Appendix B Model Kernel Int8 (Page 1/5)

GPU	Model	M	cuBLASLt Latency (μs)	cuSPARSELt Speedup Ratio							
				2:4	2:6	2:8	2:10	2:12	2:14	2:16	2:∞
A100	Llama3.2-1B	64	8.14e+01	1.16	0.84	0.81	0.70	0.63	0.65	0.71	0.65
		128	1.00e+02	1.40	1.06	1.03	0.89	0.81	0.82	0.90	0.82
		256	1.24e+02	1.37	1.08	0.98	0.88	0.81	0.85	0.87	0.79
		512	1.99e+02	1.52	1.19	1.13	0.97	0.92	0.96	1.01	0.92
		1024	3.83e+02	1.87	1.44	1.35	1.14	1.07	1.13	1.19	1.06
		2048	7.13e+02	1.79	1.34	1.24	1.09	1.01	1.09	1.11	0.96
		4096	1.38e+03	1.88	1.40	1.31	1.10	1.04	1.09	1.14	1.02
		8192	2.84e+03	2.00	1.47	1.40	1.16	1.10	1.15	1.21	1.06
		16384	5.70e+03	1.96	1.46	1.39	1.19	1.10	1.15	1.20	1.07
	BitNet-2B	64	8.50e+01	1.27	0.89	0.88	0.78	0.77	0.69	0.72	0.69
		128	9.53e+01	1.34	0.99	0.96	0.90	0.87	0.79	0.84	0.77
		256	1.29e+02	1.57	1.14	1.12	1.08	1.02	0.95	1.00	0.90
		512	2.26e+02	1.74	1.28	1.28	1.22	1.15	1.09	1.14	1.02
		1024	4.17e+02	1.90	1.39	1.39	1.27	1.24	1.17	1.19	1.08
		2048	7.46e+02	1.82	1.26	1.27	1.20	1.14	1.05	1.11	0.97
		4096	1.50e+03	1.98	1.38	1.38	1.33	1.25	1.14	1.20	1.06
		8192	3.12e+03	2.00	1.41	1.44	1.33	1.23	1.16	1.21	1.07
		16384	6.25e+03	1.99	1.41	1.38	1.30	1.21	1.14	1.19	1.05
	Llama3.2-3B	64	1.11e+02	1.27	0.95	0.89	0.81	0.76	0.68	0.80	0.69
		128	1.35e+02	1.55	1.18	1.07	1.01	0.95	0.87	0.94	0.86
		256	1.79e+02	1.45	1.15	1.05	1.02	0.91	0.83	0.92	0.82
		512	3.02e+02	1.73	1.35	1.24	1.16	1.08	0.98	1.08	0.95
		1024	5.79e+02	2.02	1.57	1.43	1.29	1.25	1.11	1.26	1.11
		2048	1.13e+03	2.03	1.58	1.41	1.28	1.22	1.08	1.24	1.07
		4096	2.28e+03	2.04	1.56	1.40	1.27	1.22	1.05	1.21	1.06
		8192	4.70e+03	2.02	1.55	1.42	1.27	1.24	1.08	1.22	1.06
		16384	9.39e+03	2.03	1.56	1.41	1.28	1.24	1.10	1.22	1.07
	Qwen2.5-7B	64	2.38e+02	1.43	1.07	1.00	0.95	0.83	0.84		
		128	2.51e+02	1.45	1.12	1.04	0.98	0.87	0.89		
		256	4.01e+02	1.82	1.37	1.29	1.22	1.02	1.08		
		512	7.49e+02	1.95	1.45	1.35	1.27	1.08	1.14		
		1024	1.36e+03	1.85	1.40	1.27	1.18	1.02	1.06		
		2048	2.69e+03	2.08	1.43	1.34	1.25	1.06	1.12		
		4096	5.39e+03	2.03	1.49	1.38	1.29	1.10	1.16		
		8192	1.07e+04	2.06	1.51	1.41	1.31	1.12	1.17		
		16384	2.14e+04	2.07	1.52	1.42	1.32	1.12	1.18		
	Qwen2.5-14B	64	2.53e+02	1.53	1.15	1.05	0.99	0.91	0.88	0.90	0.79
		128	2.74e+02	1.59	1.23	1.12	1.05	0.97	0.93	0.97	0.86
		256	4.48e+02	1.91	1.46	1.33	1.28	1.11	1.05	1.17	1.03
