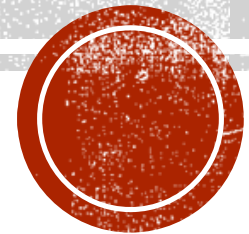


# HW AI: THE FIRST LOOK

Presenter: Cuong Pham (cuongpt12)

Nov-12-2019

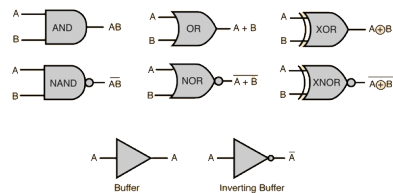


# CONTENTS

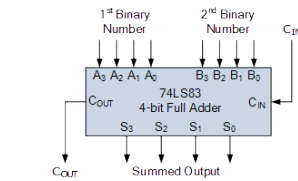
1. Overview
2. Complexity of CNN
3. Boards survey
4. Conclusion

# OVERVIEW

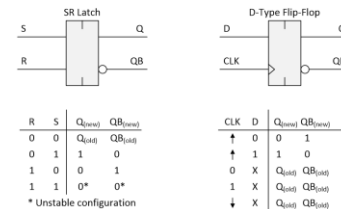
- ASIC stands for Application Specific Integrated Circuit.
- It's built from logic gate (physical layer is transistor).



