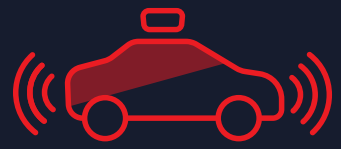
XILINX

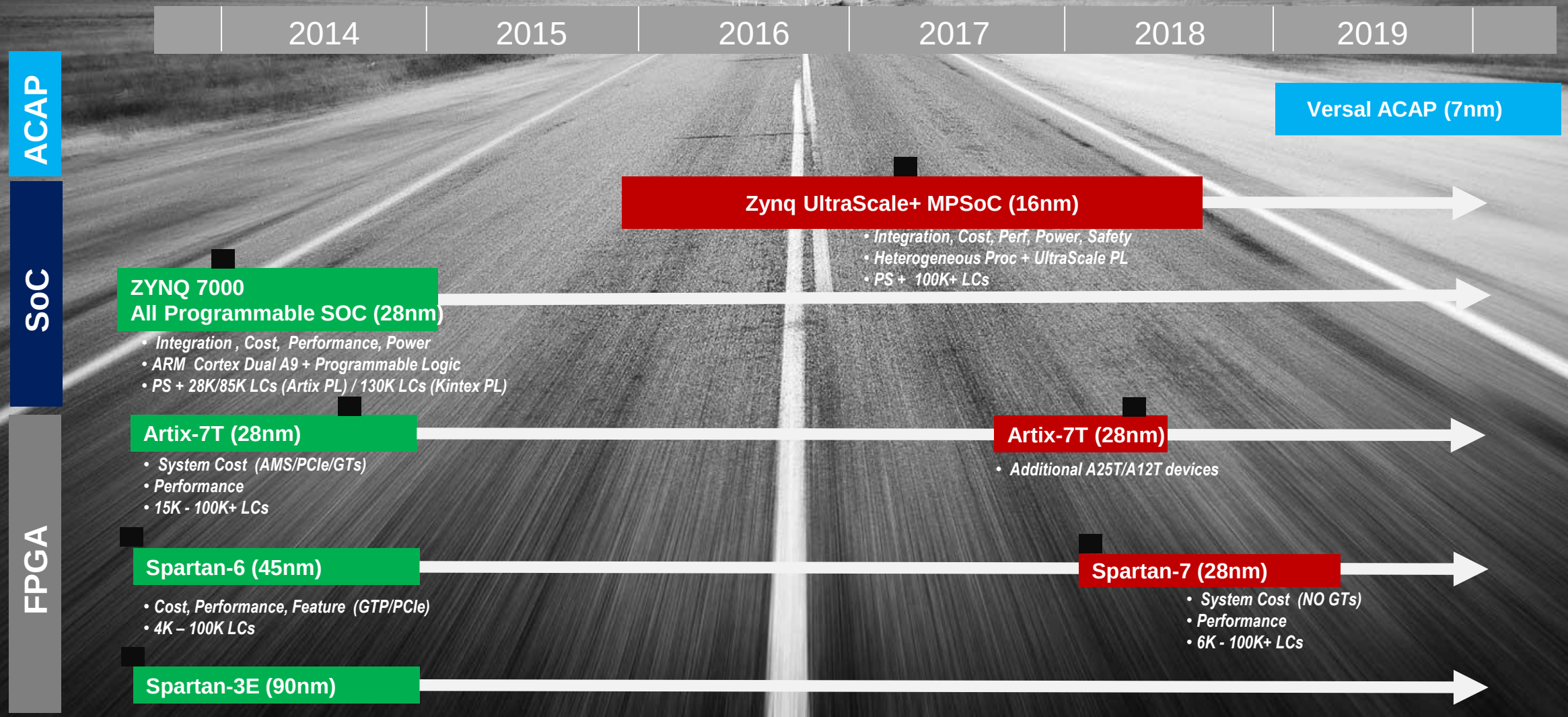
Xilinx Automotive & Versal Roadmap



March 2019

XILINX.

Xilinx Automotive (XA) Silicon Roadmap



■ 1st XA Prod

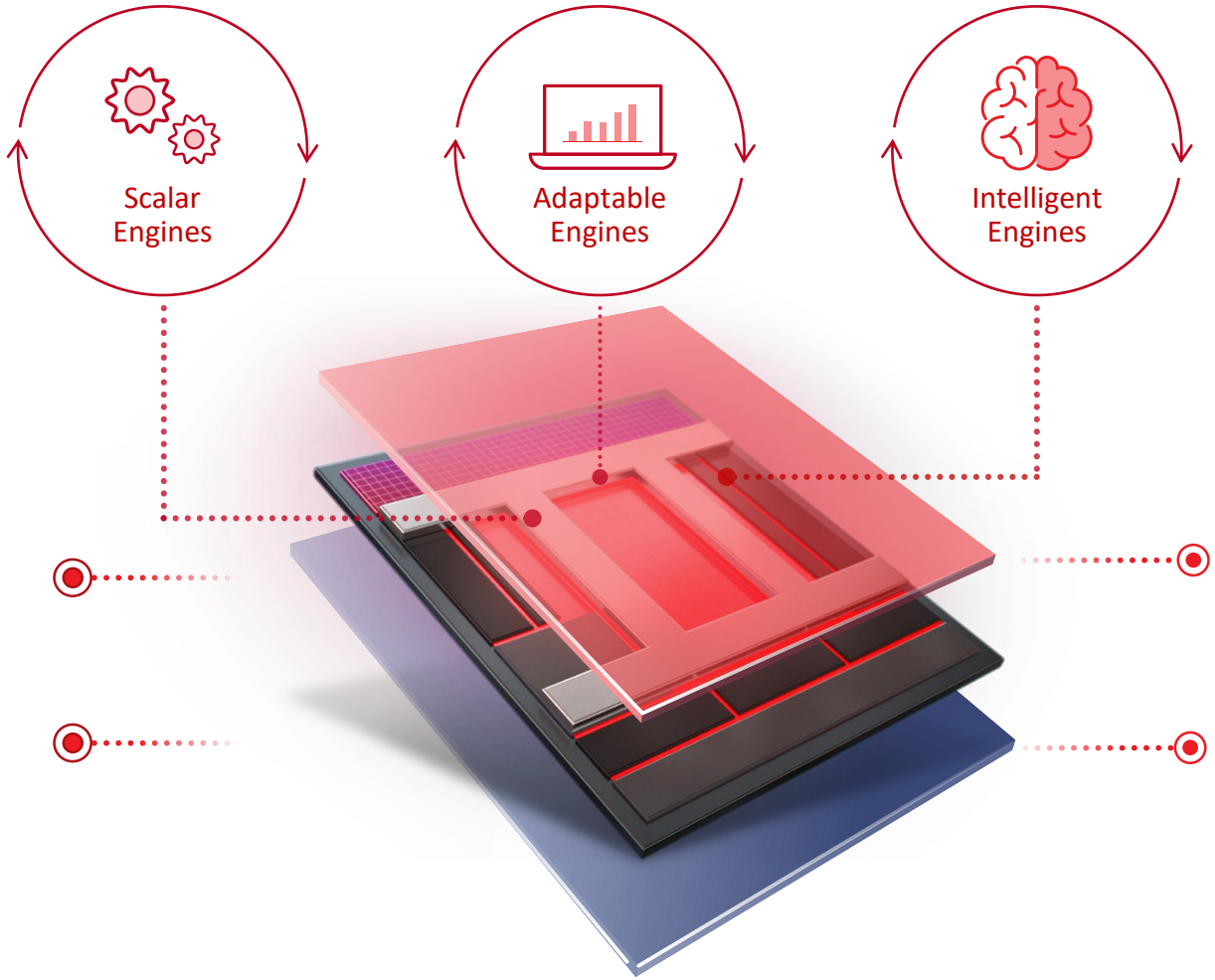
XILINX.

Agenda

- Introducing Versal: The First ACAP
- Heterogeneous Acceleration Engines
- Key Architectural Blocks
- Application Use Case
- Software Solution for Any Developer
- Product Portfolio
- Product Tables

New Device Category: Adaptive Compute Acceleration Platform

COMPUTE ACCELERATION



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

XILINX.

Introducing the World's First ACAP

XILINX
VERSAL™

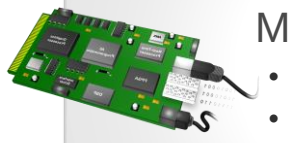
- Heterogeneous Acceleration
- For Any Application
- For Any Developer

Versal Architecture Overview

Adaptable Engines
2X compute density

Scalar Engines
- Platform Control
- Edge Compute

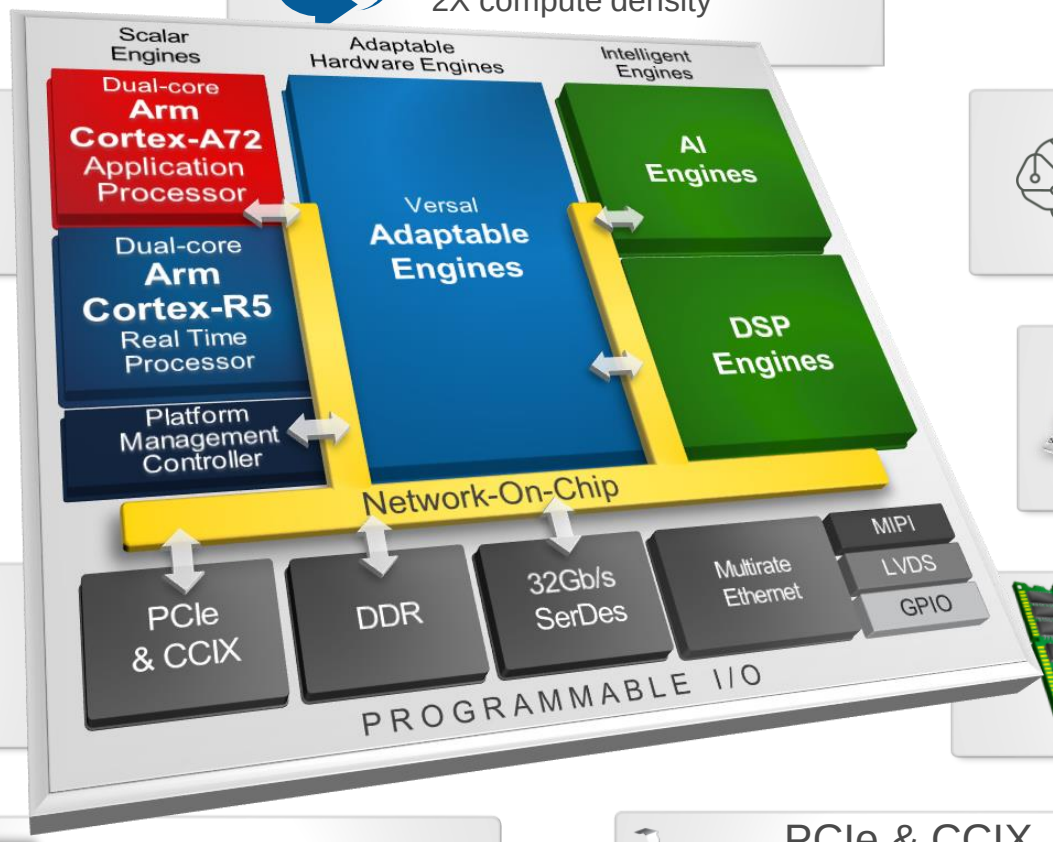
Intelligent Engines
- AI Compute
- Diverse DSP workloads

