

On Chip Customized Learning on Resistive Memory Technology for Secure Edge AI

M. Pallo^{*1,2}, S. D'Agostino^{*2}, M. Piccoli^{2,3}, D. Bonnet², N. Castellani²,
G. Piccolboni¹, M. A. Iftakher³, J.-F. Nodin², F. Andrieu²,
D. Querlioz³, G. Molas¹, L. Hutin², E. Vianello²

** These authors contributed equally to this work*

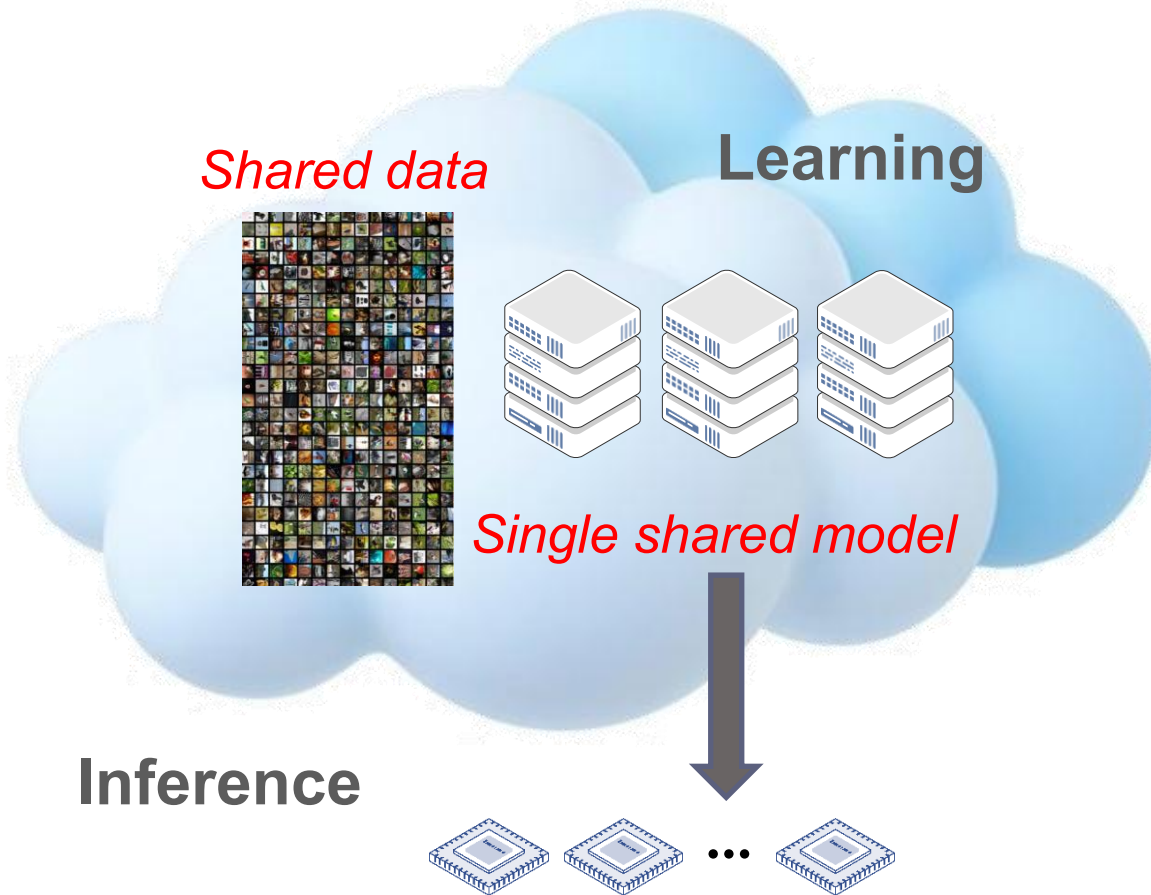
¹ Weebit Nano FR

² CEA-Leti, Univ. Grenoble Alpes

³ Univ. Paris-Saclay, CNRS



Today: Training in the Cloud, Inference at the Edge



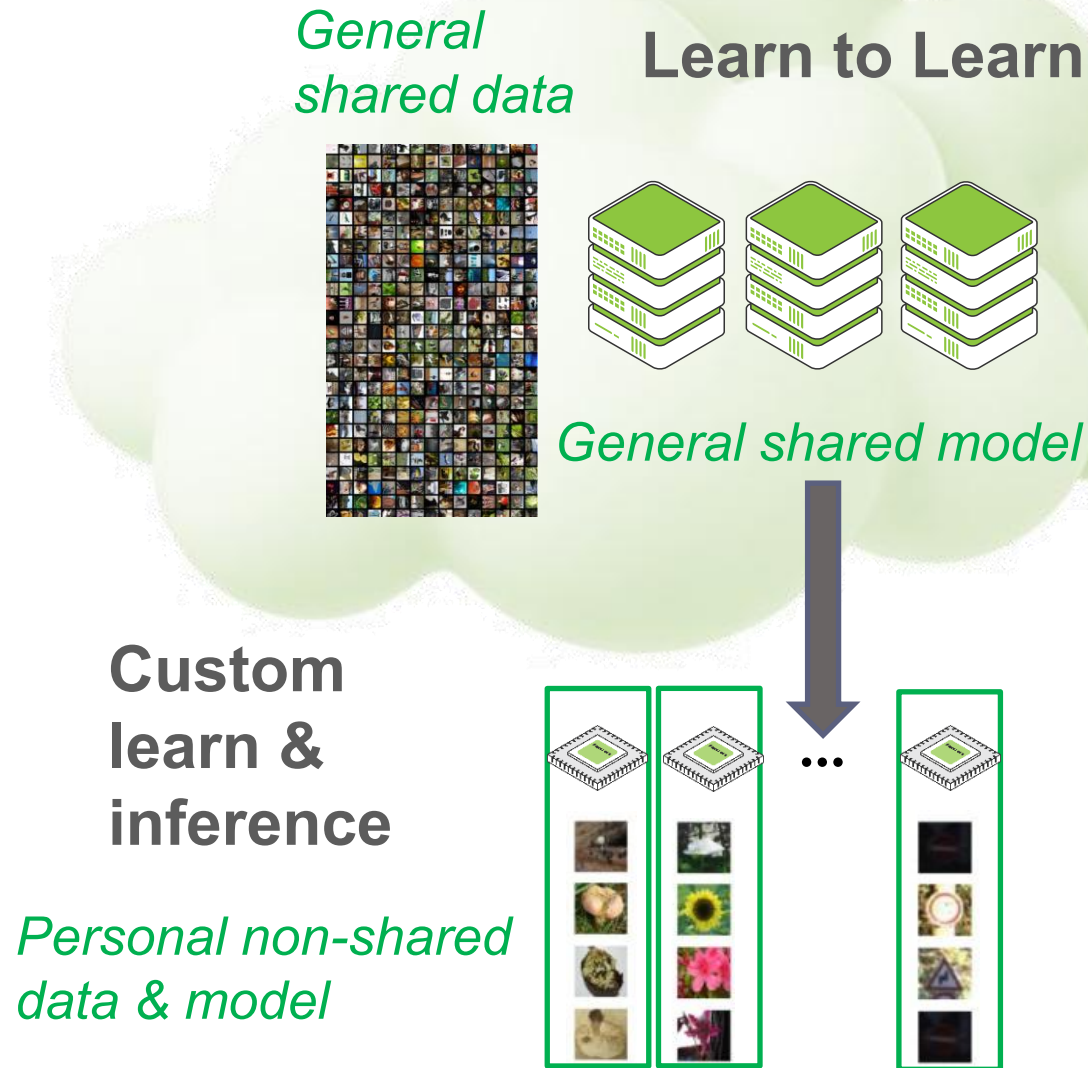
- Burden of training is centralized in the cloud

BUT

- User data shared in the cloud
- A single shared model for all users

Inadequate for confidential or customized use cases

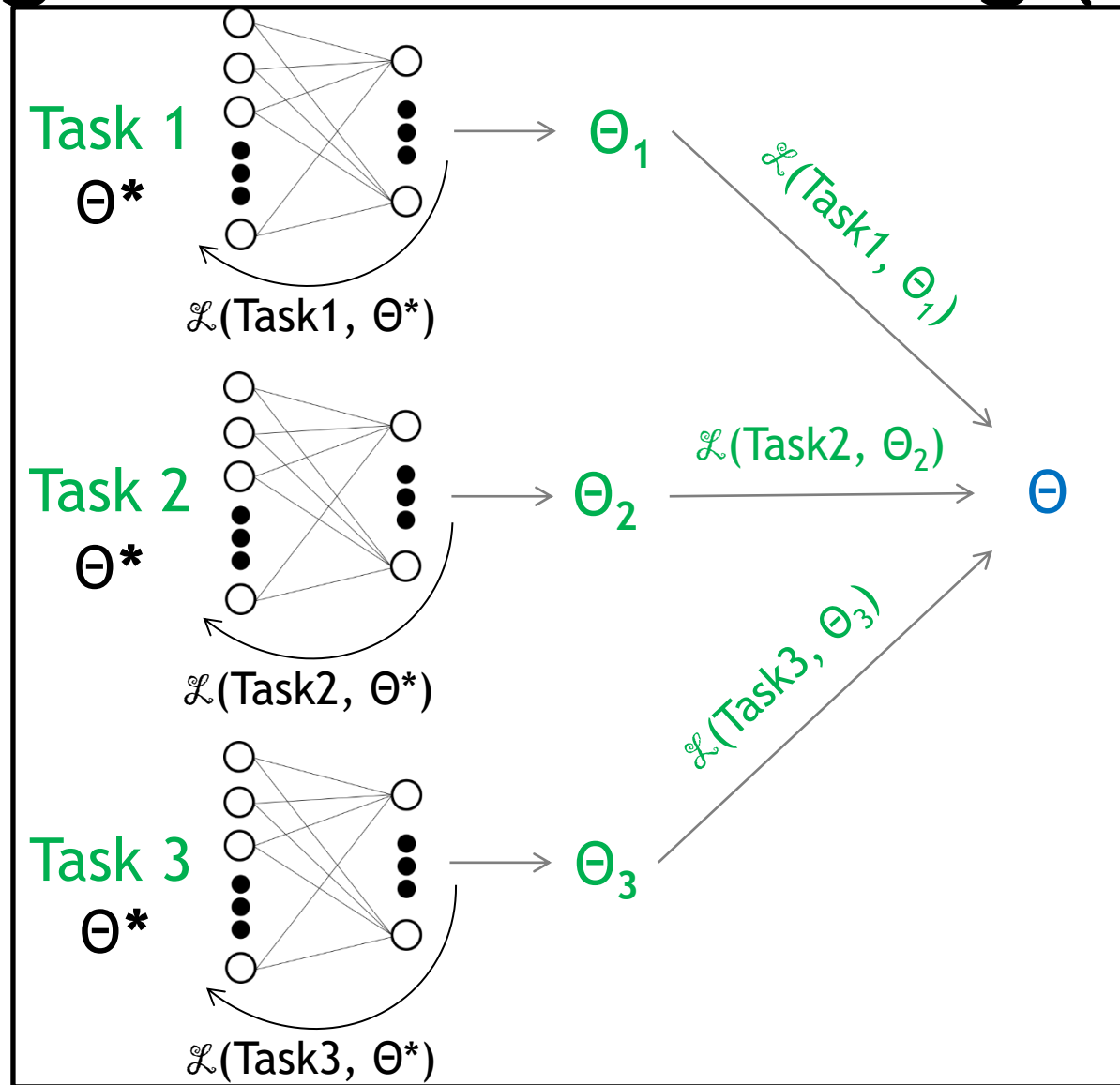
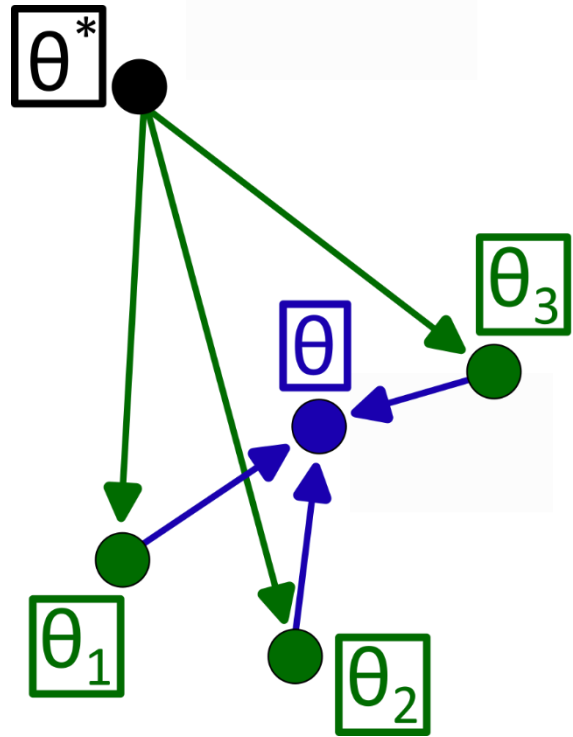
Tomorrow: Distributed Learning for Secure AI



