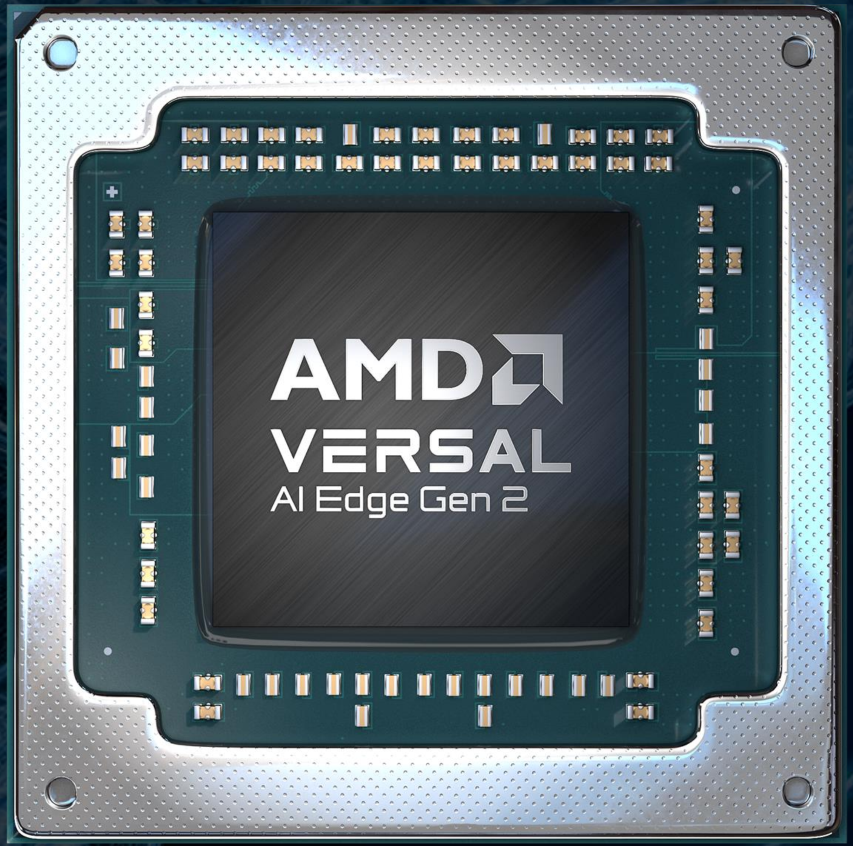
AMD Versal™ AI Edge Series Gen 2 for Vision and Automotive

Tomai Knopp and Jeffrey Chu

Co-author: Sagheer Ahmad

Hot Chips 2024



Agenda

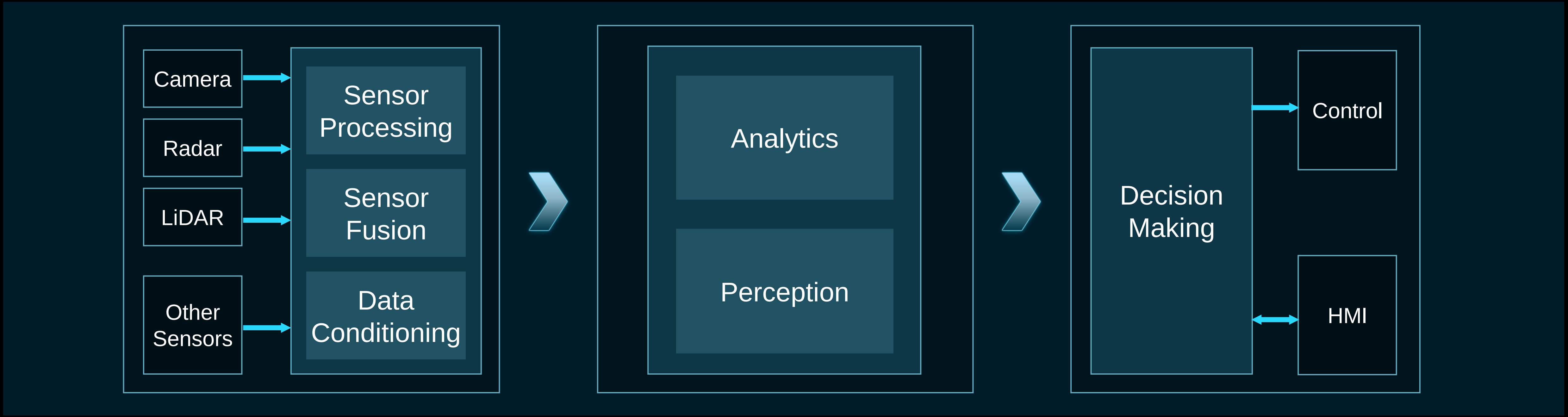
- Challenges and Motivation
- Silicon Features
- Vision Applications

