H100 Performance Report
Model: 15B, Sequence Length: 2048

Tokens/sec

Host Memory (GiB)	24	30	40	50	60	70	77
192	5459.74	5550.87	5607.27	5525.06	5756.33	5676.66	5991.17
160	5473.42	5553.46	5591.34	5522.37	5761.70	5686.27	5991.08
128	5466.33	5529.31	5570.11	5663.25	5681.35	5871.72	5922.61
112							
96							
80							

MFU

Host Memory (GiB)	24	30	40	50	60	70	77
192	0.50	0.51	0.51	0.50	0.52	0.52	0.55
160	0.50	0.51	0.51	0.50	0.53	0.52	0.55
128	0.50	0.50	0.51	0.52	0.52	0.54	0.54
112							
96							
80							

Model TFLOPS/sec

Host Memory (GiB)	24	30	40	50	60	70	77
192	492.06	500.27	505.36	497.95	518.79	511.61	539.95
160	493.29	500.51	503.92	497.70	519.27	512.48	539.95
128	492.65	498.33	502.01	510.40	512.03	529.19	533.78
112							
96							
80							

HFU

Host Memory (GiB)	24	30	40	50	60	70	77
192	0.60	0.58	0.60	0.56	0.58	0.55	0.59
160	0.61	0.58	0.59	0.56	0.58	0.55	0.59
128	0.60	0.60	0.60	0.60	0.59	0.60	0.60
112							
96							
80							

Device Memory (GiB)