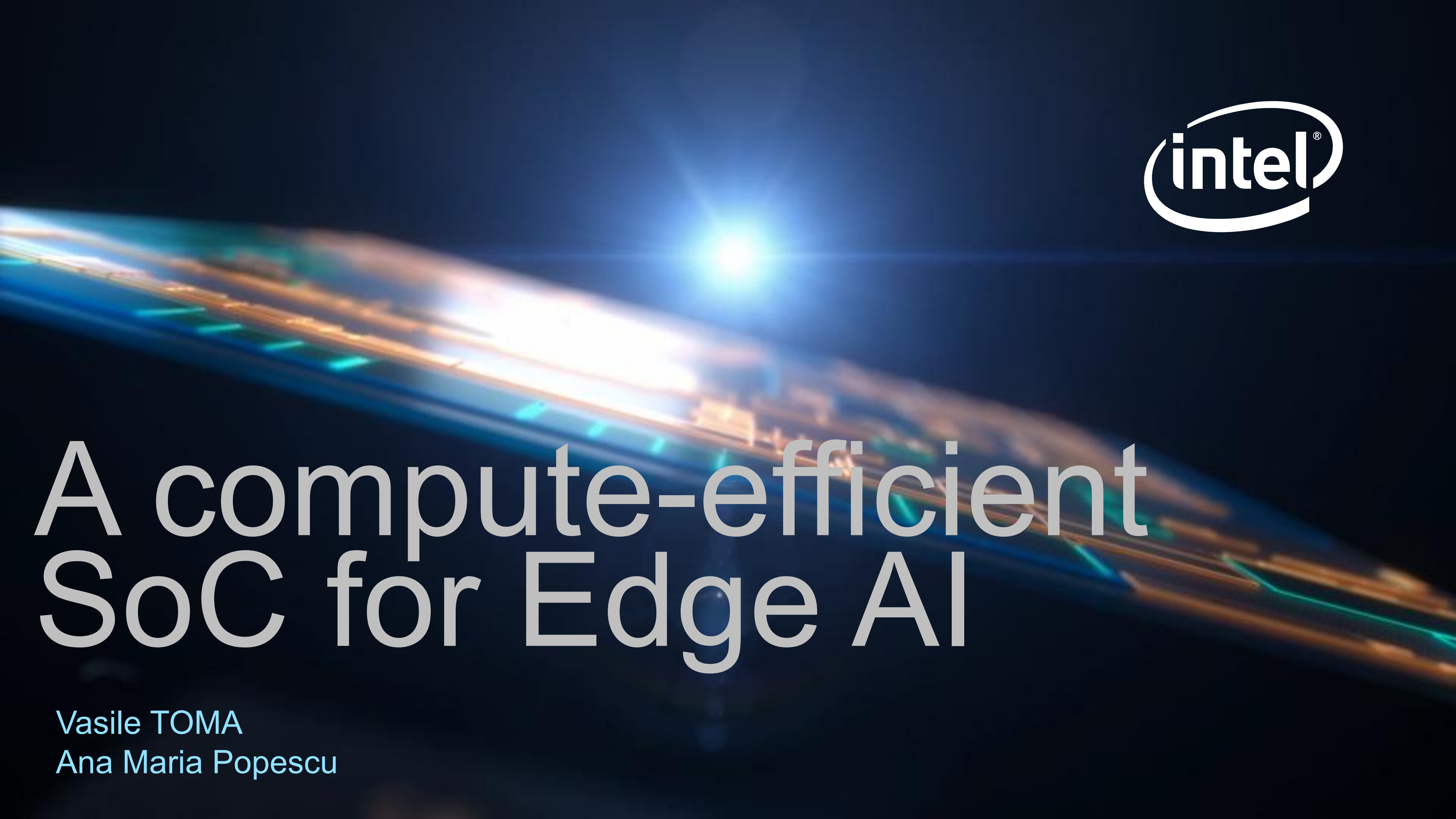


A compute-efficient SoC for Edge AI

Vasile TOMA
Ana Maria Popescu

Table of contents

- The Why ?
  - Why are we doing this (business)?
- The what?
  - What are we doing?
- The how ?
  - How are we doing this?
  - Keembay architecture overview
  - NCE acceleration
  - DPU
- Q&A

The Why ?

To help these industries make people's lives….

