清華大學电子工程系
Department of Electronic Engineering, Tsinghua University

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

IEEE/ACM 2025 INTERNATIONAL CONFERENCE ON COMPUTER-AIDED DESIGN ICCAD 44th Edition

How Do Errors Impact NN Accuracy on Non-Ideal Analog PIM? Fast Evaluation via an Error-Injected Robustness Model

Lidong Guo[1*], Zhenhua Zhu[1*], Qiushi Lin[1], Xuefei Ning[1], Yuan Xie[2], Huazhong Yang[1], Wangyang Fu[1], Yu Wang[1]

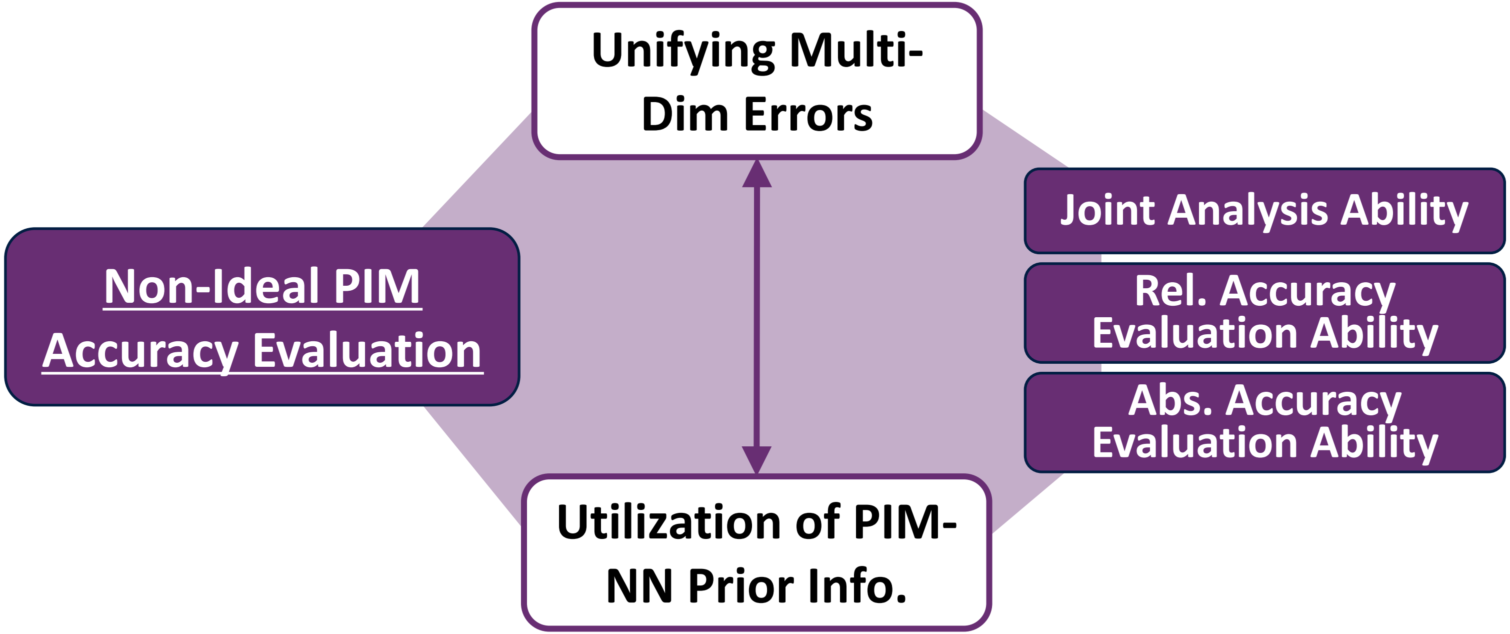
[1]Tsinghua University, [2]HKUST

E-mail: gld21@mails.tsinghua.edu.cn, zhenhuazhu@mails.tsinghua.edu.cn, yu-wang@tsinghua.edu.cn

NICS EFC

Overview

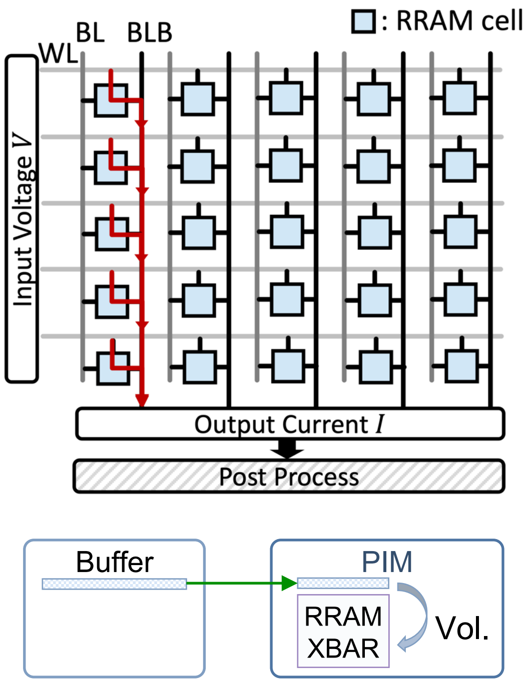
Exploring the impacts of various errors on non-ideal Processing-in-Memory (PIM) architecture's computational accuracy



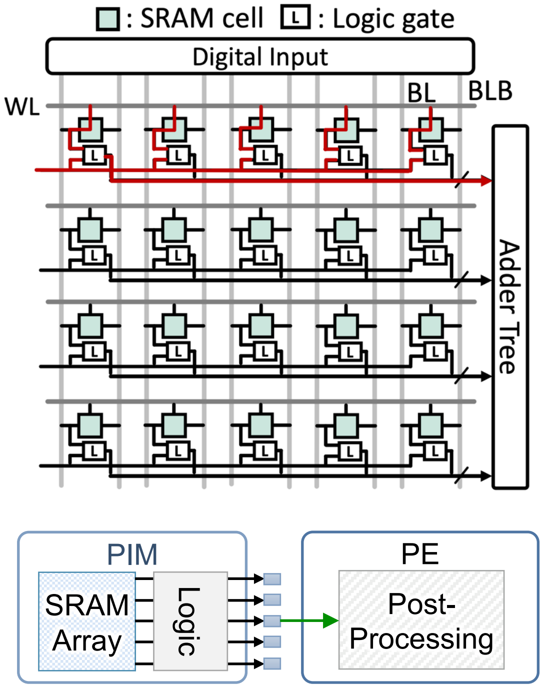
PIM for ‘Memory Wall’ Problem

Reducing the data movement overhead

Analog PIM Arch.



