ROQ: A Noise-Aware Quantization Scheme Towards Robust Optical Neural Networks with Low-bit Controls

[1]Jiaqi Gu, [1]Zheng Zhao, [1]Chenghao Feng, [2]Hanqing Zhu,
[1]Ray T. Chen, [1]David Z. Pan
[1]ECE Department, The University of Texas at Austin
[2]Microelectronics Department, Shanghai Jiao Tong University

This work is supported in part by MURI

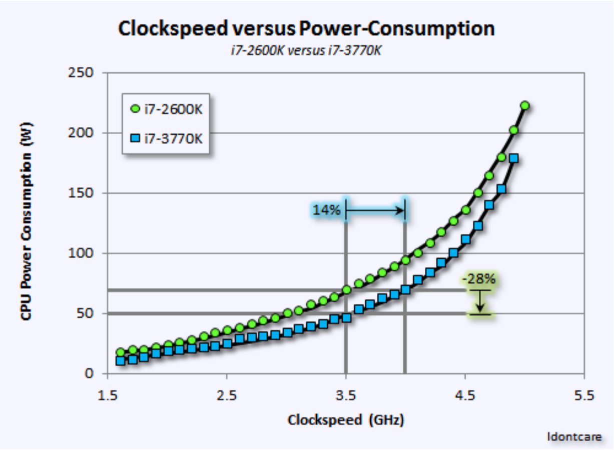
AI Acceleration and Challenges

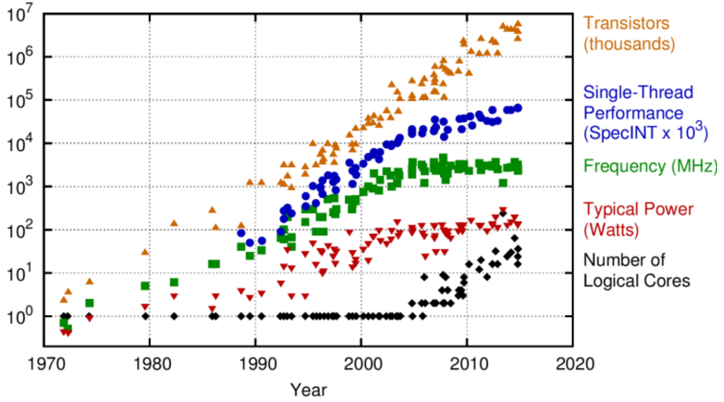
- ML models and dataset keep increasing
  - Low latency
  - Low power
  - High bandwidth

Autonomous Vehicle

Data Center

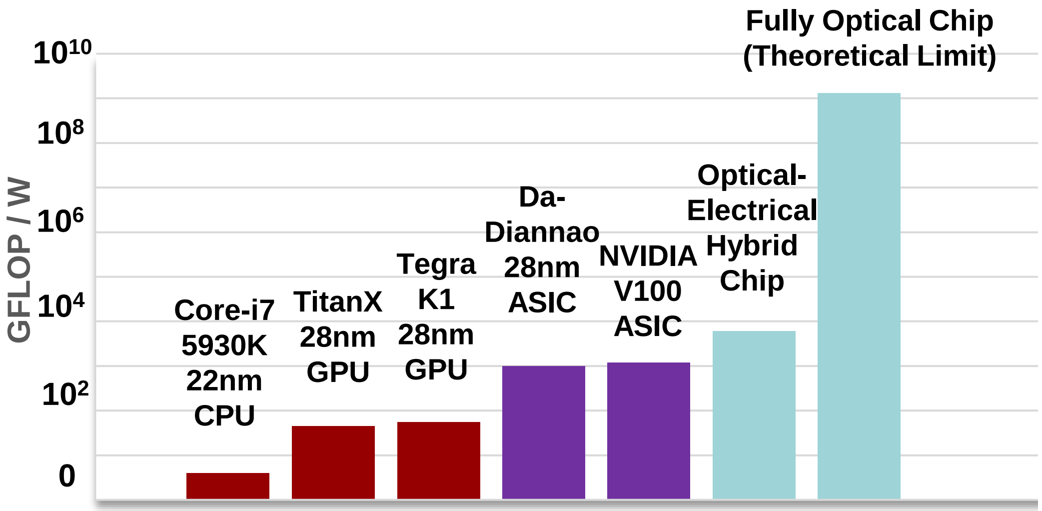
- Moore's law is challenging to provide higher-performance computations



