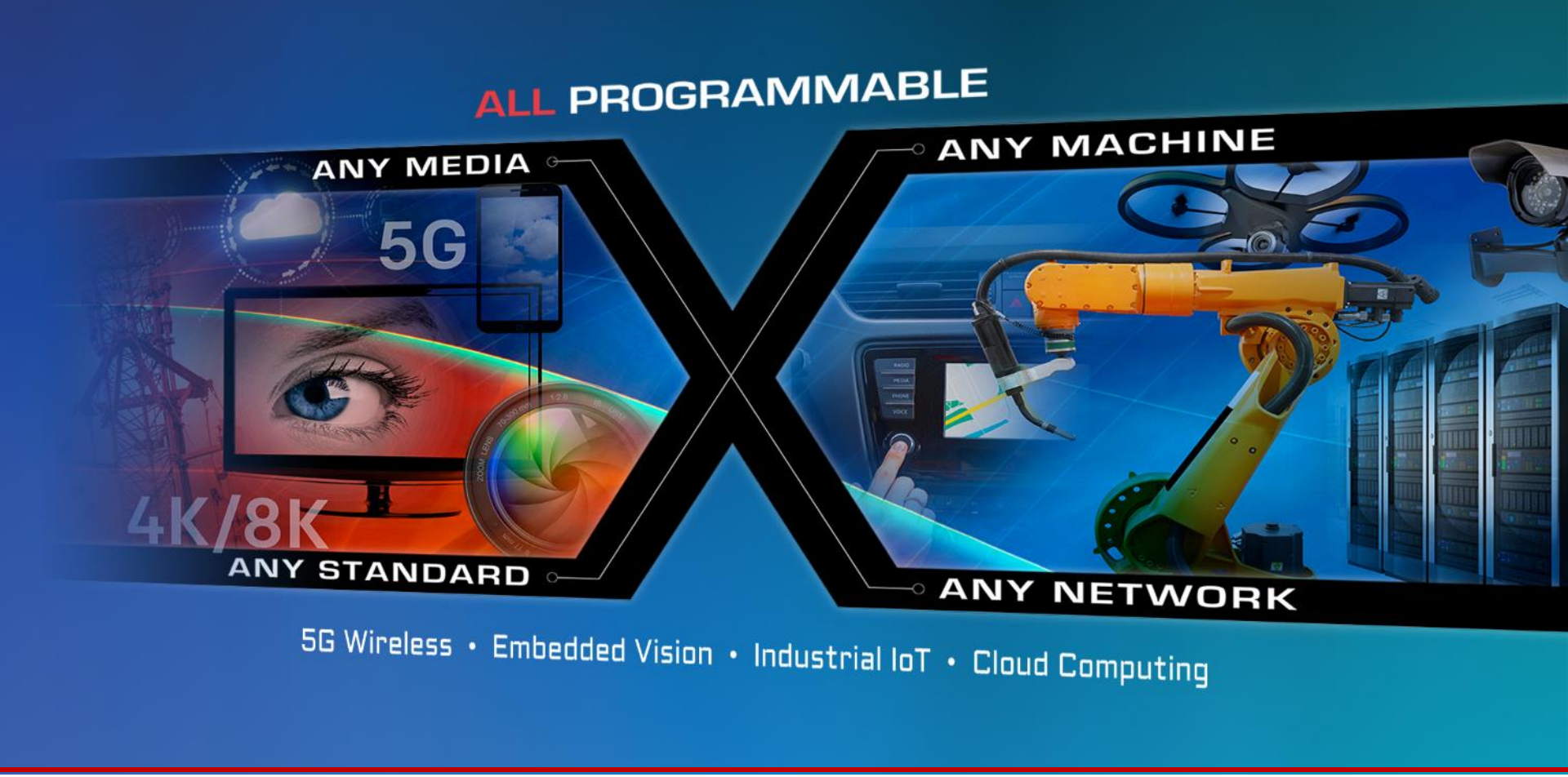


XILINX
ALL PROGRAMMABLE™

Vision based Platforms

Kees Vissers, mpsoc 2016

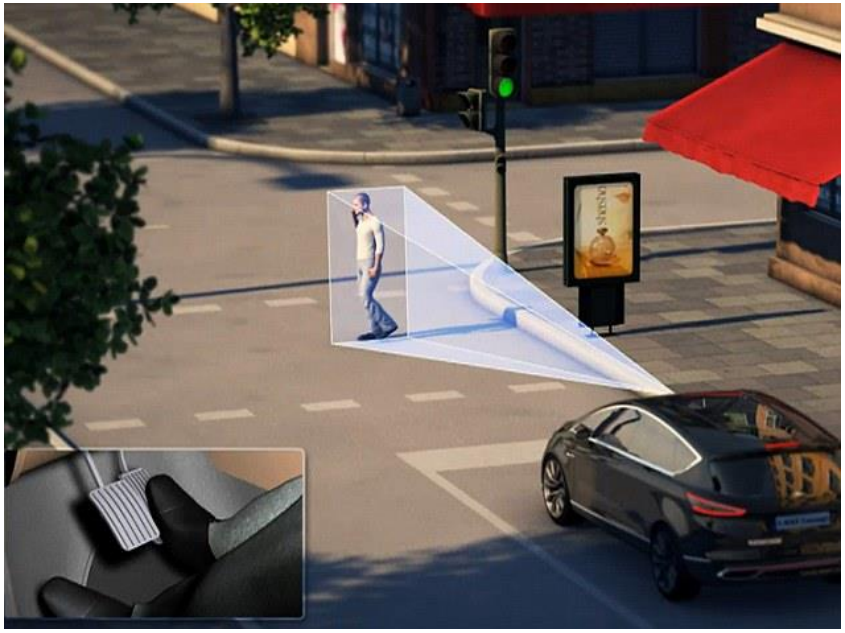
XILINX
ALL PROGRAMMABLE™

Agenda

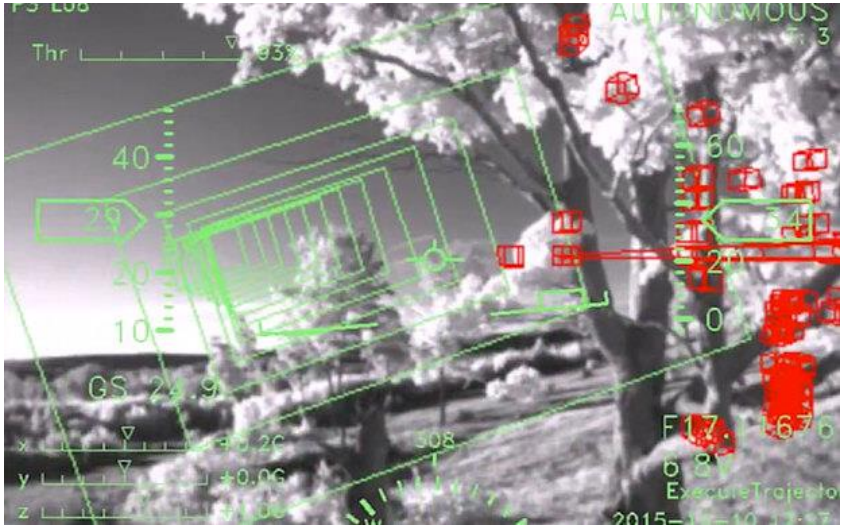
- Vision processing
- Autonomous Driving
- Roofline model
- Energy and size
- Xilinx solutions

Vision Processing

Computer Vision: pedestrians and obstacles.



Active obstacle detection and avoidance



Drone autonomously avoid obstacles

Milliseconds, versus seconds response time, no time to go to the cloud

Analysis: Image Database, Transform and Classification

3D reconstruction from drone images

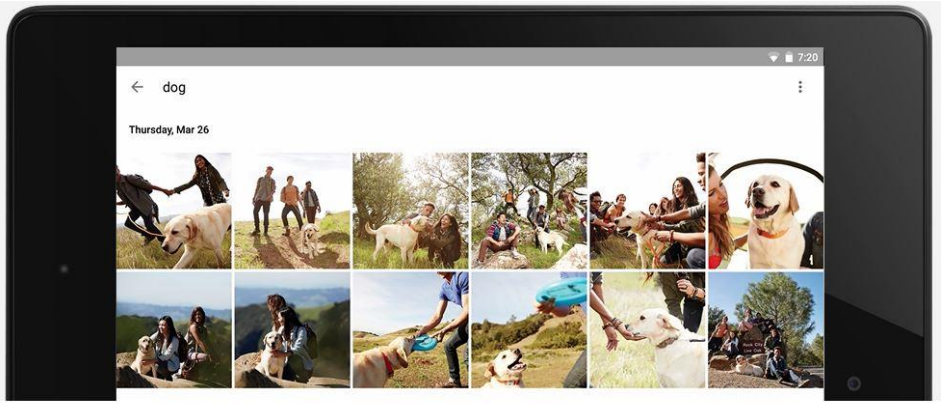
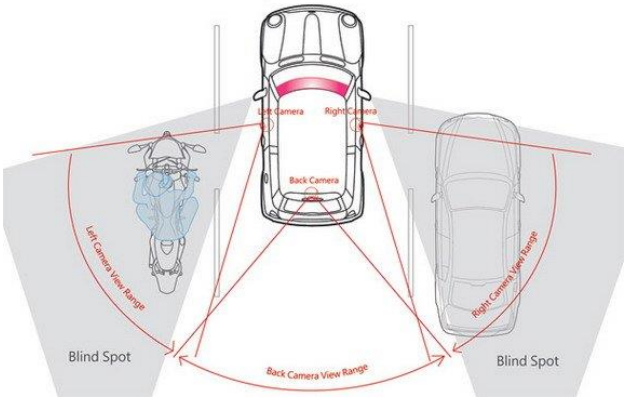


