Appendix C Prefill Int8 (Page 1/4)

GPU	Model	Batch Size	cuBLASLt	cuSPARSELt Speedup Ratio			
		M	Throughput (token/s)	2:4	2:6	2:8	2:10
A100	Llama3.2-1B	512	1.19e+04	0.88	0.89	0.84	0.88
		1024	2.46e+04	0.85	0.71	0.84	0.84
		2048	4.79e+04	0.86	0.81	0.85	0.88
		4096	9.03e+04	0.88	0.83	0.88	0.87
		8192	1.03e+05	1.37	1.21	1.23	1.11
		16384	1.06e+05	1.48	1.28	1.24	1.12
		32768	1.07e+05	1.50	1.29	1.24	1.12
	BitNet-2B	512	9.24e+03	1.02	1.05	1.08	1.09
		1024	1.83e+04	1.03	1.06	1.13	1.11
		2048	3.73e+04	1.03	1.08	1.09	1.09
		4096	5.09e+04	1.48	1.26	1.26	1.21
		8192	5.33e+04	1.52	1.26	1.26	1.21
		16384	5.43e+04	1.53	1.26	1.26	1.21
		32768	5.46e+04	1.54	1.27	1.26	1.21
	Llama3.2-3B	512	9.71e+03	1.09	1.10	1.09	1.06
		1024	2.03e+04	1.04	1.04	1.05	1.02
		2048	3.77e+04	1.13	1.15	1.13	1.14
		4096	4.09e+04	1.59	1.36	1.28	1.20
		8192	4.21e+04	1.59	1.36	1.27	1.19
		16384	4.27e+04	1.61	1.37	1.29	1.20
		32768	4.28e+04	1.62	1.37	1.29	1.20
	Qwen2.5-7B	512	1.02e+04	1.05	1.10	1.12	1.09
		1024	1.80e+04	1.24	1.25	1.18	1.14
		2048	1.98e+04	1.63	1.34	1.26	1.20
		4096	2.06e+04	1.69	1.37	1.30	1.24
		8192	2.08e+04	1.72	1.40	1.32	1.26
		16384	2.09e+04	1.75	1.41	1.34	1.27
		32768	2.08e+04	1.75	1.41	1.34	1.26
	Qwen2.5-14B	512	6.42e+03	1.03	1.04	1.01	1.06
		1024	9.68e+03	1.38	1.31	1.27	1.19
		2048	1.03e+04	1.73	1.42	1.33	1.28
		4096	1.05e+04	1.75	1.43	1.34	1.28
		8192	1.06e+04	1.77	1.43	1.33	1.27
		16384	1.06e+04	1.77	1.43	1.34	1.27
		32768	1.07e+04	1.77	1.42	1.33	1.26
RTX4090	Llama3.2-1B	512	9.21e+03	0.91	0.91	0.88	0.90
		1024	1.87e+04	0.85	0.88	0.89	0.87
		2048	3.68e+04	0.91	0.88	0.91	0.89
		4096	7.02e+04	0.92	0.90	0.93	0.91
		8192	9.14e+04	1.24	1.14	1.10	1.05
		16384	9.09e+04	1.30	1.16	1.13	1.06
		32768	9.24e+04	1.30	1.16	1.12	1.04
	BitNet-2B	512	7.79e+03	1.11	1.14	1.10	1.12
		1024	1.60e+04	1.11	1.06	1.07	1.04
		2048	3.15e+04	1.12	1.13	1.07	1.13
		4096	4.72e+04	1.43	1.16	1.23	1.13
		8192	5.36e+04	1.31	1.06	1.05	1.04
		16384	5.32e+04	1.30	1.07	1.14	1.07
		32768	5.45e+04	1.31	1.11	1.11	1.08
	Llama3.2-3B	512	6.64e+03	1.08	1.12	1.09	1.12
		1024	1.38e+04	1.01	1.12	1.07	1.11
		2048	2.78e+04	1.09	1.12	1.11	1.06
		4096	4.04e+04	1.30	1.20	1.11	1.09
		8192	3.94e+04	1.35	1.19	1.13	1.09
		16384	3.96e+04	1.34	1.18	1.12	1.09
		32768	3.98e+04	1.34	1.18	1.13	1.08
	Qwen2.5-7B	512	8.27e+03	1.06	1.06	1.07	1.06
		1024	1.63e+04	1.07	1.09	1.05	1.04
		2048	2.03e+04	1.36	1.18	1.10	1.06
		4096	2.07e+04	1.42	1.21	1.13	1.08
		8192	2.12e+04	1.39	1.18	1.11	1.08
		16384	2.13e+04	1.39	1.18	0.58	0.40
		32768	2.14e+04	0.41	0.31	0.27	0.25
	Qwen2.5-14B	512	4.49e+03	1.06	1.11	1.07	
		1024	9.26e+03	1.02	1.07	1.07	

≥1.5× ≥1.0×

Appendix C Prefill Int8 (Page 2/4)