AI Driven Embedded Processing Phases

Preprocessing
Sensor Processing & Control, Data Conditioning

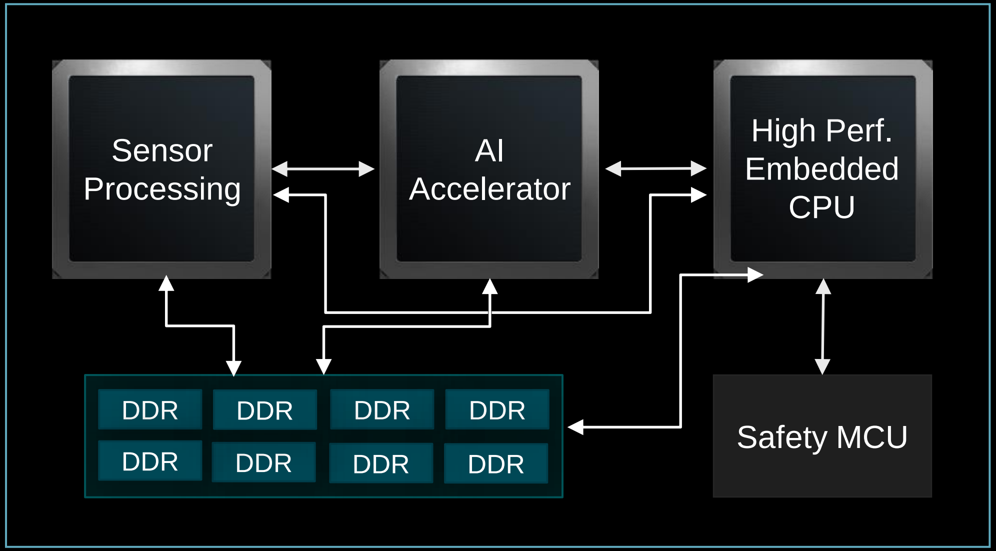
AI Inference
Perception, Analytics

Postprocessing
Decision Making, Control, Feedback

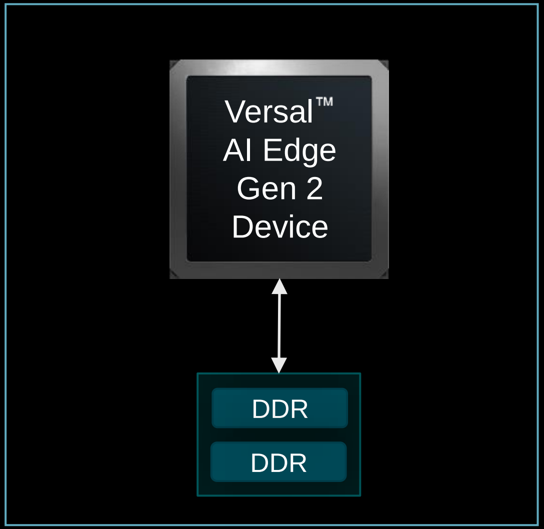


Superior Integration to Reduce System Power, Area & Complexity

Multi-Chip AI-Driven Embedded System



Single-Chip System with Versal™ AI Edge Series Gen 2



Challenging Environments

Limited Power Availability

Tight Form Factor Requirements

Real-Time Response

Security, Safety & Reliability

Long Life Cycles

AMD

AMD Versal™ AI Edge Series Gen 2 Overview

Next-Generation AI Engines for Efficient AI Inference

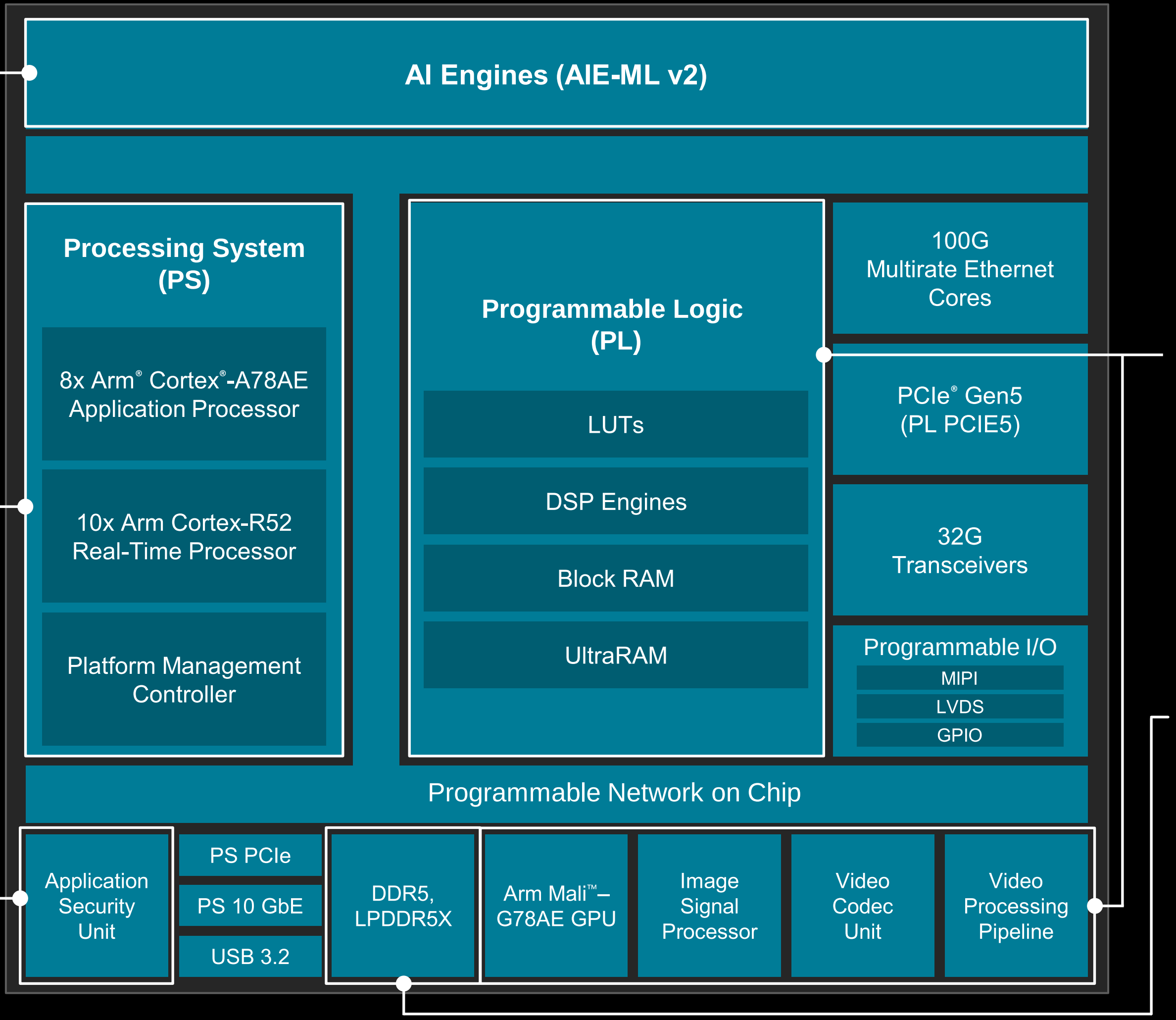
Up to 3X TOPs/Watt*

High-Performance Integrated CPUs for Postprocessing

Up to 10X Scalar Compute*

Enhanced Safety & Security

Support for ASIL D, SIL 3



World-Class Programmable Logic for Flexible, Real-Time Preprocessing

- Sensor Fusion, Data Conditioning
- New Hard Image/Video Processing

User Selectable System Memory Interfaces for Data Intensive Accesses

- Up to 170 GBytes/sec Bandwidth*

Pre-Silicon Estimated Performance

See Endnotes VER-023, VER-027

AMD Versal™ AI Edge Series Gen 2 Product Table

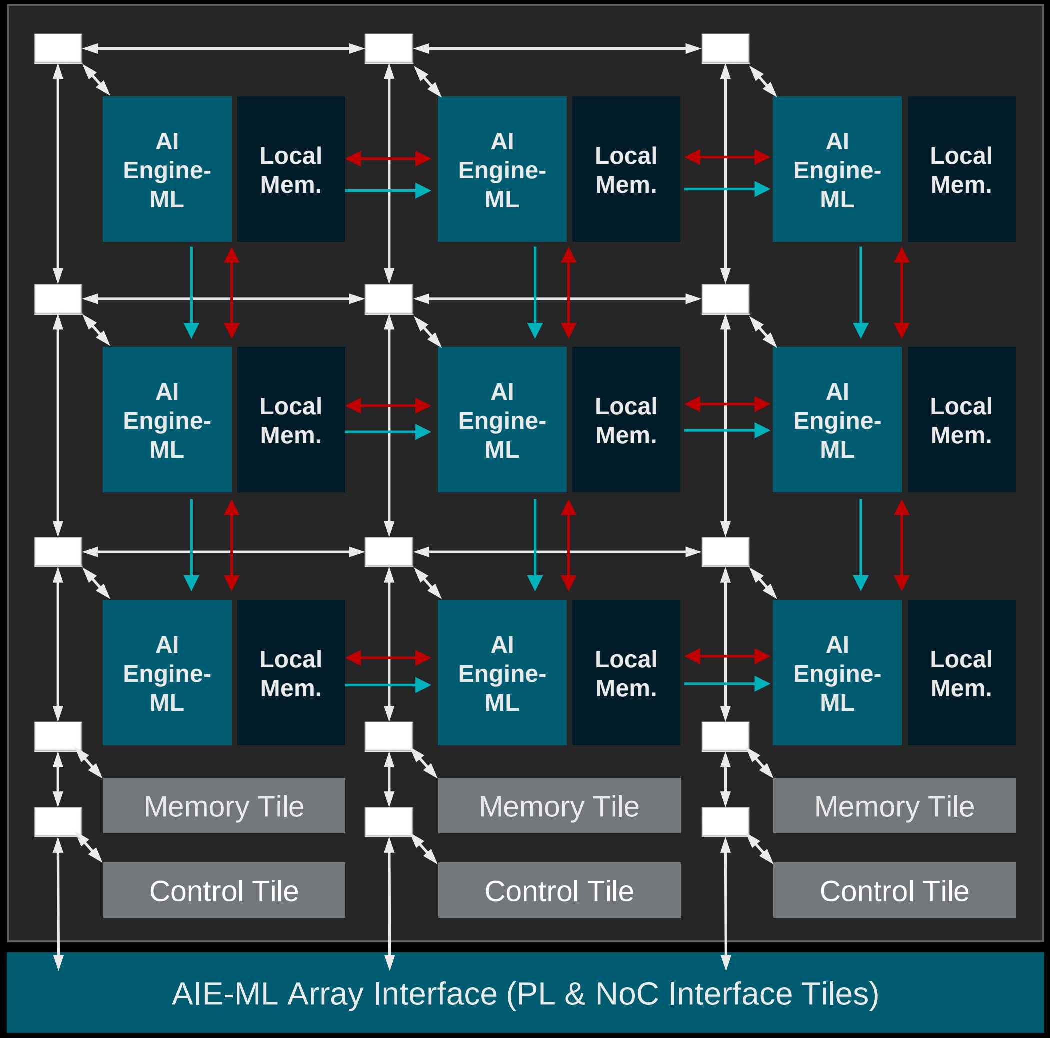
End-to-end Acceleration for AI-driven Embedded Systems

		2VE3304	2VE3358	2VE3504	2VE3558	2VE3804	Lead Device 2VE3858
Compute	AIE-ML v2 Tiles	24	24	96	96	144	144
Compute	Max. Dense INT8 TOPs	31	31	123	123	184	184
Compute	#APU Cores (Arm® Cortex®-A78AE)	4	8	4	8	4	8
Compute	# of RPU Cores (Arm Cortex-R52)	4	10	4	10	4	10
Compute	LUT6	94k	94k	225k	225k	543k	543k
Compute	DSP	184	184	700	700	2,064	2,064
Hard IP	32b Memory controller	3	3	4	4	5	5
Hard IP	Max. Memory Bandwidth (GB/s)	102	102	136	136	170	170
Hard IP	PL PCIe® (Gen4/5x4)	1	1	3	3	4	4
Hard IP	MRMAC (10/25/100GE)	1	1	1	1	3	3
Hard IP	GPU (4k60 200GFLOPs+)	1	1	1	1	1	1
Hard IP	Video Codec Unit (VCU) Tiles	-	1	-	1	-	1
Hard IP	Image Signal Processor (ISP) Tiles	-	1	-	3	-	3
Hard IP	Video Processing Pipeline (VPP) Tiles	-	1	-	-	-	-
IO	PS-Facing GTYP / PL-Facing GTYP	4 /4	4 /4	4 / 12	4 / 12	4 / 20	4 / 20
IO	X5IO / HDIO[1]	260 / 44	260 / 44	384 / 88	384 / 88	512 / 44	512 / 44
Package	A1089 (27 mm x 27 mm)	■	■				
Package	A1444 (31 mm x 31 mm)	■	■	■	■		
Package	A2112 (37.5 mm x 37.5 mm)			■	■	■	■

1. Maximum X5IO and HDIO counts may not be available in same package.

AMD

AI Engines: AIE-ML v2



TOPs in AMD Versal™ AI Edge Series Gen 2 Devices

	2VE3858	2VE3558	2VE3358	
MX6	**369**	246	61	TFLOPS
INT8 (sparse)	**369**	246	61	TOPS
INT8 (dense)	184	123	31	TOPS
FP8/MX9	184	123	31	TFLOPS
FP16/BF16	92	61	15	TFLOPS
INT16 (sparse)	92	61	15	TOPS
INT16 (dense)	46	31	8	TOPS

Subset of supported data types; values assume highest speed grade.

