UNIVERSITÄT BONN

DEEP

ErUM-Data DEEP:
Neural Networks on Hardware

Patrick Schwäbig

IHL Kickoff Meeting, Siegen
17.7.2026

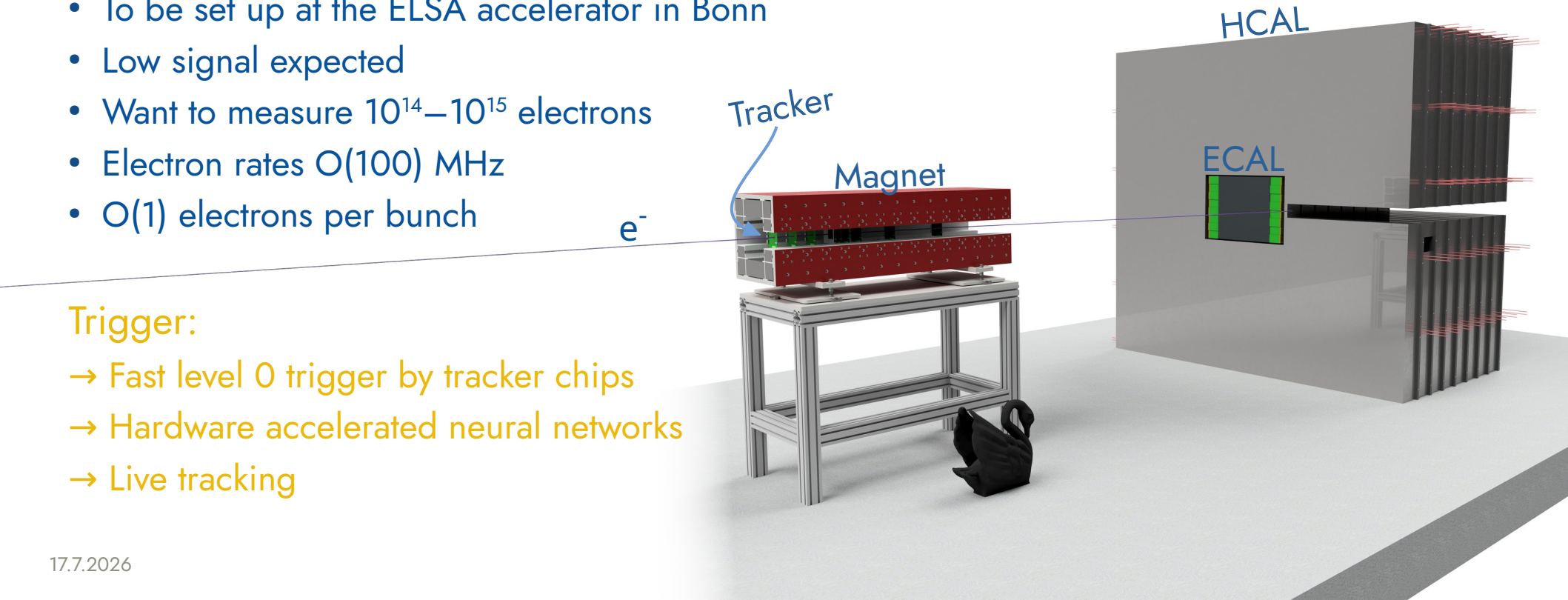
Funded by

Federal Ministry
of Research, Technology
and Space

UNIVERSITÄT BONN

WHERE I AM COMING FROM: THE LOHENGRIN EXPERIMENT

- Dark photon search
- To be set up at the ELSA accelerator in Bonn
- Low signal expected
- Want to measure 10^{14}–10^{15} electrons
- Electron rates O(100) MHz
- O(1) electrons per bunch



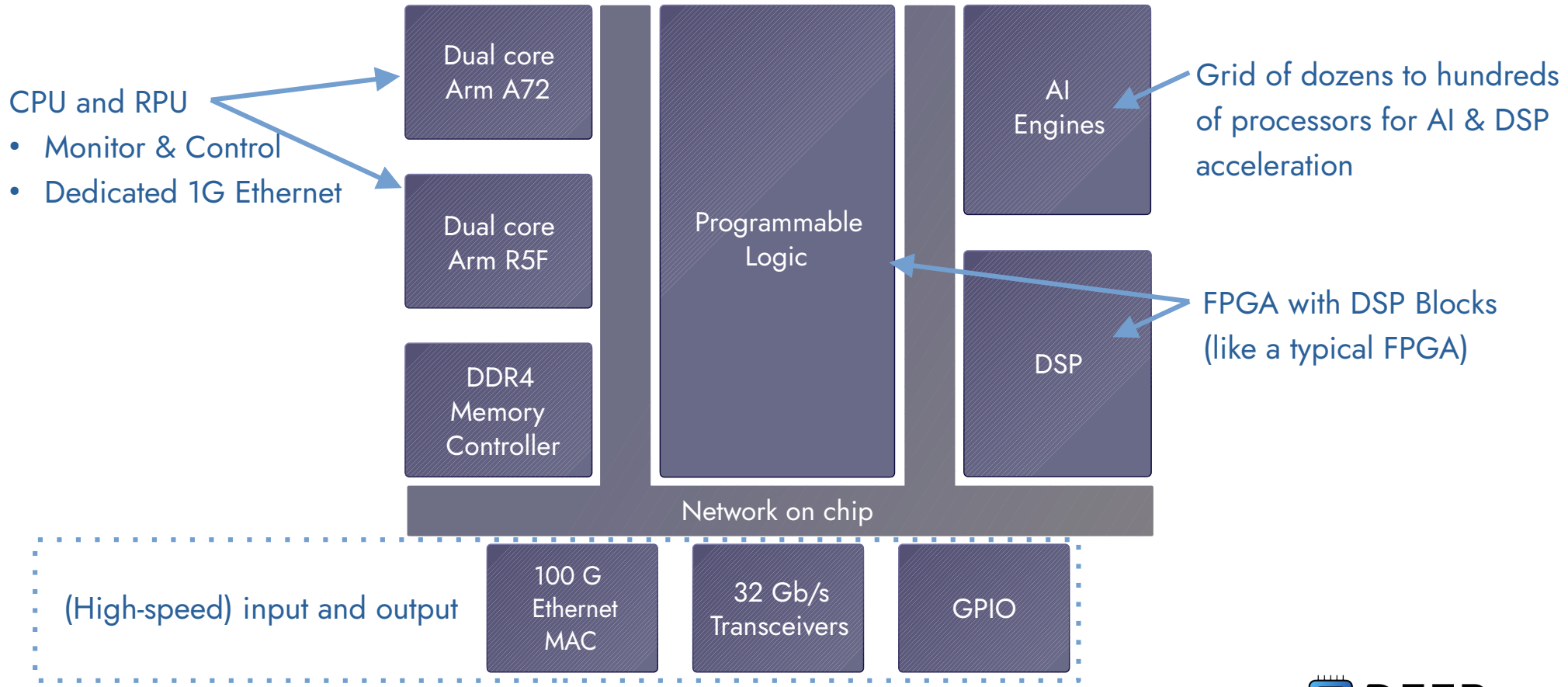
Trigger:

→ Fast level 0 trigger by tracker chips

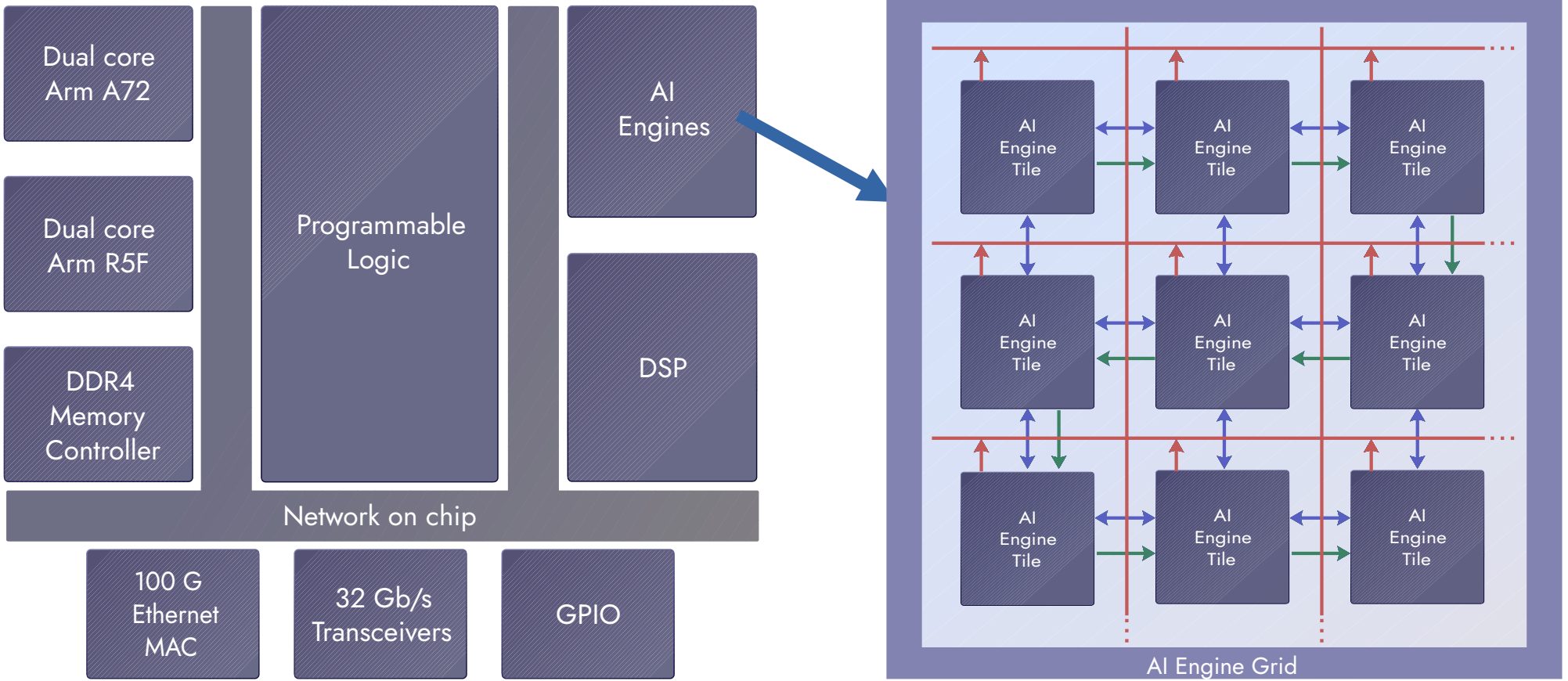
→ Hardware accelerated neural networks

→ Live tracking

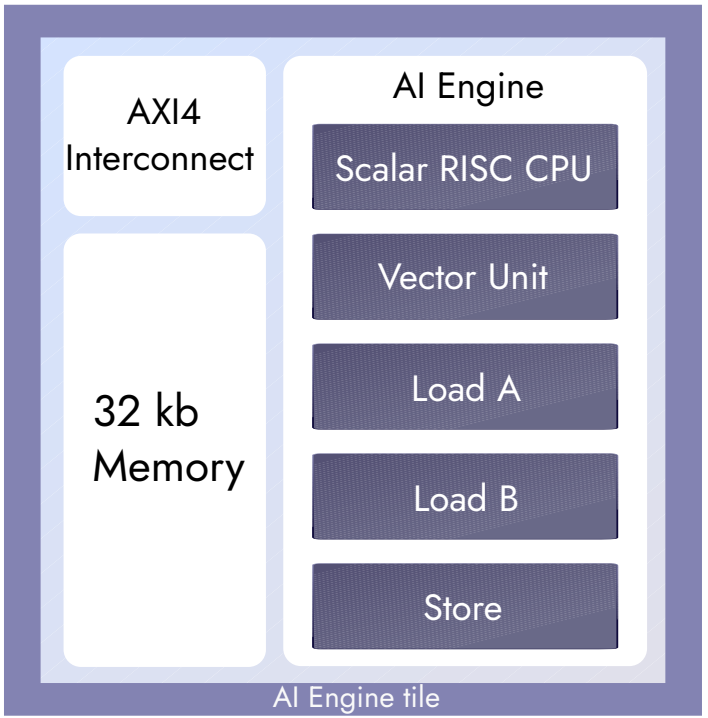
AMD VERSAL ARCHITECTURE



AMD VERSAL ARCHITECTURE



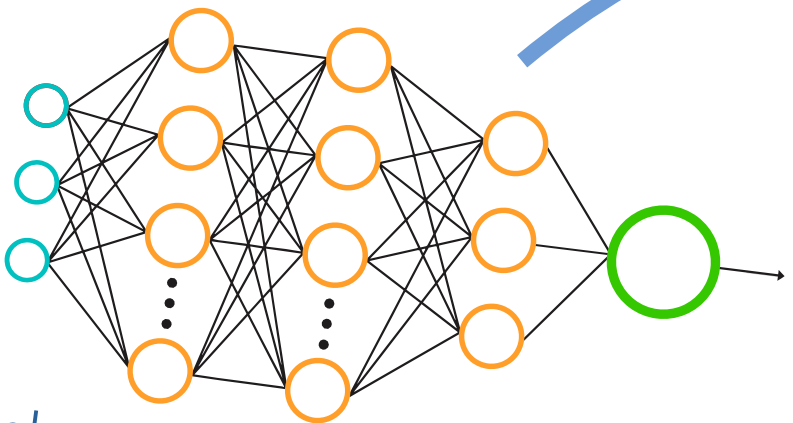
AMD VERSAL ARCHITECTURE – AI ENGINE



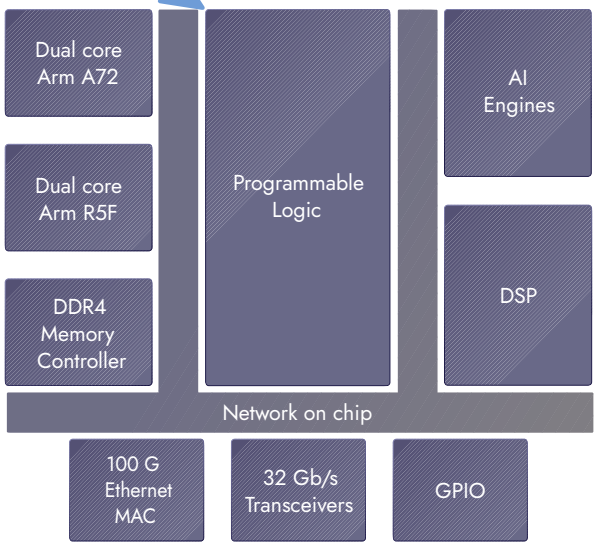
- Running at 1.25 GHz
- Instruction Level Parallelism via Very Long Instruction Words for **parallel execution** of:
  - 1 Scalar Instruction
  - 1 Vector Instruction