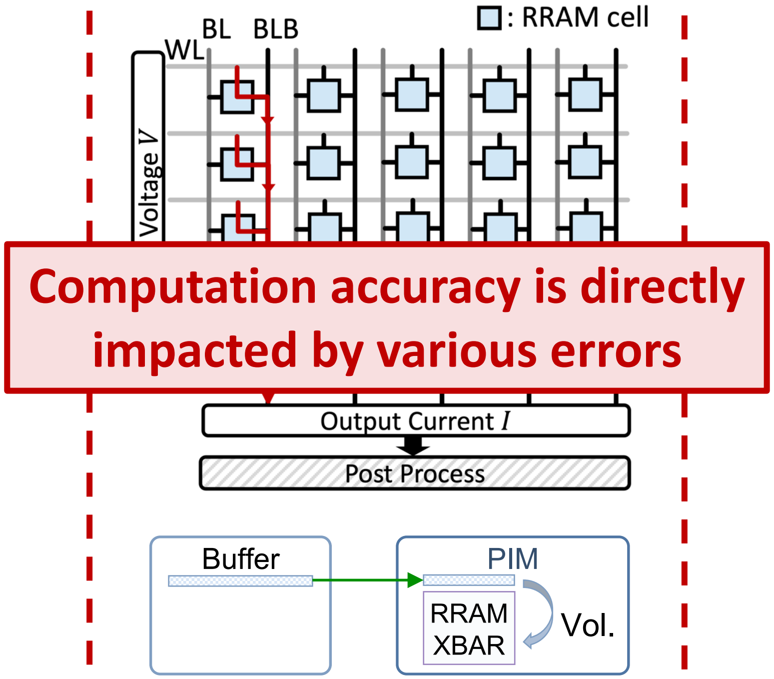
Digital PIM Arch.



PIM for 'Memory Wall' Problem

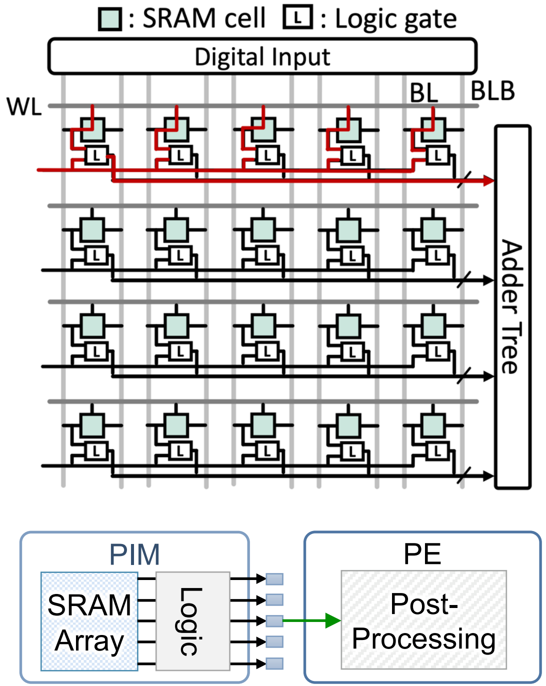
Reducing the data movement overhead

Non-Ideal Analog PIM Arch.



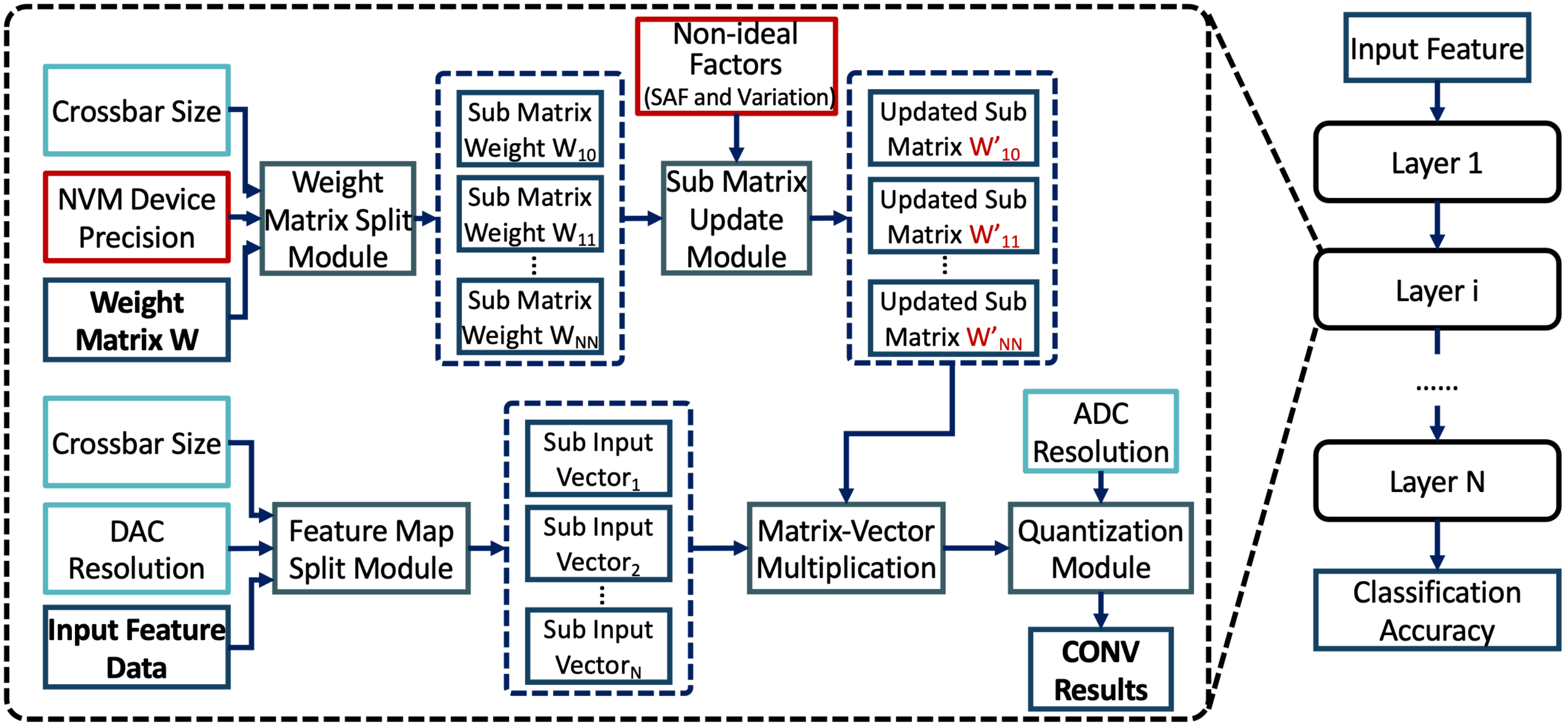
Computation accuracy is directly impacted by various errors

Digital PIM Arch.



