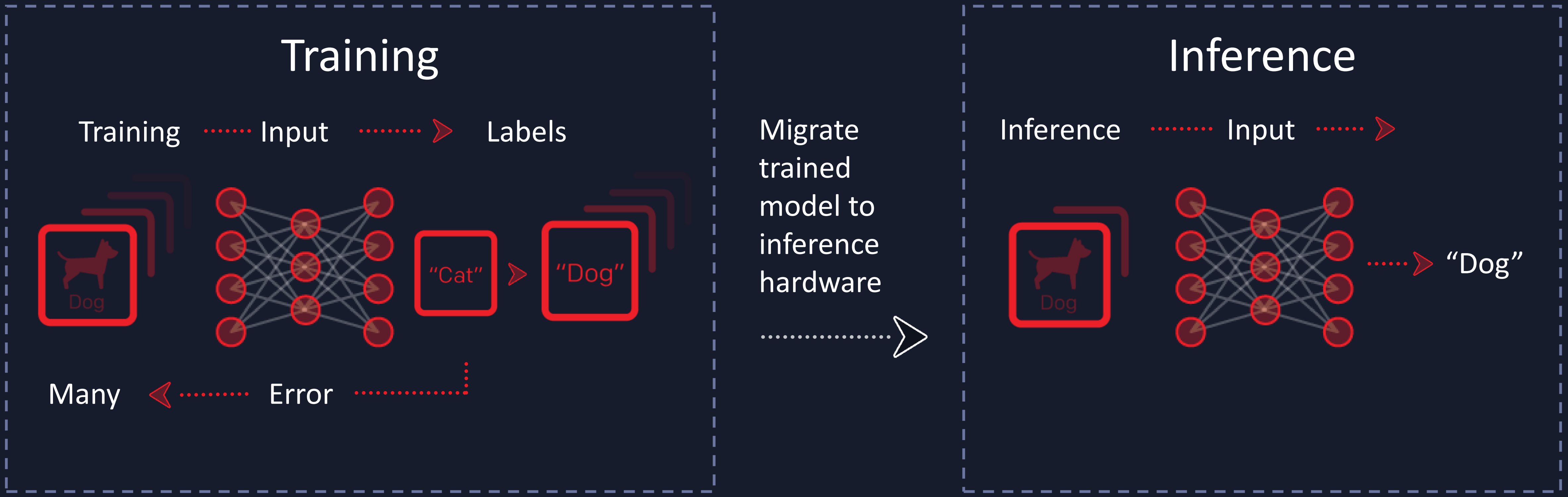
AI Acceleration

Shaun Purvis

Training vs. Inference



Training

Training Input Labels

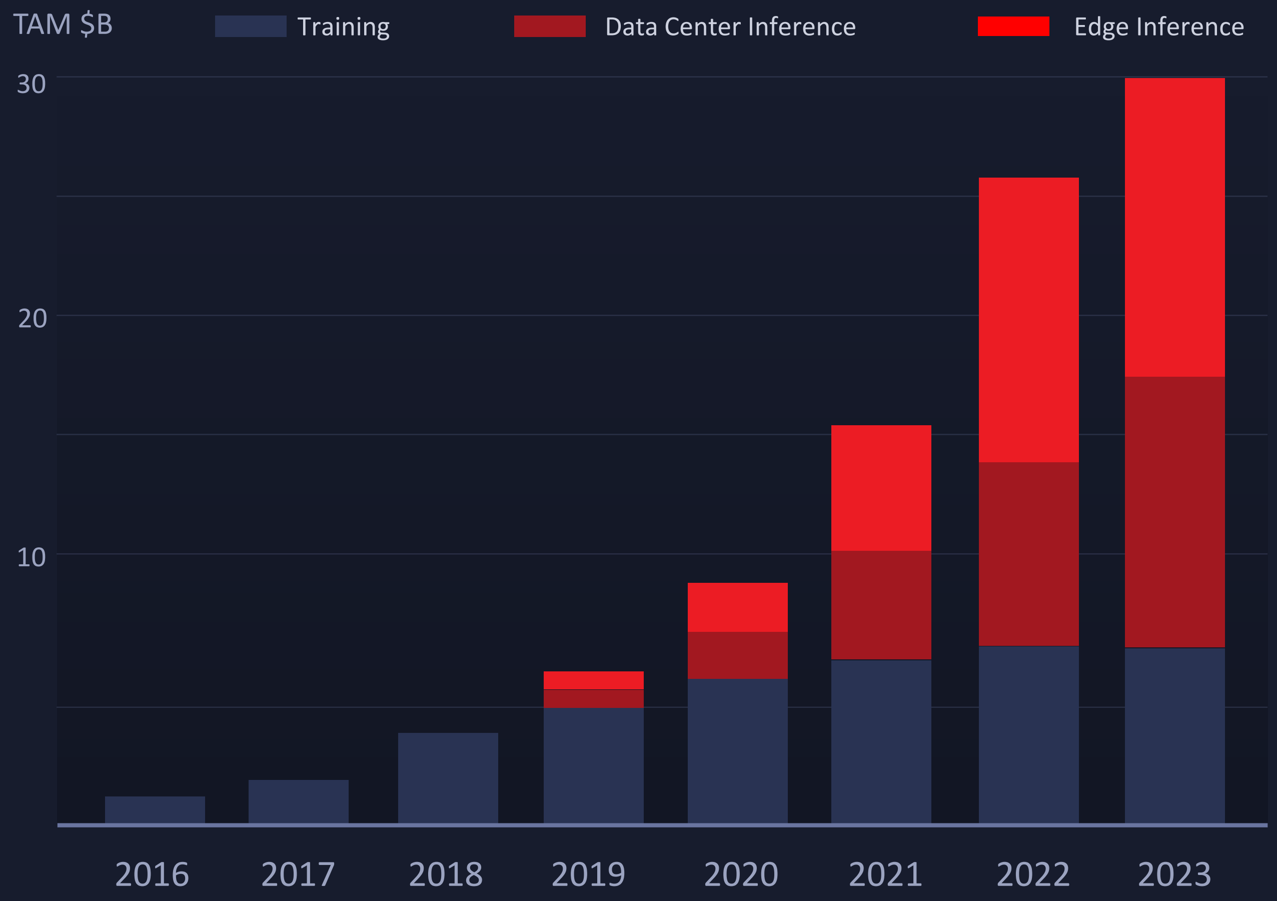
Many Error

Migrate trained model to inference hardware

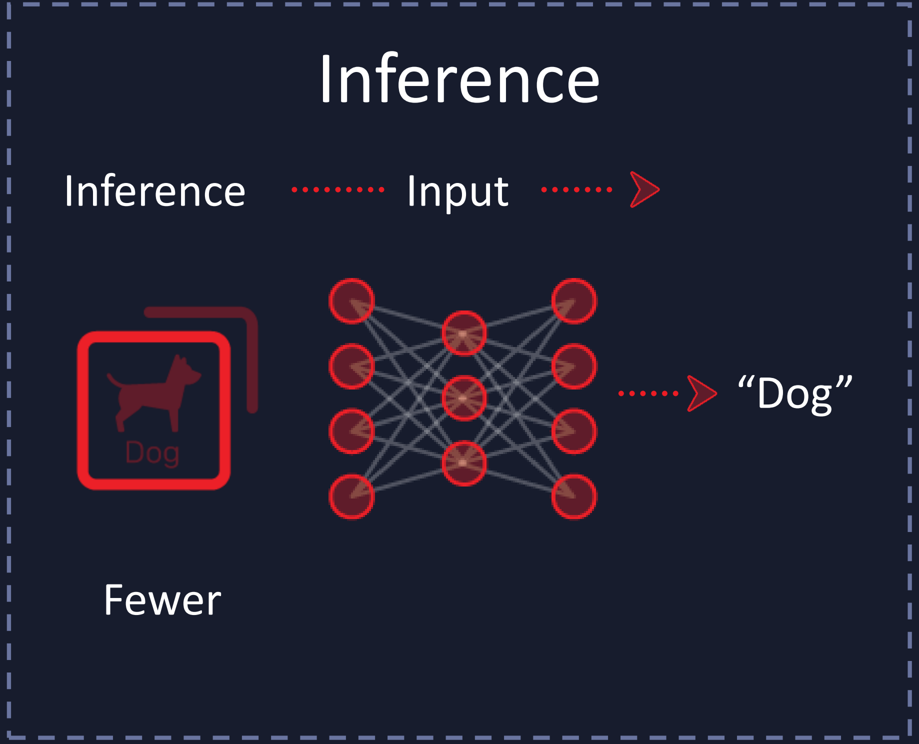
Inference

Inference Input

Inference Projected Growth