Image Classification

(Embedded) Visualization: Smart/Augmented Views

Smart Rearview Mirror



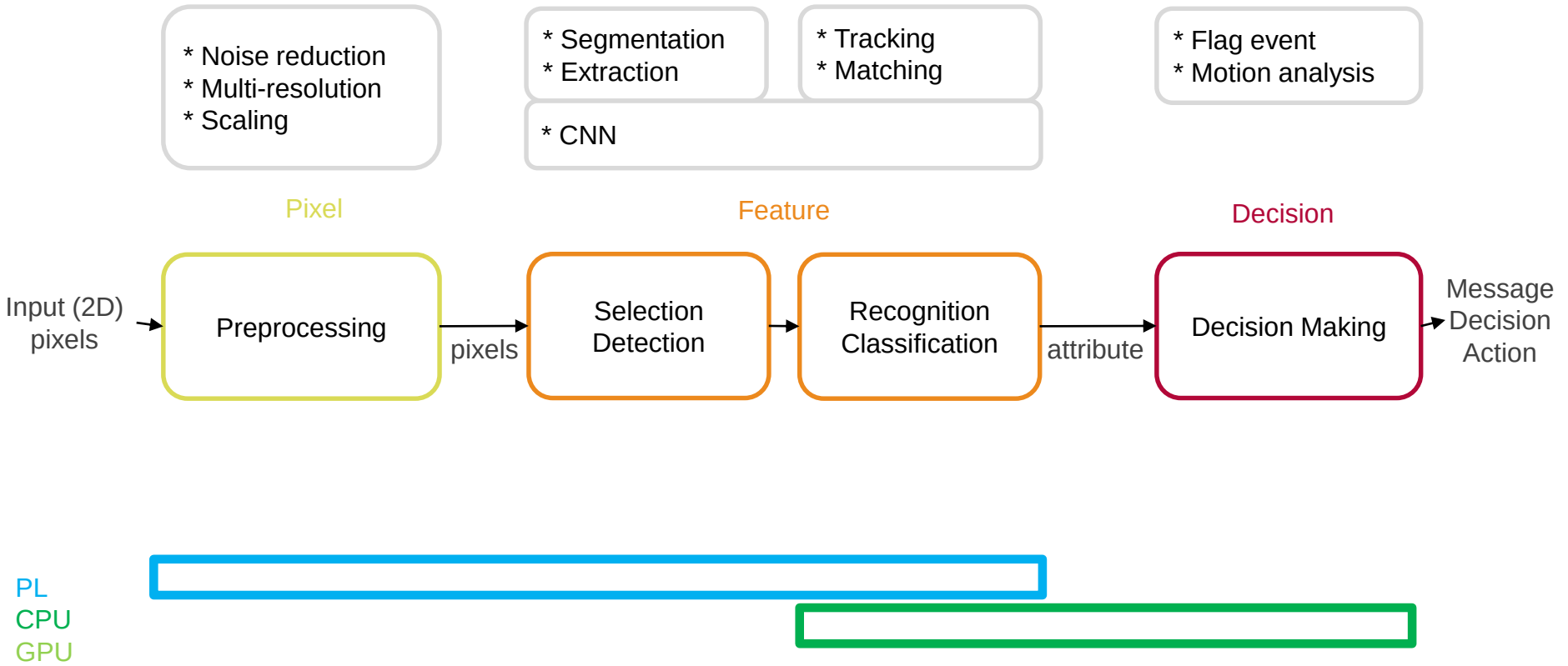
Multiple cameras

Smart Augmented Wide View

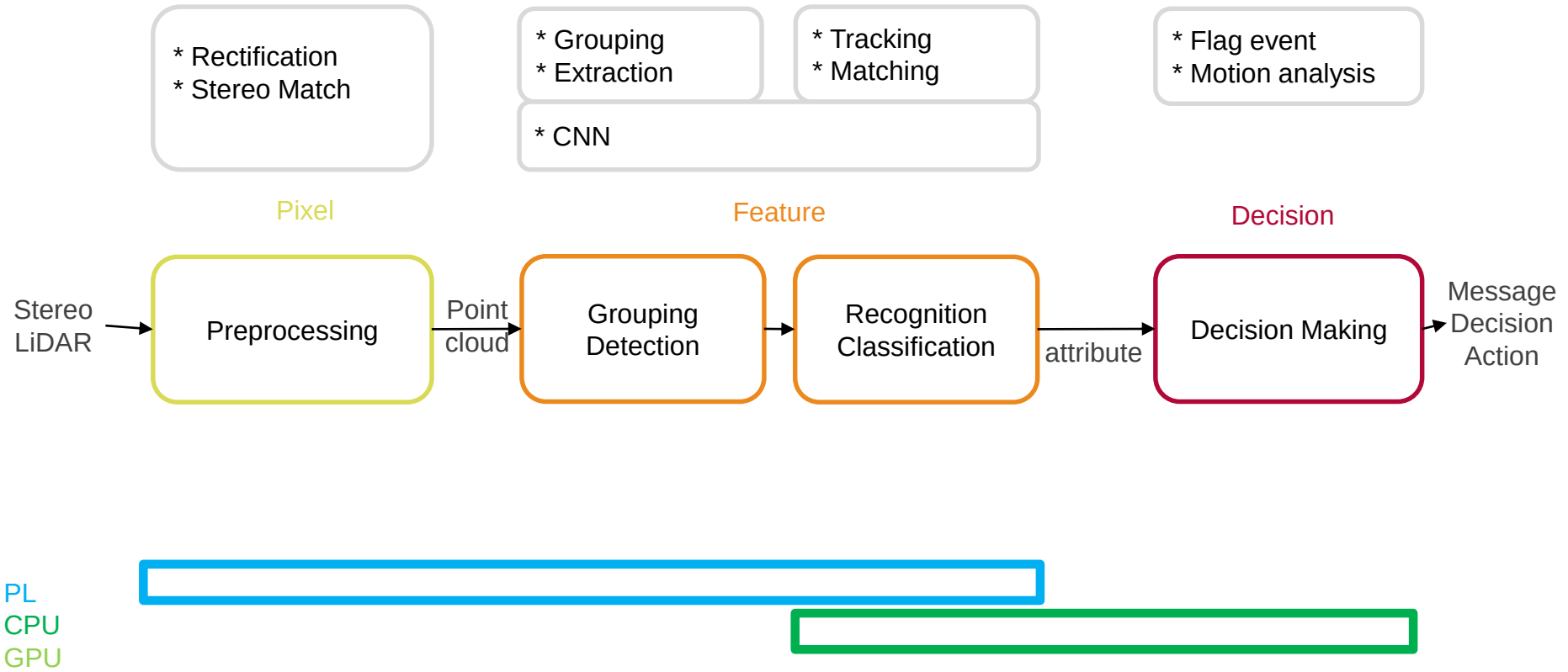
Computer Vision/Visualization

- Computer Vision: understand what is in an image
  - Pixels → Message or Decision potentially leading to automated actions
- Visualization: create images/video to communicate a message
  - Message → Pixels