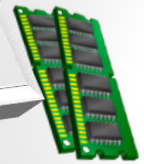


Multi-Rate Ethernet
- From 100GE to 10G ports
- Integrated error correction

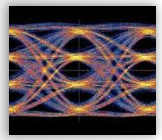
Network-on-Chip
- Guaranteed Bandwidth
- Enables SW Programmability

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





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



Transceivers
- Up to 32Gb/s
- Up to 44 SerDes blocks

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

XILINX.

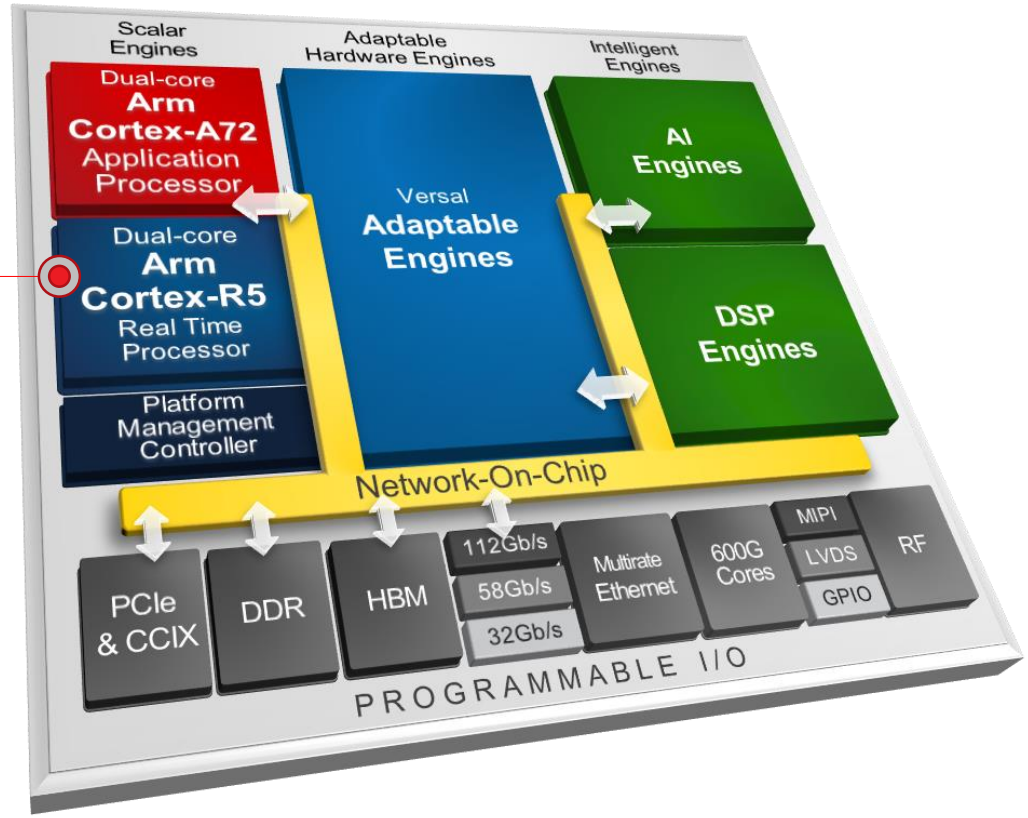
Scalar Engines

Scalar Engines

Complex Algorithms and
Decision Making for Autonomous Systems

Safety Processing and Redundancy
for Mission- and Safety-Critical

Management and Backbone that
Programs Entire Platform



Platform Management Controller

Bringing the Platform to Life & Keeping it Safe & Secure

Boot & Configuration

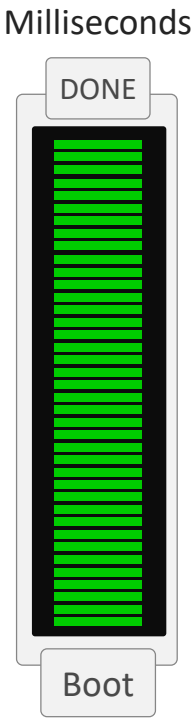
- Boots the platform in milliseconds (any engine first)
- 8X faster dynamic reconfiguration
- Advanced power & thermal management

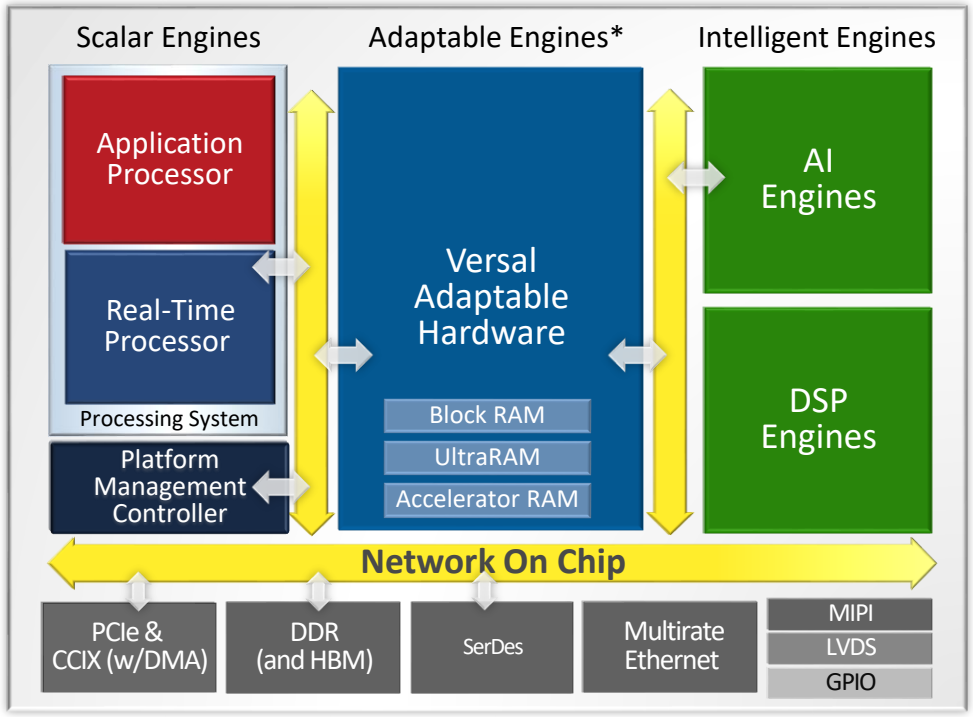
Security, Safety & Reliability Enclave

- HW Root of Trust
- Cryptographic acceleration & confidentiality
- Enhanced diagnostics, system monitoring & anti-tamper
- Error mitigation, detection & management for safety

Integrated Platform Interfaces & High Speed Debug

- Integrated flash, system & debug interfaces
- High-speed non-invasive, chip-wide debug





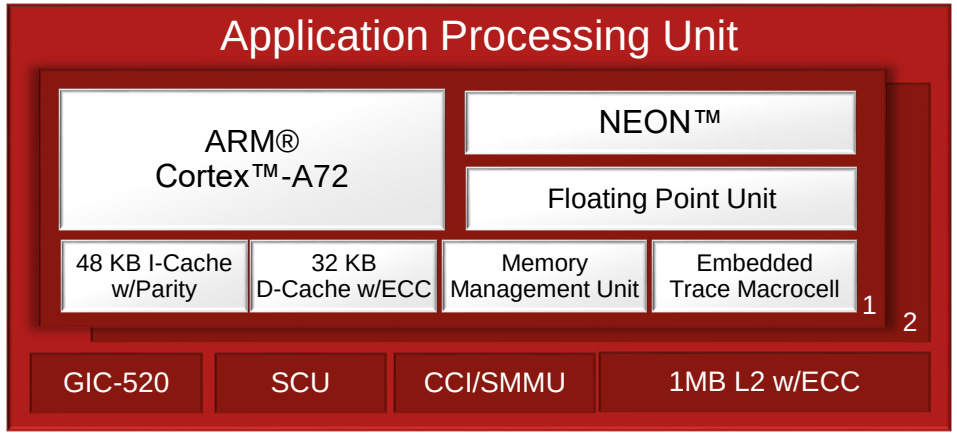
BOOT & CONFIG • SAFETY • SECURITY • DEBUG

