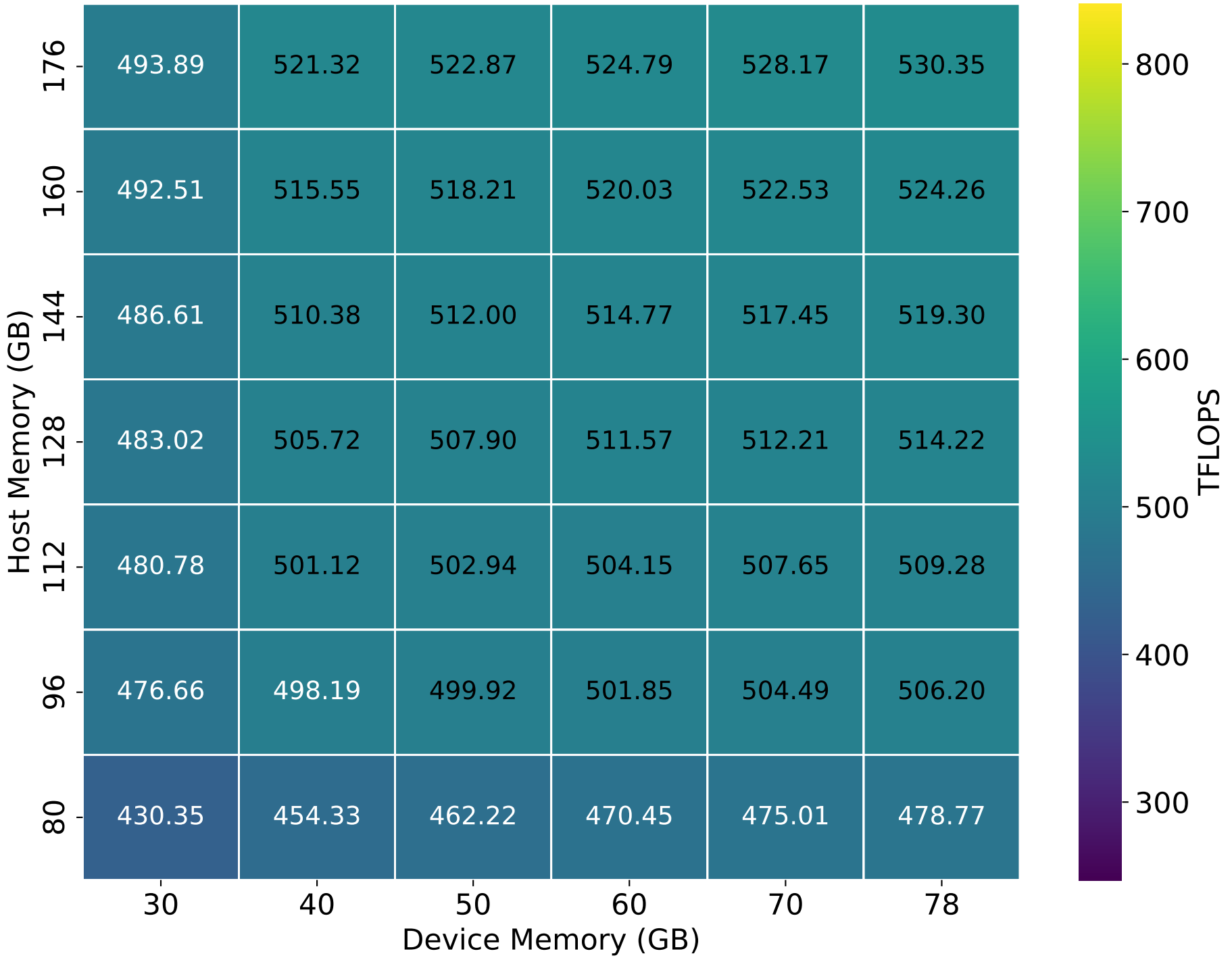
Performance for Model: 8B, Seq Len: 65536 on H100
Host Memory (GB)
176
160
144
128
112
96
80
493.89
521.32
522.87
524.79
528.17
530.35
492.51
515.55
518.21
520.03
522.53
524.26
486.61
510.38
512.00
514.77
517.45
519.30
483.02
505.72
507.90
511.57
512.21
514.22
480.78
501.12
502.94
504.15
507.65
509.28
476.66
498.19
499.92
501.85
504.49
506.20
430.35
454.33
462.22
470.45
475.01
478.77
30
40
50
60
70
78
Device Memory (GB)
TFLOPS
800
700
600
500
400
300