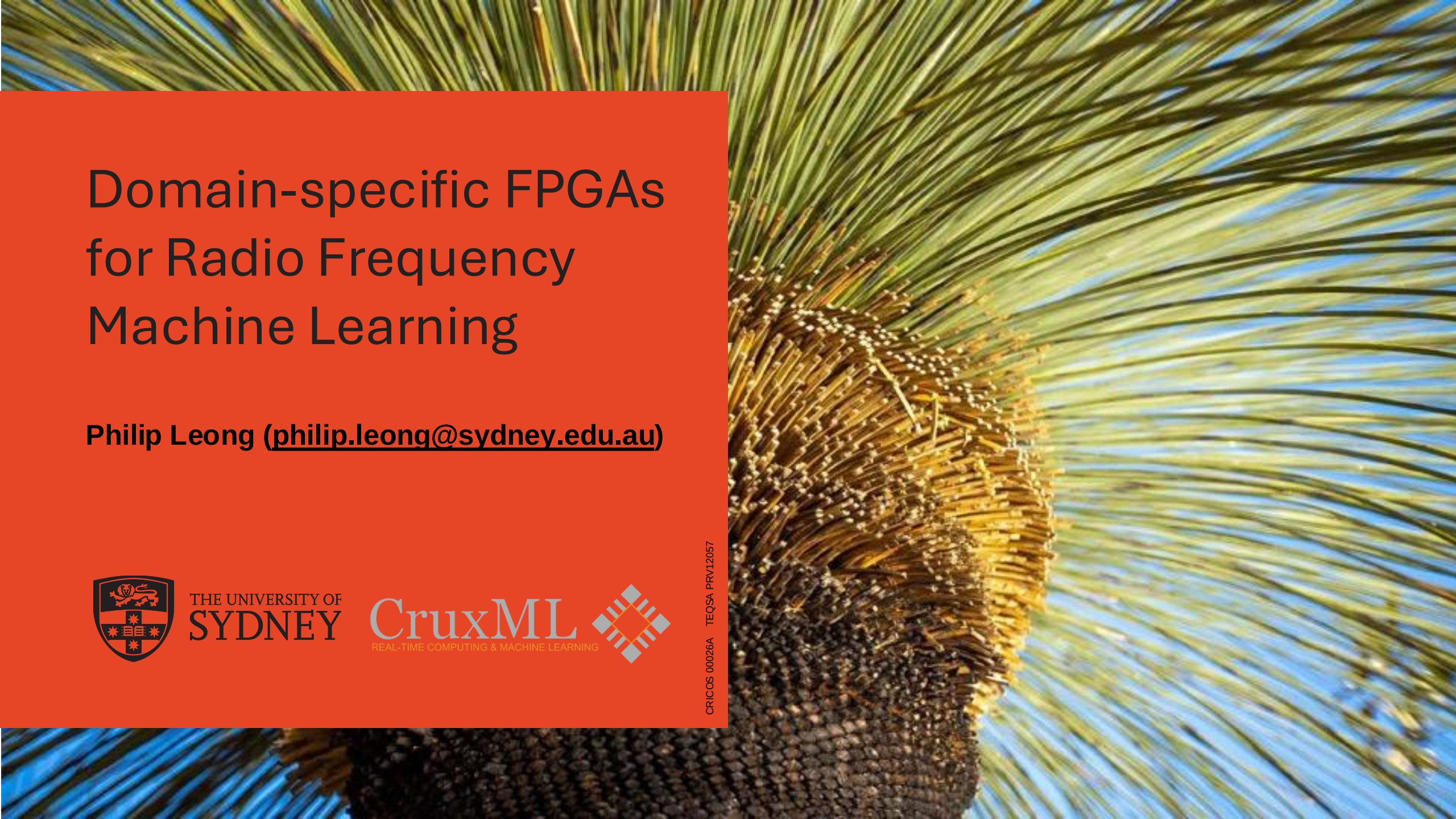
Domain-specific FPGAs for Radio Frequency Machine Learning

