XILINX®

Versal™ AI Edge Series Announcement

Rehan Tahir, Senior Product Line Manager

What’s Happening at the Edge

The Edge

Low Latency

AI Compute

Low Power

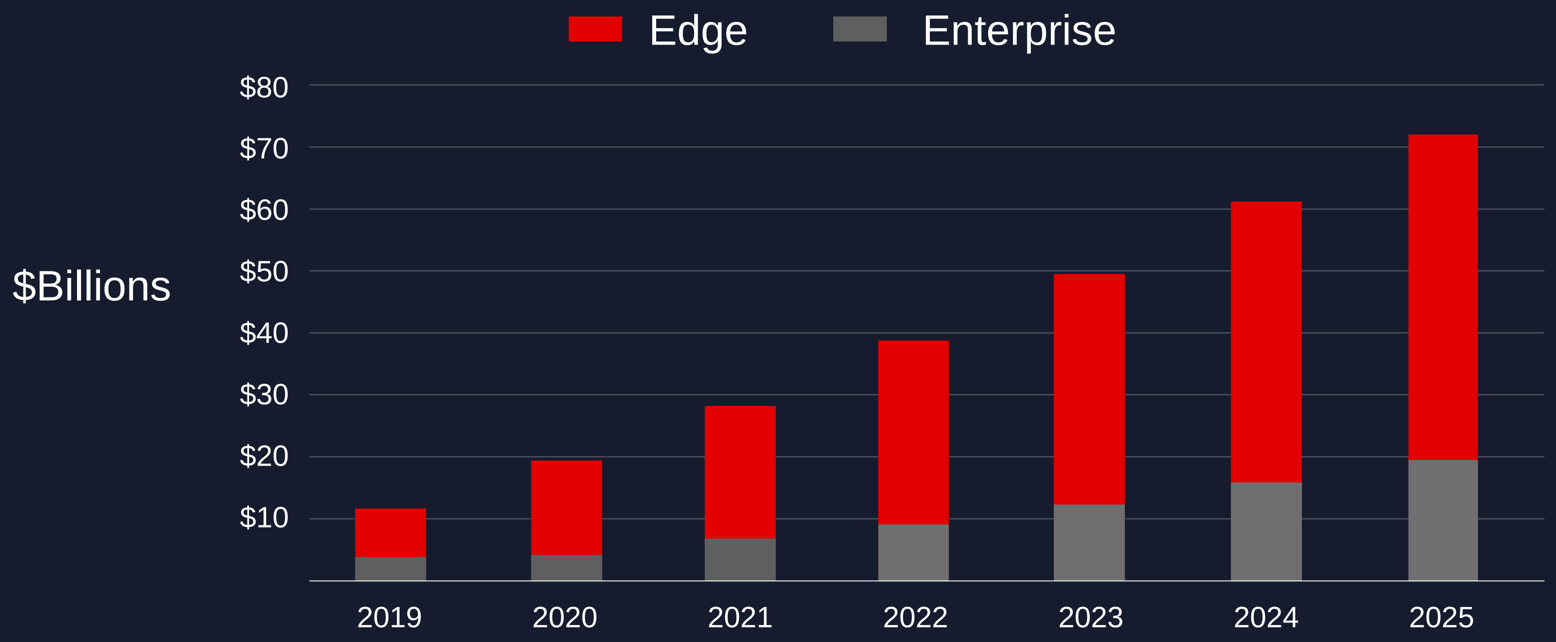
Safety and Security

Hypergrowth at the Edge

“Edge computing … solves for weaknesses of the cloud”[1]

Edge AI chipset opportunity is 3X that of data center - $65B in 2025[2]

Deep learning chipset revenue, enterprise vs. edge, world markets: 2019–25



[1] Gartner, “2021 Strategic Roadmap for Edge Computing”, November 2020
[2] Omdia, “Market Report: Deep Learning Chipsets”, July 2020

Now Bringing Versal ACAPs to the Edge

- Versal™ ACAPs first introduced breakthrough compute for the cloud and network
- Now 'miniaturizing' this technology for performance/watt at the edge

XILINX VERSAL™ | AI CORE

XILINX VERSAL™ | PREMIUM

XILINX VERSAL™ | PRIME

XILINX VERSAL™ | AI EDGE

CLOUD

NETWORK

EDGE

New Versal™ Platform for Intelligence at the Edge

XILINX VERSAL™ | AI EDGE

Smart Vision

Unmanned Aerial Vehicles

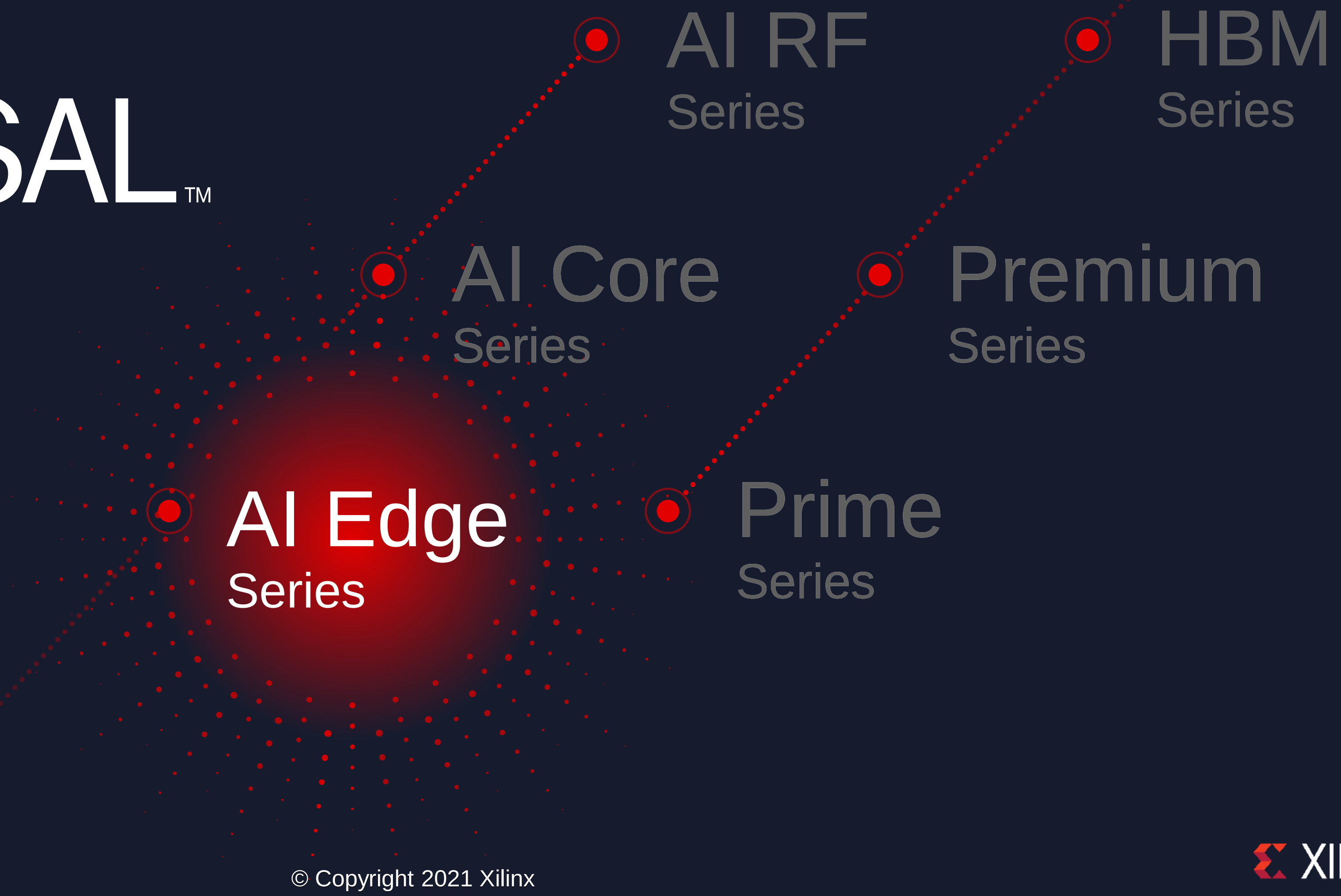
Collaborative Robotics

ADAS & Automated Drive

Endoscopy

Ultrasound

XILINX® VERSAL™



AI RF Series

HBM Series

AI Core Series

Premium Series

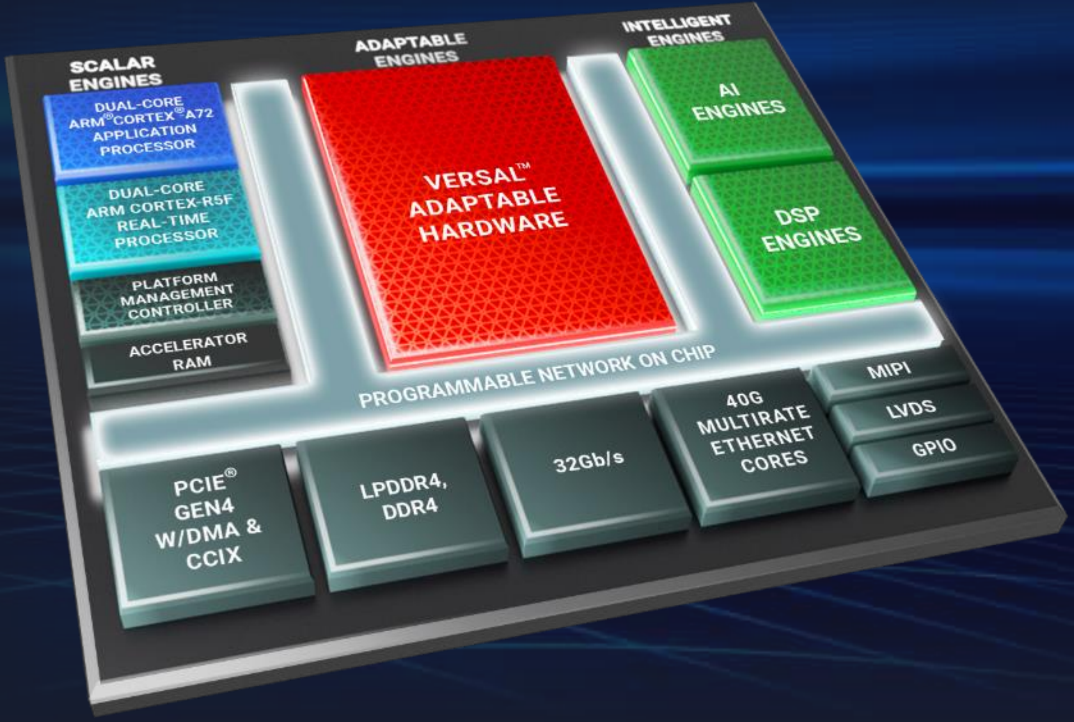
AI Edge Series

Prime Series

Versal™ AI Edge: Intelligence Unleashed

XILINX VERSAL™ | AI EDGE

- 4X AI Performance/Watt vs. GPUs[1] with Innovations in AI Engines and Memory Hierarchy
- 10X Compute Density[2] with Highest Levels of Safety and Security
- World’s Most Scalable and Adaptable Platform for Edge and Endpoint



