FINN: A Framework for Fast, Scalable Binarized Neural Network Inference on Reconfigurable Logic

Yaman Umuroglu, Nicholas J. Fraser, Giulio Gambardella, Michaela Blott,
Philip Leong, Magnus Jahre and Kees Vissers

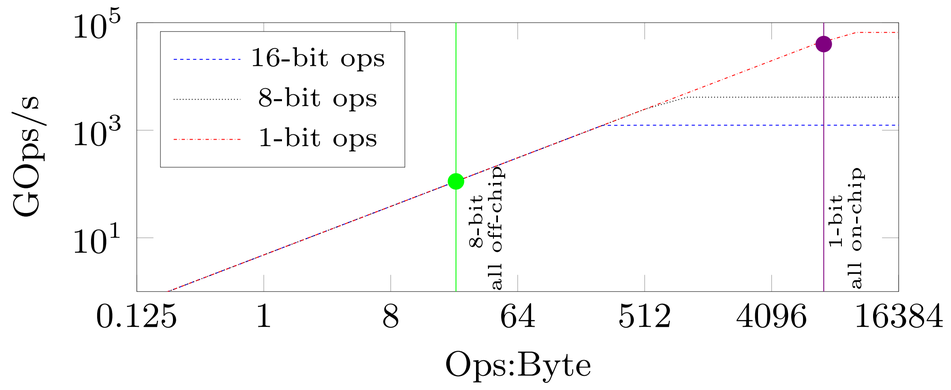
1 Binarized Neural Networks (BNNs)

- Almost all arithmetic is performed using two values: {-1, +1}
- Trained via backprop on GPU, weights constrained during training
- Convolutional, fully-connected, pooling and batchnorm layers
- Competitive accuracy for image classification tasks

Dataset	FP32	BNN
MNIST	99%	99%
SVHN	98%	97%
CIFAR-10	92%	90%
ImageNet (AlexNet arch)	80% top-5	69% top-5
ImageNet (ResNet-18 arch)	89% top-5	73% top-5
ImageNet (GoogLeNet arch)	90% top-5	86% top-5
ImageNet (DoReFa-Net)	56% top-1	50% top-1

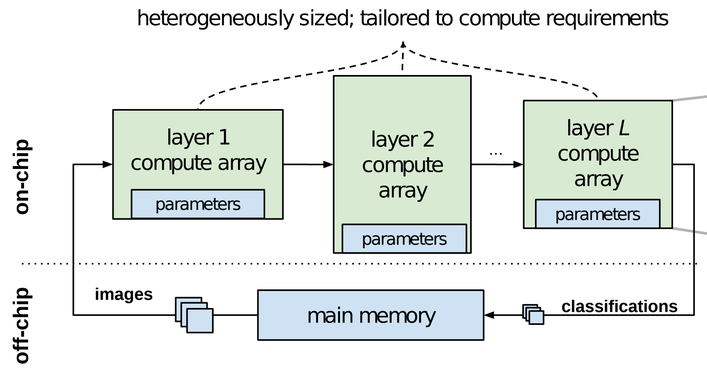
2 FPGA Potential Performance on BNNs

- Multiplications → XNOR, additions → popcount
- FPGA peak for binary ops is *much* higher than FP32 or INT8
  - ZU19EG: 66 TOPS binary, 4 TOPS INT8, 0.3 TOPS FP32
- Keeping all weights on-chip greatly increases arithmetic intensity
  - Avoid power and performance cost of most off-chip accesses



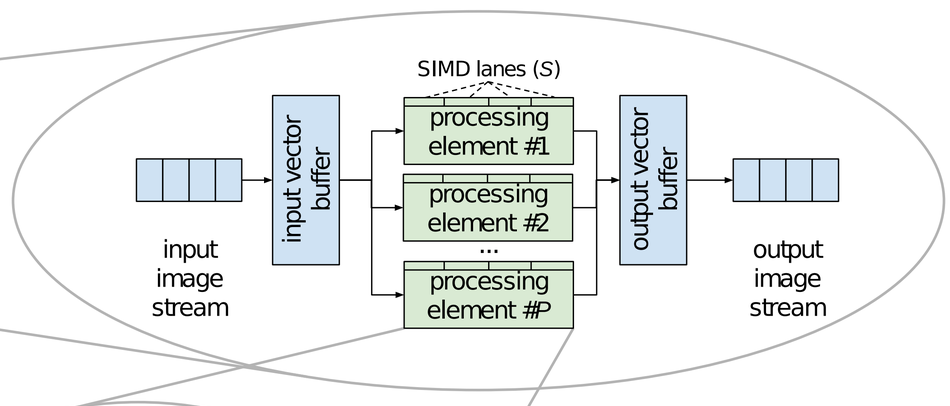
3 Generating BNN Inference Accelerators with FINN

Top Level: Heterogeneous & Streaming



- Build architecture for topology instead of compiling for fixed arch
  - Streaming architecture generated by Vivado HLS
- Compute resources heterogeneously allocated per-layer...
  - ...to balance the streaming pipeline (big & small layers)
  - ...to meet the user-specified FPS requirement (avoid waste)

Compute Arrays: SIMD & Multi-core



- Broadcast input vector to all cores
- Accumulation local to core
- Cores in layer operate in lockstep
- Exploit reuse in on-chip buffers

4 Experimental Evaluation on ZC706

BNN Topologies & Scenarios

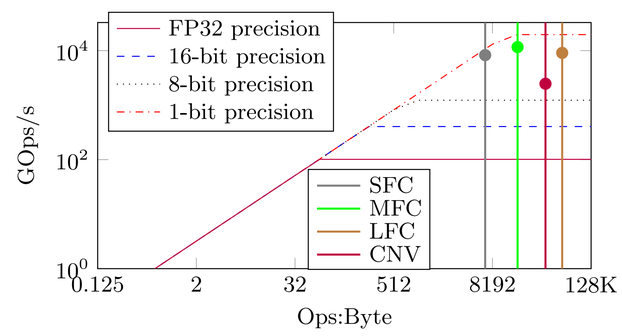
- Three BNN topologies:
  - **SFC** - fully-connected, 95.8% on MNIST
  - **LFC** - fully-connected, 98.4% on MNIST
  - **CNV** - VGG16-like convolutional
    80.1% on CIFAR-10, 94.9% on SVHN
- Two use-case scenarios:
  - **max** : maximum FPS (e.g. datacenter)
  - **fix** : 9000 FPS (e.g. embedded)

Key Metrics

Name	Thr.put (FPS)	Latency (s)	LUT	BRAM	P_{chip} (W)	P_{wall} (W)
SFC-max	12361 k	0.31	91131	4.5	7.3	21.2
LFC-max	1561 k	2.44	82988	396	8.8	22.6
CNV-max	21.9 k	283	46253	186	3.6	11.7
SFC-fix	12.2 k	240	5155	16	0.4	8.1
LFC-fix	12.2 k	282	5636	114.5	0.8	7.9
CNV-fix	11.6 k	550	29274	152.5	2.3	10

- Up to 12.3 million MNIST images per sec
- Up to 12.2 thousand CIFAR-10 images per sec

Achieved Performance vs Roofline



- 11.6 of 19.7 TOPS (68% of peak)

5 Summary

- Even mid-range FPGAs can perform trillions of binary operations per second, which can be harnessed for BNN inference
- Unprecedented image classification rates at <25 W power and <1 ms latency for MNIST and CIFAR-10 datasets
- Future work will focus on larger topologies (ImageNet), mixed precision and supporting off-chip parameters

XILINX®
THE UNIVERSITY OF SYDNEY
NTNU