WInnComm 2019

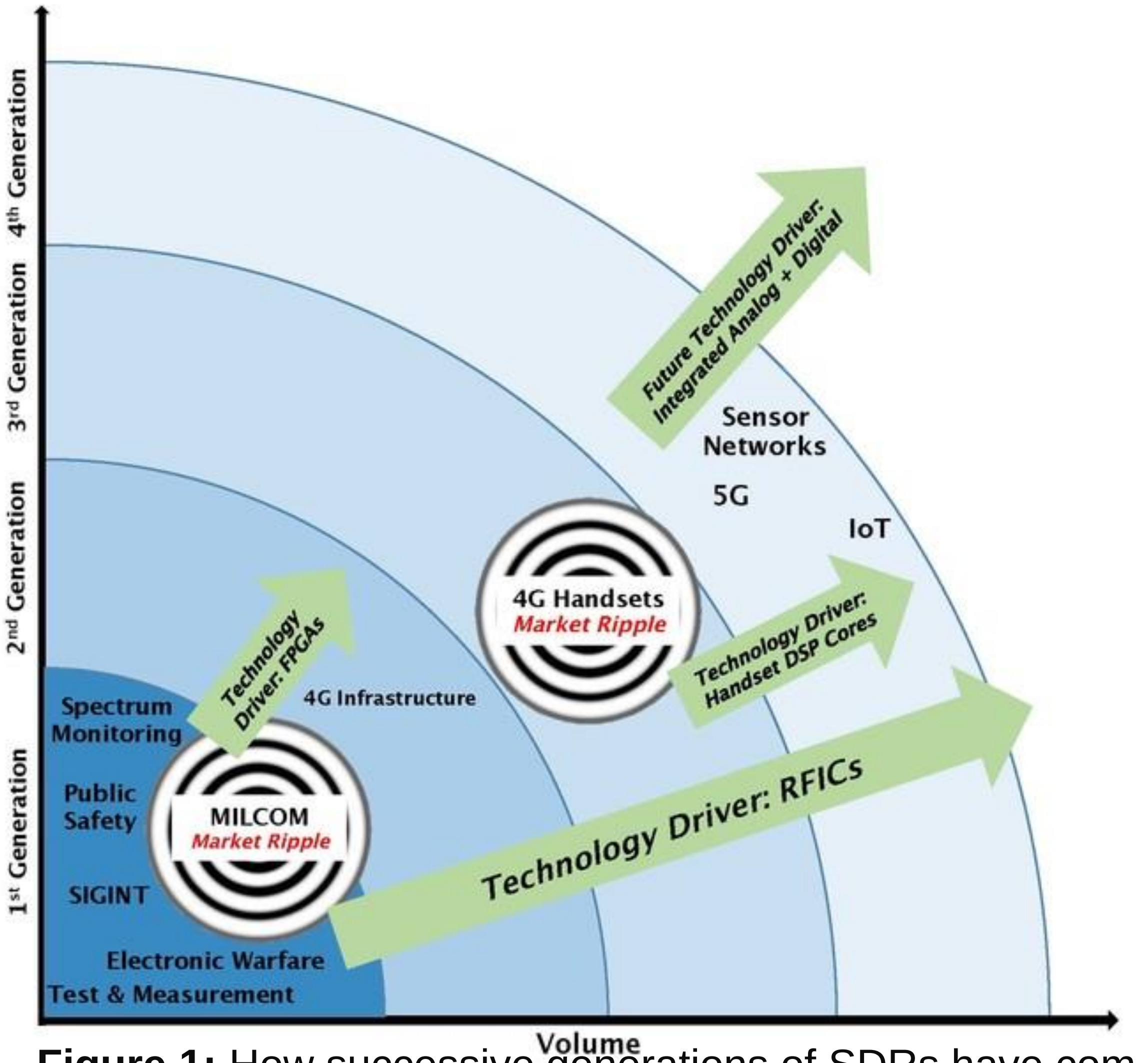
Summit: 20-21 November 2019
Group Meetings: 18-19 November 2019
Atkinson Hall, UCSD, La Jolla, CA

Software Programmable AND Hardware Adaptable: Can You Have Your Cake and Eat It, Too?

Manuel Uhm
Director, Silicon Marketing - Xilinx
Chair of the Board of Directors - Wireless Innovation Forum

XILINX

SDR Evolution



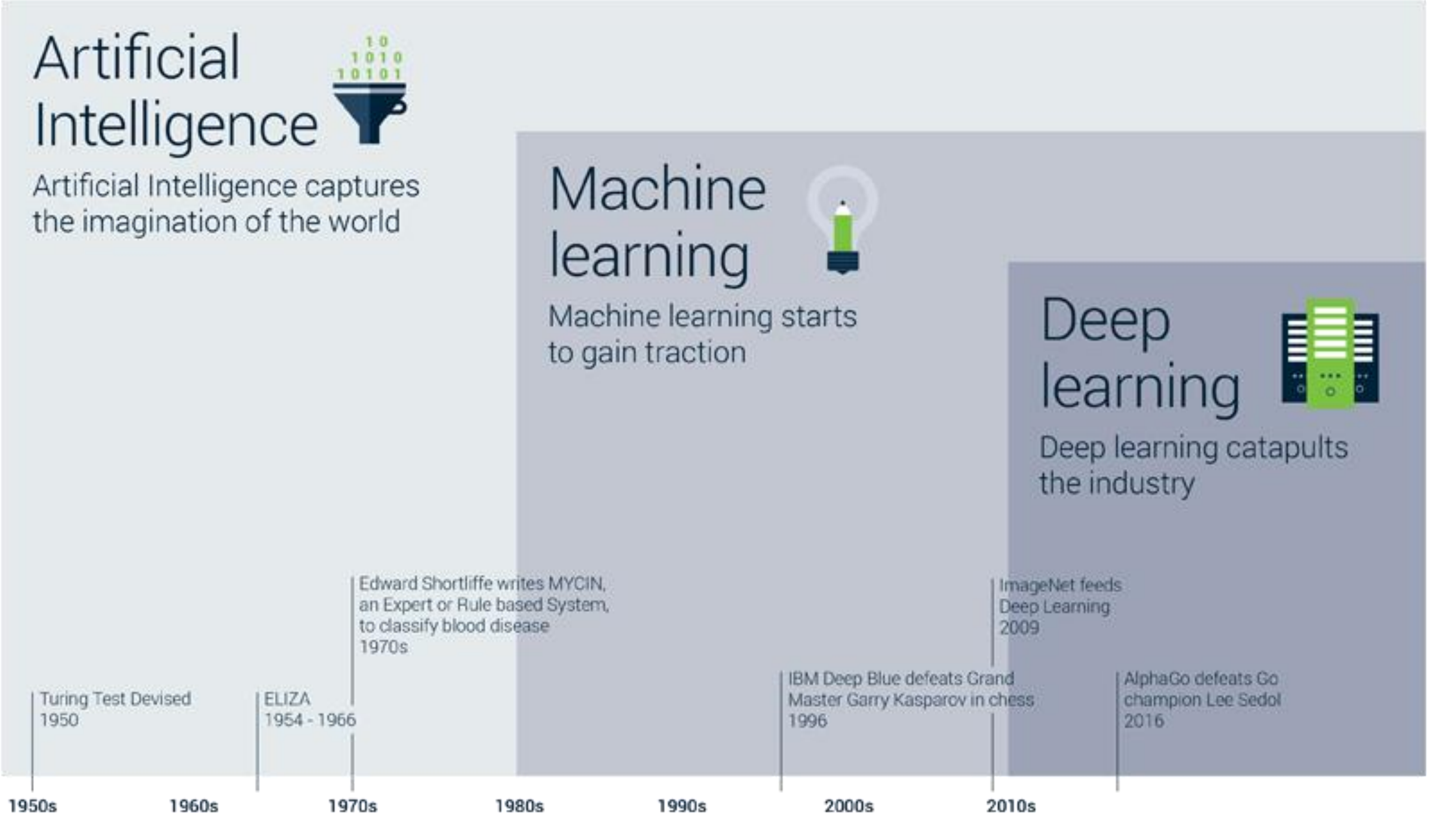
Key semiconductor technology drivers:

- 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



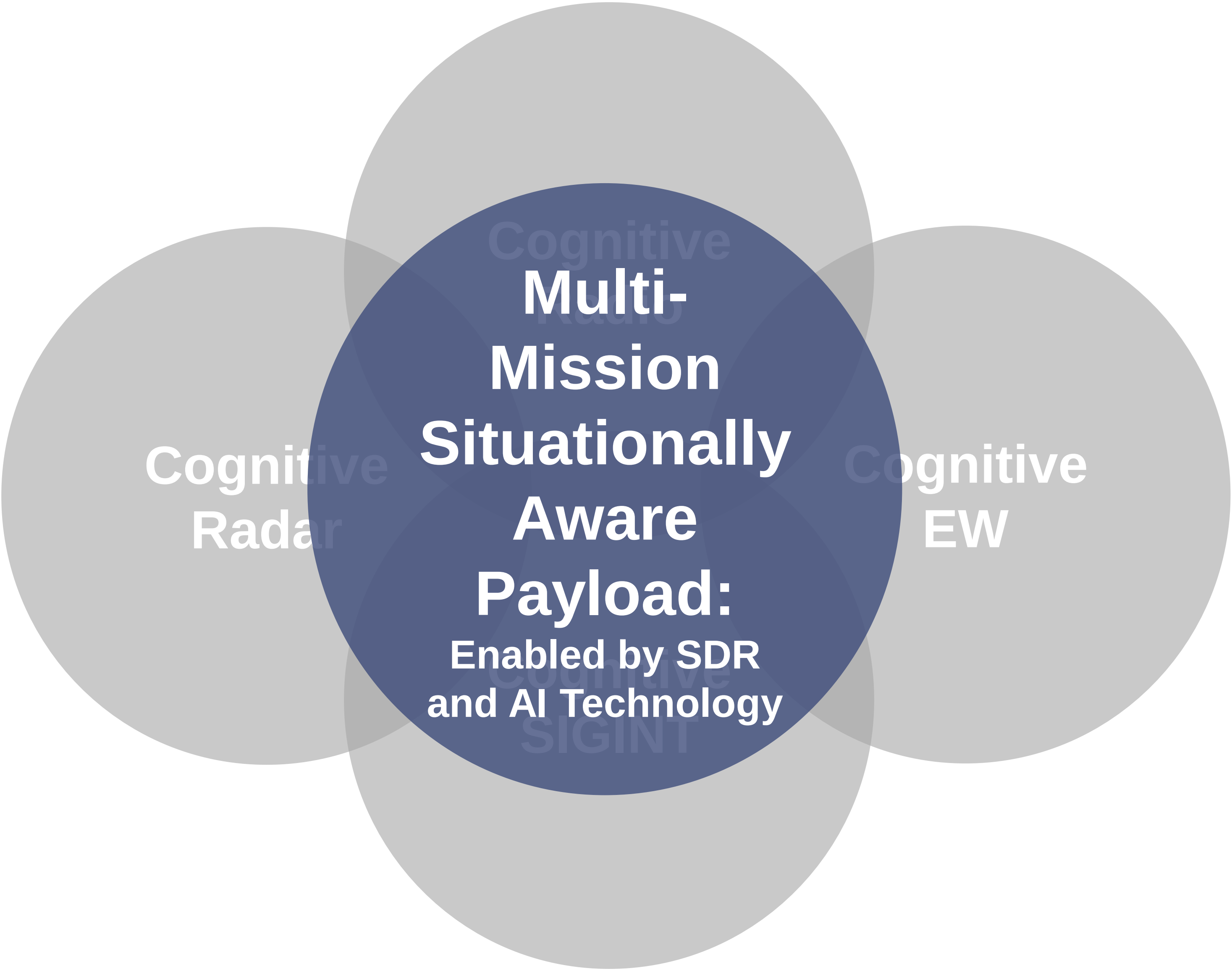
Artificial Intelligence
Artificial Intelligence captures the imagination of the world
Machine learning
Machine learning starts to gain traction
Deep learning
Deep learning catapults the industry
Turing Test Devised 1950
ELIZA 1954 - 1966
Edward Shortliffe writes MYCIN, an Expert or Rule based System, 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

