Adaptable Computing
The Future of FPGA Acceleration

Dan Gibbons, VP Software Development

June 6, 2018

Adaptable Accelerated Computing

Three Big Trends

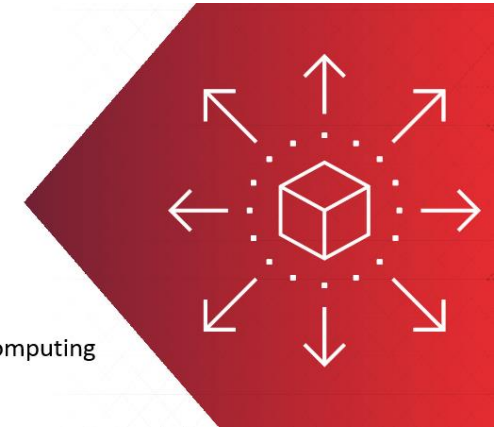
01

Explosion of Data

> 90% unstructured

> Video & image content

> Needs higher throughput & real-time computing



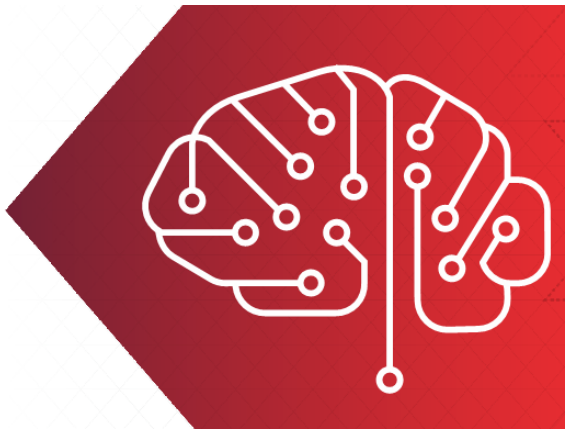
02

Dawn of AI

> Adoption across all industries

> Injecting new intelligence into apps

> From endpoints to edge to cloud



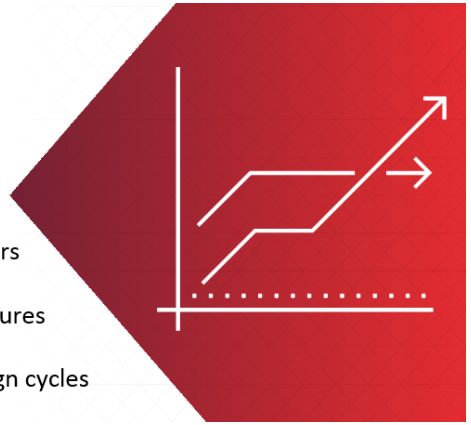
03

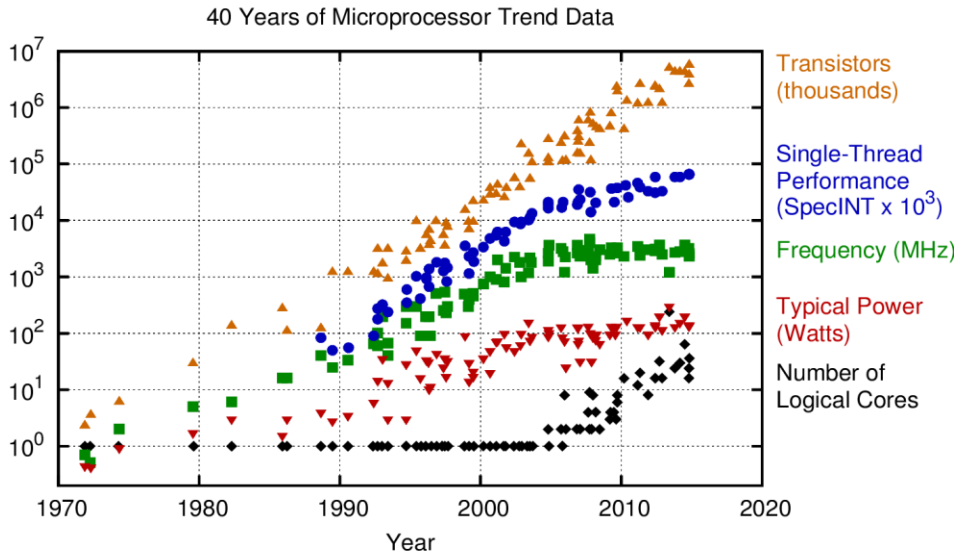
Computing After Moore's Law

> Heterogeneous computing with accelerators

> Breadth of apps require different architectures

> Speed of innovation outpacing silicon design cycles



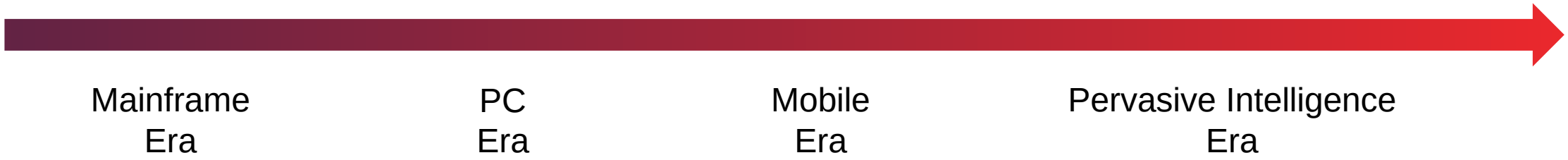


Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2015 by K. Rupp

XILINX