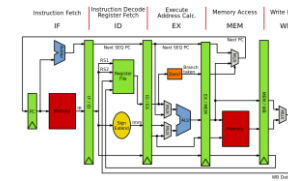
Basic gates



Logic function



Storage element



processor

Specific datapath



Memories



CPU

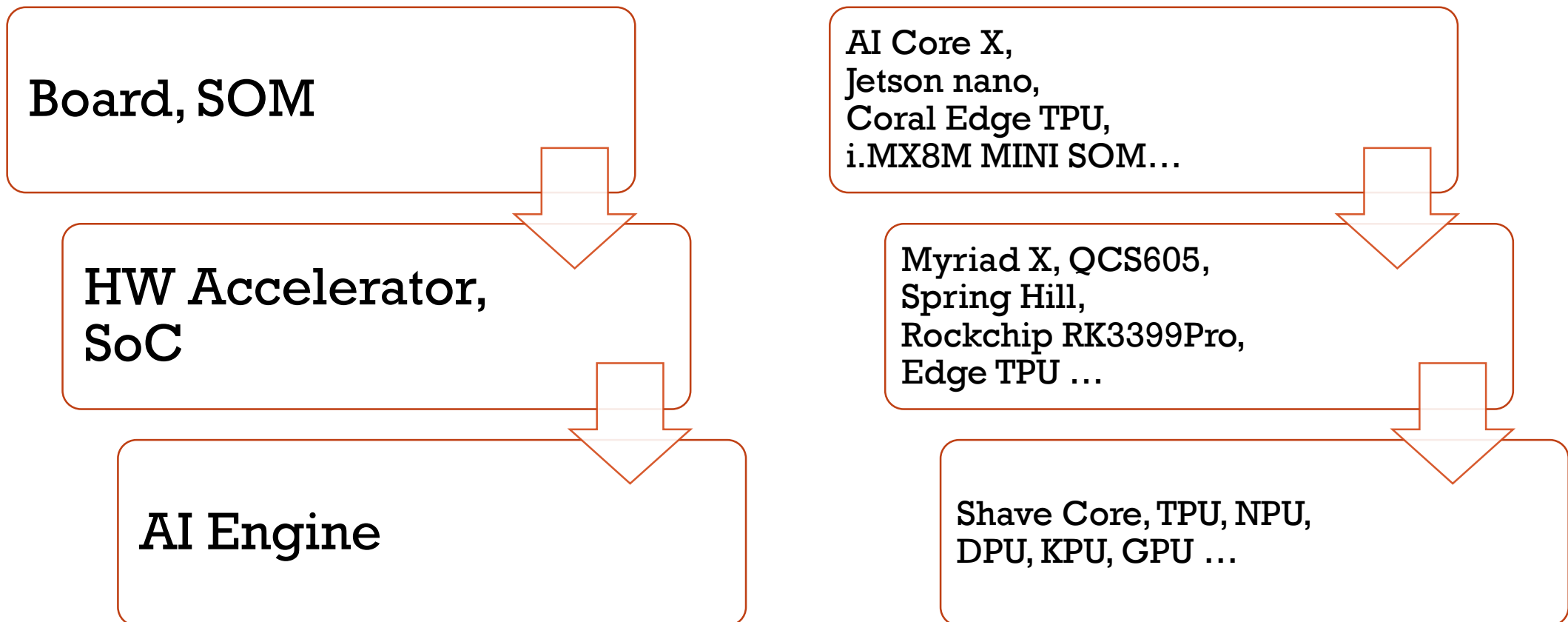


GPU



FPGA

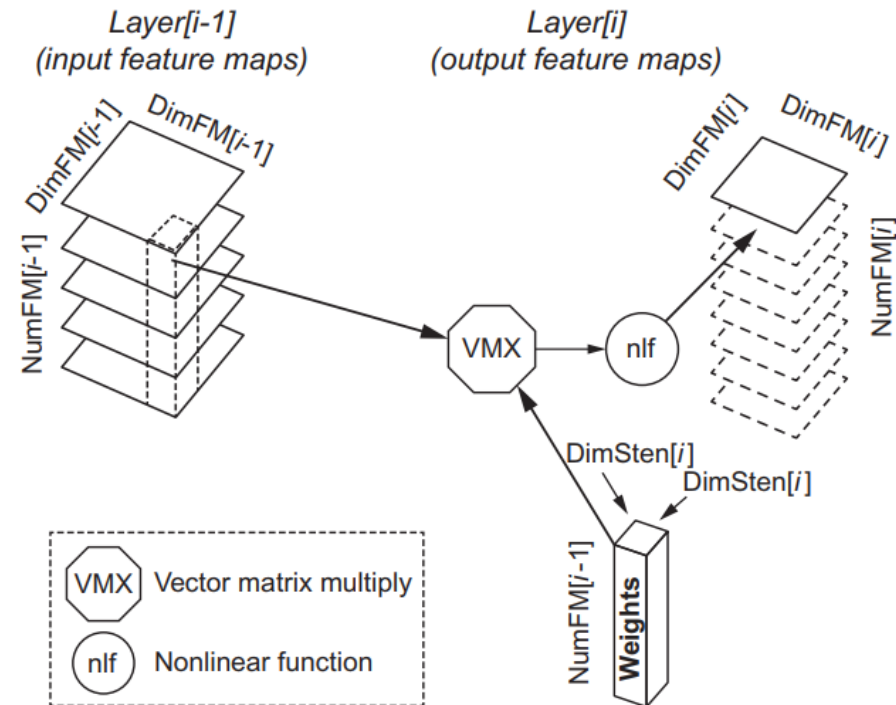
# OVERVIEW



# COMPLEXITY OF CNN

- Number of weights per output Feature Map:  $\text{NumFM}[i-1] \times \text{DimSten}[i]^2$
- Total number of weights per layer:  $\text{NumFM}[i] \times \text{Number of weights per output Feature Map}$
- Number of operations per output Feature Map:  $2 \times \text{DimFM}[i]^2 \times \text{Number of weights per output Feature Map}$
- Total number of operations per layer:  $\text{NumFM}[i] \times \text{Number of operations per output Feature Map} = 2 \times \text{DimFM}[i]^2 \times \text{NumFM}[i] \times \text{Number of weights per output Feature Map} = 2 \times \text{DimFM}[i]^2 \times \text{Total number of weights per layer}$
- Operations/Weight:  $2 \times \text{DimFM}[i]^2$

