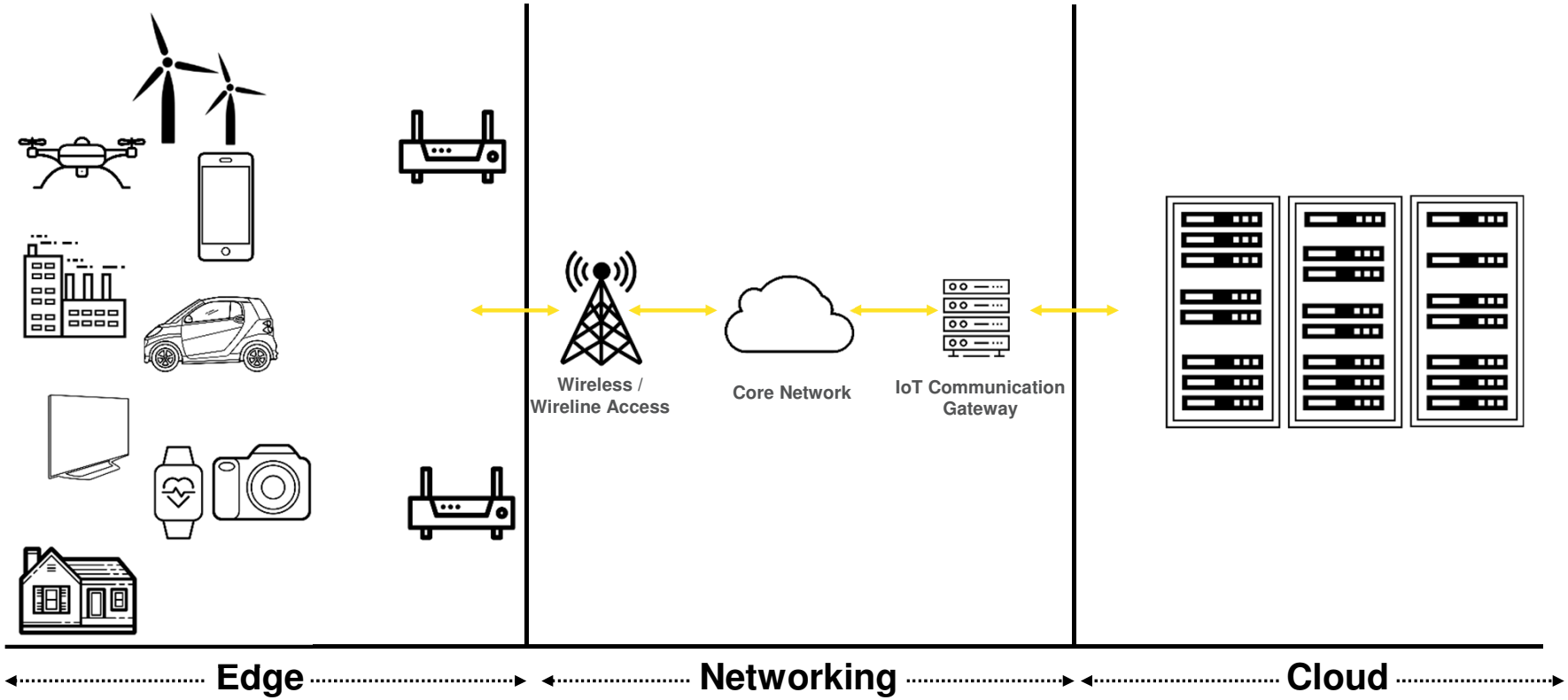
LATTICE sensAI™

Delivering Milliwatt AI to the Edge with Ultra-Low Power FPGAs

Rapidly Emerging Edge Computing Trend

Driven by Latency, Privacy, and Bandwidth Limitations



Unit growth for edge devices with AI will explode increasing over 110% CAGR over the next five years – Semico Research

Market Trends

- Most companies know AI has the power to change their business
  - But applying it effectively remains a challenge
- Many are starting to formalize their approach
  - AI Moving out of research groups and into product development
- Deployment of AI based products becoming a reality

Market Trends

- The dataset remains a significant challenge to adoption of AI:
  - Machine Learning for image recognition is only viable with a high quality set of training data
- Ecosystem developing for off-the-shelf solutions requiring no dataset
  - Pre-Trained for common applications
- “Synthetic” Data is becoming viable with computer generated data sets

LATTICE sensAI™

CUSTOM DESIGN SERVICES

Smart Home | Smart City | Smart Factory | Smart Car

REFERENCE DESIGNS / DEMOS

Face Detection | Hand Gesture Detection | Key Phrase Detection | Human Presence Detection | Face Tracking | Object Counting | Speed Sign Detection

