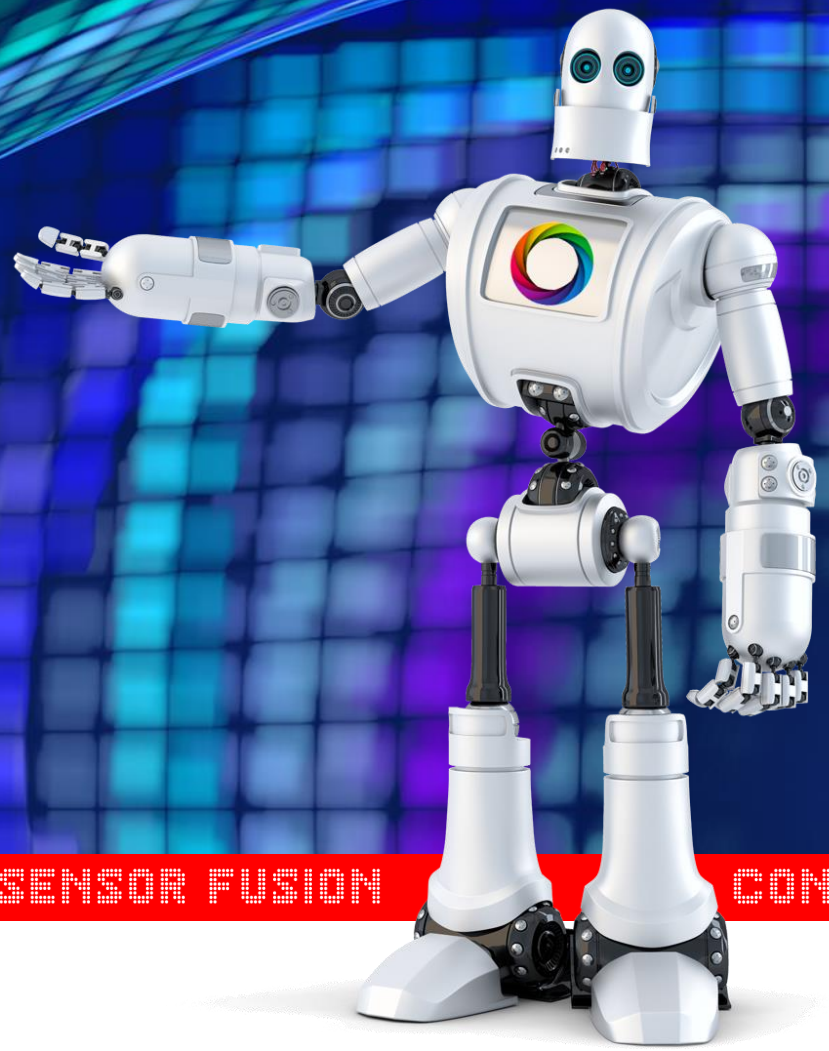
reVISION™

Responsive and Reconfigurable Vision Systems



MACHINE LEARNING | COMPUTER VISION | SENSOR FUSION | CONNECTIVITY

XILINX
ALL PROGRAMMABLE™

Accelerating your Embedded Vision / Machine Learning design with the reVISION Stack

Giles Peckham, Xilinx

Xilinx Foundation at the Edge

Vision Customers Using Xilinx

>80
ADAS Models From 23 Makers

>80
ProAV & Broadcast Suppliers

>60
Smart Camera & Visualization Suppliers

>50
Industrial Vision Equipment Makers

>10
Medical Diagnostic Suppliers

>5
VR/AR Equipment Makers

>5
Drone Suppliers

>40 Customers Designing with Machine Learning

© Copyright 2017 Xilinx

XILINX ALL PROGRAMMABLE.

Machine Learning: From the Edge to the Cloud

	Edge Resident Apps	Hybrid Solutions	Cloud Hosted Apps
Consumer/Entertainment/Retail	Personal VR/Gaming Smart Displays	Personal Assistants	Ad Targeting and E-Commerce
Transportation/Infrastructure	Autonomous Cars &Trucks	Transportation & Grid Control	Traffic & Network Analytics
Enterprise Operations	Delivery Drones, Warehouse Robots	Cyber Security	Sales, Marketing & Customer Service
Oil & Gas/Agriculture	Field Drones & Robots	Climate, Water, Energy & Flow Control	Field Sensor Data Analytics
Industrial/Military	Robots/Cobots, UAV, Inspection	Factory Control & Surveillance	Factory & Operations Analytics
Medical/Healthcare	Medical Imaging & Surgical Robots	Medical Diagnostics	Clinical Analytics & Recommendations

