embedded VISION SUMMIT 2018

Machine Learning Inference In Under 5 mW

w/ a Binarized Neural Network (BNN) in an FPGA

Abdullah Raouf

May, 2018

LATTICE SEMICONDUCTOR

Agenda

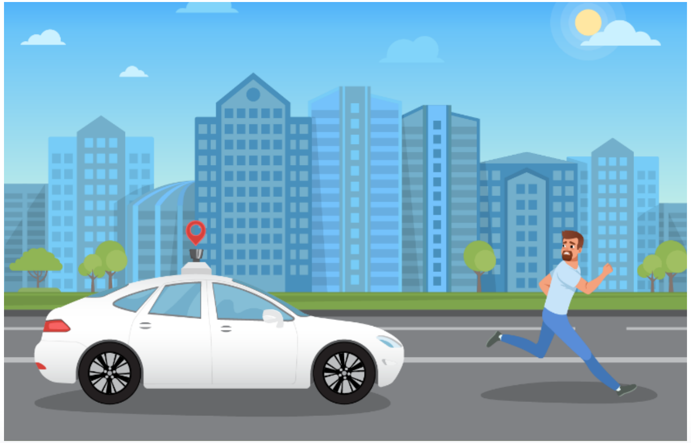
- Why Edge Intelligence?
- Introduction to Lattice & the new Lattice sensAI^{TM} stack
  - How to enable deep learning at the edge
  - Available tools and implementation methods
  - A full system example

Intelligence at the Edge Trend

Market Need: Immediate, locally processed, ML based analytics

Why:

The cloud takes too long to determine that a person is in front of you



Users do not want information sent, stored, or processed in the cloud

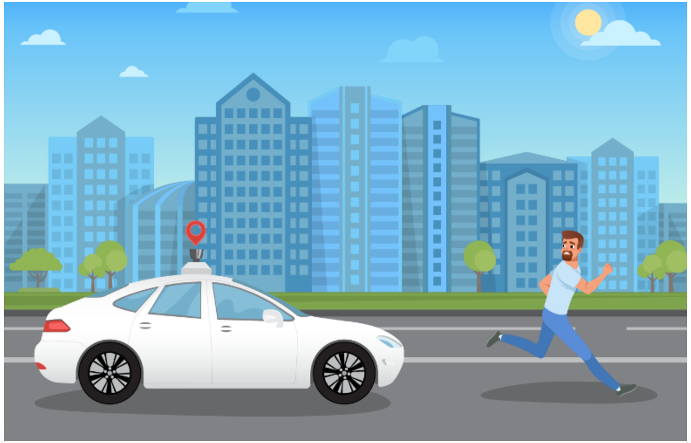
By 2019, 45% of IoT-Created Data Will Be Stored, Processed, Analyzed, and Acted Upon Close to, or at the Edge of the Network - IDC

Intelligence at the Edge Trend

Market Need: Immediate, locally processed, ML based analytics

Why:

The cloud takes too long to determine that a person is in front of you



Users do not want information sent, stored, or processed in the cloud

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

Who is Lattice?

