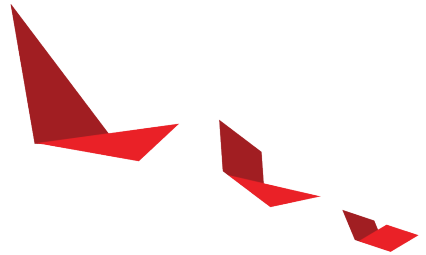
XILINX

Design Rationale of Two Generations of AI Engines

Kees Vissers
Fellow

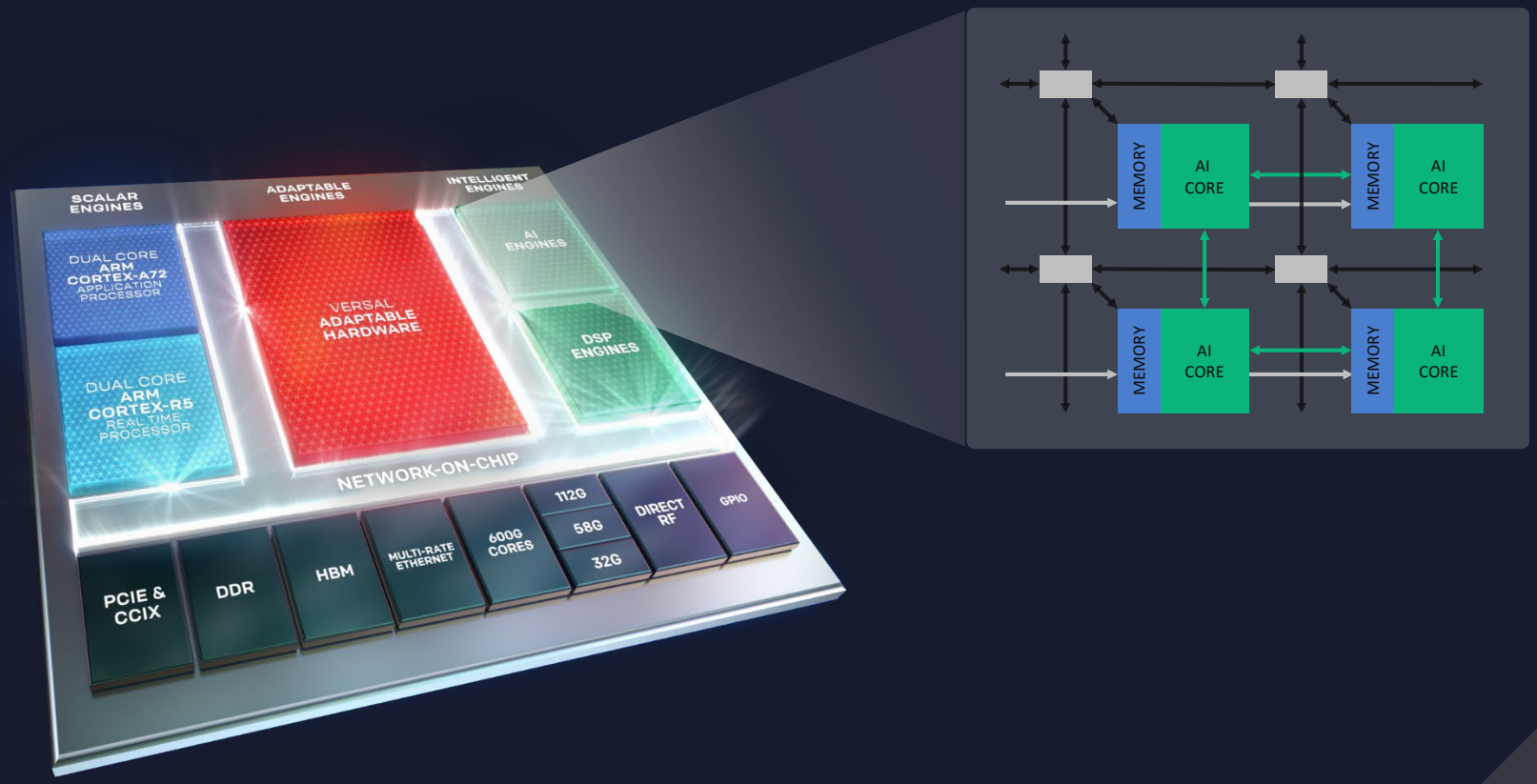
September 7, 2021

Agenda



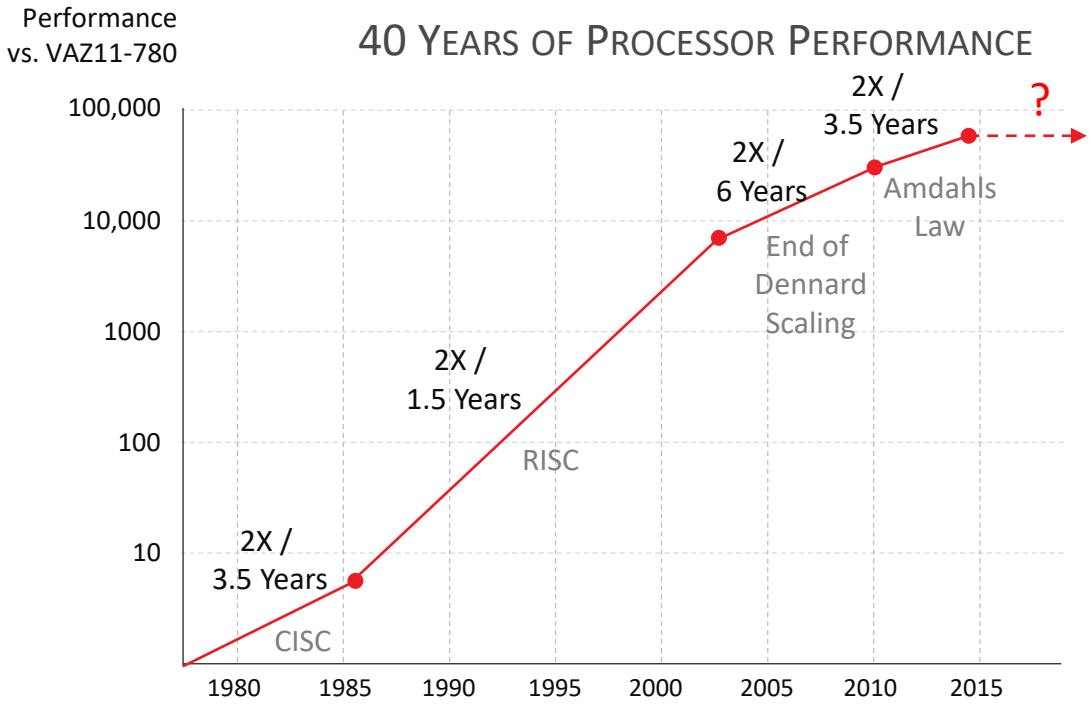
- Motivation for the first AI Engine architecture
- AI Engine architecture in more detail
- Programming
- Some Applications
- Motivation for the second generation AI Engine, AI – ML
- AI-ML in more detail
- Some Applications

Motivation for AI Engine



Computer Architecture: choices, a new golden age!

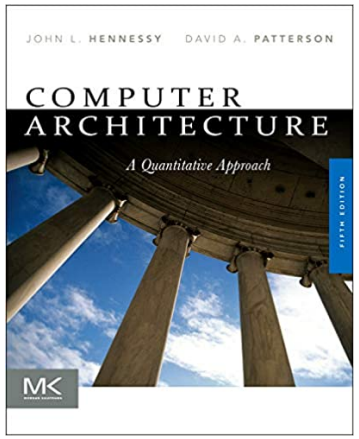
Performance is not scaling: end of Moore's Law



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

2017 Turing award winners, talk:

"A New Golden Age for Computer Architecture: Domain-Specific Hardware/Software Co-Design, Enhanced Security, Open Instruction Sets, and Agile Chip Development,"



https://www.acm.org/hennessy-patterson-turing-lecture

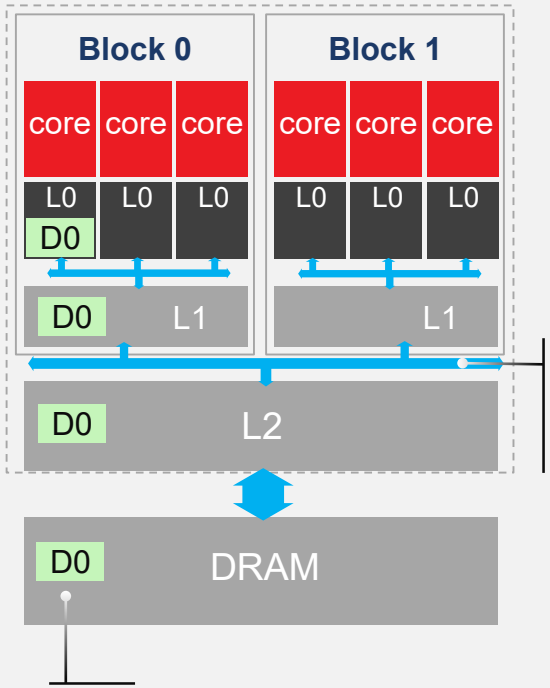
XILINX.

Original design rationale for AI engine

- **Trend towards heterogeneous architectures**
- **Can a novel array of dedicated, Xilinx designed processors, complement Programmable Logic?**
- **Can these processors provide cost and power benefit for the customers in the domains of:**
  - Wireless applications
  - General signal processing and image processing applications
  - Emerging Machine Learning applications
- **Can you raise the programming abstraction to programming environments:**
  - C/C++
  - Machine Learning Frameworks

AI Engine: Multi-Core Compute with dedicated memory

Traditional Multi-core
(cache-based architecture)

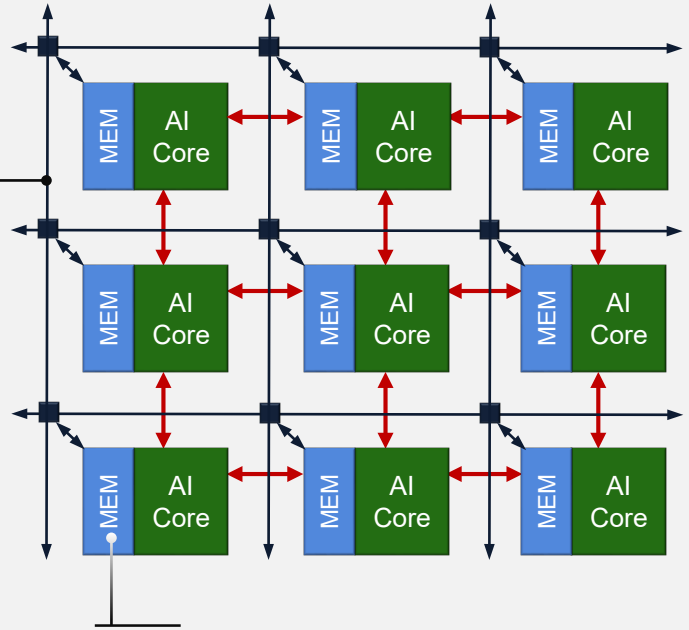


Fixed, shared Interconnect
- Blocking limits compute
- Timing not deterministic

Data Replicated
- Robs bandwidth
- Reduces capacity

AI Engine Array
(intelligent engine)

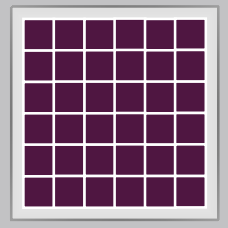
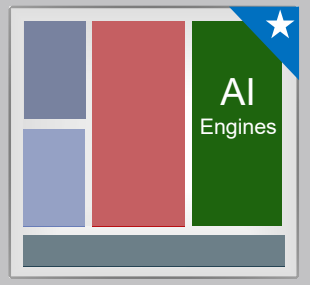
Dedicated Interconnect
- Non-blocking
- Deterministic



Local, Distributed Memory
- No cache misses
- Higher bandwidth
- Less capacity required

XILINX.

Delivering Compute Acceleration

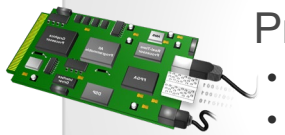
	CPU (Sequential)	GPU (Parallel)	Versal	Custom ASIC
				
SW Programmable	✓	✓	✓	✓
HW Adaptable	–	–	✓	–
Workload Flexibility	✓	✓	✓	–
Throughput vs. Latency	–	–	✓	✓
Device / Power Efficiency	–	–	✓	✓

Versal Architecture Overview

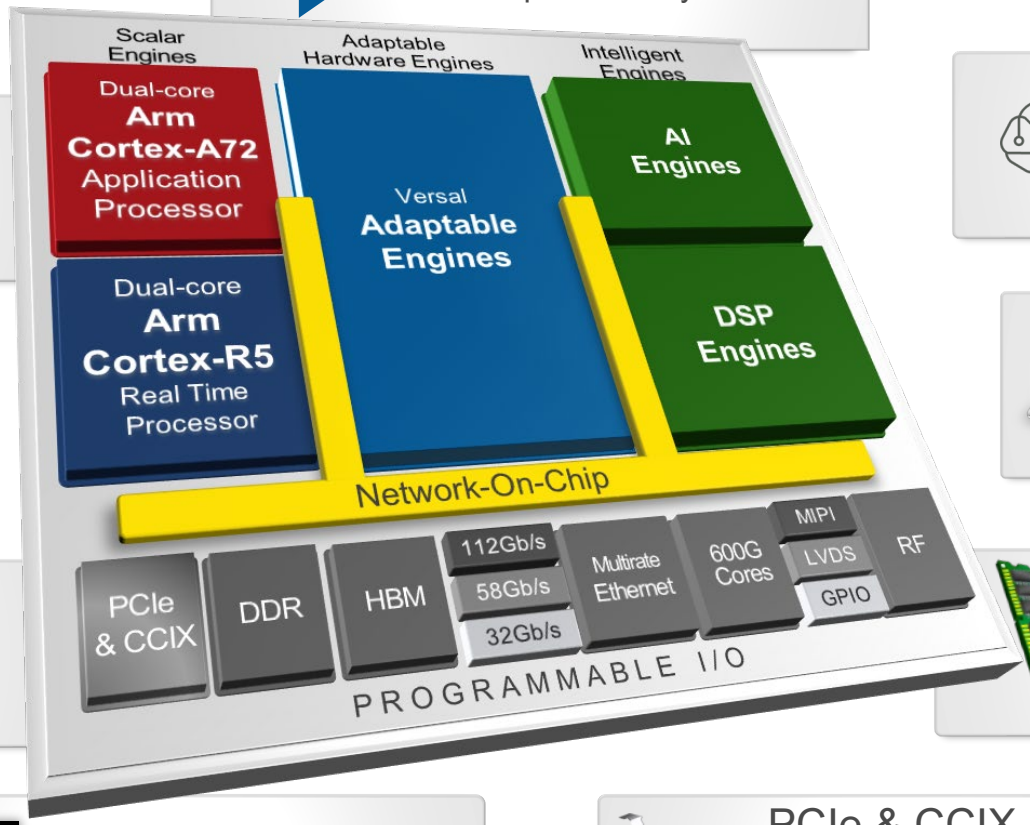
Adaptable Engines
2X compute density

Scalar Engines
- Platform Control
- Edge Compute

AI Engines
- AI Compute
- Diverse DSP workloads

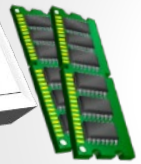


Protocol Engines
- Integrated 600G cores
- 4X encrypted bandwidth

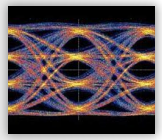


Network-on-Chip
- Guaranteed Bandwidth
- Enables SW Programmability

Programmable I/O
- Any interface or sensor
- Includes 4.2Gb/s MIPI



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



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

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

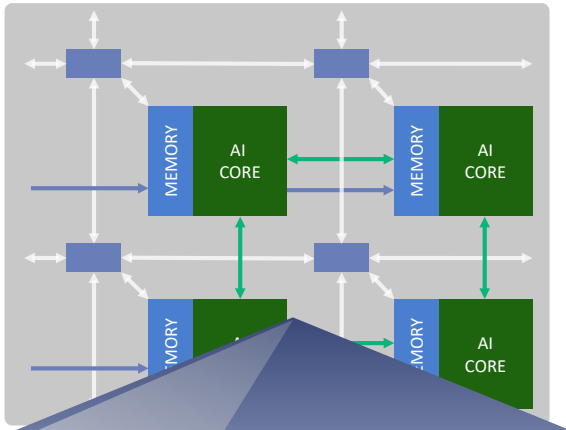
XILINX.

