ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

LAB09: End-to-end deployment

Davide Nadalini – d.nadalini@unibo.it
Luca Bompani – luca.bompani5@unibo.it
Lorenzo Lamberti – lorenzo.lamberti@unibo.it
Francesco Conti – f.conti@unibo.it

Objective of the Class

Lesson: Automated deployment on a PULP microcontroller

Programming Language: C

Lab duration: 3h

Assignment:

- Time for delivery: 2 weeks

Deadline:

Jan 2nd 2026

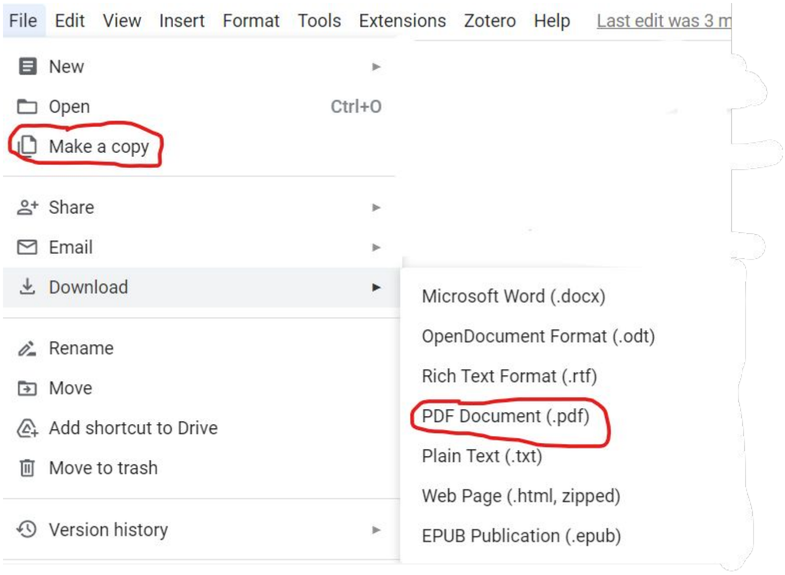
The class is meant to be interactive: coding together, on your own, and do not be afraid to ask questions!

**Hands-on Session With GAP-SDK and GAP9!
Look to the setup guide on Virtuale (carefully follow the steps!!)**

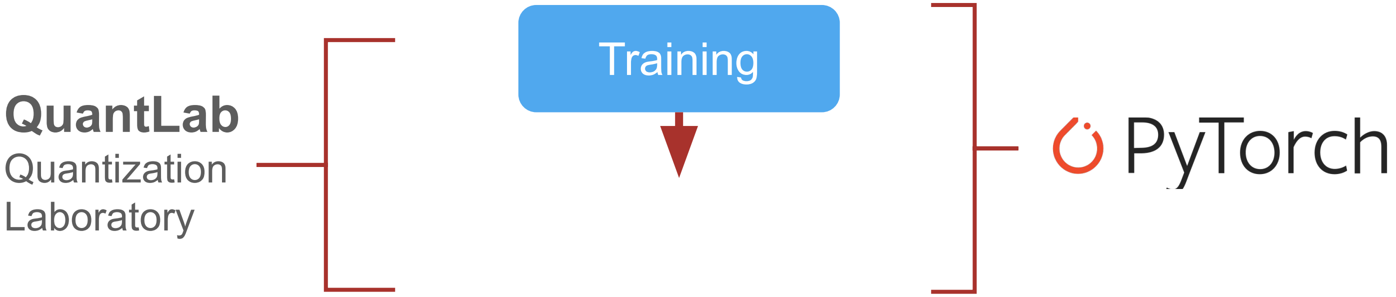
How to deliver the assignment

You will deliver ONLY the GDOC assignment, no code

- Copy the google doc to your drive, so that you can modify it. (File -> make a copy)
- Fill the tasks on this google doc.
- Export to pdf format.
- Rename the file to: LAB<number_of_the_lesson>_APAI_<your_name>.pdf
- Use Virtuale platform to load ONLY your .pdf file



