FPGAs – EPIC Benefits

Philip Leong
Director, Computer Engineering Laboratory
http://phwl.org/talks

THE UNIVERSITY OF SYDNEY

Computer Engineering Laboratory

- Focuses on how to use parallelism to solve demanding problems
  - Novel architectures, applications and design techniques using VLSI, FPGA and parallel computing technology
- Research
  - Nanoscale interfaces
  - Machine learning
  - Reconfigurable computing
- Collaborations
  - Consunet, DST Group
  - Intel, Xilinx
- Ex-students
  - Xilinx, Intel, Waymo

Overview

FPGAs

Applications

Our work

Overview

FPGAs

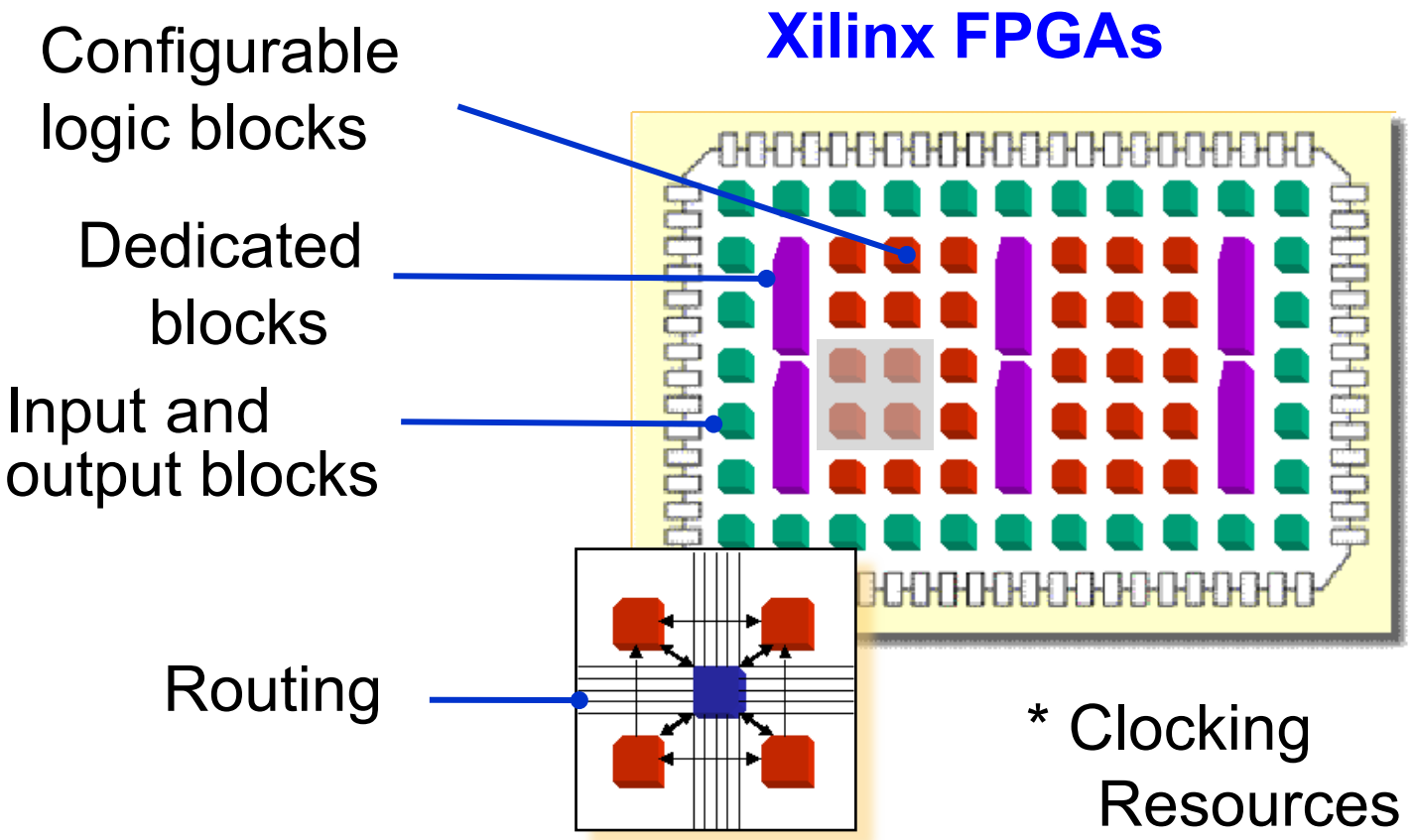
Applications

Our work

What is an FPGA?