- Models pre-trained in the cloud can quickly adapt on-chip to new tasks
- Prevents data & model sharing
- Custom model for each user

Training on-chip is mandatory for secure & personalized AI

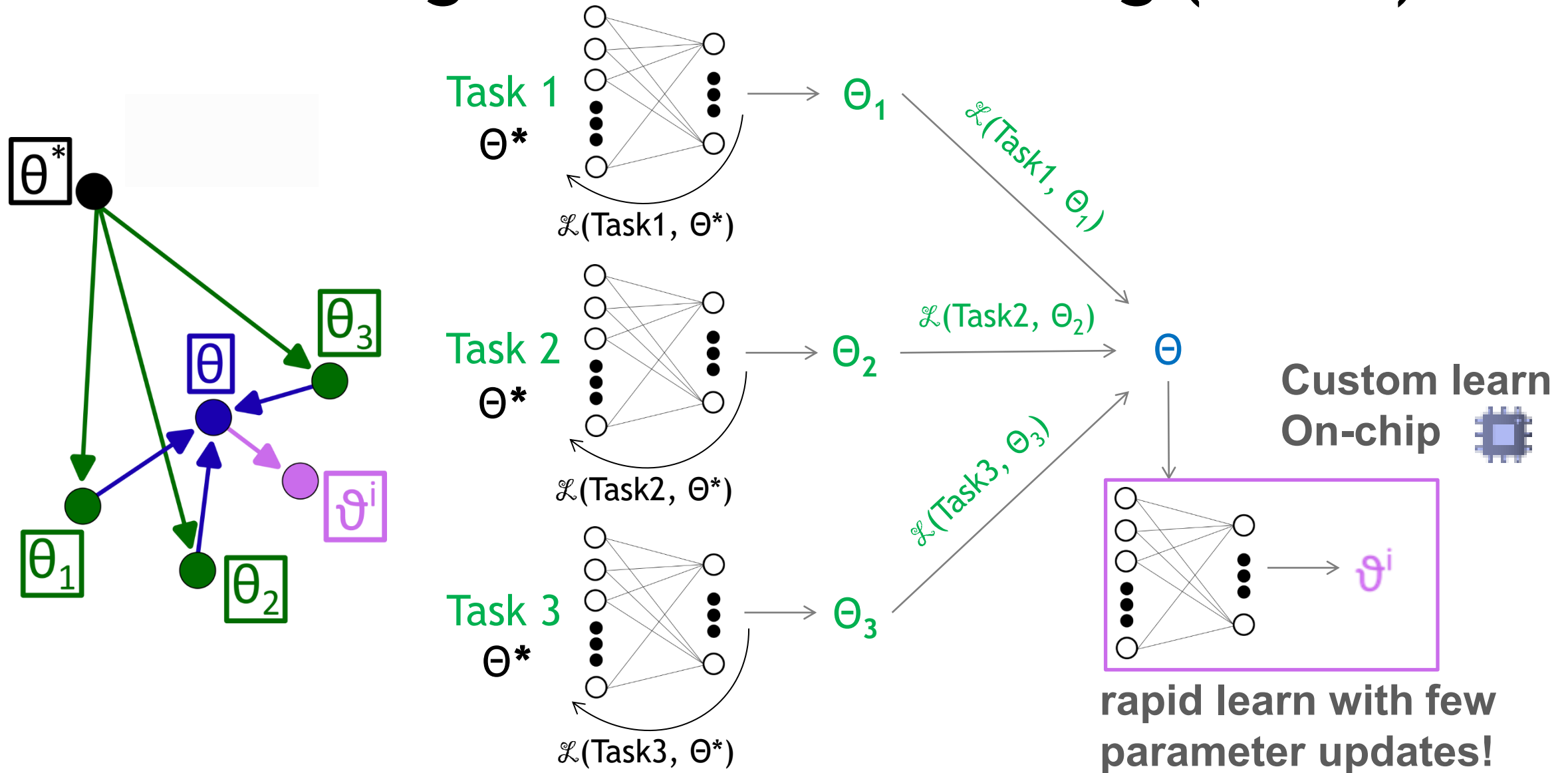
Model-Agnostic Meta-Learning (MAML)



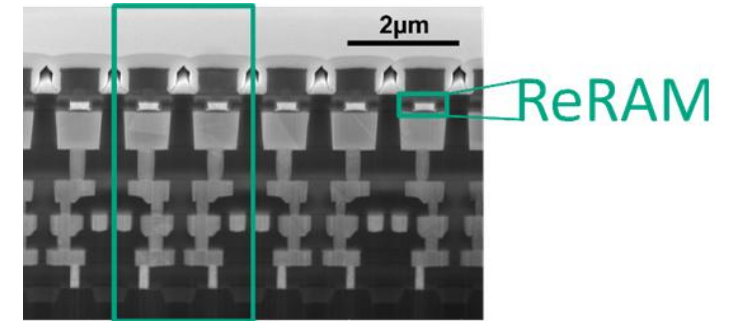
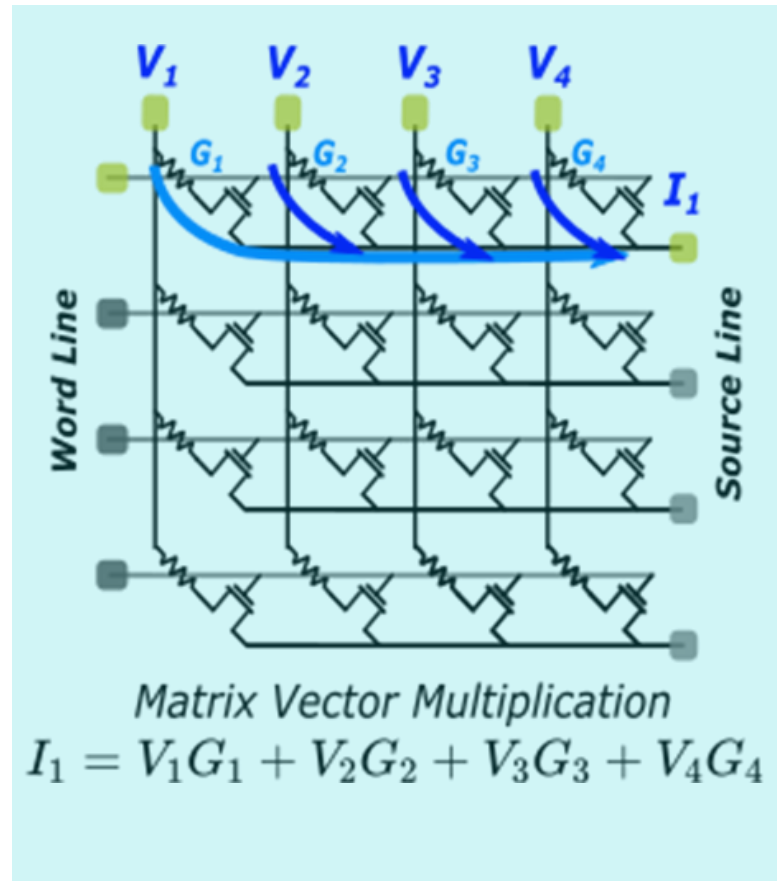
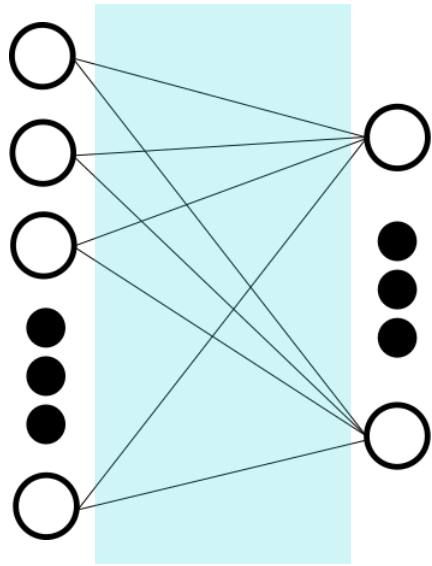
Learn to learn
Off-chip



Model-Agnostic Meta-Learning (MAML)



Resistive Memories for In-Memory-Computing



A crossbar of ReRAM naturally performs AI linear computation

- Demonstrated tens of TOPS of inference [IBM, Nature 2023] [TSMC Nature 2025]
- Zero standby power thanks to non-volatility

Objectives of this work

To equip neural networks with learning capabilities at the edge.

Our Approach:

- **Algorithm:** *Learning-to-learn* (meta-learning) – enables fast adaptation in just a few steps and is robust to hardware non-idealities.
- **Hardware:** *Neuromorphic hardware based on resistive memory* – energy-efficient due to in-memory computing (IMC) and non-volatility, ideal for normally-off systems.

