Building the Adaptable, Intelligent World

Ivo Bolsens
Senior Vice President & Chief Technology Officer

XILINX

Mountains of Unstructured Data

One Architecture Can’t Do It Alone

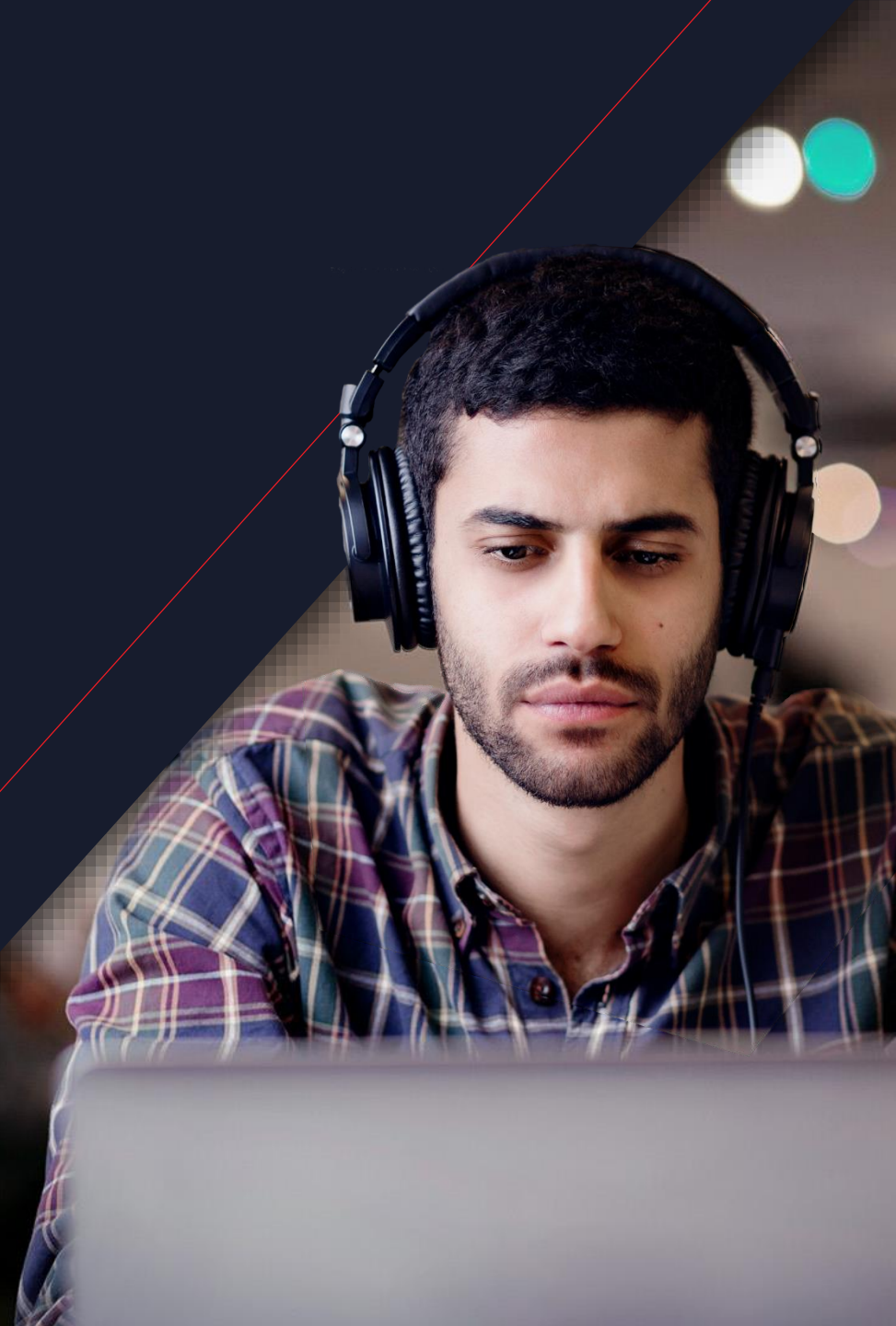
This is the Era of Heterogeneous Compute

Today’s Developer Needs

Performance for a Diverse Range of Applications

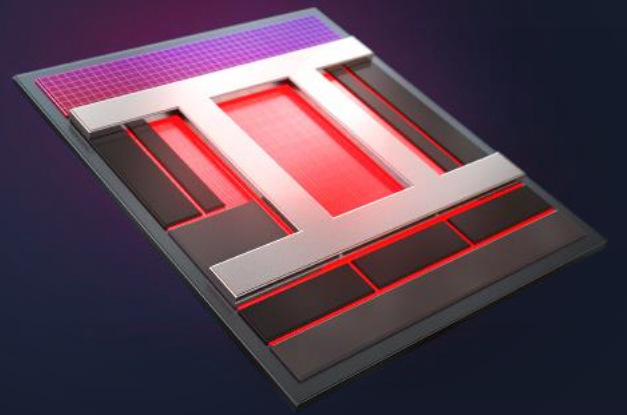
Software Programmability

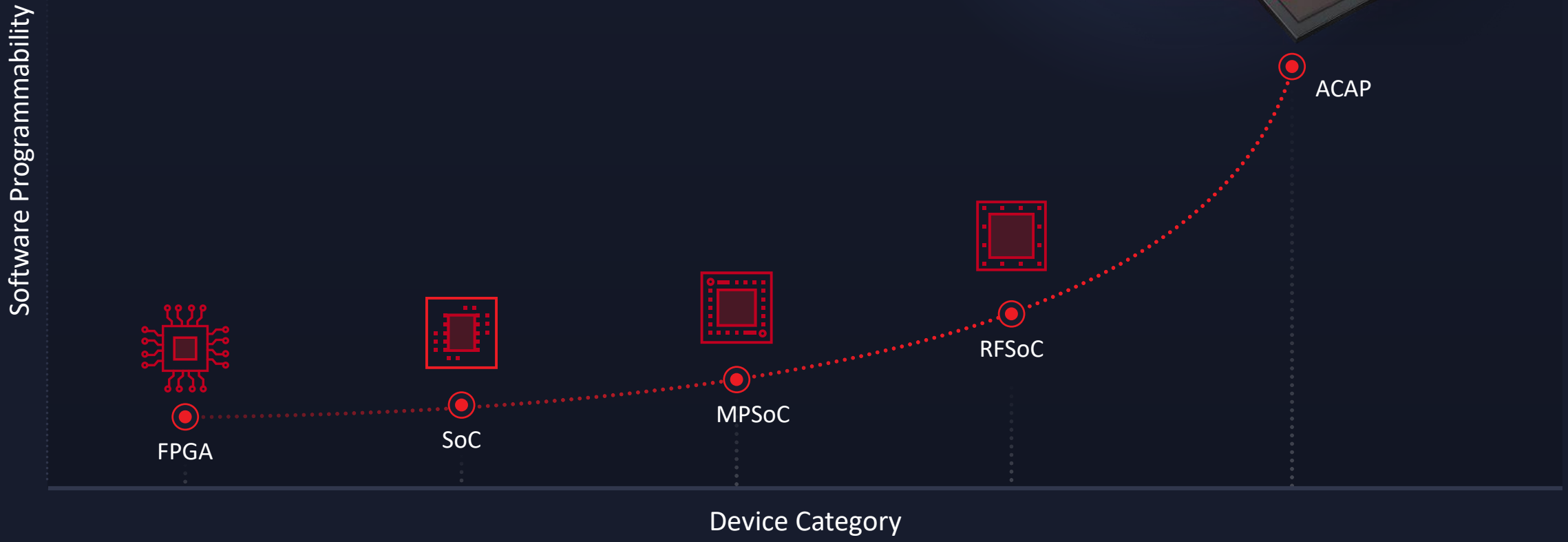
Adaptability to Keep Pace with Rapid Innovation

