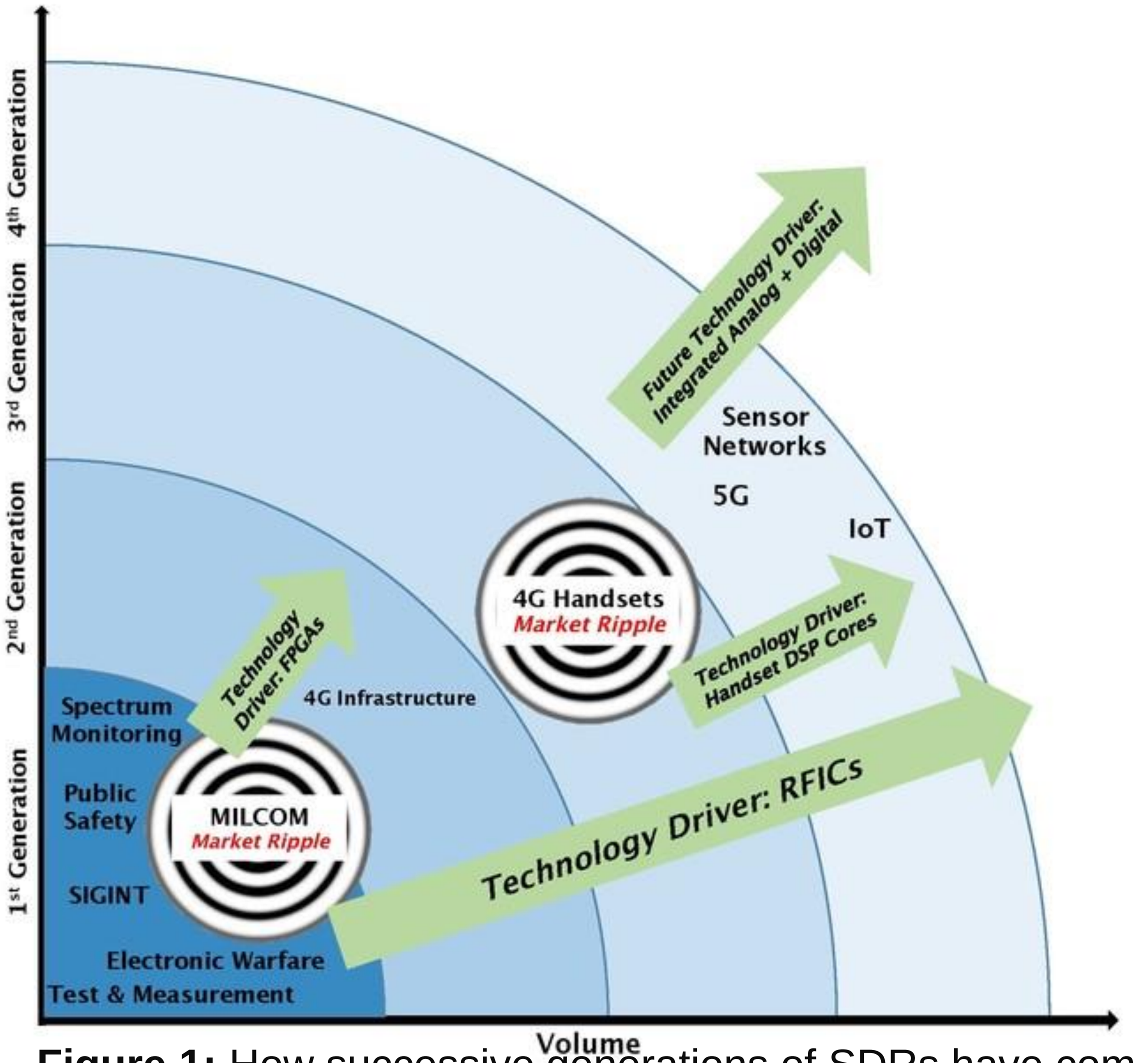
AI and SDR:
Software Meets Hardware Again…

Manuel Uhm
Director, Silicon Marketing
Chair of the Board, Wireless Innovation Forum (SDR Forum v2.0)

Jason Vidmar
Sr. System Architect – MILCOM / SATCOM / Machine Learning

SDR Evolution



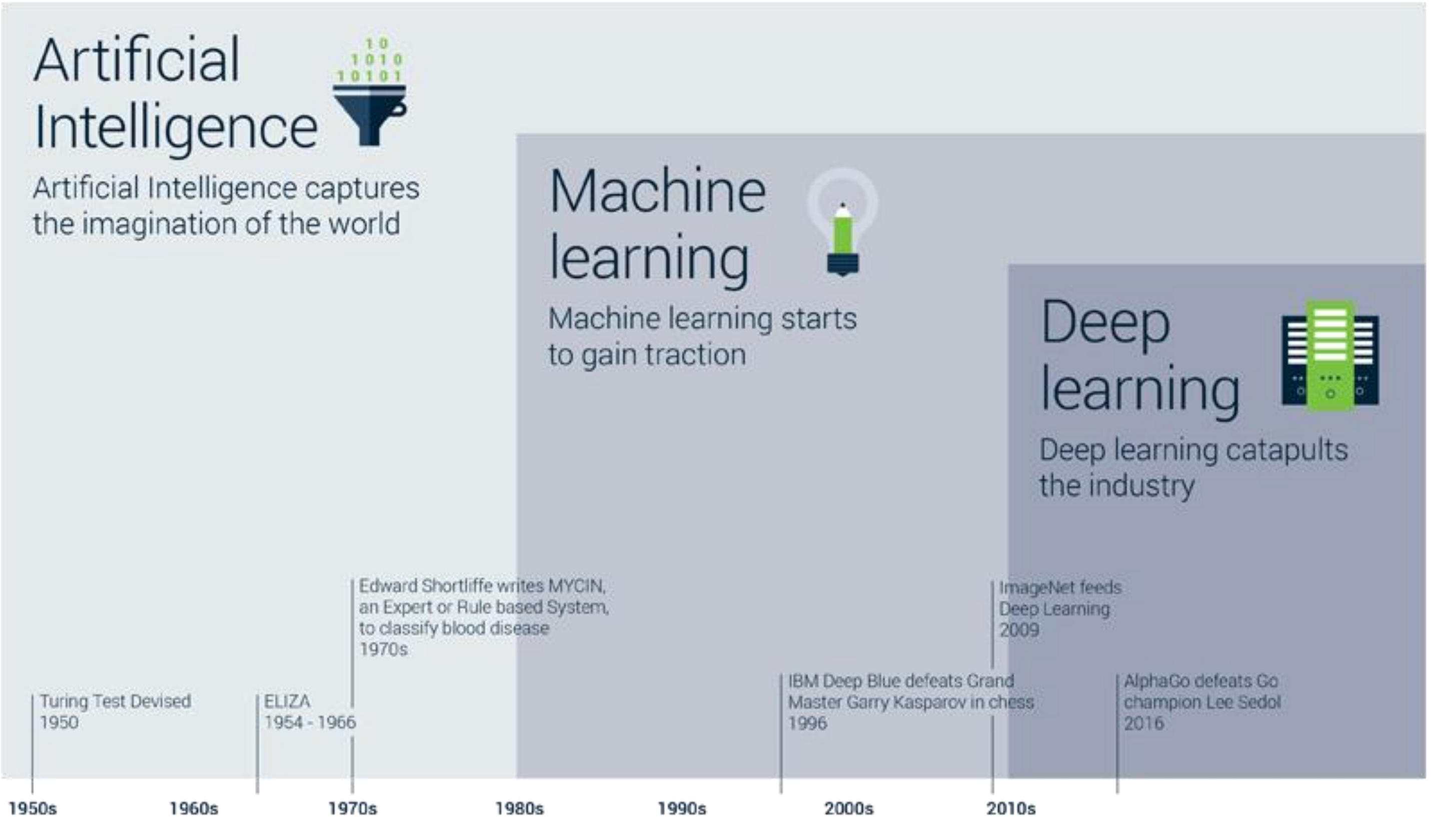
Key semiconductor technology drivers:

- Moore’s Law
- FPGAs
- RFICs
- Analog/Digital Integration

Figure 1: How successive generations of SDRs have come to dominate the radio industry and will continue to evolve.

