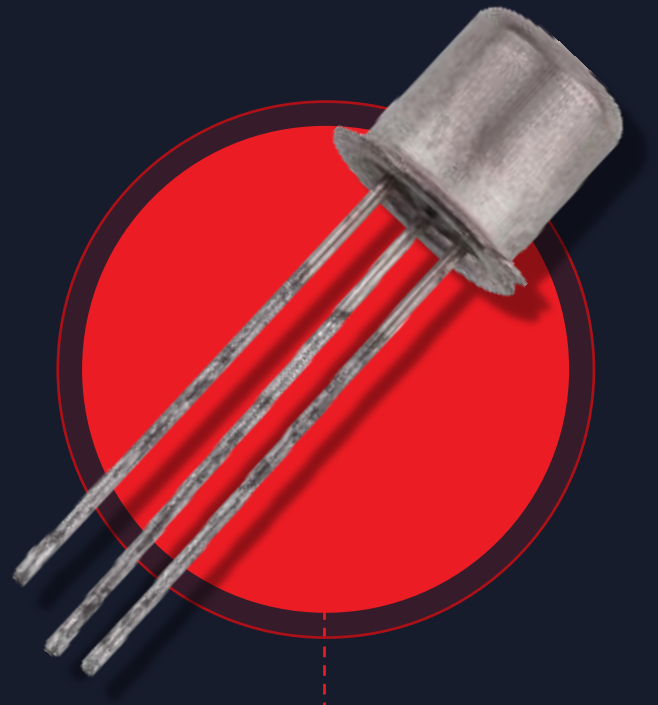
Building the Adaptable Intelligent World

Kirk Saban
Senior Director, Product & Technical Marketing

Under Embargo Until 10/2 @ 9AM PST

Disruptive Innovation



Transistor

1940s

Computing

1970s

Distributed Computing

1990s

Mountains of Unstructured Data

One Architecture Can't Do It Alone

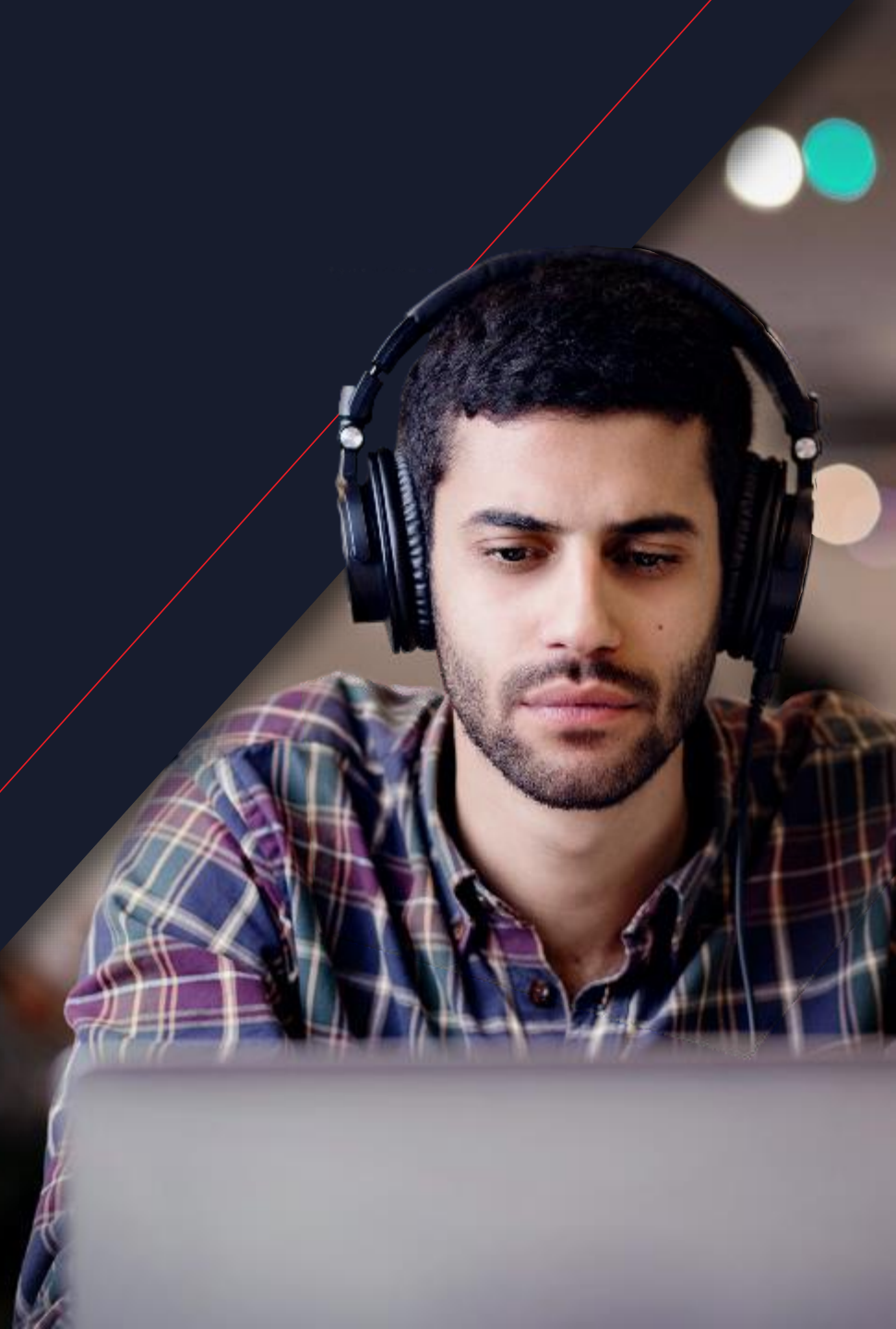
This is the Era of Heterogeneous Compute

Today's Developer Needs

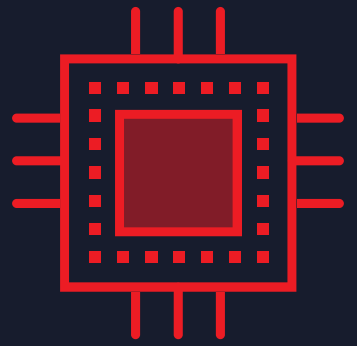
Software programmability

Performance for a diverse range of applications

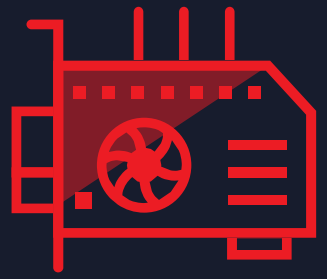
Adaptability to keep pace with rapid innovation



Today's Solutions

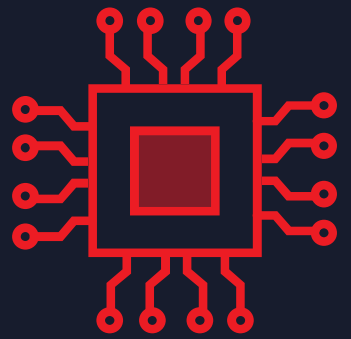


CPUs



Fixed Function Accelerators

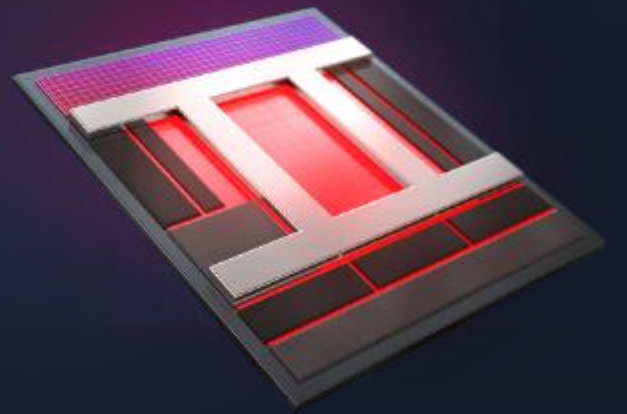
ASICs/ASSPs/GPUs

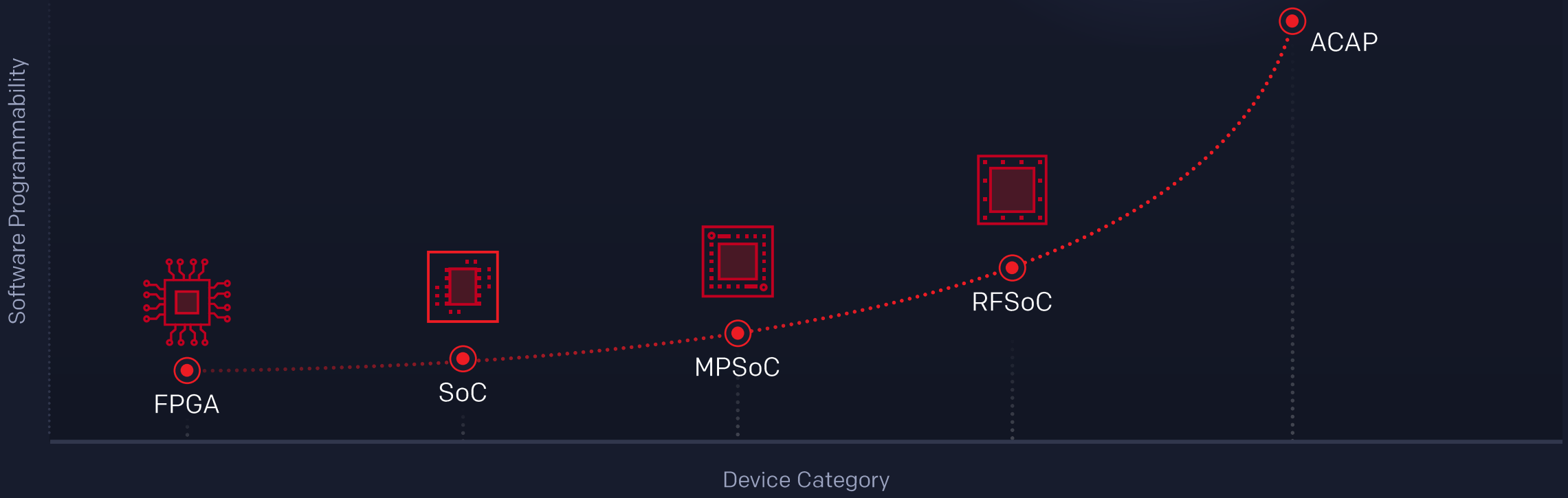


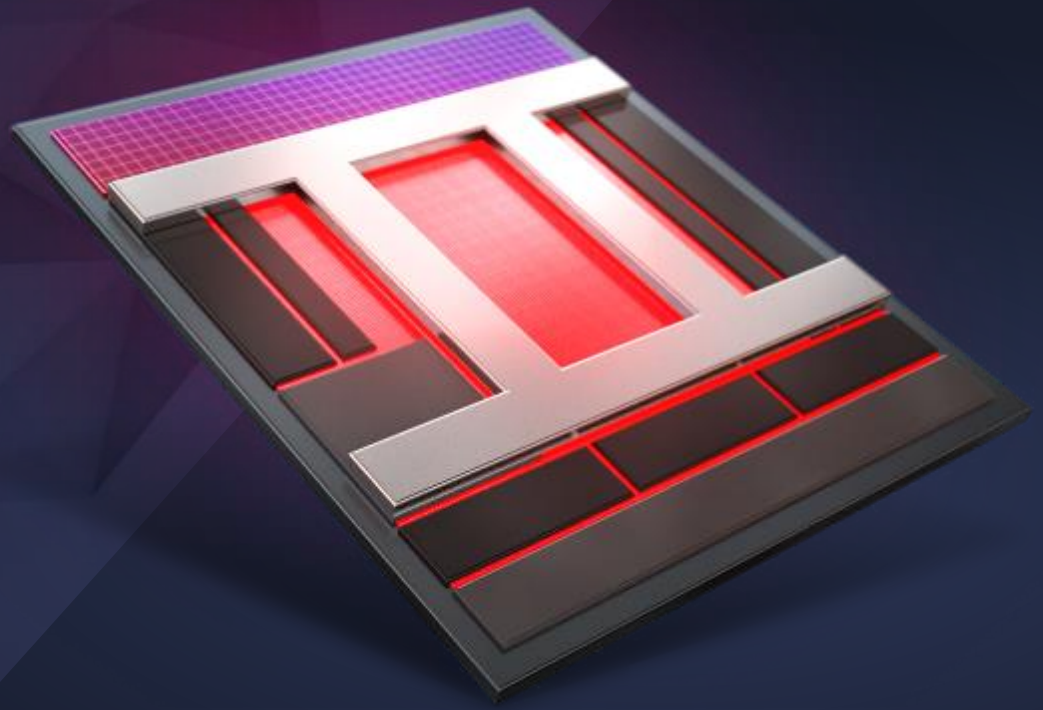
FPGAs

Disruptive Innovation Needed: Enter ACAP

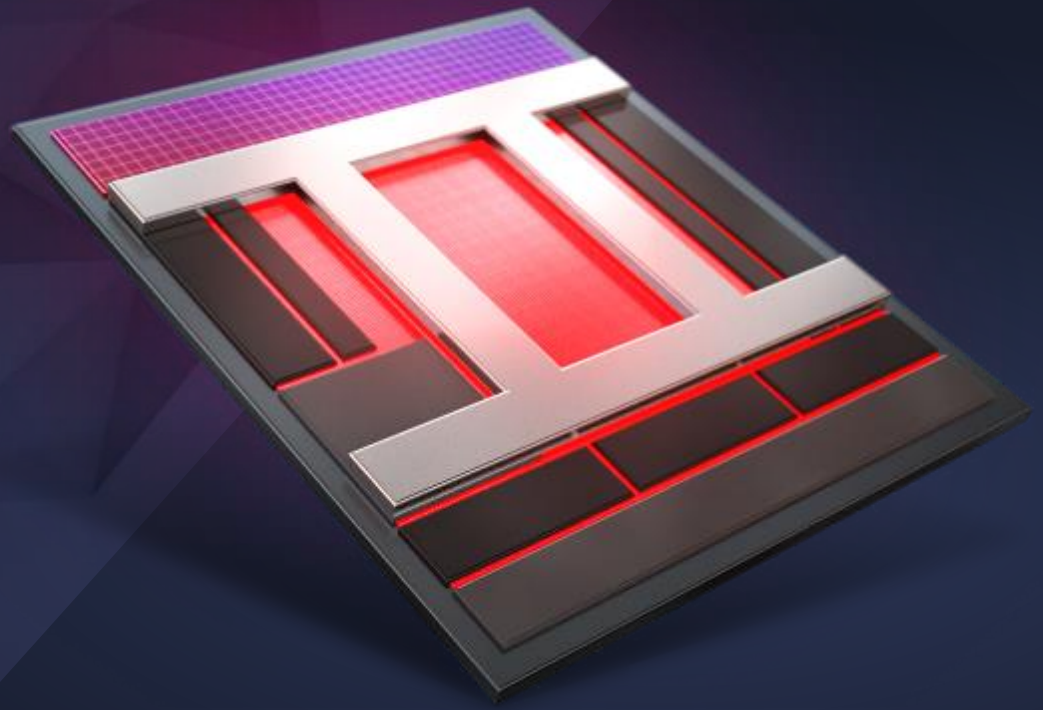
A new class of devices for today's challenges





