



Versal: AI Engine & Programming Environment

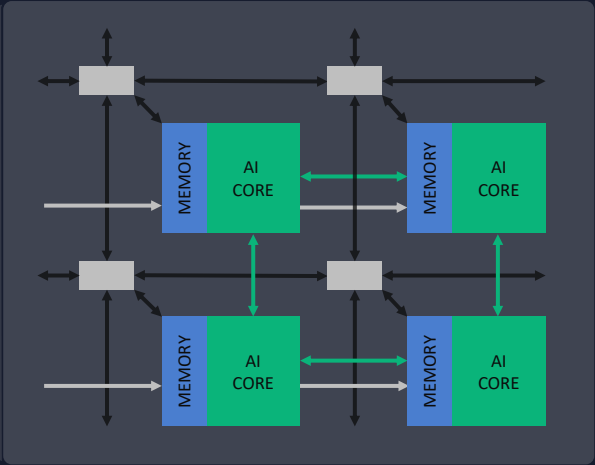
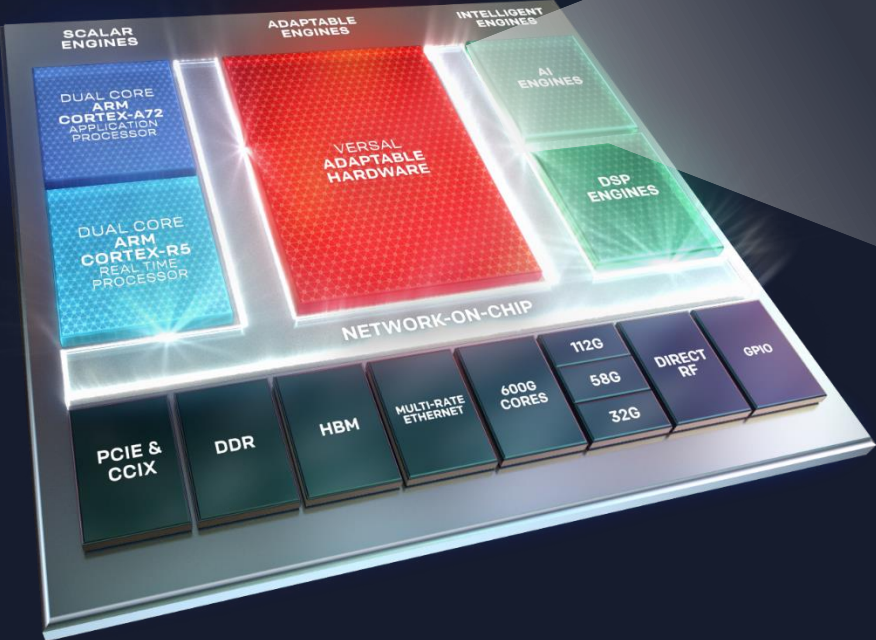
Presented By

Ambrose Finnerty
Xilinx DSP Technical Marketing Manager

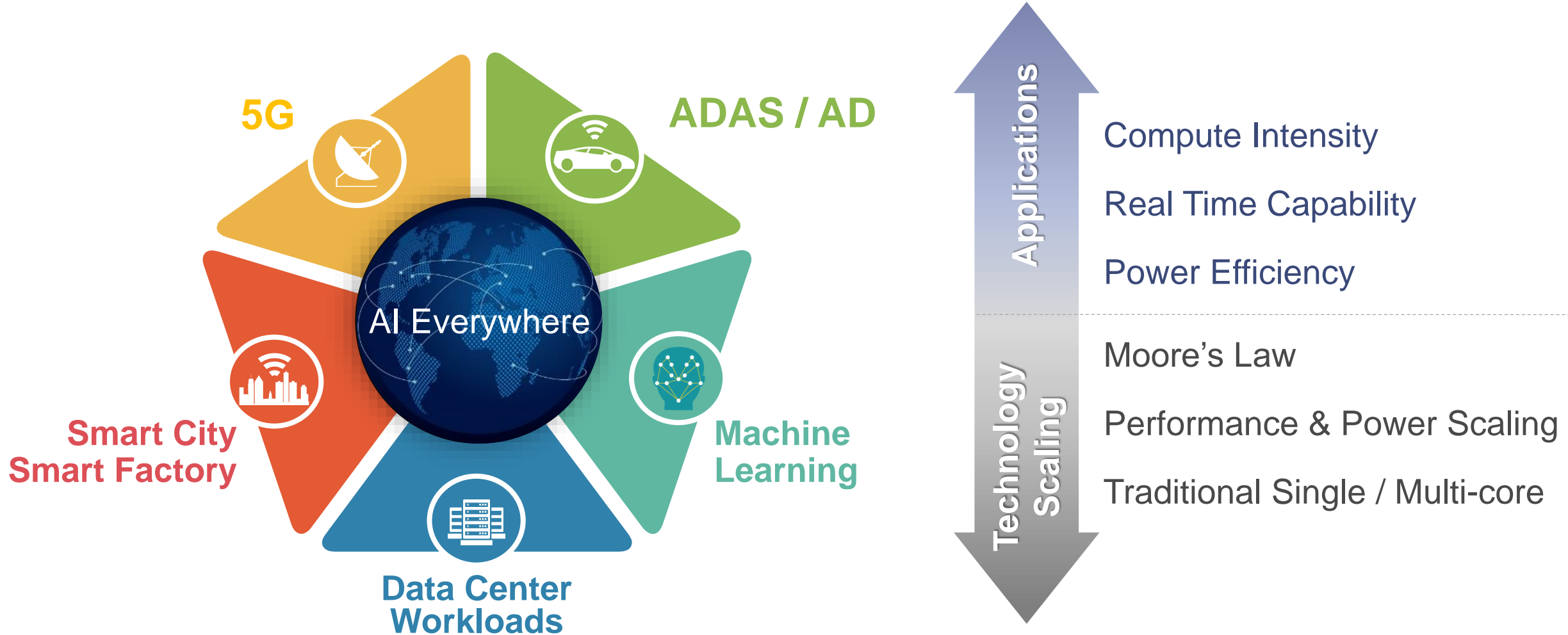
October 16, 2018



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	—	—	✓	✓

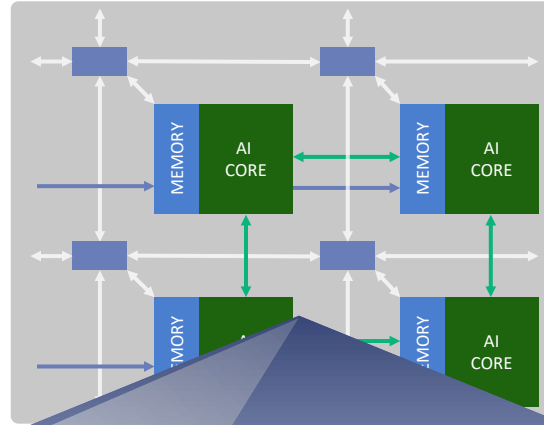
← ACAP
w/ AI Engine

Introducing the AI Engine

SW Programmable ➤

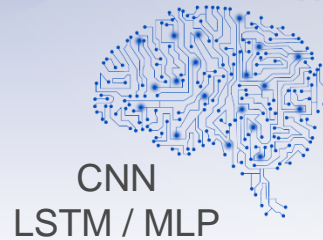
Deterministic ➤

Efficient ➤

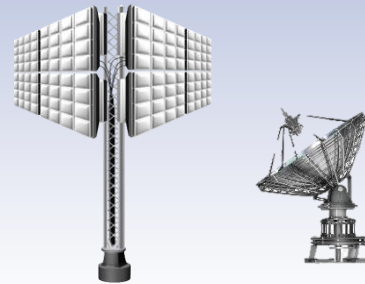


- 1GHz+ Multi-precision Vector Processor
- High bandwidth extensible memory
- Up to 400 AI Engines per device
- 8X Compute Density
- 40% Lower Power

Artificial Intelligence



Signal Processing



Computer Vision



Adaptable. **Intelligent.**

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

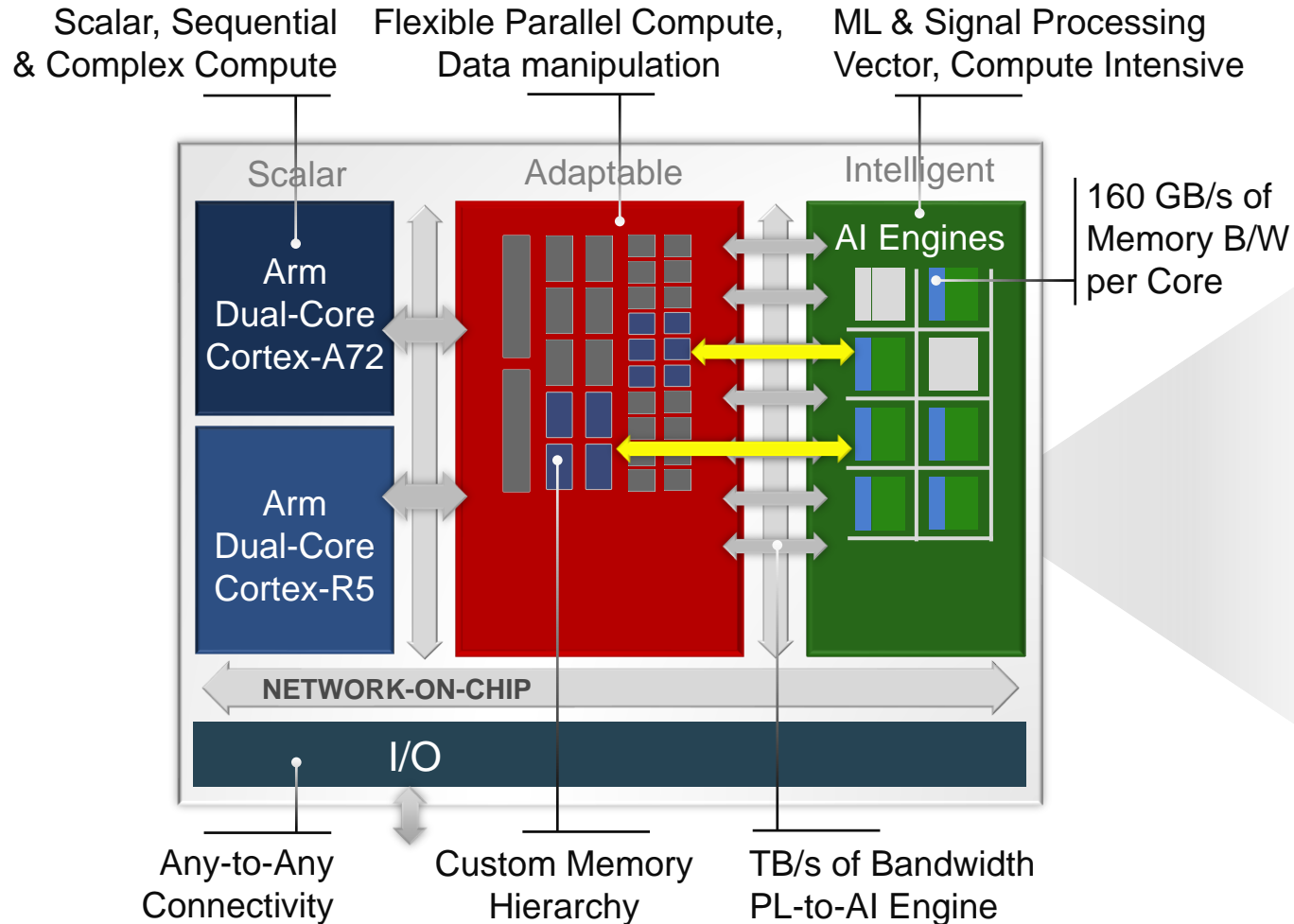
Kernel Program
Data Flow w/ user
defined libraries