Neural network deployment flow



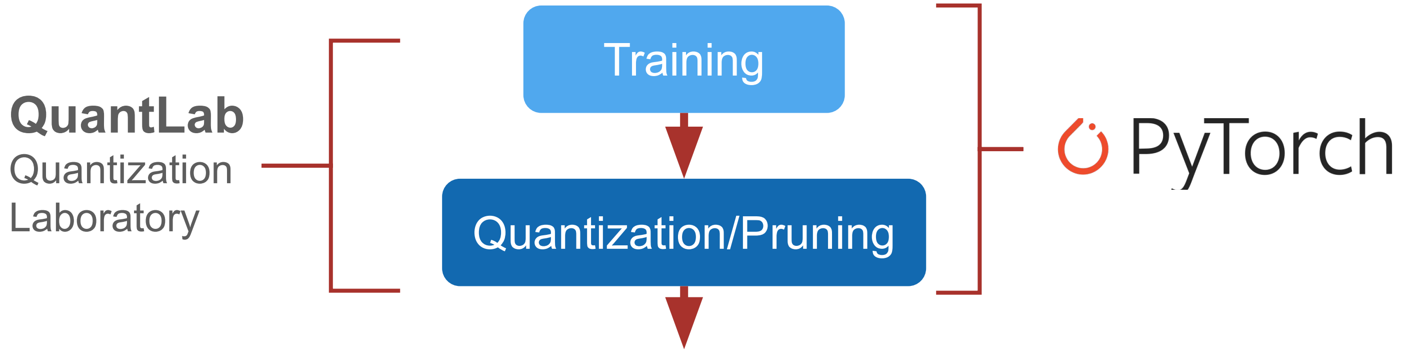
QuantLab
Quantization
Laboratory

Training

PyTorch

LAB 2: DNN definition and training

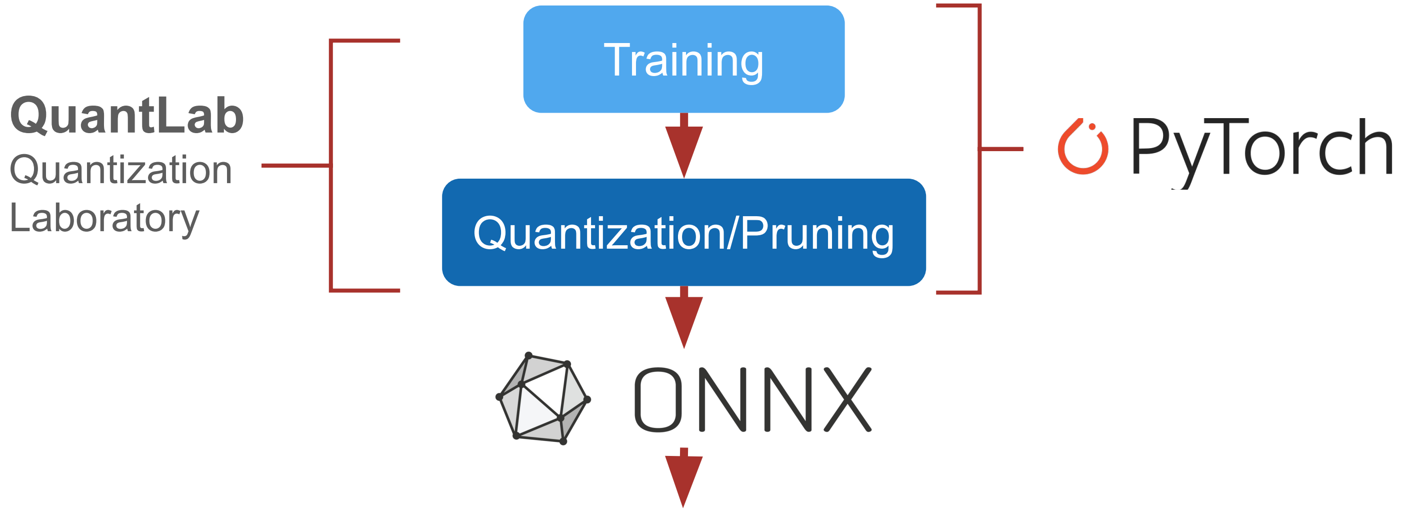
Neural network deployment flow



LAB 2: DNN definition and training

LAB 3: DNN shrinking and quantization

