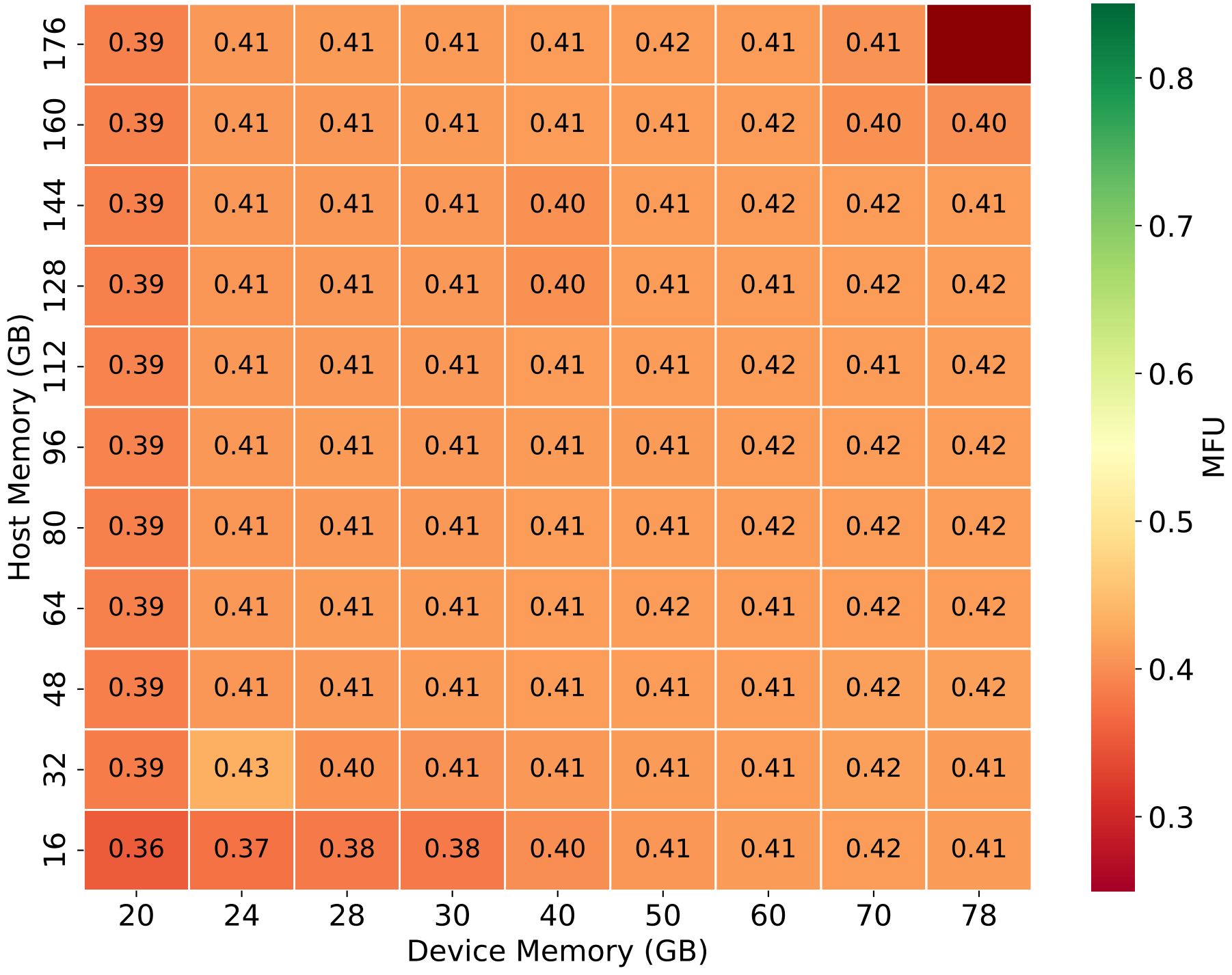
Performance for Model: 1B, Seq Len: 65536 on H100
Host Memory (GB)
176 0.39 0.41 0.41 0.41 0.41 0.42 0.41 0.41
160 0.39 0.41 0.41 0.41 0.41 0.41 0.42 0.40 0.40
144 0.39 0.41 0.41 0.41 0.40 0.41 0.42 0.42 0.41
128 0.39 0.41 0.41 0.41 0.40 0.41 0.41 0.42 0.42
112 0.39 0.41 0.41 0.41 0.41 0.41 0.42 0.41 0.42
96 0.39 0.41 0.41 0.41 0.41 0.41 0.42 0.42 0.42
80 0.39 0.41 0.41 0.41 0.41 0.41 0.42 0.42 0.42
64 0.39 0.41 0.41 0.41 0.41 0.42 0.41 0.42 0.42
48 0.39 0.41 0.41 0.41 0.41 0.41 0.41 0.42 0.42
32 0.39 0.43 0.40 0.41 0.41 0.41 0.41 0.42 0.41
16 0.36 0.37 0.38 0.38 0.40 0.41 0.41 0.42 0.41
20 24 28 30 40 50 60 70 78
Device Memory (GB)
MFU
0.8
0.7
0.6
0.5
0.4
0.3