PS

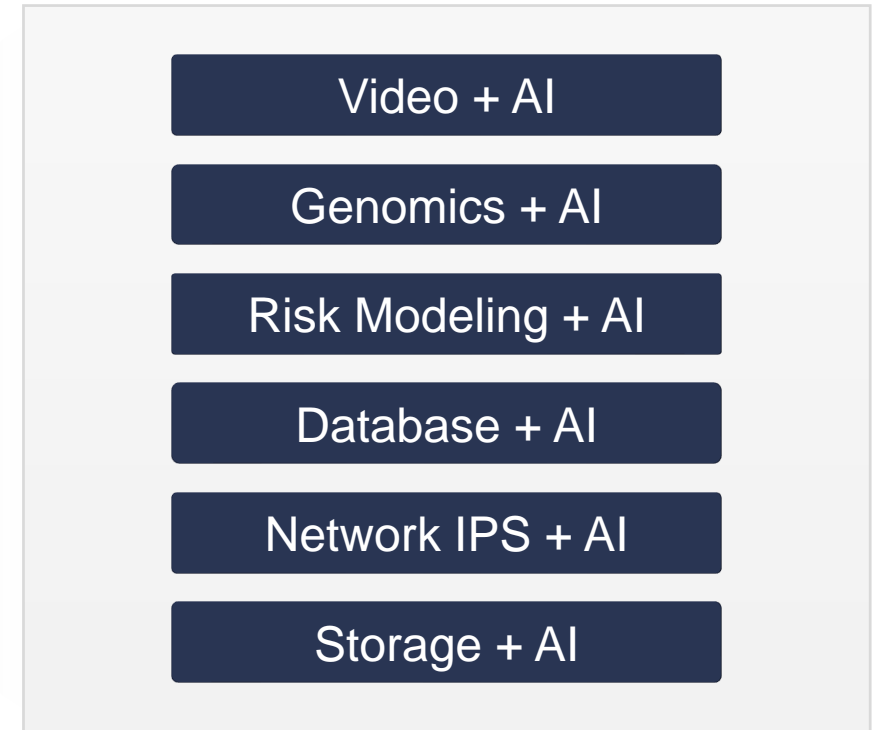
AI Engine Array
Programmable
Logic

I/O

Hardware Adaptable: Accelerating the Whole Application

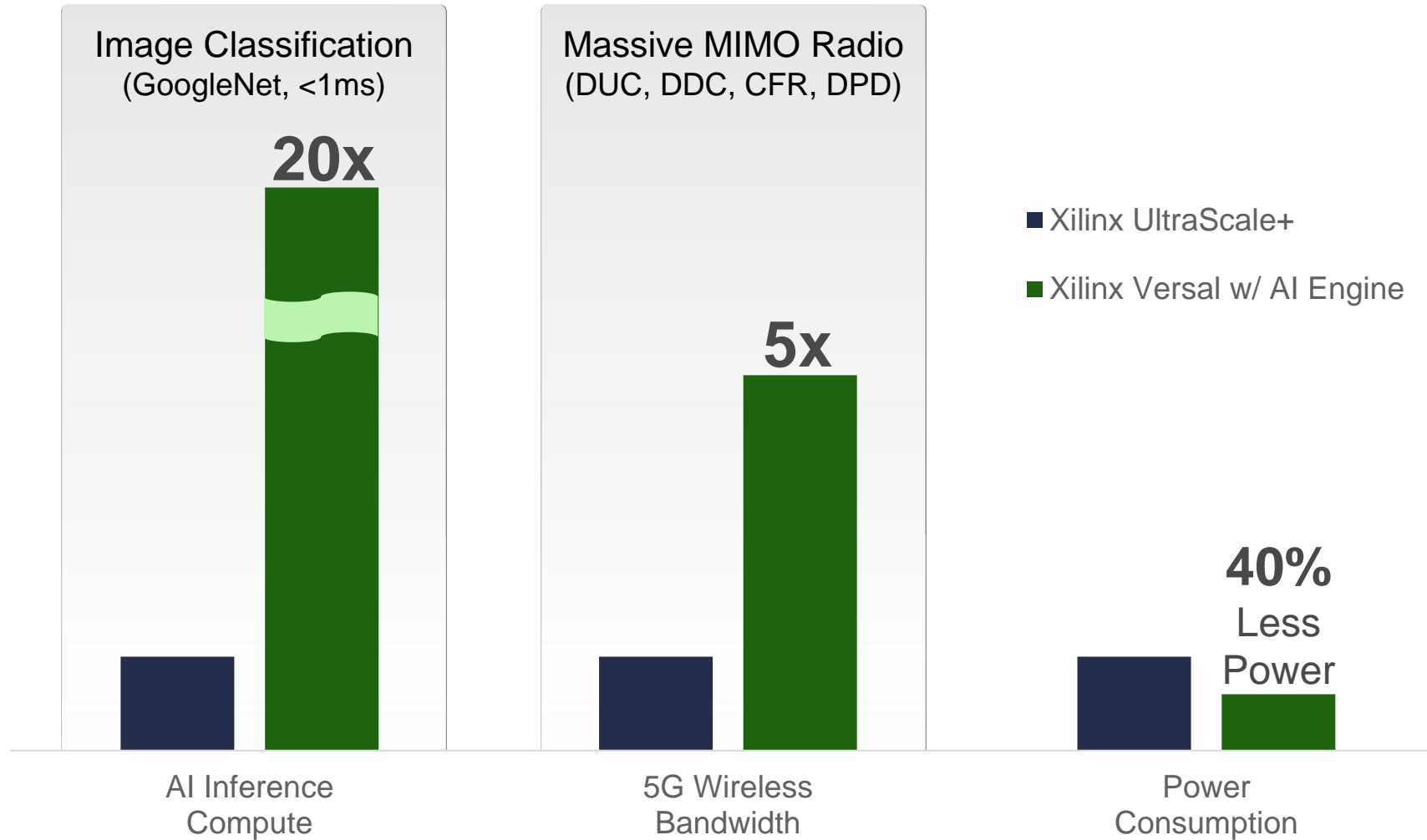


Heterogeneous Acceleration from Data Center to the Edge



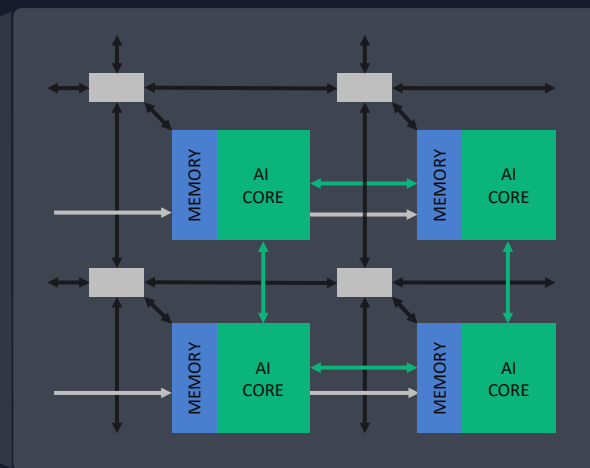
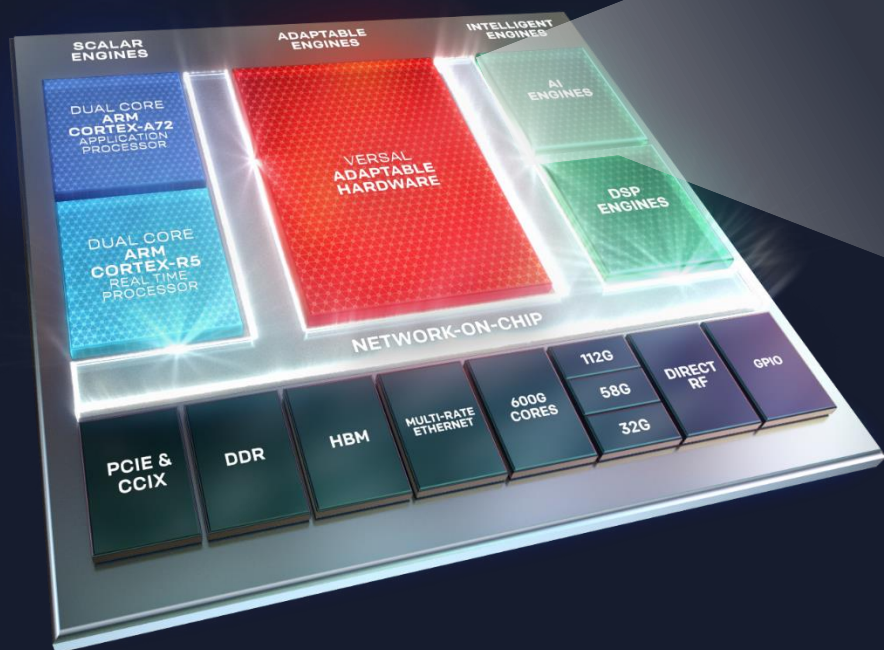
Delivering Deterministic Performance & Low Latency

AI Engine Application Performance & Power Efficiency

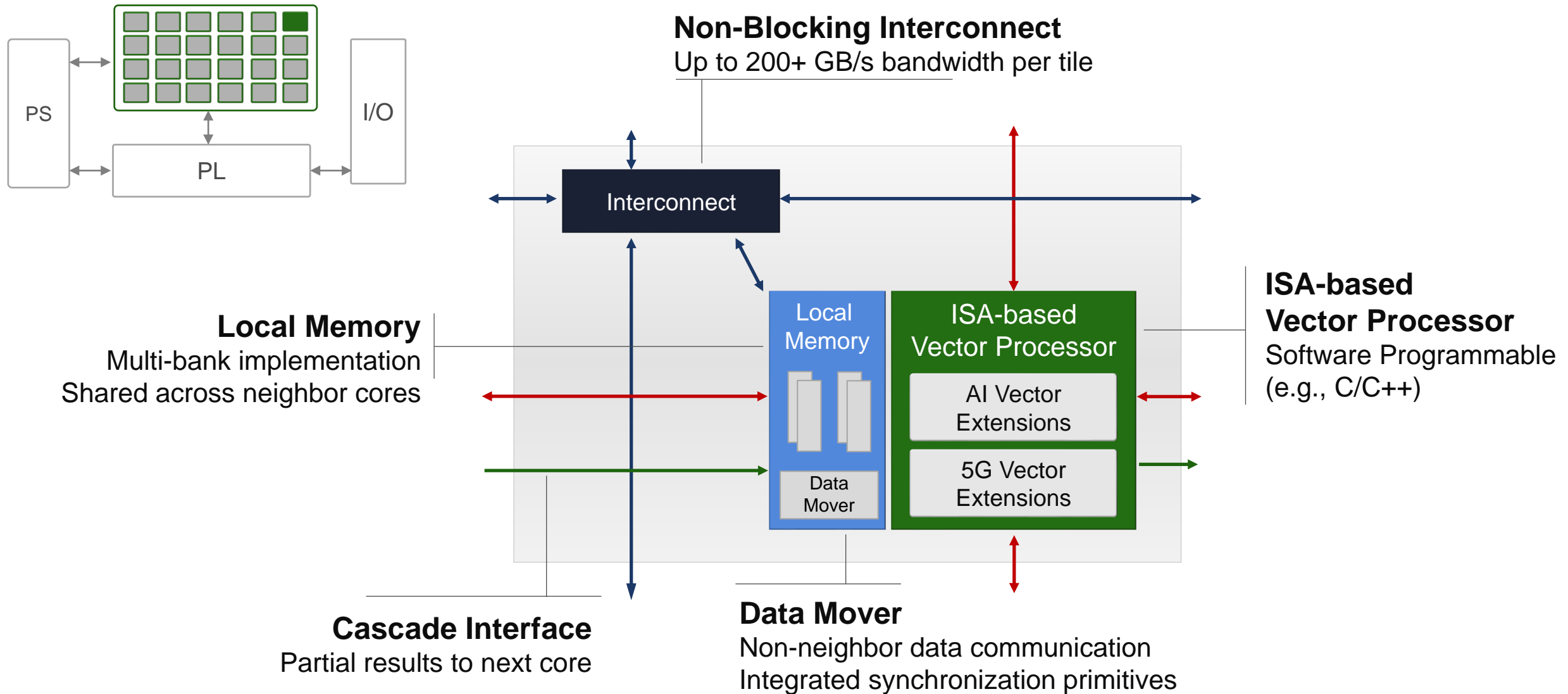


AI Engine

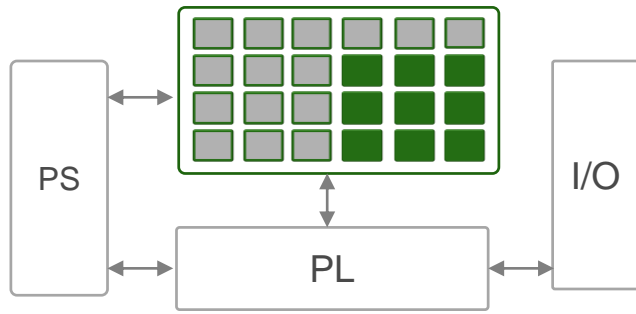
Architecture, Programming & Applications



