Appendix B Model Kernel Fp16 (Page 1/4)

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	1.23e+02	1.31	1.05	0.97	0.91	0.89	0.89	0.89	0.83
		128	1.55e+02	1.47	1.20	1.12	1.06	1.03	1.03	1.01	0.94
		256	1.88e+02	1.23	0.99	0.91	0.84	0.80	0.82	0.79	0.70
		512	3.03e+02	1.33	1.05	0.96	0.89	0.85	0.84	0.83	0.75
		1024	5.82e+02	1.49	1.09	1.01	0.93	0.90	0.88	0.87	0.79
		2048	1.13e+03	1.38	1.02	0.96	0.87	0.83	0.82	0.79	0.72
		4096	2.19e+03	1.49	1.02	0.90	0.83	0.81	0.78	0.78	0.69
		8192	4.33e+03	1.45	1.05	0.98	0.85	0.81	0.81	0.79	0.69
		16384	8.64e+03	1.47	1.02	0.92	0.85	0.81	0.82	0.79	0.69
	BitNet-2B	64	1.29e+02	1.35	1.10	1.03	1.00	0.98	0.96	0.94	0.86
		128	1.63e+02	1.47	1.17	1.08	1.04	1.01	1.00	0.98	0.87
		256	2.27e+02	1.61	1.25	1.15	1.11	1.04	1.03	1.01	0.90
		512	3.37e+02	1.43	1.11	1.00	0.96	0.90	0.88	0.89	0.78
		1024	6.48e+02	1.53	1.17	1.03	0.99	0.91	0.89	0.88	0.76
		2048	1.29e+03	1.55	1.17	1.05	0.97	0.90	0.91	0.89	0.78
		4096	2.42e+03	1.55	1.10	0.98	0.93	0.89	0.86	0.84	0.74
		8192	4.86e+03	1.49	1.07	0.97	0.91	0.85	0.84	0.83	0.73
		16384	9.73e+03	1.47	1.09	0.96	0.90	0.86	0.85	0.84	0.73
	Llama3.2-3B	64	1.78e+02	1.39	1.19	1.04	1.02	0.95	0.93	0.91	0.80
		128	2.21e+02	1.57	1.27	1.12	1.09	1.03	1.01	0.99	0.88
		256	3.29e+02	1.54	1.14	1.04	0.99	0.93	0.91	0.92	0.81
		512	4.89e+02	1.47	1.14	1.01	0.95	0.90	0.86	0.88	0.77
		1024	9.54e+02	1.62	1.21	1.07	0.95	0.95	0.91	0.91	0.77
		2048	1.86e+03	1.54	1.18	0.99	0.93	0.92	0.86	0.86	0.74
		4096	3.72e+03	1.47	1.10	0.97	0.90	0.88	0.83	0.83	0.73
		8192	7.53e+03	1.44	1.11	0.97	0.91	0.88	0.84	0.84	0.73
		16384	1.47e+04	1.41	1.07	0.96	0.90	0.86	0.82	0.82	0.72
	Qwen2.5-7B	64	3.50e+02	1.48	1.16	1.04	0.97	0.93	0.93	0.92	0.81
		128	4.13e+02	1.49	1.17	1.05	1.00	0.96	0.95	0.95	0.83
		256	6.11e+02	1.40	1.06	0.97	0.89	0.88	0.84	0.85	0.75
		512	1.17e+03	1.39	1.06	0.96	0.90	0.86	0.84	0.84	0.70
		1024	2.22e+03	1.30	1.00	0.89	0.82	0.78	0.78	0.76	0.64
		2048	4.36e+03	1.35	1.03	0.90	0.83	0.79	0.77	0.76	0.66
		4096	8.64e+03	1.38	1.03	0.92	0.87	0.81	0.80	0.78	0.66
		8192	1.69e+04	1.37	1.01	0.90	0.86	0.80	0.80	0.79	0.63
		16384	3.35e+04	1.36	1.01	0.91	0.85	0.81	0.79	0.78	0.61
	Qwen2.5-14B	64	4.19e+02	1.58	1.23	1.11	1.07	1.02	0.99	0.98	0.86
		128	4.74e+02	1.55	1.24	1.13	1.07	1.03	0.99	0.98	0.88
		256	6.91e+02	1.36	1.06	0.94	0.90	0.85	0.82	0.83	0.73
		512	1.29e+03	1.43	1.06	0.95	0.90	0.86	0.82	0.82	0.72
