Toward the Era of AI Everywhere

LOKWON KIM

CEO

DEEPX

DEEPX
FOR AI EVERYWHERE

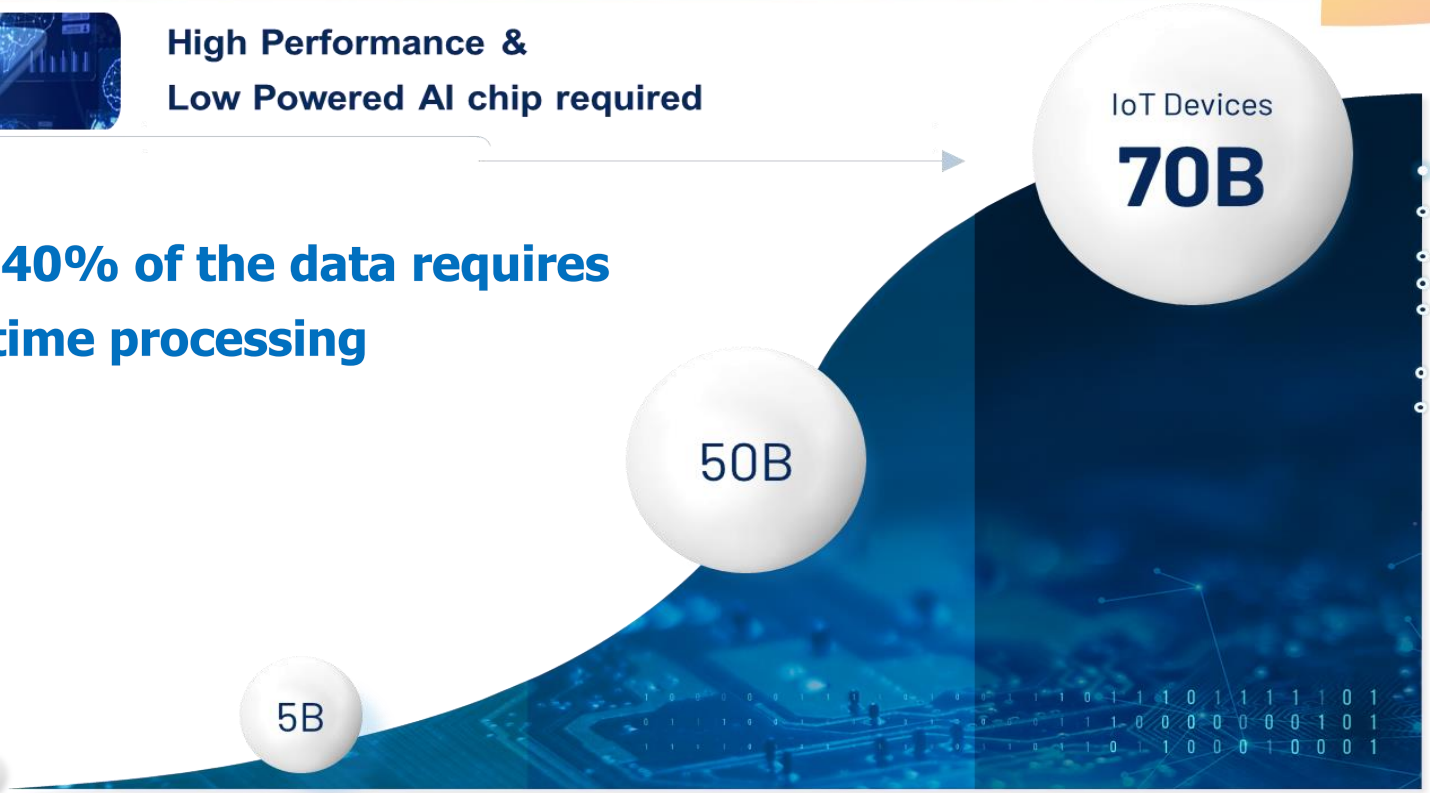
Rise of Edge AI

DEEPX
FOR AI EVERYWHERE

Hyper Connectivity

High Performance &
Low Powered AI chip required

Over 40% of the data requires real-time processing



CES2023 - Takeways

Everything with wheels has been demonstrated to move autonomously.

All camera or sensor companies demonstrated object detection solutions.

