



# DEEPX's New M1 NPU Delivers Flexibility, Accuracy, Efficiency and Performance

Jay Kim

Executive Vice President

DEEPX

**DEEPX**  
FOR AI EVERYWHERE

# Disruptive Innovation "IT'S REAL"

## NVIDIA Model: V100 16GB



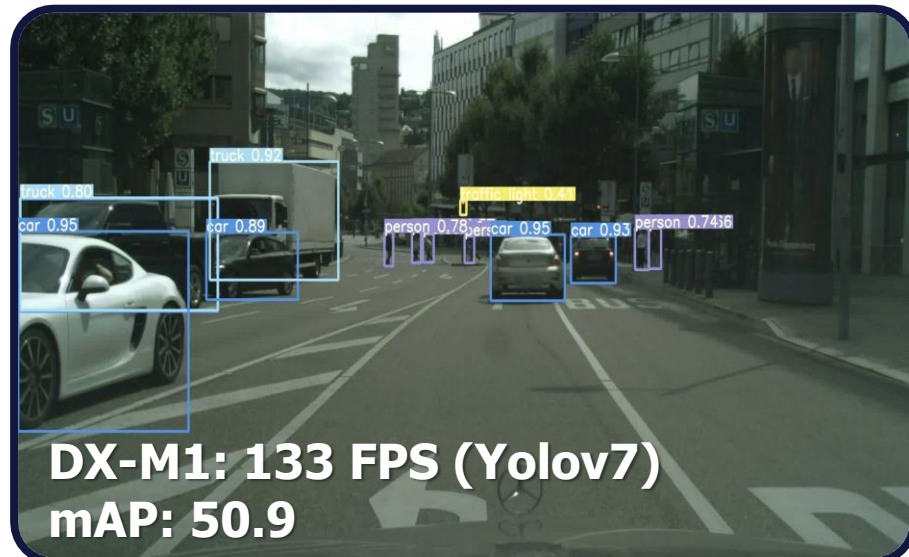
- Price: Approx. \$14,000
- Power Consumption: 300 W



## DEEPX's Flagship Model: DX-M1



- Price: Approx. \$69.99
- Power Consumption: 3~5 W



# DEEPX's Key Differentiators



## World Leading SOTA DNN Algorithms

### + Transformer Model (ViT etc.)

- ✓ densnet
- ✓ googlenet
- ✓ mnasnet
- ✓ mobileNet
- ✓ ResNet
- ✓ SSD
- ✓ Yolov3, v4, v5, v7
- ✓ EfficientNet/Det
- ✓ BiSeNet
- ✓ ShelfNet
- ✓ PIDNet
- ✓ SFA3D

+ Other AI models  
(Model Zoo: > 170 models)

## The World's First AI Accuracy Technology (mAP)

	Model	FP32 NVIDIA	INT8 Company A	INT8 DEEPX
OD*	MobileNet SSD	23	22.2	22.6
	Yolov4	49.6	41.55	49.3
	Yolov5m	44.1	39.12	43.7
	YoloXs	40.3	37.47	<b>41.1</b>
	Yolo7m	51.0	N/A	50.9
IC*	MobileNetv1	71.48	70.13	<b>72.42</b>
	ResNet50	75.94	74.69	<b>75.95</b>
	EfficientNet-B0	77.52	76.96	<b>77.62</b>
Seg*	BiSeNet	75.19	N/A	<b>75.97</b>
	PIDNet	78.76	N/A	<b>78.79</b>
	DeepLabv3+	72.07	N/A	<b>72.37</b>

## The World's best Power/Performance Efficiency

Company	TOPS/W Resnet-50	FPS/TOPS Resnet-50
DEEPX	<b>&gt; 10</b>	<b>60</b>
Company A	8.6	47
Company B	8.8	25
Company C	4.47	26
Company D	4.0	25
NVIDIA	1.8	17
Company E	0.7	29
Company F	5.0	Unknown

\* OD | Object Detection \* IC | Image Classification \* Seg | Segmentation

# DEEPX's Key Differentiators – 1. Flexibility



## The World Leading SOTA DNN Algorithms

### + Transformer Model (ViT etc.)

- ✓ densnet
- ✓ googlenet
- ✓ mnasnet
- ✓ mobileNet
- ✓ ResNet
- ✓ SSD
- ✓ Yolov3, v4, v5, v7
- ✓ EfficientNet/Det
- ✓ BiSeNet
- ✓ ShelfNet
- ✓ PIDNet
- ✓ SFA3D