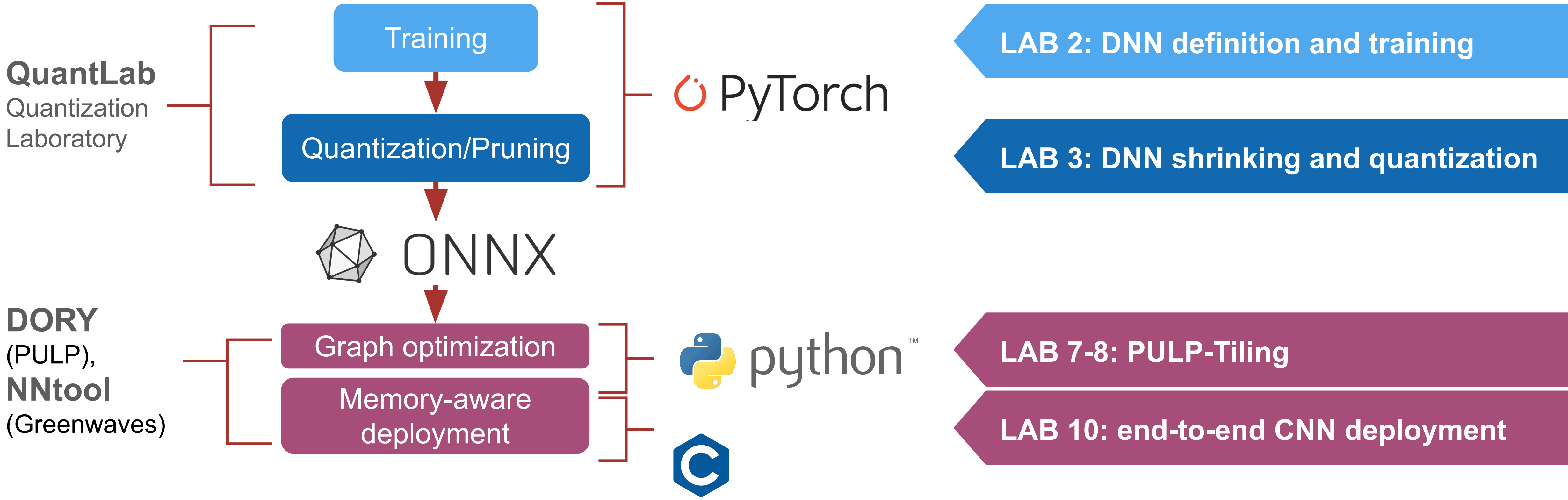
Neural network deployment flow



LAB 2: DNN definition and training

LAB 3: DNN shrinking and quantization

Neural network deployment flow



QuantLab Quantization Laboratory

Training

Quantization/Pruning

PyTorch

ONNX

DORY (PULP), **NNtool** (Greenwaves)