XILINX

The Arm Subsystem

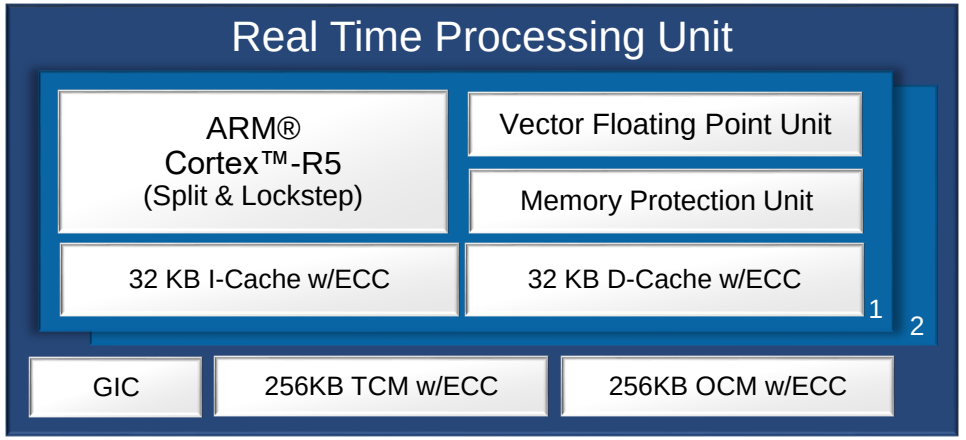
Dual-Core ARM Cortex-A72 Application Processors

- Up to 1.7GHz for 2X single-threaded performance[1]
- Cost and power optimized (half the power)
- Code compatibility (ARMv8-A architecture)
- Enables SW developers to start from a familiar place



Dual-Core ARM Cortex-R5 Real Processors

- Up to 750MHz for 1.4X greater performance[1]
- Low latency and deterministic
- Flexible operation modes: Split-Mode and Lock-Step
- Highest levels of functional safety (ASIL and SIL)



1: DMIPS vs. Zynq UltraScale+ MPSoCs

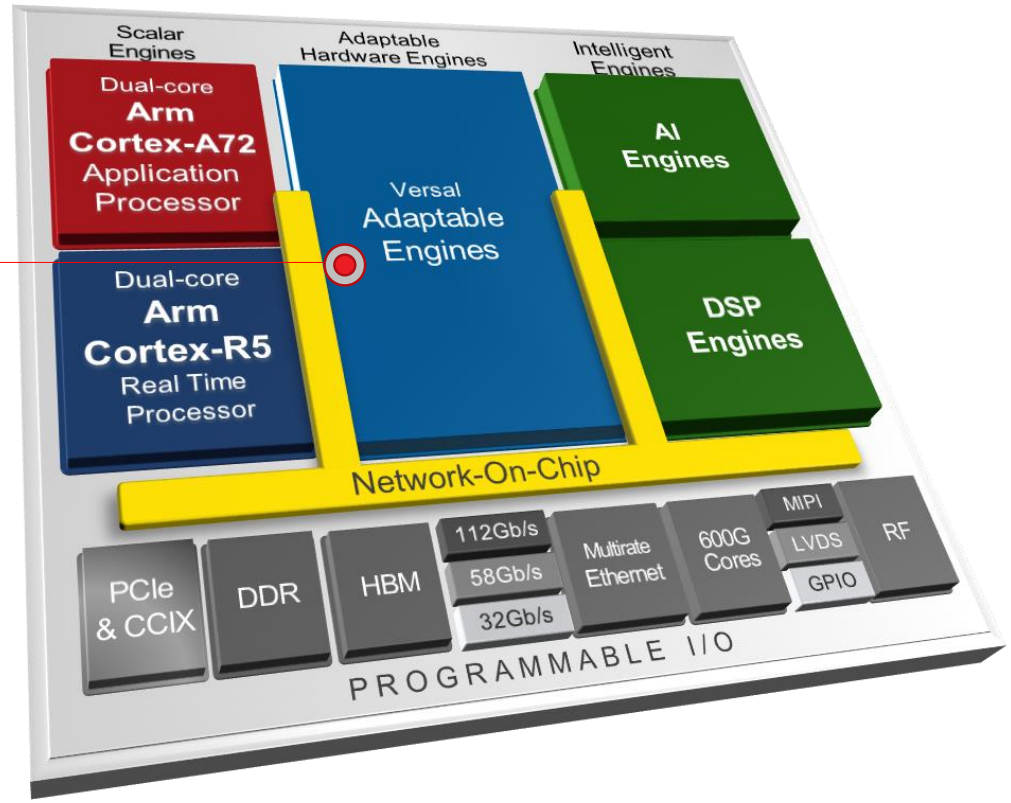
Adaptable Engines

Adaptable Hardware Engines

Programmable logic for fine-grained parallel processing, data aggregation, and sensor fusion

Programmable memory hierarchy to optimize compute efficiency

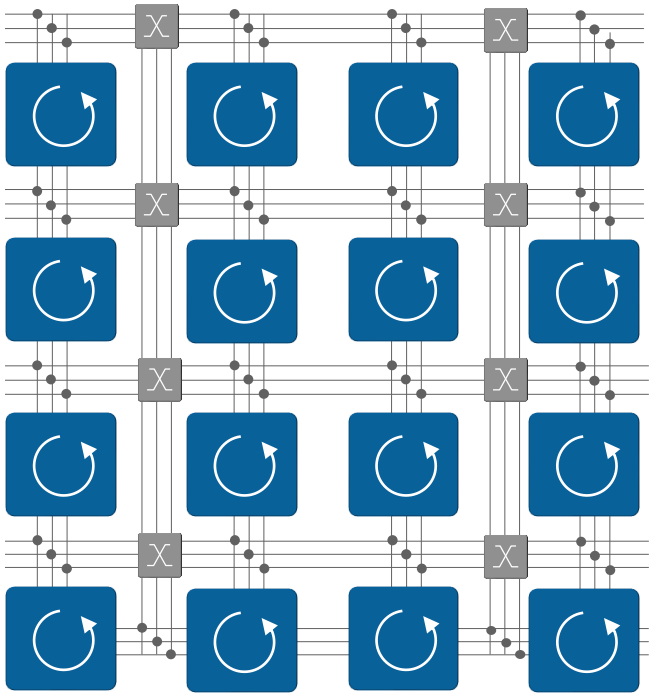
High bandwidth, low latency data movement between engines and I/O



Greater Compute Density for Any Workload

Re-Architected Hardware Fabric

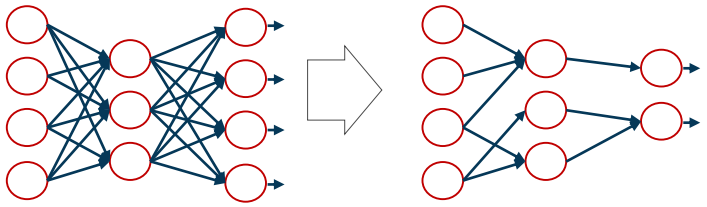
- 4X density per logic block for more compute
- Less external routing→ greater performance
- Code and IP compatible with 16nm devices



Adaptable to any Workload

- Bit-level precision (1 → 1,000) for any algorithm
- Improves ML efficiency (compression, pruning)
- Forward-compatible to lower precision neural networks, e.g., BNN

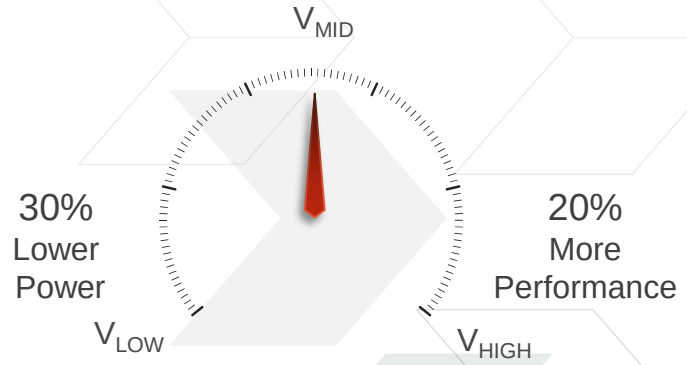
ML Inference and Optimizations (e.g., pruning)



For Any Workload, e.g., …

Tune for Power & Performance

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



XILINX

Intelligent Engines

Intelligent Engines for Diverse Compute

DSP Engines

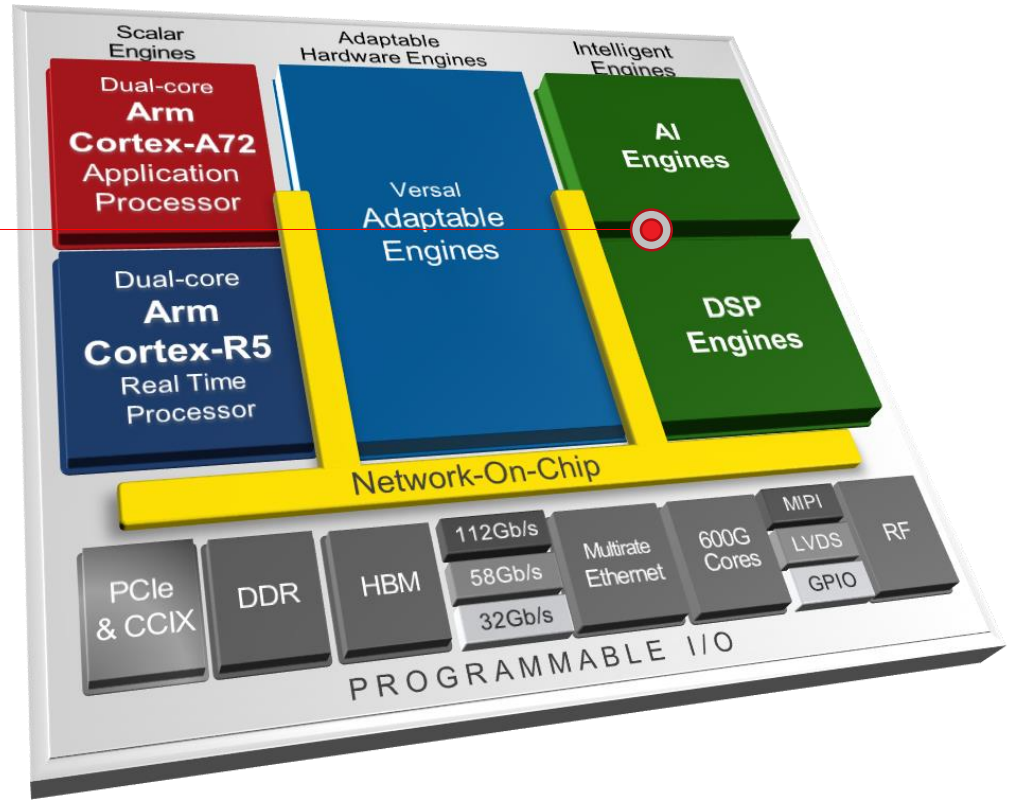
High-precision floating point & low latency