SCALAR ENGINES
DUAL-CORE ARM® CORTEX®-A72 APPLICATION PROCESSOR
DUAL-CORE ARM CORTEX-R5F REAL-TIME PROCESSOR
PLATFORM MANAGEMENT CONTROLLER
ACCELERATOR RAM
ADAPTABLE ENGINES
VERSAL™ ADAPTABLE HARDWARE
INTELLIGENT ENGINES
AI ENGINES
DSP ENGINES
PROGRAMMABLE NETWORK ON CHIP
PCIE® GEN4 W/DMA & CCIX
LPDDR4, DDR4
32Gb/s
40G MULTIRATE ETHERNET CORES
MIPI
LVDS
GPIO

[1] vs. Jetson AGX Xavier, ResNet50 224x224, batch=1, https://developer.nvidia.com/embedded/jetson-agx-xavier-dl-inference-benchmarks

[2] Compared to Zynq® UltraScale+™ MPSoCs

4X AI Performance/Watt

Proven AI Engine Architecture

Array of Compute Core

- Flexible compute: fixed- & floating-point vector processors
- HW adaptable to evolving algorithms

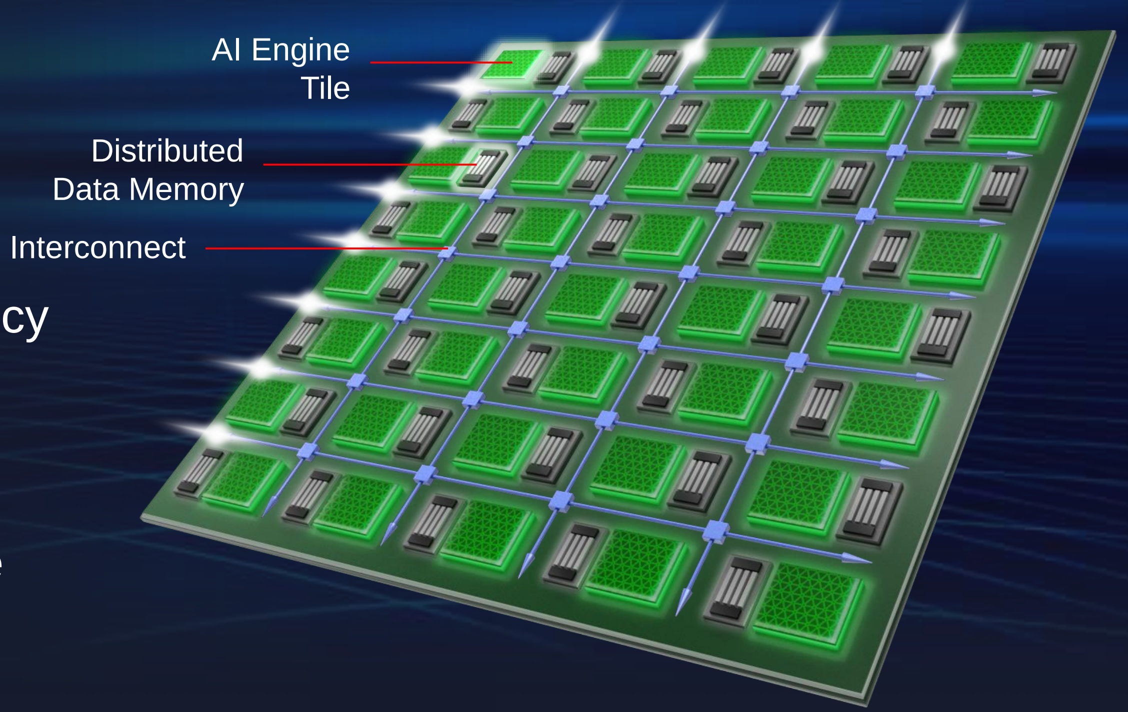
Tightly Coupled Memory

- Cache-less memory hierarchy
- Maximizes bandwidth, ensures determinism & low latency

Flexible Interconnect

- Connect any tile to any tile for custom microarchitecture
- High bandwidth

AI Engines Array
(Part of ACAP Device)



Architected for Adaptability, Low Power, and Low Latency

Optimized AI Engines-ML for Machine Learning

Optimized the compute core for ML

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

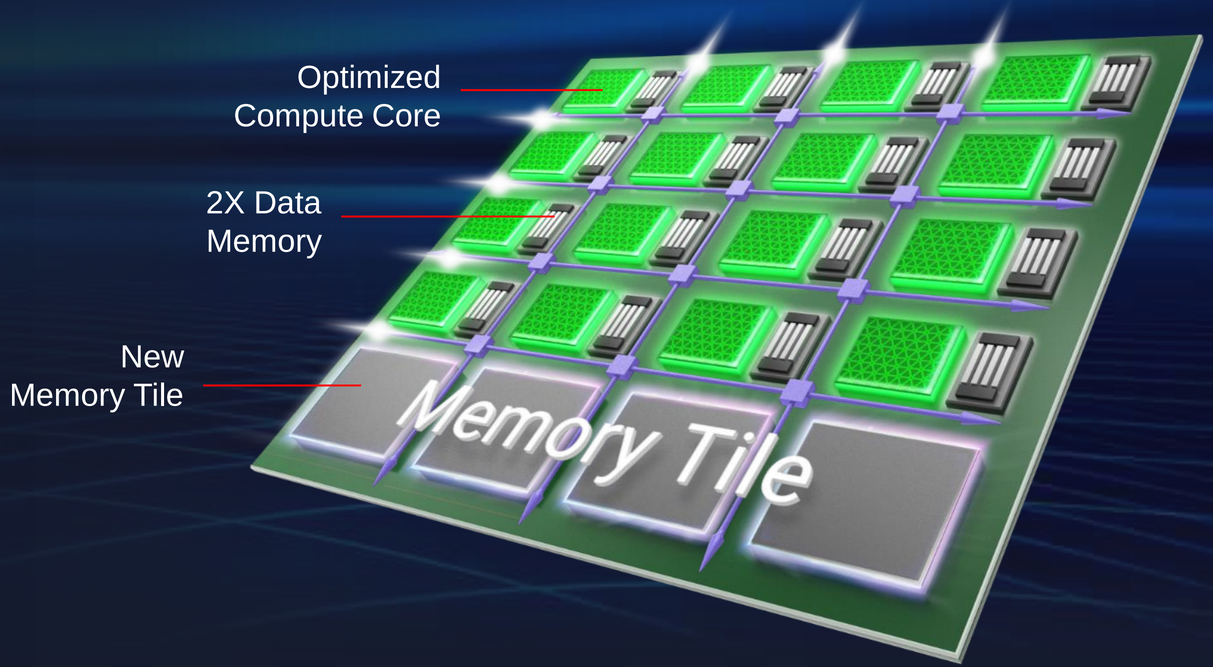
Doubled the data memory

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

New Memory Tile

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

Optimized
AI Engine-ML Array
(Part of ACAP Device)



Delivering 4X ML Compute at ½ the Latency[1]

[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

AIE-ML Complements AI Engines for Diverse Workloads

AI Engine
Architecture

Balanced for ML & DSP

▶

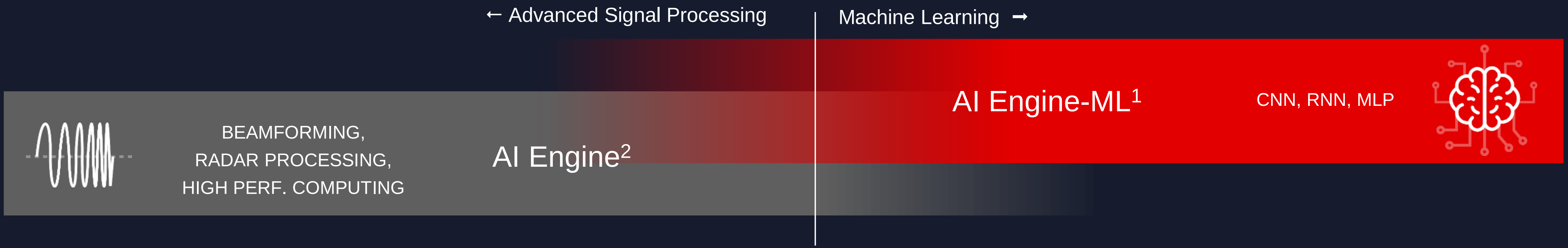
4X ML Compute

at ½ the Latency

▶

AI Engine-ML
Architecture

Optimized for ML



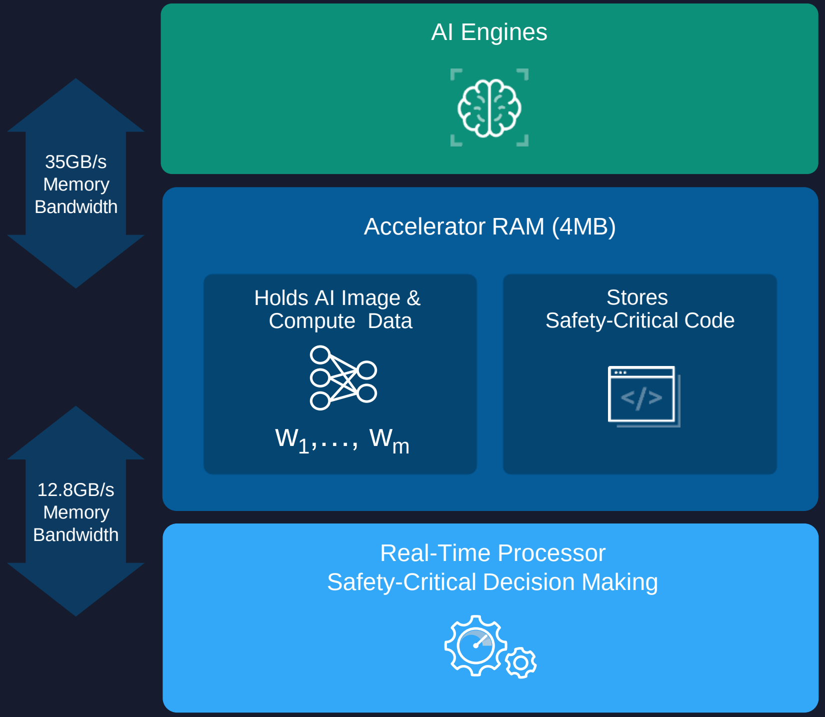
[1] AI Engine-ML delivers 2X INT8 compute, 4X INT4 compute, and 16X BFLOAT16 compute vs. AI Engine (per core)

