Why It's Critical to Have an Integrated Development Methodology for Edge AI

Sreepada V Hegade

Director, ML Software and Solutions

Lattice Semiconductor

LATTICE sensAI™

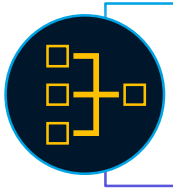
Disclaimer

Lattice makes no warranty, representation, or guarantee regarding the accuracy of information contained in this document or the suitability of its products for any particular purpose. All information herein is provided AS IS, with all faults, and all associated risk is the responsibility entirely of the recipient of the information. The information provided herein is for informational purposes only and may contain technical or other inaccuracies or omissions, and may be otherwise rendered inaccurate for many reasons, and Lattice assumes no obligation to update or otherwise correct or revise this information. Products sold by Lattice have been subject to limited testing and it is the responsibility of a buyer of Lattice products to independently determine the suitability of any products and to test and verify the same. Lattice products and services are not designed, manufactured, or tested for use in life or safety critical systems, hazardous environments, or any other environments requiring fail-safe performance, including any application in which the failure of the product or service could lead to death, personal injury, severe property damage or environmental harm (collectively, “high-risk uses”). Further, a buyer must take prudent steps to protect against product and service failures, including providing appropriate redundancies, fail-safe features, and/or shut-down mechanisms. Lattice expressly disclaims any express or implied warranty of fitness of the products or services for high-risk uses. Lattice Semiconductor Corporation, Lattice Semiconductor (& design) and specific product designations are either registered trademarks or trademarks of Lattice Semiconductor Corporation or its subsidiaries in the United States and/or other countries.

Edge AI Advancements

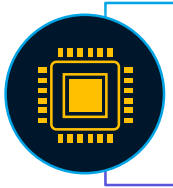
- Widespread Adoption
  - Across multiple market segments (Consumer, Industrial, Automotive, Healthcare)
- Technological Advancements
  - Compact models
  - Tiny inference devices
- Realized Benefits
  - Lower cost, power, latency
  - Enhanced security, reliability

Challenges



Complexity

How to find right trade-off between performance, power, cost and other KPIs ?



Adaptability

How can I support wide array of sensors ?

Sustainability

How do I update solution with latest innovations ?

FPGA Value Proposition

Predictable

Provide low and deterministic latency essential for many edge applications

Flexible

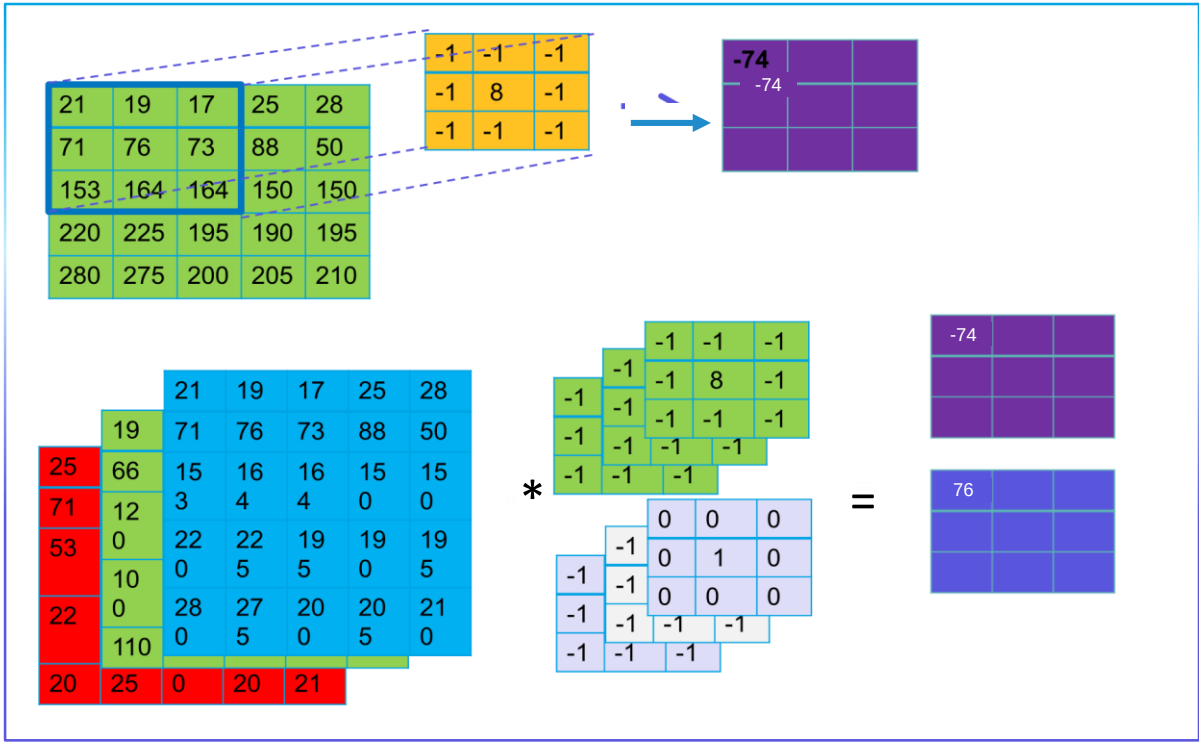
Same solution can be targeted to different devices with trade-off on power, performance and cost