GPU	Model	Batch Size	cuBLASLt	cuSPARSELt Speedup Ratio			
		M	Throughput (token/s)	2:4	2:6	2:8	2:10
		2048	1.11e+04	1.43	1.21	1.14	
		4096	1.14e+04	1.40	1.19	1.12	
		8192	1.14e+04	1.40	1.17	1.12	
		16384	1.13e+04	1.40	1.19		
		32768					
H100	Llama3.2-1B	512	1.82e+04	0.87	0.90	0.85	0.91
		1024	3.56e+04	0.88	0.90	0.87	0.90
		2048	6.84e+04	0.93	0.83	0.90	0.93
		4096	1.31e+05	0.88	0.86	0.87	0.86
		8192	1.59e+05	1.09	1.02	0.96	0.95
		16384	1.68e+05	1.20	1.12	1.12	1.02
		32768	1.75e+05	1.30	1.12	1.11	1.04
	BitNet-2B	512	1.49e+04	1.10	1.15	1.15	1.09
		1024	3.06e+04	1.10	1.14	1.08	1.10
		2048	5.67e+04	1.18	1.12	1.14	1.16
		4096	7.93e+04	1.20	1.11	1.09	1.05
		8192	8.87e+04	1.18	1.10	1.05	1.02
		16384	9.22e+04	1.25	1.09	1.05	1.03
		32768	9.41e+04	1.25	1.09	1.05	1.02
	Llama3.2-3B	512	1.42e+04	1.15	1.25	1.23	1.19
		1024	3.10e+04	1.11	1.15	1.16	1.15
		2048	5.72e+04	1.17	1.17	1.12	1.09
		4096	6.53e+04	1.31	1.16	1.12	1.07
		8192	7.19e+04	1.29	1.13	1.08	1.04
		16384	7.43e+04	1.29	1.14	1.08	1.03
		32768	7.58e+04	1.31	1.14	1.07	1.02
	Qwen2.5-7B	512	1.46e+04	1.15	1.28	1.24	1.25
		1024	2.91e+04	1.21	1.16	1.15	1.11
		2048	3.65e+04	1.30	1.09	1.03	1.00
		4096	4.00e+04	1.26	1.05	1.00	0.96
		8192	3.70e+04	1.40	1.18	1.11	1.06
		16384	3.77e+04	1.41	1.18	1.12	1.08
		32768	3.77e+04	1.42	1.19	1.12	1.06
	Qwen2.5-14B	512	1.01e+04	1.19	1.14	1.16	1.16
		1024	1.68e+04	1.34	1.13	1.07	1.04
		2048	1.82e+04	1.38	1.19	1.12	1.08
		4096	1.92e+04	1.35	1.18	1.11	1.07
		8192	1.95e+04	1.45	1.20	1.11	1.07
		16384	1.99e+04	1.43	1.19	1.11	1.07
		32768	2.00e+04	1.43	1.18	1.11	1.05
B200	Llama3.2-1B	512	2.09e+04	1.00	1.06	0.97	1.05
		1024	4.14e+04	0.89	1.01	0.93	1.02
		2048	8.15e+04	1.01	1.04	0.95	0.97
		4096	1.73e+05	0.96	0.95	0.95	0.96
		8192	3.22e+05	0.93	0.95	0.96	0.96
		16384	4.71e+05	1.00	1.00	1.00	0.95
		32768	4.83e+05	0.99	0.99	0.99	0.98
	BitNet-2B	512	1.54e+04	1.08	1.10	1.10	1.12
		1024	3.20e+04	1.02	1.06	1.06	1.05
		2048	6.41e+04	1.05	1.06	1.12	1.06
		4096	1.25e+05	1.09	1.09	1.09	1.08
		8192	2.30e+05	1.13	1.07	1.07	1.05
		16384	2.69e+05	1.20	1.04	1.04	1.01
		32768	2.82e+05	1.22	1.05	1.04	1.00
	Llama3.2-3B	512	1.60e+04	1.05	1.10	1.09	1.04
		1024	3.21e+04	1.07	1.12	1.11	1.10
		2048	6.22e+04	1.12	1.13	1.11	1.12
		4096	1.28e+05	1.04	1.08	1.12	1.11
		8192	2.10e+05	1.21	1.08	1.03	0.99
		16384	2.26e+05	1.27	1.11	1.06	1.00
		32768	2.34e+05	1.29	1.12	1.07	1.00
	Qwen2.5-7B	512	1.70e+04	1.13	1.13	1.12	1.15
		1024	3.58e+04	1.11	1.11	1.14	1.13
		2048	7.22e+04	1.03	1.11	1.10	1.09
		4096	1.11e+05	1.27	1.10	1.06	1.02

Appendix C Prefill Int8 (Page 3/4)

