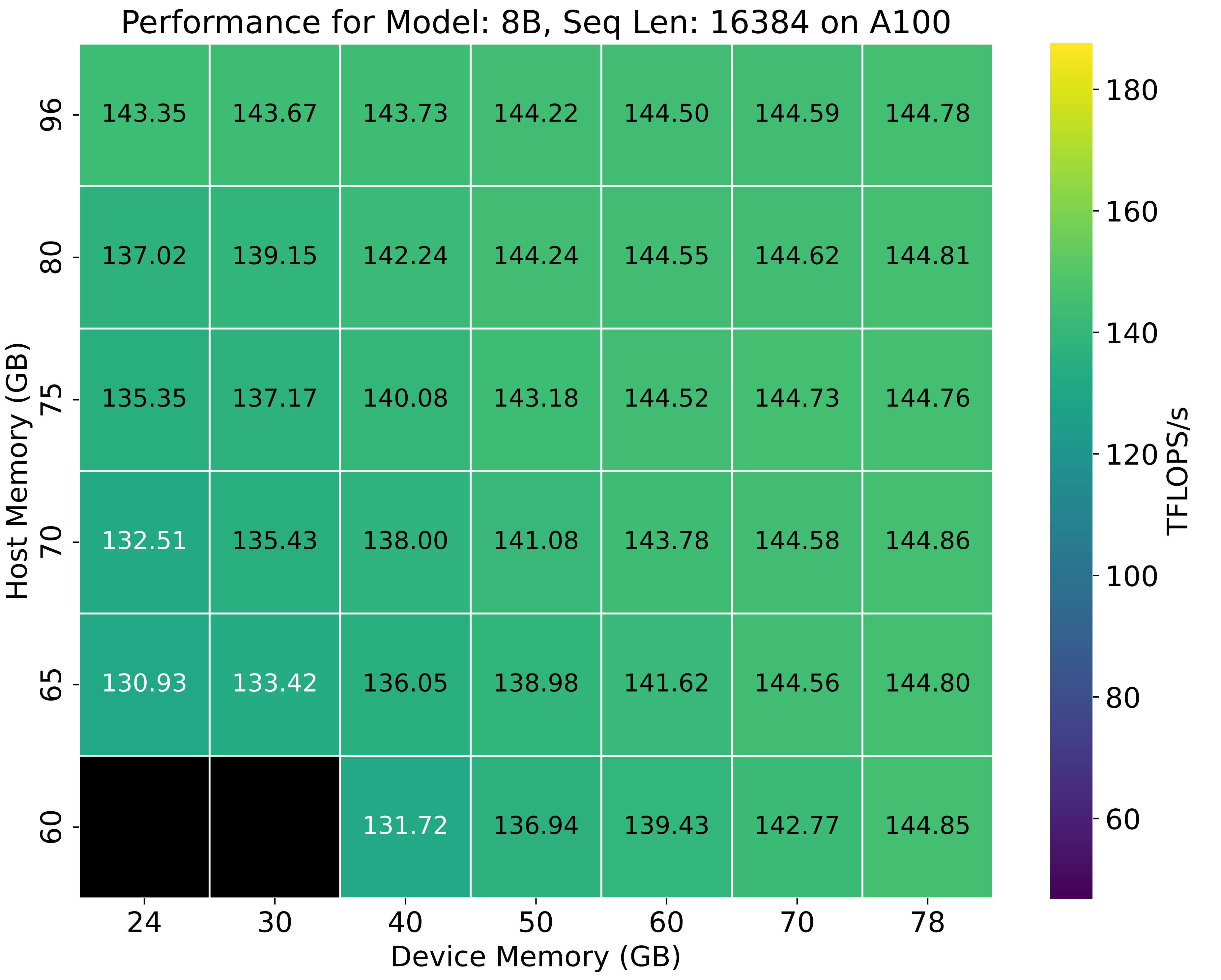
Performance for Model: 8B, Seq Len: 16384 on A100
Host Memory (GB)
Device Memory (GB)
TFLOPS/s
96
143.35 143.67 143.73 144.22 144.50 144.59 144.78
80
137.02 139.15 142.24 144.24 144.55 144.62 144.81
75
135.35 137.17 140.08 143.18 144.52 144.73 144.76
70
132.51 135.43 138.00 141.08 143.78 144.58 144.86
65
130.93 133.42 136.05 138.98 141.62 144.56 144.80
60
131.72 136.94 139.43 142.77 144.85
24
30
40
50
60
70
78
180
160
140
120
100
80
60