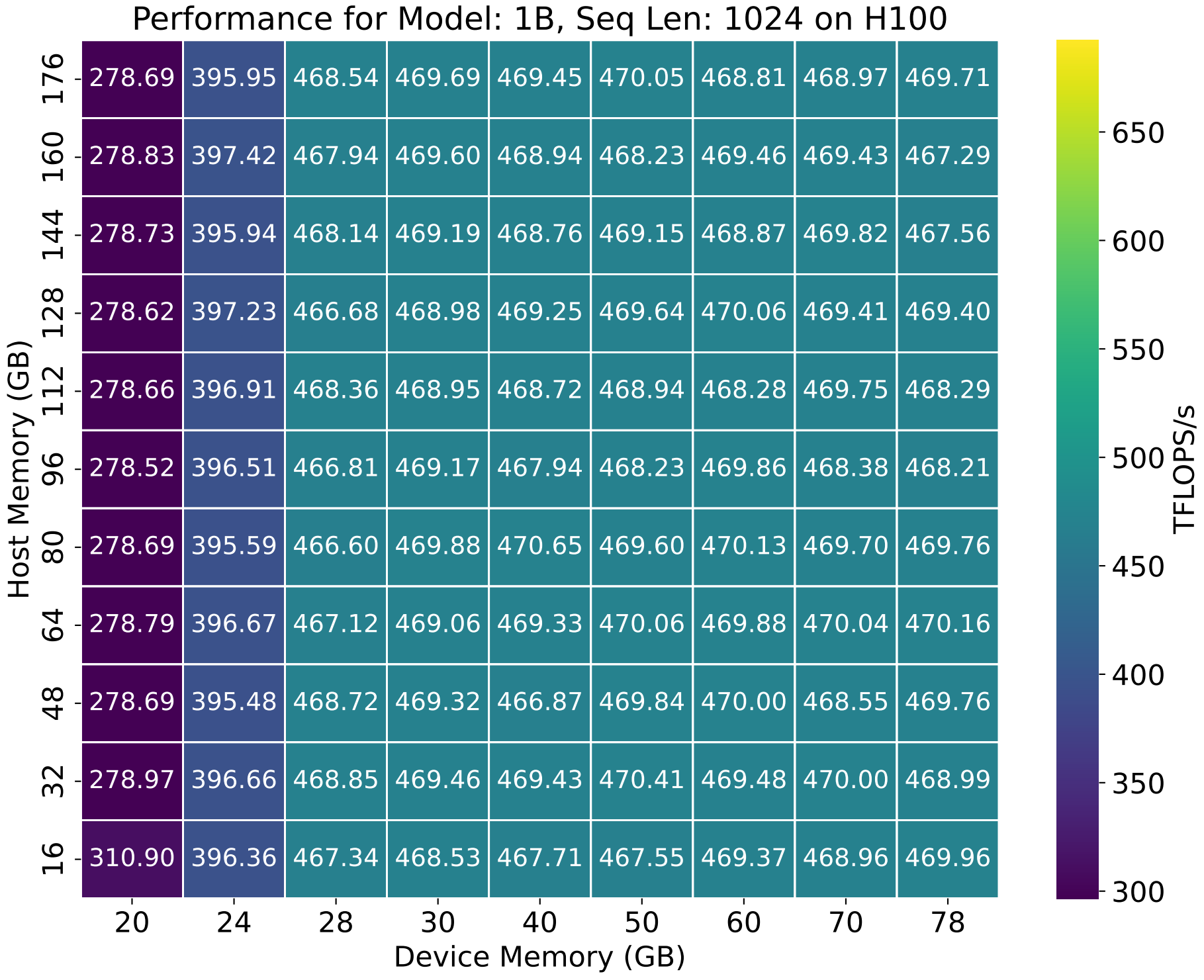
Performance for Model: 1B, Seq Len: 1024 on H100
Host Memory (GB)
176
160
144
128
112
96
80
64
48
32
16
278.69 395.95 468.54 469.69 469.45 470.05 468.81 468.97 469.71
278.83 397.42 467.94 469.60 468.94 468.23 469.46 469.43 467.29
278.73 395.94 468.14 469.19 468.76 469.15 468.87 469.82 467.56
278.62 397.23 466.68 468.98 469.25 469.64 470.06 469.41 469.40
278.66 396.91 468.36 468.95 468.72 468.94 468.28 469.75 468.29
278.52 396.51 466.81 469.17 467.94 468.23 469.86 468.38 468.21
278.69 395.59 466.60 469.88 470.65 469.60 470.13 469.70 469.76
278.79 396.67 467.12 469.06 469.33 470.06 469.88 470.04 470.16
278.69 395.48 468.72 469.32 466.87 469.84 470.00 468.55 469.76
278.97 396.66 468.85 469.46 469.43 470.41 469.48 470.00 468.99
310.90 396.36 467.34 468.53 467.71 467.55 469.37 468.96 469.96
20 24 28 30 40 50 60 70 78
Device Memory (GB)
TFLOPS/s
650
600
550
500
450
400
350
300