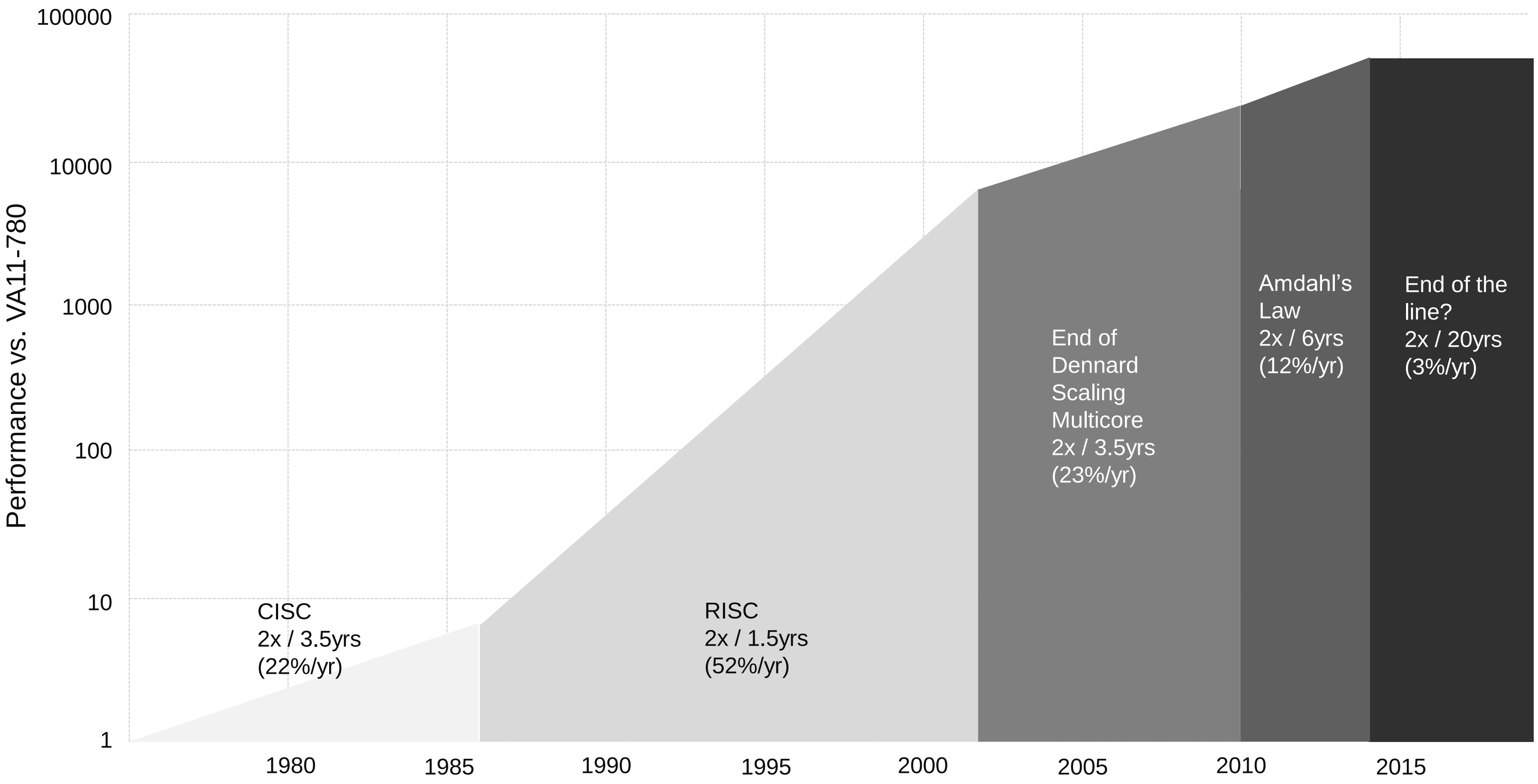
MOORE’S LAW

End of “PPA” Improvement

AMDAHL’S LAW

Multicore Hits Limit

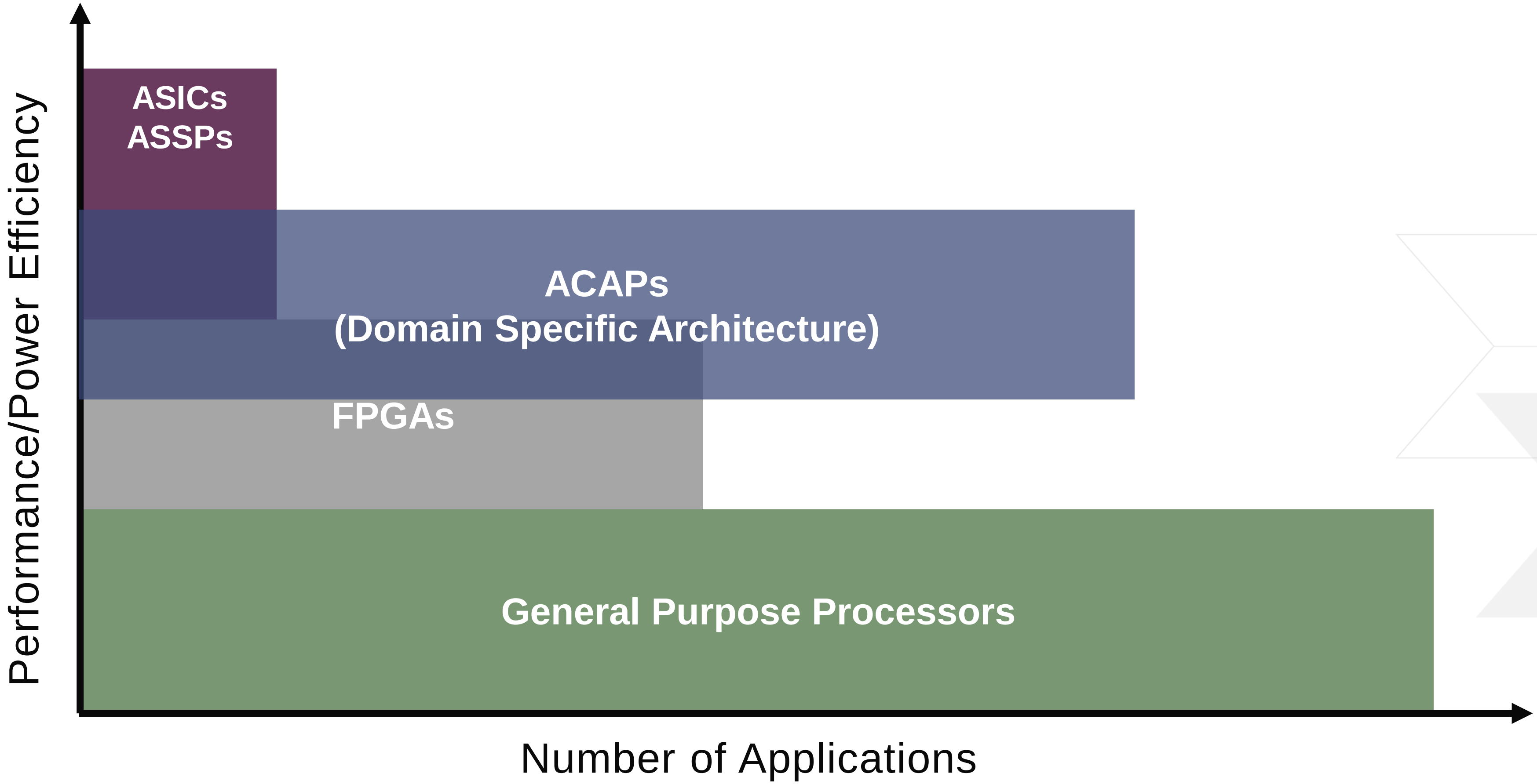
40 Years of Processor Performance



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

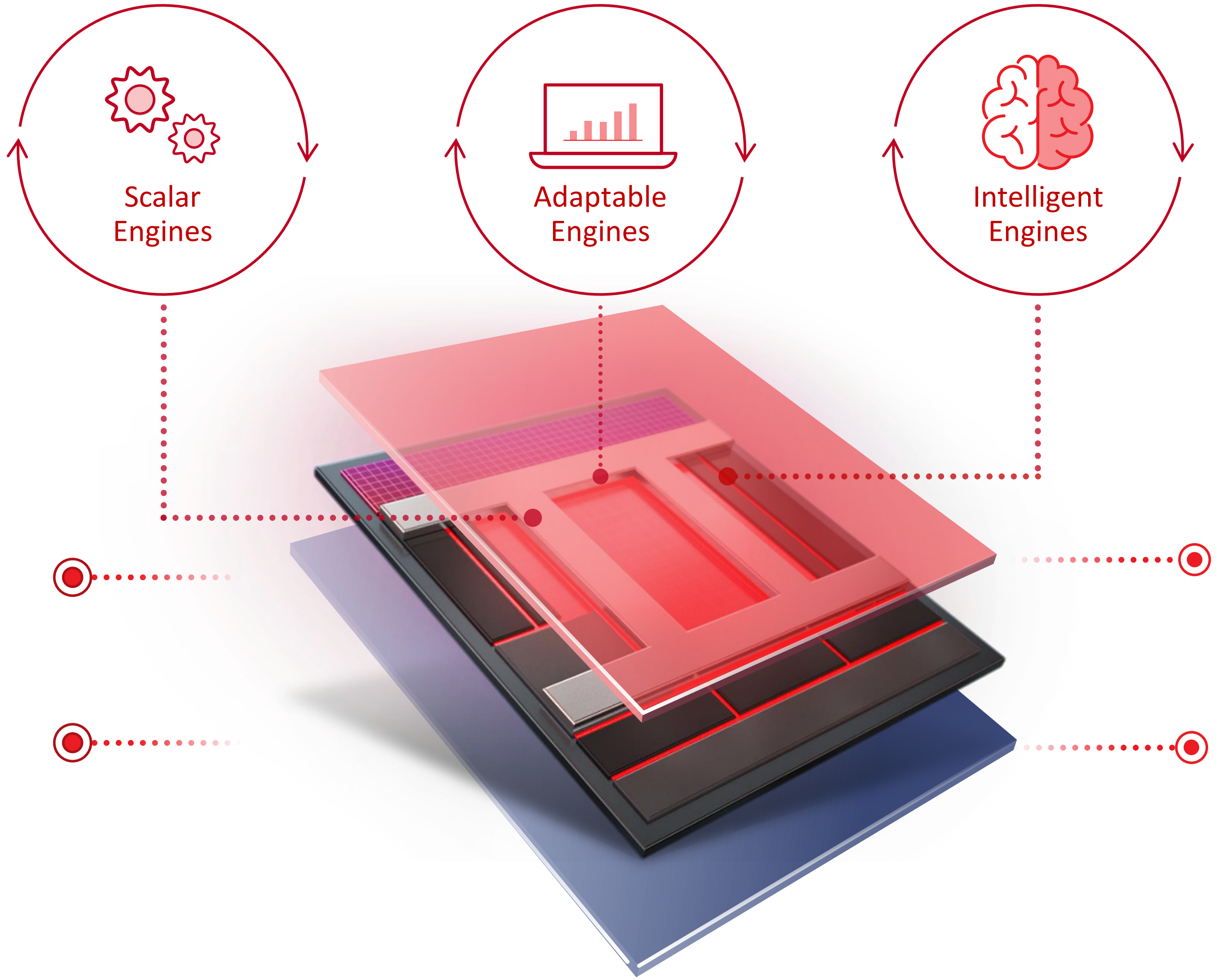
Moving Forward: Domain-Specific Architectures (DSAs)

Evolving Processor Landscape



The Adaptive Compute Acceleration Platform

COMPUTE ACCELERATION



Multi-core Processing System

Scalar Engines

Programmable Logic

Adaptable Engines

DSP (Vector-based & Fabric-based)

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

It’s About the Platform

ADAPTIVE

- > Adaptable to Diverse Workloads
- > Future-proof Algorithms

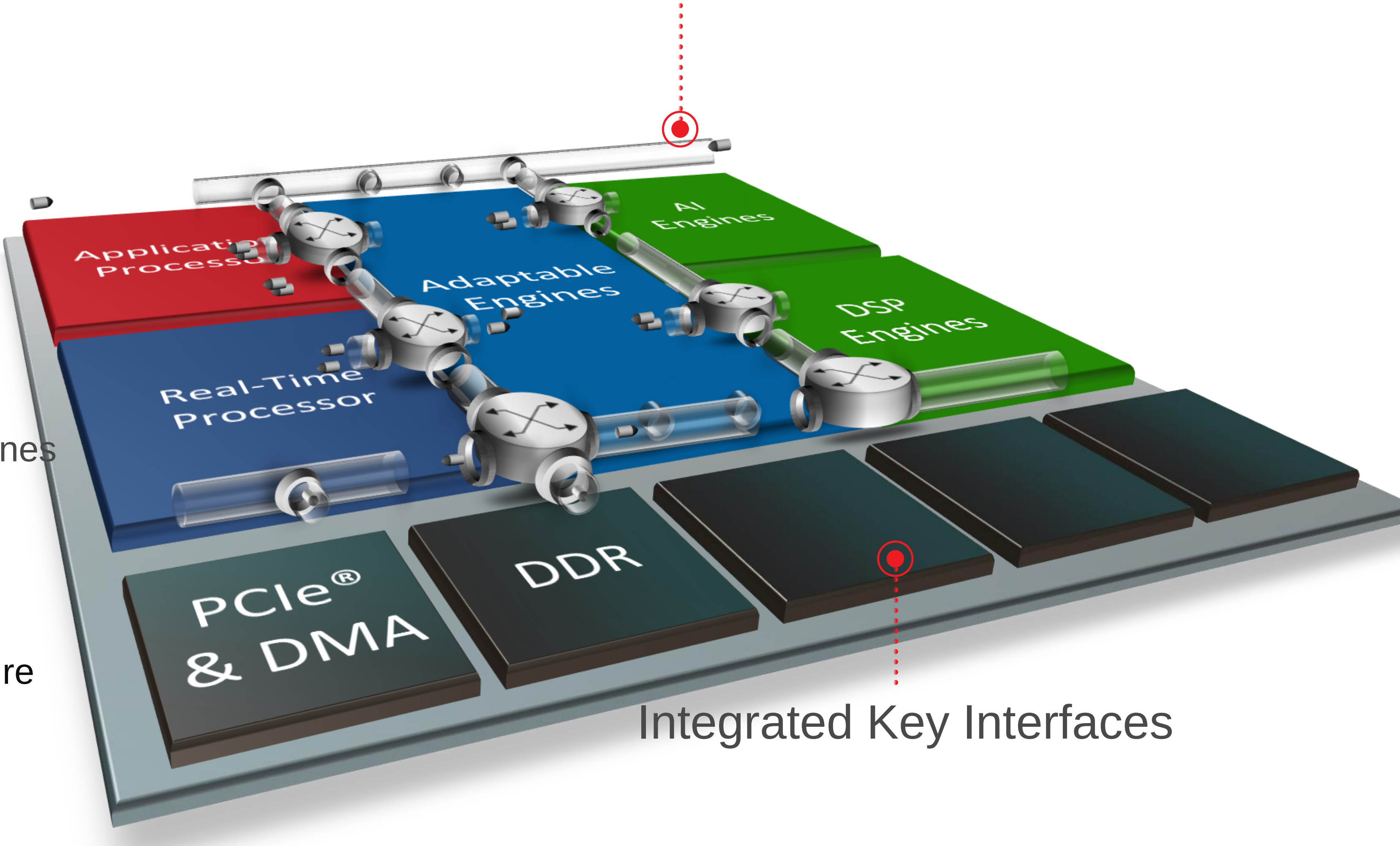
COMPUTE ACCELERATION

- > Heterogeneous architecture
- > Scalar, Adaptable, and Intelligent Engines

PLATFORM

- > SW Programmable Silicon Infrastructure
- > Integrates with Dev Tools & SW Stack

