Edge ML Acceleration with RISCV-enhanced eFPGA-SoCs

Allen Boston

B. Seyoum, L. Carloni, P.-E. Gaillardon

Department of Electrical and Computer Engineering – University of Utah

Third Workshop on Open-Source Computer Architecture Research (OSCAR)
Buenos Aires, Argentina - June 29, 2024

Introduction

- Growing demand to perform ML tasks on edge devices

Natural Language Interaction

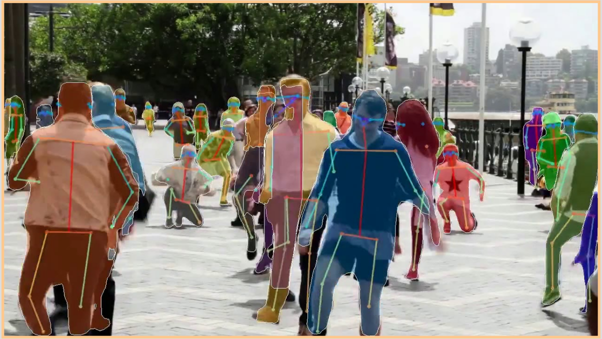
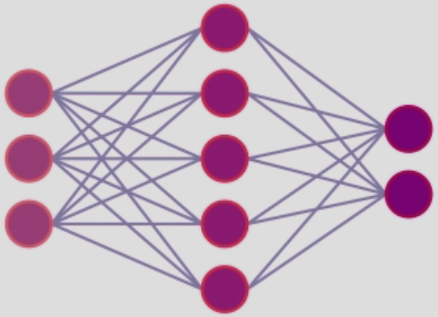


Image and Video Classification



On Device Training

- The appropriate HW deployment is:
  - *low power* - operates under severe power constraints
  - *low development effort* - easy to implement
  - *flexible* - adapt to changing ML models

These design criteria motivate innovation for specialized architectures

Motivation: Performance vs. Design Complexity

- **Cons of bespoke ML accelerator design from scratch**
  - High design effort
  - Significant software preparation
  - Low-flexibility/Fixed-function

Best Performance

High Design Complexity

Motivation: Performance vs. Design Complexity

- **Cons of bespoke ML accelerator design from scratch**
  - High design effort
  - Significant software preparation
  - Low-flexibility/Fixed-function

Best Performance
High Design Complexity

- **Role of Compilers + General Purpose Compute**
  - Streamline hardware interfacing
  - High-flexibility

Performance Suffers
Low Design Complexity

Motivation: Performance vs. Design Complexity

- **Cons of bespoke ML accelerator design from scratch**
  - High design effort
  - Significant software preparation
  - Low-flexibility/Fixed-function

Best Performance
High Design Complexity

- **Role of Compilers + General Purpose Compute**
  - Streamline hardware interfacing
  - High-flexibility

Performance Suffers
Low Design Complexity

- **Reuse in the Open-Source Domain**
  - Leverage existing solutions to reduce design complexity
  - Promotes interoperability across projects

Motivation: Performance vs. Design Complexity

- **Cons of bespoke ML accelerator design from scratch**
  - High design effort
  - Significant software preparation
  - Low-flexibility/Fixed-function

Best Performance
High Design Complexity

- **Role of Compilers + General Purpose Compute**
  - Streamline hardware interfacing
  - High-flexibility

Performance Suffers
Low Design Complexity

- **Reuse in the Open-Source Domain**
  - Leverage existing solutions to reduce design complexity
  - Promotes interoperability across projects

Is it possible to combine aspects of these general implementation approaches with specialized open-source hardware?

Background: Edge ML and Open-Source CPUs

- **TensorFlow (Lite)** open-source development of ML models
  - Easy to use - training, inference, model tuning
  - Pre-trained models - highly optimized
  - Easy portability to open-source hardware with RISC-V
- **RISC-V** open-source development of efficient general purpose computing platforms
  - Highly flexible HW and EDA tools
  - Easily programmable with software toolchain support
  - Enhanced performance with custom ISA extensions

Background: Edge ML and Open-Source CPUs

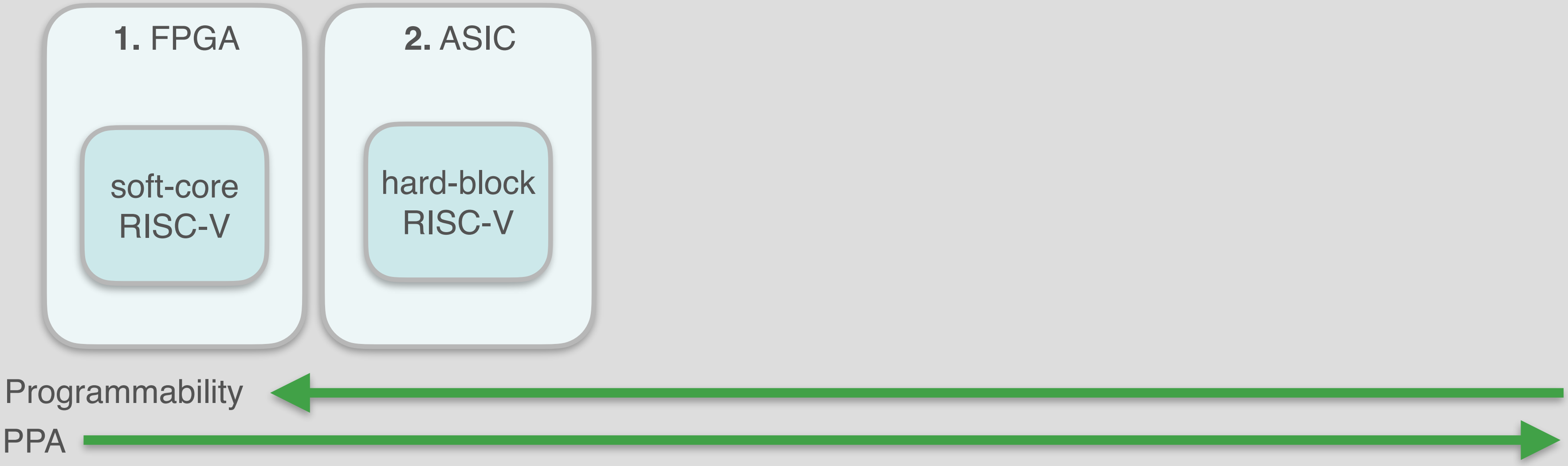
- **TensorFlow (Lite)** open-source development of ML models
  - Easy to use - training, inference, model tuning
  - Pre-trained models - highly optimized
  - Easy portability to open-source hardware with RISC-V
- **RISC-V** open-source development of efficient general purpose computing platforms
  - Highly flexible HW and EDA tools
  - Easily programmable with software toolchain support
  - Enhanced performance with custom ISA extensions

TensorFlow + RISC-V

Strong foundation for ML hardware

Related Work: TFLite + RISC-V + ISA Extensions

Architecture	TensorFlow Lite	ISA Extensions	Post-Silicon Reconfigurability	Open-Source Hardware	SoC Integration
1	✓	✗	✗	✗	N/A
2	✓	✗	✗	✗	N/A
3					
4					
5					
This Work					



1. FPGA

soft-core RISC-V

2. ASIC

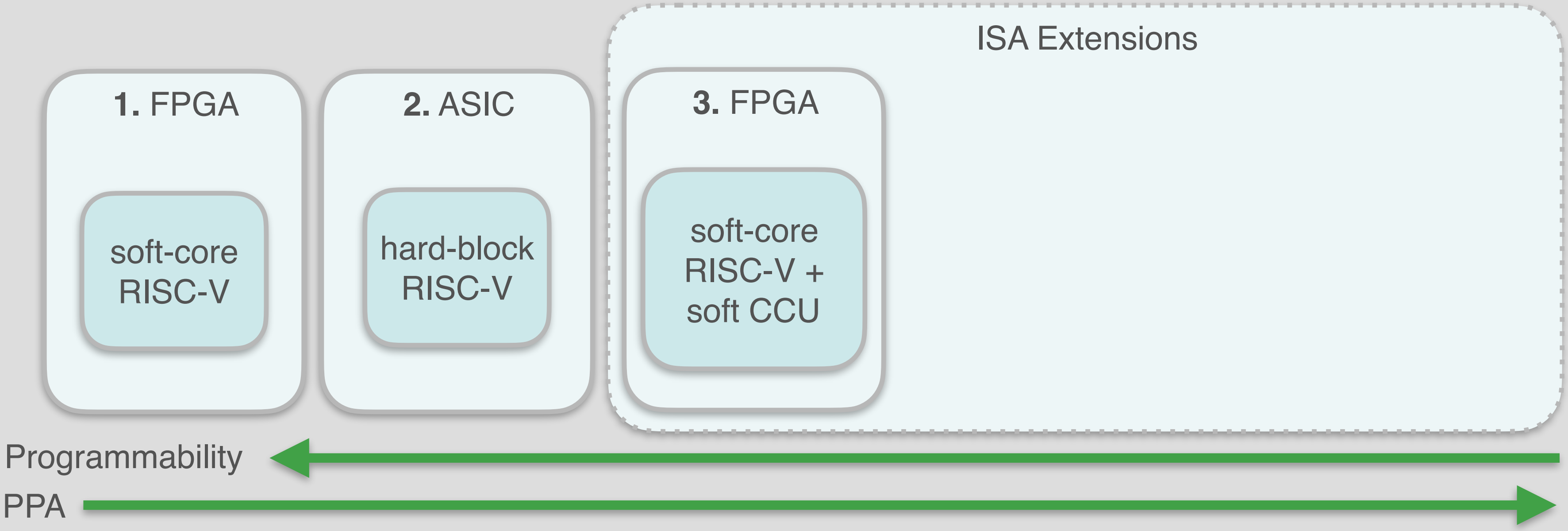
hard-block RISC-V

Programmability

PPA

Related Work: TFLite + RISC-V + ISA Extensions

Architecture	TensorFlow Lite	ISA Extensions	Post-Silicon Reconfigurability	Open-Source Hardware	SoC Integration
1	✓	✗	✗	✗	N/A
2	✓	✗	✗	✗	N/A
3	✓	✓	✓	✗	✓
4					
5					
This Work					



