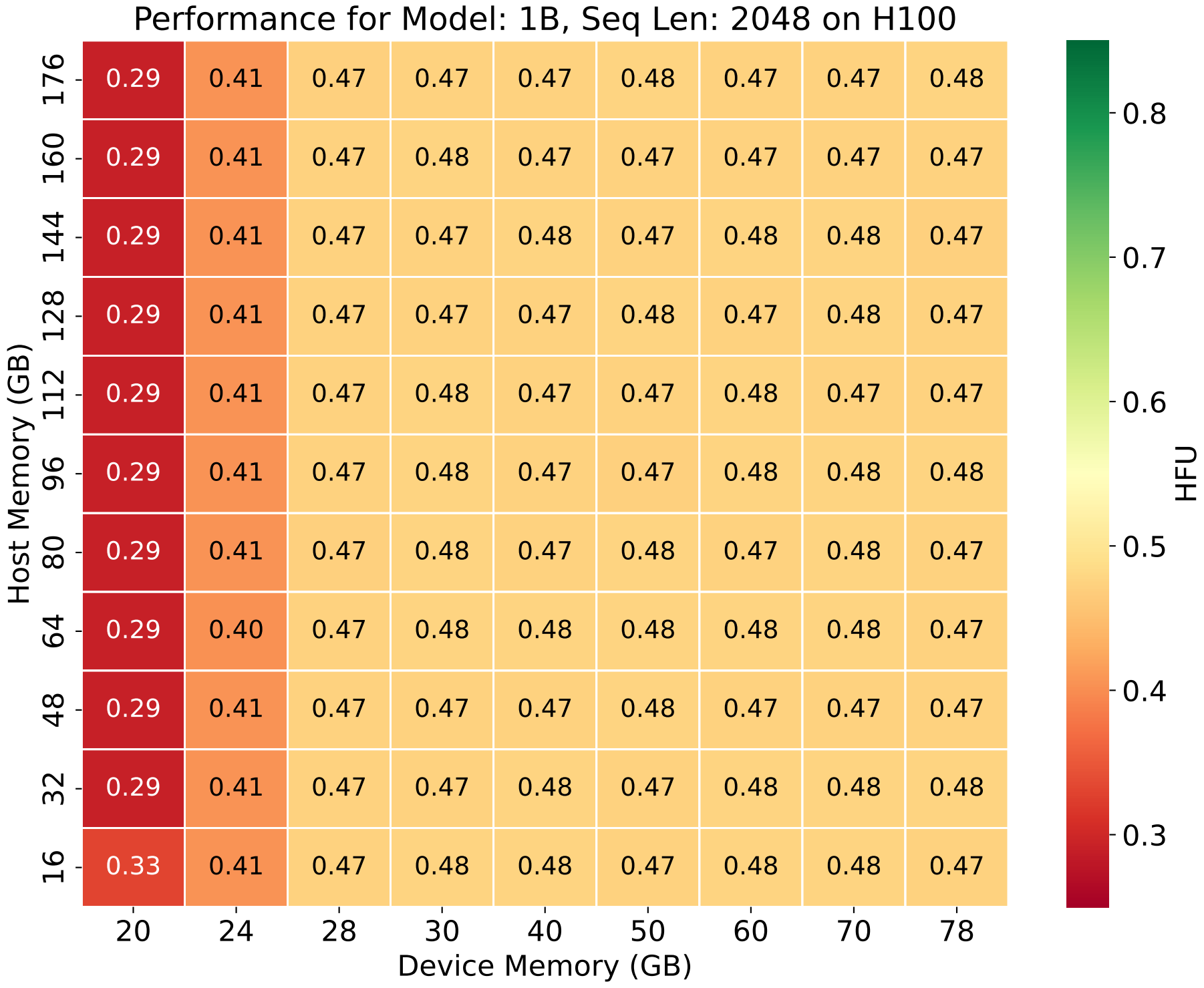
Performance for Model: 1B, Seq Len: 2048 on H100
Host Memory (GB)
176 0.29 0.41 0.47 0.47 0.47 0.48 0.47 0.47 0.48
160 0.29 0.41 0.47 0.48 0.47 0.47 0.47 0.47 0.47
144 0.29 0.41 0.47 0.47 0.48 0.47 0.48 0.48 0.47
128 0.29 0.41 0.47 0.47 0.47 0.48 0.47 0.48 0.47
112 0.29 0.41 0.47 0.48 0.47 0.47 0.48 0.47 0.47
96 0.29 0.41 0.47 0.48 0.47 0.47 0.48 0.48 0.48
80 0.29 0.41 0.47 0.48 0.47 0.48 0.47 0.48 0.47
64 0.29 0.40 0.47 0.48 0.48 0.48 0.48 0.48 0.47
48 0.29 0.41 0.47 0.47 0.47 0.48 0.47 0.47 0.47
32 0.29 0.41 0.47 0.47 0.48 0.47 0.48 0.48 0.48
16 0.33 0.41 0.47 0.48 0.48 0.47 0.48 0.48 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