Versal: The New Xilinx Adaptive Compute Acceleration Platform (ACAP) in 7nm

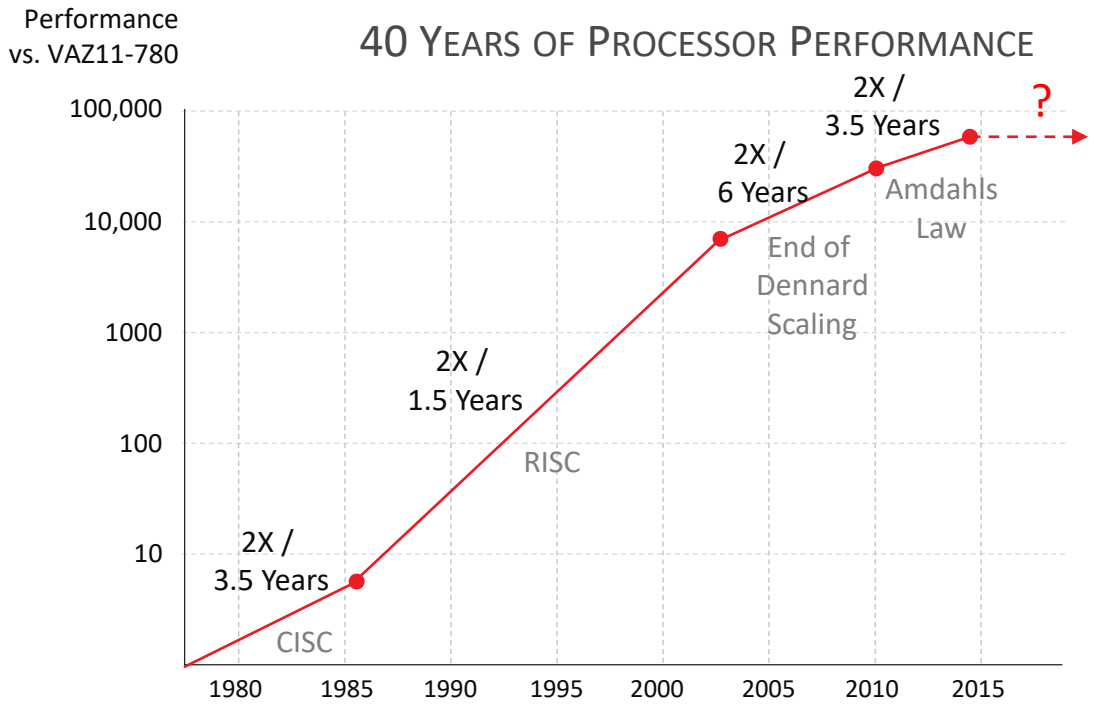
Presented By

Kees Vissers
Fellow

February 25,
FPGA 2019

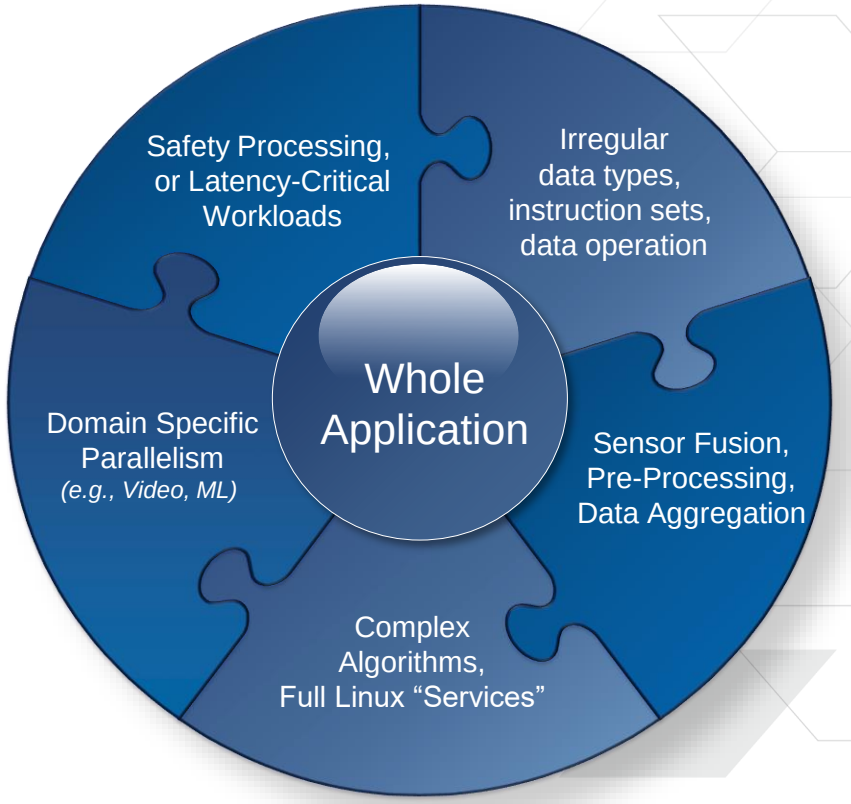
Technology scaling coming to an end

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.

Need for a New Programming Paradigm

GitHub

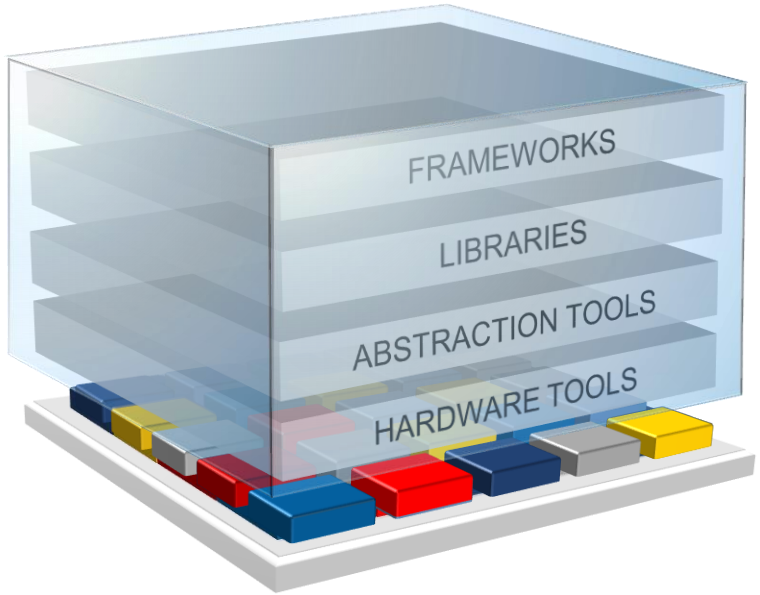
Ecosystem of Libraries

Need a Scalable, Unified Platform

Software Developer

Needs Agility and Abstraction

Modify, Design, Add Code

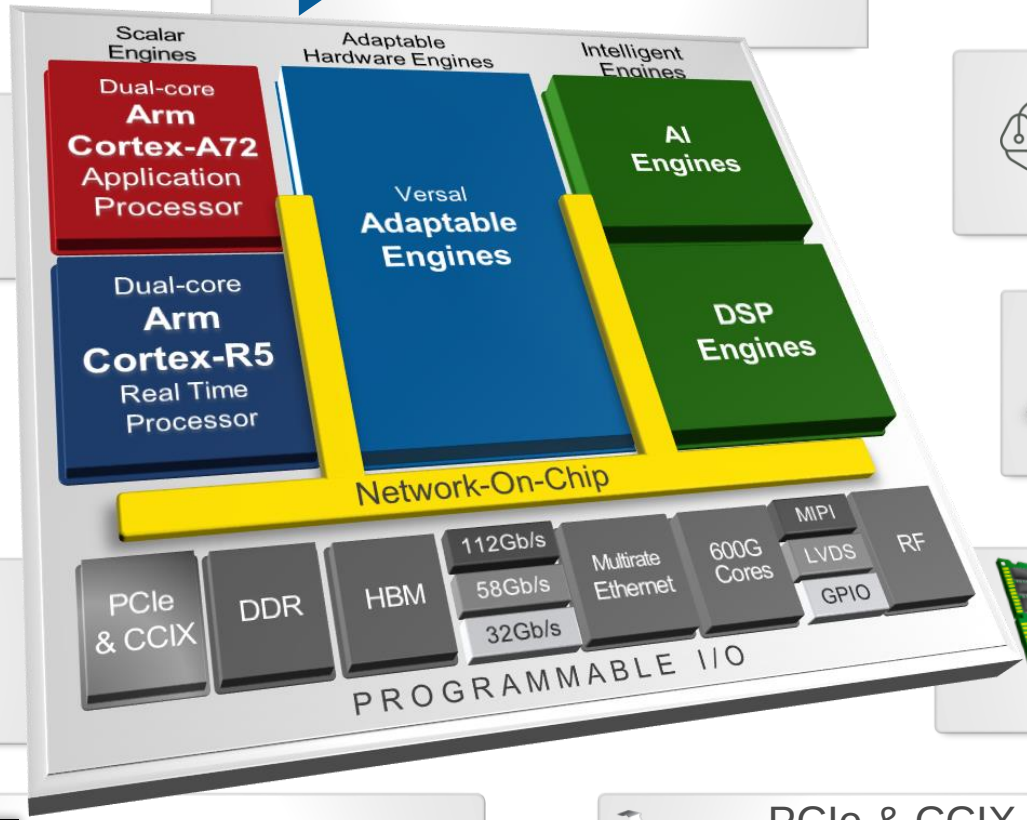


Hardware Developer

Needs Flexibility to Optimize for Performance/Power

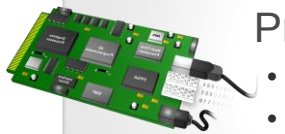
Versal Architecture Overview

Adaptable Engines
2X compute density



Scalar Engines
- Platform Control
- Edge Compute

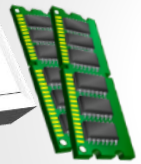
AI Engines
- AI Compute
- Diverse DSP workloads



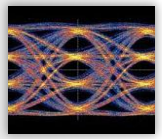
Protocol Engines
- Integrated 600G cores
- 4X encrypted bandwidth

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

Programmable I/O
- Any interface or sensor
- Includes 4.2Gb/s MIPI



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



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

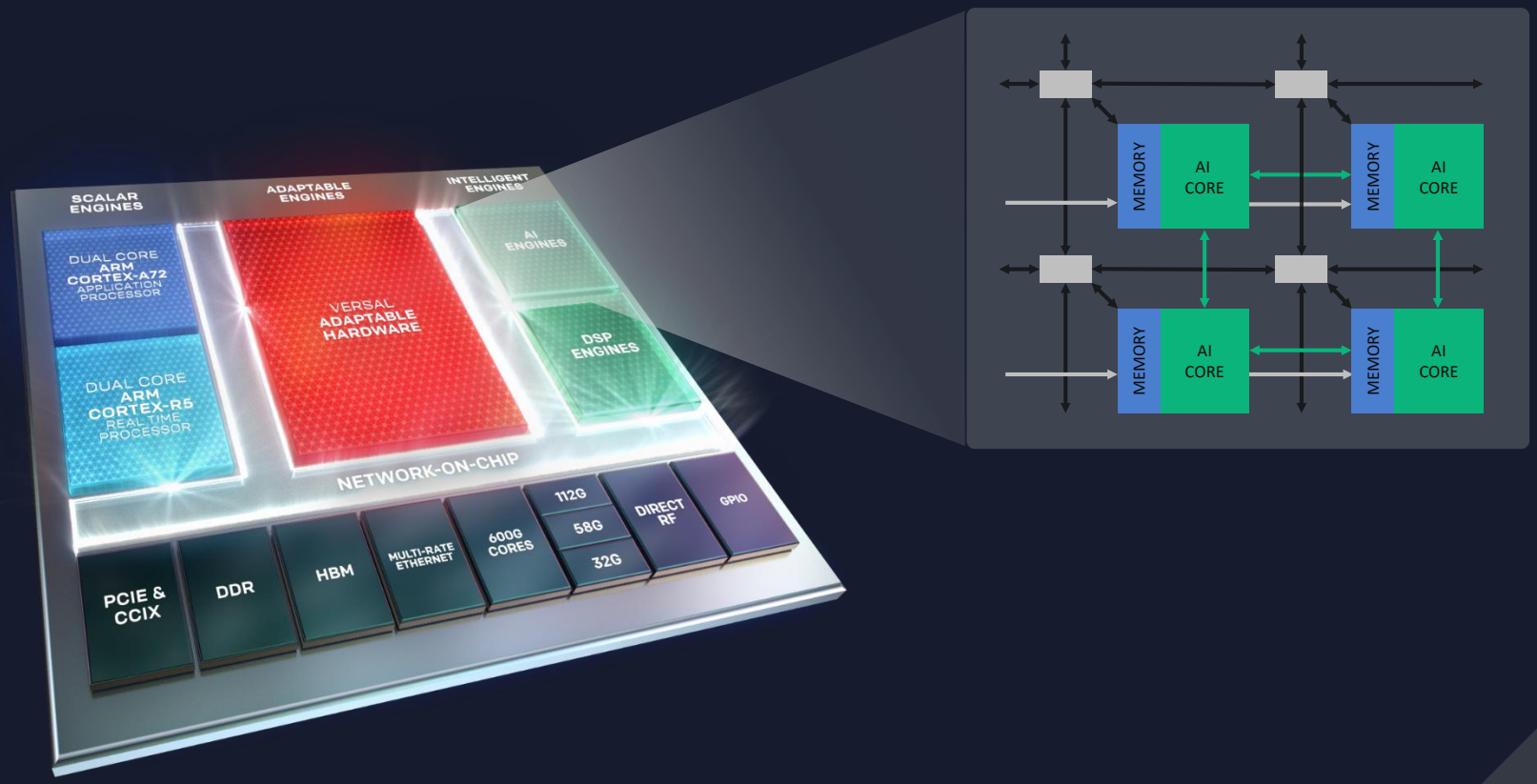
PCIe & CCIX
- 2X PCIe & DMA bandwidth
- Cache-coherent interface to accelerators

