Versal Portfolio Product Overview

Trevor Bauer
VP, Silicon Architecture
Nov 29, 2018

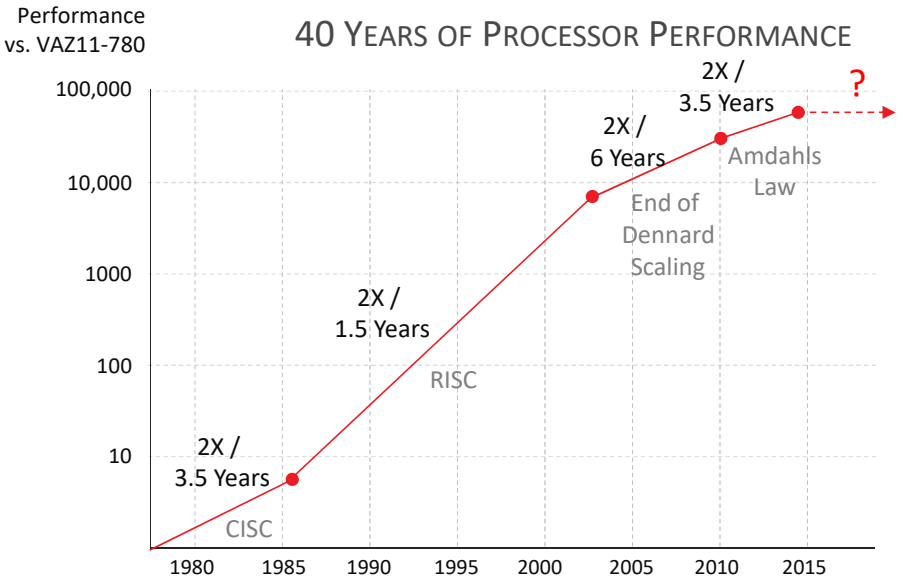
XILINX.

Agenda

- Introducing Versal: The First ACAP
- Heterogeneous Acceleration Engines
- Key Architectural Blocks
- Product Portfolio

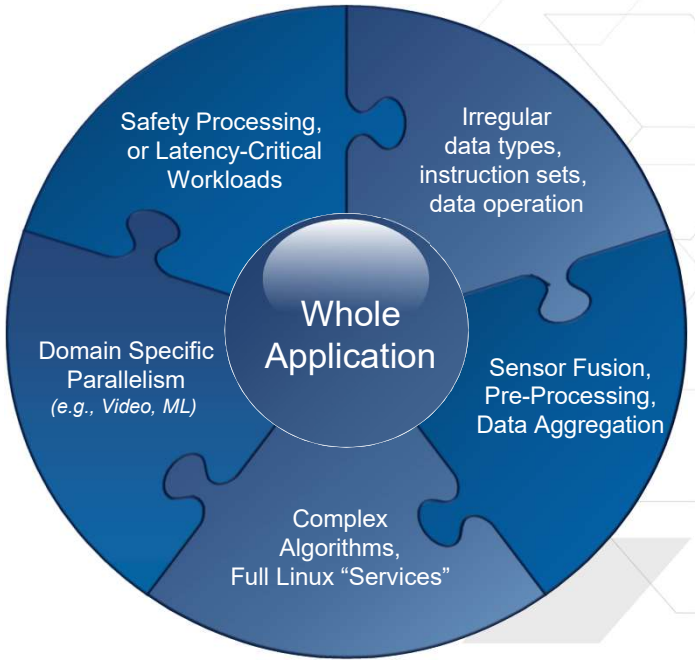
The Technology Conundrum .. And the Need for a New Compute Paradigm

Processing Architectures are Not Scaling



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

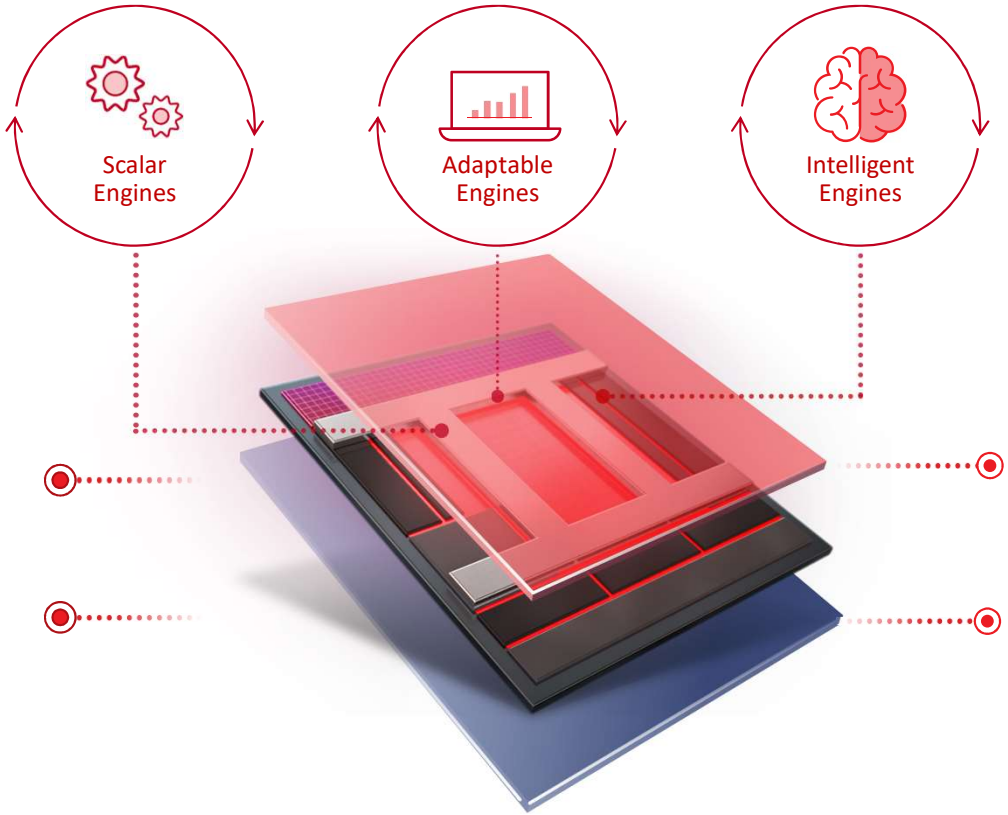
A Single Architecture Can't Do It Alone



XILINX.