Introducing the AI Engine

SW Programmable

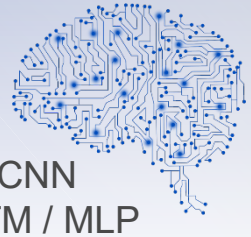
Deterministic

Efficient



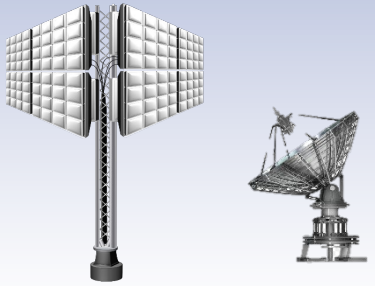
- 1GHz+ Multi-precision Vector Processor
- High bandwidth extensible memory
- Up to 400 AI Engines per device
- 8X Compute Density
- 40% Lower Power

Artificial Intelligence



CNN
LSTM / MLP

Signal Processing



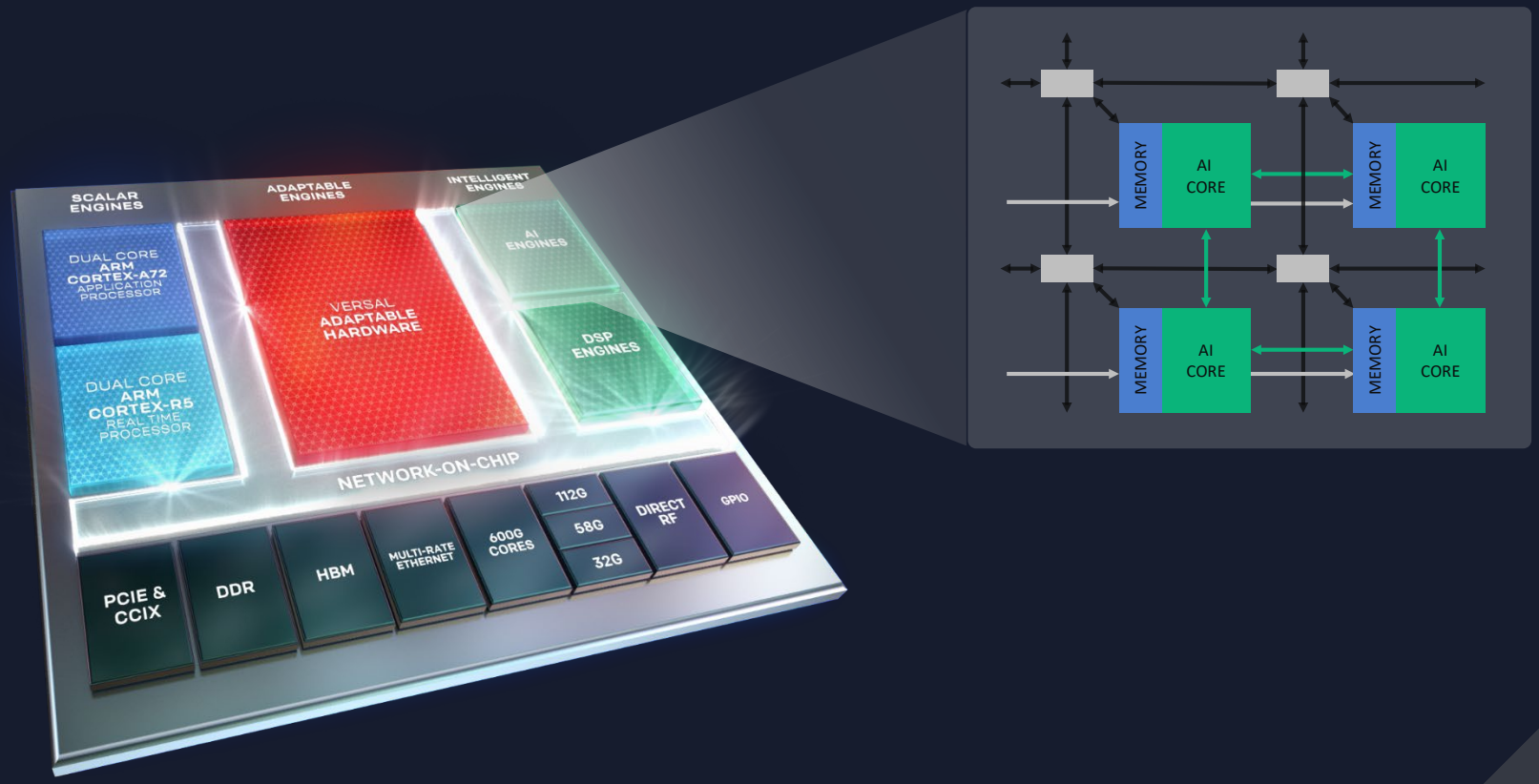
Computer Vision

High performance multi-core processor system

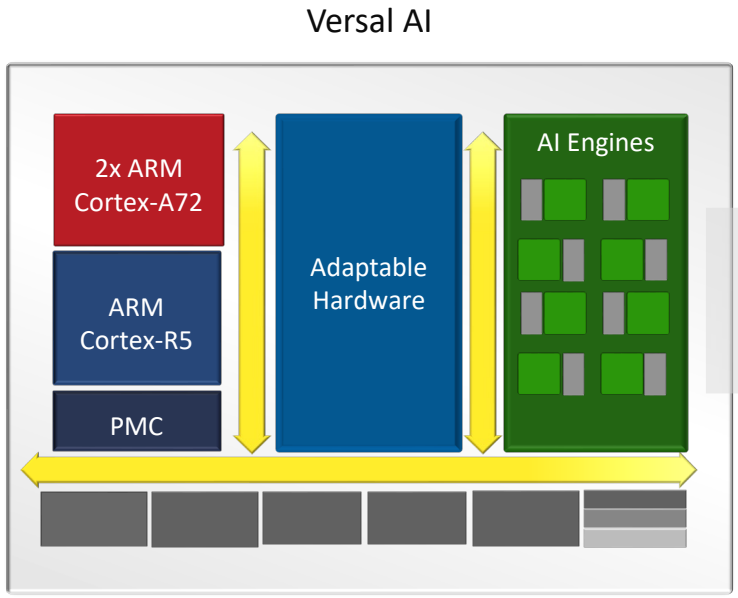
XILINX.

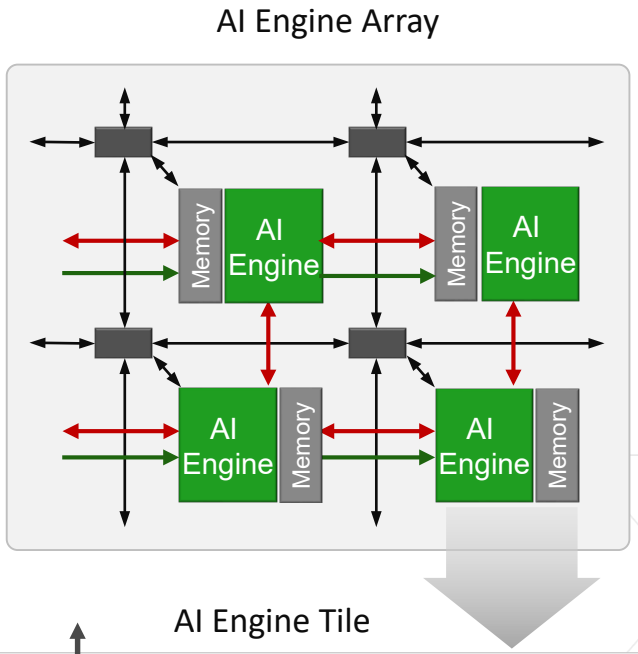
AI Engine (first generation)

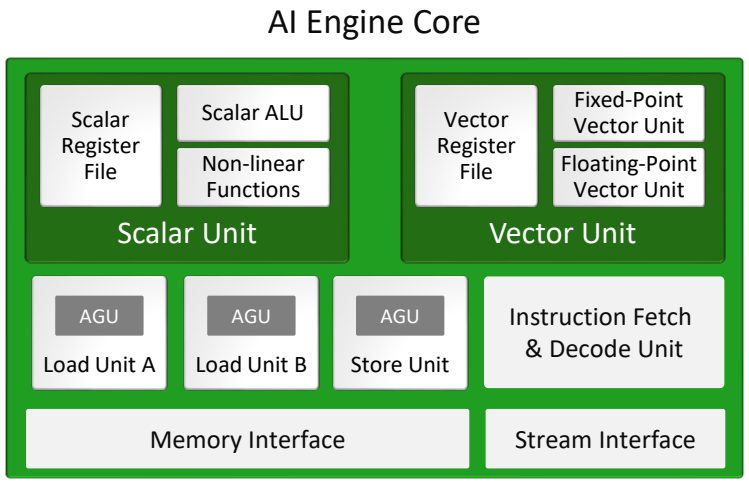
Architecture, Programming & Applications

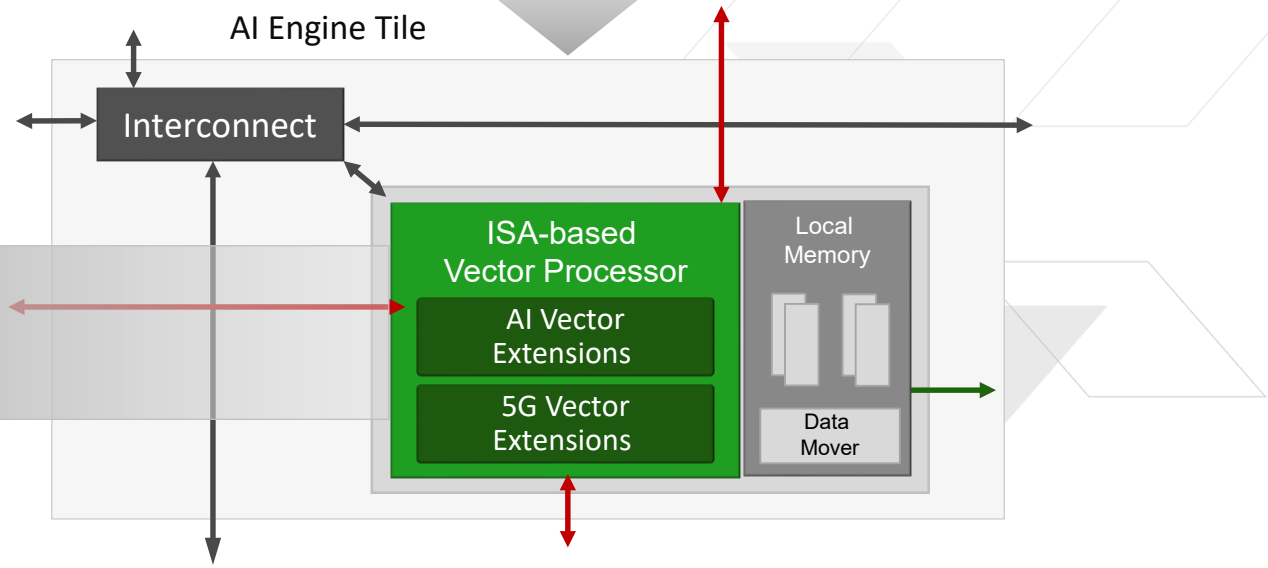


AI Engine: Terminology







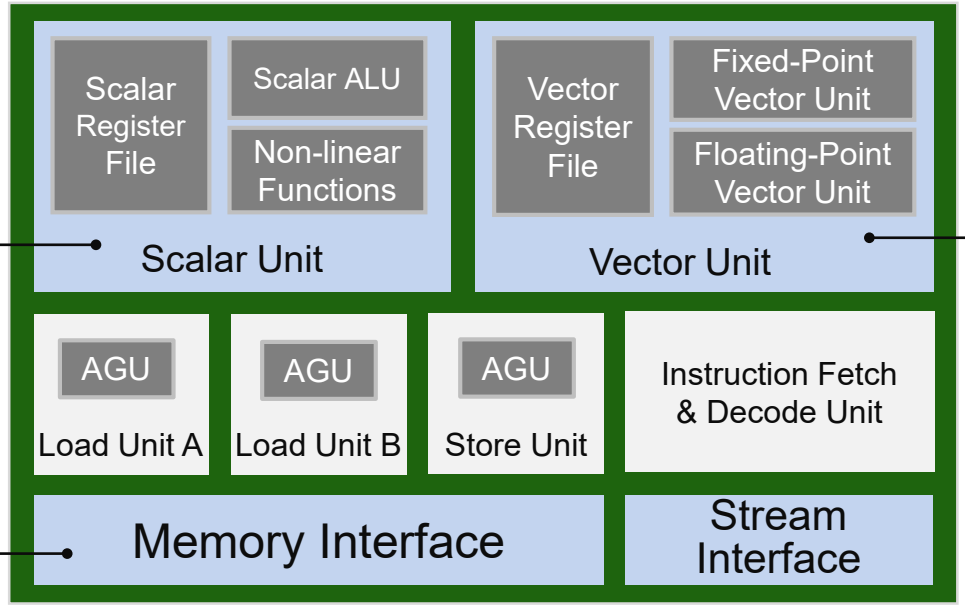


XILINX.