		512	8.59e+02	2.16	1.63	1.50	1.39	1.21	1.14	1.28	1.13
		1024	1.54e+03	1.99	1.47	1.35	1.25	1.10	1.02	1.16	1.02
		2048	3.13e+03	2.18	1.57	1.45	1.35	1.18	1.08	1.24	1.08
		4096	6.31e+03	2.12	1.58	1.43	1.33	1.16	1.07	1.23	1.08
		8192	1.26e+04	2.11	1.55	1.41	1.33	1.16	1.07	1.22	1.06
		16384	2.51e+04	2.08	1.53	1.41	1.31	1.15	1.06	1.21	1.05
RTX4090	Llama3.2-1B	64	7.71e+01	1.38	1.16	1.12	1.04	0.89	0.99	0.91	0.86
		128	8.53e+01	1.41	1.15	1.12	1.05	0.94	0.94	0.83	0.78
		256	1.08e+02	1.59	1.31	1.18	1.01	0.50	0.79	1.07	0.54
		512	1.57e+02	1.07	1.14	1.05	0.89	0.41	0.50	0.49	0.41
		1024	2.91e+02	1.28	1.30	1.19	0.98	0.44	0.41	0.67	0.43
		2048	5.65e+02	1.67	1.43	1.28	0.91	0.74	0.41	0.77	0.37
		4096	9.22e+02	1.41	1.21	1.08	0.82	0.75	0.69	0.46	0.66
		8192	1.85e+03	1.60	1.23	1.08	0.91	0.81	0.77	0.80	0.70
		16384	3.73e+03	1.60	1.22	1.08	0.90	0.83	0.90	0.92	0.81
	BitNet-2B	64	7.16e+01	1.30	0.87	1.11	0.81	1.02	0.84	1.00	0.87
		128	8.89e+01	1.21	0.94	1.17	0.85	1.06	0.98	1.03	0.90
		256	1.18e+02	1.45	1.11	1.17	1.07	1.06	0.93	1.00	0.93
		512	2.01e+02	1.64	1.19	1.20	1.11	1.09	1.02	1.05	0.92
		1024	3.70e+02	1.91	1.33	1.37	1.27	1.22	1.13	1.17	1.04
		2048	6.13e+02	1.63	1.11	1.12	1.04	1.01	0.93	0.95	0.84
		4096	1.10e+03	1.59	1.08	1.10	1.02	0.99	0.90	0.93	0.83
		8192	2.06e+03	1.56	1.04	1.07	0.99	0.97	0.89	0.91	0.82
		16384	4.20e+03	1.58	1.07	1.08	1.01	0.99	0.91	0.93	0.83
	Llama3.2-3B	64	9.24e+01	1.37	1.24	1.19	1.12	0.45	0.86	0.40	0.93
		128	1.06e+02	1.35	1.24	1.12	1.06	0.36	0.85	0.30	0.59

≥1.5× ≥1.0×

Appendix B Model Kernel Int8 (Page 2/5)

GPU	Model	M	cuBLASLt Latency (μs)	cuSPARSELt Speedup Ratio							
				2:4	2:6	2:8	2:10	2:12	2:14	2:16	2:∞
		256	1.70e+02	1.41	1.16	1.30	1.03	0.29	0.41	0.24	0.37
		512	2.69e+02	1.38	1.07	1.27	0.96	0.28	0.32	0.27	0.23
		1024	4.74e+02	1.41	1.14	1.06	1.21	0.75	0.44	0.35	0.36
		2048	8.97e+02	1.72	1.16	1.06	1.11	0.72	0.65	0.45	0.35
		4096	1.49e+03	1.58	1.19	1.06	1.02	0.62	0.57	0.86	0.66
		8192	3.01e+03	1.58	1.20	1.08	1.02	0.78	0.73	0.94	0.79
		16384	6.13e+03	1.62	1.22	1.09	1.03	0.92	0.78	0.94	0.86
	Qwen2.5-7B	64	2.56e+02	1.61	1.18	1.11	1.05	0.08	0.28	0.22	0.15
		128	2.60e+02	1.50	0.92	1.00	0.99	0.32	0.61	0.49	0.39
		256	3.54e+02	1.53	1.08	1.05	1.01	0.53	0.32	0.58	0.29
		512	5.91e+02	1.73	1.24	1.16	1.09	0.39	0.54	0.36	0.36
