2022 International Symposium on
VLSI Design, Automation and Test
April 18-21, 2022 Hsinchu, Taiwan

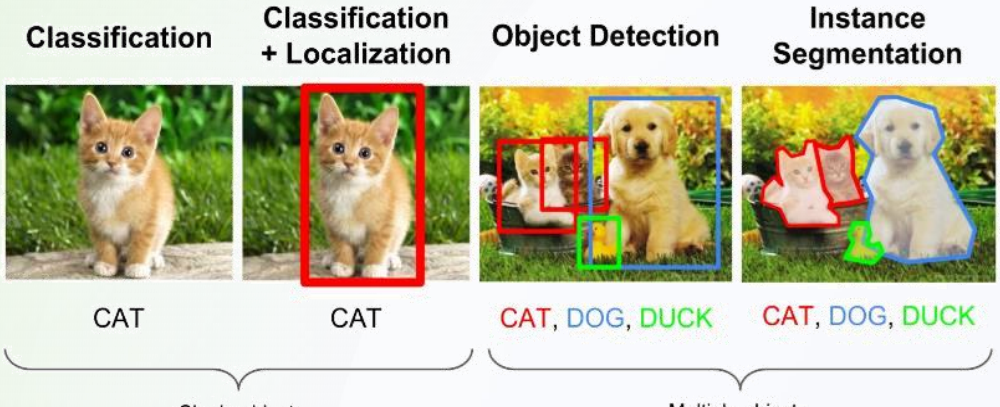
# Configurable Deep Learning Accelerator with Bitwise-accurate Training and Verification (D8.1)

**Shien-Chun Luo,** Kuo-Chiang Chang, Po-Wei Chen, and Zhao-Hong Chen
Integrated Perception and Computation Systems Division,
Electronic and Optoelectronic System Research Laboratories (EOSL), ITRI

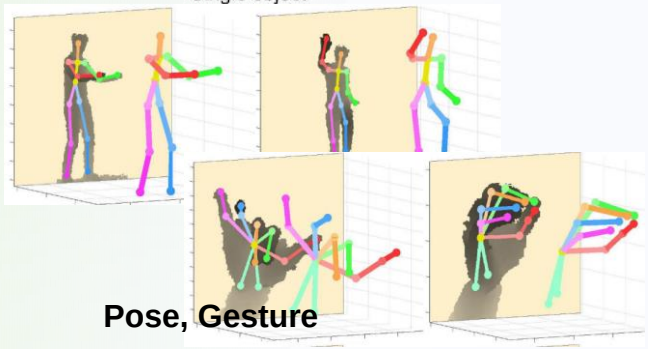
# Outline of This Presentation

- Solution of custom deep learning accelerator (DLA)
- Bit-accurate verification flow from training to inference
- Reference implementations about FPGAs and chips