The Evolution of Computing

Trend to Heterogeneous Architectures with Acceleration of New Workloads



XILINX

The Need for Adaptable Intelligence

The intelligent connected world needs adaptable accelerated computing.

Everything Intelligent & Connected

Deployed at Global Scale

Dynamic Needs & Rapid Innovation

Why it Matters – Personalized Medicine Example

Whole genome diagnosis to treat critically ill newborns

Analysis reduced from
1 day to 20 minutes

Patient-specific genomics dynamically optimized

Medical data and research needs to be securely accessed across the globe

GUINNESS WORLD RECORDS™ RECORD HOLDER

Rady Children's Hospital San Diego

edico genome

XILINX

The FPGA Advantage

The FPGA Advantage for Machine Learning Inference

Adaptive Architecture

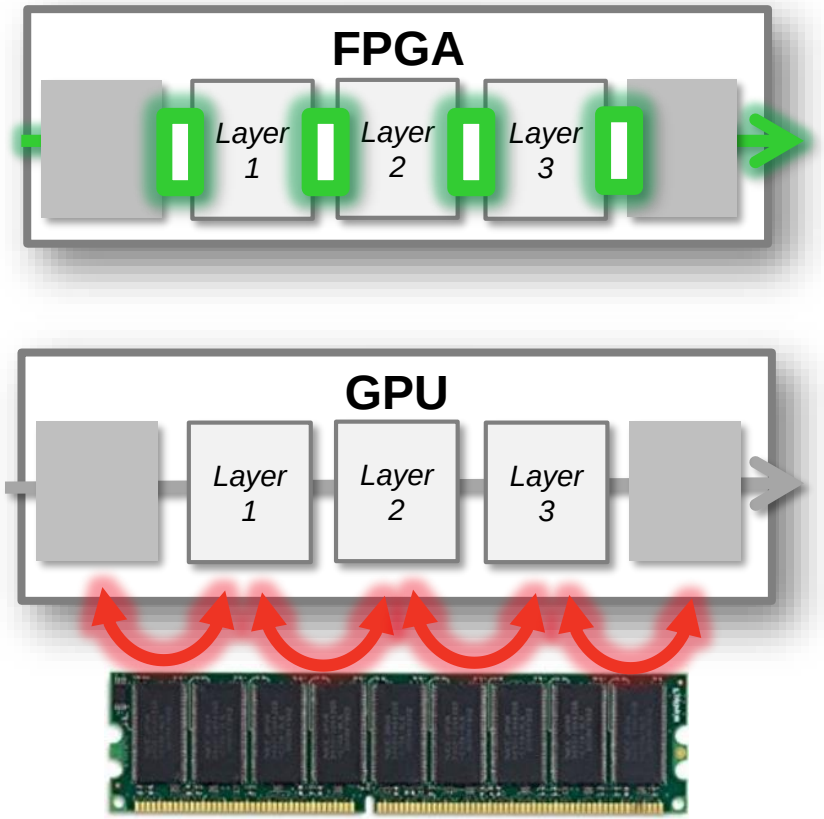
> Customer dataflow, precision, optimizations

