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

FPGA Technology

Applications

Our work

Overview

FPGA Technology

Applications

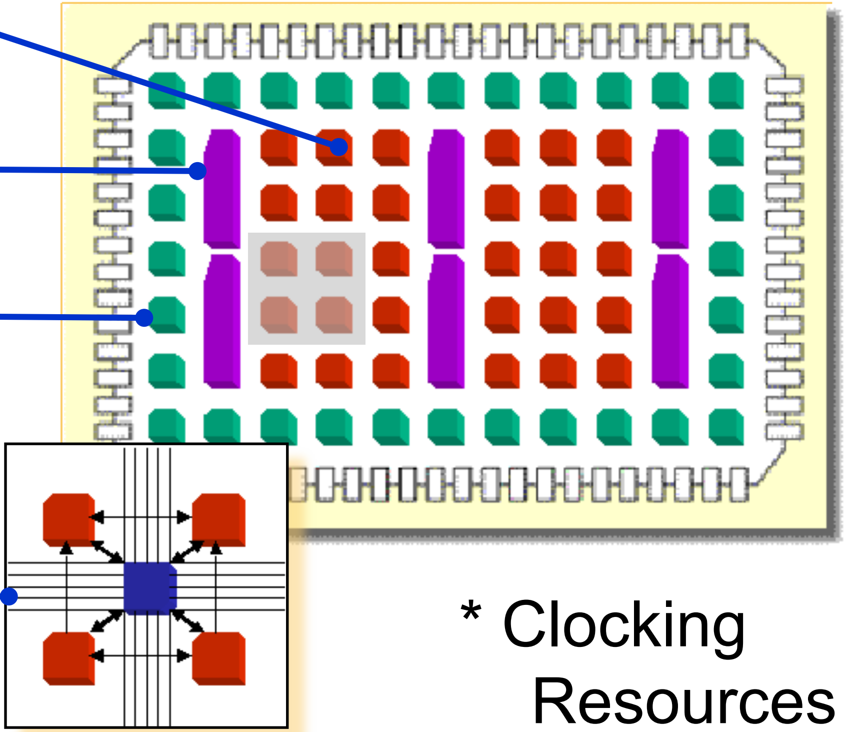
Our work

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

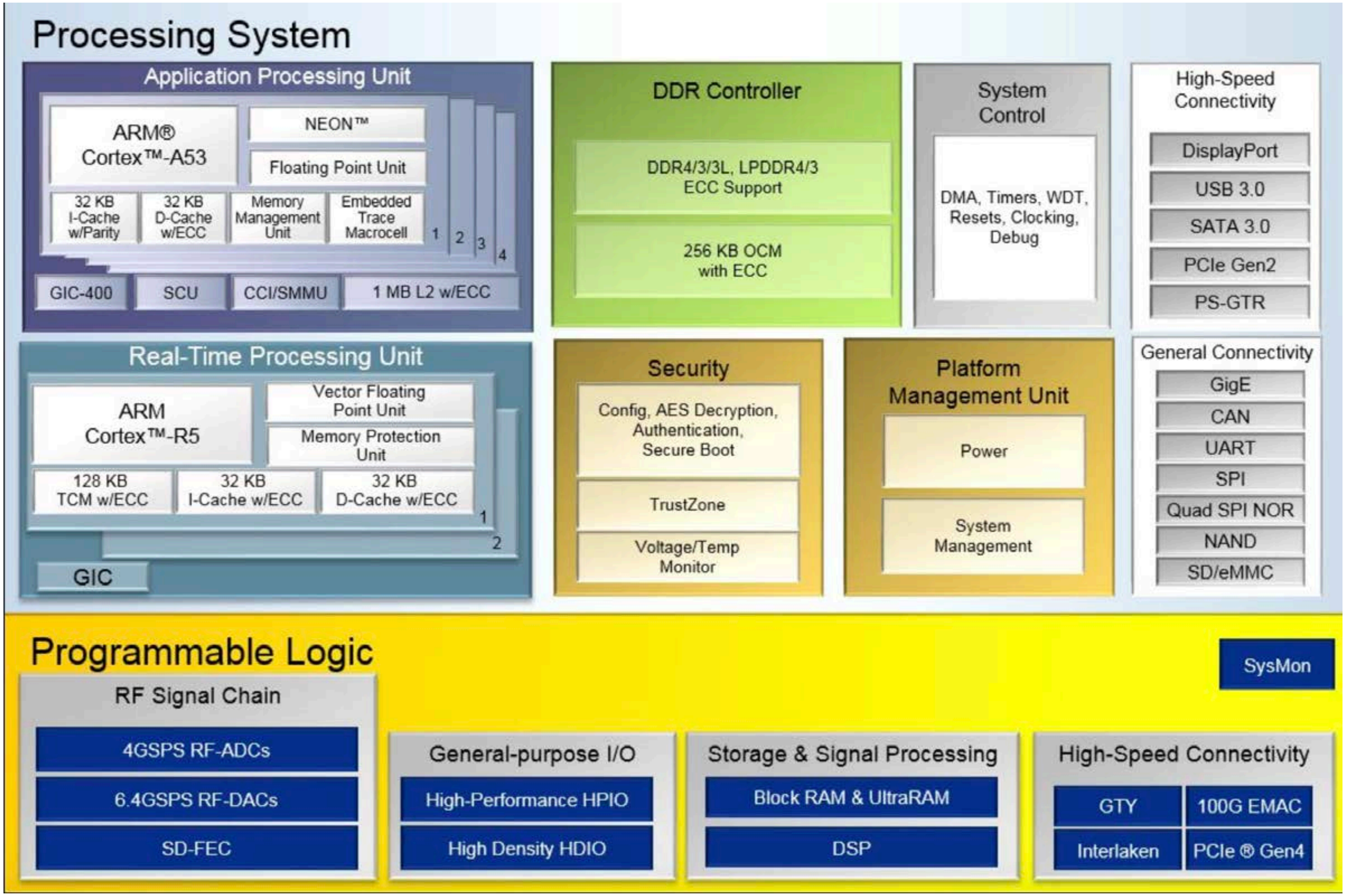
Input and output blocks

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

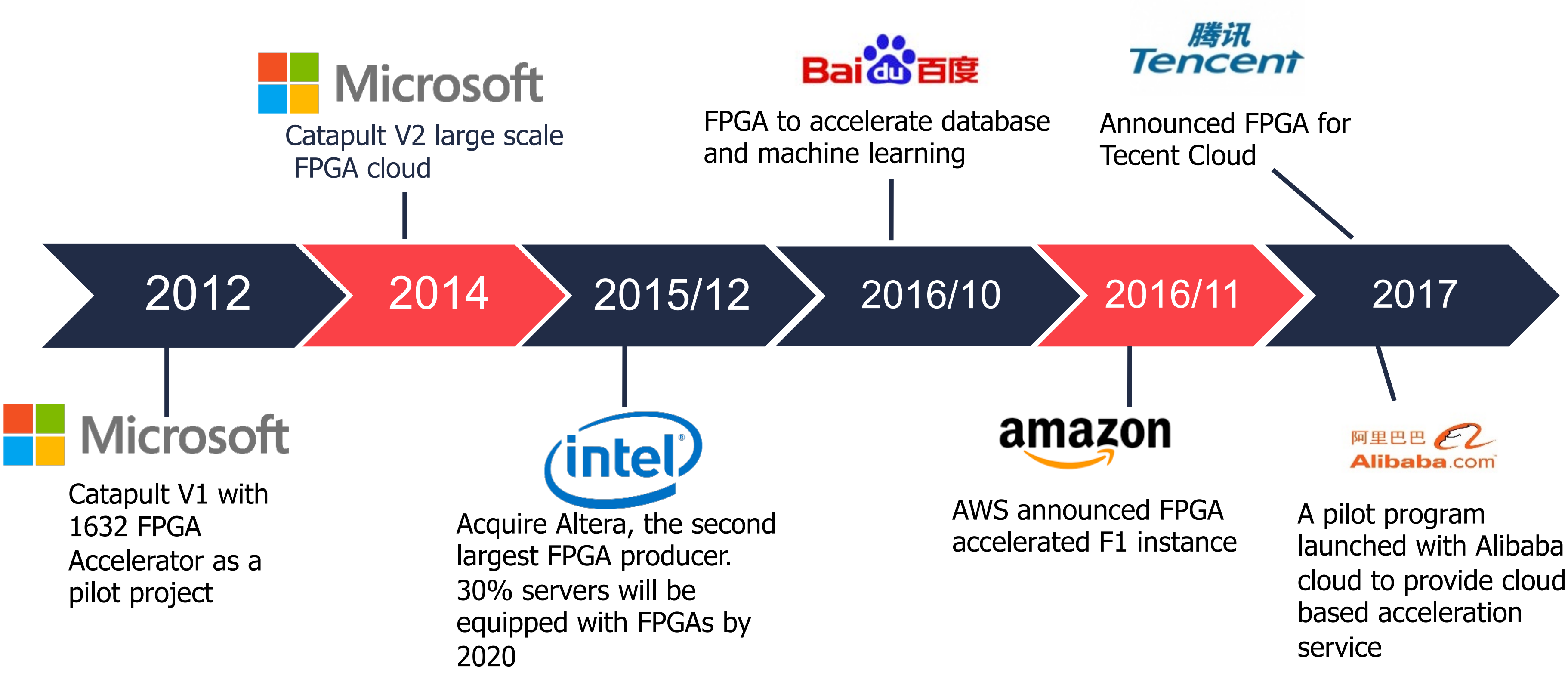
Interlaken

PCIe ® Gen4

Source: Xilinx

THE UNIVERSITY OF SYDNEY

Recent Uptake in Reconfigurable Computing



2012

Microsoft

Catapult V1 with 1632 FPGA Accelerator as a pilot project

2014

Microsoft

Catapult V2 large scale FPGA cloud

2015/12