SAFER

Smart Cities

Public Safety & Security
Traffic, Parking and LPR
Emergency Response

MORE EFFICIENT

Financial Services

People Counting
Reduce Customer Wait Time
ATM Facial Recognition

HIGHER QUALITY

Industrial

Machine Vision
Asset Inspection (i.e., Pipeline)
Augmented Reality

MORE FUN

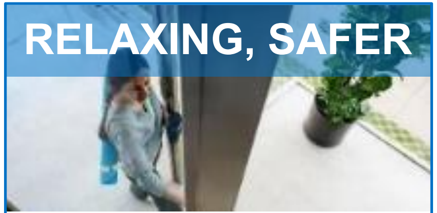
Casino Gaming

Public Safety & Security
Facial Recognition

PRODUCTIVE

Transportation

Autonomous Vehicles
Public Safety (i.e., Bus/Rail)
Traffic & People Counting



RELAXING, SAFER

Home & Retail

Security
Responsive Retail Advertising
Digital Home Assitant

MORE EFFICIENT

Robotics

Manufacturing Automation
Industrial (i.e., Pipeline Welding)

HEALTHFUL

Medical Imaging

Segmentation e.g. for MRI
Scans
to detect disease

Yea, but also…

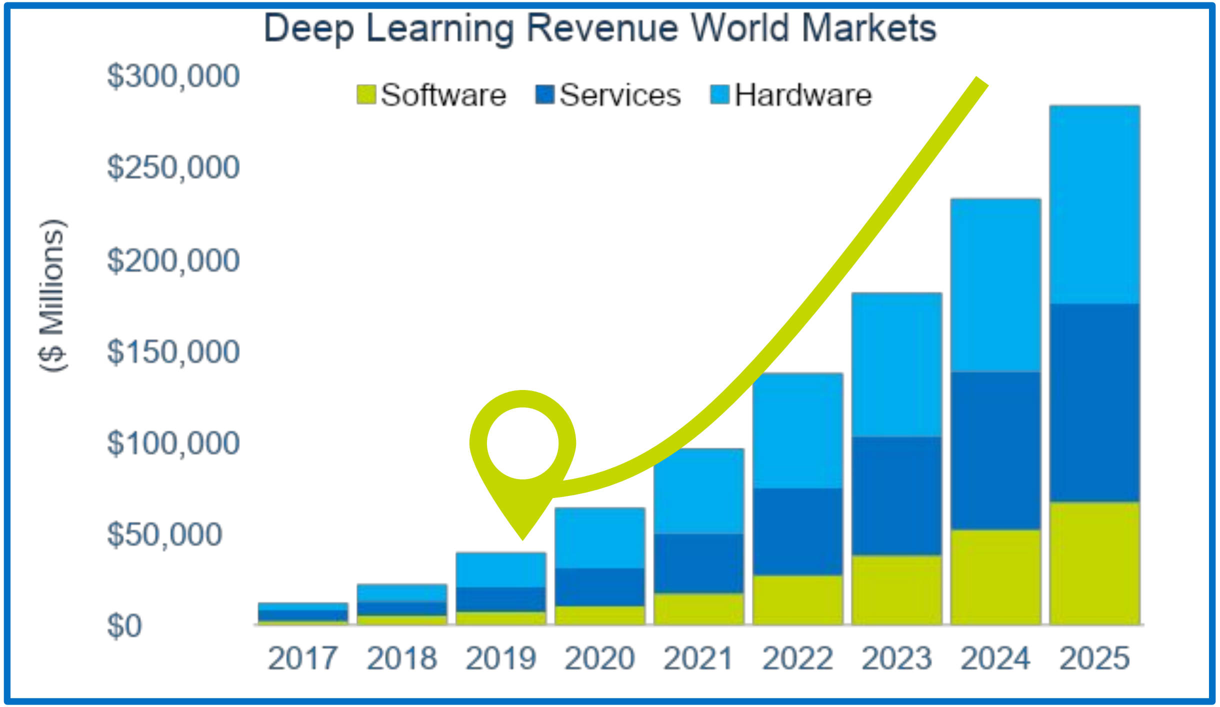


Chart Source: Source: Tractica

- AI to create $13 TRILLION additional economic activity by 2030[1]
- Machines to impact jobs—in a good way 58M NEW JOBS by 2022[2]

[1] https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-AI-frontier-modeling-the-impact-of-ai-on-the-world-economy

[2] https://www.forbes.com/sites/amitchowdhry/2018/09/18/artificial-intelligence-to-create-58-million-new-jobs-by-2022-says-report/#710234644d4b

The What…   Vision Processing Unit (VPU)?

The Intel® Movidius™ VPU is a class of processors that

Shown: Intel® Movidius™ Myriad™ X VPU

- processes sensory inputs
- computes vision or AI inference algorithms to extract meaning from multi-modal sensor data
- and has a uniquely power-efficient architecture that includes
  (1) custom accelerator engines,
  (2) programmable processors, and
  (3) a central scratchpad memory