Custom Memory Hierarchy

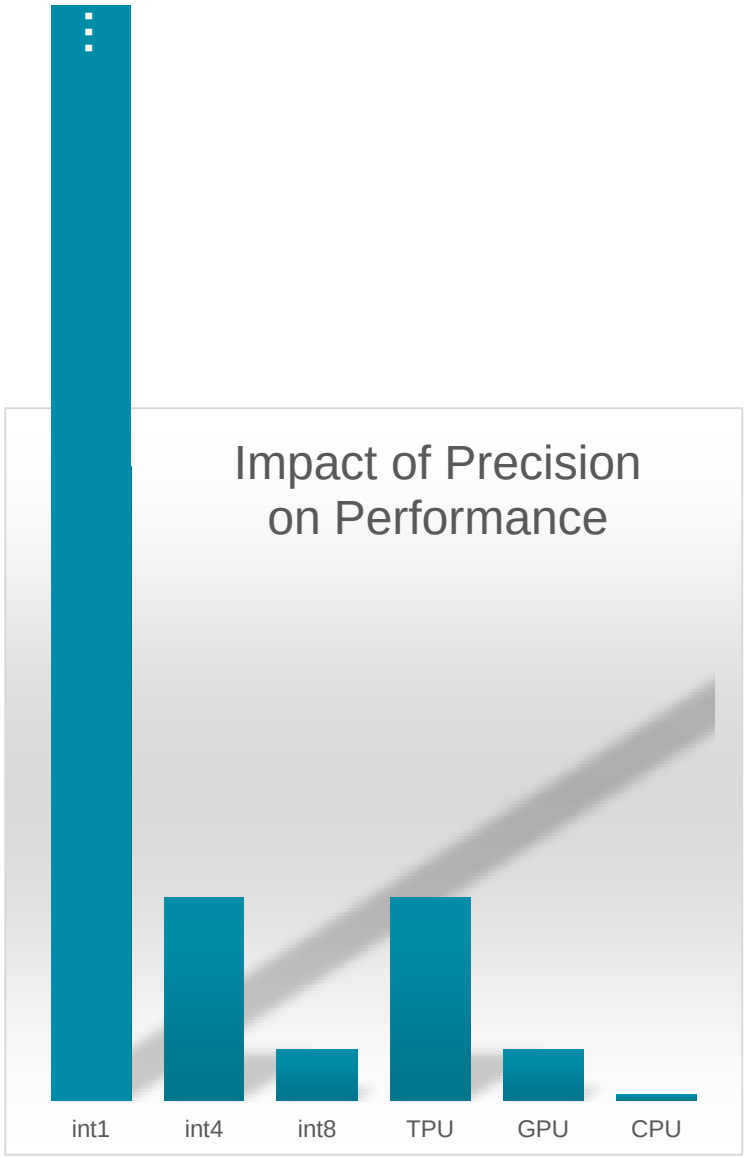
> Keeps data inside vs. external memory bottleneck

