DEEPHI 深鉴科技 Now Part of XILINX

Xilinx AI Solutions with DeePhi Technologies

Song Yao
Former Founder and CEO of DeePhi Tech
Senior Director of AI Business
songyao@xilinx.com

XILINX

DeePhi Completely Merged into Xilinx in Sep 2018

DeePhi Becomes Xilinx AI Center

Song Yao
Former Founder and CEO of DeePhi
Senior Director of AI Business

Yi Shan
Former CTO of DeePhi
Senior Director of AI R&D

(Full-time Team)

Prof. Yu Wang
Professor
Department of Electronic Engineering,
Tsinghua University

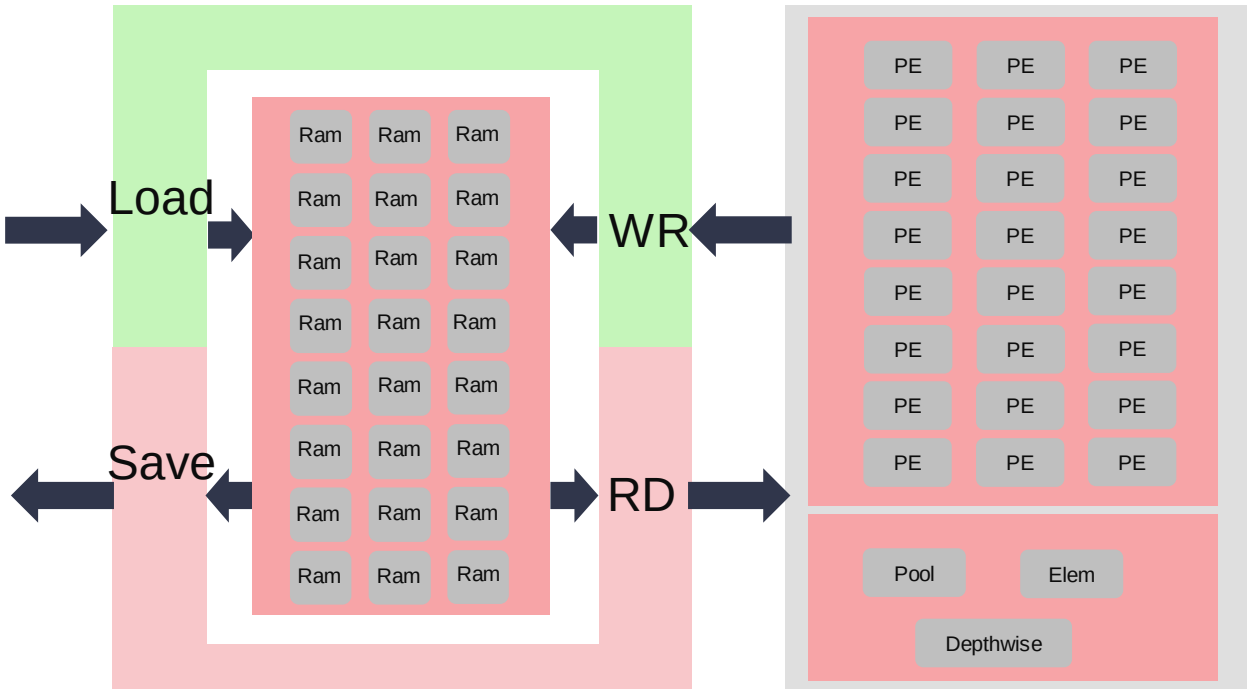
Prof. Song Han
Assistant Professor
EECS,
MIT

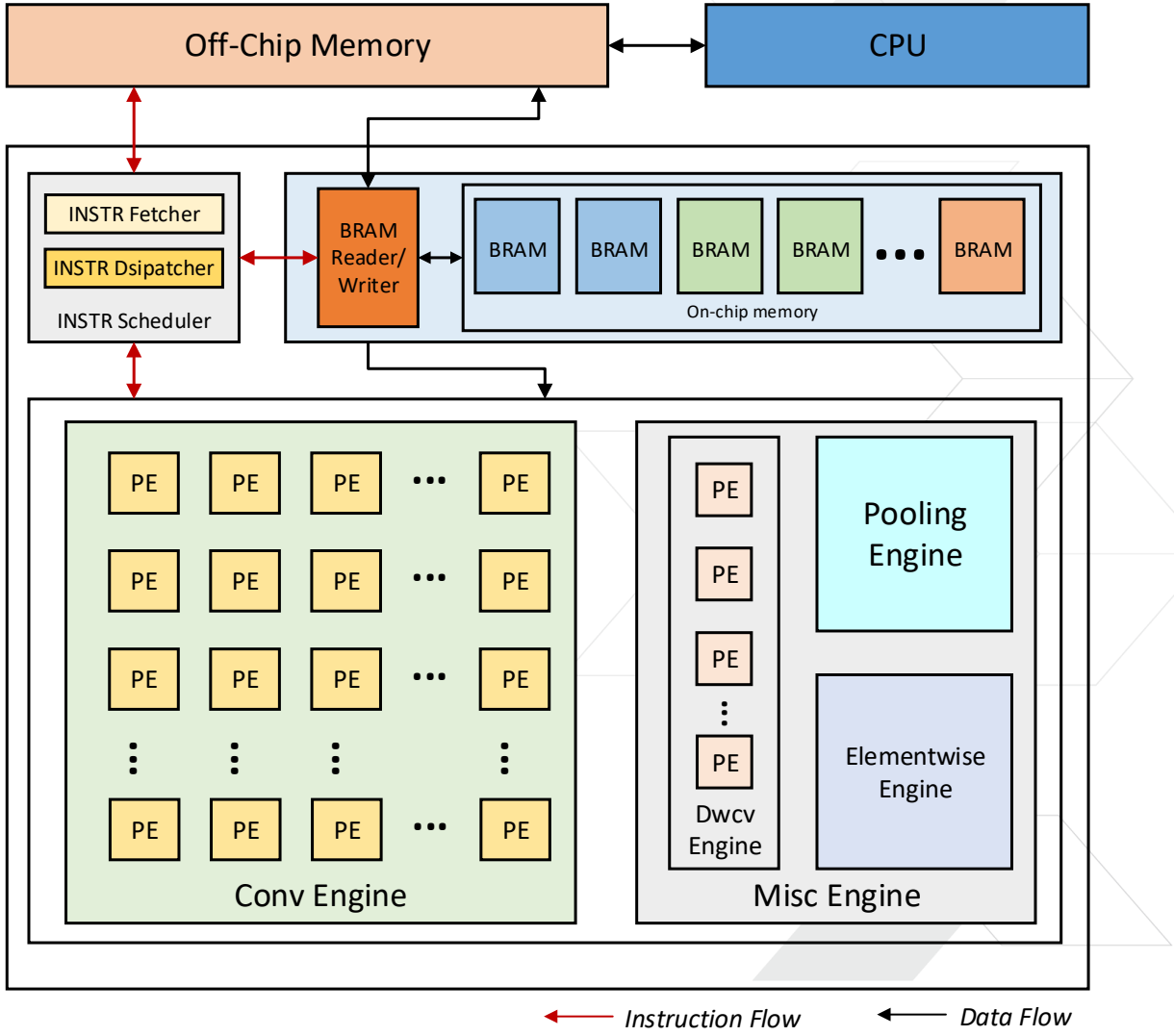
(Consultants)