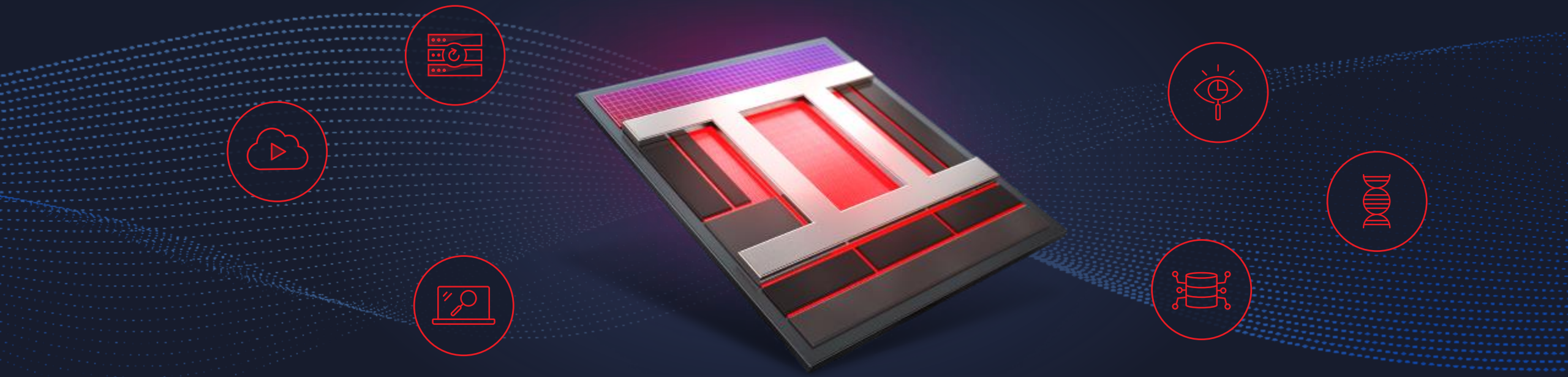


ACAP



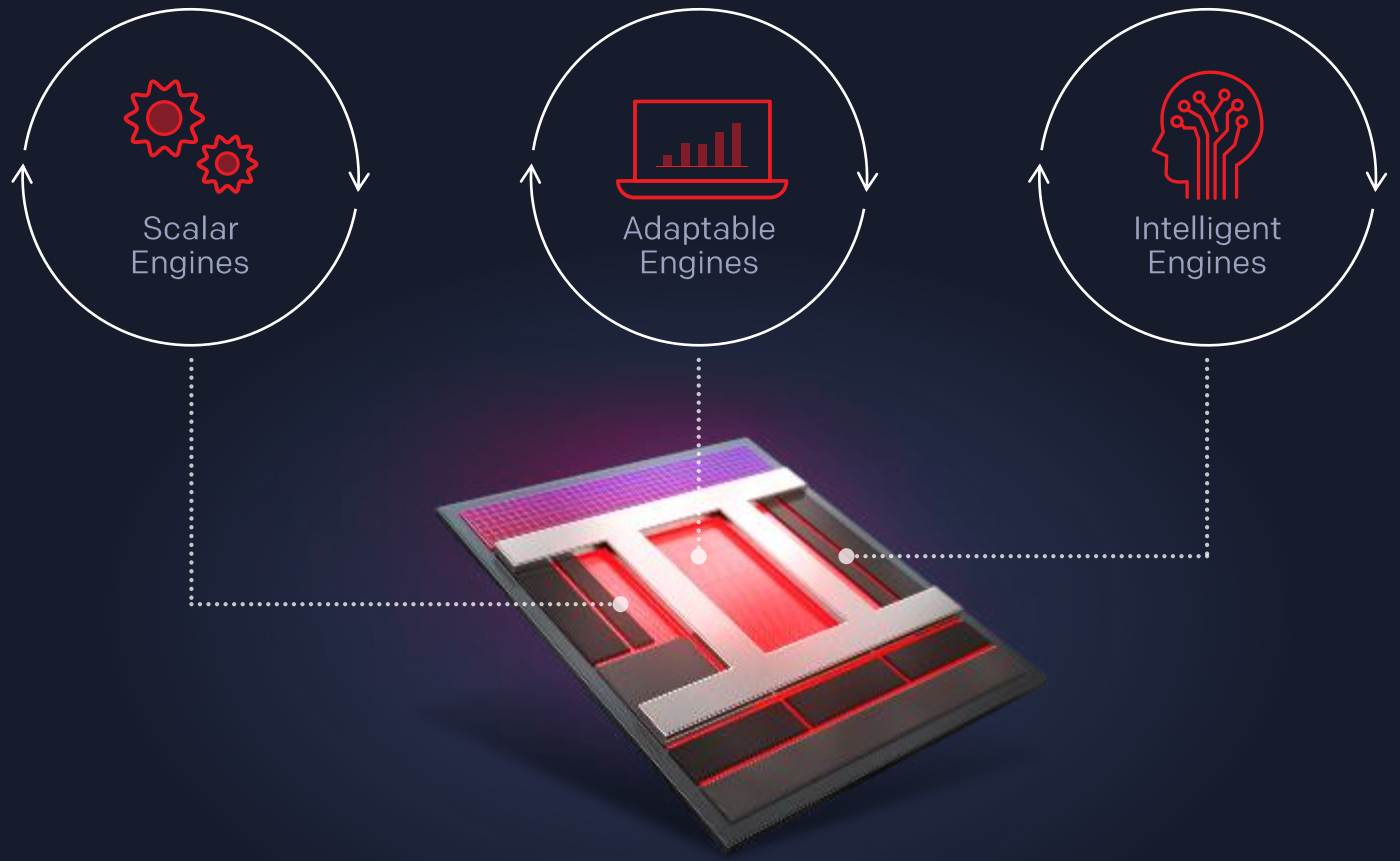
Adaptive Compute Acceleration Platform

Adaptive



Adaptive Hardware for
Domain-specific Applications

Compute Acceleration



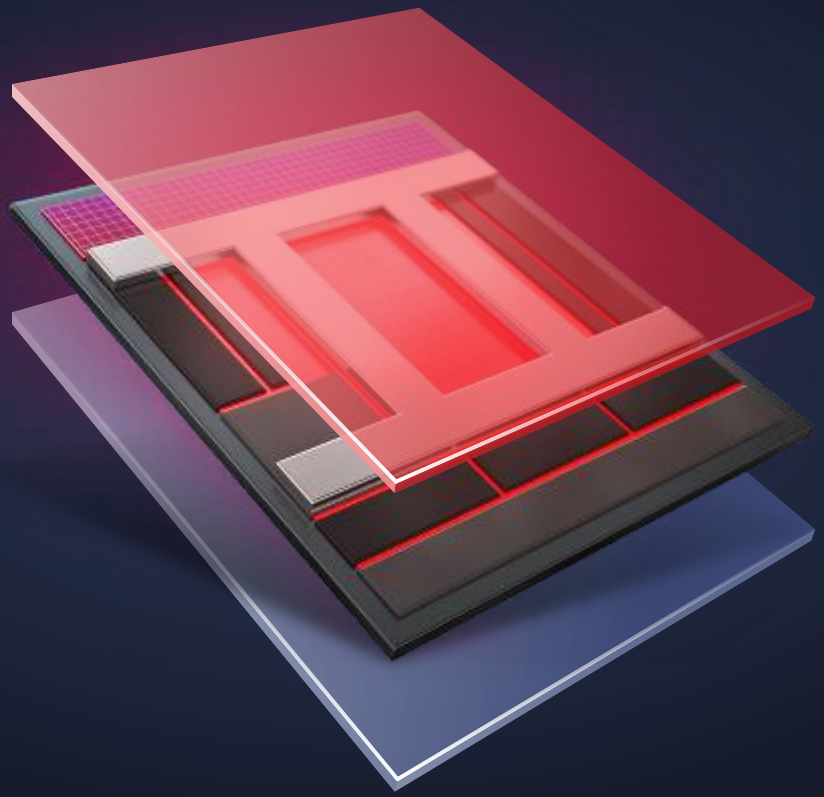
Platform

ENABLING:

Data Scientists

SW App Developers

HW Developers



Development Tools
HW/SW Libraries
Run-time Stack

SW Programmable
Silicon Infrastructure

Introducing the Industry's First ACAP

VERSATILE

UNIVERSAL

XILINX®

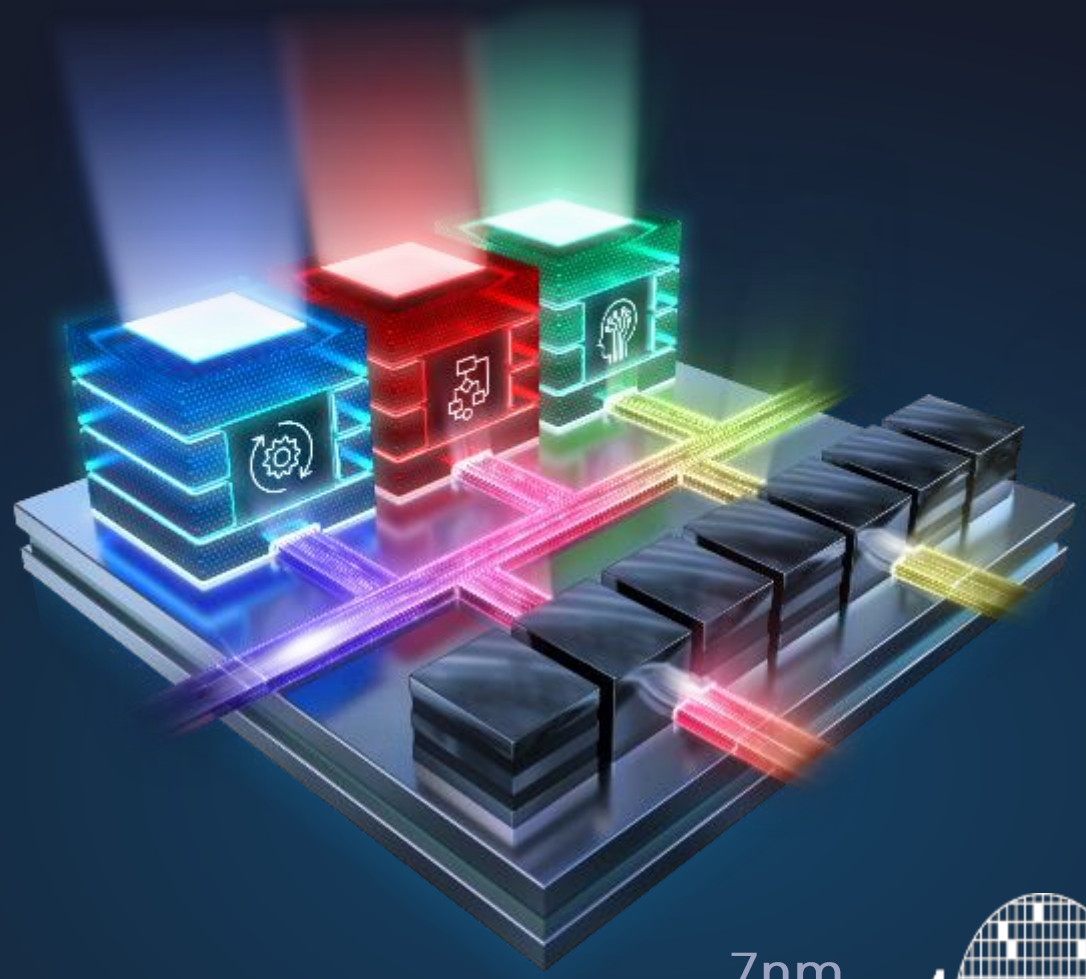
VERSAL™

The Industry's First ACAP

Heterogeneous Acceleration

For Any Application

For Any Developer

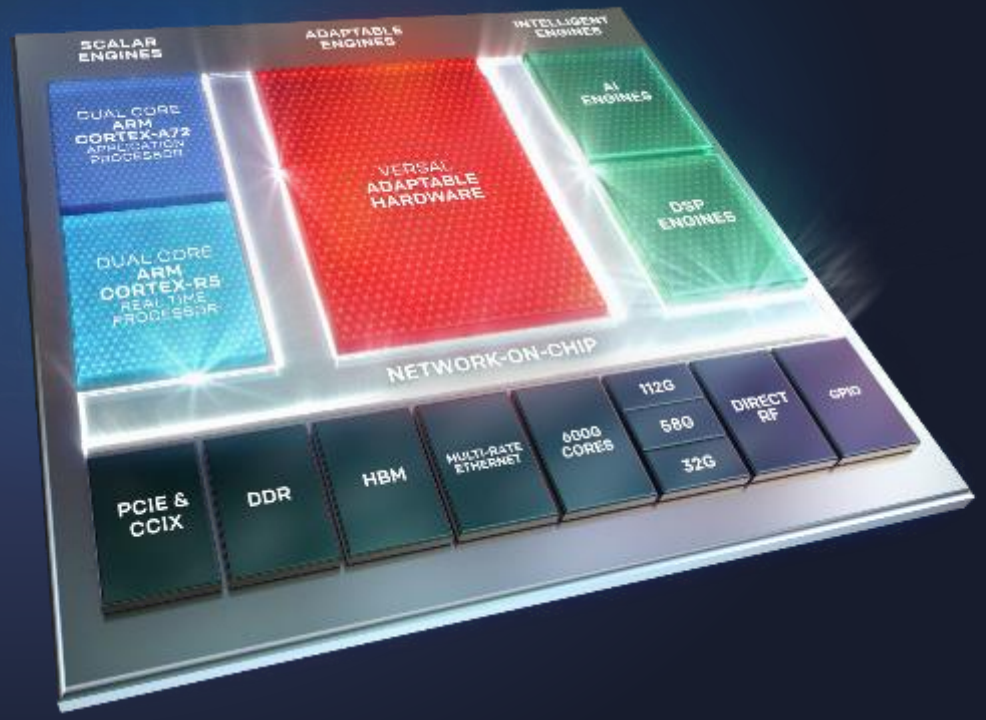


7nm
FinFET

tsmc

Versal ACAP Technology Tour

- Scalar Processing Engines
- Adaptable Hardware Engines
- Intelligent Engines
  SW Programmable, HW Adaptable
- Breakout Integration of Advanced Protocol Engines

