An Analog Computing AI Chip Architecture for Scalable and Energy-Efficient Inference

Shangqiu Li

(David)

All figures are student produced unless otherwise stated.

Background

AI development constrained by energy efficiency:

Rising Energy Demands:

1. Companies like Google, Microsoft, and Amazon are investing in whole power plants for AI
2. AI datacenters are projected to consume 652 TWh annually (equivalent to energy of 65 million homes) by 2030 [1]
3. A 10% energy efficiency increase reduces 45 million tons of CO2 and $20 billion annually in operation costs. [1]

Performance Hindered:

1. "Dark Silicon" Issue: inefficiencies lead to an increasing area in a chip is turned off with thermal throttling (50%-80%), limiting performance of the whole chip. [4]

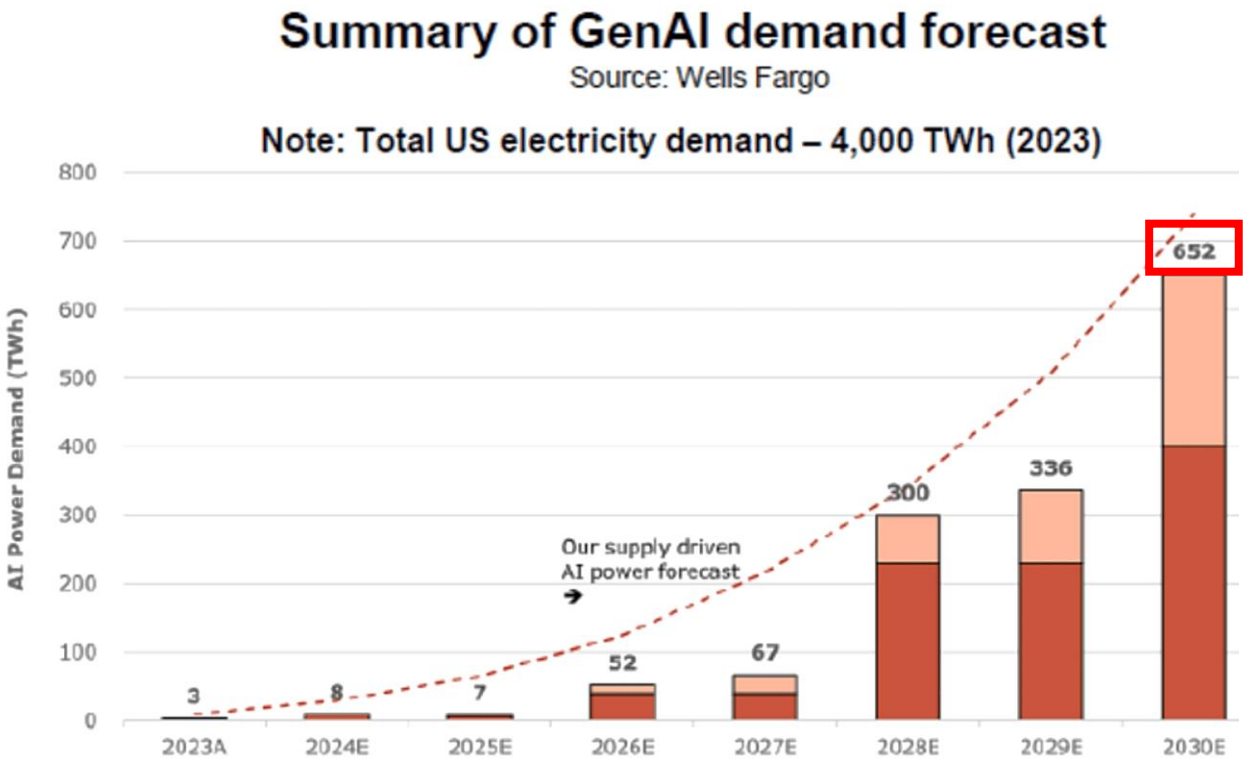
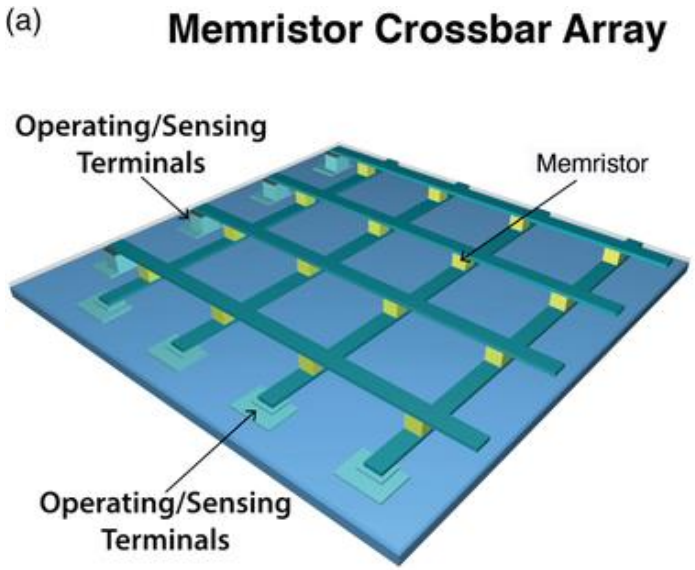


Fig. 1. AI energy demand forecast. (reproduced from [1])

Chip energy efficiency will define the evolution of AI industry

State of the Art Solutions

1. **Traditional Digital AI chip technology (GPUs):**
   - **Scalable, widely adopted**
   - **Energy inefficiency:** inherent digital switching losses
   - **Limited improvements:** Improvements from shrinking transistors are diminishing due to physical limits
2. **Shift Towards Analog Computing:**
   - **Strong research interest**: HP, IBM, MIT, Stanford
   - **Highly efficient**: Digital uses 0s and 1s; analog uses continuous values for more efficient operations.
   - **Types**: Neuromorphic, optical, memristor-crossbar, etc.
3. **Memristor-Crossbar Analog Computing:**
   - **Uniquely Easy Adoption**: Unlike other analog computing technologies, memristor computation allows existing AI models for seamless integration [5] (Mythic Inc., IBM)



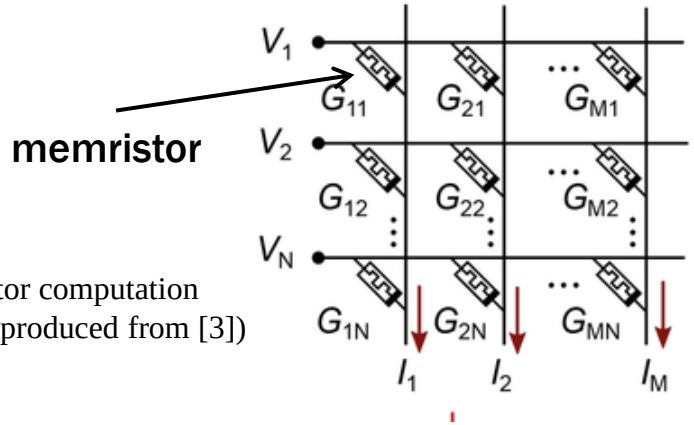


Fig. 4. Memristor computation architecture. (reproduced from [3])

