Ariane + NVDLA

Seamless Third-Party IP Integration with ESP

Davide Giri
Kuan-Lin Chiu
Guy Eichler
Paolo Mantovani
Nandhini Chandramoorthy (IBM Research)
Luca P. Carloni

CARRV 2020

Columbia University
IN THE CITY OF NEW YORK

CS@CU Computer Science

Motivation

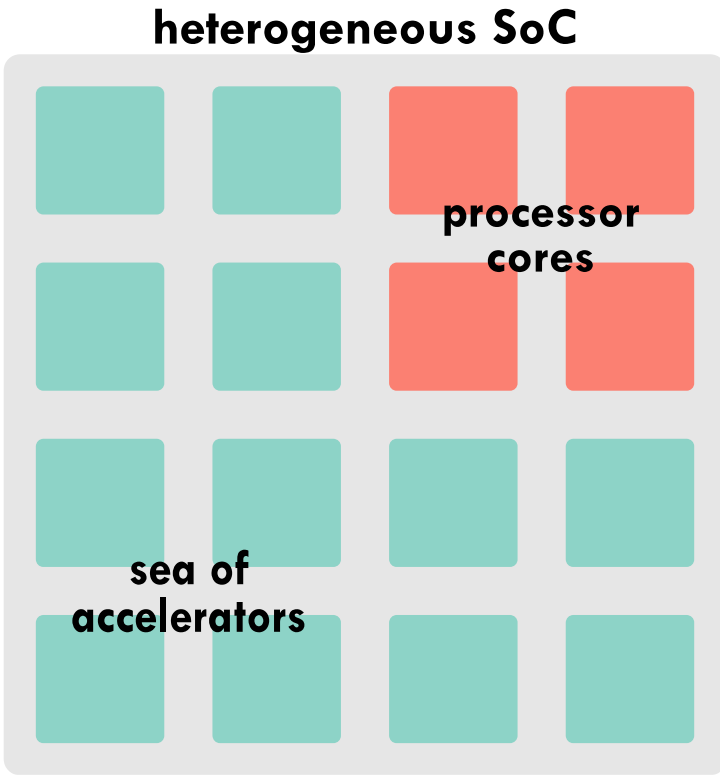
- SoCs are increasingly heterogeneous [1]
- Heterogeneity increases the engineering effort [2]

→ IP reuse enables the design of complex SoCs

- Thanks to open-source hardware (OSH) movement [3]

→ Proliferation of open-source IPs

Seamless third-party IP integration is key!



[1] Shao, SLCA'15 [2] Khailani, DAC'18 [3] Gupta, IEEE Computer'17

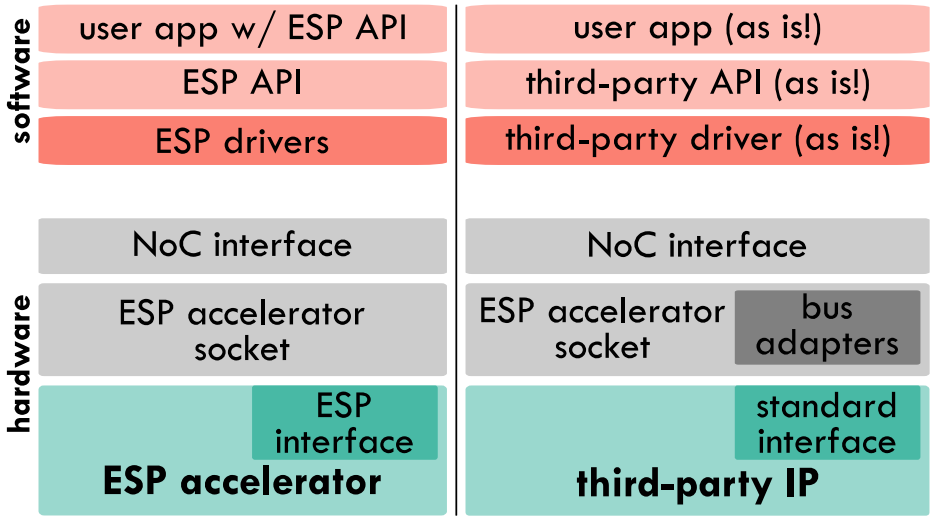
In this work

Enhance ESP with support for third-party accelerators

- ESP is our open-source platform for SoC design [4]