Blocking Points for AI Everywhere

DEEPX
FOR AI EVERYWHERE

Mismatch Between AI and Edge devices

- ✓ **Processing AI requires greater data and computations than any other algorithm created by humans.**
- ✓ **AI algorithms are rapidly evolving and becoming more intelligent, adding new mathematical operators and requiring more data and computations.**

→ **By nature, AI is ill-suited for edge device applications.**

→ **Need a solution that goes beyond limits of theory!**

#1: Conventional Solution

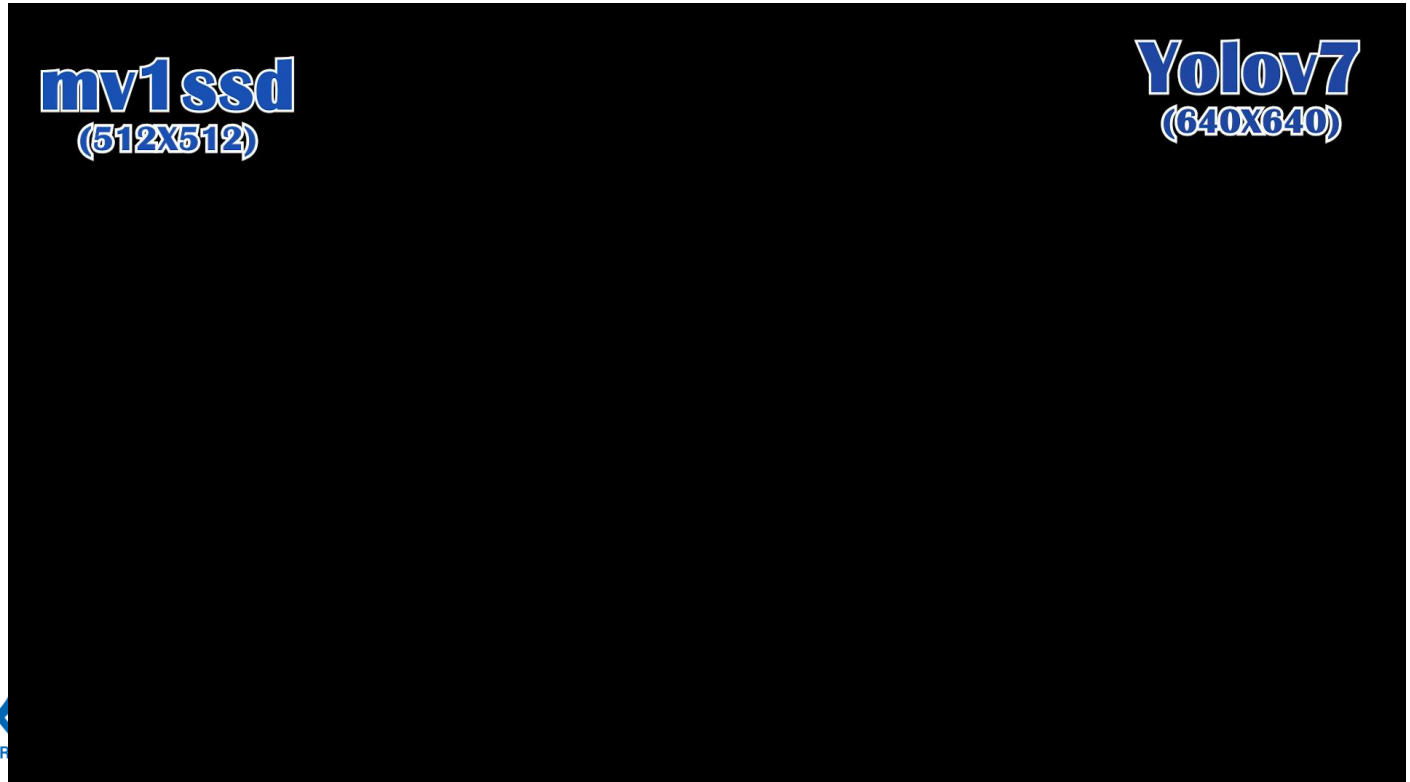
Limitations in GPU for on-device AI

- Poor AI Performance and High Cost (Over $1000)
- High Energy Consumption (30~60W) and Poor Battery Time
- Over Heat (Reliability for some applications)

DEEPX
FOR AI EVERYWHERE

#2: Edge Applications Require SOTA AI Models

Some NPU solutions improved performance and power efficiency, but …



#3: A ~1% Accuracy Loss Is Too Much!

✓ **Using 8-bit integer instead of 32-bit floating point is the key for power and area efficiency on resource-constrained edges**

→ Normally, results in AI accuracy drop compared to GPUs.

✓ **With GPUs,** edge application developers create AI models for their own applications.

→ The actual intelligence of the application **is determined by edge NPUs.**

→ **This incurs another loop of the AI accuracy validation for commercial applications which incurs longer development time and effort.**

A 1% AI accuracy drop is too much for easy deployment of edge AI.