NOTE: 1 MAC (multiply-accumulate) = 2 Operations



# COMPLEXITY OF CNN (YOLOV3)

■ Layer 0:

$$2 * (416 * 416) * [32 * (3 * 3 * 3)]$$

■ Layer 1:

$$2 * (208 * 208) * [64 * (3 * 3 * 33)]$$

■ Layer 105:

$$2 * (52 * 52) * [255 * (1 * 1 * 256)]$$

```
layer   filters  size      input                                output                                BFLOPs
0 conv   32      3 x 3      416 x 416 x 3 -> 416 x 416 x 32 0.299
1 conv   64      3 x 3 / 2  416 x 416 x 32 -> 208 x 208 x 64 1.595
-----
105 conv 255      1 x 1 / 1   52 x 52 x 256 -> 52 x 52 x 255 0.353
106 detection

truth_thresh: Using default '1.000000'
Loading weights from yolov3.weights...Done!
data/dog.jpg: Predicted in 0.029329 seconds.
dog: 99%
truck: 93%
bicycle: 99%
```

# BOARDS SURVEY



Jetson Nano



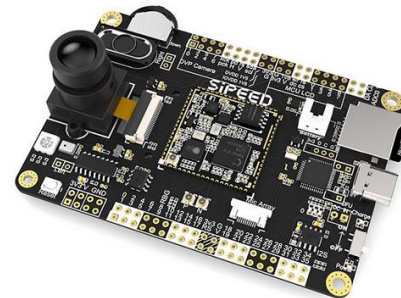
i.MX8M MINI SOM



AI Core X



Coral Edge TPU



Sipeed MAIX Go



ZCU104

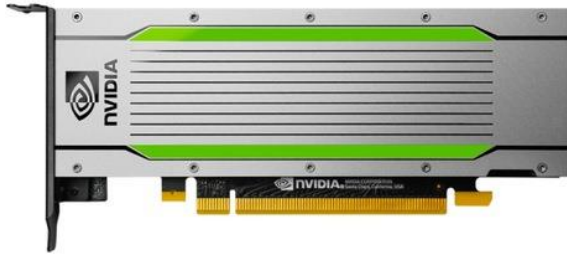


# BOARDS SURVEY

No.	Board	Vendor	feature		frameworks	Prices (eval..)	Power (W)	Power efficiently (GOP/W)
			Chipset	OPS				
1	Jetson Nano	NVIDIA	Quad-core ARM Cortex-A57 128-core NVIDIA Maxwell™ GPU	472 GFLOPS	TensorFlow, PyTorch, Caffe and MXNet	99\$	5	94.4
2	Coral Edge TPU	Google	NXP i.MX 8M SoC: Quad-core ARM Cortex-A53 + GC7000 GPU Edge TPU coprocessor	4 TOPS (int8)	TensorFlow Lite	-	2	2000
3	i.MX8M MINI SOM	Gyr Falcon	Arm Cortex A53 single/dual/quad core 1.8Ghz Lightspeeus 2803S	16.8 TOPS (int8)	Caffe, TensorFlow	132\$	-	-
4	AI Core X	AAEON (ASUS)	Intel® Movidius™ Myriad™ X VPU 2485: 2 LEON4 cores 16 SHAVE core	1 TOPS (int8)	caffe, tensorFlow	79\$	2.5	400
5	ZCU104	Xilinx	Zynq UltraScale+ (XCZU7EV-2FFVC1156)	2.4 TOPS (MAC int8)	TensorFlow, Caffe, PyTorch	895\$	—	—
6	Sipeed MAIX	Seed Studio (China)	Kendryte K210: Dua-core RV64GC, 8MB SRAM @ 400MHz, boot 800MHz KPU (Neural network processor) 64 core, APU (audio processor)	250 GOPS Max 500 GOPS (int8)	????	5\$ (chip)	0.3	833.3