Graph optimization

Memory-aware deployment

python™

C

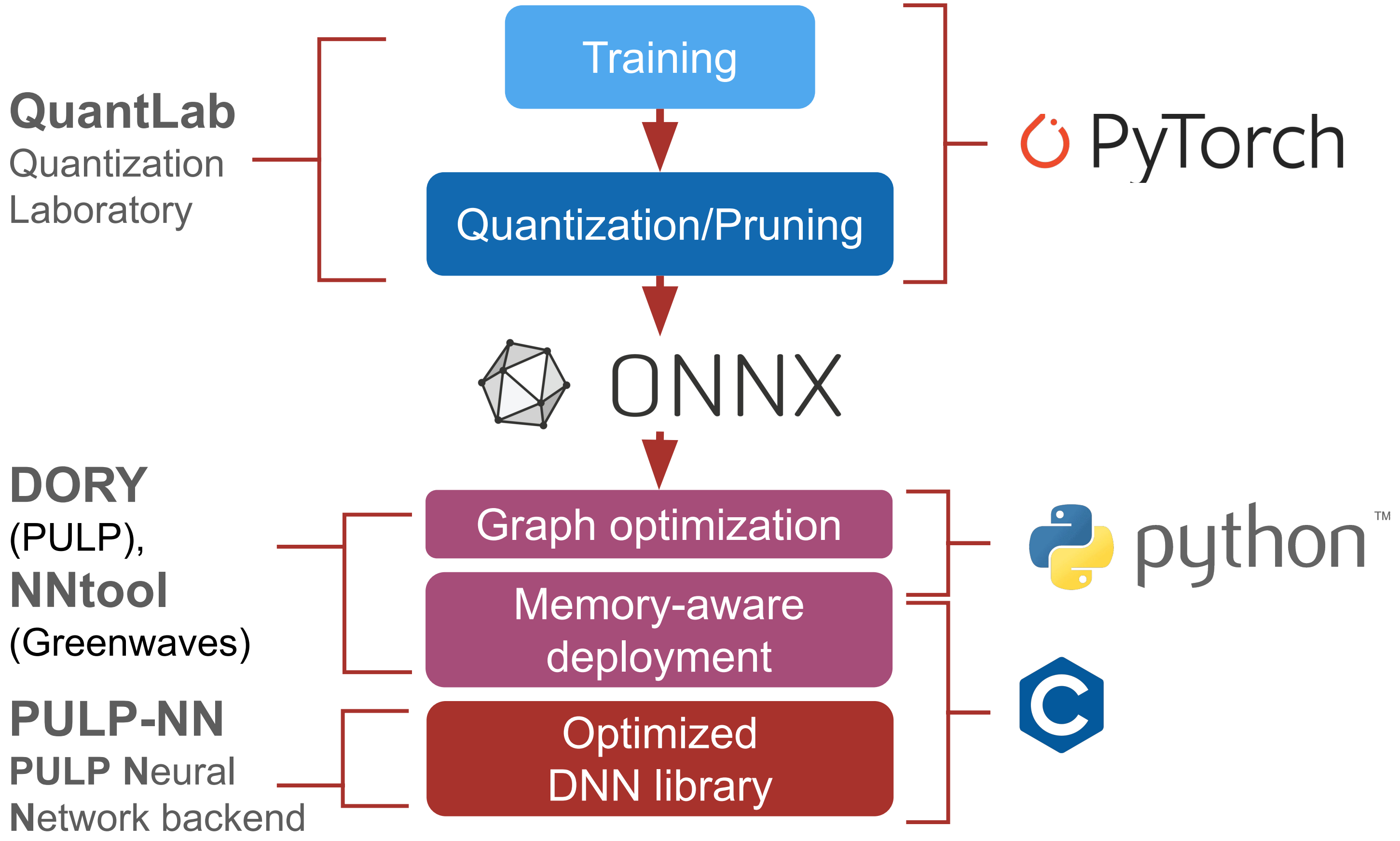
LAB 2: DNN definition and training

LAB 3: DNN shrinking and quantization

LAB 7-8: PULP-Tiling

LAB 10: end-to-end CNN deployment

Neural network deployment flow



QuantLab
Quantization
Laboratory

Training

Quantization/Pruning

PyTorch

ONNX

DORY
(PULP),
NNtool
(Greenwaves)

