Versal Portfolio Product Overview

Jason Vidmar
System Architect, MILCOM / Satcom / Machine Learning
Mar 14, 2019

XILINX.

© Copyright 2018 Xilinx

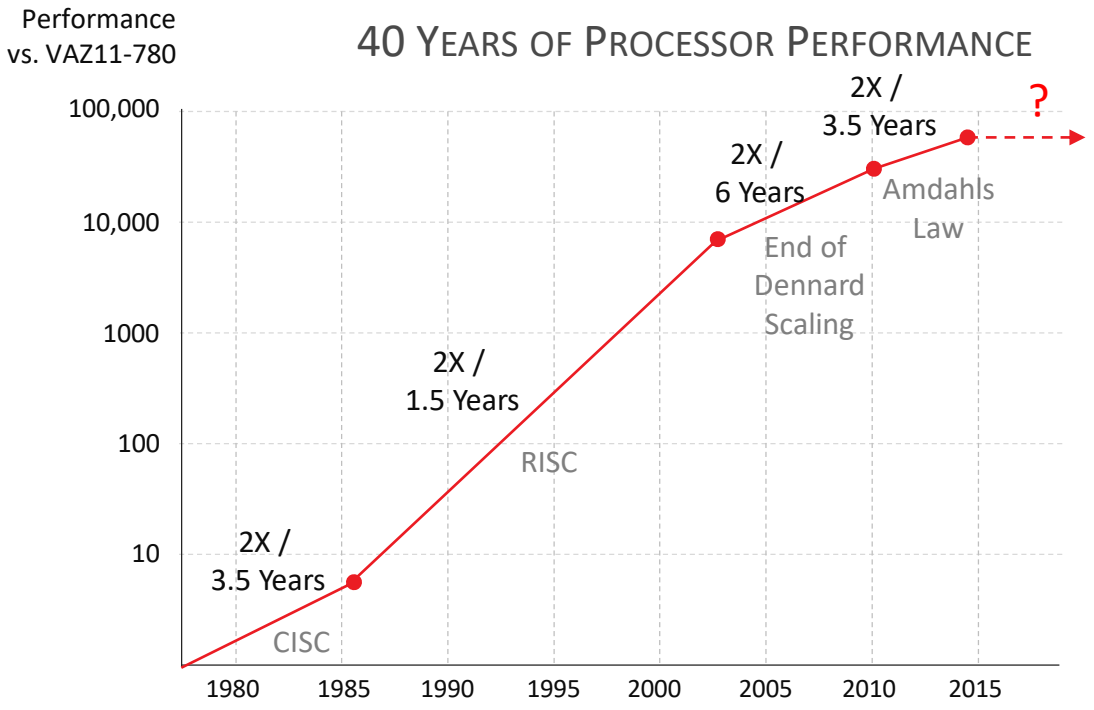
Agenda

- Introducing Versal: The First ACAP
- Heterogeneous Acceleration Engines
- Key Architectural Blocks (Focus on AI Engines)
- Product Portfolio

© Copyright 2018 Xilinx

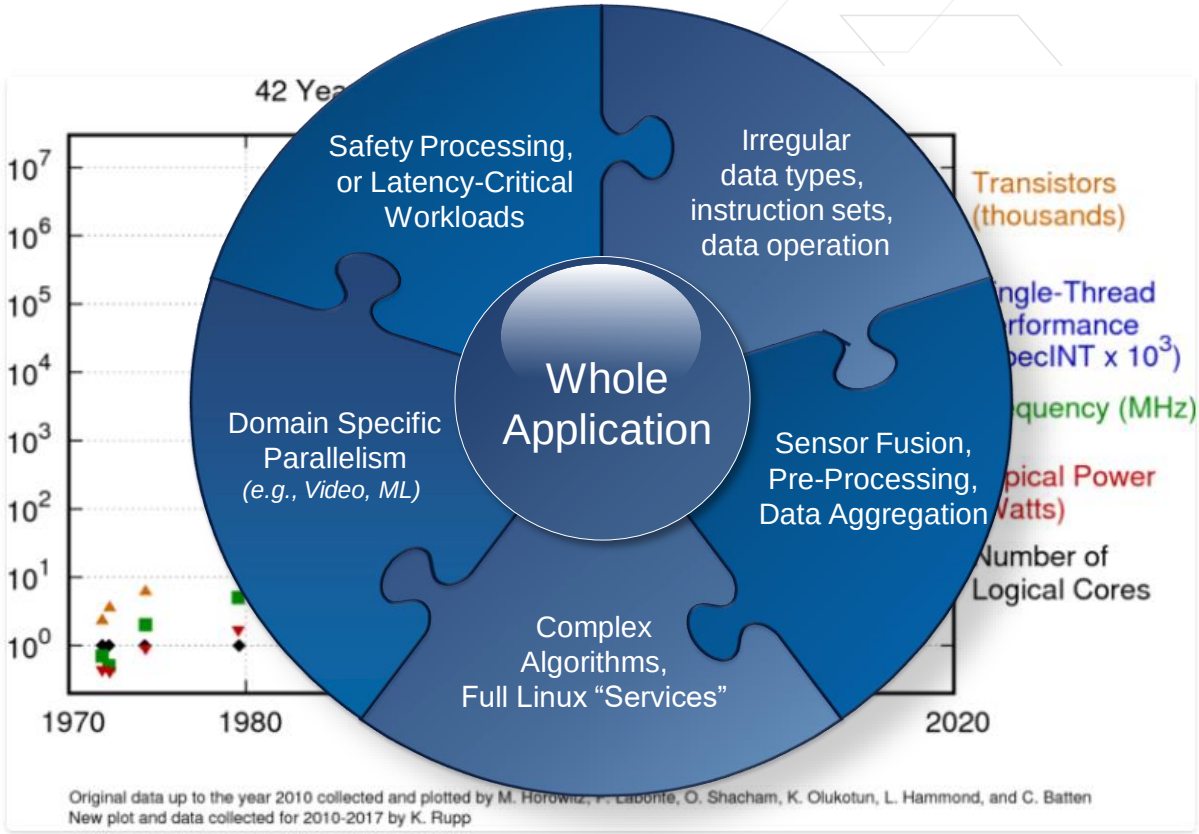
The Technology Conundrum .. And the Need for a New Compute Paradigm

Processing Architectures are Not Scaling



Source: John Hennessy and David Patterson, Computer Architecture: A Quantitative Approach, 6/e 20

A Single Architecture Can't Do It Alone



© Copyright 2018 Xilinx

XILINX

WP505: "Versal: The First Adaptive Compute Acceleration Platform."

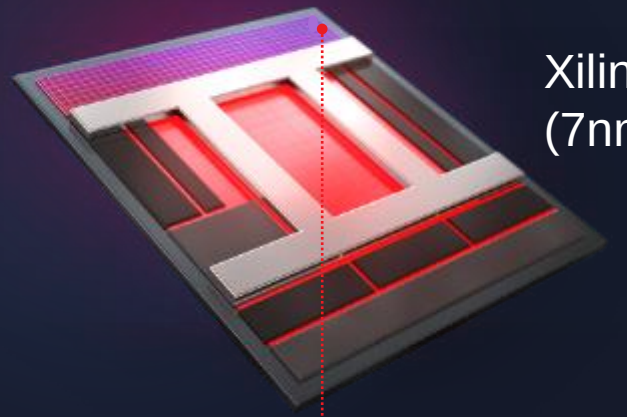
WP506: "Xilinx AI Engines and Their Applications."

Disruptive Innovation:

ACAP

Adaptive Compute Acceleration Platform

A new class of devices for today's challenges



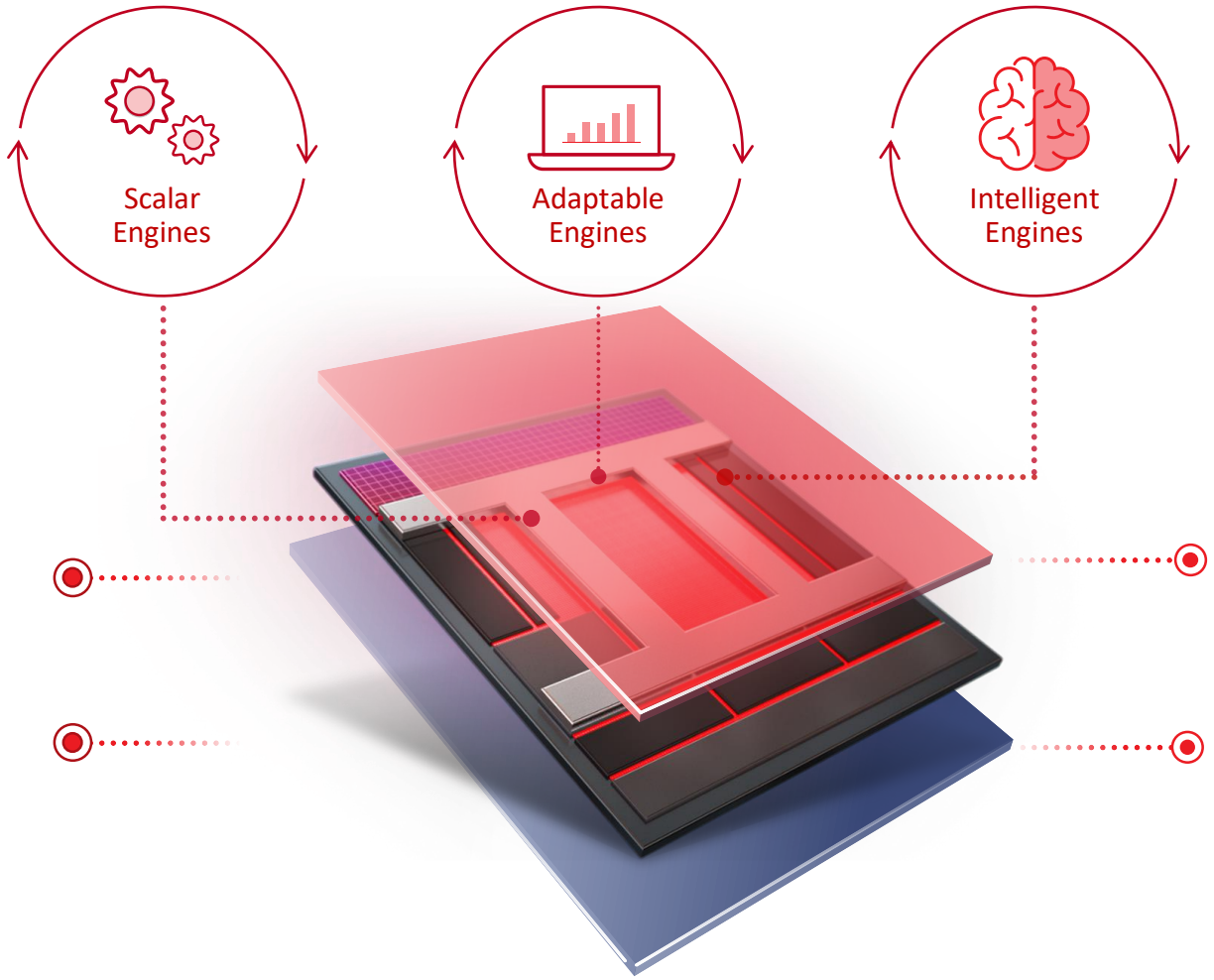
Xilinx Versal (7nm)

© Copyright 2018 Xilinx

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

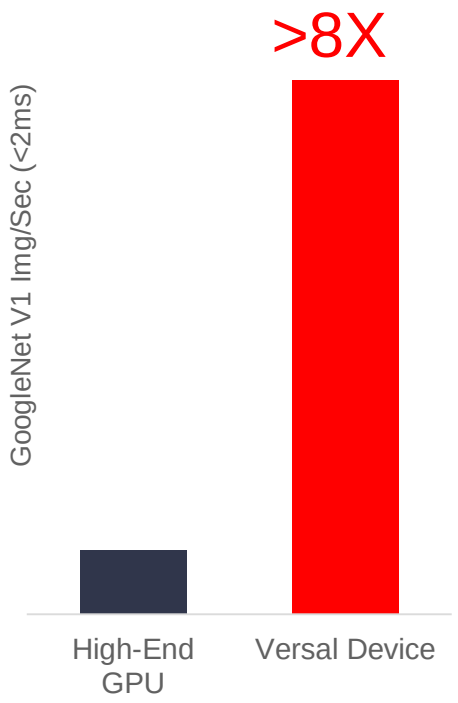
Enabling Data Scientists, SW Developers, HW Developers

