



Introducing the Versal Architecture

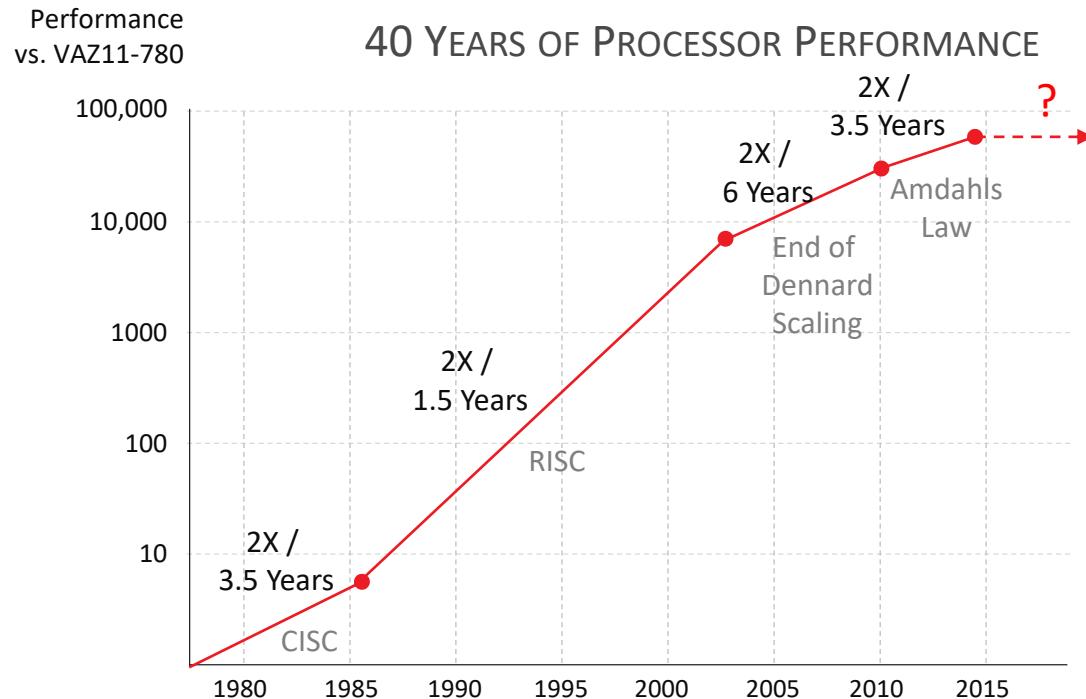
Presented By

Sumit Shah
Director of Silicon Product Marketing and Management
October 2, 2018



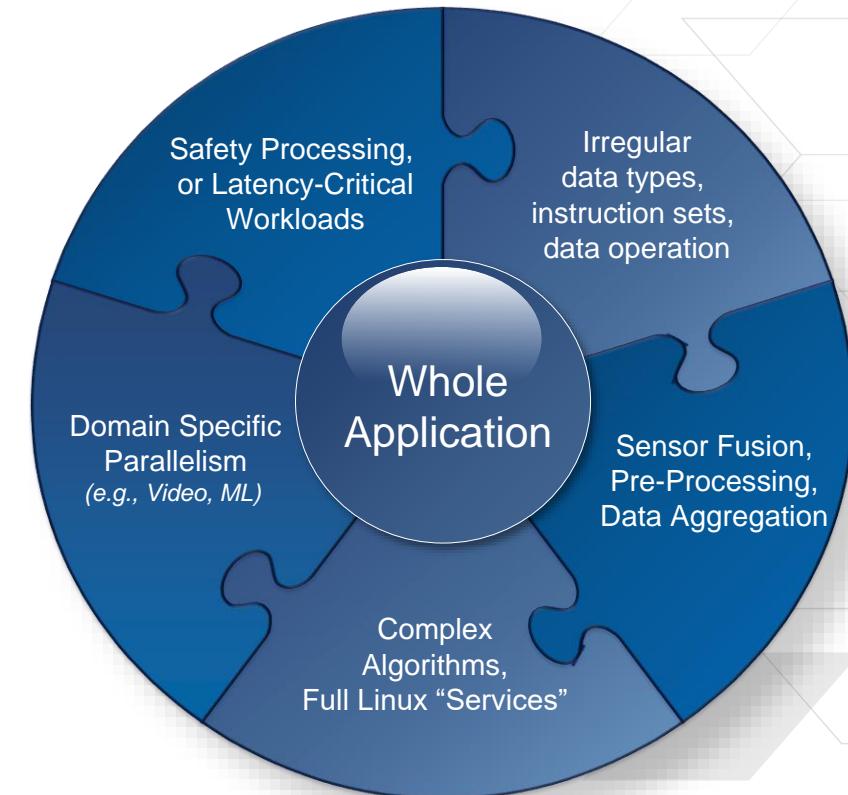
The Technology Conundrum .. And the Need for a New Compute Paradigm

Processing Architectures
are Not Scaling



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

A Single Architecture
Can't Do It Alone



Need for a New Programming Paradigm

Software Developer

Needs Agility and Abstraction



Modify,
Design,
Add Code

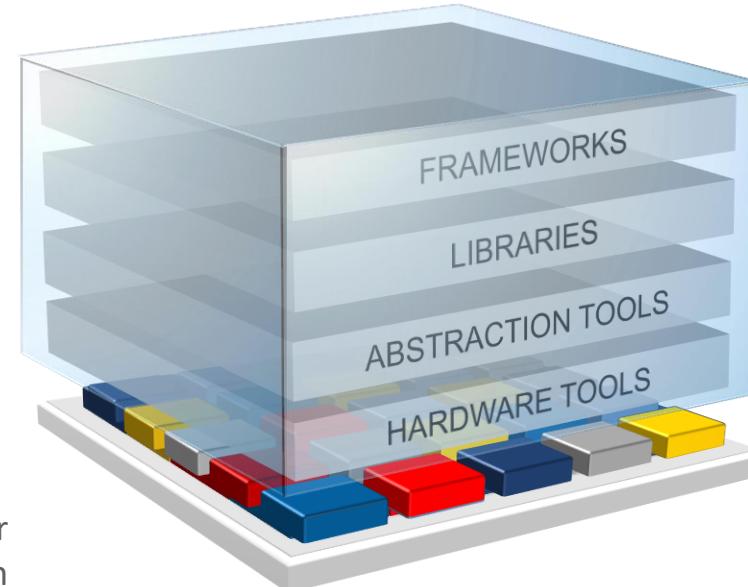
Familiar
Platform

GitHub

Ecosystem of
Libraries



Need a Scalable,
Unified Platform



Hardware Developer

Needs Flexibility to Optimize for Performance/Power



Flexible Platform
to Optimize for Performance/Power

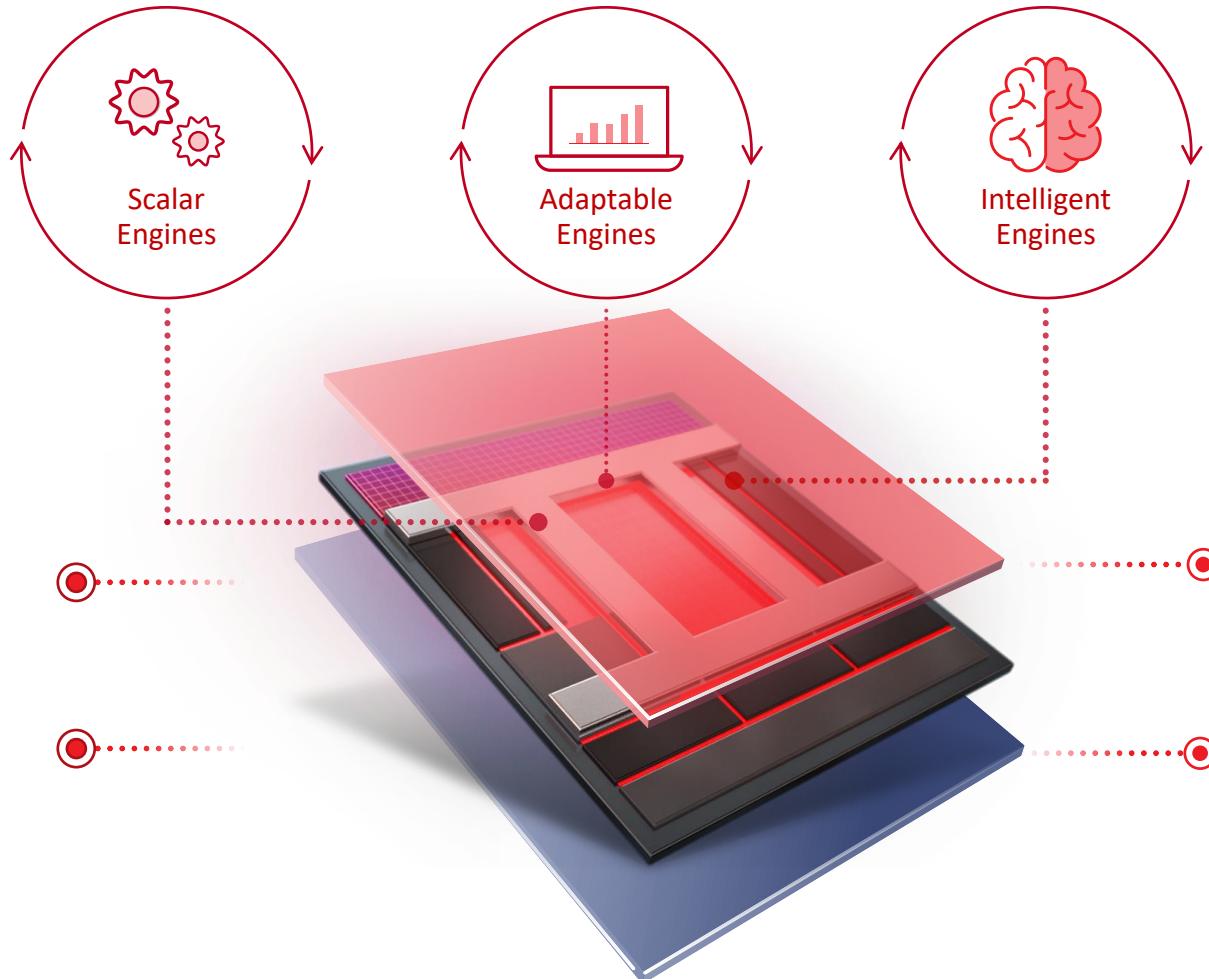
New Device Category: Adaptive Compute Acceleration Platform