Demonstrate integration capabilities of ESP

- Integration of Ariane [5] and NVDLA [6]
- Rapid FPGA prototyping



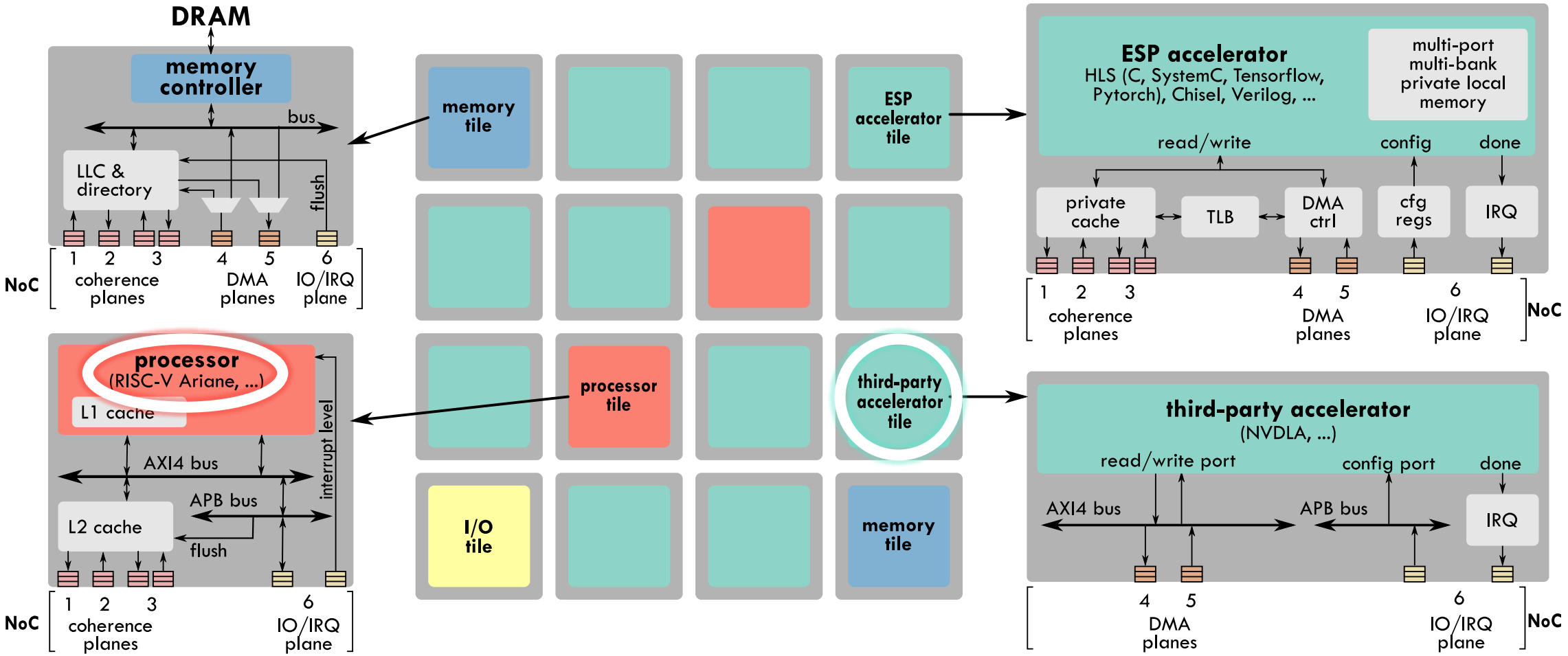
Open-source release as part of ESP

- Hands-on tutorial: esp.cs.columbia.edu/docs/thirdparty_acc

[4] ESP: esp.cs.columbia.edu [5] Ariane: github.com/pulp-platform/ariane [6] NVDLA: nvdla.org