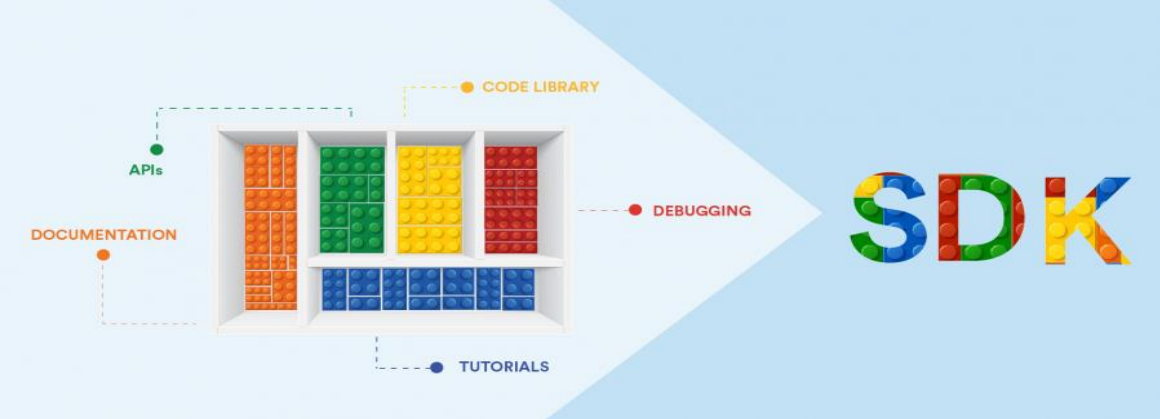
Target Accuracy Drop < 1% ?

#4: A Single One-Size-Fits-All Solution?

Intelligented by DEEPX

- ✓ There are various AI edge device applications.
- ✓ Each edge device AI application has <u>specific function, performance, and cost requirements</u>.
- → A family of chips with different functions, price, and performance is necessary for different applications.

#5: Developing Edge AI Software Takes Too Much Time and Effort



- ✓ **Edge AI application developers demand an easy-to-use SDK.**
  (In the Computer Vision Developer Survey conducted in late 2022 by the Edge AI and Vision Alliance, "**algorithm implementation/optimization**" was named as one of the most challenging aspects of product development (37%).)
- ✓ Some SDKs require several months to learn how to use.
- ✓ Some SDKs require understanding of hardware architecture and deep technical features.

→ Not easy to deploy AI!

DEEPX Disruptive Strengths

DEEPX
FOR AI EVERYWHERE

DEEPX Mission

DEEPX's Key Differentiators

World Leading
SOTA AI 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	**41.1**
	Yolo7m	51.0	N/A	50.9
IC*	MobileNetv1	71.48	70.13	**72.42**
	ResNet50	75.94	74.69	**75.95**
	EfficientNet-B0	77.52	76.96	**77.62**
Seg*	BiseNet	75.19	N/A	**75.97**
	PIDNet	78.76	N/A	**78.79**
	DeepLabv3+	72.07	N/A	**72.37**

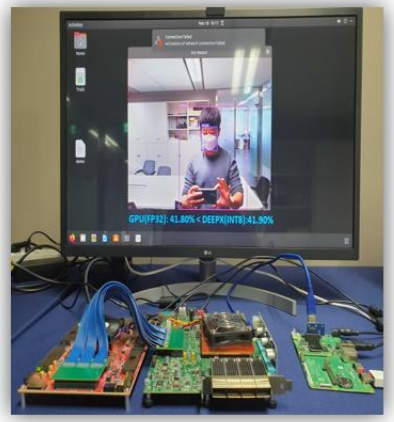
The World's best
Power/Performance Efficiency

Company	TOPS/W Resnet-50	FPS/TOPS Resnet-50
DEEPX	> 10	60
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 FOR AI EVERYWHERE

A Successful Customer Collaboration



Performance	GPGPU based edge solution	DEEPX NPU IP (FPGA)	DX-M1
			
	32TOPS **24FPS**	1TOPS **30FPS**	23TOPS **Est. 240FPS**

Accuracy	GPGPU (FP32)	DEEPX NPU (INT8)	Delta
	41.8%	**41.9%**	**0.1%↑**

- **The power/performance efficiency of DEEPX NPU is 10X higher than GPU.**
- **The AI accuracy of DEEPX NPU is higher than GPUs.**

A Comparison with a Reference Platform

Models		Throughput (FPS): NVIDIA Jetson Orin	Throughput (FPS): DX-M1	Delta (%)	NVIDIA Inference/$	DEEPX Inference/$
		GPU+DLA (200TOPS) 30W / $999	NPU (23TOPS) 5W / $30-$80			
IC	MobileNetv1 (224x224)	3530	**5751**	163%	3.53	**71.89**
	MobileNetv2 (224x224)	2072	**4586**	221%	2.07	**57.33**
	MobileNetv3-Large (224x224)	1034	**4228**	409%	1.04	**52.85**
	ResNet50 (224x224)	1367	**1337**	-2%	1.37	**16.71**
	EfficientNet-B0 (224x224)	613	**3236**	528%	0.61	**40.45**
OD	YOLOv4 (800x800)	62	**62**	-	0.06	**0.78**
	YOLOv5s (640x640)	551	**590**	7%	0.55	**7.38**

10X **more performance-power efficient**

10X **more performance-cost efficient**

DEEPX FOR AI EVERYWHERE

Product Roadmap (2023)

