AI Acceleration

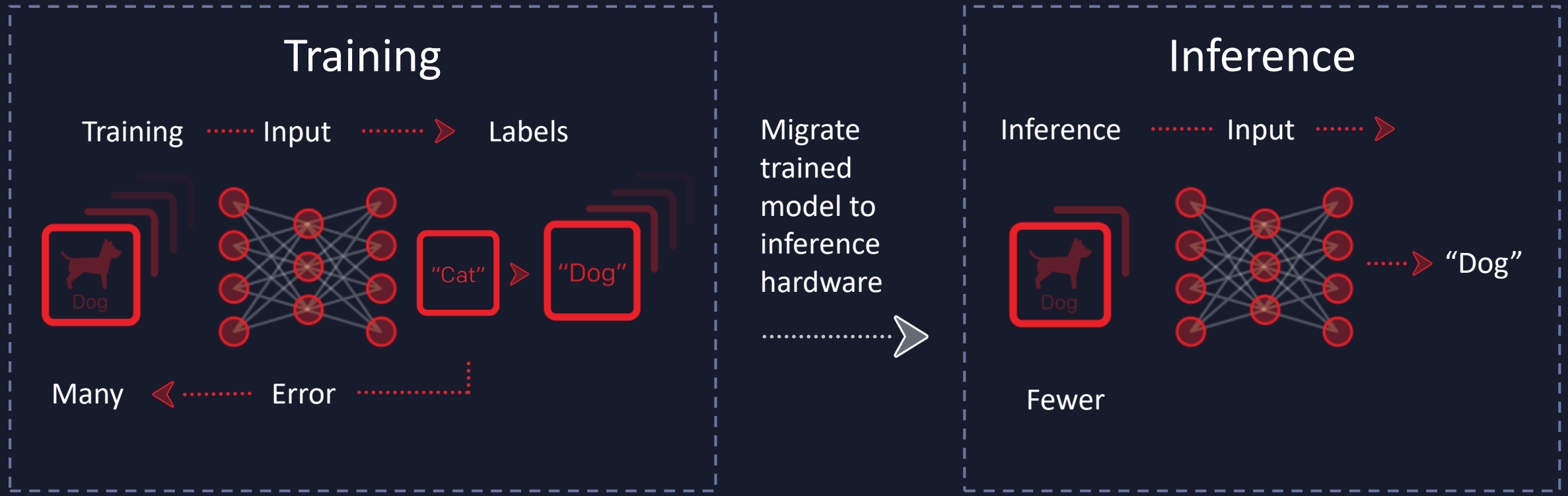
Salil Raje
Executive Vice President
Software and IP Products

XILINX

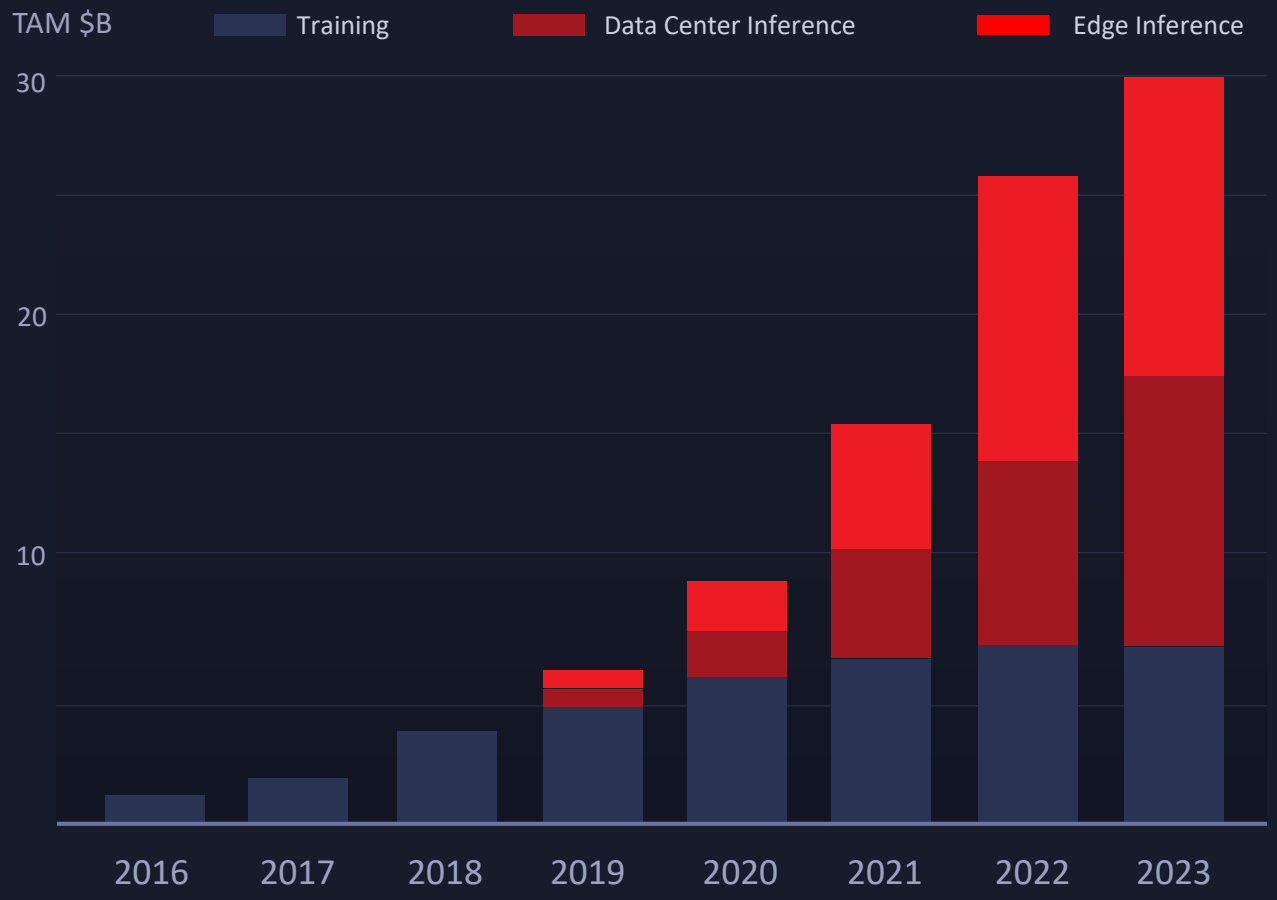
Welcome to All Developers!

Data scientists	Frameworks: Python, APIs	DEEPHi 深鉴科技, Caffe, mxnet, FFMPEG, TensorFlow
SaaS developers	FaaS Platform	aws, HUAWEI, Aliyun Alibaba Cloud Computing, NIMBIX
Application developers	SDX: C++, OpenCL, Libraries	Linux, freeRTOS, Xen
Embedded developers	Embedded Software: MPSoC	
Hardware-aware Software developers	HLS: C++ IP Functions	
System integrators	IP Integrator: System Integration	
Hardware developers	Vivado Design Suite: RTL Full Design	