Architecture Updates

XILINX

Hardware Architecture



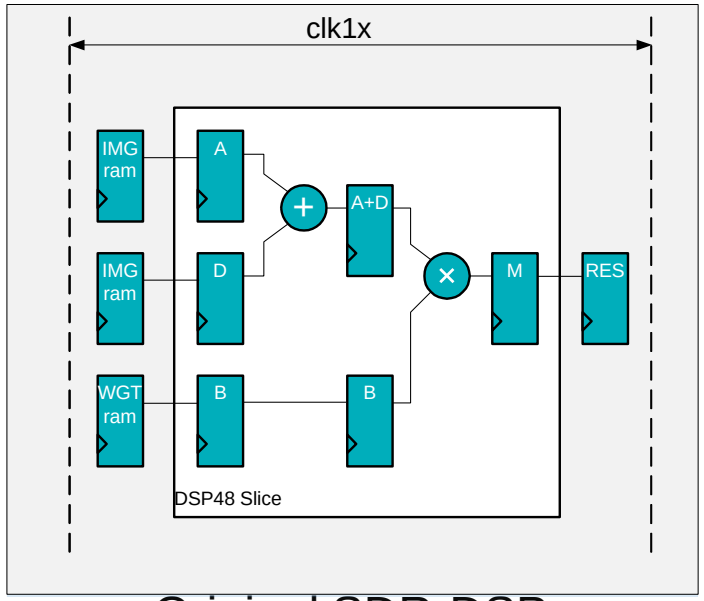


XILINX

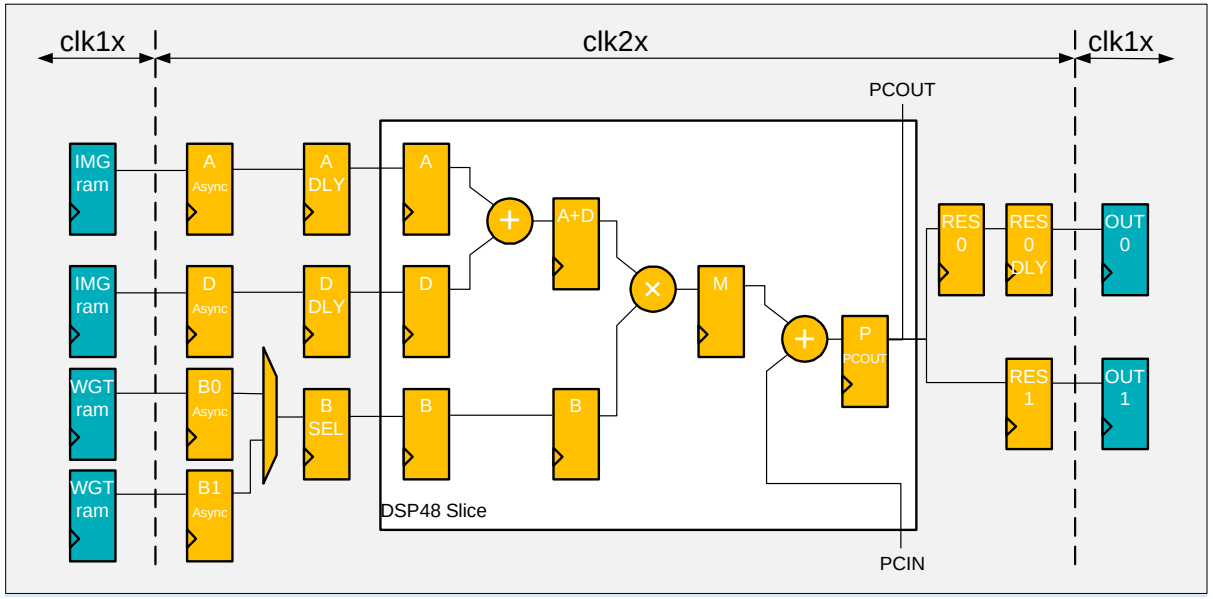
Supported Operators

- Arbitrary Input Image Size
- Conv
  - Arbitrary Conv Kernel Size
  - Arbitrary Conv Stride/Padding
  - Dilation
- Pooling
  - Avg/Max Pooling
    - Avg Pooling kernel size: 2x2~7x7
  - Arbitrary Max Pooling Size
  - Arbitrary Pooling Stride/Padding
- ReLU / pRelu / Relu6 / Leaky Relu(optional)
- Concat
- Split
- Elementwise
- FC(Int8/FP32) (optional)
- Mean scale
- Resize(optional)
- Deconv
- Depthwise conv
- BatchNormalization
- Reorg
- Softmax (optional)
- Sigmoid (optional)