[2] Native 32-bit support in AI Engines only

Innovations in Memory Hierarchy: Accelerator RAM

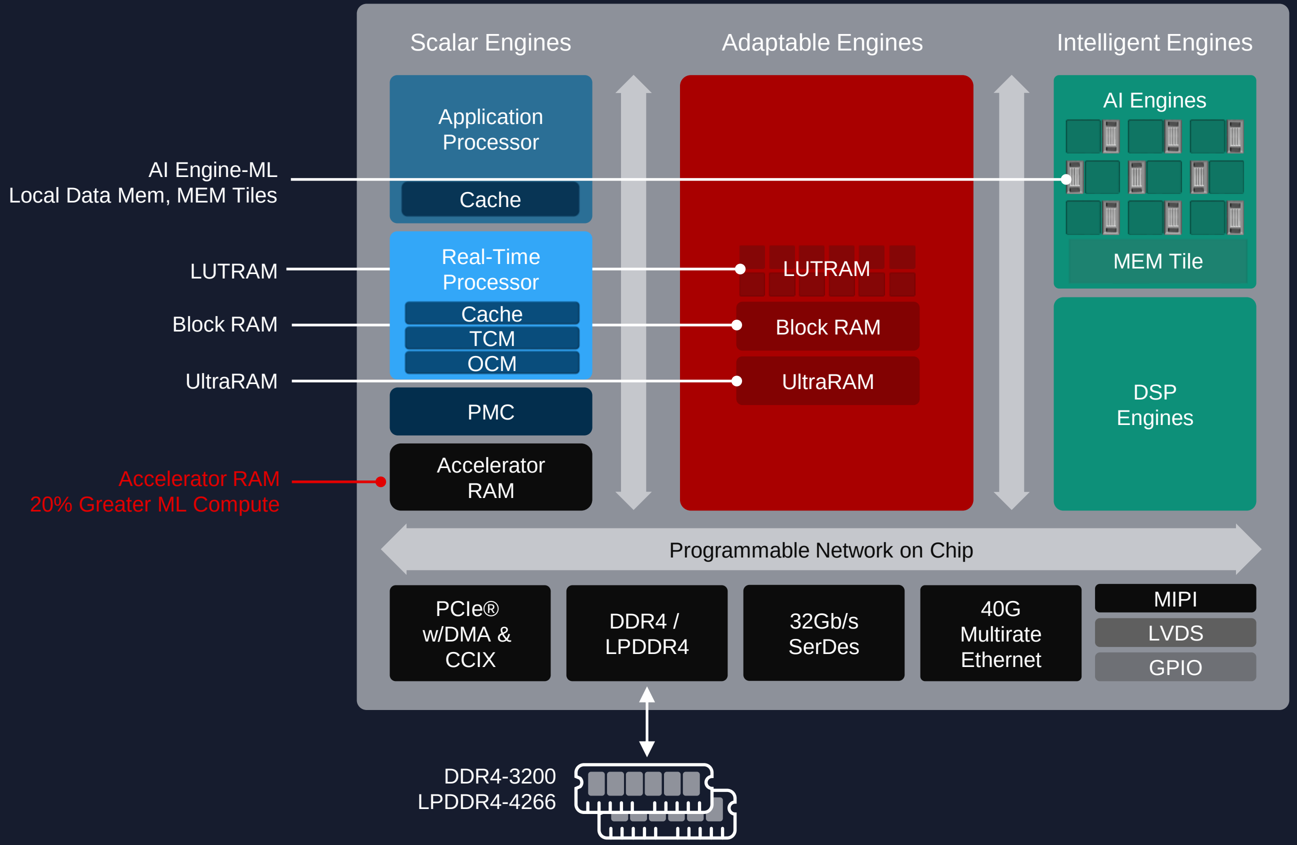
4MB of On-Chip RAM for Massive Bandwidth

Avoid DDR to store AI compute data or safety-critical code


AI Engines
35GB/s Memory Bandwidth
Accelerator RAM (4MB)
Holds AI Image & Compute Data
$W_1, \ldots, W_m$
Stores Safety-Critical Code
12.8GB/s Memory Bandwidth
Real-Time Processor Safety-Critical Decision Making

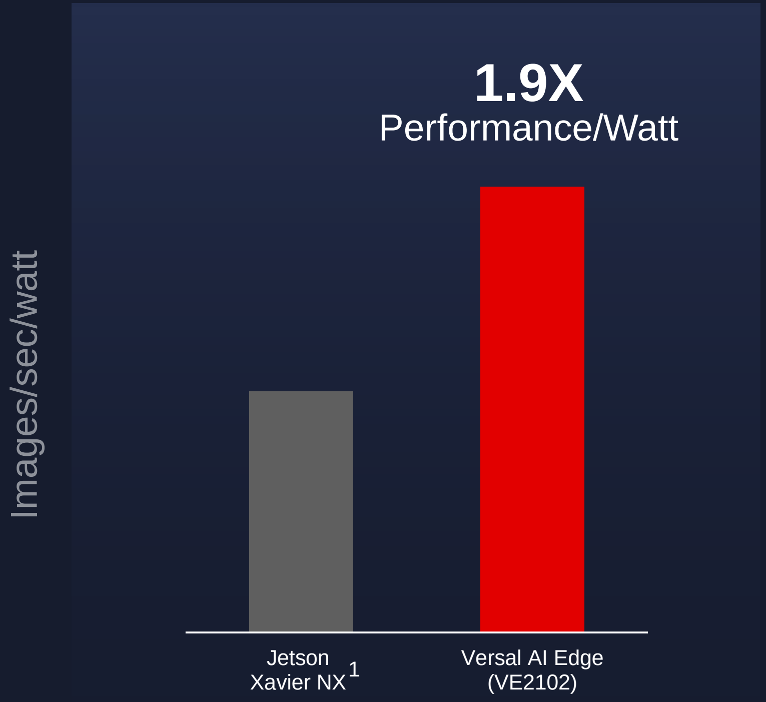
Part of the Adaptable Memory Hierarchy

Select the right memory for bandwidth requirement



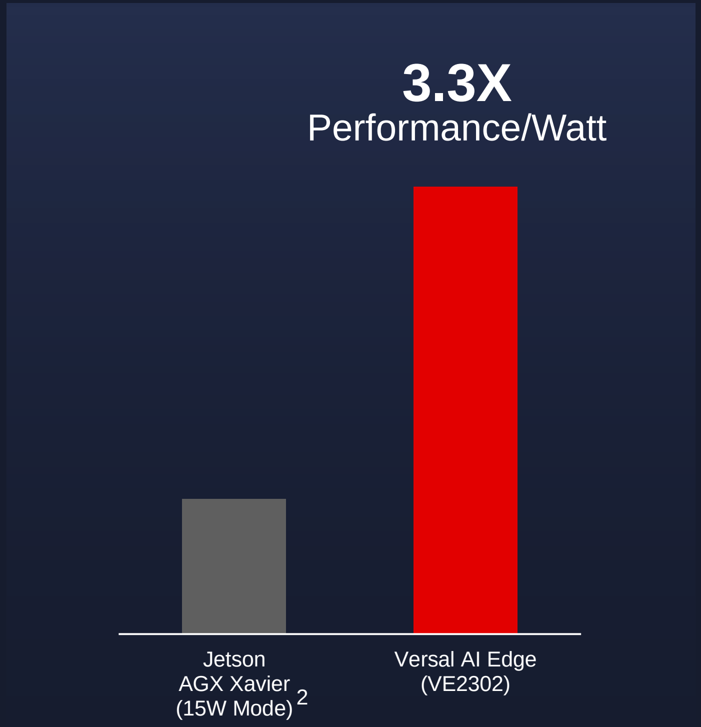
Up to 4X Performance/Watt vs. GPUs

Intelligent Edge Sensor



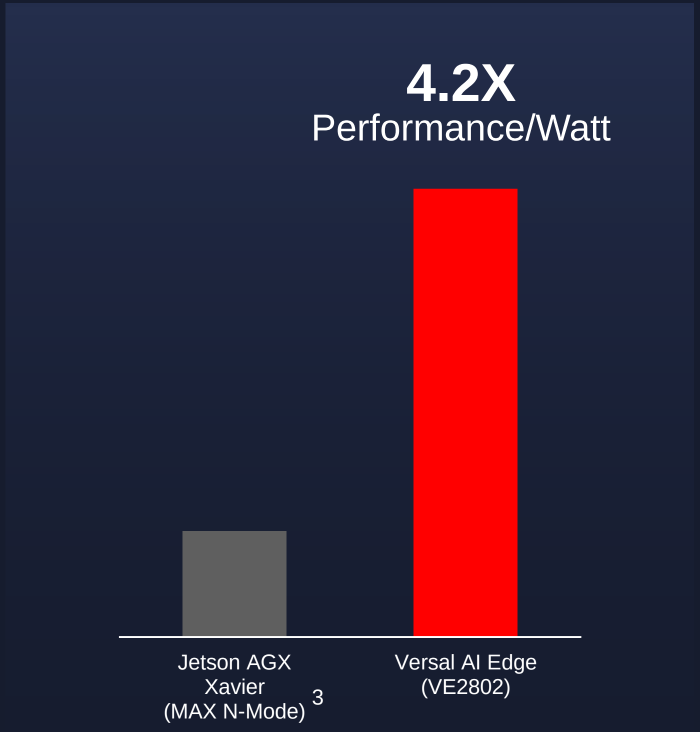
ResNet50 224x224

Autonomous System or Edge Aggregation



ResNet50 224x224, batch=1

CPU Accelerator



ResNet50 224x224, batch=1

1: Jetson NX Xavier: https://mlcommons.org/en/inference-edge-10, batch size not provided

2: Jetson AGX Xavier run in a mid-performance & power configuration, categorized as “15 W-Mode”: https://developer.nvidia.com/embedded/jetson-agx-xavier-dl-inference-benchmarks

