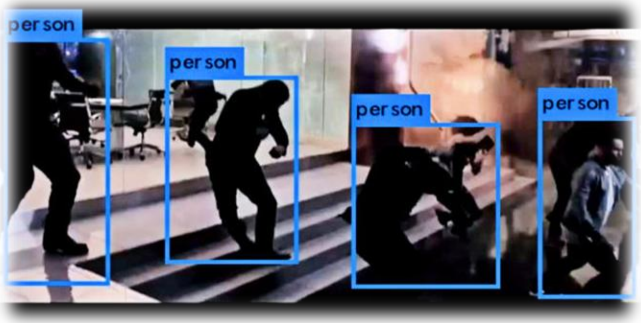
UNIMORE
UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA

Convolutional Neural Networks on embedded automotive platforms: a qualitative comparison

Paolo Burgio
paolo.burgio@unimore.it

Convolutional Neural Networks

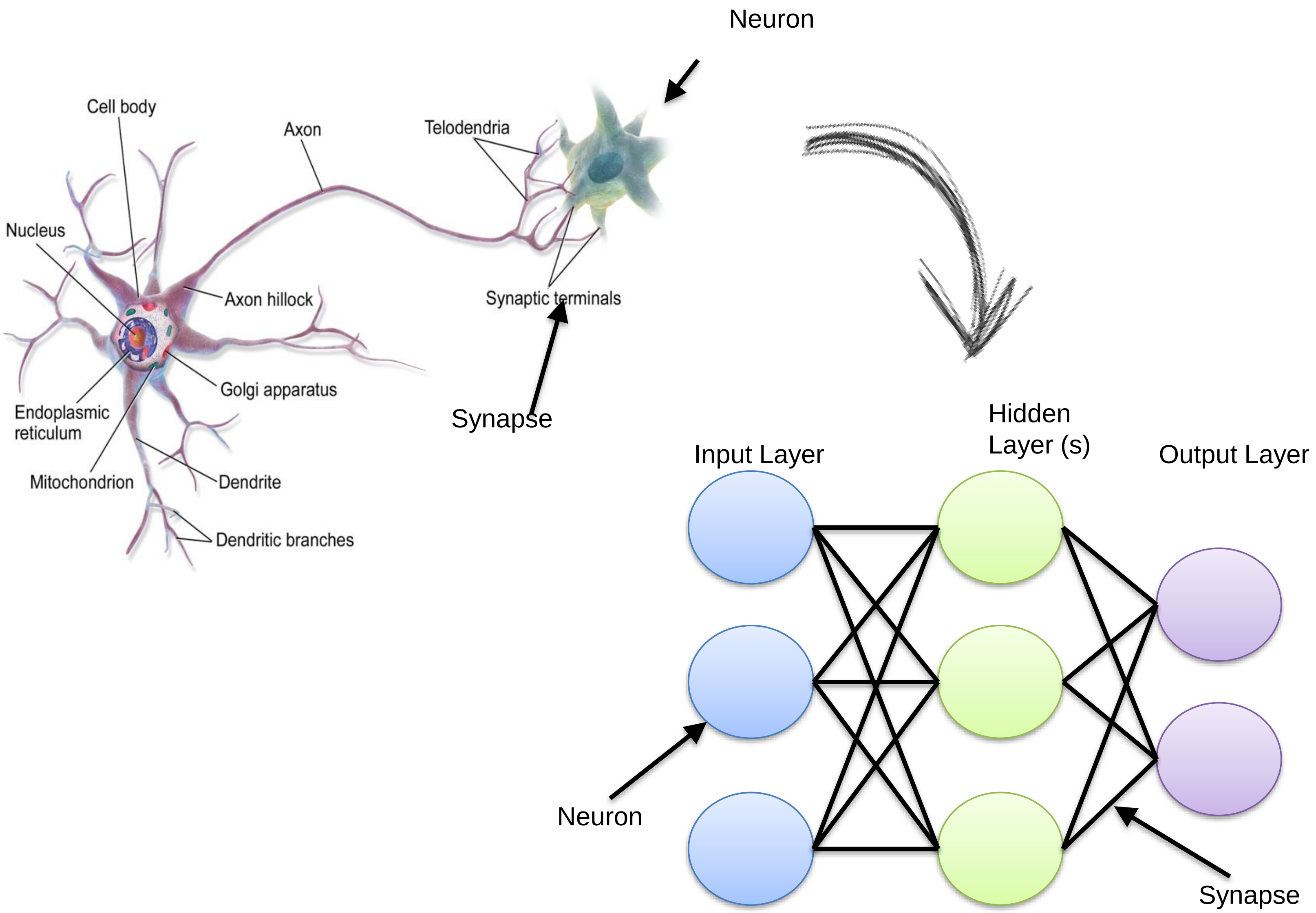
✓ Extensively adopted in the ~~embedded~~ world



✓ Computer vision and image processing tasks,
  - object categorization and labeling

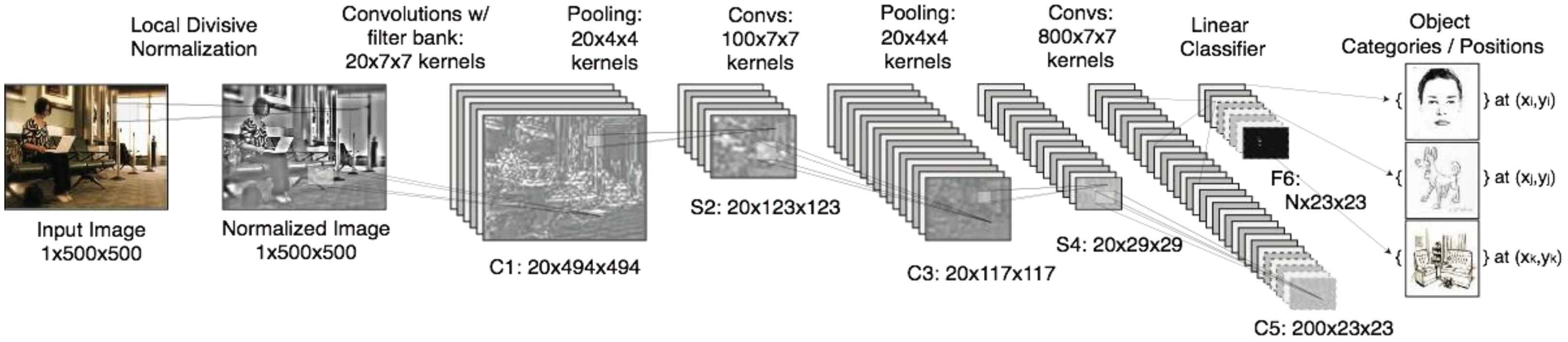
✓ Autonomous driving, industry 4.0.

