

# HETEROGENEOUS COMPUTING FOR AI AT THE EDGE

Webinar Pasca Sarjana Terapan PENS

2 September 2020

Dr. -Ing. Arif Irwansyah, S.T., M.Eng.

# Arif Irwansyah

[arif@pens.ac.id](mailto:arif@pens.ac.id)

## ■ PENDIDIKAN :

- *S1 T. Elektro ITS, Surabaya*
- *S2 Universitas Teknologi Malaysia*
- *S3 Universitaet Bielefeld, Jerman*

## ■ Pengalaman:

- *Asisten Peneliti di UTM Malaysia (2007-2009)*
- *Asisten Peneliti di Univ. Paderborn Jerman (2010-2011)*
- *Asisten Peneliti di Univ. Bielefeld Jerman (2012 – 2016)*
- *Dosen dan Peneliti di PENS (2001 sd saat ini)*

## ■ Bidang Penelitian:

- *Embedded Vision*
- *Reconfigurable Hardware (FPGA)*
- *Heterogeneous Embedded Systems*



# Background

- Nowadays, Many applications are required high performance computations.



The AI Car Computer for Self-Driving Vehicles



Drones



Robotics



Video Analytic

# Accelerating Insights

*“Now You Can Build Google’s \$1M Artificial Brain on the Cheap”*

**WIRED**

## GOOGLE DATACENTER



1,000 CPU Servers  
2,000 CPUs - 16,000 cores

600 kWatts  
\$5,000,000

## STANFORD AI LAB



3 GPU-Accelerated Servers  
12 GPUs - 18,432 cores

4 kWatts  
\$33,000

*Deep learning with COTS HPC systems, A. Coates, B. Huval, T. Wang, D. Wu, A. Ng, B. Catanzaro ICML 2013*

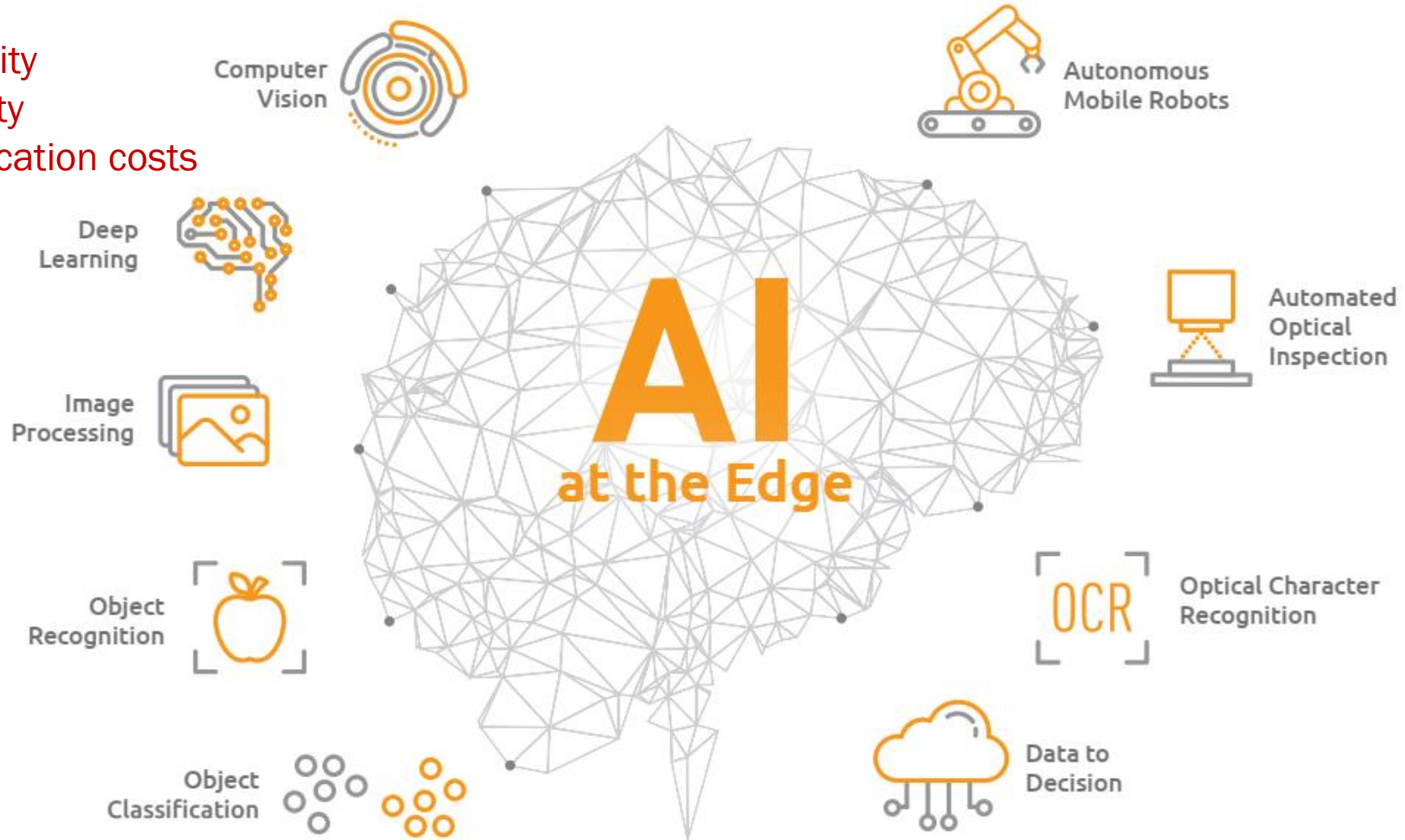
## Benefits:

Faster response

Enhanced security

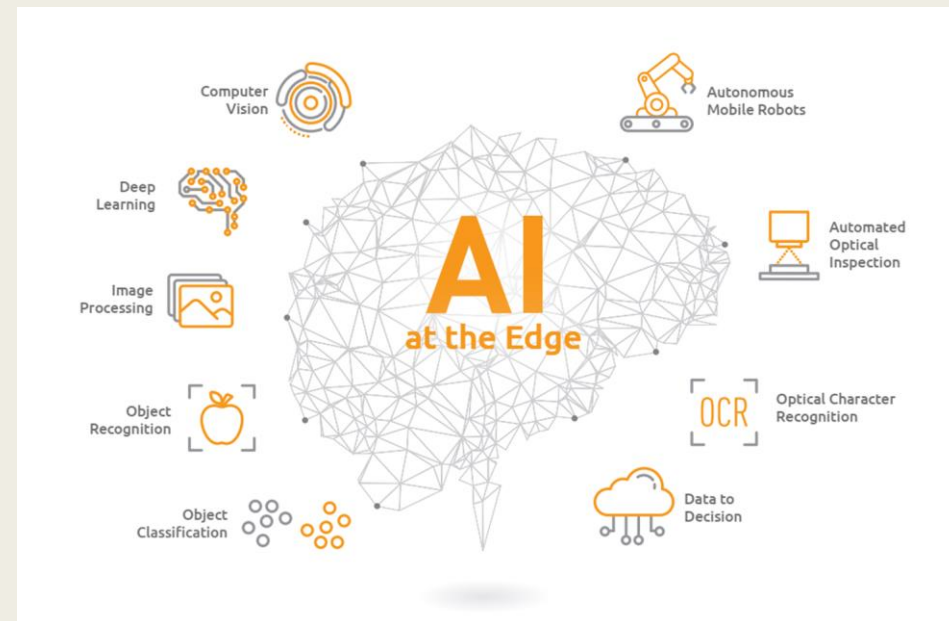
Improved mobility

Lower communication costs

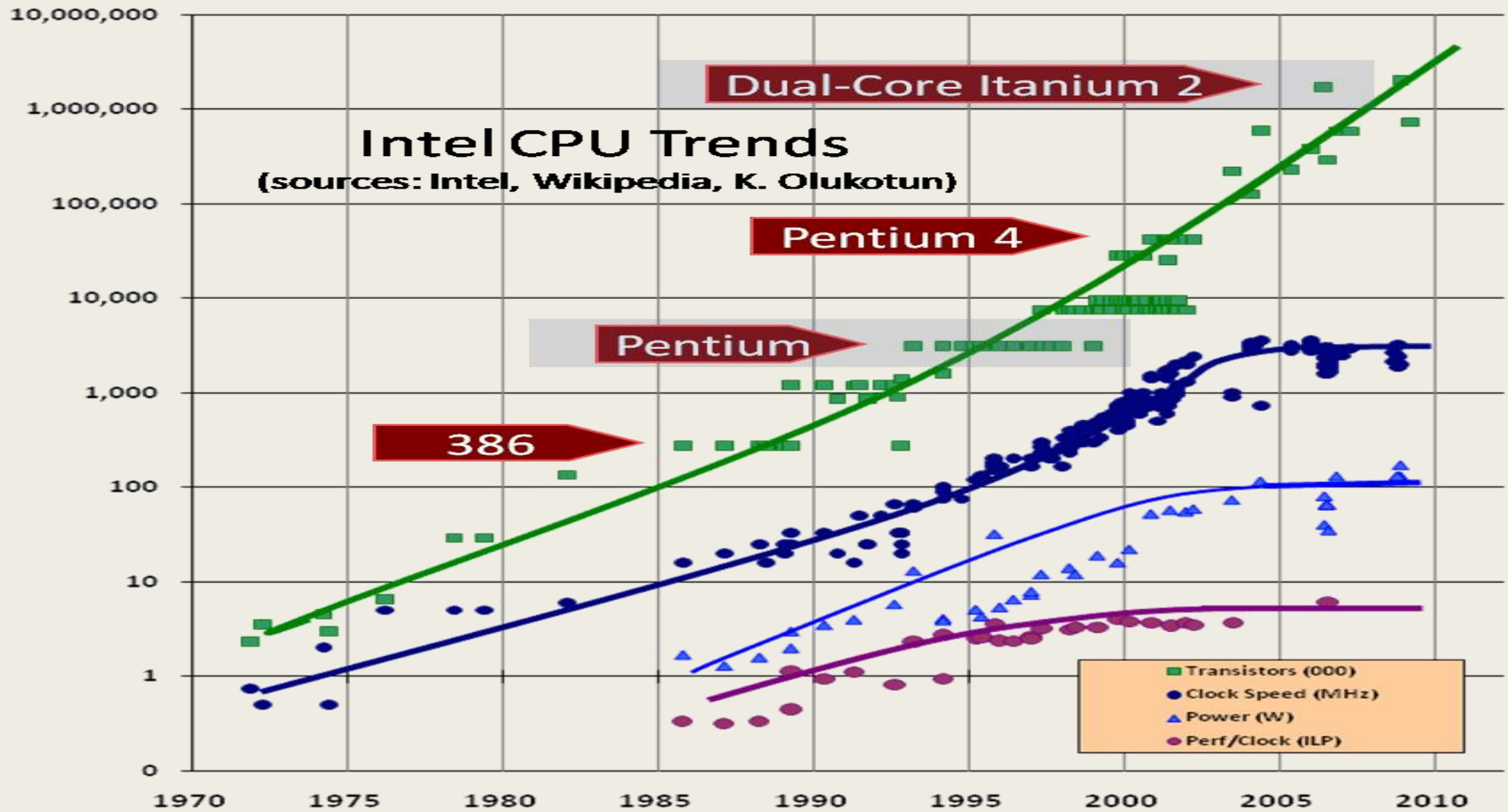


# Why Heterogeneous Computing is Needed?

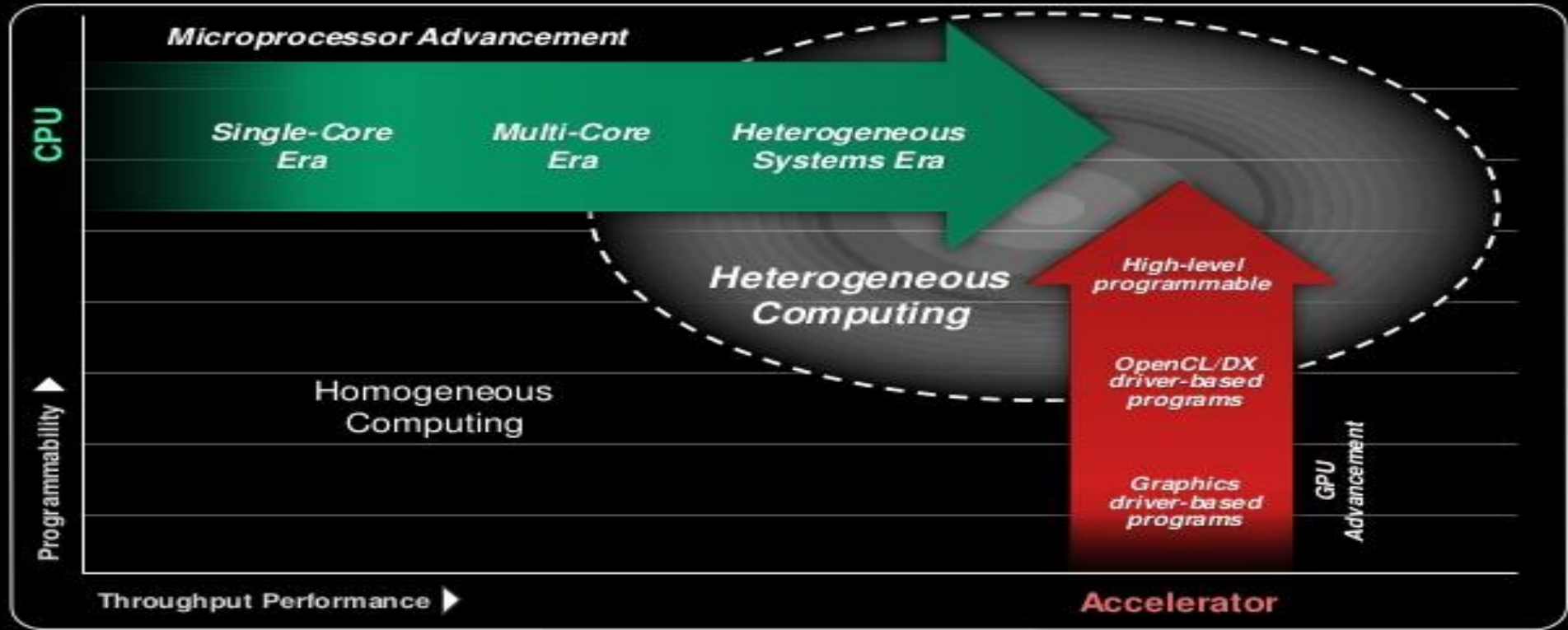
- There are some application's that cannot be fulfilled when using only one type of computing core.
- Typical Constraints:
  - *Performance / speed*
  - *Power Consumptions*
  - *Dimension / Size*



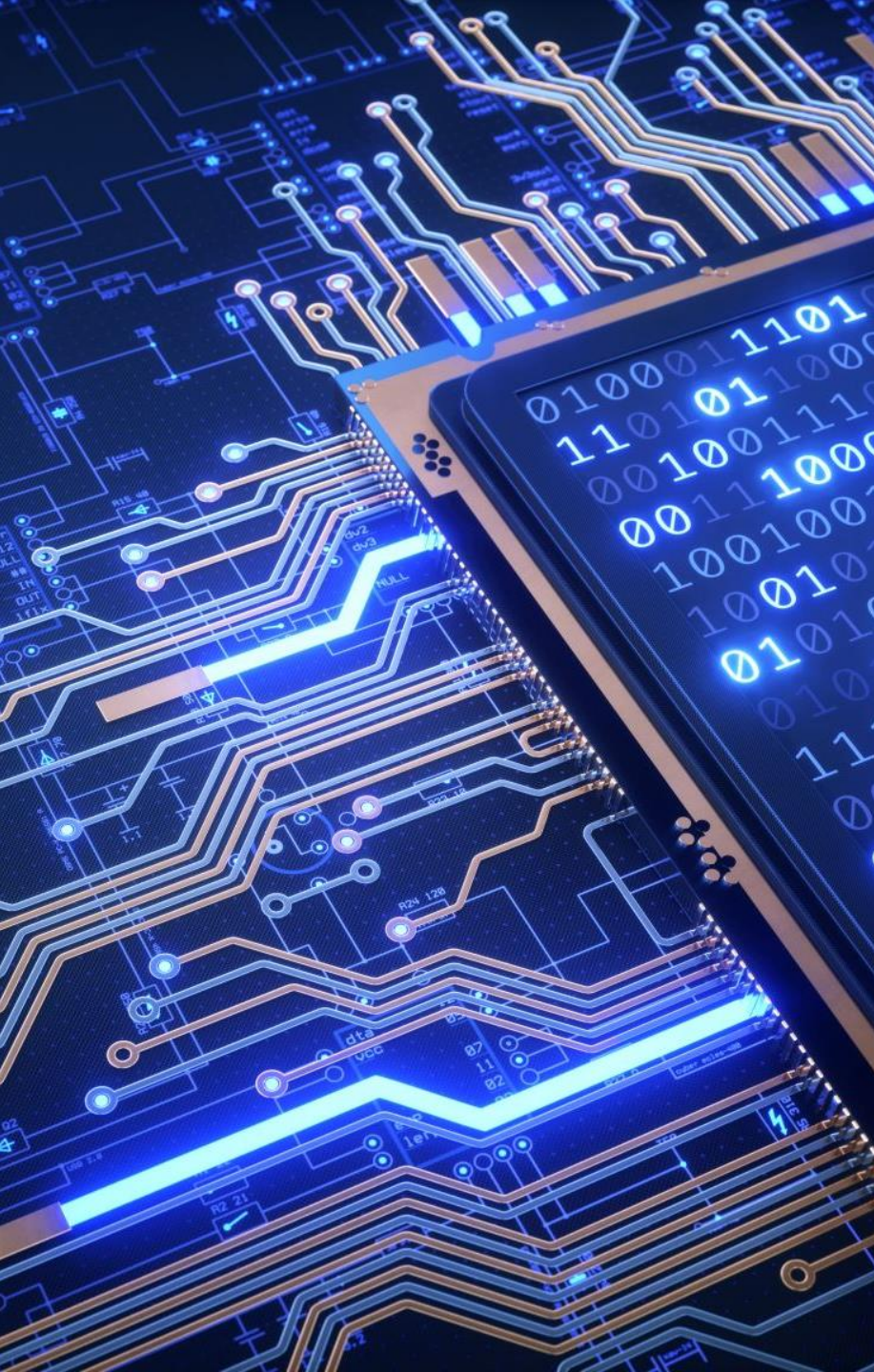
■ Growth in processor performance



# A PARADIGM SHIFT...







# Heterogeneous Computing

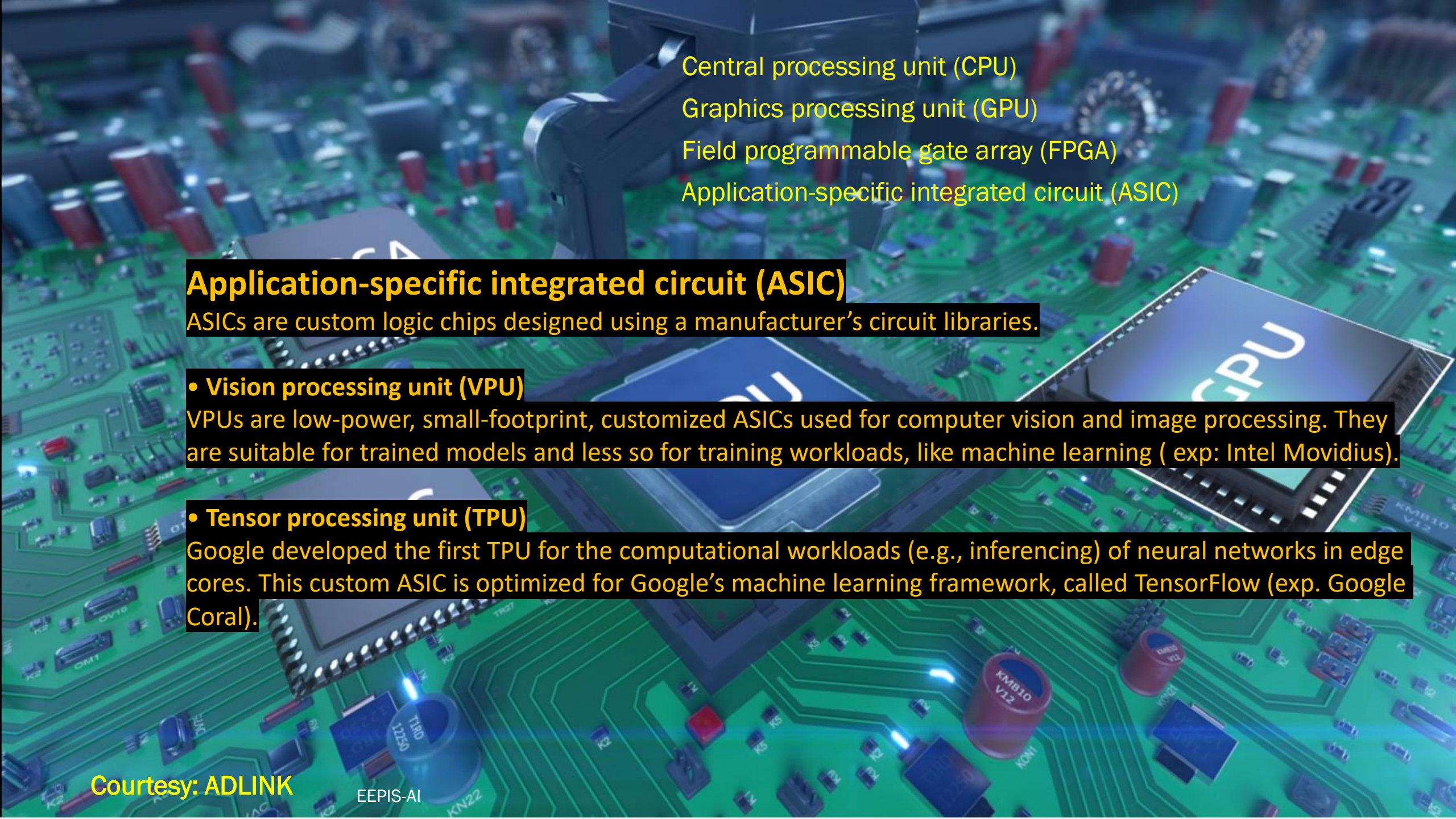
- *Systems that employ more than one different computing cores (architecture)*
- These are multi-core system that gain performance not just by adding cores, but also by incorporating specialized processing capabilities to handle particular tasks.
- *Diverse types of processors or hardware accelerators cooperate to accelerate the computational tasks*
- Potentially reduce the power consumption and maximize the computing performance



# HETEROGENEOUS COMPUTING:

Untuk mengoptimalkan setiap kelebihan pada teknologi hardware

PERFORMANCE dan EFISIENSI (Kecepatan /Daya)



Central processing unit (CPU)  
Graphics processing unit (GPU)  
Field programmable gate array (FPGA)  
Application-specific integrated circuit (ASIC)

## **Application-specific integrated circuit (ASIC)**

ASICs are custom logic chips designed using a manufacturer's circuit libraries.

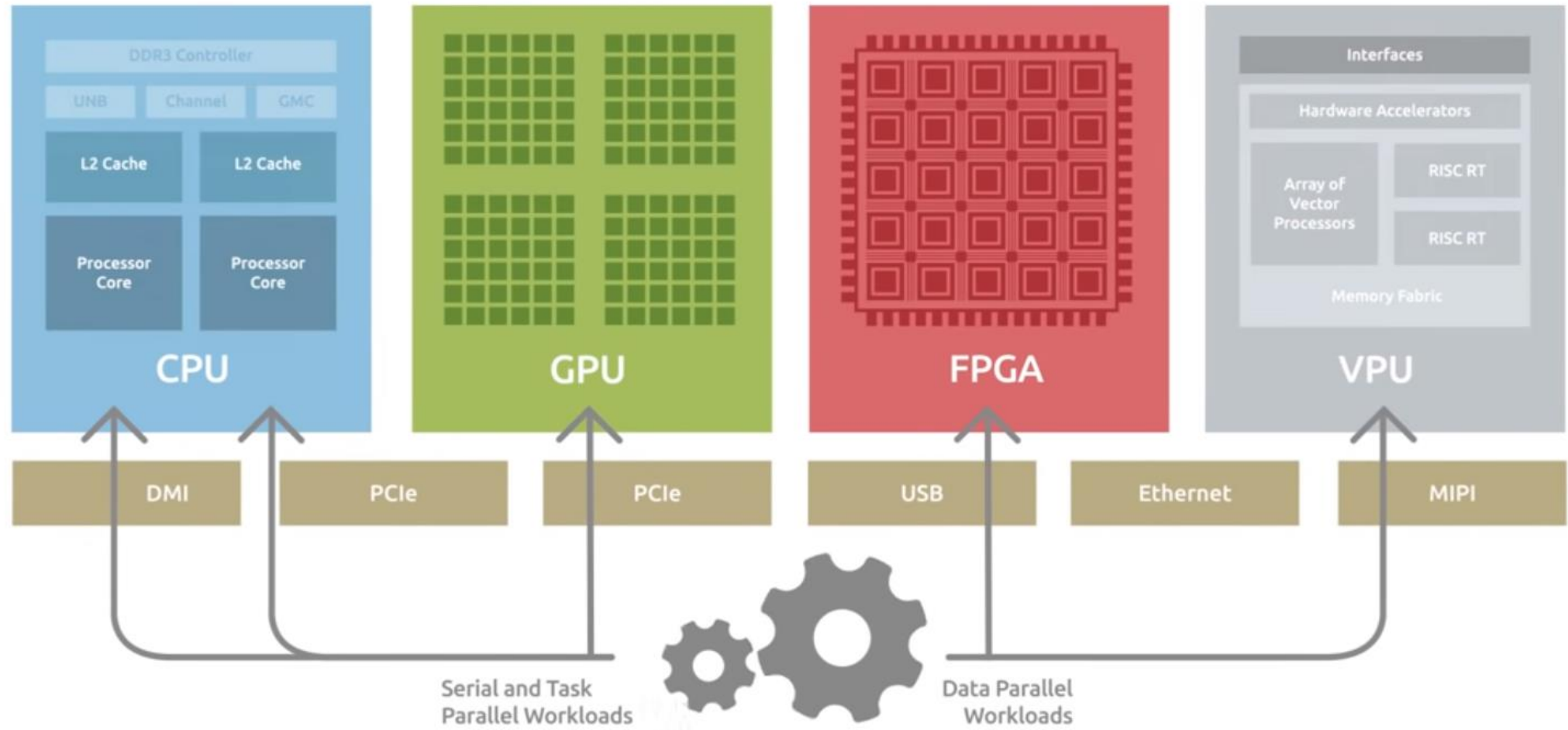
- **Vision processing unit (VPU)**

VPUs are low-power, small-footprint, customized ASICs used for computer vision and image processing. They are suitable for trained models and less so for training workloads, like machine learning (exp: Intel Movidius).

- **Tensor processing unit (TPU)**

Google developed the first TPU for the computational workloads (e.g., inferencing) of neural networks in edge cores. This custom ASIC is optimized for Google's machine learning framework, called TensorFlow (exp. Google Coral).

