Appendix C Prefill Fp8 (Page 1/3)

GPU	Model	Batch Size	cuBLASLt	cuSPARSELt Speedup Ratio			
		M	Throughput (token/s)	2:4	2:6	2:8	2:10
RTX4090	Llama3.2-1B	512	9.59e+03	0.91	0.88	0.87	0.88
		1024	1.93e+04	0.86	0.87	0.84	0.85
		2048	3.81e+04	0.88	0.87	0.86	0.86
		4096	7.14e+04	0.91	0.91	0.90	0.86
		8192	8.13e+04	1.28	1.15	1.10	1.05
		16384	7.99e+04	1.34	1.18	1.13	1.07
		32768	8.02e+04	1.36	1.19	1.14	1.08
	BitNet-2B	512	8.06e+03	1.05	1.03	1.06	1.02
		1024	1.60e+04	1.03	1.02	1.10	1.05
		2048	3.12e+04	1.07	1.07	1.01	1.05
		4096	4.61e+04	1.33	1.19	1.17	1.13
		8192	4.83e+04	1.31	1.14	1.09	1.09
		16384	4.65e+04	1.38	1.20	1.14	1.12
		32768	4.62e+04	1.38	1.22	1.14	1.13
	Llama3.2-3B	512	7.24e+03	1.02	1.04	1.05	0.98
		1024	1.47e+04	1.02	1.00	0.99	0.99
		2048	2.83e+04	1.06	1.06	1.07	1.06
		4096	3.32e+04	1.44	1.26	1.18	1.10
		8192	3.24e+04	1.44	1.25	1.18	1.12
		16384	3.27e+04	1.43	1.24	1.16	1.11
		32768	3.27e+04	1.43	1.24	1.16	1.11
	Qwen2.5-7B	512	8.30e+03	1.01	1.01	1.00	1.00
		1024	1.51e+04	1.10	1.11	1.10	1.07
		2048	1.59e+04	1.48	1.23	1.15	1.11
		4096	1.65e+04	1.50	1.25	1.17	1.12
		8192	1.63e+04	1.52	1.27	1.19	1.14
		16384	1.60e+04	1.55	1.30	0.77	0.61
		32768	1.62e+04	0.51	0.38	0.34	0.32
	Qwen2.5-14B	512	4.70e+03	1.00	1.03	1.04	
		1024	8.44e+03	1.17	1.18	1.13	
		2048	8.57e+03	1.55	1.29	1.20	
		4096	8.48e+03	1.58	1.31	1.22	
		8192	8.45e+03	1.57	1.29	1.22	
		16384	8.42e+03	1.57	1.30		
		32768					
H100	Llama3.2-1B	512	2.02e+04	0.81	0.83	0.77	0.81
		1024	3.92e+04	0.83	0.83	0.80	0.82
		2048	7.36e+04	0.84	0.88	0.84	0.84
		4096	1.35e+05	0.84	0.87	0.90	0.84
		8192	1.70e+05	1.00	0.96	0.90	0.89
		16384	1.90e+05	1.13	0.97	0.95	0.90
		32768	1.97e+05	1.14	0.98	0.97	0.90
	BitNet-2B	512	1.56e+04	1.03	1.06	1.04	1.07
		1024	3.18e+04	1.03	1.04	1.03	1.02
		2048	6.31e+04	0.99	1.03	1.01	1.08
		4096	8.74e+04	1.10	0.96	0.94	0.92
		8192	9.58e+04	1.12	0.99	0.96	0.92
		16384	9.95e+04	1.14	0.98	0.96	0.94
		32768	1.03e+05	1.15	0.99	0.96	0.93
	Llama3.2-3B	512	1.75e+04	0.99	0.97	0.95	0.99
		1024	3.49e+04	1.02	1.02	1.00	1.01
		2048	6.46e+04	1.04	1.03	0.97	0.92
		4096	7.48e+04	1.13	1.01	0.95	0.87
		8192	7.85e+04	1.17	1.01	0.96	0.91
		16384	7.95e+04	1.20	1.04	0.98	0.92
		32768	8.08e+04	1.20	1.04	0.97	0.92
	Qwen2.5-7B	512	1.84e+04	1.02	0.98	1.01	1.03
		1024	3.38e+04	1.12	0.99	0.94	0.92
		2048	3.64e+04	1.21	1.05	1.03	0.96
		4096	3.91e+04	1.24	1.05	0.99	0.93
		8192	3.96e+04	1.27	1.06	0.99	0.94
		16384	3.98e+04	1.30	1.08	1.02	0.97
		32768	4.07e+04	1.26	1.04	0.98	0.94
	Qwen2.5-14B	512	1.15e+04	1.05	0.99	1.06	0.99
		1024	1.85e+04	1.18	1.00	0.94	0.92