Current State of the Art Solutions

1. **Traditional Digital AI chip technology (GPUs):**
   - **Scalable, widely adopted**
   - **Energy inefficiency:** inherent digital switching losses
   - **Limited improvements:** Improvements from shrinking transistors are diminishing due to physical limits [4]
2. **Shift Towards Analog Computing:**
   - **Strong research interest**: HP, IBM, MIT, Stanford
   - **Highly efficient**: Digital uses 0s and 1s; analog uses continuous values for more efficient operations.
   - **Types**: Neuromorphic, optical, memristor-crossbar, etc.
3. **Memristor-Crossbar Analog Computing:**
   - **Uniquely Easy Adoption**: Unlike other analog computing technologies, memristor computation allows existing AI models for seamless integration [5] (Mythic Inc., IBM)
   - **Efficient Computation:** grid parallel computation
   - **Parameter scalability issues**: AI parameters are stored in its analog resistance. Storage reliability issues and slow parameter update limits it to small AI models. [6]

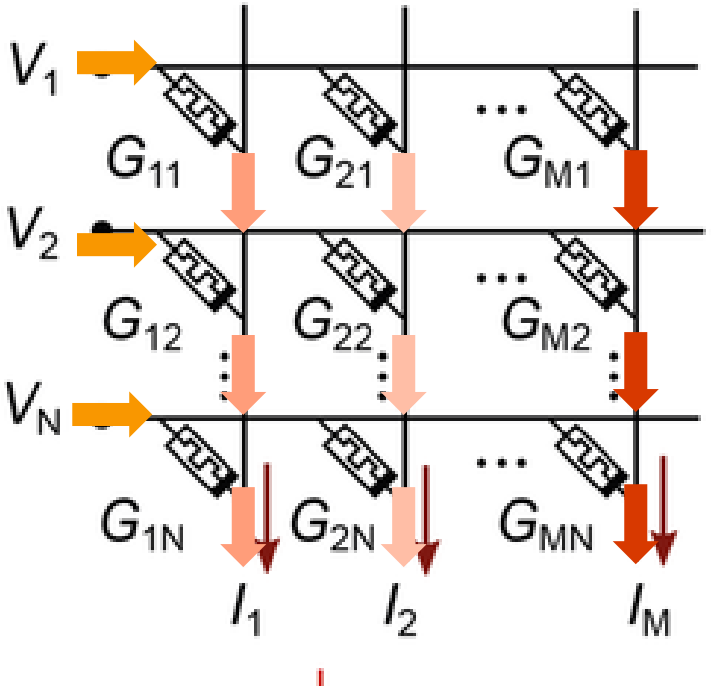


Fig. 5. Memristor computation architecture. (reproduced from [3])

Gap: Digital computation is limited by efficiency while analog computation is limited by scalability.

Research Goal

Engineering Goal:

- Develop a new analog computing architecture that combines energy efficiency of analog computing, but scalability and flexibility of digital computing.

Technical Innovation:

1. The proposed analog AI computing architecture Voltmatrix introduces a new paradigm where the **parameters are stored digitally** but the **computation is done in analog**. (compared to existing "digital store, digital compute" of digital and "analog store, analog compute" of analog)
2. Digital storage ensures reliability and scalability, while analog computation ensures energy efficiency.
3. Analog switch is used to compute multiplication efficiently both in power and complexity.

Challenges:

1. Transformation from AI software operations into discrete hardware computation
2. Maintaining noise levels and accuracy while improving efficiency

Mathematical Theory

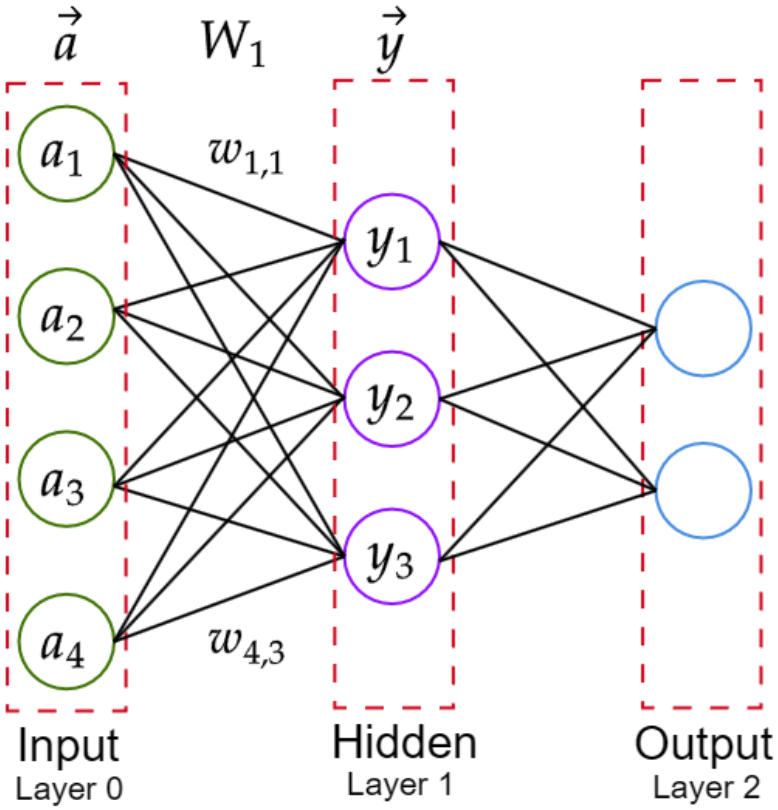


Fig. 6. Neural network architecture.

Neural network inference for one layer:

$$\vec{y}^{(n)} = \left(W^{(n)}\vec{a}^{(n-1)}\right) + \vec{b}^{(n)}$$

- Activation Vector $\vec{a}$: input vector (following the previous layer)
- Weight Matrix W: neural network parameters connecting neurons
- Output Vector $\vec{y}$: output vector

Matrix-vector-multiplication (MVM):

$$\sum_{i=0}^{m} W_{j,i}^{Q}\ \vec{a}_i^{Q}$$

- This architecture aims to perform computation on MVM operation in neural network inference between activation vectors $\vec{a}_i^{Q}$ and weight matrix $W_{j,i}^{Q}$.

