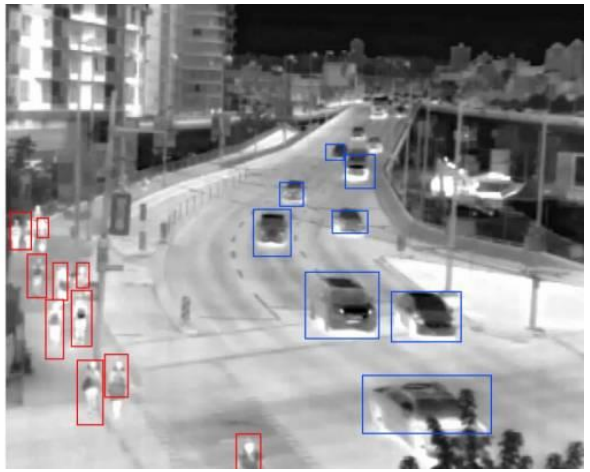
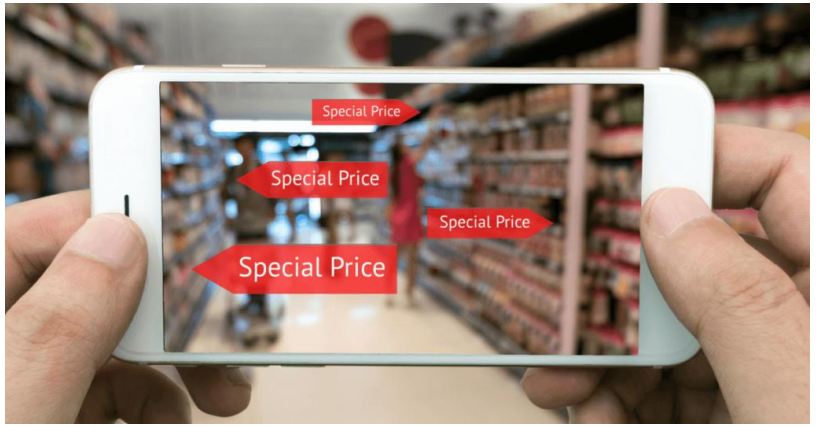
# Xilinx Machine Learning Strategies for the Edge

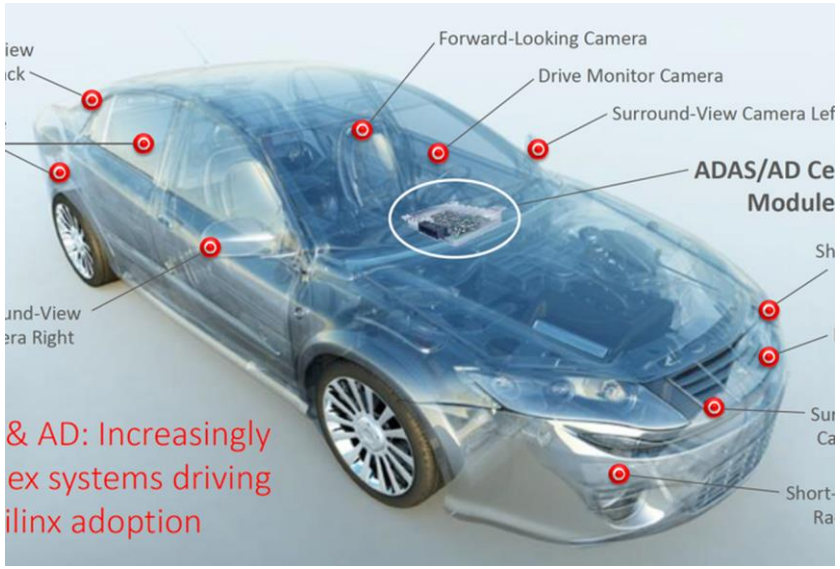
Thomas Schwing, Xilinx FAE

GET READY GET SET GO ADAPT

# AI/ML Monetization Is Here and Growing

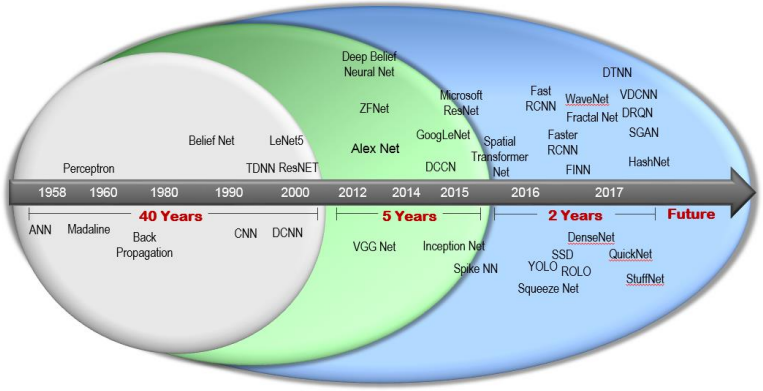





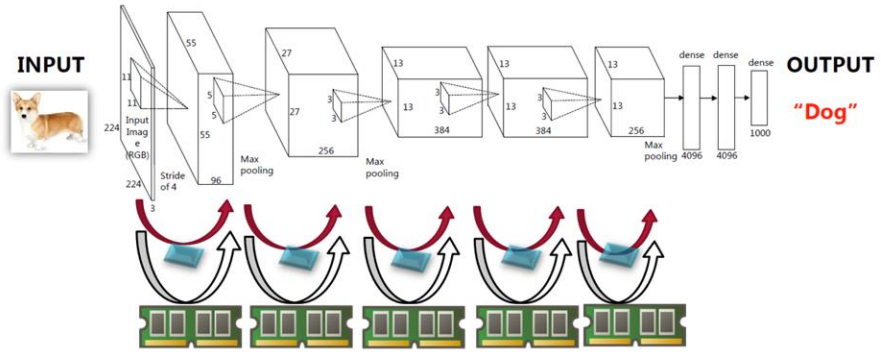




# Who is Xilinx? Why Should I Care for ML?

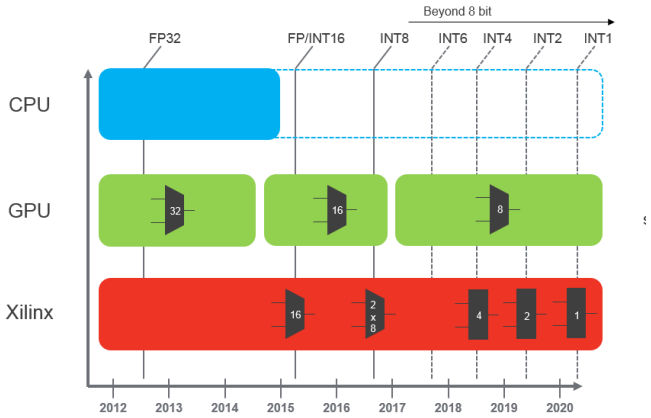
## 1 Only HW/SW configurable device for fast changing networks



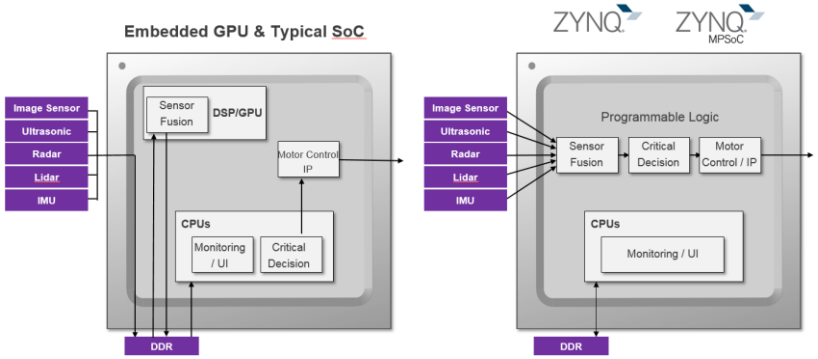

## 2 High performance / low power with custom internal memory hierarchy



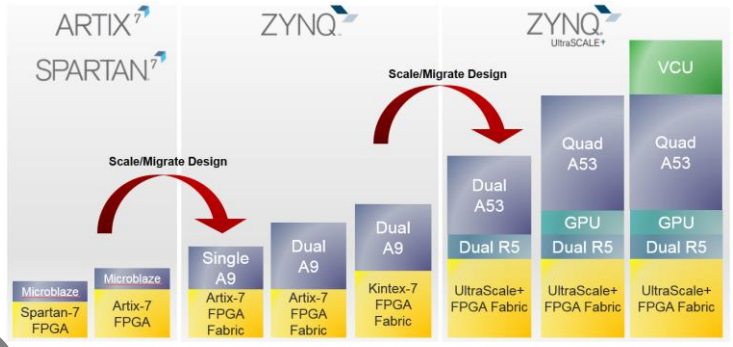

## 3 Future proof to lower precisions




## 4 Low latency end-to-end




## 5 Scalable device family for different applications




XILINX

# Integrated Xilinx AI Roadmap

Xilinx AI Development

| | Edge/Embedded | Cloud/DC |
|---|---|---|
| Models | 20+ pruned / customized / basic models | |
| Software Stack | Pruning | |
| | Quantizer | Quantizer |
| | Compiler | xfDNN Compiler |
| | Runtime | xfDNN Runtime |
| FPGA IP | DPU, LSTM | xDNN |