1. Pre-Silicon Estimated Performance vs. Previous Generation

See Endnotes VER-025, VER-026

AMD

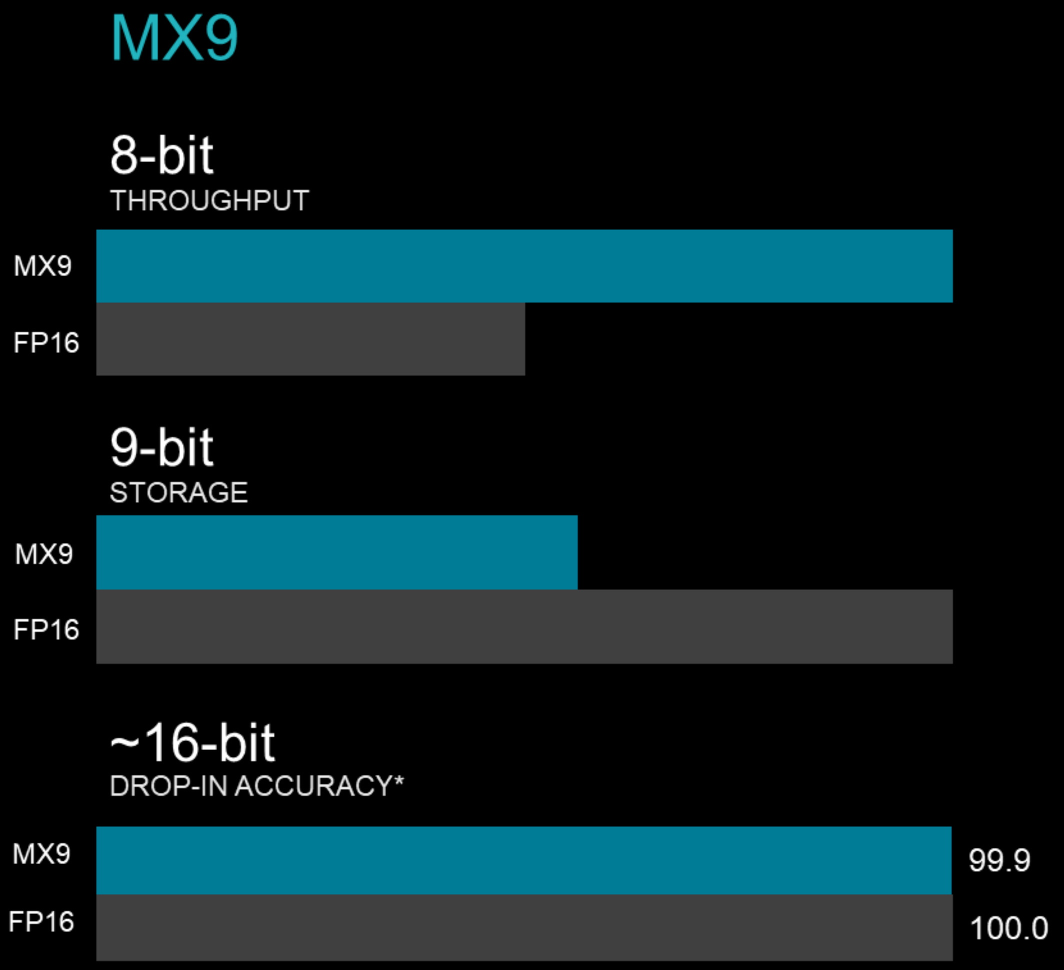
AI Engine Architecture Features

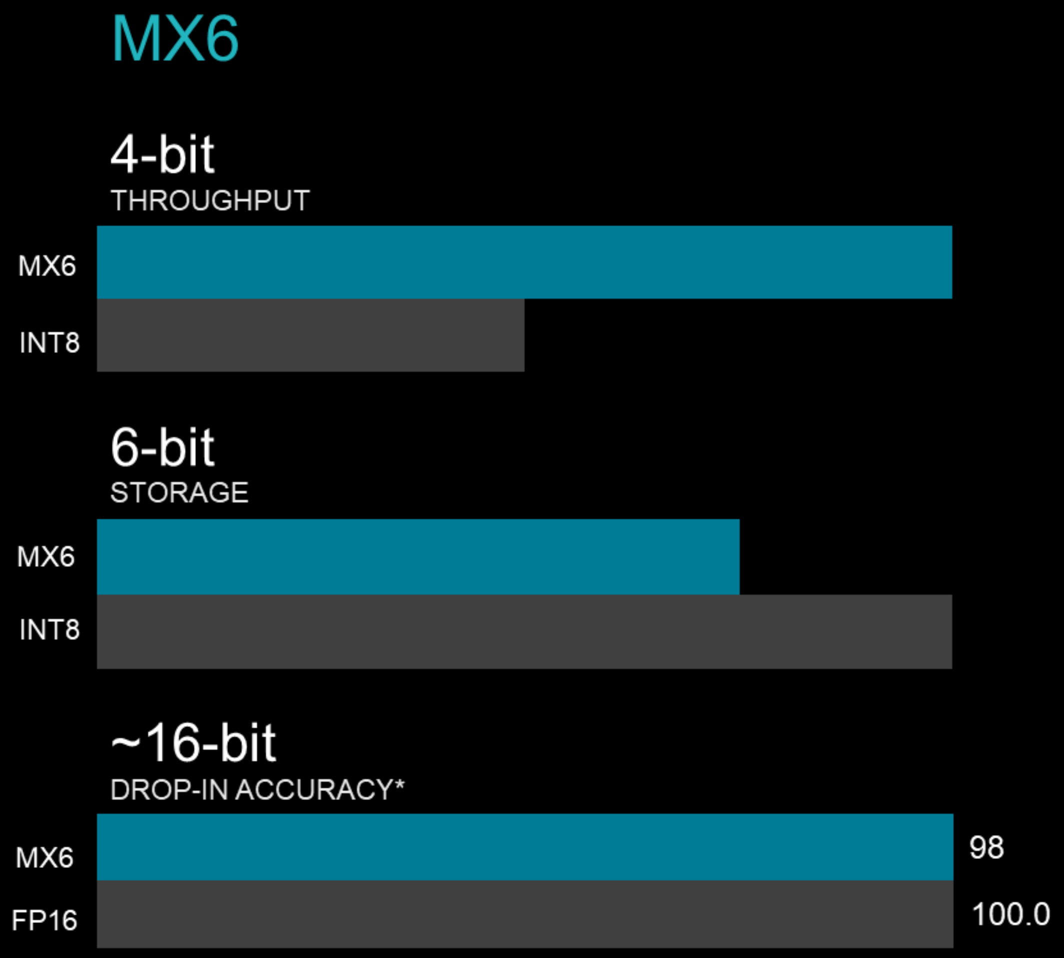
	Family	AMD Versal AI Edge Series		AMD Versal AI Edge Series Gen 2
	AIE Version	AIE-ML		AIE-MLv2
Compute (Mults / Tile)	INT8	256	2x	512
	BFLOAT16	128	2x	256
	FP8	N/A	New	512
	FP16	N/A	New	256
	MX6*	N/A	New	1024
	MX9*	N/A	New	512
	Compression and Sparsity	Yes		Yes
	AIE Array Interconnect B/W	1x (32b)	2x	2x (64b)
	Tile Local Data Memory	64 KB		64 KB
	Memory Tile	AIE Memory Tile (512KB/tile)		AIE Memory Tile (512KB/tile)
	AIE Controller	Programmable Logic Based	New	Hardended Microblaze per column

* MX6 and MX9 datatypes defined in https://arxiv.org/pdf/2302.08007, reference to Table II.

AMD

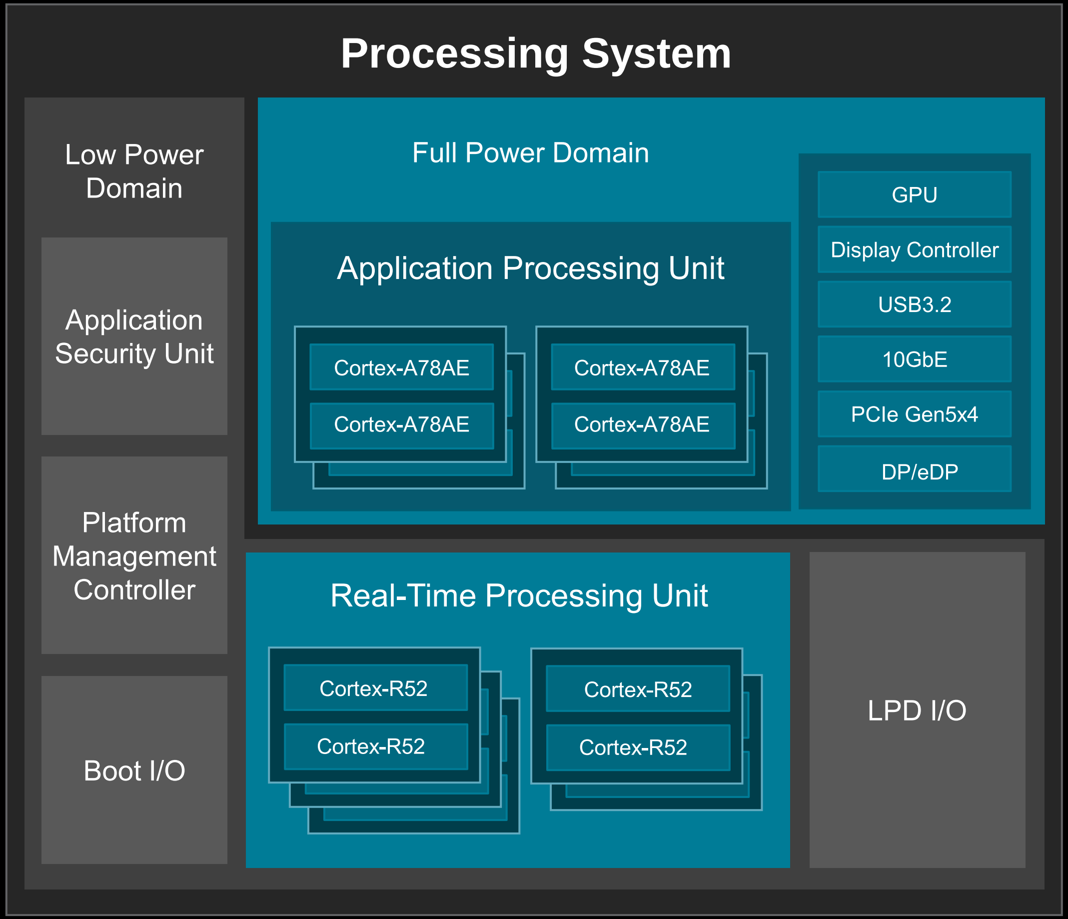
MX9 and MX6 Datatypes





*See Endnotes VER-69 and VER-70

Processing System Overview



Application Processing Unit (APU)

- Arm® Cortex®-A78AE cores
- Up to 2.2 GHz max frequency per core[1]
- Up to 200.3k DMIPs in APU[2]

Real-Time Processing Unit (RPU)