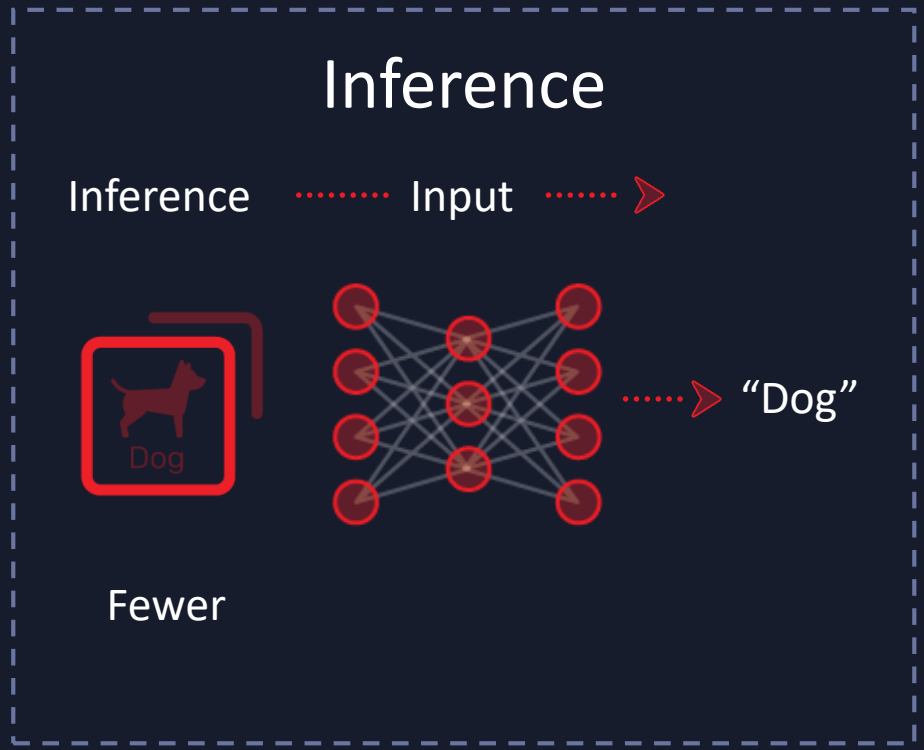
Training vs. Inference



Inference Projected Growth

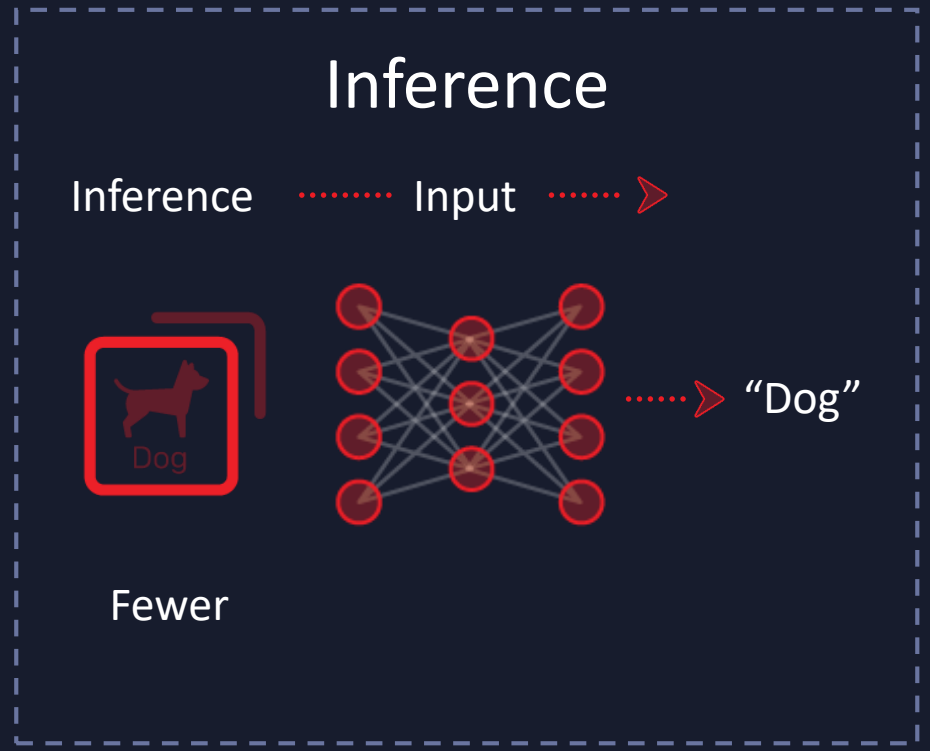


Barclays Research, Company Reports May 2018



Inference Challenges

- The rate of AI innovation
- Performance at low latency
- Low power consumption
- Whole app acceleration