© Copyright 2018 Xilinx

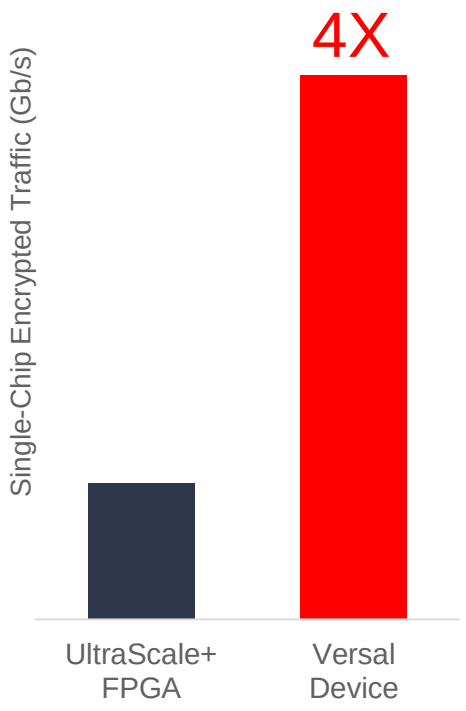
XILINX

Breakthrough Performance for Cloud, Network, and Edge

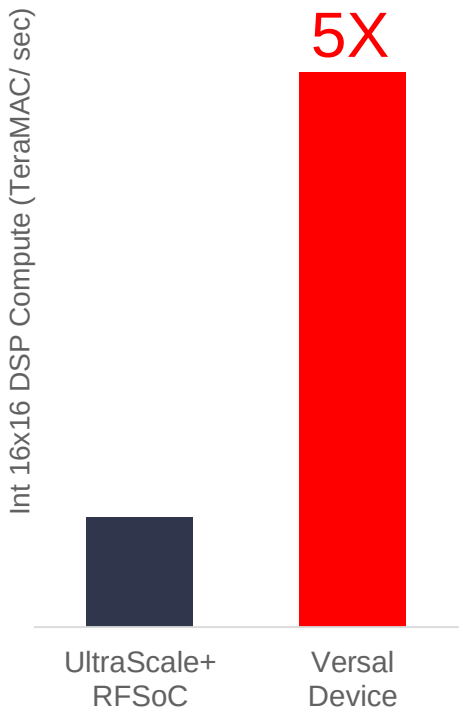
Cloud Compute
Breakthrough AI Inference



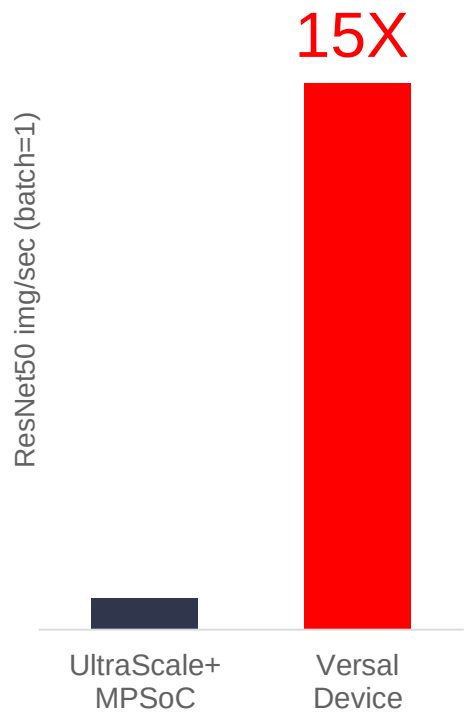
Networking
Multi-terabit Throughput



5G Wireless
Compute for Massive MIMO



Edge Compute
AI Inference at Low Power



© Copyright 2018 Xilinx

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

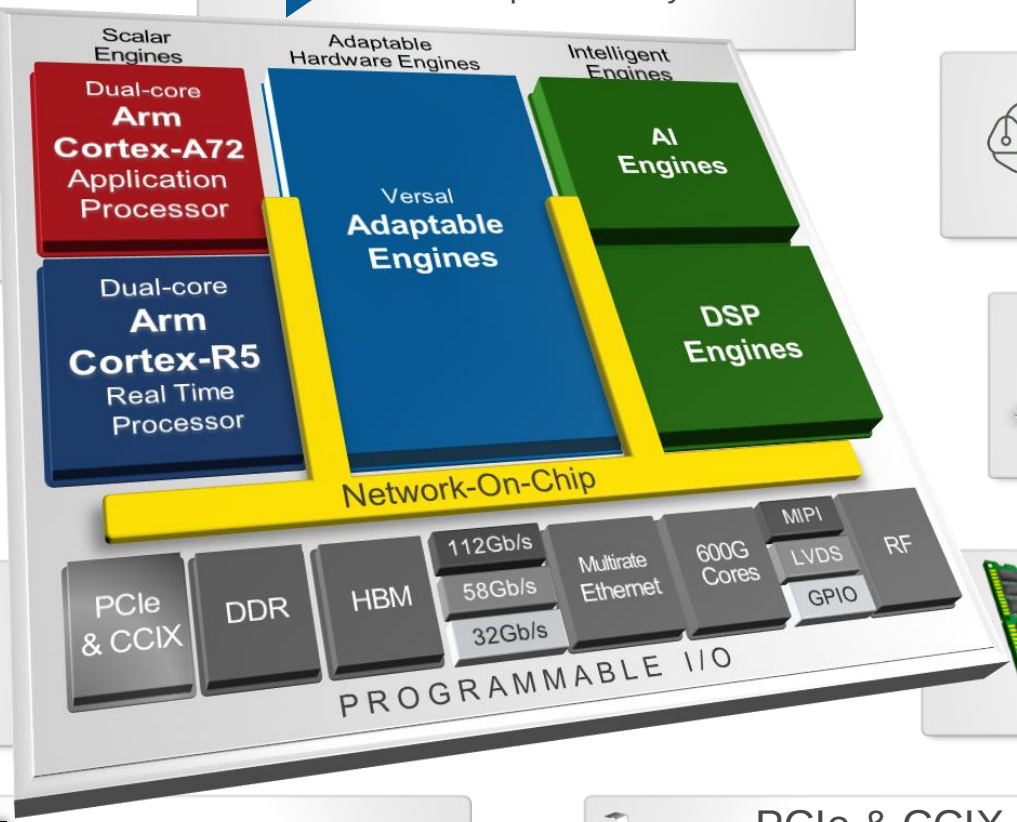
Network-on-Chip
- Guaranteed Bandwidth
- Enables SW Programmability

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

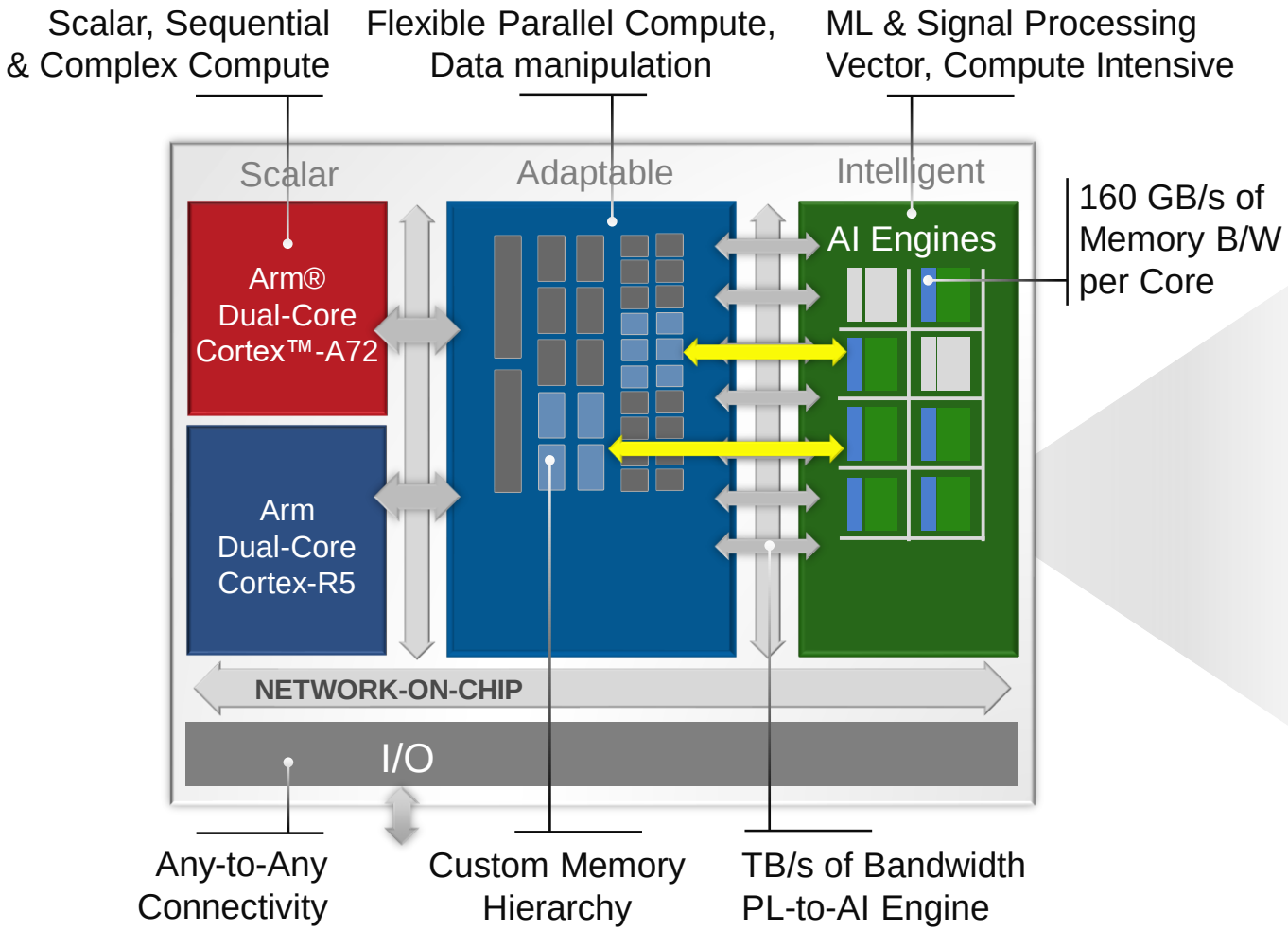
PCIe & CCIX
- 2X PCIe & DMA bandwidth
- Cache-coherent interface to accelerators



© Copyright 2018 Xilinx

XILINX

Hardware Adaptable: Accelerating the Whole Application



Heterogeneous Acceleration from Data Center to the Edge

- Video + AI
- Genomics + AI
- Risk Modeling + AI
- Database + AI
- Network IPS + AI
- Storage + AI