GPU	Model	Batch Size	cuBLASLt	cuSPARSELt Speedup Ratio			
		M	Throughput (token/s)	2:4	2:6	2:8	2:10
		8192	1.21e+05	1.31	1.09	1.05	1.01
		16384	1.21e+05	1.38	1.14	1.08	1.07
		32768	1.26e+05	1.35	1.14	1.07	1.05
	Qwen2.5-14B	512	1.01e+04	1.10	1.15	1.21	1.14
		1024	2.13e+04	1.08	1.10	1.13	1.13
		2048	4.22e+04	1.14	1.15	1.10	1.08
		4096	6.17e+04	1.31	1.11	1.04	1.02
		8192	6.46e+04	1.33	1.10	1.08	1.03
		16384	6.40e+04	1.41	1.17	1.11	1.07
		32768	6.61e+04	1.41	1.15	1.11	1.06
RTX5080	Llama3.2-1B	512	3.43e+04	0.96	1.06	0.96	1.08
		1024	7.16e+04	0.92	0.97	0.95	0.96
		2048	8.09e+04	1.30	1.20	1.15	1.06
		4096	8.25e+04	1.40	1.22	1.16	1.07
		8192	8.49e+04	1.33	1.17	1.11	1.04
		16384	8.41e+04	1.31	1.16	1.10	1.03
		32768	8.39e+04	1.31	1.16	1.10	1.03
	BitNet-2B	512	2.68e+04	1.06	1.05	1.06	1.03
		1024	4.34e+04	1.26	1.08	1.05	1.04
		2048	4.61e+04	1.35	1.10	1.10	1.08
		4096	4.49e+04	1.37	1.14	1.12	1.09
		8192	4.35e+04	1.33	1.12	1.10	1.07
		16384	4.31e+04	1.31	1.11	1.09	1.06
		32768	4.30e+04	1.31	1.11	1.09	1.06
	Llama3.2-3B	512	2.16e+03	11.74	13.37	13.13	13.33
		1024	3.26e+04	1.29	1.21	1.13	1.10
		2048	3.67e+04	1.38	1.23	1.14	1.09
		4096	3.60e+04	1.37	1.20	1.11	1.06
		8192	3.54e+04	1.35	1.18	1.10	1.05
		16384	3.51e+04	1.34	1.17	1.10	1.05
		32768	3.51e+04	1.34	1.17	1.10	1.05
	Qwen2.5-7B	512	1.52e+04	1.32	1.22	1.14	1.08
		1024	1.76e+04	1.35	1.17	1.09	1.04
		2048	1.84e+04	1.40	1.15	1.08	1.03
		4096	1.84e+04	1.38	1.16	1.09	1.03
		8192	1.81e+04	1.38	1.17	1.10	
		16384	1.82e+04	1.38			
		32768					
	Qwen2.5-14B	512					
		1024					
		2048					
		4096					
		8192					
		16384					
		32768					
GB10	Llama3.2-1B	512	2.16e+04	1.35	1.18	1.20	1.10
		1024	2.42e+04	1.32	1.19	1.17	1.10
		2048	2.64e+04	1.31	1.17	1.13	1.05
		4096	2.60e+04	1.31	1.17	1.14	1.05
		8192	2.59e+04	1.28	1.16	1.12	1.04
		16384	2.62e+04	1.27	1.14	1.11	1.04
		32768	2.66e+04	1.23	1.14	1.10	1.04
	BitNet-2B	512	1.12e+04	1.39	0.96	1.26	1.18
		1024	1.32e+04	1.32	1.18	1.16	1.12
		2048	1.37e+04	1.38	1.19	1.16	1.13
		4096	1.30e+04	1.36	1.17	1.15	1.12
		8192	1.34e+04	1.29	1.11	1.10	1.09
		16384	1.36e+04	1.28	1.11	1.09	1.07
		32768	1.37e+04	1.28	1.12	1.09	1.08
	Llama3.2-3B	512	2.27e+03	6.12	5.26	5.19	4.71
		1024	1.14e+04	1.34	1.18	1.14	1.07
		2048	1.12e+04	1.41	1.25	1.19	1.12
		4096	1.03e+04	1.45	1.31	1.24	1.17
		8192	1.11e+04	1.32	1.21	1.13	1.07
		16384	1.13e+04	1.31	1.18	1.12	1.07

Appendix C Prefill Int8 (Page 4/4)

GPU	Model	Batch Size	cuBLASLt	cuSPARSELt Speedup Ratio			
		M	Throughput (token/s)	2:4	2:6	2:8	2:10
		32768	3.84e+03	2.98	3.43	3.26	3.14
	Qwen2.5-7B	512	4.98e+03	1.24	1.07	1.06	0.97
		1024	5.48e+03	1.38	1.19	1.11	1.05
		2048	5.57e+03	1.39	1.19	1.12	1.08
		4096	5.66e+03	1.37	1.14	1.08	1.05
		8192	2.28e+03	3.35	2.82	2.54	2.53
		16384	5.80e+03	0.23	1.11	1.06	0.13
		32768	5.84e+03	1.33	1.11	1.05	1.01
	Qwen2.5-14B	512	2.76e+03	1.39	1.19	1.10	1.04
		1024	3.10e+03	1.43	1.21	1.08	1.08
		2048	3.02e+03	1.45	1.18	1.10	1.05
		4096	3.17e+03	1.39	1.12	1.06	1.01
		8192	3.22e+03	1.35	1.11	1.05	1.00
		16384	3.22e+03	1.37	1.10	1.05	1.00
		32768	3.26e+03	0.34	1.09	1.03	0.99