SOFTWARE TOOLS

LATTICE RADIANT DESIGN SOFTWARE

Neural Network Compiler

Keras | TensorFlow | Caffe

LATTICE DIAMOND DESIGN SOFTWARE

IP CORES

CNN Compact Accelerator

Neural Network Accelerators

CNN Accelerator

HARDWARE PLATFORMS

UPduino Himax Shield – iCE40 UltraPlus FPGA

Embedded Vision Dev Kit – ECP5 FPGA

1 mW, 5.5 mm², 1/8/16 bits

1 W, 100 mm², 1/8/16 bits

HOT 100 EDN 2018

THE ELECTRONICS INDUSTRY AWARDS 2019

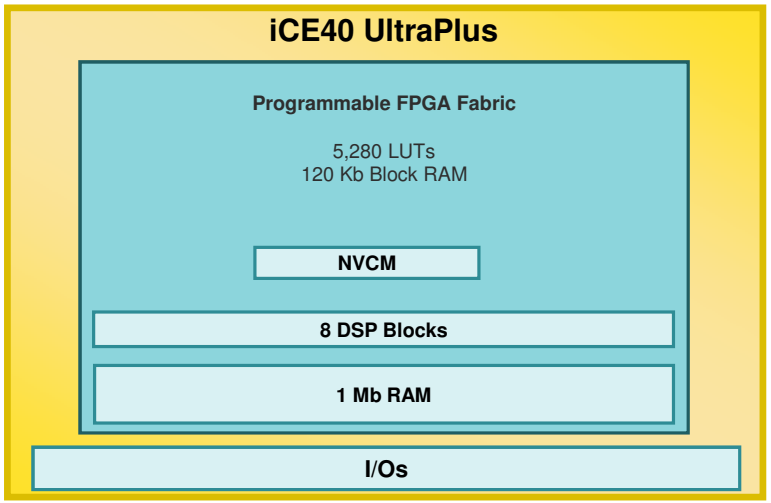
Award 2019 Assodel

AI BREAKTHROUGH AWARDS

- Ultra Low Power
- Small Form Factor
- Customizable

LATTICE SEMICONDUCTOR

iCE40 UltraPlus High Accuracy, Low Power Accelerator

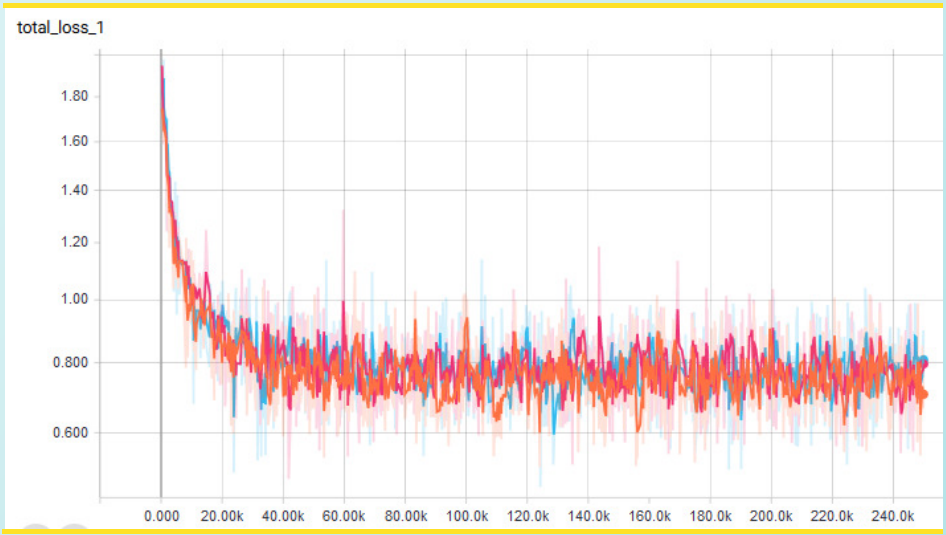


- Parallel computing capability
  - In device DSPs and 1Mbit SRAM
- Sensor agnostic flexible inferencing engine
- Single digit milli-watt power consumptions
- Lower latency
- Data pre-processing and result post post-processing in device

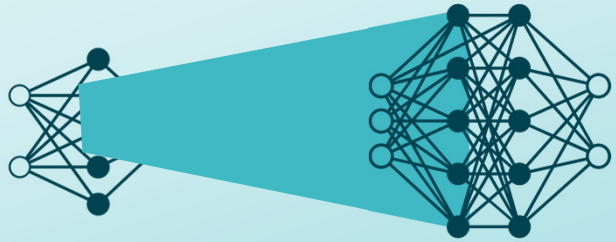
	Speed	Power	Resolution	Accuracy
Advanced MCU	1-2 FPS	50-70mW	64x64x3	Low
iCE40 UltraPlus	5-10 FPS	1-7mW	128x128x3	High

iCE40 UltraPlus FPGA: 8bit Deep Quantization Support

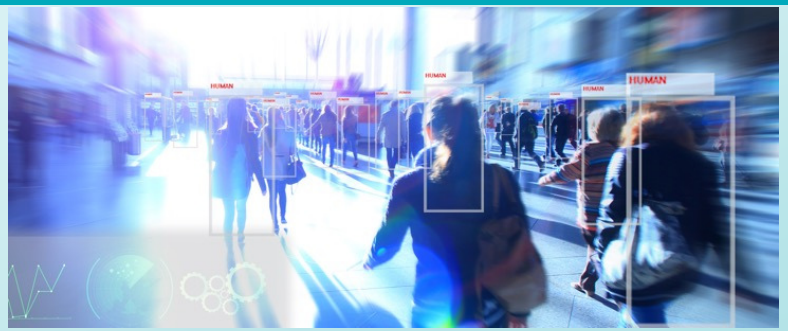
SUPPORT 8BIT QUANTIZATION



DOUBLE THE MODEL SIZE



HIGHER ACCURACY



ECP5 Enables High Speed AI Acceleration

Resolution	224x224x3
Network	VGG
Speed	6 frames per second

Previous Release

Resolution	224x224x3
Network	MobileNet v1, Resnet
Speed	17 frames per second

New Release

Focus Applications

Object Detection

Defect detection in smart security and embedded vision cameras

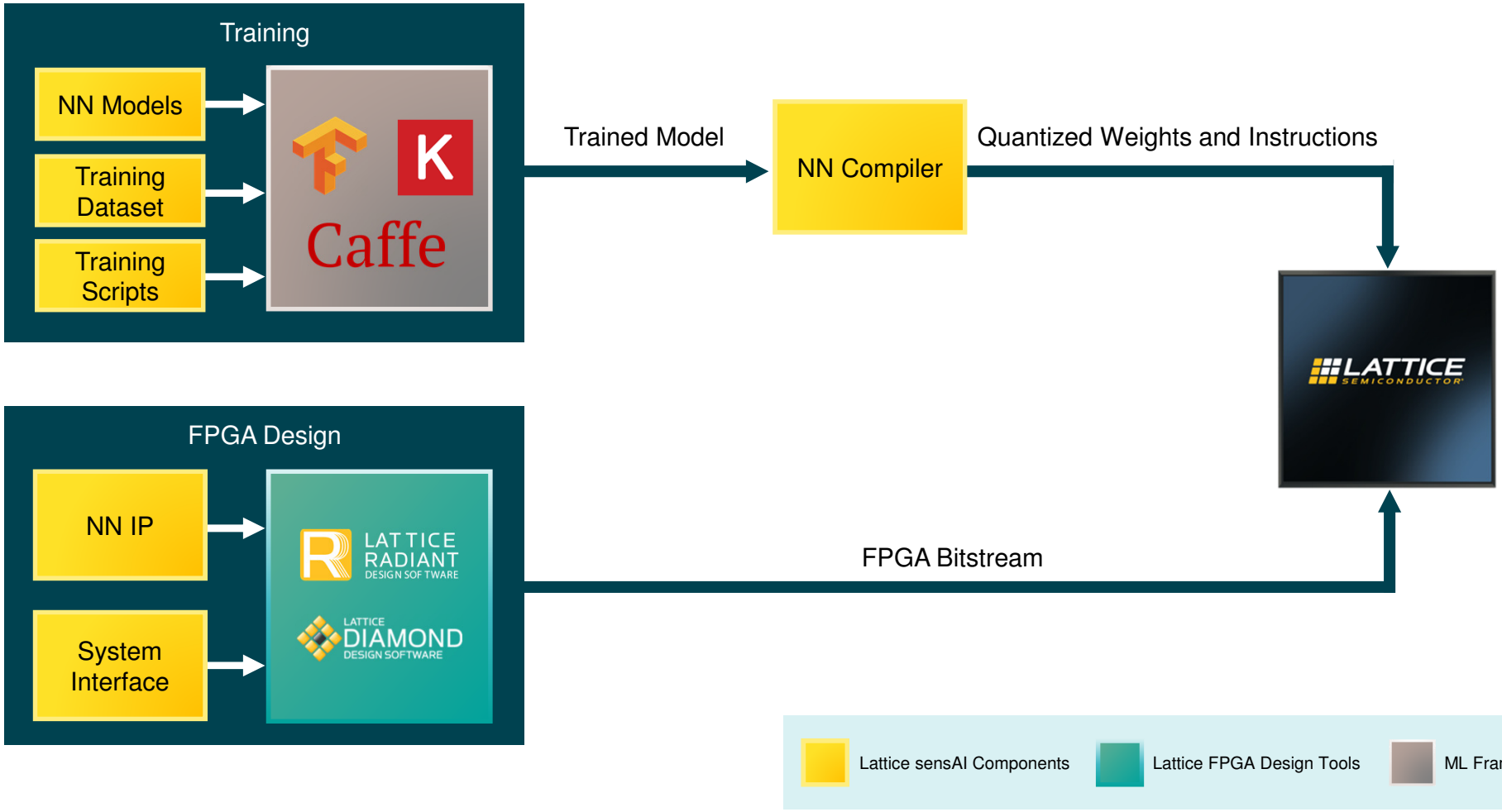
Human Machine Interface (HMI)