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

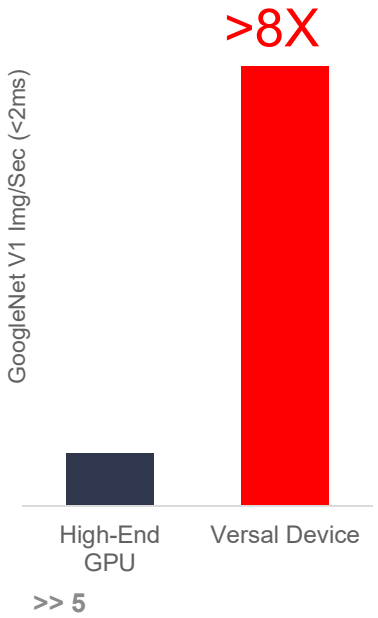
SW Programmable Silicon Infrastructure

Enabling Data Scientists, SW Developers, HW Developers

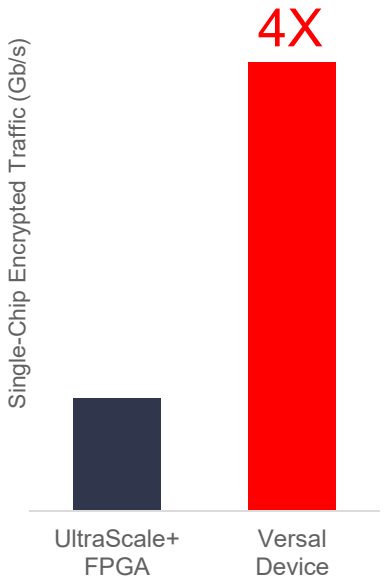
XILINX.

Breakthrough Performance for Cloud, Network, and Edge

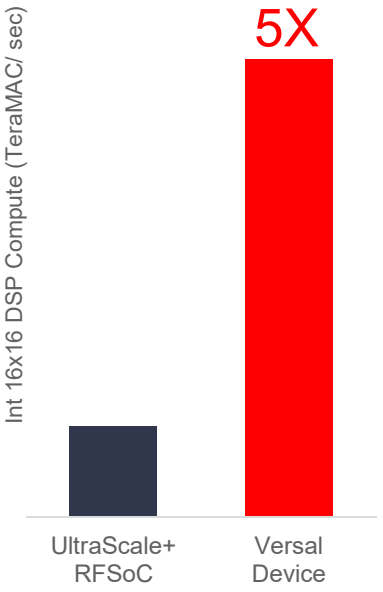
Cloud Compute
Breakthrough AI Inference



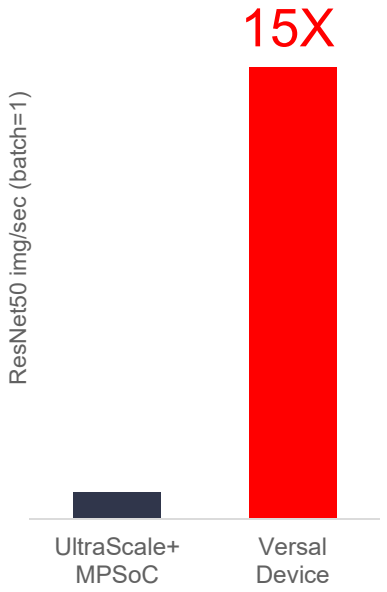
Networking
Multi-terabit Throughput



5G Wireless
Compute for Massive MIMO

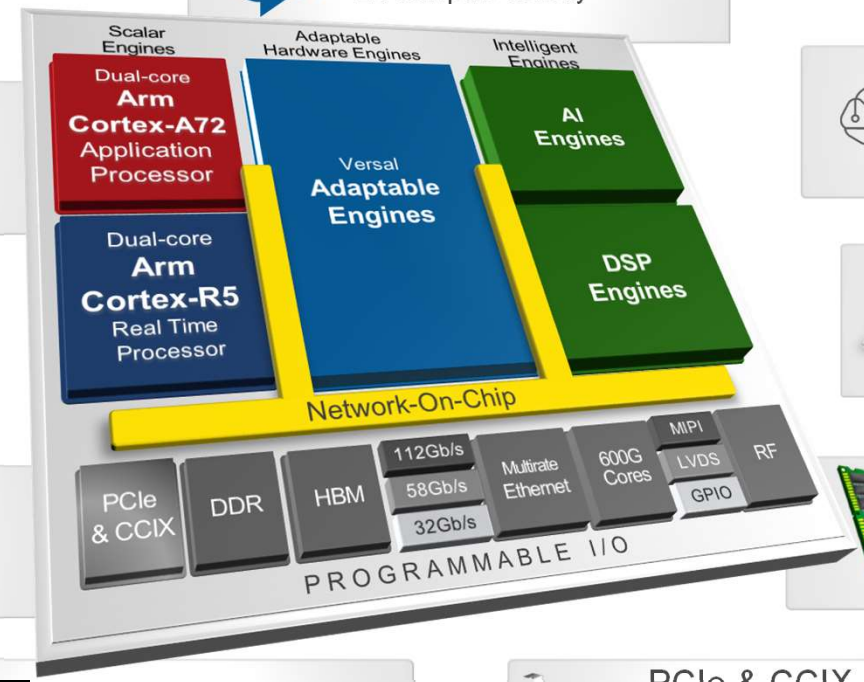


Edge Compute
AI Inference at Low Power



XILINX.