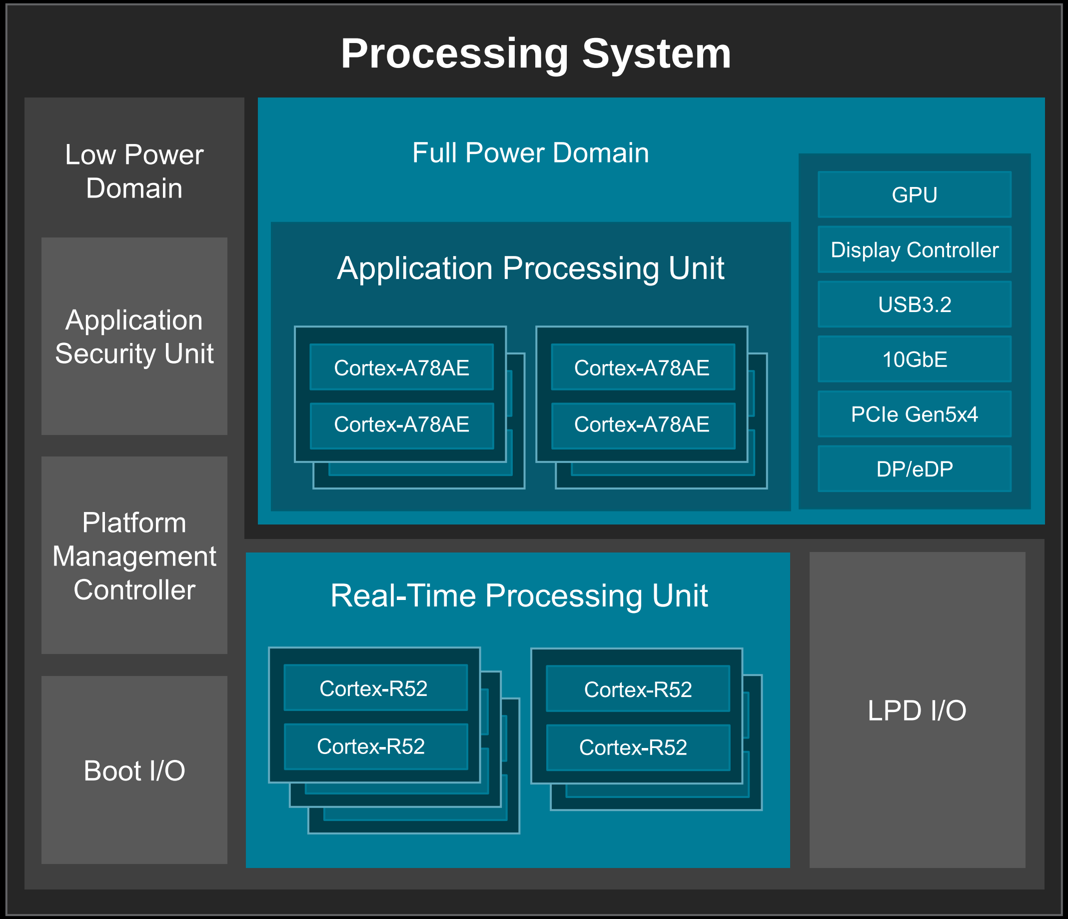
- Arm Cortex-R52 cores
- Up to 1.05 GHz max frequency per core[1]
- Up to 28.5k DMIPs in RPU[2]

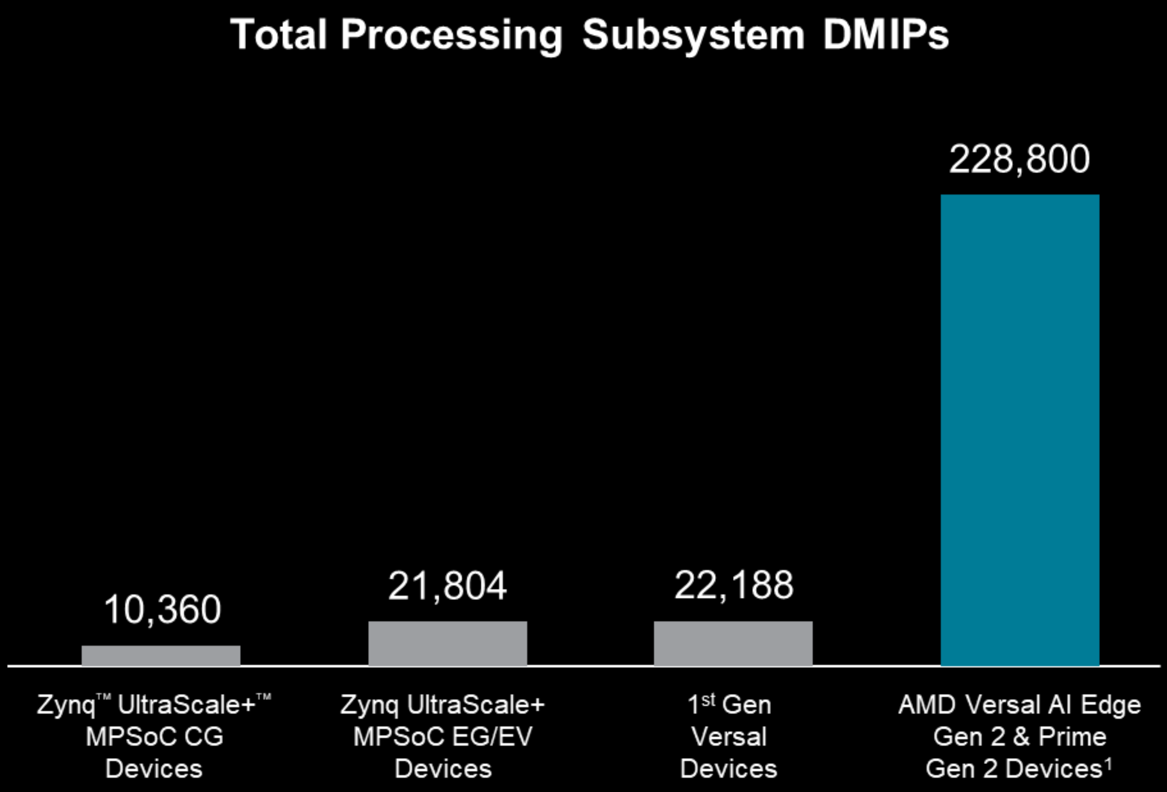
Graphics Processing Unit (GPU)

- Arm® Mali™ G78AE GPU
- Up to 1.05 GHz max frequency
- Up to 268GFlops

1: Pre-silicon estimated performance vs. prior generation. See Endnotes VER-027, VER-030

Processing System Overview





Source: AMD internal data, February 2024

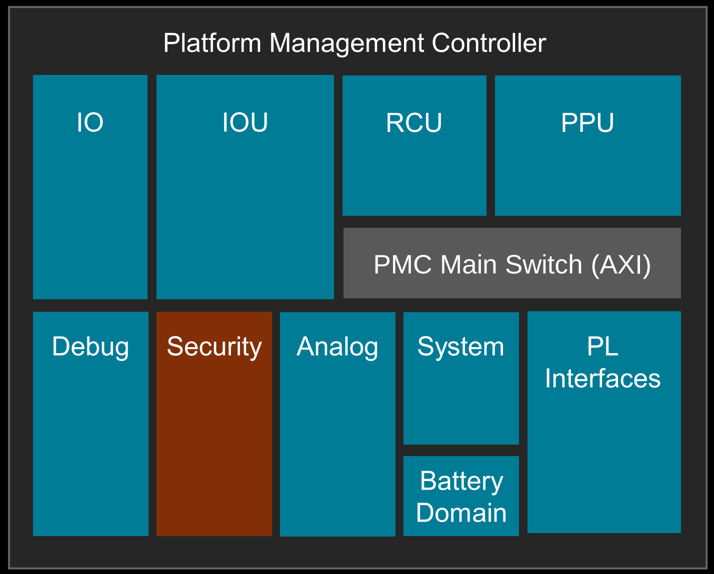
1. Pre-Silicon Estimated Performance

See Endnotes VER-027

Embedded System Security

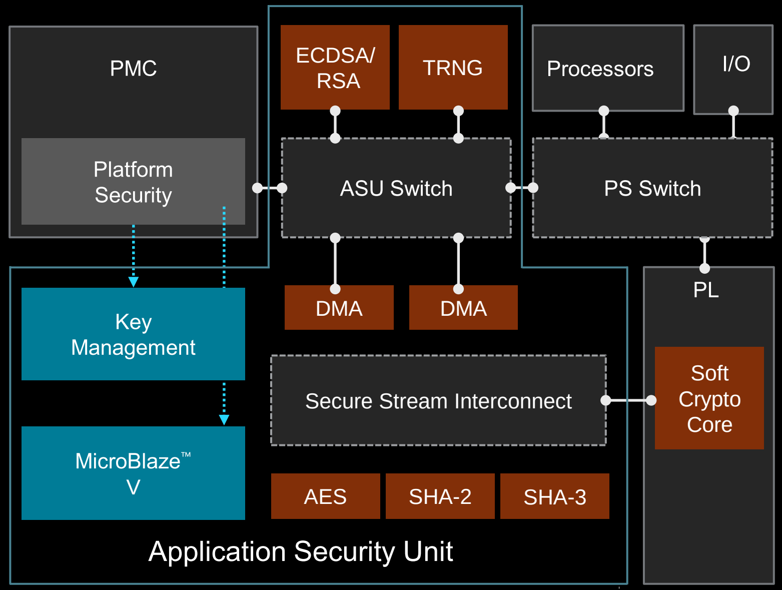
Platform Management Controller (PMC)

Manages device-level security services
(Secure Boot, HWRoT, Physical Attack Protection, etc.)