XILINX

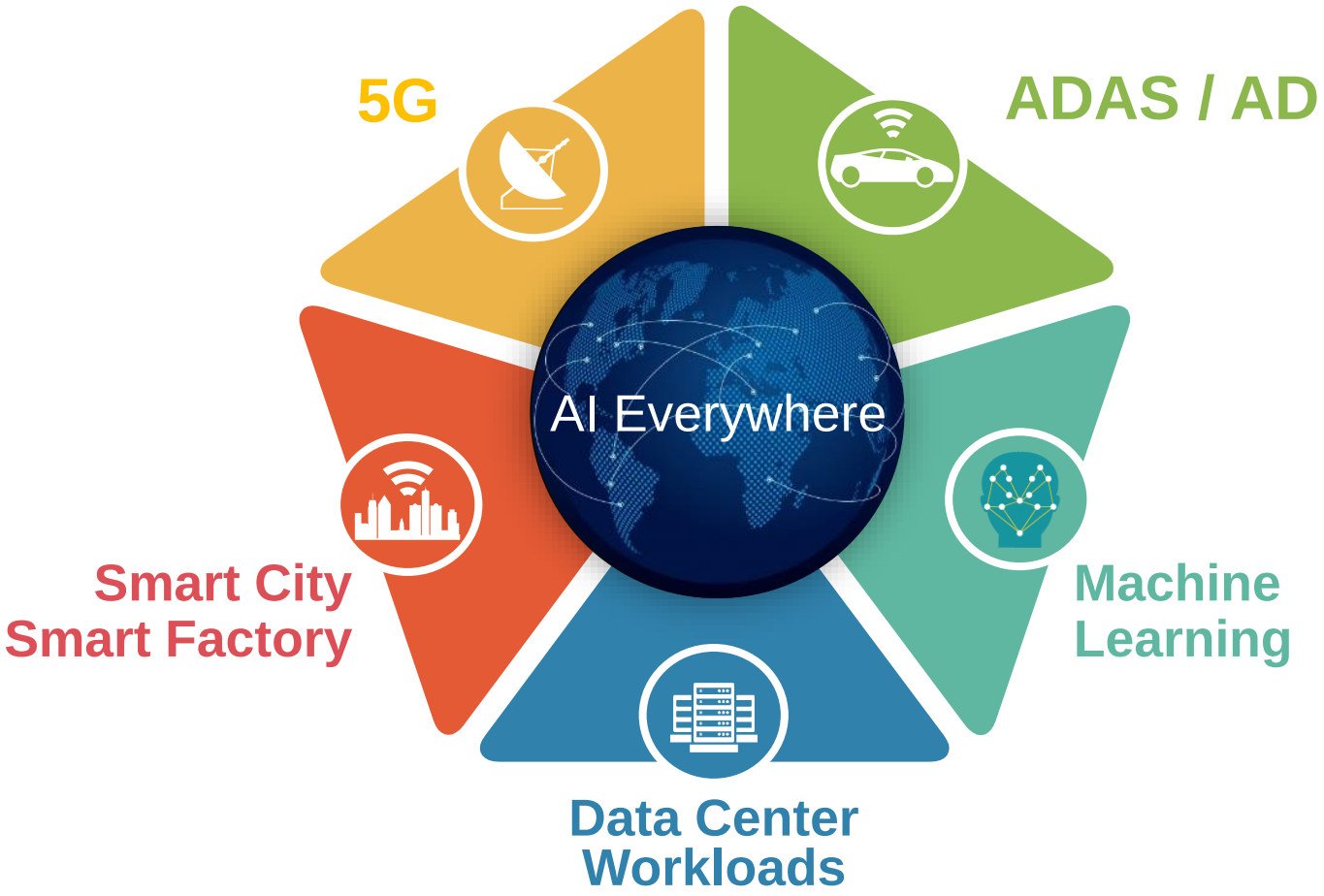
Overview

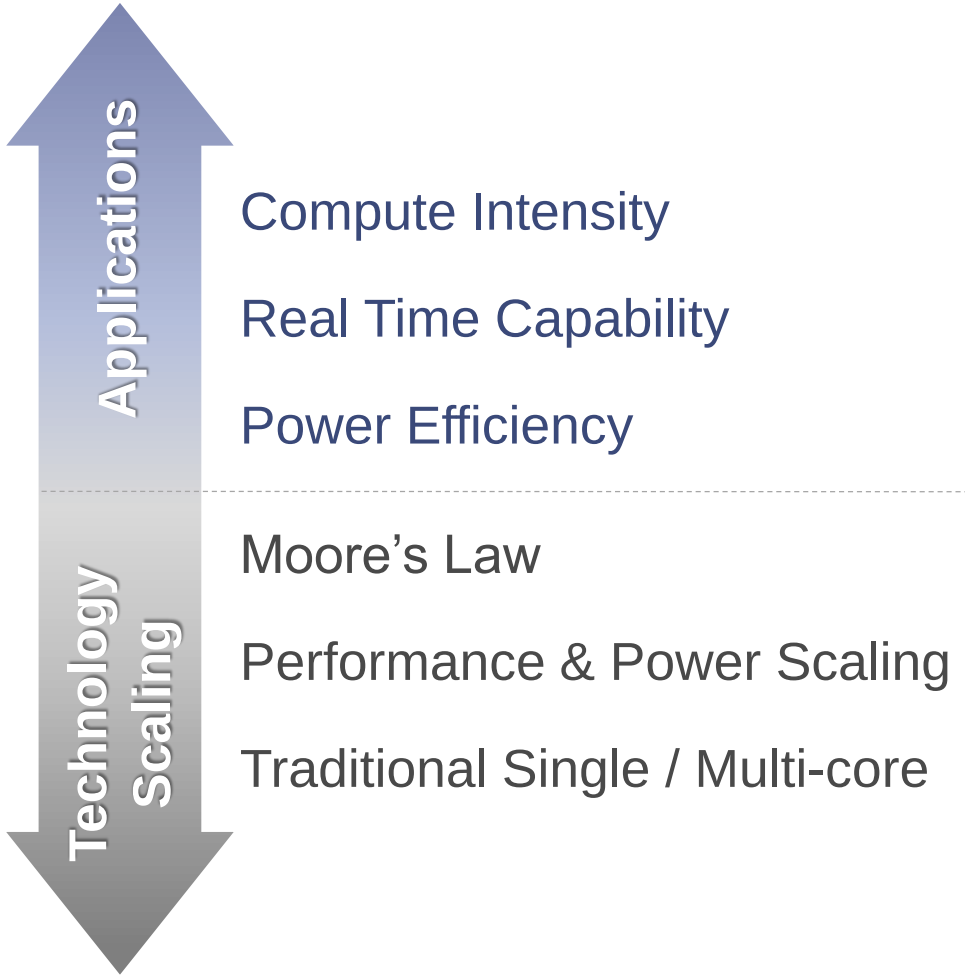
- Adaptable Engines: Brian Gaide (9.15 today, directly following this talk)
- Network on Chip: Ian Swarbrick (9.45 Tuesday)
- Rest of this Talk: Adaptable Intelligent Engines: New processors + interconnect

Motivation for AI Engine



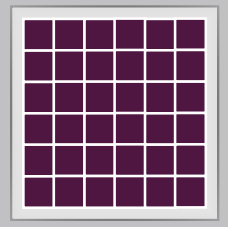
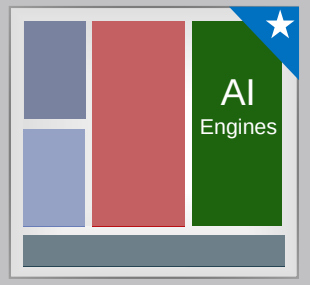
Motivation for AI Engine





Dynamic Markets Require Adaptable Compute Acceleration

Delivering Adaptable Compute Acceleration

	CPU (Sequential)	GPU (Parallel)	ACAP	Custom ASIC
				
SW Programmable	✓	✓	✓	✓
HW Adaptable	–	–	✓	–
Workload Flexibility	✓	✓	✓	–
Throughput vs. Latency	–	–	✓	✓
Device / Power Efficiency	–	–	✓	✓

ACAP
w/ AI Engine

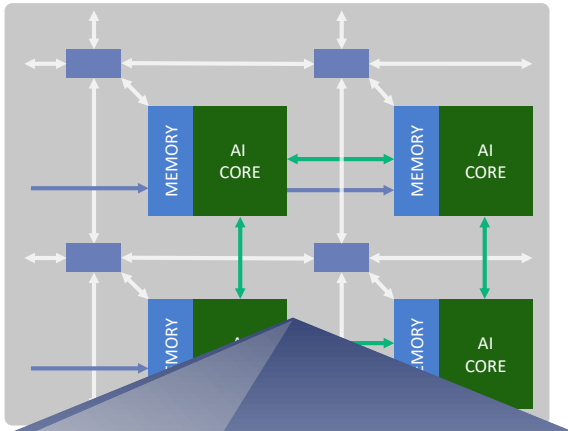
XILINX

Introducing the AI Engine

SW Programmable

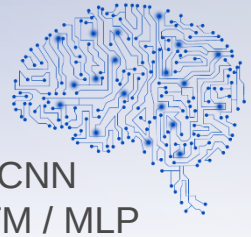
Deterministic

Efficient



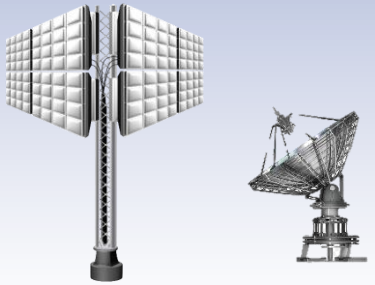
- 1GHz+ Multi-precision Vector Processor
- High bandwidth extensible memory
- Up to 400 AI Engines per device
- 8X Compute Density
- 40% Lower Power

Artificial Intelligence



CNN
LSTM / MLP

Signal Processing



Computer Vision

Adaptable. Intelligent.

XILINX.

