XILINX

Xilinx Edge Processors

AIE Engineering Team

HotChips-33 Conference, August 2021

Key Take-Aways for This Talk

Xilinx Optimized Devices for AI/ML Edge Inference Acceleration

Whole Application Acceleration

AIE-ML: Optimized AIE Architecture for Machine Learning

AI/ML Edge Inference: Use-cases and Requirements

Automotive

ADAS

Automated Driving

Industrial Robotics

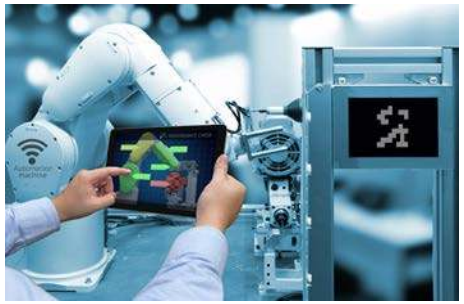
Vision guidance

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

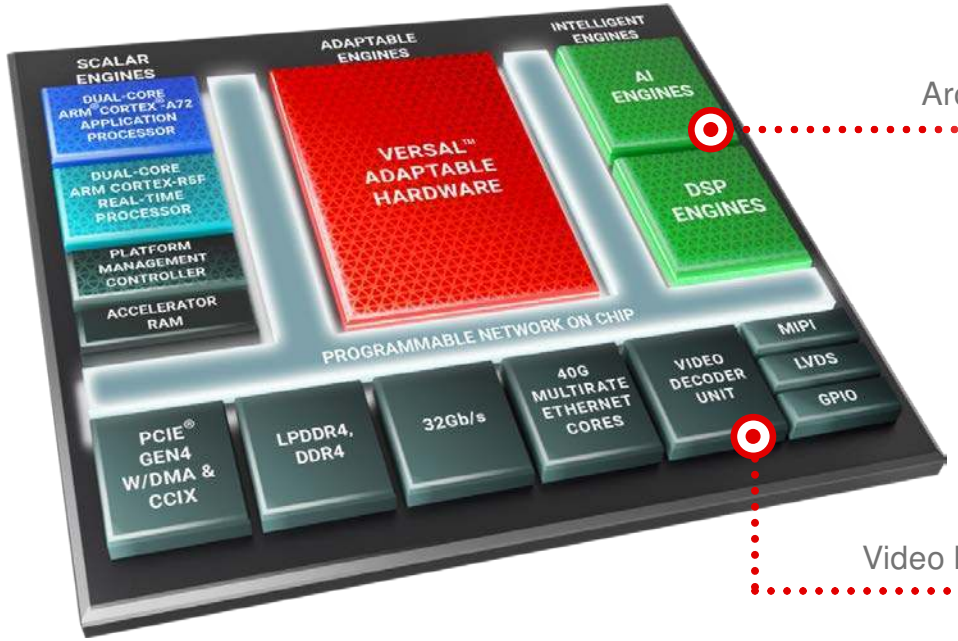
- Versal AI Core
- Versal Premium
- Versal AI Edge