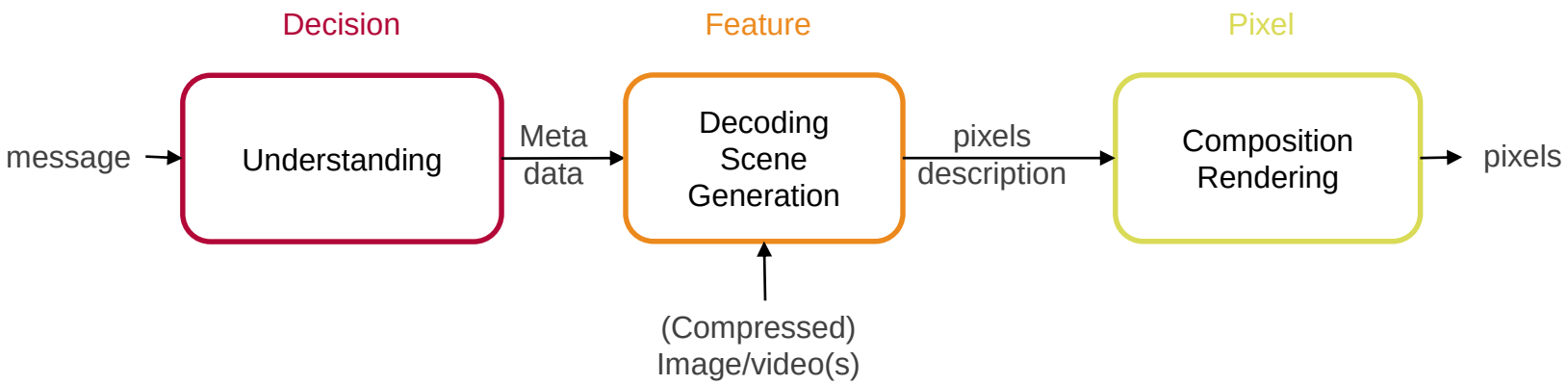
Typical Vision Pipeline on 2D Input



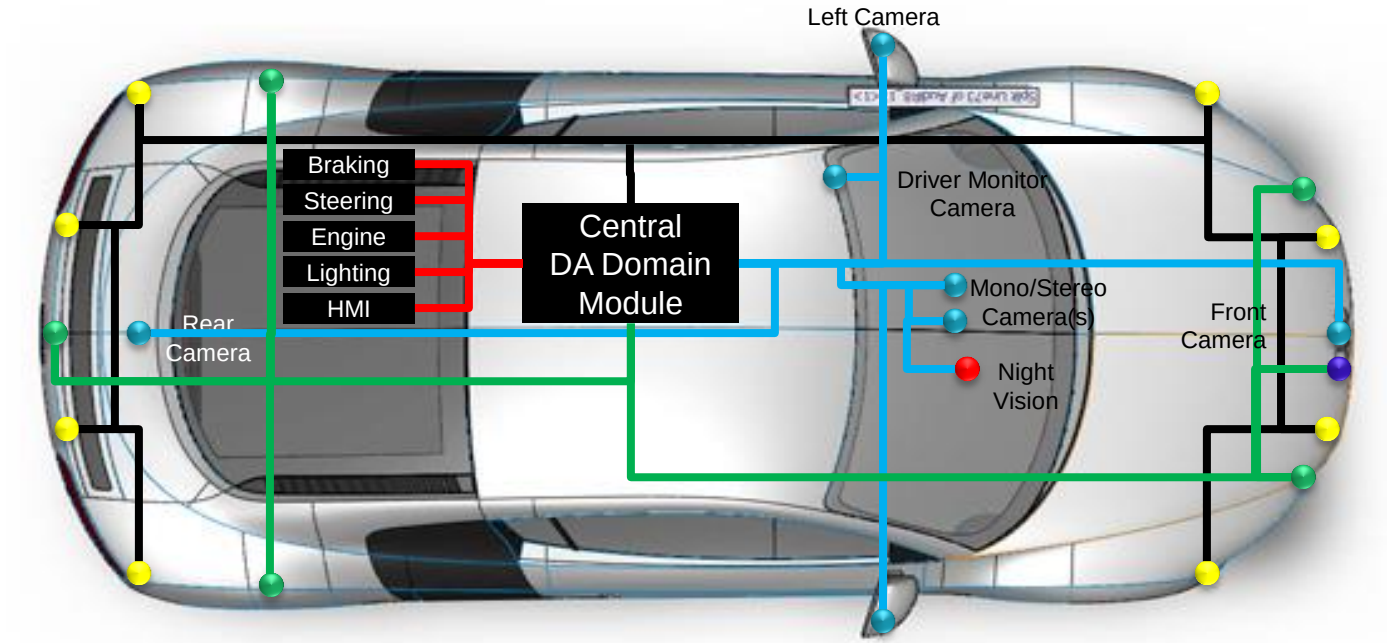
Typical Vision Pipeline on 3D Input



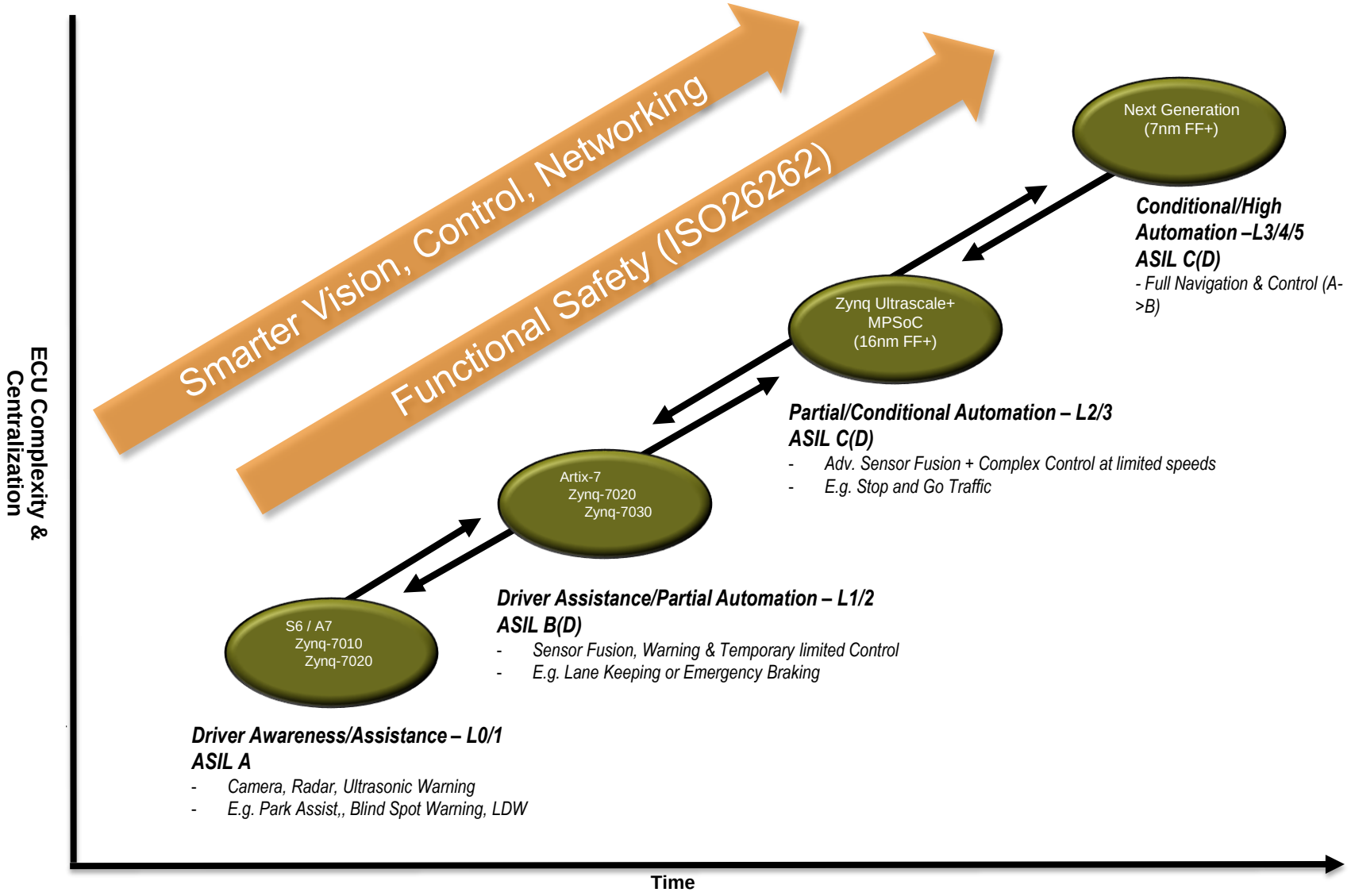
Typical Visualization Pipeline



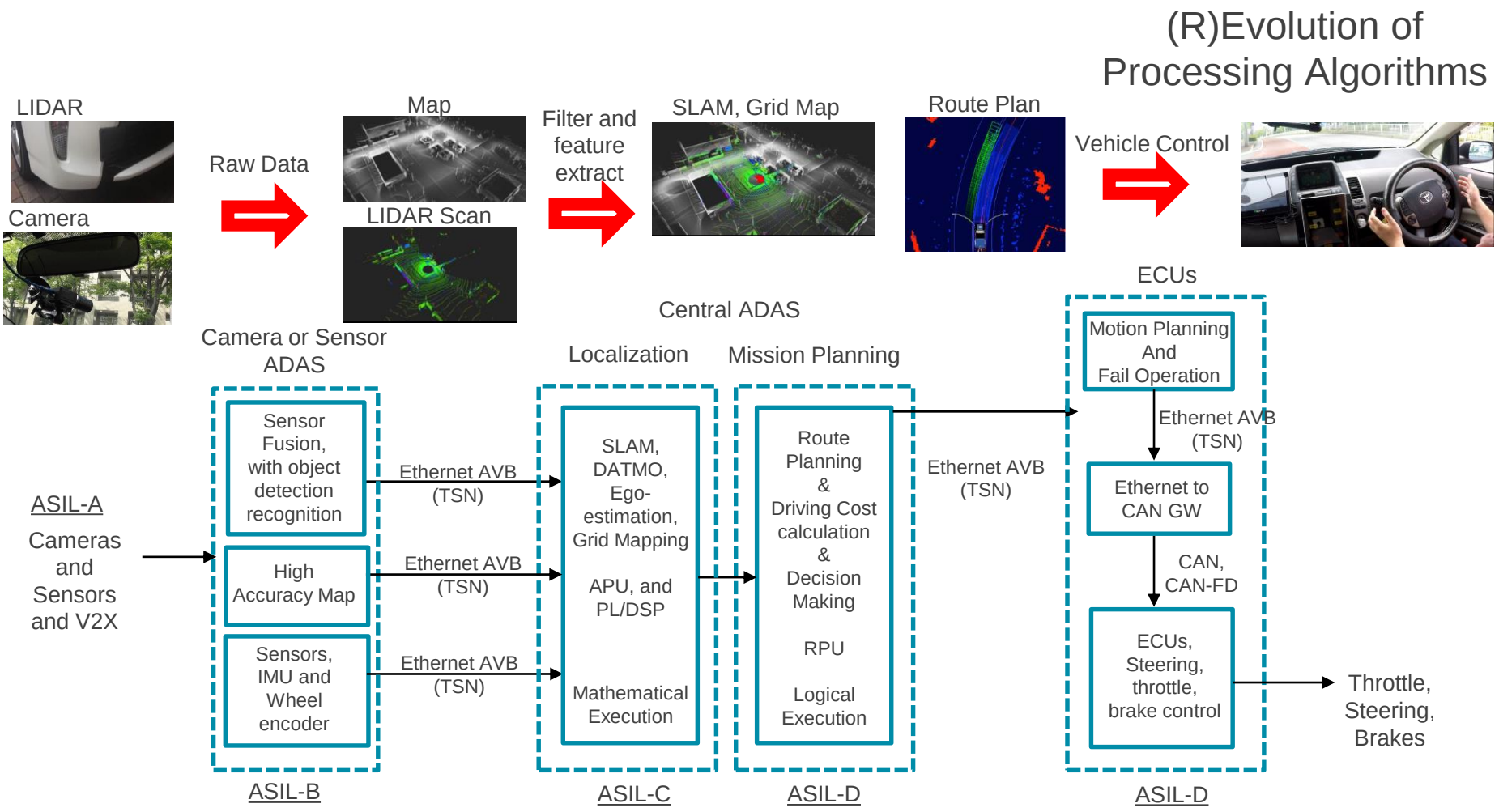
Moving Towards Autonomous Driving



Levels of Automated driving, security requirements



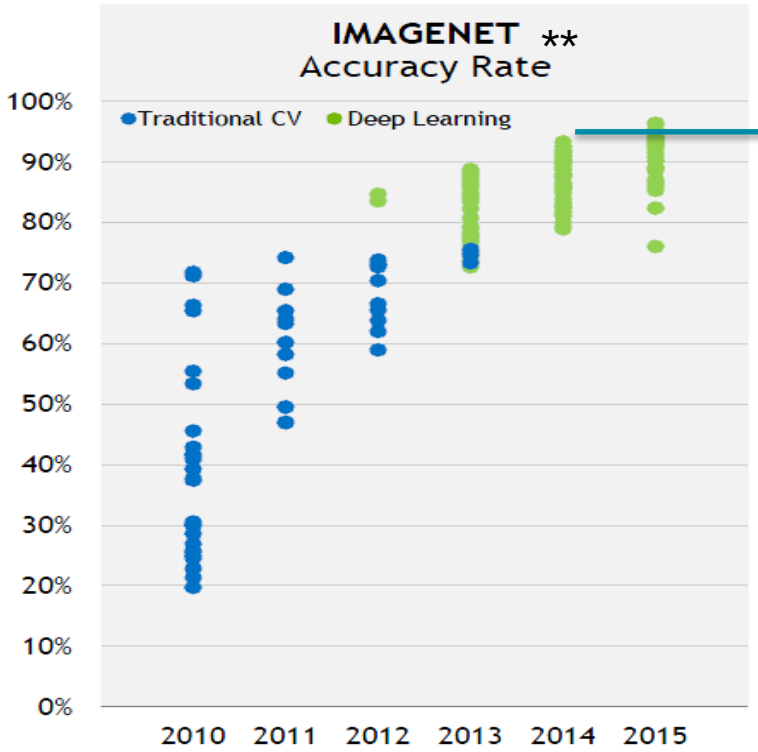
Automated Vehicle Control



Neural Networks

Top-5 Accuracy Image Classification

Image-Net Large-Scale Visual Recognition Challenge (ILSVRC*)



* http://image-net.org/challenges/LSVRC/

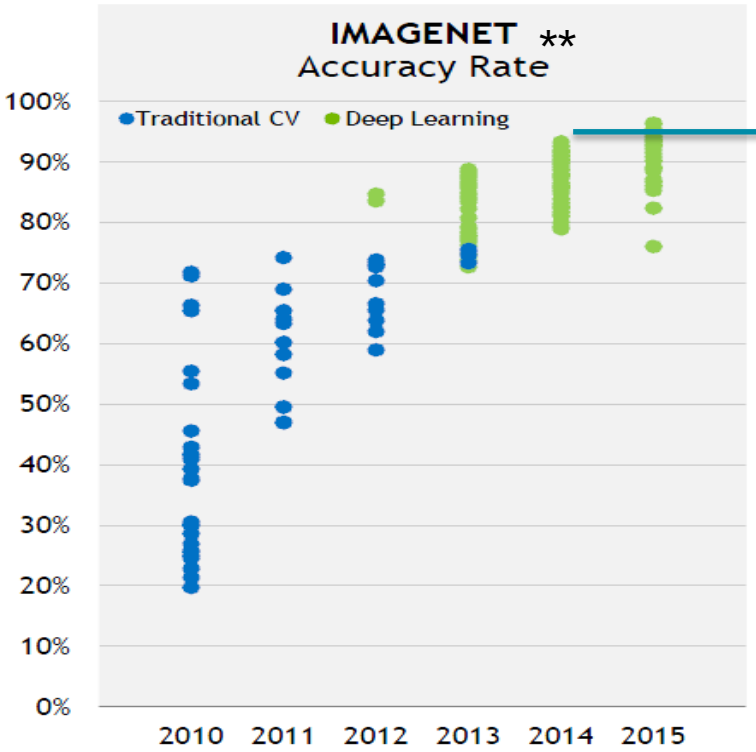
