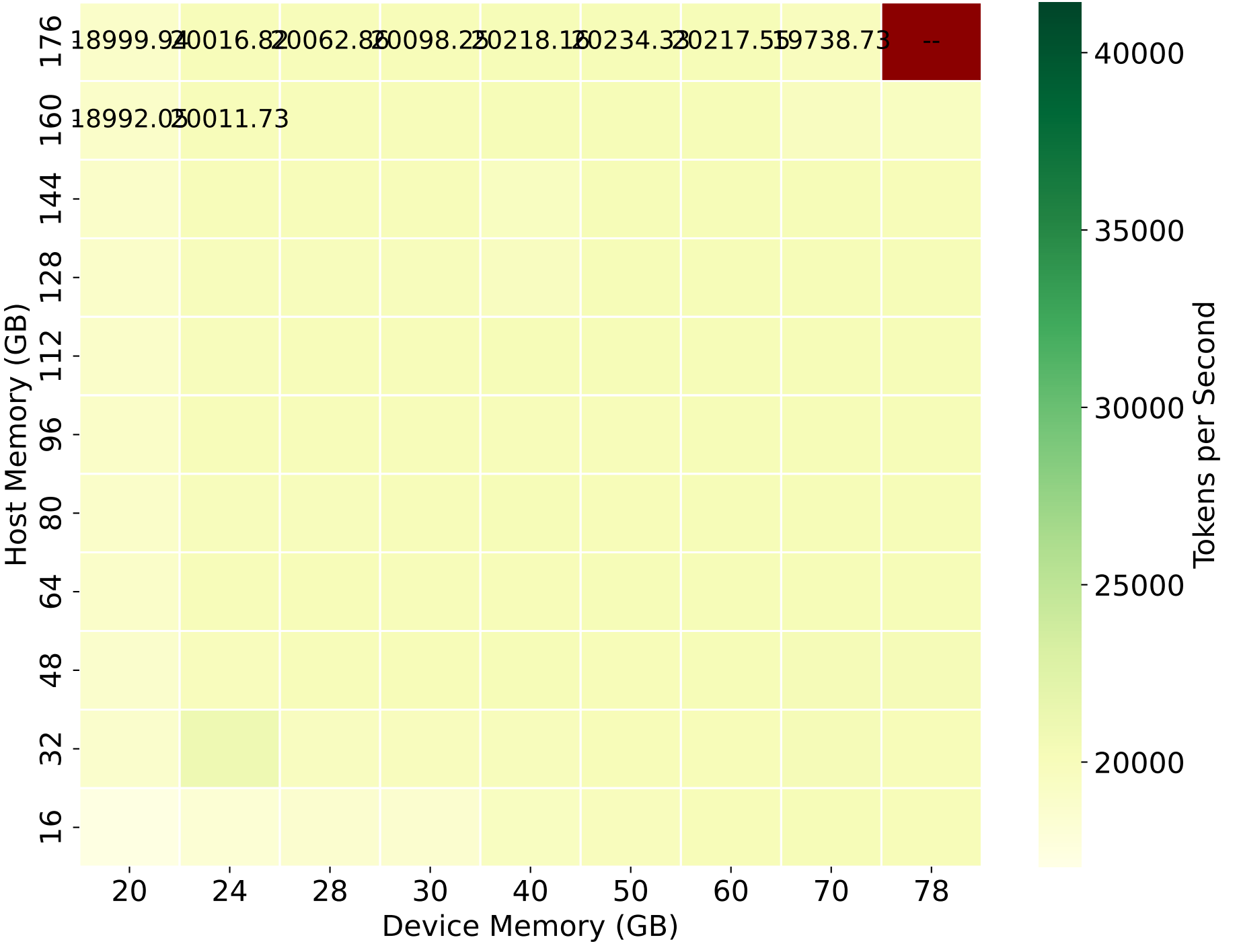
Performance for Model: 1B, Seq Len: 65536 on H100
Host Memory (GB)
176
160
144
128
112
96
80
64
48
32
16
18999.92
40016.82
20062.82
60098.25
50218.12
60234.32
30217.55
59738.73
--
18992.05
50011.73
20
24
28
30
40
50
60
70
78
Device Memory (GB)
Tokens per Second
40000
35000
30000
25000
20000