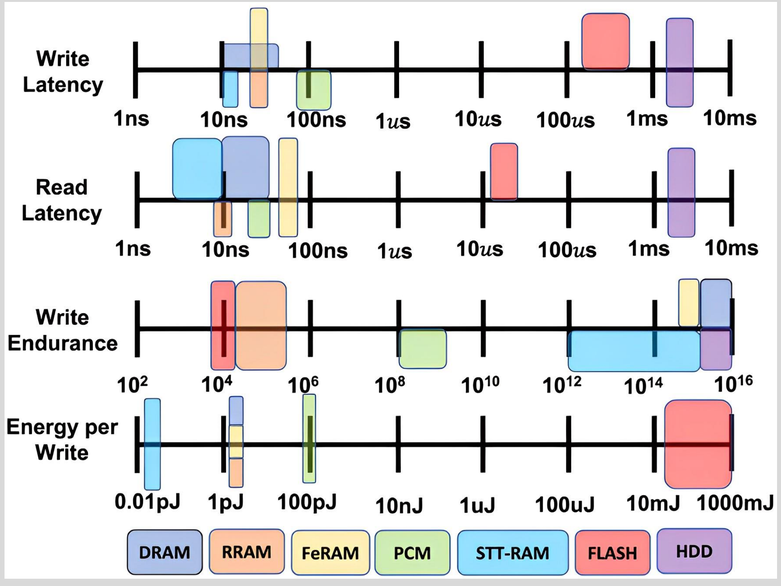
ARTS-NVM: Reconfigurable Architectures and Real-Time Systems Co-Design for Non-Volatile Main Memory

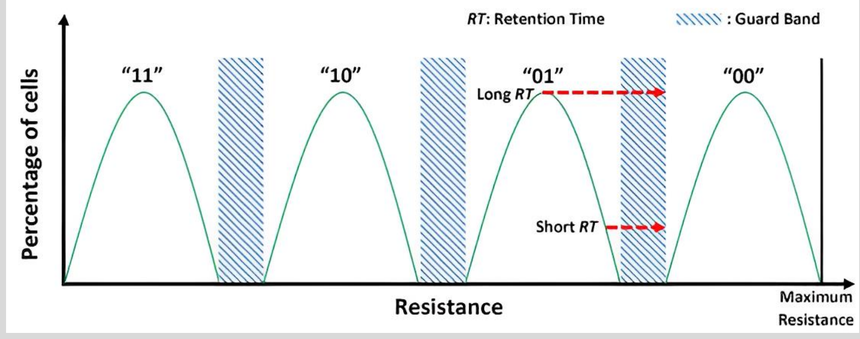
Christian Hakert, Nils Hölscher, Hassan Nassar, Tristan Seidl, Lokesh Siddhu
Lars Bauer, Jian-Jia Chen, Jörg Henkel

Challenges and Chances of NVMs in Embedded Systems

- **Many Types of emerging non-volatile memories (NVMs)**
  - Phase Change memory (**PCM**): Phase (Amorphous/ crystalline) of a chalcogenide glass
  - Resistive RAM (**RRAM**): Resistance
  - Magneto-resistive RAM (**MRAM**): Magnetic field
  - …



[Kargar-DPD-2022]



[Pan-TECS-2017]

Multi-level Cells (**MLCs**): the stored values 'drift' over time and eventually pass the guard band to the next cell value → **Retention time (RT) violation**

- **Main Features of NVMs**
  - Low leakage power, larger capacity (multi-level cell), non-volatility (persistence)
  - Suitable for compute-in-memory architectures

- **Key Limitations**
  - **Endurance**: restricted number of writes per memory cell
  - **Write performance**

Potential Write Modes:

Fast
Medium
Slow

The Trade-off between Latency and Retention of MLC PCM Write Operations

Write Type	Latency (in ns.)	Retention (in seconds)	Current (in μA)	Normalized Energy
3-SETs-Write	550	2.01	42	0.84
4-SETs-Write	700	24.05	37	0.869
5-SETs-Write	850	104.4	35	0.972
6-SETs-Write	1000	991.4	32	0.975
7-SETs-Write	1150	3054.9	30	1

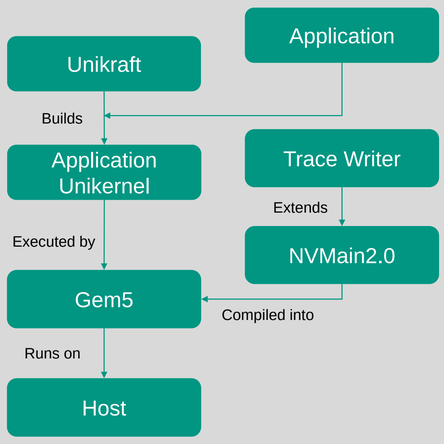
[Zhang-TC-2019]

Trade-off: A lower number of write pulses at an increased current per pulse allows to **reduce the write latency** at the cost of a **reduced retention time** (due to less precise resistance programming)

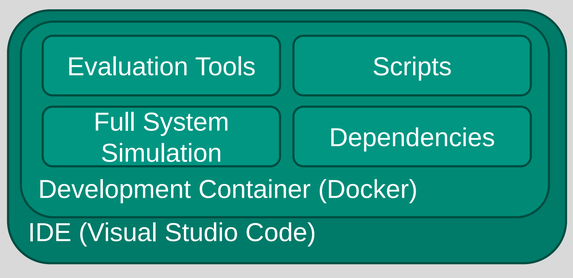
Toolchain for SLC/MLC NVM simulation and evaluation

Toolchain objectives: Easy usage, configurable, comparable results, real data

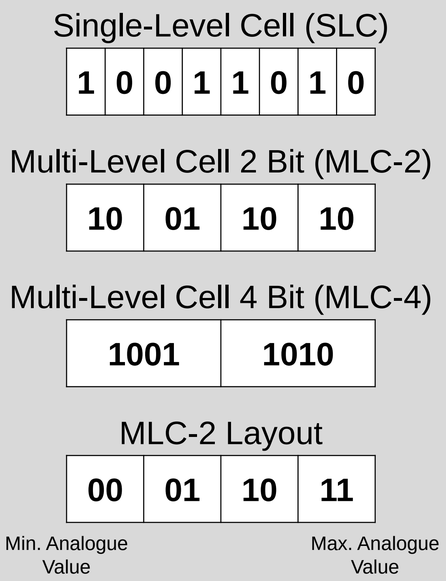
SLC NVM



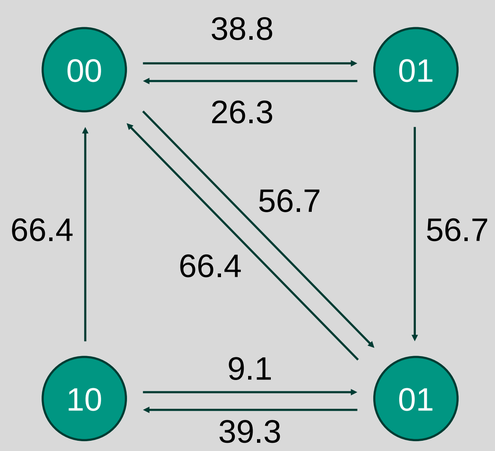
- Record **number of bitflips** based on access information
- **Analysis script** used for evaluation of simulated data
- Environment **already setup** in docker container