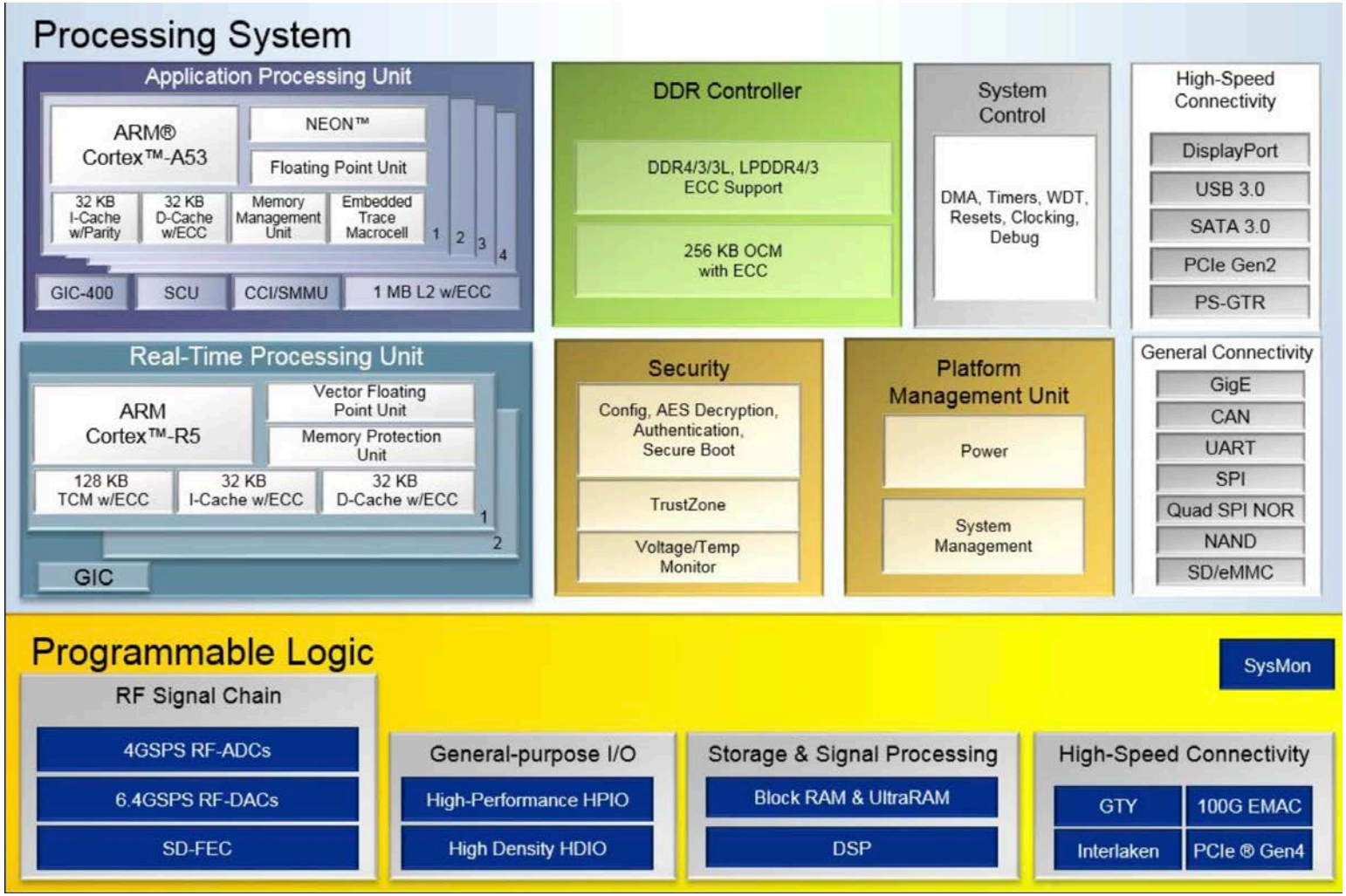
User-customisable integrated circuit

› Dedicated blocks: memory, transceivers and MAC, PLLs, DSPs, ARM cores



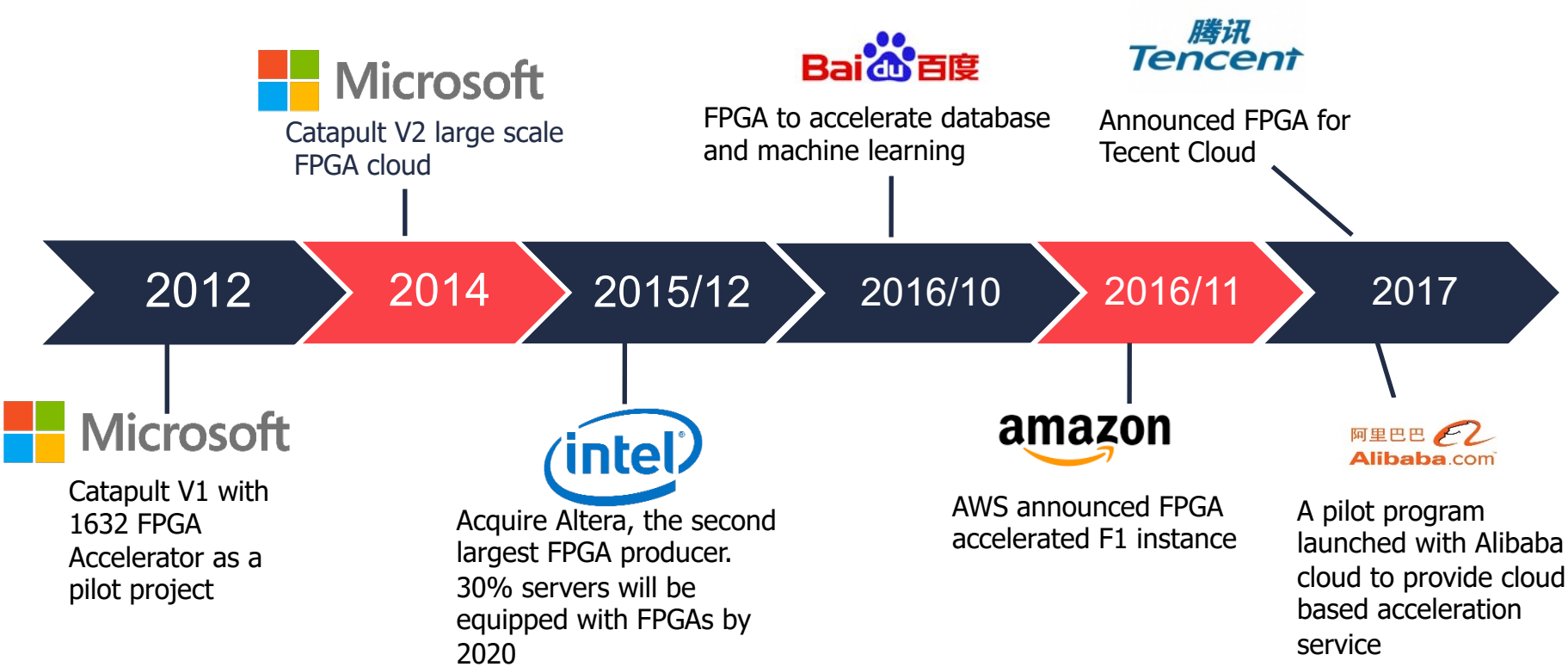
Source: Xilinx

Xilinx RFSoc Device



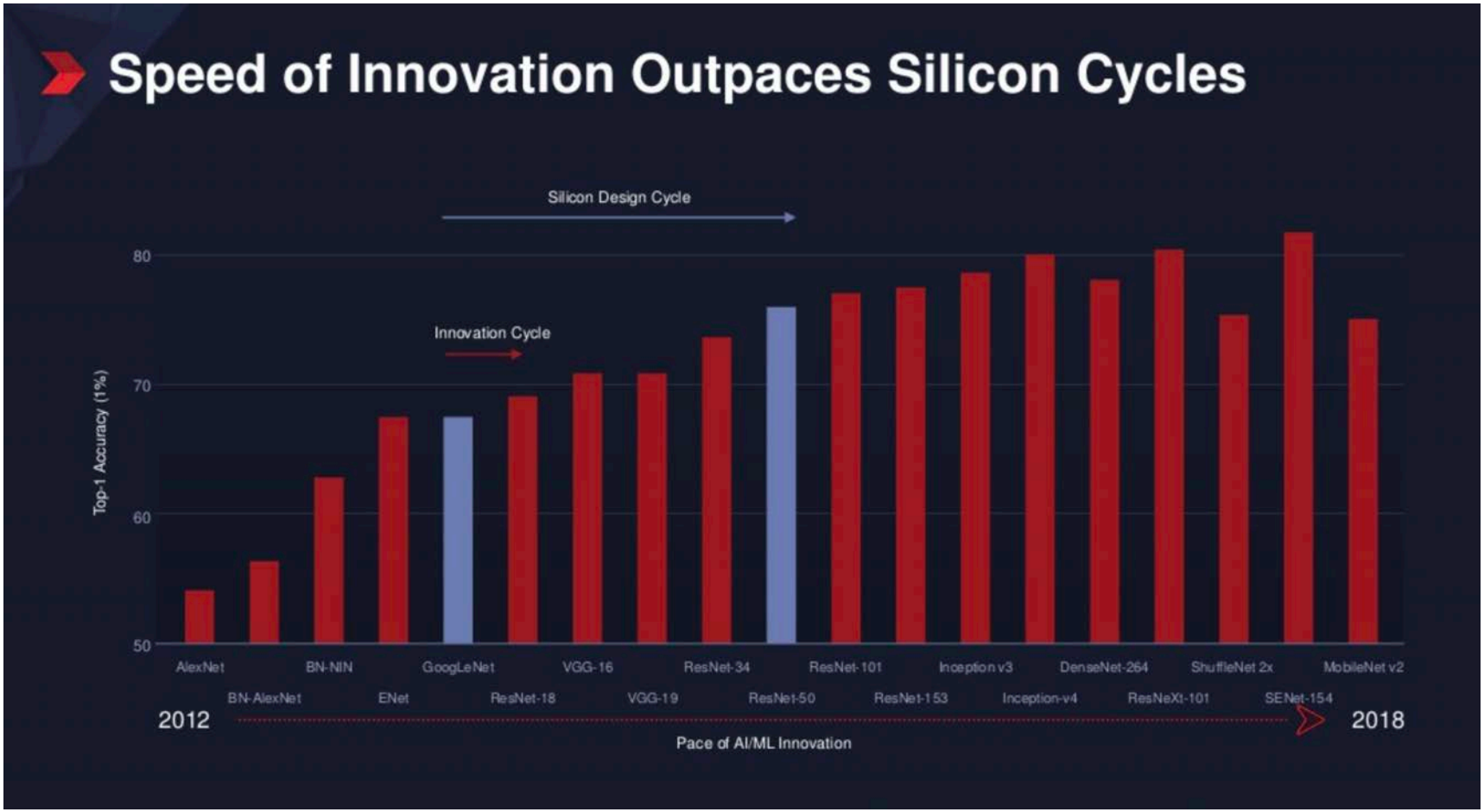
Source: Xilinx

Recent Uptake in Reconfigurable Computing



Source: CTAccel

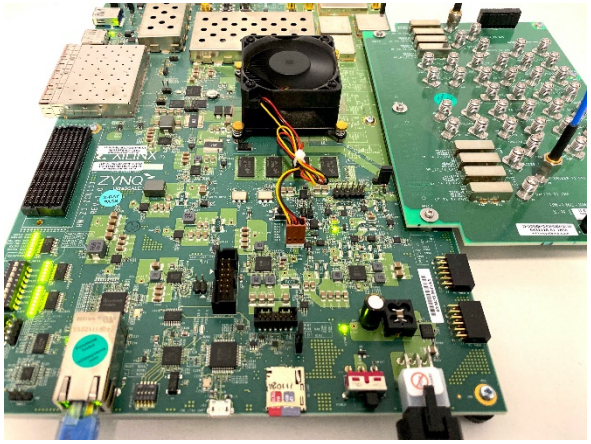
Speed of Innovation



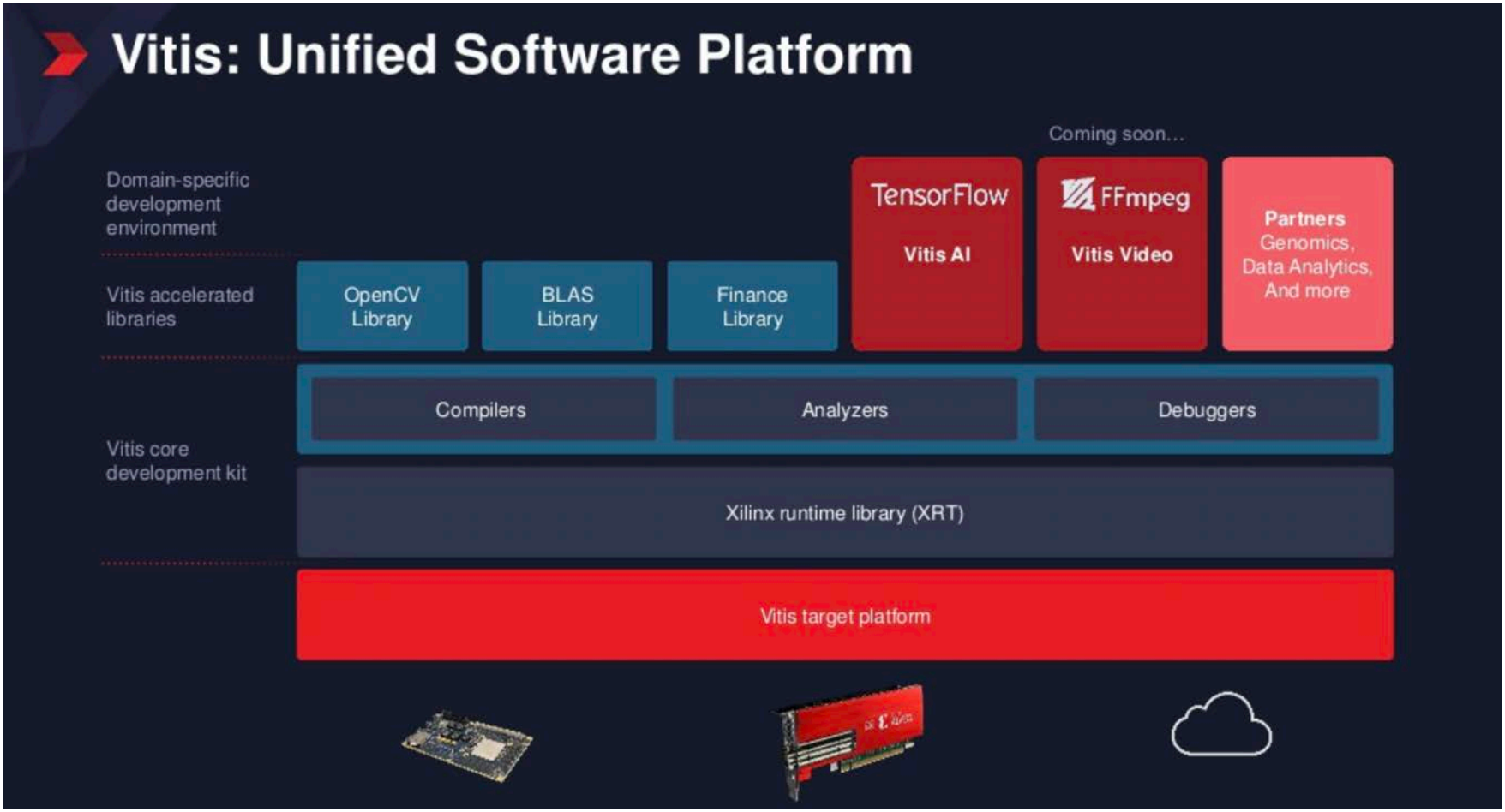
Source: Xilinx

Motivation for FPGAs (EPIC)

- FPGAs commercial off-the-shelf
- They offer an opportunity to implement complex algorithms with higher throughput, lower latency and lower power through
  - **E**xploration– easily try different ideas to arrive at a good solution