Scalable

Helps to address scalability issues like integration with multiple sensors of different types

Low and Deterministic Latency

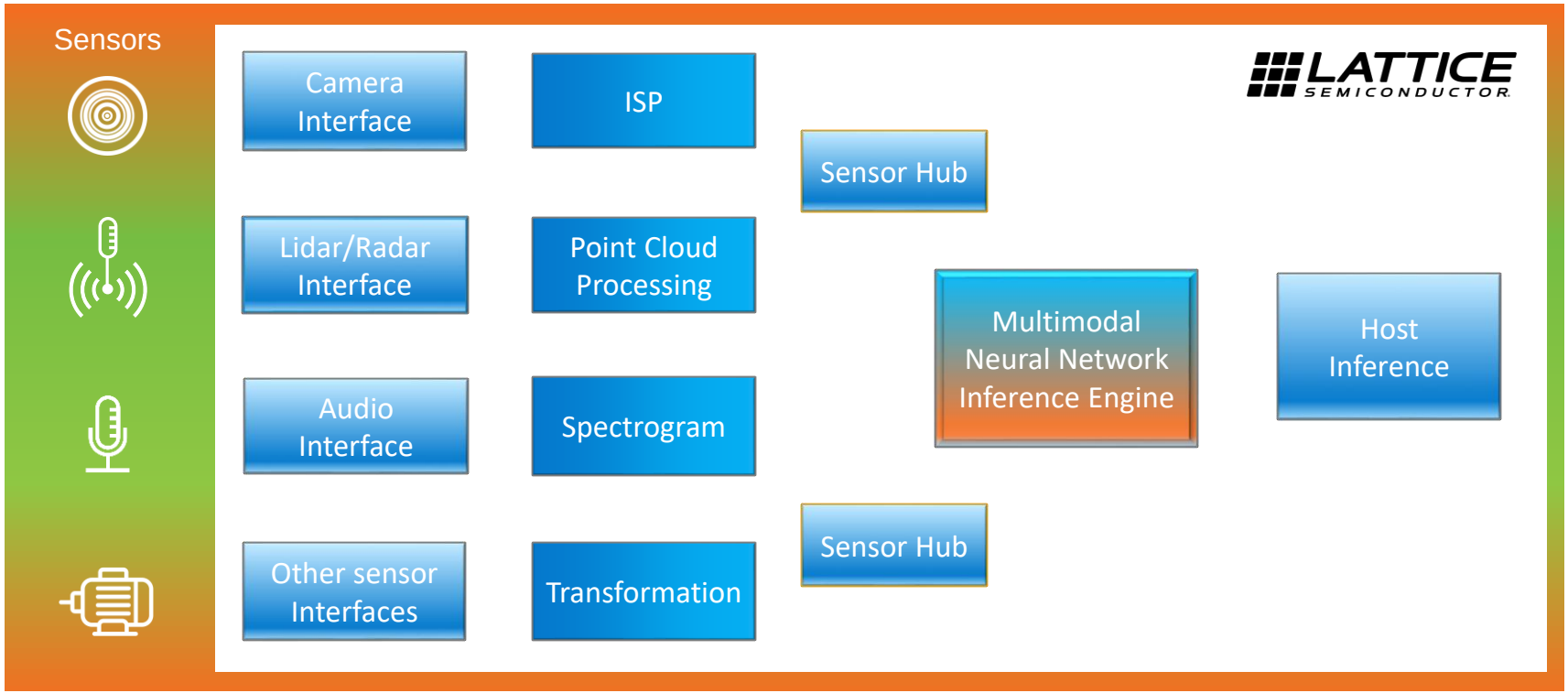
- Time Complexity of Convolution
  - Image is w x h
  - Filter is k x k
  - Each output requires $O(k^2)$ operations
  - Total time complexity:
    - $O(whk^2)$
  - Number of output entries:
    - (w+k-1)(h+k-1)