intel

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

2016/10

Baidu 百度

FPGA to accelerate database and machine learning

2016/11

amazon

AWS announced FPGA accelerated F1 instance

2017

腾讯 Tencent

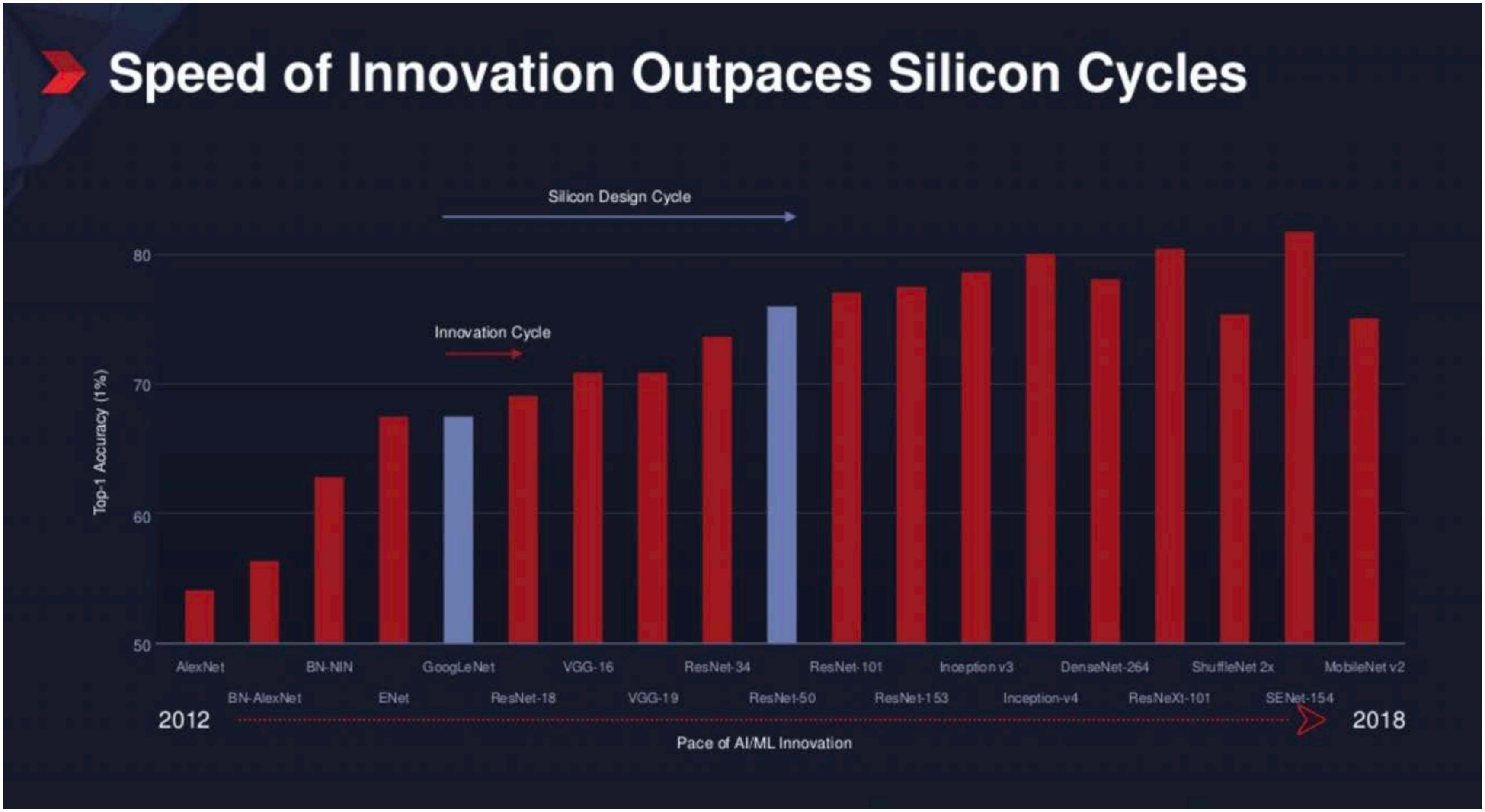
Announced FPGA for Tecent Cloud

阿里巴巴 Alibaba.com

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

Source: CTAccel

Speed of Innovation

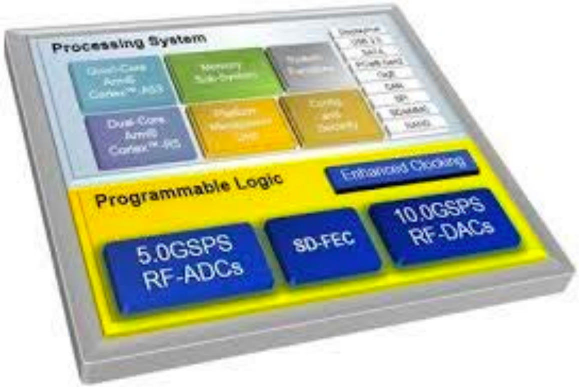


Speed of Innovation Outpaces Silicon Cycles

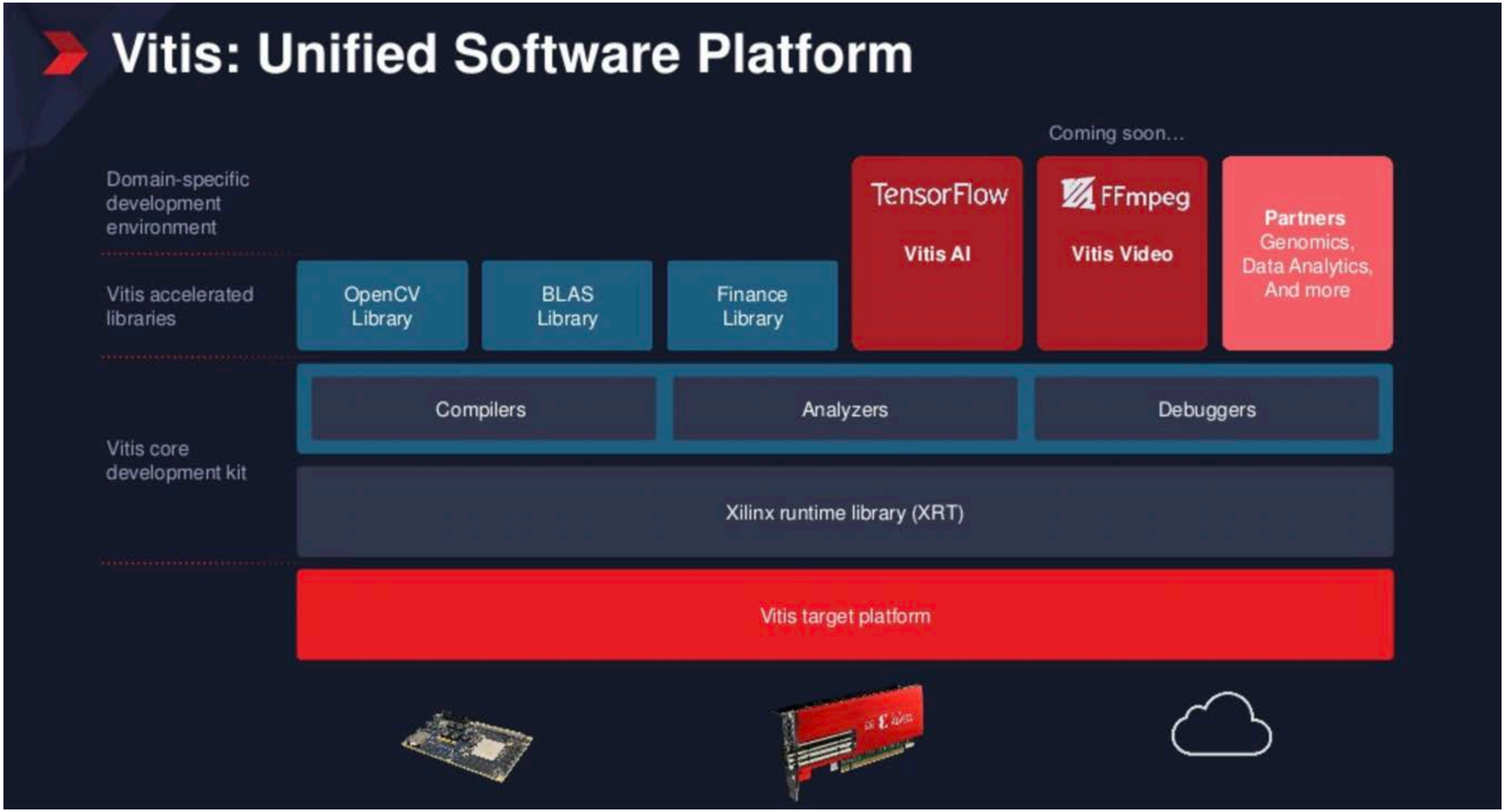
Source: Xilinx

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

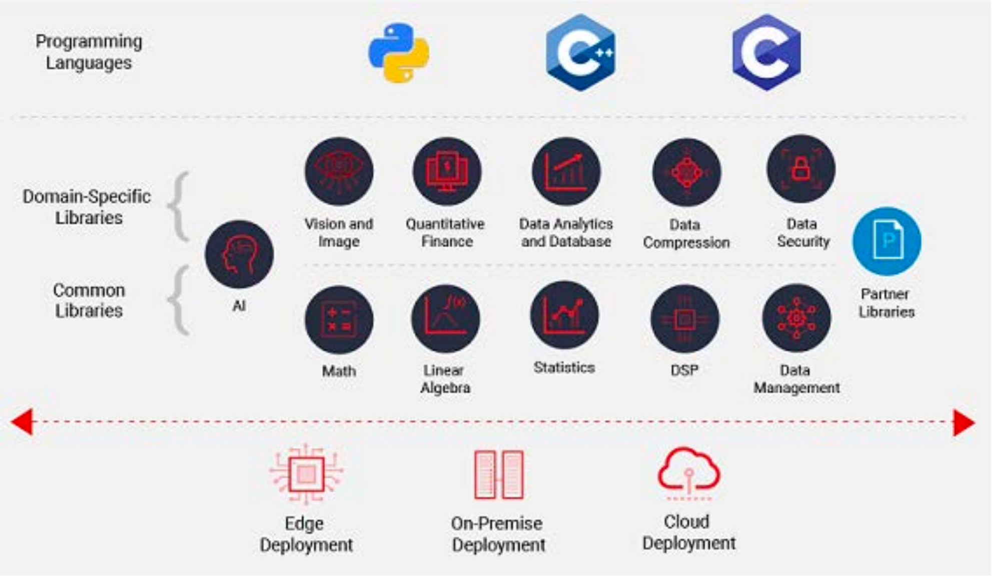
Vitis core development kit