ADAPTIVE

Diverse Workloads in
Milliseconds

Future-Proof for
New Algorithms

COMPUTE ACCELERATION

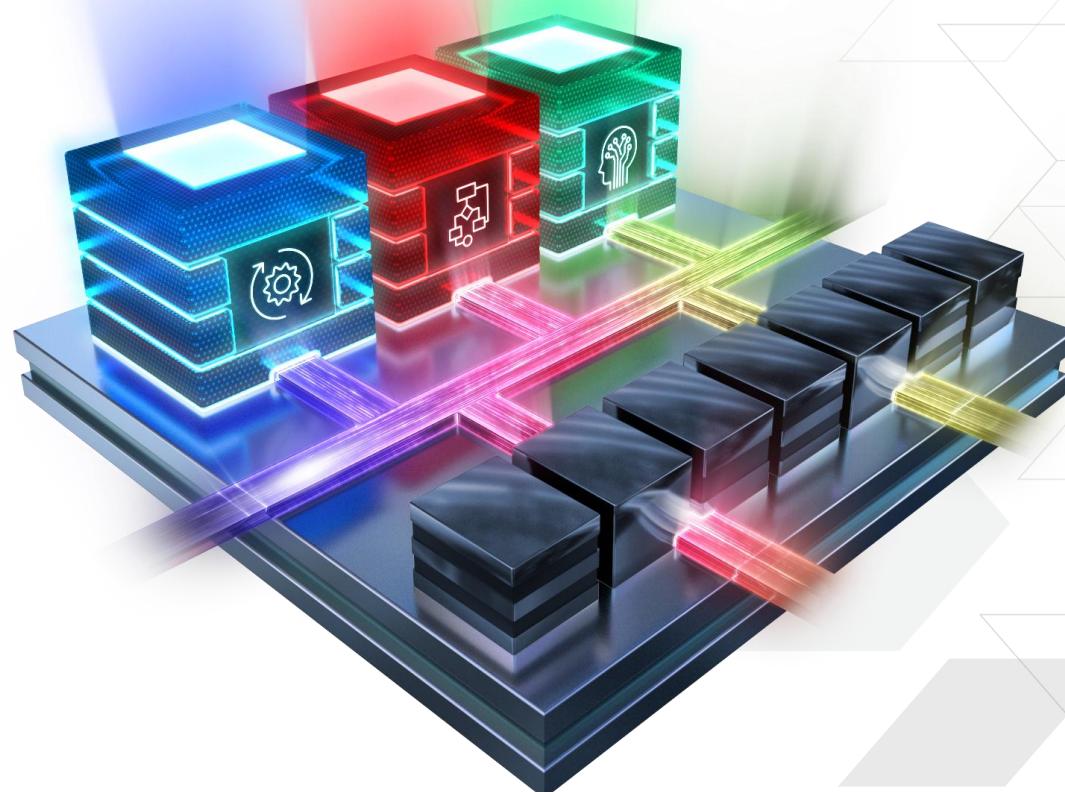


Introducing the World's First ACAP

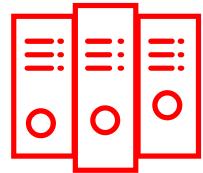


XILINX
VERSAL™

- > Heterogeneous Acceleration
- > For Any Application
- > For Any Developer



Breakthrough Performance for Cloud, Network, and Edge



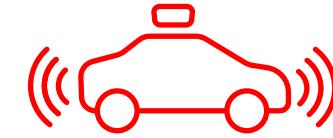
Cloud Compute
Breakthrough AI Inference



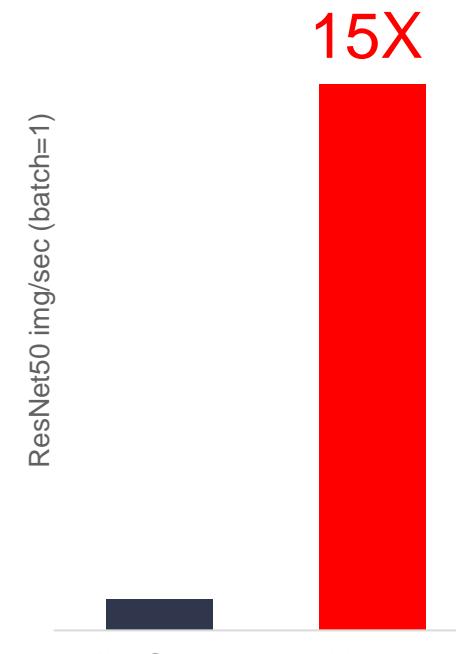
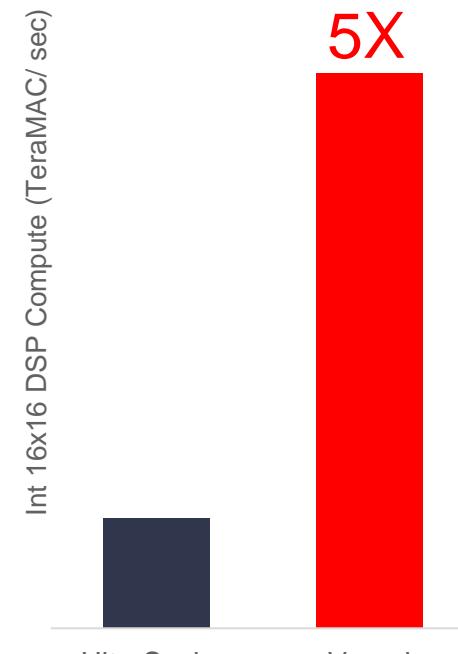
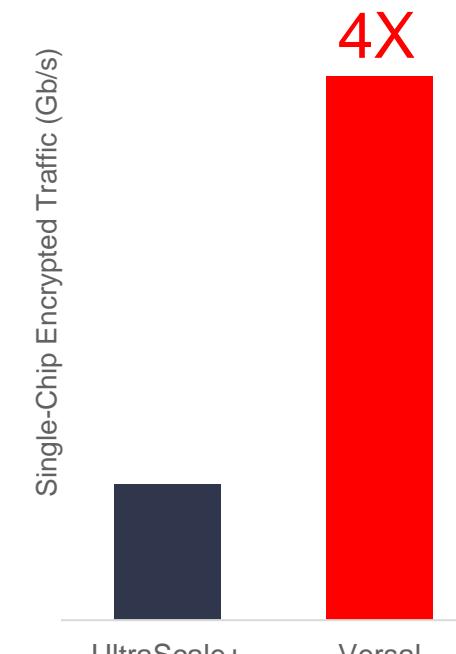
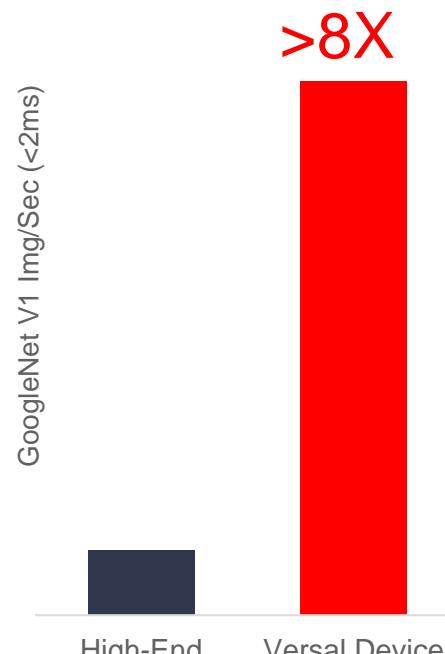
Networking
Multi-terabit Throughput



