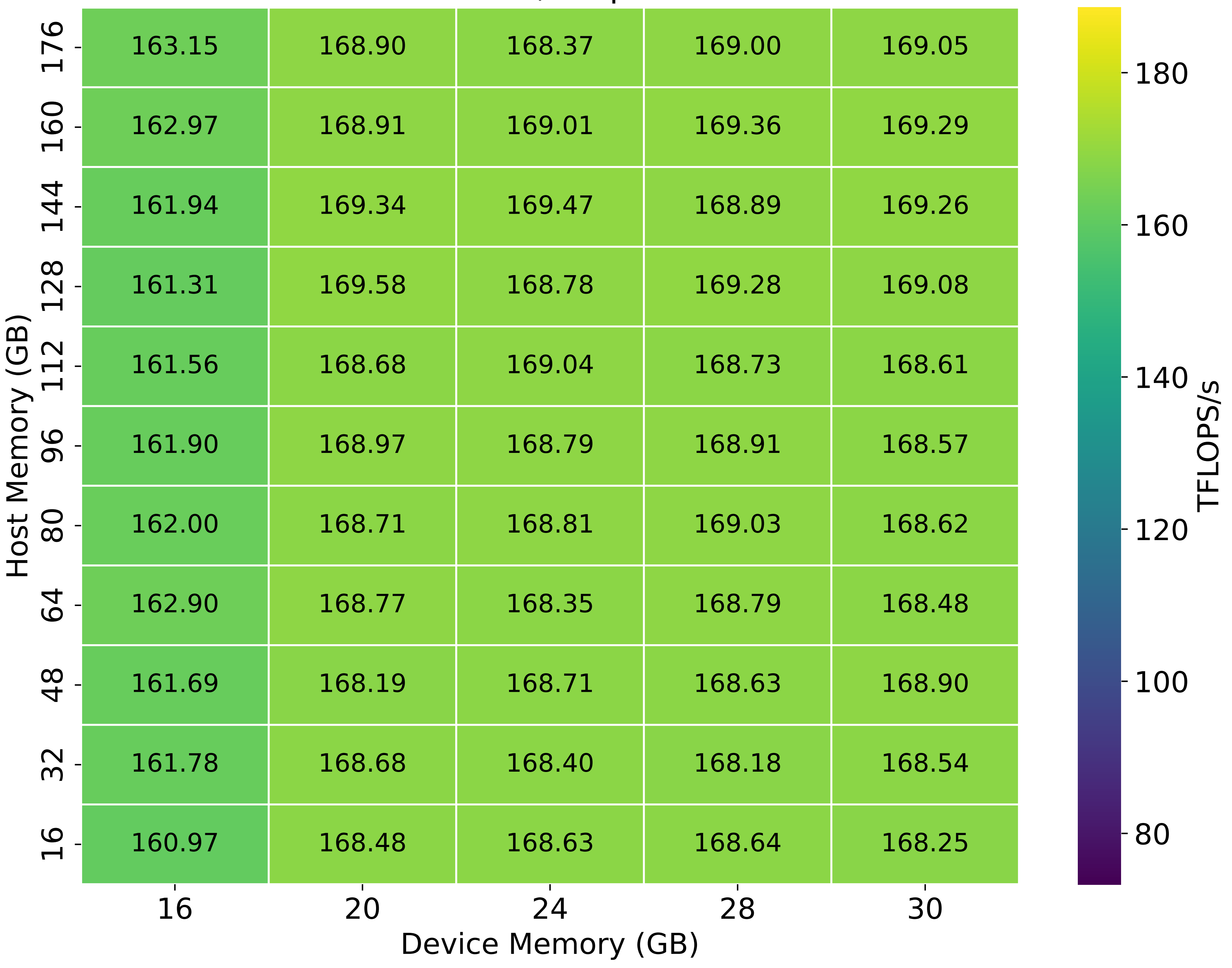
Performance for Model: 1B, Seq Len: 1024 on RTX5090



Host Memory (GB) \ Device Memory (GB)	16	20	24	28	30
176	163.15	168.90	168.37	169.00	169.05
160	162.97	168.91	169.01	169.36	169.29
144	161.94	169.34	169.47	168.89	169.26
128	161.31	169.58	168.78	169.28	169.08
112	161.56	168.68	169.04	168.73	168.61
96	161.90	168.97	168.79	168.91	168.57
80	162.00	168.71	168.81	169.03	168.62
64	162.90	168.77	168.35	168.79	168.48
48	161.69	168.19	168.71	168.63	168.90
32	161.78	168.68	168.40	168.18	168.54
16	160.97	168.48	168.63	168.64	168.25