Source: Machine Learning Landscape from Moor Insights & Strategy

© Copyright 2017 Xilinx

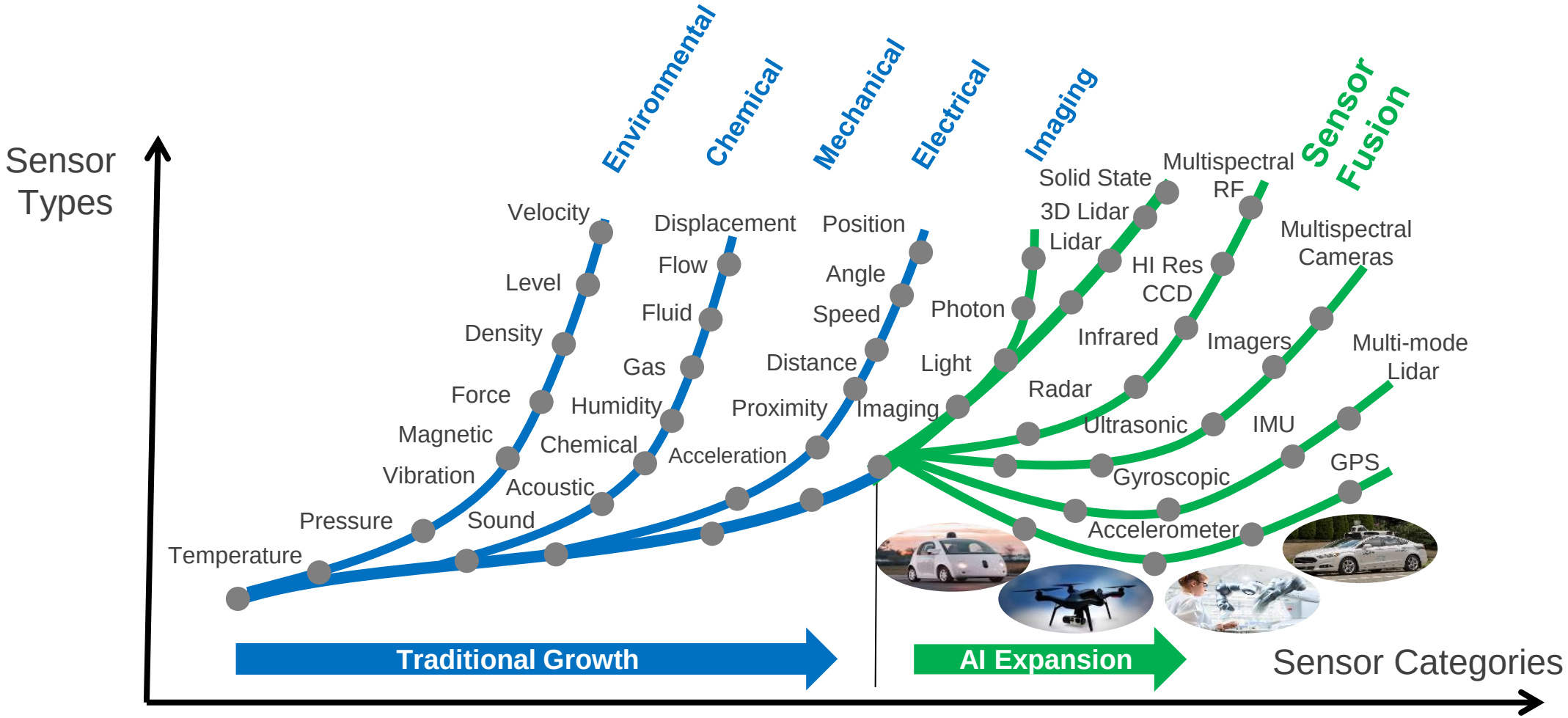
XILINX ALL PROGRAMMABLE.

Applications: Wide Range of Rapidly Changing Vision Guided Systems

Embedded Vision Systems →	Vision Guided Autonomous Systems
▪ Factory Robotics	▪ Vision Guided ‘Cobots’
▪ Camera Equipped Aircraft	▪ ‘Sense and Avoid’ & Autonomous Drones
▪ Physical Displays and HMI	▪ Augmented Reality and HUDs
▪ Forward Auto Camera	▪ Autonomous Vehicles
▪ Video Security Cams	▪ Automated Surveillance
▪ Medical Imaging and Human Eye	▪ Automated Medical Diagnostics

