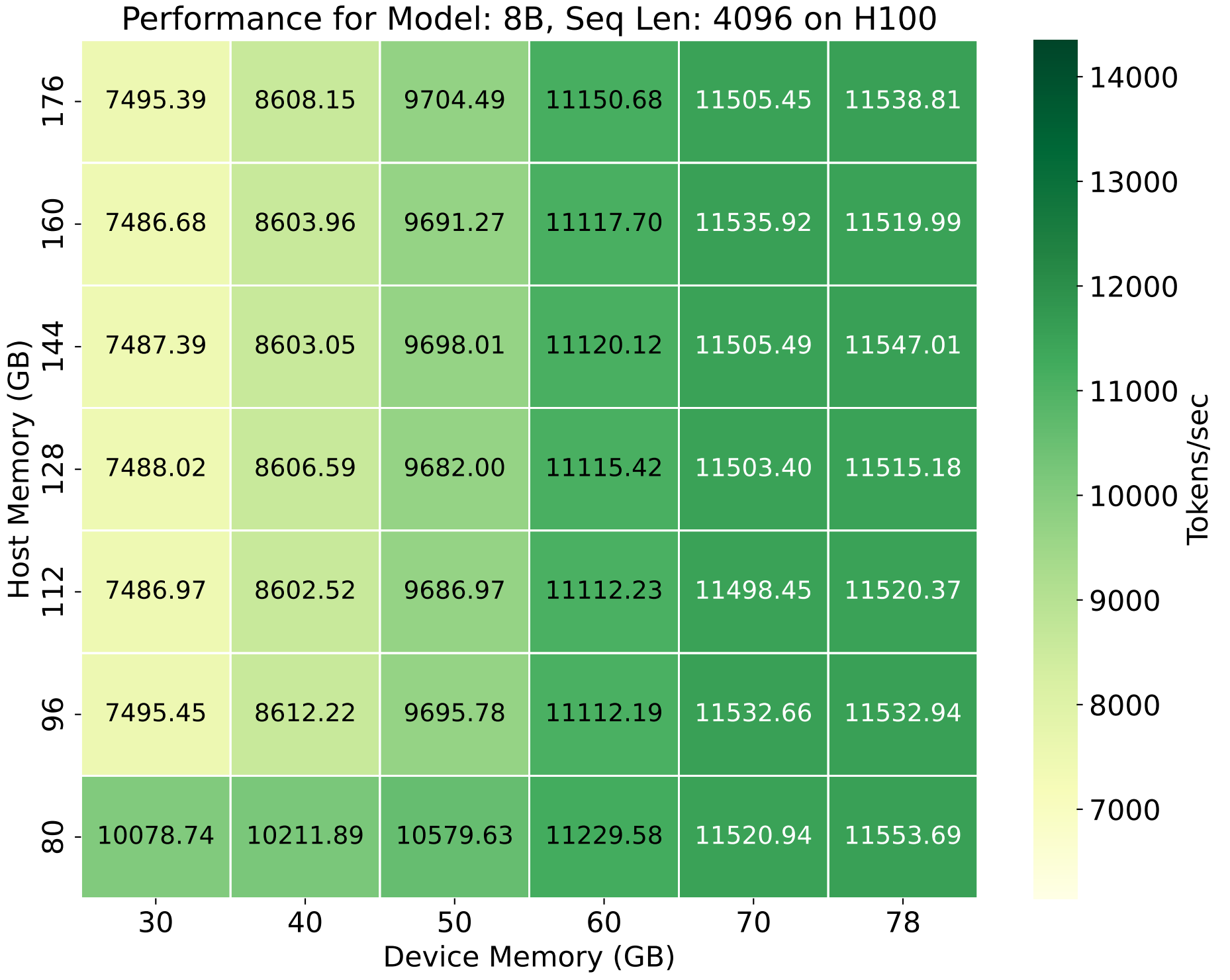
Performance for Model: 8B, Seq Len: 4096 on H100
Host Memory (GB)
176
160
144
128
112
96
80
7495.39
8608.15
9704.49
11150.68
11505.45
11538.81
7486.68
8603.96
9691.27
11117.70
11535.92
11519.99
7487.39
8603.05
9698.01
11120.12
11505.49
11547.01
7488.02
8606.59
9682.00
11115.42
11503.40
11515.18
7486.97
8602.52
9686.97
11112.23
11498.45
11520.37
7495.45
8612.22
9695.78
11112.19
11532.66
11532.94
10078.74
10211.89
10579.63
11229.58
11520.94
11553.69
30
40
50
60
70
78
Device Memory (GB)
14000
13000
12000
11000
10000
9000
8000
7000
Tokens/sec