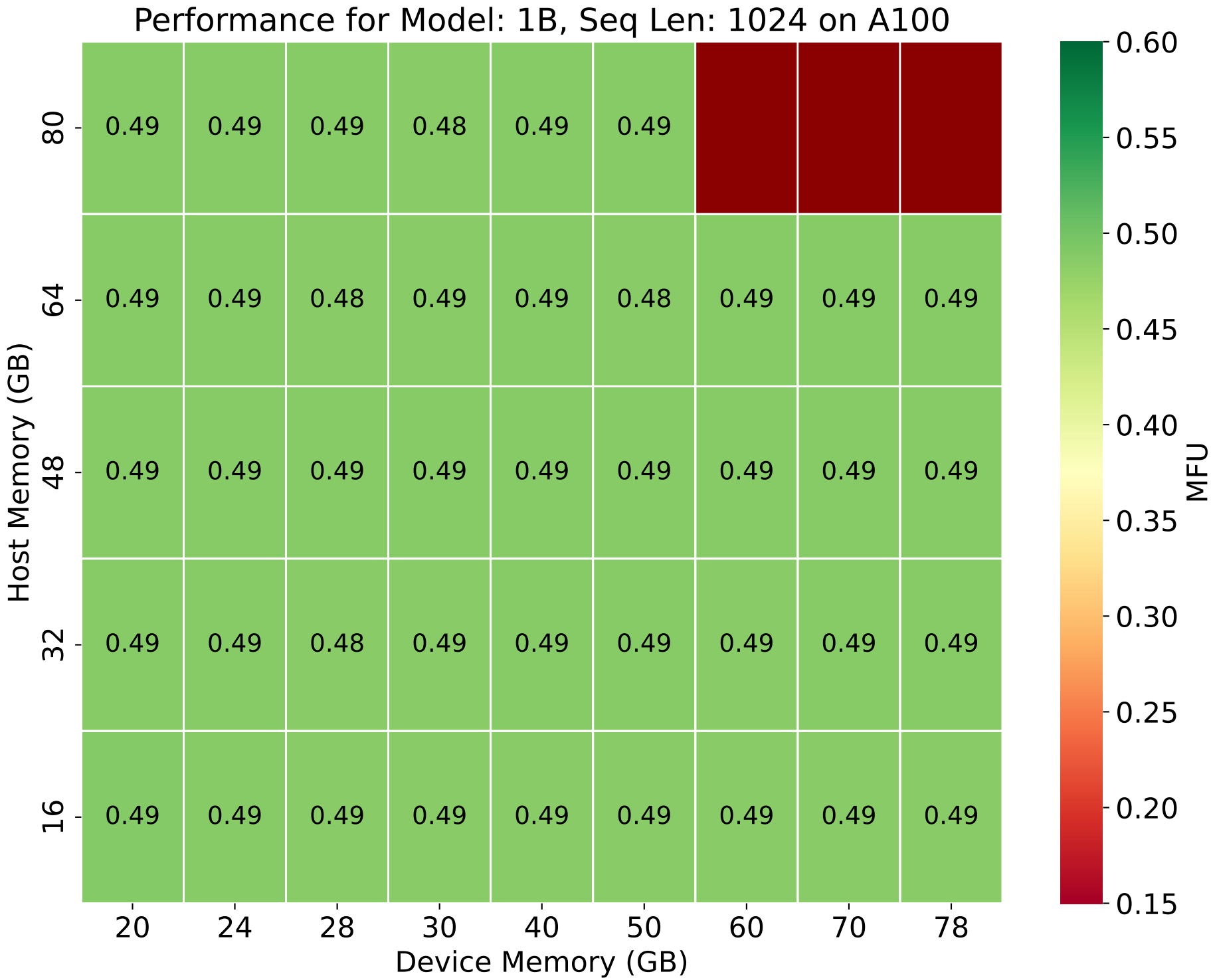
Performance for Model: 1B, Seq Len: 1024 on A100
Host Memory (GB)
80
64
48
32
16
0.49 0.49 0.49 0.48 0.49 0.49
0.49 0.49 0.48 0.49 0.49 0.48 0.49 0.49 0.49
0.49 0.49 0.49 0.49 0.49 0.49 0.49 0.49 0.49
0.49 0.49 0.48 0.49 0.49 0.49 0.49 0.49 0.49
0.49 0.49 0.49 0.49 0.49 0.49 0.49 0.49 0.49
20 24 28 30 40 50 60 70 78
Device Memory (GB)
MFU
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15