Outline

I. Analog ReRAM for Meta-learning

II. Few Shot on Chip Learning Experiments

III. Conclusions

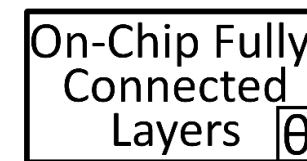
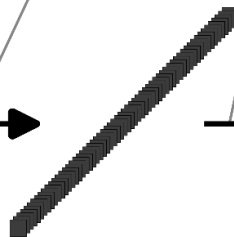
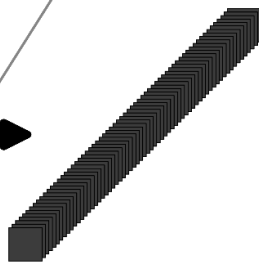
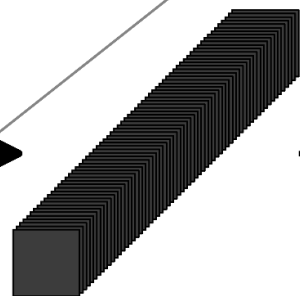
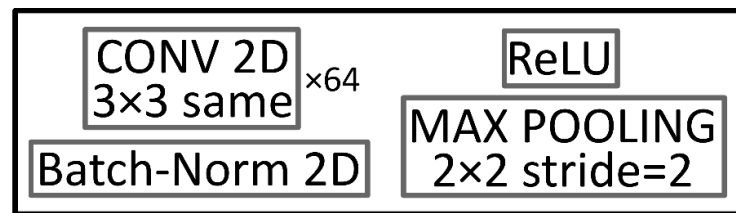
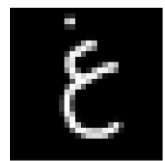
Our network

Objective: learn to recognize new characters from the Omniglot dataset

4 convolutional layers: *hard-wired* (non-trainable)

2 fully connected layers: *trainable* (updatable)

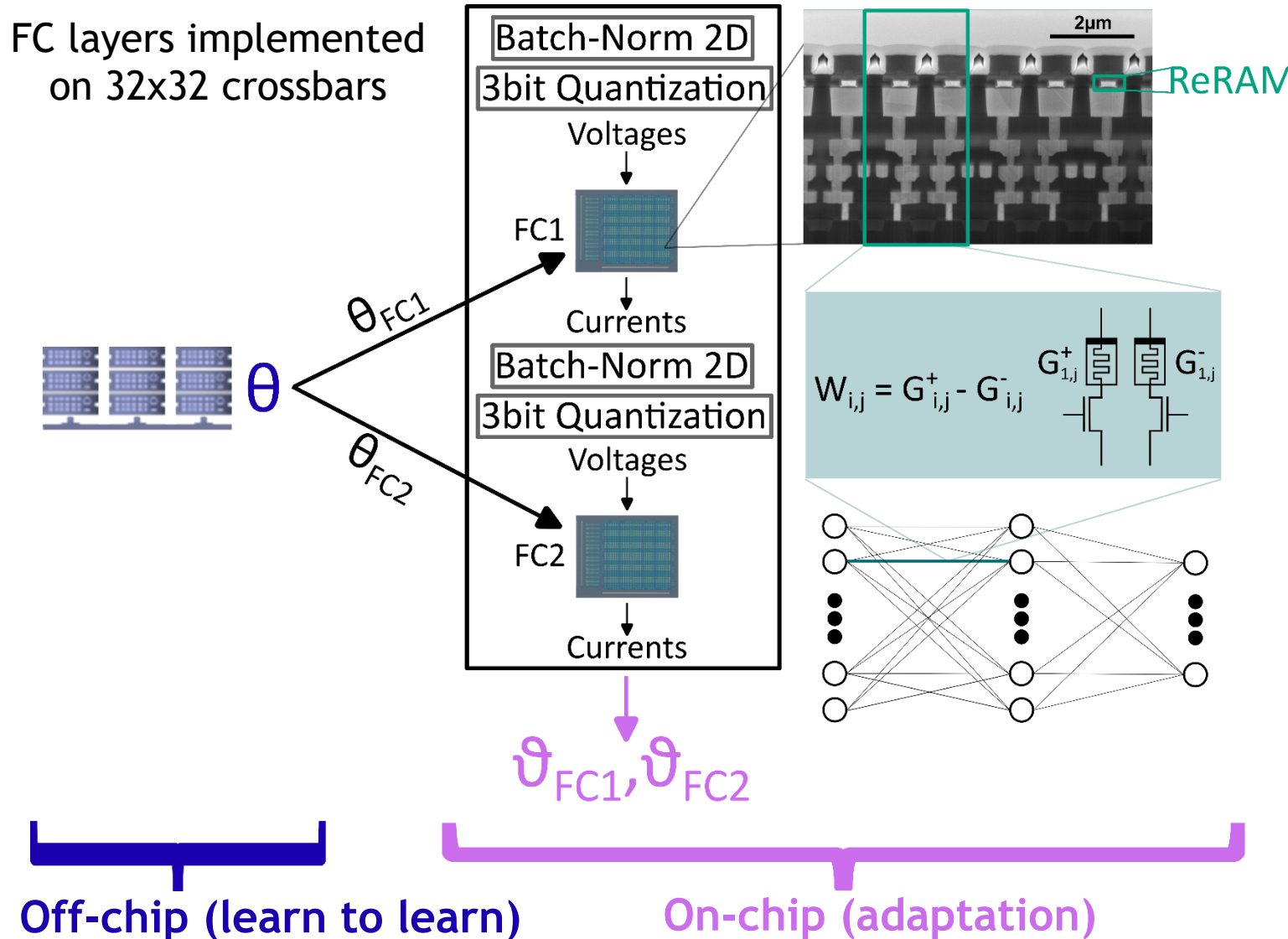
उ ष ङ ञ ऒ
क ञ ष ऒ उ
ष ञ ऒ उ क
ऋ ऒ ष क उ
क ष उ ऒ ञ



Pred

On-chip learning with ReRAM

FC layers implemented on 32x32 crossbars

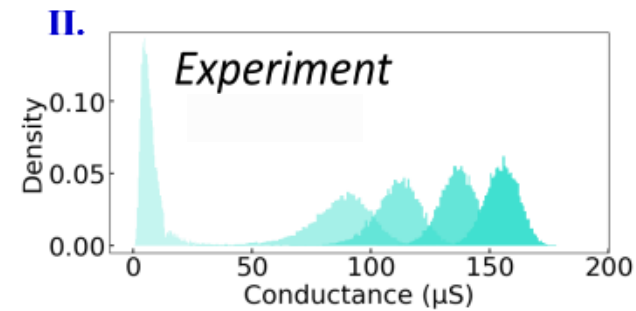
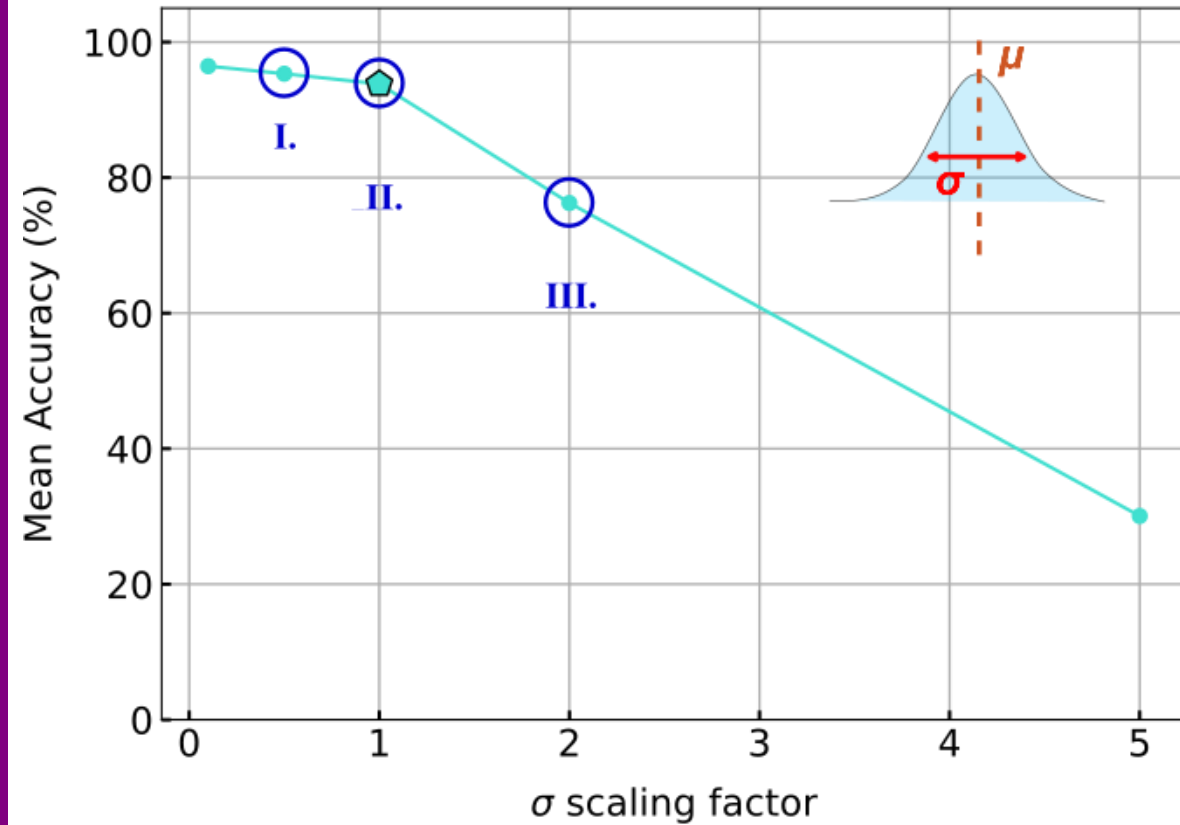


- After off-chip meta-training, the model is programmed into multi-level ReRAM crossbar arrays.

