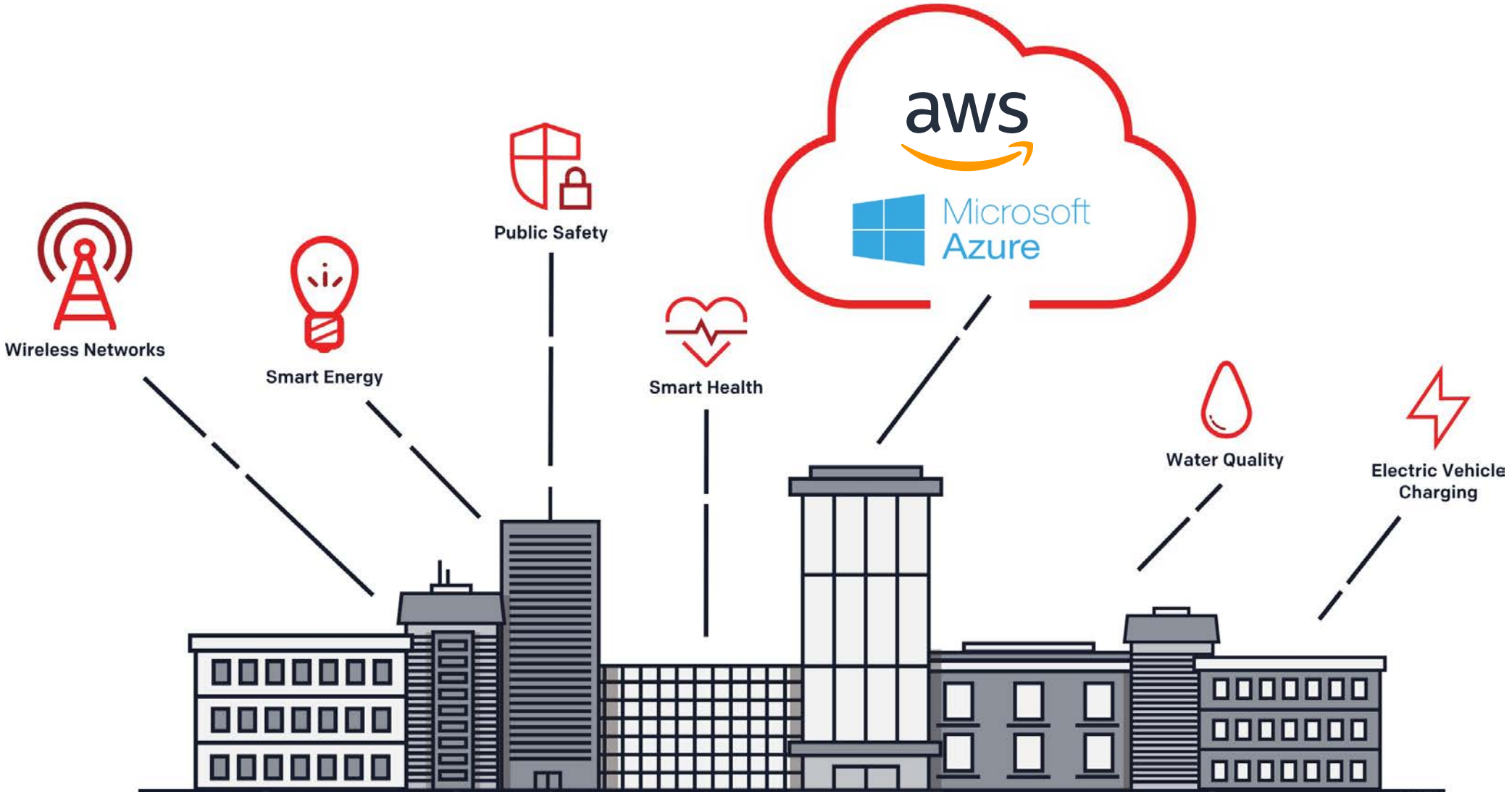
Adaptable Intelligence
The Next Computing Era

Hot Chips, August 21, 2018

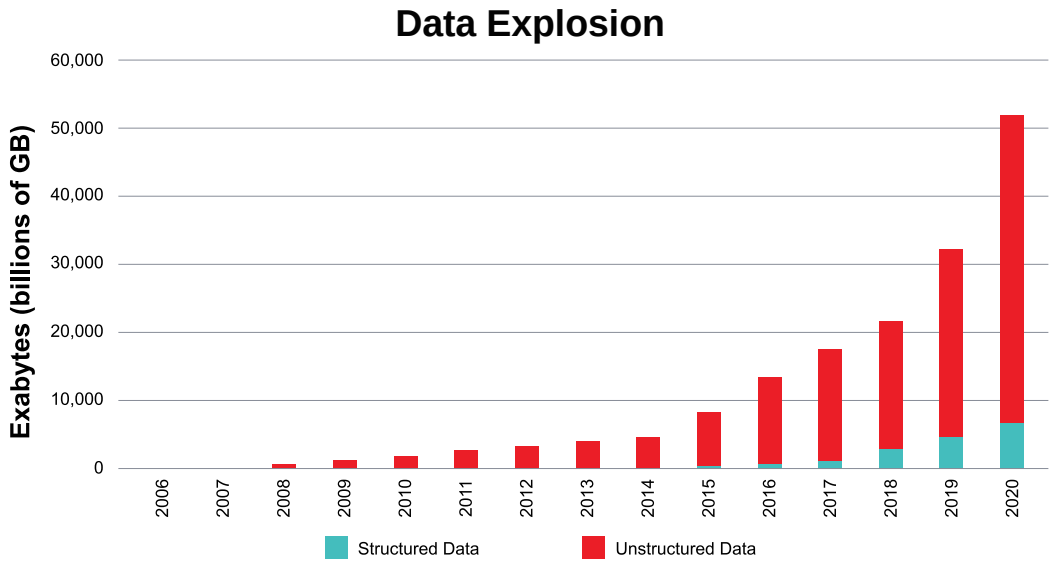
Victor Peng, CEO, Xilinx

XILINX

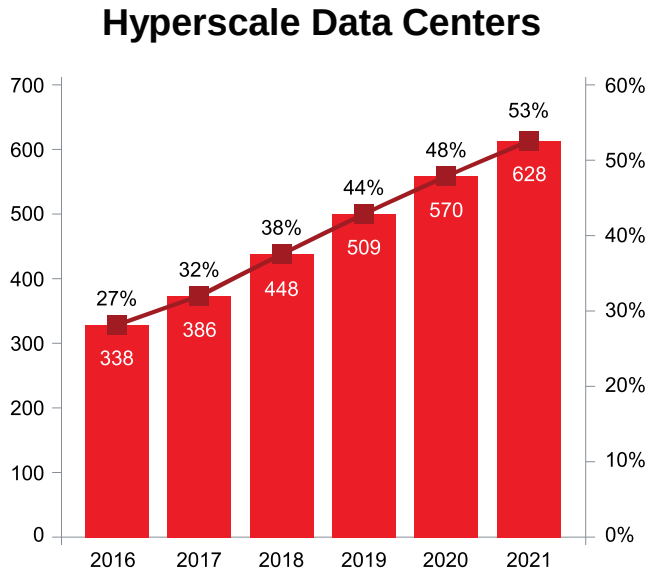
Pervasive Intelligence from Cloud to Edge to Endpoints



