

Per-Layer

Batch Size * Sequence Length			32	64	128	256	512	1024	2048	4096	8192	16384
800M Model	Q·K·V	NO HT	0.42	0.44	0.82	1.31	1.25	1.66	1.78	1.92	1.90	1.88
		HT	0.40	0.42	0.78	1.22	1.16	1.60	1.62	1.73	1.76	1.73
	O	NO HT	0.21	0.22	0.34	0.46	0.73	1.21	1.31	1.45	1.58	1.55
		HT	0.20	0.21	0.30	0.42	0.65	1.00	1.08	1.18	1.28	1.20
	Gate·Up	NO HT	0.60	0.80	1.21	1.26	1.70	1.78	2.02	2.01	2.01	2.01
		HT	0.58	0.77	1.13	1.20	1.61	1.69	1.91	1.92	1.94	1.92
	Down	NO HT	0.28	0.26	0.41	0.73	1.29	2.05	2.20	2.35	2.17	2.25
		HT	0.27	0.25	0.38	0.66	1.08	1.57	1.65	1.66	1.56	1.58
7B Model	Q·K·V	NO HT	2.55	2.44	3.04	2.18	2.70	3.18	3.39	3.55	3.59	3.61
		HT	2.47	2.34	2.86	2.05	2.51	2.93	3.11	3.31	3.33	3.25
	O	NO HT	0.36	0.41	0.70	1.08	1.92	2.19	2.49	2.83	2.85	2.72
		HT	0.35	0.39	0.66	0.98	1.65	1.86	2.08	2.30	2.26	2.17
	Gate·Up	NO HT	2.82	3.19	3.04	2.66	3.04	3.48	3.67	3.84	3.88	3.90
		HT	2.76	3.10	2.91	2.55	2.88	3.31	3.52	3.70	3.71	3.66
	Down	NO HT	1.30	1.35	1.34	1.69	2.83	3.26	3.51	3.47	3.52	2.88
		HT	1.26	1.28	1.24	1.48	2.31	2.59	2.74	2.68	2.65	2.27

**End-to-End Prefill (800M)**

Batch Size	Sequence Len	32	64	128	256	512 (max)
1	NO HT	0.76	0.78	0.90	1.04	1.21
	HT	0.75	0.76	0.87	0.99	1.15
2	NO HT	0.78	0.89	1.04	1.19	1.53
	HT	0.76	0.86	0.98	1.11	1.40
4	NO HT	0.89	1.04	1.22	1.55	1.60
	HT	0.86	0.98	1.14	1.42	1.47
8	NO HT	1.04	1.22	1.56	1.61	1.51
	HT	0.99	1.13	1.42	1.47	1.41
16	NO HT	1.22	1.56	1.62	1.53	OOM
	HT	1.14	1.42	1.48	1.41	OOM
32	NO HT	1.55	1.62	1.53	1.39	OOM
	HT	1.41	1.48	1.42	1.31	OOM
64	NO HT	1.62	1.53	1.39	OOM	OOM
	HT	1.47	1.42	1.31	OOM	OOM
128	NO HT	1.53	1.39	OOM	OOM	OOM
	HT	1.42	1.31	OOM	OOM	OOM
256	NO HT	1.39	OOM	OOM	OOM	OOM
	HT	1.31	OOM	OOM	OOM	OOM
512	NO HT	OOM	OOM	OOM	OOM	OOM
	HT	OOM	OOM	OOM	OOM	OOM