Barclays Research, Company Reports May 2018



Inference Space

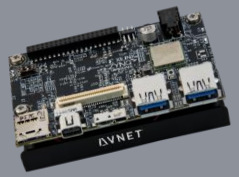
Cloud / Data-Center / On-Premise

FaaS

XILINX VERSAL™

Edge / Embedded / IoT

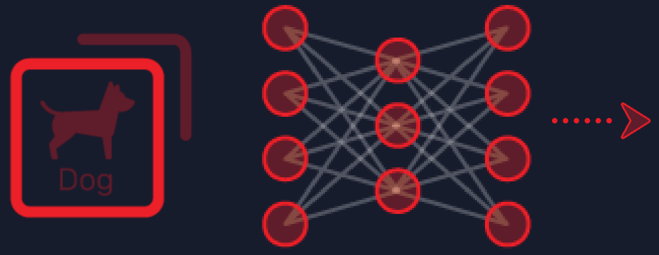
ZYNQ



XILINX VERSAL™

Inference

Inference ······ Input ······>

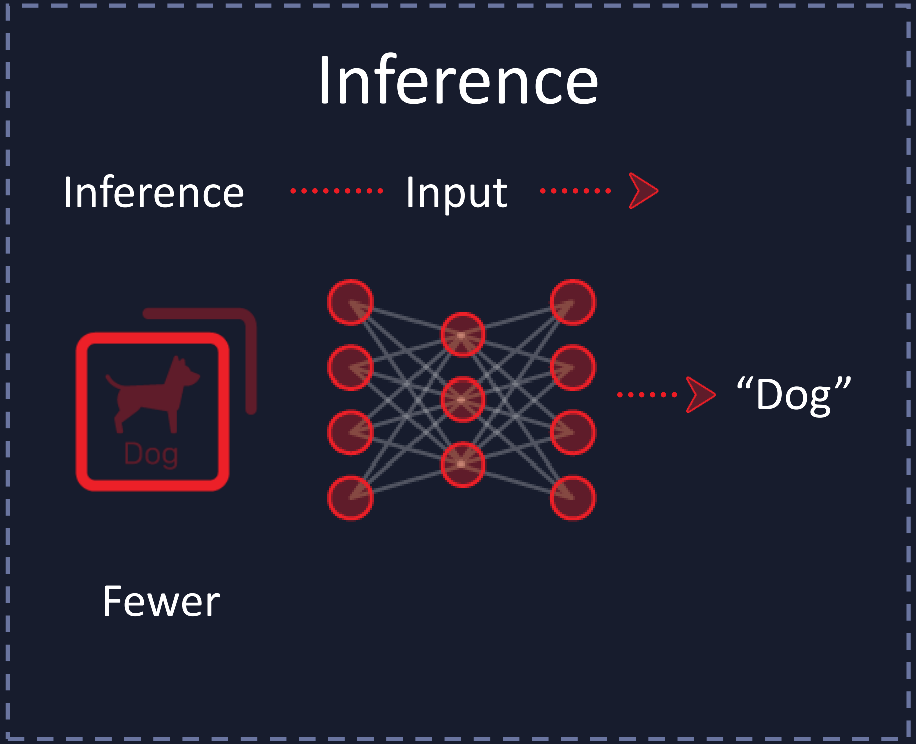


······> "Dog"