Software Programmable: Any Developer

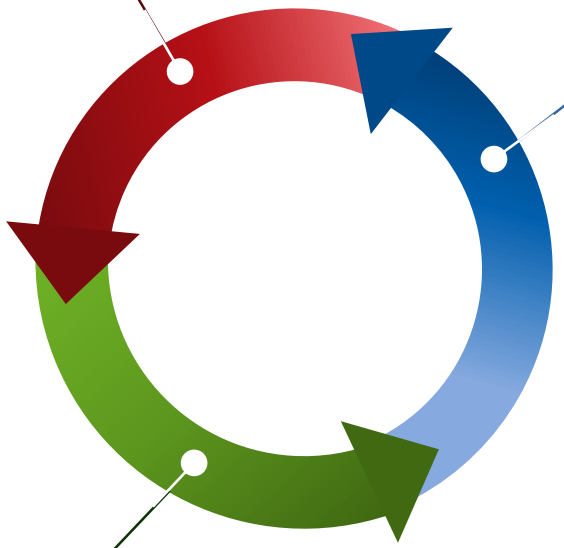
1 Design

C/C++

Frameworks

mxnet TensorFlow Caffe

4G/5G/Radar Library | AI Library | Vision Library



2 Compile

AI Engine Compiler

Run 3

Programming Abstraction Levels

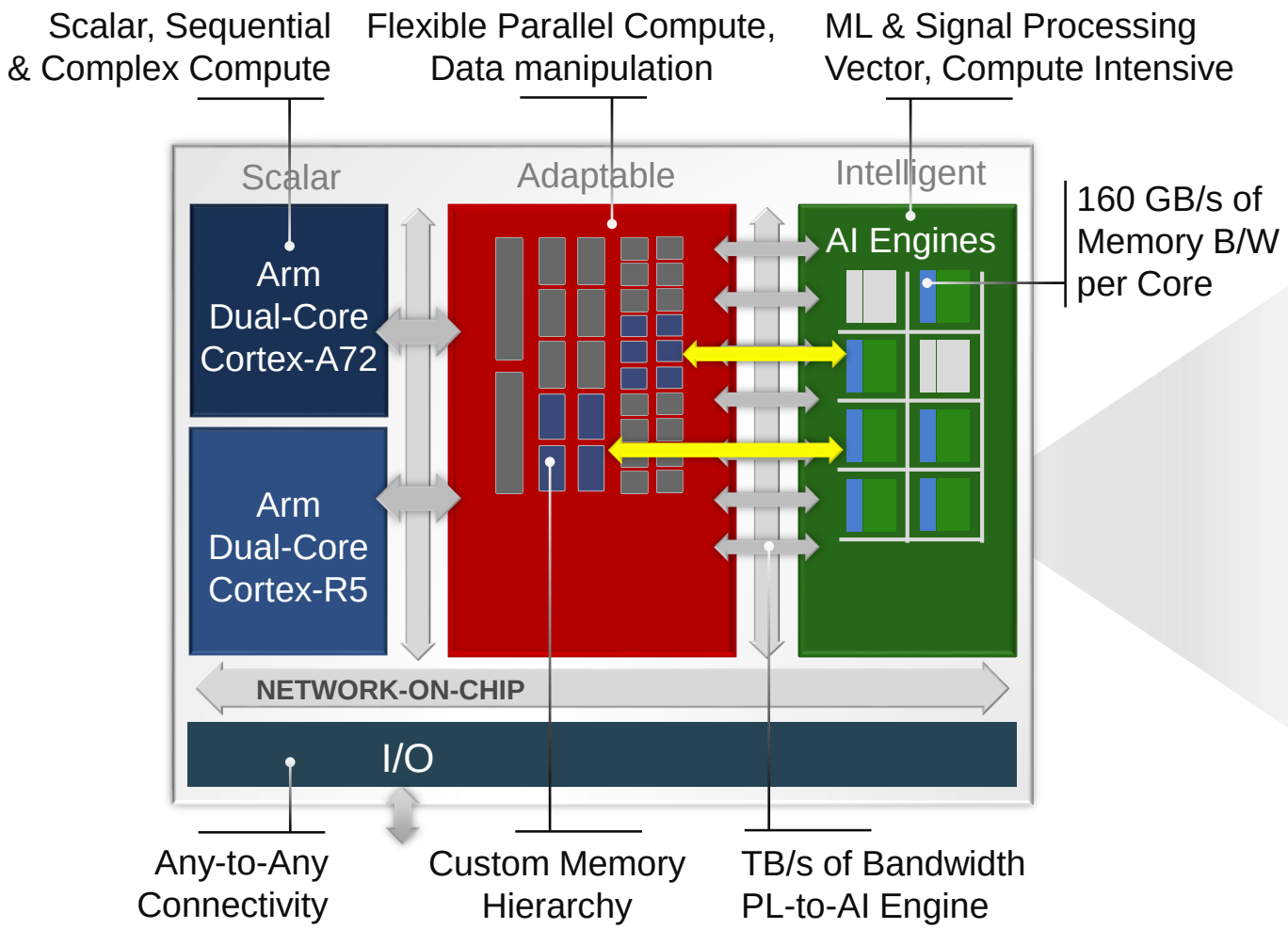
Architecture Overlay

Data Flow
w/ Xilinx libraries

Kernel Program
Data Flow w/ user defined libraries

XILINX

Hardware Adaptable: Accelerating the Whole Application

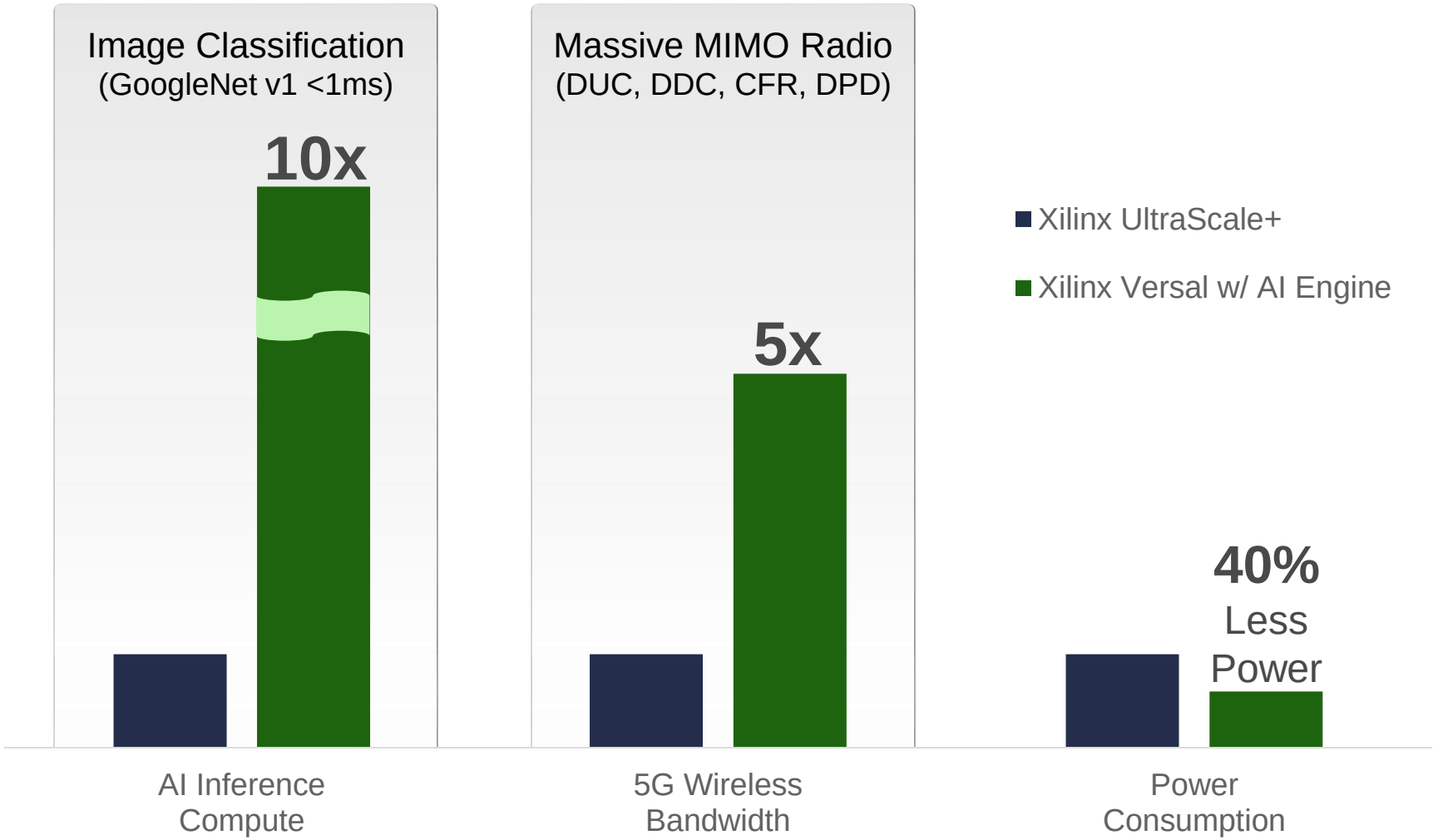


Heterogeneous Acceleration from Data Center to the Edge

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

Delivering Deterministic Performance & Low Latency

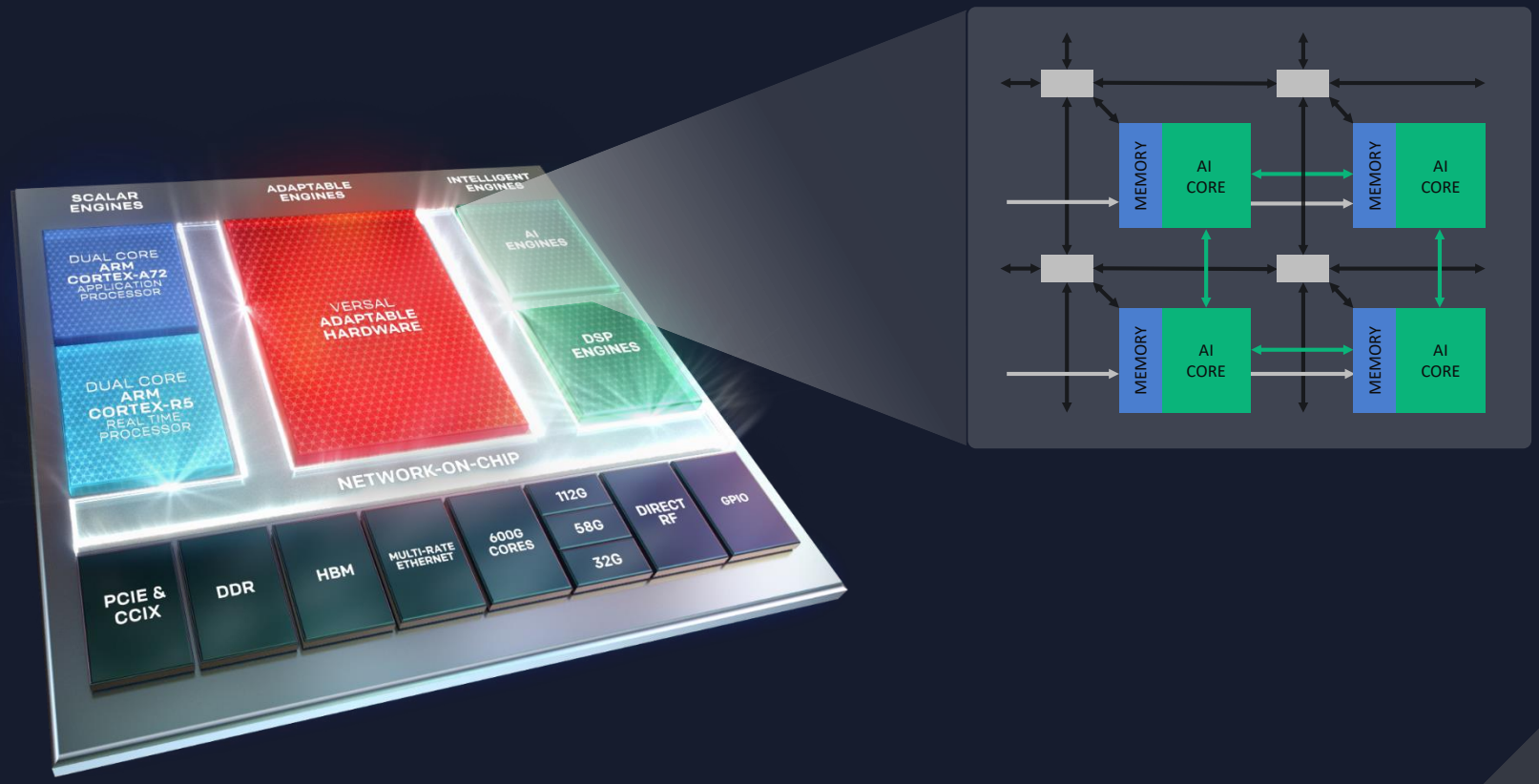
AI Engine Application Performance & Power Efficiency



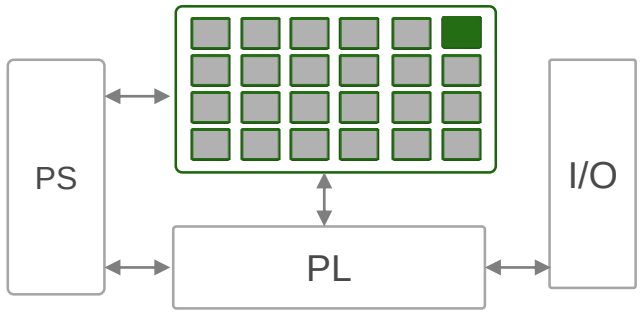
XILINX.

AI Engine

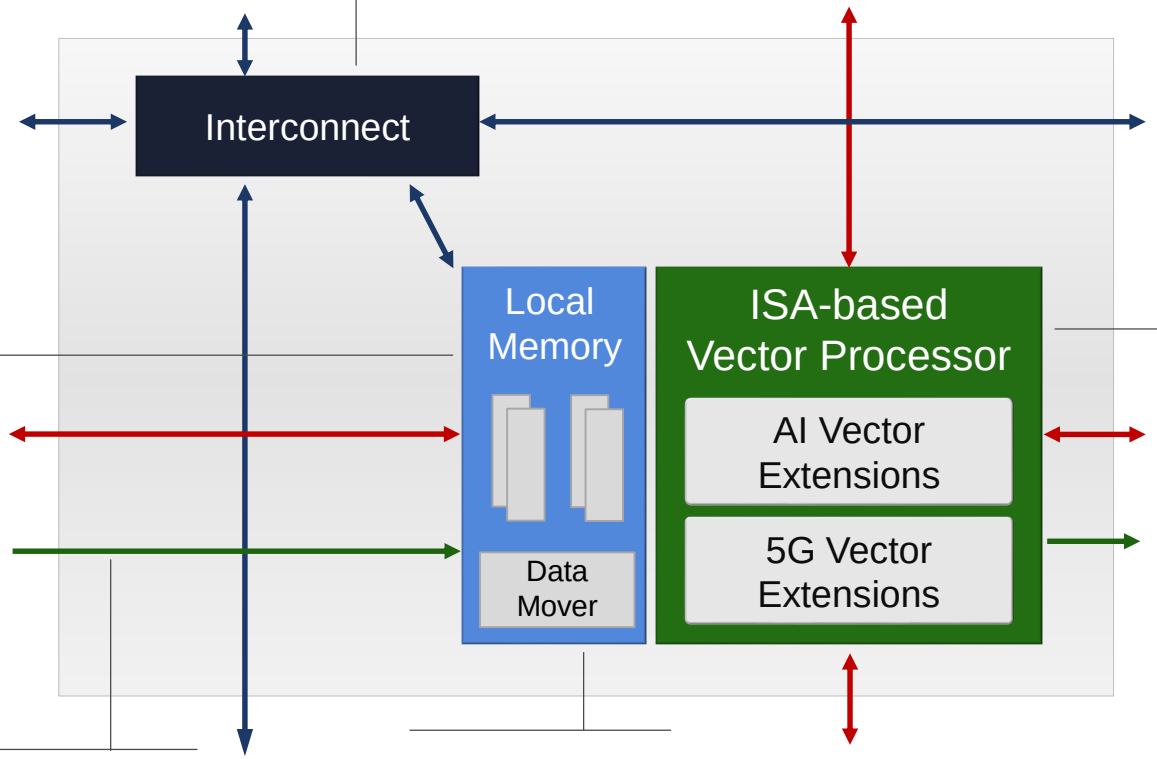
Architecture, Programming & Applications