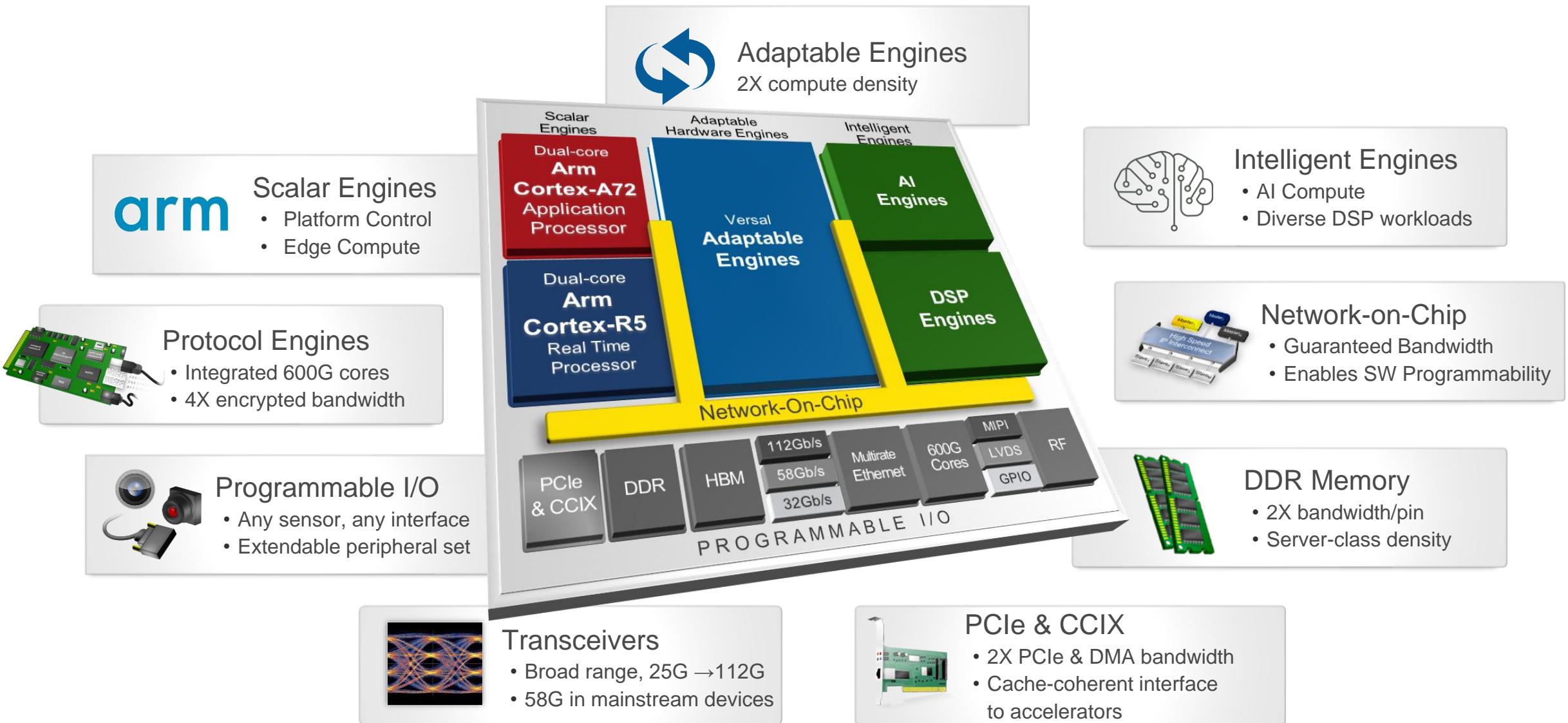
5G Wireless
Compute for Massive MIMO



Edge Compute
AI Inference at Low Power



Versal Architecture Overview

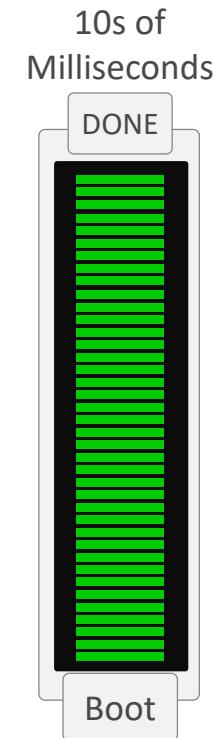


Platform Management Controller

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

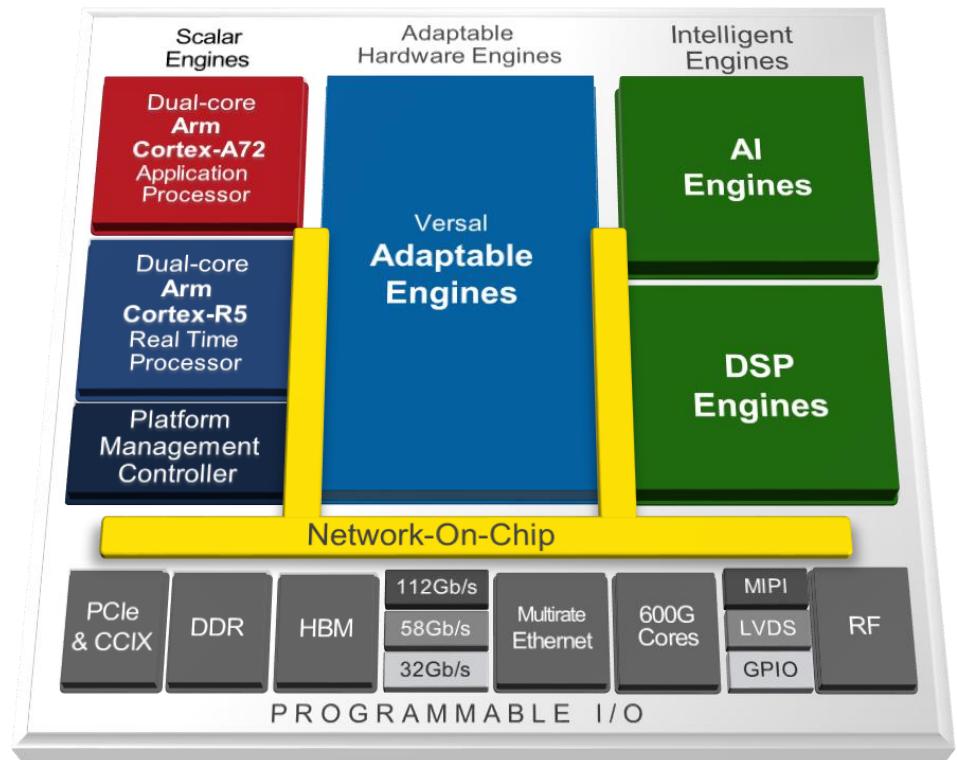


Security, Safety & Reliability Enclave

- > 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

A Processor in Every Device

Diverse Use Models for Scalar Processing



Edge Compute & Autonomous Systems

Complex processing for intelligent edge and endpoint



Control Plane Processing in the Network & Cloud

Data path management & device management



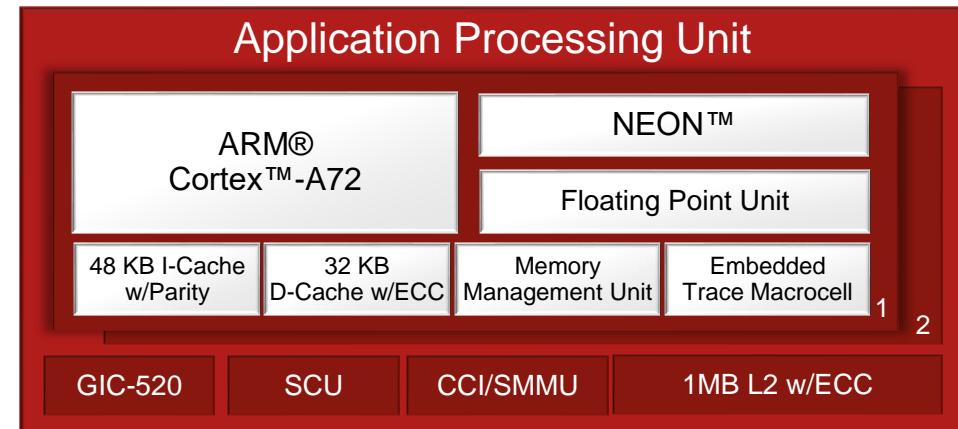
Operation & Management in Communication Systems

Board level control monitoring

The Arm Subsystem

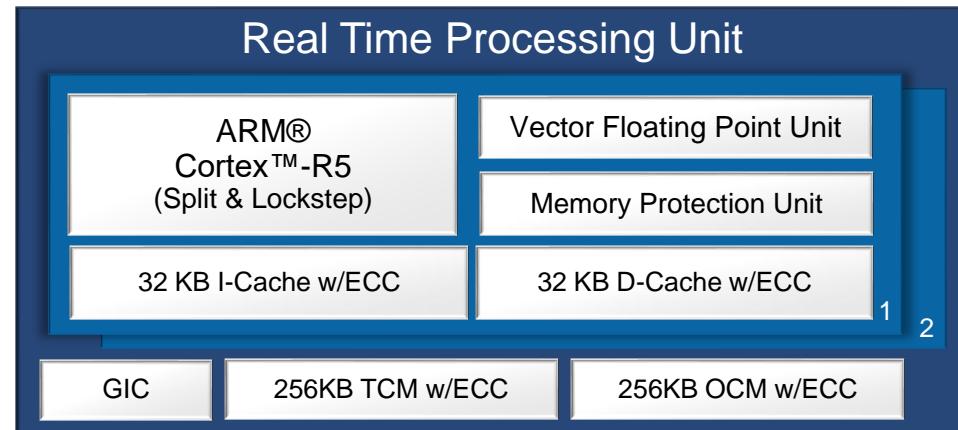
Dual-Core ARM Cortex-A72 Application Processors

- > Up to 1.7GHz for 2X single-threaded performance¹
- > 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 performance¹
- > 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