Exponential Growth and Opportunities

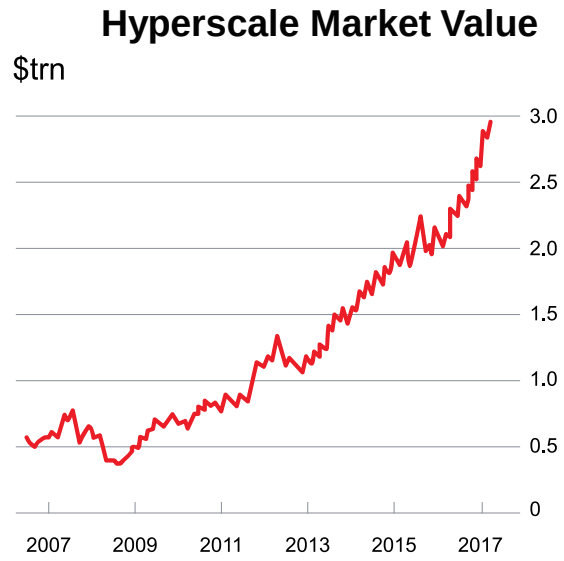


Source: Patrick Cheesman



% of All Data Center Servers

Source: Cisco Global Cloud Index, 2016-2021

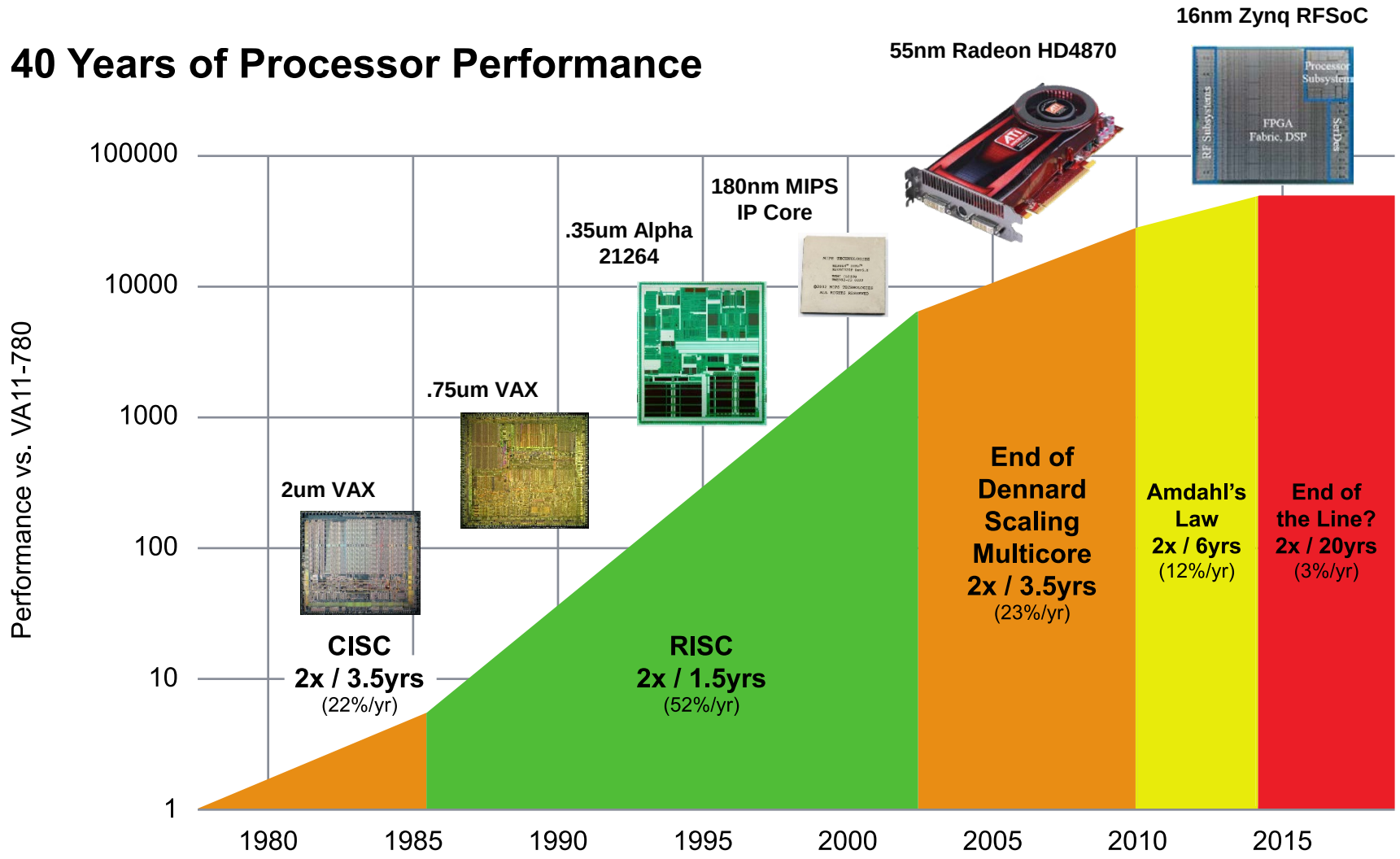


Combined Market Capitalisation

Source: Bloomberg; company reports; The Economist estimates

XILINX

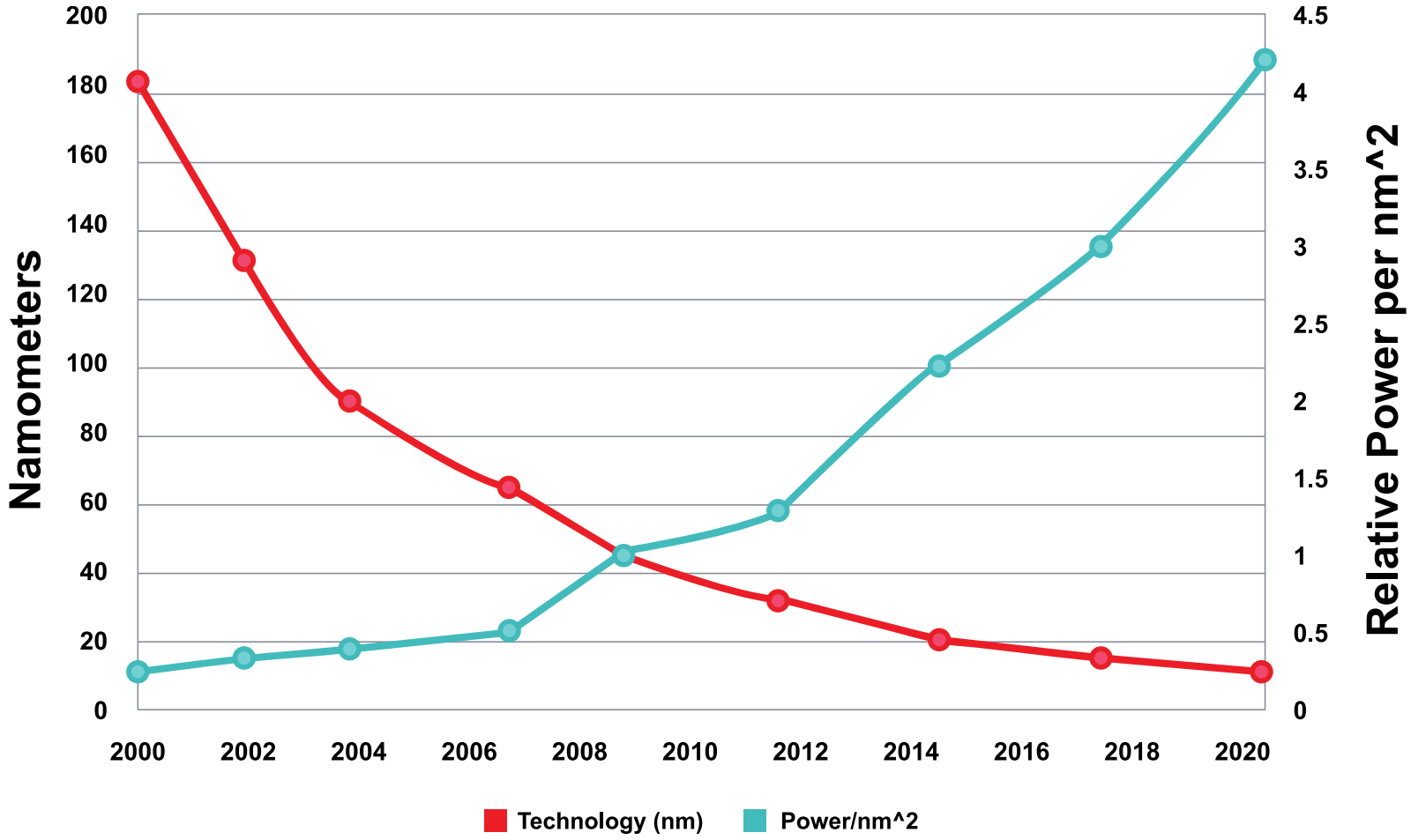
Challenges: The End of Moore's Law and Scaling



Source: John Hennessy and David Patterson, Computer Architecture: A Quantitative Approach, 6/e 2018

XILINX.

Challenge: Exponential Power Density Growth



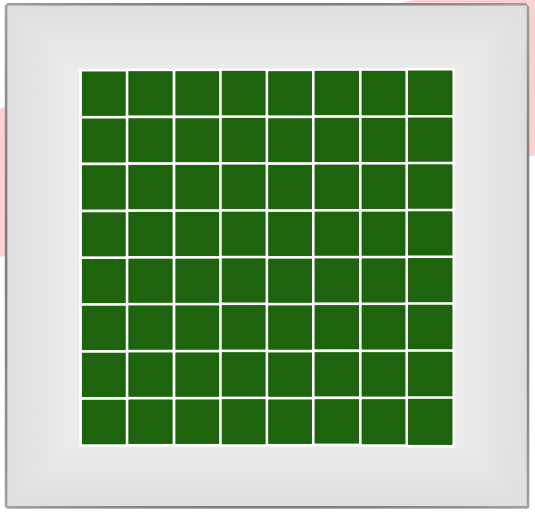
Source: John Hennessy and David Patterson: A New Golden Age for Computer Architecture
Domain-Specific Hardware/Software Co-Design, Enhanced Security, Open Instruction Sets, and Agile Chip Development

Power consumption based on models in "Dark Silicon and the End of Multicore Scaling"
Hadi Esmaelizadeh, *ISCA,* 2011