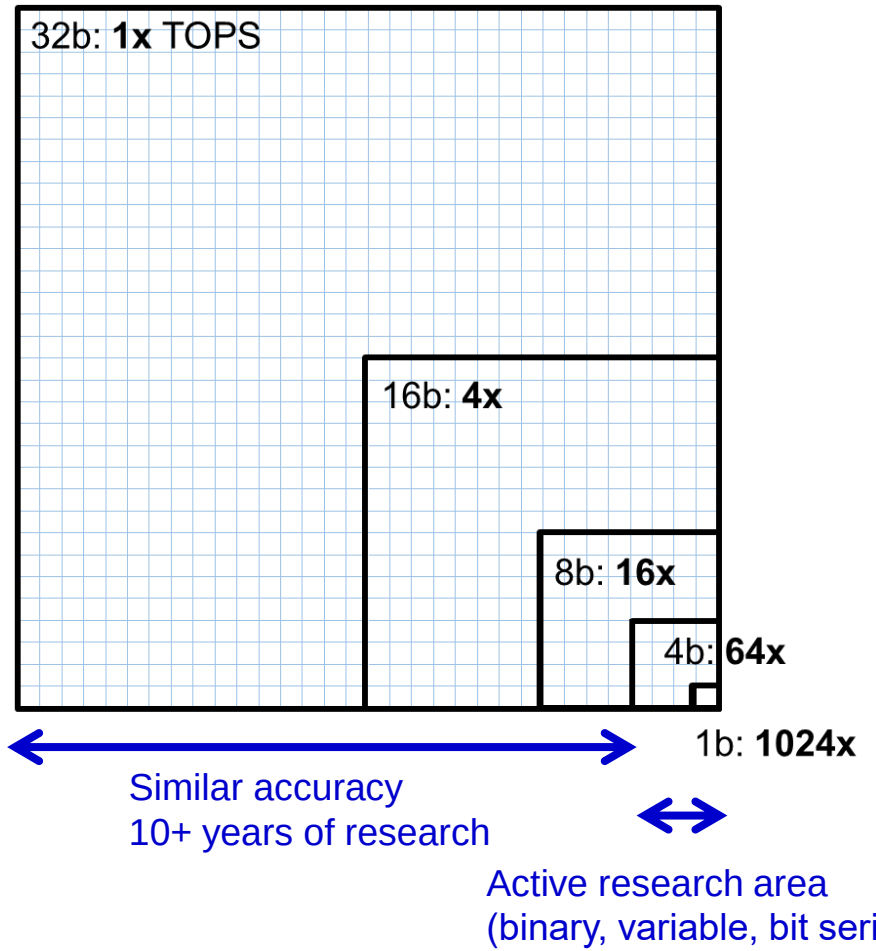
Workload + ML Inference

> Unleashes the power of on-chip system dataflow



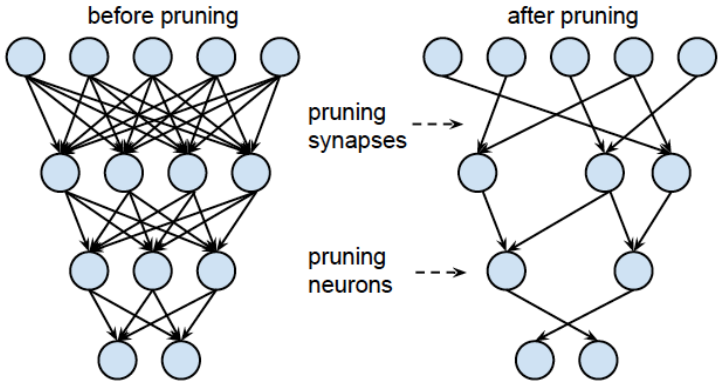
Powerful FPGA Optimizations: Precision

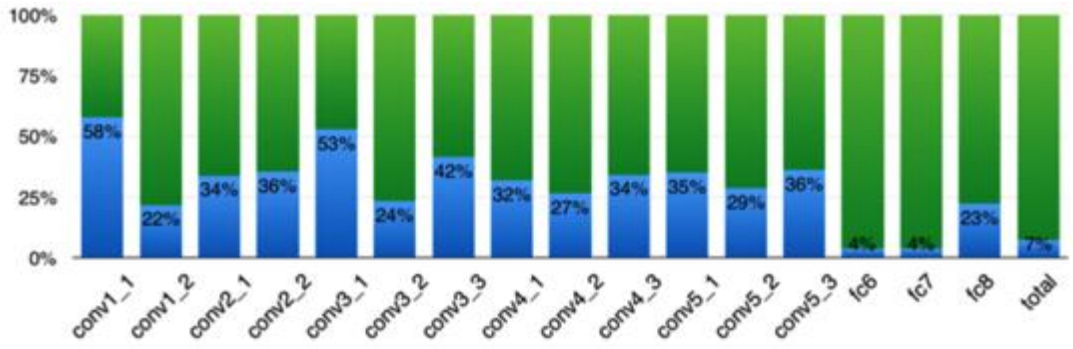




XILINX

Powerful FPGA Optimizations: Compression





30x to 50x compression rate without impacting accuracy (AlexNet)

