Appendix D Decode Int8 (Page 1/2)

GPU	Model	Concurrency	cuBLASLt	cuSPARSELt Speedup Ratio			
		M	Throughput (token/s)	2:4	2:6	2:8	2:10
A100	Llama3.2-1B	64	1.12e+04	1.10	1.00	0.91	0.96
		128	1.58e+04	1.05	0.98	1.01	0.97
		256	2.30e+04	1.07	1.03	1.01	0.99
		512	2.29e+04	1.09	1.06	1.06	1.02
	BitNet-2B	64	7.09e+03	1.16	1.11	1.10	1.07
		128	1.09e+04	1.14	1.07	1.05	1.03
		256	1.60e+04	1.14	1.06	1.05	1.03
		512	1.59e+04	1.17	1.09	1.08	1.07
	Llama3.2-3B	64	6.69e+03	1.09	1.04	1.01	0.99
		128	1.01e+04	1.16	1.08	1.06	1.05
		256	1.42e+04	1.13	1.06	1.05	1.03
		512	1.42e+04	1.18	1.10	1.07	1.04
	Qwen2.5-7B	64	5.00e+03	1.25	1.14	1.12	1.08
		128	8.17e+03	1.21	1.09	1.05	1.03
		256	1.03e+04	1.26	1.12	1.08	1.04
		512	9.82e+03	1.31	1.16	1.12	1.09
	Qwen2.5-14B	64	3.08e+03	1.28	1.14	1.10	1.08
		128	4.84e+03	1.24	1.12	1.09	1.04
		256	5.77e+03	1.35	1.19	1.13	1.10
		512	5.39e+03	1.40	1.23	1.17	1.14
RTX4090	Llama3.2-1B	64	9.37e+03	1.15	1.06	0.97	0.94
		128	1.40e+04	1.03	1.01	1.00	0.96
		256	2.02e+04	1.08	0.97	1.03	1.01
		512	2.01e+04	1.08	1.04	1.02	1.00
	BitNet-2B	64	6.99e+03	1.05	0.98	0.96	0.95
		128	1.06e+04	1.17	1.05	1.10	1.07
		256	1.46e+04	1.19	1.13	1.13	1.06
		512	1.32e+04	1.34	1.26	1.27	1.18
	Llama3.2-3B	64	5.68e+03	0.98	0.97	0.96	0.92
		128	8.19e+03	1.20	1.15	1.12	1.10
		256	1.15e+04	1.16	1.10	1.09	1.07
		512	1.16e+04	1.17	1.09	1.05	1.03
	Qwen2.5-7B	64	3.99e+03	1.22	1.12	0.94	0.90
		128	6.33e+03	1.32	1.18	1.04	1.01
		256	8.69e+03	1.23	1.13	1.01	1.00
		512	8.70e+03	1.26	1.14	1.06	1.04
	Qwen2.5-14B	64	1.53e+03	1.94	1.14	0.64	
		128	2.38e+03	1.75	1.14	0.79	0.40
		256	2.43e+03	1.67	1.12	0.58	
		512	1.37e+03	3.04	1.23		
H100	Llama3.2-1B	64	1.31e+04	1.11	1.09	0.98	1.04
		128	1.83e+04	1.05	1.11	1.06	1.11
		256	2.91e+04	0.92	1.02	1.01	0.98
		512	2.99e+04	1.09	1.05	1.05	1.03
	BitNet-2B	64	8.11e+03	1.10	1.10	1.08	1.05
		128	1.32e+04	1.16	1.11	1.09	1.08
		256	2.03e+04	1.10	1.03	1.03	1.00
		512	2.07e+04	1.13	1.06	1.08	1.04
	Llama3.2-3B	64	7.34e+03	1.11	1.12	1.11	1.07
		128	1.14e+04	1.23	1.20	1.16	1.16
		256	1.80e+04	1.02	1.03	0.99	0.95
		512	1.81e+04	1.11	1.09	1.07	1.04
	Qwen2.5-7B	64	5.72e+03	1.24	1.15	1.15	1.10
		128	9.93e+03	1.24	1.13	1.12	1.09
		256	1.46e+04	1.12	1.04	1.01	0.98
		512	1.47e+04	1.19	1.05	1.07	1.03
	Qwen2.5-14B	64	3.31e+03	1.49	1.27	1.27	1.24
		128	5.67e+03	1.32	1.19	1.16	1.13
		256	8.29e+03	1.21	1.07	1.04	0.96
		512	8.08e+03	1.28	1.14	1.10	1.06
B200	Llama3.2-1B	64	1.71e+04	1.17	1.18	1.16	1.14
		128	2.54e+04	1.13	1.11	1.10	1.10
		256	4.00e+04	1.13	0.95	0.98	1.07
		512	4.36e+04	1.12	1.09	1.09	1.08
	BitNet-2B	64	9.97e+03	1.05	1.10	1.11	1.10

