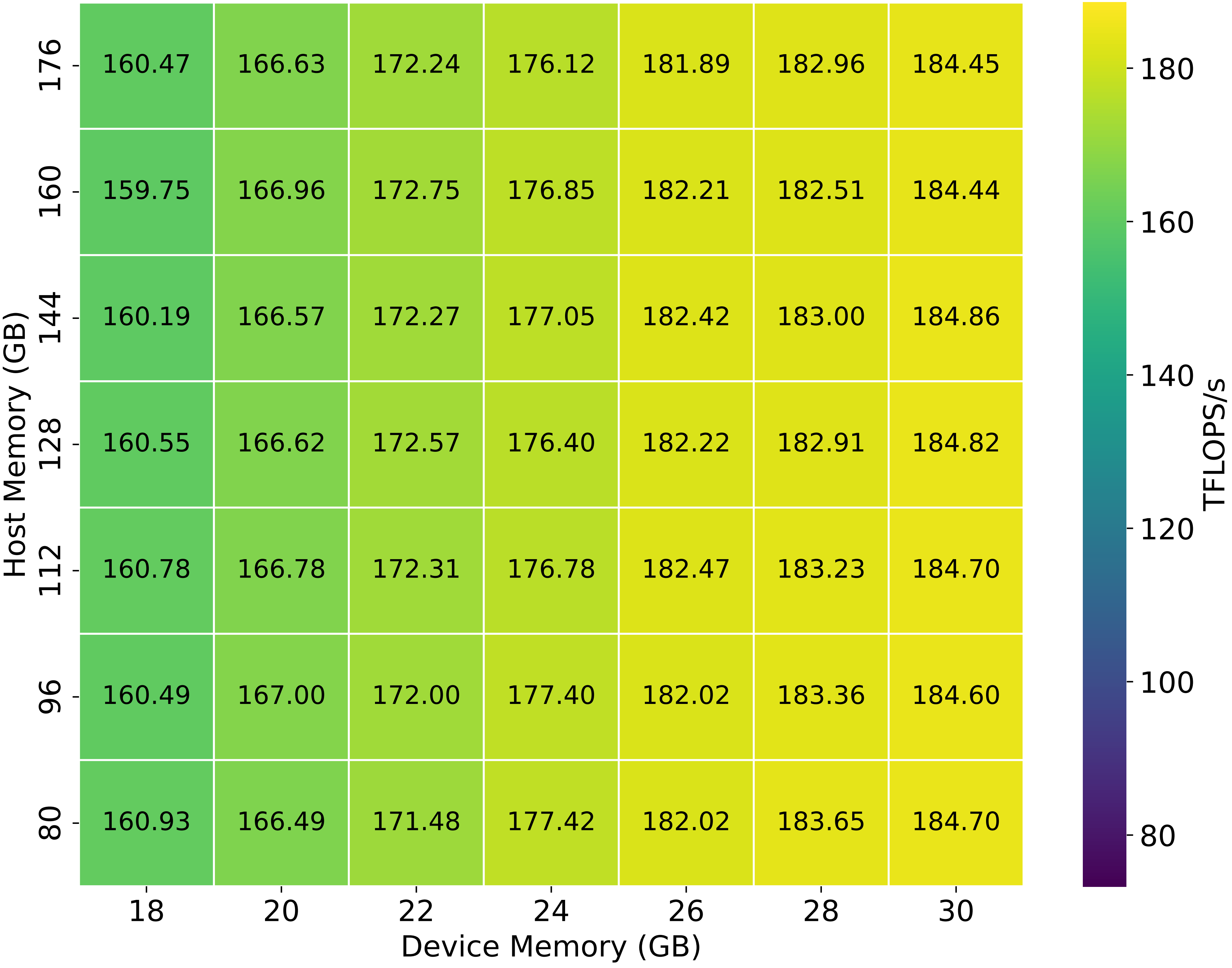

Performance for Model: 8B, Seq Len: 2048 on RTX5090

Host Memory (GB) \ Device Memory (GB)	18	20	22	24	26	28	30
176	160.47	166.63	172.24	176.12	181.89	182.96	184.45
160	159.75	166.96	172.75	176.85	182.21	182.51	184.44
144	160.19	166.57	172.27	177.05	182.42	183.00	184.86
128	160.55	166.62	172.57	176.40	182.22	182.91	184.82
112	160.78	166.78	172.31	176.78	182.47	183.23	184.70
96	160.49	167.00	172.00	177.40	182.02	183.36	184.60
80	160.93	166.49	171.48	177.42	182.02	183.65	184.70