Versal Architecture Overview



Adaptable Engines
2X compute density

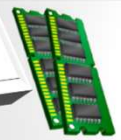
Scalar Engines
- Platform Control
- Edge Compute

Intelligent Engines
- AI Compute
- Diverse DSP workloads

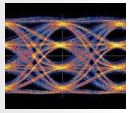
Protocol Engines
- Integrated 600G cores
- 4X encrypted bandwidth

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

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



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



Transceivers
- Broad range, 25G →112G
- 58G in mainstream devices

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

XILINX.

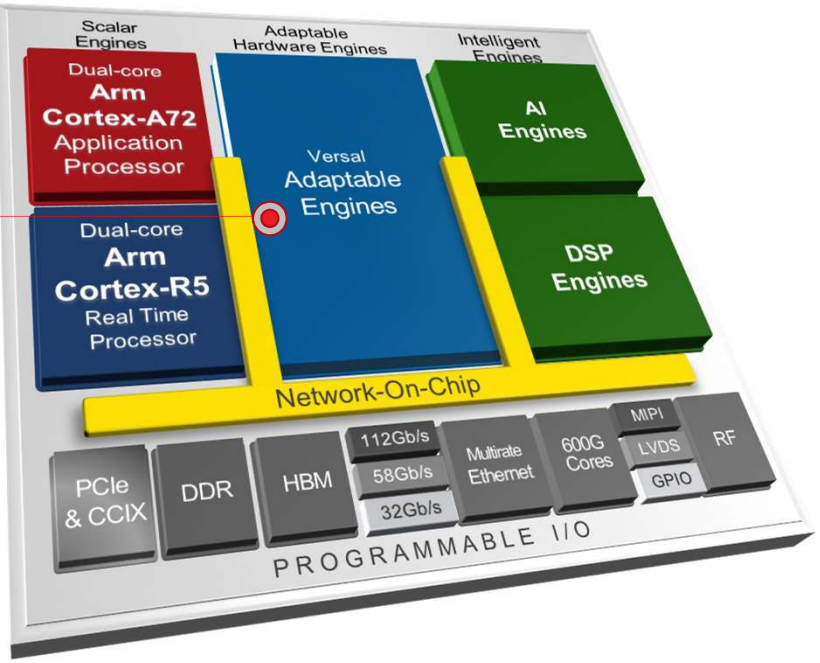
Adaptable Engines

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

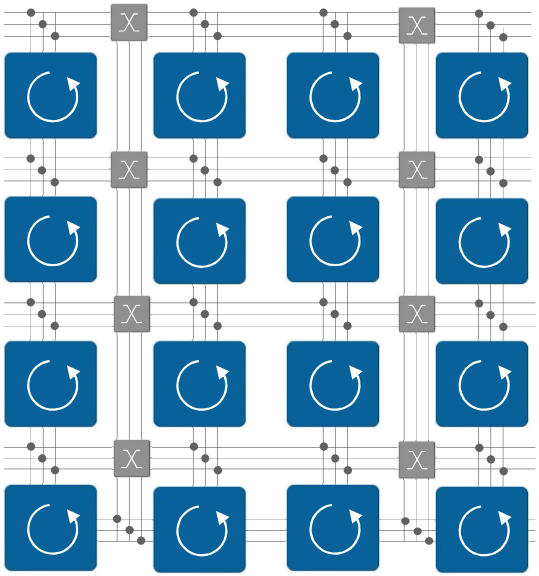


XILINX.

Greater Compute Density for Any Workload

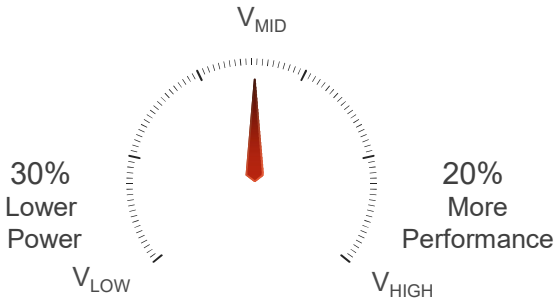
Re-Architected Hardware Fabric

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



Tune for Power & Performance

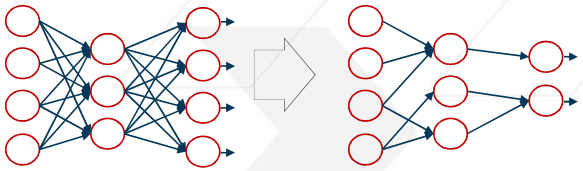
- Three operating voltages to choose from
- Balance power/performance for target app
- Equivalent to 3 speed grades in one device



Adaptable to any Workload

- Bit-level precision (1 → 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., …

XILINX.

