RTX5090 Performance Report
Model: 8B, Sequence Length: 1024

Tokens/sec

Host Memory (GB) \ Device Memory (GB)	18	20	22	24	26	28	30
96	3360.38	3433.78	3554.25	3623.55	3768.61	3905.22	3952.42
80	3353.17	3434.84	3519.03	3615.48	3777.48	3896.52	3961.62
75	3358.90	3446.66	3530.45	3617.85	3779.93	3912.19	3960.74
70	3517.19	3738.82	3730.51	3768.68	3861.13	3909.28	3953.61
65	3591.45	3678.85	3765.73	3777.41	3846.63	3853.69	3909.04
60	3541.24	3559.96	3609.37	3630.88	3685.54	3725.75	3746.98

MFU

Host Memory (GB) \ Device Memory (GB)	18	20	22	24	26	28	30
96	0.74	0.75	0.78	0.79	0.82	0.85	0.86
80	0.73	0.75	0.77	0.79	0.83	0.85	0.87
75	0.73	0.75	0.77	0.79	0.83	0.86	0.87
70	0.77	0.82	0.82	0.82	0.84	0.86	0.86
65	0.79	0.80	0.82	0.83	0.84	0.84	0.86
60	0.77	0.78	0.79	0.79	0.81	0.82	0.82

TFLOPS/s

Host Memory (GB) \ Device Memory (GB)	18	20	22	24	26	28	30
96	154.02	157.38	162.90	166.08	172.73	178.99	181.15
80	153.69	157.43	161.29	165.71	173.13	178.59	181.57
75	153.95	157.97	161.81	165.82	173.25	179.31	181.53
70	161.20	171.36	170.98	172.73	176.97	179.17	181.21
65	164.61	168.61	172.60	173.13	176.30	176.63	179.16
60	162.31	163.16	165.43	166.41	168.92	170.76	171.74

HFU

Host Memory (GB) \ Device Memory (GB)	18	20	22	24	26	28	30
96	0.74	0.75	0.78	0.80	0.83	0.86	0.87
80	0.74	0.75	0.77	0.79	0.83	0.85	0.87
75	0.74	0.76	0.77	0.79	0.83	0.86	0.87
70	0.82	0.86	0.86	0.85	0.86	0.87	0.87
65	0.88	0.90	0.91	0.91	0.91	0.91	0.91
60	0.91	0.91	0.92	0.92	0.92	0.92	0.92

Host Memory (GB)

Device Memory (GB)