Appendix B Model Kernel Bf16 (Page 1/4)

GPU	Model	M	cuBLASLt	cuSPARSELt Speedup Ratio							
			Latency (μs)	2:4	2:6	2:8	2:10	2:12	2:14	2:16	2:∞
A100	Llama3.2-1B	64	1.19e+02	1.25	0.98	0.92	0.88	0.84	0.86	0.85	0.80
		128	1.41e+02	1.33	1.08	1.01	0.97	0.94	0.94	0.92	0.85
		256	1.88e+02	1.22	0.95	0.89	0.84	0.80	0.81	0.77	0.70
		512	3.20e+02	1.40	1.09	1.01	0.93	0.89	0.88	0.88	0.78
		1024	5.90e+02	1.46	1.10	1.03	0.94	0.89	0.90	0.89	0.79
		2048	1.14e+03	1.39	1.02	0.95	0.87	0.82	0.81	0.80	0.71
		4096	2.27e+03	1.48	1.05	0.93	0.85	0.82	0.80	0.79	0.71
		8192	4.42e+03	1.46	1.05	0.94	0.86	0.82	0.82	0.79	0.71
		16384	8.85e+03	1.43	1.03	0.93	0.86	0.81	0.82	0.79	0.70
	BitNet-2B	64	1.28e+02	1.34	1.10	1.03	1.00	0.96	0.95	0.94	0.85
		128	1.50e+02	1.35	1.07	1.00	0.95	0.91	0.90	0.90	0.80
		256	1.95e+02	1.37	1.06	0.98	0.94	0.90	0.87	0.86	0.77
		512	3.59e+02	1.50	1.16	1.07	1.02	0.95	0.93	0.93	0.82
		1024	6.52e+02	1.52	1.17	1.05	0.98	0.91	0.89	0.90	0.77
		2048	1.26e+03	1.50	1.12	0.98	0.94	0.86	0.86	0.83	0.72
		4096	2.47e+03	1.52	1.14	0.98	0.94	0.90	0.86	0.86	0.75
		8192	4.93e+03	1.50	1.10	0.97	0.91	0.87	0.84	0.83	0.74
		16384	9.82e+03	1.46	1.10	0.96	0.90	0.86	0.83	0.82	0.72
	Llama3.2-3B	64	1.73e+02	1.35	1.14	1.01	0.98	0.93	0.90	0.88	0.79
		128	2.02e+02	1.44	1.16	1.03	1.00	0.95	0.92	0.92	0.80
		256	2.92e+02	1.34	1.03	0.91	0.87	0.84	0.81	0.79	0.70
		512	5.01e+02	1.49	1.15	1.02	0.98	0.93	0.90	0.88	0.76
		1024	9.60e+02	1.63	1.22	1.06	1.00	0.93	0.90	0.89	0.76
		2048	1.92e+03	1.57	1.18	1.02	0.96	0.93	0.87	0.89	0.76
		4096	3.78e+03	1.48	1.11	0.98	0.91	0.88	0.84	0.84	0.74
		8192	7.54e+03	1.42	1.09	0.96	0.90	0.87	0.83	0.84	0.73
		16384	1.49e+04	1.40	1.08	0.96	0.90	0.87	0.82	0.83	0.72
	Qwen2.5-7B	64	3.50e+02	1.49	1.16	1.03	0.98	0.93	0.93	0.92	0.81
		128	4.03e+02	1.44	1.13	1.02	0.97	0.93	0.93	0.92	0.80
		256	6.18e+02	1.41	1.09	0.99	0.93	0.90	0.87	0.85	0.77
		512	1.15e+03	1.35	1.04	0.93	0.88	0.83	0.82	0.82	0.68
		1024	2.19e+03	1.25	0.98	0.88	0.80	0.77	0.75	0.75	0.63
		2048	4.39e+03	1.35	1.01	0.90	0.82	0.78	0.77	0.77	0.65
		4096	8.62e+03	1.36	1.02	0.90	0.84	0.80	0.78	0.77	0.63
		8192	1.71e+04	1.36	1.02	0.91	0.84	0.81	0.79	0.78	0.62