The Third Wave: Domain Specific Architectures on Adaptable HW

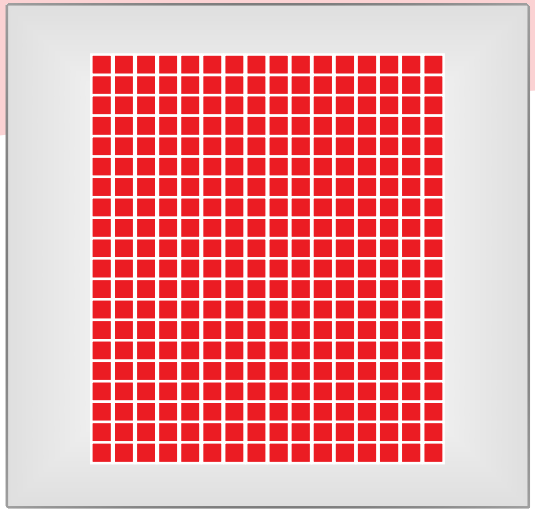
CPU

CISC → RISC → Multi-Core



Fixed HW Accelerators

GPU, ASSP, ASIC

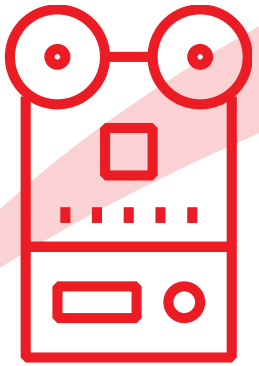


DSA's on Adaptable Platforms

FPGA, Zynq, ACAP

Massive Scale Out Requires DSA's and Adaptable Platforms

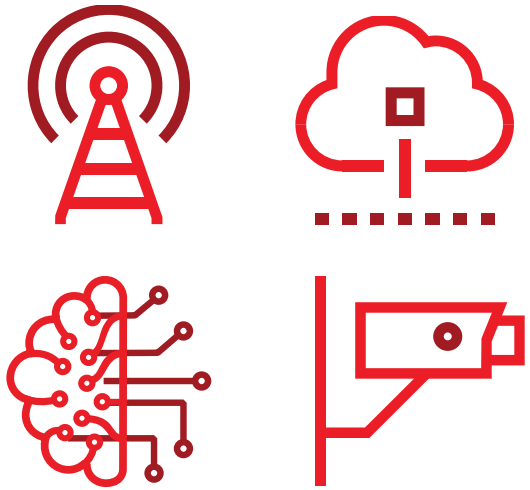
Integration of AI



Mainframe Era
Units (M)

PC Era
Units (100M's)

Mobile Era
Units (B's)

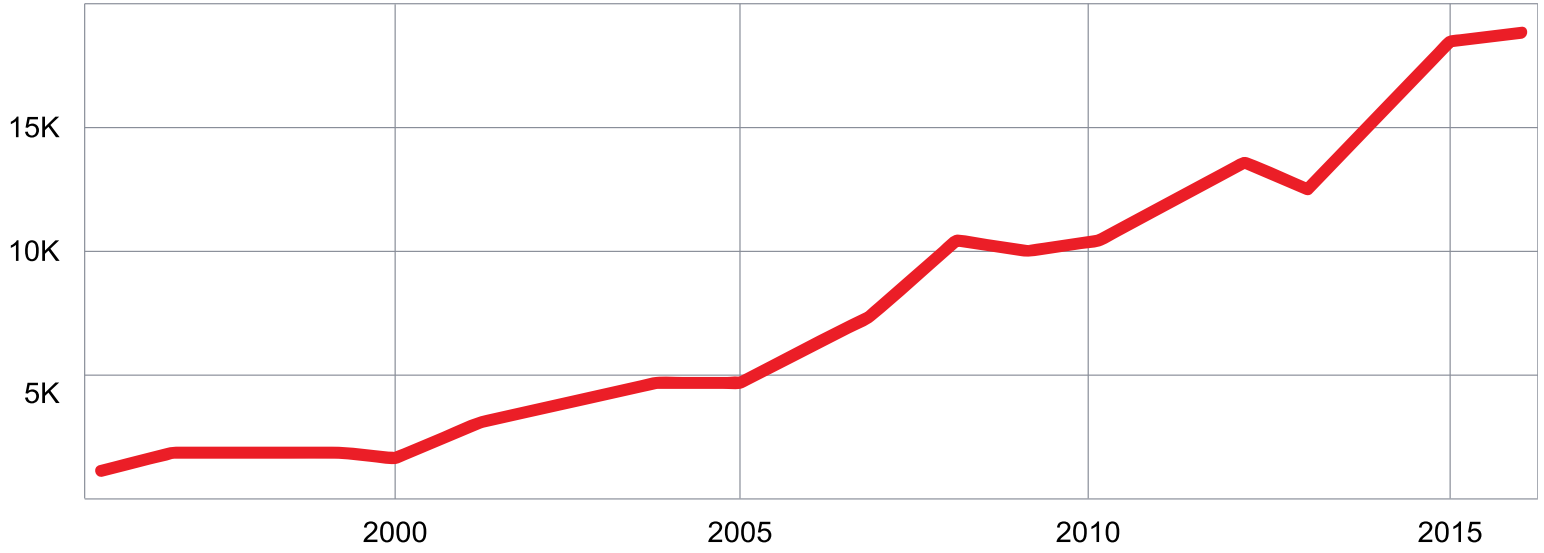


Pervasive Intelligence Era
Units (50B's)

XILINX.

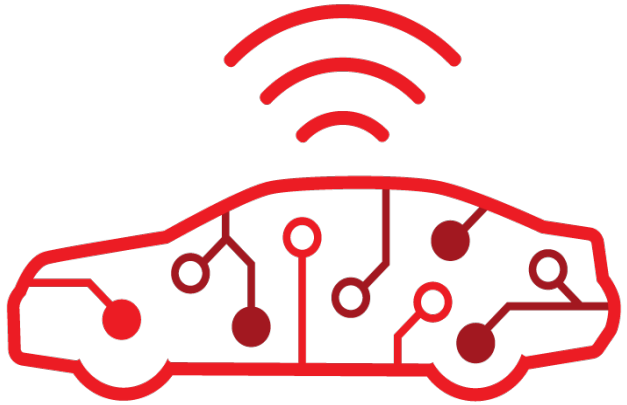
The Innovation to Deployment Acceleration Imperative

