embedded VISION SUMMIT 2018

Programmable CNN Acceleration in Under 1 Watt

Gordon Hands

May 22, 2018

LATTICE SEMICONDUCTOR

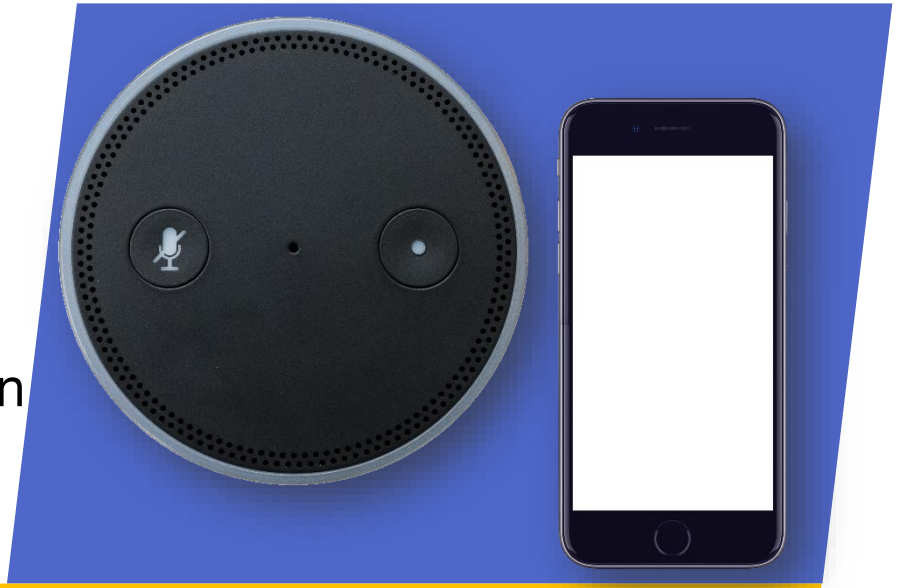
Drivers for Artificial Intelligence at the Edge

- Improving Privacy
- Simplified Regulation Compliance
- Reducing Bandwidth Required
- Optimizing Use of Cloud Computing
- Minimizing Latency