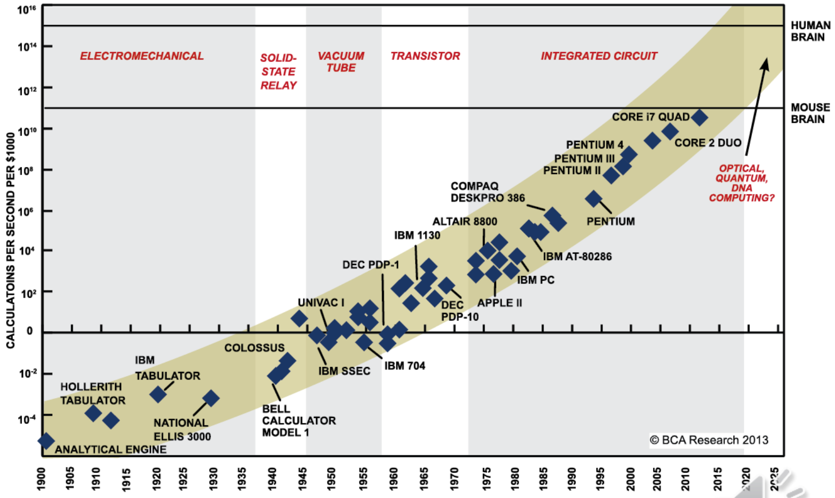


AI Acceleration and Challenges

- Using light to continue Moore's Law
- Promising technology for next-generation AI accelerator



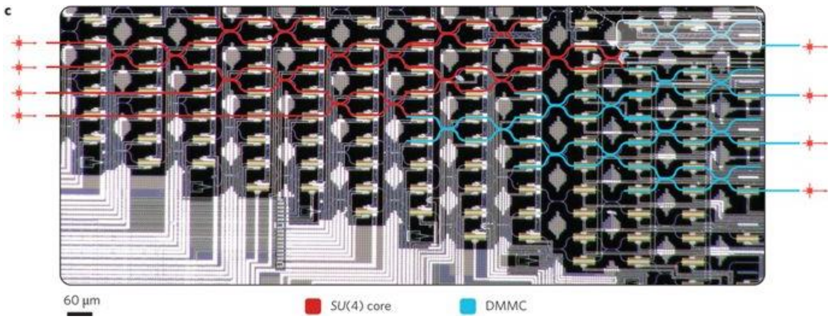
[Shen+, *Nature Photonics* 2017]

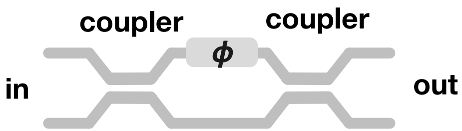


SOURCE: RAY KURZWEIL, "THE SINGULARITY IS NEAR: WHEN HUMANS TRANSCEND BIOLOGY", P.67, *THE VIKING PRESS*, 2006. DATAPOINTS BETWEEN 2000 AND 2012 REPRESENT BCA ESTIMATES.