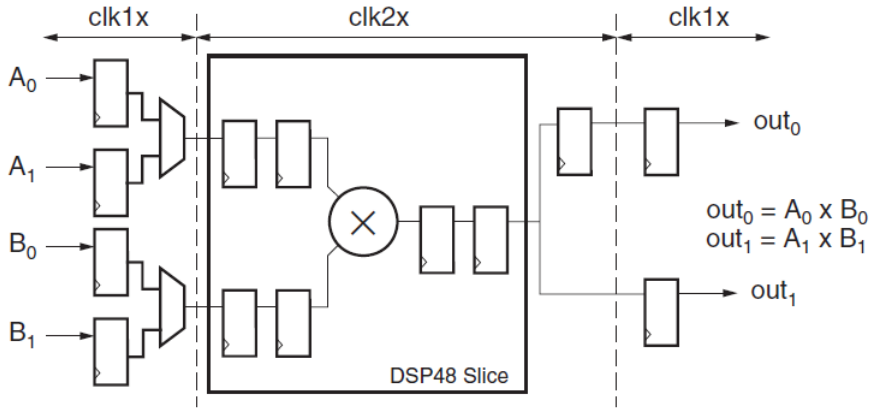
Double-Data-Rate DSP



Original SDR-DSP



Enhanced to DDR-DSP



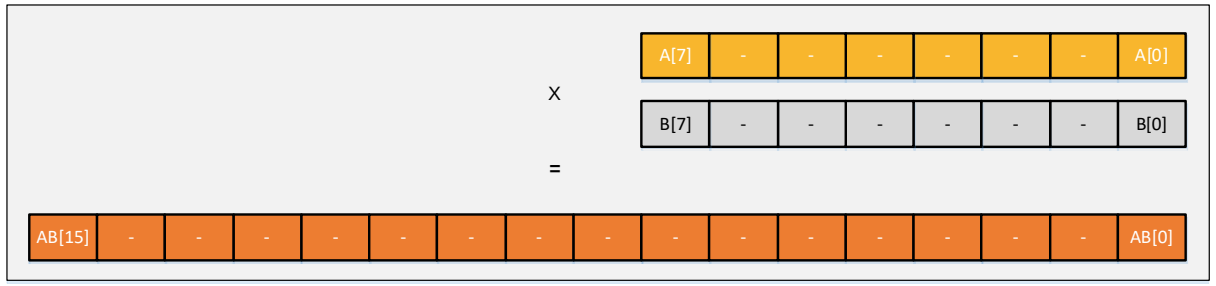
XAPP706[1]

LESS ultilization, but MORE power consumption

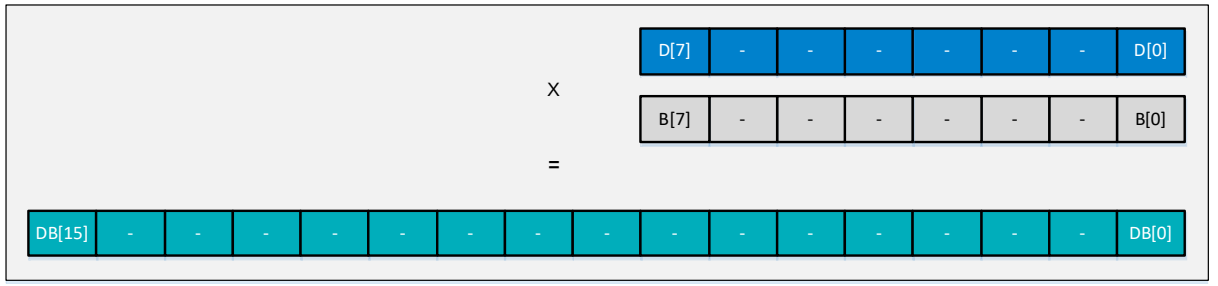
Board	Arch	LUT (SDR/DDR)	DSP (SDR/DDR)	PWR_MAX (SDR/DDR)
dpb1100v102	B1152	39230/27379	220/146	-/-
zcu102v100	B4096	54748/40930	1026/514	8.7W/9.8W

[1] Reed P. Tidwell. XAPP706 (v1.0) Alpha Blending Two Data Streams Using a DSP48 DDR Technique. Xilinx Inc., March 31, 2005.

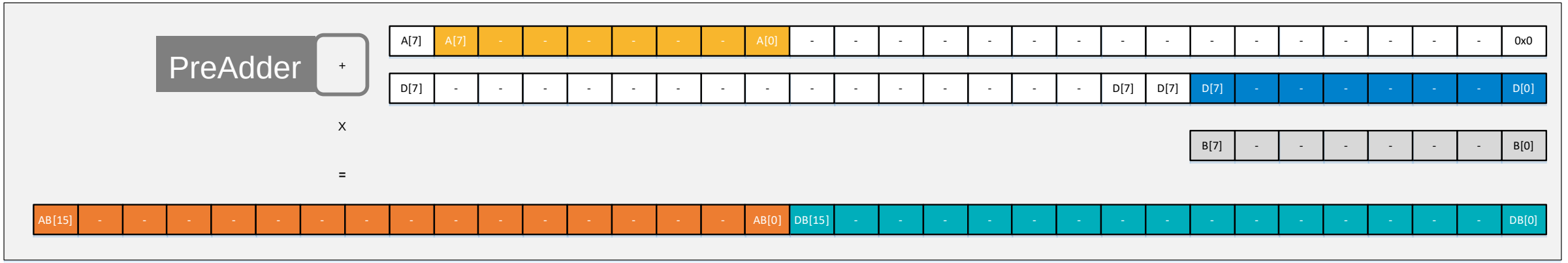
Packing Two INT8 Mults into One DSP48E1 Slice