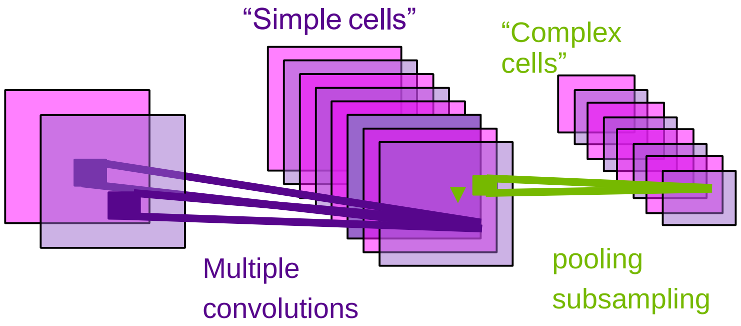
Neural Networks



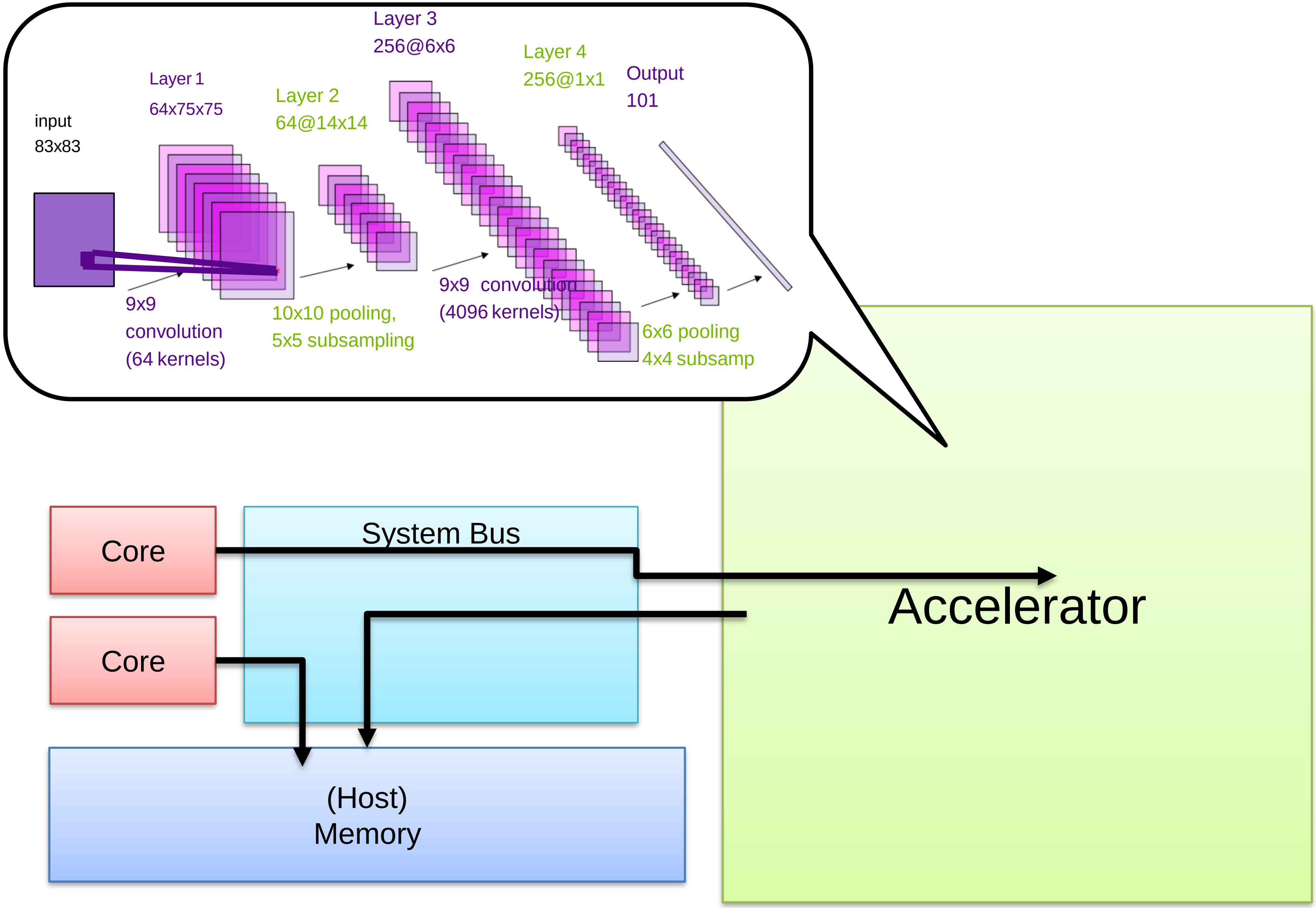
Convolutional Neural Networks

- ✓ Multiple bidimensional layers
- ✓ Huge number of multiply-accumulate (MAC) operation
- ✓ on thousands of pixel of an input image





CNN on highly parallel architectures



input
83x83

Layer 1
64x75x75

9x9
convolution
(64 kernels)

Layer 2
64@14x14

10x10 pooling,
5x5 subsampling

Layer 3
256@6x6

9x9 convolution
(4096 kernels)