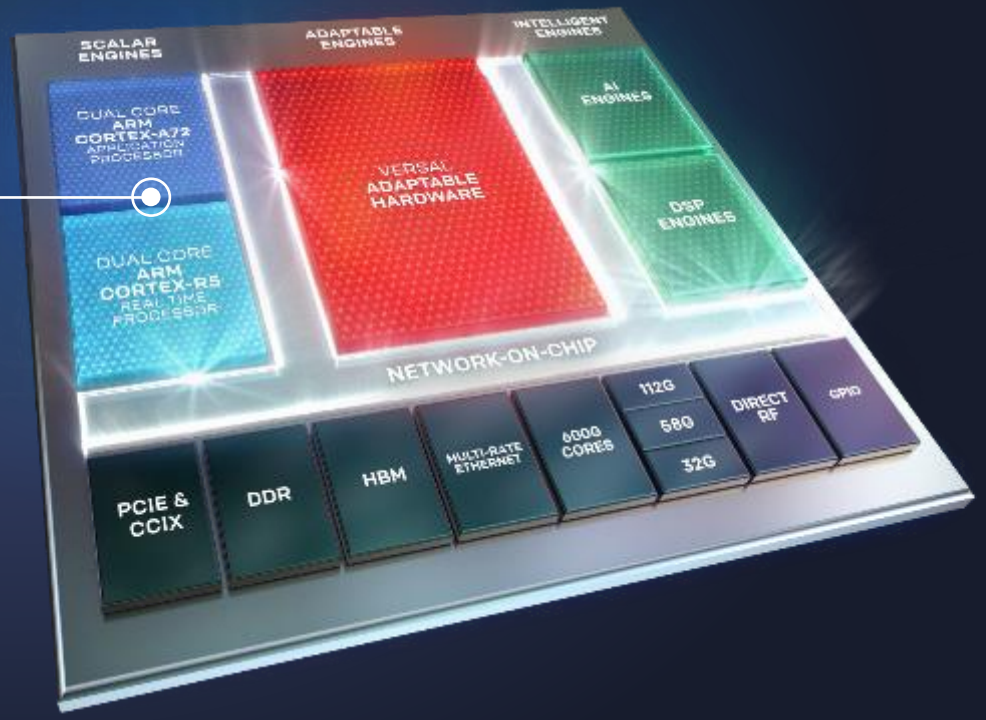


Scalar Processing Engines

Arm Cortex-A72
Application Processor

Arm Cortex-R5
Real-Time Processor

Platform Management Controller

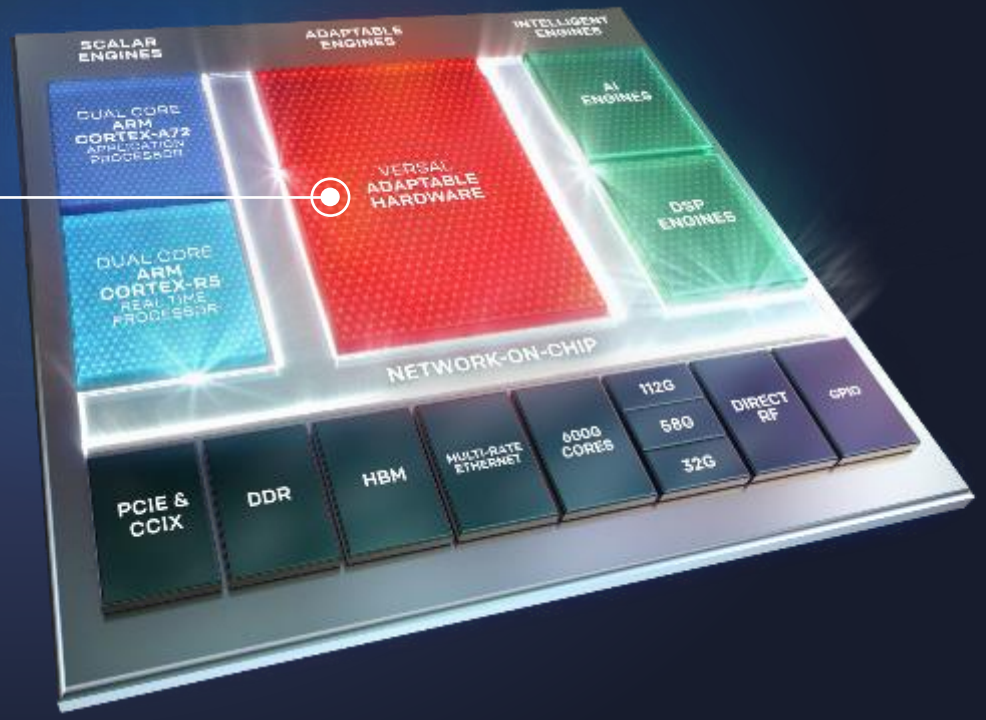


Adaptable Hardware Engines

Re-architected foundational HW fabric for greater compute density

Enables custom memory hierarchy

8X Faster Dynamic Reconfiguration (“on-the-fly”)



Intelligent Engines

DSP Engines

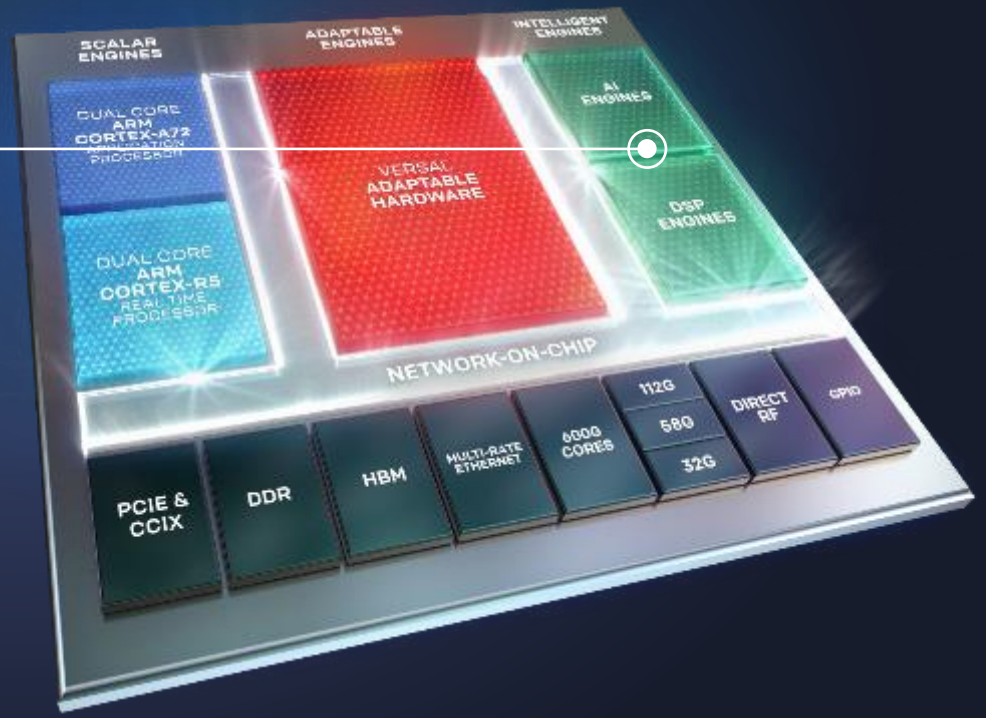
High-precision floating point & low latency

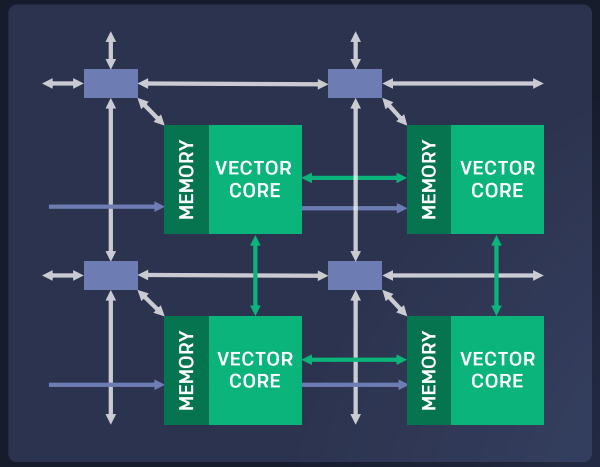
Granular control for customized datapaths

AI Engines

High throughput, low latency, and power efficient

Ideal for AI inference and advanced signal processing

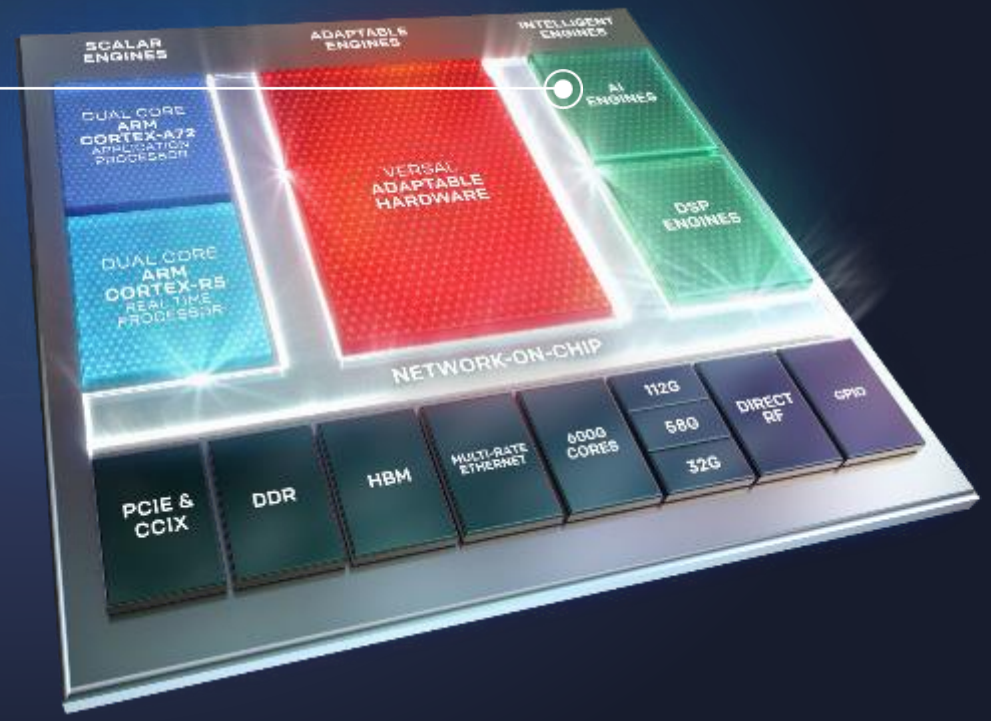


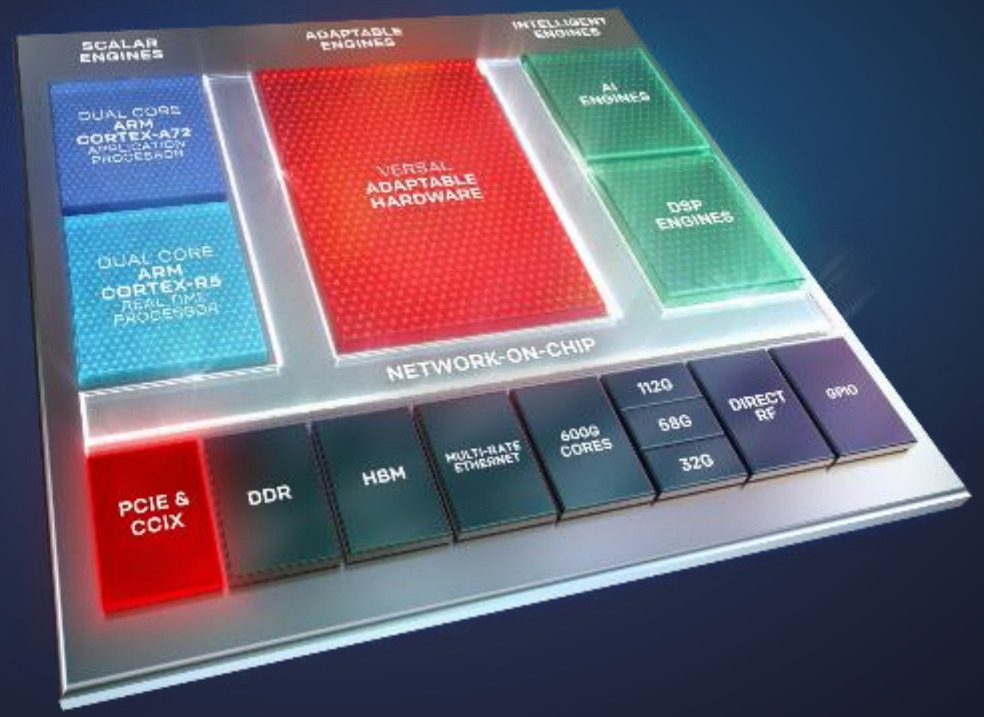


AI Engines

Optimized for AI Inference and Advanced Signal Processing Workloads

- >1GHz VLIW/SIMD vector processor cores
- Massive array of interconnected cores with local memory
- Tightly coupled to adaptable hardware enabling custom memory hierarchy
- Software programmable with hardware adaptability