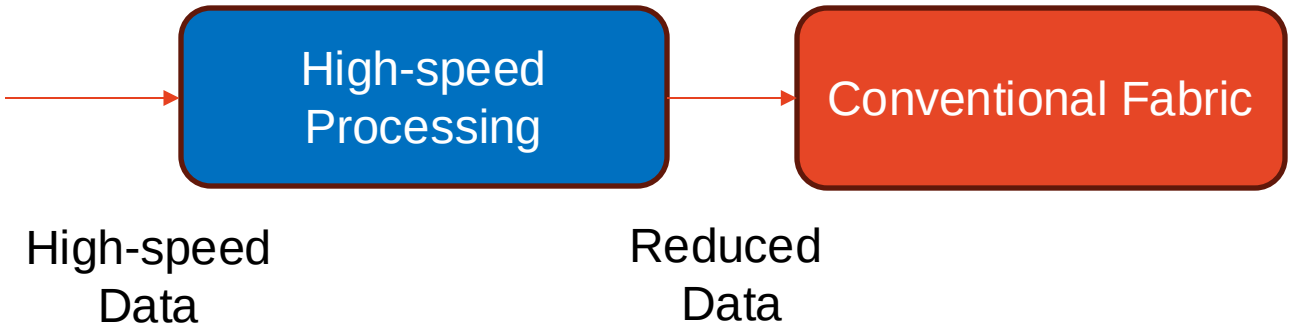
Philip Leong (philip.leong@sydney.edu.au)

THE UNIVERSITY OF SYDNEY

CruxML
REAL-TIME COMPUTING & MACHINE LEARNING



Domain

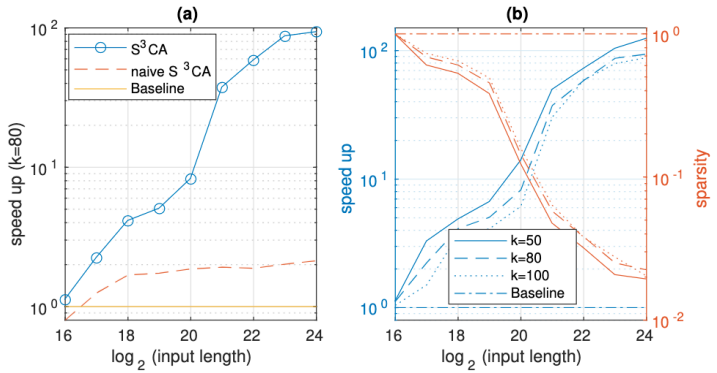


Source: https://sballiance.net.au

- Can we improve architectures for radio frequency machine learning?
- Applications include 6G, radio astronomy, cognitive radio

Spectral Correlation Function (SCF)

- Used as features for DNN
- SCF is correlation of downconverted narrowband spectral components at $f \pm A/2$
  - $A \neq 0$ are *cycle frequencies*



S3CA (fast algorithm)