ESP overview

ESP architecture



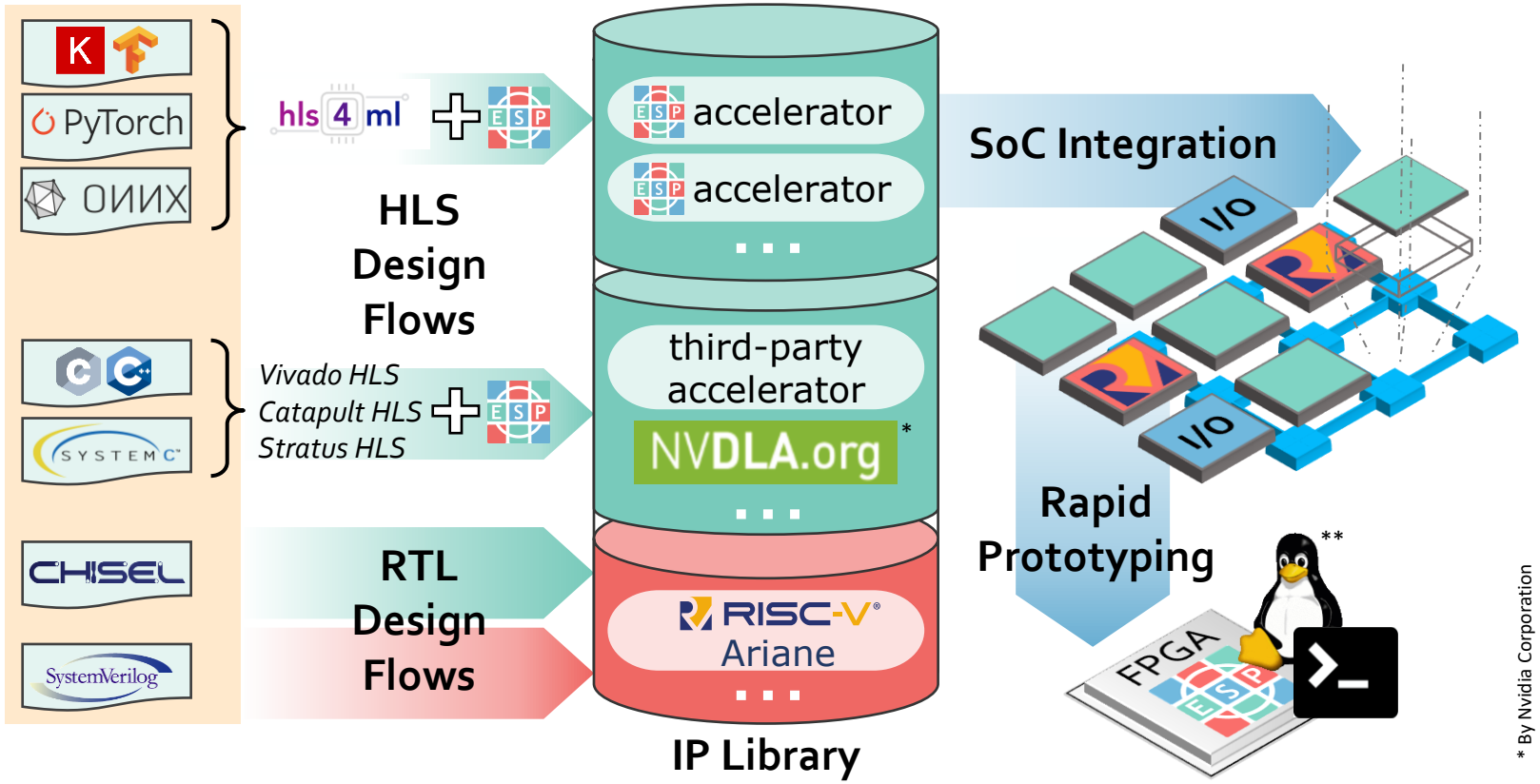
ESP methodology

Accelerator Flow

- Simplified design
- Automated integration