AI Papers Published



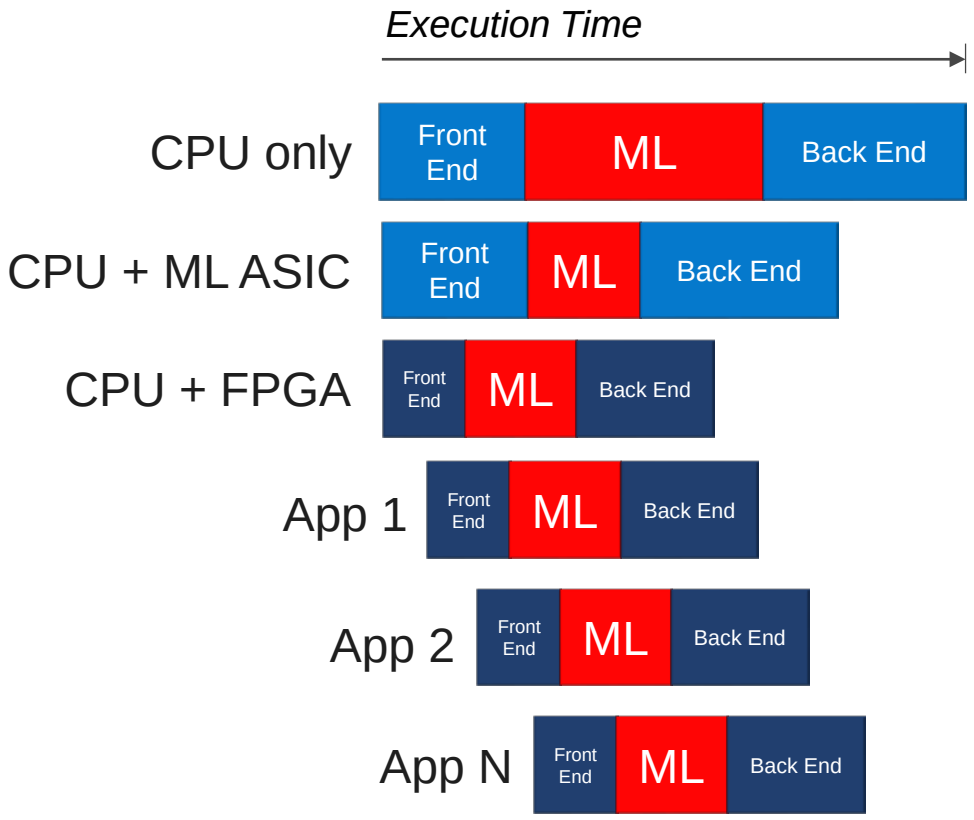
Source: Scopus

Over-the-air Updates

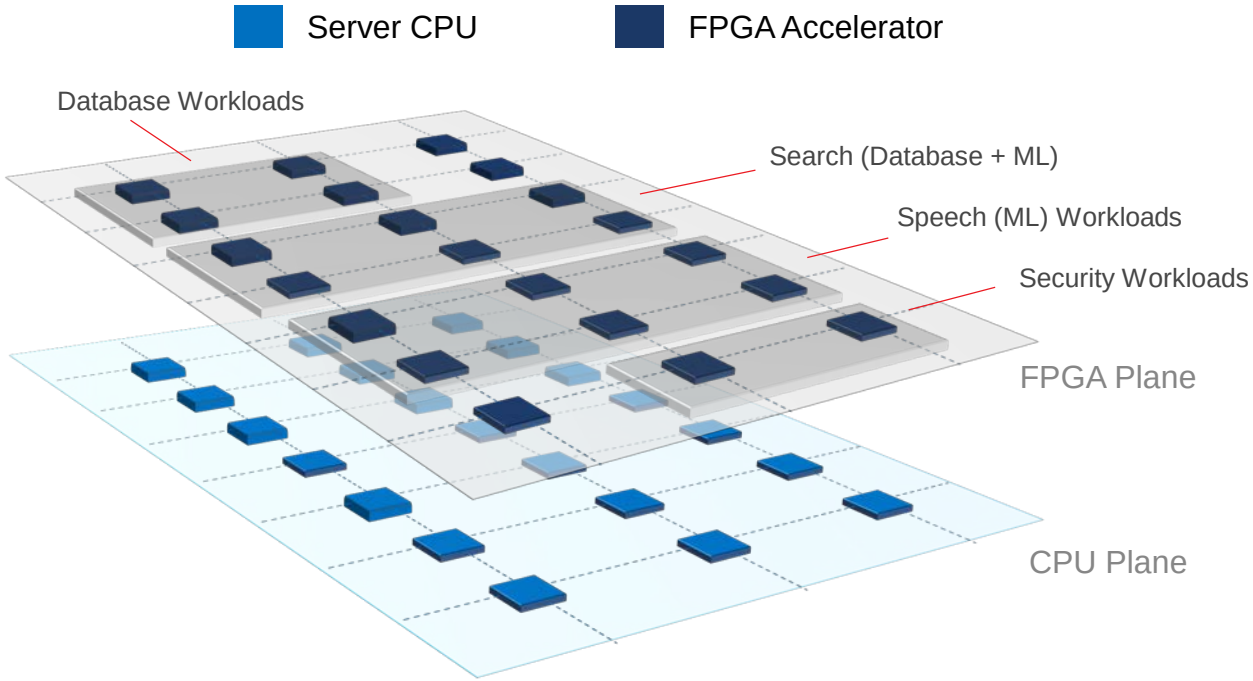


Application Acceleration with DSA's on FPGA Platforms

FPGA's Accelerate Entire Application



Hyperscale Data Centers with FPGA Accelerators

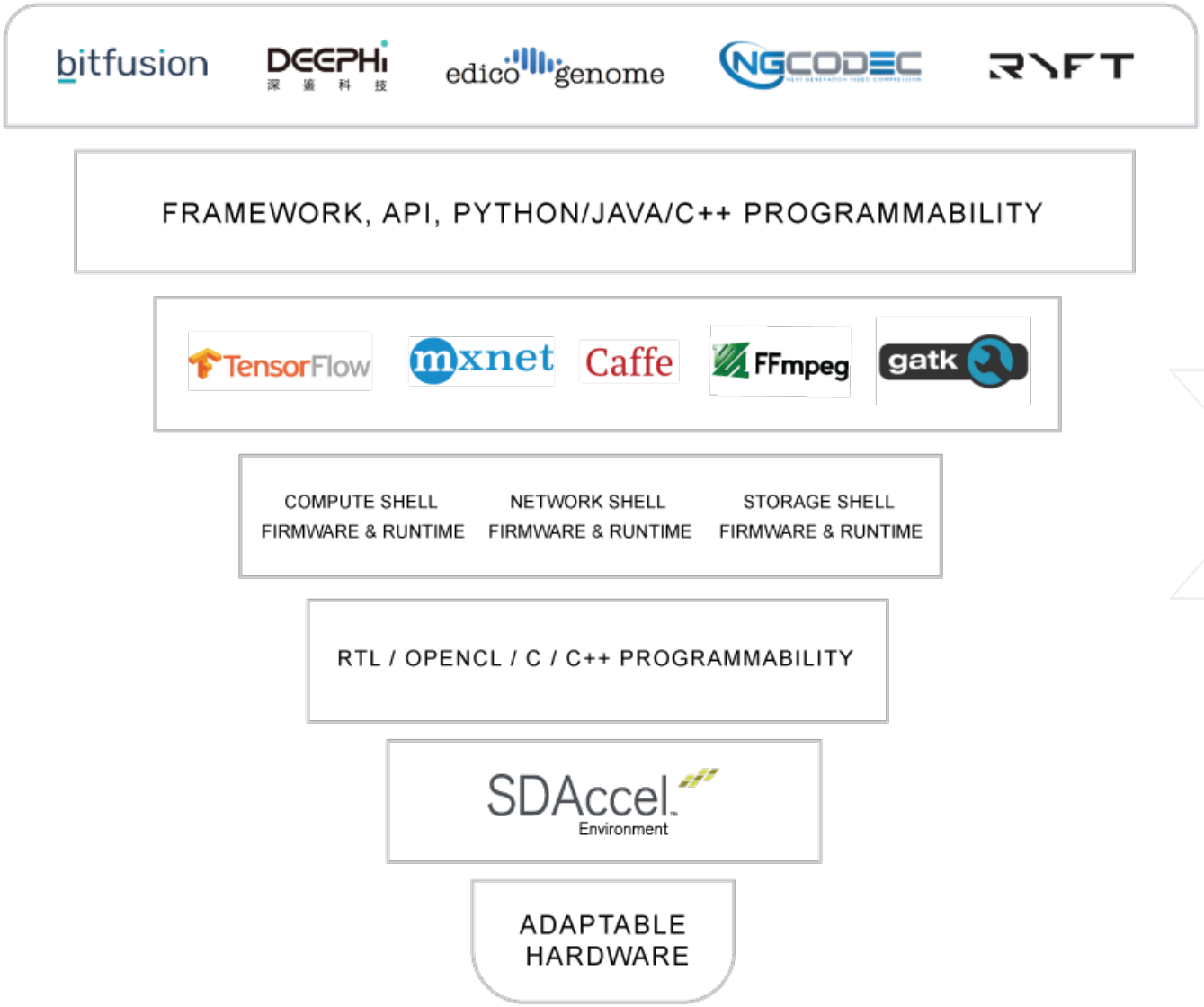


Virtualized & Scaled Out Adaptable Acceleration

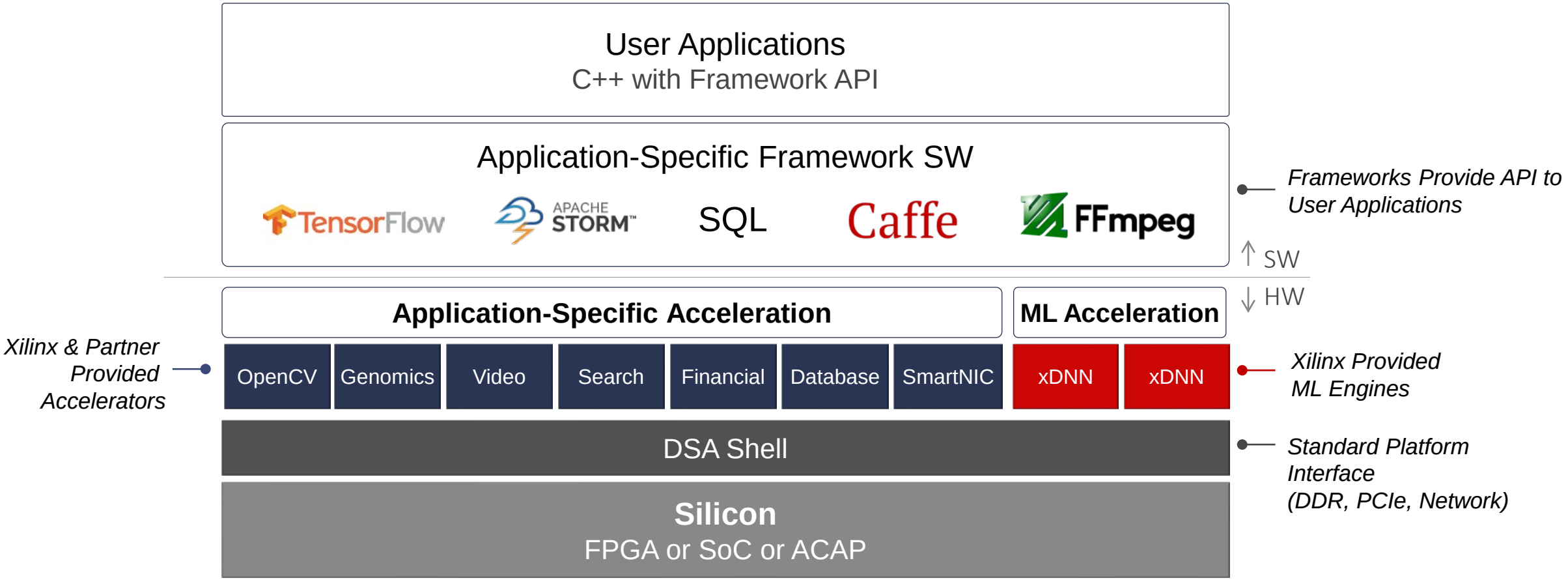
Dynamic Optimization for Changing Workloads & Mix

XILINX.