AB



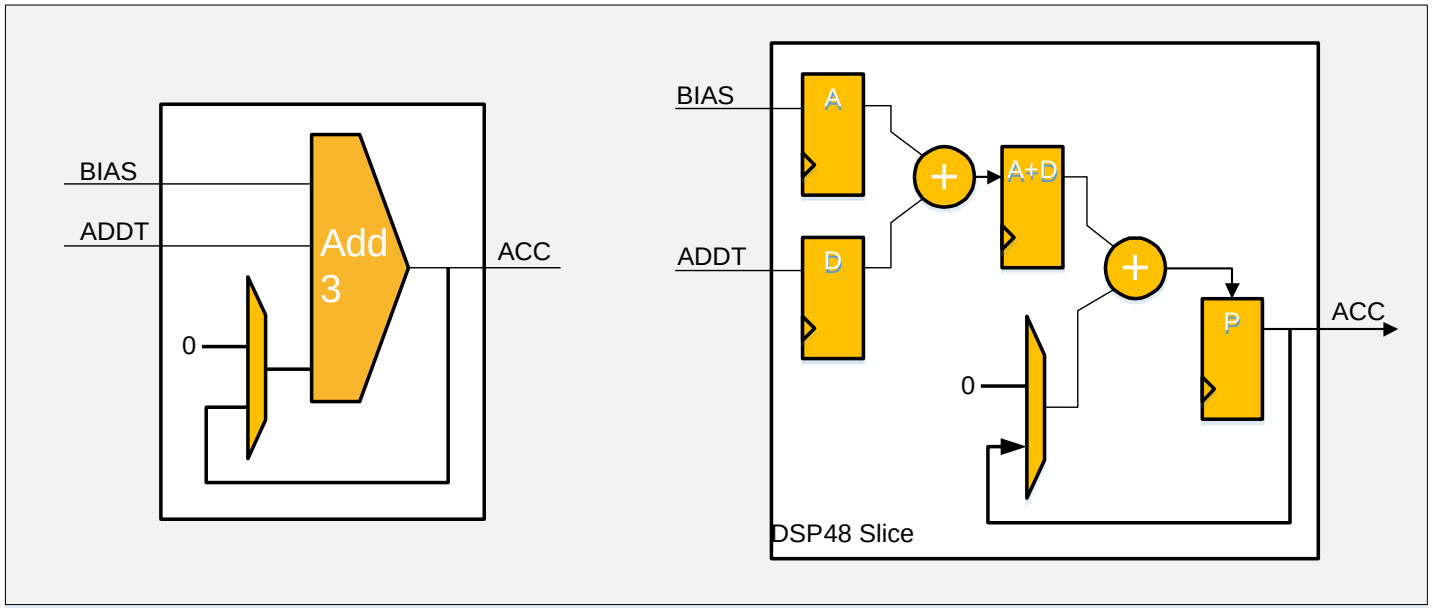
DB



(A*2^16+D)*B=AB*2^16+DB

XILINX

Accumulator via DSP



Original
Accu. via LUT

Enhanced to
Accu. via DSP[3][4]

Ability to compromise between
LUT and DSP resources

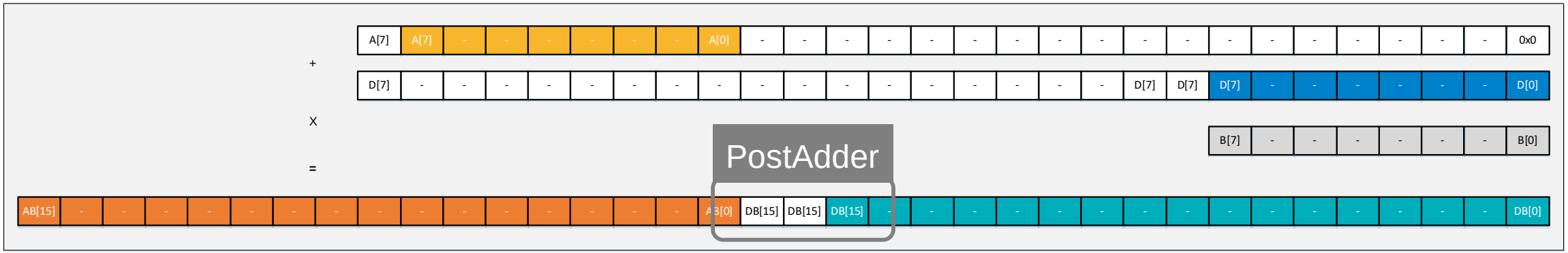
AdderTree via LUT could also be enhanced to DSP

Less ultilization, but a bit larger latency

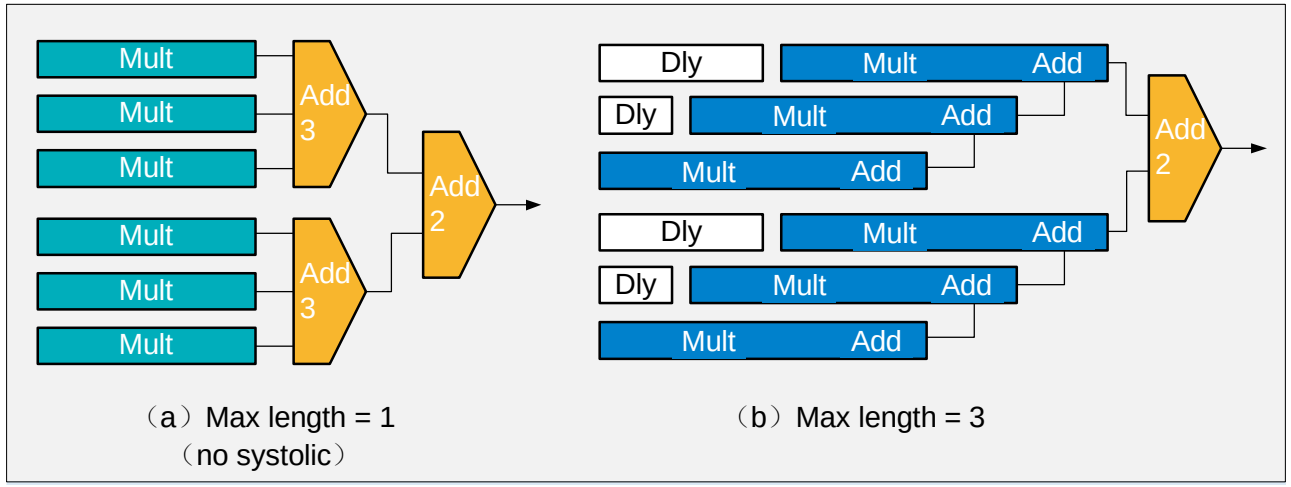
AccEn	LUT	FF	DSP	Resnet50 GOPS
Disable	40722	88778	514	554.114
Enable	37038	82134	642	553.474

[3] UG579 (v1.7) UltraScale Architecture DSP Slice. Xilinx Inc., June 4, 2018.
[4] UG479 (v1.10) 7 Series DSP48E1 Slice. Xilinx Inc., March 27, 2018.

