

Deep Learning using FPGA Devices

Philip Leong

Director, Computer Engineering Laboratory

<http://phwl.org/talks>



THE UNIVERSITY OF
SYDNEY

Computer Engineering Laboratory

- › Focuses on how to use parallelism to solve demanding problems
 - Novel architectures, applications and design techniques using VLSI, FPGA and parallel computing technology
- › Research
 - Machine learning
 - Reconfigurable computing
 - Nanoscale Interfaces
- › Ex-students
 - AMD/Xilinx, Intel, Waymo, Amazon, Qualcomm



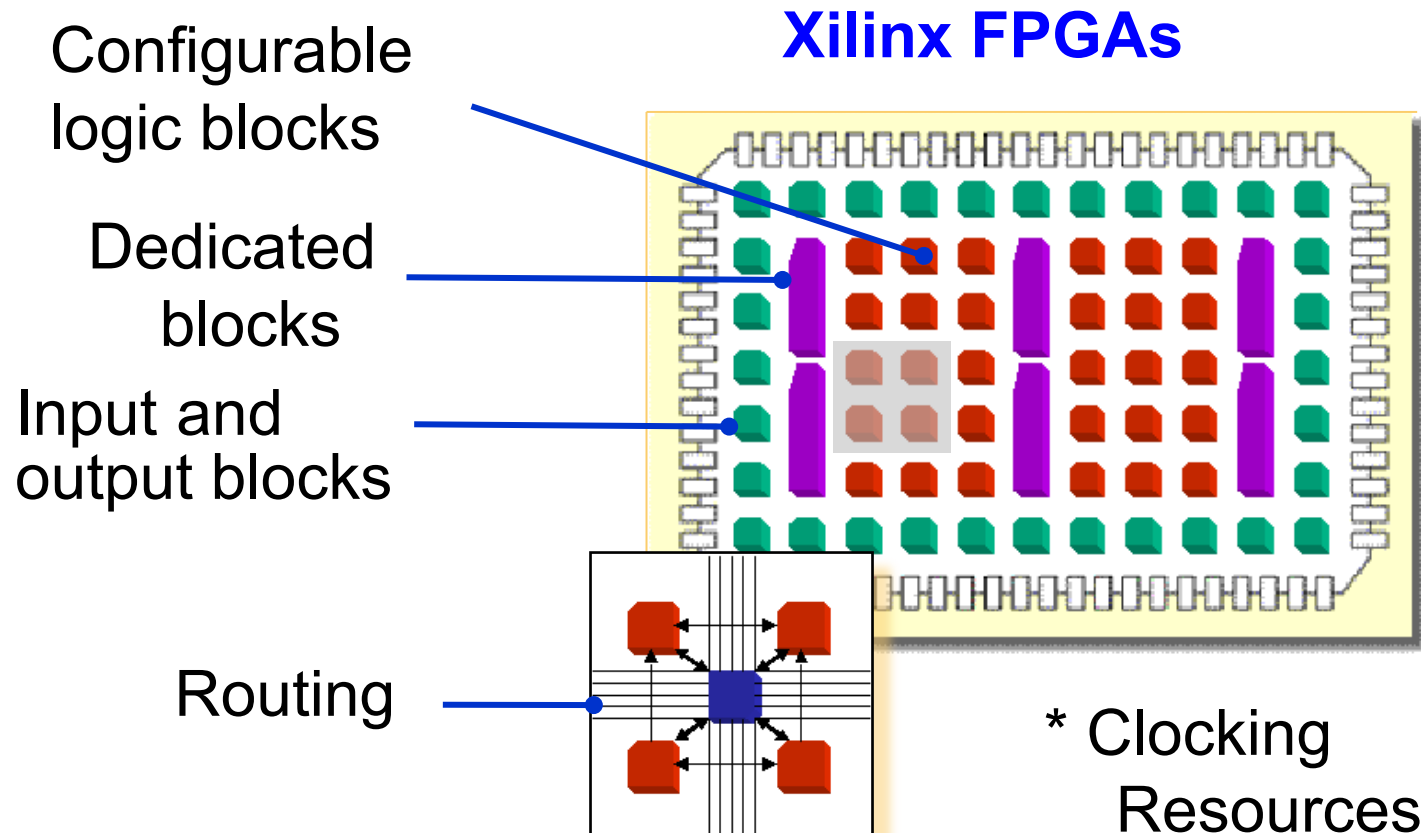
- › GPUs widely used for DNNs (optimized for throughput)
- › FPGAs achieve better **throughput**, **latency** and **power** through (EPIC)
 - Exploration –try different ideas to arrive at a good solution
 - Parallelism – arrive at an answer faster
 - Integration – so interfaces are not a bottleneck
 - Customisation – problem-specific designs to improve efficiency (power, speed, density)
- › This talk: describe our work in using FPGAs for DNN acceleration



What is an FPGA?