VoltMatrix Architecture

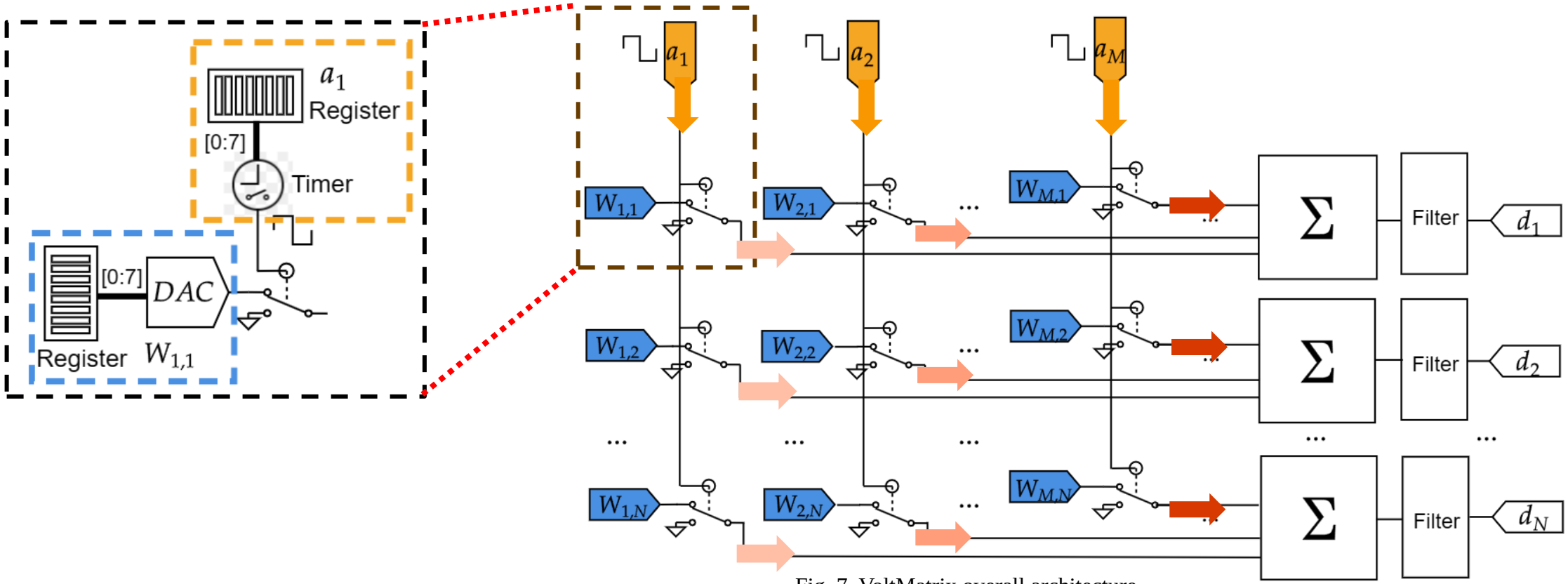


Fig. 7. VoltMatrix overall architecture.

- Activation vectors $\vec{a}_i$ and weight matrix parameters $W_{j,i}$ are stored in digital registers (digital storage)
- $\vec{a}_i^Q$ are encoded with PWM signals generated by a timer, and $W_{j,i}^Q$ encoded with analog voltage signals generated by DAC
- The two signals are multiplied with analog switches ($V_{out} = D \times V_w$) in a parallel grid fashion.
- Activation signal shared column wise. Summed and filtered row-wise for output $\vec{d}_j$.

Architecture Circuit Design

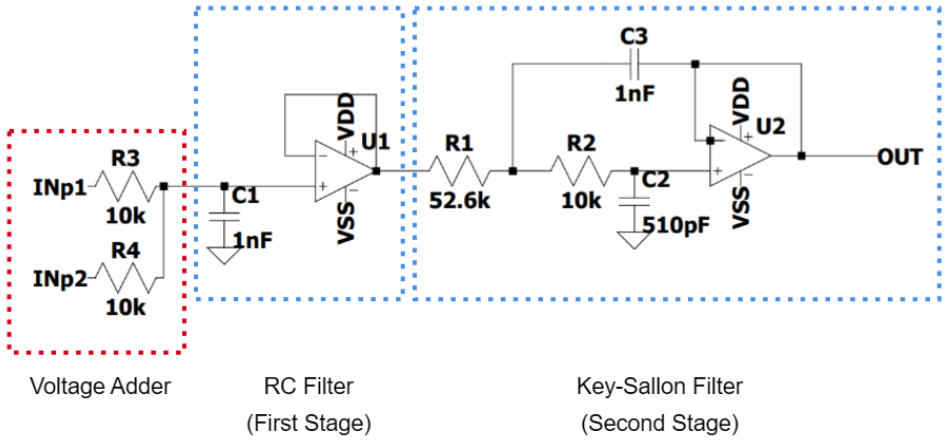


Fig. 8. Summation and filter circuits.

- Voltage adder: perform summation computation
- Two stage filter: 1st stage RC filter, 2nd stage Sallen-Key filter