Granular control for customized data paths

AI Engines

High throughput, low latency, and power efficient

Ideal for AI inference and advanced signal processing



DSP Engines

Versatility and Granular Control of Data Path

Enhanced Compute architecture

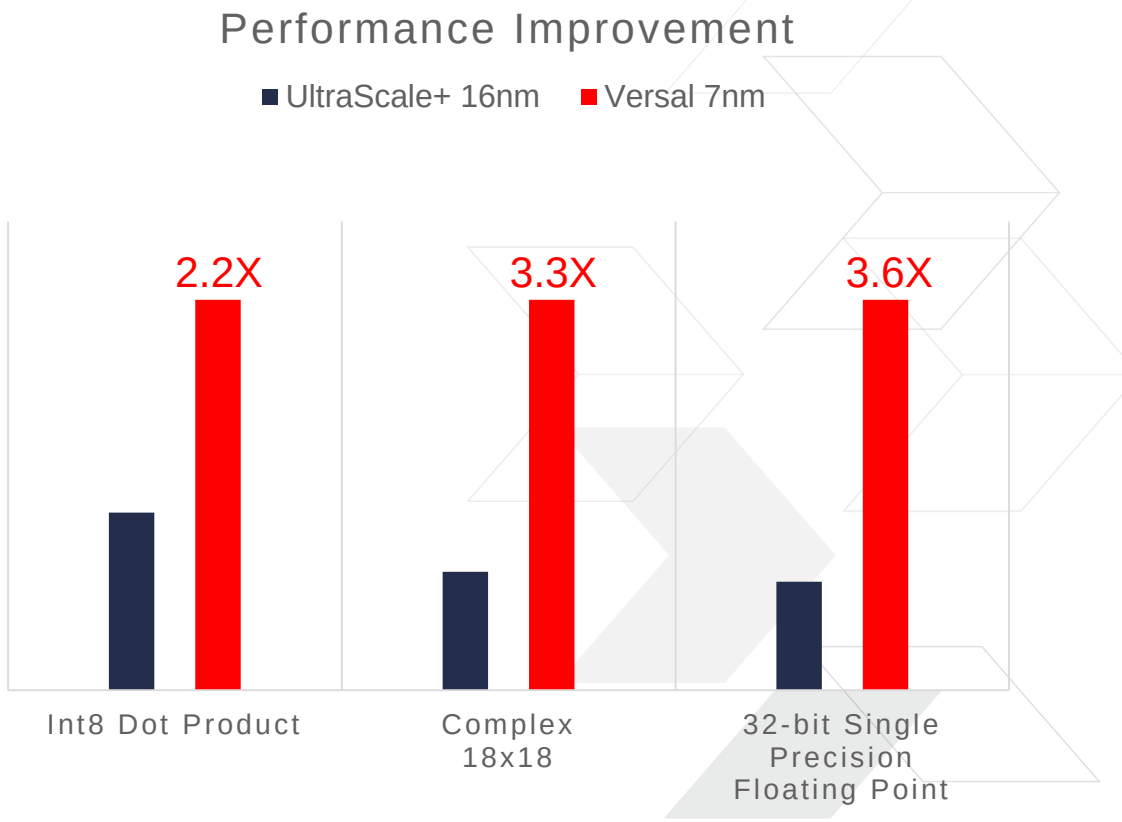
- Greater than 1GHz of performance

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



XILINX.

AI Engines

Massive AI Inference Throughput and Wireless Compute

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

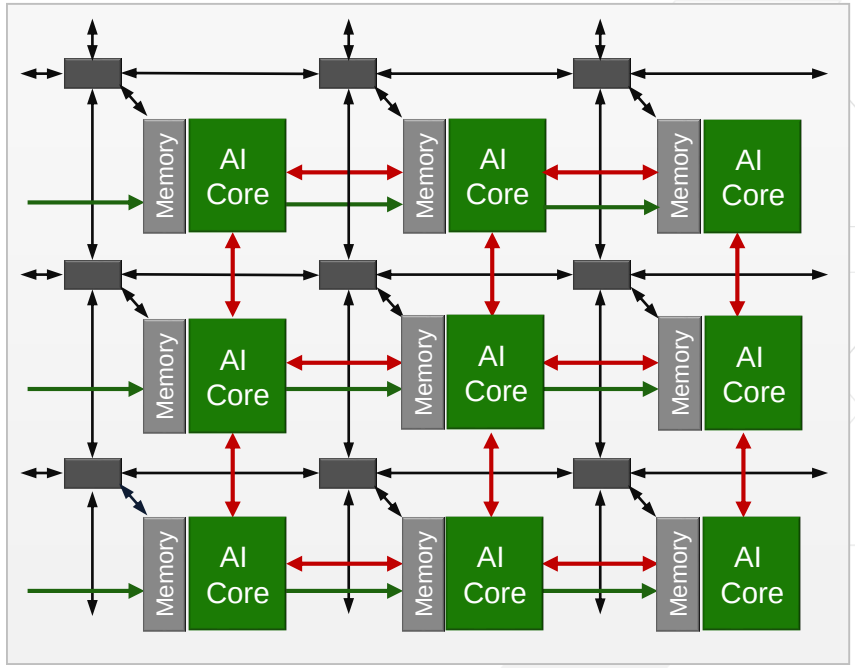
> Direct, massive throughput to adaptable HW engines

> Implement core application with AI for “Whole App Acceleration”

SW programmable for any developer

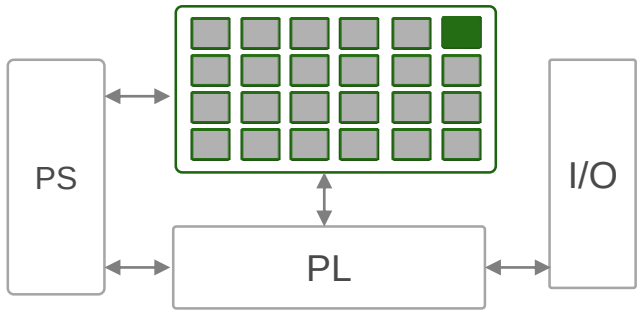
> C programmable, compile in minutes

> Library-based design for ML framework developers

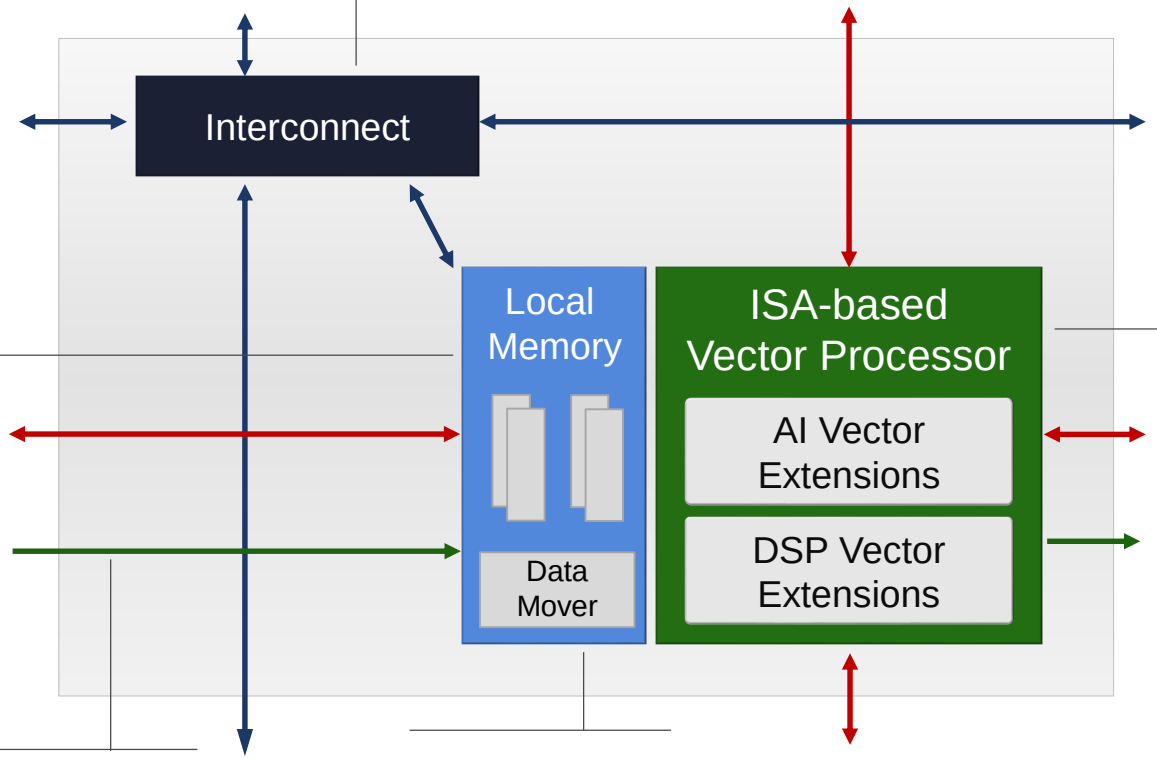


XILINX.