Optical Neural Networks (ONN)

- Emergence of neuromorphic platforms for AI acceleration
- Optical neural networks (ONNs)
  - Ultra-fast inference speed (**~ 100 ps**)
  - >100 GHz photo-detection rate
  - Near-zero energy consumption (**< 1 fJ / MAC**)
- Unsatisfactory non-ideal effects
  - Limited voltage control resolution -> **Low precision** phase encoding
  - Device-level noise and variation -> **Noise robustness** issue



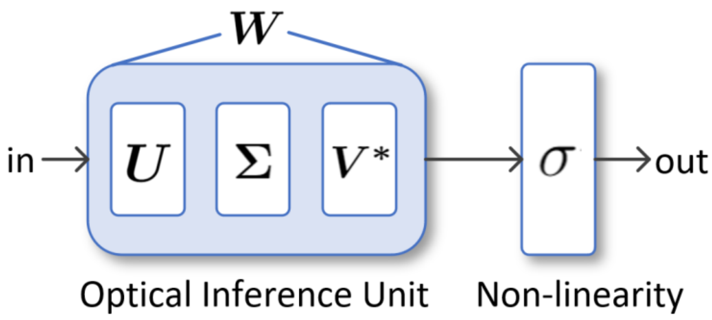


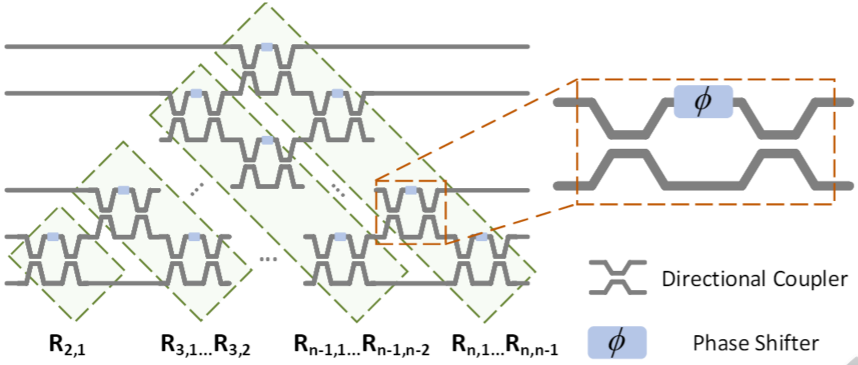
[Shen+, *Nature Photonics* 2017]

Classical ONN Architecture

- Map weight matrix to MZI arrays
- Singular value decomposition
  - $W = U\Sigma V^*$
  - **U and V*** are square unitary matrices
  - **Σ** is diagonal matrix
- Unitary group parametrization:
  - $$U(n) = D \prod_{i=n}^{2} \prod_{j=1}^{i-1} R_{ij}$$
  - R_{ij} is planar rotation matrix
  - R_{ij} with phase ϕ can be implemented by an MZI