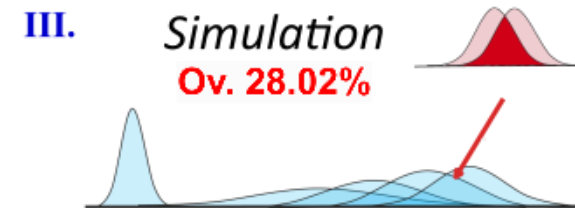
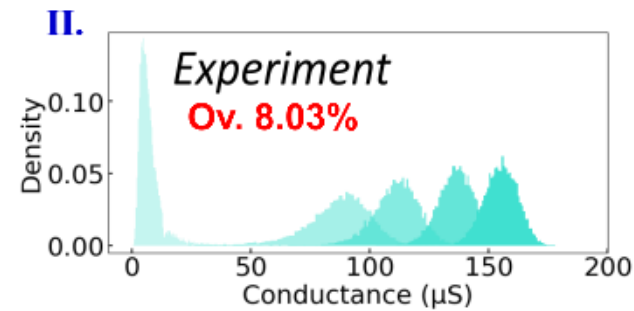
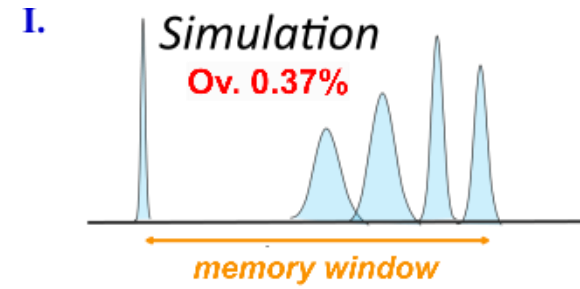
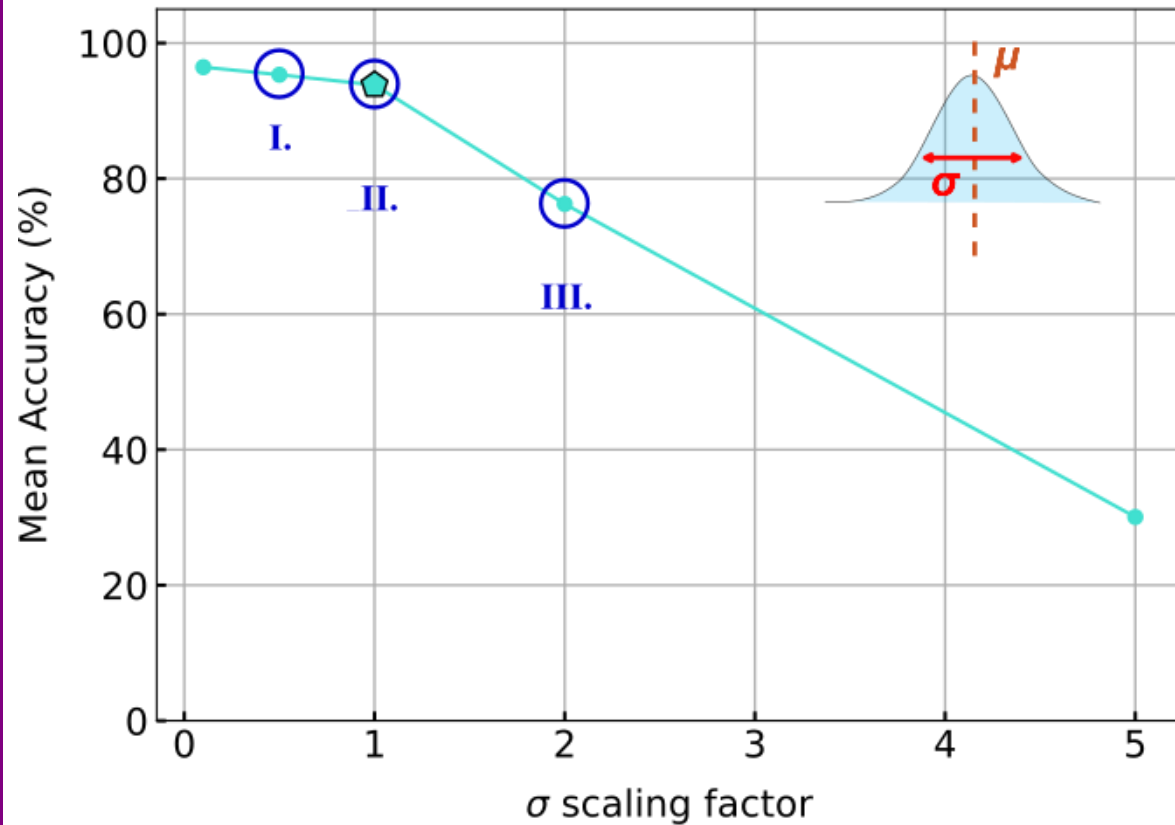
- During the adaptation phase, updates are performed directly on the ReRAM, enabling on-chip learning.

Key question: What is the optimal multi-level programming strategy?

Impact of Overlap Between Adjacent States

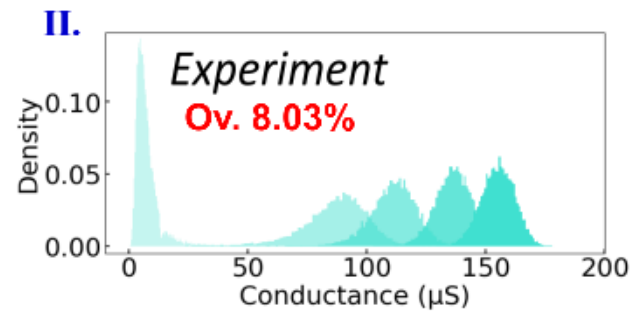
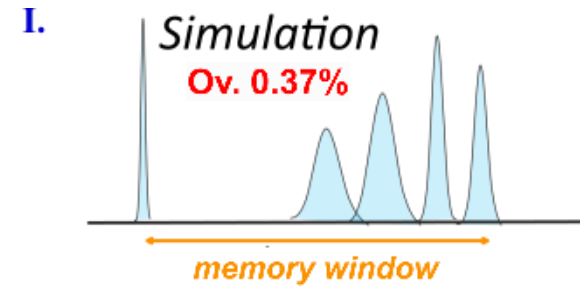
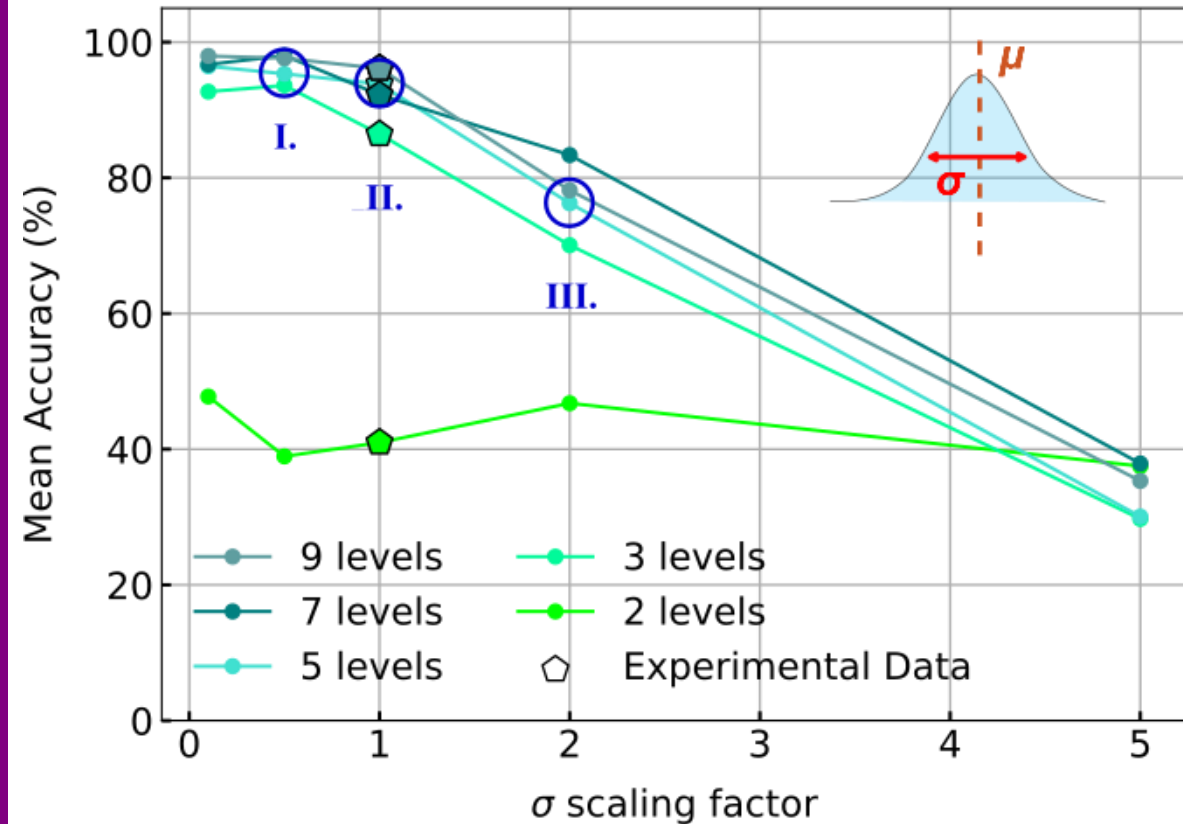


Impact of Overlap Between Adjacent States



Drop in accuracy associated to increased overlap between adjacent states

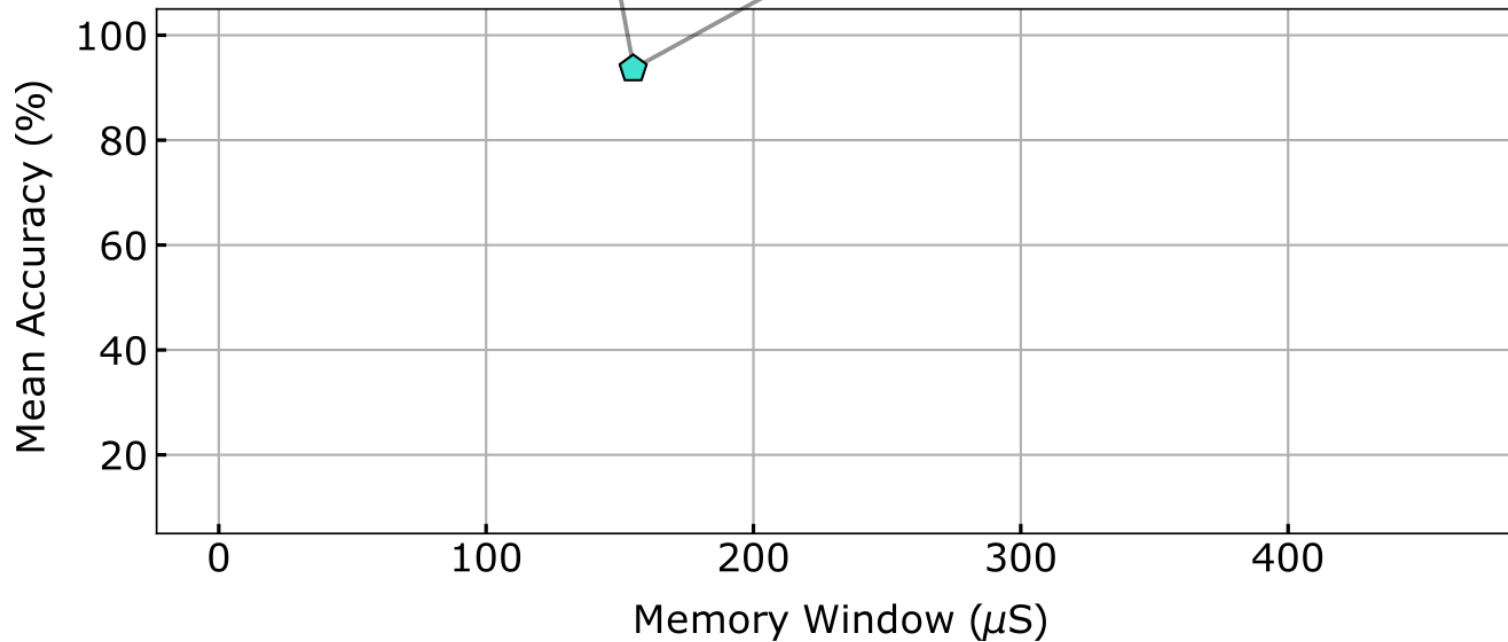
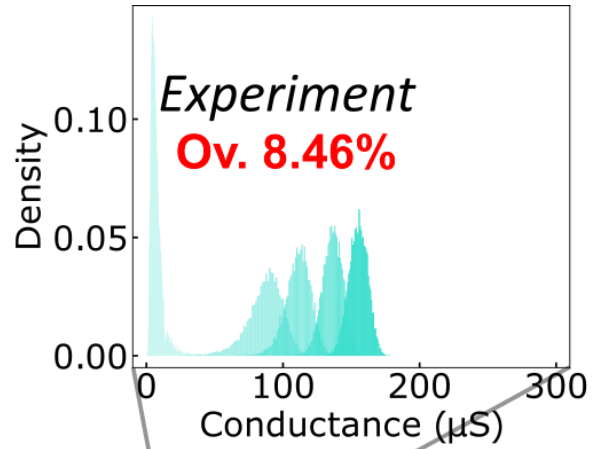
Impact of Overlap Between Adjacent States