MLC NVM

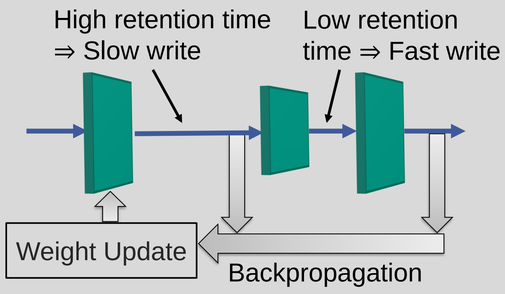


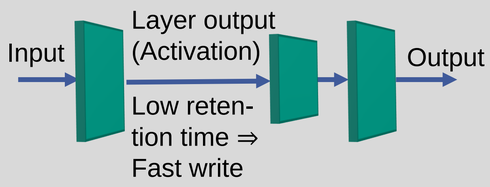
- # Memory accesses **insufficient** for wear indication
- **MLC extension** for toolchain
- **Transition model** to provide MLC wear behavior



Swift-CNN: Using write modes to accelerate Neural Networks

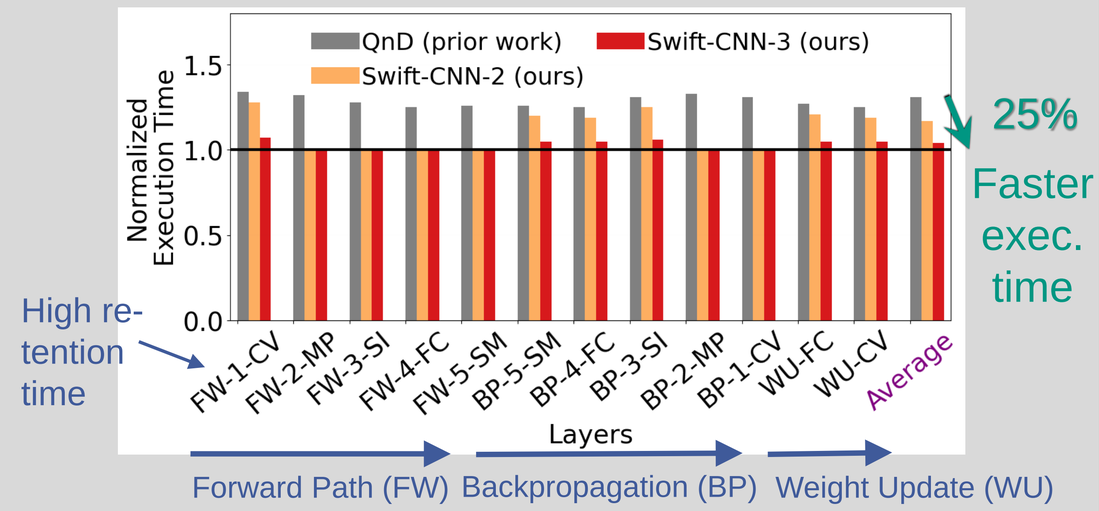
- **Main Idea:** use 'fast writes' for variables with short retention-time requirements, and 'slow writes' for others
  - Use **profiling** to extract application requirements
  - Add a new **'fast write' assembly instruction** to the ISA to use it





Inferences: Layer outputs/ activations can use fast writes

Training: Layers close to the output can use fast writes; Layers further away need to use slow writes, as the activations are utilized to calculate error gradients



- **Evaluation:** Avg. **25% faster application execution time** at no overhead (other than adding the 'fast write' assembly instruction
  - Higher gains for CNN layers closer to the output
  - Training benefits from having more than two write modes

[ESL'23] Lokesh Siddhu, et al. "Swift-CNN: Leveraging PCM Memory's Fast Write Mode to Accelerate CNNs", in IEEE Embedded Systems Letters (accepted)
[CASES'23] Jörg Henkel et al. "Non-Volatile Memories: Challenges and Opportunities for Embedded System Architectures with Focus on Machine Learning Applications" in Int. Conf. on Compilers, Architectures, and Synthesis for Embedded Systems (accepted)

KIT – University of the State of Baden-Württemberg and
National Research Center of the Helmholtz Association

Department of Computer Science,
TU Dortmund University