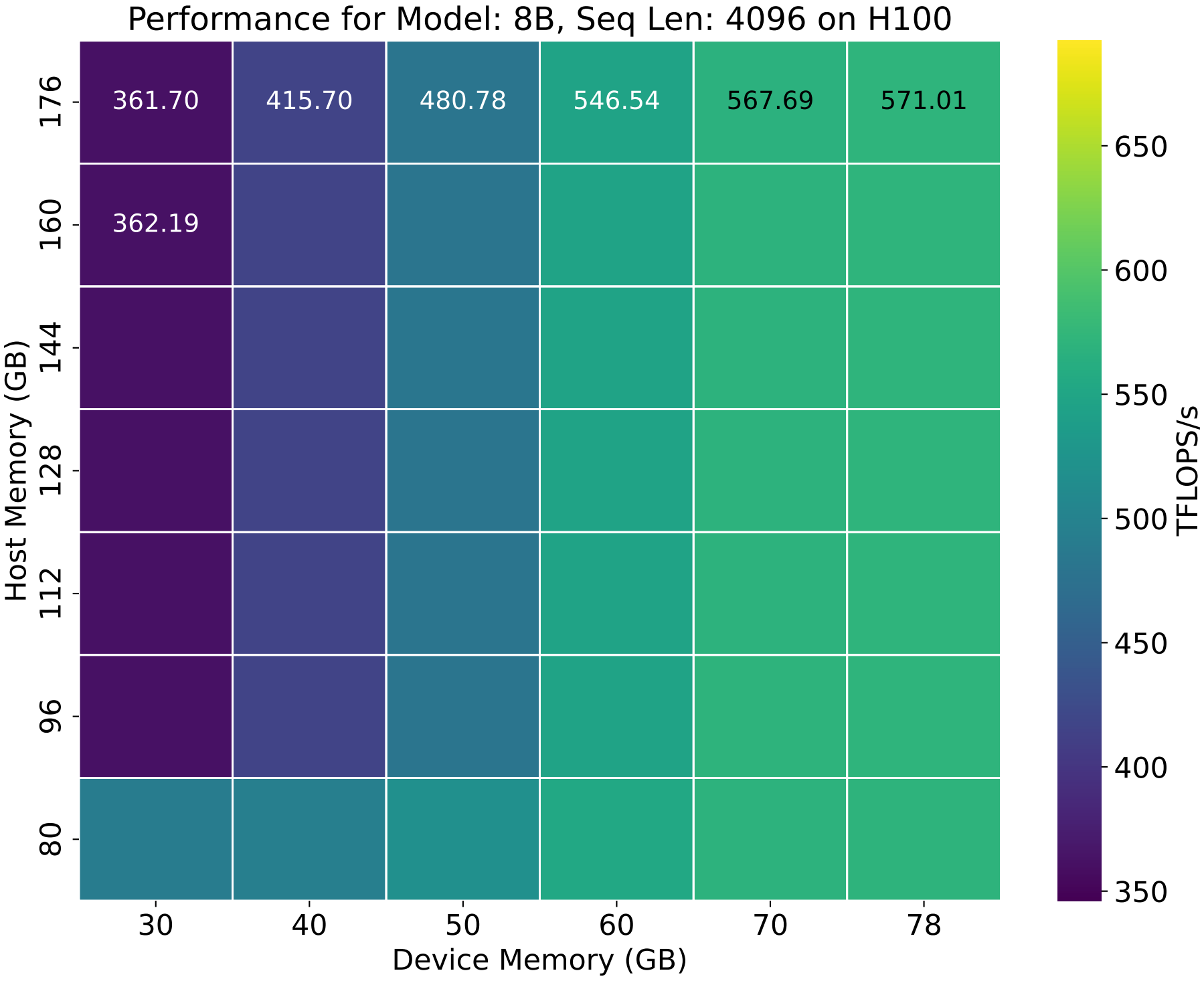
Performance for Model: 8B, Seq Len: 4096 on H100
361.70
415.70
480.78
546.54
567.69
571.01
362.19
176
160
144
128
112
96
80
Host Memory (GB)
30
40
50
60
70
78
Device Memory (GB)
650
600
550
500
450
400
350
TFLOPS/s