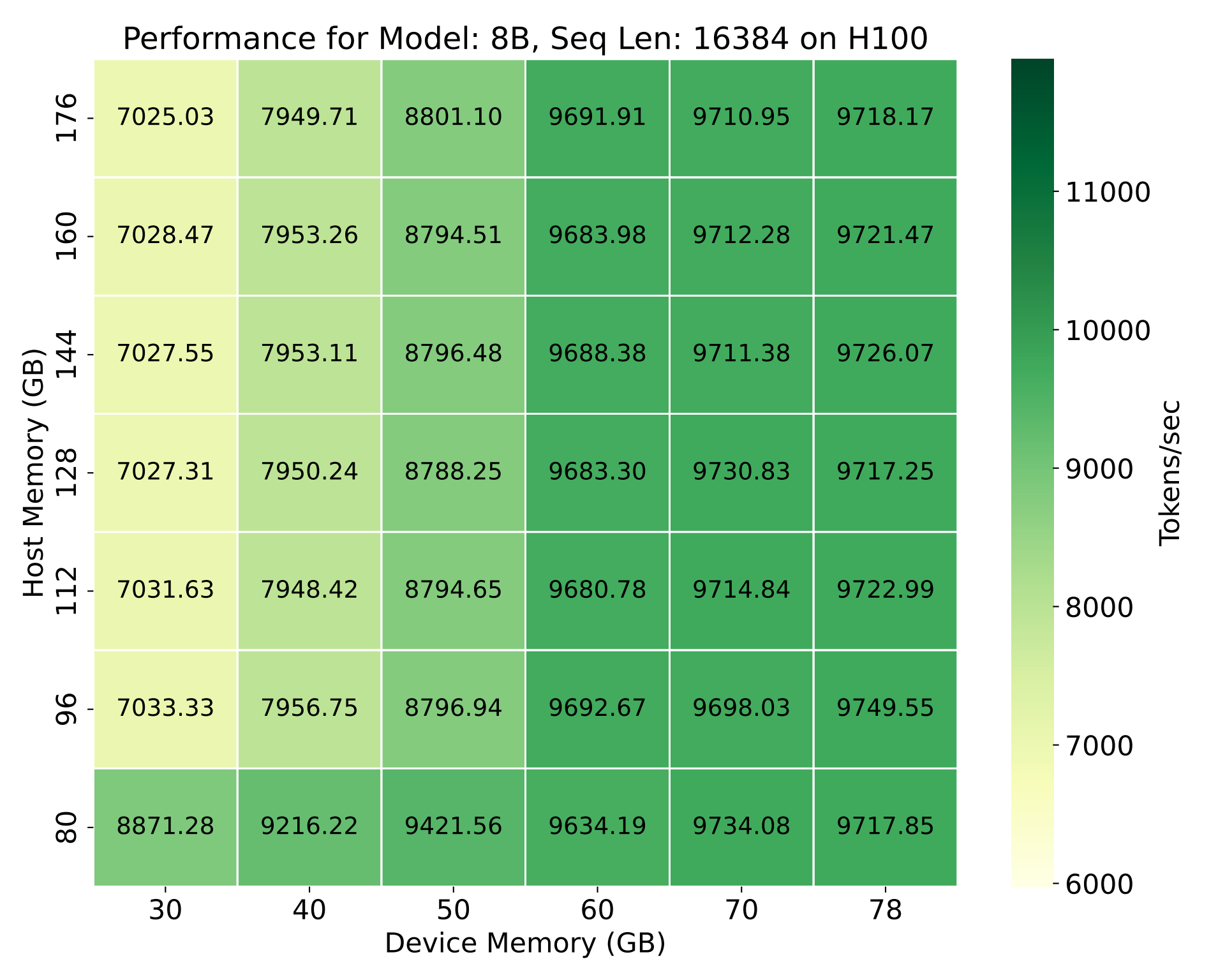
Performance for Model: 8B, Seq Len: 16384 on H100
Host Memory (GB)
176
160
144
128
112
96
80
7025.03
7949.71
8801.10
9691.91
9710.95
9718.17
7028.47
7953.26
8794.51
9683.98
9712.28
9721.47
7027.55
7953.11
8796.48
9688.38
9711.38
9726.07
7027.31
7950.24
8788.25
9683.30
9730.83
9717.25
7031.63
7948.42
8794.65
9680.78
9714.84
9722.99
7033.33
7956.75
8796.94
9692.67
9698.03
9749.55
8871.28
9216.22
9421.56
9634.19
9734.08
9717.85
30
40
50
60
70
78
Device Memory (GB)
Tokens/sec
11000
10000
9000
8000
7000
6000