AI Engine: Tile-Based Architecture



Non-Blocking Interconnect
high GB/s bandwidth per tile



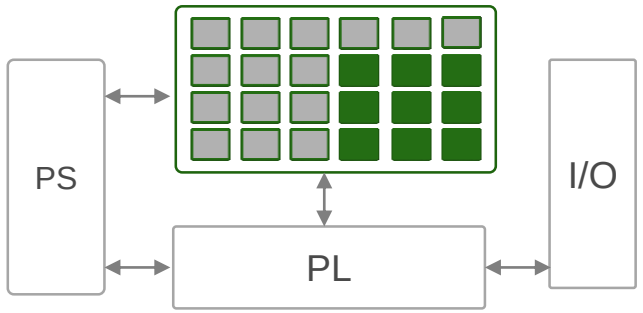
Local Memory
Multi-bank implementation
Shared across neighbor cores

ISA-based Vector Processor
Software Programmable
(e.g., C/C++)

Cascade Interface
Partial results to next core

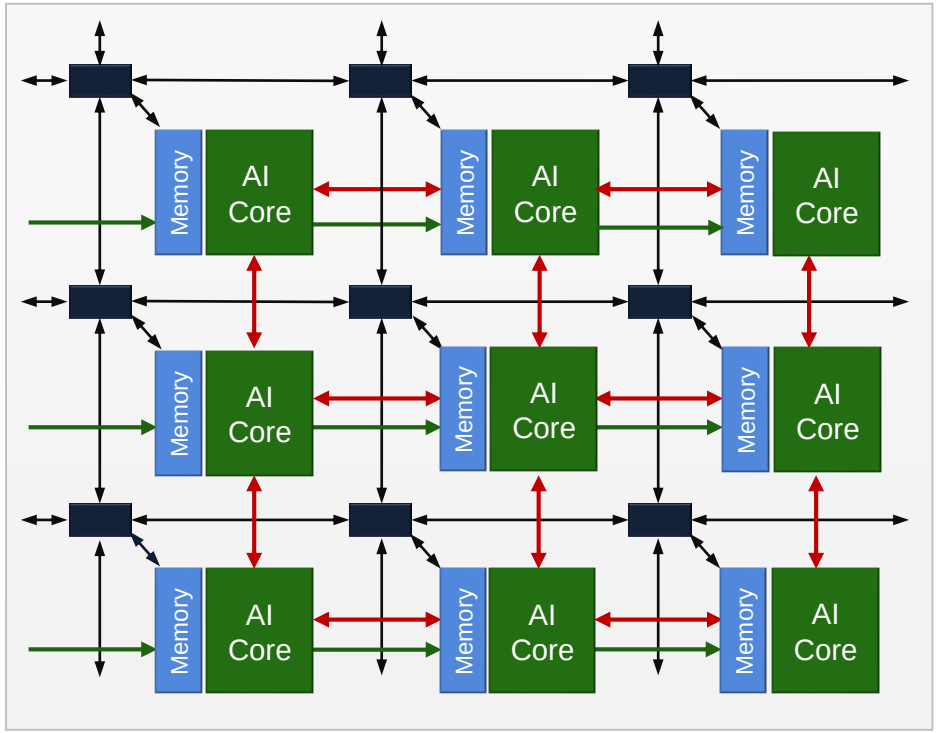
Data Mover
Non-neighbor data communication
Integrated synchronization primitives

AI Engine: Array Architecture



Array of AI Engines

- Increase in compute, memory and communication bandwidth



Modular and scalable architecture

- More tiles = more compute
- Up to 400 per device
  - Versal AI Core VC1902 device

Distributed memory hierarchy

Maximize memory bandwidth

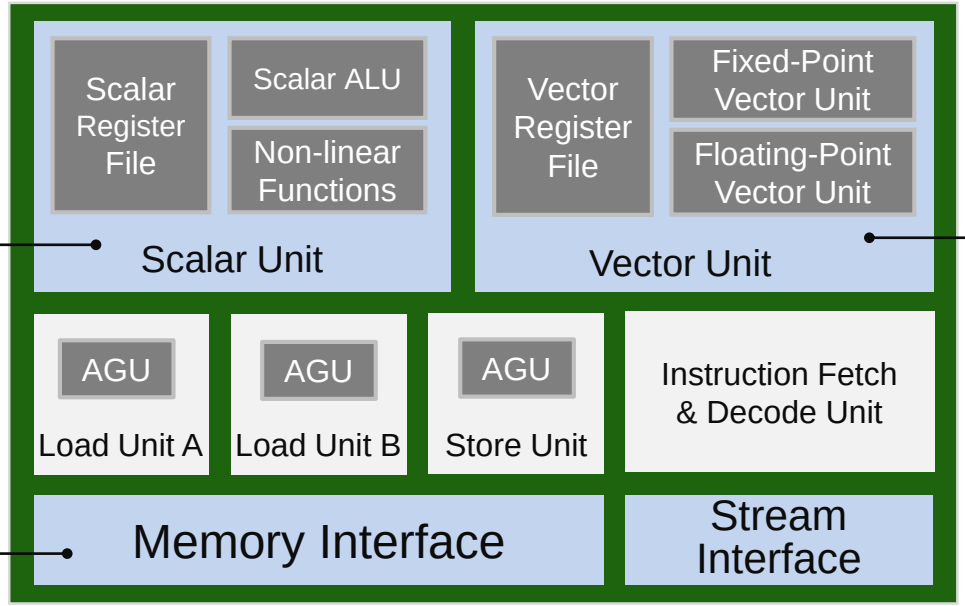
Deterministic Performance & Low Latency

XILINX.

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

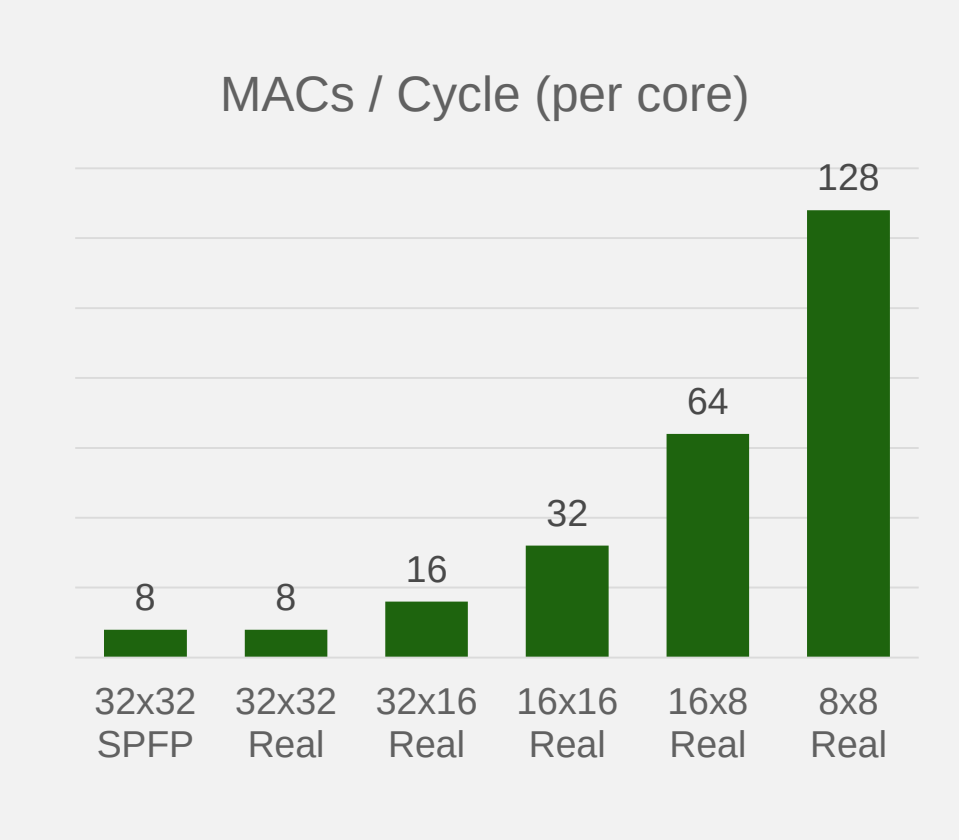
Data Parallelism: SIMD

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

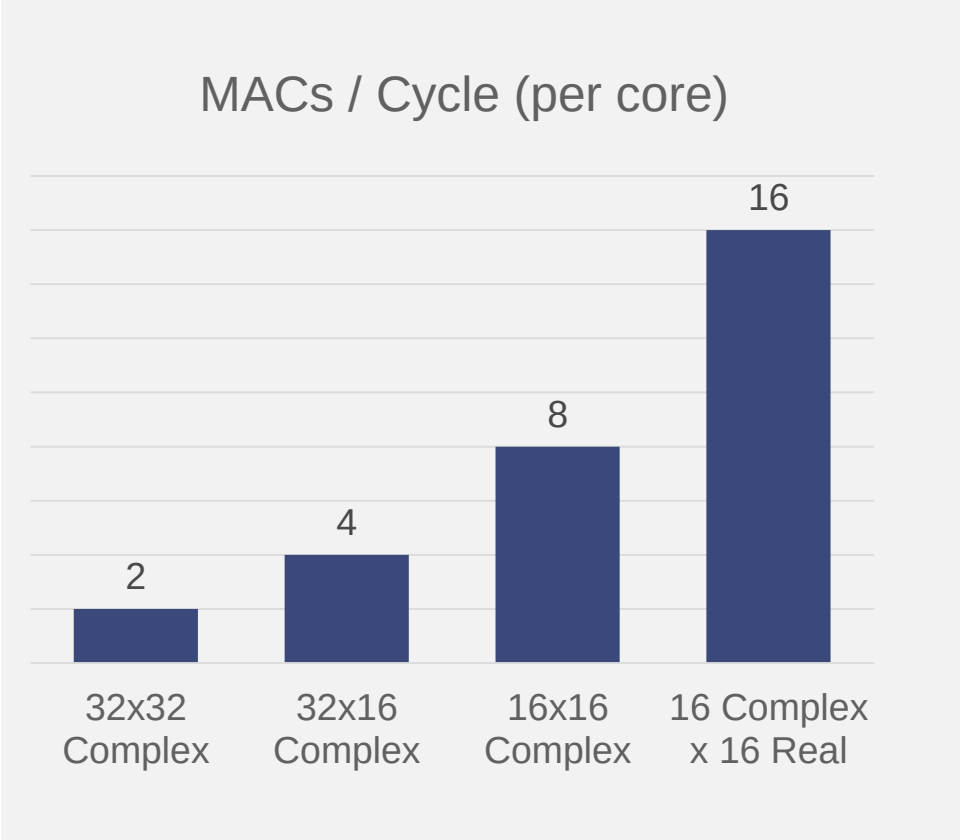
Up to 128 MACs / Clock Cycle per Core (INT 8)

Multi-Precision Support

AI Data Types



Signal Processing Data Types

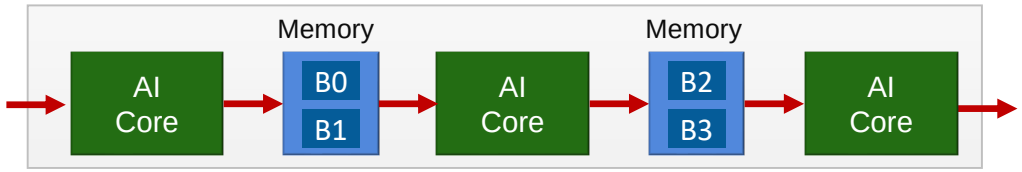


XILINX.