AI Engine: Processor Core

32-bit Scalar RISC Processor

Local, Shareable Memory
- 32KB Local, 128KB Addressable



Vector Processor
512-bit SIMD Datapath

Instruction Parallelism: VLIW

7+ operations / clock cycle
- 2 Vector Loads / 1 Mult / 1 Store
- 2 Scalar Ops / Stream Access

Highly Parallel

Data Parallelism: SIMD

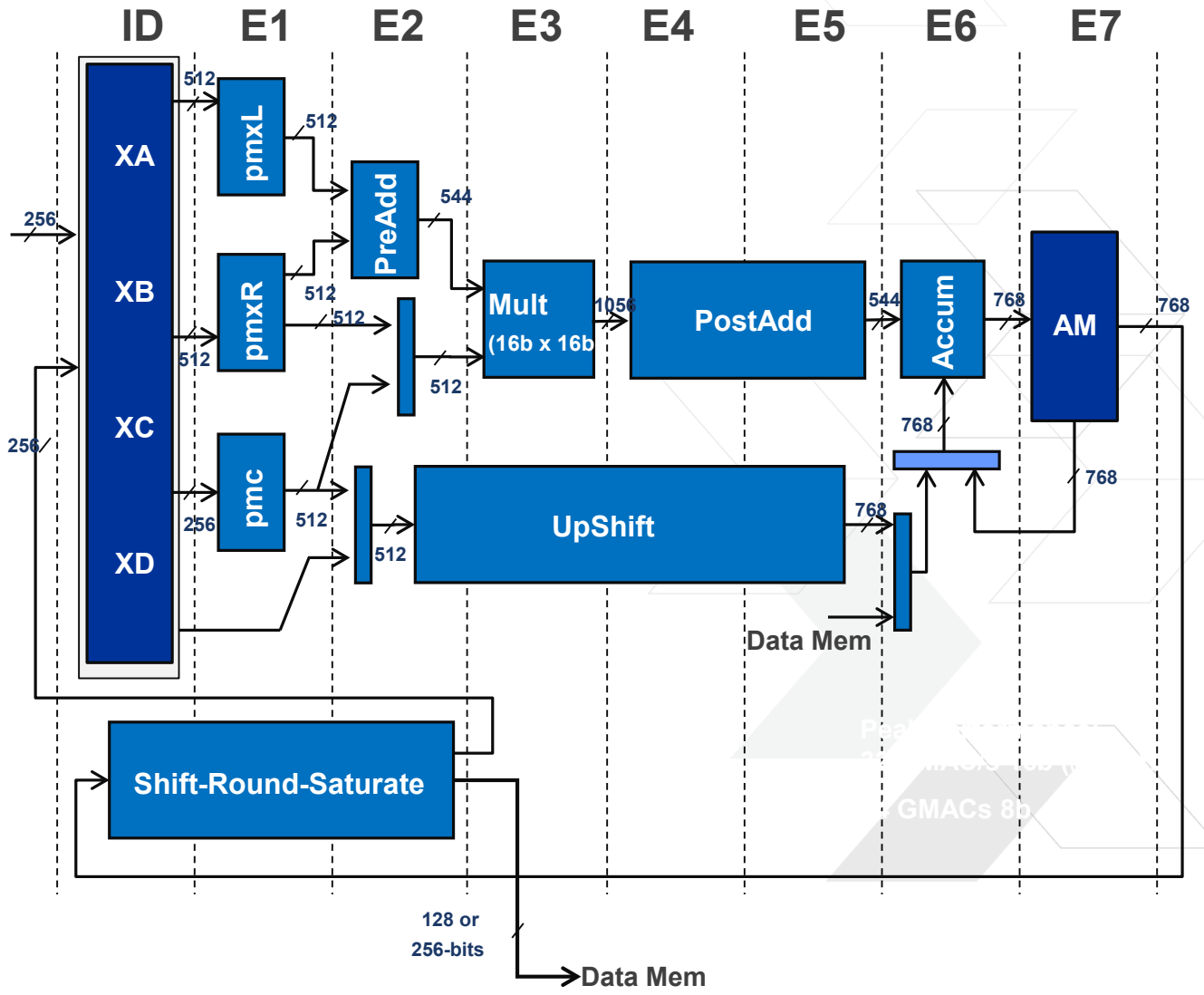
Multiple vector lanes
- Vector Datapath
- 8 / 16 / 32-bit & SPFP operands

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

XILINX.

AIE Core: Fixed-Point Vector Unit

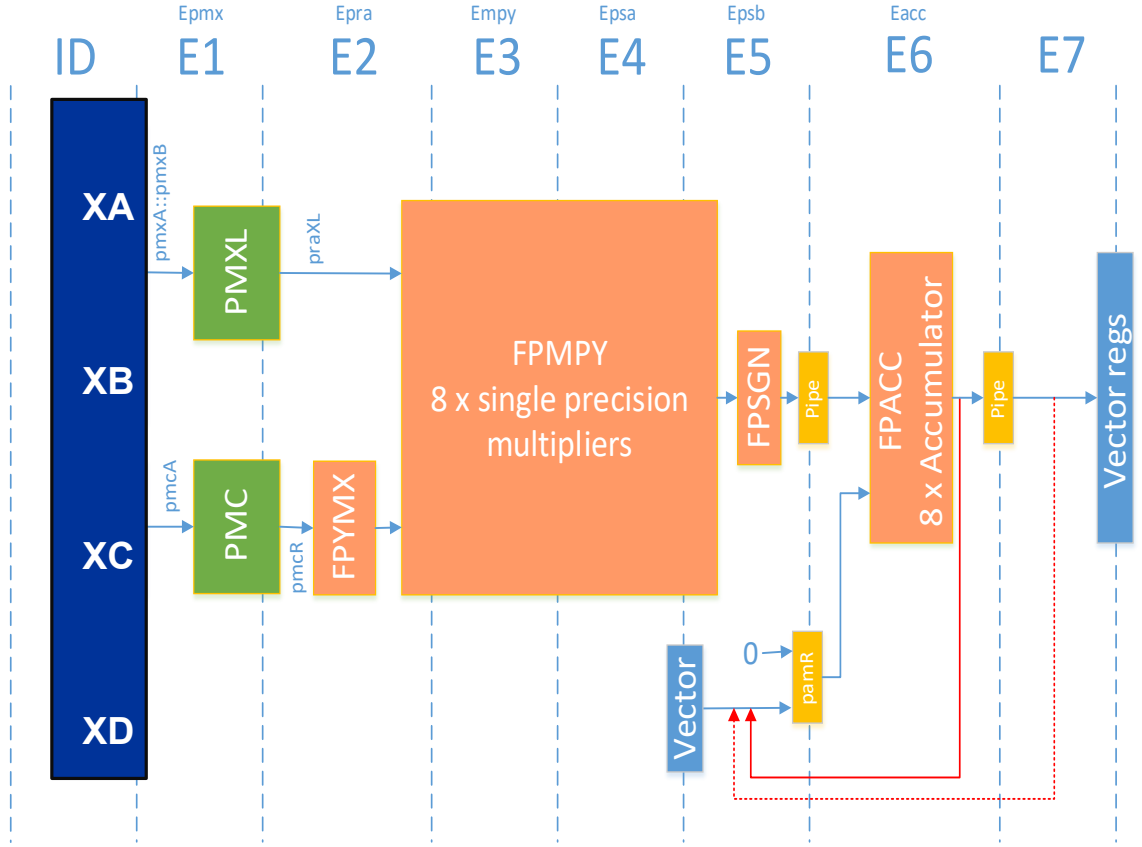
- **Vector fixed-point/integer unit**
  - Multiple precision for complex and real operand
  - Full permute unit (32-bit granularity)
  - Pre-adder
  - Multiplier
  - Two-step post-adding
  - Accumulator
  - Shift, round, saturate



XILINX.

AIE Core: Floating-point Vector Unit

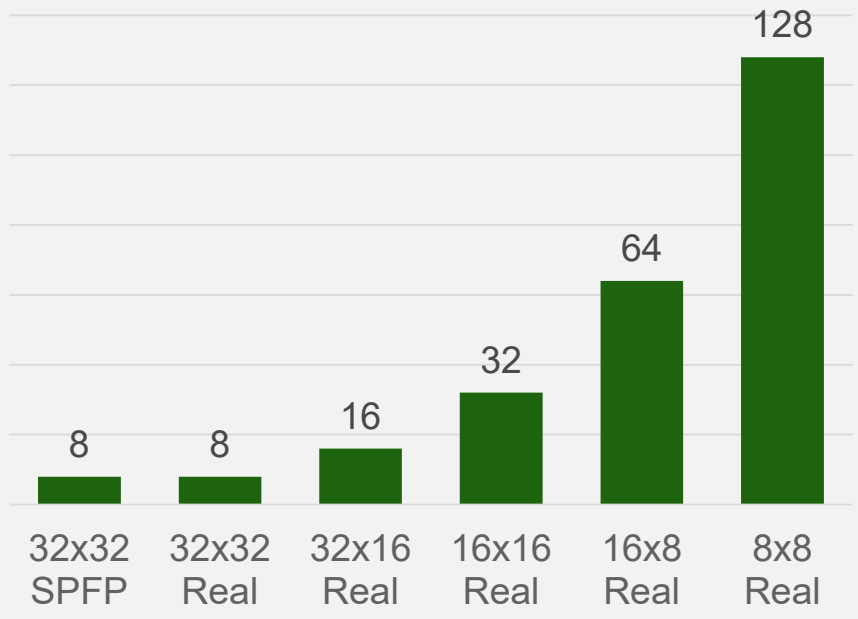
- **Vector floating-point Unit**
  - Single precision
  - 8 multiply-accumulate per cycle
  - Sign change (FPSGN) is on per-lane basis
- **Exceptions**
  - Per-lane exceptions
  - ZERO, INFINITY, INEXACT, HUGE_INT
  - TINY, HUGE, INVALID , DIVIDE_BY_ZERO
  - The latter can be converted into events broadcasted to PL-AI interface and sent as interrupts to PS/PMC



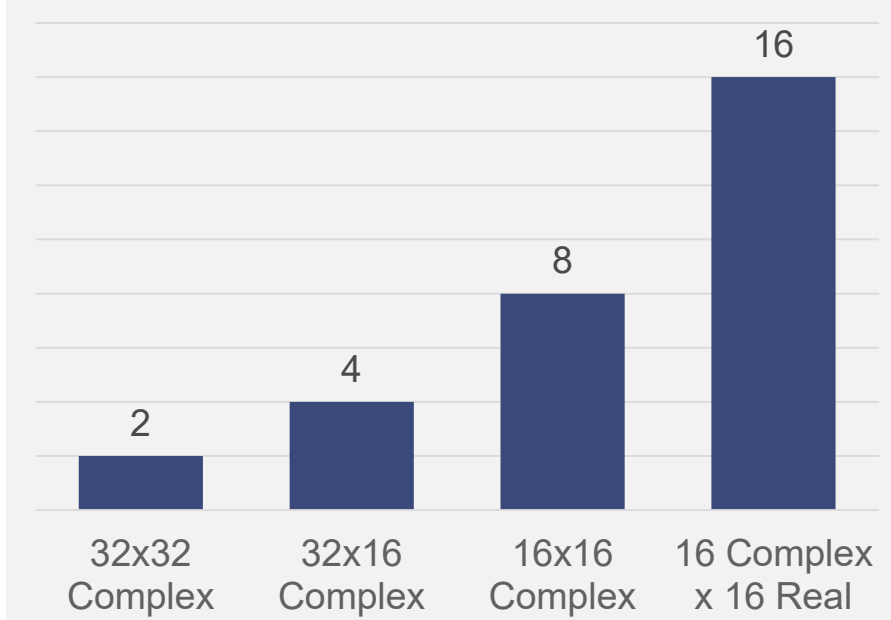
XILINX.

Multi-Precision Support

AI Data Types



Signal Processing Data Types

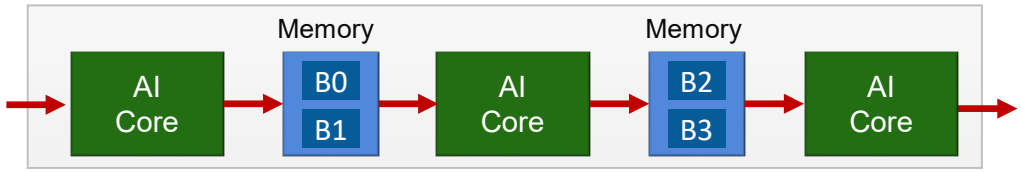


XILINX.