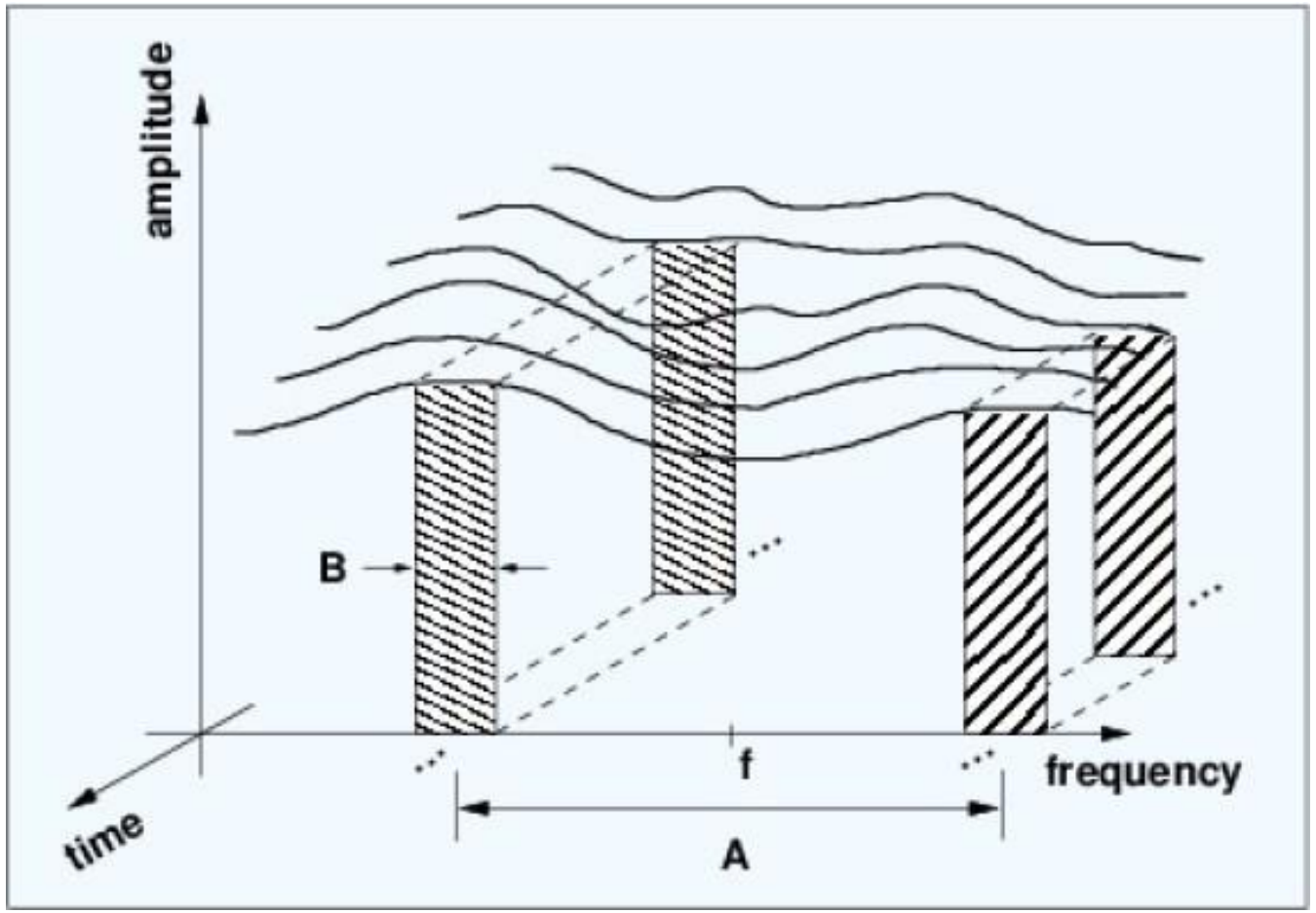
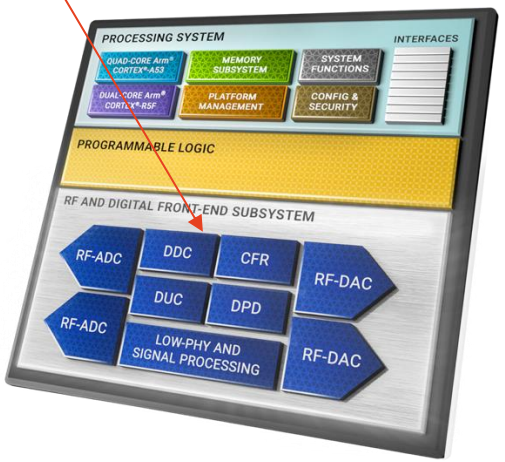


Figure: Chad Spooner cyclostationary.blog

FPGA Features

- FPGAs equipped with high-speed converters
  - Enabling technology for
    - Quantum sensing & computing
    - Test and measurement
    - Radio applications
  - ML used to enhance all applications

