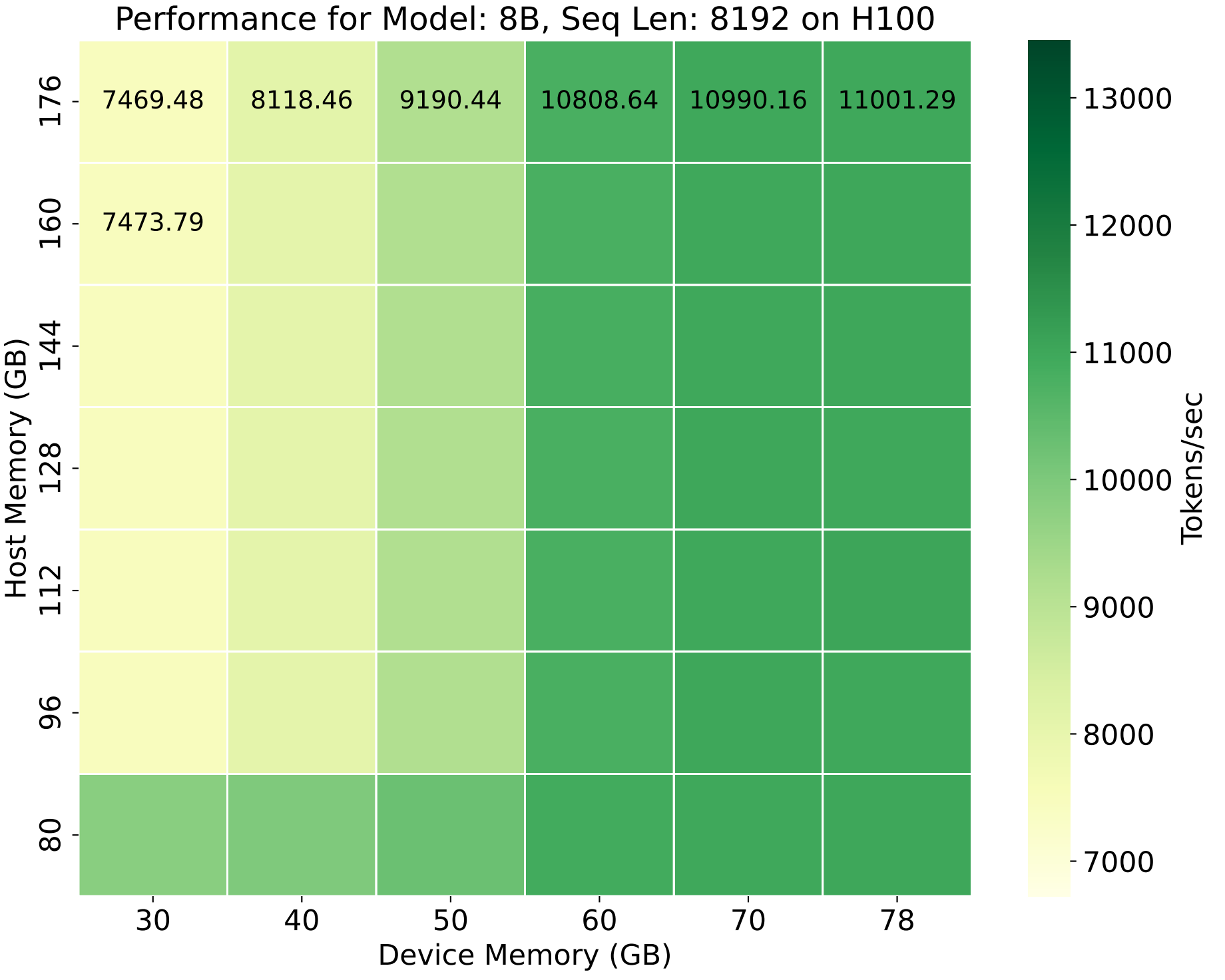
Performance for Model: 8B, Seq Len: 8192 on H100
Host Memory (GB)
176
160
144
128
112
96
80
7469.48
8118.46
9190.44
10808.64
10990.16
11001.29
7473.79
30
40
50
60
70
78
Device Memory (GB)
Tokens/sec
13000
12000
11000
10000
9000
8000
7000