COMPUTE ACCELERATION

- Scalar Engines
- Adaptable Engines
- Intelligent Engines

PLATFORM

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



AIE-ML Architecture

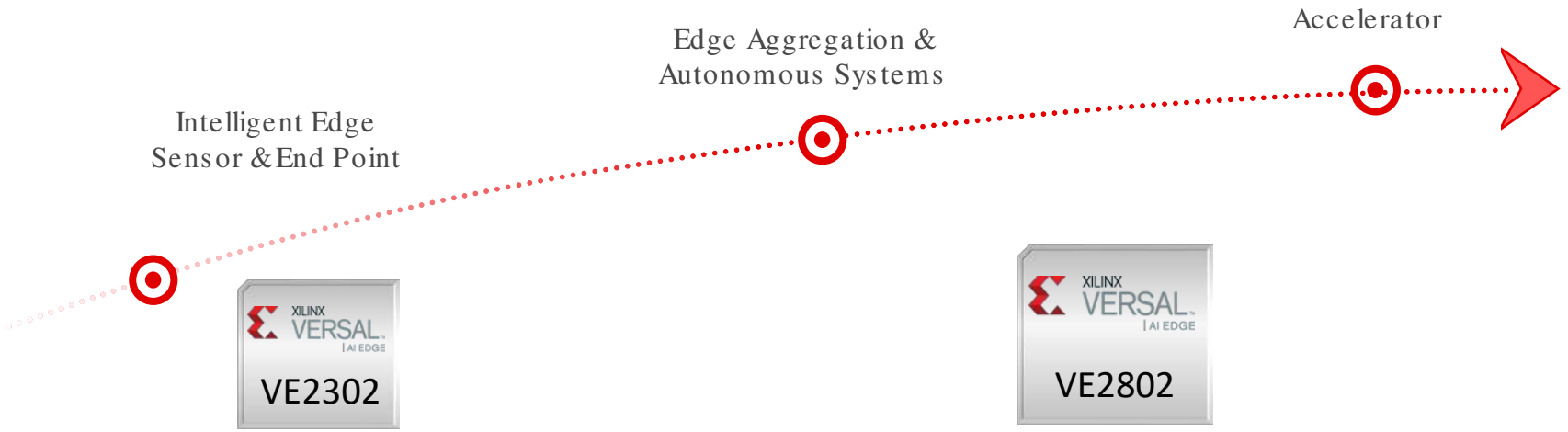
Video Decode Unit

Technology Node: TSMC 7nm FinFET

tsmc 7nm

XILINX

Optimized Devices for AI/ML Edge Inference

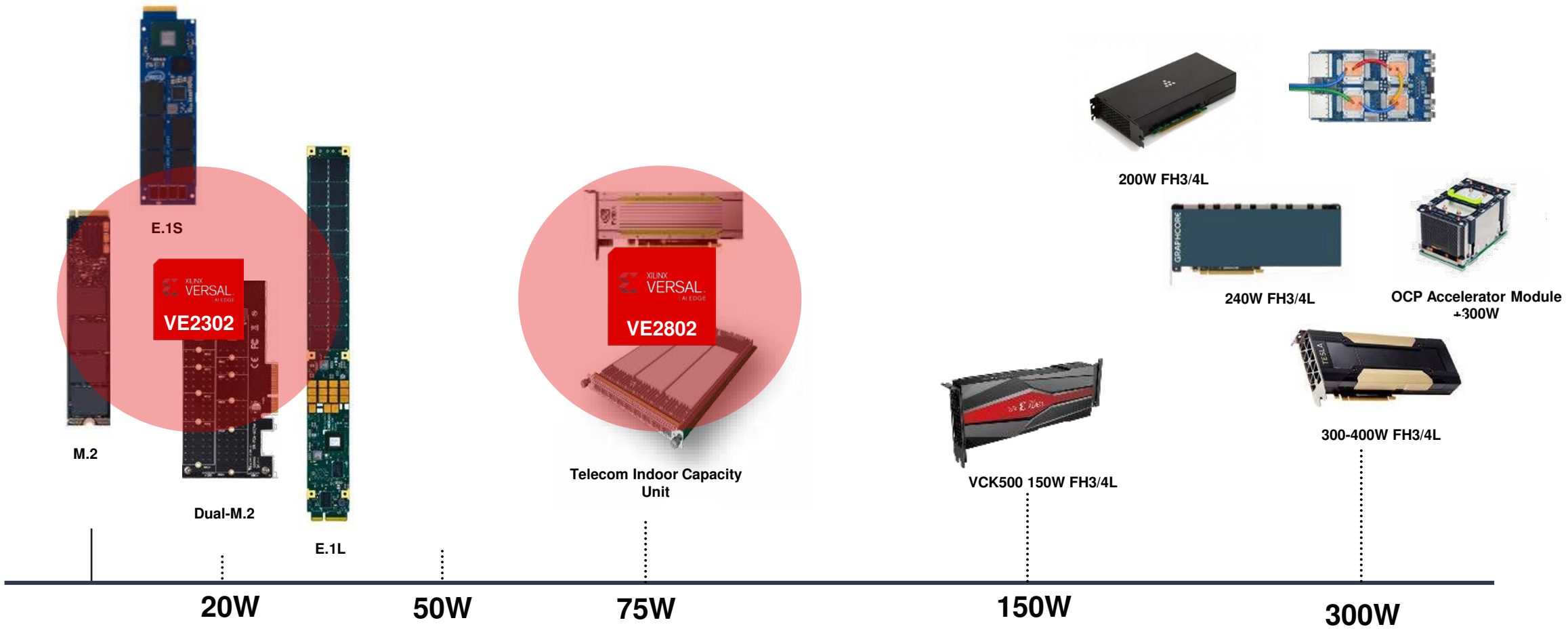


Engines	AI Compute (INT8x4)[1]	67 TOPS	479 TOPS
	AI Compute (INT8)[1]	31 TOPS	228 TOPS
	AIE-ML Tiles	34	304
	Adaptable Engines	150K LUTs	521K LUTs
	Processing Subsystem	Dual-Core Arm® Cortex®-A72 Application Processing Unit / Dual-Core Arm Cortex-R5F Real-Time Processing Unit	
RAM	Accelerator RAM (4MB)	✓	-
	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

Inference Form Factor



Small Module Form Factor
High Degrees of Scalability, Thermal Challenges

HHHL PCIe Form Factor
Lowest Common Denominator
Deployments in Core, AI Edge Cloud

200W+ FH3/4L Form Factor
Some restrictions due power and form factor

Optimized AIE-ML for Machine Learning

Optimized the compute core for ML

- Doubled the multipliers, doubled INT8 performance
- Native support for INT4 and BFLOAT16

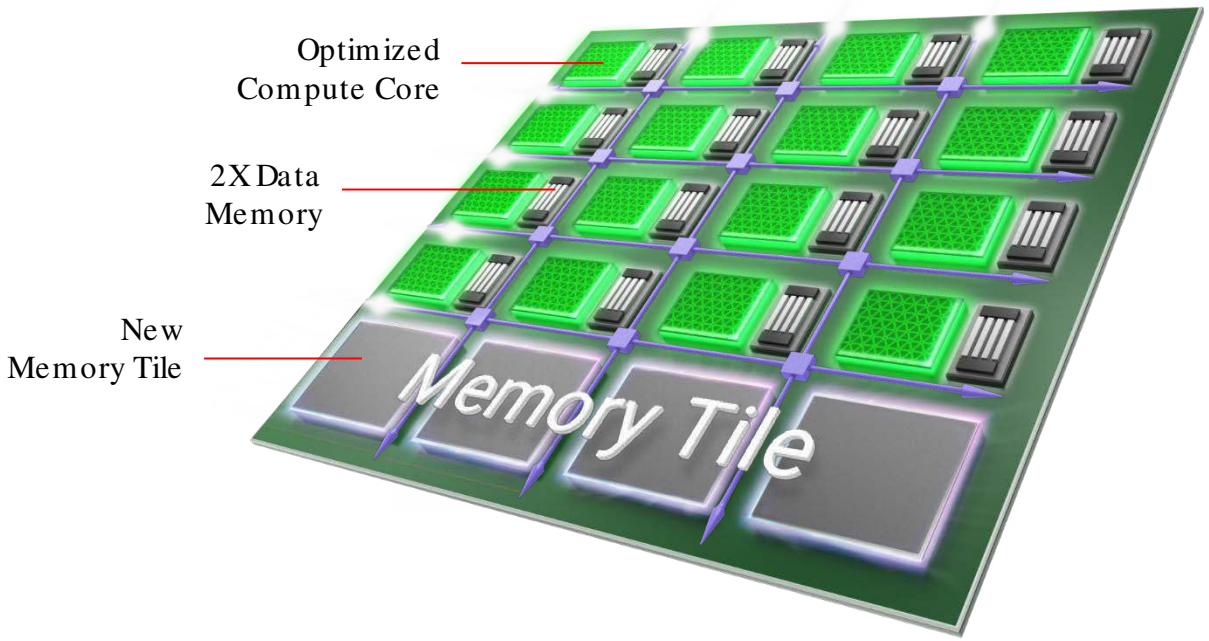
Doubled the data memory

- From 32kB to 64 kB
- Improved localization of data

New Memory Tile

- Up to 38 Megabytes across the AIE array
- Higher bandwidth memory access

Optimized
AIE-ML Array
(Part of Versal AI Edge devices)

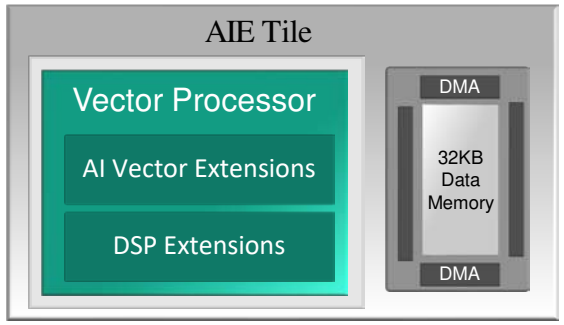


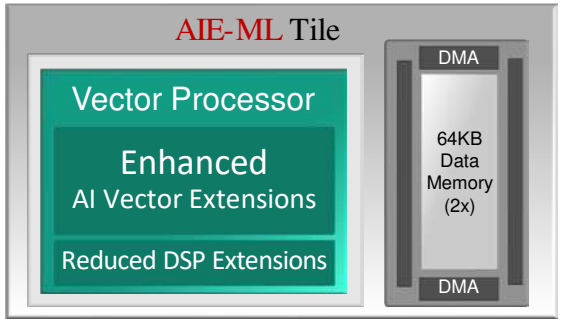
Scalable and Modular Architecture

1: AI Engine-ML delivers 2X INT8 compute, 4X INT4 compute, and 16X BFLOAT16 compute vs. AI Engine (per core)
2: Native 32-bit support in AI Engines only

XILINX.

