

Appendix A Square Fp8

GPU	M	cuBLASLt Latency (μ s)	cusPARSElt Speedup Ratio							
			2:4	2:6	2:8	2:10	2:12	2:14	2:16	2: ∞
RTX4090	64	1.13e+01	1.12	1.18	1.06	1.22	0.40	0.40	0.65	0.40
	128	1.17e+01	1.30	1.36	1.11	1.31	0.33	0.33	1.23	0.34
	256	1.10e+01	1.20	1.10	1.22	1.07	0.25	0.25	1.14	0.25
	512	1.22e+01	1.31	1.16	1.21	1.27	0.20	0.21	1.08	0.19
	1024	1.24e+01	1.03	0.90	1.01	0.97	0.12	0.11	0.24	0.11
	2048	5.78e+01	1.75	1.36	1.26	1.14	0.10	0.10	0.10	0.12
	4096	4.20e+02	1.87	1.40	1.27	1.18	0.71	1.10	1.08	0.95
	8192	3.42e+03	1.99	1.50	1.36	1.26	1.20	1.13	1.15	1.03
	16384	2.84e+04	2.08	1.51	1.37	1.28	1.22	1.20	1.18	1.04
H100	64	4.61e+00	0.95	0.89	0.91	0.93	0.90	0.90	0.89	0.90
	128	4.70e+00	0.93	0.90	0.90	0.89	0.47	0.85	0.85	0.87
	256	4.61e+00	0.91	0.83	0.85	0.82	0.82	0.82	0.81	0.83
	512	4.62e+00	0.82	0.77	0.76	0.75	0.72	0.75	0.75	0.74
	1024	6.36e+00	0.73	0.61	0.65	0.62	0.59	0.57	0.62	0.60
	2048	2.12e+01	1.07	0.84	0.80	0.72	0.71	0.75	0.74	0.68
	4096	1.56e+02	1.53	1.10	0.94	0.86	0.84	0.82	0.83	0.73
	8192	1.41e+03	1.54	1.15	1.05	0.85	0.93	0.86	0.92	0.79
	16384	1.28e+04	1.73	1.12	1.08	1.02	1.08	1.02	1.03	0.91
B200	64	5.97e+00	0.96	0.96	0.97	0.97	0.96	0.97	0.96	0.96
	128	5.58e+00	0.90	0.90	0.90	0.90	0.90	0.90	0.92	0.90
	256	5.67e+00	0.91	0.91	0.91	0.91	0.91	0.91	0.93	0.91
	512	5.64e+00	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91
	1024	5.64e+00	0.91	0.79	0.90	0.87	0.73	0.68	0.83	0.74
	2048	1.04e+01	1.00	0.84	0.84	0.84	0.75	0.84	0.84	0.84
	4096	4.55e+01	1.51	1.06	1.10	0.90	0.95	0.96	0.96	0.83
	8192	3.40e+02	1.72	1.14	1.16	0.97	0.88	0.82	1.00	0.86
	16384	3.03e+03	1.85	1.07	1.07	1.00	0.91	0.88	0.91	0.79
RTX5080	64	3.34e+00	0.81	0.81	0.80	0.81	0.81	0.81	0.81	0.80
	128	3.32e+00	0.80	0.80	0.81	0.80	0.80	0.80	0.80	0.80
	256	3.37e+00	0.82	0.55	0.55	0.55	0.55	0.55	0.55	0.54
	512	4.19e+00	0.68	0.68	0.68	0.68	0.66	0.67	0.68	0.65
	1024	1.44e+01	1.40	1.00	1.00	0.88	0.89	0.88	0.88	0.81
	2048	7.99e+01	1.56	1.17	1.08	0.93	0.87	0.93	0.95	0.85
	4096	5.92e+02	1.83	1.31	1.20	1.11	1.07	1.03	1.03	0.91
	8192	4.56e+03	1.76	1.33	1.17	1.10	1.05	1.03	1.01	0.88
	16384	3.64e+04	1.74	1.31	1.17	1.10	1.07	1.03	1.00	0.88
GB10	64	5.16e+00	0.96	0.95	0.93	0.96	0.96	0.96	0.95	0.97
	128	5.03e+00	1.00	1.03	1.03	0.98	0.96	0.97	0.97	1.04
	256	3.86e+00	0.75	0.63	0.63	0.63	0.62	0.63	0.62	0.63
	512	6.23e+00	0.98	0.76	0.77	0.76	0.75	0.75	0.76	0.75
	1024	1.98e+01	1.07	0.88	0.80	0.75	0.76	0.75	0.74	0.64
	2048	1.02e+02	1.19	0.95	0.84	0.81	0.74	0.76	0.73	0.65
	4096	7.32e+02	1.21	0.92	0.84	0.79	0.77	0.73	0.71	0.63
	8192	6.01e+03	1.26	0.97	0.85	0.80	0.75	0.75	0.73	0.64
	16384	5.37e+04	1.41	0.89	0.90	0.83	0.73	0.72	0.76	0.66

■ $\geq 1.5 \times$ ■ $\geq 1.0 \times$