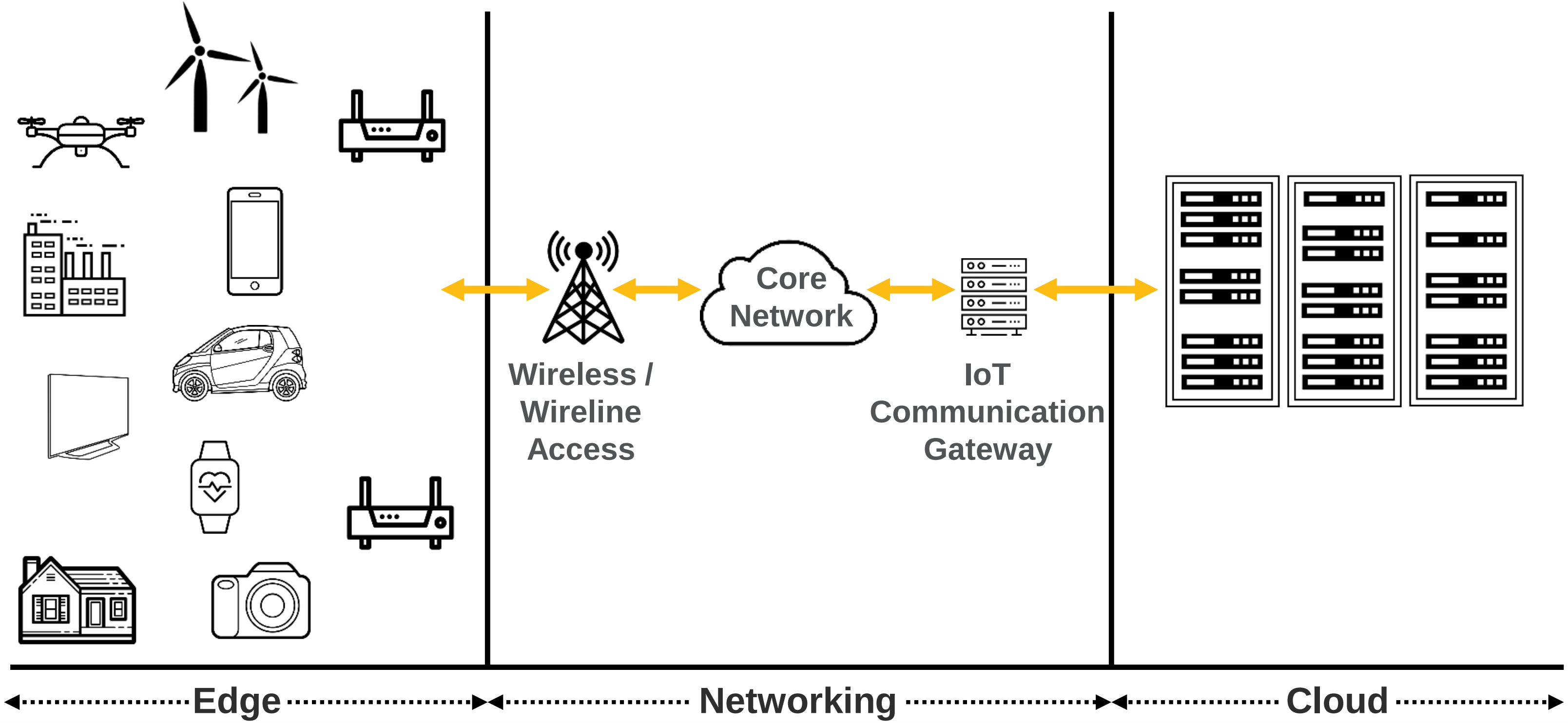
Architecting Always-On, Context-Aware, On-Device AI Using Flexible Low-power FPGAs

Deepak Boppana – Senior Director Product & Segment Marketing
Gordon Hands – Director Solutions Marketing

LATTICE SEMICONDUCTOR™

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*

Always-on, On-device AI Applications

Human Presence Detection Example

Smart Home Appliance

LCD turns on when needed

Consumer Electronics

TV turns off when no one is 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

Always-on, On-device AI Applications

Other Examples

Smart speakers
Key phrase detection

Retail store cameras
Face tracking

Selfie drones
Face tracking

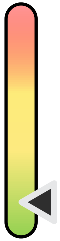
Toll gate camera
Vehicle classification

Machine vision
Object counting

After market automotive cameras
Speed sign detection

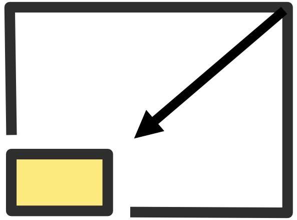
Always-on, On-device AI Requirements

Unmet Need for Ultra-Low Power, Scalable, and Flexible Inferencing

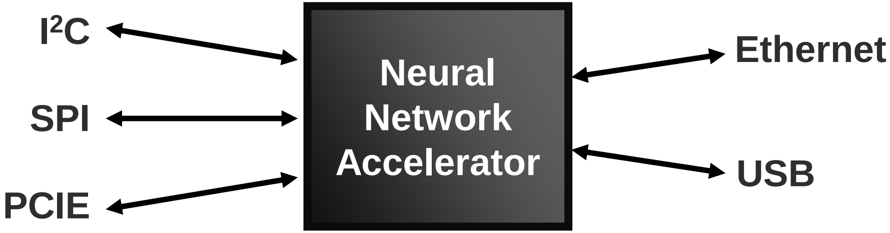


Few mWs of Power Consumption

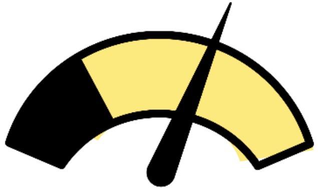
Few $s of BOM Cost Adder



Few mm² of Board Area



Flexible Legacy Interface Support



Customized Performance/Accuracy

LATTICE sensAI

CUSTOM DESIGN SERVICES

Mobile | 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

Caffe | TensorFlow

LATTICE DIAMOND DESIGN SOFTWARE

IP CORES

Neural Network Accelerators

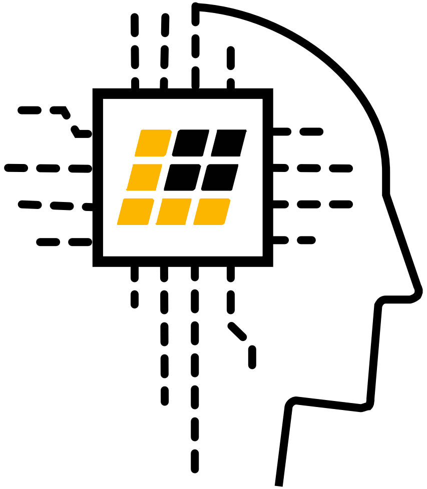
CNN Compact Accelerator | CNN Accelerator