Platforms


Z7020 Board


Z7020 SOM


ZU2/3 SOM


ZU2/3 Card

ZU9 Card


ZCU102


ZCU104


Ultra96


Xilinx U200, U250, U280



XILINX

# Xilinx Network Development

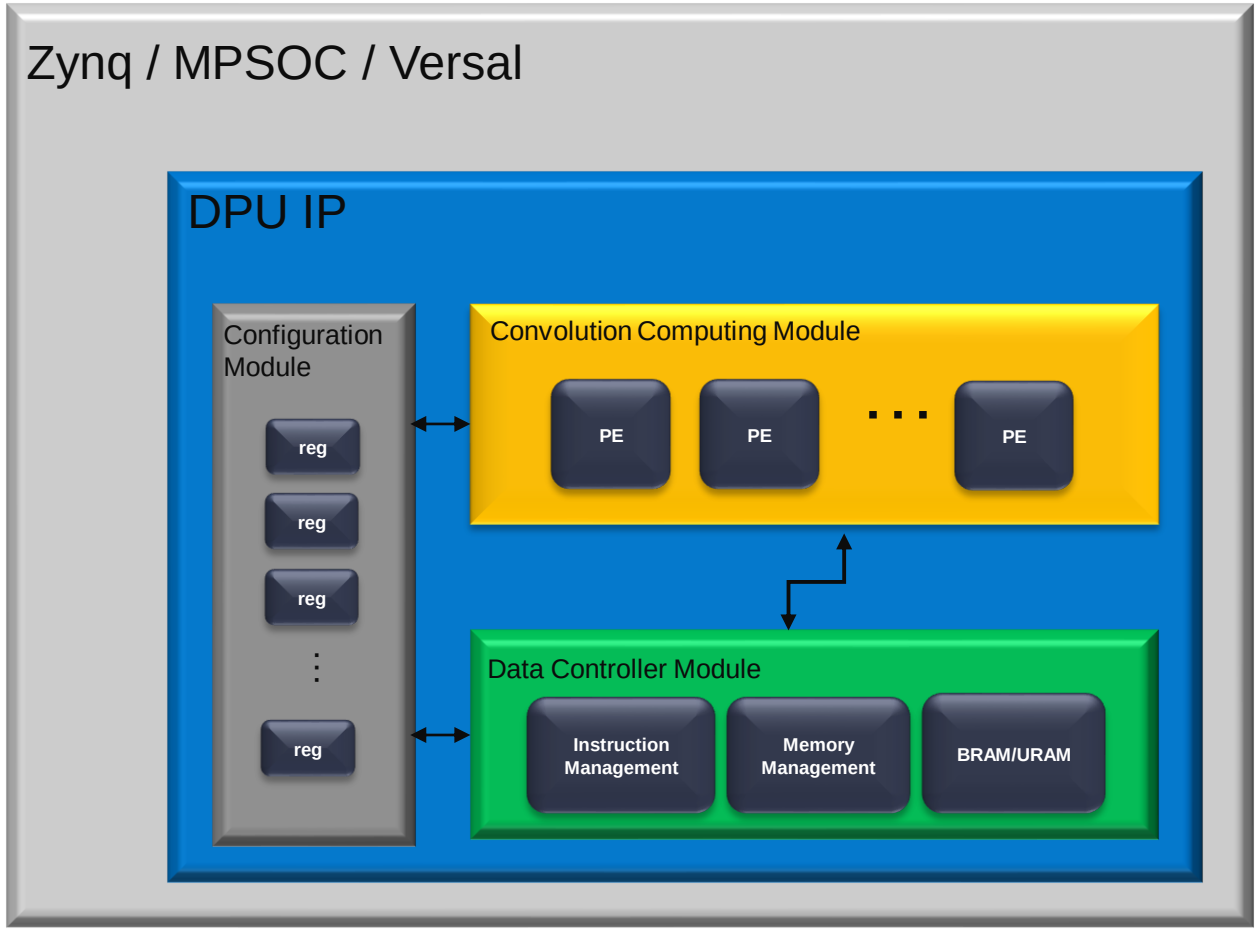
| Application | Module | Algorithm | Model Development | Compression | Deployment |
|---|---|---|---|---|---|
| Face | Face detection | SSD, Densebox | ✔ | ✔ | ✔ |
| | Landmark Localization | Coordinates Regression | ✔ | N / A | ✔ |
| | Face recognition | ResNet + Triplet / A-softmax Loss | ✔ | ✔ | ✔ |
| | Face attributes recognition | Classification and regression | ✔ | N / A | ✔ |
| Pedestrian | Pedestrian Detection | SSD | ✔ | ✔ | ✔ |
| | Pose Estimation | Coordinates Regression | ✔ | ✔ | ✔ |
| | Person Re-identification | ResNet + Loss Fusion | ✔ | | |
| Video Analytics | Object detection | SSD, RefineDet | ✔ | ✔ | ✔ |
| | Pedestrian Attributes Recognition | GoogleNet | ✔ | ✔ | ✔ |
| | Car Attributes Recognition | GoogleNet | ✔ | ✔ | ✔ |
| | Car Logo Detection | DenseBox | ✔ | ✔ | |
| | Car Logo Recognition | GoogleNet + Loss Fusion | ✔ | ✔ | |
| | License Plate Detection | Modified DenseBox | ✔ | ✔ | ✔ |
| | License Plate Recognition | GoogleNet + Multi-task Learning | ✔ | ✔ | ✔ |
| ADAS/AD | Object Detection | SSD, YOLOv2, YOLOv3 | ✔ | ✔ | ✔ |
| | 3D Car Detection | F-PointNet, AVOD-FPN | ✔ | | |
| | Lane Detection | VPGNet | ✔ | ✔ | ✔ |
| | Traffic Sign Detection | Modified SSD | ✔ | | |
| | Semantic Segmentation | FPN | ✔ | ✔ | ✔ |
| | Drivable Space Detection | MobilenetV2-FPN | ✔ | | |
| | Multi-task (Detection+Segmentation) | Xilinx | ✔ | | |

# DPU IP

XILINX

# What is the DPU?

- **Deep learning Processor Unit: Optimized for convolutional neural networks**
- **Consists of 3 Main modules**
  - Configuration module
  - Data Controller module
  - Convolution Computing module
- **Instruction Set**
  - Tensor based instructions
  - Up to 268,435,456 MACs/instruction
- **DPU Targets the Zynq Device Family**
  - APU required:
    - Interrupt handling
    - Data transfers
    - Unsupported operations



# DPU V1.3.0: Supported CNN Operations

- **Arbitrary input image size**
- **Convolution**
  - Arbitrary kernel size
  - Arbitrary stride/padding
  - Dilation
  - Depthwise (optional)
- **Pooling**
  - Max pooling
  - Arbitrary pooling size
  - Arbitrary stride/padding
  - Average Pooling (optional)
- **ReLU/Leaky ReLU/**Relu6(optional)
- **Concat**
- **Split**
- **Elementwise sum**
- **Fully Connected (FC)**
- **Deconv**
- **Batch normalization**
- **Mean scale**
- **Reorg**
- **Resize (optional)**
- **Softmax (optional)**
- **Sigmoid (optional)**

Operations not listed here *must* be mapped to the ARM CPU

# Supported Layers

| Layer Type \ Next Layer | Conv | Deconv | Depth-wise Conv | Inner Product | Max Pooling | Ave Pooling | BN | ReLU | LeakyReLU | Element-wise | Concat | As Input | As Output |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Conv | ● | ● | ○ | ● | ● | ○ | ● | ● | ○ | ● | ● | ● | ● |
| Deconv | ● | ● | ○ | ● | ● | ○ | ● | ● | ○ | ● | ● | ● | ● |
| Depth-wise Conv | ● | ● | ○ | ● | ● | ○ | ● | ● | ○ | ● | ● | ● | ● |
| Inner Product | ● | ● | ○ | ● | ● | ○ | ● | ● | ○ | ● | ● | ● | ● |
| Max Pooling | ● | ● | ○ | ● | ● | ○ | ○ | × | × | ● | ● | ● | ● |
| Ave Pooling | ○ | ○ | ○ | ○ | ○ | ○ | ○ | × | × | ○ | ○ | ○ | ○ |
| BN | ● | ● | ○ | ● | ● | ○ | ○ | ● | × | ● | ● | ○ | ○ |
| ReLU | ● | ● | ○ | ● | ● | ○ | ○ | × | × | ● | ● | --- | ● |
| LeakyReLU | ○ | ○ | ○ | ○ | ○ | ○ | ○ | × | × | ○ | ○ | --- | ○ |
| Element-wise | ● | ● | ○ | ● | ● | ○ | ○ | ● | ○ | ● | ● | --- | ● |
| Concat | ● | ● | ○ | ● | ● | ○ | ○ | × | × | ● | ● | --- | ● |

●: Support　　×: Not support　　○: Supported with specific DPU Versions

# DPU Support CNN Operation Revision History

- **DPU V1.3.0**
  - Conv + Max Pooling + Relu / Leaky Relu + FC + BN + Concat + Elementwise + Split + Reorg + Up sampling
- **DPU V1.3.1**
  - V1.3.0 + Average Pooling
- **DPU V1.3.7**
  - V1.3.0 + Depthwise conv + Average Pooling + Relu6
- **DPU V1.4.0**
  - Unified version with all feature configuration

# DPU Evolution

XILINX

# DPU Optimizations

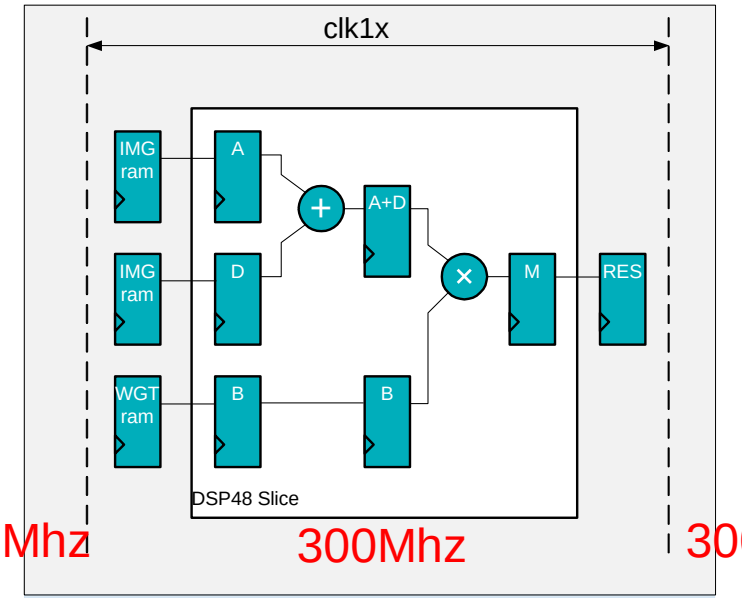
## DPU

- One clock domain
- Lower power than DPU_EU
- Instructions: Convolution, Deconvolution, Depthwise Convolution, MaxPool, AveragePool, Elementwise,Softmax,Sigmoid…

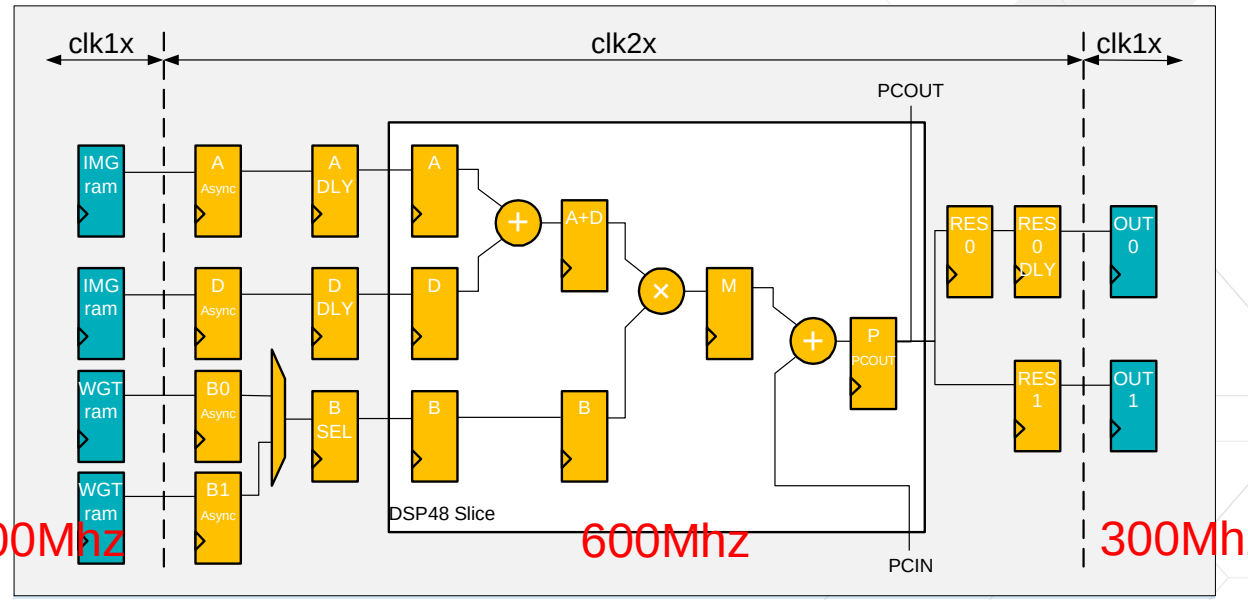
## DPU_EU

- Two clock domains
- Smaller DSP footprint than DPU
  - Uses DSP DDR technique
  - Uses DSP cascading to reduce resources
  - Optional ability to trade DSPs for LUTs to reduce DSP utilization
- Low power version (DPU_EU_LP) uses gated clocks to reduce power consumption (but is still higher power than DPU)

