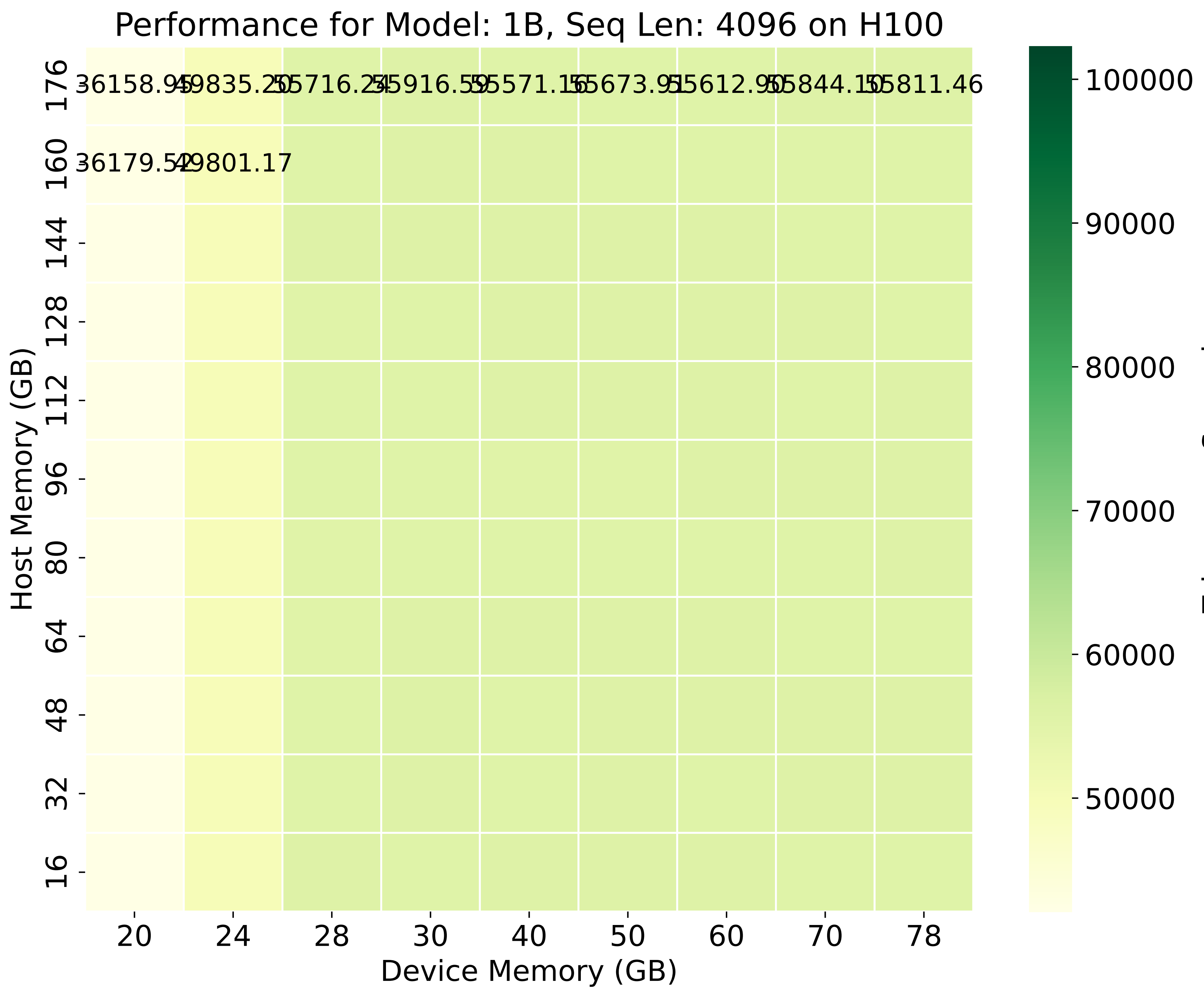
Performance for Model: 1B, Seq Len: 4096 on H100
36158.95
49835.20
55716.24
55916.59
55571.16
55673.91
55612.90
55844.10
55811.46
36179.52
49801.17
Host Memory (GB)
176
160
144
128
112
96
80
64
48
32
16
Device Memory (GB)
20
24
28
30
40
50
60
70
78
Tokens per Second
100000
90000
80000
70000
60000
50000