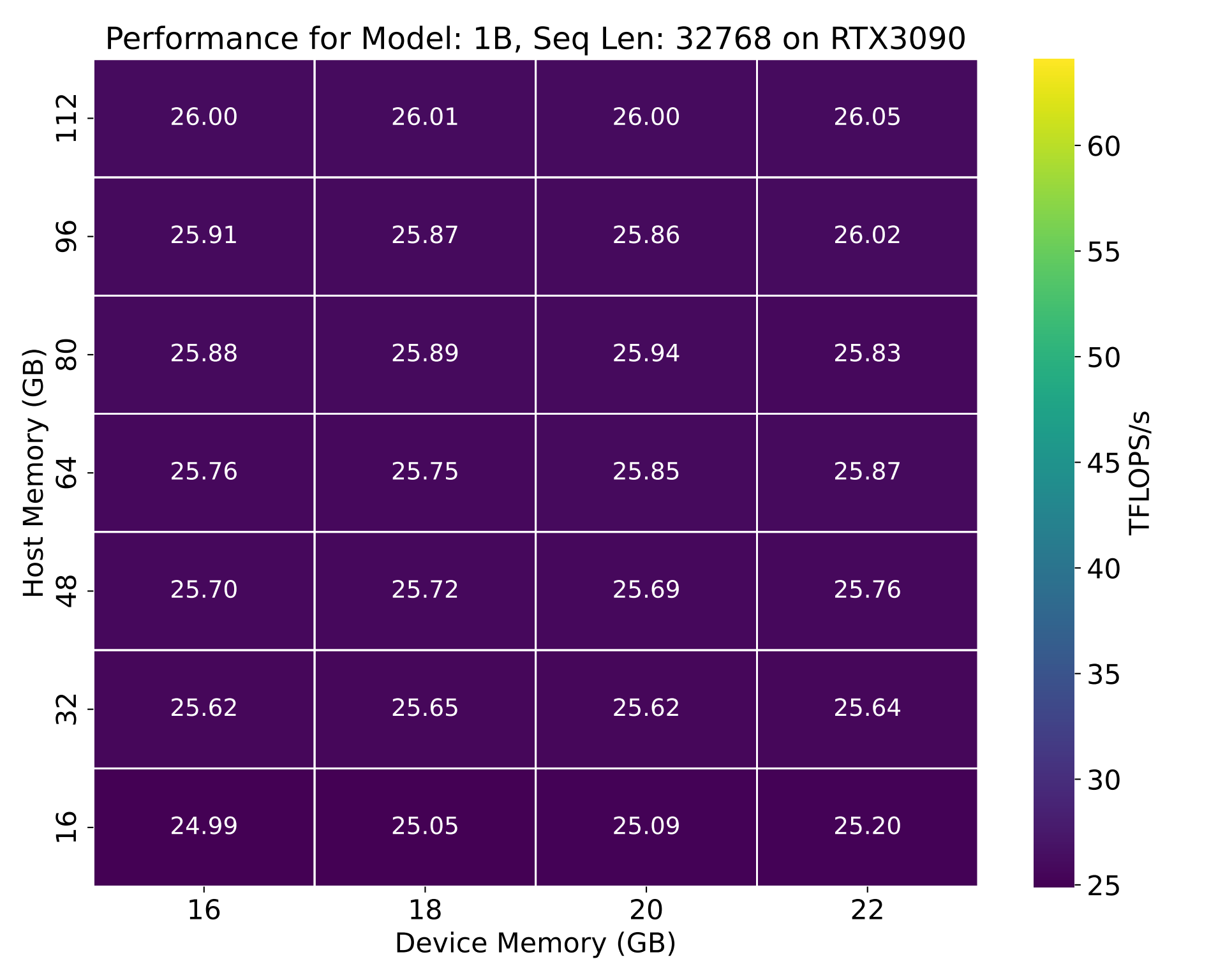
Performance for Model: 1B, Seq Len: 32768 on RTX3090
112
96
80
64
48
32
16
Host Memory (GB)
26.00
26.01
26.00
26.05
25.91
25.87
25.86
26.02
25.88
25.89
25.94
25.83
25.76
25.75
25.85
25.87
25.70
25.72
25.69
25.76
25.62
25.65
25.62
25.64
24.99
25.05
25.09
25.20
16
18
20
22
Device Memory (GB)
60
55
50
45
40
35
30
25
TFLOPS/s