Application Security Unit (ASU)

Provides run-time HSM security
(encryption / authentication / key management)



Memory Controller Inline Crypto (ILC)

Built-in inline encryption within DDR5 / LPDDR5X
memory controllers (AES-XTS or AES-GCM)

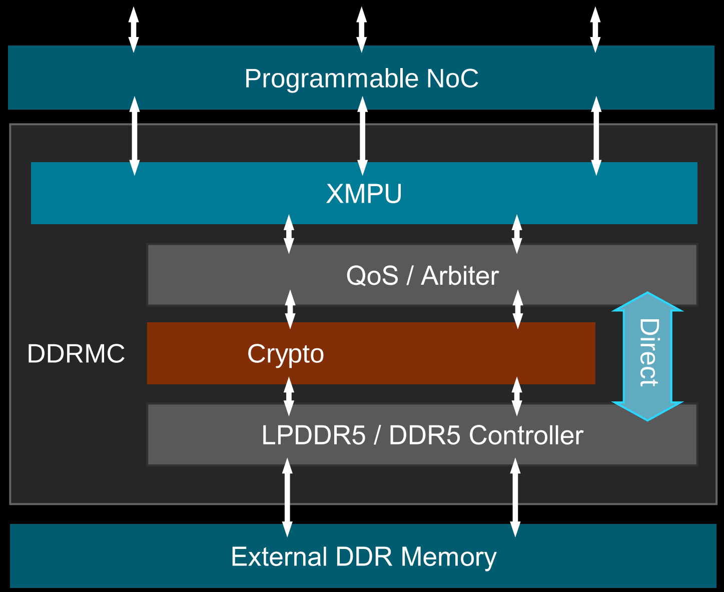
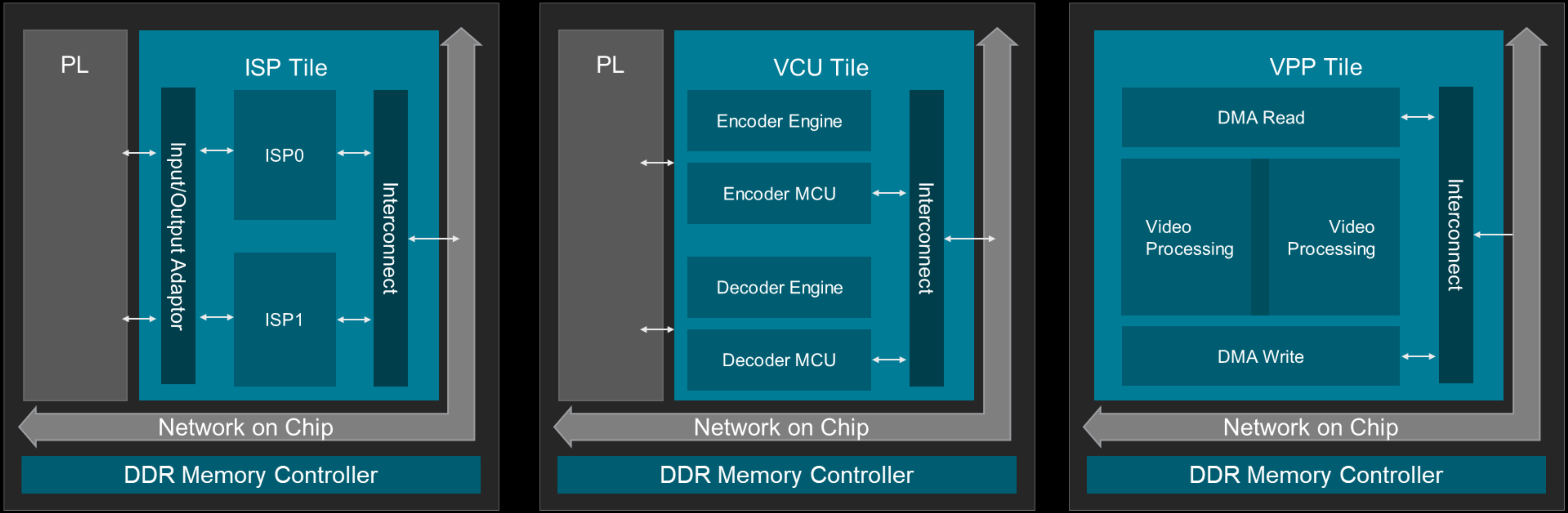


Image and Video Processing IP



2 ISP instances per ISP Tile

1GPIX/s with up to 5 Cameras per tile

Encoder/Decoder per VCU Tile

4Kp60 H.265/264 supported

8Kp30 with dual VCU tiles

Up to 8Kp60 or dual 8Kp30

Supports all common video processing functions

Functional Safety

QM Systematic & Random Fault Integrity

ASIL D/SIL3 Systematic Fault Integrity

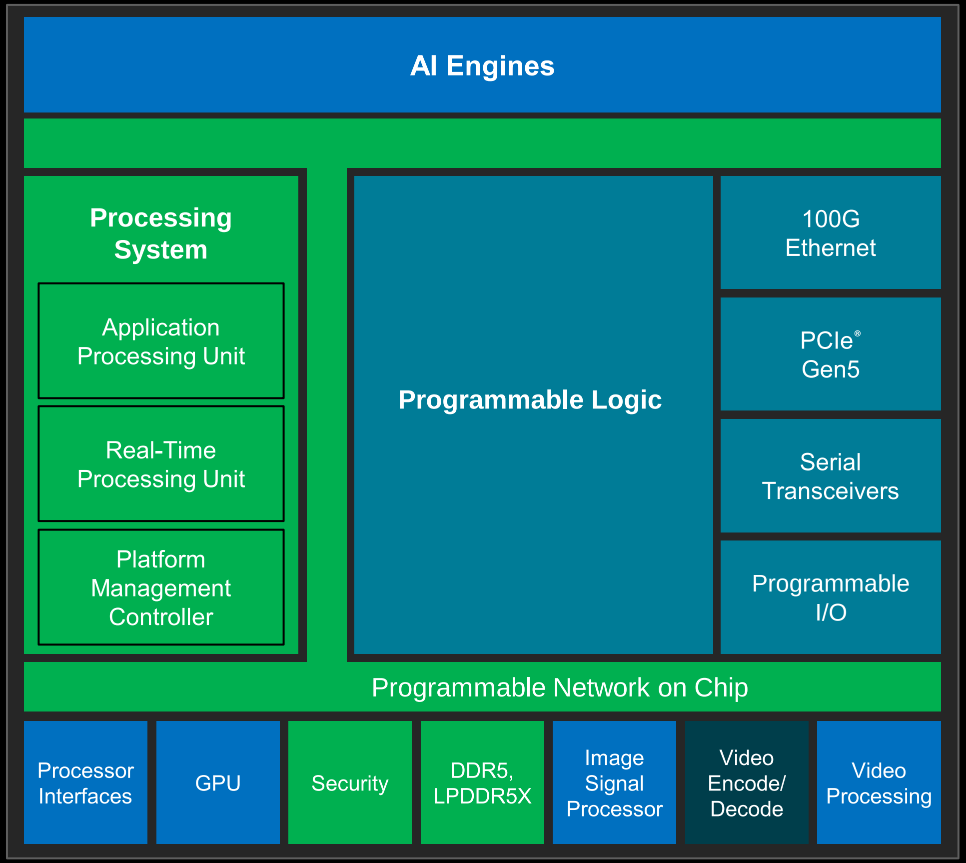
Including Application/Real Time/Video and Acceleration Channels

Video/Acceleration channel

ASIL B/SIL2 Random Hardware Fault Integrity

Application/Real Time channels

ASIL D/SIL3 Random Hardware Fault Integrity



ISO 26262
Automotive Safety

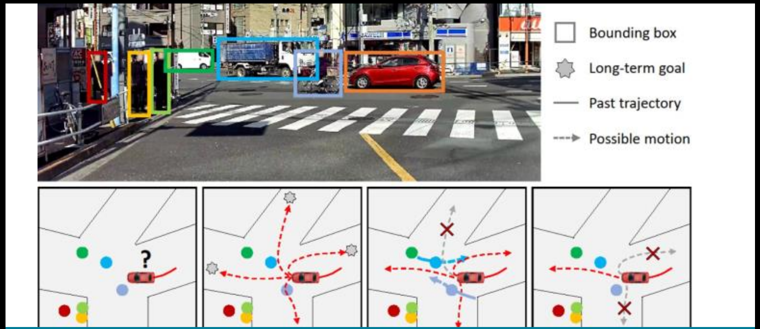
IEC 61508
Functional Safety

ISO 13849
Machine Safety

Automotive AI and Vision Applications

DMS/OMS

- Face recognition
- Eye gaze
- Pose estimation
- Hand gesture
- Health monitoring



Exterior Imaging

- Surround view monitoring
- Image enhancement
- Object detection
- Perception

Productivity

- Smart Assistant
- Video conferencing
- Face detection/tracking
- Background blur/replacement