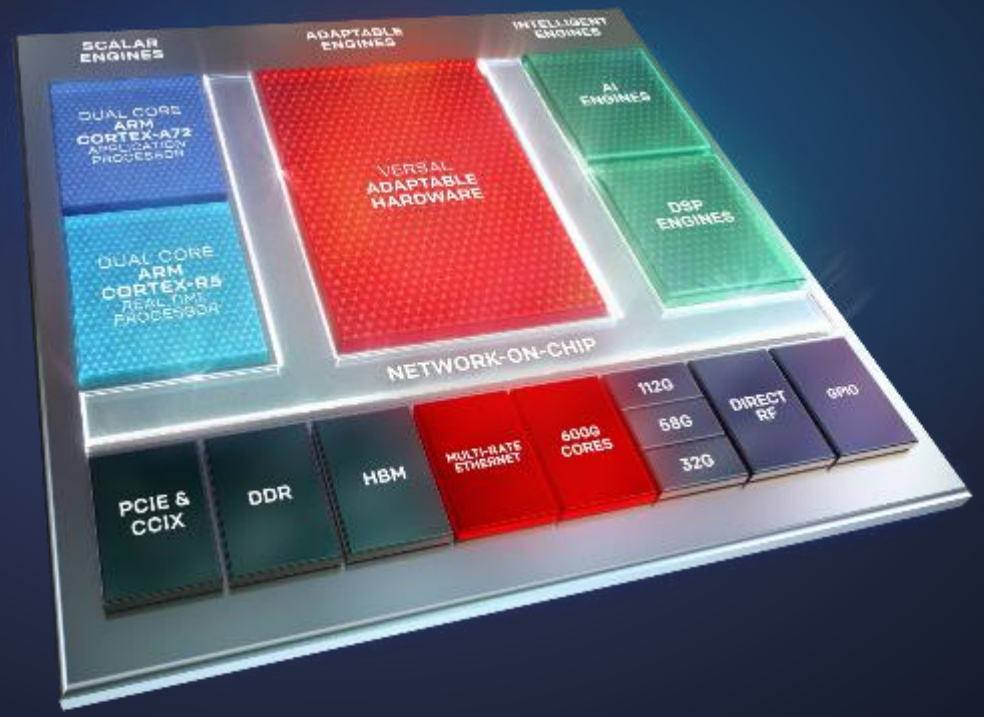


Integrated Host Interfaces

- PCIe Gen4x16
- Integrated AXI-DMA
- CCIX for seamless acceleration of server-class CPUs

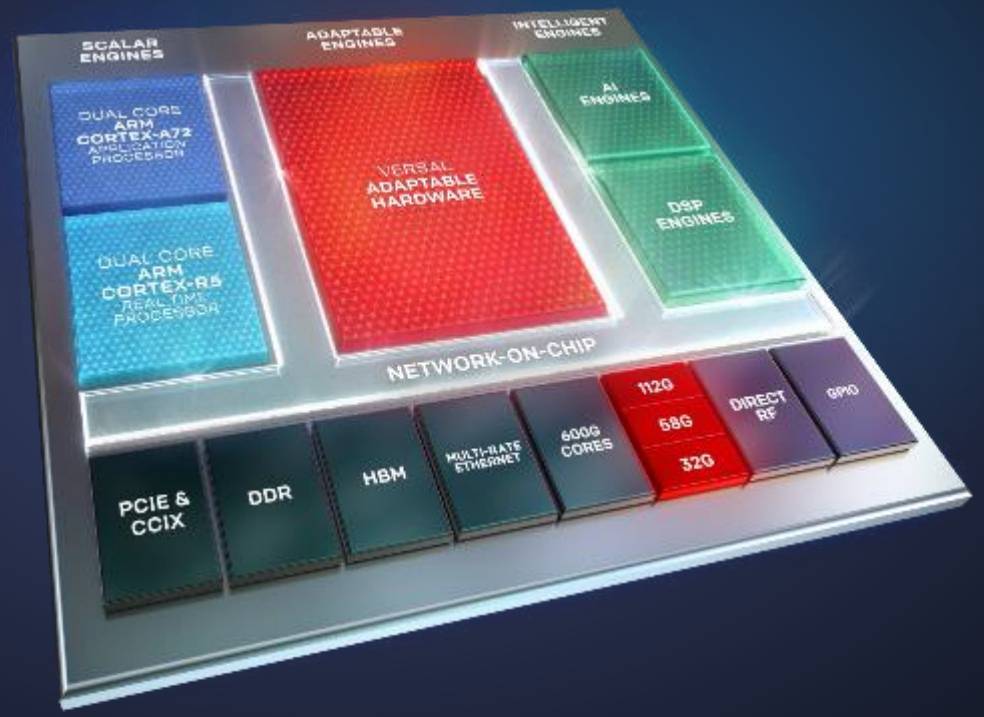
Scalable, Integrated Memory Controllers

- DDR4-3200
- LPDDR4-4266
- High Bandwidth Memory (HBM)



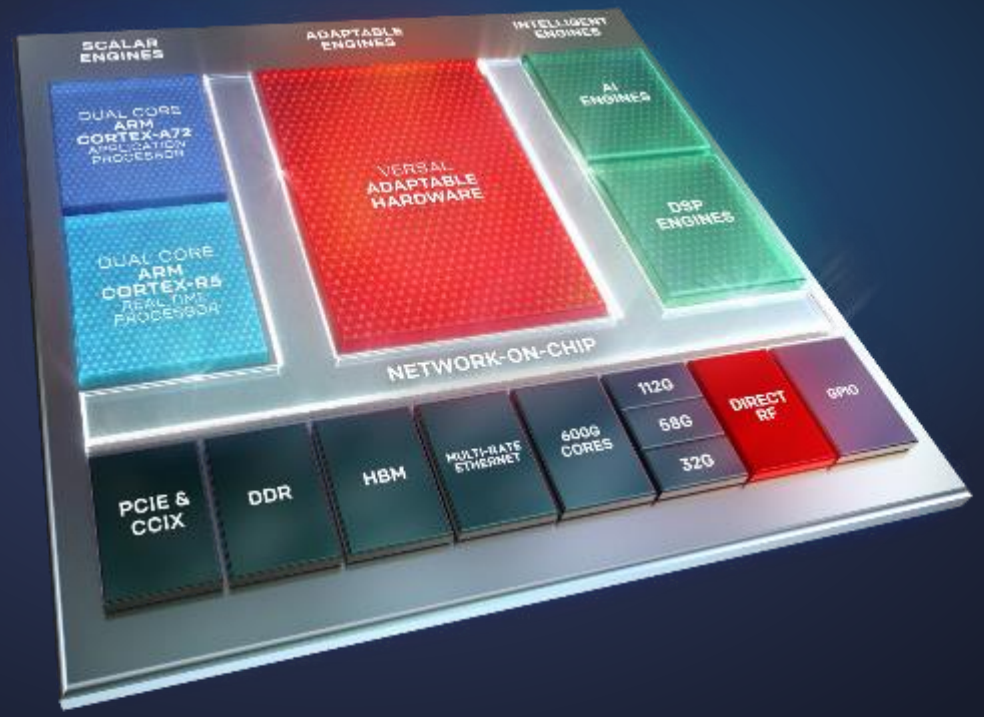
Integrated Protocol Engines

- 100G Multirate Ethernet
- 600G Ethernet and Interlaken
- 600G Cryptographic Engines (AES/IPSEC/MACSEC)



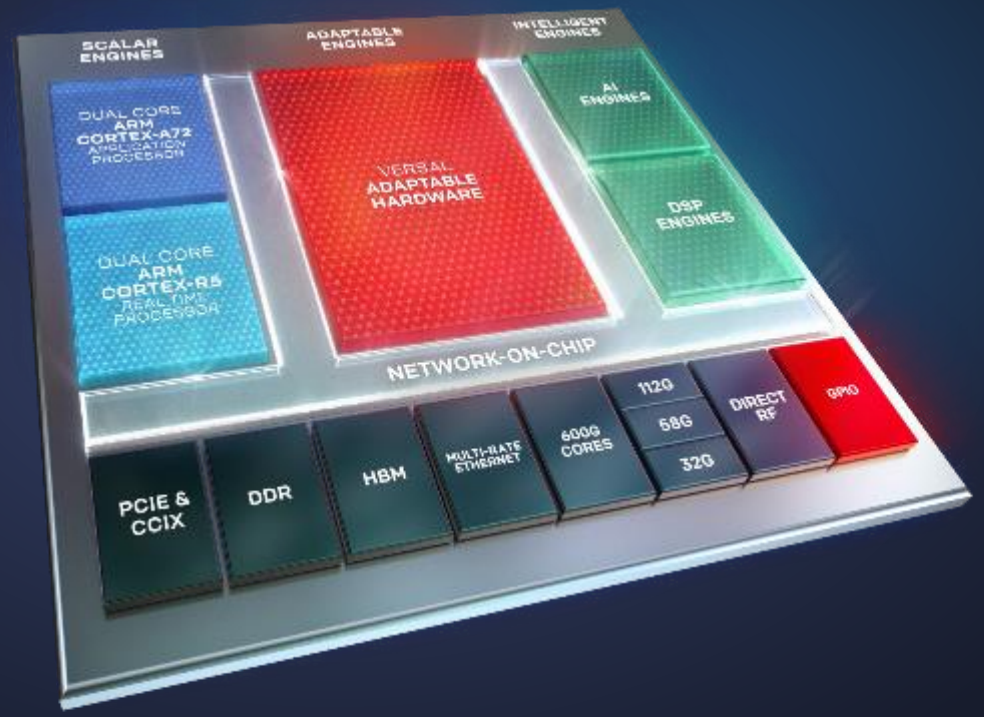
Broadest Range of Transceivers

- 32G power optimized for edge applications
- 58G PAM4 in mainstream devices
- 112G PAM4—Industry's highest performance



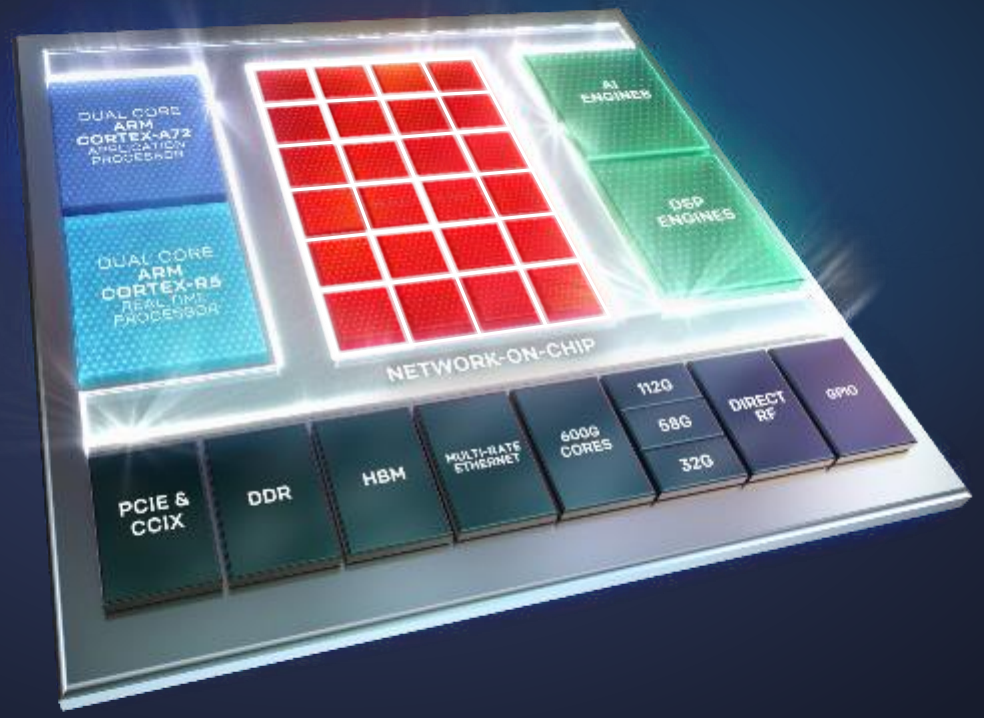
Integrated RF Signal Chain

- Next-generation multi-GSPS direct RF-ADC/DAC
- Integrated DDC/DUC
- SD-FEC for 5G and DOCSIS



Programmable I/O Interfaces

- MIPI D-PHY >3Gb/s for sensors
- NAND and storage-class memory
- LVDS and general-purpose I/O