Key Phrase detection to control smart appliances

Object Identification

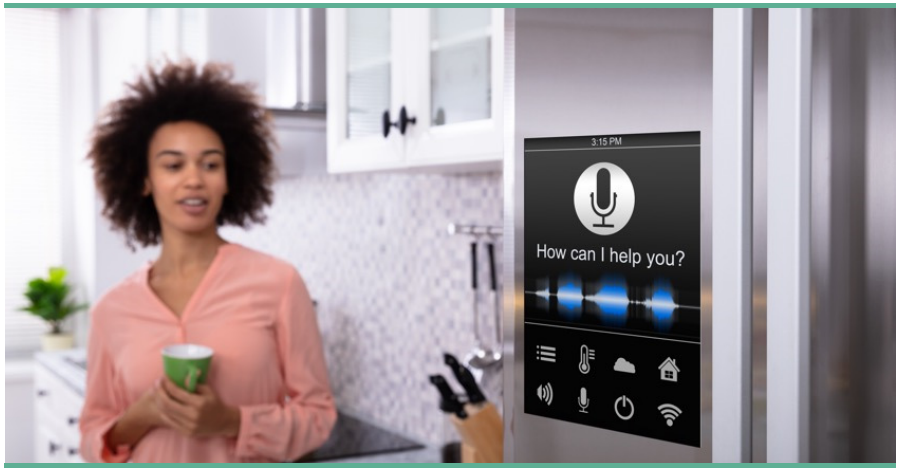
Feature extraction enabling navigation of robots

Customizable Reference Designs



Reference Design / Demo – Key Phrase Detection

FEATURES	
Sensor	Microphones
Network	VGG8
Speed	40 Evaluations per Second
Power	7 mW on iCE40 UltraPlus

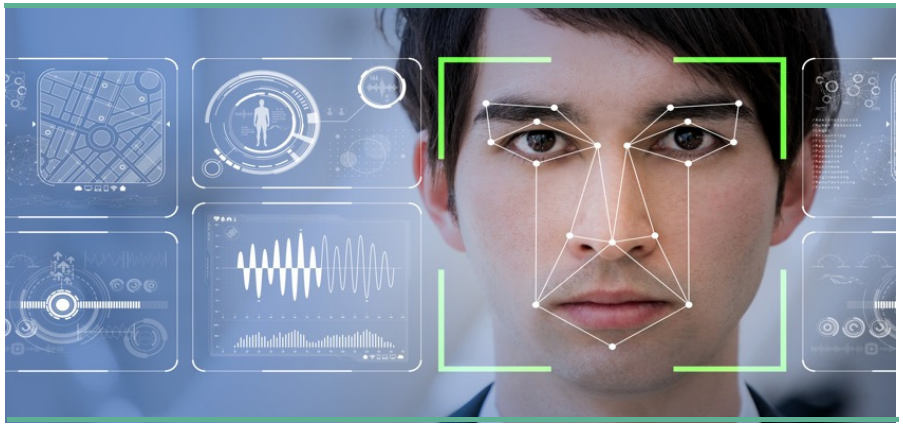


SMART APPLIANCE HMI VIA VOICE




Reference Design / Demo - Human Face Identification

FEATURES	
Sensor	CMOS image sensor
Speed	2 frames per second
Power	850 mW on ECP5-85K



HUMAN IDENTIFICATION IN VIDEO SECURITY DEVICES

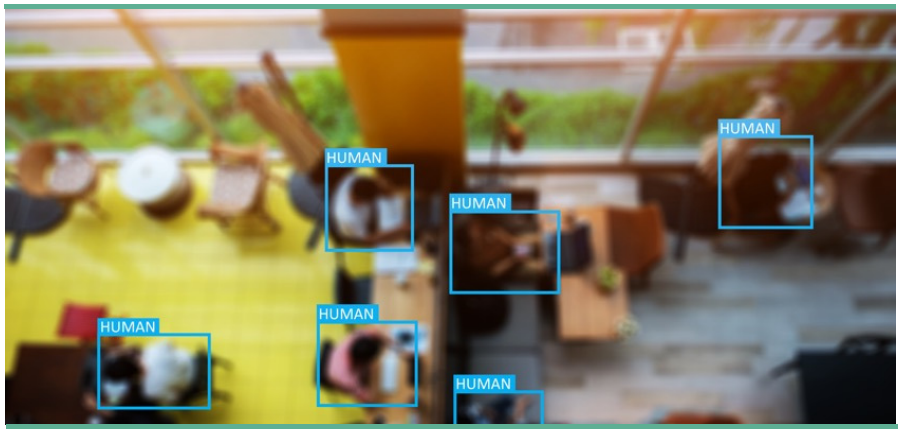
USER IDENTIFICATION IN SMART TOYS

SLAM FOR CLEANING ROBOTS

IN SYSTEM OBJECT REGISTRATION WITHOUT RETRAINING

Reference Design / Demo -– Human Presence Detection

FEATURES	
Sensor	CMOS image sensor
Speed	5 frames per second
Power	7 mW on iCE40 UltraPlus

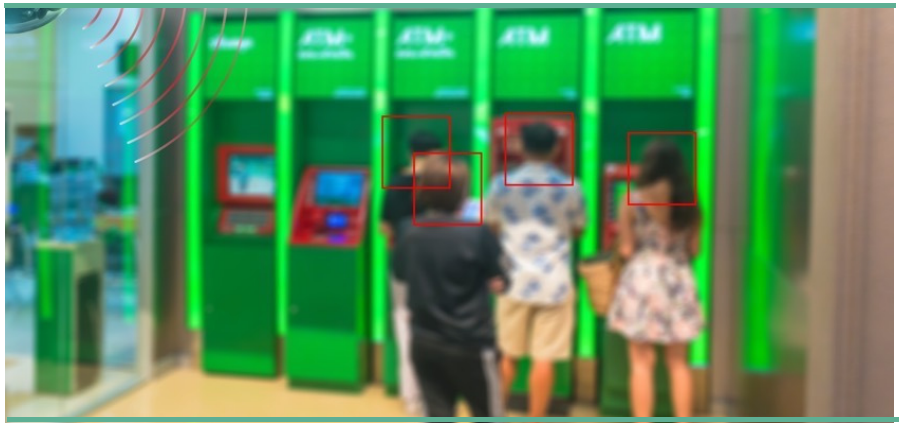


ALWAYS ON HUMAN DETECTION IN APPLIANCE

LOW POWER HUMAN DETECTION FOR WAKE ON APPROACH FOR LAPTOPS AND PRINTERS

Reference Design / Demo Object Counting

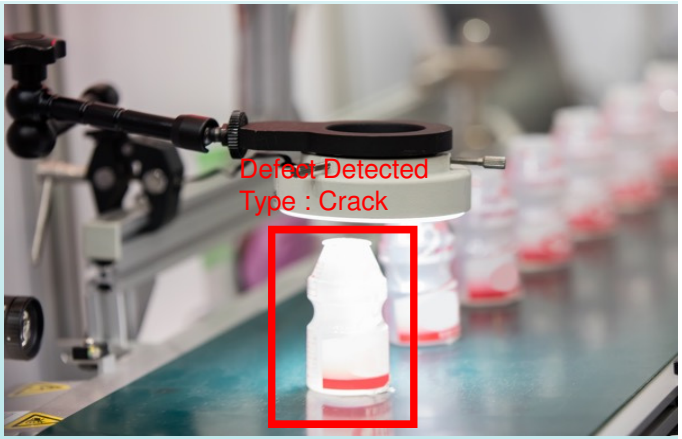
FEATURES	
Sensor	CMOS image sensor
Speed	17 frames per second - Lower Latency
Power	850 mW on ECP5-85K