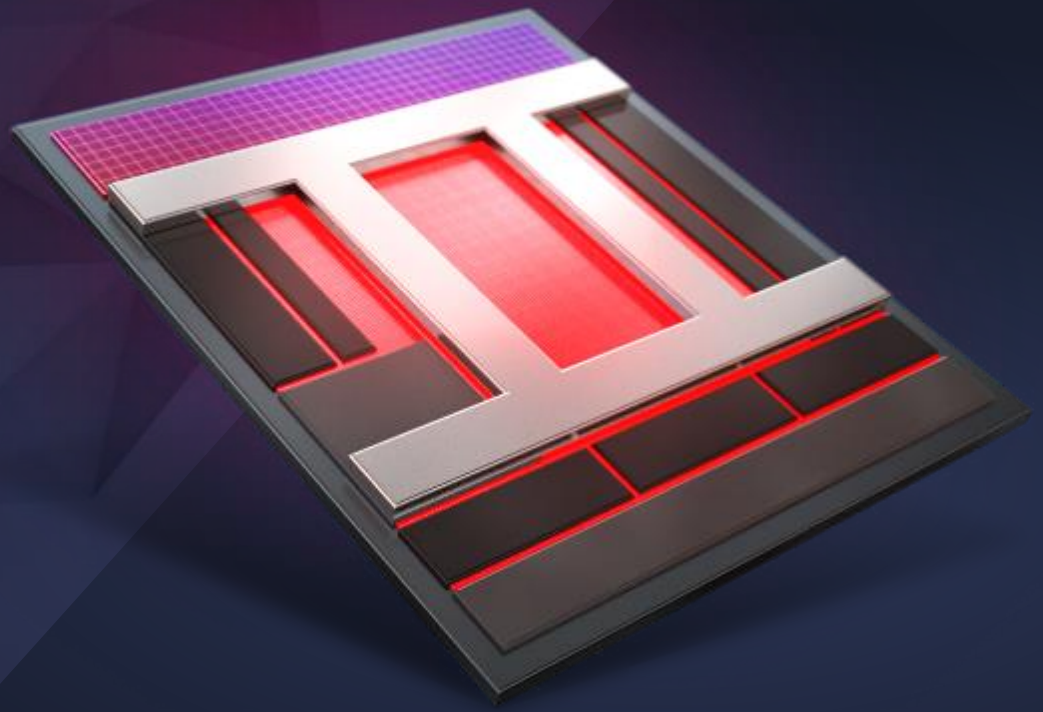


Enter the ACAP

A New Class of Devices for Today's Challenges

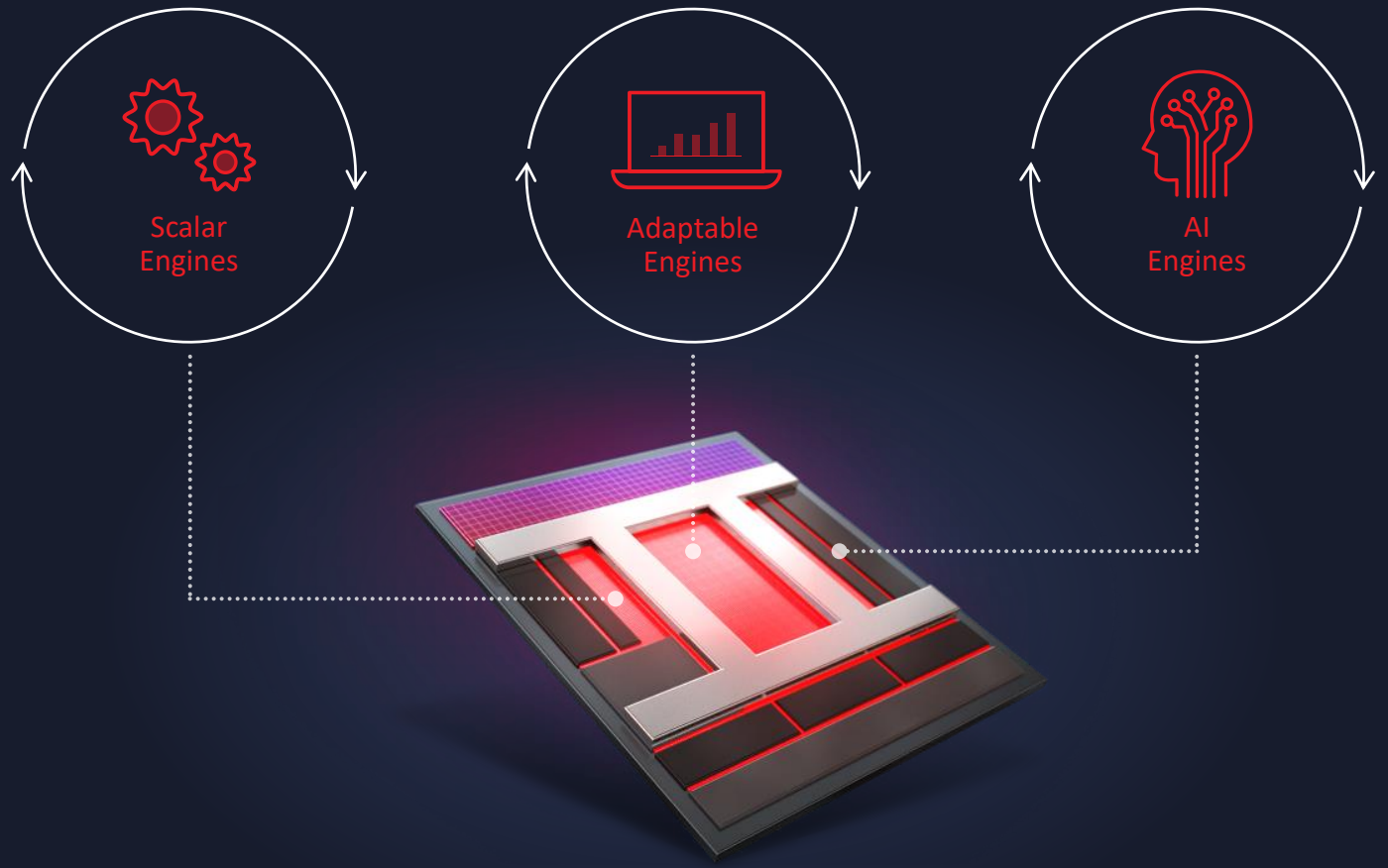






Adaptive Compute Acceleration Platform

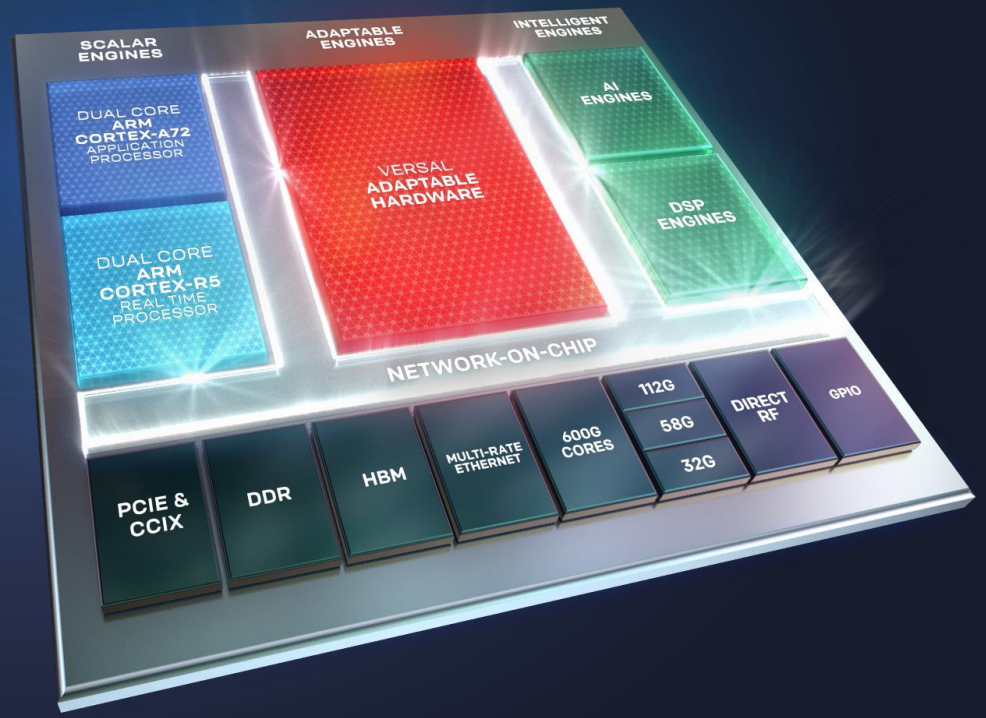
Compute Acceleration

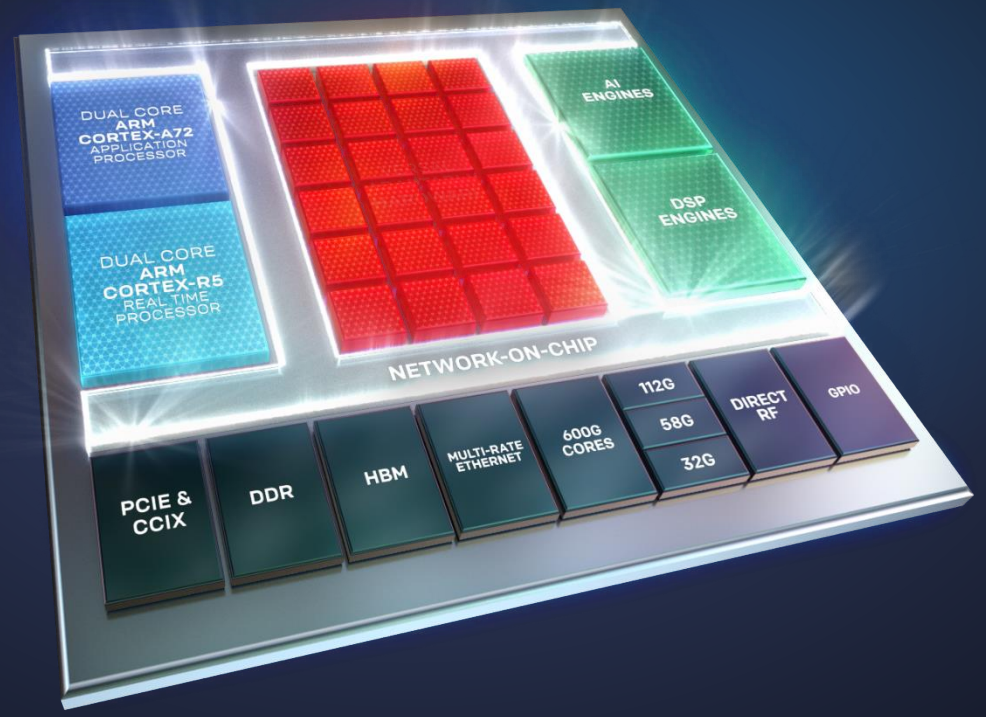


XILINX
VERSAL™

The Industry's First ACAP

- Heterogeneous
- Parallel
- Scalable
- SW Programmable
- HW Adaptable





Network-on-Chip (NoC)