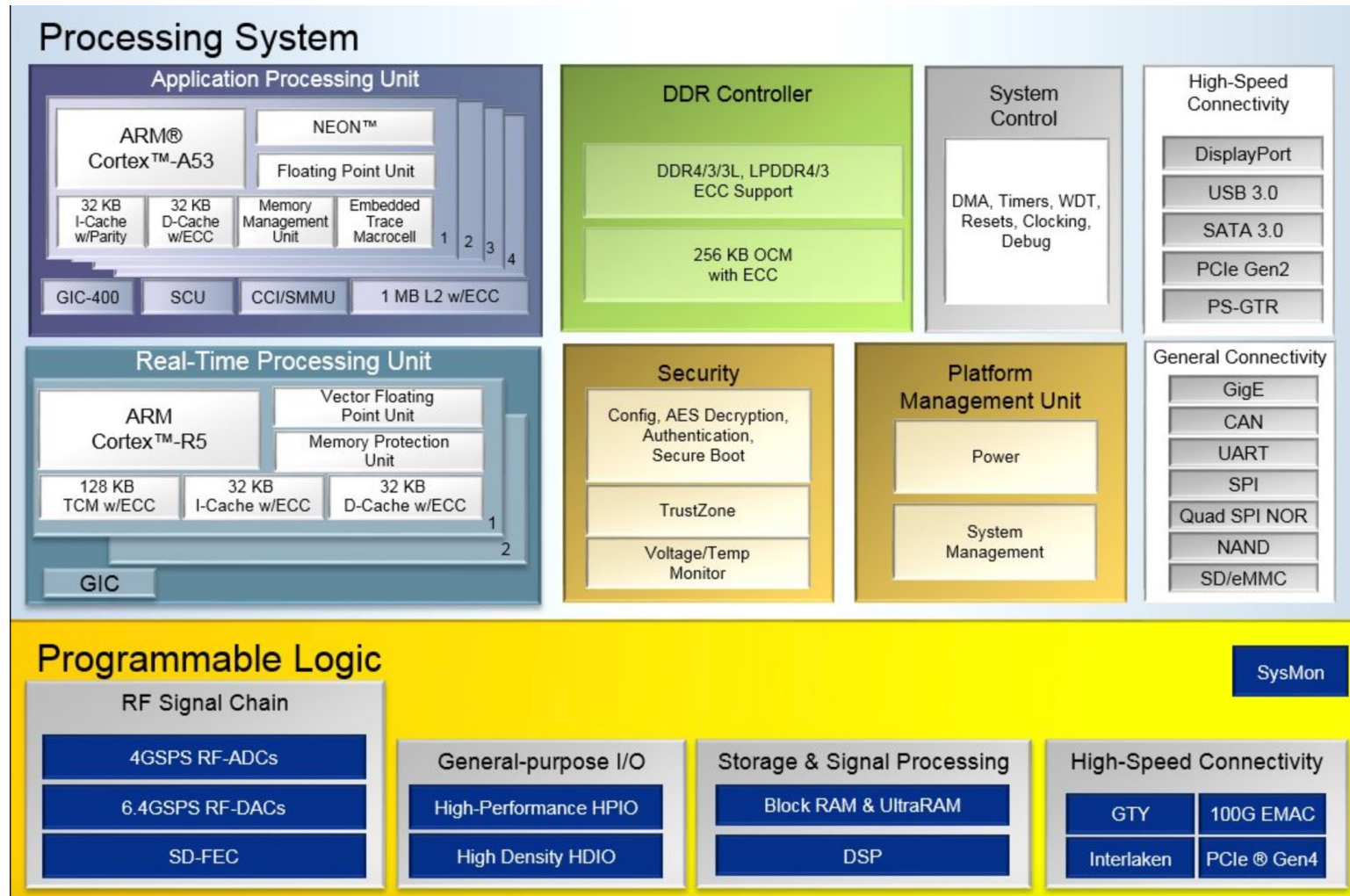
User-customisable integrated circuit

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



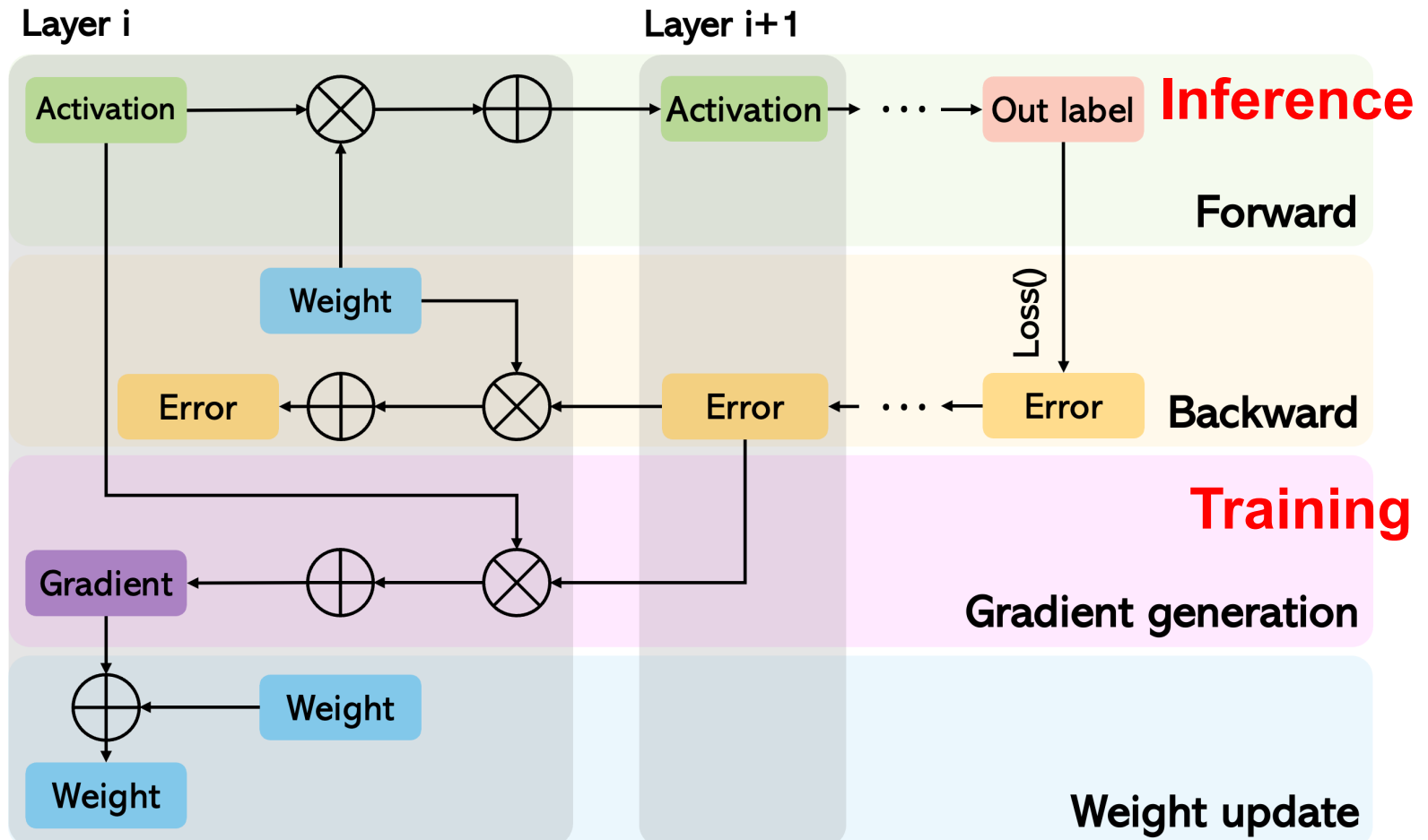