Layer 4
256@1x1

6x6 pooling
4x4 subsamp

Output
101

Core

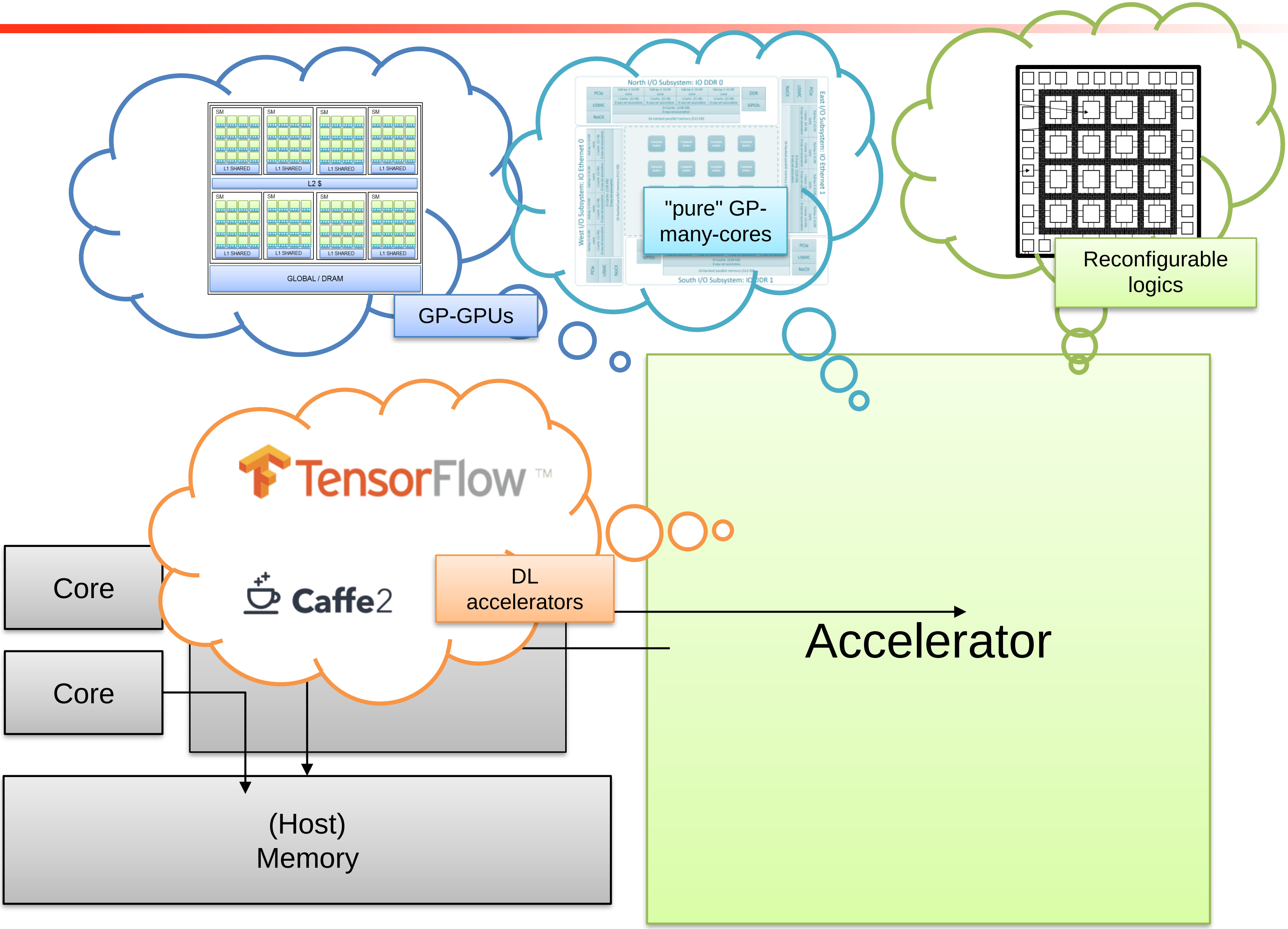
Core

System Bus

(Host)
Memory

Accelerator

…which architecture?


SM
L1 SHARED
L2 $
GLOBAL / DRAM
North I/O Subsystem: IO DDR 0
South I/O Subsystem: IO DDR 1
West I/O Subsystem: IO Ethernet 0
East I/O Subsystem: IO Ethernet 1
TensorFlow
Caffe2

GP-GPUs

"pure" GP-many-cores

Reconfigurable logics

DL accelerators

Core

Core

(Host) Memory

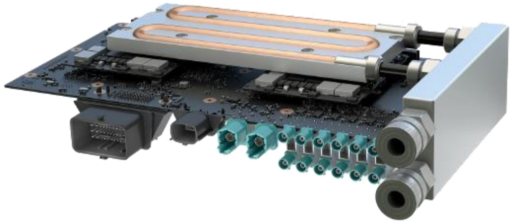
Accelerator

Our aim

Assessing the performance
of representative CNN packages
on state-of-the-art ADAS platforms

- ✓ E2E latency (time per image)
- ✓ Throughput (images per time)
- ✓ Power consumption (Watt)

Current (ongoing) work

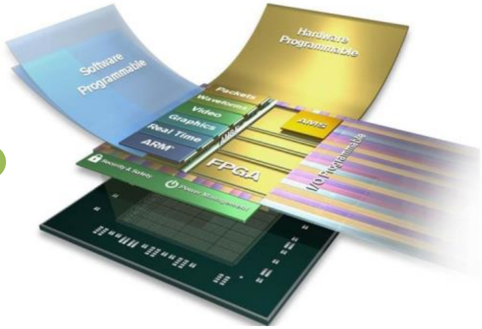


Nvidia Parker SoC

- ✓ Drive PX2 for autonomous driving
- ✓ 4 x ARM Cortex 57 + 2 x Denver
- ✓ Pascal GPU

Xilinx Zynq Ultrascale+

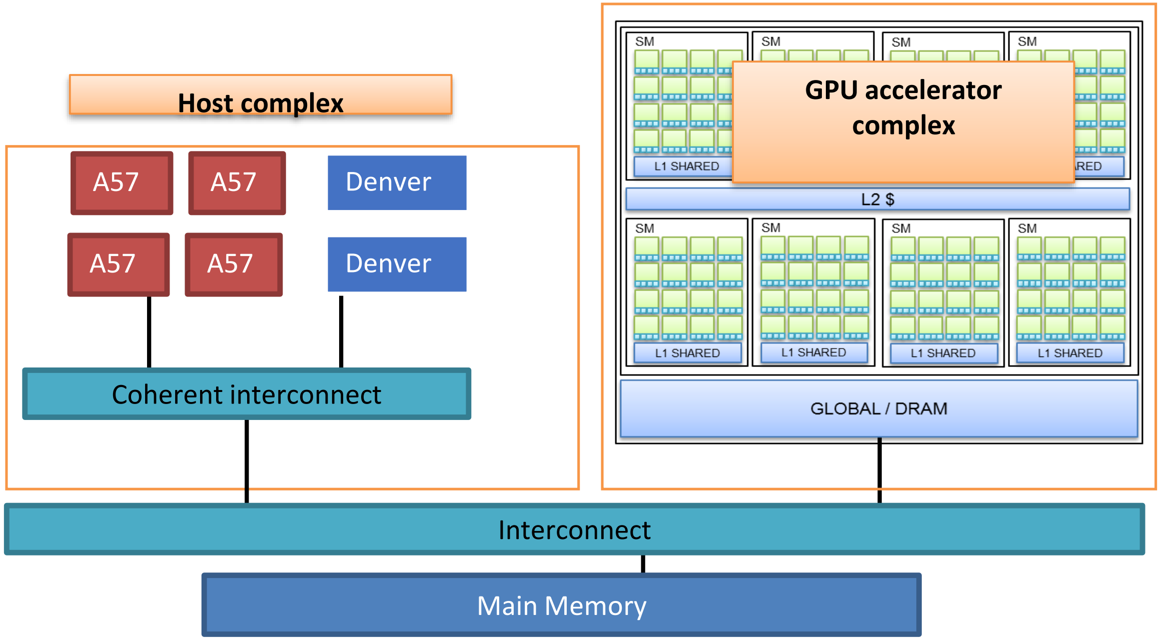
- ✓ 4 x ARM Cortex A53 + 2 x R5
- ✓ Mali GPU
- ✓ FPGA fabric



