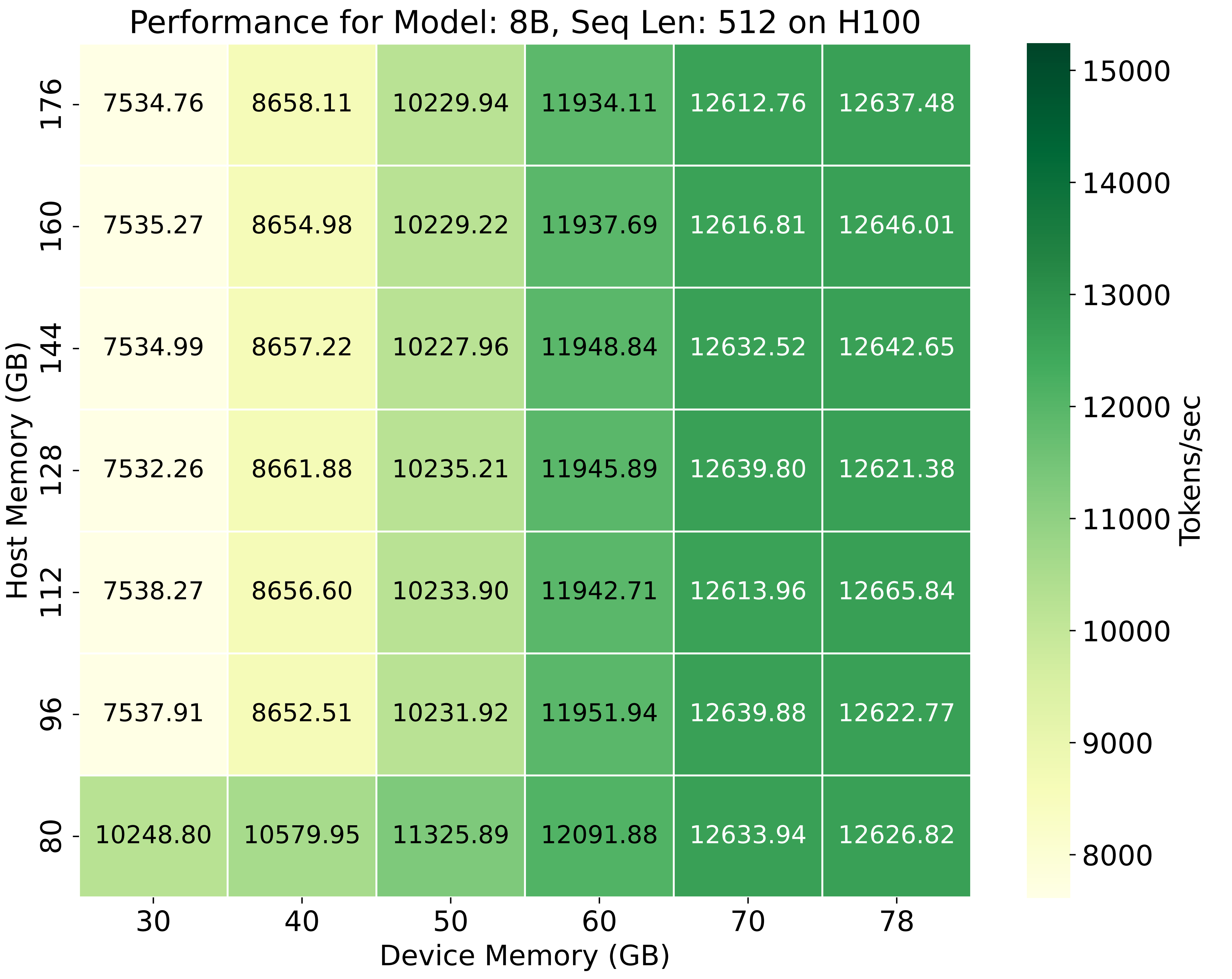
Performance for Model: 8B, Seq Len: 512 on H100
Host Memory (GB)
176
7534.76
8658.11
10229.94
11934.11
12612.76
12637.48
160
7535.27
8654.98
10229.22
11937.69
12616.81
12646.01
144
7534.99
8657.22
10227.96
11948.84
12632.52
12642.65
128
7532.26
8661.88
10235.21
11945.89
12639.80
12621.38
112
7538.27
8656.60
10233.90
11942.71
12613.96
12665.84
96
7537.91
8652.51
10231.92
11951.94
12639.88
12622.77
80
10248.80
10579.95
11325.89
12091.88
12633.94
12626.82
30
40
50
60
70
78
Device Memory (GB)
15000
14000
13000
12000
11000
10000
9000
8000
Tokens/sec