Channel filter, DUC/DDC, Mixer, Crest Factor Reduction, DPD



Source: AMD

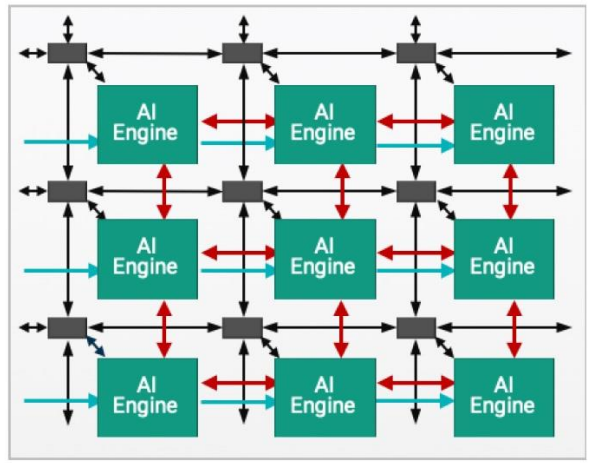
High Frequency Machine Learning Bottlenecks

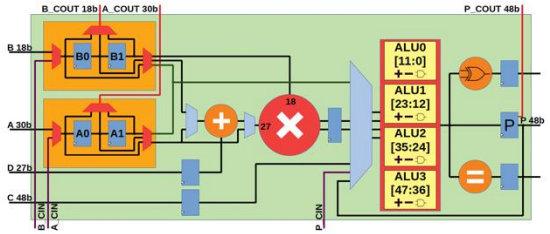
1. Front-end Data rates
2. Power consumption

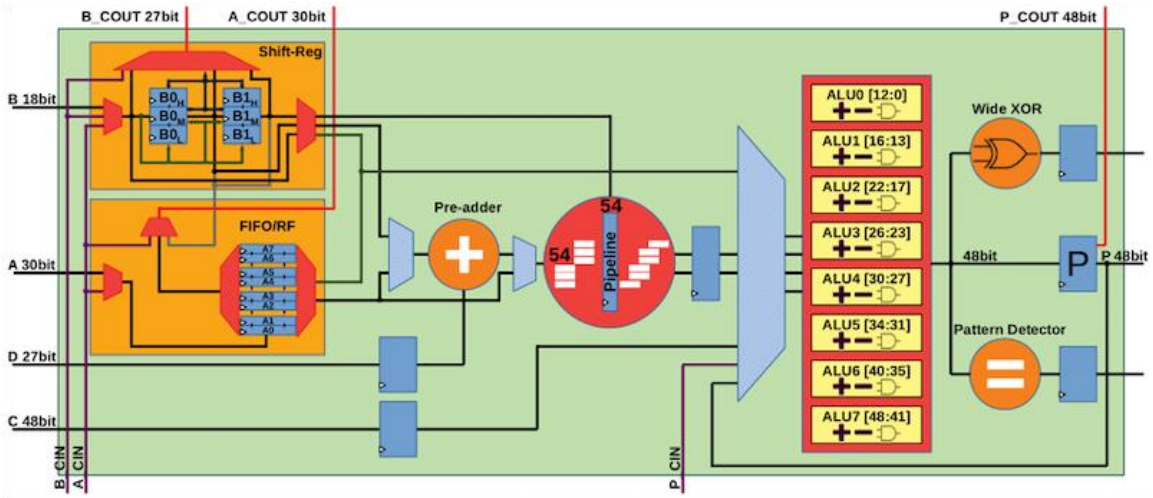
1. Data Rates - Compute hard block

- Data rates high (Gbps)
- Data is I/Q streams
- Clock frequency low
- Coarse-grained block
  - Higher level
  - Local memory
  - Streaming
  - Complex arithmetic
  - Word-based fabric
  - High frequency