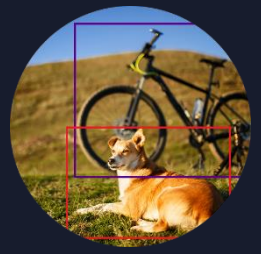


The Rate of AI Model Innovation

APPLICATIONS

Classification

Object Detection



Segmentation

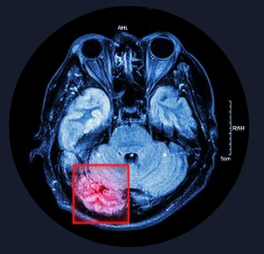
CNN

Speech Recognition

RNN, LSTM

Recommendation Engine

Anomaly Detection

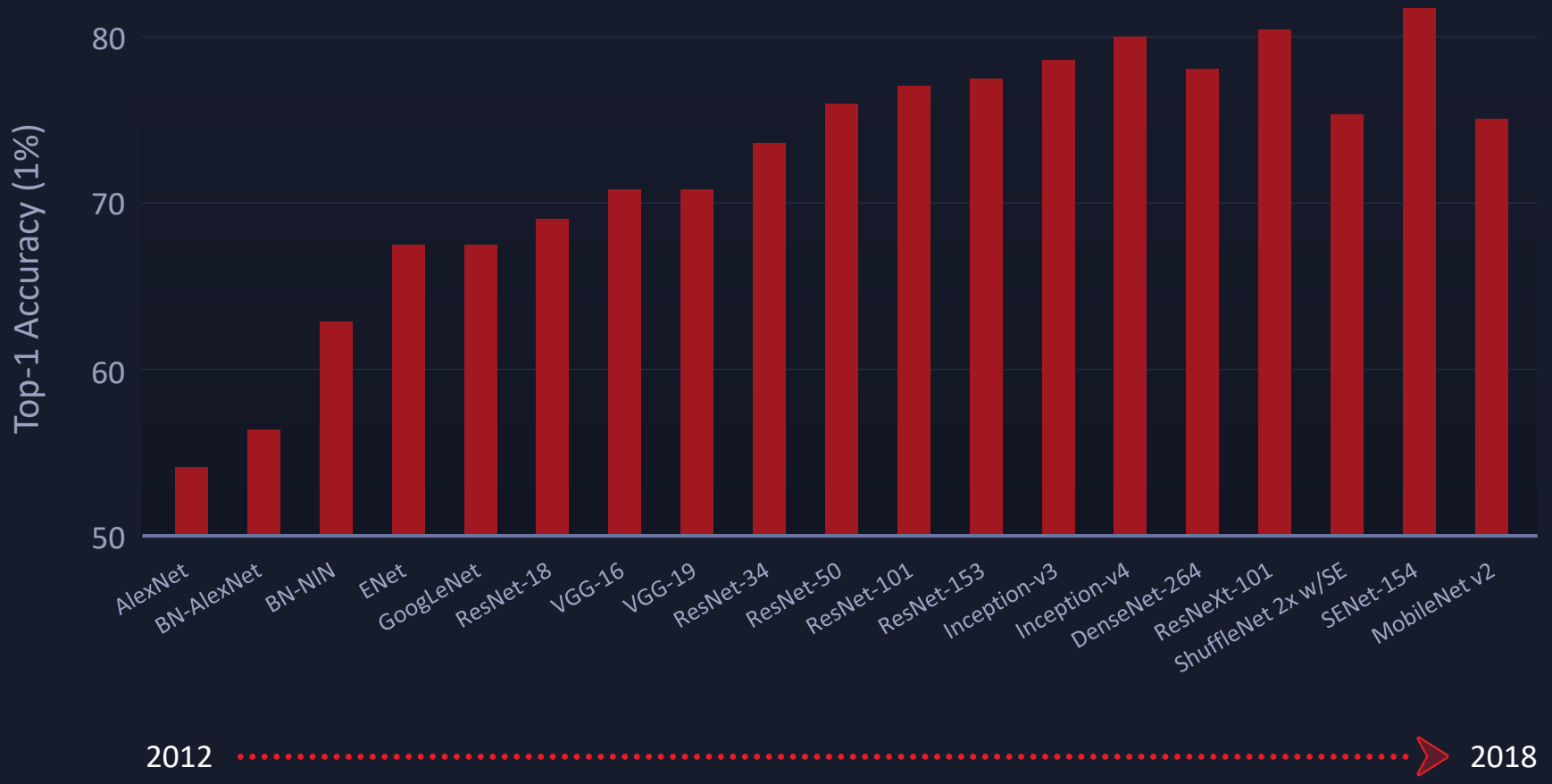


MLP

Diverse models over a broad range of applications

The Rate of AI Model Innovation: Classification

Classification

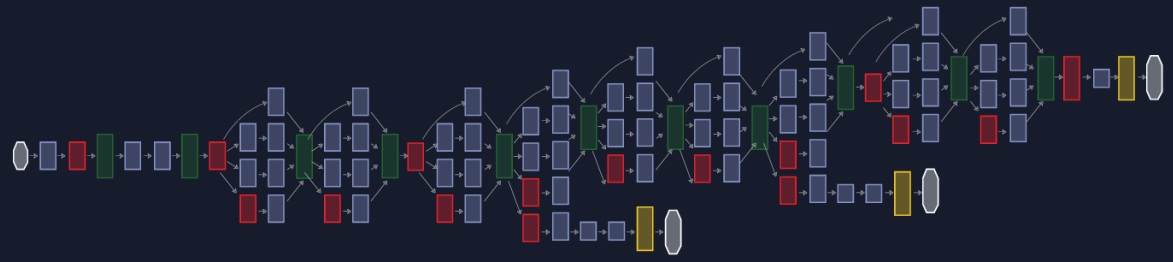


Source:
https://arxiv.org/pdf/1605.07678.pdf https://arxiv.org/pdf/1608.06993.pdf
https://arxiv.org/pdf/1709.01507.pdf https://arxiv.org/pdf/1611.05431.pdf

Network Complexity is Growing

AlexNet

GoogLeNet



DenseNet

Inference is Moving to Lower Precision

RELATIVE ENERGY COST

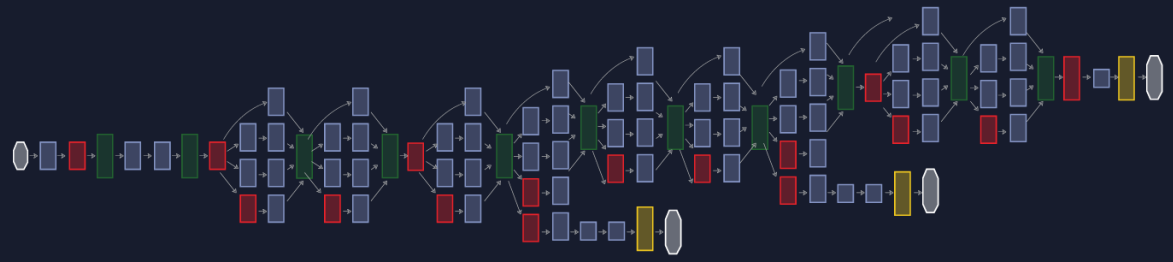
Operation:	Energy (pJ)
8b Add	0.03
16b Add	0.05
32b Add	0.1
16b FP Add	0.4
32b FP Add	0.9