Edge Device AI Requirements

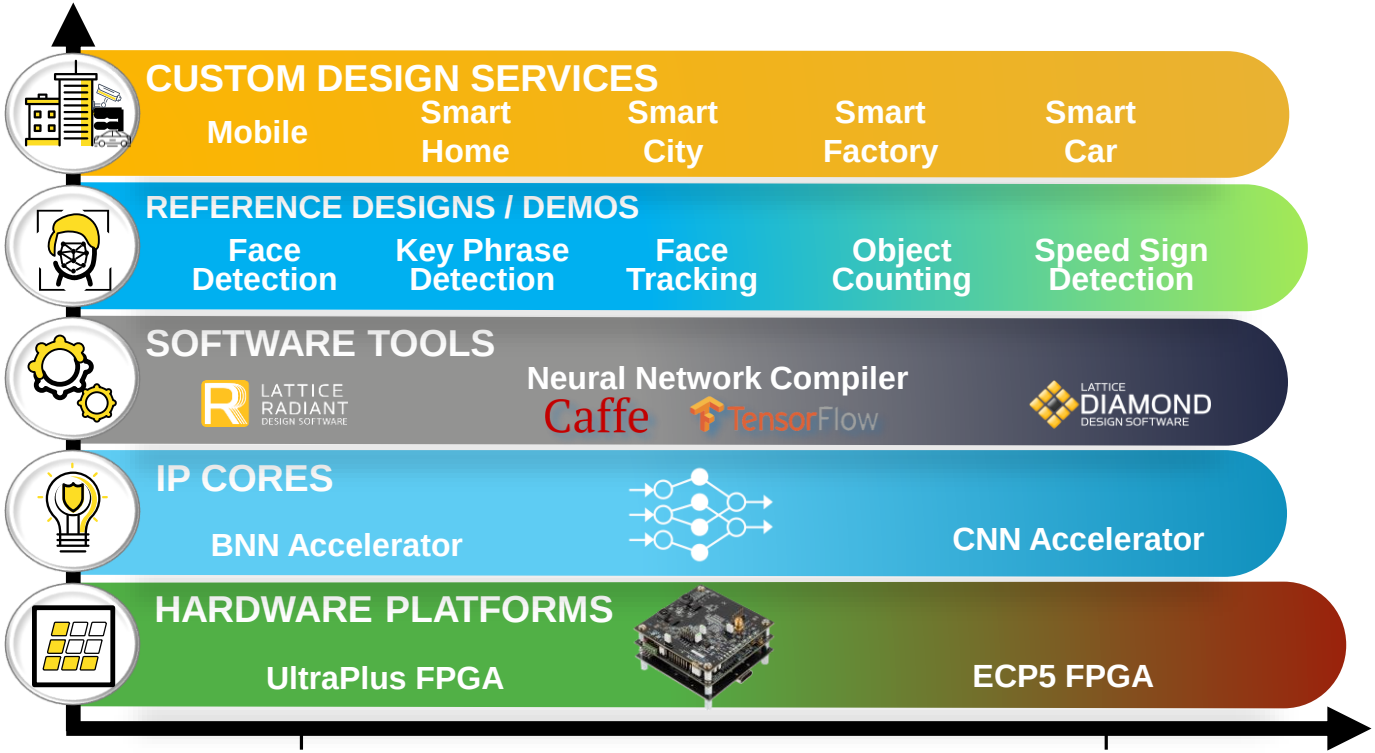
Edge Device Requirements

- Low Power
- Integration for Small Form Factor
- Fast Development
- Low Cost for High Volume Production
- Moderate Performance Inferencing



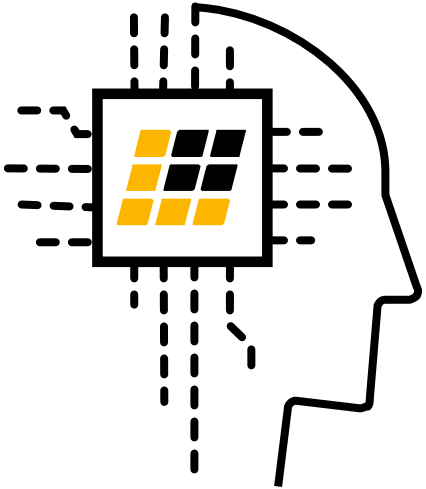
Lattice is Focused on Adding AI Capability to its Flexible Low Cost, Low-power Production Priced FPGA Solutions

