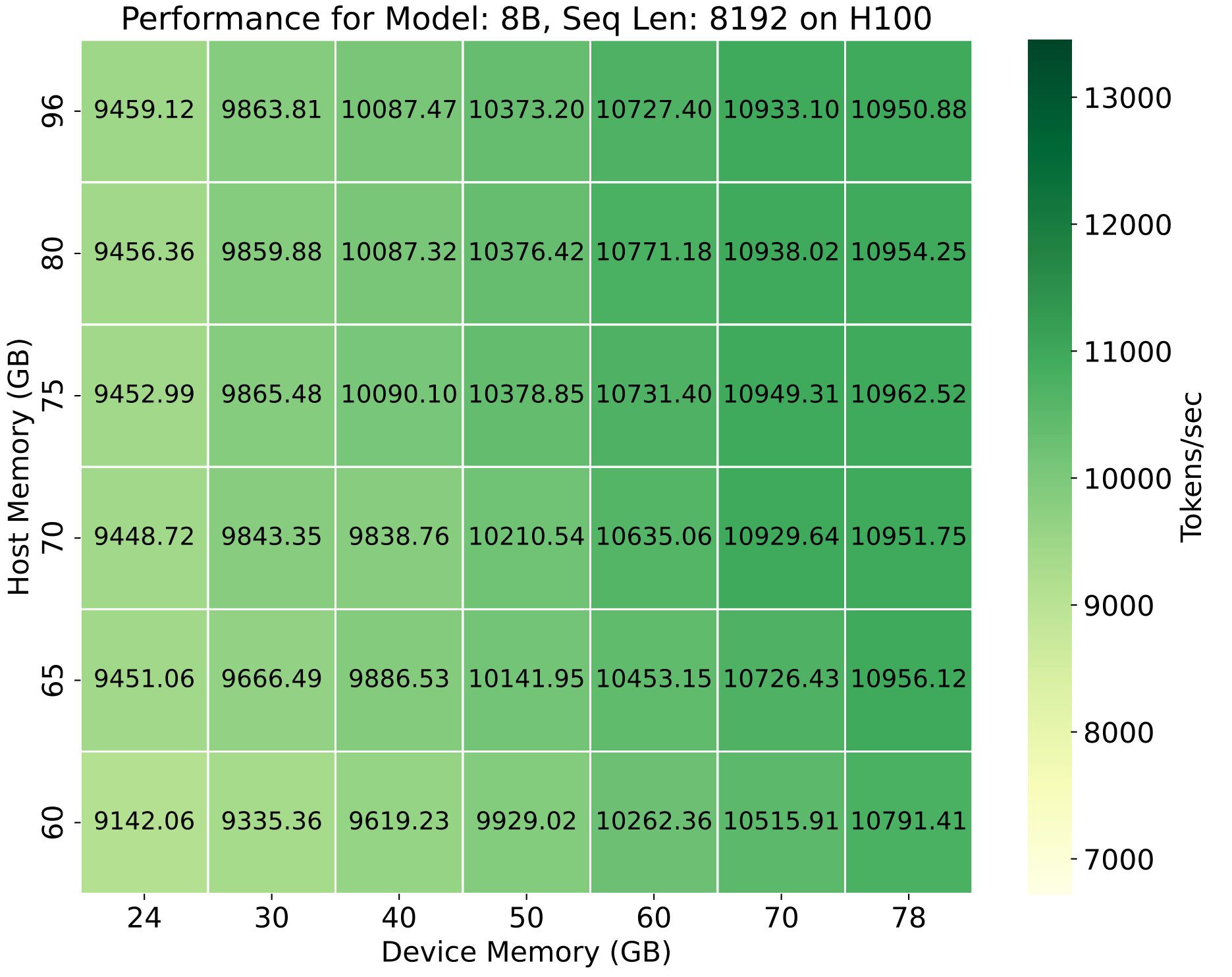
Performance for Model: 8B, Seq Len: 8192 on H100
Host Memory (GB)
96
80
75
70
65
60
9459.12 9863.81 10087.47 10373.20 10727.40 10933.10 10950.88
9456.36 9859.88 10087.32 10376.42 10771.18 10938.02 10954.25
9452.99 9865.48 10090.10 10378.85 10731.40 10949.31 10962.52
9448.72 9843.35 9838.76 10210.54 10635.06 10929.64 10951.75
9451.06 9666.49 9886.53 10141.95 10453.15 10726.43 10956.12
9142.06 9335.36 9619.23 9929.02 10262.36 10515.91 10791.41
24
30
40
50
60
70
78
Device Memory (GB)
Tokens/sec
13000
12000
11000
10000
9000
8000
7000