Supplier of the worlds smallest FPGA

As small as 1.6 mm x 1.6 mm

Smallest SIZE

Sub $1.00 FPGAs

Production PRICED

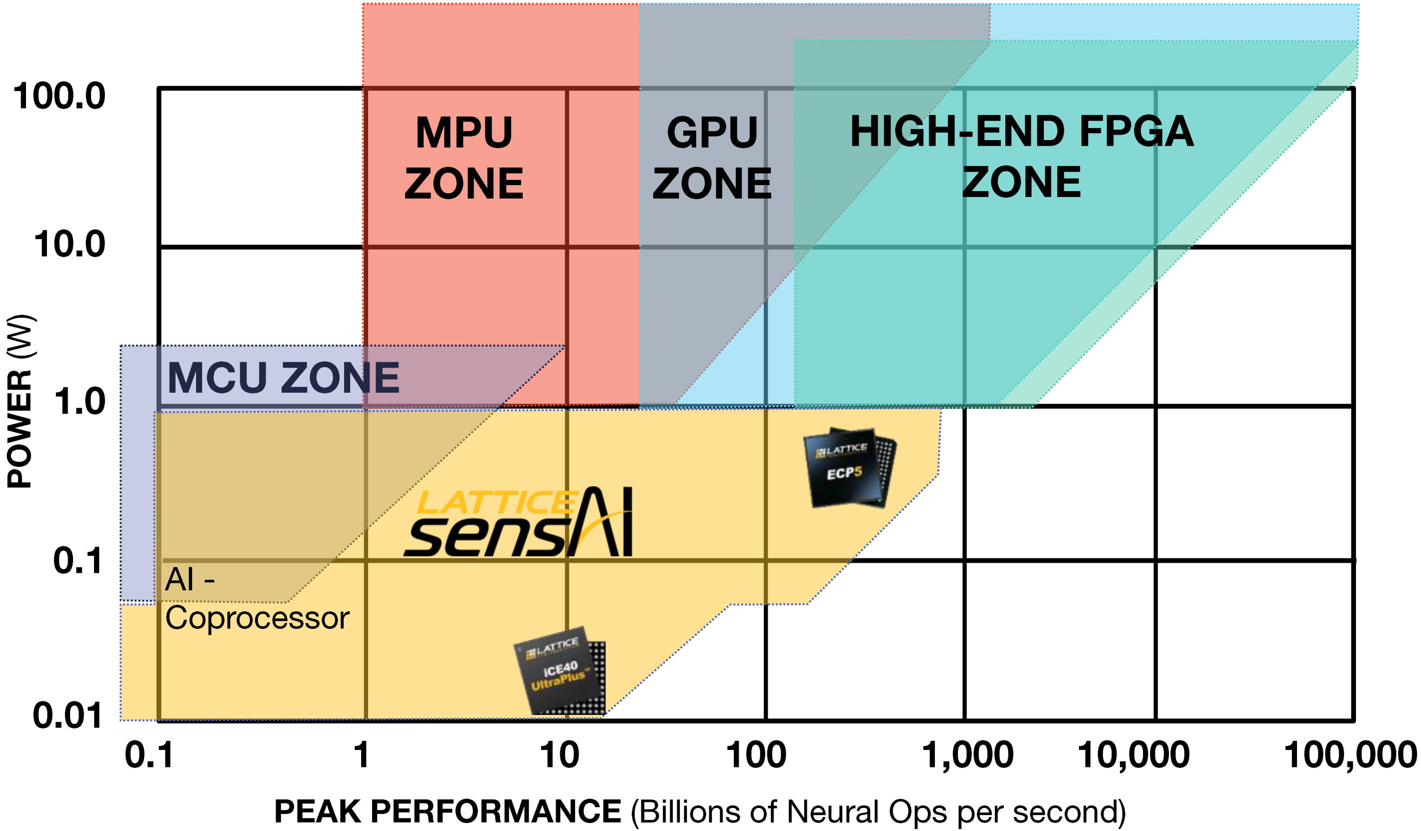
Lowest POWER

Fastest to MARKET

Single object detection using deep neural network @ 847 uW

LATTICE

Edge Device AI – Competitive Landscape



Introducing Lattice sensAI

Complete Technology Stack for Ultra-Low Power, Flexible Inferencing

CUSTOM DESIGN SERVICES

Mobile | Smart Home | Smart City | Smart I

REFERENCE DESIGNS / DEMOS

Face Detection | Key Phrase Detection | Face Tracking | Object

SOFTWARE TOOLS

Neural Network Compiler

LATTICE RADIANT DESIGN SOFTWARE | Caffe | Tens

IP CORES

Neural Network Accelerators

BNN Accelerator | CNN Accelerator

HARDWARE PLATFORMS

Mobile Developme

Platform – UltraPlus FPGA

Video Interface Platform – ECP5 FPGA

