DATE 26
DESIGN, AUTOMATION AND TEST IN EUROPE
THE EUROPEAN EVENT FOR ELECTRONIC SYSTEM DESIGN & TEST
20 – 22 APRIL 2026
VERONA, ITALY
PALAZZO DELLA GRAN GUARDIA

TrainDeeploy: Hardware-Accelerated Parameter-Efficient Fine-Tuning of Small Transformer Models at the Extreme Edge

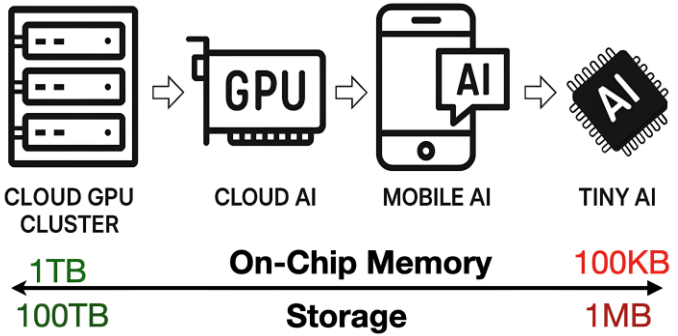
Run Wang*, Victor J. B. Jung*, Philip Wiese*,
Francesco Conti‡, Alessio Burrello†, Luca Benini*‡
***Integrated Systems Laboratory, ETH Zürich**
‡ DEI, University of Bologna
†DAUIN, Politecnico di Torino