Development for DC Compute, Storage, Network Apps



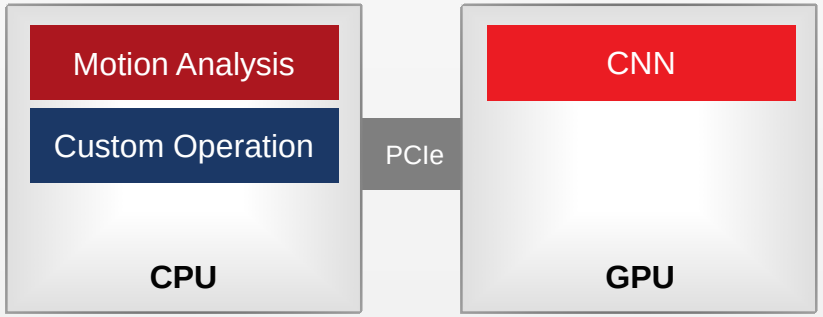
Stack for Application Acceleration including ML



XILINX

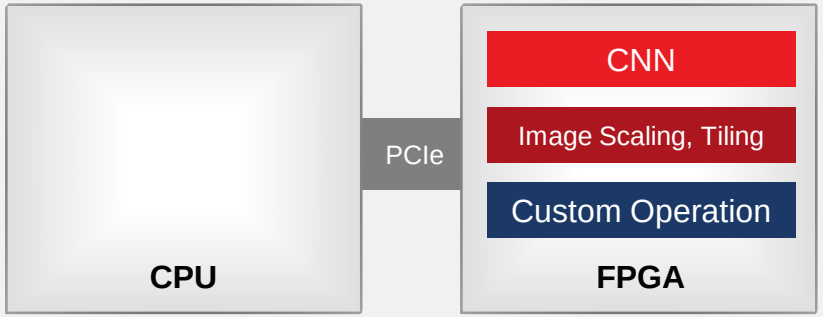
Cloud: Latency-sensitive High Resolution Imaging

CPU/GPU

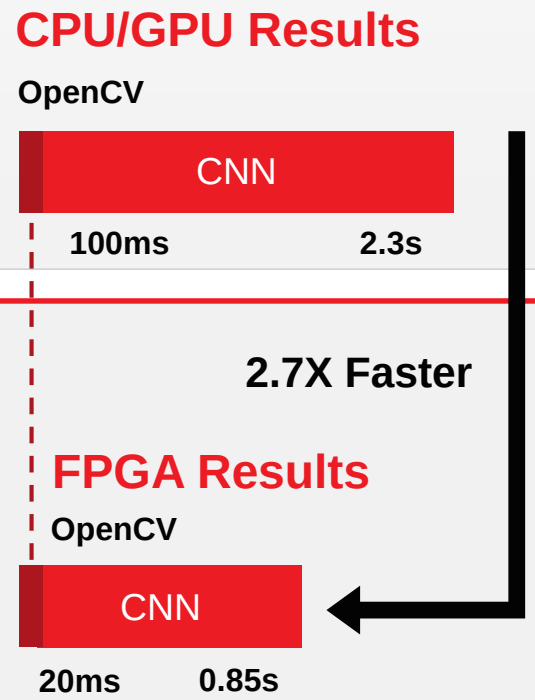


- OpenCV: Pre-processing a Bottleneck
- CNN: Object Detection/Feature Extraction
- Data Sharing Between CPU and GPU

Xilinx FPGA



- OpenCV: 5x Faster
- CNN: Up to 3x Faster
  Based on Image Size
- Model Parallelism for High Res Processing
- Lower Power

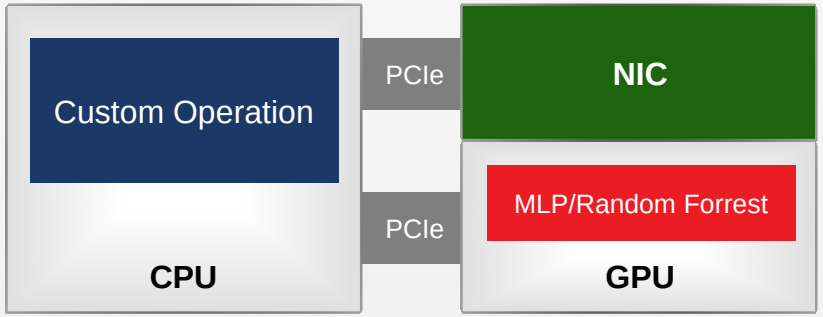


2.7X Faster and Lower Power

XILINX

Cloud: Security / Anomaly Detection

CPU/GPU



- High Performance NIC: 100+ G
- MLP: MLP/RF Acceleration
- Security Vulnerability: Data Travels thru CPU to GPU

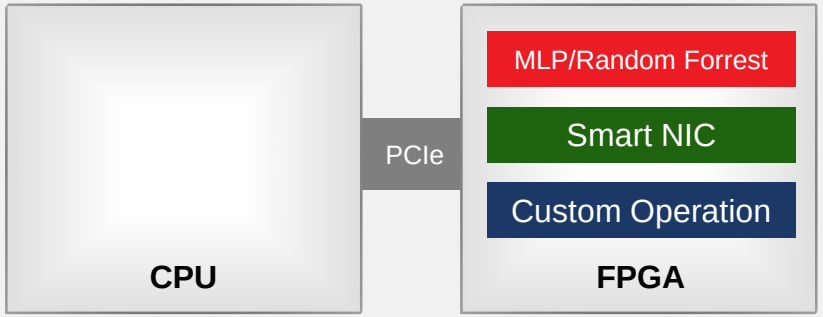
CPU/GPU Results

OpenCV

MLP

50 µs

Xilinx FPGA



- Integrated Single Card/Chip Solution
- High Speed Smart NIC
- Real-time MLP/RF at 5x Lower Latency
- TCO and Power Advantage: 1 Card vs 2 Cards
- In-line Security Detection is Much Higher Security

5X Lower Latency

FPGA Results

OpenCV

MLP

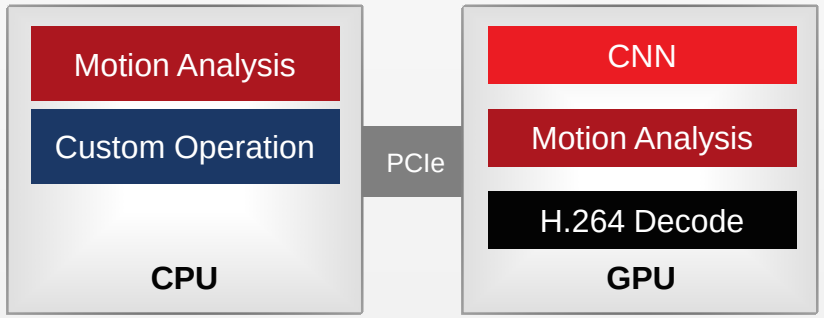
10 µs

5X Lower Latency with High Security

XILINX

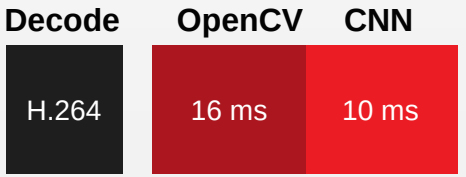
Cloud: Smart City / Security

CPU/GPU

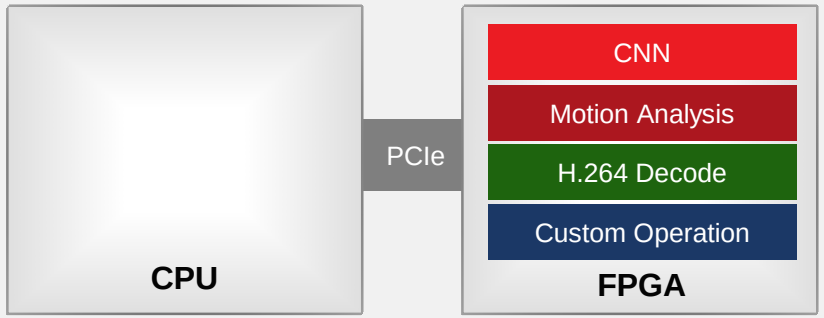


- H.265 Decode: 4x 1080p Streams (P4)
- OpenCV: Pre-Processing/Motion Analysis on CPU
- CNN: Object Detection on GPU
- Data Sharing Between CPU and GPU