© Copyright 2018 Xilinx

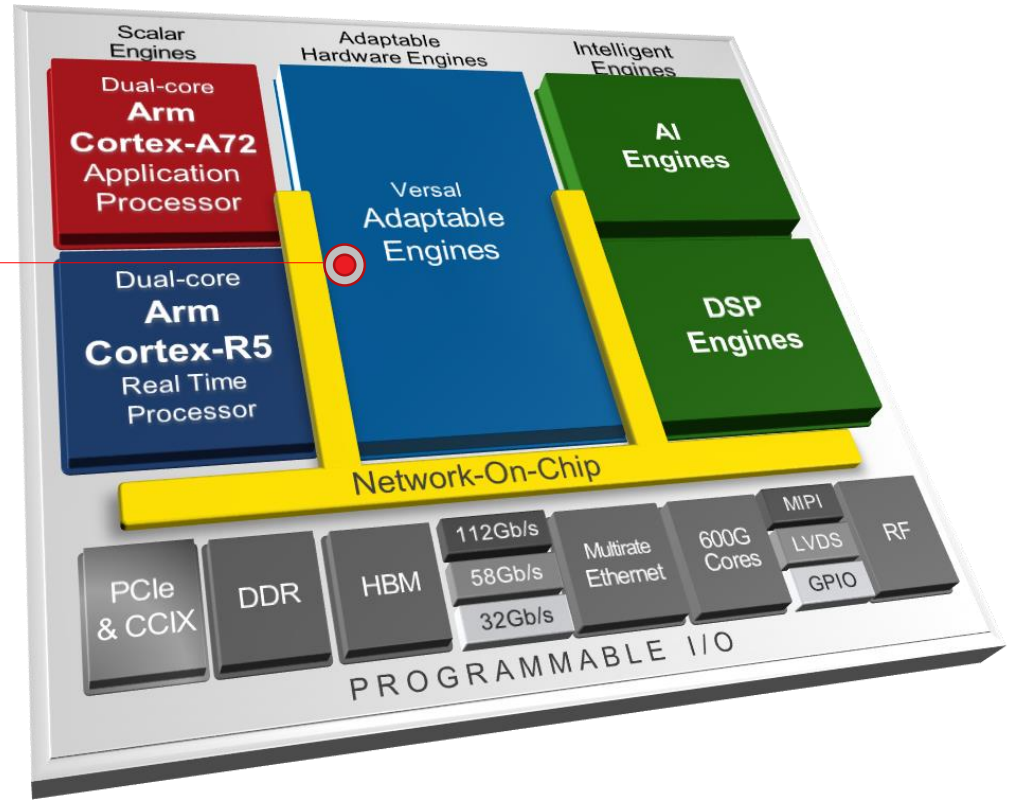
Adaptable Engines

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



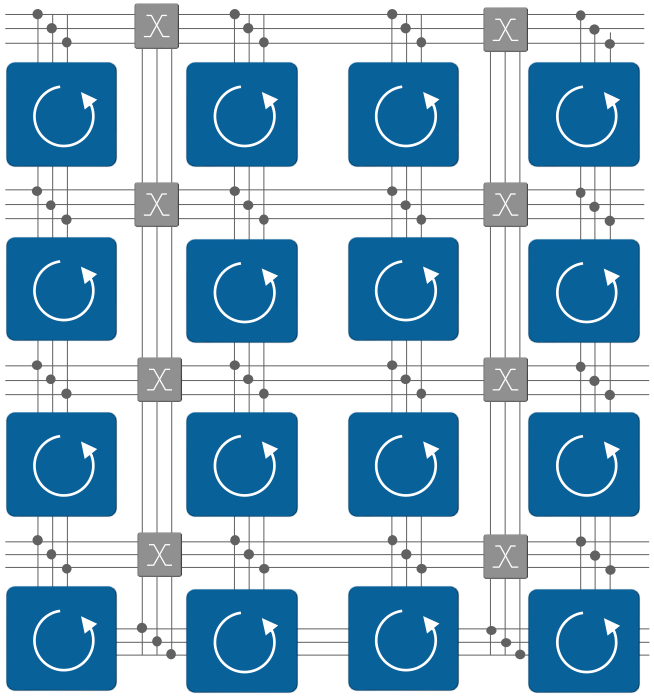
© Copyright 2018 Xilinx

XILINX.

Adaptable Engines: Greater Compute Density for Any Workload

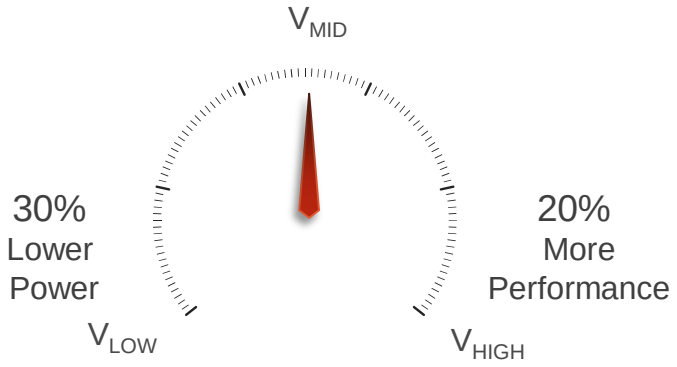
Re-Architected Hardware Fabric

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



Tune for Power & Performance

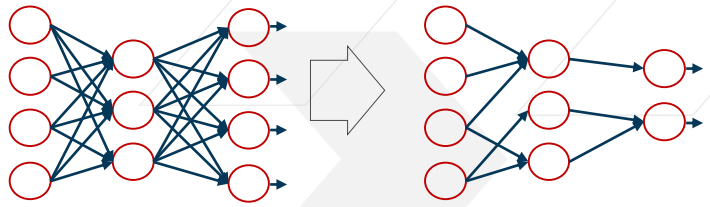
- Three operating voltages to choose from
- Balance power/performance for target app
- Equivalent to 3 speed grades in one device



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)



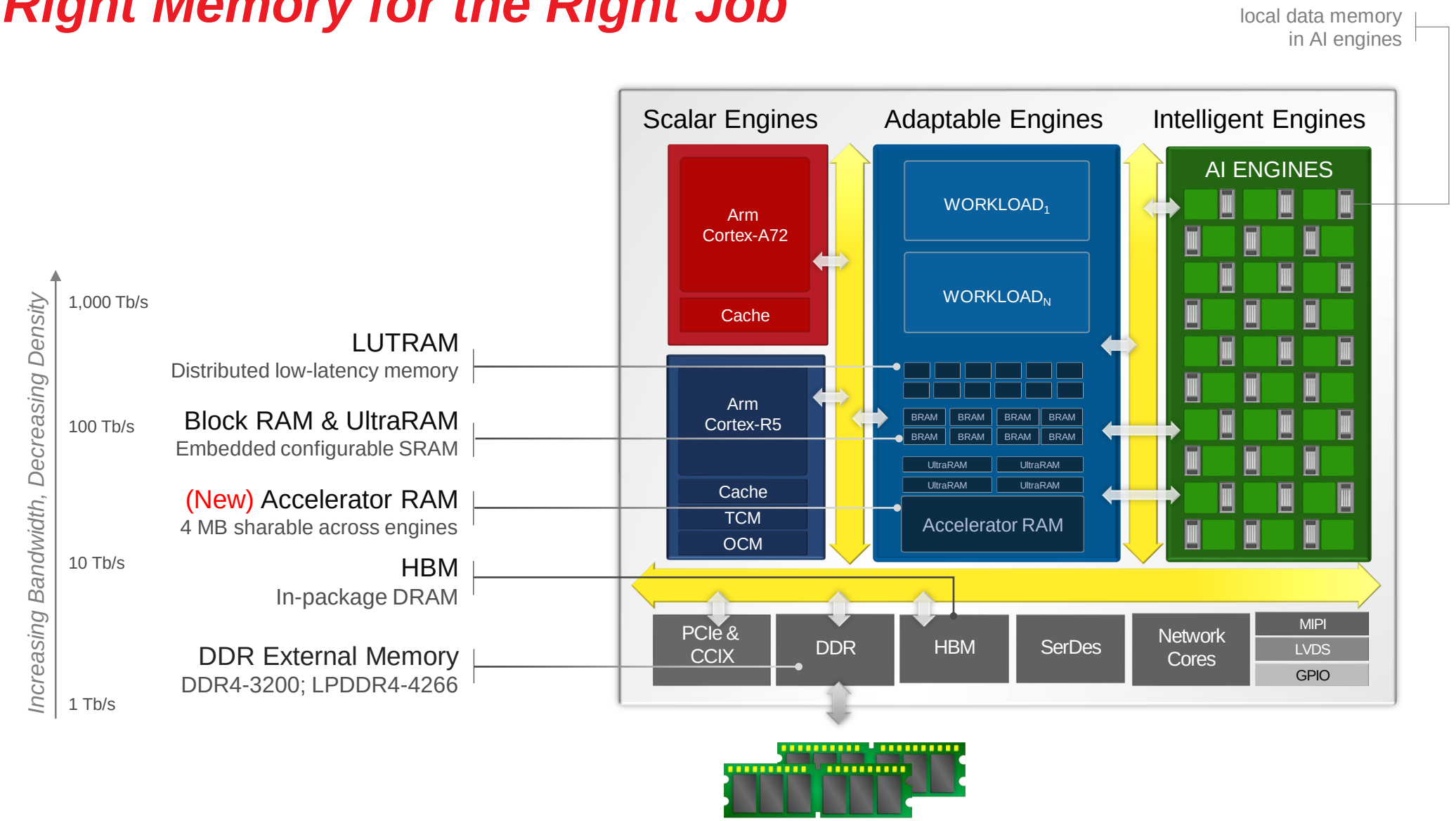
For Any Workload, e.g., …

© Copyright 2018 Xilinx

XILINX

Adaptable Memory Hierarchy

The Right Memory for the Right Job



© Copyright 2018 Xilinx

XILINX.

Intelligent Engines

Intelligent Engines for Diverse Compute

DSP Engines

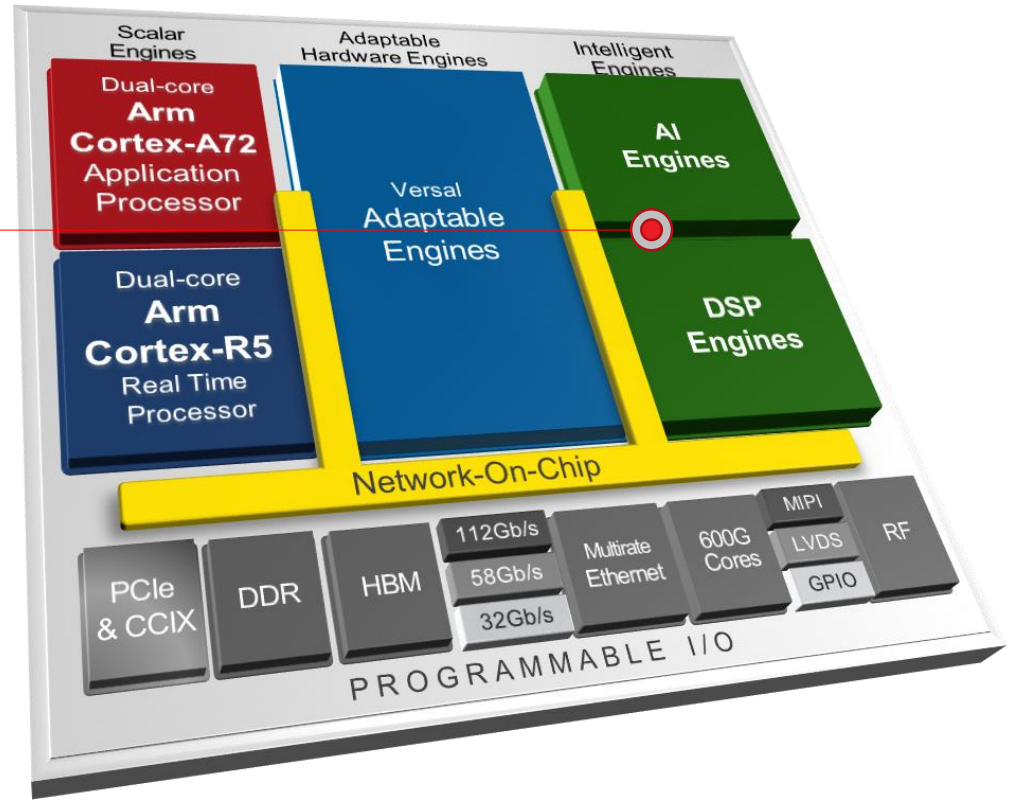
High-precision floating point & low latency

Granular control for customized data paths

AI Engines

High throughput, low latency, and power efficient

Ideal for AI inference and advanced signal processing



© Copyright 2018 Xilinx

XILINX.