		16384	3.38e+04	1.33	1.01	0.90	0.84	0.80	0.78	0.77	0.61
	Qwen2.5-14B	64	4.40e+02	1.66	1.30	1.18	1.13	1.07	1.04	1.04	0.91
		128	4.72e+02	1.52	1.22	1.12	1.07	1.02	0.99	0.98	0.87
		256	7.12e+02	1.40	1.08	0.96	0.92	0.88	0.86	0.84	0.75
		512	1.30e+03	1.39	1.04	0.92	0.88	0.83	0.79	0.79	0.69
		1024	2.51e+03	1.36	1.01	0.88	0.83	0.79	0.75	0.77	0.67
		2048	4.99e+03	1.41	1.05	0.93	0.87	0.83	0.79	0.81	0.70
		4096	9.88e+03	1.36	1.03	0.92	0.86	0.82	0.79	0.79	0.69
		8192	1.98e+04	1.36	1.02	0.91	0.86	0.81	0.78	0.79	0.69
		16384	3.96e+04	1.35	1.02	0.91	0.84	0.81	0.78	0.79	0.68
RTX4090	Llama3.2-1B	64	9.71e+01	1.27	1.04	0.99	0.93	0.97	1.01	0.89	0.86
		128	1.31e+02	1.28	1.09	1.06	0.97	0.99	0.95	0.98	0.71
		256	2.22e+02	1.64	1.33	1.20	1.14	0.99	1.07	0.97	0.85
		512	4.12e+02	1.73	1.37	1.24	1.12	1.02	1.04	1.05	0.85
		1024	7.97e+02	1.86	1.41	0.99	1.18	1.10	1.07	0.84	0.91
		2048	1.54e+03	1.85	1.20	1.28	1.18	1.11	1.07	1.10	0.95
		4096	3.09e+03	1.88	1.44	1.30	1.20	1.11	1.09	1.10	0.95
		8192	6.19e+03	1.89	1.45	1.30	1.20	1.14	1.12	1.10	0.94
		16384	1.23e+04	1.88	1.43	1.29	1.19	1.16	1.12	1.12	0.96
	Llama3.2-3B	64	1.92e+02	2.09	1.73	1.21	1.09	1.05	1.00	0.99	0.89
		128	2.37e+02	1.81	1.32	1.07	1.04	0.97	0.96	0.76	0.77
		256	3.78e+02	1.83	1.36	1.22	1.18	1.12	0.98	0.94	0.90
		512	6.86e+02	1.77	1.33	1.20	1.14	1.05	1.04	1.05	0.85
		1024	1.29e+03	1.80	1.34	1.19	1.14	0.82	1.06	1.03	0.91
		2048	2.53e+03	1.86	1.36	1.20	1.14	1.09	1.07	1.06	0.93
		4096	5.11e+03	1.92	1.43	1.27	1.20	1.15	1.11	1.11	0.97
		8192	1.02e+04	1.95	1.43	1.26	1.20	1.14	1.12	1.10	0.94
		16384	2.03e+04	1.94	1.42	1.25	1.18	1.13	1.10	1.09	0.96
	Qwen2.5-7B	64	5.24e+02	1.87	1.32	1.15	1.09	1.05	0.93	0.84	0.89
		128	5.66e+02	1.54	1.15	0.98	0.92	0.85	0.78	0.80	0.65

🟩 ≥1.5× 🟩 ≥1.0×

Appendix B Model Kernel Bf16 (Page 2/4)

GPU	Model	M	cuBLASLt	cuSPARSELt Speedup Ratio							
			Latency (µs)	2:4	2:6	2:8	2:10	2:12	2:14	2:16	2:∞
		256	8.82e+02	1.41	1.08	0.98	0.90	0.66	0.81	0.81	0.70
		512	1.60e+03	1.40	1.08	0.98	0.89	0.86	0.82	0.69	0.71
		1024	3.10e+03	1.39	1.06	0.96	0.88	0.86	0.82	0.80	0.70
		2048	5.98e+03	1.37	1.05	0.95	0.88	0.84	0.81	0.80	0.70
		4096	1.18e+04	1.38	1.05	0.95	0.88	0.85	0.82	0.81	0.69