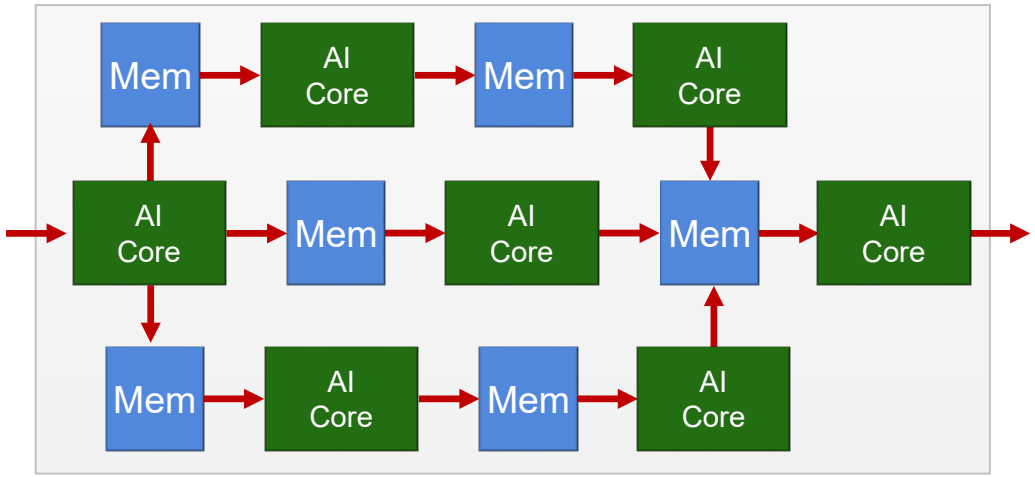
Data Movement Architecture

Memory Communication

Dataflow Pipeline



Dataflow Graph



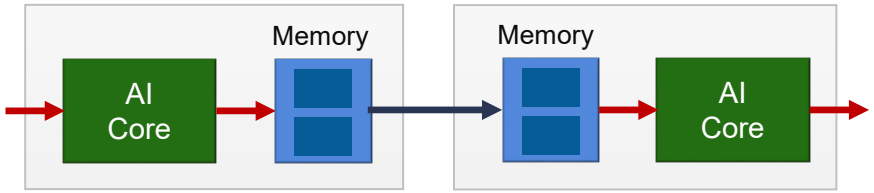
Memory Interface

Stream Interface

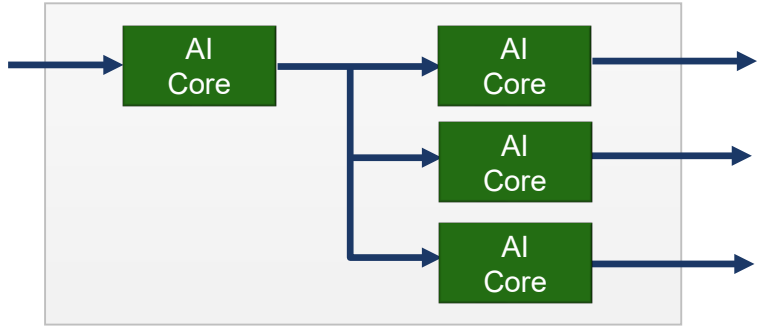
Cascade Interface

Streaming Communication

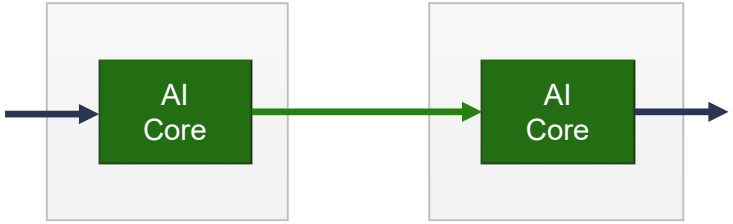
Non-Neighbor



Streaming Multicast

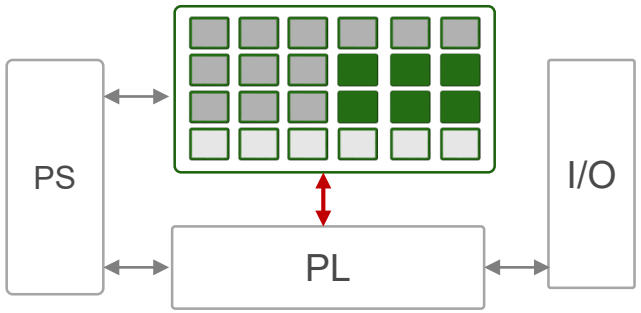


Cascade Streaming

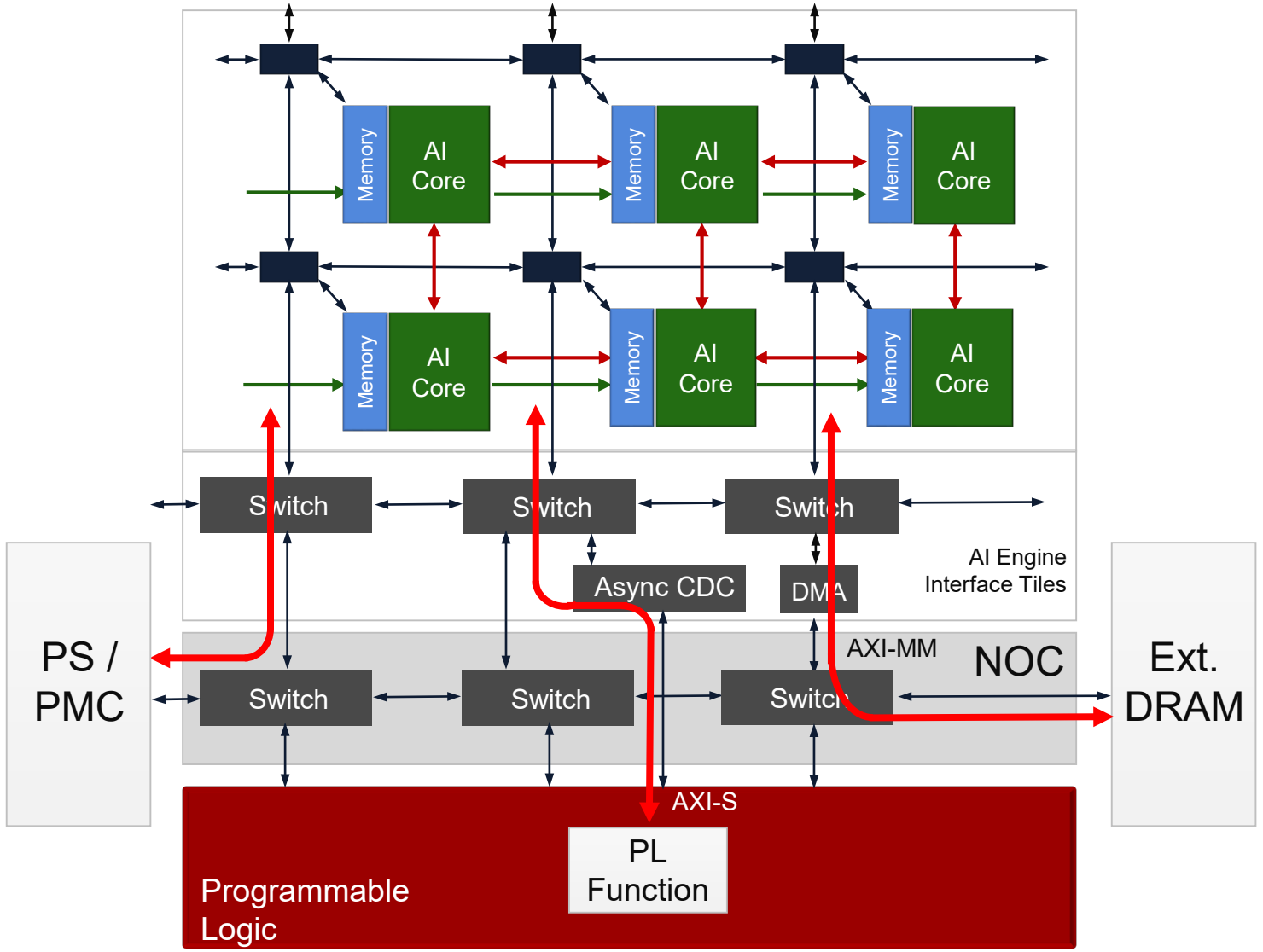


XILINX

AI Engine Integration with Versal



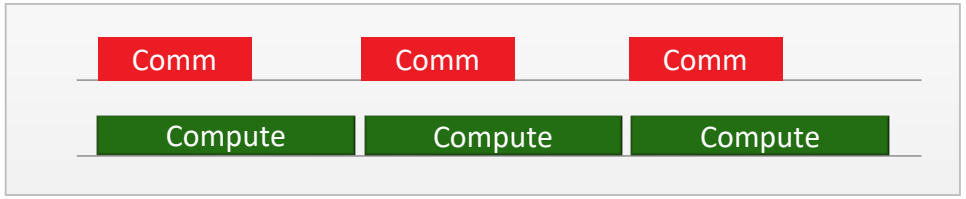
- **TB/s of Interface Bandwidth**
  - AI Engine to Programmable Logic
  - AI Engine to NOC
- **Leveraging NOC connectivity**
  - PS manages Config / Debug / Trace
  - AI Engine to DRAM (no PL req'd)



XILINX.

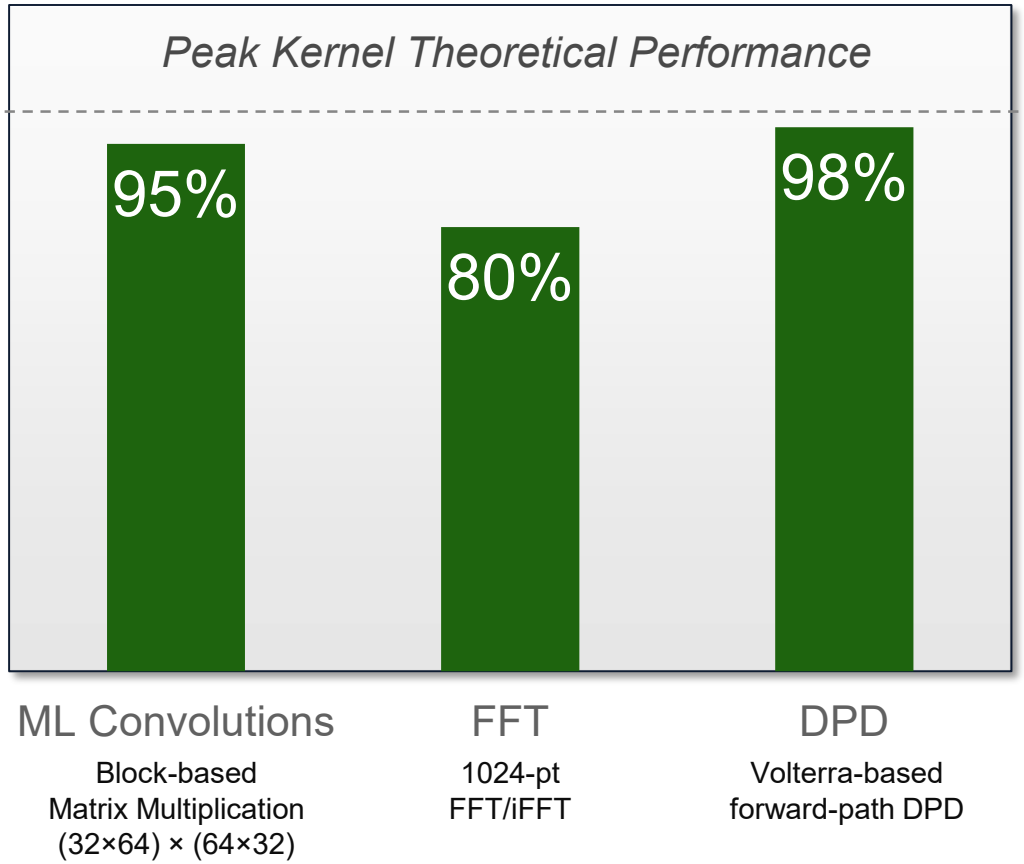
AI Engine Delivers High Compute Efficiency

- **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

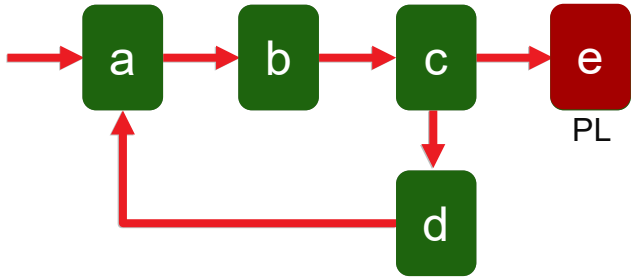
Vector Processor Efficiency



XILINX.

AI Engine Programming Experience: Dataflow Model

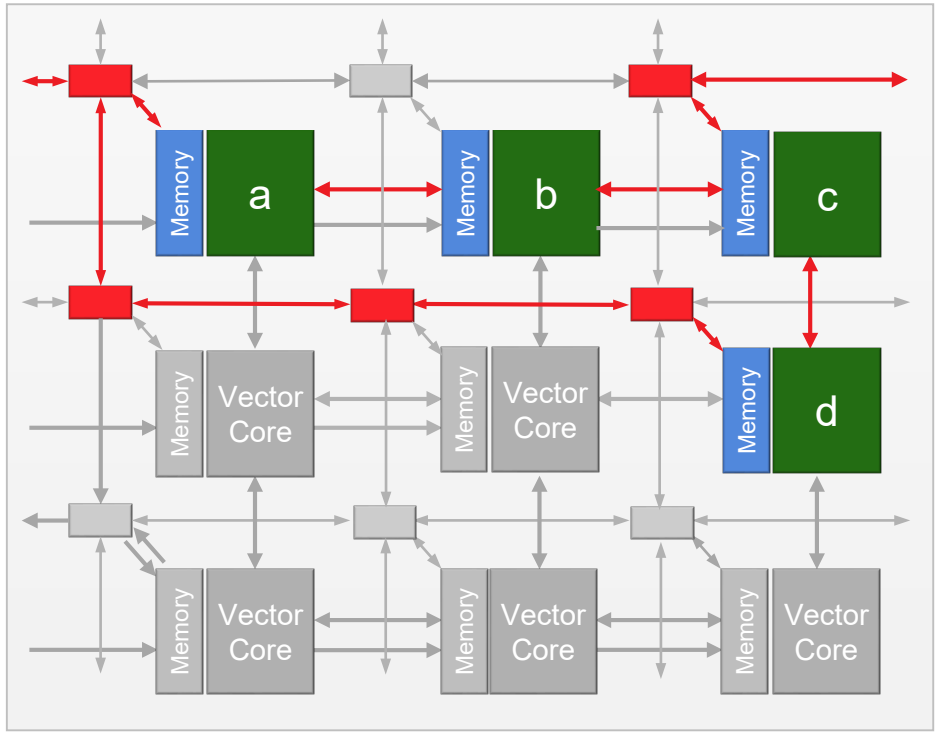
1 User defines dataflow logic



2 User describes dataflow graph using C/C++ APIs

3 Compiler transparently manages placement & interconnect

Physical Mapping to AI Engines



to e

XILINX.

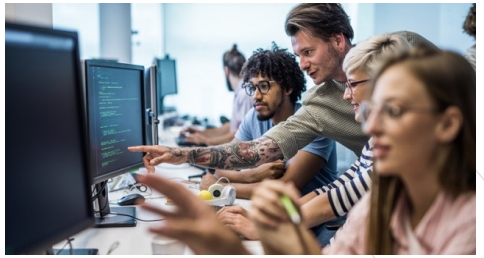
All Developers Can Build and Deploy on All Platforms

Build

Embedded Developers

Enterprise Application Developers

Enterprise Infrastructure Developers

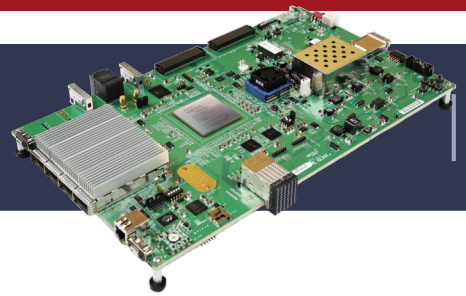


Data & AI Scientists

XILINX VITIS™

Deploy

Zynq-7000



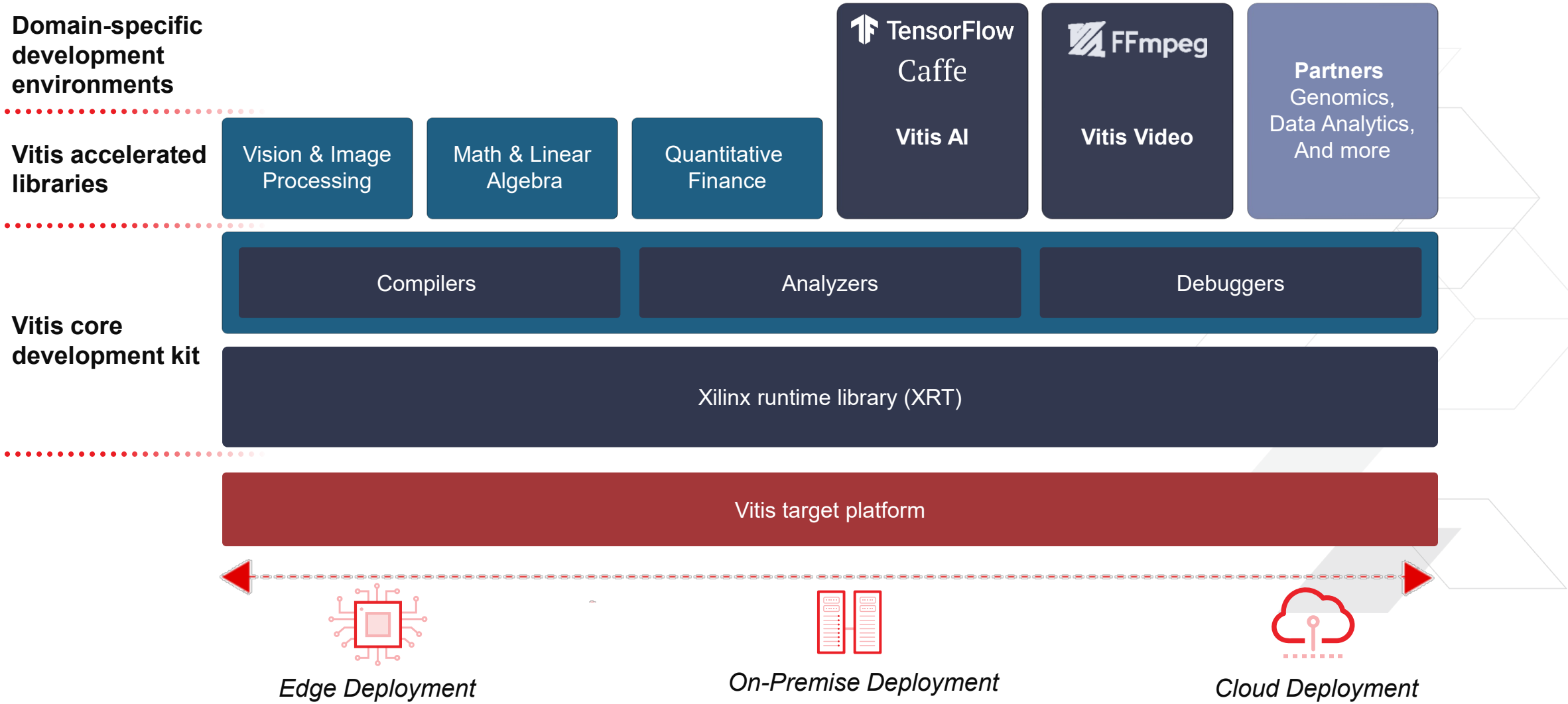
Zynq UltraScale+ MPSoC

Alveo

Versal ACAPs

XILINX

Vitis Unified Software Platform



XILINX.

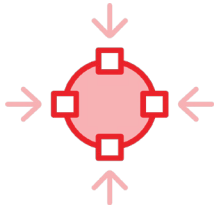
Vitis libraries, including AIE Libraries

Domain-Specific Libraries

Vision & Image

Quantitative Finance

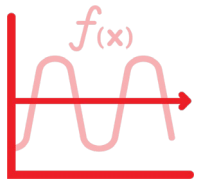
Data Analytics & Database



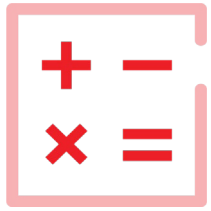
Data Compression

Data Security

Common Libraries

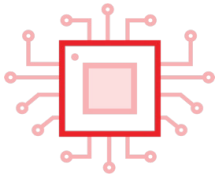


Math



Linear Algebra

Statistics



DSP

Data Management

500+ functions across multiple libraries for performance-optimized out-of-the-box acceleration

XILINX.