Source: AMD

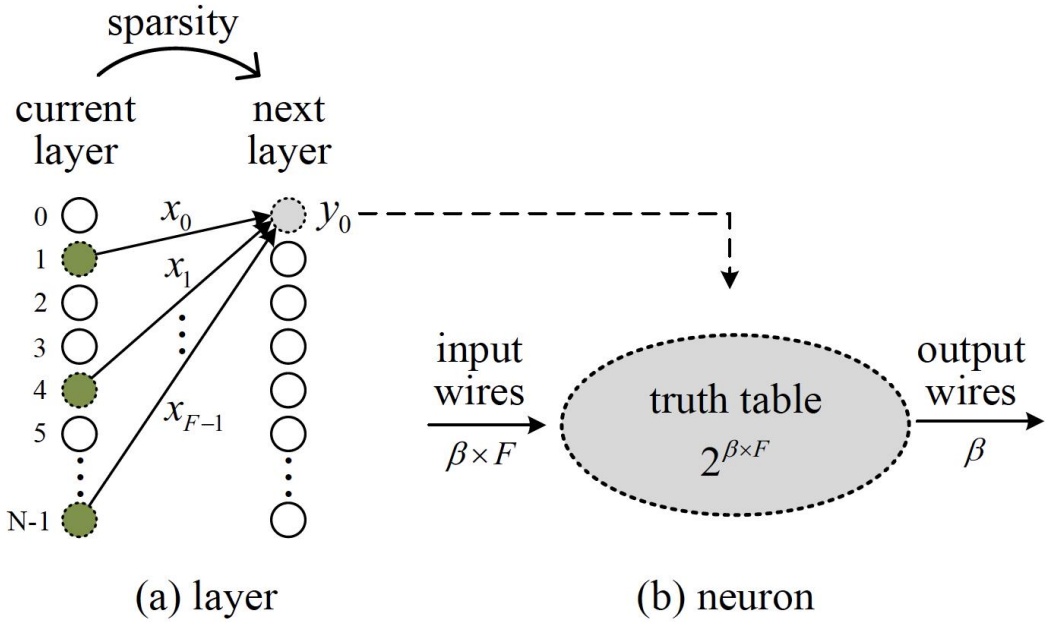




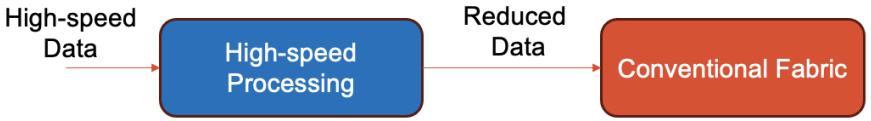


1. Data Rates - Deep Neural Networks

- LUTnets/LogicNets can achieve high data rates & low-latency
- Accuracy compromised by sparsity
- Only for small networks
- Compute hard block with large fan-in
- Word-based fabric
- Higher level of abstraction
  - Programmable machine



2. Power Consumption



- Customisation options via chiplets
- Integration reduces data movement
- “High-speed Processing” built from custom IP + compute hard block