		1024	1.14e+03	1.91	1.45	1.30	1.22	0.45	0.52	0.39	0.54
		2048	1.83e+03	1.60	1.19	1.08	1.01	0.47	0.59	0.66	0.55
		4096	3.57e+03	1.59	1.18	1.09	1.01	0.70	0.82	0.73	0.60
		8192	6.92e+03	1.58	1.19	1.07	1.00	0.77	0.90	0.92	0.80
		16384	1.40e+04	1.60	1.21	1.07	1.02	0.80	0.91	0.92	0.83
	Qwen2.5-14B	64	2.59e+02	1.61	1.17	1.09	1.05	0.22	0.18	0.23	0.45
		128	2.81e+02	1.50	0.98	0.88	0.97	0.28	0.26	0.35	0.12
		256	4.62e+02	1.90	1.05	1.14	1.19	0.28	0.33	0.27	0.44
		512	7.37e+02	1.99	1.11	1.34	1.26	0.24	0.30	0.22	0.25
		1024	1.27e+03	1.74	1.15	1.17	0.98	0.56	0.70	0.81	0.71
		2048	2.18e+03	1.66	1.12	1.08	1.00	0.64	0.55	0.59	0.58
		4096	4.01e+03	1.52	1.08	1.07	0.98	0.69	0.62	0.81	0.72
		8192	8.03e+03	1.57	1.20	1.08	1.02	0.81	0.74	0.94	0.80
		16384	1.63e+04	1.65	1.22	1.10	1.03	0.81	0.77	0.95	0.83
H100	Llama3.2-1B	64	6.67e+01	1.67	1.44	1.37	1.31	1.25	1.25	1.30	1.07
		128	7.43e+01	1.69	1.40	1.37	1.22	1.16	1.18	1.25	1.03
		256	8.15e+01	1.42	1.18	1.13	1.01	0.94	0.97	1.01	0.90
		512	1.25e+02	1.55	1.20	1.16	1.02	1.01	0.99	1.06	0.95
		1024	1.60e+02	1.26	1.02	1.01	0.88	0.81	0.83	0.84	0.79
		2048	3.12e+02	1.48	1.07	1.01	0.91	0.90	0.90	0.89	0.77
		4096	5.47e+02	1.30	1.03	0.96	0.88	0.84	0.83	0.82	0.75
		8192	1.02e+03	1.36	1.00	0.95	0.84	0.76	0.80	0.80	0.70
		16384	2.09e+03	1.34	0.99	0.92	0.82	0.76	0.78	0.81	0.71
	BitNet-2B	64	6.28e+01	1.63	1.35	1.37	1.33	1.26	1.20	1.24	0.98
		128	7.15e+01	1.70	1.34	1.42	1.35	1.27	1.19	1.22	0.99
		256	7.90e+01	1.47	1.19	1.18	1.09	1.04	0.94	0.97	0.86
		512	1.25e+02	1.59	1.26	1.20	1.13	1.10	1.06	1.05	0.94
		1024	1.78e+02	1.32	1.05	1.01	0.93	0.88	0.85	0.84	0.79
		2048	3.27e+02	1.40	1.04	0.96	0.94	0.91	0.83	0.85	0.78
		4096	5.76e+02	1.31	1.02	0.92	0.88	0.84	0.78	0.80	0.72
		8192	1.13e+03	1.33	0.96	0.93	0.86	0.81	0.78	0.81	0.71
		16384	2.31e+03	1.29	0.94	0.90	0.83	0.79	0.75	0.77	0.67
	Llama3.2-3B	64	9.00e+01	1.96	1.54	1.29	1.22	1.19	1.15	1.18	1.09
		128	9.61e+01	1.85	1.42	1.27	1.18	1.17	1.08	1.15	1.03
		256	1.25e+02	1.76	1.37	1.24	1.15	1.13	1.04	1.11	1.02
		512	1.70e+02	1.62	1.24	1.18	1.06	1.03	0.97	1.04	0.91
		1024	2.43e+02	1.39	1.16	0.99	0.95	0.93	0.88	0.88	0.80
		2048	5.00e+02	1.59	1.25	1.16	1.05	1.03	0.94	0.99	0.85
		4096	8.84e+02	1.43	1.11	1.00	0.93	0.89	0.83	0.86	0.76
		8192	1.75e+03	1.50	1.14	1.01	0.90	0.87	0.77	0.86	0.74
		16384	3.67e+03	1.46	1.14	1.00	0.93	0.87	0.80	0.82	0.75