PULP-NN
PULP Neural
Network backend

Graph optimization

Memory-aware
deployment

Optimized
DNN library

python™

C

LAB 2: DNN definition and training

LAB 3: DNN shrinking and quantization

LAB 7-8: PULP-Tiling

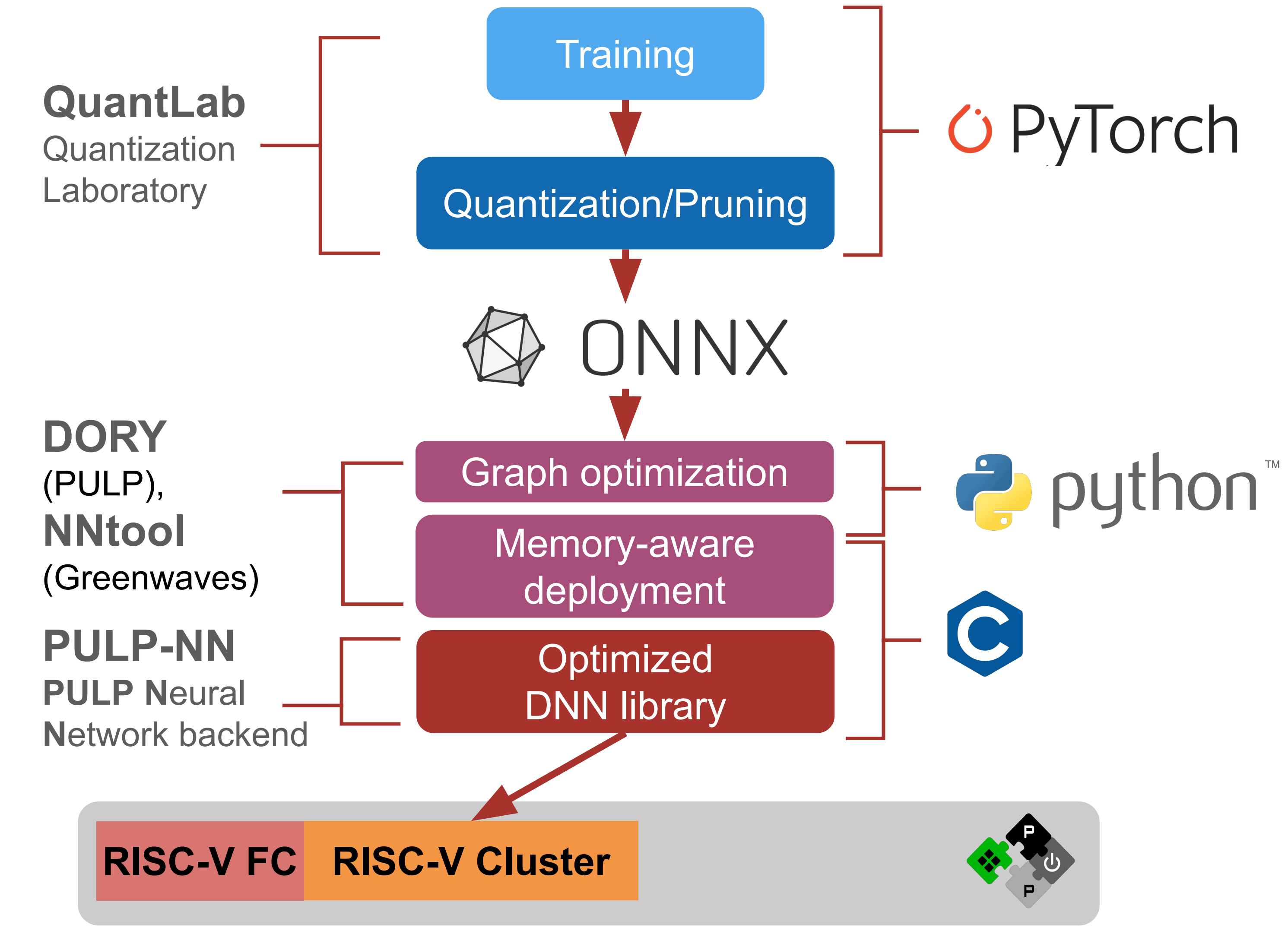
LAB 10: end-to-end CNN deployment

LAB 4: PULP-NN

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

ETH zürich

Neural network deployment flow



QuantLab
Quantization Laboratory

Training

Quantization/Pruning

PyTorch

ONNX

DORY (PULP), NNtool (Greenwaves)