**http://www.slideshare.net/NVIDIA/nvidia-ces-2016-press-conference, pg 10

*** Russakovsky, et al 2014, http://arxiv.org/pdf/1409.0575.pdf

Top-5 Accuracy Image Classification

Image-Net Large-Scale Visual Recognition Challenge (ILSVRC*)

- CNNs outperform classical algorithms
- CNNs deliver super-human accuracy
- At 'Infinite' compute resources.



Super Human

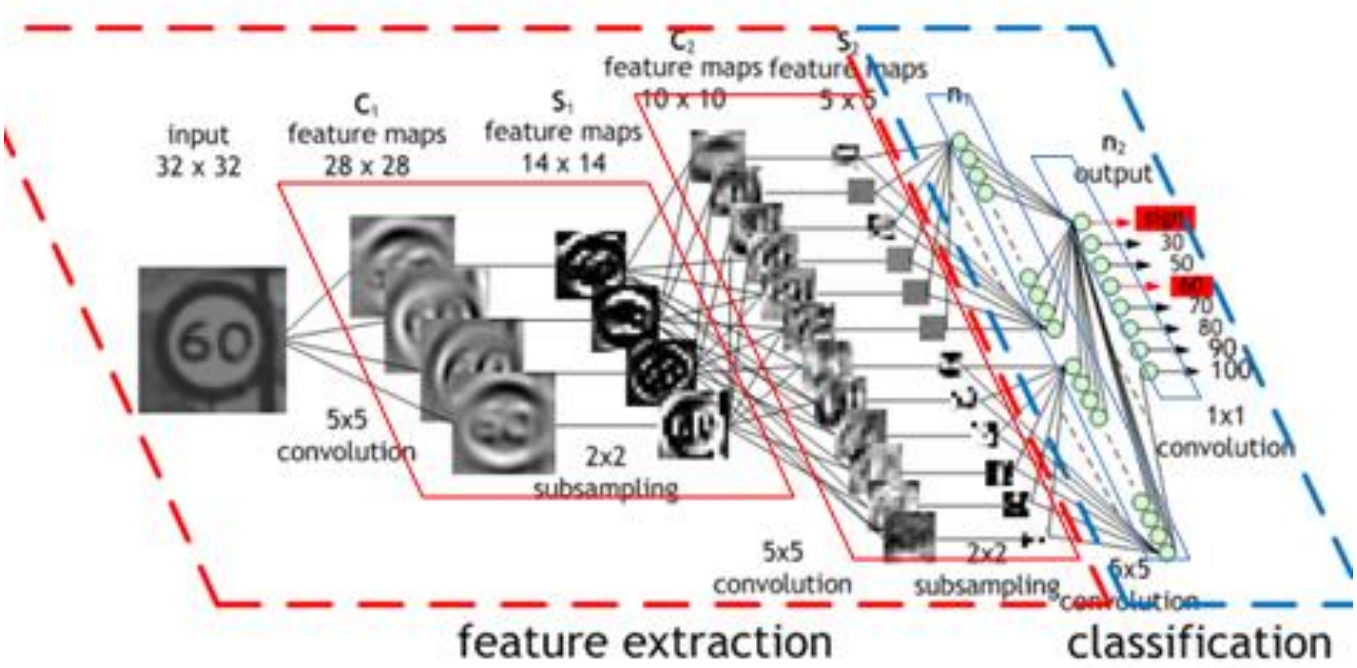
Humans: ~95%***

* http://image-net.org/challenges/LSVRC/

**http://www.slideshare.net/NVIDIA/nvidia-ces-2016-press-conference, pg 10

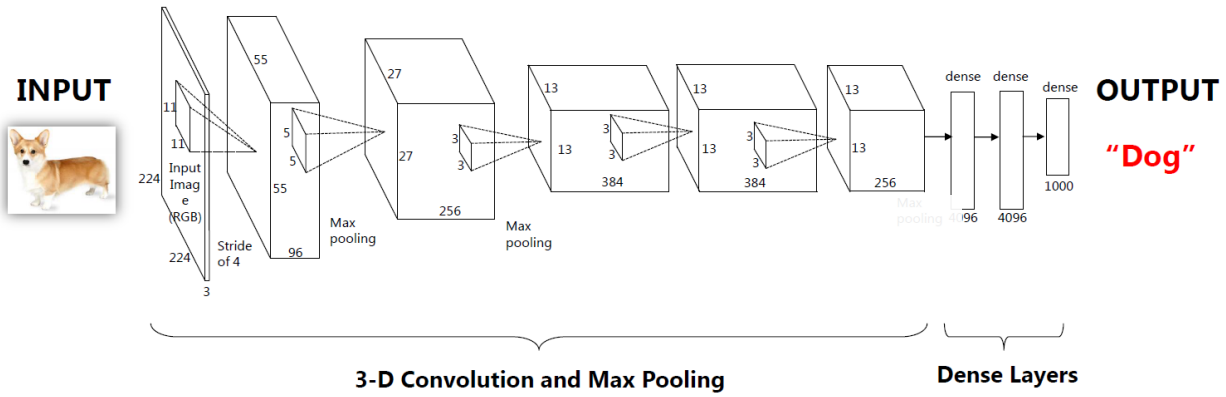
*** Russakovsky, et al 2014, http://arxiv.org/pdf/1409.0575.pdf