Xilinx AIE Tile: Architecture Features



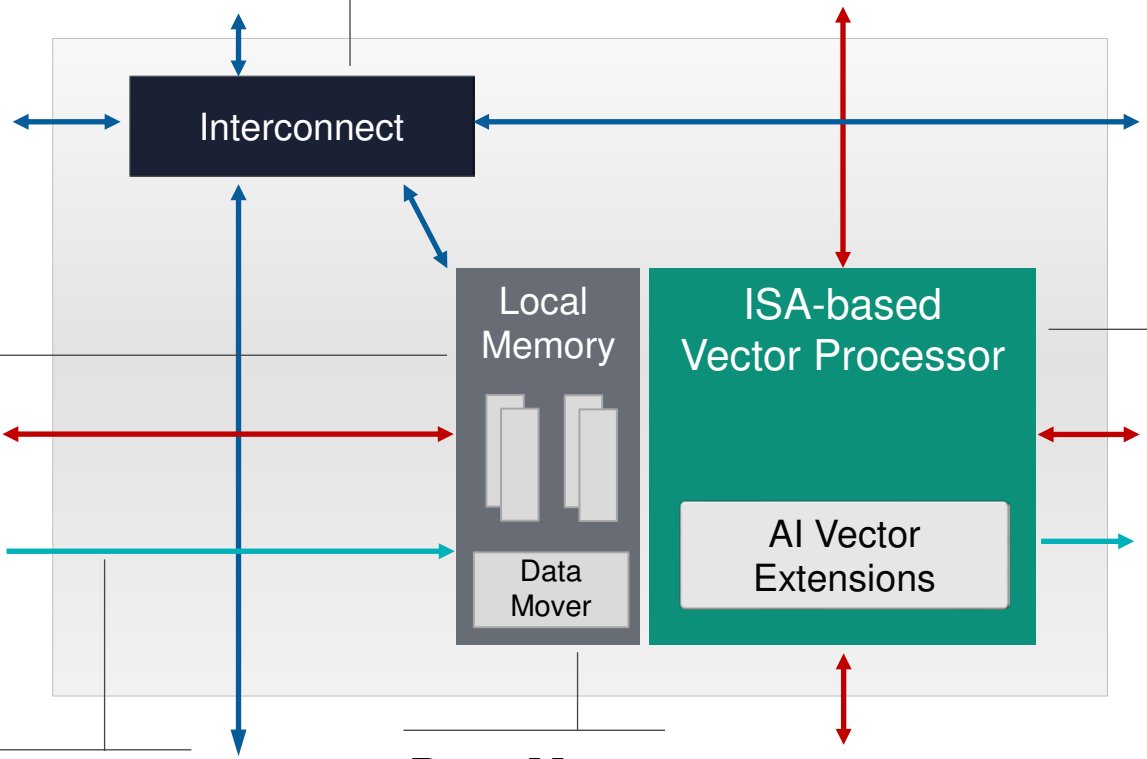


		AIE Tile		AIE-ML Tile
Target Markets		Wireless 5G, AI/ML inference, A&D		AI/ML inference
Compute (Mults / tile)	BFLOAT16	—	New	128
	INT8	128	2X	256
	INT4	—	New	512
Tile Local Data Memory		32 KB	2X	64 KB
AIE Array Interconnect B/W		1X	1X	1X
Compression and Sparsity		No		Yes
Scratchpad On-Chip Memory	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

XILINX.

AIE-ML Tile Architecture

Non-Blocking Interconnect
Up to 200+ GB/s bandwidth per tile



Local Memory
Multi-bank implementation
Shared across neighbor cores

ISA-based Vector Processor
Software Programmable
(e.g., C/C++)

Cascade Interface
Partial results to next core

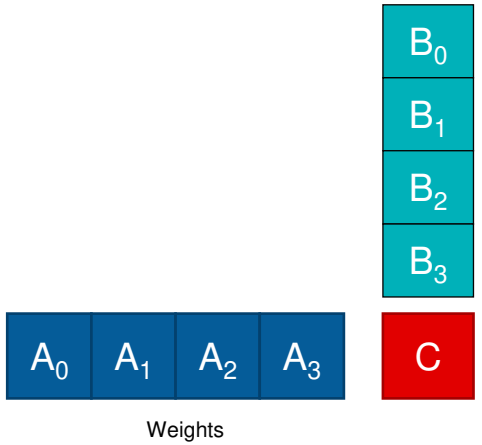
Data Mover
Non-neighbor data communication
Integrated synchronization primitives

