Performance for Model: 8B, Seq Len: 8192 on H100 16000 7454.32 9147.92 8076.75 10493.35 10759.92 10777.28 14000 7453.46 8075.41 9149.43 10480.00 10754.79 10755.14 7451.89 8069.53 9136.15 10464.02 10738.06 10759.64 **Tokens per Second** 12000 7453.37 8076.92 9138.68 10478.79 10737.27 10746.98 10000 7453.54 8070.59 9135.89 10472.62 10738.90 10740.55 -8000 7459.05 8081.81 9151.16 10497.16 10761.26 10771.79 - 6000 9614.47 9905.43 10212.02 10546.01 10755.43 10761.29 70 30 40 50 60 78

Device Memory (GB)

176

160

Host Memory (GB) 12 128 144