≥1.5× ≥1.0×

Appendix C Prefill Fp8 (Page 2/3)

GPU	Model	Batch Size	cuBLASLt	cuSPARSELt Speedup Ratio			
		M	Throughput (token/s)	2:4	2:6	2:8	2:10
		2048	1.97e+04	1.24	1.06	0.98	0.96
		4096	2.03e+04	1.30	1.07	1.00	0.97
		8192	2.08e+04	1.30	1.08	1.01	0.96
		16384	2.09e+04	1.31	1.07	1.00	0.96
		32768	2.11e+04	1.30	1.07	1.00	0.95
B200	Llama3.2-1B	512	2.47e+04	0.90	0.89	0.85	0.88
		1024	4.83e+04	0.87	0.88	0.87	0.88
		2048	8.66e+04	1.01	0.92	0.98	0.98
		4096	1.83e+05	0.90	0.91	0.89	0.92
		8192	3.42e+05	0.90	0.92	0.92	0.87
		16384	4.68e+05	0.99	0.98	1.00	0.98
		32768	4.72e+05	1.01	1.00	1.00	1.01
	BitNet-2B	512	1.67e+04	1.01	1.00	1.01	1.01
		1024	3.40e+04	0.99	0.99	0.98	0.98
		2048	6.68e+04	1.04	1.05	1.01	0.98
		4096	1.31e+05	1.03	1.03	1.02	1.03
		8192	2.42e+05	1.04	1.00	0.97	0.98
		16384	2.92e+05	1.10	0.95	0.93	0.91
		32768	2.97e+05	1.13	0.98	0.96	0.93
	Llama3.2-3B	512	1.78e+04	1.01	0.99	0.98	0.97
		1024	3.62e+04	1.02	0.96	0.98	1.02
		2048	6.82e+04	1.01	1.01	1.01	1.02
		4096	1.37e+05	1.01	1.02	1.03	1.03
		8192	2.25e+05	1.10	0.98	0.95	0.88
		16384	2.38e+05	1.15	1.02	0.97	0.91
		32768	2.44e+05	1.20	1.04	0.99	0.92
	Qwen2.5-7B	512	1.85e+04	1.06	1.08	1.05	1.05
		1024	3.86e+04	1.06	1.06	1.06	1.05
		2048	7.12e+04	1.05	1.08	1.07	1.10
		4096	1.17e+05	1.19	1.01	0.98	0.93
		8192	1.23e+05	1.23	1.02	1.00	0.96
		16384	1.29e+05	1.25	1.03	0.99	0.94
		32768	1.31e+05	1.26	1.04	0.99	0.95
	Qwen2.5-14B	512	1.12e+04	1.06	1.02	1.05	1.05
		1024	2.31e+04	1.04	1.02	1.05	1.04
		2048	4.63e+04	1.07	1.00	1.01	1.05
		4096	6.26e+04	1.23	1.02	0.99	0.93
		8192	6.59e+04	1.21	1.05	1.00	0.96
		16384	6.78e+04	1.27	0.93	0.99	0.95
		32768	6.89e+04	1.28	1.03	1.00	0.97
RTX5080	Llama3.2-1B	512	3.77e+04	0.97	0.99	0.93	1.00
		1024	6.82e+04	1.03	1.03	1.00	0.95
		2048	7.10e+04	1.33	1.14	1.09	1.04
		4096	7.26e+04	1.33	1.15	1.08	1.04
		8192	7.23e+04	1.33	1.14	1.07	1.04
		16384	7.14e+04	1.32	1.14	1.07	1.03
		32768	7.10e+04	1.32	1.14	1.07	1.03
	BitNet-2B	512	2.62e+04	1.10	1.10	1.11	1.08
		1024	3.62e+04	1.27	1.08	1.02	0.95
		2048	3.67e+04	1.35	1.17	1.10	1.04
		4096	3.67e+04	1.38	1.17	1.10	1.06
		8192	3.53e+04	1.36	1.16	1.09	1.05
		16384	3.51e+04	1.35	1.14	1.08	1.04
		32768	3.50e+04	1.34	1.15	1.08	1.04
	Llama3.2-3B	512	2.47e+04	1.19	1.12	1.05	1.01
		1024	2.86e+04	1.31	1.13	1.05	0.99
		2048	2.84e+04	1.47	1.23	1.13	1.08
		4096	2.85e+04	1.41	1.20	1.10	1.05
		8192	2.80e+04	1.40	1.19	1.10	1.04
		16384	2.79e+04	1.39	1.18	1.10	1.04
		32768	2.78e+04	1.39	1.18	1.10	1.04
	Qwen2.5-7B	512	1.30e+04	1.42	1.18	1.10	1.04
		1024	1.35e+04	1.48	1.19	1.12	1.06
		2048	1.34e+04	1.52	1.24	1.14	1.08
		4096	1.36e+04	1.47	1.23	1.12	1.06