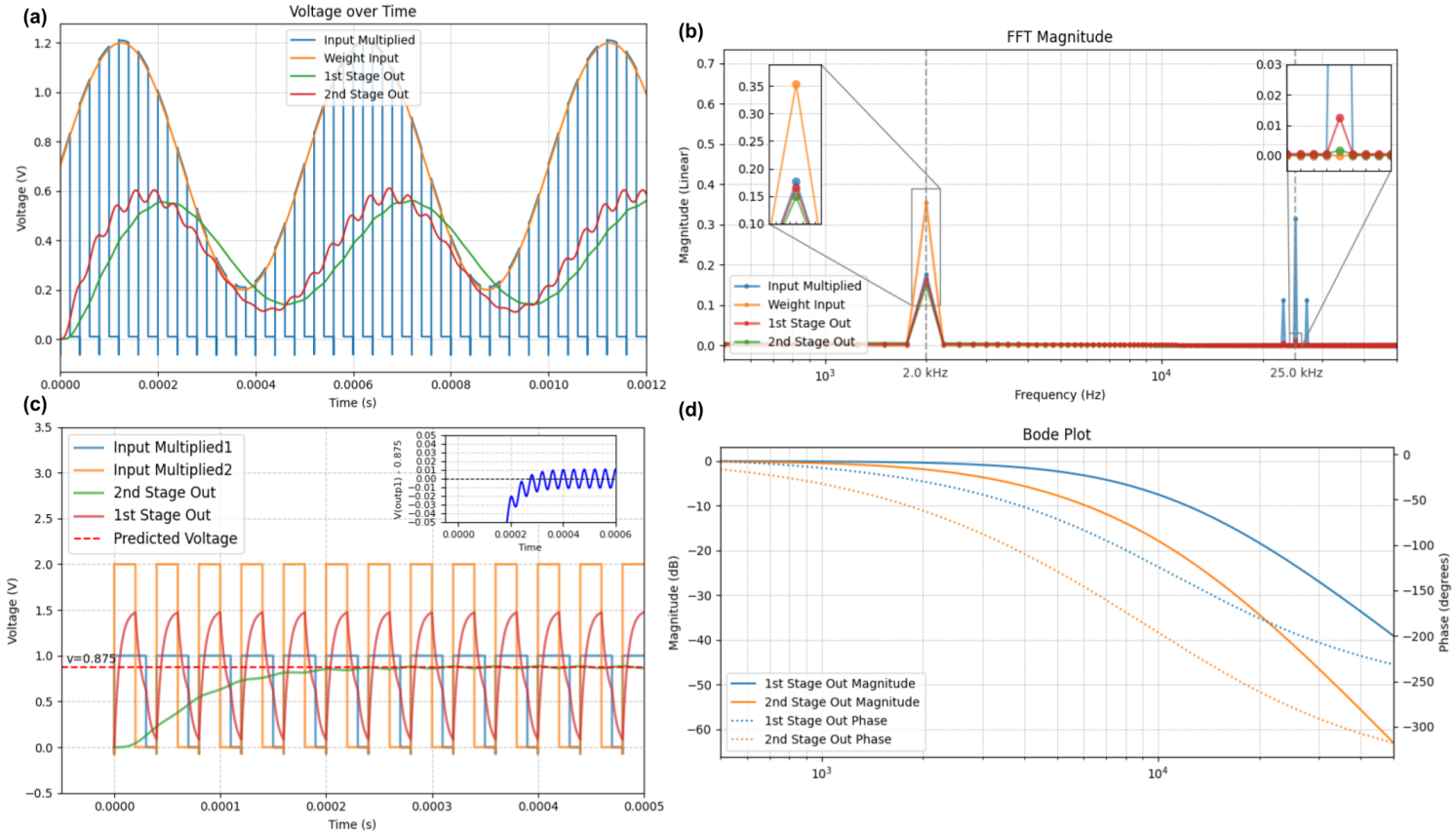
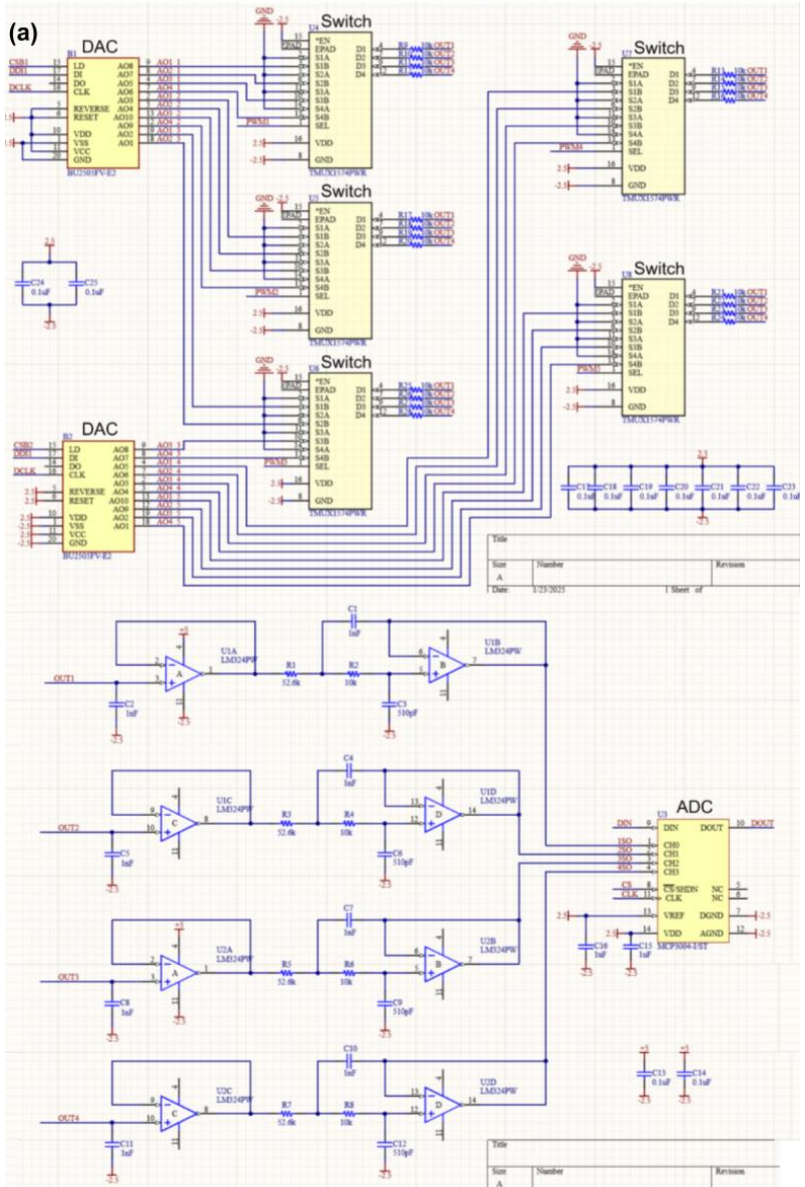


Fig. 9. VoltMatrix architecture circuit simulations.

Overall: 0.2ms input rise time, <0.01V output fluctuation

Prototype Design



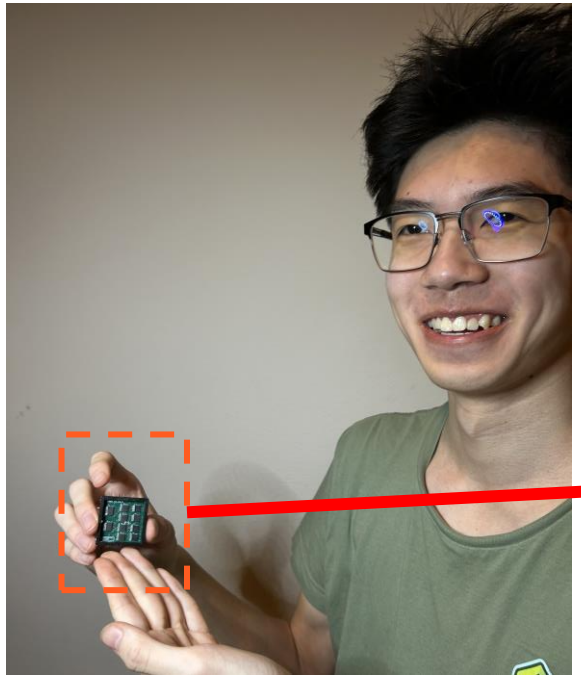


Fig. 10. Analog AI module.