Keem Bay VPU SKUs support accelerator or standalone use

Summary	Performance Optimized, Edge AI Processor (Accelerator mode)	Performance Optimized Smart Camera SOC (Camera mode)	Performance Optimized, Edge AI Processor	Performance Optimized Smart Camera SOC
Process **Clock Frequency**	12 nm TSMC 500 MHz (Nominal)		12 nm TSMC 700 MHz (Nominal)	
ResNet-50 Performance; **Max TOPS (AI Inference)**	406 inferences/sec 5.1 TOPS (5x4=20DPU)	318 inferences/sec 4.0 TOPS (4x4=16DPU (Max))	565 inferences/sec 7.1 TOPS (5x4=20DPU)	341 inferences/sec 4.3 TOPS (3x4=12DPU)
Performance Measurement Configuration Details	INT8; Batch Size=1; employing native optimizations; ResNet-50 running standalone; ResNet-50 model trained using weight sparsity at 50%; Optimizations for max performance; measured with pre-production silicon and tools; results may vary based on tools release used.			
Computer Vision Support	CV/Warp Acceleration 1.4 GP/s; Stereo Depth 720p @ 180; 6DOF Motion Mask support; Motion Estimation 1080p60 ± 32			
Video CODEC	4K75 (Encode) 4K60 (Decode); Decode: 10 channels of 1080 30fps			
ISP **(Available capabilities)**		Up to 6 Cameras 500 MP/s HDR, TNF		Up to 8 Cameras 700 MP/s HDR, TNF
SHAVE Processors included	16	12	16	12
CPU	4x ARM* A53, OS Supported: Yocto 2.7.1 (codenamed Warrior) with Linux* Kernel 5.3			
Security	Secure boot, Trust Zone Enabled A53, TEE, Crypto HW (AES, SHA, ECC); Fuse based Key & SVN			
Interfaces	PCIE Gen4 x2, USB3.1/2 eMMC, SPI, MIPI RX, MIPI TX, SLVS, I3C, I2S		PCIE Gen4 x2, USB3.1/2 eMMC, SPI	PCIE Gen4 x2, USB3.1/2 eMMC, SPI, MIPI RX, MIPI TX, SLVS, I3C, I2S
DRAM	2x 32-bit LP4 1600-2133 MHz			
Production (Expected PRQ)	Early Q2’20		Late Q3’20	

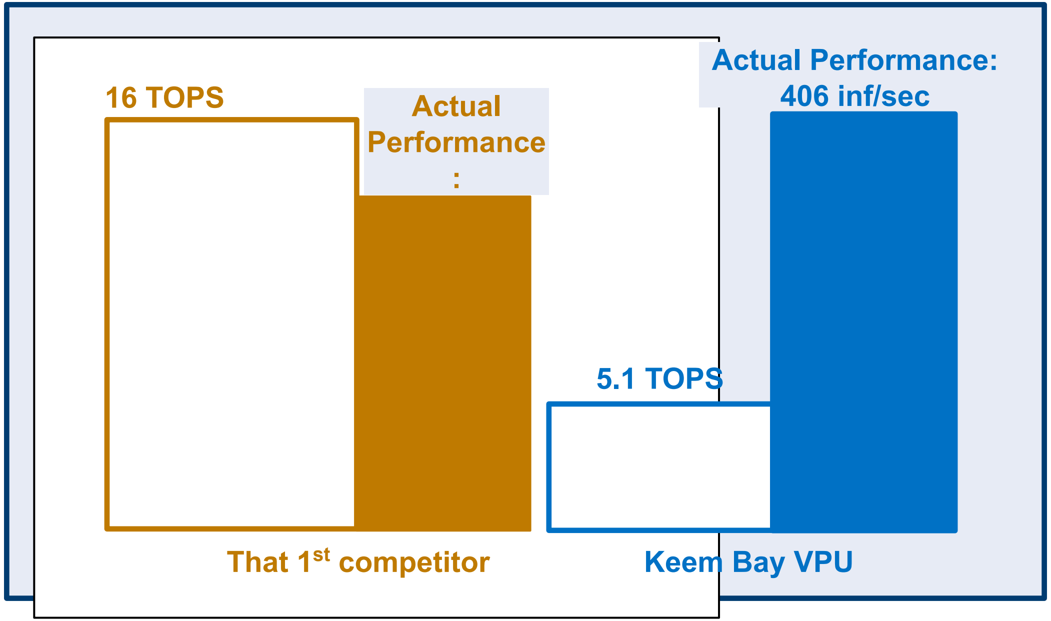
What? … Are the results ?

DNN	Performance measured SKU: 3400VE	Power Efficiency SKU: 3400VE
ResNet-50	406 inferences/sec	139 inf/sec/W

What is the competition doing ?

Comparison	Usage	ResNet-50 Performance SKU: 3400VE	Performance per Watt SKU: 3400VE	Keem Bay VPU Efficiency is:
Keem Bay VPU	**IP Camera, AI Appliance**	**406 inferences/sec**	**139 inf/sec/W**	-
1st Competitor	AI Appliance Only	319 inferences/sec	40 inf/sec/W	**3.5x** vs. C1
2nd Competitor	IP Camera, AI Appliance	95 inferences/sec	9.5 inf/sec/W	**14.6x** vs. C2
3ed Competitor	AI Appliance Only	438 inferences/sec	15 inf/sec/W	**9.3x** vs. C3