Flexibility

	Lattice iCE40 UltraPlus™	Lattice CrossLink™-NX 17	Lattice CrossLink™-NX 40	Lattice CrossLink™NX 33	Lattice CertusPro™-NX	Lattice Avant™
Footprint (mm)	2.15 x 2.55	3.7 x 4.1	6 x 6	3.1 x 7.4	9 x 9	10 x 10 – 27 x 27
# of DSPs (18 x 18)	8	24	56	64	156	320 - 1800
# of 8 x 8 multipliers per DSP	1x	2x	2x	2x	2x	4x
Distributed Memory (kbits)	120	432	1512	1512	3400	-
SPRAM (kbits)	1024	2560	1024	2560	3072	-
# of Cores	Compact	1	2	2	6	12 - 72
Speed (MHz)	40	150	150	-	150	650
Power (W)	0.02	0.05	0.2	0.150	0.45	1

Scalability



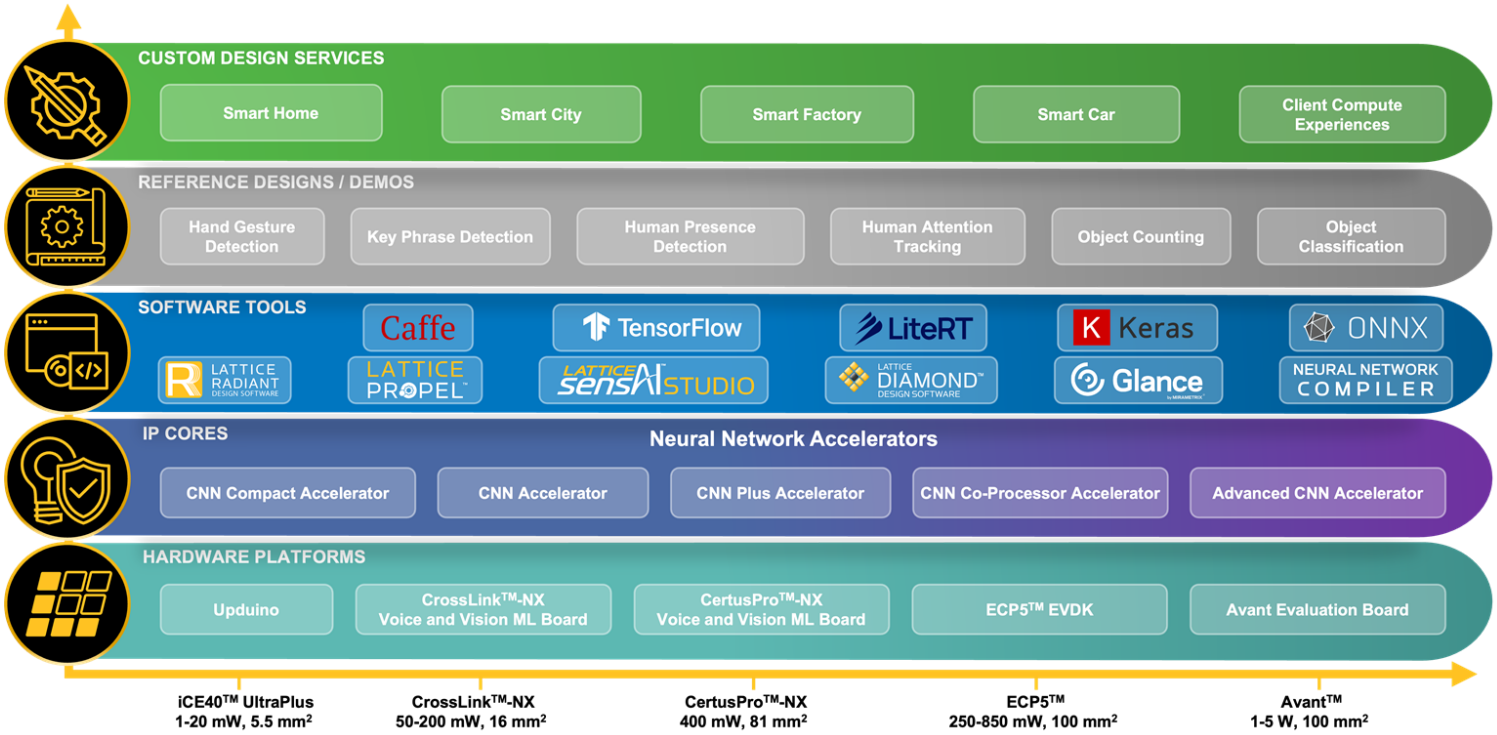
Key Observations

Implementation of efficient edge AI solution requires **hardware aware model development and training**