01 DX-M1

DX M1

- Performance: 200 eTOPS**
- Type: AI Accelerator in M.2
- Features: PCIe, ARM CPU
  LPDDR5
- Launch Date: 23.2Q

02 DX-H1

DX H1

- Performance: 1,600 eTOPS**
- Type: AI Accelerator in PCIe Card
- Features: PCIe Card
- Launch Date: 23.2Q

03 DX-L1

DX L1

- Performance: 12 eTOPS**
- Type: SoC
- Features: RISC-V CPU, ISP, MIPI
  LPDDR4, Video Codec
- Launch Date: 23.2Q

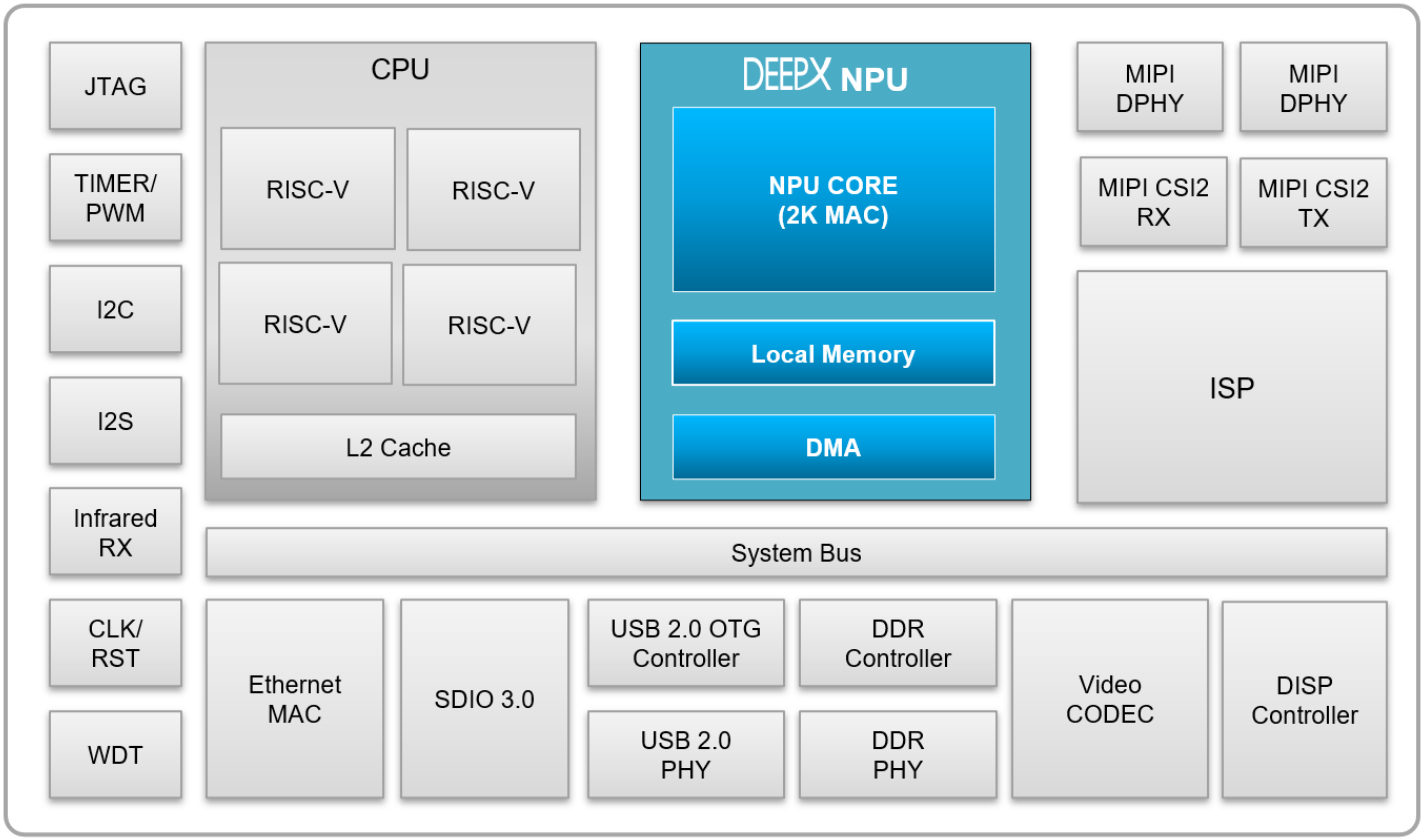
04 DX-L2

DX L2

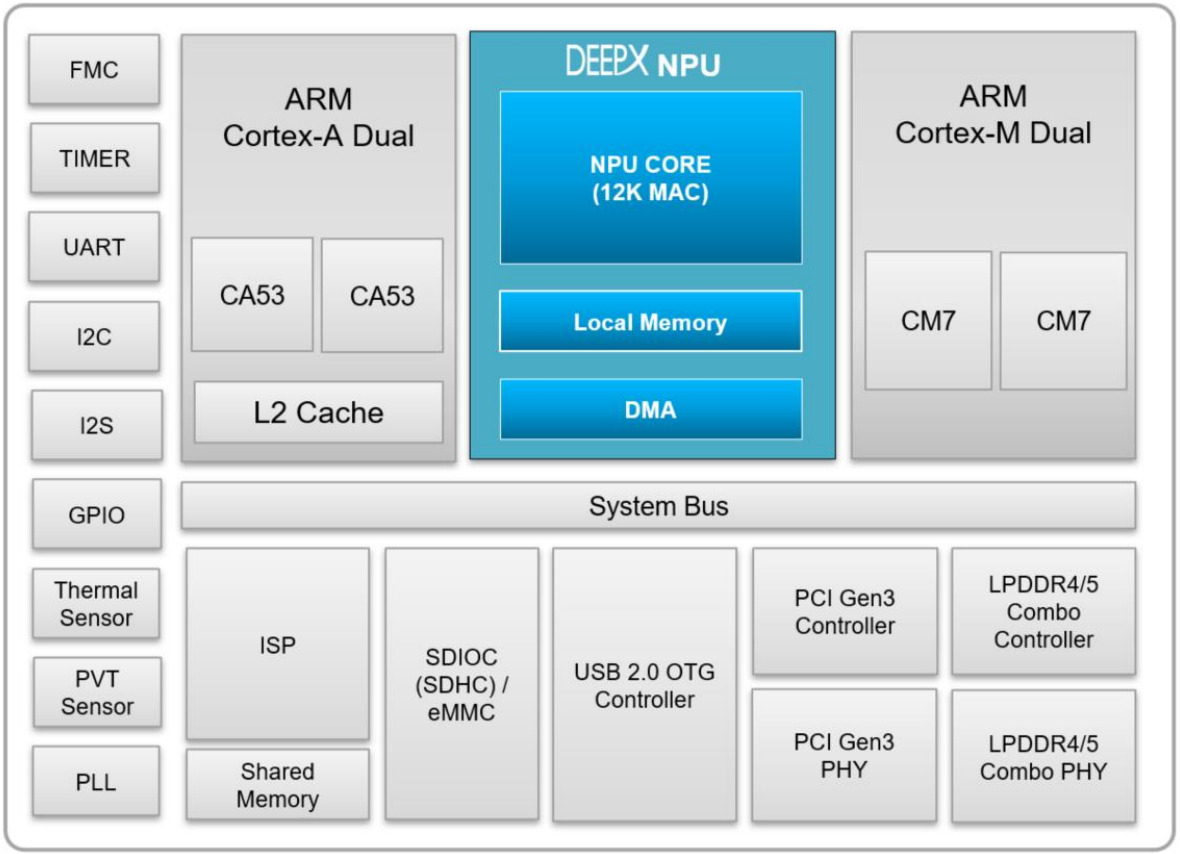
- Performance: 38 eTOPS**
- Type: SoC
- Features: RISC-V CPU, ISP, MIPI
  LPDDR4, Video Codec
- Launch Date: 23.2Q

** A performance metric equivalent to GPU's TOPS

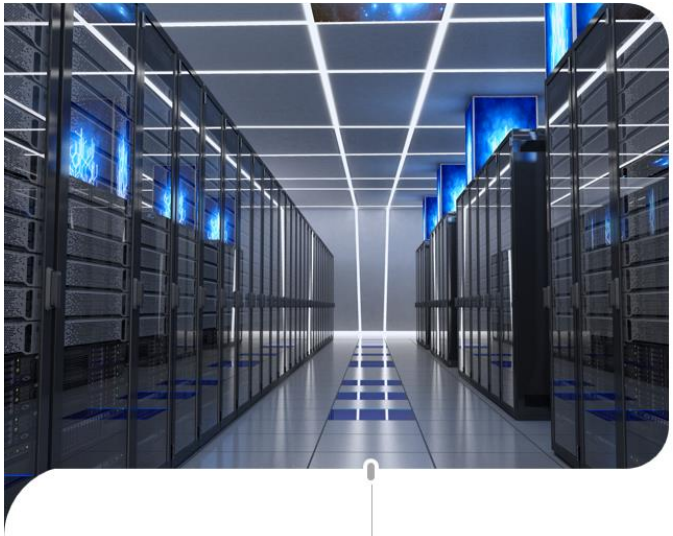
DX-L1 Targets Single-camera AI Applications