Source: Manuel Uhm, *Software-Defined Radio: To Infinity and Beyond*, Military Embedded Systems, October 2016

AI Evolution

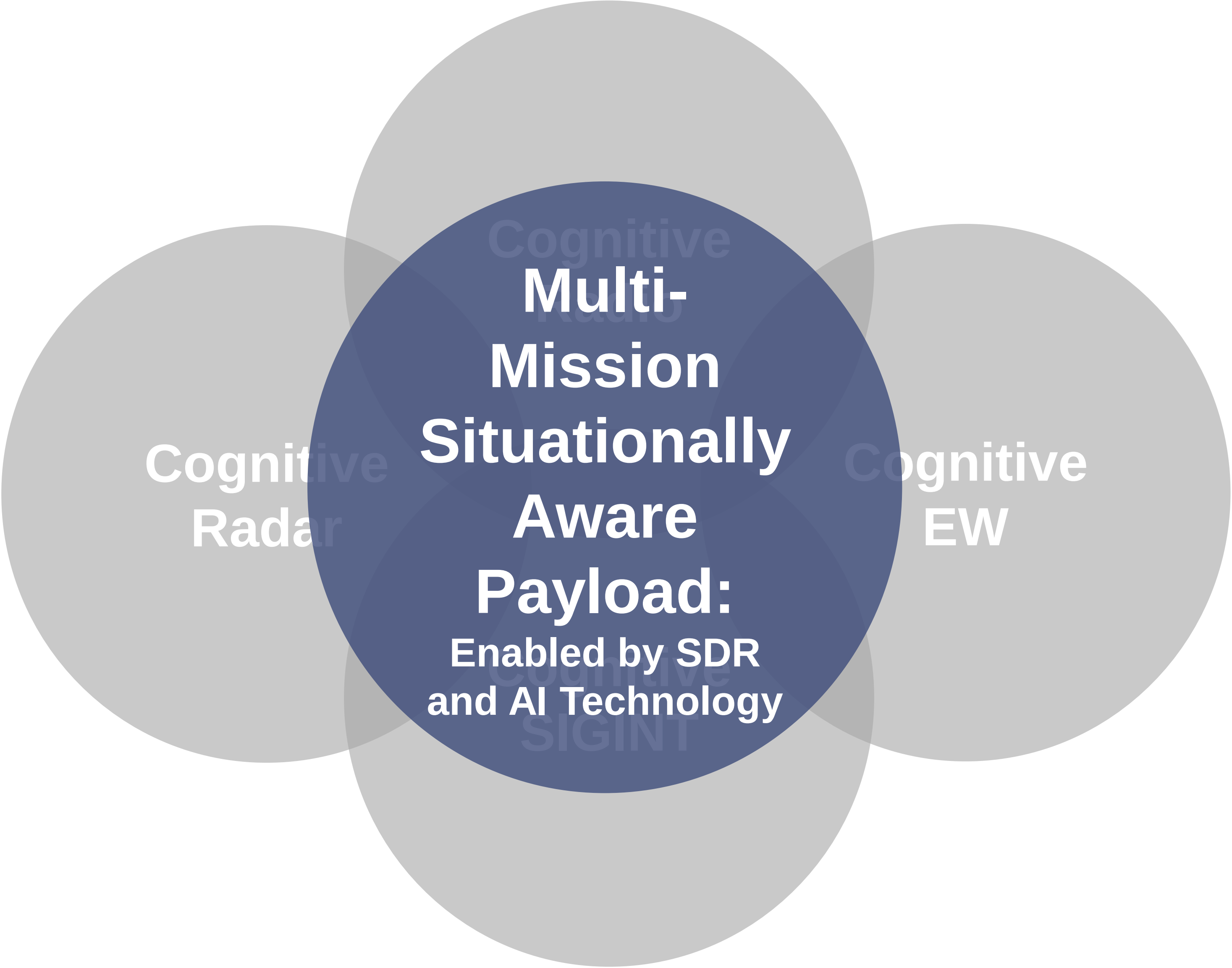

Artificial Intelligence
Artificial Intelligence captures the imagination of the world
Machine learning
Machine learning starts to gain traction
Deep learning
Deep learning catapults the industry
Turing Test Devised 1950
ELIZA 1954 - 1966
Edward Shortliffe writes MYCIN, an Expert or Rule based System, to classify blood disease 1970s
IBM Deep Blue defeats Grand Master Garry Kasparov in chess 1996
ImageNet feeds Deep Learning 2009
AlphaGo defeats Go champion Lee Sedol 2016
1950s
1960s
1970s
1980s
1990s
2000s
2010s

Key semiconductor technology drivers:

- Moore’s Law
- GPUs
- FPGAs
- ASICs

Source: Verhaert, *2019 Perspective on Artificial Intelligence Evolution*

SDR & AI Payload Convergence


Cognitive Radio
Cognitive Radar
Cognitive EW
Cognitive SIGINT
Multi-Mission Situationally Aware Payload:
Enabled by SDR and AI Technology

End of the Line for Processor Performance?