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$





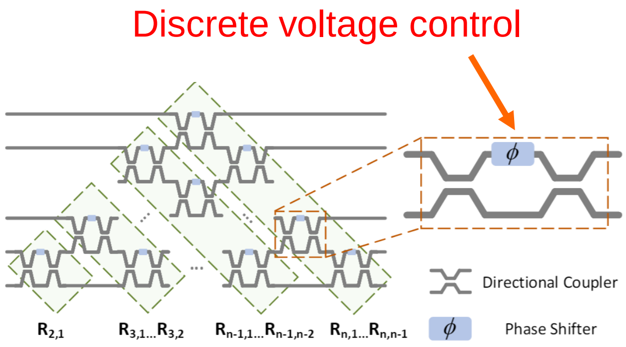
Non-ideality: Low-bit Control

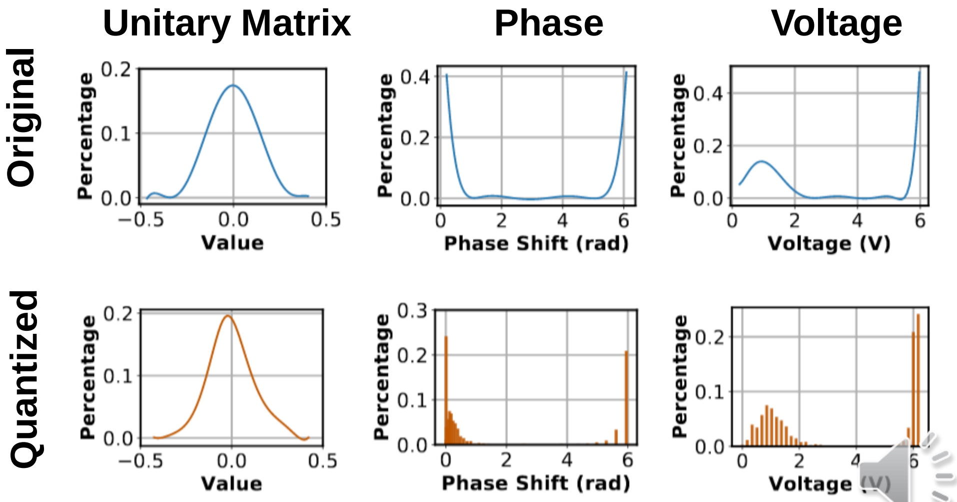
- Low control precision
  - Control complexity consideration
  - Voltage control has limited bitwidths

$$\Delta_v = v_{max}/(2^b - 1)$$

- Challenge
  - Non-uniform phase quantization
  - Expensive for gradient calculation

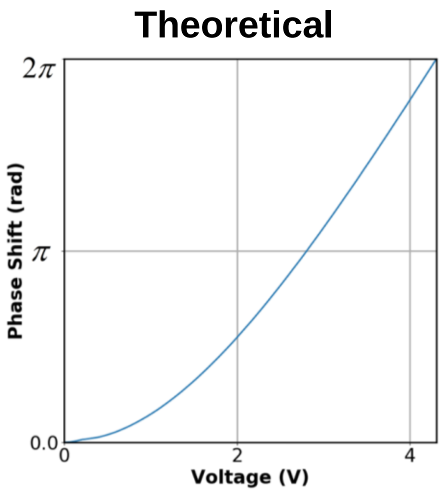
$$\frac{\partial U}{\partial \phi_{ij}} = \boldsymbol{D}\boldsymbol{R}_{n1}\boldsymbol{R}_{n2}\boldsymbol{R}_{n3}\cdots\frac{\partial \boldsymbol{R}_{ij}}{\partial \phi_{ij}}\cdots\boldsymbol{R}_{31}\boldsymbol{R}_{32}\boldsymbol{R}_{21}$$

