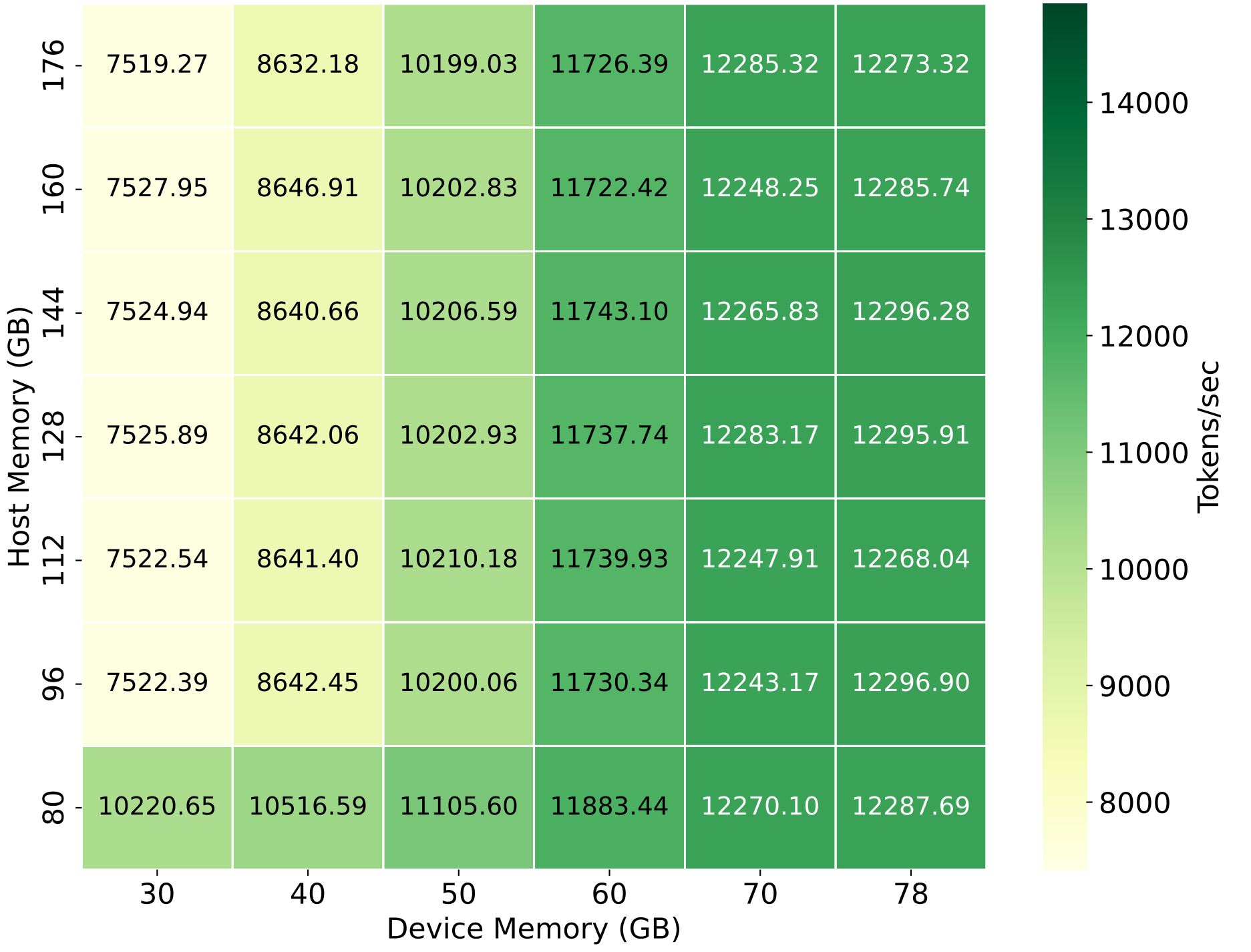
Performance for Model: 8B, Seq Len: 2048 on H100
Host Memory (GB)
176
160
144
128
112
96
80
7519.27
8632.18
10199.03
11726.39
12285.32
12273.32
7527.95
8646.91
10202.83
11722.42
12248.25
12285.74
7524.94
8640.66
10206.59
11743.10
12265.83
12296.28
7525.89
8642.06
10202.93
11737.74
12283.17
12295.91
7522.54
8641.40
10210.18
11739.93
12247.91
12268.04
7522.39
8642.45
10200.06
11730.34
12243.17
12296.90
10220.65
10516.59
11105.60
11883.44
12270.10
12287.69
30
40
50
60
70
78
Device Memory (GB)
Tokens/sec
14000
13000
12000
11000
10000
9000
8000