© Copyright 2017 Xilinx

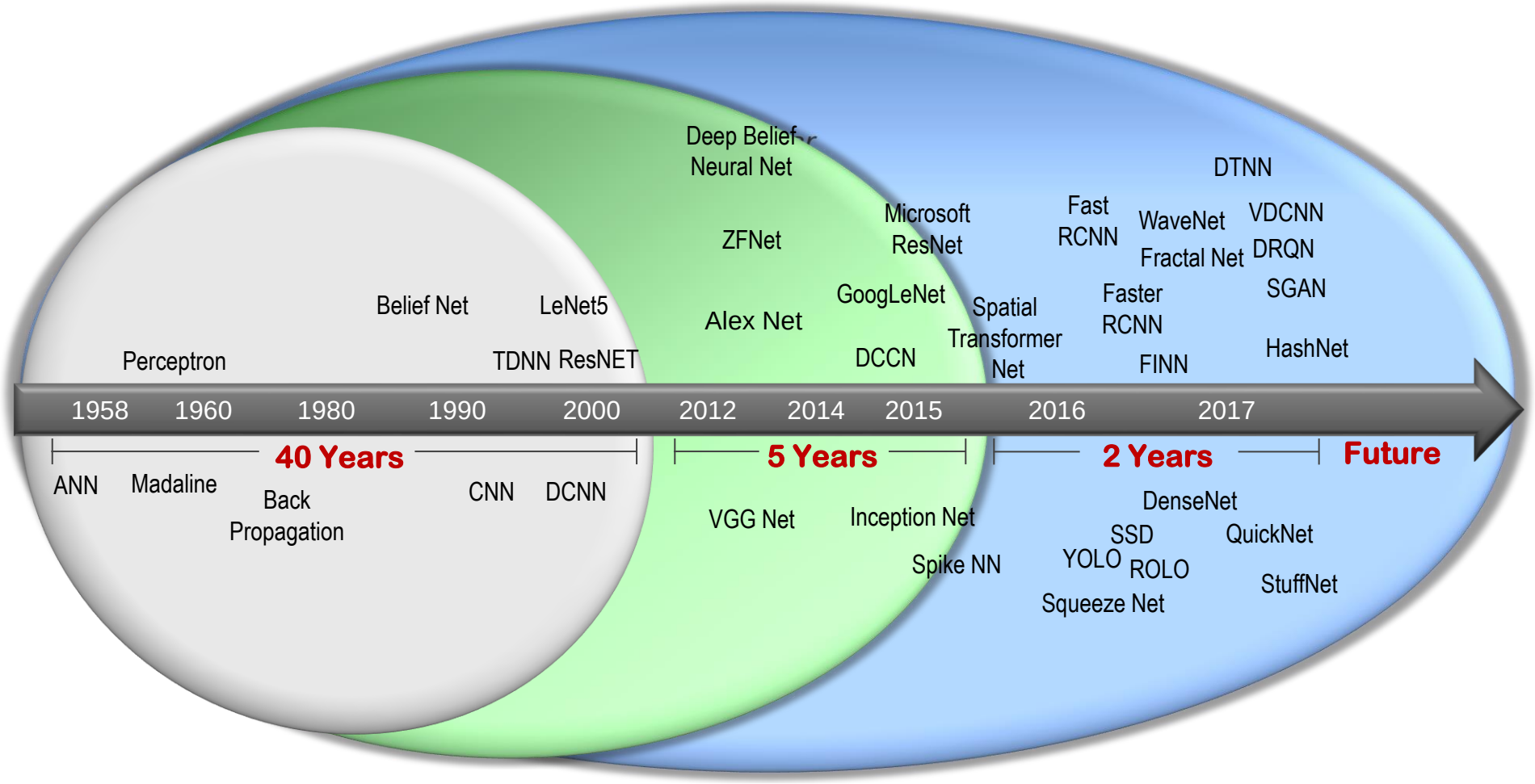
Must Keep Up with Sensor Fusion Evolution



© Copyright 2017 Xilinx

XILINX ALL PROGRAMMABLE.

Must Keep Up with Neural Network Evolution



Floating Point → **8-bit to 1-bit and Variable Precision Inference**

© Copyright 2017 Xilinx

XILINX ALL PROGRAMMABLE.

Xilinx Unique Application Advantages

Optimized from Sensors to Inference & Control

Reconfigurable for Latest Networks & Sensors

Any-to-Any Connectivity

© Copyright 2017 Xilinx

XILINX ALL PROGRAMMABLE.

Xilinx Unique Application Advantages

Responsive

Optimized from Sensors to <8-bit Inference & Control

Reconfigurable

Reconfigurable for Latest Networks & Sensors

Connected

Any-to-Any Connectivity

Barrier to Broad Adoption:

Software Defined Programming, Libraries and Frameworks

© Copyright 2017 Xilinx

XILINX ALL PROGRAMMABLE.