# BOARDS SURVEY (NVIDIA)



NVIDIA T4



NVIDIA TESLA V100



NVIDIA QUADRO P4000



GEFORCE RTX 2080 Ti

# BOARDS SURVEY (NVIDIA)

	NVIDIA T4	NVIDIA TESLA V100	NVIDIA QUADRO P4000	GEFORCE RTX 2080 Ti
NVIDIA CUDA Cores	2560	5120	1792	4352
NVIDIA Tensor Cores	320	640	-	-
Double-Precision Performance	-	7 TFLOPS		
Single-Precision Performance	8.1 TFLOPS	14 TFLOPS	3 TFLOPS (estimated)	7.1 TFLOPS (estimated)
Tensor Performance	65 TFLOPS (FP16) 130 TOPS (int8)	112 TFLOPS (FP16)	-	-
GPU Memory	16 GB GDDR6	32/16 GB HBM2	8 GB GDDR5	11 GB GDDR6
Memory Bandwidth	300 GB/s	900 GB/s	243 GB/s	616 GB/s
Interconnect Bandwidth	32 GB/s	32 GB/s	32 GB/s	
System Interface	x16 PCIe Gen3	PCIe Gen3	x16 PCIe Gen3	
Max Power Consumption	70 W	250 W	105 W	260 W
Compute APIs	CUDA, NVIDIA TensorRT, ONNX	CUDA, DirectCompute, OpenCL, OpenACC	CUDA, DirectCompute, OpenCL	

# BOARDS SURVEY (PERFORMANCE)

id	Neural Network	input size	MAC (GOPs)	Performance l (fps)	Percentage l	performance m (fps)	Percentage m	estimate peak MAC (GOPs)
<b>ZCU104</b>								
1	ResNet-18	224x224	3.65	206.7	31.44%	<b>428.6</b>	65.18%	2400
2	ResNet-50	224x224	7.7	82.5	26.47%	<b>151.8</b>	48.70%	2400
3	YOLOV3_ADAS	512x256	5.5	95.2	21.82%	<b>228.4</b>	52.34%	2400
4	YOLOV3_ADAS	512x288	53.7	14.3	32.00%	<b>31.7</b>	70.93%	2400
5	YOLOV3_VOC	416x416	65.4	15.1	41.15%	<b>33</b>	89.93%	2400
6	YOLOV3_VOC_TF	416x416	65.6	15	41.00%	<b>32.8</b>	89.65%	2400
7	YOLOV2_BASELINE	448x448	34	26.6	37.68%	<b>63.8</b>	90.38%	2400
8	Inception-v2	224x224	4	158	26.33%	<b>310.2</b>	51.70%	2400
9	Inception V4	299x299	24.5	30.9	31.54%	<b>64.6</b>	65.95%	2400
10	SSD Mobilenet-V2	480x360	6.6	26.5	7.29%	<b>116.4</b>	32.01%	2400
11	ResNet-50	224x224	7.7	82.5	26.47%	<b>151.8</b>	48.70%	2400
<b>jetson nano</b>								
18	Tiny YOLOV3	416x416	2.78			<b>25</b>	29.45%	236
19	Inception V4	299x299				<b>11</b>		236
20	SSD Mobilenet-V2	480x360				<b>39</b>		236
21	ResNet-50	224x224				<b>36</b>		236
22	VGG-19	224x224				<b>10</b>		236
<b>coral edge TPU (destop host: 64-bit Intel(R) Xeon(R) Gold 6154 CPU @ 3.00GHz)</b>								
28	Inception V4	299x299				<b>11.7</b>		2000
29	Inception v1	224x224				<b>294.1</b>		2000
30	VGG-19	224x224				<b>3.2</b>		2000

