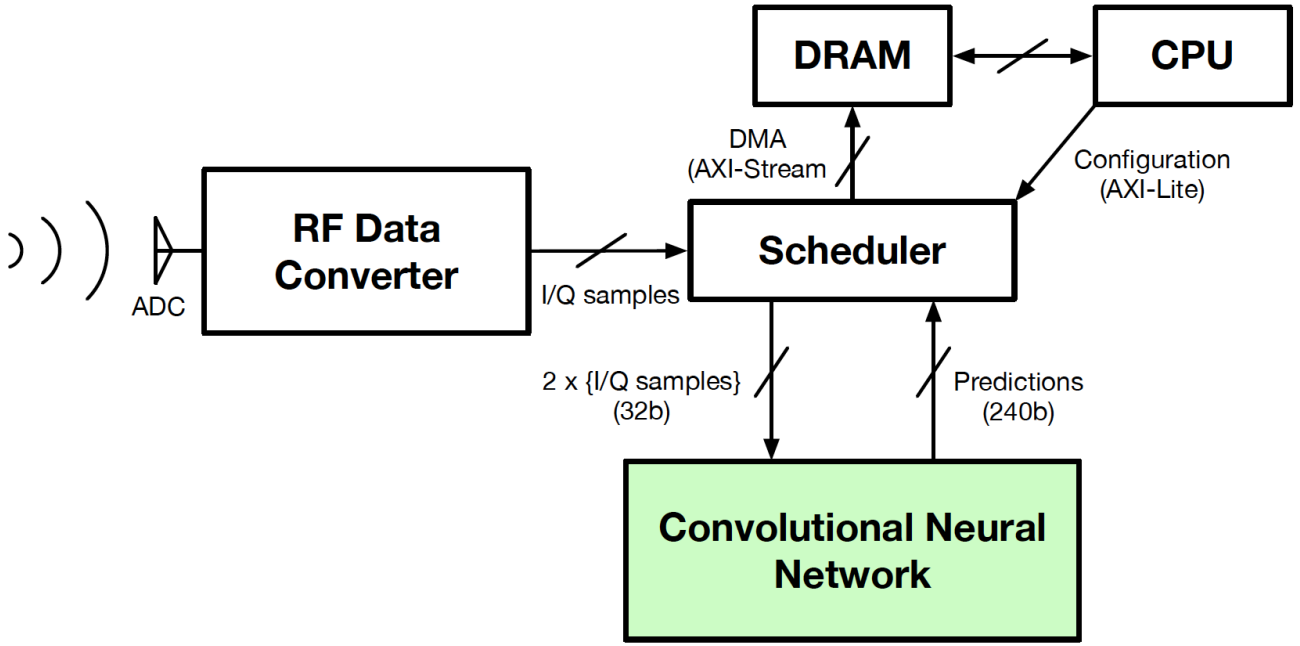
Real-time Automatic Modulation Classification using RFSoC

Stephen Tridgell (Waymo*) - stridge@waymo.com
David Boland (University of Sydney) – david.boland@sydney.edu.au
Philip H.W. Leong (University of Sydney) - philip.leong@sydney.edu.au
Ryan Kastner (UCSD) - kastner@ucsd.edu
Alireza Khodamoradi (UCSD) - alirezak@eng.ucsd.edu
Siddhartha (University of Sydney) - siddhartha.siddhartha@sydney.edu.au
*work done while at University of Sydney