HARDWARE PLATFORMS

Mobile Development Platform – iCE40 UltraPlus FPGA

Video Interface Platform – ECP5 FPGA

1 mW, 5.5 mm^2, 1/16 bits → 1 W, 100 mm^2, 1/8/16 bits



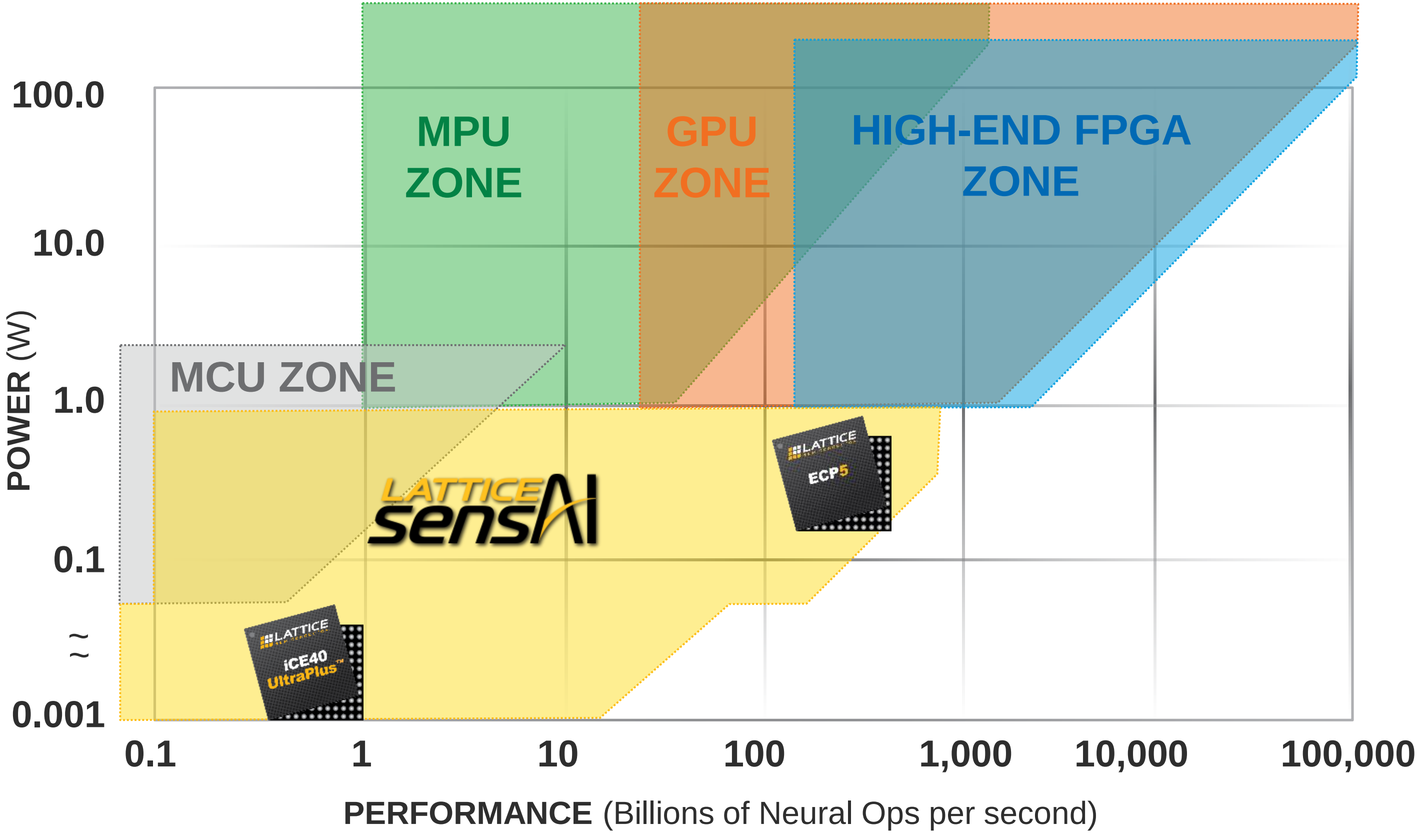
Ultra Low Power

Small Form Factor

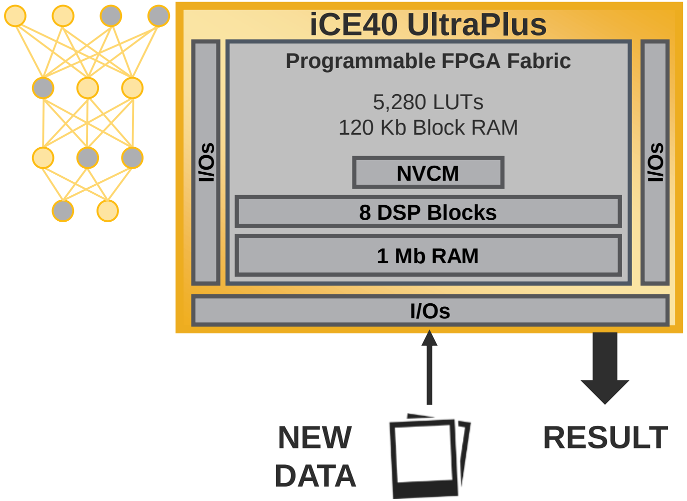
Customizable

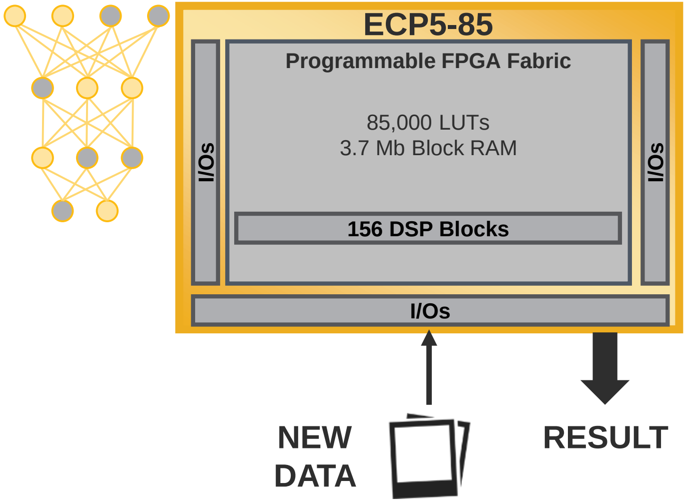
Flexible and Scalable Inferencing at the Edge