# Optimize for the Applications that Edge AI Needs





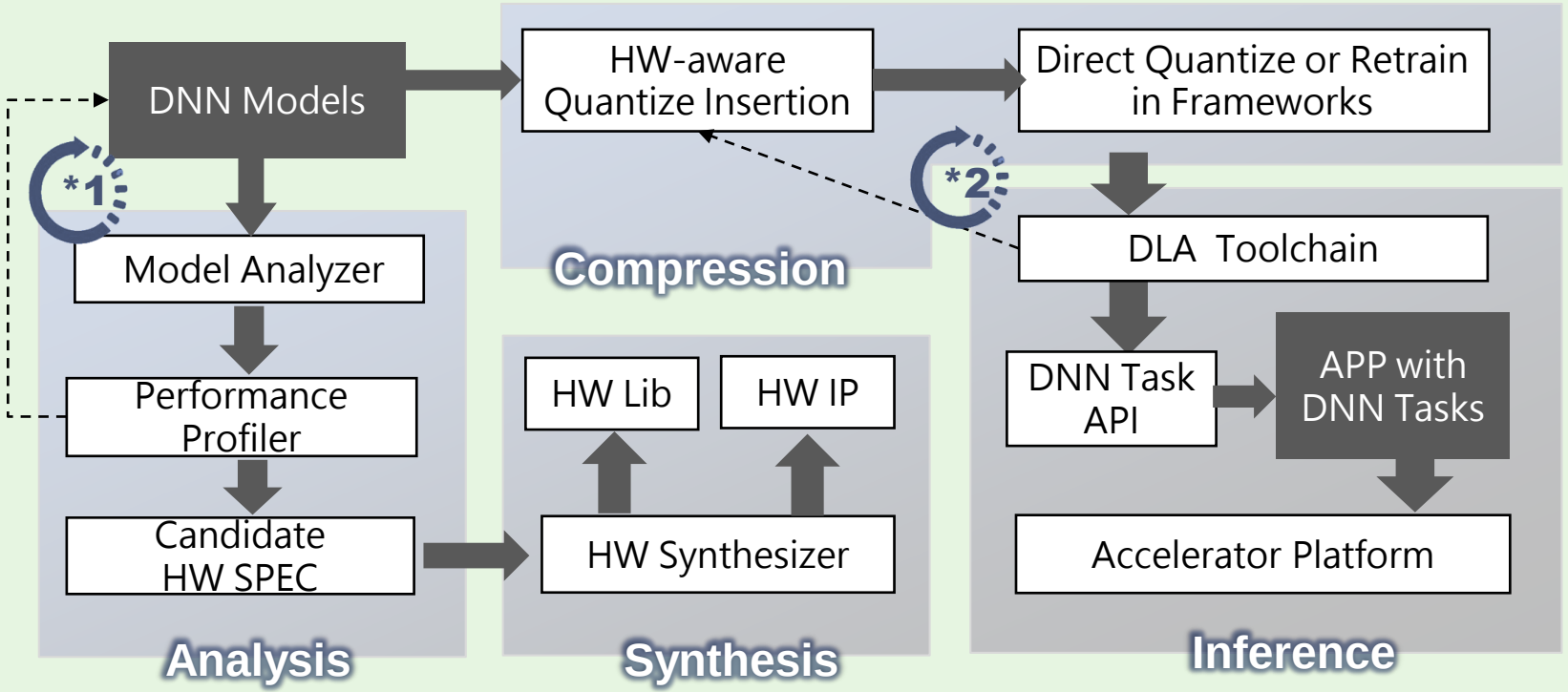




**Various performance and power consumption budgets**

# Our Total Solution

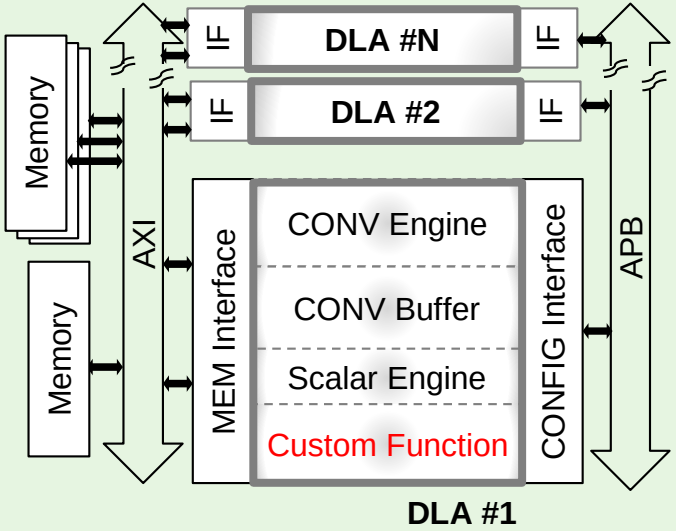
- Can select a HW SPEC by given models
- Can fine-tune model by given HW SPECs
- Proposes a bit-precise verification flow



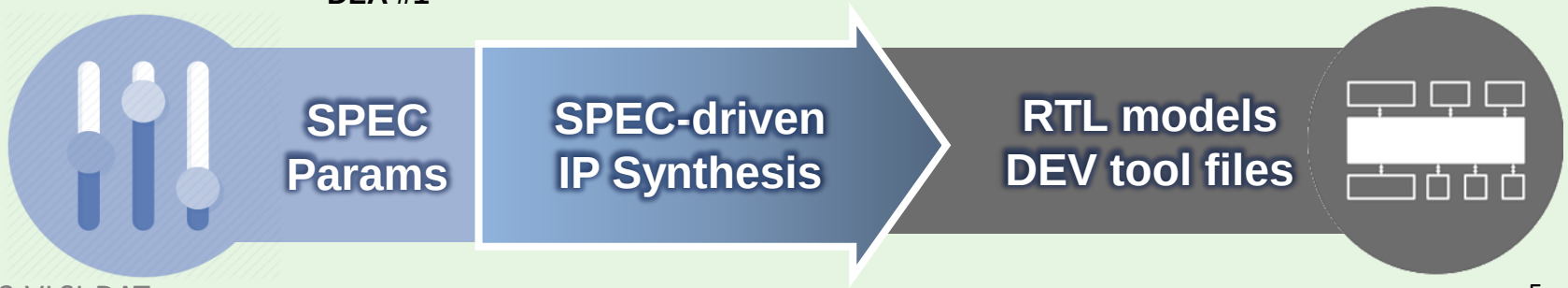

*1. Network architecture search *2. HW-precise quantize modification

# Configure an Accelerator

- Generate an DLA by finite SPEC params
- Scalable DLAs using standard ARM bus

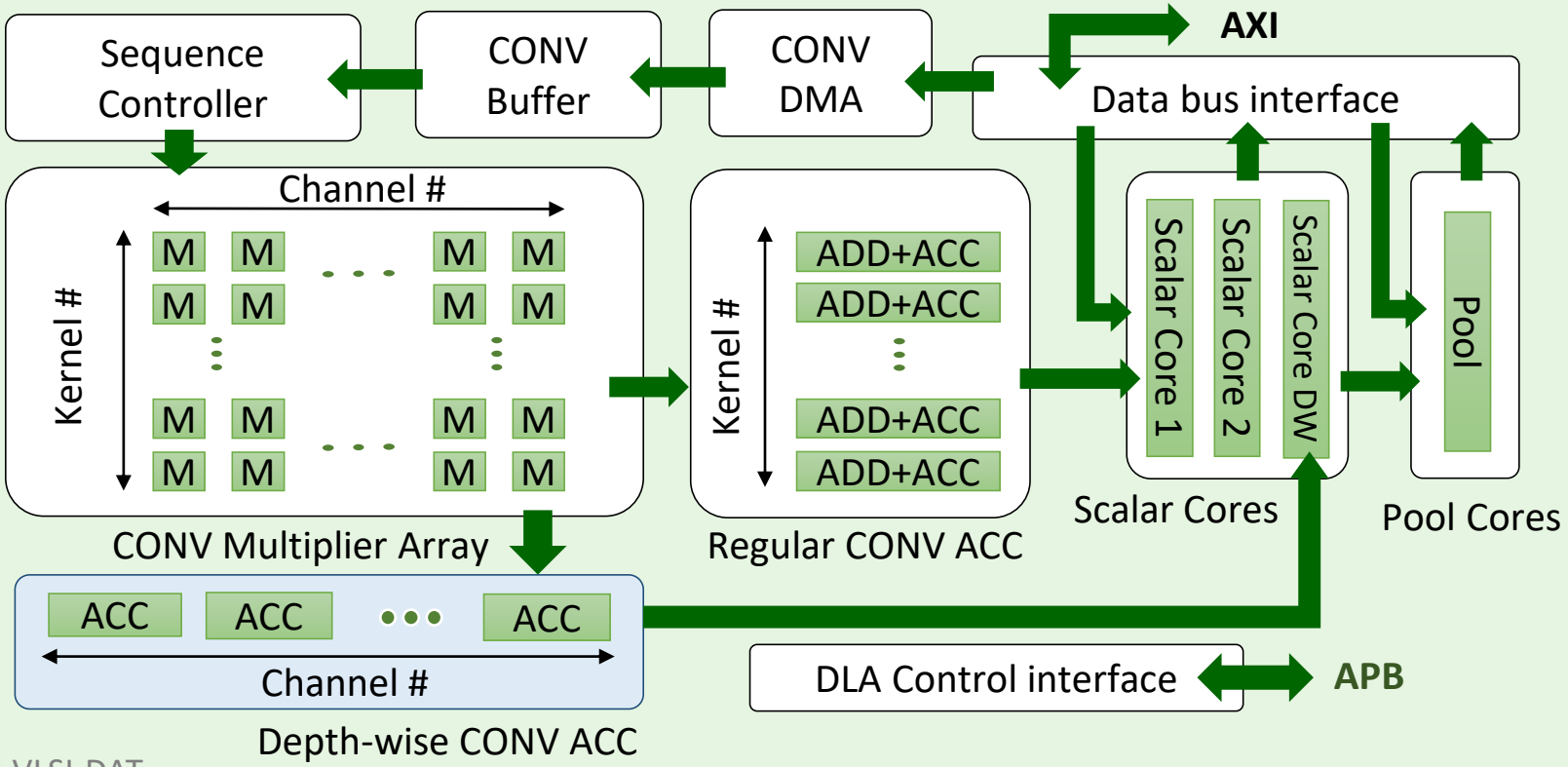



| Single DLA | SPEC Options |
|---|---|
| CONV PE Number | 64 ~ 2048 |
| CONV Buffer Size | 32KB ~ 512KB |
| Scalar Engine Config | Cascading engines = 2, 3<br>Parallel number = 1, 2, 4, 8 |
| Custom Function Config | Enable / Disable |

