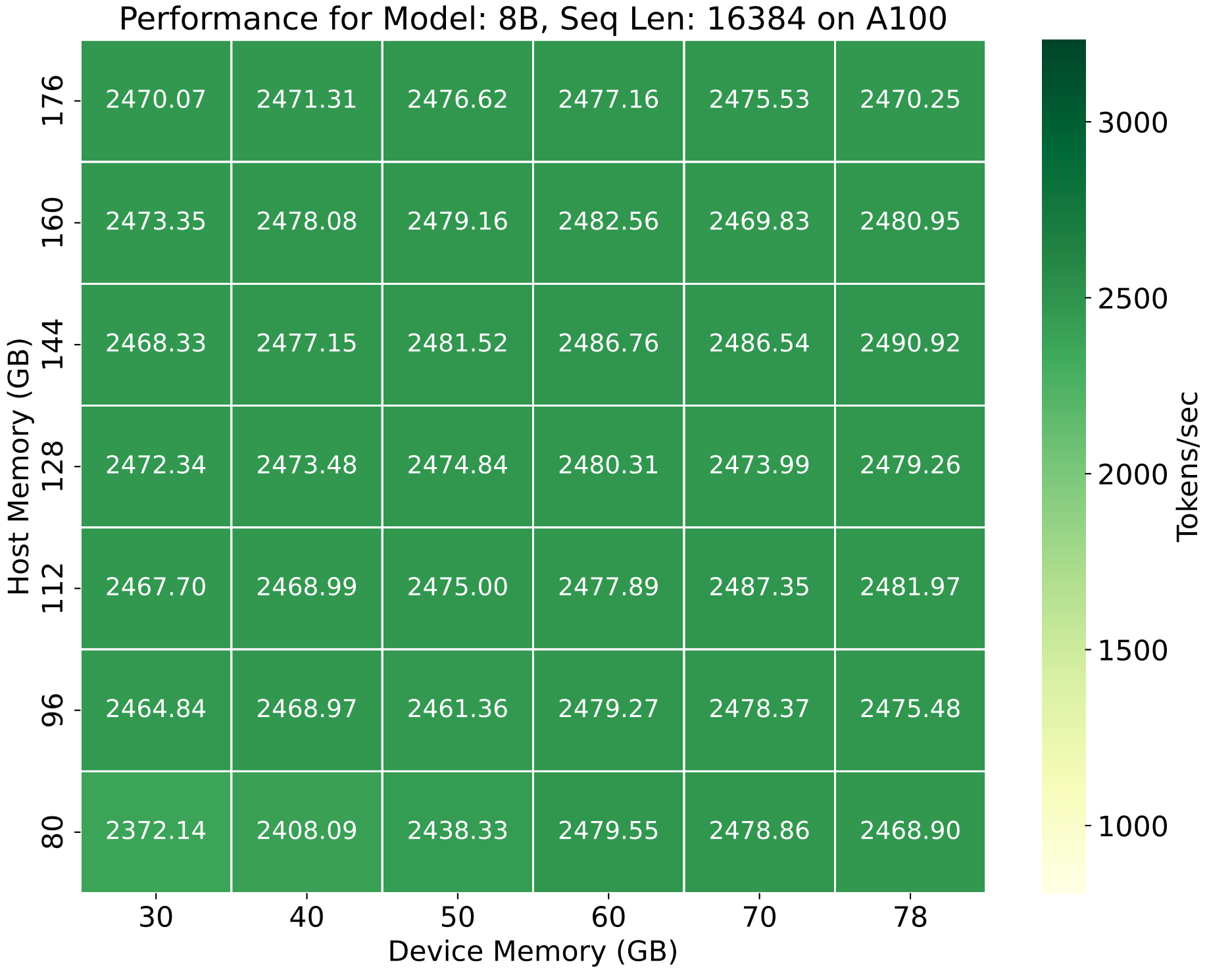
Performance for Model: 8B, Seq Len: 16384 on A100
Host Memory (GB)
176
160
144
128
112
96
80
2470.07
2471.31
2476.62
2477.16
2475.53
2470.25
2473.35
2478.08
2479.16
2482.56
2469.83
2480.95
2468.33
2477.15
2481.52
2486.76
2486.54
2490.92
2472.34
2473.48
2474.84
2480.31
2473.99
2479.26
2467.70
2468.99
2475.00
2477.89
2487.35
2481.97
2464.84
2468.97
2461.36
2479.27
2478.37
2475.48
2372.14
2408.09
2438.33
2479.55
2478.86
2468.90
30
40
50
60
70
78
Device Memory (GB)
Tokens/sec
3000
2500
2000
1500
1000