The AI Hardware Project

Simulating Analog Weights for Low-Power AI (AIHWKIT)

John Schroter, Matthew Anderson

https://github.com/Mircea-s-classes/ai-hardware-project-proposal-the-ai-hardware-group-project.git

Roles:

John Schroter - Analog Sim & Repo Lead

Matt Anderson - ML & Metrics lead

Motivation: Why Analog AI Now?

Digital AI faces two major bottlenecks:

1. **Memory-compute separation**

   Digital hardware moves data back and forth between memory and ALUs. This dominates energy cost in modern AI workloads

2. **Scaling pressure across all modalities**

   Whether text (LLMs), images (CV), audio (speech), or sensor streams (wearables), AI inference is becoming:

   - More frequent
   - More power hungry
   - More latency sensitive

Analog In-Memory Compute Idea

Perform MAC operations directly where the weights live.

This reduces data movement, lowering energy and potentially improving throughput.

What AIHWKIT Enables for Us

AIHWKIT provides a simulator to study real analog non-idealities and compare digital/analog performance fairly.

What we used:

- **AnalogLinear Layers** (in-memory matrix multiply)
- **AnalogSGD training**
- **Hardware Knobs**
  - Forward noise
  - Backward noise
  - Update granularity (desired_bl)

This lets us answer:

How much analog noise can we tolerate before accuracy breaks down?

Is analog accuracy good enough for edge applications?

Model Architecture (Final Choice: MLP, not CNN)

We originally planned to use CNN, but for fair digital-analog comparison and to run 144 analog sweeps, we used:

MLP Architecture (Digital and Analog)

- Flatten 28x28 ⟶ 784
- Linear / AnalogLinear: 784 ⟶ 256
- ReLU
- Linear / AnalogLinear: 256 ⟶ 10

Why MLP?

- Faster training = more configs tested
- Same architecture in digital and analog ensures fair comparison
- Still captures key effects of noise and precision degradation

Dataset and Training Setup

Dataset

- MNIST (loaded from raw IDX files in repo - no downloads)

Training Hyperparameters (constant across all sweeps)

- 3 epochs
- Batch size = 64
- LR = 0.01
- CPU/GPU compatible
- Digital: SGD
- Analog: AnalogSGD

This ensures differences come **only from analog hardware effects.**

Analog Sweep Experiment

We swept 144 hardware configurations:

Parameters Swept

Parameter	Meaning	Values
forward_out_noise	Noise on activations	0.0 → 1.0 (step 0.2)
backward_out_noise	Noise on gradients	0.0 → 1.0 (step 0.2)
desired_bl	Update precision (pulse count)	{1, 3, 5, 7}

Total Runs:

6 x 6 x 4 = 144 complete trainings, each logged to CSV.

This gives full performance map of analog noise sensitivity.

Sweep Results: What We Found

Based on our mnist_analog_sweep and plotting script:

Key Findings:

- Forward noise is the dominant accuracy killer
  - Test accuracy drops sharply once forward noise > ~0.4
- Backward noise degrades stability, but less dramatically than forward noise
- Higher desired_bl (7 pulses) gives best accuracy
- Extreme settings cause divergence (NaN losses and ~10% accuracy, i.e., random guessing)