Graph optimization

Memory-aware deployment

python™

PULP-NN
PULP Neural Network backend

Optimized DNN library

RISC-V FC

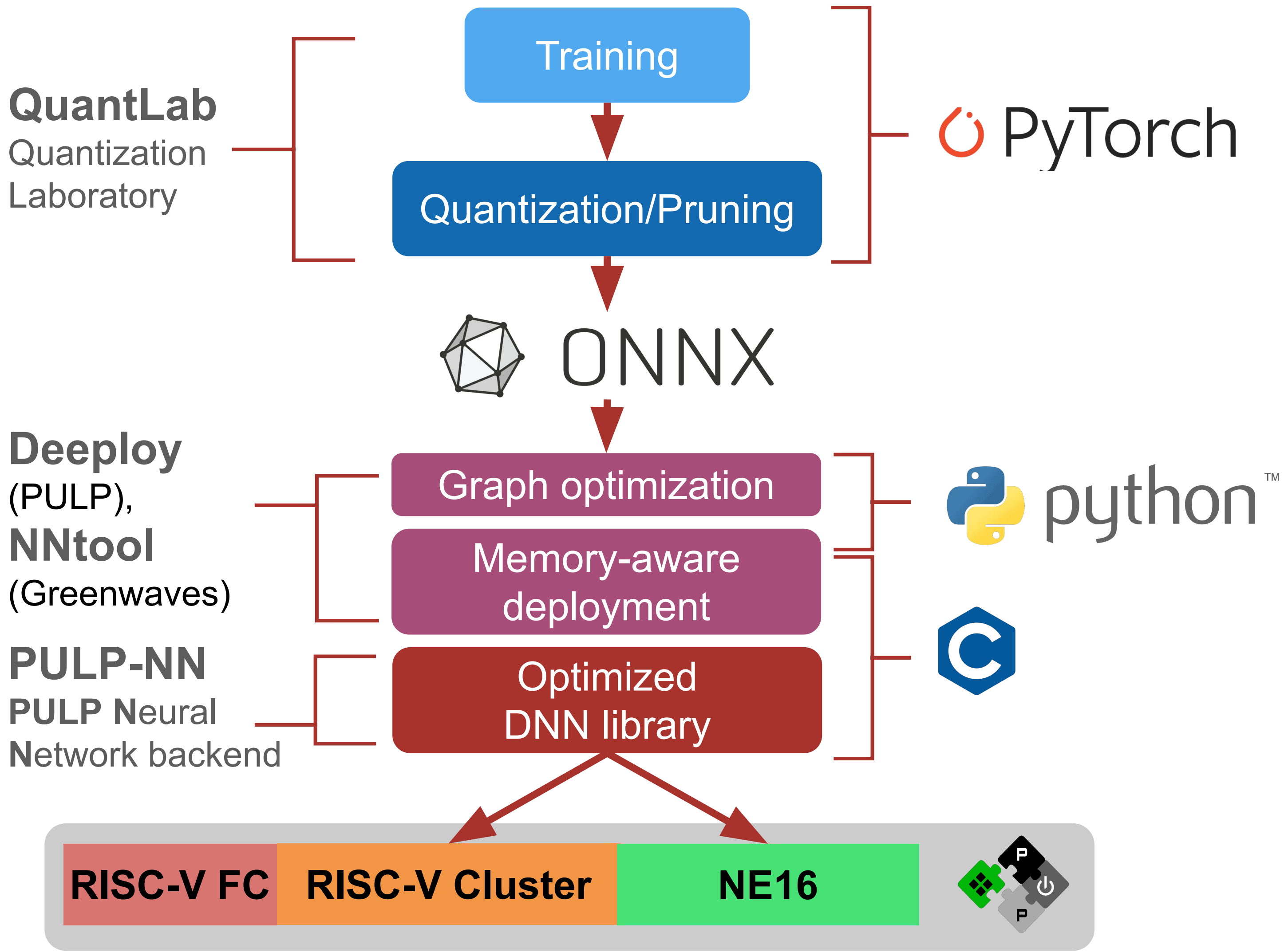
RISC-V Cluster

- LAB 2: DNN definition and training
- LAB 3: DNN shrinking and quantization
- LAB 7-8: PULP-Tiling
- LAB 10: end-to-end CNN deployment
- LAB 4: PULP-NN
- LAB 1: PULP embedded programming