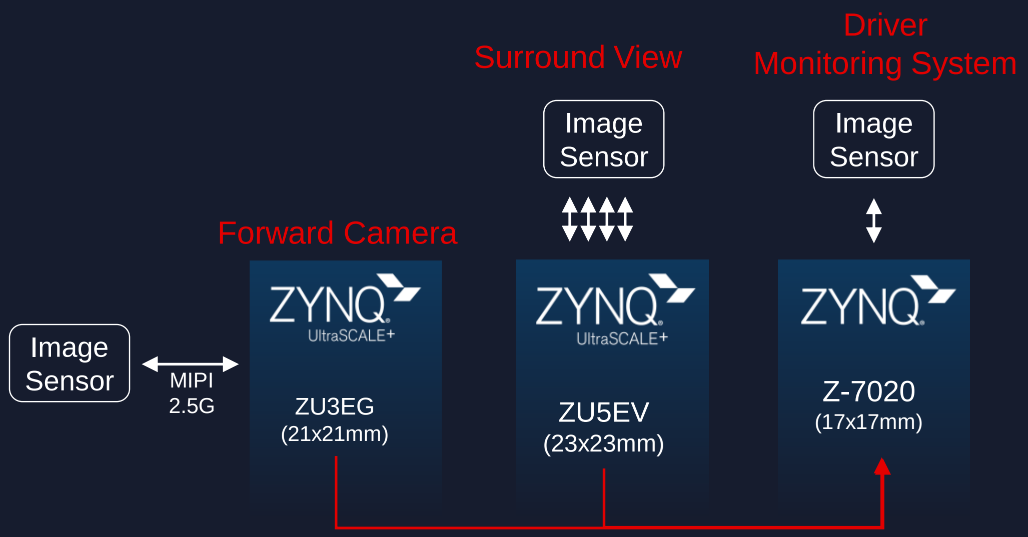
3: Jetson AGX Xavier MAX N-Mode and Versal™ VE2802 ACAP represent the highest performing device configuration in their respective portfolios

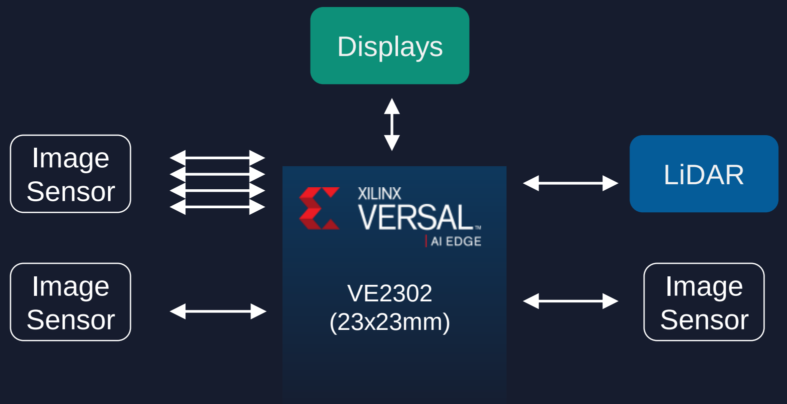
Jetson Xavier device power estimated by subtracting published memory & I/O power from total module power

All charts are normalized

10X Compute Density with Highest Levels of Safety and Security

10X Compute Density: Level 3 Semi-Automated Driving

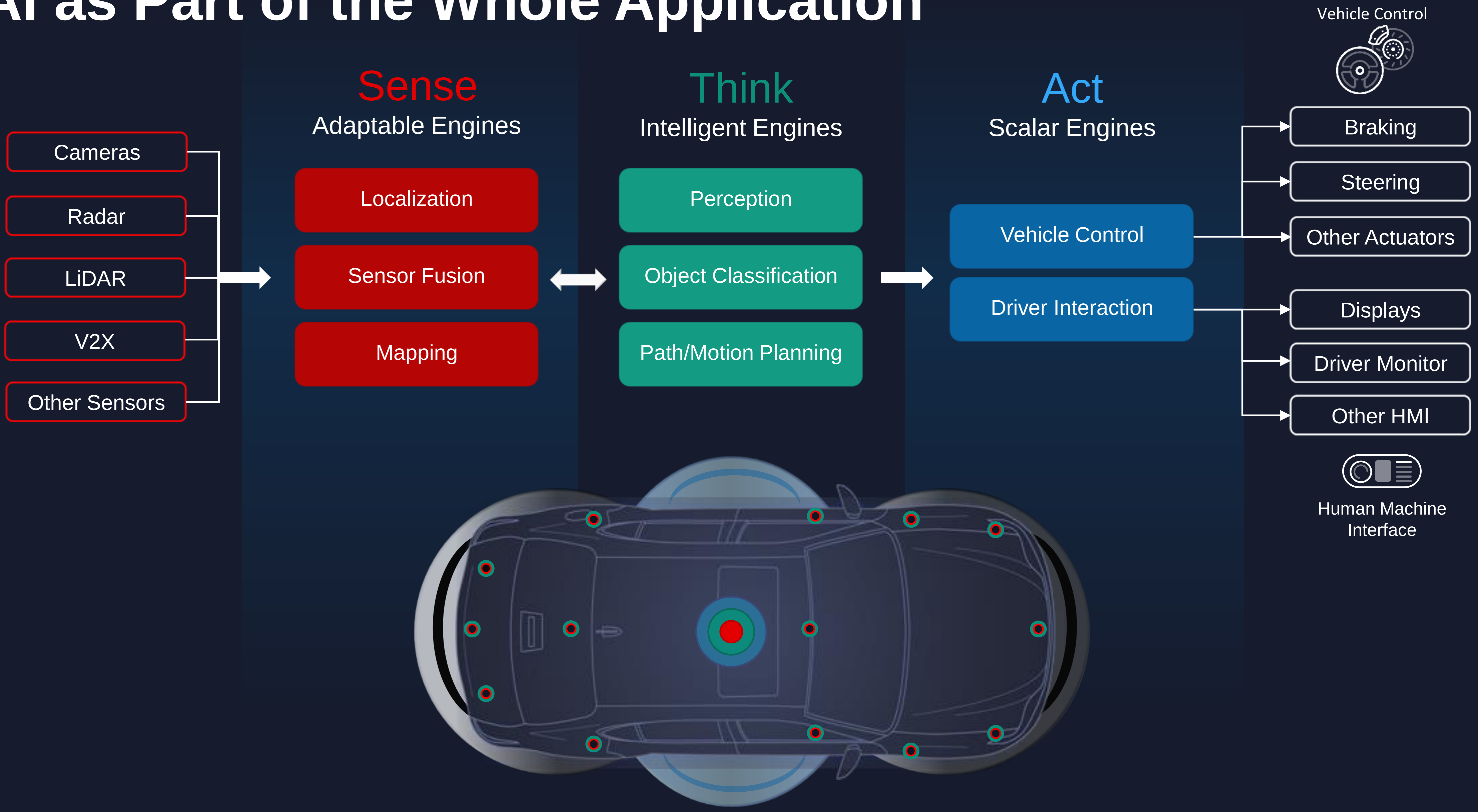




	Previous Gen Adaptive SoC		Versal™ ACAP
Compute	6x cameras (2MP, 4MP) + AI (4TOPs)	▶ 4.4X ▶	6x cameras (2MP, 8MP) + AI (17.4TOPs)
Area	3 devices = 1,259mm^2	▶ 58% Less ▶	1 device = 529mm^2
Power	ZU3(6W) + ZU5(10W) + Z-7020(5W)	▶ **~1X*** ▶	~20W

*Power levels are typical, approximate, and estimated at room temp

AI as Part of the Whole Application



Cameras
Radar
LiDAR
V2X
Other Sensors

Sense
Adaptable Engines
- Localization
- Sensor Fusion
- Mapping

Think
Intelligent Engines
- Perception
- Object Classification
- Path/Motion Planning

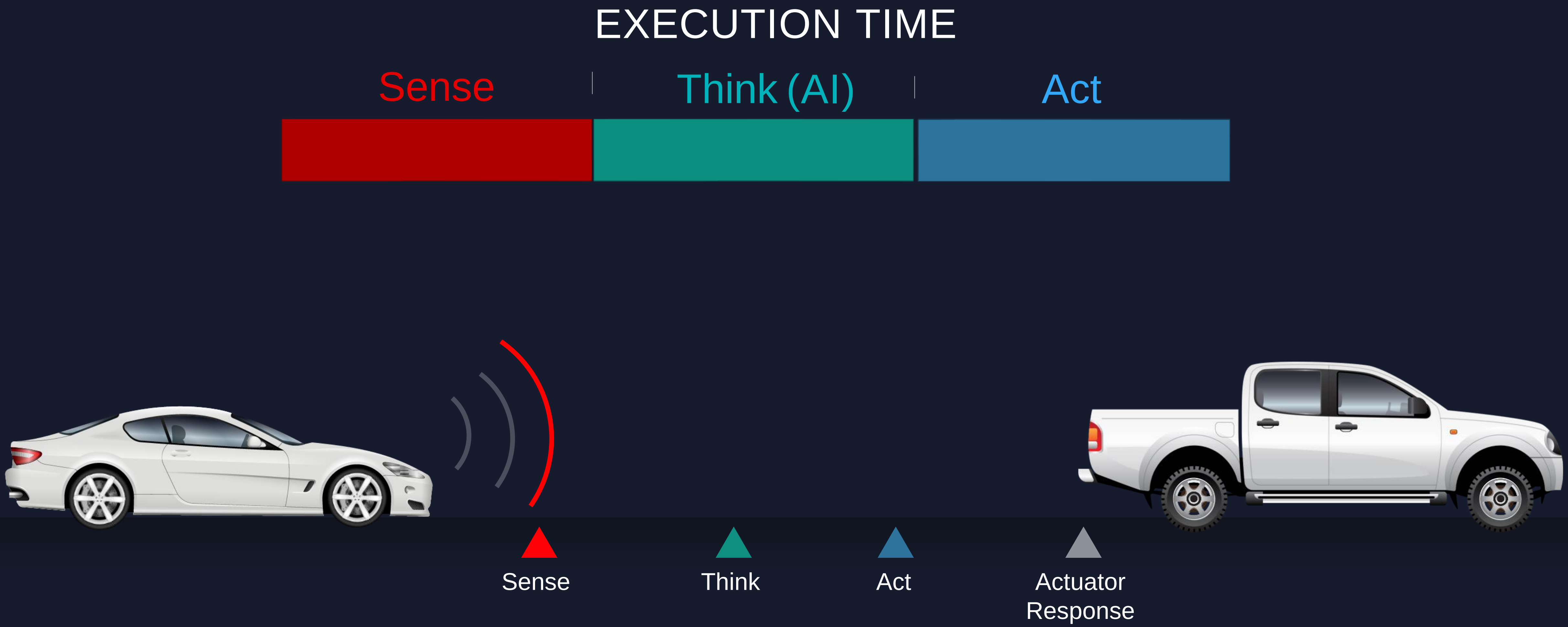
Act
Scalar Engines
- Vehicle Control
- Driver Interaction