Xilinx RFSoc Device





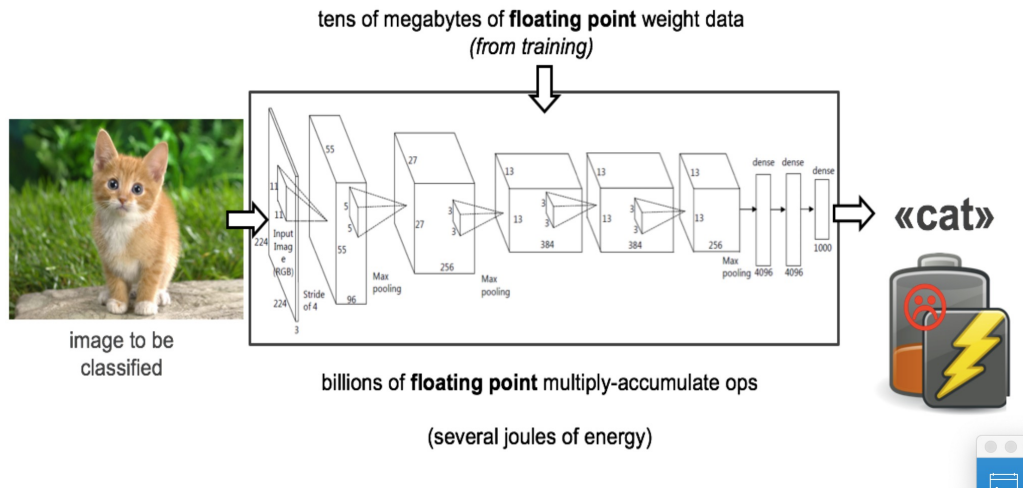
Convolutional Neural Networks (CNNs)





Binarized Neural Networks (FPGA17 [1])

