CruxML

REAL-TIME COMPUTING & MACHINE LEARNING

THE UNIVERSITY OF SYDNEY

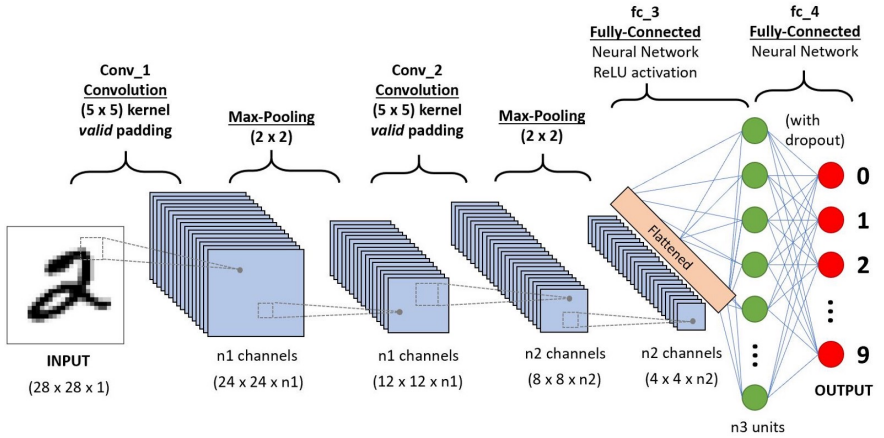
FPGA-based Machine Learning for Communications Applications: A Tutorial

Prof. Philip Leong | 2022

www.cruxml.com

Artificial Intelligence / Machine Learning

- First AI Generation
  - Machines like Deep Blue using expert-engineered features
- Second AI Generation
  - Machines that can learn from massive amounts of data
  - Successful for ill-defined tasks like machine translation, speech recognition and computer vision



Some Challenges in Communications Systems

1. Limited spectrum (defence spectrum requirements increasing exponentially)
2. Understand RF scenes (for surveillance, IED detection, EW)
3. Authentication (security and privacy)

Radio frequency machine learning (RFML) on FPGAs is a promising technology for addressing these challenges with improved SWaP-C

Overview

FPGAs

Challenges

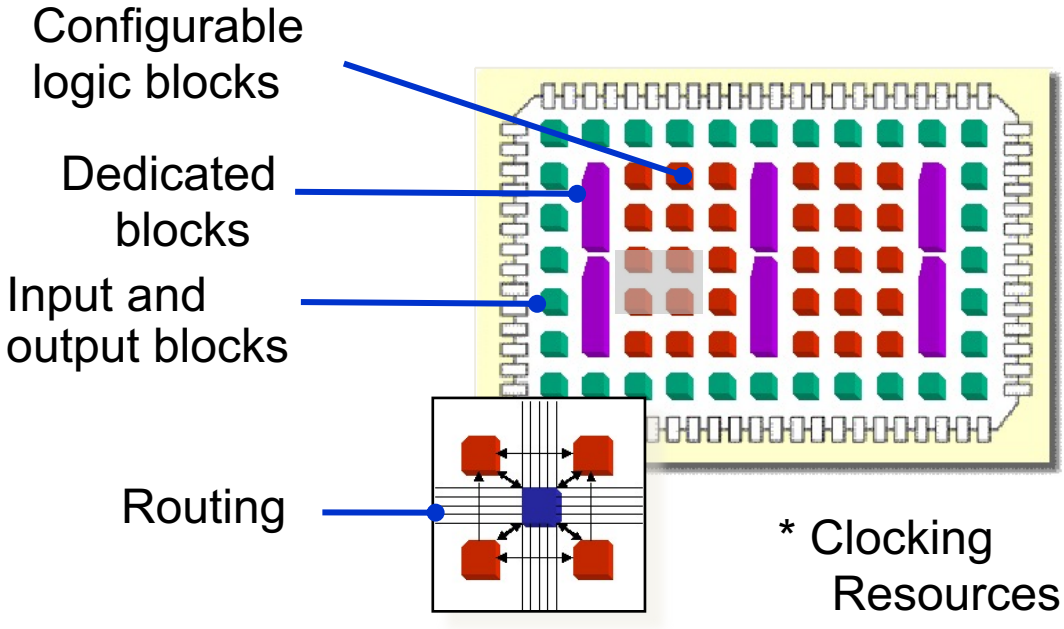
Emerging Technologies

Summary

Background: What is an FPGA?