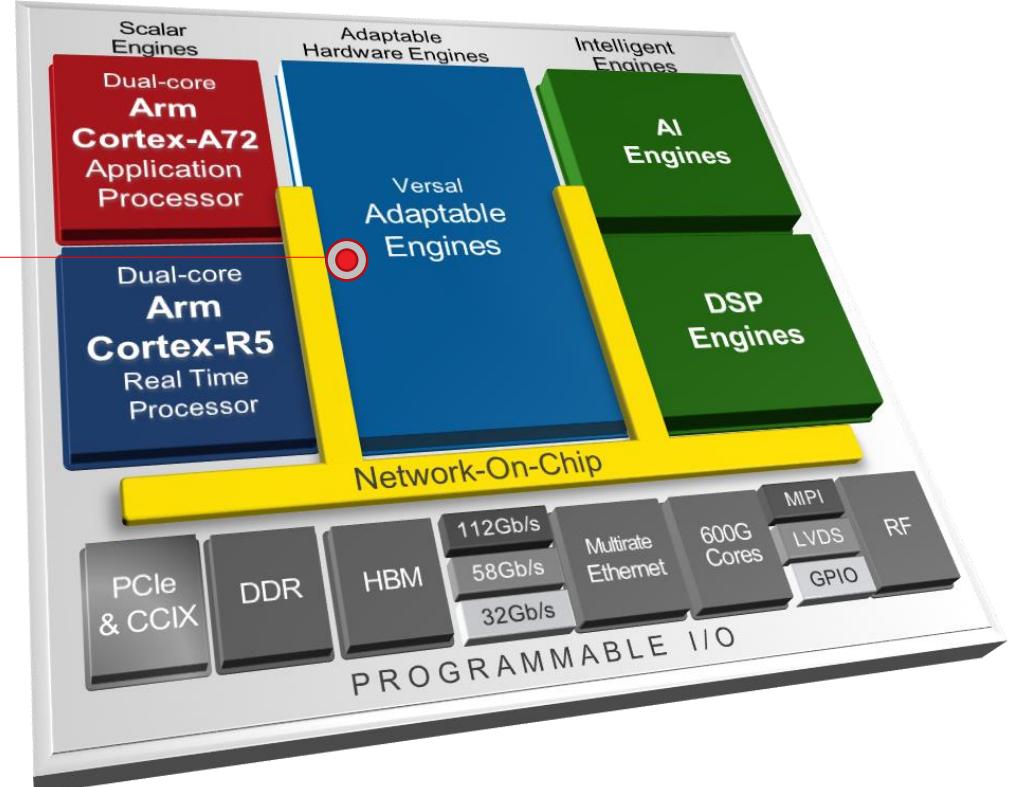
Adaptable Engines

Adaptable Hardware Engines

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

Programmable memory hierarchy to optimize compute efficiency

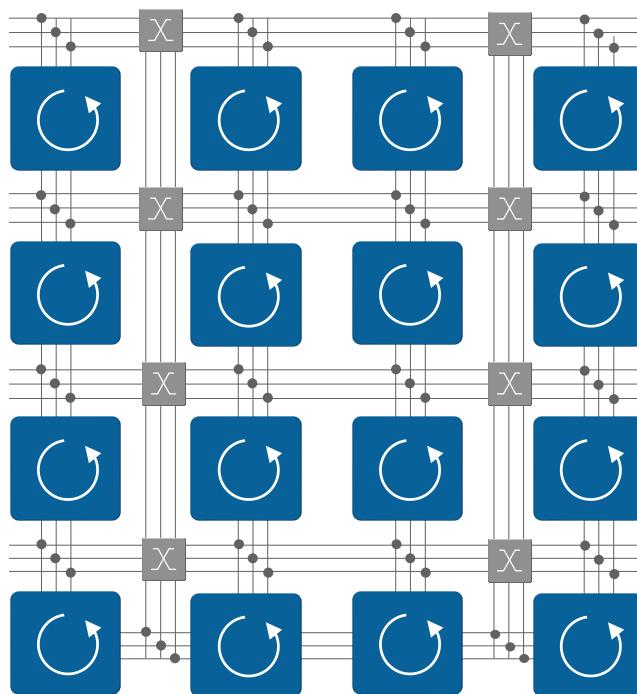
High bandwidth, low latency data movement between engines and I/O