AI Engine: Tile-Based Architecture



AI Engine: Array Architecture



Modular and scalable architecture

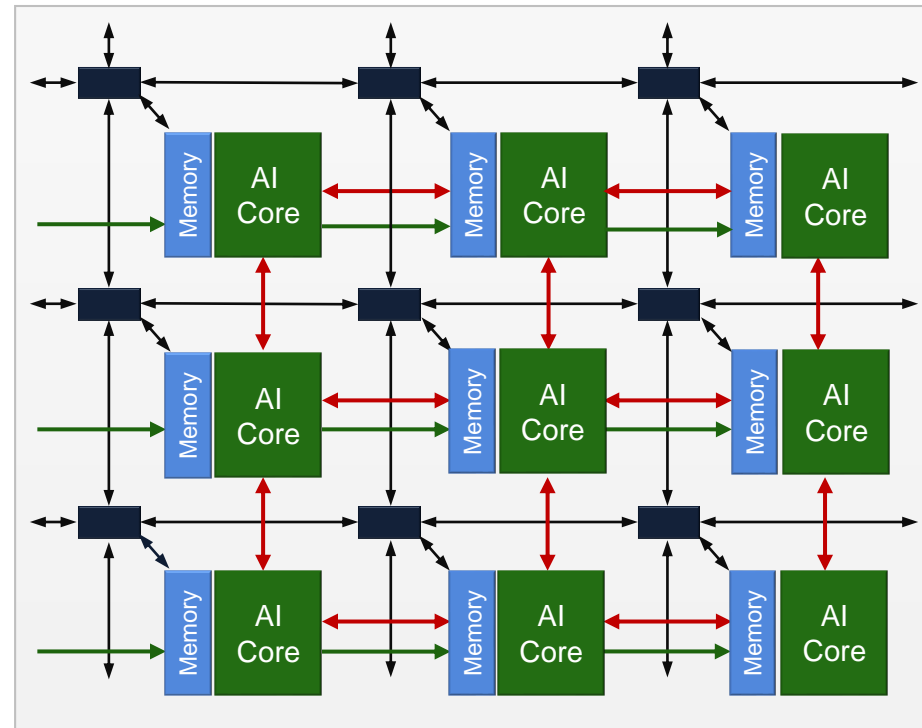
- More tiles = more compute
- Up to 400 per device
 - Versal AI Core VC1902 device

Distributed memory hierarchy

Maximize memory bandwidth

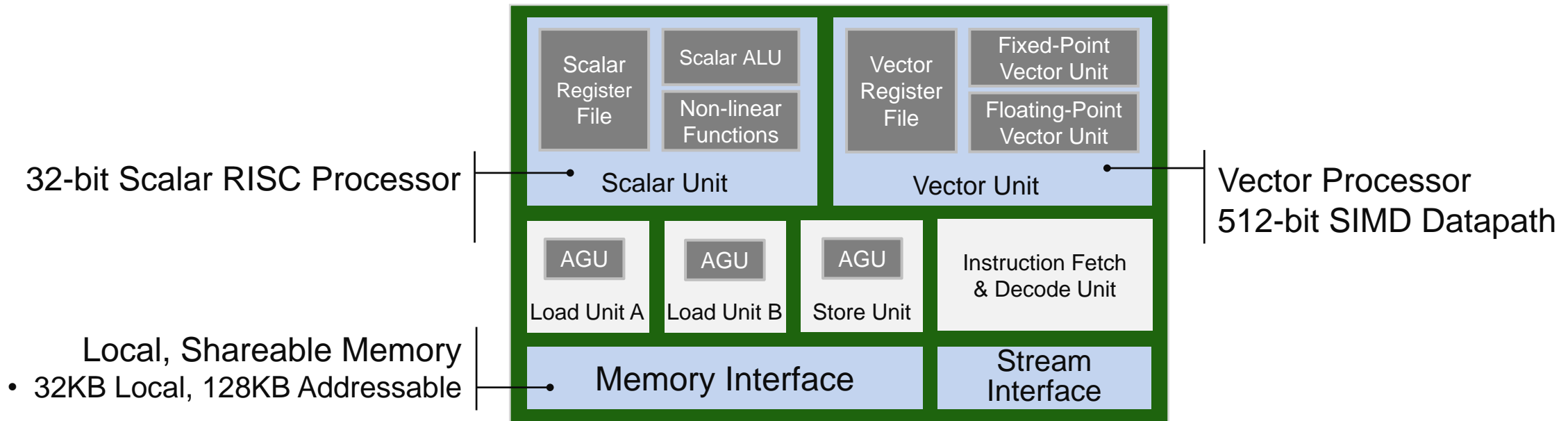
Array of AI Engines

- Increase in compute, memory and communication bandwidth



Deterministic Performance & Low Latency

AI Engine: Processor Core



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

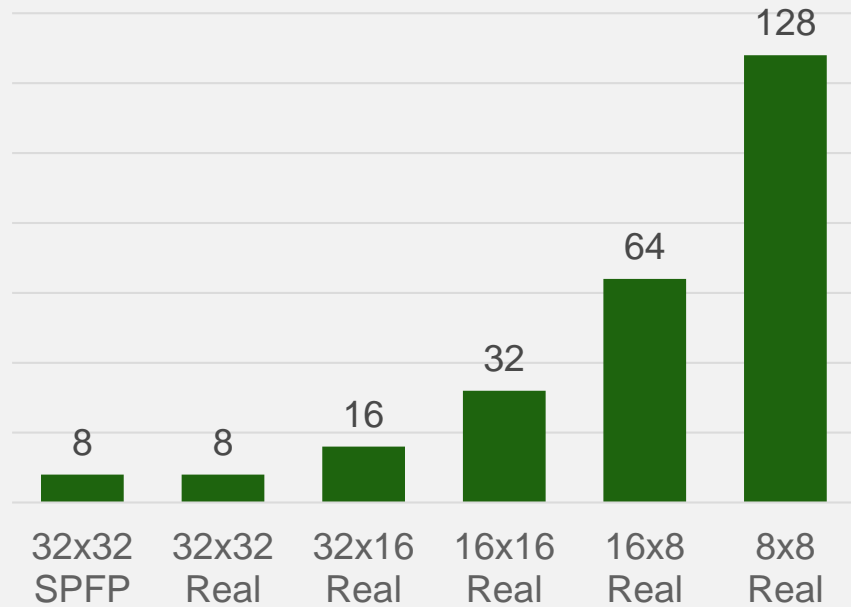
- Vector Datapath
- 8 / 16 / 32-bit & SPFP operands

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

Multi-Precision Support

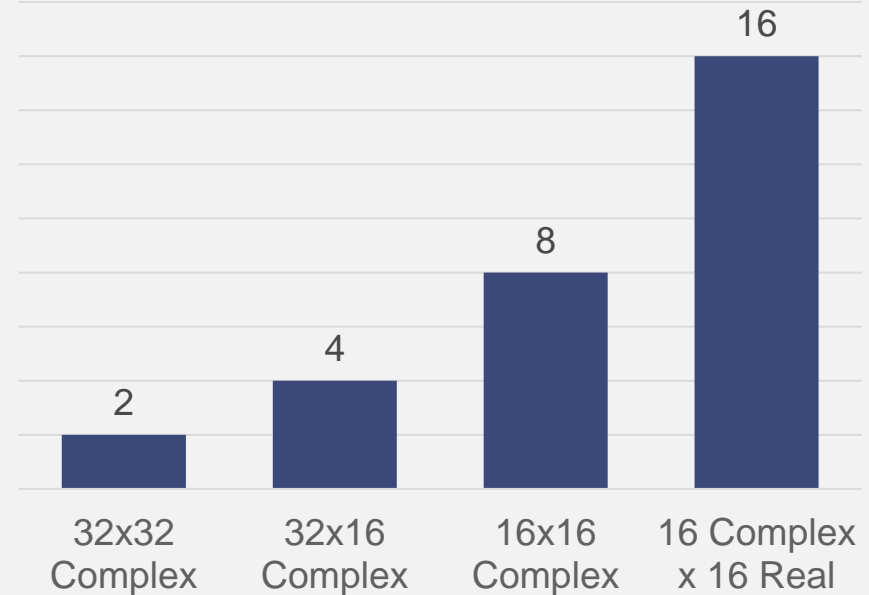
AI Data Types

MACs / Cycle (per core)



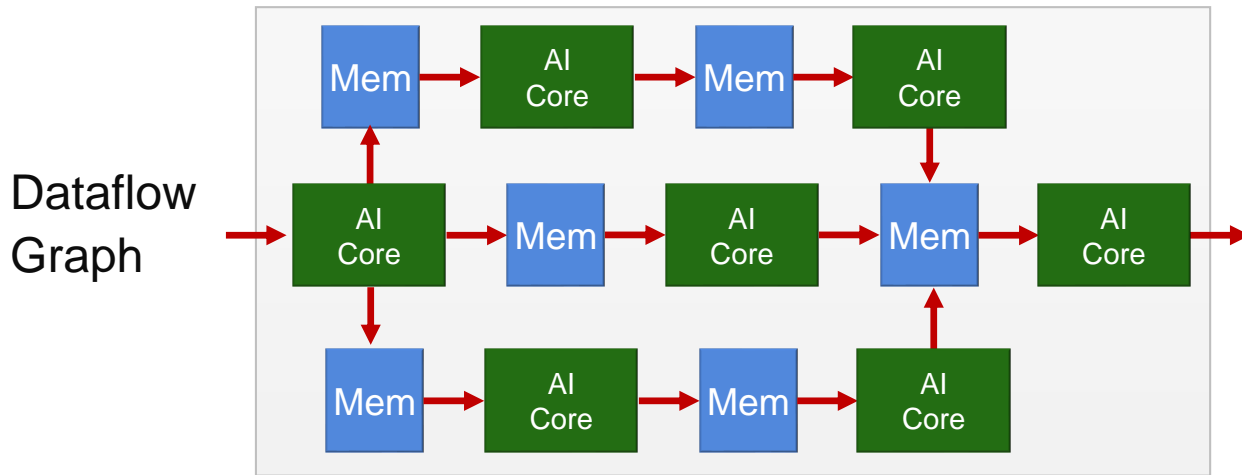
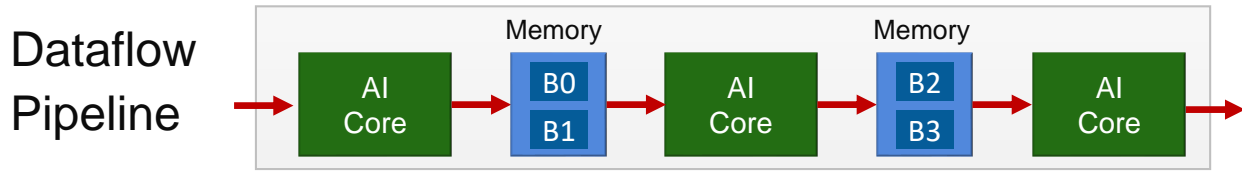
Signal Processing Data Types

MACs / Cycle (per core)



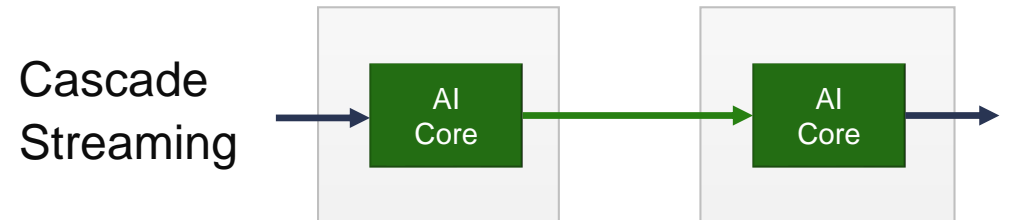
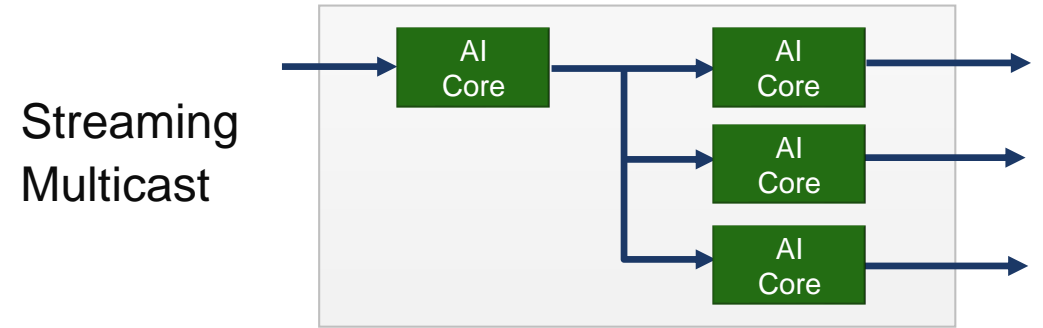
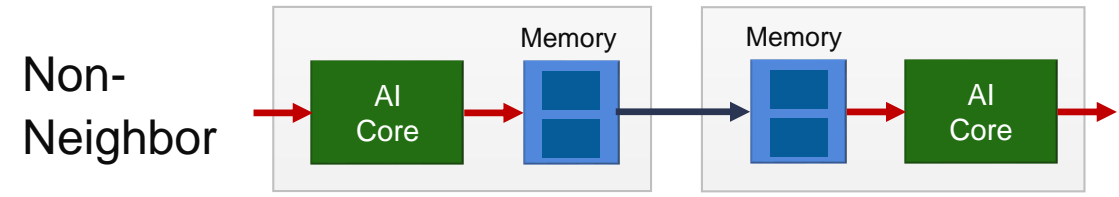
Data Movement Architecture