Programmable Network on Chip

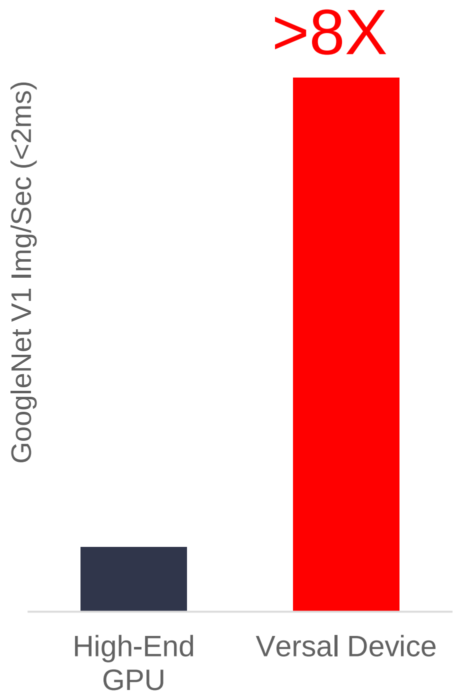


Integrated Key Interfaces

Breakthrough Performance for Cloud, Network, and Edge

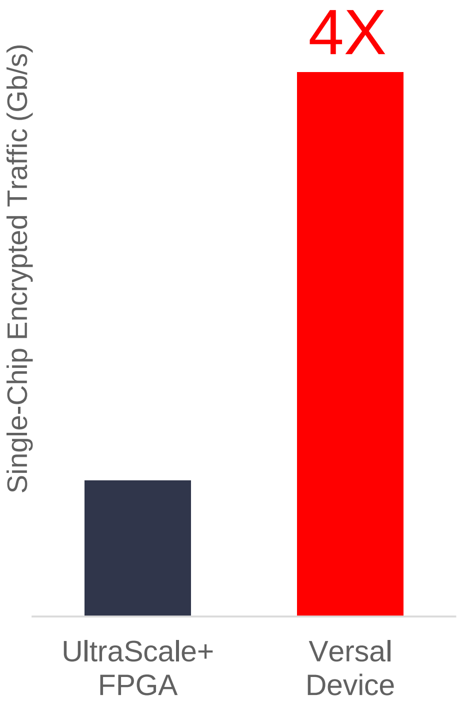
Cloud Compute

Breakthrough AI Inference



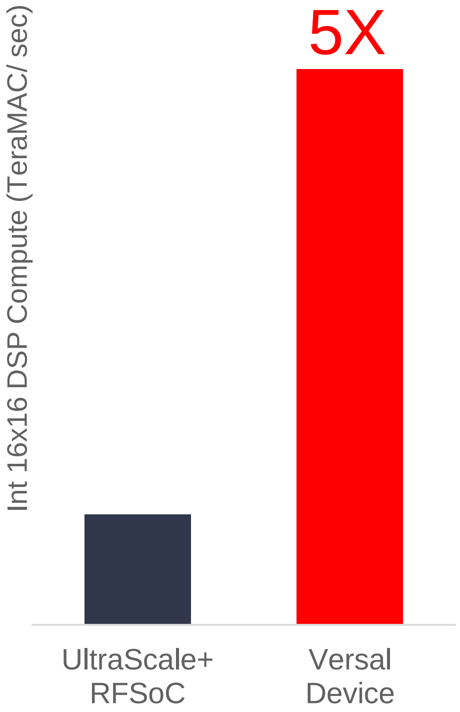
Networking

Multi-terabit Throughput



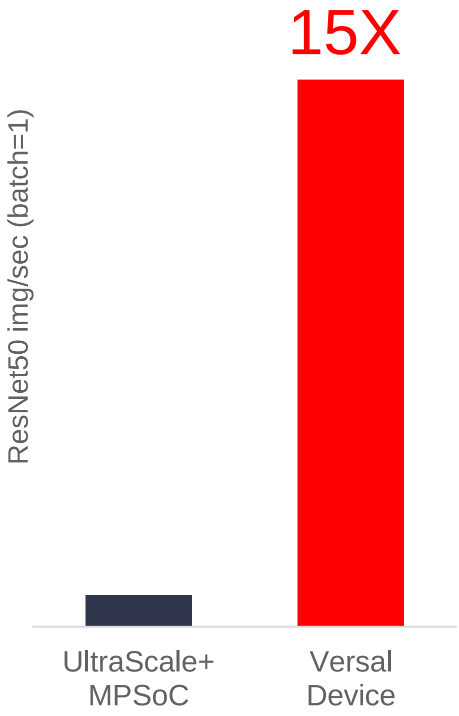
5G Wireless

Compute for Massive MIMO

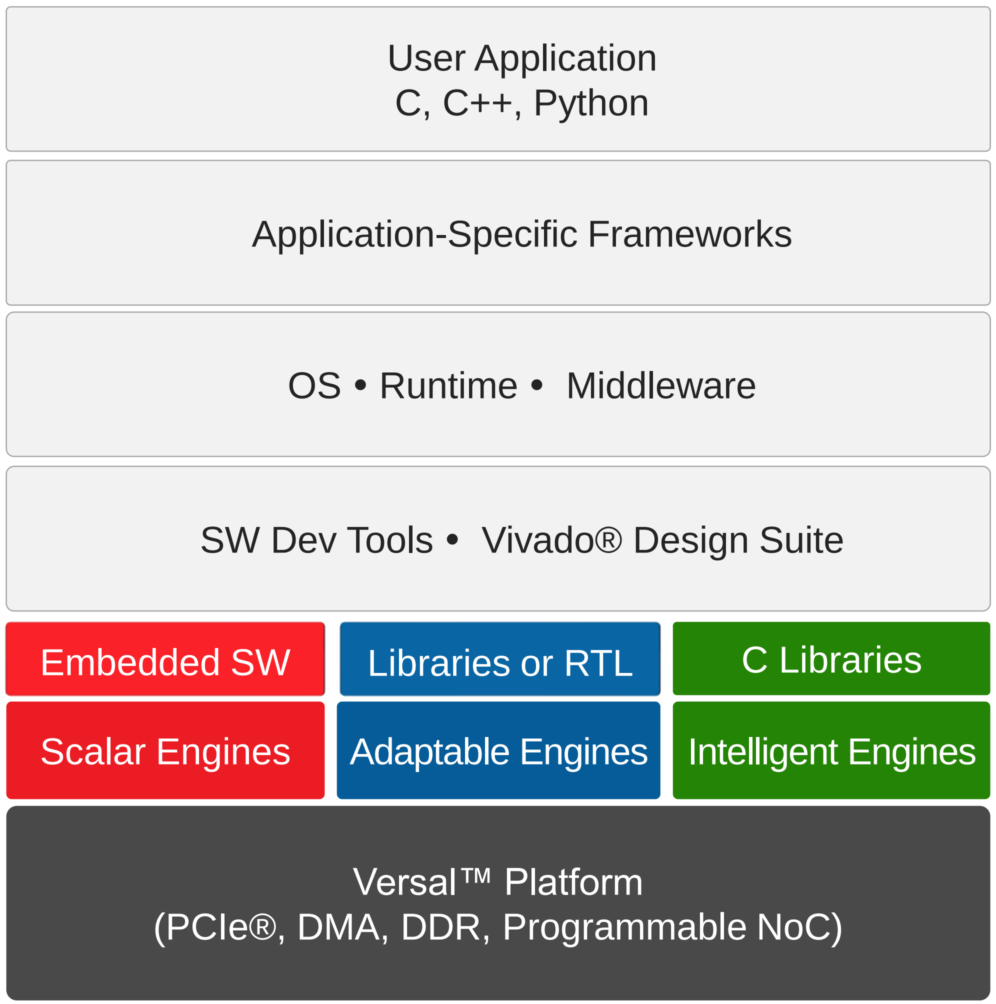


Edge Compute

AI Inference at Low Power



Hardware: The Foundation for the Software Stack



User Application
C, C++, Python

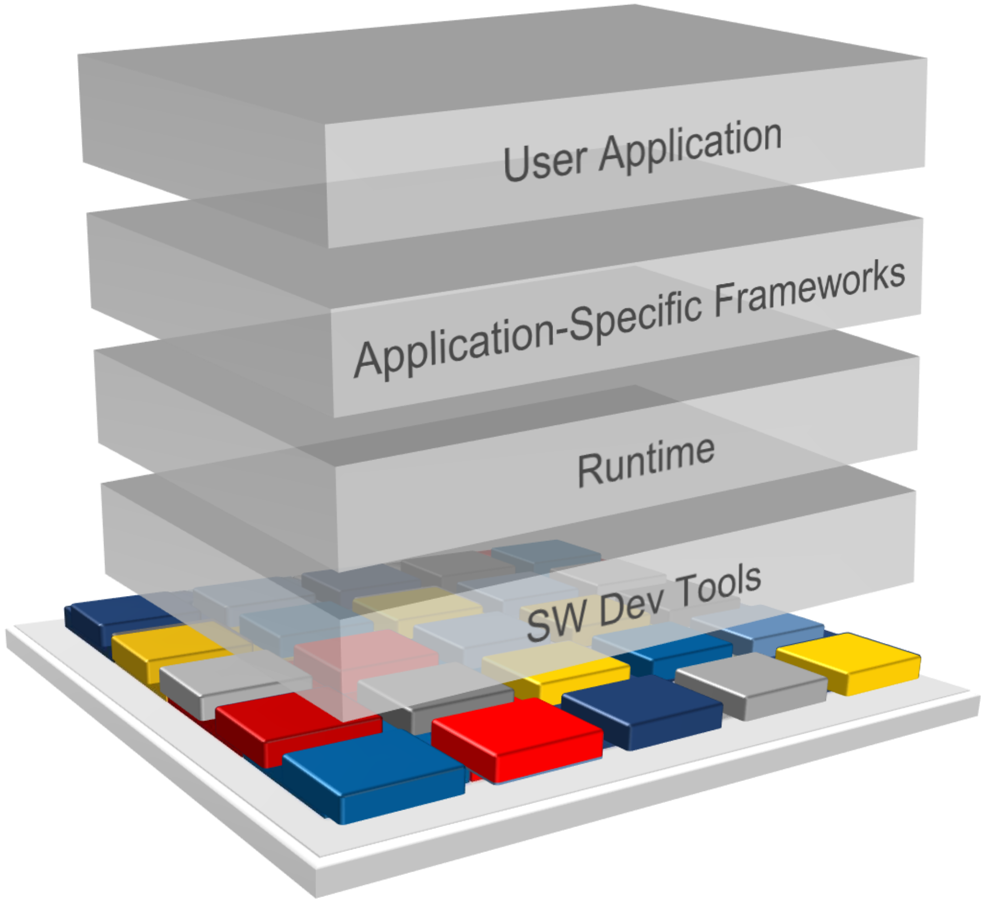
Application-Specific Frameworks

OS • Runtime • Middleware

SW Dev Tools • Vivado® Design Suite

Embedded SW	Libraries or RTL	C Libraries
Scalar Engines	Adaptable Engines	Intelligent Engines