# Heterogeneous Computing



	CPU	GPU	FPGA
<b>Parallelism</b>	Limited by the number of cores	Supported by SIMT (single instruction multiple threads) approach	High parallelism with a customized design approach
<b>Power efficiency (performance / watt)</b>	Low	High	Very high
<b>Interfaces</b>	Support different interfaces	Limited or dependent on the interface with CPU	Customizable, including direct connection to cameras
<b>Development time</b>	Short	Medium	Long

ASIC		Low	Custom logic designed with libraries	<ul style="list-style-type: none"> <li>Fast and low power consumption</li> <li>Small footprint</li> </ul>	<ul style="list-style-type: none"> <li>Fixed function</li> <li>Expensive custom design</li> </ul>
	VPU	Ultra-low	Image and vision processor/co-processor	<ul style="list-style-type: none"> <li>Low power and small footprint</li> <li>Dedicated to image and vision acceleration</li> </ul>	<ul style="list-style-type: none"> <li>Limited dataset and batch size</li> <li>Limited network support</li> </ul>
	TPU	Low to medium	Custom ASIC developed by Google	<ul style="list-style-type: none"> <li>Specialized tool support</li> <li>Optimized for TensorFlow</li> </ul>	<ul style="list-style-type: none"> <li>Proprietary design</li> <li>Very limited framework support</li> </ul>

# CPU VS GPU VS FPGA VS ASIC

# CPU

- CPU, GPU, and FPGA hardware accelerators has different advantages compared to the others.
- A CPU is a general purpose processor that executes an instruction in a computer program, such as a computational operation, along with input/output operations.
- A CPU has various advantageous.
  - *First, it is ideal for complex scalar processing and very suitable for executing complicated operations on a single or a few streams of data.*
  - *Second, a CPU is able to accommodate its integration with various operating systems (OS).*
  - *Third, it also provides a well-known software development environment and I/O port access for sensors and devices (e.g., a camera, display, or network).*
- As a result, CPUs perform essential roles within computing systems.
- CPU's weakness → its parallel processing capabilities are limited by the number of processing cores



# Hardware accelerators for HCS

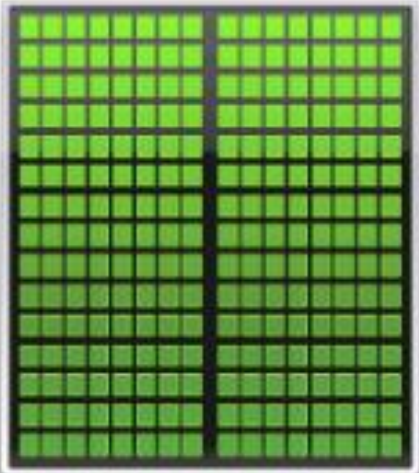
- In contrast with CPU, GPUs and FPGAs are specialized devices with highly parallel architectures.
- GPU (Graphics Processing Unit) and FPGAs (Field Programmable Gate Arrays) can enhance the computing performance for some algorithms.
- The GPU consists of hundreds or even thousands of small yet efficient cores, designed to handle multiple tasks (threads) simultaneously.
- Meanwhile, the FPGAs offers a parallel hardware structure that is re-programmable according to a specific user application.

# Graphics Processing Unit

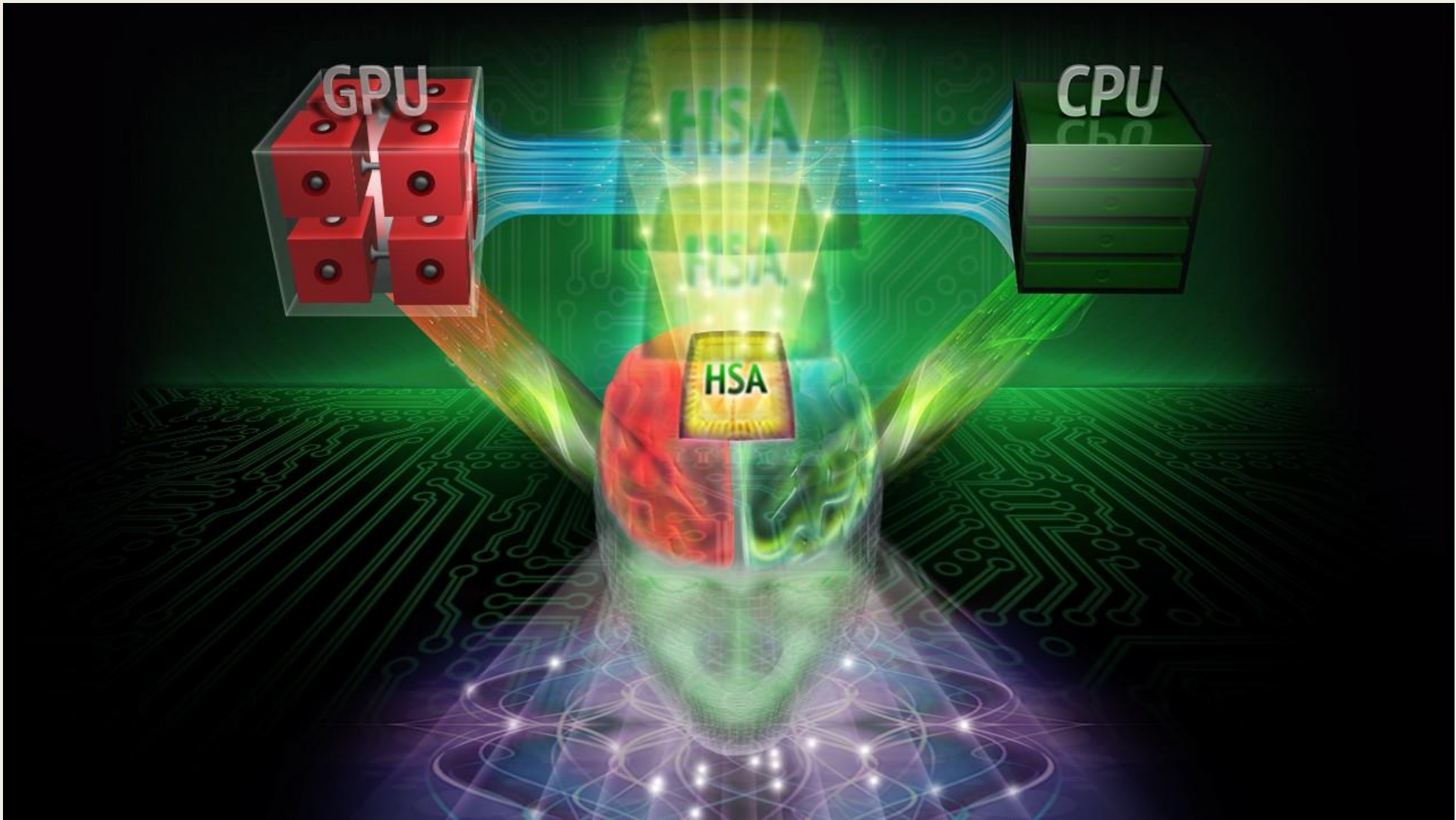
**CPU**  
MULTIPLE  
CORES



**GPU**  
THOUSANDS OF  
CORES







- Advanced drones and unmanned aerial vehicles (UAVs) that uses the power of deep learning algorithms to understand and react to the world around them. These autonomous machines are driving exciting new capabilities—from streamlining warehouses and inspecting lengths of hard to access power lines in real time to aiding in search and rescue operations in difficult terrain. NVIDIA® Jetson™ is the platform that makes it possible.