Why Adaptable SoCs

Enables Differentiation

Ideal for emerging applications

OMS, LiDAR, 4D Imaging Radar

Futureproofing

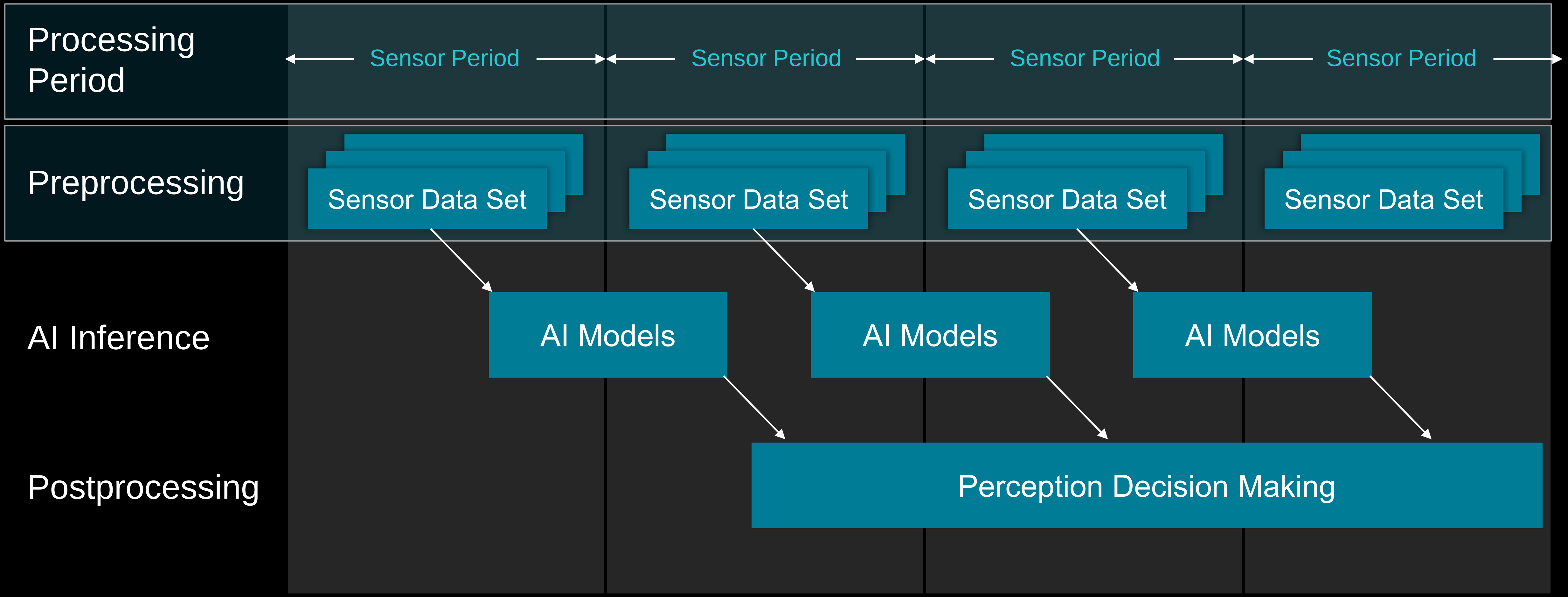
Changes in Sensors

Transition from CV to AI

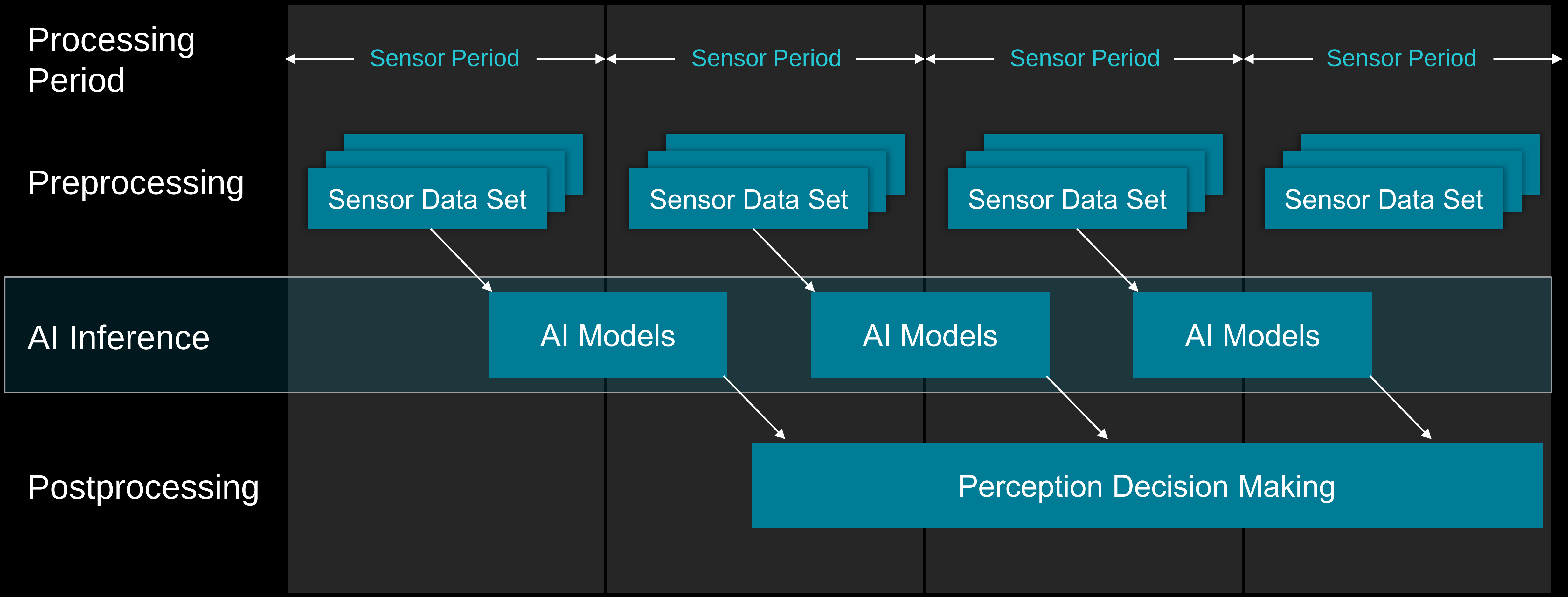
Low Latency Processing

Parallelization and/or Isolation of critical processing pipelines

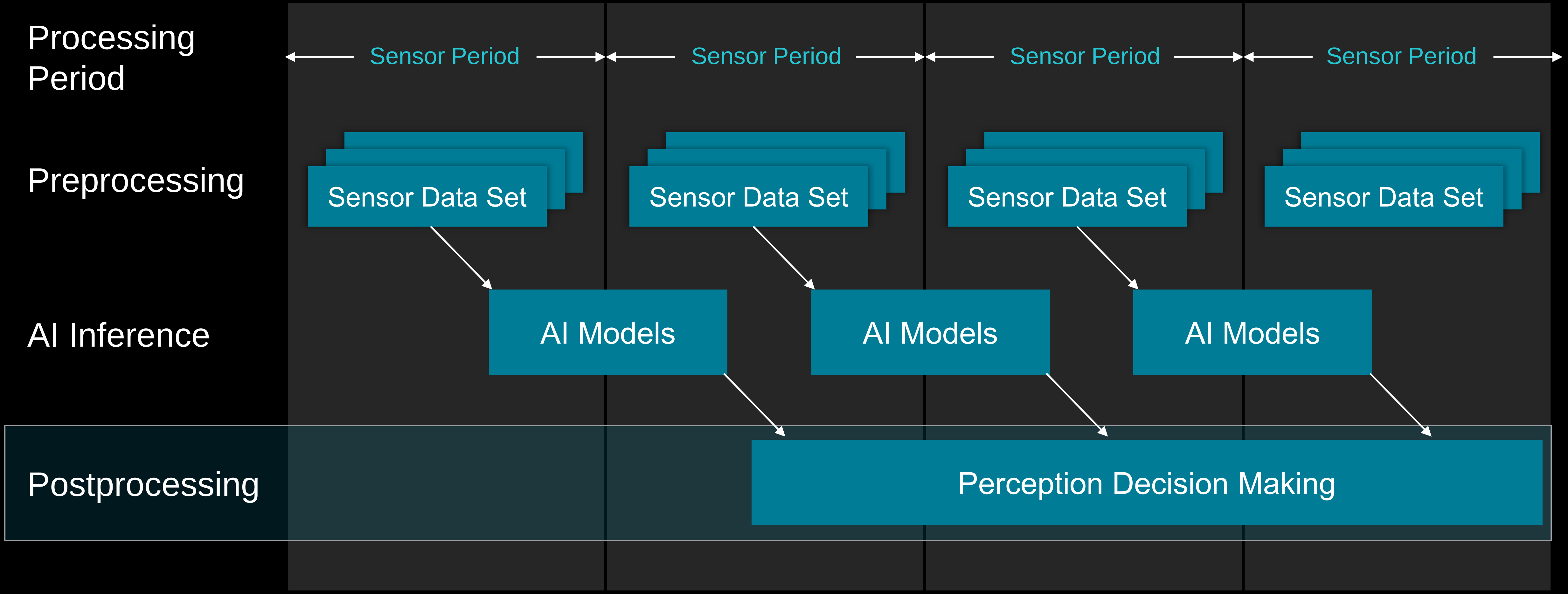
AI And Vision Processing Pipeline



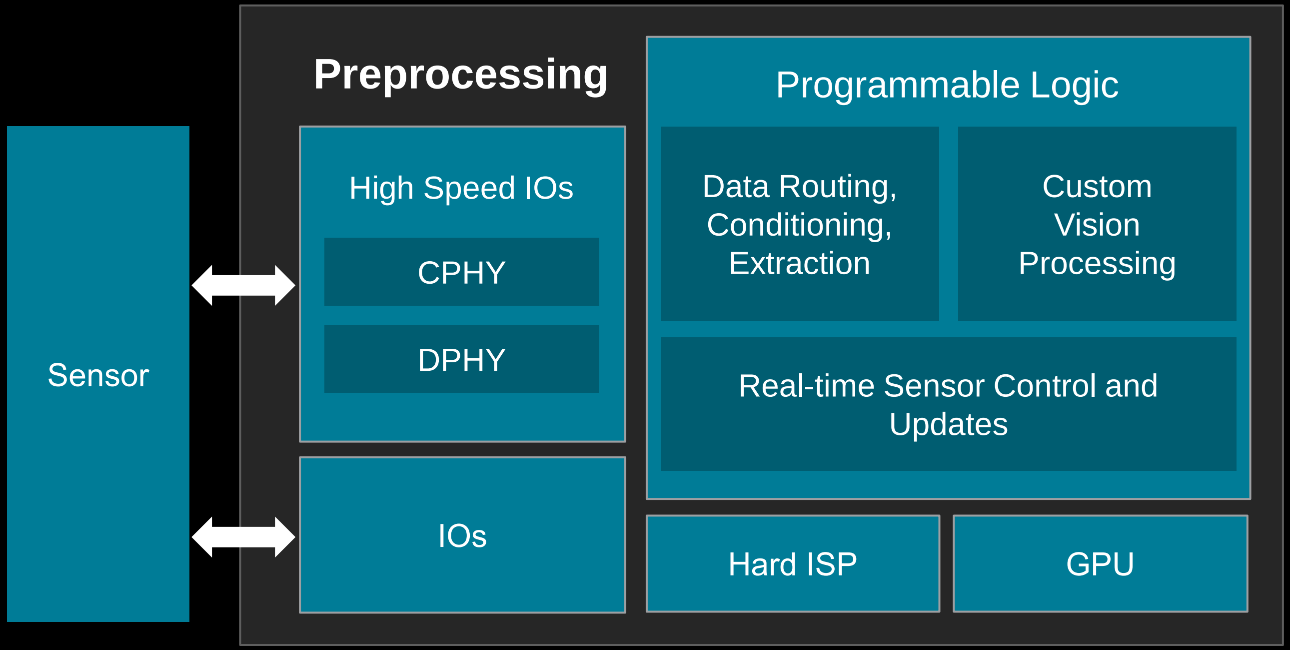
AI And Vision Processing Pipeline



AI And Vision Processing Pipeline



Sensor Preprocessing

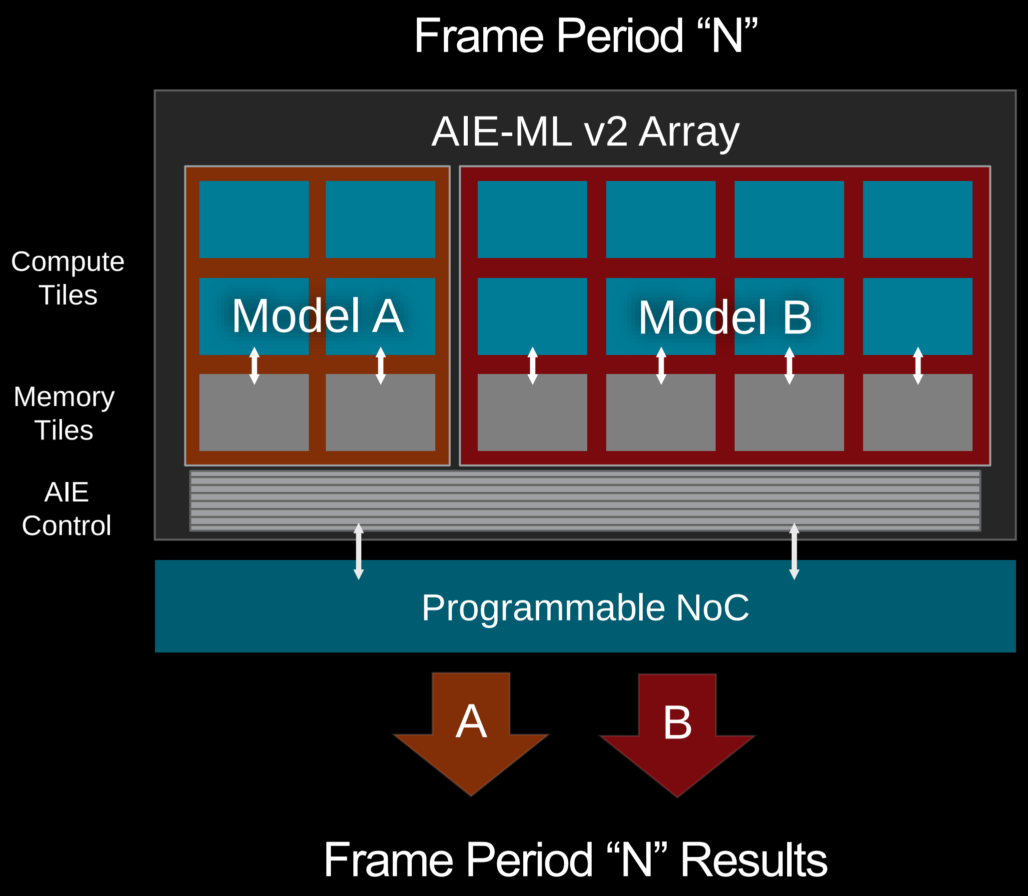


Function	Adaptive SoC	Other SoCs or ASICs
