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

THE UNIVERSITY OF SYDNEY

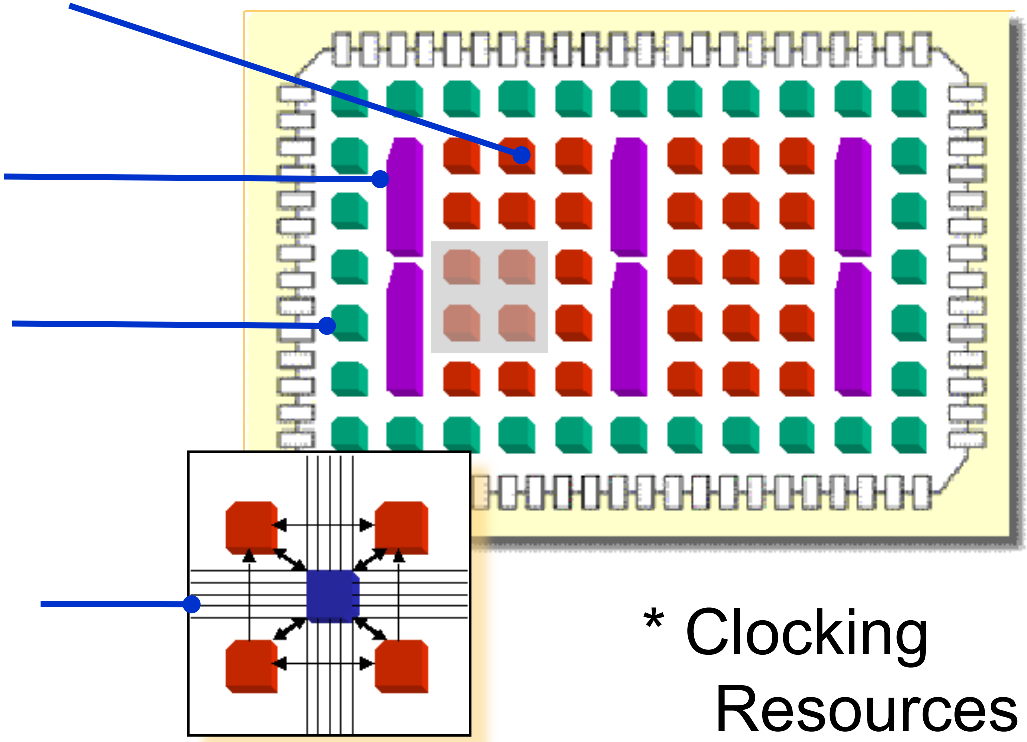
THE UNIVERSITY OF SYDNEY

What is an FPGA?