		8192	2.36e+04	1.43	1.07	0.98	0.90	0.87	0.84	0.83	0.71
		16384	4.72e+04	1.42	1.08	0.97	0.90	0.86	0.84	0.82	0.71
	Qwen2.5-14B	64	5.69e+02	1.87	1.32	1.16	1.10	0.85	1.01	0.99	0.75
		128	6.45e+02	1.71	1.29	1.15	1.09	1.03	0.92	0.95	0.84
		256	1.03e+03	1.57	1.23	1.03	0.99	1.00	0.96	0.91	0.82
		512	1.91e+03	1.53	1.20	1.06	0.99	0.97	0.94	0.90	0.79
		1024	3.71e+03	1.55	1.20	1.05	0.99	0.97	0.94	0.89	0.80
		2048	7.45e+03	1.60	1.24	1.09	1.02	1.01	0.97	0.93	0.83
		4096	1.49e+04	1.69	1.31	1.14	1.07	1.03	1.01	0.97	0.85
		8192	3.00e+04	1.70	1.31	1.14	1.08	1.06	1.03	0.97	0.87
		16384	6.01e+04	1.71	1.31	1.15	1.07	1.05	1.02	0.97	0.86
H100	Llama3.2-1B	64	7.57e+01	0.74	0.61	0.57	0.53	0.51	0.48	0.50	0.42
		128	8.10e+01	0.77	0.65	0.60	0.56	0.53	0.50	0.53	0.46
		256	9.37e+01	0.87	0.72	0.66	0.61	0.59	0.57	0.58	0.52
		512	1.43e+02	1.09	0.83	0.76	0.71	0.69	0.65	0.67	0.60
		1024	2.63e+02	1.27	0.98	0.89	0.83	0.80	0.75	0.77	0.62
		2048	5.40e+02	1.41	1.09	0.98	0.90	0.87	0.76	0.84	0.65
		4096	1.05e+03	1.43	1.09	0.98	0.91	0.85	0.82	0.84	0.64
		8192	2.16e+03	1.52	1.11	1.00	0.91	0.87	0.84	0.83	0.70
		16384	4.64e+03	1.59	1.17	1.06	0.97	0.92	0.87	0.90	0.70
	Llama3.2-3B	64	1.22e+02	0.93	0.72	0.66	0.63	0.61	0.59	0.58	0.53
		128	1.25e+02	0.93	0.73	0.66	0.64	0.63	0.60	0.60	0.55
		256	1.42e+02	1.03	0.81	0.73	0.70	0.69	0.66	0.66	0.59
		512	2.25e+02	1.21	0.95	0.84	0.81	0.79	0.74	0.73	0.66
		1024	4.46e+02	1.48	1.12	0.99	0.94	0.91	0.86	0.85	0.75
		2048	8.93e+02	1.56	1.18	1.04	0.97	0.95	0.88	0.88	0.78
		4096	1.80e+03	1.60	1.15	1.01	0.94	0.91	0.87	0.86	0.73
		8192	3.77e+03	1.61	1.17	1.03	0.99	0.93	0.89	0.86	0.75
		16384	7.48e+03	1.53	1.13	1.01	0.91	0.91	0.84	0.83	0.73
	Qwen2.5-7B	64	2.69e+02	1.03	0.86	0.80	0.76	0.73	0.73	0.71	0.62
		128	2.76e+02	1.13	0.90	0.85	0.80	0.77	0.76	0.74	0.66
		256	3.05e+02	1.25	0.98	0.90	0.85	0.81	0.80	0.78	0.69
		512	5.39e+02	1.43	1.09	0.97	0.93	0.87	0.86	0.84	0.74
		1024	1.14e+03	1.74	1.27	1.15	1.07	1.02	1.01	0.98	0.76
		2048	2.02e+03	1.54	1.11	0.97	0.91	0.83	0.81	0.80	0.71
		4096	4.76e+03	1.65	1.26	1.11	1.03	0.92	0.93	0.92	0.78
		8192	9.70e+03	1.58	1.24	1.09	0.99	0.99	0.95	0.92	0.80
		16384	1.95e+04	1.71	1.20	1.10	0.99	0.90	0.89	0.88	0.76