Introducing Lattice sensAI

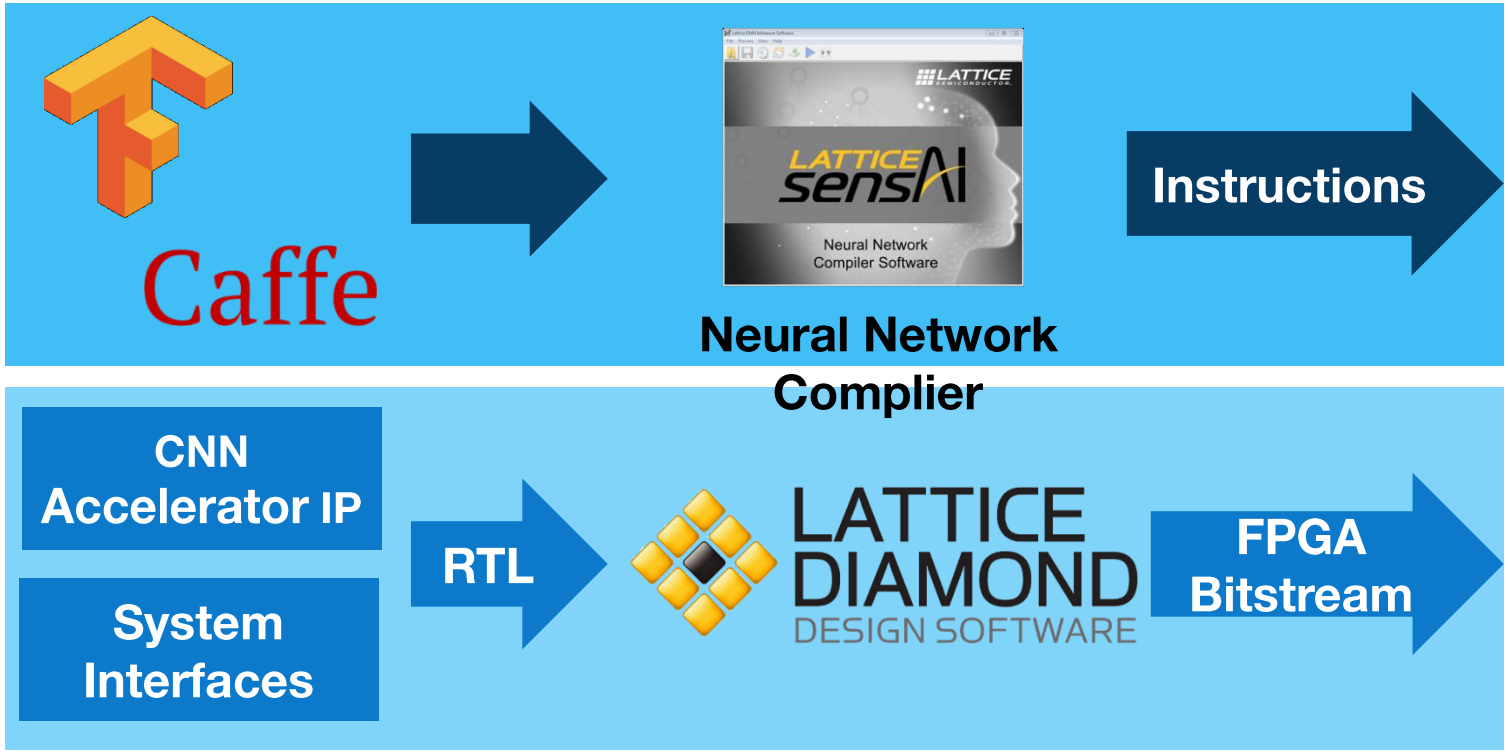


1 mW, 5.5 mm^2, 1 bit, ~$1

1 W, 100 mm^2, 8/16 bits, ~$10

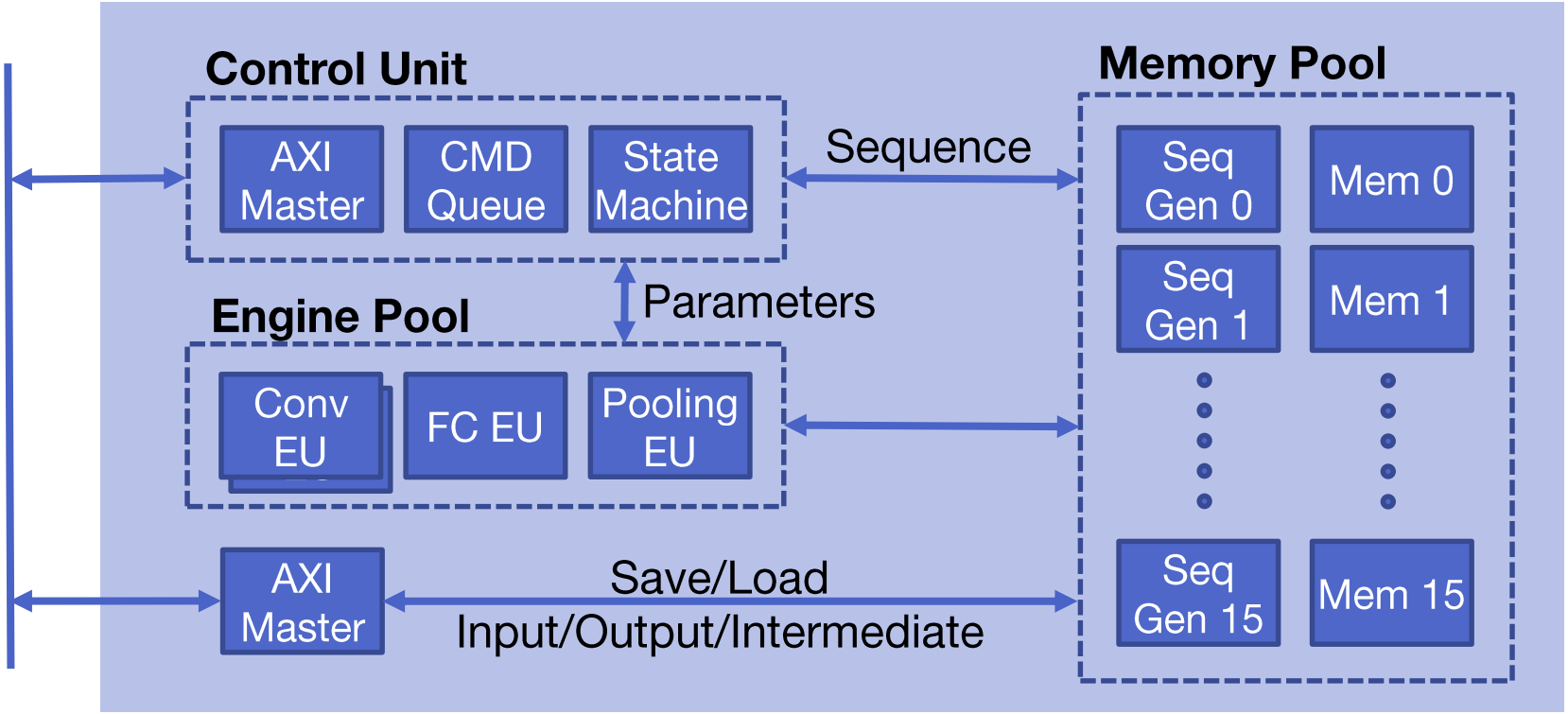


Delivering Edge CNN Acceleration in Lattice FPGA