AIE Designs for massive MIMO and DFE Platforms

A set of highly efficient and scalable AIE designs of product quality

- Advanced design techniques published on Xilinx.com as application notes with source code

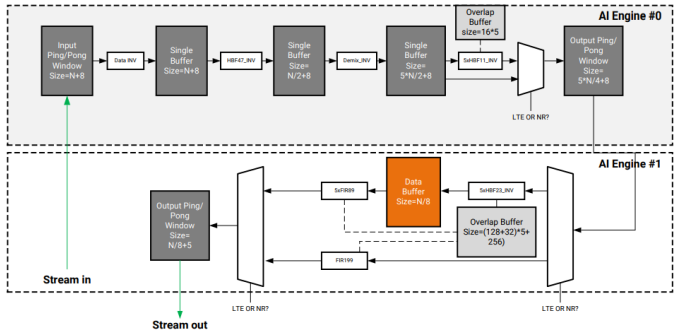
Application Note: Versal™ AI Core Devices

XILINX

Digital Down-conversion Chain Implementation on AI Engine

XAPP1351 (v1.0) February 15, 2021

Figure 17: **DDC Data Buffers With Data Memory Sharing (Proposed)**



XAPP 1351
Multi-carrier Multi-rate Filters

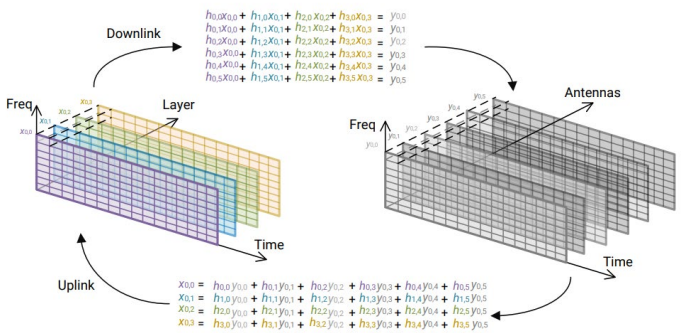
Application Note: Versal™ AI Core Devices

XILINX

Beamforming Implementation on AI Engine

XAPP1352 (v1.0) January 11, 2021

Figure 1: **Beamforming in OFDM Systems**



XAPP 1352
Scalable beamforming

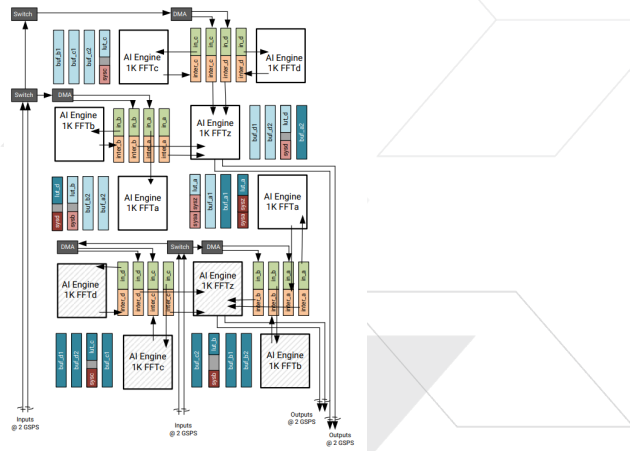
Application Note: Versal™ AI Core Devices

XILINX

Block-by-Block Configurable Fast Fourier Transform Implementation on AI Engine

XAPP1356 (v1.0) January 11, 2021

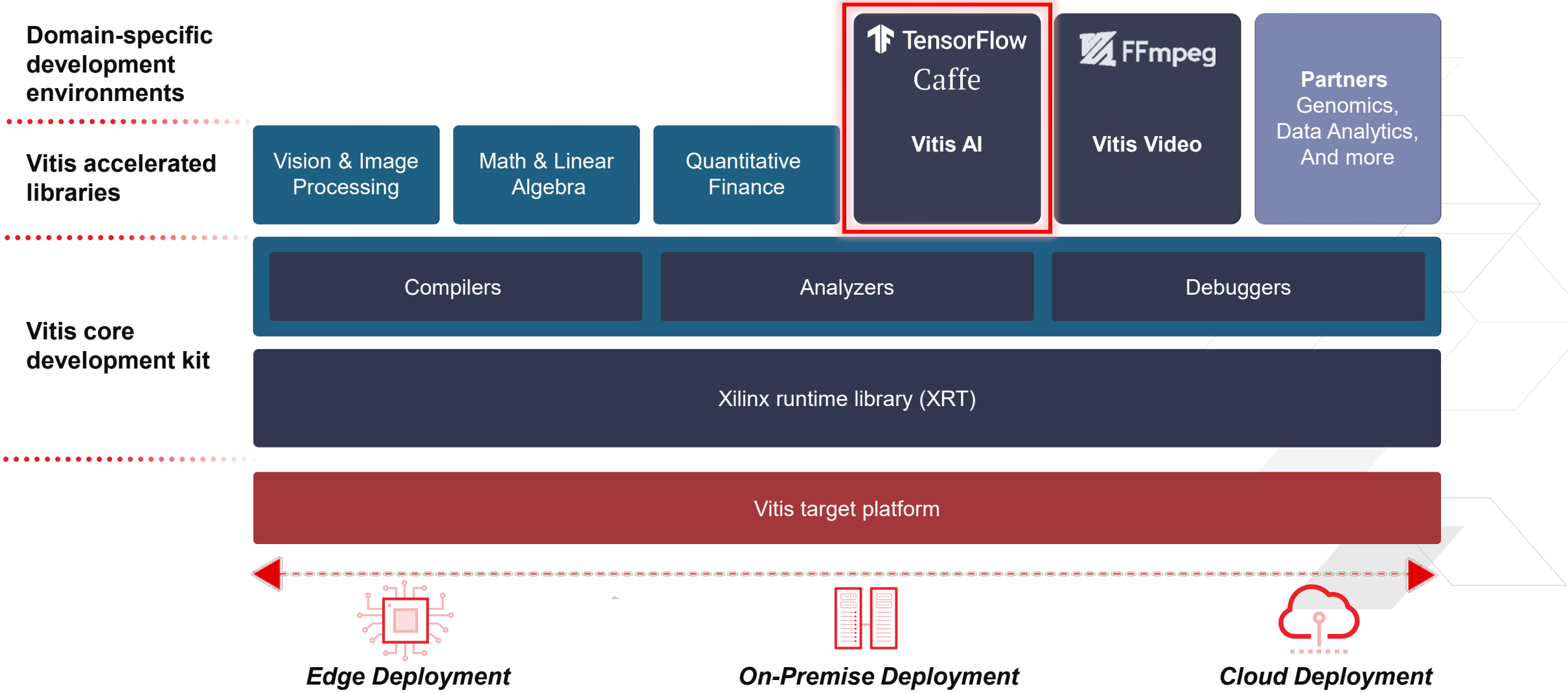
Figure 7: **Two FFT Modules Packed into a 5x2 AI Engine Array**



XAPP 1356
Configurable FFT/IFFT

XILINX

Vitis Unified Software Platform



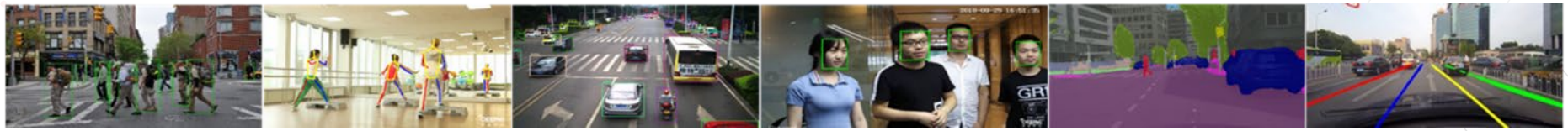
XILINX.

Vitis AI: Deep Learning Acceleration

Frameworks

TensorFlow Caffe PyTorch

Vitis AI models



Vitis AI development kit

AI Optimizer | AI Quantizer | AI Compiler | AI Profiler | AI Library

Xilinx runtime library (XRT)

Deep Learning Processing Unit (DPU)

CNN DPU | LSTM DPU | MLP DPU

Edge Deployment

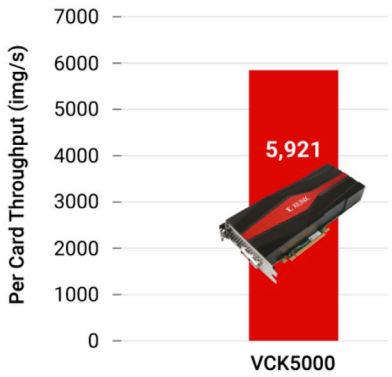
On-Premise Deployment

Cloud Deployment

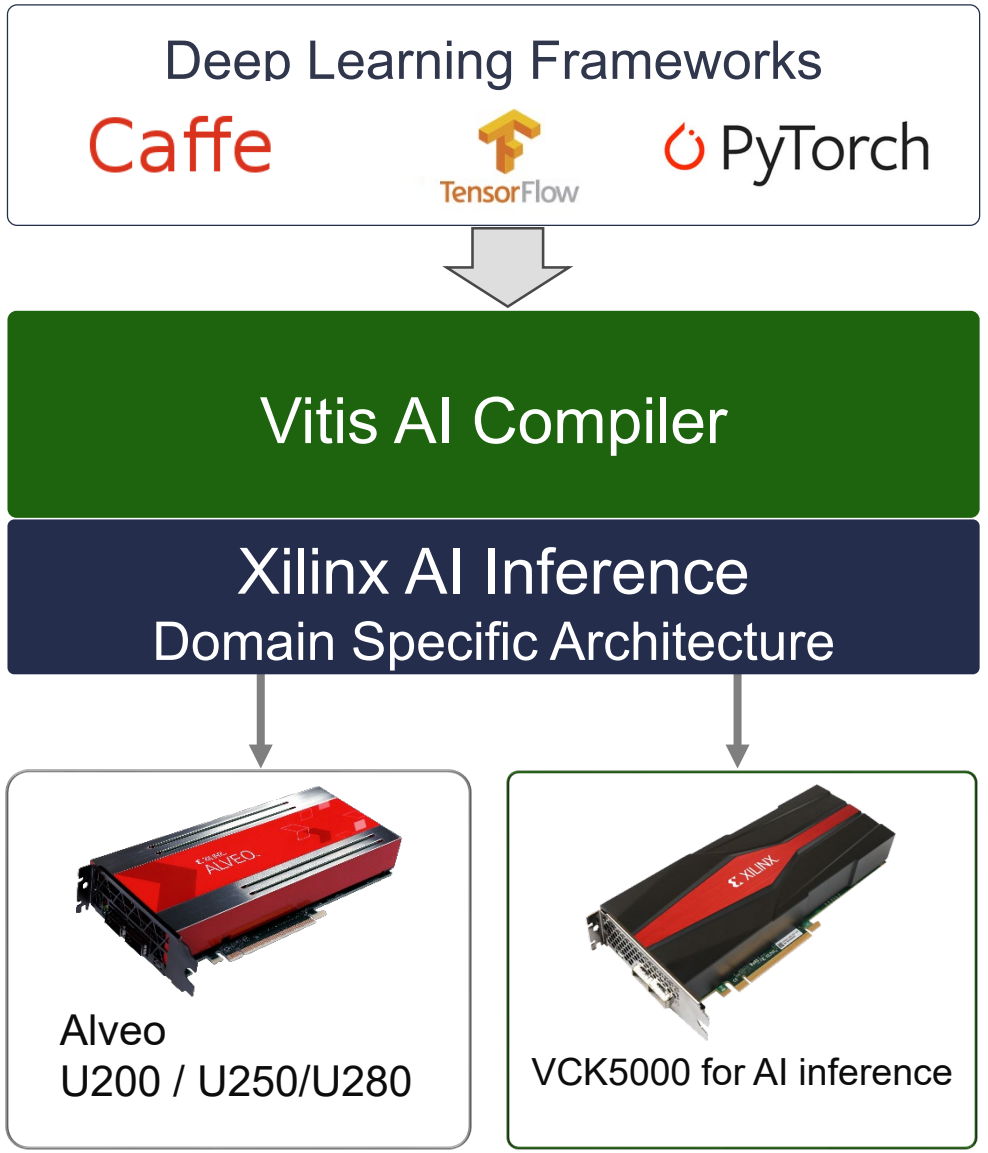
XILINX

Vitis AI: Inference in the Data Center

1. User works in Framework of choice
   - Develop & train custom network
   - User provides trained model
2. Xilinx Vitis AI Compiler implements network
   - Targets AI Inference Domain Specific Architecture
   - Quantize, merge layers, prune
   - Compile to AI Engines
3. Scalable across hardware targets
   - AI engine performance benefit 8x -10x



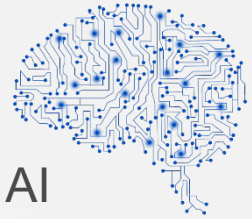
ML commons v1.0 Data Center Closed Division Server Resnet-50