CNNs Explained



Conclusion: We can recognize what we have trained for -> Database of Images

Compute and Memory Requirements



CNN for ImageNet datasets	Memory (SP) [MB]	Operations [MOPS]	Operational Intensity [OPS:B]
AlexNet – convolutions only	9.3	1332	**143**
AlexNet – complete	244	1456	**5.97**
VGG-16	552	30823	**55.84**
GoogleNet	27.2	1502	**55.24**

CNNs are highly compute and highly memory intensive

Binary and Almost Binary Networks

Accuracy (published & reproduced results)

Dataset	FP32	BNN	Source
MNIST	99%	99%	[1]
CIFAR-10	92%	90%	[1]
ImageNet (GoogleNet arch)	90% top-5	86% top-5	[2] binary weights
ImageNet (DoReFaNet)	56% top-1	50% top-1	[4] 2-bit activations

[1] Courbariaux, Matthieu, and Yoshua Bengio. "BinaryNet: Training deep neural networks with weights and activations constrained to+ 1 or-1." arXiv preprint arXiv:1602.02830 (2016).

[2] Rastegari, Mohammad, et al. "XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks." arXiv preprint arXiv:1603.05279 (2016).

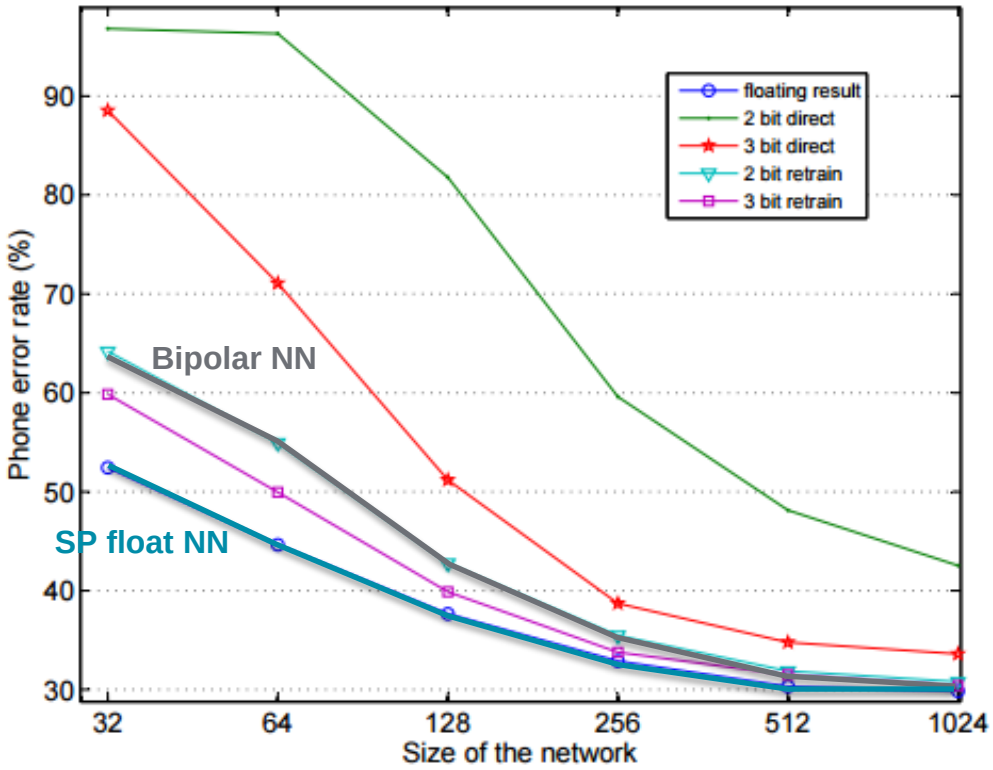
[3] Xundong Wu: High Performance Binarized Neural Networks trained on the ImageNet Classification Task" arXiv:1604.03058

[4] S. Zhou, z.Ni, X. Zhou, H.Wen, Y.Wu, Y. Zou: "DoReFa-Net: Training Low Bitwidth Convolutional Neural Networks with Low Bitwidth Gradients", http://arxiv.org/abs/1606.06160#

Emerging: Low-Precision Networks

- **Reducing precision is shown today to work to 6b**
  - 50x reduction in model size (no external memory needed) [1]
- **Reducing to the extreme: binary neural networks (BNNs)**

"The performance gap between the floating-point and {ternary fixed-point: -1, 0, +1} networks almost vanishes {with a big/complex enough network}." [2]



[1] Iandola et al. "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 1MB model size." (2016).

[2] Sung et al., "Resiliency of Deep Neural Networks Under Quantization", ICLR'16 (fully connected network layers for phoneme recognition)

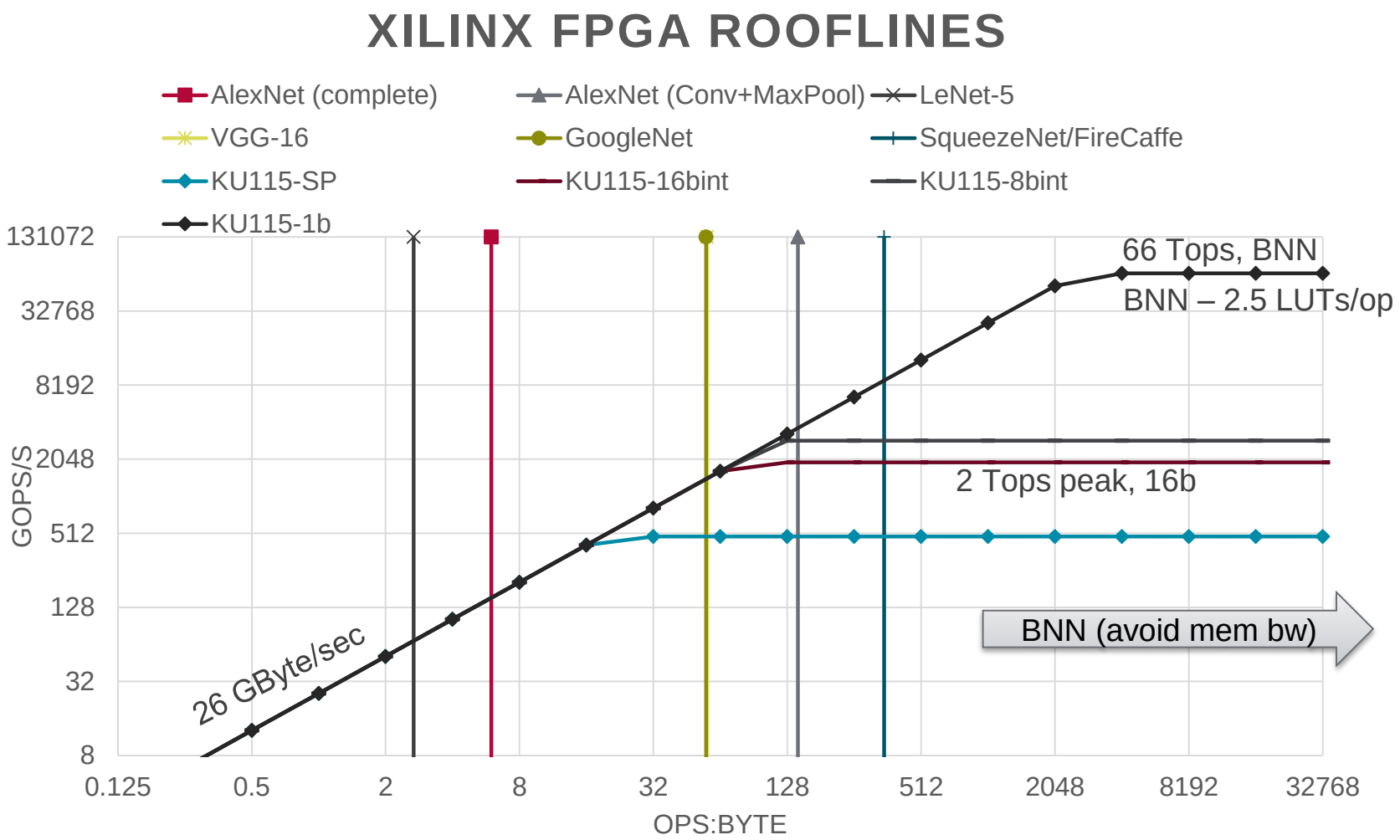
Roofline model

Xilinx Kintex® UltraScale™ KU115

- 5520 DSP Cores, 663 KLuts
- 2Tops or more, 16 bit
- 4 GB, 26Gbyte/sec bandwidth
- Single Slot PCIe, Low Profile
- Total card power ~ 40W