Flexible Dataflow: AIE-ML Array Examples

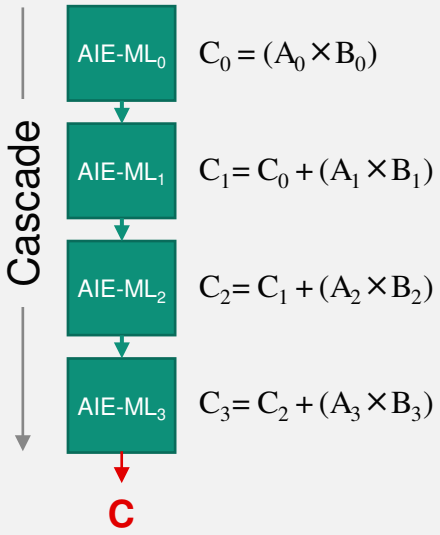
DATA REDUCTION

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

COMPUTE **SINGLE** OUTPUT MATRIX

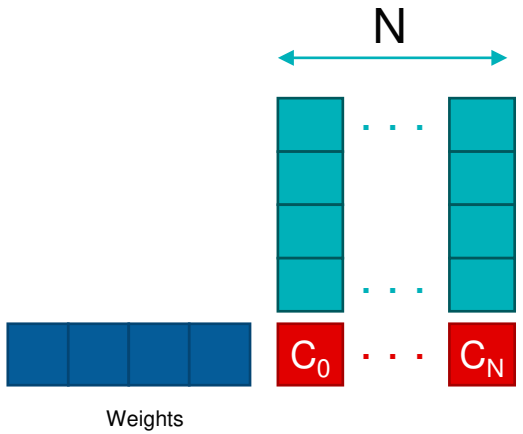


AIE-ML IMPLEMENTATION

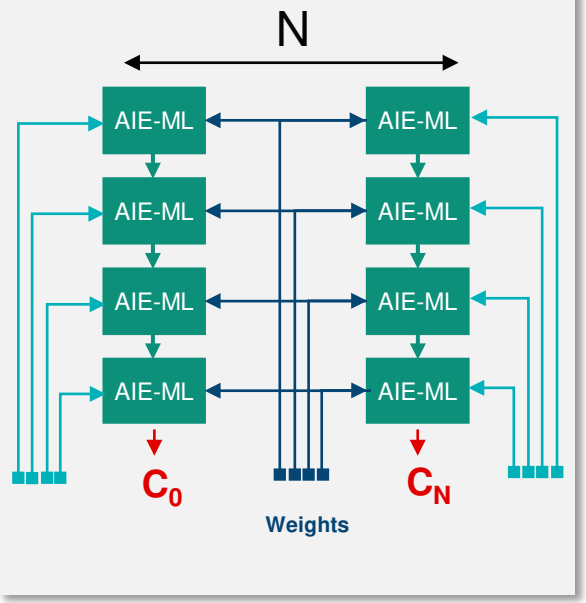


DATA MULTICASTING

COMPUTE **MULTIPLE** OUTPUT MATRICES

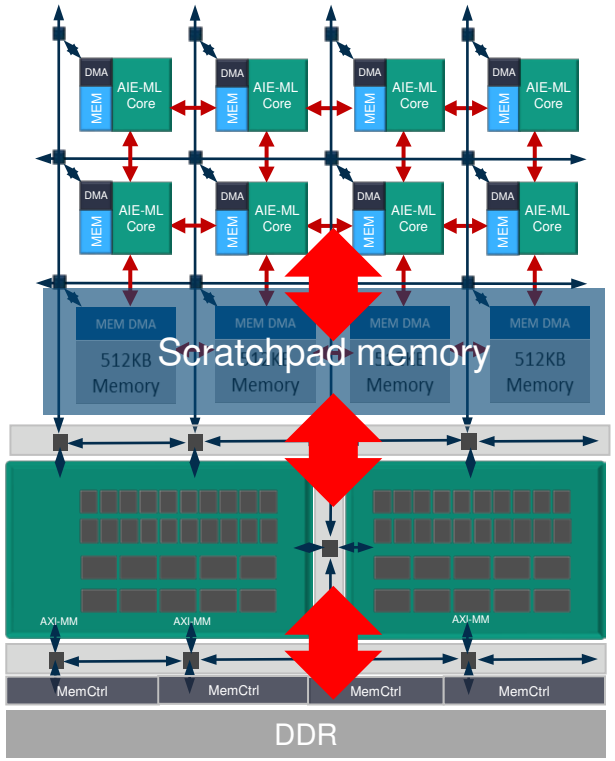


AIE-ML IMPLEMENTATION

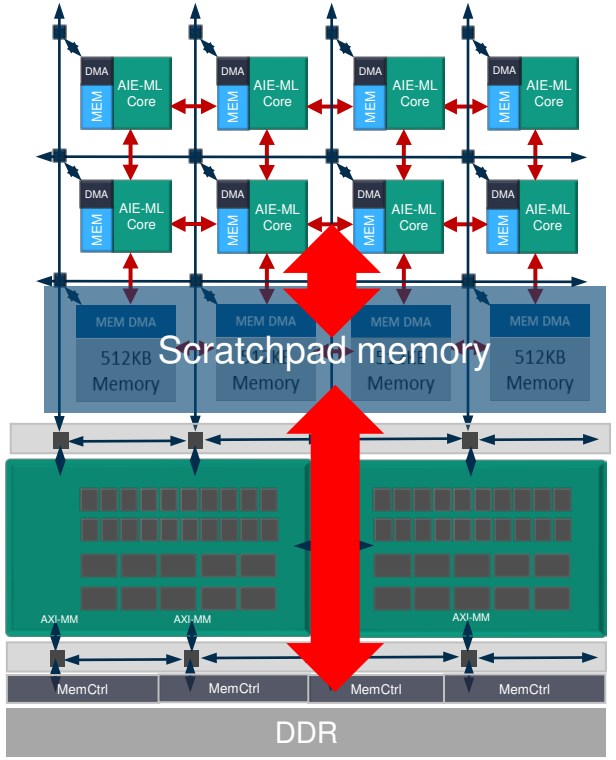


XILINX.

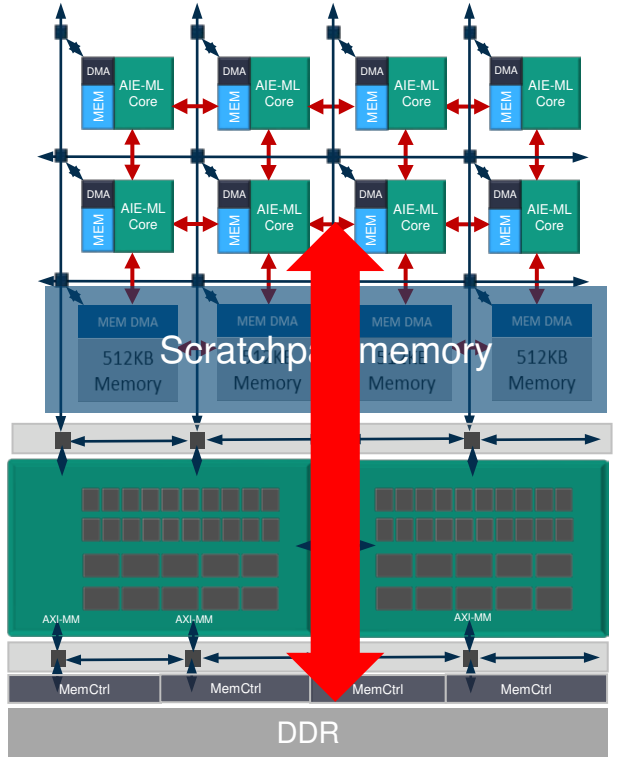
Flexible Dataflow: Device-level Examples