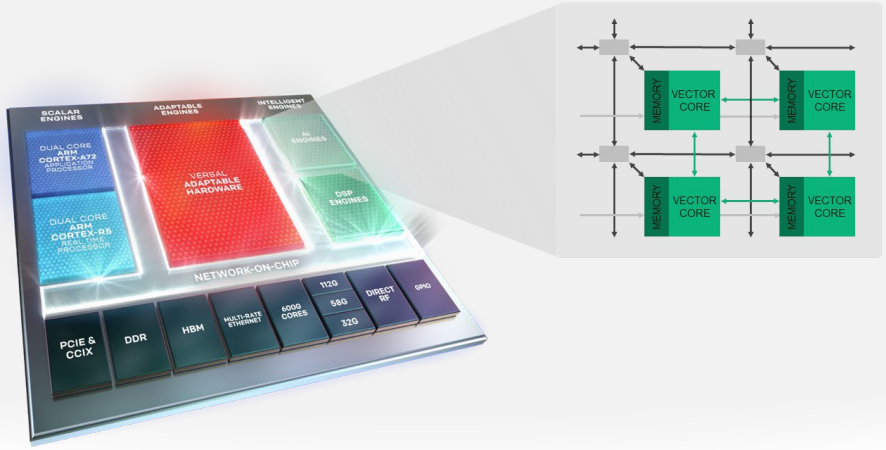
XILINX

AI Engine: Accelerating AI Inference & Signal Processing

10x



AI
Inference



5x

Signal
Processing

Software Programmable

- Frameworks & C/C++
- SW Compile, Debug & Deploy

Deterministic

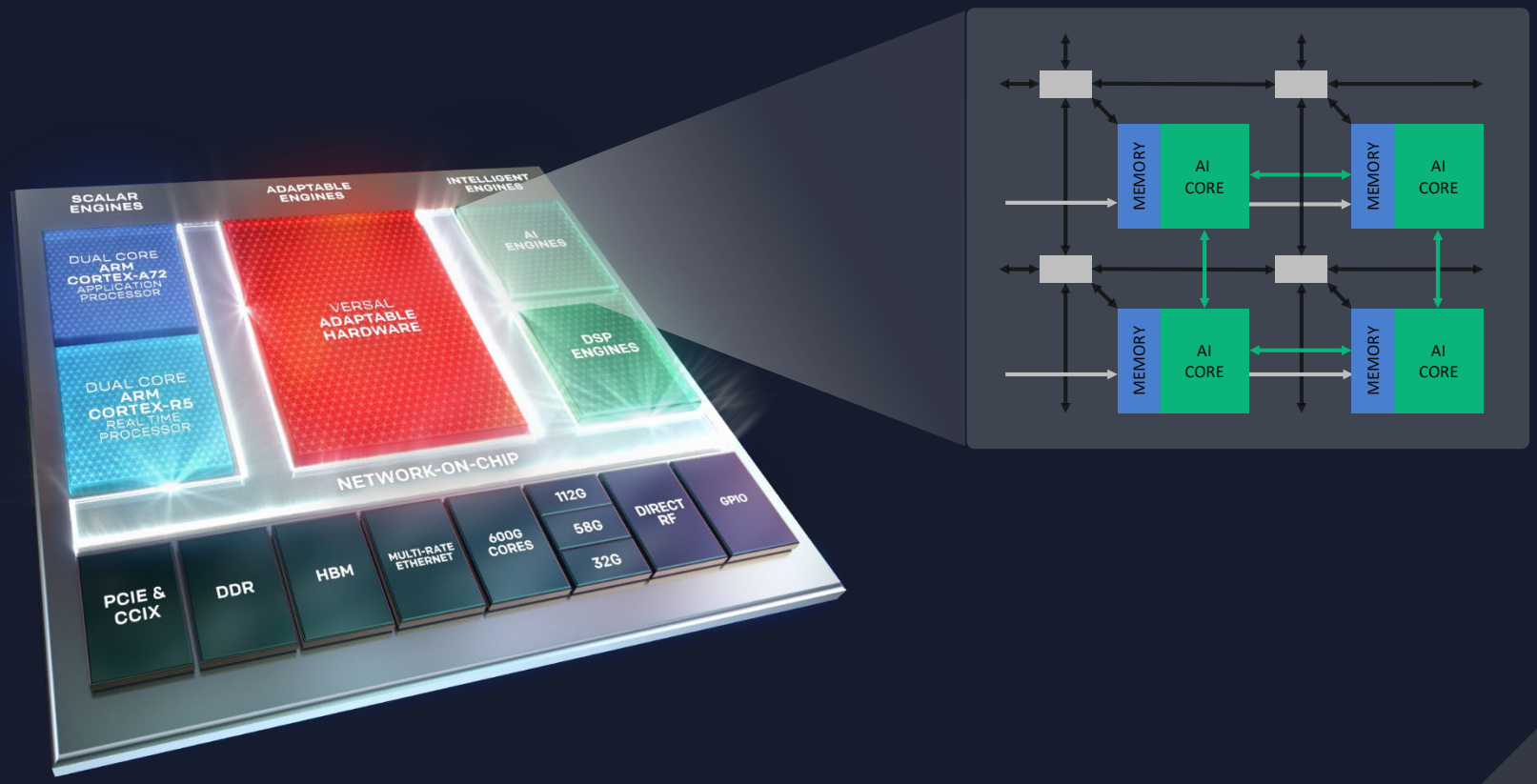
- Max throughput w/ low latency
- Real-time inference leadership

Efficient

- Up to 8X compute density
- At ~40% lower power

XILINX.

Motivation for second generation: AI - ML



AI/ML Edge Inference: Use-cases and Requirements

Automotive

ADAS

Automated Driving

Industrial Robotics

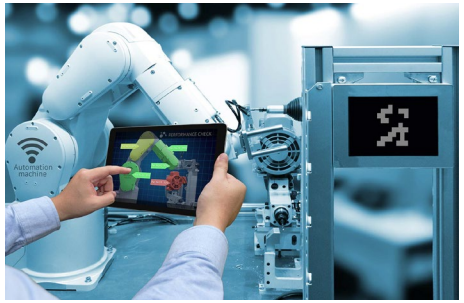
Vision guidance

Environmental awareness

Medical

Ultrasound

Endoscopy, Surgical Robotics



Industrial IoT

Industrial PCs

Smart Grid Controllers

Vision & Smart City

Machine Vision Camera

Edge AI Box

Aerospace & Defense

Unmanned Aerial Vehicles

Multi-Mission Payload Systems

Challenge: Delivering massive AI compute at low-latency and low-power

Versal™ AI Edge: Architecture Overview

XILINX VERSAL™ AI CORE

XILINX VERSAL™ PREMIUM

XILINX VERSAL™ AI EDGE

ADAPTABLE TO MULTIPLE WORKLOADS

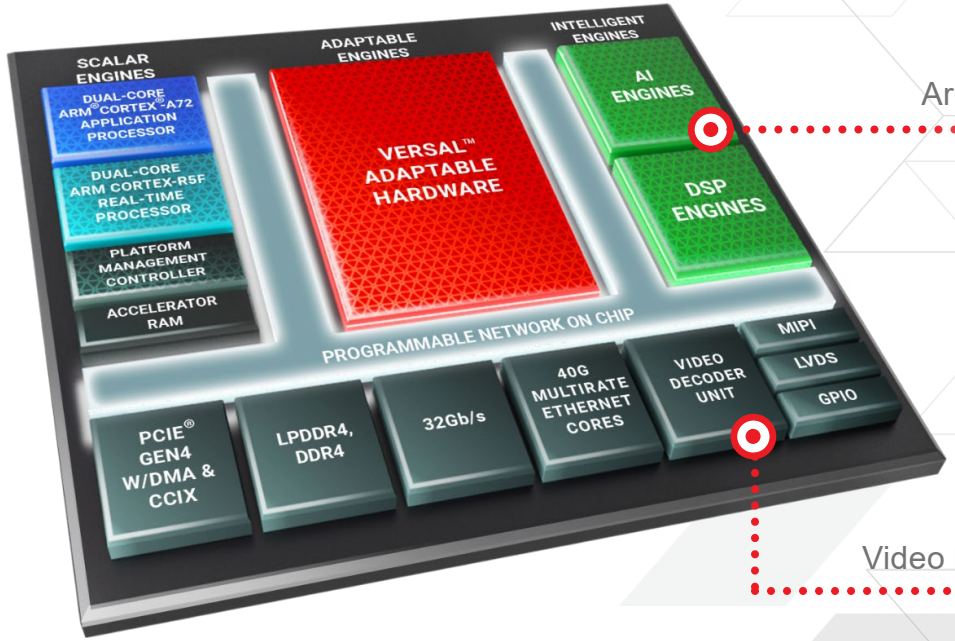
- Versal AI Core
- Versal Premium
- Versal AI Edge

COMPUTE ACCELERATION

- Scalar Engines
- Adaptable Engines
- Intelligent Engines

PLATFORM

- Software programmable NoC
- Platform Management Controller
- Dedicated Interfaces (e.g., PCIe, DDR)



AIE-ML Architecture

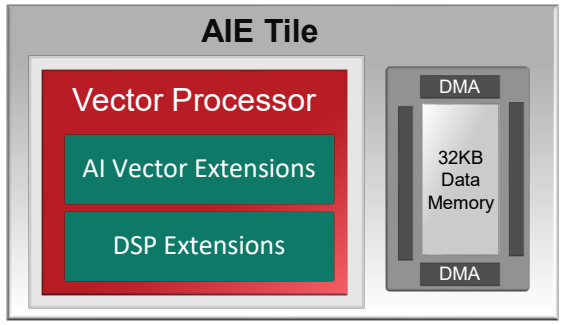
Video Decode Unit

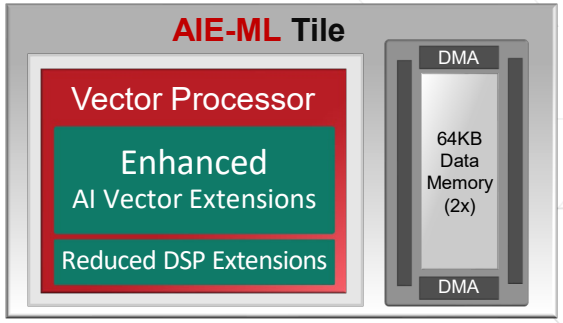
Technology Node: TSMC 7nm FinFET

tsmc 7nm

XILINX

Xilinx AIE Tile: Architecture Features

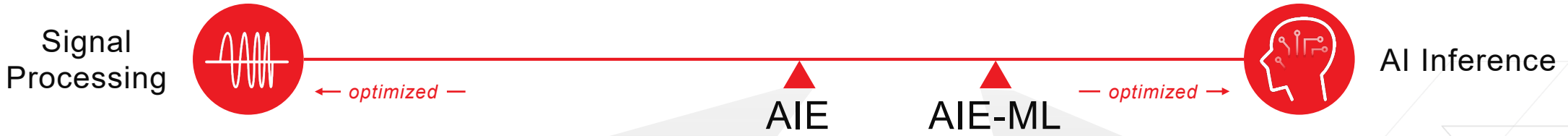




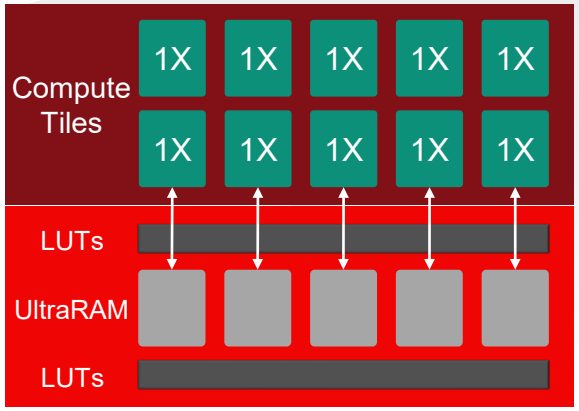
		AIE Tile		AIE-ML Tile
Target Markets		Wireless 5G, AI/ML inference, A&D		AI/ML inference
Compute (Mults / tile)	BFLOAT16	—	New	128
	INT8	128	2X	256
	INT4	—	New	512
Tile Local Data Memory		32 KB	2X	64 KB
AIE Array Interconnect B/W		1X	1X	1X
Compression and Sparsity		No		Yes
Scratchpad On-Chip Memory		PL uRAM		AIE Memory (512KB/tile)

XILINX.

Intelligent Engines Optimized for Any AI Application



AI Engine Architecture



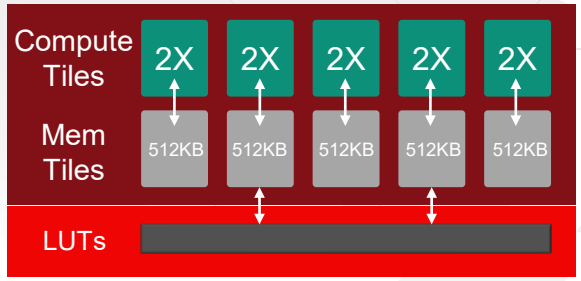
- Optimized for signal processing AND ML
- Flexibility for high performance DSP applications
- Native support for INT8, INT16, FP32

AIE	OPS / Tile	AIE-ML
256	INT4	1024
256	INT8	512
64	INT16	128
16	BFLOAT16	256
16	INT32	
16	FP32	42*

*Via software emulation

KB / Tile		
32	Data Memory	64
16	Program Memory	16

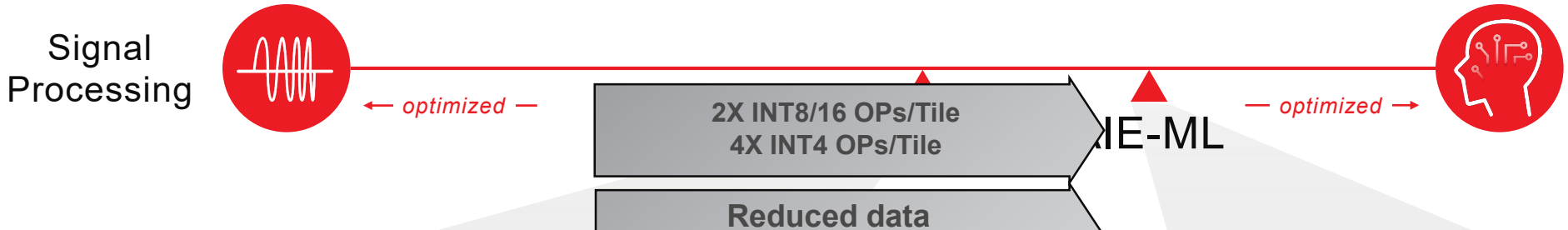
AIE-ML Architecture



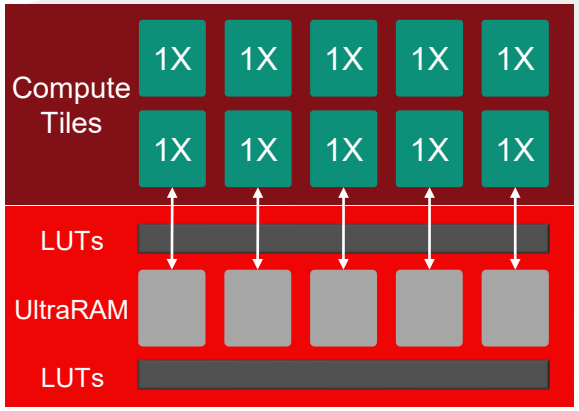
- Optimized for ML Inference Applications
- Maximum AI/ML compute with reduced footprint
- Native support for INT4, INT8, INT16, bfloat16
- Fine grained sparsity HW optimization
- Enhanced FFT & complex math support

XILINX

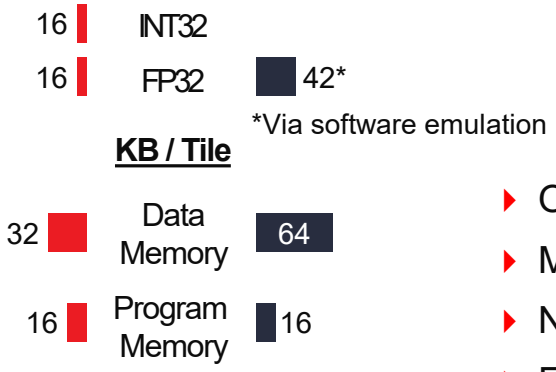
Intelligent Engines Optimized for Any AI Application



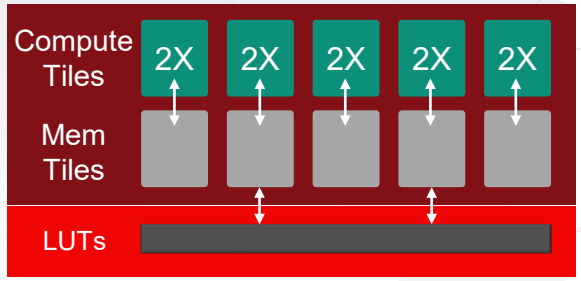
AI Engine Architecture



- Optimized for signal processing AND ML
- Flexibility for high performance DSP applications
- Native support for INT8, INT16, FP32



AIE-ML Architecture



- Optimized for ML Inference Applications
- Maximum AI/ML compute with reduced footprint
- Native support for INT4, INT8, INT16, bfloat16
- Fine grained sparsity HW optimization
- Enhanced FFT & complex math support