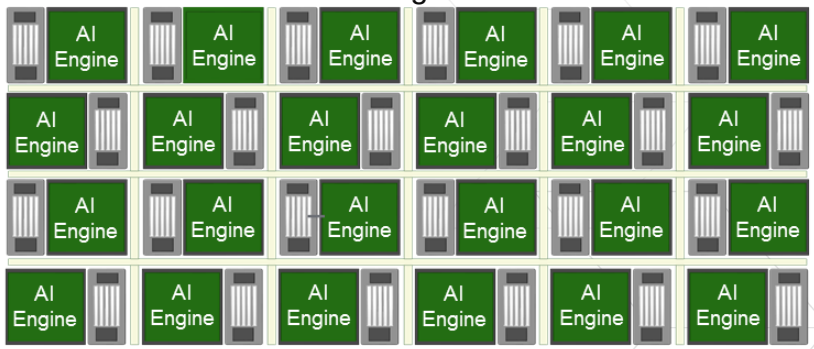
Intelligent Engines: Digital Signal Processing Capability

Function	DSP48E2	DSP58
DSP Tile/Slice Type	DSP48E2	DSP58
Multiplier and MACC	27x18	27x24
32b/16b Single Precision Floating Point Multiply-Add	Soft	√
Complex 18b x Complex 18b	N/A	2 x DSP58
3 x Int8 Dot Product	N/A	√

AI Engine 2D Array

VLIW and SIMD Architecture

C/C++ Programmable

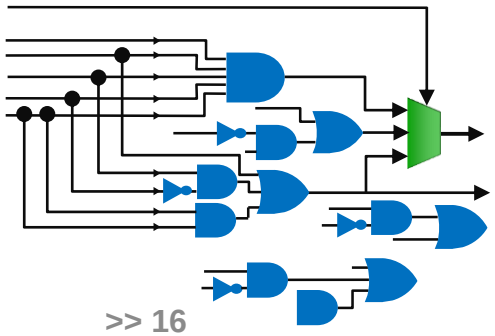


Capability

FPGA Fabric DSP

LUT and Memory

RTL Entry



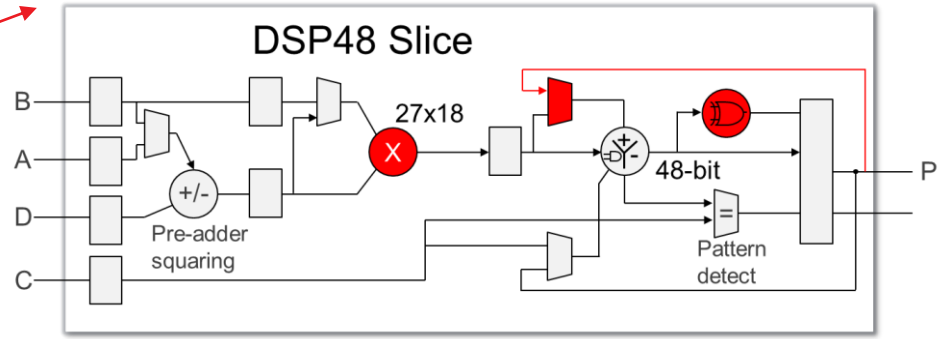
Capability

DSP48E2 Slice

Hardened MULT & ADDERS

ACC = ACC + (A × B)

RTL Entry

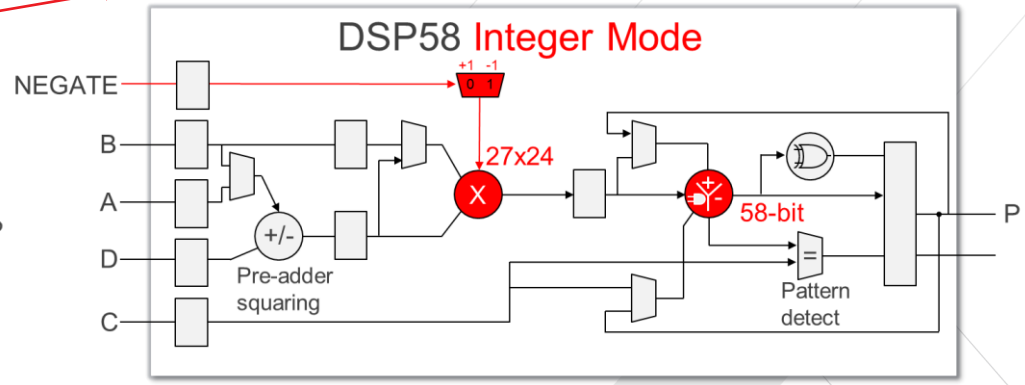


Capability

DSP58

Additional features

RTL Entry



© Copyright 2018 Xilinx

XILINX

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

High bandwidth terabit network-on-chip

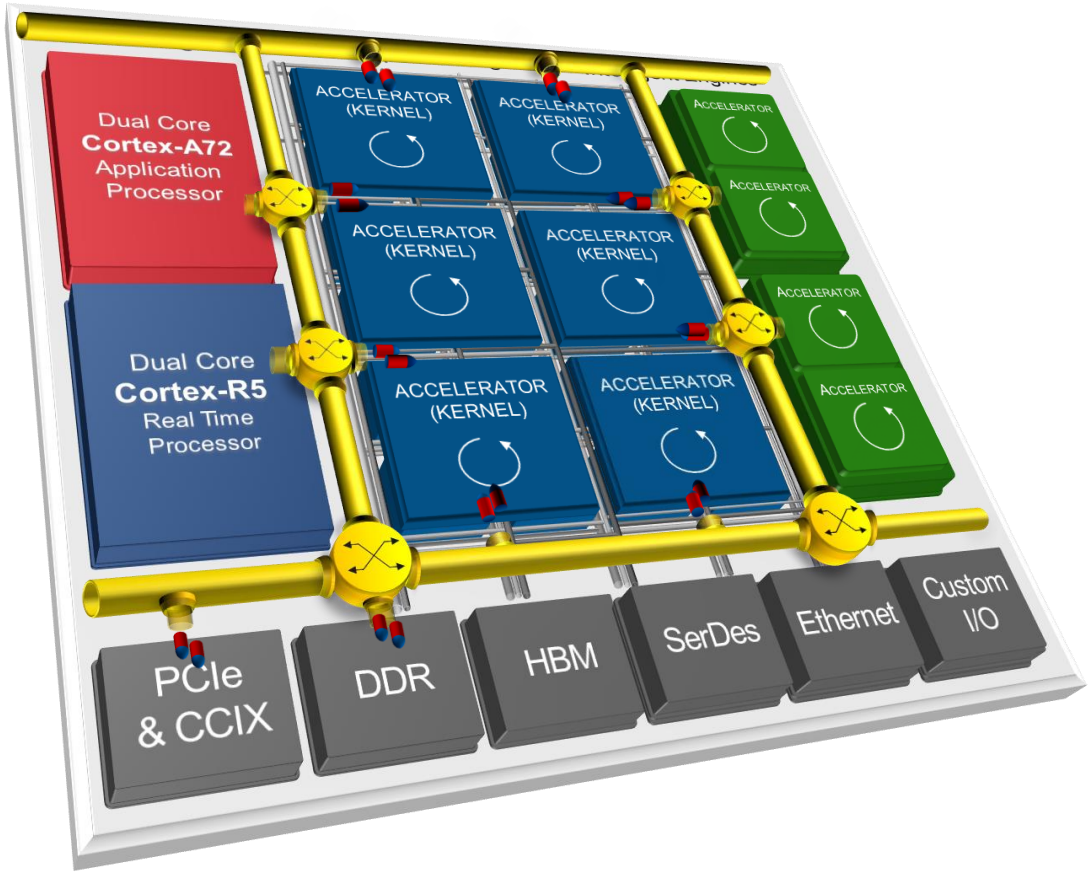
- Memory mapped access to all resources
- Built-in arbitration between engines and memory

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



© Copyright 2018 Xilinx

XILINX.

Introducing the “Integrated Shell”

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

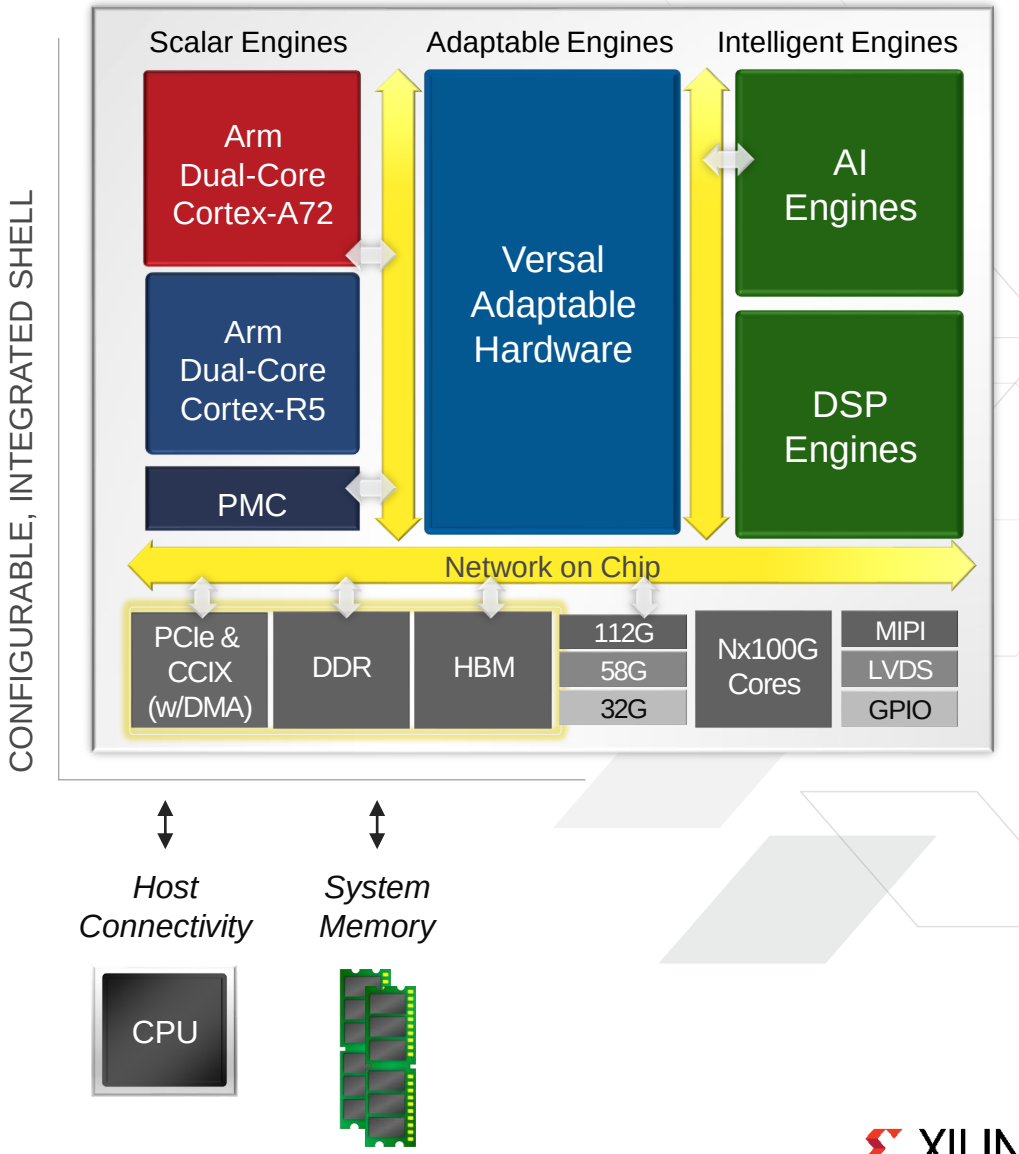
- External host interface
- Memory subsystem
- Basic interfaces (e.g., JTAG, USB, GbE)

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



© Copyright 2018 Xilinx

XILINX.