Bit/XBAR-Slicing-based ACC Evaluation

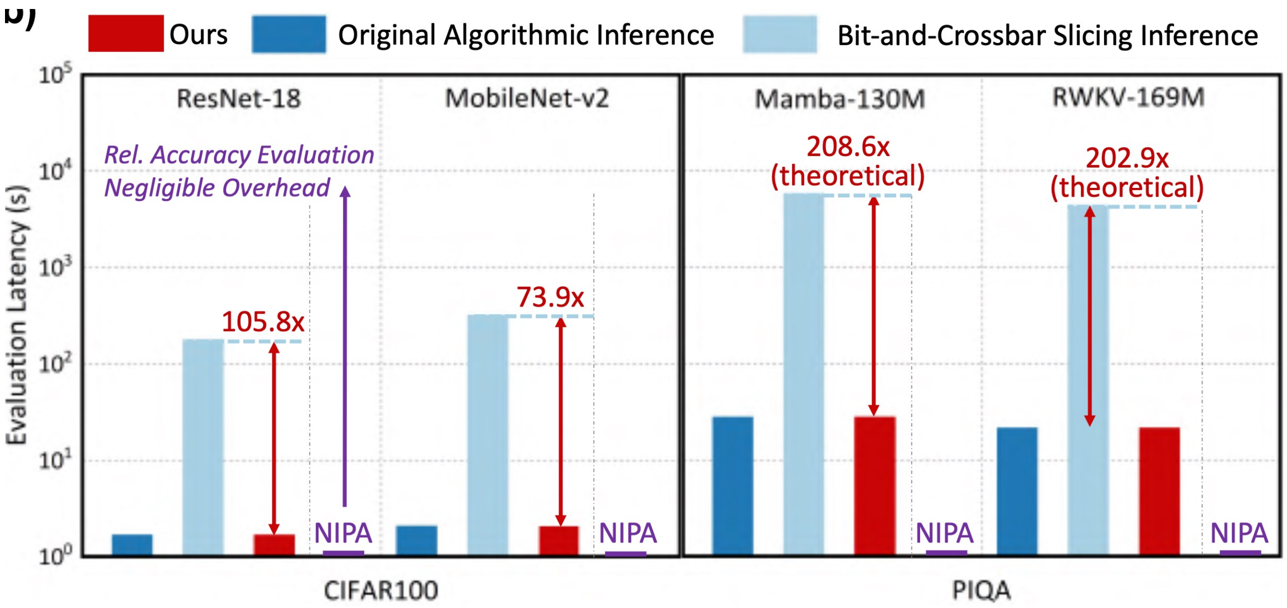
- Absolute Acc of NN on analog PIM arch. can be accurately evaluated by slicing paradigm



MNSIM 2.0

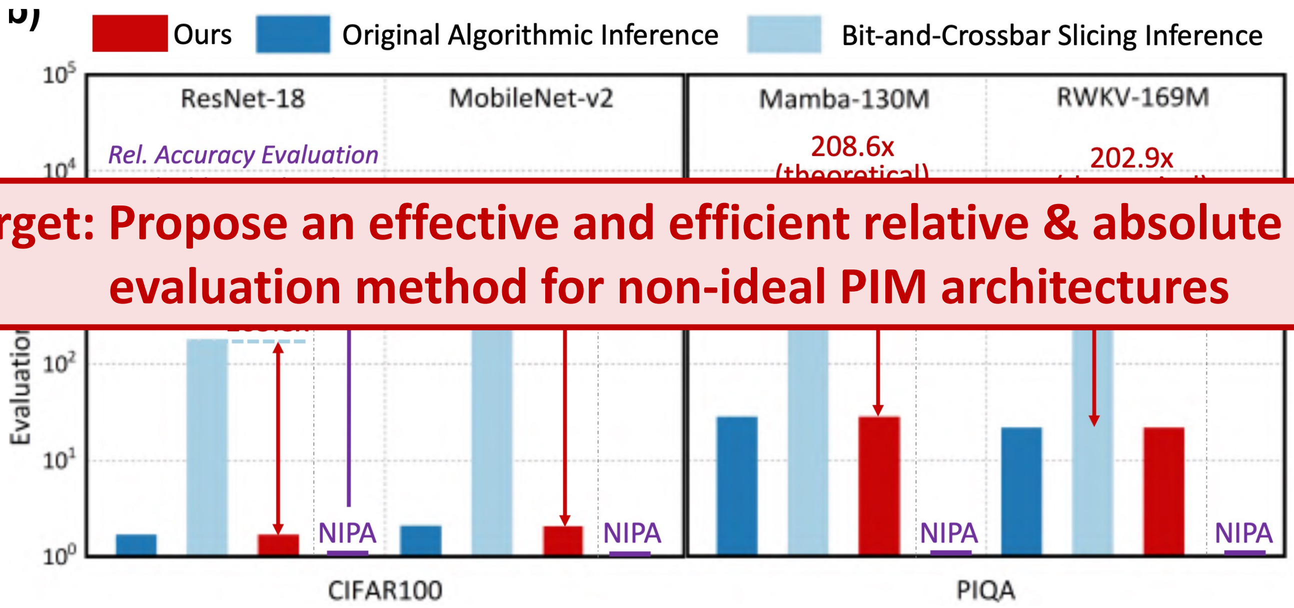
Bit/XBAR-Slicing-based ACC Evaluation

- **Slicing-based evaluation paradigm incurs additional intolerable overhead**
  - Inference latency increases more than **100 times**
  - E.g. **10,000 seconds** for an end-to-end accuracy evaluation process
  - NN-based predictor / a methematically derived metric only support relative acc evaluation