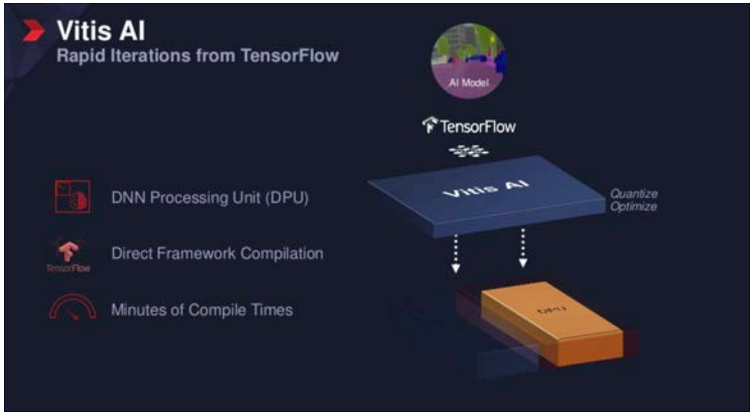
- Compilers
- Analyzers
- Debuggers
- Xilinx runtime library (XRT)

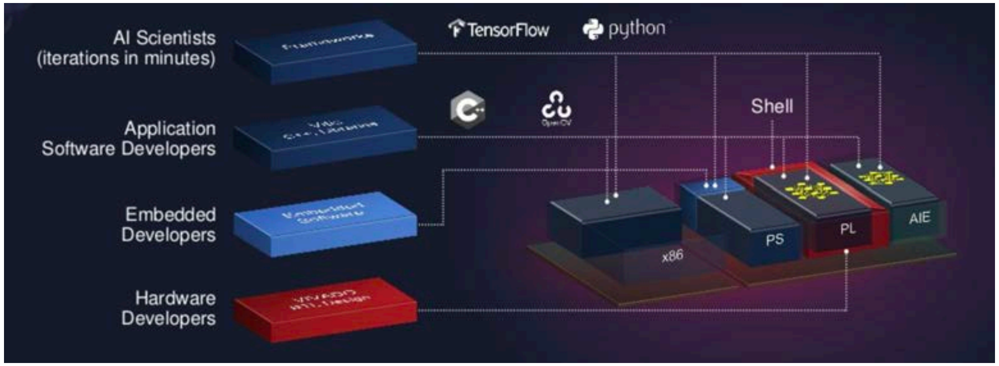
Vitis target platform

Source: Xilinx

Xilinx Vitis Unified Software Platform







https://github.com/Xilinx/Vitis_Libraries

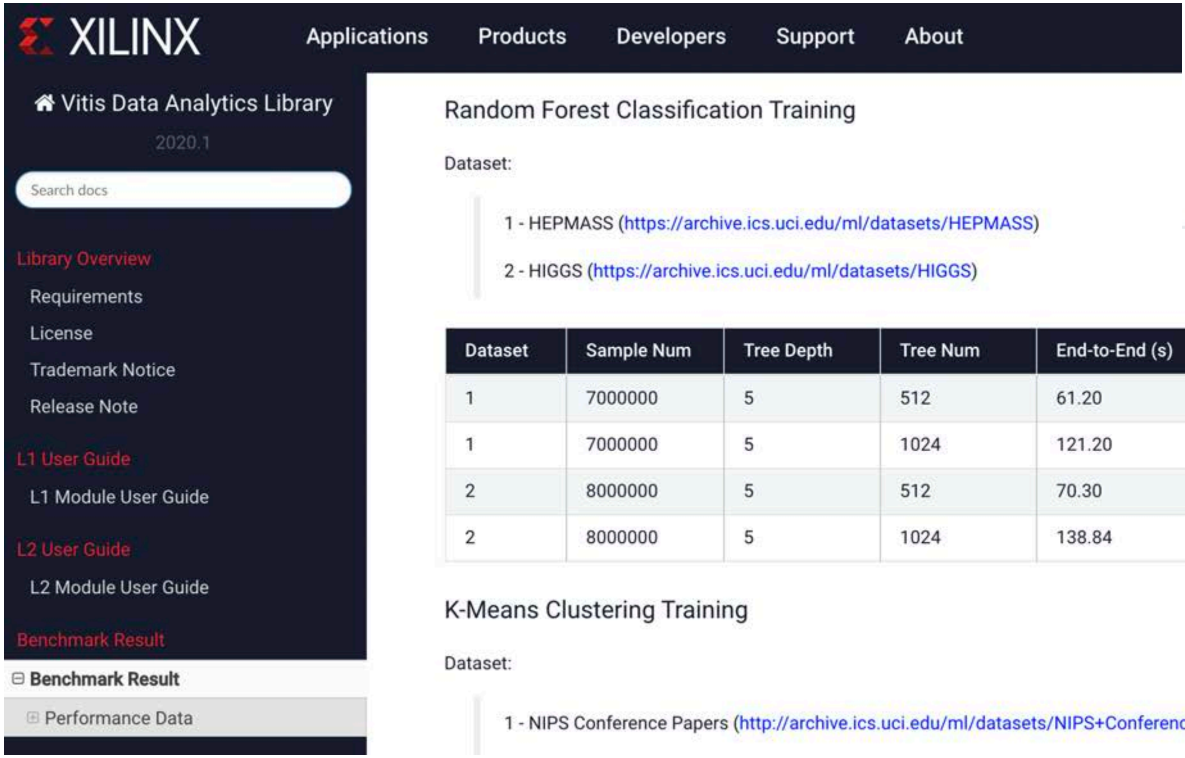
Source: Xilinx

THE UNIVERSITY OF SYDNEY

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



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

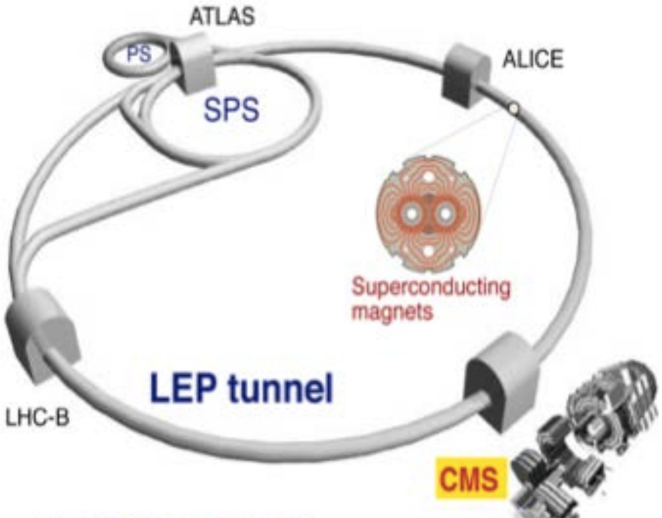
FPGA Technology

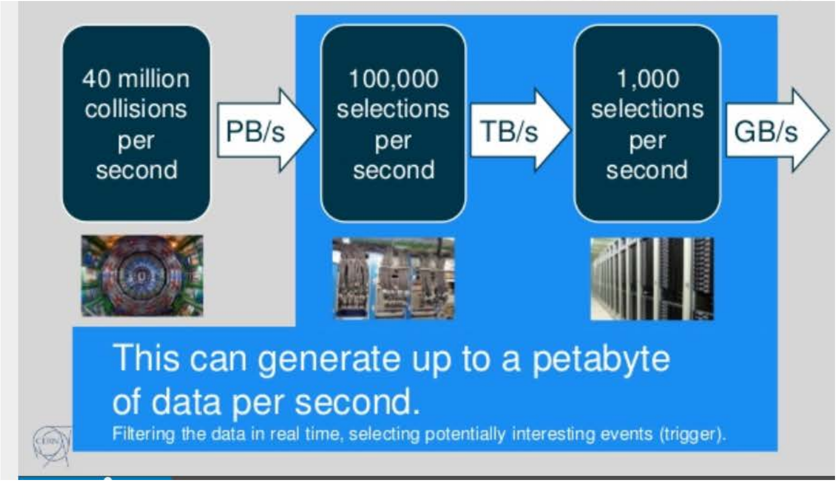
Applications

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)




