RTX5090 Performance Report
Model: 1B, Sequence Length: 1024

Tokens/sec

Host Memory (GiB)	16	20	24	28	30
48	20278.32	21782.10	21752.72	21755.65	21809.12
40	20310.68	21728.66	21735.74	21836.70	21766.94
32	20211.88	21720.24	21789.39	21665.65	21762.04
24	20358.39	21787.49	21739.49	21818.88	21760.67
16	20397.40	21782.01	21688.04	21848.98	21660.10

MFU

Host Memory (GiB)	16	20	24	28	30
48	0.74	0.79	0.79	0.79	0.79
40	0.74	0.79	0.79	0.79	0.79
32	0.73	0.79	0.79	0.79	0.79
24	0.74	0.79	0.79	0.79	0.79
16	0.74	0.79	0.79	0.79	0.79

Model TFLOPS/sec

Host Memory (GiB)	16	20	24	28	30
48	154.44	165.89	165.66	165.69	166.09
40	154.68	165.48	165.54	166.30	165.77
32	153.93	165.42	165.94	165.00	165.74
24	155.05	165.93	165.56	166.17	165.73
16	155.34	165.89	165.17	166.40	164.96

HFU

Host Memory (GiB)	16	20	24	28	30
48	0.74	0.80	0.79	0.79	0.80
40	0.74	0.79	0.79	0.80	0.79
32	0.74	0.79	0.80	0.79	0.79
24	0.74	0.80	0.79	0.80	0.79
16	0.74	0.80	0.79	0.80	0.79

Device Memory (GiB)