reVISION Stack

OpenVX™ Vision Acceleration | Caffe

Application Development

OpenCV | SDSoC™ Environment

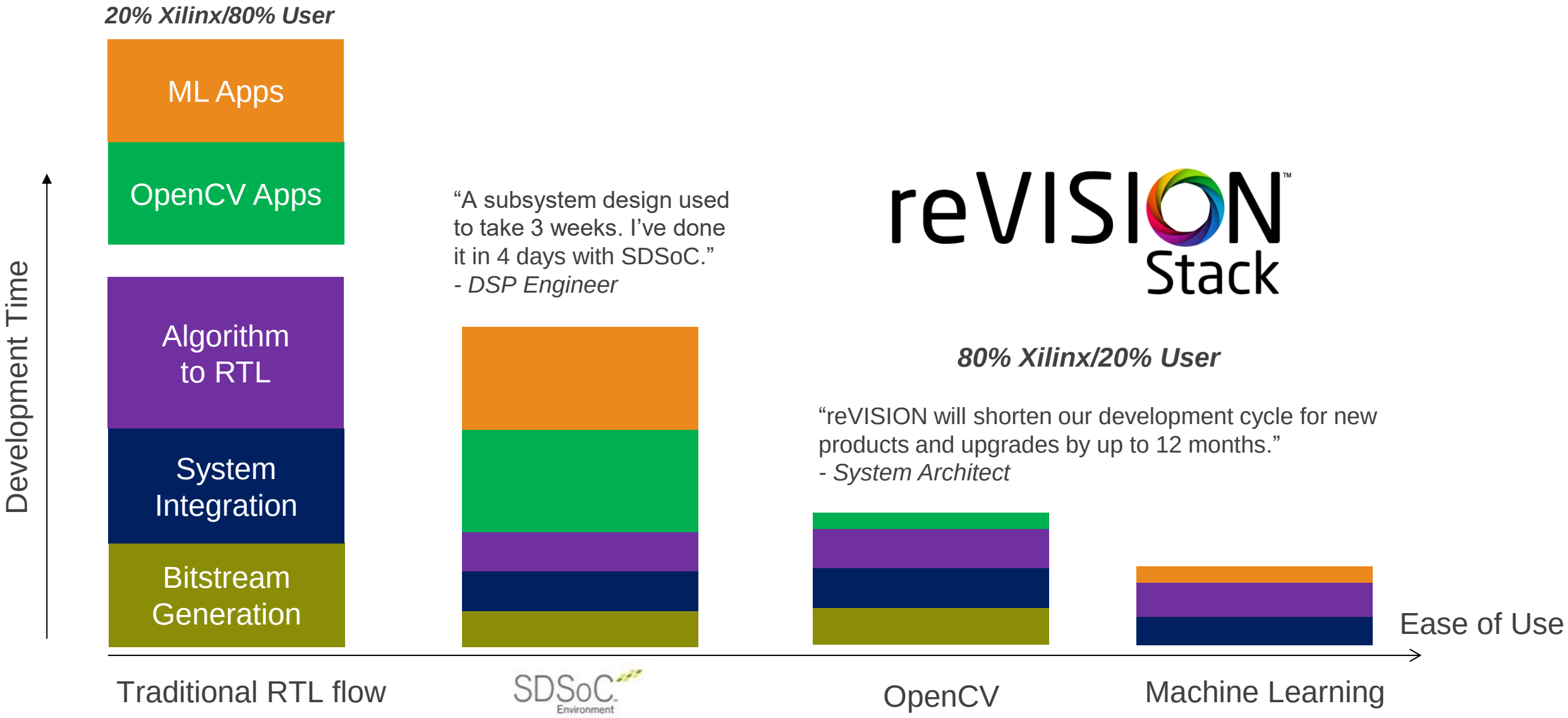
DNN
CNN
GoogLeNet
SSD
FCN …

Algorithm Development

Platform Development

© Copyright 2017 Xilinx

Removing the Barrier to Broad Adoption: reVISION Stack



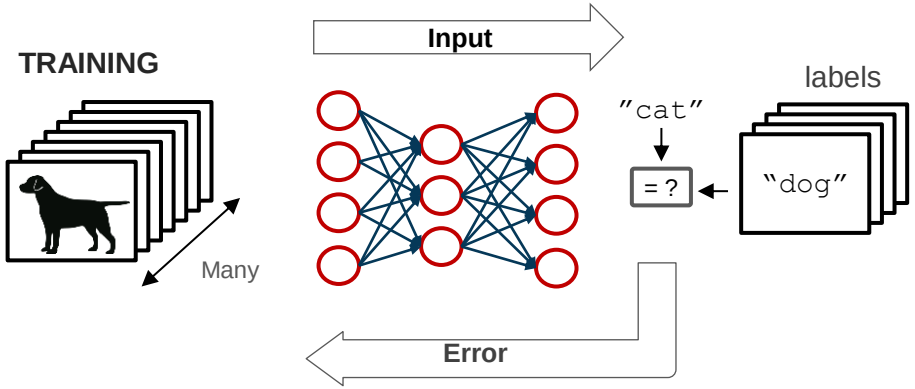
© Copyright 2017 Xilinx

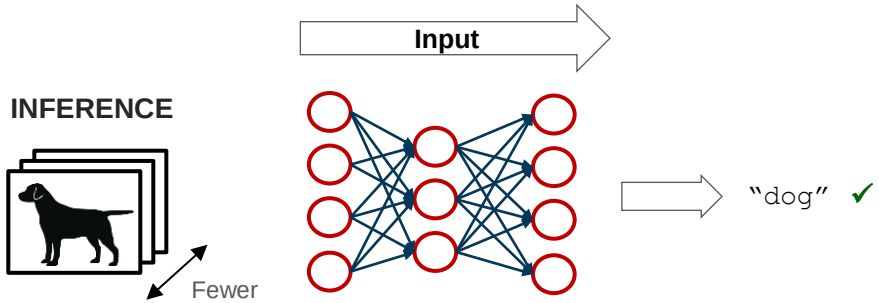
XILINX ALL PROGRAMMABLE.

Machine Learning Inference

© Copyright 2017 Xilinx

The Divergence of Training and Inference in Machine Learning





Training: Process for machine to "learn" and optimize model from data

Inference: Using trained models to predict/estimate outcomes from new observations in efficient deployments

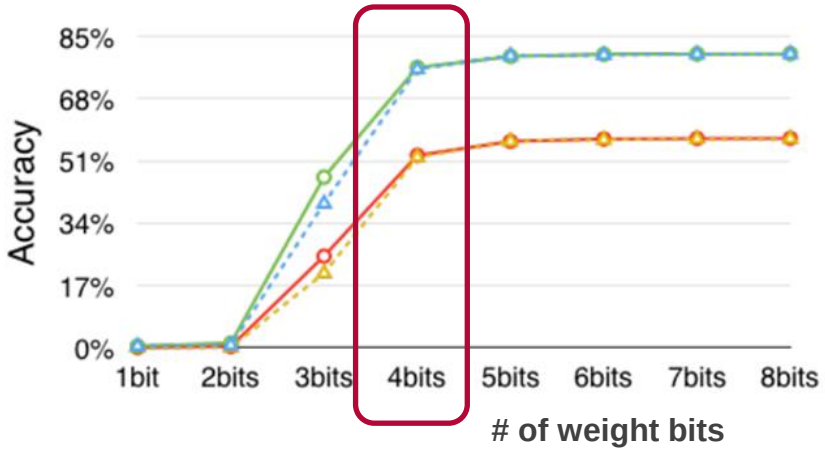
Inference now 8 bit and below for maximum efficiency