  - 2 Loads
  - 2 Register Moves
  - 1 Store
- **Single Instruction Multiple Data** (SIMD):
  - e.g. 8 Float32 or 128 8-bit integer MACs
  - Pre-adder

THE PIECES OF THE PUZZLE ARE THERE

Translation to hard- & software tuned to ultra low latencies

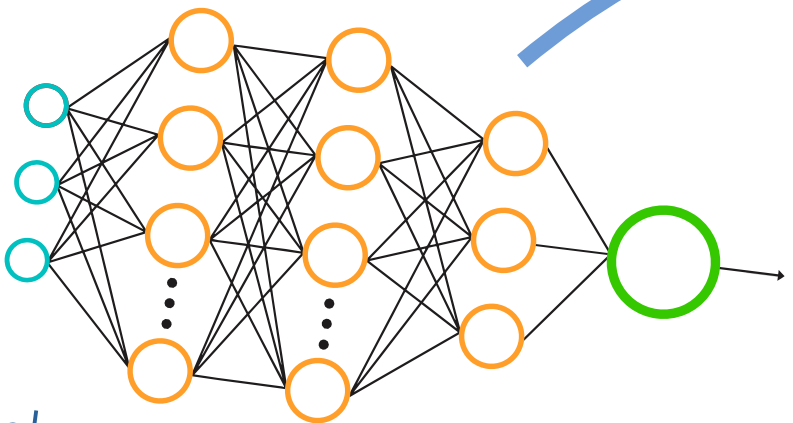


High speed algorithm development

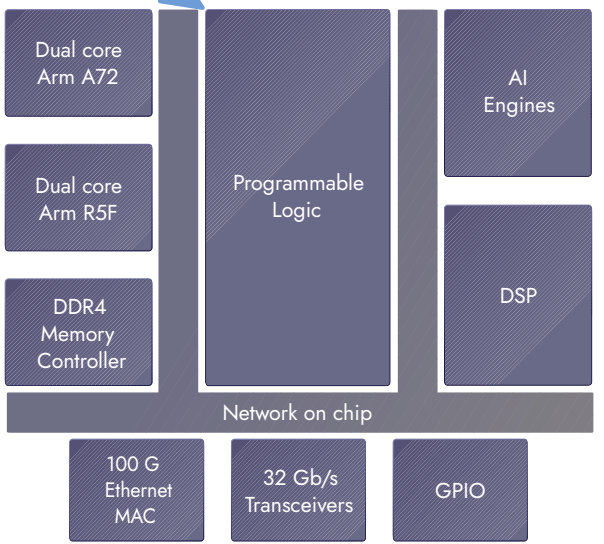


THE PIECES OF THE PUZZLE ARE THERE

Translation to hard- & software tuned to ultra low latencies



High speed algorithm development



Not a one person job!