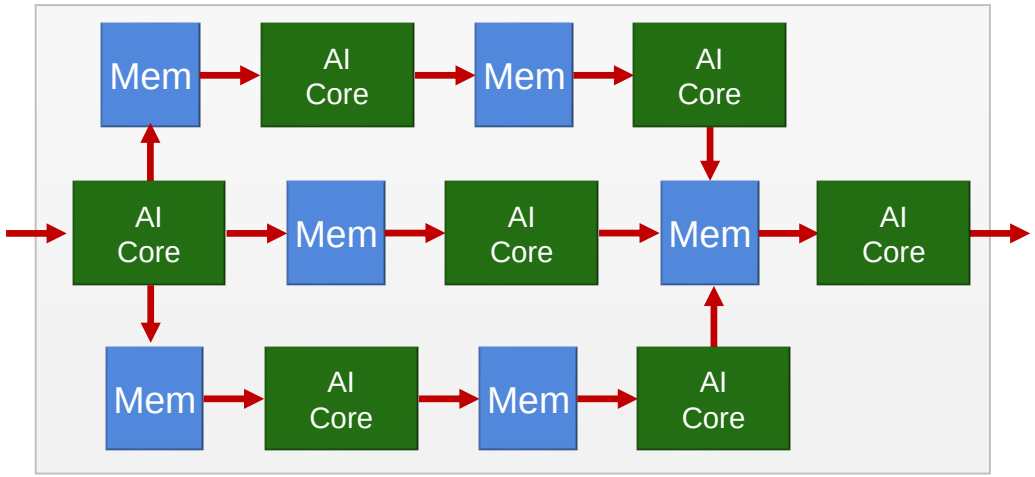
Data Movement Architecture

Memory Communication

Dataflow Pipeline



Dataflow Graph



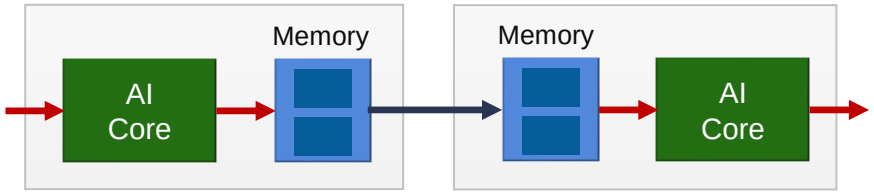
Memory Interface

Stream Interface

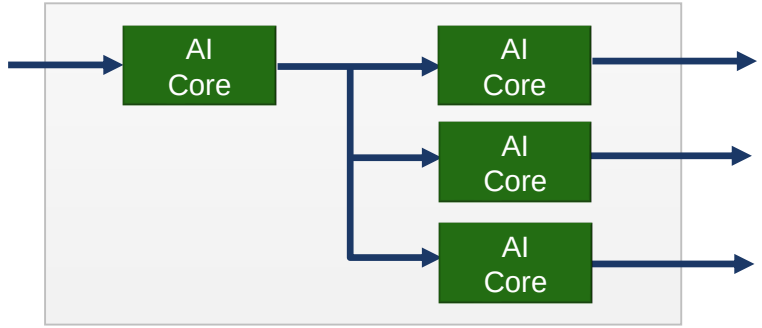
Cascade Interface

Streaming Communication

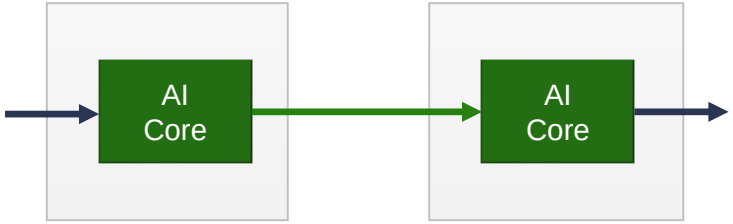
Non-Neighbor



Streaming Multicast

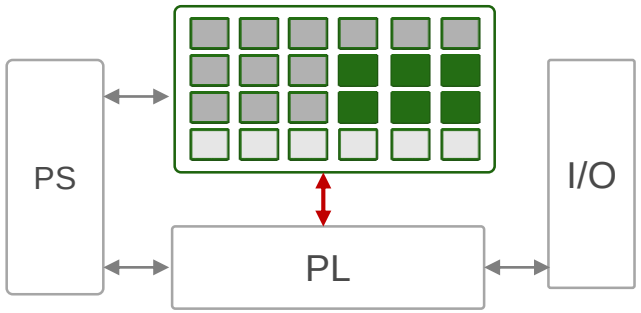


Cascade Streaming

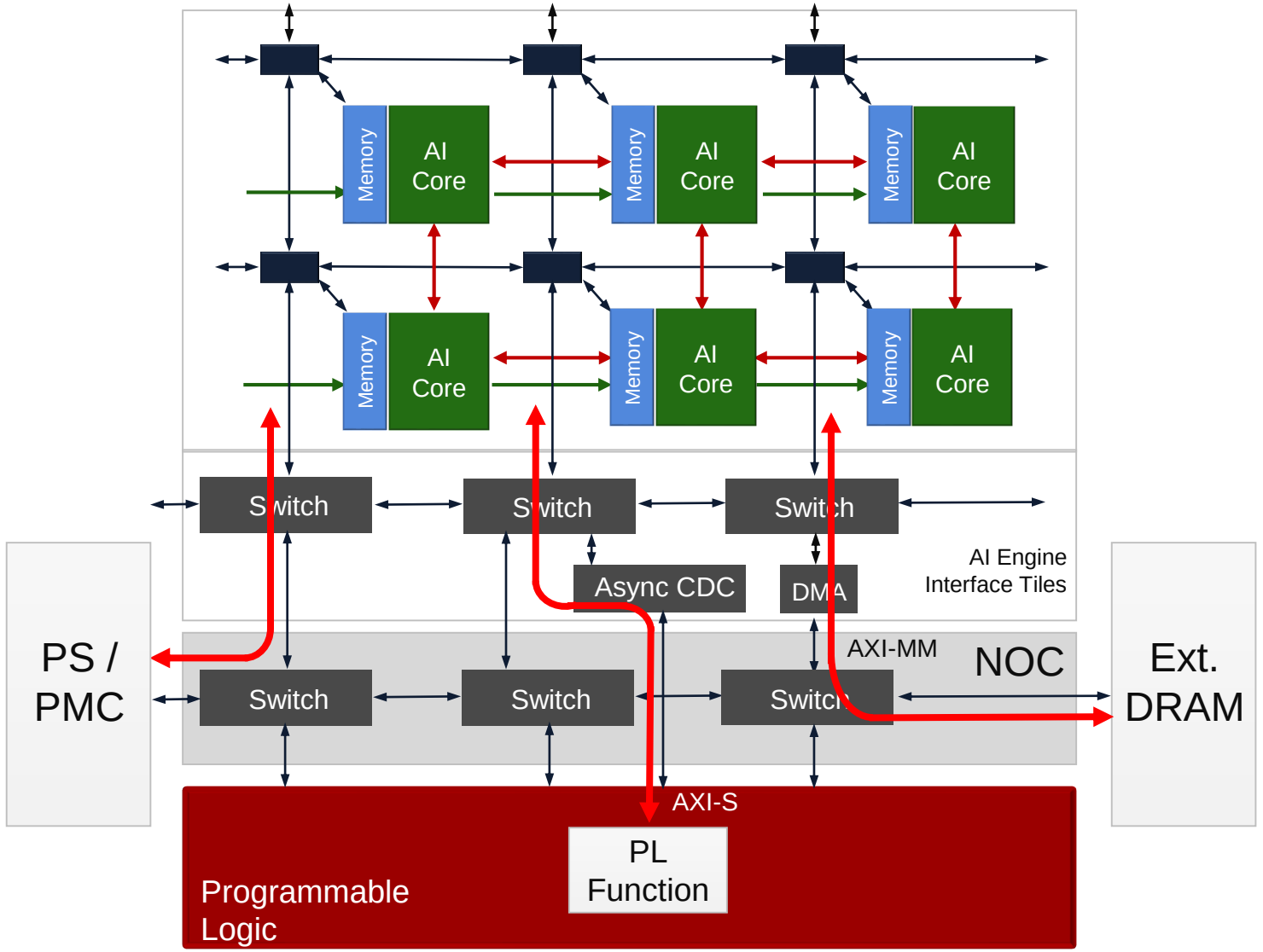


XILINX

AI Engine Integration with Versal ACAP

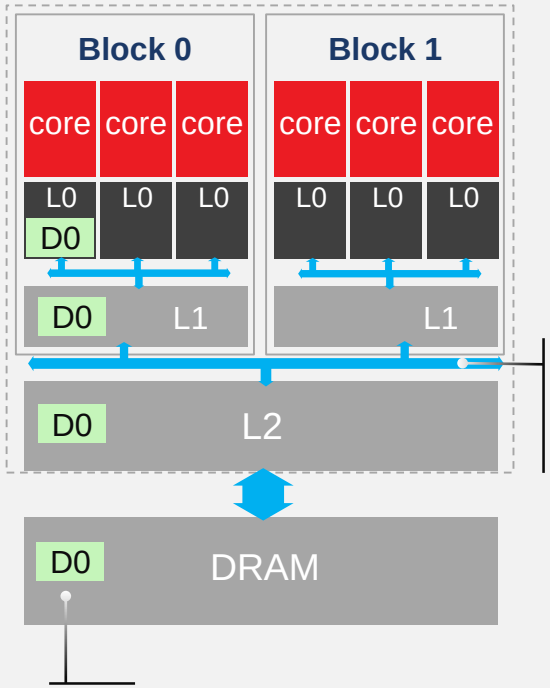


- **TB/s of Interface Bandwidth**
  - AI Engine to Programmable Logic
  - AI Engine to NOC
- **Leveraging NOC connectivity**
  - PS manages Config / Debug / Trace
  - AI Engine to DRAM (no PL req'd)



AI Engine: Multi-Core Compute with dedicated memory

Traditional Multi-core
(cache-based architecture)

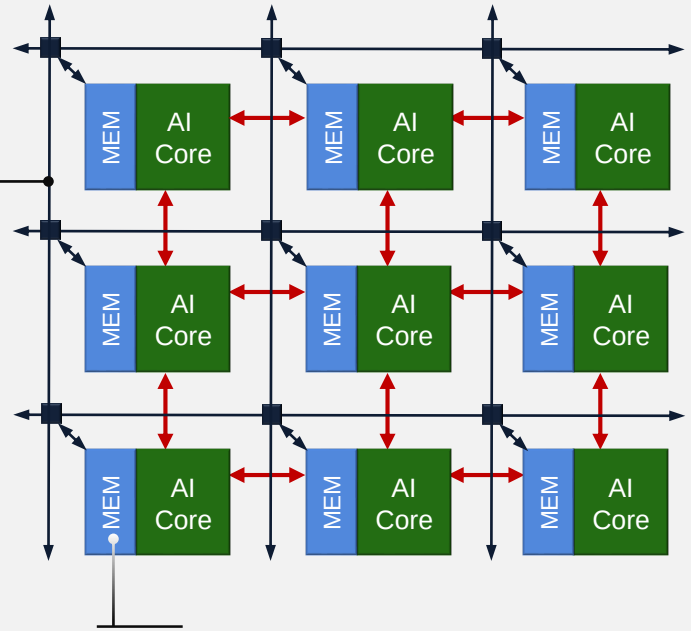


Fixed, shared Interconnect
- Blocking limits compute
- Timing not deterministic

Data Replicated
- Robs bandwidth
- Reduces capacity

AI Engine Array
(intelligent engine)

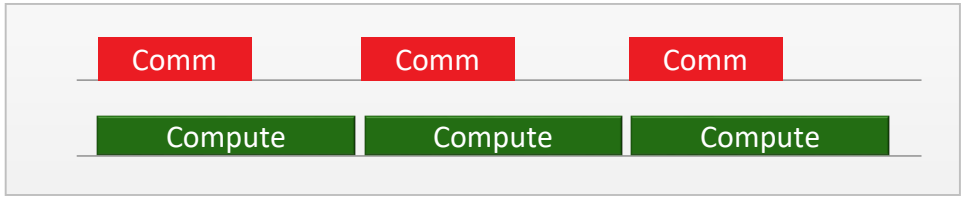
Dedicated Interconnect
- Non-blocking
- Deterministic



Local, Distributed Memory
- No cache misses
- Higher bandwidth
- Less capacity required

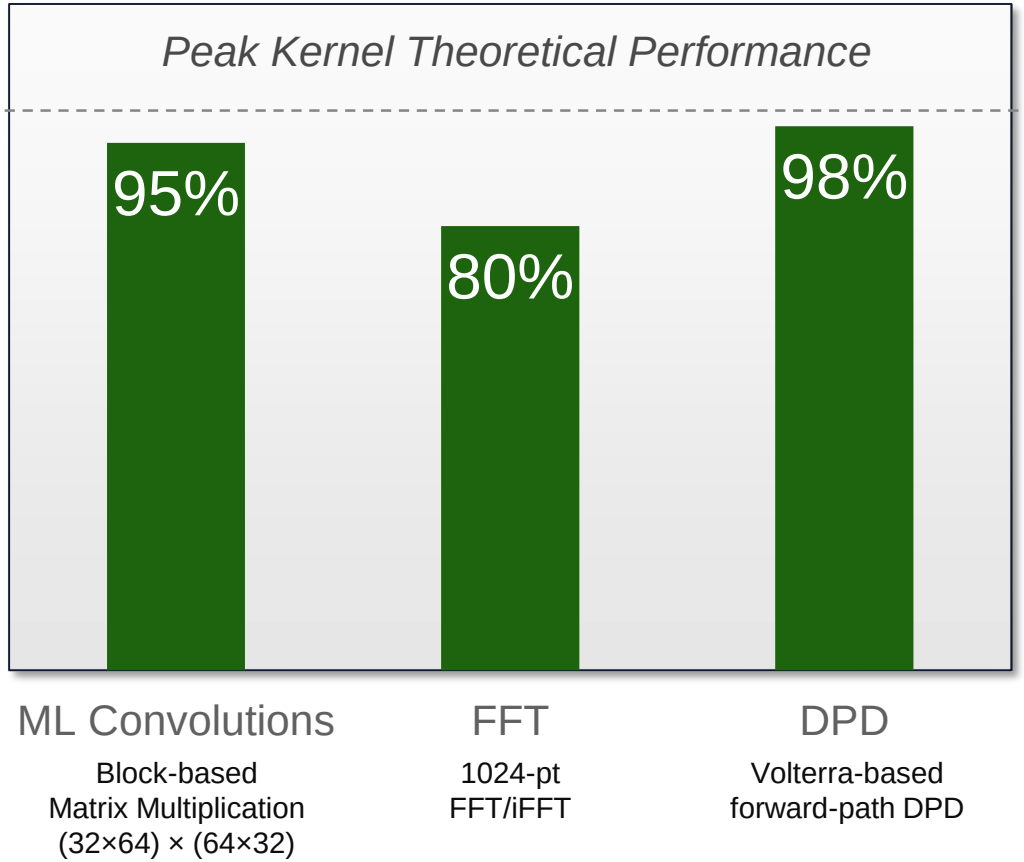
AI Engine Delivers High Compute Efficiency

