Performance for Model: 8B, Seq Len: 1024 on RTX5090 - 180 154.02 166.08 157.38 162.90 172.73 178.99 181.15 153.69 157.43 161.29 165.71 173.13 178.59 181.57 -160 153.95 157.97 161.81 165.82 179.31 173.25 181.53 140 161.20 171.36 170.98 172.73 176.97 179.17 181.21 - 120 164.61 168.61 172.60 173.13 176.30 176.63 179.16 - 100 162.31 163.16 165.43 166.41 168.92 170.76 171.74 - 80 18 20 22 24 26 30

Device Memory (GB)

28

96

80

Host Memory (GB) 70 75

65