From under 1 mW to 1 W with Lattice sensAI



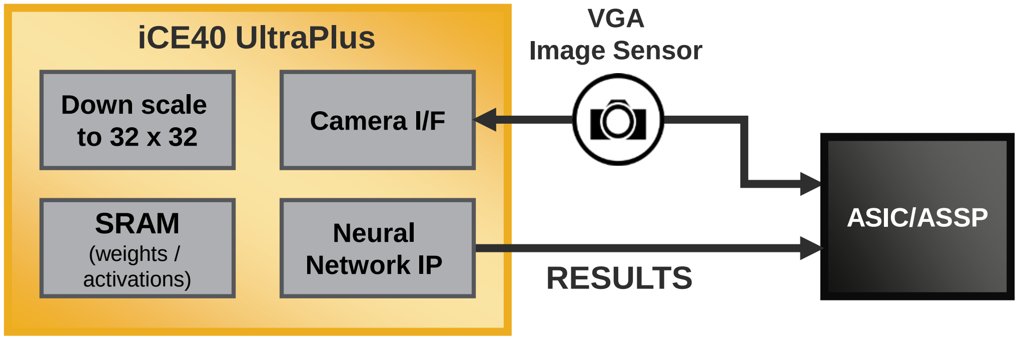
Stand-alone, Integrated FPGA Solution





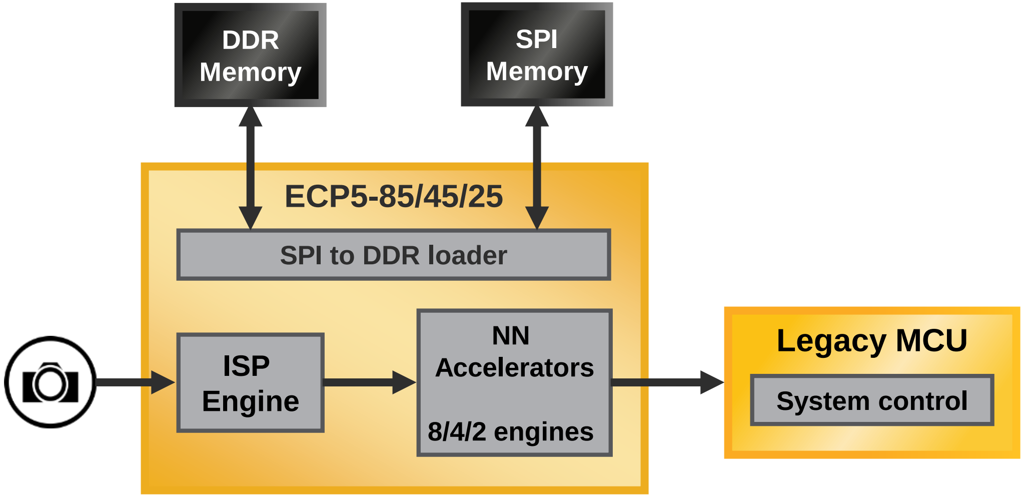
- Always-on, integrated solutions on ECP5 or iCE40 UltraPlus FPGA
- Low latency and secure implementation
- Small form factor packages from 5.5 mm^2 to 100 mm^2

FPGA as Activity Gate to ASIC/ASSP



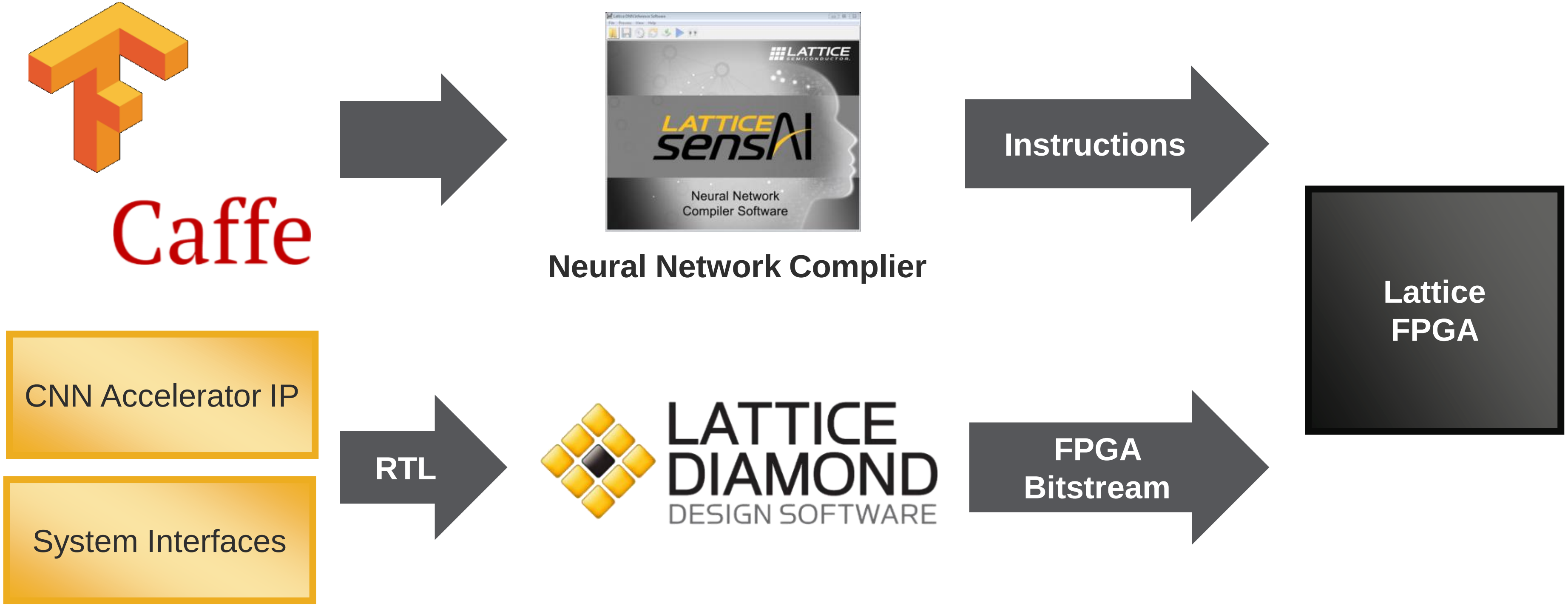
- iCE40 UltraPlus FPGA for always-on detection of key-phrases or objects
- Wakes-up a high performance ASIC/ASSP for further analytics only when required
- Reduces overall system power consumption

FPGA as a Co-Processor to MCU



- Scalable performance/power with ECP5 based neural network acceleration
- ECP5 based IO flexibility to seamlessly interface to on-board legacy devices including sensors
- Low-end MCU for flexible system control