# Optimizations Using DSP DDR Technique (DPU_EU)



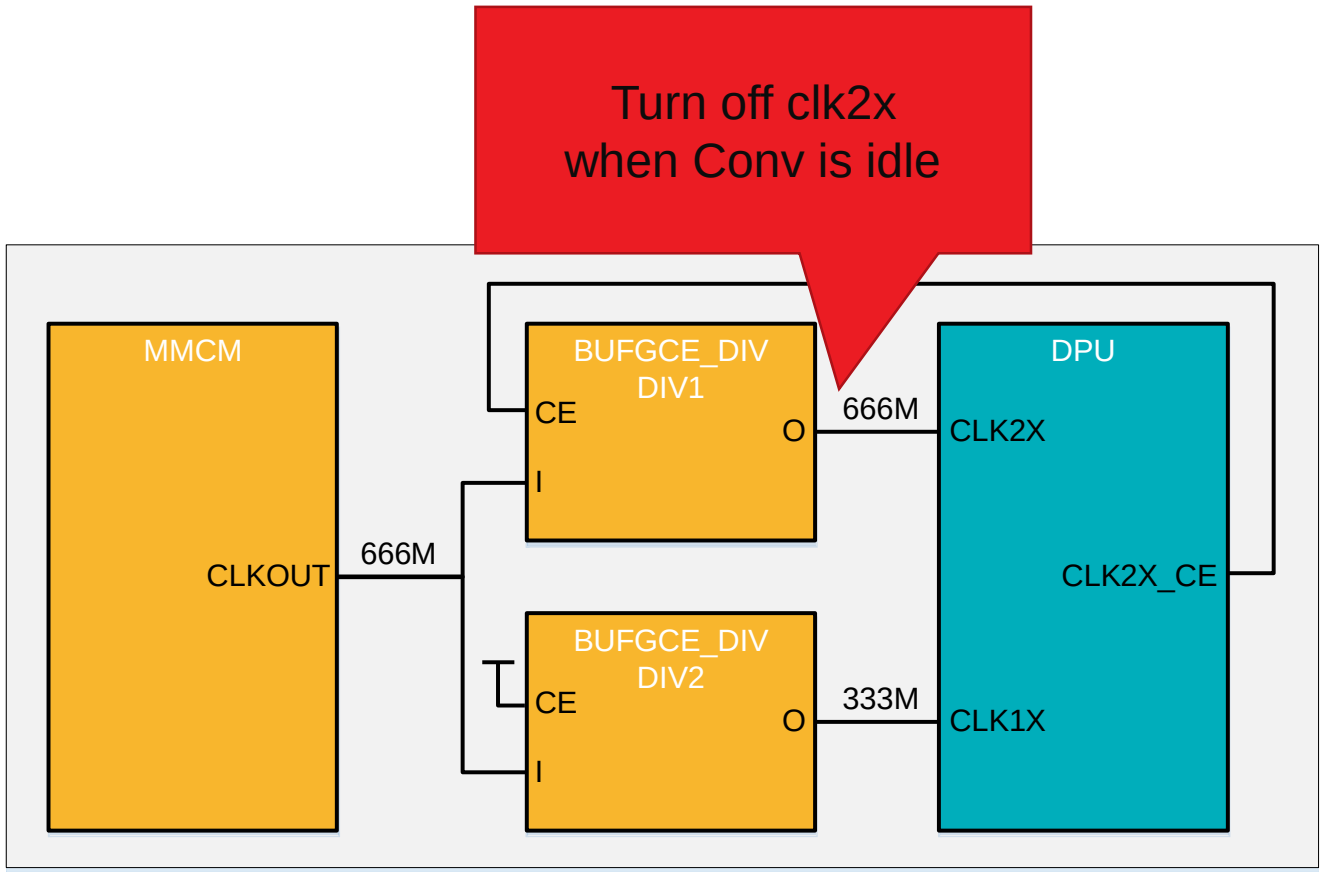

Original SDR-DSP



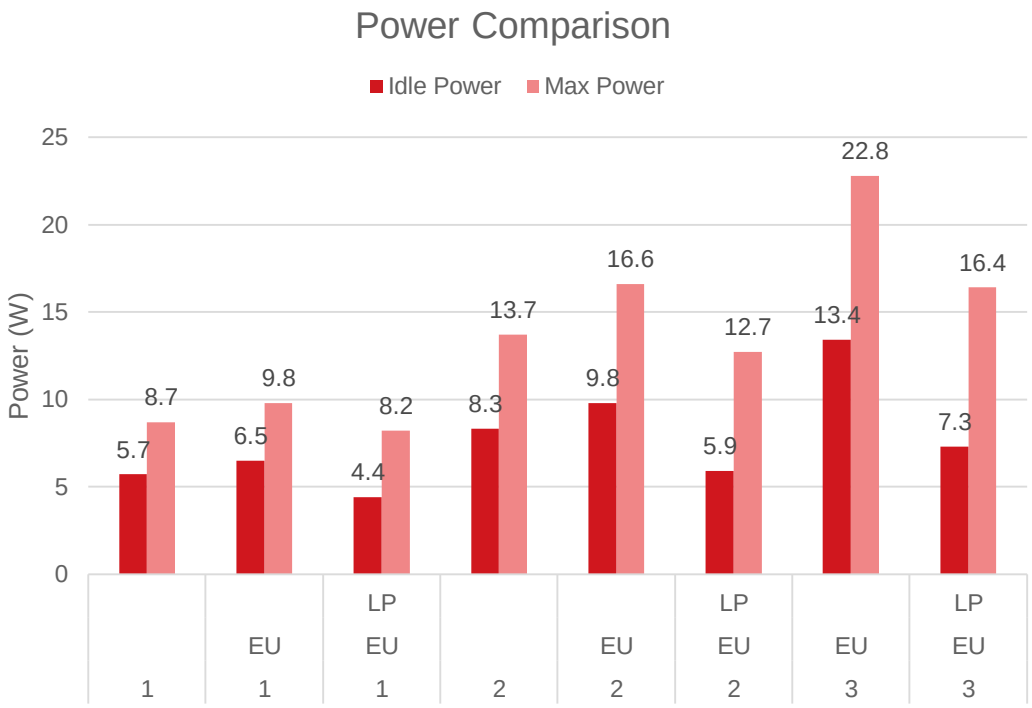

DDR-DSP

| Platform | Arch | LUT (SDR/DDR) | DSP (SDR/DDR) | PWR_MAX (SDR/DDR) | Freq |
|---|---|---|---|---|---|
| Zynq7020 | B1152 | 39230/27379 | 220/146 | -/- | 200/400MHz |
| ZU9EG | B4096 | 54748/40930 | 1026/514 | 8.7W/9.8W | 330/660MHz |

# DPU_EU_LP: Using the Gated Clock to Reduce Power




Enhanced clk2x to Gated clock




* Measured via Xilinx Zynq UltraScale+ MPSoc Power Management Tool on zcu102 B4096 when running Resnet50 @333MHz

***Reduce the power consumption by turning off the clock of computing modules when they are in the idle status.***

# DPU_EU: DSP vs LUT Preference

*We provide an option for customers depending on DSP vs LUT preference*

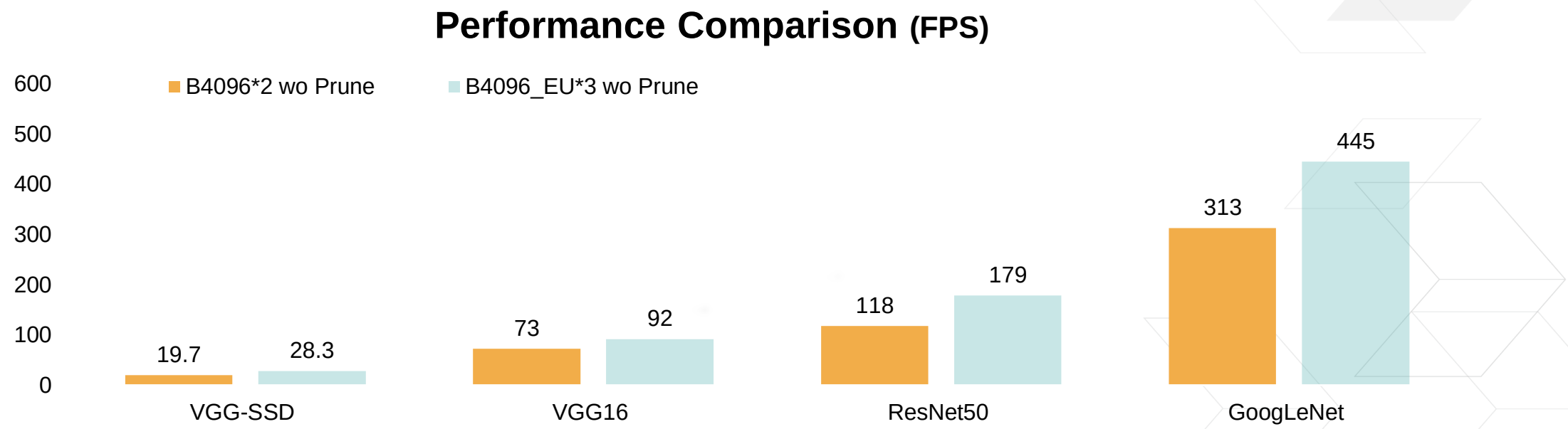
*More DSP*

| Arch | LUTs | Registers | BRAM | DSP |
|---|---|---|---|---|
| B512 | 20177 | 31782 | 69.5 | 98 |
| B800 | 19903 | 31673 | 87 | 142 |
| B1024 | 27377 | 46241 | 101.5 | 194 |
| B1152 | 28698 | 46906 | 117.5 | 194 |
| B1600 | 30877 | 56267 | 123 | 282 |
| B2304 | 34379 | 67481 | 161.5 | 386 |
| B3136 | 38555 | 79867 | 203.5 | 506 |
| B4096 | 40865 | 92630 | 249.5 | 642 |

*More LUT*

| Arch | LUTs | Registers | BRAM | DSP |
|---|---|---|---|---|
| B512 | 20759 | 33572 | 69.5 | 66 |
| B800 | 21050 | 33752 | 87 | 102 |
| B1024 | 29155 | 49823 | 101.5 | 130 |
| B1152 | 30043 | 49588 | 117.5 | 146 |
| B1600 | 33130 | 60739 | 123 | 202 |
| B2304 | 37055 | 72850 | 161.5 | 290 |
| B3136 | 41714 | 86132 | 203.5 | 394 |
| B4096 | 44583 | 99791 | 249.5 | 514 |

# Perf Improvement on ZCU102




*The FPS of VGG-SSD of end to end performance
*The FPS of VGG16/ResNet50/GoogLeNet is of CONV part (w/o FC layer)

## Resource Utilization Comparison

| | DSP | LUT | FF | BRAM | Frequency |
|---|---|---|---|---|---|
| B4096*2 | 2048 | 156744 | 224650 | 501 | 330MHz |
| B4096_EU*3 | 1926 | 110311 | 255020 | 748.5 | 330/660MHz |

