Performance for Model: 8B, Seq Len: 32768 on H100 6879.56 7135.56 7364.15 7657.83 7688.22 7687.26 9000 6880.47 7125.71 7356.93 7657.66 7670.25 7682.57 8000 6875.74 7123.67 7354.81 7636.88 7663.78 7699.04 Tokens/sec 7000 7069.83 7228.35 7375.54 7651.23 7662.20 7666.94 7310.28 7368.34 7445.38 7528.95 7606.36 7667.42 6000 7135.81 7210.88 7292.65 7373.40 7446.32 7506.95 - 5000 6969.44 7050.38 7119.98 7178.06 7265.79 7315.90 30 40 50 60 70 78

Device Memory (GB)

176

160

Host Memory (GB) 12 128 144

96