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

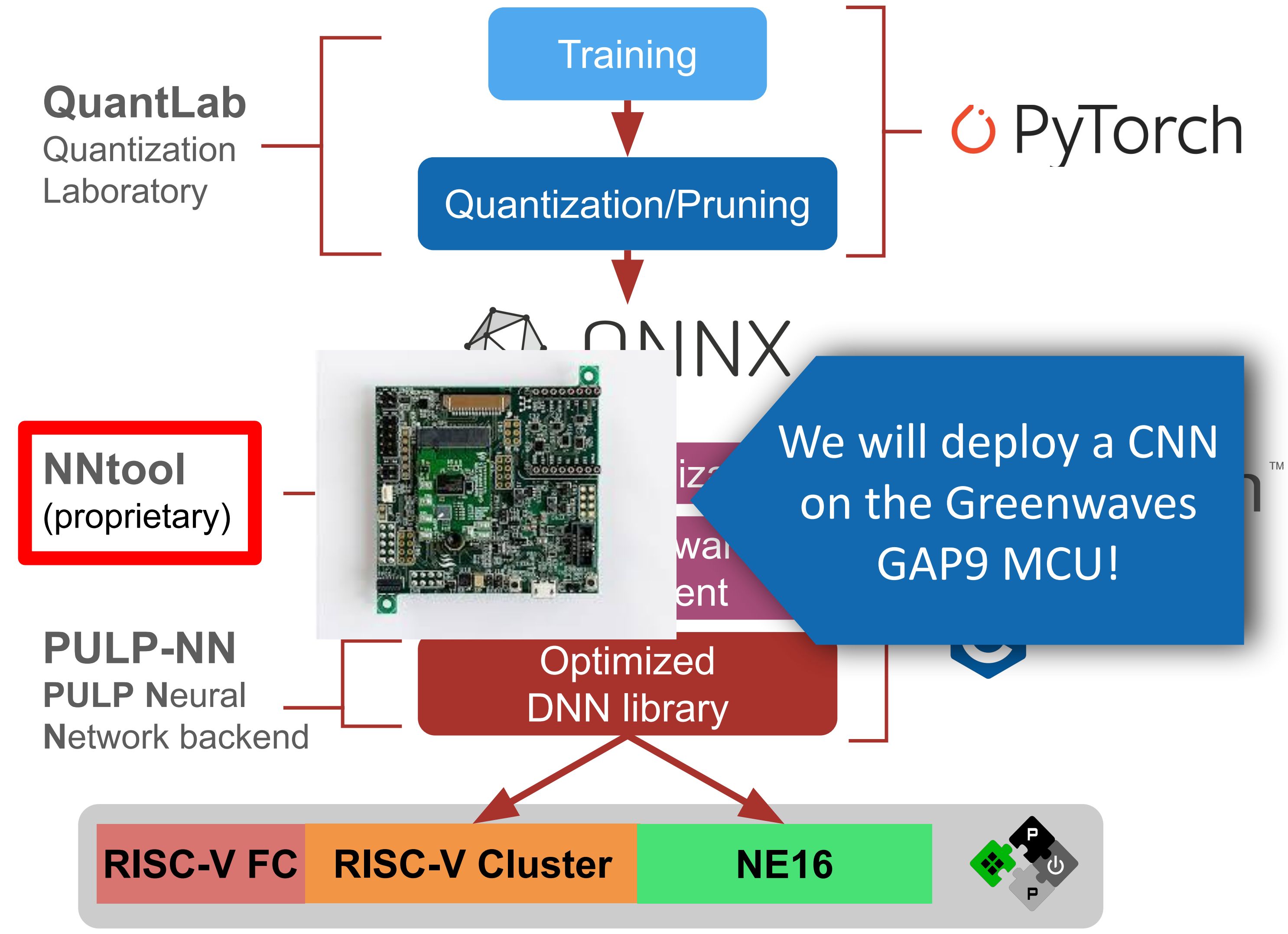
ETH zürich

Neural network deployment flow



- LAB 2: DNN definition and training
- LAB 3: DNN shrinking and quantization
- LAB 7-8: PULP-Tiling
- LAB 10: end-to-end CNN deployment
- LAB 4: PULP-NN
- LAB 1: PULP embedded programming
- LAB 9: NE16

Neural network deployment flow (PULP)



We will deploy a CNN on the Greenwaves GAP9 MCU!

