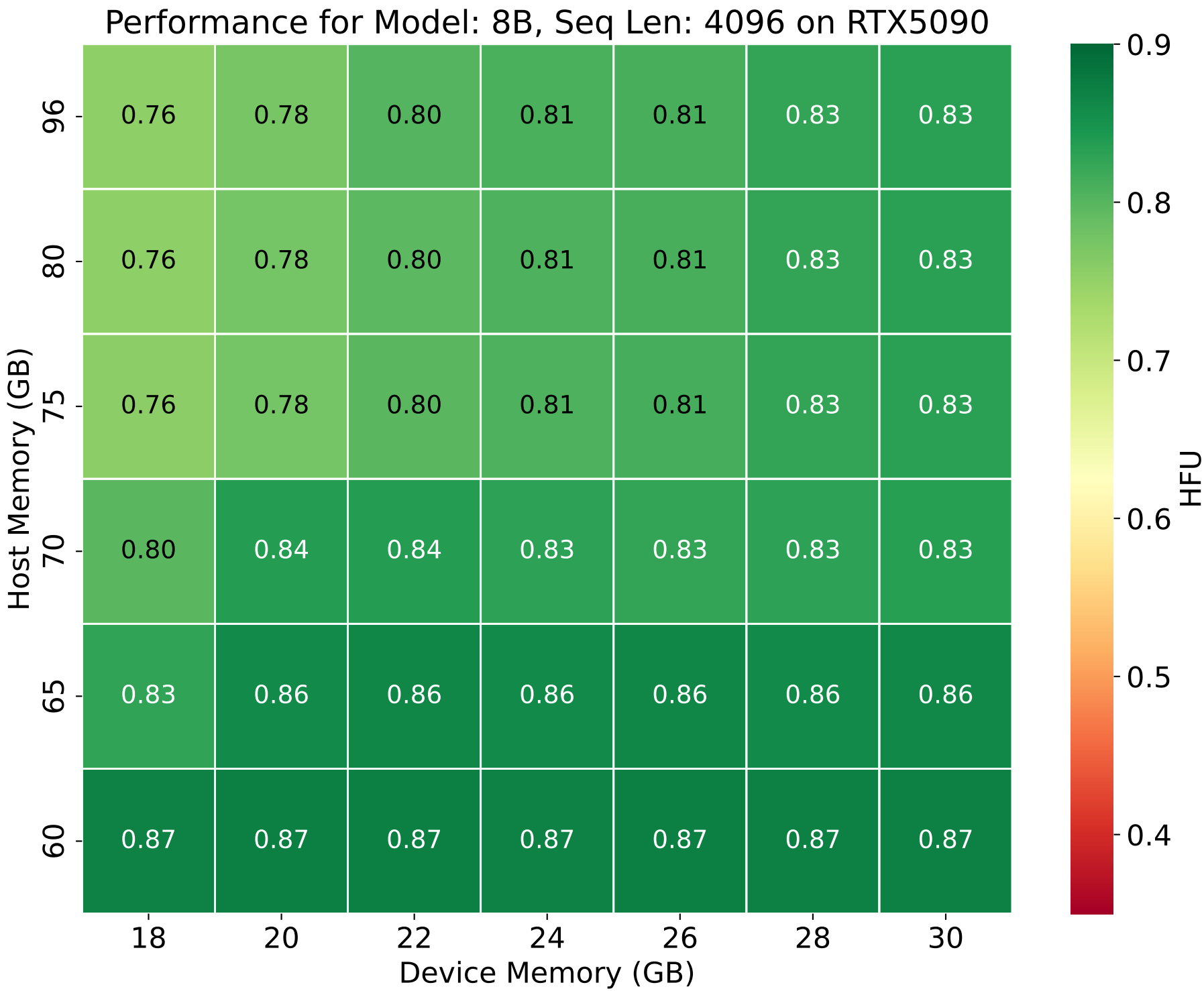
Performance for Model: 8B, Seq Len: 4096 on RTX5090
Host Memory (GB)
96
80
75
70
65
60
0.76 0.78 0.80 0.81 0.81 0.83 0.83
0.76 0.78 0.80 0.81 0.81 0.83 0.83
0.76 0.78 0.80 0.81 0.81 0.83 0.83
0.80 0.84 0.84 0.83 0.83 0.83 0.83
0.83 0.86 0.86 0.86 0.86 0.86 0.86
0.87 0.87 0.87 0.87 0.87 0.87 0.87
18 20 22 24 26 28 30
Device Memory (GB)
HFU
0.9
0.8
0.7
0.6
0.5
0.4