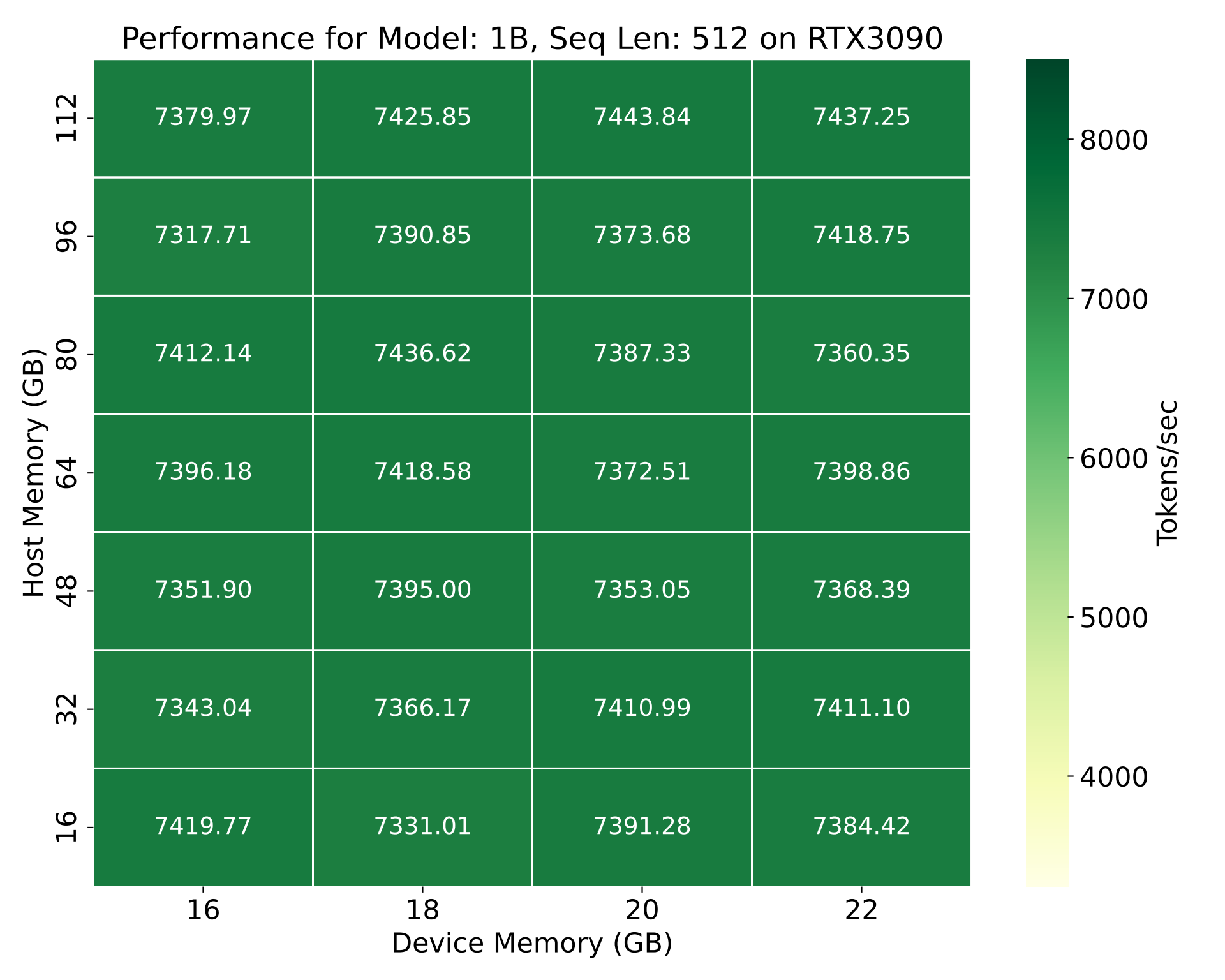
Performance for Model: 1B, Seq Len: 512 on RTX3090
Host Memory (GB)
112
96
80
64
48
32
16
7379.97
7425.85
7443.84
7437.25
7317.71
7390.85
7373.68
7418.75
7412.14
7436.62
7387.33
7360.35
7396.18
7418.58
7372.51
7398.86
7351.90
7395.00
7353.05
7368.39
7343.04
7366.17
7410.99
7411.10
7419.77
7331.01
7391.28
7384.42
16
18
20
22
Device Memory (GB)
Tokens/sec
8000
7000
6000
5000
4000