Network-on-Chip (NoC)

Ease of Use

Inherently software programmable
Available at boot, no place-and-route required

High Bandwidth and Low Latency

Multi-terabit/sec throughput
Guaranteed QoS

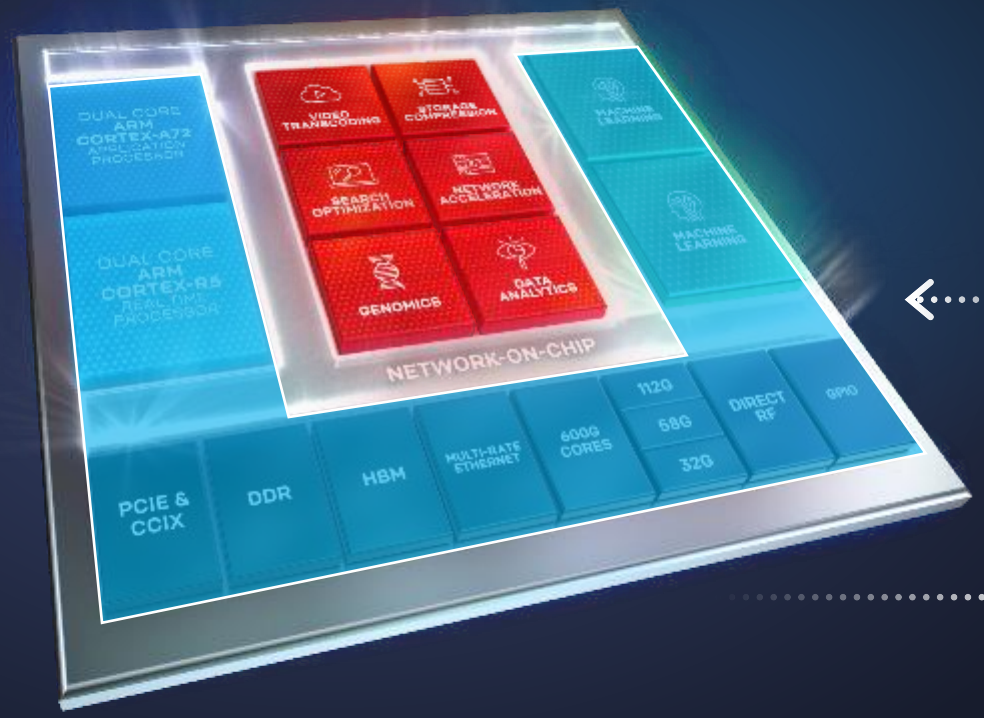
Power Efficiency

8X power efficiency vs. soft implementations
Arbitration across heterogeneous engines

NoC Enables Software Programmability

Data Transfer between Engines and Memory

NoC Enables Software Programmability



Data Transfer between Engines and Memory

Seamless Integration

For Any Application

Versal for Multi-Market Applications

CLOUD

NETWORK

EDGE

Data Center

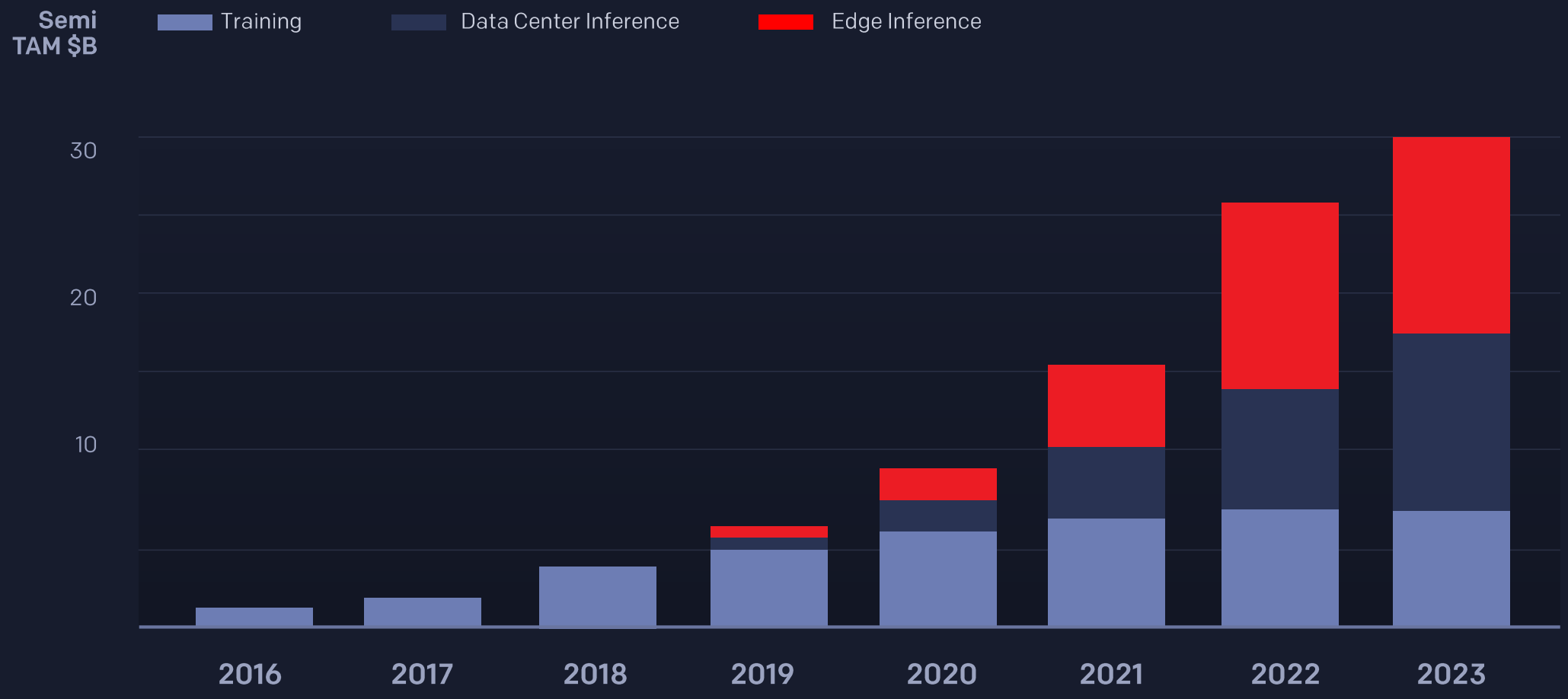
Wired

Wireless

Endpoints

AI ADOPTION ACROSS MARKETS

Pervasiveness of AI and Inference



1: Barclays Research, Company Reports May 2018

XILINX®

VERSAL™

AI RF
Series

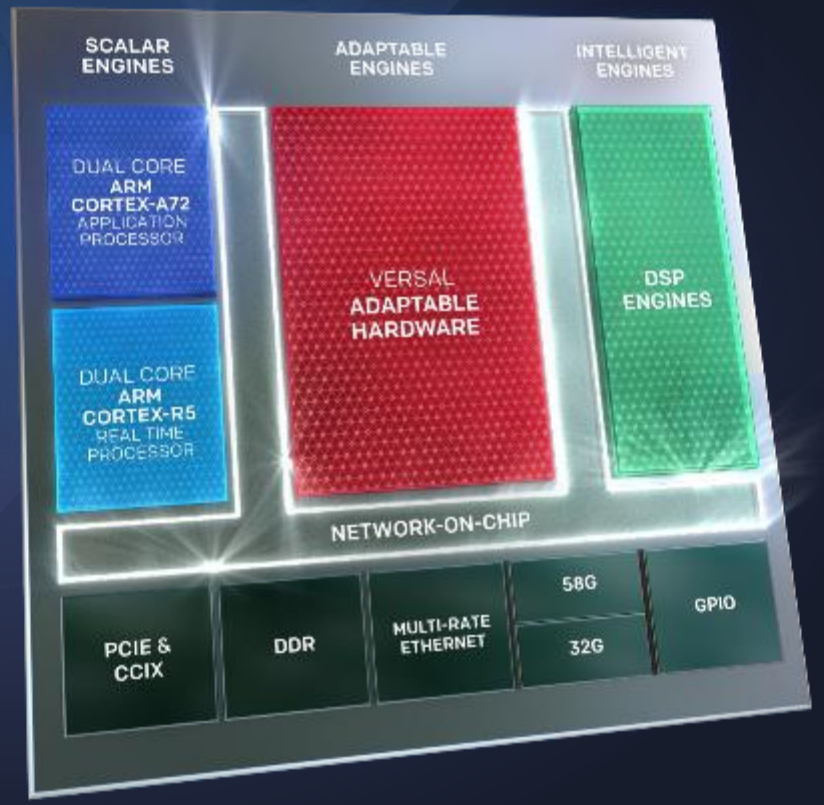
HBM
Series

AI Core
Series

Premium
Series

AI Edge
Series

Prime
Series



VERSAL

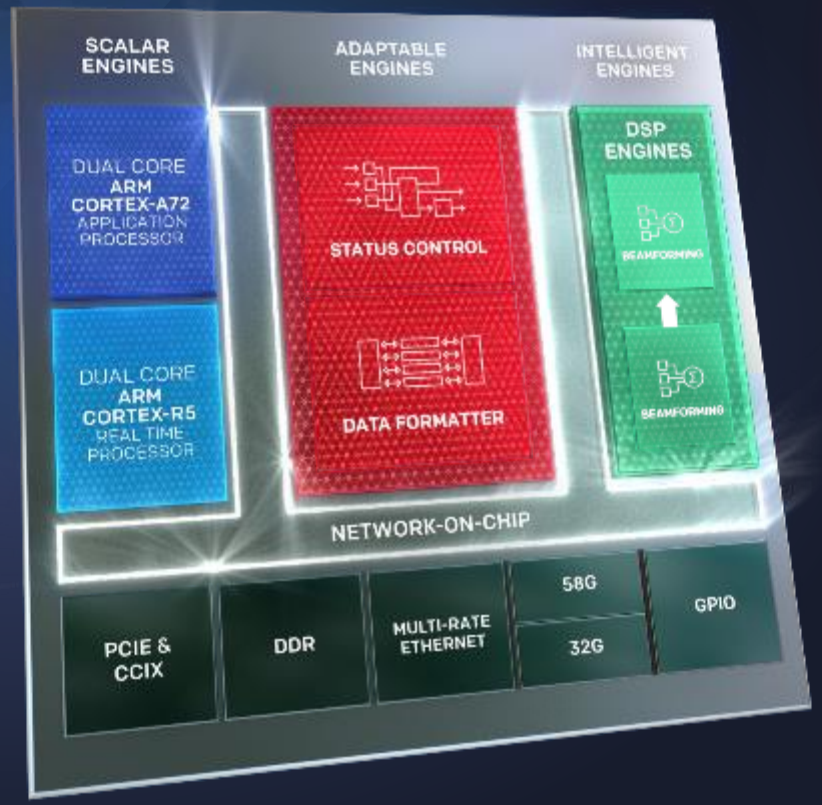
Prime Series

Broad Applicability Across Multiple Markets

Mid-range series in the Versal portfolio

Optimized for connectivity

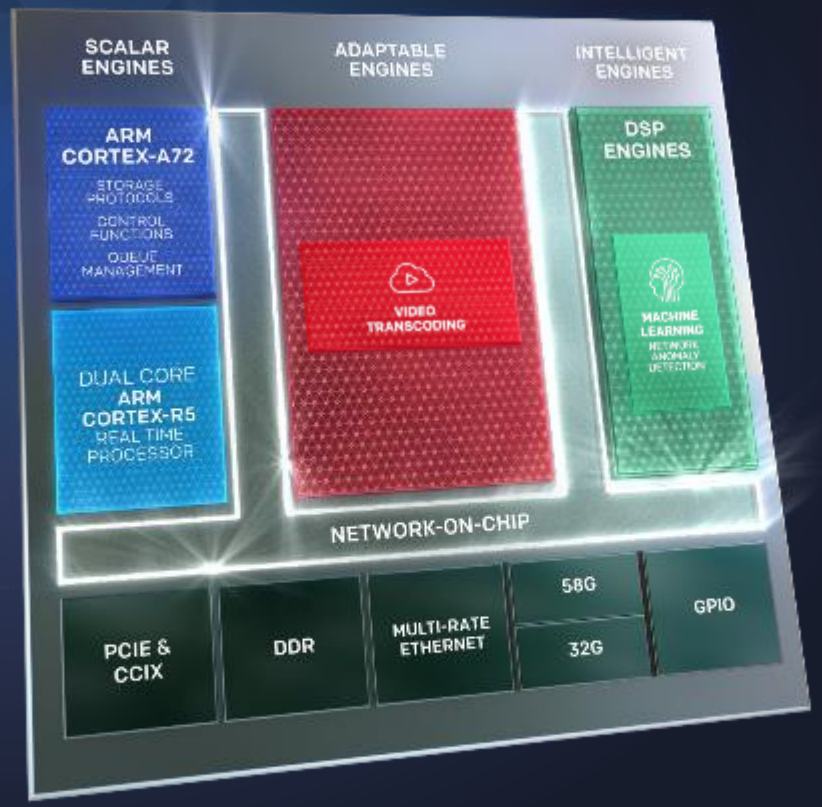
For inline acceleration and diverse workloads