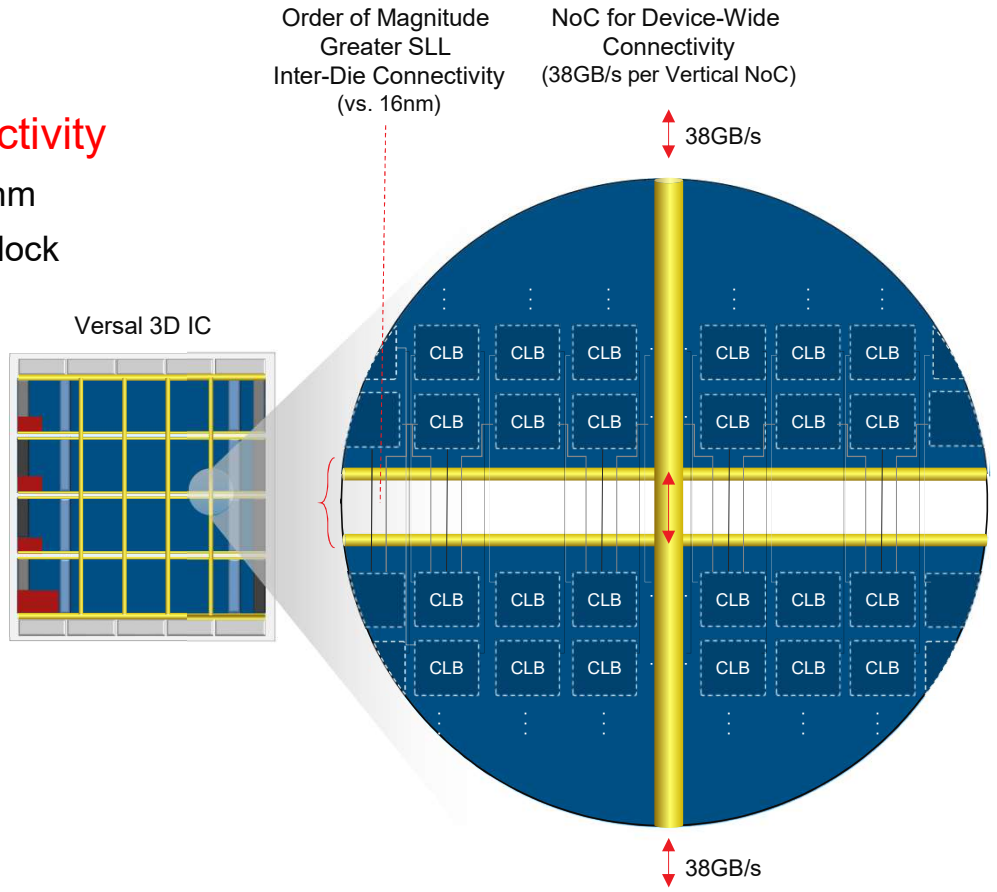
4th Generation Stacked Silicon Interconnect
Versal 3D IC Devices

Enhanced Super Long-Line (SLL) inter-die connectivity

- In 16nm, only 2 columns of registers per clock region in 16nm
- In Versal Premium: dedicated SLL circuitry for *every* logic block
- For both inter-die and intra-die bandwidth

NoC maximizes 3D IC bandwidth device-wide

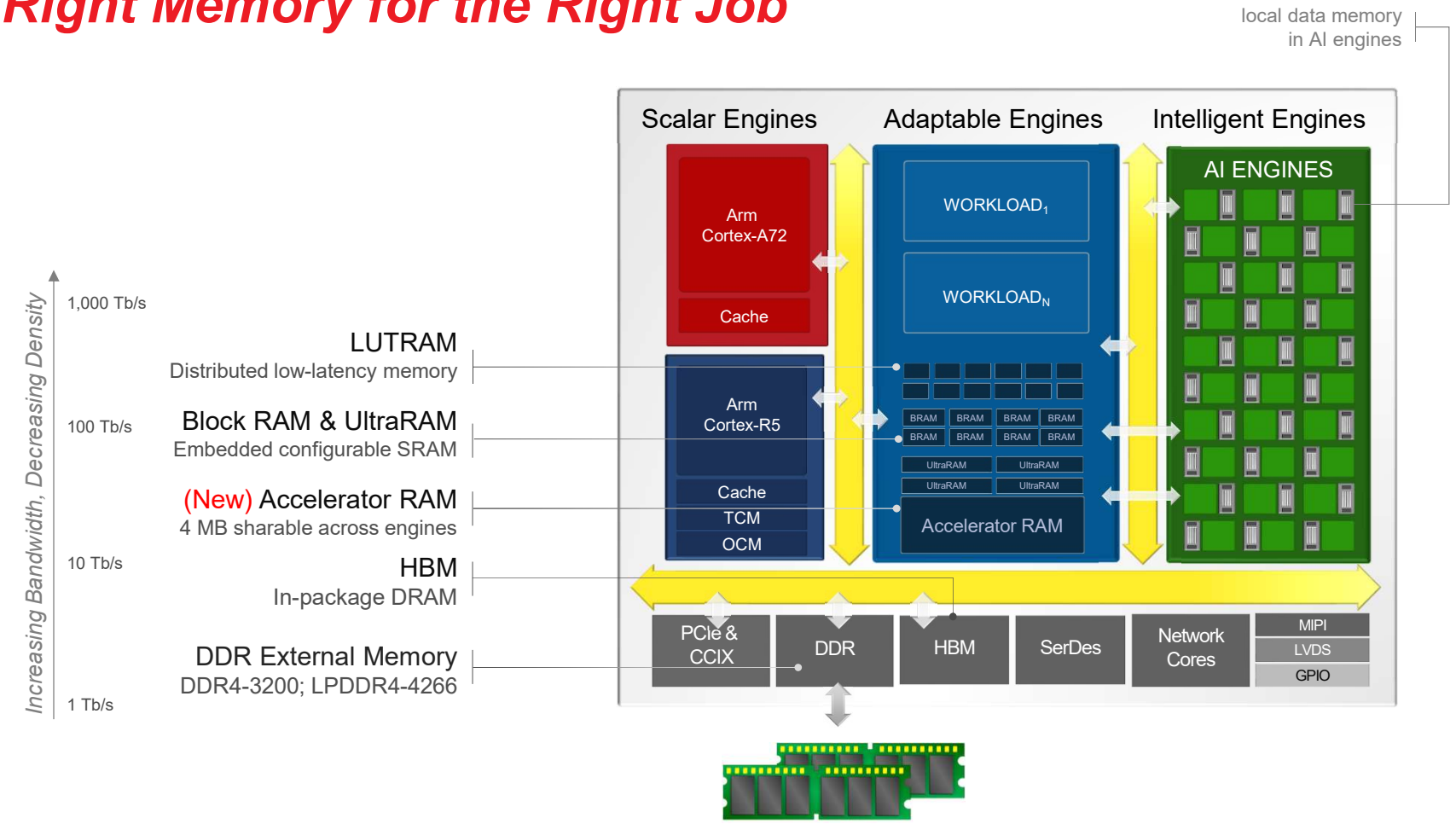
- Reduces P&R times
- Simplified floor planning
- Saves fabric resources
- High bandwidth across boundaries