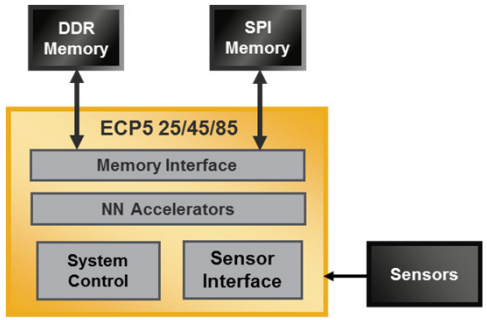
HUMAN DETECTION IN VIDEO SECURITY DEVICES

HUMAN COUNTING IN RETAIL CAMERA APPLICATIONS

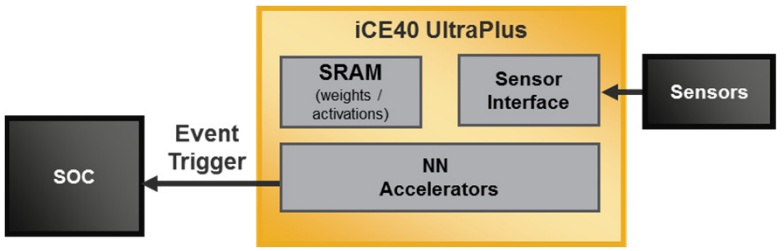
DEFECT DETECTION AND OPERATOR COMPLIANCE IN SMART FACTORY CAMERAS



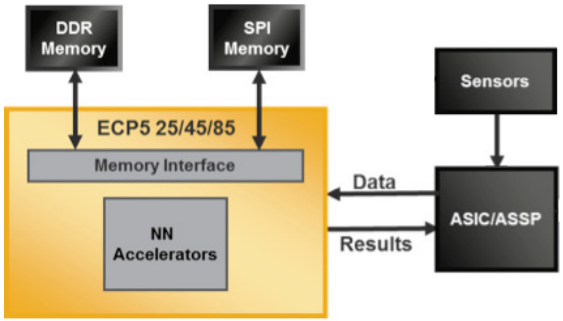
Popular sensAI Accelerator Use Cases



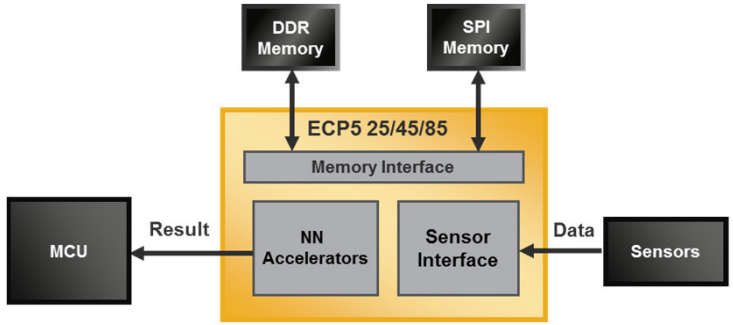
Stand-alone



Preprocessing



Post Processing

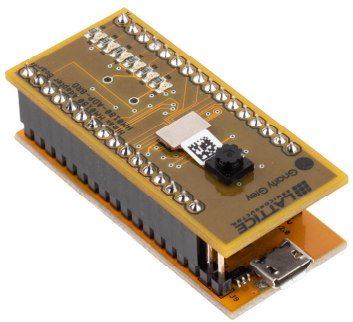


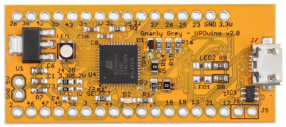
Preprocessing

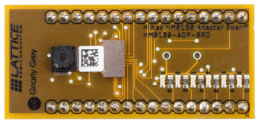
Hardware Platforms

Modular Platforms for Rapid Prototyping

HM01B0 UPduino Shield Board



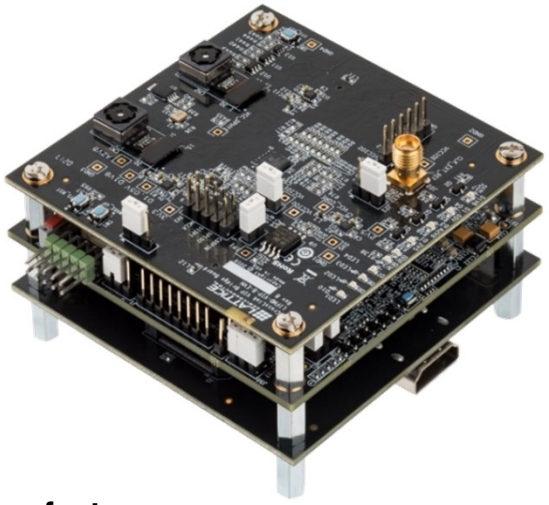




Key features

- Video and Audio sensors
- Compact 22 x 50 mm
- Includes HM01B0 image sensor board
- Arduino Micro form factor UltraPlus board

Embedded Vision Development Kit



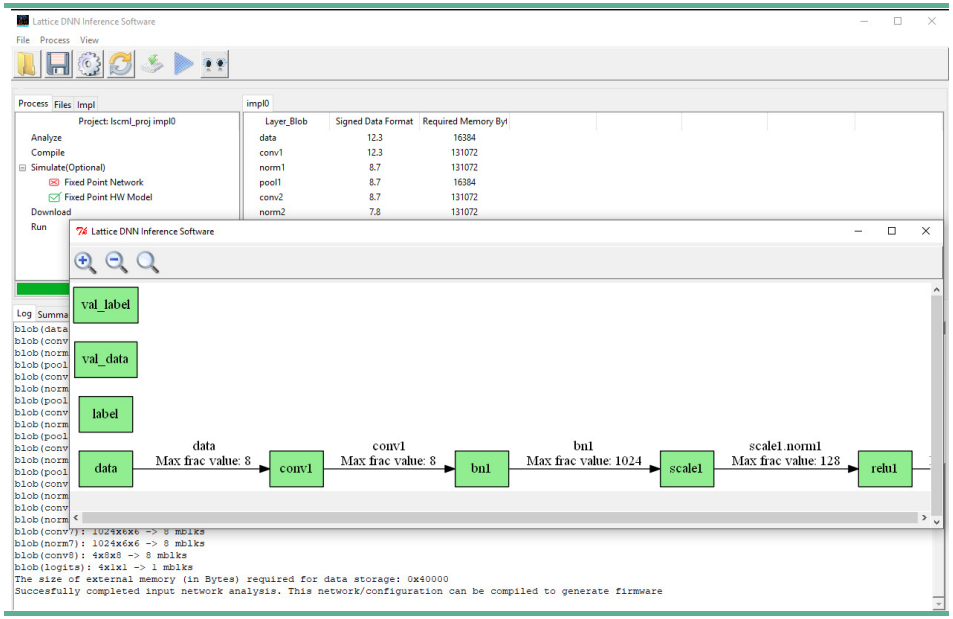
China Electronics Market
2017 Editor Choice Award