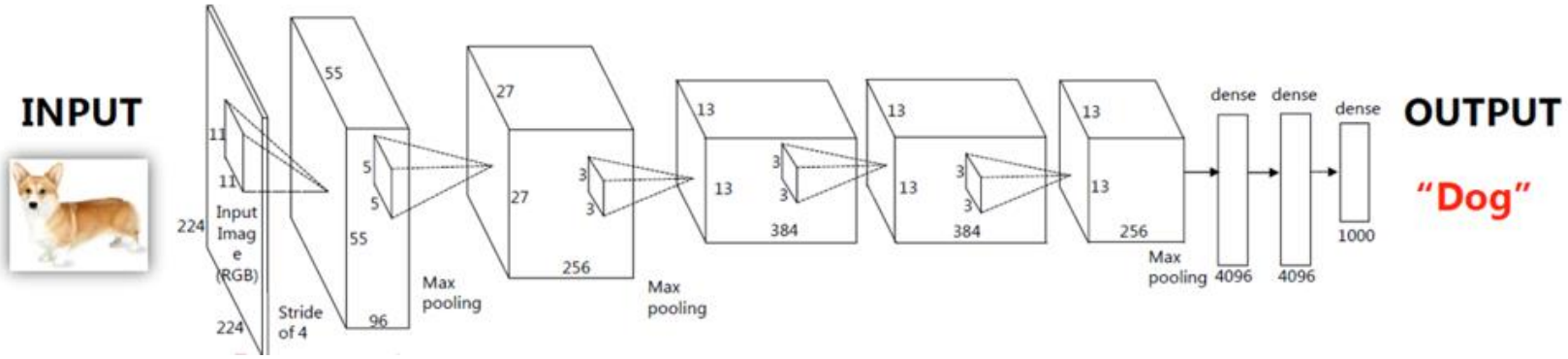
Top-5 Accuracy

	FP-32	FIXED-16 (INT16)	FIXED-8 (INT8)	Difference vs FP32
VGG-16	86.6%	86.6%	86.4%	(0.2%)
GoogLeNet	88.6%	88.5%	85.7%	(2.9%)
SqueezeNet	81.4%	81.4%	80.3%	(1.1%)