DENNARD SCALING

Power Density Rises

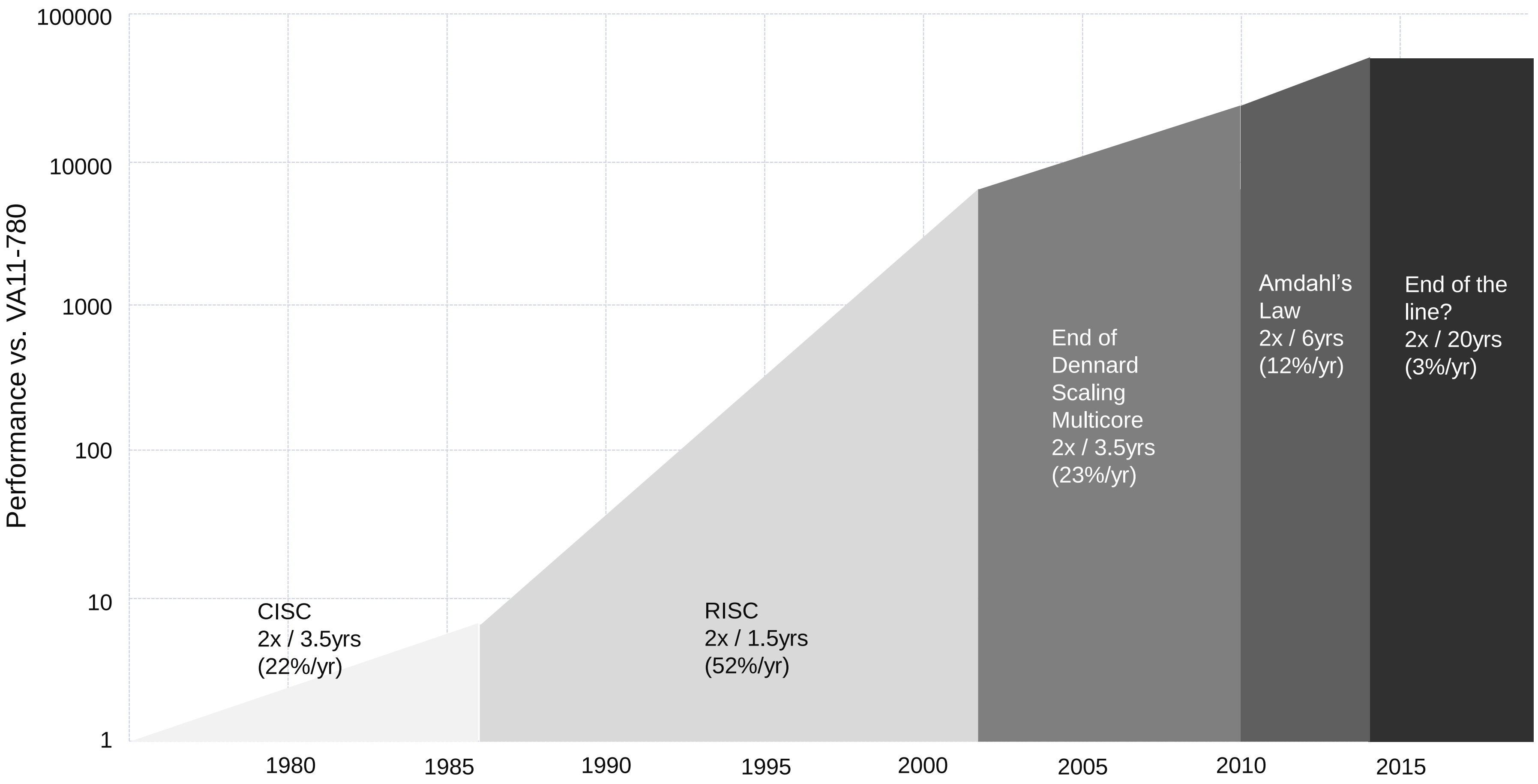
MOORE’S LAW

End of “PPA” Improvement

AMDAHL’S LAW

Multicore Hits Limit

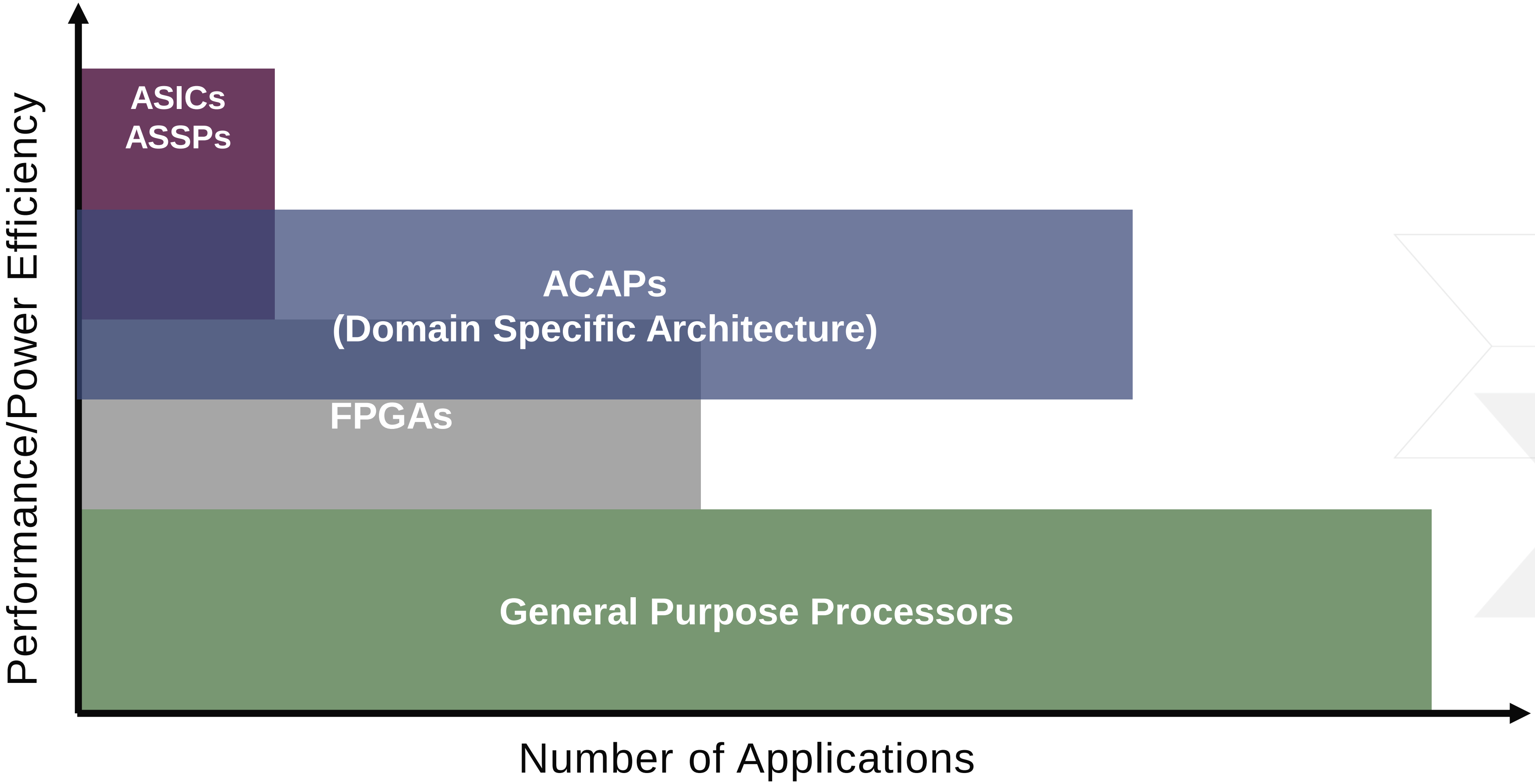
40 Years of Processor Performance



Source: John Hennessy and David Patterson, *Computer Architecture: A Quantitative Approach*, 6/e. 2018

Moving Forward: Domain-Specific Architectures (DSAs)

Evolving Processor Landscape


ASICs
ASSPs
ACAPs
(Domain Specific Architecture)
FPGAs
General Purpose Processors
Performance/Power Efficiency
Number of Applications

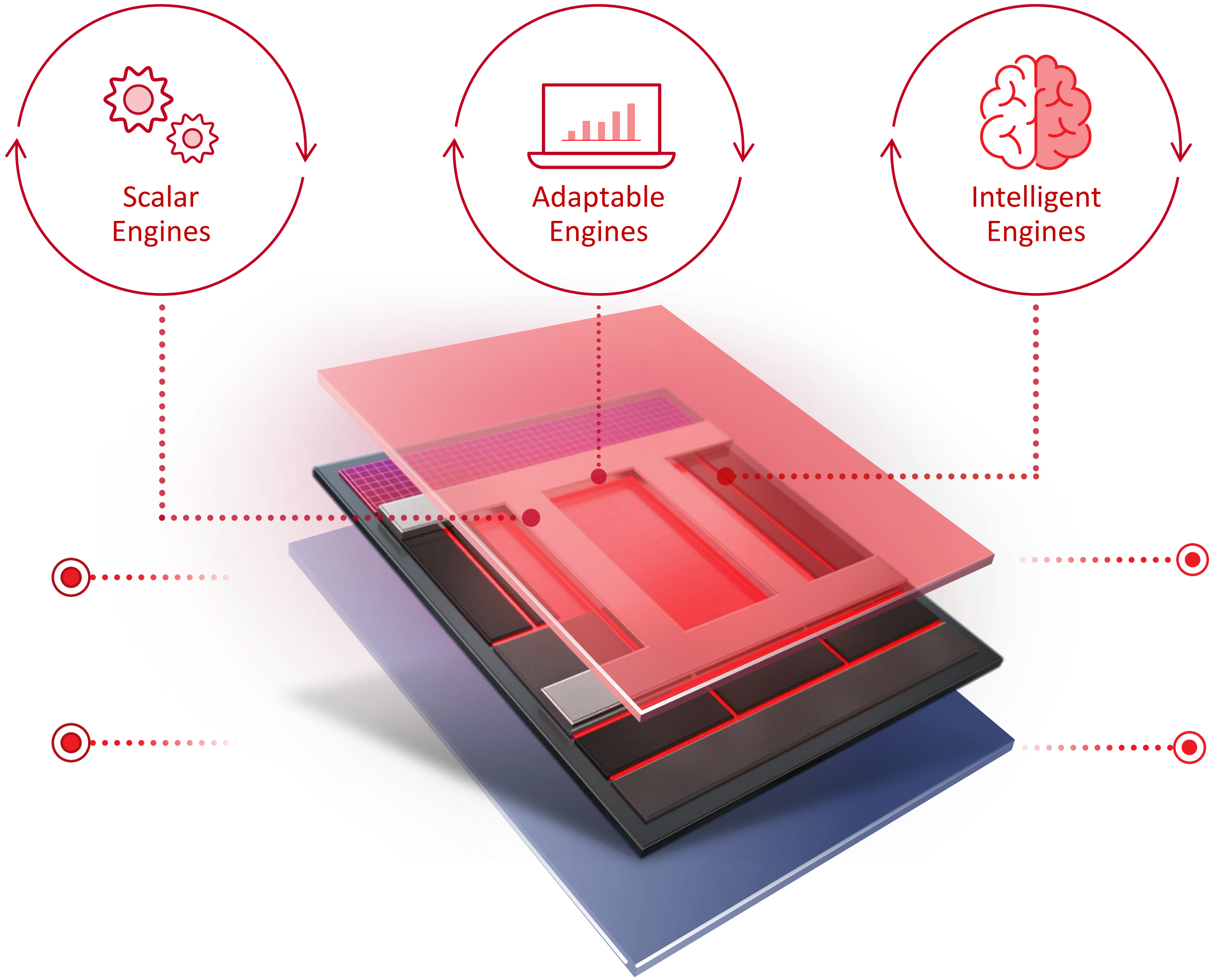
The Adaptive Compute Acceleration Platform

COMPUTE ACCELERATION

Multi-core Processing System

Programmable Logic

DSP (Vector-based & Fabric-based)



Scalar Engines

Adaptable Engines

Intelligent Engines

ADAPTIVE

Diverse Workloads in Milliseconds

Future-Proof for New Algorithms

PLATFORM

Development Tools
HW/SW Libraries
Run-time Stack

SW Programmable Silicon Infrastructure

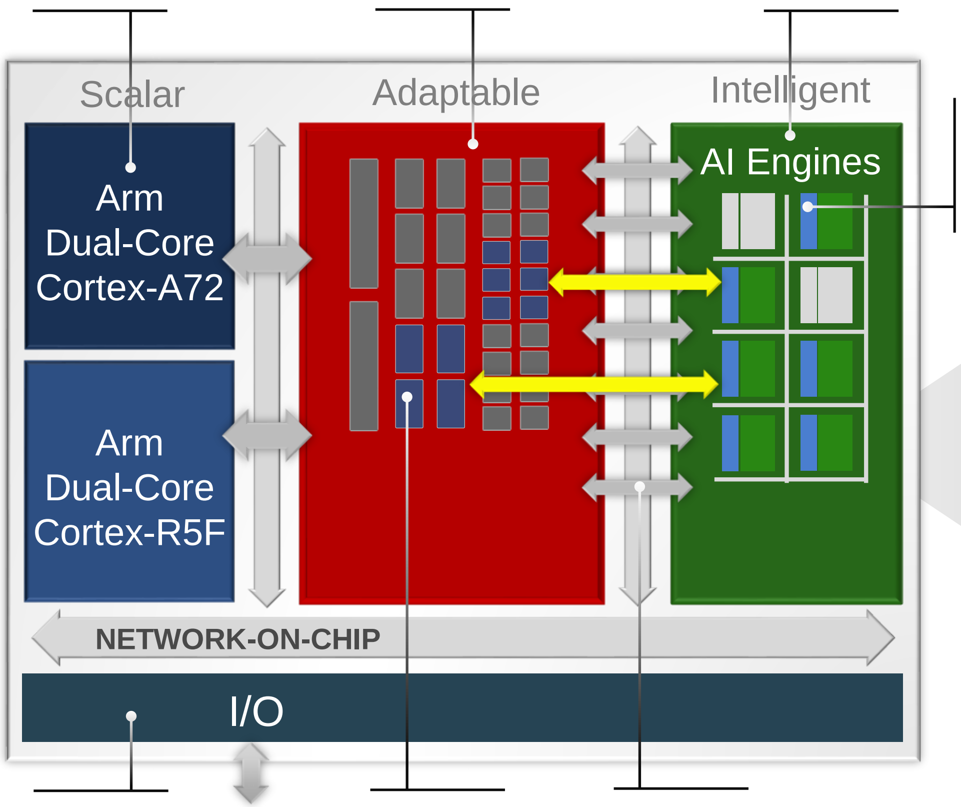
Enabling Data Scientists, SW Developers, HW Developers

Hardware Adaptable: Accelerating the Whole Application

Scalar, Sequential & App Processing

Flexible Parallel Compute, Data manipulation

ML & Signal Processing Vector, Compute Intensive



128 GB/s of Memory B/W per Core

Any-to-Any Connectivity

Custom Memory Hierarchy

TB/s of Bandwidth PL-to-AI Engine

Robust Device & Run-time Security

Heterogeneous Processing For Tactical Edge Systems

(Example Applications)

