Theoretical Benchmarking Summary: NeRF MLP and Rendering Acceleration

Objective

To estimate the latency and performance gains of offloading parts of a simplified Neural Radiance Fields (NeRF) implementation to hardware, specifically: the MLP inference block (which dominates runtime) and optionally, the volume rendering integration loop. We compare three configurations:

1. Software-only (PyTorch on CPU/GPU)
2. Digital hardware accelerator (e.g., FPGA, TPU)
3. Analog hardware accelerator (e.g., memristor crossbar)

NeRF Workload Overview

Image resolution: 32x32 = 1024 rays
Samples per ray: 64
Total sample points: 65,536
MLP input size: 90 (63 pos + 27 dir)
MLP structure: 5x256-layer (pos), 2x128-layer (dir)
Output per sample: RGB (3) + sigma (1) = 4 values
Goal: Per-frame latency and speedup estimation

Software Runtime Estimation

Based on empirical profiling and literature:
- Forward-only MLP (PyTorch GPU): ~1015 ms
- Volume rendering loop: ~15 ms
- Total forward time per frame: ~2530 ms

MLP complexity estimate:
- 200K FLOPs per sample
- Total: ~13.1 GFLOPs/frame

Hardware Acceleration Candidates

1. Digital MLP Accelerator (TPU or FPGA)
- Pipelined matrix-vector multipliers, ~520 TFLOPs/s
- Latency: ~1.3 ms MLP + 0.20.5 ms transfer
- Total MLP latency: ~1.52 ms

Theoretical Benchmarking Summary: NeRF MLP and Rendering Acceleration

2. Analog MLP Accelerator (Memristor Crossbar)

- In-memory compute with ultra-fast MACs
- MLP: ~100200 us, transfer + ADC/DAC: ~100500 us
- Total latency: ~0.20.3 ms

Optional: Hardware Volume Rendering

Volume rendering operations:

- Cumulative product for transmittance
- Weighted sum for RGB integration
- Total latency in hardware: ~100 us (pipelined)

Latency and Speedup Summary Table

Configuration	MLP Latency	Rendering Latency	Total Latency	Estimated Speedup
Software (PyTorch)	~10 ms	~15 ms	~25 ms	1x (baseline)
Digital HW	~1.3 ms	~1 ms	~2.3 ms	~10x
Analog HW	~0.2 ms	~0.1 ms	~0.3 ms	~80x

Summary of Findings

- The MLP forward pass is the dominant bottleneck in NeRF inference.
- Digital accelerators (TPU, FPGA) offer ~10x speedup.
- Analog accelerators (memristors) offer up to ~80100x speedup.
- Volume rendering is less intensive but still benefits from acceleration.
- Optimal software/hardware boundary is: software = ray sampling & rendering loop, hardware = MLP inference.