Collaboration with Xilinx



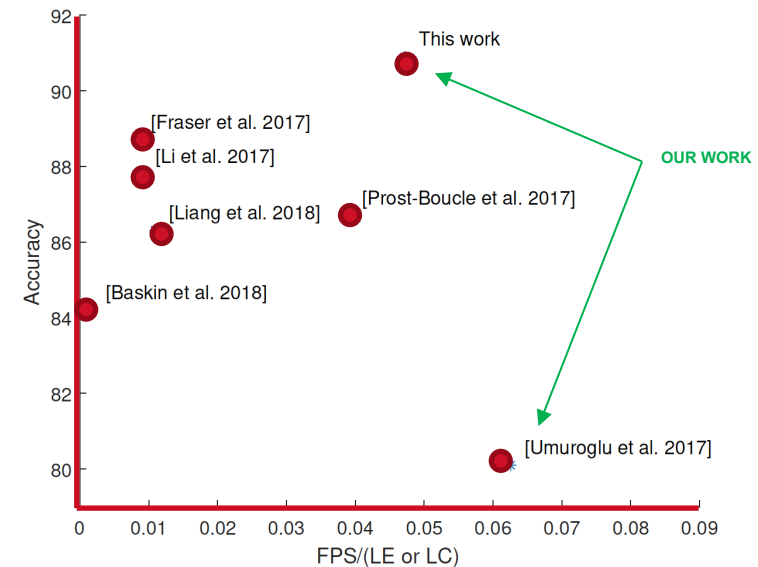
-0.4	-0.4	0.9
0.9	0.4	0.8
0.4	-0.4	-0.4

W

≈ 0.2

-1	-1	1
1	1	1
1	-1	-1

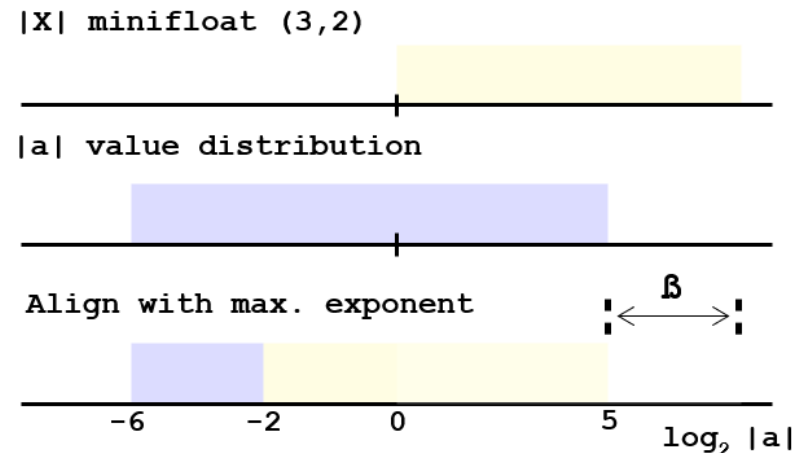
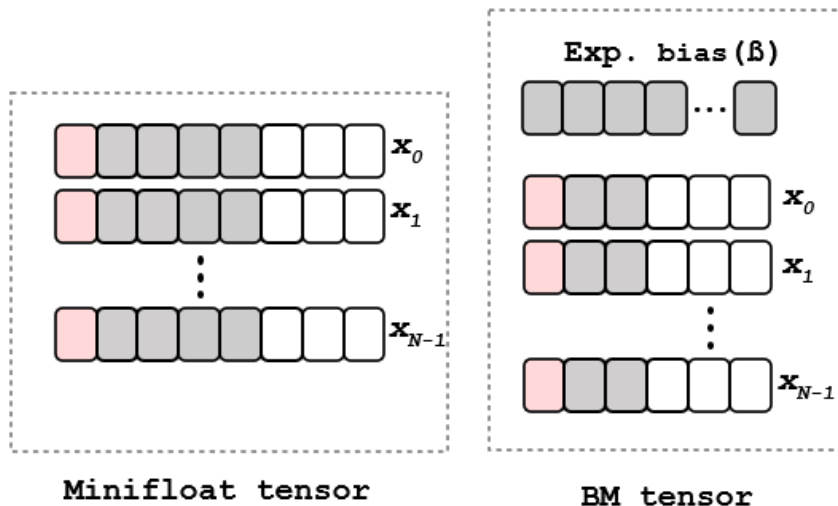
αW^B



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

Block Minifloat for Training (ICLR21 [2])