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

Rate of Innovation Outpaces Silicon Cycles

AlexNet

GoogLeNet



DenseNet

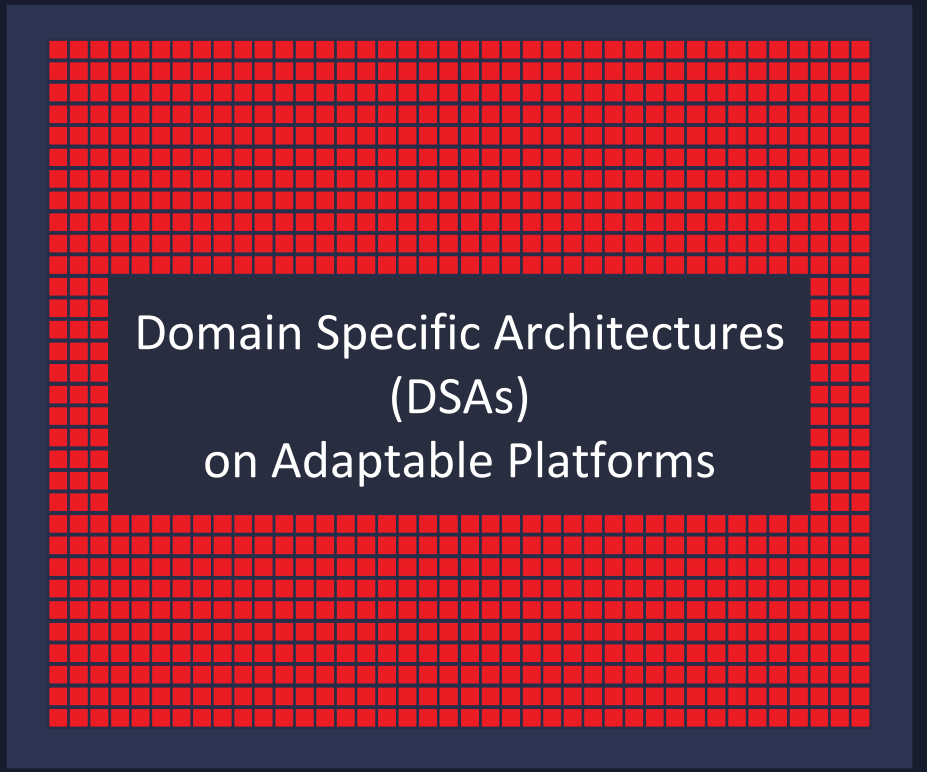
Silicon lifecycle

Only Adaptable Hardware Addresses Inference Challenges

Custom data flow

Custom memory hierarchy

Custom precision



Xilinx Acquires DeePhi

Custom data flow

Custom memory hierarchy

Custom precision

DEEPHI 深鉴科技 Now Part of XILINX

Pruning Quantization

Patented Compression Technology

- Reduces DL accelerator footprint
- Increases performance per watt

Example: DeePhi LSTM

Custom data flow

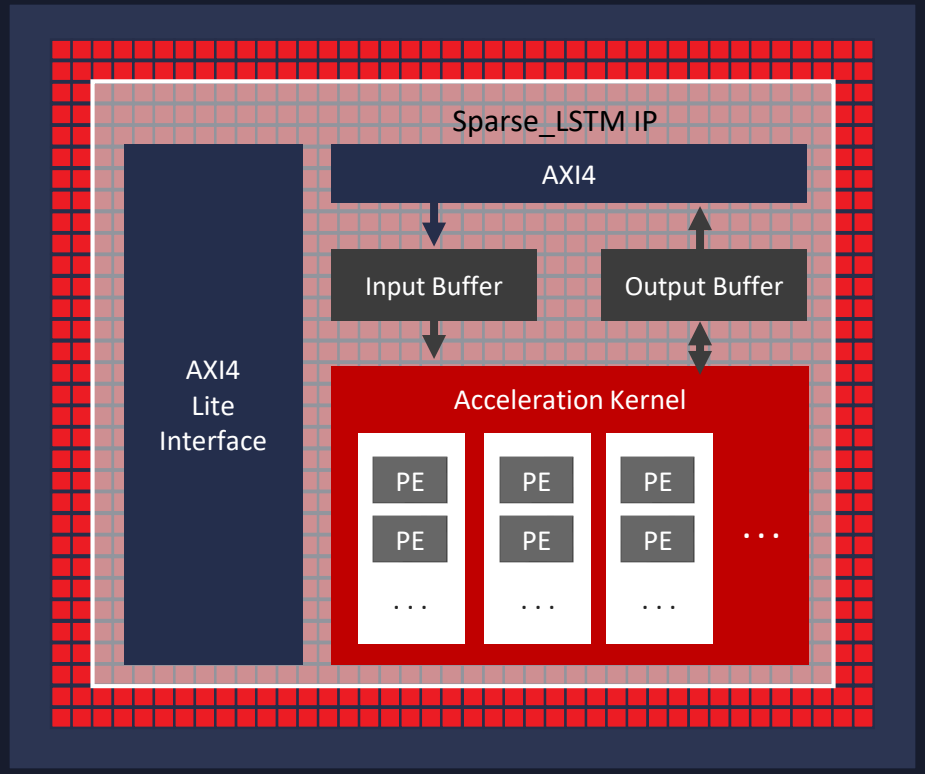
LSTM for speech recognition

Custom memory hierarchy

Sparse matrix implementation in memory

Custom precision

12 bit weights, 16 bit activations



Example: xDNN

Custom data flow

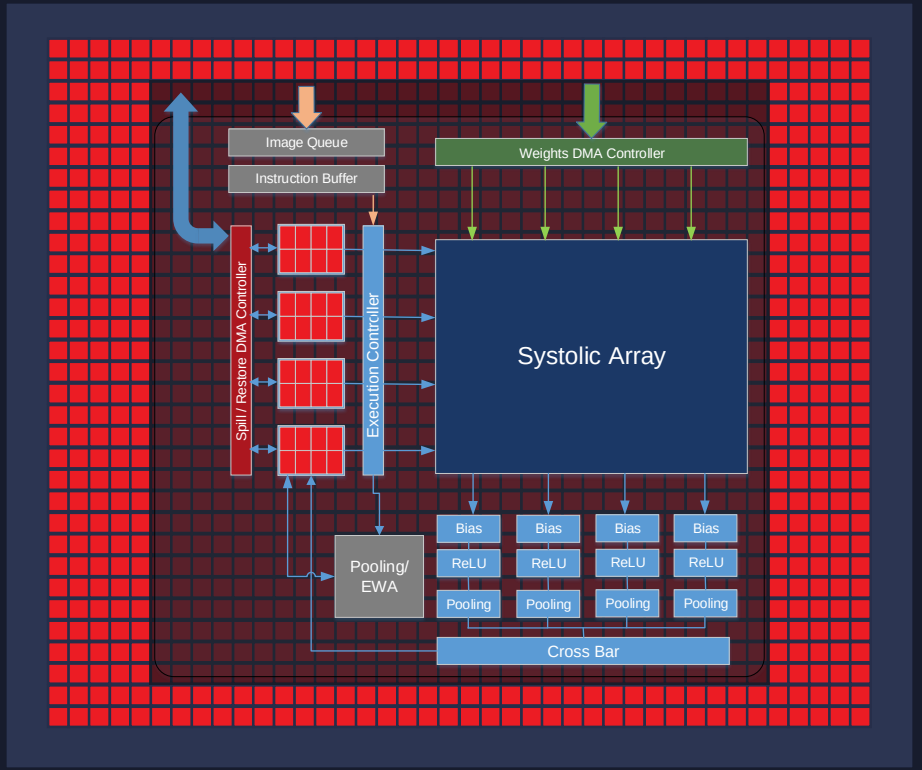
Optimized for latest CNN

Custom memory hierarchy

Optimized on-chip memory

Custom precision

Int8



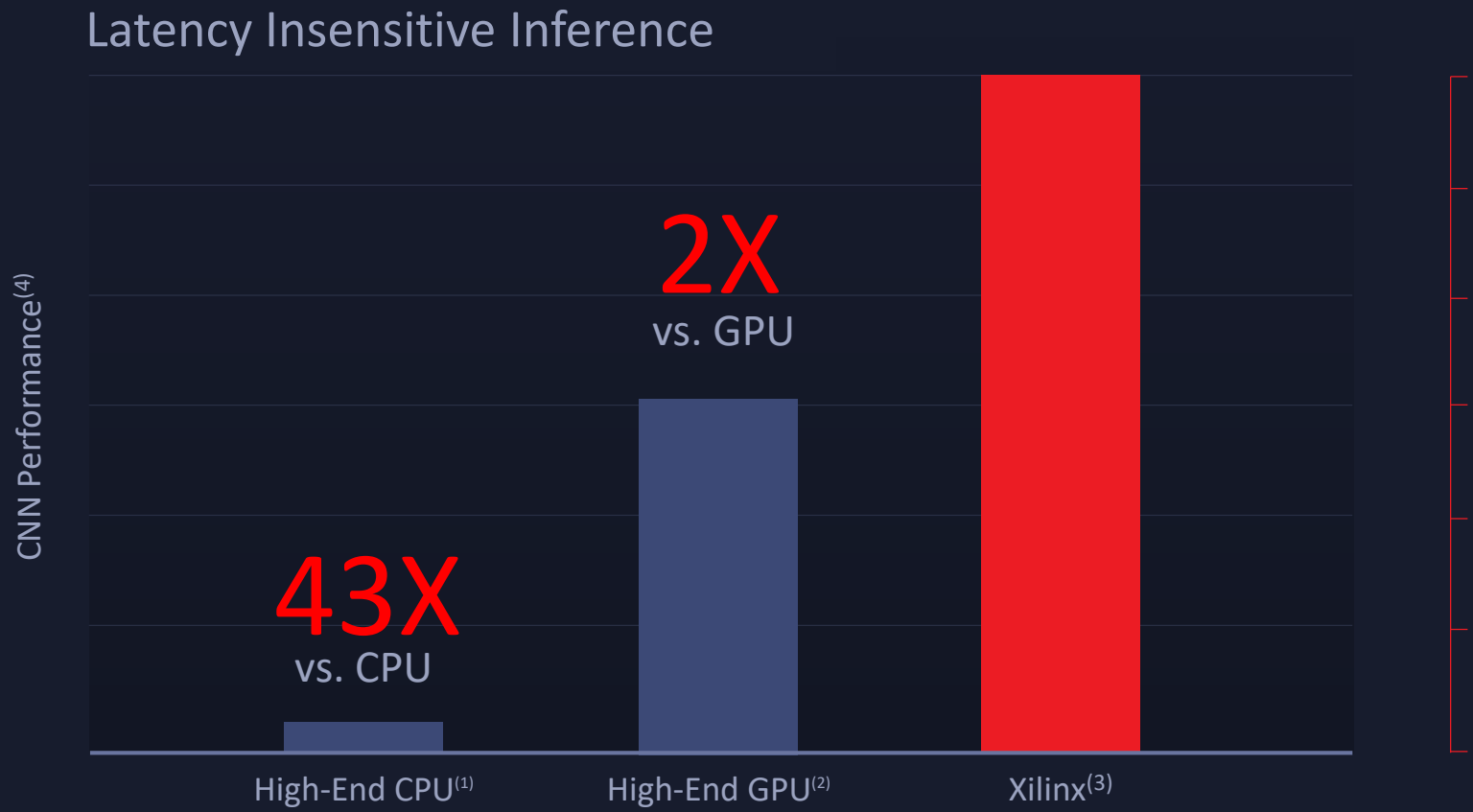
Low Latency is Critical for Inference

	CPU/GPU		FPGA/ACAP
Input 1		Result 1	
Input 2		Result 2	
Input 3		Result 3	
Input 4		Result 4	