1mW, 5.5mm², 1 bit, ~$1 → 1W, 100mm², 8/16 bits, ~$10

iCE40 UltraPlus

Programmable FPGA Fabric

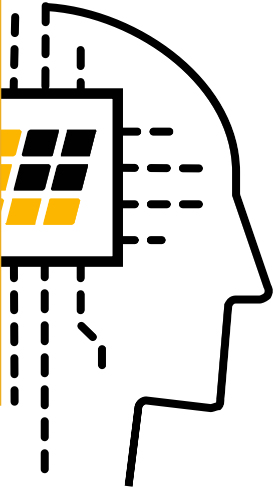
5,280 LUTs
120 Kb Block RAM

NVCM

8 DSP Blocks

1 Mb RAM

I/O's



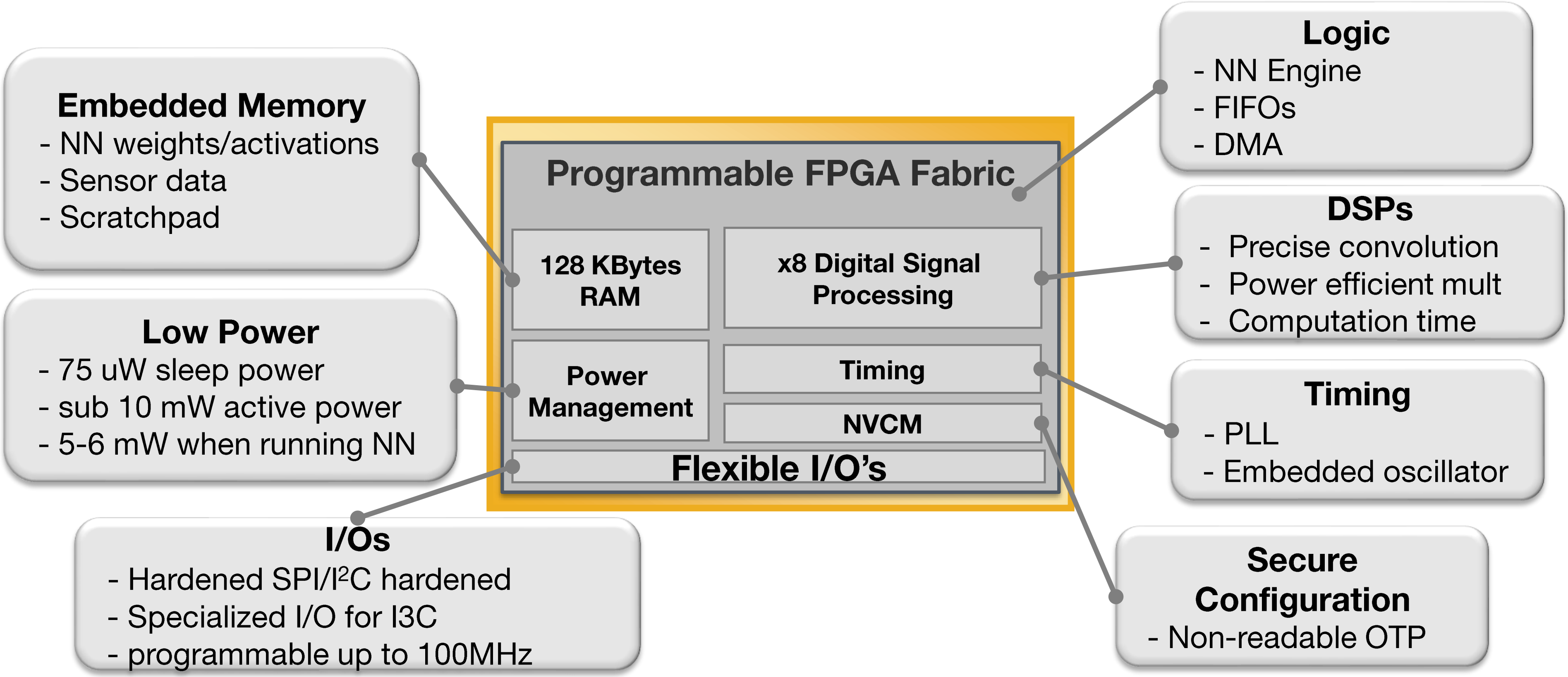
Ultra-Low Power

Small-Form Factor

Customizable

Production Priced

Introducing iCE40 UltraPlus FPGA



Embedded Memory
- NN weights/activations
- Sensor data
- Scratchpad

Low Power
- 75 uW sleep power
- sub 10 mW active power
- 5-6 mW when running NN

I/Os
- Hardened SPI/I^{2}C hardened
- Specialized I/O for I3C
- programmable up to 100MHz

Programmable FPGA Fabric

128 KBytes RAM

x8 Digital Signal Processing

Power Management

Timing

NVCM

Flexible I/O's

Logic
- NN Engine
- FIFOs
- DMA

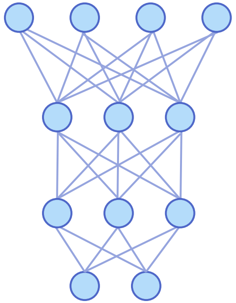
DSPs
- Precise convolution
- Power efficient mult
- Computation time

