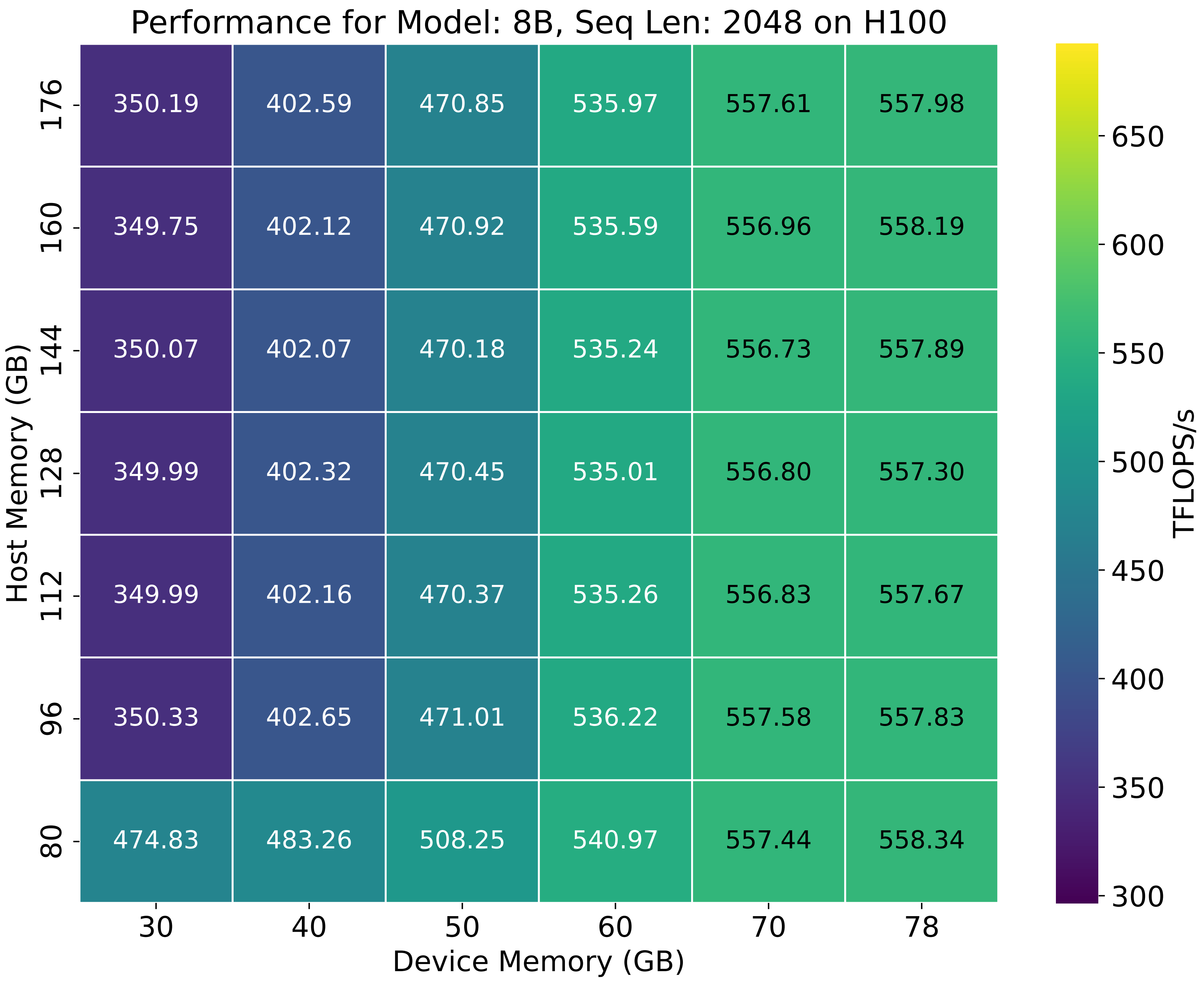
Performance for Model: 8B, Seq Len: 2048 on H100
Host Memory (GB)
176
160
144
128
112
96
80
Device Memory (GB)
30
40
50
60
70
78
350.19
402.59
470.85
535.97
557.61
557.98
349.75
402.12
470.92
535.59
556.96
558.19
350.07
402.07
470.18
535.24
556.73
557.89
349.99
402.32
470.45
535.01
556.80
557.30
349.99
402.16
470.37
535.26
556.83
557.67
350.33
402.65
471.01
536.22
557.58
557.83
474.83
483.26
508.25
540.97
557.44
558.34
TFLOPS/s
650
600
550
500
450
400
350
300