User-customisable integrated circuit

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

Xilinx FPGAs



Configurable logic blocks

Dedicated blocks

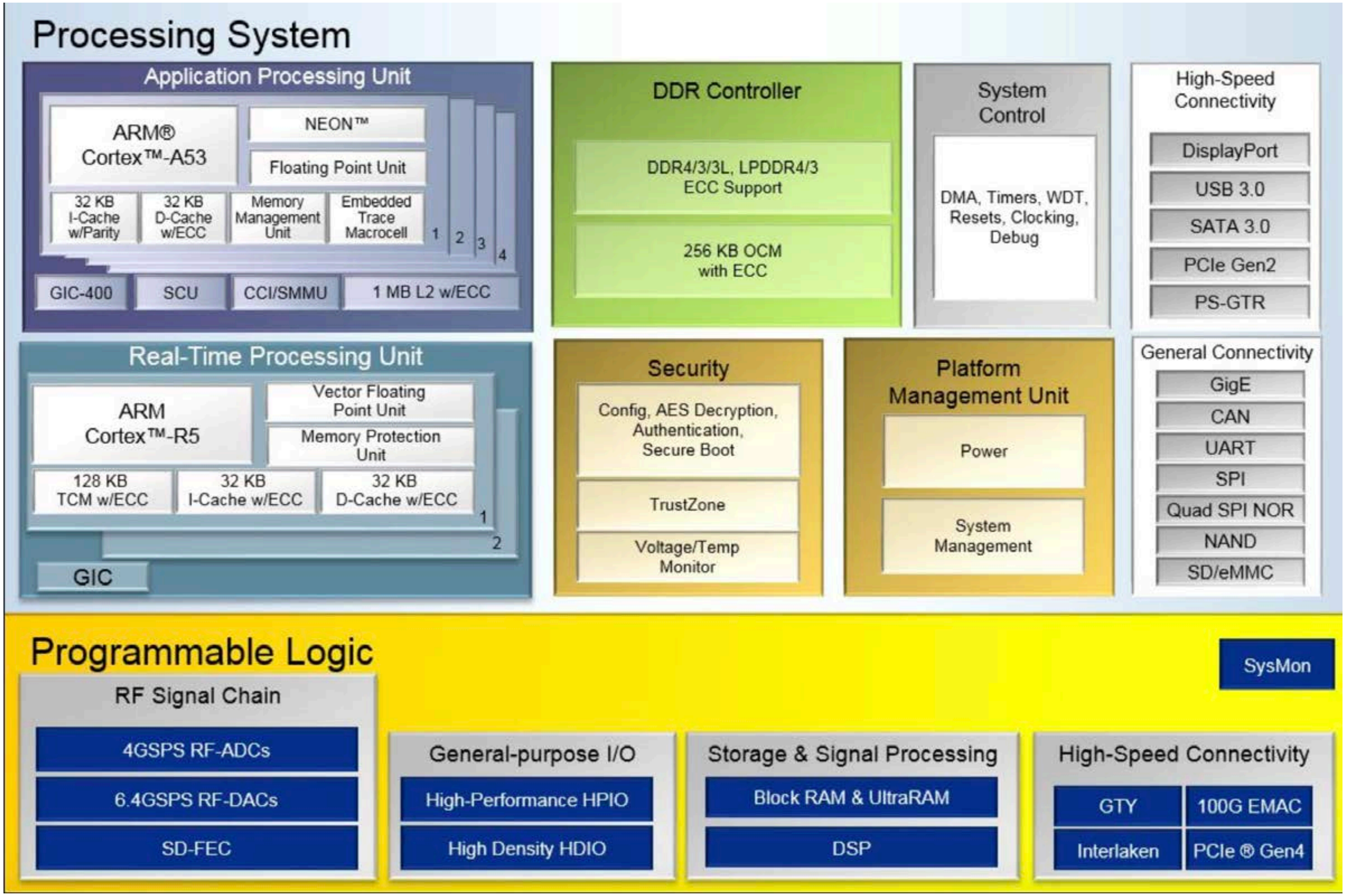
Input and output blocks

Routing

* Clocking Resources

Source: Xilinx

Xilinx RFSoc Device



Processing System

Application Processing Unit

ARM® Cortex™-A53

NEON™

Floating Point Unit

32 KB I-Cache w/Parity

32 KB D-Cache w/ECC

Memory Management Unit

Embedded Trace Macrocell

1 2 3 4

GIC-400

SCU

CCI/SMMU

1 MB L2 w/ECC

Real-Time Processing Unit

ARM Cortex™-R5

Vector Floating Point Unit

Memory Protection Unit

128 KB TCM w/ECC

32 KB I-Cache w/ECC

32 KB D-Cache w/ECC

1 2

GIC

DDR Controller

DDR4/3/3L, LPDDR4/3 ECC Support

256 KB OCM with ECC

System Control

DMA, Timers, WDT, Resets, Clocking, Debug

High-Speed Connectivity

DisplayPort

USB 3.0

SATA 3.0

PCIe Gen2

PS-GTR

Security

Config, AES Decryption, Authentication, Secure Boot

TrustZone

Voltage/Temp Monitor

Platform Management Unit

Power

System Management

General Connectivity

GigE

CAN

UART

SPI

Quad SPI NOR

NAND

SD/eMMC

Programmable Logic

SysMon

RF Signal Chain

4GSPS RF-ADCs

6.4GSPS RF-DACs

SD-FEC

General-purpose I/O

High-Performance HPIO

High Density HDIO

Storage & Signal Processing

Block RAM & UltraRAM

DSP

High-Speed Connectivity

GTY

100G EMAC

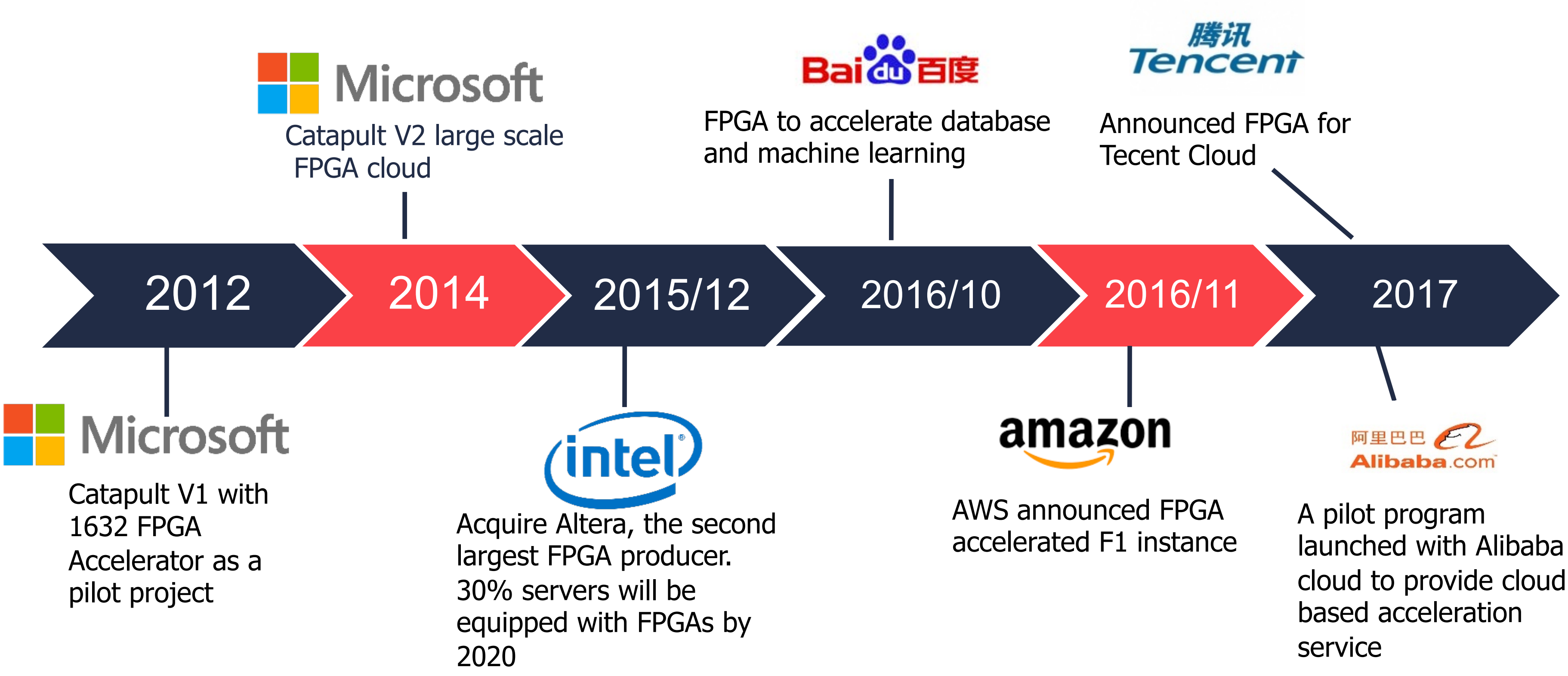
Interlaken

PCIe ® Gen4

Source: Xilinx

THE UNIVERSITY OF SYDNEY

Recent Uptake in Reconfigurable Computing



Catapult V1 with 1632 FPGA Accelerator as a pilot project

Catapult V2 large scale FPGA cloud

Acquire Altera, the second largest FPGA producer. 30% servers will be equipped with FPGAs by 2020

FPGA to accelerate database and machine learning

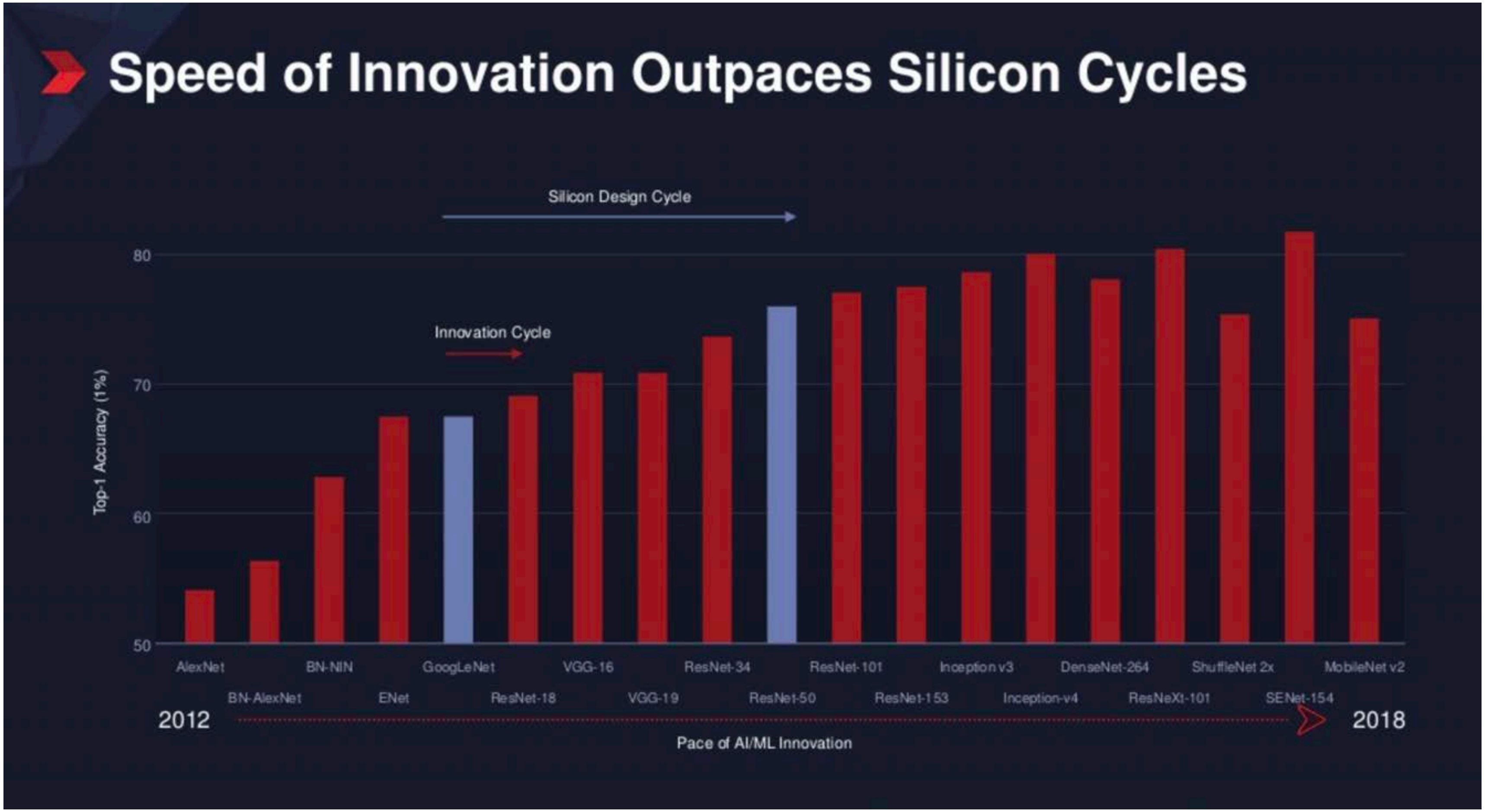
AWS announced FPGA accelerated F1 instance

Announced FPGA for Tecent Cloud

A pilot program launched with Alibaba cloud to provide cloud based acceleration service