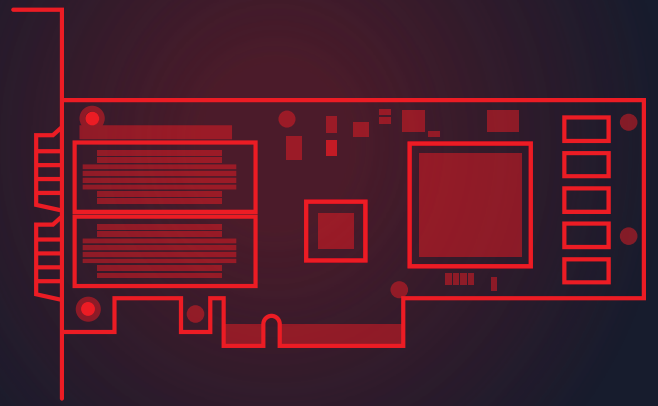


Versal Prime Series

Intelligent Engines in Radar Beamforming

DSP Engines for diverse, fixed & floating point signal processing workloads

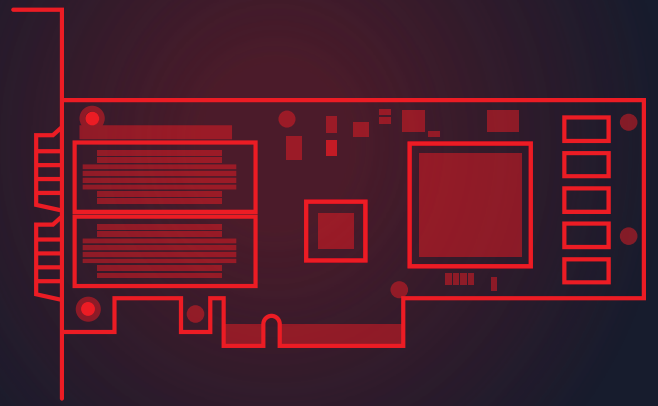




Network Attached Acceleration

Support for multiple network-attached workloads

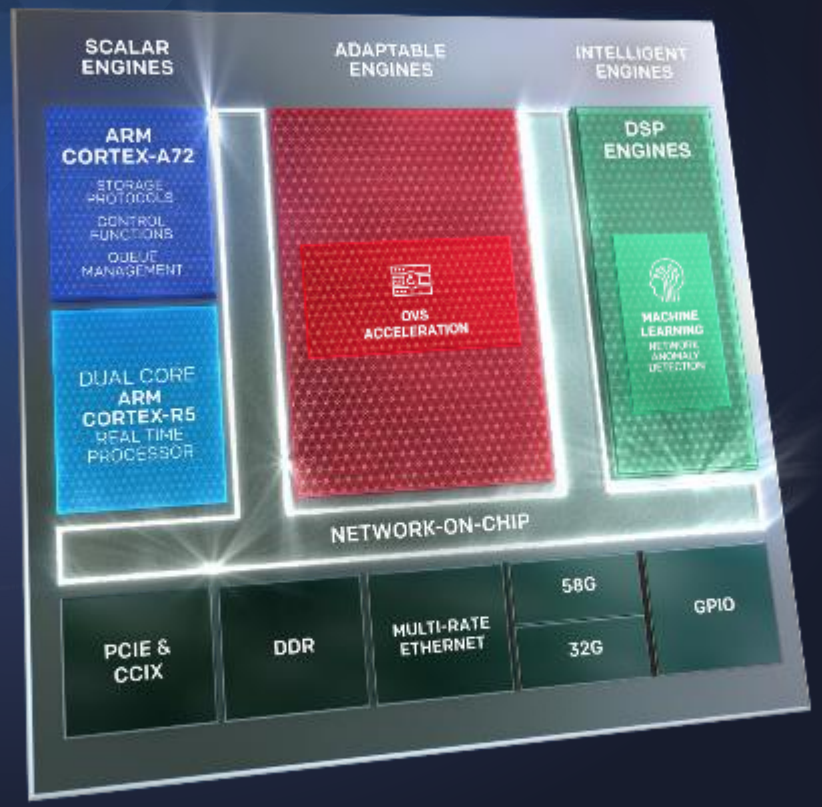
Ability to combine workloads with AI inference

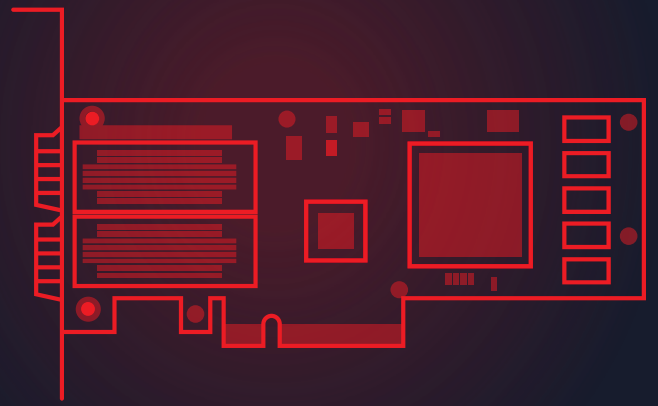


Network Attached Acceleration

Support for multiple network-attached workloads

Ability to combine workloads with AI inference



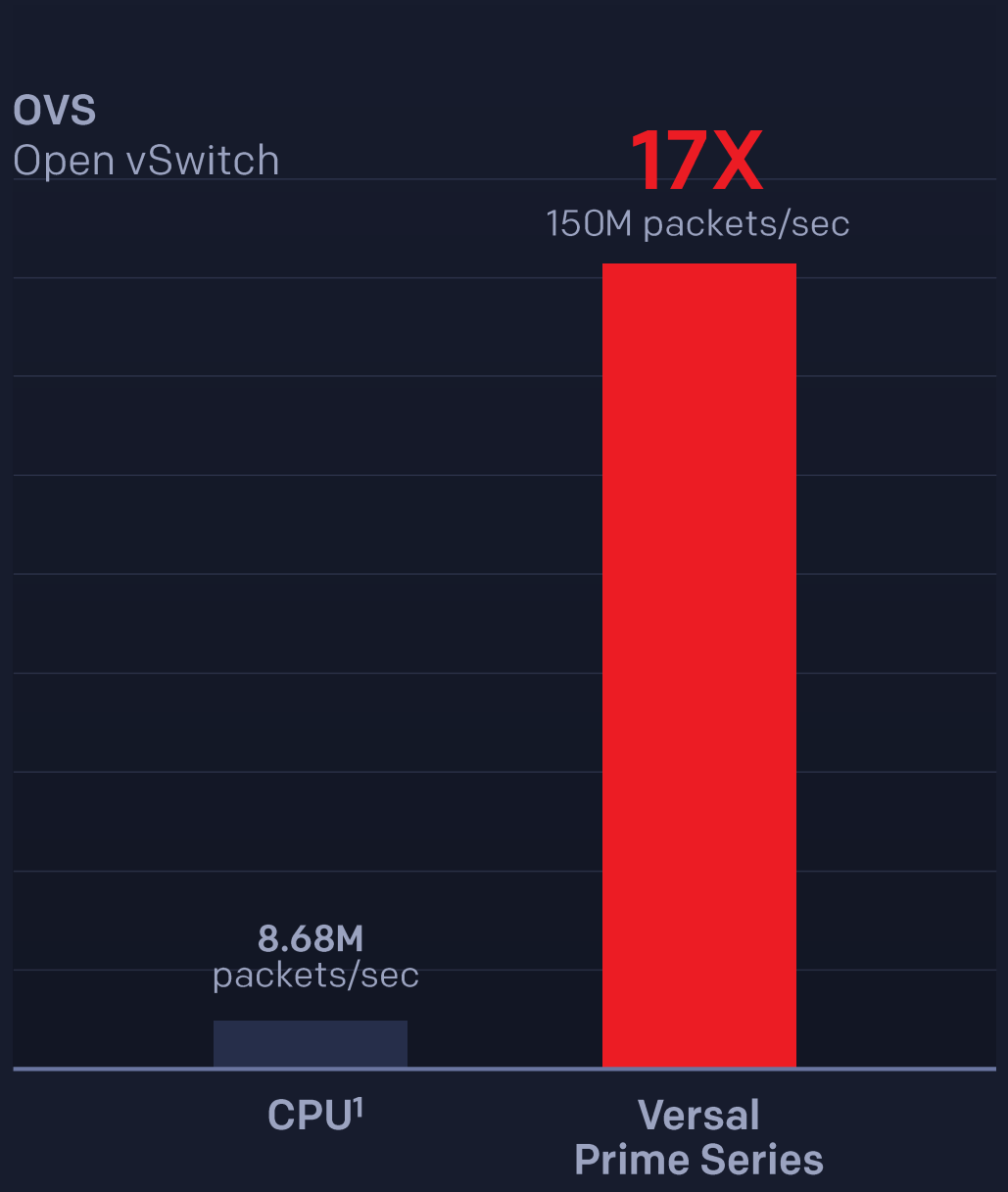


Network Attached Acceleration

Support for multiple network-attached workloads

Ability to combine workloads with AI inference

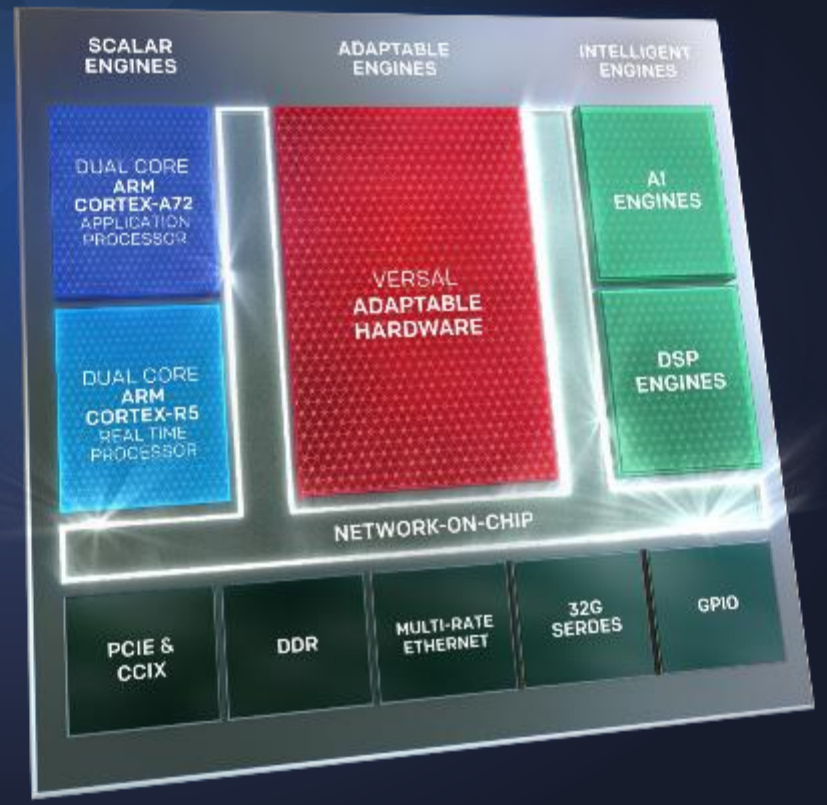
Network Attached Accelerator Workloads



1: Assuming 4 Xeon cores at 2.17/Mp/s per core for zero packet loss; Source: "Red Hat's Perspective on OVS HW Offload Status ", Open vSwitch Fall Conference 2017

More Applications with Versal Prime Series

- Communications Test Equipment
- Data Center Network and Storage Acceleration
- Nx100G Ethernet and OTN Networking
- Broadcast Switches
- Medical Imaging
- Avionics Control