DX-M1 Targets Clustered Cameras or Workloads



AI Inference Server Solution for ESG/TCO

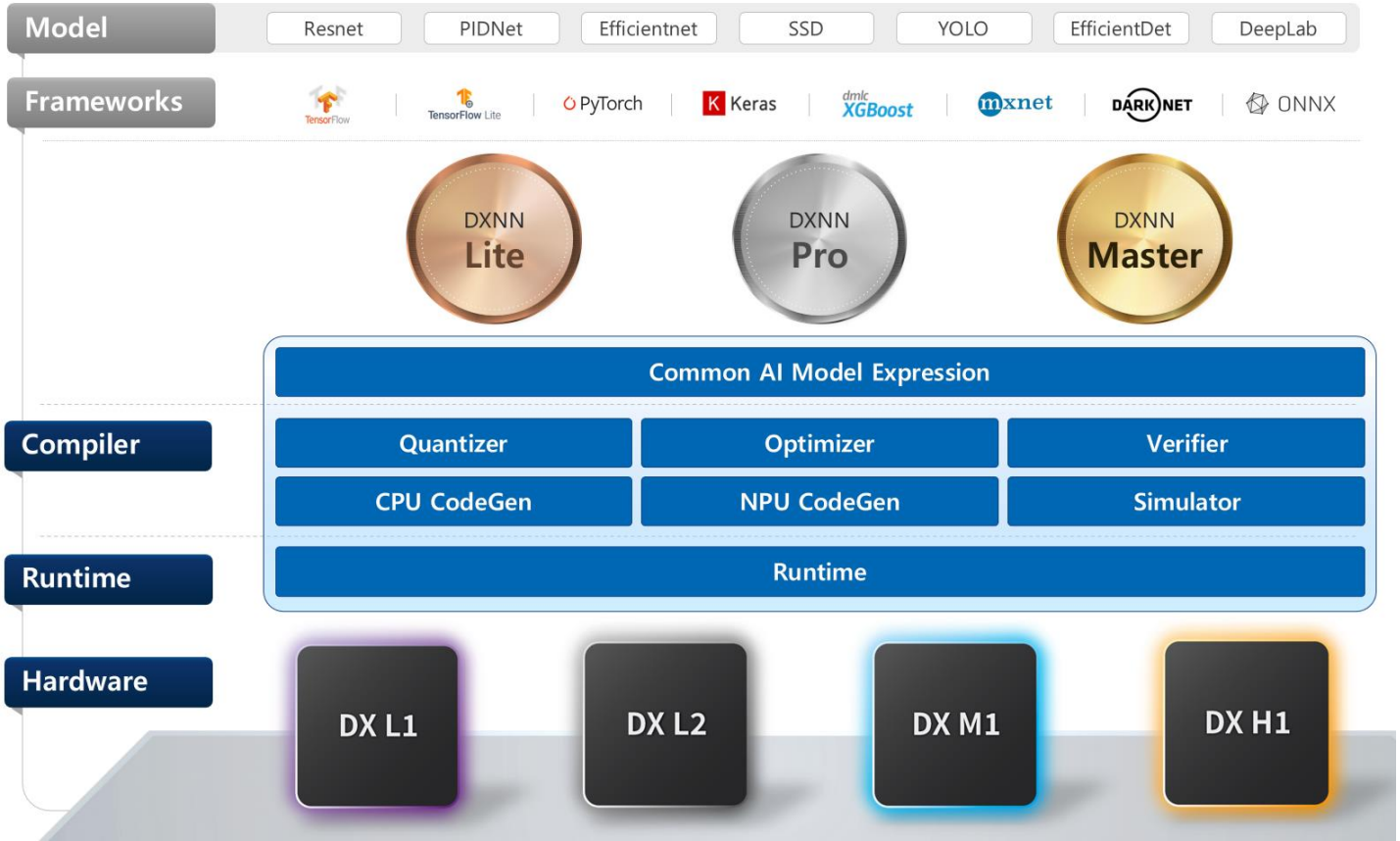


DX-H1

- Performance: Up to 18POPS
- Features: PCIE Card, Server, Rack
- Launch Date: 23. 3Q