- Share exponent bias across **blocks** of $N \times N$ minifloat numbers

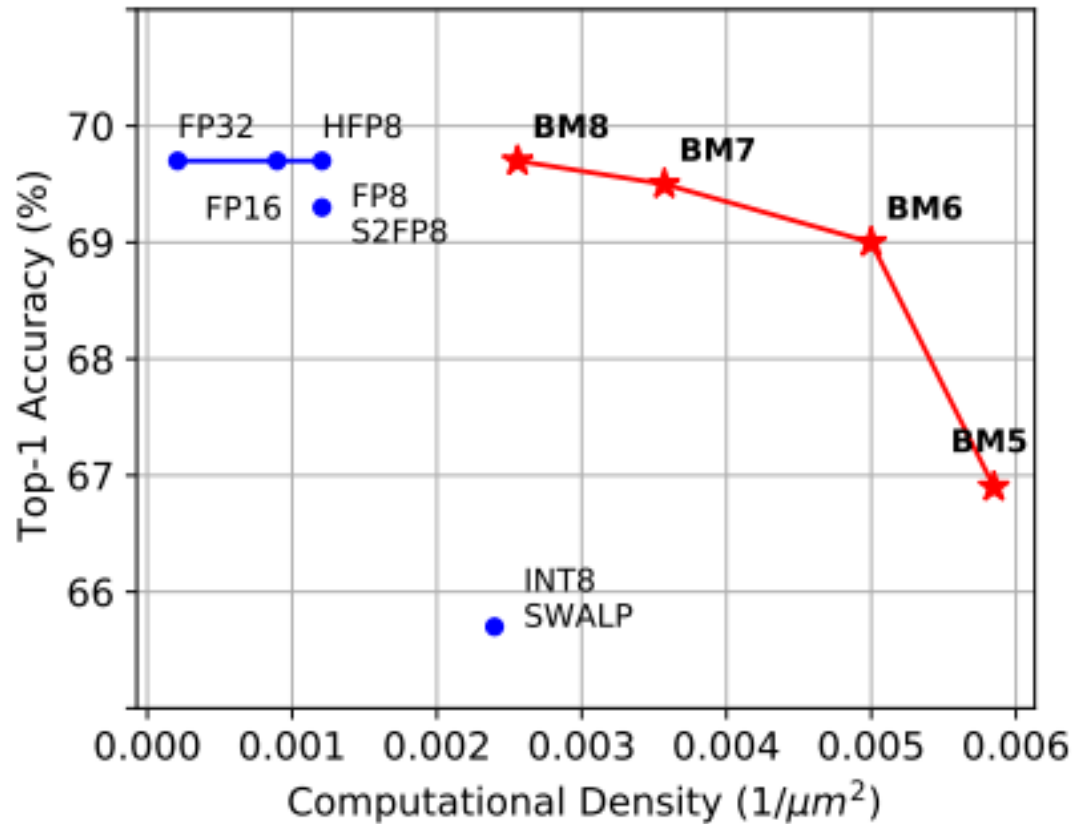


- Dynamic range (with fewer bits)
- Denser dot-products in hardware

- Align with **max** exponent
- Underflow is tolerated



Imagenet Training using Block minifloat (ICLR21 [2])