Timing
- PLL
- Embedded oscillator

Secure Configuration
- Non-readable OTP

Edge Acceleration in Lattice FPGAs

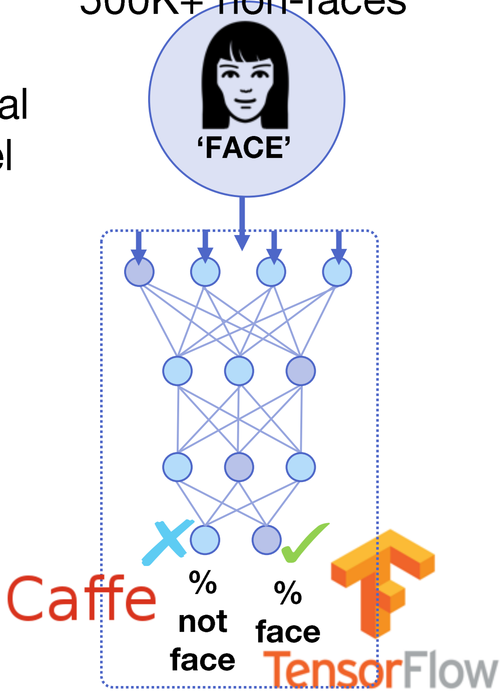
Untrained Neural Network Model



TRAINING DATASET

500K+ faces

500K+ non-faces

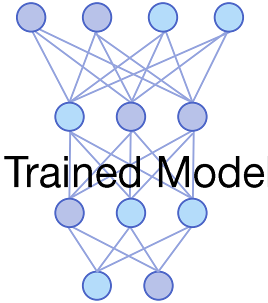


LATTICE SEMICONDUCTOR

LATTICE sensAI

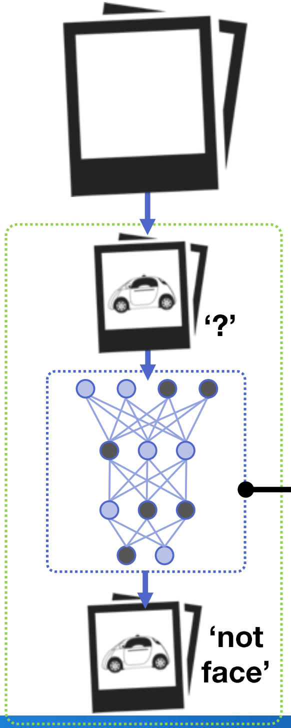
Neural Network Compiler Software

Neural Network Complier



LATTICE RADIANT DESIGN SOFTWARE

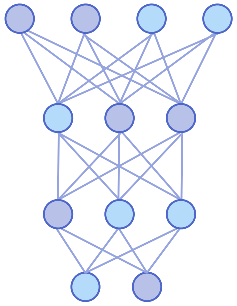
NEW DATA



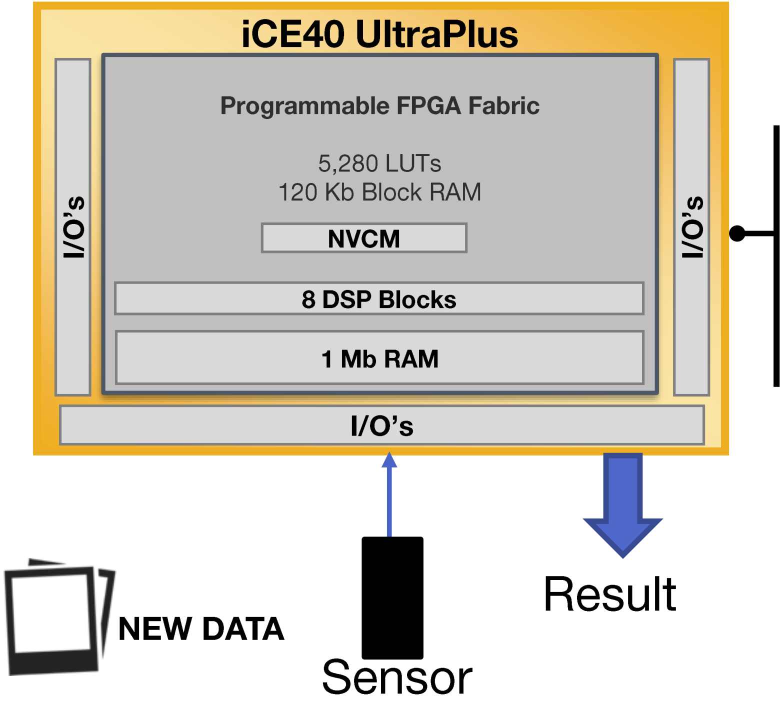
Trained Model

Optimized for performance

Edge Acceleration in iCE40 UP

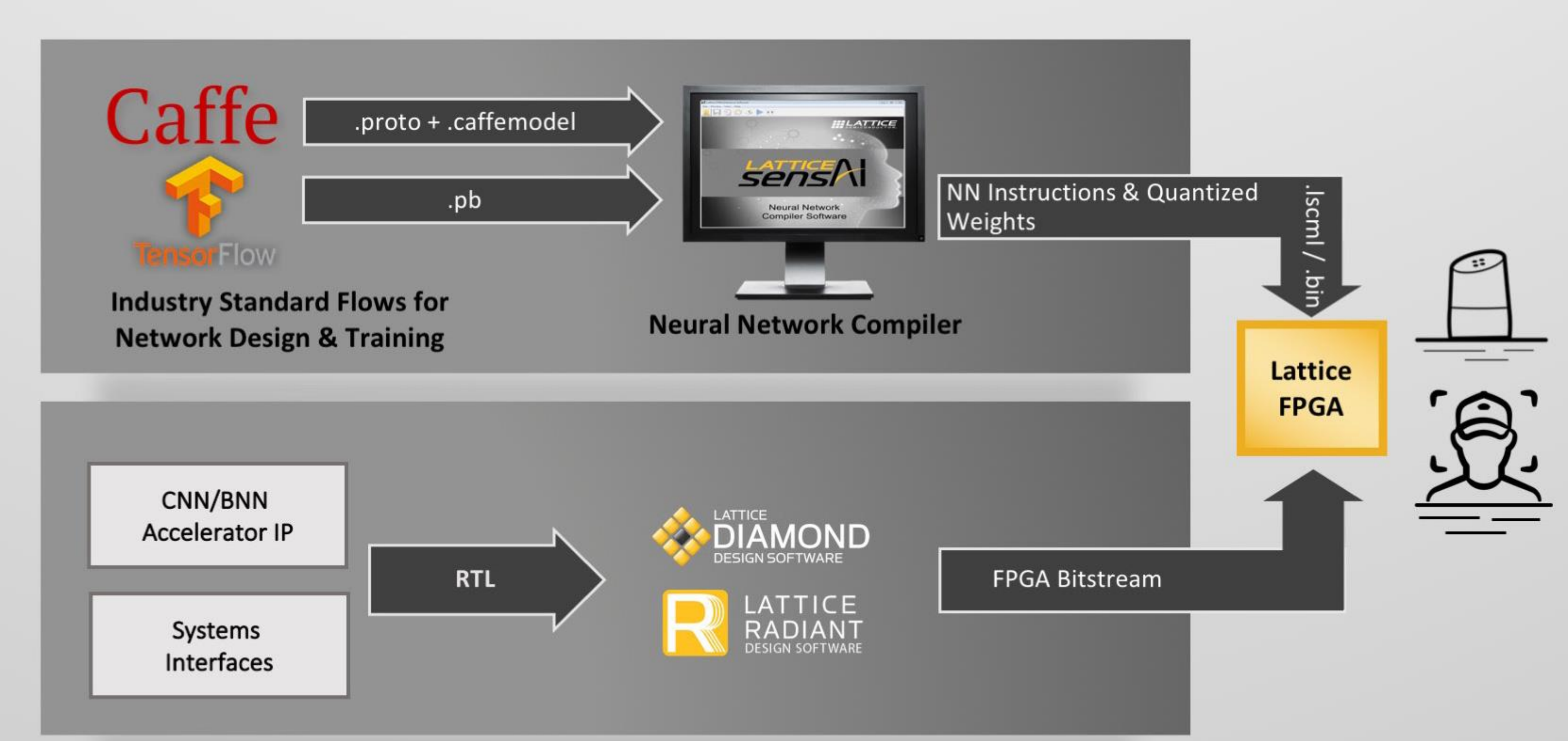


LATTICE RADIANT DESIGN SOFTWARE



Small size:	2.15 mm x 2.55 mm
Inferencing capability:	1.1 T ops/W
Quantization:	1bit W & A
Activation Layer:	tanh
# of parallel engines:	16
# of cycles per frame:	85 K
Estimated power:	4 to 6 mW
Computation time:	10 ms