  - **P**arallelism – so we can arrive at an answer faster
  - **I**ntegration – so interfaces are not a bottleneck
  - **C**ustomisation – problem-specific designs to improve efficiency (power, speed, density)



Unified Environment 2019



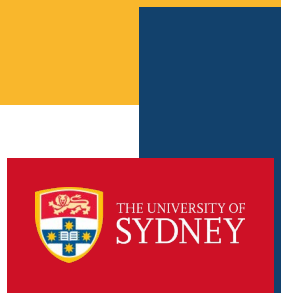
Source: Xilinx

Overview

FPGAs

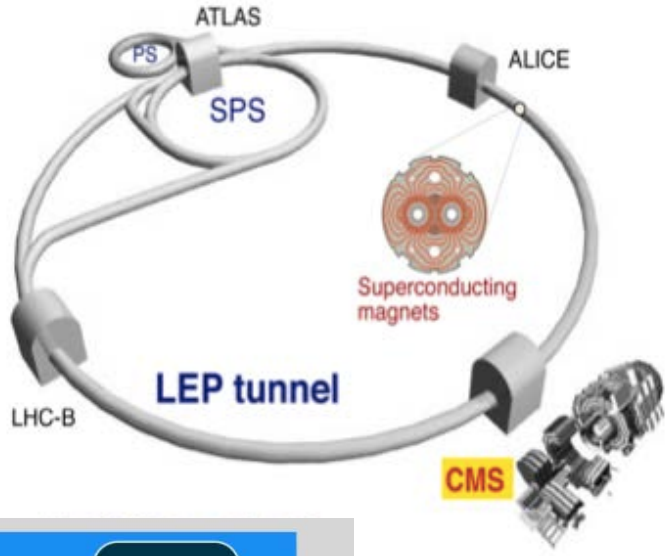
Applications

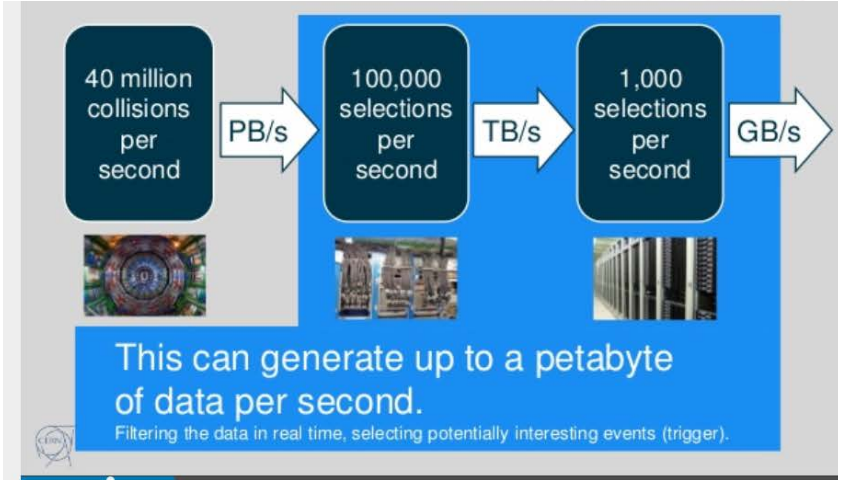
Our work



CERN Large Hadron Collider

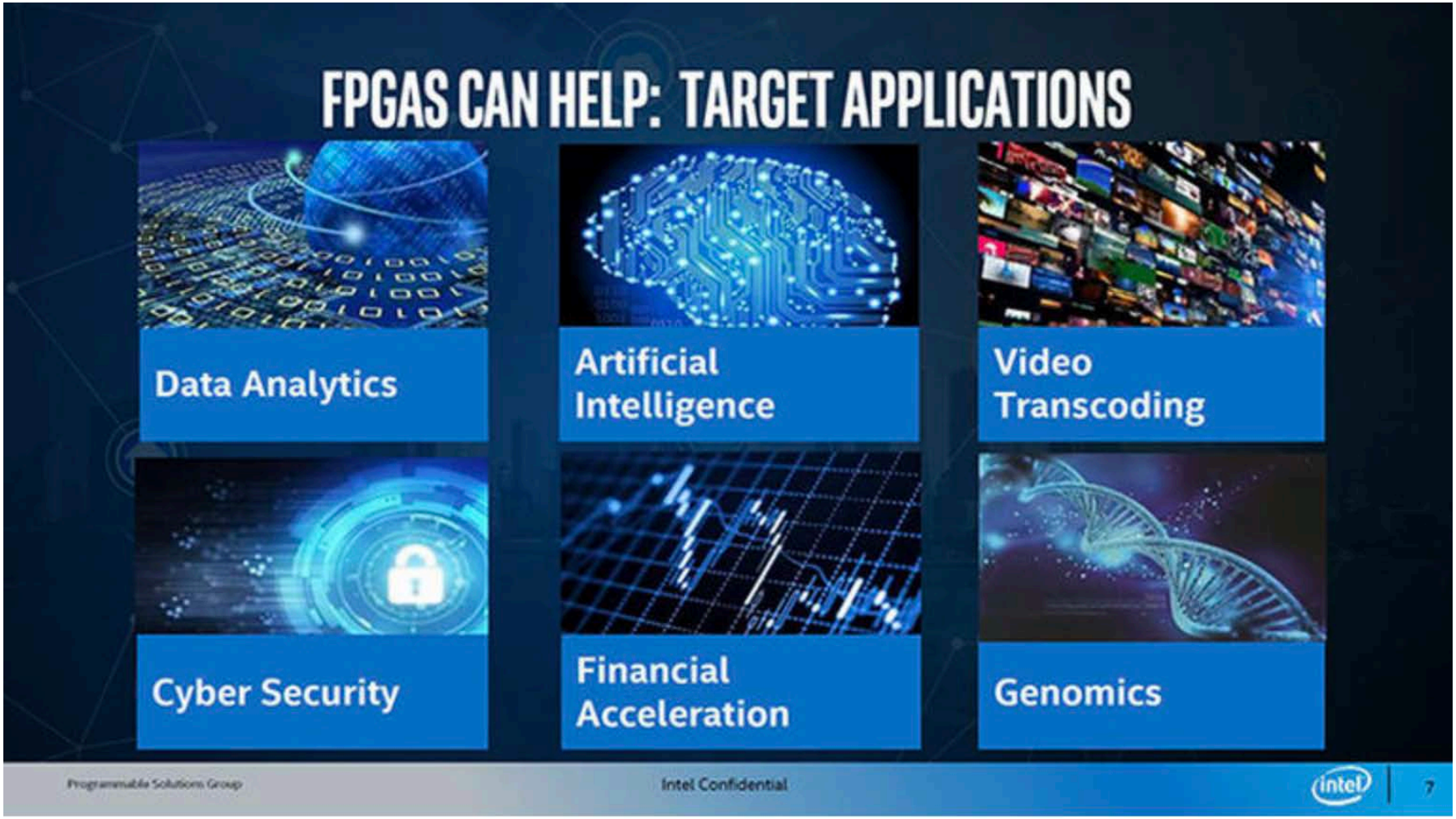
- Compact Muon Solenoid
  - Few interesting events ~100 Higgs events/year
  - 1.5Tb/s real-time DSP problem
  - (2014) More than 500 Virtex and Spartan FPGAs used in real-time trigger
  - (2019 doing FPGA-based DNN inference using Vivado HLS)