ETH zürich

Introduction: On Device Training on SoC

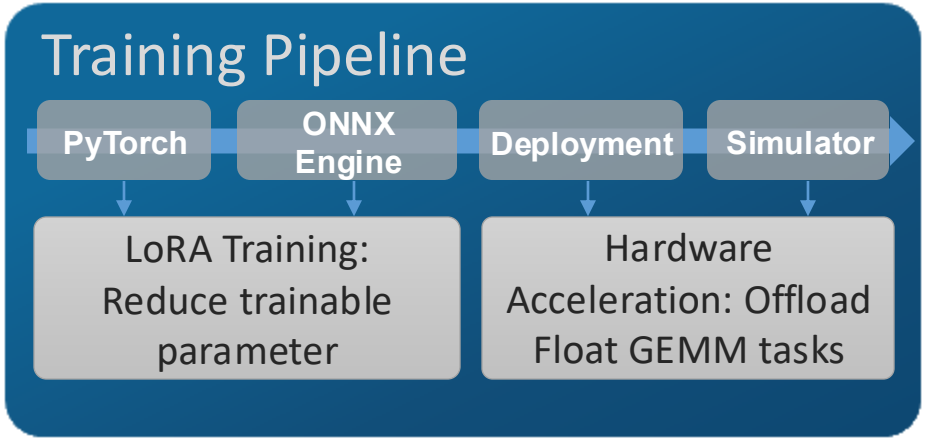
On Device Training Benefits

- Privacy
- Personalization
- Communication

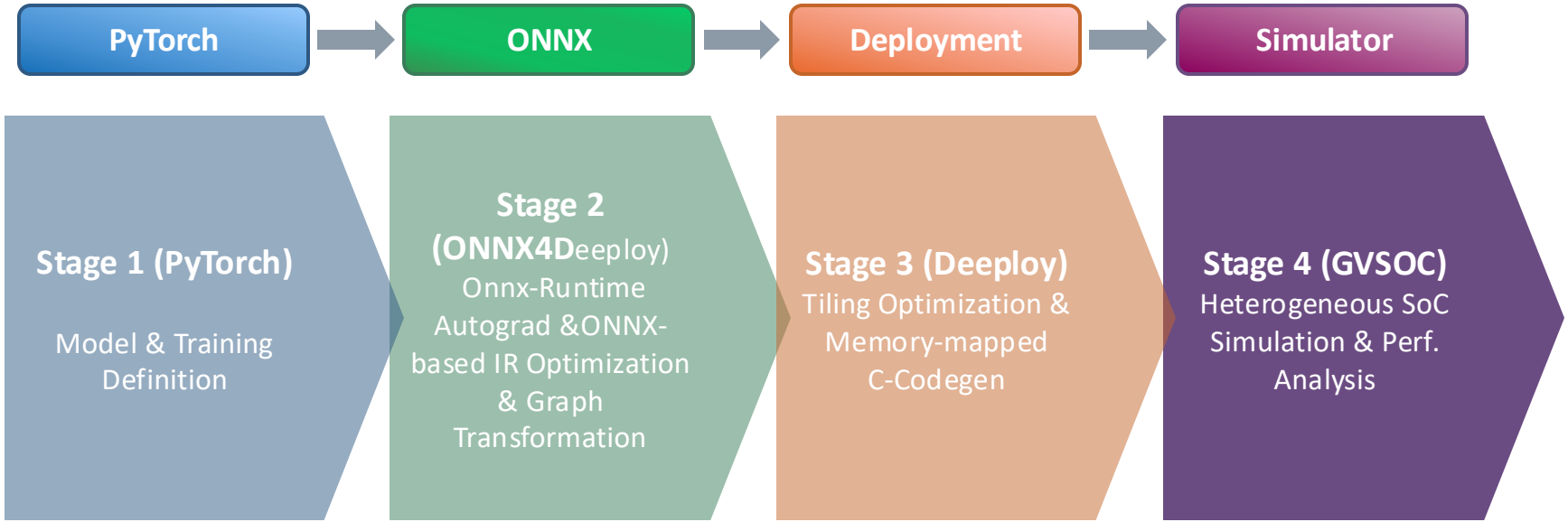


Core Challenges

- **Memory Bottleneck**
  - 10x more memory than inference
  - Activation stored for backpropagation
- **Computational Overhead**
  - Backpropagation and gradient update



Method: Training Deployment Pipeline

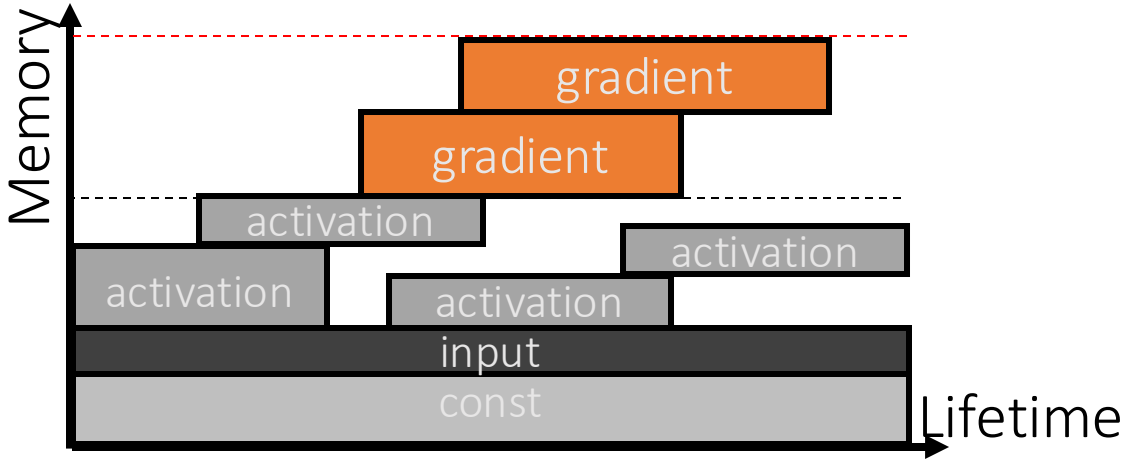


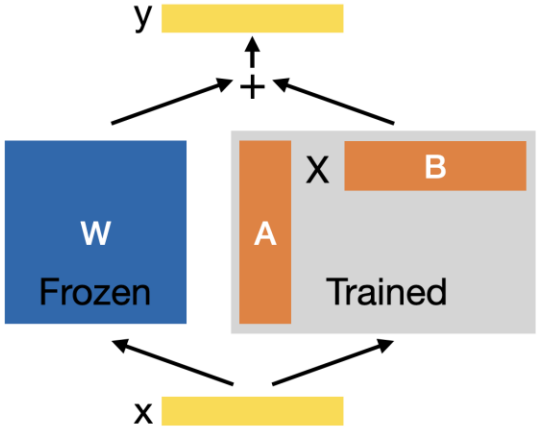
[1] https://github.com/pulp-platform/Onnx4Deeploy
[2] https://github.com/pulp-platform/deeploy
[3] https://github.com/gvsoc/gvsoc

