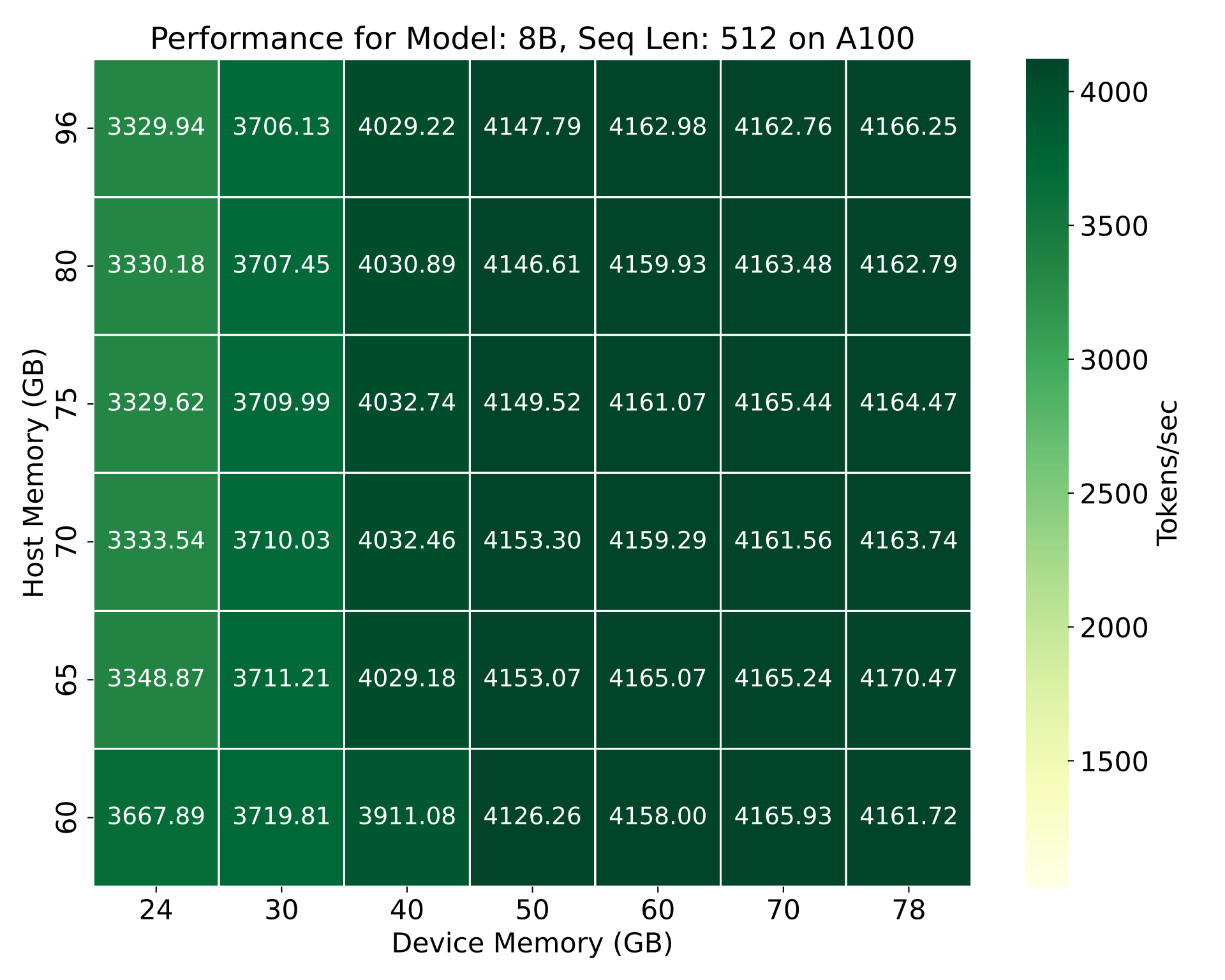
Performance for Model: 8B, Seq Len: 512 on A100
Host Memory (GB)
96
80
75
70
65
60
3329.94 3706.13 4029.22 4147.79 4162.98 4162.76 4166.25
3330.18 3707.45 4030.89 4146.61 4159.93 4163.48 4162.79
3329.62 3709.99 4032.74 4149.52 4161.07 4165.44 4164.47
3333.54 3710.03 4032.46 4153.30 4159.29 4161.56 4163.74
3348.87 3711.21 4029.18 4153.07 4165.07 4165.24 4170.47
3667.89 3719.81 3911.08 4126.26 4158.00 4165.93 4161.72
24
30
40
50
60
70
78
Device Memory (GB)
Tokens/sec
4000
3500
3000
2500
2000
1500