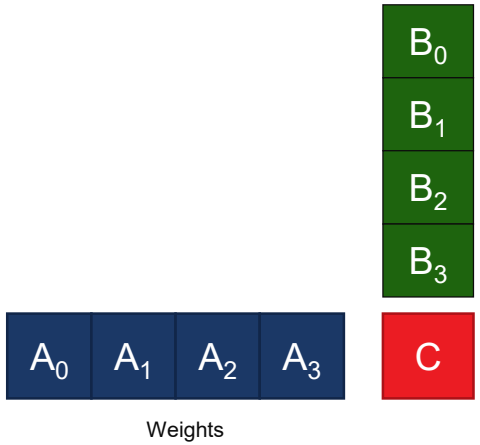
XILINX

Flexible Dataflow: AIE-ML Array Examples

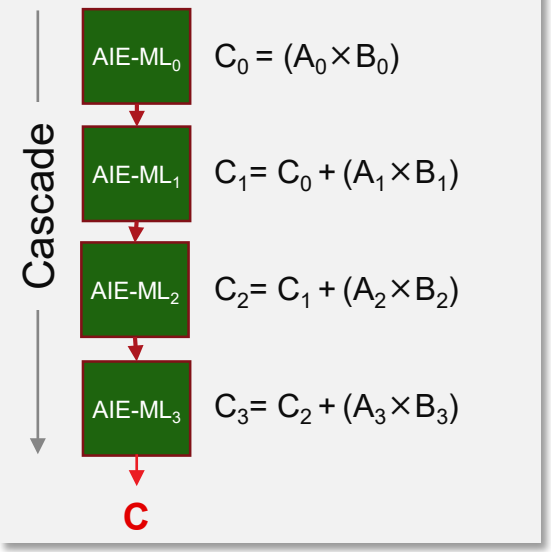
DATA REDUCTION

$$C = (A_0 \times B_0) + (A_1 \times B_1) + (A_2 \times B_2) + (A_3 \times B_3)$$

COMPUTE **SINGLE** OUTPUT MATRIX

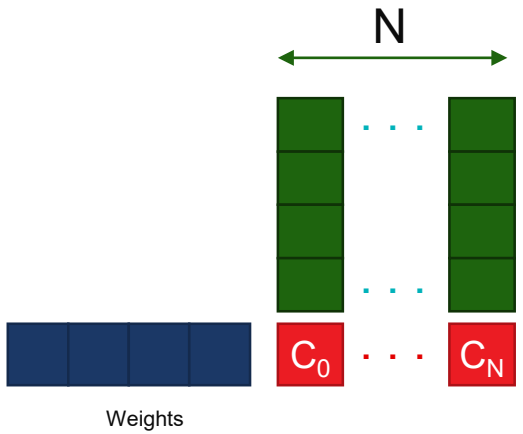


AIE-ML IMPLEMENTATION

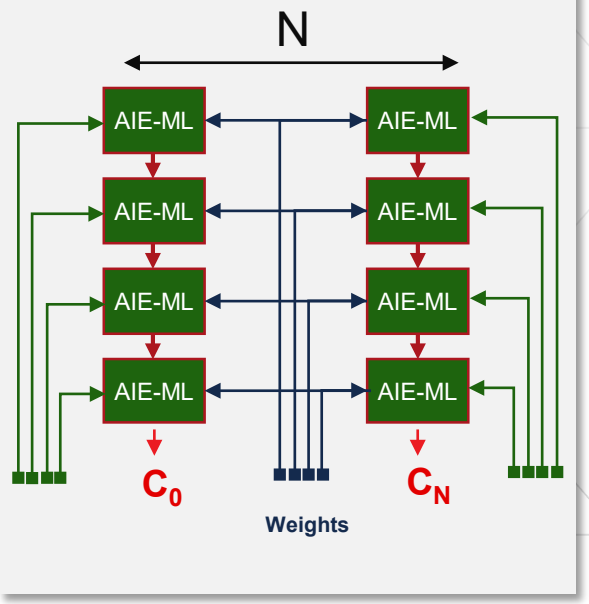


DATA MULTICASTING

COMPUTE **MULTIPLE** OUTPUT MATRICES

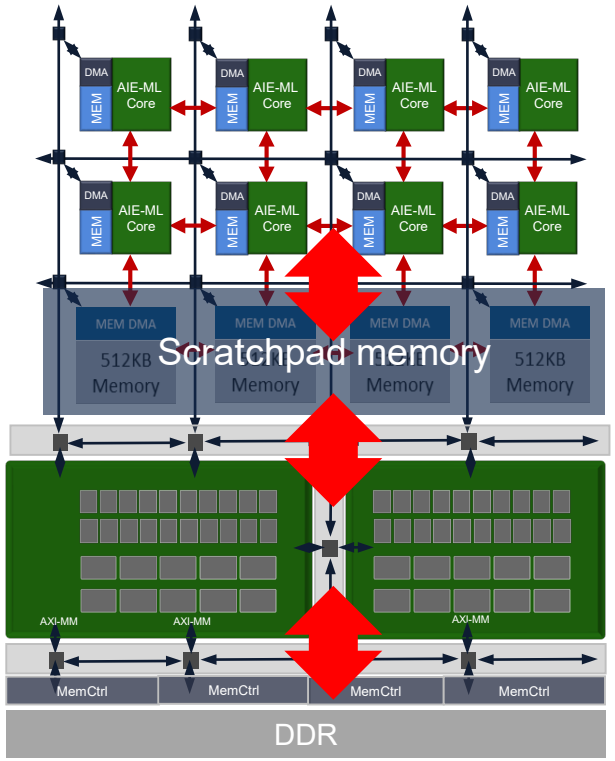


AIE-ML IMPLEMENTATION

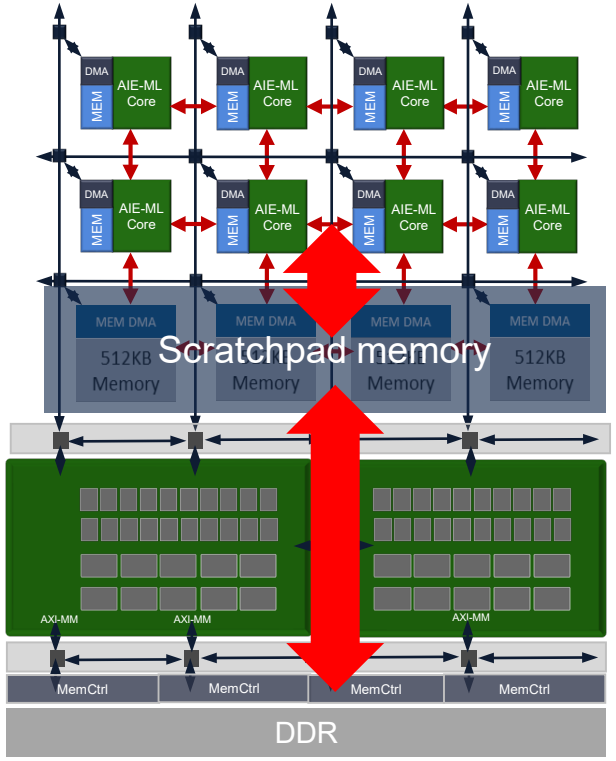


XILINX.

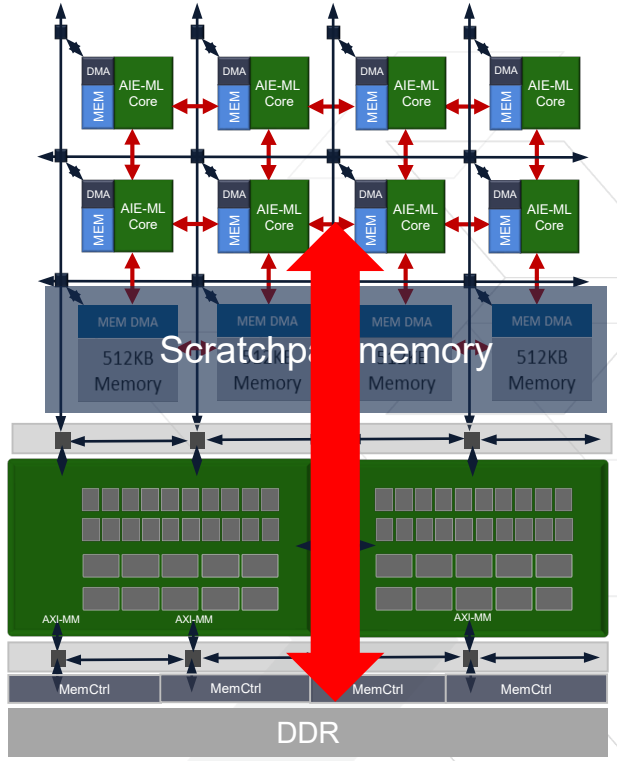
Flexible Dataflow: Device-level Examples



DDR ↔ Scratchpad using Programmable Logic



DDR ↔ Scratchpad using NoC

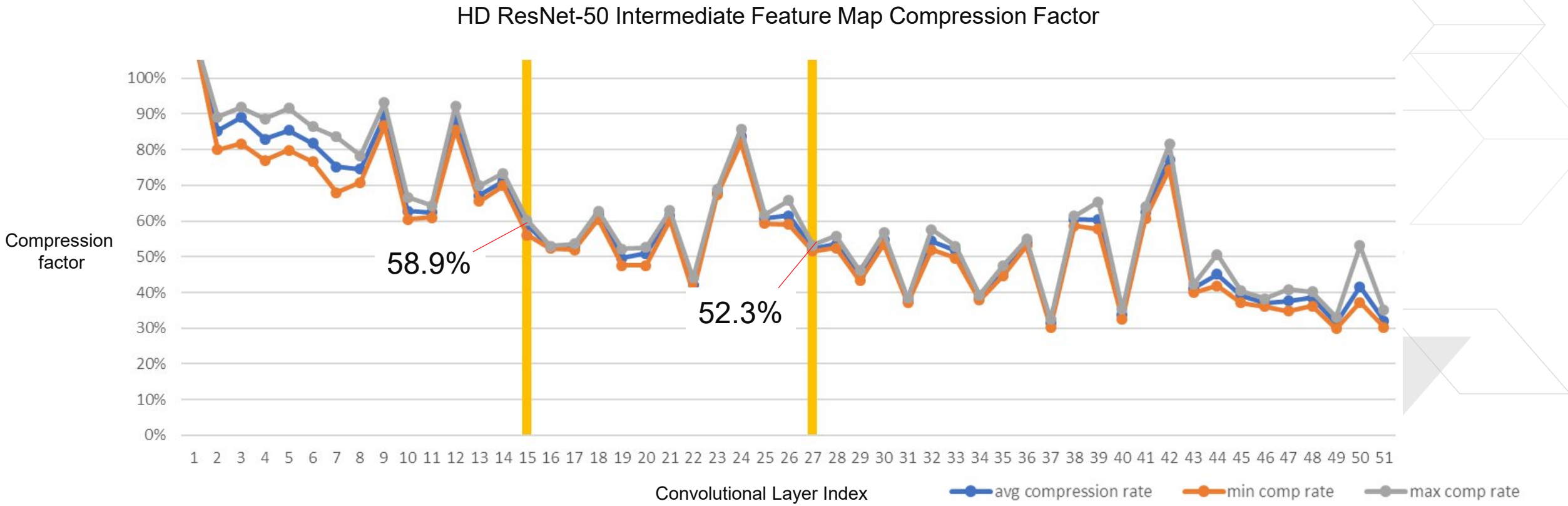


DDR ↔ AIE-ML Tiles using NoC

XILINX.

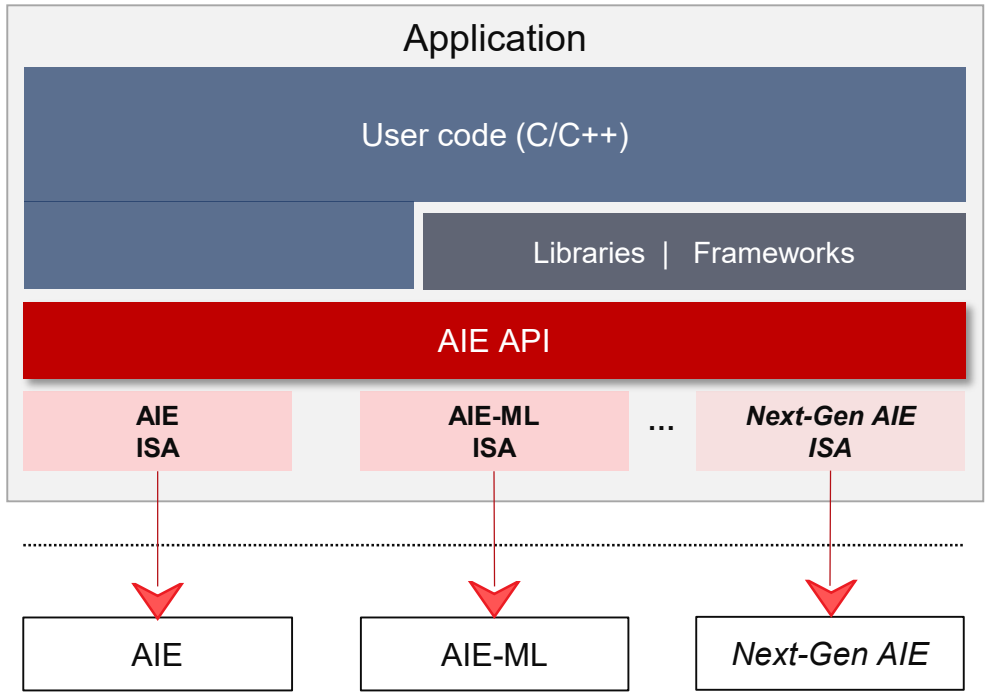
Compression to DDR: Acceleration of activation Spill/Restore

- **Reduction in external memory bandwidth using compression**
  - Quantized int8 weights, int8 activations
  - HD images require tiling, with spill/restore to DDR
    - Without compression **~56GB/s**
    - With compression **~29GB/s**



XILINX.

AIE Architecture: Source Code Portability



- **Backward compatible**
- **Evolving**
  - **New devices can add new functionality**
- **Efficient**
  - **Applications can be optimized to maximize compute efficiency**

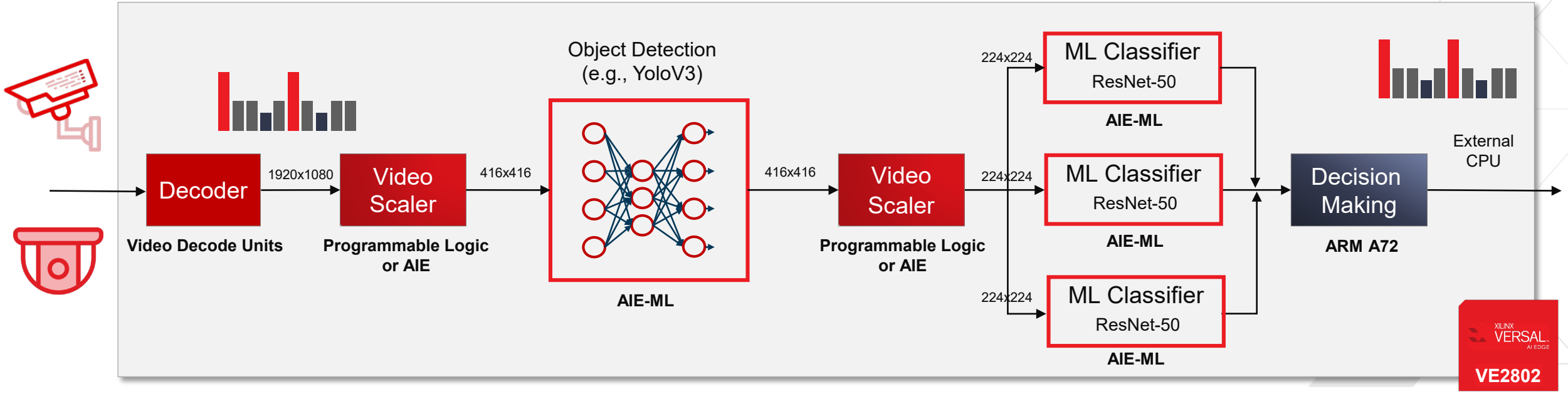
Portable common high-level API across the AIE architecture