Fewer

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

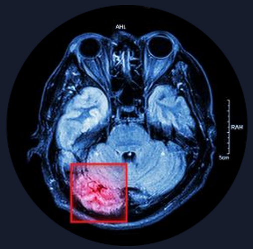
Object Detection

Segmentation

Speech Recognition

Recommendation Engine

Anomaly Detection



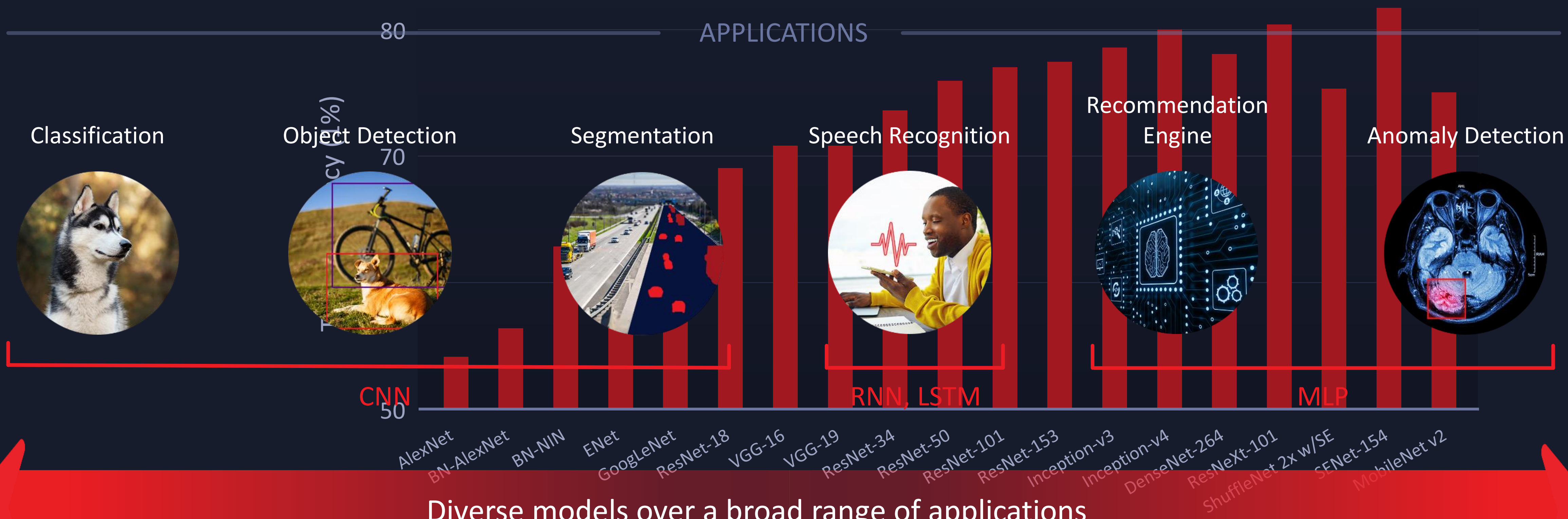
CNN

RNN, LSTM

MLP

Diverse models over a broad range of applications

The Rate of AI Model Innovation: Classification



APPLICATIONS

Classification

Object Detection

Segmentation

Speech Recognition

Recommendation Engine

Anomaly Detection

CNN

RNN, LSTM

MLP

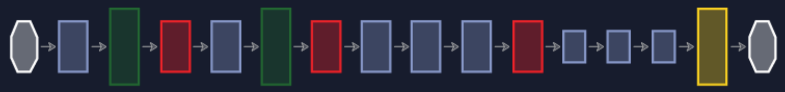
Diverse models over a broad range of applications

2012 → 2018

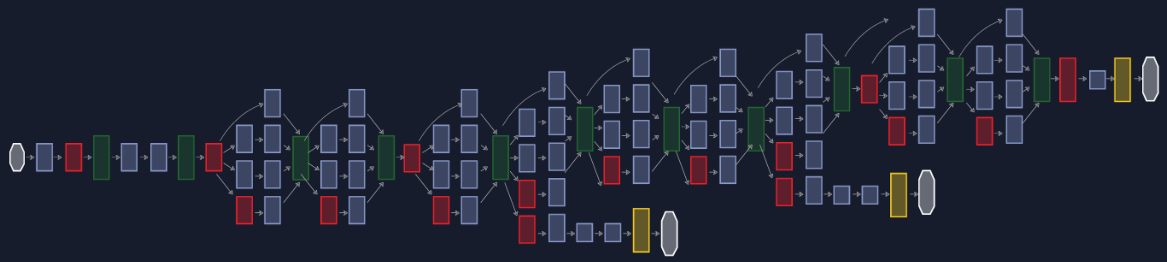
Source:
https://arxiv.org/pdf/1605.07678.pdf https://arxiv.org/pdf/1608.06993.pdf
https://arxiv.org/pdf/1709.01507.pdf https://arxiv.org/pdf/1611.05431.pdf

Network Complexity is Growing

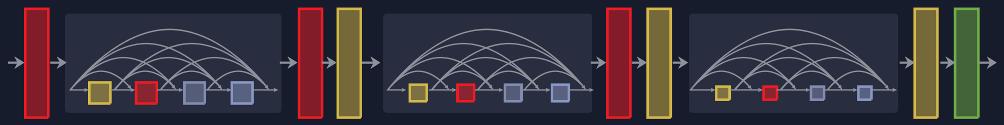
AlexNet



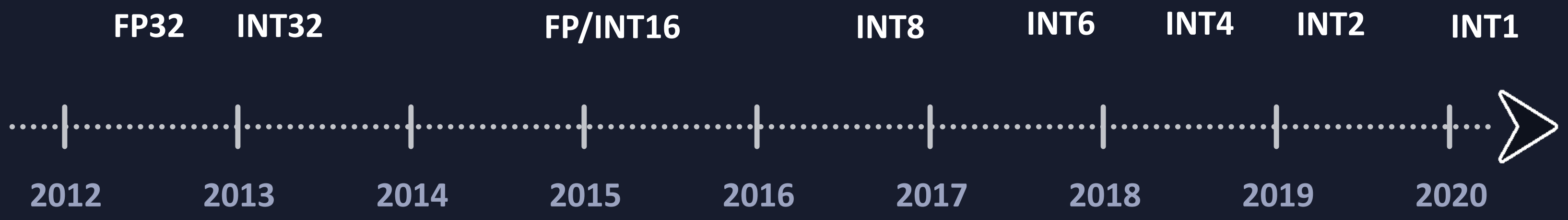
GoogLeNet



DenseNet



Inference is Moving to Lower Precision



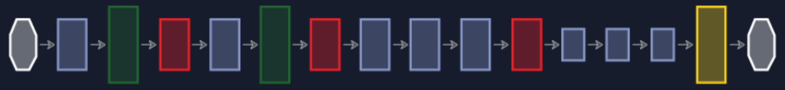
RELATIVE ENERGY COST

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

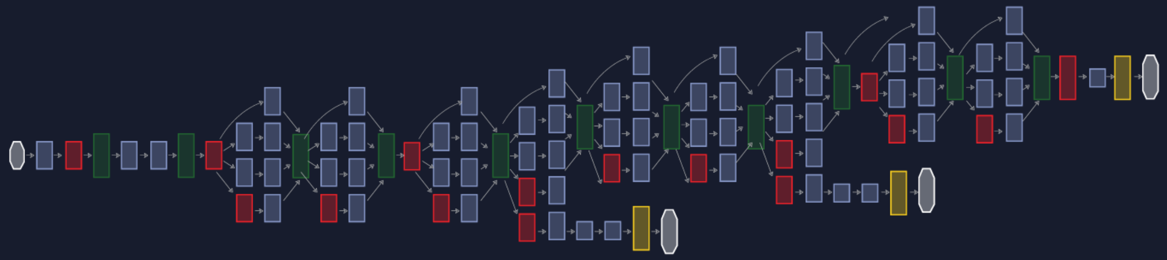
Source: Bill Dally (Stanford), Cadence Embedded Neural Network Summit, February 1, 2017

