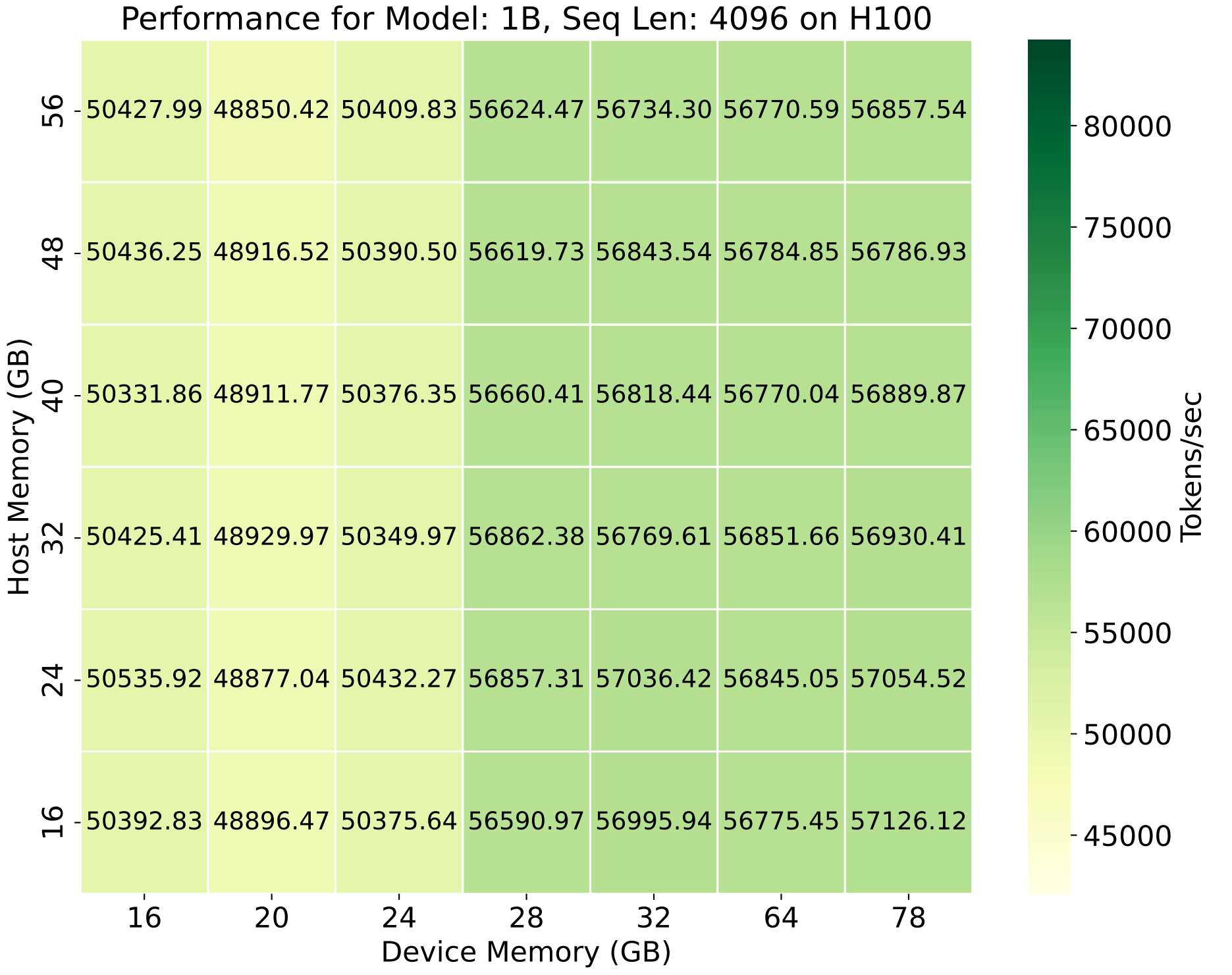
Performance for Model: 1B, Seq Len: 4096 on H100
Host Memory (GB)
56
48
40
32
24
16
50427.99 48850.42 50409.83 56624.47 56734.30 56770.59 56857.54
50436.25 48916.52 50390.50 56619.73 56843.54 56784.85 56786.93
50331.86 48911.77 50376.35 56660.41 56818.44 56770.04 56889.87
50425.41 48929.97 50349.97 56862.38 56769.61 56851.66 56930.41
50535.92 48877.04 50432.27 56857.31 57036.42 56845.05 57054.52
50392.83 48896.47 50375.64 56590.97 56995.94 56775.45 57126.12
16
20
24
28
32
64
78
Device Memory (GB)
Tokens/sec
80000
75000
70000
65000
60000
55000
50000
45000