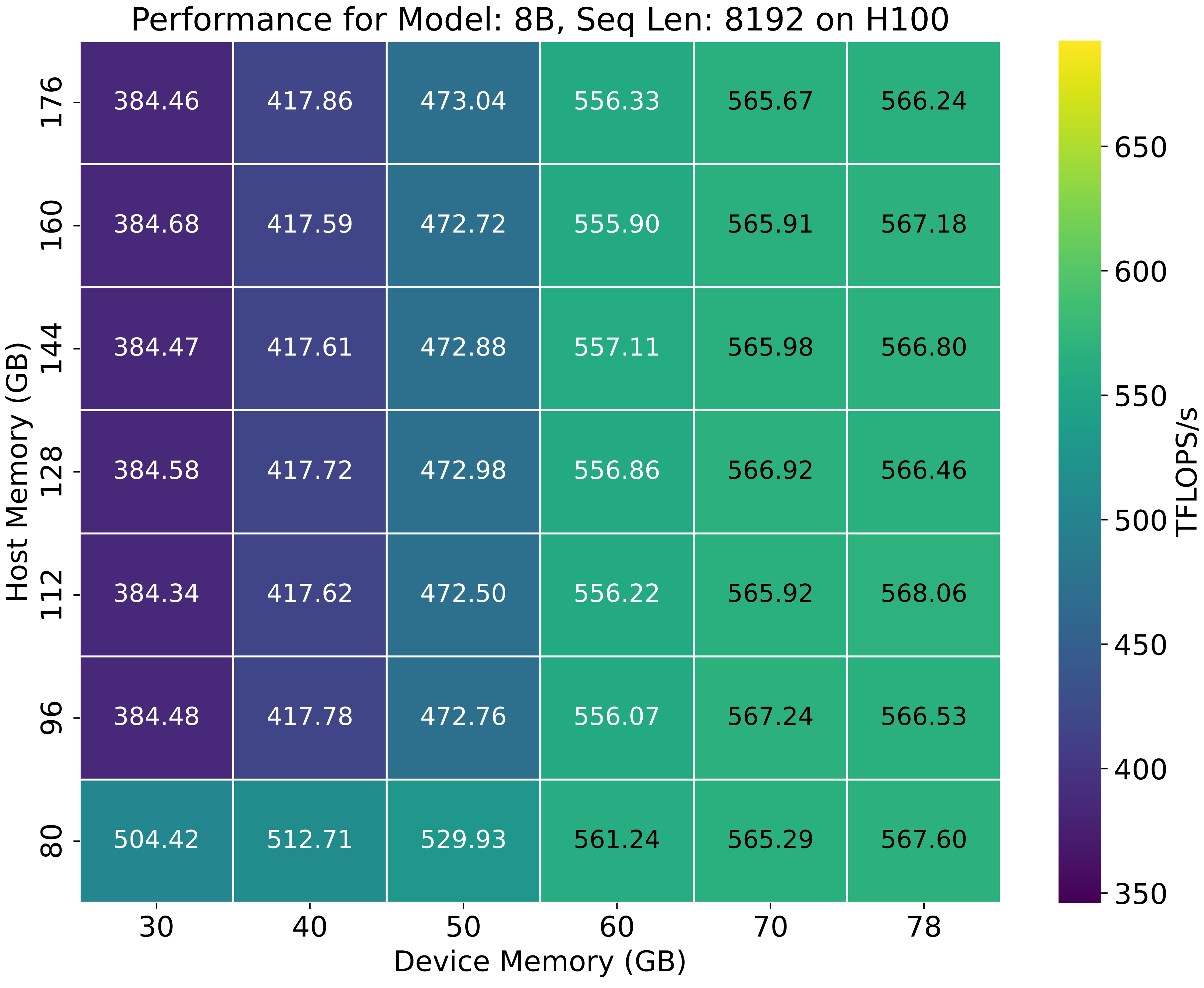Performance for Model: 8B, Seq Len: 8192 on H100
Host Memory (GB)
176
384.46
417.86
473.04
556.33
565.67
566.24
160
384.68
417.59
472.72
555.90
565.91
567.18
144
384.47
417.61
472.88
557.11
565.98
566.80
128
384.58
417.72
472.98
556.86
566.92
566.46
112
384.34
417.62
472.50
556.22
565.92
568.06
96
384.48
417.78
472.76
556.07
567.24
566.53
80
504.42
512.71
529.93
561.24
565.29
567.60
30
40
50
60
70
78
Device Memory (GB)
TFLOPS/s
650
600
550
500
450
400
350