AI Engine: Tile-Based 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++)

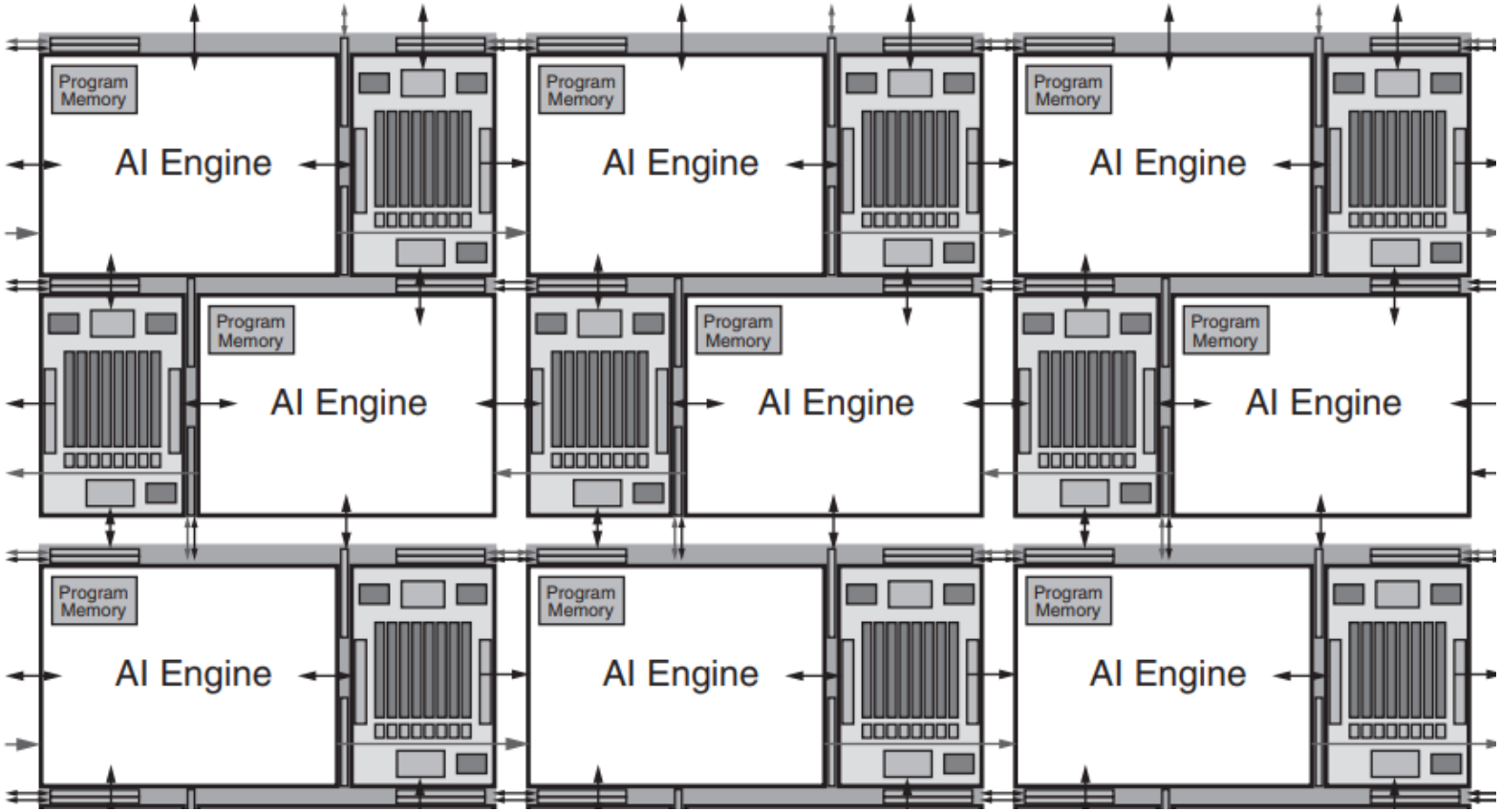
Cascade Interface
Partial results to next core

Data Mover
Non-neighbor data communication
Integrated synchronization primitives

XILINX

AI Engine Array (3x3 Example)

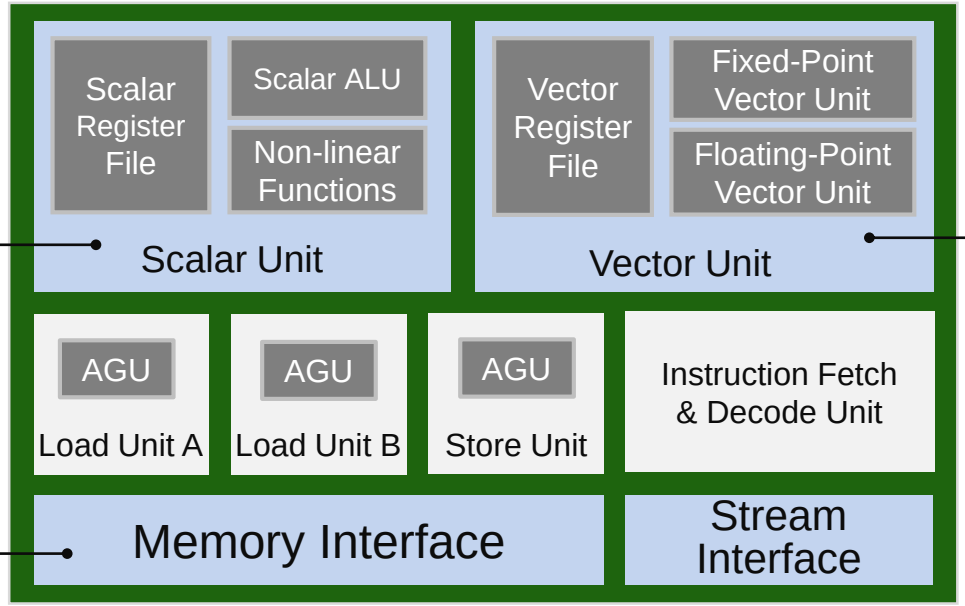
- 128KB Addressable Data Memory (32KB * 4 neighboring memory tiles)
- 16KB Program Memory per AI Engine



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

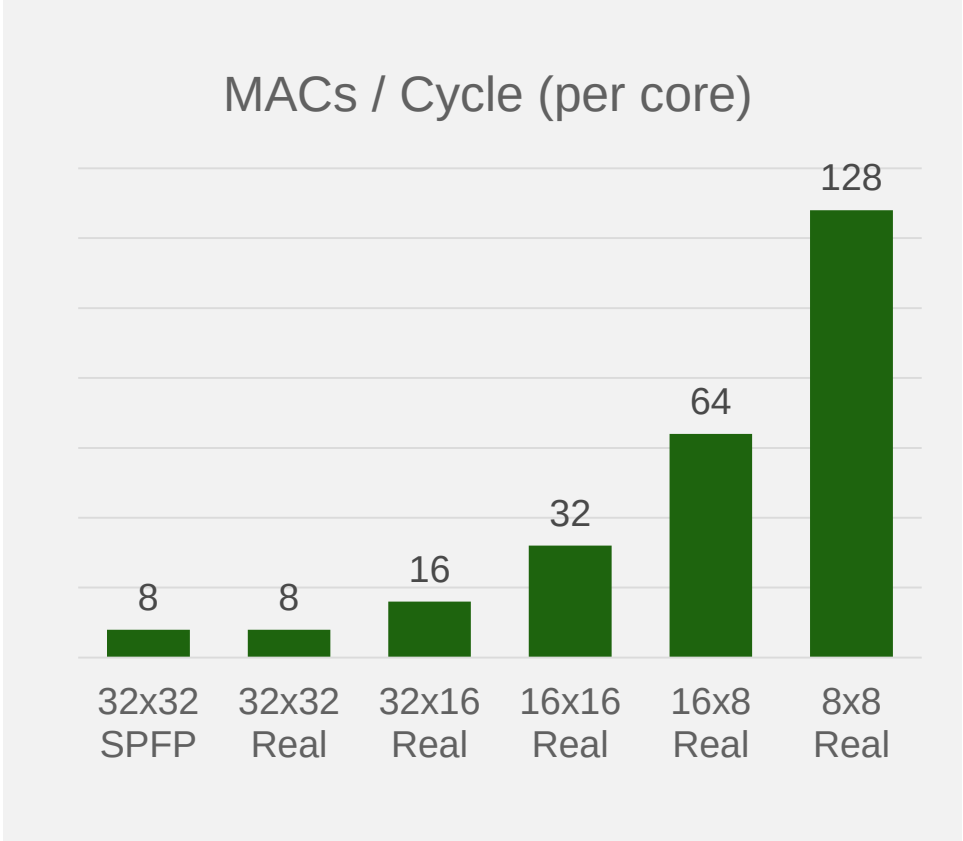
Multiple vector lanes
- Vector Datapath
- 8 / 16 / 32-bit & SPFP operands

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

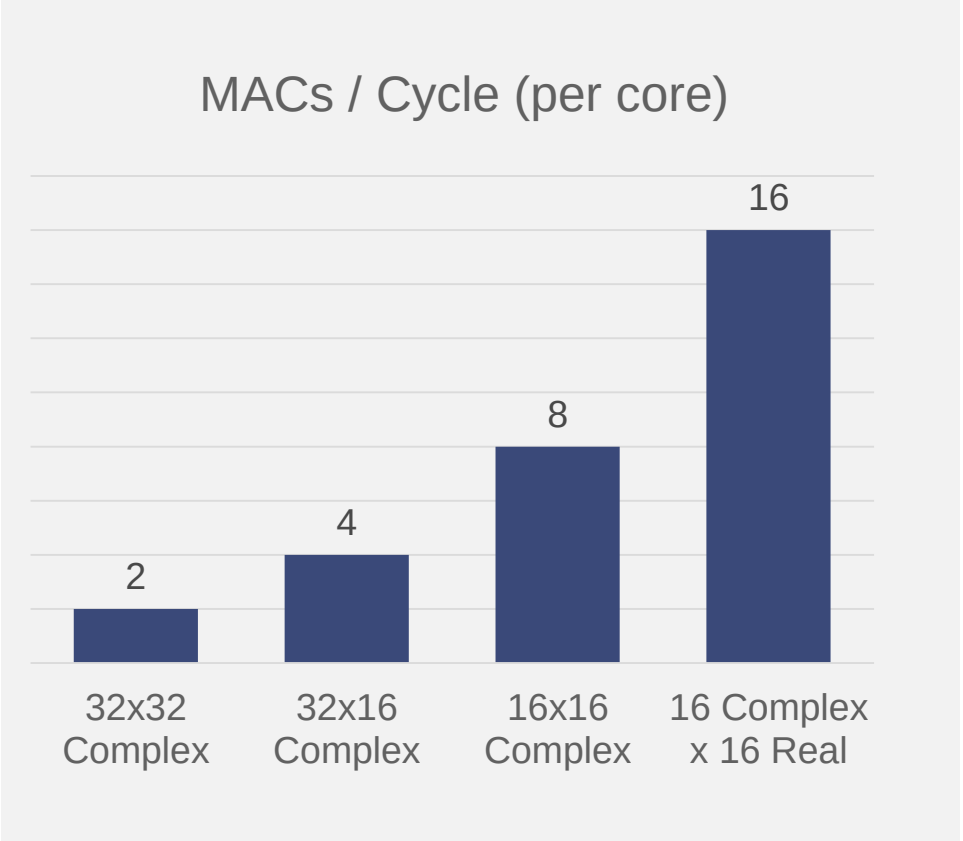
XILINX.

