H100 Performance Report
Model: 32B, Sequence Length: 2048

Tokens/sec

Host Memory (GiB)	30	40	50	60	70	77
356		2664.41	2669.40	2710.71	2733.76	2729.08
320		2659.29	2667.96	2711.83	2727.35	2730.42
288		2598.25	2675.69	2692.63	2692.24	2687.37
256		2588.86	2602.03	2627.66	2642.41	2660.15
224						

MFU

Host Memory (GiB)	30	40	50	60	70	77
356		0.53	0.53	0.54	0.54	0.54
320		0.53	0.53	0.54	0.54	0.54
288		0.51	0.53	0.53	0.53	0.53
256		0.51	0.52	0.52	0.52	0.53
224						

Model TFLOPS/sec

Host Memory (GiB)	30	40	50	60	70	77
356		522.04	523.02	531.12	535.63	534.71
320		521.04	522.74	531.33	534.38	534.98
288		509.08	524.25	527.57	527.50	526.54
256		507.24	509.82	514.84	517.73	521.21
224						

HFU

Host Memory (GiB)	30	40	50	60	70	77
356		0.60	0.59	0.60	0.60	0.59
320		0.60	0.59	0.60	0.60	0.59
288		0.60	0.61	0.61	0.61	0.60
256		0.62	0.62	0.62	0.62	0.62
224						

Device Memory (GiB)