User-customisable integrated circuit

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

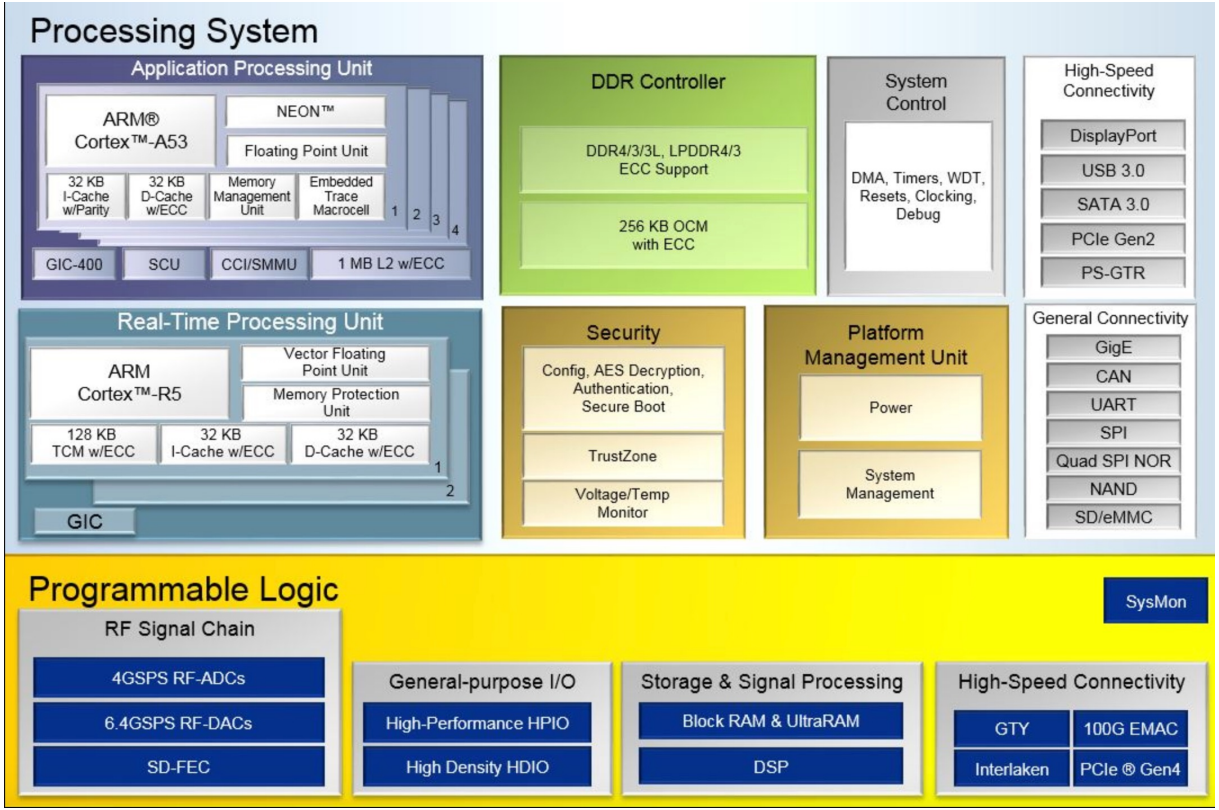


Source: Prof Steve Wilton, UBC

Xilinx RFSoc Device

Most interesting part for RFML is inclusion of

- **4 GSPS ADCs**
- **6 GSPS DACs**
- **FEC**



Source: Xilinx

Motivation for FPGAs (EPIC)

FPGAs commercial off-the-shelf

They offer an opportunity to implement complex algorithms with lower latency, improved SWaP-C and higher reliability

Exploration – agile approach to find solution

Parallelism – arrive at an answer faster

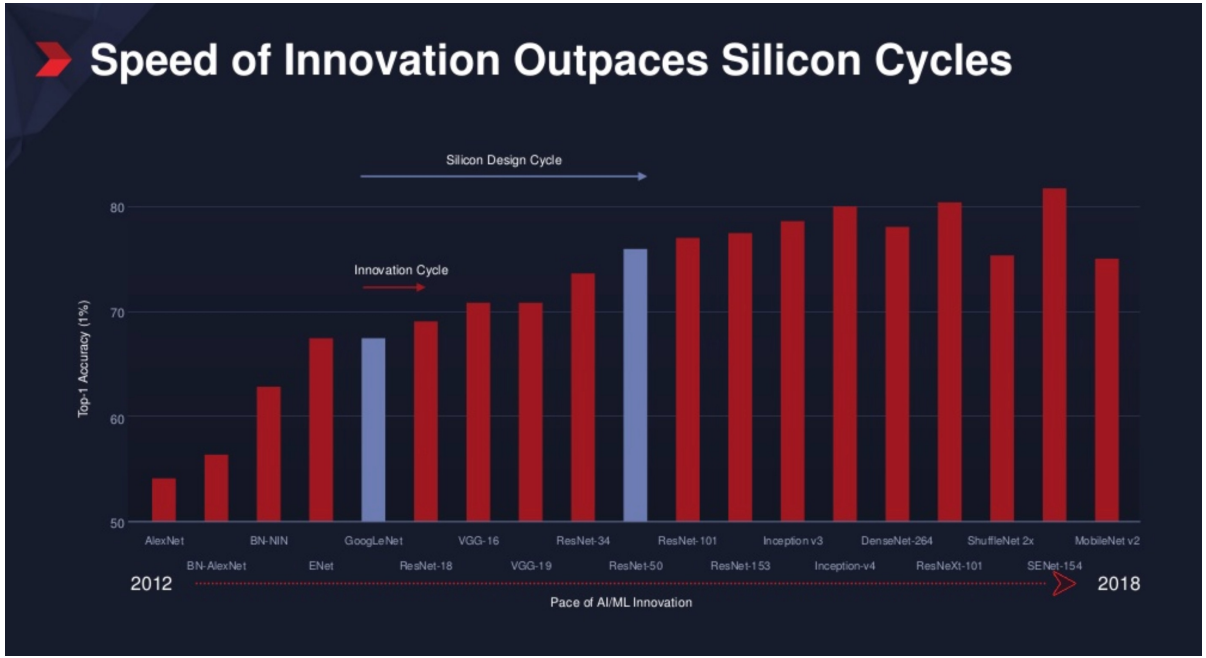
Integration – interfaces are not a bottleneck

Customisation – problem-specific designs to improve efficiency (power, speed, density)

https://cruxml.com/wp-content/uploads/2021/11/FPGA_based_Machine_Learning.pdf

EPIC: Exploration

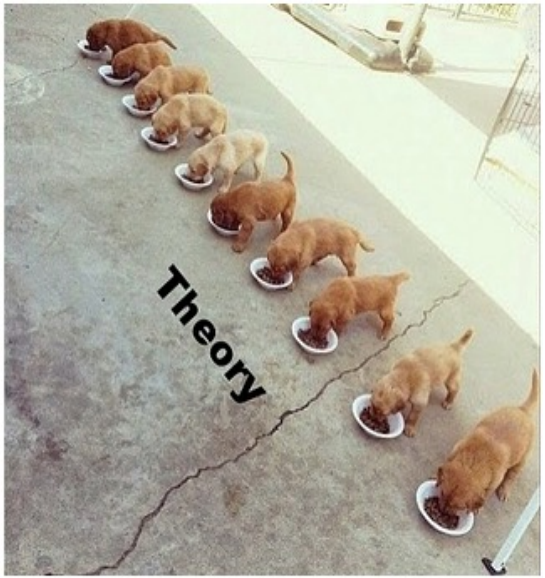
- ML algorithms are improving at a much faster pace than silicon designs can be made
- FPGAs ideal for adapting to rapidly evolving algorithms