Configurable Physical Interfaces	HW/SW Customizable	Fixed
Low Latency Control & Synchronization	HW/SW Customizable	Fixed or SW
Vision Pipeline	HW/SW Customizable	Fixed
Sensor Fusion	HW/SW Customizable	SW Only

Programmable Logic and IO enable wide range of user adaptability in Hardware

Inference - Spatial Sharing

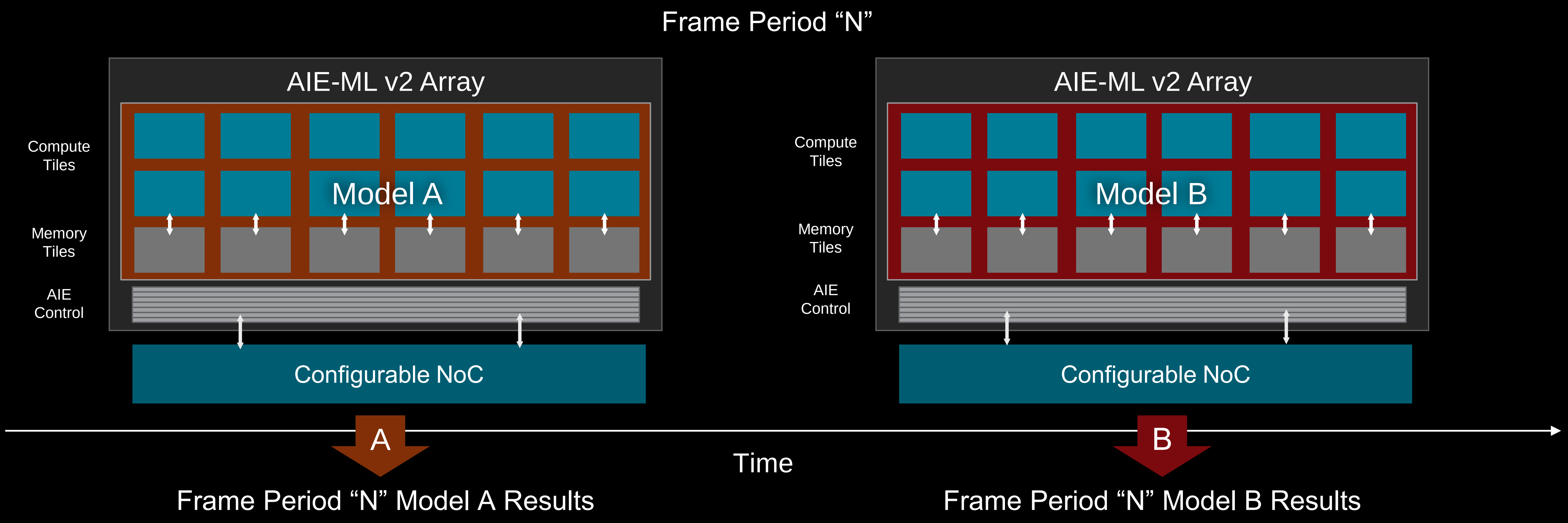
Run Models Concurrently within the AIE-ML array



Inference – Temporal Sharing

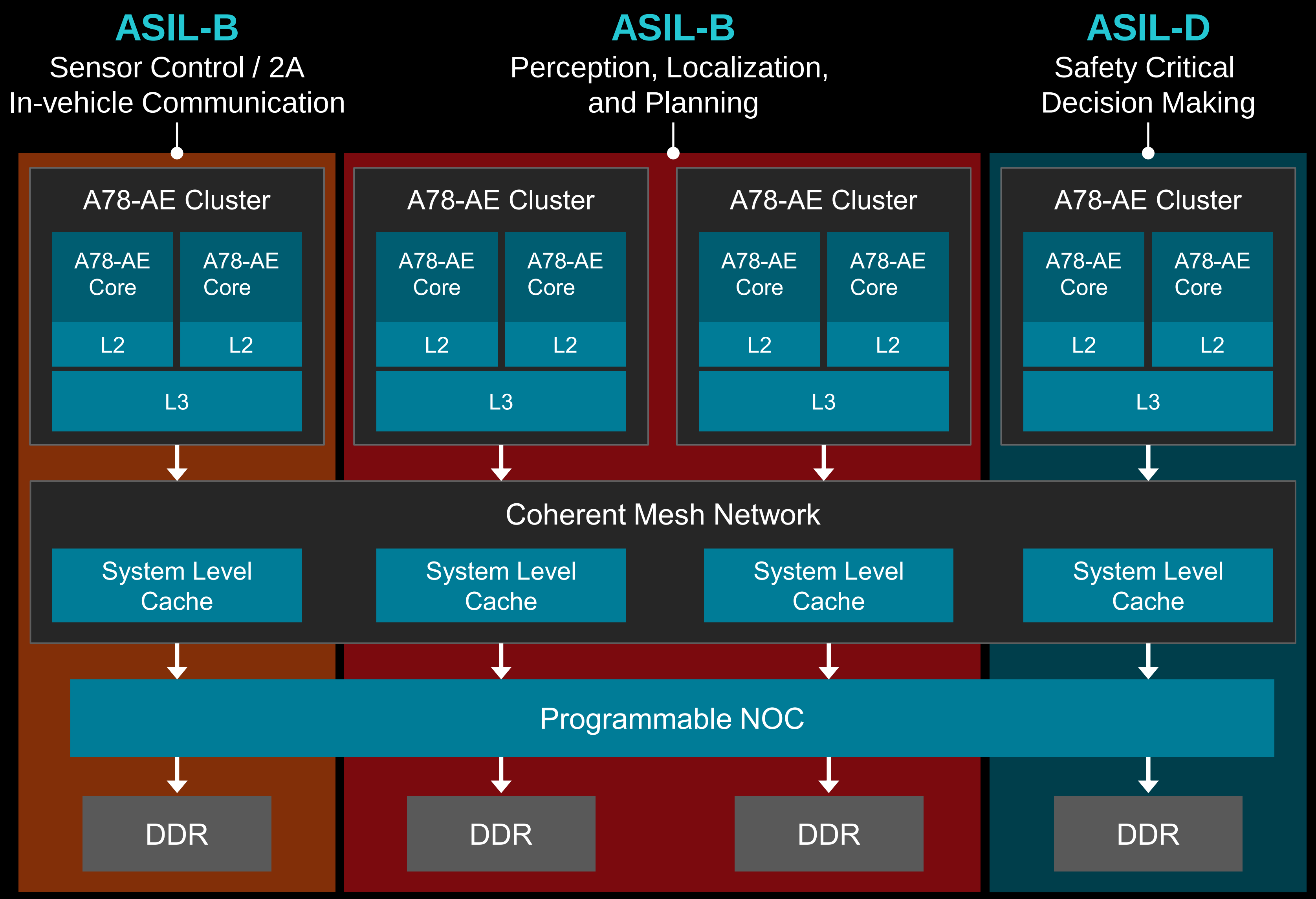
Alternatively, multiple AI model workloads can timeshare AIE-ML

- Context Switching between multiple AI Models
- Enable prioritization of order of AI Model results for post processing