LAB 2: DNN definition and training

LAB 3: DNN shrinking and quantization

TODAY (part 1)

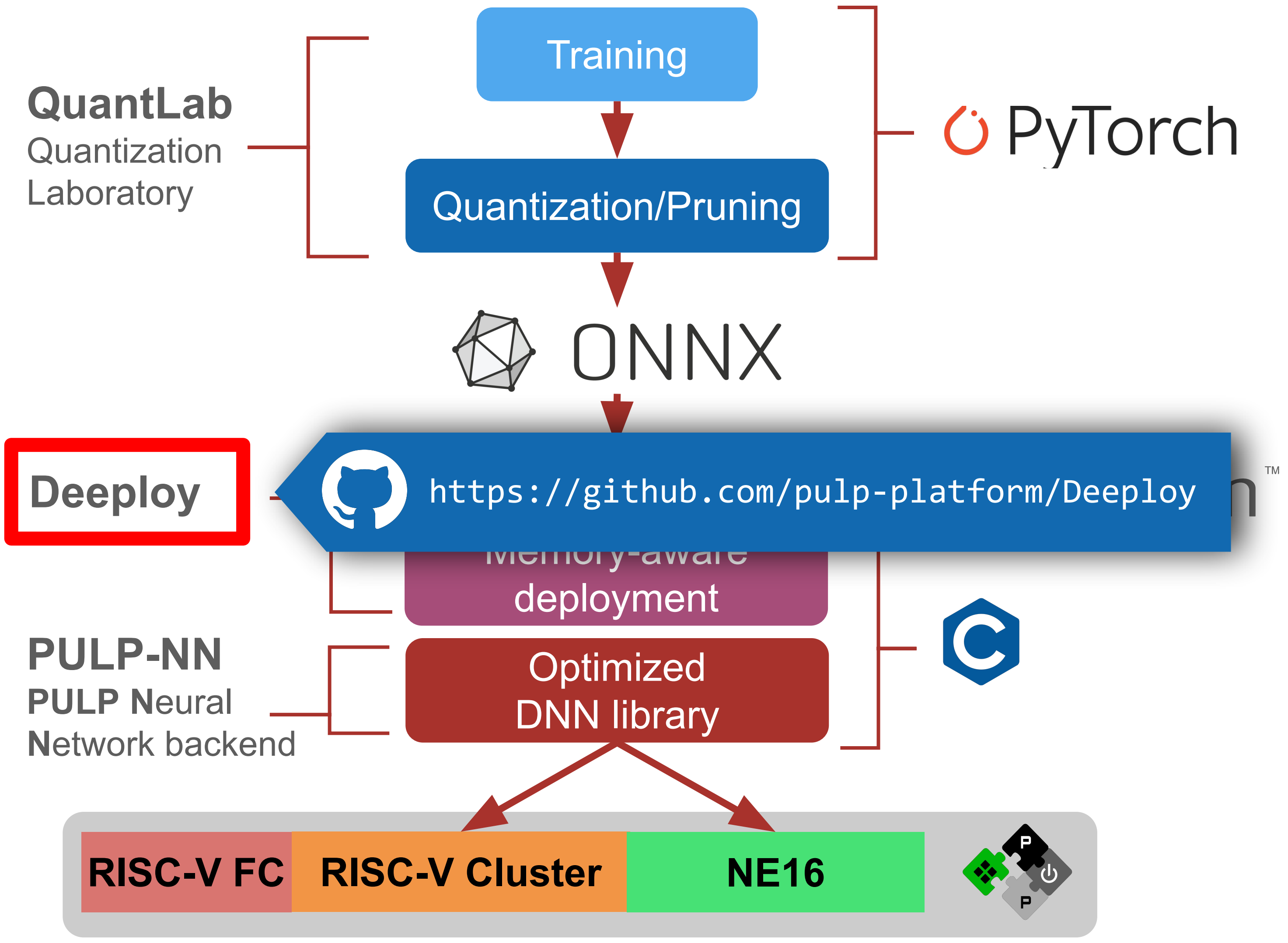
LAB 10: end-to-end CNN deployment

LAB 4: PULP-NN

LAB 1: PULP embedded programming

LAB 9: NE16

Neural network deployment flow (PULP)



LAB 2: DNN definition and training

LAB 3: DNN shrinking and quantization

TODAY (part 2)

LAB 10: end-to-end CNN deployment

LAB 4: PULP-NN

LAB 1: PULP embedded programming

LAB 9: NE16

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

BACKUP

DEI – Università di Bologna