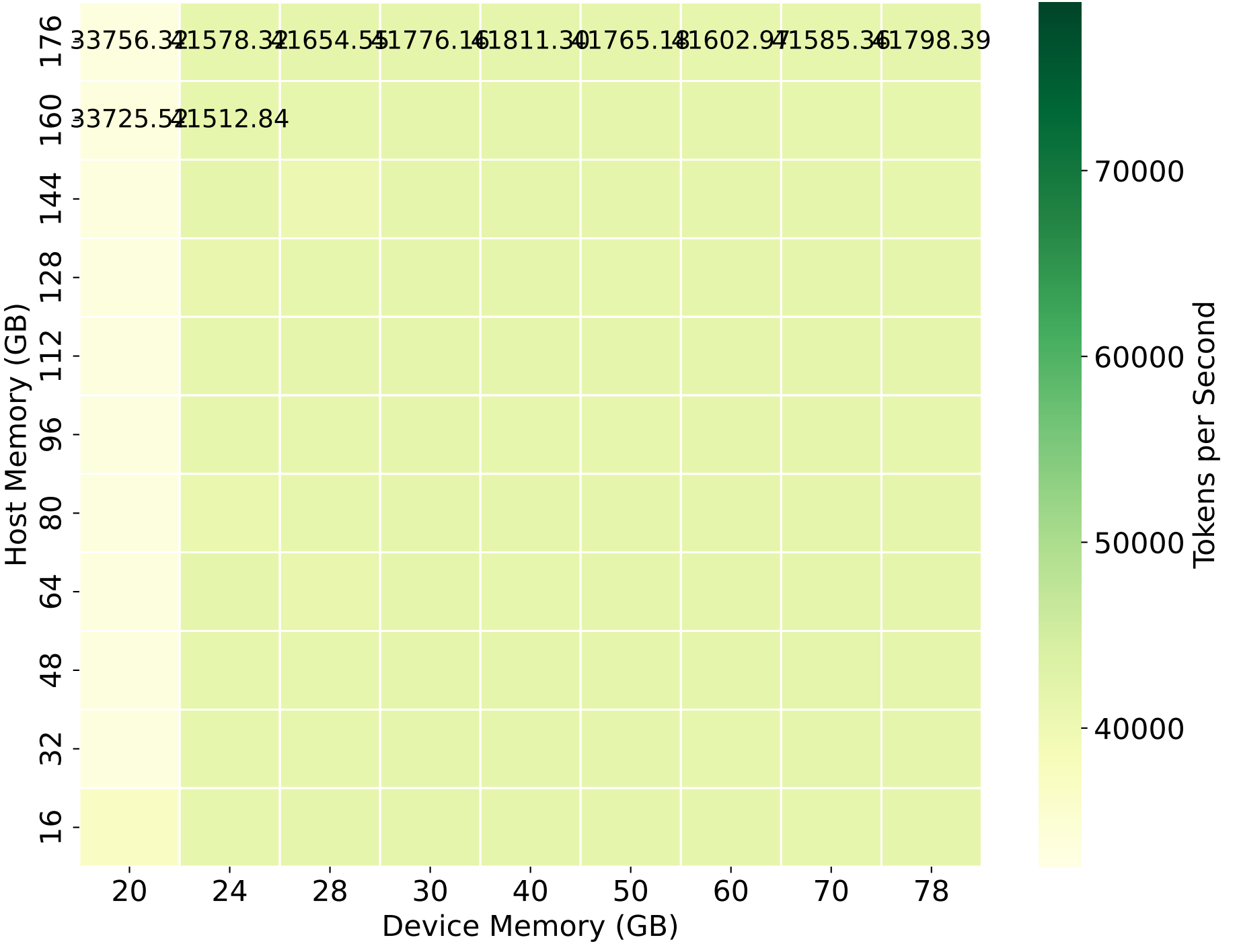
Performance for Model: 1B, Seq Len: 16384 on H100
Host Memory (GB)
176
160
144
128
112
96
80
64
48
32
16
33756.34
21578.34
21654.54
21776.14
21811.30
21765.14
21602.94
21585.36
21798.39
33725.54
21512.84
20
24
28
30
40
50
60
70
78
Device Memory (GB)
Tokens per Second
70000
60000
50000
40000