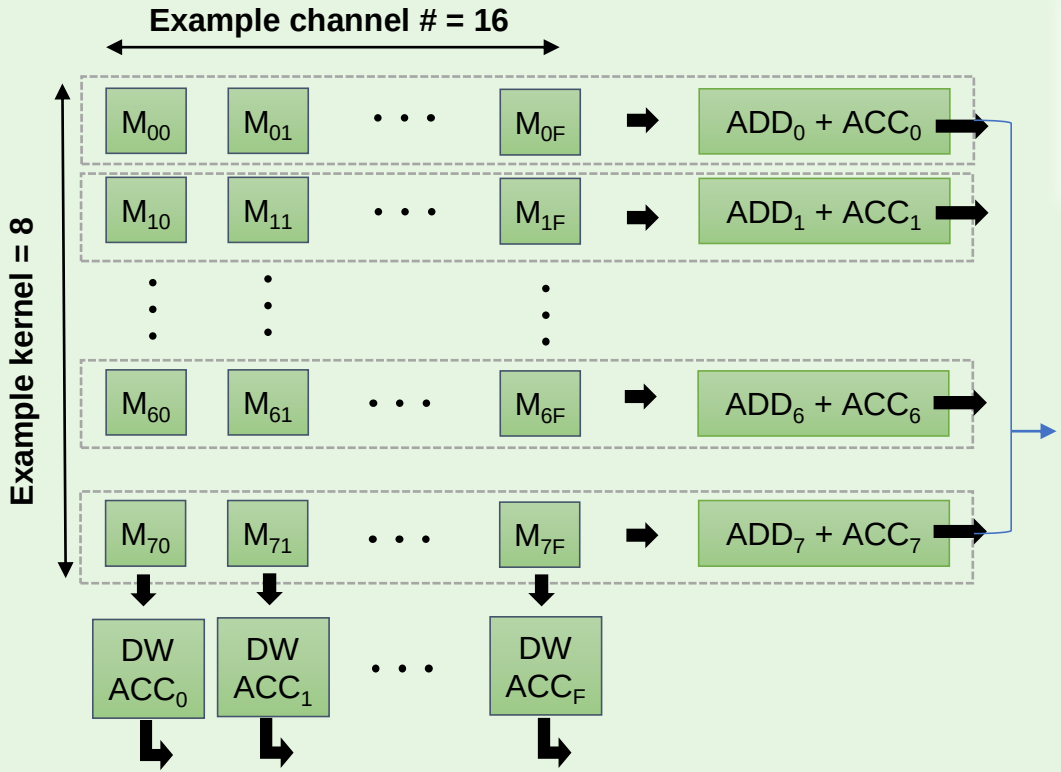


# Share the Hardware of Convolutions

- MUL array provides 2 data reuse strategies
- Data flow meets generic CNN flow



# Example of Regular and DW CONV Flows




**Assume**

Common Input = 7 x 7 x 256

1. Regular Kernel = 3 x 3 x 256 x 512
2. DW Kernel = 3 x 3 x 256

**8 Regular CONV Output Points after**

$3 \times 3 \times \frac{256}{16}$ cycles

**Complete Needs**

$\frac{512}{8}$ runs

**16 DW CONV Output after**

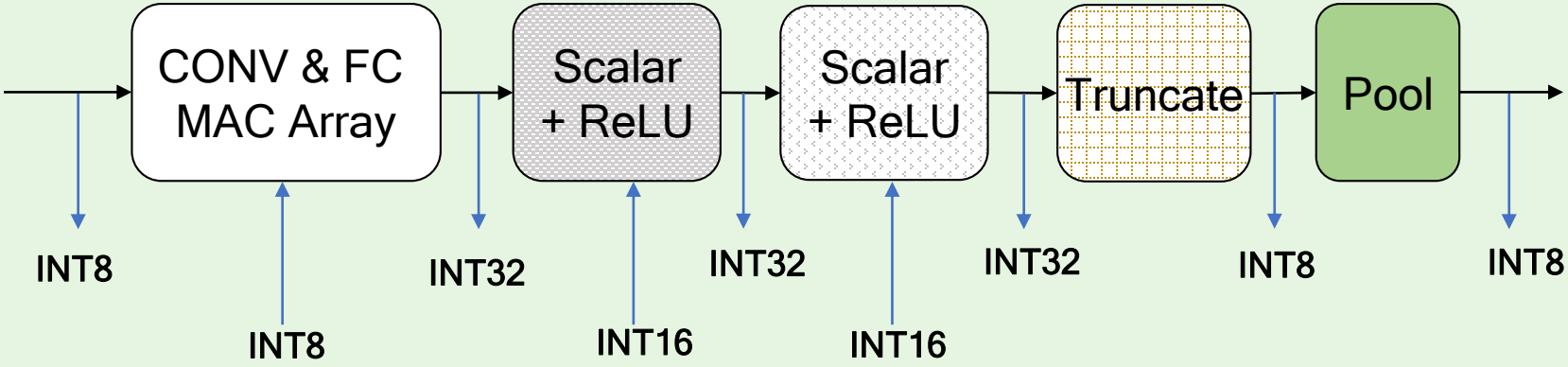