40 million collisions per second

PB/s

100,000 selections per second

TB/s

1,000 selections per second

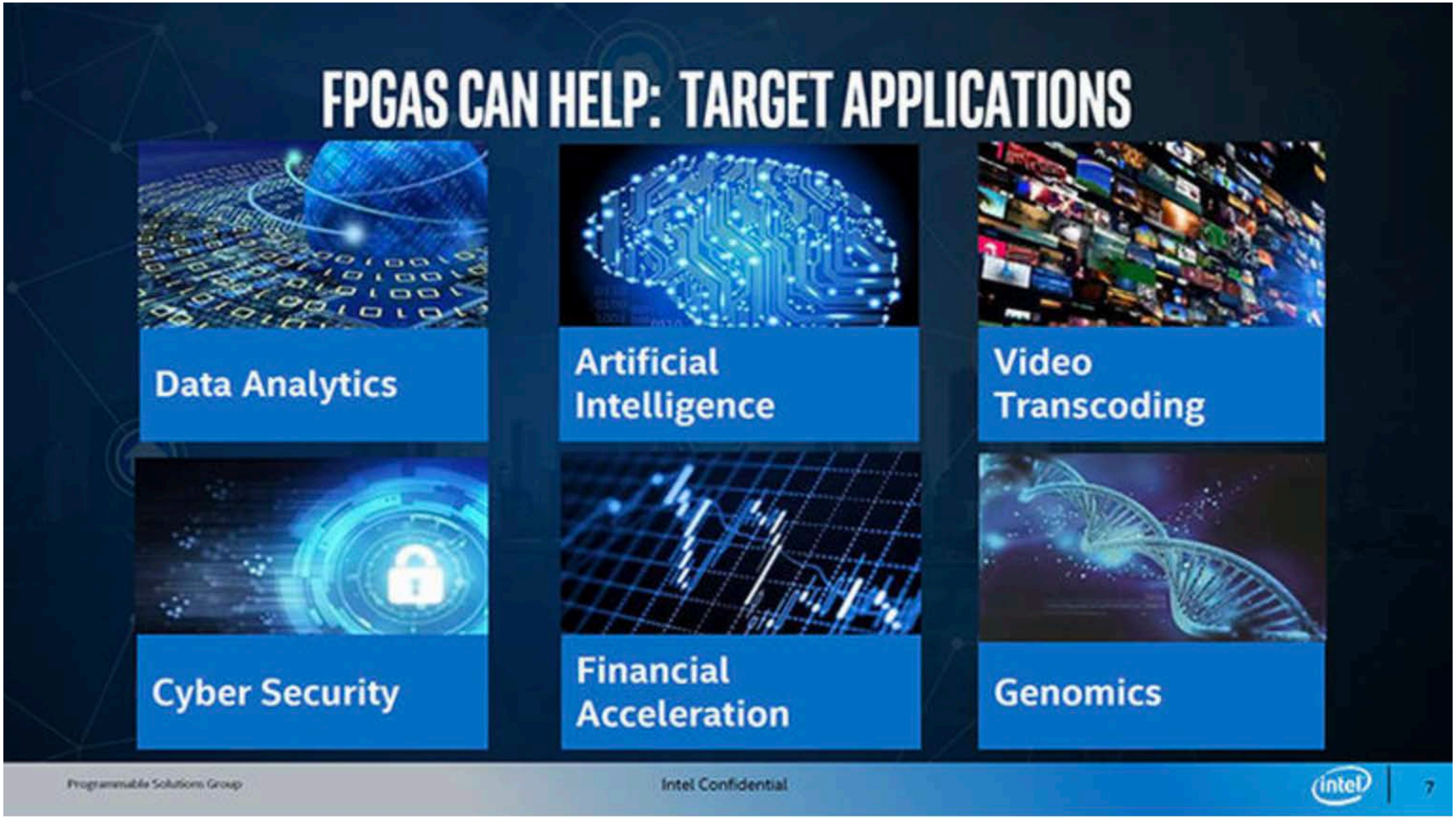
GB/s

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

intel

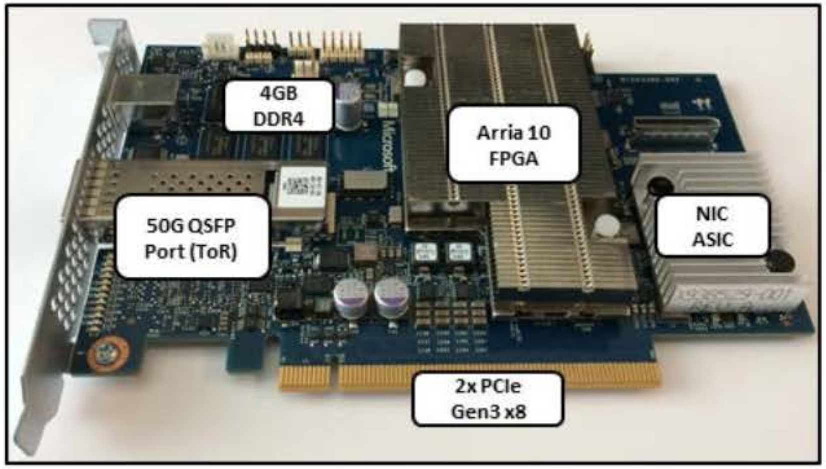
7

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.



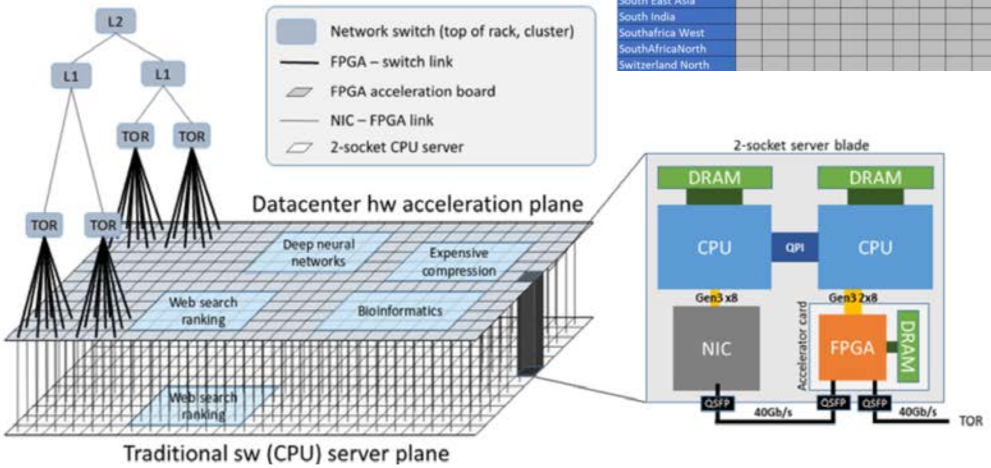
Source: Microsoft

Microsoft Azure Cloud Network

World’s fastest cloud network

Round trip time between Azure regions (milliseconds)	Australia Central	Australia Central2	Australia East	AustraliaSouthEast	Brazil South	Canada Central	Canada East	Central India	Central US	East Asia	East US	East US2	France Central	FranceSouth	Germany North	Germany West Central	Japan East	Japan West	Korea Central	Korea South
Australia Central		2	7	12	312	198	208	144	179	120	205	200	242	228	253	244	132	138	159	158
Australia Central2			8	12	314	200	208	144	180	120	205	200	242	228	252	244	132	138	158	158
Australia East				14	308	194	204	139	184	116	202	198	236	223	248	240	154	162	154	152
AustraliaSouthEast					319	206	214	137	197	120	214	209	234	222	246	238	154	161	158	162
Brazil South						130	139	304	144	321	117	114	186	200	200	190	262	268	302	300
Canada Central							12	210	20	198	25	28	92	106	108	100	153	160	178	178
Canada East								220	30	208	34	38	102	116	116	108	162	170	188	188
Central India									226	86	198	200	118	106	130	122	123	128	126	132
Central US										177	24	28	102	110	116	106	132	139	158	158
East Asia											202	196	184	170	194	186	49	52	43	49
East US												6	80	90	94	86	154	162	182	182
East US2													80	90	98	88	150	156	178	176
France Central														11	20	10	218	224	224	228
FranceSouth															24	16	206	212	210	216
Germany North																10	230	236	234	240
Germany West Central																	221	228	226	232
Japan East																		8	30	30
Japan West																			36	36
Korea Central																				8
Korea South																				
North Central US																				
North Europe																				
Norway East																				
Norway West																				
South Central US																				
South East Asia																				
South India																				
Southafrica West																				
SouthAfricaNorth																				
Switzerland North																				