Source: Geoff Hall, Imperial College

Target Commercial Applications

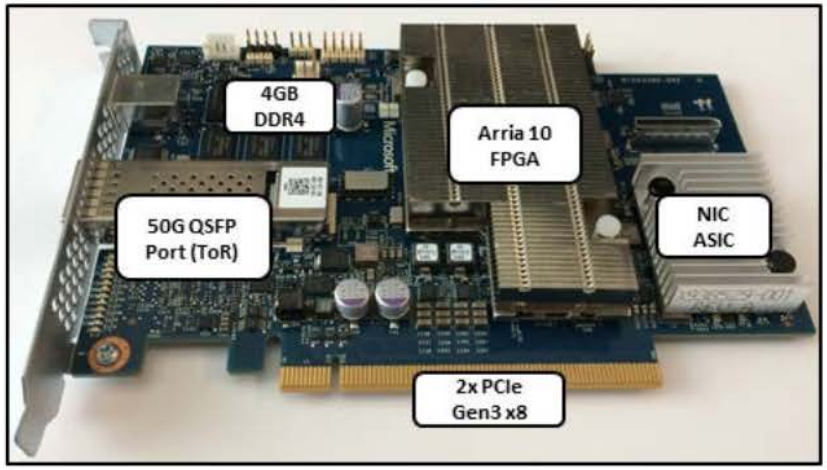


Source: Intel

Microsoft Project Catapult

› Uses FPGAs for DNNs, Bing search, and software defined networking (SDN) acceleration to reduce latency, while freeing CPUs for other tasks

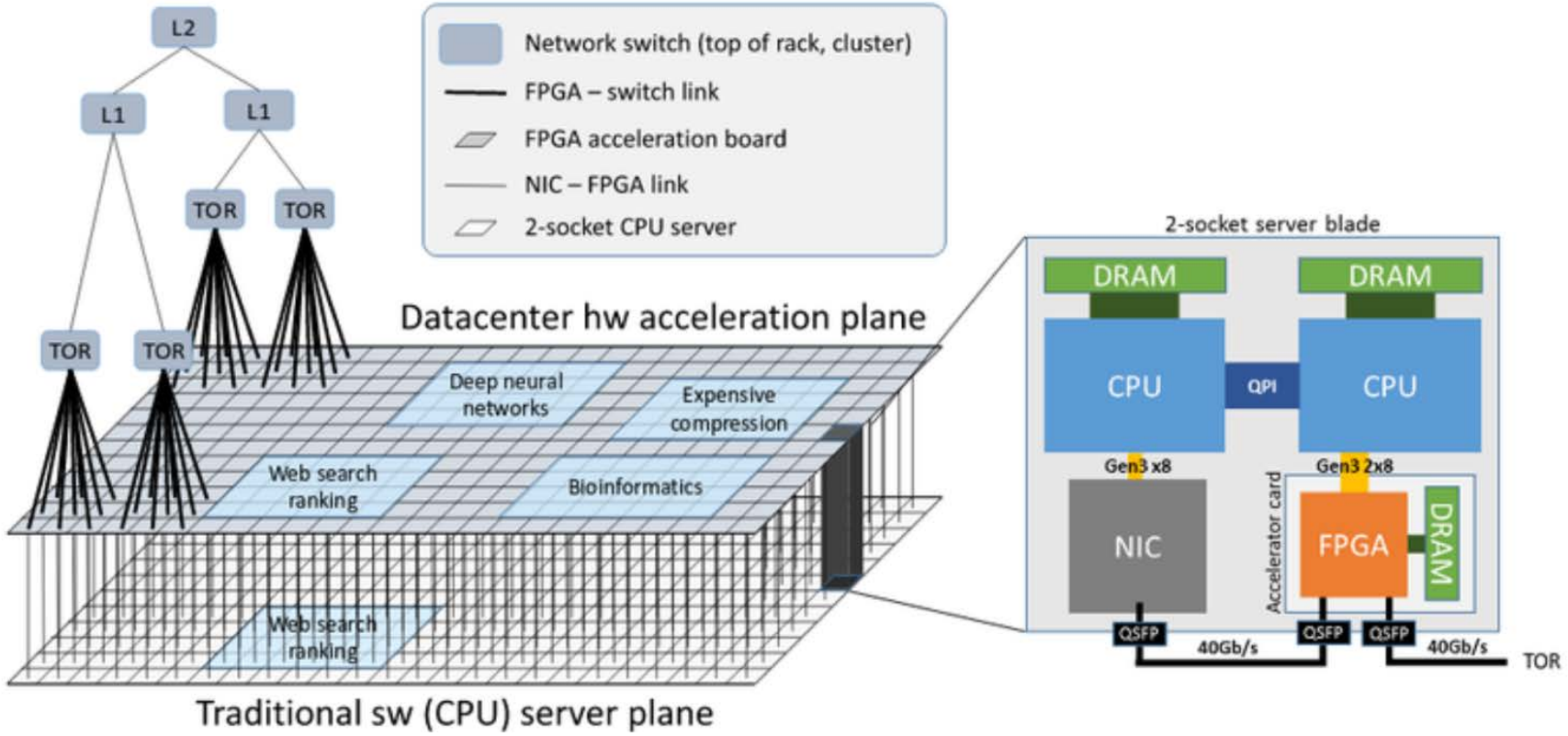
- 2010: MSR study FPGAs to accelerate Web search
- 2012: Project Catapult's scale pilot of 1,632 FPGA servers deployed
- 2013: Bing decision-tree algorithms 40x faster than CPUs
- 2015: FPGAs deployed at scale in Bing and Azure datacenters (> 1M) - enabled 50% ↑ throughput, 25% ↓ latency.



Source: Microsoft

Microsoft Azure Cloud Network

World's fastest cloud network



Source: Microsoft

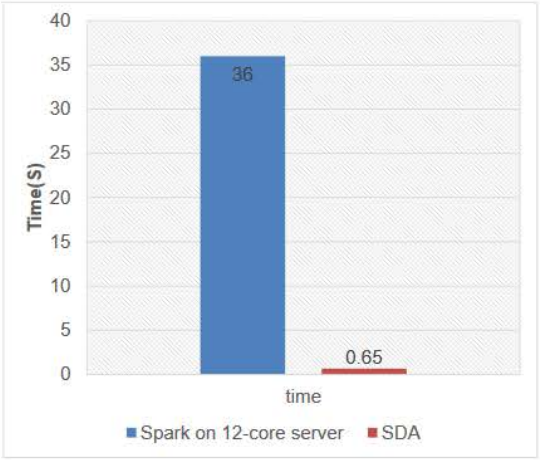
Baidu

› Accelerator for SQL Queries (40% of their data analysis)

Total data:	~1EB
Processing data :	~100PB/day
Total web pages:	~1000 Billion
Web pages updated:	~10Billion/day
Requests:	~10Billion/day
Total logs :	~100PB
Logs updated:	~1PB/day