≥1.5× ≥1.0×

Appendix D Decode Int8 (Page 2/2)

GPU	Model	Concurrency	cuBLASLt	cuSPARSELt Speedup Ratio			
		M	Throughput (token/s)	2:4	2:6	2:8	2:10
		128	1.72e+04	1.24	1.19	1.19	1.19
		256	2.65e+04	1.26	1.11	1.12	1.12
		512	3.11e+04	1.19	1.13	1.13	1.13
	Llama3.2-3B	64	1.11e+04	1.14	1.11	1.09	1.07
		128	1.72e+04	1.24	1.19	1.18	1.17
		256	2.59e+04	1.15	1.12	1.12	1.07
		512	3.04e+04	1.17	1.14	1.13	1.10
	Qwen2.5-7B	64	9.57e+03	1.28	1.25	1.23	1.21
		128	1.54e+04	1.30	1.21	1.21	1.18
		256	2.47e+04	1.17	1.10	1.08	1.06
		512	2.65e+04	1.24	1.15	1.14	1.12
	Qwen2.5-14B	64	5.67e+03	1.36	1.38	1.35	1.31
		128	1.00e+04	1.31	1.22	1.21	1.19
		256	1.66e+04	1.13	1.08	1.10	1.04
		512	1.74e+04	1.25	1.15	1.12	1.11
RTX5080	Llama3.2-1B	64	1.40e+04	1.18	1.11	1.07	1.05
		128	2.04e+04	1.13	1.07	1.06	1.04
		256	2.23e+04	1.15	1.09	1.03	1.05
		512	2.25e+04	1.11	1.06	1.04	1.01
	BitNet-2B	64	8.37e+03	1.11	1.09	1.04	1.02
		128	1.33e+04	1.18	1.09	1.07	1.04
		256	1.55e+04	1.16	1.08	1.03	1.06
		512	1.42e+04	1.15	1.05	1.03	1.02
	Llama3.2-3B	64	6.95e+03	1.30	1.18	1.18	1.14
		128	1.07e+04	1.27	1.18	1.14	1.11
		256	1.24e+04	1.21	1.11	1.07	1.03
		512	1.07e+04	1.19	1.10	1.06	1.04
	Qwen2.5-7B	64	4.49e+03	1.38	1.24	1.17	1.09
		128	7.50e+03	1.43	1.24	1.17	1.10
		256	6.75e+03	1.80	1.25	1.02	0.93
		512	6.33e+03	1.52	1.19	0.77	
	Qwen2.5-14B	64					
		128					
		256					
		512					
GB10	Llama3.2-1B	64	4.22e+03	1.16	1.00	1.00	0.95
		128	6.23e+03	1.18	1.04	1.05	1.01
		256	7.86e+03	1.11	1.00	1.03	0.97
		512	7.99e+03	1.13	1.04	1.05	0.99
	BitNet-2B	64	2.46e+03	1.22	0.99	1.07	0.98
		128	3.71e+03	1.18	1.05	1.06	1.03
		256	4.81e+03	1.13	1.03	1.03	1.02
		512	4.77e+03	1.17	1.07	1.07	1.06
	Llama3.2-3B	64	2.83e+02	8.56	7.82	7.43	6.58
		128	3.07e+03	1.21	1.09	1.07	0.96
		256	3.97e+03	1.12	1.05	1.03	0.96
		512	4.01e+03	1.15	1.07	1.04	0.98
	Qwen2.5-7B	64	1.14e+03	1.36	1.08	1.15	1.08
		128	1.94e+03	1.38	1.05	1.11	1.08
		256	2.55e+03	1.36	1.06	1.10	1.09
		512	2.71e+03	1.28	1.05	1.07	0.99
	Qwen2.5-14B	64	6.17e+02	1.41	1.05	1.07	1.02
		128	1.02e+03	1.40	1.06	1.09	1.04
		256	1.38e+03	1.32	1.06	1.07	1.01
		512	1.44e+03	1.28	1.03	1.05	1.02