ISA Extensions

1. FPGA

soft-core RISC-V

2. ASIC

hard-block RISC-V

3. FPGA

soft-core RISC-V + soft CCU

Programmability

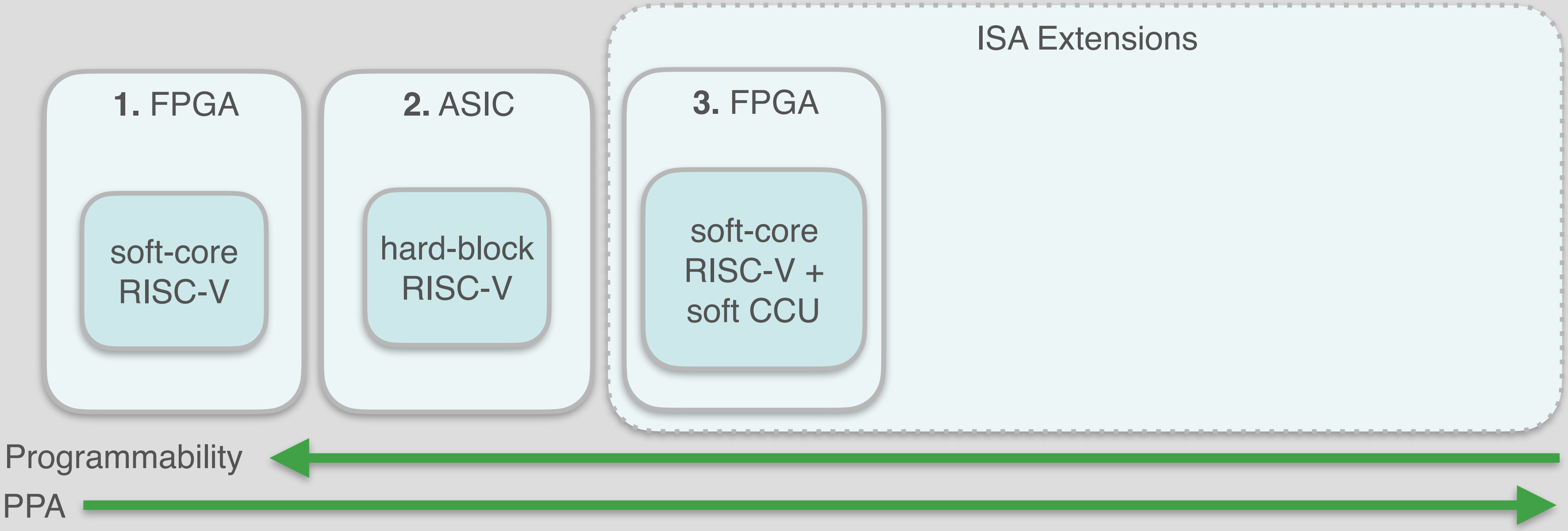
PPA

Related Work: TFLite + RISC-V + ISA Extensions

Architecture	TensorFlow Lite	ISA Extensions	Post-Silicon Reconfigurability	Open-Source Hardware	SoC Integration
1	✓	✗	✗	✗	N/A
2	✓	✗	✗	✗	N/A
3	✓	✓	✓	✗	✓
4					
5					
This Work					

Acknowledgements:

- CFU-Playground
- Efinix TinyMl Platform



ISA Extensions

1. FPGA

soft-core RISC-V

2. ASIC

hard-block RISC-V

3. FPGA

soft-core RISC-V + soft CCU

Programmability

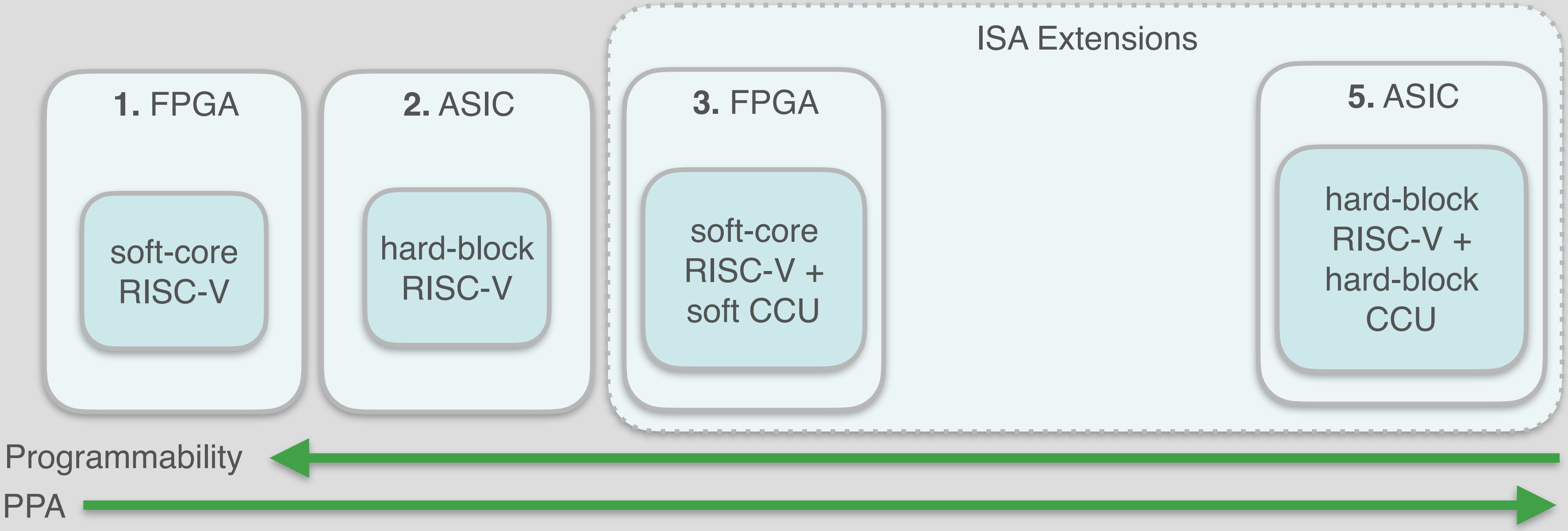
PPA

Related Work: TFLite + RISC-V + ISA Extensions

Architecture	TensorFlow Lite	ISA Extensions	Post-Silicon Reconfigurability	Open-Source Hardware	SoC Integration
1	✓	✗	✗	✗	N/A
2	✓	✗	✗	✗	N/A
3	✓	✓	✓	✗	✓
4					
5	✓	✓	✗	✓	✓
This Work					

Acknowledgements:

- CFU-Playground
- Efinix TinyMl Platform

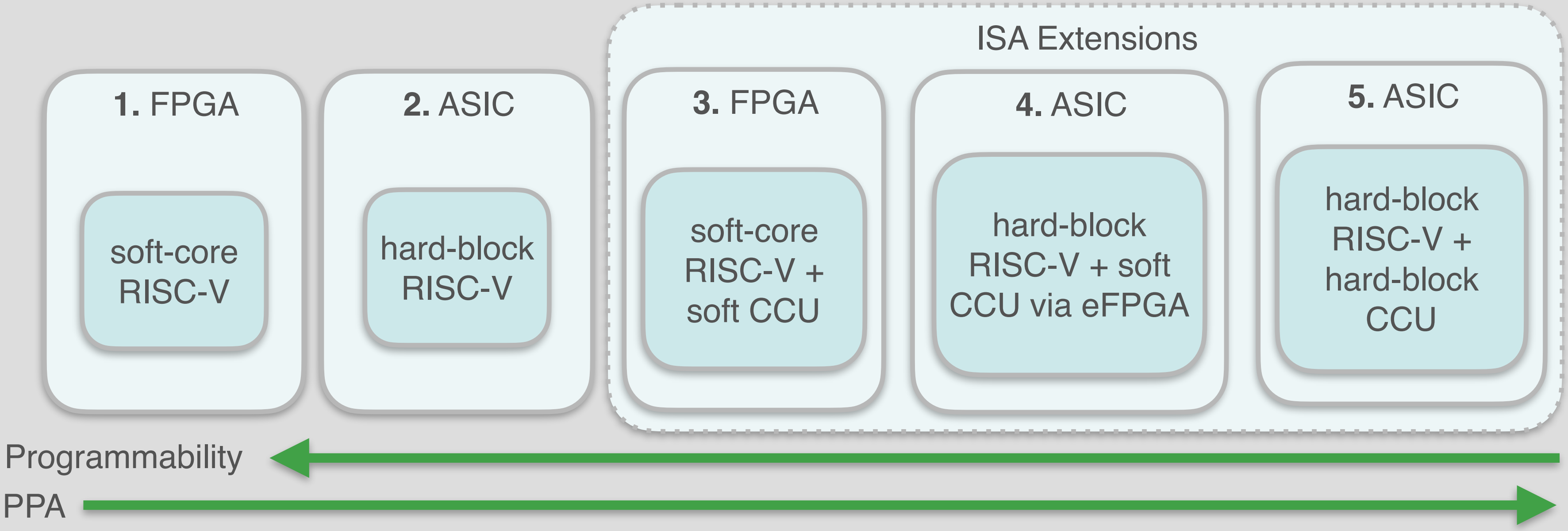


Related Work: TFLite + RISC-V + ISA Extensions

Architecture	TensorFlow Lite	ISA Extensions	Post-Silicon Reconfigurability	Open-Source Hardware	SoC Integration
1	✓	✗	✗	✗	N/A
2	✓	✗	✗	✗	N/A
3	✓	✓	✓	✗	✓
4	✗	✓	✓	✓	✓
5	✓	✓	✗	✓	✓
This Work					

Acknowledgements:

- CFU-Playground
- Efinix TinyMl Platform



ISA Extensions

1. FPGA: soft-core RISC-V

2. ASIC: hard-block RISC-V

3. FPGA: soft-core RISC-V + soft CCU

4. ASIC: hard-block RISC-V + soft CCU via eFPGA

5. ASIC: hard-block RISC-V + hard-block CCU

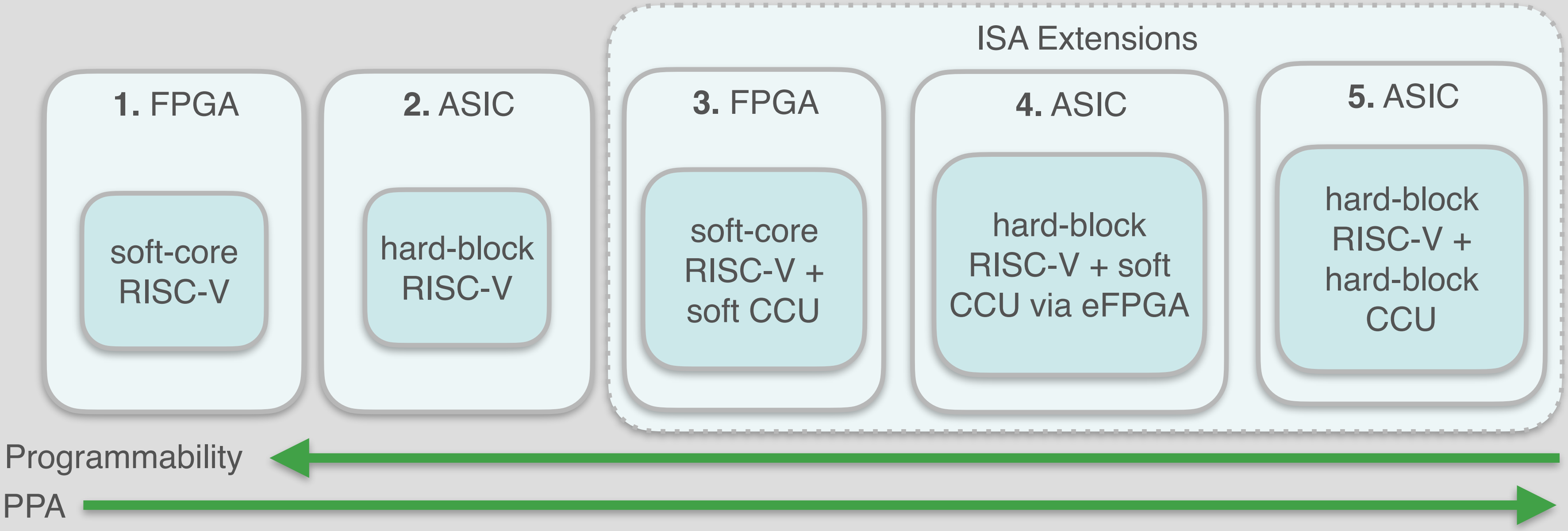
Related Work: TFLite + RISC-V + ISA Extensions

Architecture	TensorFlow Lite	ISA Extensions	Post-Silicon Reconfigurability	Open-Source Hardware	SoC Integration
1	✓	✗	✗	✗	N/A
2	✓	✗	✗	✗	N/A
3	✓	✓	✓	✗	✓
4	✗	✓	✓	✓	✓
5	✓	✓	✗	✓	✓
This Work					

Acknowledgements:

- CFU-Playground
- Efinix TinyMl Platform

- Flexbex - Fabulous
- Arnold - eFPGA-Augmented RISC-V SoC



ISA Extensions

1. FPGA: soft-core RISC-V

2. ASIC: hard-block RISC-V

3. FPGA: soft-core RISC-V + soft CCU

4. ASIC: hard-block RISC-V + soft CCU via eFPGA

5. ASIC: hard-block RISC-V + hard-block CCU

Programmability

PPA

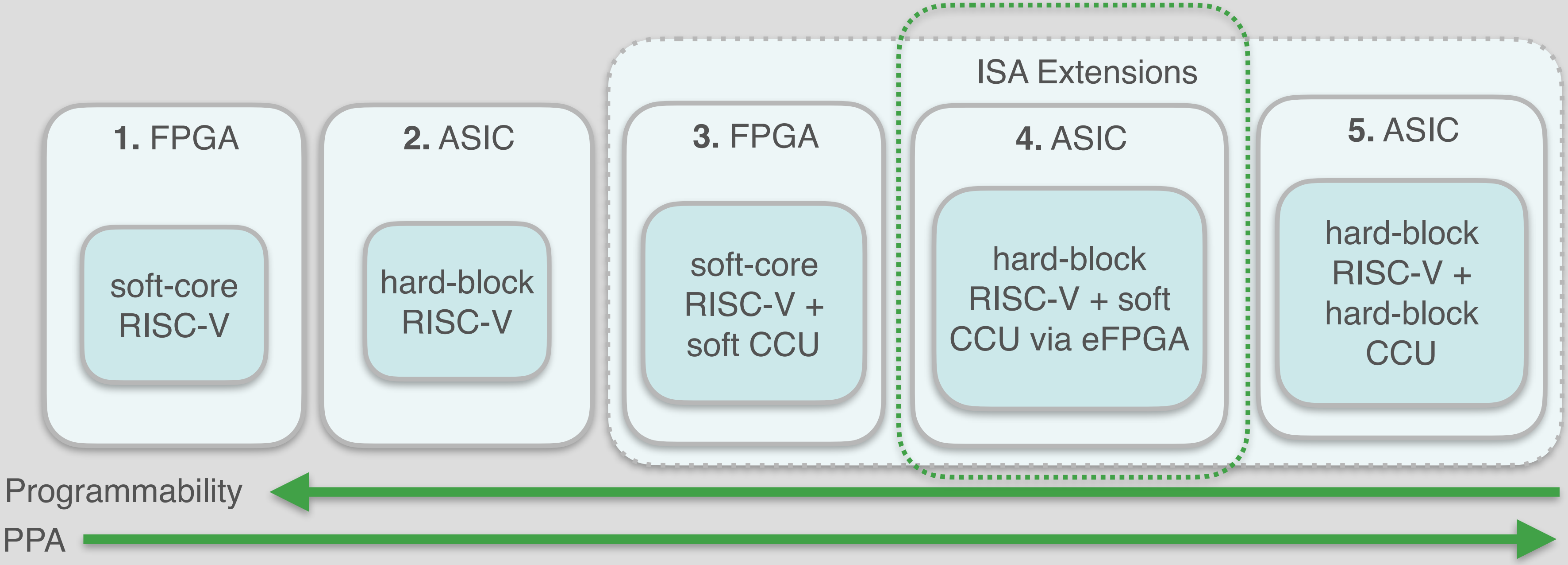
Related Work: TFLite + RISC-V + ISA Extensions

Architecture	TensorFlow Lite	ISA Extensions	Post-Silicon Reconfigurability	Open-Source Hardware	SoC Integration
1	✓	✗	✗	✗	N/A
2	✓	✗	✗	✗	N/A
3	✓	✓	✓	✗	✓
4	✗	✓	✓	✓	✓
5	✓	✓	✗	✓	✓
This Work	✓	✓	✓	✓	✓

Acknowledgements:

- CFU-Playground
- Efinix TinyMl Platform

- Flexbex - Fabulous
- Arnold - eFPGA-Augmented RISC-V SoC



Proposal

- Assemble “hard-block” Ibex RISC-V and OpenFPGA eFPGA to support ML ISA extensions
- Integrate architecture as loosely-coupled programmable accelerator in SoC with ESP
- Evaluate inference of ML Models on heterogeneous SoC

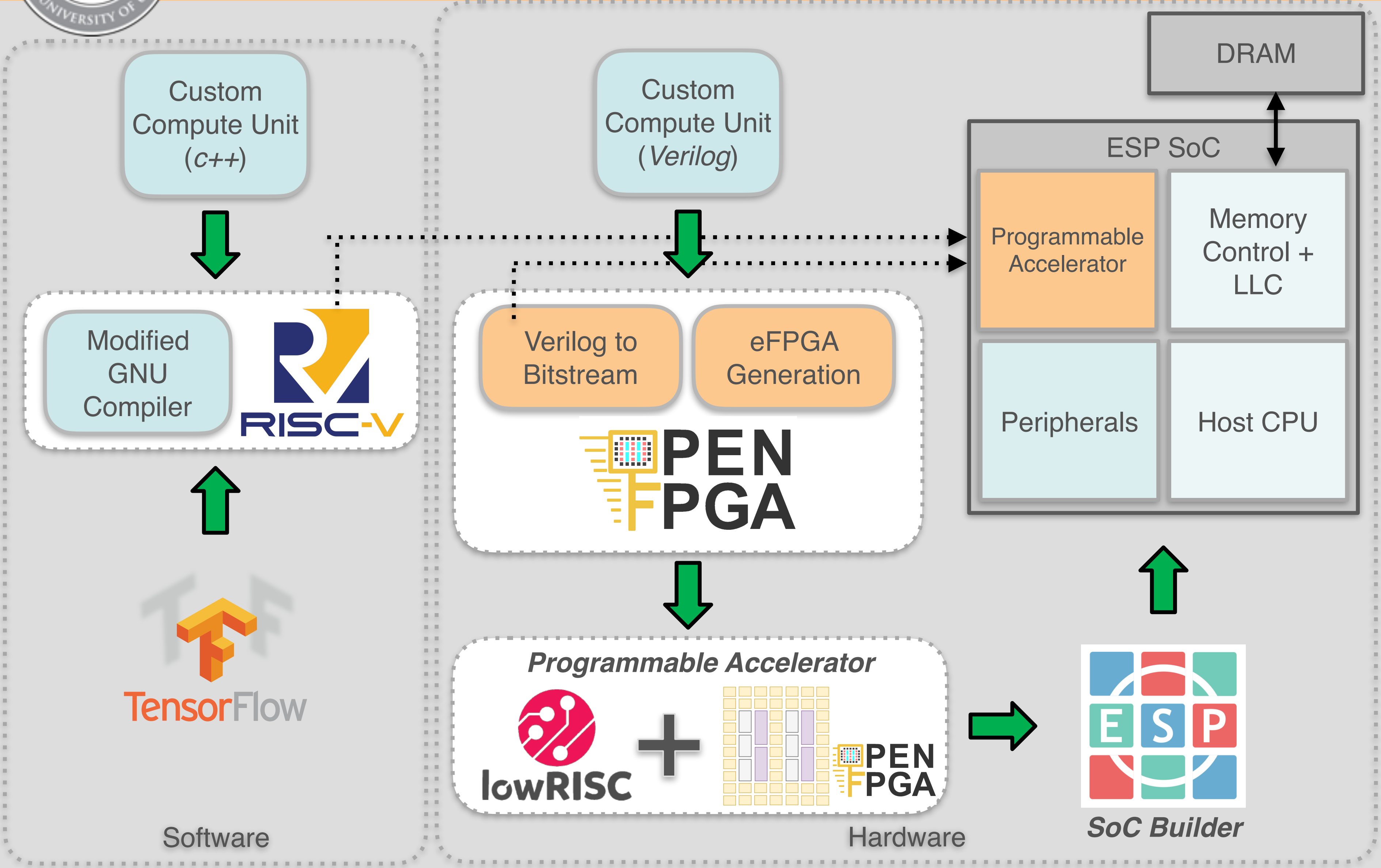
Proposal

- Assemble “hard-block” Ibex RISC-V and OpenFPGA eFPGA to support ML ISA extensions
- Integrate architecture as loosely-coupled programmable accelerator in SoC with ESP
- Evaluate inference of ML Models on heterogeneous SoC

Contributions:

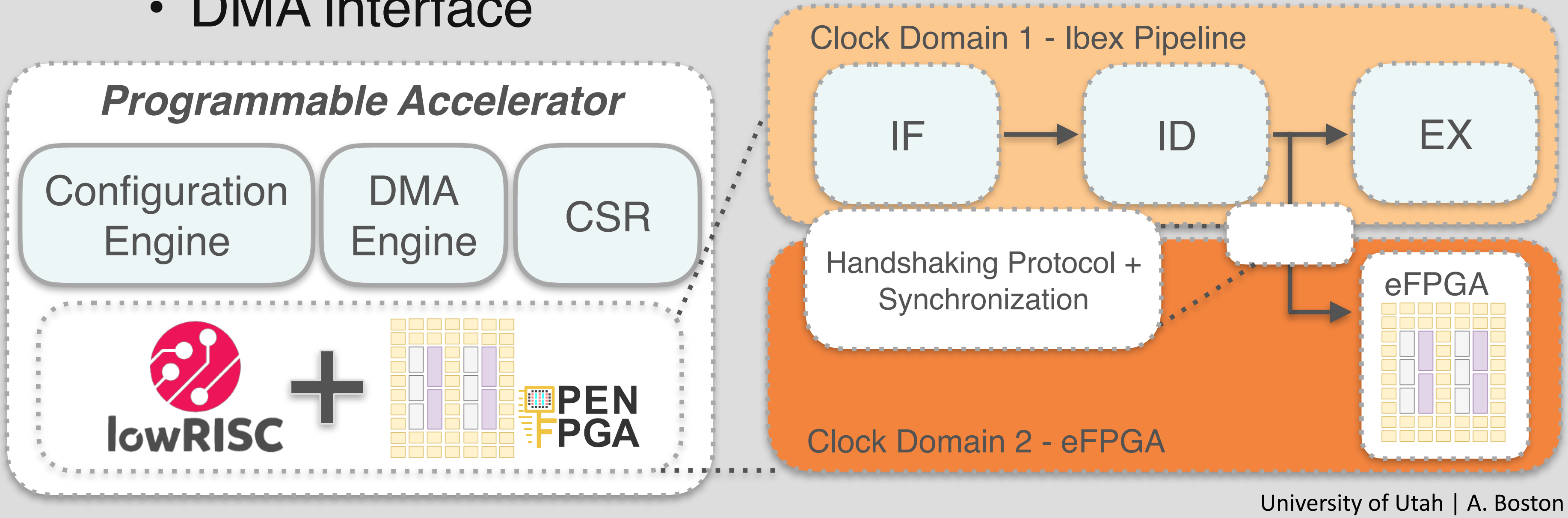
- Tightly-coupled RISC-V + eFPGA programmable accelerator
- Integration into ESP-based heterogeneous SoC
- Acceleration of TensorFlow Lite workloads

Open-Source Tool Flow



Programmable Accelerator Architecture

- Tightly-coupled Ibex RISC-V and OpenFPGA eFPGA
  - eFPGA inserted into RISC-V pipeline
  - RISC-V pipeline stalls when eFPGA invoked
  - Multi-Clock Domain management
  - eFPGA wrapper
    - Configuration engine
    - Control and status register
    - DMA interface

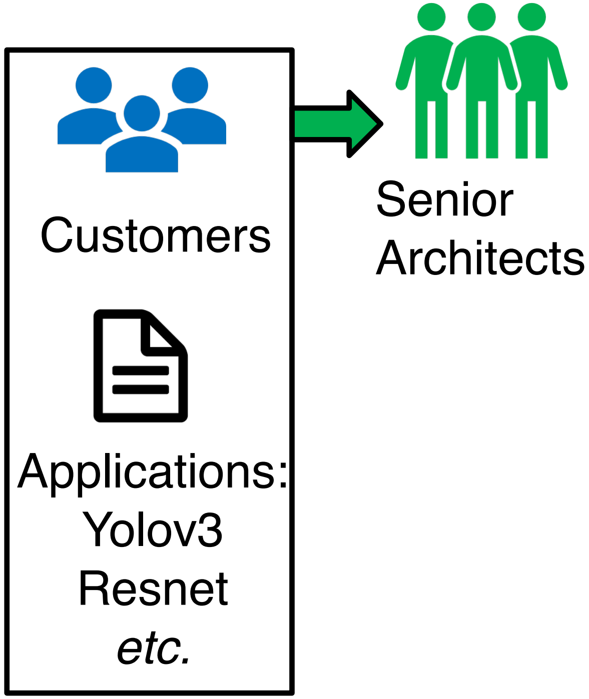


OpenFPGA - In a nutshell

Customers

Applications:
Yolov3
Resnet
etc.

OpenFPGA - In a nutshell

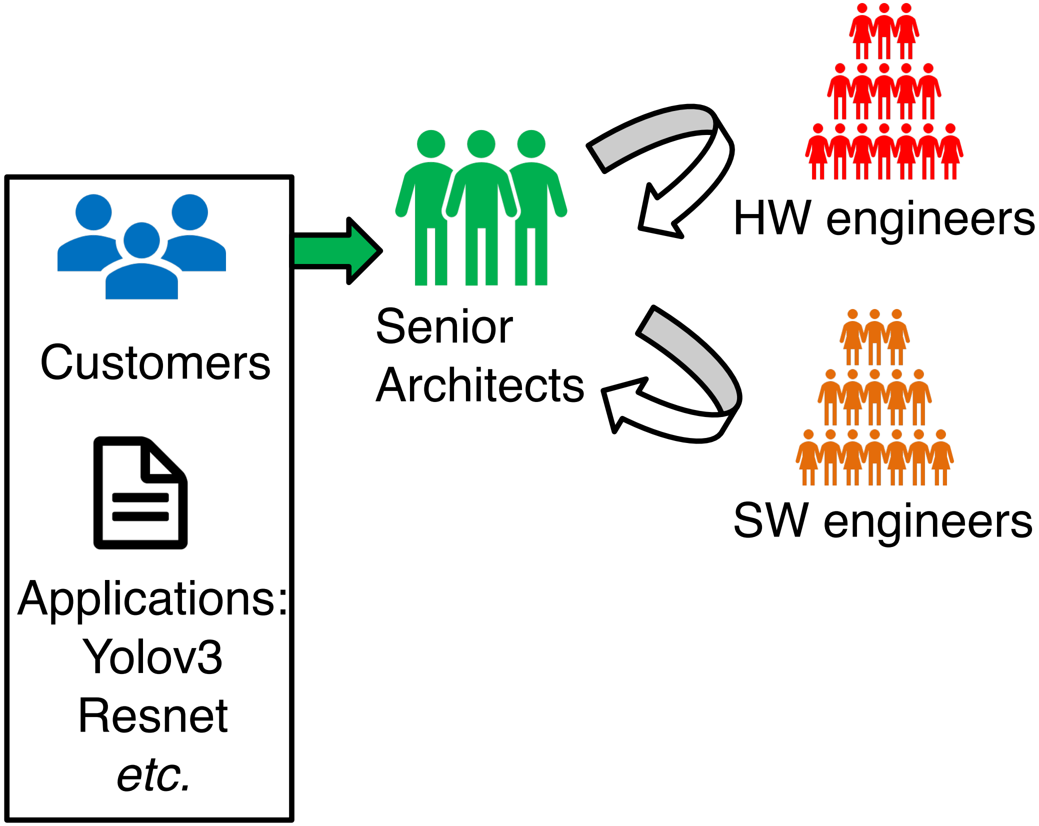


Customers

Applications:
Yolov3
Resnet
etc.

Senior
Architects

OpenFPGA - In a nutshell



Customers

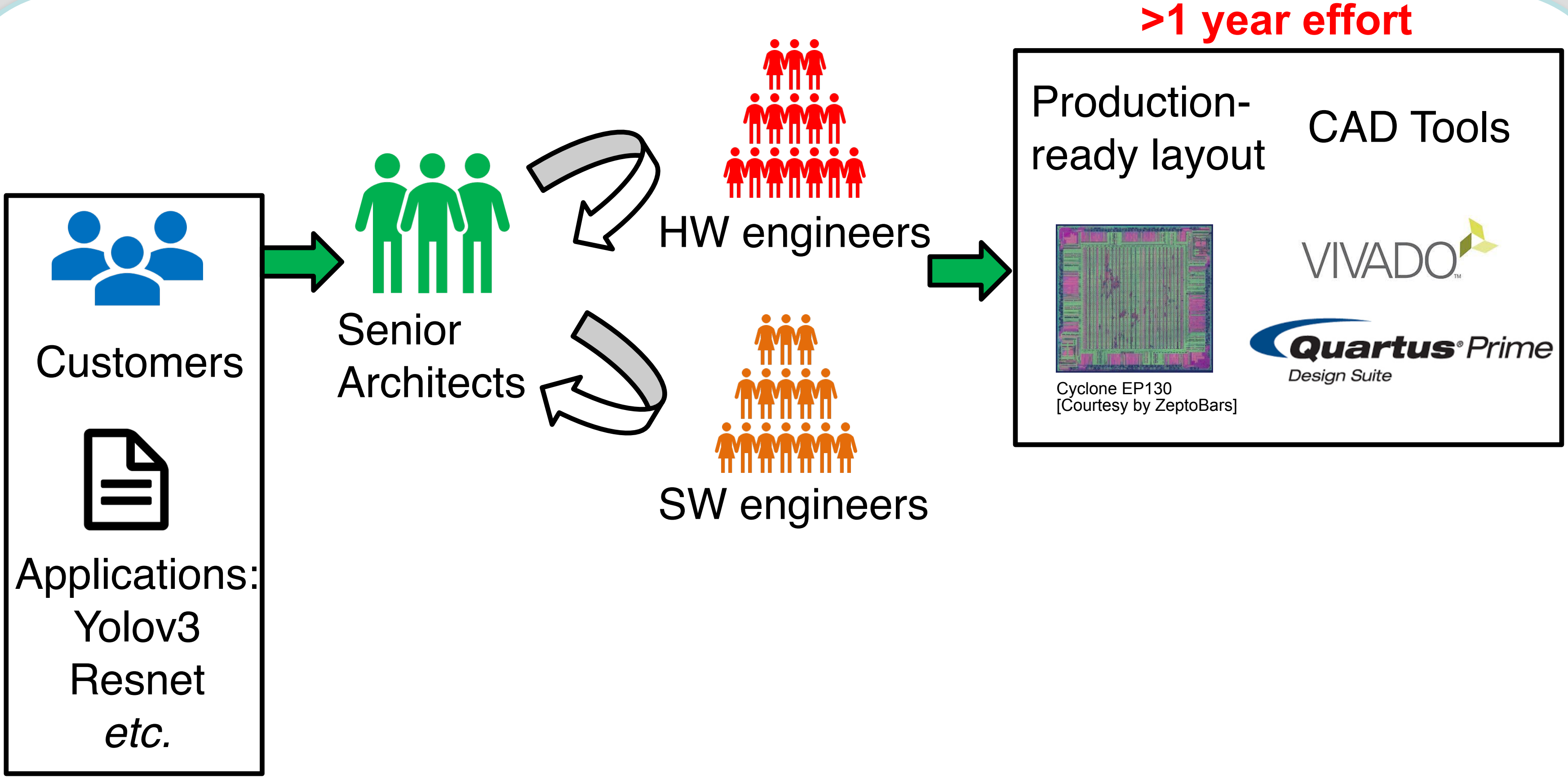
Applications:
Yolov3
Resnet
etc.