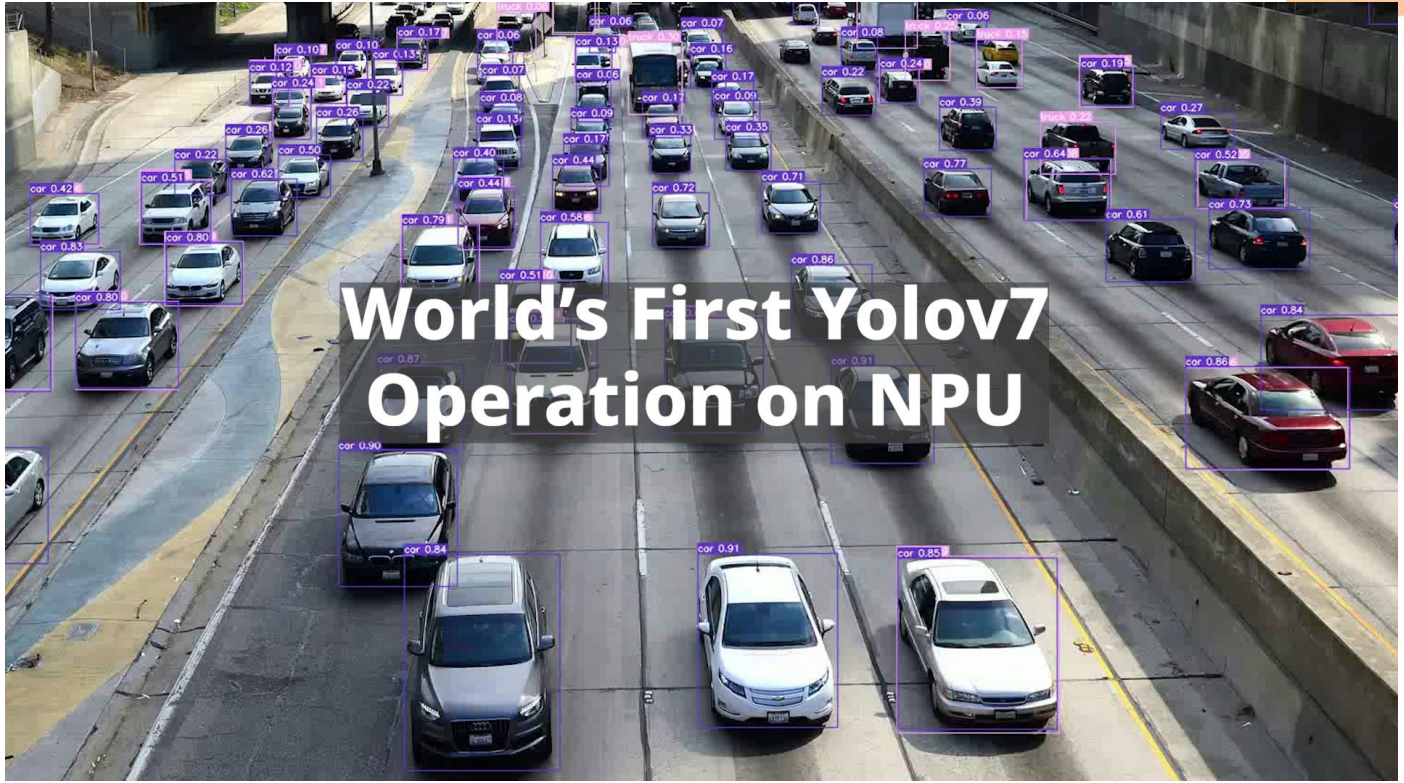
* DX H1C: PCIe Card (184TOPS)
* DX H1R: DX H1S x 6ea (18POPS)
* DX H1S: DX H1C x 10ea (1.8POPS)

DXNN – SDK for DEEPX NPUs



DEEPX
FOR AI EVERYWHERE

Enjoy the SOTA AI on All DEEPX Solutions



DEEPX
FOR AI EVERYWHERE

Representative Project #1

Hyundai Motors

Mobility of Everything

- **Smart Camera Sensor or AP (PoC)**
  - ✓ AI Performance
  - ✓ Price

DX M1

DX L2

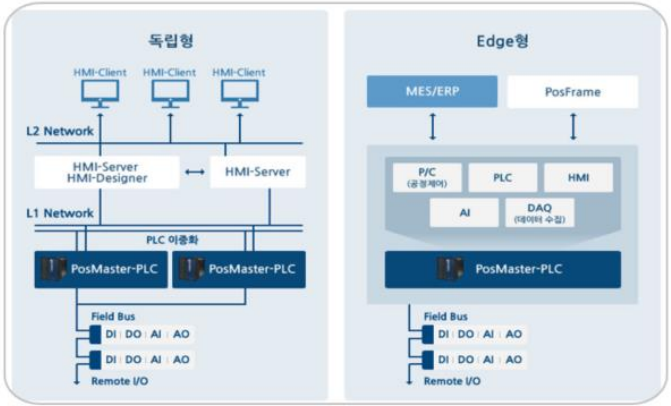
DX L1

Representative Project #2

POSCO DX

Edge Computing Solution for Smart Factory

- **AI based PosMaster (Machine Vision & Machine Maintenance**
  (In Development)