DEEPHI Compression

A Survey of Model Compression and Acceleration for Deep Neural Networks

Yu Cheng, Duo Wang, Pan Zhou, *Member, IEEE,* and Tao Zhang, *Senior Member, IEEE*

With the world's leading research in neural network model compression, DeePhi develops DECENT (DEep ComprEssioN Tool). It firstly introduces pruning, quantization, weight-sharing and Huffman encoding to reduce model size from 5x to 50x without loss of accuracy.

XILINX

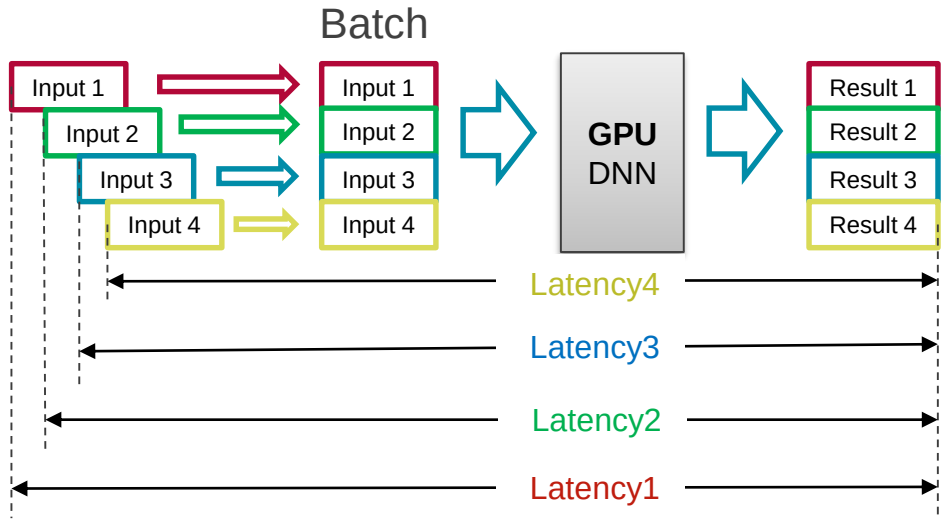
