Performance for Model: 1B, Seq Len: 512 on RTX5090 176 135.28 168.45 168.49 168.41 168.46 160 135.41 168.45 168.51 168.17 168.60 -160 144 136.31 168.42 168.29 168.58 168.34 128 -140 135.39 168.26 168.09 168.58 168.53 Host Memory (GB) 134.56 168.38 168.43 168.10 168.27 120 135.39 168.61 168.30 168.24 168.29 134.90 168.54 168.04 168.39 167.95 - 100 64 135.84 168.35 168.38 167.94 168.42 48 133.90 168.33 167.97 168.41 168.38 80 32 133.69 168.40 167.91 168.01 168.24 60 16 135.98 168.15 168.43 168.46 168.31 16 20 28 24 30 Device Memory (GB)