- Adaptive Beamforming
- AJ
- Tactical Networking
- SAR Backprojection
- Spectrum Processing
- Machine Learning

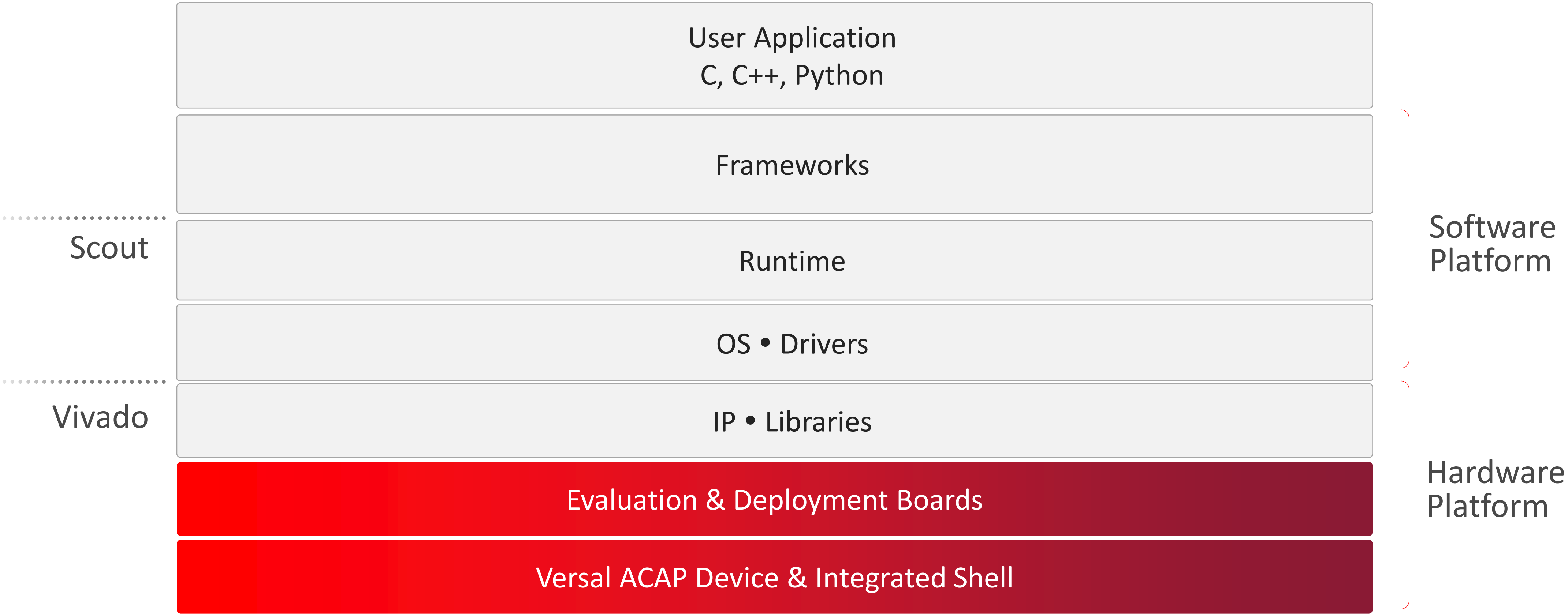
Applications are combined into Domain Specific Architectures (DSAs)

Delivering Deterministic Performance & Low Latency

Versal ACAP:
A Platform for Software *and* Hardware Developers

Fully Software Programmable

with Hardware Design Path



User Application
C, C++, Python

Frameworks

Runtime

OS • Drivers

IP • Libraries

Evaluation & Deployment Boards

Versal ACAP Device & Integrated Shell

Scout

Vivado

Software Platform

Hardware Platform

Possible Platform Example: Multi-Mission Situationally Aware UAV Payload with Versal ACAP

UAV Platform

Multi-Mission Applications: Comms, Radar, SIGINT, EW

Frameworks

TensorFlow PYTORCH

Xilinx Runtime (XRT)

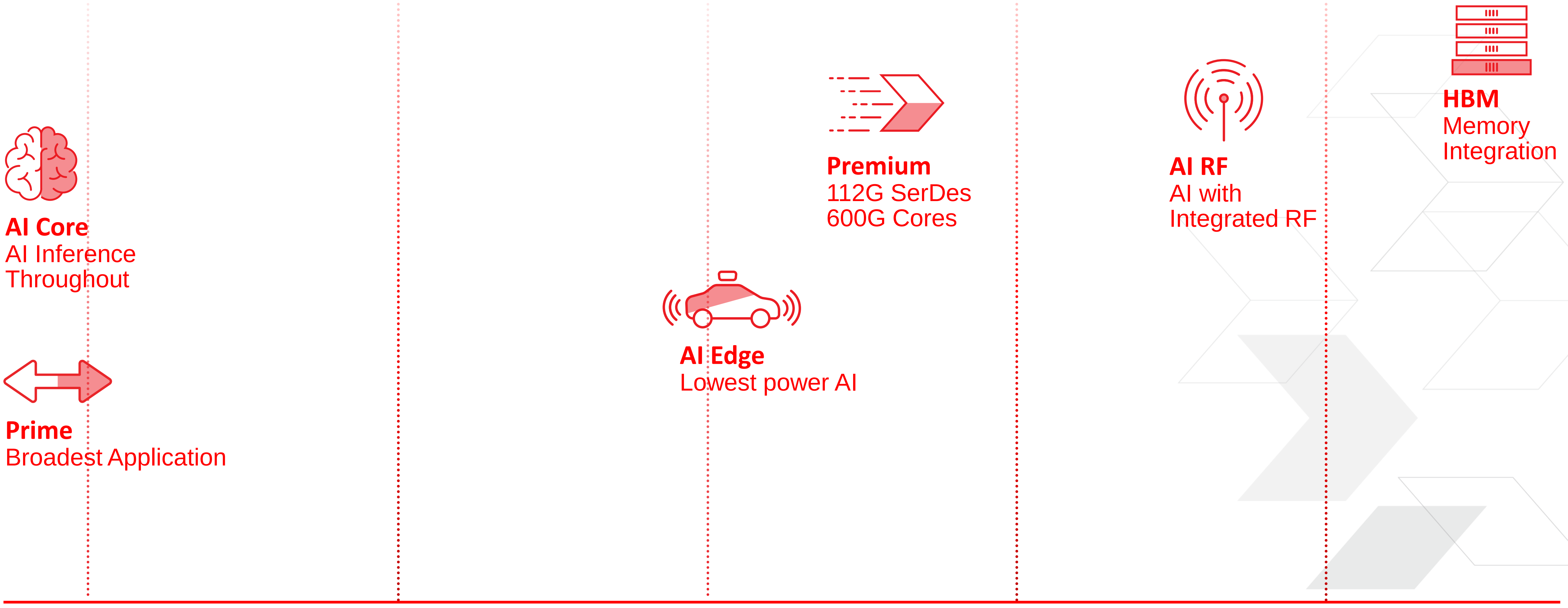
VxWorks | INTEGRITY SECURITY SERVICES | QNX SOFTWARE SYSTEMS | xfopenCV | DSPlib | ML Overlay

Scalar Engines | Adaptable Engines | AI Engines

Versal ACAP Eval Board

VERSAL ACAP

Versal ACAP Roadmap



AI Core
AI Inference
Throughout

Prime
Broadest Application

AI Edge
Lowest power AI

Premium
112G SerDes
600G Cores

AI RF
AI with
Integrated RF

HBM
Memory
Integration

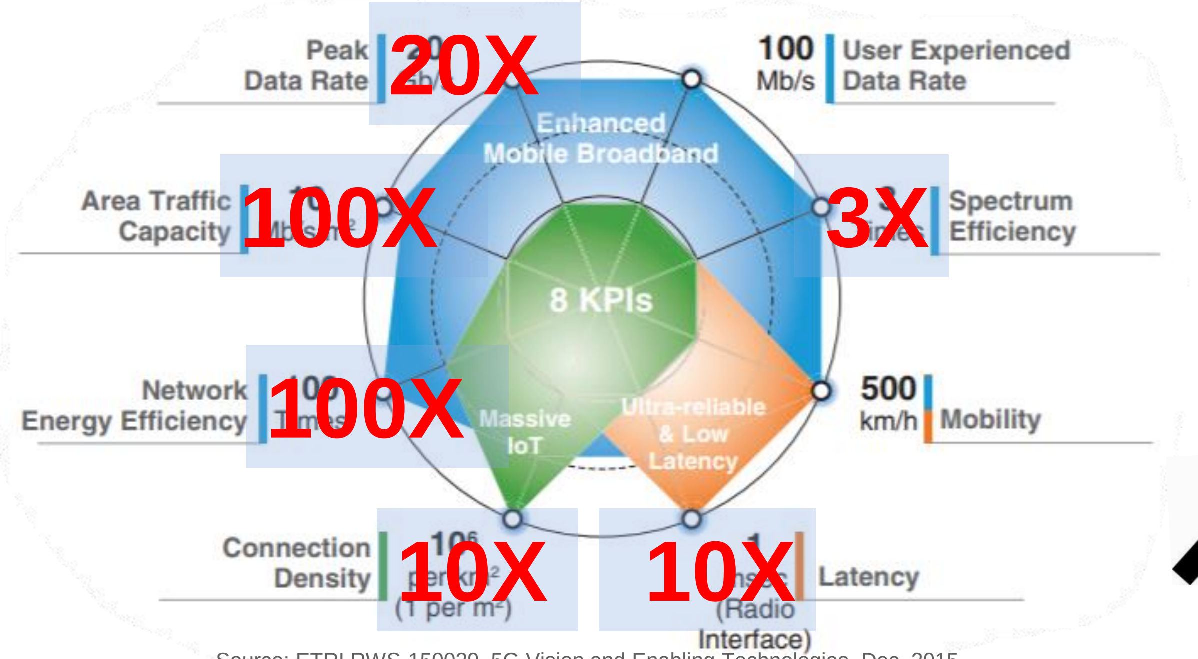
Advanced SDR: Technologies and Challenges