FPGA Advantage: Deterministic Latency

"Batch" Inference

> Parallel batch of data to feed SIMD

> High batch => low latency, higher throughput

> Lower compute efficiency at low batch

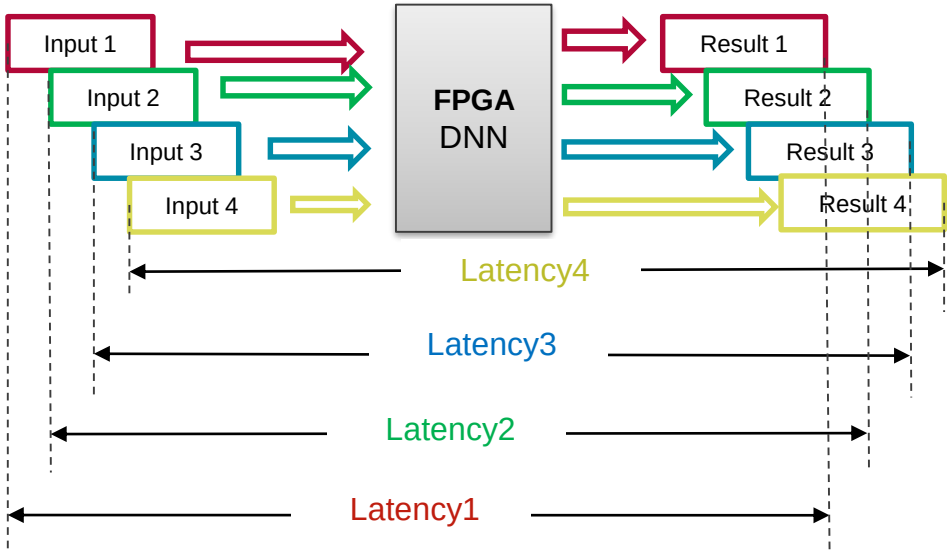


"Batch-less" Inference

> Low and deterministic latency

> High throughput regardless of batch size

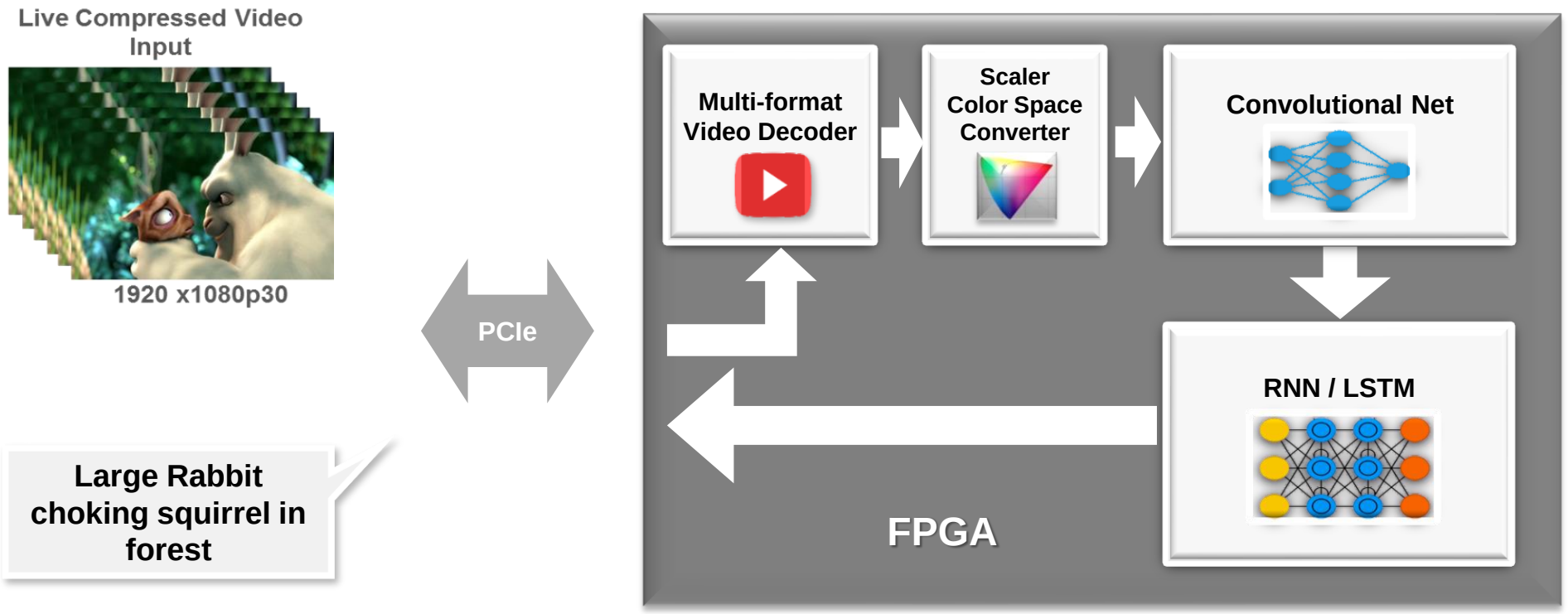
> Consistent compute efficiency



XILINX

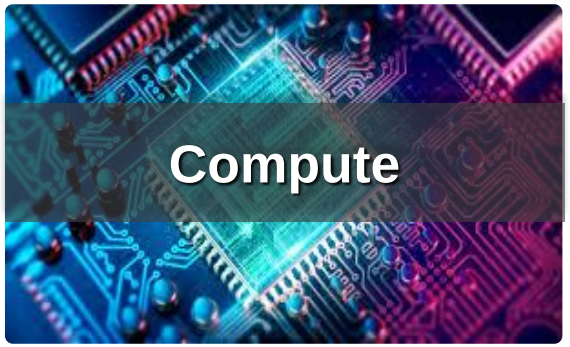
ML Inference Integrated with Other Workloads

