Performance for Model: 1B, Seq Len: 16384 on RTX5090 176 145.92 141.84 145.17 146.58 146.68 160 141.65 146.15 146.62 146.44 145.08 -160 144 146.69 141.55 145.57 145.84 146.58 128 -140 141.22 145.22 145.85 146.47 146.47 Host Memory (GB) 141.21 145.11 145.70 146.31 146.50 TFLOPS/s 141.31 144.85 145.64 146.41 146.56 141.22 144.89 145.65 146.38 146.41 - 100 64 140.91 144.78 145.79 146.32 146.31 48 141.57 146.32 144.99 145.45 146.39 80 32 141.15 146.04 146.14 144.95 145.51 - 60 16 146.10 140.22 144.49 145.28 145.81 16 20 24 28 30 Device Memory (GB)