Key features

- ECP5 FPGA consuming under 1 W of power consumption
- Flexible video connectivity with support for MIPI CSI-2, eDP, HDMI, GigE Vision, USB 3.0, and more

Software Tools
Neural Network Compiler

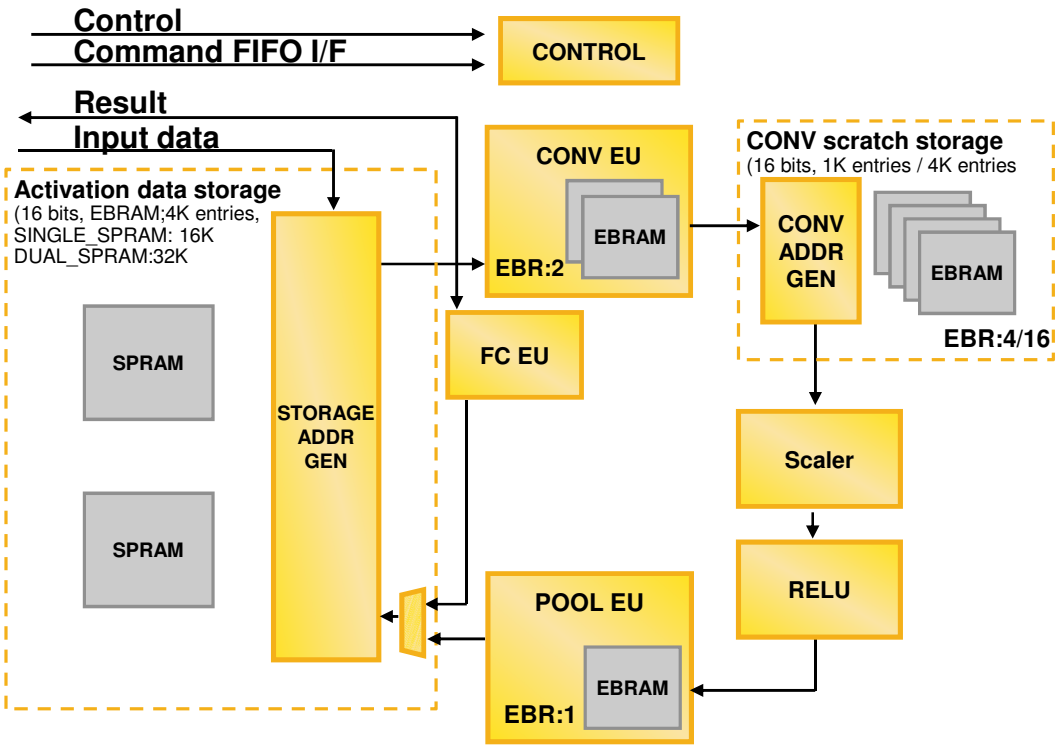
Caffe TensorFlow Keras



Key features

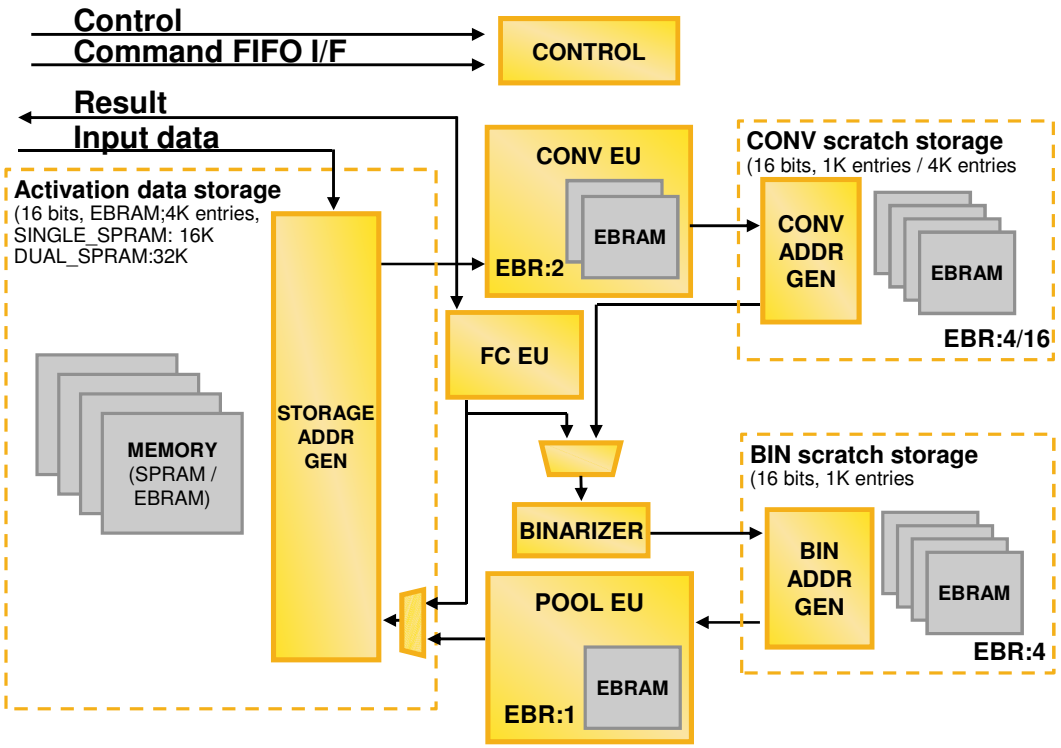
- Implement networks developed using standard frameworks into Lattice FPGAs without prior RTL experience
- Rapidly analyze, simulate, and compile CNNs/BNNs for implementation on Lattice sensAI IP cores

Engine Structure



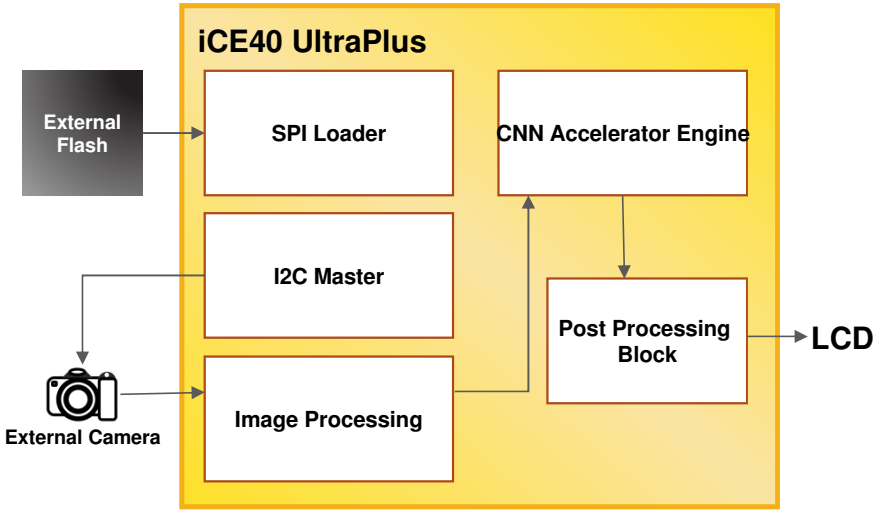
- **Hand crafted and pre-designed, not HLS based**
- **HW engines compute ALL NN functions of one layer**
  - No CPU involvement in NN computation
  - All layers have the corresponding HW engines