Systolic Adder Tree



Enhanced to (A*2^18+D)*B=AB*2^18+DB on DSP48E2 Slice[5]

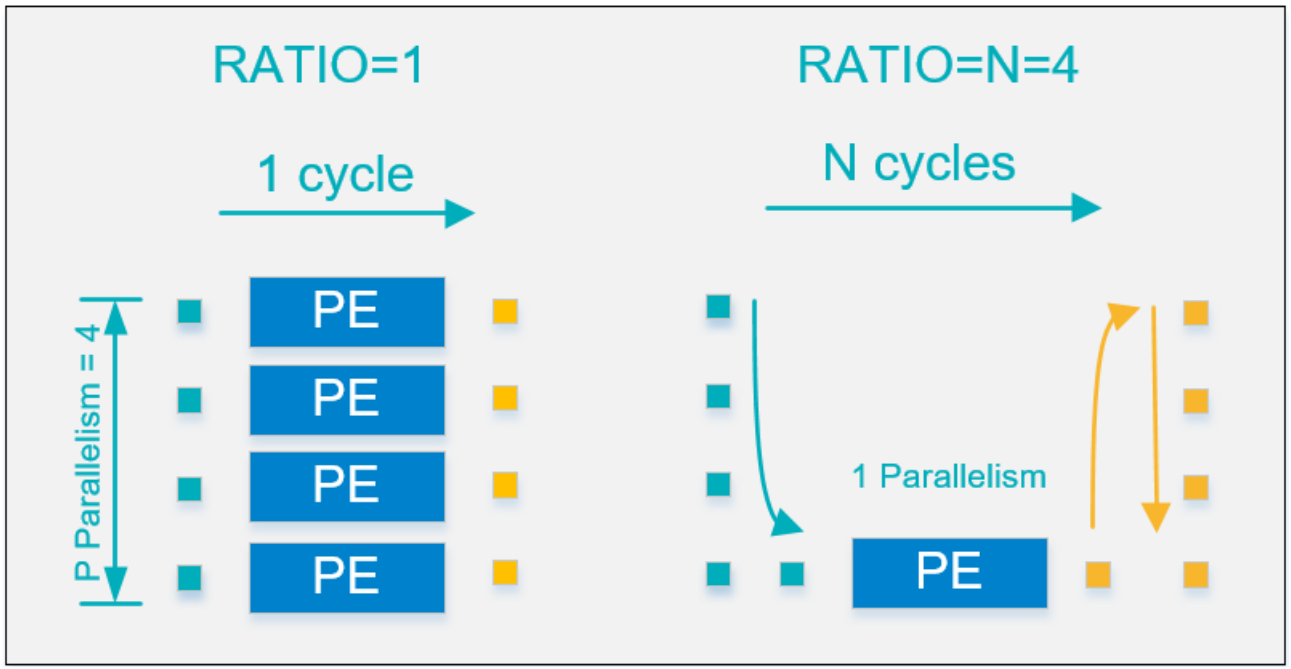


Cascaded Length	LUT	FF	Resnet50 GOPS
1	39349	95381	552.419
4	38172	83285	553.375
6	37038	82134	553.474

Enhanced to Systolic Adder Tree, via cascaded DSP Slice

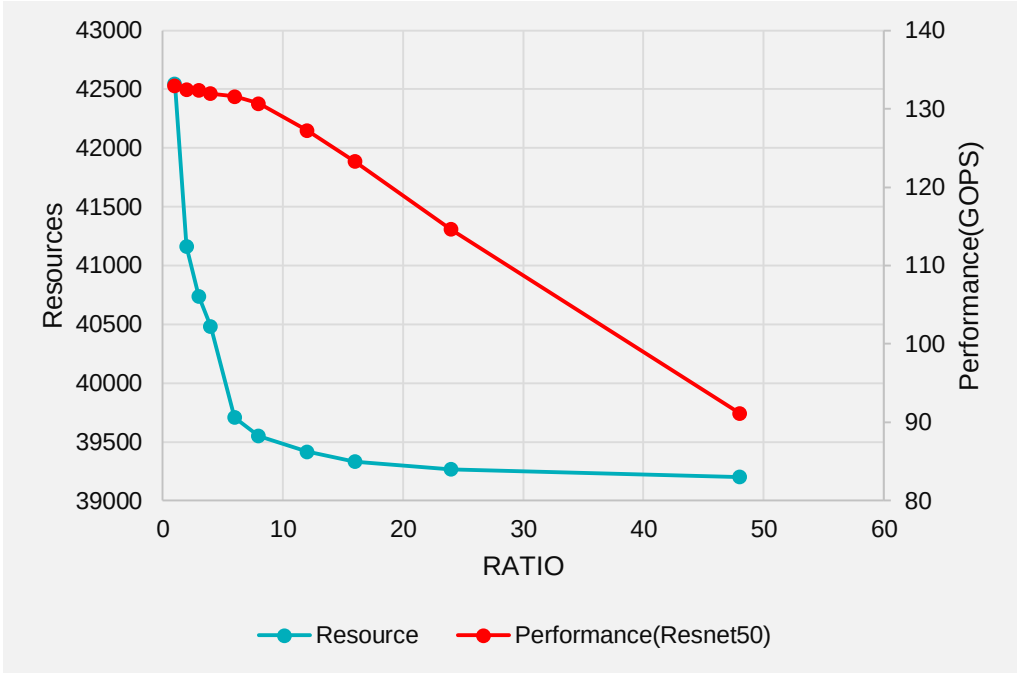
(Max Cascaded Length = 7)

‘SERDES’-like Optimization



Original

Enhanced to SERDES



B1152 resources on zynq7020

- CONV Nonlinear SERDES
  - Nonlinear PE runs at P/N parallelism for N cycles
  - Decrease utilization of Nonlinear function to N/P
  - Ability to compromise between LUT resources and Performance