XILINX.

Adaptable Memory Hierarchy

The Right Memory for the Right Job



XILINX.

Intelligent Engines

Intelligent Engines for Diverse Compute

DSP Engines

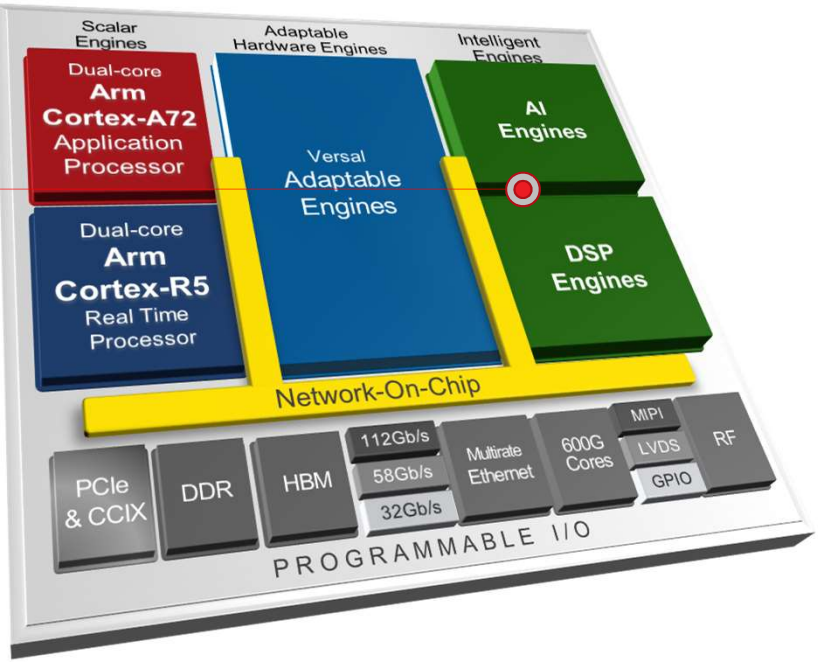
High-precision floating point & low latency

Granular control for customized data paths

AI Engines

High throughput, low latency, and power efficient

Ideal for AI inference and advanced signal processing



XILINX.

DSP Engines

Versatility and Granular Control of Data Path

Enhanced Compute architecture

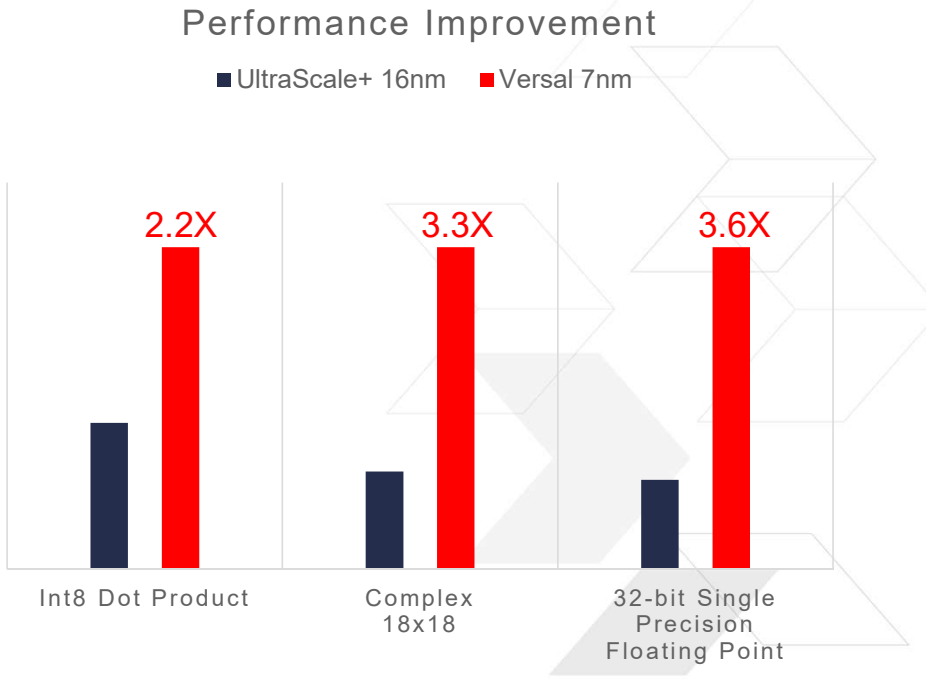
> Greater than 1GHz of performance

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.

