Appendix A Square Fp16

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	4.01e+00	0.70	0.64		0.66	0.65	0.65	0.66	0.65
	128	4.28e+00	0.69	0.64		0.61	0.61	0.61	0.61	0.61
	256	5.20e+00	0.76	0.67		0.65	0.65	0.66	0.66	0.63
	512	7.17e+00	0.87	0.77		0.72	0.67	0.70	0.70	0.66
	1024	1.98e+01	1.24	1.04		0.97	0.95	0.91	0.93	0.86
	2048	7.45e+01	1.08	0.85		0.78	0.76	0.73	0.73	0.56
	4096	5.90e+02	1.81	1.24		1.16	1.07	1.00	0.95	0.83
	8192	4.68e+03	1.52	1.00		0.97	0.82	0.81	0.81	0.68
	16384	3.74e+04	1.22	0.91		0.79	0.77	0.72	0.71	0.62
RTX4090	64	9.44e+00	1.00	1.01	0.99	1.08	0.90	1.01	1.01	0.24
	128	9.32e+00	1.01	1.00	0.90	0.92	0.24	0.74	0.87	0.19
	256	9.20e+00	0.88	0.90	0.89	0.80	0.14	0.36	0.84	0.15
	512	1.03e+01	0.97	0.89	0.84	0.84	0.10	0.10	0.10	0.17
	1024	1.96e+01	1.28	1.18	0.98	0.88	0.11	0.11	0.11	0.84
	2048	1.15e+02	1.86	1.44	1.30	1.21	0.10	0.10	0.11	1.06
	4096	8.18e+02	1.83	1.39	1.25	1.19	1.12	0.85	1.13	0.99
	8192	6.64e+03	1.95	1.47	1.31	1.23	1.17	1.15	1.14	0.99
	16384	5.52e+04	1.90	1.43	1.27	1.19	1.17	1.13	1.10	0.97
H100	64	4.47e+00								
	128	4.48e+00								
	256	4.49e+00								
	512	4.75e+00								
	1024	7.77e+00								
	2048	3.59e+01								
	4096	2.81e+02								
	8192	2.48e+03								
	16384	2.21e+04								
B200	64	5.61e+00	0.90	0.91	0.91	0.91	0.90	0.91	0.91	0.91
	128	5.25e+00	0.84	0.85	0.84	0.84	0.85	0.85	0.84	0.84
	256	5.29e+00	0.85	0.85	0.86	0.85	0.85	0.85	0.85	0.85
	512	5.25e+00	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84
	1024	6.28e+00	0.86	0.76	0.76	0.76	0.76	0.76	0.76	0.76
	2048	1.65e+01	1.27	1.14	1.00	1.00	1.00	1.00	1.00	0.89
	4096	8.98e+01	1.62	1.14	1.08	0.94	0.94	0.92	0.93	0.81
	8192	6.54e+02	1.60	1.18	1.06	0.98	0.92	0.85	0.86	0.82
	16384	5.95e+03	1.63	1.16	1.09	1.03	0.98	0.95	0.94	0.83
RTX5080	64	2.12e+00	0.52	0.52	0.51	0.52	0.52	0.54	0.52	0.52
	128	2.30e+00	0.56	0.56	0.55	0.37	0.37	0.37	0.37	0.38
	256	4.06e+00	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66
	512	6.13e+00	0.99	0.75	0.74	0.74	0.74	0.73	0.71	0.60
	1024	2.67e+01	1.63	1.30	1.18	1.18	1.08	1.08	1.08	0.98
	2048	1.54e+02	1.53	1.17	1.06	0.99	0.96	0.94	0.92	0.81
	4096	1.15e+03	1.81	1.37	1.23	1.15	1.10	1.08	1.06	0.93
	8192	9.05e+03	1.81	1.34	1.16	1.13	1.06	1.04	1.01	0.92
	16384	7.27e+04	1.53	1.08	0.98	0.93	0.89	0.88	0.85	0.75
GB10	64	3.45e+00	0.82	0.79	0.77	0.77	0.78	0.79	0.79	0.79
	128	3.21e+00	0.69	0.52	0.47	0.48	0.50	0.52	0.50	0.52
	256	4.18e+00	0.61	0.65	0.67	0.61	0.65	0.67	0.62	0.63
	512	8.54e+00	1.13	0.97	1.02	0.98	0.88	0.94	0.91	0.83
	1024	3.35e+01	1.25	1.01	0.96	0.86	0.85	0.84	0.82	0.75
	2048	2.18e+02	1.57	1.20	1.10	0.99	0.98	0.98	0.88	0.79
	4096	1.85e+03	1.61	0.67	0.69	0.51	0.57	0.58	0.58	0.48
	8192	1.36e+04	0.54	0.40	0.35	0.33	0.25	0.24	0.31	0.27
	16384	1.07e+05	0.53	0.30	0.35	0.33	0.32	0.31	0.30	0.26

≥1.5× ≥1.0×