DPU Peak Perf & Power

	LUT	Flip-Flops	Block RAM	DSP[1)]	DPU Config	Peak[3)] performance	Frequency	Board Power[4)]
7020	53200	106400	4.9Mb	220	1xB1152	230GOPS	200MHz	3.5 / 6.5W
ZU2	47000	94000	5.3Mb	240	1xB1152	576GOPS	500MHz	5.5 / 9.1W
ZU3	71000	141000	7.6Mb	360	1xB2304	852GOPS	370MHz	N/A
ZU5	117000	234000	5.1Mb+18Mb	1248	1xB4096	1.433TOPS	350MHz	N/A
ZU7EV	230000	461000	11Mb+27Mb	1728	2xB4096	2.7TOPS	330MHz	N/A
ZU9	274000	548000	32.1Mb	2520	3xB4096	4.05TOPS	330MHz	17 / 31W

1) One DSP48E is used for two int8 multiplication
2) MACs is constructed by DSP and LUT (if DSP is not enough)
3) Peak performance is calculated by MACs: GOPS = 2*MACs*Frequency
4) The power listed here is average and peak power

DPU Utilization

Do not enable LeakyRelu

Arch	LUTs	Registers	BRAM	DSP
B512	17951	28280	69.5	97
B800	20617	35065	87	141
B1024	22327	39000	101.5	193
B1152	22796	40276	117.5	193
B1600	26270	50005	123	281
B2304	29592	57549	161.5	385
B3136	33266	69110	203.5	505
B4096	37495	84157	249.5	641

Enable LeakyRelu

Arch	LUTs	Registers	BRAM	DSP
B512	18371	28292	69.5	97
B800	21162	35079	87	141
B1024	22759	39012	101.5	193
B1152	23453	40292	117.5	193
B1600	26817	50019	123	281
B2304	30268	57565	161.5	385
B3136	34032	69125	203.5	505
B4096	38392	84173	249.5	641

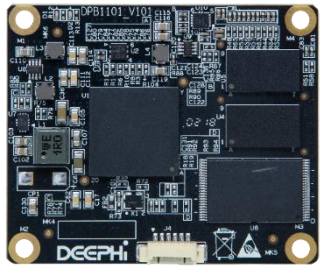
DEEPHi 深鉴科技 Now Part of XILINX

More Mature Products and Collaboration

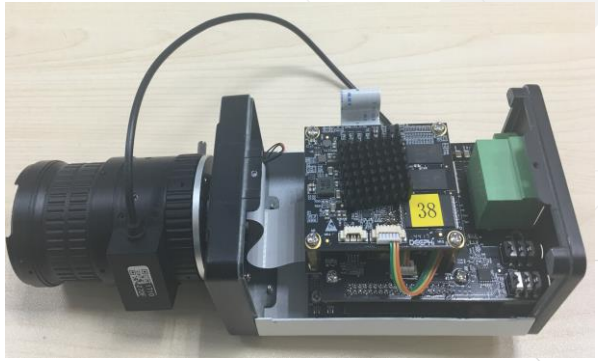
XILINX

Video Surveillance AI Solutions(1)

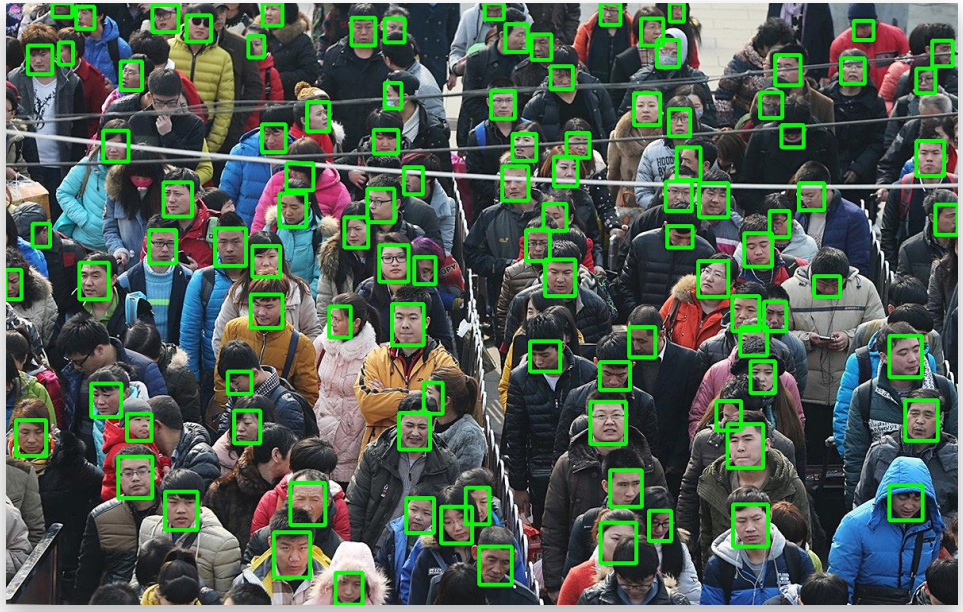
Size: 50*60 mm
DPU : B1152
Peak perf.: 230Gops (200Mhz)
Power: 3.5W (whole board)