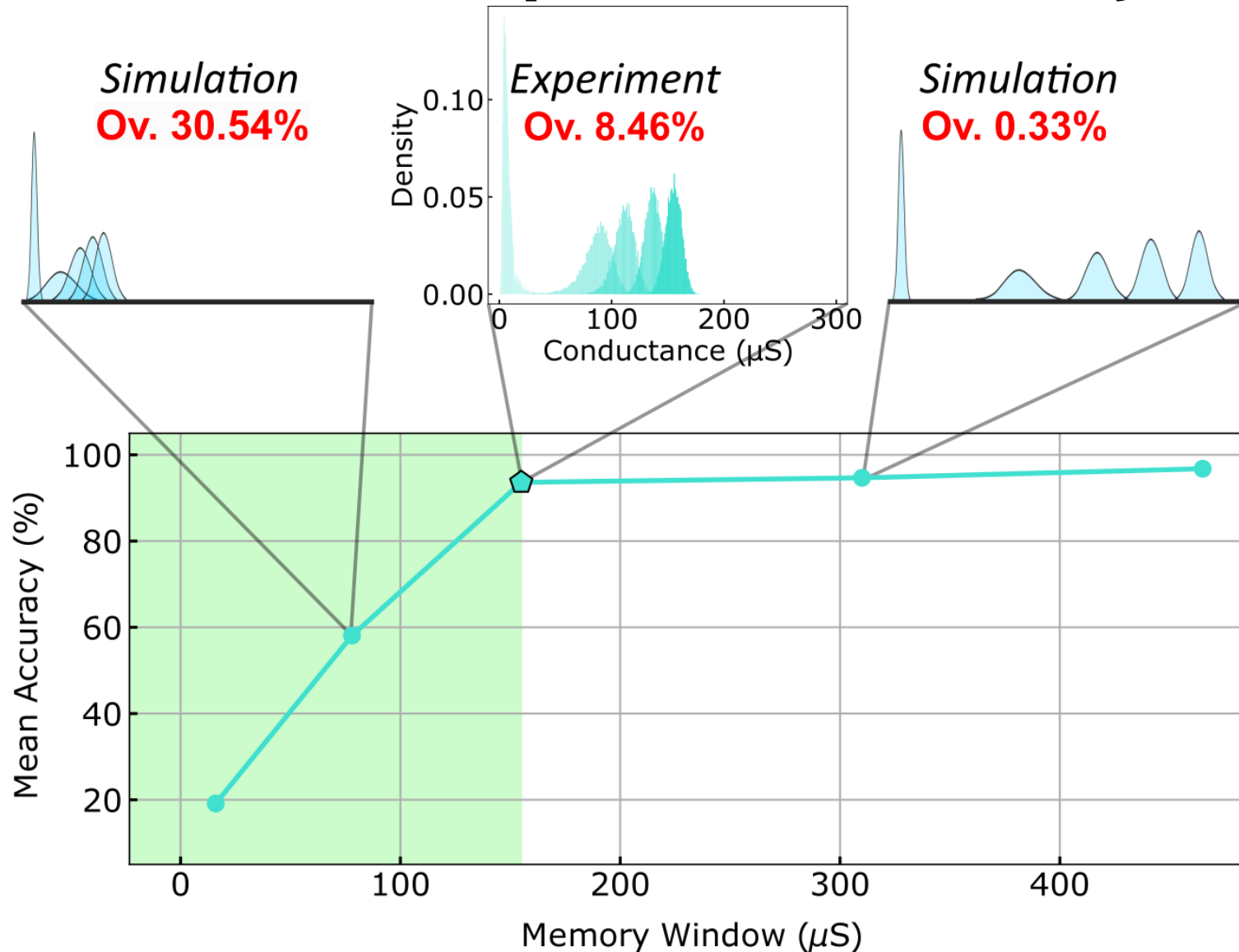
- Increased overlap between adjacent conductance states leads to a **drop in accuracy**.
- More conductance levels can slightly **improve accuracy**, but also raise the risk of state overlap.

Impact of Memory Window

ReRAM multi-level programming

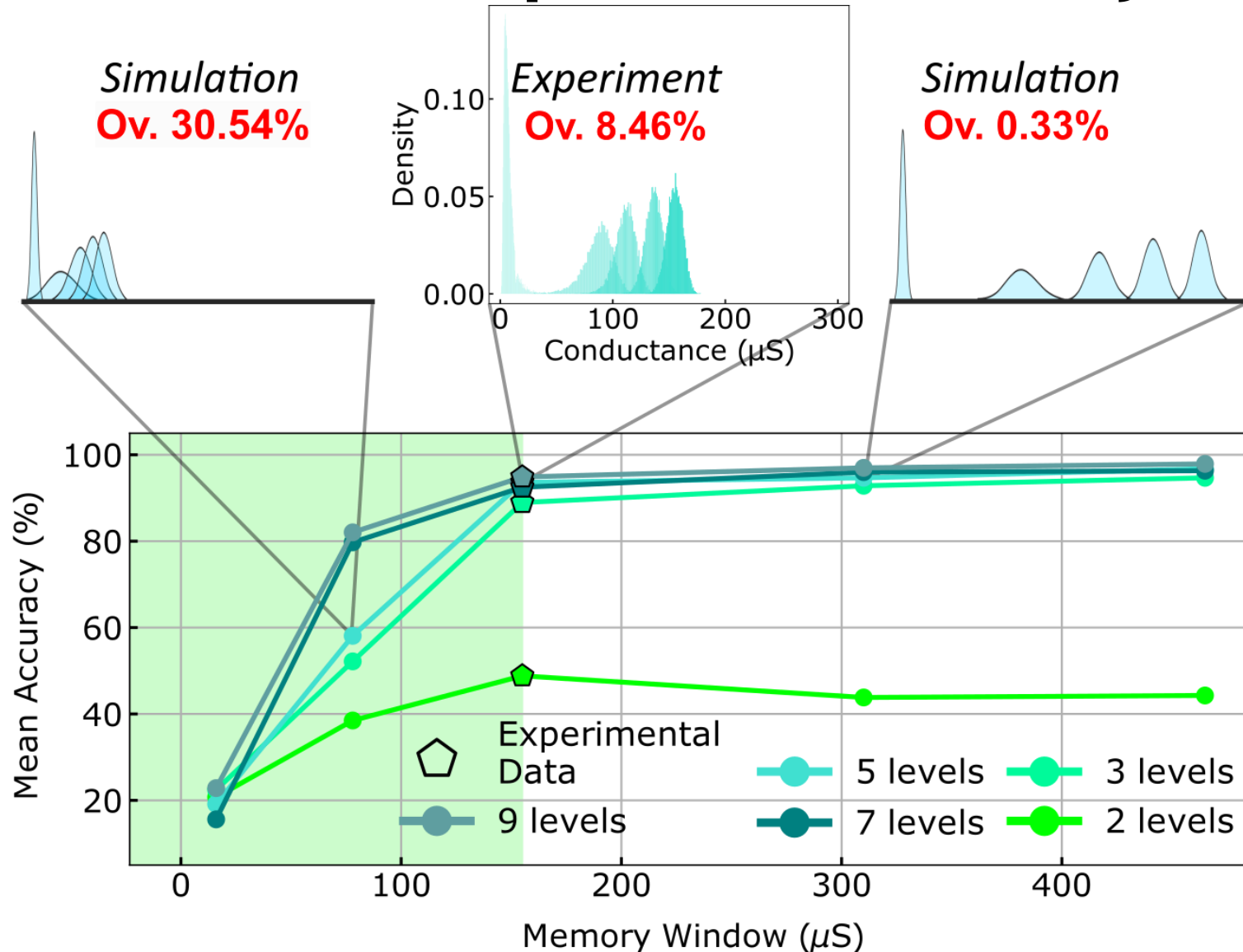


Impact of Memory Window



- Expanding the memory window improves accuracy by increasing state separation.

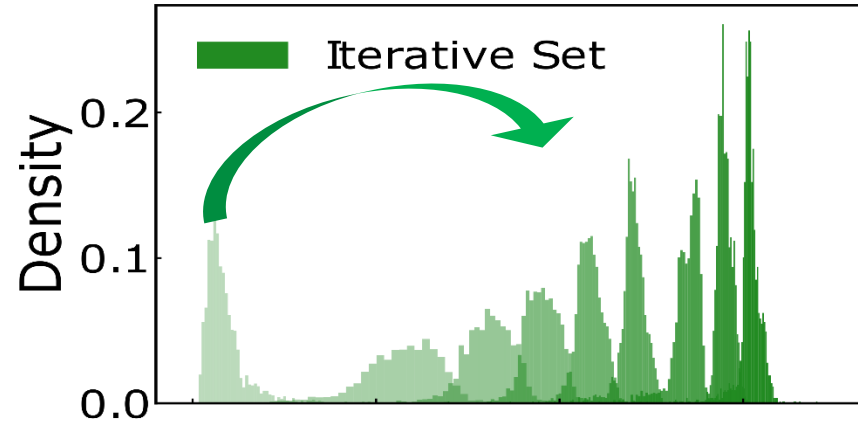
Impact of Memory Window



- Expanding the memory window improves accuracy by increasing state separation.
- Increasing the number of conductance levels slightly reduces the demand on the memory window.

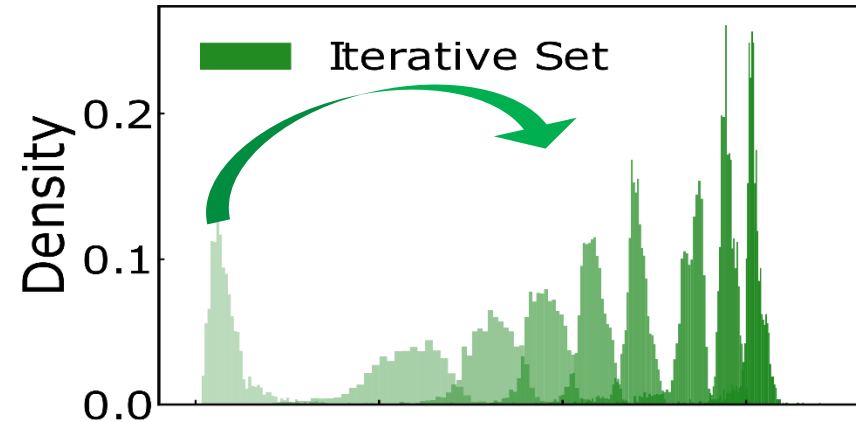
Multi-level SET vs Multi-level RESET

Smart SET drives the device to a HCS starting from a LCS.