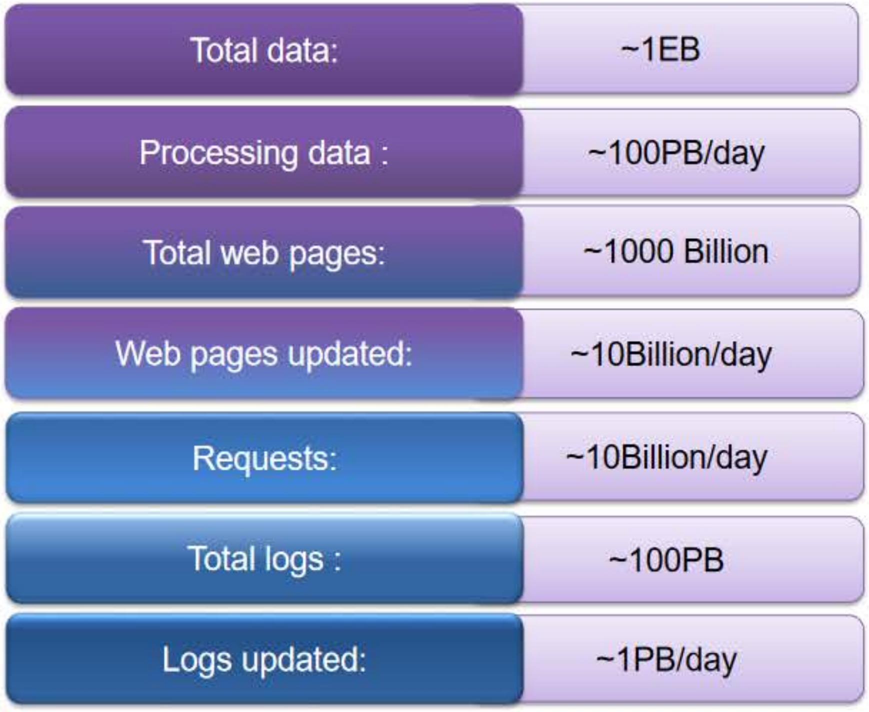
Round trip time between Azure regions (milliseconds)	North Central US	North Europe	Norway East	Norway West	South Central US	South East Asia	South India	Southafrica West	SouthAfricaNorth	Switzerland North	Switzerland West	UAE Central	UAE North	UK South	UK West	West Central US	West Europe	West India	West US	West US2
Australia Central	186	271	272	264	174	92	124	400	390	238	236	172	170	246	252	164	250	146	142	162
Australia Central2	186	272	272	264	176	92	124	400	390	238	236	170	170	246	250	166	250	146	143	164
Australia East	190	266	266	258	172	88	120	396	384	234	230	166	166	242	246	170	246	140	146	168
AustraliaSouthEast	204	250	264	256	184	86	118	394	382	232	228	164	164	239	244	182	244	138	158	180
Brazil South	138	172	208	202	140	330	324	324	352	196	194	294	298	180	182	158	188	302	172	180
Canada Central	14	80	116	110	42	218	232	232	258	104	100	200	204	86	88	36	95	208	58	56
Canada East	24	88	126	118	52	226	240	240	268	114	110	210	214	96	98	44	104	218	67	66
Central India	219	133	148	140	226	54	24	278	266	116	112	30	28	123	128	240	128	4	222	214
Central US	8	88	124	124	22	196	230	246	274	112	110	216	219	96	97	14	102	223	40	36
East Asia	184	197	213	206	171	34	66	342	332	180	176	113	112	188	193	162	192	88	148	142
East US	19	66	102	98	32	219	220	220	246	90	88	188	192	74	76	40	81	196	64	61
East US2	22	71	106	100	26	215	222	224	252	94	88	190	194	78	80	46	86	198	58	68
France Central	96	16	31	24	106	150	138	152	174	14	10	108	112	7	9	118	10	116	138	140
FranceSouth	104	26	40	36	116	136	126	174	162	10	8	96	98	16	18	126	20	102	148	146
Germany North	110	29	20	26	124	162	150	168	186	16	19	120	124	24	26	130	12	128	156	150
Germany West Central	102	20	24	20	114	152	142	160	178	6	10	112	114	14	16	124	8	120	146	146
Japan East	140	220	248	246	124	68	102	372	366	216	212	148	148	226	228	118	229	128	105	97
Japan West	146	228	254	246	130	76	108	380	372	222	218	154	154	231	236	124	234	128	106	104
Korea Central	166	248	254	246	152	74	106	382	372	220	216	154	154	228	234	144	232	128	130	122
Korea South	165	248	258	252	152	80	112	388	376	226	222	159	162	234	238	144	238	132	130	122
North Central US		84	118	118	30	204	238	240	267	108	104	208	212	90	91	24	100	216	52	46
North Europe			38	32	97	164	153	156	185	26	24	123	127	11	12	106	16	130	129	126
Norway East				8	132	180	168	172	204	30	34	138	142	30	32	138	22	146	164	159
Norway West					128	172	162	166	198	26	30	130	134	22	22	138	15	138	160	160
South Central US						190	224	250	280	120	116	217	220	104	106	22	112	224	34	44
South East Asia							34	310	298	148	144	80	80	154	160	182	160	56	169	161
South India								299	286	136	132	50	48	144	148	214	148	26	200	193
Southafrica West									18	164	160	268	272	146	149	260	156	275	282	282
SouthAfricaNorth										172	168	256	260	176	176	288	184	264	310	310
Switzerland North											4	106	110	20	22	130	12	114	152	151



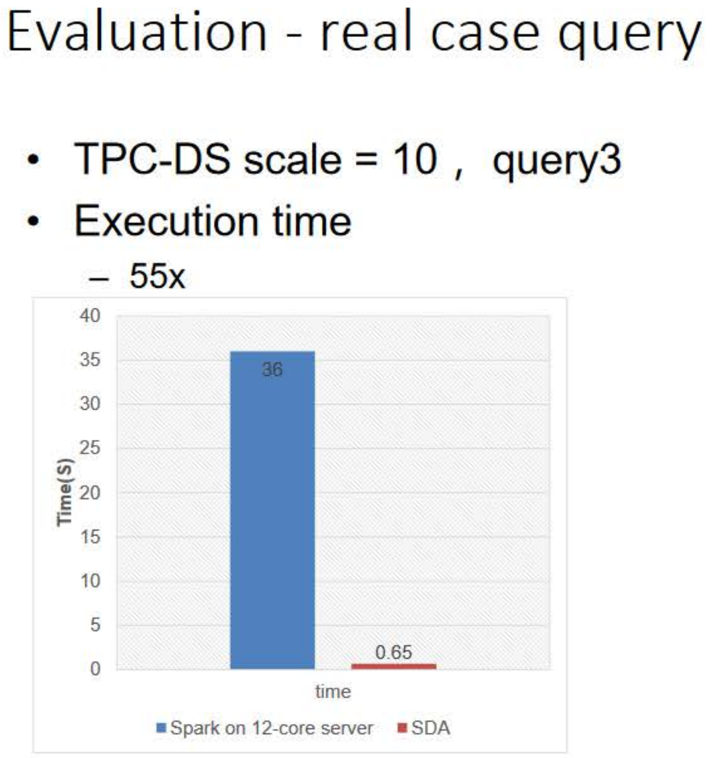
Source: Microsoft https://docs.microsoft.com/en-us/azure/networking/azure-network-latency

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

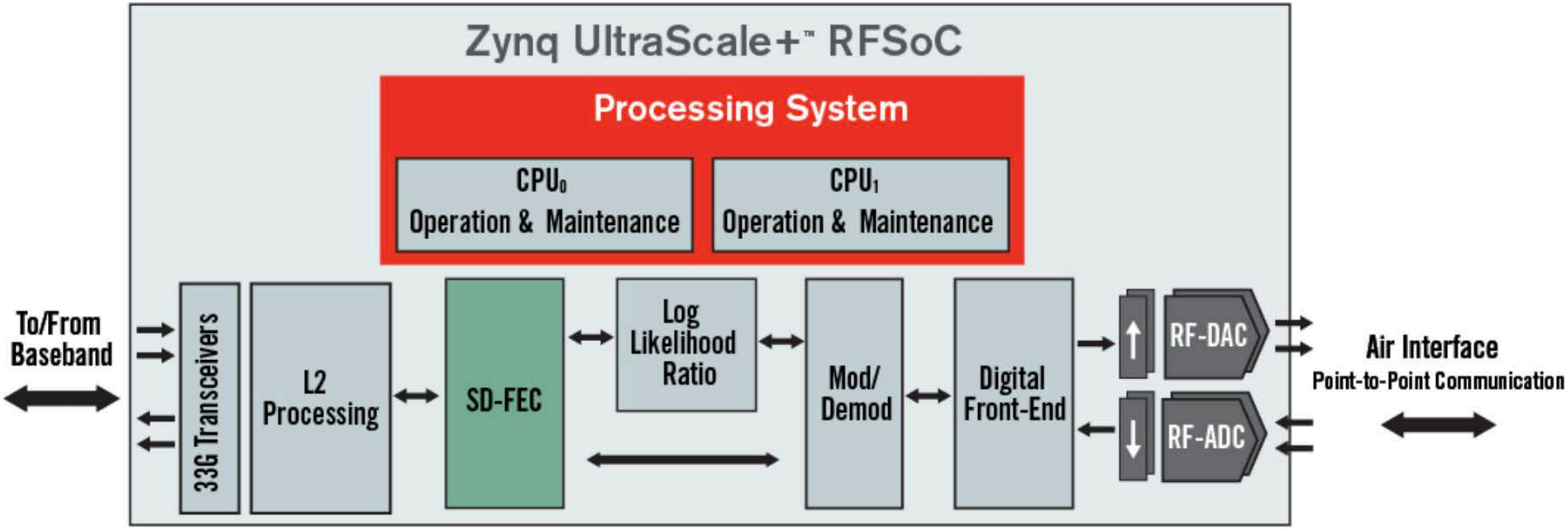
- TPC-DS scale = 10， query3
- Execution time
  - 55x