$3 \times 3$ cycles

**Complete Needs**

$\frac{256}{16}$ runs

# Embed HW Control Knobs into Training

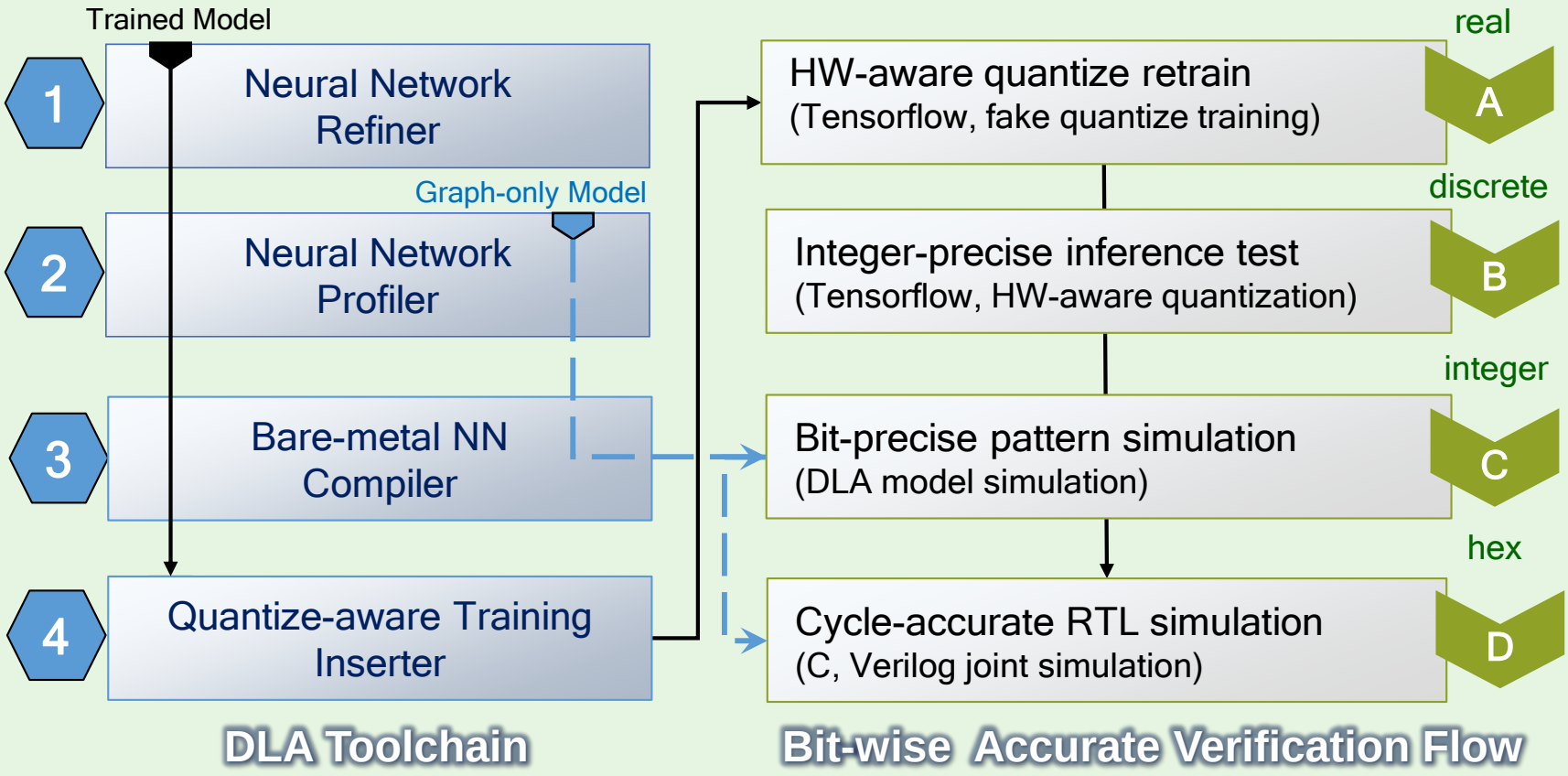
Data precision propagation in the hardware setup




Resource allocation and quantization sections → wrapped into an API of TF

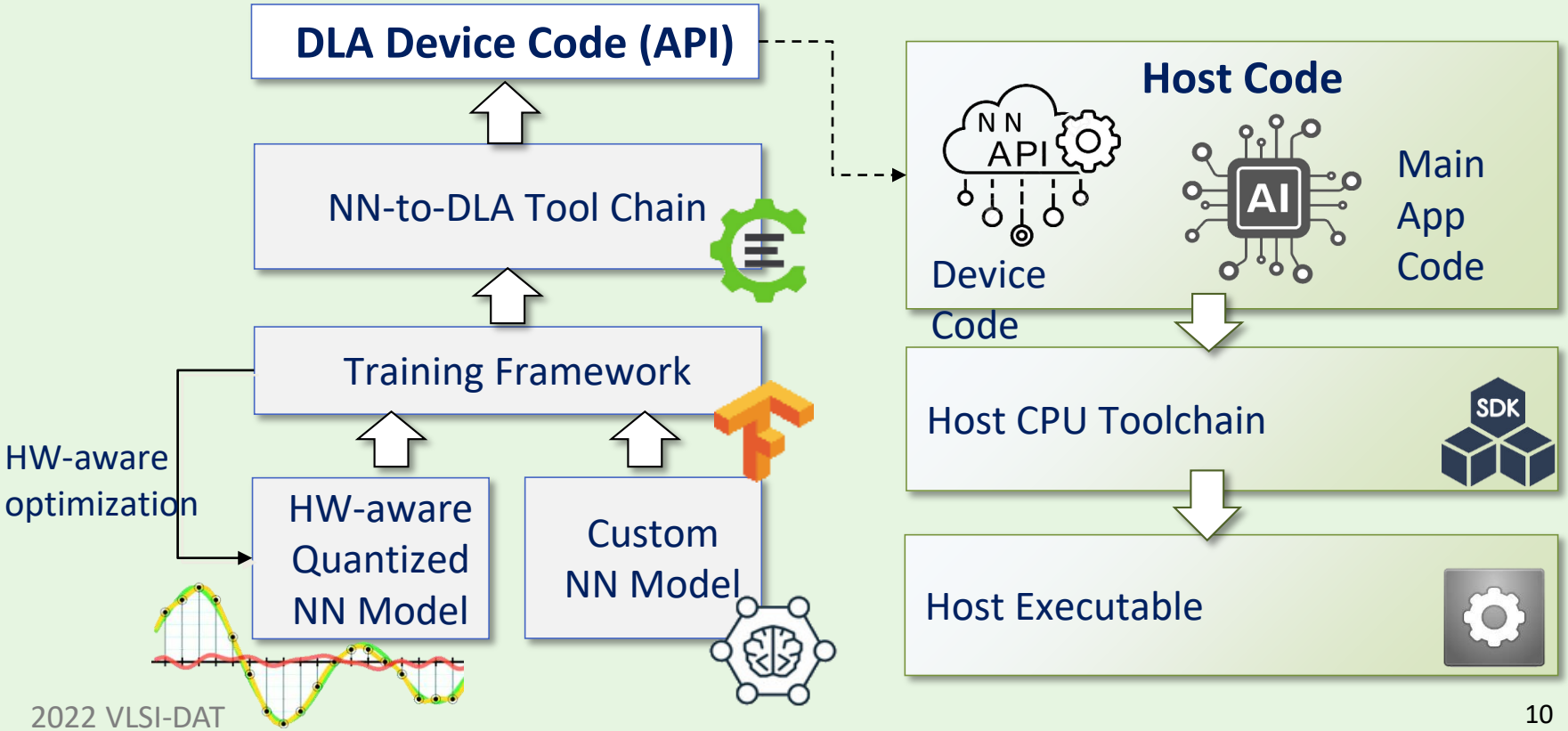
| Convolution | Bias | Scale | ReLU | Scale | Truncate | Max Pool |
|---|---|---|---|---|---|---|
| Folded BN | Folded BN | Leaky ReLU | Leaky ReLU | Quantize | Quantize | |