Evaluation - real case query

- TPC-DS scale = 10 , query3
- Execution time
  - 55x

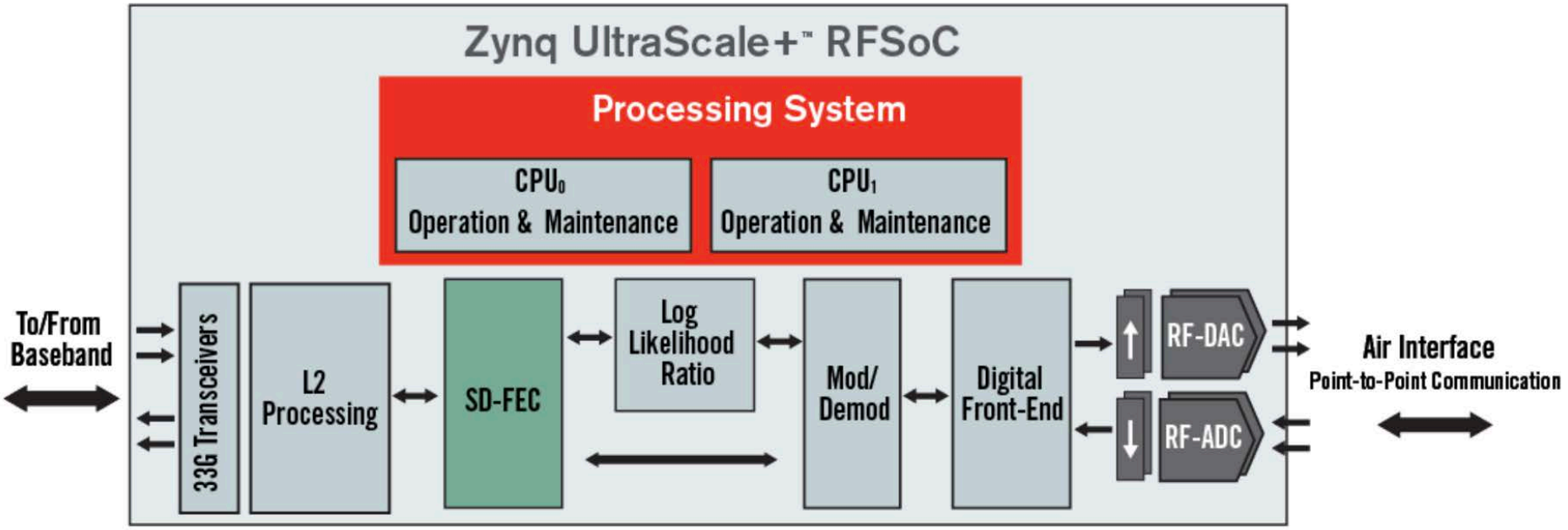


Source: Baidu 2016 Hot Chips

Mobile Backhaul

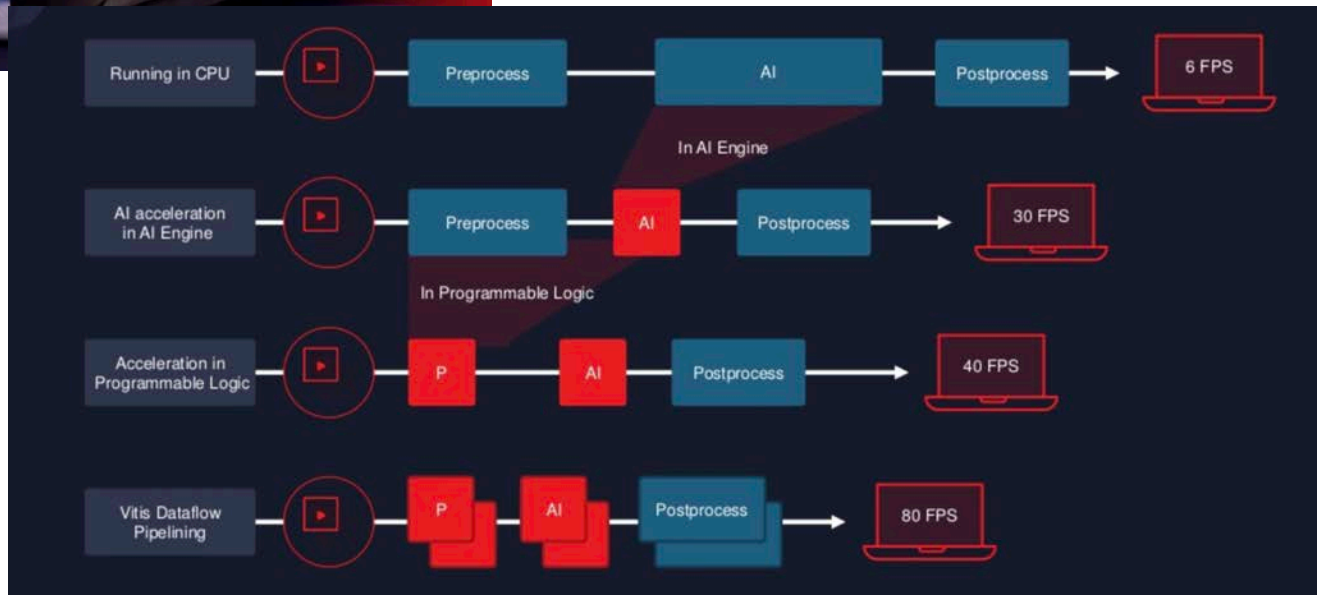
Key Zynq UltraScale+ RFSoC Benefits:

- Integrated Direct RF data converters for 4x4 TX/RX mobile backhaul architectures
- Multi-Level LDPC codec (SD-FEC) to meet 5G standards and support for custom codes
- Turbo Decode (SD-FEC) for 4G LTE-Advanced and 4G LTE Pro
- DSP48-rich fabric (6,620 GMACs) provides high-performance filtering and encoding/decoding
- 33 Gb/s transceivers for 12.2G CPRI and expansion into 16G & 25G CPRI