THE UNIVERSITY OF SYDNEY

FPGA Implementation

- Automatic Modulation classifier: 488K class/s, 8us latency for CNN



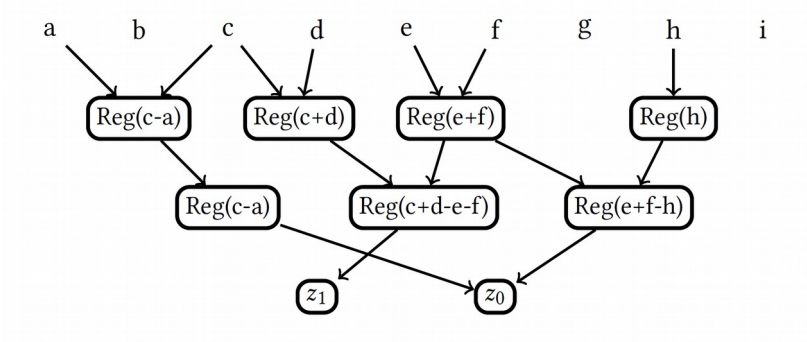
Ternary Weight Networks on an FPGA

- Hard to make fully parallel implementations of a NN on contemporary FPGA due to size
- Fit entire DNN on FPGA by exploiting unstructured sparsity and the following techniques:
  1. Buffering of streaming inputs in a pipelined manner
  2. Ternary weights implemented as pruned adder trees
  3. Common subexpression merging
  4. 16-bit bit serial arithmetic to minimize accuracy loss with low area
  5. Sparsity control
- Apply this to AMC

Common Subexpression Elimination

- Weights are ternary
  - Reduces convolution to constructing adder tree
  - Subexpression merged to reduce implementation

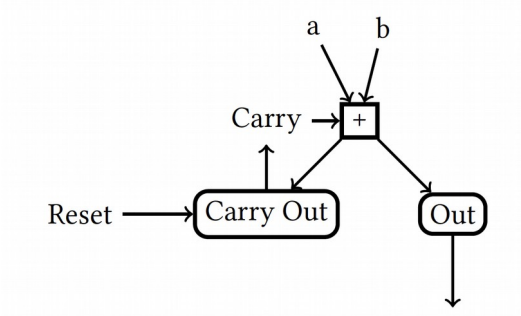
$$\begin{array}{ccc} a \times (-1) & b \times 0 & c \times 1 \\ d \times 0 & e \times 1 & f \times 1 \\ g \times 0 & h \times (-1) & i \times 0 \end{array}$$