AI Engines

© Copyright 2018 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

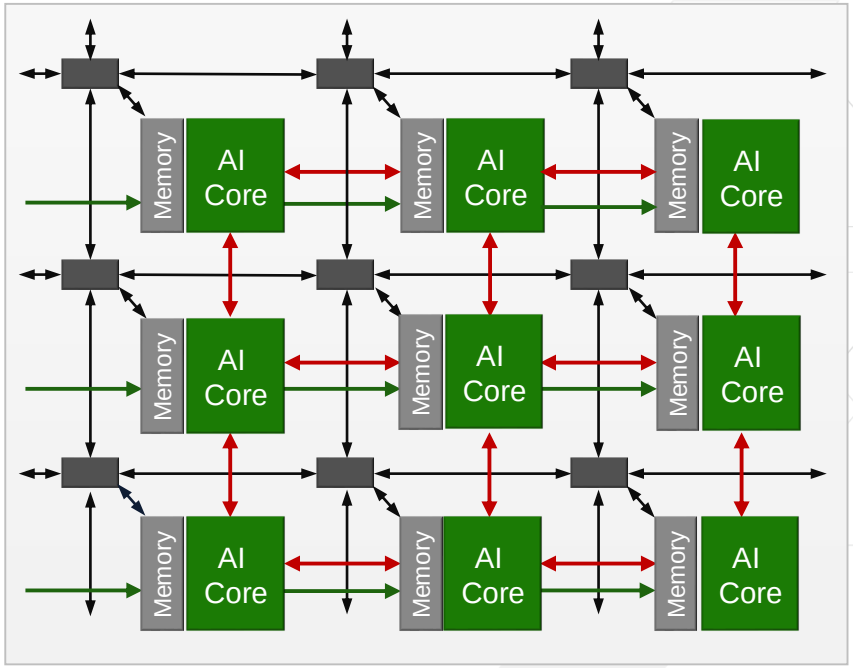
> Direct, massive throughput to adaptable HW engines

> 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



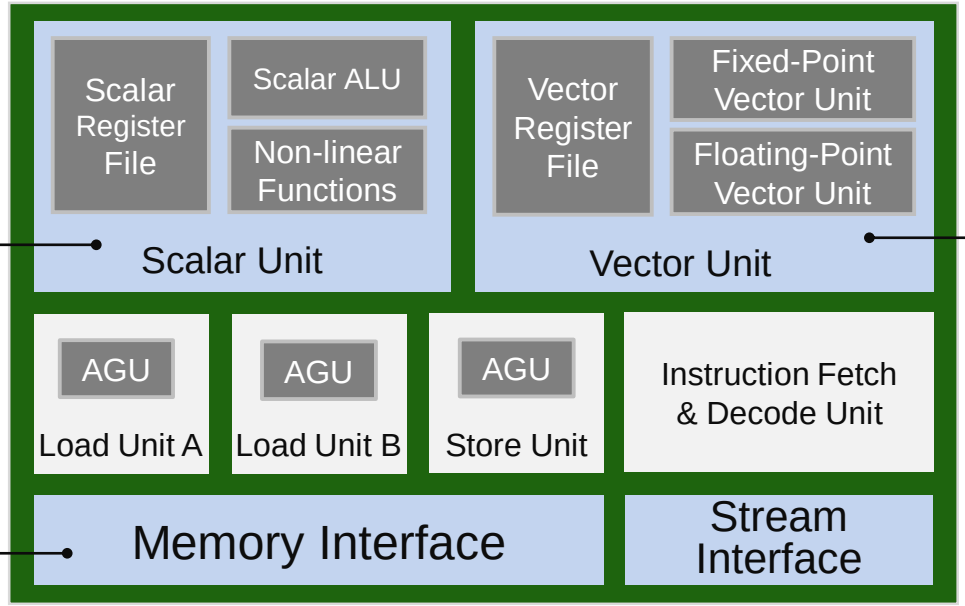
© Copyright 2018 Xilinx

XILINX

AI Engine: Scalar Unit, Vector Unit, Load Units and Memory

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

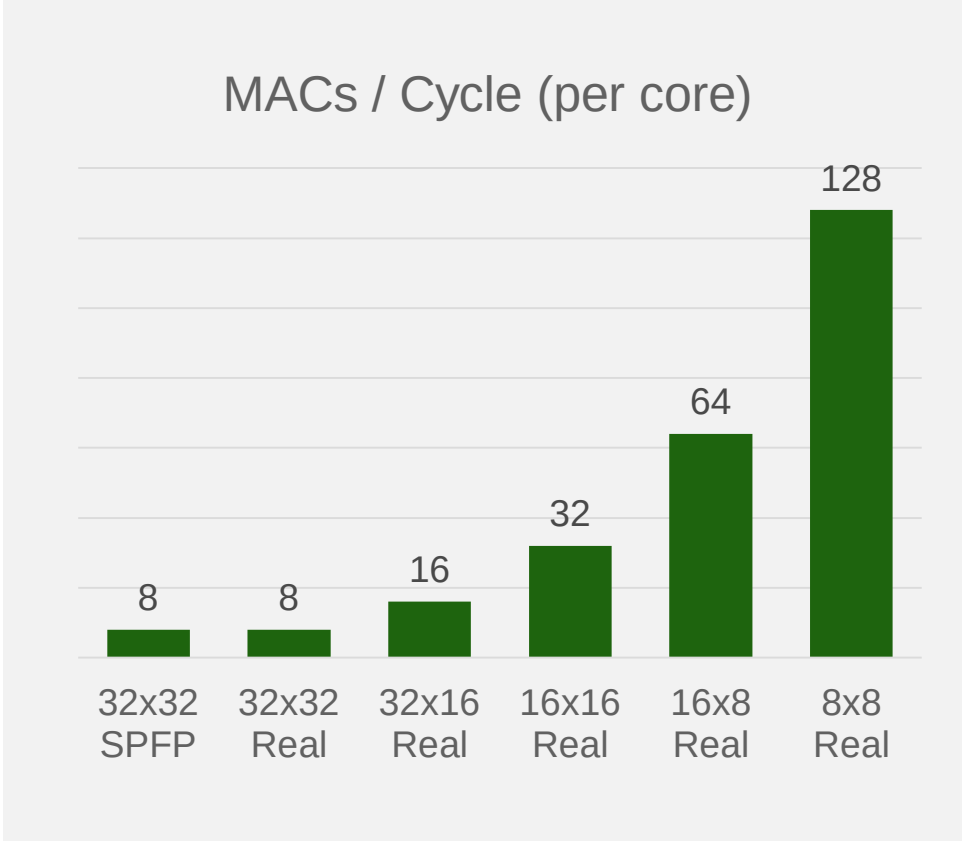
Multiple vector lanes
- Vector Datapath
- 8 / 16 / 32-bit & SPFP operands

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

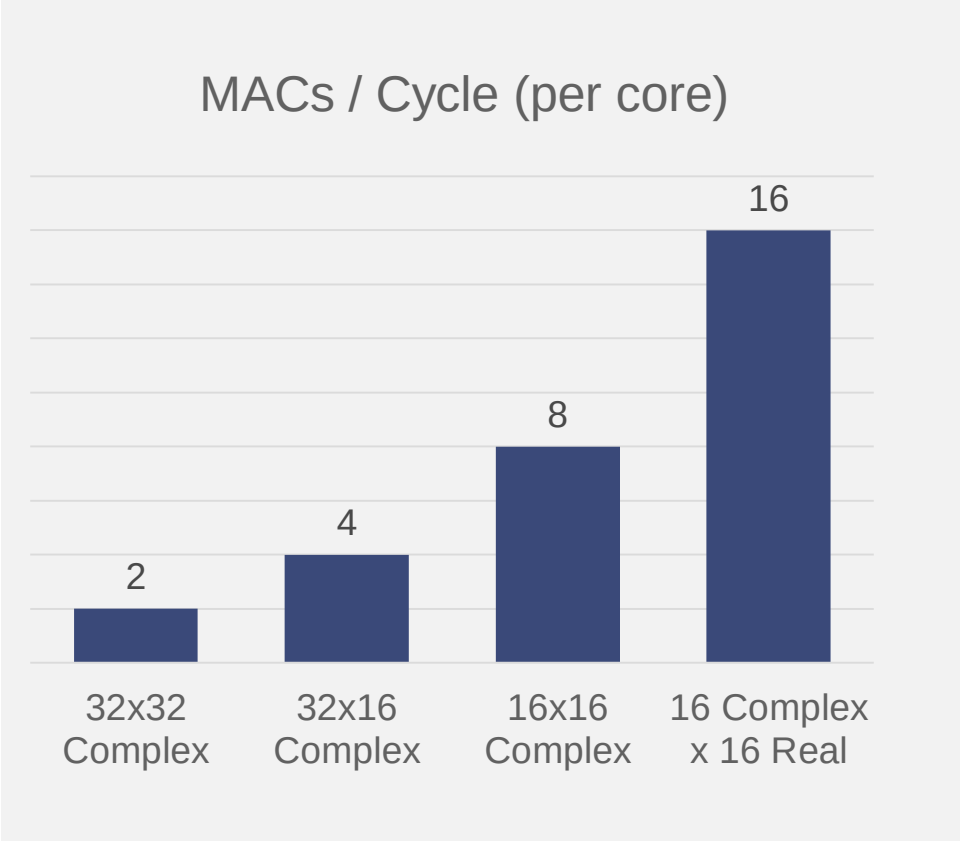
© Copyright 2018 Xilinx

Multi-Precision Support

AI Data Types



Signal Processing Data Types

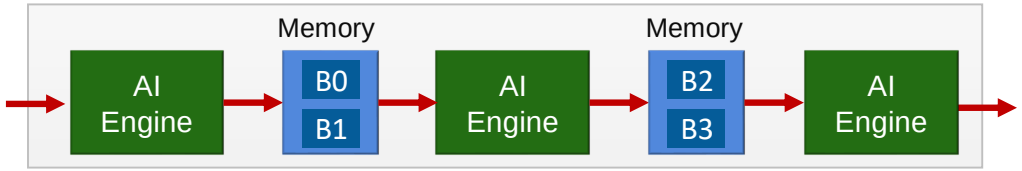


© Copyright 2018 Xilinx

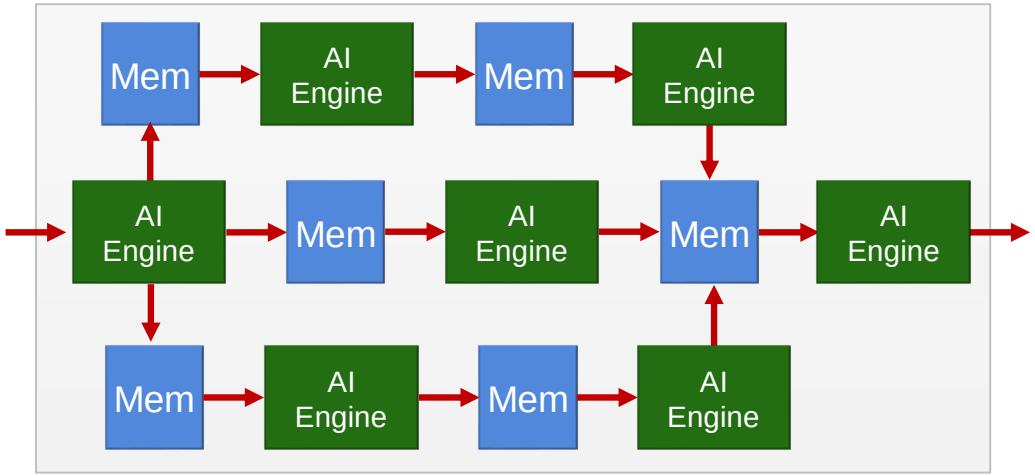
Data Movement Architecture

Memory Communication (neighbor)