Rate of Innovation Outpaces Silicon Cycles

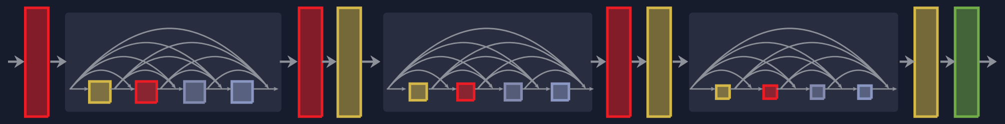
AlexNet



GoogLeNet

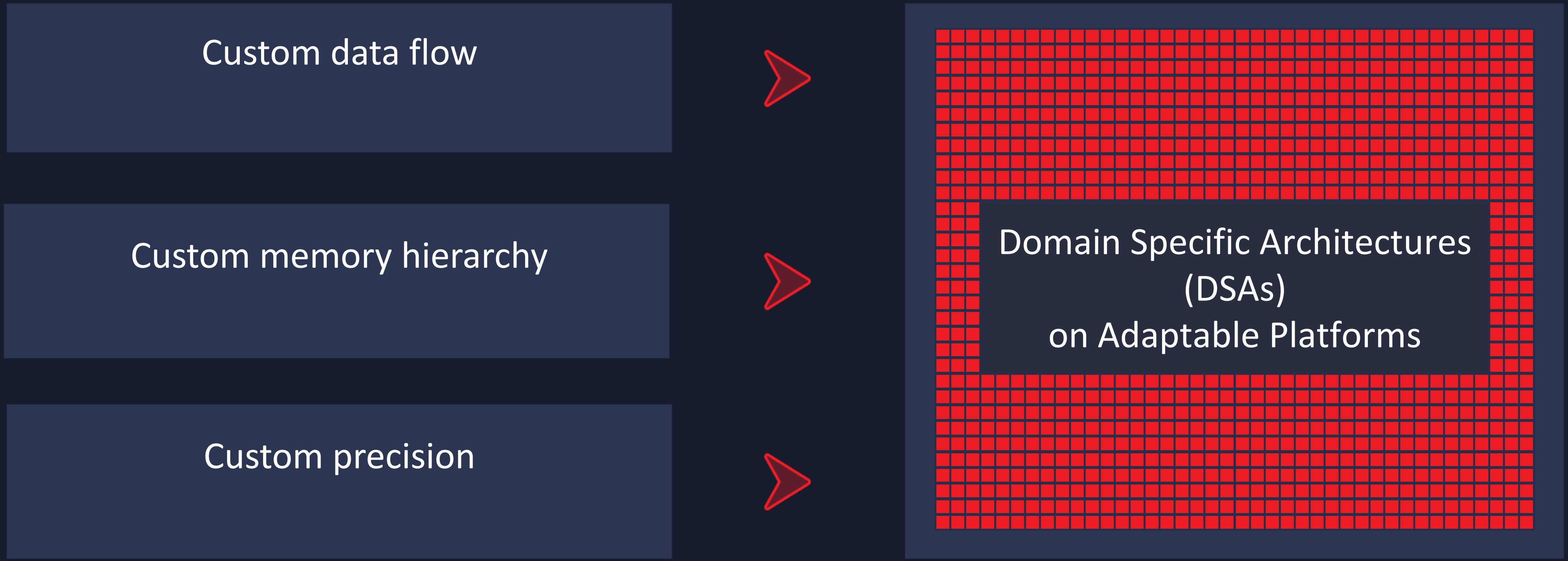


DenseNet



Silicon lifecycle

Only Adaptable Hardware Addresses Inference Challenges



Custom data flow

Custom memory hierarchy

Custom precision

Domain Specific Architectures (DSAs)
on Adaptable Platforms

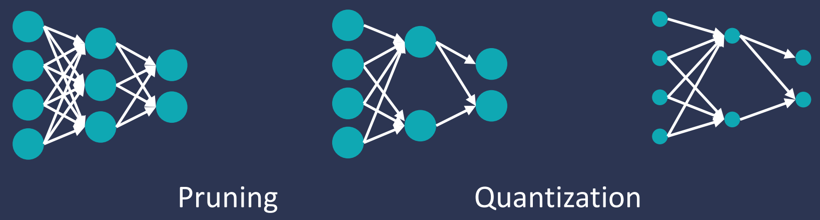
DeePhi Joins Xilinx

Custom data flow

Custom memory hierarchy

Custom precision

DEEPHi 深 鉴 科 技 Now Part of XILINX®



Patented Compression Technology

- Reduces DL accelerator footprint
- Increases performance per watt

Example: DeePhi LSTM

Custom data flow