Lattice FPGA

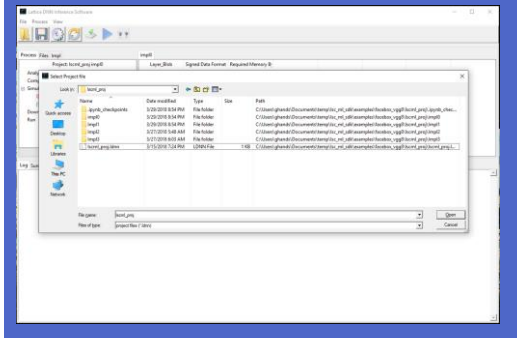
CNN Accelerator IP Architecture



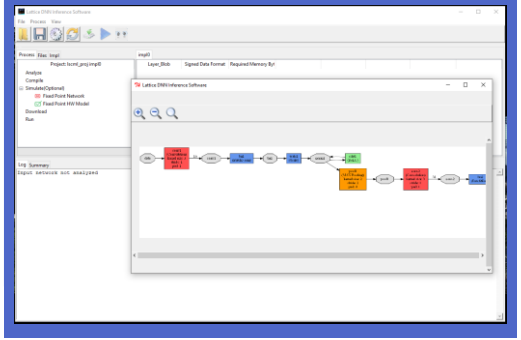
Translating Trained Neural Network Into Lattice CNN Accelerator Instructions