Source: CTAccel

Speed of Innovation

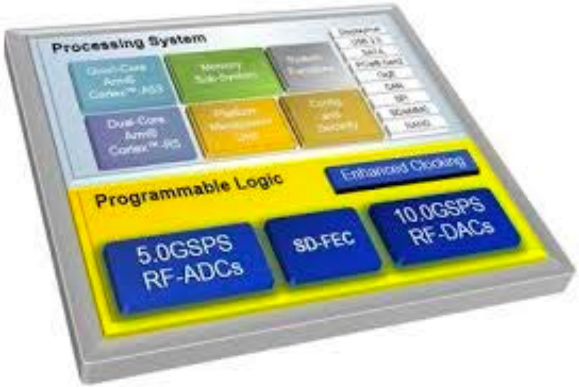


Speed of Innovation Outpaces Silicon Cycles

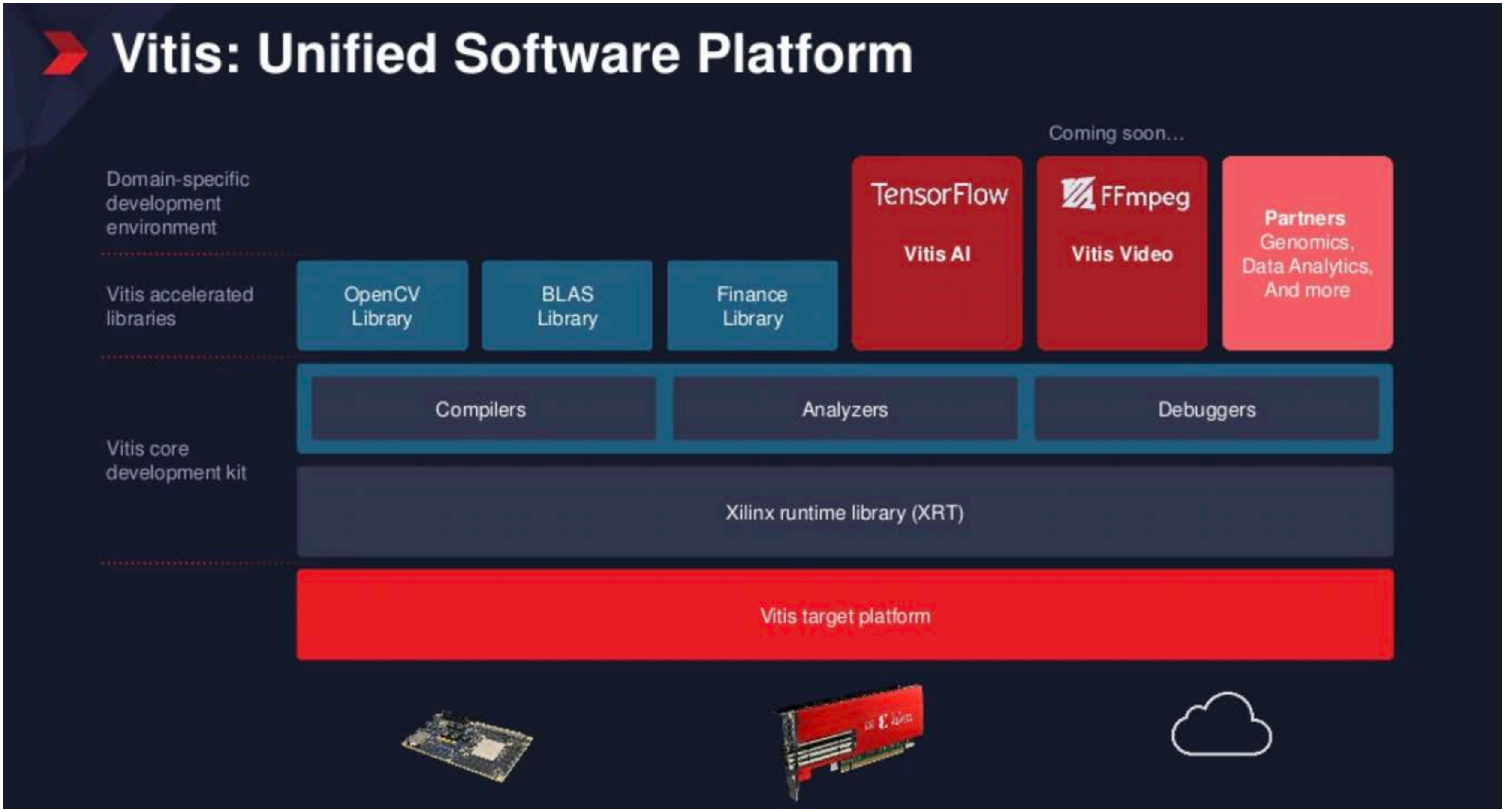
Source: Xilinx

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



Vitis: Unified Software Platform

Coming soon…

Domain-specific development environment

TensorFlow
Vitis AI

FFmpeg
Vitis Video

Partners
Genomics,
Data Analytics,
And more

Vitis accelerated libraries

OpenCV Library

BLAS Library

Finance Library

Vitis core development kit

Compilers

Analyzers

Debuggers

Xilinx runtime library (XRT)

Vitis target platform

Source: Xilinx

Overview

FPGAs

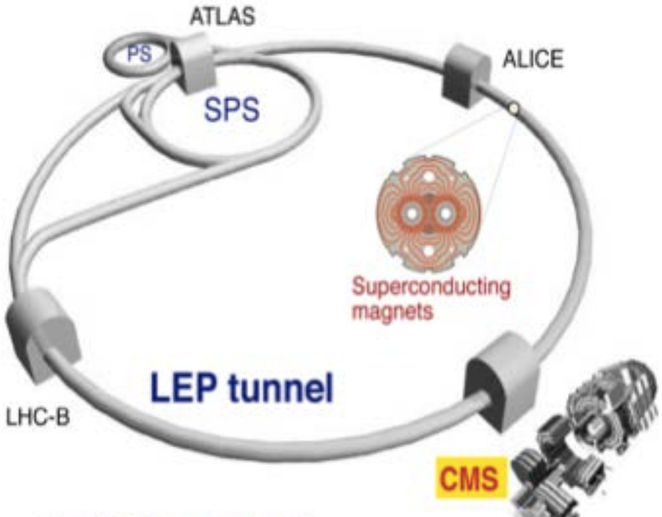
Applications

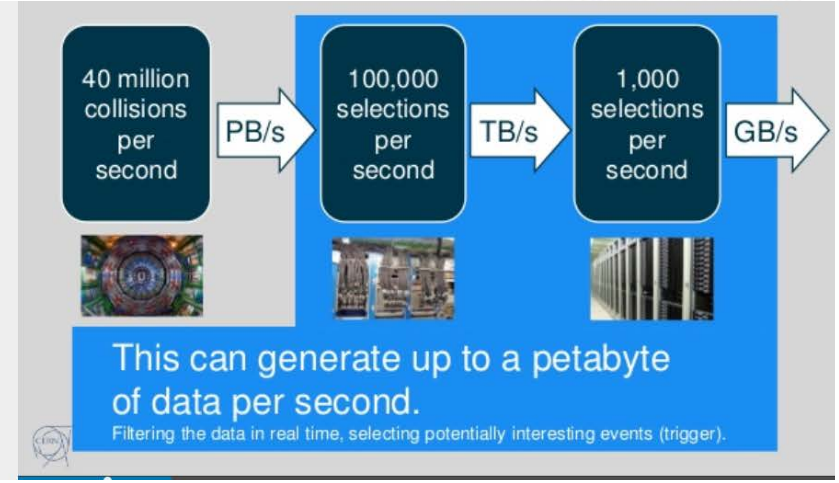
Our work

THE UNIVERSITY OF SYDNEY

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)





40 million collisions per second

PB/s

100,000 selections per second

TB/s

1,000 selections per second

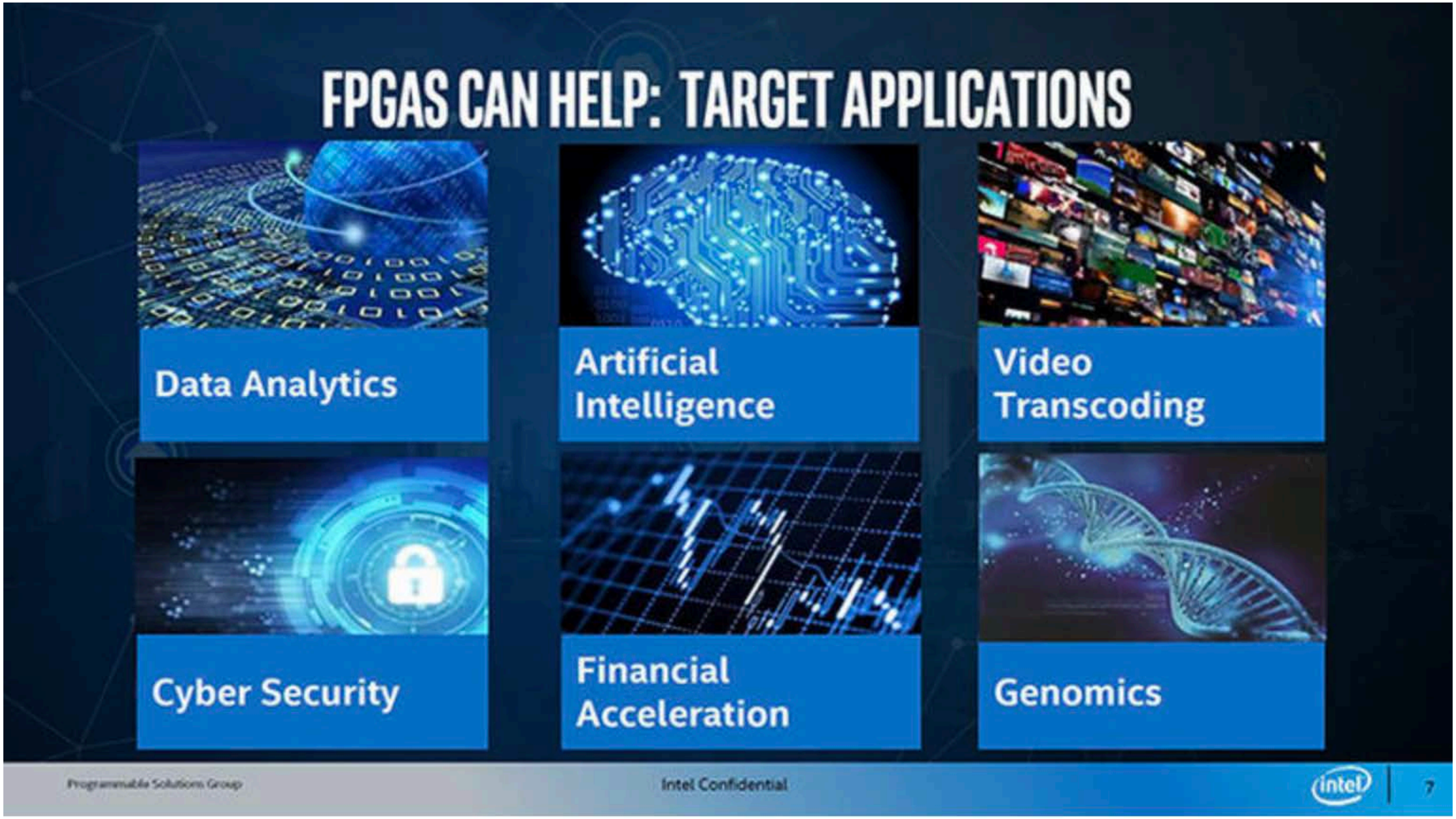
GB/s

This can generate up to a petabyte of data per second.

Filtering the data in real time, selecting potentially interesting events (trigger).