Source: Xilinx

EPIC: Parallelism

- Do what would take many cycles on uP in several (instruction level parallelism)
- Execute independent tasks in parallel (spatial parallelism, multithreading)
- Tradeoff latency with throughput by doing things in stages (pipelining)

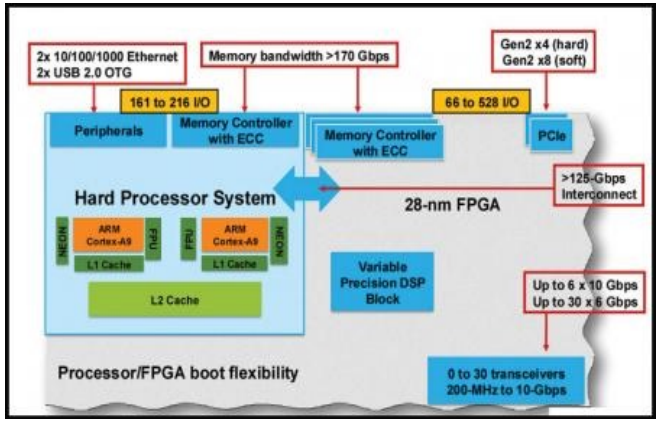


EPIC: Parallelism

- Do what would take many cycles on uP in several (instruction level parallelism)
- Execute independent tasks in parallel (spatial parallelism, multithreading)
- Tradeoff latency with throughput by doing things in stages (pipelining)

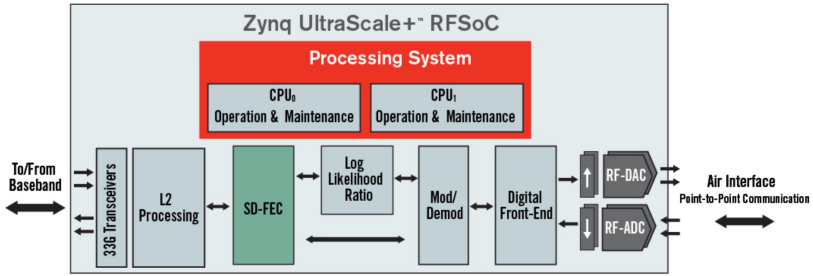
EPIC: Integration

- Networking, chip IO, data conversion and computation on same device
  - Reduction of buffering can help latency
  - Single chip operation massive interconnect within chip exploited
  - Multiple (small) memories within FPGA offer enormous memory bandwidth
  - Integrated data converters enable low-power RFML



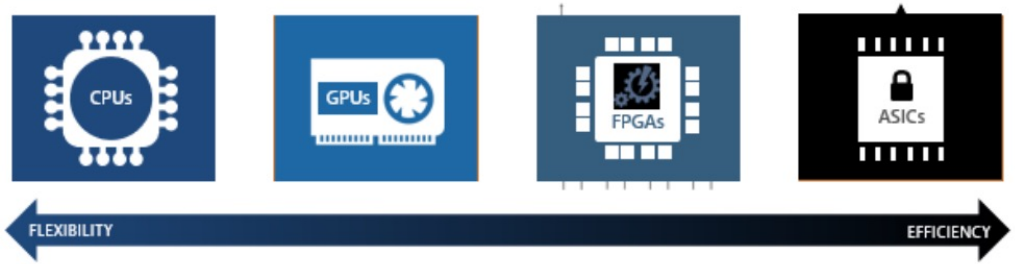
Key Zynq UltraScale+ RFSoC Benefits:

- 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



EPIC: Customisation

- More specific functions can be implemented more efficiently (approx. 10x each step to the right)
- Too expensive to design ASIC to perform very specialised function
- FPGAs can be heavily customised due to their programmability i.e. so only does one thing efficiently