AlphaData ADM-PCIE-8K5

Roofline model for KU115 (ADM-PCIE-8K5) & CNNs



First prototype in Xilinx labs Dublin* shows 4.8TOPS/sec, 1Mfps for MNIST

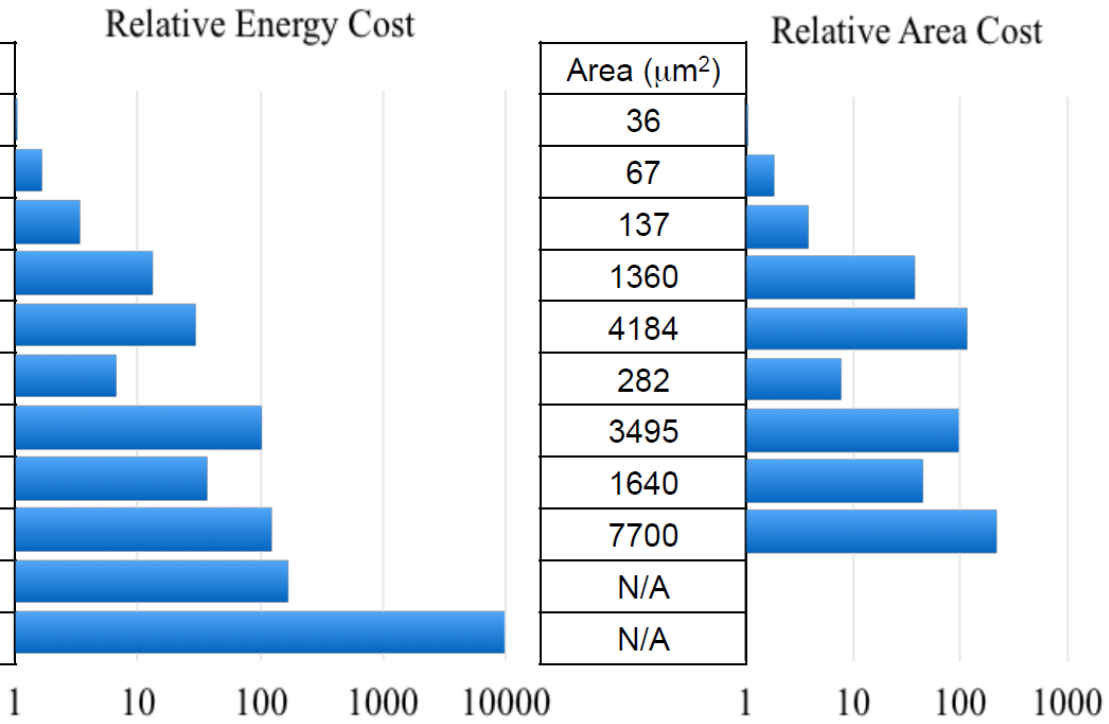
*Yaman Umuroglu (NTNU, Xilinx); Nicholas Fraser (University of Sydney, Xilinx); Michaela Blott (Xilinx Research)

How to reach peak performance

- Use minimal resources, i.e 1DSP per 16 bit Multiply-Accumulate (MAC).
- Use a reduced number of coefficients net (move to the right), e.g. Squeeze net
- Use the bit-accuracy that makes sense: e.g. 16bit, 8bit, 1 bit
- Use local buffering to avoid spending energy in memory interface or hitting memory bandwidth limitations
- Peak performance in Tops range possible today
- Best in class performance/watt of programmable devices (~10x lower Gops/W for 16bit operation compared to GPUs)

Energy and Si area of Computation

Operation:	Energy (pJ)	Area (μm²)
8b Add	0.03	36
16b Add	0.05	67
32b Add	0.1	137
16b FP Add	0.4	1360
32b FP Add	0.9	4184
8b Mult	0.2	282
32b Mult	3.1	3495
16b FP Mult	1.1	1640
32b FP Mult	3.7	7700
32b SRAM Read (8KB)	5	N/A
32b DRAM Read	640	N/A



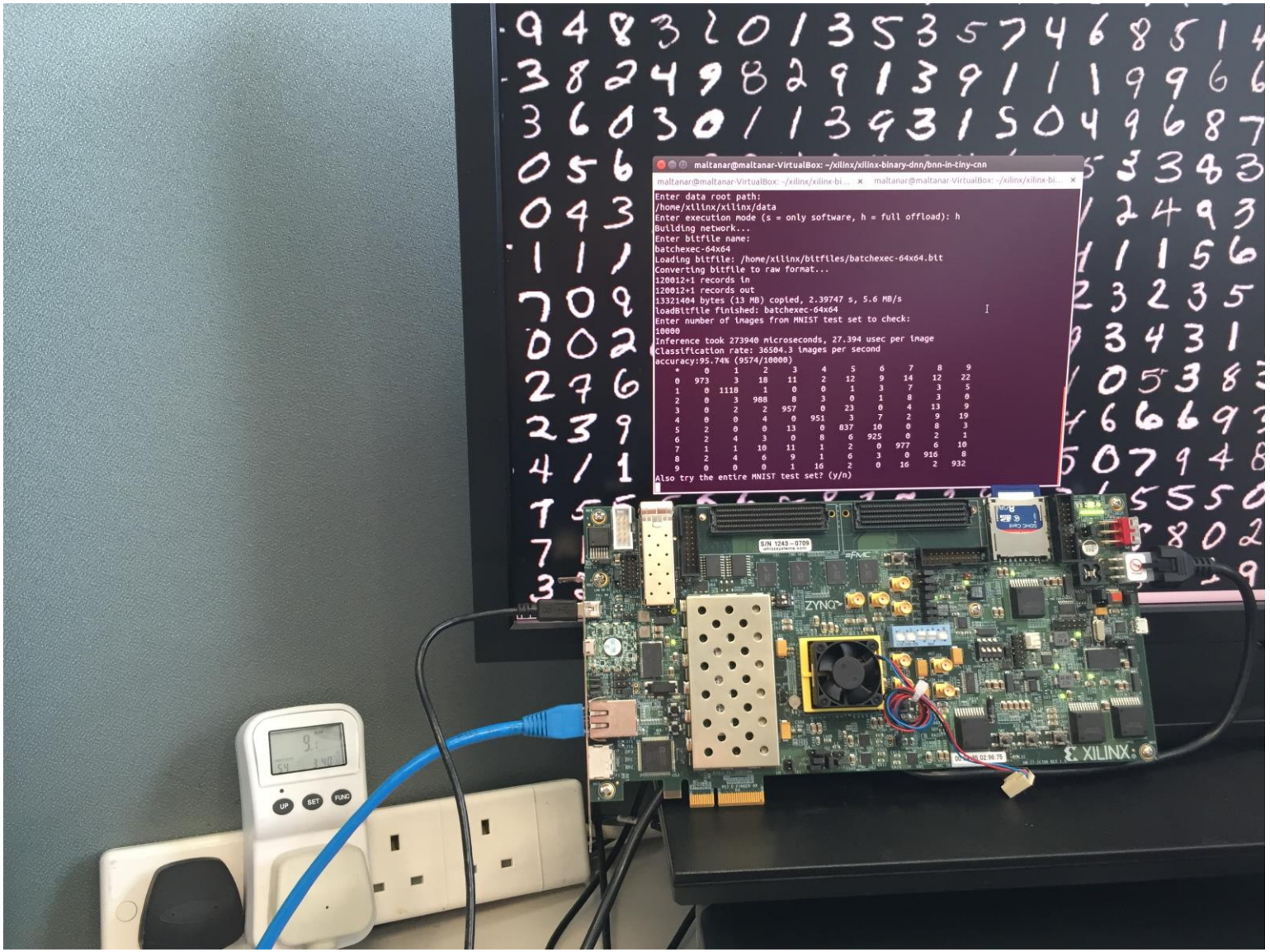
Energy numbers are from Mark Horowitz "Computing's Energy Problem (and what we can do about it)", ISSCC 2014
Area numbers are from synthesized result using Design Compiler under TSMC 45nm tech node. FP units used DesignWare Library.