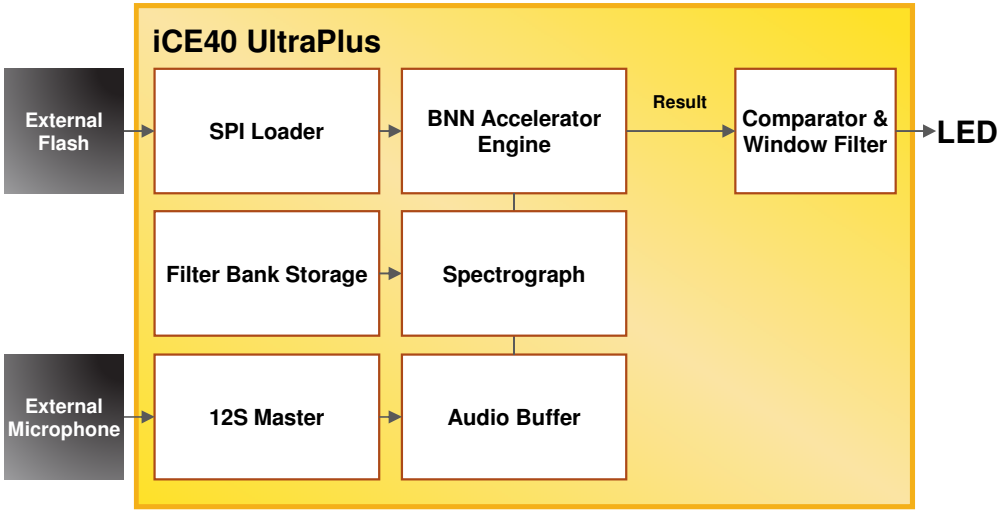
Engine Structure



- **Multiple engines** for various different network topologies
  - Reprogram different engines per network
- **Focus** on HW efficiency and exploit re-programmability