0:00:38

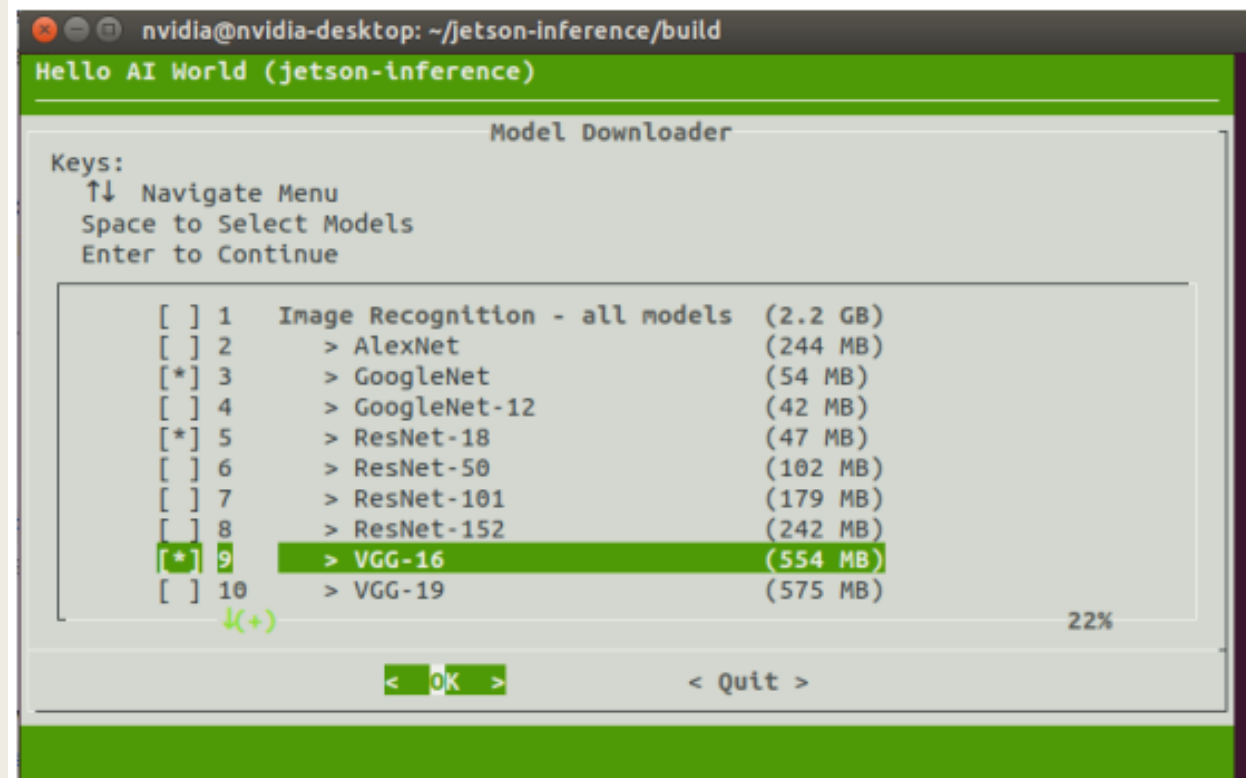
0:01:22



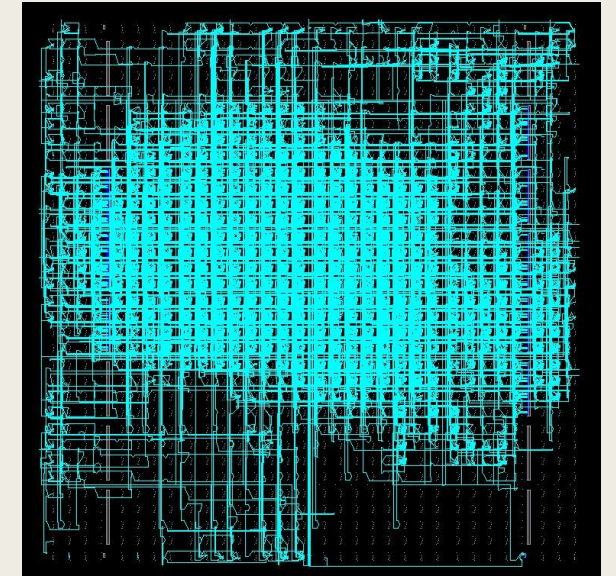
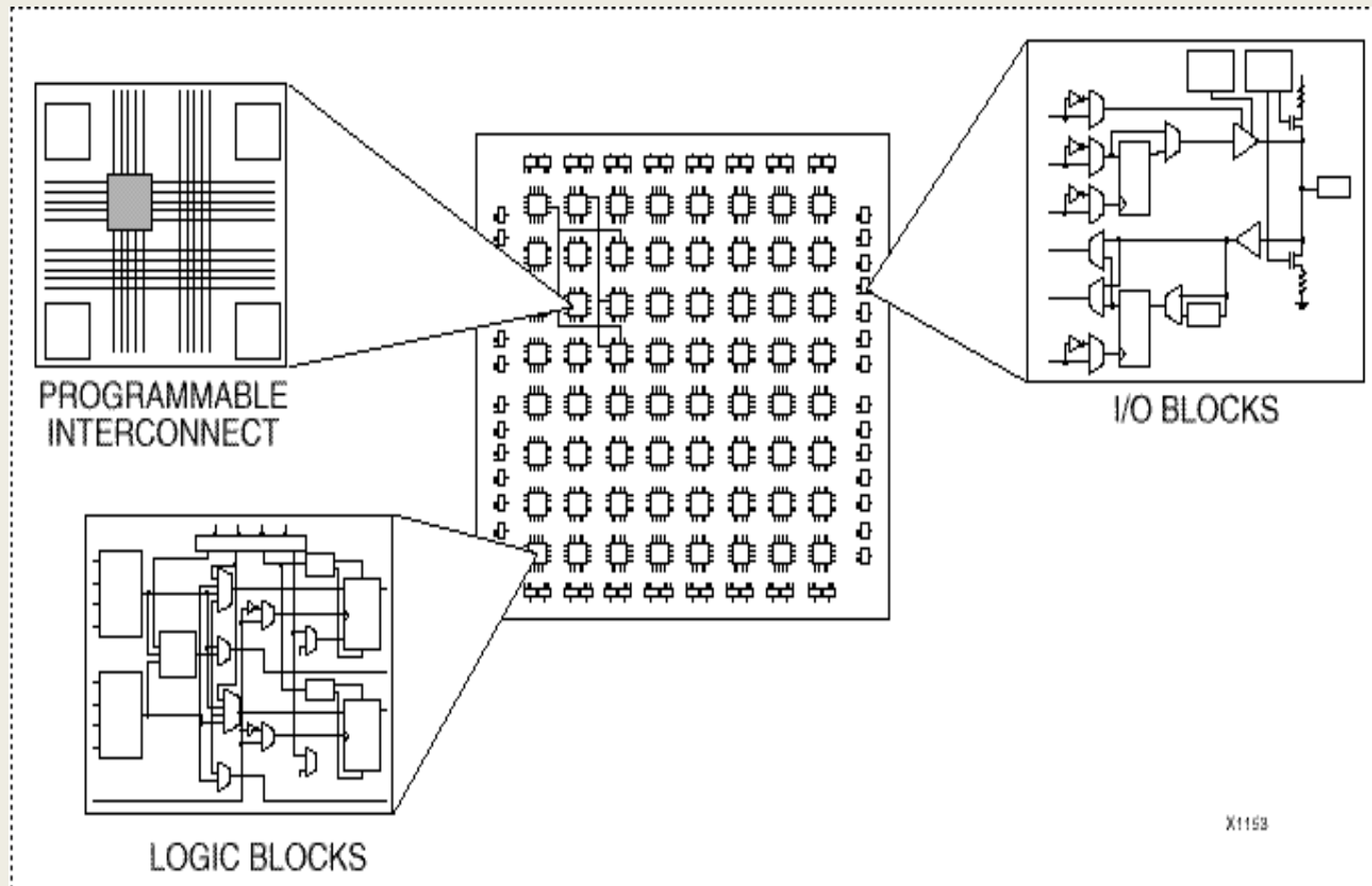
EEPIS-AI