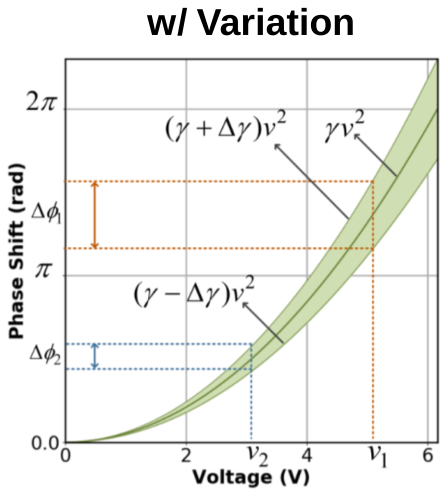




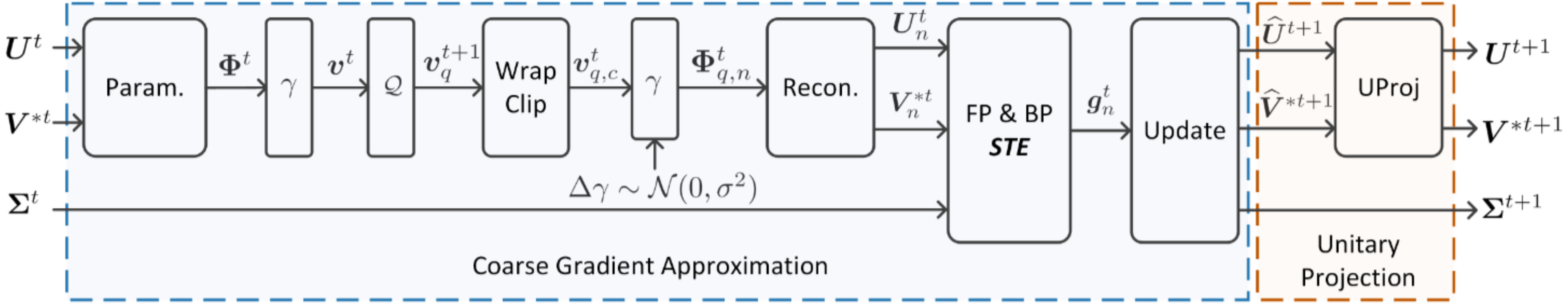
Non-ideality: Device Variation

- Phase shifter Gamma noise => Phase encoding error => Acc. degradation
- Non-ideal phase shifter response curve
  - Theoretical: $\phi = \gamma v^2$
  - Practical: gamma noise $\Delta\gamma \sim \mathcal{N}(0, \sigma^2)$
    - Environmental changes
    - Manufacturing variations
    - Temperature changes
    - ...
  - Larger phase is more noise sensitive





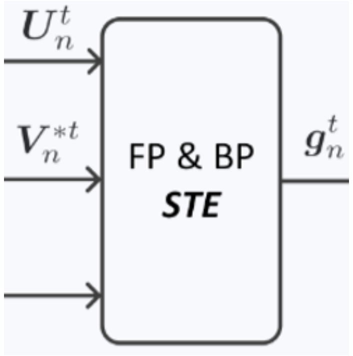
Quantization Scheme



- **Coarse Gradient Approximation**
  - Gradient propagation for voltage quantization
- **Unitary projection**
  - Map matrix U, V* to unitary planes
- Based on blocking matrix multiplication
  - Better scalability

Coarse Gradient Approximation

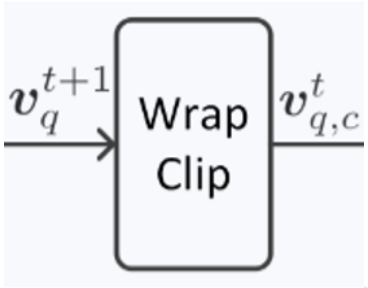
- Model voltage-domain quantization $\boldsymbol{U}_q^t = \mathcal{Q}_b(\boldsymbol{U}^t)$ as STE
  - No intermediate gradient computation
  - Efficient coarse gradient propagation



$$g_q^t = \frac{\partial L^t}{\partial U^t} = \frac{\partial L^t}{\partial U_q^t}$$

- Wrap clipping
  - Invalid large phases will be clipped

$$v_{q,c} = \texttt{WrapClip}(v_q) = \begin{cases} v_q, & \text{if } 0 \leq v_q < v_{2\pi} \\ 0, & \text{if } v_q \geq v_{2\pi}. \end{cases}$$



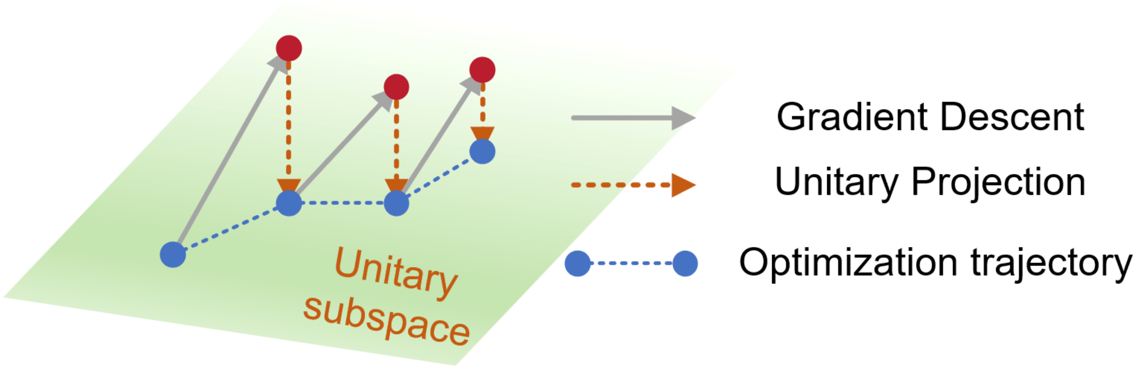
  - Wrapping will reduce phase error and noise sensitivity