Smart World Application: Use-Case Mapping

Smart Retail

Smart Hospital

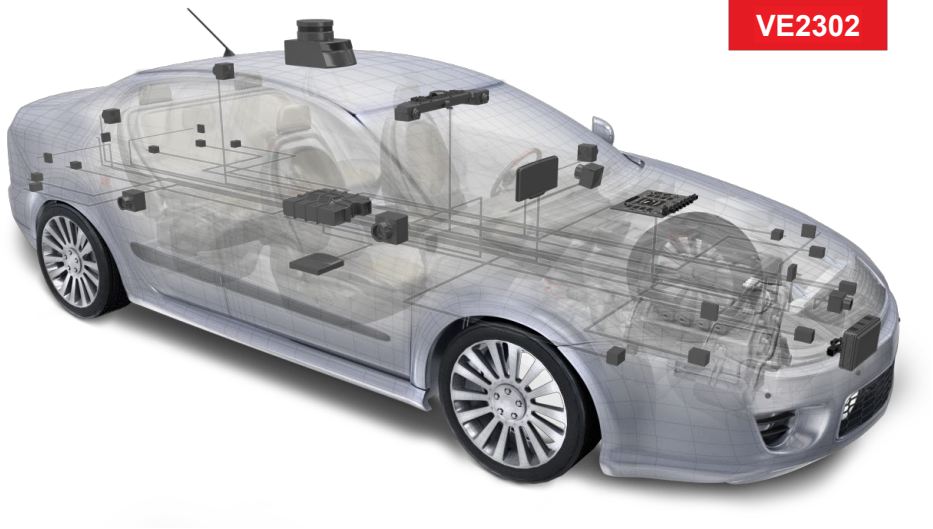
Smart City



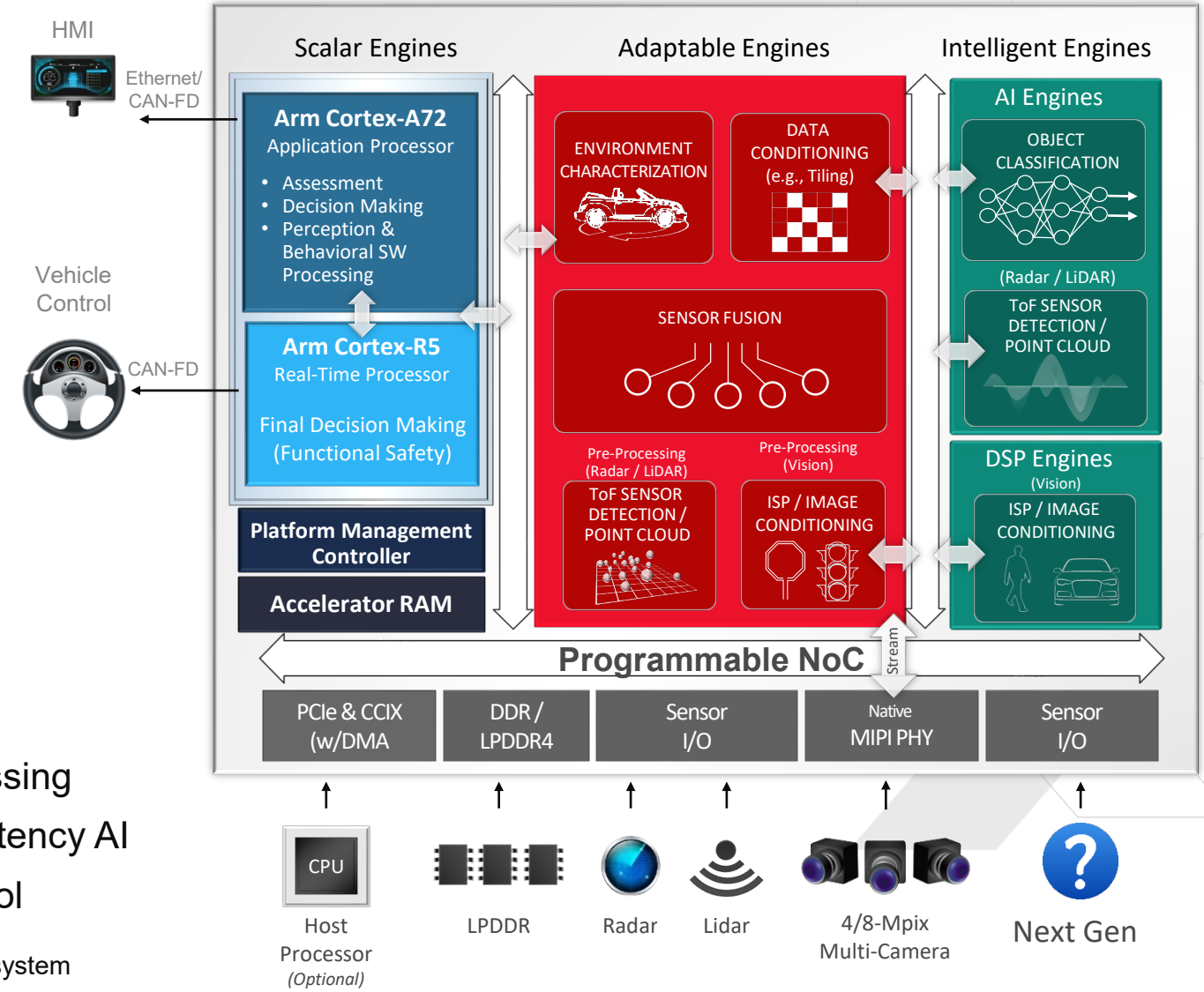
Accelerating the Whole Application: From Video Decode to Decision Making

XILINX

Versal AI Edge in ADAS and Automated Driving



XILINX VERSAL AI EDGE
VE2302



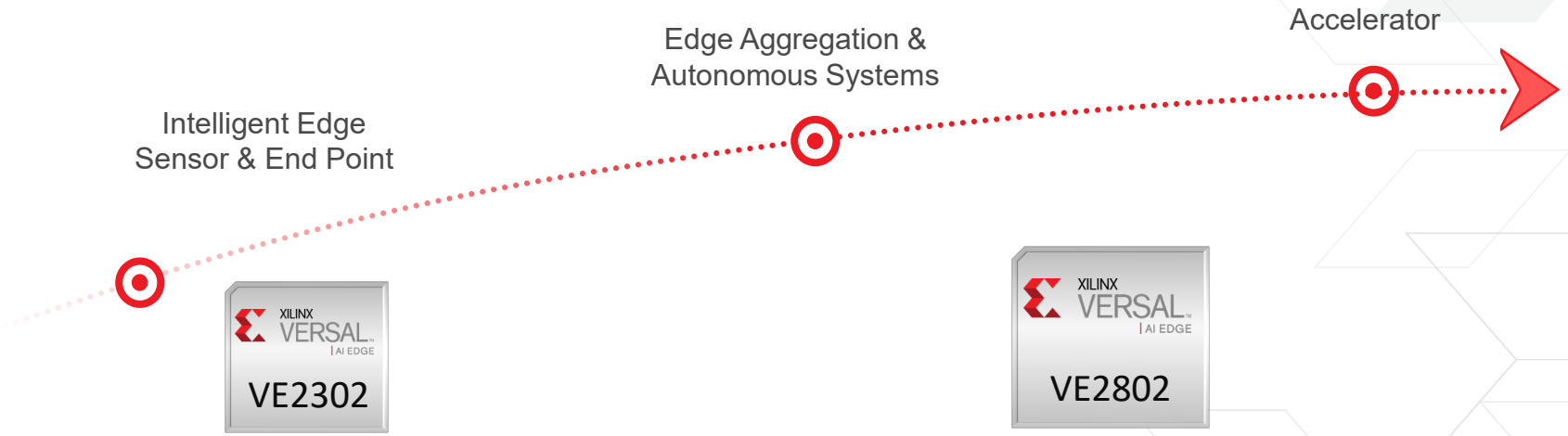
Accelerating the Whole Application From Sensor to AI to Real-Time Control

- Adaptable Engines for sensor fusion and pre-processing
- Intelligent Engines for signal conditioning and low latency AI
- Scalar engine for decision making and vehicle control

1: Demonstrates capabilities of architecture, not representing a single chip AD system

XILINX

Range of efficient Devices for AI/ML

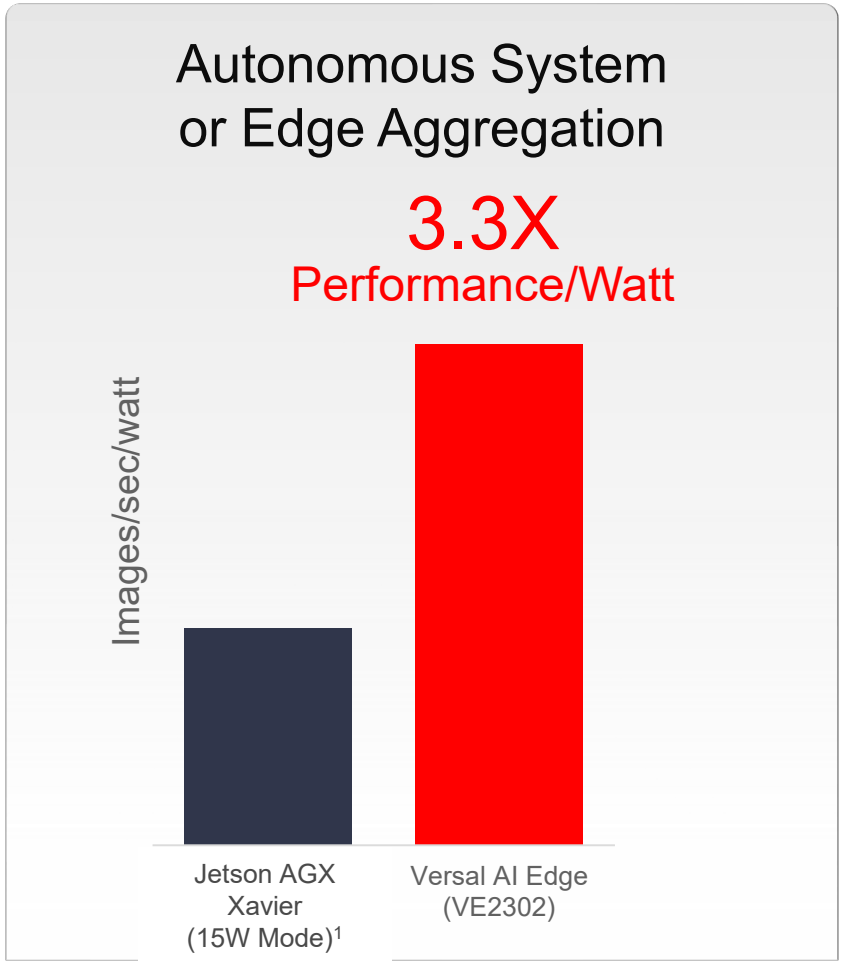


Engines	AI Compute (INT8x4)[1]	**67 TOPS**	**479 TOPS**
	AI Compute (INT8)[1]	**31 TOPS**	**228 TOPS**
	AIE-ML Tiles	34	304
	Adaptable Engines	150K LUTs	521K LUTs
	Processing Subsystem	Dual-Core Arm® Cortex®-A72 Application Processing Unit / Dual-Core Arm Cortex-R5F Real-Time Processing Unit	
RAM	Accelerator RAM (4MB)	✓	-
	Total Memory	172Mb	575Mb
	32G Transceivers	8	32
	PCIe®	✓	✓ (PCIe gen5 w/ DMA)
	Video Decode Unit (VDU)	-	✓
	Power [2]	**15-20W**	**75W**

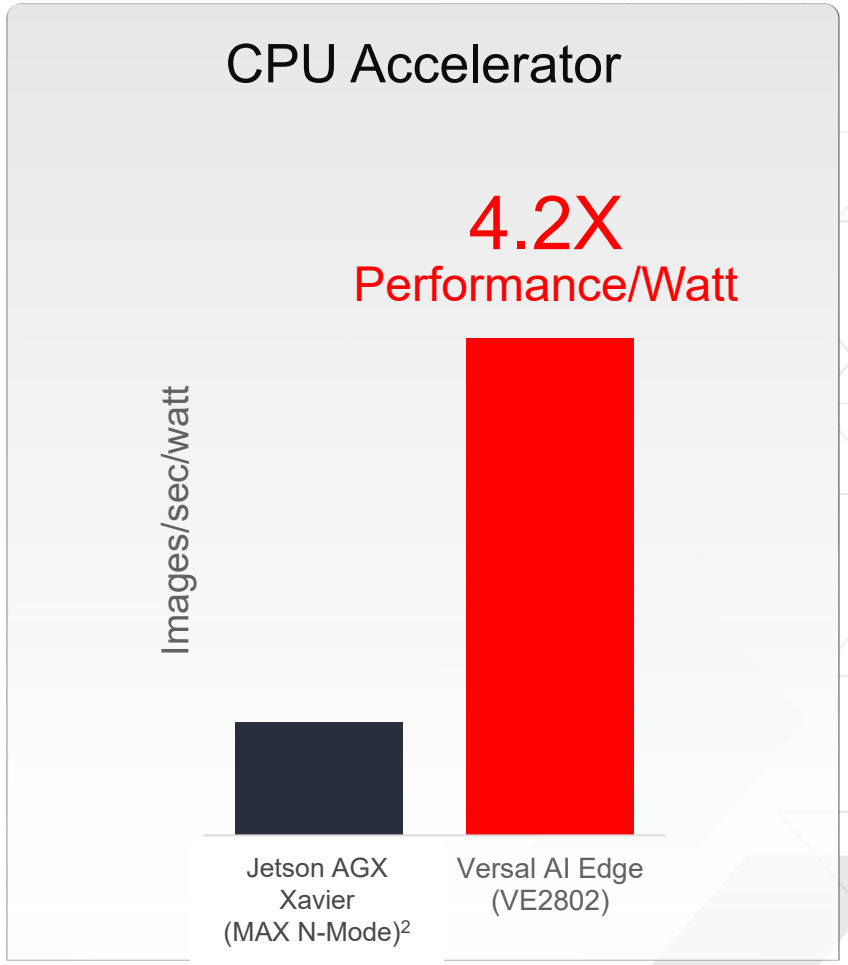
1: Total AI compute includes AI Engines, DSP Engines, and Adaptable Engines
2: Power Projections

XILINX

Performance/Watt Results



ResNet50 224x224, batch=1

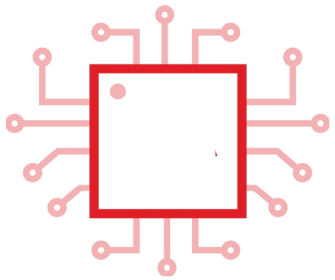


ResNet50 224x224, batch=1

1: Jetson AGX Xavier: https://developer.nvidia.com/embedded/jetson-agx-xavier-dl-inference-benchmarks,
2: Jetson AGX Xavier MAX N-Mode and VE2802 represent the highest performing device configuration in their respective portfolios
Jetson AGX Xavier device power estimated by subtracting published memory and IO power from total module power

XILINX.

Conclusion for 2 generations of AI Engines

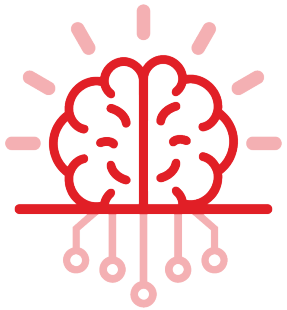


Xilinx Optimized Devices for Wireless and AI/ML Acceleration

- Built using Xilinx 7nm Versal™ Architecture
- First Generation: Versal AI Core series with AIE
- Second Generation: Versal AI Edge series with AI-ML

Whole Application Acceleration

- Wireless Beamforming and Digital Frontend, Video processing with ML
- Autonomous driving, Smart City for Edge Server Accelerator



Multi-generation continuous improvements

- Major improvements in TOPs/W over multiple generations
- ML-optimized features (e.g., optimized datatypes and memory hierarchy)