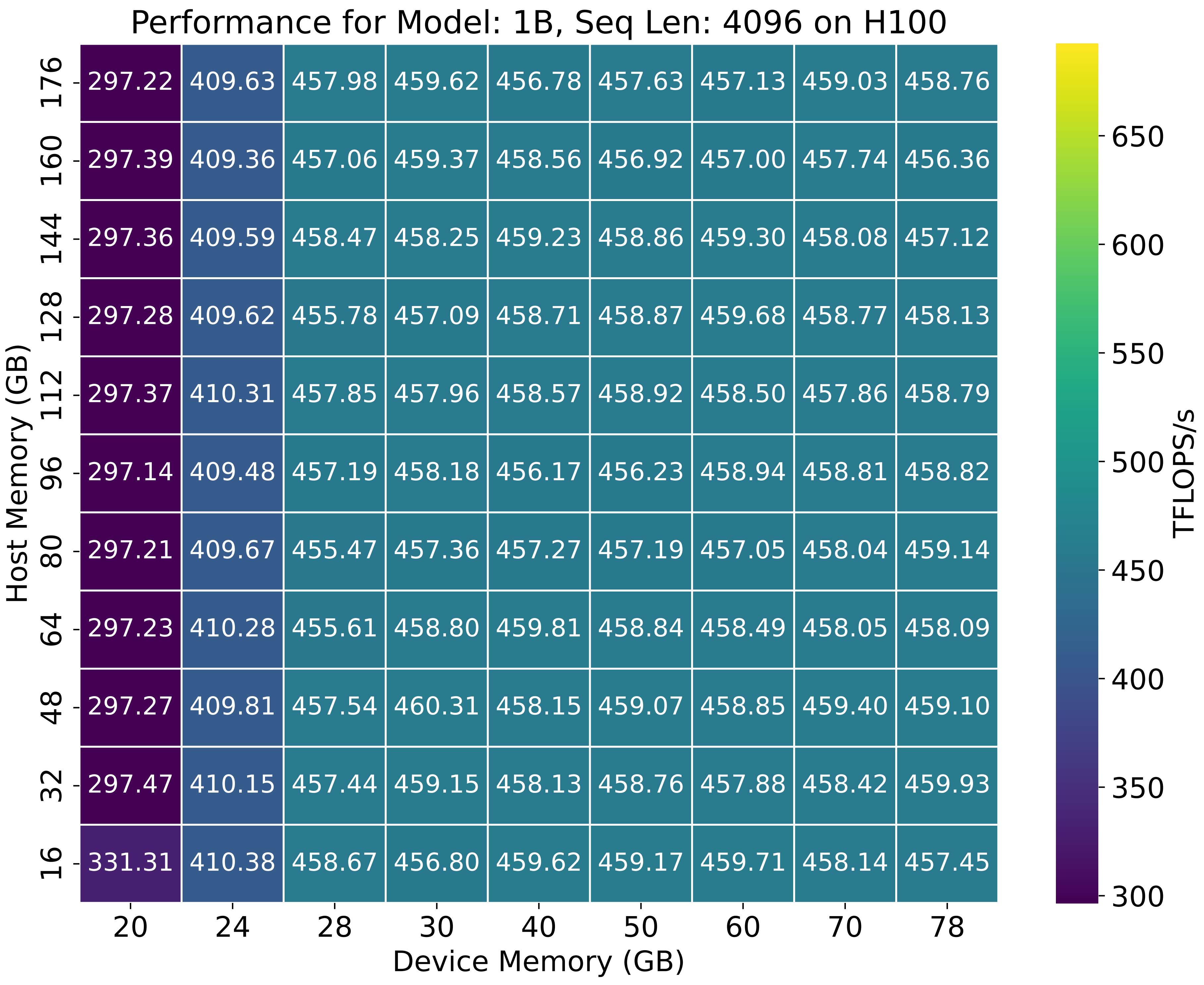

Performance for Model: 1B, Seq Len: 4096 on H100

Host Memory (GB) \ Device Memory (GB)	20	24	28	30	40	50	60	70	78
176	297.22	409.63	457.98	459.62	456.78	457.63	457.13	459.03	458.76
160	297.39	409.36	457.06	459.37	458.56	456.92	457.00	457.74	456.36
144	297.36	409.59	458.47	458.25	459.23	458.86	459.30	458.08	457.12
128	297.28	409.62	455.78	457.09	458.71	458.87	459.68	458.77	458.13
112	297.37	410.31	457.85	457.96	458.57	458.92	458.50	457.86	458.79
96	297.14	409.48	457.19	458.18	456.17	456.23	458.94	458.81	458.82
80	297.21	409.67	455.47	457.36	457.27	457.19	457.05	458.04	459.14
64	297.23	410.28	455.61	458.80	459.81	458.84	458.49	458.05	458.09
48	297.27	409.81	457.54	460.31	458.15	459.07	458.85	459.40	459.10
32	297.47	410.15	457.44	459.15	458.13	458.76	457.88	458.42	459.93
16	331.31	410.38	458.67	456.80	459.62	459.17	459.71	458.14	457.45