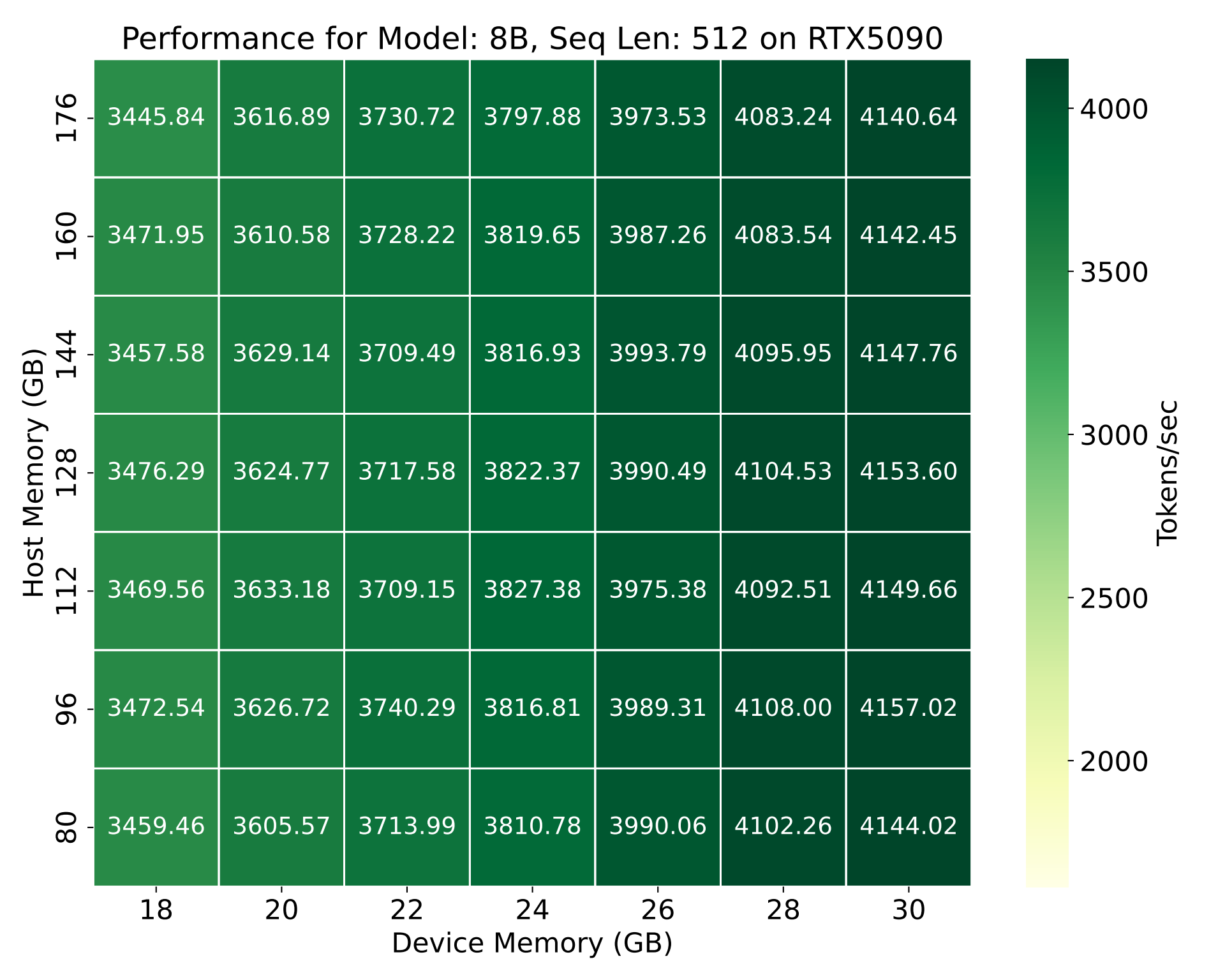
Performance for Model: 8B, Seq Len: 512 on RTX5090
Host Memory (GB)
176
160
144
128
112
96
80
3445.84 3616.89 3730.72 3797.88 3973.53 4083.24 4140.64
3471.95 3610.58 3728.22 3819.65 3987.26 4083.54 4142.45
3457.58 3629.14 3709.49 3816.93 3993.79 4095.95 4147.76
3476.29 3624.77 3717.58 3822.37 3990.49 4104.53 4153.60
3469.56 3633.18 3709.15 3827.38 3975.38 4092.51 4149.66
3472.54 3626.72 3740.29 3816.81 3989.31 4108.00 4157.02
3459.46 3605.57 3713.99 3810.78 3990.06 4102.26 4144.02
18 20 22 24 26 28 30
Device Memory (GB)
Tokens/sec
4000
3500
3000
2500
2000