Configurable and flexible hardware is essential to achieve optimal solution for a given application

Integrated design **tools and methodology critical to successfully co-design model and hardware** for edge AI

Lattice sensAI Solution



CES INNOVATION AWARDS 2022 HONOREE

AI ARTIFICIAL INTELLIGENCE EXCELLENCE AWARD 2023

AI ARTIFICIAL INTELLIGENCE EXCELLENCE AWARD 2025

EDGE AWARDS Honoree

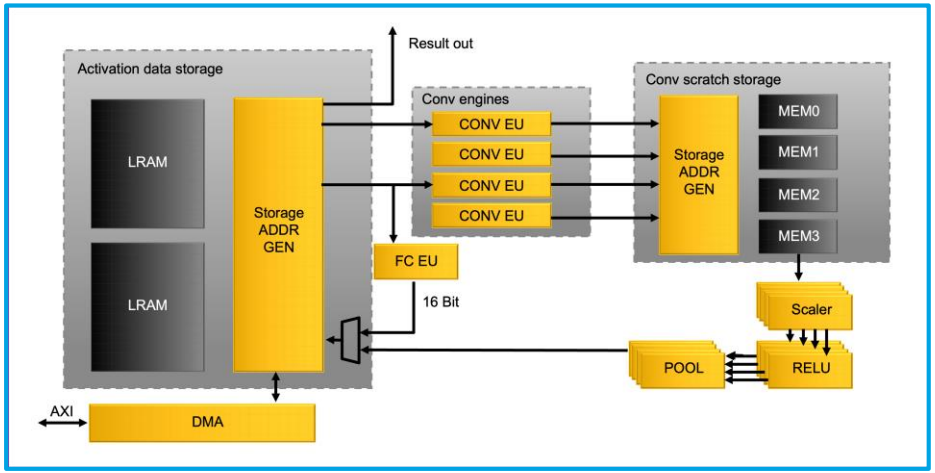
电子发烧友 www.elecfans.com

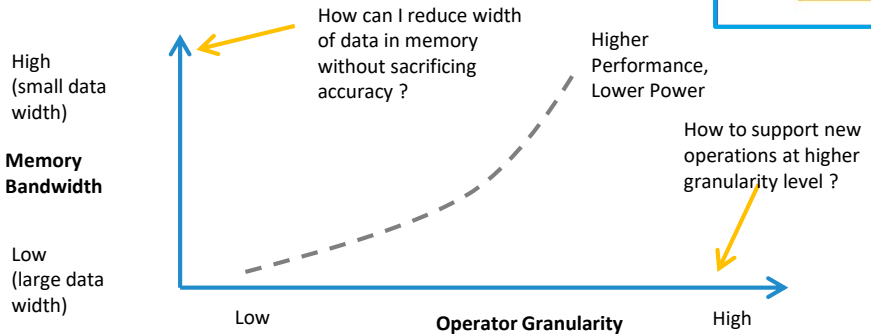
2021第二届中国人工智能卓越创新奖

最具创新价值产品

Customizable ML Accelerator

- Reduce memory access
- Customize engine with absolutely required operators
- Configure data width and precision

