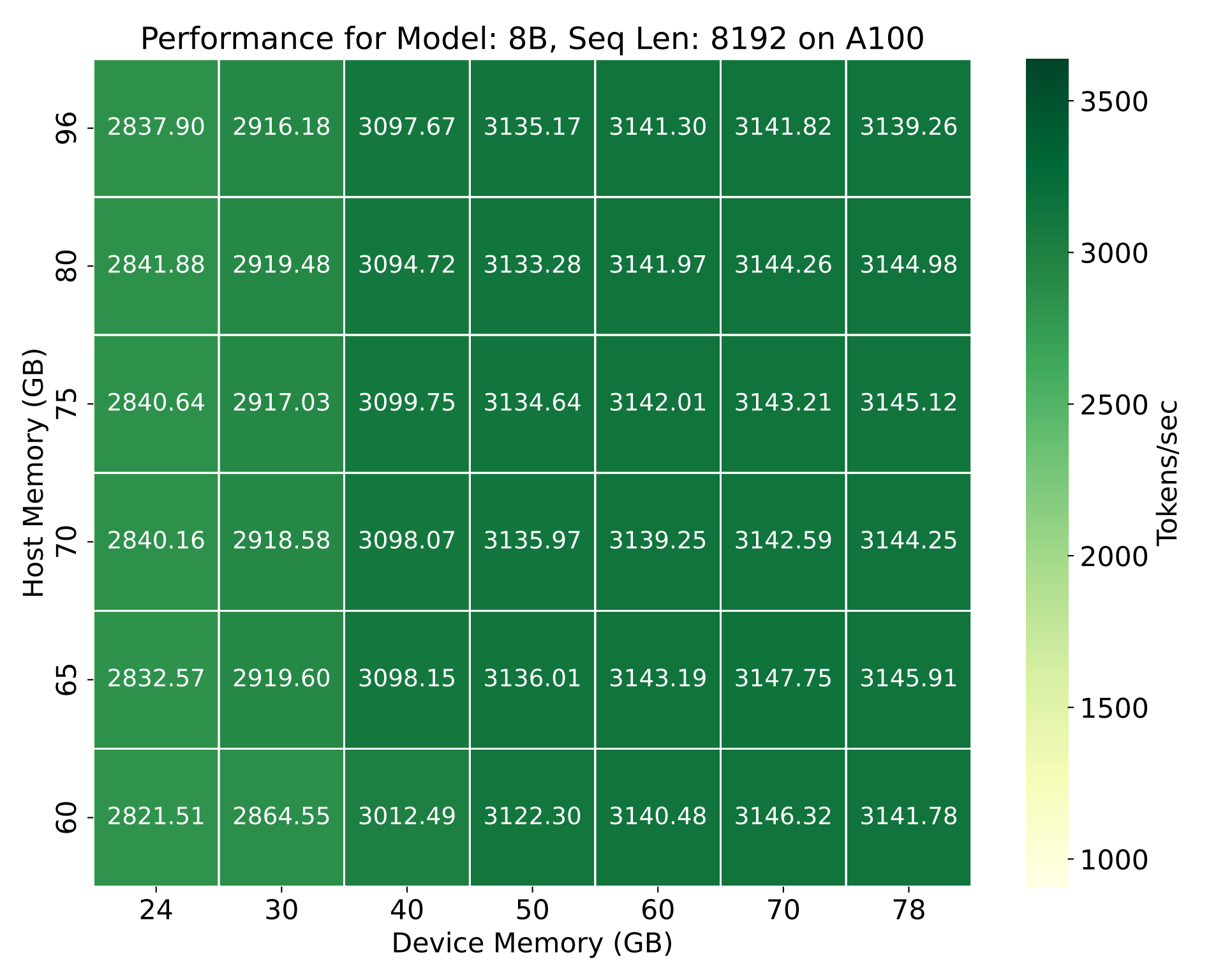
Performance for Model: 8B, Seq Len: 8192 on A100
Host Memory (GB)
96
80
75
70
65
60
2837.90 2916.18 3097.67 3135.17 3141.30 3141.82 3139.26
2841.88 2919.48 3094.72 3133.28 3141.97 3144.26 3144.98
2840.64 2917.03 3099.75 3134.64 3142.01 3143.21 3145.12
2840.16 2918.58 3098.07 3135.97 3139.25 3142.59 3144.25
2832.57 2919.60 3098.15 3136.01 3143.19 3147.75 3145.91
2821.51 2864.55 3012.49 3122.30 3140.48 3146.32 3141.78
24
30
40
50
60
70
78
Device Memory (GB)
Tokens/sec
3500
3000
2500
2000
1500
1000