Senior
Architects

HW engineers

SW engineers

OpenFPGA - In a nutshell



Customers

Applications:
Yolov3
Resnet
etc.

Senior
Architects

HW engineers

SW engineers

>1 year effort

Production-ready layout

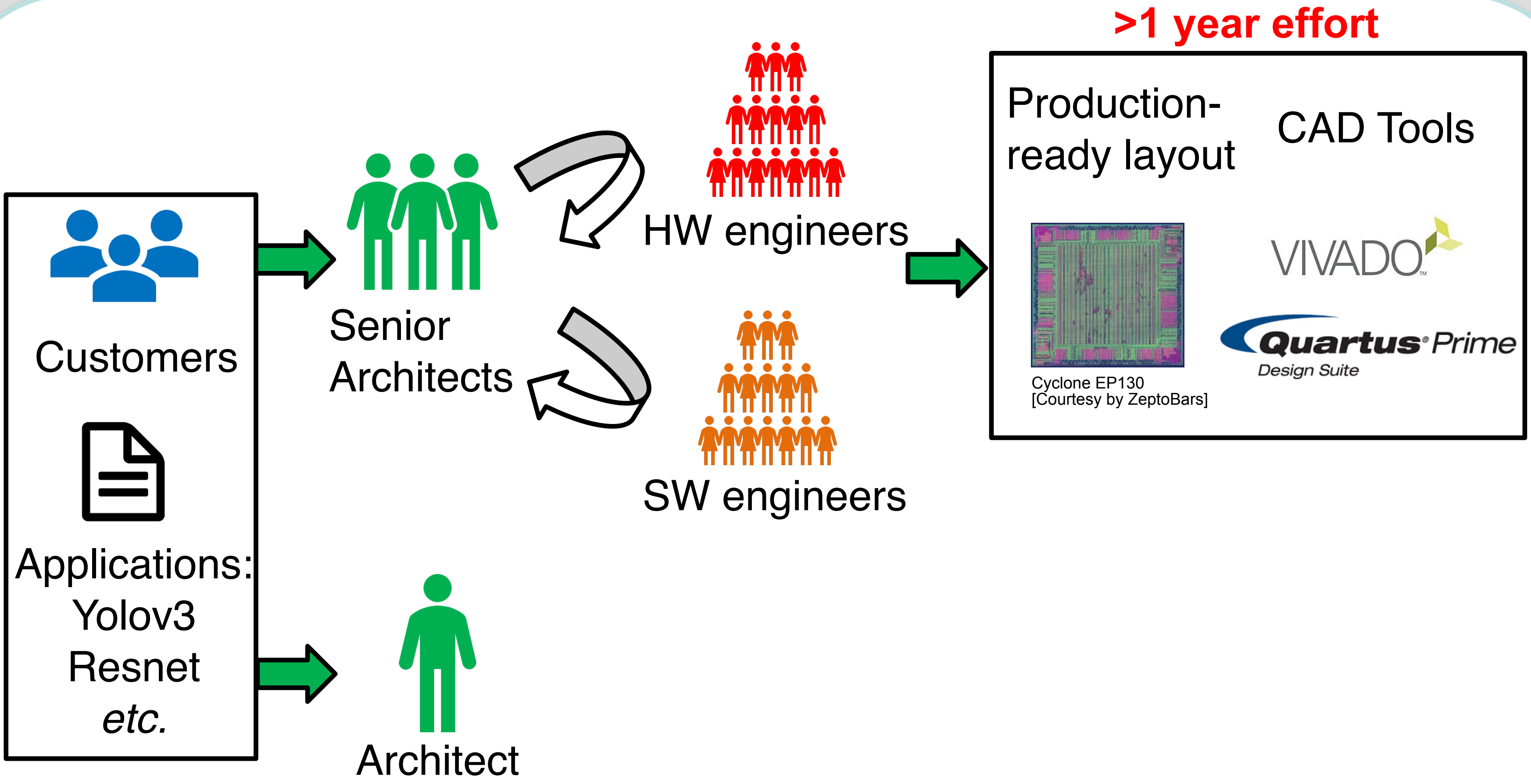
CAD Tools

Cyclone EP130
[Courtesy by ZeptoBars]

VIVADO

Quartus Prime
Design Suite

OpenFPGA - In a nutshell



>1 year effort

Customers

Applications:
Yolov3
Resnet
etc.

Senior Architects

HW engineers

SW engineers

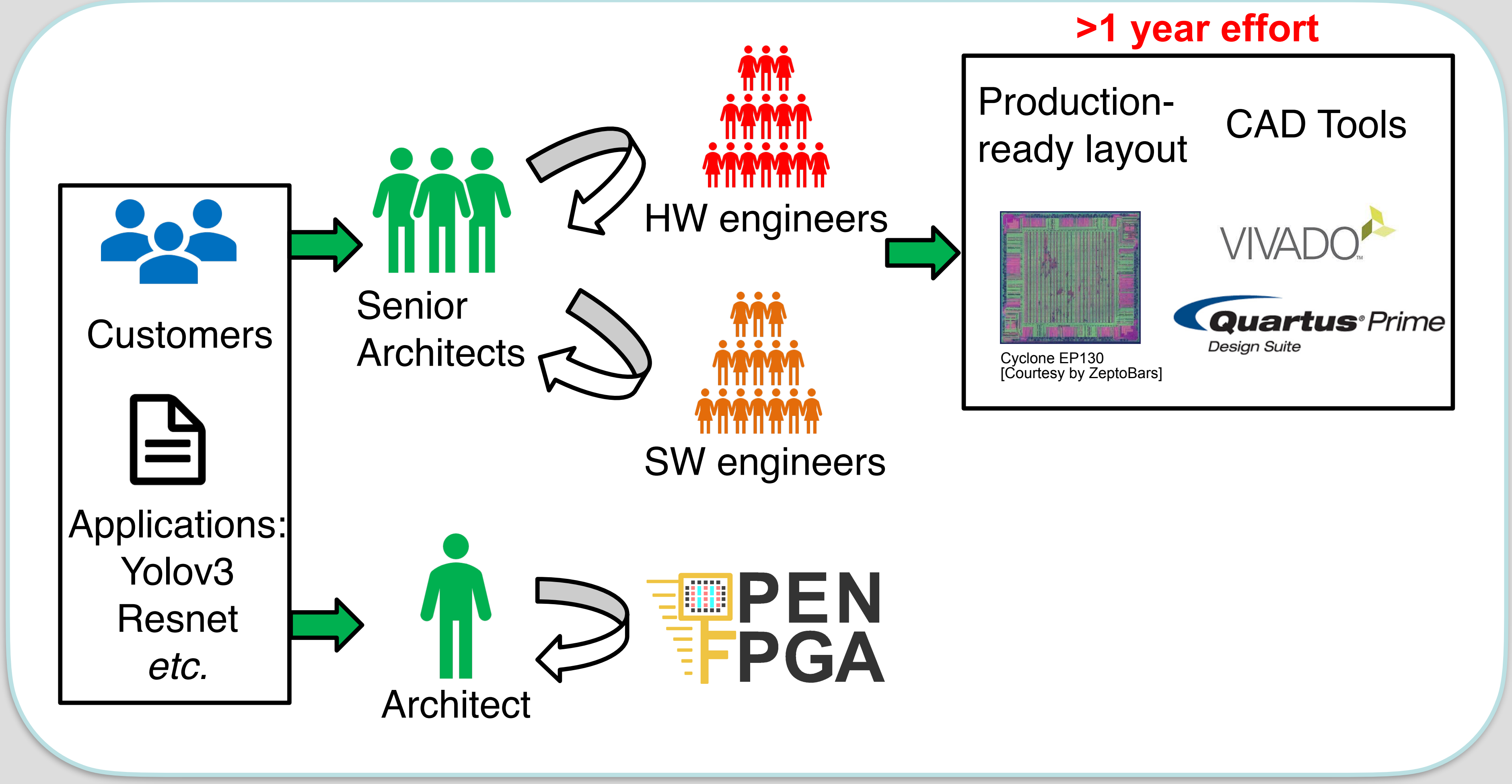
Production-ready layout

CAD Tools

Cyclone EP130
[Courtesy by ZeptoBars]

Architect

OpenFPGA - In a nutshell



>1 year effort

Customers

Applications:
Yolov3
Resnet
etc.