NVIDIA Tegra X2

Embedded computing platform for the automotive market

- ✓ Esa-core host with Big.SUPER configuration
  - – 4 ARM A57 + 2 Denver
- ✓ Pascal GPU with 2 NVIDIA Streaming Multiprocessors (SM)
  - – 256 CUDA cores. The
- ✓ 1 TFLOP of computing power, @20 Watts.
- ✓ Qualified according to Functional Safety and Road Vehicles Standard
  - – ISO 26262's ASIL-B level)
  - – marketed as Drive PX2



Host complex

A57

A57

Denver

A57

A57

Denver

Coherent interconnect

GPU accelerator complex

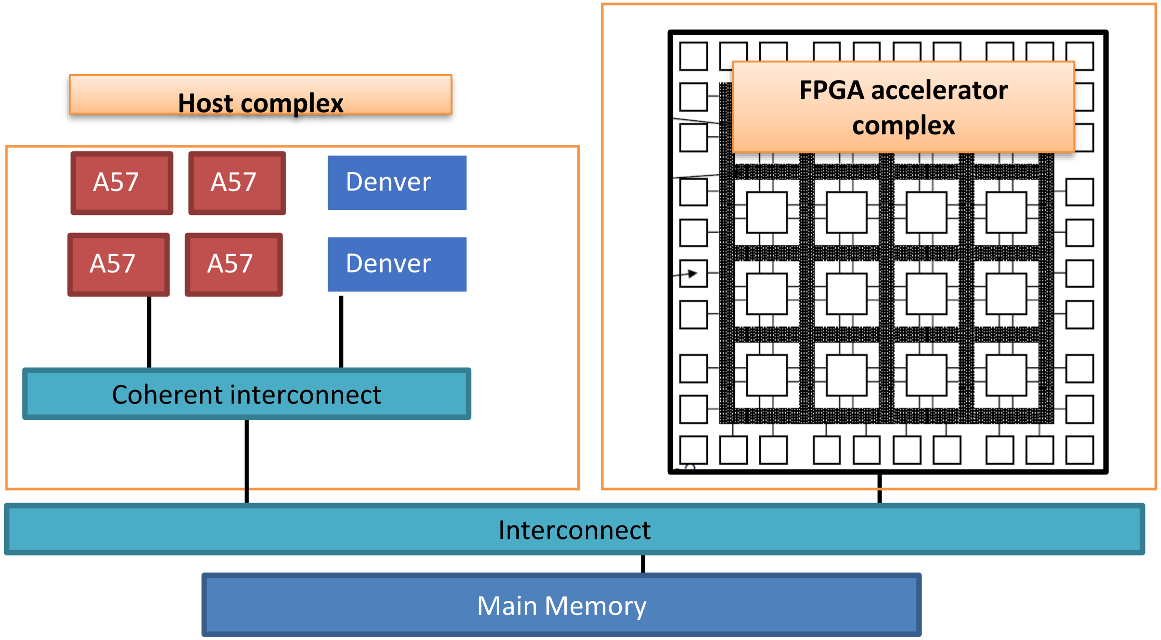
Interconnect

Main Memory

Xilinx Ultrascale+

Embedded computing platform for the automotive market

- ✓ Esa-core host
  - 4 x ARM Cortex A53 + 2 x R5
- ~~✓ Mali GPU~~
  - Poor programmability (not a GP-GPU)
- ✓ Programmable FPGA fabric



DISCLAIMER

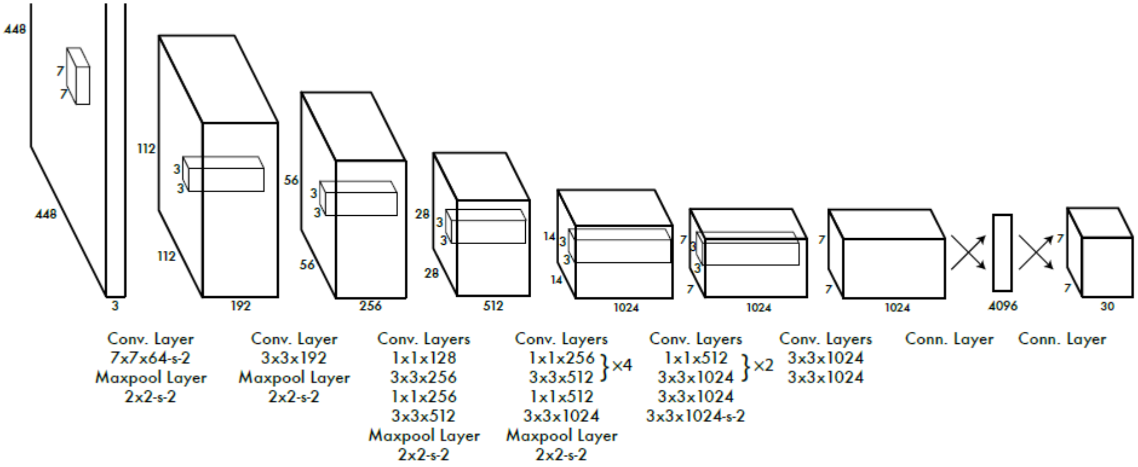
Currently targeted networks

- ✓ You-only-look-once (Yolo)
  - GPU ✓
  - FPGA (Darknet) ✗
- ✓ ZynqNet
  - GPU ✓
  - FPGA ✓