Solution HW structure

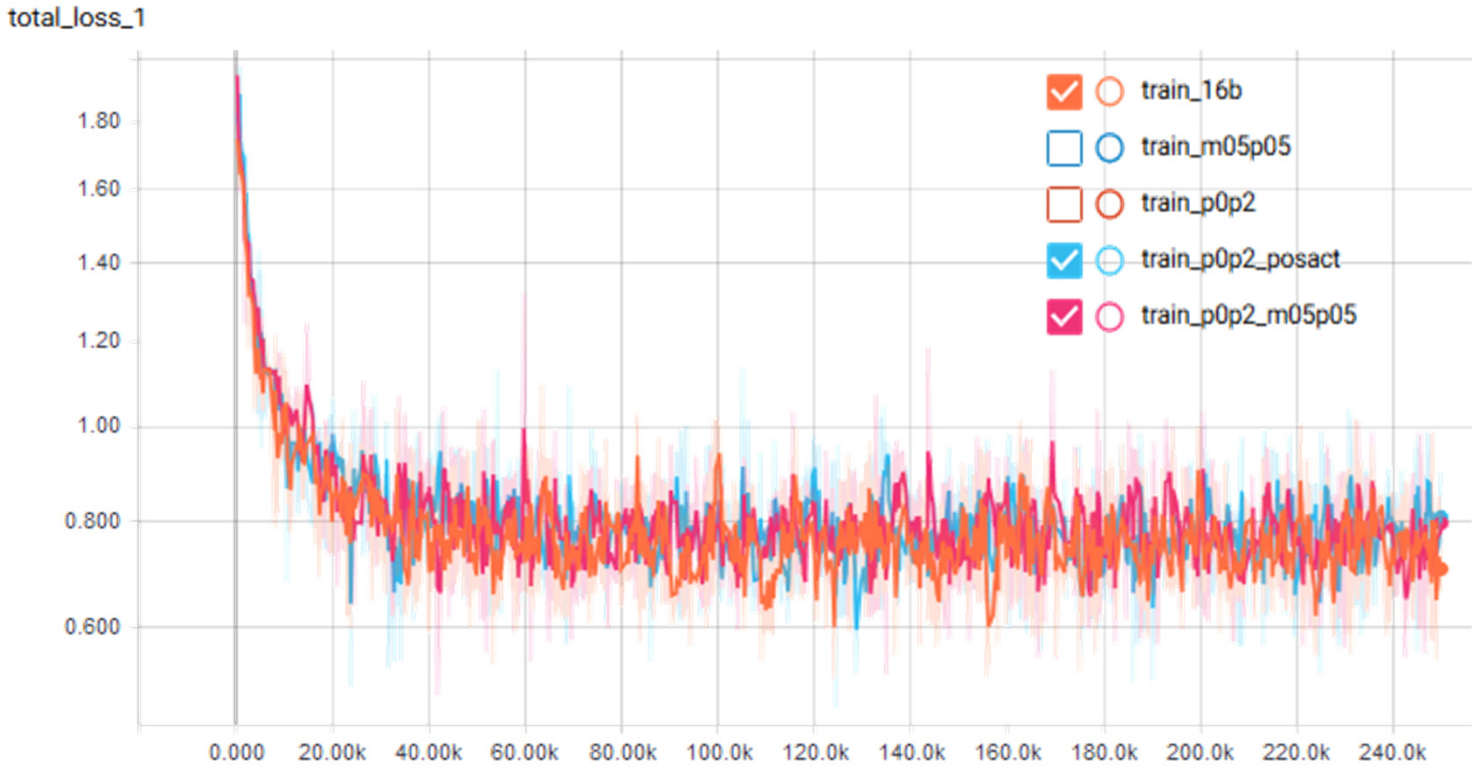




FPGA runs not only ML engine but also all the pre/post processors

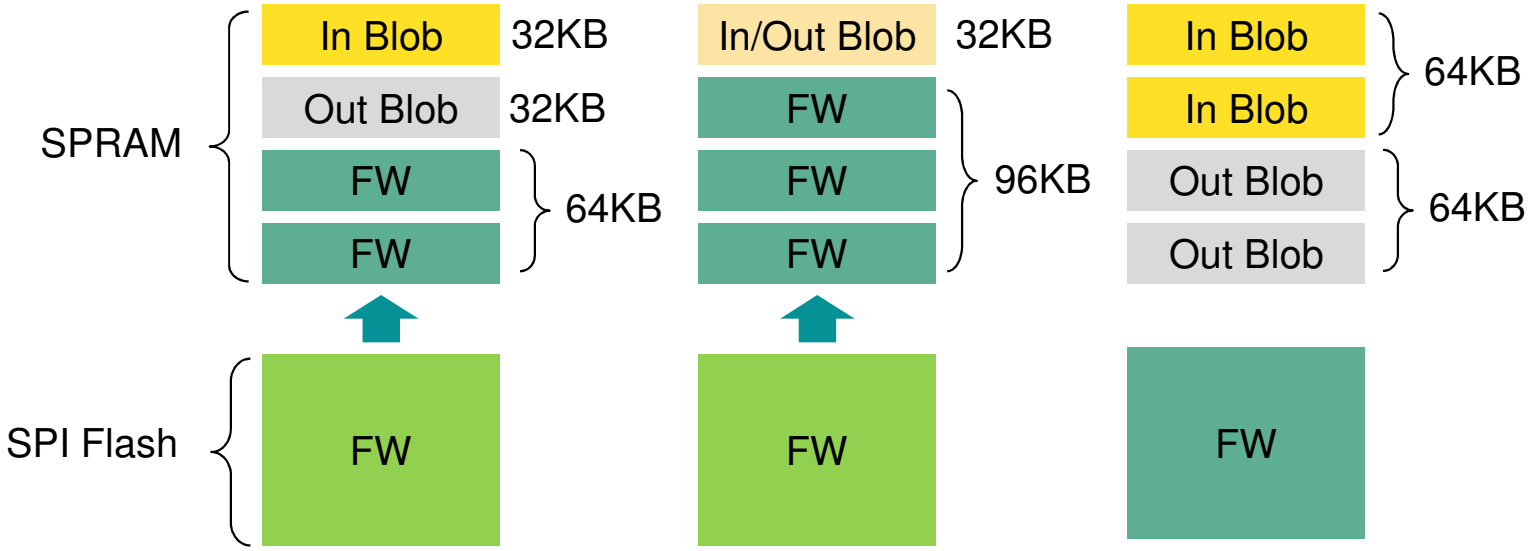
- Camera control, image processing (e.g., ISP, down scaler), post processing part
- MIC control, I2S master, audio data buffer, spectrograph (timed FFT), output time filter

Optimization - Quantization



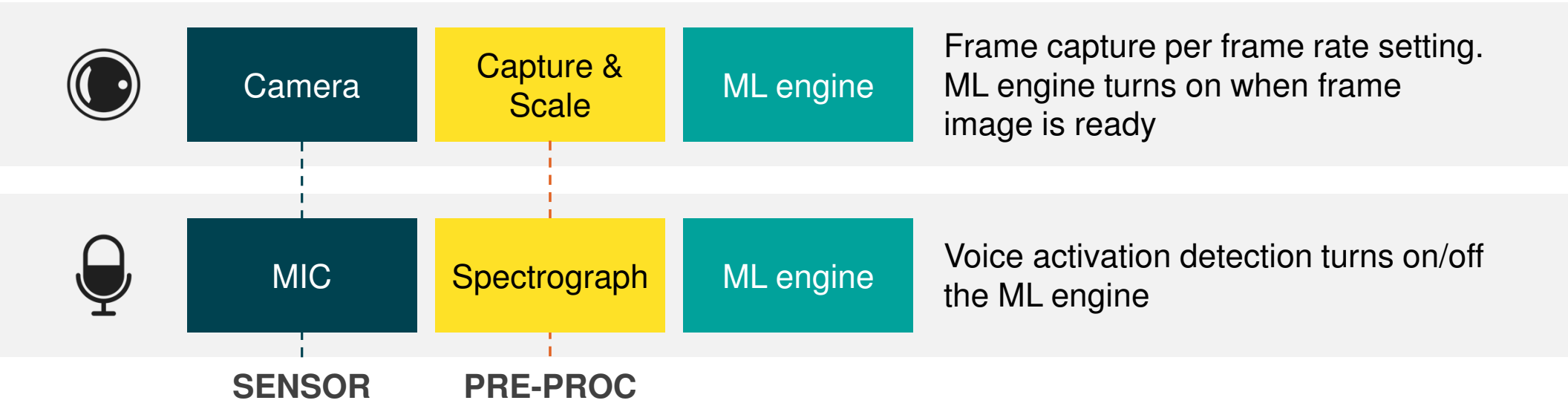
- "Quantization during training" instead of "Quantization after training"
  - Put the quantization layer in the training and let neurons know that they are 8b instead of floating point. Neurons/weights will evolve to find out the best values (8b values) that minimize the error in training process
  - Extendible to deeper quantization (4b, 2b, etc.)

Optimization – Memory Assignment



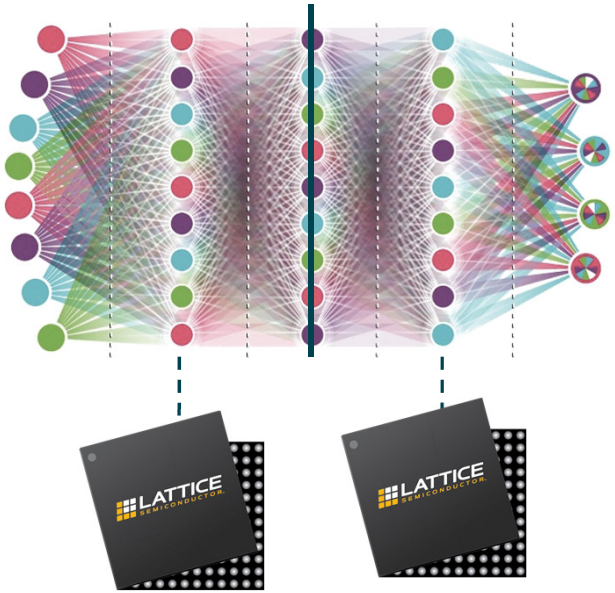
- Different memory assignments for different blob sizes and FW (weight) sizes
  - Choose different engines per the network requirements (blob size, weight size) and power constraint

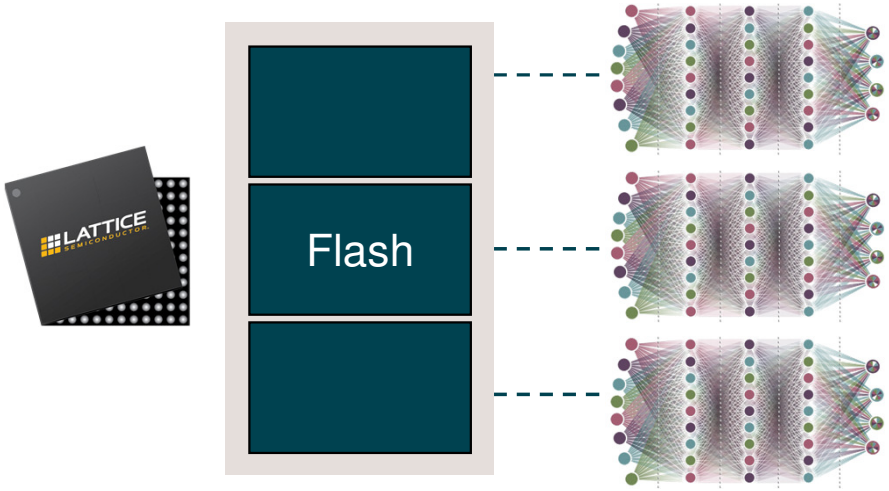
Optimization – Power Optimization



- **Minimize the activation of ML engine**
  - Clock gating of ML engine when preprocessor collecting data to process
  - Run engine as fast as possible and turn off clock and/or go to low power mode