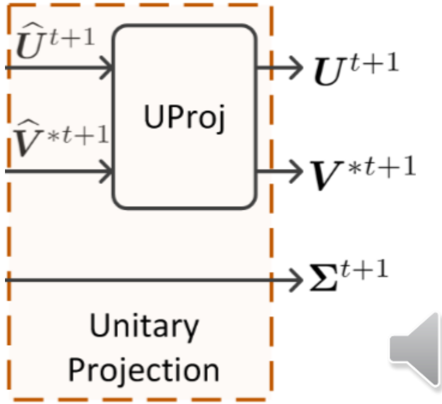
Unitary Projection

- Satisfy orthogonality constraint for unitary matrix U and V*

$$U = \texttt{UProj}(\widehat{U}) \quad \left\{ \begin{array}{l} \boldsymbol{P}\boldsymbol{S}\boldsymbol{Q}^* = \text{SVD}(\widehat{\boldsymbol{U}}) \\ \boldsymbol{U} = \boldsymbol{P}\boldsymbol{Q}^*. \end{array} \right.$$

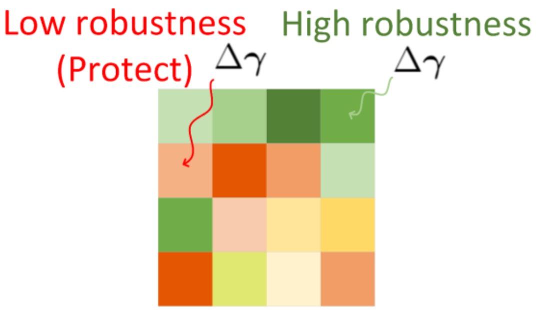
- SVD-based projection method minimizes projection error
- Projected gradient descent: project onto unitary plane each iteration





Noise-Aware Training

- Protective group Lasso regularization (PGL)
  - Penalize less robust weight blocks



$$\mathcal{L}_{PGL} = \sum_{l=1}^{L} \sum_{i=1}^{p^l} \sum_{j=1}^{q^l} P_{ij}^l \sqrt{1/\beta_{ij}^l}\, ||\boldsymbol{W}_{ij}^l\,||_2^2$$

  - Protective coefficient is dynamically learnable
    - Gamma noise injection: $\boldsymbol{\Phi}_{q,n} = (\gamma + \Delta\gamma)\boldsymbol{v}_{q,c}^2$
    - Dynamic robustness evaluation

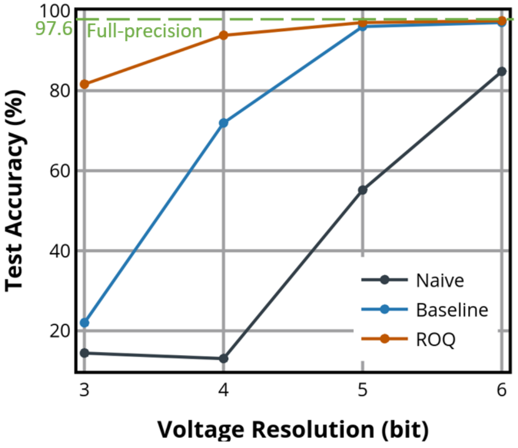
$$P_{ij}^l = \frac{d(\boldsymbol{W}_{ij,q}^l, \boldsymbol{W}_{ij,q,n}^l)}{\max_{i,j}\left(d(\boldsymbol{W}_{ij,q}^l, \boldsymbol{W}_{ij,q,n}^l)\right)}$$