Greater Compute Density for Any Workload

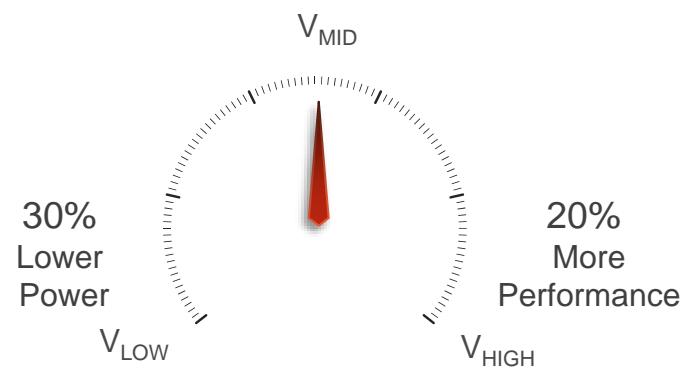
Re-Architected Hardware Fabric

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



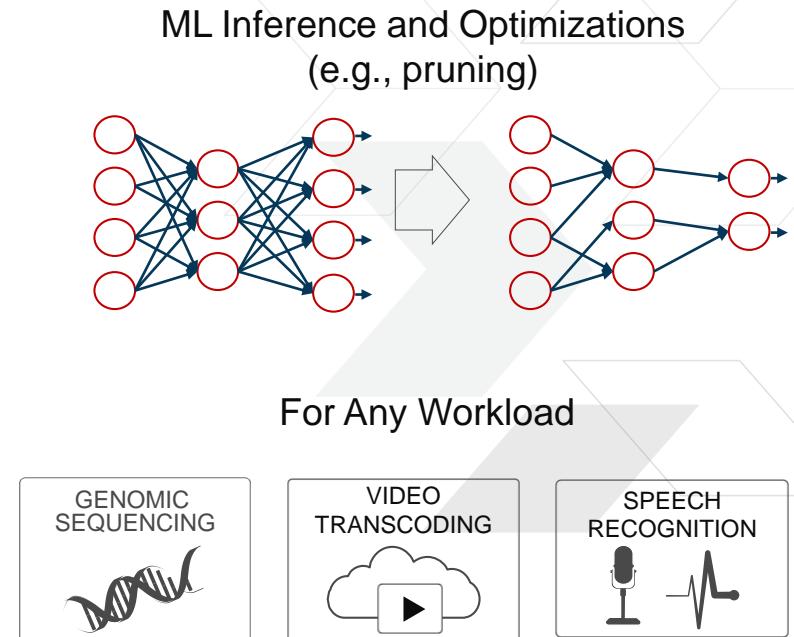
Tune for Power & Performance

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



Adaptable to any Workload

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



Intelligent Engines

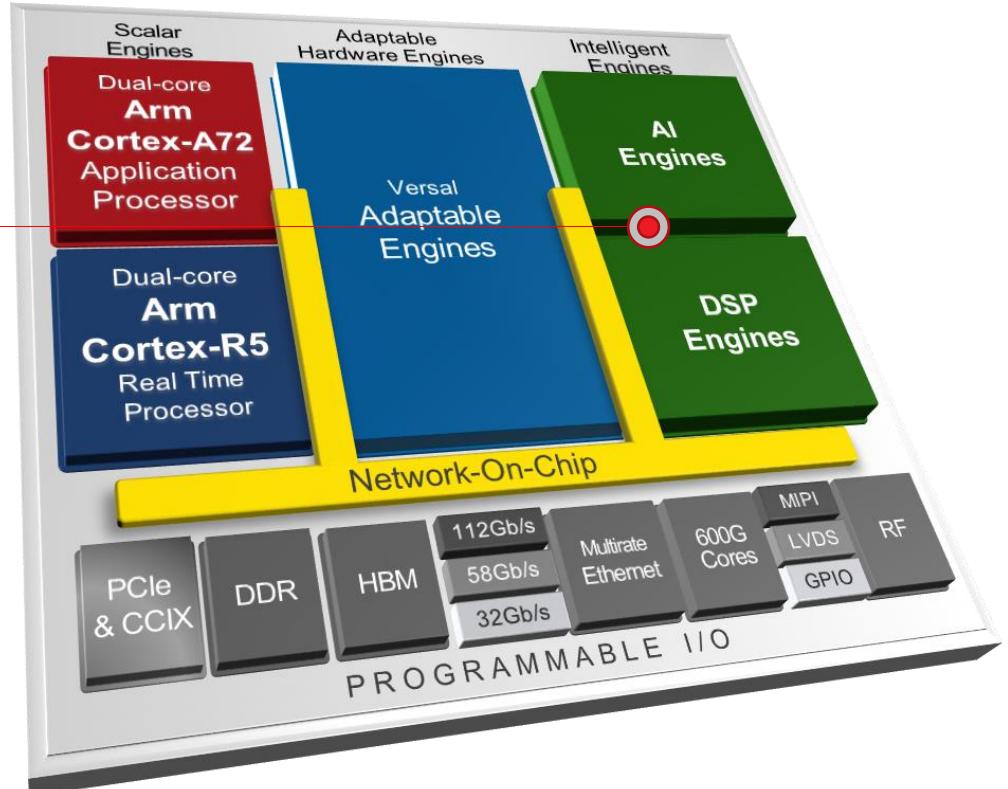
Intelligent Engines for Diverse Compute

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

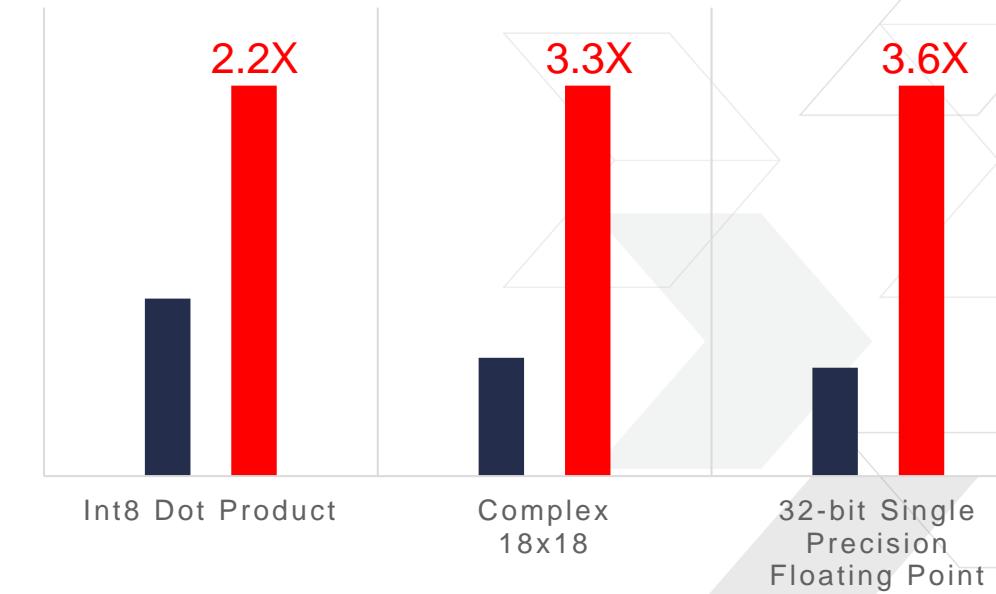
- > Integrated FP32, FP16 floating point, INT24 (HPC)
- > Integrated complex 18x18 operation (wireless, cable access)
- > Double the performance in INT8 operation (AI inference)

Code Portability for UltraScale+ 16nm designs

- > Support for legacy IP and LogiCore libraries
- > Compatibility with SysGen, Model Composer, HLS tools

Performance Improvement

■ UltraScale+ 16nm ■ Versal 7nm



Intelligent 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

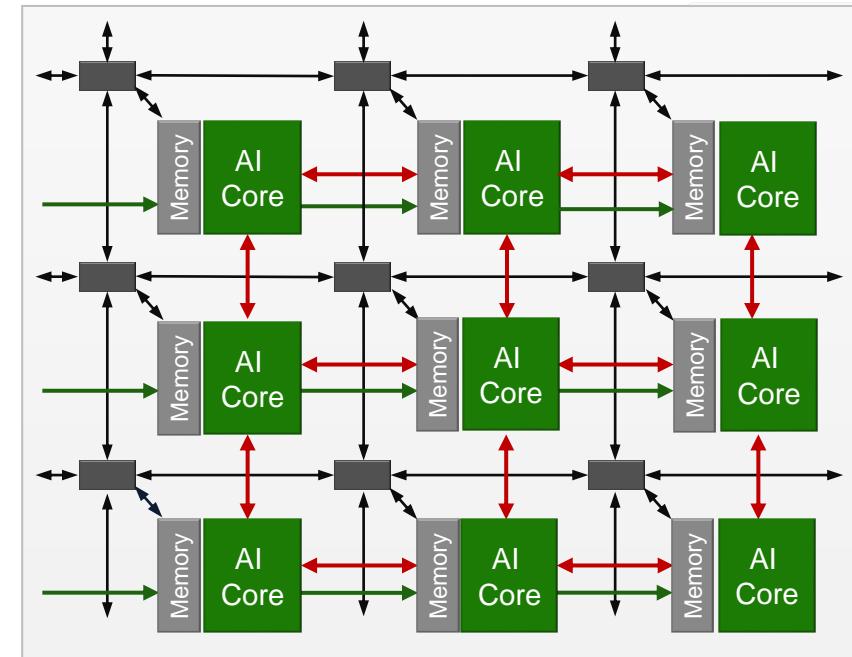
- > Instantiate multiple tiles (10s to 100s) for scalable compute

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



NoC for Ease of Use, Guaranteed Bandwidth, and Power Efficiency

High bandwidth terabit network-on-chip

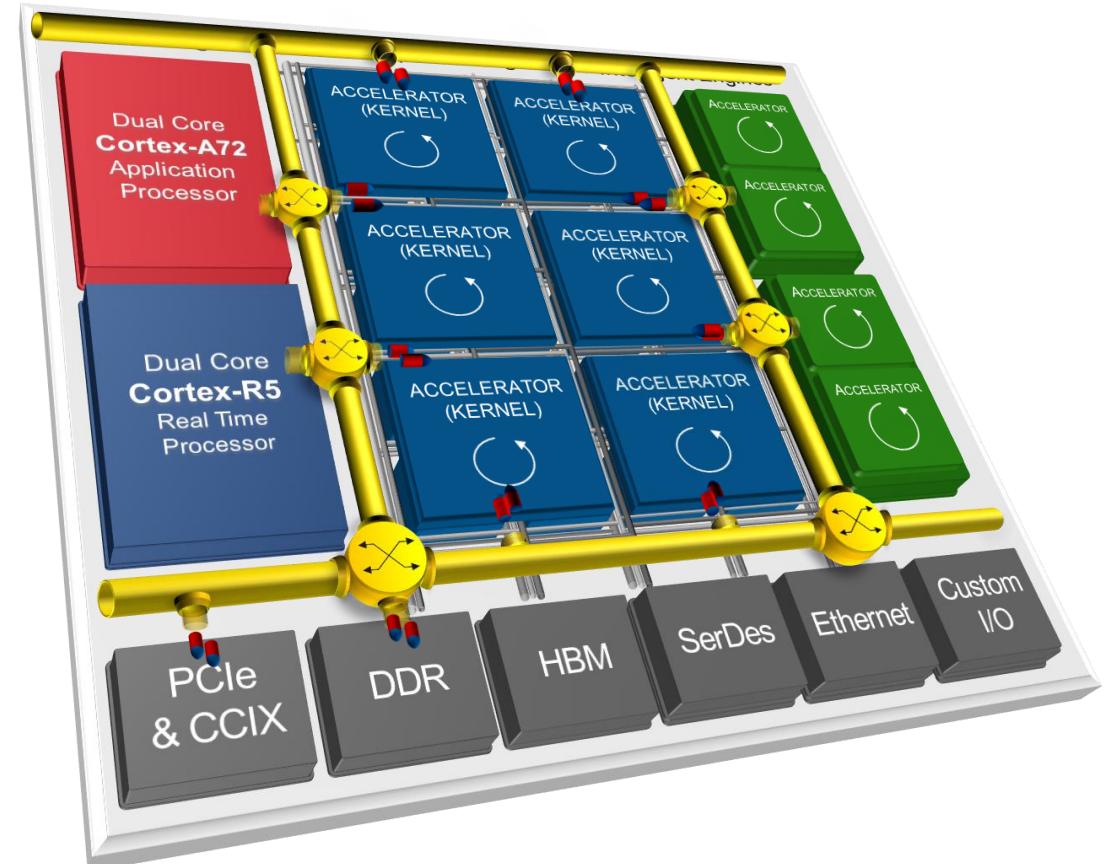
- > Memory mapped access to all resources
- > Built-in arbitration between engines and memory