Vehicle Control
- Braking
- Steering
- Other Actuators

Human Machine Interface
- Displays
- Driver Monitor
- Other HMI

Whole Application Acceleration for Real-Time Systems

From Sensor to AI to Real-Time Control



EXECUTION TIME

Sense | Think (AI) | Act

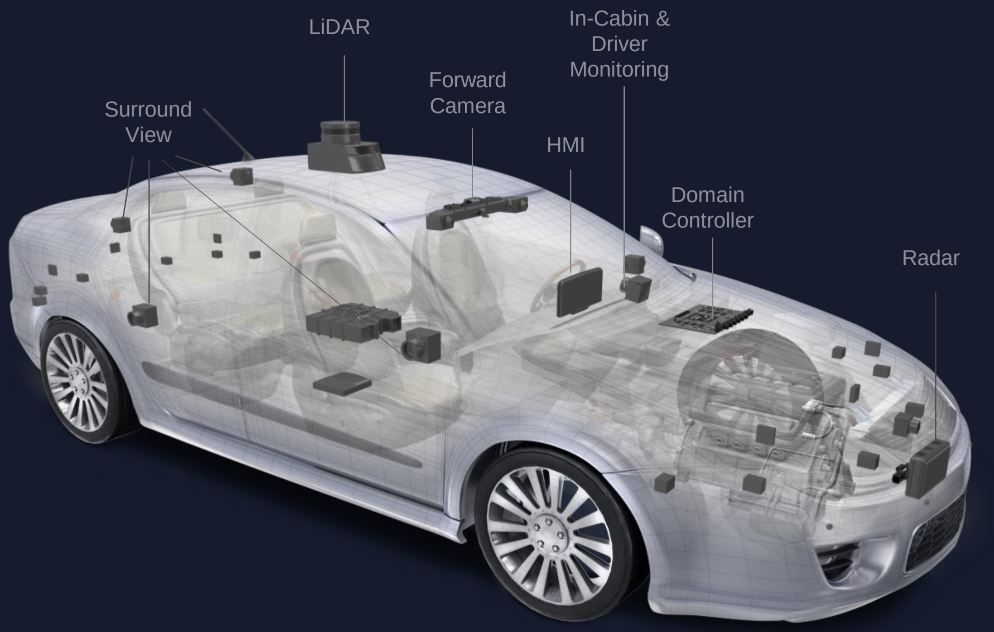
Sense

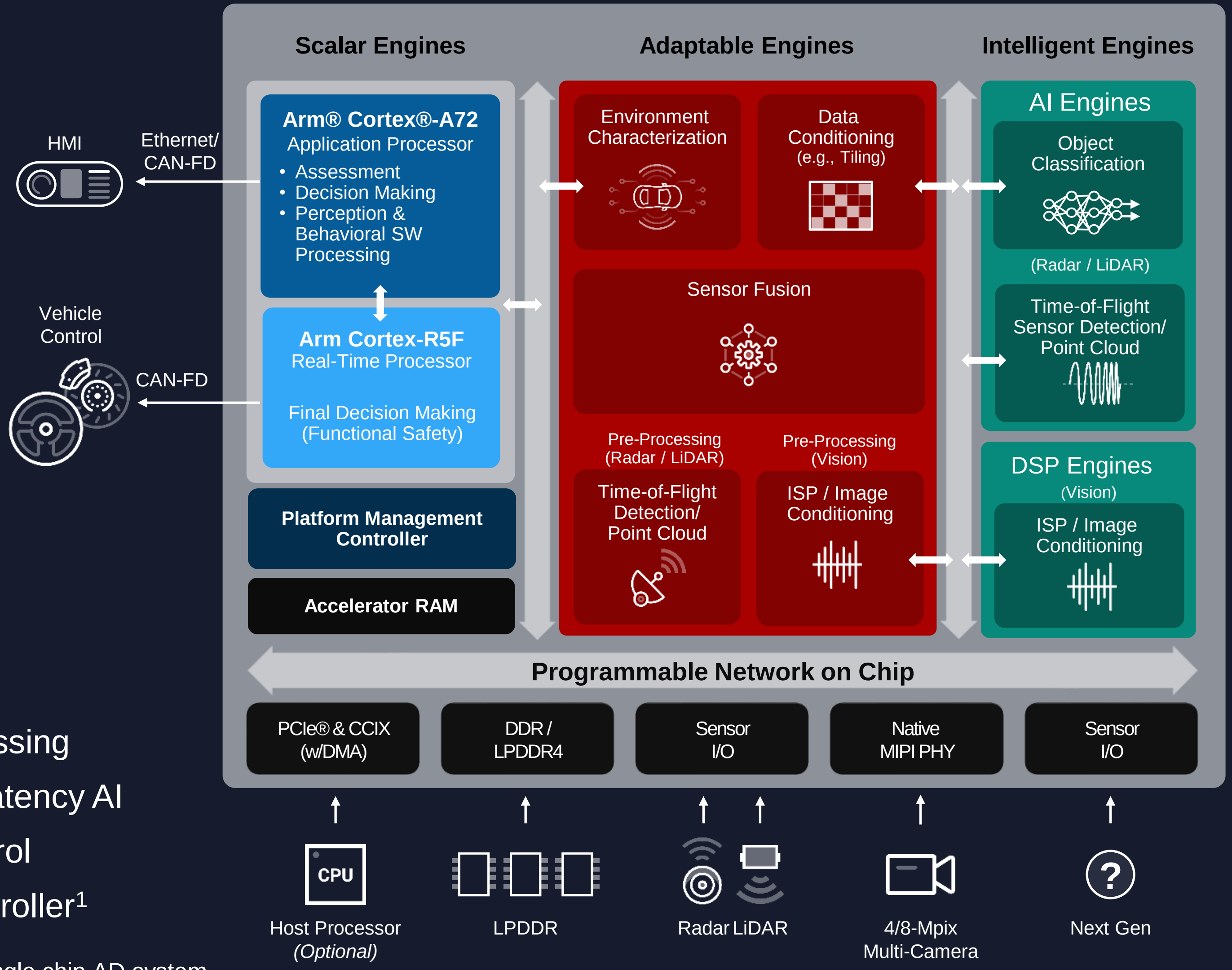
Think

Act

Actuator Response

Versal™ AI Edge ACAP in ADAS and Automated Driving





HMI

Ethernet/
CAN-FD

Vehicle
Control

CAN-FD

Scalar Engines

Arm® Cortex®-A72
Application Processor

- Assessment
- Decision Making
- Perception & Behavioral SW Processing

Arm Cortex-R5F
Real-Time Processor

Final Decision Making
(Functional Safety)

Platform Management Controller

Accelerator RAM

Adaptable Engines

Environment Characterization

Data Conditioning
(e.g., Tiling)

Sensor Fusion

Pre-Processing
(Radar / LiDAR)

Time-of-Flight Detection/
Point Cloud

Pre-Processing
(Vision)

ISP / Image Conditioning

Intelligent Engines

AI Engines

Object Classification

(Radar / LiDAR)

Time-of-Flight Sensor Detection/
Point Cloud

DSP Engines
(Vision)

ISP / Image Conditioning

Programmable Network on Chip

PCIe® & CCIX
(w/DMA)

DDR /
LPDDR4

Sensor
I/O

Native
MIPI PHY

Sensor
I/O

Host Processor
(Optional)

LPDDR

Radar LiDAR

4/8-Mpix
Multi-Camera

Next Gen

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
- Scalable compute from edge sensor to domain controller[1]

[1] Diagram demonstrates capabilities of architecture; does not represent a single chip AD system

