Performance for Model: 8B, Seq Len: 1024 on RTX5090 161.88 165.20 171.25 173.68 182.09 185.02 186.39 180 160.46 164.97 170.03 174.18 182.23 186.26 184.63 -160 160.81 171.41 165.38 174.94 182.05 185.12 186.48 140 161.46 165.78 170.94 174.50 181.81 184.45 186.33 120 161.20 166.20 170.58 174.66 181.65 184.81 185.88 - 100 161.67 165.64 169.71 174.20 181.71 185.53 186.36 161.27 165.39 169.80 174.60 182.44 184.23 185.96 - 80 18 20 22 24 30 26 28

Device Memory (GB)

176

160

Host Memory (GB) 12 128 144

96