Demo interface circuitry

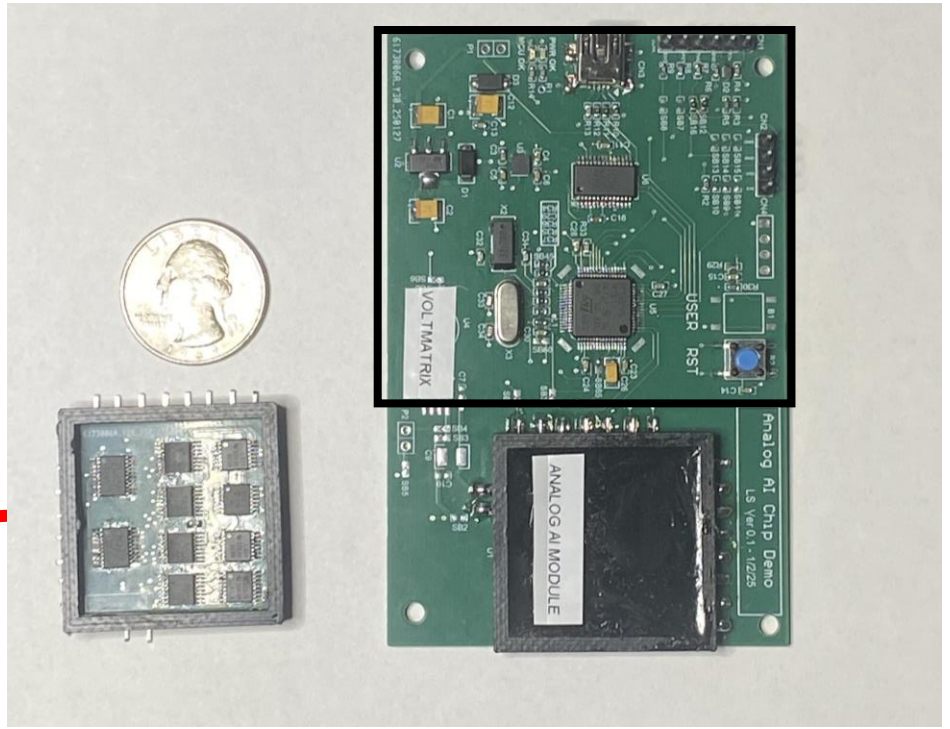


Fig. 11. Application demo unit.

- ✓ AI computation capability: 100k operations per second
- ✓ Comparable to Microchip SAM D processors used in voice detection

Prototype Testing: Multiplication

Multiplication test results:

- Multiplication operation $a \times b$ both ranging from -128 to 128.
- Average 1.7% multiplication error (compared to 4-10% by memristors)
- Roughly symmetrical distribution in raw integer error → most noise canceled out in the summation during the neural network computation

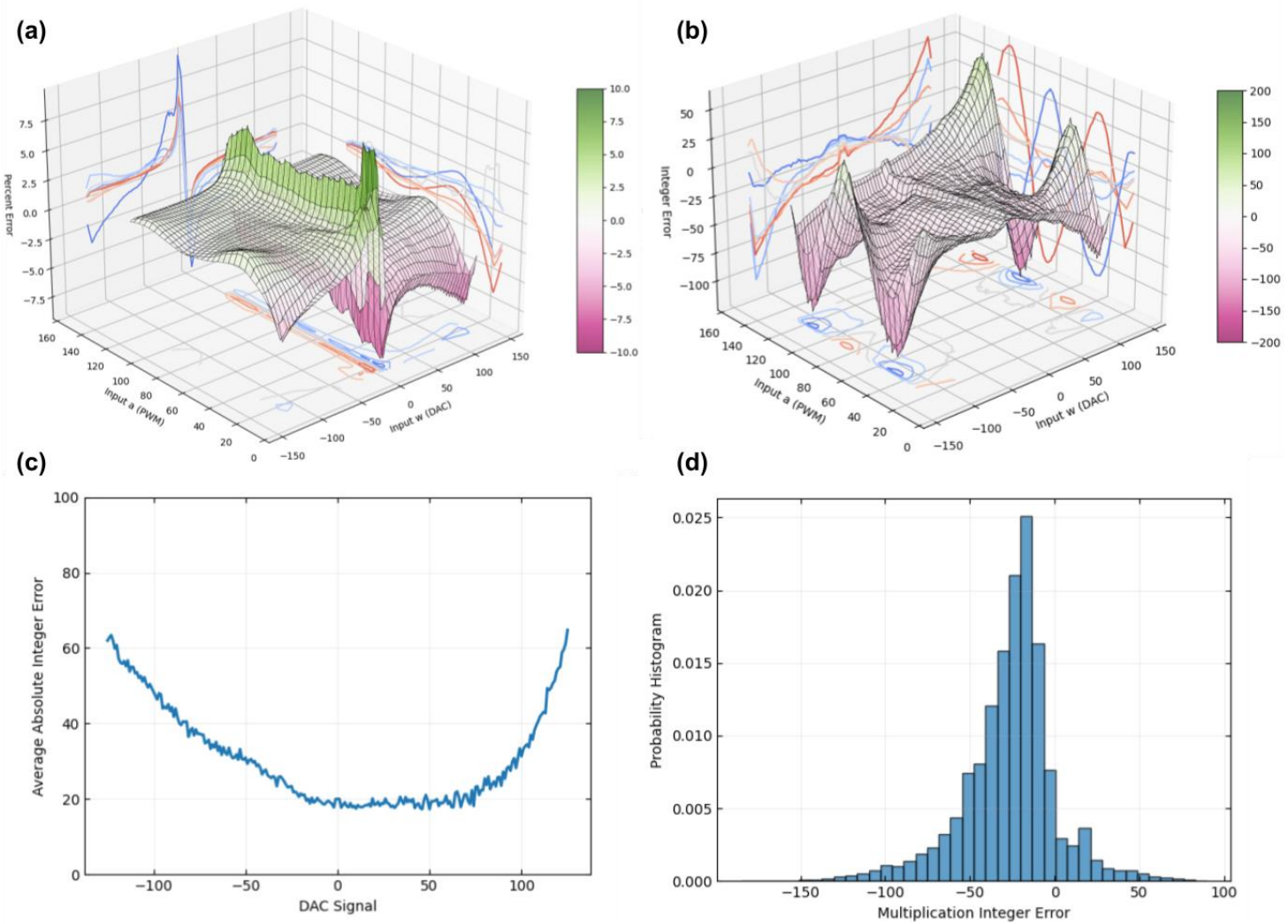


Fig. 12. (a) Multiplication test percentage error (b) Multiplication test raw integer error. (c) Average absolute integer error across DAC signal. (d) Distribution of multiplication integer error