Ease of Use

Inherently Software Programmable

Available at Boot, No Place-and-Route Required

High Bandwidth and Low Latency

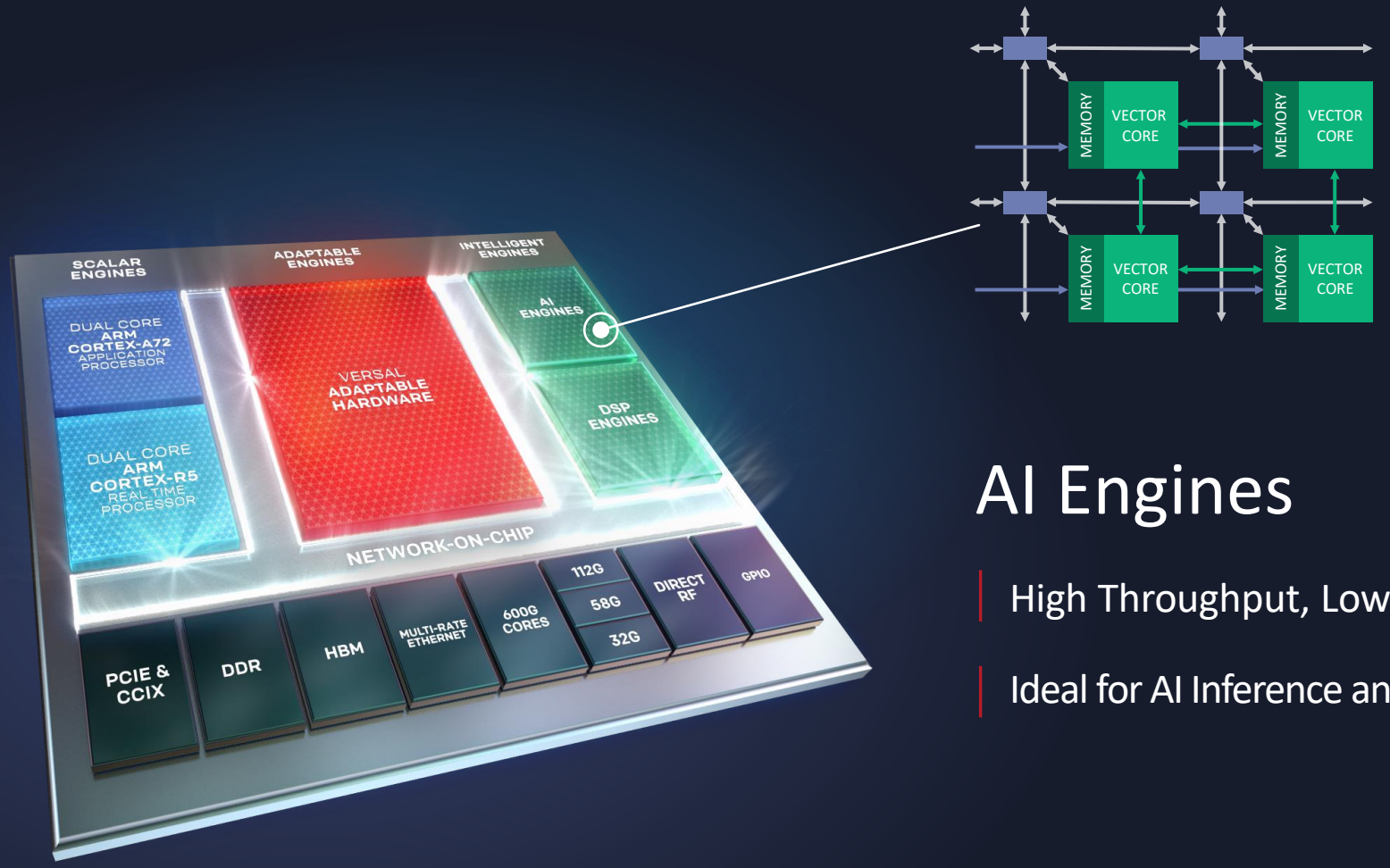
Multi-Terabit/Sec Throughput

Guaranteed QoS

Power Efficiency

8x Power Efficiency vs. Soft Implementations

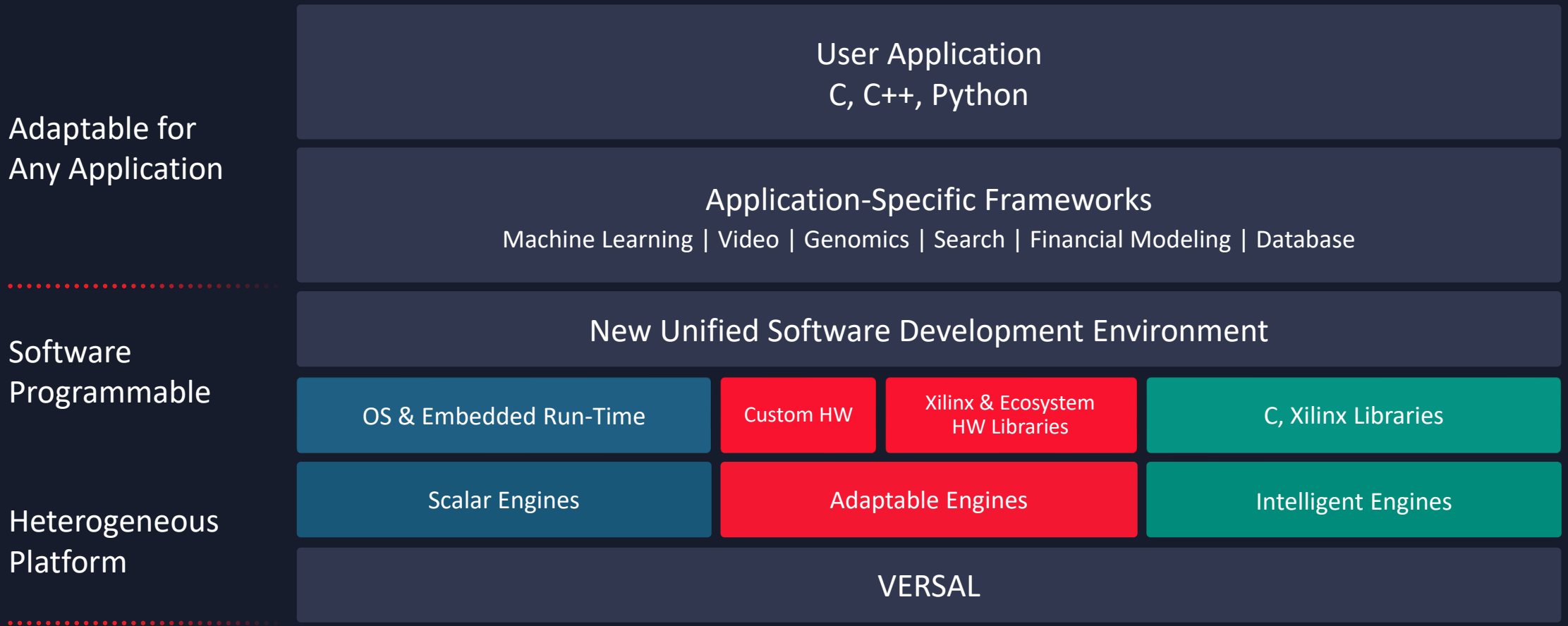
Arbitration Across Heterogeneous Engines



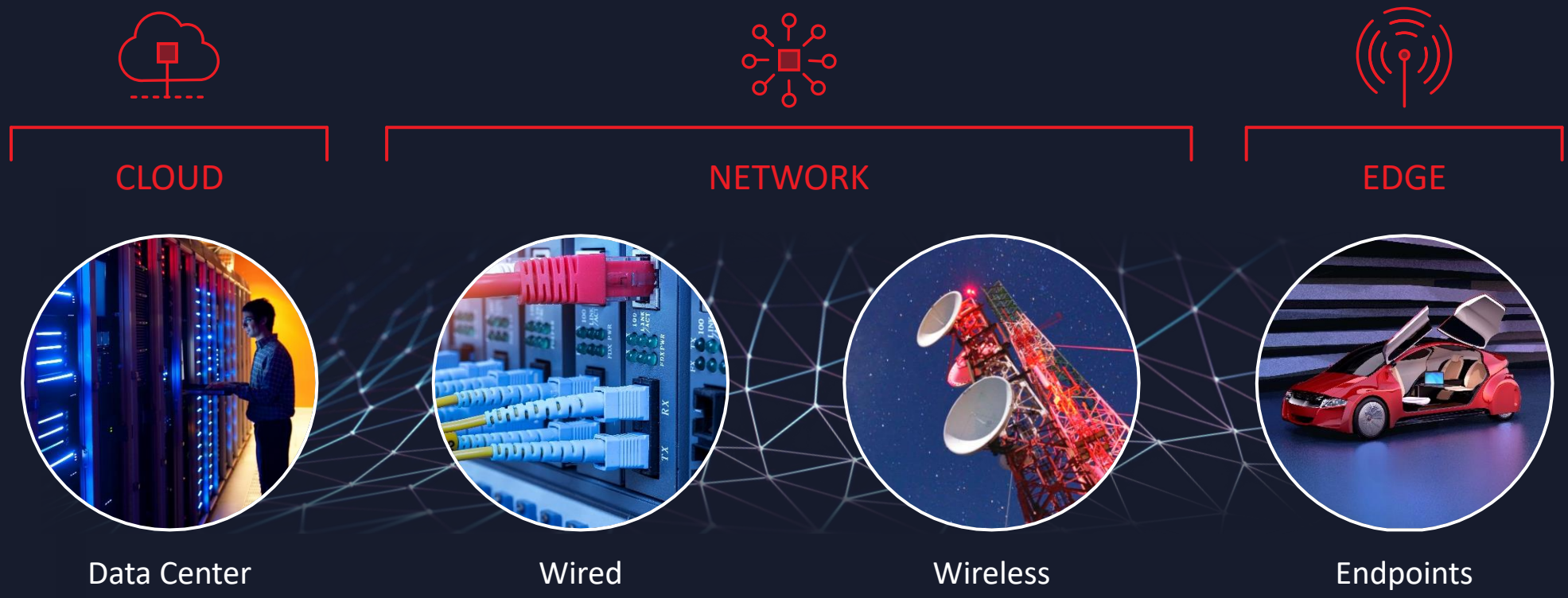
AI Engines

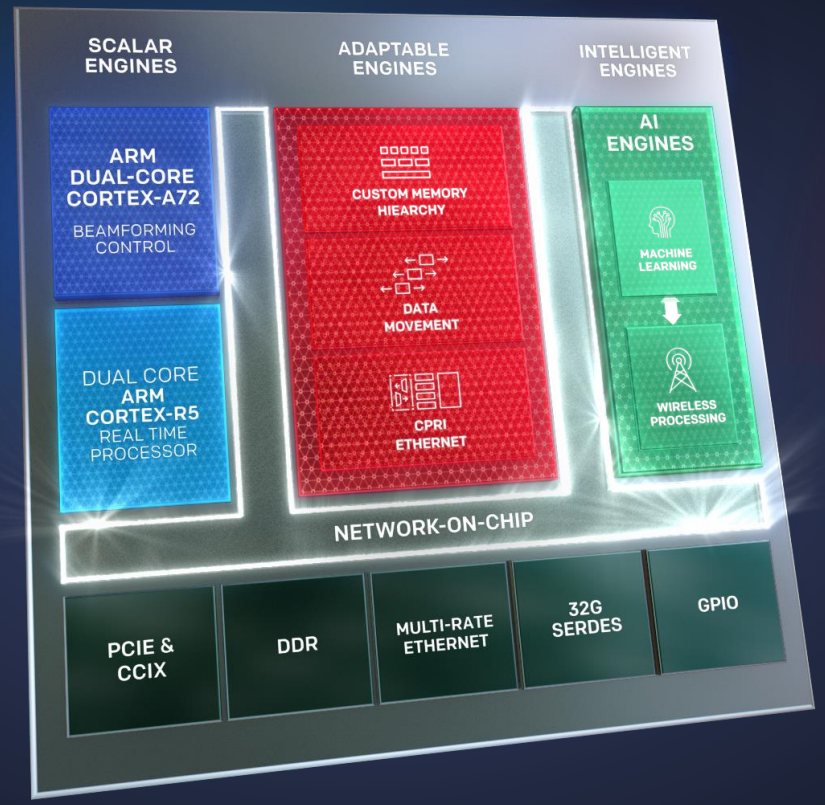
- High Throughput, Low Latency, and Power Efficient
- Ideal for AI Inference and Advanced Signal Processing