Fully Automotive-Qualified and Safety Certified

Architected to Meet Stringent ISO 26262 Requirements

Supporting Multiple Safety Standards

Versal™ AI Edge ACAP

ISO 26262

Automotive

IEC 61508

Safety across All Industries

DO-254/178

Avionics HW/SW

IEC 61511

Process Industry

IEC 61800

Electrical Drives

ISO 13849

Machinery Control

IEC 62061

Machinery

EN 60601

Medical

Collaborative Robotics: AI-Based Systems Need to be Safe and Secure

Real-Time Precision and Control to Augment AI

Deterministic response, AI to navigate unpredictable movement of workers

Environmental Awareness and Perception

Sensor fusion for perception, self-learning to improve capabilities over time

Predictive Maintenance

Analyze sensor data for actionable insights to reduce downtime

Safety and Security are Connected Matters

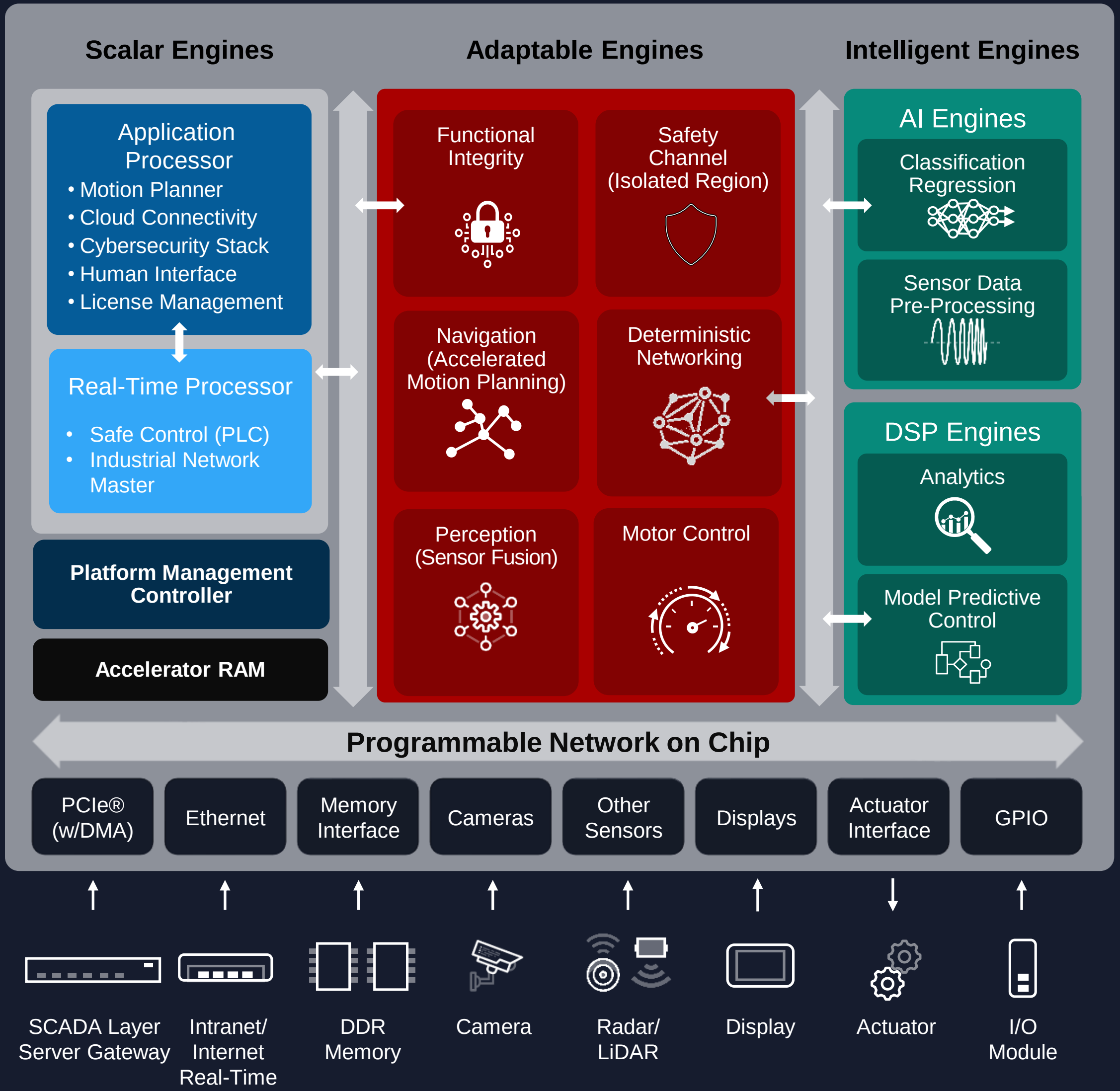
Cyber-attack creates safety and data privacy risks, robotic systems require IEC 62443 compliance

Human Safety
Real-Time Control
Sensor Data

Whole Application Acceleration for Collaborative Robotics

Robotic Perception Systems for Real-Time Control, Safety Critical, and Predictive Maintenance

- Adaptable Engines for perception, control/networking, navigation
- AI to augment control for dynamic execution, predictive maintenance
- Scalar Engines for cybersecurity (IEC 62443), safety control, UI



Scalar Engines

Application Processor
- Motion Planner
- Cloud Connectivity
- Cybersecurity Stack
- Human Interface
- License Management

Real-Time Processor
- Safe Control (PLC)
- Industrial Network Master

Platform Management Controller

Accelerator RAM

Adaptable Engines

Functional Integrity

Safety Channel (Isolated Region)

Navigation (Accelerated Motion Planning)

Deterministic Networking

Perception (Sensor Fusion)

Motor Control

Intelligent Engines

AI Engines

Classification Regression

Sensor Data Pre-Processing

DSP Engines

Analytics

Model Predictive Control

Programmable Network on Chip

PCIe® (w/DMA) | Ethernet | Memory Interface | Cameras | Other Sensors | Displays | Actuator Interface | GPIO

SCADA Layer Server Gateway | Intranet/ Internet Real-Time | DDR Memory | Camera | Radar/ LiDAR | Display | Actuator | I/O Module

AI-Enabled Multi-Mission Payloads for UAVs

AI with Software Defined Radio (SDR), Signal Intelligence (SIGINT), Image/Video Processing

Vision AI for Real-Time Analysis and Response

Autonomous flight control, optimize navigation paths

Cognitive RF

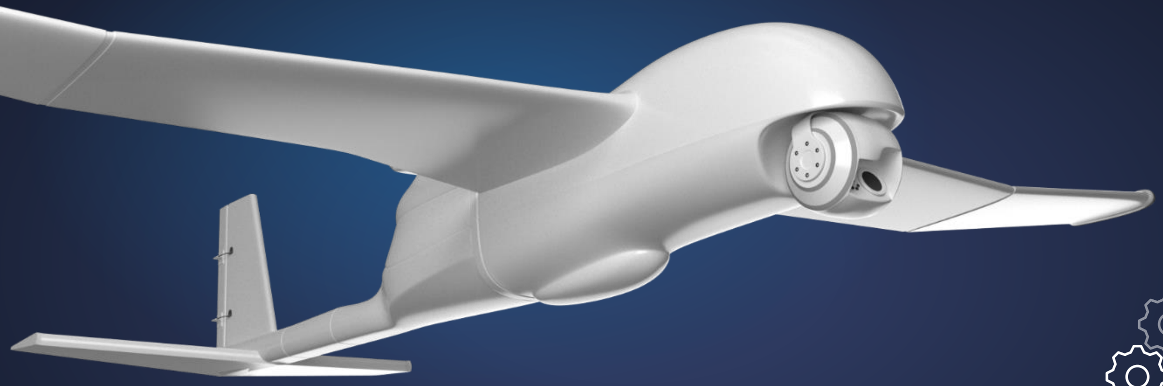
Optimizing radio communication and protecting against malicious intrusion

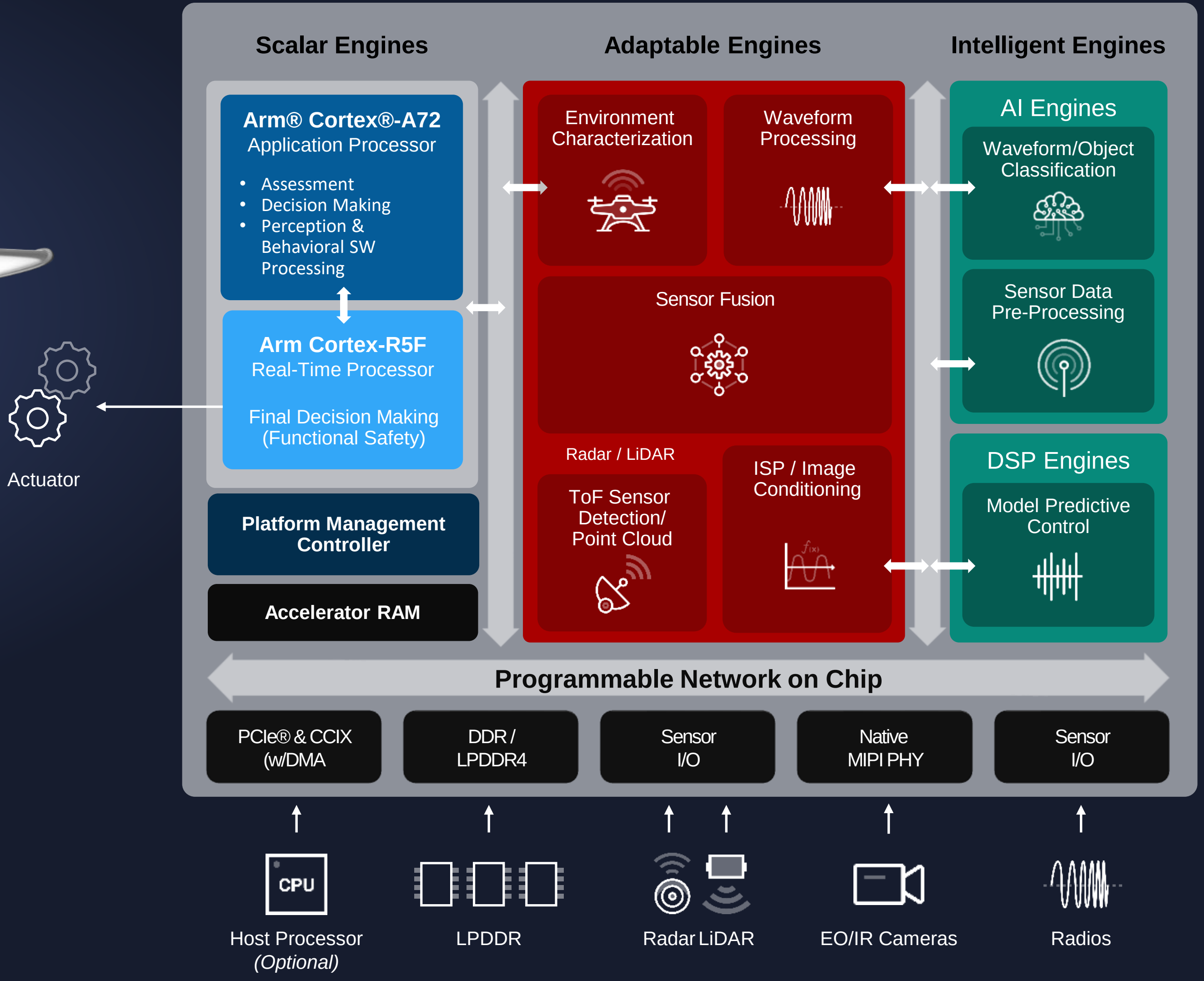
Diverse and Emerging Forms of AI

AI is rapidly evolving in tactical applications and vendors will need to adapt over time

Need AI Compute in Limited Size, Weight, and Power (SWaP) and Thermal Envelope

Versal AI Edge for Unmanned Aerial Vehicles





Multi-Mission and Situationally Aware UAVs with Low SWaP

- Adaptable Engines for sensor fusion and pre-processing
- Intelligent Engines for low power, low latency AI and signal conditioning
- Scalar Engines for command and control
- Ruggedized packaging and military-temp grade (XQ)