Senior Architects

HW engineers

SW engineers

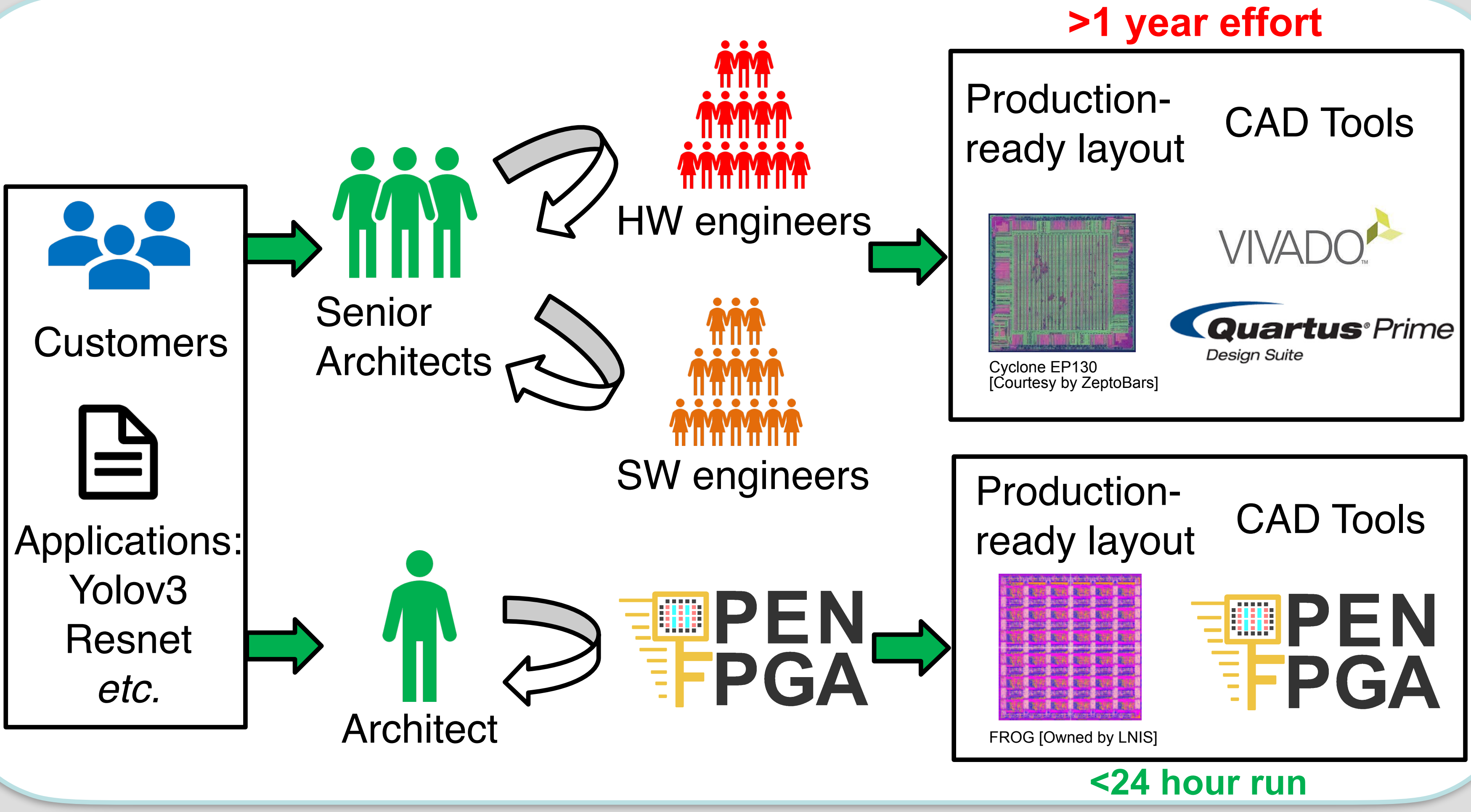
Production-ready layout

CAD Tools

Cyclone EP130
[Courtesy by ZeptoBars]

Architect

OpenFPGA - In a nutshell



>1 year effort

Customers

Applications:
Yolov3
Resnet
etc.

Senior
Architects

HW engineers

SW engineers

Production-
ready layout

CAD Tools

Cyclone EP130
[Courtesy by ZeptoBars]

Architect

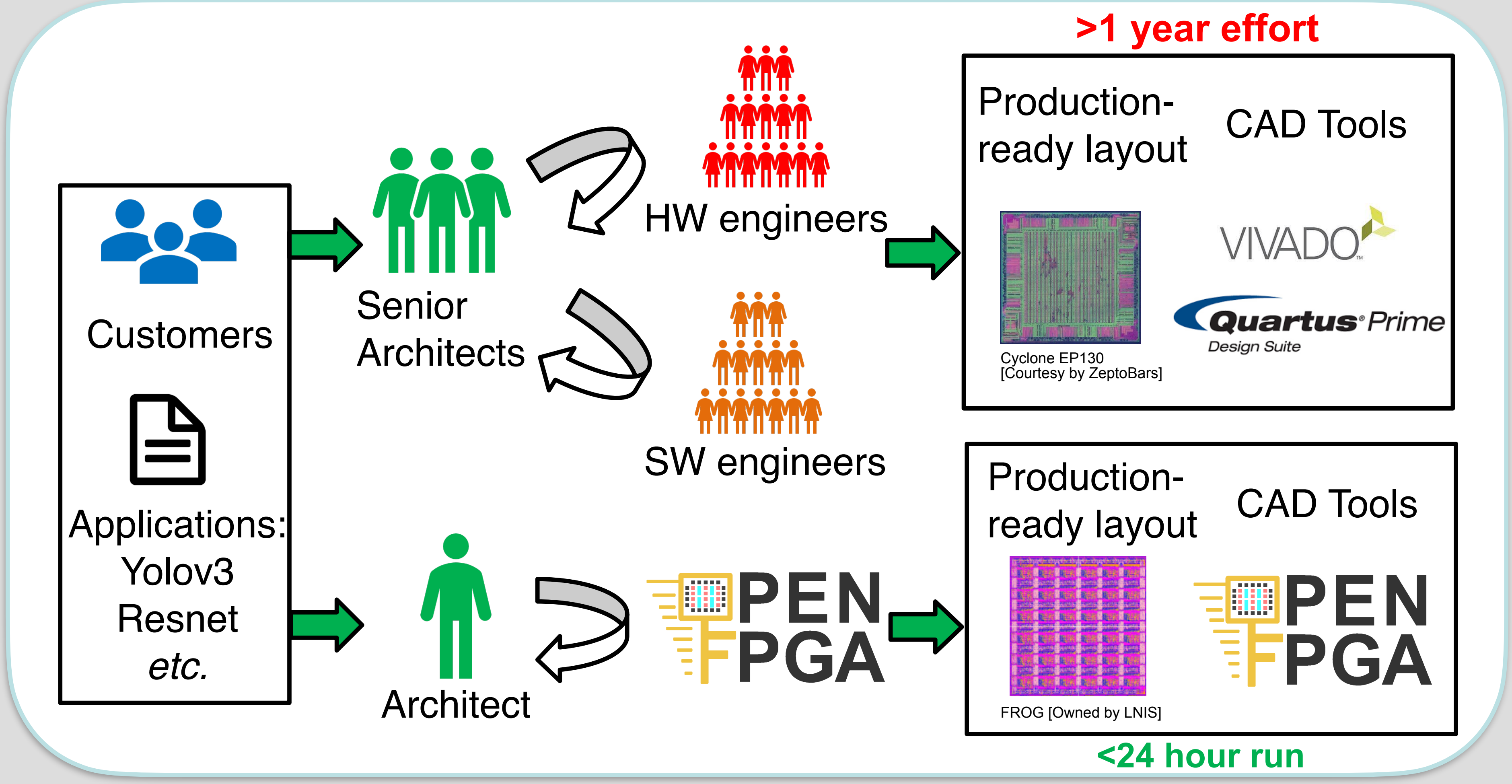
Production-
ready layout

CAD Tools

FROG [Owned by LNIS]

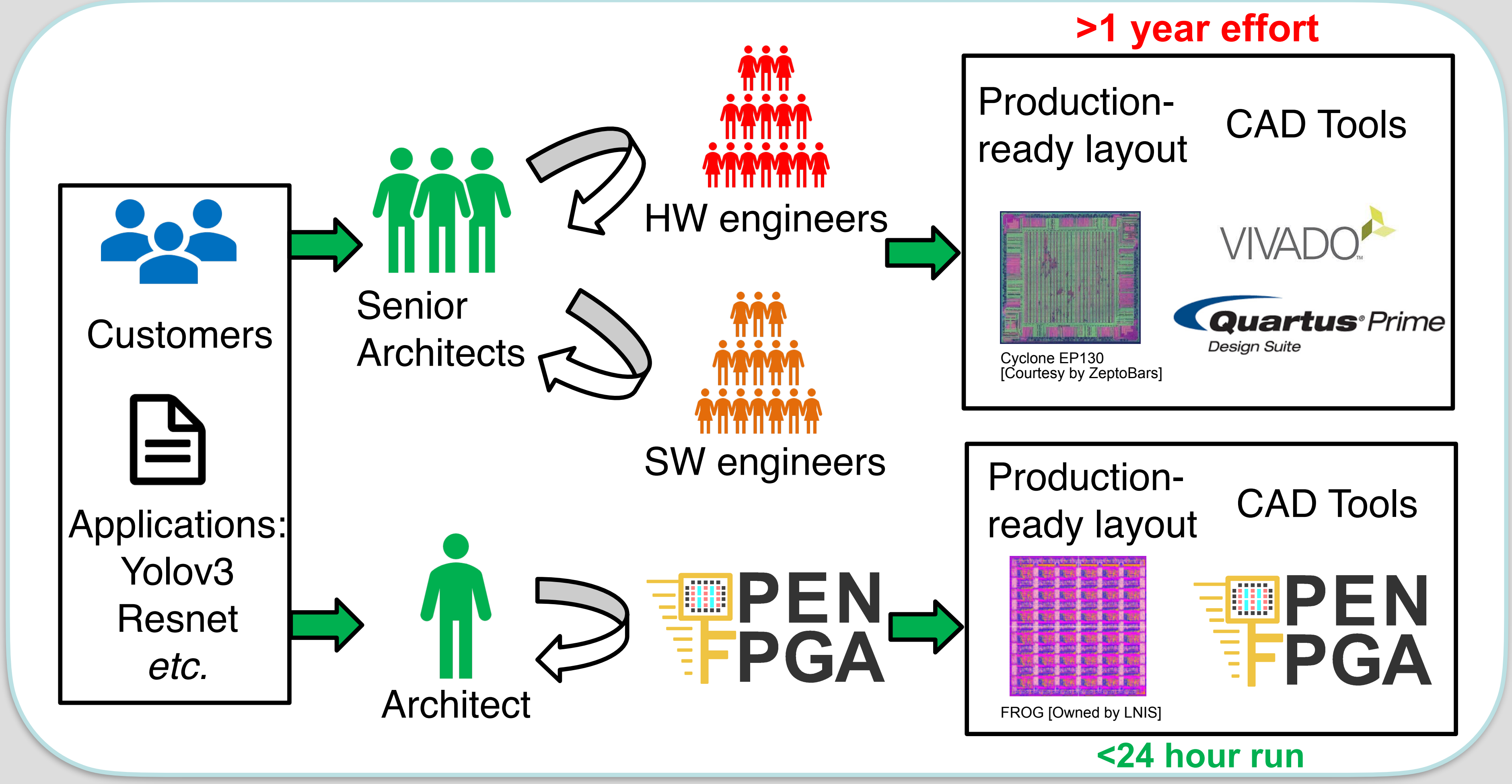
<24 hour run

OpenFPGA - In a nutshell



✓ Complete FPGA and eFPGA generation (10+ commercial and academic tape-outs)

OpenFPGA - In a nutshell



>1 year effort

Customers

Applications:
Yolov3
Resnet
etc.