XILINX

Trends in SDR Pushing the Compute Boundary

[CAPACITY]

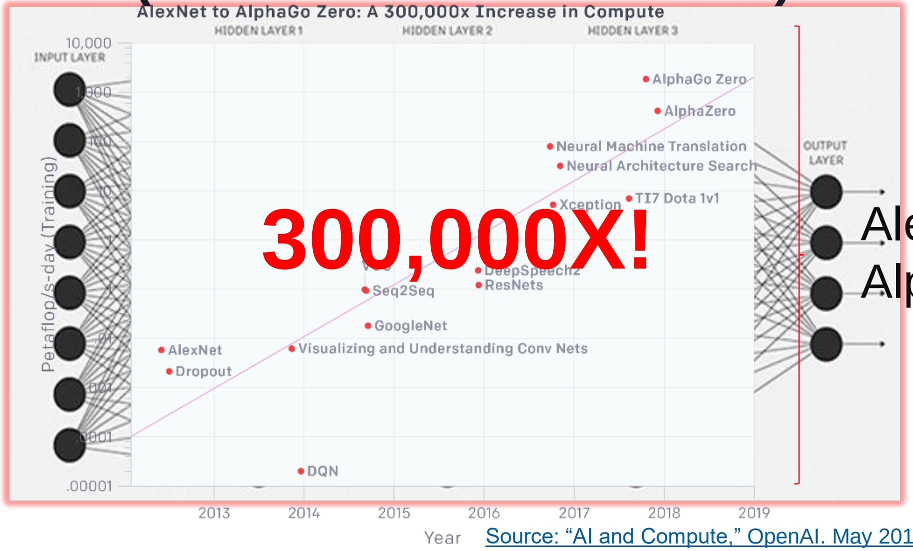
5G 100X Complexity[1] vs. 4G



Source: ETRI RWS-150029, 5G Vision and Enabling Technologies, Dec. 2015.

[AUTONOMY]

Rise of Deep Learning (Dawn of Next Wave of AI)



Source: "AI and Compute," OpenAI. May 2018.

AlexNet to AlphaGo Zero

[RESILIENCY]

Operations in Contested Spectrum

Enabling Technologies

- Direct-RF / High-IF Sampling Data Converters
- Array Antennas
- Compute Optimizations for Deep Learning

Deep Learning Classification

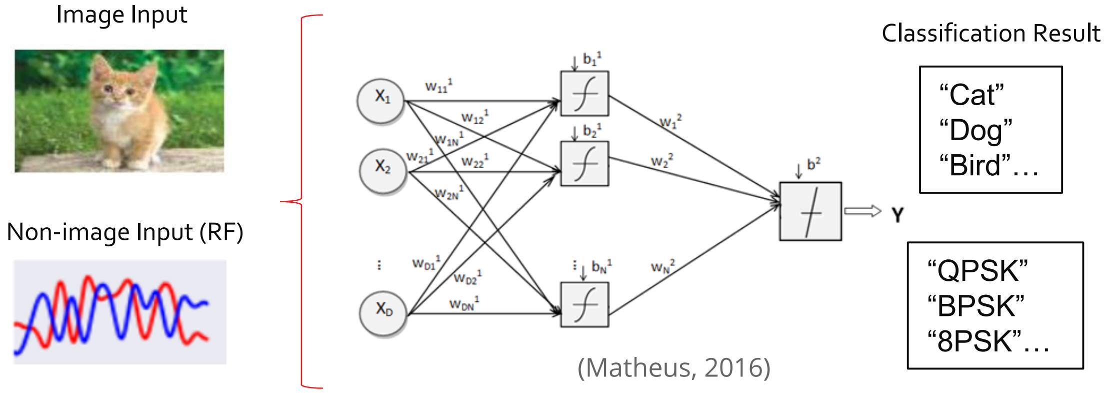


Image Input

Non-image Input (RF)

Classification Result

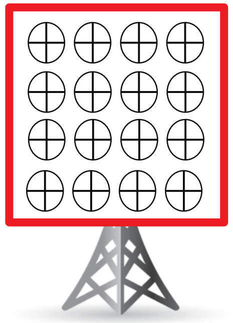
"Cat"
"Dog"
"Bird"…

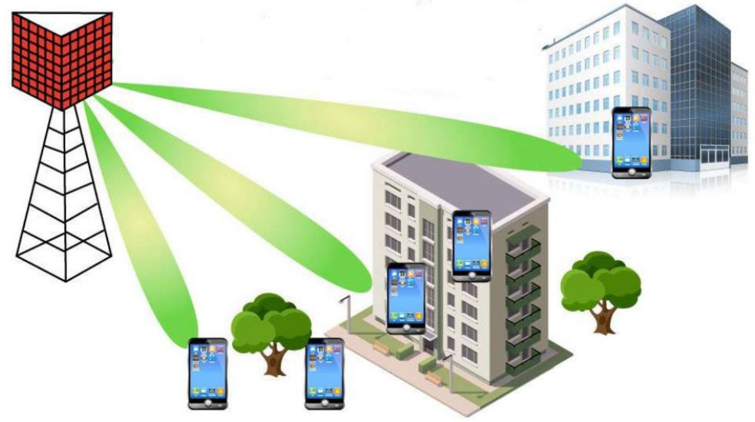
"QPSK"
"BPSK"
"8PSK"…

(Matheus, 2016)

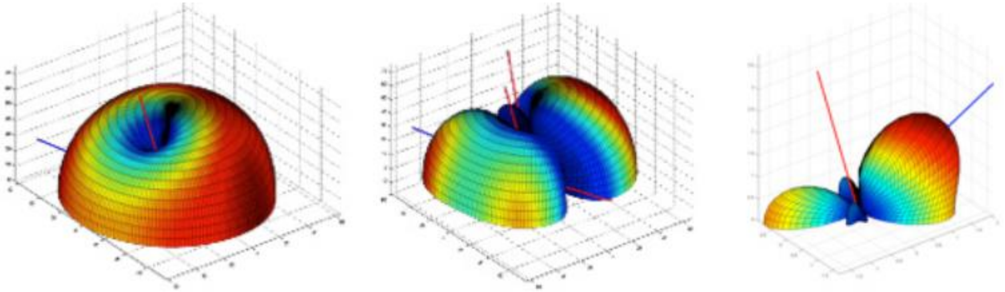
Animation credit: Philip Leone, Univ. of Sydney. Presentation.

Array Antennas





mMIMO Spatial Multiplexing and Beamforming (5G).

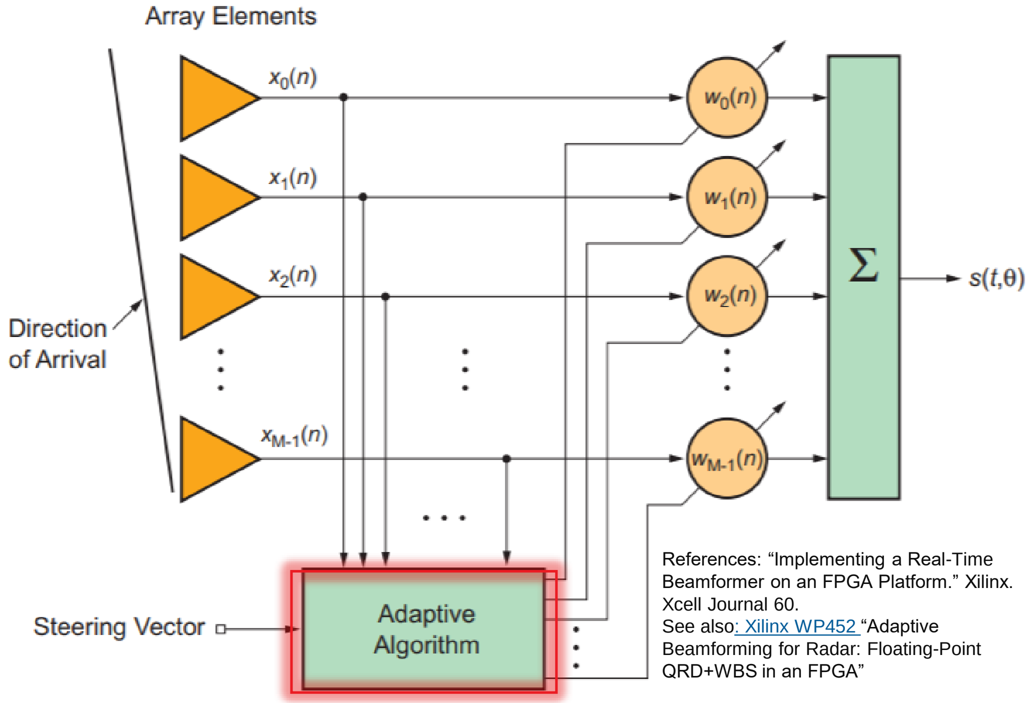


Controlled Reception Pattern Array (CRPA) beam patterns.
(source: gpsworld.com)

