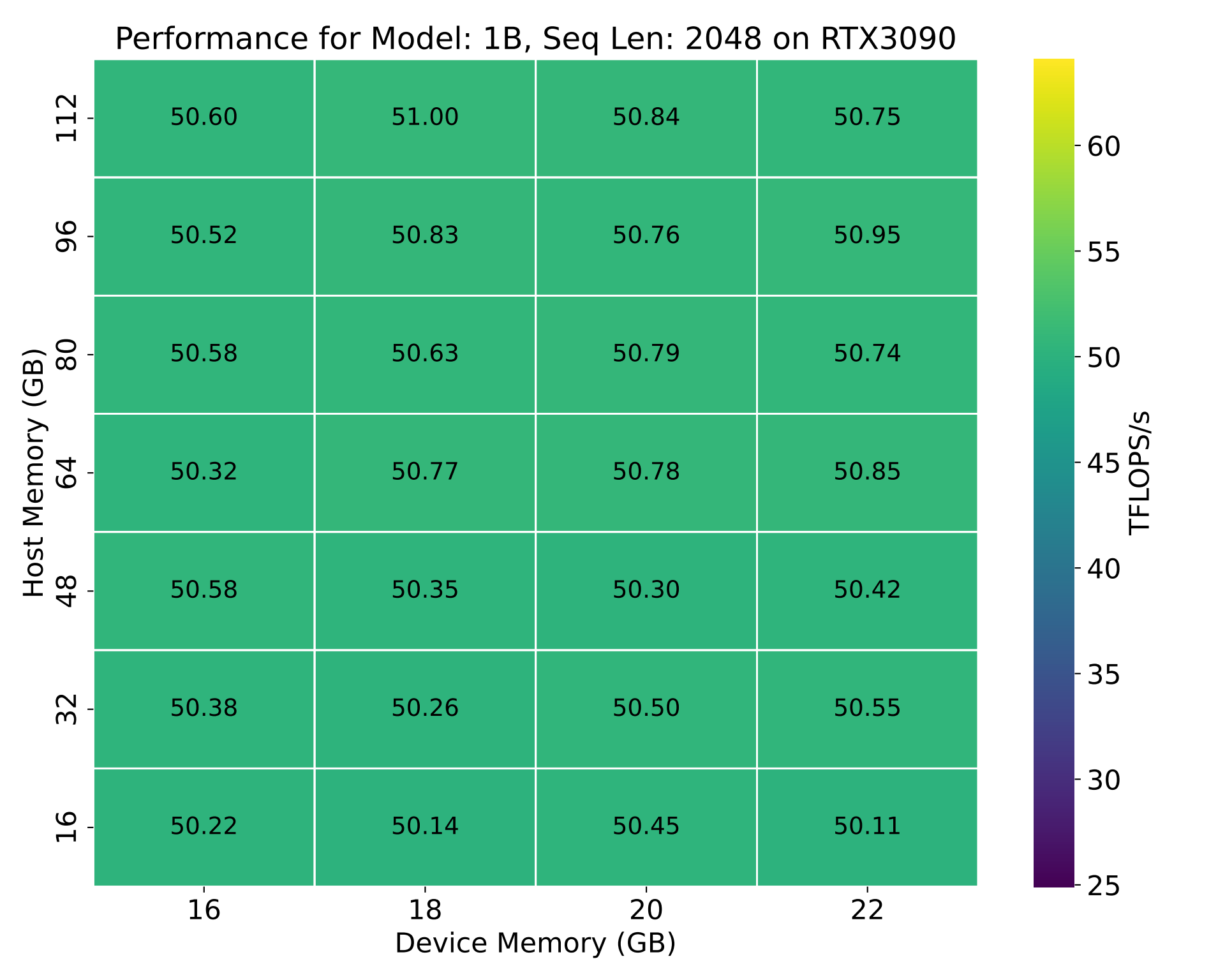
Performance for Model: 1B, Seq Len: 2048 on RTX3090
Host Memory (GB)
112
96
80
64
48
32
16
50.60
51.00
50.84
50.75
50.52
50.83
50.76
50.95
50.58
50.63
50.79
50.74
50.32
50.77
50.78
50.85
50.58
50.35
50.30
50.42
50.38
50.26
50.50
50.55
50.22
50.14
50.45
50.11
16
18
20
22
Device Memory (GB)
TFLOPS/s
60
55
50
45
40
35
30
25