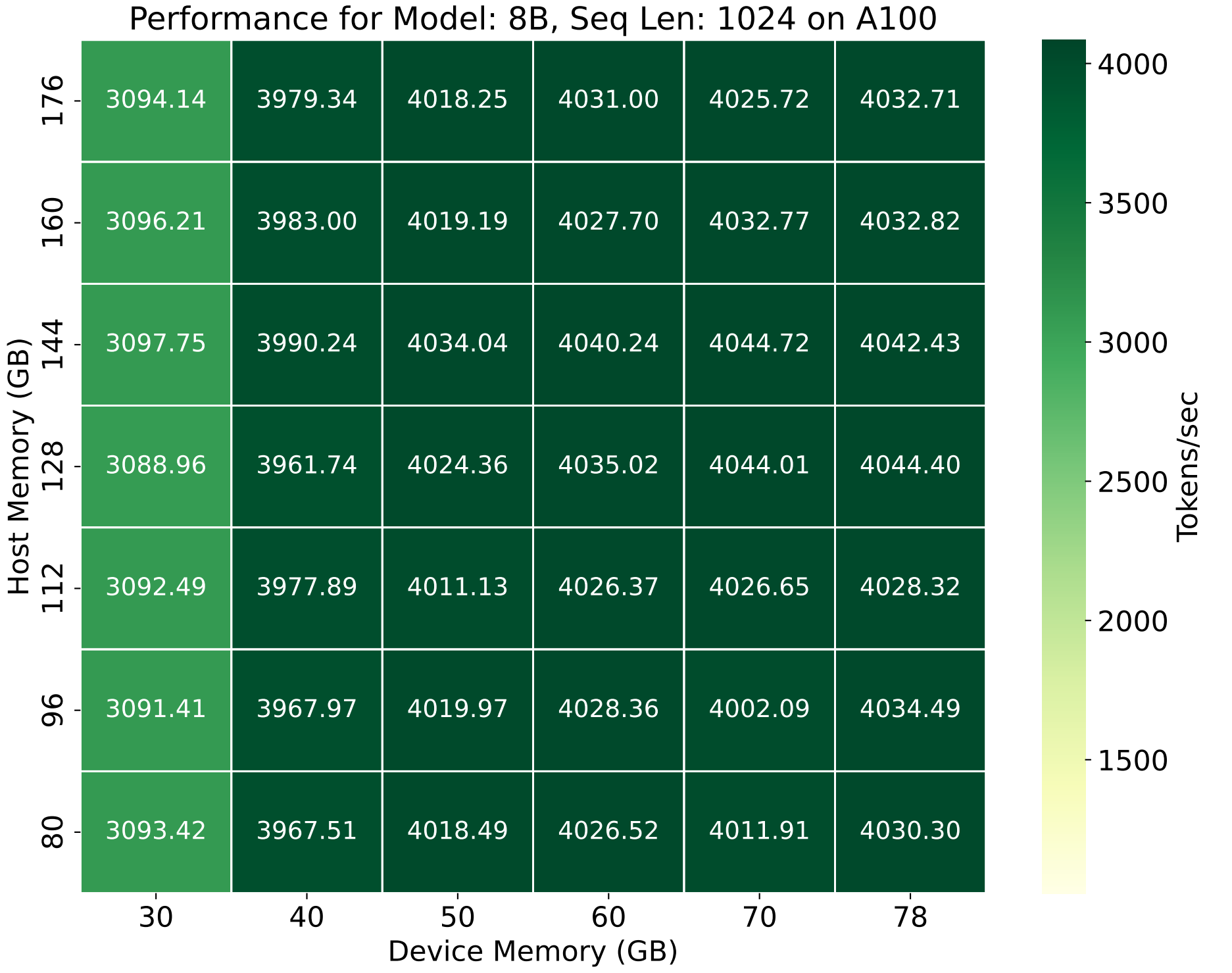
Performance for Model: 8B, Seq Len: 1024 on A100
Host Memory (GB)
176
160
144
128
112
96
80
3094.14
3979.34
4018.25
4031.00
4025.72
4032.71
3096.21
3983.00
4019.19
4027.70
4032.77
4032.82
3097.75
3990.24
4034.04
4040.24
4044.72
4042.43
3088.96
3961.74
4024.36
4035.02
4044.01
4044.40
3092.49
3977.89
4011.13
4026.37
4026.65
4028.32
3091.41
3967.97
4019.97
4028.36
4002.09
4034.49
3093.42
3967.51
4018.49
4026.52
4011.91
4030.30
30
40
50
60
70
78
Device Memory (GB)
Tokens/sec
4000
3500
3000
2500
2000
1500