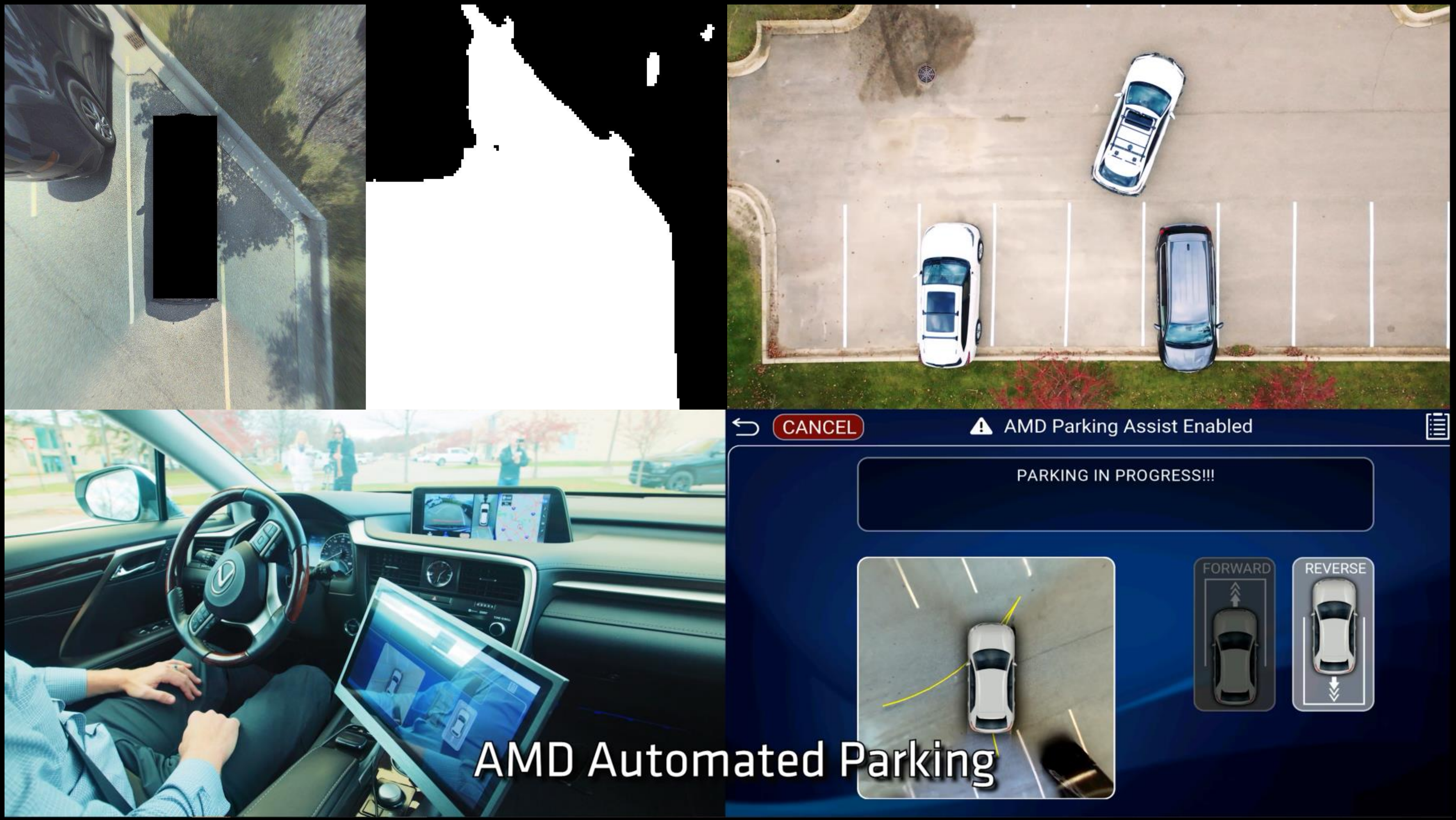


Postprocessing

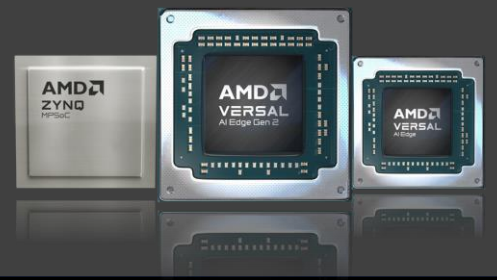
Processing cluster resources configurable based on application specific needs



Automated Parking

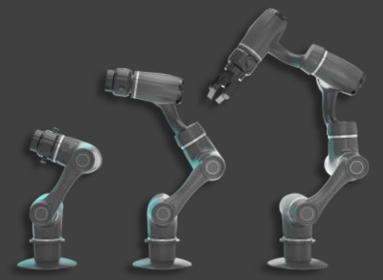


AMD Versal AI Edge Series Gen 2 –
Adaptable and Scalable for Vision and Automotive



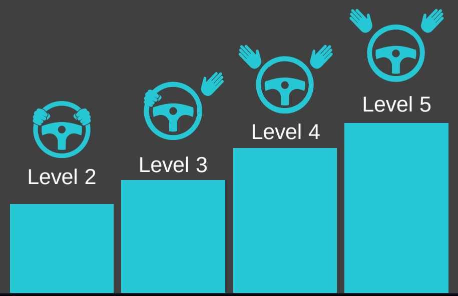
End-to-end acceleration of AI-driven embedded systems from a single-chip

Highly integrated device with hardened compute accelerators, programmable logic, and built-in high reliability and safety



Single Architecture Enabling Cost Efficient to High Performance Processing

Design once and scale performance needs with the same tools, SW, ecosystem, and safety certification



Optimized for User Configurability in Vision and Automotive Applications

Customize for user application needs, considering total compute, custom processing and functional safety

Endnotes

VER-023

Based on AMD internal pre-silicon performance estimates and power projections for the AIE-ML v2 compute tile architecture featured in the Versal AI Edge series Gen 2 using the MX6 data type compared to production performance specifications and AMD Power Design Manager power estimates for the AIE-ML compute tile architecture featured in the first generation Versal AI Edge series using the INT8 data type. Operating conditions: 1 GHz Fmax, 0.7V AIE operating voltage, 85C junction temperature, typical process, 60% vector load, % activations = 0 < 10%. Actual performance will vary when final products are released in market. Performance projections as of February 2024.

VER-025

Based on AMD internal pre-silicon performance estimates and power projections for the AIE-ML v2 compute tile architecture featured in the Versal AI Edge series Gen 2 for the MX6 data type compared to the INT8 data type. Operating conditions: 1 GHz Fmax, 0.7V AIE operating voltage, 85C junction temperature, typical process, 60% vector load, % activations = 0 < 10%. Actual performance will vary when final products are released in market.

VER-026

Based on AMD internal pre-silicon performance estimates for the AIE-ML v2 compute tile architecture featured in the Versal AI Edge series Gen 2 using the MX6 data type compared to the INT8 data type. Operating conditions: 1 GHz Fmax, 0.7V AIE operating voltage. Actual performance will vary when final products are released in market.

VER-027

Based on AMD internal pre-silicon performance estimates for combined total DMIPs of the Versal AI Edge series Gen 2 and Versal Prime series Gen 2 processing system when configured with 8 Arm Cortex-A78AE applications cores @2.2 GHz & 10 Arm Cortex-R52 real-time cores @1.05 GHz, compared to the published combined total DMIPs of the processing system in the first generation Versal AI Edge series and Versal Prime series. Versal AI Edge series Gen 2 and Prime series Gen 2 operating conditions: highest available speed grade, 0.88V PS operating voltage, split-mode operation, maximum supported operating frequency. First generation Versal AI Edge series and Prime series operating conditions: highest available speed grade, 0.88V PS operating voltage, maximum supported operating frequency. Actual DMIPs performance will vary when final products are released in market.

VER-030

Based on Arm product specifications for a Versal AI Edge series Gen 2 and a Versal Prime series Gen 2 configured with a 4 core Mali-G78AE GPU with a maximum operating frequency of 1050 MHz, 64 FP32 per ops/clock/core, and 4 texels per ops/clock/core. Actual Versal product performance will vary when final products are released in market.

Endnotes

VER-69

Based on AMD internal accuracy testing in July 2024, AMD evaluated the MX6 datatype as drop-in replacement across a subset of market-relevant FP32 AI models (25/44) in various categories (speech, automotive, CNN and vision).Configuration for internal accuracy testing: AMD EPYC 73F3 CPU with 8x AMD Instinct MI250 GPUs. OS ver: Ubuntu-20.04. ML Env: Pytorch v2.1.2, ROCm 5.6.0.50600-76~20.04. Results may vary and are based on several factors, including design, device, configuration, AI model, and ML software.

VER-70

Based on AMD internal accuracy testing in July 2024, AMD evaluated the MX9 datatype as a drop-in replacement across a subset of market-relevant FP32 AI models (35/44) in various categories (speech, automotive, CNN and vision).Configuration for internal accuracy testing: AMD EPYC 73F3 CPU with 8x AMD Instinct MI250 GPUs. OS ver: Ubuntu-20.04. ML Env: Pytorch v2.1.2, ROCm 5.6.0.50600-76~20.04. Results may vary and are based on several factors, including design, device, configuration, AI model, and ML software.

Disclaimer and Attributions

DISCLAIMER
Timelines, roadmaps, and/or product release dates shown in these slides are plans only and subject to change.

The information contained herein is for informational purposes only and is subject to change without notice. While every precaution has been taken in the preparation of this document, it may contain technical inaccuracies, omissions and typographical errors, and AMD is under no obligation to update or otherwise correct this information. Advanced Micro Devices, Inc. makes no representations or warranties with respect to the accuracy or completeness of the contents of this document, and assumes no liability of any kind, including the implied warranties of noninfringement, merchantability or fitness for particular purposes, with respect to the operation or use of AMD hardware, software or other products described herein. No license, including implied or arising by estoppel, to any intellectual property rights is granted by this document. Terms and limitations applicable to the purchase or use of AMD's products are as set forth in a signed agreement between the parties or in AMD's Standard Terms and Conditions of Sale. GD-18

AMD