Source: Xilinx 2019

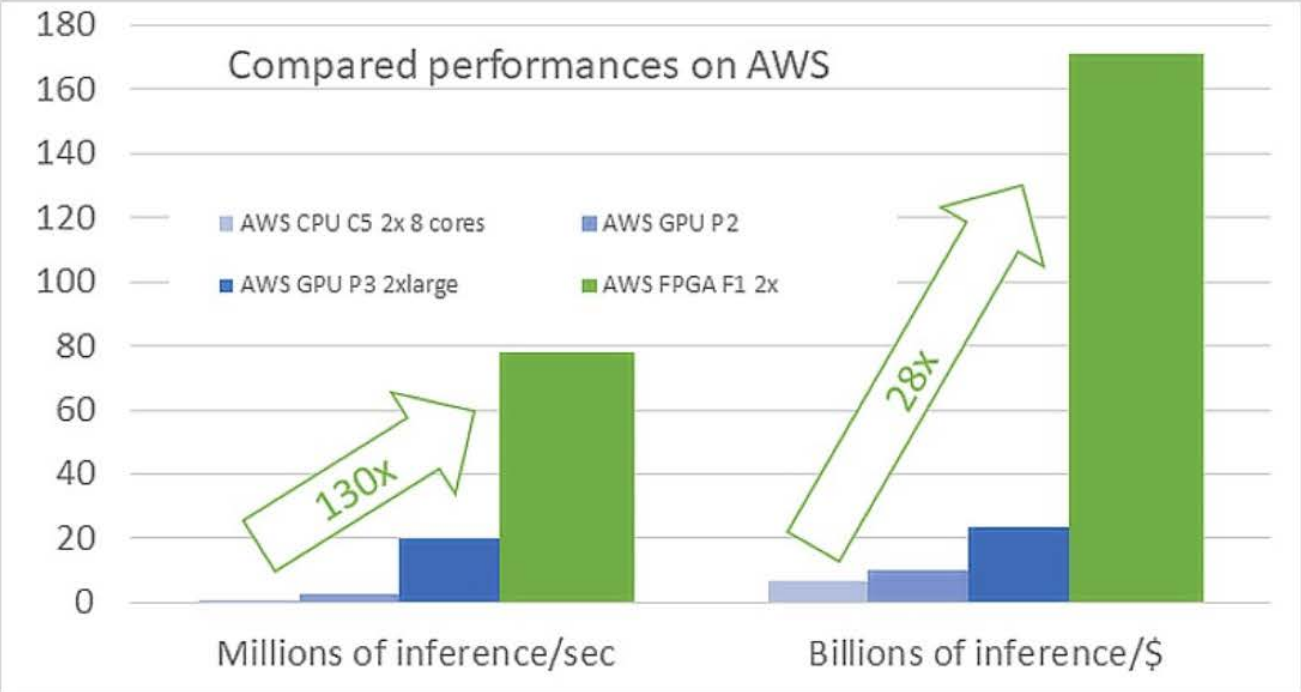
Smart City Example



Source: Xilinx

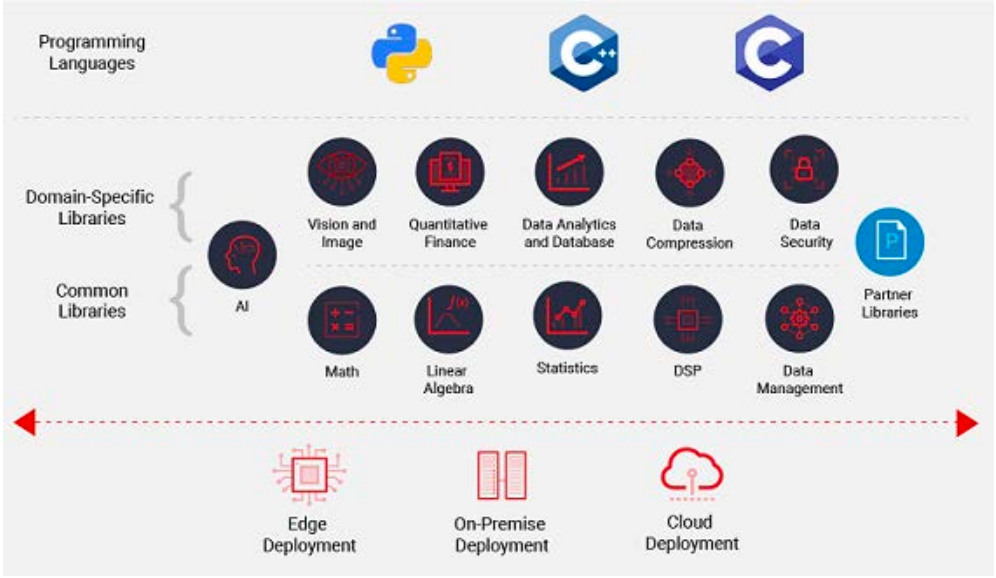
Amadeus

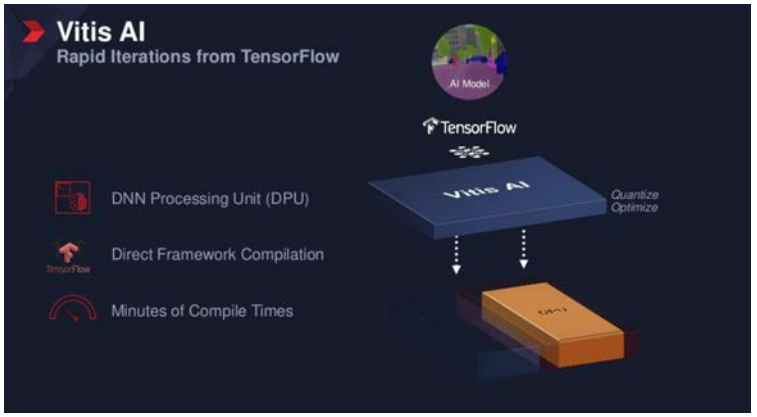
- Amadeus IT Group S.A adjusted profit €1.27B in 2019
- Accelerated inference of gradient boosted decision trees for search queries and quantified cost

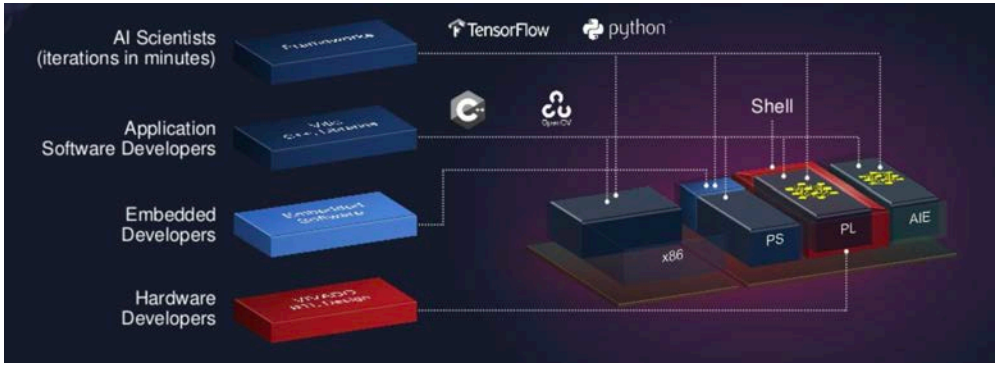


Source: VLDB 2020 http://www.vldb.org/pvldb/vol13/p71-owaida.pdf

Xilinx Vitis Unified Software Platform







https://github.com/Xilinx/Vitis_Libraries

Source: Xilinx

Vitis Data Analytics Library

XILINX Applications Products Developers Support About

Vitis Data Analytics Library

2020.1

Search docs

Library Overview
Requirements
License
Trademark Notice
Release Note

L1 User Guide
L1 Module User Guide

L2 User Guide
L2 Module User Guide

Benchmark Result
Benchmark Result
Performance Data

Random Forest Classification Training

Dataset:

1 - HEPMASS (https://archive.ics.uci.edu/ml/datasets/HEPMASS)

2 - HIGGS (https://archive.ics.uci.edu/ml/datasets/HIGGS)

Dataset	Sample Num	Tree Depth	Tree Num	End-to-End (s)	Speedup	Thread num	Spark (s)
1	7000000	5	512	61.20	10.2	28	622.30
1	7000000	5	1024	121.20	15.3	16	1849.724
2	8000000	5	512	70.30	13.3	28	933.83
2	8000000	5	1024	138.84	15.5	16	2154

K-Means Clustering Training

Dataset:

1 - NIPS Conference Papers (http://archive.ics.uci.edu/ml/datasets/NIPS+Conference+Papers+1987-2015)

https://xilinx.github.io/Vitis_Libraries/data_analytics/2020.1/benchmark/result.html

Source: Xilinx

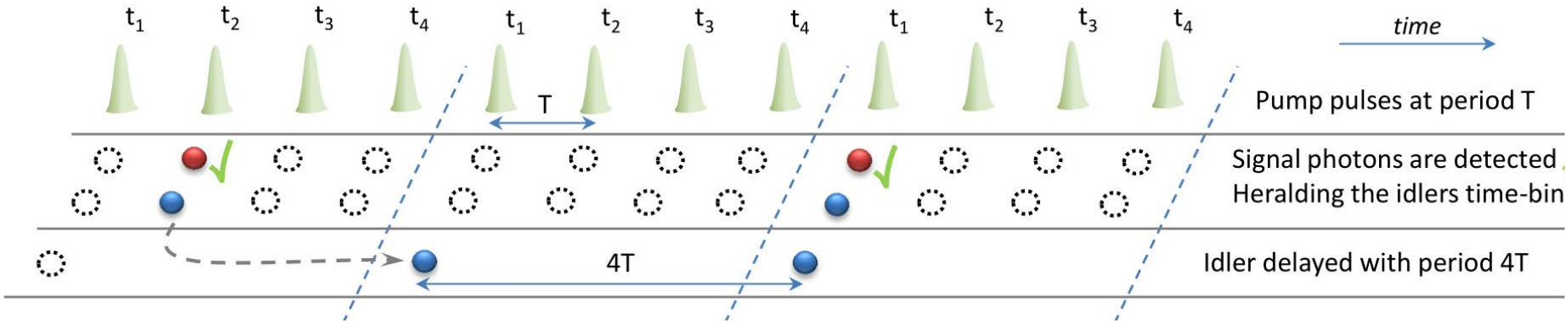
Overview

FPGAs

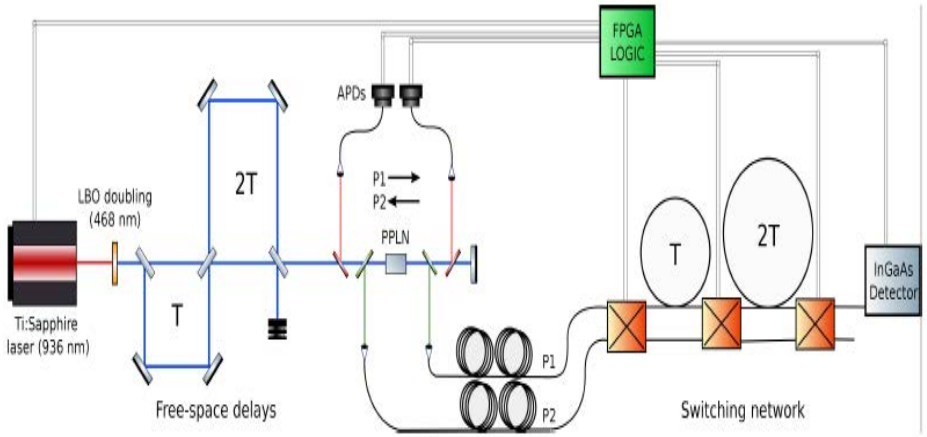
Applications

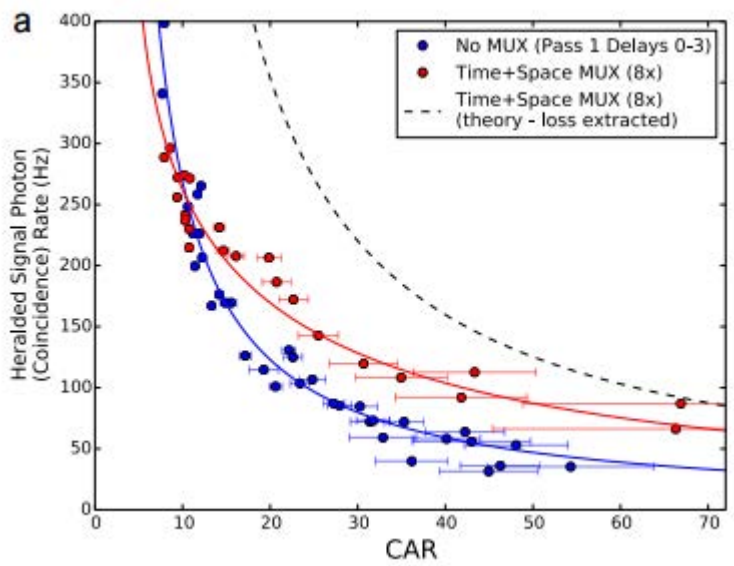
Our work

Time domain multiplexing of single photons



Initially expectation : Heralded single photon rate should enhance significantly without degrading coincidence to accidental ratio (CAR)