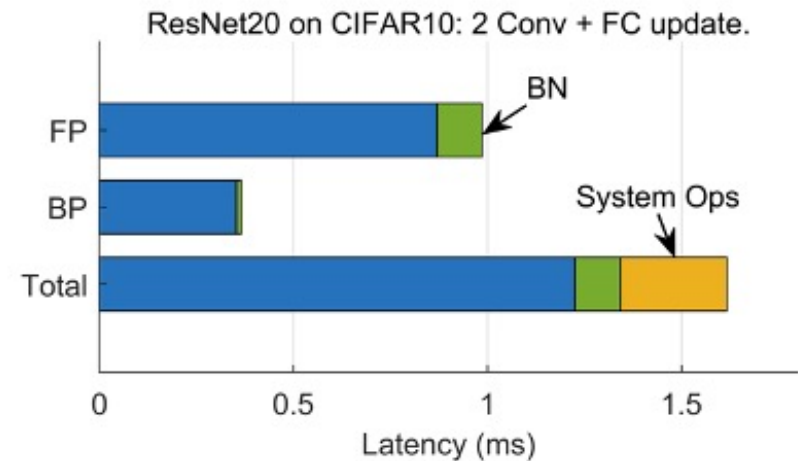
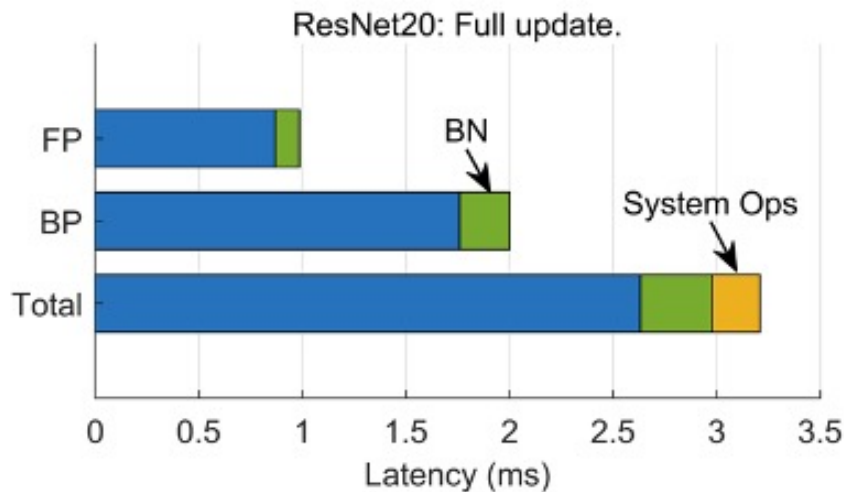


BM units are:

- **Smaller**
- **Consume less Power**

Model: ResNet-18
Dataset: ImageNet

CNN Transfer Learning on FPGA (ICCAD23 [3])