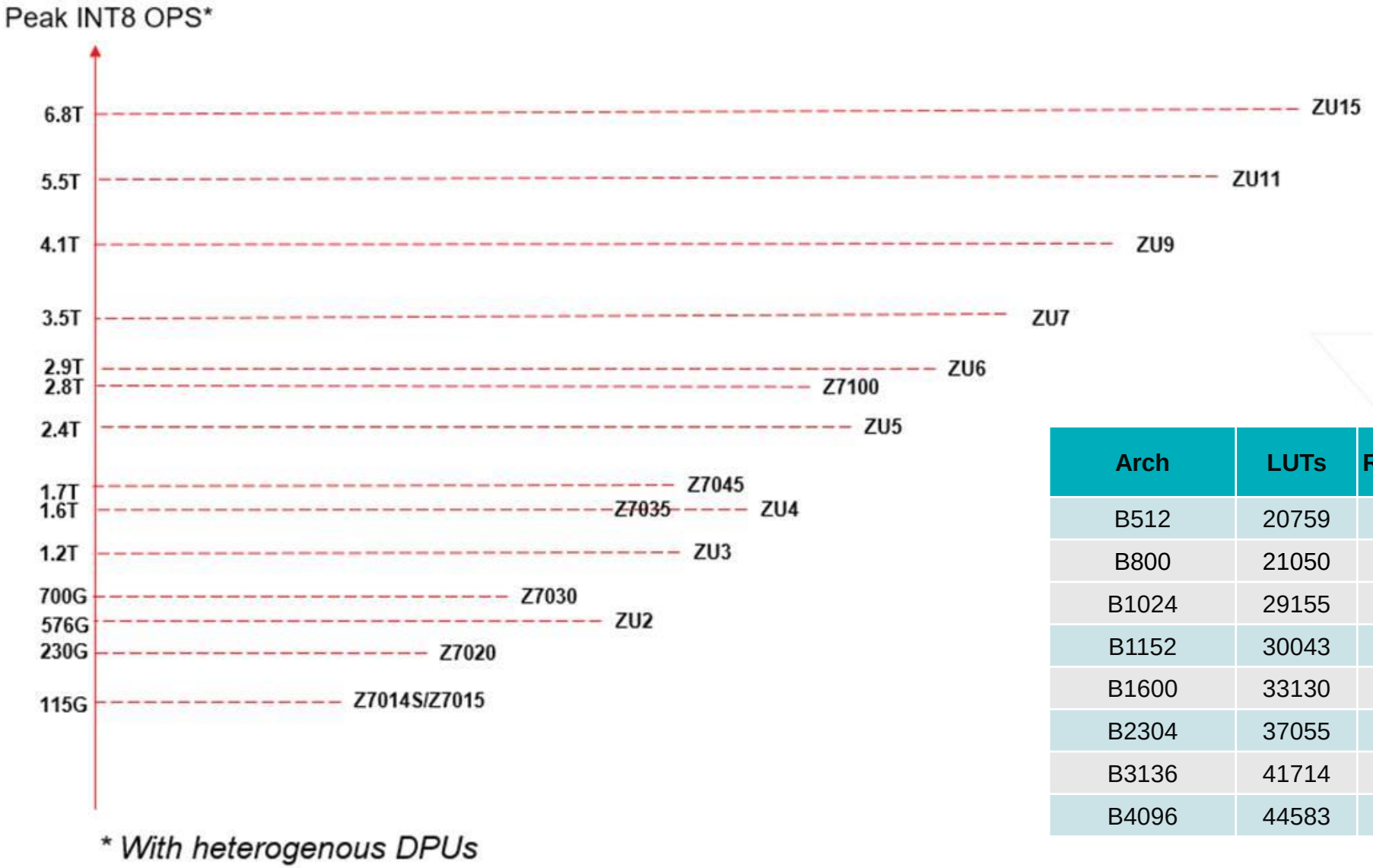
# End-to-end Solution of Different Models

| | Operation | Input Image | FPS |
|---|---|---|---|
| Inception-v1 | 3.2G | 224*224 | 405 |
| SqueezeNet | 0.698G | 224*224 | 1048 |
| Tiny-YOLO | 6.97G | 448*448 | 220 |
| YOLO-V2 | 82.5G | 640*640 | 24 |
| YOLO-V2[1)] | 18.4G | 640*640 | 120 |
| YOLO-V3 | 53.7G | 512*256 | 43 |
| YOLO-V3[1)] | 4G | 512*256 | 115 |

*1) DeePhi's deep compression technique is applied on these networks*

*3 x B4096_EU @330MHz on ZCU102*

# DPU Scalability




** With heterogenous DPUs*

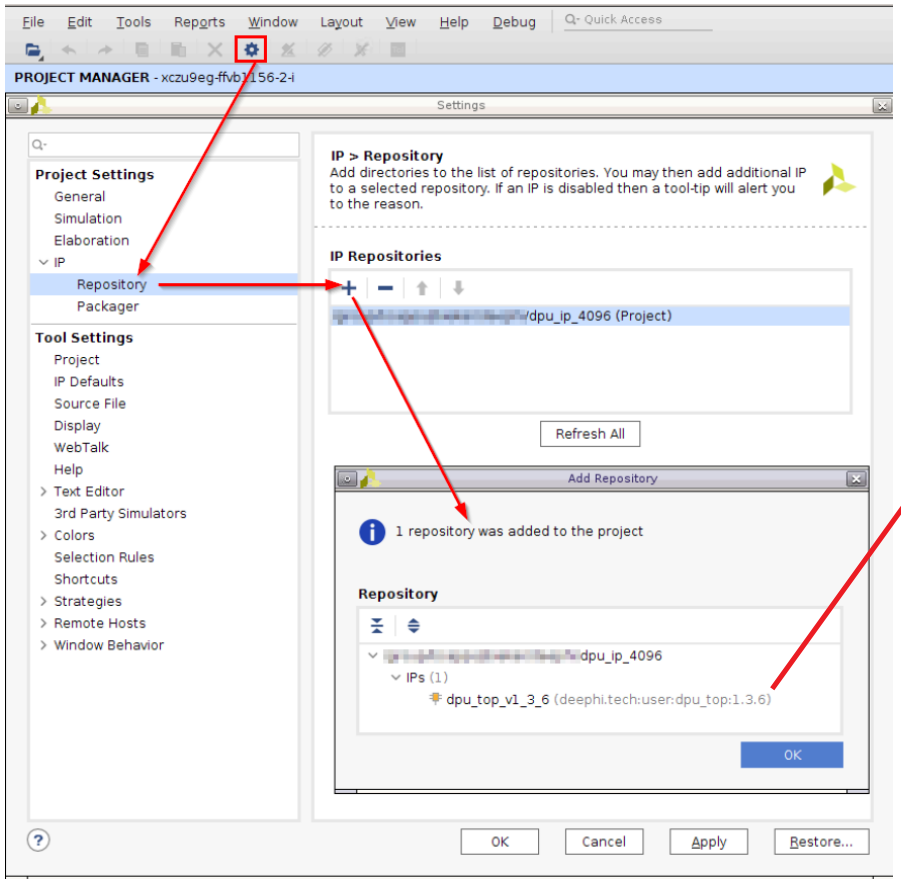
| Arch | LUTs | Registers | BRAM | DSP |
|---|---|---|---|---|
| B512 | 20759 | 33572 | 69.5 | 66 |
| B800 | 21050 | 33752 | 87 | 102 |
| B1024 | 29155 | 49823 | 101.5 | 130 |
| B1152 | 30043 | 49588 | 117.5 | 146 |
| B1600 | 33130 | 60739 | 123 | 202 |
| B2304 | 37055 | 72850 | 161.5 | 290 |
| B3136 | 41714 | 86132 | 203.5 | 394 |
| B4096 | 44583 | 99791 | 249.5 | 514 |