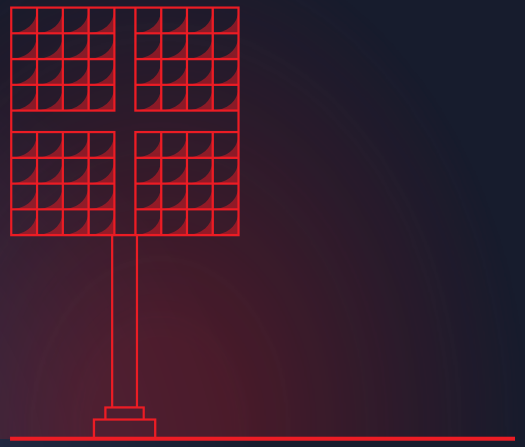
Platform for Any Developer



Versal Multi-Market Platform







VERSAL AI Core Series

For 5G Beamforming & CloudRAN

AI Engines Provides >5X Compute Density for Advanced Wireless Compute

Versal Multi-Market Platform

CLOUD

NETWORK

EDGE

Data Center

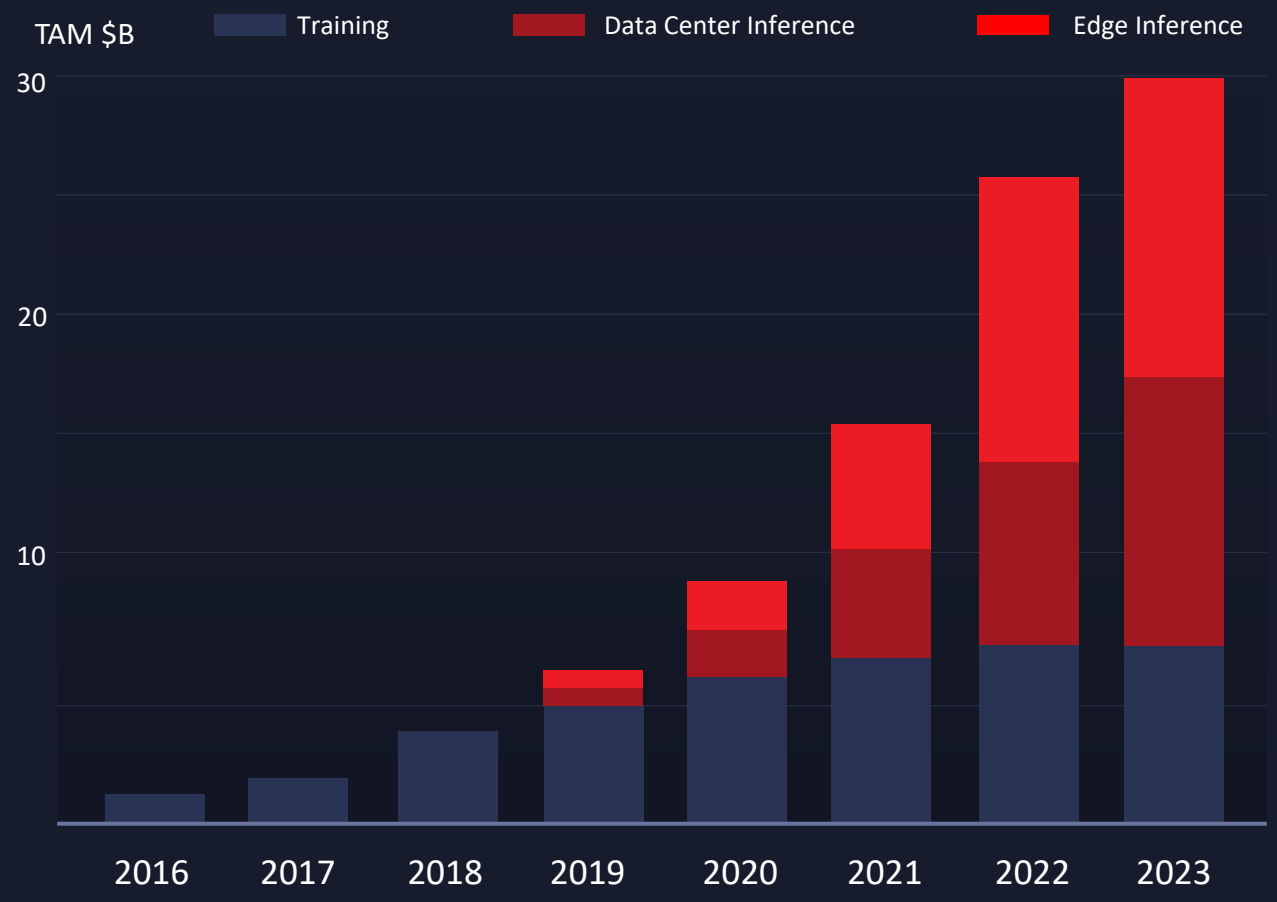
Wired

Wireless

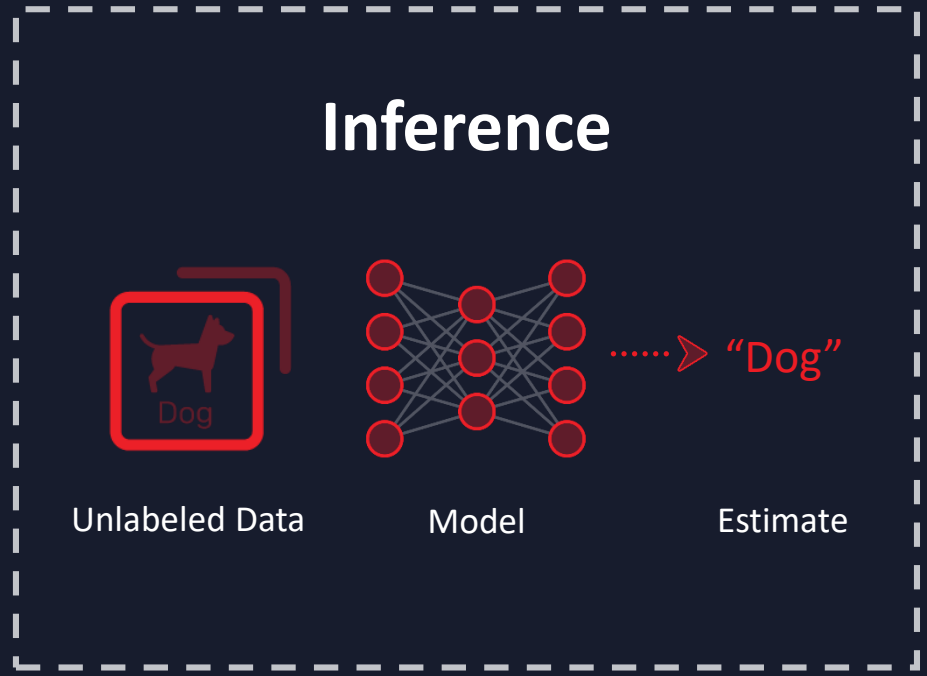
Endpoints

AI ADOPTION ACROSS MARKETS

Projected Growth in AI Inference

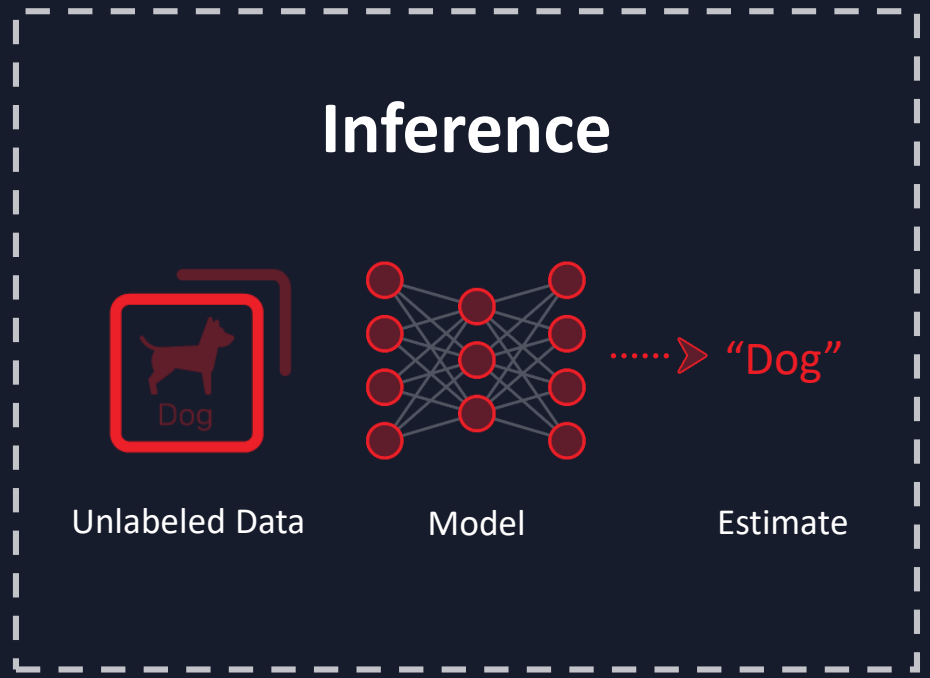


Barclays Research, Company Reports May 2018



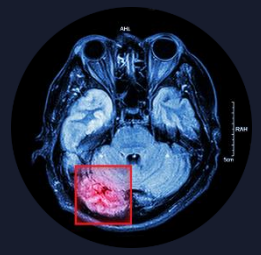
Challenges

- The Rate of AI Innovation
- Performance at Low Latency
- Low Power Consumption
- Whole App Acceleration



The Rate of AI Model Innovation

APPLICATIONS

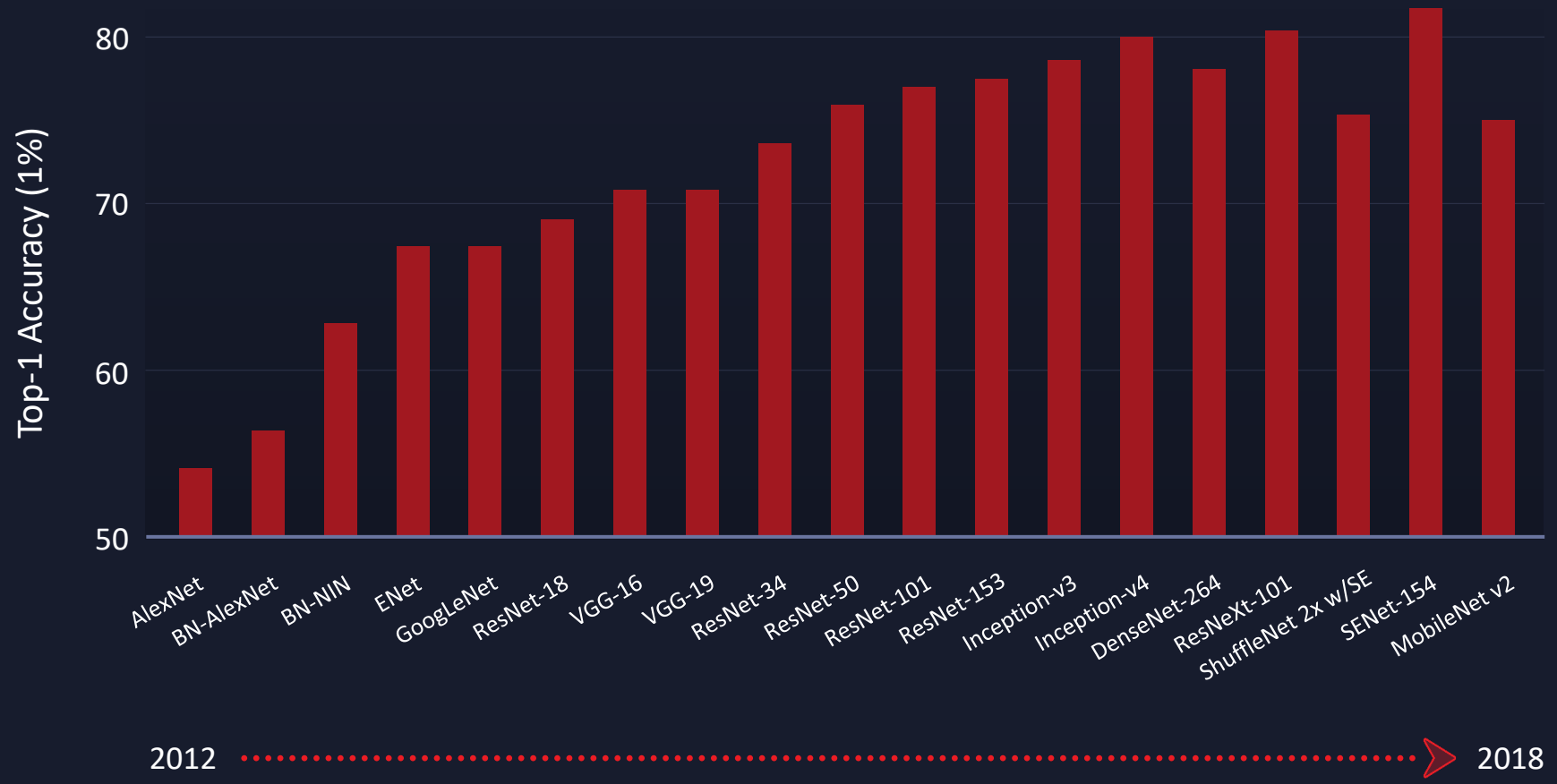
Classification	Object Detection	Segmentation	Speech Recognition	Recommendation Engine	Anomaly Detection
					