© Copyright 2017 Xilinx

Inference Precisions Moving to Lower and Variable Precision



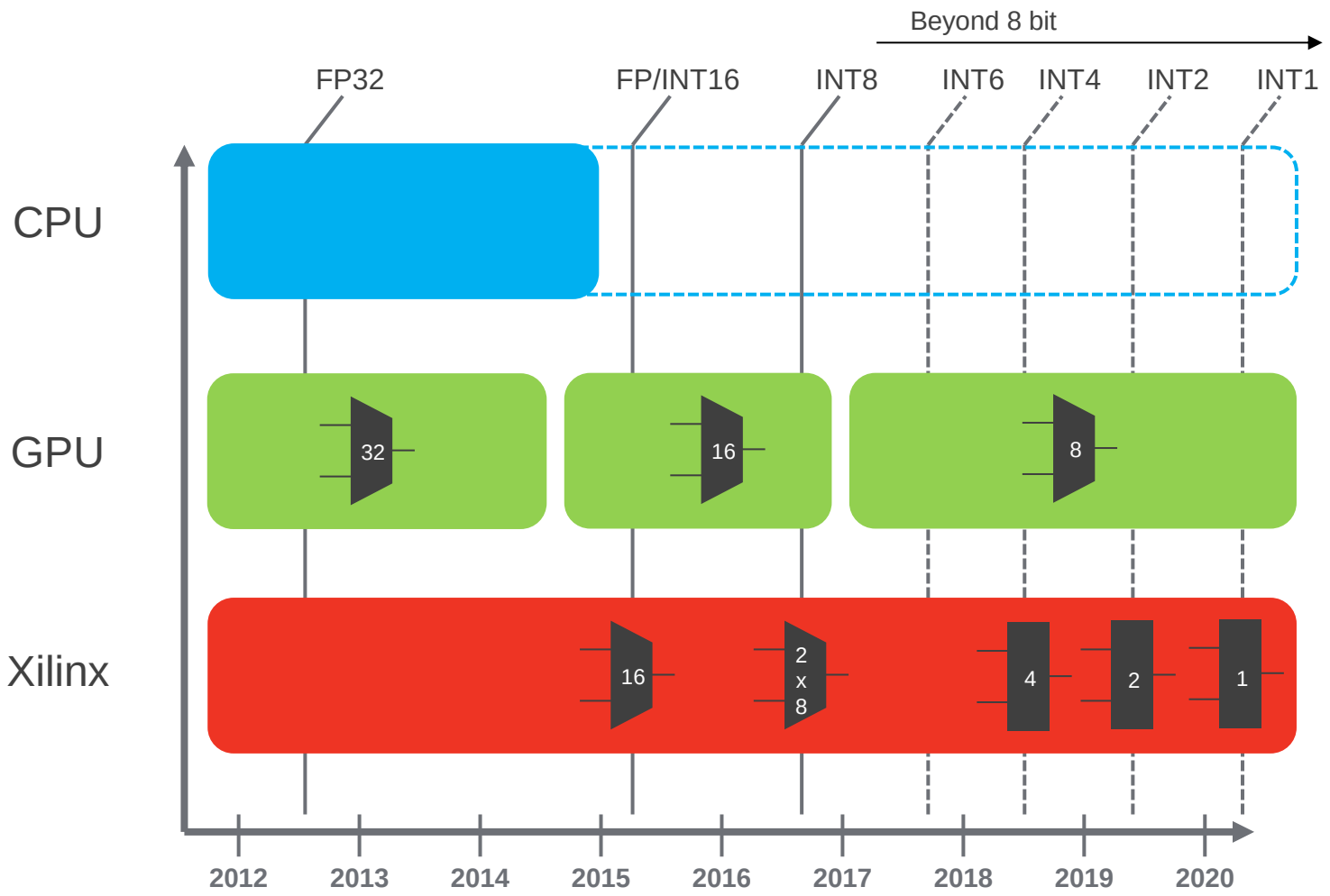


Citation: https://arxiv.org/pdf/1510.00149.pdf

© Copyright 2017 Xilinx

XILINX ALL PROGRAMMABLE.

Xilinx: Future Proof Architecture for Any Precisions



Limited to 32 bit operations

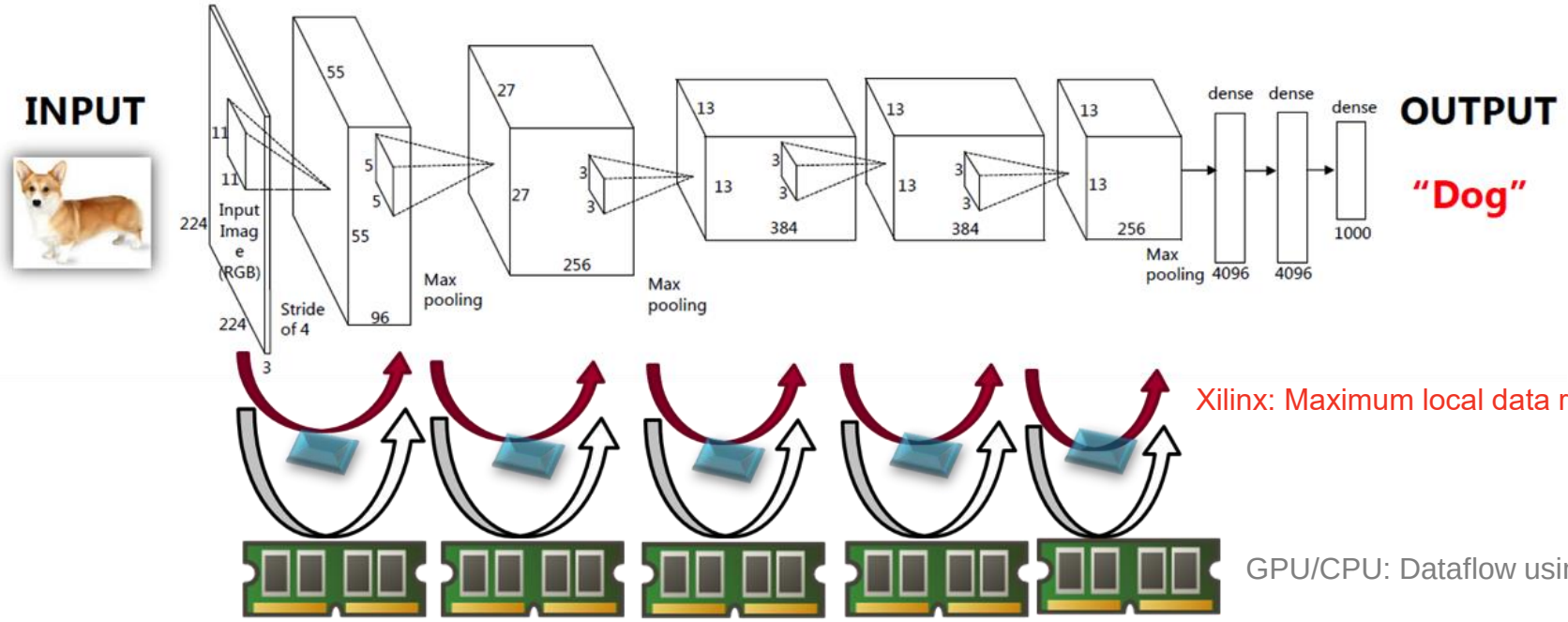
New devices required to support change in precision efficiently

Reconfigurable to scale and optimize for different precisions

© Copyright 2017 Xilinx

XILINX ALL PROGRAMMABLE.