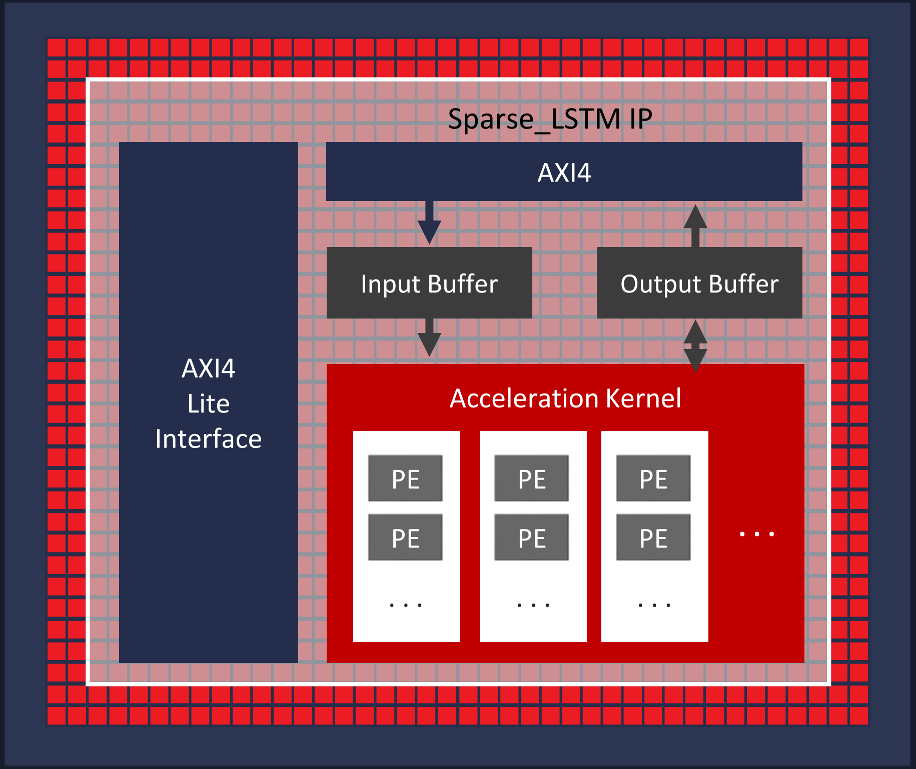
LSTM for speech recognition

Custom memory hierarchy

Sparse matrix implementation in memory

Custom precision

12 bit weights, 16 bit activations



Example: xDNN

Custom data flow

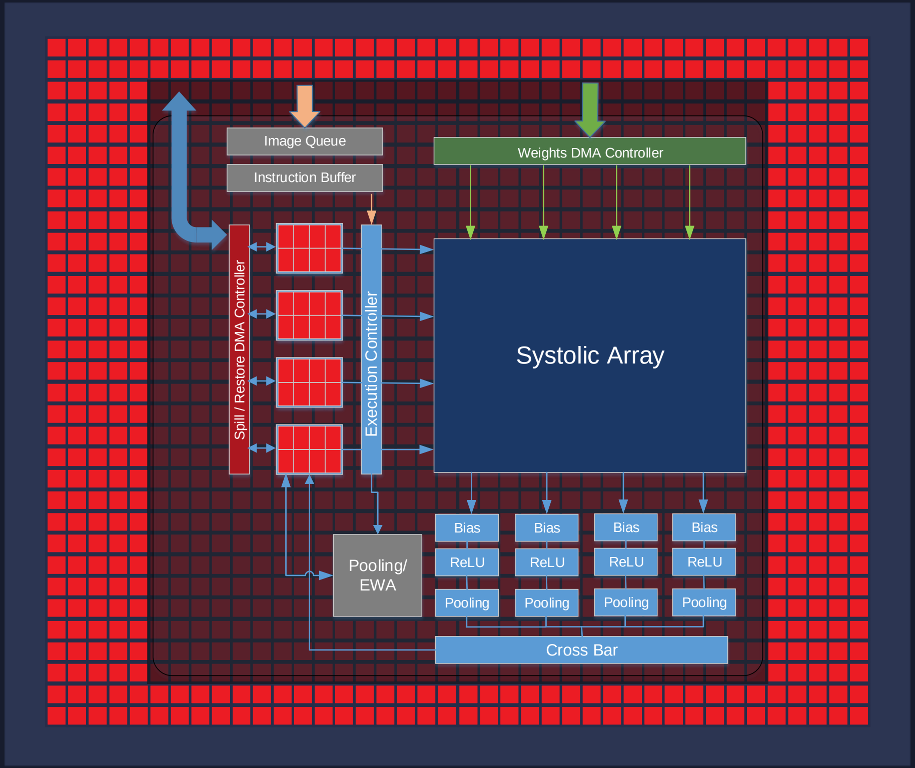
Optimized for latest CNN

Custom memory hierarchy

Optimized on-chip memory

Custom precision

Int8



Low Latency is Critical for Inference

50ms latency response

3ms latency response

CPU/GPU

Input 1 | Result 1
Input 2 | Result 2
Input 3 | Result 3
Input 4 | Result 4

High throughput OR low latency

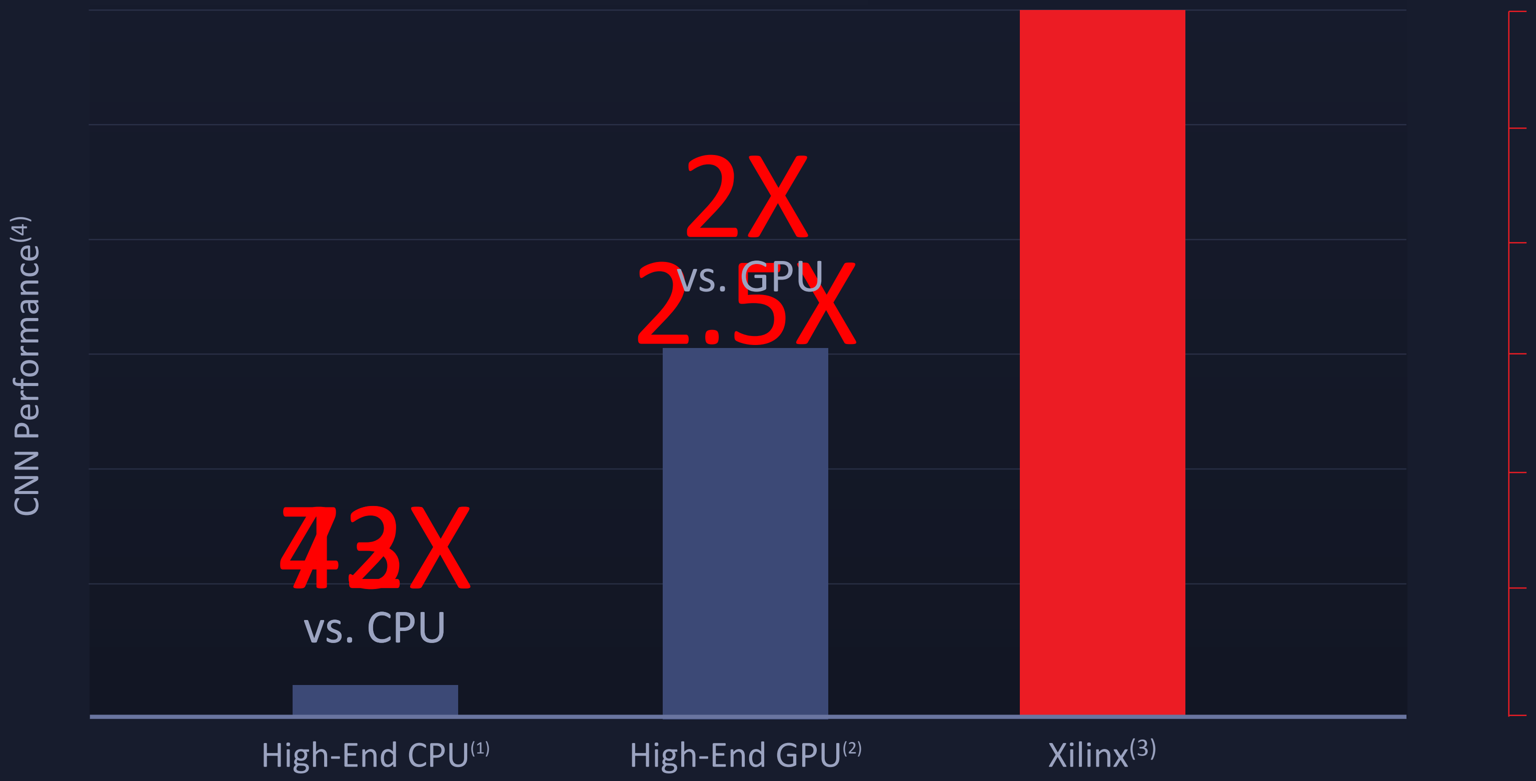
FPGA/ACAP

Input 1 | Result 1
Input 2 | Result 2
Input 3 | Result 3
Input 4 | Result 4