Senior
Architects

HW engineers

SW engineers

Production-ready layout

CAD Tools

Cyclone EP130
[Courtesy by ZeptoBars]

VIVADO

Quartus Prime
Design Suite

Architect

OPENFPGA

Production-ready layout

CAD Tools

FROG [Owned by LNIS]

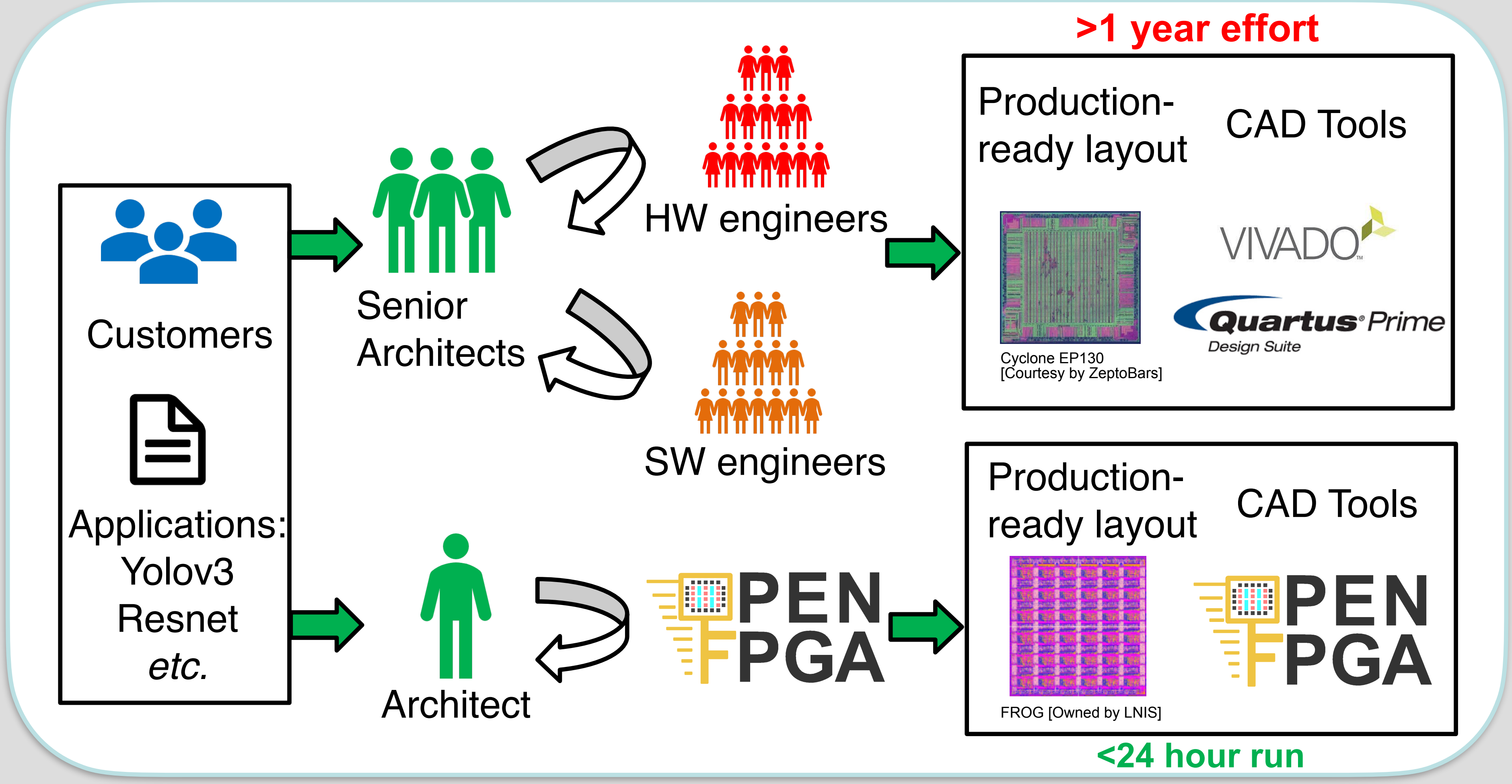
OPENFPGA

<24 hour run

✓ Complete FPGA and eFPGA generation (10+ commercial and academic tape-outs)

✓ Fully customizable modern architecture (100+ tested)

OpenFPGA - In a nutshell



Customers

Applications:
Yolov3
Resnet
etc.

Senior
Architects

HW engineers

SW engineers

Architect

>1 year effort

Production-
ready layout

CAD Tools

Cyclone EP130
[Courtesy by ZeptoBars]

Production-
ready layout

CAD Tools

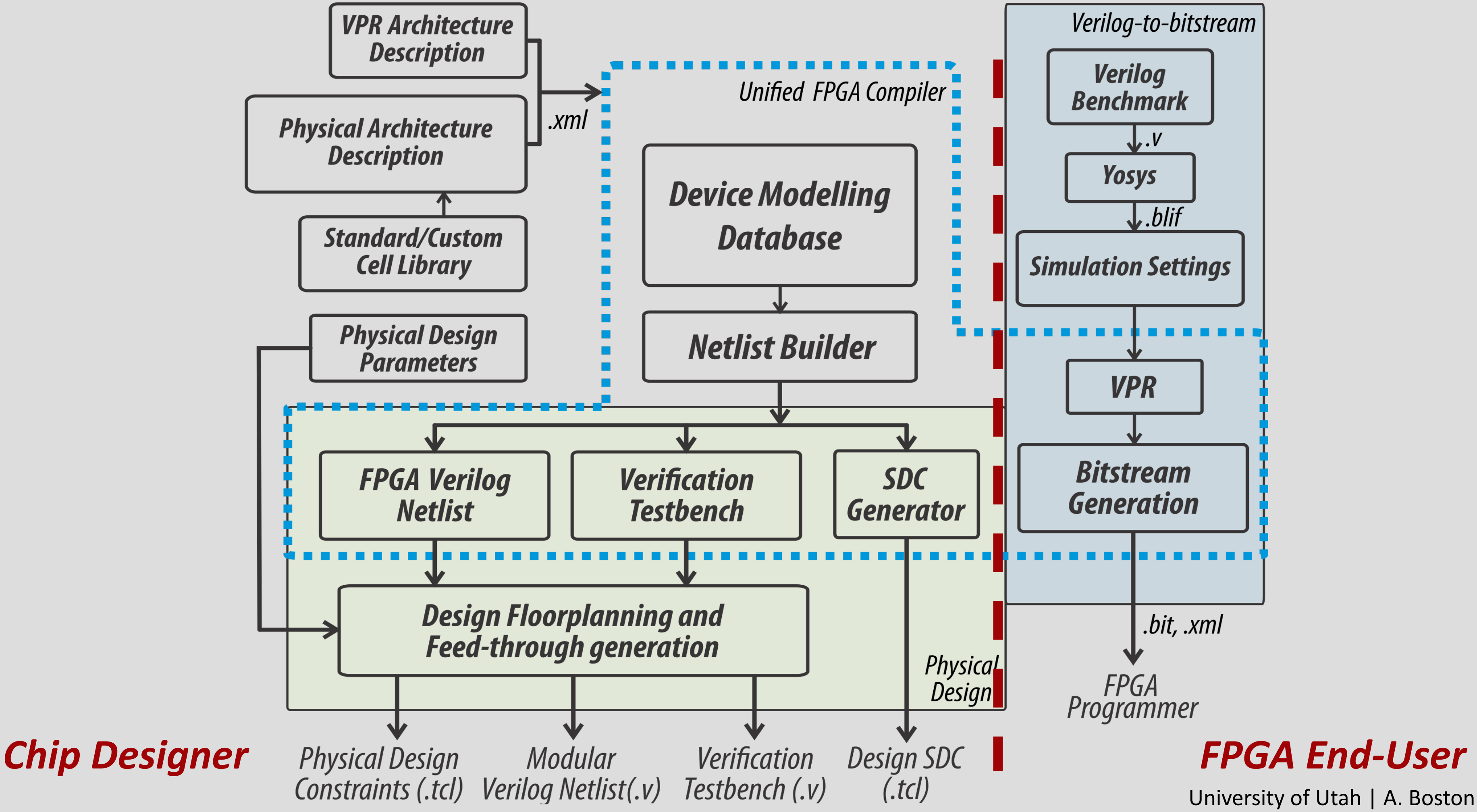
FROG [Owned by LNIS]

<24 hour run

- ✓ Complete FPGA and eFPGA generation (10+ commercial and academic tape-outs)
- ✓ Fully customizable modern architecture (100+ tested)
- ✓ Optimized for fast physical design (150k-LUT+ FPGA < 24 hr)

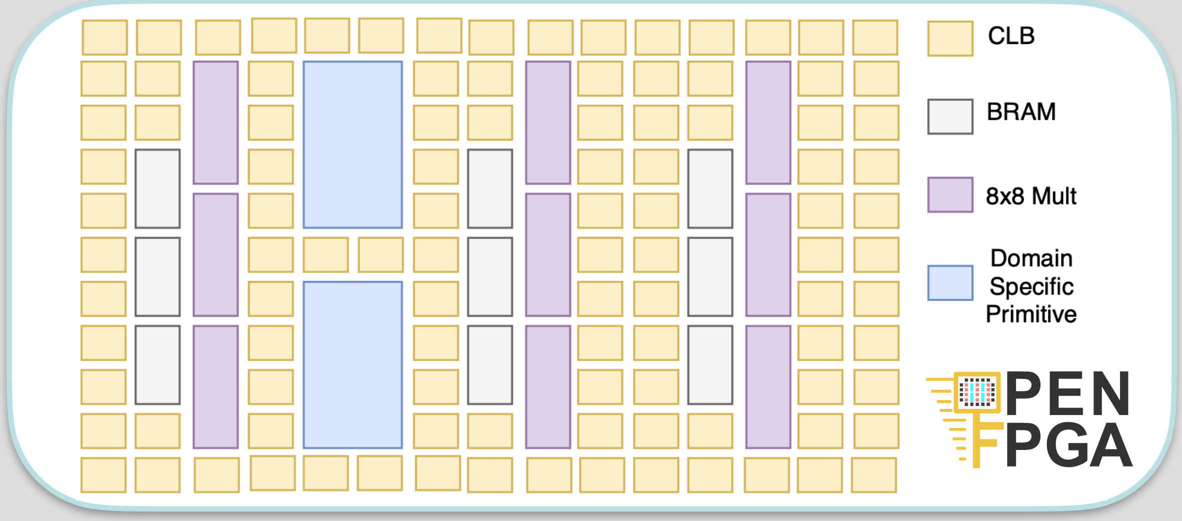
OpenFPGA - In a nutshell

- Unified code base for fabric generation, design verification and end-user bitstream generation


VPR Architecture Description
Physical Architecture Description
.xml
Standard/Custom Cell Library
Physical Design Parameters
Unified FPGA Compiler
Device Modelling Database
Netlist Builder
FPGA Verilog Netlist
Verification Testbench
SDC Generator
Design Floorplanning and Feed-through generation
Physical Design
Verilog-to-bitstream
Verilog Benchmark
.v
Yosys
.blif
Simulation Settings
VPR
Bitstream Generation
.bit, .xml
FPGA Programmer
Chip Designer
Physical Design Constraints (.tcl)
Modular Verilog Netlist(.v)
Verification Testbench (.v)
Design SDC (.tcl)
FPGA End-User