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

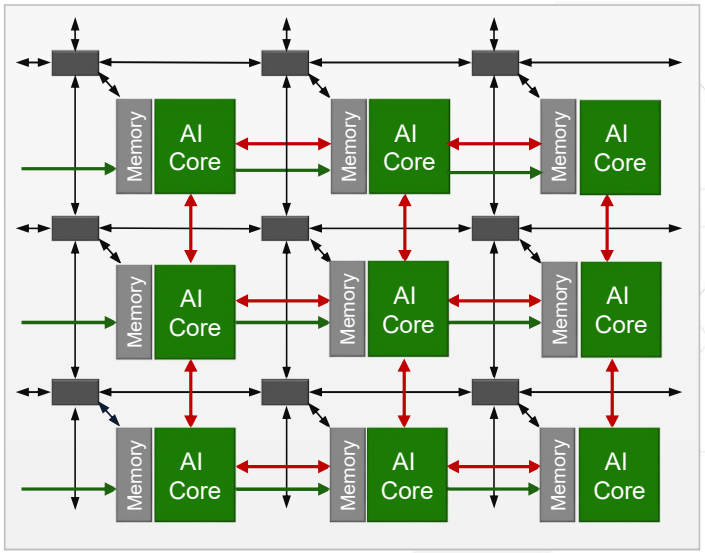
> Direct, massive throughput to adaptable HW engines

> 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



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, Low power

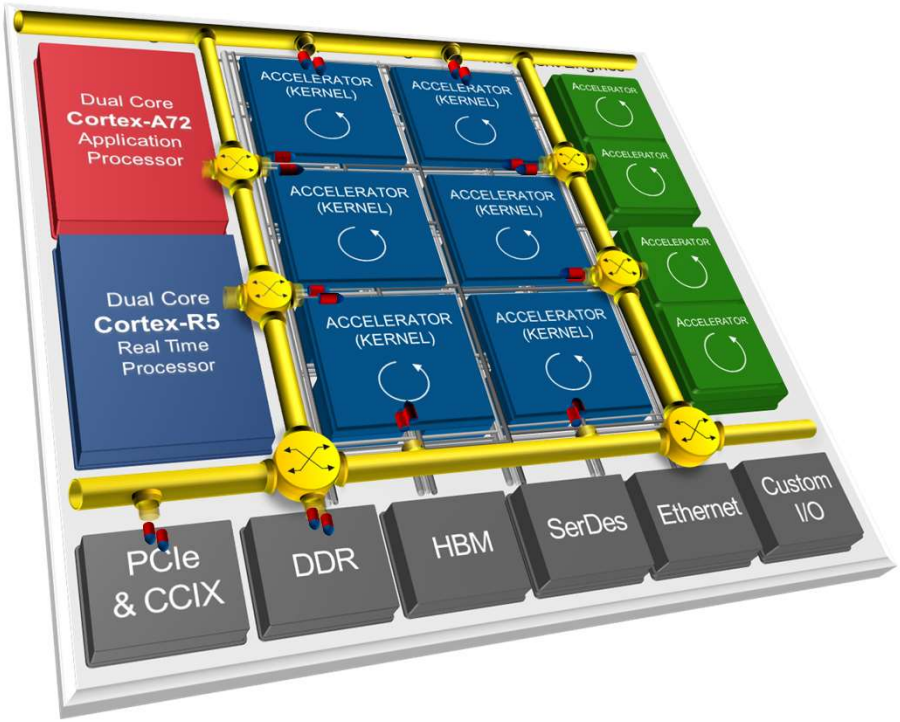
> Guaranteed QoS

> 8X power efficiency vs. FPGA implementations

Eases Kernel Placement

> Easily swap kernels at NoC port boundaries

> Simplifies connectivity between kernels

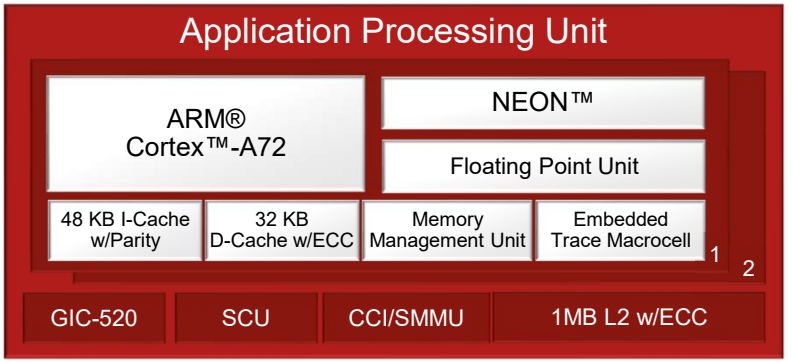


XILINX.

The Arm Subsystem

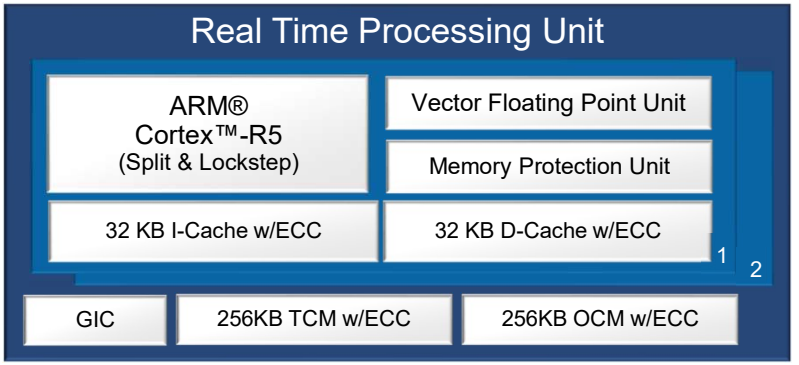
Dual-Core ARM Cortex-A72 Application Processors

- 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

Introducing the “Integrated Shell”