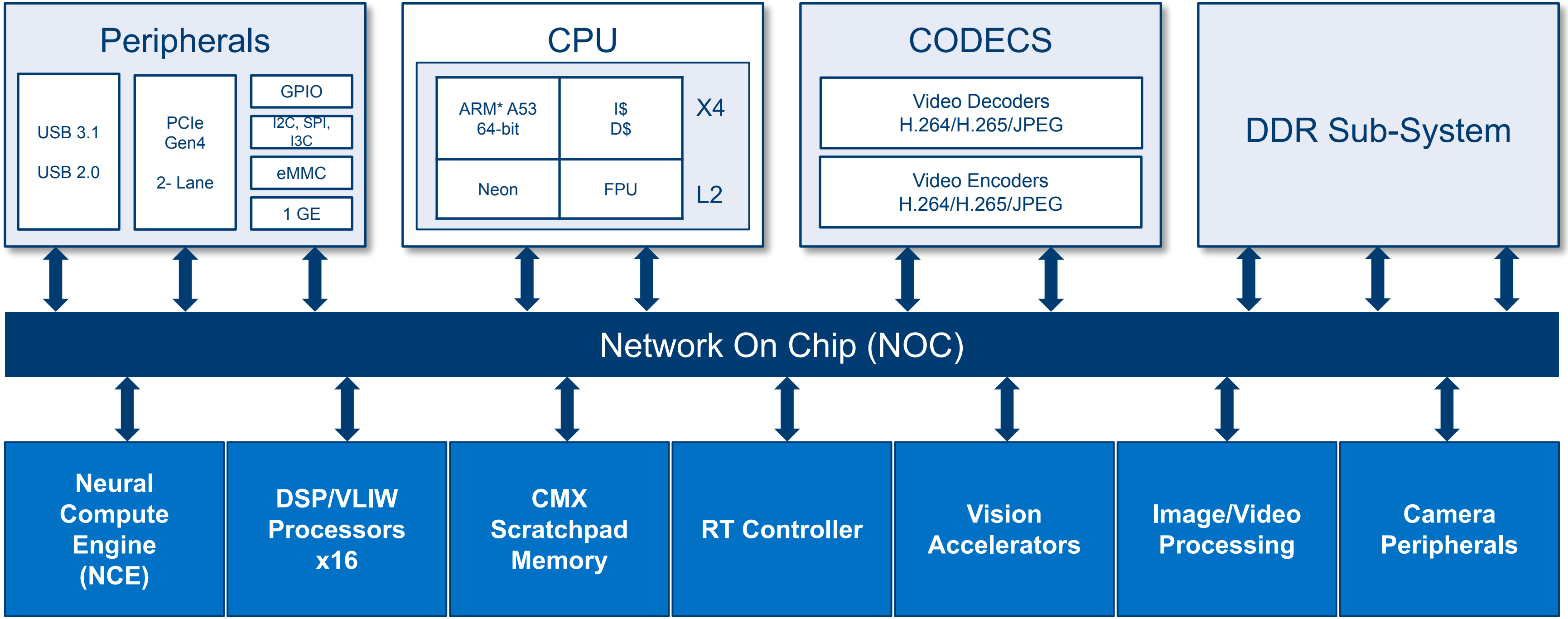
When comparing architectures for DL inference, actual workload performance matters more than peak TOPS



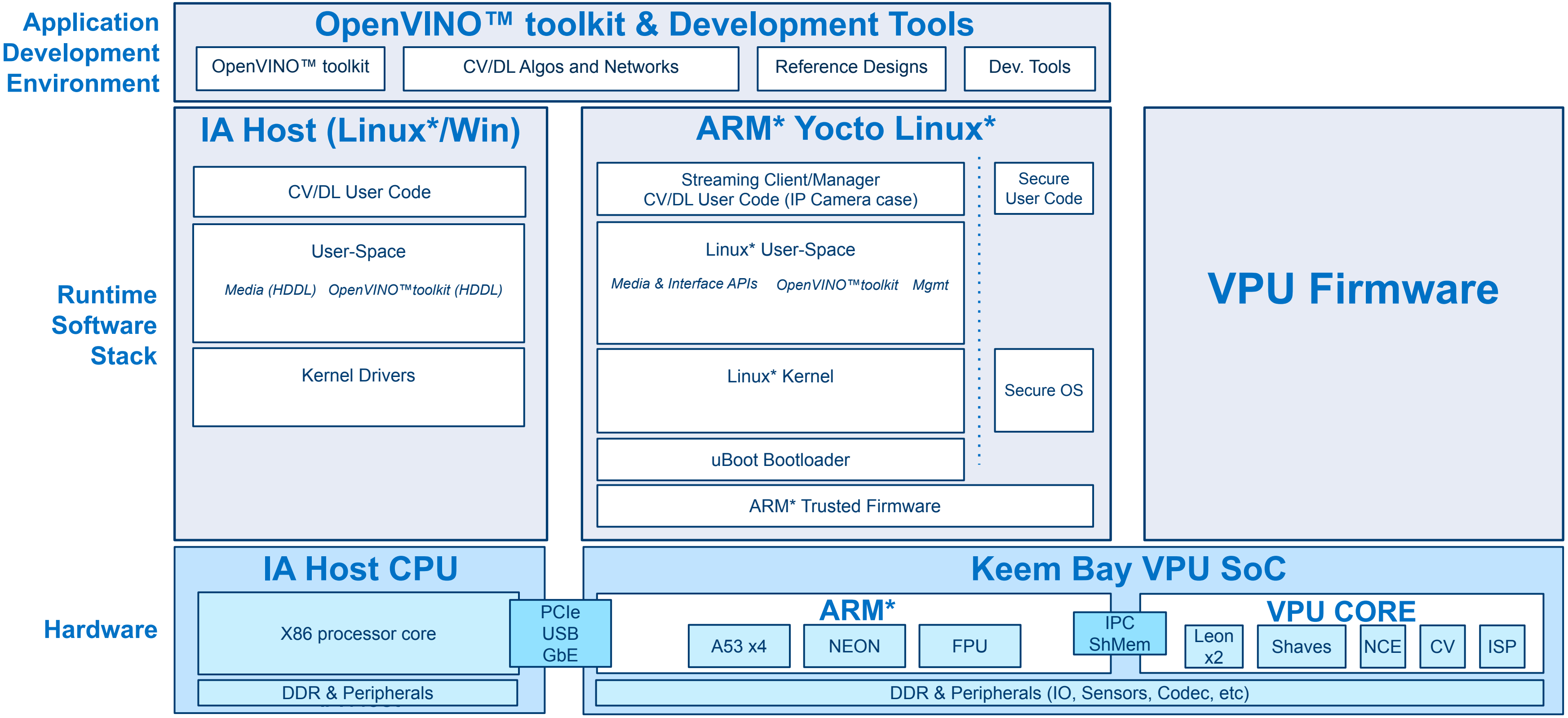
Keem Bay VPU architecture is efficient.
With 1/3 the TOPS compute available, Keem Bay VPU still gets >1.25x the performance, which means it's 4x the compute efficiency.
This comes from Keem Bay VPU's architectural advantages such as:

- Native optimizations to speed-up DNNs
- Memory architecture
- Dataflow between compute units

The How?... Keem Bay VPU SoC Architecture



Keem Bay VPU software stack allows efficient development of vision and DL application pipelines



Application Development Environment

OpenVINO™ toolkit & Development Tools

OpenVINO™ toolkit

CV/DL Algos and Networks

Reference Designs

Dev. Tools

Runtime Software Stack

IA Host (Linux*/Win)

CV/DL User Code

User-Space

Media (HDDL) *OpenVINO™toolkit (HDDL)*

Kernel Drivers

ARM* Yocto Linux*

Streaming Client/Manager
CV/DL User Code (IP Camera case)

Secure User Code

Linux* User-Space

Media & Interface APIs *OpenVINO™toolkit* *Mgmt*

Linux* Kernel

Secure OS

uBoot Bootloader

ARM* Trusted Firmware

VPU Firmware

Hardware

IA Host CPU

X86 processor core

PCIe
USB
GbE

DDR & Peripherals

Keem Bay VPU SoC

ARM*

A53 x4

NEON

FPU

IPC
ShMem

VPU CORE

Leon x2

Shaves

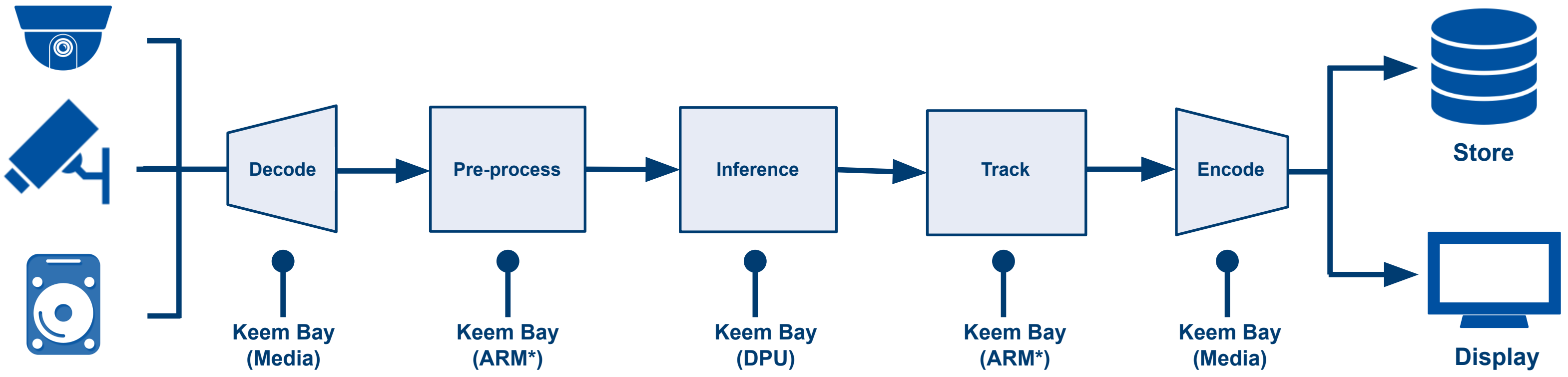
NCE

CV

ISP

DDR & Peripherals (IO, Sensors, Codec, etc)