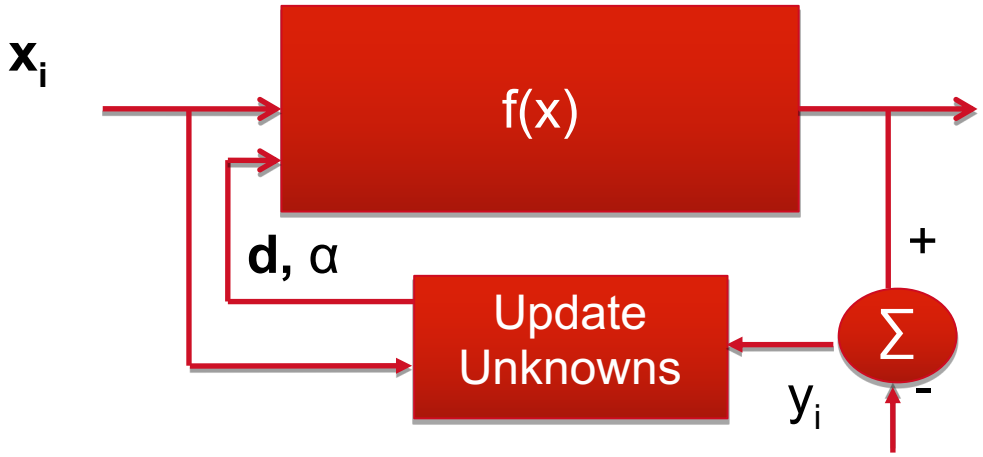
Enhancement : 33%~59%

THE UNIVERSITY OF SYDNEY

Time Multiplexing of Single Photons

Exploration: Kernel Methods

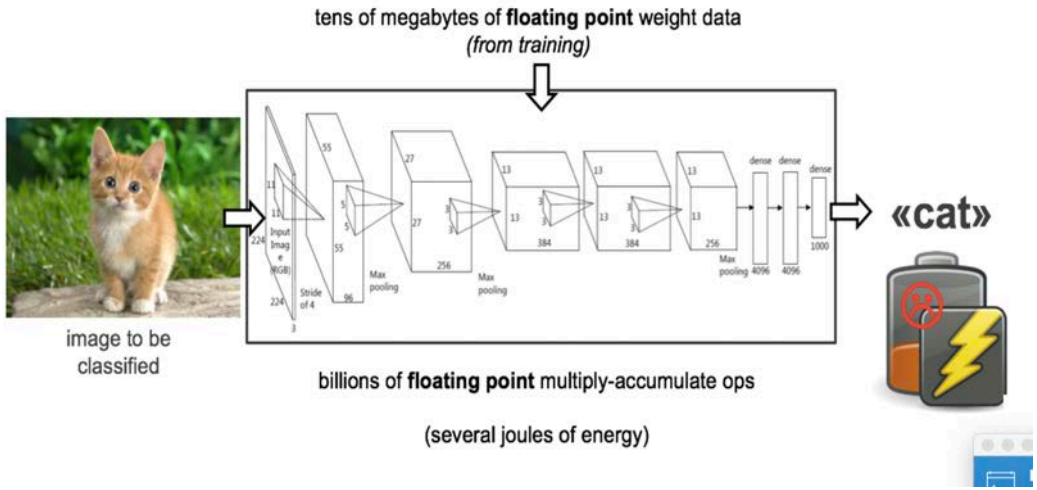
ARC Linkage with Exablaze

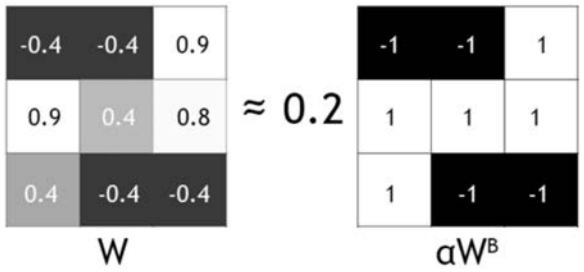


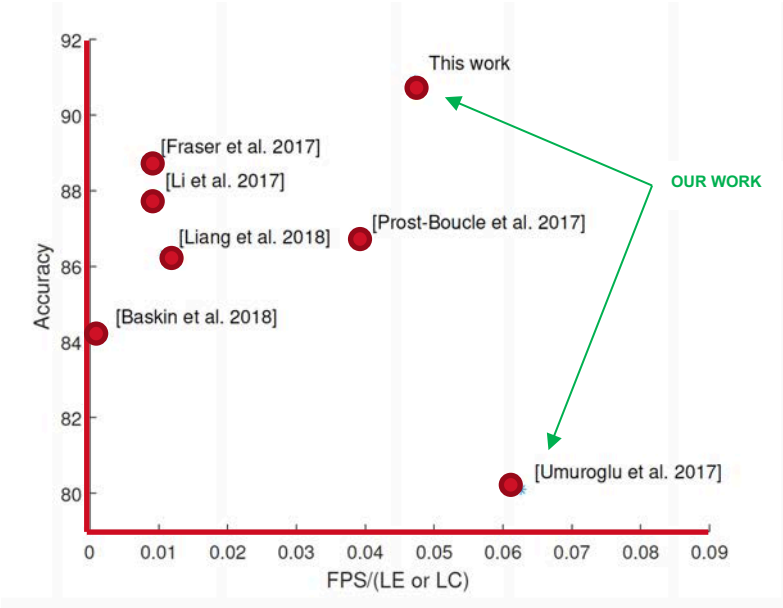
- A family of kernel methods that can do simultaneous learning and inference
  - Highest reported throughput 80 Gbps (TRETS'17)
  - Lowest reported latency 80 ns (FPT'15)
  - Higest capacity (FPGA'18)

Parallelism: Binarized Neural Networks

Collaboration with Xilinx





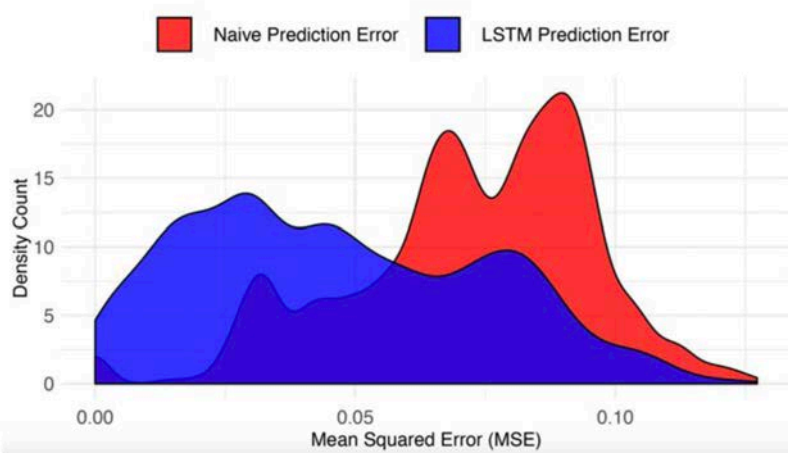


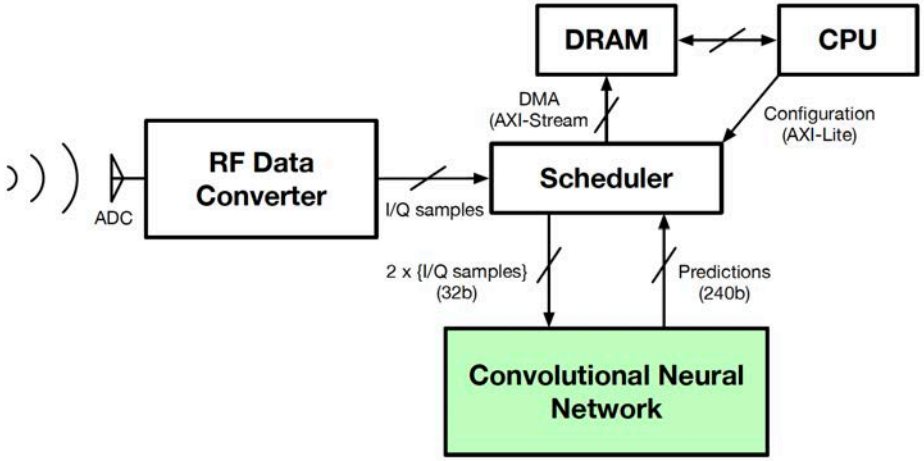
Ours is the most accurate and fastest reported FPGA-based CNN inference implementation
CIFAR10: 90.9% acc, 122K fps (TRETS'19)