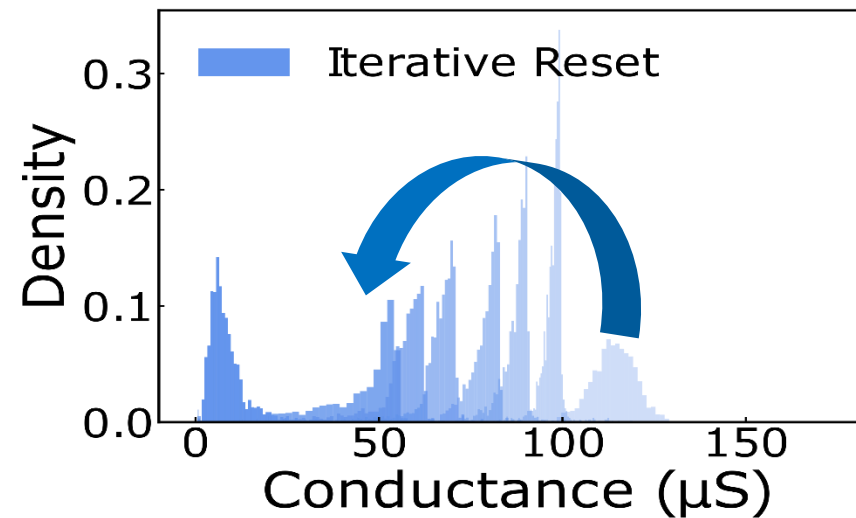


Multi-level SET vs Multi-level RESET

Smart SET drives the device to a HCS starting from a LCS.

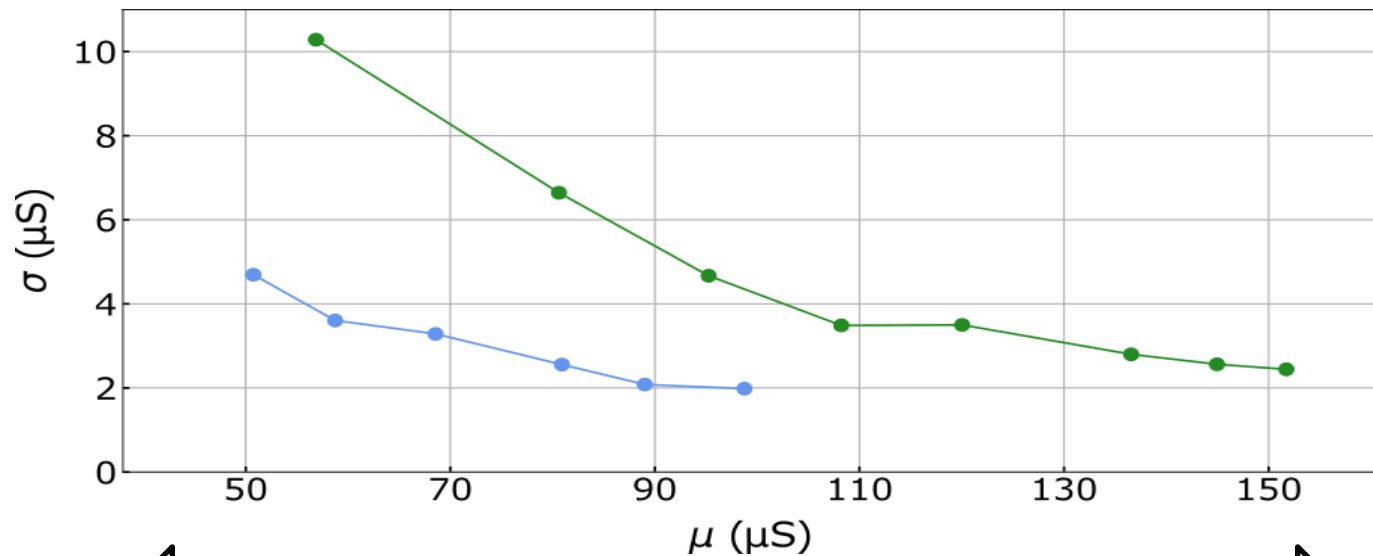
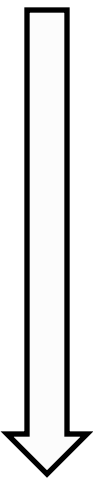


Smart RESET drives the device to a LCS starting from a HCS.

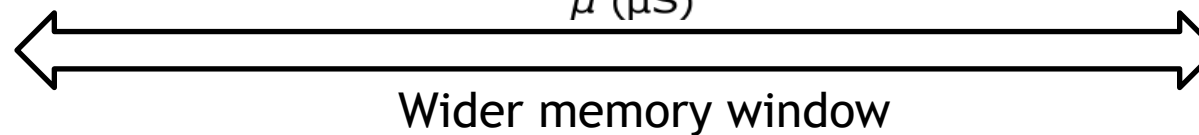


Hybrid programming strategy

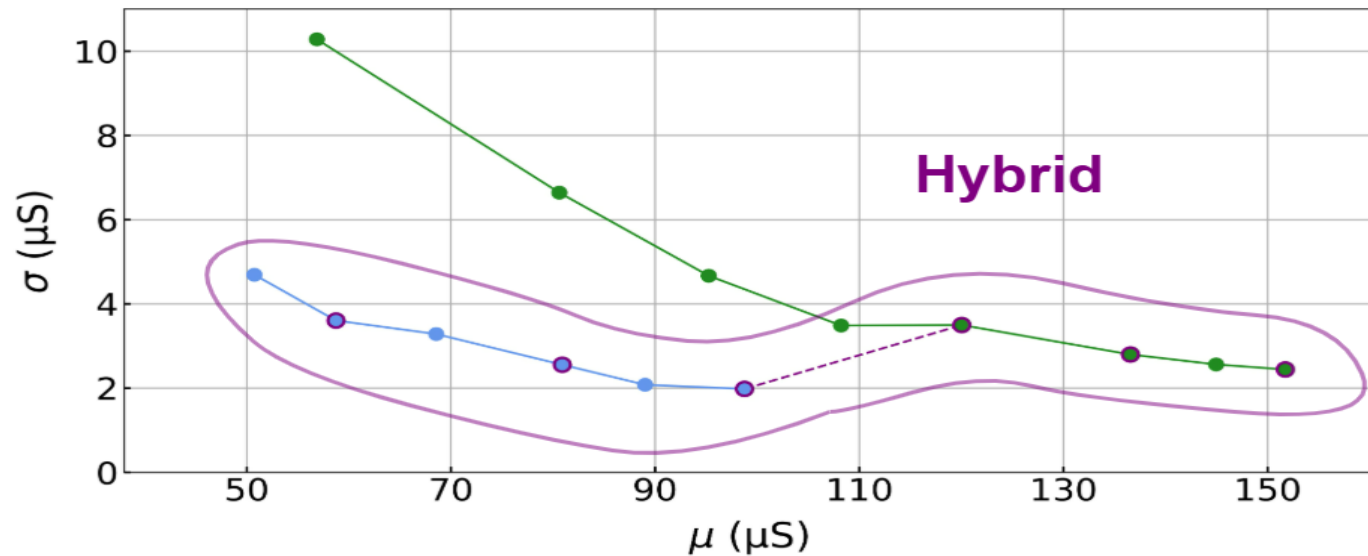
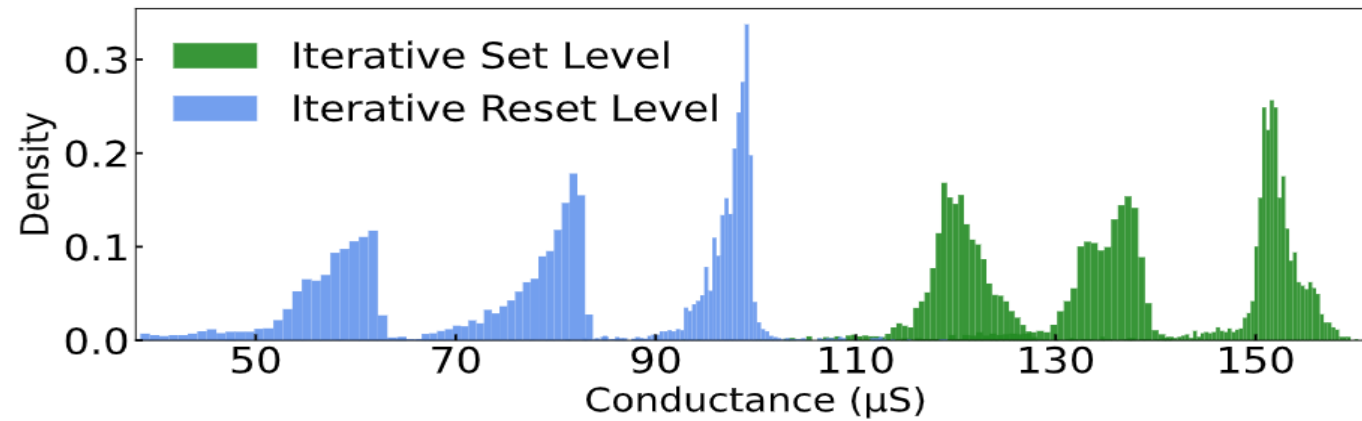
Overlap reduction