Advanced SDR: Compute Comparisons

Space Time Adaptive Processing

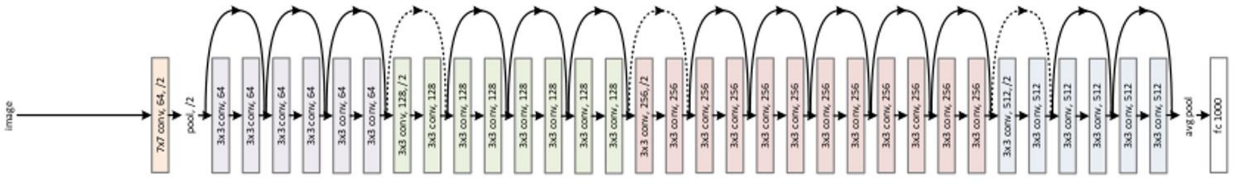
Application Example: Beamforming/Nulling (Comms / Anti-Jam)



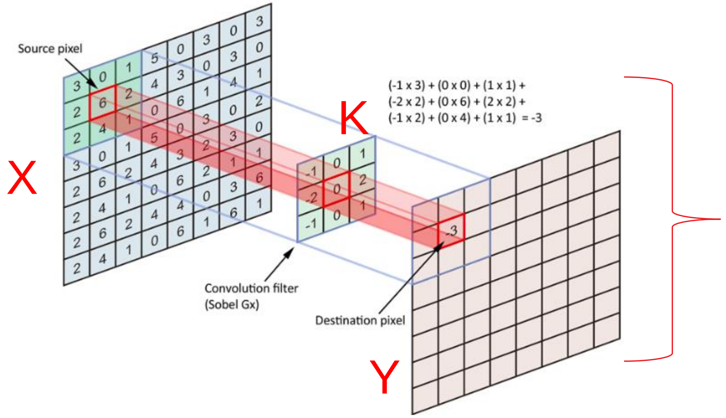
References: “Implementing a Real-Time Beamformer on an FPGA Platform.” Xilinx. Xcell Journal 60.
See also: Xilinx WP452 “Adaptive Beamforming for Radar: Floating-Point QRD+WBS in an FPGA”

Deep Learning Inference (Conv. Nets)

Application: Modulation Recognition, Waveform Classification



Resnet-50 visualization. Kaggle.com



Source pixel
X
K
Y
Convolution filter (Sobel Gx)
Destination pixel
(-1 x 3) + (0 x 0) + (1 x 1) +
(-2 x 2) + (0 x 6) + (2 x 2) +
(-1 x 2) + (0 x 4) + (1 x 1) = -3

Convolutional Layer Processing

References: “Applied Deep Learning - Part 4: Convolutional Neural Networks”, Towards Data Science (blog).

$$W = \underbrace{R_{xx}^{-1}}_{} * b$$

Steering Vector

Covariance Matrix Decomposition: QR, Cholesky, etc.

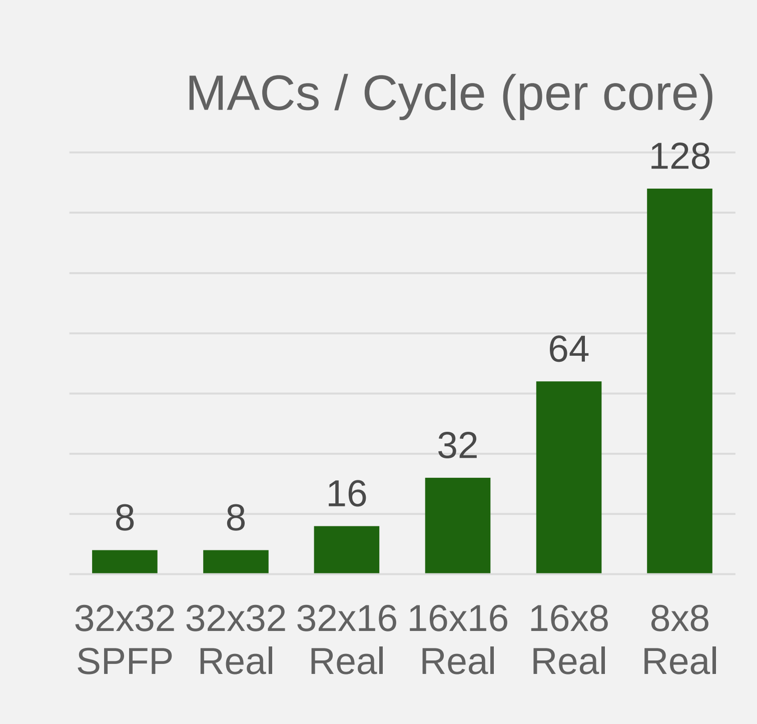
$$Y = X * K$$

Complex-valued
Higher Precision Desirable (e.g., SPFP32)
Typical FLOPS: up to **MFLOPS** per Decomposition

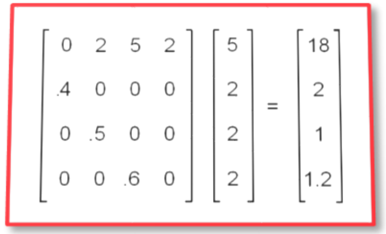
Real-valued
Lower Precision Desirable (e.g., INT8)
Typical OPS: 7.6 **GOPS** (Resnet-50 unpruned)

AI Engine: Multi-Precision Math Support

Real Data Types



Optimized For:

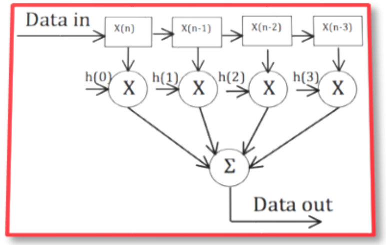


$$\begin{bmatrix} 0 & 2 & 5 & 2 \\ .4 & 0 & 0 & 0 \\ 0 & .5 & 0 & 0 \\ 0 & 0 & .6 & 0 \end{bmatrix} \begin{bmatrix} 5 \\ 2 \\ 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 18 \\ 2 \\ 1 \\ 1.2 \end{bmatrix}$$

Linear Algebra

Matrix-Matrix Mult

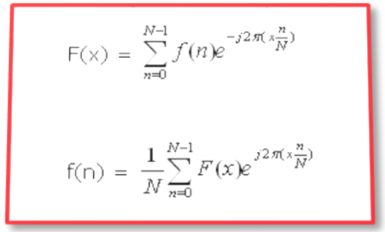
Matrix-Vector Mult



Convolution

FIR Filters

2-D Filters



$$F(x) = \sum_{n=0}^{N-1} f(n) e^{-j2\pi(x\frac{n}{N})}$$

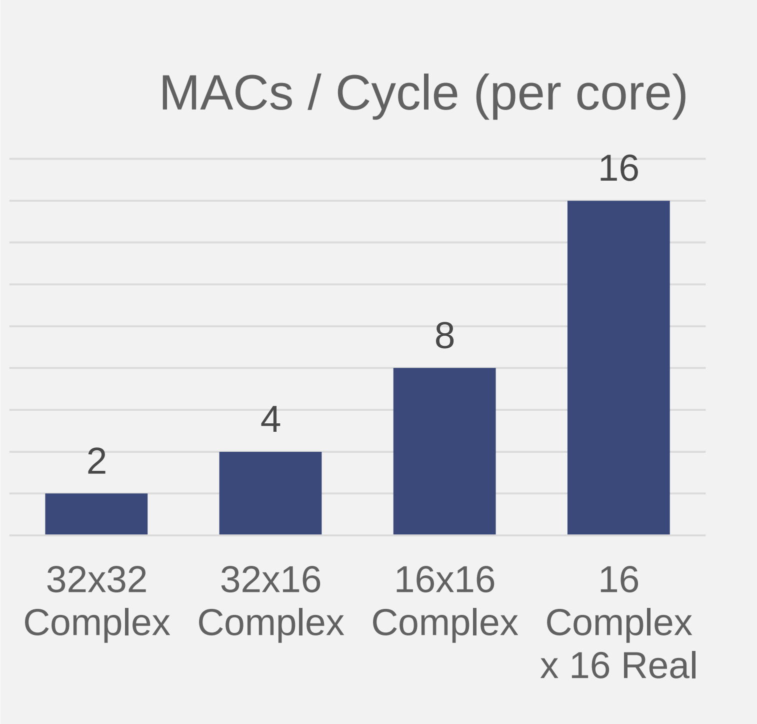
$$f(n) = \frac{1}{N} \sum_{n=0}^{N-1} F(x) e^{j2\pi(x\frac{n}{N})}$$