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

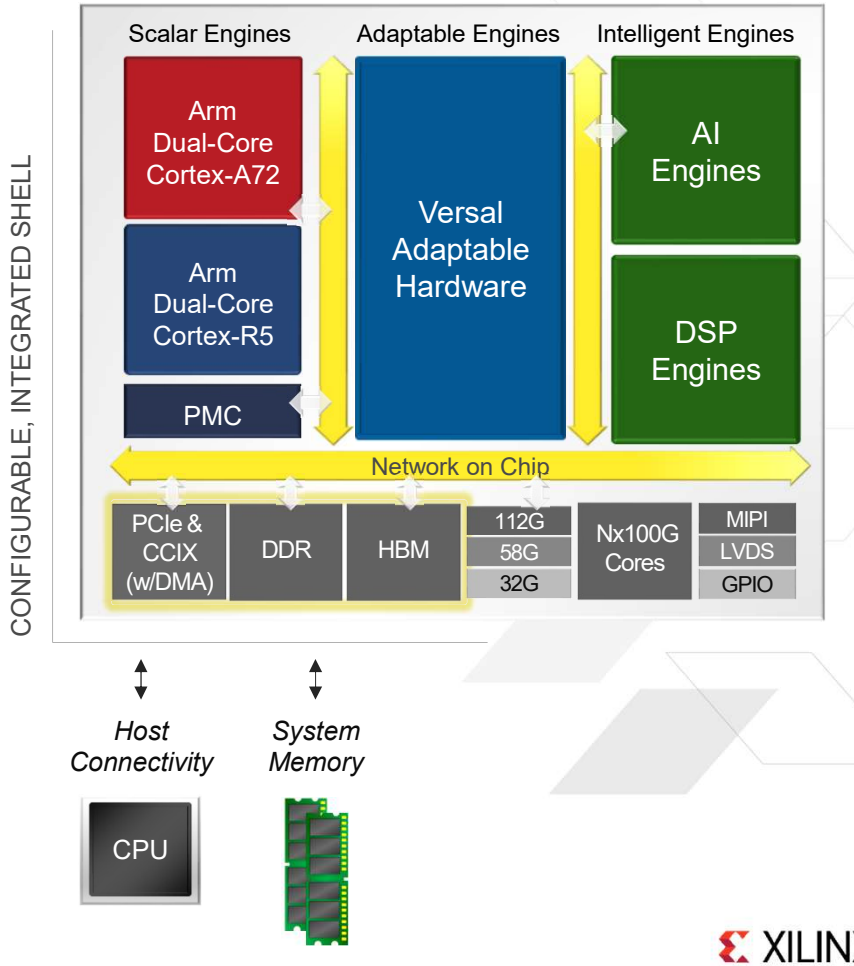
- External host interface
- Memory subsystem
- Basic interfaces (e.g., JTAG, USB, GbE)

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

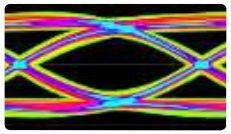
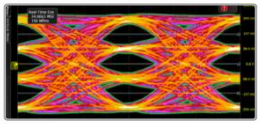
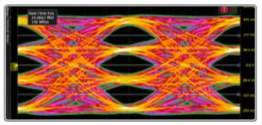
Greater Performance, Device Utilization, and Productivity

- 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



XILINX.

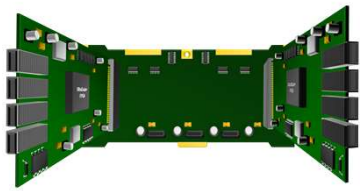
Transceivers: Robust and Scalable Connectivity