# Toolchain and Verification



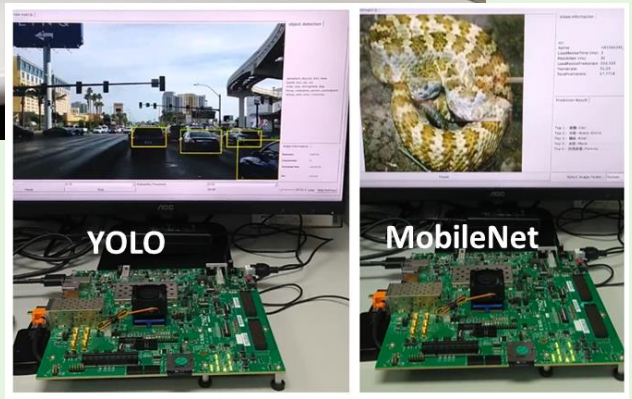
# Inference Flow Overview

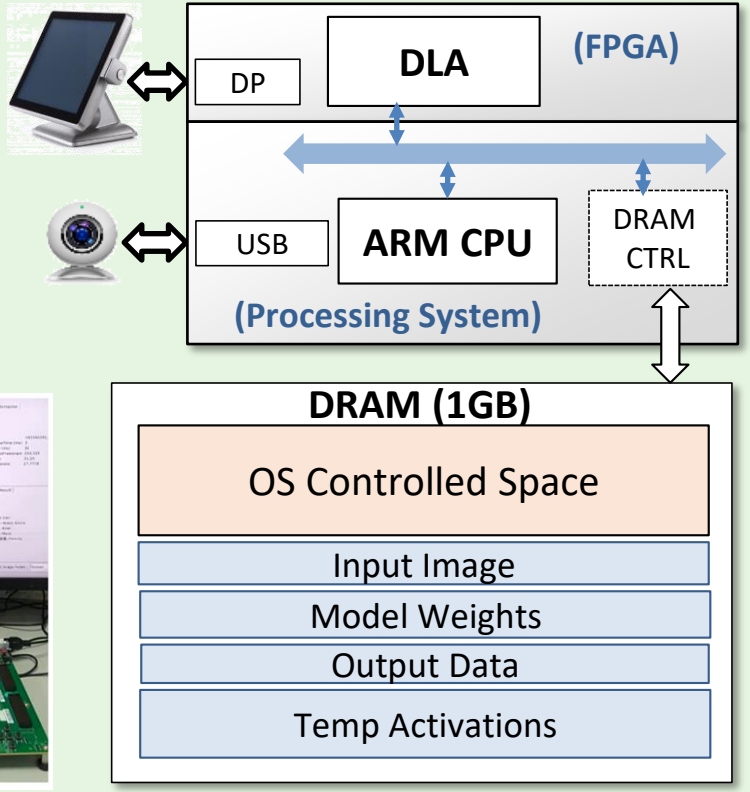
Each neural network model will be wrapped into a device code or API.



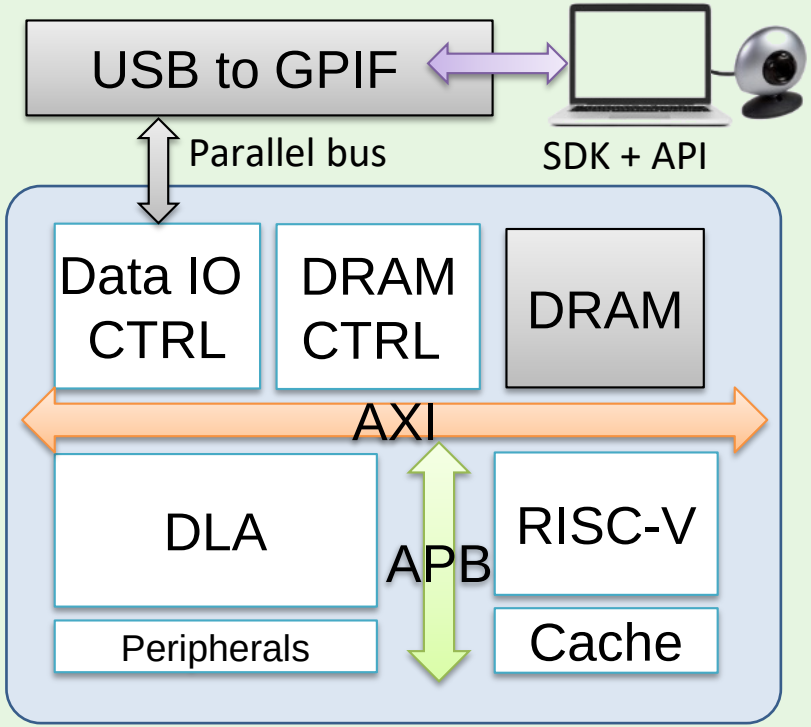
# FPGA Verifications - Standalone Device

Tiny YOLO running on
DLA2048M (ZCU102)



# FPGA Verifications - USB Accelerator




ZCU102 + Cypress FX3

VCU118 + Cypress FX3

CESYS EF03

# FPGA Verifications - Reference Performance

Utilization depends on (1) CNN dimension (2) DLA & AXI data bandwidth (3) Operators