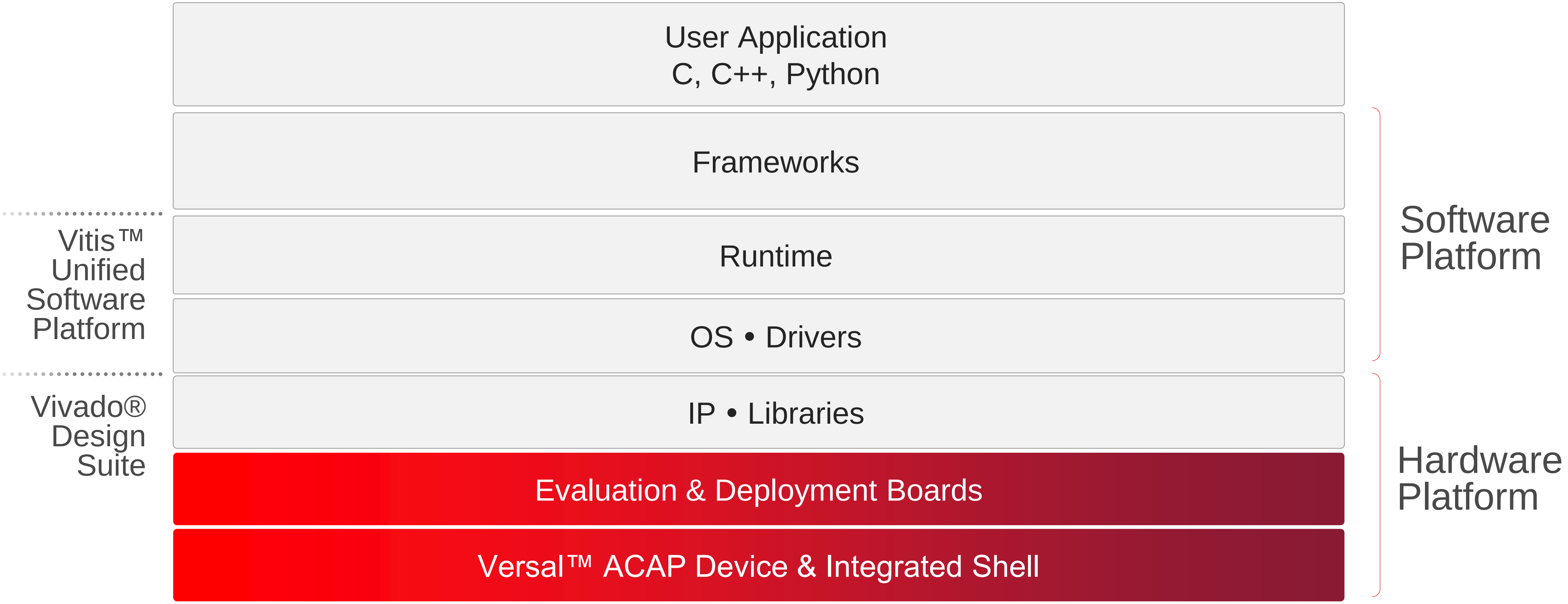
Versal™ Platform
(PCIe®, DMA, DDR, Programmable NoC)



Versal ACAP: A Platform for SW *and* HW 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

Vitis™ Unified Software Platform

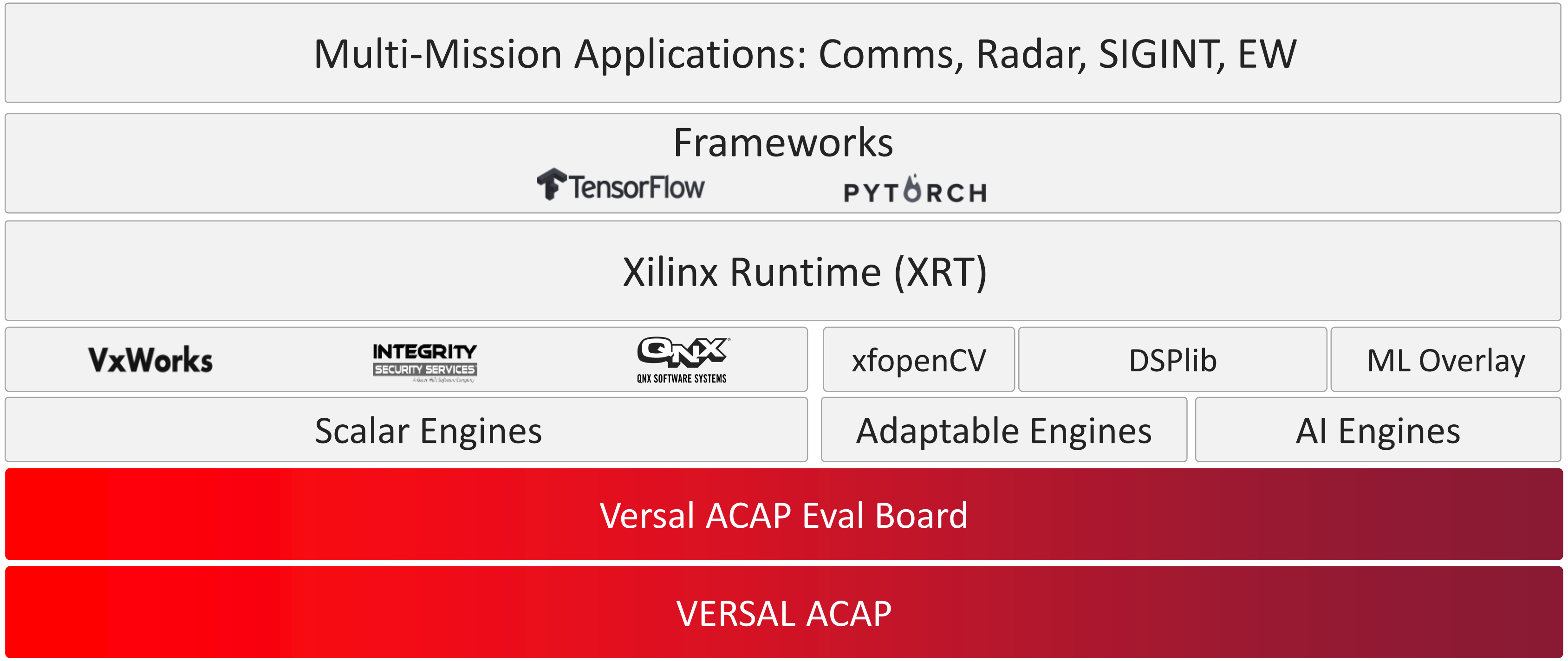
Vivado® Design Suite

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

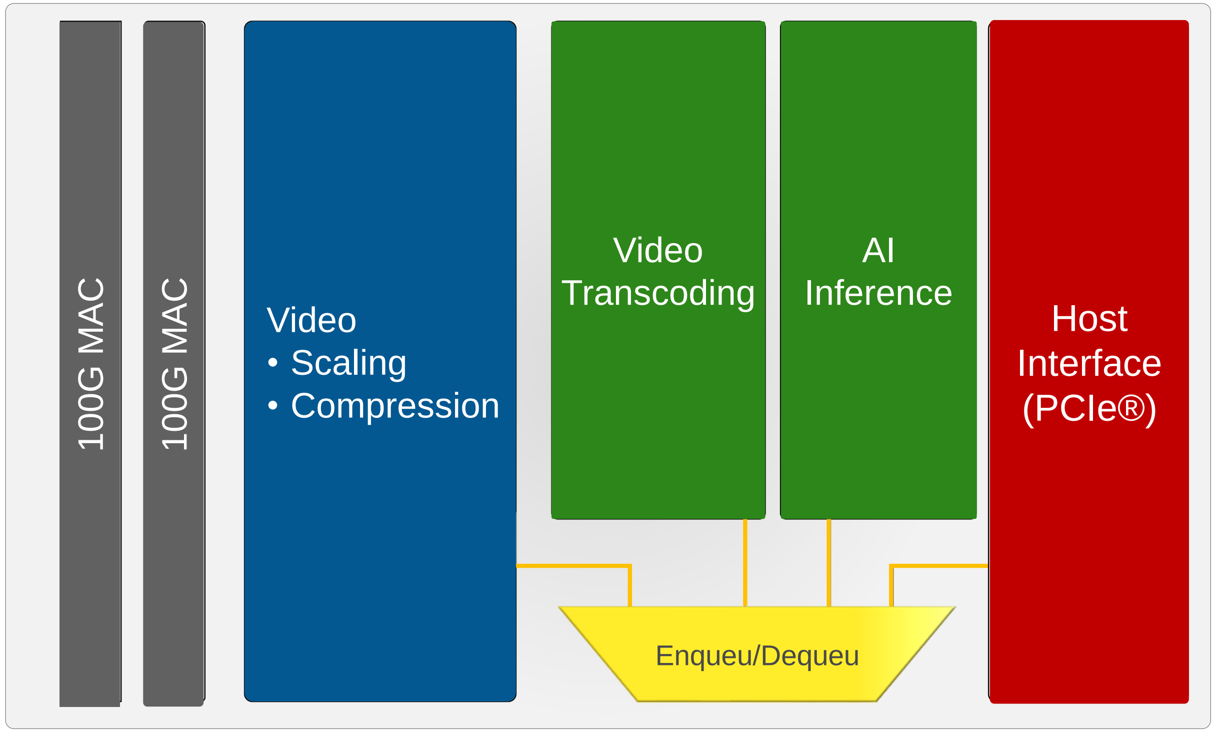
Versal ACAP Eval Board

VERSAL ACAP

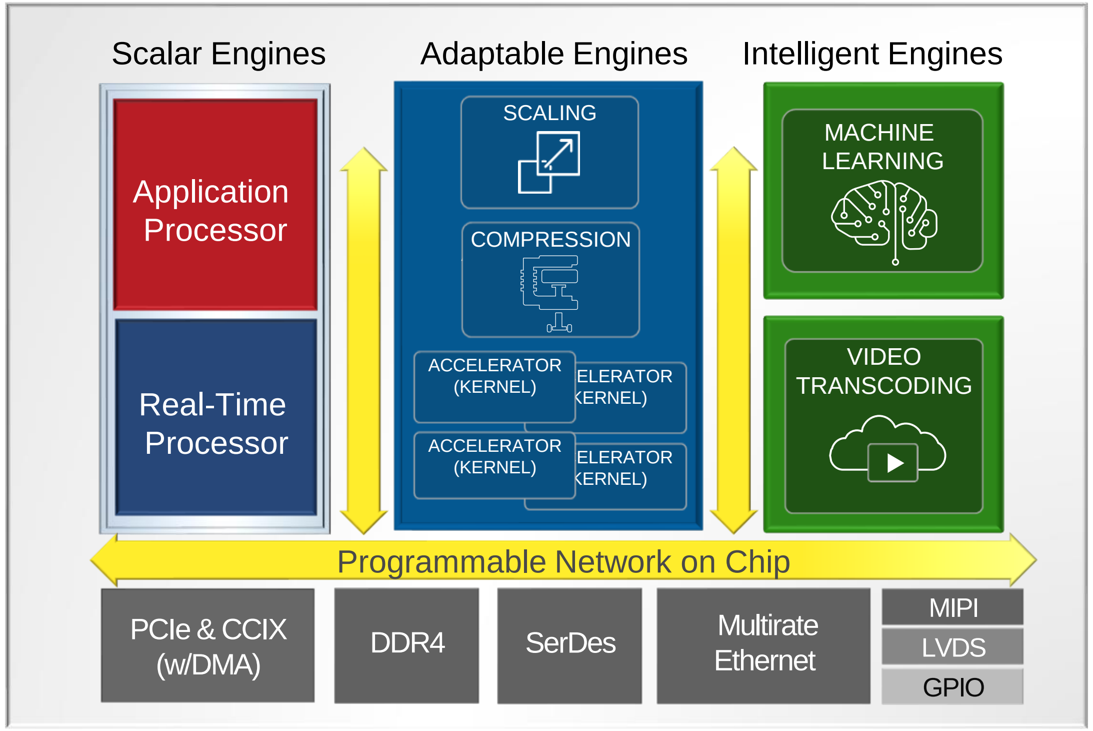
Key to ACAP is Mapping to Embedded IP & Accelerators

Application Example: Video Security (Surveillance)

FPGA



ACAP



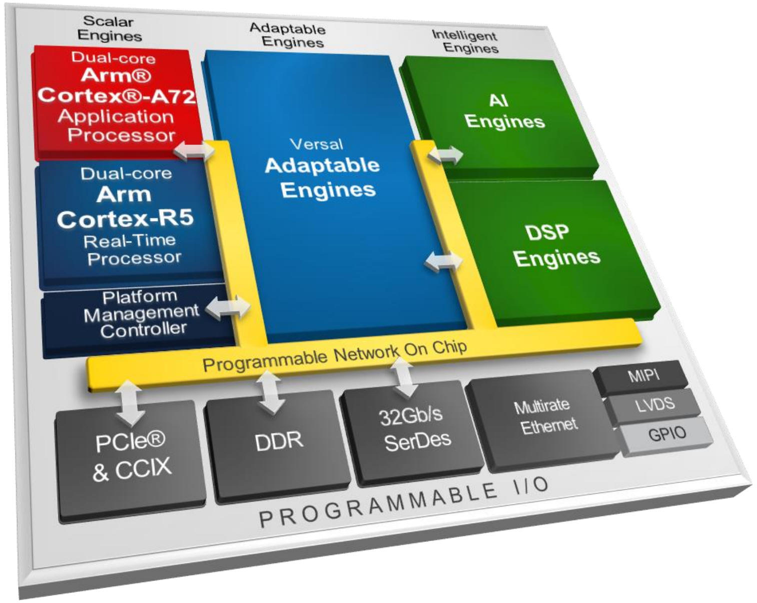
> All workloads are in logic, no room for additional differentiation
> Need for the arbitration between workloads and memory

> Workloads are mapped to the right engines
> High bandwidth, guaranteed QoS via Programmable NoC
> Power efficiency and greater performance

Part of a Comprehensive Product Portfolio

	28nm	20nm	16nm	7nm
Cloud Applications				XILINX VERSAL AI Core
RF Applications			ZYNQ RFSoC	XILINX VERSAL AI RF
Edge Applications	ZYNQ Zynq-7000 ZYNQ Zynq-7000S		ZYNQ UltraSCALE+	XILINX VERSAL AI Edge
Broad Application	VIRTEX 7 KINTEX 7 ARTIX 7 SPARTAN 7	VIRTEX UltraSCALE KINTEX UltraSCALE	VIRTEX UltraSCALE+ HBM VIRTEX UltraSCALE+ KINTEX UltraSCALE+	XILINX VERSAL HBM XILINX VERSAL Premium XILINX VERSAL Prime

Versal™ AI Core Series



Breakthrough AI Inference Throughput

- Portfolio's highest throughput for low latency inference
- Optimized for cloud, networking, and autonomous applications
- For highest dynamic range of AI and workload acceleration