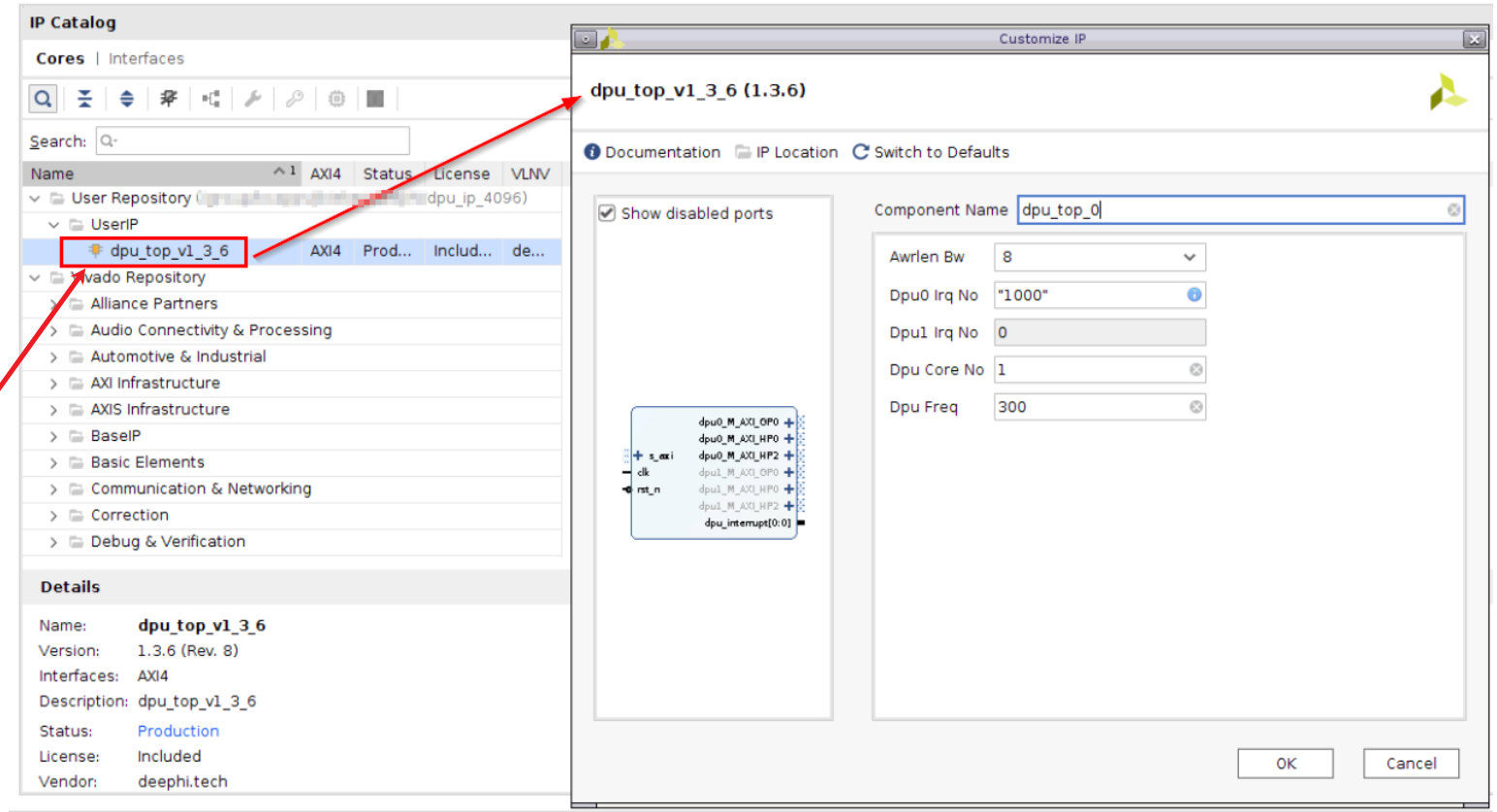
# DPU Design Integration

XILINX

# DPU Design Analysis

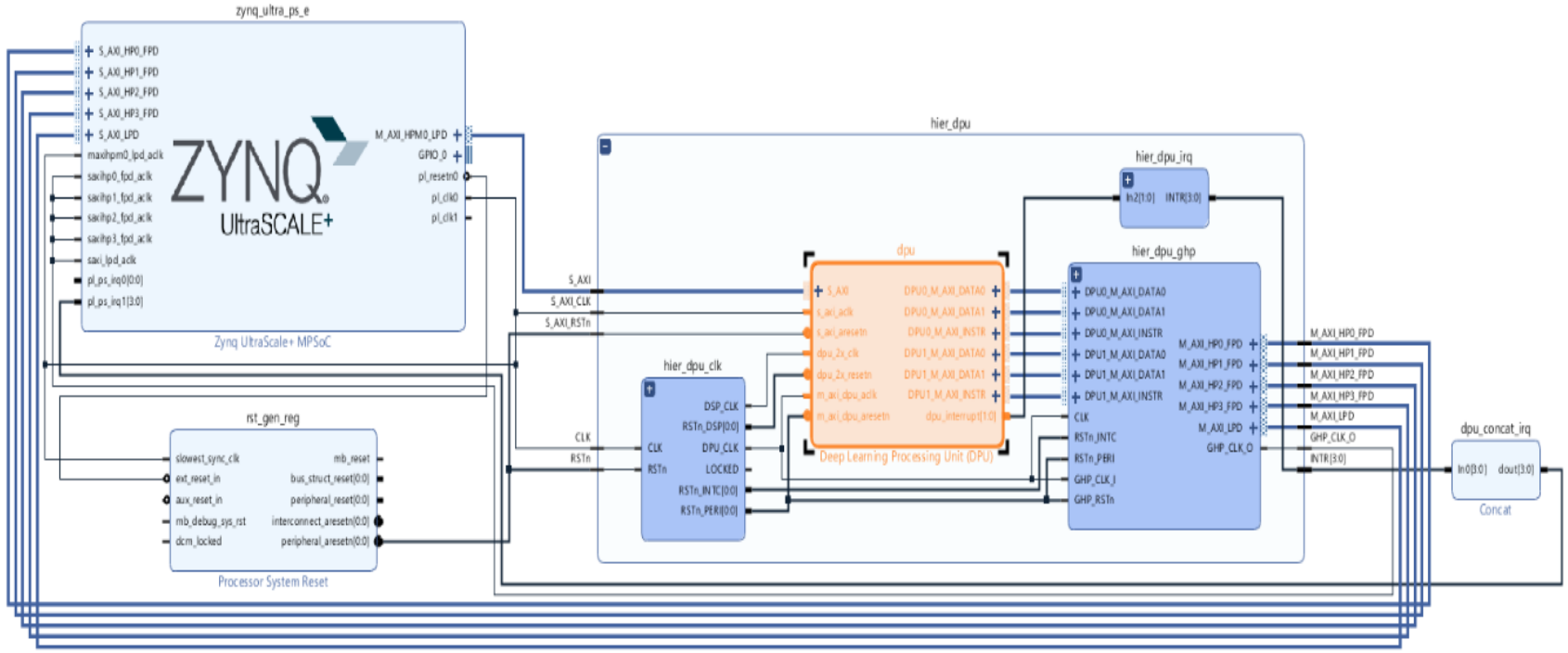
- **DPU repository pulled into IP Catalog**



# Zynq UltraScale+ MPSoC DPU TRD

- **Primary goal of this TRD is to demonstrate the capabilities of the DPU IP on Zynq UltraScale+ MPSoC devices. The TRD serves as a platform for the user to have a quick look at how to use the DPU IP. The TRD uses Vivado IP Integrator (IPI) flow for building the hardware design and Xilinx Yocto PetaLinux flow for software design.**




XILINX.

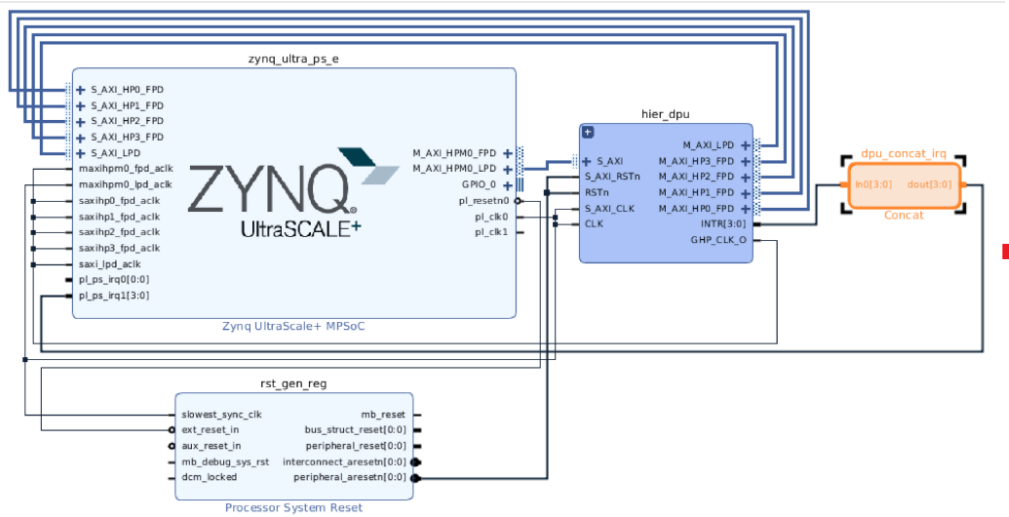
# Zynq UltraScale+ MPSoC DPU TRD

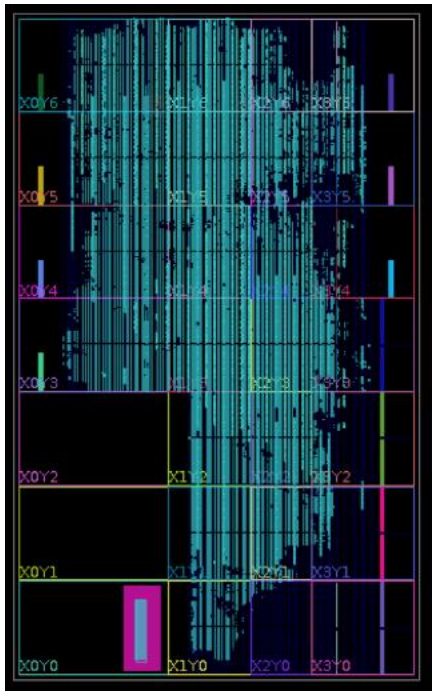
- **Device: xczu9eg-ffvb1156-2-I**
- **DPU: 2xB4096**
  - DPU Clock: 333 MHz, DSP Clock: 666 MHz
  - DPU0 Data Bus 2x128 => Zynq S_AXI_HP0/HP1
  - DPU1 Data Bus 2x128 => Zynq S_AXI_HP2/HP3
  - DPU00/1 Instr Bus => AXI Interconnect => Zynq S_AXI_LPD

| Site Type | Used | Fixed | Available | Util% |
|---|---|---|---|---|
| CLB LUTs | 78462 | 0 | 274080 | 28.63 |
| LUT as Logic | 70310 | 0 | 274080 | 25.65 |
| LUT as Memory | 8152 | 0 | 144000 | 5.66 |
| LUT as Distributed RAM | 3044 | 0 | | |
| LUT as Shift Register | 5108 | 0 | | |
| CLB Registers | 170558 | 0 | 548160 | 31.11 |
| Register as Flip Flop | 170558 | 0 | 548160 | 31.11 |
| Register as Latch | 0 | 0 | 548160 | 0.00 |
| CARRY8 | 3062 | 0 | 34260 | 8.94 |
| F7 Muxes | 3507 | 0 | 137040 | 2.56 |
| F8 Muxes | 45 | 0 | 68520 | 0.07 |
| F9 Muxes | 0 | 0 | 34260 | 0.00 |

| Site Type | Used | Fixed | Available | Util% |
|---|---|---|---|---|
| Block RAM Tile | 499 | 0 | 912 | 54.71 |
| RAMB36/FIFO* | 466 | 0 | 912 | 51.10 |
| RAMB36E2 only | 466 | | | |
| RAMB18 | 66 | 0 | 1824 | 3.62 |
| RAMB18E2 only | 66 | | | |

| Site Type | Used | Fixed | Available | Util% |
|---|---|---|---|---|
| DSPs | 1282 | 0 | 2520 | 50.87 |
| DSP48E2 only | 1282 | | | |



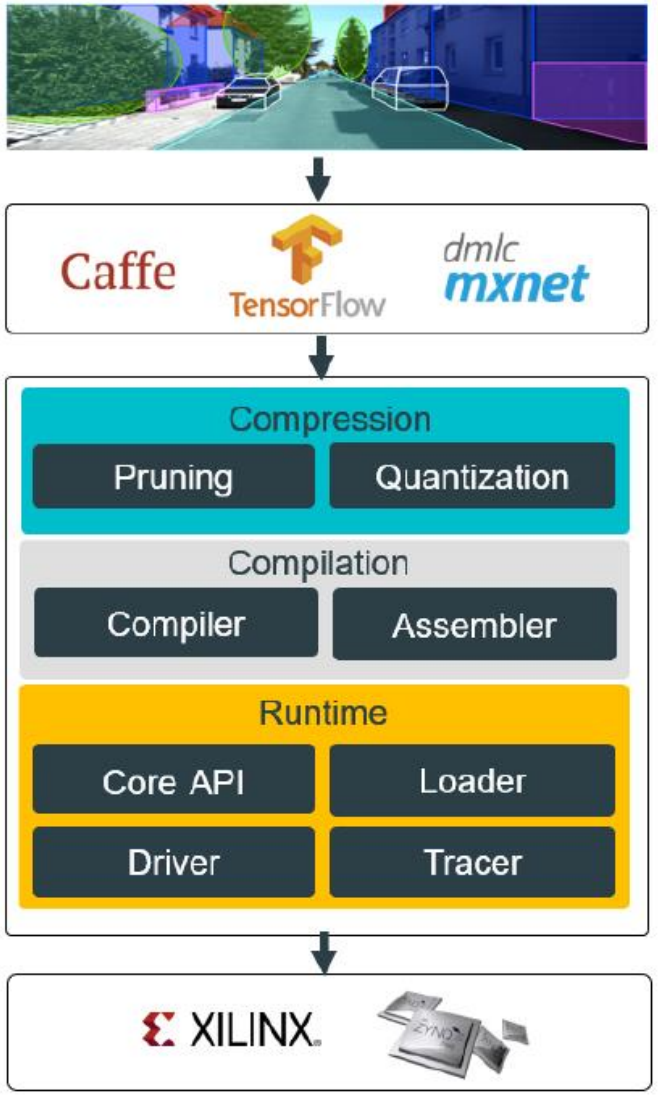




XILINX

# DNNDK

# DNNDK Tool Chain for Convolutional Networks

- **Deep Learning Solution Stack**
  - Compression
  - Compilation
  - Runtime tools
  - C++ APIs
  - DPU IP

