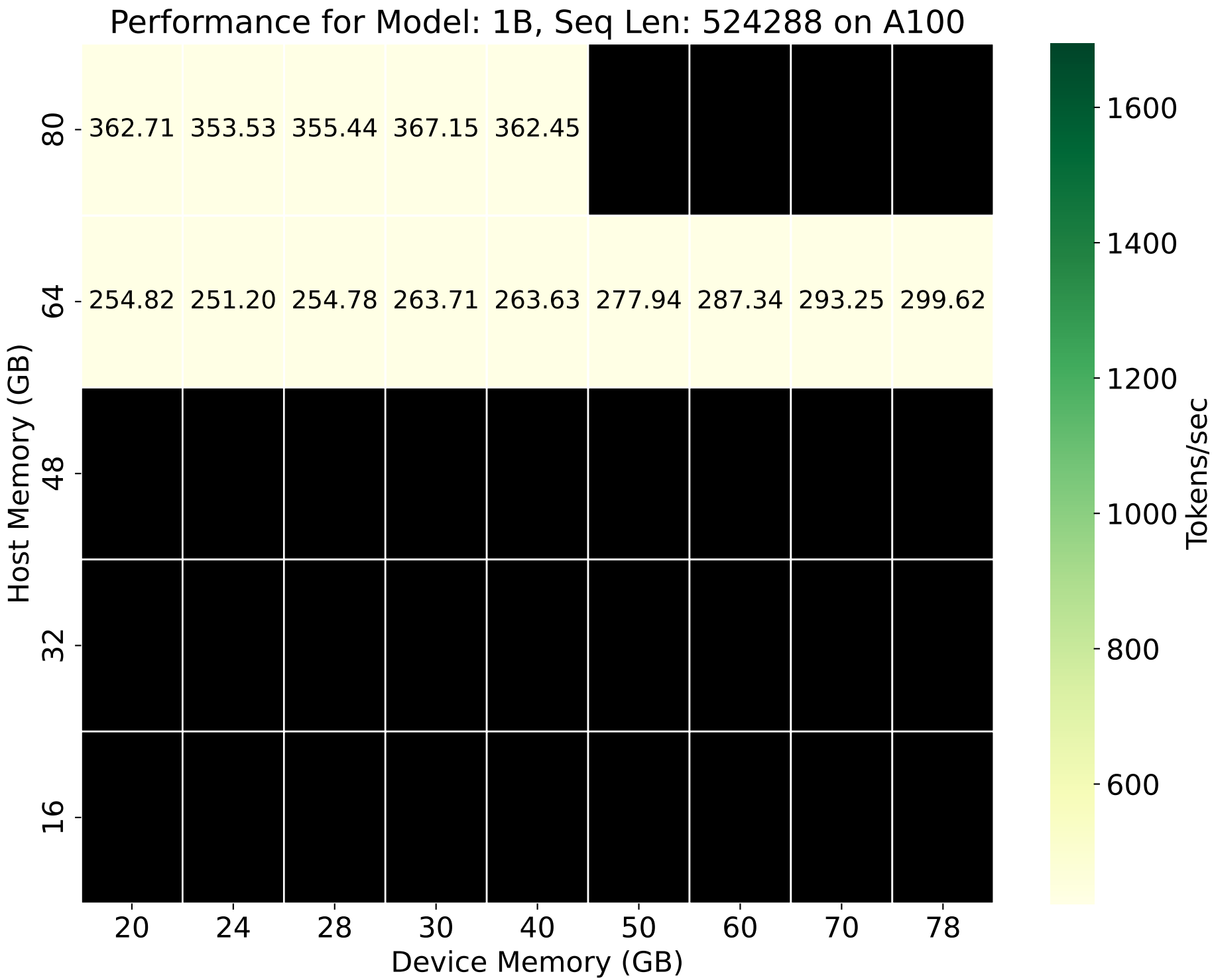
Performance for Model: 1B, Seq Len: 524288 on A100
Host Memory (GB)
80
64
48
32
16
362.71
353.53
355.44
367.15
362.45
254.82
251.20
254.78
263.71
263.63
277.94
287.34
293.25
299.62
20
24
28
30
40
50
60
70
78
Device Memory (GB)
Tokens/sec
1600
1400
1200
1000
800
600