High Bandwidth, Low Latency, Low power

- > Guaranteed QoS
- > 8X power efficiency vs. FPGA implementations

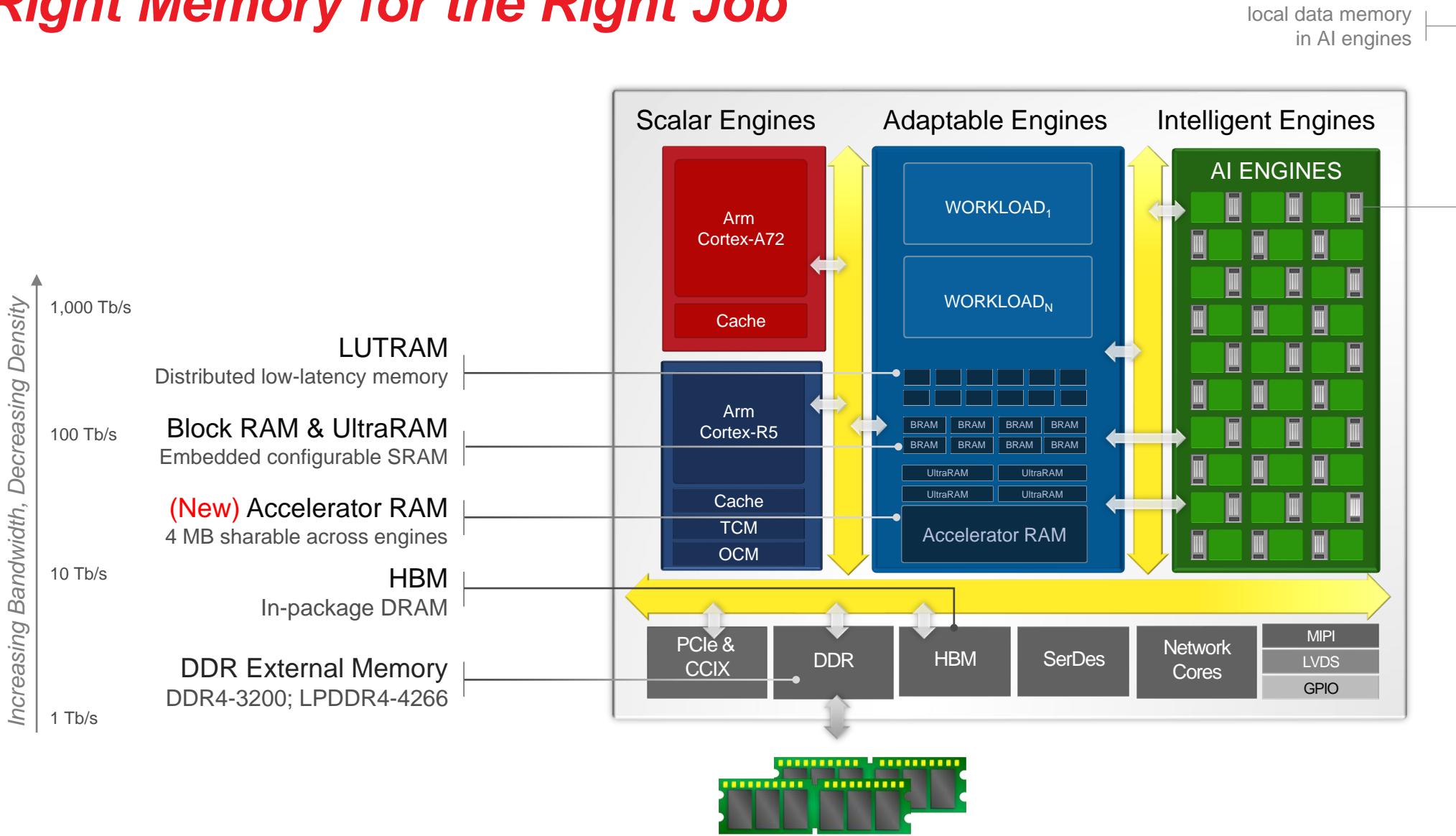
Eases Kernel Placement

- > Easily swap kernels at NoC port boundaries
- > Simplifies connectivity between kernels



Adaptable Memory Hierarchy

The Right Memory for the Right Job



Introducing the “Integrated Shell”

‘Shell’: Pre-Built Core Infrastructure & System Connectivity

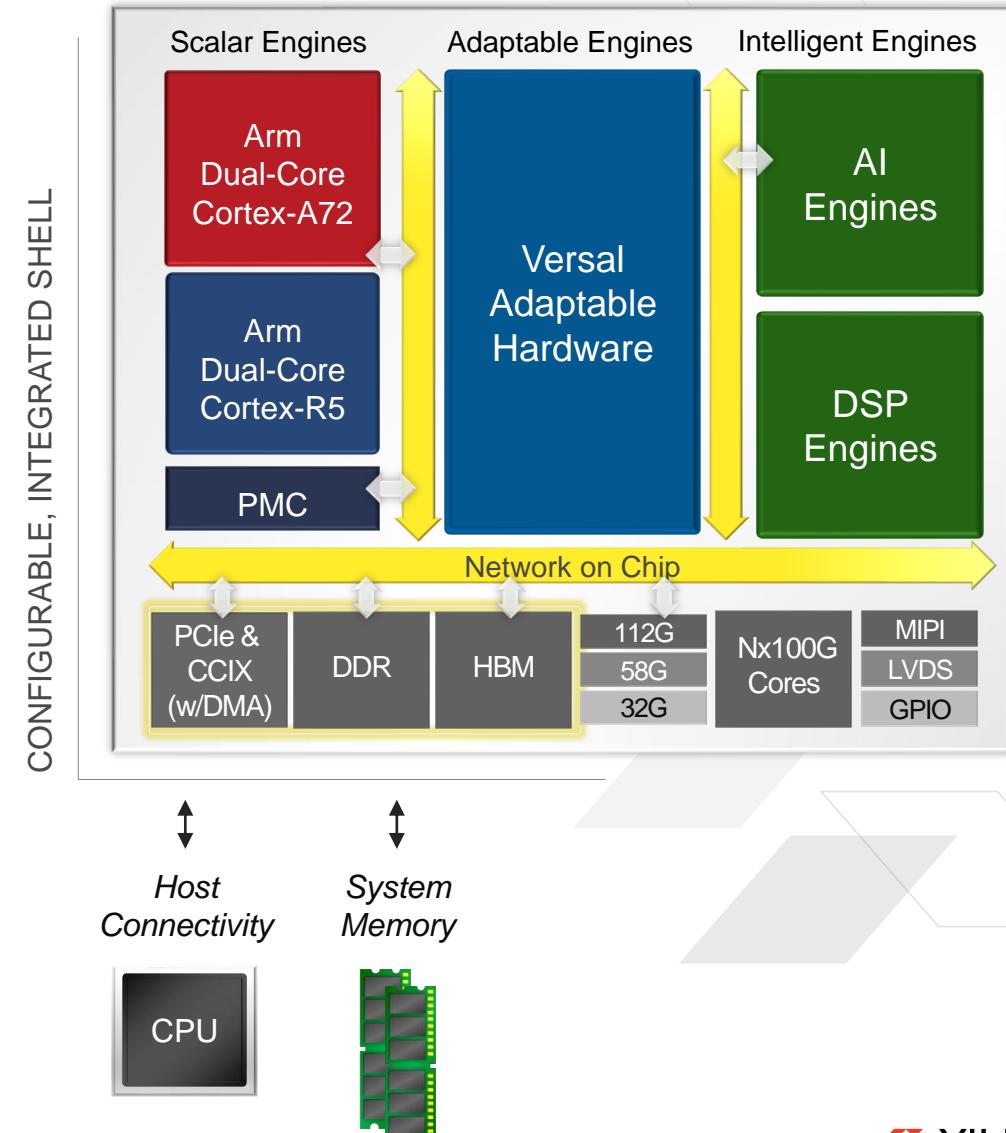
- > External host interface
- > Memory subsystem
- > Basic interfaces (e.g., JTAG, USB, GbE)

Key Architectural Elements of the Shell

- > Platform Management Controller (PMC)
- > Integrated host interfaces: PCIe & CCIX, DMA
- > Scalable Memory Subsystem: DDR4 & LPRDDR4
- > Network-on-Chip for connectivity and arbitration

Greater Performance, Device Utilization, and Productivity

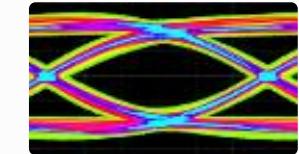
- > More of the platform available for application’s workload(s)
- > Target application runs faster with less device congestion
- > Turn-key, pre-engineered timing closure – no debug



Transceivers: Robust and Scalable Connectivity

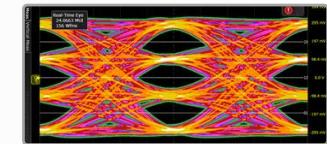
32G
NRZ Transceivers

Optimized for latency and power



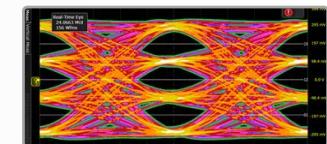
58G
PAM4 Transceivers

Tuned for the latest copper cable,
backplane & optical interfaces



112G
PAM4 Transceivers

Industry-leading performance for
copper cable, backplane, optical



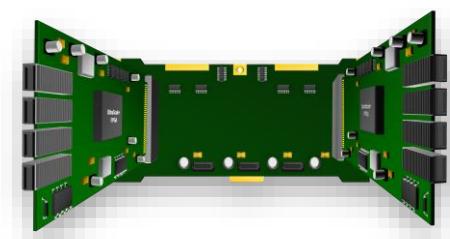
COPPER CABLE



OPTICS

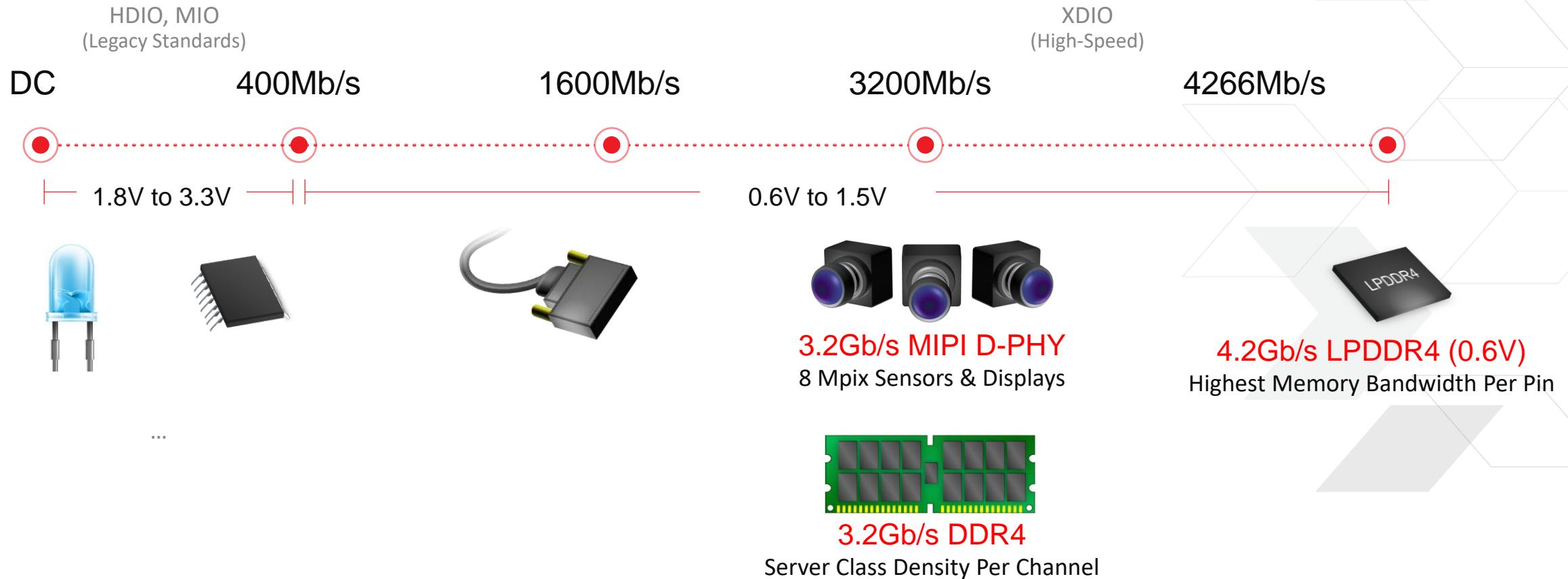


BACKPLANE



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



Versal Core Series Enables “Smart Cities”