High throughput OR low latency

High throughput AND low latency

Low Latency: Xilinx's Unique Advantage



AI Inference Acceleration

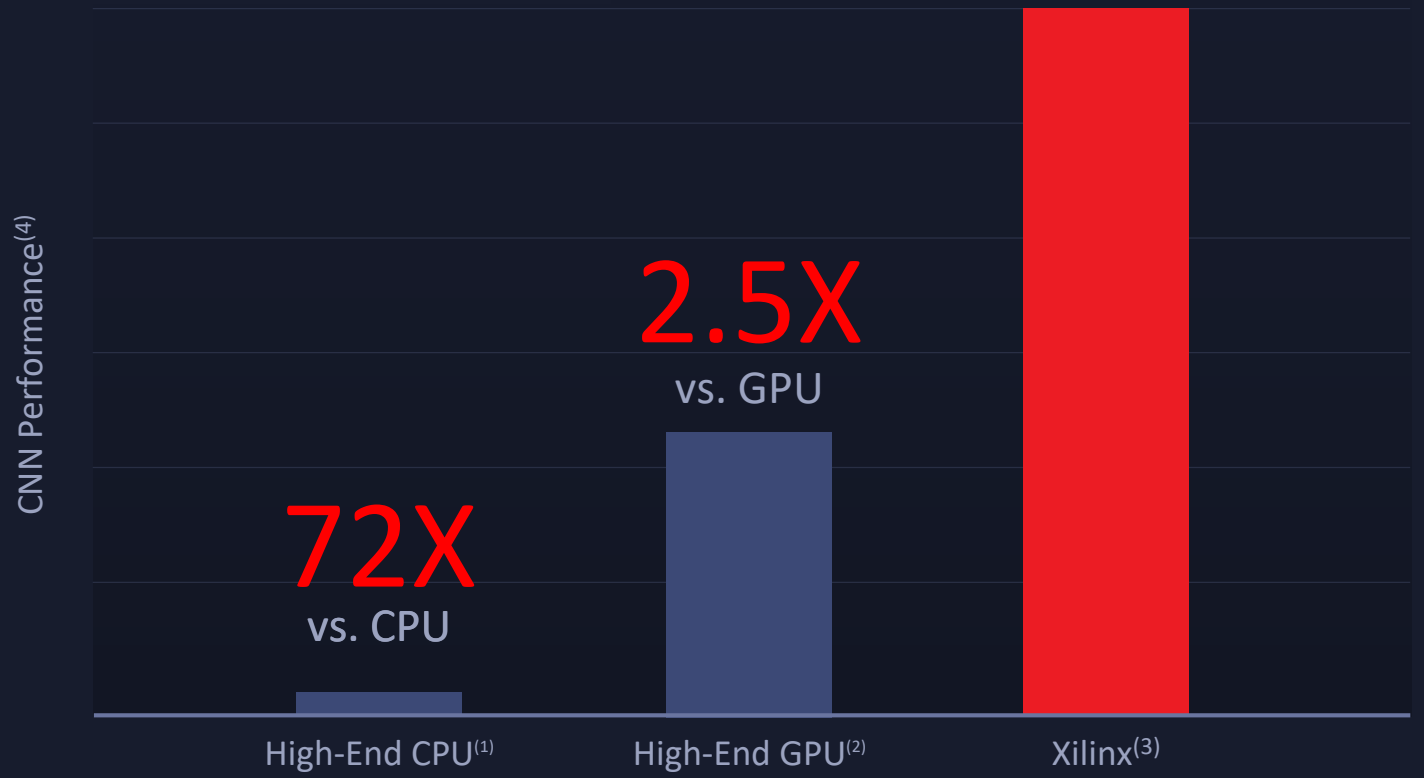
Leveraging AI Engines

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 numbers taken from Nvidia Technical Overview, "Deep Learning Platform, Giant Leaps in Performance and Efficiency for AI Services"
(3) Versal Core Series
(4) GoogLeNet V1 throughput (Img/sec)

Low Latency: Xilinx's Unique Advantage

Sub – 7ms Latency



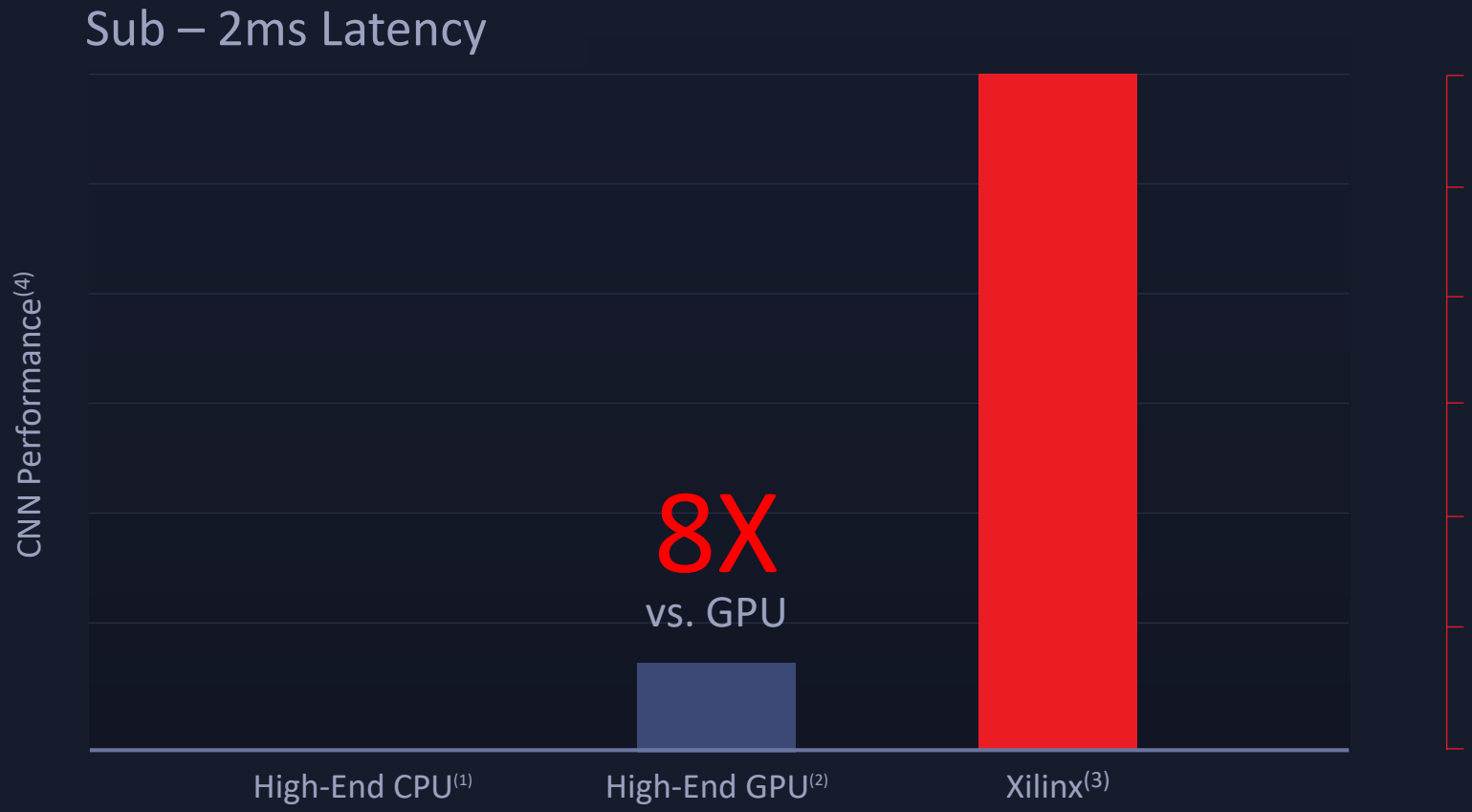
AI Inference Acceleration

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 numbers taken from Nvidia Technical Overview, "Deep Learning Platform, Giant Leaps in Performance and Efficiency for AI Services"
(3) Versal Core Series
(4) GoogLeNet V1 throughput (Img/sec)