Source: Baidu 2016 Hot Chips

Mobile Backhaul

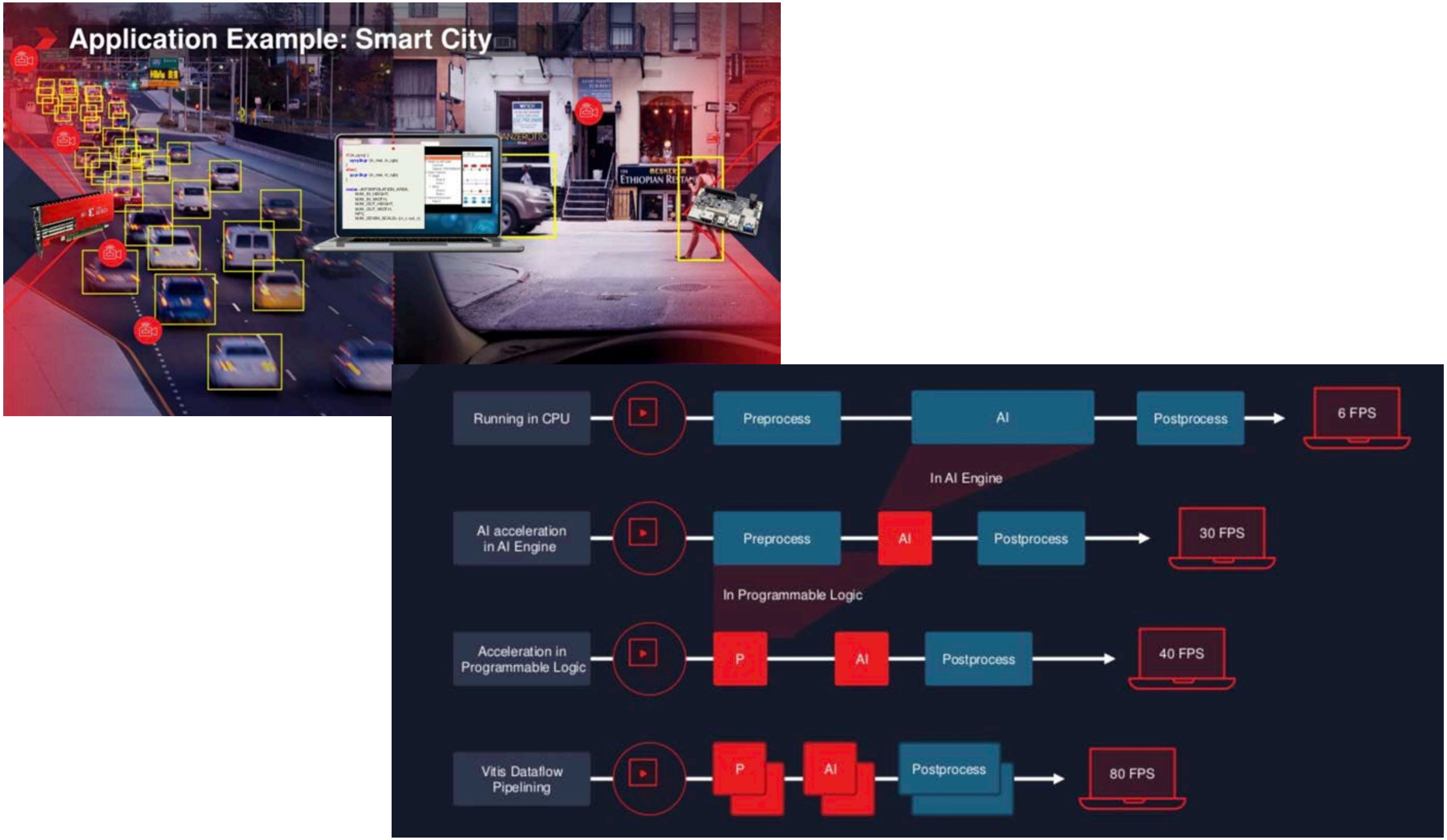
Key Zynq UltraScale+ RFSoC Benefits:

- 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



Application Example: Smart City

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

In AI Engine

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

In Programmable Logic

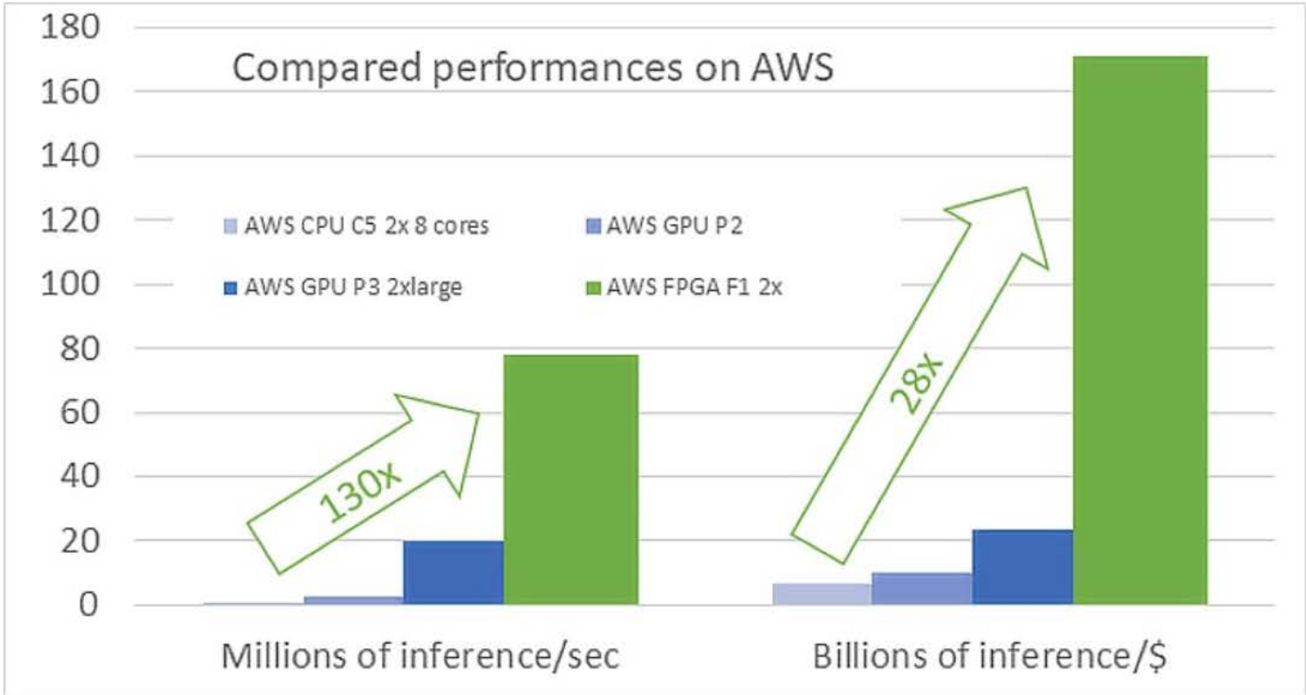
Acceleration in Programmable Logic → P → AI → Postprocess → 40 FPS