Live video summary using CNN & RNN



XILINX

Adaptable Compute Use Cases Across the Datacenter



- ✓ ML Inference
- ✓ Database / Big Data Analytics
- ✓ Video Transcoding
- ✓ Financial Services Analytics
- ✓ Genomics

- ✓ Compression
- ✓ Encryption
- ✓ Key-Value Store
- ✓ ML Inference
- ✓ Database / Big Data Analytics

- ✓ IPSec/SSL
- ✓ OVS Offload
- ✓ Bare Metal Services
- ✓ Security
- ✓ Monitoring

XILINX

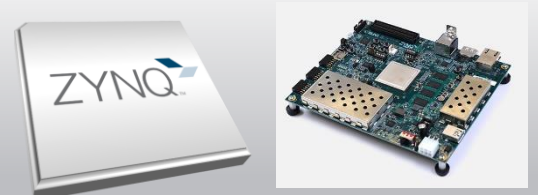
Zynq SoCs: Adaptable Computing on the Edge

Caffe

OpenCV

DEEPHi DNNDK

SDSoC Environment



ZCU102 Development Platform

- ✓ 4 CNN Models
- ✓ 3 Live Inputs + File IO
- ✓ Under 10 Watts!

XILINX

Xilinx Enables Adaptable Accelerated Computing

XILINX ‘FPGA as a Service’ goes wide

XILINX

Towards Software as a Service (SaaS)

Enterprise

awsmarketplace

SaaS

SW API

Accelerated SW

SDAccel

FaaS

MAYO CLINIC

illumina

edico genome

aws

NIMBIX

Tencent Cloud

Baidu

Alibaba Cloud

HUAWEI

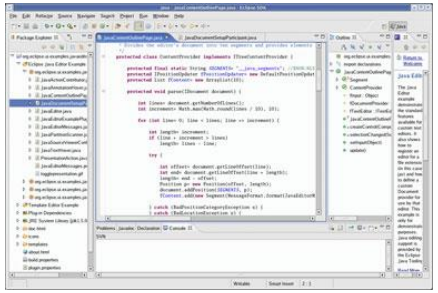
XILINX

SDAccel™ Environment Breakout in Programming for Acceleration

Optimal acceleration results requires platform performance, compiler efficiency and programming proficiency