+ Other AI models  
(Model Zoo: > 170 models)

- ✓ Latest: Latest DNN algorithms (SOTA)
    - EfficientNet , YoloX, YoloV7, PIDNet ...
  - ✓ Wide Range: Various kinds of algorithms
    - Classification, object detection, segmentation, pose estimation, anomaly detection ...
  - ✓ Complex Algorithms: Normally too complex for edge NPUs
    - DeepX NPU can run PIDNet, SFA3D (Sensor Fusion)
- **Customers can use our NPU longer and wider.**



# Evidence #1: SOTA Support (Object Detection)



# DEEPX's Key Differentiators – 2. Accuracy



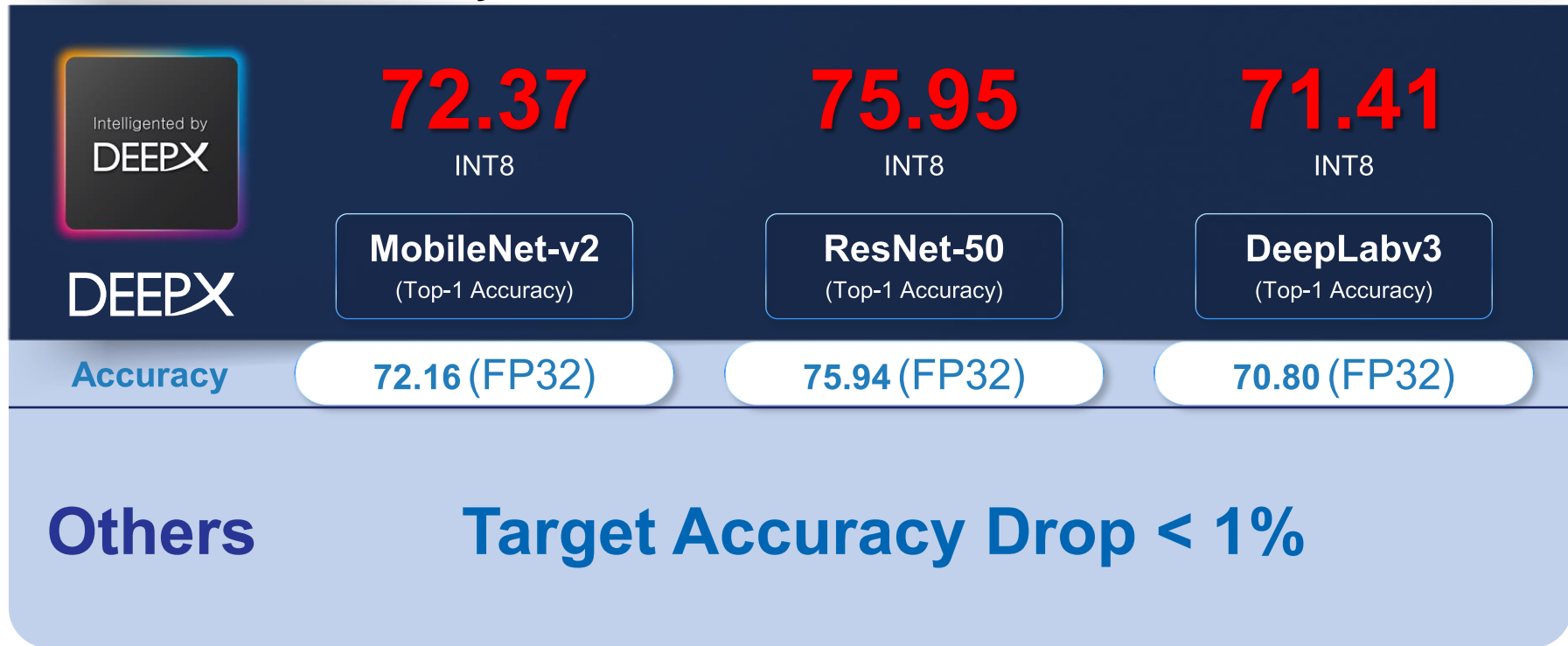
## The World's First AI Accuracy Technology