Source: Geoff Hall, Imperial College

Target Commercial Applications



FPGAS CAN HELP: TARGET APPLICATIONS

- Data Analytics
- Artificial Intelligence
- Video Transcoding
- Cyber Security
- Financial Acceleration
- Genomics

Programmable Solutions Group

Intel Confidential

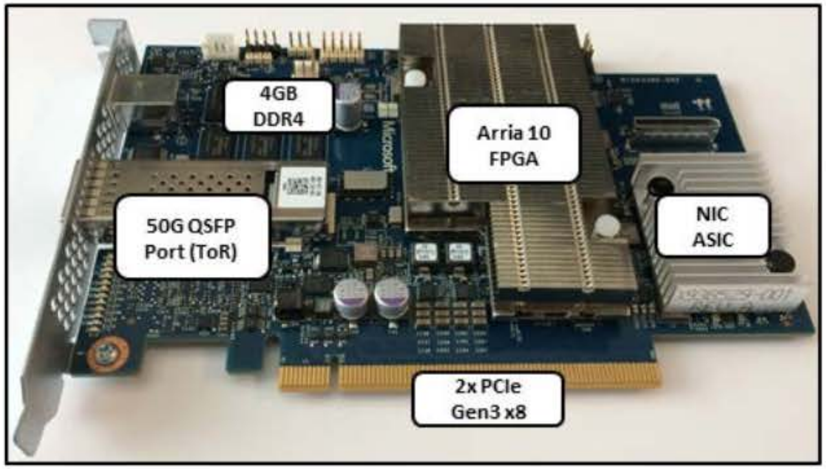
7

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

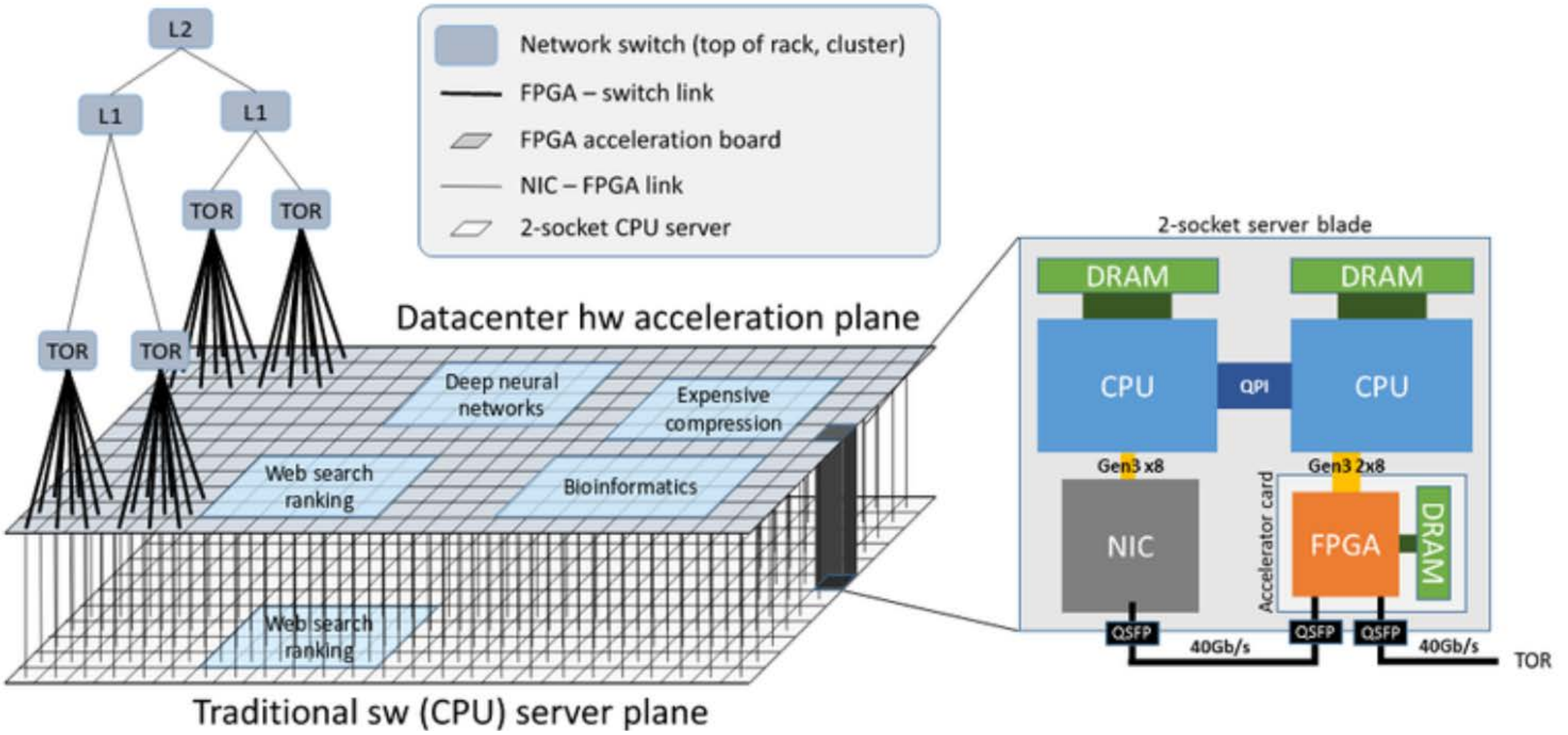
- 2010: MSR study FPGAs to accelerate Web search
- 2012: Project Catapult's scale pilot of 1,632 FPGA servers deployed
- 2013: Bing decision-tree algorithms 40x faster than CPUs
- 2015: FPGAs deployed at scale in Bing and Azure datacenters (> 1M) - enabled 50% ↑ throughput, 25% ↓ latency.