CPU/GPU Results



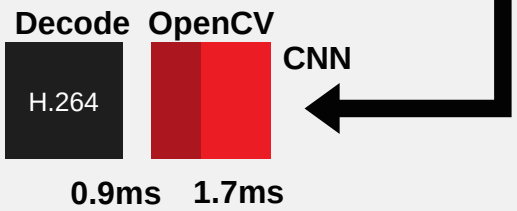
CPU/Xilinx FPGA



- H.265 Decode: 4x 1080p Streams (P4)
- OpenCV: Up to 20x Higher Performance
- CNN: 5ms Object Detection (xDNN Acceleration)
- Integrated Single Chip Solution: Cloud (VU9/13P)
- Lower Power

10X Lower Latency

FPGA Results



10x Lower Latency and Lower Power

XILINX

Xilinx Programmable Architecture Milestones

First FPGA Introduced

1980

First Virtex FPGA

1990

Virtex-2 Pro

2000

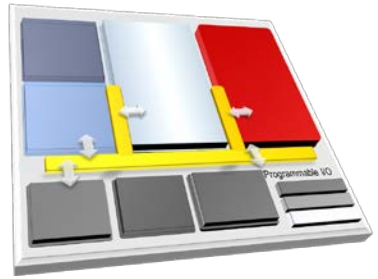
First 3D FPGA & HW/SW Programmable SoC

2010

First MPSoC & RFSoC

2016

ACAP



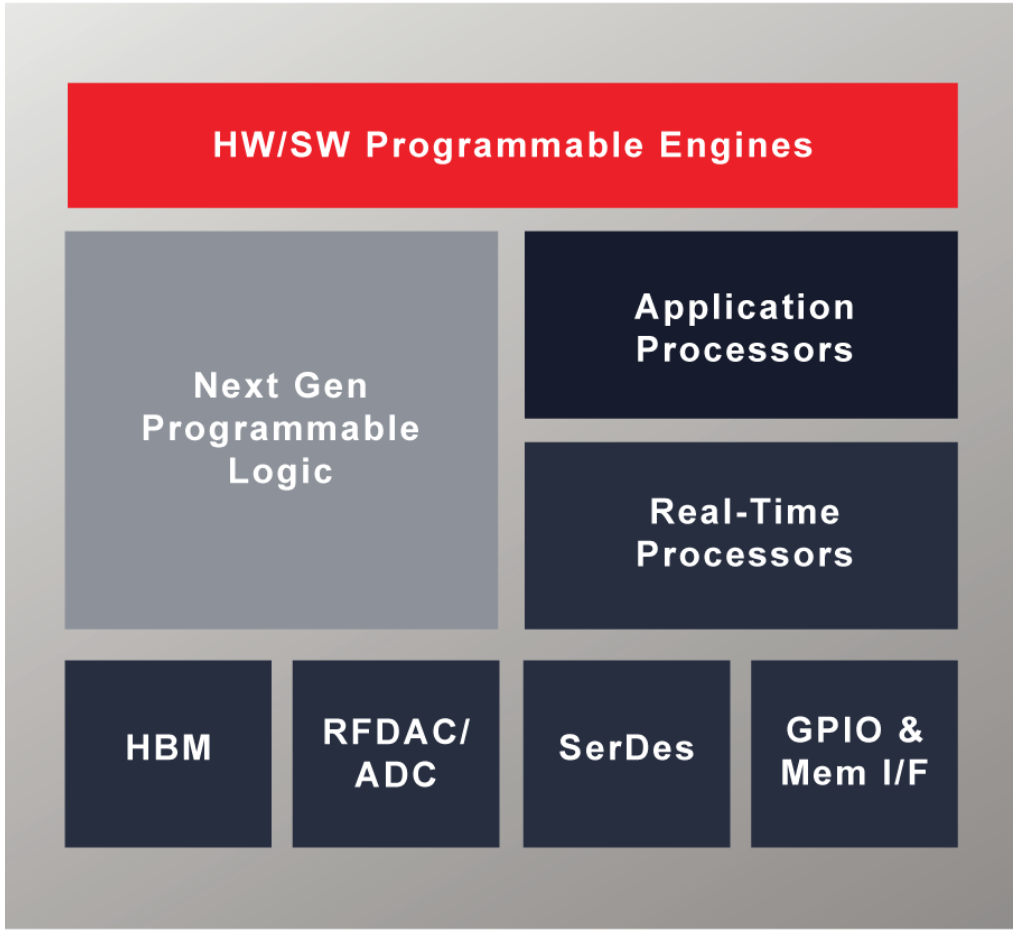
2019

XILINX

From FPGA to Adaptive Compute Acceleration Platform

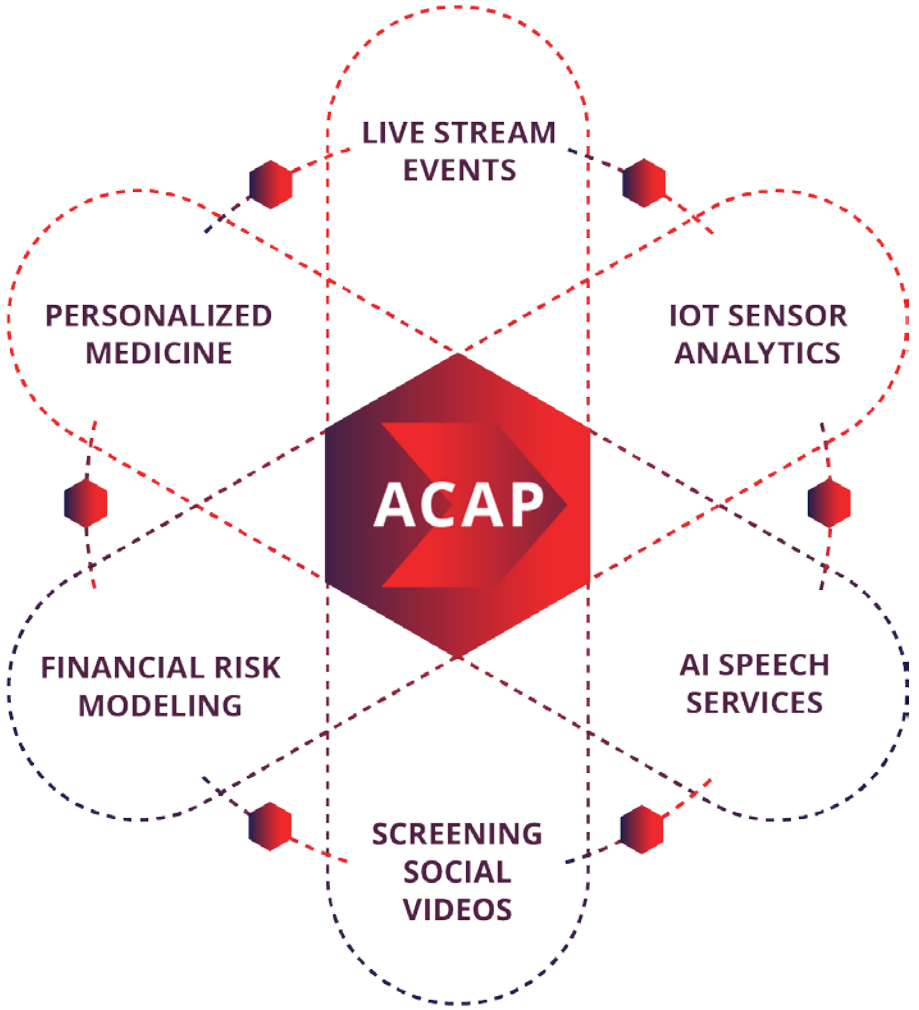
New Device Category for Adaptive Workload-specific Acceleration

- HW/SW Programmable Engines
- IP Subsystems and a Network-on-Chip
- Platform Offerings for Compute / Storage / Networking