CNN			RNN, LSTM	MLP	

DIVERSE MODELS OVER A BROAD RANGE OF APPLICATIONS

The Rate of AI Model Innovation

Classification

Classification

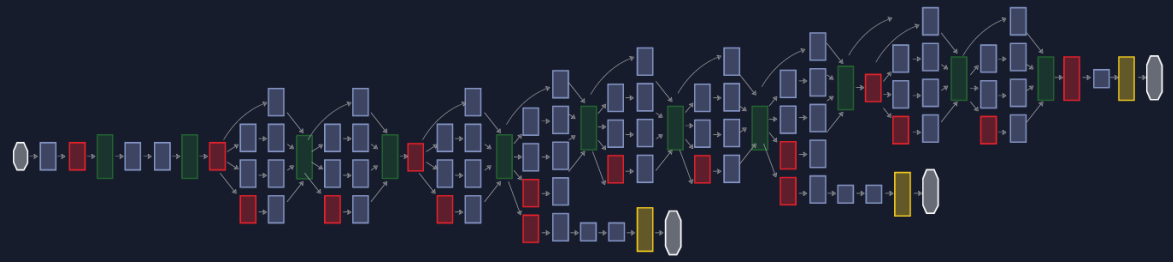


Source:
https://arxiv.org/pdf/1605.07678.pdf https://arxiv.org/pdf/1608.06993.pdf
https://arxiv.org/pdf/1709.01507.pdf https://arxiv.org/pdf/1611.05431.pdf

Rate of Innovation Outpaces Silicon Cycles

AlexNet

GoogLeNet



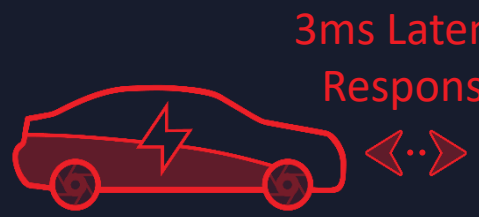
DenseNet

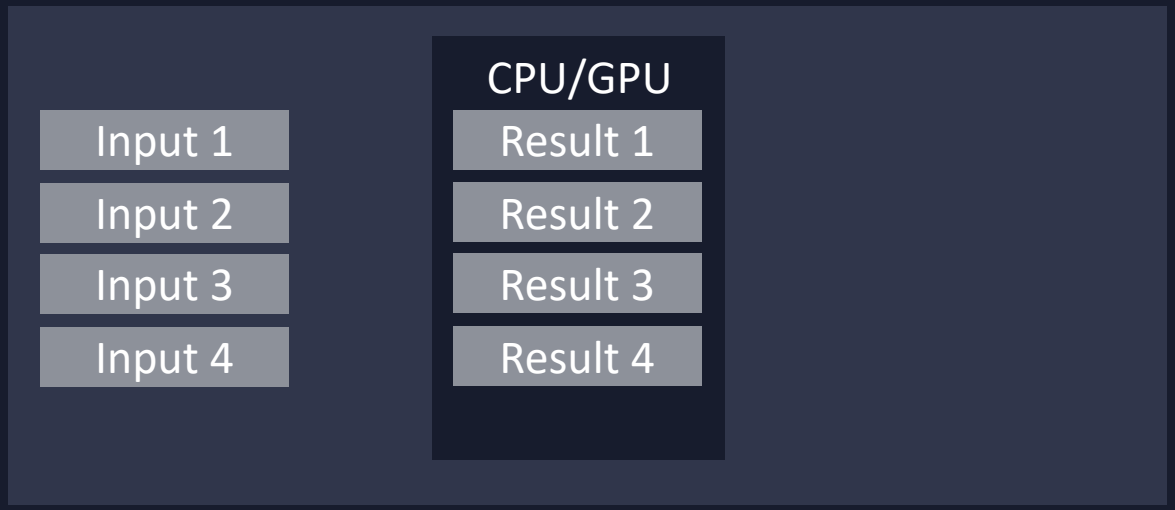
Start Design

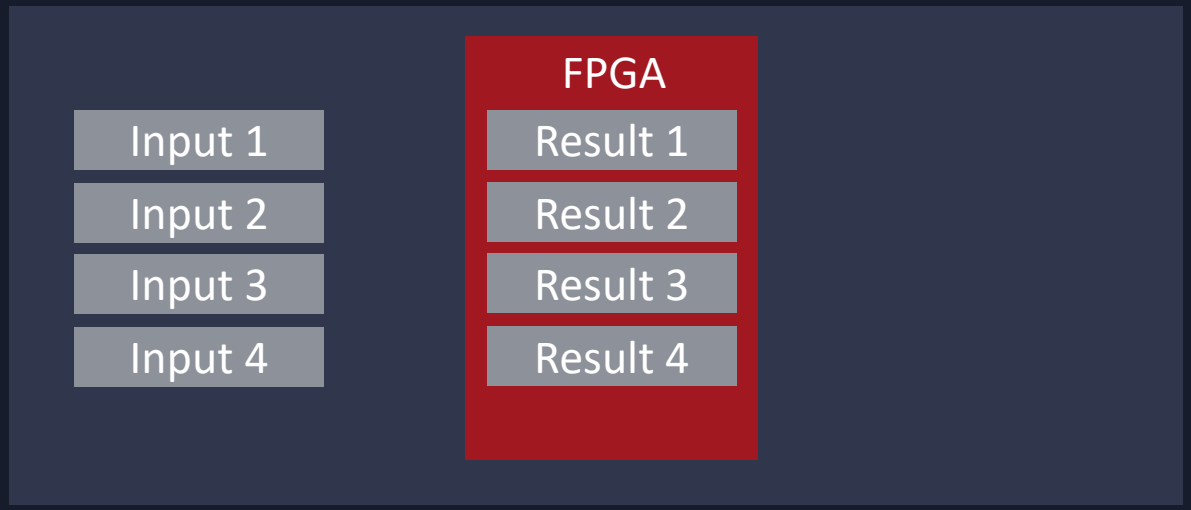
Silicon Design Cycle (time)