Transforms

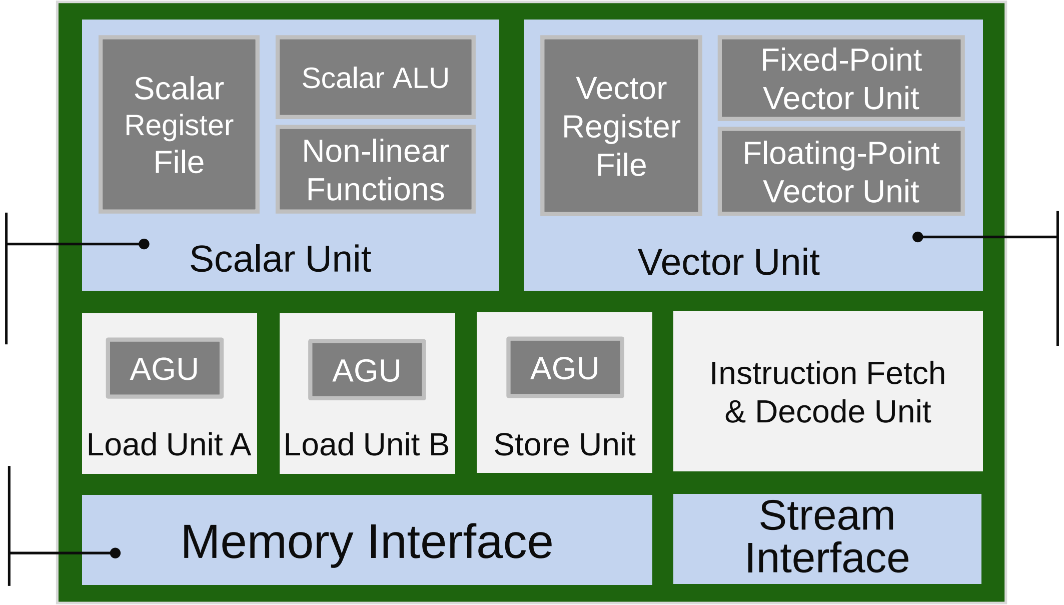
FFTs/IFFTs

DCT, etc

Complex Data Types



AI Engine: Scalar Unit, Vector Unit, Load Units and Memory



32-bit Scalar RISC Processor

Vector Processor
512-bit SIMD Datapath

Local, Shareable Memory

- 32KB Local, 128KB Addressable

Instruction Parallelism: VLIW

7+ operations / clock cycle

- 2 Vector Loads / 1 Mult / 1 Store
- 2 Scalar Ops / Stream Access

Highly Parallel

Data Parallelism: SIMD

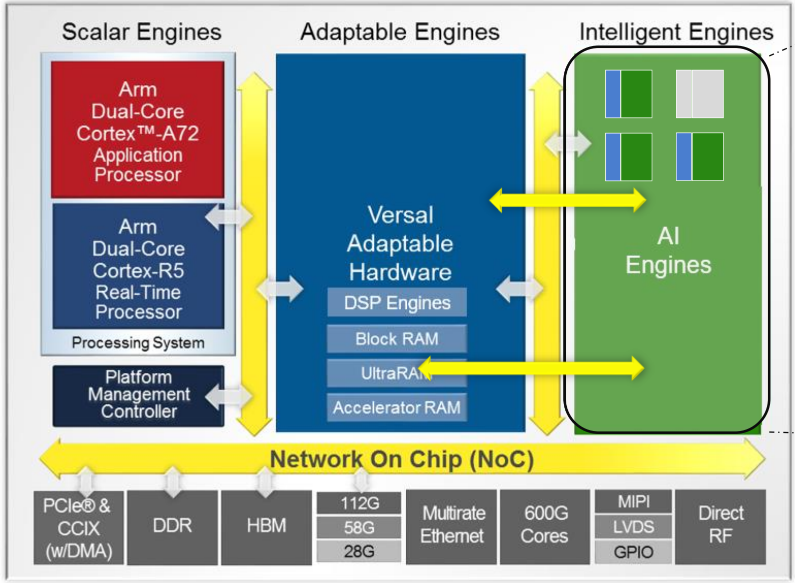
Multiple vector lanes

- Vector Datapath
- 8 / 16 / 32-bit & SPFP operands

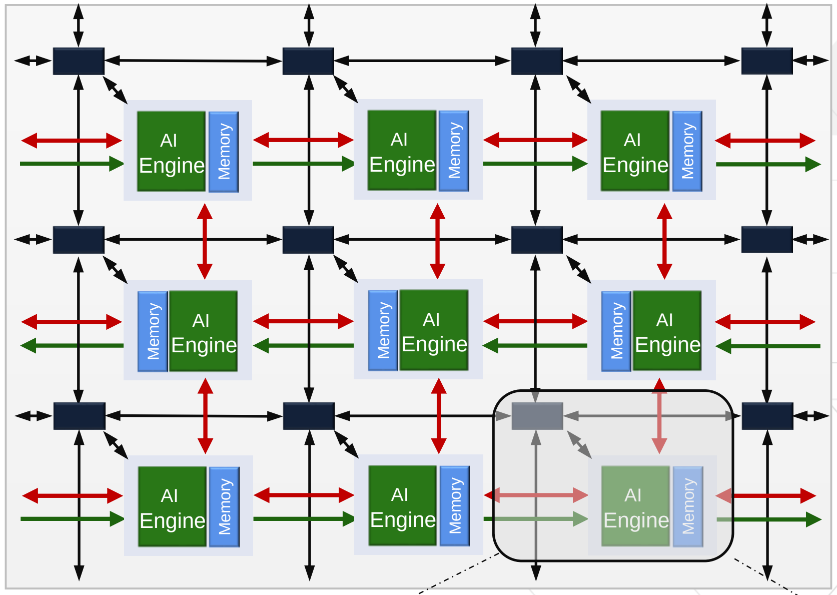
Up to 128 MACs / Clock Cycle per Core (INT 8)
8 FLOPs / Clock Cycle (32SPFP)