XILINX

Versal ACAP Development Experience for All Developers

HW Developer

VIVADO

SW Developer

XILINX VITIS

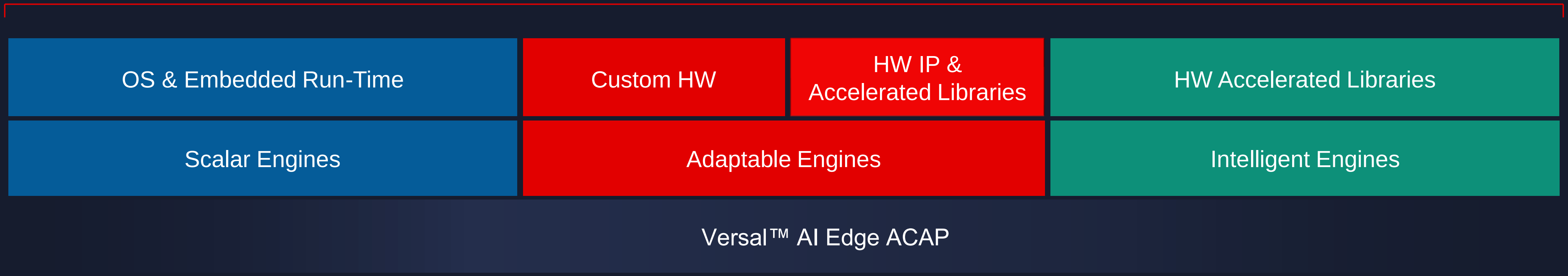
C, C++, Python

Data Scientist

XILINX VITIS | AI

TensorFlow PyTorch Caffe

Spark FFmpeg mxnet



OS & Embedded Run-Time

Custom HW

HW IP & Accelerated Libraries

HW Accelerated Libraries

Scalar Engines

Adaptable Engines

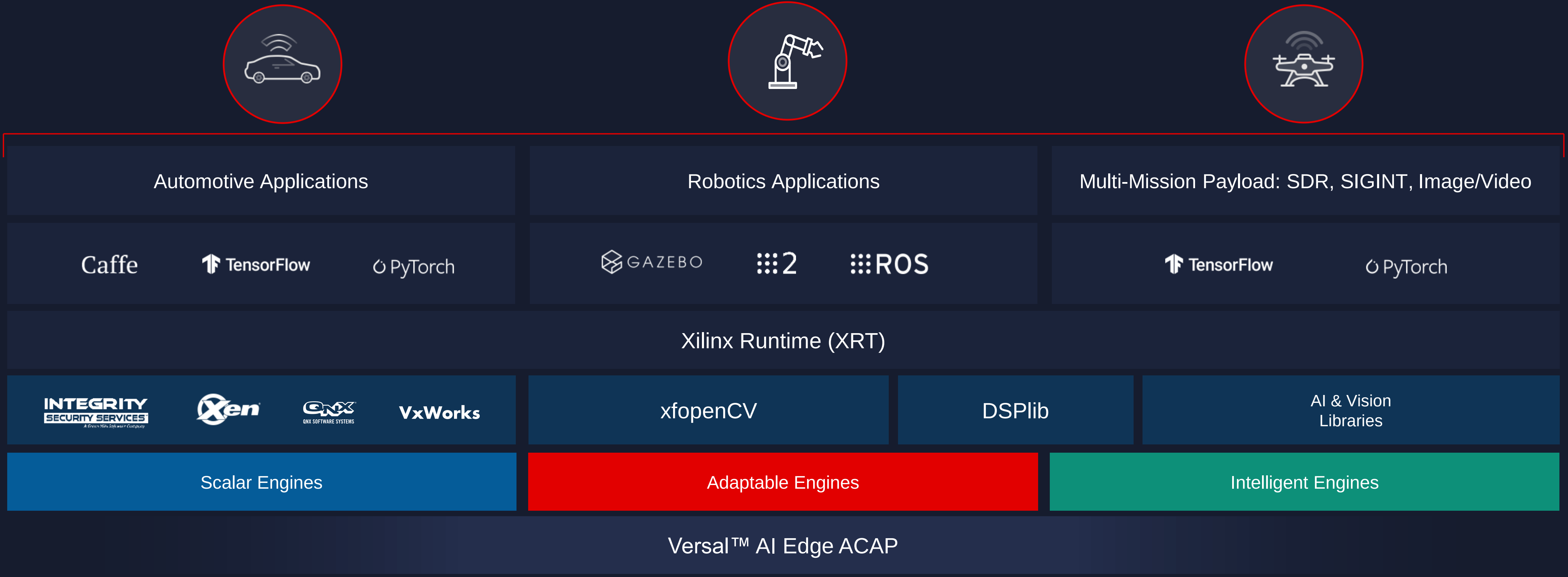
Intelligent Engines

Versal™ AI Edge ACAP

Market-Specific Application Stacks

Examples for Automotive, Robotics, and Multi-Mission Payload Applications

- One platform with market-specific libraries, frameworks, and ecosystem to enable all developers
- Following industry standards for developing safety critical software on silicon



Automotive Applications

Caffe TensorFlow PyTorch

Robotics Applications

GAZEBO ROS 2 ROS

Multi-Mission Payload: SDR, SIGINT, Image/Video

TensorFlow PyTorch

Xilinx Runtime (XRT)

INTEGRITY SECURITY SERVICES Xen QNX VxWorks

xfopenCV

DSPlib

AI & Vision Libraries

Scalar Engines

Adaptable Engines

Intelligent Engines

Versal™ AI Edge ACAP

World’s Most Adaptable and Scalable Edge Platform

Adaptability: From Domain Specific Architectures (DSAs) to Dynamic Function Exchange

DSAs for Diverse Platform Requirements

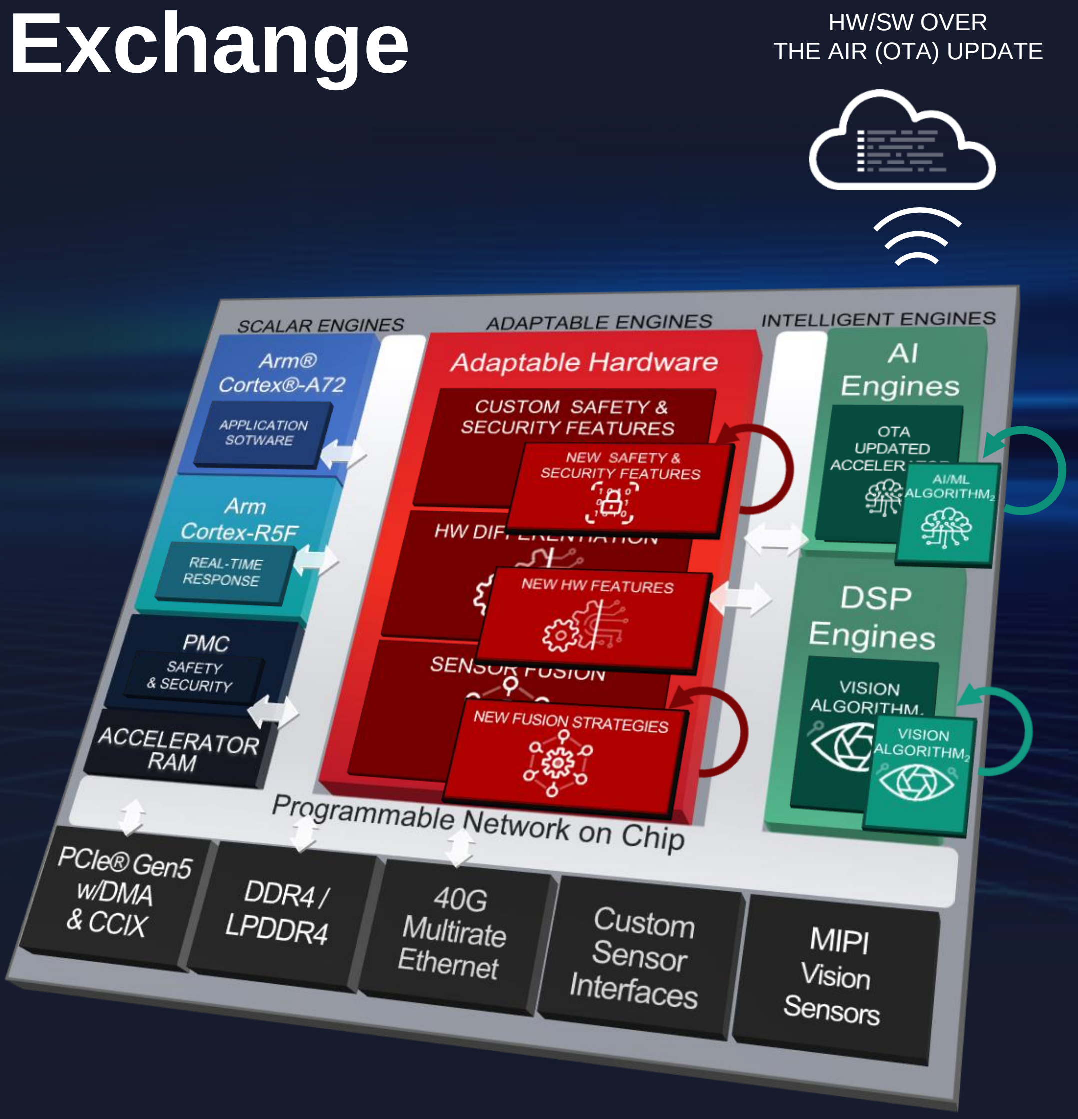
- Implement custom AI, vision, sensor strategies
- Design for different safety and security targets
- One platform for diverse end-customers' requirements

Hardware/Software Over-the-Air Updates

- Update your AI accelerator or fusion algorithms
- Future proof for emerging security threats
- Avoid recalls or costly re-deployment

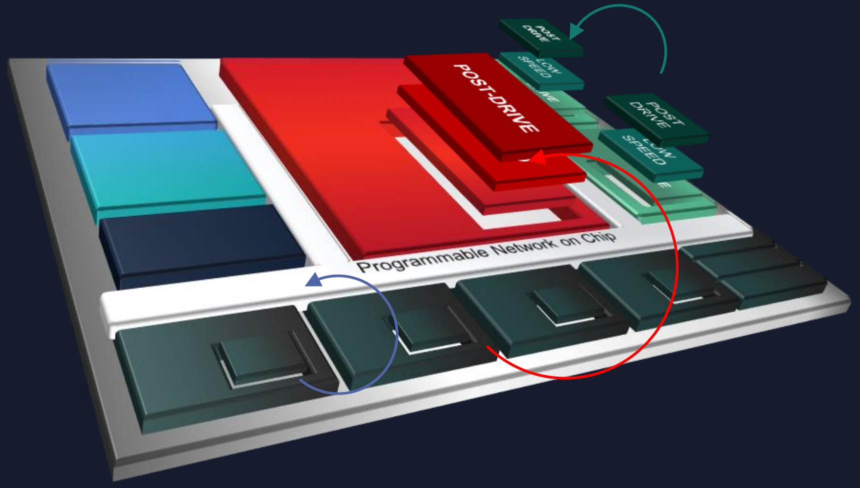
Dynamic Function Exchange (DFx)

