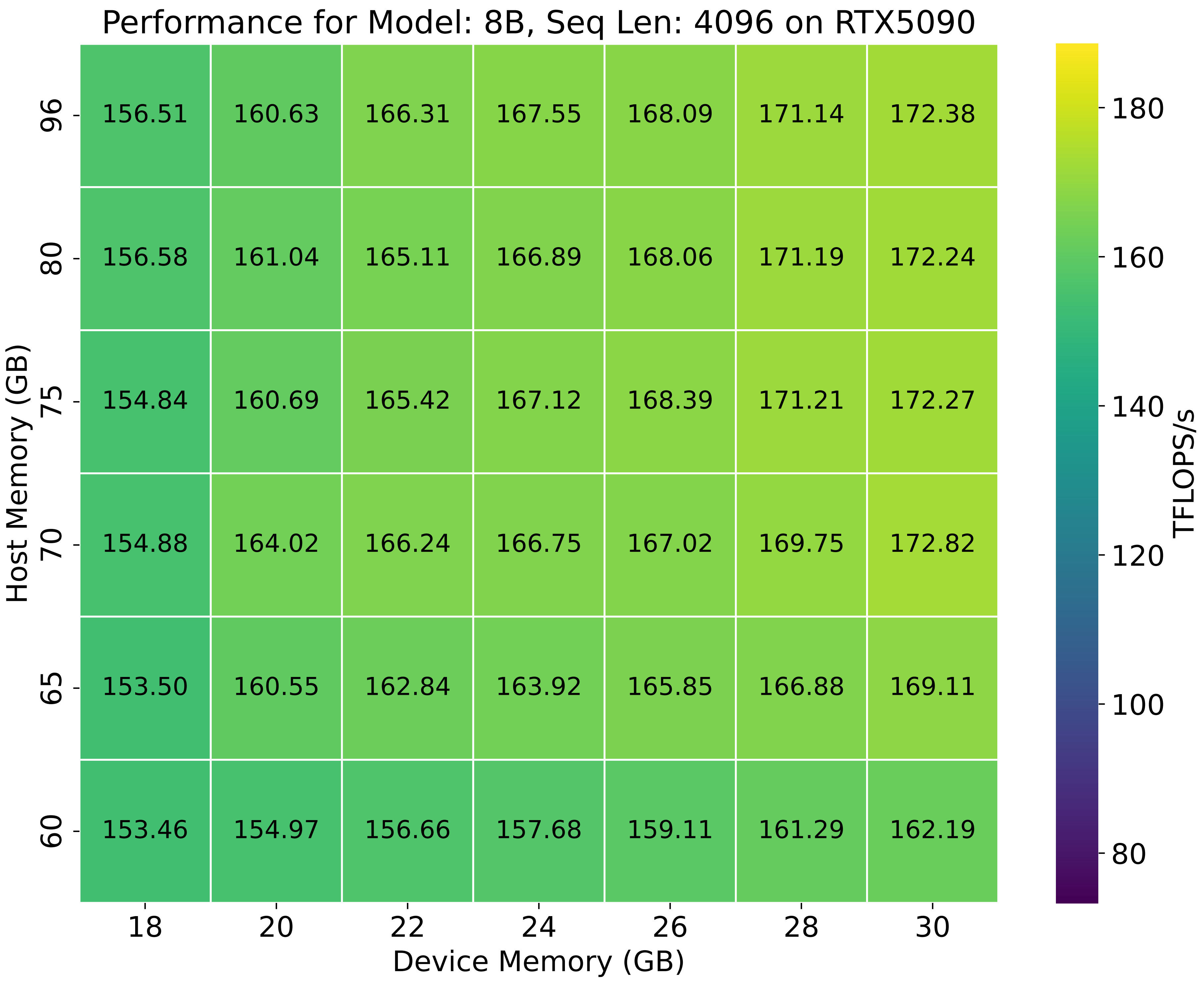

Performance for Model: 8B, Seq Len: 4096 on RTX5090

Host Memory (GB) \ Device Memory (GB)	18	20	22	24	26	28	30
96	156.51	160.63	166.31	167.55	168.09	171.14	172.38
80	156.58	161.04	165.11	166.89	168.06	171.19	172.24
75	154.84	160.69	165.42	167.12	168.39	171.21	172.27
70	154.88	164.02	166.24	166.75	167.02	169.75	172.82
65	153.50	160.55	162.84	163.92	165.85	166.88	169.11
60	153.46	154.97	156.66	157.68	159.11	161.29	162.19