Production Design

Low Latency is Critical for Inference



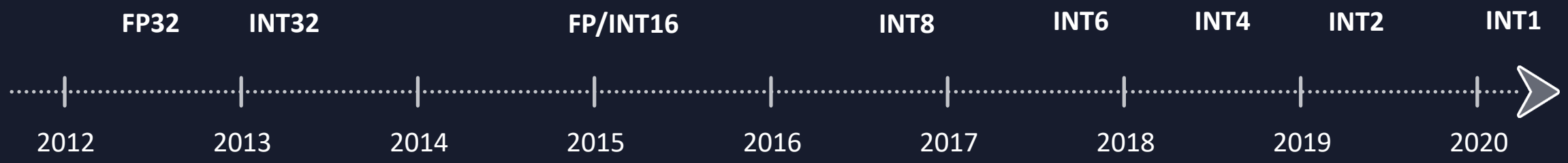




High Throughput OR Low Latency

High Throughput AND Low Latency

Inference Moving to Lower Precision

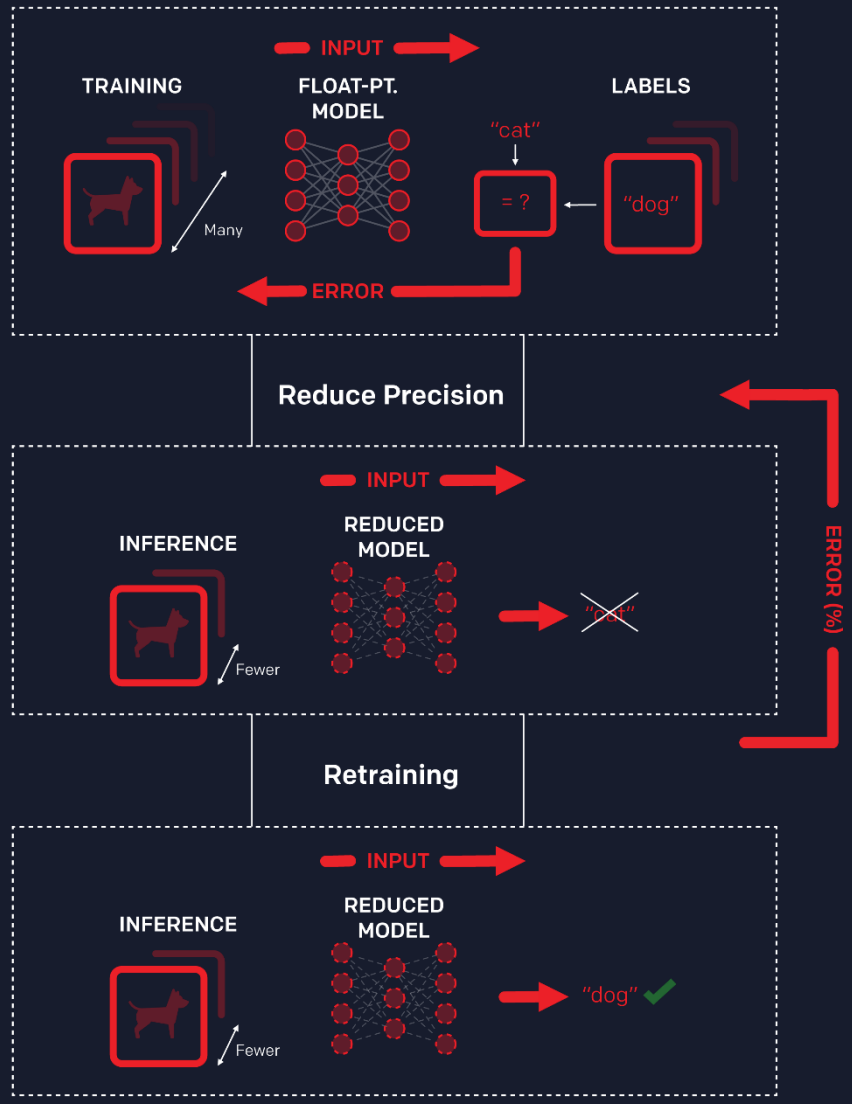


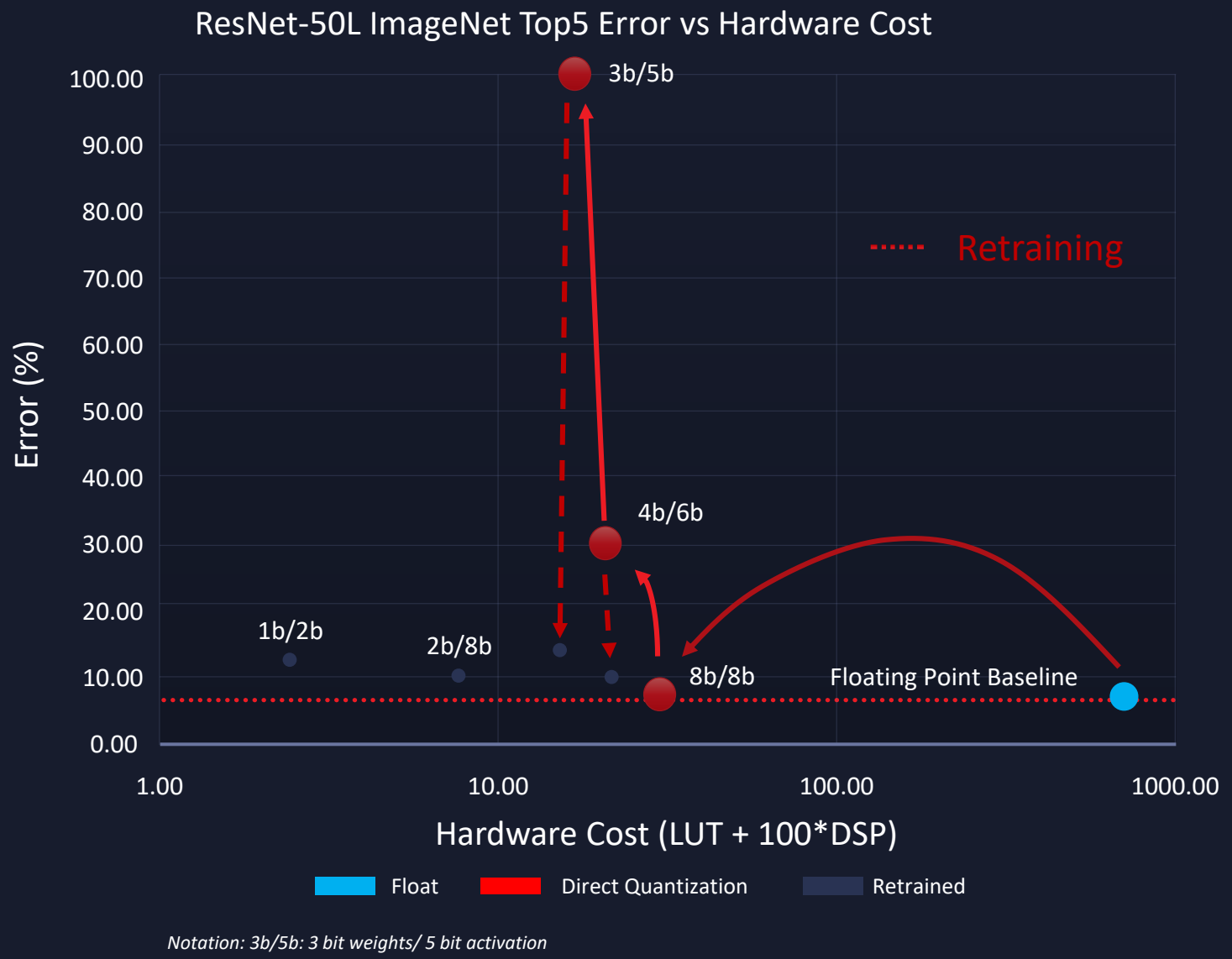
RELATIVE ENERGY COST

Operation:	Energy (pJ)
8b Add	0.03
16b Add	0.05
32b Add	0.1
16b FP Add	0.4
32b FP Add	0.9

Source: Bill Dally (Stanford), Cadence Embedded Neural Network Summit, February 1, 2017

Reduced Precision Arithmetic





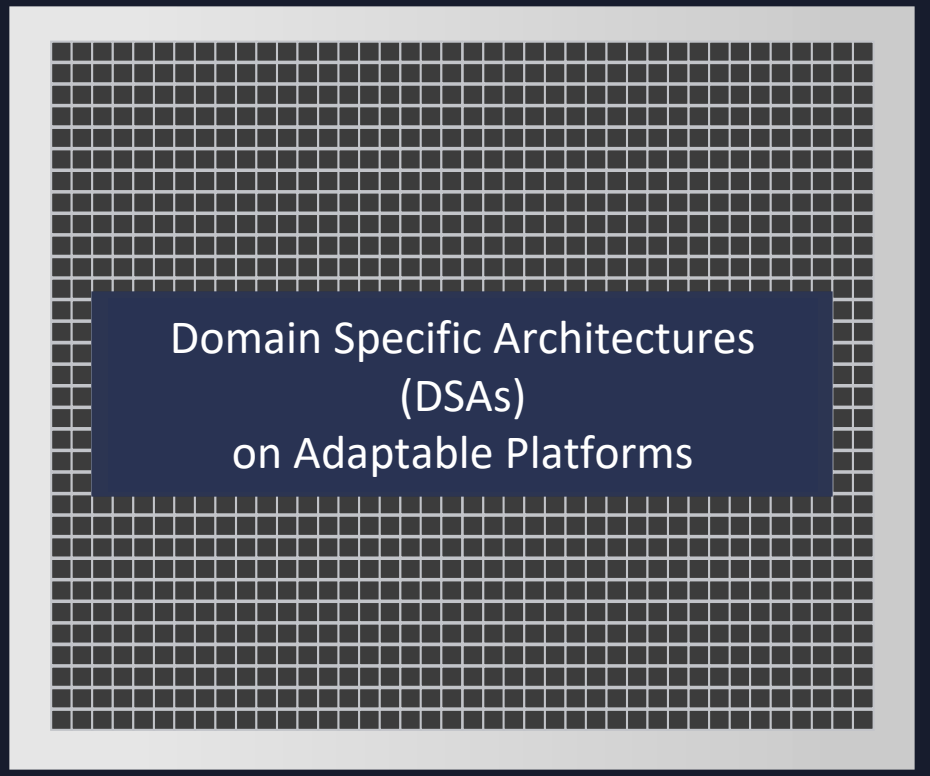
Notation: 3b/5b: 3 bit weights/ 5 bit activation