Bit/XBAR-Slicing-based ACC Evaluation

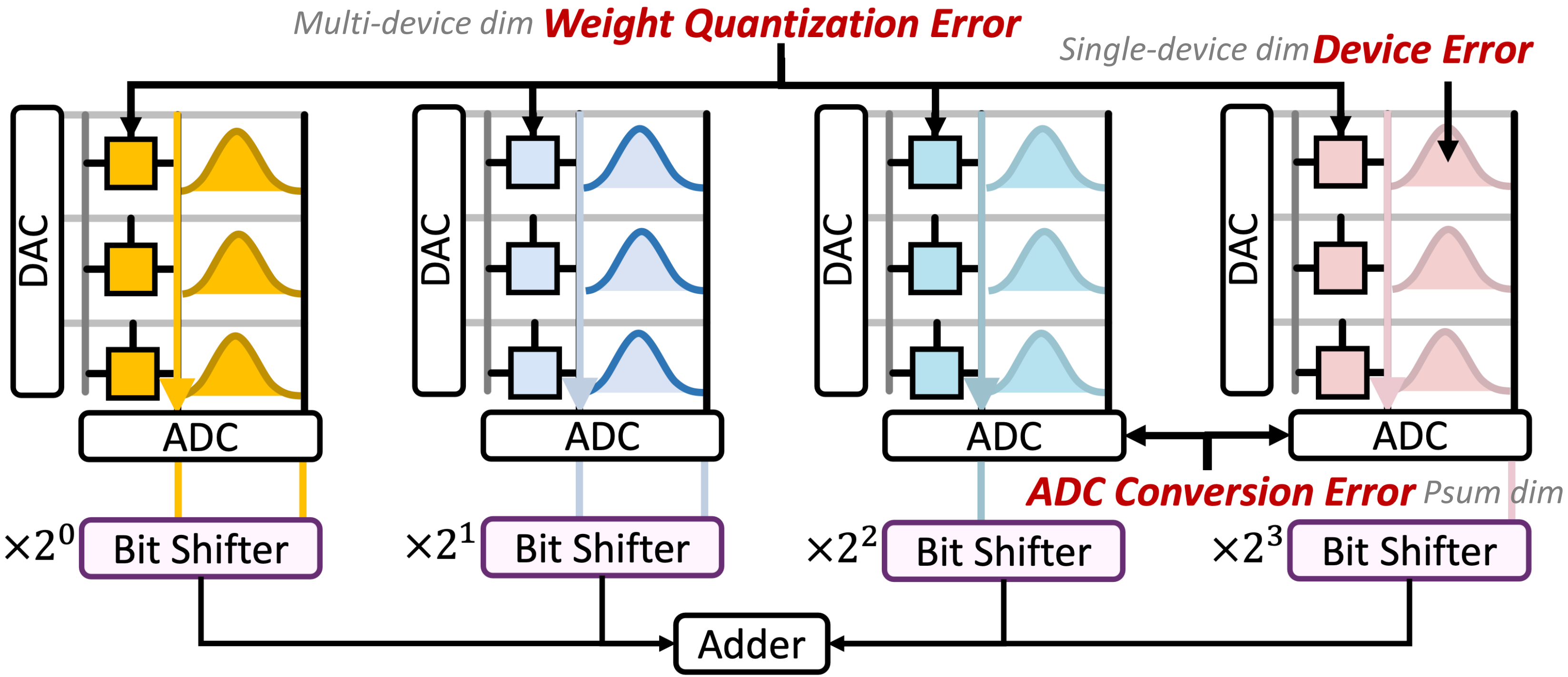
- **Slicing-based evaluation paradigm incurs** <u>additional intolerable overhead</u>
  - Inference latency increases more than **100 times**
  - E.g. **10,000 seconds** for an end-to-end accuracy evaluation process
  - NN-based predictor / a methematically derived metric only support relative acc evaluation



Key Target: Propose an effective and efficient relative & absolute accuracy evaluation method for non-ideal PIM architectures

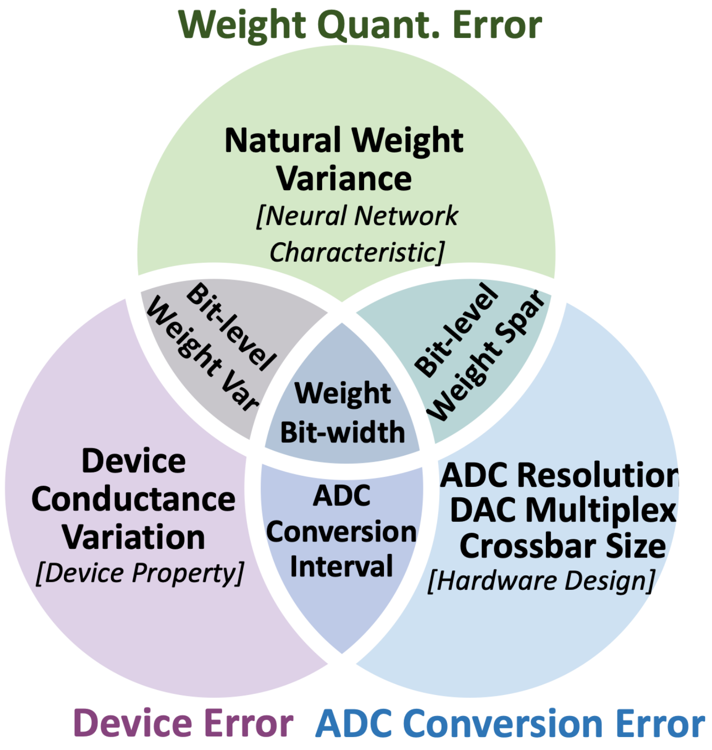
Challenges Arise in Error Analysis

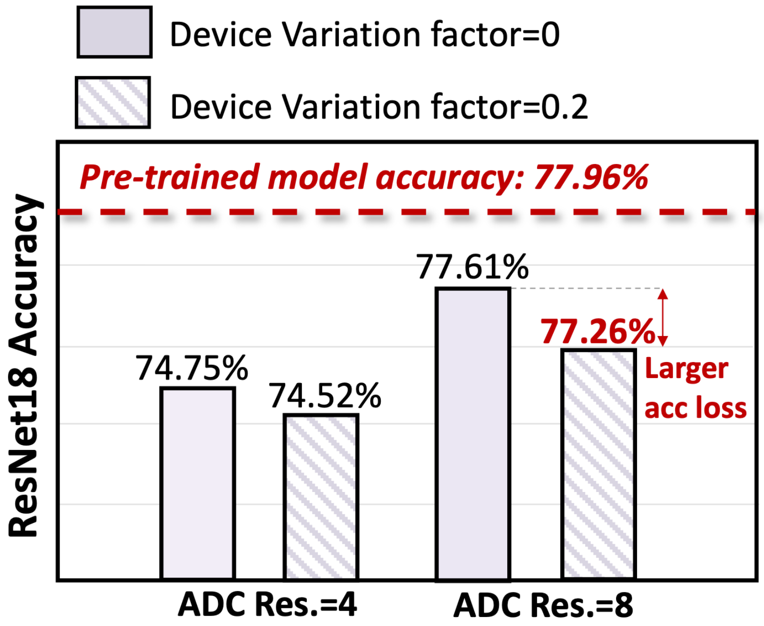
First Challenge: ***Errors are injected across different dimensions***