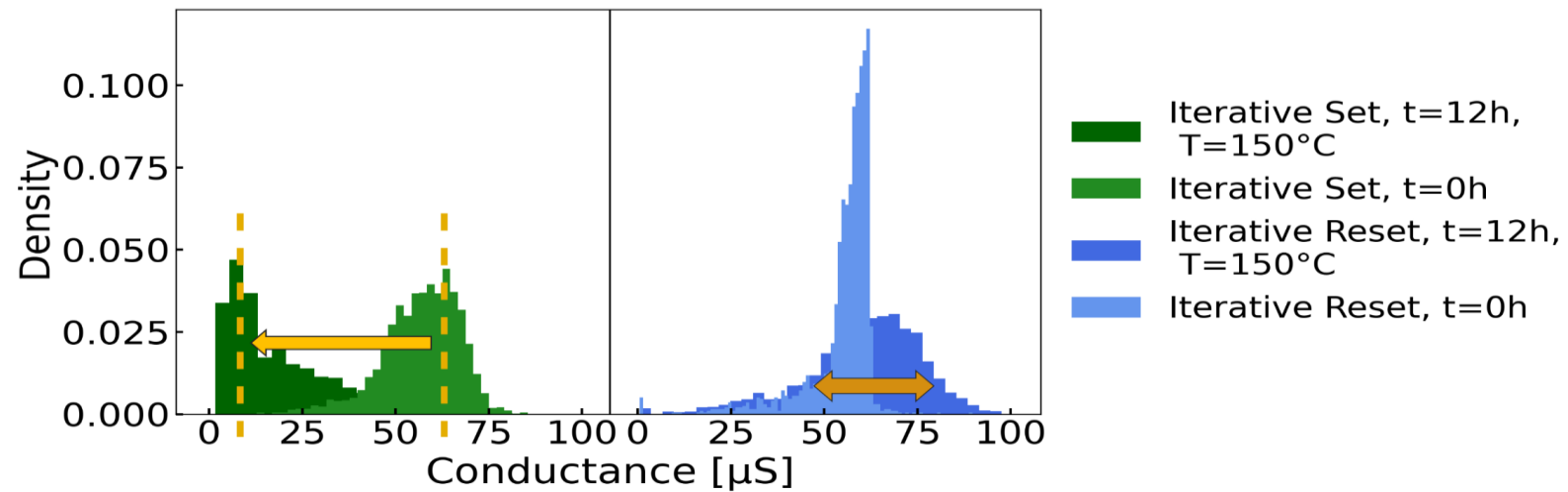
Wider memory window



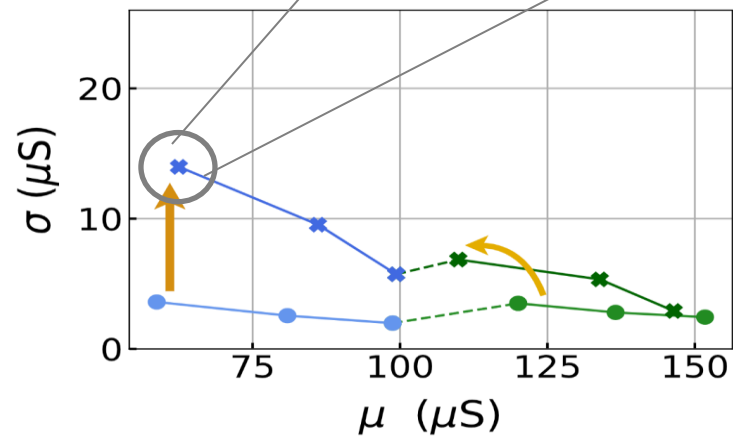
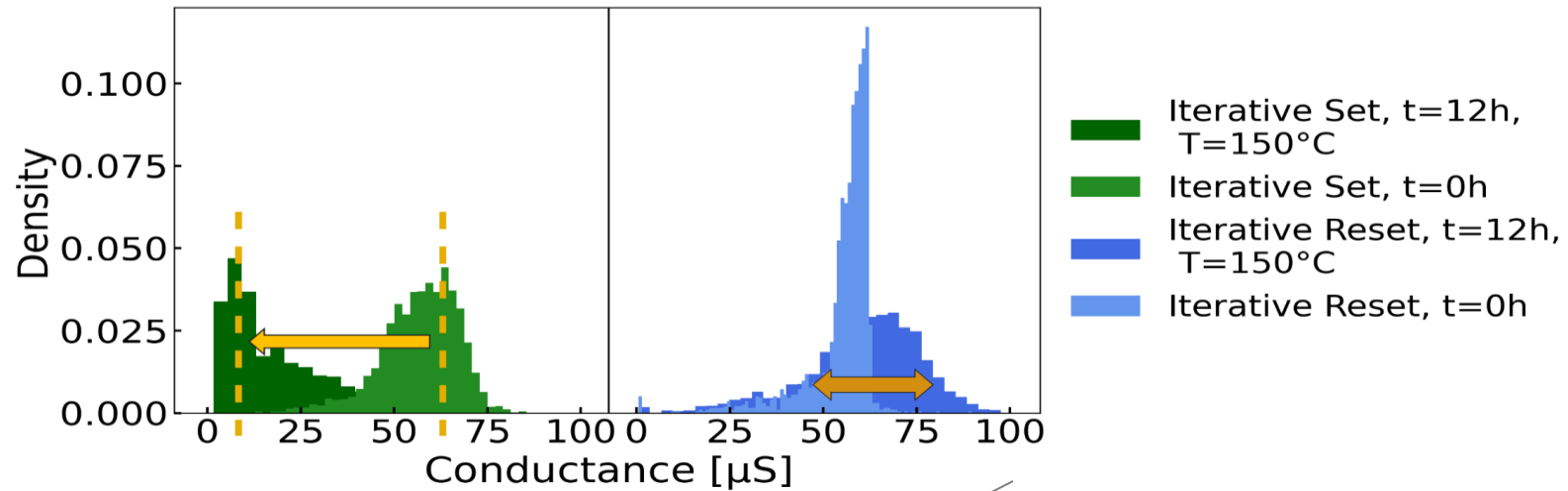
Hybrid programming strategy



Hybrid improves drift



Hybrid improves drift



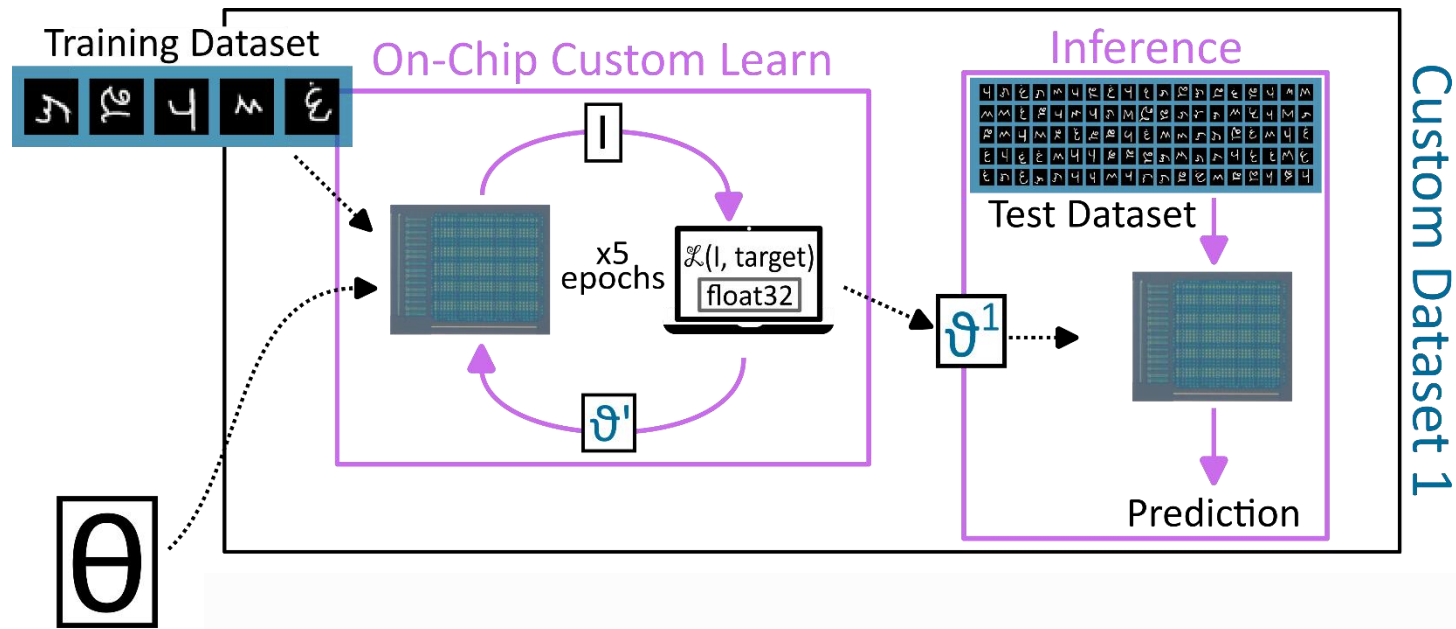
Outline

I. Analog ReRAM for Meta-learning

II. Few Shot on Chip Learning Experiments

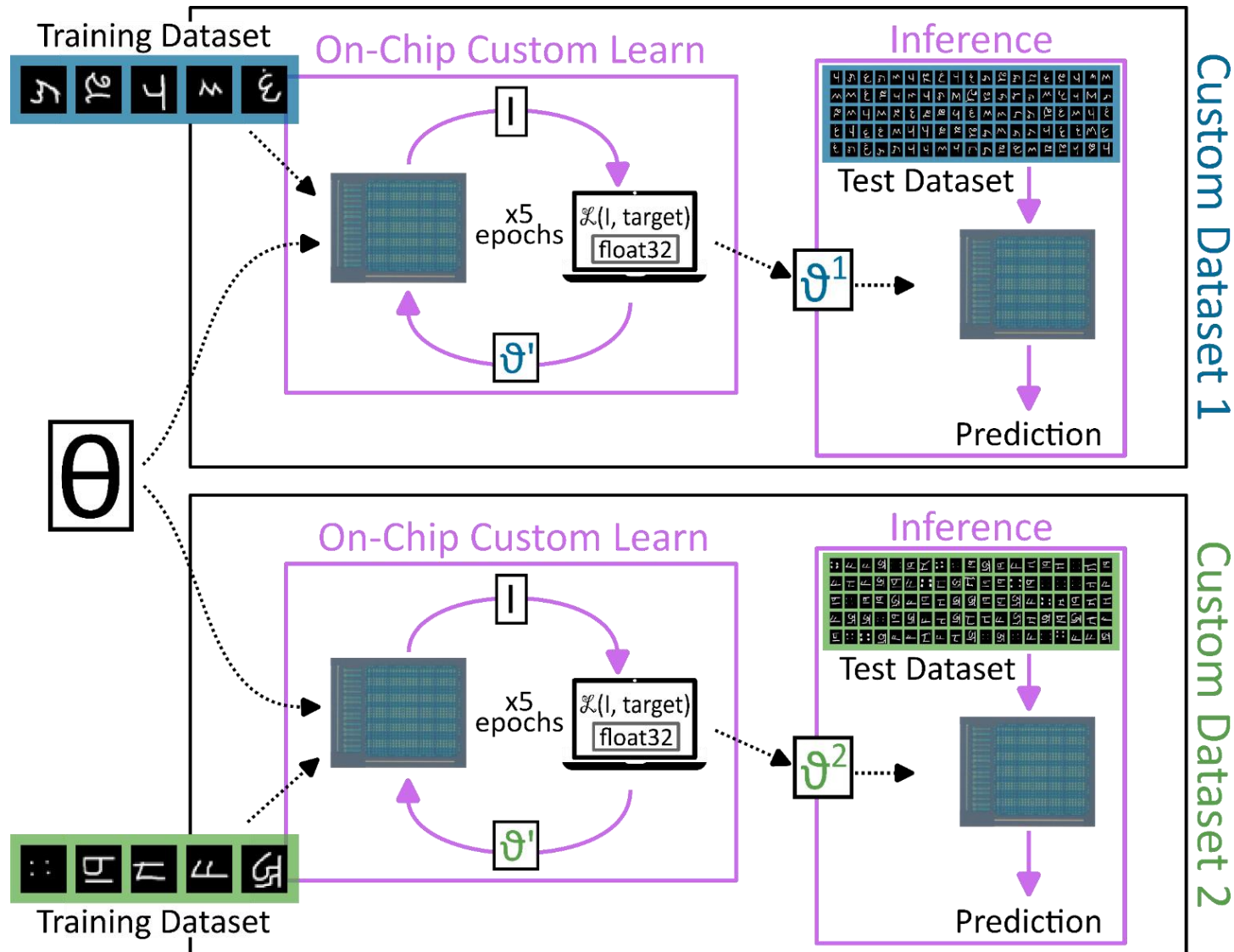
III. Conclusions

On-chip learning experimental set-up



A computer-in-the-loop system configures voltages to read and program ReRAM during both training and inference