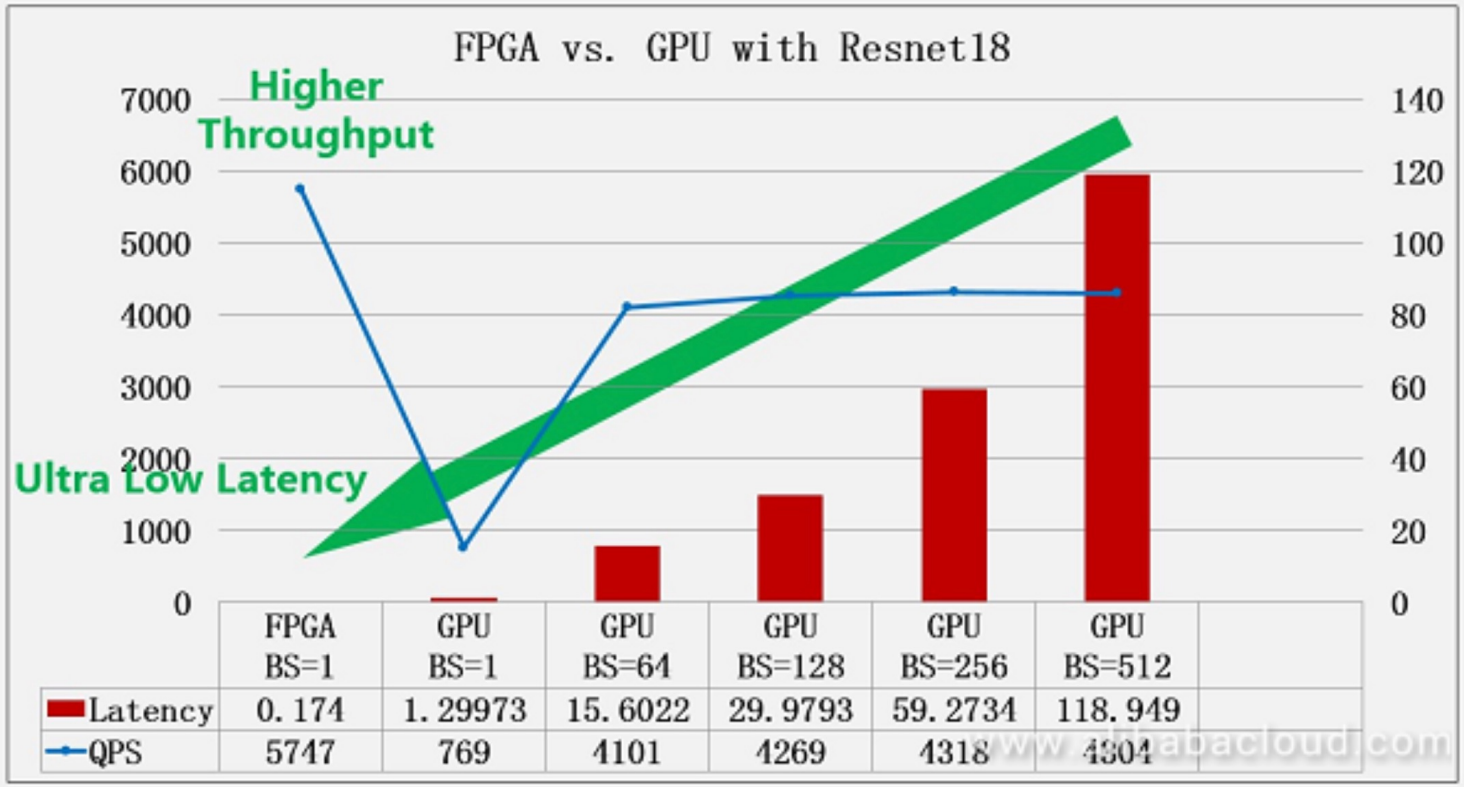


Source: Microsoft

Background: Latency vs Throughput

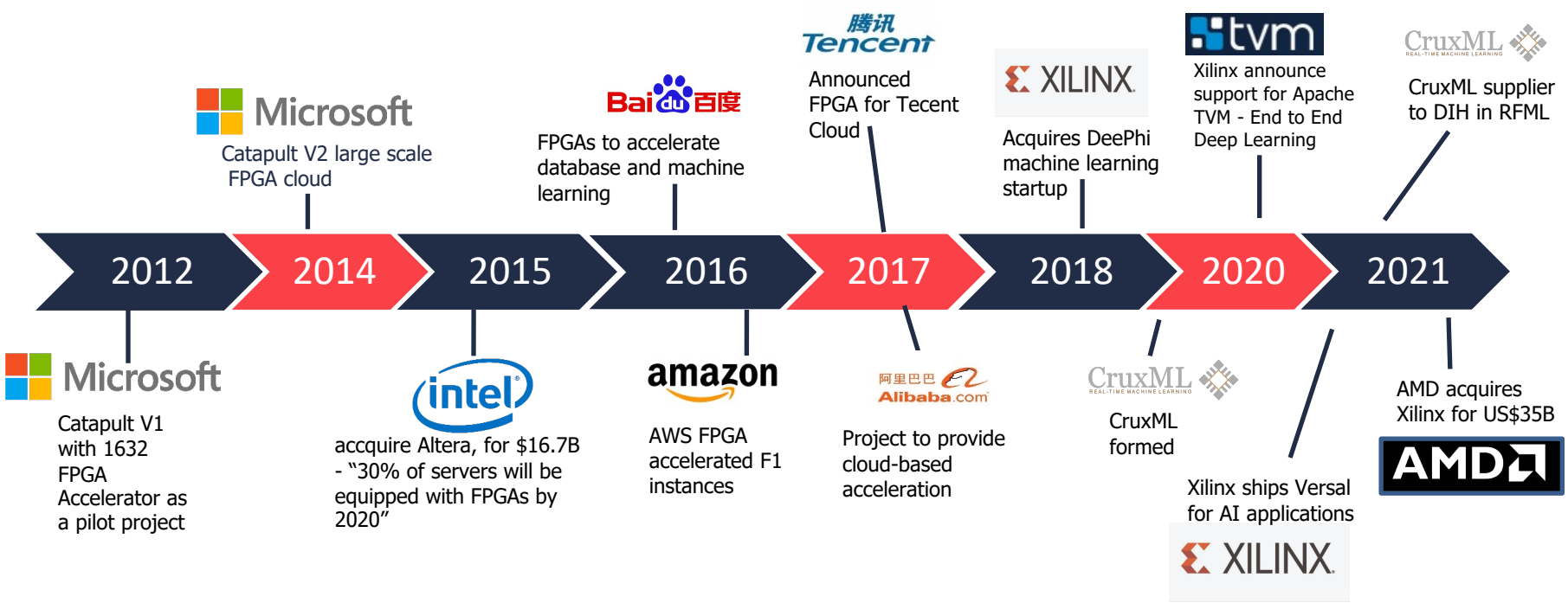
- Latency is reponse time. Throughput (or bandwidth) is capacity (messages per second)
- E.g. London to NY flight: Concorde had lowest latency (~ 2x); max throughput Boeing 747 (~ 4x) (travel time vs passengers per hour)
- Latency crucial in communications and defence applications

FPGA vs GPU Throughput and Latency



Source: https://www.alibabacloud.com/blog/ultra-low-latency-and-high-performance-deep-learning-processor-with-fpga_593942