Challenges Arise in Error Analysis

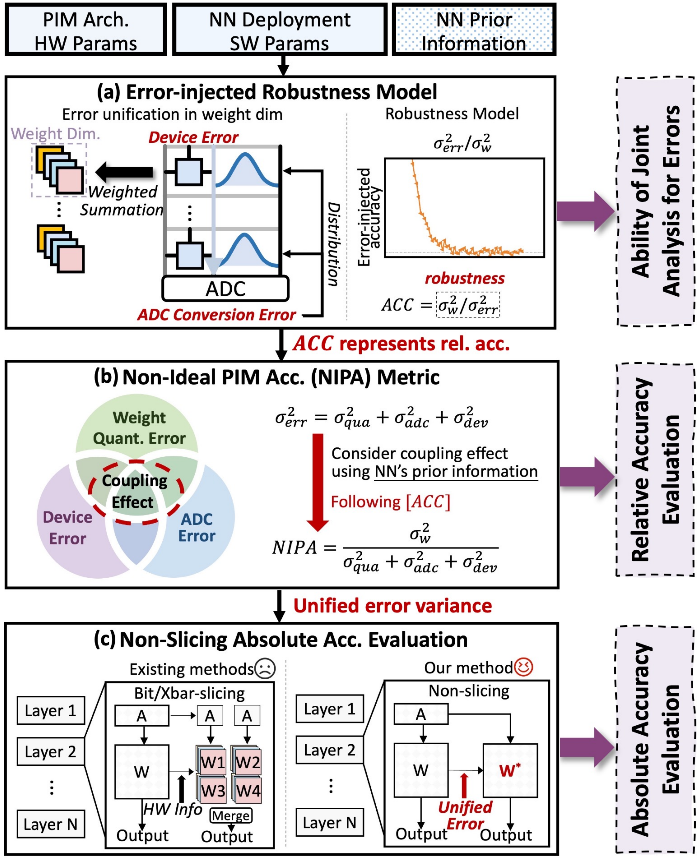
Second Challenge: ***Complex coupling effects exist among various errors***





E.g., tightly coupled device error and ADC Conversion Error

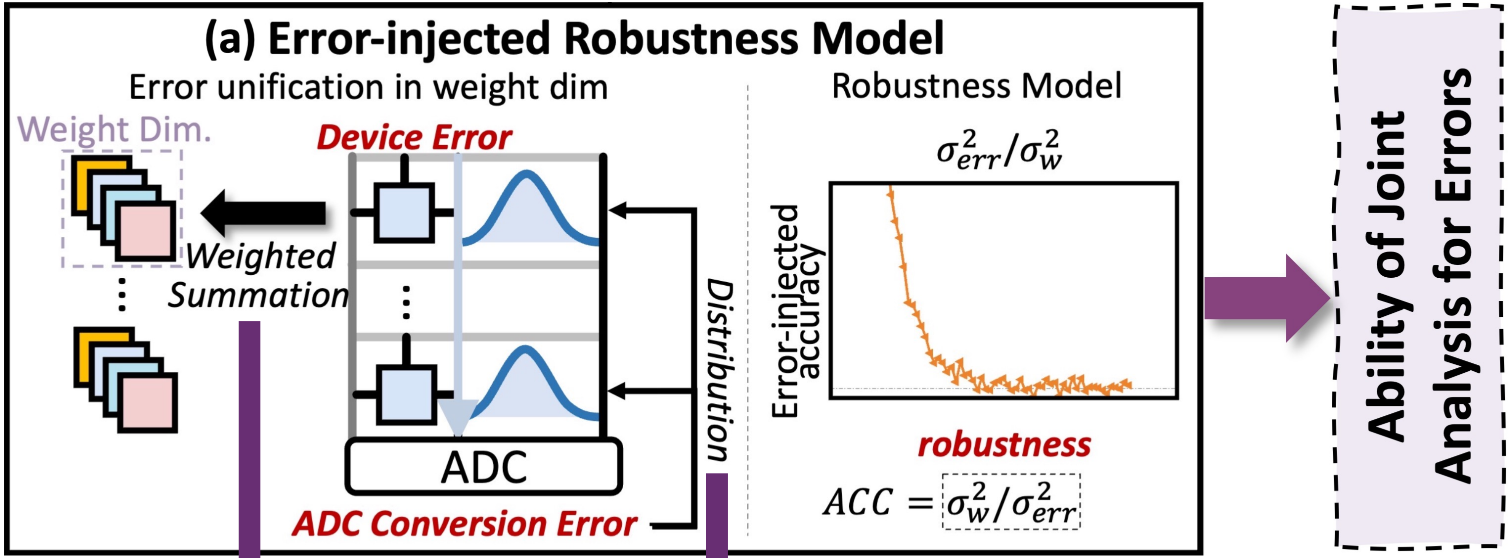
Methodology Overview



Error-Injected Robustness Model

- **Fundamental Insight: Errors injected in different dims can be mapped to weight dim through distribution and weighted summation**