	Qwen2.5-14B	64	3.19e+02	1.20	0.92	0.83	0.78				
		128	3.28e+02	1.27	0.96	0.88	0.83				
		256	3.68e+02	1.37	1.03	0.94	0.90				
		512	5.96e+02	1.41	1.08	0.97	0.91				
		1024	1.29e+03	1.64	1.24	1.10	1.02				
		2048	2.54e+03	1.62	1.18	1.08	0.99				
		4096	5.50e+03	1.66	1.25	1.06	1.03				
		8192	1.14e+04	1.66	1.21	1.09	1.04				
		16384	2.32e+04	1.60	1.18	1.13	1.00				
B200	Llama3.2-1B	64	4.46e+01	1.14	1.03	1.03	0.93				
		128	4.85e+01	1.18	1.09	1.07	0.96				
		256	6.07e+01	1.38	1.25	1.13	1.07				
		512	6.49e+01	1.24	1.05	0.98	0.90				
		1024	1.01e+02	1.36	1.09	1.03	0.92				
		2048	1.71e+02	1.41	1.13	1.03	0.94				
		4096	3.21e+02	1.50	1.16	1.07	0.95				
		8192	6.49e+02	1.59	1.18	1.08	0.96				
		16384	1.34e+03	1.67	1.29	1.20	1.04				
	Llama3.2-3B	64	5.68e+01	1.37	1.18	1.15	1.10				
		128	5.80e+01	1.29	1.12	1.04	0.97				
		256	6.21e+01	1.20	1.00	0.97	0.87				
		512	9.06e+01	1.33	1.07	0.98	0.92				

Appendix B Model Kernel Bf16 (Page 3/4)

GPU	Model	M	cuBLASLt	cuSPARSELt Speedup Ratio							
			Latency (μs)	2:4	2:6	2:8	2:10	2:12	2:14	2:16	2:∞
		1024	1.47e+02	1.33	1.07	0.97	0.90				
		2048	2.67e+02	1.43	1.11	1.00	0.92				
		4096	5.22e+02	1.59	1.20	1.08	1.00				
		8192	1.07e+03	1.67	1.25	1.11	1.05				
		16384	2.13e+03	1.68	1.30	1.14	1.04				
	Qwen2.5-7B	64	1.02e+02	1.48	1.19	1.08	1.02				
		128	1.05e+02	1.36	1.12	1.01	0.97				
		256	1.22e+02	1.40	1.07	1.01	0.97				
		512	1.82e+02	1.41	1.10	1.02	0.96				
		1024	3.18e+02	1.53	1.18	1.09	1.02				
		2048	6.37e+02	1.66	1.26	1.14	1.08				
		4096	1.27e+03	1.70	1.27	1.17	1.10				
		8192	2.64e+03	1.81	1.38	1.24	1.14				
		16384	5.27e+03	1.81	1.33	1.20	1.10				
	Qwen2.5-14B	64	1.13e+02	1.57	1.20	1.10	1.03				
		128	1.16e+02	1.43	1.10	1.05	0.97				
		256	1.36e+02	1.32	1.05	0.98	0.92				
		512	2.02e+02	1.35	1.06	0.98	0.92				
		1024	4.00e+02	1.60	1.21	1.10	1.04				
		2048	7.21e+02	1.59	1.18	1.08	1.00				
		4096	1.44e+03	1.65	1.23	1.09	1.01				
		8192	3.04e+03	1.80	1.31	1.20	1.08				
		16384	6.14e+03	1.77	1.28	1.13	1.06				
RTX5080	Llama3.2-1B	64	1.13e+02	1.31	1.05	0.99	0.97	0.94	0.95	0.90	0.67
		128	1.98e+02	1.64	1.30	1.18	1.14	1.07	1.05	1.05	0.91
		256	3.60e+02	1.79	1.40	1.26	1.19	1.14	1.13	1.11	0.97