Need for Adaptable Hardware

Custom Data Flow

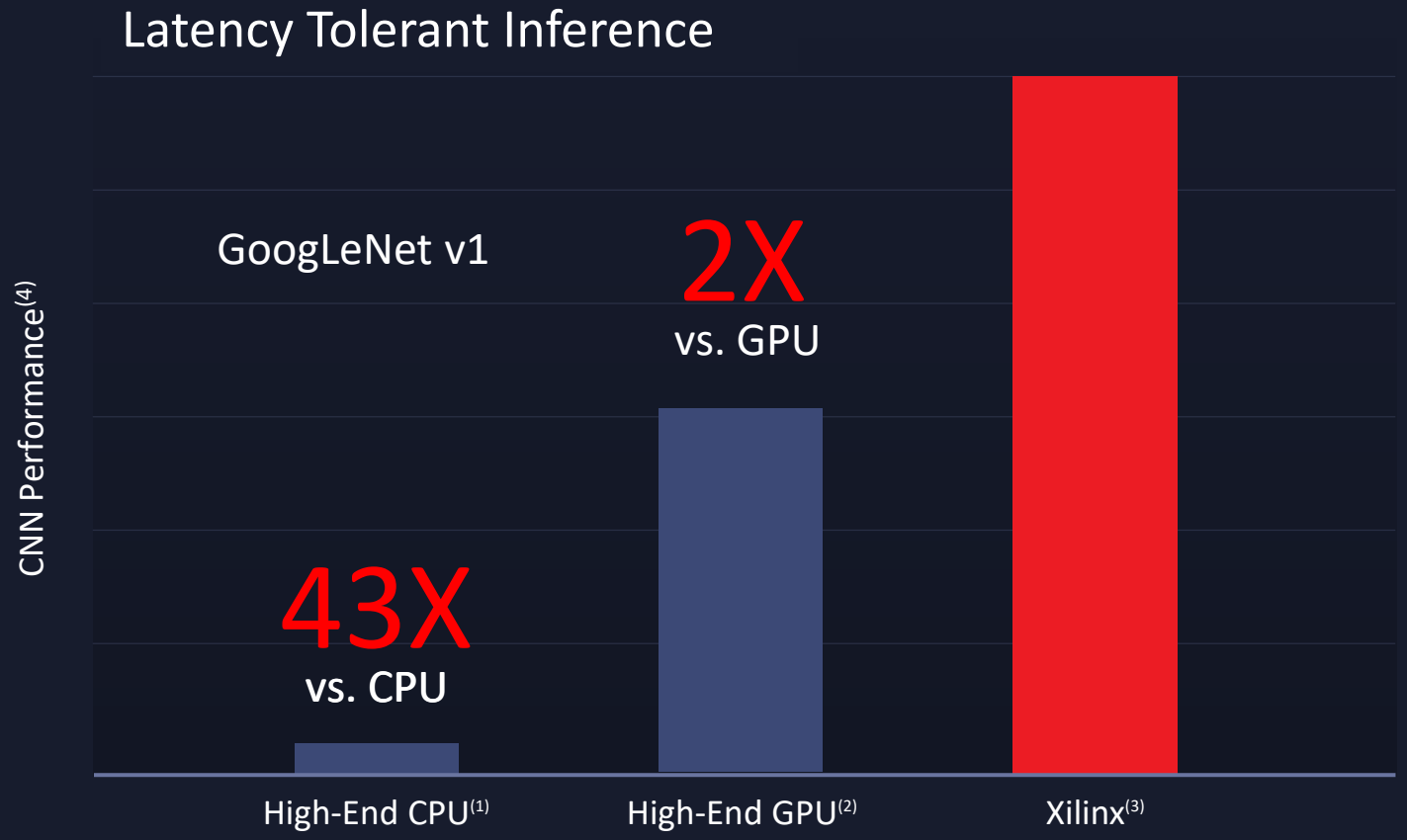
Custom Memory Hierarchy

Custom Precision



Low Latency

Xilinx's Unique Advantage



AI Inference Acceleration

Leveraging AI Engines

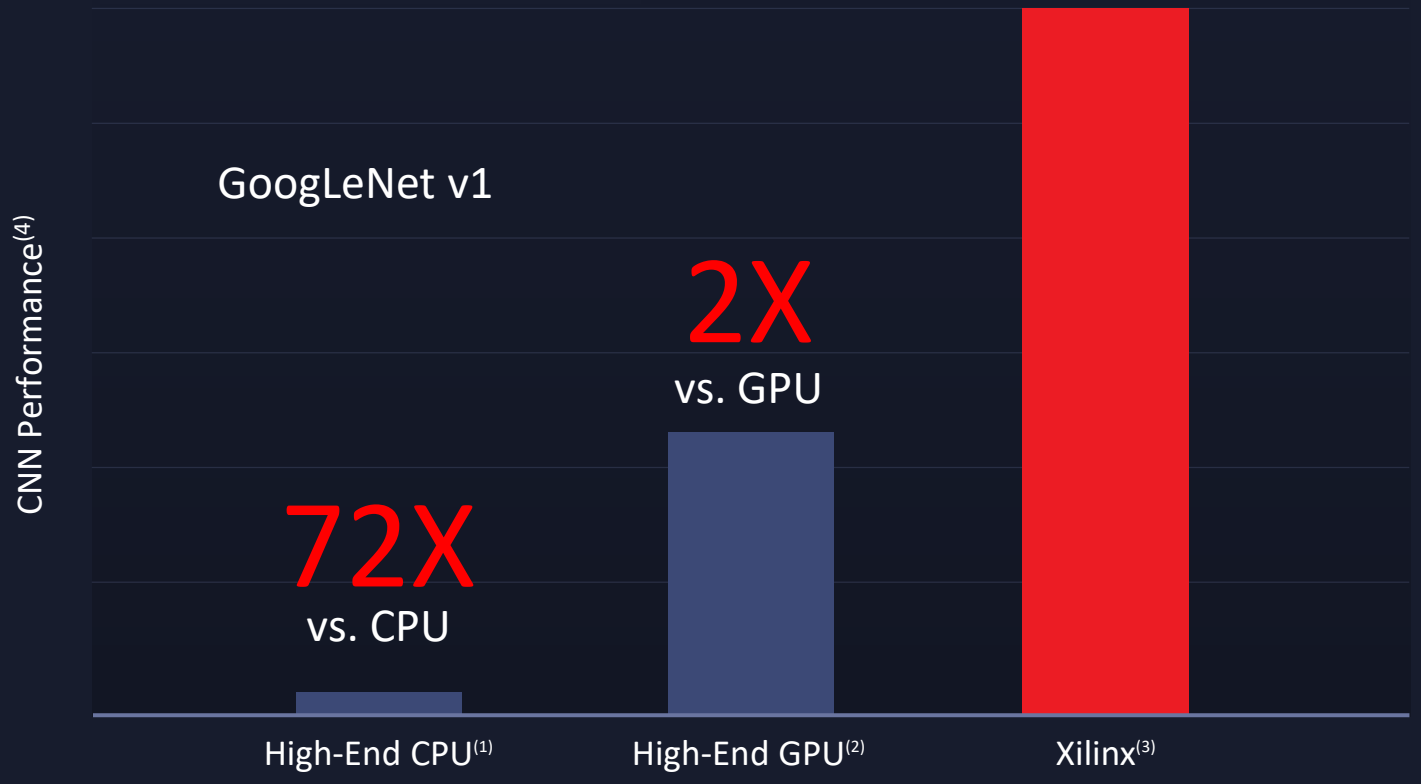
Majority of Adaptable & Scalar Engines Available for Whole App Acceleration

(1) *Measured on EC2 Xeon Platinum 8124 Skylake, c5.18xlarge AWS instance, Intel Caffe: https://github.com/intel/caffe*
(2) *V100 results taken from Oct 9th updates on www.Nvidia.com*
(3) *Versal Core Series*
(4) *GoogLeNet V1 throughput (Img/sec)*

Low Latency

Xilinx's Unique Advantage

Sub – 7ms Latency



AI Inference Acceleration

Leveraging AI Engines

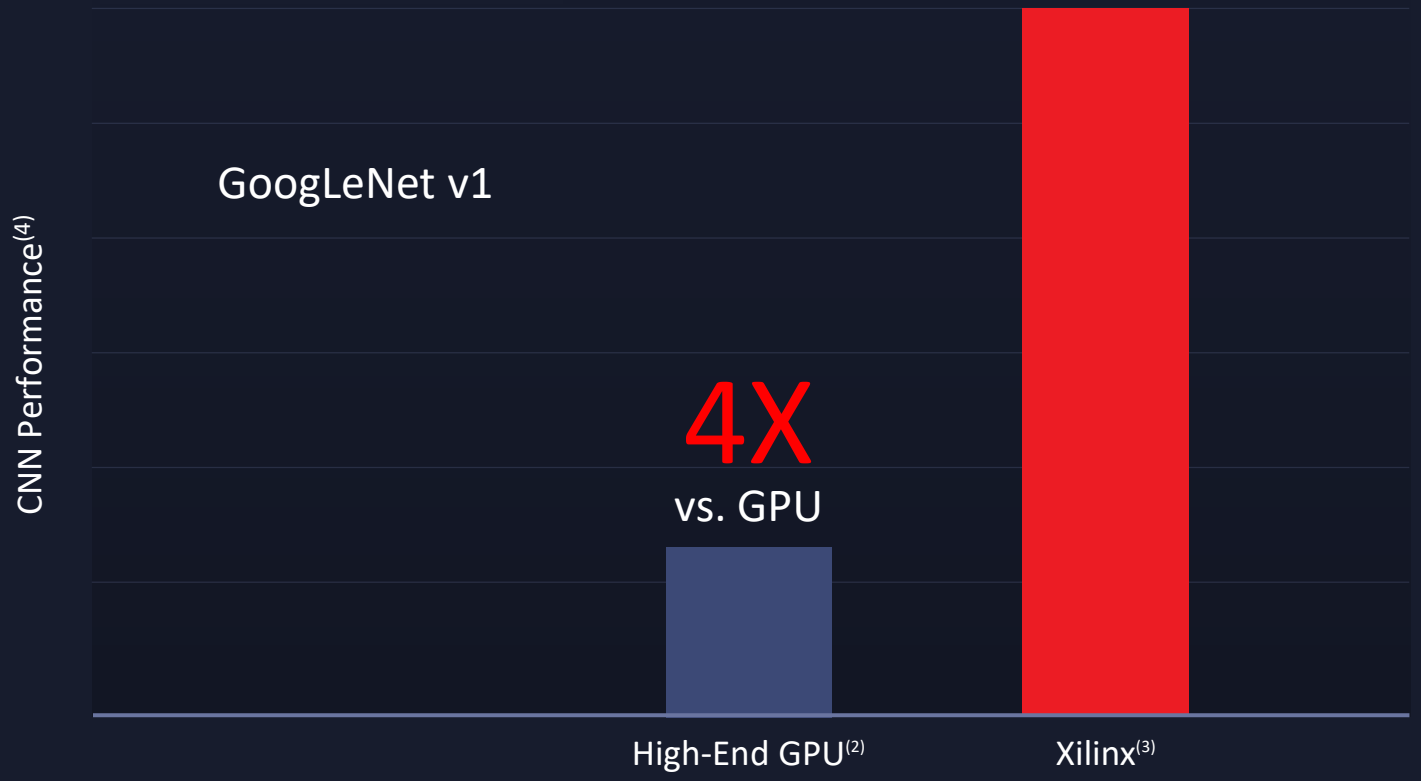
Majority of Adaptable & Scalar Engines Available for Whole App Acceleration

(1) *Measured on EC2 Xeon Platinum 8124 Skylake, c5.18xlarge AWS instance, Intel Caffe: https://github.com/intel/caffe*
(2) *V100 results taken from Oct 9th updates on www.Nvidia.com*
(3) *Versal Core Series*
(4) *GoogLeNet V1 throughput (Img/sec)*

Low Latency

Xilinx's Unique Advantage

Sub – 2ms Latency



AI Inference Acceleration

Leveraging AI Engines

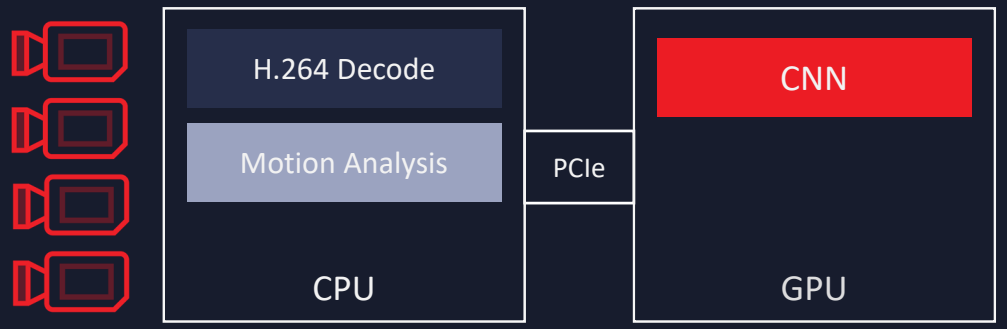
Majority of Adaptable & Scalar Engines Available for Whole App Acceleration

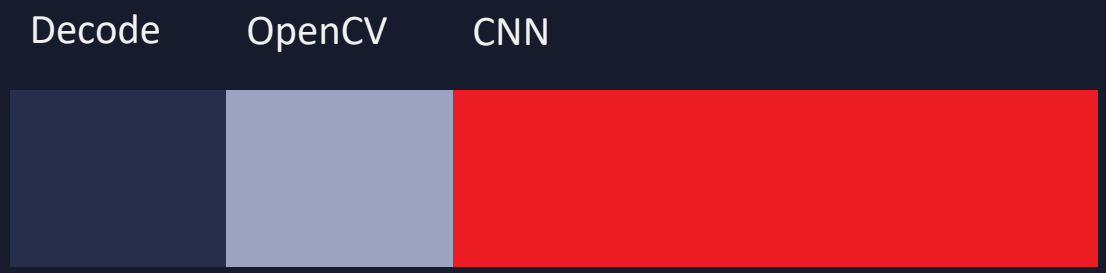
(1) *Measured on EC2 Xeon Platinum 8124 Skylake, c5.18xlarge AWS instance, Intel Caffe: https://github.com/intel/caffe*
(2) *V100 results taken from Oct 9th updates on www.Nvidia.com*
(3) *Versal Core Series*
(4) *GoogLeNet V1 throughput (Img/sec)*