		1024	2.65e+03	1.47	1.10	0.97	0.91	0.87	0.81	0.83	0.72
		2048	4.92e+03	1.41	1.05	0.93	0.86	0.84	0.80	0.80	0.70
		4096	9.90e+03	1.40	1.05	0.94	0.87	0.84	0.81	0.81	0.70
		8192	1.97e+04	1.39	1.02	0.91	0.86	0.83	0.80	0.80	0.69
		16384	3.92e+04	1.37	1.02	0.91	0.84	0.82	0.78	0.79	0.69
RTX4090	Llama3.2-1B	64	8.90e+01	1.18	0.93	0.88	0.88	0.94	0.84	0.94	0.77
		128	1.27e+02	1.31	1.07	0.98	0.98	0.95	0.95	0.94	0.77
		256	2.25e+02	1.68	1.33	1.21	1.13	1.07	1.01	0.90	0.92
		512	4.04e+02	1.73	1.35	1.21	1.13	1.06	1.04	0.94	0.88
		1024	7.69e+02	1.71	1.36	1.23	1.15	1.10	1.04	1.05	0.90
		2048	1.52e+03	1.83	1.41	1.27	1.18	1.13	1.06	1.08	0.93
		4096	3.02e+03	1.85	1.43	1.28	1.19	1.11	1.05	1.08	0.93
		8192	6.05e+03	1.86	1.42	1.28	1.19	1.15	1.11	1.08	0.94
		16384	1.21e+04	1.86	1.42	1.28	1.18	1.14	1.12	1.10	0.95
	Llama3.2-3B	64	1.88e+02	1.94	1.66	1.15	1.12	1.03	0.97	0.99	0.76
		128	2.36e+02	1.74	1.33	1.07	1.03	0.87	0.79	0.79	0.75
		256	3.88e+02	1.82	1.42	1.27	1.22	1.00	0.97	1.07	0.68
		512	6.82e+02	1.76	1.34	1.20	1.15	1.00	1.02	1.04	0.92
		1024	1.30e+03	1.81	1.37	1.21	1.15	1.09	1.07	1.05	0.93
		2048	2.57e+03	1.89	1.41	1.23	1.18	1.11	1.09	1.08	0.94
		4096	5.15e+03	1.98	1.46	1.28	1.22	1.16	1.14	1.13	0.98
		8192	1.02e+04	1.96	1.45	1.26	1.22	1.10	1.13	1.09	0.97
		16384	2.05e+04	1.97	1.45	1.27	1.20	1.14	1.12	1.11	0.97
	Qwen2.5-7B	64	5.25e+02	1.80	1.27	1.18	1.09	1.04	0.91	1.00	0.85
		128	5.66e+02	1.55	1.15	0.99	0.92	0.85	0.77	0.74	0.67

≥1.5× ≥1.0×

Appendix B Model Kernel Fp16 (Page 2/4)

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	8.71e+02	1.30	1.06	0.96	0.90	0.84	0.74	0.81	0.66
		512	1.59e+03	1.41	1.09	0.98	0.90	0.87	0.84	0.82	0.61
		1024	3.07e+03	1.39	1.05	0.95	0.88	0.84	0.82	0.80	0.70
		2048	5.96e+03	1.39	1.05	0.95	0.88	0.84	0.82	0.79	0.70
		4096	1.18e+04	1.40	1.06	0.96	0.88	0.85	0.81	0.81	0.71
		8192	2.40e+04	1.46	1.10	1.00	0.92	0.89	0.86	0.84	0.73
		16384	4.83e+04	1.46	1.11	1.00	0.92	0.89	0.86	0.84	0.73
	Qwen2.5-14B	64	5.85e+02	1.90	1.36	1.20	1.12	0.99	1.03	1.02	0.81
		128	6.59e+02	1.76	1.33	1.17	1.10	1.05	1.02	0.98	0.85
		256	1.00e+03	1.51	1.21	1.05	0.97	0.95	0.95	0.87	0.77
		512	1.87e+03	1.51	1.19	1.04	0.96	0.95	0.92	0.87	0.78
		1024	3.59e+03	1.52	1.18	1.02	0.95	0.95	0.91	0.71	0.77
		2048	7.15e+03	1.55	1.20	1.05	0.97	0.97	0.94	0.88	0.79
		4096	1.42e+04	1.62	1.24	1.09	1.01	0.99	0.95	0.92	0.81