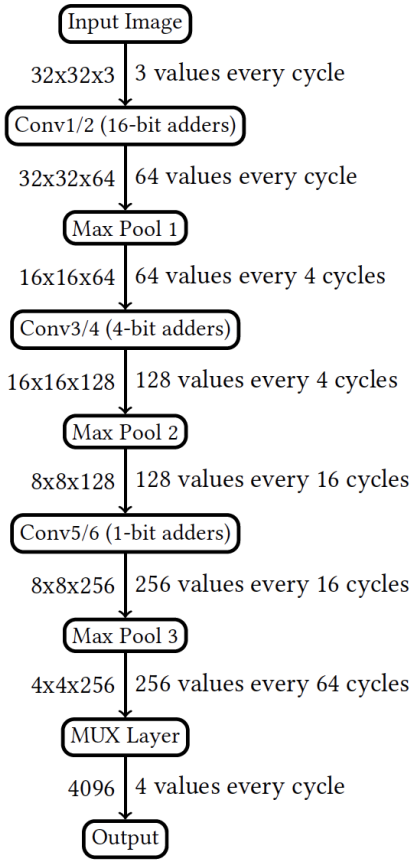


Computing $z_0 = c + e + f - (a + h)$ and $z_1 = c + d - e - f$

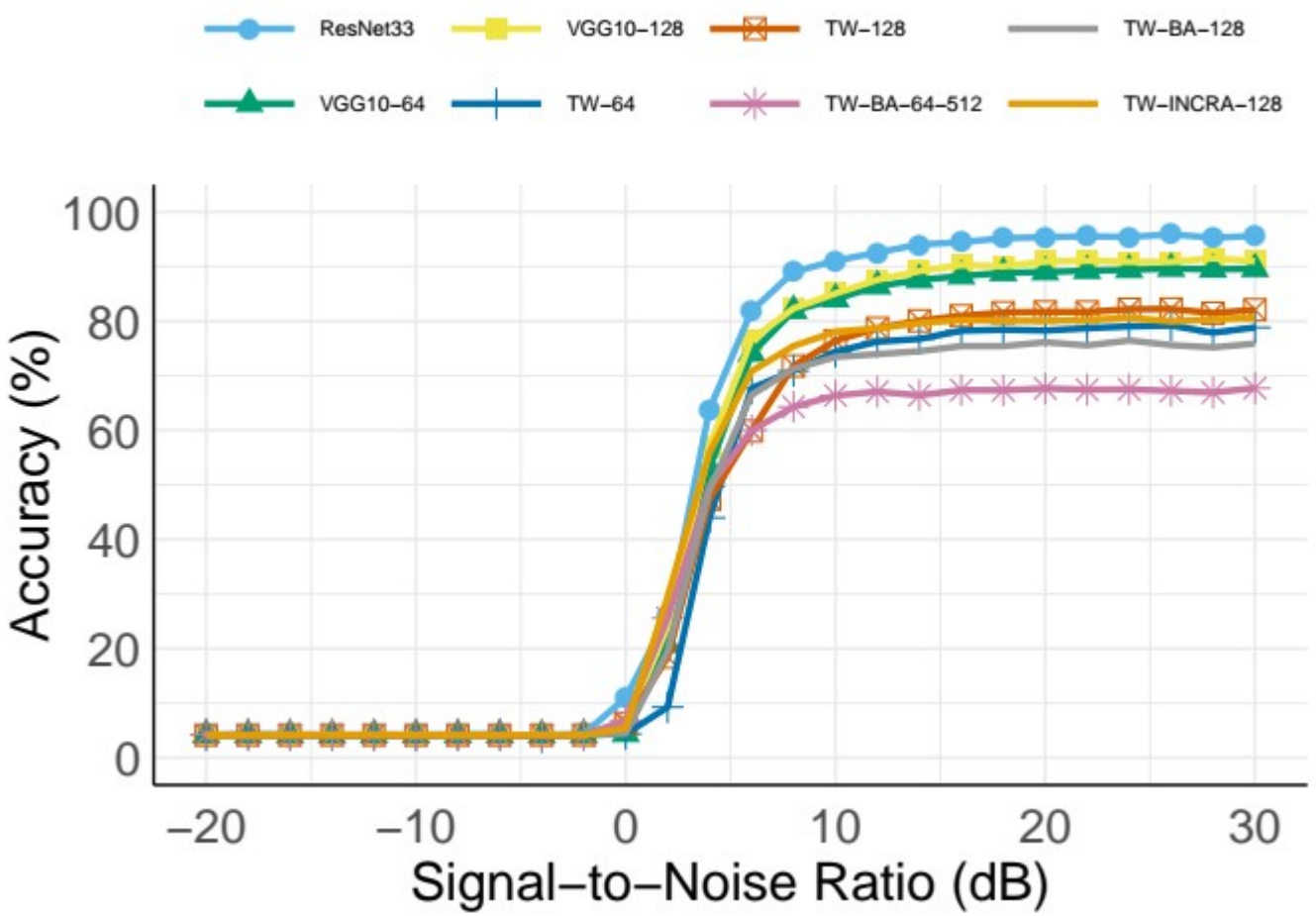
Throughput matching with serial adders

- Activations are 16-bit
  - Not all layers have same throughput
  - Use digit serial to make more compact
  - 4-bit digit serial has 1/4 area
  - 1-bit bit serial has 1/16 area