Data Center Compute

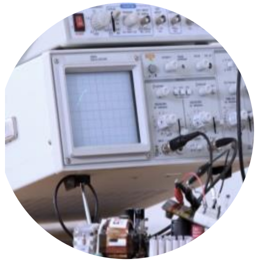
5G Radio & Beamforming

ADAS, AD Prototyping

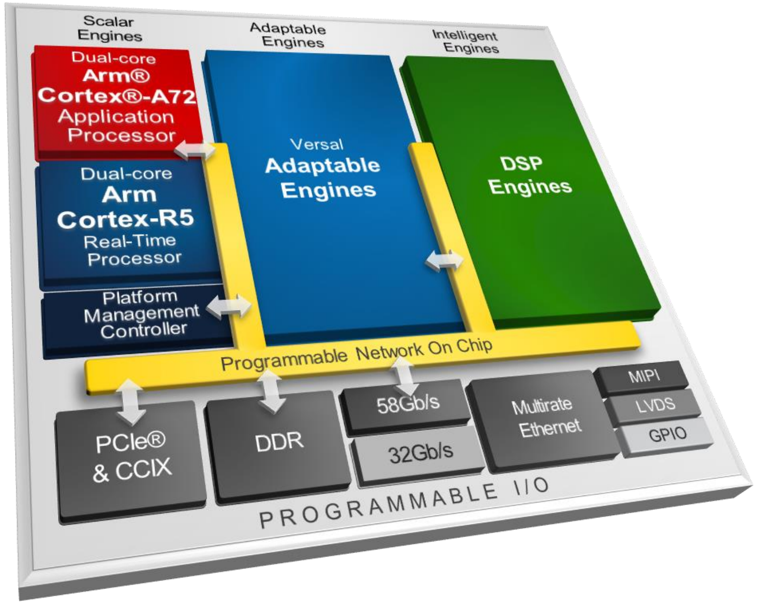
Cable Access

A&D

Wireless Test Equipment



Versal Prime Series



Broad Applicability across Multiple Markets

- Mid-range series in the Versal™ portfolio
- Optimized for connectivity
- For inline acceleration and diverse workloads

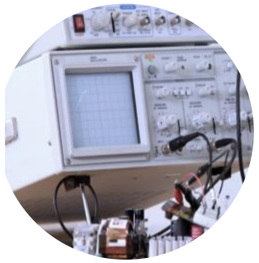
Nx100G Ethernet & OTN Networking

Data Center Network & Storage

Avionics Control

Broadcast Switches

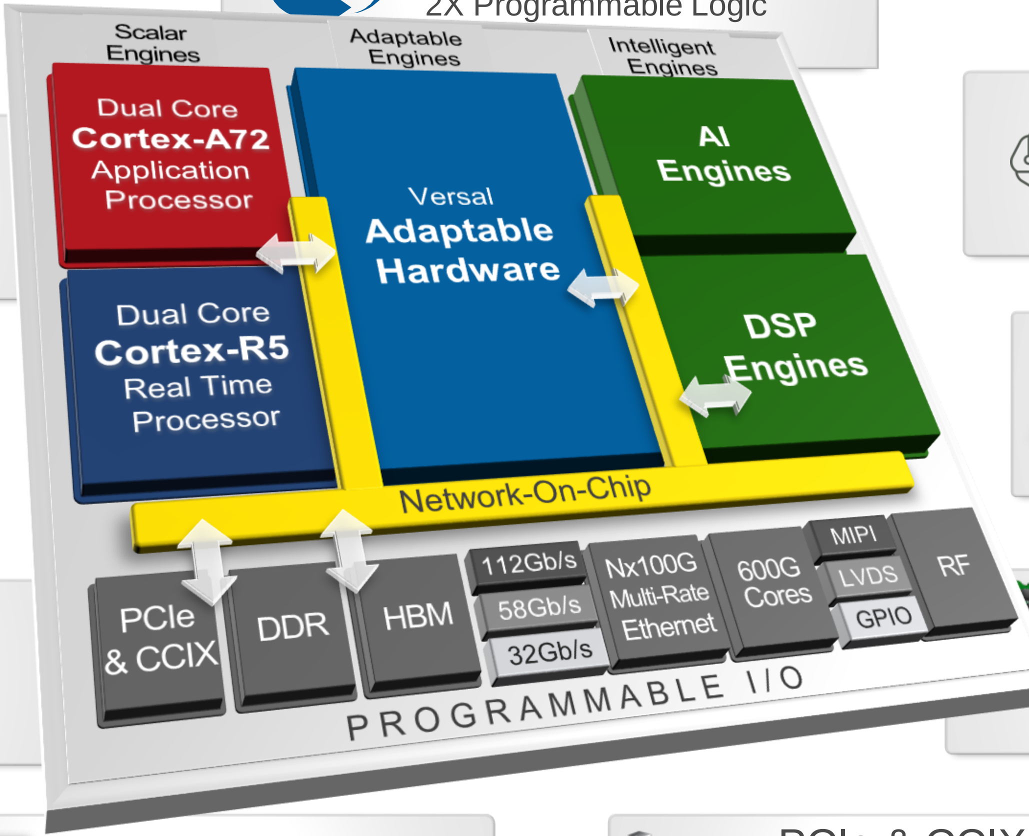
Network Test Equipment



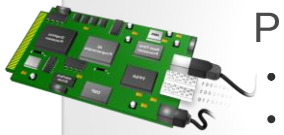
Medical Imaging

Versal Architecture Overview

Adaptable Engines
2X Programmable Logic



Scalar Engines
- Platform Control
- Edge Compute

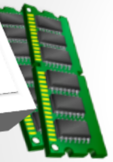


Protocol Engines
- Integrated 600G cores
- 4X encrypted bandwidth

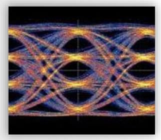
Programmable I/O
- Any interface or sensor
- Includes 3.2Gb/s MIPI

Intelligent Engines
- AI Compute
- Diverse DSP workloads

Network-on-Chip
- Guaranteed Bandwidth
- Enables SW Programmability



DDR Memory
- 3200-DDR4, 4266-LPDDR4
- 2X bandwidth/pin



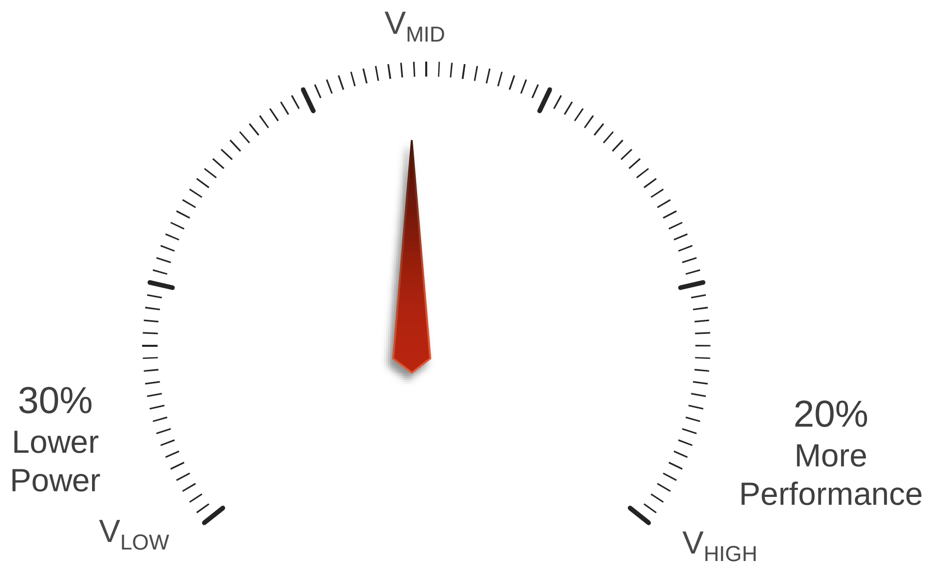
Transceivers
- Broad range, 25G →112G
- 58G in mainstream devices

PCIe & CCIX
- 2X PCIe & DMA bandwidth
- Cache-coherent interface to accelerators

Tune for Power & Performance in Versal ACAP

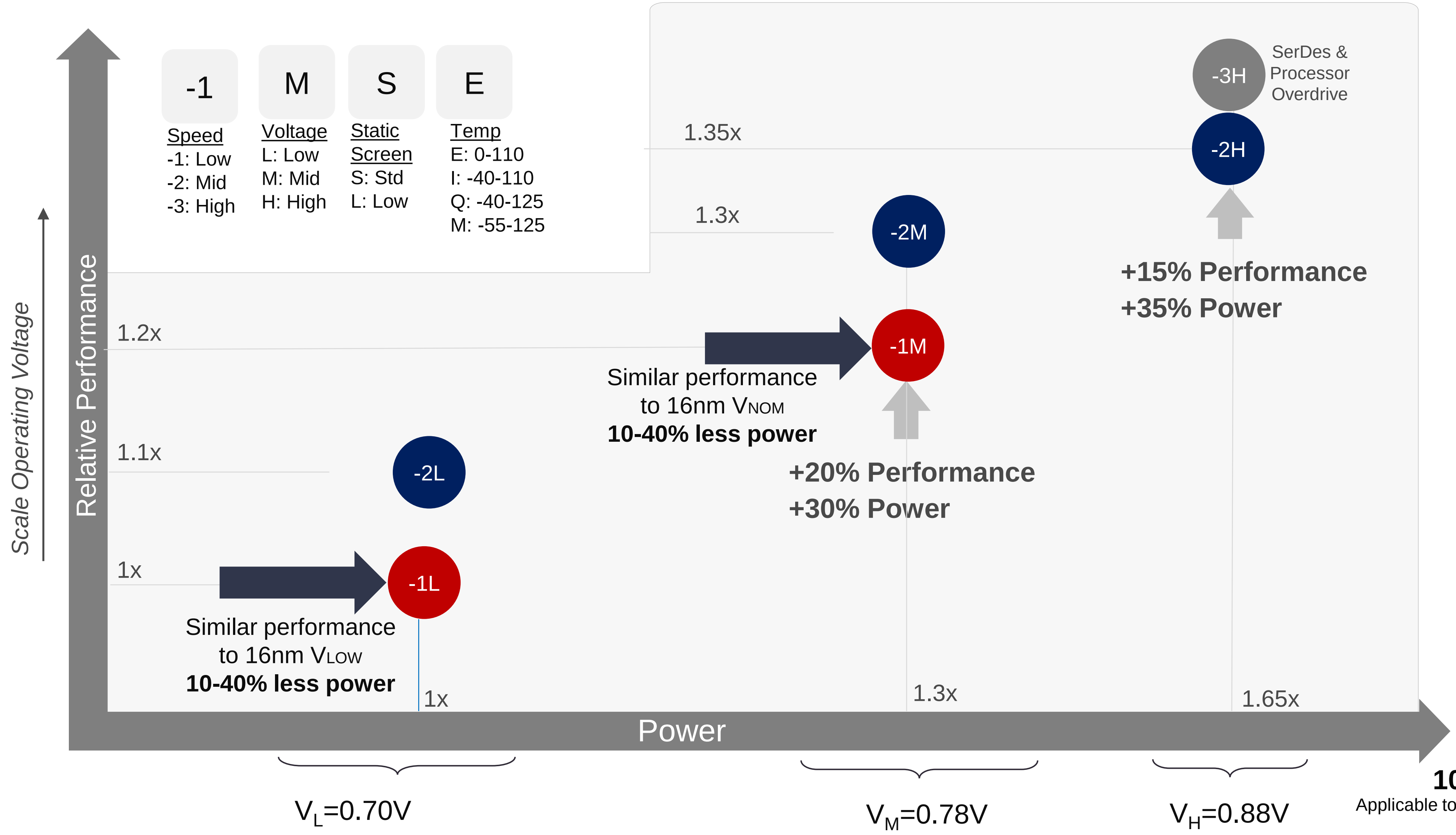
- Three operating voltages to choose from
- Balance power/performance for target app
- Equivalent to 3 speed grades in one device

-1	M	S	E
Speed	Voltage	Static Screen	Temp
-1: Low	L: Low	S: Std	E: 0-110
-2: Mid	M: Mid	L: Low	I: -40-110
-3: High	H: High		Q: -40-125
			M: -55-125



Granular Control of Power vs. Performance

Voltage Scaling with Speed Grade Options



-1	M	S	E
Speed	Voltage	Static Screen	Temp
-1: Low -2: Mid -3: High	L: Low M: Mid H: High	S: Std L: Low	E: 0-110 I: -40-110 Q: -40-125 M: -55-125

Similar performance to 16nm VLOW
10-40% less power

Similar performance to 16nm VNOM
10-40% less power

+20% Performance
+30% Power

+15% Performance
+35% Power

● Speed Grade 3
● Speed Grade 2
● Speed Grade 1

Versal (7nm)

V_H = 0.88V
V_M = 0.80V
V_L = 0.72V

UltraScale+ (16nm)

V_{NOM} = 0.85V
V_{LOW} = 0.72V

100 to 110ºC 3% Lifetime
Applicable to all E and I grades except -1E and -3E

New Intelligent Engines

Massive AI Inference Throughput and Wireless Compute

Up to 1.3GHz VLIW / SIMD vector processors

- Versatile core for ML and other advanced DSP workloads

Massive array of interconnected cores

- Instantiate multiple tiles (10s to 100s) for scalable compute

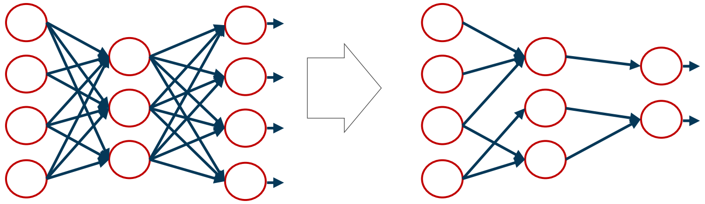
Terabytes/sec of interface bandwidth to other engines