- Swap functionality in milliseconds
- Available in Adaptable Engines, DSP, AI Engines
- Fewer system components → reduce power and cost

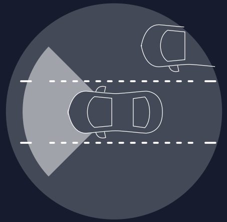


Dynamic Function Exchange (DFx) in Automotive

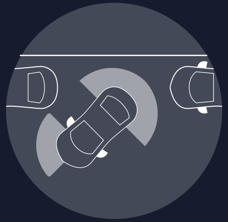
Swap Functionality in
Milliseconds



Dynamic Regions
(Engines, Integrated Cores, I/O)



Drive Mode
(Lane Departure
Warning)

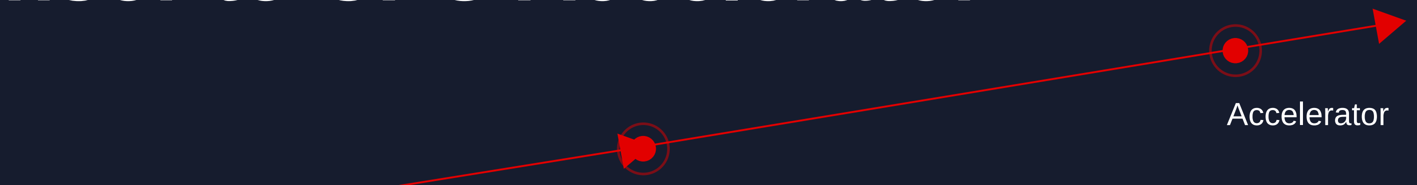


Low Speed Mode
(Parking Assist)

Post-Drive Mode
(Dog Left Behind)

Fewer Devices to Reduce System-Wide Power and Cost

Scale from Edge Sensor to CPU Accelerator



Intelligent Edge Sensor & End Point

Edge Aggregation & Autonomous Systems

Accelerator

Lead Device

		VE2002	VE2102	VE2202	VE2302	VE2602	VE1752	VE2802
Engines	Total AI Compute (INT4)	14 TOPS	22 TOPS	47 TOPS	67 TOPS	256 TOPS	166 TOPS	479 TOPS
	Total AI Compute (INT8)	7 TOPS	10 TOPS	21 TOPS	31 TOPS	120 TOPS	124 TOPS	228 TOPS
	AIE / AIE-ML[1]	8	12	24	34	152	304	304
	Adaptable Engines	20K LUTs	37K LUTs	105K LUTs	150K LUTs	375K LUTs	449K 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	95Mb	103Mb	156Mb	172Mb	554Mb	253Mb	575Mb
	32G Transceivers	-	-	8	8	32	44	32
	PCIe®	-	-	✓	✓	-	-	-
	PCIe + CCIX	-	-	-	-	✓	✓	✓
	Estimated Power	6–9W	7–10W	15–20W	~20W	50–60W	50–60W	75W

[1] VE2xx2 based on AIE-ML, VE1752 device base based on AIE

The Only Edge AI Platform that Scales from Sensor to Accelerator on a Single Architecture[1,2]

1–100 Watts

	0–10TOPS	10–25TOPS	25–75TOPS	100+ TOPS
XILINX VERSAL AI EDGE	●	●	●	●
NVIDIA (Jetson)	●	●	●	
NVIDIA (T4)				●
NVIDIA (ORIN)			●	●
mobileye An Intel Company	●	●	●	
Texas Instruments	●	●	●	
Qualcomm snapdragon	●	●		
Qualcomm® Cloud AI 100			●	●
NXP (iMX8)	●			
RENESAS	●		●	

[1] Shown in INT8 TOPS
[2] Based on published sources

Scalable for Different Requirements and Product Features

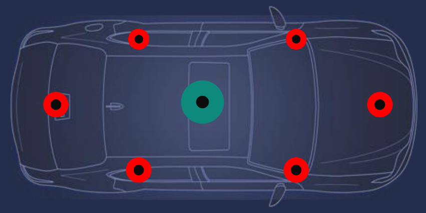
Level 2
Level 3
Level 4
Level 5

Scale for Varying Levels of Compute and Safety

- Scale number of sensors, AI compute, vision and video processing
- e.g., Scale from Level-3 ADAS to Level-5 automated drive on a single platform

Scale a Low-End to High-End End-Product Portfolio

- Design once, scale with same tools, SW, ecosystem, safety certification
- Scale for different price points and capabilities



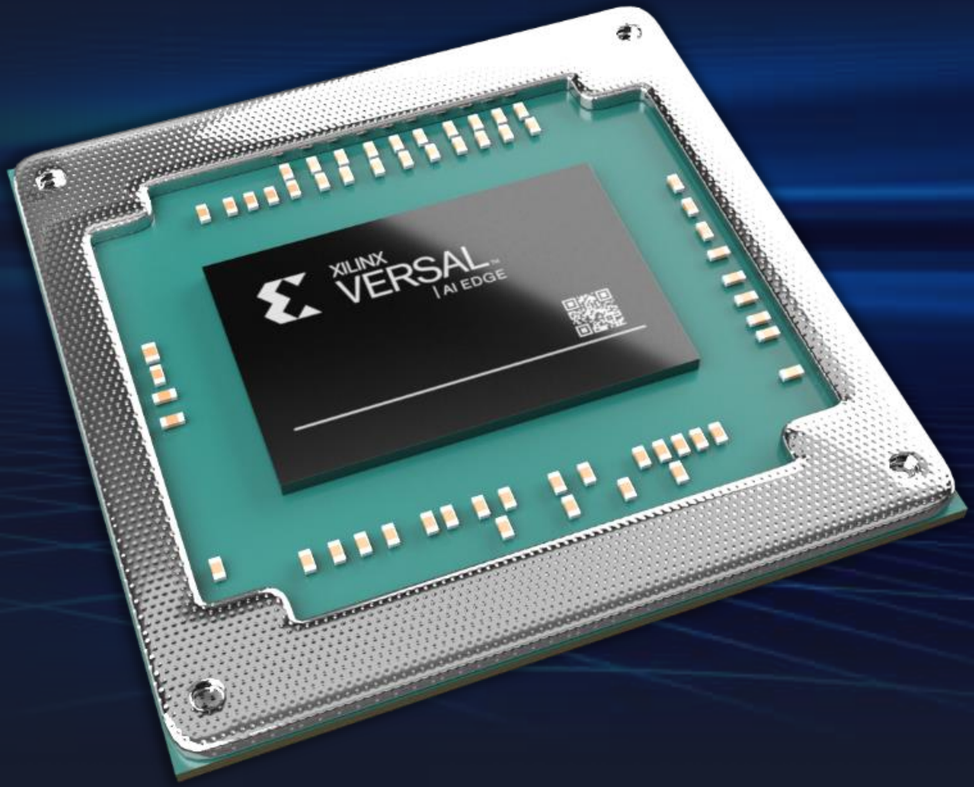
Explore Distributed vs. Centralized Architectures

- "Load Balance" across the system
- Shift compute from edge sensor to central compute across a single system

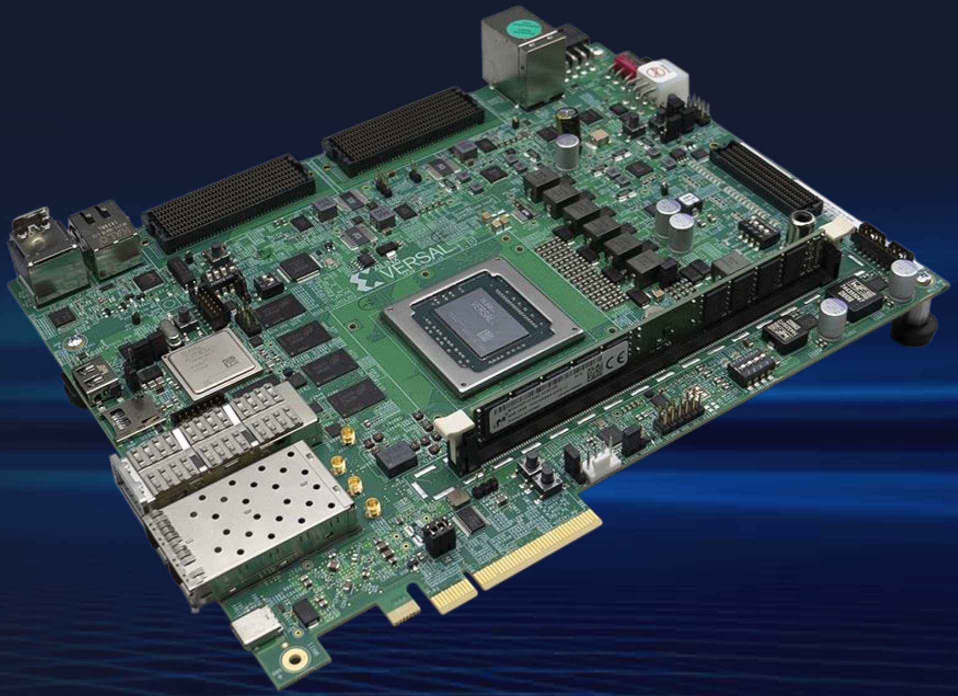
How Customers Can Get Started

Availability

- Documentation Available Now
- Tools Available in 2nd Half of ‘21
- ES & Production Silicon in 1st Half ‘22
- Versal™ AI Edge ACAP Eval Kit in 2nd Half ‘22



Start Prototyping Now



Start Now with Versal AI Core ACAP VCK190 Evaluation Kit

Migrate Later to Versal AI Edge Device

Evaluate Key Blocks in Versal™ AI Edge

Leverage Vitis™ Accelerated Libraries

Breadth of Interfaces for System Testing

System-Design Methodology Guides

Guided Flows in Vitis and Vivado® Tools

www.xilinx.com/vck190

Versal AI Edge ACAP: Intelligence Unleashed

From Sensor to AI to Real-Time Control

- 4X AI Performance/Watt vs. GPUs[1] with Innovations in AI Engines and Memory Hierarchy
- 10X Compute Density[2] with Highest Levels of Safety and Security
- World's Most Scalable and Adaptable Platform for Edge Systems

Silicon Sampling in 1st Half 2022

[1] vs. Jetson AGX Xavier (MAX N-Mode), ResNet50 224x224, batch=1, https://developer.nvidia.com/embedded/jetson-agx-xavier-dl-inference-benchmarks
[2] Compared to Zynq® UltraScale+™ MPSoCs

Thank You