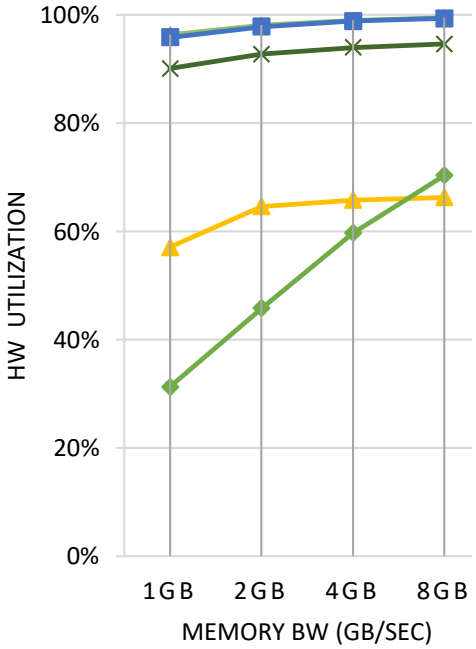
| DLA Series | DLA256M | DLA2048M |
|---|---|---|
| MAC Number | 256 | 2048 |
| CONV SRAM | 128KB | 512KB |
| Norm. Engine | 1X | 8X |
| Pool Engine | 1X | 4X |
| AXI Data BW | 1X | 8X |
| ZCU102 FPGA Resource % | 43% Logic, 3%DSP | 78% Logic, 86% DSP |
| ZCU102 FPGA Frequency | 200 MHz | 125 MHz |

| Reference Model | 200 MHz **DLA256M** | 125 MHz **DLA2048M** |
|---|---|---|
| Tiny YOLO v1 | 14.4 | 28.7 |
| Tiny YOLO v2 | 7.5 | 28.7 |
| Tiny YOLO v3 | 12.3 | 32.3 |
| Full YOLO v3 | 1.07 | 4.5 |
| Full YOLO v4 | 1.04 | 4.4 |
| Resnet50 | 6.4 | 17.6 |
| Mobilenet v1 | 33 | DW turn-off |

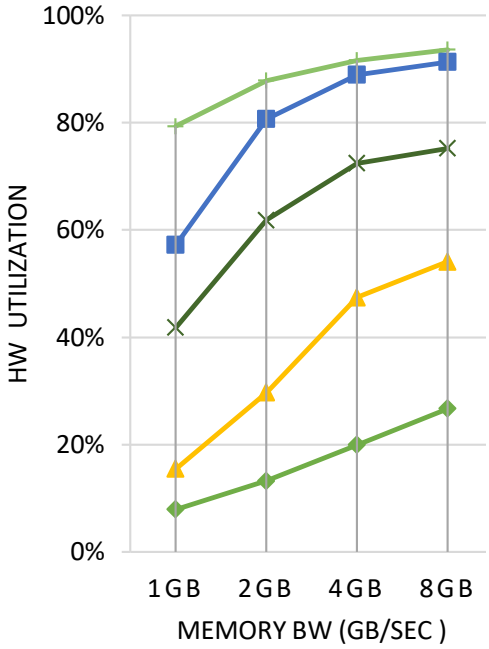
Unit: inference per second

# Data bandwidth and Performance Profiles

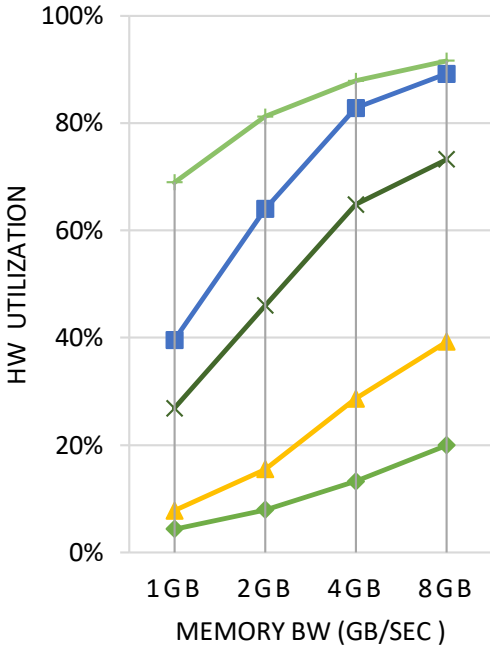
Utilization depends on (1) CNN dimension (2) DLA & AXI data bandwidth (3) Operators



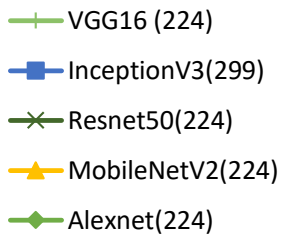

DLA64M (64 PE)
with 128KB buffer




DLA256M (256 PE)
with 128KB buffer




DLA512M (512 PE)
with 512KB buffer

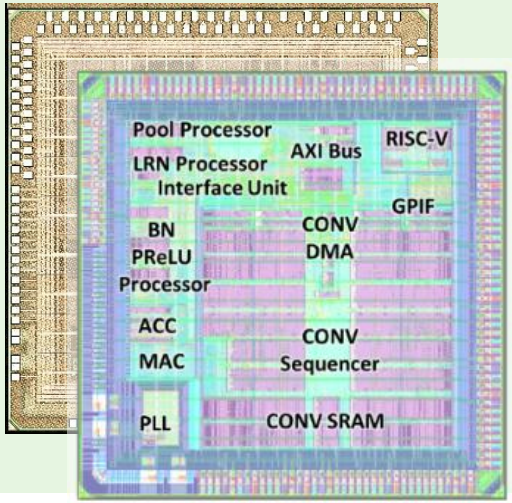



Factors of low utilization

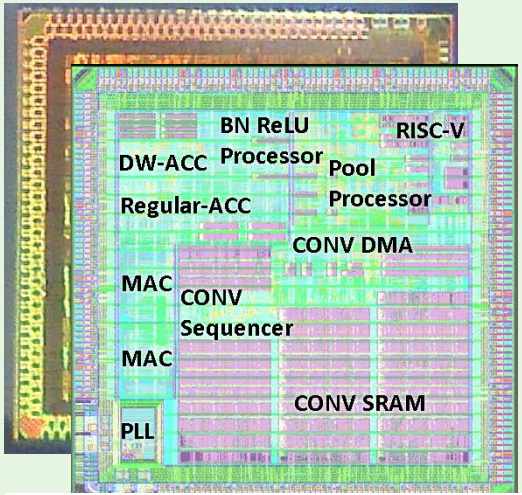
1. Individual pool
2. Arithmetic
3. Fully connection
4. Small channel

All these DLAs are operated at 150MHz.

