Performance for Model: 1B, Seq Len: 4096 on RTX3090 112 5576.43 5578.65 5579.28 5572.94 7000 96 5533.61 5567.10 5558.03 5566.31 Host Memory (GB) 48 64 80 5536.66 5555.18 5560.34 5535.45 6000 5518.42 5528.01 5561.78 5534.44 5000 5513.43 5516.32 5522.16 5545.08 32 5532.92 5510.95 5517.83 5521.98 - 4000 16 5526.04 5497.35 5502.86 5494.60 20 16 18 22 Device Memory (GB)