Media Pipeline Programming on Keem Bay made easier with Gstreamer*



Decode

Pre-process

Inference

Track

Encode

Store

Display

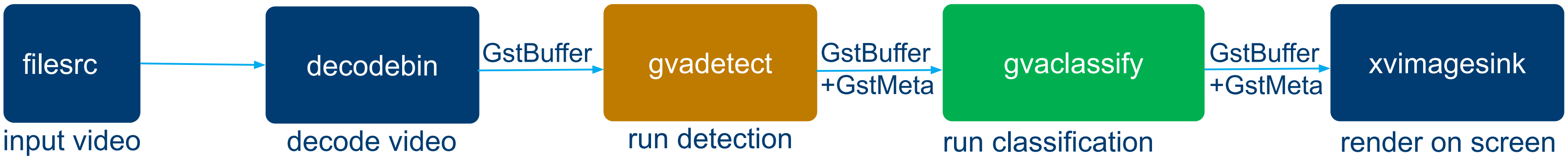
Keem Bay (Media)

Keem Bay (ARM*)

Keem Bay (DPU)

Keem Bay (ARM*)

Keem Bay (Media)



filesrc

input video

decodebin

decode video

GstBuffer

gvadetect

run detection

GstBuffer +GstMeta

gvaclassify

run classification

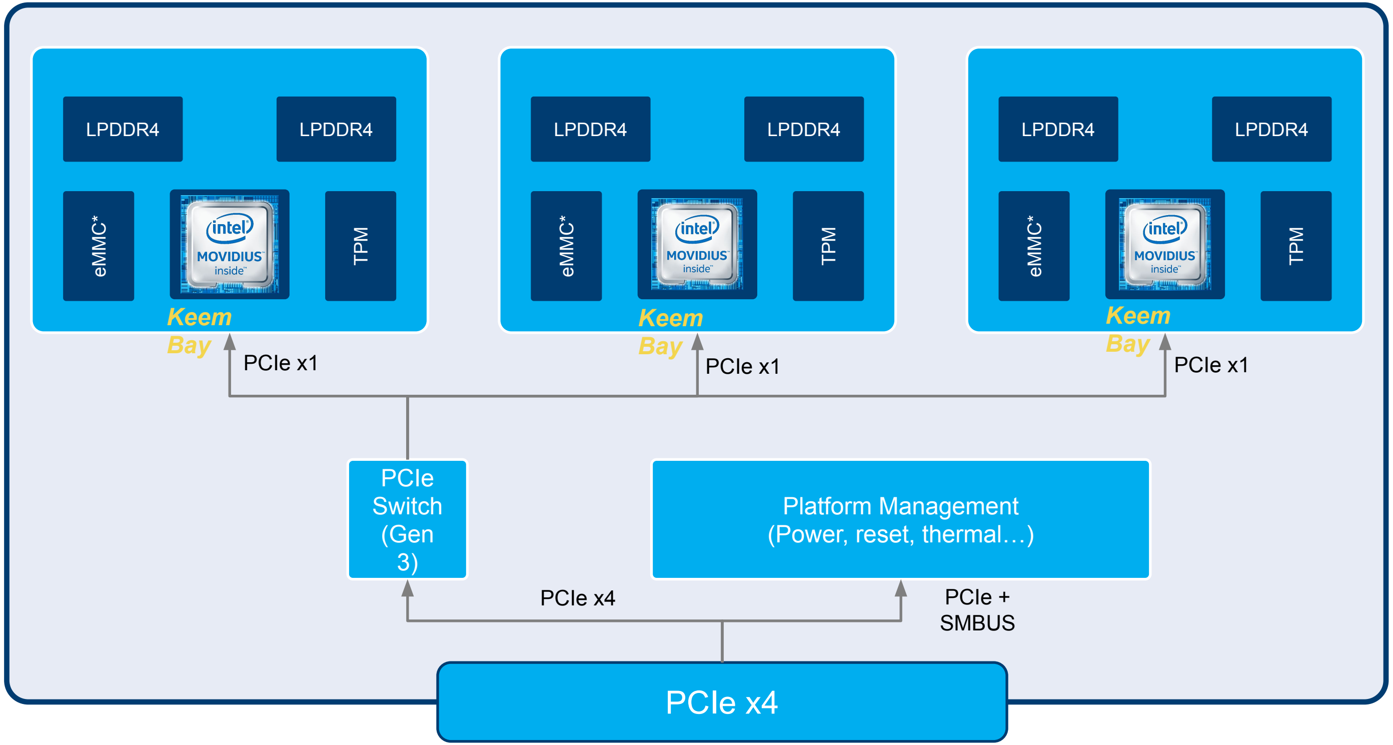
GstBuffer +GstMeta

xvimagesink

render on screen

Example: Video analytics pipeline with detect & classify with Gstreamer*

The Intel® Vision Accelerator Design with x3 Keem Bay VPU provides significant DL inference performance in a power-efficient PCIe form factor