Prototype Testing: Digit recognition application

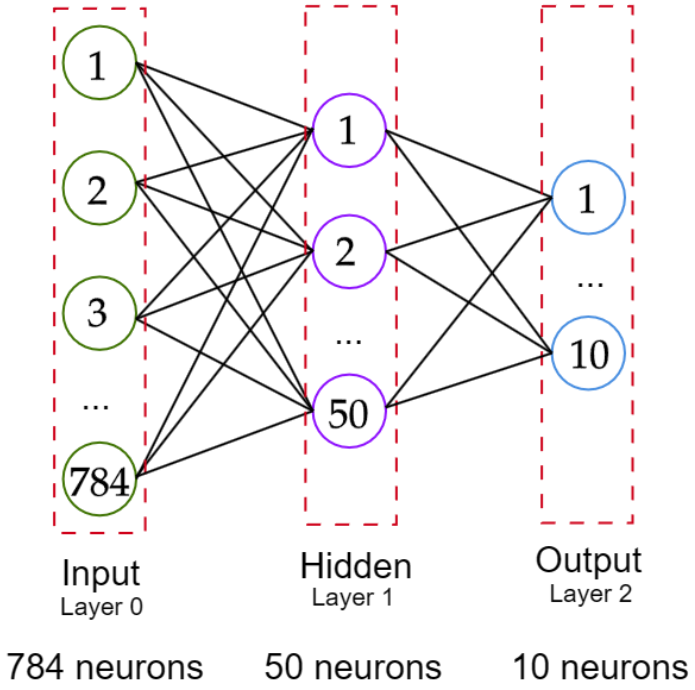


Fig. 13. Digit recognition neural network structure

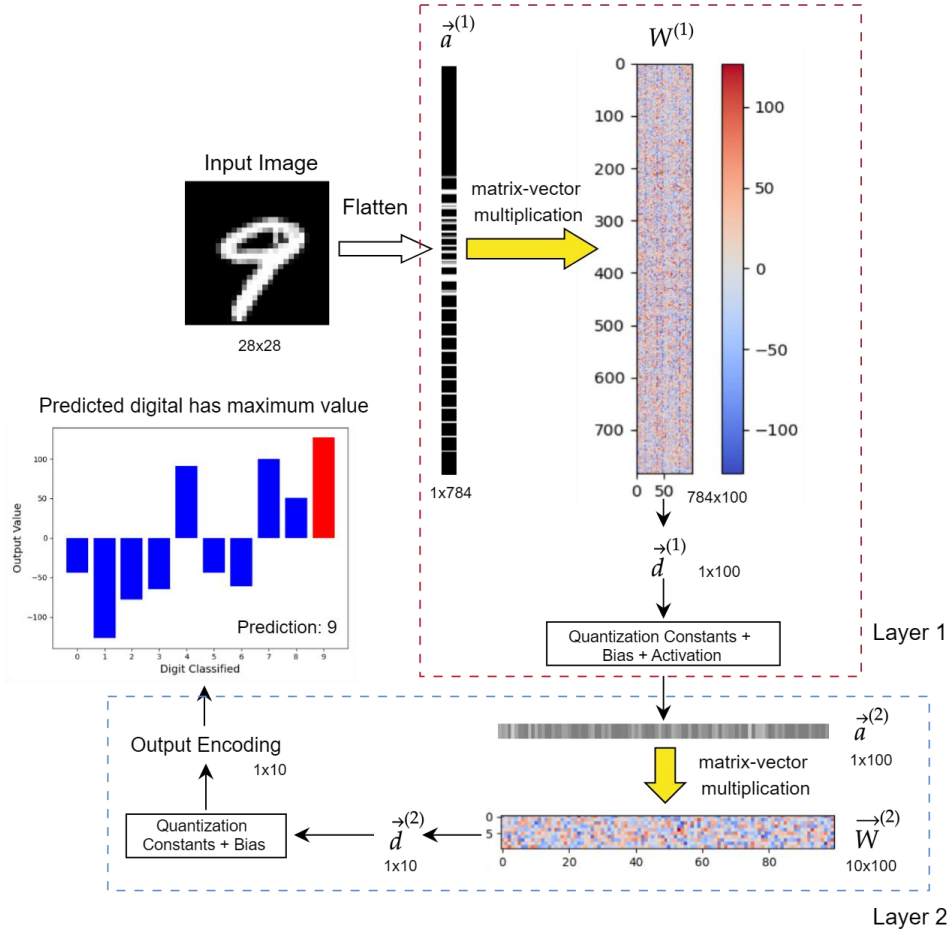


Fig. 14. Neural network interference process

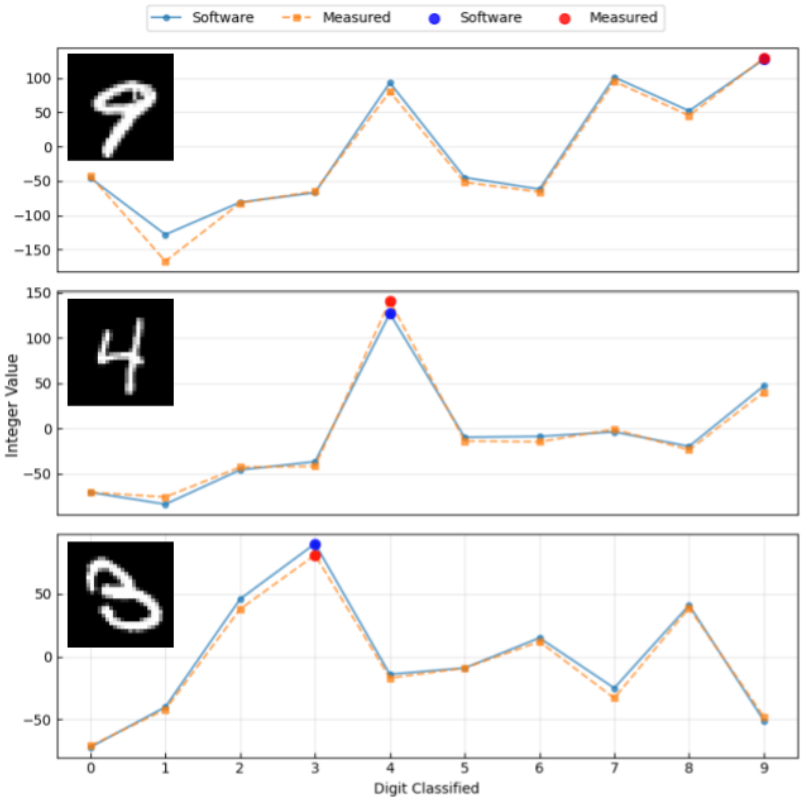


Fig. 15. Sample comparisons of output confidence vectors

Software accuracy: 96.5%
Hardware prototype accuracy: 94.8%.

