Performance for Model: 8B, Seq Len: 512 on H100

Host Memory (GB)	30	40	50	60	70	78
176	7523.81	8645.64	10205.87	11744.06	12272.46	12320.83
160	7514.22	8635.04	10197.63	11733.94	12278.76	12324.62
144	7518.30	8642.48	10191.83	11741.81	12282.04	12305.23
128	7516.02	8638.25	10202.08	11752.74	12288.68	12303.61
112	7523.81	8642.03	10195.69	11735.41	12288.67	12304.35
96	7523.61	8647.34	10210.83	11739.78	12302.34	12313.90
80	10183.36	10527.10	11102.16	11841.51	12294.88	12324.10

Device Memory (GB)