High throughput AND low latency

Low Latency: Xilinx's Unique Advantage

Latency Insensitive Inference



AI Inference Acceleration

Leveraging AI Engines

Majority of Adaptable & Scalar Engines available for Whole App Acceleration

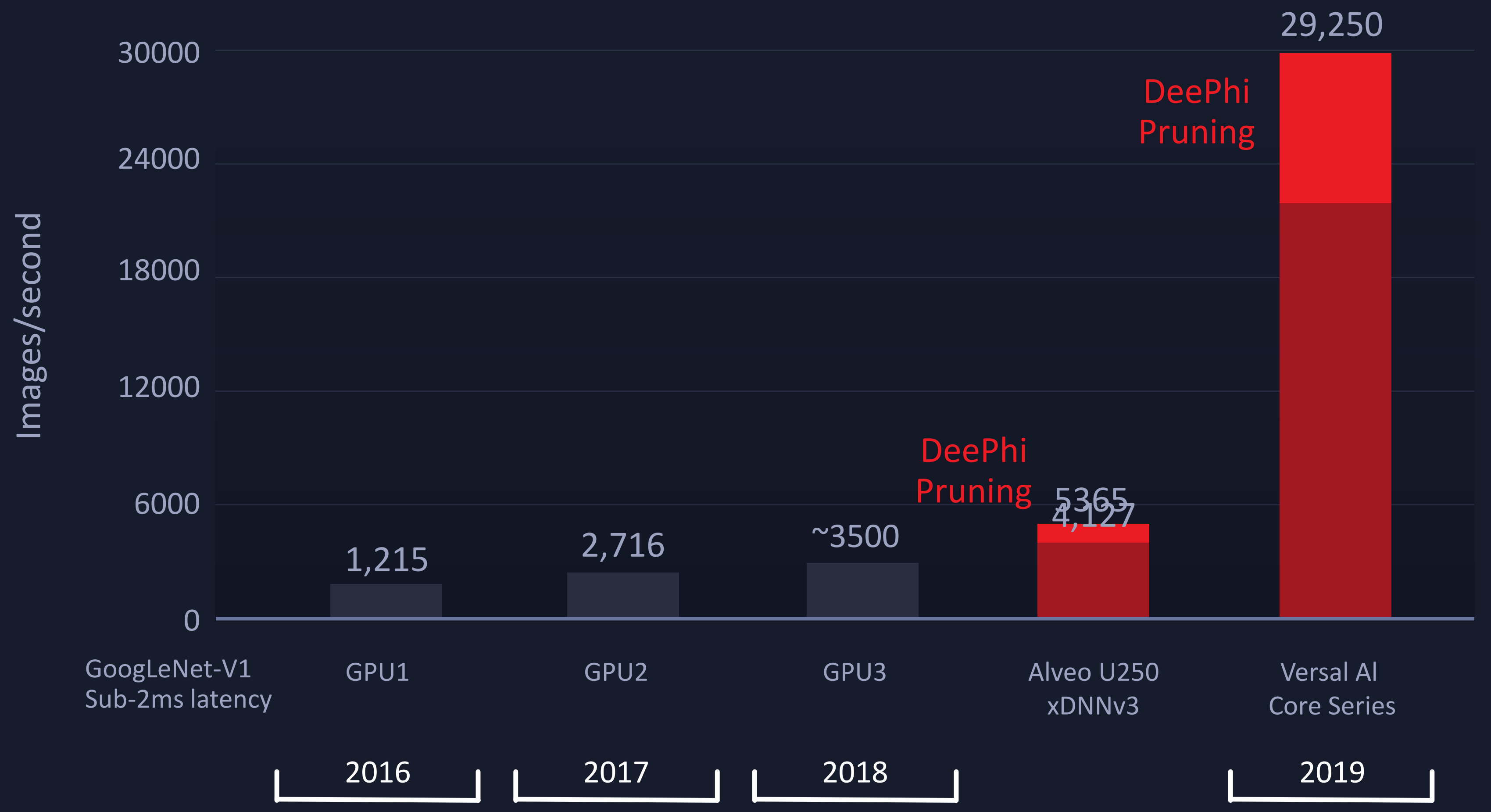
[1] Measured on EC2 Xeon Platinum 8124 Skylake, c5.18xlarge AWS instance, Intel Caffe: https://github.com/intel/caffe

[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



DeePhi Pruning Technology

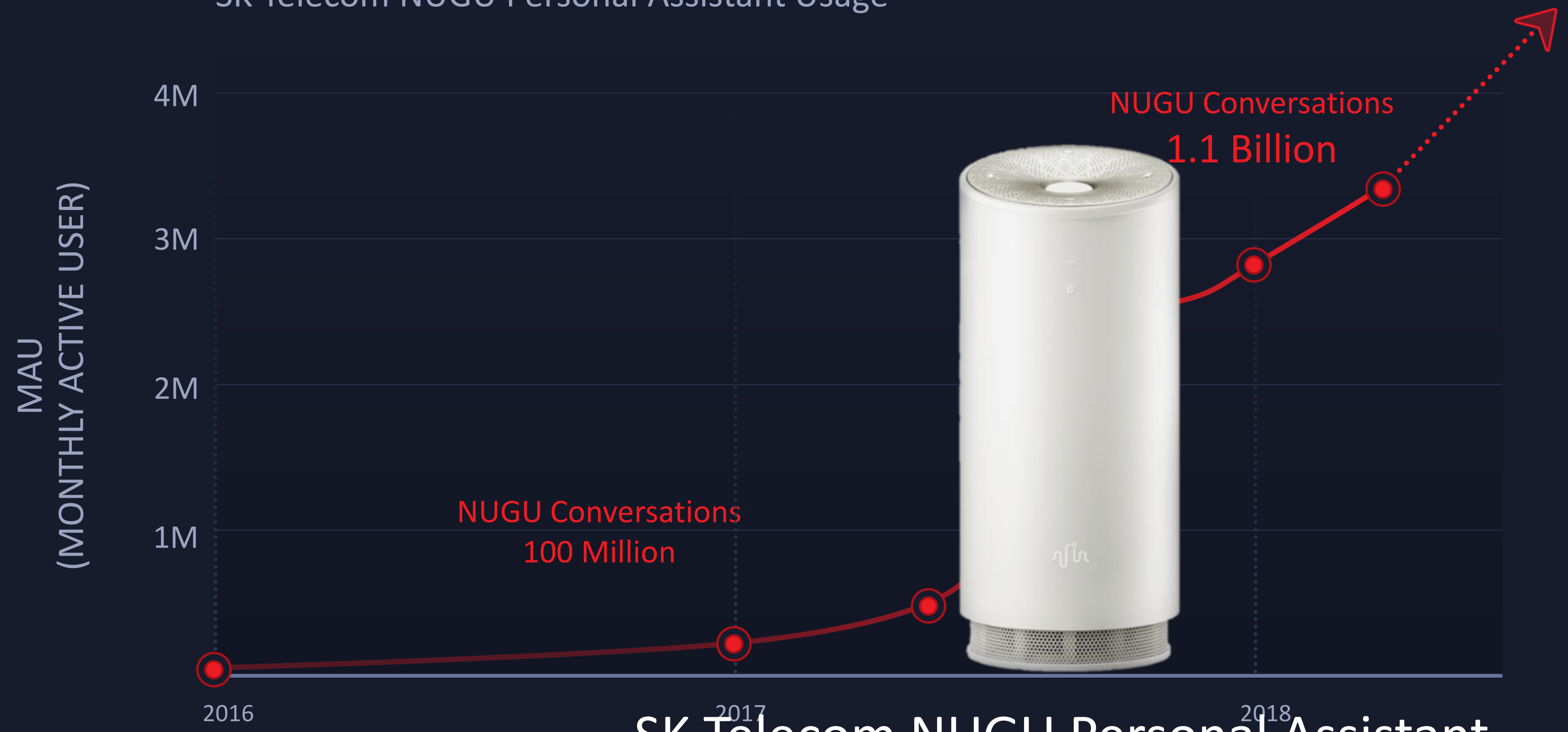
1.3x-8x

Performance improvement based on the network

Sources: Alveo - Published (INT8); Versal - Proiected (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