		8192	2.84e+04	1.62	1.25	1.09	1.02	0.98	0.98	0.92	0.82
		16384	5.69e+04	1.62	1.25	1.09	1.03	1.00	0.97	0.92	0.82
H100	Llama3.2-1B	64	7.65e+01								
		128	7.88e+01								
		256	8.89e+01								
		512	1.41e+02								
		1024	2.62e+02								
		2048	5.21e+02								
		4096	1.09e+03								
		8192	2.22e+03								
		16384	4.52e+03								
	Llama3.2-3B	64	1.20e+02								
		128	1.23e+02								
		256	1.39e+02								
		512	2.24e+02								
		1024	4.28e+02								
		2048	8.30e+02								
		4096	1.81e+03								
		8192	3.76e+03								
		16384	7.83e+03								
	Qwen2.5-7B	64	4.19e+02								
		128	2.76e+02								
		256	3.03e+02								
		512	5.30e+02								
		1024	1.07e+03								
		2048	2.14e+03								
		4096	4.57e+03								
		8192	9.30e+03								
		16384	1.90e+04								
	Qwen2.5-14B	64	3.20e+02								
		128	3.24e+02								
		256	3.61e+02								
		512	6.05e+02								
		1024	1.22e+03								
		2048	2.65e+03								
		4096	5.20e+03								
		8192	1.12e+04								
		16384	2.28e+04								
B200	Llama3.2-1B	64	4.26e+01	1.08	0.98	0.98	0.91				
		128	4.10e+01	0.99	0.93	0.90	0.81				
		256	5.17e+01	1.18	1.06	0.96	0.89				
		512	6.62e+01	1.27	1.07	1.01	0.92				
		1024	1.01e+02	1.37	1.09	1.01	0.92				
		2048	1.71e+02	1.44	1.13	1.03	0.95				
		4096	3.16e+02	1.48	1.15	1.05	0.94				
		8192	6.48e+02	1.58	1.20	1.10	1.00				
		16384	1.27e+03	1.63	1.23	1.11	1.02				
	Llama3.2-3B	64	5.38e+01	1.30	1.11	1.08	1.05				
		128	5.78e+01	1.31	1.12	1.04	0.99				
		256	6.82e+01	1.32	1.14	1.06	0.95				
		512	9.27e+01	1.37	1.10	1.01	0.94				

Appendix B Model Kernel Fp16 (Page 3/4)

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	1.49e+02	1.35	1.09	0.98	0.91				
		2048	2.62e+02	1.41	1.09	1.00	0.92				
		4096	5.37e+02	1.65	1.26	1.12	1.03				
		8192	1.03e+03	1.60	1.23	1.10	1.02				
		16384	2.12e+03	1.69	1.31	1.14	1.07				
	Qwen2.5-7B	64	1.03e+02	1.49	1.22	1.11	1.02				
		128	1.06e+02	1.36	1.14	1.02	0.99				
		256	1.21e+02	1.39	1.06	1.00	0.96				
		512	1.75e+02	1.35	1.07	0.98	0.93				
		1024	3.20e+02	1.54	1.20	1.10	1.04				
		2048	6.17e+02	1.63	1.22	1.11	1.05				
		4096	1.26e+03	1.74	1.26	1.18	1.12				
		8192	2.54e+03	1.82	1.33	1.24	1.13				
		16384	5.20e+03	1.80	1.31	1.19	1.09				
	Qwen2.5-14B	64	1.11e+02	1.52	1.15	1.08	1.02				
		128	1.17e+02	1.43	1.12	1.05	0.99				
		256	1.36e+02	1.31	1.05	0.97	0.91				