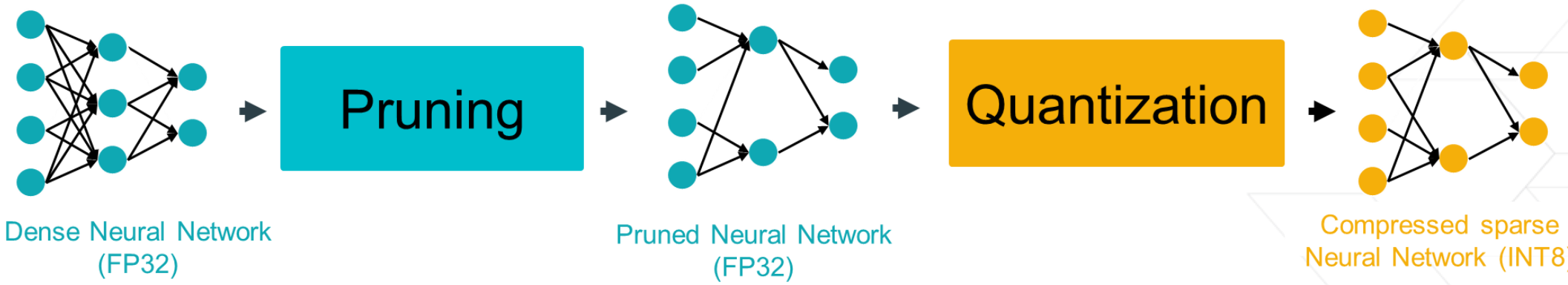


# Compression Tool

decent – Deep compression Tool

decent_q – Quantization Tool

decent_p – Pruning Tool



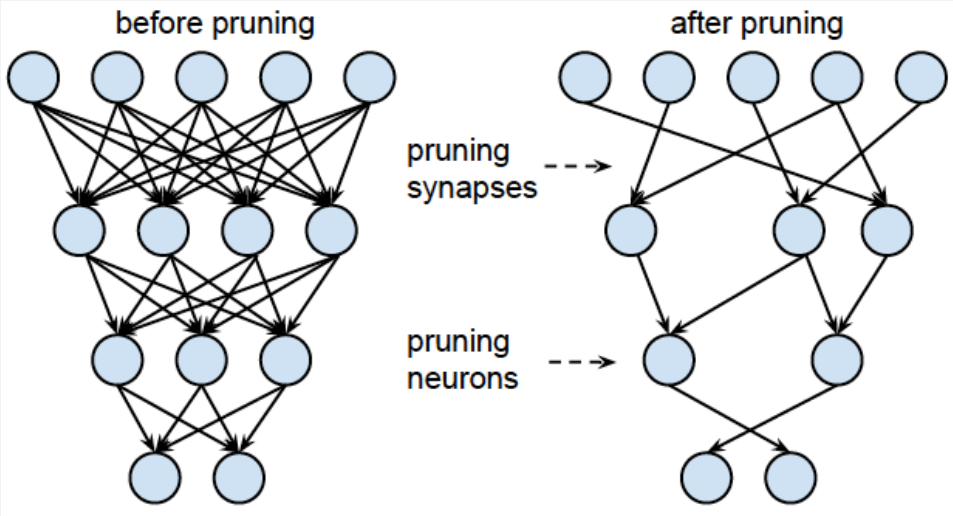

- Consists of two separate tools
  - Quantization Tool
  - Pruning Tool

- Effects
  - Compress model size 5x – 100x
  - Compress running time 1.5x – 10x

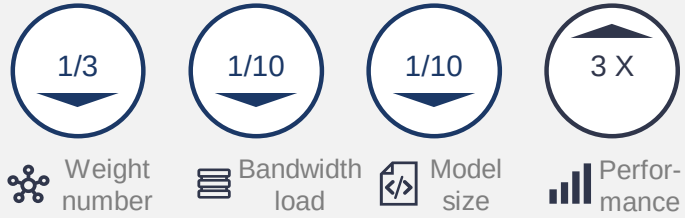
- Platform
  - Caffe, darknet
  - TensorFlow
    - Quantization Tool Beta version
    - Pruning Tool Internal version

# Core advantage | Deep compression algorithm

## Deep compression
## Makes algorithm smaller and lighter




**Highlight**




**Compression efficiency**

Deep Compression Tool can achieve significant compression on **CNN** and **RNN**

**Accuracy**

Algorithm can be **compressed 7 times without losing accuracy** under SSD object detection framework

**Easy to use**

Simple software development kit need only **50 lines of code** to run ResNet-50 network

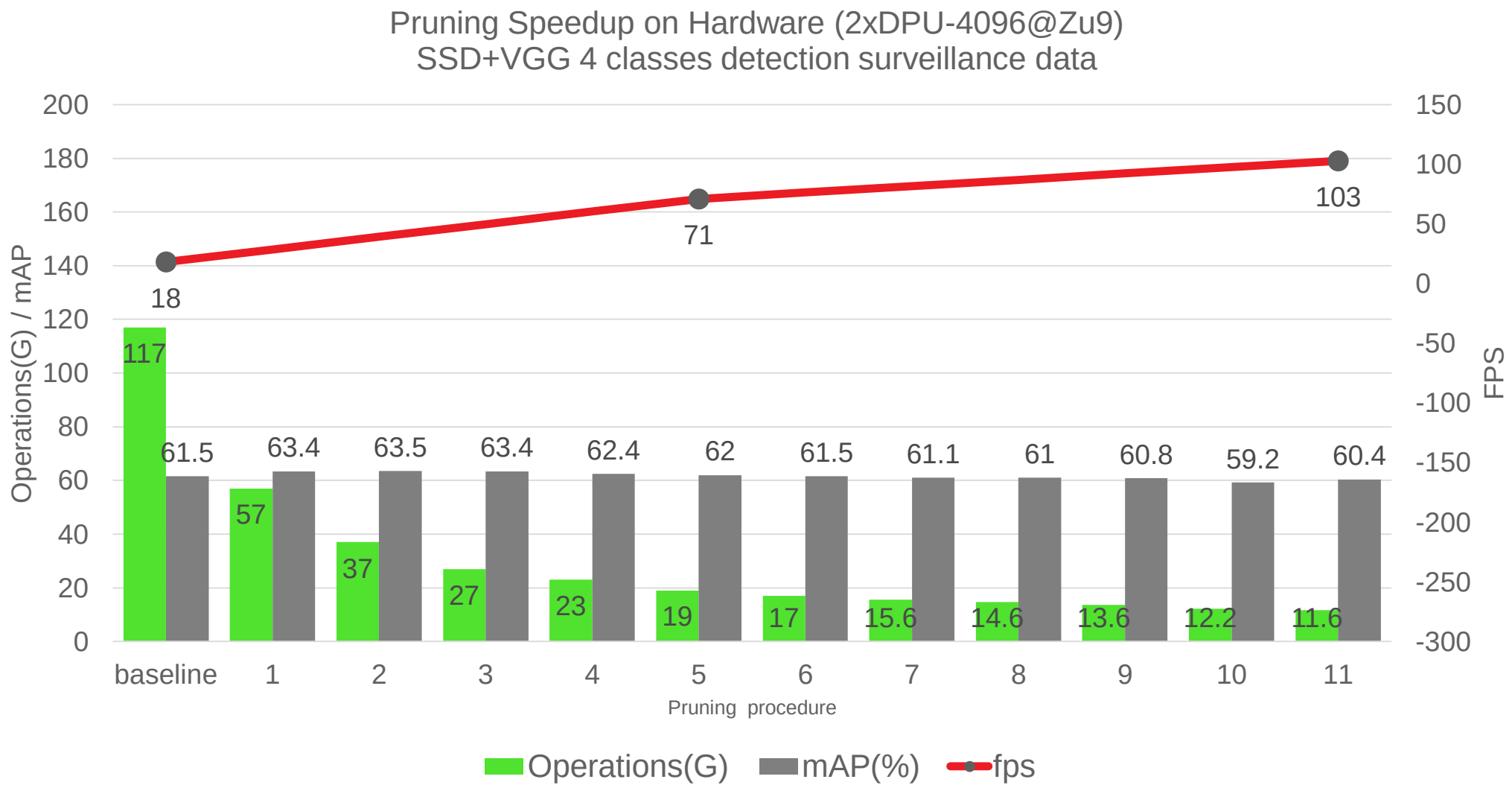
# Pruning Results

| Classification Networks | Baseline | Pruning Result 1 | | | Pruning Result 2 | | |
|---|---|---|---|---|---|---|---|
| | Top-5 | Top-5 | ΔTop5 | ratio | Top-5 | ΔTop5 | ratio |
| Resnet50 [7.7G] | 91.65% | 91.23% | -0.42% | 40% | 90.79% | -0.86% | 32% |
| Inception_v1 [3.2G] | 89.60% | 89.02% | -0.58% | 80% | 88.58% | -1.02% | 72% |
| Inception_v2 [4.0G] | 91.07% | 90.37% | -0.70% | 60% | 90.07% | -1.00% | 55% |
| SqueezeNet [778M] | 83.19% | 82.46% | -0.73% | 89% | 81.57% | -1.62% | 75% |

| Detection Networks | Baseline mAP | Pruning Result 1 | | | Pruning Result 2 | | |
|---|---|---|---|---|---|---|---|
| | | mAP | ΔmAP | ratio | mAP | ΔmAP | ratio |
| DetectNet [17.5G] | 44.46 | 45.7 | +1.24 | 63% | 45.12 | +0.66 | 50% |
| SSD+VGG [ 117G] | 61.5 | 62.0 | +0.5 | 16% | 60.4 | -1.1 | 10% |
| [A] SSD+VGG [ 173G] | 57.1 | 58.7 | +1.6 | 40% | 56.6 | -0.5 | 12% |
| [B] Yolov2 [ 198G] | 80.4 | 81.9 | +1.5 | 28% | 79.2 | -1.2 | 7% |

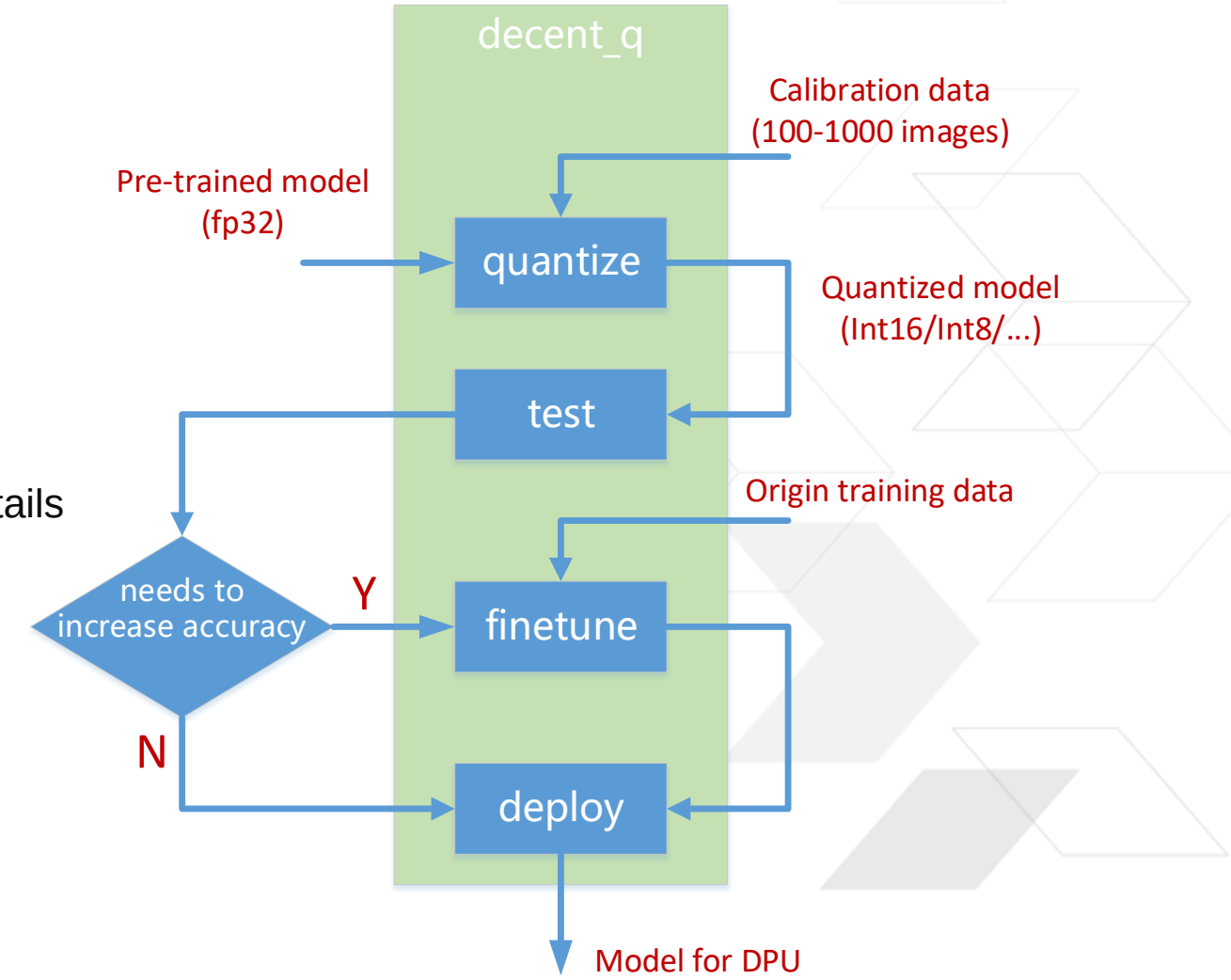
| Segmentation Networks | Baseline | Pruning Result 1 | | | Pruning Result 2 | | |
|---|---|---|---|---|---|---|---|
| | mIoU | mIoU | ΔmIoU | ratio | mIoU | ΔmIoU | ratio |
| FPN [163G] | 65.69% | 65.21% | -0.48% | 80% | 64.07% | -1.62% | 60% |