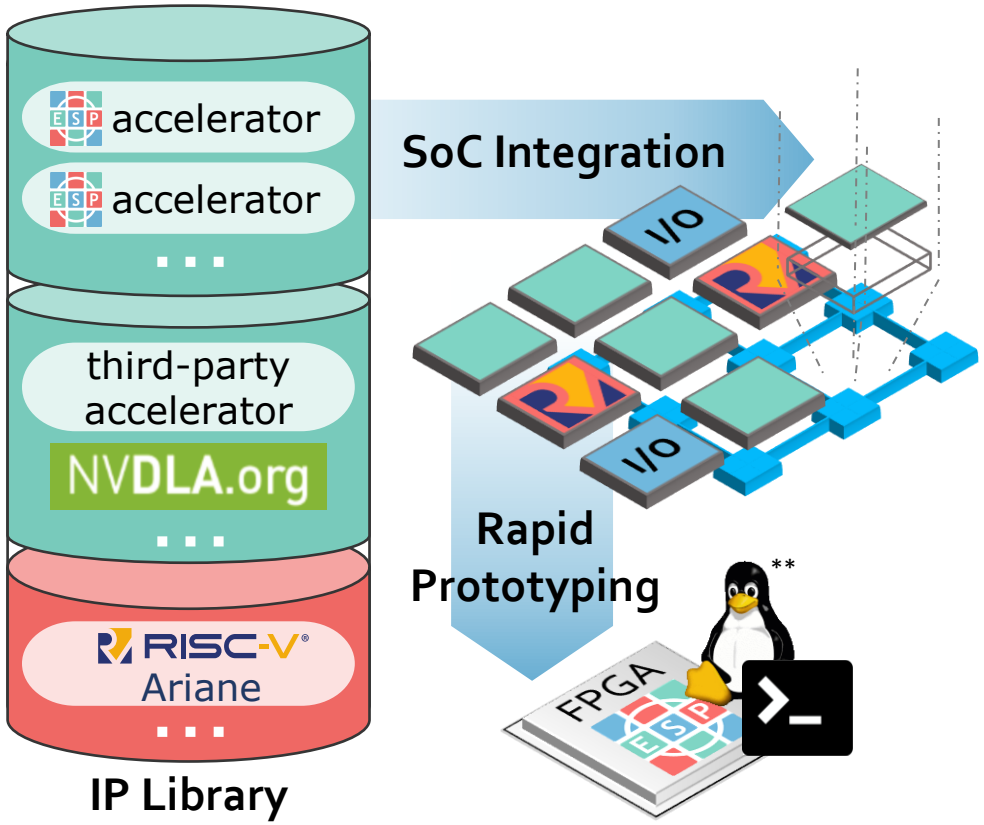
SoC Flow

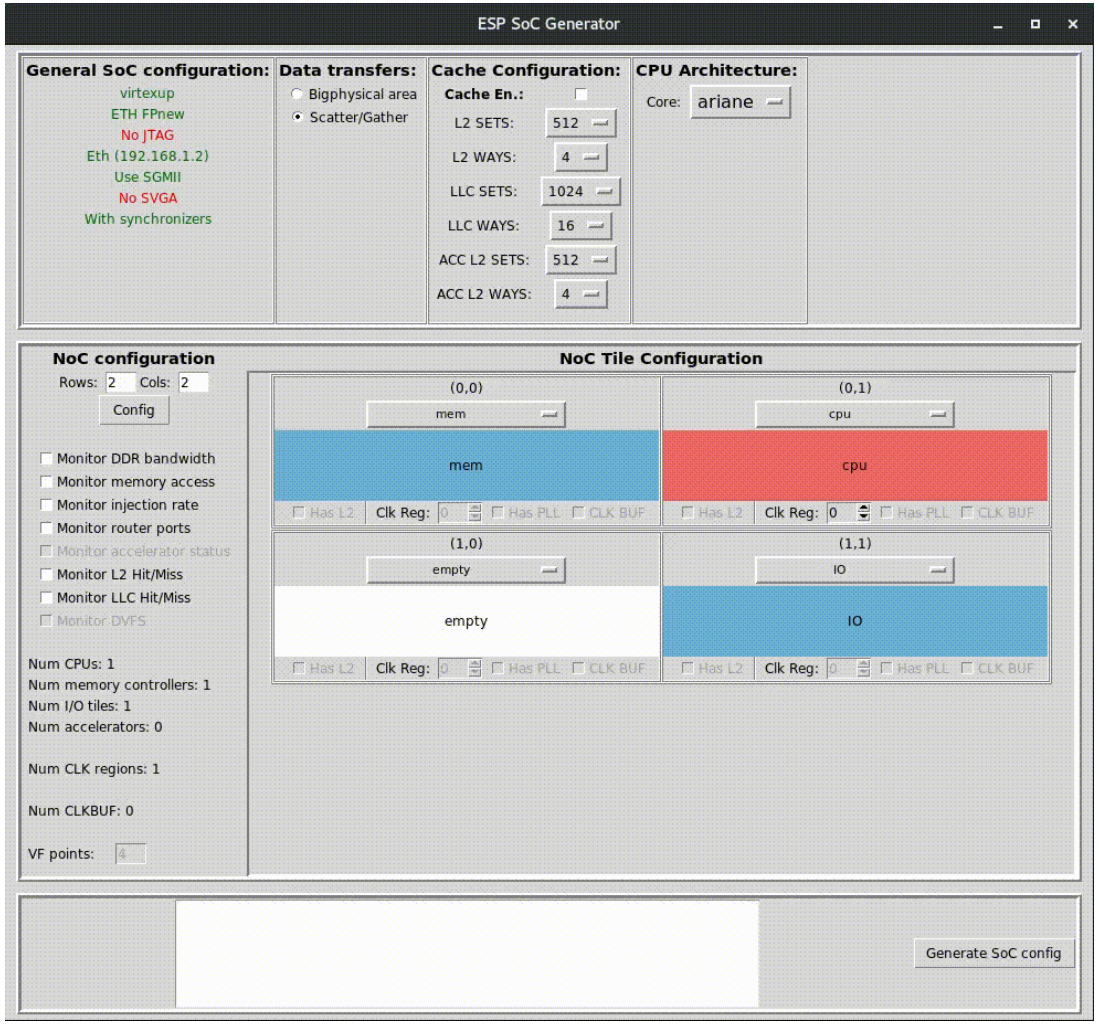
- Mix&match floorplanning GUI
- Rapid FPGA prototyping



