

Appendix D Decode Fp8 (Page 1/2)

GPU	Model	Concurrency M	cuBLASLt Throughput (token/s)	cuSPARSELt Speedup Ratio			
				2:4	2:6	2:8	2:10
RTX4090	Llama3.2-1B	64	9.78e+03	1.05	0.86	0.87	0.86
		128	1.37e+04	1.00	0.96	0.97	0.96
		256	1.88e+04	1.13	1.01	1.06	1.03
		512	1.97e+04	1.04	0.99	0.98	0.97
	BitNet-2B	64	7.54e+03	0.90	0.86	0.83	0.80
		128	9.96e+03	1.12	1.04	1.07	1.07
		256	1.54e+04	1.09	0.98	0.97	0.96
		512	1.52e+04	1.07	1.01	1.00	0.95
	Llama3.2-3B	64	5.86e+03	0.98	0.89	0.86	0.82
		128	8.73e+03	1.06	0.99	0.96	0.96
		256	1.15e+04	1.13	1.05	1.02	0.96
		512	1.10e+04	1.17	1.07	1.02	1.00
	Qwen2.5-7B	64	3.87e+03	1.15	0.99	0.93	0.98
		128	6.34e+03	1.20	1.05	0.97	0.96
		256	8.20e+03	1.20	1.05	1.01	0.99
		512	7.83e+03	1.25	1.13	1.07	1.02
	Qwen2.5-14B	64	1.56e+03	1.84	1.08	0.58	
		128	2.28e+03	1.88	1.11	0.77	0.39
		256	2.26e+03	1.69	1.11	0.56	
		512	1.32e+03	2.83	1.17		
H100	Llama3.2-1B	64	1.49e+04	1.00	0.96	0.84	0.94
		128	2.12e+04	0.96	0.95	0.95	0.95
		256	3.14e+04	0.96	0.94	0.95	0.90
		512	3.19e+04	0.99	0.99	0.98	0.97
	BitNet-2B	64	9.65e+03	0.98	0.95	0.92	0.90
		128	1.44e+04	1.04	1.00	0.99	0.99
		256	2.18e+04	1.00	0.96	0.96	0.94
		512	2.28e+04	1.02	0.98	0.97	0.94
	Llama3.2-3B	64	8.64e+03	1.00	0.93	0.93	0.90
		128	1.35e+04	1.03	1.01	0.99	0.96
		256	1.95e+04	1.00	0.92	0.93	0.90
		512	2.03e+04	1.01	0.94	0.95	0.90
	Qwen2.5-7B	64	7.00e+03	1.09	0.98	0.94	0.91
		128	1.12e+04	1.10	0.99	1.00	0.98
		256	1.50e+04	1.13	1.01	1.00	0.95
		512	1.61e+04	1.09	0.98	0.96	0.92
	Qwen2.5-14B	64	3.78e+03	1.30	1.12	1.11	1.08
		128	6.22e+03	1.21	1.03	1.06	1.01
		256	8.89e+03	1.13	1.01	0.96	0.91
		512	8.66e+03	1.18	1.04	1.02	0.98
	Llama3.2-1B	64	2.05e+04	1.00	0.98	0.86	0.97
		128	2.93e+04	1.00	0.98	0.96	0.97
		256	4.49e+04	1.01	0.86	0.87	0.83
		512	4.88e+04	1.00	0.97	0.98	0.97
	BitNet-2B	64	1.34e+04	0.88	0.85	0.84	0.84
		128	2.04e+04	1.03	0.99	0.99	0.98
		256	3.06e+04	0.99	0.97	0.95	0.95
		512	3.59e+04	1.02	0.98	0.98	0.97
	Llama3.2-3B	64	1.40e+04	0.94	0.91	0.90	0.89
		128	2.08e+04	1.02	0.98	0.98	0.97
		256	3.28e+04	1.00	0.97	0.93	0.95
		512	3.48e+04	1.03	1.00	0.98	0.97
	Qwen2.5-7B	64	1.19e+04	1.11	1.02	1.00	0.99
		128	1.88e+04	1.06	1.00	0.98	0.88
		256	2.76e+04	1.05	0.97	0.99	0.93
		512	3.08e+04	1.07	1.00	0.98	0.97
	Qwen2.5-14B	64	7.35e+03	1.16	1.06	1.04	1.02
		128	1.22e+04	1.08	1.01	0.99	0.99
		256	1.81e+04	1.09	0.98	0.94	0.95
		512	1.89e+04	1.15	1.06	1.03	1.03
RTX5080	Llama3.2-1B	64	1.46e+04	1.11	1.02	0.99	0.94
		128	2.08e+04	1.05	1.00	0.98	0.97
		256	2.14e+04	1.08	1.06	1.04	0.97
		512	2.14e+04	1.10	1.03	1.01	0.98
	BitNet-2B	64	9.45e+03	0.98	0.89	0.86	0.82

 ≥1.5x
 ≥1.0x

Appendix D Decode Fp8 (Page 2/2)

GPU	Model	Concurrency	cuBLASLt	cuSPARSELt Speedup Ratio			
		M	Throughput (token/s)	2:4	2:6	2:8	2:10
		128	1.38e+04	1.05	0.98	0.94	0.93
		256	1.46e+04	1.13	1.04	1.02	0.99
		512	1.30e+04	1.16	1.05	1.03	0.99
	Llama3.2-3B	64	7.81e+03	1.16	0.99	0.97	0.94
		128	1.14e+04	1.12	1.03	0.99	0.96
		256	1.17e+04	1.17	1.08	1.04	1.00
		512	9.99e+03	1.18	1.08	1.03	0.99
	Qwen2.5-7B	64	5.09e+03	1.23	1.05	0.99	0.94
		128	7.94e+03	1.28	1.11	1.05	1.00
		256	6.67e+03	1.61	1.41	1.13	0.99
		512	5.89e+03	1.43	1.18	1.08	0.99
	Qwen2.5-14B	64					
		128					
		256					
		512					
GB10	Llama3.2-1B	64	4.42e+03	1.12	0.97	0.94	0.92
		128	6.65e+03	1.05	0.99	0.96	0.92
		256	8.20e+03	1.05	0.97	0.97	0.93
		512	8.65e+03	1.02	0.95	0.95	0.92
	BitNet-2B	64	2.52e+03	1.15	0.98	1.00	0.97
		128	3.93e+03	1.15	1.01	0.99	0.97
		256	5.00e+03	1.10	1.00	1.00	0.97
		512	5.18e+03	1.06	0.99	0.98	0.96
	Llama3.2-3B	64	2.11e+03	1.18	1.02	1.03	0.93
		128	3.15e+03	1.17	1.06	1.05	0.97
		256	4.02e+03	1.11	1.04	1.03	0.97
		512	4.20e+03	1.08	1.02	0.99	0.95
	Qwen2.5-7B	64	1.26e+03	1.33	1.07	1.05	0.97
		128	2.11e+03	1.29	1.05	1.04	1.00
		256	2.91e+03	1.20	1.01	1.02	0.98
		512	2.94e+03	1.19	1.02	1.00	0.96
	Qwen2.5-14B	64	6.49e+02	1.34	1.08	1.08	1.03
		128	1.10e+03	1.34	1.07	1.06	1.00
		256	1.48e+03	1.25	1.07	1.07	1.02
		512	1.53e+03	1.22	1.06	1.01	1.00