Appendix A Square Bf16

GPU	M	cuBLASLt Latency (μs)	cuSPARSELt Speedup Ratio							
			2:4	2:6	2:8	2:10	2:12	2:14	2:16	2:∞
A100	64	4.32e+00	0.76	0.70	0.69	0.71	0.71	0.72	0.71	0.70
	128	4.57e+00	0.71	0.69	0.65	0.65	0.65	0.66	0.65	0.66
	256	5.80e+00	0.83	0.75	0.76	0.72	0.73	0.74	0.74	0.70
	512	7.58e+00	0.90	0.81	0.78	0.76	0.71	0.74	0.74	0.69
	1024	1.91e+01	1.19	1.01	0.97	0.93	0.91	0.88	0.90	0.83
	2048	8.21e+01	1.18	0.93	0.91	0.86	0.83	0.83	0.80	0.62
	4096	5.86e+02	1.71	1.25	1.12	1.01	0.99	0.96	0.95	0.82
	8192	4.76e+03	1.52	1.02	0.98	0.93	0.80	0.80	0.84	0.70
	16384	3.80e+04	1.22	0.91	0.81	0.77	0.74	0.72	0.71	0.62
RTX4090	64	9.54e+00	0.99	0.90	0.92	0.92	0.39	0.94	0.43	0.97
	128	9.93e+00	1.15	1.08	1.09	1.00	0.20	0.32	0.20	0.20
	256	1.00e+01	1.03	0.91	0.98	0.95	0.15	0.15	0.15	0.19
	512	1.18e+01	1.14	1.00	0.96	0.95	0.12	0.12	0.12	0.13
	1024	1.92e+01	1.46	1.16	1.11	1.01	0.11	0.11	0.11	0.90
	2048	1.06e+02	1.72	1.33	1.20	1.13	0.09	0.09	0.12	0.97
	4096	8.24e+02	1.84	1.40	1.25	1.18	0.67	1.16	0.93	0.99
	8192	6.90e+03	2.01	1.51	1.33	1.25	1.22	1.18	1.17	1.00
	16384	5.73e+04	1.97	1.46	1.31	1.22	1.19	1.16	1.13	0.99
H100	64	4.66e+00	0.80	0.76	0.76	0.78	0.74	0.75	0.74	0.74
	128	4.59e+00	0.69	0.71	0.71	0.68	0.66	0.65	0.66	0.66
	256	4.57e+00	0.67	0.61	0.62	0.61	0.59	0.59	0.59	0.58
	512	4.71e+00	0.60	0.56	0.54	0.53	0.50	0.51	0.51	0.49
	1024	7.80e+00	0.75	0.67	0.64	0.60	0.59	0.59	0.58	0.55
	2048	3.50e+01	0.97	0.78	0.72	0.65	0.63	0.62	0.62	0.54
	4096	2.89e+02	1.53	1.13	0.98	0.95	0.95	0.94	0.86	0.79
	8192	2.59e+03	1.59	1.18	1.05	0.95	0.93	0.92	0.90	0.79
	16384	2.23e+04	1.45	1.05	0.93	0.88	0.92	0.89	0.88	0.77
B200	64	5.89e+00	1.15	0.95	0.95	0.95	0.96	1.13	1.13	0.95
	128	5.40e+00	0.87	0.87	0.87	0.87	0.87	0.87	0.92	0.87
	256	5.51e+00	0.89	0.89	0.89	0.89	0.88	0.89	0.89	0.89
	512	5.55e+00	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89
	1024	6.09e+00	0.83	0.74	0.74	0.74	0.74	0.73	0.73	0.74
	2048	1.65e+01	1.28	1.14	1.00	1.00	0.99	1.00	1.00	0.89
	4096	9.09e+01	1.61	1.13	1.07	0.96	0.95	0.95	0.92	0.81
	8192	6.69e+02	1.64	1.20	1.06	1.00	0.92	0.88	0.93	0.83
	16384	5.97e+03	1.61	1.14	1.08	1.01	0.96	0.94	0.93	0.81
RTX5080	64	2.13e+00	0.52	0.52	0.52	0.52	0.51	0.52	0.35	0.52
	128	2.41e+00	0.58	0.58	0.58	0.39	0.39	0.39	0.38	0.39
	256	4.06e+00	0.66	0.66	0.66	0.65	0.65	0.65	0.65	0.66
	512	8.19e+00	1.32	1.00	0.99	0.99	0.95	0.90	0.78	0.80
	1024	2.67e+01	1.63	1.30	1.18	1.18	1.08	1.08	1.08	1.00
	2048	1.90e+02	1.89	1.45	1.31	1.22	1.18	1.16	1.15	1.01
	4096	1.22e+03	1.93	1.45	1.30	1.22	1.17	1.14	1.12	0.98
	8192	9.10e+03	1.81	1.34	1.17	1.13	1.06	1.04	1.00	0.92
	16384	7.28e+04	1.53	1.13	0.98	0.94	0.90	0.88	0.85	0.75
GB10	64	3.03e+00	0.73	0.66	0.73	0.66	0.69	0.70	0.68	0.70
	128	4.29e+00	0.90	0.67	0.69	0.53	0.69	0.70	0.57	0.54
	256	5.30e+00	0.85	0.82	0.86	0.81	0.84	0.85	0.83	0.81
	512	1.24e+01	1.58	1.41	1.38	1.41	1.18	1.00	1.32	1.09
	1024	3.72e+01	1.38	1.13	1.03	0.93	0.95	0.95	0.92	0.82
	2048	1.96e+02	1.38	1.08	1.00	0.92	0.90	0.88	0.79	0.71
	4096	1.67e+03	1.36	0.63	0.63	0.47	0.53	0.52	0.51	0.43
	8192	1.36e+04	0.54	0.40	0.36	0.33	0.25	0.24	0.31	0.27
	16384	1.03e+05	0.51	0.29	0.34	0.32	0.31	0.30	0.29	0.26

≥1.5× ≥1.0×