	Model	FP32 NVIDIA	INT8 Company A	INT8 DEEPX
OD*	MobileNet SSD	23	22.2	22.6
	Yolov4	49.6	41.55	49.3
	Yolov5m	44.1	39.12	43.7
	YoloXs	40.3	37.47	<b>41.1</b>
	Yolo7m	51.0	N/A	50.9
IC*	MobileNetv1	71.48	70.13	<b>72.42</b>
	ResNet50	75.94	74.69	<b>75.95</b>
	EfficientNet-B0	77.52	76.96	<b>77.62</b>
	BiseNet	75.19	N/A	<b>75.97</b>
Seg*	PIDNet	78.76	N/A	<b>78.79</b>
	DeepLabv3+	72.07	N/A	<b>72.37</b>

- ✓ Use int 8-bit instead of 32-bit floating point operations
    - Expect accuracy drop comparing to GPU
  - ✓ Almost the same accuracy as a GPU can get
    - Normally same or less than 1% accuracy drop expected
  - ✓ Wow factor: Better accuracy than GPU in some cases
    - YoloXs, MobileNetv1, ResNet50, EfficientNet-B0 and etc
- **Customers can get the most accurate result with our NPU.**

# Evidence #2: AI Accuracy Comparison

DEEPX accuracy is better than FP32 & other latest NPUs



# DEEPX's Key Differentiators – 3. Efficiency



## The World's best Power/Performance Efficiency

Company	TOPS/W Resnet-50	FPS/TOPS Resnet-50	FPS/W Resnet-50
DEEPX	> 10	60	> 600
Company A	8.6	47	404.2
Company B	8.8	25	220
Company C	4.47	26	116.22
Company D	4.0	25	100
 NVIDIA	1.8	17	30.6
Company E	0.7	29	20.3
Company F	5.0	Unknown	Unknown

- ✓ Popular power efficiency factor (TOPS/W) with Resnet50
    - World top power efficiency: over 10 TOPS/W
  - ✓ Prefer other efficiency factor (FPS/TOPS)
    - Actual result with 1 TOPS
  - ✓ Finally pursue the effective power efficiency such as FPS/W
    - Actual result with 1 Watt
- **Customers can get the max performance with the lowest power.**



# Extreme Case: Ultra Low Power NPU

## CMOS IMAGE SENSOR

### Intelligent CMOS Image Sensor

- Image Enhancement NPU (PoC)
- Face Recognition Function NPU (PoC)
  - ✓ **1 mm X 1 mm @40 nm**
  - ✓ **Lower than 10 mW & 10 fps**
  - ✓ Face recognition accuracy



