Predictive Cost Modelling & Empirical Characterisation of ZKML Circuits

Huy Quang Le[1], Wan Ki (Thomas) Wong[2], Dr. Rik Sarkar[2]

[1]School of Mathematics, University of Bristol [2]ANGLE Lab, School of Informatics, University of Edinburgh

EPSRC Centre for Doctoral Training in Machine Learning Systems

1. What is ZKML & The Research Challenge

Zero-Knowledge Machine Learning (ZKML) allows someone to mathematically prove they ran a specific AI model correctly, *without* revealing the secret model weights or the private input data.

While this is a breakthrough for verifiable, privacy-preserving AI, deploying it faces a major barrier: the highly unpredictable computational cost of generating these proofs. Traditional cost estimations assume linear additivity, where total prover time equals the sum of individual layer latencies.

By isolating mathematical operations using the **EZKL/Halo2 backend**, this project demonstrates that ZKML proving cost is strictly **non-additive**. Instead, time is governed by cryptographic matrix constraints like discrete grid bounds and Fast Fourier Transform (FFT) domain complexity.

2. Methodology & Cost Equation

All benchmarks and cluster sweeps were executed on the **University of Edinburgh's Eddie HPC Cluster**.

- **Data Generation:** Executed thousands of compiled ZK circuits across 8 distinct neural network architectures.
- **Mathematical Modelling:** Extracted pure circuit features (grid density, lookup spans) and applied Non-Negative Least Squares (NNLS) to isolate the underlying $\mathcal{O}(N \log N)$ scaling laws, yielding the following predictive formula:

$$\widehat{T}_{\text{prove}} \approx (2.94\text{e-}8) D \log_2 D + (1.51\text{e-}4) A_{\text{total}} + (9.77\text{e-}4) L_{\text{span}} + (6.63\text{e-}2) C_{\text{size}}$$

Where $\mathbf{D}$*: FFT Domain Size,* $\mathbf{A}_{total}$*: Active Grid Assignments,* $\mathbf{L}_{span}$*: Non-linear Lookup Span, and* $\mathbf{C}_{size}$*: Static Parameter Constraints.*

- **Validation:** Achieved a Mean Absolute Percentage Error (MAPE) of just **12.69%**, maintaining a highly stable $13.60\% \pm 4.10\%$ error rate under Repeated 5-Fold Cross-Validation.

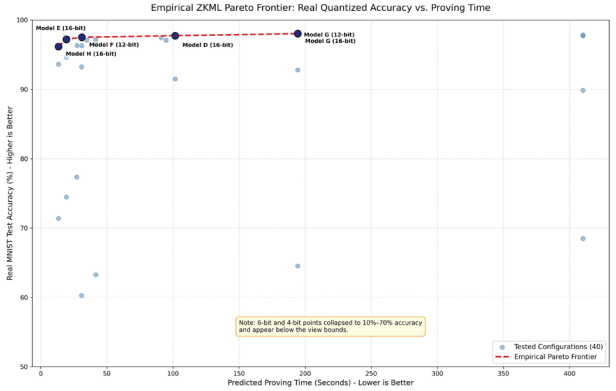
From Axioms to Applications

As a first-year mathematics student, the proofs I have studied are traditionally rooted in axioms yielding absolute certainty. Discovering the probabilistic and cryptographic nature of Zero-Knowledge Proofs (ZKPs) for the first time offered a completely new perspective.

3. The Empirical Pareto Frontier

To demonstrate the practical application of the mathematical cost model, I constructed an empirical Pareto frontier mapping real-world test accuracy against predicted proving time on the MNIST dataset.

By evaluating 8 different architectures across 5 bit-width quantisation scales (4-bit to 16-bit), I identified 40 unique ZKML configurations.



Definitive Architectural Guidelines Derived:

1. **The 12-Bit Sweet Spot:** Because non-linear operations (GELU/Sigmoid) require physical lookup tables that scale by 2^{bits}, reducing precision from 16-bit to 12-bit slashed proving time by **66%** with zero accuracy loss.
2. **Depth Beats Width:** In standard ML, wide networks compute quickly on GPUs. In ZKML, wide layers create dense, costly mathematical constraints. A deep, narrow network achieved 97.2% accuracy in **19.46s**, whereas a shallow, wide network took **41.67s**.
3. **The Quantisation Cliff:** Standard post-training quantisation holds strong down to 8 bits, making it a reliable baseline. However, crossing into 6-bit and 4-bit models caused a catastrophic accuracy collapse (dropping to 10%–30%).
4. **Convolutional Overhead:** Spatial convolutions incur massive mathematical penalties due to dense cross-multiplication. A medium CNN took $\sim 10\times$ longer to prove than a Deep Linear Network (MLP) for less than a 0.8% gain in accuracy.

4. Zero to One: Skills Developed

Prior to this placement, I had minimal exposure to research literature, terminal environments, or version control. Over the course of 8 weeks, I underwent rapid technical development:

- **High-Performance Computing:** Learned to navigate Linux terminals, utilise GitHub for version control, and execute complex PyTorch workflows on the Eddie cluster.
- **Informatics Foundations:** Expanded my theoretical knowledge by completing university modules in Data Science, Cryptography, and Machine Learning.
- **Academic Communication:** Presented my findings to established academics whose feedback improved my model and challenged me to think more deeply about the subject.

5. The Edinburgh Experience

Coming to the University of Edinburgh made me realise there are truly "levels" to academic research. From attending visiting lectures and seeing theoretical research implemented in the real world, to working alongside industry leaders, the environment was profoundly inspiring.

The welcoming culture of the School of Informatics, from engaging in group discussions to eating out, visiting the Forth Bridge, and watching the World Cup together, made this an unforgettable summer experience.

6. Future Aspirations & Acknowledgments

Future Aspirations:
Working alongside other incredible interns and researchers has given me a strong ambition to aim much higher upon returning to Bristol. In the immediate term, I am continuing this project with the ANGLE Lab to publish our benchmarking framework at an upcoming conference. To scale this ongoing research, I plan to leverage the University of Bristol's Isambard-AI supercomputer facility.

Acknowledgments:
I would like to express my deep gratitude to my supervisors, **Wan Ki (Thomas) Wong** and **Dr. Rik Sarkar**, for their invaluable guidance.

This project was made possible by the funding and support of the **EPSRC Centre for Doctoral Training in Machine Learning Systems**.