Whole Application Acceleration

Intelligent Video Analytics

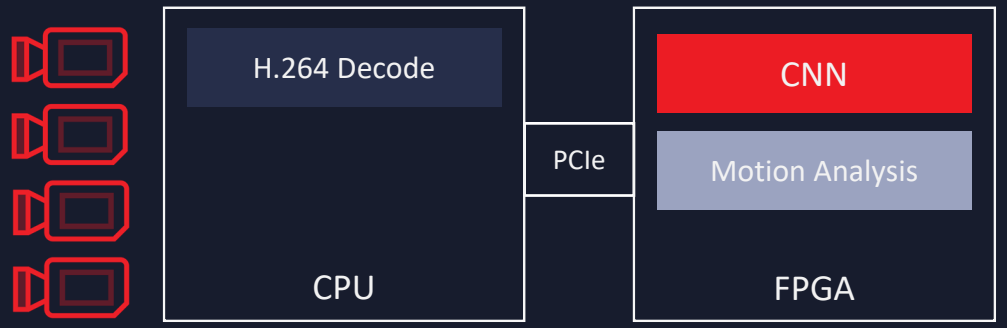
CPU-GPU

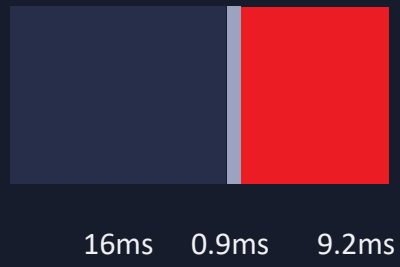




> Power: 75W > Latency: 82 ms > Throughput: 4x12 fps

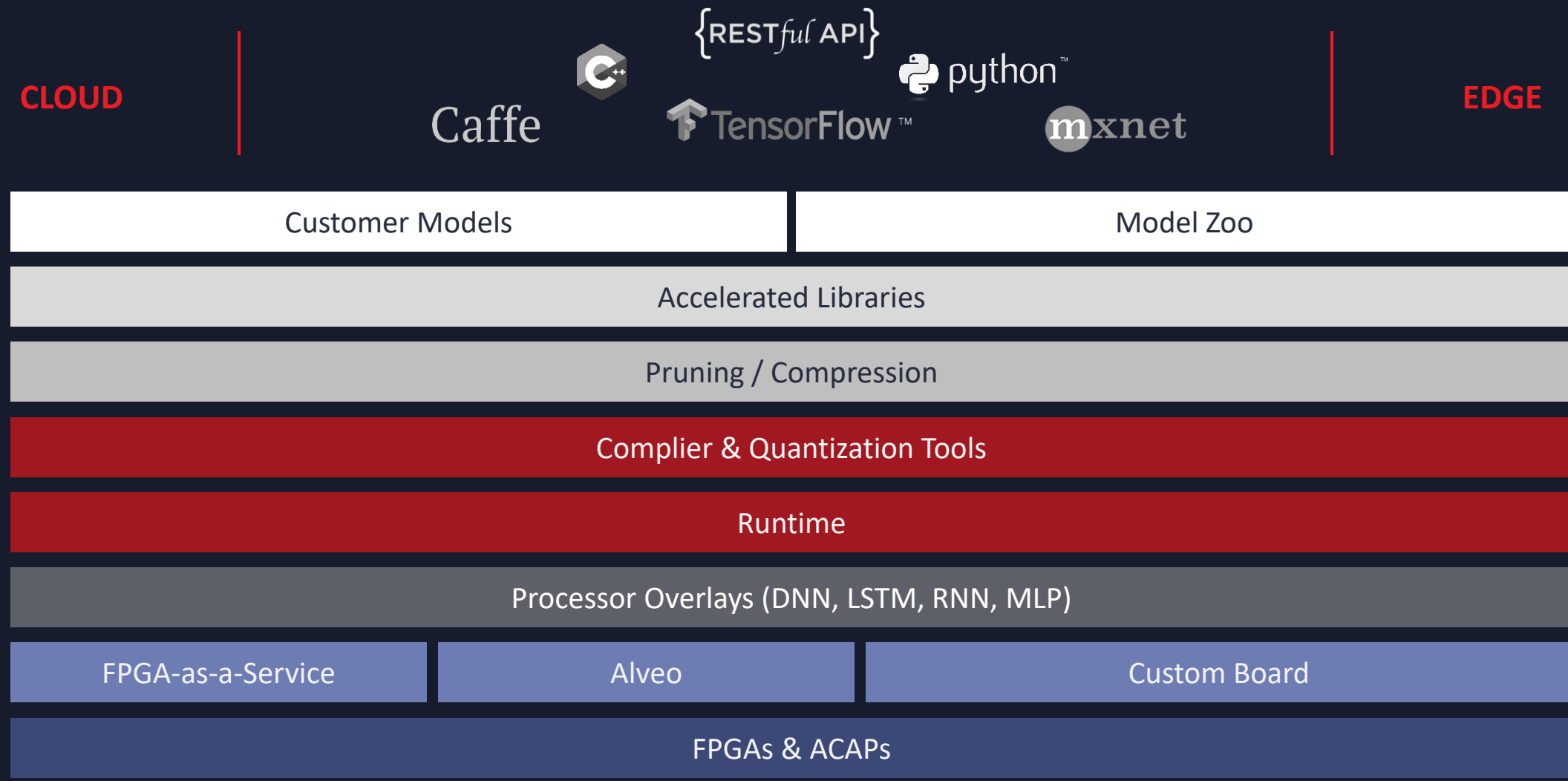
CPU-Xilinx FPGA





> Power: 50W > Latency: 26.1 ms > Throughput: 4x38 fps

Enabling the Development Community



Platforms for Every Developer

Data Scientists	Frameworks: Python, APIs	DEEPHI, Caffe, mxnet, FFMPEG, TensorFlow
SaaS Developers	FaaS Platform	aws, HUAWEI, Aliyun, NIMBIX
Application Developers	SDX: C++, OpenCL, Libraries, XRT open source runtime	Linux, freeRTOS, Xen
Embedded Developers	MPSoC Software Environment	
Hardware-Aware Software Developers	HLS: C++ IP Functions	
System Integrators	IP Integrator	
Hardware Developers	Vivado Design Suite: RTL Full Design	

XILINX
ALVEO™

Fast

Faster than CPUs & GPUs

Latency Advantage Over GPUs

Adaptable

Optimized for Any Workload

Adapt to Changing Algorithms

Accessible

Deploy in the Cloud or On-Premises

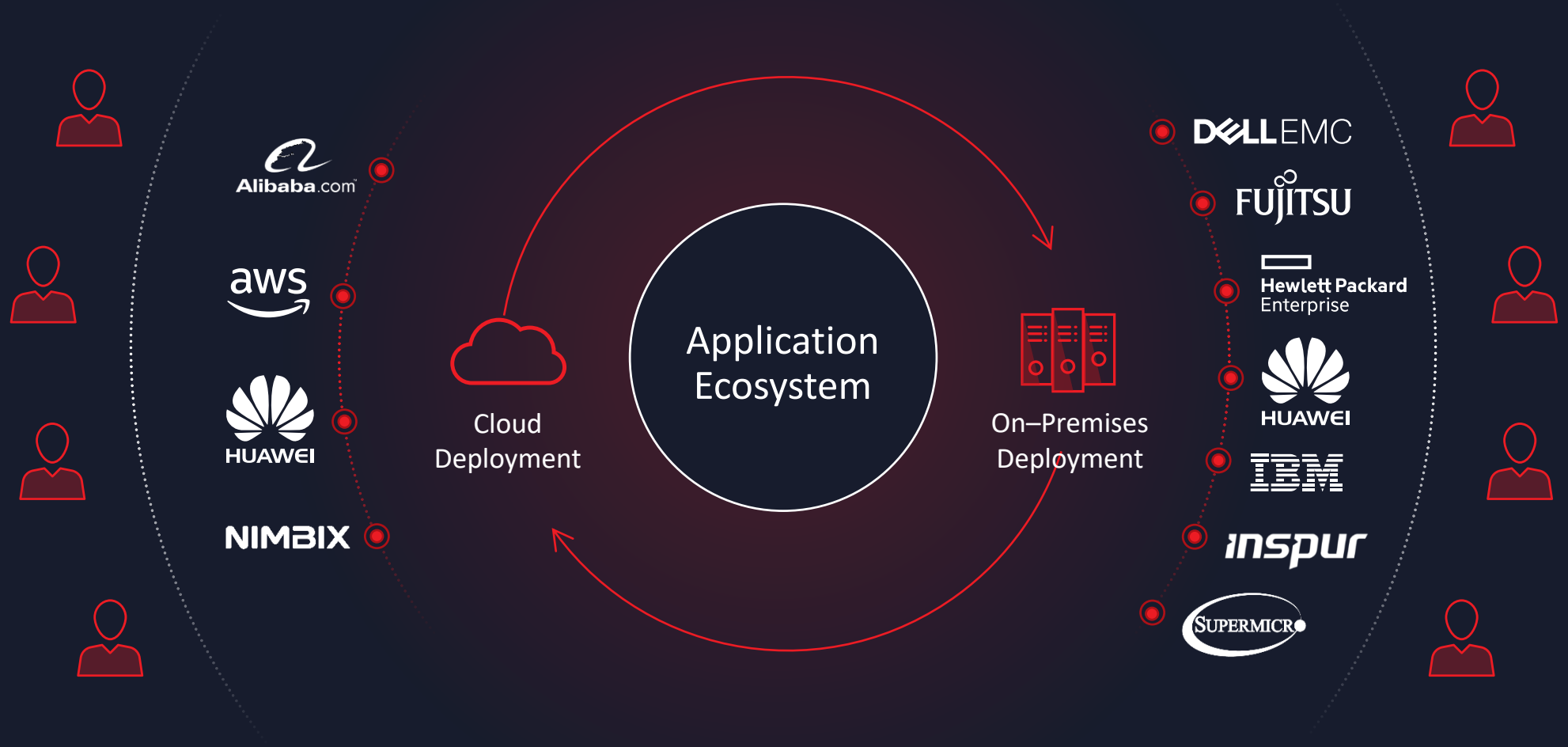
Applications Available Now

XILINX®
ALVEO™

Xilinx Acceleration Platform

TOOLS	USER
Frameworks	Data Scientists and AI Developers
Libraries, Compilers, Middleware	Application Developers
Firmware and Runtime	Software Developers
Integrated Development Environment	Hardware and Software Developers

Accessible: Cloud & On-Premise



Building the Adaptable, Intelligent World

XILINX