Recent Uptake in FPGA Technology



Overview

FPGAs

Challenges

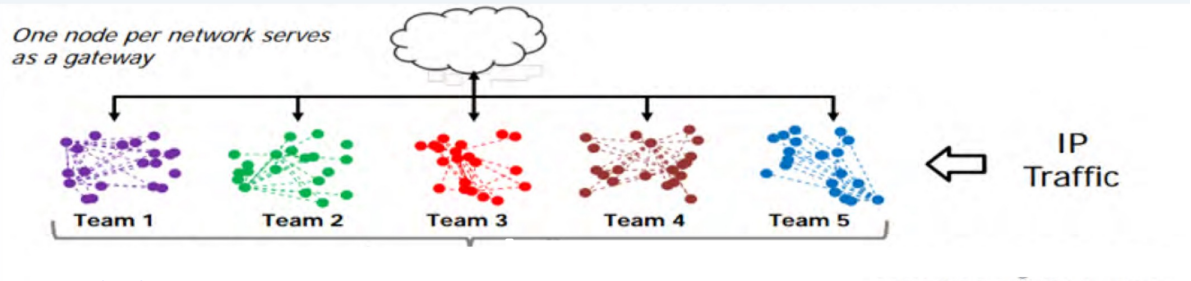
Emerging Technologies

Summary

1. Limited spectrum
2. Understand RF scenes
3. Authentication

1. Spectrum - DAPRA Spectrum Challenge 2 (SC2, 2019)

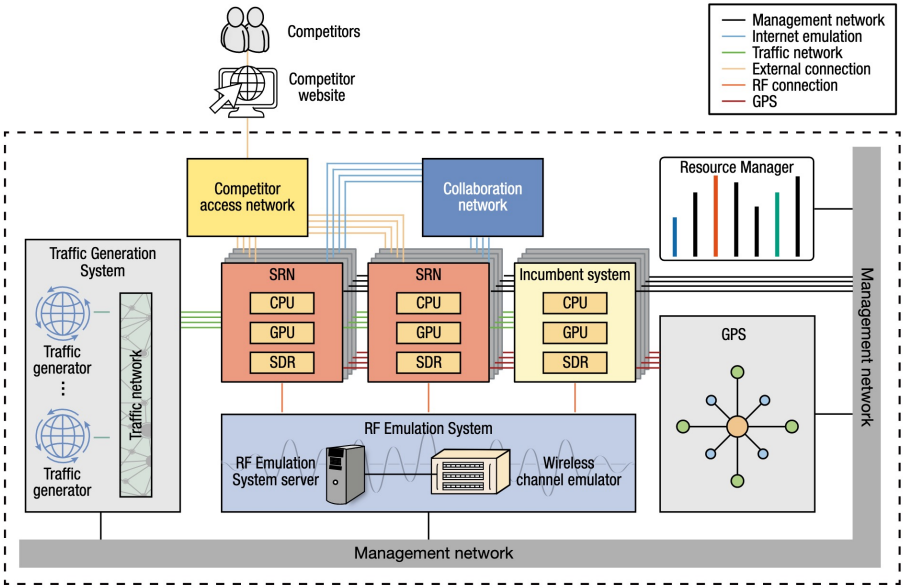
- Collaborative machine-learning competition to address increasingly crowded RF spectrum. Take humans out of loop and shrink time scales from years to seconds
- Score points by
  - Delivering traffic flows
  - Achieving cooperative objective (score limited to lowest score of all times unless all teams exceed threshold)
- Share information about channel usage, radio locations and performance
  - Teams don't need to report their true score



https://idstch.com/technology/ict/dod-developing-spectrum-management-tools-for-ensuring-access-in-congested-and-contested-environment/

1. Spectrum - DAPRA Spectrum Challenge 2 (SC2, 2019)

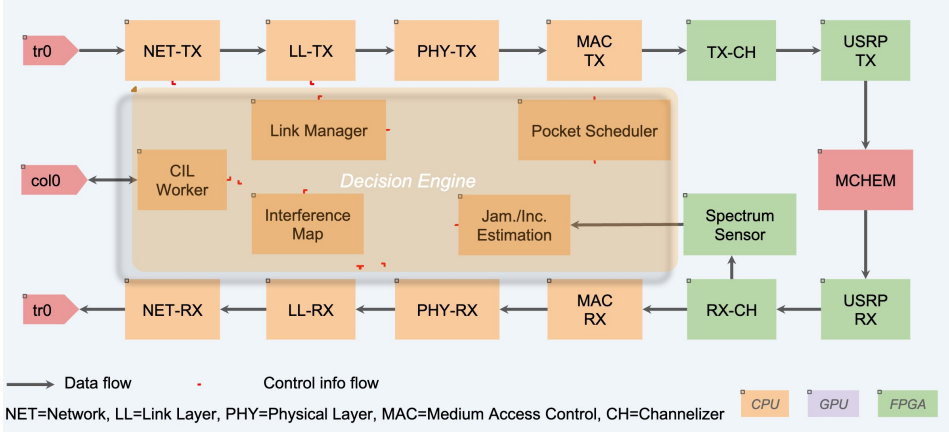
- DARPA Colosseum (21 rack) RF simulation environment
  - Largest RF test bed, Ettus X310 SDRs
  - Large scale channel emulator on heterogeneous FPGA/GPU/CPU cluster

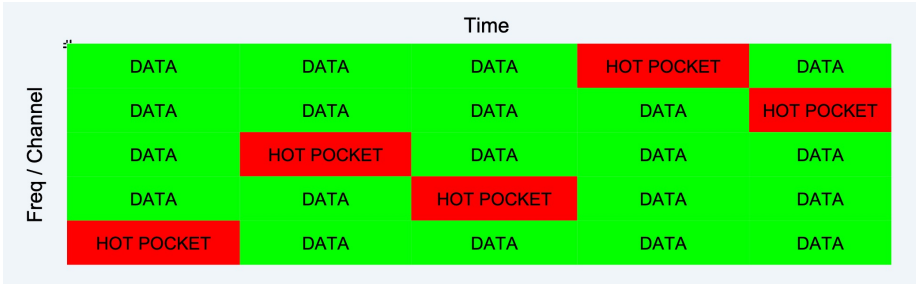


https://idstch.com/technology/ict/dod-developing-spectrum-management-tools-for-ensuring-access-in-congested-and-contested-environment/

1. Spectrum – Winner University of Florida

- Reinforcement learning
  - Rewarded when it succeeds, penalised when it fails
- Everything is adaptive:
  - PHY: Acquisition, Modulation, Coding, TX Power, RX Gain
  - LL: Channels and Time Slots/Channel, Mapping of SRCs to Time Slots
  - NET: Supported flows, admission control granularity down to individual files/bursts
  - Other: Channels to jam