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

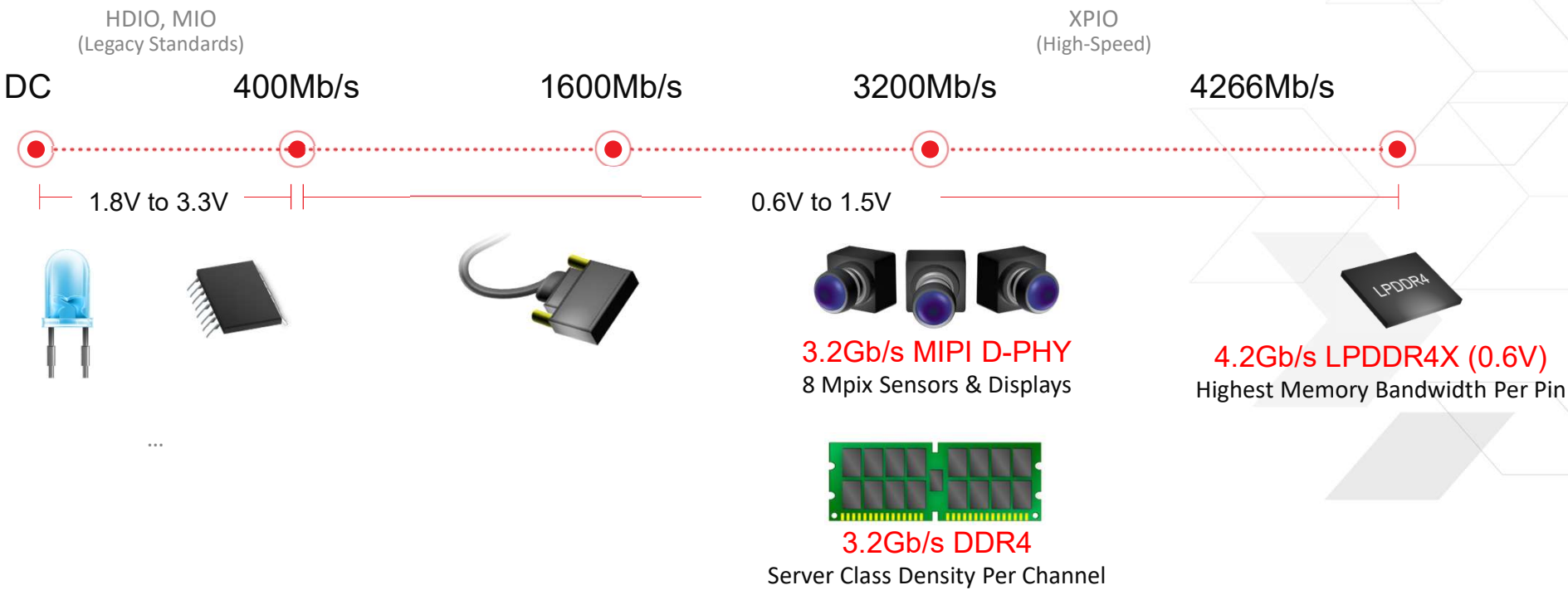
BACKPLANE



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



XILINX.

XILINX®

VERSAL™

AI RF Series

HBM Series

AI Core Series

Premium Series

AI Edge Series

Prime Series

Announcing the First Two Series of the Versal Portfolio

AI Core Series

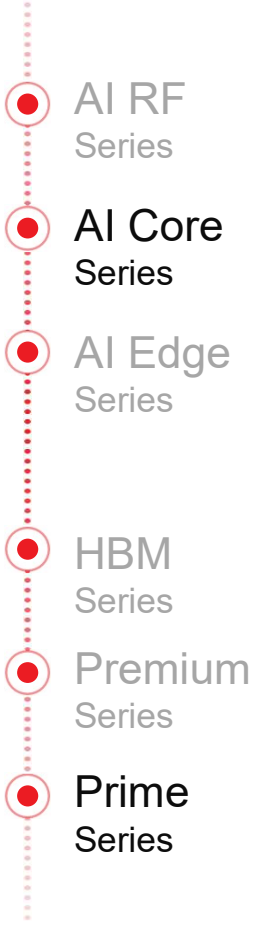
Breakthrough AI Inference Throughput

- 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



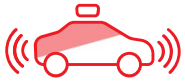
XILINX.

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

XILINX.

Getting Started

Visit www.xilinx.com/versal

> Watch ACAP Intro video

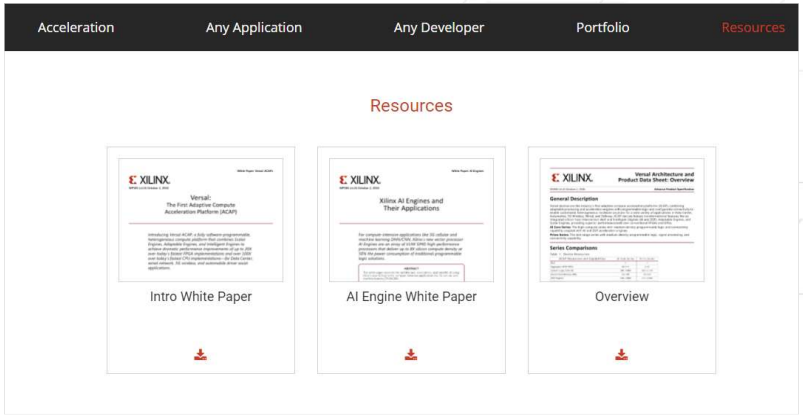
> Subscribe to mailing list for the latest news

View documentation and resources

> Data Sheet Overview

> Product Tables

> Versal Architecture and AI Engine White Papers



XILINX

Key Take-Aways

Versal: The First ACAP

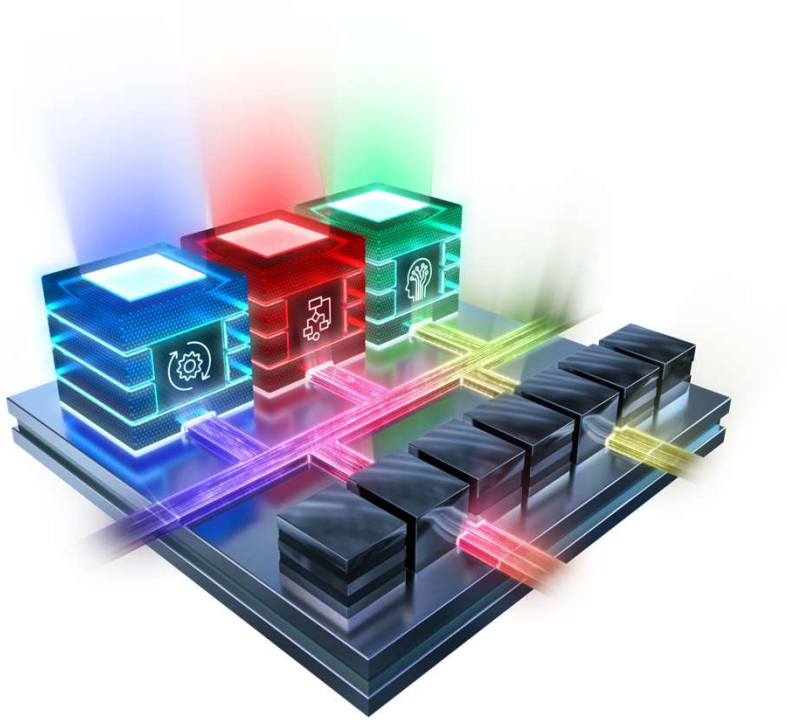
- Heterogeneous Acceleration
- For Any Application
- For Any Developer

Announcing Two Device Series

- 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



Adaptable.
Intelligent.

XILINX.