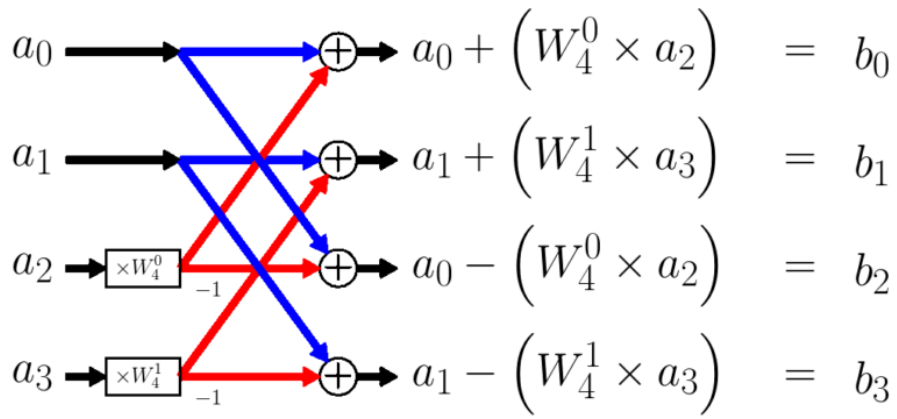
ADC/DAC Chiplet 64 Gsps

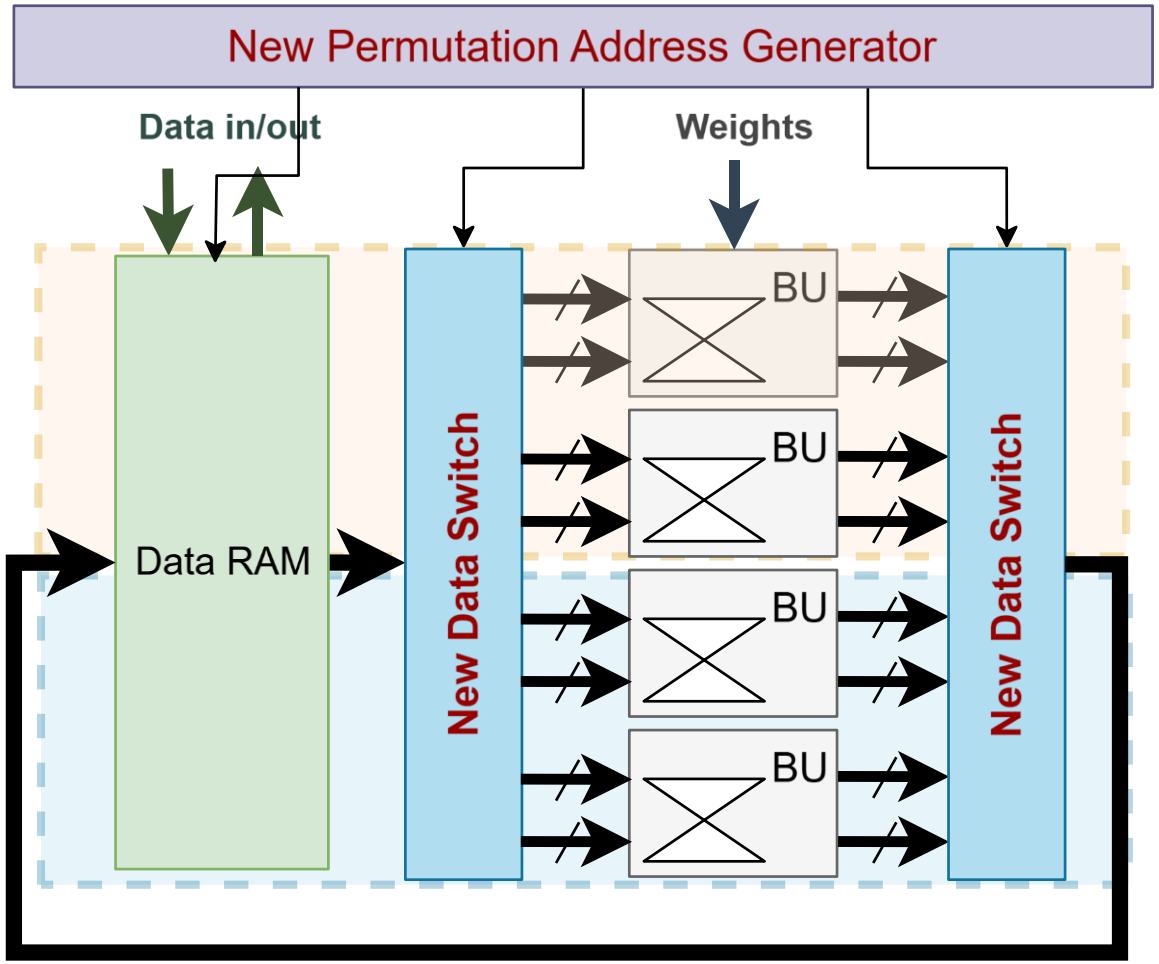
Source: Intel

	Full SCD[1]			Full SCD[2]
	GPU [2]	GPU [2]	FPGA+GPU [2]	Optimized
Platform	Tegra K1	Tesla K20	ZedBoard+Tegra K1	ZCU111
Initiation Interval (ms)	111.61	8.98	50.95	**0.164**
Throughput (MS/s)	0.018	0.228	0.040	**12.5**
Speedup	1	12.3	2.1	**677.6**
Computational Performance (GOPS)	0.14	1.75	0.30	**460**
Power (W)	3.5	51	5	**35(6.10)**[5]
Energy Efficiency (MOPS/W)	40	34	60	**13,143**
Signal-to-quantization noise ratio (dB)	-	-	-	**73**[6]