Multi-Precision Support

AI Data Types



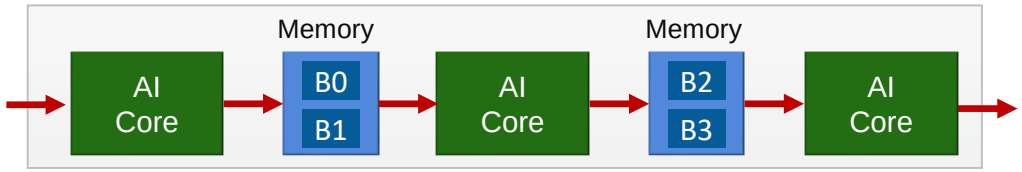
Signal Processing Data Types



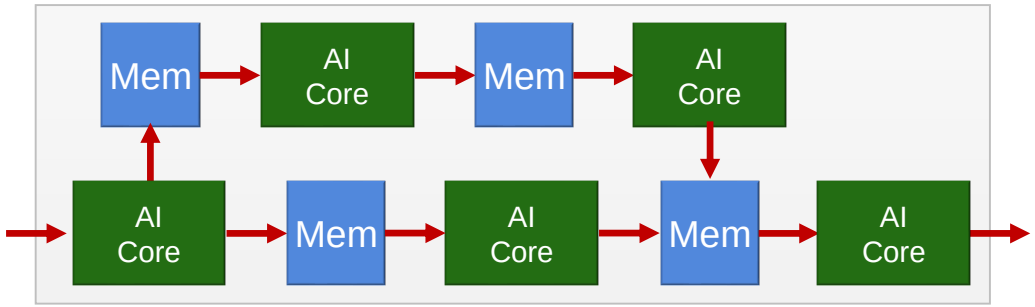
Data Movement Architecture

Memory Communication

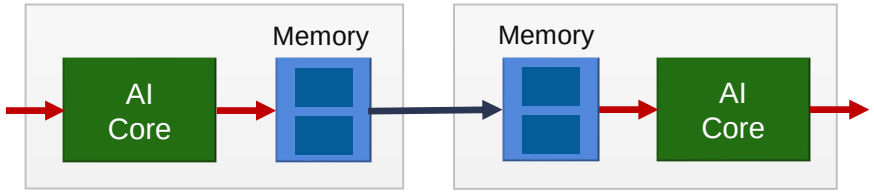
Dataflow Pipeline



Dataflow Graph

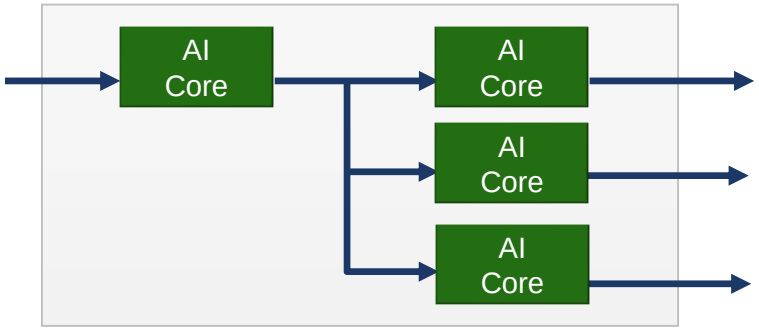


Non-Neighbor

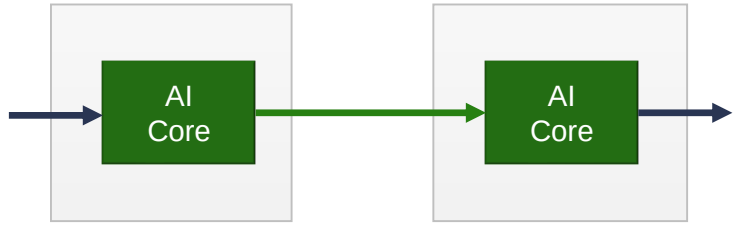


Streaming Communication

Streaming Multicast



Cascade Streaming



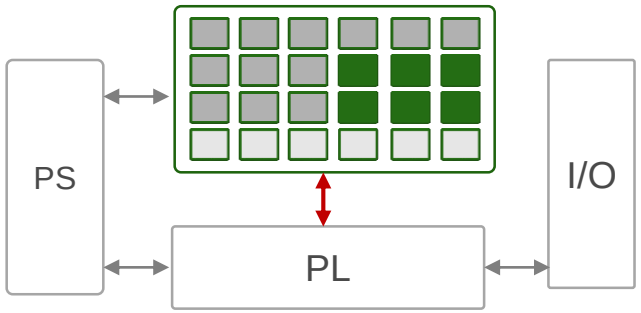
Memory Interface

Stream Interface

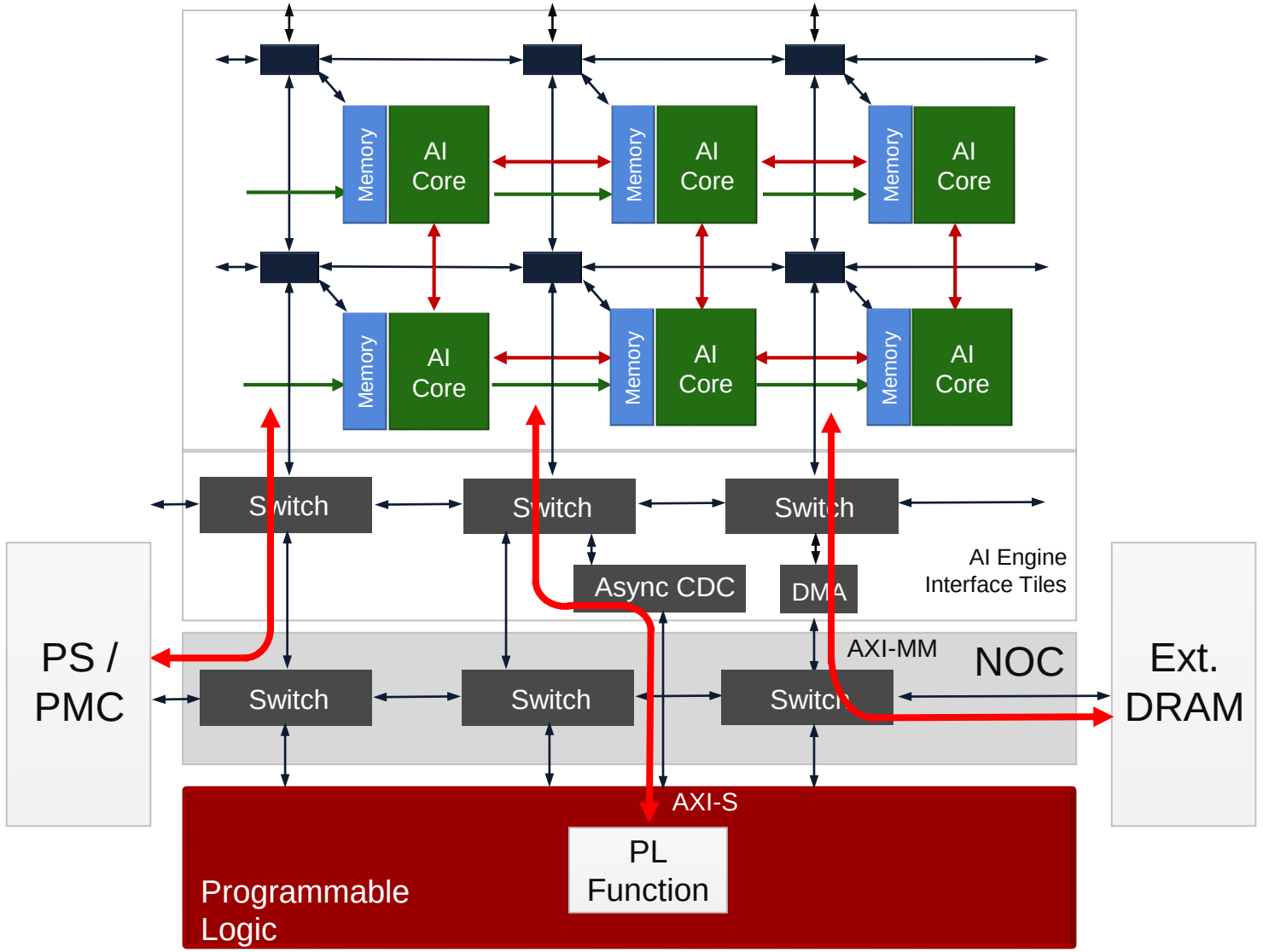
Cascade Interface

XILINX

AI Engine Integration with Versal ACAP



- **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.