		512	6.38e+02	1.81	1.40	1.27	1.18	1.14	1.12	1.10	0.98
		1024	1.23e+03	1.91	1.47	1.33	1.24	1.20	1.17	1.15	1.02
		2048	2.31e+03	1.87	1.44	1.29	1.20	1.16	1.13	1.12	0.98
		4096	4.41e+03	1.85	1.41	1.28	1.18	1.14	1.11	1.09	0.96
		8192	8.63e+03	1.87	1.42	1.28	1.18	1.13	1.11	1.09	0.95
		16384	1.70e+04	1.87	1.43	1.27	1.17	1.13	1.11	1.09	0.95
	BitNet-2B	64	2.47e+02	2.96	2.46	2.28	2.19	2.03	1.88	1.82	1.43
		128	2.91e+02	2.39	1.87	1.72	1.62	1.55	1.46	1.43	1.22
		256	3.60e+02	1.78	1.39	1.25	1.17	1.12	1.09	1.06	0.94
		512	7.01e+02	1.91	1.45	1.31	1.23	1.17	1.14	1.13	0.98
		1024	1.29e+03	1.89	1.43	1.29	1.22	1.16	1.12	1.10	0.96
		2048	2.49e+03	1.87	1.43	1.29	1.21	1.16	1.12	1.11	0.97
		4096	4.95e+03	1.89	1.43	1.29	1.20	1.14	1.10	1.09	0.95
		8192	9.82e+03	1.89	1.43	1.28	1.20	1.14	1.10	1.09	0.95
		16384	1.93e+04	1.89	1.42	1.28	1.19	1.13	1.09	1.08	0.94
	Llama3.2-3B	64	3.80e+02	3.24	2.23	2.04	1.89	1.85	1.80	1.80	1.61
		128	4.26e+02	2.55	1.86	1.72	1.60	1.54	1.50	1.48	1.33
		256	5.61e+02	1.88	1.40	1.27	1.20	1.14	1.09	1.05	0.92
		512	1.05e+03	1.90	1.47	1.31	1.23	1.17	1.12	1.08	0.90
		1024	1.91e+03	1.90	1.46	1.30	1.20	1.14	1.09	1.05	0.88
		2048	3.63e+03	1.90	1.43	1.27	1.18	1.13	1.07	1.04	0.87
		4096	7.19e+03	1.90	1.41	1.26	1.17	1.12	1.06	1.04	0.87
		8192	1.40e+04	1.89	1.40	1.25	1.14	1.09	1.04	1.02	0.87
		16384	2.78e+04	1.89	1.40	1.23	1.12	1.08	1.02	1.01	0.87
	Qwen2.5-7B	64	8.98e+02	2.92	2.13	1.91	1.79	1.69	1.68	1.63	1.42
		128	6.30e+02	1.78	1.31	1.17	1.10	1.05	1.02	1.00	0.87
		256	1.13e+03	1.66	1.27	1.11	1.03	0.96	0.94	0.94	0.80
		512	2.16e+03	1.69	1.28	1.13	1.05	0.98	0.97	0.94	0.81
		1024	4.25e+03	1.72	1.29	1.15	1.07	1.01	0.99	0.96	0.83
		2048	8.18e+03	1.69	1.27	1.13	1.06	0.99	0.97	0.95	0.81
		4096	1.62e+04	1.70	1.29	1.15	1.07	1.00	0.99	0.95	0.82
		8192	3.21e+04	1.69	1.28	1.14	1.07	1.00	0.98	0.96	0.82
		16384	6.39e+04	1.69	1.28	1.14	1.07	1.01	0.98	0.96	0.81
	Qwen2.5-14B	64	1.07e+03	3.08	2.30	2.04	1.94	1.86	1.76	1.72	1.43
		128	1.03e+03	2.45	1.80	1.64	1.54	1.49	1.41	1.40	1.20
		256	1.38e+03	1.79	1.31	1.19	1.10	1.06	1.02	1.01	0.88
		512	2.51e+03	1.73	1.26	1.13	1.04	1.02	0.98	0.96	0.84
		1024	4.88e+03	1.73	1.25	1.13	1.05	1.01	0.97	0.96	0.84
