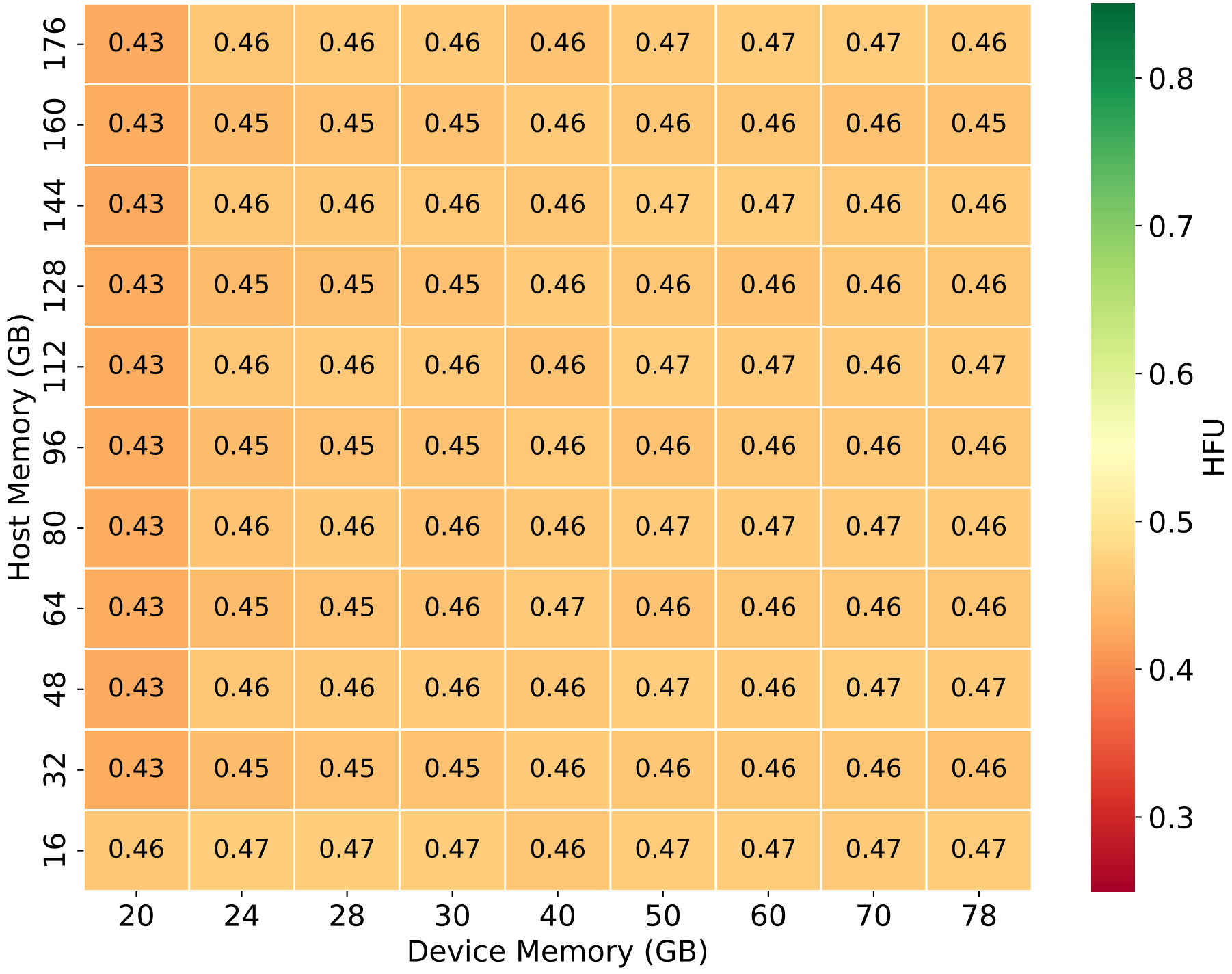
Performance for Model: 1B, Seq Len: 32768 on H100
Host Memory (GB)
176 0.43 0.46 0.46 0.46 0.46 0.47 0.47 0.47 0.46
160 0.43 0.45 0.45 0.45 0.46 0.46 0.46 0.46 0.45
144 0.43 0.46 0.46 0.46 0.46 0.47 0.47 0.46 0.46
128 0.43 0.45 0.45 0.45 0.46 0.46 0.46 0.46 0.46
112 0.43 0.46 0.46 0.46 0.46 0.47 0.47 0.46 0.47
96 0.43 0.45 0.45 0.45 0.46 0.46 0.46 0.46 0.46
80 0.43 0.46 0.46 0.46 0.46 0.47 0.47 0.47 0.46
64 0.43 0.45 0.45 0.46 0.47 0.46 0.46 0.46 0.46
48 0.43 0.46 0.46 0.46 0.46 0.47 0.46 0.47 0.47
32 0.43 0.45 0.45 0.45 0.46 0.46 0.46 0.46 0.46
16 0.46 0.47 0.47 0.47 0.46 0.47 0.47 0.47 0.47
20 24 28 30 40 50 60 70 78
Device Memory (GB)
HFU
0.8
0.7
0.6
0.5
0.4
0.3