

Appendix B Model Kernel Fp4

GPU	Model	M	cuBLASLt Latency (μs)	cuSPARSELt Speedup Ratio							
				2:4	2:6	2:8	2:10	2:12	2:14	2:16	2:∞
GB10	Llama3.2-1B	64									
		128									
		256									
		512									
		1024									
		2048									
		4096									
		8192									
	Llama3.2-3B	64									
		128									
		256									
		512									
		1024									
		2048									
	Qwen2.5-7B	4096									
		64									
		128									
		256									
		512									
		1024									
	Qwen2.5-14B	2048									
		4096									
		8192									
		64									
		128									
		256									
		512									
		1024									
		2048									
		4096									
		8192									
		16384									

≥1.5× ≥1.0×