		512	2.05e+02	1.37	1.09	1.00	0.95				
		1024	3.79e+02	1.52	1.17	1.06	1.00				
		2048	7.30e+02	1.63	1.23	1.09	1.03				
		4096	1.46e+03	1.69	1.26	1.11	1.04				
		8192	2.99e+03	1.77	1.33	1.17	1.09				
		16384	6.04e+03	1.78	1.26	1.12	1.06				
RTX5080	Llama3.2-1B	64	1.04e+02	1.21	0.95	0.91	0.89	0.86	0.86	0.84	0.62
		128	1.74e+02	1.44	1.14	1.04	0.98	0.96	0.95	0.92	0.79
		256	3.01e+02	1.50	1.18	1.06	1.00	0.97	0.95	0.93	0.81
		512	5.61e+02	1.60	1.23	1.12	1.04	1.01	0.99	0.97	0.86
		1024	1.09e+03	1.69	1.29	1.17	1.09	1.06	1.04	1.02	0.90
		2048	2.12e+03	1.73	1.32	1.19	1.11	1.08	1.05	1.03	0.91
		4096	4.23e+03	1.79	1.36	1.22	1.13	1.10	1.07	1.05	0.93
		8192	8.38e+03	1.82	1.39	1.25	1.15	1.12	1.09	1.07	0.93
		16384	1.68e+04	1.87	1.41	1.27	1.17	1.13	1.11	1.09	0.94
	BitNet-2B	64	1.39e+02	1.67	1.39	1.30	1.23	1.15	1.08	1.02	0.81
		128	1.81e+02	1.51	1.18	1.07	1.01	0.97	0.91	0.88	0.76
		256	3.24e+02	1.60	1.25	1.12	1.06	1.01	0.98	0.96	0.85
		512	6.24e+02	1.69	1.30	1.16	1.10	1.05	1.02	1.01	0.87
		1024	1.23e+03	1.81	1.38	1.24	1.16	1.11	1.07	1.05	0.93
		2048	2.41e+03	1.82	1.39	1.25	1.17	1.12	1.09	1.07	0.94
		4096	4.83e+03	1.84	1.39	1.25	1.18	1.13	1.08	1.07	0.93
		8192	9.66e+03	1.86	1.41	1.27	1.19	1.13	1.09	1.08	0.94
		16384	1.93e+04	1.89	1.43	1.28	1.20	1.14	1.10	1.08	0.94
	Llama3.2-3B	64	2.05e+02	1.75	1.18	1.11	1.01	1.00	0.97	0.97	0.85
		128	2.65e+02	1.56	1.16	1.06	1.00	0.96	0.94	0.93	0.82
		256	4.76e+02	1.59	1.20	1.08	1.01	0.97	0.93	0.90	0.78
		512	8.99e+02	1.64	1.26	1.13	1.05	1.00	0.96	0.93	0.78
		1024	1.77e+03	1.77	1.35	1.20	1.12	1.06	1.01	0.98	0.82
		2048	3.50e+03	1.83	1.39	1.23	1.15	1.09	1.03	1.01	0.84
		4096	6.96e+03	1.85	1.38	1.22	1.14	1.09	1.02	1.00	0.85
		8192	1.39e+04	1.88	1.40	1.24	1.15	1.10	1.04	1.01	0.86
		16384	2.76e+04	1.88	1.41	1.24	1.13	1.09	1.02	1.01	0.87
	Qwen2.5-7B	64	5.32e+02	1.73	1.26	1.12	1.07	0.99	0.99	0.96	0.85
		128	6.05e+02	1.72	1.25	1.12	1.05	0.99	0.98	0.96	0.84
		256	1.08e+03	1.59	1.22	1.06	0.98	0.91	0.92	0.90	0.78
		512	2.08e+03	1.63	1.24	1.08	1.00	0.94	0.94	0.91	0.79
		1024	4.08e+03	1.65	1.24	1.10	1.03	0.96	0.96	0.93	0.80
		2048	8.07e+03	1.67	1.26	1.12	1.04	0.97	0.97	0.94	0.80
		4096	1.60e+04	1.69	1.27	1.14	1.06	1.00	0.98	0.95	0.81
		8192	3.20e+04	1.69	1.28	1.14	1.06	1.01	0.99	0.96	0.81
		16384	6.36e+04	1.69	1.28	1.14	1.07	1.00	0.99	0.96	0.81