$$E_{qua} + E_{adc\_weight} + E_{dev\_weight}$$

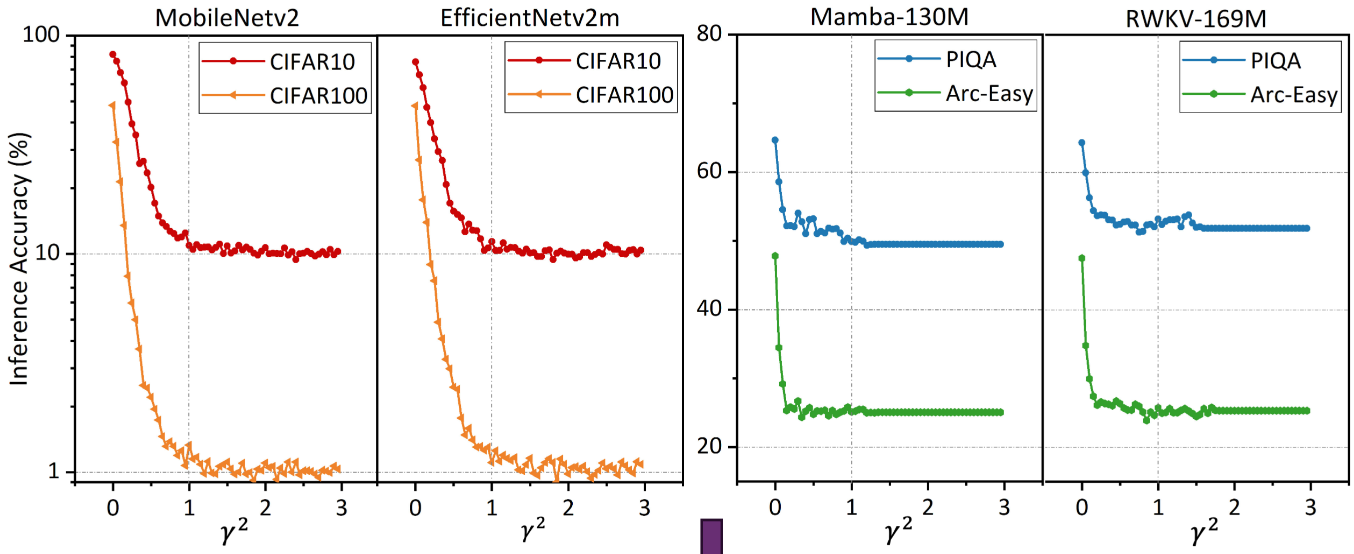


$$E_{dev\_weight} = s * \sum_{i=0}^{b_w-1} 2^i \cdot E_{dev_i} \qquad E_{adc} \Rightarrow \sum_{i=0}^{S-1} E_{adc\_dev_i}$$

Error-Injected Robustness Model

➢ **Oracle Exp: Explore the relationship between weight error E and model accuracy**

$$\mathbf{W}^* = \mathbf{W} + \mathbf{W} \cdot \mathcal{N}(0, \gamma^2)$$



$$ACC = \sigma_w^2 / \sigma_{err}^2$$

NIPA: Non-Ideal PIM Acc Metric

- **Mathematically derive the NIPA metric based on robustness Model**

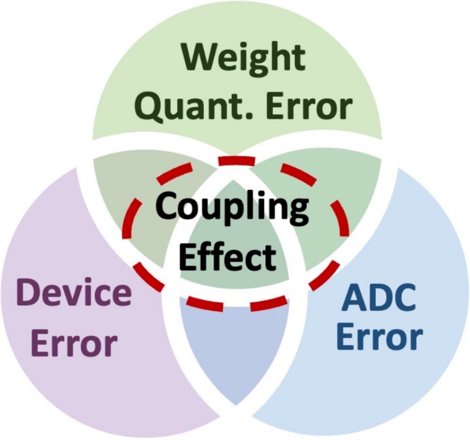
$$\sigma_w^2 = s^2 \cdot \sum_{i=0}^{b_w-1} 2^{2i} \sigma_{w_i}^2$$

$$\sigma_{adc_i}^2 = \frac{S \cdot (1 - sp_i)^2}{12 \cdot 2^{2a_r}} \cdot k$$

$$\sigma_{qua}^2 = \int_{-\frac{Q_{INR}}{2}}^{\frac{Q_{INR}}{2}} \frac{1}{Q_{INR}} \cdot x^2 dx = 3\sigma_w^2 / 2^{2b_w}$$

$$\sigma_{dev}^2 = s^2 \cdot \sum_{i=0}^{b_w-1} 2^{2i} \cdot \gamma^2 \cdot \sigma_{w_i}^2 \cdot A_i$$

(b) Non-Ideal PIM Acc. (NIPA) Metric



$$\sigma_{err}^2 = \sigma_{qua}^2 + \sigma_{adc}^2 + \sigma_{dev}^2$$

Consider coupling effect using NN's prior information

Following $[ACC]$

$$NIPA = \frac{\sigma_w^2}{\sigma_{qua}^2 + \sigma_{adc}^2 + \sigma_{dev}^2}$$

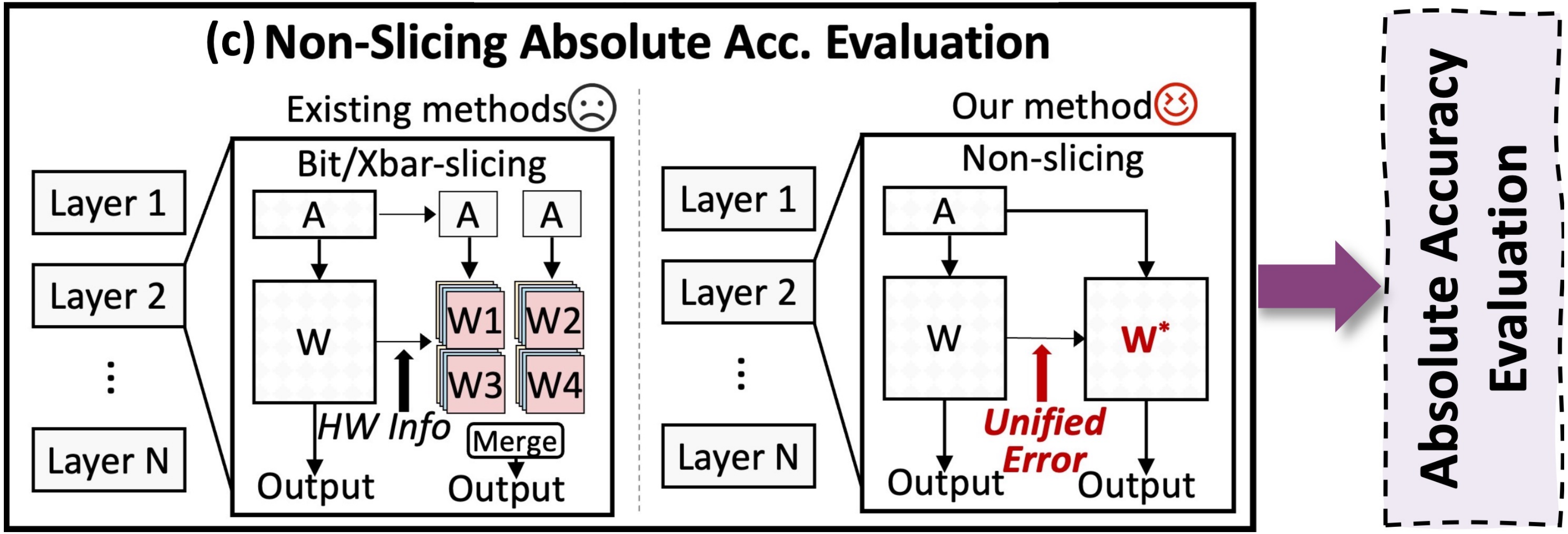
Relative Accuracy Evaluation

Non-Slicing Absolute Acc Evaluation

- **Unified error is directly injected to pre-trained NN weight values**

  The calculation and injection of error in the absolute accuracy evaluation part is performed layer by layer.