* By Nvidia Corporation
** By lewing@isc.tamu.edu Larry Ewing and The GIMP

ESP methodology: SoC flow

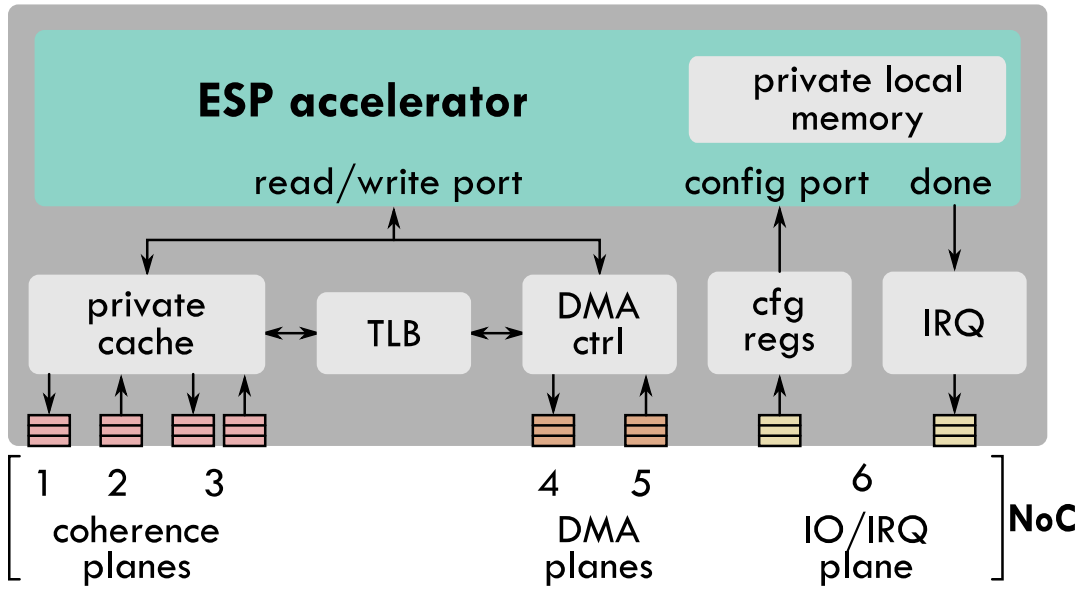


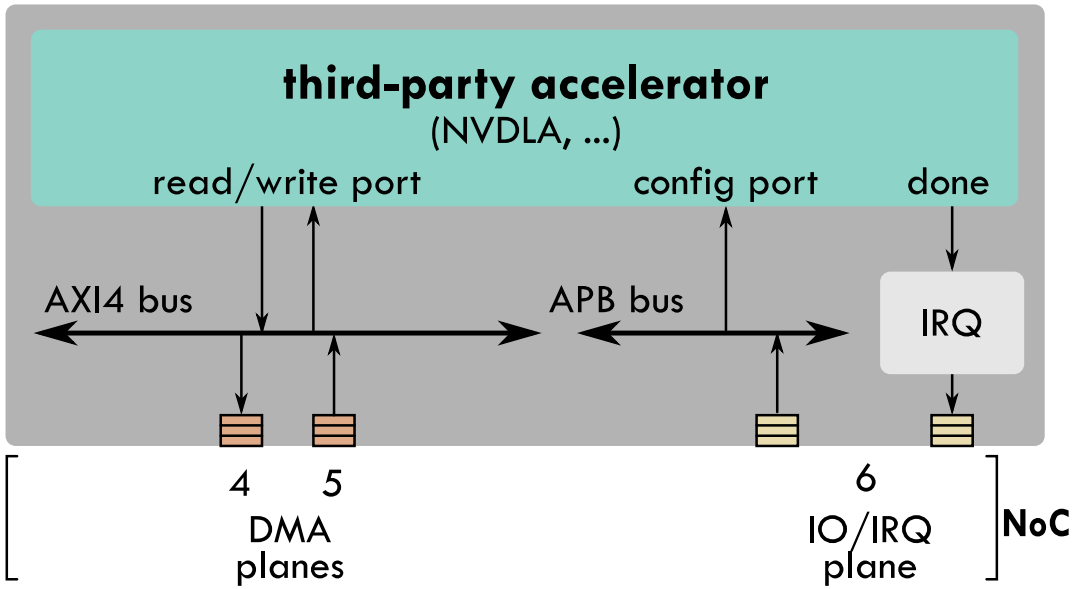


** By lewing@isc.tamu.edu Larry Ewing and The GIMP

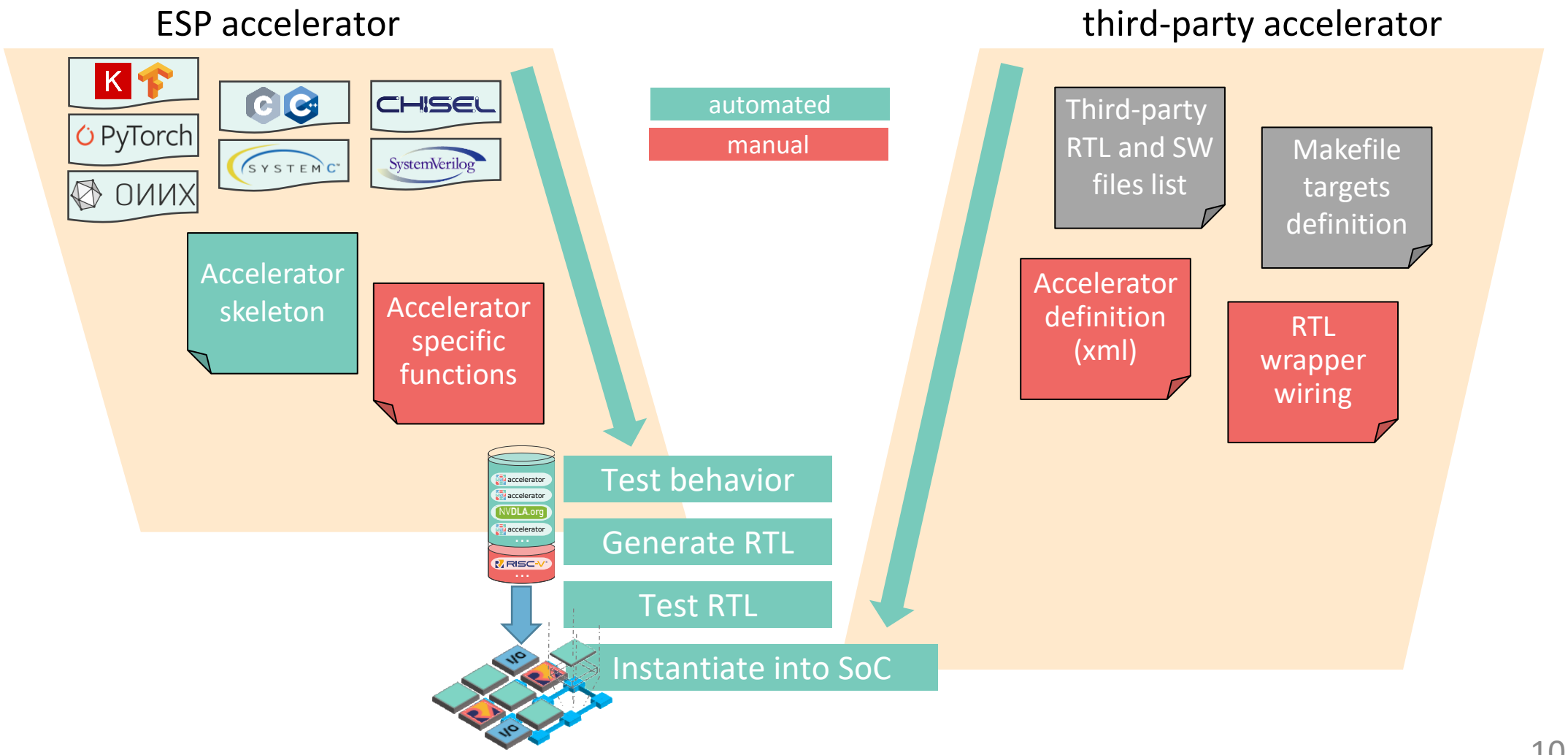
Third-party IP integration with ESP

ESP accelerator tile





ESP accelerator flow