Delivering Edge CNN Acceleration in Lattice FPGA



Caffe

CNN Accelerator IP

System Interfaces

RTL

LATTICE sensAI

Neural Network Compiler Software

Neural Network Complier

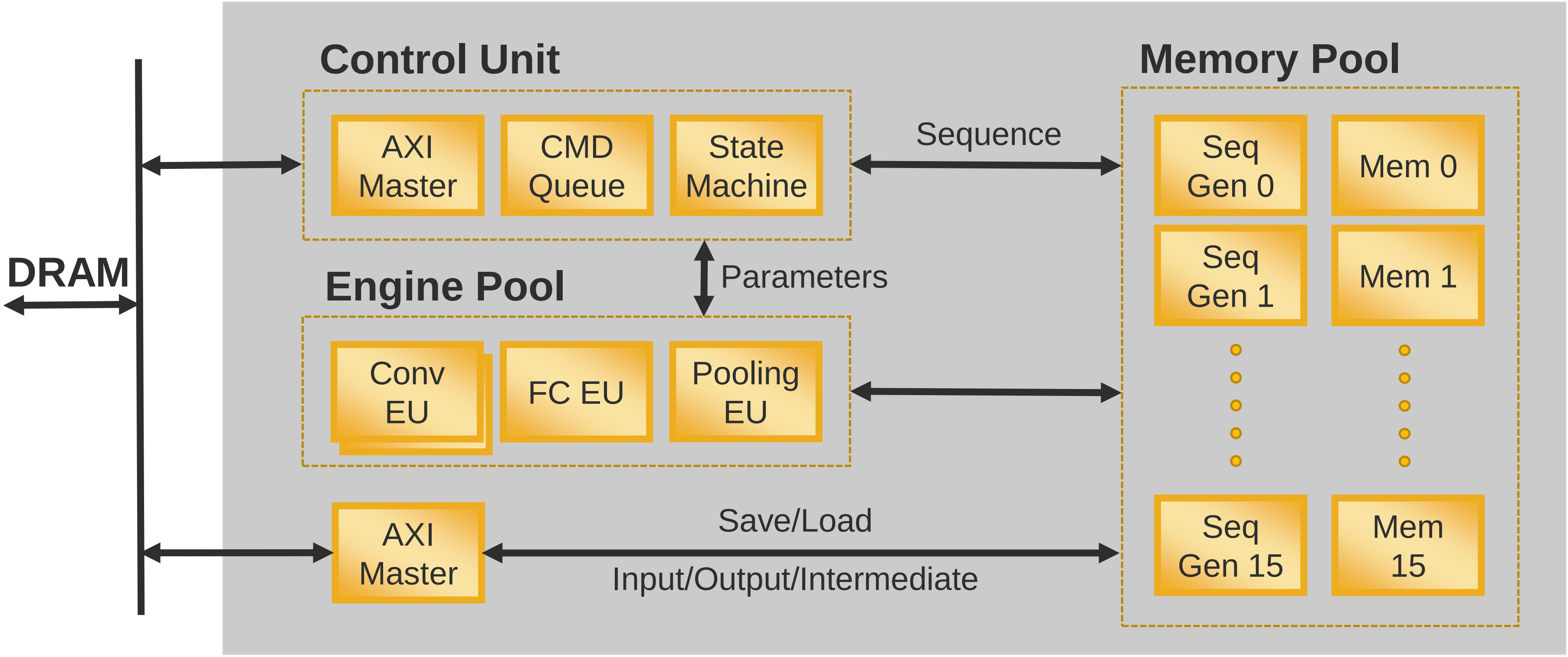
LATTICE DIAMOND DESIGN SOFTWARE

Instructions

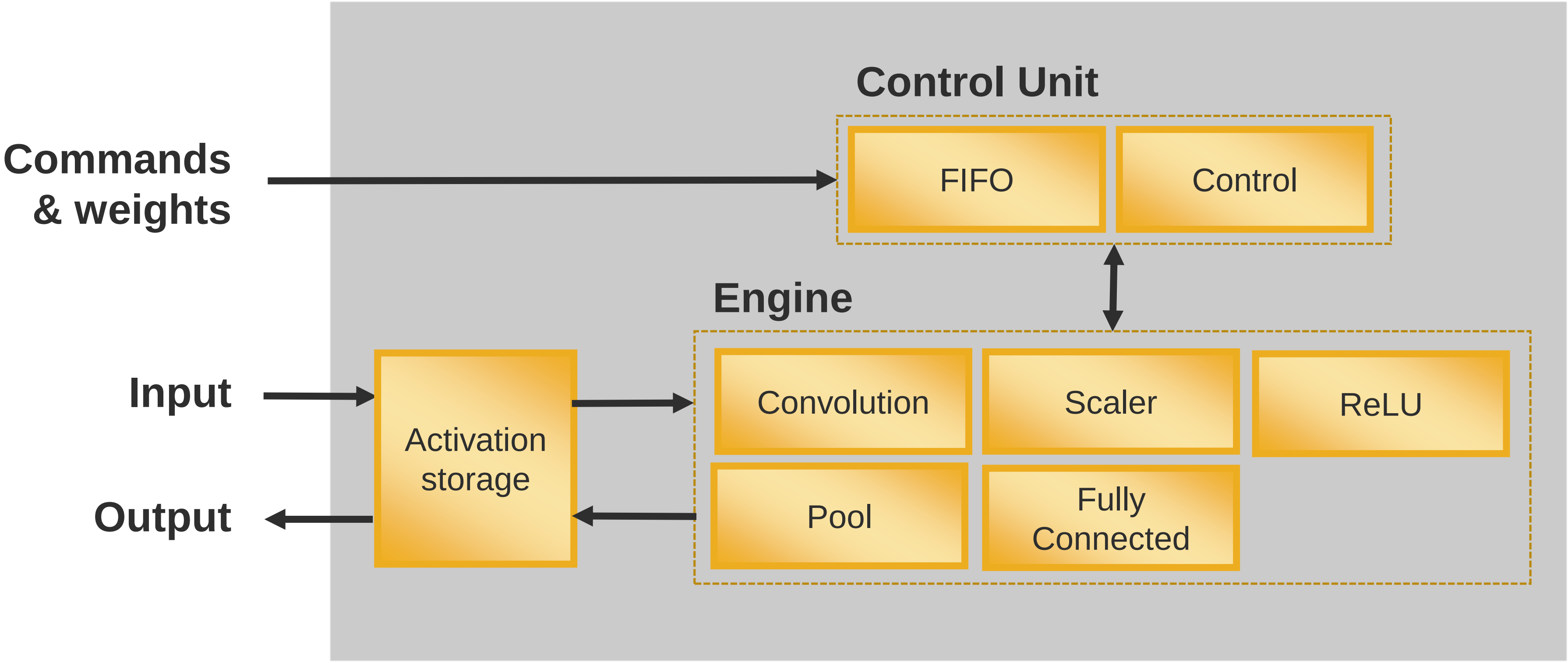
FPGA Bitstream

Lattice FPGA

CNN Accelerator IP Architecture

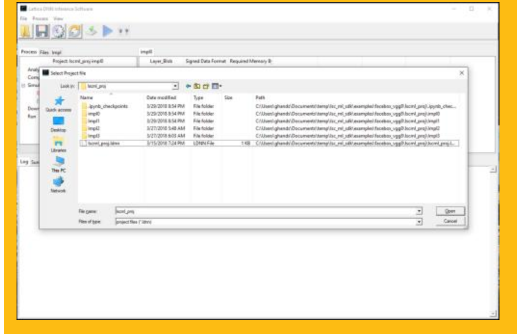


CNN Compact Accelerator IP Architecture



Translating Trained Neural Network Into Lattice CNN Accelerator Instructions

1. Load

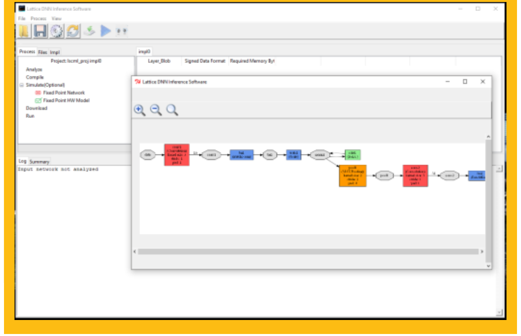


Name	Date modified	Type	Size	Path
.ipynb_checkpoints	3/29/2018 8:54 PM	File folder		C:\Users\ghands\Documents\temp\lsc_ml_sdk\examples\facebox_vgg8\lscml_proj\.ipynb_chec...
impl0	3/29/2018 8:54 PM	File folder		C:\Users\ghands\Documents\temp\lsc_ml_sdk\examples\facebox_vgg8\lscml_proj\impl0
impl1	3/29/2018 8:54 PM	File folder		C:\Users\ghands\Documents\temp\lsc_ml_sdk\examples\facebox_vgg8\lscml_proj\impl1
impl2	3/27/2018 5:48 AM	File folder		C:\Users\ghands\Documents\temp\lsc_ml_sdk\examples\facebox_vgg8\lscml_proj\impl2
impl3	3/27/2018 6:03 AM	File folder		C:\Users\ghands\Documents\temp\lsc_ml_sdk\examples\facebox_vgg8\lscml_proj\impl3
lscml_proj.ldnn	3/15/2018 7:24 PM	LDNN File	1 KB	C:\Users\ghands\Documents\temp\lsc_ml_sdk\examples\facebox_vgg8\lscml_proj\lscml_proj.l...