O’Shea et. al. Deepsig dataset

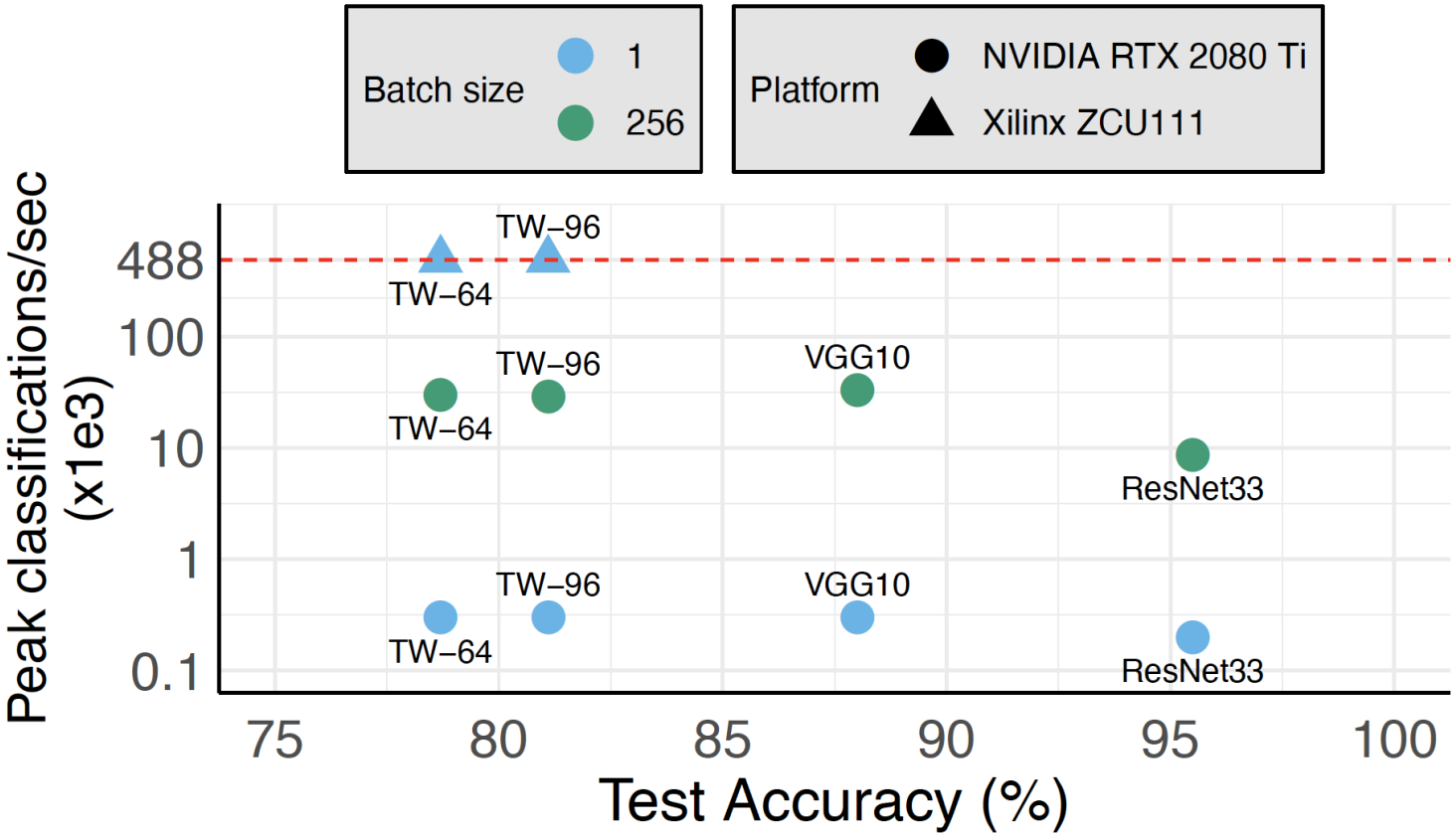


Incremental Precision

- Use incremental precision activations instead of 16 bit
  - Use bit serial adders everywhere
  - Adjust precision to match the throughput
  - Same area as binary activations
  - Almost 5% accuracy gain over binary activations