	Qwen2.5-14B	64	6.32e+02	1.83	1.36	1.20	1.14	1.11	1.04	1.02	0.85
		128	7.08e+02	1.69	1.25	1.12	1.05	1.03	0.97	0.97	0.83
		256	1.25e+03	1.62	1.19	1.06	0.99	0.96	0.93	0.91	0.80
		512	2.43e+03	1.69	1.22	1.10	1.02	0.99	0.96	0.94	0.82
		1024	4.81e+03	1.71	1.23	1.11	1.02	1.00	0.97	0.95	0.83
		2048	9.57e+03	1.73	1.23	1.11	1.02	1.00	0.97	0.95	0.83

Appendix B Model Kernel Fp16 (Page 4/4)

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	1.89e+04	1.71	1.23	1.10	1.00	0.99	0.95	0.94	0.82
		8192	3.75e+04	1.72	1.22	1.10	1.01	0.99	0.95	0.94	0.83
		16384	7.50e+04	1.72	1.22	1.10	1.01	0.99	0.96	0.94	0.83
GB10	Llama3.2-1B	64	4.86e+02	2.22	1.34	1.20	0.96	0.85	0.91	1.02	0.82
		128	5.39e+02	2.05	1.33	1.11	0.87	0.81	0.87	0.92	0.78
		256	6.13e+02	1.20	0.80	0.73	0.55	0.54	0.59	0.61	0.52
		512	9.14e+02	0.90	0.63	0.57	0.44	0.40	0.47	0.49	0.40
		1024	1.51e+03	0.77	0.56	0.49	0.36	0.34	0.39	0.43	0.34
		2048	2.97e+03	0.77	0.54	0.48	0.36	0.33	0.39	0.40	0.35
		4096	5.91e+03	0.76	0.54	0.47	0.37	0.33	0.40	0.42	0.34
		8192	1.12e+04	0.74	0.50	0.44	0.34	0.32	0.37	0.39	0.32
		16384	2.19e+04	0.70	0.52	0.47	0.34	0.31	0.37	0.39	0.32
	Llama3.2-3B	64	8.79e+02	2.12	1.41	1.17	1.13	0.96	0.81	0.97	0.84
		128	9.16e+02	1.82	1.26	1.11	0.98	0.90	0.73	0.91	0.79
		256	1.06e+03	1.18	0.78	0.72	0.68	0.61	0.49	0.62	0.53
		512	1.51e+03	0.87	0.61	0.55	0.50	0.44	0.35	0.42	0.38
		1024	2.54e+03	0.71	0.51	0.43	0.39	0.37	0.29	0.37	0.32
		2048	5.01e+03	0.74	0.52	0.45	0.42	0.38	0.30	0.38	0.32
		4096	1.13e+04	0.85	0.59	0.52	0.47	0.42	0.33	0.43	0.37
		8192	2.18e+04	0.83	0.57	0.51	0.46	0.43	0.33	0.43	0.36
		16384	4.35e+04	0.80	0.57	0.51	0.45	0.40	0.32	0.42	0.35
	Qwen2.5-7B	64	2.12e+03	1.70	1.26	1.10	1.02				
		128	2.21e+03	1.62	1.21	1.04	0.97				
		256	2.52e+03	1.01	0.75	0.67	0.61				
		512	3.60e+03	0.76	0.55	0.49	0.45				
		1024	6.22e+03	0.68	0.48	0.43	0.40				
		2048	1.22e+04	0.68	0.49	0.44	0.40				
		4096	2.63e+04	0.72	0.53	0.47	0.43				
		8192	5.12e+04	0.71	0.52	0.46	0.43				
		16384	1.00e+05	0.70	0.51	0.45	0.42				
	Qwen2.5-14B	64	2.62e+03	1.71	1.24	1.11	1.02				
		128	2.70e+03	1.59	1.21	1.06	0.99				
		256	2.98e+03	0.93	0.69	0.61	0.57				
		512	4.27e+03	0.65	0.50	0.45	0.41				
		1024	7.62e+03	0.61	0.45	0.41	0.37				
		2048	1.59e+04	0.64	0.47	0.43	0.39				
		4096	3.06e+04	0.63	0.46	0.41	0.38				
		8192	5.93e+04	0.61	0.45	0.40	0.37				
		16384	1.16e+05	0.61	0.44	0.39	0.36				