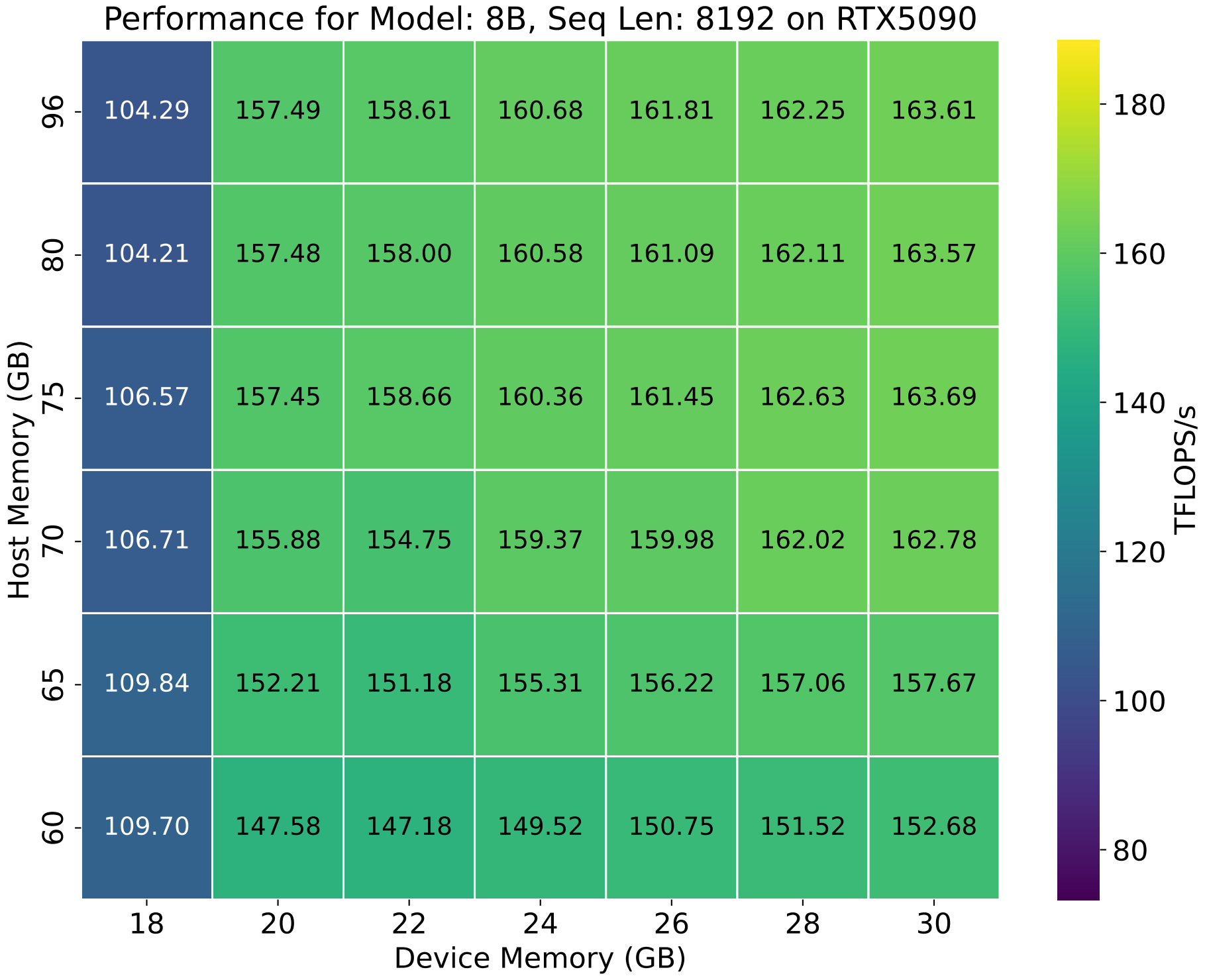
Performance for Model: 8B, Seq Len: 8192 on RTX5090
Host Memory (GB)
96
80
75
70
65
60
104.29 157.49 158.61 160.68 161.81 162.25 163.61
104.21 157.48 158.00 160.58 161.09 162.11 163.57
106.57 157.45 158.66 160.36 161.45 162.63 163.69
106.71 155.88 154.75 159.37 159.98 162.02 162.78
109.84 152.21 151.18 155.31 156.22 157.06 157.67
109.70 147.58 147.18 149.52 150.75 151.52 152.68
18
20
22
24
26
28
30
Device Memory (GB)
180
160
140
120
100
80
TFLOPS/s