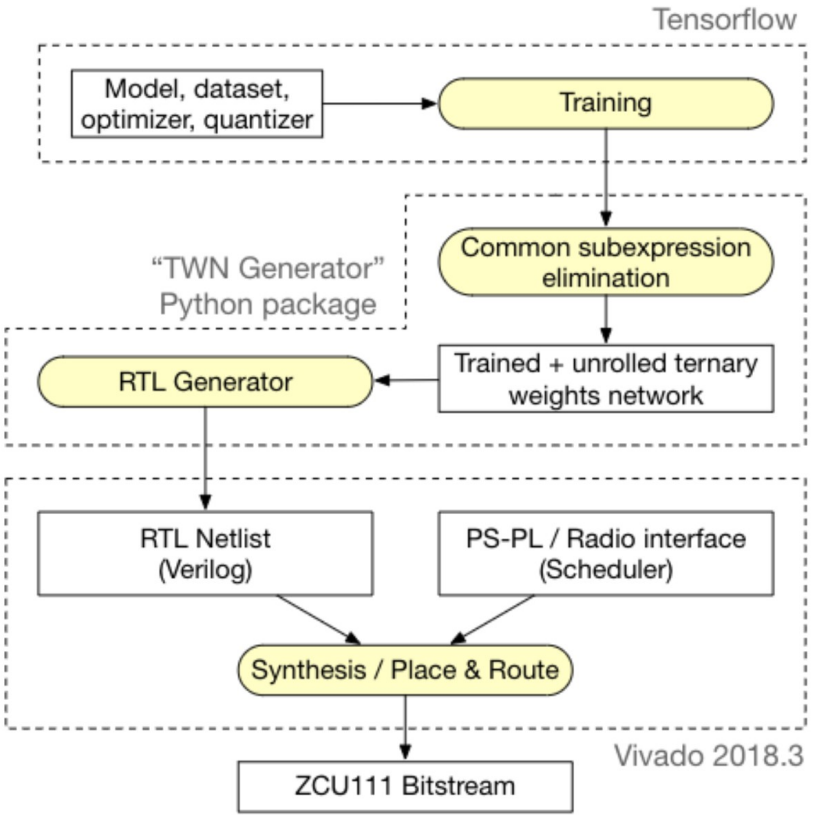
Model	TW-64	TW-96	TW-BA-128	TW-INCRA-128
CLBs	28k (53.5%)	47k (89.3%)	43k (80.7%)	42k (80.2%)
LUTs	124k (29.1%)	232k (54.7%)	234k (55.1%)	211k (49.6%)
FFs	217k (25.5%)	369k (43.4%)	333k (39.2%)	324k (38.1%)
BRAMs	524 (48.5%)	524 (48.5%)	523 (48.4%)	512.2 (48.3%)
DSPs	1496 (35%)	1207 (28.3%)	1408 (33.0%)	1407 (32.9%)
Accr	78.7	81.1	75.9	80.2

Accuracy


Batch size
1
256
Platform
NVIDIA RTX 2080 Ti
Xilinx ZCU111
Peak classifications/sec (x1e3)
488
100
10
1
0.1
TW–96
TW–64
TW–96
TW–64
VGG10
ResNet33
TW–96
TW–64
VGG10
ResNet33
75
80
85
90
95
100
Test Accuracy (%)

Implementation

- System implemented on ZCU111 RFSoC
- https://github.com/da-steve101/radio_modulation
- Open Source Verilog generator
- https://github.com/da-steve101/twn_generator



THE UNIVERSITY OF SYDNEY

Conclusion

- Presented an optimized network for AMC which
  - Applies common subexpression elimination and digit serial arithmetic to a fully unrolled ternary network
  - Integrates the entire design on a single chip for a low-latency batch size 1 implementation
- These serve to achieve a level of performance higher than previously reported
- Challenge of achieving state of the art accuracy remains