AI Engine: Terminology

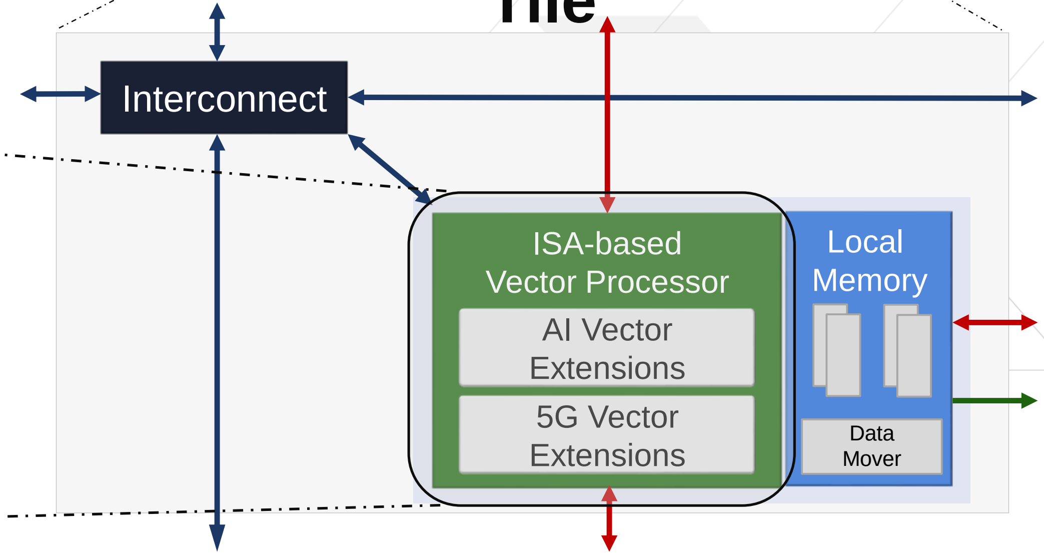
Versal ACAP



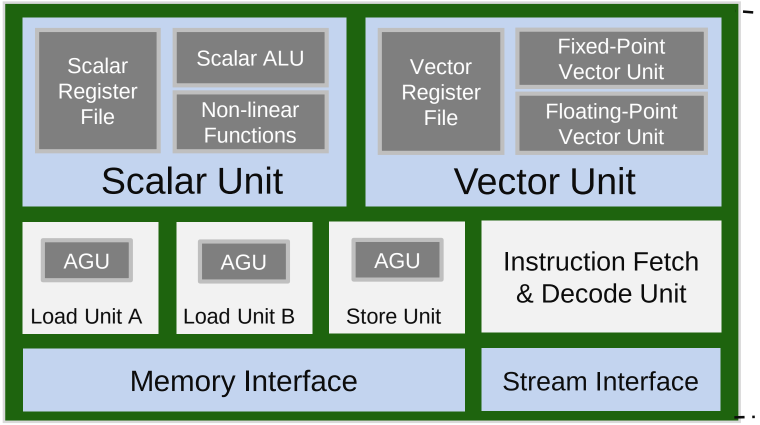
AI Engine Array



AI Engine Tile



1GHz+ VLIW / SIMD vector processor



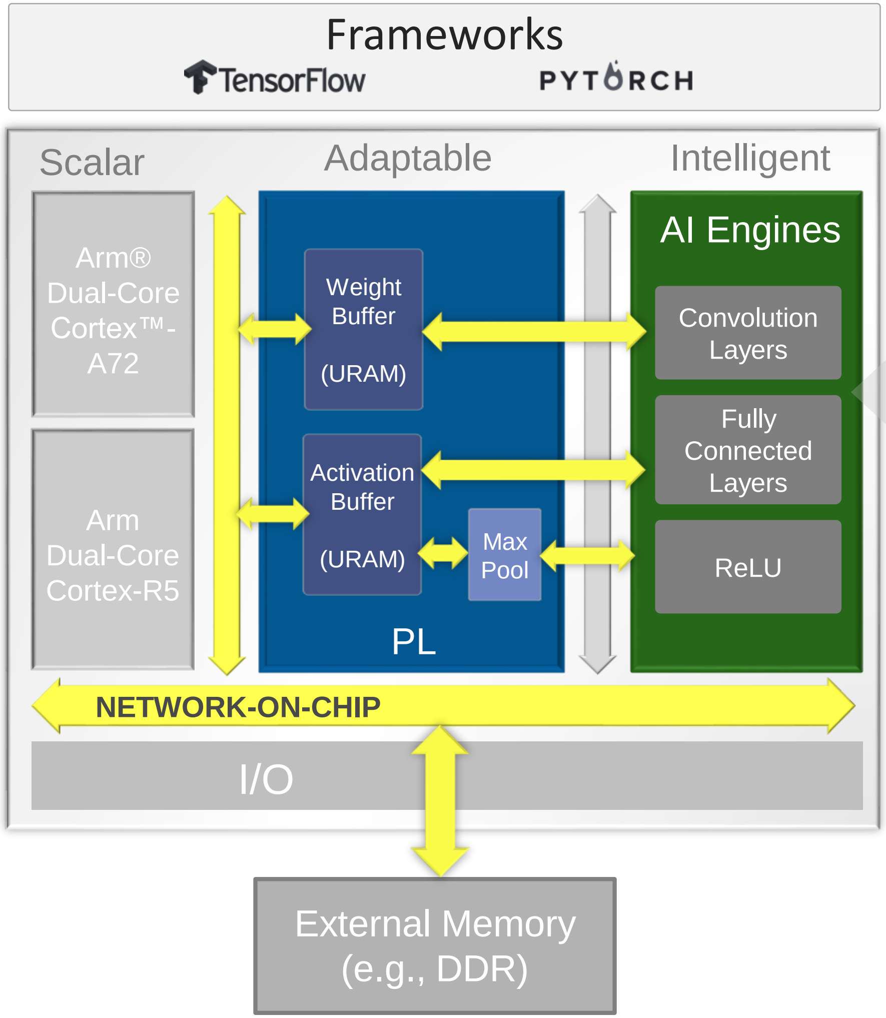
AI Engine

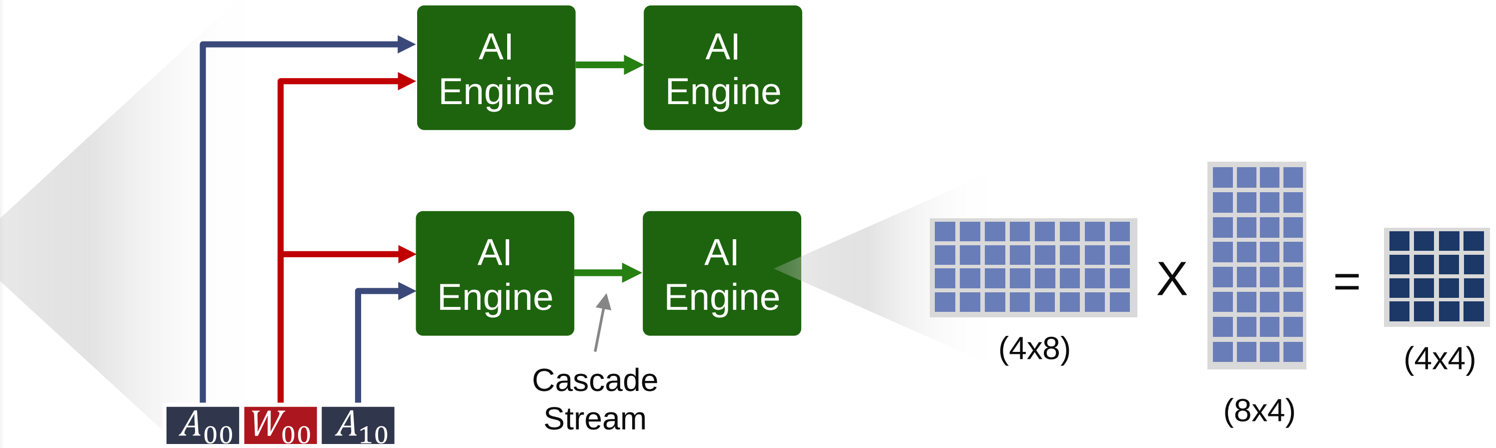
AI Inference Mapping on Versal™ ACAP

Program Directly From High-level ML Frameworks

A = Activations
W = Weights

$$\begin{bmatrix} A_{00} & A_{01} \\ A_{10} & A_{11} \end{bmatrix} \times \begin{bmatrix} W_{00} & W_{01} \\ W_{10} & W_{11} \end{bmatrix} = \begin{bmatrix} A_{00} \times \boldsymbol{W_{00}} + A_{01} \times W_{10} & \ldots \\ A_{10} \times \boldsymbol{W_{00}} + A_{11} \times W_{10} & \ldots \end{bmatrix}$$

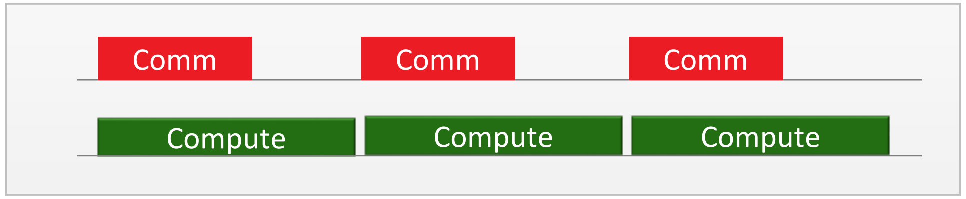




- Custom memory hierarchy
  - Buffer on-chip vs off-chip; Reduce latency and power
- Stream Multi-cast on AI interconnect
  - Weights and Activations
  - Read once: reduce memory bandwidth
- AI-optimized vector instructions (128 INT8 mults/cycle)

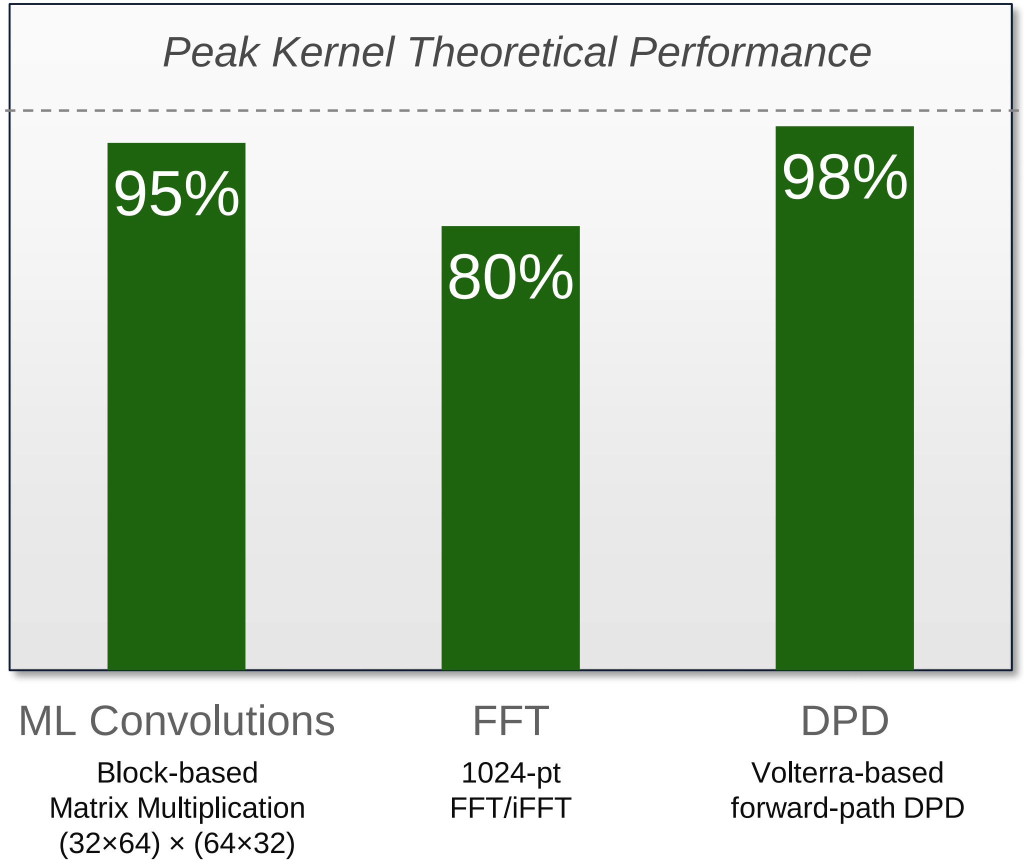
AI Engine Delivers High Compute Efficiency

- Adaptable, non-blocking interconnect
  - Flexible data movement architecture
  - Avoids interconnect “bottlenecks”
- Adaptable memory hierarchy
  - Local, distributed, shareable = extreme bandwidth
  - No cache misses or data replication
  - Extend to PL memory (BRAM, URAM)
- Transfer data while AI Engine Computes



Overlap Compute and Communication

Vector Processor Efficiency



Summary

- The evolution of processing for AI is following a similar track to SDR where hardware and software need to be tightly coupled
- The drive for more Capacity, Autonomy and Resiliency in advanced SDRs carry high compute demands and mixed precision processing capabilities
- Moore’s Law is running out of steam which means the goal of a SWaP-friendly multi-mission situationally aware payload requires advancements in processing beyond just process technology
- ACAPs are a response to this new reality

Visit https://www.xilinx.com/products/silicon-devices/acap/versal.html for datasheets, whitepapers, and product tables.

Xilinx VC1902 Versal ACAP with 400 AI Engines.
First shipment June 2019.

Adaptable.
Intelligent.

THANK YOU!

Contact Info:
manuelu@xilinx.com
jasonv@xilinx.com

XILINX®