Low Latency Inference by Layer to Layer Dataflow On Chip



Xilinx: Maximum local data reuse "layer merging" between layers

GPU/CPU: Dataflow using Off chip memory

On-chip Memory	Nvidia Tegra X1 (GPU L2 Cache)	Xilinx ZU7 (BRAM + URAM)
	2 Mb	38 Mb

Up to 19x More On-chip Memory than SoCs and eGPUs

*Nvidia TX1 spec: http://wccftech.com/nvidia-tegra-x1-super-chip-announced-ces-2015-features-maxwell-core-architecture-256-cuda-cores/

© Copyright 2017 Xilinx

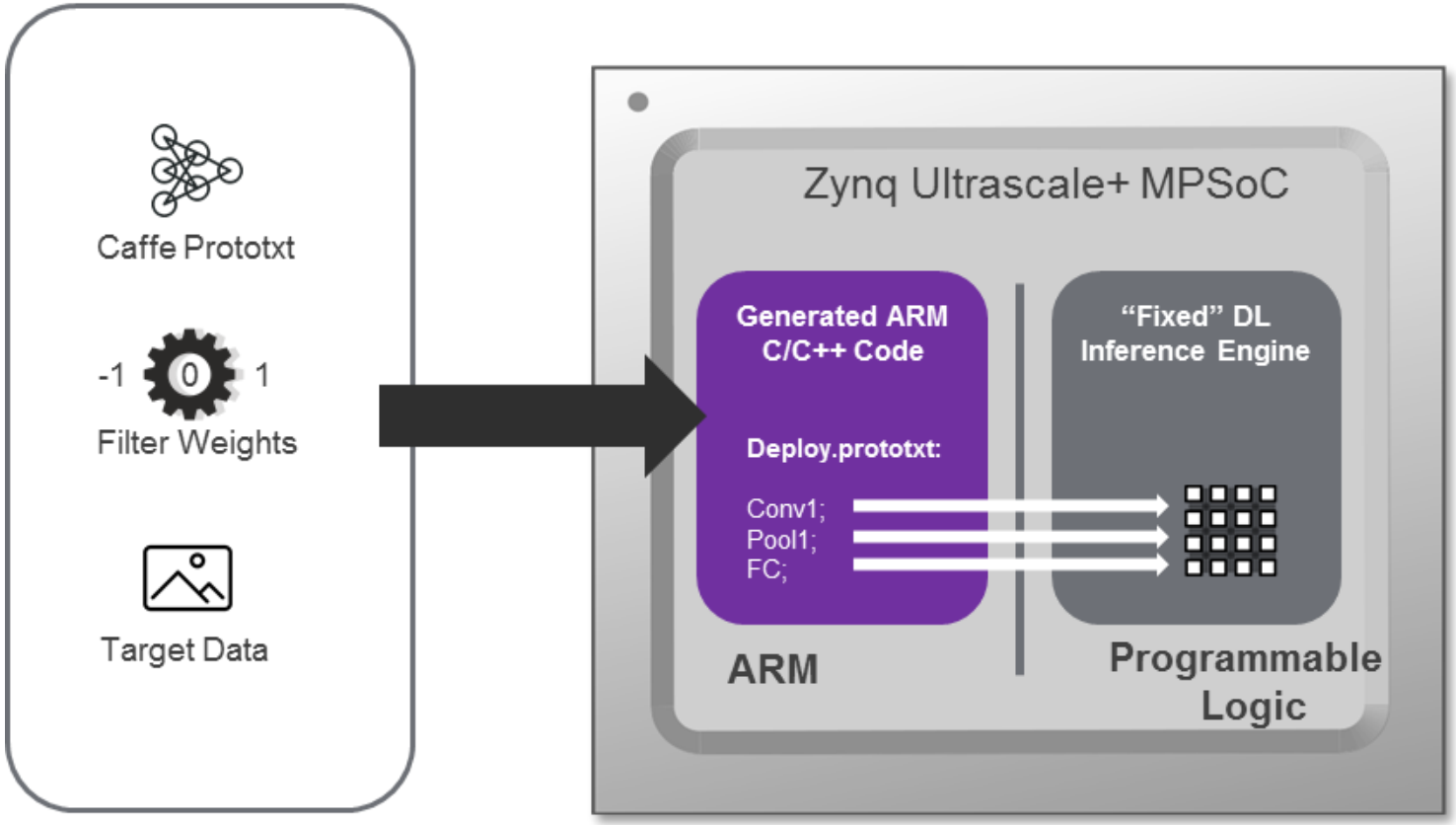
XILINX ALL PROGRAMMABLE.

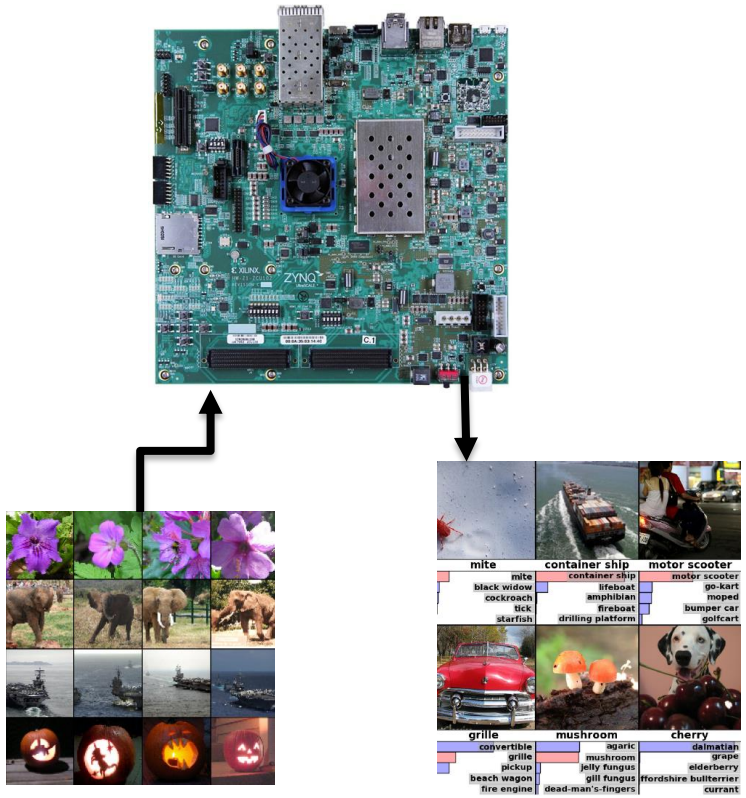
xFdnn: Direct Deep Learning Inference from Caffe

