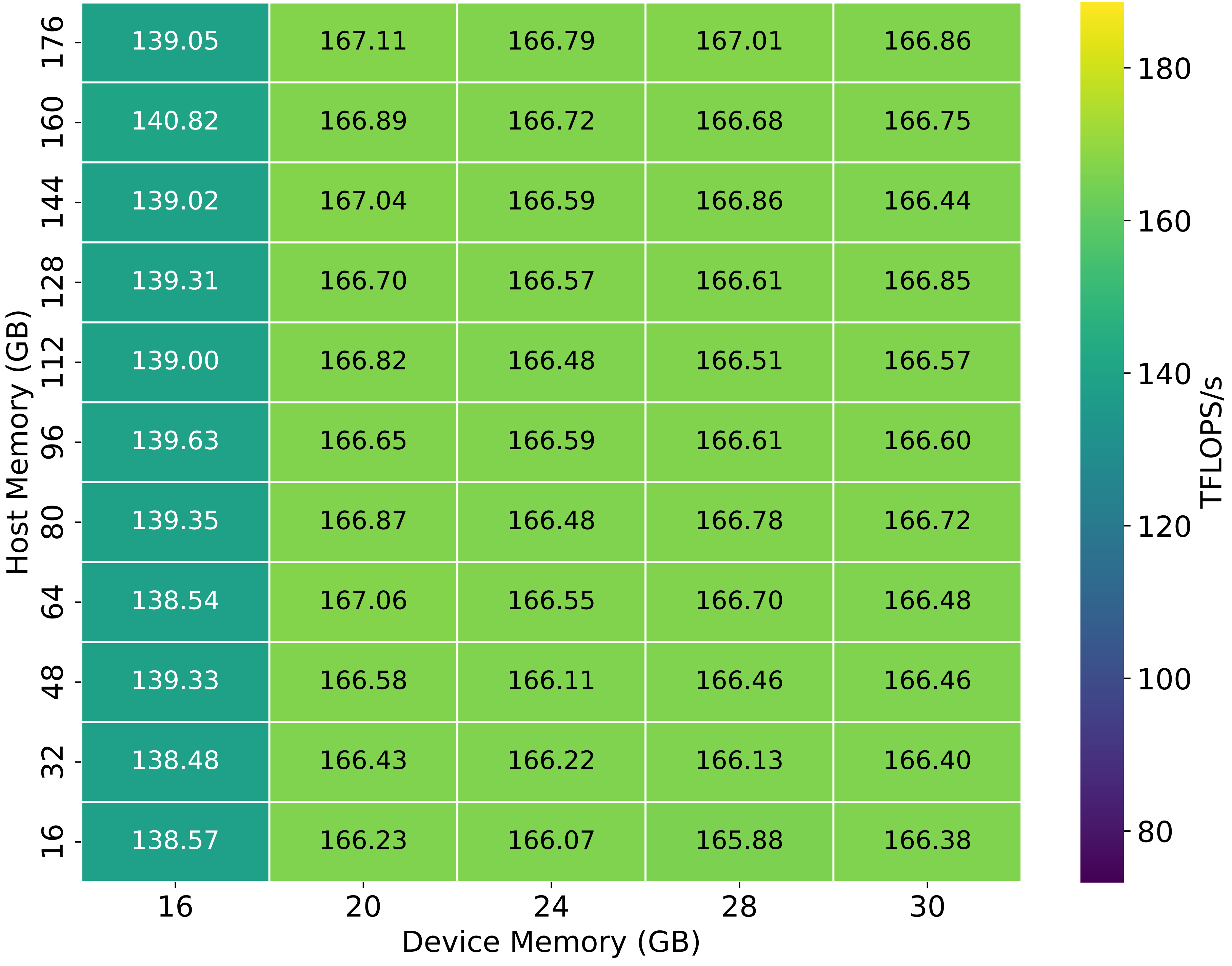
Performance for Model: 1B, Seq Len: 2048 on RTX5090



Host Memory (GB)	16	20	24	28	30
176	139.05	167.11	166.79	167.01	166.86
160	140.82	166.89	166.72	166.68	166.75
144	139.02	167.04	166.59	166.86	166.44
128	139.31	166.70	166.57	166.61	166.85
112	139.00	166.82	166.48	166.51	166.57
96	139.63	166.65	166.59	166.61	166.60
80	139.35	166.87	166.48	166.78	166.72
64	138.54	167.06	166.55	166.70	166.48
48	139.33	166.58	166.11	166.46	166.46
32	138.48	166.43	166.22	166.13	166.40
16	138.57	166.23	166.07	165.88	166.38