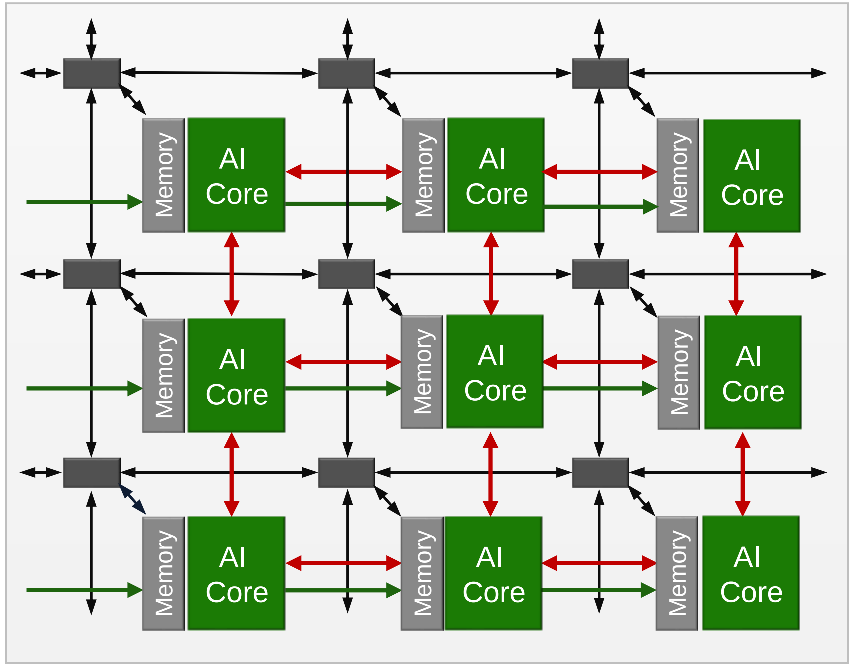
- Direct, massive throughput to adaptable HW engines
- Implement core application with AI for “Whole App Acceleration”

SW programmable for any developer

- C programmable, compile in minutes
- Library-based design for ML framework developers

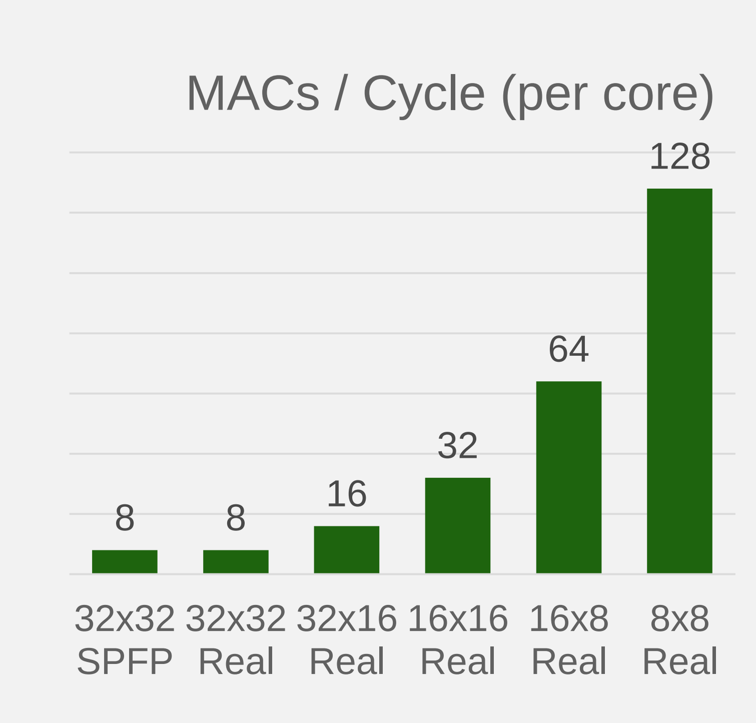
ML Inference and Optimizations



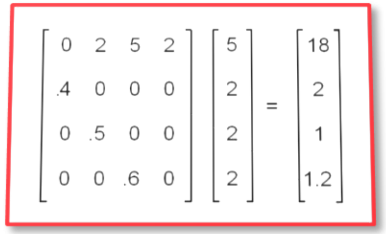


AI Engine: Multi-Precision Math Support

Real Data Types



Optimized For:

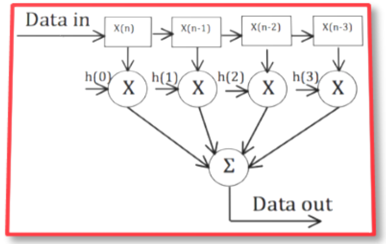


$$\begin{bmatrix} 0 & 2 & 5 & 2 \\ .4 & 0 & 0 & 0 \\ 0 & .5 & 0 & 0 \\ 0 & 0 & .6 & 0 \end{bmatrix} \begin{bmatrix} 5 \\ 2 \\ 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 18 \\ 2 \\ 1 \\ 1.2 \end{bmatrix}$$

Linear Algebra

Matrix-Matrix Mult

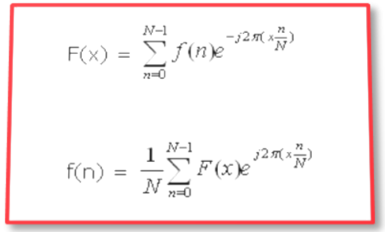
Matrix-Vector Mult



Convolution

FIR Filters

2-D Filters



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

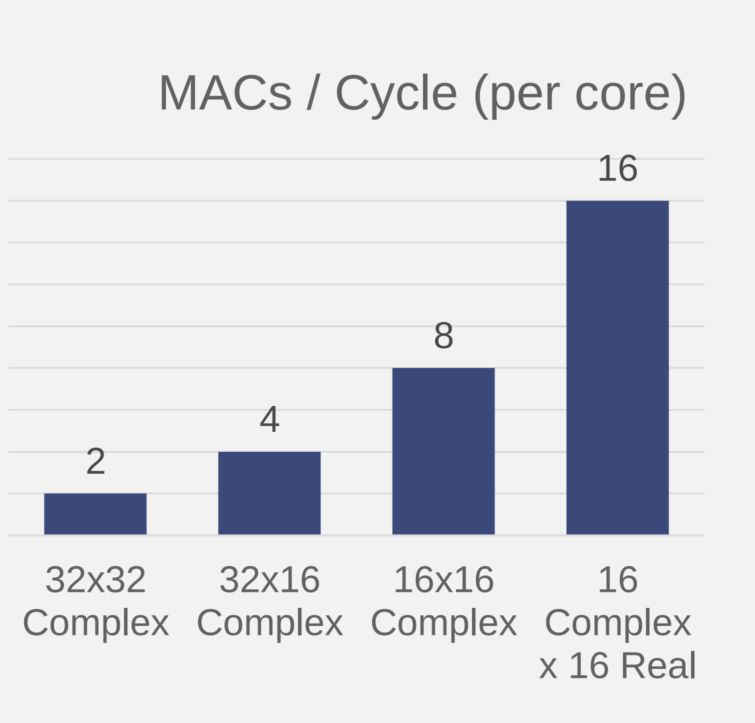
$$f(n) = \frac{1}{N} \sum_{n=0}^{N-1} F(x) e^{j2\pi(x\frac{n}{N})}$$

Transforms

FFTs/IFFTs

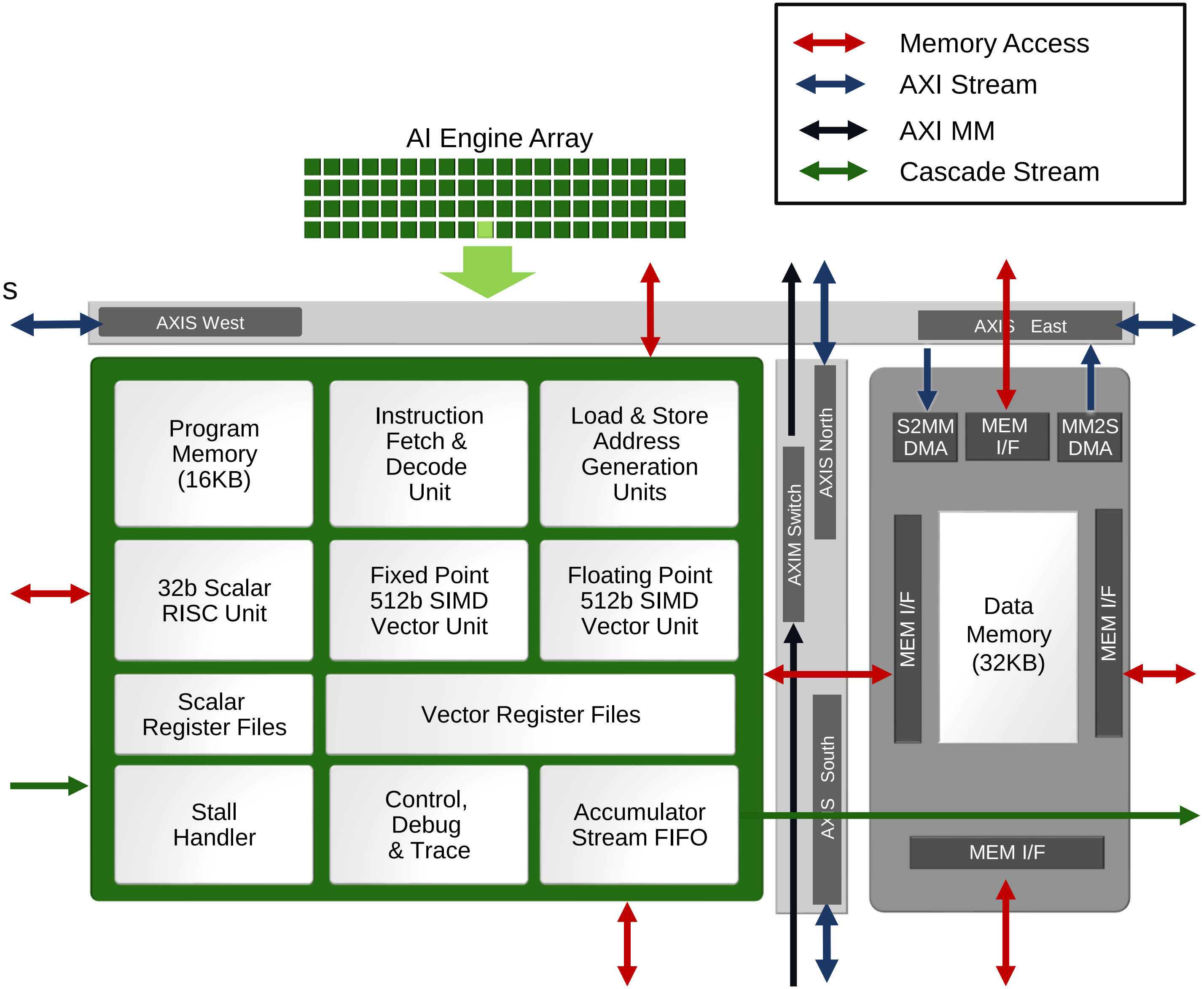
DCT, etc

Complex Data Types



AI Engine Tile

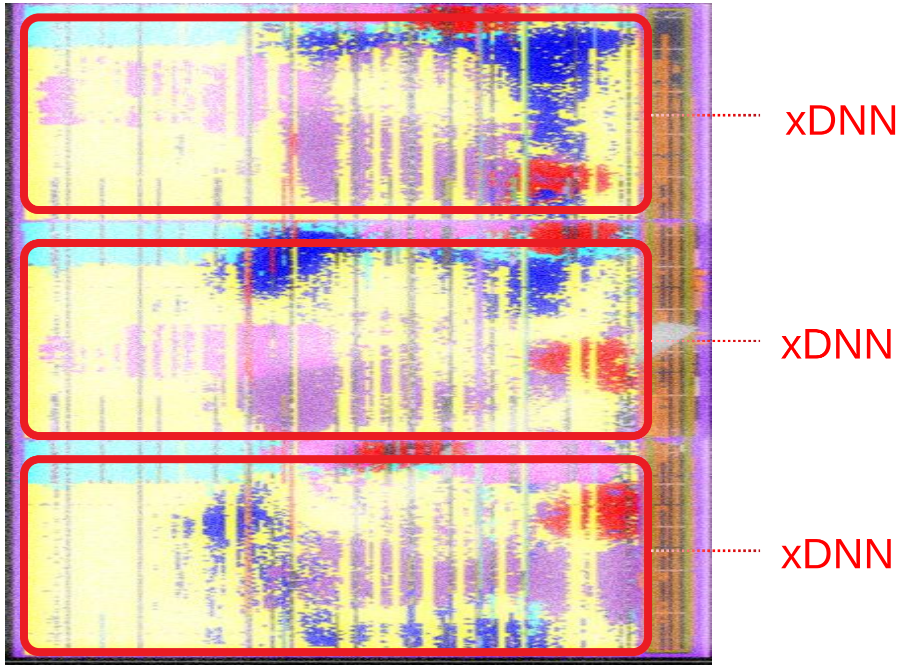
- AI Engine core
  - 512b SIMD vector units
    - Both fixed and floating point
    - 16KB program memory
  - 32b scalar RISC processor
  - 256-bit load (x2) and store units with individual AGUs
- 128KB direct core memory access
  - 32KB local
  - 32KB north, south, east & west
- Streaming interconnects
  - AXI Memory Mapped (AXI-MM) switch
    - Configuration, control and debug
  - AXI-Stream crossbar switch
    - Routing N/S/E & west around the array
- Debug/Trace/Profile functionality
  - Debug using memory-mapped AXI4 i/f
  - Connect to PMC via JTAG or HSDP



