Performance for Model: 8B, Seq Len: 2048 on H100 350.19 402.59 470.85 535.97 557.61 557.98 800 349.75 402.12 470.92 535.59 556.96 558.19 700 350.07 402.07 470.18 535.24 556.73 557.89 600 349.99 402.32 470.45 535.01 556.80 557.30 500 349.99 402.16 470.37 535.26 556.83 557.67 400 350.33 471.01 536.22 402.65 557.58 557.83 - 300 474.83 483.26 508.25 540.97 557.44 558.34

60

78

70

50

Device Memory (GB)

176

160

Host Memory (GB) 12 128 144

96

80

30

40