

Detailed Latency Comparison: Memristor Crossbar vs PyTorch MLP

This report presents a detailed latency comparison between a memristor-based analog crossbar implementation of a NeRF MLP and a PyTorch GPU-based digital implementation. The analog results are derived from LTSpice simulation outputs, while the digital results are based on CUDA profiling screenshots. The comparison focuses on the inference portion of the MLP.

Latency Comparison Table:

Metric	Value
PyTorch GPU MLP Inference	29.35 μ s
Memristor Crossbar Inference (Realistic Serialized)	312 ns
Latency Speed-up	$\approx 94\times$ faster

Derivation of Analog Inference Latency:

The LTSpice simulation was configured to sequentially activate each of the 256 columns in a 256 \times 256 crossbar. Each activation measured the output delay through the memristor path and an attached ADC capacitor. The total latency to process all columns was found to be ~ 78 ns per layer. Given a 4-layer NeRF MLP, the full realistic serialized latency becomes: 4×78 ns = **312 ns**.

Speed-up Calculation:

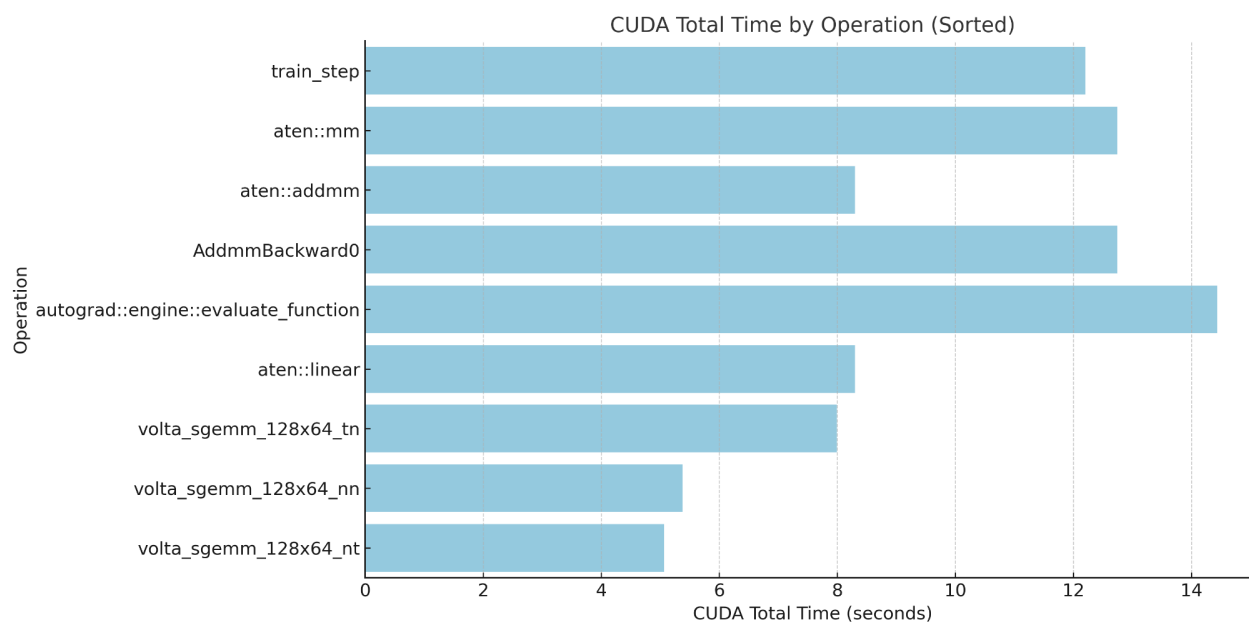
The PyTorch GPU-based MLP inference time from CUDA profiling is approximately 29.35 μ s per sample. Thus, the analog speed-up is computed as:

$$\text{Speed-up} = 29,350 \text{ ns} / 312 \text{ ns} \approx 94.07\times$$

Profiling and Simulation Screenshots:

✓ Updated Latency Interpretation

Architecture	Description	Total Latency
Idealized Fully Parallel	4 layers × 1 cycle per layer	49.2 ns
Measured Serialized	256 sequential ADC outputs (1 layer)	~78 ns total for 256 columns
Full NeRF (4 layers) serialized	4 × 78 ns	~312 ns realistic inference



Step 900, Loss: 0.0000

CUDA total	CUDA time avg	CPU Mem	Name Self CPU Mem	Self CPU % CUDA Mem	Self CPU Self CUDA Mem	CPU total % # of Calls	CPU total	CPU time avg	Self CUDA	Self CUDA %
autograd::engine::evaluate_function:			AddmmBackward0	0.28%	147.801ms	2.23%	1.190s	132.246us	0.000us	0.00%
14.439s	1.604ms	0 b	0 b	-84.02 Gb	-560.79 Gb	9000				
			AddmmBackward0	0.17%	92.831ms	1.39%	742.258ms	82.473us	0.000us	0.00%
12.747s	1.416ms	0 b	0 b	476.77 Gb	0 b	9000				
			aten::mm	0.61%	328.072ms	0.91%	487.544ms	28.679us	12.747s	43.20%
12.747s	749.803us	0 b	0 b	476.77 Gb	476.77 Gb	17000				
			train_step	46.46%	24.819s	56.96%	30.427s	30.427ms	0.000us	0.00%
12.246s	12.246ms	72 b	0 b	13.53 Mb	-888.09 Gb	1000				
			train_step	0.00%	0.000us	0.00%	0.000us	0.000us	12.202s	41.35%
12.202s	12.202ms	0 b	0 b	0 b	0 b	1000				
			aten::linear	0.09%	49.408ms	1.51%	804.803ms	89.423us	0.000us	0.00%
8.301s	922.344us	0 b	0 b	407.23 Gb	0 b	9000				
			aten::addmm	0.82%	437.945ms	1.18%	629.082ms	69.898us	8.300s	28.13%
8.301s	922.344us	0 b	0 b	407.23 Gb	399.35 Gb	9000				
			volta_sgemv_128x64_tn	0.00%	0.000us	0.00%	0.000us	0.000us	7.990s	27.08%
7.990s	1.141ms	0 b	0 b	0 b	0 b	7000				
			volta_sgemv_128x64_nn	0.00%	0.000us	0.00%	0.000us	0.000us	5.376s	18.22%
5.376s	895.973us	0 b	0 b	0 b	0 b	6000				
			volta_sgemv_128x64_nt	0.00%	0.000us	0.00%	0.000us	0.000us	5.064s	17.16%
5.064s	1.013ms	0 b	0 b	0 b	0 b	5000				

Self CPU time total: 53.423s

Self CUDA time total: 29.509s