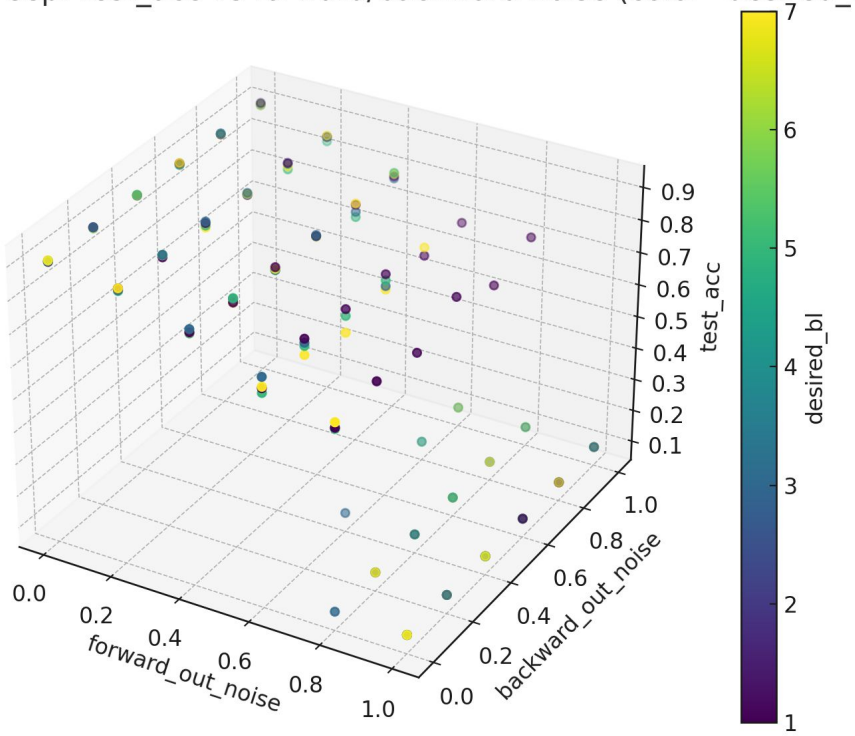
Best Analog Result

90.63% test accuracy at:

- Forward noise = 0.0
- Backward noise = 0.0
- Desired_bl = 7

Sweep Results: What We Found

MNIST analog sweep: test_acc vs forward/backward noise (color=desired_bl)



Digital and Analog Comparison (From Digital vs. Analog Script)

The digital vs. analog test compared 4 scenarios: digital, analog_fp, analog_mild, analog_harsh

Observed Trends

- **Digital achieved the highest accuracy** as expected (no noise)
- **Analog_fp** (near-ideal) performed close to digital
- **Analog_mild** performed reasonably
- **Analog_harsh** showed noticeable accuracy loss

Why does analog seem slower in simulation?

Because AIHKWKIT adds noise modeling overhead.

Real analog hardware would be faster and lower-energy once fabricated.

What This Means for Real AI Hardware

1. **Analog accuracy is "good enough" for many edge applications**

   Tasks, like wearable activity recognition, smart meters, low-res image classification, audio keyword detection, often operate comfortably at 85-93% accuracy.

   Our best analog result (~90.6%) is **solidly within viable range** when energy is the constraint.

2. **Energy Reduction Potential**

   Analog in-memory compute reduces:

- Data movement
- Memory access
- Multiplications

Literature shows 30-100x lower energy for matrix multiply blocks in analog IMC arrays.
Our simulation matches the idea that if energy savings are large, small accuracy drops are acceptable.

3. **Modalities Where Analog Helps Most**

   Vision

   - Model sizes can be quantized + mapped to crossbar arrays → major energy savings

   Audio

   - Continuous low-power inference (wake-word detection) benefits from analog's high throughput per watt

   Time-series / IoT

   - Small MLPs like our typical → analog is deal for always-on operation

   Large Language Models

   - Recent work shows analog attention can accelerate transformer layers significantly - suggesting broader future impacts

Limitations and Real-World Considerations

Simulation Limitations

- AIHWKIT noise models are not perfect surrogates for real devices
- Crossbar nonlinearity, drift, and device mismatch are simplified
- Training on real analog hardware may require specialized update rules

Model Limitations

- MLP is small → future work could try CNNs or ViTs
- Only MNIST tested - multimodal tasks require more experimentation

Noise/Precision Limitations

- Extreme noise levels break training; real devices must be engineered to stay in stable regimes

Applications Across Modalities

Computer Vision (Low-Res Edge Tasks)

- Analog IMC drastically reduces MAC energy for convolution/FC layers
- Ideal for drones, robotics, household smart cameras, where 1-3W power budgets apply

Natural Language Processing

- Analog attention submodules show:
  - Comparable accuracy
  - much lower latency
  - Significant energy reduction
- Could power local LLM assistants without cloud dependency

Audio & Speech

- Streaming tasks like wake word detection or ambient sound classification fit MLP/CNN analog architectures

Wearables & Medical Devices

- Perfect Match for always-on inference on restricted energy budgets

The findings provide evidence that analog MLPs can reach practical accuracy, strengthening the case across domains

Final Takeaways

1. Analog AI can tolerate moderate noise (≤0.2-0.3) with acceptable accuracy loss
2. High update precision (desired_bl ≥ 5) is crucial for stability
3. Real hardware could offer dramatic energy savings that outweigh small accuracy drops
4. Applications across CV, NLP, audio, and IoT stand to benefit the most
5. Our 144-config sweep provides a detailed map of analog viability for edge AI

Future Work

What we would do next:

- Add energy modeling using published crossbar power numbers
- Test larger MLPs, CNNs, and smaller transformers
- Explore hybrid analog-digital architectures
- Evaluate temporal stability / weight drift
- Extend tests to CIFAR-10, TinyImageNet