UNIVERSITÄT BONN

ERUM-DATA DEEP : DATA EFFICIENCY IN EMBEDDED PROCESSORS

Collaborating across universities, companies…

KIT Karlsruher Institut für Technologie

universität freiburg

TUHH Technische Universität Hamburg

DESY.

UNIVERSITÄT BONN

SICK

SiMa ai

UNIVERSITÄT BONN

ERUM-DATA DEEP : DATA EFFICIENCY IN EMBEDDED PROCESSORS

... and experiments

Belle II

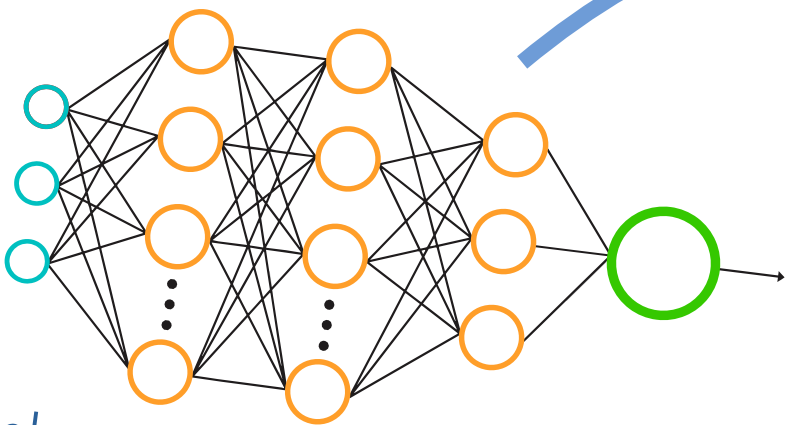
European XFEL

LHCb

PETRA III

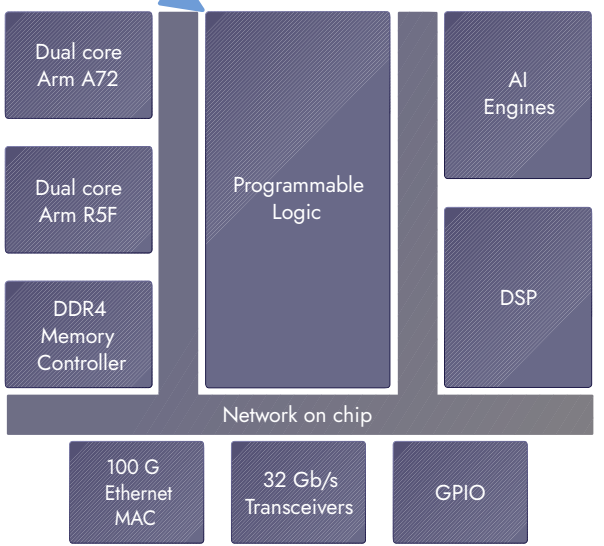
FRM II
Forschungs-Neutronenquelle
Heinz Maier-Leibnitz