Optimization – Multiple FPGAs & Chaining of multiple networks





Network is partitioned and mapped into multiple FPGAs for better throughput

- Blob value is transferred

Multiple networks are stored in a Flash and run in serial

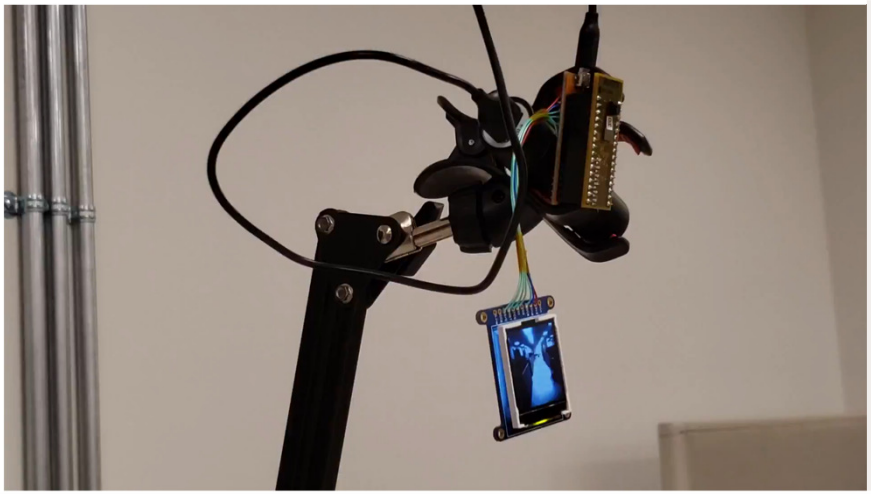
- Output of each network is aggregated or used for the next network invoking

Network Design for Edge Applications

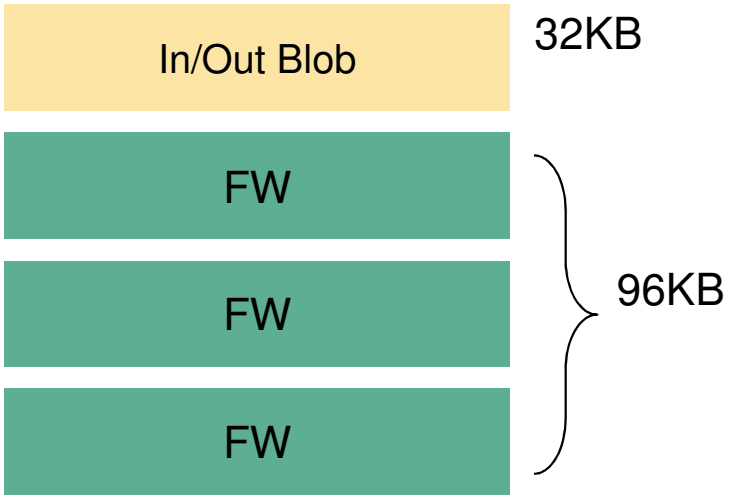
- **Don’t try to run reference models** in the web site as is
  - Hundred of layers is not needed/suitable for low power edge applications
- **Most of applications can be covered** by 8~15 CONV layers
  - Not much benefit from residual net/dense net
- **Mostly VGG type and MobileNet type**
- **Optimization process**
  - Start from a known reference network with a given training set
  - Optimize network (reducing depth and width) with monitoring accuracy
  - Dataset clean up
  - Small network is more sensitive to the quality of training set
  - Augmentation to reflect the sensor characteristics
  - Add quantization in training

Object Detection – Human Presence Detection

- 64*64*3 input
  - 6 zone searching to cover 128*128*3
- VGG8 like – 8*(Conv, BatchNorm) + 4*Pooling
- 10FPS; 6~7mW@5FPS



Green box in LCD and LEDs on the board indicates the detection of human



Lattice ECP5 FPGA vs SOC and ASICs

- ECP5 has more flexible I/Os and Interfaces
- ECP5 can reconfigure itself from one application to the other
- ECP5 can support changing ML topologies
- SOC:
  - Has more horsepower but consumers more power
- ASICs:
  - Lack the flexibility in topology selection and modification

Lattice iCE40 UltraPlus vs MCUs

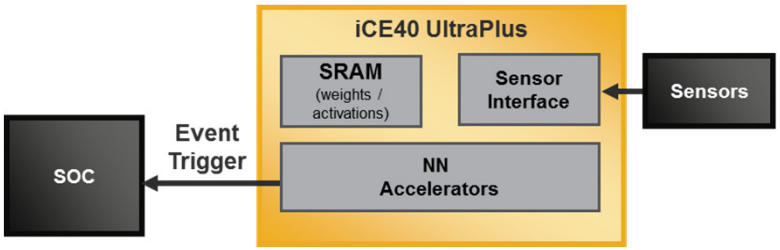
- MCUs generally suffer from performance
  - Need ARM Cortex M7 class processors to do image based NN acceleration with good performance
- 10X higher power consumption for most applications
  - MCU runs at higher clock frequency ~200 – 500 MHz
- MCU has higher latency ~500ms
- iCE provider higher efficiency in preprocessing and post processing
- Designers are comfortable with MCU environment
- Acceleration – use lower class MCU + FPGA (co-exist)

Lowest Power, Performance Optimized

Always-on human presence detection

- 128x128x3 (RGB)
- 5 frames/sec

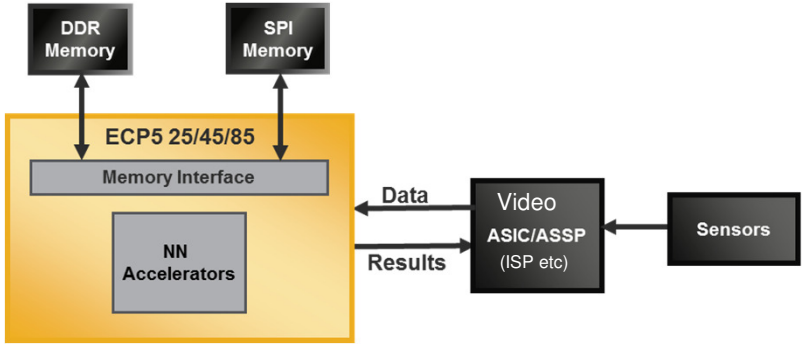
	Performance	Power	Cost
MCU	~1-2 FPS	~100mW	$
Lattice iCE40 UltraPlus	~5 FPS	~7 mW	$



Always-on human counting

- 1080p downscaled to 224x224x3 (RGB)
- 17 frames/sec

	Performance	Power	Cost
Vision AI SoC	30+ FPS	~2W	$$
Lattice ECP5	17 FPS	~400-850 mW	$



Summary of Latest sensAI Updates

HIGHER ACCURACY

iCE40 UltraPlus, the ultra-low power edge AI accelerator now delivers higher accuracy

HIGHER SPEED

ECP5 FPGA extends support to MobileNet and Resnet for higher speed processing at high accuracy

FOCUS APPLICATIONS

Expanded focus applications including object identification and HMI

NEW REFERENCE DESIGNS

New and updated demos and end-to-end reference designs

- Key Phrase Detection
- Human Identification
- Human Presence Detection
- Object Counting with MobileNet

LATTICE SEMICONDUCTOR®

The Low Power Programmable Leader

THANK YOU