Low Latency: Xilinx's Unique Advantage



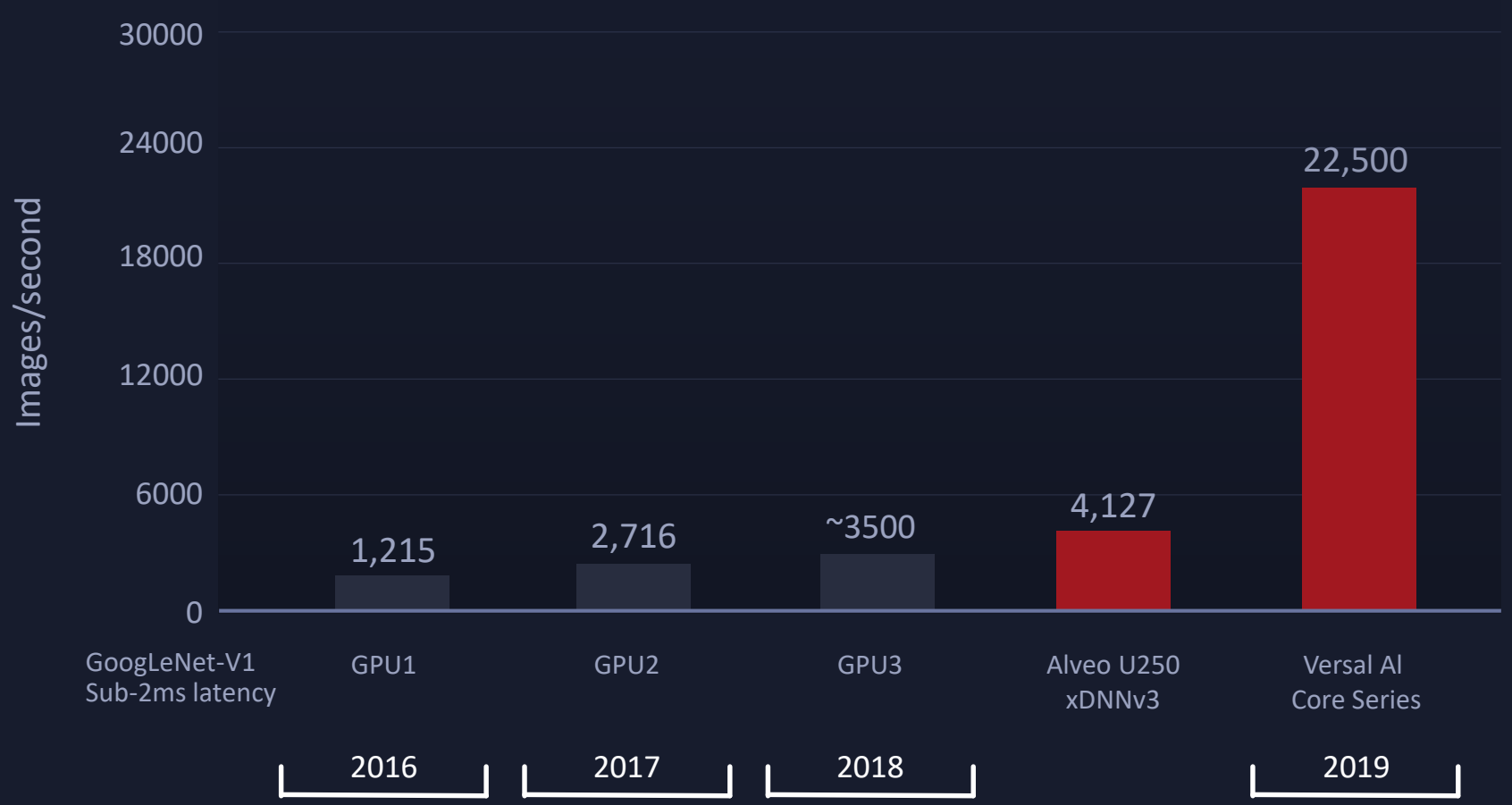
AI Inference Acceleration

Leveraging AI Engines

Majority of Adaptable & Scalar Engines available for Whole App Acceleration

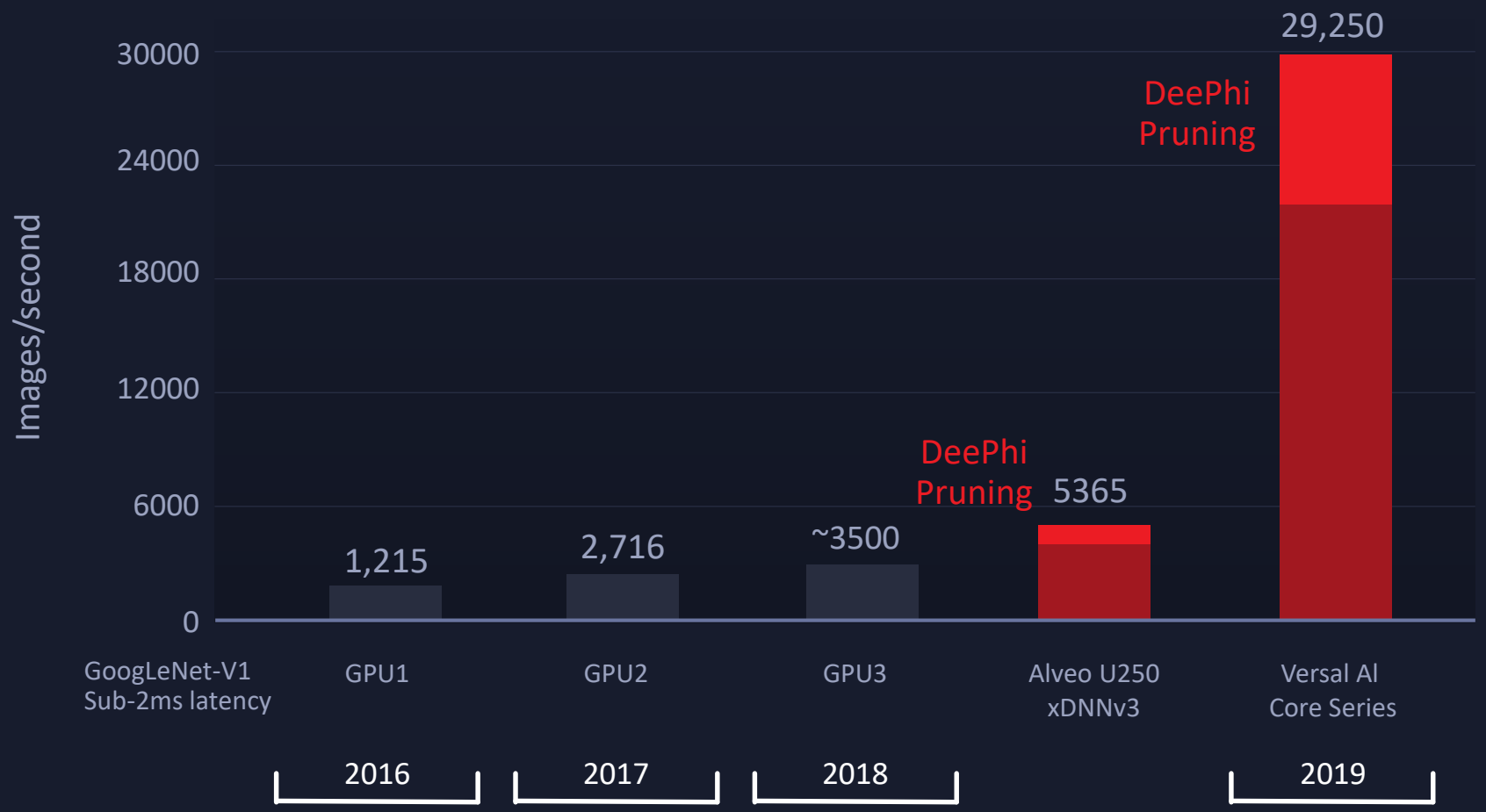
(1) Measured on EC2 Xeon Platinum 8124 Skylake, c5.18xlarge AWS instance, Intel Caffe: https://github.com/intel/caffe
(2) V100 numbers taken from Nvidia Technical Overview, "Deep Learning Platform, Giant Leaps in Performance and Efficiency for AI Services"
(3) Versal Core Series
(4) GoogLeNet V1 throughput (Img/sec)