	Qwen2.5-7B	64	2.98e+02	2.94	2.26	2.10	1.96	1.90	1.90	1.84	1.66
		128	3.06e+02	2.70	2.08	1.98	1.85	1.65	1.72	1.74	1.55
		256	2.51e+02	1.80	1.34	1.31	1.23	1.12	1.14	1.11	1.02
		512	3.33e+02	1.52	1.18	1.14	1.06	0.96	0.98	0.94	0.84
		1024	5.43e+02	1.55	1.15	1.06	1.01	0.92	0.93	0.94	0.83
		2048	1.06e+03	1.53	1.16	1.07	1.00	0.89	0.92	0.91	0.84
		4096	2.38e+03	1.70	1.28	1.15	1.06	0.93	0.95	0.94	0.82
		8192	4.59e+03	1.59	1.15	1.08	0.99	0.87	0.87	0.90	0.81
		16384	9.16e+03	1.54	1.13	1.06	1.02	0.88	0.88	0.91	0.81
	Qwen2.5-14B	64	1.96e+02	1.85	1.41	1.26	1.26	1.14	1.10	1.14	0.97
		128	2.65e+02	2.28	1.69	1.62	1.54	1.36	1.28	1.39	1.21
		256	2.91e+02	2.03	1.50	1.44	1.36	1.24	1.18	1.24	1.11
		512	3.44e+02	1.55	1.18	1.09	1.02	0.92	0.89	0.93	0.82

Appendix B Model Kernel Int8 (Page 3/5)

GPU	Model	M	cuBLASLt Latency (μs)	cuSPARSELt Speedup Ratio							
				2:4	2:6	2:8	2:10	2:12	2:14	2:16	2:∞
		1024	6.80e+02	1.62	1.25	1.14	1.07	0.99	0.94	1.00	0.89
		2048	1.25e+03	1.65	1.24	1.03	0.92	0.92	0.89	0.97	0.81
		4096	2.62e+03	1.66	1.20	1.05	1.04	0.92	0.88	0.90	0.79
		8192	5.41e+03	1.66	1.23	1.07	1.03	0.95	0.87	0.94	0.80
		16384	1.35e+04	2.01	1.51	1.36	1.26	1.13	1.05	1.13	1.01
B200	Llama3.2-1B	64	5.78e+01	1.86	1.64	1.66	1.51	1.40	1.47	1.55	1.47
		128	6.39e+01	2.05	1.71	1.77	1.63	1.55	1.55	1.62	1.55
		256	7.51e+01	2.26	1.83	1.82	1.74	1.66	1.73	1.74	1.69
		512	9.99e+01	2.50	2.20	2.17	1.90	1.75	1.92	2.01	1.89
		1024	1.71e+02	3.43	2.77	2.75	2.44	2.38	2.42	2.57	2.44
		2048	3.18e+02	4.54	3.56	3.56	3.06	2.88	2.98	3.18	2.91
		4096	5.83e+02	5.08	3.83	3.75	3.16	2.91	3.00	3.29	2.99
		8192	1.14e+03	5.57	4.15	3.96	3.39	3.18	3.26	3.50	3.12
		16384	2.25e+03	5.97	4.23	4.22	3.43	3.27	3.35	3.57	3.25
	BitNet-2B	64	6.09e+01	2.04	1.67	1.69	1.64	1.64	1.55	1.64	1.62
		128	6.80e+01	2.21	1.83	1.80	1.74	1.73	1.65	1.73	1.73
		256	7.57e+01	2.07	1.91	1.84	1.83	1.75	1.67	1.76	1.69
		512	1.25e+02	2.99	2.37	2.51	2.39	2.25	2.17	2.30	2.17
		1024	2.07e+02	3.64	2.88	2.99	2.95	2.74	2.64	2.73	2.57
		2048	3.57e+02	4.48	3.31	3.42	3.27	3.02	2.73	3.03	2.80
		4096	6.56e+02	4.99	3.78	3.74	3.56	3.27	3.06	3.27	2.91
		8192	1.29e+03	5.79	4.06	4.08	3.86	3.51	3.31	3.51	3.17
		16384	2.53e+03	5.87	4.15	4.20	3.95	3.59	3.31	3.61	3.20
	Llama3.2-3B	64	6.60e+01	2.00	1.77	1.75	1.68	1.60	1.47	1.68	1.60