DDR ↔ Scratchpad using Programmable Logic



DDR ↔ Scratchpad using NoC

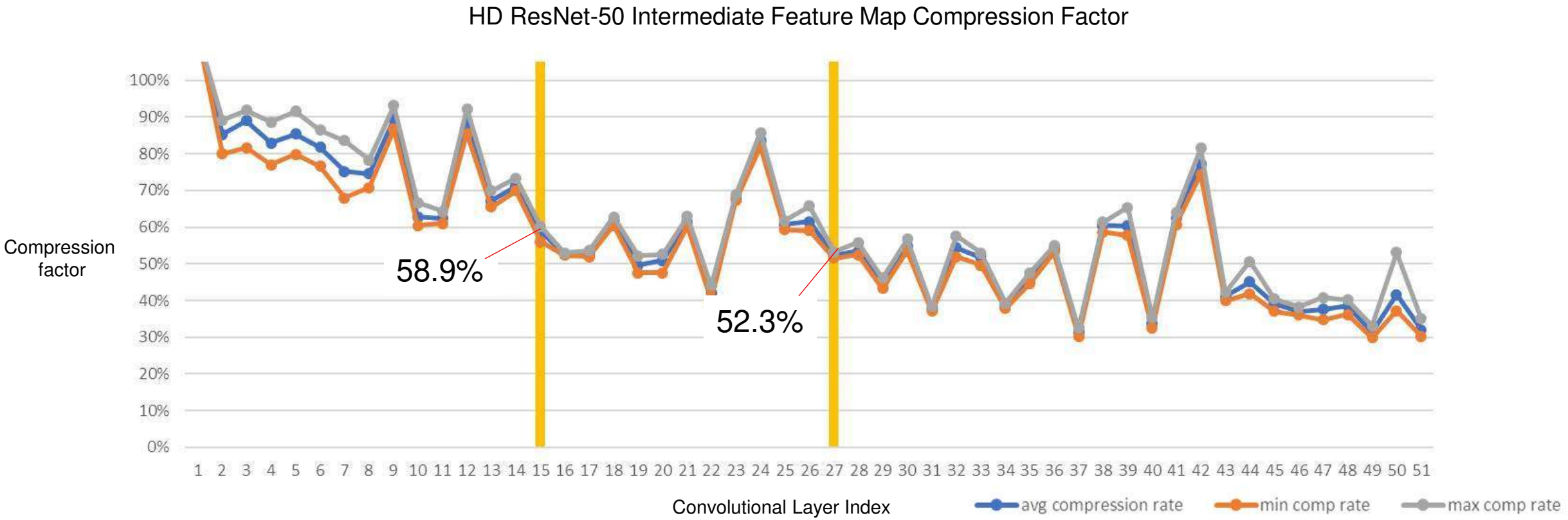


DDR ↔ AIE-ML Tiles using NoC

XILINX.

Compression to DDR: Acceleration of activation Spill/Restore

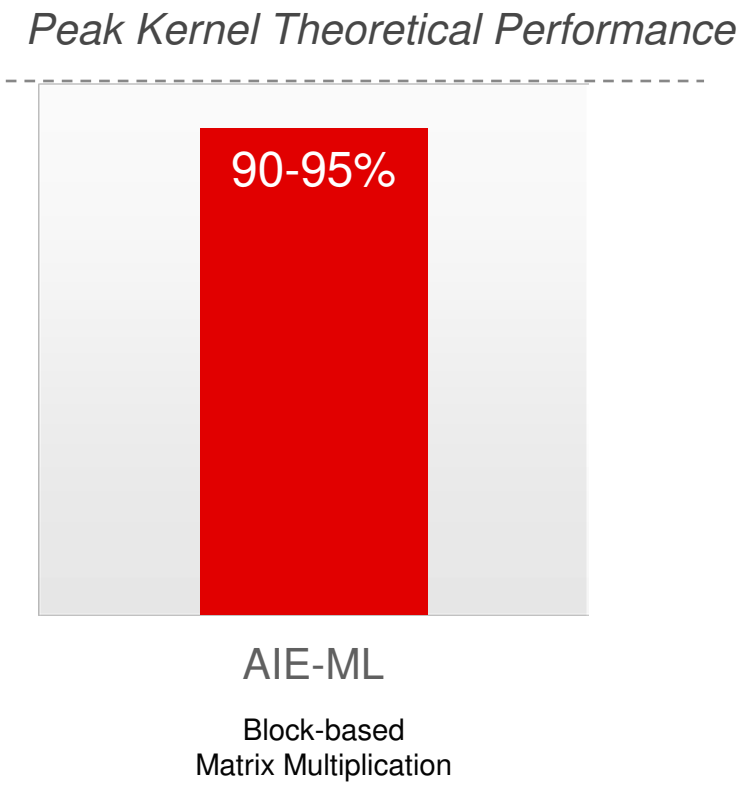
- Reduction in external memory bandwidth using compression
  - Quantized int8 weights, int8 activations
  - HD images require tiling, with spill/restore to DDR
    - Without compression ~**56GB/s**
    - With compression ~**29GB/s**



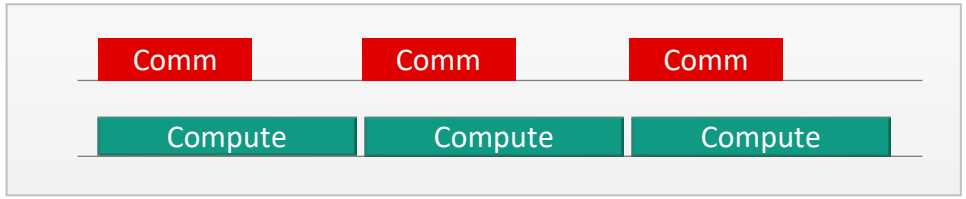
XILINX.

AIE-ML Delivers High Compute Efficiency

Vector Processor 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 AIE-ML computes in parallel