Integration: Radio Frequency Signals

Next Generation Technology Fund

- Processing RF signals remains a challenge
  - FPGAs allow integration of radio, machine learning and signal processing





LSTM Spectral prediction: 4.3 µs latency on Ettus X310 XC7K410T (MILCOM'18)

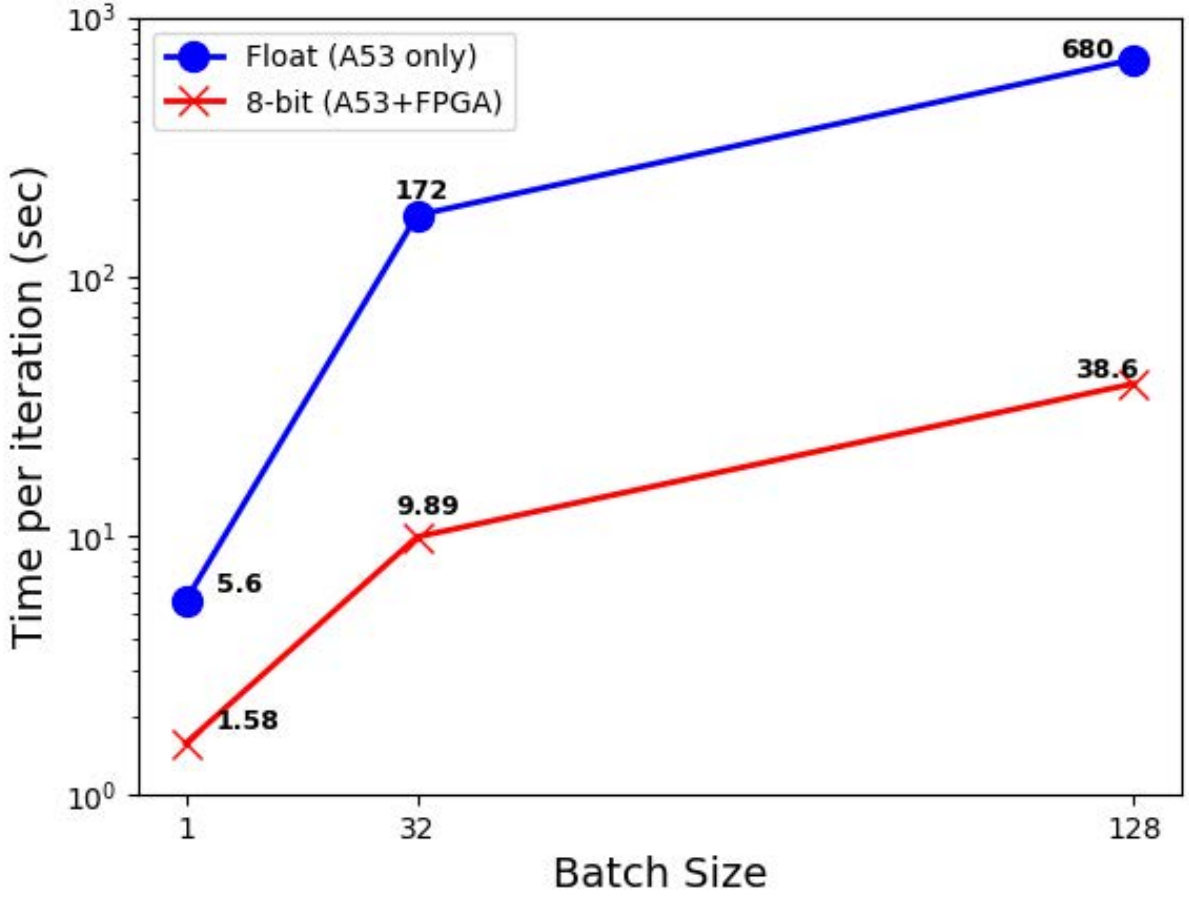
Ternary Modulation classifier: 488K class/s, 8us latency, Xilinx ZCU111 RFSoC (FPT'19)

Customisation: High Dynamic Range Signals

Defence Innovation Hub

- Implementation of a neuromorphic high dynamic range camera-based object detector on FPGAs
- Significantly improved accuracy in high contrast situations

On-FPGA Training



- 17x speed-up over ARM

Overview

FPGAs

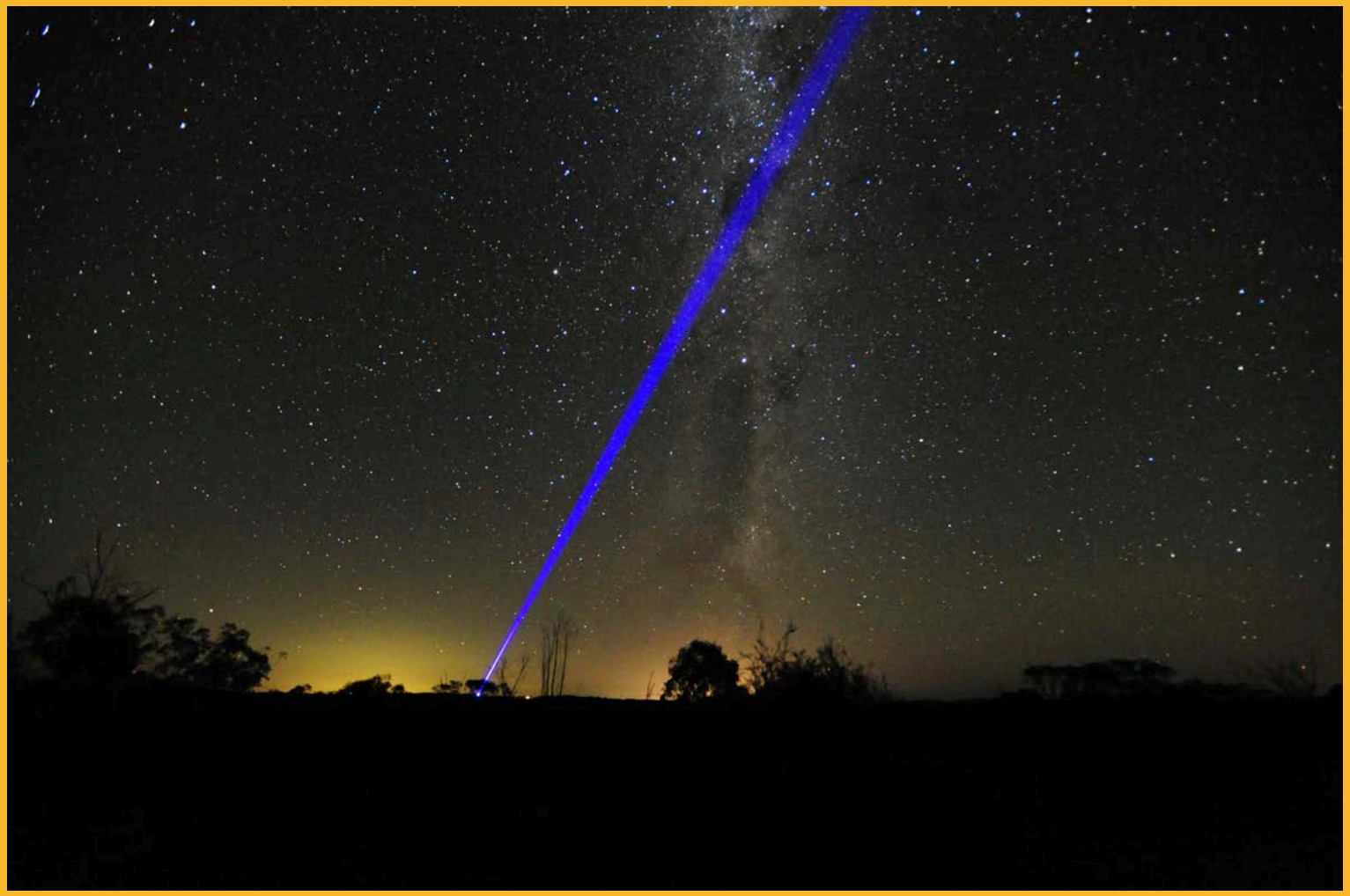
Applications

Our work

Summary

- Industry Trends
  - Cloud/edge unification
  - **More** Sensors (video and hyperspectral); **more** nodes (edge devices/servers) generating data; **more** computation (DNNs, Monte Carlo methods); **more** bandwidth
  - Real-time AI and data science applied at all levels
- FPGAs has advantages for these types of problems

Figure: Microsoft



THE UNIVERSITY OF SYDNEY