Chip Layout and Simulation

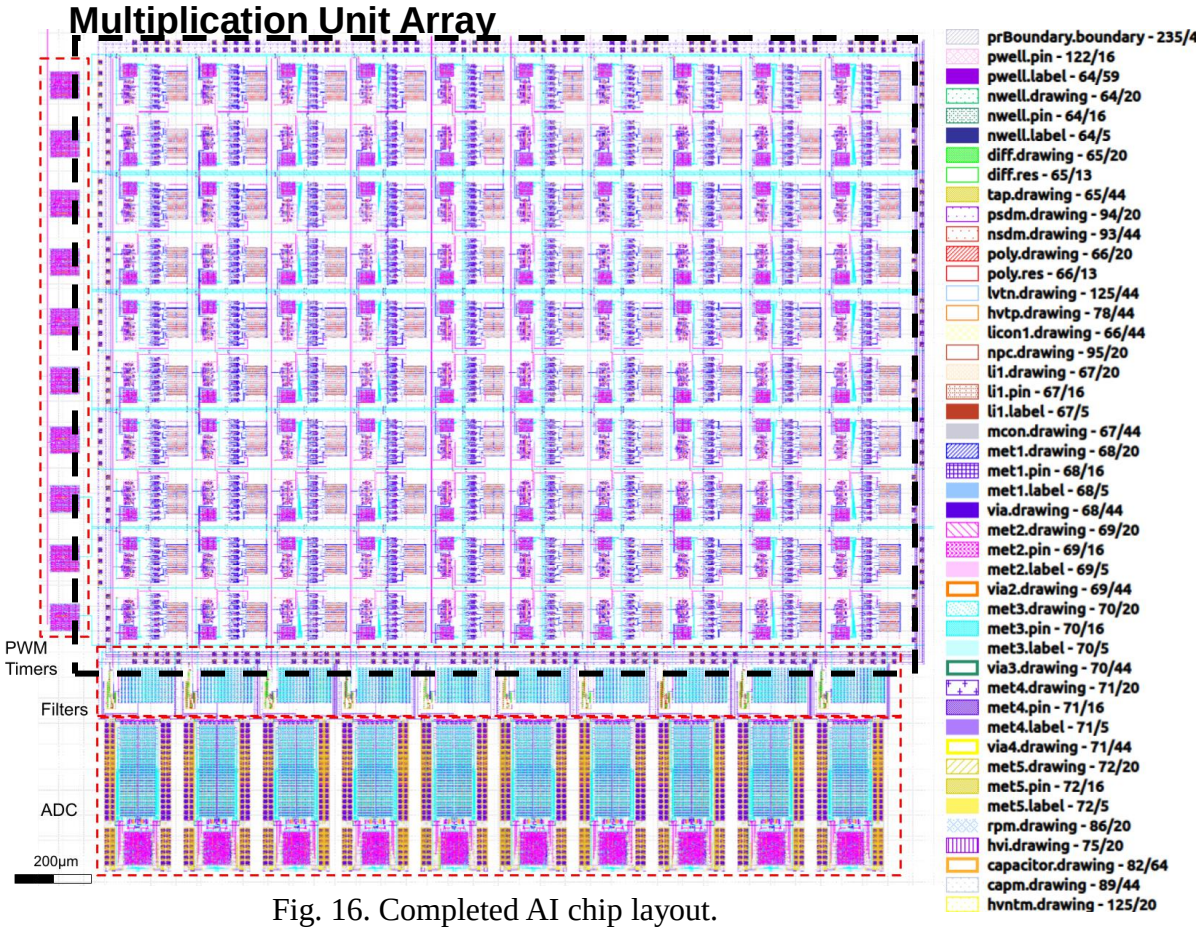


Fig. 16. Completed AI chip layout.

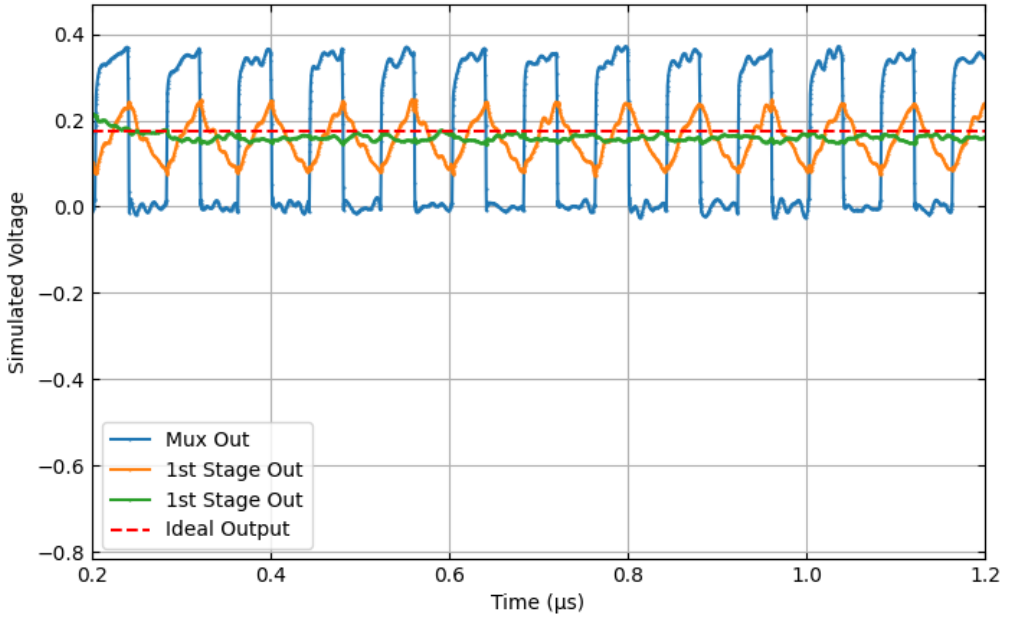


Fig. 17. Simulation Result performing one multiplication with PWM=50% and V_w^H=0.351V.

- 100-multiplier chip layout designed using Magic and Klayout with Skywater 180nm node
- SPICE circuit equivalent model was extracted with Magic (including parasitic) and simulated with NgSpice, achieving average multiplication error of 3.2%.