Overview of Development Flow



Caffe

TensorFlow

.proto + .caffemodel

.pb

Industry Standard Flows for Network Design & Training

LATTICE sensAI

Neural Network Compiler Software

Neural Network Compiler

NN Instructions & Quantized Weights

.lscml / .bin

Lattice FPGA

CNN/BNN Accelerator IP

Systems Interfaces

RTL

LATTICE DIAMOND DESIGN SOFTWARE

LATTICE RADIANT DESIGN SOFTWARE

FPGA Bitstream

Real Example of using VGG with 32 x 32 RGB input

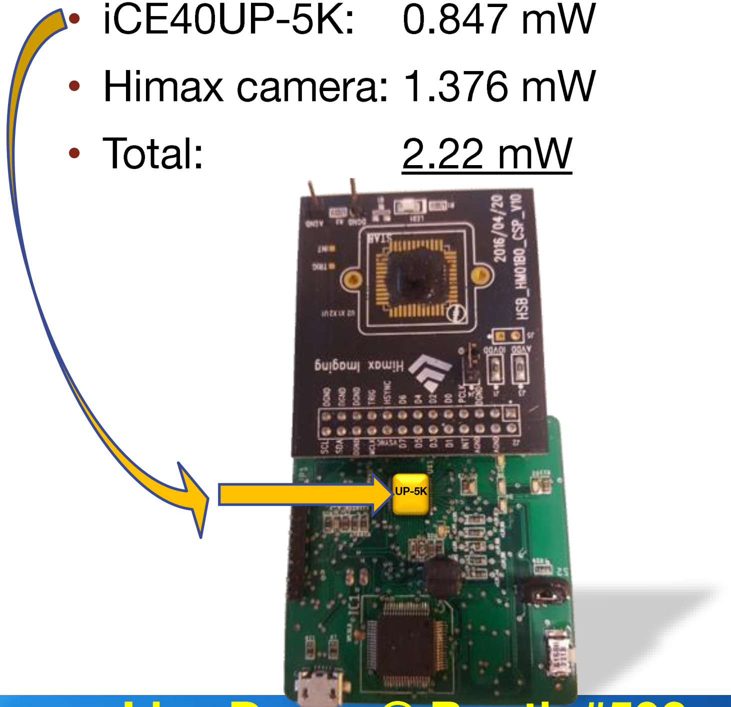
- 7 layers of BNN
- 2 class classification (Face or no Face)
- 32 x 32 RGB input

Face Detection

Face Det	MAC	Activation (mem in KB)		Weight (mem in KB)	
Layers	**# (M)**	**# (K)**	**1b**	**# (K)**	**1b**
Input		3	3		
Conv1	2	66	8	2	.22
Pool1		16	2		
Conv2	9	16	2	37	4.61
Pool2		4	1		
Conv3	5	8	1	74	9.22
Pool3		2	0		
FC9	0	0	0	4	.51
Total	**16**	**116**	**17**	**116**	**15**

- Power at 5 frame per sec speed

  - iCE40UP-5K: 0.847 mW
  - Himax camera: 1.376 mW
  - Total: <u>2.22 mW</u>

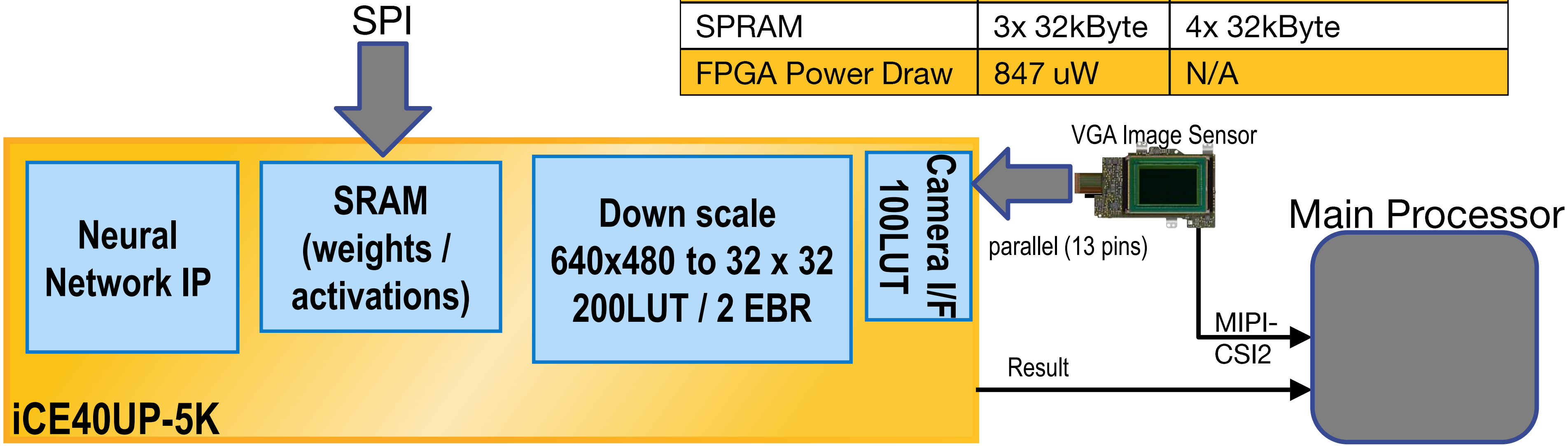


Live Demo @ Booth #502

System Block Diagram & FPGA Utilization

FPGA Resource Utilization

Resources Required		Available Resources
LUTs	3,931	5,280
DSP (16x16 Mult)	0	8
BRAM	20x 4kbit	30x 4kbit
SPRAM	3x 32kByte	4x 32kByte
FPGA Power Draw	847 uW	N/A