Cloud Inference
SK Telecom NUGU Personal Assistant Usage



NUGU Conversations
100 Million

NUGU Conversations
1.1 Billion

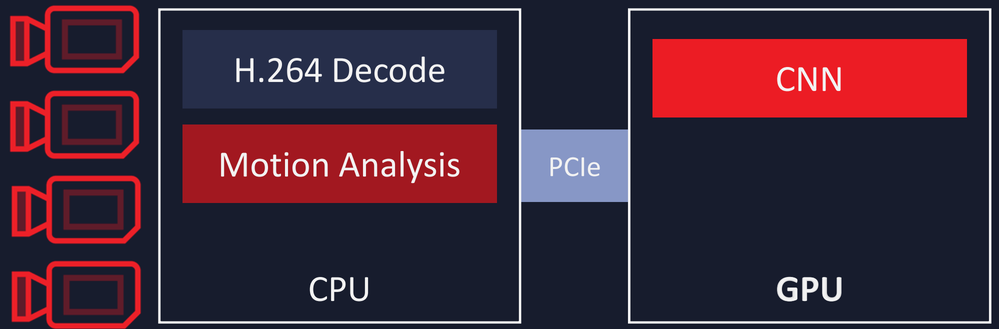
SK Telecom NUGU Personal Assistant

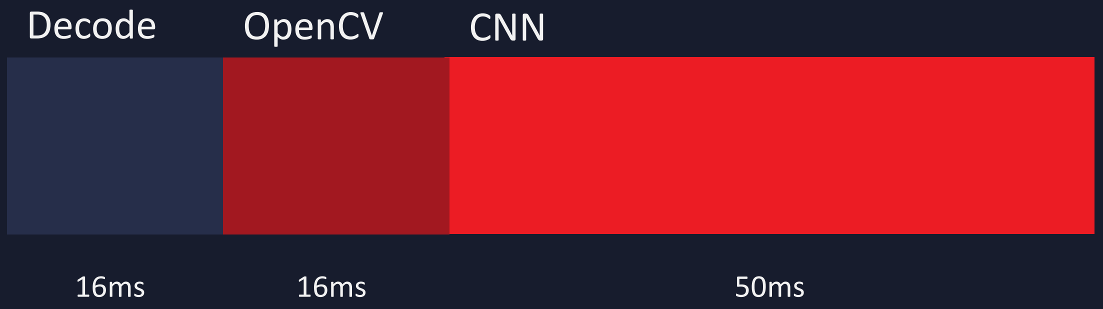
16x

Perf/watt
vs. GPU

Whole Application Acceleration: Smart City / Security

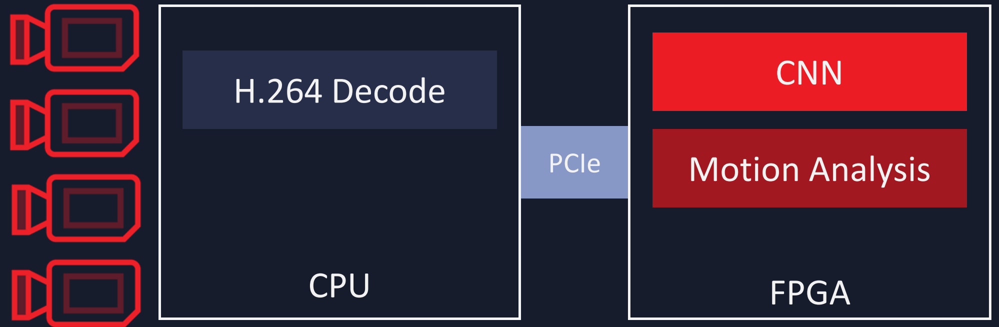
CPU-GPU

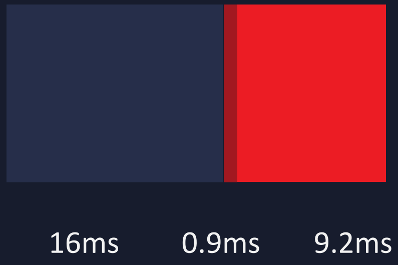




- Power: 75W
- Latency: 82 ms
- Throughput: 4x12 fps

CPU-Xilinx FPGA





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

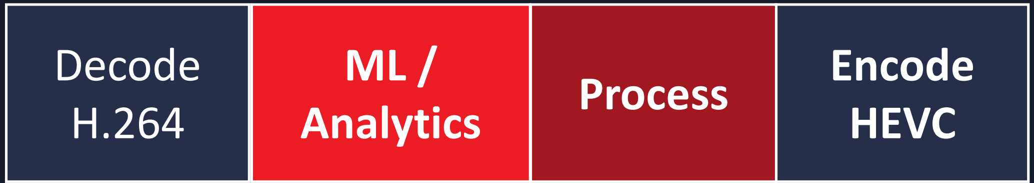
Whole Application Acceleration: Online Video Streaming

欢聚时代 YY Inc.