Memory Communication

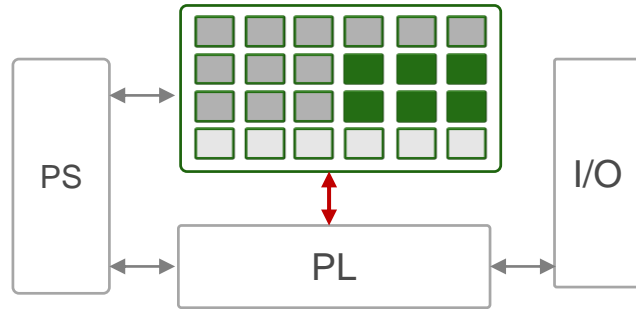


- Memory Interface
- Stream Interface
- Cascade Interface

Streaming Communication



AI Engine Integration with Versal ACAP

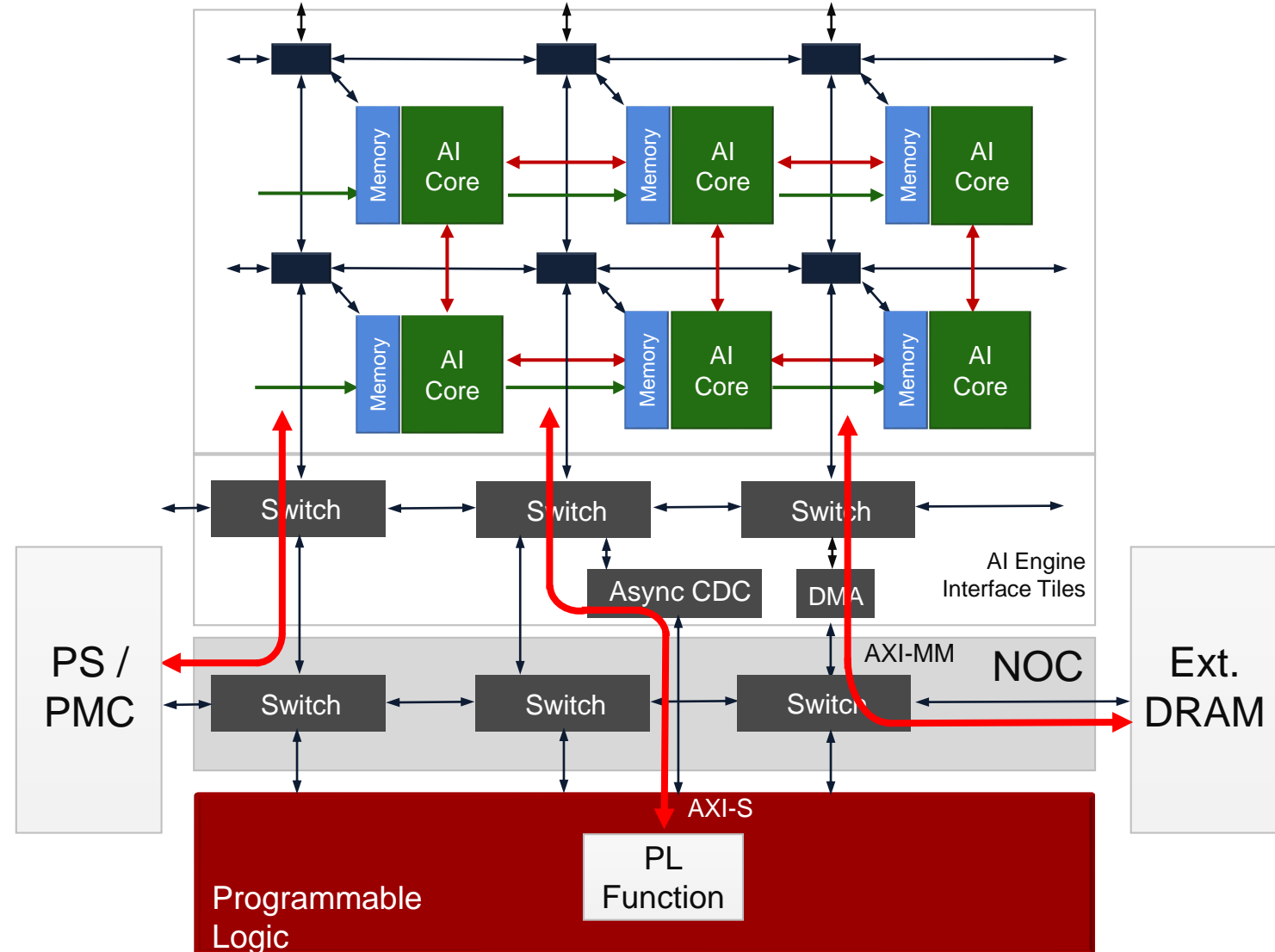


> TB/s of Interface Bandwidth

- >> AI Engine to Programmable Logic
- >> AI Engine to NOC

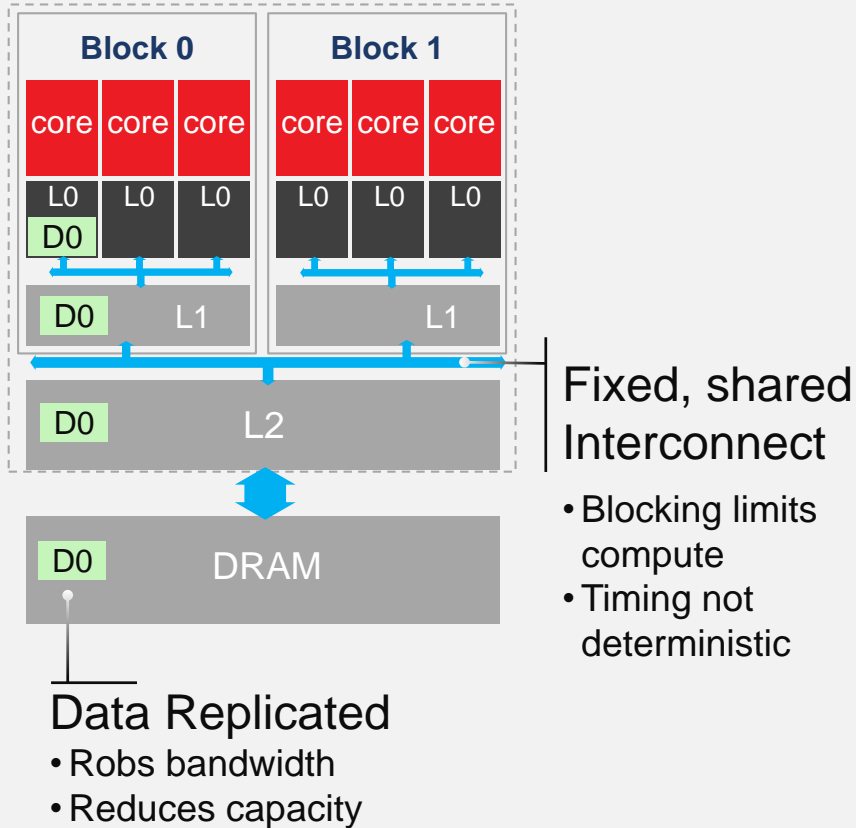
> Leveraging NOC connectivity

- >> PS manages Config / Debug / Trace
- >> AI Engine to DRAM (no PL req'd)



AI Engine: Xilinx Reinvents Multi-Core Compute

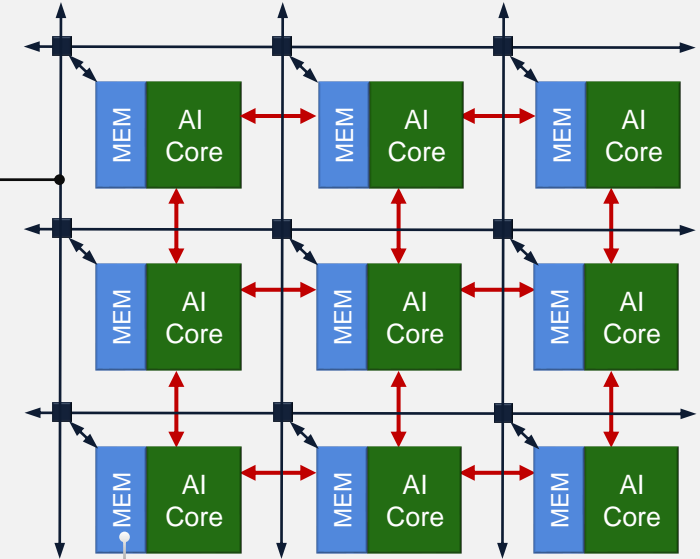
Traditional Multi-core (cache-based architecture)



AI Engine Array (intelligent engine)

Dedicated Interconnect

- Non-blocking
- Deterministic

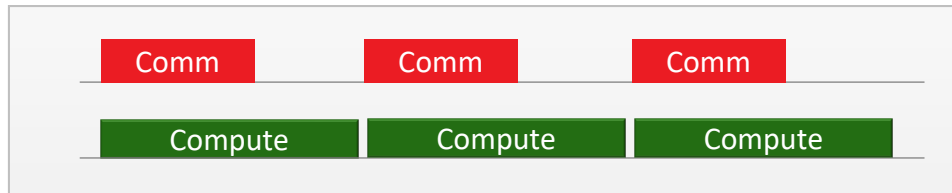


Local, Distributed Memory

- No cache misses
- Higher bandwidth
- Less capacity required

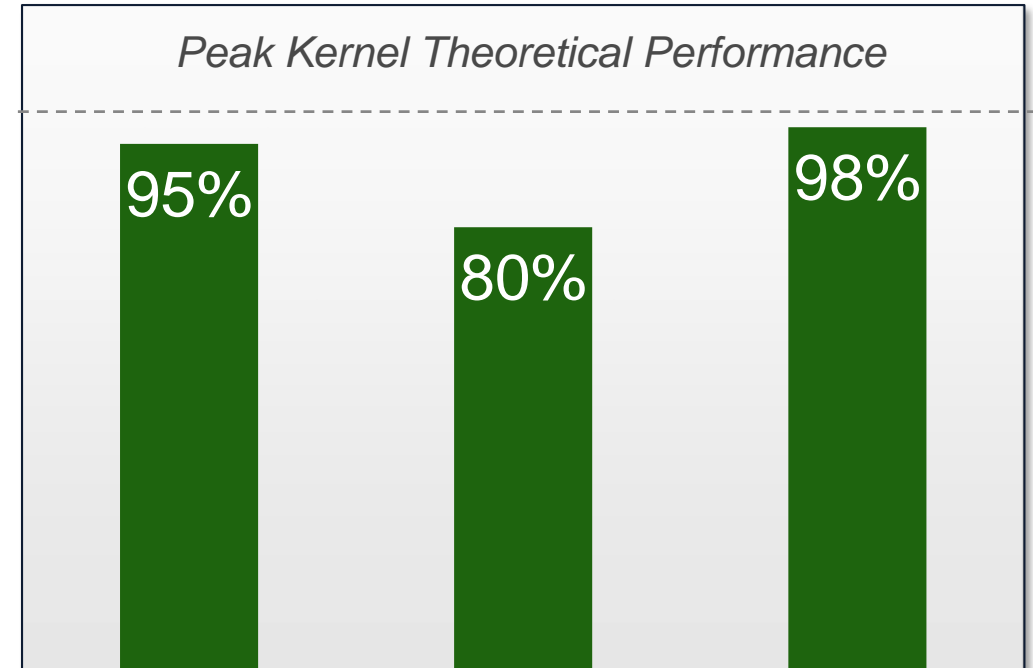
AI Engine Delivers High Compute Efficiency

- > **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



ML Convolutions

Block-based
Matrix Multiplication
(32x64) x (64x32)