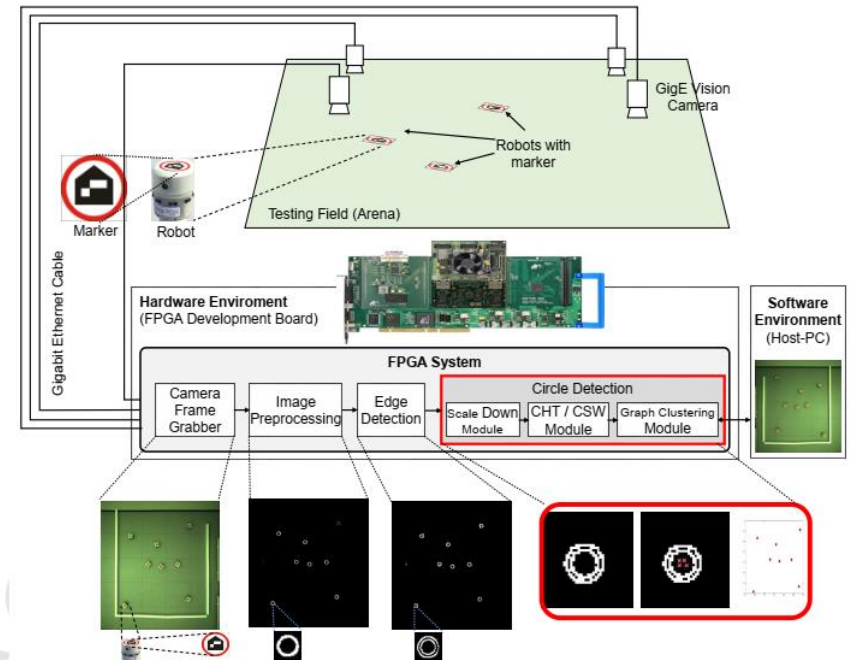
- This device uses computer vision and machine learning to help visually impaired people read, navigate, and recognize friends.





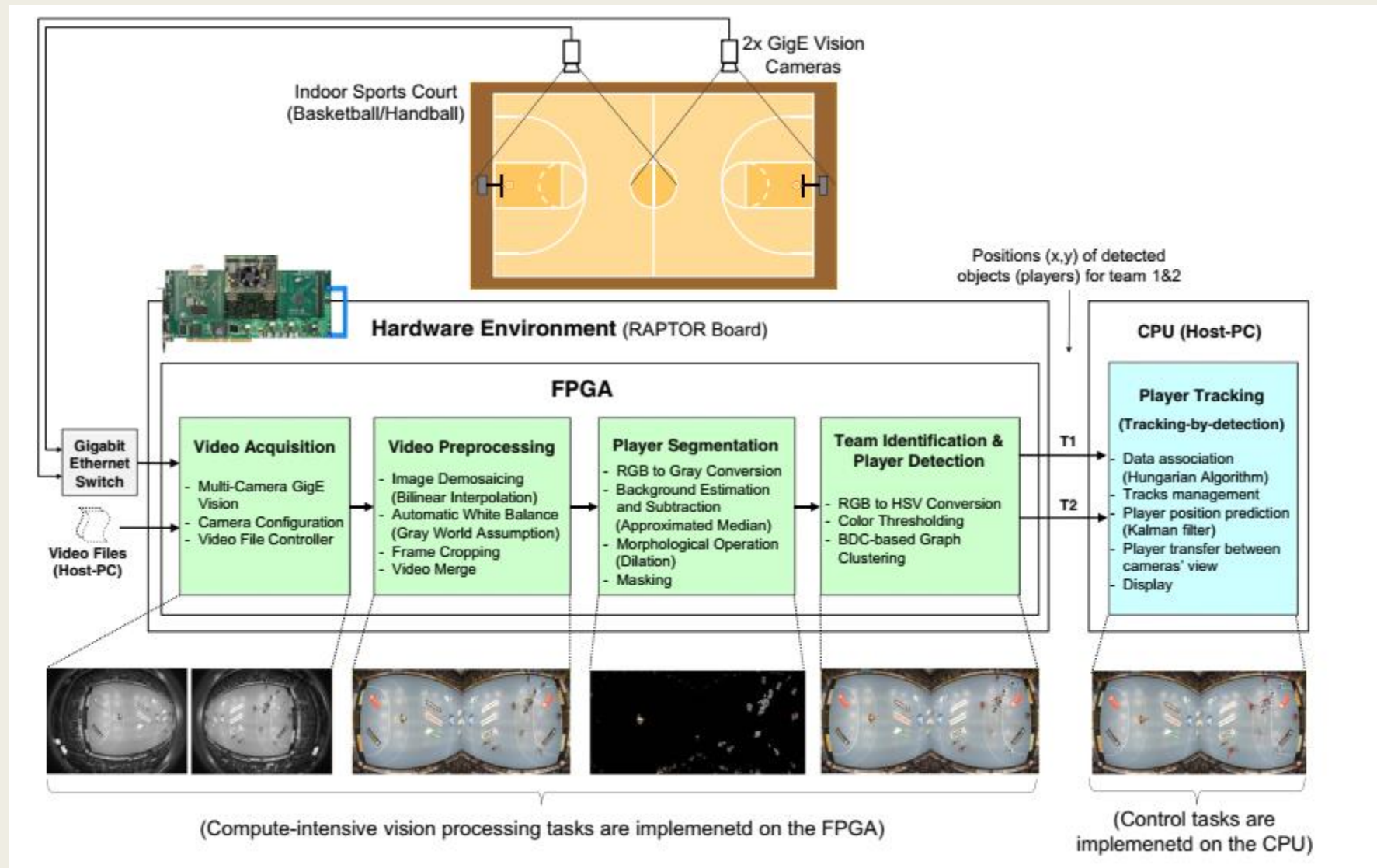
# Field Programmable Gate Array(FPGA)