Low-Latency CNN Inference Performance



Sources: Alveo - Published (INT8); Versal - Projected (INT8), 65% PL reserved for whole application; GPU 1 - P4 Published (INT8); GPU 2 - V100 Published (FP16/FP32); GPU 3 - T4 Projected

Low-Latency CNN Inference Performance



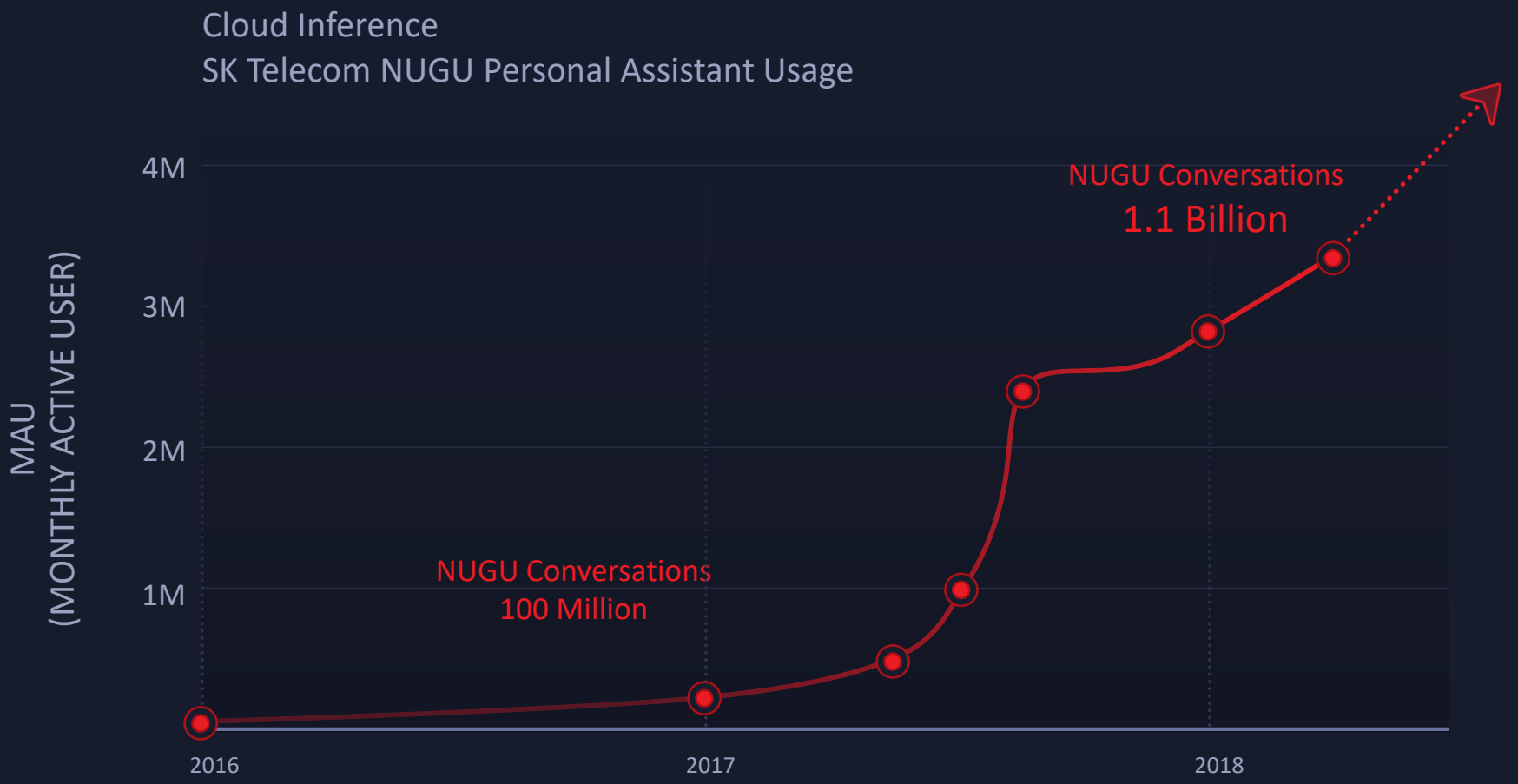
DeePhi Pruning Technology

1.3x-8x

Performance improvement based on the network

Sources: Alveo - Published (INT8); Versal - Projected (INT8), 65% PL reserved for whole application; GPU 1 - P4 Published (INT8); GPU 2 - V100 Published (FP16/FP32); GPU 3 - T4 Projected

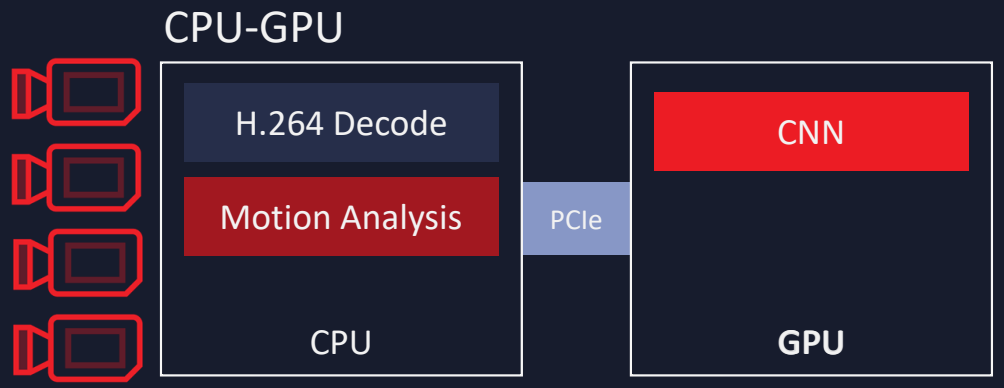
Power Is Critical for Inference Applications