FFT

1024-pt
FFT/iFFT



DPD

Volterra-based
forward-path DPD

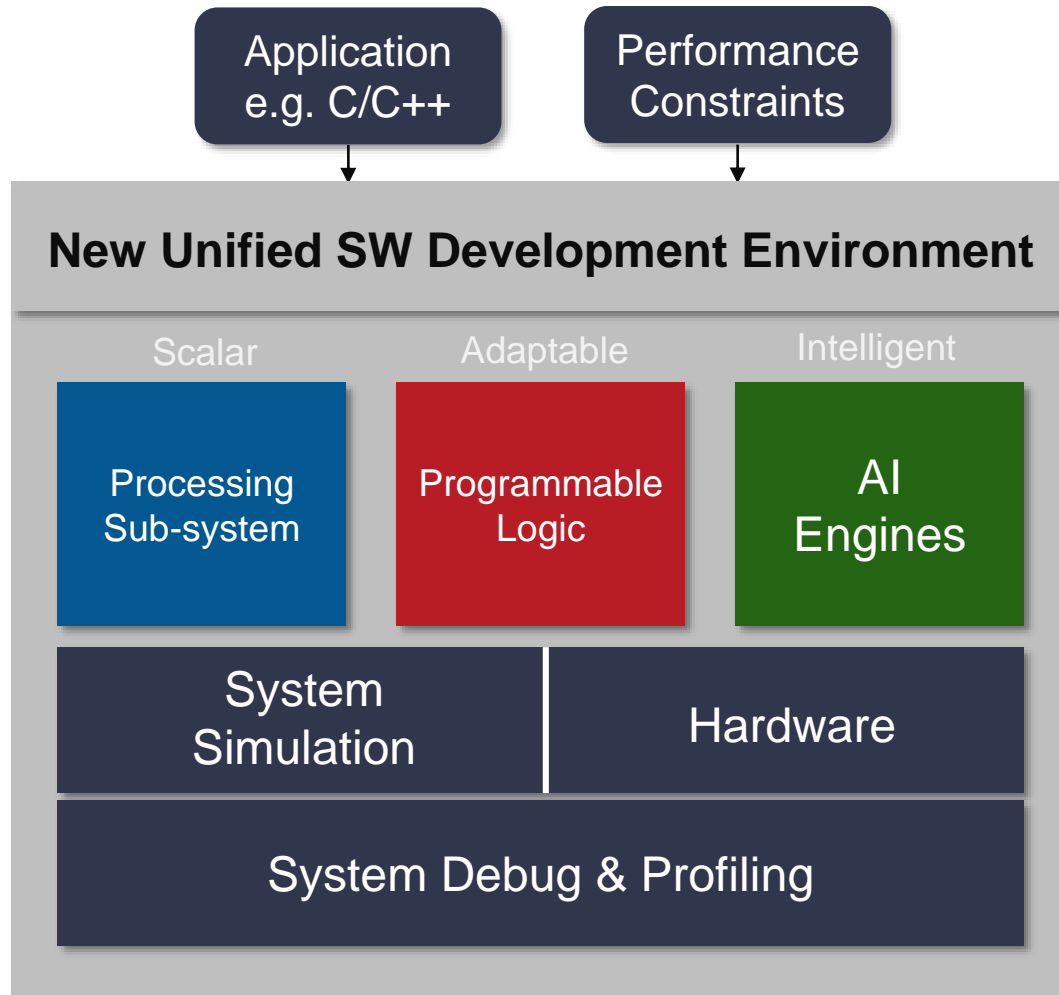
Versal ACAP Development Tools:

Any Application, Any Developer

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	

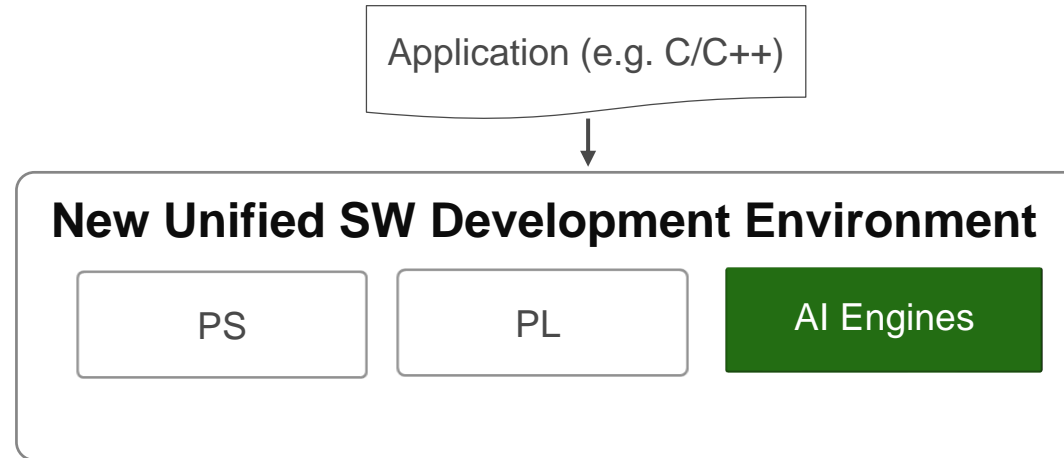


Software Development Environment

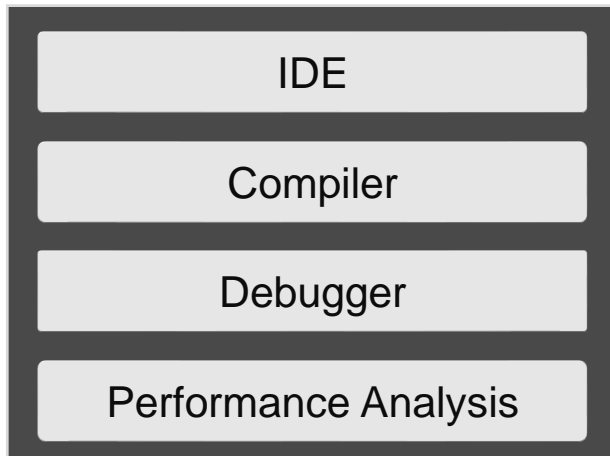


- > **Unified development environment**
 - >> Full chip programming
- > **SW programmable for whole application**
 - >> Heterogeneous SW acceleration
- > **Full system simulation, debug & profiling**
 - >> Software development experience

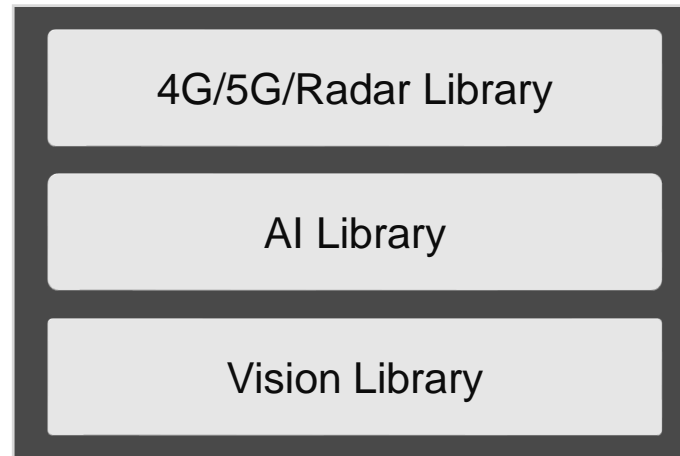
AI Engine Programming Environment



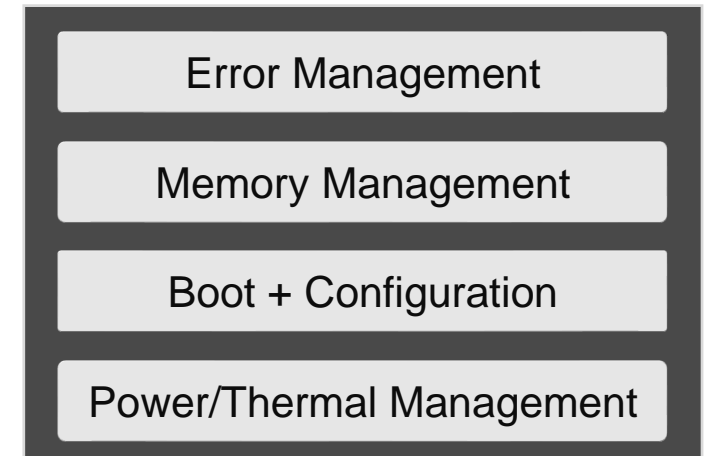
Full SW Programming Tool Chain *(Single-engine and Multi-engine)*



Performance-Optimized Software Libraries *(Examples)*

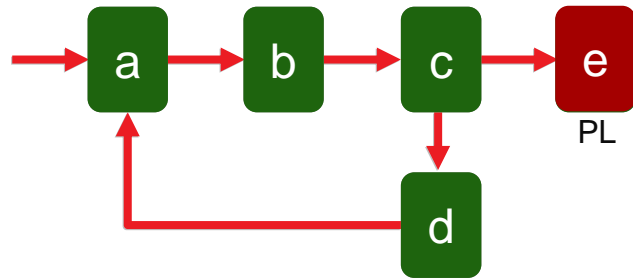


Run-Time Software *(Examples)*



AI Engine Programming Experience: Dataflow Model

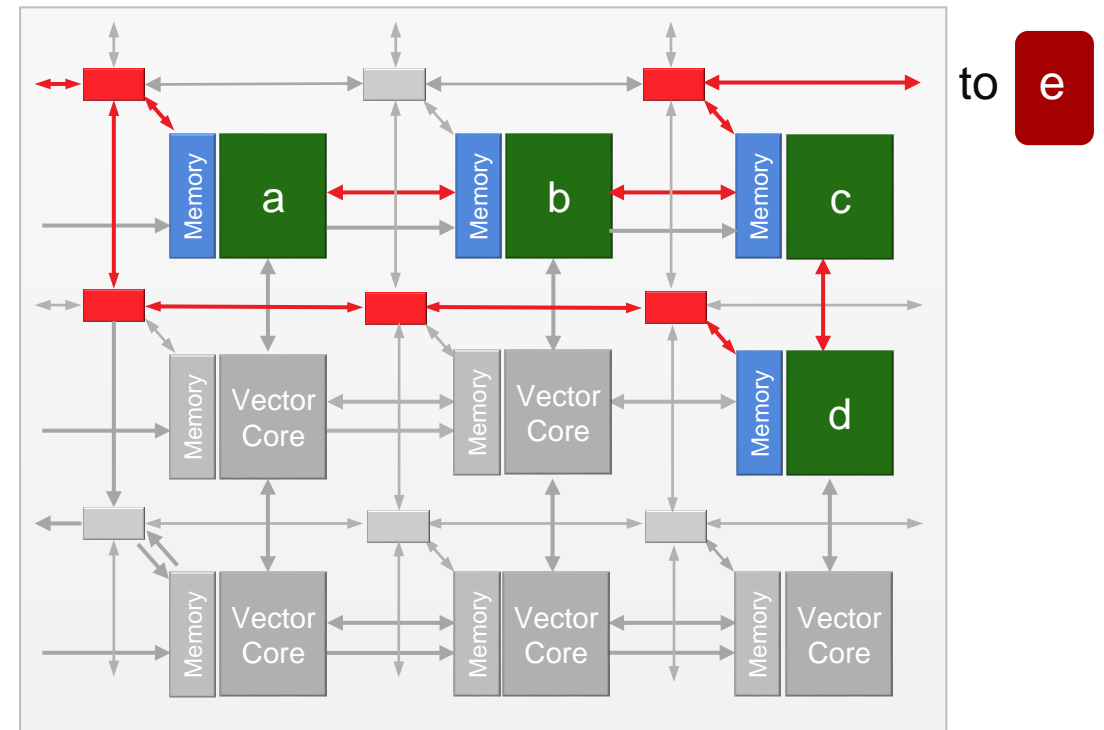
1 User defines dataflow logic



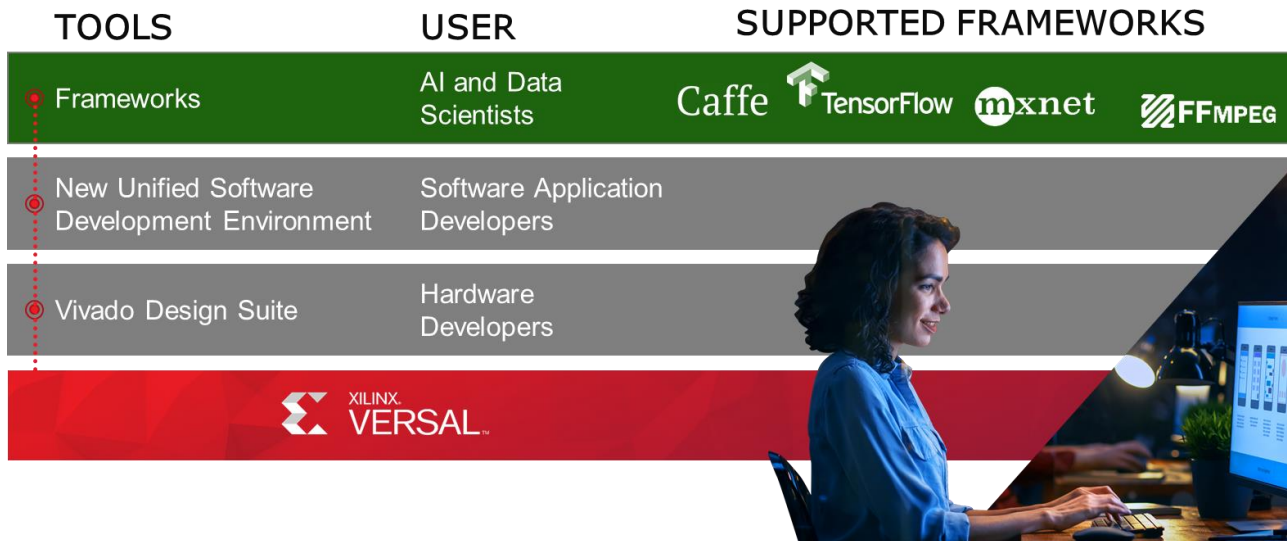
2 User describes dataflow graph using C/C++ APIs