		128	7.58e+01	2.18	1.93	1.93	1.83	1.83	1.63	1.88	1.70
		256	1.01e+02	2.57	2.33	2.22	2.12	2.08	1.88	2.13	1.96
		512	1.62e+02	3.33	2.92	2.78	2.54	2.54	2.25	2.55	2.39
		1024	2.62e+02	3.78	3.34	3.01	2.83	2.83	2.59	2.78	2.54
		2048	4.97e+02	4.98	4.02	3.66	3.32	3.27	2.92	3.25	2.92
		4096	9.41e+02	5.72	4.47	4.06	3.67	3.66	3.16	3.54	3.16
		8192	1.82e+03	6.02	4.53	4.08	3.56	3.55	3.15	3.59	3.16
		16384	3.64e+03	6.15	4.75	4.28	3.72	3.75	3.27	3.62	3.22
	Qwen2.5-7B	64	1.17e+02	2.58	2.10	2.00	1.89	1.68	1.78	1.78	1.63
		128	1.52e+02	3.21	2.53	2.39	2.35	2.04	2.15	2.15	1.95
		256	2.01e+02	3.84	3.04	2.85	2.67	2.35	2.44	2.56	2.28
		512	3.27e+02	4.52	3.52	3.45	3.25	2.70	2.93	2.99	2.66
		1024	5.94e+02	5.33	4.09	3.90	3.69	3.14	3.43	3.41	3.09
		2048	1.10e+03	5.55	4.04	3.98	3.71	2.90	3.25	3.46	3.03
		4096	2.13e+03	5.81	4.31	4.09	3.87	3.19	3.32	3.57	3.18
		8192	4.21e+03	6.17	4.38	4.18	3.99	3.30	3.49	3.51	3.18
		16384	8.48e+03	6.25	4.35	4.32	4.07	3.24	3.47	3.74	3.18
	Qwen2.5-14B	64	1.17e+02	2.69	1.90	1.82	1.72	1.57	1.49	1.65	1.53
		128	1.33e+02	2.83	2.08	2.03	1.94	1.72	1.61	1.83	1.62
		256	2.32e+02	3.88	3.04	2.94	2.81	2.42	2.30	2.57	2.28
		512	3.82e+02	4.53	3.57	3.46	3.27	2.87	2.70	3.00	2.69
		1024	6.72e+02	5.10	3.73	3.62	3.43	2.91	2.67	3.17	2.83
		2048	1.25e+03	5.49	4.14	3.90	3.71	3.20	2.99	3.40	3.00
		4096	2.46e+03	6.19	4.41	4.06	3.89	3.35	3.15	3.55	3.14
		8192	4.86e+03	6.35	4.47	4.21	3.92	3.46	3.19	3.51	3.12
		16384	9.89e+03	6.38	4.43	4.19	3.95	3.46	3.22	3.59	3.16
RTX5080	Llama3.2-1B	64	6.94e+01	1.41	1.17	1.02	0.91	0.90	0.94	0.94	0.85
		128	8.52e+01	1.54	1.26	1.12	1.04	0.98	1.03	1.02	0.92
		256	1.17e+02	1.55	1.19	1.08	0.97	0.86	0.91	0.97	0.89
		512	2.13e+02	1.71	1.35	1.22	1.06	0.99	1.03	1.07	0.97
		1024	3.88e+02	1.68	1.31	1.16	1.02	0.94	0.99	1.03	0.92
		2048	7.07e+02	1.62	1.24	1.11	0.97	0.90	0.94	0.96	0.86
		4096	1.34e+03	1.62	1.24	1.10	0.97	0.89	0.93	0.96	0.85
		8192	2.58e+03	1.57	1.19	1.08	0.95	0.88	0.92	0.95	0.84
		16384	5.10e+03	1.55	1.17	1.07	0.94	0.87	0.91	0.93	0.83
	BitNet-2B	64	7.19e+01	1.46	1.07	1.06	1.00	0.95	0.92	0.92	0.81
		128	8.21e+01	1.43	1.00	1.03	0.98	0.93	0.89	0.90	0.80
		256	1.27e+02	1.59	1.11	1.09	1.10	1.01	0.94	0.97	0.90
		512	2.12e+02	1.54	1.11	1.05	1.04	0.97	0.94	0.95	0.85
		1024	4.15e+02	1.70	1.21	1.17	1.13	1.07	1.02	1.03	0.91
		2048	7.61e+02	1.62	1.14	1.10	1.06	1.00	0.95	0.97	0.86