A LOT OF MANUAL LABOUR



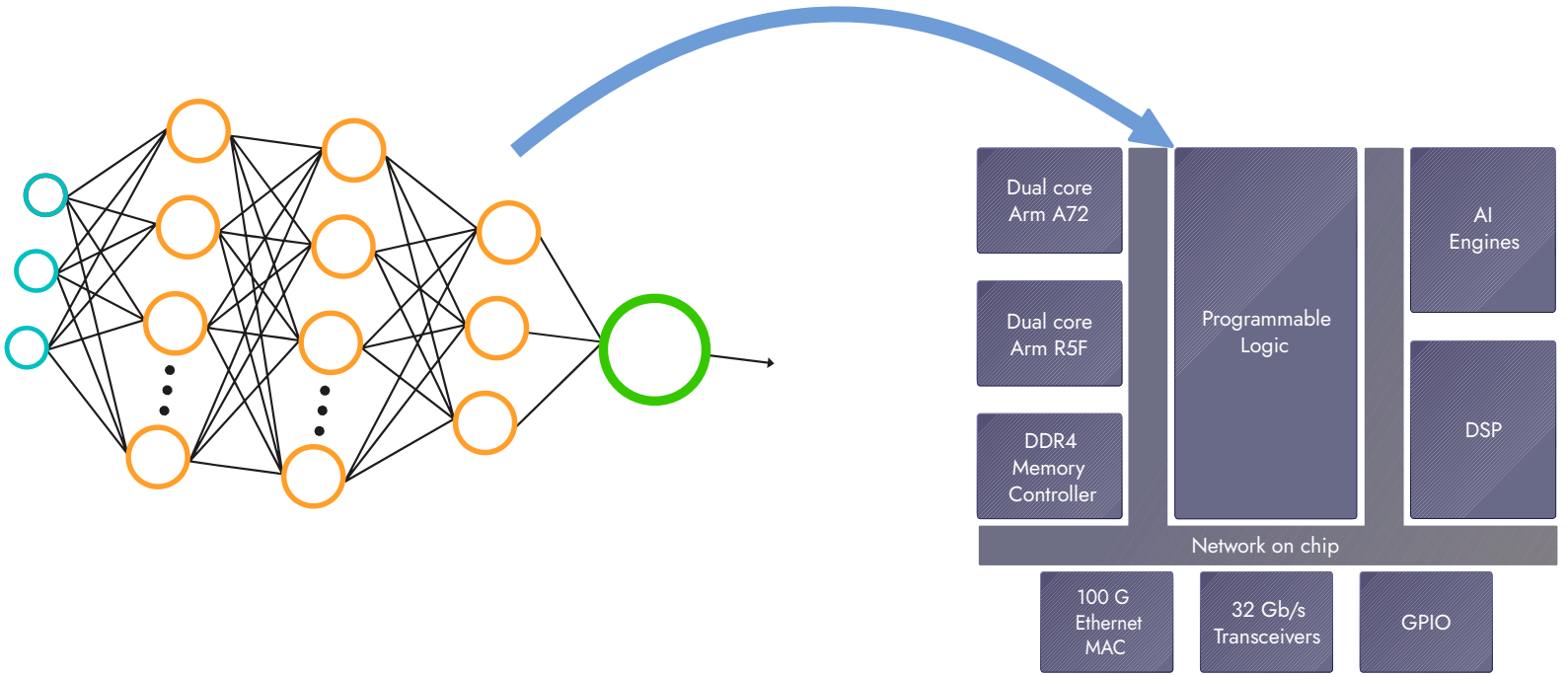
High speed algorithm development

Translation to hard- & software tuned to ultra low latencies

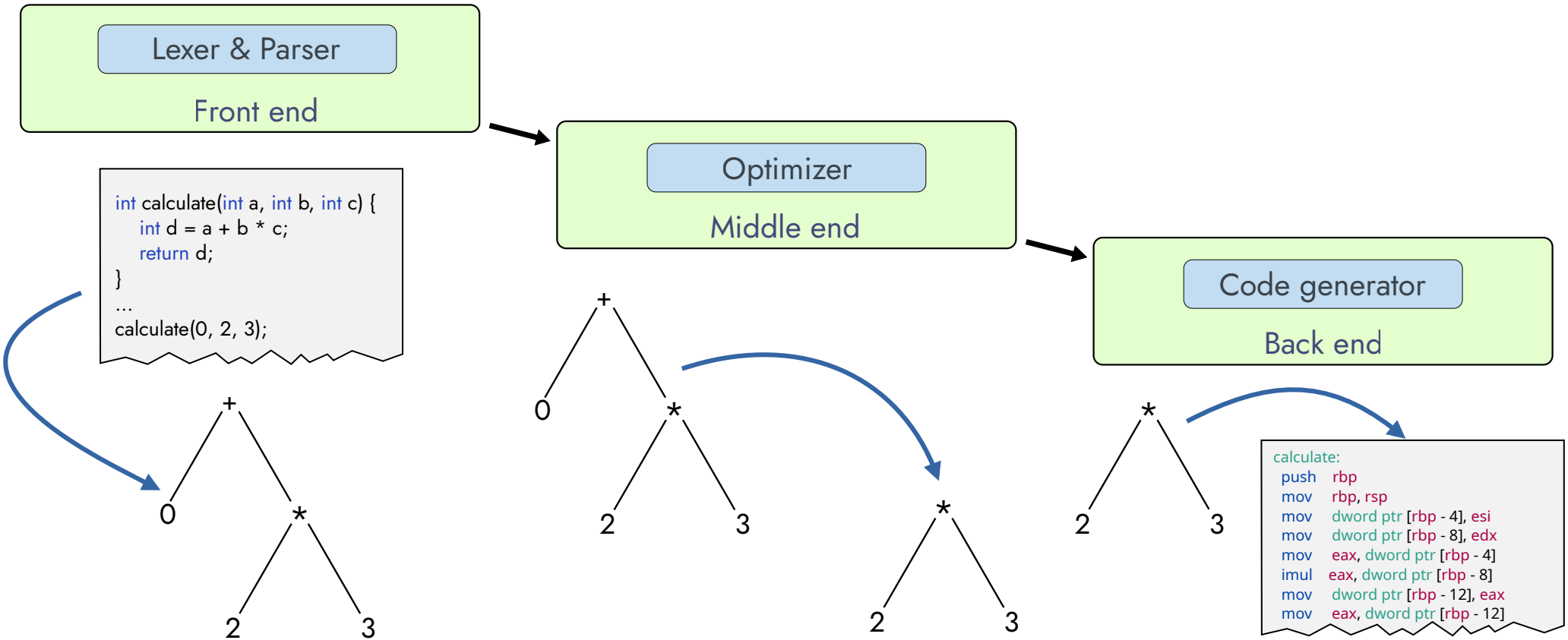


SPEEDING UP THE TRANSLATION PROCESS

→ ONNX to Versal Compiler



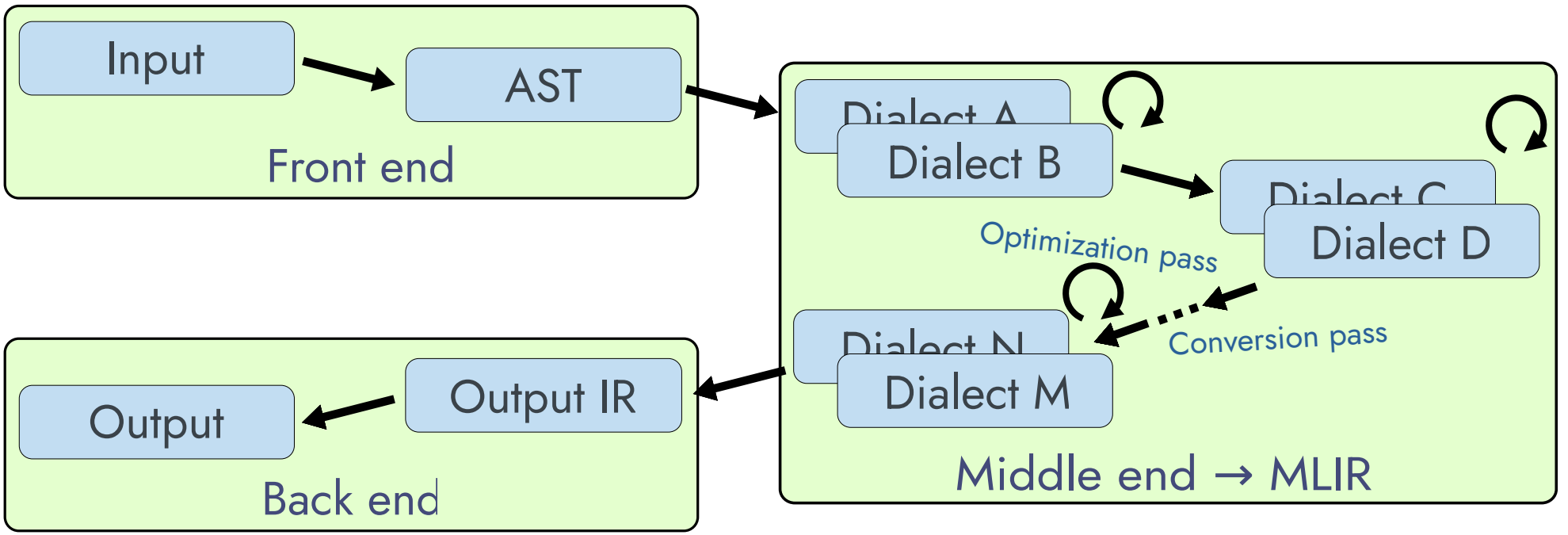
HOW TO BUILD A COMPILER?