3 Compiler transparently manages placement & interconnect

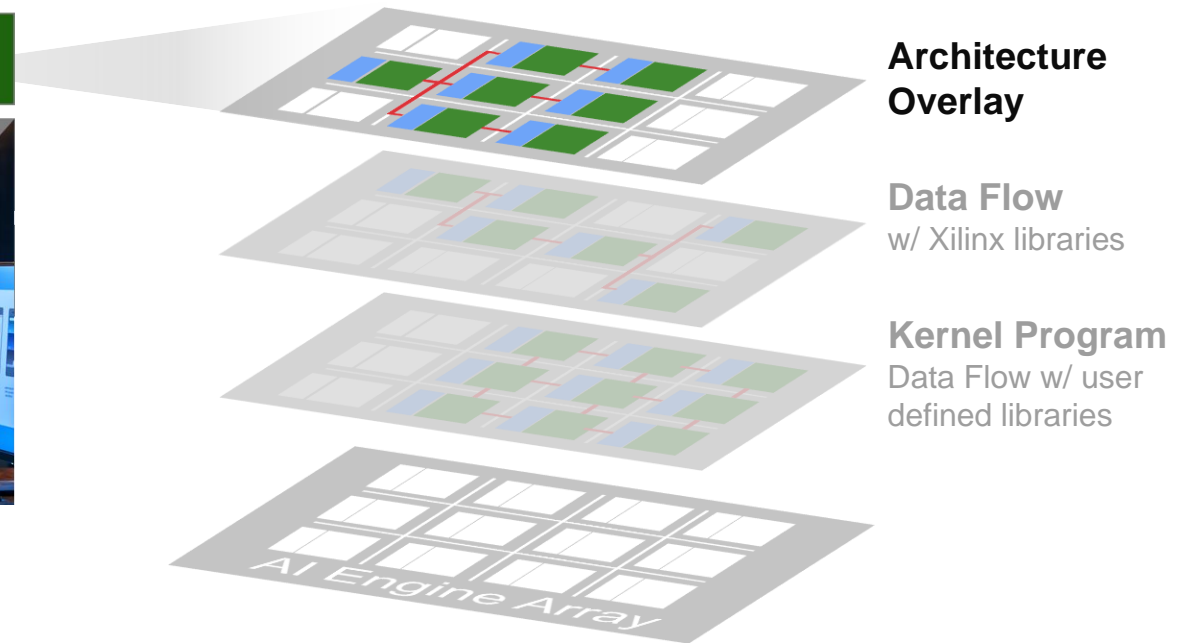
Physical Mapping to AI Engines



Frameworks for Any Developer



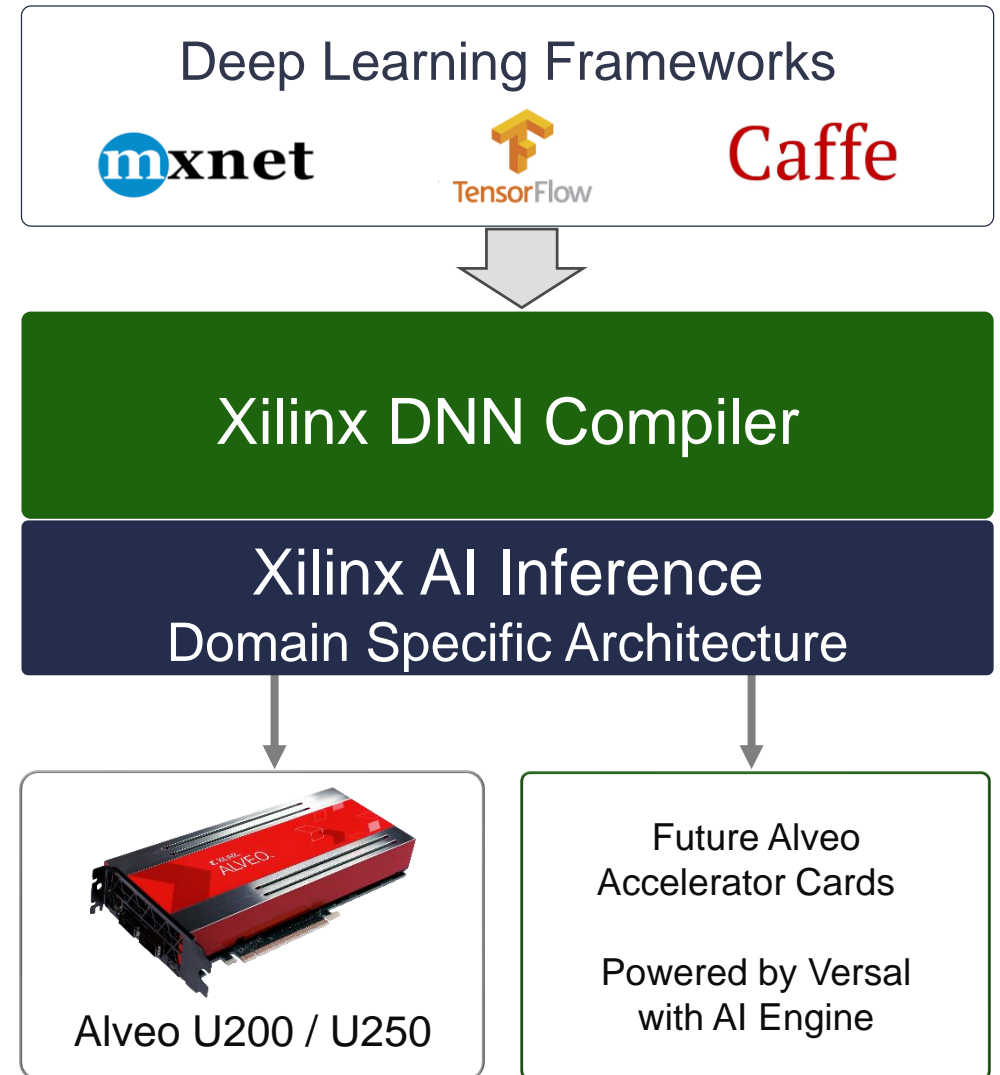
Domain Specific Architecture
(e.g. AI Inference)



Target Domain Specific Architectures – No HW Design Experience Required

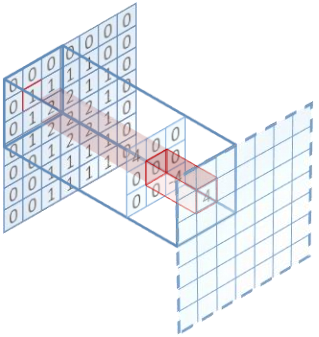
Accelerating AI Inference in the Data Center

- 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

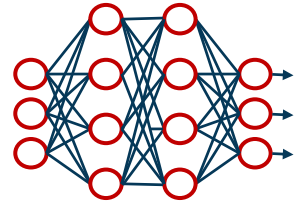


AI Inference on Versal ACAP

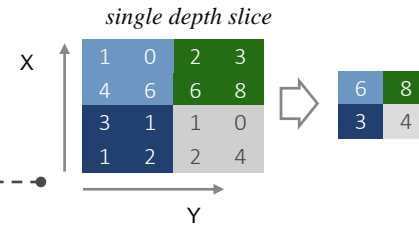
Convolutions



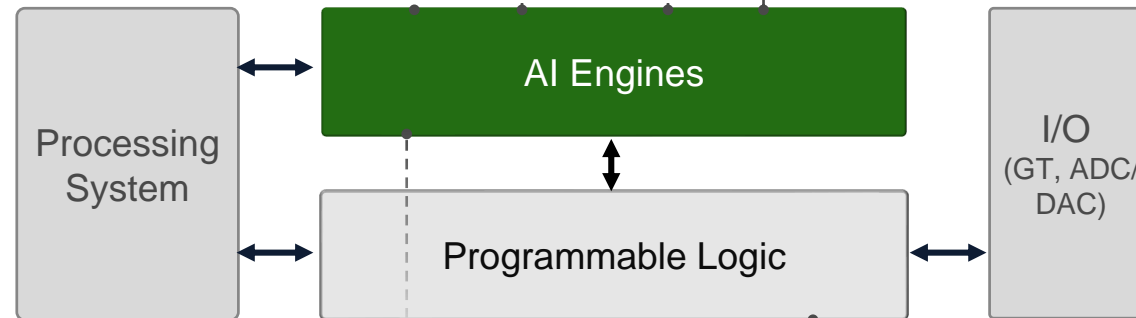
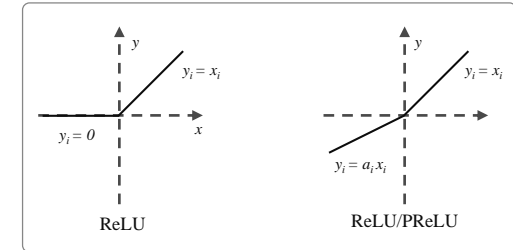
Fully Connected Layers