A DSP/DNN Hard Block

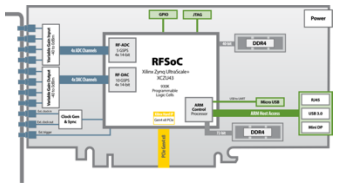
- Integrates large FFTs with neural network inference for DSP + DNN
- Software programmable
- Streaming interface



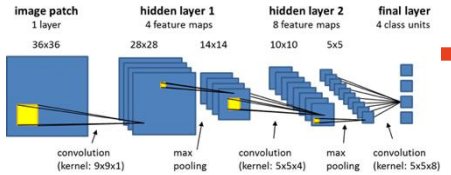


Example System for RF Scene Understanding

Combine real-time FPGA HW, cyclostationary analysis, ML and AR

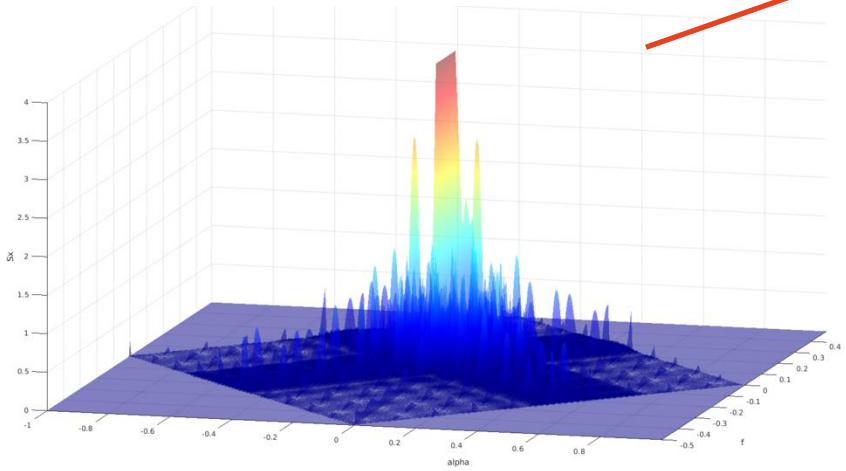


Emitter Localization

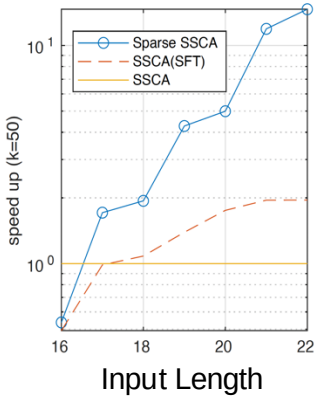


ML on FPGA

AR Display



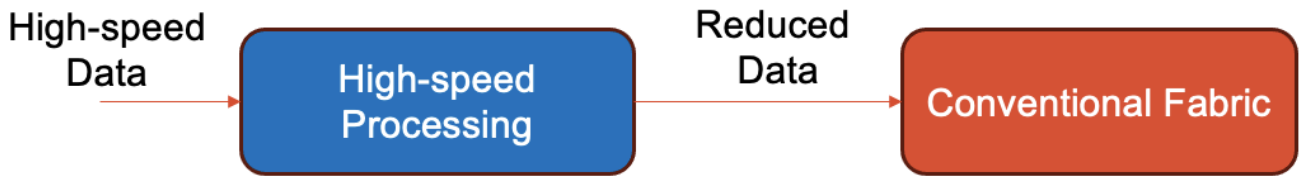
Cyclostationary Signal Processing (CSP)



10x Acceleration of CSP in SW

RF Fingerprinting on FPGA

Summary



- Always need more speed
- Domain-specific Architectures can improve high-speed processing
  - New hard blocks that move DSP to higher abstraction (programmable, local mem, arithmetic, word-based)
  - Streaming interfaces
  - Heterogeneous programming tools
- Questions?
  - Can blocks be generalized from RF to any application?
  - Should it be a hard block or chiplet?

Thank you!

THE UNIVERSITY OF SYDNEY

The University of Sydney

philip.leong@sydney.edu.au