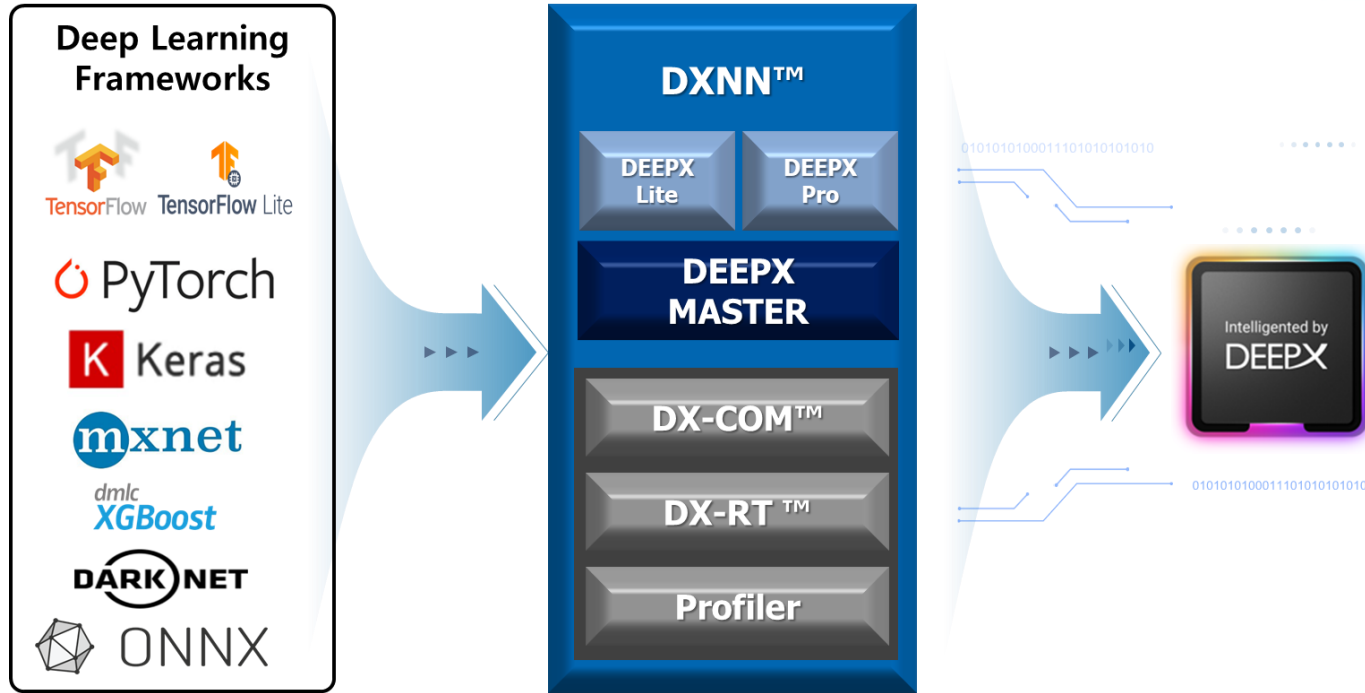
Image Enhancement



Face ID






# Important Factor for Commercial Success: DXNN™ – DEEPX NPU SDK



❖ **Beta version of DXNN has been released to customers.**



# Customer Success Story

# Success Story: Actual Throughput (1<sup>st</sup> Wow)

	Jetson AGX Xavier Dev-Kit	DEEPX NPU IP (FPGA)	DX-M1
Type			
	MSRP: \$699.00		MSRP: \$70.00
AI Performance	32 TOPS <b>24 FPS</b> Customer Algorithm	1 TOPS <b>30 FPS</b> Customer Algorithm	23 TOPS <b>Est. 240 FPS</b> Customer Algorithm

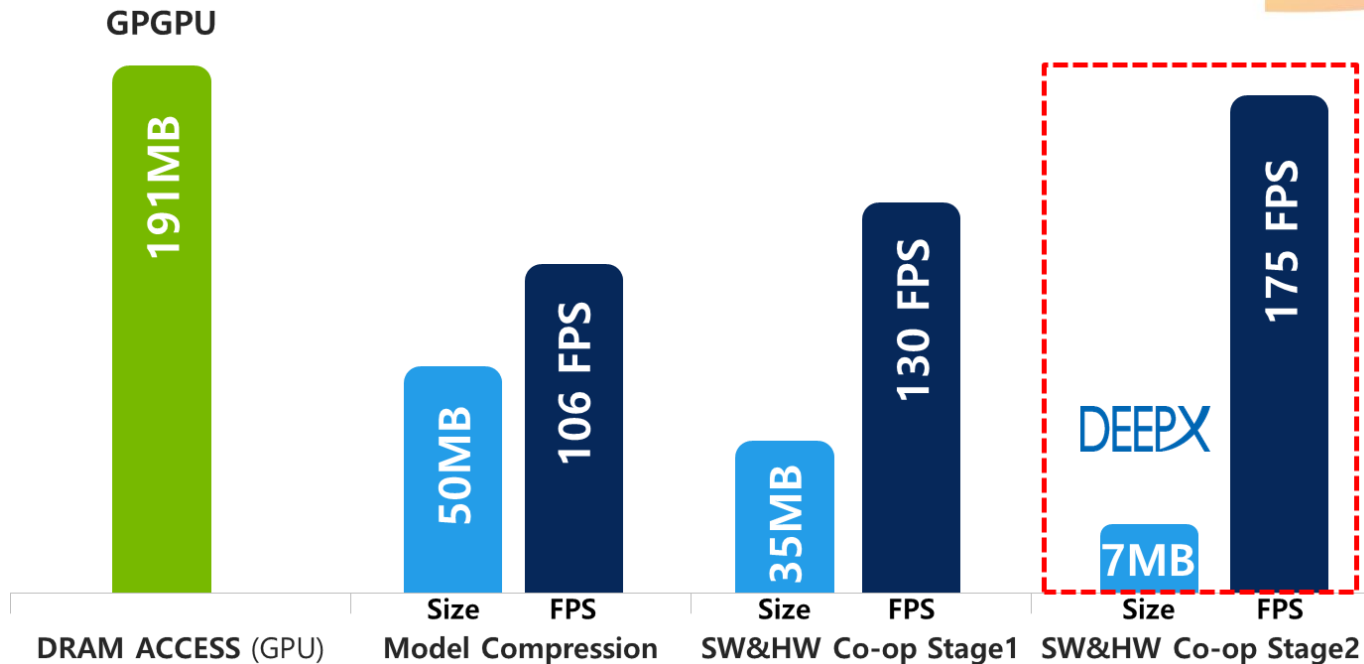
Performance **X 10**  COST **X 10**  = Efficiency **X 100** 