- **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**



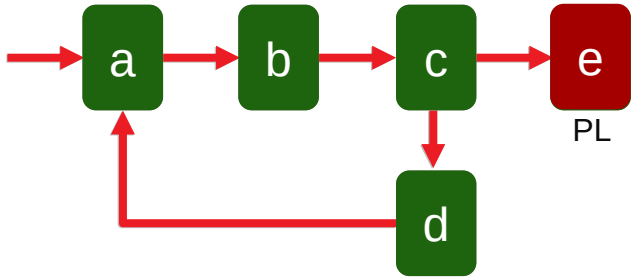
Overlap Compute and Communication

Vector Processor Efficiency



AI Engine Programming Experience: Dataflow Model

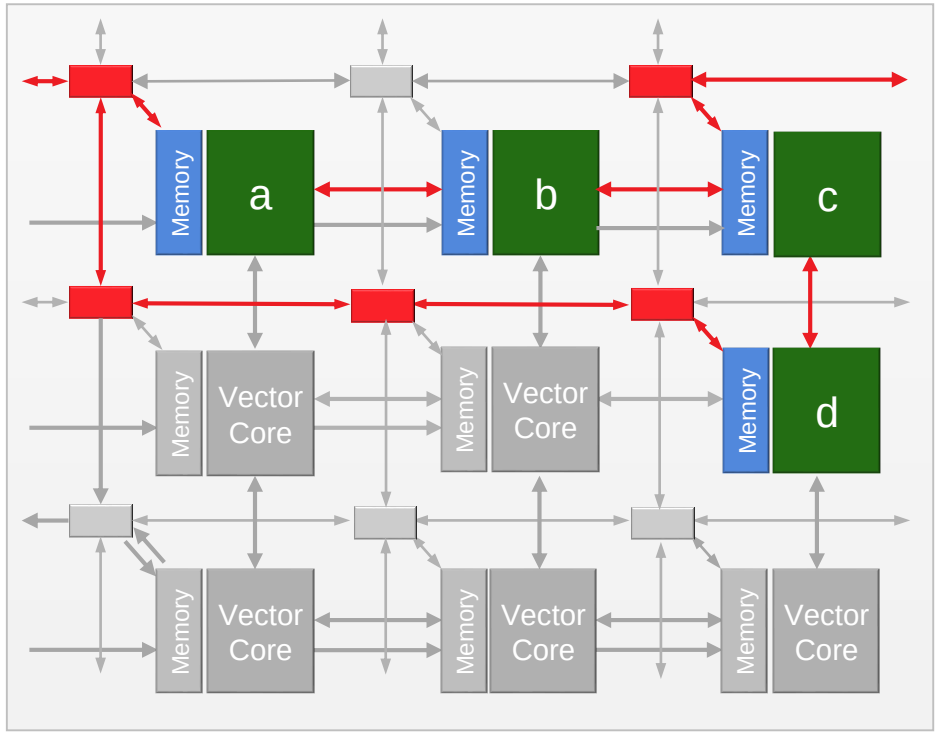
1 User defines dataflow logic



2 User describes dataflow graph using C/C++ APIs

3 Compiler transparently manages placement & interconnect

Physical Mapping to AI Engines



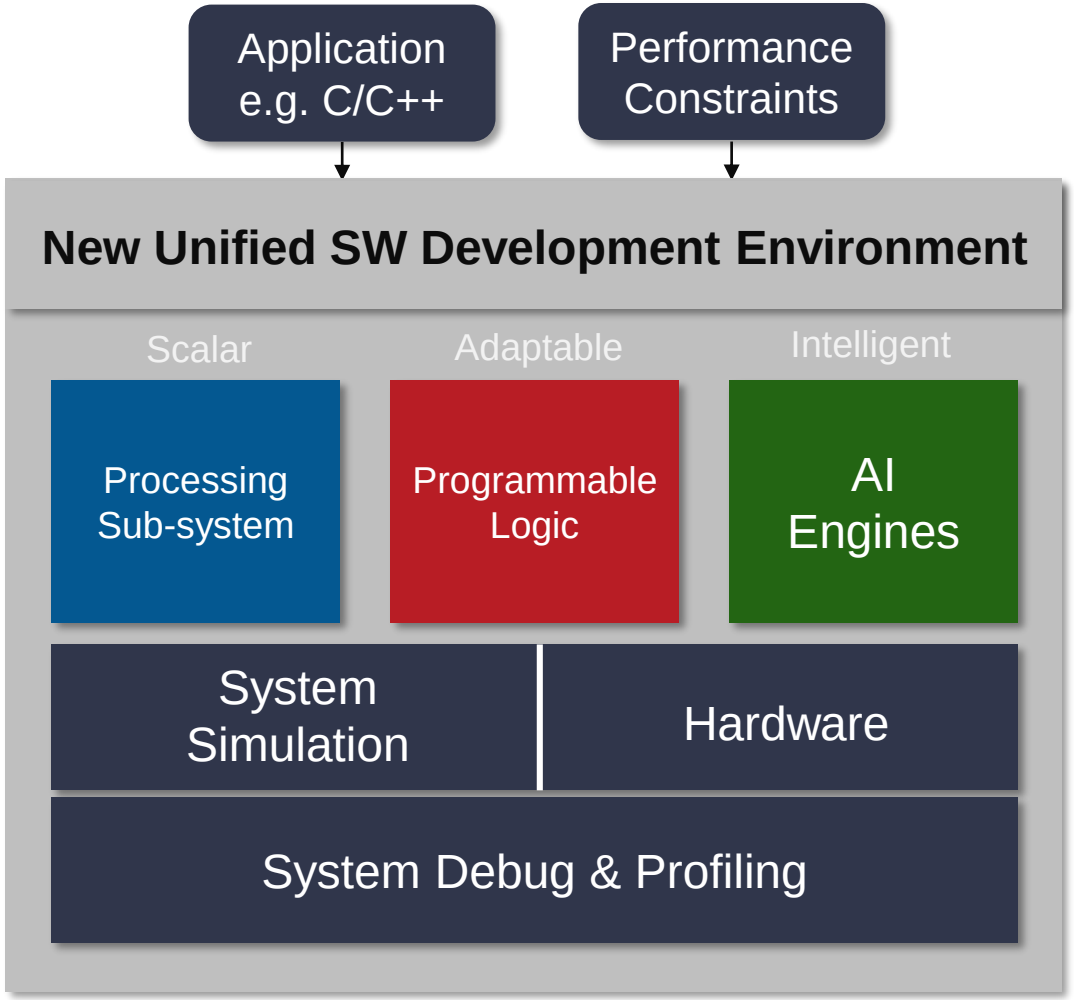
to e

XILINX

Versal ACAP Development Tools:

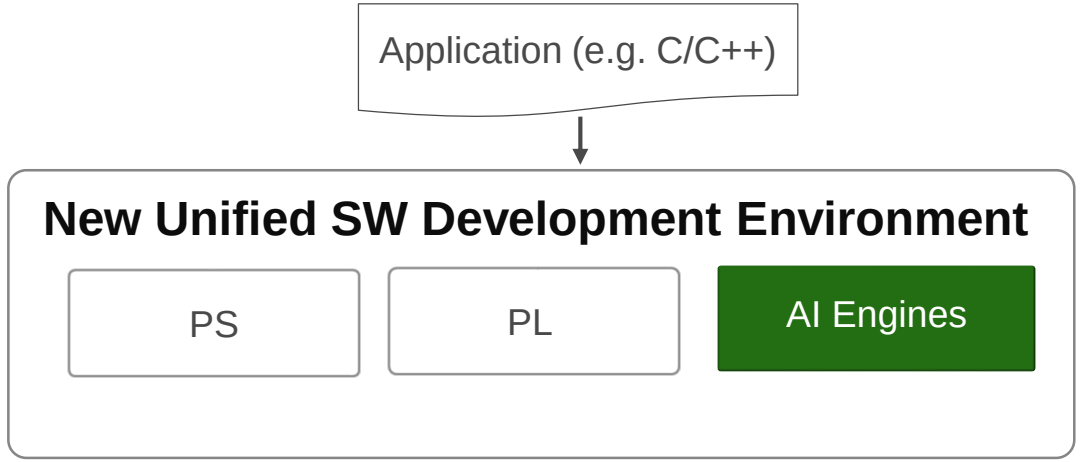
TOOLS	USER	SUPPORTED FRAMEWORKS
Frameworks	AI and Data Scientists	Caffe, TensorFlow, mxnet, FFMPEG
New Unified Software Development Environment	Software Application Developers	
Vivado Design Suite	Hardware Developers	
XILINX VERSAL		

Software Development Environment



- **Unified development environment**
  - Full chip programming
- **SW programmable for whole application**
  - Heterogeneous SW acceleration
- **Full system simulation, debug & profiling**
  - Software development experience

AI Engine Programming Environment



Full SW Programming Tool Chain
(Single-engine and Multi-engine)

- IDE
- Compiler
- Debugger
- Performance Analysis

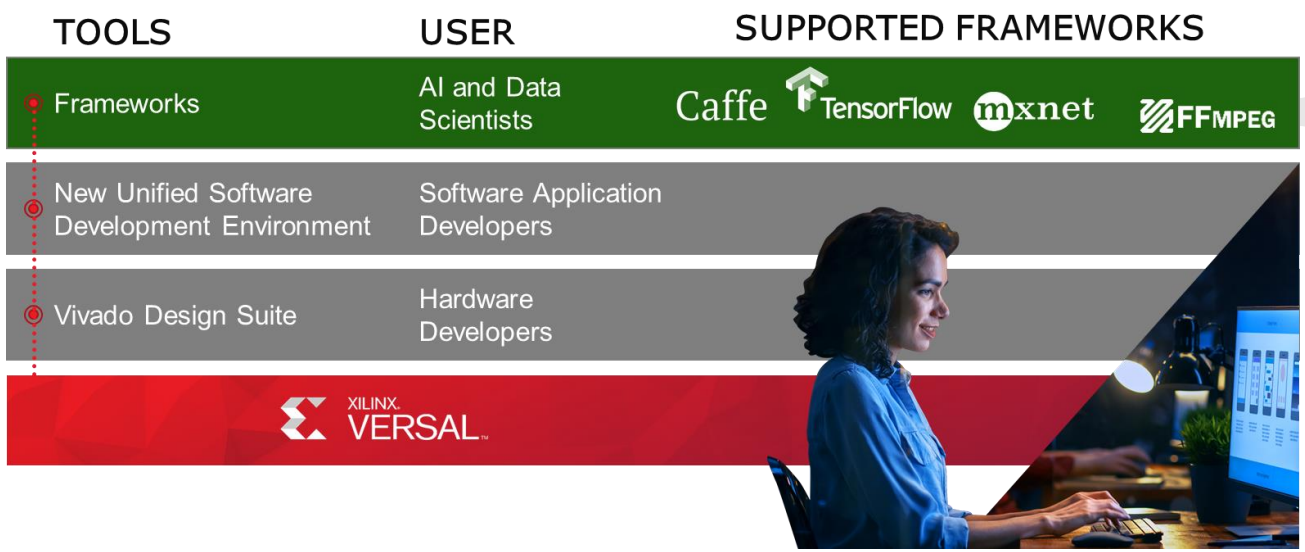
Performance-Optimized Software Libraries
(Examples)

- 4G/5G/Radar Library
- AI Library
- Vision Library

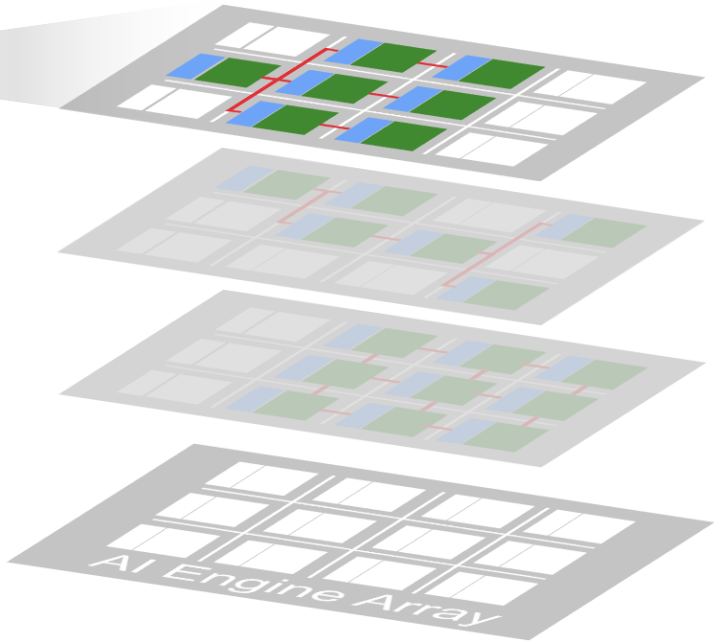
Run-Time Software
(Examples)