Video Surveillance with Machine Learning

Host Connectivity and Network Connectivity

Integrated PCIe and Ethernet

Adaptable Engines Optimize Compute/Watt

Video scaling, custom memory hierarchy, compression

DSP Engines for Video Transcode

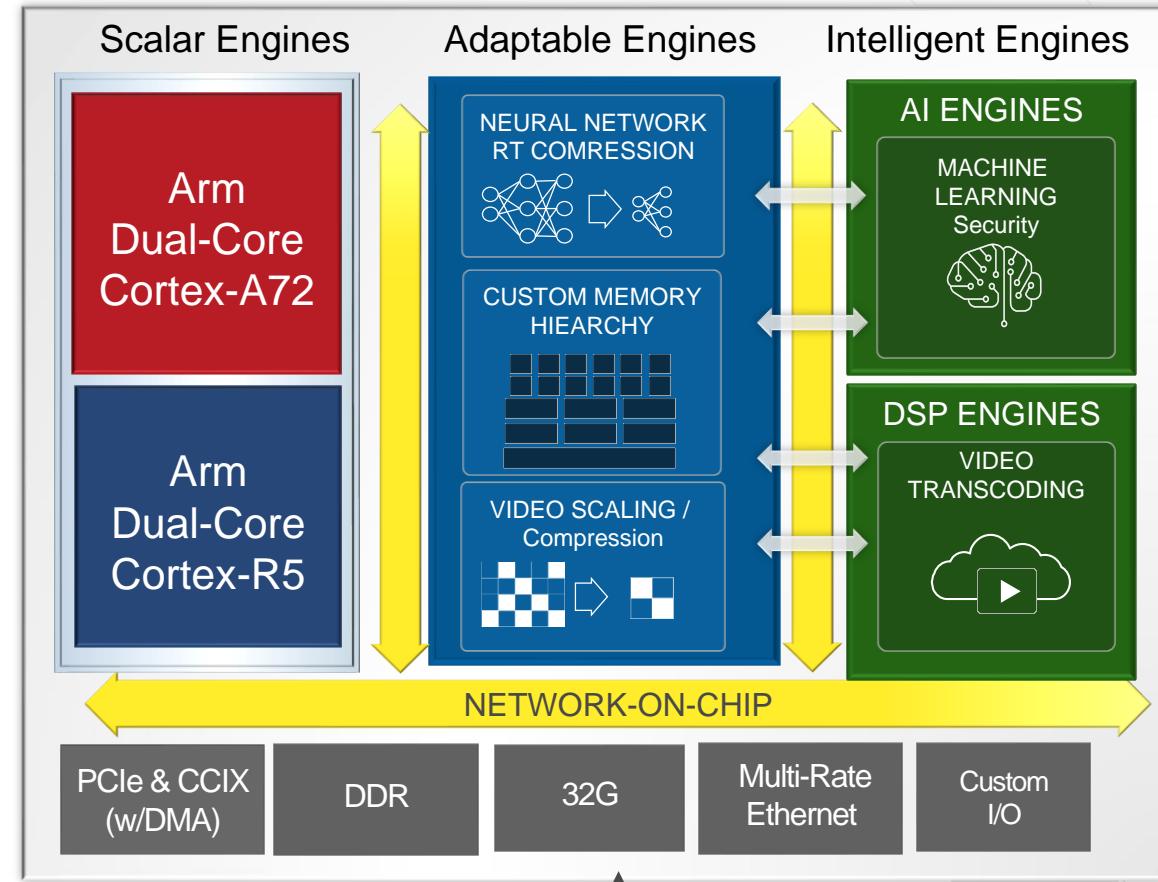
Scalable for legacy and emerging video formats

AI Engines for Real-Time Image Recognition

License Plate/Facial Recognition

Network On Chip and Memory Subsystem

Interconnects memory and compute Engines



For Any Developer

Frameworks



Caffe



AI & Data
Scientists

New Unified Software
Development Environment

Software Application
Developers

Embedded Run-Time



Embedded
Developers

Vivado Design Suite

Hardware
Developers





XILINX.[®]
VERSAL™

AI Edge
Series

AI Core
Series

AI RF
Series

Prime
Series

Premium
Series

HBM
Series

Part of the Xilinx Product & Technology Portfolio

DEVICE CATEGORY	FPGA	SoC	ACAP
FEATURED PRODUCTS	Spartan Artix Kintex Virtex	Zynq-7000 Zynq UltraScale+ MPSoC Zynq UltraScale+ RFSoC	Versal

Announcing the First Two Series of the Versal Portfolio

➤ 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

		VC1352	VC1502	VC1702	VC1802	VC1902	First Available Device
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	
I/O	SD-FEC	2	0	5	0	0	
	Programmable I/O	500	500	500	770	770	
	Transceivers	8	44	24	44	44	

Scalable DDR
128b – 256b w/ECC

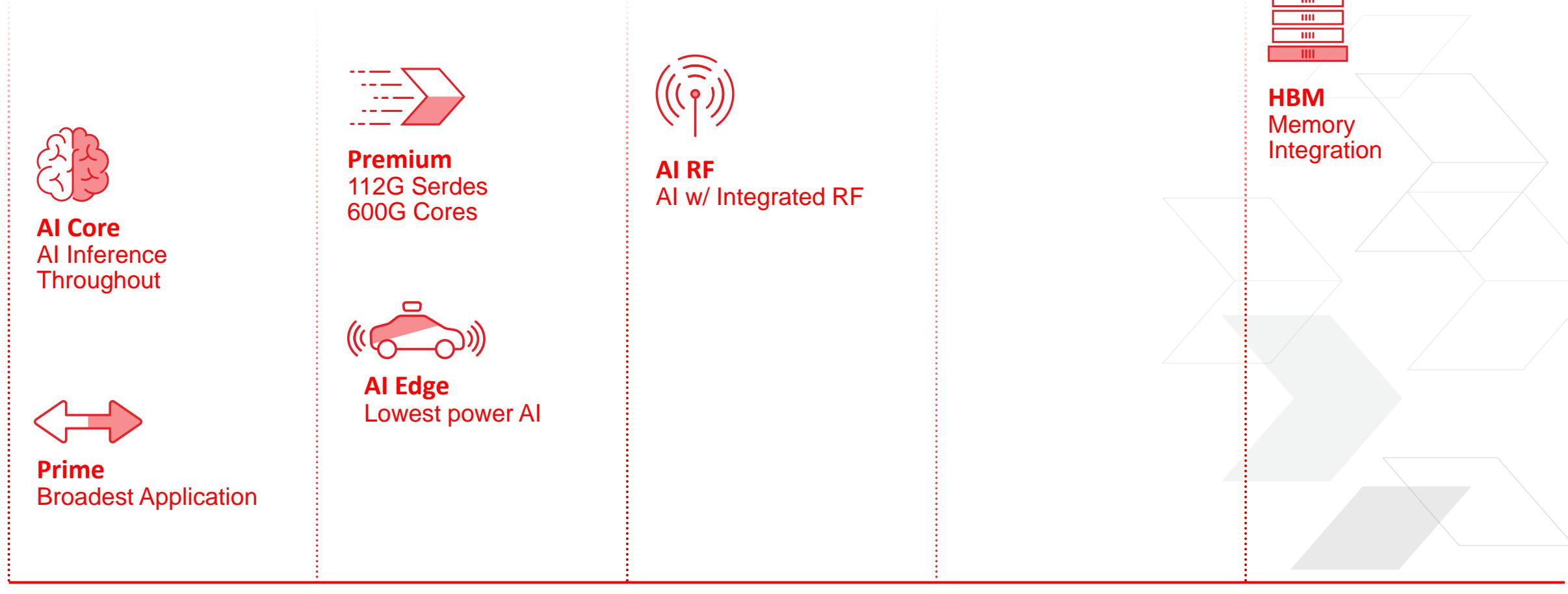
Enabling Ethernet at
10G/25G/50G/100G

256Gb/s PCIe & CCIX
Bandwidth to Host

Versal Prime Series

		VM1102	VM1302	VM1402	VM1502	VM1802	VM2502	VM2602	VM2702	VM2902
Intelligent Engines	DSP Engines	472	736	1,504	1,312	1,968	3,984	1,880	2,500	3,080
Adaptable Engines	System Logic Cells (K)	352	572	1,002	797	1,968	2,030	1,263	1,805	2,154
	Total SRAM Capacity (Mb)	35	63	87	80	164	245	174	243	294
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	5	16	16	14	28	28	16	26	26
	DDR Bus Widths	64	128	256	128	256	288	384	384	384
	DDR Memory Controllers	1	2	4	2	4	5	6	6	6
	CCIX & PCIe® w/DMA (CPM)	-	-	-	1 x Gen4x16, CCIX	1 x Gen4x16, CCIX	1 x Gen4x16, CCIX	1 x Gen4x16, CCIX	1 x Gen4x16, CCIX	1 x Gen4x16, CCIX
	PCI Express®	1 x Gen4x8	2 x Gen4x8	2 x Gen4x8	4 x Gen4x8	4 x Gen4x8	1 x Gen4x8	1 x Gen4x8	2 x Gen4x8	2 x Gen4x8
	Multirate Ethernet MAC	1	2	2	4	4	1	2	2	2
I/O	Programmable I/O	316	554	770	500	770	824	802	824	824
	Transceiver	12	24	24	44	44	44	52	76	92
	Integrated DDR Controllers & Protocol Engines						I/O Count Optimized			
							Transceiver Bandwidth Optimized			