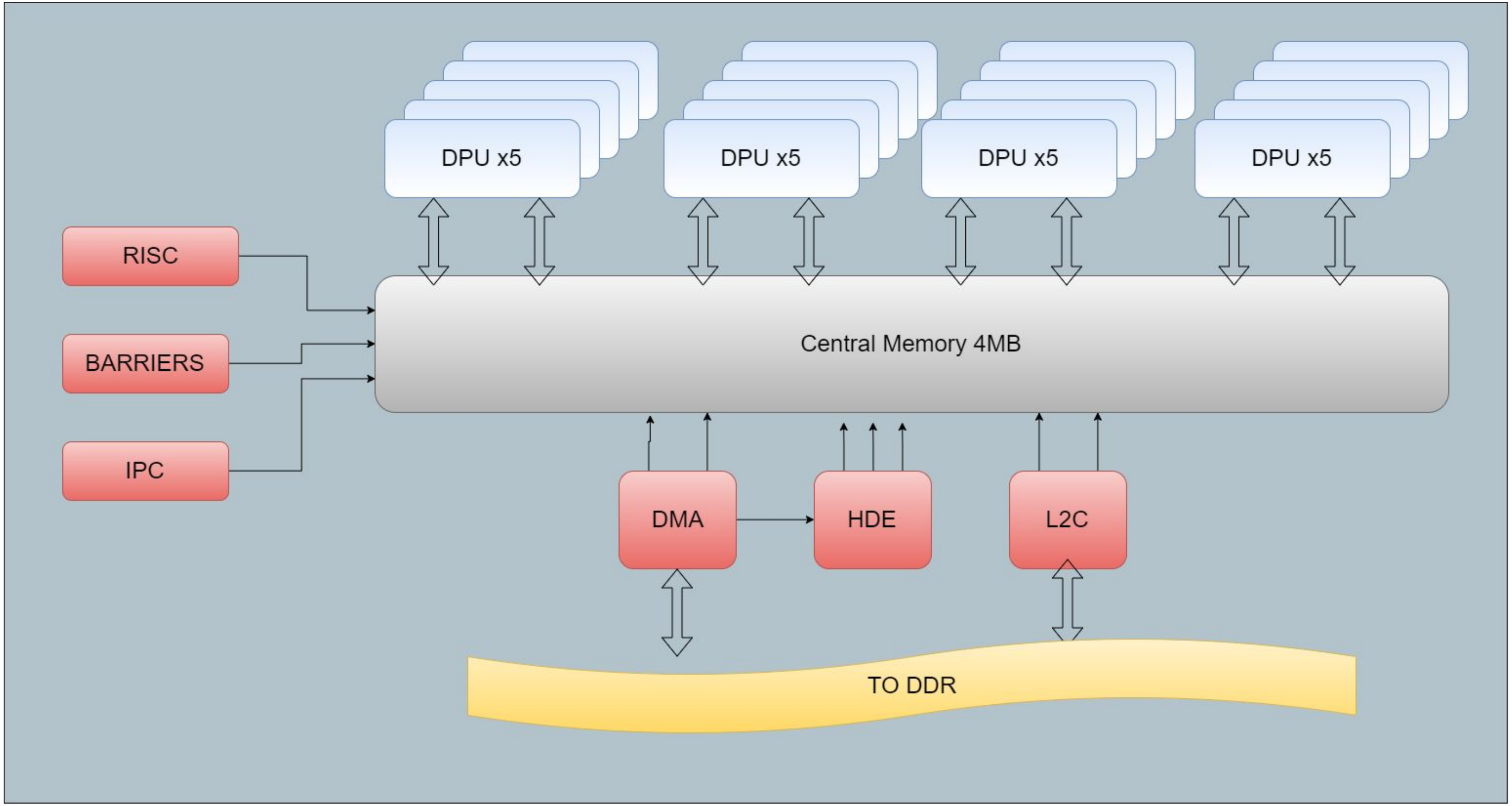


Keem Bay VPU x3 Design: Half-height half-length PCIe

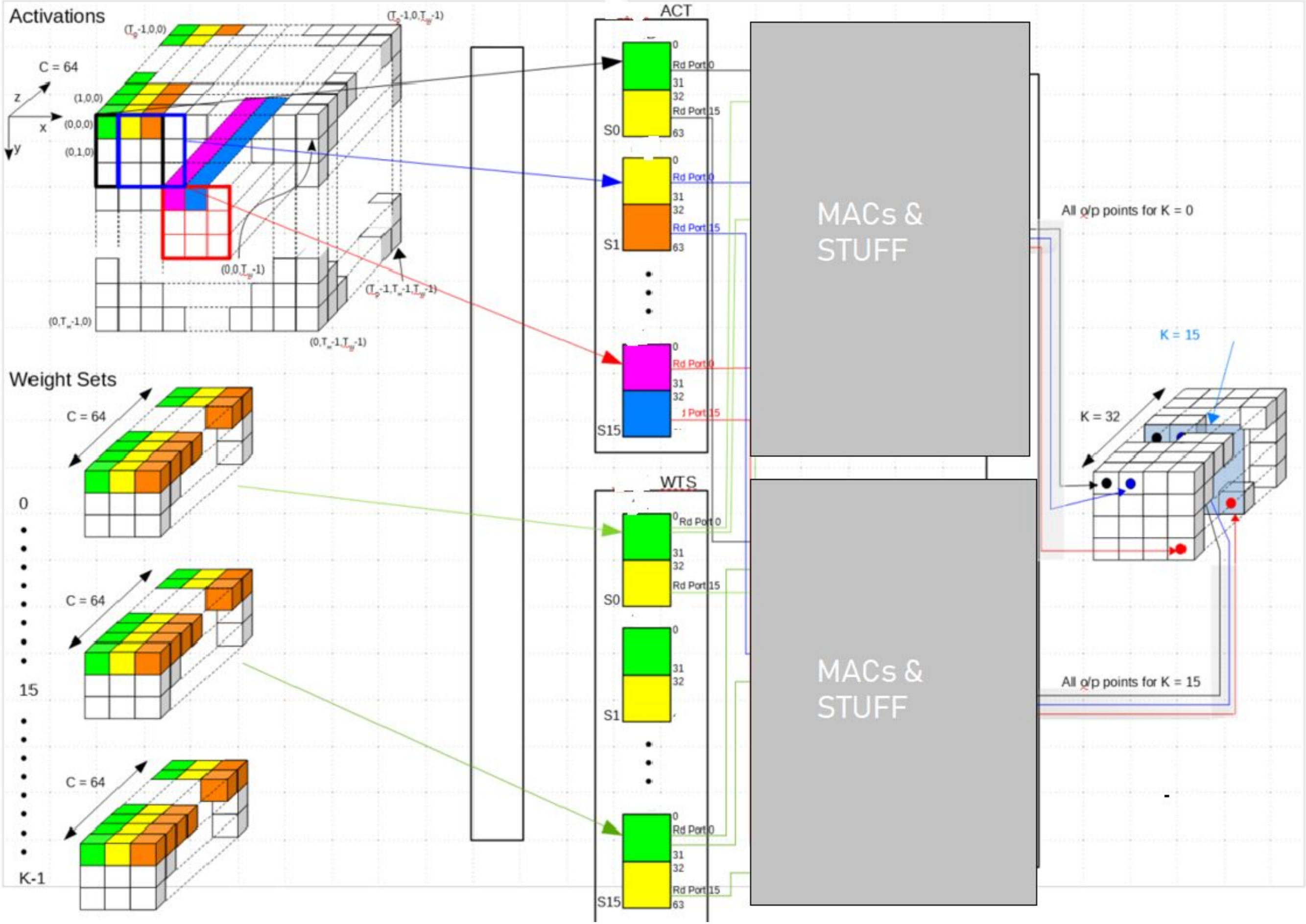
What CNN actualy does?

- Convolution Kernel size of M*N where N,M are up to 11
- Optimised HW support for standard, depthwise and planer convolutions
- Element-wise operations
- Flexible activation and scaler compute with micro-code engine
- Convolution/Pooling Stride: 1-8
- FP16 operation mode with ¼ FLOPS support
- Support for quantized 8-bit networks
- Low bit rate modes (4, 2, 1):
  - 4, 2 and 1 bit linear (i4, i2, i1) representation for both Activation's and Weights
- Channel dimensions up to 8192H, 8192W, 8192D, up to a maximum of 256MB
- Palettised u8/FP16 mode for weight storage reduction
- Broad activation function support including arbitrary piecewise linear mode
- Max, Average pooling
- Weight sharing for batching
- Supports Huffman compression for Weights achieving up to 25% weight compression
- 85% efficiency at MAC utilization on RESNET-50, Mobilenets, Tini-Yolo @ DPU level

NCE Architecture



Is a DPU JOB



Activations
$(T_g-1,0,0)$
$(T_g-1,0,T_w-1)$
C = 64
z
x
y
(1,0,0)
(0,0,0)
(0,1,0)
$(0,0,T_w-1)$
(T_g-1,T_h-1,T_w-1)
$(0,T_h-1,0)$
$(0,T_h-1,T_w-1)$
Weight Sets
C = 64
0
C = 64
15
C = 64
K-1
ACT
0
Rd Port 0
31
32
Rd Port 15
63
S0
0
Rd Port 0
31
32
Rd Port 15
63
S1
0
Rd Port 0
31
32
Rd Port 15
S15
WTS
0
Rd Port 0
31
32
Rd Port 15
S0
0
31
32
S1
0
Rd Port 0
31
32
Rd Port 15
63
S15
MACs & STUFF
MACs & STUFF
All o/p points for K = 0
K = 15
K = 32
All o/p points for K = 15

Q&A