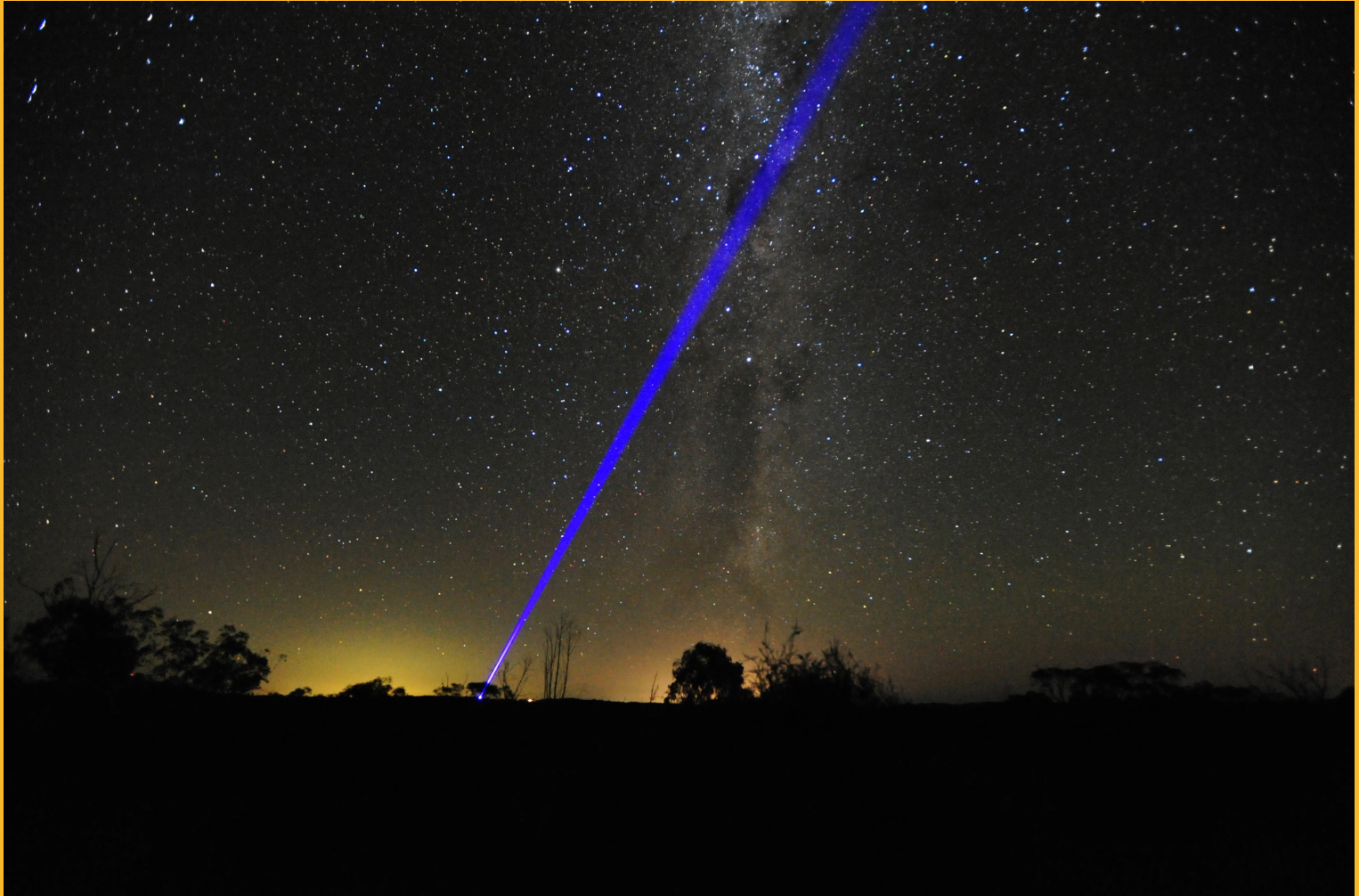


- › Applied block minifloat to transfer learning
- › Implemented on FPGA using high-level synthesis
 - Reduced backpropagation time (4x faster)
 - Overall latency reduced (2x faster)

- › Deep neural network acceleration is important for many real-time applications (self-driving cars, communications systems, radar)
- › CPUs and GPUs can achieve good performance but latency usually high
- › FPGAs allow everything to be integrated on a single device and achieve the lowest latency
- › We are working on using the techniques developed for radio frequency machine learning

Available at. <https://phwl.org/assets/papers/papers>

1. Yaman Umuroglu, Nicholas J. Fraser, Giulio Gambardella, Michaela Blott, Philip Leong, Magnus Jahre, and Kees Vissers. FINN: a framework for fast, scalable binarized neural network inference. In *Proc. ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA)*, 65–74. 2017.
 2. Sean Fox, Seyedramin Rasoulinezhad, Julian Faraone, and David Boland Philip H.W. Leong. A block minifloat representation for training deep neural networks. In *Proc. of The International Conference on Learning Representations (ICLR)*. 2021.
 3. Chuliang Guo, Binglei Lou, Xueyuan Liu, David Boland, Philip H.W. Leong, and Cheng Zhuo. BOOST: block minifloat-based on-device CNN training accelerator with transfer learning. In *Proc. ICCAD*, to appear. 2023.
-



THE UNIVERSITY OF
SYDNEY