Leveraging AI Engines for Compute-Intensive Applications

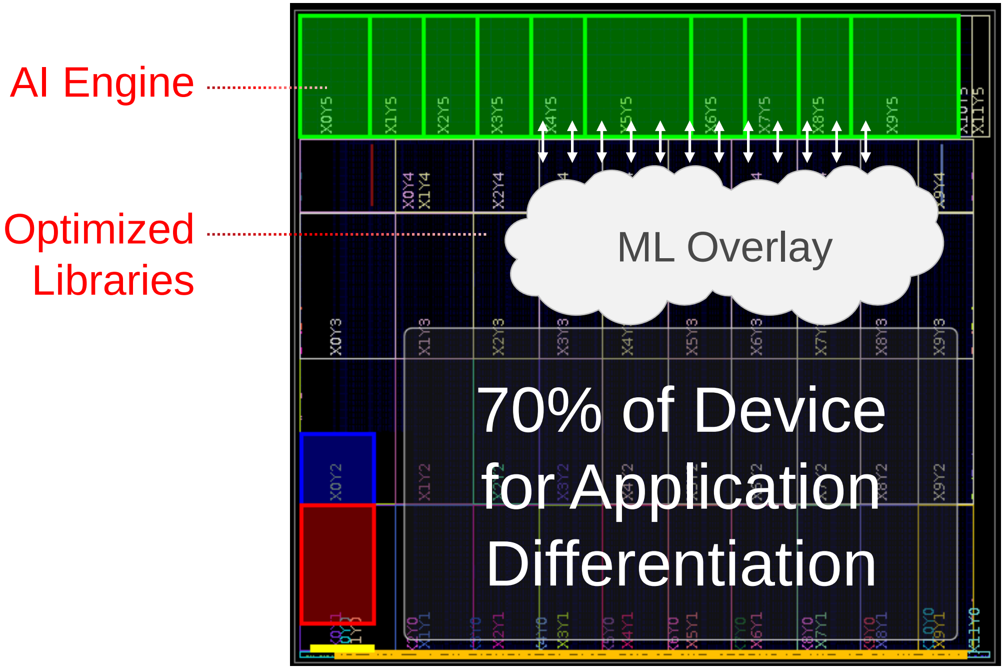
Virtex® UltraScale+™ VU9P

100% of Logic for ML Acceleration (20 TOPS INT8)



Versal™ AI Core Series

>5X More Compute w/AI Engines (133 TOPS INT8)



DSP Engines

Versatility and Granular Control of Datapath

Enhanced Compute architecture

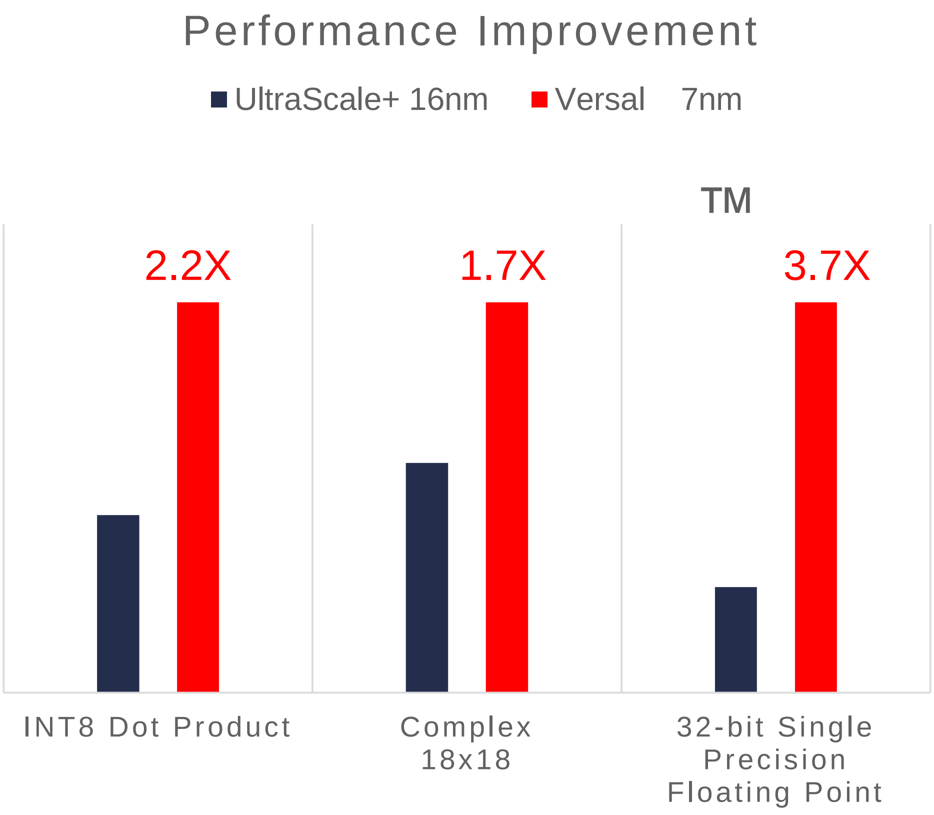
- > Greater than 1GHz of performance

Versatility for Wireless, ML, HPC, and more

- > Integrated FP32, FP16 floating point, INT24 (HPC)
- > Integrated complex 18x18 operation (wireless, cable access)
- > Double the performance in INT8 operation (AI inference)

Code Portability for UltraScale+™ 16nm designs

- > Support for legacy IP and LogiCORE™ libraries
- > Compatibility with SysGen, Model Composer, HLS tools

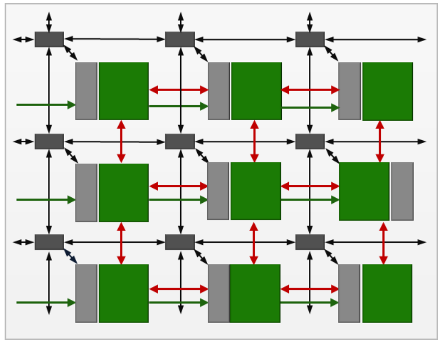


AI and DSP Engines

AI Engine 2D Array

VLIW and SIMD Architecture

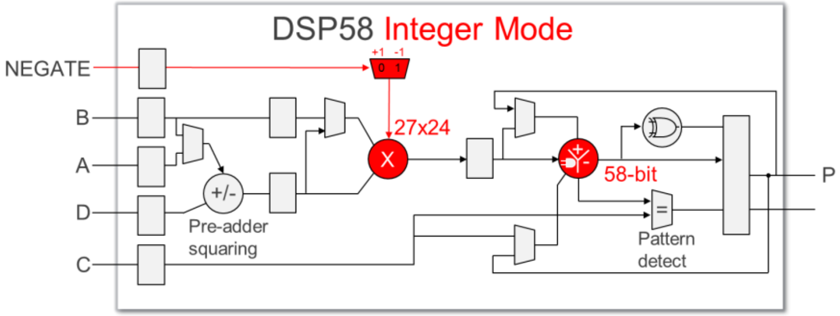
C/C++ Programmable

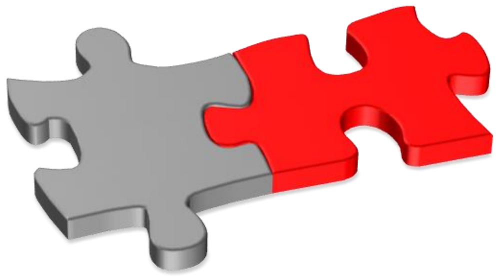


DSP Engine

Additional Features

RTL Entry





Why AI Engine?

- > Massive compute performance
- > S/W programmable (C/C++)
- > Fast compile – increase productivity

Why DSP Engine?

- > Existing RTL/HLS IP usage
- > Additional features not available in AI Engine, e.g., 58-bit logic unit
- > Pre/Post processing to/from AI Engine

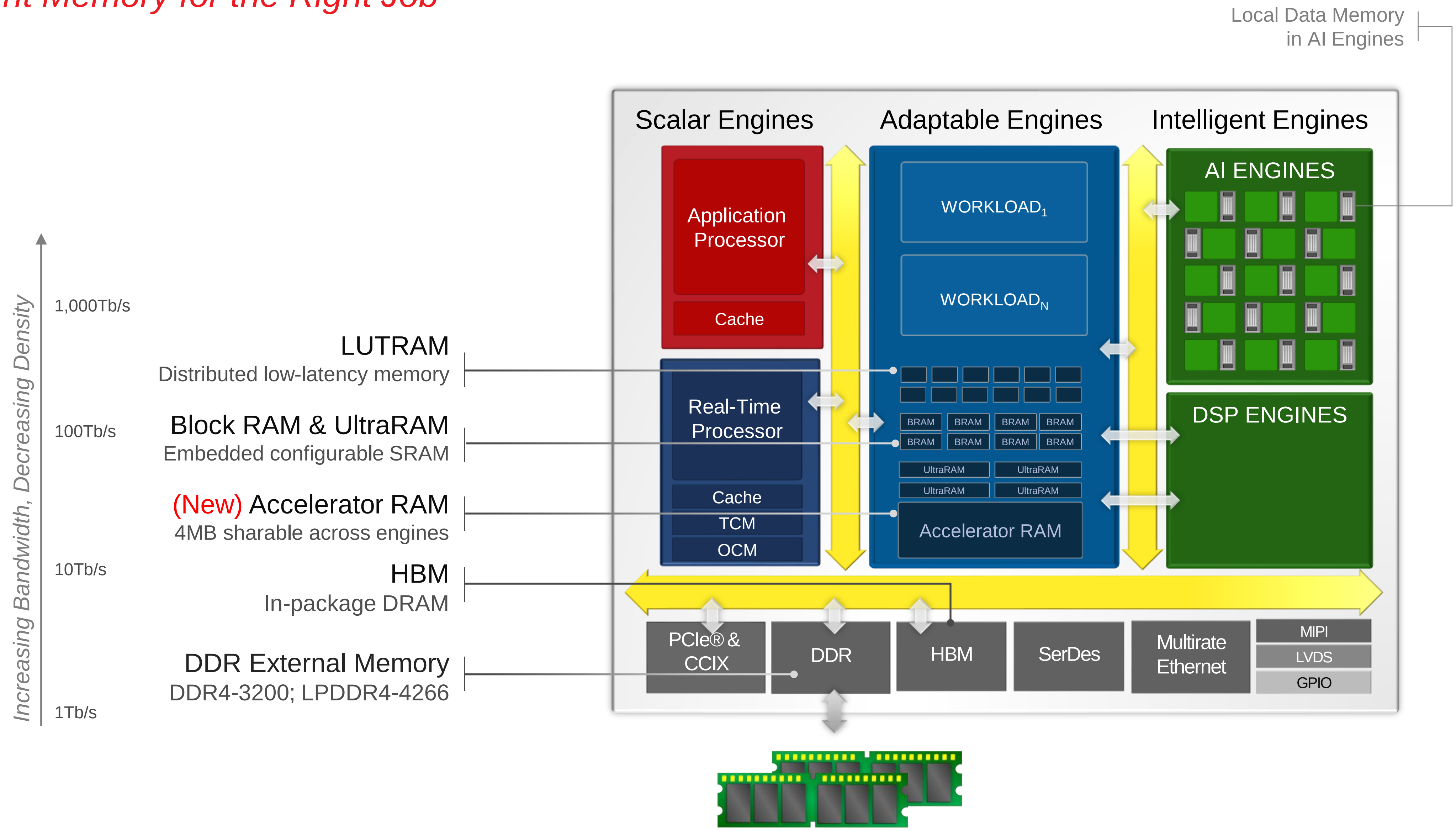
Why both?