Appendix C Prefill Fp8 (Page 3/3)

GPU	Model	Batch Size	cuBLASLt	cuSPARSELt Speedup Ratio			
		M	Throughput (token/s)	2:4	2:6	2:8	2:10
		8192	1.35e+04	1.48	1.22	1.12	1.06
		16384	1.35e+04	1.47	1.22	1.12	1.06
		32768					
	Qwen2.5-14B	512					
		1024					
		2048					
		4096					
		8192					
		16384					
		32768					
GB10	Llama3.2-1B	512	2.70e+04	1.10	0.98	0.92	0.92
		1024	3.07e+04	1.04	0.93	0.90	0.86
		2048	3.23e+04	1.06	0.94	0.88	0.86
		4096	3.19e+04	1.06	0.93	0.91	0.85
		8192	3.13e+04	1.05	0.94	0.91	0.87
		16384	3.13e+04	1.06	0.97	0.93	0.89
		32768	3.19e+04	1.06	0.97	0.92	0.89
	BitNet-2B	512	1.40e+04	1.08	1.02	1.01	0.95
		1024	1.61e+04	1.07	0.96	0.93	0.91
		2048	1.71e+04	1.07	0.93	0.89	0.86
		4096	1.66e+04	1.03	0.90	0.87	0.85
		8192	1.61e+04	1.07	0.93	0.90	0.87
		16384	1.63e+04	1.08	0.94	0.91	0.87
		32768	1.64e+04	1.05	0.93	0.88	0.88
	Llama3.2-3B	512	1.20e+04	1.16	1.01	1.00	0.96
		1024	1.39e+04	1.10	0.96	0.91	0.86
		2048	1.39e+04	1.08	0.97	0.91	0.86
		4096	1.36e+04	1.06	0.95	0.89	0.85
		8192	1.35e+04	1.08	0.96	0.88	0.87
		16384	1.36e+04	1.08	0.96	0.90	0.87
		32768	1.35e+04	1.08	0.96	0.87	0.88
	Qwen2.5-7B	512	5.83e+03	1.09	1.02	0.97	0.90
		1024	6.78e+03	1.12	0.91	0.88	0.85
		2048	6.91e+03	1.10	0.90	0.86	0.83
		4096	7.06e+03	1.04	0.86	0.83	0.79
		8192	6.79e+03	1.10	0.91	0.88	0.85
		16384	6.80e+03	1.10	0.92	0.88	0.85
		32768	6.81e+03	1.12	0.93	0.88	0.85
	Qwen2.5-14B	512	3.17e+03	1.29	1.07	1.03	0.97
		1024	3.75e+03	1.15	0.92	0.86	0.83
		2048	3.66e+03	1.14	0.92	0.88	0.84
		4096	3.73e+03	1.15	0.94	0.87	0.83
		8192	3.71e+03	1.13	0.94	0.88	0.84
		16384	1.10e+03	1.03	3.18	2.98	2.79
		32768	3.70e+03	1.12	0.93	0.85	0.82