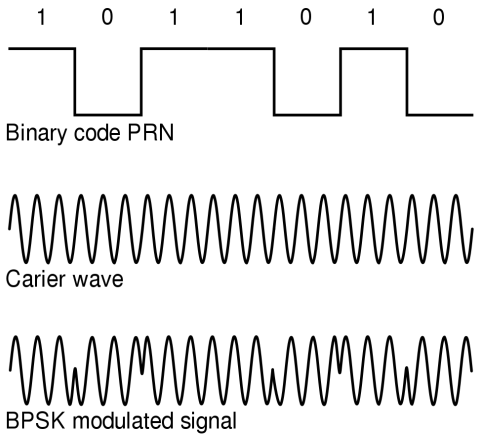


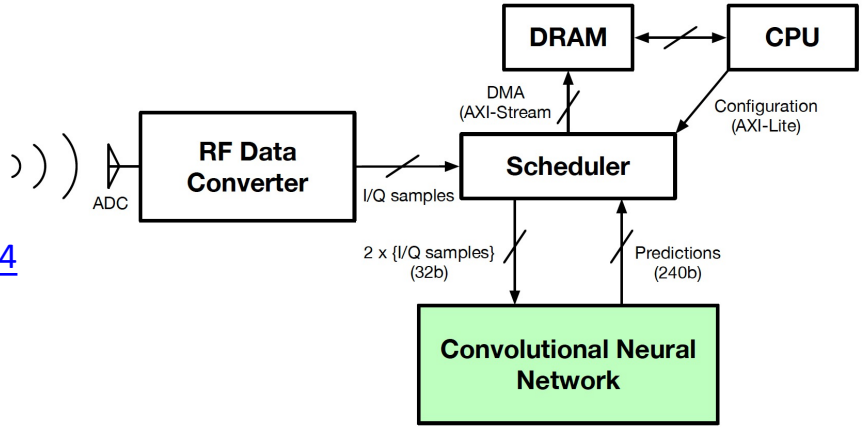


https://par.nsf.gov/servlets/purl/10202115

2. Understand RF scenes – Modulation Classification

- Automatic modulation classification (AMC) from raw IQ samples
- Application: e.g. IED detection
- Very difficult under low SNR conditions
- ITU Challenge on same problem
  https://challenge.aiforgood.itu.int/match/matchitem/34



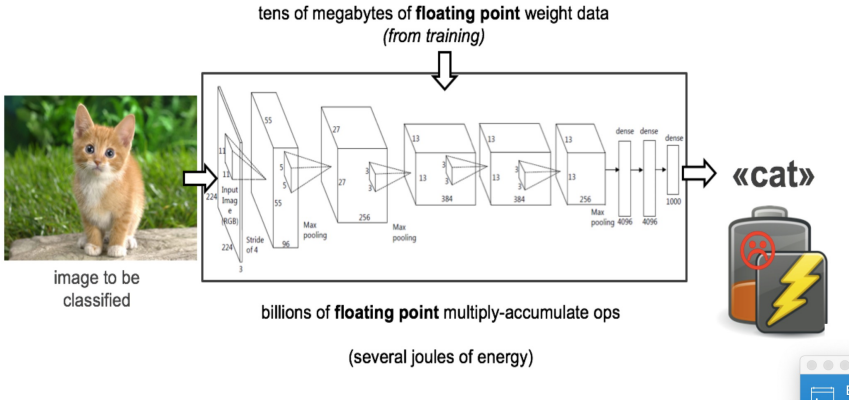


Ternary Modulation classifier: 488K class/s, 8us latency, Xilinx ZCU111 RFSoC (FPT'19)

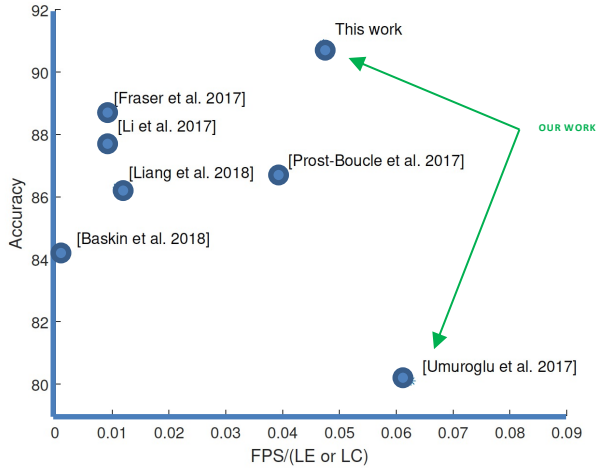
http://phwl.org/assets/papers/amc_raw20.pdf