# Success Story: Accuracy (2<sup>nd</sup> Wow)

Model / Data	NVIDIA GPU (FP32)	DEEPX NPU (INT8)	Delta
 <b>HYUNDAI</b> (Robot)	41.8%	41.9%	<u>0.1%</u> ↑
 <b>eyenix</b> the great eyes (CCTV)	87.6%	89%	<u>1.4%</u> ↑

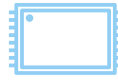
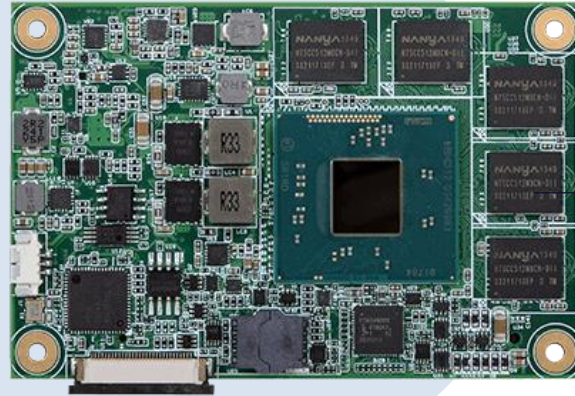


# Success Story: SH/HW Co-optimization (3<sup>rd</sup> Wow)

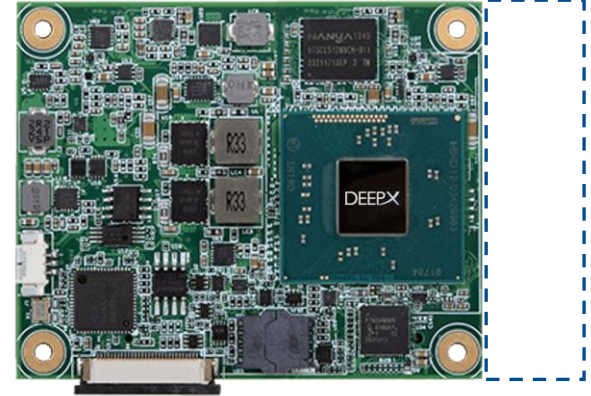


With DEEPX's Optimization, we can maximize AI performance and minimize DRAM Access

# Success Story: BOM Cost Reduction (4<sup>th</sup> Wow)



DEEPX DRAM Access  
Optimization



With DEEPX's Optimization, we can maximize AI performance and minimize DRAM Access

**I** Small DRAM footprint ► **Performance and energy efficiency**

**II** Less DRAM chipset required

**III** Small form factor

**IV** Cost saving & space saving

## DEEPX NPU can:

1. Run more SOTA AI models.
2. Get the most accurate results.
3. Achieve the best power efficiency.
4. Provide a cost efficient solution (including BOM).

Please visit our demo booth and check!

**Thank you !!**

1. Demo Booth: #103
2. DEEPX Homepage  
<https://www.deepx.ai>
3. LinkedIn & Youtube



## 2023 Embedded Vision Summit

Additional Talks from DEEPX:

1. "Toward the Era of AI Everywhere"  
(Lokwon Kim, May 24, 10:50 am)
  
2. "State-of-the-Art Model  
Quantization and Optimization for  
Efficient Edge AI"  
(Hyunjin Kim, May 24, 12:00 pm)