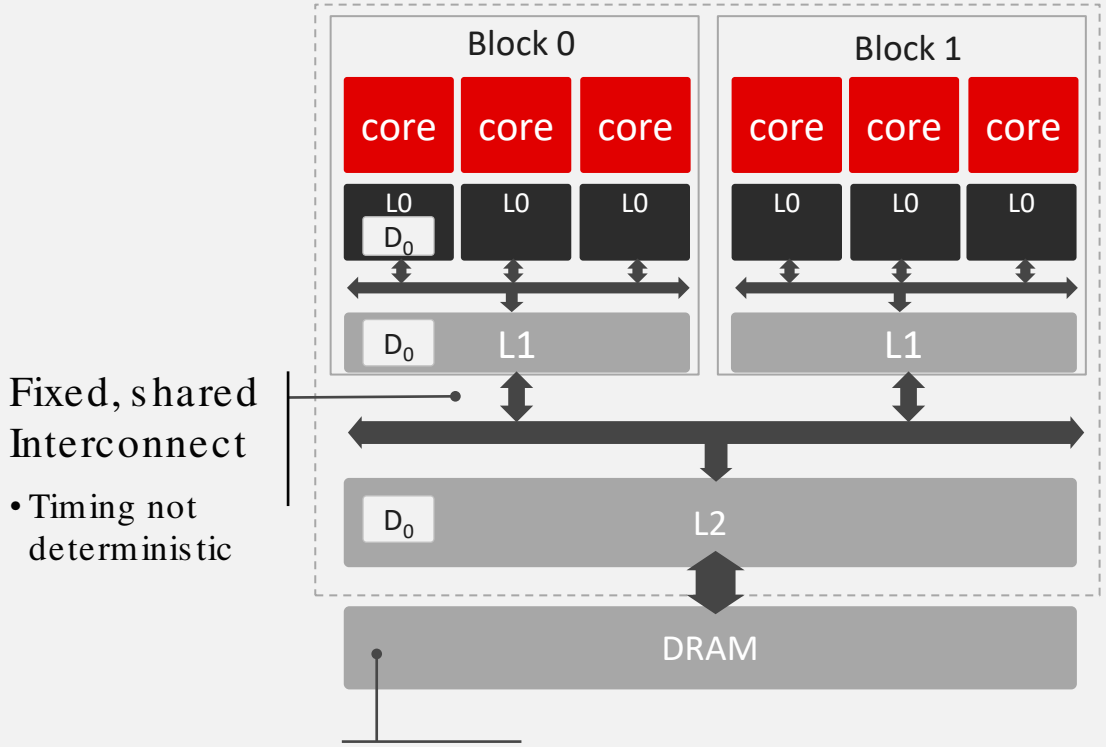


Overlap Compute and Communication

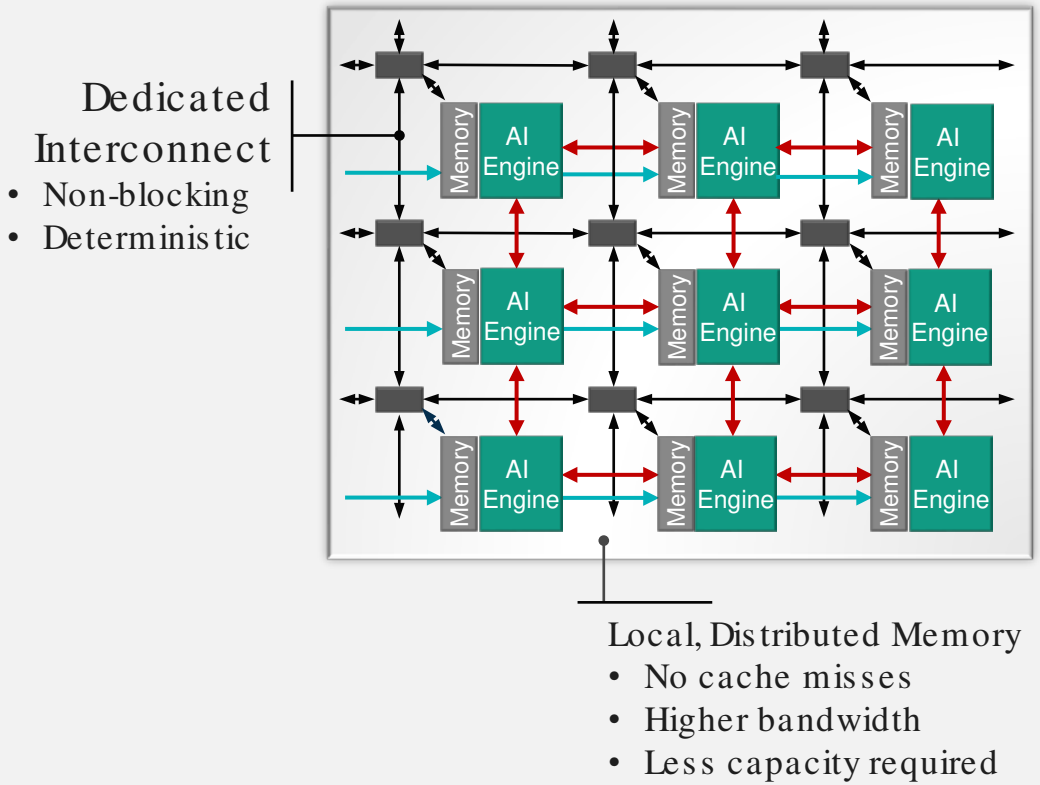
XILINX.

AIE-ML: Multi-Core Architecture Comparison

Traditional Multi-core
(cache-based architecture)

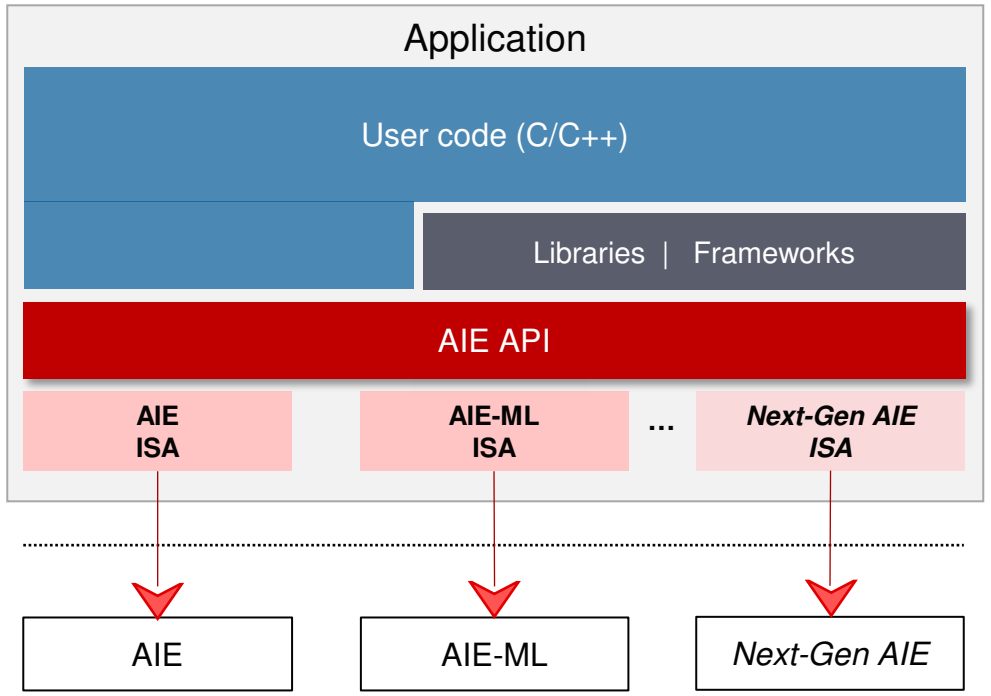


AIE-ML Array
(intelligent engine)