2. Understand RF scenes – Low-Precision Neural Networks

Collaboration with Xilinx



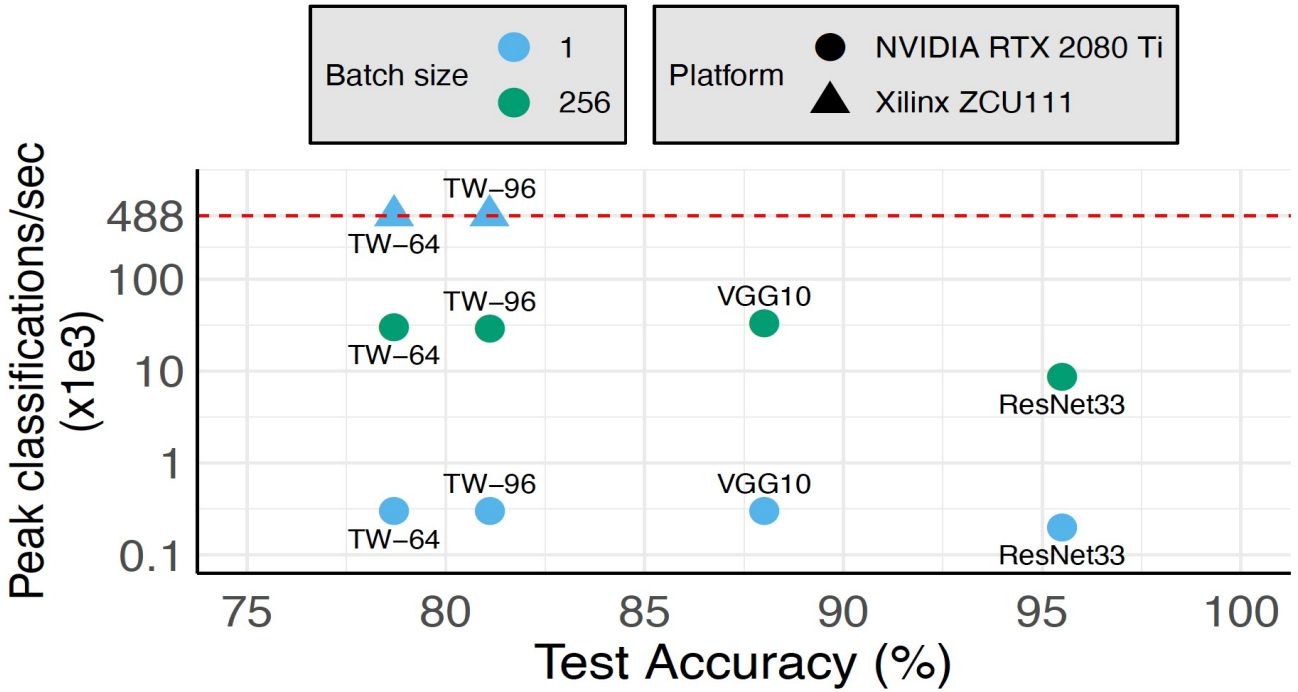
Source: Xilinx



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

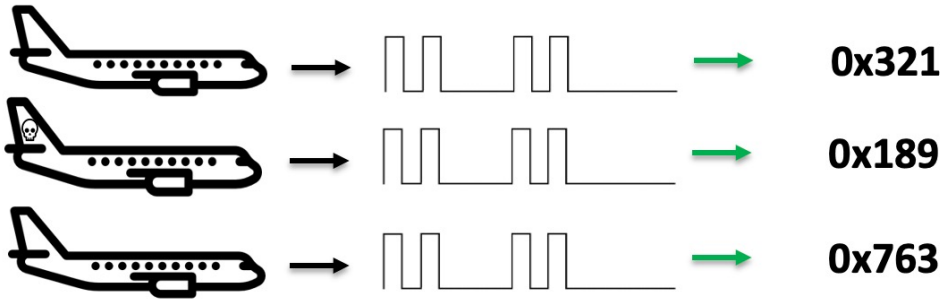
http://phwl.org/assets/papers/bnn_fpga17.pdf

2. Understand RF scenes – Accuracy



3. Authentication - Specific Emitter Identification

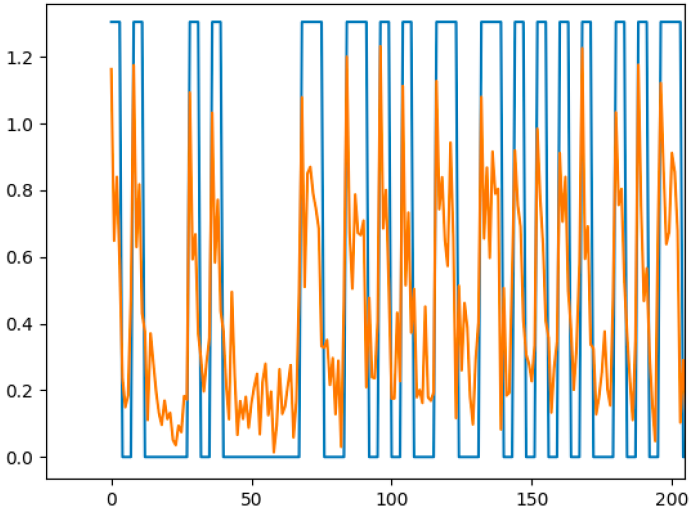
- Differentiating emitter sources from their raw waveforms
- Possible due to detectable differences the transmitter characteristics of individual radios
- Analogous to commonly used biometric fingerprinting and retinal scans for people

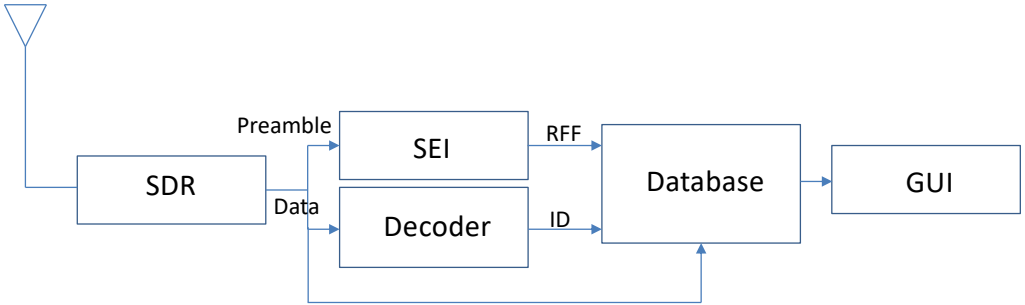


3. Authentication - CruxSEI

Defence Innovation Hub

- Integrates software defined radio, SEI and conventional decoder on single FPGA
- Uses low precision deep neural network to achieve > 80% accuracy, goal ~ 1 ms latency





Overview

FPGAs

Challenges

Emerging Technologies

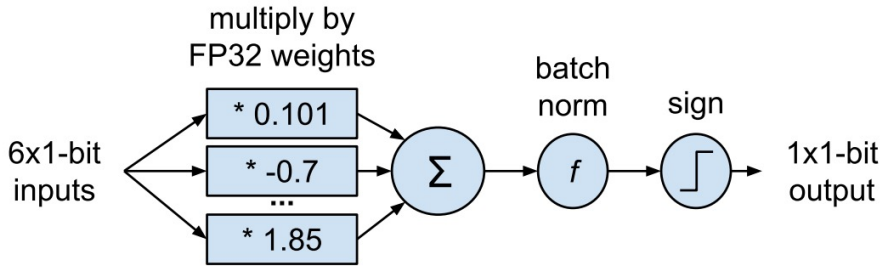
Summary

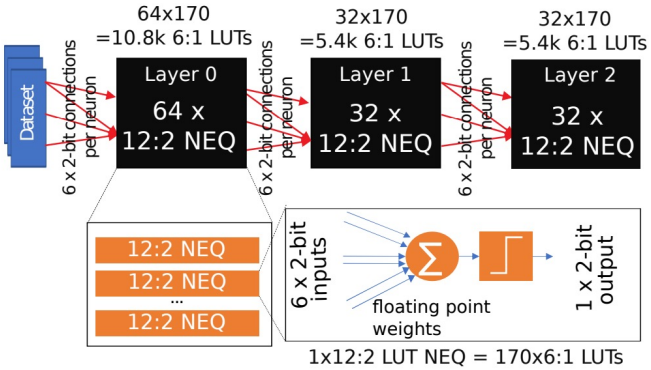
Inference: Sub-100ns Inference Engines (FPL'20)