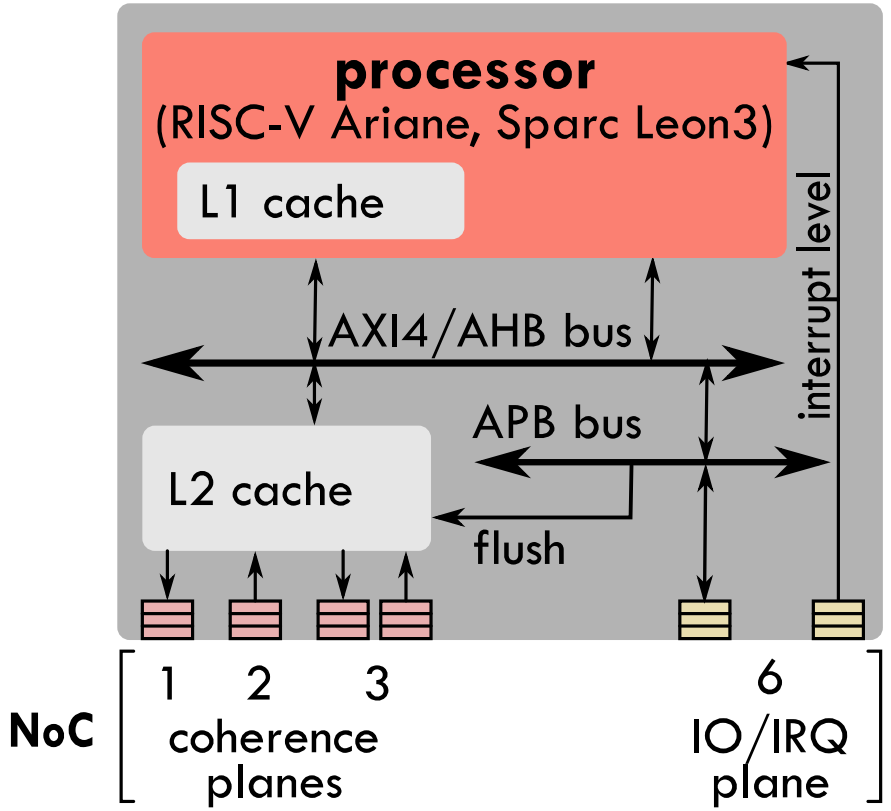


Ariane + NVDLA with ESP

Integration of Ariane

ESP processor tile

- RISC-V Ariane (new!) or Sparc-v8 Leon3
- Boot unmodified Linux
- AXI4 (new!) or AHB bus to access memory
- APB bus to access peripherals
- Optional L2 private cache
- Processor-specific interrupt controller placed in the I/O tile