Source: William Dally, "High Performance Hardware for Machine Learning"

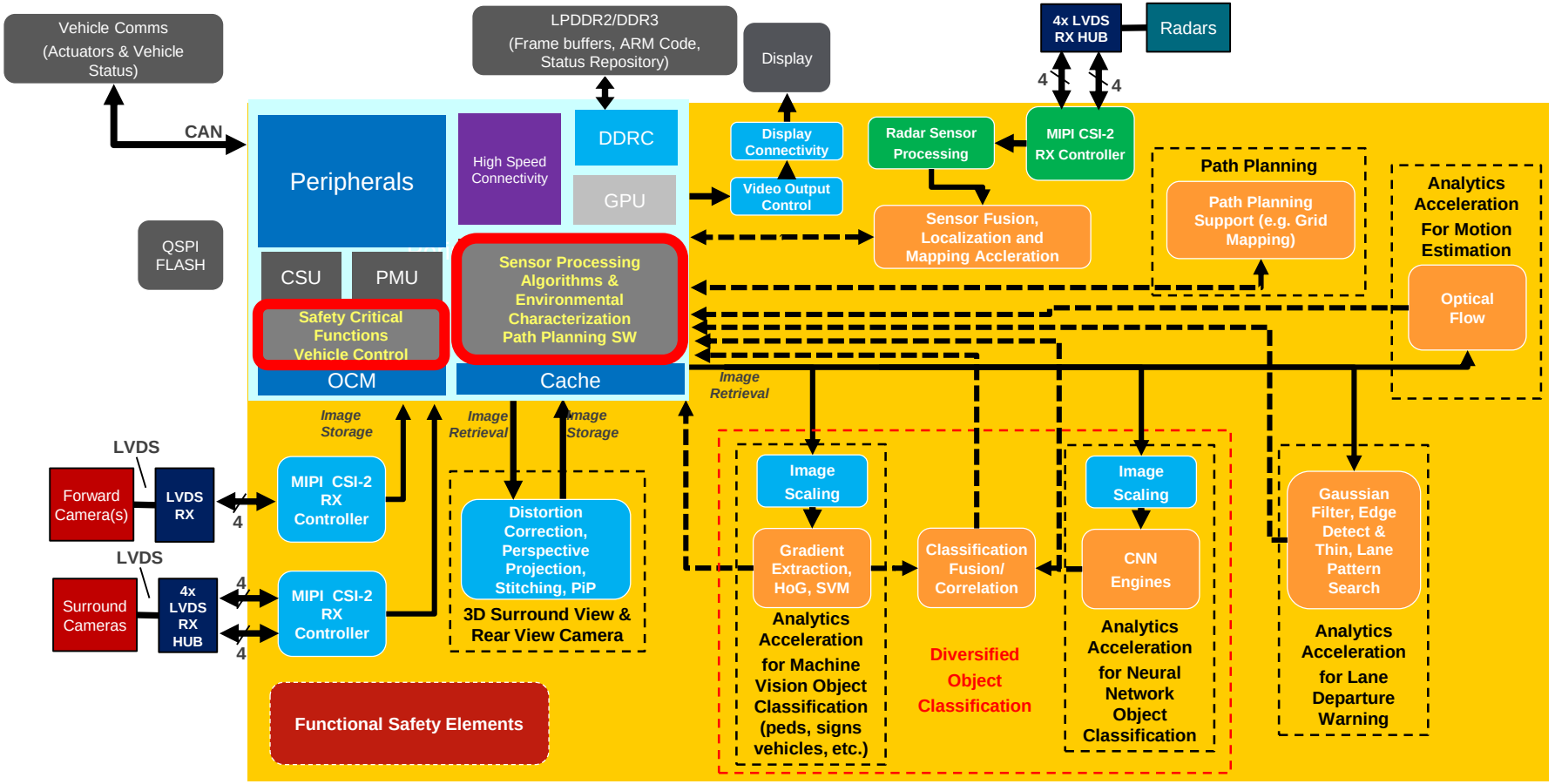
Energy Considerations

- Use local on chip buffering to avoid going to External Memory
- Streaming based solutions are very power efficient
- Threading, and Cache misses are designed for Random access programs
- Streaming and explicit (pre)fetching fit the characteristics of Neural Networks, sliding window style processing is possible.

Lab setup using the MNIST dataset, Zynq chip



ZU+ MPSoC-based Multi-Feature Module



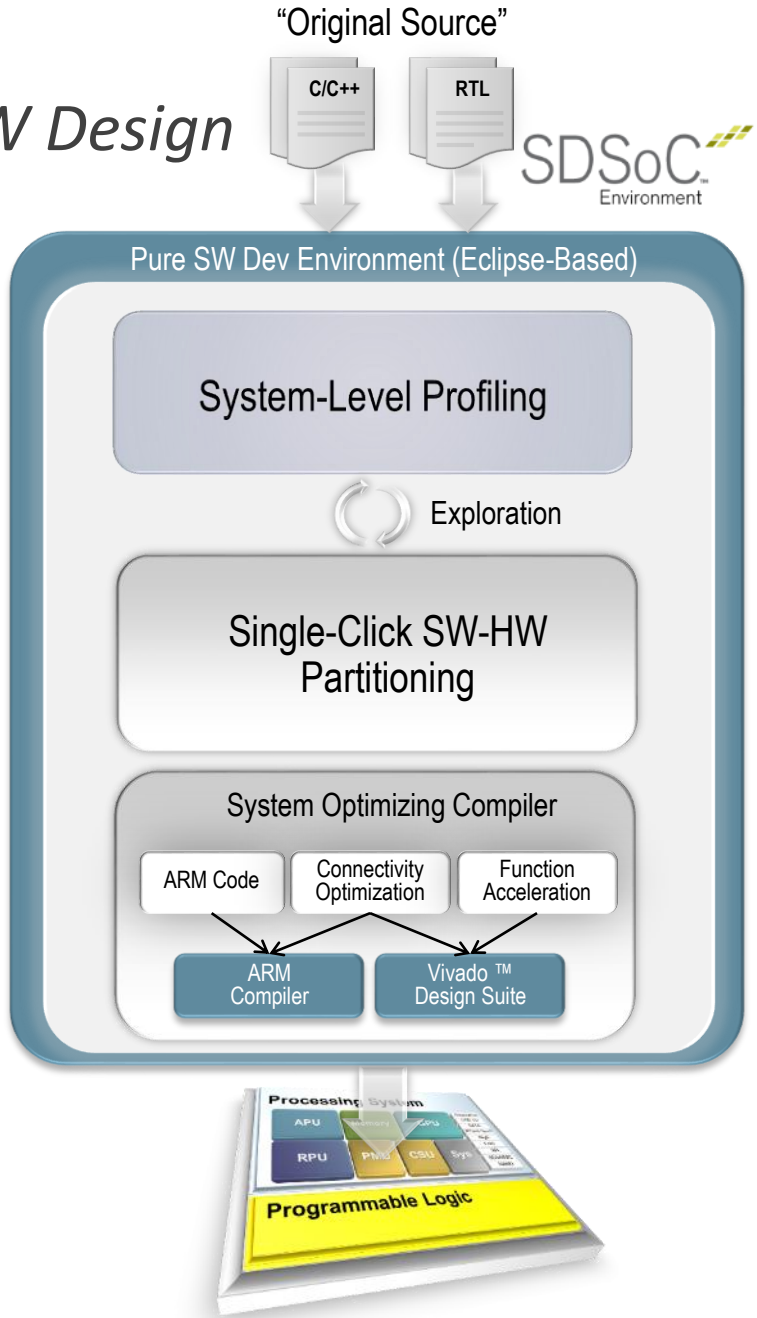
Conclusions

- External Memory transfers are 100-200x more energy then a basic operation: Build architectures that minimize external memory transfers, use on chip local storage.
- Multiple High-resolution cameras and sensors will be deployed in Automotive systems.
- Complete vision processing includes standard algorithms and Neural Networks.
- Reduced precision in Neural Networks is a very promising research direction
- Compute load will be several Tops, in a few Watts for the complete device.
- What really works in vision processing is active research: build real systems!
- MPSoC devices (ZU3, ZU4, ZU5, ZU7, ZU9, ZU15) will do (Automotive) system vision processing in a few Watts, including Neural Networks.
- A complete programming environments is available, including OS, High-level Synthesis, and SDSoC.