Source: Microsoft

Microsoft Azure Cloud Network

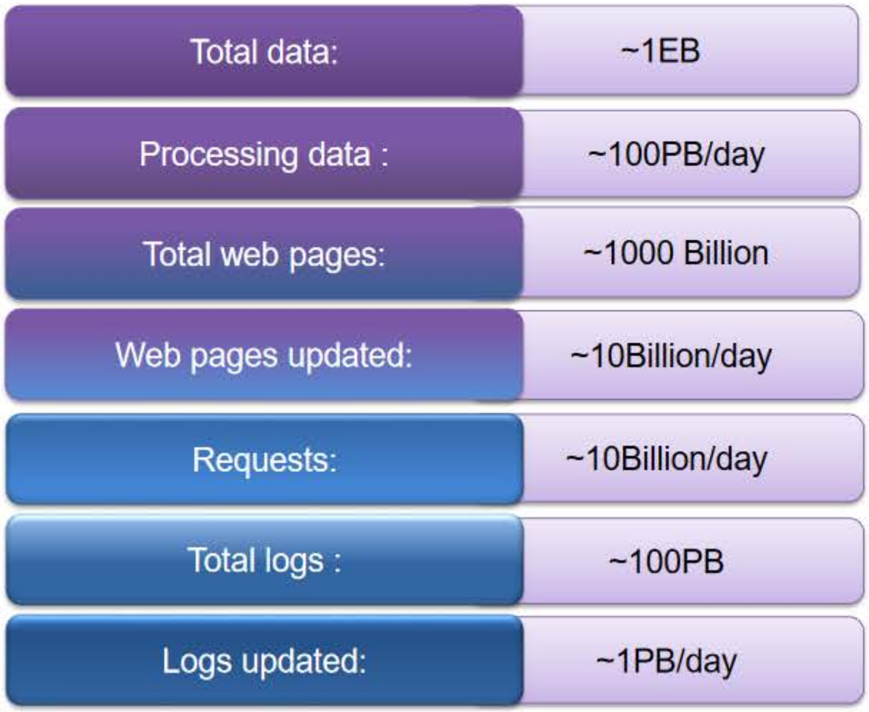
World's fastest cloud network



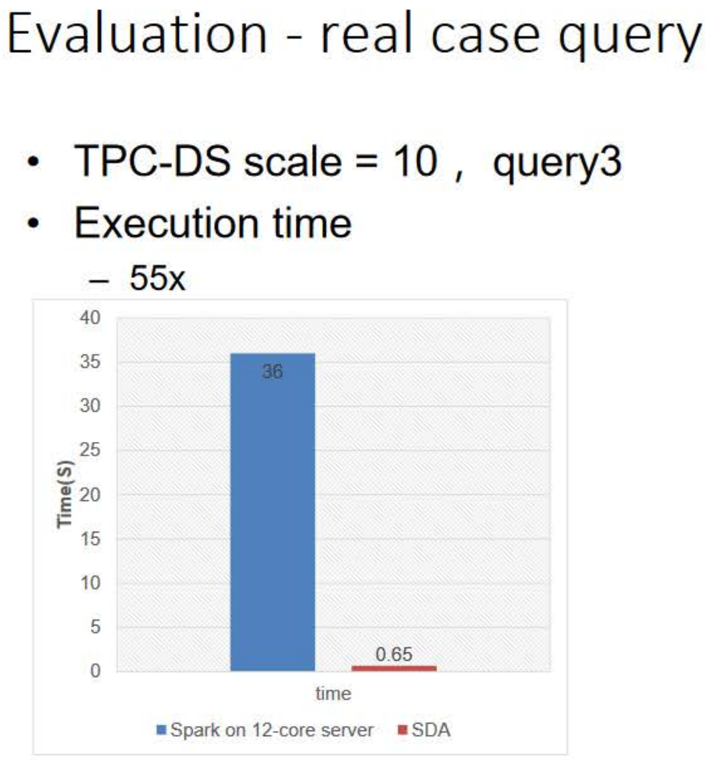
Source: Microsoft

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

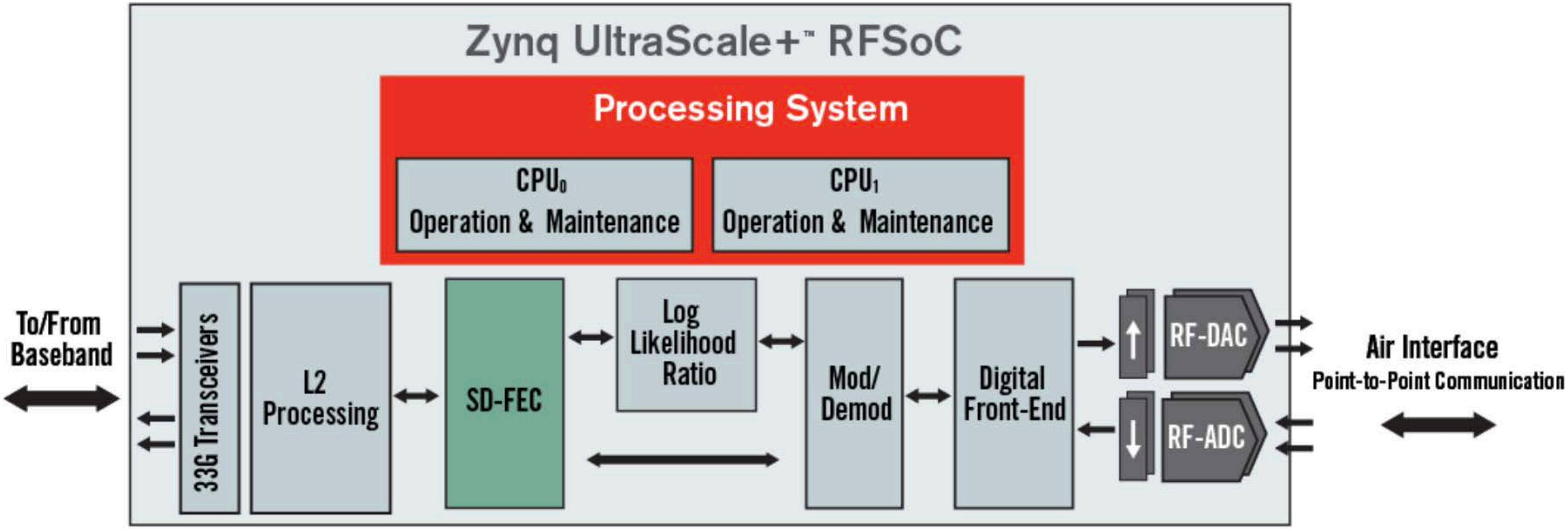
- TPC-DS scale = 10， query3
- Execution time
  - 55x

Source: Baidu 2016 Hot Chips

Mobile Backhaul

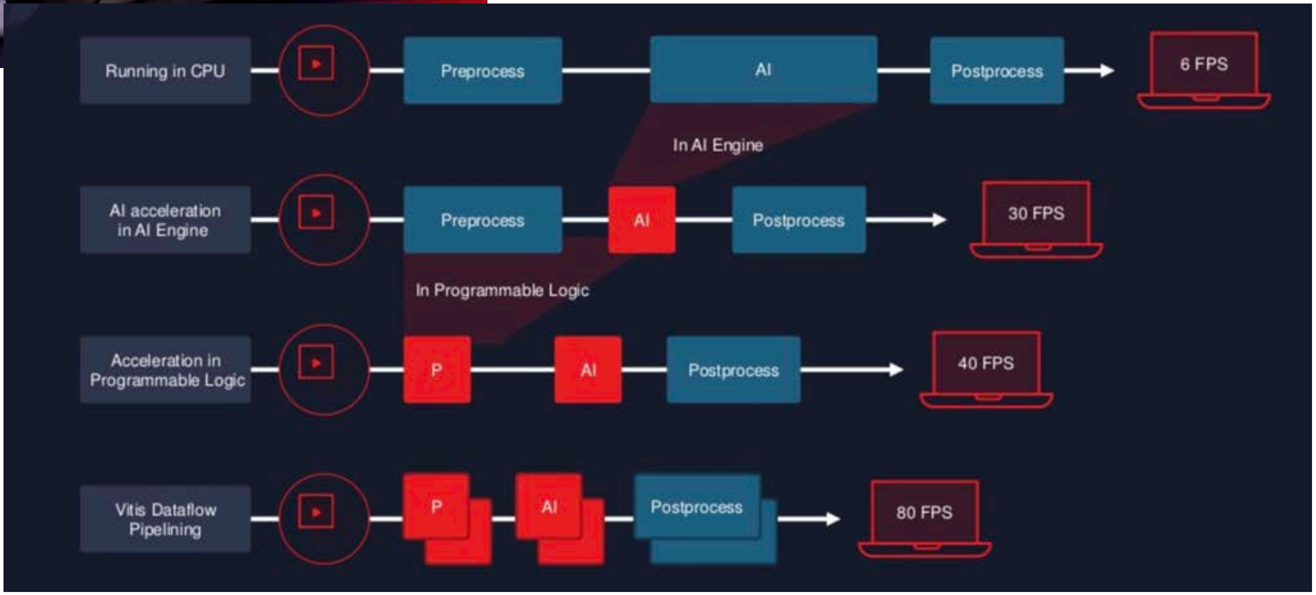
Key Zynq UltraScale+ RFSoC Benefits:

- 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


Running in CPU → Preprocess → AI → Postprocess → 6 FPS

In AI Engine

AI acceleration in AI Engine → Preprocess → AI → Postprocess → 30 FPS

In Programmable Logic

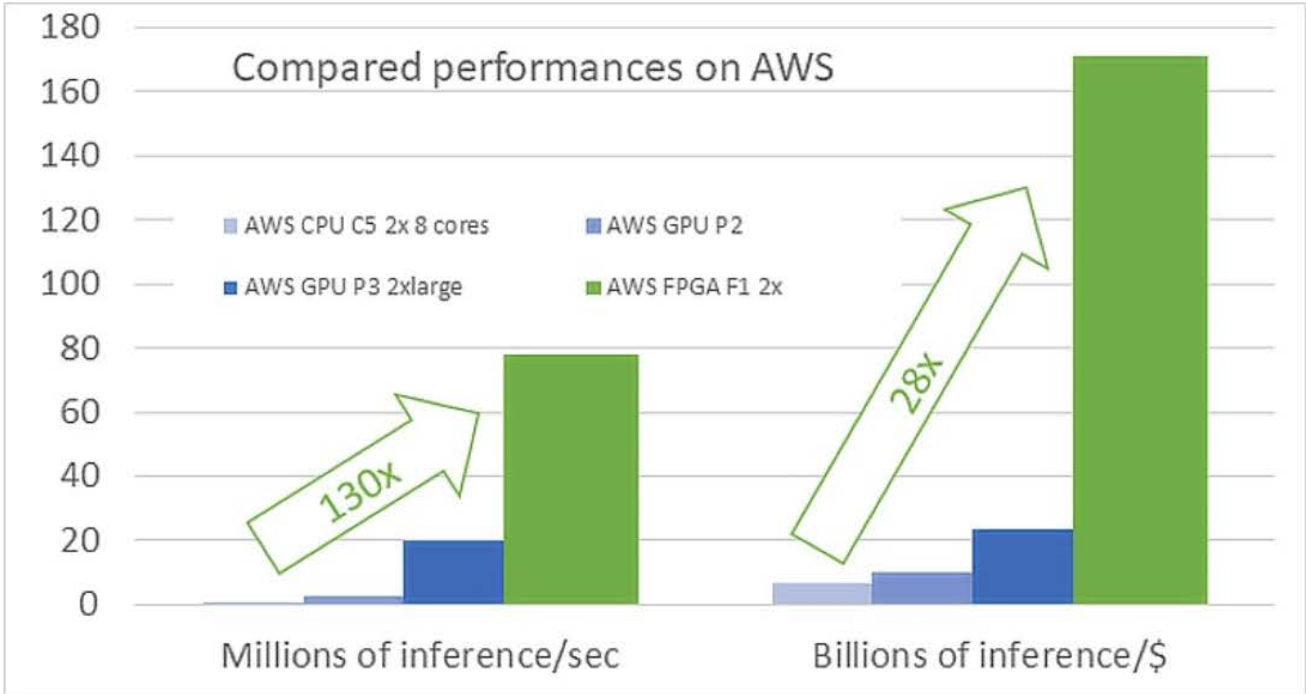
Acceleration in Programmable Logic → P → AI → Postprocess → 40 FPS