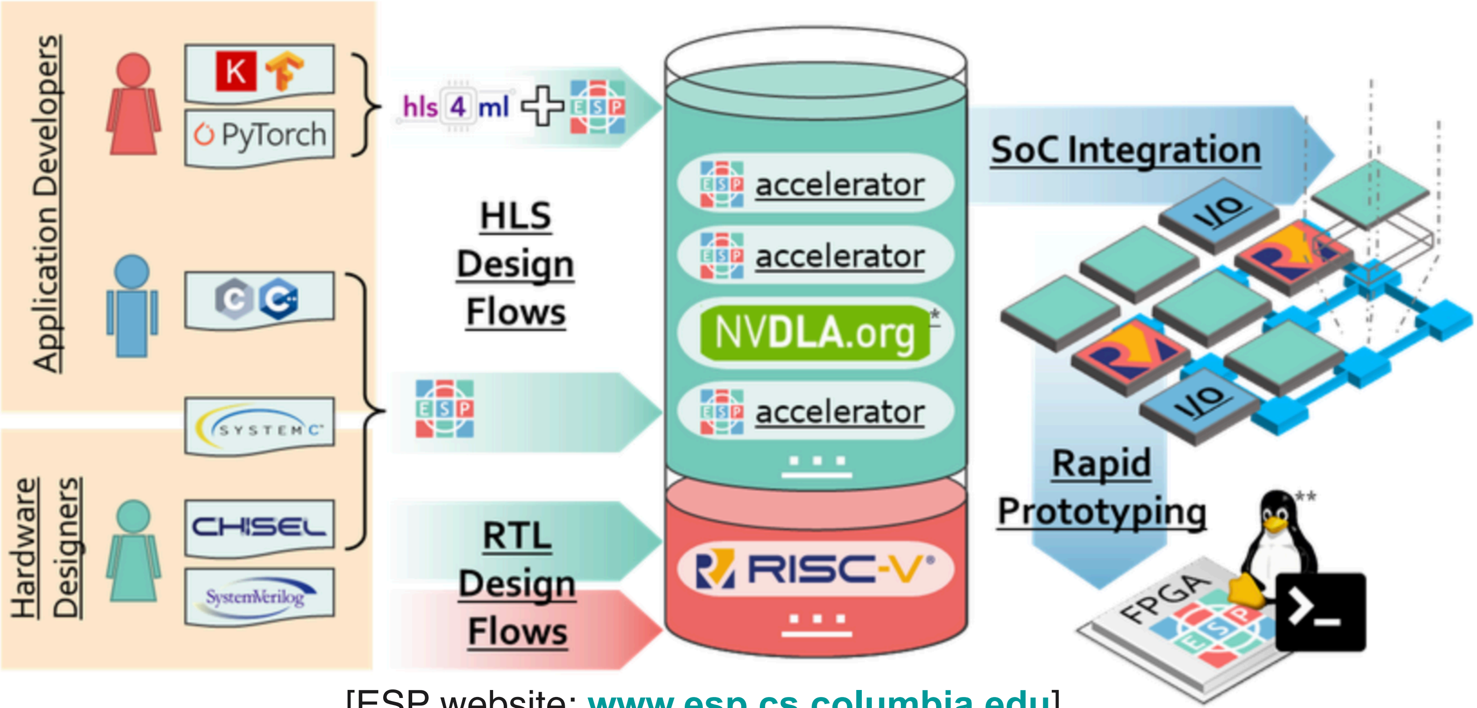
Domain-Specific eFPGAs

- OpenFPGA enables DSE and prototyping of highly customizable eFPGA fabrics
  - Best fit fabric architecture - k6_n8 vs. k4_n8
  - Appropriately sized fabric - 16×16 vs. 32×32
  - Balanced compute, memory, and routing resources
  - Domain-specific primitives to improve PPA
  - <u>Optimizing eFPGA lowers PPA trade-off for post-silicon reconfigurability</u>



SoC integration with ESP

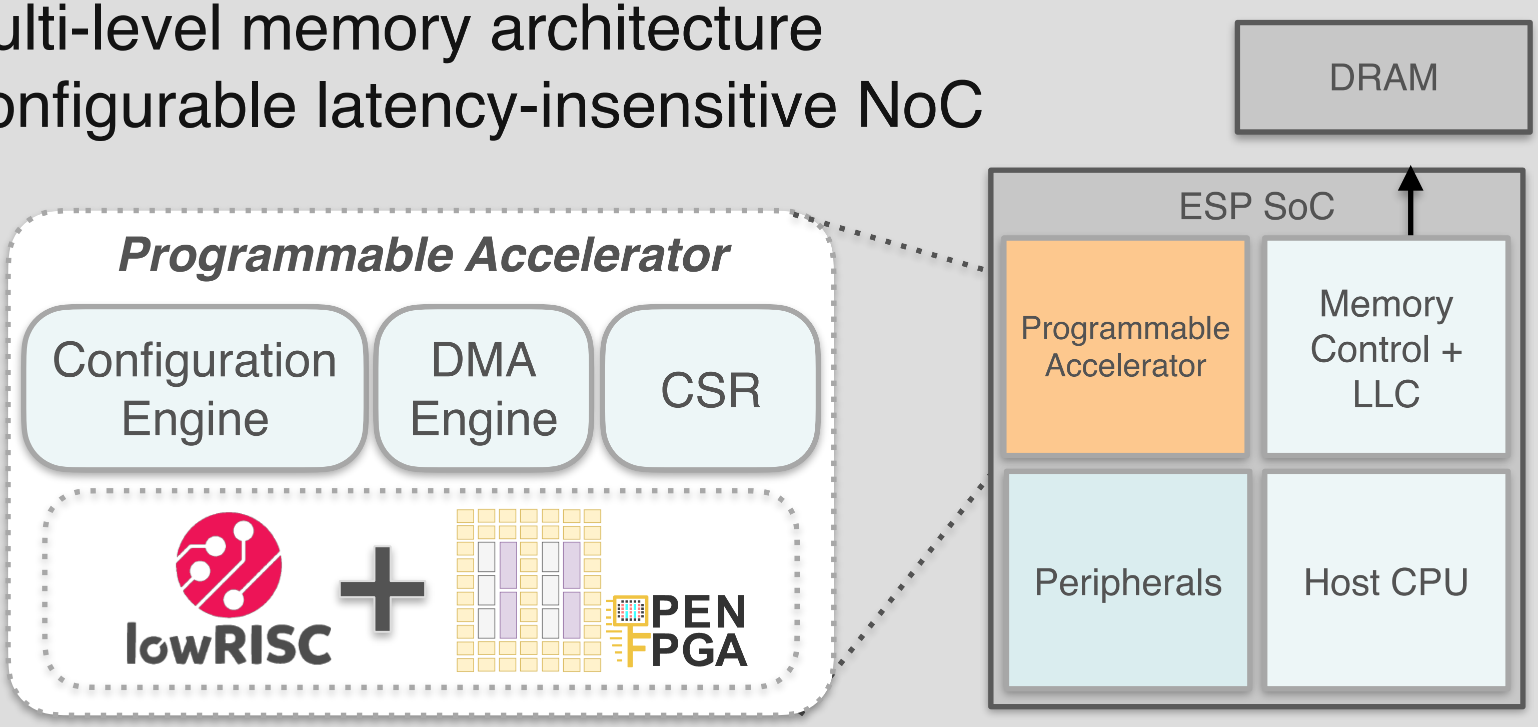
- ESP provides automated design flow for SoC development
- Supports different accelerator designs
- Push-button SoC generation and application mapping
- Rapid FPGA prototyping



[ESP website: www.esp.cs.columbia.edu]

2×2 SoC Architecture

- Programmable Accelerator Tile
  - RISC-V + eFPGA
- Host CPU
  - Manages SoC
  - Invokes accelerators
- Multi-level memory architecture
- Configurable latency-insensitive NoC



Conclusion

Contributions:

- Tightly-coupled RISC-V + eFPGA programmable accelerator
- Integration into ESP-based heterogeneous SoC
- Acceleration of TensorFlow Lite workloads

- Approach combines aspects of user-friendly general purpose compute with ASIC performance and parallelization
- TensorFlow Lite → Heterogeneous SoC
- Toolchain enables HW DSE and PPA evaluation for RISC-V ML ISA extension architectures

Conclusion

Contributions:

- Tightly-coupled RISC-V + eFPGA programmable accelerator
- Integration into ESP-based heterogeneous SoC
- Acceleration of TensorFlow Lite workloads

- Approach combines aspects of user-friendly general purpose compute with ASIC performance and parallelization
- TensorFlow Lite → Heterogeneous SoC
- Toolchain enables HW DSE and PPA evaluation for RISC-V ML ISA extension architectures

- *low power* - improved HW PPA compared to fully “soft-core”
- *flexible* - maintains post-silicon reconfigurability of CCU
- *low development effort* - TensorFlow Lite interoperability

Future Work: eFPGA Evaluation

- **DSE of eFPGA based on ISA extensions**
  - Determine the best fit eFPGA architecture
  - Appropriately size fabric to reduce overhead
  - Identify domain-specific primitives to improve PPA

- **Compare RISC-V coupling strategies**
  - Should the RISC-V hard-block be embedded directly in the eFPGA fabric or located outside of the eFPGA fabric?

- **Demonstrate the advantages of hard-block RISC-V**
  - Compare soft-core RISC-V on baseline FPGA architecture to a hard-block implementation in terms of power, area, and operating frequency

Future Work: System Level Evaluation

- **Execute TensorFlow Lite Models on proposed architecture**
  - Compare standalone RISC-V to proposed eFPGA enhanced architecture
    - Execute ML models with and without ISA extensions
    - Is the slow down of eFPGA worth the parallelization across all TensorFlow Lite benchmarks?
- **Show flexibility of eFPGA to accomodate several varying ML models**
  - Perform inference of varying ML tasks using the same architecture
- **Leverage open-source tool flow to evaluate hardware, *i.e.*, area, power, energy per inference**
  - SOTA works often evaluate only speedup of ML models with ISA extensions

Thank you for your attention

Questions?

allen.boston@utah.edu

THE UNIVERSITY OF UTAH®

LNIS

Laboratory for NanoIntegrated Systems

Laboratory for NanoIntegrated Systems
Department of Electrical and Computer Engineering
SMBB Building – University of Utah – Salt Lake City – UT – USA