# Pruning Speedup Example – SSD



# Quantization Tool – decent

- **4 steps in decent quantization**
  - quantize – quantize network
    - Calibration images required
  - test – test network accuracy/mAP
  - finetune – finetune quantized network
    - Usually not needed
    - Requires entire training data set
    - Not documented contact factory for more details
  - deploy – generate model for DPU
    - This is input to the dnnc compiler
- **Data**
  - Calibration data – quantize activation
  - Training data – further increase accuracy



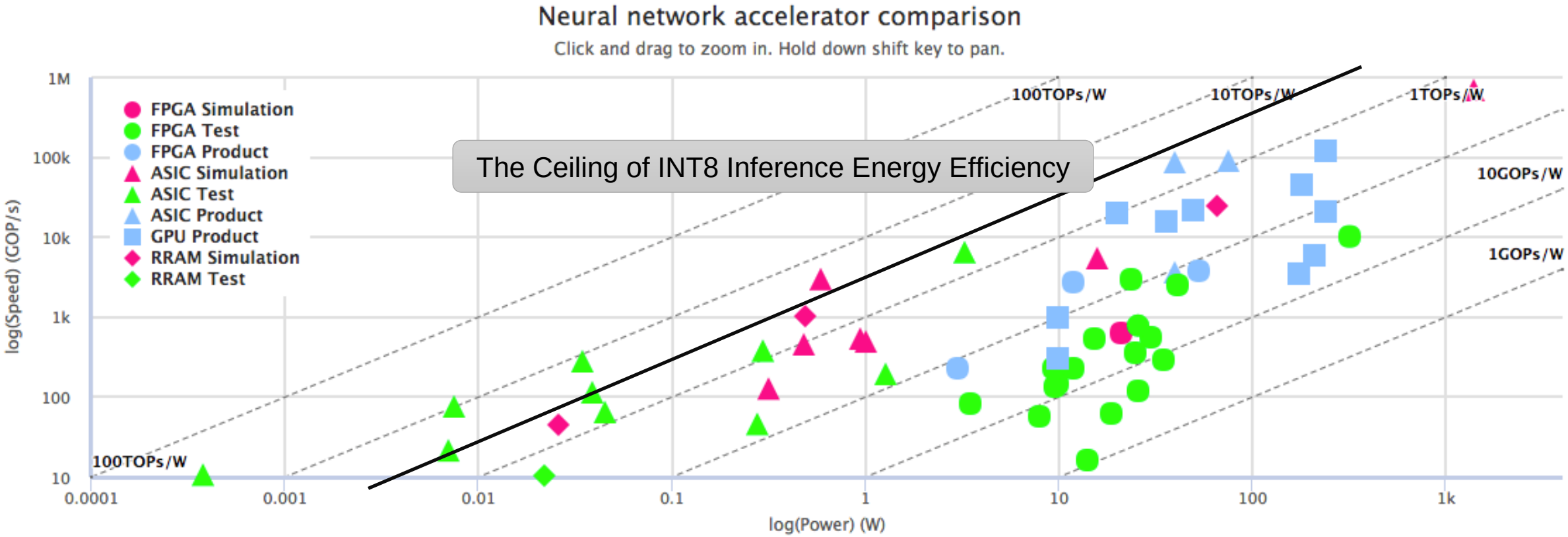
# Quantization Results for Popular Networks

| Classification | float | | 8-bit fix | |
|---|---|---|---|---|
| | Top1 | Top5 | ΔTop1 | ΔTop5 |
| Inception_v1 | 66.90% | 87.68% | -0.28% | -0.10% |
| Inception_v2 | 72.78% | 91.04% | -0.38% | -0.23% |
| Inception_v3 | 77.01% | 93.29% | -0.45% | -0.29% |
| Inception_v4 | 79.74% | 94.80% | -0.32% | -0.16% |
| ResNet-50 | 74.76% | 92.09% | -0.17% | -0.14% |
| ResNet-50-v2 | 75.39% | 92.45% | -0.60% | -0.33% |
| VGG16-3fc-float | 70.97% | 89.85% | -0.23% | -0.06% |
| VGG16-1fc-float | 70.97% | 89.85% | -0.20% | -0.09% |
| Inception-ResNet-v2 | 79.95% | 95.13% | -0.51% | -0.16% |
| **Detection** | Float mAP | | 8-bit fix mAP | |
| SSD_VGG | 76.47% | | -0.20% | |

# Network Precision

XILINX

# Current Ceiling of CNN Architecture




*Source:http://nics-efc.org/projects/neural-network-accelerator/*

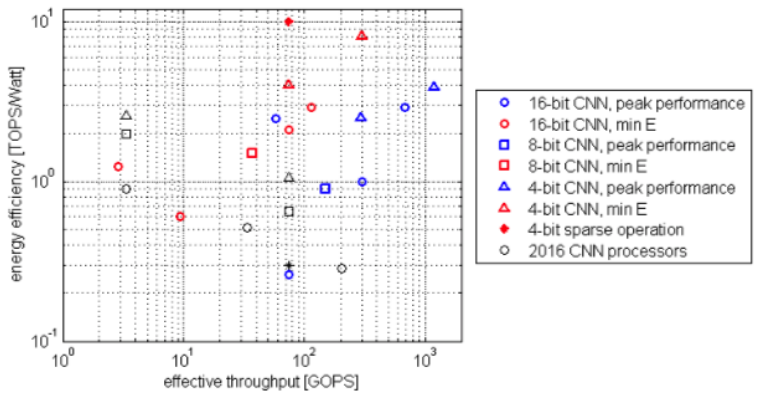
**INT8 improvements are slowing down and approaching the ceiling.**

**Solutions**

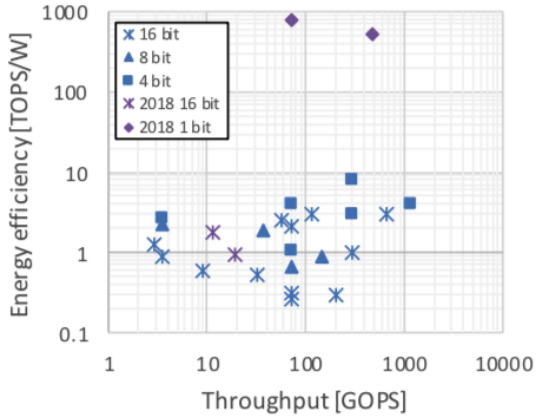
- **Sparsity**
- **Low Precision**

XILINX

# Potentials of low precision




ISSCC,2017




ISSCC,2018

- Scales performance
- Reduces hardware resources
- Less bandwidth/on-chip memory requirement
- Regular memory access pattern and calculating pattern

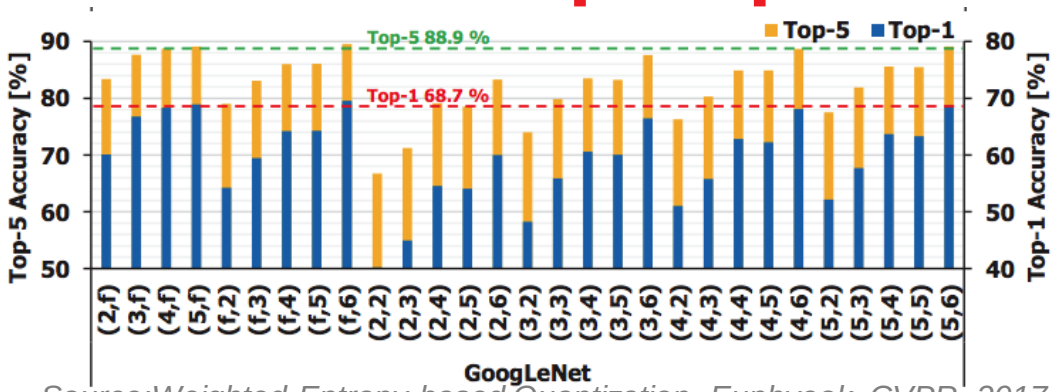
FPGA benefits a lot from low-precision.

## Low Precision Becomes Popular

| Energy Cost | |
|---|---|
| Operation | Energy(pJ) |
| 1bit Fixed-point MAC | 0.118 |
| 4bit Fixed-point MAC | 0.517 |
| 8bit Fixed-point MAC | 0.865 |
| 16bit Fixed-point MAC | 1.64 |
| *65nm process,200Mhz,1.2v,25°C | |

| Model Size(ResNet-50) | |
|---|---|
| Precision | Size(MB) |
| 1b | 3.2 |
| 8b | 25.5 |
| 32b | 102.5 |

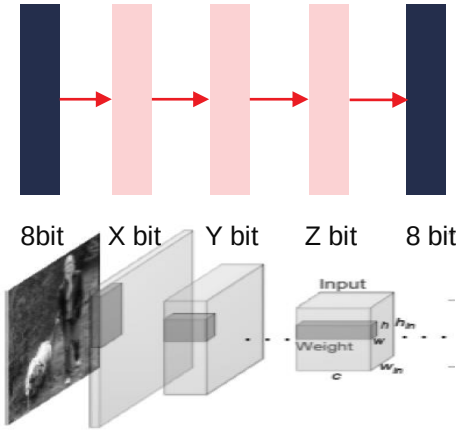
# Architecture perspective: Mixed Low-Precision

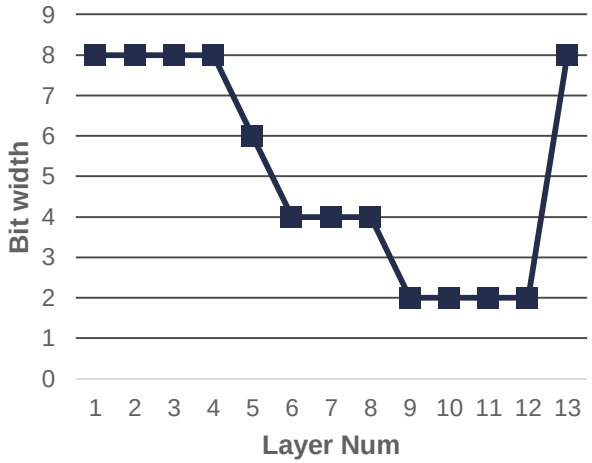


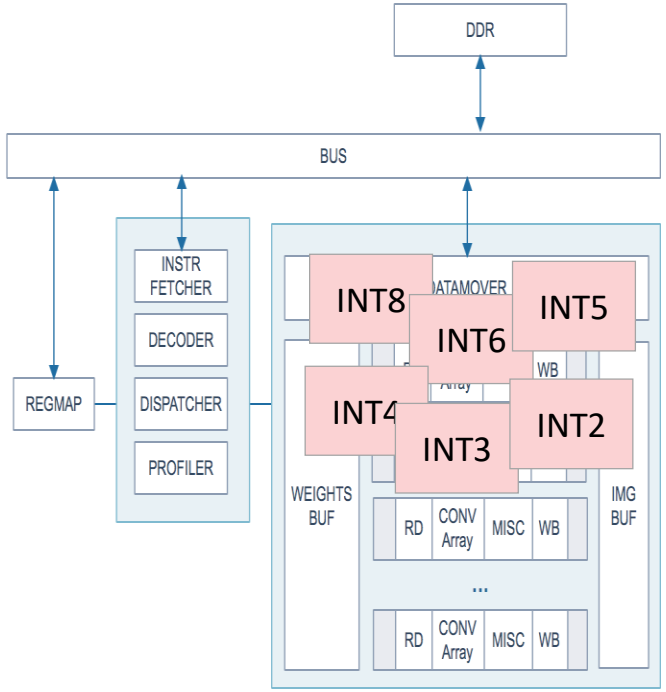

*Source:Weighted-Entropy-based Quantization, Eunhyeok, CVPR, 2017-*

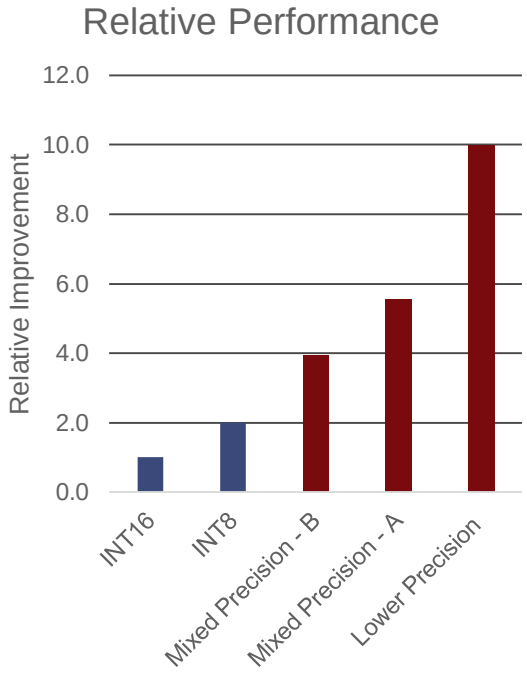
**Fixed** low-precision quantization already showed competitive results.