Market Needs: Local Person Detection

Smart Home Appliance

LCD turns on when needed

Consumer Electronics

TV turns off when nobody present

Smart DoorBell

Rings automatically when needed

Vending Machine

LCD turns on when needed

Security Camera

Alerts when intruder present, not a cat

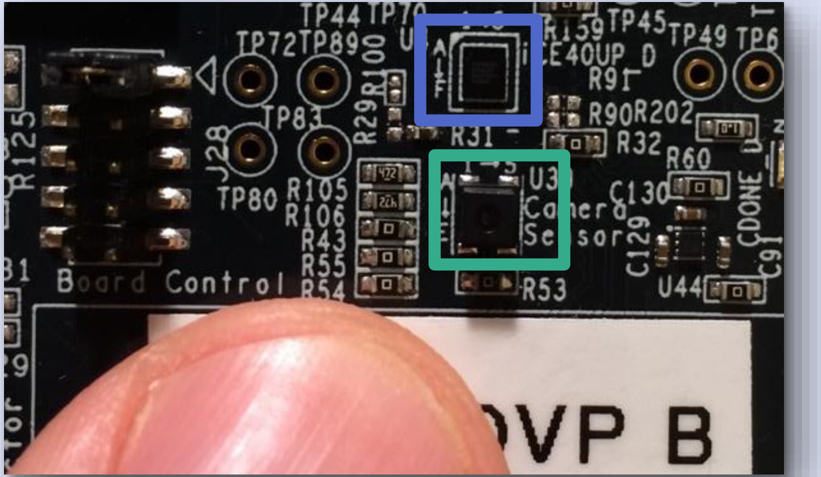
Smart Doors

Opens when person is present

Complete Face Detection Solution

iCE40 UltraPlus FPGA (2.15 mm x 2.55 mm)

Omnivision OVM7692 Camera



Example

Human Face Detection (Without the Cloud)

Live Demo @ Booth #502

Summary of Solution

- User 1 Bit Weights / Activations
- 0.8 mW Power Consumption at 5fps
- Standalone
- 2.15x2.55 mm Single Chip Solution
- 99% Accuracy