File name: lscml_proj

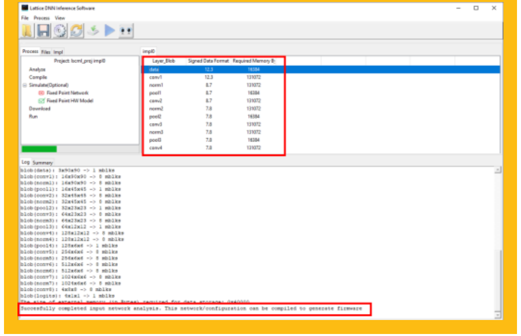
Files of type: project files (*.ldnn)

2. Review



```
Input network not analyzed
```

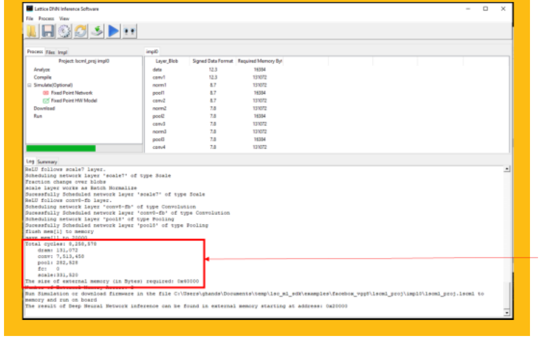
3. Analyze



Layer_Blob	Signed Data Format	Required Memory B
data	12.3	16384
conv1	12.3	131072
norm1	8.7	131072
pool1	8.7	16384
conv2	8.7	131072
norm2	7.8	131072
pool2	7.8	16384
conv3	7.8	131072
norm3	7.8	131072
pool3	7.8	16384
conv4	7.8	131072

```
blob(data): 3x90x90 -> 1 mblks
blob(conv1): 16x90x90 -> 8 mblks
blob(norm1): 16x90x90 -> 8 mblks
blob(pool1): 16x45x45 -> 1 mblks
blob(conv2): 32x45x45 -> 8 mblks
blob(norm2): 32x45x45 -> 8 mblks
blob(pool2): 32x23x23 -> 1 mblks
blob(conv3): 64x23x23 -> 8 mblks
blob(norm3): 64x23x23 -> 8 mblks
blob(pool3): 64x12x12 -> 1 mblks
blob(conv4): 128x12x12 -> 8 mblks
blob(norm4): 128x12x12 -> 8 mblks
blob(pool4): 128x6x6 -> 1 mblks
blob(conv5): 256x6x6 -> 8 mblks
blob(norm5): 256x6x6 -> 8 mblks
blob(conv6): 512x6x6 -> 8 mblks
blob(norm6): 512x6x6 -> 8 mblks
blob(conv7): 1024x6x6 -> 8 mblks
blob(norm7): 1024x6x6 -> 8 mblks
blob(conv8): 4x6x6 -> 8 mblks
blob(logits): 4x1x1 -> 1 mblks
The size of external memory (in Bytes) required for data storage: 0x40000
Successfully completed input network analysis. This network/configuration can be compiled to generate firmware
```