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

Vitis AI: Full AI Software Stack Support

Frameworks

PyTorch TensorFlow Caffe tvm

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

Platforms

XILINX VERSAL™ AI CORE

XILINX VERSAL™ AI EDGE

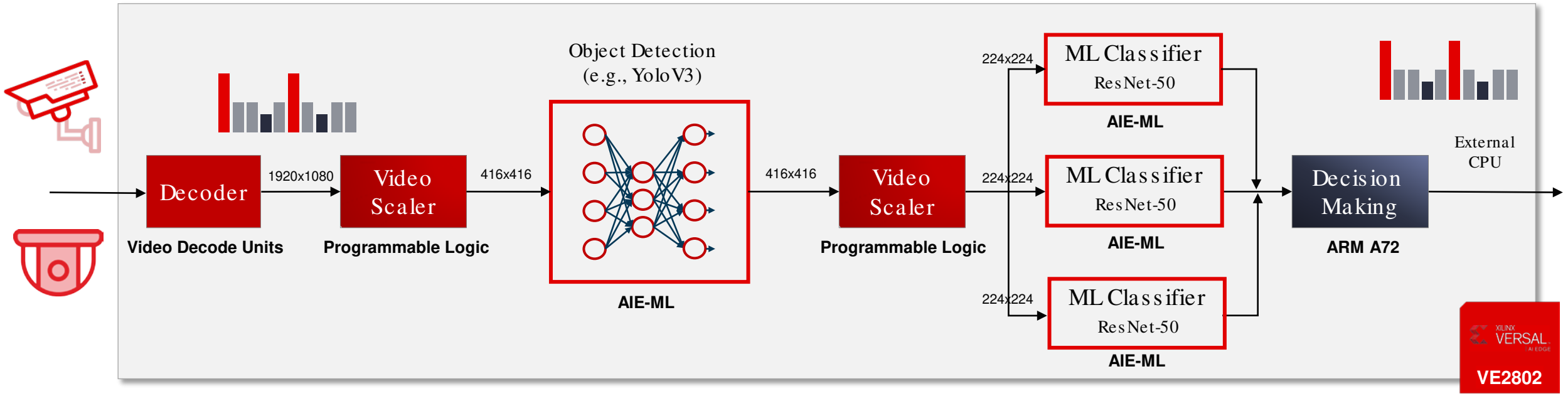
XILINX.