Method: LoRA – Parameter Efficient Training

LoRA Fine-Tuning

- **Parameter Reduction:** Train only low-rank matrices (A, B) instead of weights (W)
- **Memory Savings:** Reduce optimizer states and gradient storage

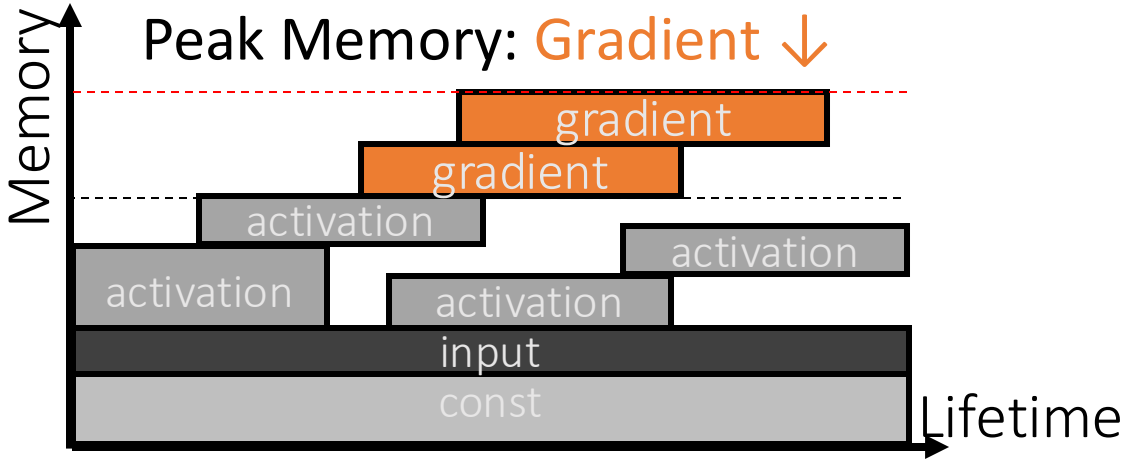


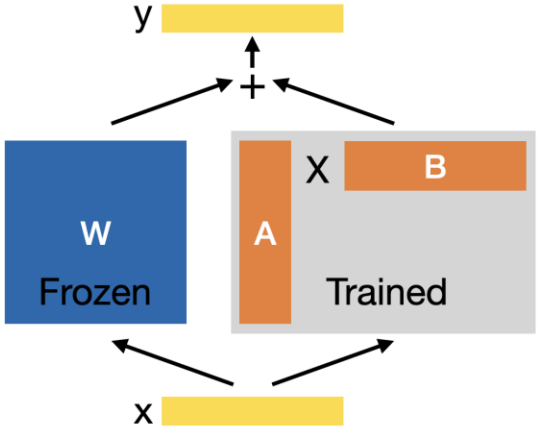


Method: LoRA – Parameter Efficient Training

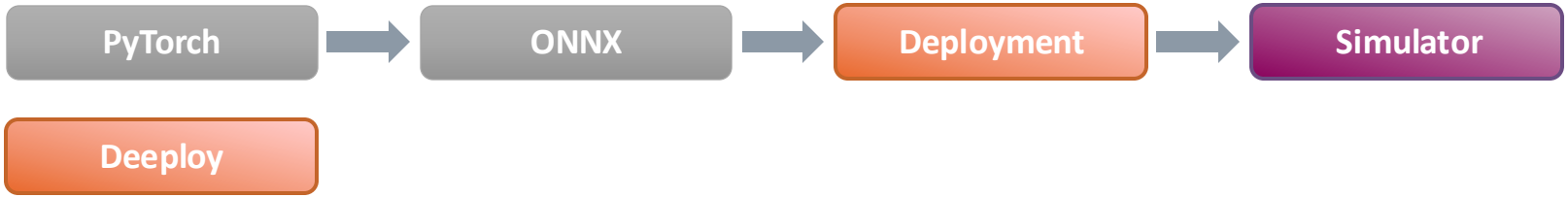
LoRA Fine-Tuning

- **Parameter Reduction:** Train only low-rank matrices (A, B) instead of weights (W)
- **Memory Savings:** Reduce optimizer states and gradient storage