To Learn More about LATTICE sensAI

Stop by and talk to our team @ Booth 502

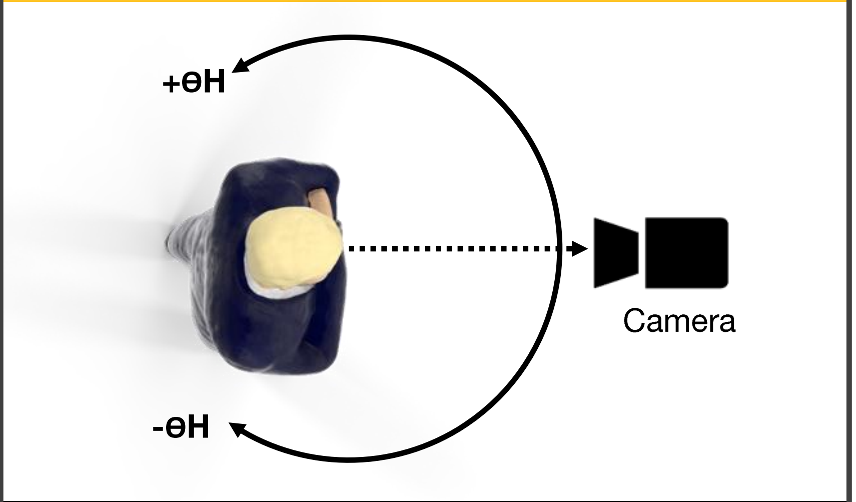
Visit our website @ www.latticesemi.com

Contact me @ Abdullah.Raouf@lscc.com

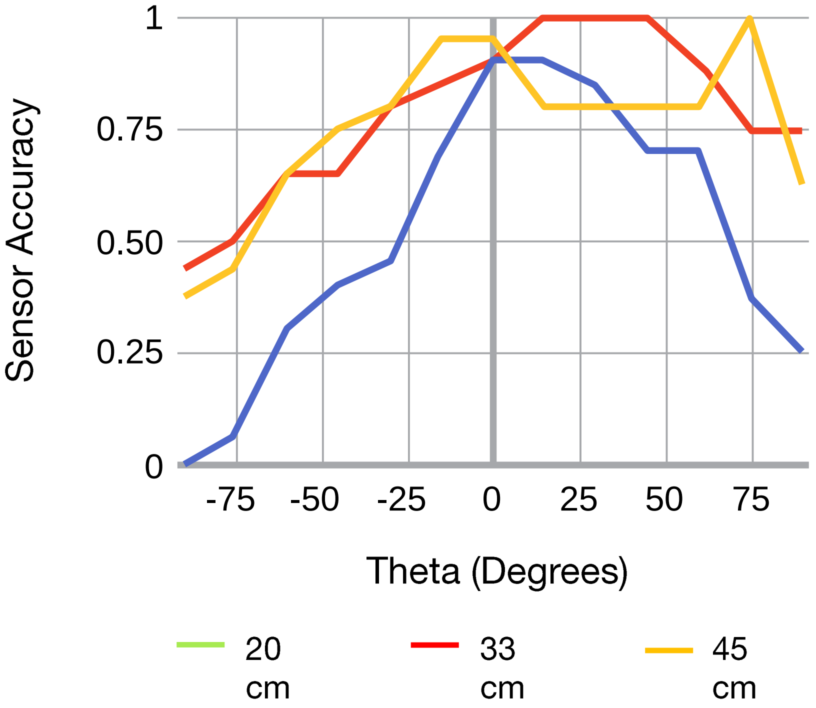
Summary of Results

Inference Engine Capabilities

Horizontal Camera Sweep (YAW)

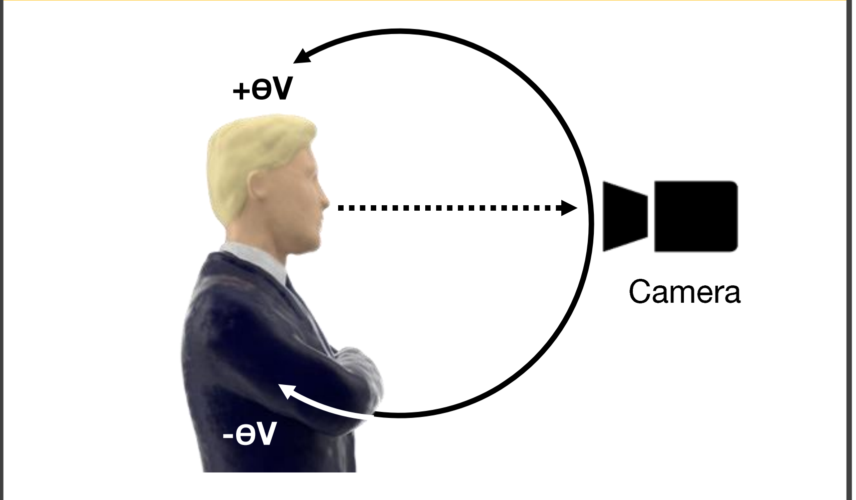


Item	Lattice Response
Detect range	20 cm to 45 cm
Face angle	Yaw: -12 to 35°

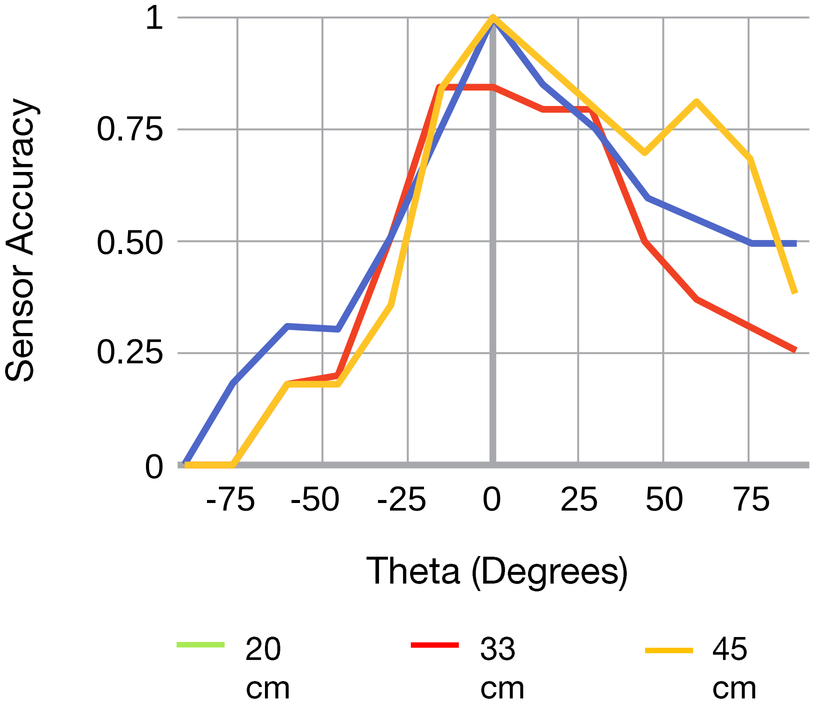


Inference Engine Capabilities

Vertical Camera Sweep (PITCH)