embedded VISION SUMMIT 2018

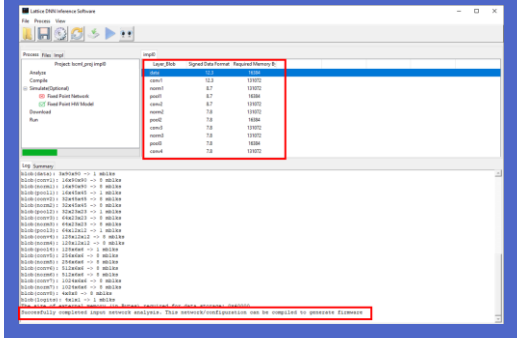
1. Load



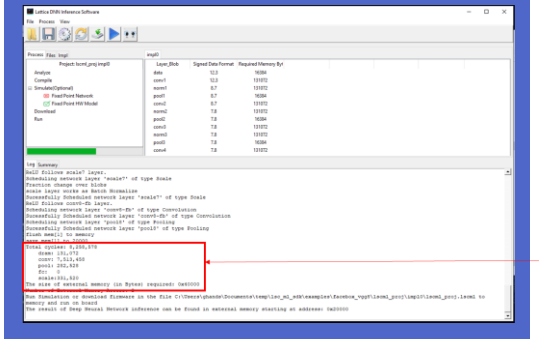
2. Review



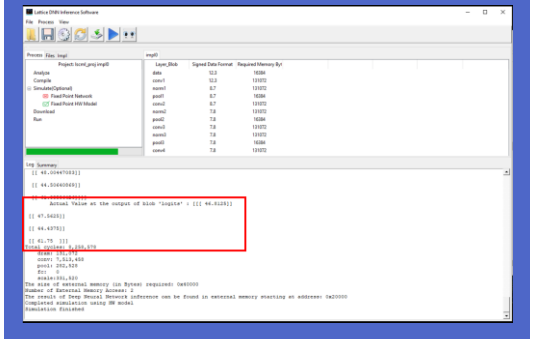
3. Analyze



4. Compile



5. Simulate



Edge AI – Complex Optimization

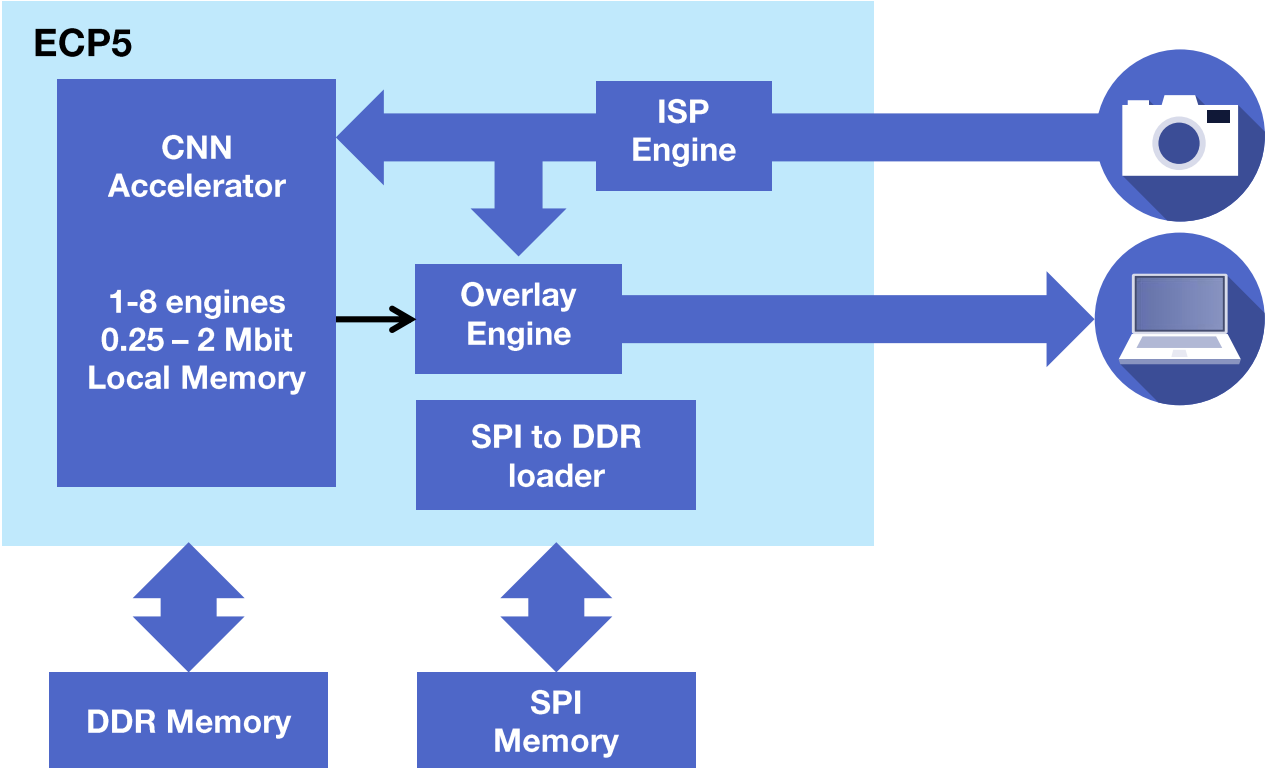
Design Factors / Attributes	Device		Network		
	# of Engines	Local Memory	Input Size	Number of Multipliers	Bit Widths
Power (W)					
Cost ($)					
Performance (fps)					
Accuracy (%)					
Small Object (% fov)					