Appendix B Model Kernel Int8 (Page 4/5)

GPU	Model	M	cuBLASLt Latency (μs)	cuSPARSELt Speedup Ratio							
				2:4	2:6	2:8	2:10	2:12	2:14	2:16	2:∞
		4096	1.37e+03	1.49	1.05	1.01	0.96	0.90	0.87	0.87	0.78
		8192	2.78e+03	1.51	1.06	1.03	0.97	0.92	0.88	0.88	0.78
		16384	5.70e+03	1.54	1.09	1.04	1.00	0.94	0.89	0.91	0.80
	Llama3.2-3B	64	8.20e+01	1.38	1.14	1.05	1.00	0.92	0.81	0.90	0.80
		128	1.07e+02	1.53	1.24	1.13	1.05	1.00	0.86	0.97	0.86
		256	1.96e+02	1.71	1.40	1.24	1.15	1.12	0.99	1.07	0.95
		512	3.33e+02	1.61	1.25	1.12	1.05	0.99	0.86	0.96	0.84
		1024	5.88e+02	1.63	1.27	1.12	1.04	0.97	0.85	0.96	0.85
		2048	1.10e+03	1.65	1.27	1.13	1.05	0.98	0.87	0.97	0.86
		4096	2.01e+03	1.51	1.17	1.03	0.98	0.91	0.80	0.90	0.79
		8192	4.07e+03	1.54	1.19	1.05	0.99	0.92	0.82	0.91	0.80
		16384	8.20e+03	1.56	1.21	1.06	1.00	0.93	0.83	0.92	0.81
	Qwen2.5-7B	64	2.54e+02	1.58	1.21	1.07	0.97	0.95	0.85	0.84	0.73
		128	2.74e+02	1.64	1.25	1.10	0.93	0.90	0.87	0.87	0.75
		256	3.99e+02	1.71	1.33	1.18	1.08	0.99	1.02	1.01	0.88
		512	6.82e+02	1.54	1.16	1.03	0.96	0.83	0.87	0.89	0.79
		1024	1.29e+03	1.64	1.24	1.12	1.03	0.90	0.95	0.96	0.85
		2048	2.38e+03	1.53	1.16	1.04	0.95	0.83	0.88	0.90	0.79
		4096	4.70e+03	1.51	1.16	1.03	0.95	0.83	0.88	0.90	0.79
		8192	9.27e+03	1.49	1.14	1.02	0.95	0.82	0.87	0.89	0.78
		16384	1.88e+04	1.50	1.16	1.02	0.97	0.84	0.89	0.90	0.80
	Qwen2.5-14B	64	2.62e+02	1.62	1.24	1.11	0.91	0.84	0.82	0.81	0.71
		128	2.98e+02	1.63	1.21	1.09	0.95	0.88	0.85	0.85	0.75
		256	4.28e+02	1.63	1.24	1.09	1.05	0.90	0.87	0.94	0.83
		512	7.95e+02	1.72	1.28	1.15	1.10	0.95	0.90	1.01	0.88
		1024	1.49e+03	1.66	1.25	1.11	1.06	0.93	0.87	0.97	0.86
		2048	2.69e+03	1.51	1.13	1.00	0.96	0.83	0.79	0.88	0.77
		4096	5.36e+03	1.50	1.12	0.99	0.94	0.82	0.79	0.88	0.77
		8192	1.09e+04	1.52	1.15	1.02	0.97	0.84	0.81	0.90	0.79
		16384	2.18e+04	1.52	1.14	1.02	0.97	0.85	0.81	0.90	0.79
GB10	Llama3.2-1B	64	2.14e+02	2.39	1.31	1.23	0.97	0.88	0.96	1.00	0.82
		128	2.30e+02	2.22	1.37	1.29	0.83	0.82	0.90	1.00	0.83
		256	3.38e+02	2.29	1.52	1.54	1.03	1.00	1.04	1.09	1.06
		512	4.63e+02	1.67	1.18	1.17	0.95	0.91	0.86	0.93	0.85
		1024	8.19e+02	1.54	1.23	1.08	0.96	0.89	0.95	0.99	0.85
		2048	1.40e+03	1.39	1.06	0.97	0.81	0.76	0.80	0.76	0.67
		4096	2.79e+03	1.44	1.03	0.93	0.80	0.76	0.79	0.82	0.73