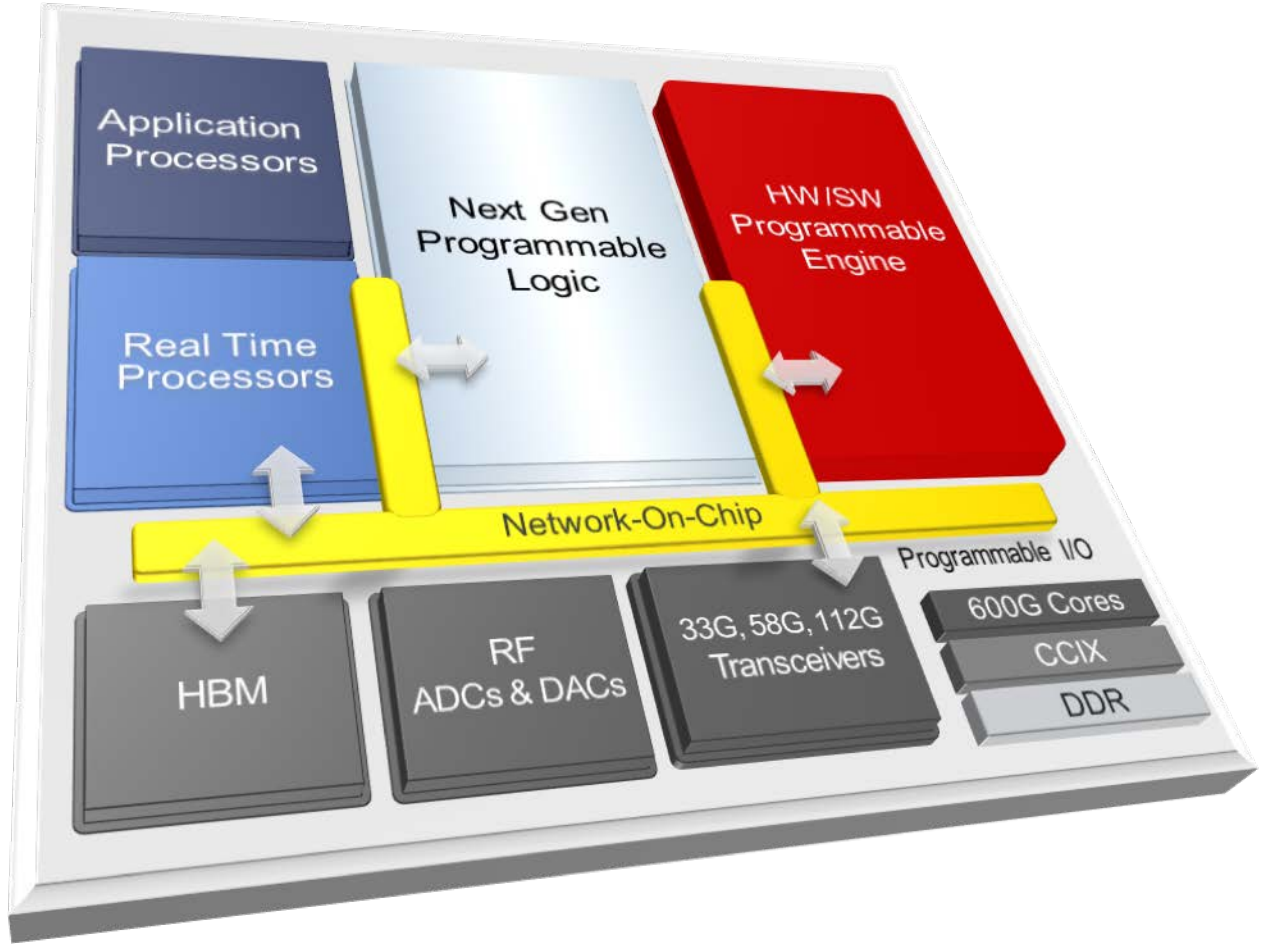
Adaptive Compute Acceleration Platform

- Dynamically Adaptable to Workloads
- Exponential Increase in Acceleration
- Software Programmable



Adaptive Compute Acceleration Platform

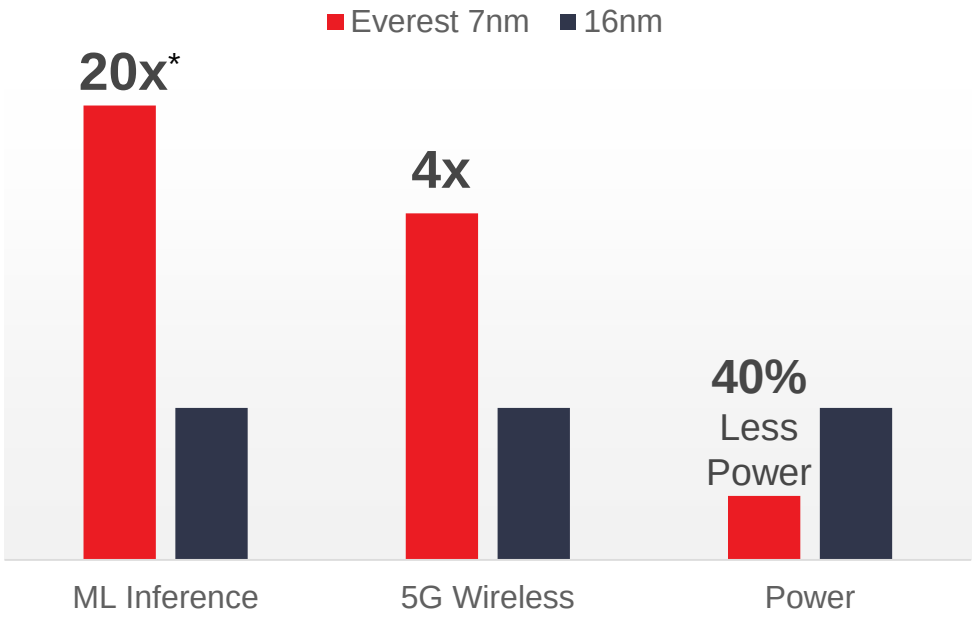
- **20x** ML Inference Performance
- **4x** 5G Communications Bandwidth
- **112G** Transceivers



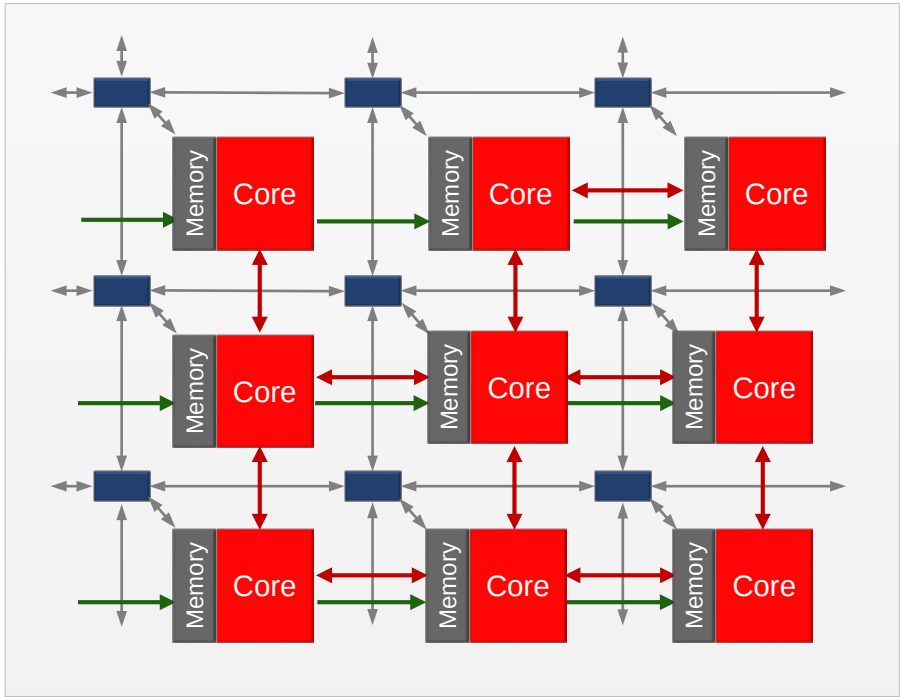
XILINX.

New HW/SW Programmable Architecture

Application-level Performance Enabled by SW Programmable Engine



Array Architecture



Compute Efficiency

> Domain Specific Engine
> Greater Compute Density
> Xilinx 7nm Everest

Multiple Applications

> ML Inference for Cloud DC
> Wireless 5G: Radio, Baseband
> ADAS/AD Embedded Vision
> Wired: DOCSIS Cable Access

Heterogenous Architecture

> PE Throughout and Efficiency
> PL Flexibility
> Customized Memory Hierarchy

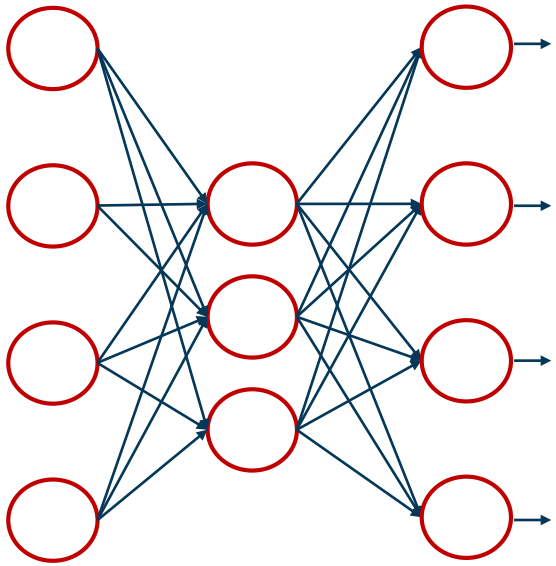
SW Programmable

> SW Programmable (e.g., C/C++)
> Compile, Execute, Debug
> Increased Productivity