- > AI Engine for efficiency
- > PL & DSP for flexibility

Versal™ ACAPs Accelerate the Complete Application

Adaptable Cache-less Memory Hierarchy

The Right Memory for the Right Job



LUTRAM
Distributed low-latency memory

Block RAM & UltraRAM
Embedded configurable SRAM

(New) Accelerator RAM
4MB sharable across engines

HBM
In-package DRAM

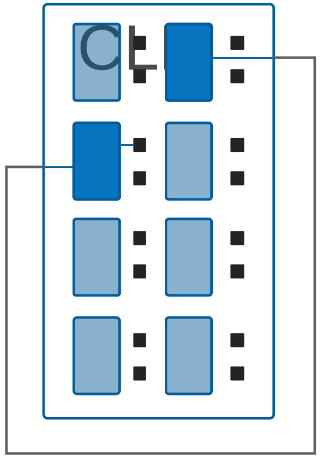
DDR External Memory
DDR4-3200; LPDDR4-4266

Local Data Memory
in AI Engines

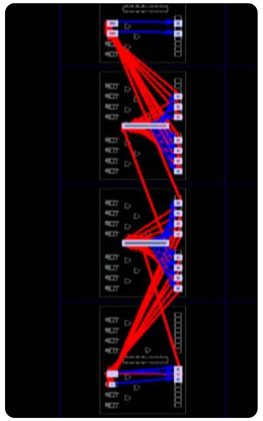
Re-Architected Hardware Logic for 4X Compute Density

16nm UltraScale+

UltraScale+™

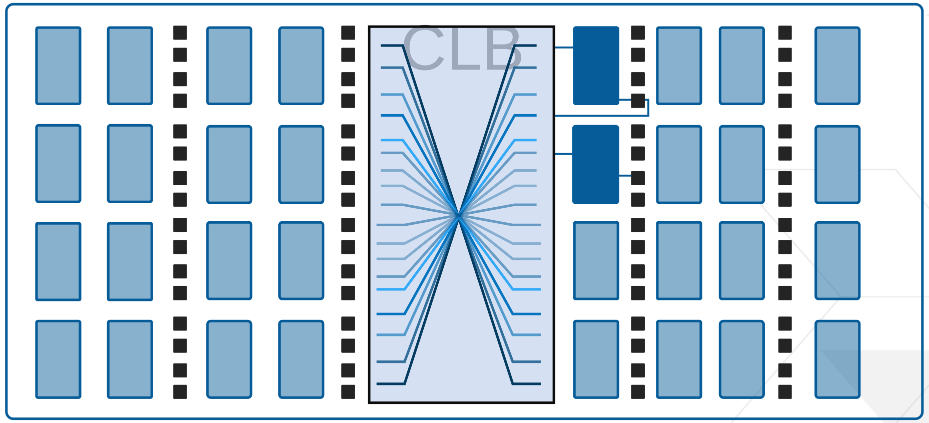


16b_counter
16 Global routes used

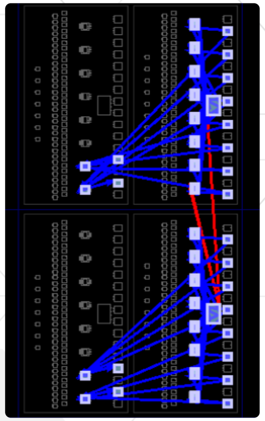


7nm ACAP – 4X logic resources per CLB

Versal™ ACAP



16b_counter
2 Global routes used



□ LUT ■ FF

— Global route (slow)

— Local route (fast)

New CLB Interconnect Reduces Need for Global Interconnect

NoC for Ease of Use, Guaranteed Bandwidth, and Power Efficiency

High bandwidth terabit network-on-chip

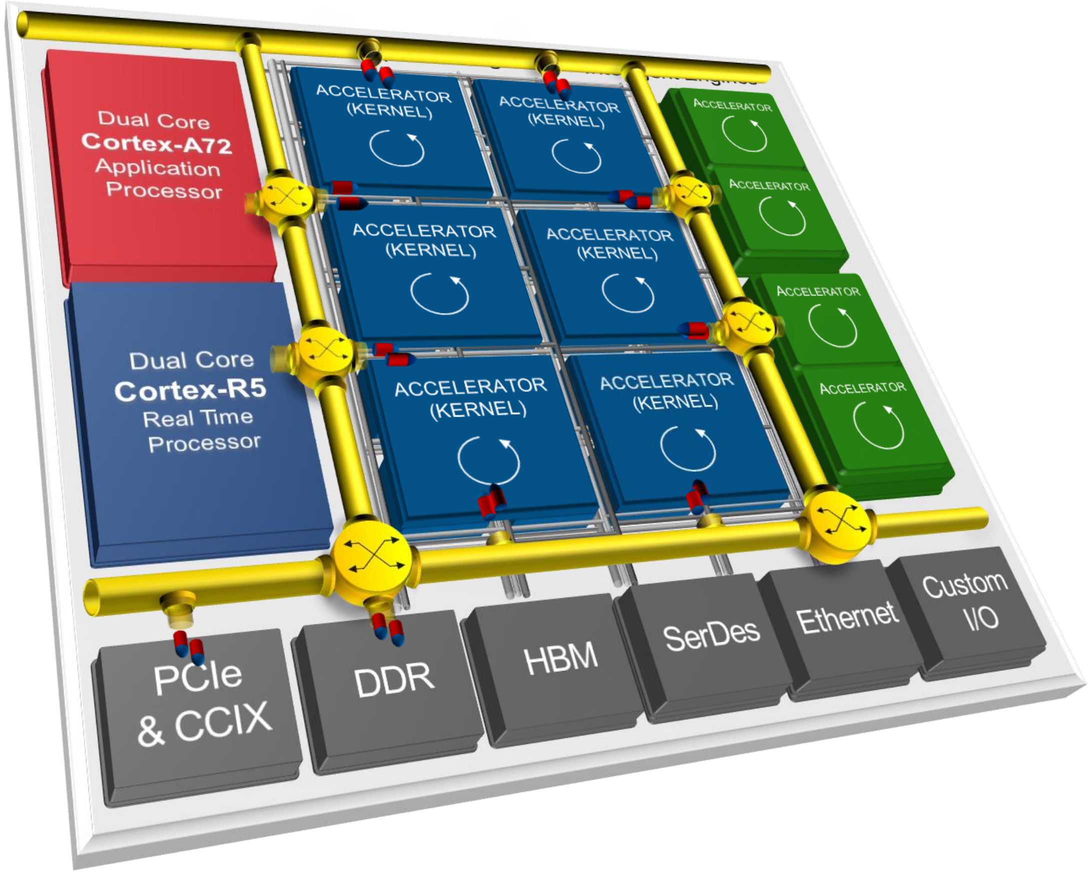
- Memory mapped access to all resources
- Built-in arbitration between engines and memory

High Bandwidth, Low Latency, Low power

- Guaranteed QoS

Eases Kernel Placement

- Easily swap kernels at NoC port boundaries
- Simplifies connectivity between kernels



Programmable NoC vs. Logic Utilization using UltraScale+

Simple Test Case – 4 AXI Traffic Generators Connected to 4 Block RAM

UltraScale+™ FPGA

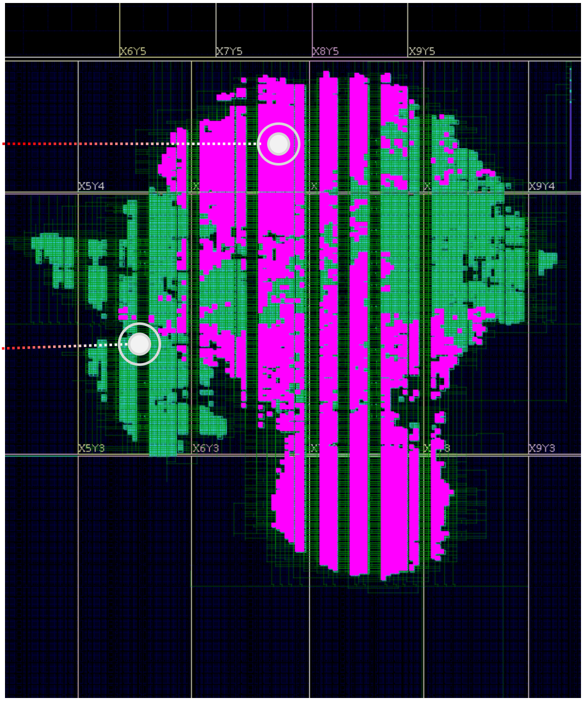
Logic Resources for Connectivity (SmartConnect)

42,012 FF

37,015 LUTs

Connectivity (pink)

Design (Green)



Design Size Reduction

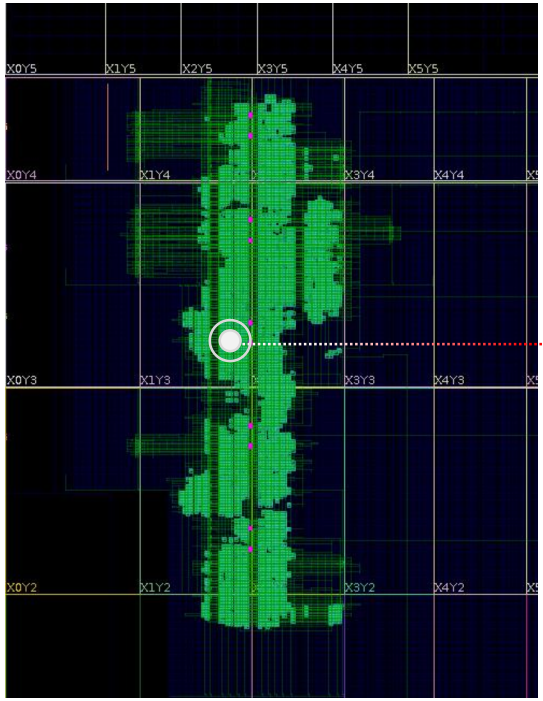
~60% FF
~50% LUT

Versal™ ACAP

Logic Resources for Connectivity

0 FF

0 LUTs



Design (Green)

Introducing the “Integrated Shell”

‘Shell’: Pre-Built Core Infrastructure & System Connectivity

> External host interface

> Memory subsystem

> Basic interfaces (e.g., JTAG, USB, GbE)

Key Architectural Elements of the Shell

> Platform Management Controller (PMC)

> Integrated host interfaces: PCIe & CCIX, DMA

> Scalable Memory Subsystem: DDR4 & LPRDDR4

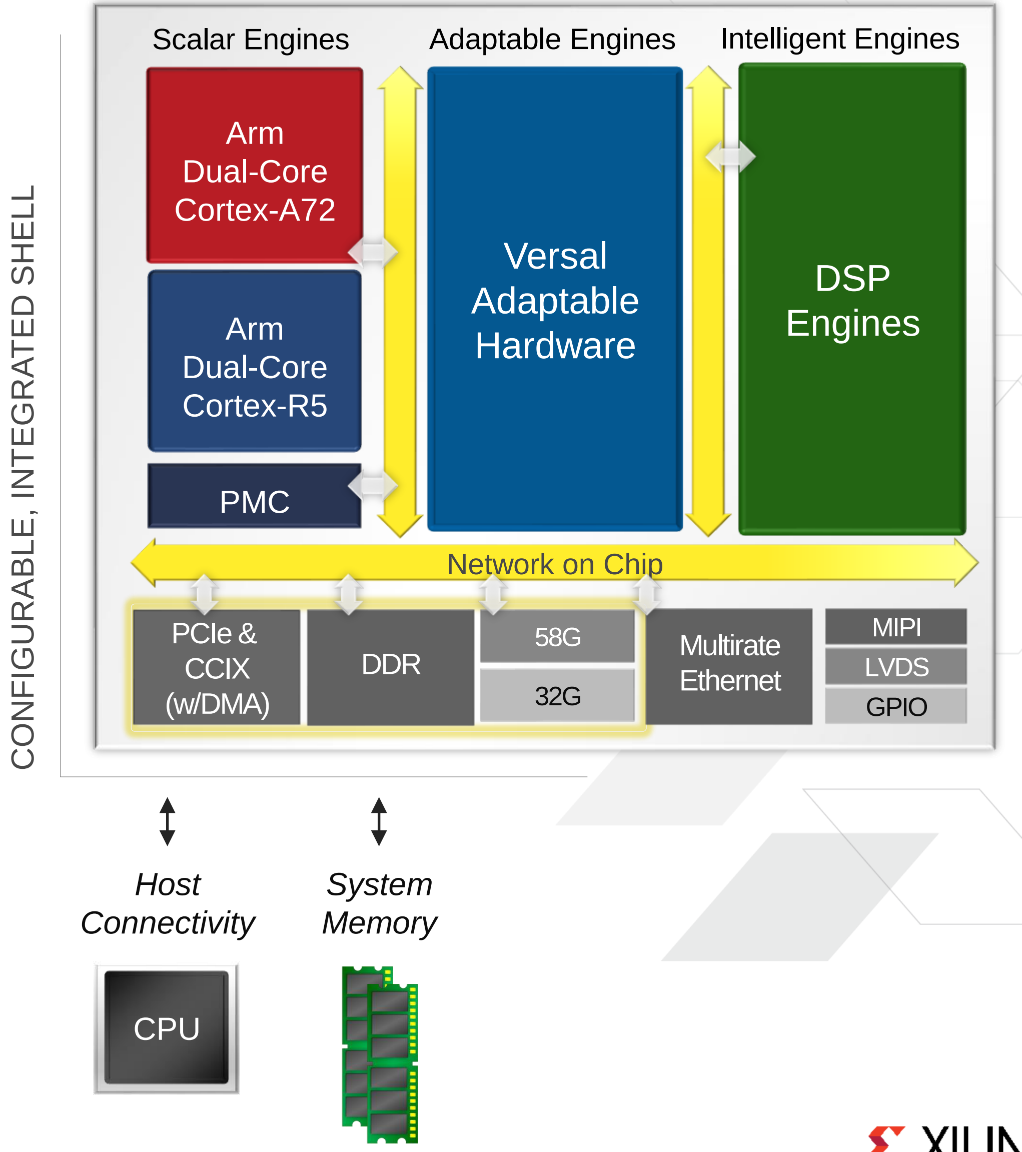
> Network-on-Chip for connectivity and arbitration

Greater Performance, Device Utilization, and Productivity

> More of the platform available for application’s workload(s)

> Target application runs faster with less device congestion

> Turn-key, pre-engineered timing closure – no debug



Summary

- Heterogeneous processing is required for the future as there is no single processor type that is ideally suited for all algorithms and applications
- This can be a challenge for design development and productivity
- ACAPs are a response to this new reality
- Embracing multiple levels of abstraction through a unified platform allows developers to use the tools and languages that they are familiar with while simplifying debugging

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.

Building the Adaptable, Intelligent World

XILINX