Performance for Model: 8B, Seq Len: 65536 on H100 7000 5114.41 5414.52 5434.48 5492.03 5398.53 5469.44 6500 5100.17 5338.73 5366.34 5385.12 5411.06 5428.95 6000 5039.02 5285.25 5301.95 5330.67 5358.40 5377.59 5500 5001.94 5236.96 5259.50 5297.56 5304.13 5324.96 5000 4978.72 5189.37 5208.21 5220.68 5256.95 5273.87 - 4500 4936.03 5158.94 5176.92 5196.91 5224.24 5241.90 - 4000 4456.45 4704.76 4786.51 4871.75 4918.89 4957.85 30 40 50 60 70 78

Device Memory (GB)

176

160

Host Memory (GB) 12 128 144

96