Yolo

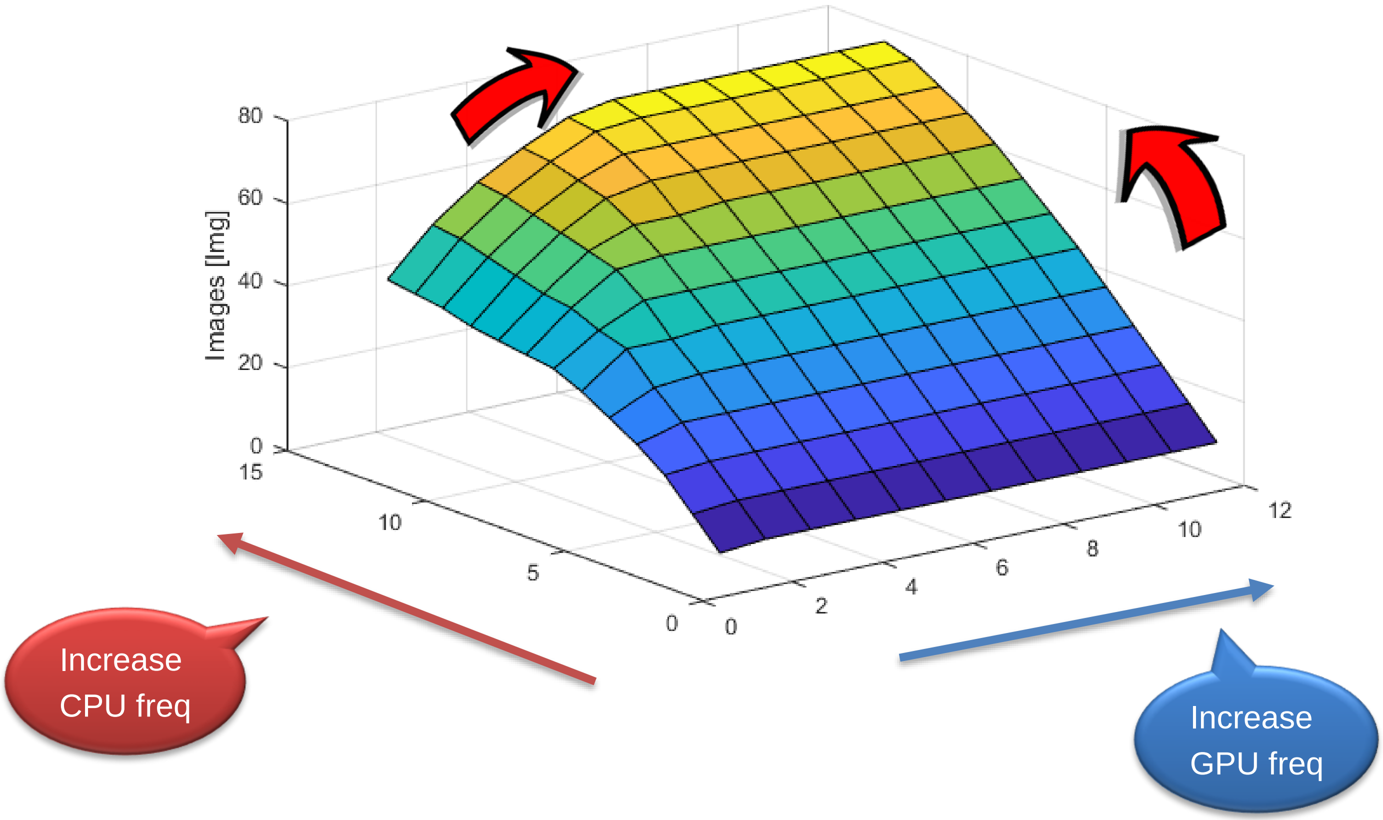
"You-only-look-once"

- ✓ 24 convolutional layers followed by 2 fully connected layers
  - – Inspired by the GoogleNet project
- ✓ Performs object categorization, detection and segmentation
  - – ImageNet dataset, 1k classes
- ✓ FPGA version is based on Darknet (wip)


448
7
7
448
3
112
3
3
112
192
56
3
3
56
256
28
3
3
28
512
14
3
3
14
1024
7
3
3
7
1024
7
7
1024
4096
7
7
30
Conv. Layer 7x7x64-s-2 Maxpool Layer 2x2-s-2
Conv. Layer 3x3x192 Maxpool Layer 2x2-s-2
Conv. Layers 1x1x128 3x3x256 1x1x256 3x3x512 Maxpool Layer 2x2-s-2
Conv. Layers 1x1x256 3x3x512 }×4 1x1x512 3x3x1024 Maxpool Layer 2x2-s-2
Conv. Layers 1x1x512 3x3x1024 }×2 3x3x1024 3x3x1024-s-2
Conv. Layers 3x3x1024 3x3x1024
Conn. Layer
Conn. Layer

Yolo on TX2: throughput

- ✓ Explore CPU/GPU frequency scaling
- ✓ "only" 8 FPS (classification + detection + segmentation)

