Appendix A Square Fp4

GPU	M	cuBLASLt Latency (μs)	cuSPARSELt Speedup Ratio							
			2:4	2:6	2:8	2:10	2:12	2:14	2:16	2:∞
B200	64	8.42e+00	1.37	1.39	1.36	1.51	1.39	1.36	1.38	1.38
	128	8.47e+00	1.37	1.36	1.36	1.51	1.37	1.36	1.36	1.37
	256	8.48e+00	1.37	1.36	1.36	1.37	1.36	1.36	1.36	1.36
	512	8.47e+00	1.36	1.35	1.36	1.35	1.35	1.36	1.35	1.35
	1024	8.46e+00	1.35	1.20	1.22	1.09	1.07	1.11	1.10	1.03
	2048	8.70e+00	0.84	0.82	0.84	0.84	0.83	0.70	0.84	0.84
	4096	1.86e+01	0.81	0.65	0.65	0.53	0.53	0.59	0.60	0.53
	8192	9.31e+01	0.81	0.54	0.57	0.46	0.48	0.50	0.49	0.42
	16384	6.83e+02	0.75	0.54	0.50	0.43	0.39	0.37	0.40	0.35
RTX5080	64	4.20e+00	1.03	1.02	1.02	1.02	1.02	1.02	1.02	1.02
	128	4.17e+00	1.01	1.02	1.01	1.02	1.02	1.01	1.02	1.01
	256	4.19e+00	1.02	1.02	1.01	1.01	1.01	1.01	1.01	1.01
	512	4.21e+00	0.69	0.69	0.69	0.68	0.68	0.68	0.69	0.68
	1024	6.20e+00	1.01	0.76	0.76	0.76	0.76	0.76	0.76	0.76
GB10	64	6.17e+00	1.05	1.08	1.05	1.06	1.07	1.10	1.09	1.12
	128	6.20e+00	1.07	1.00	1.02	1.01	1.06	1.01	1.07	1.07
	256	6.82e+00	1.11	1.11	1.10	1.11	1.10	1.11	1.11	1.10
	512	6.21e+00	1.01	1.00	0.98	0.77	0.77	0.77	0.79	0.77
	1024	1.03e+01	0.84	0.64	0.71	0.63	0.56	0.63	0.63	0.63
	2048	3.29e+01	0.76	0.48	0.55	0.50	0.46	0.40	0.48	0.42
	4096	2.41e+02	0.78	0.54	0.55	0.42	0.41	0.44	0.49	0.43
	8192	1.70e+03	0.73	0.48	0.50	0.40	0.44	0.44	0.43	0.38

≥1.5× ≥1.0×