$$\mathbf{W}^* = \mathbf{W} + \mathbf{W} \cdot \mathcal{N}(0, \frac{\sigma_{err}^2}{\sigma_w^2})$$



Without the need of time-consuming bit/crossbar slicing process

Comparison with existing work

	Abs. Acc	Rel. Acc	Coupling Effect	Error-injected Dim.	Models	Non-ideal Factors
DNN+NeuroSim [4]	Slow	✖	Strong	Device/Psum-level	CNN	Quant./ADC/Device*
MNSIM2.0 [5]	Slow	✖	Strong	Device/Psum-level	CNN	Quant./ADC/Device*
DL-RSIM [6]	Slow	✖	Weak	Psum-level	CNN	Quant./Device*
Geniex [8]	Slow	✖	Strong	Device/Psum-level	CNN	Quant./ADC/Xbar
PytorX [9]	Slow	✖	Weak	Device-level	CNN	Crossbar/Device*
MICSim [10]	Slow	✖	Weak	Device/Psum-level	CNN/LLM	Quant./ADC
CoMN [11]	Slow	✖	Strong	Device/Psum-level	CNN	Quant./ADC/Xbar/Device*
Swordfish [7]	Slow	✖	Strong	Device/Psum-level	Basecaller	Quant./ADC/Device*
RxNN [23]	Slow	✖	Strong	Device/Psum-level	CNN	Quant./ADC/Device*
Yan, et al. [12]	Fast	✖	✖	Device-level	CNN	Device*
Gibbon [13]	✖	Predictor	Weak	Device/Psum-level	CNN	Quant./ADC
Unified-QCN [14]	✖	Metric	Weak	Device/Psum-level	CNN	Quant./ADC/Device*
This Work	**Fast**	**Metric**	**Strong**	**Weight-level**	**CNN/LLM**	**Quant./ADC/Xbar/Device***

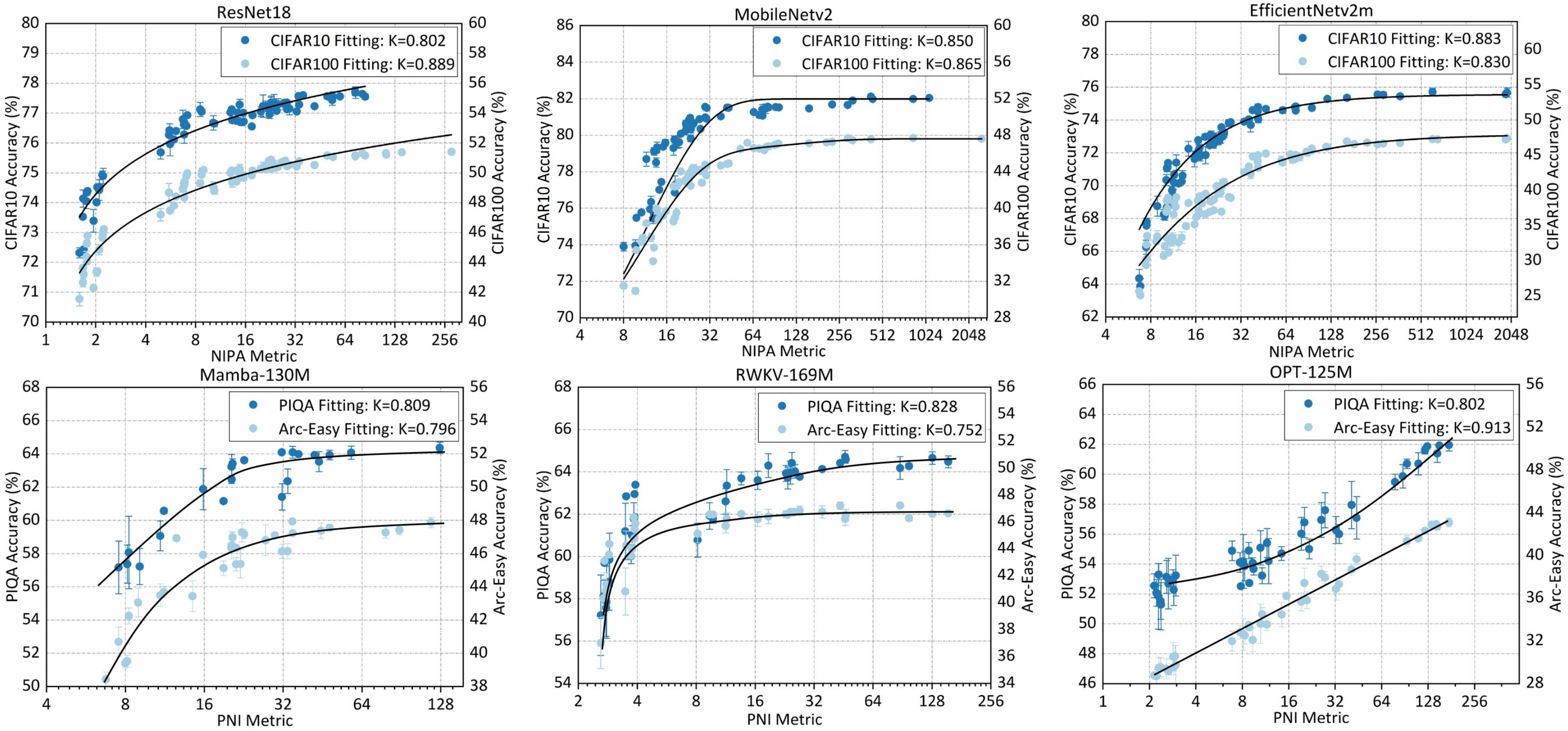
Device*: different types of device conductance deviation caused by inaccurate programming, Stuck-at-Fault(SAF), conductance retention, and so on.
Strong: comprehensive consideration of coupling effects is achieved by injecting all errors into the slicing computation or directly analyzing their interactions in derivation.
Weak: insufficient consideration of coupling effects due to the incomplete consideration of non-ideal factors or the indirect modeling by NN-based predictor.

