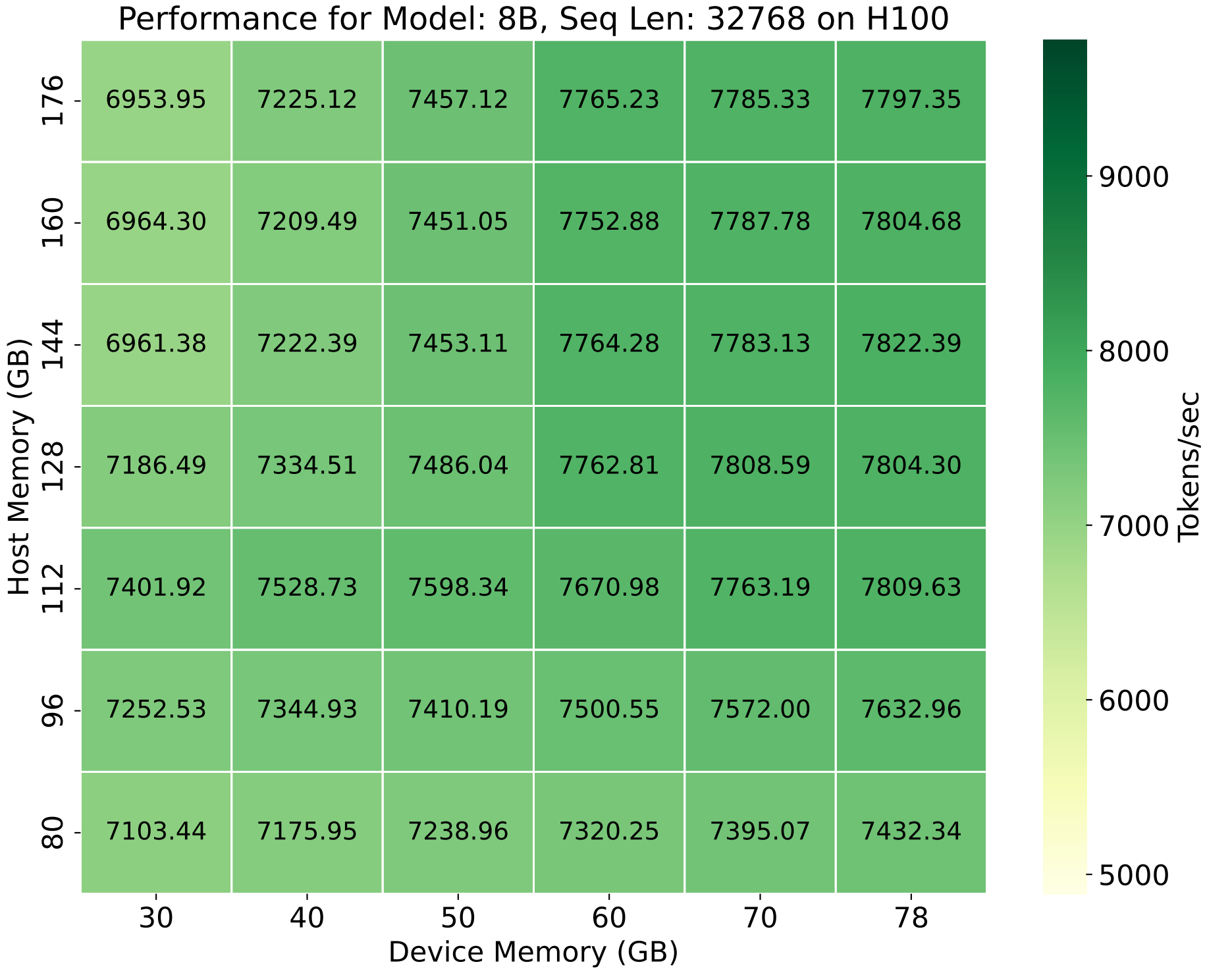
Performance for Model: 8B, Seq Len: 32768 on H100
Host Memory (GB)
176
160
144
128
112
96
80
6953.95
7225.12
7457.12
7765.23
7785.33
7797.35
6964.30
7209.49
7451.05
7752.88
7787.78
7804.68
6961.38
7222.39
7453.11
7764.28
7783.13
7822.39
7186.49
7334.51
7486.04
7762.81
7808.59
7804.30
7401.92
7528.73
7598.34
7670.98
7763.19
7809.63
7252.53
7344.93
7410.19
7500.55
7572.00
7632.96
7103.44
7175.95
7238.96
7320.25
7395.07
7432.34
30
40
50
60
70
78
Device Memory (GB)
9000
8000
7000
6000
5000
Tokens/sec