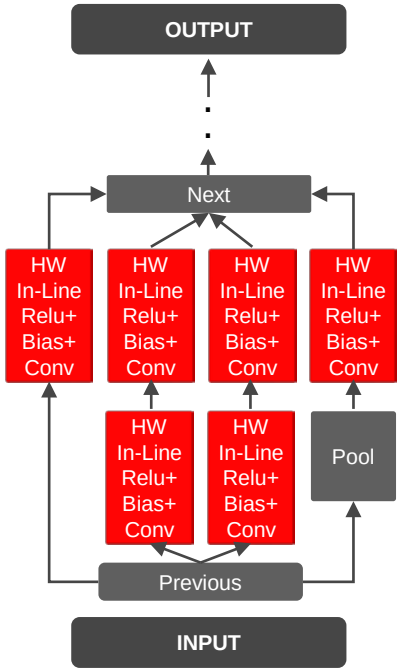
XILINX.

xDNN: Adaptable Overlay DNN Processor for Xilinx FPGA

Trained Model

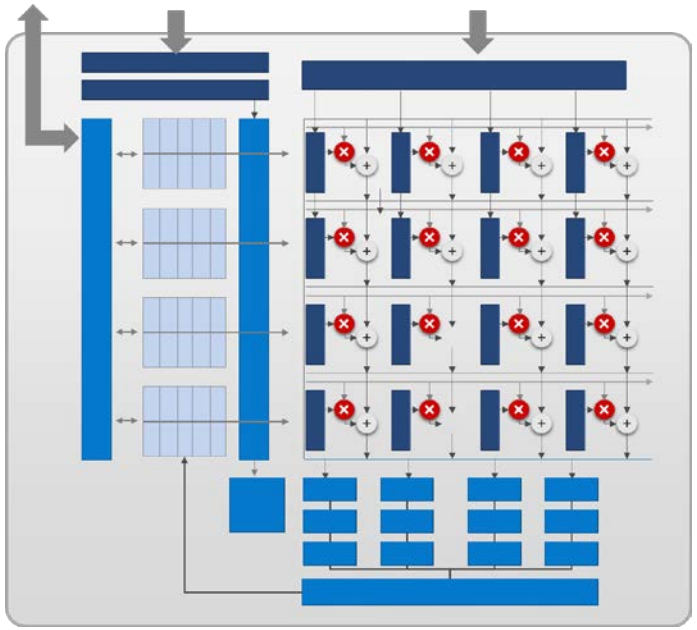


Compiler + Runtime



Overlay DNN Processor

60-80% Computation Efficiency



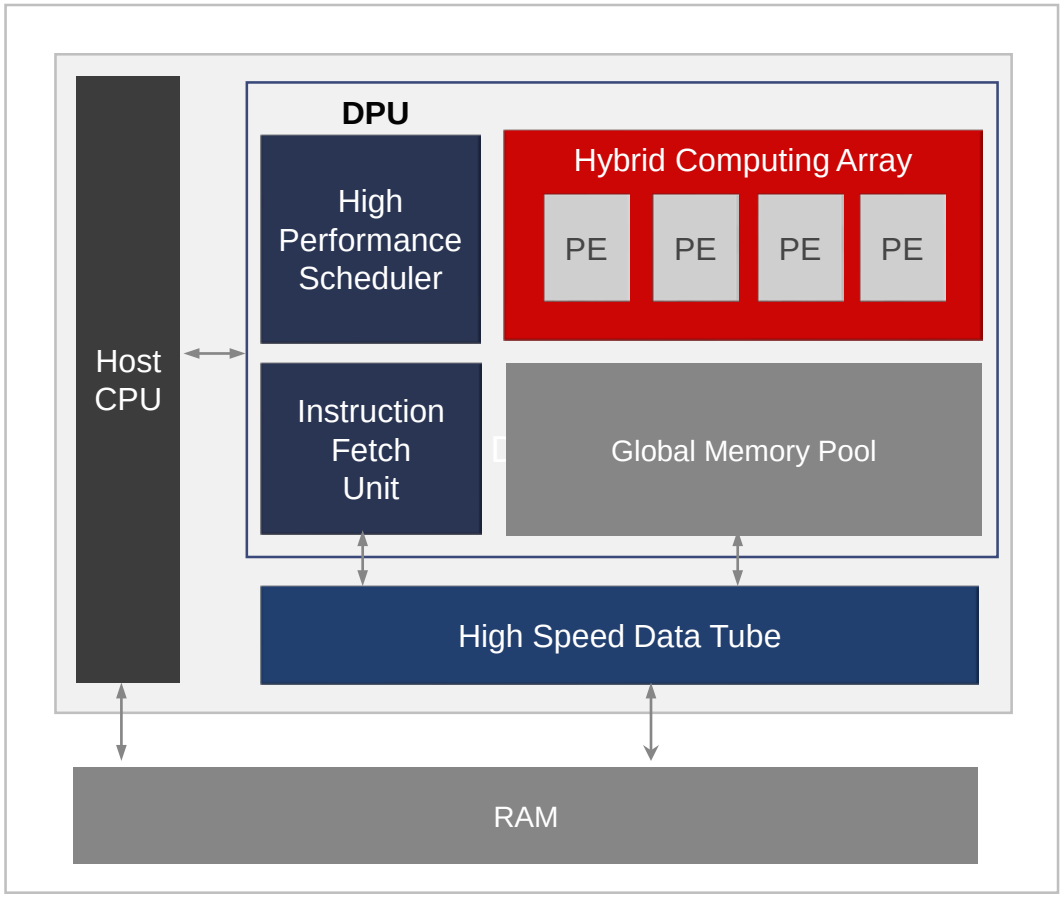
Low Latency & High Throughput
(Batch = 1)

No FPGA Expertise Needed

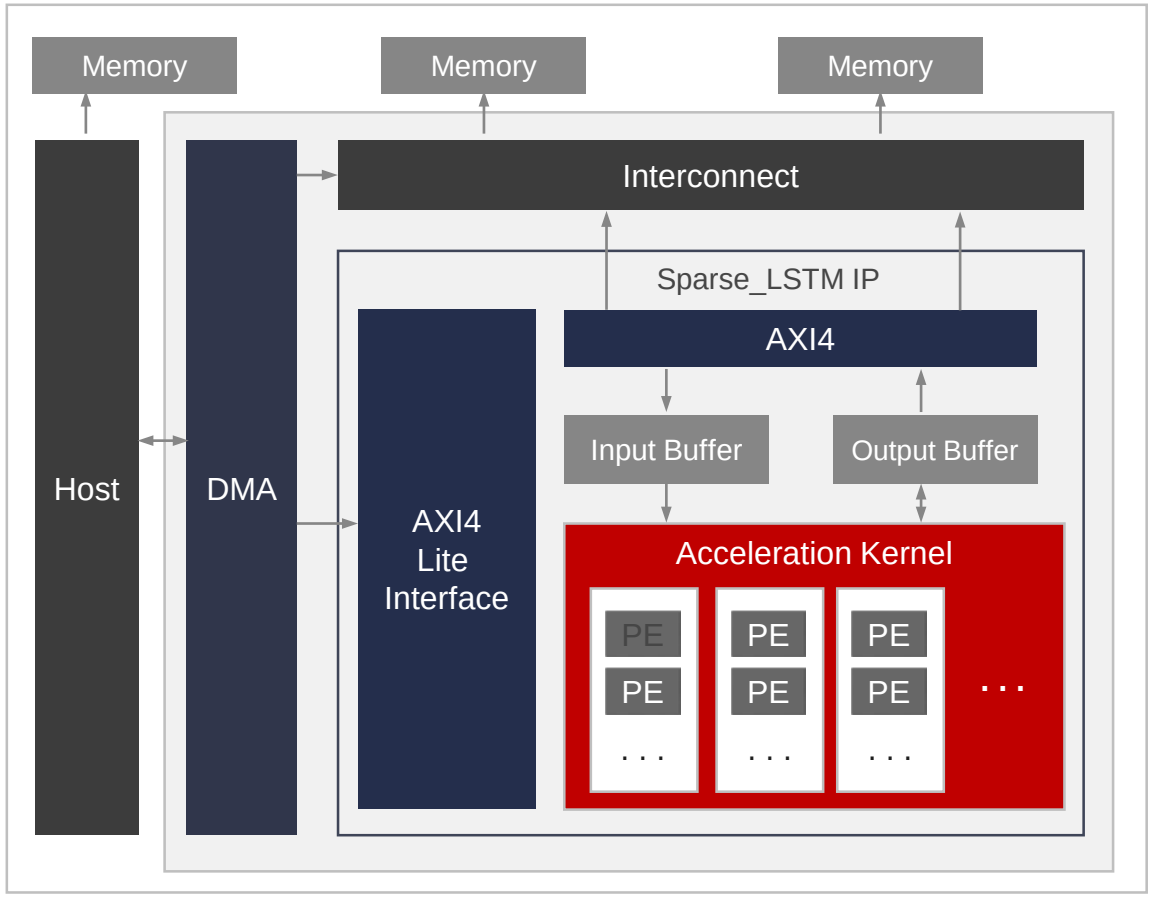
XILINX.

Soft Overlay Architectures for ML

DeePhi Aristotle Architecture



DeePhi Descartes Architecture



Building the Adaptable, Intelligent World

XILINX