Evaluations

- **Experiment Setup**
  - **Models:**
    - **Basic CNN:** ResNet18
    - **Lightweight CNNs:** MobileNetv2, EfficientNetv2m
    - **Attention-free LLMs:** Mamba-130M, RWKV-169M
    - **Attention-based LLM:** OPT-125M
  - **Datasets:** CIFAR10/100, PIQA, Arc-Easy
  - **Metrics:** KenDall Rankng Correlation for rel. acc. Evaluation, Mean Absolute Error (MAE) for abs. acc. Evaluation
  - **Ground Truth:** DNN+NeuroSim
  - **Parameter Space:**
    weight bit-width {4,6,8}, device conductance variation factor {0,0.1,0.2},
    ADC resolution {4,5,6,7,8}, crossbar size {128,256}

Evaluations

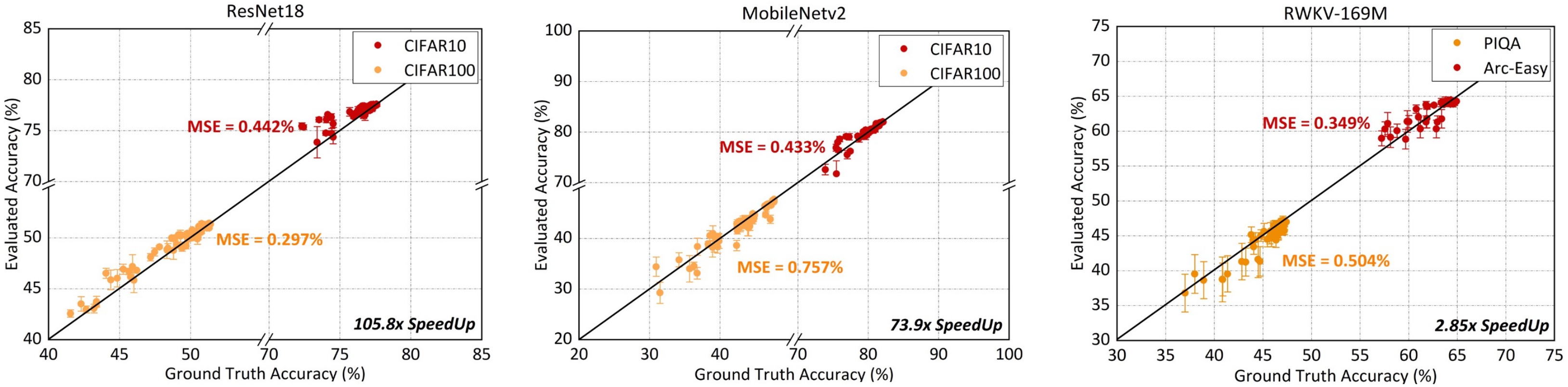
➢ Performance of NIPA Metric



High correlations between NIPA and ground truth accuracy provided by bit-and-crossbar slicing simulations.

Evaluations

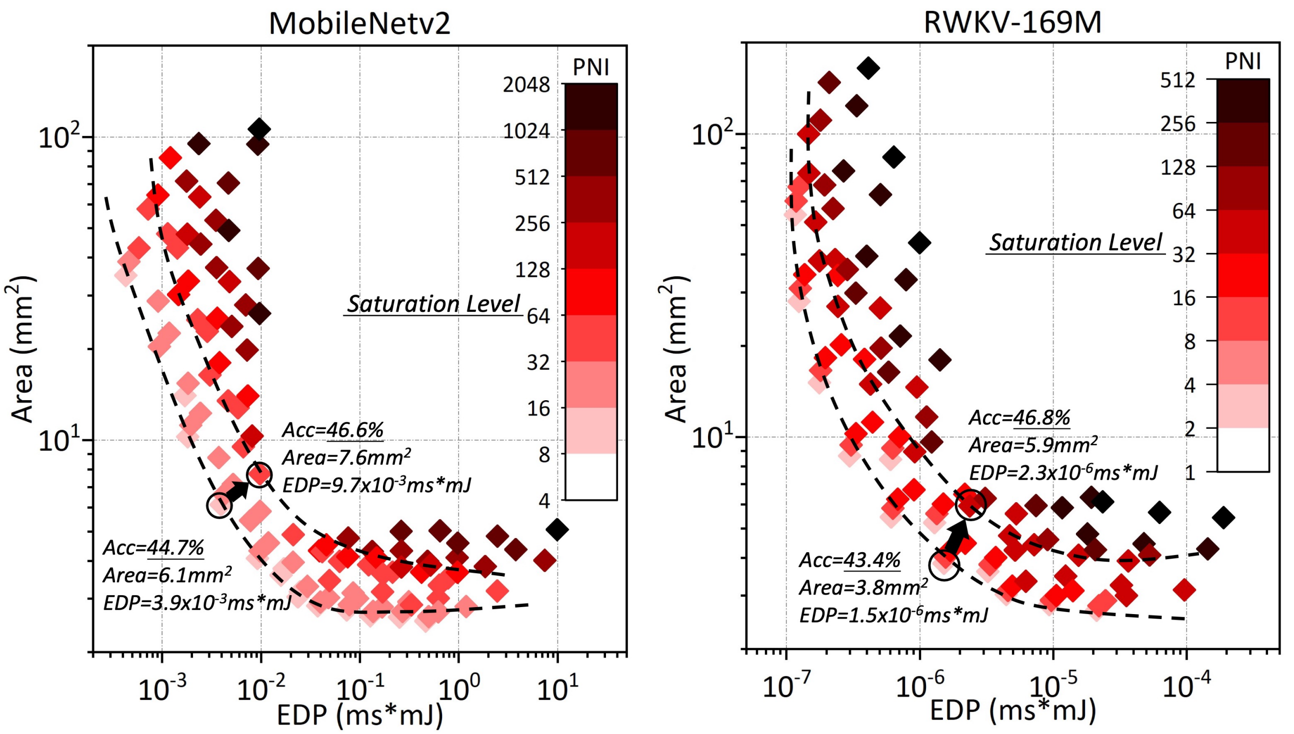
➢ **Non-Slicing Absolute Accuracy Evaluation Performance**



MAE between ground truth accuracy and results of our non-slicing absolute accuracy evaluation method.

Evaluations

➢ **Co-Exploration Results Embedded with NIPA Metric**



NIPA helps to achieve a better trade-off between algorithm and hardware performance.

Thanks for your attention!
Q&A

Lidong Guo[1*], Zhenhua Zhu[1*], Qiushi Lin[1], Xuefei Ning[1], Yuan Xie[2], Huazhong Yang[1], Wangyang Fu[1], Yu Wang[1]

[1]Tsinghua University, [2]HKUST
E-mail: gld21@mails.tsinghua.edu.cn, zhenhuazhu@mails.tsinghua.edu.cn, yu-wang@tsinghua.edu.cn