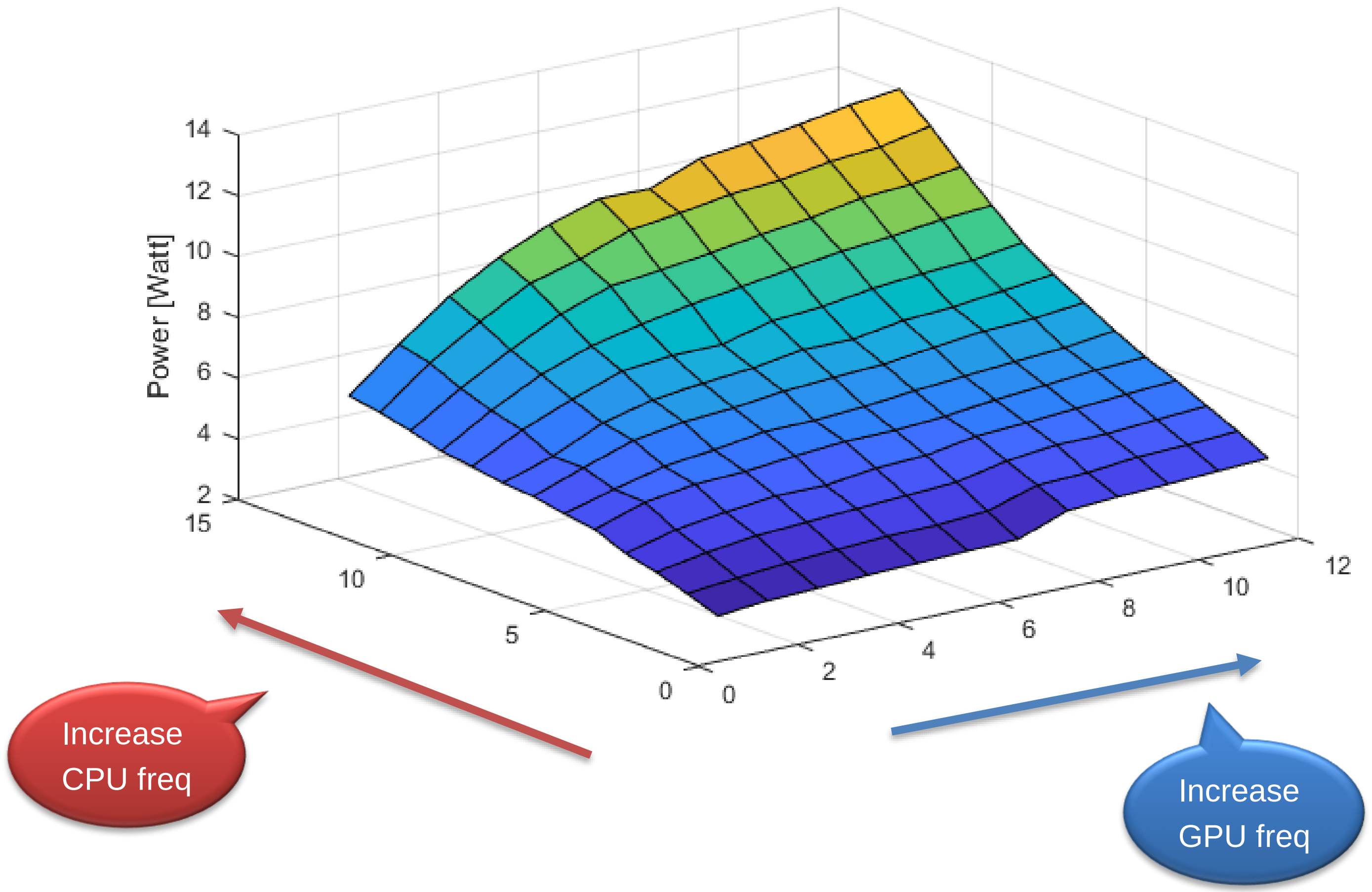


Increase CPU freq

Increase GPU freq

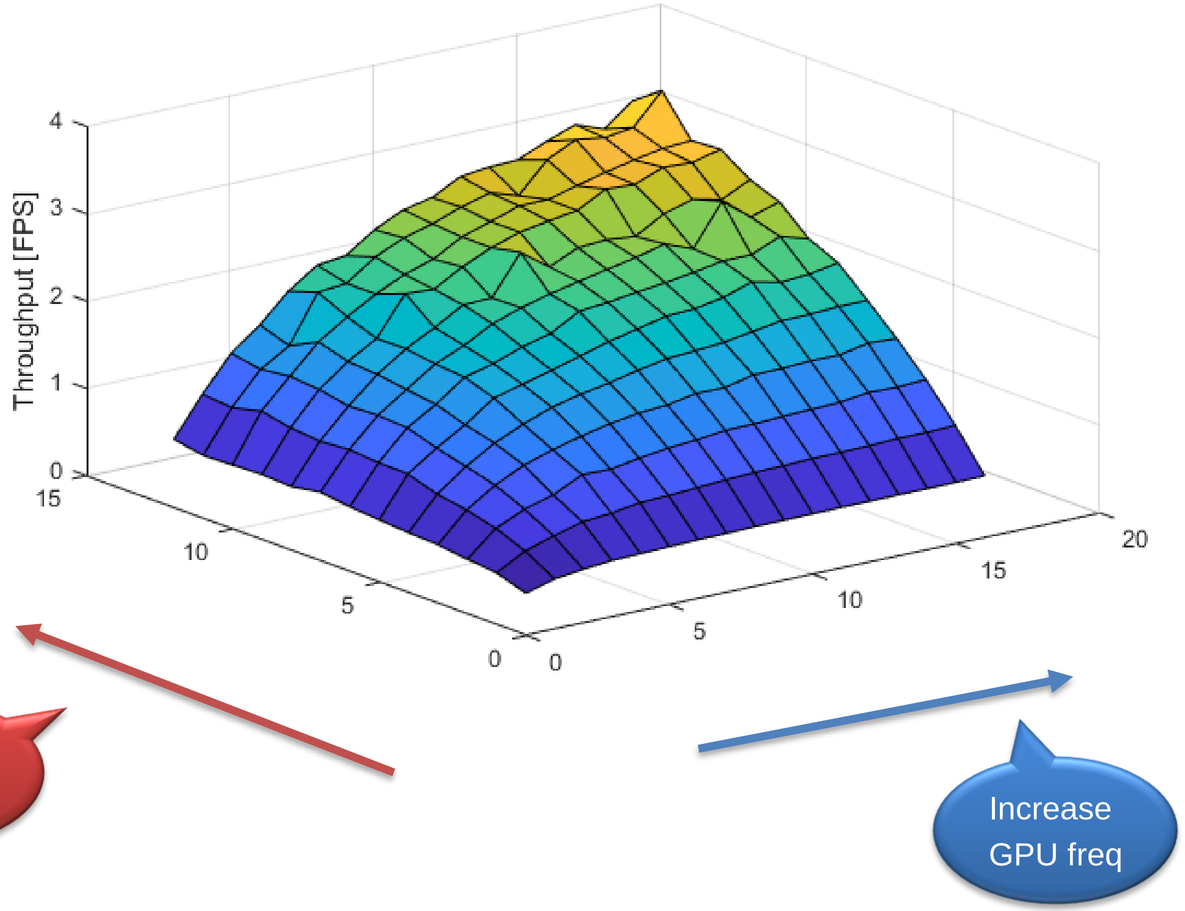
Yolo on TX2: power consumption

- ✓ Watt (TX2 claims 20W max)