Vitis Dataflow Pipelining → P → AI → Postprocess → 80 FPS

Source: Xilinx

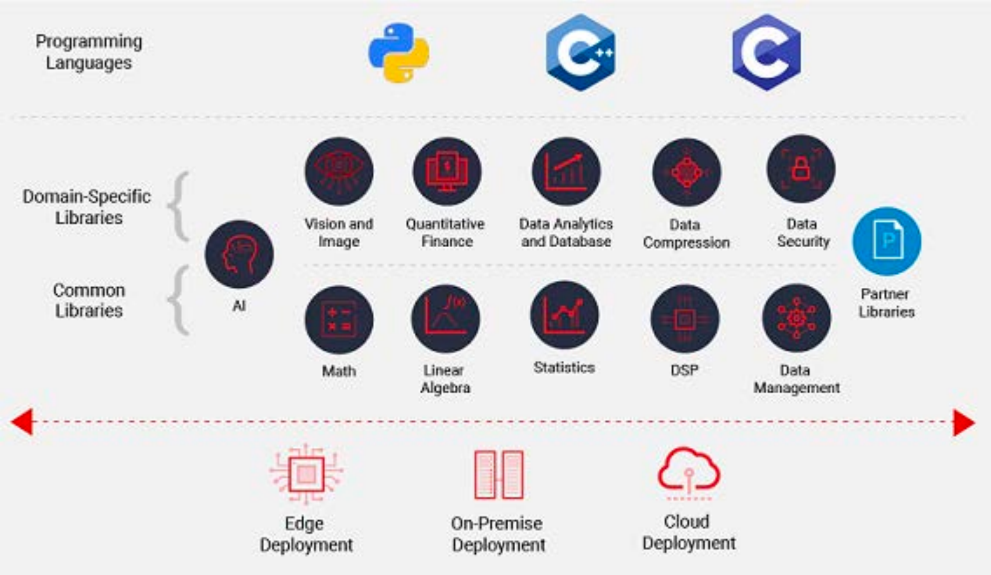
Amadeus

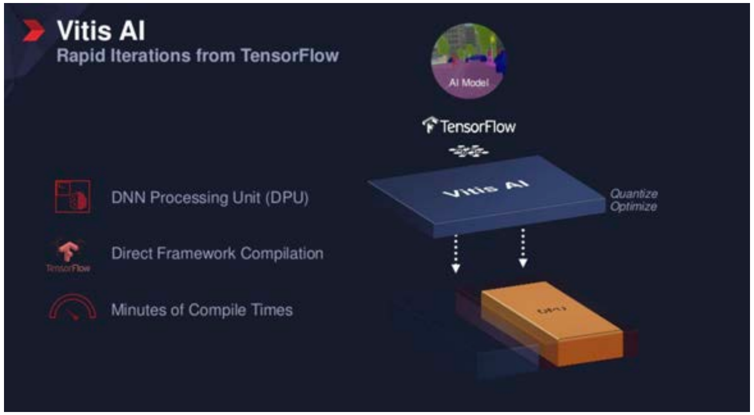
- Amadeus IT Group S.A adjusted profit €1.27B in 2019
- Accelerated inference of gradient boosted decision trees for search queries and quantified cost

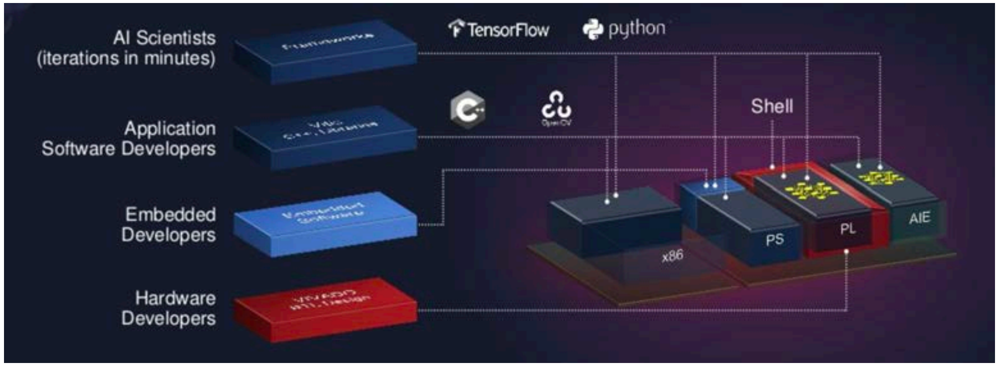


Source: VLDB 2020 http://www.vldb.org/pvldb/vol13/p71-owaida.pdf

Xilinx Vitis Unified Software Platform



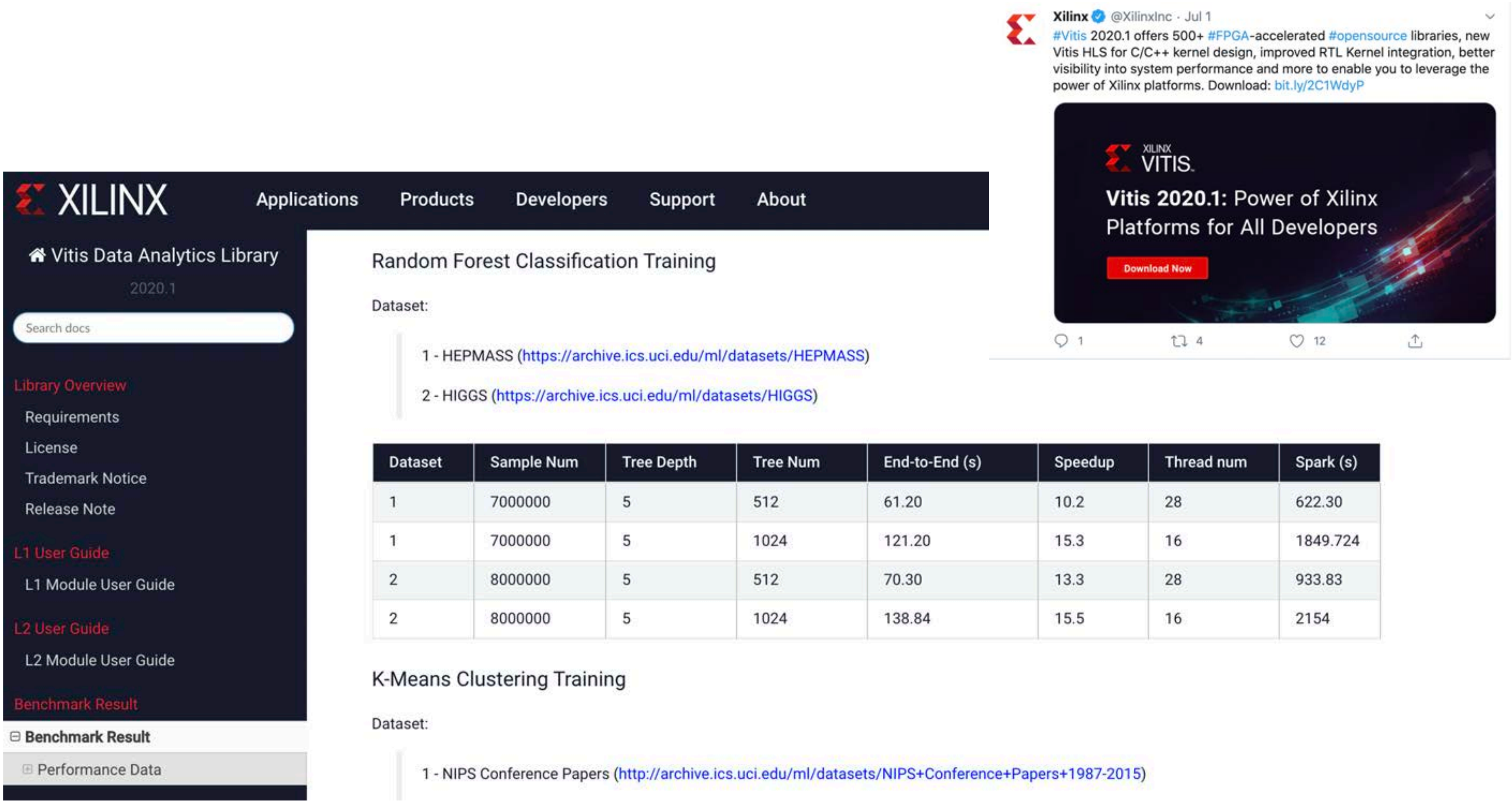




https://github.com/Xilinx/Vitis_Libraries

Source: Xilinx

Vitis Data Analytics Library



Xilinx @XilinxInc · Jul 1

#Vitis 2020.1 offers 500+ #FPGA-accelerated #opensource libraries, new Vitis HLS for C/C++ kernel design, improved RTL Kernel integration, better visibility into system performance and more to enable you to leverage the power of Xilinx platforms. Download: bit.ly/2C1WdyP

XILINX VITIS.