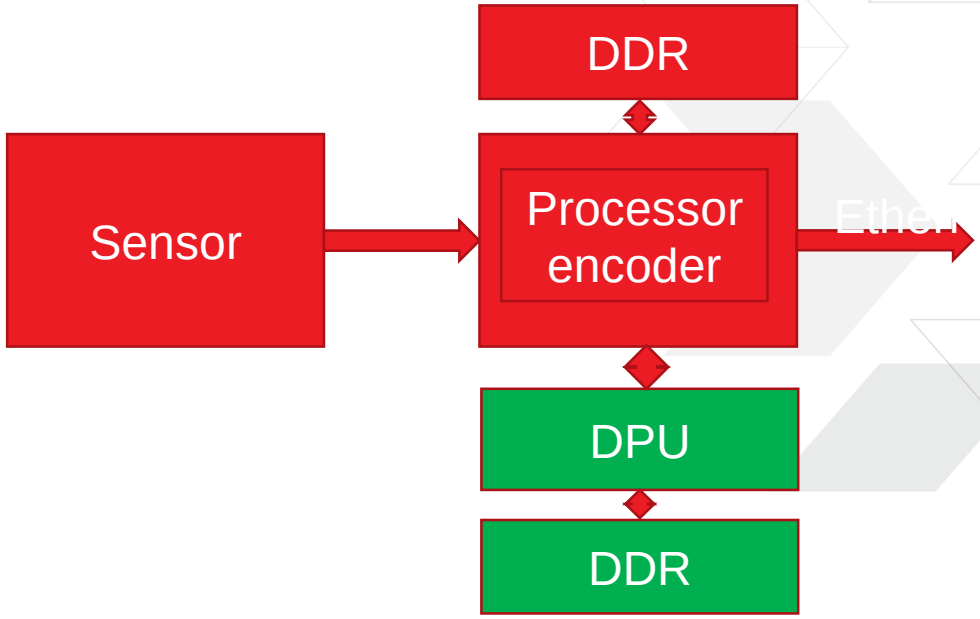


Intelligent IP
Camera Solution



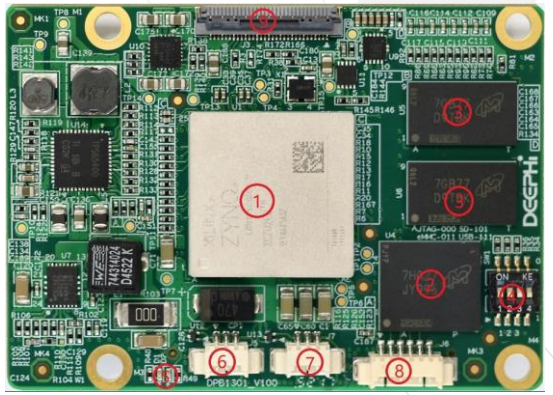
Face recognition camera
with Zynq7020





Video Surveillance ML Solutions(2)

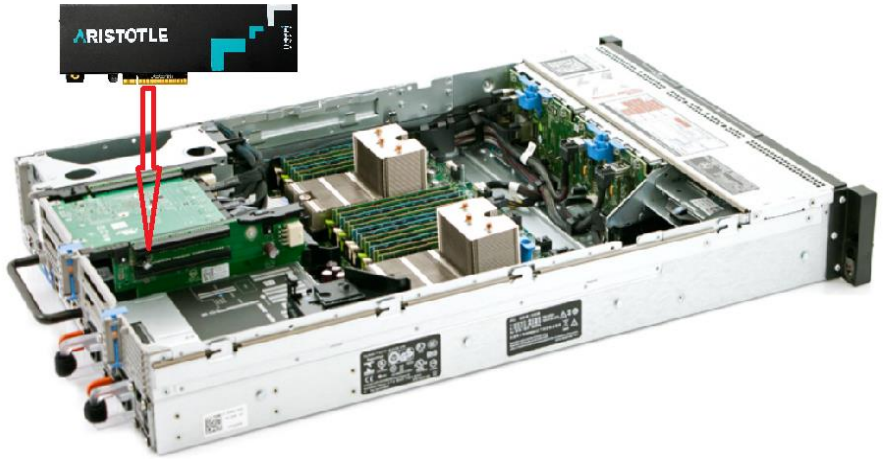
- **Platform**：MPSoC ZU2EG
- **Size:** 50*70 mm
- **DPU**：B1152
- **Peak perf.:** 576GOPS (500MHz)

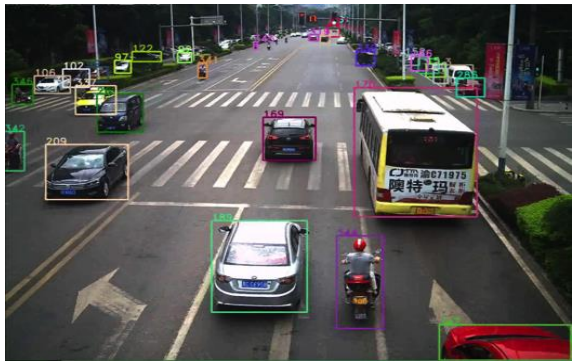


	GOP	200MHz		250MHz		300MHz		350MHz		400MHz	
		FPS	Power	FPS	Power	FPS	Power	FPS	Power	FPS	Power
ResNet-50	7.7	15.3	5.3	18.38	5.55	22.3	5.9	24.87	6.32	27	6.6
ResNet-50[1)]	3.8	23.6	5.29	28.24	5.54	33.5	5.89	37.8	6.2	39.6	6.52
GoogLeNet	3.2	36.5	5.26	45.90	5.54	53.8	5.97	61.7	6.43	68.2	6.74
GoogLeNet[1)]	1.6	62	5.24	77.11	5.59	93	6.03	109.4	6.41	116	6.72
SSD	117	1.63	5.4	2.03	5.71	2.44	6.22	2.835	6.52	3.23	7.05
SSD [1)]	11.6	13.2	5.47	16.35	5.77	19.34	6.18	22.43	6.65	25.3	7.05

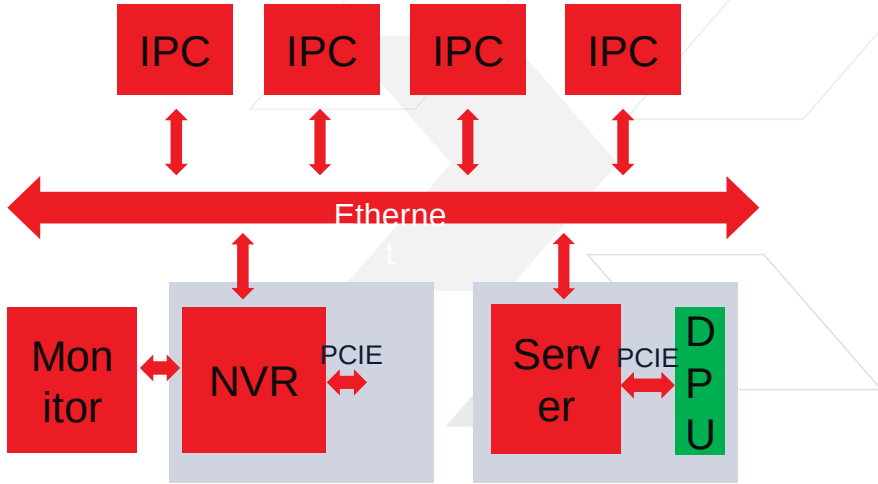
Video Surveillance ML Solutions(3)