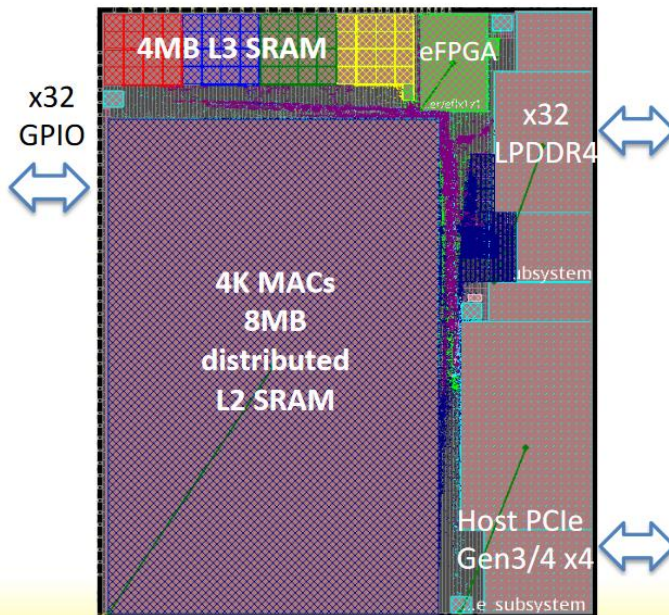
# CHIPS SURVEY

id	company	Machine learning processors	type
1	Intel	NNP	both
2	Intel	Myriad	inference
3	Intel	EyeQ	inference
4	Intel	GNA	inference
5	Intel	Spring Hill	inference
6	ARM	ML Processor	inference
7	Google	TPU	both
8	Apple	Neural Engine	
9	Nvidia	NVDLA	inference
10	Nvidia	Xavier	inference
11	Samsung	Neural Processing Unit	
12	Tesla	FSD Chip	
13	Alibaba	Ali-Npu	
14	Amazon	AWS Inferentia	
15	Huawei	Ascend	

id	company	Machine learning processors	type
16	Baidu	Kunlun	both
17	Bitmain	Sophon	both
18	Cambricon	MLU	
19	Flex Logix	InferX	inference
20	Nepes	NM500	
21	GreenWaves	GAP8	
22	Gyr Falcon Technology	Lightspeed	inference
23	Graphcore	IPU	
24	Hailo	Hailo-8	
25	Kendryte	K210	inference
26	Mythic	Template:mythic	
27	NationalChip	Neural Processing Unit	
28	Synaptics	SyNAP	
29	Wave Computing	DPU	

# FLEX-LOGIX

## | InferX X1 – Running Models in Emulation



- Q4 tape-out
- <50mm<sup>2</sup> TSMC 16FFC
- 1.067GHz Operation
- 4K MACs @ INT8x8/16x8
  - 2K MACs @ INT16x16/BF16
  - Winograd acceleration for INT8
- 8MB L2 SRAM + 4MB L3 SRAM
- x32 LPDDR4 (16GB/s peak BW)
- Partners: TSMC, GUC, Synopsys, Arteris, Analog Bits, Cadence, Mentor
- Available as Chip & PCIe Board in Q1

# FLEX-LOGIX

Model	Image Size	Frames/Second	DRAM Bandwidth	MAC Utilization
YOLO v3	2048x1024	12.3	4.6 GB/sec	68%
YOLO v3	608x608	66	6.1 GB/sec	65%
YOLO v2	2048x1024	36	7.3 GB/sec	66%
ResNet50 v1	2048x1204	20.8	6.4 GB/sec	60%
ResNet50 v1	224x224	293	11.1 GB/sec	20%
ResNet152 v2	299x299	139	10.2 GB/sec	50%
ResNet101 v2	299x299	181	10.7 GB/sec	43%
MobileNet v1	224x224	1188	9.7 GB/sec	17%