Xilinx Research Labs LogicNets (https://arxiv.org/pdf/2004.03021.pdf)

Low-precision neural network matched to FPGA architecture

Applications with 15 ns of latency reported



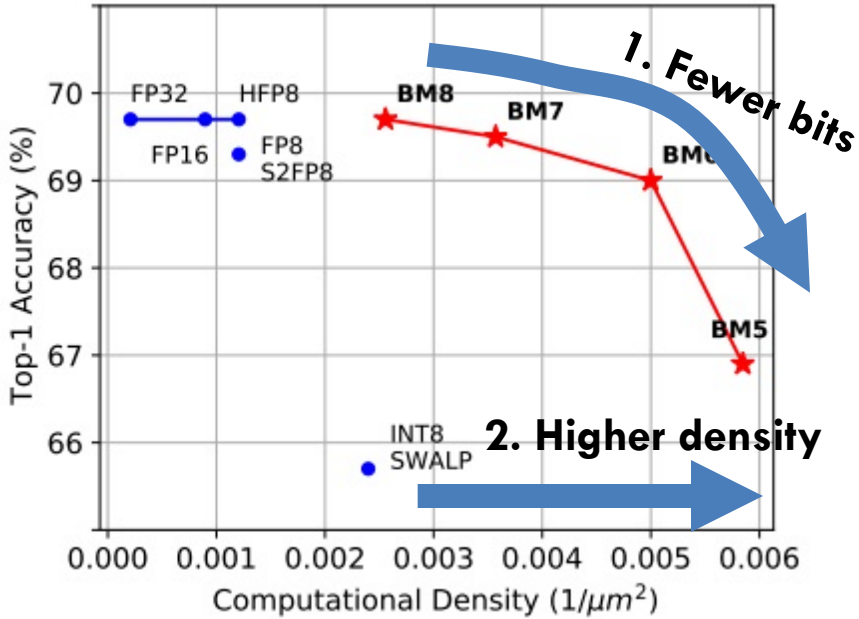


Source: Xilinx

Can this be used in ultra high-speed communications applications?

Training: Low-Precision Training: Block Minifloats (ICLR'21)

Block minifloats which can train ImageNet with 8-bit precision



Model: ResNet-18 Dataset: ImageNet

http://phwl.org/assets/papers/bm_iclr21.pdf

- This work may be particularly advantageous in moving training into **Edge devices**
- github.com/sfox14/block_minifloat
- Can we use in online applications which adapt to changing conditions?

Data: Data-centric AI - A Shift to Good Data

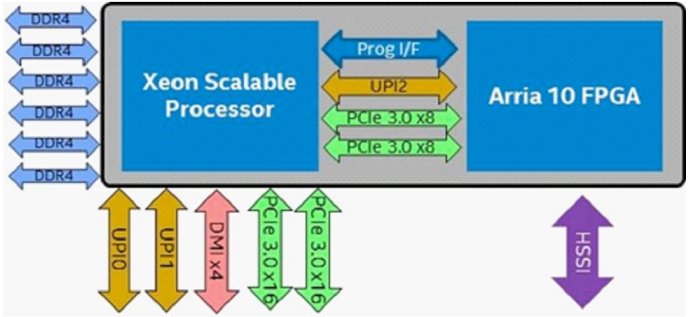
"In many industries where giant data sets simply don't exist, I think the focus has to shift from big data to good data. Having 50 thoughtfully engineered examples can be sufficient to explain to the neural network what you want it to learn."
—Andrew Ng, CEO & Founder, Landing AI

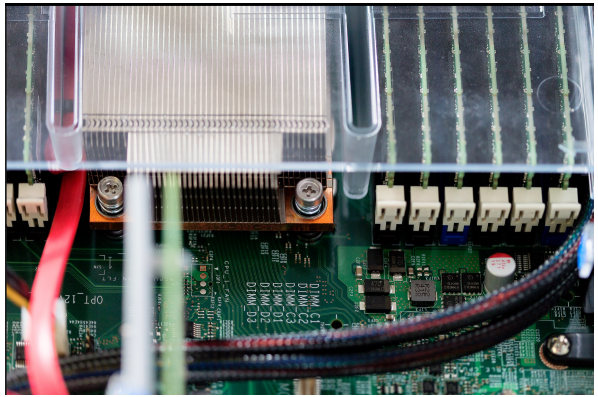
- Data-centric AI is the discipline of systematically engineering the data needed to successfully build an AI system
  - dominant paradigm was to focus on improving the code (NN architecture)
  - for many applications the neural network architecture is a solved problem
  - in many applications more productive to fix NN architecture and improve data
- See also NVIDIA Omniverse synthetic data
- Do we really need massive amounts of data?

https://spectrum.ieee.org/andrew-ng-data-centric-ai
https://developer.nvidia.com/nvidia-omniverse-platform

Heterogeneous (CPU+FPGA+GPU) Processors

- Common bottleneck is CPU/FPGA interface
- Intel Xeon CPU-Arria 10 FPGA (15 core Xeon in 2016) about 50x speedup over floating point for GEMM http://phwl.org/assets/papers/cmm_fpga18.pdf
- AMD's acquisition of Xilinx likely to result in tightly coupled CPU+FPGA+GPU devices
  - New era in heterogeneous edge computing





Overview

FPGAs

Challenges

Emerging Technologies

Summary

Summary

- FPGAs have **Exploration**, **Parallelism**, **Interface** and **Customisation** (EPIC) benefits
- Highlighted examples in **spectrum management**, **RF scene understanding** and **authentication**
- Emerging technologies in low-latency **inference**, online **training**, dealing with **limited data**, and **heterogeneous** devices

1. **Dramatic advances in both ML and FPGA technology continue to be made.**
2. **Together they enable RFML which will improve every aspect of communications systems.**

Thank you!

www.cruxml.com