Vitis 2020.1: Power of Xilinx Platforms for All Developers

Download Now

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

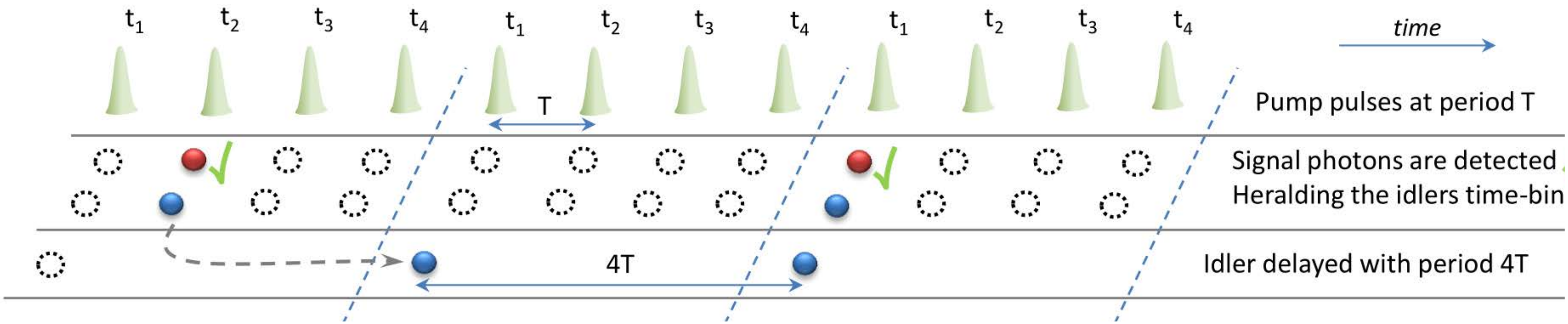
FPGAs

Applications

Our work

THE UNIVERSITY OF SYDNEY

Time domain multiplexing of single photons

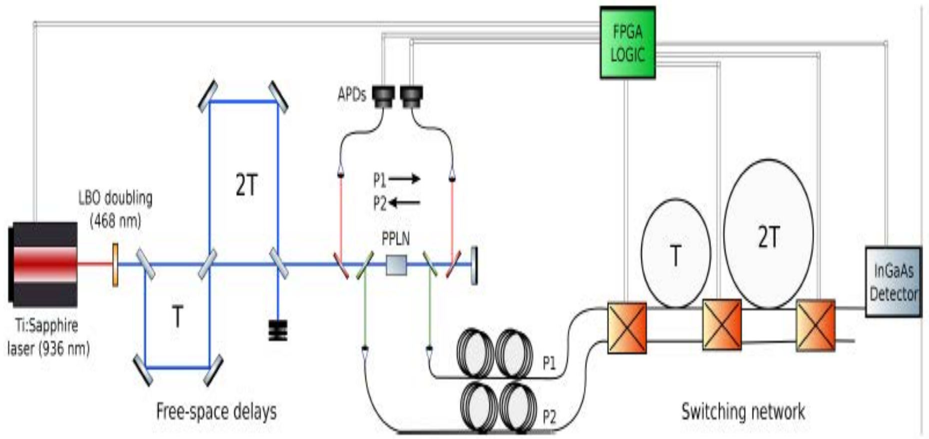


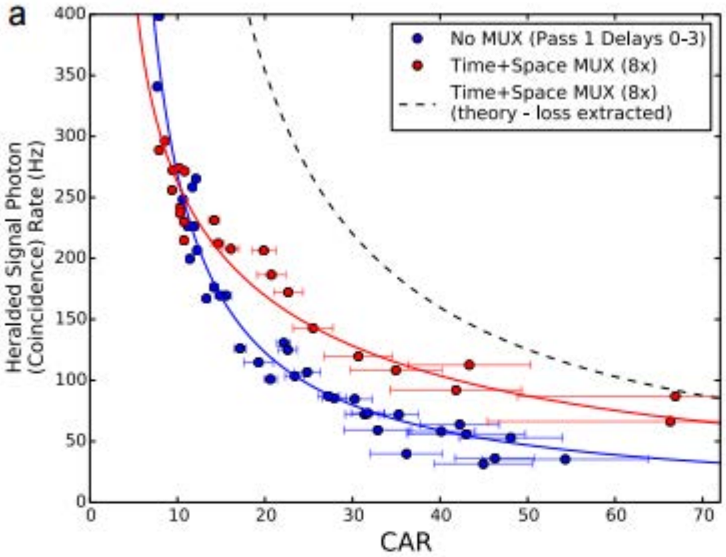
Pump pulses at period T

Signal photons are detected ,
Heralding the idlers time-bin

Idler delayed with period 4T

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



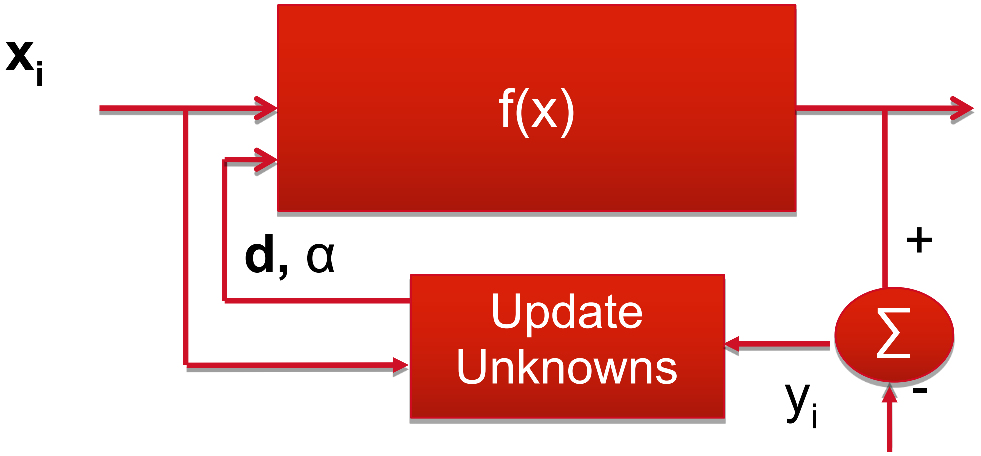


Enhancement : 33%~59%

Time Multiplexing of Single Photons

Exploration: Kernel Methods

ARC Linkage with Exablaze

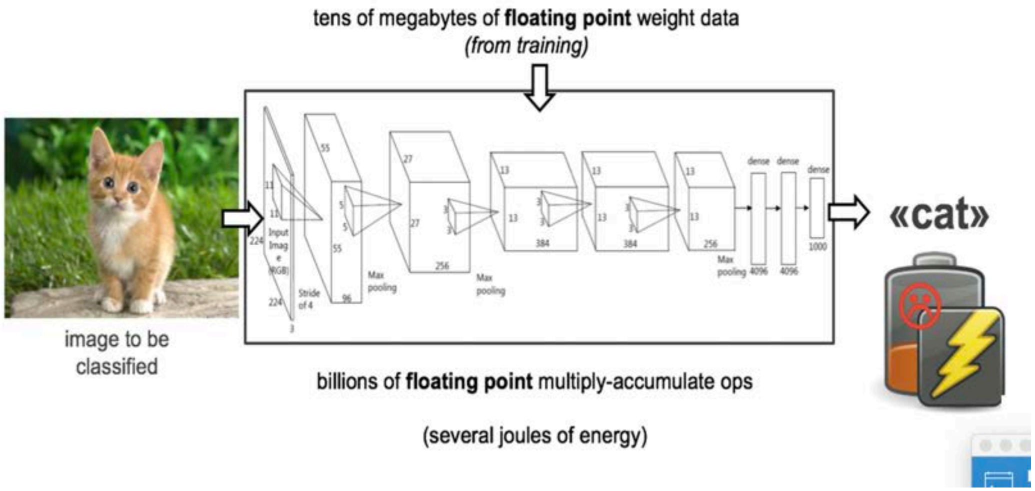


- A family of kernel methods that can do simultaneous learning and inference
  - Highest reported throughput 80 Gbps (TRETS’17)
  - Lowest reported latency 80 ns (FPT’15)
  - Higest capacity (FPGA’18)

THE UNIVERSITY OF SYDNEY

Parallelism: Binarized Neural Networks

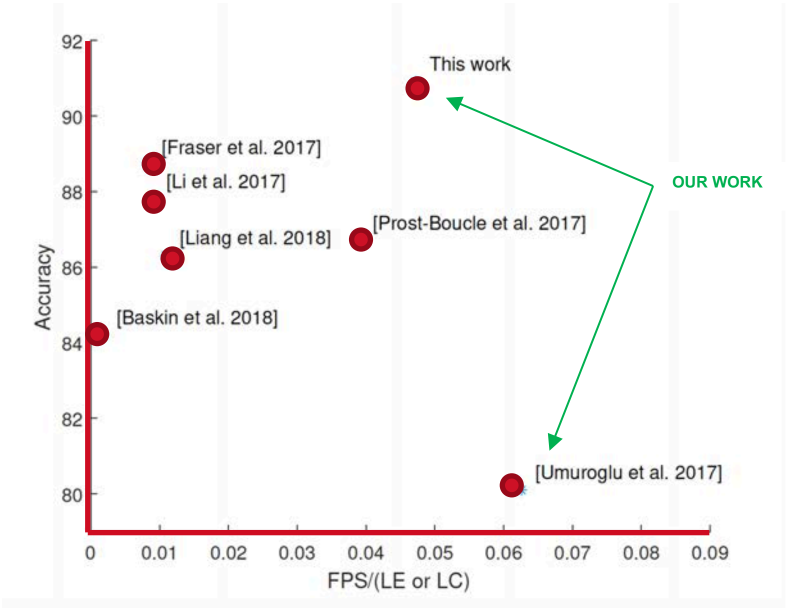
Collaboration with Xilinx


tens of megabytes of **floating point** weight data *(from training)*

billions of **floating point** multiply-accumulate ops

(several joules of energy)

$$\underbrace{\begin{bmatrix}-0.4 & -0.4 & 0.9\\ 0.9 & 0.4 & 0.8\\ 0.4 & -0.4 & -0.4\end{bmatrix}}_{W} \approx 0.2 \underbrace{\begin{bmatrix}-1 & -1 & 1\\ 1 & 1 & 1\\ 1 & -1 & -1\end{bmatrix}}_{\alpha W^{B}}$$



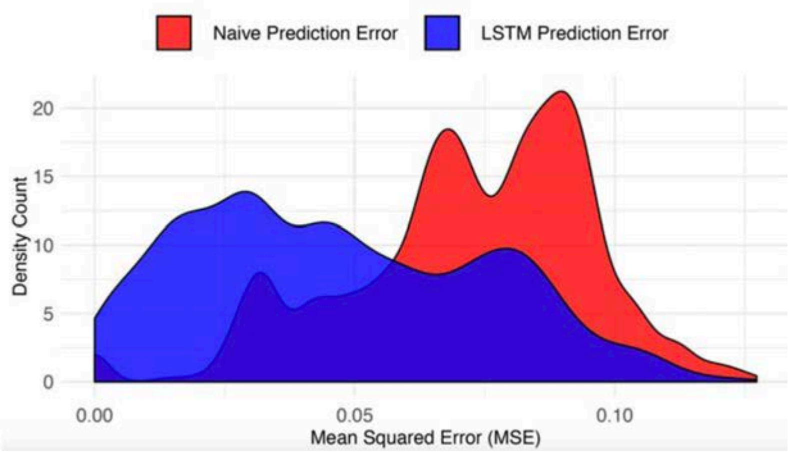
Ours is the most accurate and fastest reported FPGA-based CNN inference implementation
CIFAR10: 90.9% acc, 122K fps (TRETS'19)