VERSAL

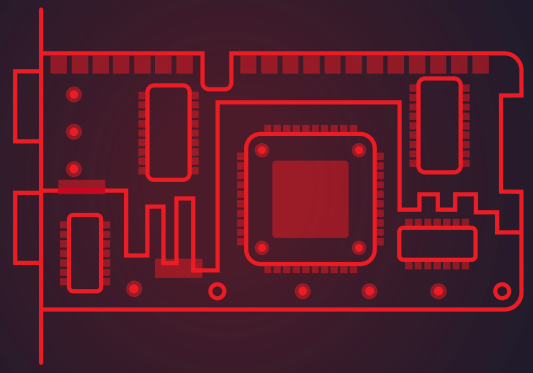
AI Core Series

Breakthrough AI Inference Throughput

Portfolio's highest throughput for low latency inference

Optimized for cloud, networking, and autonomous applications

For highest dynamic range of AI and workload acceleration

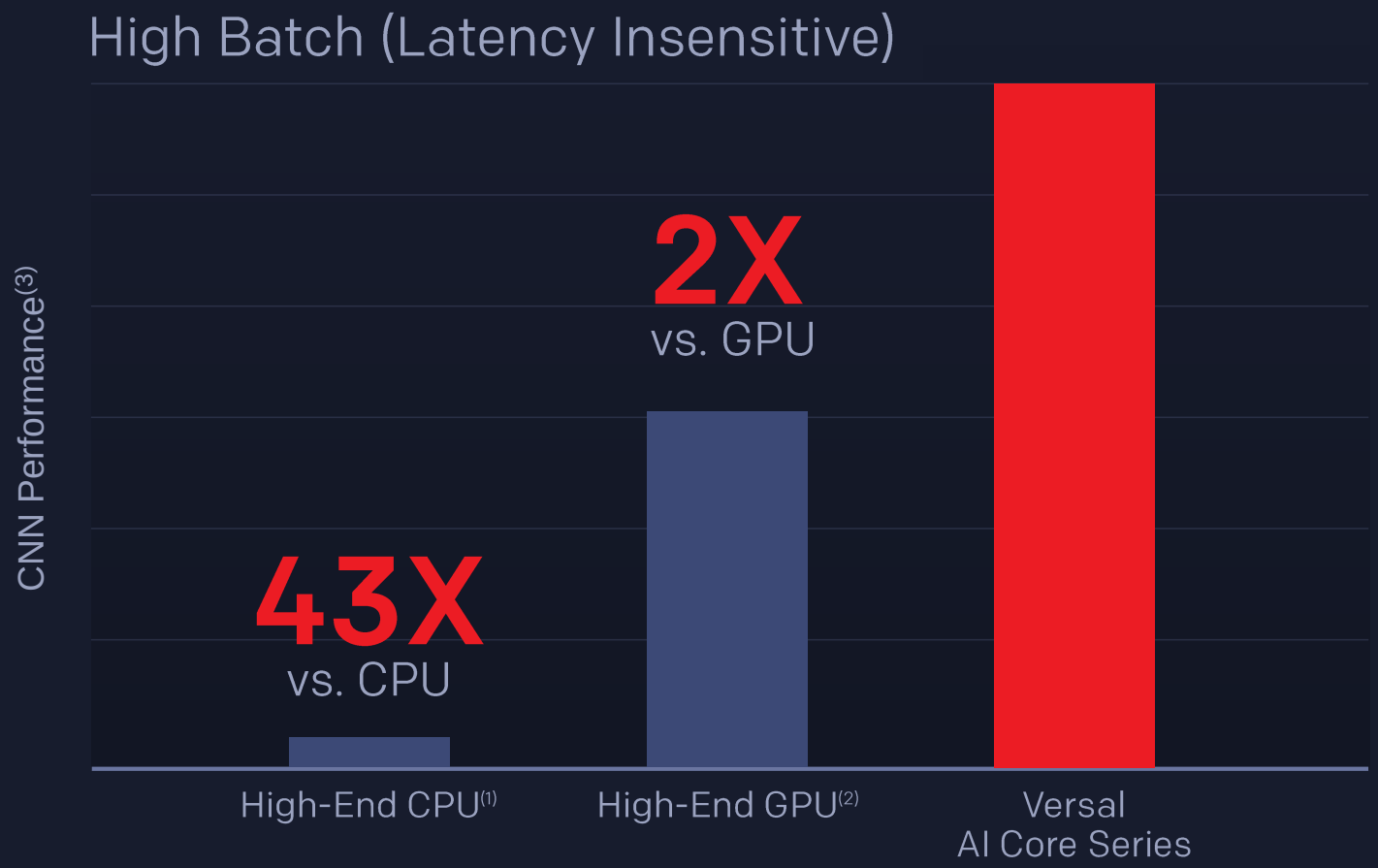


VERSAL AI Core Series

AI Engines and Adaptable Hardware Maximize AI Inference

Massive bandwidth across heterogeneous engines for optimal performance

AI Compute Compared to CPUs and GPUs



Inference Performance

Leveraging AI Engines

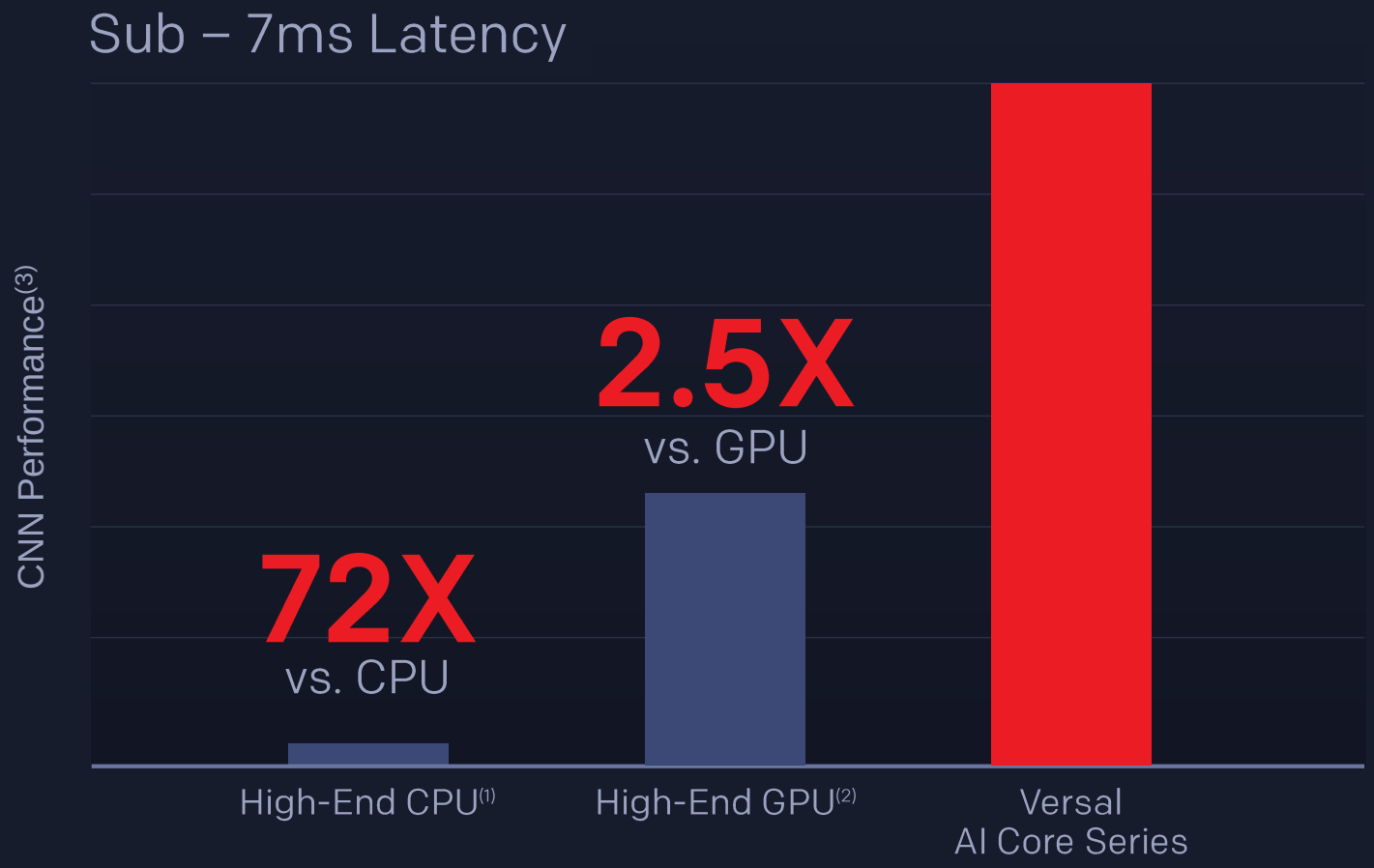
Majority of Adaptable & Scalar Engines available for Whole Application Acceleration

Whole Application Acceleration

Ability to combine workloads and trade-off resources between AI and workload

(1) Measured on EC2 Xeon Platinum 8124 Skylake, c5.18xlarge AWS instance, Intel Caffe: https://github.com/intel/caffe
(2) V100 numbers taken from Nvidia Technical Overview, "Deep Learning Platform, Giant Leaps in Performance and Efficiency for AI Services"
(3) GoogLeNet V1 throughput (Img/sec)

AI Compute Compared to CPUs and GPUs



Inference Performance

Leveraging AI Engines

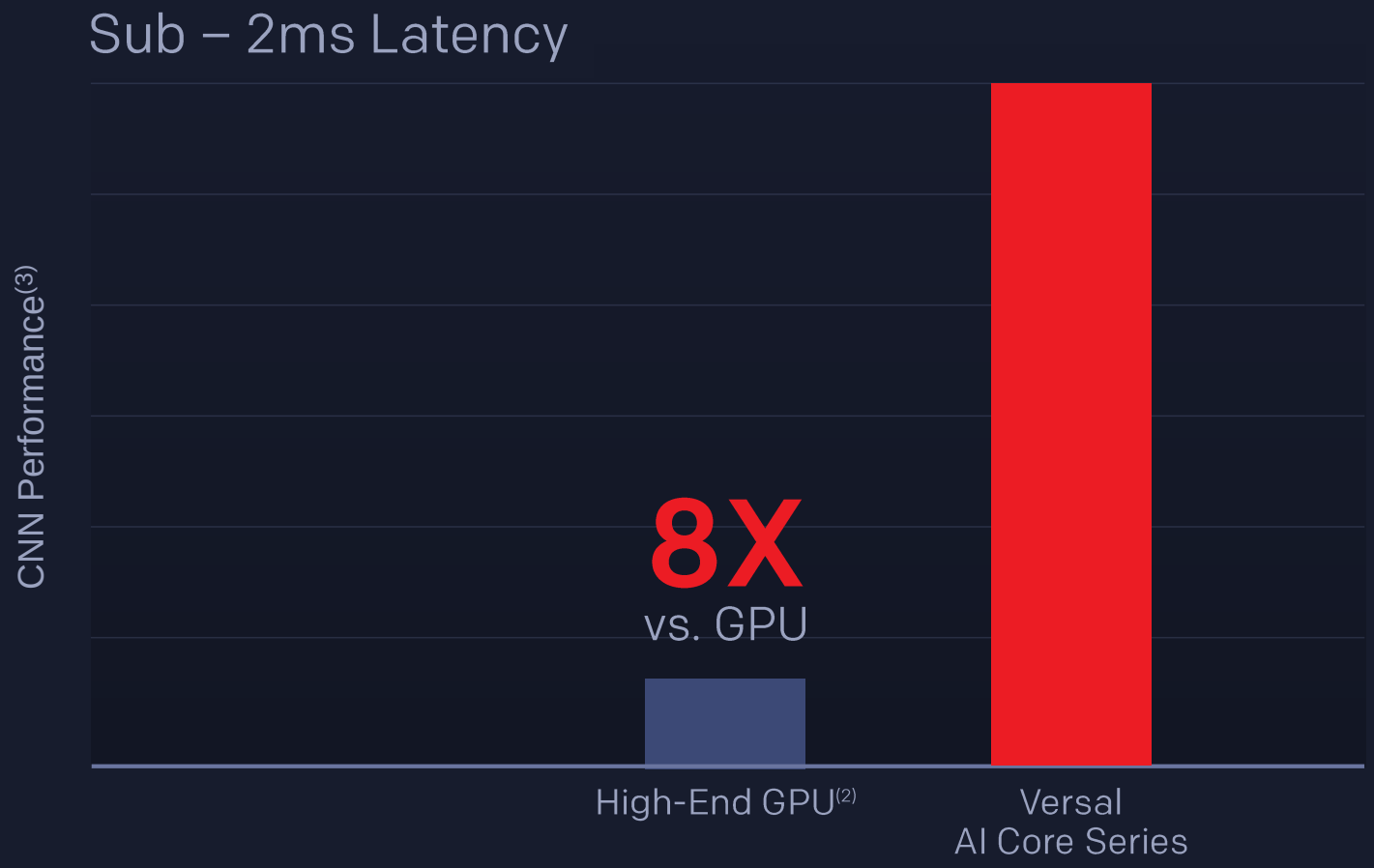
Majority of Adaptable & Scalar Engines available for Whole Application Acceleration

Whole Application Acceleration

Ability to combine workloads and trade-off resources between AI and workload

(1) Measured on EC2 Xeon Platinum 8124 Skylake, c5.18xlarge AWS instance, Intel Caffe: https://github.com/intel/caffe
(2) V100 numbers taken from Nvidia Technical Overview, "Deep Learning Platform, Giant Leaps in Performance and Efficiency for AI Services"
(3) GoogLeNet V1 throughput (Img/sec)

AI Compute Compared to CPUs and GPUs



Inference Performance

Leveraging AI Engines

Majority of Adaptable & Scalar Engines available for Whole Application Acceleration

