Appendix A Square Int8

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	5.57e+00	1.04			1.02	1.02	0.97	1.03	1.02
	128	5.86e+00	1.04			0.95	0.97	0.97	0.98	0.99
	256	6.14e+00	1.05			0.88	0.91	0.92	0.92	0.92
	512	7.11e+00	1.05			0.89	0.84	0.93	0.93	0.89
	1024	1.36e+01	1.18			0.95	0.92	0.86	0.93	0.88
	2048	5.54e+01	1.42			0.96	0.93	1.04	0.94	0.85
	4096	3.47e+02	2.06			1.13	1.25	1.23	1.19	1.08
	8192	3.05e+03	2.19			1.29	1.11	1.07	1.23	1.08
	16384	2.51e+04	2.18			1.36	1.22	1.23	1.22	0.91
RTX4090	64	9.52e+00	1.05	1.09	1.06	0.99	0.38	0.42	0.35	0.35
	128	9.91e+00	0.48	1.13	1.00	1.17	0.31	0.31	0.31	0.32
	256	1.01e+01	1.04	1.11	1.05	1.08	0.27	0.27	0.27	0.26
	512	1.06e+01	1.21	1.14	1.08	1.12	0.20	0.21	0.21	0.20
	1024	1.16e+01	1.02	0.92	0.99	0.91	0.13	0.11	0.12	0.11
	2048	5.35e+01	2.06	1.74	1.49	1.24	0.14	0.16	0.15	0.14
	4096	2.36e+02	1.49	0.96	1.06	0.80	0.10	0.51	0.67	0.59
	8192	1.94e+03	1.60	1.21	1.07	1.01	0.79	0.77	0.94	0.82
	16384	1.53e+04	1.59	0.95	1.04	0.98	0.91	0.88	0.89	0.76
H100	64	4.41e+00	0.87	0.88	0.87	0.88	0.85	0.84	0.85	0.84
	128	4.44e+00	0.86	0.85	0.84	0.84	0.81	0.81	0.81	0.83
	256	4.86e+00	0.93	0.87	0.89	0.87	0.86	0.85	0.86	0.64
	512	5.77e+00	0.96	0.97	0.97	0.95	0.90	0.93	0.93	0.91
	1024	1.12e+01	1.30	1.05	1.16	1.12	1.05	0.98	1.09	1.04
	2048	3.13e+01	1.58	1.25	1.17	1.08	1.05	1.08	1.10	1.00
	4096	1.36e+02	1.28	0.98	0.94	0.85	0.87	0.83	0.79	0.73
	8192	1.26e+03	1.71	1.24	0.98	1.00	0.90	0.80	0.91	0.75
	16384	1.25e+04	1.79	1.18	1.33	1.03	1.12	1.11	1.07	0.94
B200	64	4.79e+00	0.77	0.78	0.77	0.77	0.77	0.77	0.77	0.77
	128	4.85e+00	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78
	256	4.84e+00	0.78	0.81	0.78	0.78	0.78	0.78	0.78	0.79
	512	6.25e+00	1.01	1.01	1.01	1.01	1.00	1.01	1.00	1.01
	1024	8.31e+00	1.34	1.23	1.33	1.33	1.19	1.01	1.29	1.15
	2048	2.74e+01	2.65	2.21	2.22	1.97	1.90	2.21	2.21	2.18
	4096	1.54e+02	5.34	3.75	3.98	3.26	3.41	3.50	3.43	3.01
	8192	1.18e+03	6.47	4.46	4.31	3.62	3.21	3.11	3.56	3.09
	16384	9.67e+03	6.11	3.83	3.82	3.57	3.13	3.12	3.21	2.73
RTX5080	64	4.16e+00	1.02	1.01	1.02	1.03	1.01	1.02	1.01	1.02
	128	4.15e+00	1.03	0.76	1.01	1.01	1.01	1.01	1.01	1.18
	256	4.17e+00	1.01	1.02	1.02	1.01	1.01	0.89	1.01	1.01
	512	5.94e+00	1.46	1.09	1.44	1.43	0.96	1.02	1.01	0.97
	1024	1.23e+01	1.51	1.49	1.50	1.21	1.21	1.20	1.20	1.20
	2048	5.95e+01	1.61	1.25	1.16	0.98	0.94	1.00	1.00	0.91
	4096	3.55e+02	1.63	1.12	1.11	0.94	0.99	0.97	0.96	0.85
	8192	2.60e+03	1.61	1.18	1.04	0.98	0.82	0.80	0.90	0.80
	16384	2.07e+04	1.57	1.03	1.04	0.98	0.93	0.90	0.90	0.80
GB10	64	4.18e+00	1.00	1.02	0.94	1.00	1.04	1.01	1.00	1.00
	128	4.21e+00	0.97	0.98	1.01	1.00	1.01	1.00	1.01	0.87
	256	4.95e+00	1.19	1.15	0.77	0.83	0.92	0.92	0.90	0.92
	512	6.26e+00	1.01	1.01	0.84	1.01	0.90	0.93	1.01	1.01
	1024	2.26e+01	1.37	1.09	1.05	1.06	1.01	0.93	1.00	0.92
	2048	9.97e+01	1.23	1.06	0.95	0.77	0.65	0.87	0.84	0.74
	4096	7.70e+02	1.52	1.01	1.00	0.83	0.90	0.86	0.88	0.75
	8192	6.04e+03	1.46	1.09	0.99	0.88	0.74	0.72	0.84	0.70
	16384	5.18e+04	1.55	0.75	0.78	0.72	0.63	0.61	0.63	0.49

≥1.5× ≥1.0×