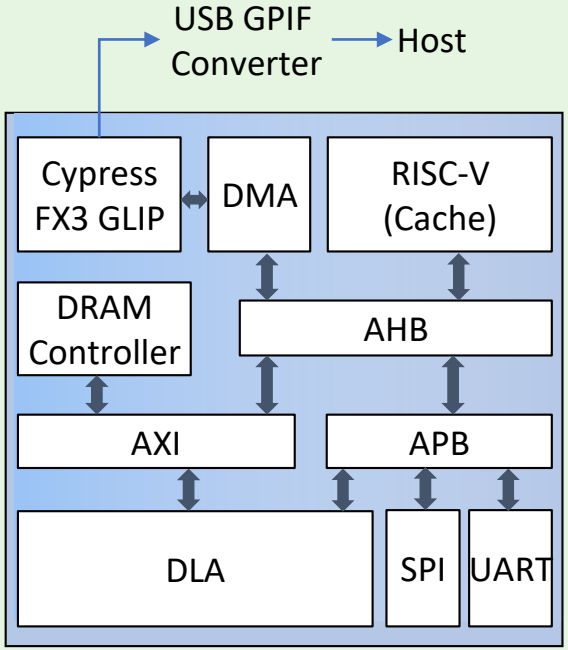
# ASIC Verifications




64MAC, 128KB @65nm,
3.2mm side,
50 GOPs / 60mW, 0.8TOPs/W

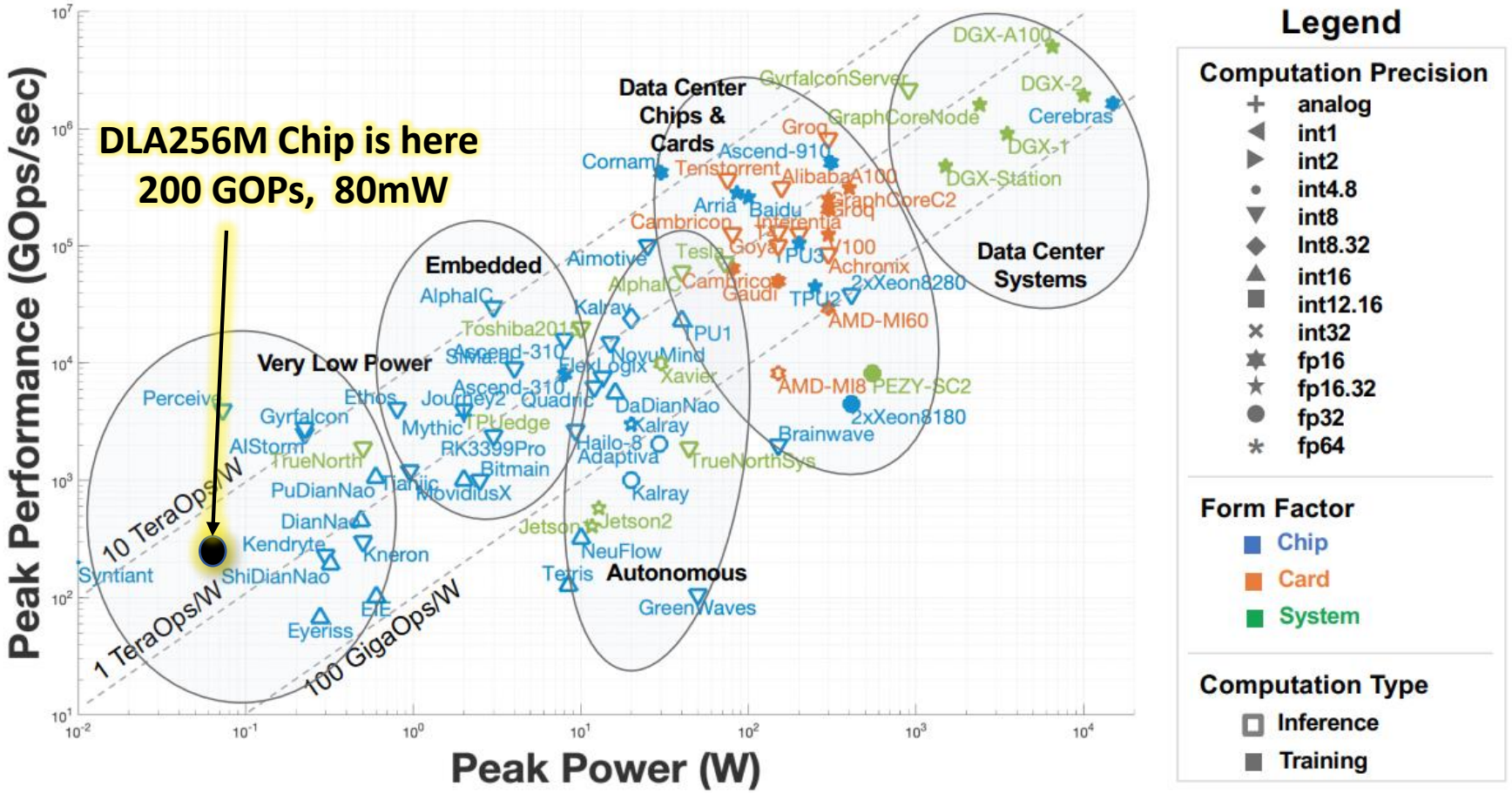



256MAC, 128KB @65nm,
3.6mm side,
200 GOPs / 80mW, 2.5 TOPs/W




Functional Block View

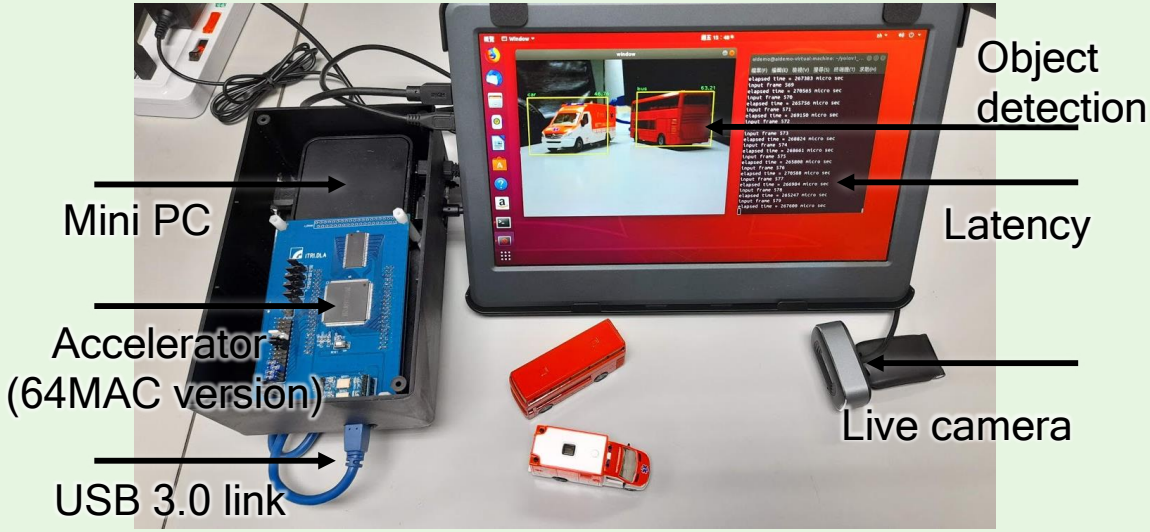
Peak power consumption measured by continuous high-utilization convolution task

