Performance for Model: 1B, Seq Len: 65536 on RTX5090 176 6351.97 6404.77 6406.48 6426.67 6435.03 160 8000 6343.01 6406.17 6410.41 6419.02 6438.71 144 6336.60 6402.23 6416.12 6422.82 6433.09 7000 128 6410.41 6341.45 6402.81 6418.57 6434.07 Second Host Memory (GB) 6339.31 6392.19 6413.98 6426.38 6433.33 6000 **Tokens** per 6403.22 6337.35 6409.93 6417.20 6425.74 6426.81 6338.43 6389.52 6400.46 6406.87 5000 64 6335.52 6387.69 6394.92 6409.30 6417.88 48 6272.06 6344.39 6377.53 6400.88 6421.40 - 4000 6169.63 6335.87 6356.12 6294.66 6317.28 - 3000 16 5815.48 6034.19 6119.75 6246.34 6203.99 16 30 28 20 24 Device Memory (GB)