THE MULTI-LEVEL INTERMEDIATE REPRESENTATION COMPILER FRAMEWORK (MLIR)

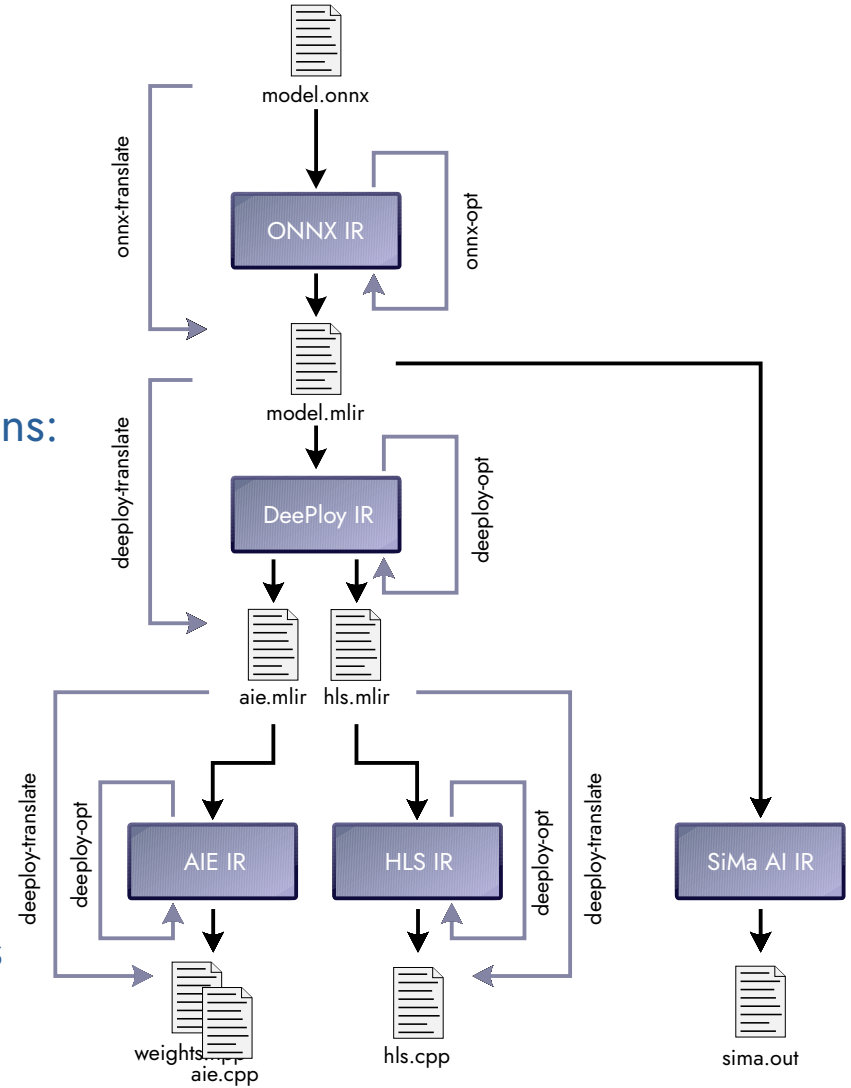
- Framework to build compilers
- Part of the LLVM project
- Primarily middle end
- Simple interfacing to existing front- and back ends
- Code represented through **intermediate representations** (IR)
- IRs can consist of variety of **dialects**
- Translating from input IR to output IR via **dialect conversions**
- Optimizations on individual IRs

THE MULTI-LEVEL INTERMEDIATE REPRESENTATION COMPILER FRAMEWORK (MLIR)



DEEPLOY TOOLCHAIN

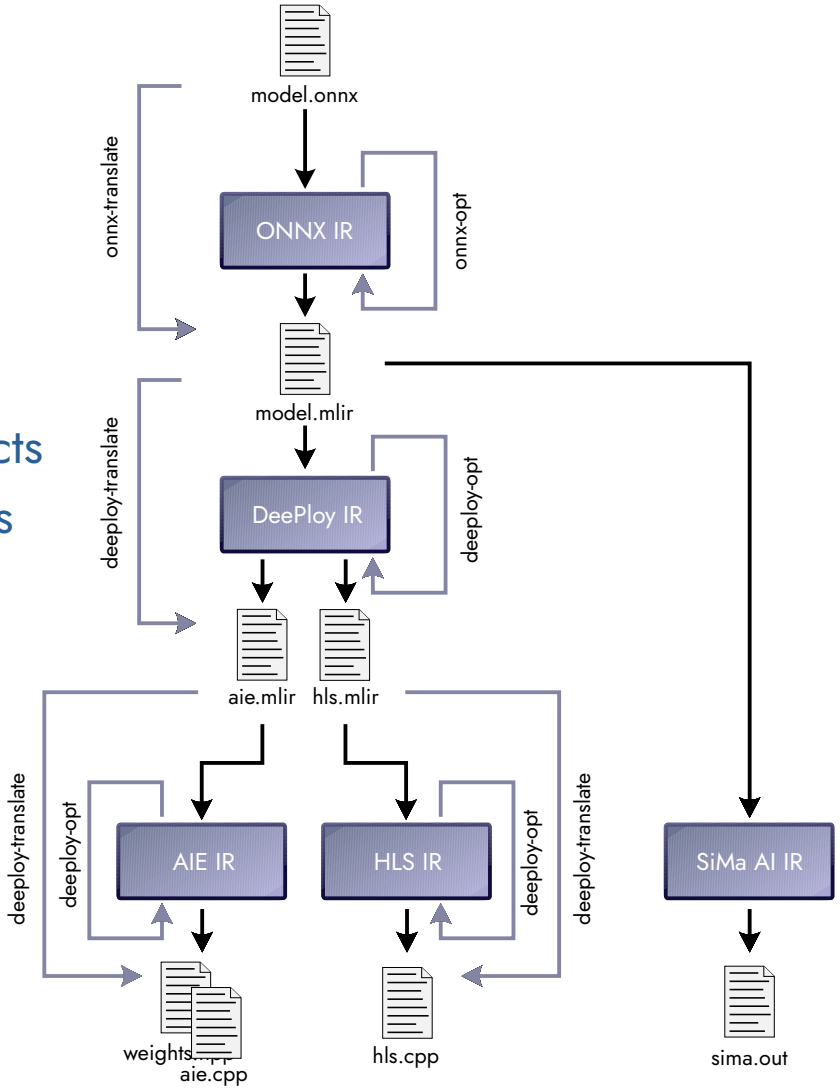
- ONNX-MLIR for input translation to MLIR IR
- Three new dialects for different lowering levels & domains:
  - **HLS dialect** for operators mapped to the PL
  - **AIE dialect** for operators mapped to the AIE
  - **DeePloy dialect** as intermediate dialect
- Each with support for all operators necessary for developed algorithms
- Support for quantized operators
  → Extending ONNX-MLIR to support QONNX operators