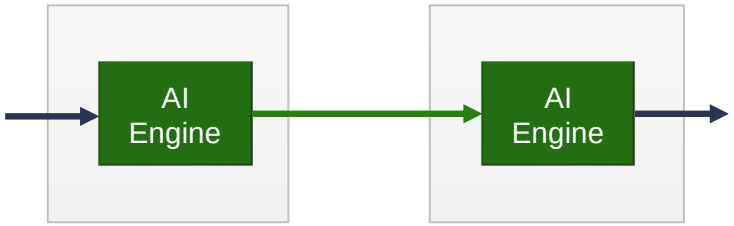
Dataflow Pipeline



Dataflow Graph

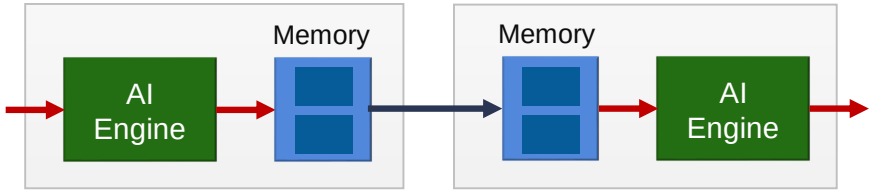


Cascade Streaming

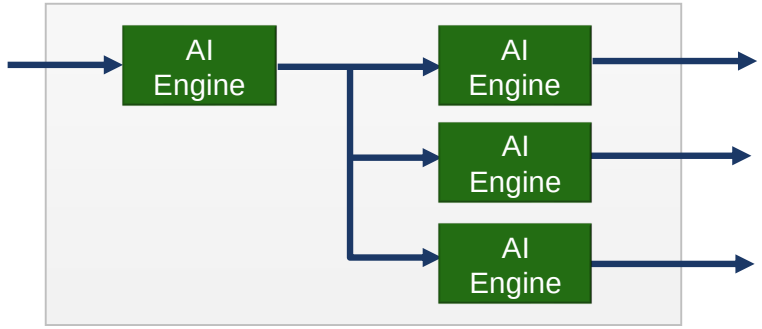


Streaming Communication (non-neighbor)

Non-Neighbor



Streaming Multicast

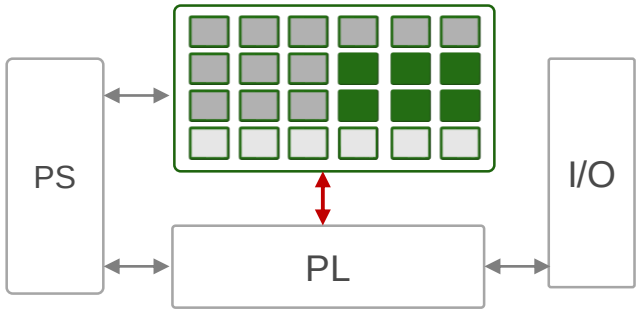


Memory Interface
Stream Interface
Cascade Interface

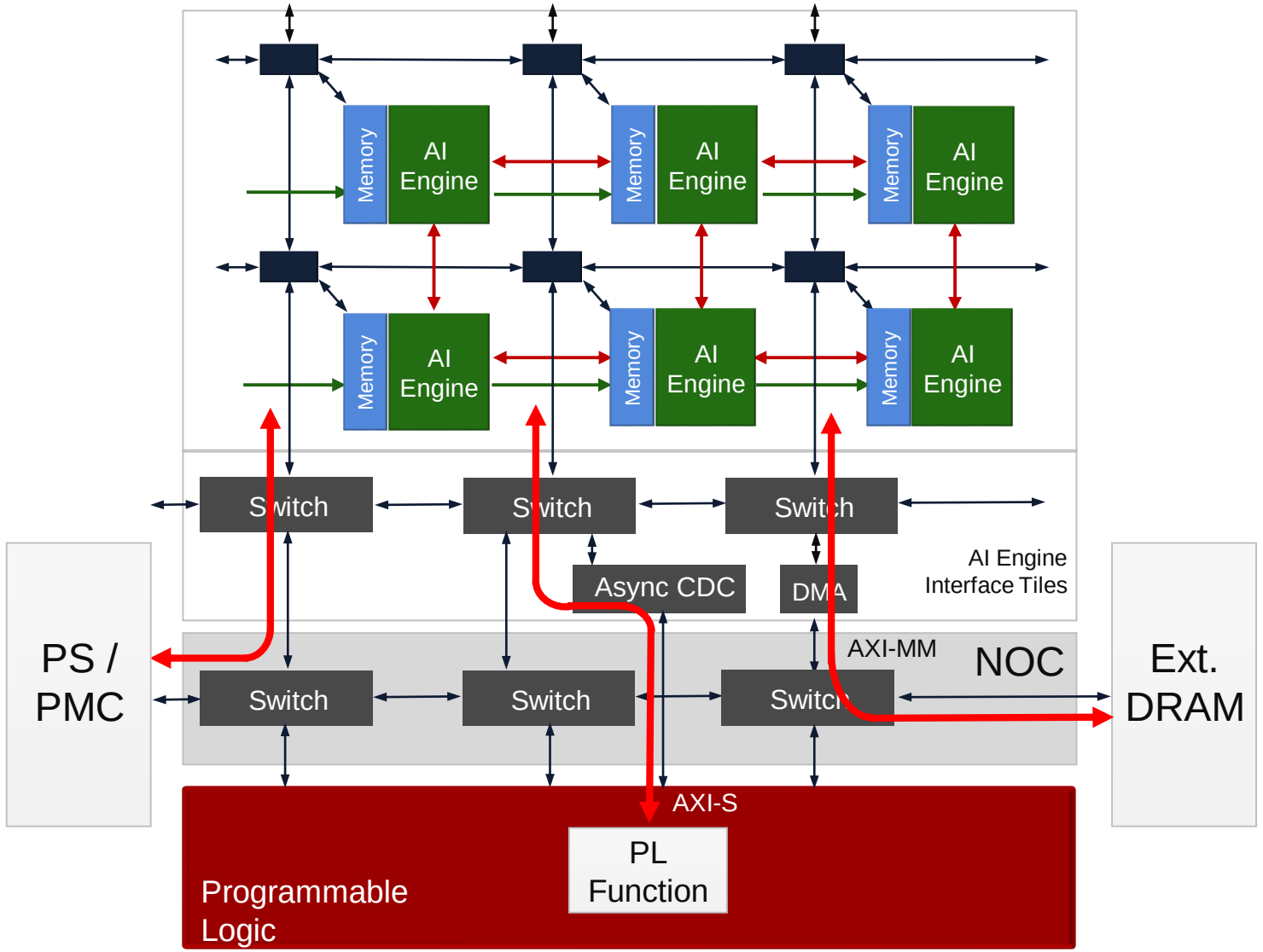
© Copyright 2018 Xilinx

XILINX

AI Engine Integration with Versal™ ACAP



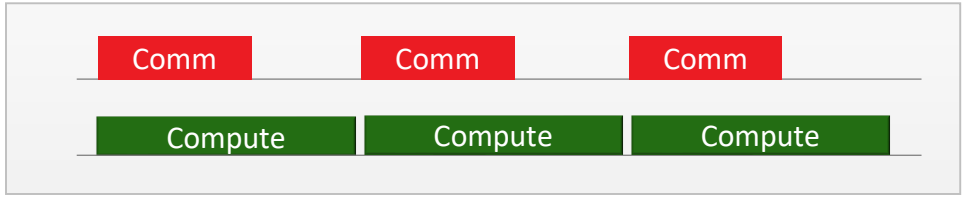
- **TB/s of Interface Bandwidth**
  - AI Engine to Programmable Logic
  - AI Engine to NOC
- **Leveraging NOC connectivity**
  - PS manages Config / Debug / Trace
  - AI Engine to DRAM (no PL req'd)



© Copyright 2018 Xilinx

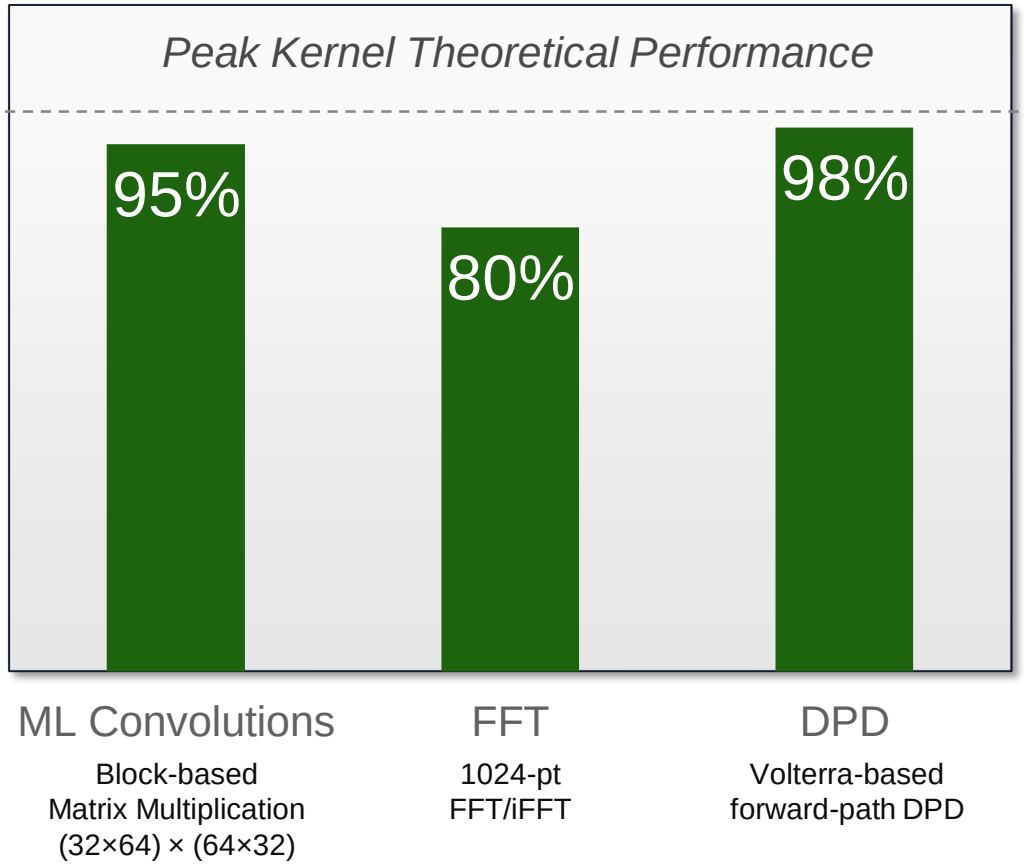
AI Engine Delivers High Compute Efficiency

- **Adaptable, non-blocking interconnect**
  - Flexible data movement architecture
  - Avoids interconnect “bottlenecks”
- **Adaptable memory hierarchy**
  - Local, distributed, shareable = extreme bandwidth
  - No cache misses or data replication
  - Extend to PL memory (BRAM, URAM)
- **Transfer data while AI Engine Computes**



