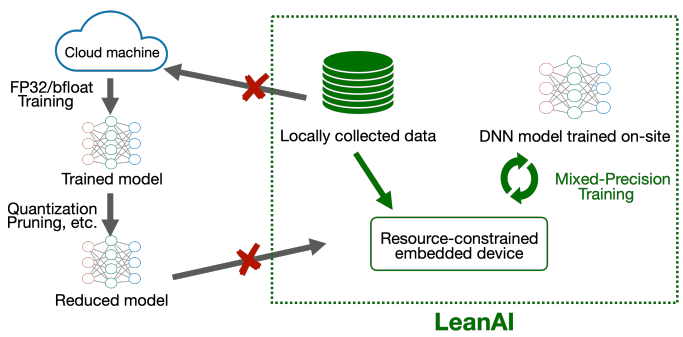
LeanAI: Dynamic Precision Training on the Edge

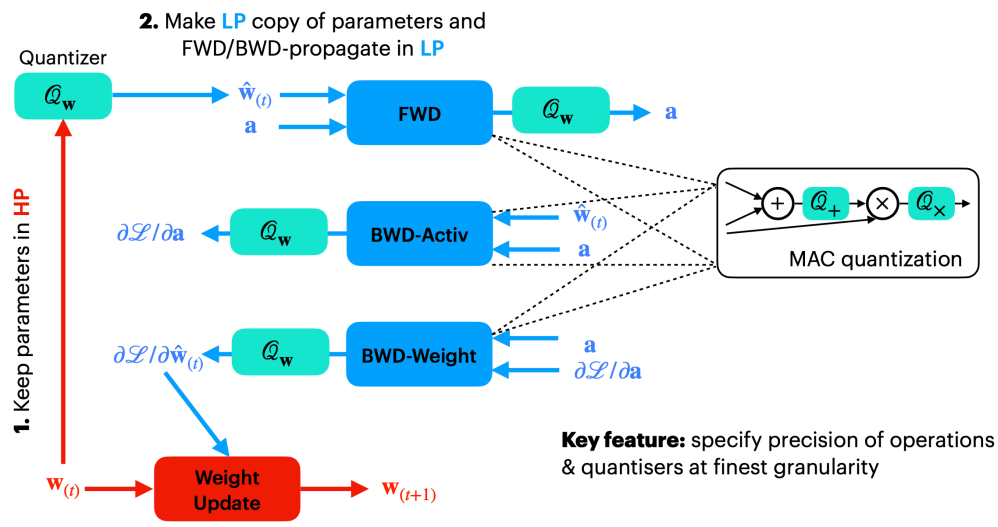
IRISA, TARAN
LS2N, OGRE
LIP, OKHAM

Context and objectives



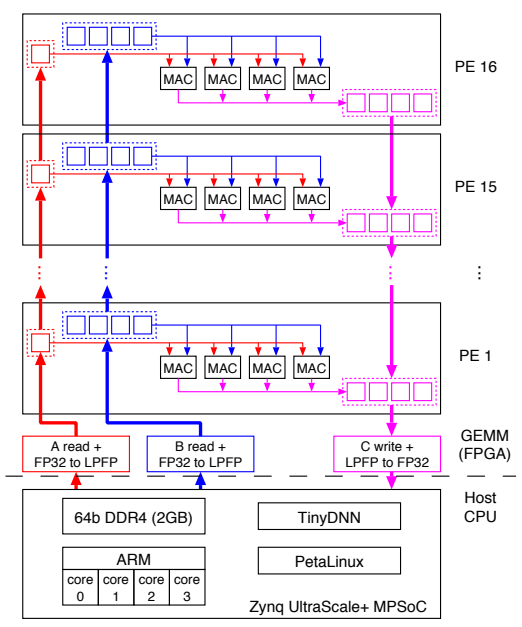
- Need for **learning acceleration** mechanisms in both ***cloud*** (for large-scale models) and ***on-site*** settings (e.g. autonomous driving, privacy).
- Working on both arithmetic and algorithmic levels
- Design of dedicated HW operators

MPTorch: mixed-precision arithmetic simulator



Key feature: specify precision of operations & quantisers at finest granularity

Archimedes-MPO and GEMM kernels



Simulation framework

- extends TinyDNN C++ framework

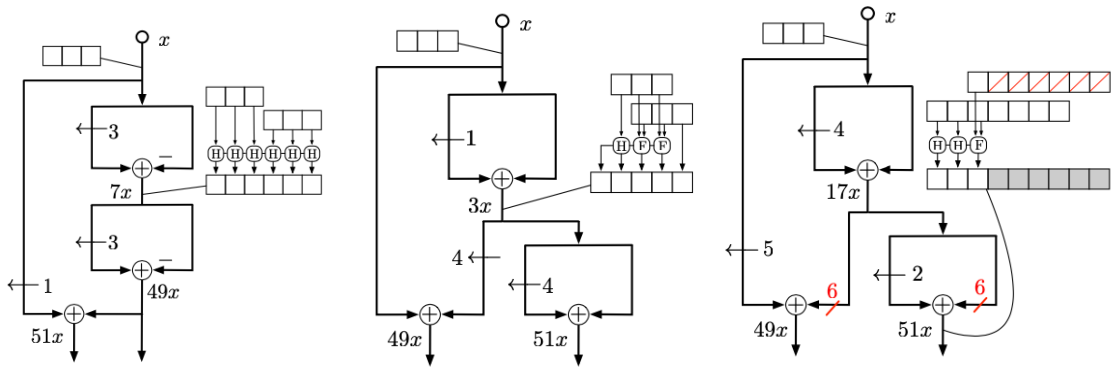
GEMM kernel on FPGA

- custom precision and operators
- parametrizable architecture
- Xilinx ZCU104 board

GEMM kernel on GPU

- Bit-accurate with FPGA
- Convenient deployment & testing

Basic brick: multiplication by multiple constants



Classic MCM

Metric: #additions

Realistic: 24 one-bit adders

Proposed MCM

Metric: #one-bit adders

Realistic: 9 one-bit adders

Truncated MCM

Metric: #one-bit adders & error

Realistic: 4 one-bit adders

Enables fully-parallel unrolled inference at high throughput and low power

Mixed-precision inference for hardware neural network controllers

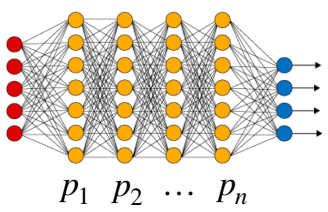
Neural Network Controllers

- Approximate control law and replace costly solvers
- Up to 100 layers
- Require ultra-fast HW
- High safety requirements (error, controllability)
- Formally verified but no guarantee on finite precision

Example: Automated Cruise Control

Mixed-precision assignment

- Use ILP to assign precisions under a priori rounding error constraint
- Can use truncated MCM on each dense layer



Hardware-friendly quantization-aware training

- Fine-tune an FP32 model for small coefficients
- Key idea: select coefficients that have pre-defined low adder-cost for an MCM fully parallel implementation
- Result: can retrain for cost-1 coefficients while staying safe

Project status

Hardware accelerator in active development

- Generic MAC units in fixed/floating-point
- Fully-parallel fixed-point basic bricks for GEMM

Quantization for inference

- ILP-based for small networks
- Heuristic search enabled by MPTorch

Training algorithms

- Trust-region based algorithms (WIP)
- Attention-layer retraining for NLP (collaboration with CIFRE by Valeuriad company)

Project publications: