Award Ceremony

2021 Challenge

Lightning-Fast Modulation Classification with Hardware-Efficient Neural Networks

© Copyright 2021 Xilinx

Part of

XILINX

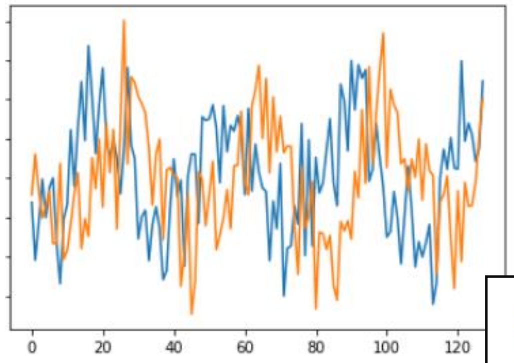
© Copyright 2021 Xilinx

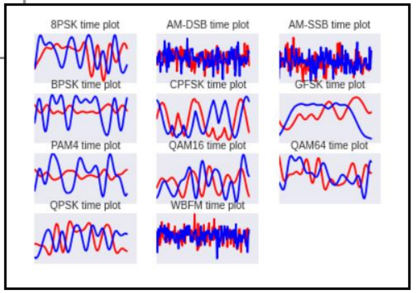
Agenda

- Introduction & Statistics
- Awards ceremony with Xilinx CTO Ivo Bolsens
- Lightning talks from top submissions
- RadioML on FPGAs

© Copyright 2021 Xilinx

Modulation Classification: what's in my RF spectrum?





Which RF modulation is currently in use?

- Rapidly label + understand RF spectrum
- Key enabler for...
  - spectrum interference monitoring
  - radio fault detection
  - dynamic spectrum access
  - numerous regulatory and defense applications
- DNNs promising for modulation classification
  - Especially for short-time observations

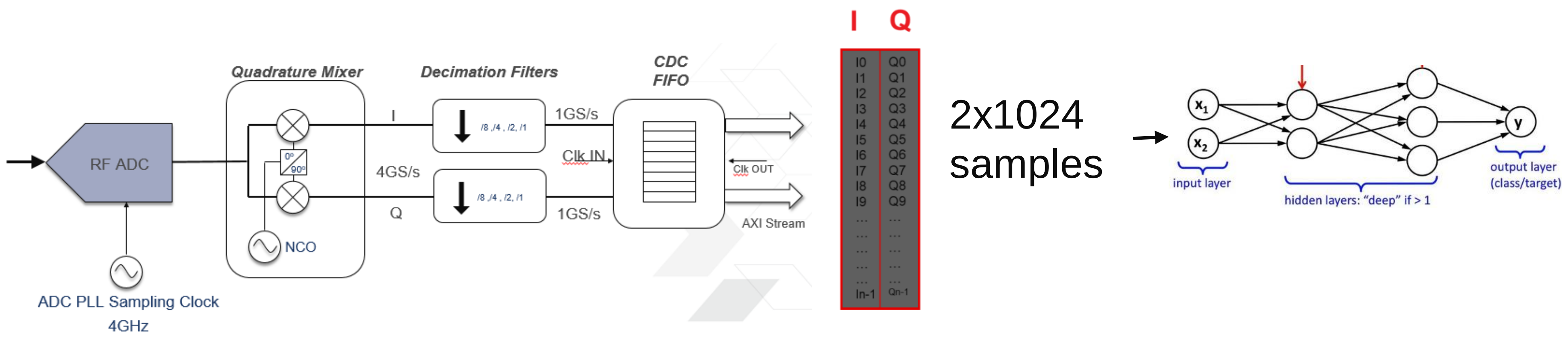
[O'Shea et al., Over the Air Deep Learning Based Radio Signal Classification, IEEE JSTSP'17]

© Copyright 2021 Xilinx

Challenges: Inference Throughput

RF IN

RF System



Sample

Mix and Down Convert

Samples => "Frame"

Mod. Classification

Output: Modulation Type

Date rate in ~Gigasamples/s

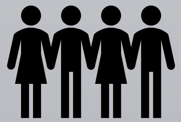
Sample-rate inference throughput is a challenge

© Copyright 2021 Xilinx

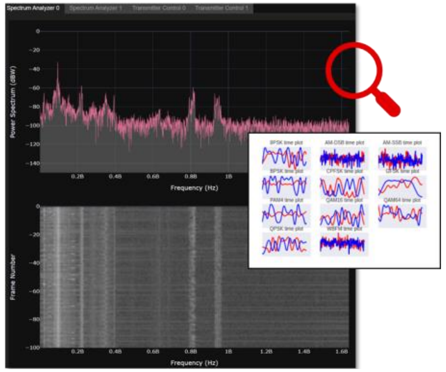
Goal: Enabling many future DNN-based RF applications with extreme throughput and ultra-low latency

© Copyright 2021 Xilinx

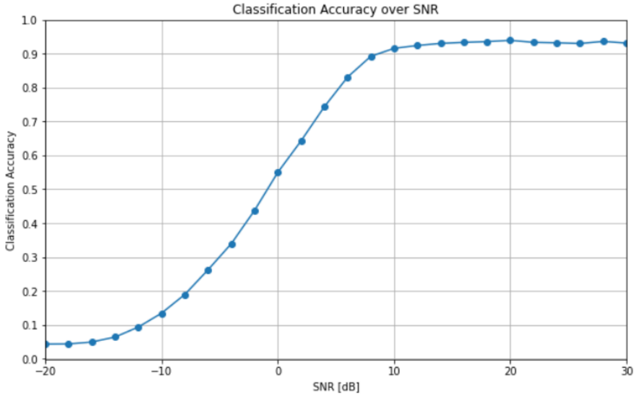
The Challenge



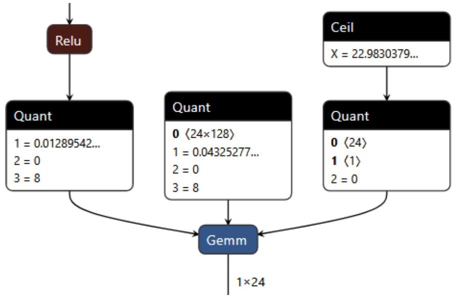
Your team



Train a DNN on **RadioML 2018.01A**

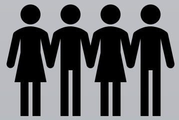


Achieve at least **56.000%** average accuracy over full SNR range



Minimize **inference cost**

XILINX



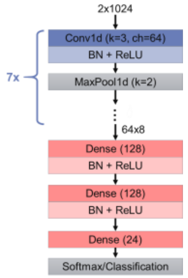
Our team

Brevitas

PyTorch

docker jupyter

Sandbox environment for **quantization-aware training**



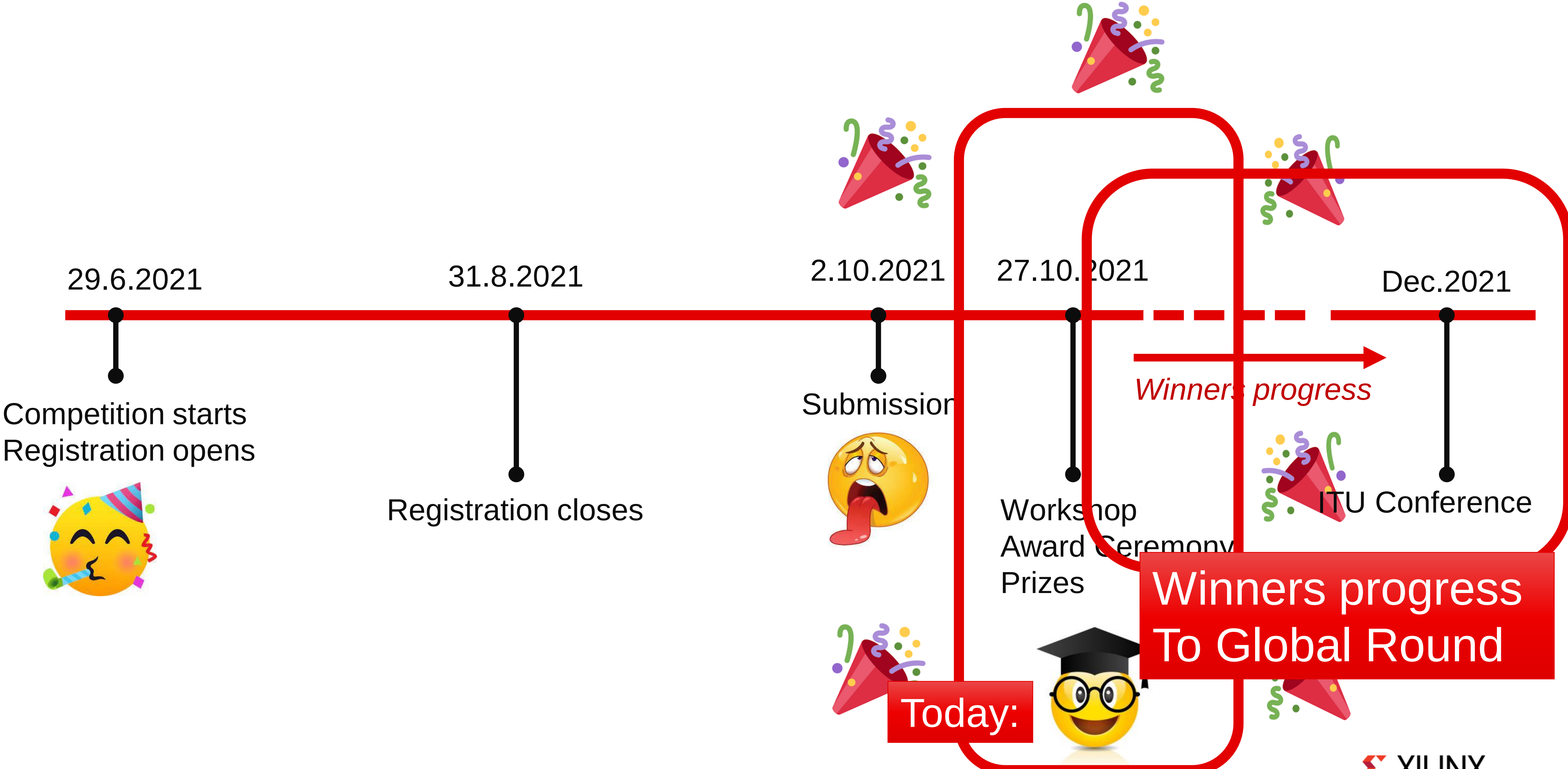
Provide training script for **baseline model**

Support, evaluate and rank the **submissions** on basis of inference cost

© Copyright 2021 Xilinx

XILINX

Timeline



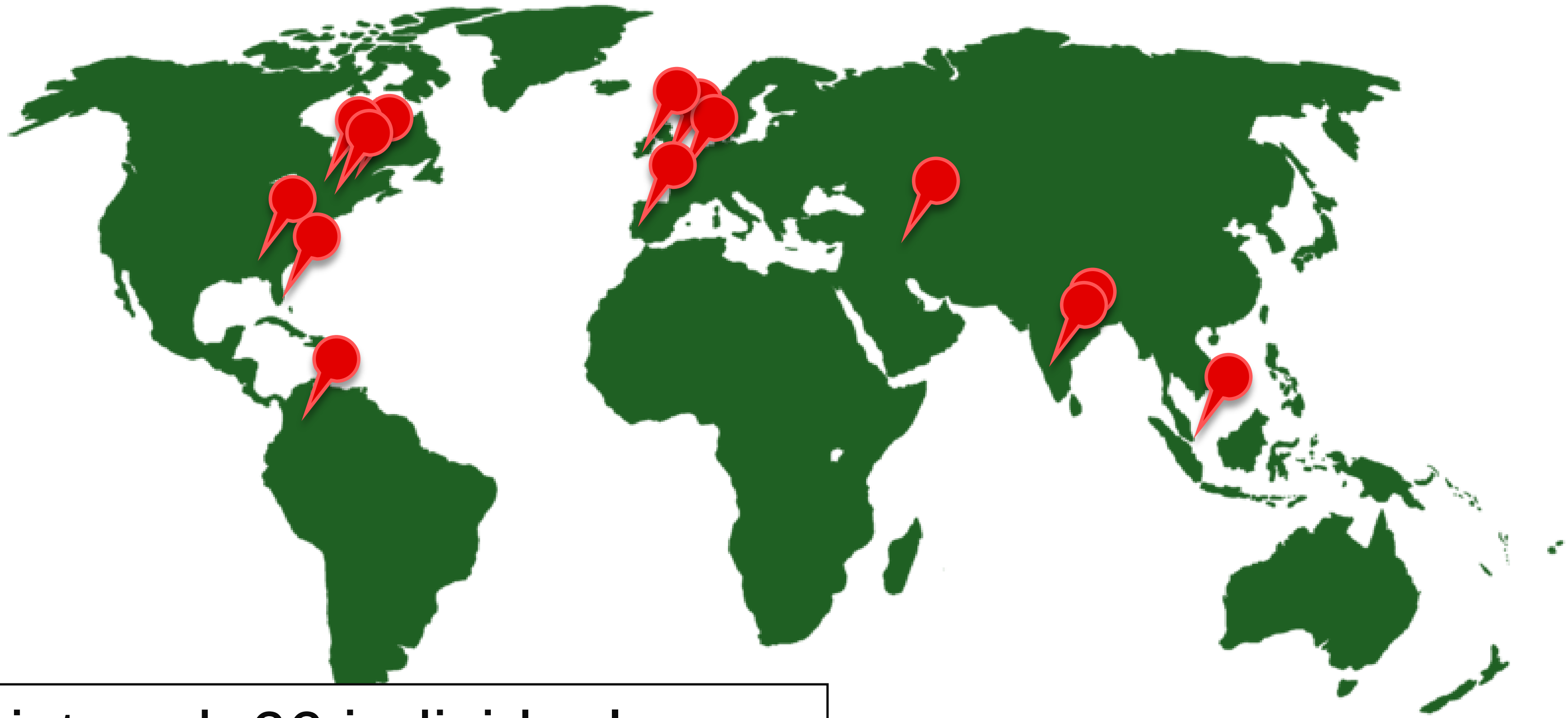
© Copyright 2021 Xilinx

Statistics

© Copyright 2021 Xilinx

Statistics and Geography of Participants

- North America
- South America
- Europe
- Middle East
- South-East Asia



41 teams registered, 66 individuals
13 teams with 32 participants submitted

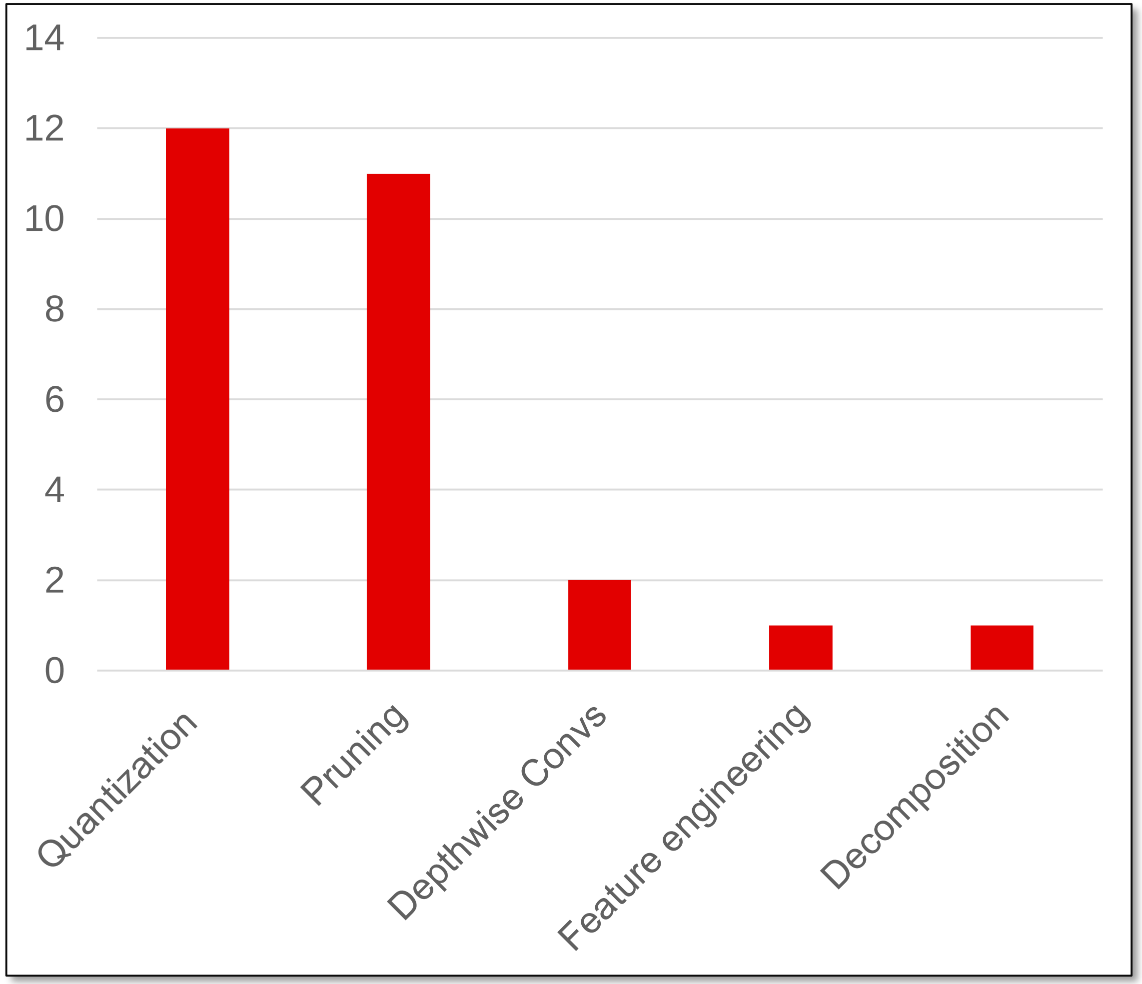
© Copyright 2021 Xilinx

Techniques Adopted

- Most teams adopted variations of the baseline topology with a mixture of quantization and pruning

- Some teams switched to more efficient MobileNet-like topologies

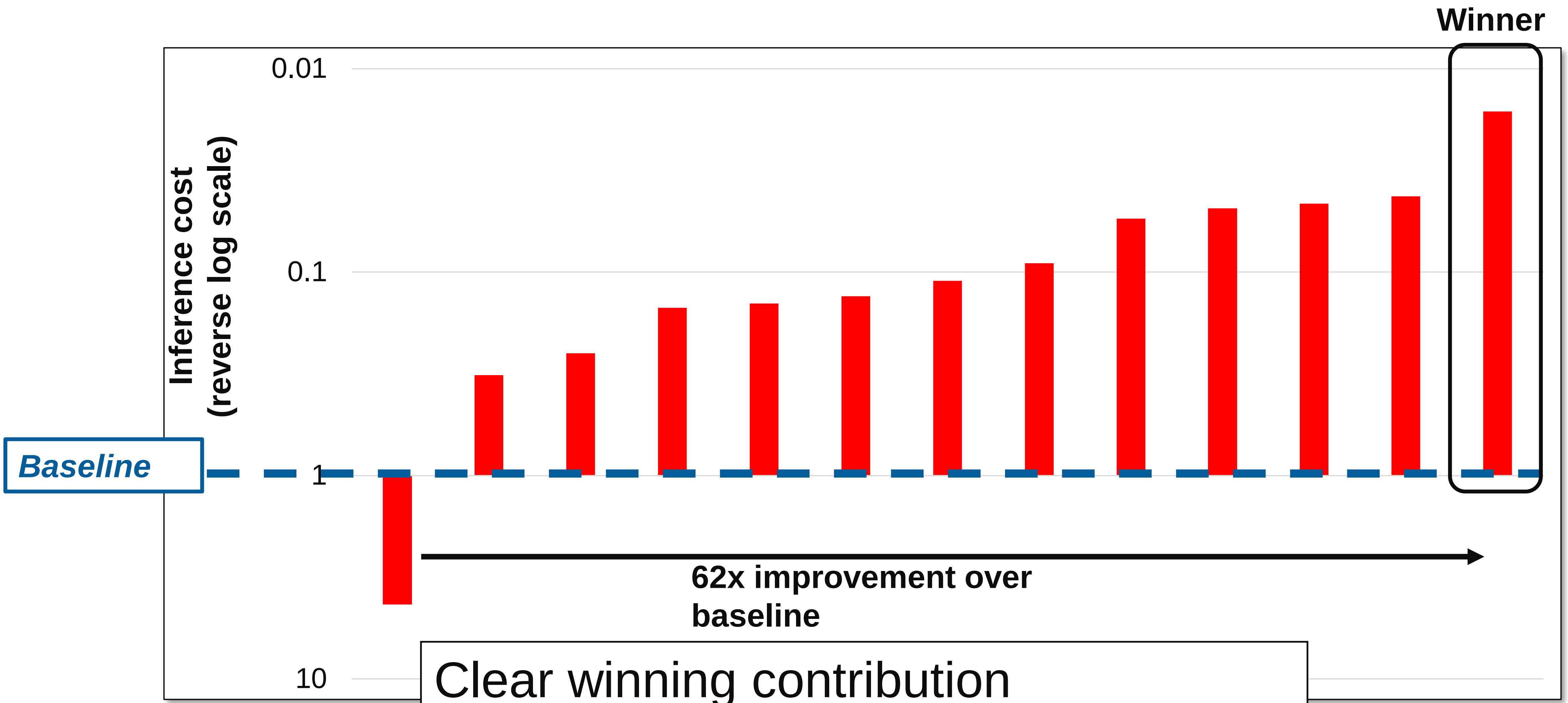
- The winning team is a mixture of the above techniques



XILINX.

© Copyright 2021 Xilinx

Results Across Teams



Clear winning contribution
Close competition for 2nd and 3rd place

© Copyright 2021 Xilinx

Results

© Copyright 2021 Xilinx

Third Place: Aaronica

- **Inference cost score 0.046007**
  - **22x better than baseline!**
- **Team members**
  - **Mohammad Chegini**
  - **Pouya Shiri**

XILINX

© Copyright 2021 Xilinx

Second place: The A(MC) Team

- Inference cost score 0.042467
  - 24x better than baseline!

- Team members:
  - Jakob Krzyston
  - Rajib Bhattacharjea
  - Andrew Stark

© Copyright 2021 Xilinx

First place: BacalhauNET

- Inference cost score **0.016211**
  - **62x better than baseline!**

- Team members:
  - José Rosa
  - Guilherme Carvalho
  - Daniel Granhão
  - Tiago Gonçalves

XILINX

© Copyright 2021 Xilinx

Full Ranking

https://bit.ly/brevitas-radioml-challenge-21-results

Rank	Team Name	Inference cost	Accuracy
1	**BacalhauNET**	**0.016211**	**0.56241**
2	**The A(MC) Team**	**0.042467**	**0.562543**
3	**Aaronica**	**0.046007**	**0.561585**
4	Red Gecko	0.048649	0.5604
5	Wolf	0.054512	0.560464
6	LightNeting (iSmart)	0.090334	0.566972
7	Imperial_IPC	0.1106	0.564
8	ANTENNAE	0.131579	0.566405
9	TCD	0.143013	0.563407
10	TeamX	0.149641	0.563247
11	sing-rb	0.250492	0.563508
12	Team Velocity	0.321127	0.5607
13	FAU-CA-AI	4.307883	0.5719

© Copyright 2021 Xilinx

Team Presentations

© Copyright 2021 Xilinx

Team Presentations

1. Imperial_IPC
2. Aaronica
3. The A(MC) Team
4. BacalhauNet

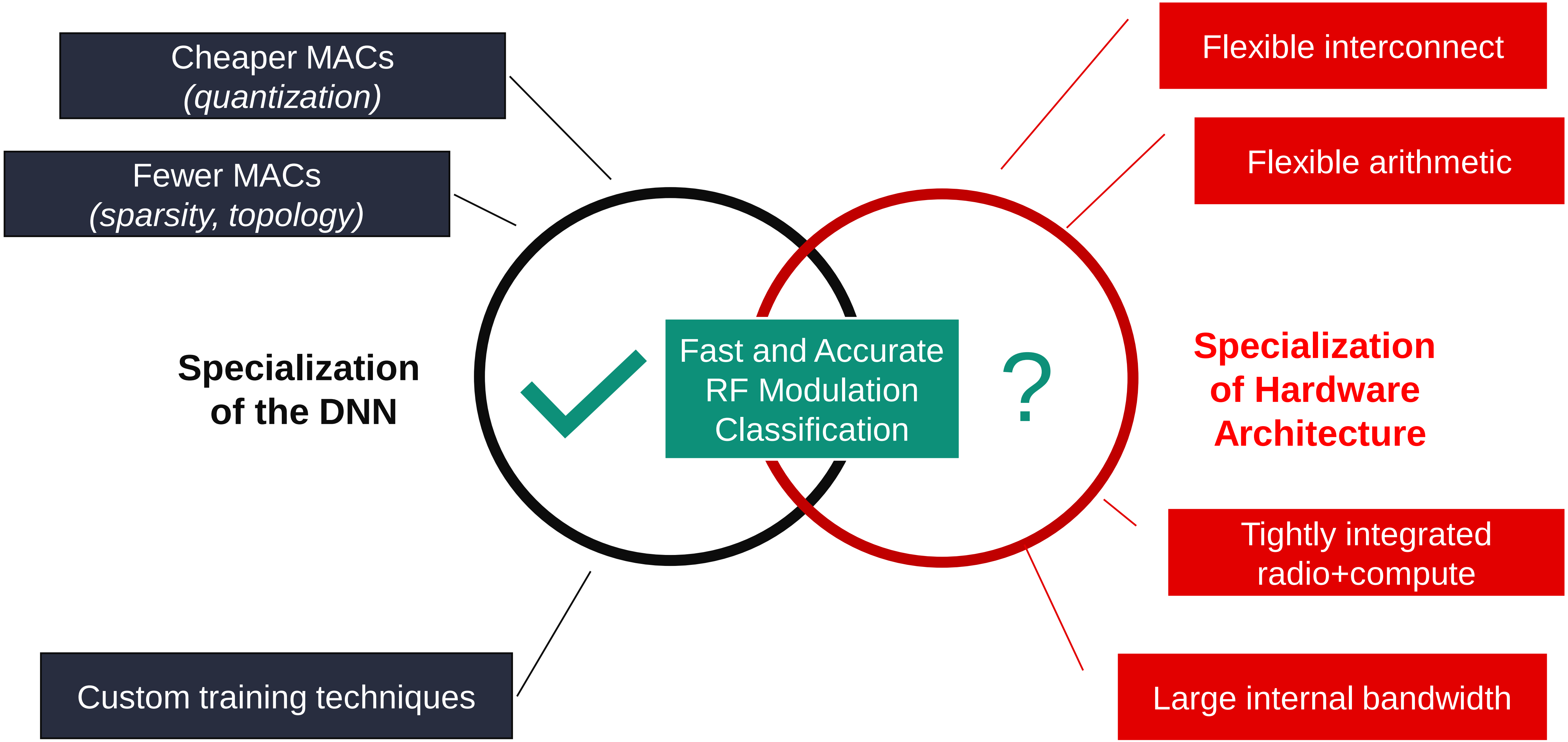
© Copyright 2021 Xilinx

FINN for RadioML on FPGAs

Yaman Umuroglu, Xilinx Research Labs
Felix Jentzsch, Paderborn University

© Copyright 2021 Xilinx

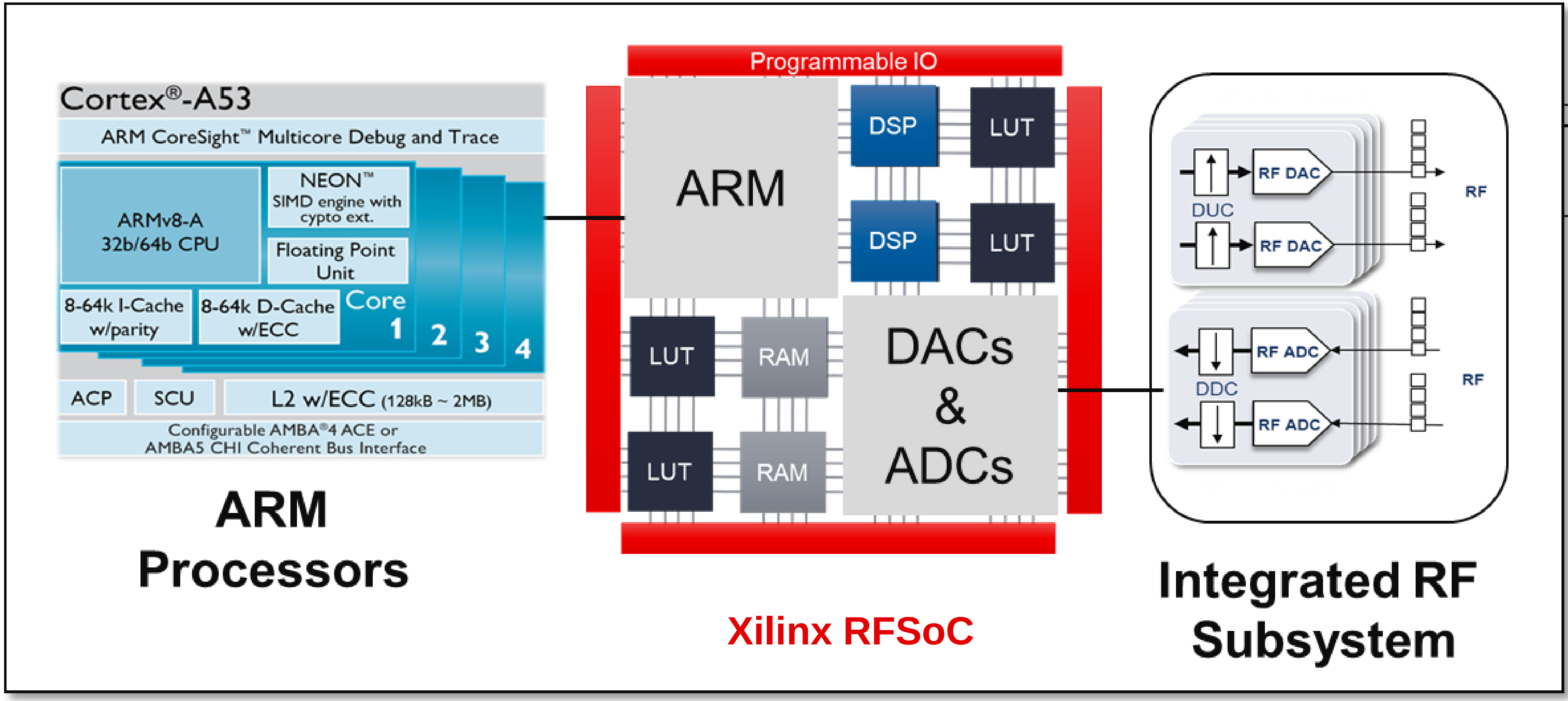
Where do we go from here?



© Copyright 2021 Xilinx

A Refresher on FPGAs and Xilinx RFSoC
Customizable, Programmable Hardware Architectures

Customizes IO interfaces, compute architectures, memory subsystems
to meet the specific application requirements



Tables)

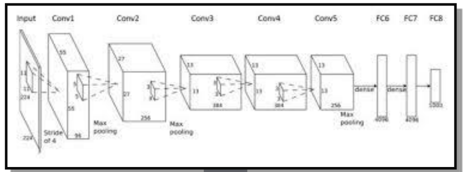
© Copyright 2021 Xilinx

XILINX

Customized DNN to FPGA Solution Stack

FINN

GitHub https://xilinx.github.io/finn/



Brevitas / QKeras*
Training with
algorithmic optimizations

- Train highly efficient, customized DNNs

ONNX Intermediate Representation

?

FINN compiler
Specializations of
hardware architecture

- Hardware optimizations
- Generates dedicated FPGA accelerator

Deployment

- Integrate accelerator into system and deploy

** coming soon*

© Copyright 2021 Xilinx

XILINX

Harnessing FPGA Specialization with FINN

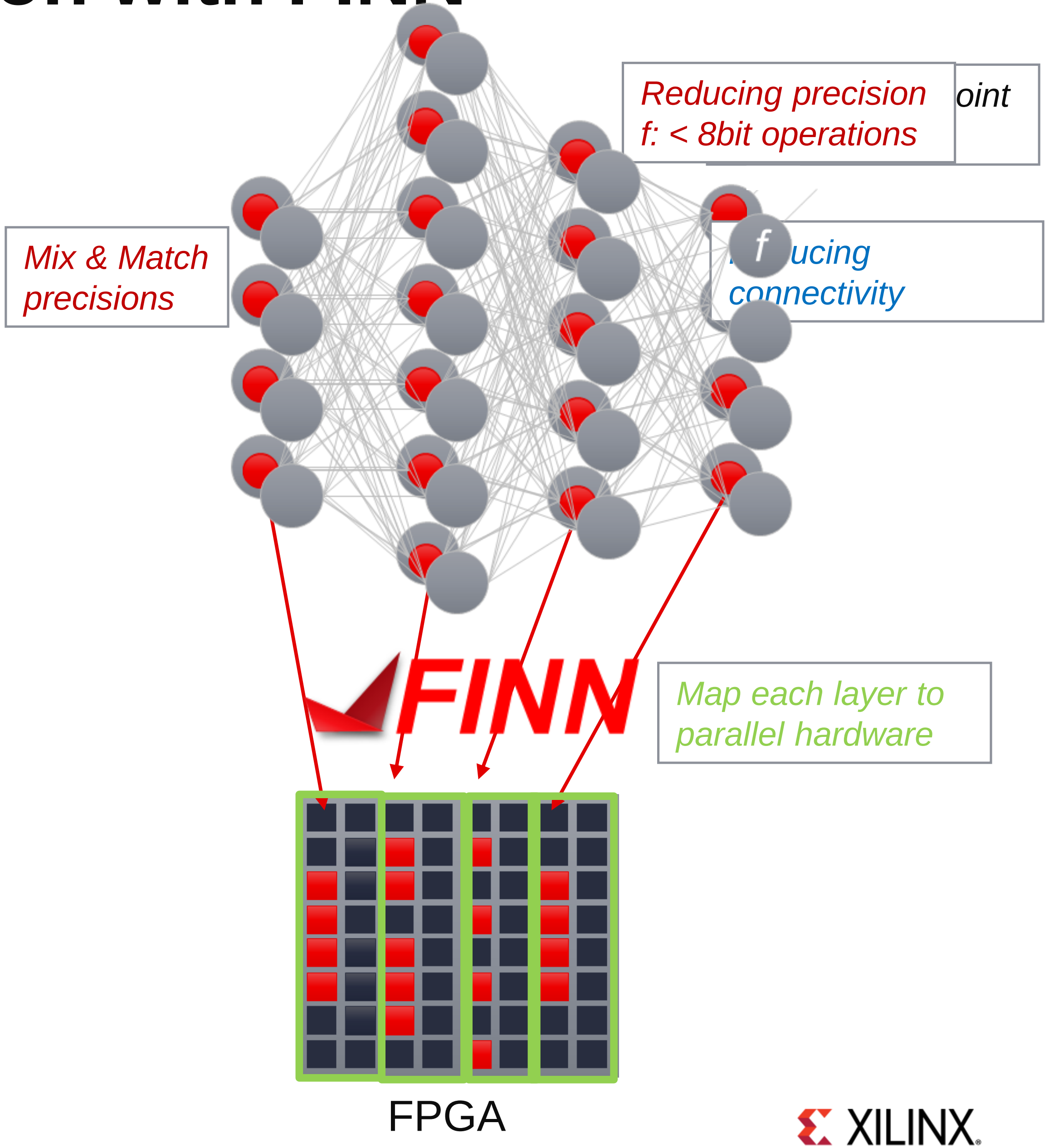
- FPGAs can scale DNN performance through extreme specialization

FINN

- **Reduced precision quantized arithmetic**
  - **Arbitrary bitwidth**
  - **Mix & match bitwidths between layers**

Precision	MAC cost (LUTs)
1-bit	~1.1
2-bit	~4.4
4-bit	~17.6
8-bit	~70.4

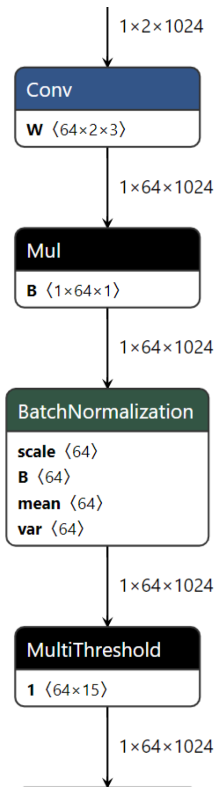
- **Fine-grained sparsity**
- **Layer-parallel dataflow implementation**



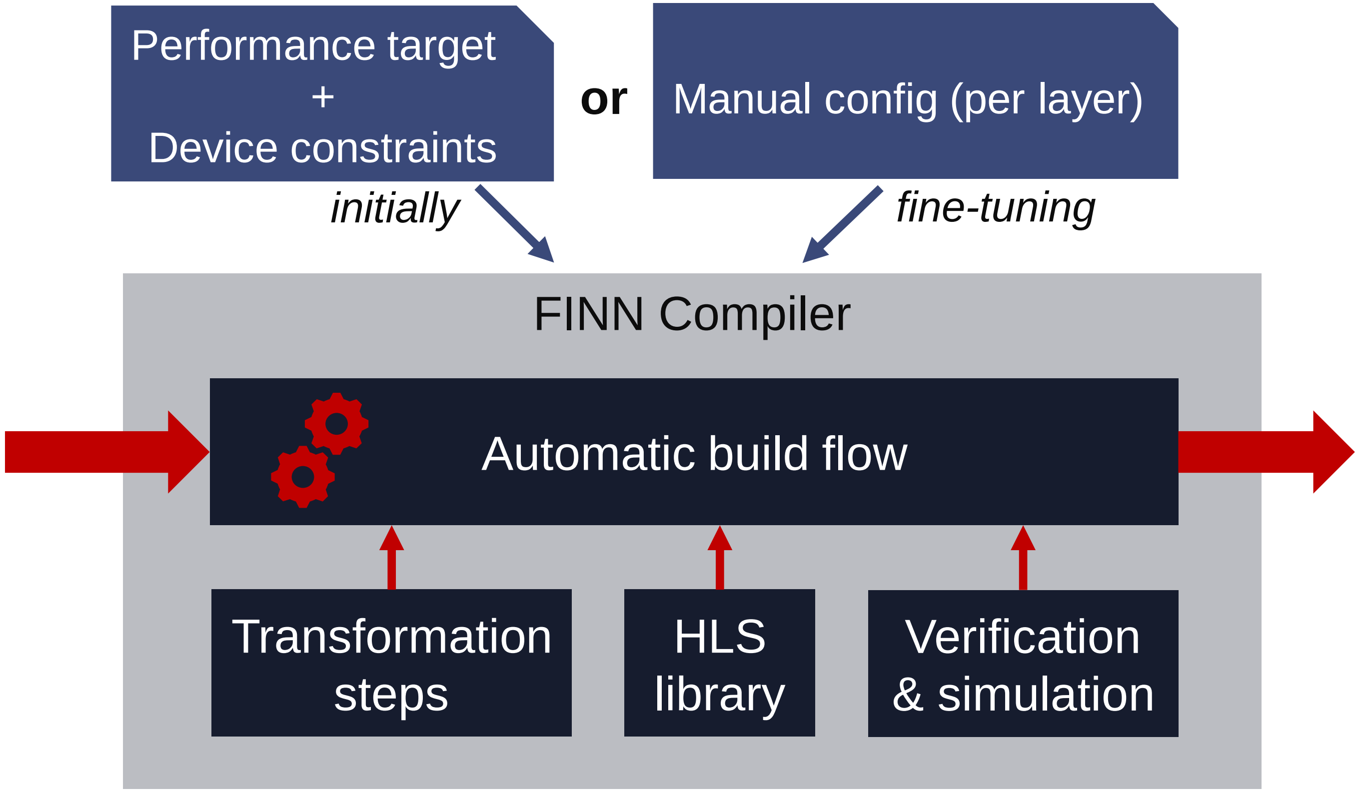
XILINX

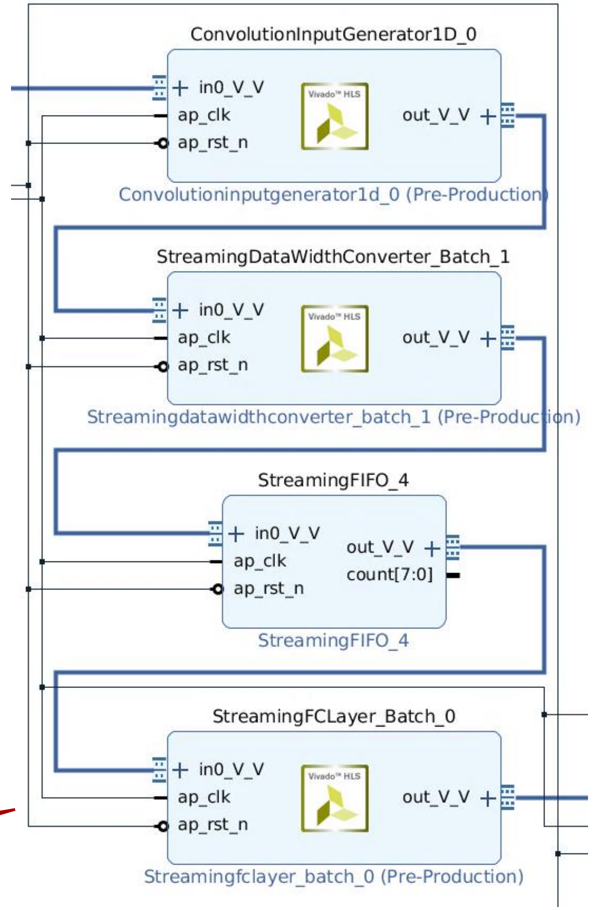
© Copyright 2021 Xilinx

FINN compiler flow: at a glance



Trained ONNX model
(exported from Brevitas)





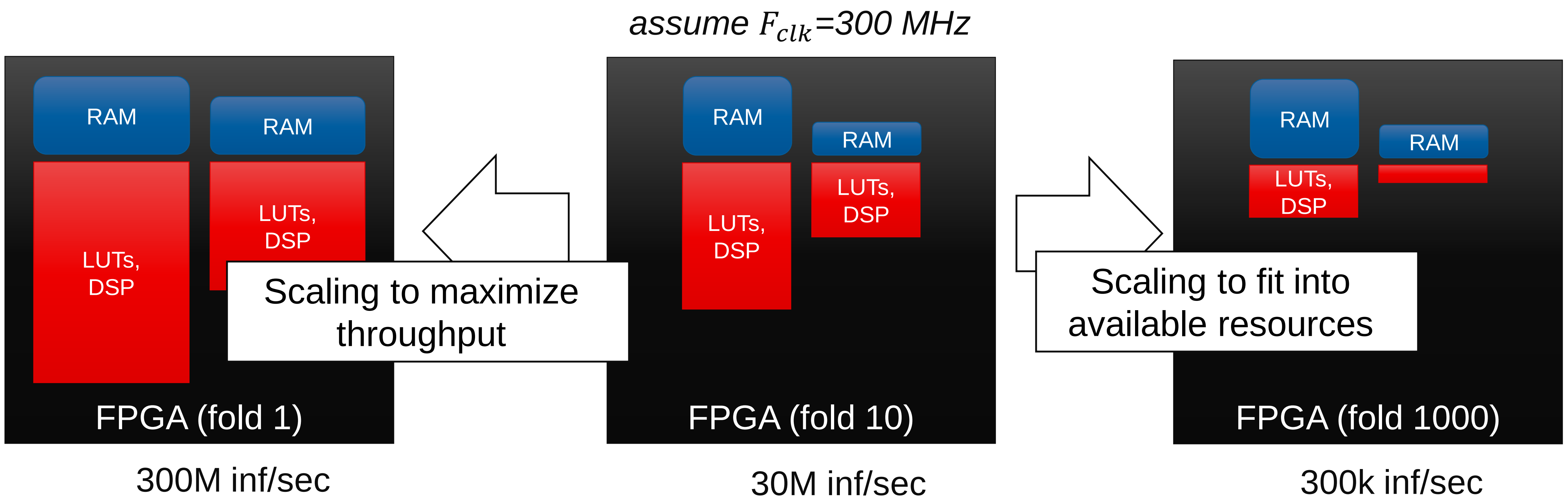
Each layer gets its own parallel hardware!
On-chip streaming dataflow with FIFO links

Vivado IPI design
+ PYNQ deployment

XILINX

© Copyright 2021 Xilinx

Dataflow Hardware Generation with the FINN Compiler:
Scaling to Meet Performance & Resource Requirements



1. Scale performance & resources to meet the application requirements
2. If resources allow, we can completely unfold to create a circuit that inferences at clock speed *(not practical for RadioML-sized circuits)*

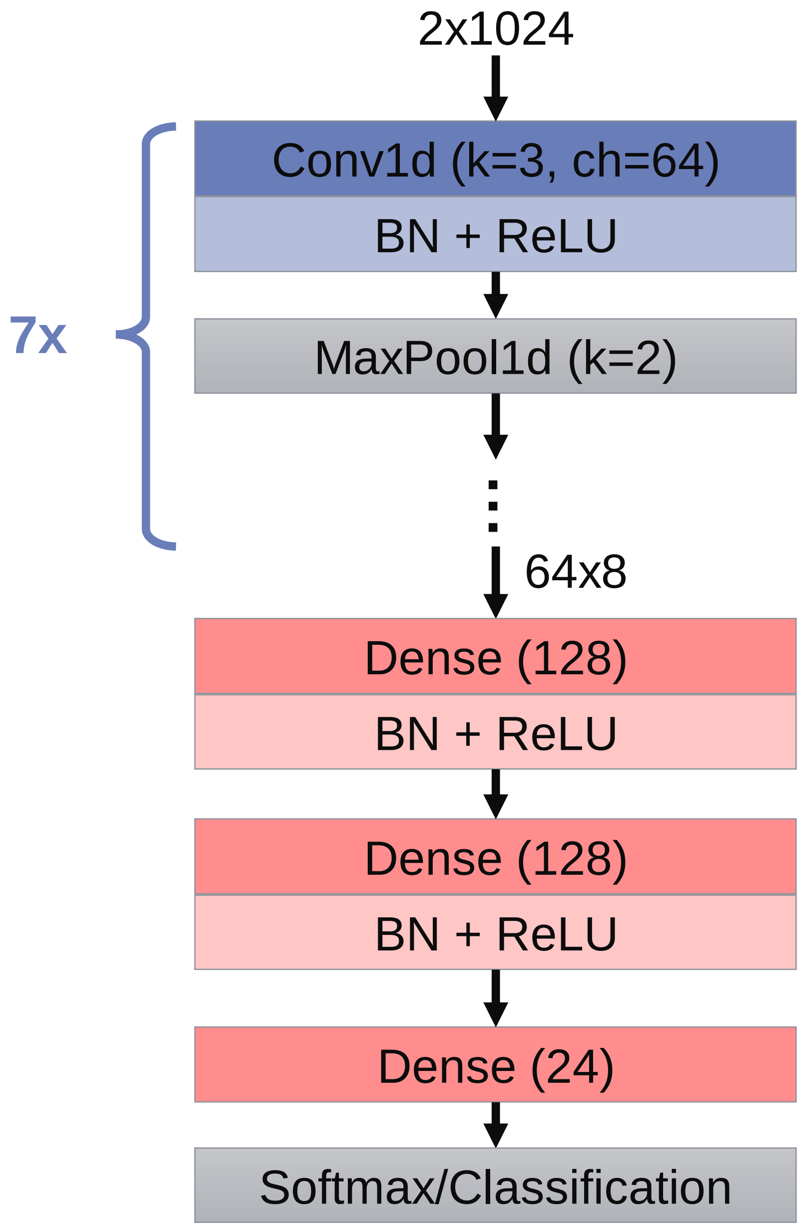
© Copyright 2021 Xilinx

First FINN RadioML Prototypes

© Copyright 2021 Xilinx

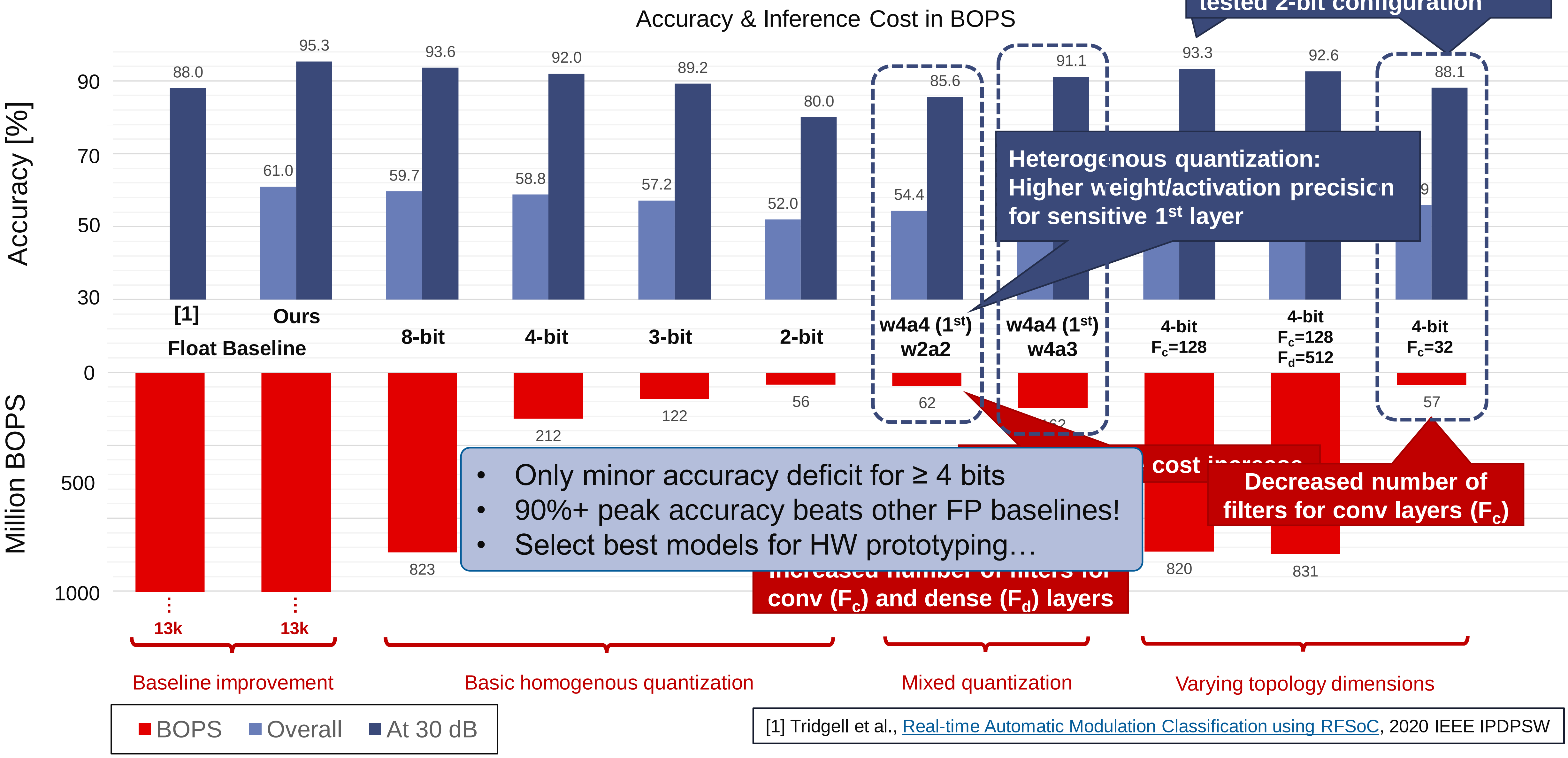
Initial Topology for our Experiments

- 1-dimensional convolutional neural net (**VGG10**)
  - Proposed by dataset creators along 2018 dataset
  - Proven performance against traditional classification techniques
  - Simple starting point without residual connections
- Baseline topology for competition
- Brevitas training setup
  - 8-bit input quantization to fixed range
    - ~98% percentile across whole dataset at high SNR
  - Using all available training data (whole SNR range)
    - Focus on overall accuracy
    - If only high SNR accuracy is of interest: Train on high SNR for ~1-2% gain



© Copyright 2021 Xilinx

Training Results



[1] Tridgell et al., Real-time Automatic Modulation Classification using RFSoC, 2020 IEEE IPDPSW

XILINX

© Copyright 2021 Xilinx

Current FINN Hardware Prototypes

- Working prototypes on ZCU111 @ 200 MHz:

	A	B	C
Topology	VGG10	VGG10	VGG10, F_c=32
Quantization	w4a4+w4a3	w4a4+w2a2	w4a4
Accuracy overall	58.5%	54.4%	55.9
Accuracy @ 30 dB	91.1%	85.6%	88.1
Throughput [samples/s]	190M	190M	190M
Latency [us]	16	16	16
LUT (util.)	196k (46%)	105k (25%)	82k (19%)
BRAM18 (util.)	358 (17%)	176 (8%)	116 (5%)

Same parallelism ► same performance

Inference at ~ 1 sample/cycle!

Not limited by resource constraints (scale-up WIP)

XILINX

© Copyright 2021 Xilinx

Your Submissions vs. our Prototype

Unstructured pruning is challenging to exploit

Team	W Bits	BOPS	Sparsity	Cost Score
Prototype A (4/3-bit)	536k	162M	~10%	0.316
Prototype B (4/2-bit)	137k	53M	~40%	0.088
Prototype C (4-bit, Fc=32)	**237k**	**50M**	**~10%**	**0.126**
3. Aaronica	**53k**	**40M**	**~30%**	**0.046**
2. The A(MC) Team	**68k**	**24M**	**~90%**	**0.042**
1. BacalhauNET	**11k**	**19M**	**~80%**	**0.016**

BRAM	LUT
358	196k
176	105k
116	**82k**

?

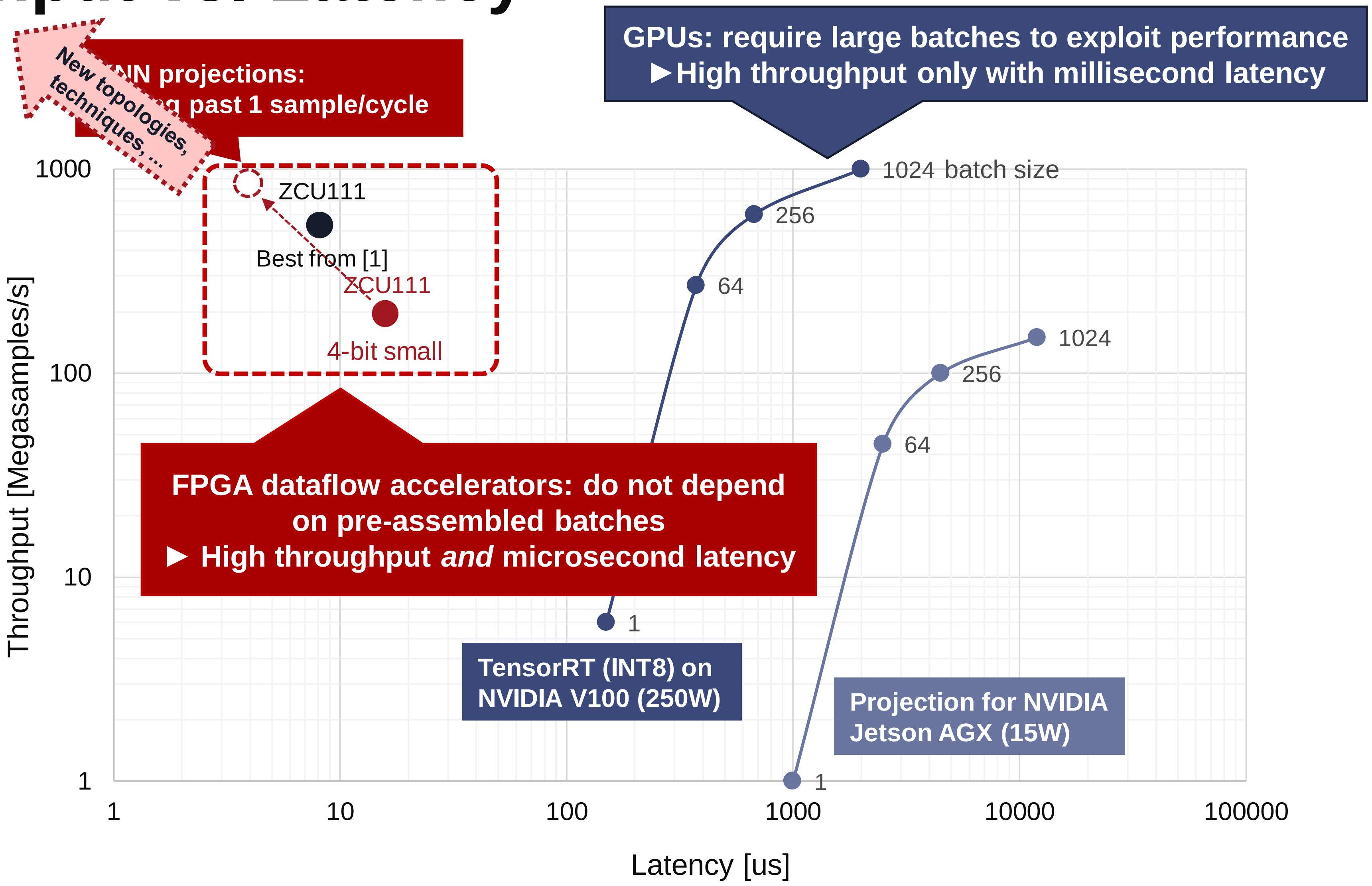
reach ~56% overall accuracy

Winning submissions look very promising

© Copyright 2021 Xilinx

XILINX

Throughput vs. Latency



[1] Tridgell et al., Real-time Automatic Modulation Classification using RFSoC, 2020 IEEE IPDPSW

XILINX.

© Copyright 2021 Xilinx

Conclusion & Future Work

© Copyright 2021 Xilinx

Conclusion

- DL in communications demands extreme throughput & latency
  - Your submissions show how much smaller the models could be
- Specialized DNNs + streaming dataflow showcase what's possible with FPGAs
  - High-throughput, low-latency, streaming inference without batching
  - Brevitas + FINN well-suited for exploration & implementation
- Initial FINN prototypes for RF modulation classification
  - Already achieving near 200M samples/sec at <20 us latency, with 90%+ accuracy
  - Soon available as part of finn-examples GitHub repository

XILINX

© Copyright 2021 Xilinx

Future Work

- Further FINN improvements geared towards 1D/time-series networks
- Explore advanced topologies, leveraging recent work on ResNets in FINN
- Enable you to get your models running on FPGAs

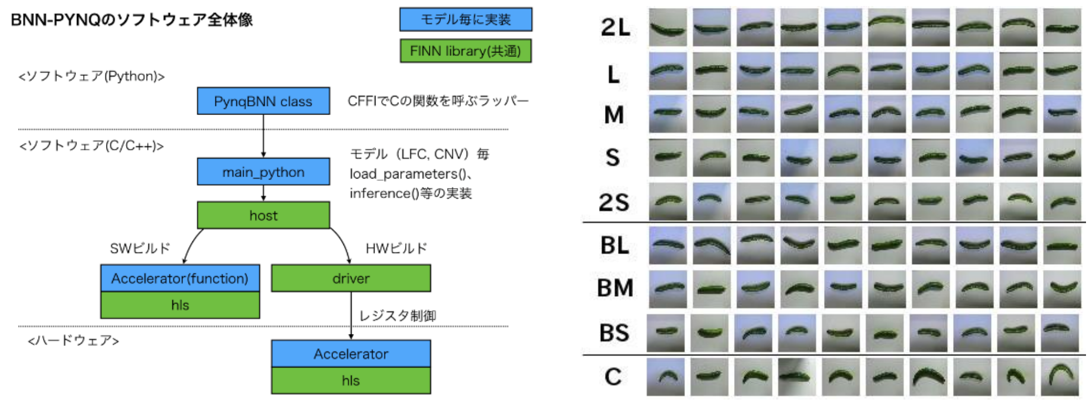
XILINX

© Copyright 2021 Xilinx

Join the growing FINN community!

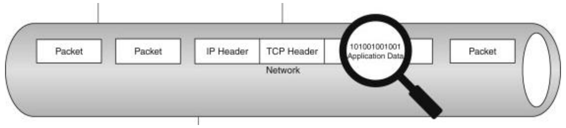
https://github.com/Xilinx/finn/discussions
https://gitter.im/xilinx-finn/community

Face mask detection (TU Munich, BMW)

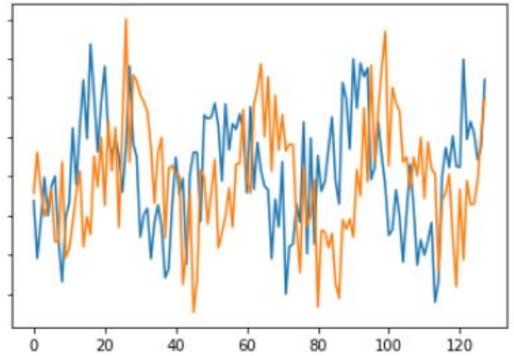


Documentation in Japanese + cucumber sorting demo (tkat0, ykshr)

Speech recognition (TU Delft)



Network intrusion detection (KU Leuven)



<your RadioML FPGA accelerator here!>

XILINX

© Copyright 2021 Xilinx

Thank You

Please fill out the feedback form, it only takes a minute:
https://bit.ly/brevitas-radioml-challenge-21-feedback

© Copyright 2021 Xilinx