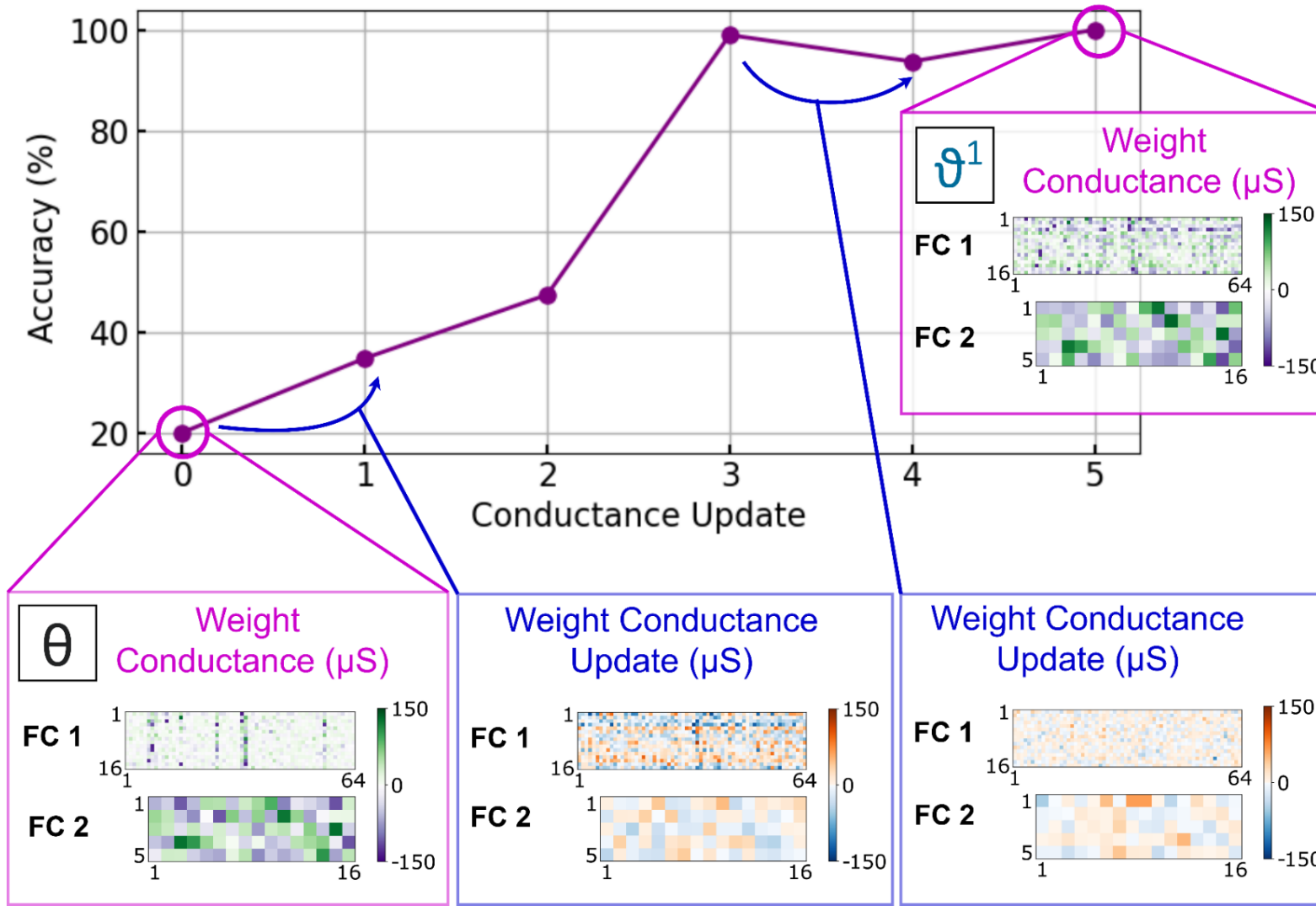
On-chip learning experimental set-up



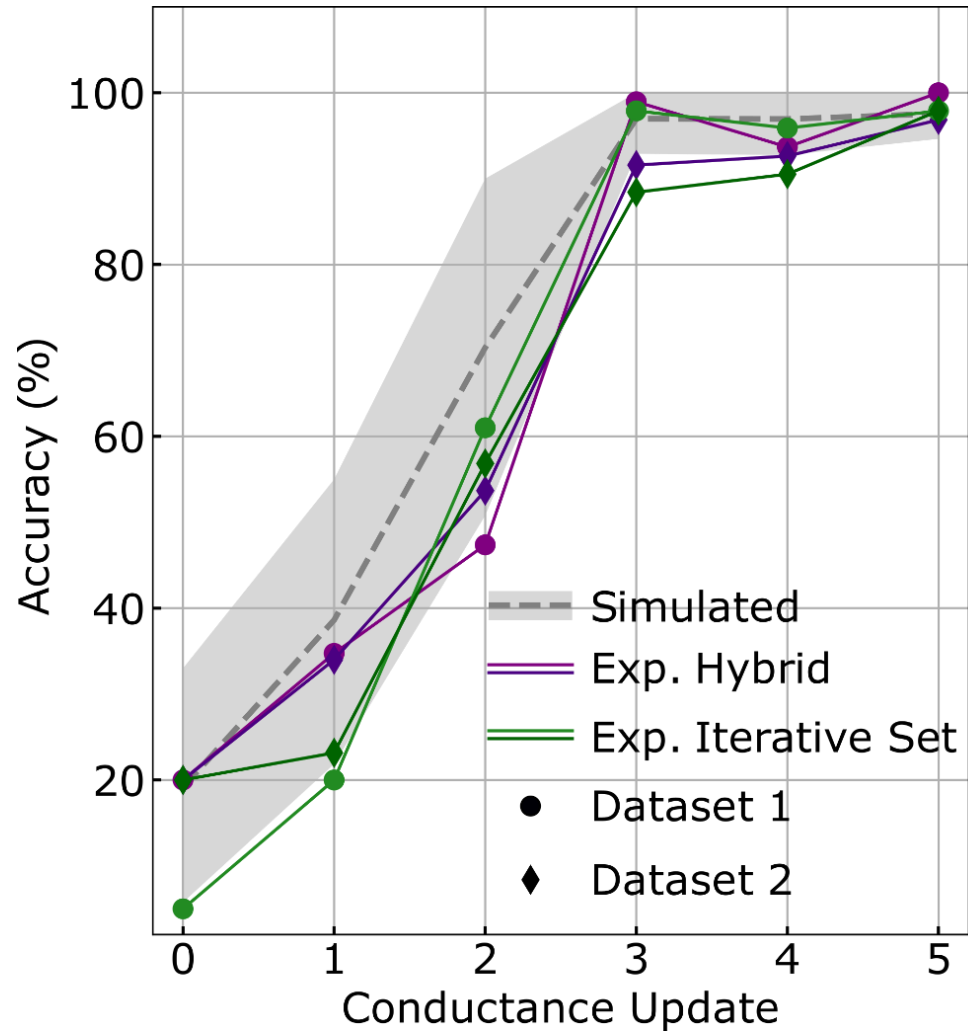
A computer-in-the-loop system configures voltages to read and program ReRAM during both training and inference

The off-chip trained parameters are programmed onto different ReRAM platforms, each of which is adapted to a different set of characters with 5 gradient updates.

Experimental on-chip learning curve

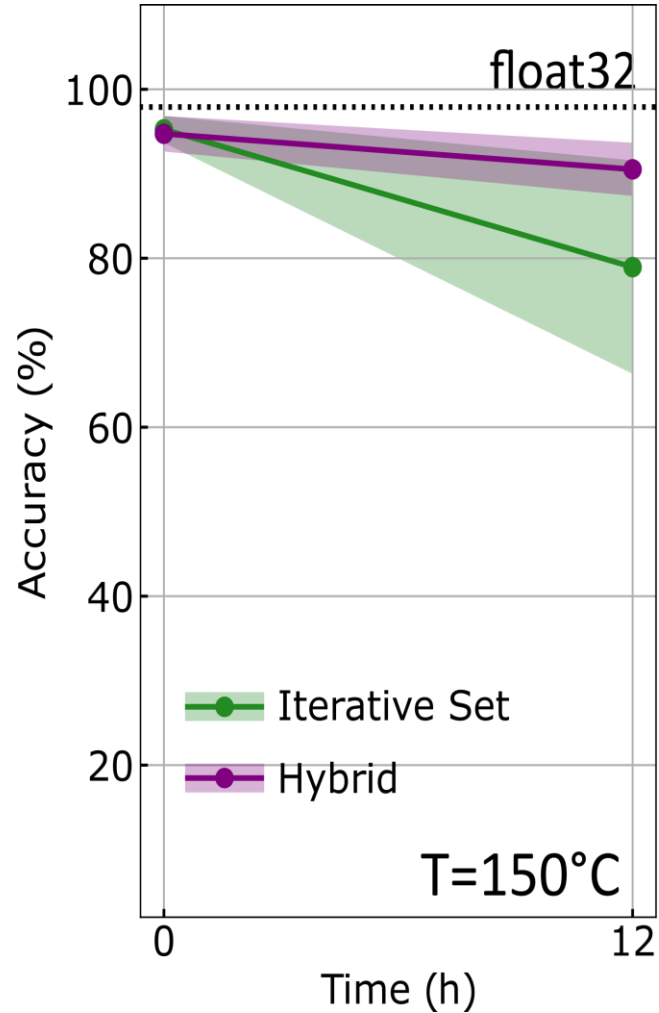


On-chip Learning on Different Datasets



- Repeated on-chip training shows stable learning curves across tasks for both Iterative Set and Hybrid programming methods.
- Results closely align with hardware-aware simulations using 16 custom training datasets.

Accuracy before and after a 12 h bake at 150 °C



Improved retention under *Hybrid* programming

Outline

I. Analog ReRAM for Meta-learning

II. Few Shot on Chip Learning Experiments

III. Conclusions

Conclusion

- Achieved over **97% accuracy** in character recognition using only **five weight updates** on an **in-memory ReRAM platform**.
- **Off-chip training cost** can be **shared across users**, reducing local compute burden during personalization.
- **Retention verified at 150 °C**

This approach enables secure and energy-efficient edge AI with resistive memory.

Thank You!