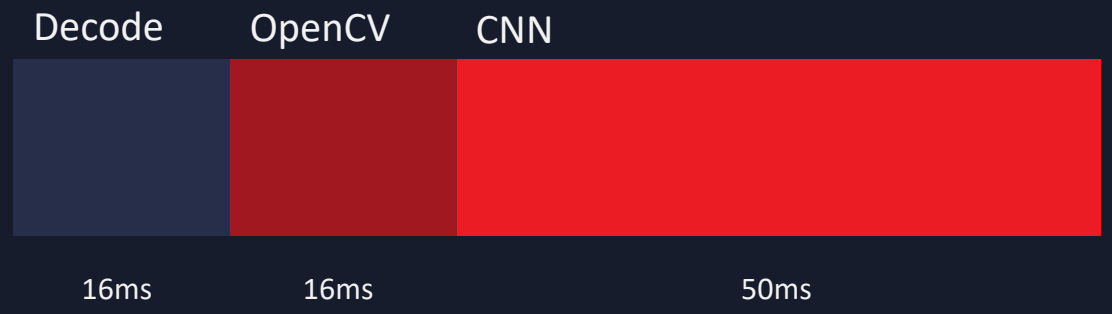


16x

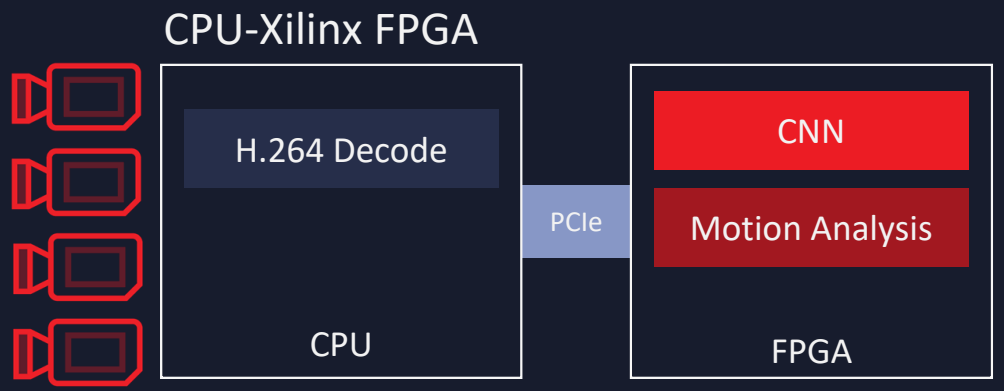
Perf/watt vs. GPU

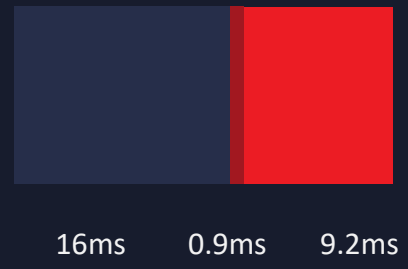
Whole Application Acceleration: Smart City / Security





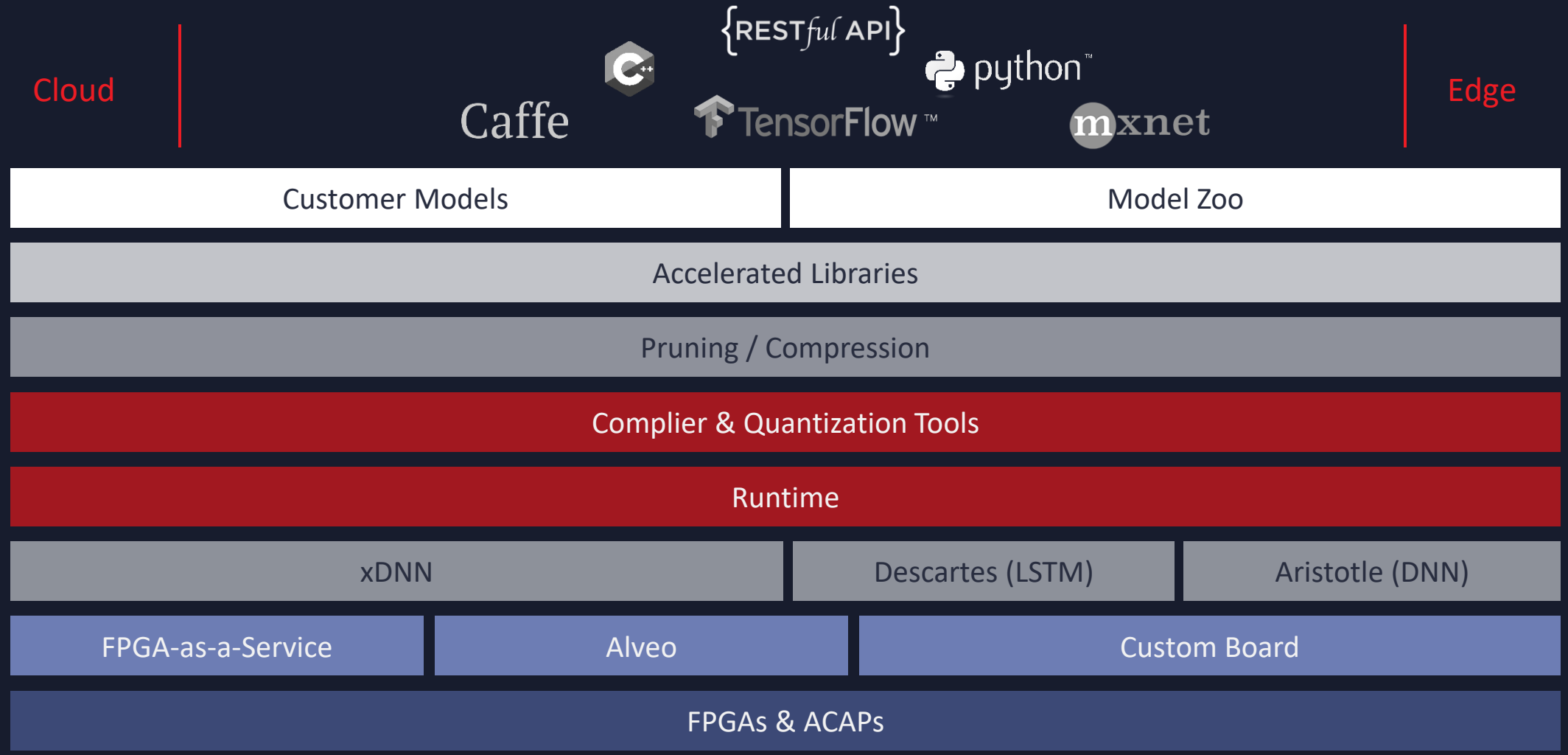
> Power: 75W > Latency: 82 ms > Throughput: 4x12 fps





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

Enabling the Development Community



IN SUMMARY

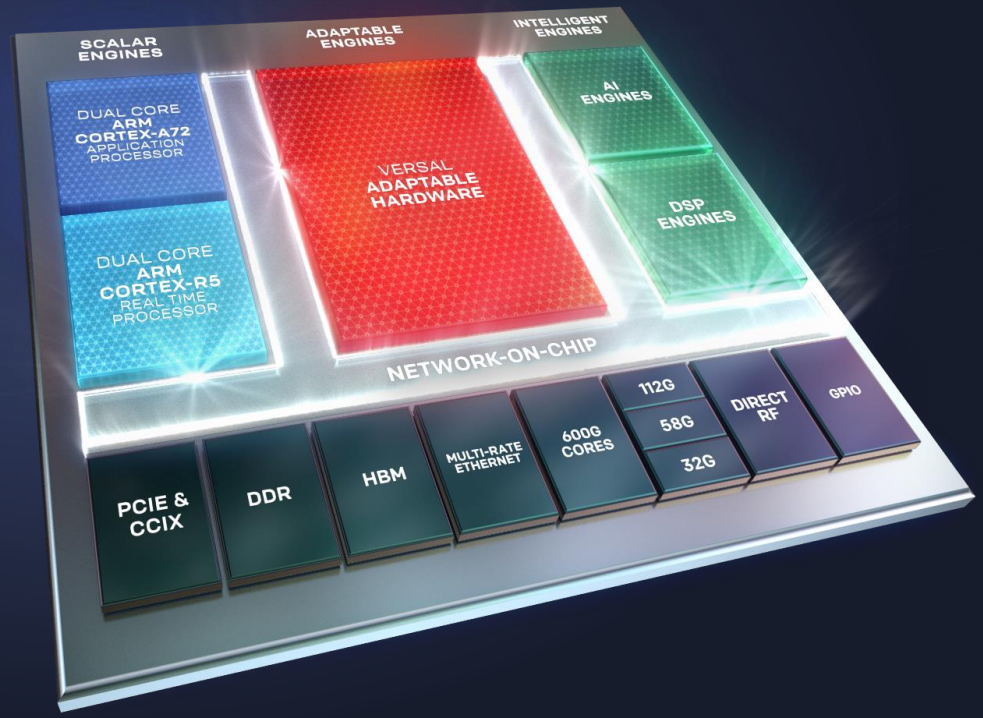
Only Xilinx Adaptable Devices Can:

Match the speed of AI innovation

Give the best performance at low latency

Give the best power results

Accelerate the whole application



Xilinx

Building the Adaptable, Intelligent World