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

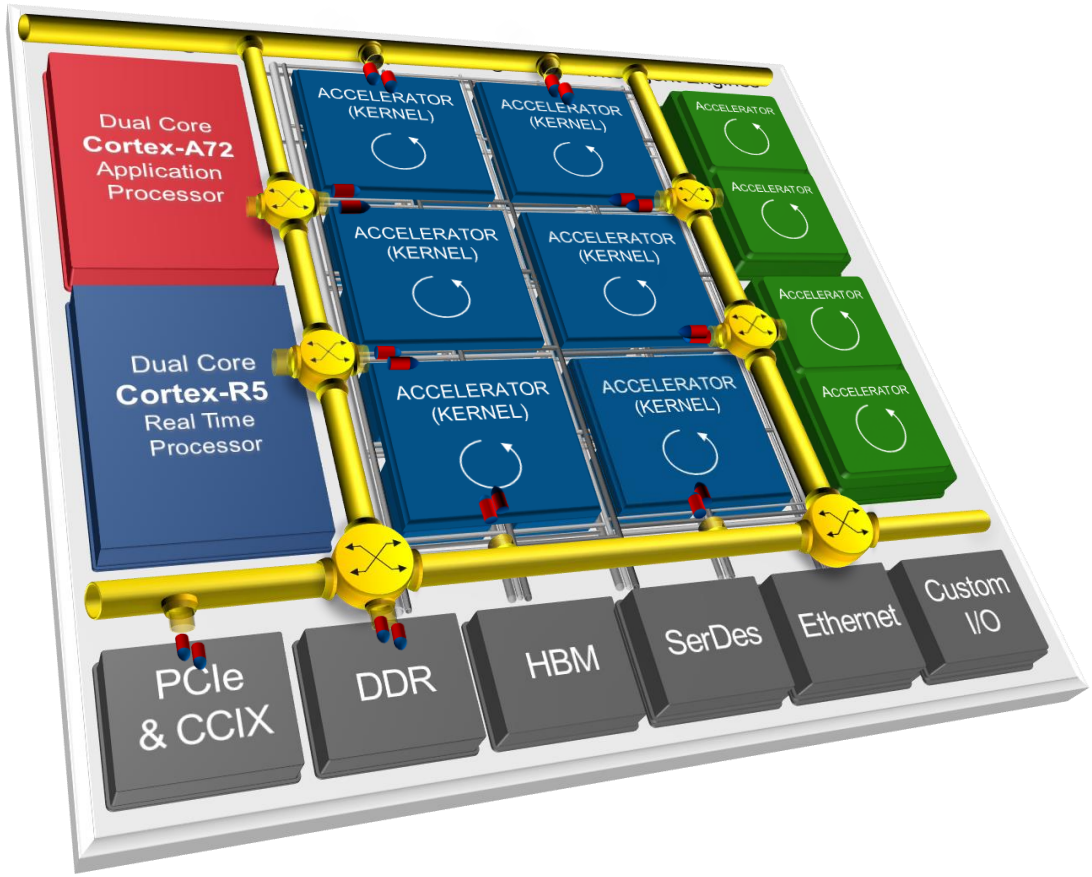
High Bandwidth, Low Latency

> Guaranteed QoS

Eases Kernel Placement

> Easily swap kernels at NoC port boundaries

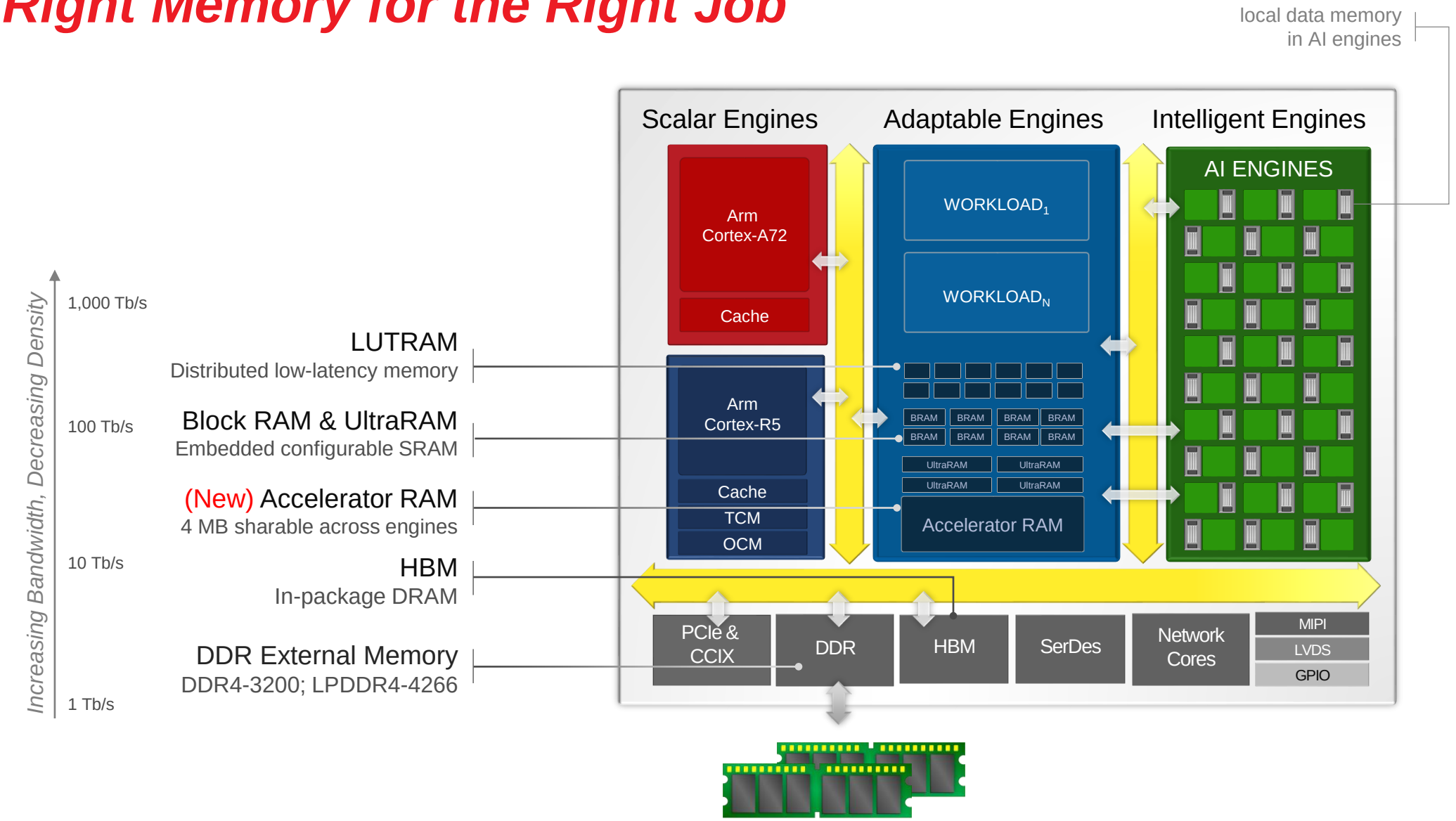
> Simplifies connectivity between kernels



XILINX.

Adaptable Memory Hierarchy

The Right Memory for the Right Job



XILINX

Programmable I/O for Any Sensor, Interface, or Memory

- Different IO types provide a wide range of speeds and voltages
- Configure the same I/O for either memory or sensor interfaces per application requirements

HDIO, MIO
(Legacy Standards)

XPIO
(High-Speed)

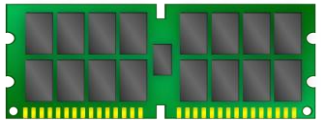
DC | 400Mb/s | 1600Mb/s | 3200Mb/s | 4266Mb/s

1.8V to 3.3V

0.6V to 1.5V

3.2Gb/s MIPI D-PHY
8 Mpix Sensors & Displays

4.2Gb/s LPDDR4X (0.6V)
Highest Memory Bandwidth Per Pin



3.2Gb/s DDR4
Server Class Density Per Channel

Versal in Automated Driving & Smart Sensors

Programmable I/O to Integrate Any Sensor
e.g., radar, LiDAR, multi-camera, GPS, digital map

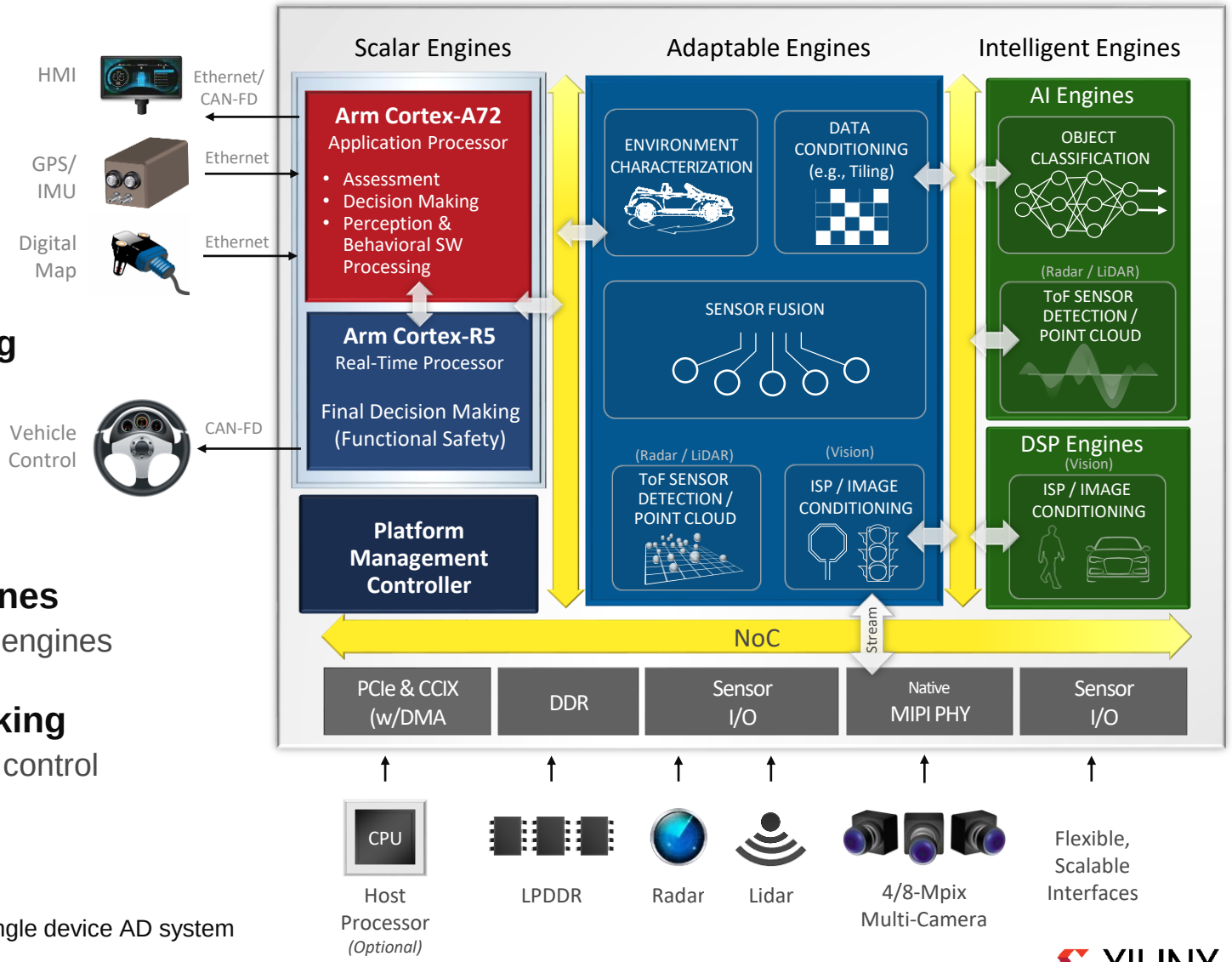
Adaptable/Intelligent Engines for Pre-Processing
e.g., image conditioning, point cloud for radar/LiDAR