Overlap Compute and Communication

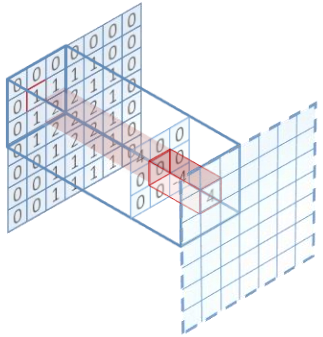
Vector Processor Efficiency



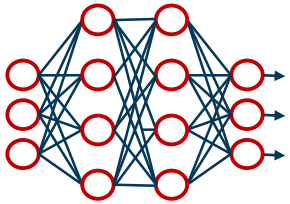
© Copyright 2018 Xilinx

AI Inference on Versal™ ACAP

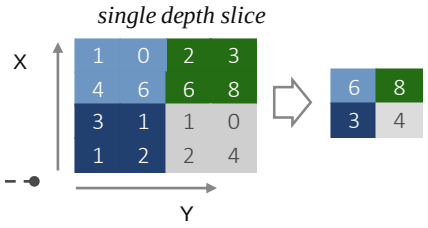
Convolutions



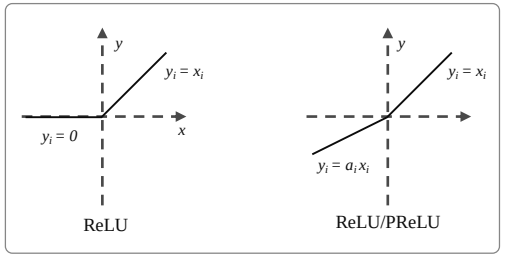
Fully Connected Layers



Pooling



Activations



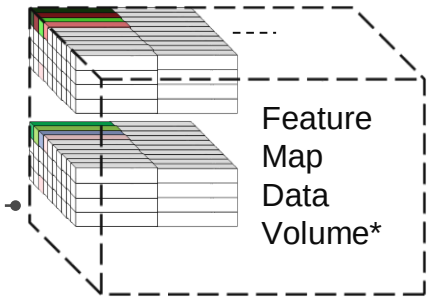
Processing System

AI Engines

Programmable Logic

I/O (GT, ADC/ DAC)

- Video
- Genomics
- Storage
- Database
- Network IPS
- Risk modeling



Custom Memory Hierarchy

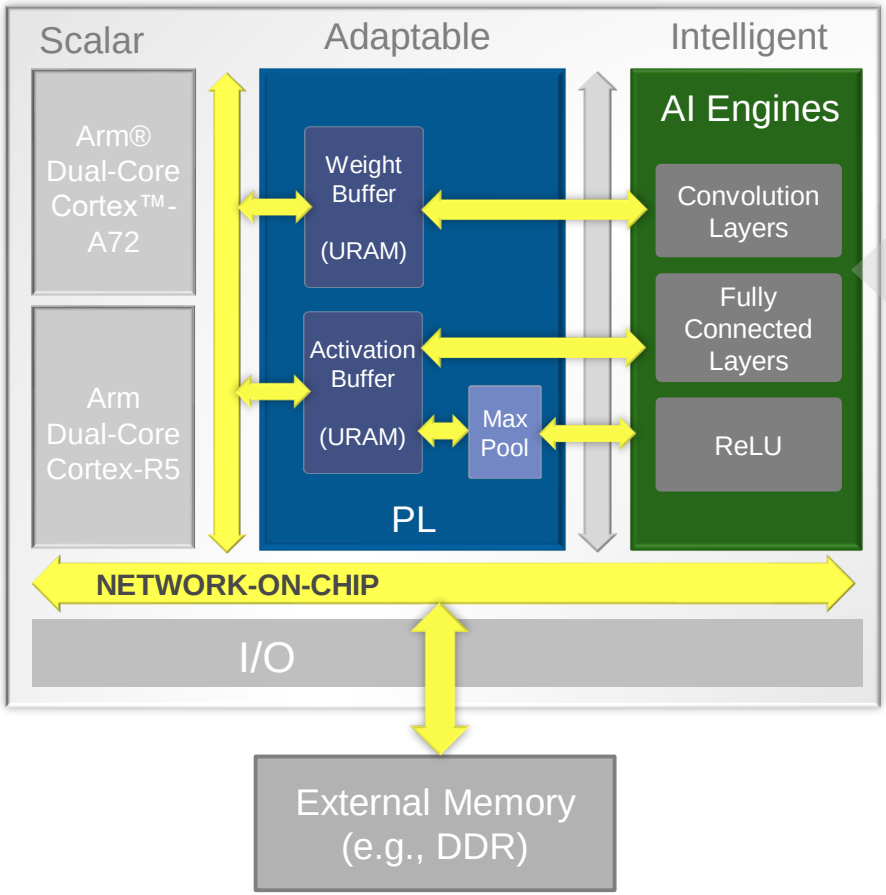
*Figure credit: https://en.wikipedia.org/wiki/Convolutional_neural_network

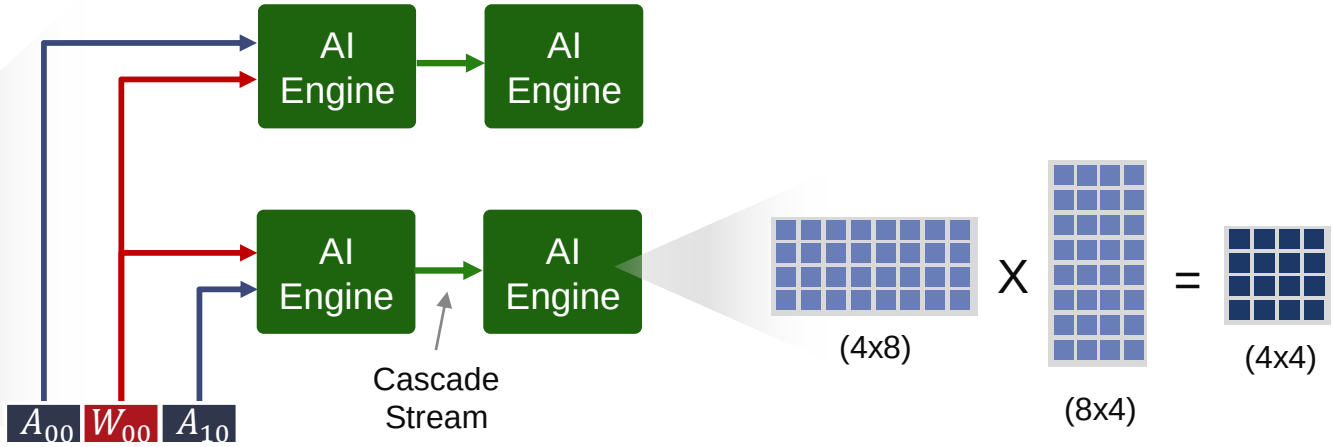
© Copyright 2018 Xilinx

AI Inference Mapping on Versal™ ACAP

A = Activations
W = Weights

$$\begin{bmatrix} A_{00} & A_{01} \\ A_{10} & A_{11} \end{bmatrix} \times \begin{bmatrix} W_{00} & W_{01} \\ W_{10} & W_{11} \end{bmatrix} = \begin{bmatrix} A_{00} \times \boldsymbol{W_{00}} + A_{01} \times W_{10} & \dots \\ A_{10} \times \boldsymbol{W_{00}} + A_{11} \times W_{10} & \dots \end{bmatrix}$$

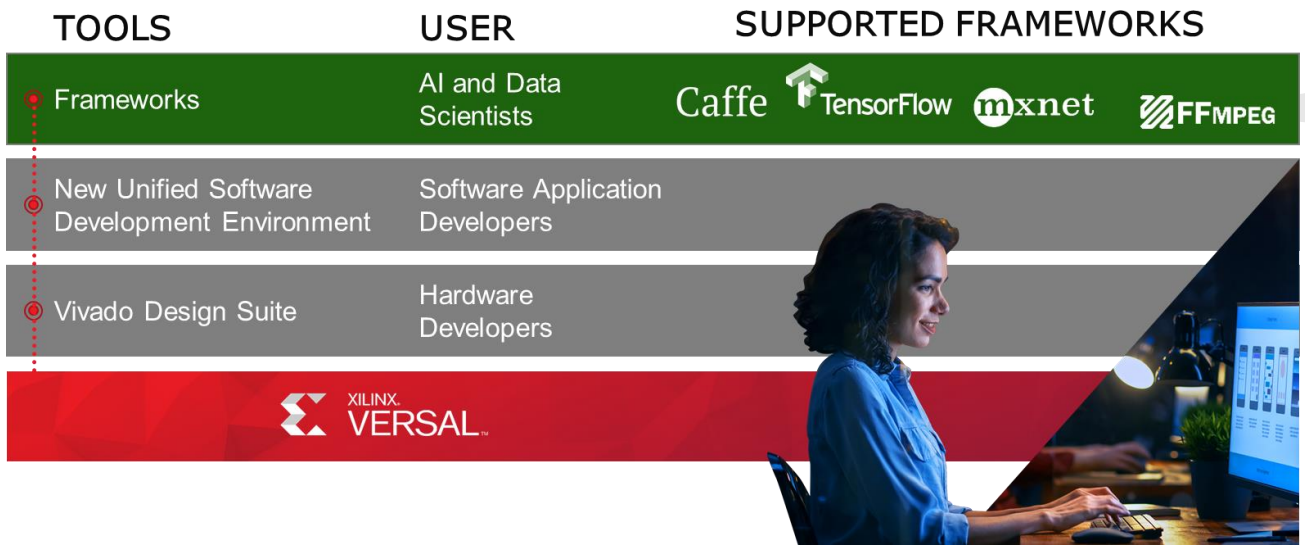




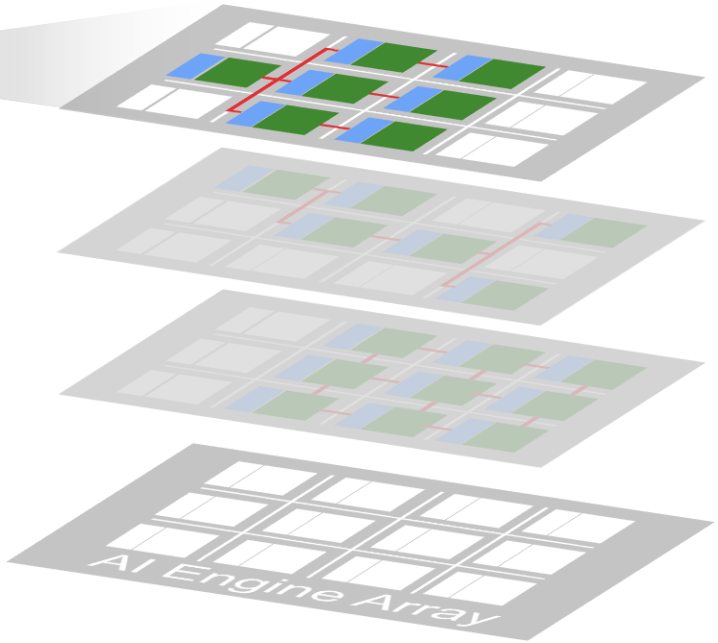
- Custom memory hierarchy
  - Buffer on-chip vs off-chip; Reduce latency and power
- Stream Multi-cast on AI interconnect
  - Weights and Activations
  - Read once: reduce memory bandwidth
- AI-optimized vector instructions (128 INT8 mults/cycle)

© Copyright 2018 Xilinx

Frameworks for Any Developer



Domain Specific Architecture
(e.g. AI Inference)