Versal Roadmap

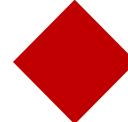


2H 2019

2020

2021

Getting Started



Visit www.xilinx.com/versal

- > Check out the Media Kit
- > Watch ACAP Intro video
- > Subscribe to mailing list for the latest news



View documentation and resources

- > Data Sheet Overview
- > Product Tables
- > Versal Architecture and AI Engine White Papers

XILINX VERSAL™
Industry's First ACAP
Adapative Compute Acceleration Platform

Watch Video > Media Kit > Stay Informed >

Acceleration Any Application Any Developer Portfolio Resources

Resources

Intro White Paper AI Engine White Paper Overview

Key Take-Aways

- **Versal: The First ACAP**
 - > Heterogeneous Acceleration
 - > For Any Application
 - > For Any Developer
- **Announcing Two Families**
 - > Versal Prime Series for Broad Application
 - > Versal AI Core Series for Highest AI Throughput
- **Availability**
 - > Early Access Program for SW and tools
 - > Devices Available 2H 2019





Versal AI Core Series

		VC1352	VC1502	VC1702	VC1802	VC1902	First Available Device
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	
I/O	Multirate Ethernet MAC	1	4	3	4	4	
	SD-FEC	2	0	5	0	0	
	Programmable I/O (4G, 3.3V)	378, 122	378, 122	378, 122	648, 122	648, 122	
	Transceivers (32G)	8	44	24	44	44	

Scalable DDR
128b – 256b w/ECC

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

256Gb/s PCIe & CCIX
Bandwidth to Host

Versal Prime Series

		VM1102	VM1302	VM1402	VM1502	VM1802	VM2502	VM2602	VM2702	VM2902
Intelligent Engines	DSP Engines	472	736	1,504	1,312	1,968	3,984	1,880	2,500	3,080
Adaptable Engines	System Logic Cells (K)	352	572	1,002	797	1,968	2,030	1,263	1,805	2,154
	Total SRAM Capacity (Mb)	35	63	87	80	164	245	174	243	294
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	5	16	16	14	28	28	16	26	26
	DDR Bus Widths	64	128	256	128	256	288	384	384	384
	DDR Memory Controllers	1	2	4	2	4	5	6	6	6
	CCIX & PCIe® w/DMA (CPM)	-	-	-	1 x Gen4x16, CCIX	1 x Gen4x16, CCIX	1 x Gen4x16, CCIX	1 x Gen4x16, CCIX	1 x Gen4x16, CCIX	1 x Gen4x16, CCIX
	PCI Express®	1 x Gen4x8	2 x Gen4x8	2 x Gen4x8	4 x Gen4x8	4 x Gen4x8	1 x Gen4x8	1 x Gen4x8	2 x Gen4x8	2 x Gen4x8
	Multirate Ethernet MAC	1	2	2	4	4	1	2	2	2
I/O	Programmable I/O (4G, 3.3V)	216, 100	432, 122	648, 122	378, 122	648, 122	702, 122	702, 100	702, 122	702, 122
	Transceiver (32G, 58G)	12, 0	24, 0	24, 0	44, 0	44, 0	16, 28	20, 32	32, 44	40, 52
		Integrated DDR Controllers & Protocol Engines					I/O Count Optimized		Transceiver Bandwidth Optimized	

Versal AI Core Series

	VC1352	VC1502	VC1702	VC1802	VC1902
Intelligent Engines	AI Engines	128	217	310	300
	AI Engine Data Memory Blocks (#)	1024	1736	2480	2400
	AI Engine Data Memory (Mb)	32	54.25	77.5	75
Adaptable Engines	DSP Engines	928	1,312	1,272	1,600
	System Logic Cells (K)	540	797	1,021	1,586
	LUTs	246,784	364,544	466,688	725,000
Memory	Distributed RAM (Mb)	8	11	14	22
	Total Block RAM (Mb)	18	19	29	28
	UltraRAM (Mb)	42	60	113	91
Scalar Engines	Accelerator RAM (Mb)	32	0	32	0
	Total SRAM Capacity (Mb)	92	80	174	120
	Application Processing Unit	Dual-core Arm® Cortex-A72, 48KB/32KB L1 Cache w/ parity & ECC; 1MB L2 Cache w/ ECC			
Foundational Platform	Real-time Processing Unit	Dual-core Arm Cortex-R5, 32KB/32KB L1 Cache, and 256KB TCM w/ECC			
	Memory	256KB On-Chip Memory w/ECC			
	Connectivity	Ethernet (x2); UART (x2); CAN-FD (x2); USB 2.0 (x1); SPI (x2); I2C (x2)			
Platform Management Controller	NoC Master / NoC Slave Ports	10	14	18	28
	DDR Bus Width	128	128	128	256
	DDR Memory Controllers	2	2	2	4
	CCIX & PCIe® w/DMA (CPM)	–	1 x Gen4x16, CCIX	–	1 x Gen4x16, CCIX
	PCI Express®	1 x Gen4x8	4 x Gen4x8	1 x Gen4x8	4 x Gen4x8
	Multirate Ethernet MAC	1	4	3	4
	SD-FEC	2	0	5	0
	Platform Management Controller	Boot, Security, Safety, Monitoring, and High Speed Debug			
Package Footprint	Package Dimensions	Ball Pitch	XPIO, HDIO, MIO, GTY	XPIO, HDIO, MIO, GTY	XPIO, HDIO, MIO, GTY
A1024	31x31	0.92	378, 22, 78, 8	378, 22, 78, 8	378, 44, 78, 24
E1369	35x35	0.92	378, 44, 78, 8	378, 44, 78, 32	378, 44, 78, 16
A1596	37.5x37.5	0.92	378, 44, 78, 32	378, 44, 78, 16	378, 44, 78, 32
D1760	40x40	0.92	378, 44, 78, 44	378, 44, 78, 32	648, 44, 78, 24
A2197	45x45	0.92	378, 44, 78, 44	648, 44, 78, 44	648, 44, 78, 44

Versal Prime Series

		VM1102	VM1302	VM1402	VM1502	VM1802	VM2502	VM2602	VM2702	VM2902
Intelligent Engines	DSP Engines	472	736	1,504	1,312	1,968	3,984	1,880	2,500	3,080
Adaptable Engines	System Logic Cells (K)	352	572	1,002	797	1,968	2,030	1,263	1,805	2,154
	LUTs	161,024	261,376	457,984	364,544	899,840	927,872	577,536	825,000	984,576
Memory	Distributed RAM (Mb)	5	8	14	11	27	28	18	25	30
	Total Block RAM (Mb)	8	16	40	19	34	48	55	74	90
	Total UltraRAM (Mb)	27	47	47	60	130	197	119	169	204
Scalar Engines	Total SRAM Capacity (Mb)	35	63	87	80	164	245	174	243	294
	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, and 256KB TCM w/ECC								
	Memory	256KB On-Chip Memory w/ECC								
	Connectivity	Ethernet (x2); USB 2.0 (x1); UART (x2); SPI (x2); I2C (x2); CAN-FD (x2)								
Foundational Platform	NoC Master/NoC Slave Ports	5	16	16	14	28	28	16	26	26
	DDR Bus Widths	64	128	256	128	256	288	384	384	384
	DDR Memory Controllers	1	2	4	2	4	5	6	6	6
	CCIX & PCIe® w/DMA (CPM)	-	-	-	1 x Gen4x16, CCIX					
	PCI Express®	1 x Gen4x8	2 x Gen4x8	2 x Gen4x8	4 x Gen4x8	4 x Gen4x8	1 x Gen4x8	1 x Gen4x8	2 x Gen4x8	2 x Gen4x8
	Multirate Ethernet MAC	1	2	2	4	4	1	2	2	2
Package Footprint	Package Dimensions	XPIO, HDIO, MIO GTY, GTM	XPIO, HDIO, MIO GTY, GTM	XPIO, HDIO, MIO GTY, GTM	XPIO, HDIO, MIO GTY, GTM	XPIO, HDIO, MIO GTY, GTM	XPIO, HDIO, MIO GTY, GTM	XPIO, HDIO, MIO GTY, GTM	XPIO, HDIO, MIO GTY, GTM	XPIO, HDIO, MIO GTY, GTM
B625	21x21	216, 22, 78, 4, 0								
B1024	31x31	216, 22, 78, 12, 0	216, 44, 78, 16, 0	324, 44, 78, 16, 0						
B1369	35x35		216, 44, 78, 24, 0	324, 44, 78, 24, 0	324, 44, 78, 24, 0					
A1760	40x40		432, 22, 78, 24, 0	648, 22, 78, 24, 0			756, 22, 78, 20, 0			
C1760	40x40			378, 44, 78, 44, 0	378, 44, 78, 44, 0		378, 44, 78, 20, 32	378, 44, 78, 24, 32	378, 44, 78, 24, 32	
D1760	40x40				648, 44, 78, 24, 0					
A2197	45x45				648, 44, 78, 44, 0	648, 44, 78, 16, 16				
A2785	50x50					702, 44, 78, 16, 28	702, 44, 78, 20, 32	702, 44, 78, 32, 44	702, 44, 78, 40, 52	