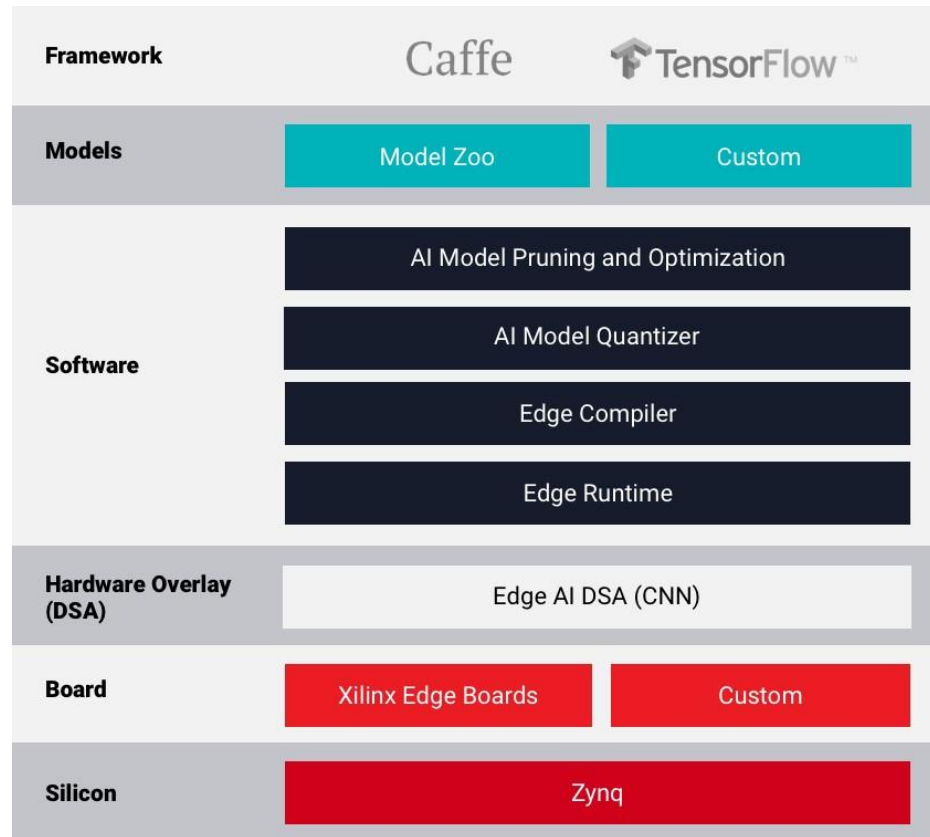


### FPGA-based multi-robot tracking

A Irwansyah, OW Ibraheem, J Hagemeyer, M Pormann... - Journal of Parallel and Distributed Computing, 2017



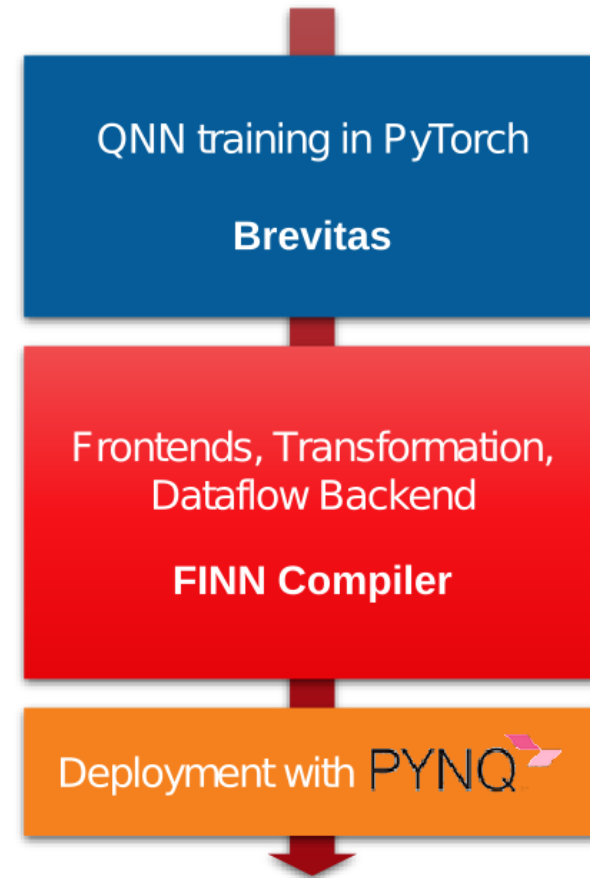


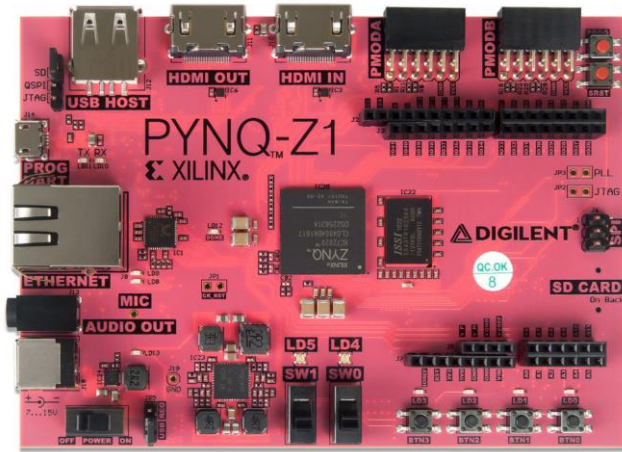


Customization of Algorithm

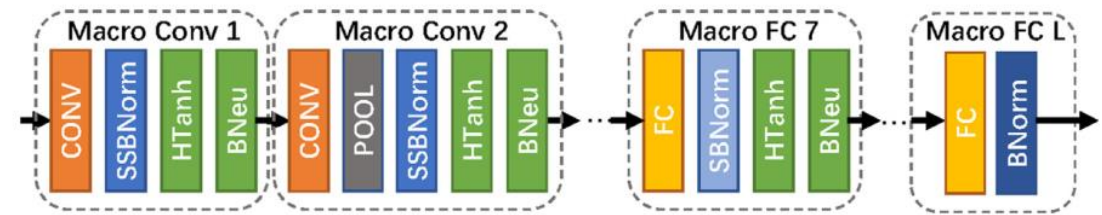


Customization of Hardware Architecture





*S. Liang et al./Neurocomputing 275 (2018) 1072–1086*



Shuang Liang, Shouyi Yin, Leibo Liu, Wayne Luk, Shaojun Wei,

FP-BNN: Binarized neural network on FPGA, Journal of Neurocomputing, Volume 275, 2018,

Pages 1072-1086, ISSN 0925-2312,

EEPIS-AI

# Heterogeneous Computing

- The collaboration between a CPU and some hardware accelerator (FPGA or GPU or ASICs) can increase the parallel computational capability of the computing system.
- These hybrid systems potentially reduce the power consumption and maximize the computing performance.

- Question ?