Architecture Overlay

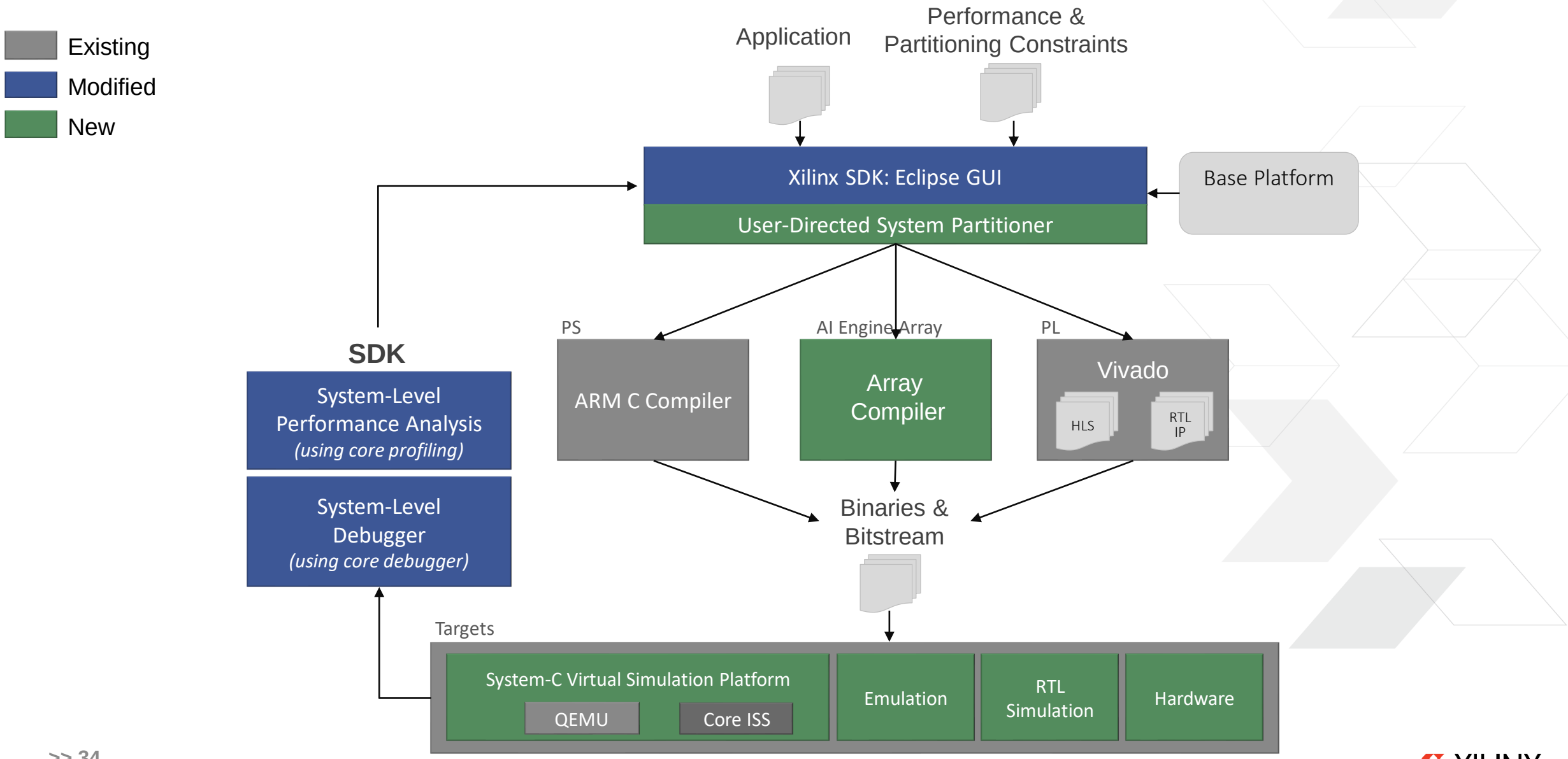
Data Flow
w/ Xilinx libraries

Kernel Program
Data Flow w/ user defined libraries

Target Domain Specific Architectures – No HW Design Experience Required

© Copyright 2018 Xilinx

Unified Tool Chain for Device Programming



© Copyright 2018 Xilinx

XILINX®

VERSAL™

AI RF Series

HBM Series

AI Core Series

Premium Series

AI Edge Series

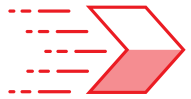
Prime Series

© Copyright 2018 Xilinx

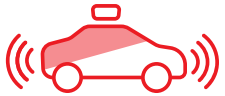
Versal Roadmap

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

© Copyright 2018 Xilinx

XILINX.

Getting Started

Visit www.xilinx.com/versal

> Watch ACAP Intro video

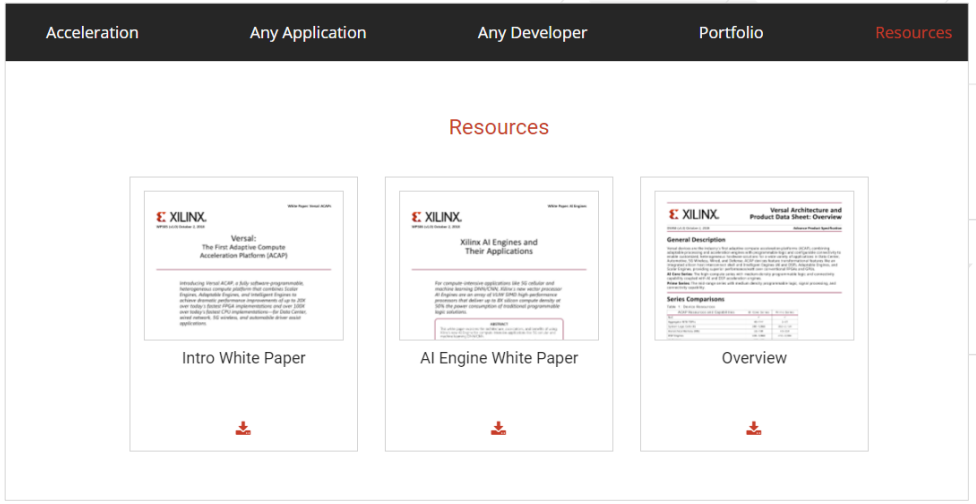
> Subscribe to mailing list for the latest news

View documentation and resources

> Data Sheet Overview

> Product Tables

> Versal Architecture and AI Engine White Papers



© Copyright 2018 Xilinx

Key Take-Aways

Versal: The First ACAP

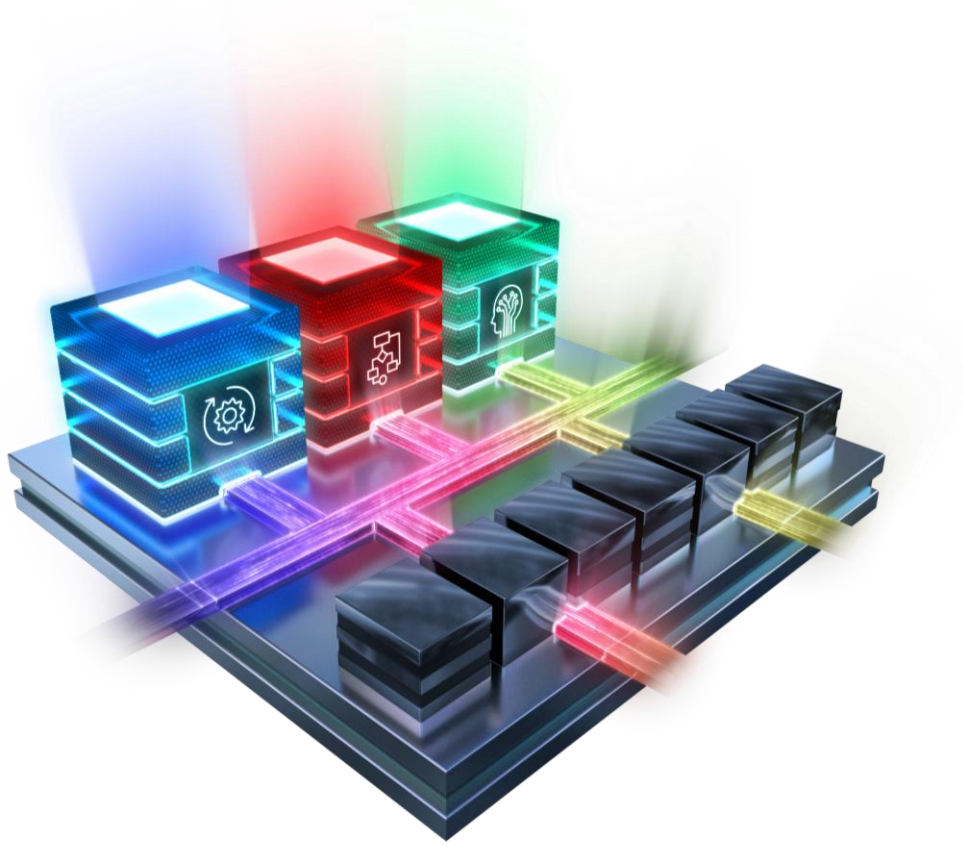
- Heterogeneous Acceleration
- For Any Application
- For Any Developer

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



© Copyright 2018 Xilinx

Building the Adaptable, Intelligent World

Thank you!

Contact Info:
jasonv@xilinx.com

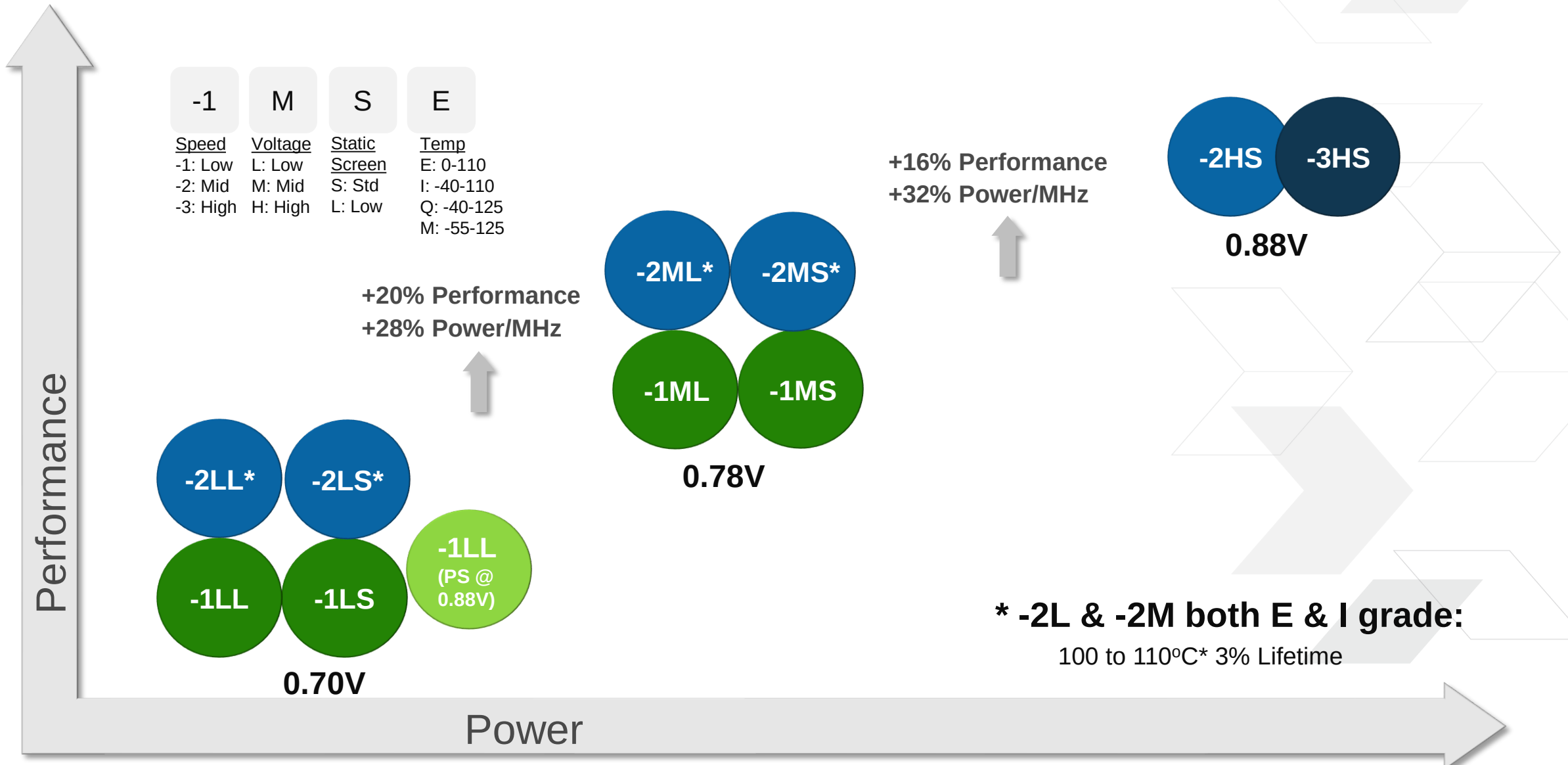
XILINX

© Copyright 2019 Xilinx

Appendix

© Copyright 2018 Xilinx

Versal™ Speed Grade Performance vs. Power



*** -2L & -2M both E & I grade:**

100 to 110°C* 3% Lifetime

© Copyright 2018 Xilinx

XILINX.