DEEPLOY TOOLCHAIN

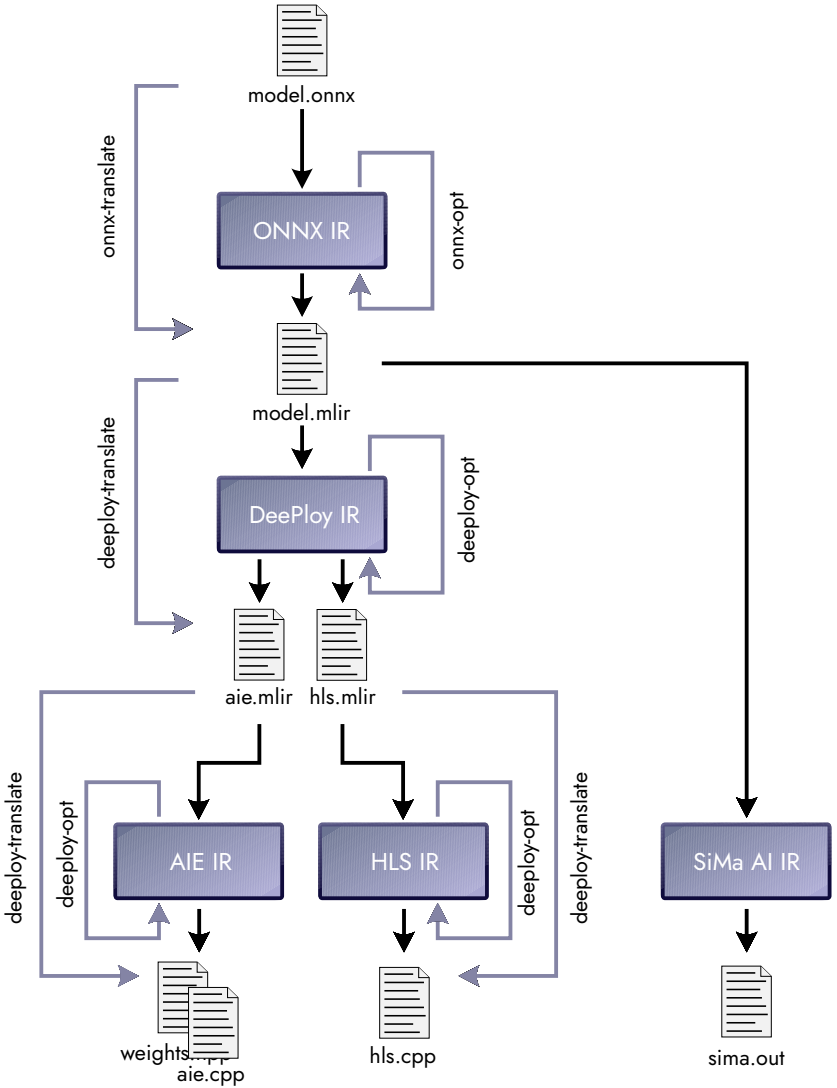
Two MLIR-based tools working with these dialects:

- **deeploy-translate** for conversion passes between dialects
  - Conversion from DeePloy dialect to AIE & HLS dialects
  - Translation from AIE & HLS dialects to source code
- **deeploy-opt** for optimization passes on dialects
  - Mapping of operators to PL or AIE using MILP solver
  - Architecture specific optimizations for AIE & PL