Smart World Application: Use-Case Mapping

Smart Retail

Smart Hospital

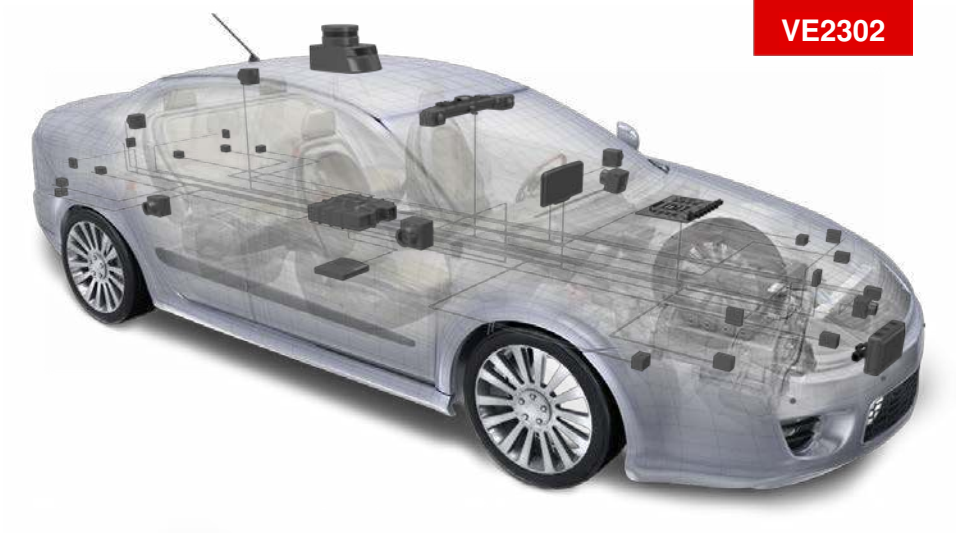
Smart City

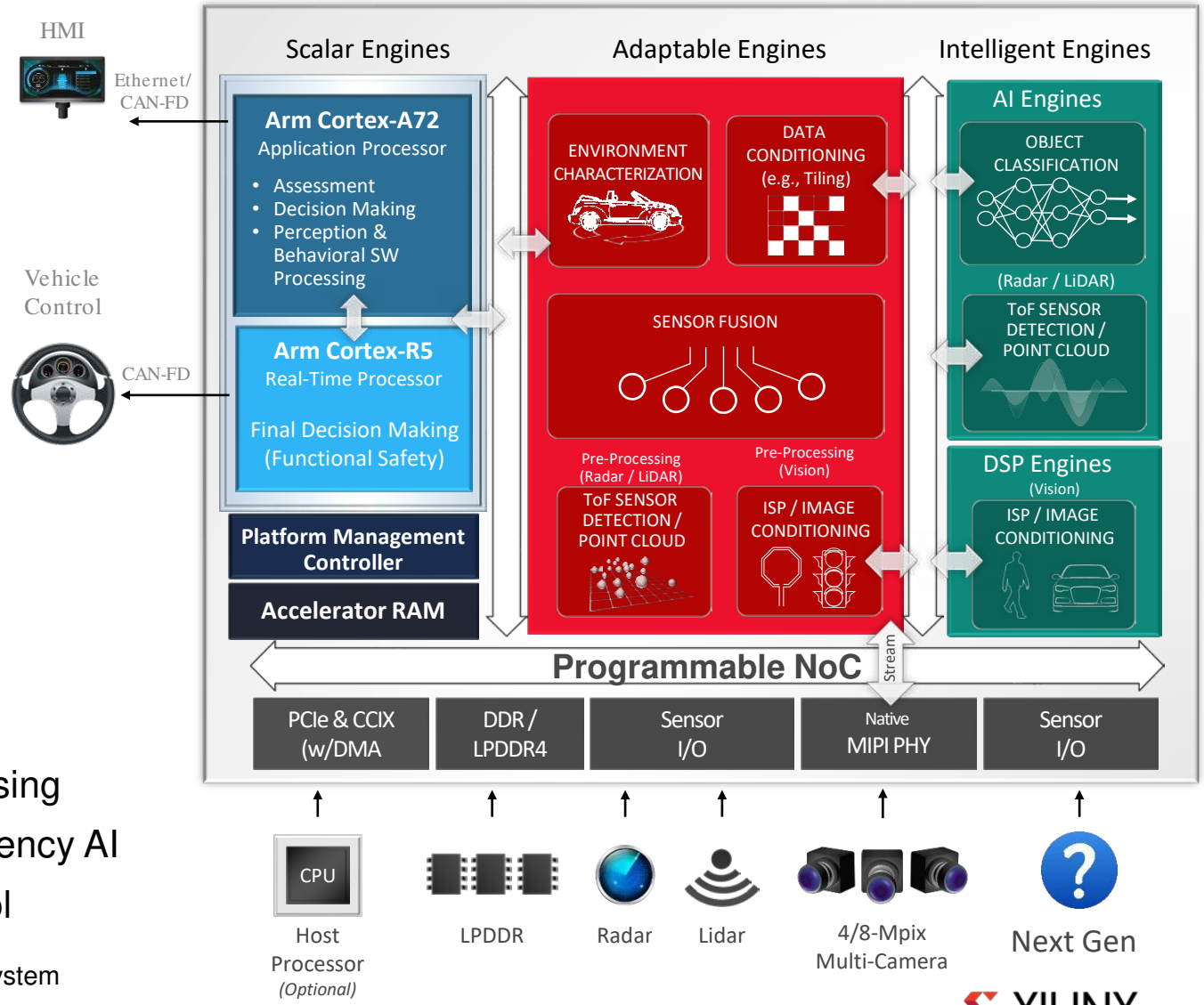


Accelerating the Whole Application: From Video Decode to Decision Making

XILINX

Versal AI Edge in ADAS and Automated Driving





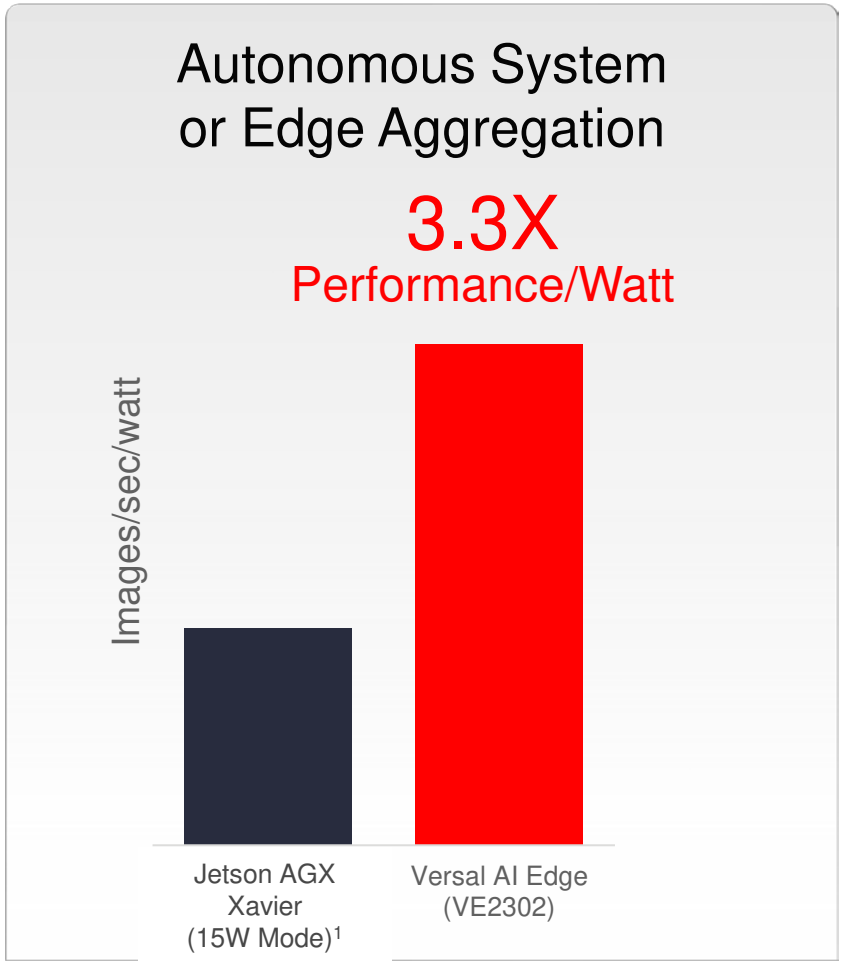
Accelerating the Whole Application From Sensor to AI to Real-Time Control

- 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

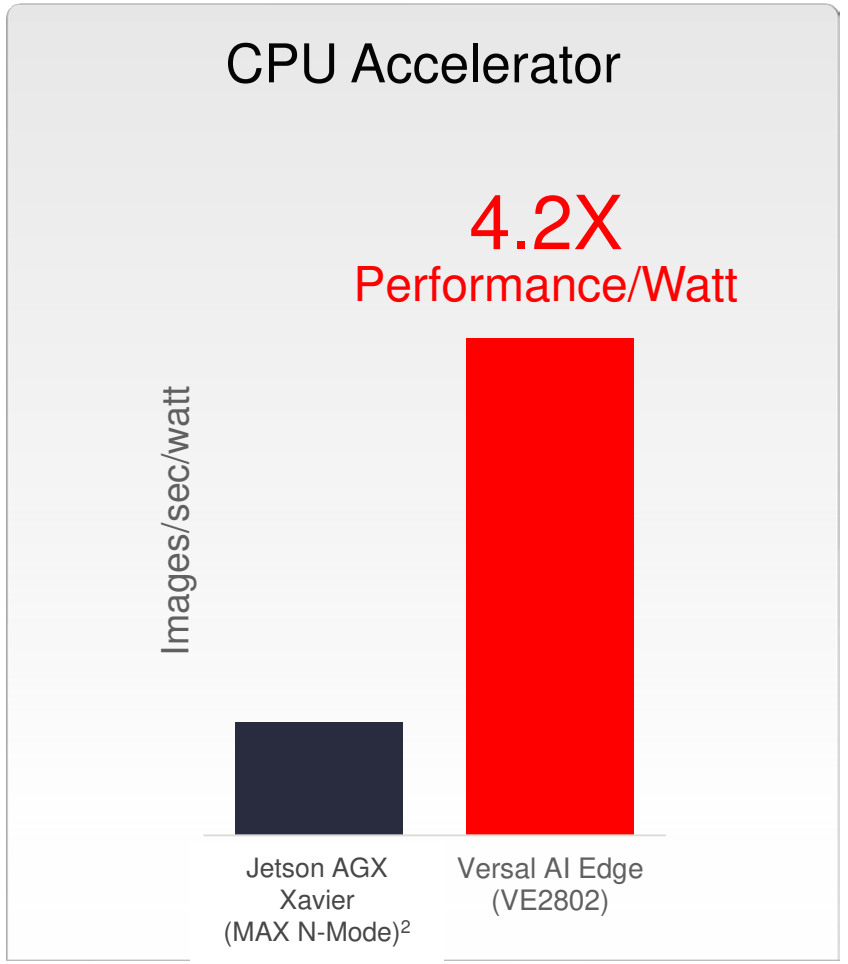
1: Demonstrates capabilities of architecture, not representing a single chip AD system

XILINX

Performance/Watt Results



ResNet50 224x224, batch=1

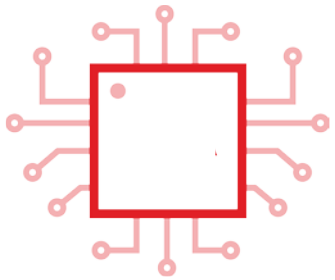


ResNet50 224x224, batch=1

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.

Summary

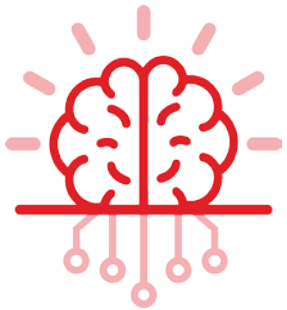


Xilinx Optimized Devices for AI/ML Edge Inference Acceleration

- Built using Xilinx 7nm Versal™ Architecture
- Versal AI Edge VE2302: 15-20W, small form factor
- Versal AI Edge VE2802: 75W, HHHL PCIe form factor

Whole Application Acceleration

- Edge Aggregation & Autonomous Driving
- Smart City for Edge Server Accelerator



AIE-ML: Optimized AIE Architecture for Machine Learning

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

Thank You