Adaptable Engines for data fusion & conditioning
Preparing data for AI engine

AI Engines for Object Classification
Perception processing, sent to Adaptable Engines for Evn

Environment Characterization in Adaptable Engines
e.g., localization and mapping, also spread to scalar and AI engines

Scalar engines for assessment and decision making
Final decision done in Arm Cortex-R5, sent to HMI / vehicle control



Note: Demonstrates capabilities of Versal Architecture, not representing a single device AD system

XILINX.

XILINX®
VERSAL™

- AI RF Series
- HBM Series
- AI Core Series
- Premium Series
- AI Edge Series
- Prime Series

The First Two Series of the Versal Portfolio (Announced)

AI Core Series

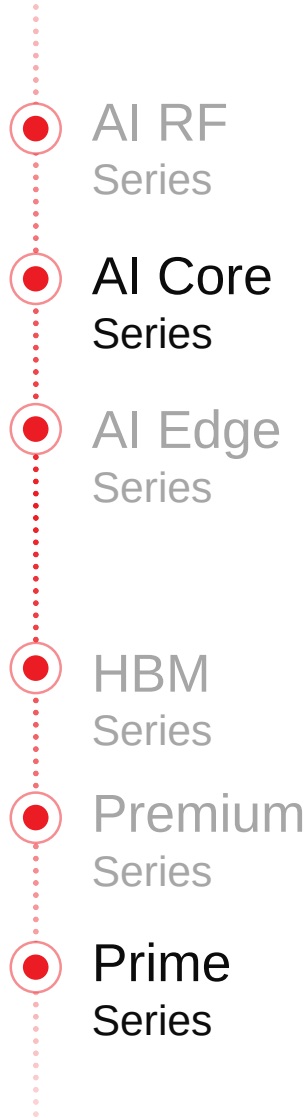
Breakthrough AI Inference Throughput

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

Prime Series

Broad Applicability Across Multiple Markets

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



Versal AI Core Series

Highest AI Inference Throughput
50 – 150 INT8 TOPs

First Available Device

		VC1352	VC1502	VC1702	VC1802	VC1902
Intelligent Engines	AI Engines	128	217	310	300	400
	DSP Engines	928	1,312	1,272	1,600	1,968
Adaptable Engines	System Logic Cells (K)	540	797	1,021	1,586	1,968
	Accelerator RAM (Mb)	32	0	32	0	0
	Total SRAM Capacity (Mb)	92	80	174	120	164
Scalar Engines	Application Processing Unit	Dual-core Arm® Cortex-A72, 48KB/32KB L1 Cache w/ parity & ECC; 1MB L2 Cache w/ ECC				
	Real-time Processing Unit	Dual-core Arm Cortex-R5, 32KB/32KB L1 Cache, 256KB TCM w/ECC and 256KB OCM w/ECC				
Foundational Platform	NoC Master / NoC Slave Ports	10	14	18	28	28
	DDR Memory Controllers	2	2	2	4	4
	CCIX & PCIe® w/DMA (CPM)	–	1 x Gen4x16, CCIX	–	1 x Gen4x16, CCIX	1 x Gen4x16, CCIX
	PCI Express®	1 x Gen4x8	4 x Gen4x8	1 x Gen4x8	4 x Gen4x8	4 x Gen4x8
	Multirate Ethernet MAC	1	4	3	4	4
	SD-FEC	2	0	5	0	0
I/O	Programmable I/O	500	500	500	770	770
	Transceivers	8	44	24	44	44

Enabling Ethernet at 10G/25G/50G/100G

Scalable DDR
128b – 256b w/ECC

256Gb/s PCIe & CCIX Bandwidth to Host

XILINX.

For Any Developer

Frameworks	TensorFlow, Caffe, mxnet, Spark, FFmpeg	AI & Data Scientists
New Unified Software Development Environment		Software Application Developers
Embedded Run-Time	Linux, Xen, freeRTOS	Embedded Developers
Vivado Design Suite		Hardware Developers

XILINX VERSAL™

Learn More

Visit www.xilinx.com/versal

> Watch ACAP Intro video

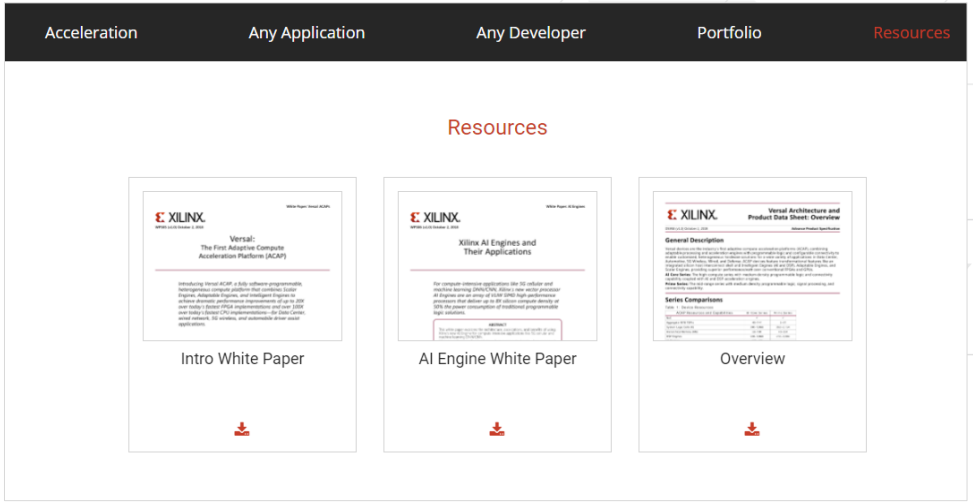
> Subscribe to mailing list for the latest news

View documentation and resources

> Data Sheet Overview

> Product Tables (Versal AI Core & Versal Prime Series)

> Versal Architecture and AI Engine White Papers



XILINX

Take-Aways

Versal: The First ACAP

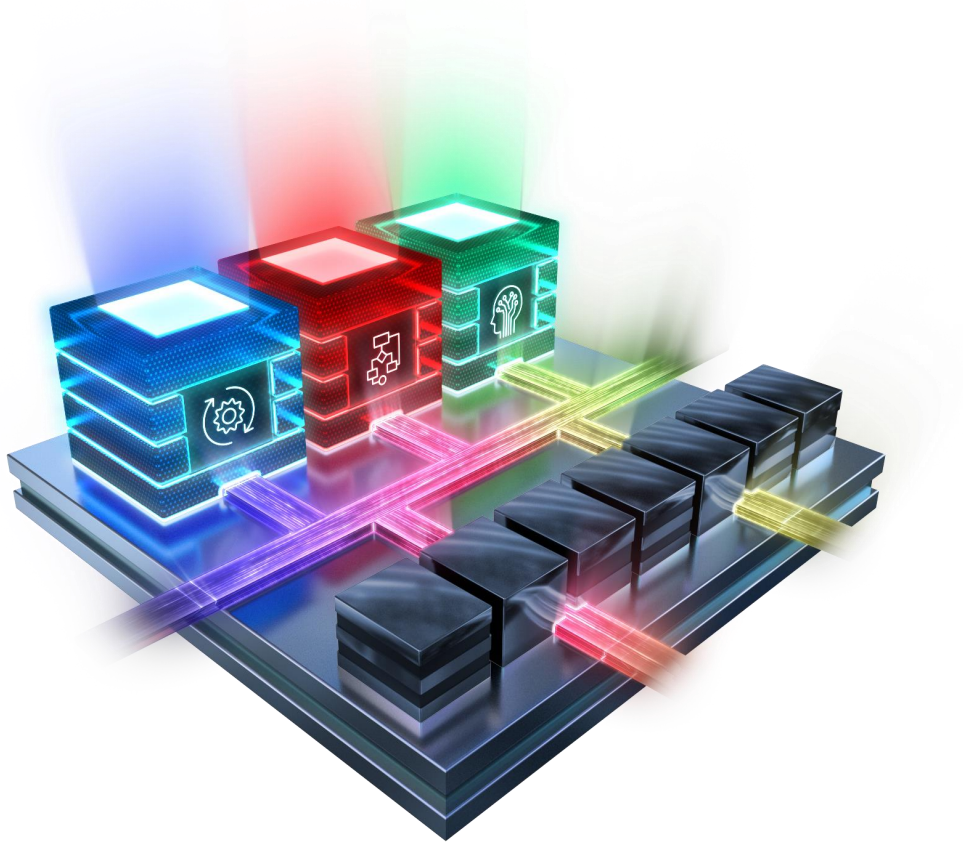
- Heterogeneous Acceleration
- For Any Application
- For Any Developer

Two Device Series Announced

- Versal Prime Series for Broad Application
- Versal AI Core Series for Highest AI Throughput

Availability

- Early Access Program for SW and tools
- First Devices Available 2H 2019



XILINX