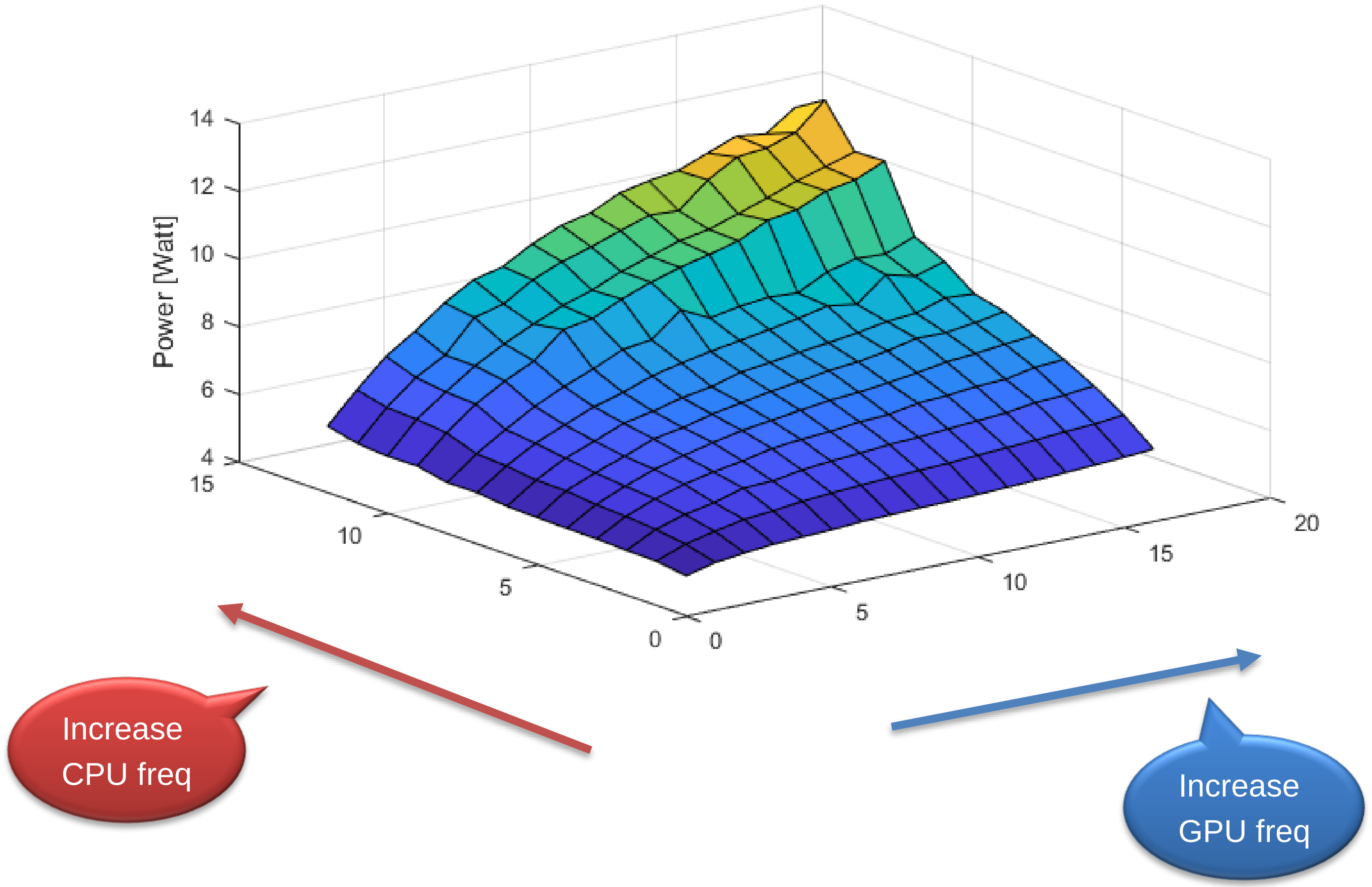
Yolo on TX1: throughput

- ✓ Same Host, Maxwell GPU (one generation older)
- ✓ Half perf than TX2



Yolo on TX1: power consumption

- ✓ Same power consumption



Yolo on XU+

- ✓ Hard to find an equivalent model and to synthesize it on FPGA
- ✓ "We're working for you"

ZynqNet

ZynqNet

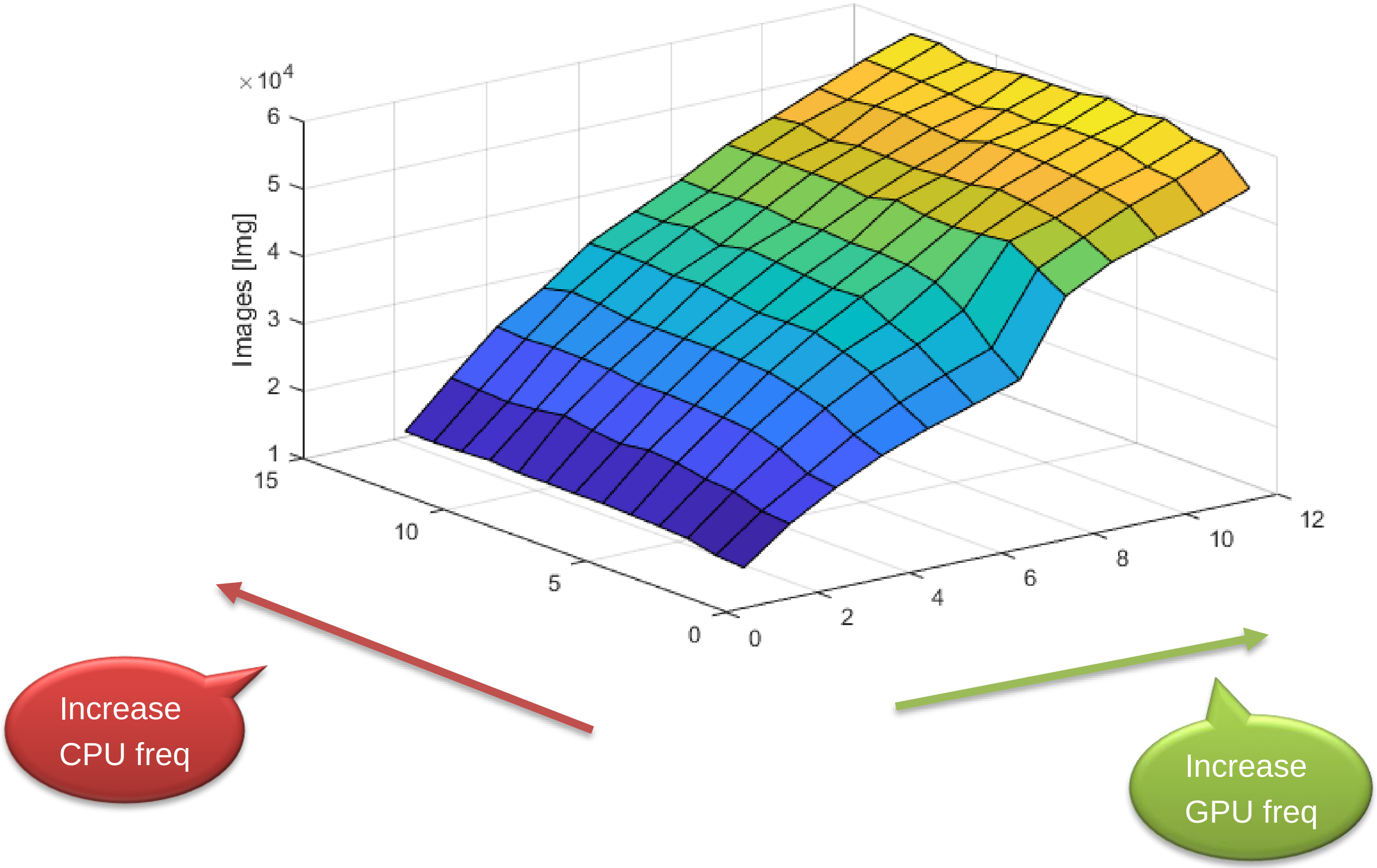
- ✓ Master thesis of David Gschwend @ETH Zurich
  - https://github.com/dgschwend/zynqnet
  - Performs classification
  - ImageNet dataset, 1k classes