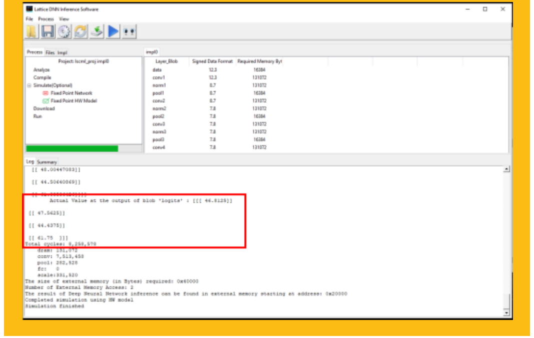
4. Compile



Layer_Blob	Signed Data Format	Required Memory By
data	12.3	16384
conv1	12.3	131072
norm1	8.7	131072
pool1	8.7	16384
conv2	8.7	131072
norm2	7.8	131072
pool2	7.8	16384
conv3	7.8	131072
norm3	7.8	131072
pool3	7.8	16384
conv4	7.8	131072

```
ReLU follows scale7 layer.
Scheduling network layer 'scale7' of type Scale
Fraction change over blobs
scale layer works as Batch Normalize
Sucessfully Scheduled network layer 'scale7' of type Scale
ReLU follows conv8-fb layer.
Scheduling network layer 'conv8-fb' of type Convolution
Sucessfully Scheduled network layer 'conv8-fb' of type Convolution
Scheduling network layer 'pool8' of type Pooling
Sucessfully Scheduled network layer 'pool8' of type Pooling
flush mem[1] to memory
save mem[1] to 20000
Total cycles: 8,258,578
    dram: 131,072
    conv: 7,513,458
    pool: 282,528
    fc:   0
    scale:331,520
The size of external memory (in Bytes) required: 0x40000
Number of External Memory Access: 2
Run Simulation or download firmware in the file C:\Users\ghands\Documents\temp\lsc_ml_sdk\examples\facebox_vgg8\lscml_proj\impl0\lscml_proj.lscml to
memory and run on board
The result of Deep Neural Network inference can be found in external memory starting at address: 0x20000
```

5. Simulate



Layer_Blob	Signed Data Format	Required Memory By
data	12.3	16384
conv1	12.3	131072
norm1	8.7	131072
pool1	8.7	16384
conv2	8.7	131072
norm2	7.8	131072
pool2	7.8	16384
conv3	7.8	131072
norm3	7.8	131072
pool3	7.8	16384
conv4	7.8	131072

```
  [[ 48.00447083]]

  [[ 44.50640869]]

  [[ 62.00000000]]]]
        Actual Value at the output of blob 'logits' : [[[ 46.8125]]

 [[ 47.5625]]

 [[ 44.4375]]

 [[ 61.75  ]]]
Total cycles: 8,258,578
    dram: 131,072
    conv: 7,513,458
    pool: 282,528
    fc:   0
    scale:331,520
The size of external memory (in Bytes) required: 0x40000
Number of External Memory Access: 2
The result of Deep Neural Network inference can be found in external memory starting at address: 0x20000
Completed simulation using HW model
Simulation finished
```

On-device AI – Complex Optimization

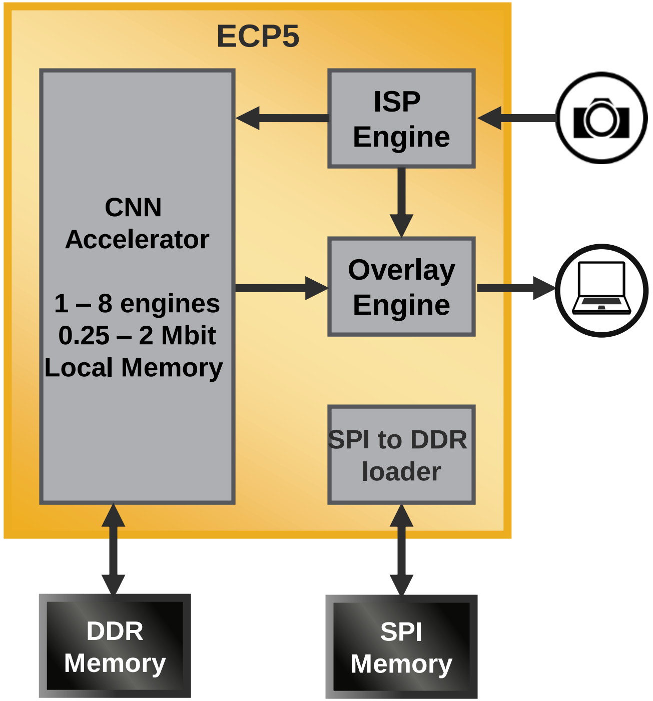
Design Factors Attributes	Device		Network		
	# of Engines	Local Memory	Input Size	Number of Multipliers	Bit Widths
Power (W)	🟥	🟩			
Device Size	🟥	🟥			
Performance (fps)	🟩	🟩	🟥	🟥	🟥
Accuracy (%)			🟩	🟩	🟩
Small Object (% fov)			🟩		

Correlation Between Design Factors and Product Attributes

Examples for Illustration

	Architecture	Number of Multiplications	Input Size	Quantization
Face Detection	VGG style	290,816	32*32*3	16-bit fixed point
	VGG style	14,353,920	90*90*3	16-bit fixed point
Human Presence Detection	VGG style	8,570,880	64*64*3	16-bit fixed point
	VGG style	338,558,976	128*128*3	16-bit fixed point

Image Based Neural Networks on Lattice FPGAs



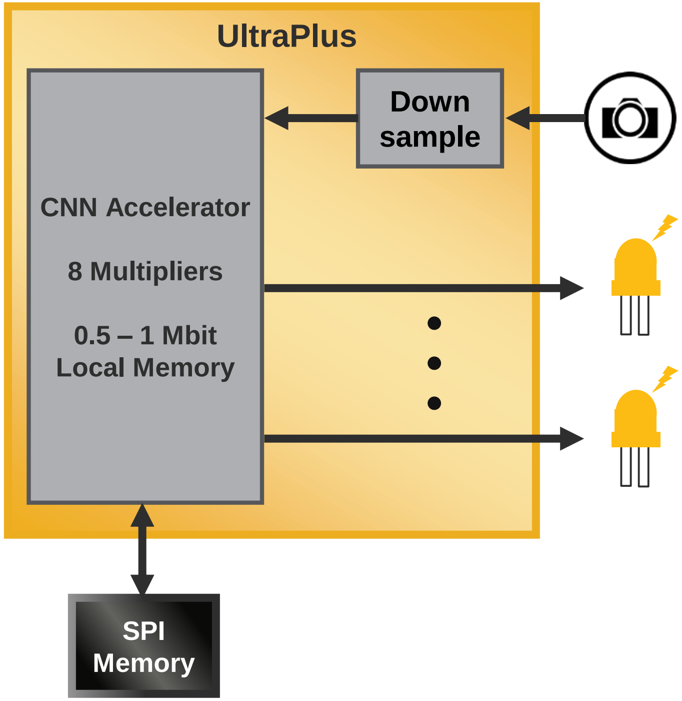
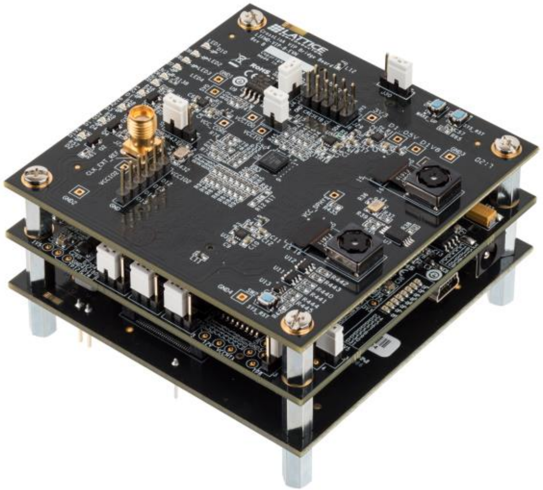