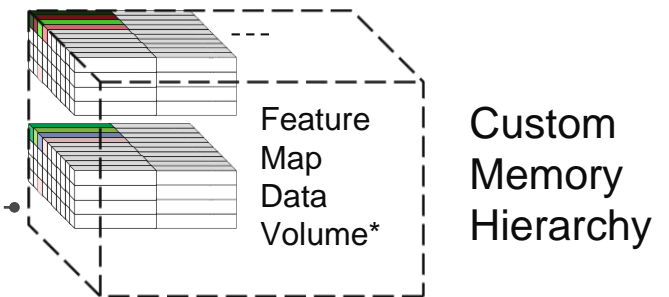
Pooling



Activations



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

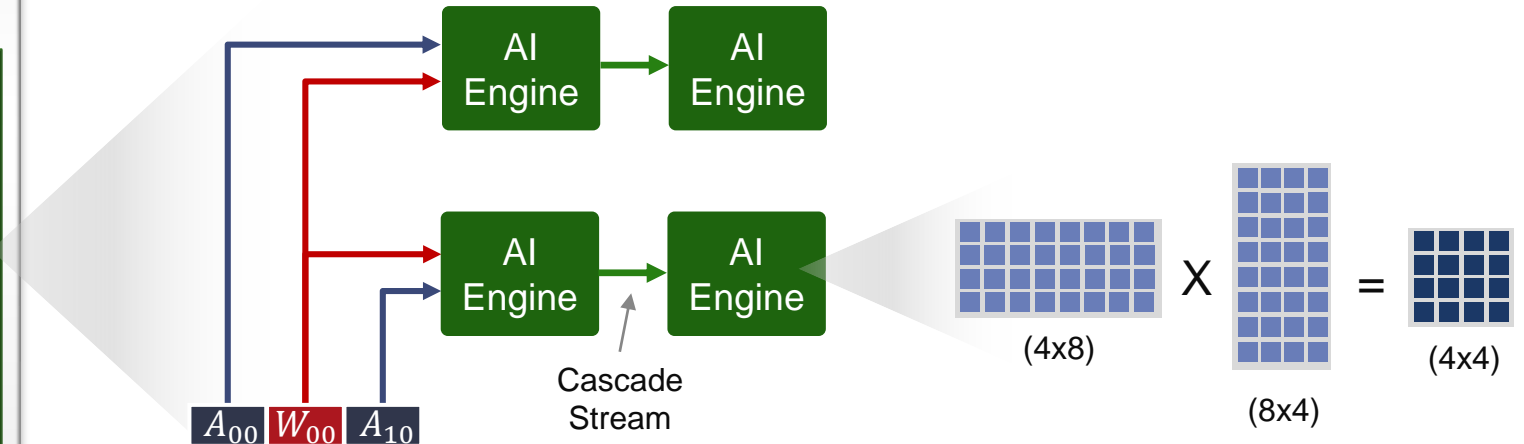
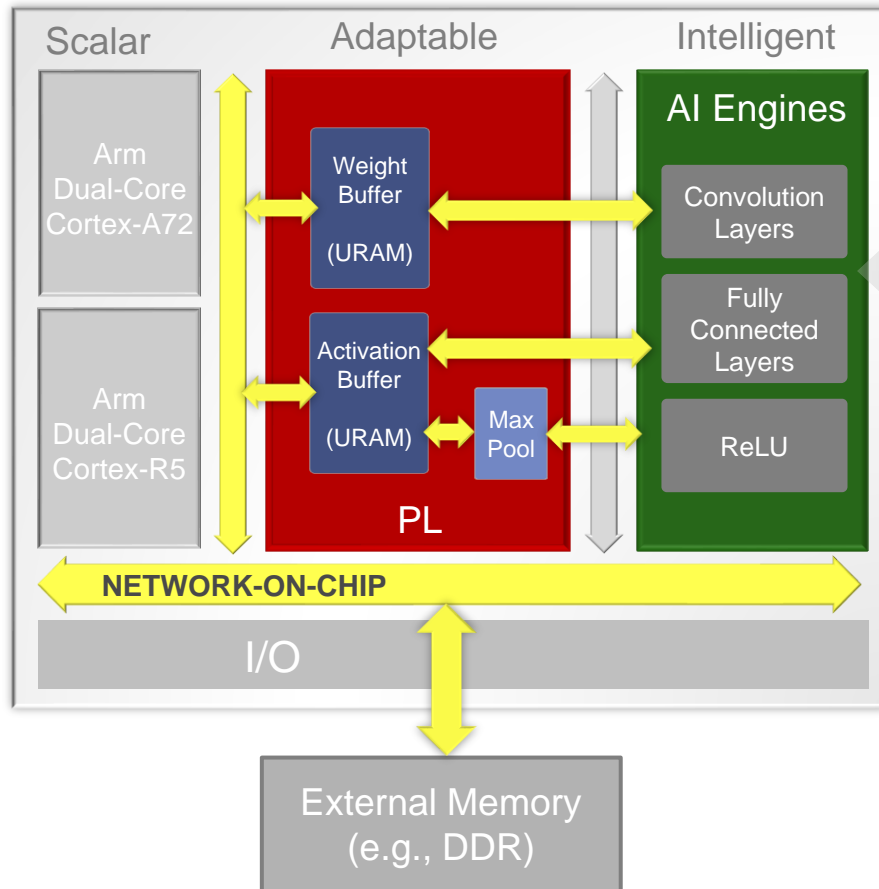


*Figure credit: https://en.wikipedia.org/wiki/Convolutional_neural_network

AI Inference Mapping on Versal ACAP

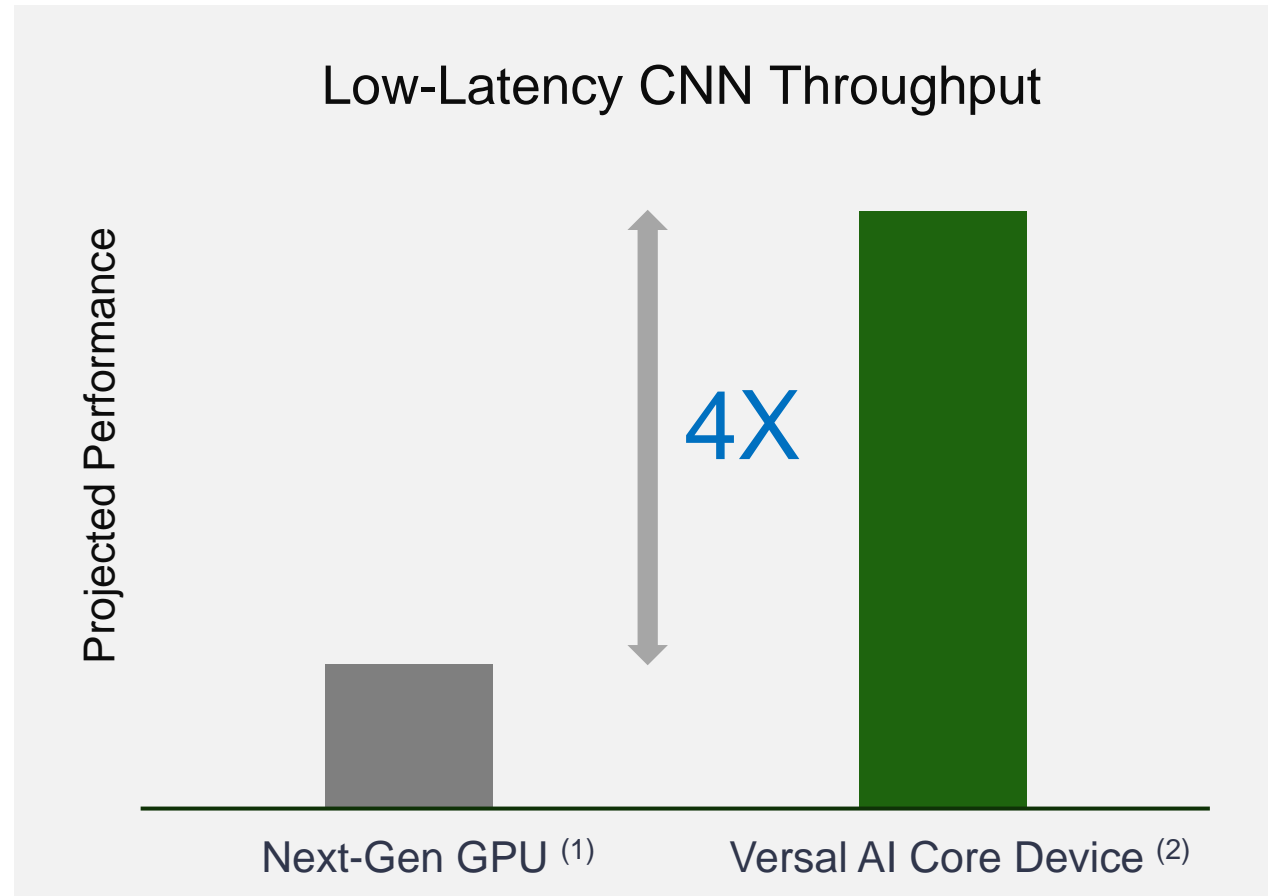
A = Activations
W = Weights

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



- > 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)



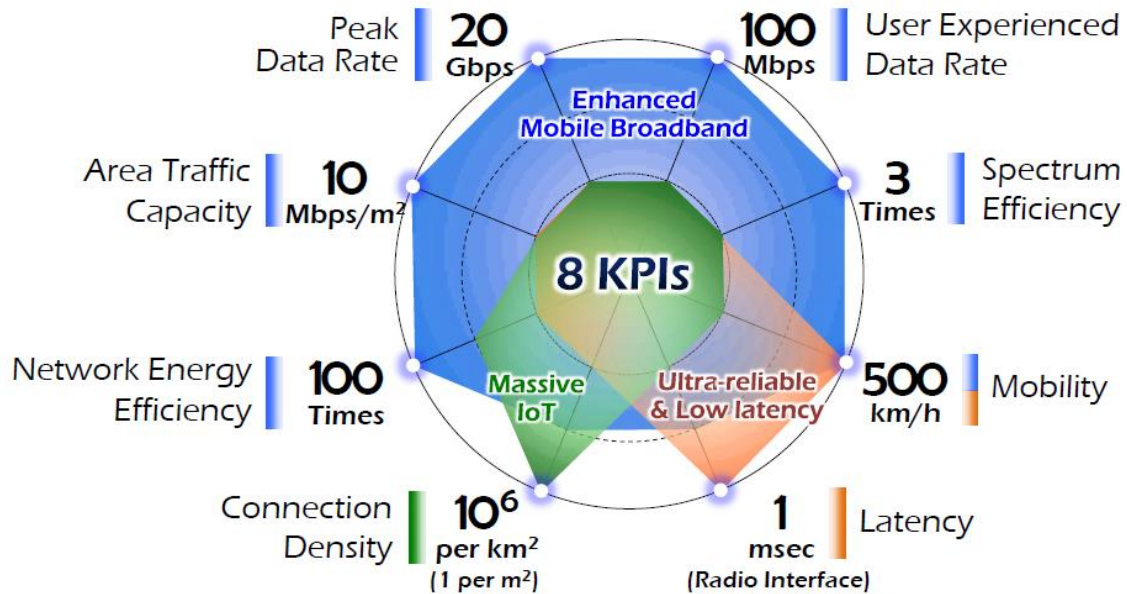
Note:
Versal device achieves 8X performance increase in 150W power envelope

(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

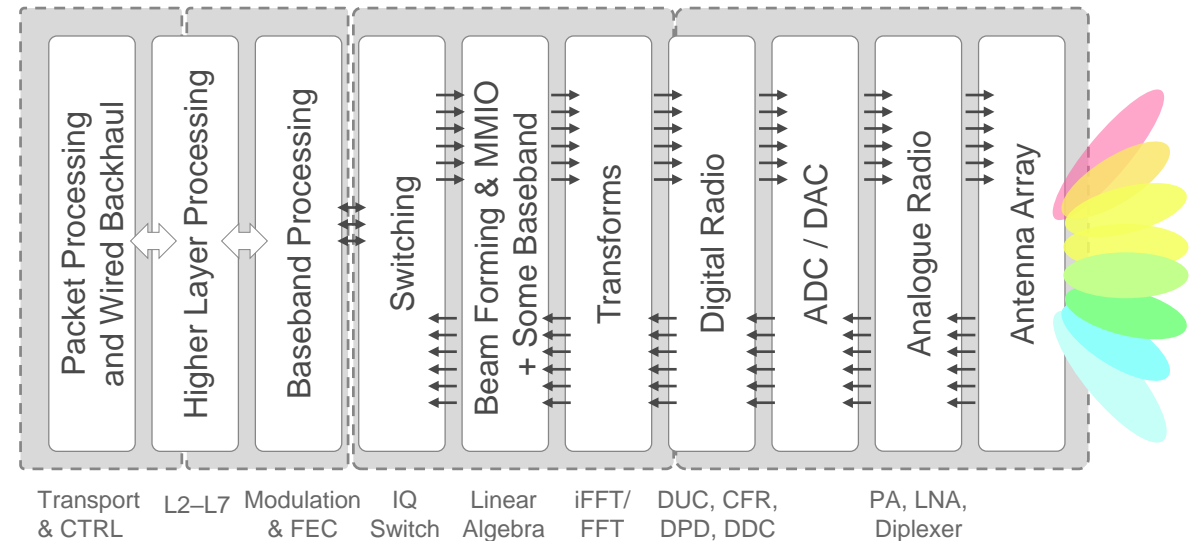
5G Complexity is 100X that of 4G Still Evolving Standard



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

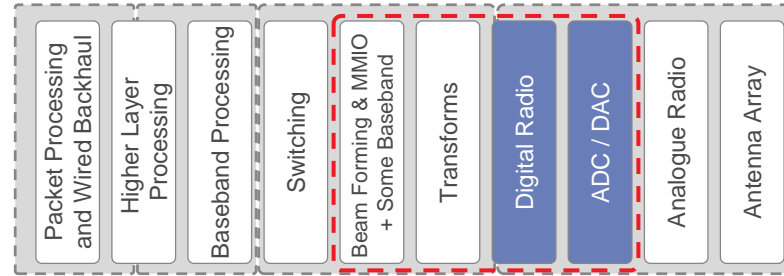
New Technologies in 5G

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

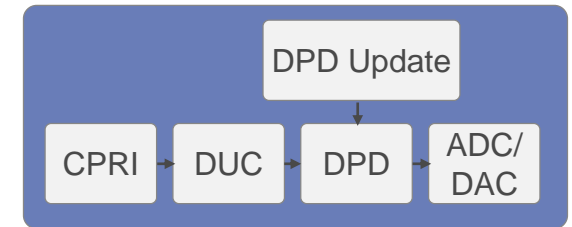


5G Wireless on Versal ACAP

5G Wireless Infrastructure (i.e., base-station)