Correlation Between Design Factors and Product Attributes

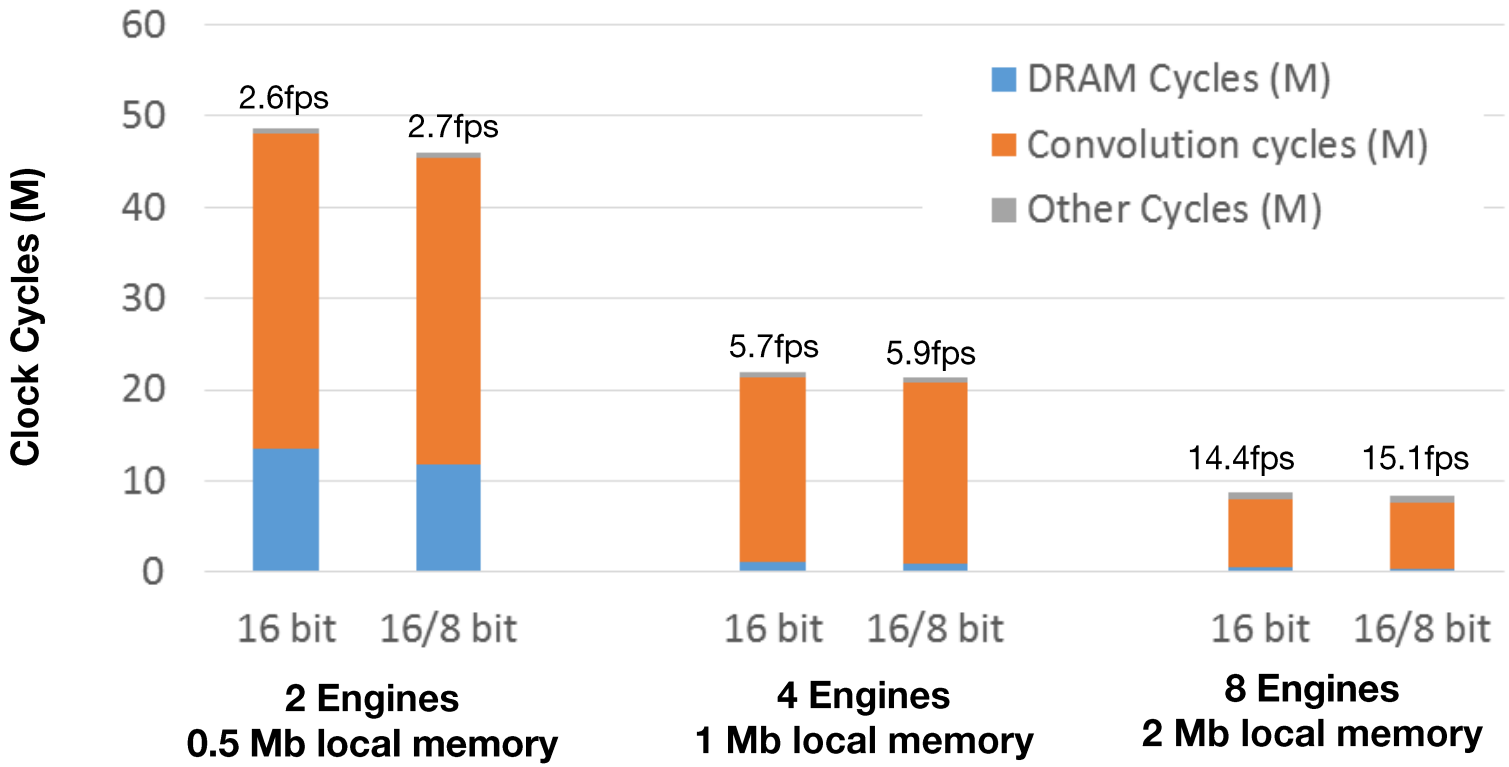
Examples for Illustration

		Architecture	Number of Multiplications	Input Size	Quantization
Face Tracking		Modified VGG8	256M	90x90	16 bit fixed
					8/16 bit fixed
Speed Sign Detect		Modified VGG8	146M	128x128	16 bit fixed
					8/16 bit fixed