		8192	5.43e+03	1.37	1.04	0.92	0.80	0.73	0.76	0.80	0.69
		16384	1.10e+04	1.36	1.04	0.96	0.81	0.76	0.78	0.81	0.69
	BitNet-2B	64	2.19e+02	2.81	1.13	1.20	1.06	0.75	0.71	0.92	0.84
		128	2.47e+02	2.26	1.21	1.26	1.09	0.90	0.74	0.97	0.76
		256	3.48e+02	2.13	1.28	1.21	1.14	0.99	0.94	0.95	0.83
		512	5.17e+02	1.77	1.11	1.15	1.11	0.96	0.93	0.97	0.83
		1024	8.53e+02	1.41	1.07	1.02	0.96	0.95	0.89	0.87	0.76
		2048	1.64e+03	1.48	1.06	1.02	0.96	0.92	0.87	0.87	0.77
		4096	3.22e+03	1.45	1.01	1.00	0.93	0.87	0.79	0.81	0.73
		8192	6.24e+03	1.39	0.99	0.96	0.91	0.86	0.81	0.82	0.73
		16384	1.24e+04	1.38	0.98	0.96	0.90	0.84	0.79	0.81	0.72
	Llama3.2-3B	64	3.62e+02	2.41	1.30	1.22	0.91	0.96	0.75	0.99	0.76
		128	3.90e+02	2.04	1.22	1.09	0.88	0.92	0.78	0.94	0.72
		256	5.11e+02	1.93	1.34	1.18	0.95	1.01	0.80	1.03	0.81
		512	7.60e+02	1.81	1.36	1.20	1.02	0.97	0.92	1.06	0.84
		1024	1.28e+03	1.63	1.23	1.06	1.03	0.97	0.87	0.94	0.81
		2048	2.40e+03	1.57	1.20	1.07	0.98	0.91	0.79	0.90	0.74
		4096	4.65e+03	1.50	1.13	1.02	0.91	0.85	0.78	0.86	0.75
		8192	9.08e+03	1.46	1.11	1.00	0.92	0.87	0.76	0.84	0.73
		16384	1.78e+04	1.41	1.09	0.98	0.88	0.84	0.75	0.83	0.72
	Qwen2.5-7B	64	1.03e+03	1.58	0.90	1.01	0.93	0.71	0.82	0.87	0.77
		128	1.22e+03	1.76	1.04	1.15	1.03	0.80	0.90	0.99	0.86
		256	1.46e+03	1.69	1.11	1.16	1.07	0.84	0.93	0.96	0.89
		512	1.90e+03	1.74	1.13	1.13	1.04	0.88	0.95	0.99	0.86
		1024	3.01e+03	1.56	1.07	1.00	0.91	0.79	0.82	0.84	0.76
		2048	5.67e+03	1.43	0.98	0.91	0.88	0.74	0.74	0.79	0.70
		4096	1.10e+04	1.42	1.00	0.93	0.85	0.72	0.76	0.78	0.70
		8192	2.11e+04	1.39	0.94	0.86	0.82	0.70	0.73	0.76	0.66

Appendix B Model Kernel Int8 (Page 5/5)

GPU	Model	M	cuBLASLt Latency (μs)	cuSPARSELt Speedup Ratio							
				2:4	2:6	2:8	2:10	2:12	2:14	2:16	2:∞
		16384	4.19e+04	1.33	0.94	0.86	0.78	0.69	0.72	0.72	0.60
	Qwen2.5-14B	64	1.31e+03	1.82	1.07	1.09	0.96	0.77	0.70	0.92	0.75
		128	1.44e+03	1.86	1.09	1.14	1.03	0.82	0.74	0.97	0.82
		256	1.68e+03	1.92	1.15	1.16	1.07	0.89	0.80	1.00	0.83
		512	2.29e+03	1.94	1.29	1.21	1.12	0.98	0.90	1.01	0.89
		1024	3.79e+03	1.82	1.29	1.10	1.04	0.91	0.83	0.94	0.83
		2048	6.67e+03	1.58	1.12	0.98	0.92	0.81	0.72	0.84	0.72
		4096	1.27e+04	1.50	1.09	0.93	0.87	0.76	0.70	0.79	0.70
		8192	2.46e+04	1.45	1.03	0.91	0.85	0.75	0.69	0.77	0.67
		16384	4.80e+04	1.44	1.01	0.89	0.83	0.74	0.67	0.75	0.66