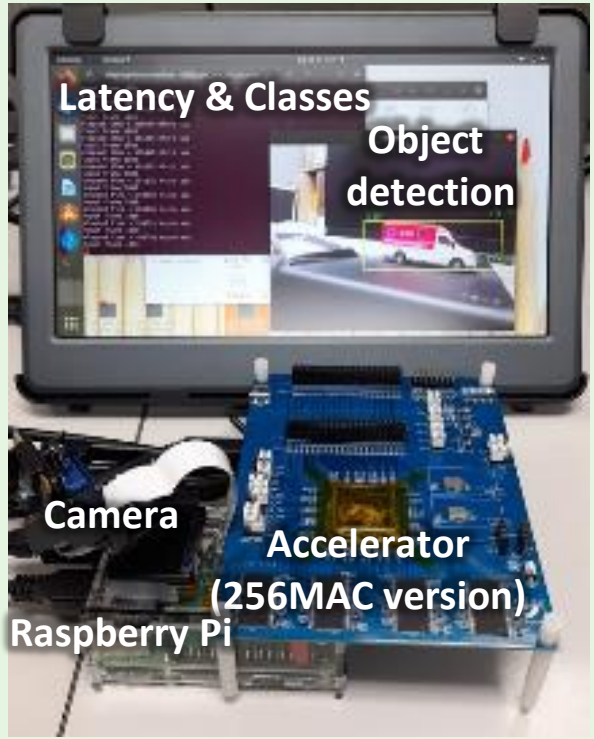


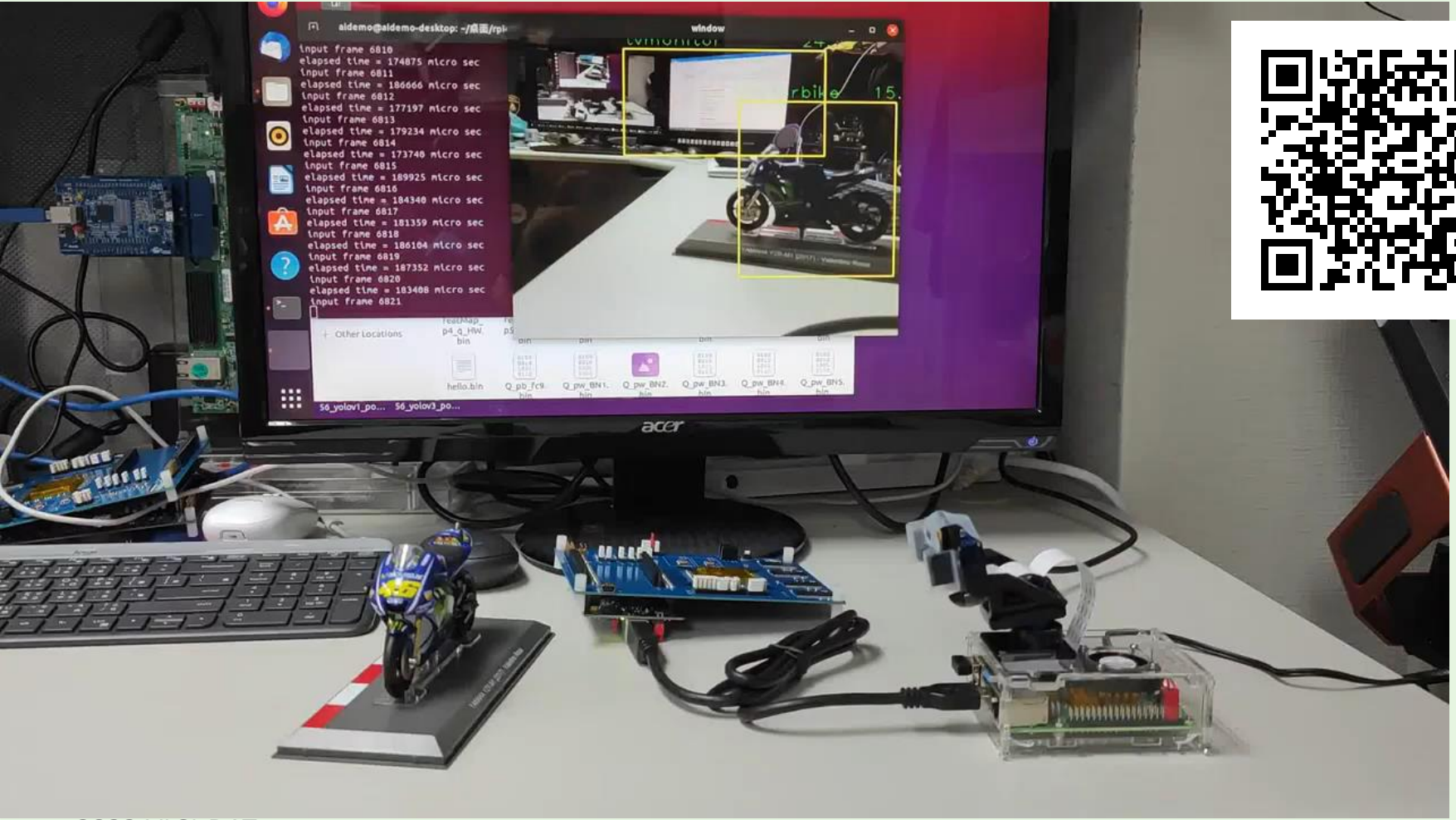



Survey and Benchmarking of Machine Learning Accelerators
https://arxiv.org/abs/1908.11348

# Evaluation Board and Environment




aidemo@aidemo-desktop: ~/桌面/rpi
window
input frame 6810
elapsed time = 174875 micro sec
input frame 6811
elapsed time = 186666 micro sec
input frame 6812
elapsed time = 177197 micro sec
input frame 6813
elapsed time = 179234 micro sec
input frame 6814
elapsed time = 173740 micro sec
input frame 6815
elapsed time = 189925 micro sec
input frame 6816
elapsed time = 184340 micro sec
input frame 6817
elapsed time = 181359 micro sec
input frame 6818
elapsed time = 186104 micro sec
input frame 6819
elapsed time = 187352 micro sec
input frame 6820
elapsed time = 183408 micro sec
input frame 6821
Other Locations
acer

Demo Video

# Brief Summaries

1. We developed a custom DLA SPEC to RTL solution.
2. The DLA toolchain provides bit-precise verifications from training to RTL simulation.
3. Reference FPGA and chip implementations introduce the performance and power consumption for edge AI products.