- Error Management
- Memory Management
- Boot + Configuration
- Power/Thermal Management

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

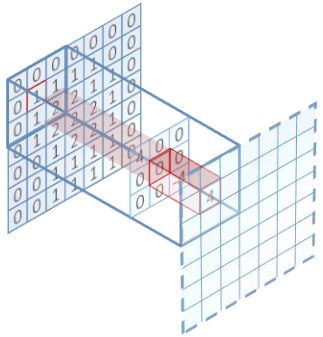
Accelerating AI Inference in the Data Center

1. User works in Framework of choice
   - Develop & train custom network
   - User provides trained model
2. Xilinx DNN Compiler implements network
   - Targets AI Inference Domain Specific Architecture
   - Quantize, merge layers, prune
   - Compile to AI Engines
3. Scalable across hardware targets
   - Start with Alveo today

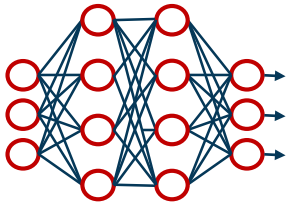
XILINX

AI Inference on Versal ACAP

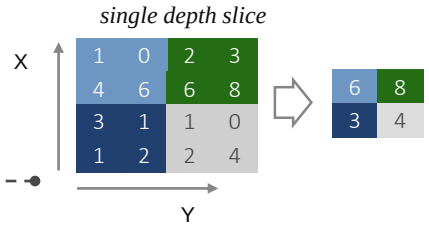
Convolutions



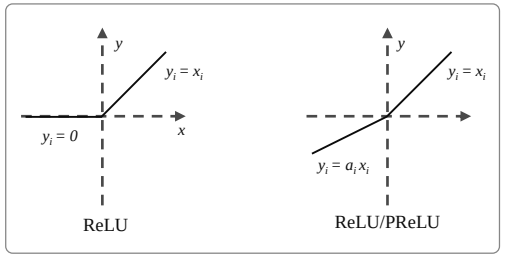
Fully Connected Layers



Pooling



Activations



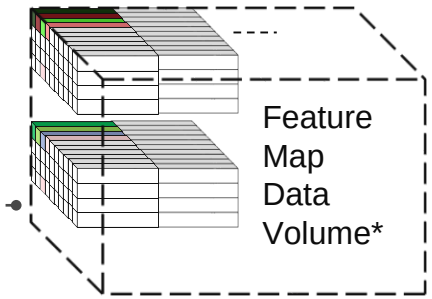
Processing System

AI Engines

Programmable Logic

I/O (GT, ADC/ DAC)

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



Custom Memory Hierarchy

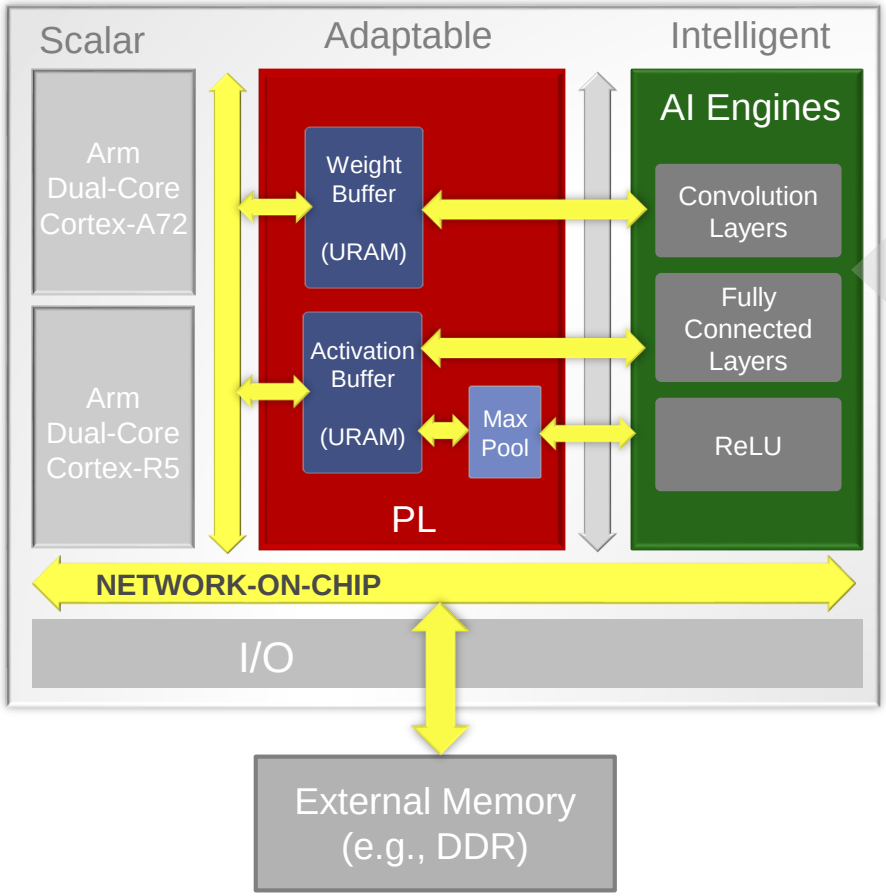
*Figure credit: https://en.wikipedia.org/wiki/Convolutional_neural_network

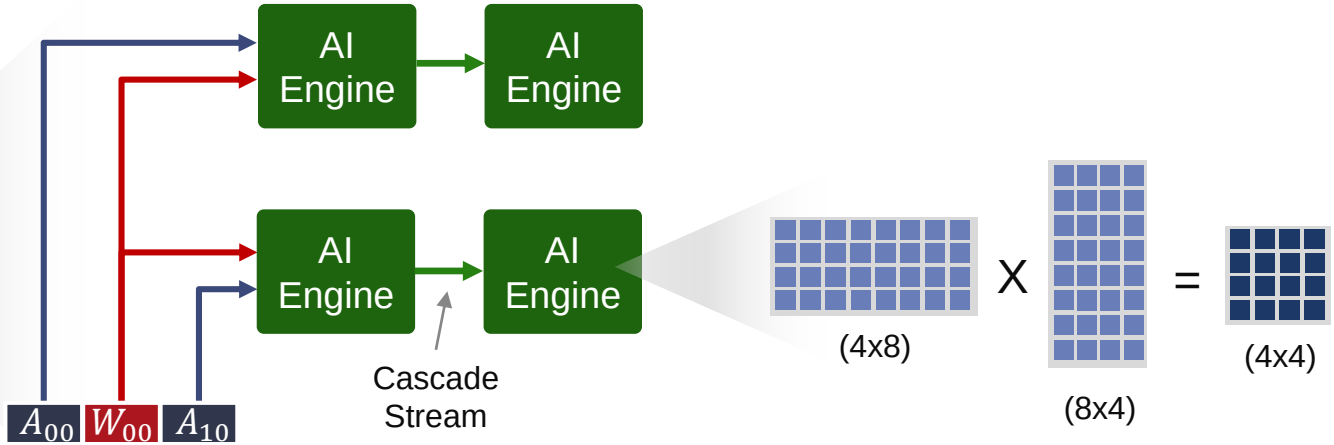
XILINX

AI Inference Mapping on Versal ACAP

A = Activations
W = Weights

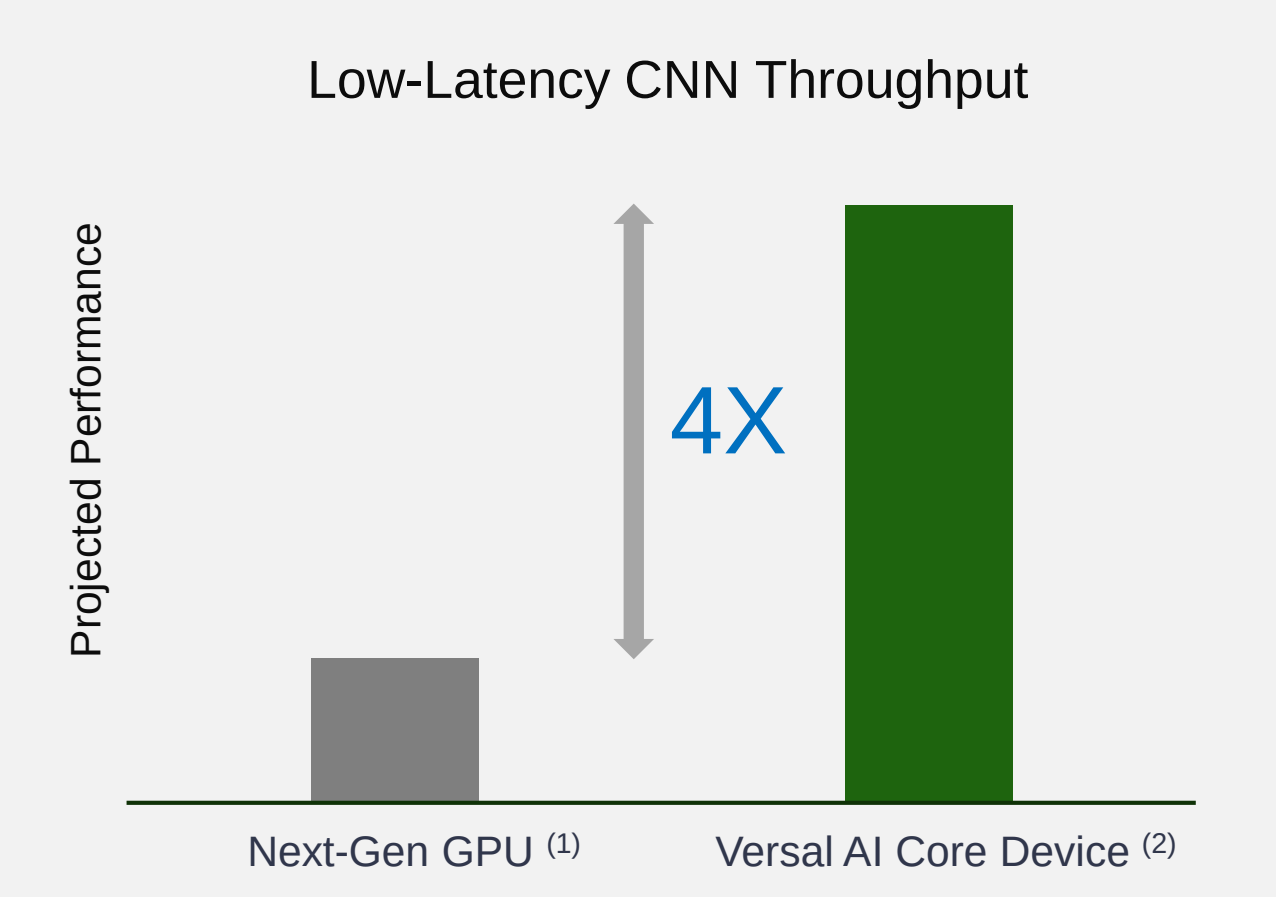
$$\begin{bmatrix} A_{00} & A_{01} \\ A_{10} & A_{11} \end{bmatrix} \times \begin{bmatrix} W_{00} & W_{01} \\ W_{10} & W_{11} \end{bmatrix} = \begin{bmatrix} A_{00} \times \boldsymbol{W_{00}} + A_{01} \times W_{10} & \dots \\ A_{10} \times \boldsymbol{W_{00}} + A_{11} \times W_{10} & \dots \end{bmatrix}$$





- 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)

AI Engine Delivers Real-time Inference Leadership
(75W Power Envelope)



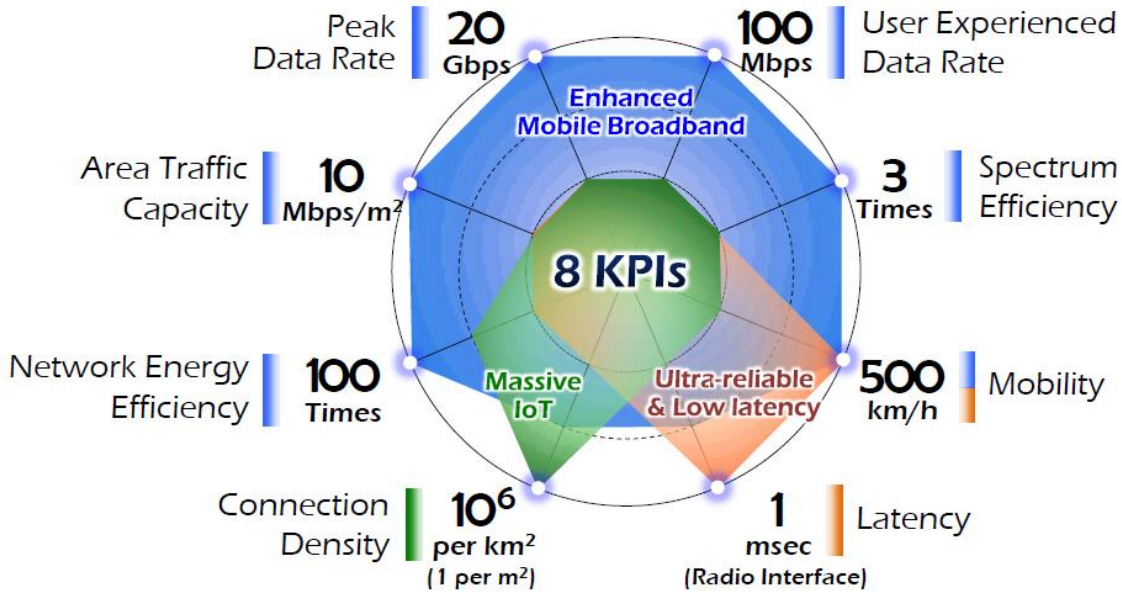
Note:
Versal device achieves 8X performance increase in 150W power envelope