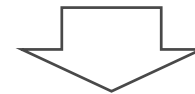


Digital Radio with ADC/DAC

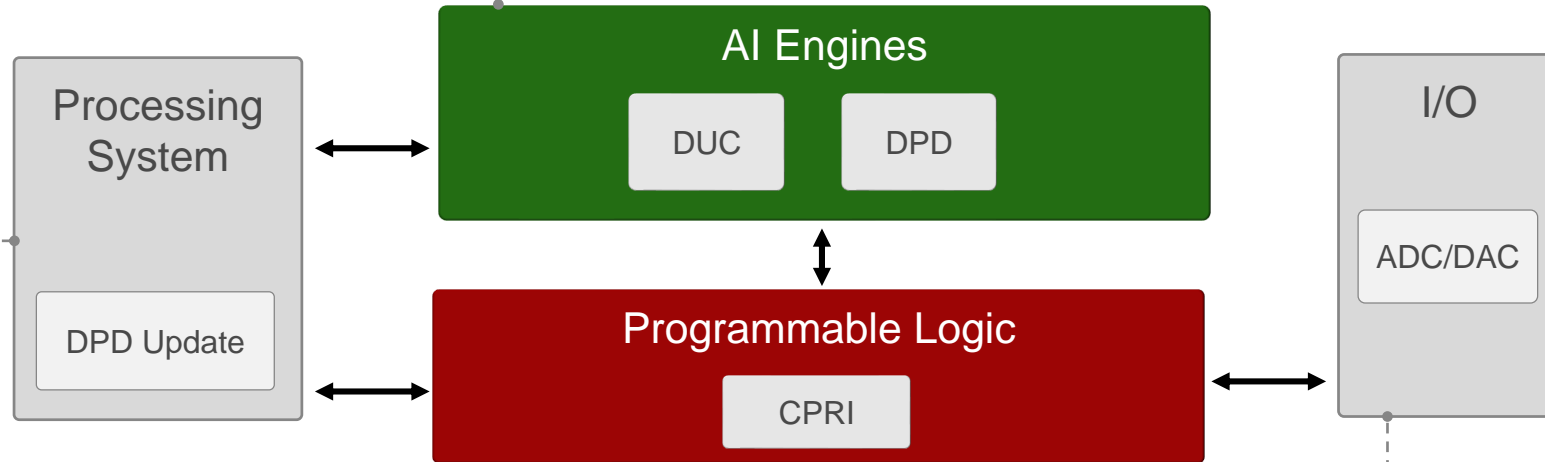


Compute Maps to AI Engine

Mapping Example



Control Maps to PS

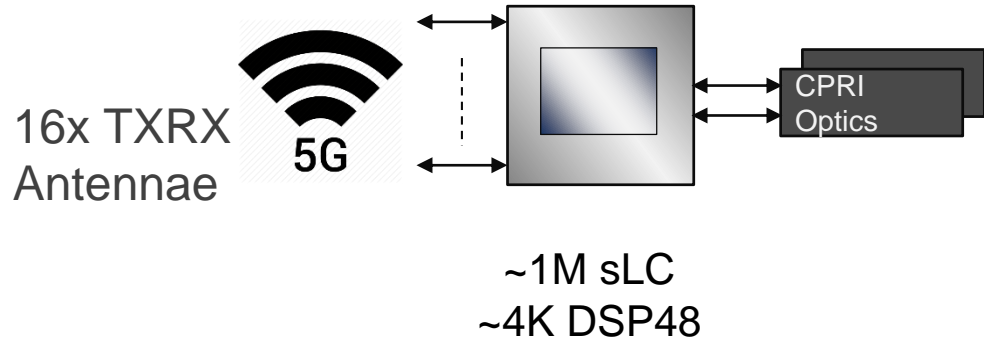


I/O Maps to PL

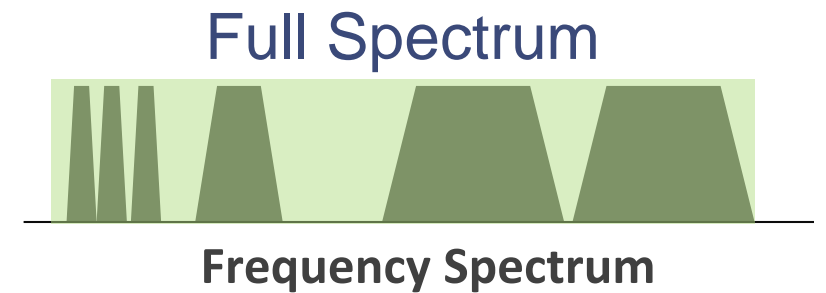
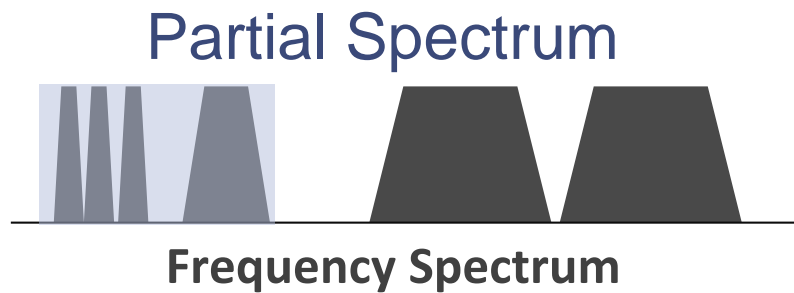
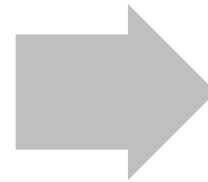
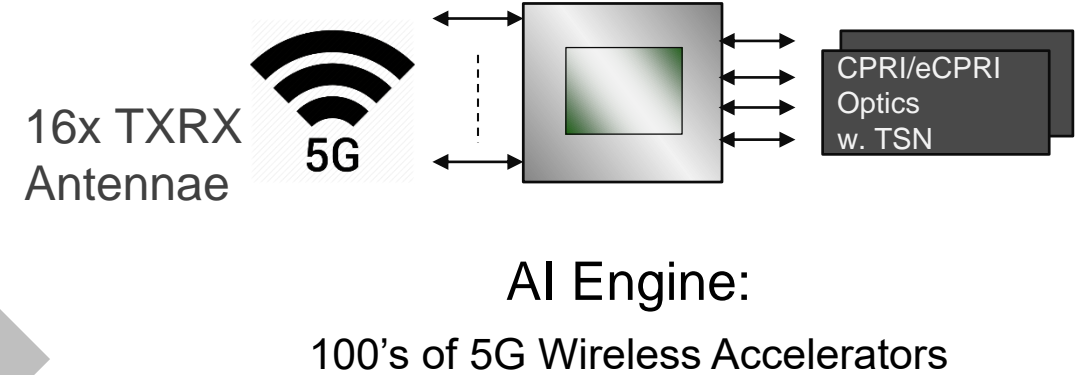
- 1: DUC: Digital Up Converter
- 2: DPD: Digital Pre-Distortion
- 3: Direct RF: ADC/DAC
- 4: CPRI: Common Public Radio Interface

AI Engine Delivers 5X more 5G Wireless Compute

Xilinx Zynq UltraScale+ RFSoc

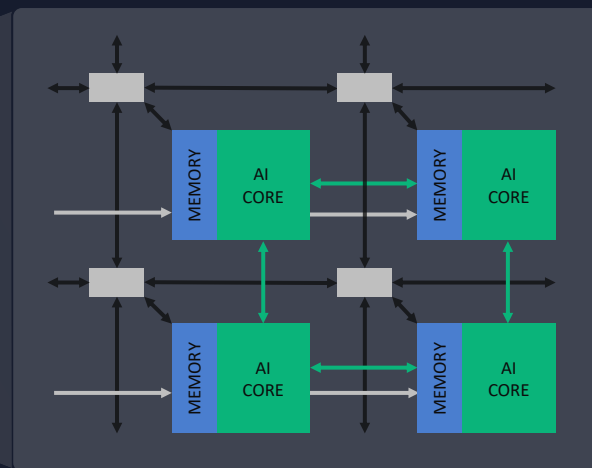
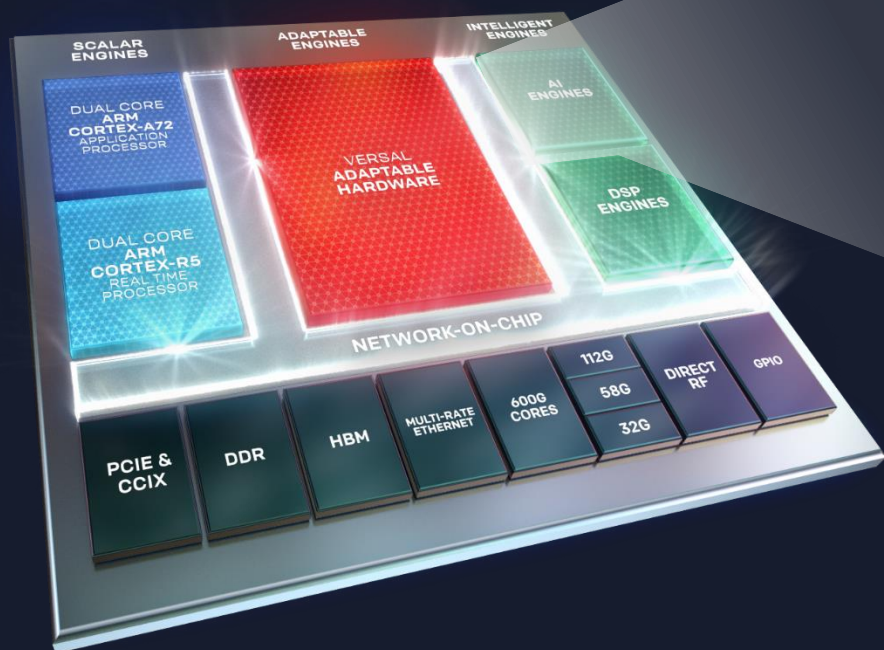


Xilinx Versal with AI Engine



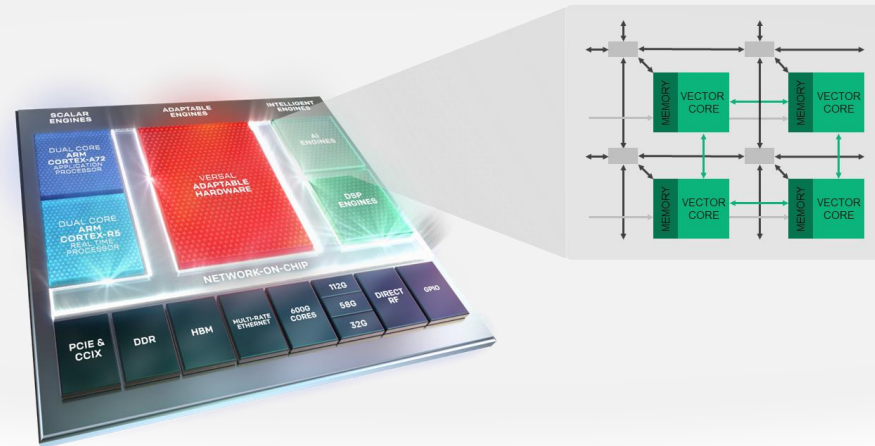
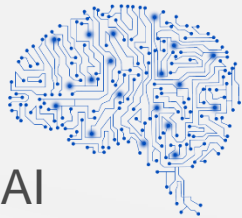
Enabling Single Chip Massive MIMO 16x16 800 MHz Radio

Wrapping Up

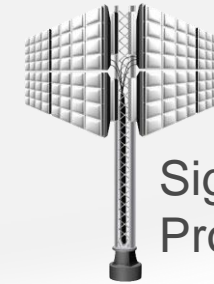


AI Engine: Accelerating AI Inference & Signal Processing

20x
AI Inference



5x
Signal Processing



Software Programmable

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

Deterministic

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

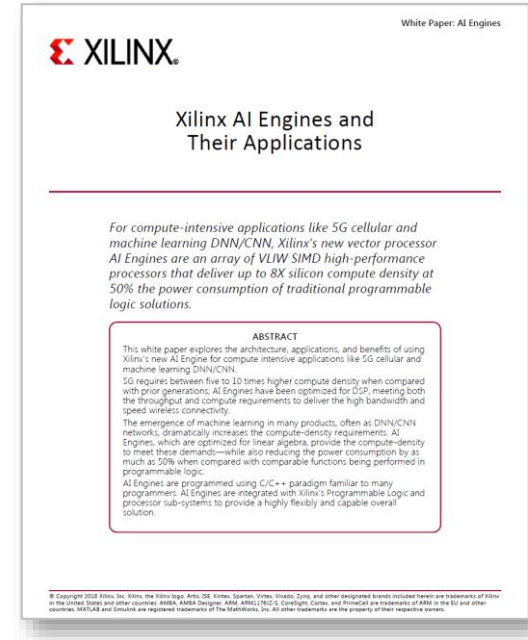
Efficient

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

See What AI Engine Can Do For You

Read the AI Engine White Paper

- Visit: www.xilinx.com/versal



Find out more about AI Engine

- Early Access Program Open Now
- Contact Your Xilinx Sales Representative

The logo consists of a red chevron pointing right, followed by the letters 'XDF' in a white, bold, sans-serif font.

XILINX
DEVELOPER
FORUM