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

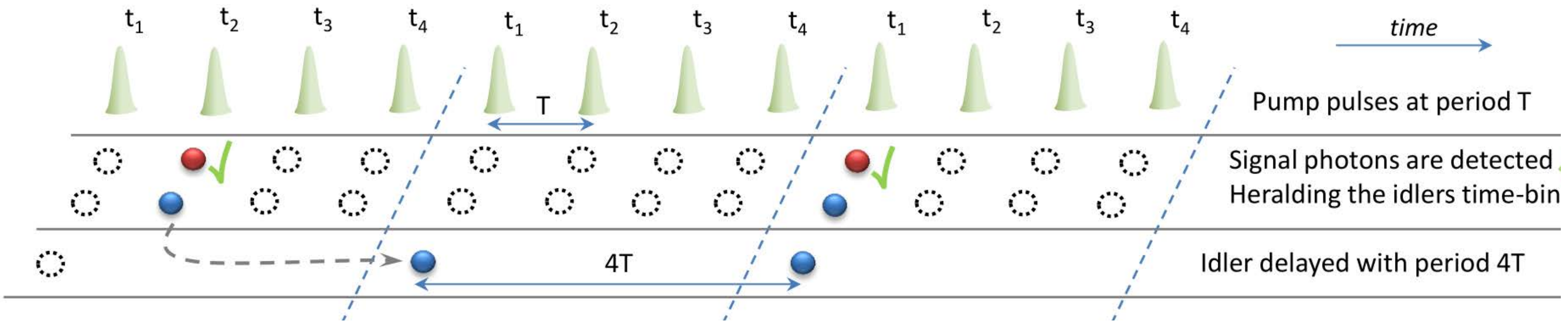
Overview

FPGA Technology

Applications

Our work

Time domain multiplexing of single photons

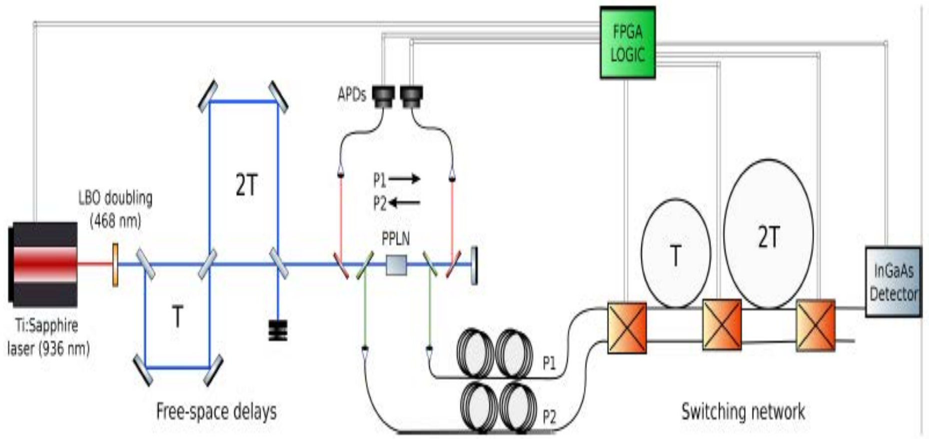


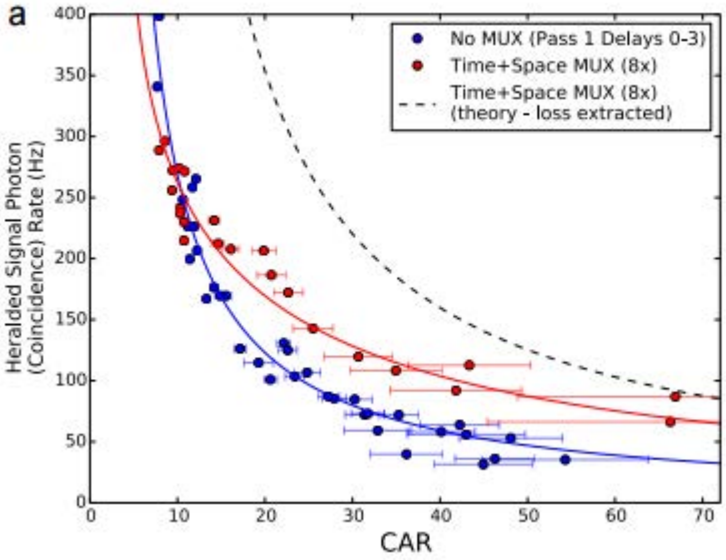
Pump pulses at period T

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%

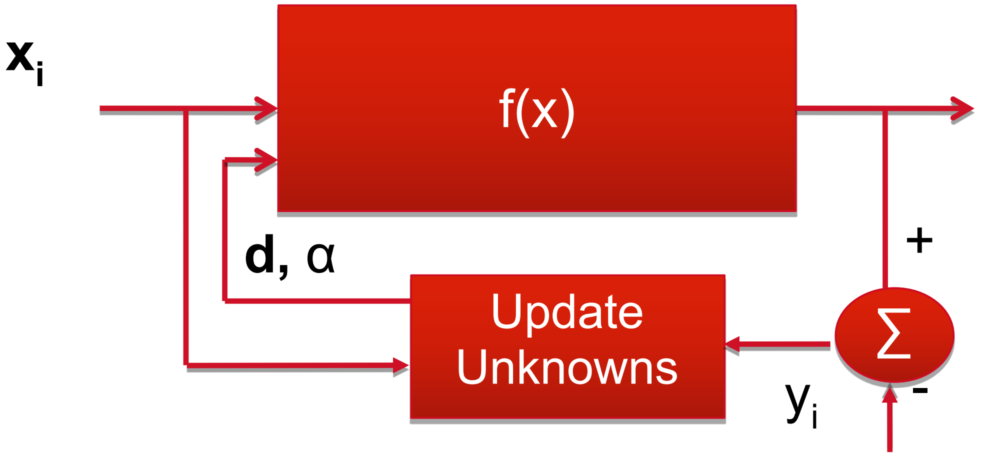
http://phwl.org/assets/papers/atm_ncomms16.pdf

Time Multiplexing of Single Photons

THE UNIVERSITY OF SYDNEY

Exploration: Kernel Methods

ARC Linkage with Exablaze



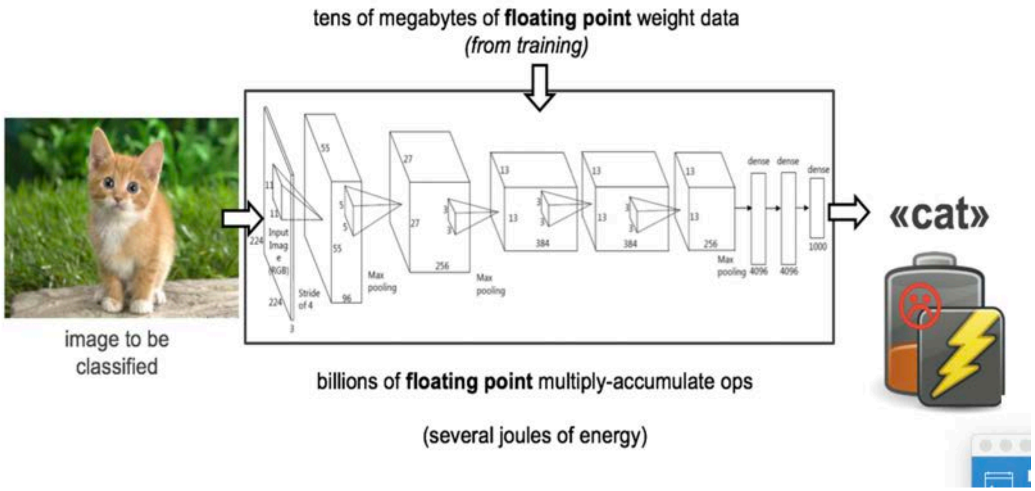
› A family of kernel methods that can do simultaneous learning and inference

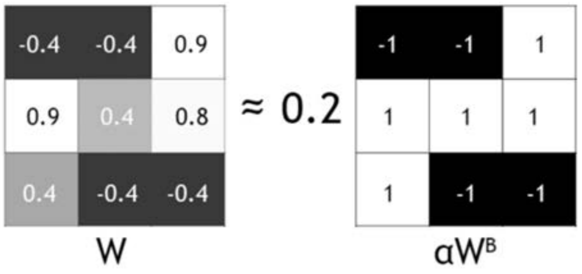
- Highest reported throughput 80 Gbps (TRETS’17)
- Lowest reported latency 80 ns (FPT’15)
- Highest capacity (FPGA’18)

http://phwl.org/assets/papers/dknlms_trets20.pdf

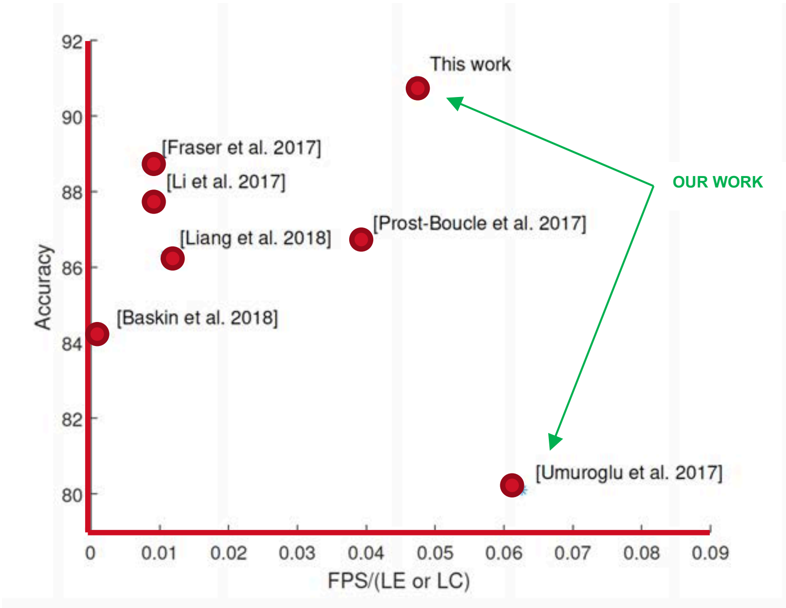
Parallelism: Binarized Neural Networks

Collaboration with Xilinx


tens of megabytes of **floating point** weight data
(from training)
image to be classified
«cat»
billions of **floating point** multiply-accumulate ops
(several joules of energy)



$$\begin{bmatrix} -0.4 & -0.4 & 0.9 \\ 0.9 & 0.4 & 0.8 \\ 0.4 & -0.4 & -0.4 \end{bmatrix} \approx 0.2 \begin{bmatrix} -1 & -1 & 1 \\ 1 & 1 & 1 \\ 1 & -1 & -1 \end{bmatrix}$$