(1) 12-nanometer T4 GPU device, Projected Batch=1 performance based on currently available vendor benchmarks
(2) 7-nanometer Versal AI Core Series VC1902 Device, 75W card power figures based on 2018.3 XPE power estimates, Latency <500us

Market Requirements and Trends: Wireless 5G

5G Complexity is 100X that of 4G

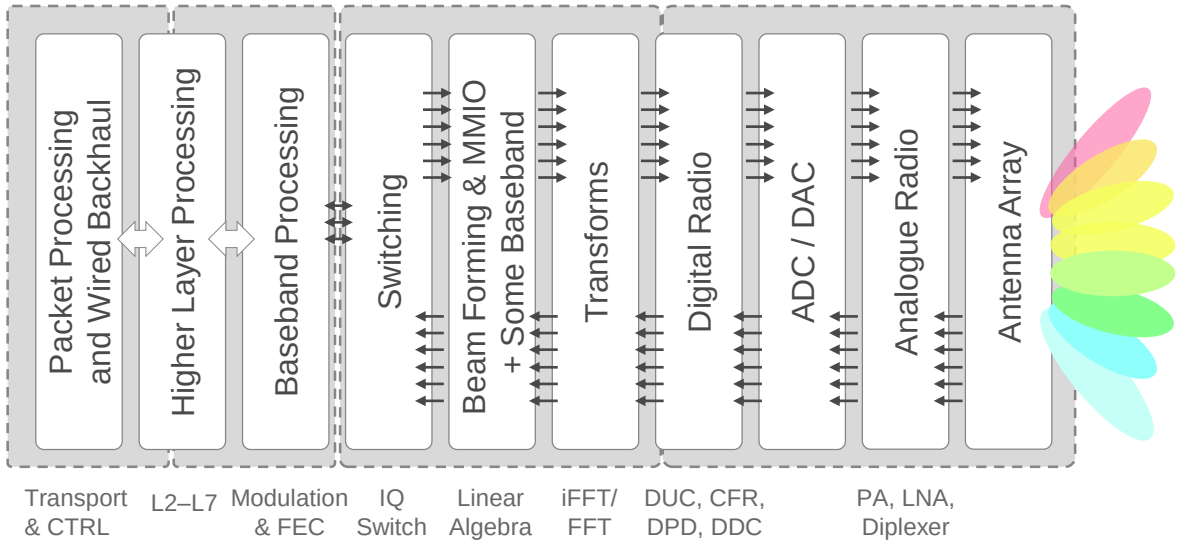
Still Evolving Standard



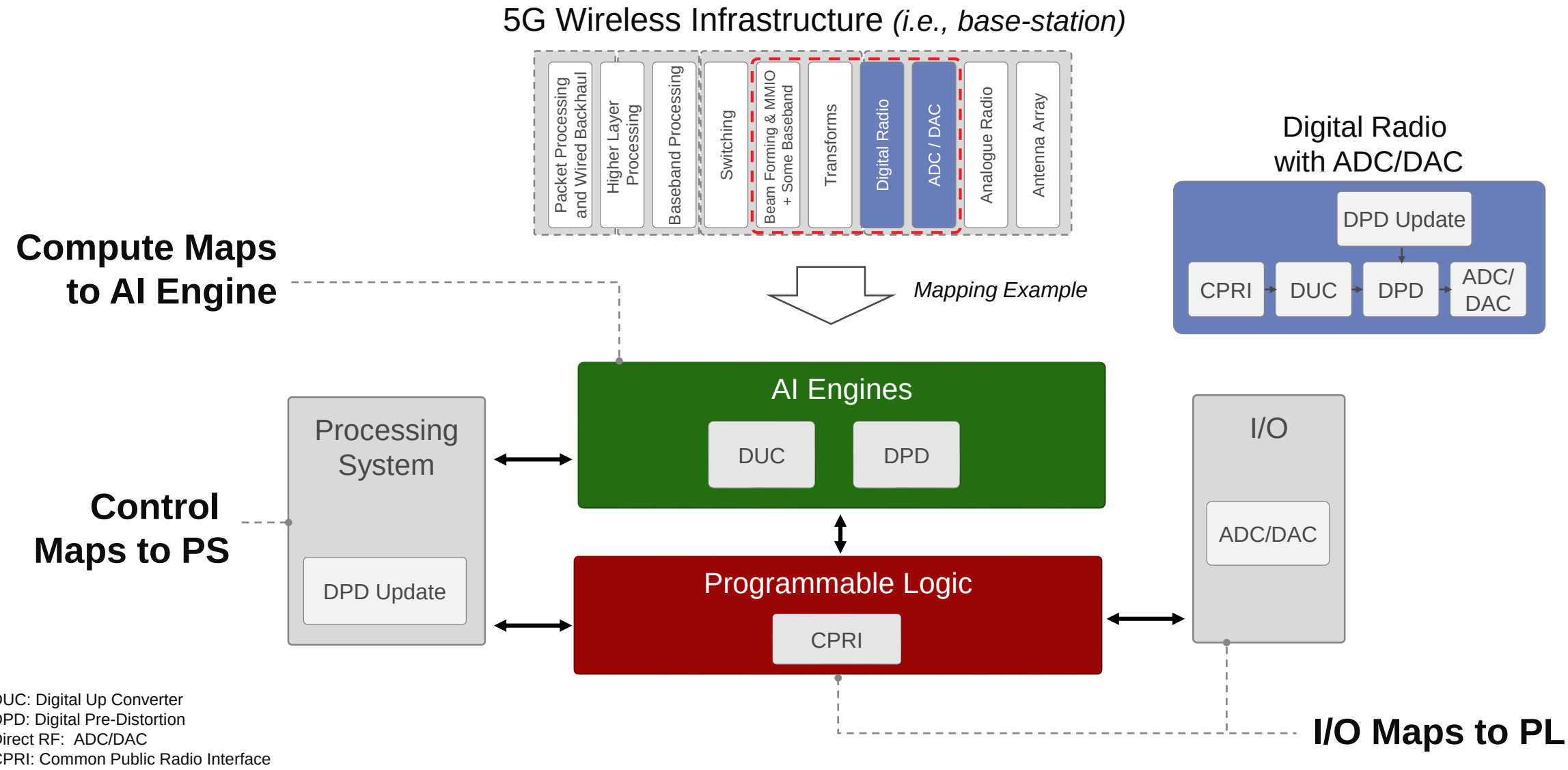
ETRI RWS-150029,
5G Vision and Enabling Technologies: ETRI Perspective 3GPP RAN Workshop
Phoenix, Dec. 2015
http://www.3gpp.org/ftp/tsg_ran/TSG_RAN/TSGR_70/Docs

New Technologies in 5G

- Massive MIMO
- Multiple antenna, frequency bands
- Changing functional partitioning



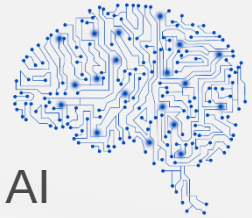
5G Wireless on Versal ACAP



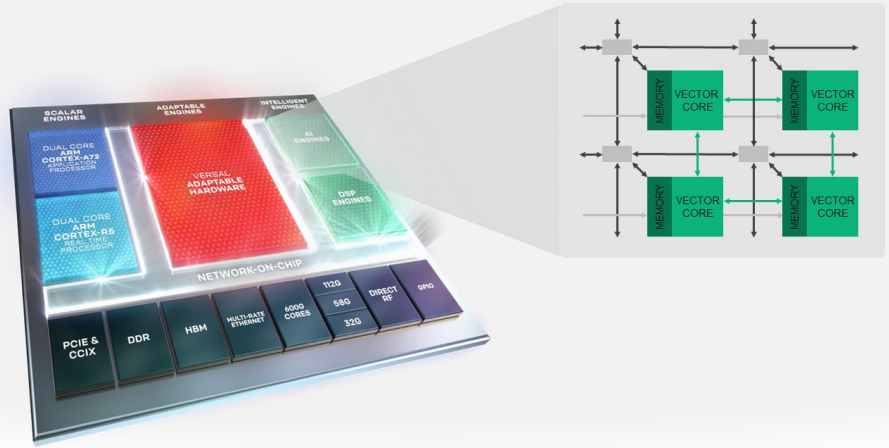
1: DUC: Digital Up Converter
2: DPD: Digital Pre-Distortion
3: Direct RF: ADC/DAC
4: CPRI: Common Public Radio Interface

AI Engine: Accelerating AI Inference & Signal Processing

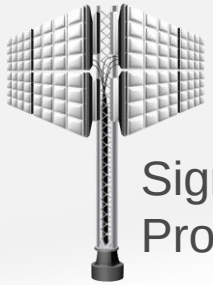
10x



AI Inference



5x



Signal Processing

Software Programmable

- Frameworks & C/C++
- SW Compile, Debug & Deploy

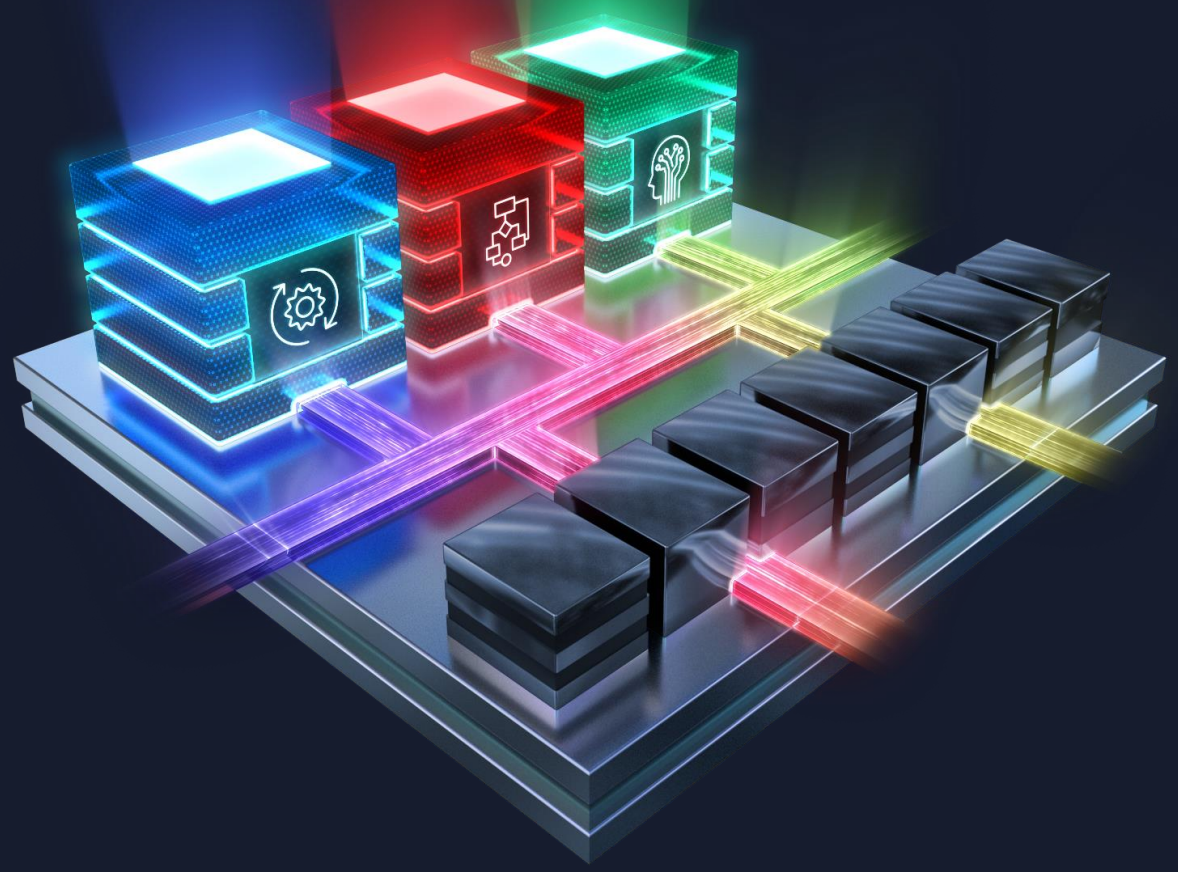
Deterministic

- Max throughput w/ low latency
- Real-time inference leadership

Efficient

- Up to 8X compute density
- At ~40% lower power

XILINX



www.xilinx.com/versal

VC1902:133TOPS (int8 peak)