DX H1

DX M1

DX L2

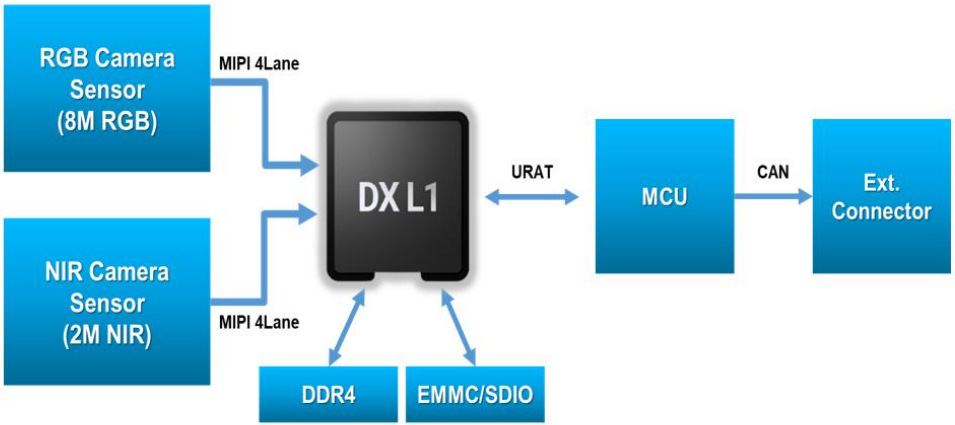
DX L1

Representative Project #3

Jahwa

Smart Camera Module Development for Automotive

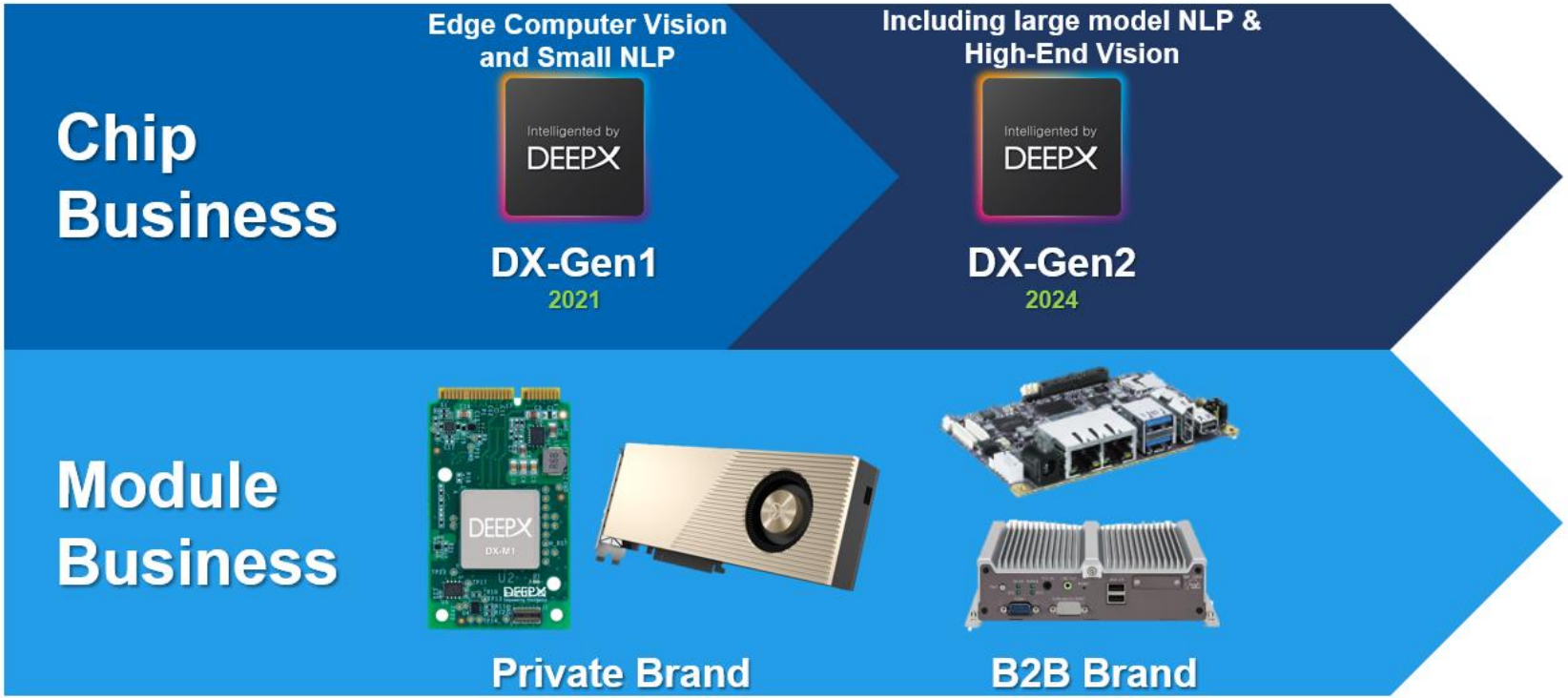
- **DSM & Smart Camera Module** (In Development)



DX L2

DX L1

Product Strategy



Summary

Edge AI requires performance, low power, low cost, SOTA AI algorithms, AI accuracy, and ease of use.

DEEPX NPU can:

1. Run **SOTA AI models**.
2. Get the **best AI accuracy**.
3. Achieve the **highest performance, power efficiency, and cost efficiency** (including BOM cost).

Please visit our demo booth and check it out!

Thank you!!

Company Info, Demo and Additional Talk

1. Demo Booth: #103

2. DEEPX Homepage

https://www.deepx.ai

3. Linkedin & Youtube

2023 Embedded Vision Summit

Additional Talks from DEEPX

"State-of-the-Art Model Quantization and Optimization for Efficient Edge AI"

(Hyunjin Kim, **Wednesday, May 24, 12:00 pm**)