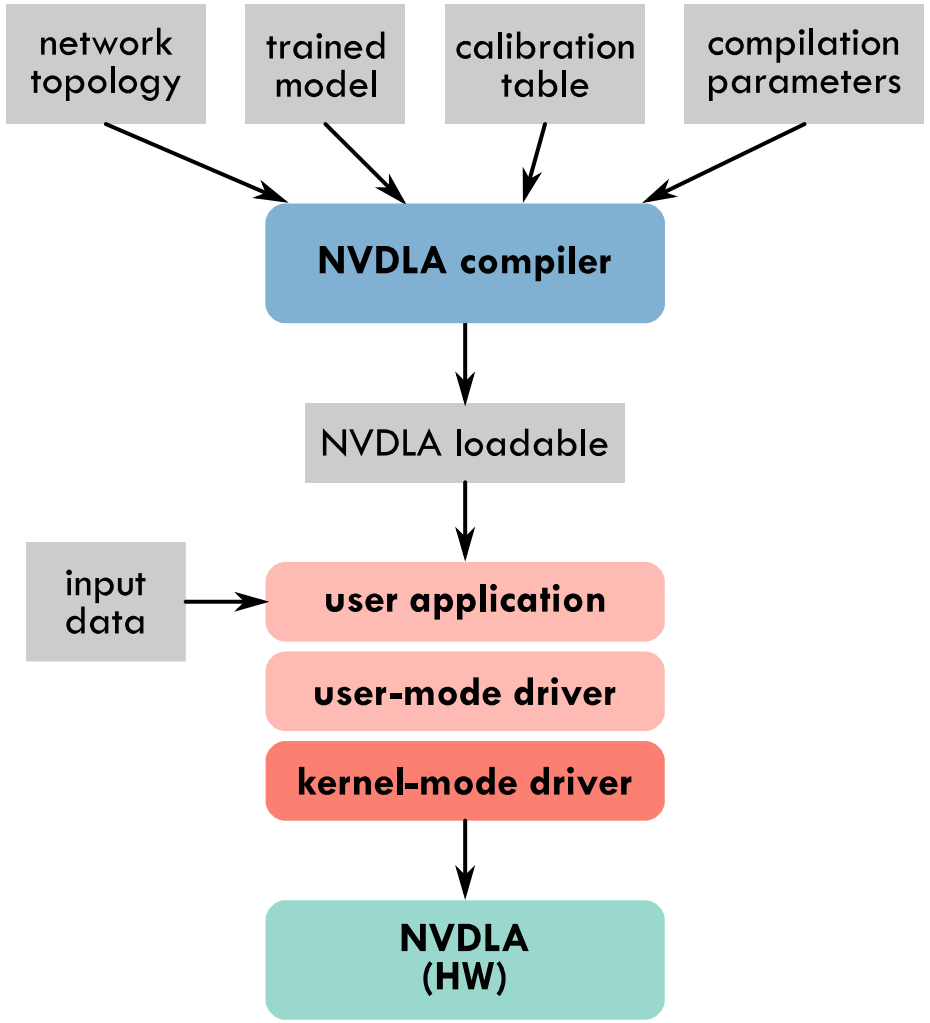
NVDLA

NVIDIA Deep Learning Accelerator

- Open source
- Fixed function
- Highly configurable

NVDLA *small*

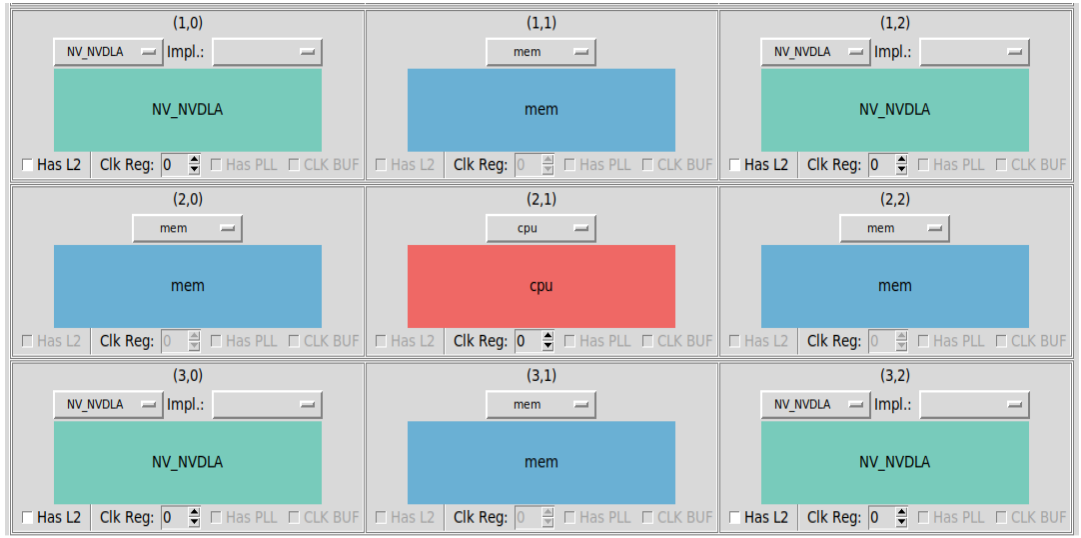
- 8-bit integer precision
- 64 MAC units
- 128 KB local memory



Evaluation: setup

SoCs evaluated on FPGA (Xilinx XCVU440)

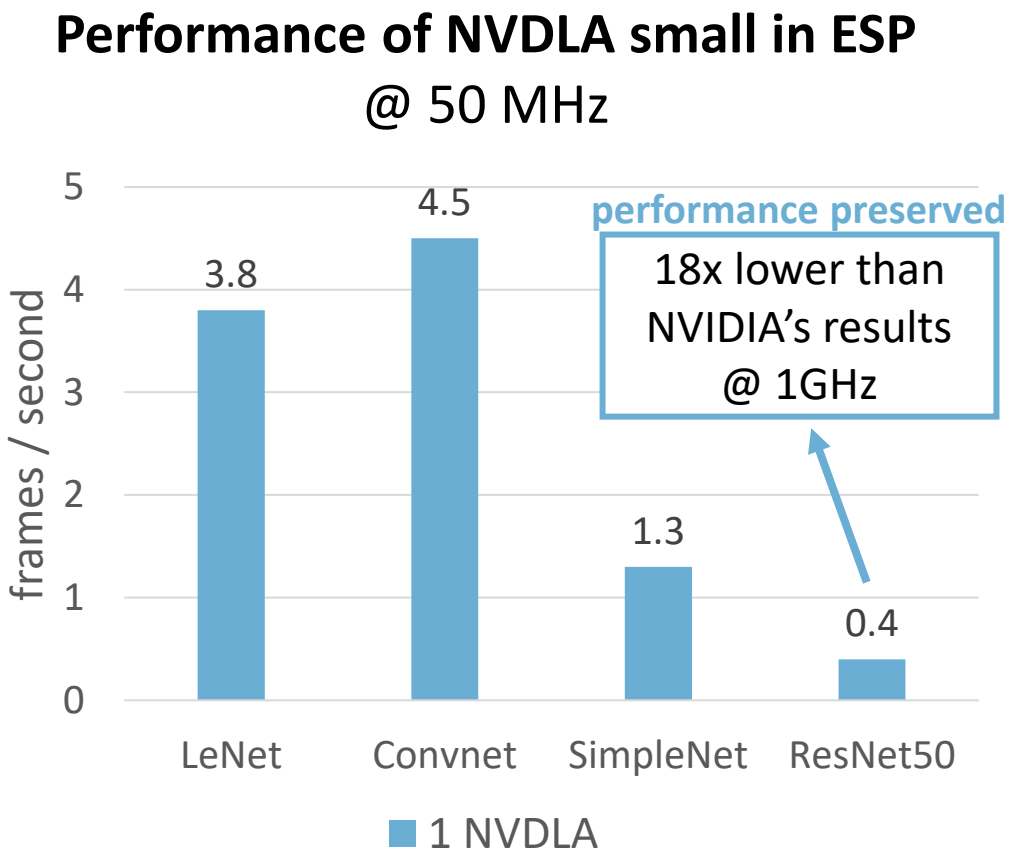
- Ariane core
- 1-4 NVDLA tiles
- 1-4 memory channels