Next generation: **Variable** precision of activation/weights among layers



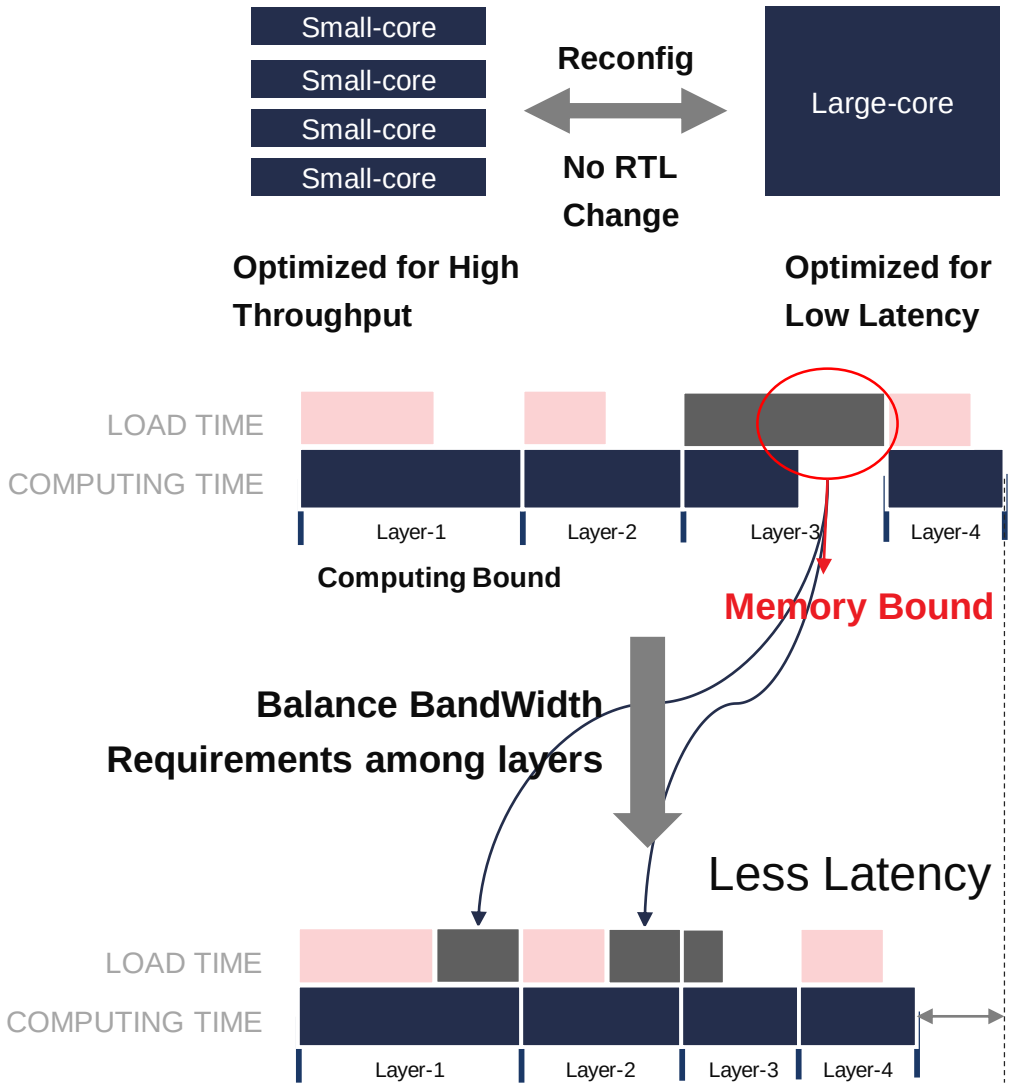




XILINX

# Architecture perspective: Mixed Low-Precision CNN

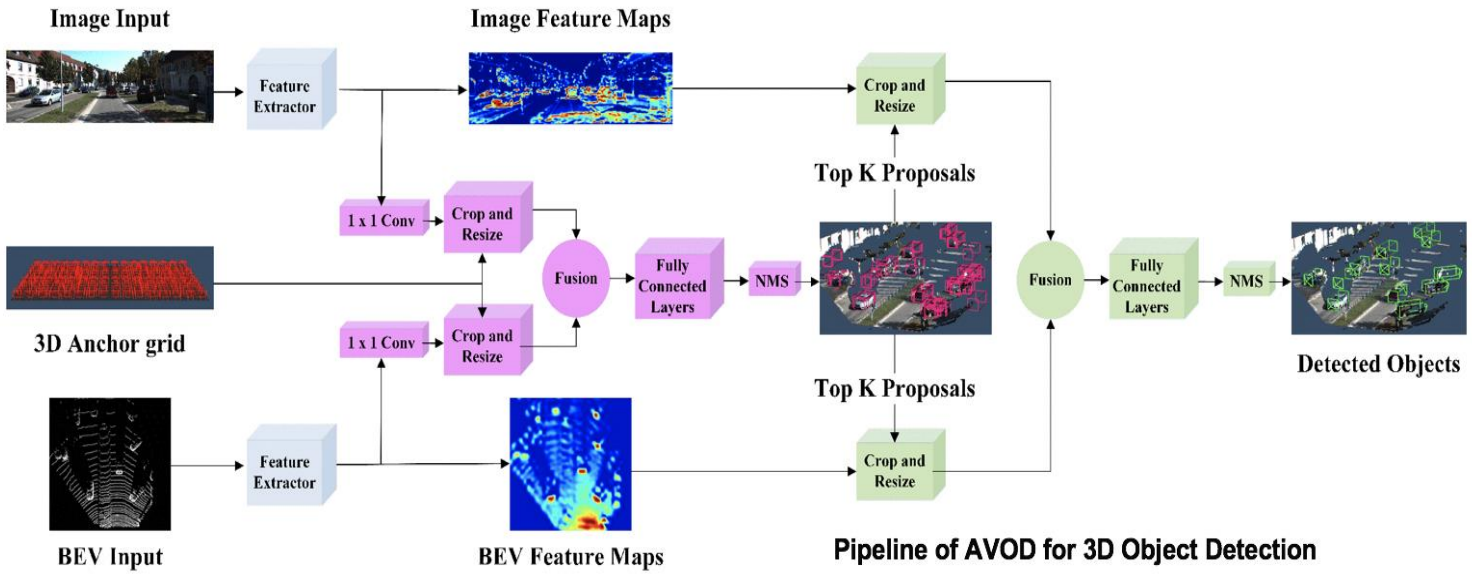
- **Mixed Precision Support**
  - INT8/6/5/4/3/2
- **Flexible Between Throughput and Latency**
  - Switch between Throughput-Opt-Mode and Latency-Opt-Mode without RTL change
- **Enhanced Dataflow Techniques**
  - Make the balance among different layers. Do NOT require the model can be fully placed on chip, but load the data at the right time.
  - Physical-aware data flow design to meet higher frequency.
  - Supports high-resolution images at high utilization.
- **Performance Target** (googlenet_v1)
  - 3103 FPS (INT8)
  - 5320 FPS (INT8/4/2 mixed)
  - 12412 FPS (INT2 only)



# Application

# System perspective: schedule ADAS tasks in single FPGA

- **Multi-task Models**
  - Training:
    - Knowledge sharing
    - Reduce computation cost
  - Pruning:
    - Balance different objective functions
- **Sensor Fusion**
  - Sensor alignment & Data Fusion
- **Task scheduling**
  - Resource constrained scheduling: Serialization & Parallelization
  - Task scheduling and memory management framework with low context-switching cost
  - Support new operations with runtime variable parameter by software and hardware co-design



# System perspective: Video Surveillance in single FPGA

- **Platform：ZU4EV**
- **DPU：B2304_EU**
- **Peak perf.: 921Gops (400Mhz)**
- **Power: 7.7W (XPE)**

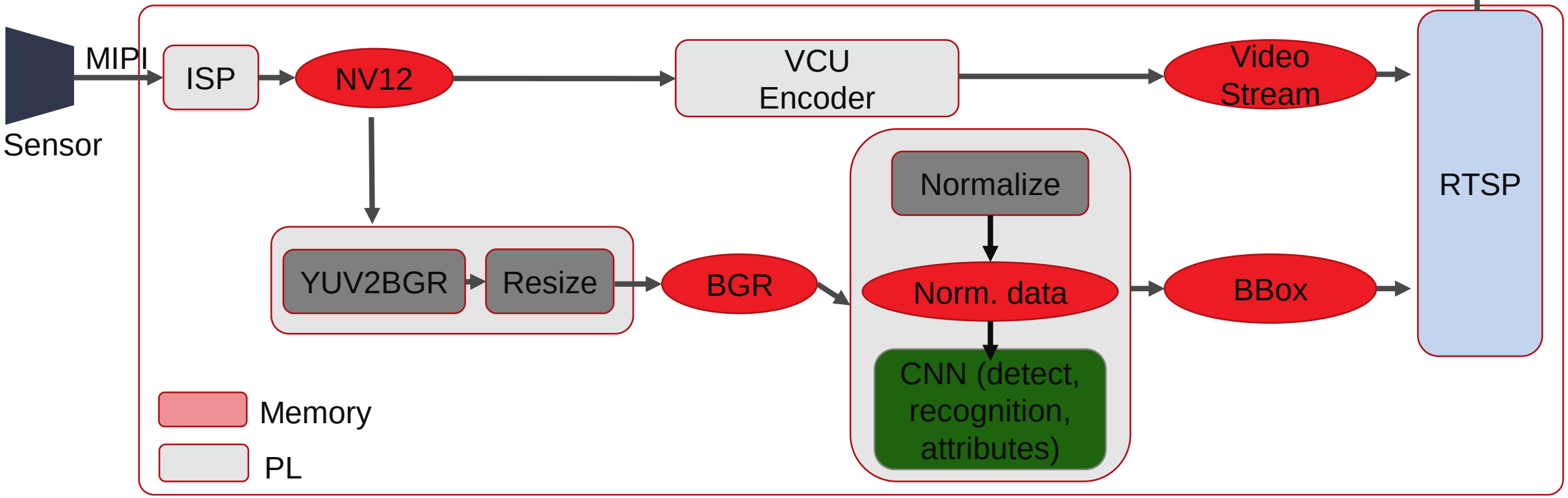
ML+X

Computer

HUB

Ethernet

*Single Chip Solution*




This solution needs to further enhance ISP functionality

# Adaptable.
# Intelligent.