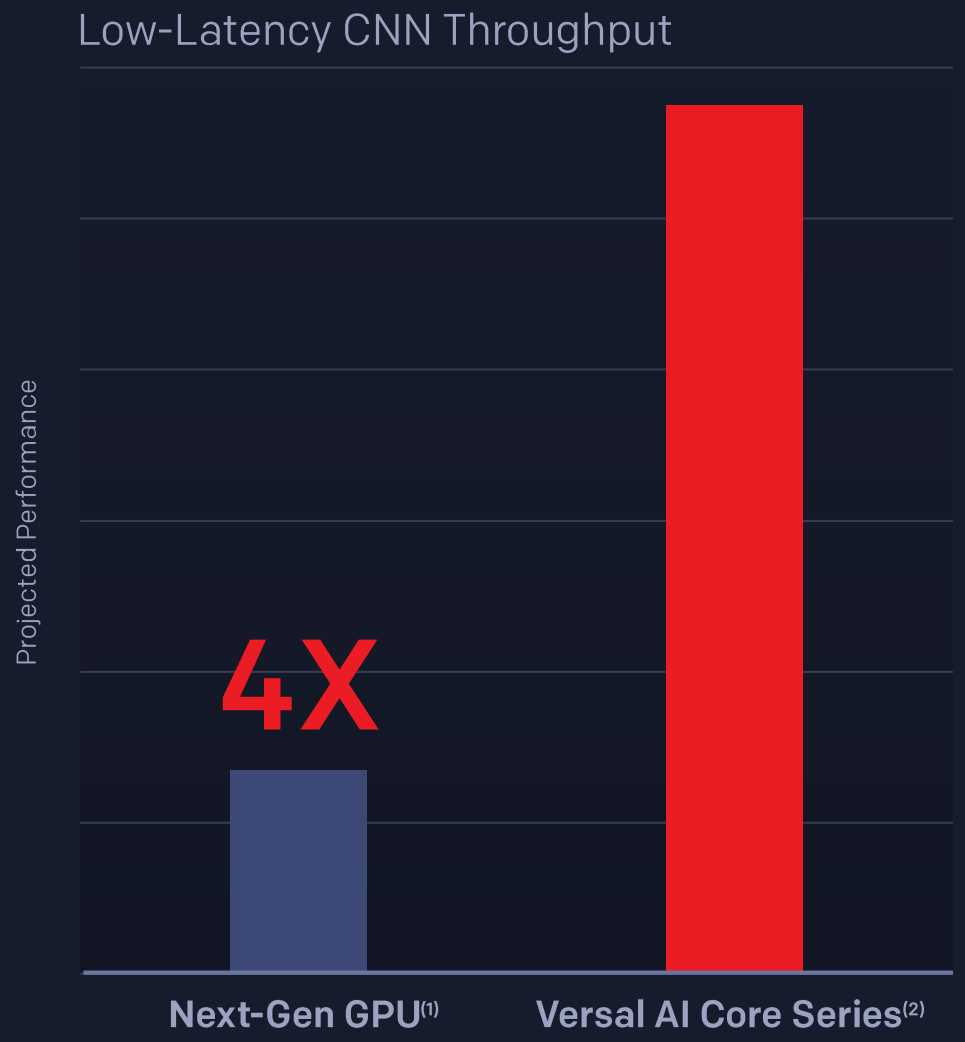
Whole Application Acceleration

Ability to combine workloads and trade-off resources between AI and workload

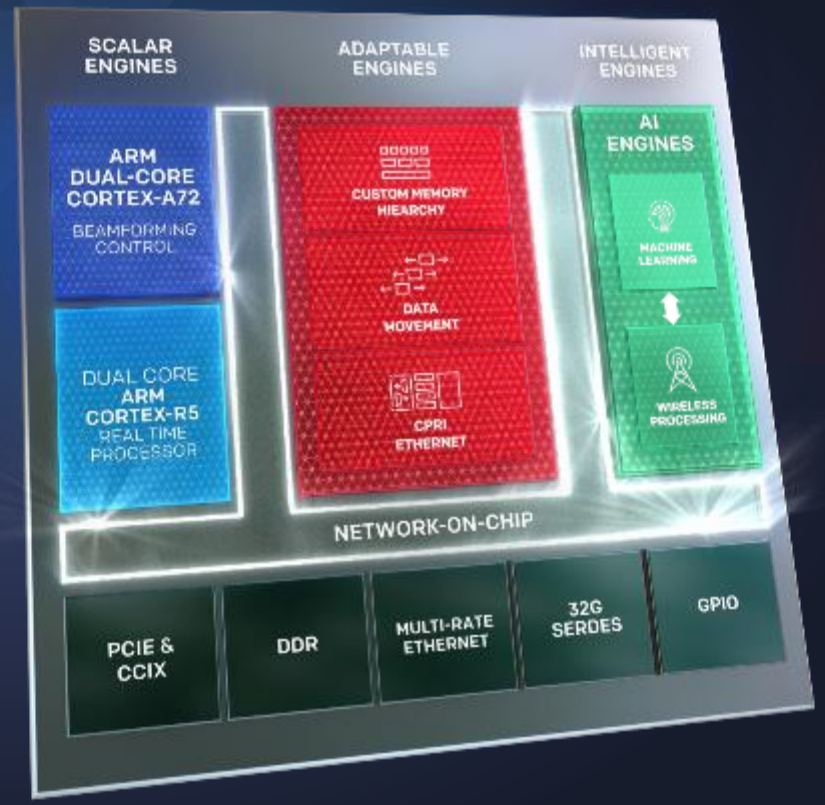
(2) V100 numbers taken from Nvidia Technical Overview, "Deep Learning Platform, Giant Leaps in Performance and Efficiency for AI Services"
(3) GoogLeNet V1 throughput (Img/sec)

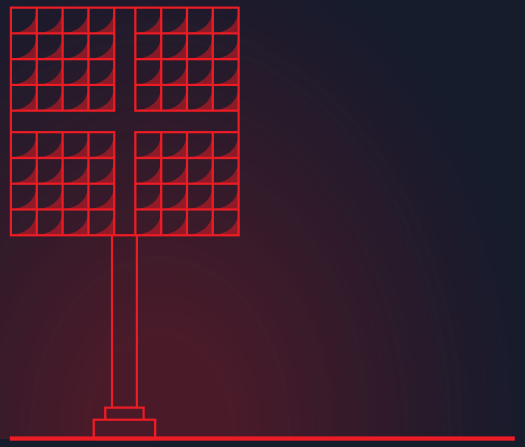
AI Inference Power Efficiency Advantage over GPUs

4X the Throughput in the Same Power Envelope (75W)



(1) 12-nanometer T4 GPU device, Projected Batch=1 performance based on currently available vendor benchmarks
(2) 7-nanometer Xilinx Versal ACAP device, Latency ~500us





VERSAL AI Core Series

For 5G Wireless Compute with AI Inference

AI Engines have ability to combine inference with wireless compute

Mixed Workloads on AI Engine

UNDER DEVELOPER CONTROL

Combine Workloads on AI Engine

5X

MAX Wireless Compute[1]

- Systolic Array
- CloudRAN
- Baseband Processing

8X

MAX AI Inference[1]

- Self-Organizing Networks
- Anomaly Detection
- Scheduling

1: Compared to UltraScale+

Mixed Workloads on AI Engine

UNDER DEVELOPER CONTROL

Combine Workloads on AI Engine

5X

MAX Wireless Compute[1]

- Systolic Array
- CloudRAN
- Baseband Processing

8X

MAX AI Inference[1]

- Self-Organizing Networks
- Anomaly Detection
- Scheduling

1: Compared to UltraScale+

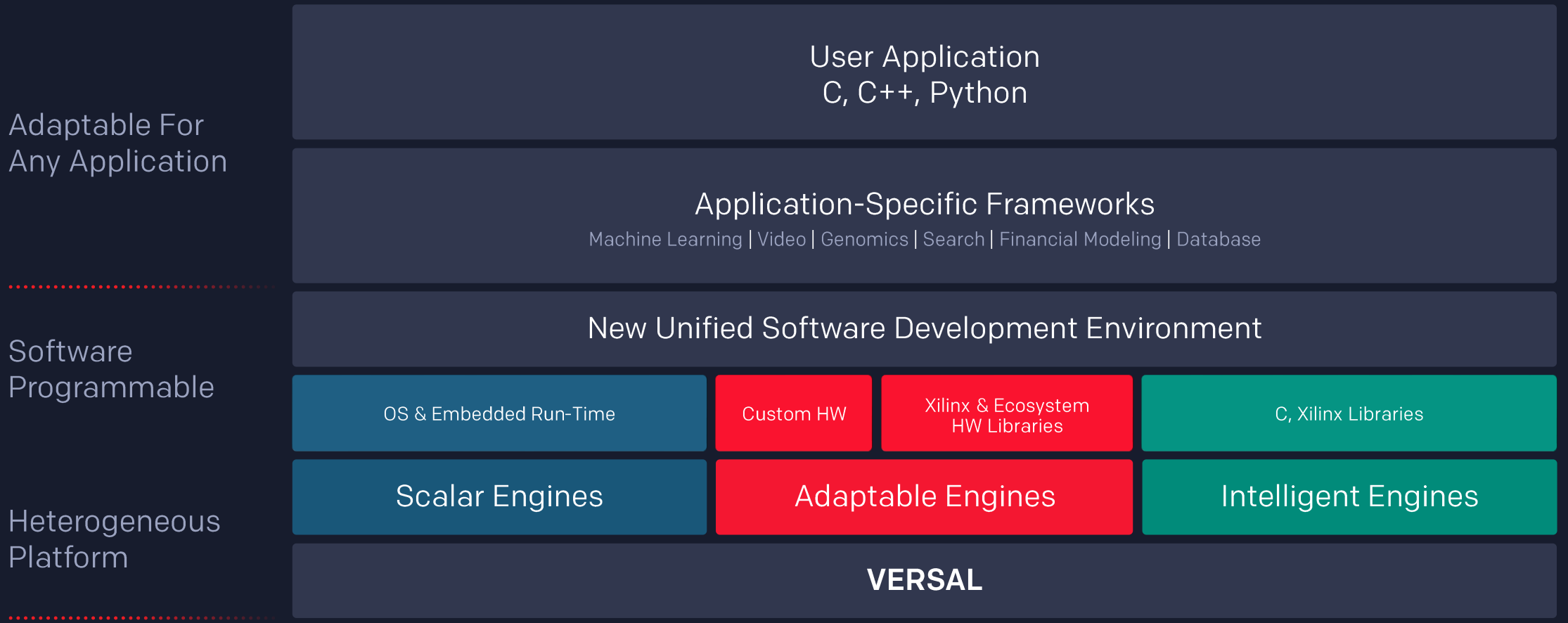
For Any Developer

Comprehensive Tool Chain

TOOLS	USER	SUPPORTED ENTRY METHODS
Frameworks	Data Scientists & AI Developers	TensorFlow Caffe mxnet Spark FFMPEG
New Unified Software Development Environment	Application Developers	
Embedded Run-Time	Embedded Developers	Linux Xen FreeRTOS
Vivado Design Suite	Hardware Developers	

XILINX VERSAL

Versal Development Experience



What’s Ahead

XILINX®

VERSAL™

AI RF
Series

HBM
Series

AI Core
Series

Premium
Series

AI Edge
Series

Prime
Series

Versal Roadmap

AI Core
AI Inference
Throughout

Prime
Broadest
Application

Premium
112G Serdes
600G Cores

AI Edge
Lowest
power AI

AI RF
AI w/
Integrated RF

HBM
Memory
Integration

2H 2019 | 2020 | 2021

IN SUMMARY

Versal ACAP

Heterogeneous Acceleration

For Any Application

For Any Developer

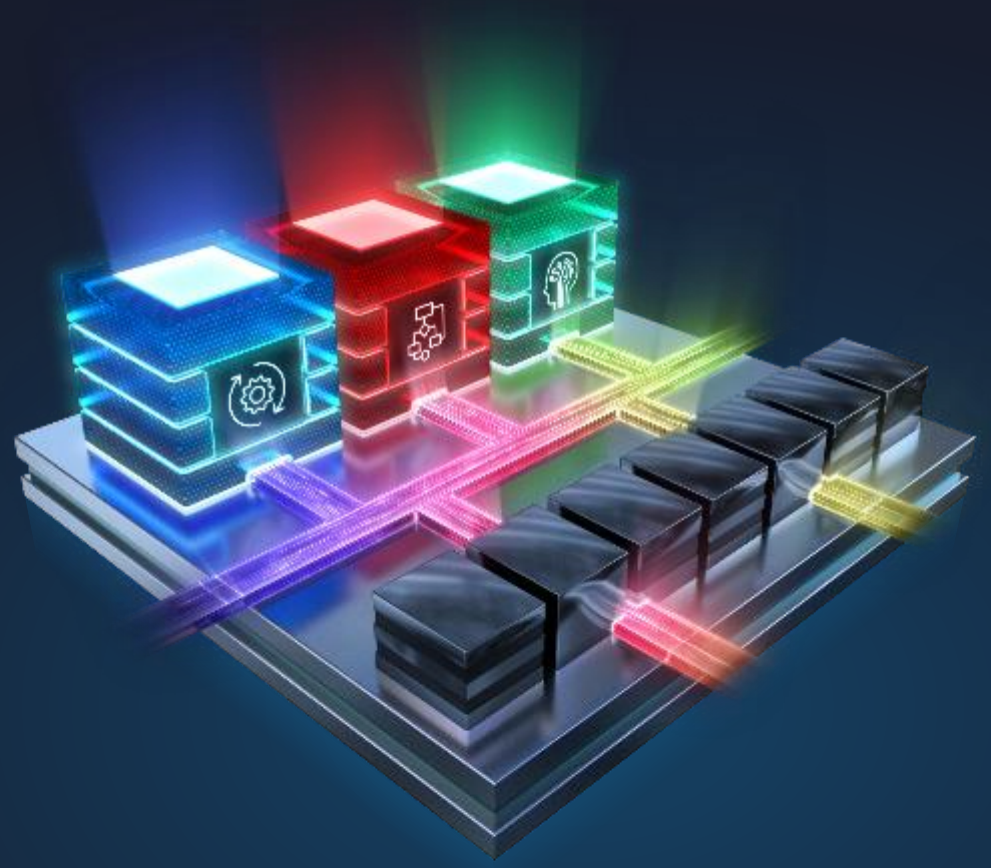
Delivers

Disruptive Innovation

Software Programmability

Hardware Adaptability

Whole Application Acceleration



Adaptable
Intelligent

Versal Prime Series — Resources

		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
	Distributed RAM (Mb)	5	8	14	11	27	28	18	25	30
Memory	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
	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, 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	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
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, 44, 78, 24, 0	648, 44, 78, 24, 0				756, 22, 78, 20, 0		
C1760	40x40				378, 44, 78, 44, 0	378, 44, 78, 44, 0		378, 22, 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, 22, 78, 20, 32	702, 44, 78, 32, 44	702, 44, 78, 40, 52

All parameters listed are maximum values. Verify all data in this document with the device data sheets or product guides found at: www.xilinx.com.

Versal AI Core Series — Resources

			VC1352	VC1502	VC1702	VC1802	VC1902
Intelligent Engines	AI Engines		128	217	310	300	400
	AI Engine Data Memory Blocks (#)		1024	1736	2480	2400	3200
	AI Engine Data Memory (Mb)		32	54.25	77.5	75	100
	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
	LUTs		246,784	364,544	466,688	725,000	899,840
	Distributed RAM (Mb)		8	11	14	22	27
Memory	Total Block RAM (Mb)		18	19	29	28	34
	UltraRAM (Mb)		42	60	113	91	130
	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, 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)				
Foundational Platform	NoC Master / NoC Slave Ports		10	14	18	28	28
	DDR Bus Width		128	128	128	256	256
	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
	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	XPIO, HDIO, MIO, GTY	XPIO, HDIO, MIO, GTY
A1024	31x31	0.92	378, 22, 78, 8	378, 22, 78, 8			
E1369	35x35	0.92	378, 44, 78, 8		378, 44, 78, 24		
A1596	37.5x37.5	0.92		378, 44, 78, 32	378, 44, 78,16	378, 44, 78, 32	378, 44, 78, 32
D1760	40x40	0.92					648, 44, 78, 24
A2197	45x45	0.92		378, 44, 78, 44		648, 44, 78, 44	648, 44, 78, 44

All parameters listed are maximum values. Verify all data in this document with the device data sheets or product guides found at: www.xilinx.com.