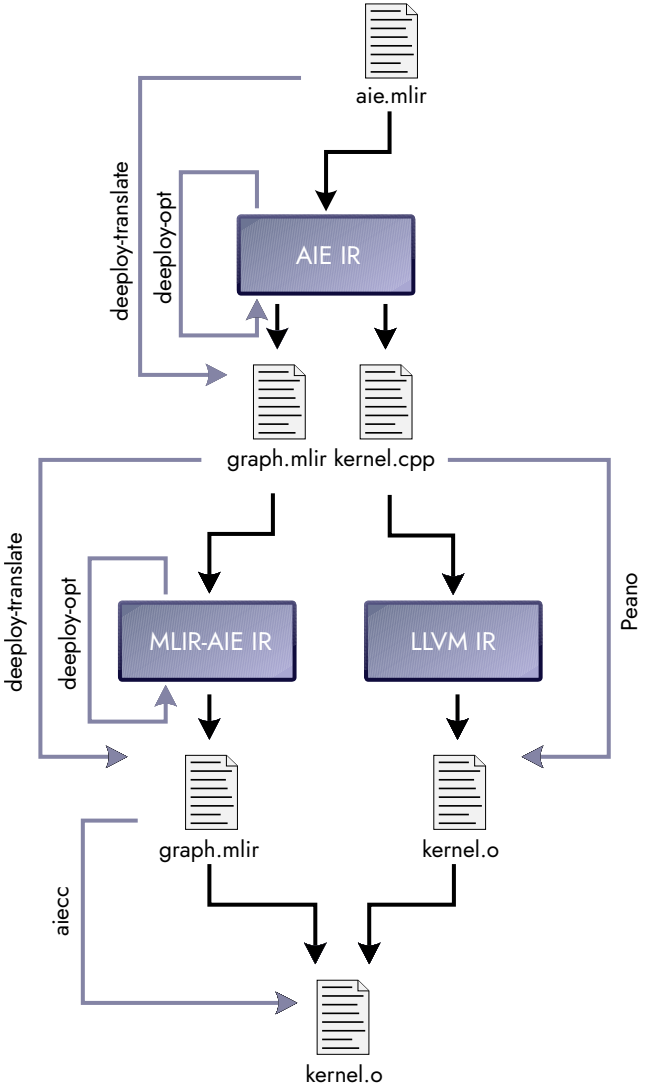
DEEPLOY TOOLCHAIN

- Output generator is the EmitC emitter of MLIR
  → Generated output: **HLS & AIE C++ source code**
- Using highly optimized pre-implemented operators
- Final compilation step via
  - AMD Vitis HLS compiler
  - AMD Vitis AIE compiler



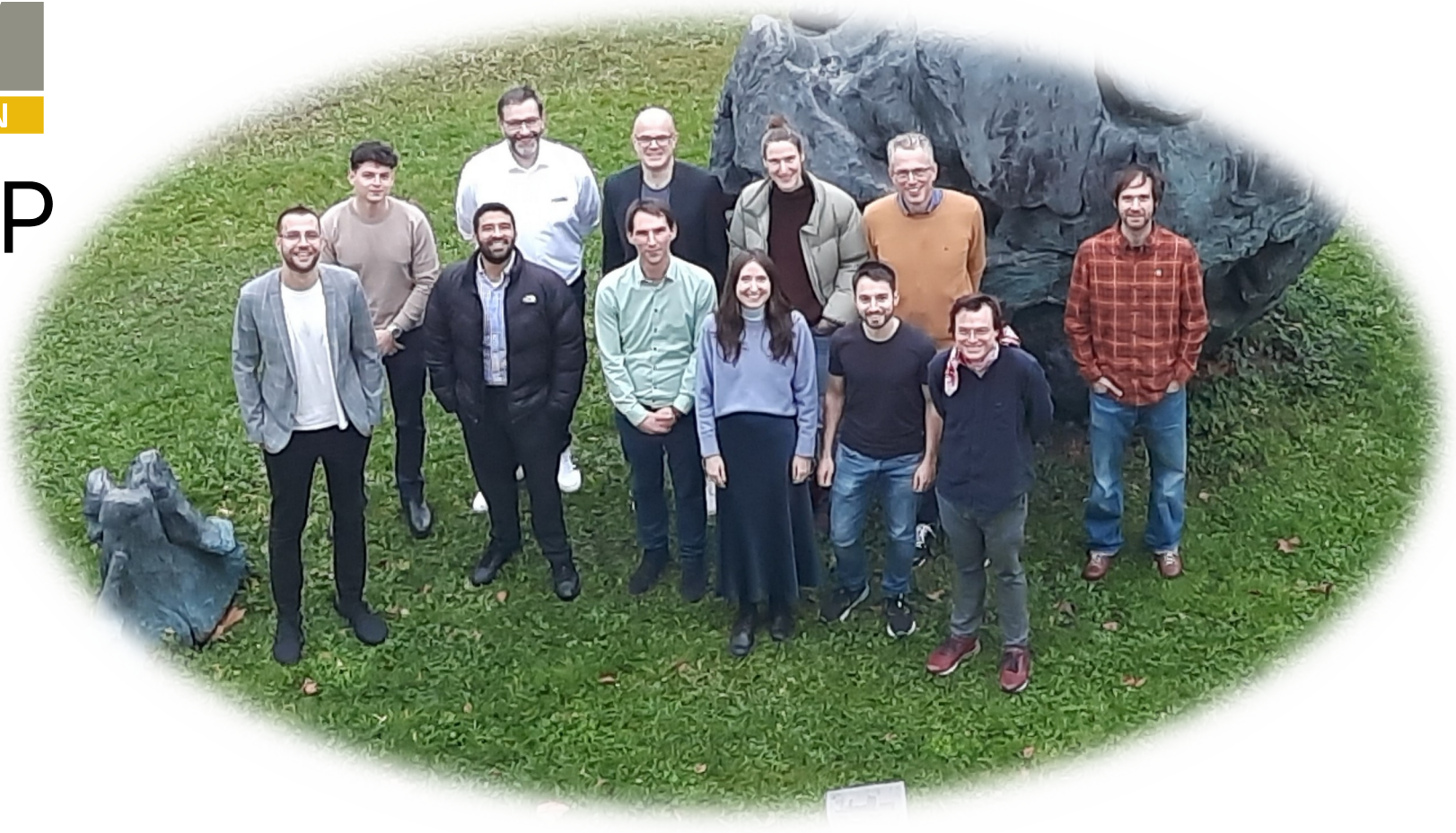
WHATS NEXT: NEAR & MORE DISTANT FUTURE

- Getting a **full conversion flow** up and running
- Adding **more operators** → support for larger model variety
- **Helper tool**: Chains individual passes
  → Simple to use for model developers
  → "one-click" conversion from ONNX to firmware
- Possibly: Replacing Vitis AIE compiler with MLIR-AIE & Peano developed by AMD



UNIVERSITÄT BONN

DEEP



Thanks for your attention

www.erumdata-deep.de