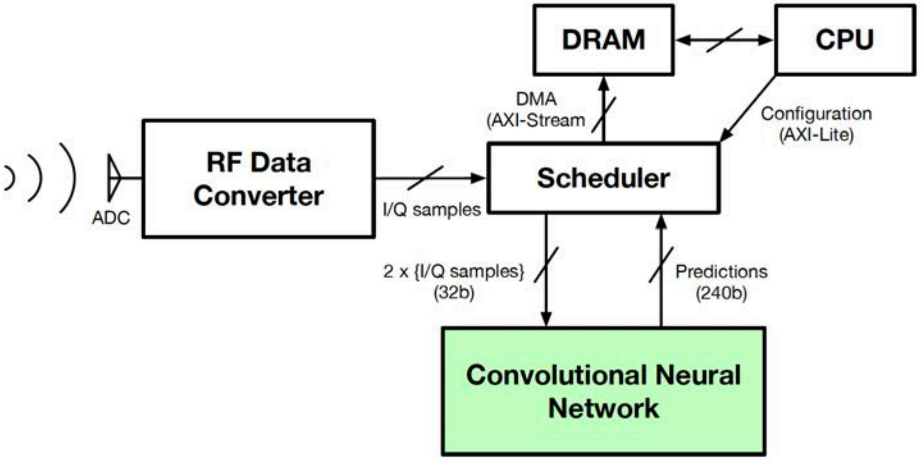
Integration: Radio Frequency Signals

Next Generation Technology Fund

› Processing RF signals remains a challenge

- FPGAs allow integration of radio, machine learning and signal processing





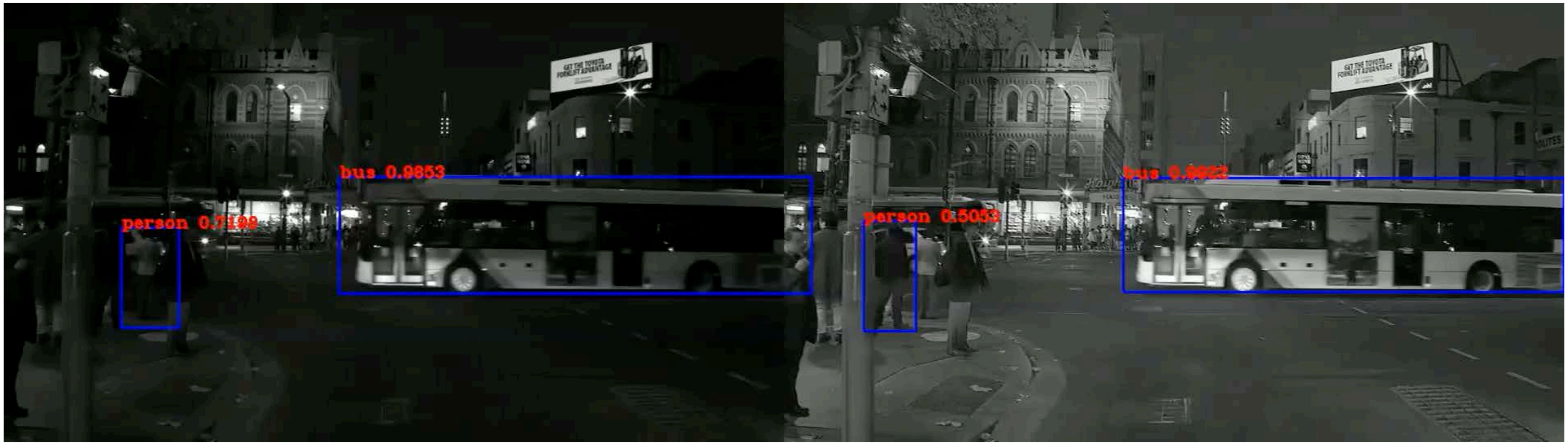
LSTM Spectral prediction: 4.3 µs latency on Ettus X310 XC7K410T (MILCOM'18)

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



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

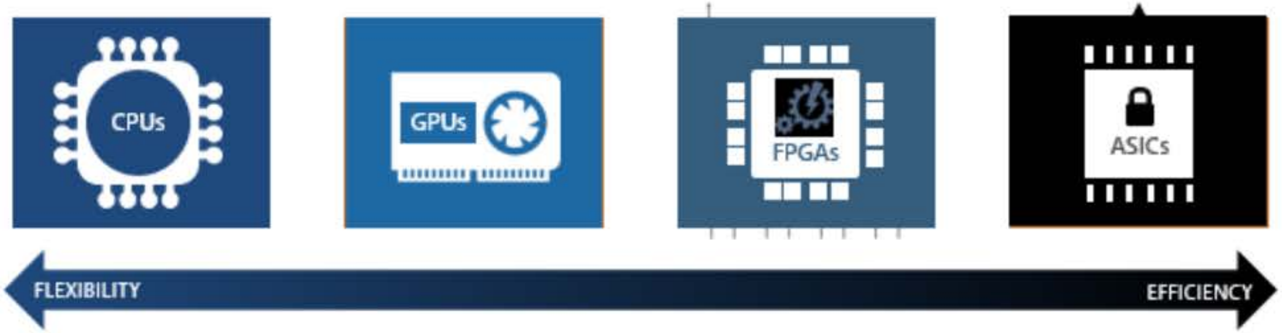
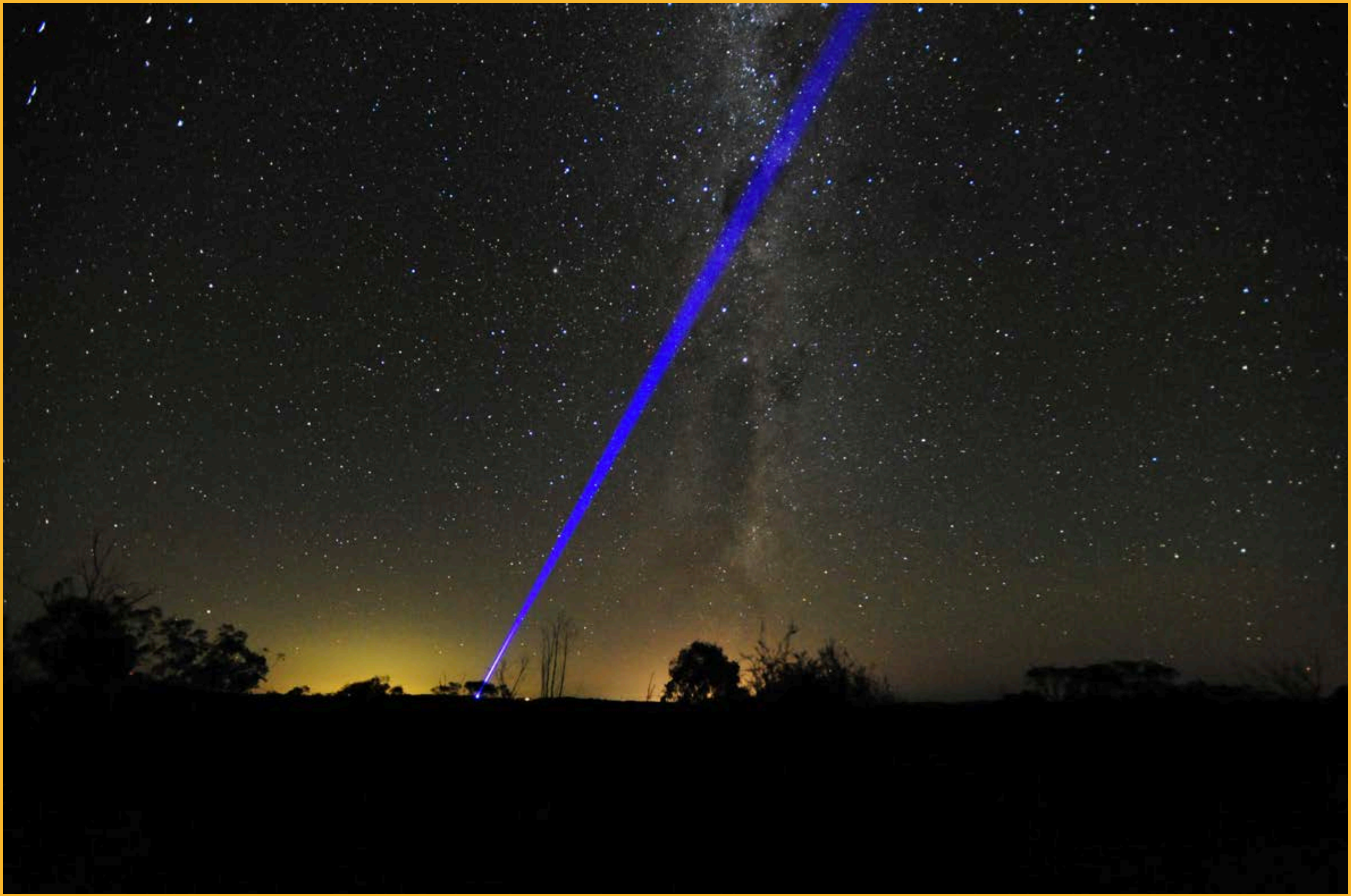


Figure: Microsoft



THE UNIVERSITY OF
SYDNEY