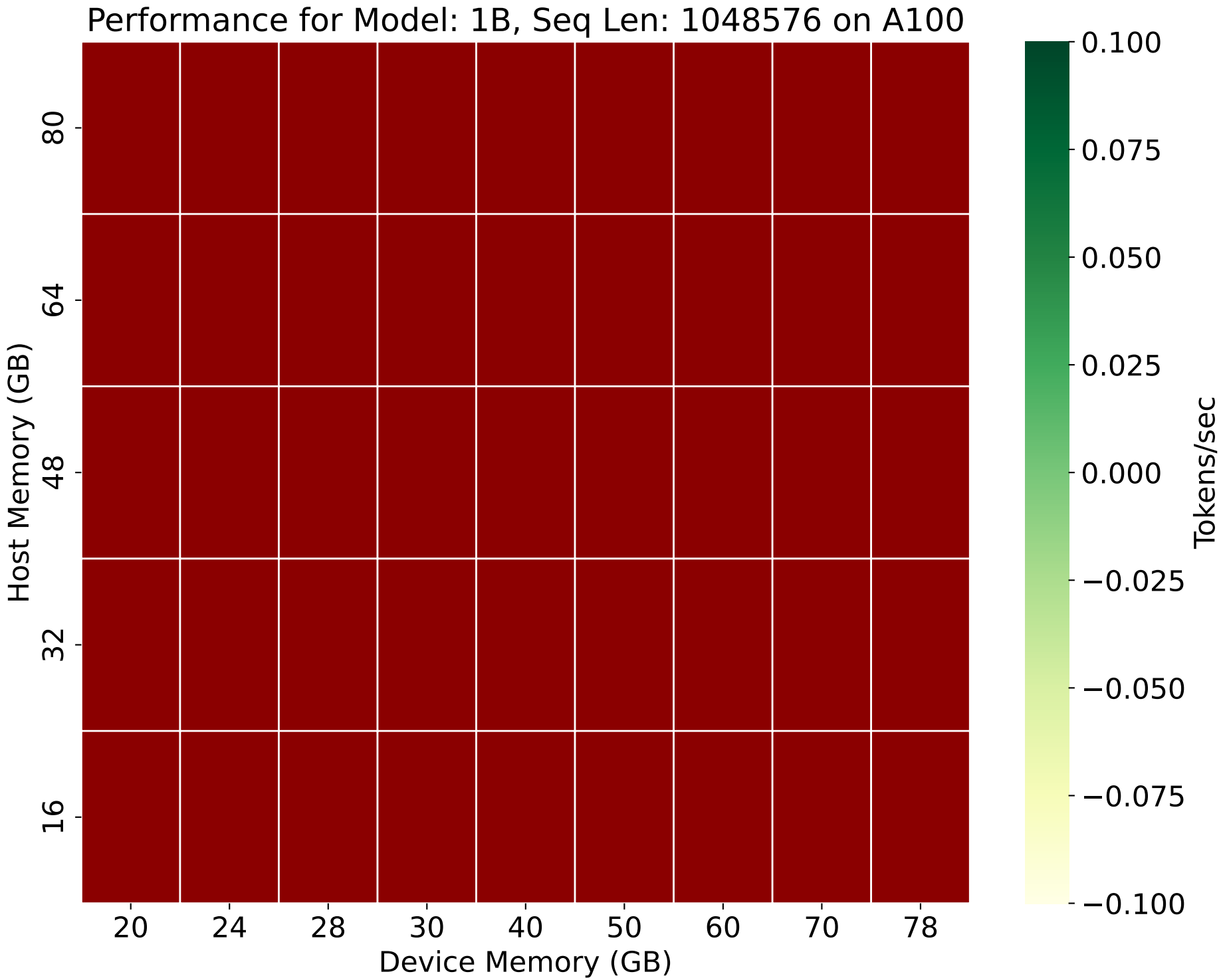
Performance for Model: 1B, Seq Len: 1048576 on A100
Host Memory (GB)
80
64
48
32
16
20
24
28
30
40
50
60
70
78
Device Memory (GB)
0.100
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
−0.100
Tokens/sec