Image Based Neural Networks Lattice Hardware

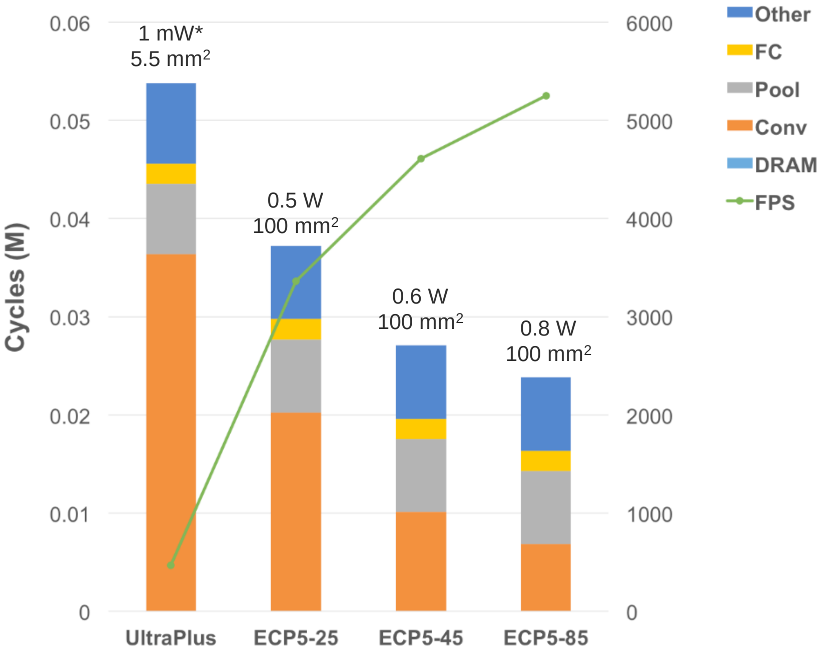
Himax HM01B0 UPduino Shield



Embedded Vision Development Kit

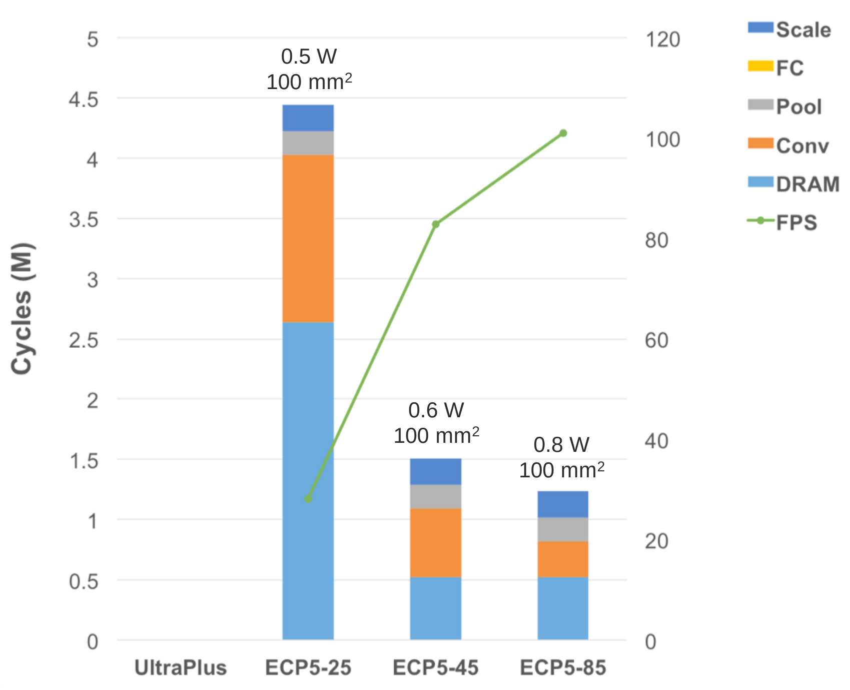
Face Detect Implementations

32 x 32 Input

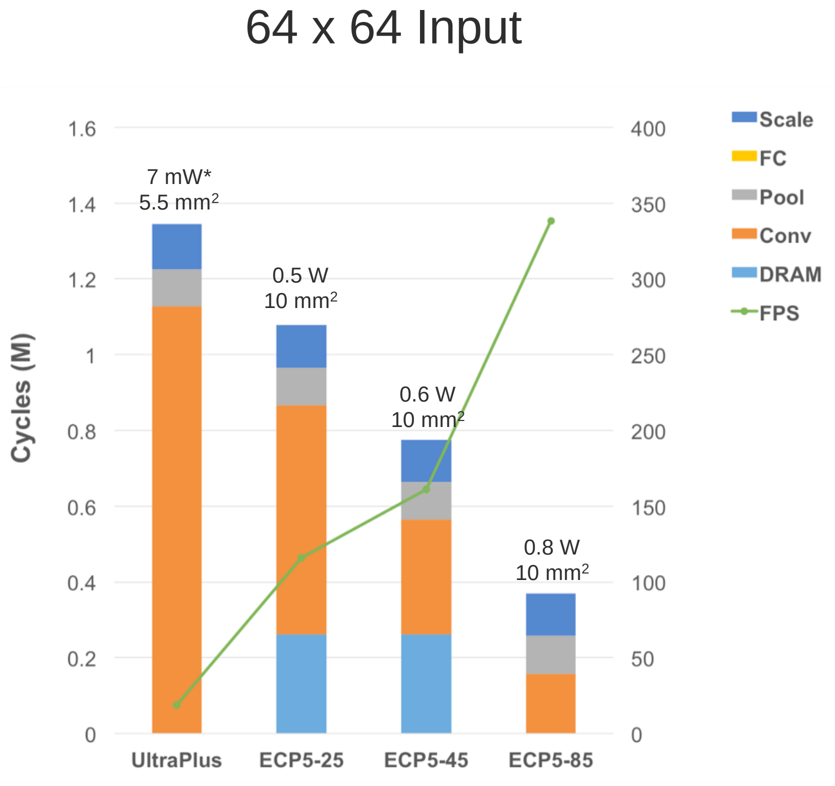


* Running at 5 frames per second

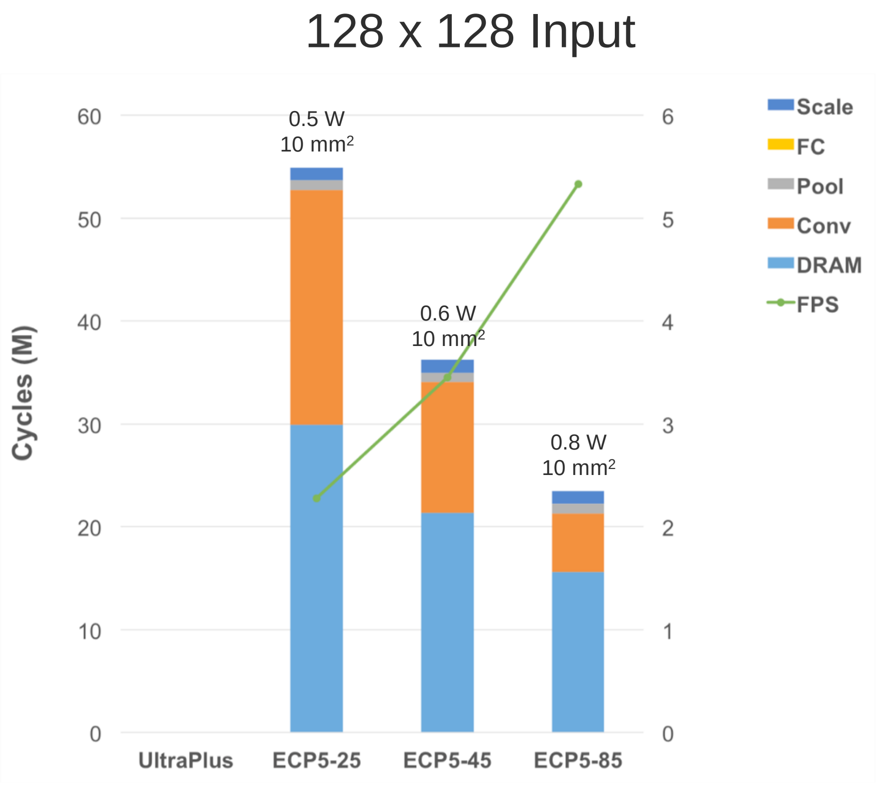
90 x 90 Input



Human Presence Detect Implementations



* Running at 5 frames per second



Bringing It Together

		Device Size / Power / Performance			
Network	**Smallest Object**	**UltraPlus** **1 – 7 mW*** **5.5 mm^2**	**ECP5-25** **0.5 W** **100 mm^2**	**ECP5-45** **0.6 W** **100 mm^2**	**ECP5-85** **0.8 W** **100 mm^2**
Face Detection 32 x 32 Input	50%	465	3360	4511	5251
Face Face Detection 90 x 90 Input	20%	--	28	82	101
Human Presence Detect 64 x 64 Input	20%	18	115	161	338
Human Presence Detect 128 x 128 Input	10%	--	2.3	3.5	5.4

* Running at 5 frames per second

Summary

- AI at the edge solves real world problems
- FPGAs can implement AI standalone or in conjunction with other components
- sensAI stack components provide edge AI building blocks
  - Silicon, soft IP, tools, development boards & reference designs
- Configurable engine size and bit widths coupled with multiple target devices allows system optimization
  - 1 mW – 1 W
  - 5.5 mm^2 – 100 mm^2

Resources

Please visit latticesemi.com/sensAI for more information and downloads

- 4 ECP5 Based Reference Designs / Demonstrations – Free
- 4 iCE40 Based Reference Designs / Demonstrations – Free
- CNN Accelerator IP – Free Evaluation