- **Platform**： MPSoC ZU9EG
- **Size:** 179*68.9*14.47 mm
- **DPU**： 2 * B4096
- **Peak perf.:** 2.7Tops (330Mhz)
- **Power**: 21W (whole board)





	FPS	Avg Power
JPEG analysis **For face detection**	126	19.7W
9 channels Video **For face detection and recognition**	267	18.7W
12 channels Video **For video analysis**	105	21.5W
Idle status	N/A	14.3W



*Input : 1080P Jpeg or H264 video

XILINX

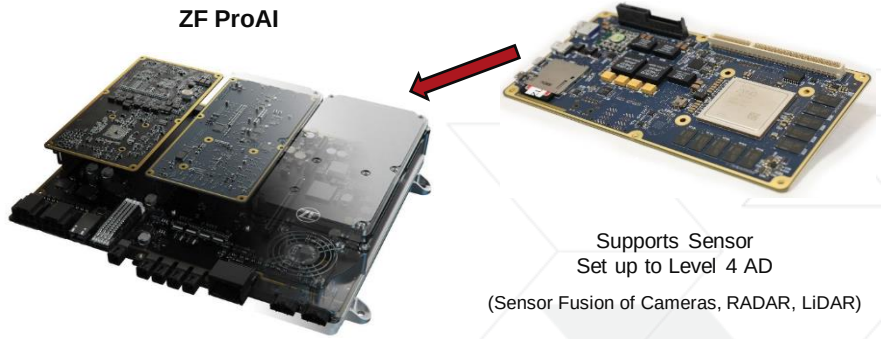
Xilinx AI in Automotive Applications

Daimler Selects Xilinx for AI-based Auto Applications

"Xilinx is providing technology that will enable us to deliver **very low latency and power-efficient solutions for vehicles that must operate in thermally constrained environments**. We have been very impressed by Xilinx's heritage and selected the company as a trusted partner for our future products."

Georges Massing, Director, Daimler AG

Xilinx powers ZF's artificial intelligence (AI)-based automotive control unit



ZF ProAI is its modular hardware concept and open software architecture. This approach is unique compared to other systems on the market, which use a fixed combination of hardware and software architecture (which can limit functionality and add more cost)

DEEPHi 深鉴科技 Now Part of XILINX

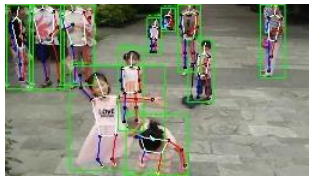
DNNDK Becomes Xilinx AI SDK

XILINX

Platform Stack for Edge AI

Models


Face detection


Pose estimation


Video analytics


Lane detection


Object detection


Segmentation

Framework

Caffe

Darknet

TensorFlow

Tools & IP

DEEPHi DNNDK

Compression
- Pruning
- Quantization

Compilation
- Compiler
- Assembler

Runtime
- Core API
- Loader
- Driver
- Tracer

DEEPHi DPU

VIVADO

SDSoC Environment

AI HW Platforms


Z7020 Board


Z7020 SOM


ZU2 SOM


ZU2/3 Card

ZU9 Card


ZCU102

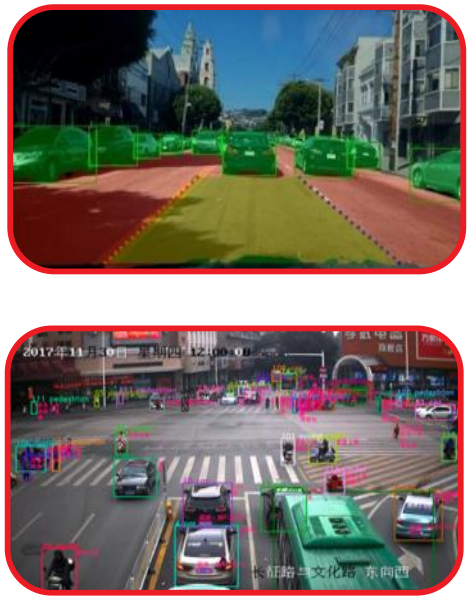

ZCU104


Ultra96

XILINX

Xilinx AI SDK

- **Xilinx AI SDK for different applications**
  - High-level API based libraries across different vision tasks: classification, detection, segmentation and etc.
  - Reference applications to help fast prototyping
  - Very few codes to construct AI application



Very few codes to run YOLOv3 model on video

```
#include <xilinx/yolov3/yolov3.hpp>
#include <iostream>
#include <memory>
#include <glog/logging.h>
#include <opencv2/core.hpp>
#include <opencv2/highgui.hpp>
#include <opencv2/imgproc.hpp>
#include <xilinx/demo/demo.hpp>
#include "./process_result.hpp"
using namespace std;
int main(int argc, char *argv[]) {
  return xilinx::demo::main_for_video_demo(
      argc, argv, [] {
    return xilinx::yolov3::YOLOv3::create(
        xilinx::yolov3::VOC_416x416);
  },
  process_result);
}
```

Xilinx AI SDK v1.0 Features

27 models included
Classification, detection and segmentation
5 TensorFlow models

Performance optimization
Post-processing code optimized by 10X at most

Boards supporting
ZCU102, ZCU104 and Ultra96

Easy demo construction
Simple demo for every libraries
Supporting image and video input

Many demos based on AI SDK

- **8 channel vehicle and pedestrian demo**
- **4 channel mixed demo**
  - Segmentation and pose detection by video input
  - Face and pedestrian detection by camera input
- **4 channel segmentation + road-line detection demo**
- **…**

DEEPHI 深鉴科技 Now Part of XILINX

Adaptable.
Intelligent.

XILINX