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

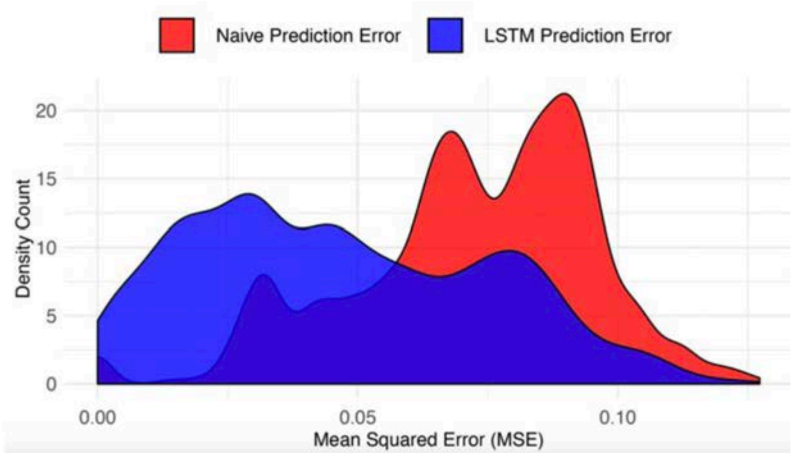
http://phwl.org/assets/papers/bnn_fpga17.pdf

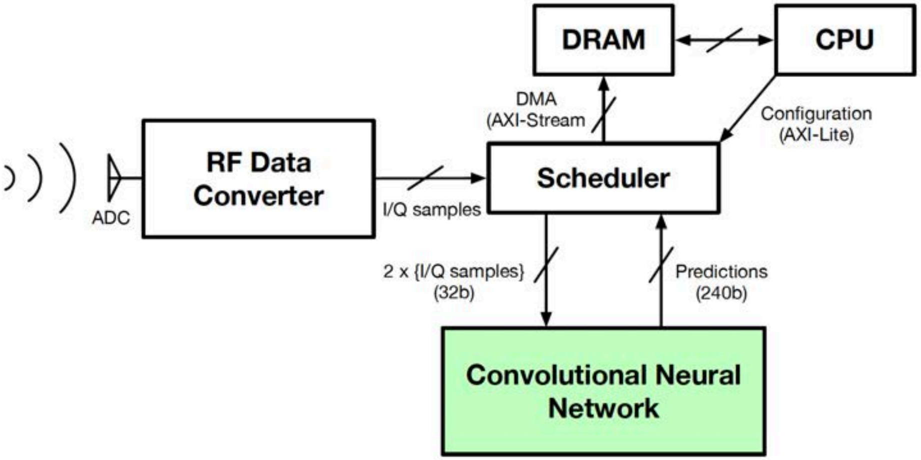
Integration: Radio Frequency Signals

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)

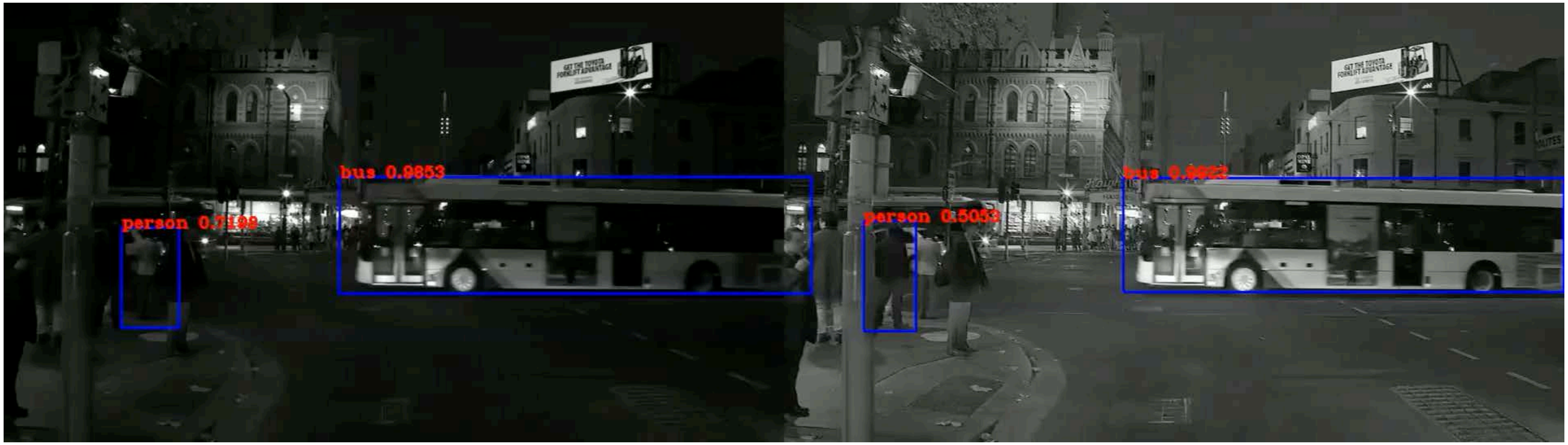
Ternary Modulation classifier: 488K class/s, 8us latency, Xilinx ZCU111 RFSoC (FPT'19)

http://phwl.org/assets/papers/amc_raw20.pdf

Customisation: High Dynamic Range Signals

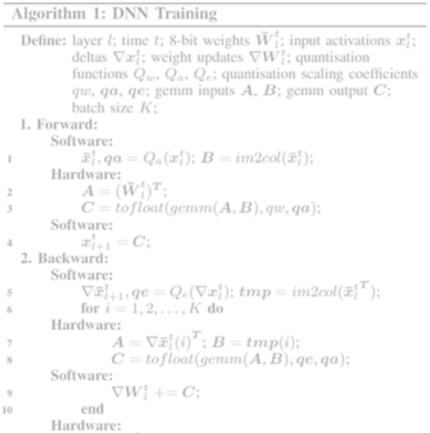
Defence Innovation Hub

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

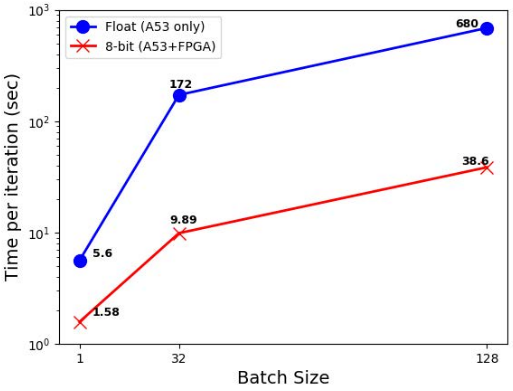


On-FPGA Training

See paper for details



```
Algorithm 1: DNN Training
Define: layer $l$; time $t$; 8-bit weights $\bar{W}_l^t$; input activations $x_l^t$;
        deltas $\nabla x_l^t$; weight updates $\nabla W_l^t$; quantisation
        functions $Q_w$, $Q_a$, $Q_e$; quantisation scaling coefficients
        $qw$, $qa$, $qe$; gemm inputs $A$, $B$; gemm output $C$;
        batch size $K$;
1. Forward:
    Software:
        $\bar{x}_l^t, qa = Q_a(x_l^t)$; $B = im2col(\bar{x}_l^t)$;
    Hardware:
        $A = (\bar{W}_l^t)^T$;
        $C = tofloat(gemm(A, B), qw, qa)$;
    Software:
        $x_{l+1}^t = C$;
2. Backward:
    Software:
        $\nabla \bar{x}_{l+1}^t, qe = Q_e(\nabla x_l^t)$; $tmp = im2col(\bar{x}_l^{t\,T})$;
        for $i = 1, 2, \ldots, K$ do
    Hardware:
            $A = \nabla \bar{x}_l^t(i)^T$; $B = tmp(i)$;
            $C = tofloat(gemm(A, B), qe, qa)$;
    Software:
            $\nabla W_l^t$ += $C$;
        end
    Hardware:
        $A = \bar{W}_l^t$; $B = \nabla \bar{x}_l^t$
        $C = tofloat(gemm$
    Software:
        $\nabla x_l^t = col2im(C)$;
```



- 17x speed-up over ARM

FPGA

- Low-Precision (8-bit)
  - All matrix multiplications
  - >95% of DNN operations

ARM

- High-Precision
  - Everything else!
  - Of particular importance is the **weight update and gradient accumulator**
- Suits a Zynq platform
  - Fast DDR, shared between PL and floating-point

http://phwl.org/assets/papers/lptrain_fpt19.pdf

Overview

FPGA Technology

Applications

Our work

THE UNIVERSITY OF SYDNEY

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

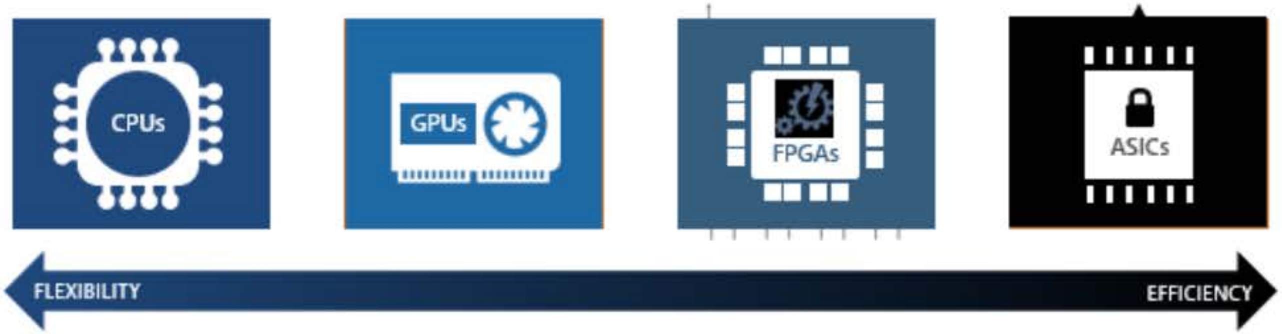


Figure: Microsoft