		2048	9.66e+03	1.74	1.24	1.12	1.04	1.00	0.96	0.95	0.83

Appendix B Model Kernel Bf16 (Page 4/4)

GPU	Model	M	cuBLASLt	cuSPARSELt Speedup Ratio							
			Latency (μs)	2:4	2:6	2:8	2:10	2:12	2:14	2:16	2:∞
		4096	1.92e+04	1.73	1.24	1.11	1.03	1.00	0.96	0.95	0.83
		8192	3.78e+04	1.72	1.23	1.10	1.02	0.99	0.95	0.94	0.83
		16384	7.49e+04	1.71	1.22	1.10	1.01	0.99	0.95	0.94	0.82
GB10	Llama3.2-1B	64	5.78e+02	2.66	1.75	1.45	1.08	0.97	1.09	1.12	0.96
		128	7.01e+02	2.69	1.84	1.47	1.11	1.04	1.17	1.19	1.05
		256	7.17e+02	1.45	1.01	0.81	0.64	0.60	0.68	0.70	0.59
		512	9.84e+02	1.06	0.67	0.64	0.47	0.42	0.48	0.50	0.43
		1024	1.61e+03	0.81	0.59	0.51	0.38	0.37	0.41	0.43	0.37
		2048	2.97e+03	0.79	0.52	0.50	0.36	0.33	0.39	0.40	0.34
		4096	5.90e+03	0.78	0.54	0.48	0.36	0.33	0.39	0.42	0.34
		8192	1.12e+04	0.72	0.51	0.45	0.34	0.31	0.38	0.38	0.33
		16384	2.20e+04	0.75	0.51	0.44	0.34	0.33	0.37	0.39	0.32
	Llama3.2-3B	64	9.97e+02	2.50	1.59	1.34	1.27	1.12	0.90	1.10	0.91
		128	1.17e+03	2.34	1.57	1.40	1.28	1.12	0.94	1.16	1.02
		256	1.20e+03	1.37	0.91	0.80	0.72	0.68	0.57	0.70	0.60
		512	1.44e+03	0.85	0.56	0.49	0.46	0.41	0.33	0.43	0.35
		1024	2.56e+03	0.74	0.51	0.45	0.41	0.37	0.30	0.38	0.33
		2048	4.99e+03	0.72	0.50	0.45	0.42	0.38	0.30	0.38	0.32
		4096	9.88e+03	0.73	0.51	0.45	0.42	0.38	0.30	0.38	0.32
		8192	1.86e+04	0.71	0.49	0.42	0.40	0.36	0.28	0.36	0.30
		16384	3.68e+04	0.70	0.48	0.42	0.39	0.36	0.28	0.36	0.30
	Qwen2.5-7B	64	2.38e+03	1.95	1.41	1.25	1.15				
		128	2.59e+03	1.92	1.40	1.24	1.15				
		256	2.87e+03	1.18	0.85	0.76	0.69				
		512	3.75e+03	0.79	0.58	0.51	0.47				
		1024	6.16e+03	0.66	0.48	0.43	0.40				
		2048	1.19e+04	0.66	0.48	0.42	0.39				
		4096	2.26e+04	0.63	0.46	0.41	0.37				
		8192	4.29e+04	0.59	0.44	0.38	0.36				
		16384	8.46e+04	0.59	0.43	0.38	0.35				
	Qwen2.5-14B	64	3.02e+03	1.99	1.46	1.29	1.20				
		128	3.14e+03	1.89	1.33	1.23	1.12				
		256	3.40e+03	1.09	0.77	0.69	0.66				
		512	4.50e+03	0.70	0.52	0.47	0.44				
		1024	7.64e+03	0.61	0.45	0.40	0.37				
		2048	1.55e+04	0.64	0.46	0.41	0.38				
		4096	2.88e+04	0.60	0.43	0.39	0.36				
		8192	5.55e+04	0.58	0.42	0.37	0.35				
		16384	1.09e+05	0.56	0.41	0.36	0.34				