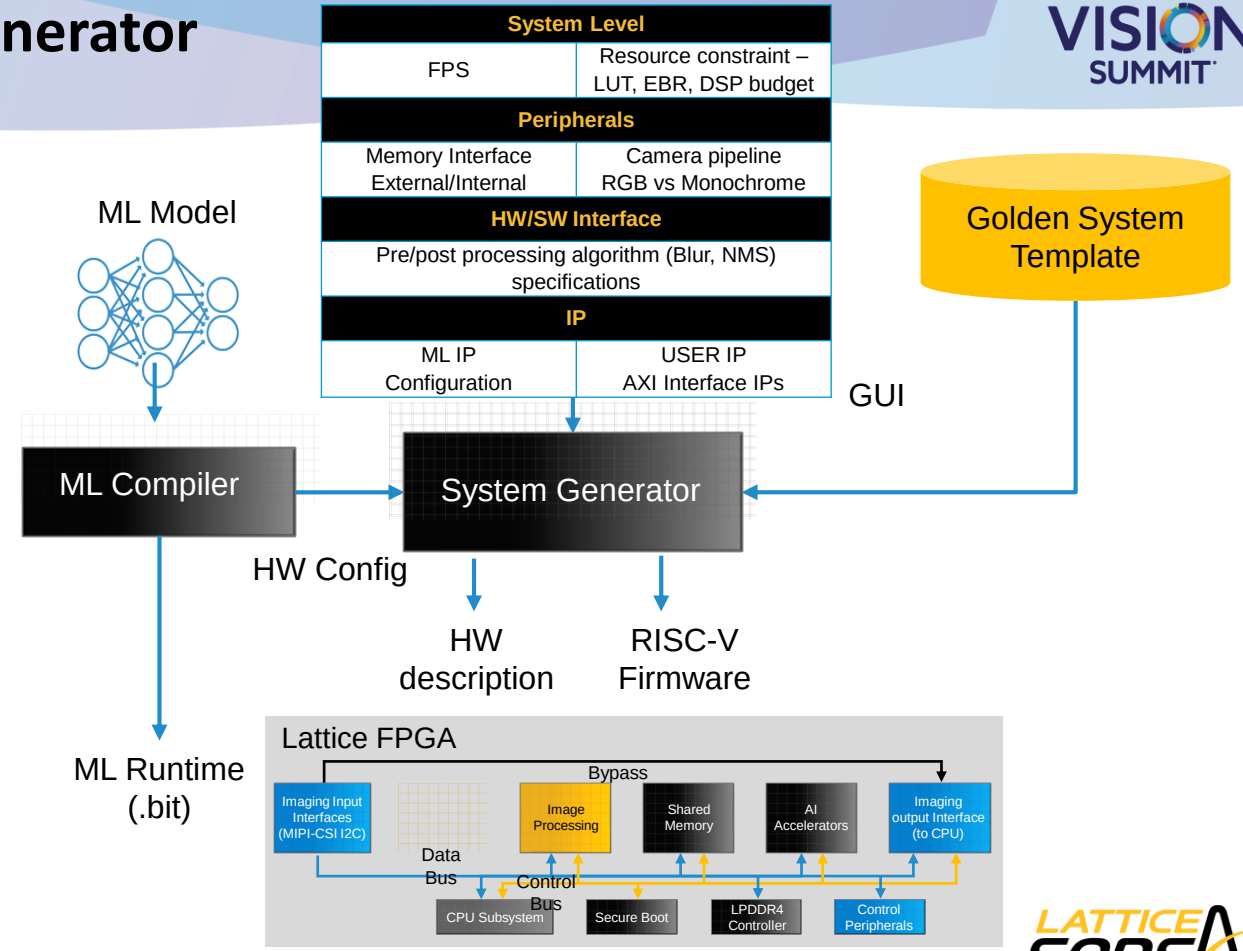




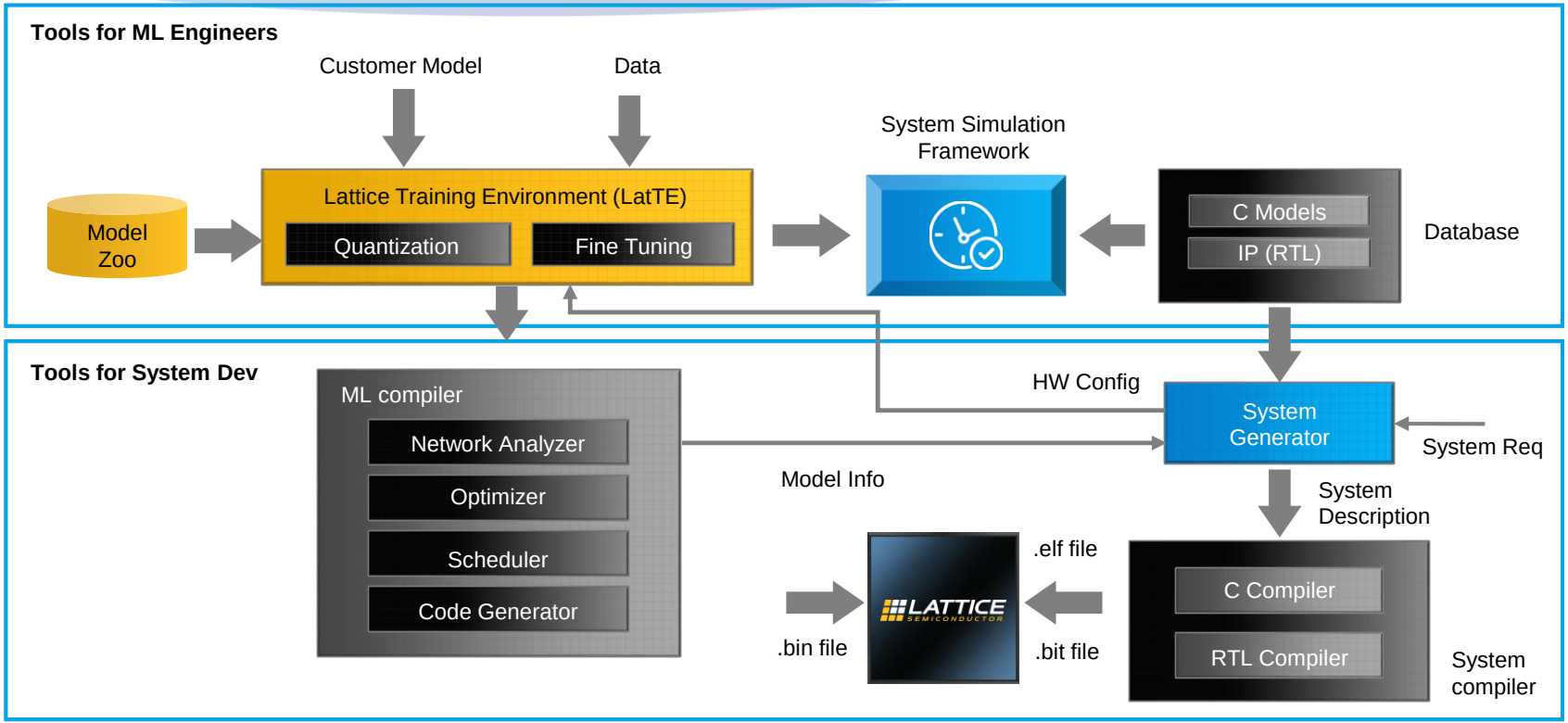
Zero Code System Generator

Configure ML System by extracting parameters from trained neural networks

Tool and flow that makes system easy to use for ML engineers

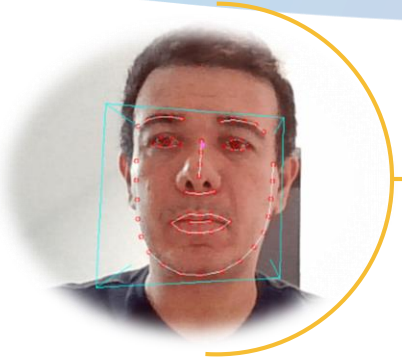


Integrated Tool and Flow



Case Studies

Face Landmark Detection



- Variation of MobileNet V2 Model
- Model input resolution 96x96
- Detects 23 landmarks on the face

Variation	Quantization	Landmark Pixel MAE
1	float32	1.09952
2	Fixed Step Size Quantization (int8)	1.28866
3	Learned Step Size Quantization (int8)	1.11622

Learned Step Size Quantization (LSQ) Helps Improve Accuracy

Barcode Detection

- YOLOv5
- Input Resolution 320x240 (QVGA)
- Detect barcode(s) on packages that are moving on a conveyer belt

Network Type	YOLOv5
mAP with fixed quantization	77.43%
mAP with Learned Step Quantization	96.79%
Few layers quantized to 4-bit data	89.73%

Learned Step Quantization helped reduce data width, thereby saving memory utilization

Summary

Lattice low power FPGAs are perfect for edge AI model inferences

Production-proven solutions and tools offer a good trade-off when developing applications

Hardware can be configured for optimal inference of a given network topology

Reference designs and models are a great a starting point for building your application

Thank You!

Visit Lattice and see our technologies at Booth #416

More information can be found at

https://www.latticesemi.com/en/Solutions/Solutions/SolutionsDetails02/sensAI