- ✓ Written for Zynq arch
  - Not (yet) optimized
  - We ported on XU+

- ✓ Coffee model
  - To get GPU implementation for TX2

ZynqNet on TX2: throughput

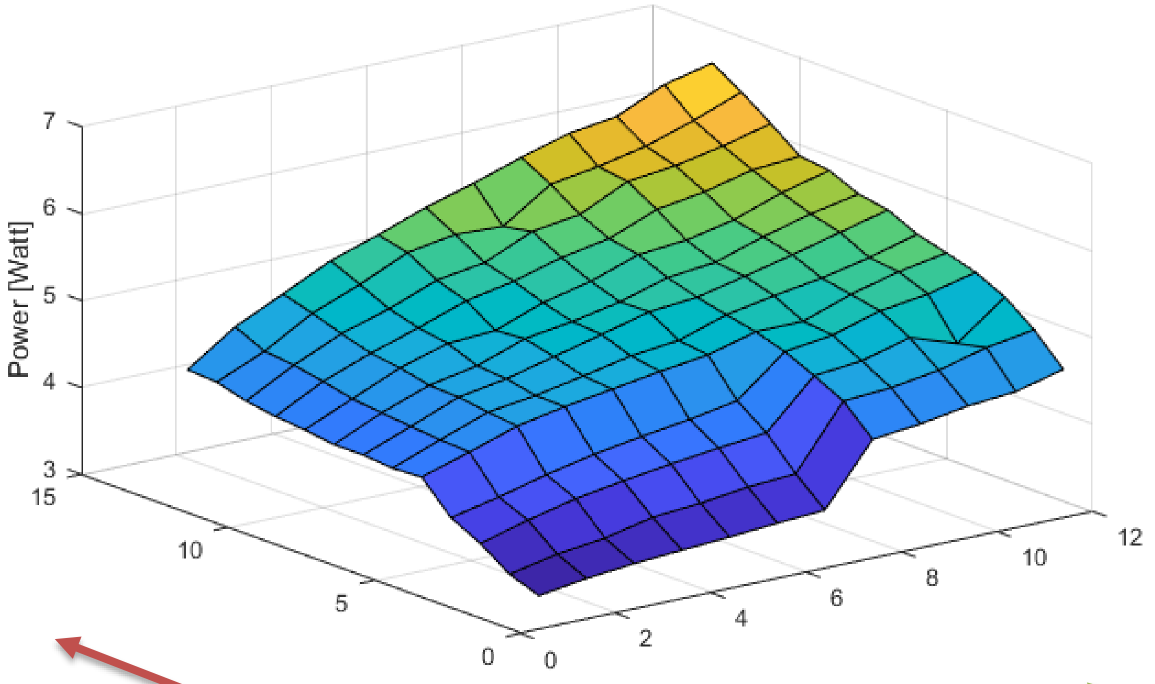
- ✓ #images in 5 minutes (…)
- ✓ Up to 200 FPS!!!!

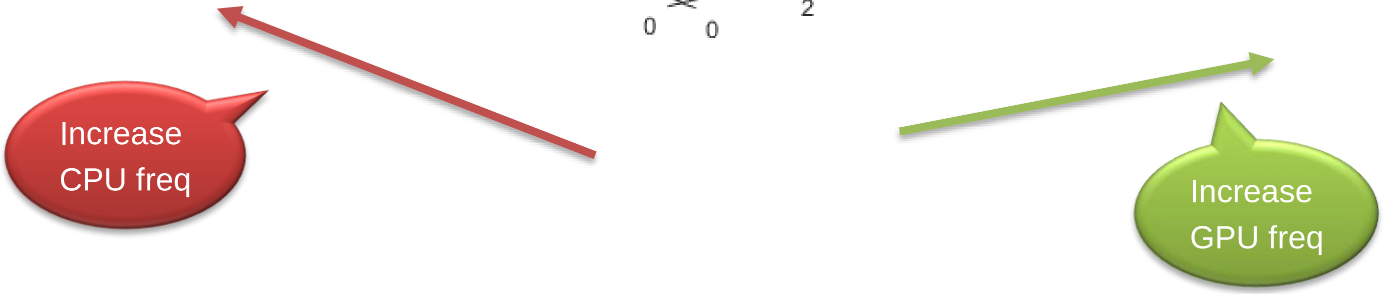


Increase CPU freq

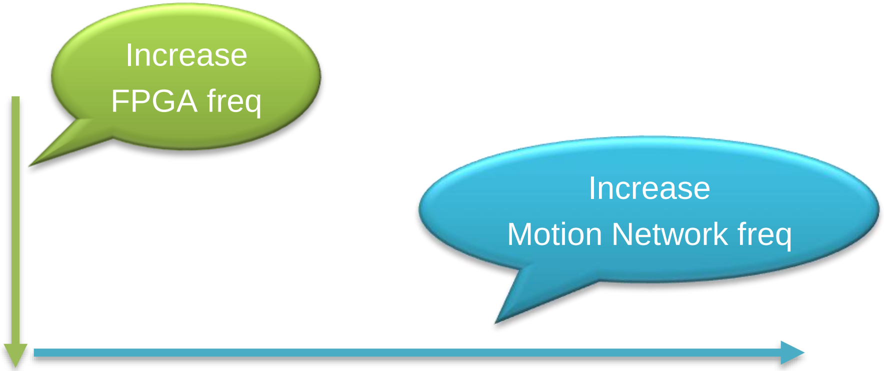
Increase GPU freq

ZynqNet on TX2: power consumption





ZynqNet on FPGA



Avg E2E latency (s)

	100MHz	150MHz	200MHz
100MHz	2,4752	2,4752	2,4751
150MHz	2,1782	2,1783	2,178
200MHz	1,9498	1,9487	1,9489

Accuracy (Top 5)

	100MHz	150MHz	200MHz
100MHz	390	390	390
150MHz	390	196	390
200MHz	390	390	390

Average power (Watt)

	100MHz	150MHz	200MHz
100MHz	19,8376	20,2953	20,8902
150MHz	19,8929	20,3305	20,9532
200MHz	19,9874	20,4512	20,9117

Energy consumption (J)

	100MHz	150MHz	200MHz
100MHz	25781	22688	20968
150MHz	25210	22382	21024
200MHz	25338	22885	21198

Throughput (Images/5min)

	100MHz	150MHz	200MHz
100MHz	121	121	121
150MHz	137	137	137
200MHz	153	153	153

What's left?

Future works

- Yolo
  - …
- ZynqNet
  - Optimizing the FPGA code to achieve comparable performance
  - Use fixed point datatypes

Currently, GP-GPU SoCs have **outstanding** performance

- Other platforms
  - Kalray MPPA -> KaNN Kalray-NN
  - ASIC -> TPU? PoliTo's?
- Other networks
  - PipeCNN
  - NEURaghe from UniCa/UniBo
  - Our (UniMoRe) CNN that reaches 40FPS @3 channels on ZedBoard

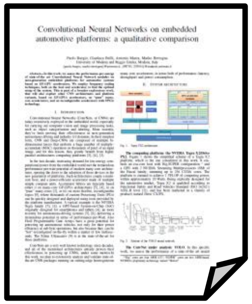
Want to contribute?

- ✓ Are you developing CNN packages for embedded accelerators?
- ✓ Are you developing embdedded accelerators?
- ✓ Want to compare?

Join our effort!

Thank you!

And..see you in Dresden!
(and hopefully Vancouver)



Convolutional Neural Networks on embedded automotive platforms: a qualitative comparison

DATE 18

RAW
Reconfigurable Architectures Workshop