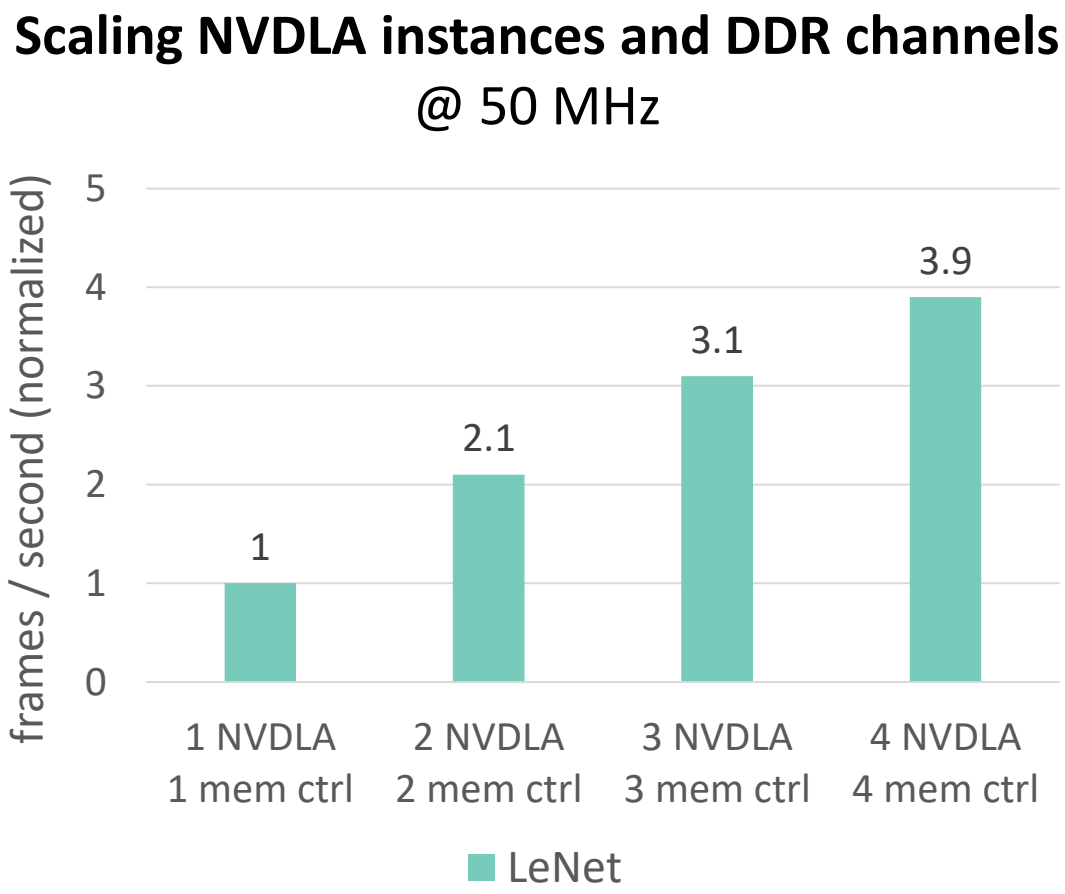


Evaluation networks

Model	Dataset	Layers	Input	Model Size
LeNet	MNIST	9	1x28x28	1.7MB
Convnet	CIFAR10	13	3x32x32	572KB
SimpleNet	MNIST	44	1x28x28	21MB
ResNet50	ILSVRC2012	229	3x224x224	98MB

Evaluation: results





Thank you from the ESP team!

sld.cs.columbia.edu


ColumbiaSld

esp.cs.columbia.edu


sld-columbia/esp


ESP channel

Ariane + NVDLA

Seamless Third-Party IP Integration with **ESP**

Davide Giri
Kuan-lin Chiu
Guy Eichler
Paolo Mantovani
Nandhini Chandramoorthy (IBM)
Luca P. Carloni

CARRV 2020