Aupera
Making Video Alive

1
Aup2603

48 ZU7EV



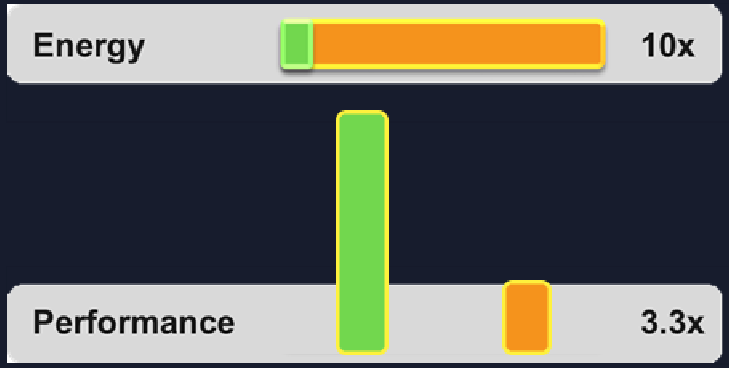
Decode H.264

ML / Analytics

Process

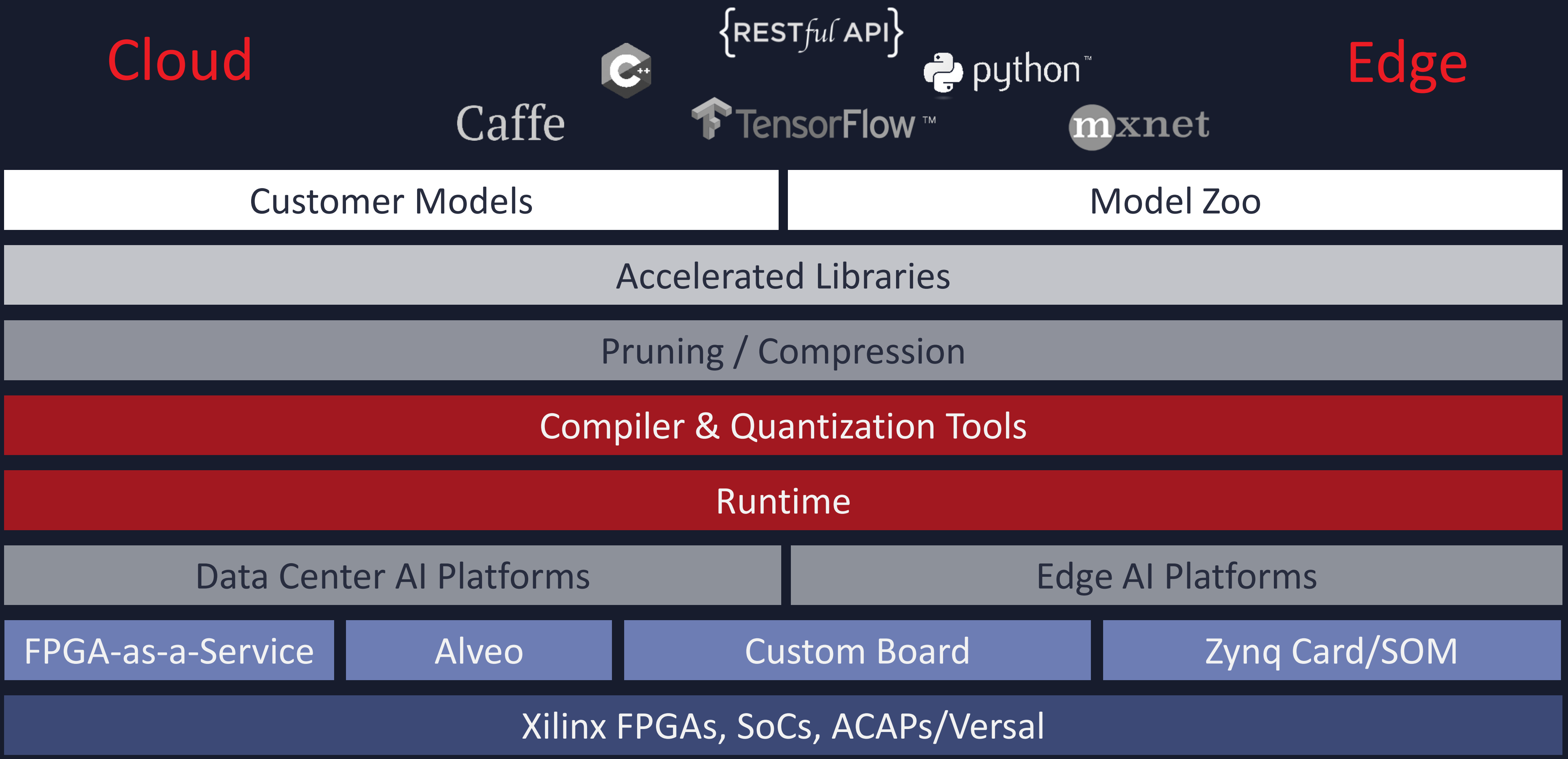
Encode HEVC

Video transcoding + AI analytics



30
E5 Servers

AI Solution Stack



Cloud

Edge

Caffe

C++

{RESTful API}

TensorFlow™

python™

mxnet

Customer Models | Model Zoo

Accelerated Libraries

Pruning / Compression

Compiler & Quantization Tools

Runtime

Data Center AI Platforms | Edge AI Platforms

FPGA-as-a-Service | Alveo | Custom Board | Zynq Card/SOM

Xilinx FPGAs, SoCs, ACAPs/Versal

IN SUMMARY

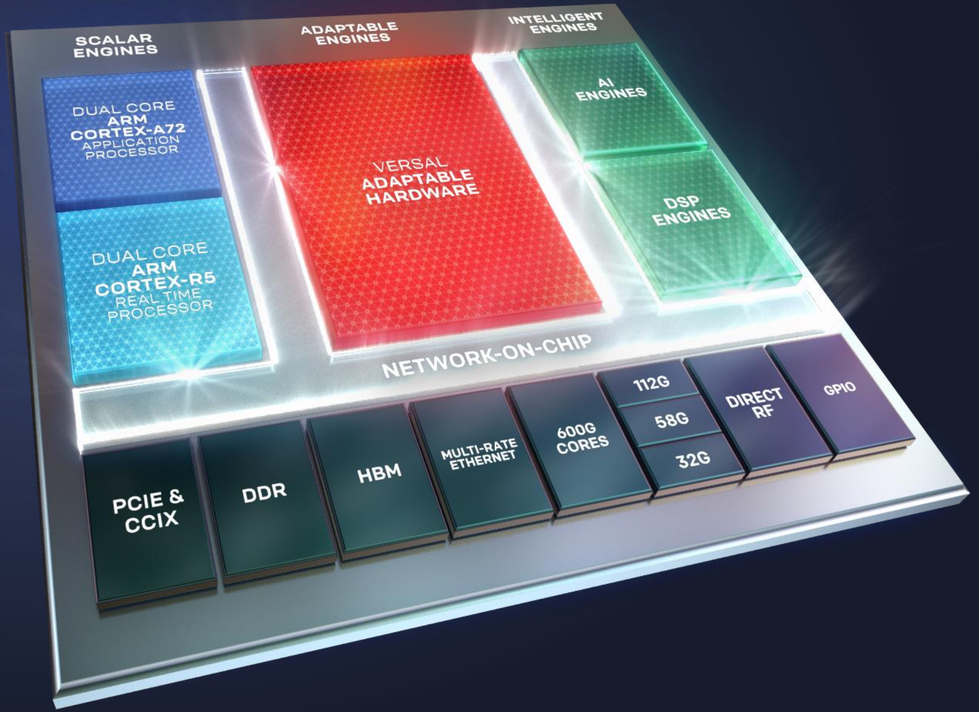
Only Xilinx Adaptable Devices Can:

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