High Performance Platform

Advanced Compiler

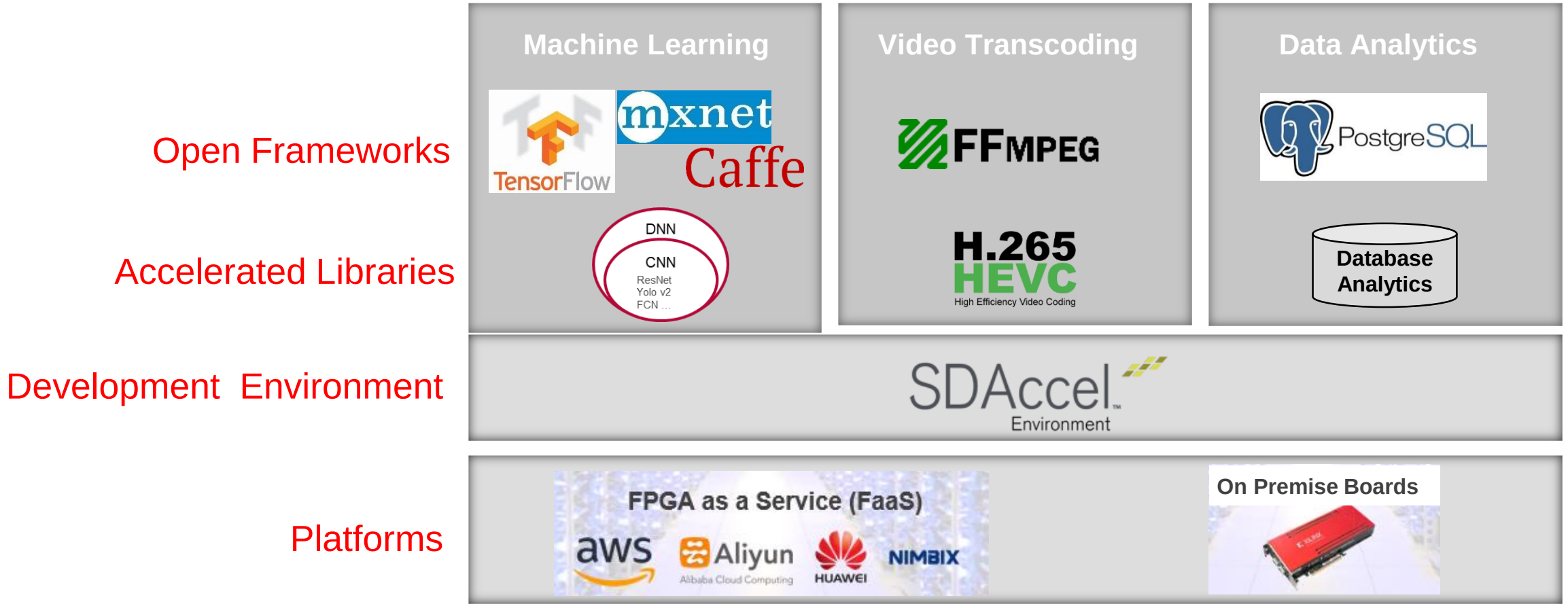


Productive IDE & Optimized Libraries

User Onboarding

XILINX

Rich Stack Integrated with Frameworks



XILINX

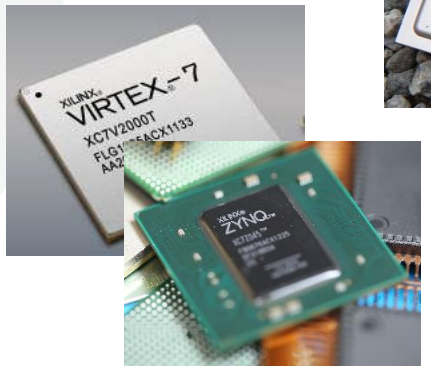
Transformation Through Innovation

World's
First FPGA

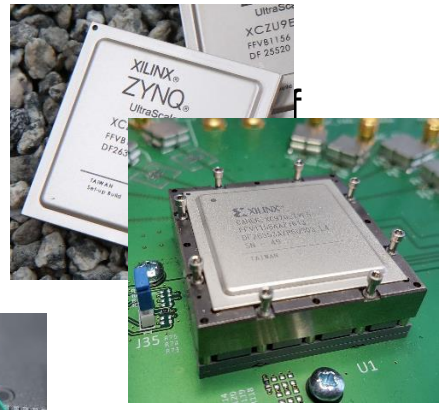
First
Virtex FPGA

Virtex-2 Pro

First 3D FPGA
& HW/SW
Programmable SoC



First
MPSoC & RFSoC



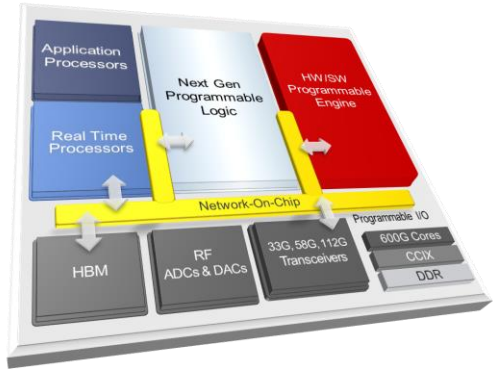
ACAP

1980 1990 2000 2010 2020

XILINX

The Era of Heterogeneous Computing Architectures is Here

FPGA's are uniquely suited for adaptable accelerated computing



Xilinx is leading the way with platforms, tools, applications and FaaS

SDAccel™ Environment

Now is the opportunity for application development and deployment

NIMBIX

XILINX