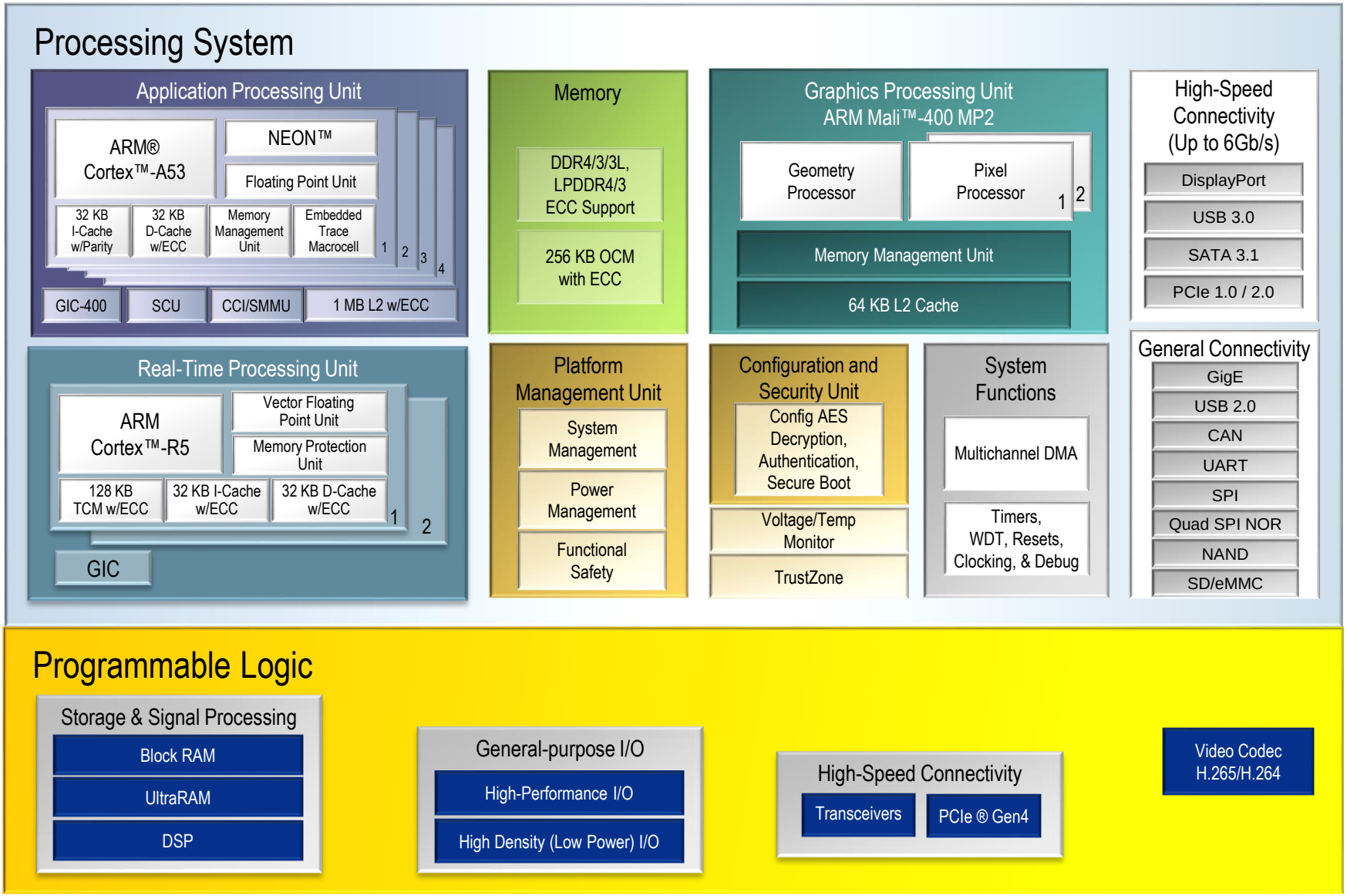
SDSoC Dev Environment

Complete Environment for HW and SW Design

Feature	Benefit
Pure SW Dev Environment	• From C/C++ to fully functional AP SoC without any HW code • Familiar, Eclipse-based, SW-centric design environment
System Level Profiling	• PS/PL performance, area, data transfer estimates in minutes • Quickly identify optimal total system architecture
Single-Click SW/HW Partitioning	• Quickly explore different system connectivity in C/C++ • Find data mover and PS-PL interface for optimal dataflow
System Optimizing Compiler	• Automated function acceleration in logic fabric • Generates both ARM software and FPGA bitstream • Examine throughput, latency and area tradeoffs



Zynq UltraScale+ Block Diagram



Zynq® UltraScale+™ MPSoCs: EG Devices

			Smarter Control and Vision					Smarter Network					
		Device Name[1]	ZU2EG	ZU3EG	ZU4EG	ZU5EG	ZU7EG	ZU6EG	ZU9EG	ZU15EG	ZU11EG	ZU17EG	ZU19EG
Processing System (PS)	Application Processor Unit	Processor Core	**Quad-core** ARM® Cortex™-A53 MPCore™ up to 1.5GHz										
		Memory w/ECC	L1 Cache 32KB I / D per core, L2 Cache 1MB, on-chip Memory 256KB										
	Real-Time Processor Unit	Processor Core	**Dual-core** ARM Cortex-R5 MPCore™ up to 600MHz										
		Memory w/ECC	L1 Cache 32KB I / D per core, Tightly Coupled Memory 128KB per core										
	Graphic & Video Acceleration	Graphics Processing Unit	Mali™-400 MP2 up to 667MHz										
		Memory	L2 Cache 64KB										
	External Memory	Dynamic Memory Interface	x32/x64: DDR4, LPDDR4, DDR3, DDR3L, LPDDR3 with ECC										
		Static Memory Interfaces	NAND, 2x Quad-SPI										
	Connectivity	High-Speed Connectivity	PCIe® Gen2 x4, 2x USB3.0, SATA 3.1, DisplayPort, 4x Tri-mode Gigabit Ethernet										
		General Connectivity	2xUSB 2.0, 2x SD/SDIO, 2x UART, 2x CAN 2.0B, 2x I2C, 2x SPI, 4x 32b GPIO										
	Integrated Block Functionality	Power Management	Full / Low / PL / Battery Power Domains										
		Security	RSA, AES, and SHA										
		AMS - System Monitor	10-bit, 1MSPS - Temperature, Voltage, and Current Monitor										
PS to PL Interface			12 x 32/64/128b AXI Ports										
Programmable Logic (PL)	Programmable Functionality	System Logic Cells (K)	103	154	192	256	504	469	600	747	653	926	1,143
		CLB Flip-Flops (K)	94	141	176	234	461	429	548	682	597	847	1,045
		CLB LUTs (K)	47	71	88	117	230	215	274	341	299	423	523
	Memory	Max. Distributed RAM (Mb)	1.2	1.8	2.6	3.5	6.2	6.9	8.8	11.3	9.1	8.0	9.8
		Total Block RAM (Mb)	5.3	7.6	4.5	5.1	11.0	25.1	32.1	26.2	21.1	28.0	34.6
		UltraRAM (Mb)	-	-	14.0	18.0	27.0	-	-	31.5	22.5	28.7	36.0
	Clocking	Clock Management Tiles (CMTs)	3	3	4	4	8	4	4	4	8	11	11
	Integrated IP	DSP Slices	240	360	728	1,056	1,728	1,973	2,520	3,528	2,928	1,590	1,968
		PCI Express® Gen 3x16 / Gen4x8	-	-	2	2	2	-	-	-	4	4	5
		150G Interlaken	-	-	-	-	-	-	-	-	2	2	4
		100G Ethernet MAC/PCS w/RS-FEC	-	-	-	-	-	-	-	-	1	2	4
		AMS - System Monitor	1	1	1	1	1	1	1	1	1	1	1
	Speed Grades	Extended[2]	-1 -2L -3					-1 -2L -3					
		Industrial	-1 -1L -2					-1 -1L -2					

Notes:
1. For full part number details, see the Ordering Information section in DS891, *Zynq UltraScale+ MPSoC Overview*.
2.-2LE (Tj = 0°C to 110°C). For more details, see the Ordering Information section in DS891, *Zynq UltraScale+ MPSoC Overview*.

Zynq® UltraScale+™ MPSoCs: EV Devices

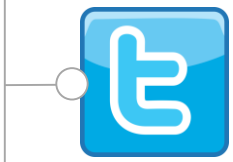
			Smarter Vision		
		Device Name[1]	ZU4EV	ZU5EV	ZU7EV
Processing System (PS)	Application Processor Unit	Processor Core	**Quad-core** ARM® Cortex™-A53 MPCore™ up to 1.5GHz		
		Memory w/ECC	L1 Cache 32KB I / D per core, L2 Cache 1MB, on-chip Memory 256KB		
	Real-Time Processor Unit	Processor Core	**Dual-core** ARM Cortex-R5 MPCore™ up to 600MHz		
		Memory w/ECC	L1 Cache 32KB I / D per core, Tightly Coupled Memory 128KB per core		
	Graphic & Video Acceleration	Graphics Processing Unit	Mali™-400 MP2 up to 667MHz		
		Memory	L2 Cache 64KB		
	External Memory	Dynamic Memory Interface	x32/x64: DDR4, LPDDR4, DDR3, DDR3L, LPDDR3 with ECC		
		Static Memory Interfaces	NAND, 2x Quad-SPI		
	Connectivity	High-Speed Connectivity	PCIe® Gen2 x4, 2x USB3.0, SATA 3.1, DisplayPort, 4x Tri-mode Gigabit Ethernet		
		General Connectivity	2xUSB 2.0, 2x SD/SDIO, 2x UART, 2x CAN 2.0B, 2x I2C, 2x SPI, 4x 32b GPIO		
	Integrated Block Functionality	Power Management	Full / Low / PL / Battery Power Domains		
		Security	RSA, AES, and SHA		
		AMS - System Monitor	10-bit, 1MSPS - Temperature, Voltage, and Current Monitor		
PS to PL Interface			12 x 32/64/128b AXI Ports		
Programmable Logic (PL)	Programmable Functionality	System Logic Cells (K)	192	256	504
		CLB Flip-Flops (K)	176	234	461
		CLB LUTs (K)	88	117	230
	Memory	Max. Distributed RAM (Mb)	2.6	3.5	6.2
		Total Block RAM (Mb)	4.5	5.1	11.0
		UltraRAM (Mb)	14.0	18.0	27.0
	Clocking	Clock Management Tiles (CMTs)	4	4	8
	Integrated IP	DSP Slices	728	1,056	1,728
		Video Codec Unit (VCU)	1	1	1
		PCI Express® Gen 3x16 / Gen4x8	2	2	2
		150G Interlaken	-	-	-
		100G Ethernet MAC/PCS w/RS-FEC	-	-	-
		AMS - System Monitor	1	1	1
	Speed Grades	Extended[2]	-1 -2L -3		
		Industrial	-1 -1L -2		

Notes:
1. For full part number details, see the Ordering Information section in DS891, *Zynq UltraScale+ MPSoC Overview*.
2.-2LE (Tj = 0°C to 110°C). For more details, see the Ordering Information section in DS891, *Zynq UltraScale+ MPSoC Overview*.

Follow Xilinx

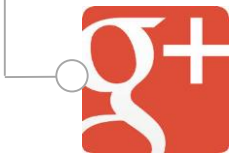
XILINX
ALL PROGRAMMABLE™


facebook.com/XilinxInc


twitter.com/XilinxInc

youtube.com/XilinxInc


linkedin.com/company/Xilinx


plus.google.com/+Xilinx