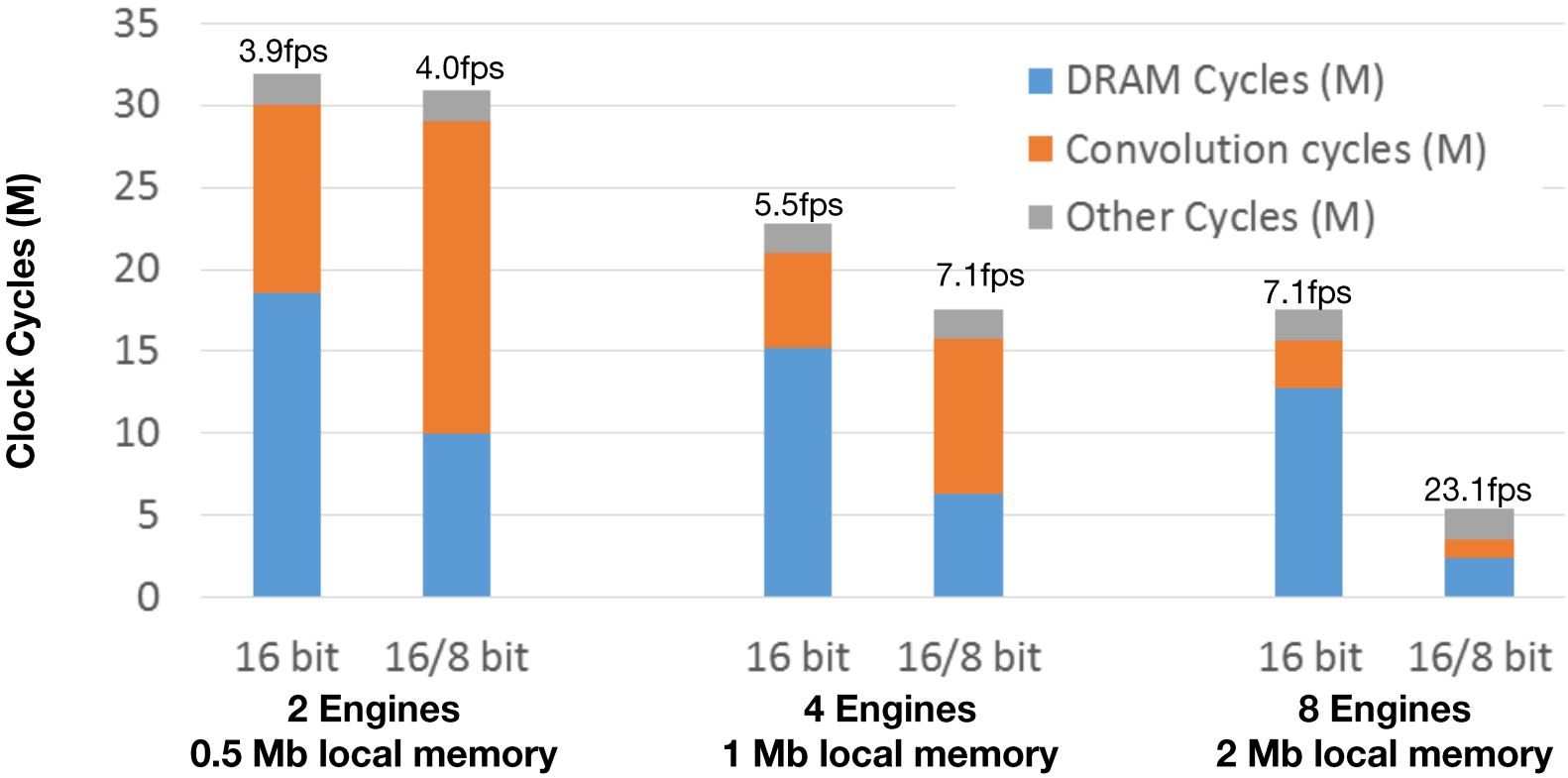
System Block Diagram



Face Tracking Implementations



Speed Sign Implementations



Bringing It Together

		Device Cost / Power / Performance		
Network	**Smallest Object**	**ECP5-25 Cost x0.25 0.5 W**	**ECP5-45 Cost x0.5 0.53 - 0.62 W**	**ECP5-85 Cost x1.0 0.58 – 0.8 W**
Face Tracking 16 bit	20 % of image height	2.6 fps	2.6 – 5.7 fps	2.6 – 14.4fps
Face Tracking 8/16 bit		2.7 fps	2.7 – 5.9 fps	2.6 – 15.1fps
Speed Sign Detect 16 bit	15% of image height	3.9 fps	3.9 - 5.5 fps	3.9 – 7.1 fps
Speed Sign Detect 8/16 bit		4.0 fps	4.0 – 7.1 fps	4.0 – 23.1 fps

Summary

- AI at the edge solves real world problems
- ECP5 sensAI Stack Components Provide Edge AI Building Bocks
  - Silicon, Soft IP, Tools, Development Boards & Reference Designs
- Configurable Engine Size and Bit widths Coupled with Multiple Devices Allows System Optimization
  - 0.5 – 0.8 W, 10x10 mm^2, < $10

Resources

- Please Visit www.latticesemi.com for More Information and Downloads
  - 3 ECP5 Based Reference Designs / Demonstrations -- Free
  - CNN Accelerator IP – Free Evaluation
  - NN Compiler – Free
  - Video Interface Board – Currently $199 Promotional Price

- Please Visit the Lattice Booth in the Showcase
  - 8+ Intelligence At The Edge Demonstrations