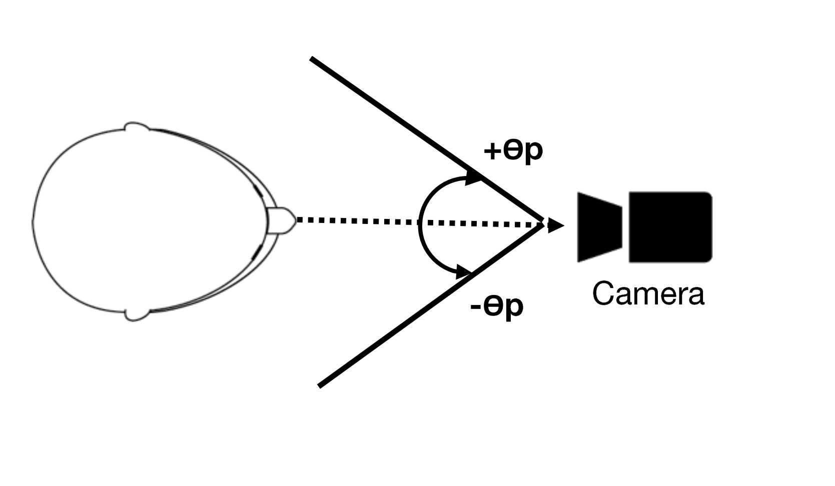


Item	Lattice Response
Detect range	20 cm to 45 cm
Face angle	Pitch: -25 to 25°

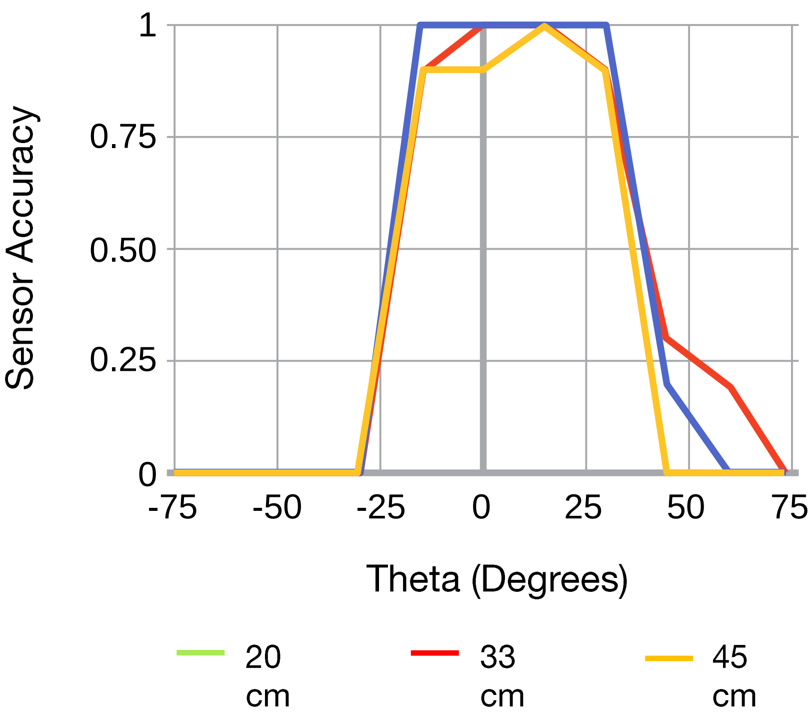


Inference Engine Capabilities

Camera Pan (Roll)

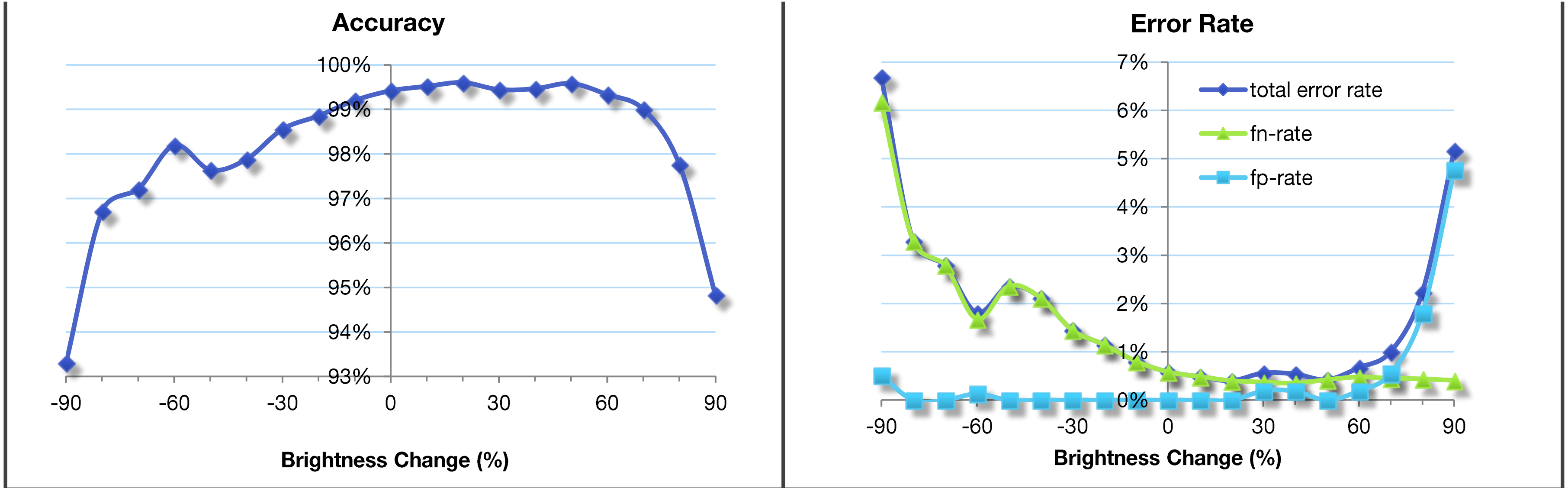


Item	Lattice Response
Detect range	20 cm to 45 cm
Face angle	Roll: -20 to 35°



Inference Engine Capabilities

True and False Positives with Variable Lighting Conditions



Item	Lattice Response
Detection Rate(True Positive)	>95%
Error Rate(False Positive)	<0.1%
Light condition	See plot above