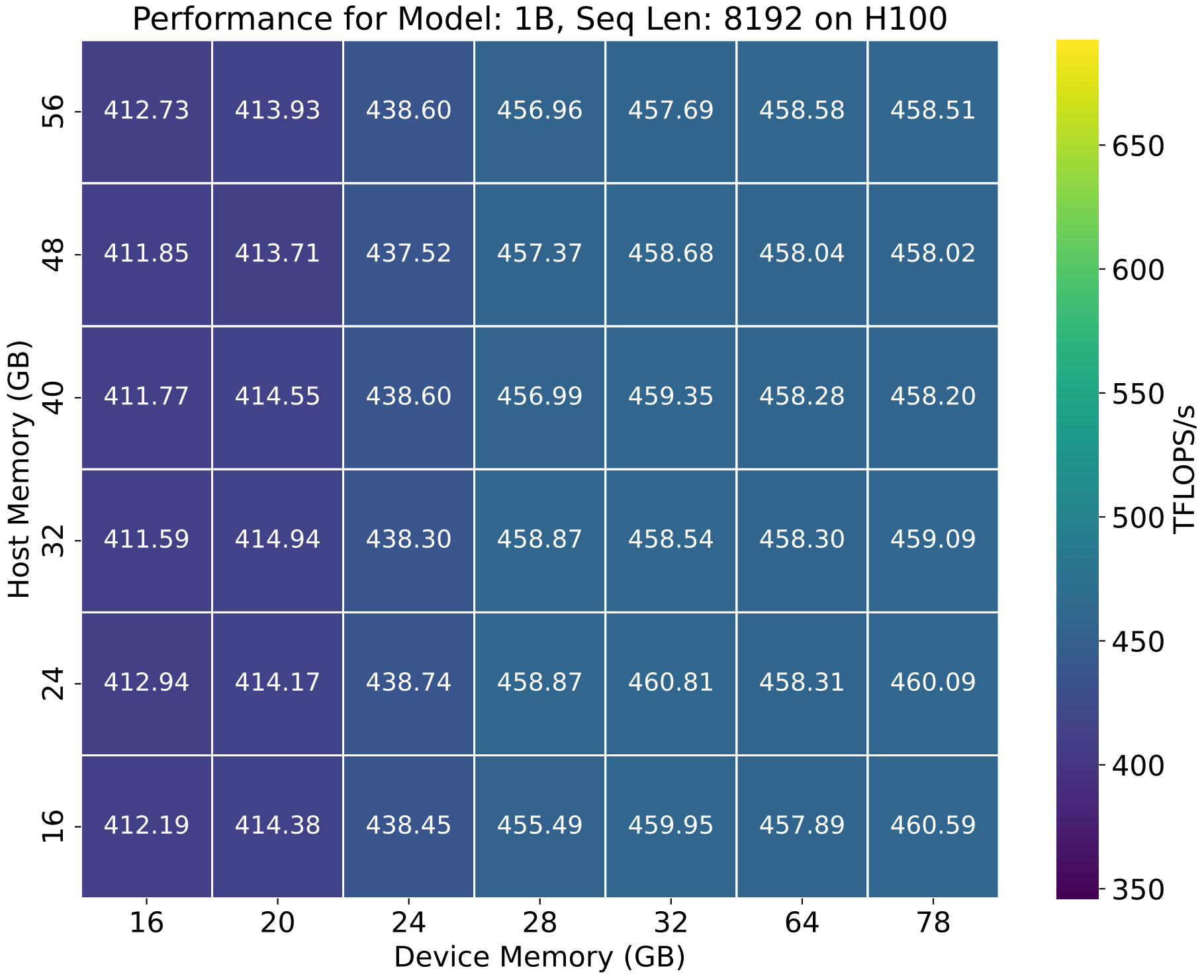
Performance for Model: 1B, Seq Len: 8192 on H100
Host Memory (GB)
56
48
40
32
24
16
412.73 413.93 438.60 456.96 457.69 458.58 458.51
411.85 413.71 437.52 457.37 458.68 458.04 458.02
411.77 414.55 438.60 456.99 459.35 458.28 458.20
411.59 414.94 438.30 458.87 458.54 458.30 459.09
412.94 414.17 438.74 458.87 460.81 458.31 460.09
412.19 414.38 438.45 455.49 459.95 457.89 460.59
16 20 24 28 32 64 78
Device Memory (GB)
TFLOPS/s
650
600
550
500
450
400
350