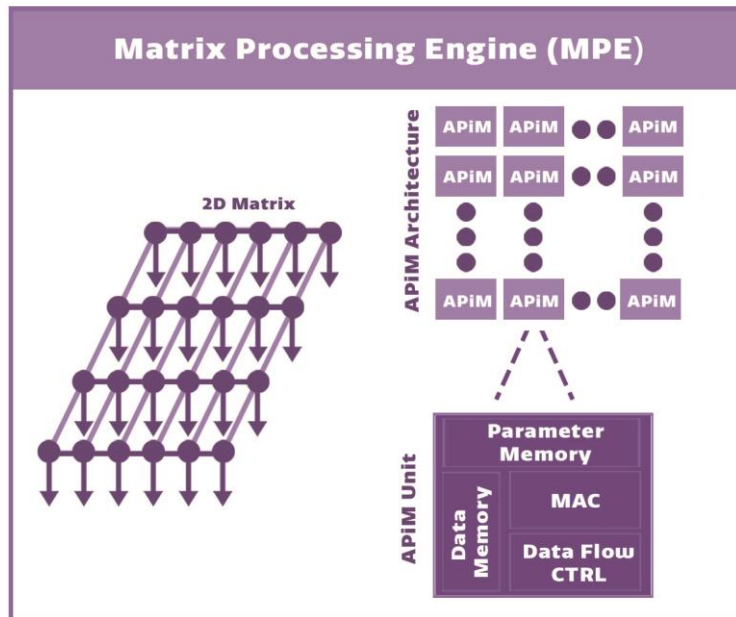
*All benchmarks are  
Single X1,  
batch size = 1,  
multi-layer,  
INT8, no pruning.*

September 2019

Flip Chip BGA	21 x 21 mm
TSMC 16FFC	~50 mm <sup>2</sup>
TOPS (MACs x 1.067GHz x 2)	8.5
On-chip Compute & Memory	4K MACs, 8MB SRAM
x32 LPDDR4 DRAM	1
Interfaces	x4 PCIe Gen3/4 & x32 GPIO
Typical power @ TT 0.8V 85°C	~2.5 W (ResNet-50)
PCIe Card TDP (Worst-case)	~10 W (YOLOv3)



# GYRFALCON



PLAI PLUG 2801



PLAI PLUG 2803



PLAI PLUG 5801



i.MX8M MINI SOM



# GYRFALCON

	Lightspeeur 5801	Lightspeeur 2801	Lightspeeur 2803
<b>computational power</b>	2.8 TOPs	5.6 TOPs	16.8 TOPs
<b>power</b>	224 mW -> (12.6 TOPs/W)	600 mW -> (9.3 TOPs/W)	700 mW -> (24 TOPs/W)
<b>freq</b>	200 MHz	100 MHz	250 MHz
<b>size</b>	6x6	7x7	9x9
<b>Applications</b>	Mobile Edge Computing	Image & Video Classification	Image & Video Classification
	Vision Systems	Object & Facial Recognition	Object & Facial Recognition
	Smart Toys/Robotics/Home	Voice Authentication	Voice Authentication
	Augmented Reality/Virtual Reality	OCR Optical Character Recognition	OCR Optical Character Recognition
	Face Detection/Recognition	Natural Language Processing	Natural Language Processing
	Speech/Voice Recognition		
	Natural Language Processing		
	Deep Learning Enabled Devices		
<b>use case</b>	Edge Inference Sytems		
	Smart home	Smart home	Smart home
	Manufacturing	Manufacturing	Manufacturing
	Security/Surveillance	Security/Surveillance	Security/Surveillance
	Medical	Medical	Medical
	Retail	Retail	Retail
	mobile, embedded, and IoT	AI devices for edge, desktop and data center deployments	advanced edge, desktop and data center deployments
<b>supported Frameworks</b>	Caffe, TensorFlow	Caffe, PyTorch, TensorFlow	Caffe, TensorFlow
<b>price (PLAI)</b>	39.99\$	49.99\$	69.99\$
<b>SOC</b>		rockchip RK3399Pro	

# HAIO

# CONCLUSION

- Edge tpu offer huge computational power, but we only use small piece of it.
- Myriad chip in board AI core X is the best choice with 1 TOPS, low power, low cost and good platform.
- ZCU104 is the best FPGA for AI at the edge with 2.4 MAC TOPS (~4.8 TOPS) at the cost 895\$
- Gyrfalcon offer good chip for AI acceleration.