Chip Layout and Simulation

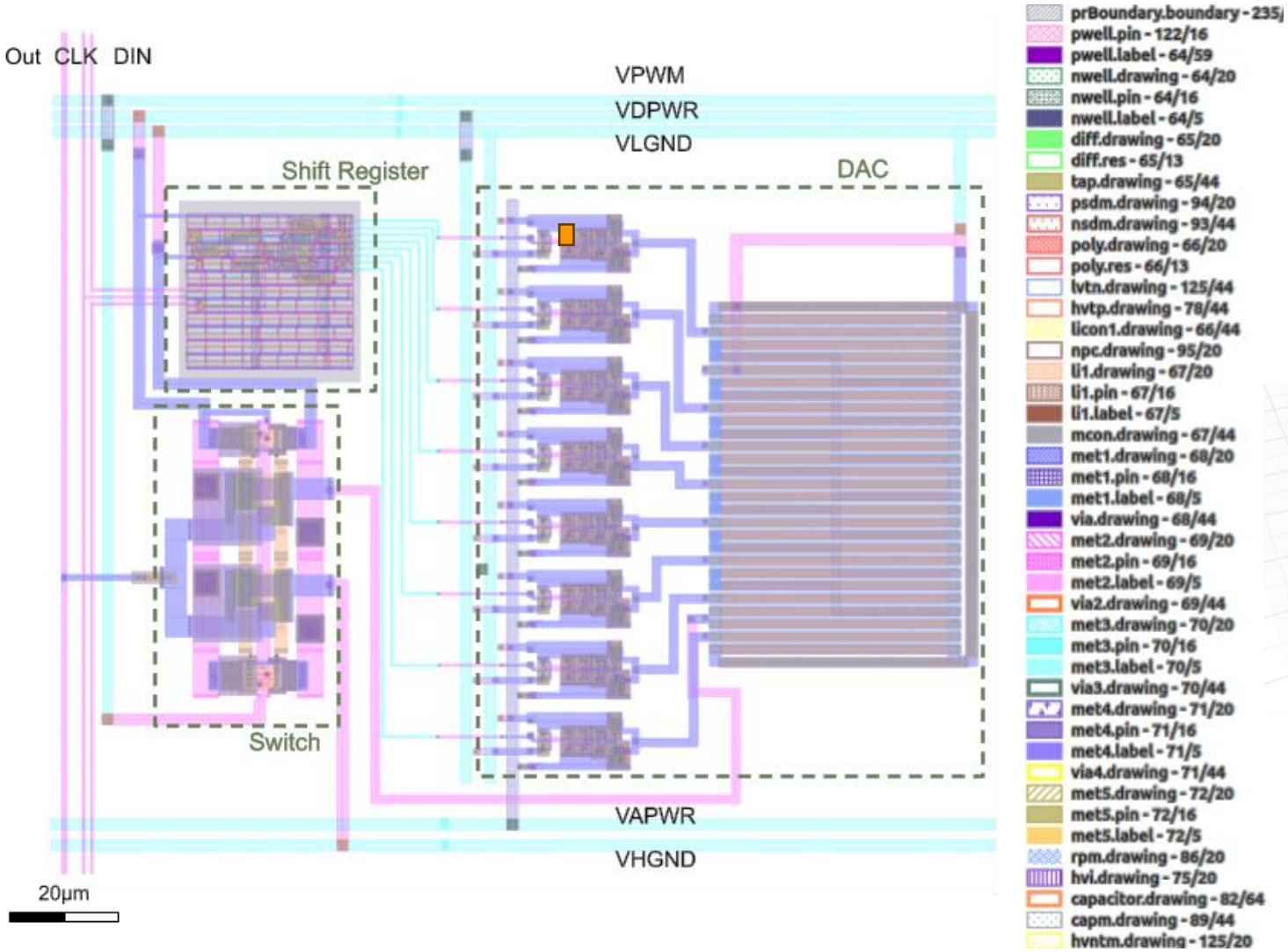


Fig. 18. Multiplication subunit layout.

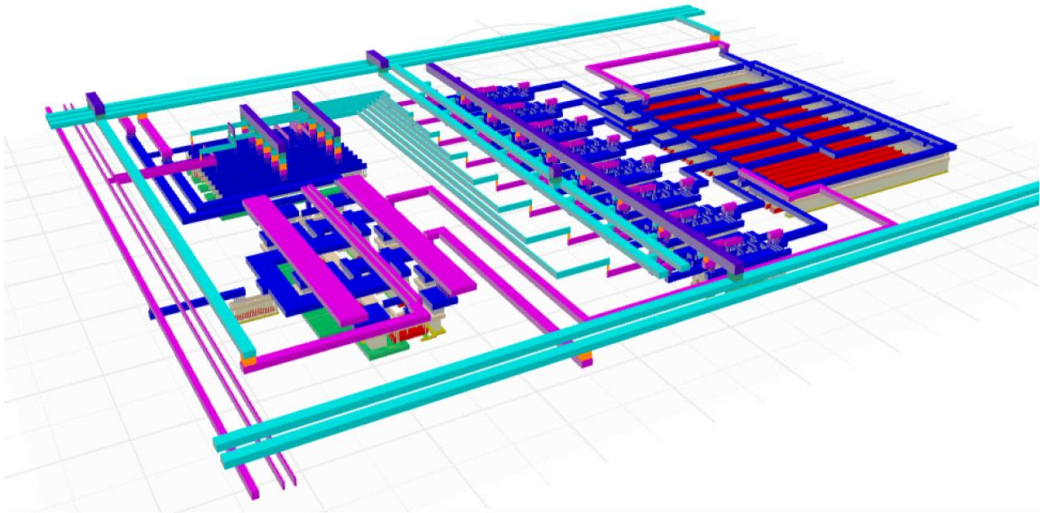


Fig. 19. Multiplication subunit 3D view.

Chip Layout Results

TABLE I
POWER EFFICIENCY COMPARISONS

Design	Power Efficiency (TOPS/W)
VoltMatrix	**5.04**
Complementary pass transistor logic (Digital) [29][1]	1.75
Vedic mathematic algorithm (Digital) [30][1]	0.85
Intelligent Edge SoC neural processor (Digital) [31][1]	1.5
Memristor crossbar dot product engine (Analog) [7][2]	115

TSMC 180nm process
Custom 5nm-40nm hybrid process with palladium and silver

TABLE I III
NEURAL NETWORK COMPARISONS

Design	MNIST accuracy rate
VoltMatrix	**94.8%**
Memristor crossbar dot product engine (Analog) [7]	82%
Digital Designs (Software)	96.5%

- VoltMatrix has a ~3x power efficiency than comparable digital designs with similar 180nm node.
- VoltMatrix has a <100ns storage update time while analog memristor takes 1-10s.

Discussion/Conclusions

- **Validity** – Mathematical model and the prototype results showed minimal deviation. A digit recognition neural network was tested, achieving 94,8% accuracy, similar to software predicted accuracy.
- **Advantage** – The proposed VoltMatrix architecture achieved ~3x higher power efficiency than comparable digital designs. It achieved higher accuracy and flexibility with its more reliable digital storage, demonstrating scalability compared to analog memristors.
- **Applicability** - The design and simulation of the AI chip layout showed VoltMatrix can be realized within constraints of VLSI fabrication while maintaining signal integrity, with extracted simulation achieving an average multiplication error of 3.2%
- **Future work**
  - Manufacture the chip layout design with TinyTapeout.
  - Scaling up the chip with more advanced technology nodes (e.g. TSMC 3nm technology).for commercialization
- **Outlook** - This project serves as a pioneering work to a more energy-efficient computation using “digital store, analog compute” paradigm, paving the way to a more economical and sustainable AI infrastructure.

Acknowledgements: Thank you to Mr. Lei Jiang for his incredible guidance and direction.

REFERENCES

[1] MarketWatch, "AI could demand a shocking amount of electricity — check out this chart," MarketWatch, Feb. 28, 2024. [Online]. Available: https://www.marketwatch.com/story/ai-could-demand-a-shocking-amount-of-electricity-check-out-this-chart-e91e306d[2] NVIDIA, "GeForce RTX Graphics Cards," NVIDIA, 2024. [Online]. Available: https://www.nvidia.com/en-us/geforce/rtx/
[3] K. J. Prabhu and R. Kumar, "Artificial intelligence applications in smart grid," in Machine Learning and Optimization Models for Optimization in Cloud, Cham: Springer, 2022, pp. 753–772. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-05230-9_51
[4] A. Mirhoseini, "Introduction to artificial intelligence and machine learning," University of California, Irvine, 2024. [Online]. Available: https://ics.uci.edu/~amirr1/docs/Introduction.pdf
[5] H. Esmaeilzadeh, "Efficient deep learning processing with specialized accelerators," in IEEE Transactions on Computers, vol. 69, no. 11, pp. 1632–1645, Nov. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8865106
[6] J. Wu et al., "Nanomaterials for artificial intelligence and machine learning: A review," Nanoscale Research Letters, vol. 15, no. 1, p. 183, 2020. [Online]. Available: https://link.springer.com/article/10.1186/s11671-020-03299-9