Method: Hardware Acceleration Mapping & Modeling

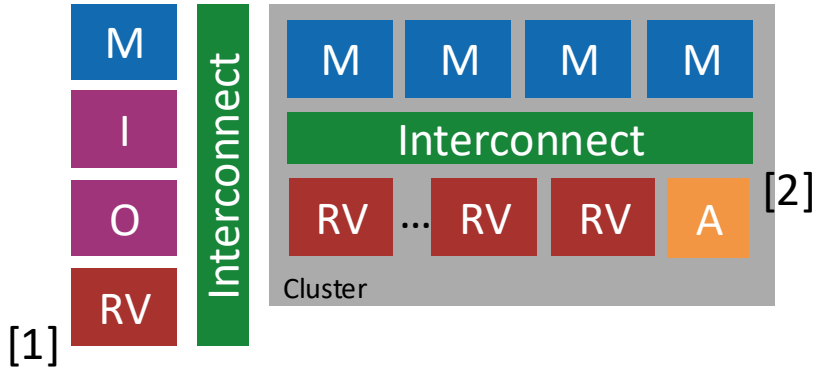


- Kernel mapping for heterogeneous accelerator offloading

- Hierarchical memory management with tiling and static allocation

GVSOC

- Heteorgenous RISC-V SoC[1] with Float Tensor Accelerator[2]



[1]

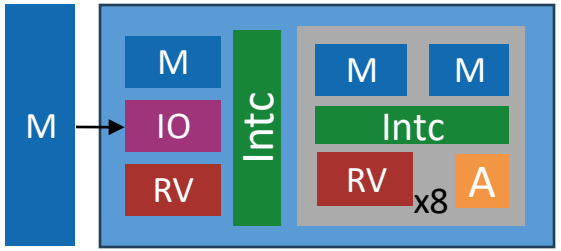
[2]

[1] Prasad et al., "Siracusa: A 16 nm Heterogenous RISC-V SoC for Extended Reality With At-MRAM Neural Engine", JSSC 2024
[2] Tortorella et al. "RedMule: A mixed-precision matrix–matrix operation engine for flexible and energy-efficient on-chip linear algebra and TinyML training acceleration", FGCS 2023

Results

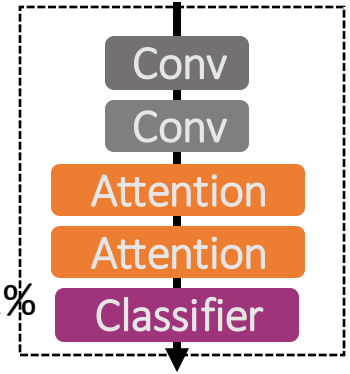
Hardware Target

- TCDM 128KB
- On-Chip SRAM 2MB
- Off-Chip 32MB



Workload

- CCT-2 [1]
  - 0.28M parameters
- CIFAR-10 -> MNIST
- ACC: LoRA vs Full Train +1%



Memory Efficiency via LoRA

- 15x reduction in trainable parameters
- 23% decrease in peak memory usage
- 1.6x reduction in off-chip memory transfer

Hardware Accelerated Performance

- 3.5x speedup with accelerator
- 11 gradient updates/s @ 360MHz
- 13.4 FLOPs/cycle peak performance

[1] A. Hassani *et al.*, "Compact Convolutional Transformers for Lightweight Image Classification," CVPR Workshops, 2021.

TrainDeeploy

Welcome to the poster for more details and demos!

RISC-V Europe
Submit Demo 2026

IJCNN TinyML 2026,
(Extended Work on CNN
and on-board)

Onnx4Deeploy

https://github.com/pulp-platform/Onnx4Deeploy

Onnx4Deeploy

Deeploy

https://github.com/pulp-platform/deeploy

Deeploy