1 Import .prototxt and trained weights

2 Call prototxt runtime API in your application

3 Cross-compile for Cortex-A53 and run on a board





Compiles only ARM software code in minutes. No hardware compilation

© Copyright 2017 Xilinx

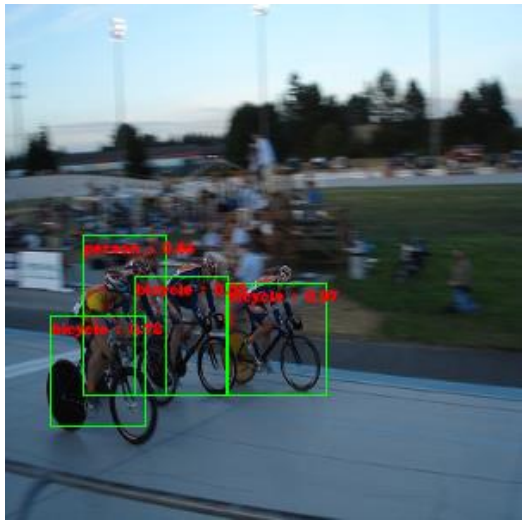
XILINX ALL PROGRAMMABLE

Deep Learning Design Examples

		Mar 2017	Roadmap
GoogLeNet @ batch = 1	Images/s	**115**	**370**
	Power (W)	6.0	7.0
	Images/s/watt	**19.2**	**53.0**

		Mar 2017	Roadmap
SSD @ batch = 1	Images/s	**6.3**	Coming Soon
	Power (W)	6.5	Coming Soon
	Images/s/watt	**1.0**	Coming Soon

		Mar 2017	Roadmap
FCN-AlexNet @ batch = 1	Images/s	**7.0**	Coming Soon
	Power (W)	6.5	Coming Soon
	Images/s/watt	**1.1**	Coming Soon



© Copyright 2017 Xilinx

XILINX ALL PROGRAMMABLE.

Performance: More efficient than Nvidia Tegra X1

6x
Images/sec/watt

Machine Learning *Inference*

Xilinx Benchmark

42x
Frames/sec/watt

Computer Vision

Xilinx Benchmark

		Xilinx ZU9	Xilinx ZU5	Nvidia TX1
GoogLeNet @ batch = 1	Images/s	**370.0**	**155.0**	70
	Power (W)	**7.0**	**4.5**	7.9
	Images/s/watt	**53.0**	**34.5**	8.9

		Xilinx ZU9	Xilinx ZU5	nVidia TX1
CV:: StereoLBM @1080p	Frames/s	**700**	**296**	28
	Power (W)	**4.8**	**3.3**	7.9
	Frames/s/watt	**145.8**	**89.7**	3.5

		Xilinx ZU9	Xilinx ZU5	nVidia TX1
CV:: LK Dense Optical Flow @720p	Frames/s	**170**	**73**	7
	Power (W)	**4.8**	**3.3**	7.9
	Frames/s/watt	**35.4**	**22.1**	0.9

- nVidia GoogLeNet performance: https://devblogs.nvidia.com/parallelforall/jetpack-doubles-jetson-tx1-deep-learning-inference/
- ML based on Xilinx GoogleNet performance roadmap in 2H2017: 180 img/s in May 2017, 370 img/s in 2H17
- LK Dense Optical Flow using pyramid = 5, iteration = 5
- All benchmarks utilize as much resources as possible on GPU (~99%) and programmable logic (~70%)

© Copyright 2017 Xilinx

Latency: Xilinx Provides Fastest Response Time for Any Batch Size

1/5

Latency (ms)

Real Time Applications
Latency

Xilinx Benchmark

GoogLeNet @ batch = 1		**Xilinx ZU9**	Xilinx ZU5	Nvidia TX1
	Images/s	**370.0**	155.0	70
	Latency (ms)	**2.7**	6.4	14.2

GoogLeNet @ batch = 8		**Xilinx ZU9**	Xilinx ZU5	Nvidia TX1
	Images/s	**370.0**	155.0	163
	Latency (ms)	**2.7**	6.4	49.0

For large batch, Nvidia latency increases significantly

*nVidia GoogLeNet performance: https://devblogs.nvidia.com/parallelforall/jetpack-doubles-jetson-tx1-deep-learning-inference/

© Copyright 2017 Xilinx

XILINX ALL PROGRAMMABLE.

- ✓ Removing the Barriers for Expansion into a Wide Range of Vision Guided Machine Learning Applications
- ✓ Broadening the Development and Deployment of Machine Learning Applications from the Edge to the Cloud

© Copyright 2017 Xilinx