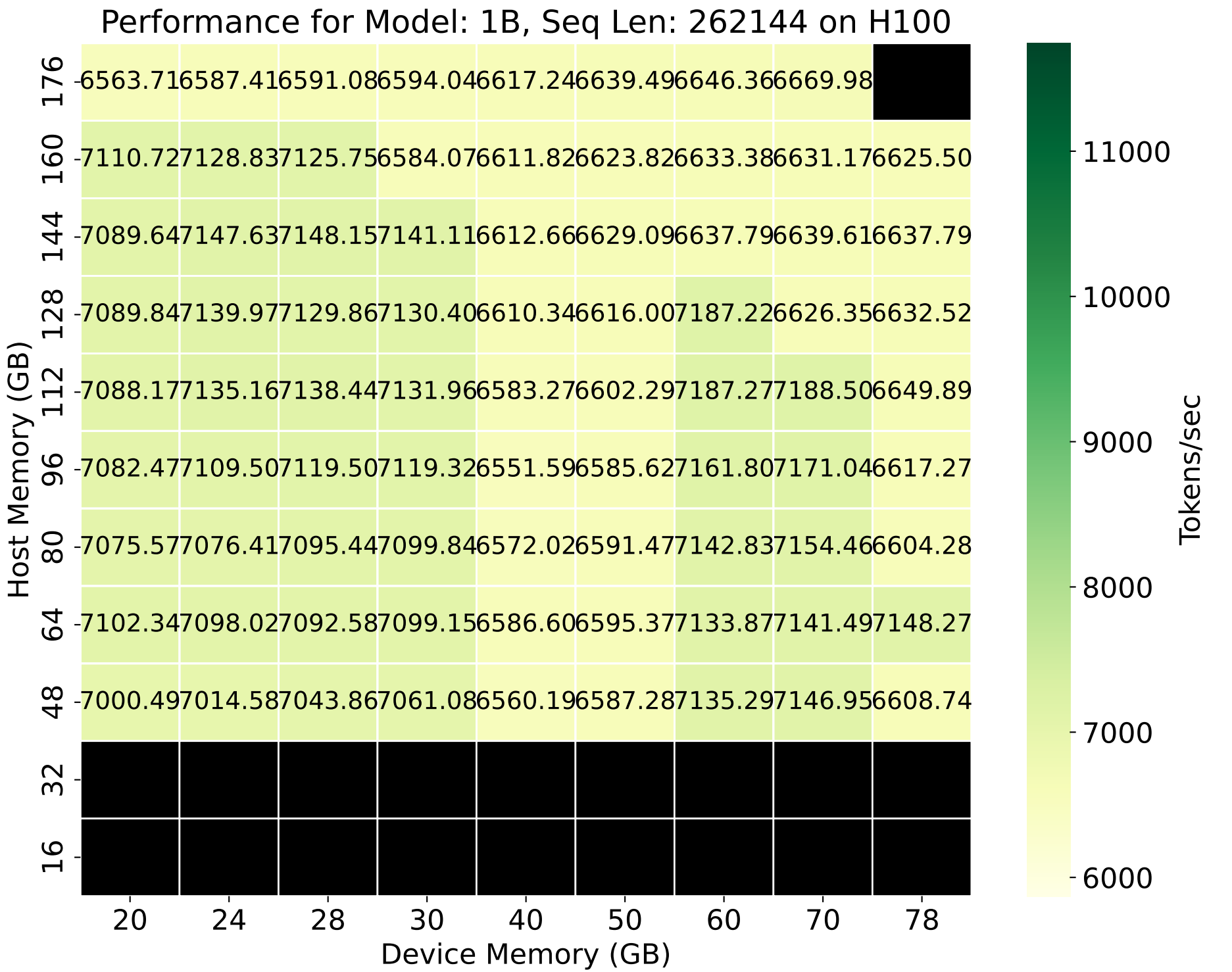
Performance for Model: 1B, Seq Len: 262144 on H100
Host Memory (GB)
176
6563.71
6587.41
6591.08
6594.04
6617.24
6639.49
6646.36
6669.98
160
7110.72
7128.83
7125.75
6584.07
6611.82
6623.82
6633.38
6631.17
6625.50
144
7089.64
7147.63
7148.15
7141.11
6612.66
6629.09
6637.79
6639.61
6637.79
128
7089.84
7139.97
7129.86
7130.40
6610.34
6616.00
7187.22
6626.35
6632.52
112
7088.17
7135.16
7138.44
7131.96
6583.27
6602.29
7187.27
7188.50
6649.89
96
7082.47
7109.50
7119.50
7119.32
6551.59
6585.62
7161.80
7171.04
6617.27
80
7075.57
7076.41
7095.44
7099.84
6572.02
6591.47
7142.83
7154.46
6604.28
64
7102.34
7098.02
7092.58
7099.15
6586.60
6595.37
7133.87
7141.49
7148.27
48
7000.49
7014.58
7043.86
7061.08
6560.19
6587.28
7135.29
7146.95
6608.74
32
16
20
24
28
30
40
50
60
70
78
Device Memory (GB)
Tokens/sec
11000
10000
9000
8000
7000
6000