- CNN Compact Accelerator IP – Free
- Neural Network Compiler – Free
- Embedded Vision Development Kit – $199 Promotional Price
- Himax HM01B0 UPduino Shield – Available November ~$49

Empowering Product Creators to Harness Embedded Vision

The Embedded Vision Alliance (www.Embedded-Vision.com) is a partnership of 90+ leading embedded vision technology and services suppliers, and solutions providers

Mission: Inspire and empower product creators to incorporate visual intelligence into their products

The Alliance provides low-cost, high-quality technical educational resources for product developers

Register for updates at www.Embedded-Vision.com

The Alliance enables vision technology providers to grow their businesses through leads, ecosystem partnerships, and insights

For membership, email us: membership@Embedded-Vision.com

Embedded Vision Insights

The Latest Developments on Designing Machines that See

Join us at the Embedded Vision Summit

May 20-23, 2019—Santa Clara, California

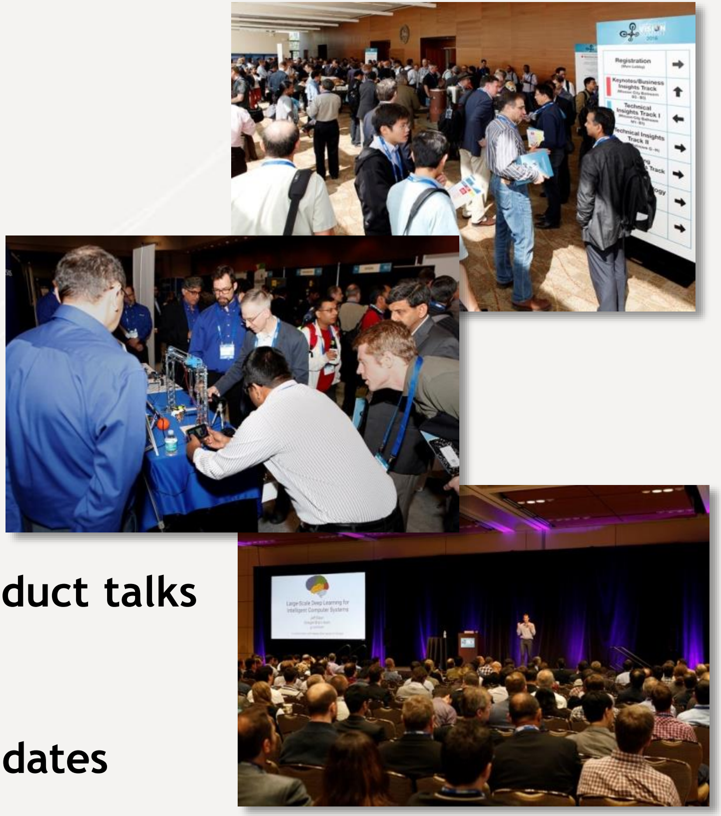
The only industry event focused on enabling product creators to create "machines that see"

- *"Awesome! I was very inspired!"*
- *"Fantastic. Learned a lot and met great people."*
- *"Wonderful speakers and informative exhibits!"*

Embedded Vision Summit 2019 highlights:

- **Inspiring keynotes** by leading innovators
- High-quality, practical **technical, business and product talks**
- Exciting **demos** of the latest apps and technologies

Visit www.EmbeddedVisionSummit.com **to sign up for updates**



Q & A

Thank you

LATTICE SEMICONDUCTOR™