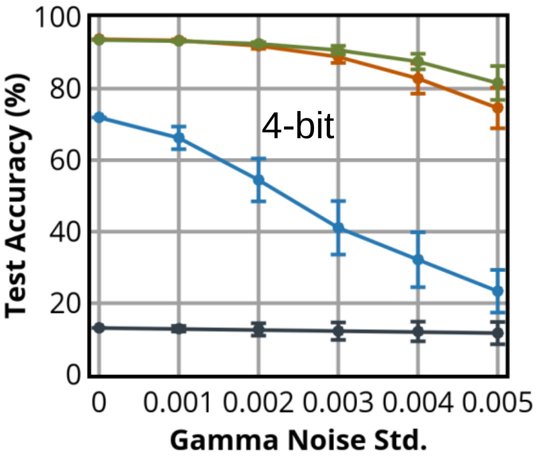
    - Learnable coefficient via EMA: $\widehat{P}_{ij}^{l(t)} = \eta \widehat{P}_{ij}^{l(t-1)} + (1-\eta) P_{ij}^{l(t)}$

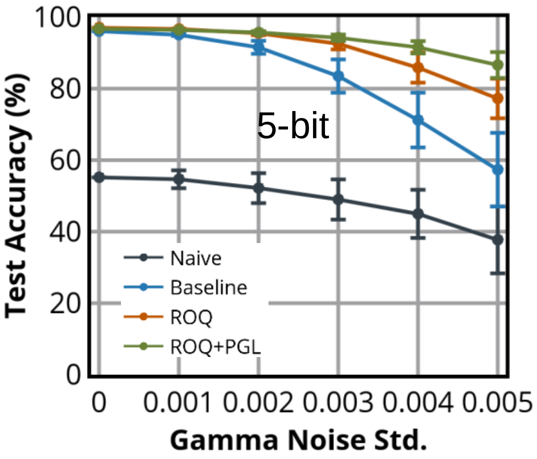
Experimental Results

- Better Noise-robustness under low-bit voltage controls (3 ~ 6 bits)

	Bitwidth	Test Acc.	Test Acc. w/ variation
Full-precision	**High**	**97%**	**89%**
Previous method	**Low**	**72%**	**41%**
ROQ	**Low**	**94%**	**91%**







Contribution of This Work

- Voltage-domain quantization scheme for ONN
  - Efficient quantized ONN training methodology
  - ~90% accuracy under low-bit voltage controls

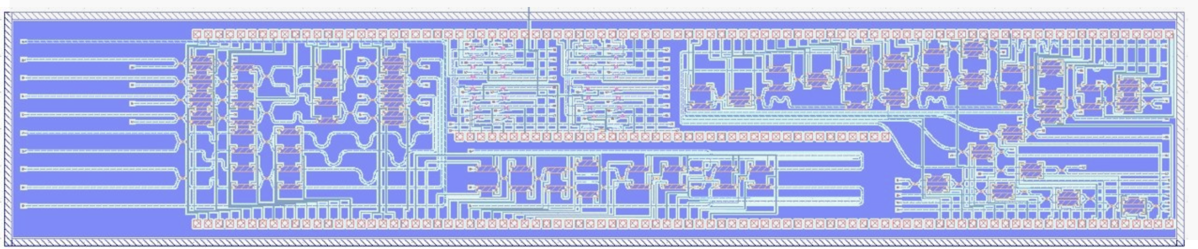
- Noise-aware training method
  - Protective Group Lasso regularization technique is proposed to boost noise-robustness of quantized ONNs
  - >80% inference accuracy under 3-bit control and 5e-3 gamma noise, compared to ~20% for baseline method
  - Lower accuracy variance under gamma noise

Future Directions

Investigate other robustness issues: thermal crosstalk

Integration with On-chip training and other ONN architectures

Chip tapeout and experimental evaluation



Thanks
Q&A