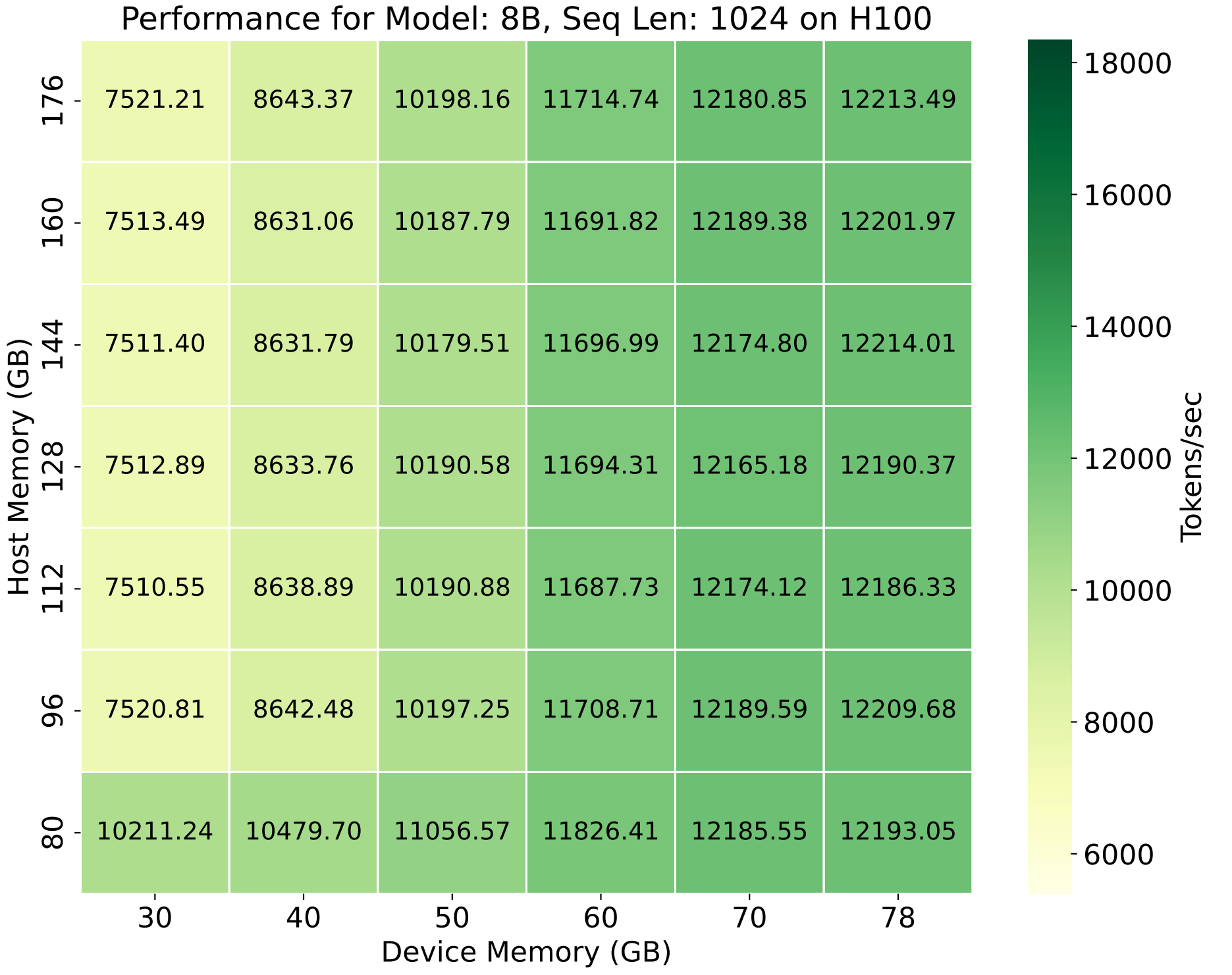
Performance for Model: 8B, Seq Len: 1024 on H100
Host Memory (GB)
176
160
144
128
112
96
80
7521.21
8643.37
10198.16
11714.74
12180.85
12213.49
7513.49
8631.06
10187.79
11691.82
12189.38
12201.97
7511.40
8631.79
10179.51
11696.99
12174.80
12214.01
7512.89
8633.76
10190.58
11694.31
12165.18
12190.37
7510.55
8638.89
10190.88
11687.73
12174.12
12186.33
7520.81
8642.48
10197.25
11708.71
12189.59
12209.68
10211.24
10479.70
11056.57
11826.41
12185.55
12193.05
30
40
50
60
70
78
Device Memory (GB)
Tokens/sec
18000
16000
14000
12000
10000
8000
6000