

tinyML[®] Talks

Enabling Ultra-low Power Machine Learning at the Edge

“Neural Architecture Search for Tiny Devices”

Philip Leong – Chief Technology Officer, CruxML Pty Ltd
Professor, Computer Systems

School of Electrical and Information Engineering, University of Sydney

April 20, 2023



www.tinyML.org



Thank you, **tinyML Strategic Partners**,
for committing to take tinyML to the next Level, together



Executive Strategic Partners

T I N Y



TALKS
webcast



EDGE IMPULSE

The Leading Development Platform for Edge ML

edgeimpulse.com

Qualcomm
AI research

Advancing AI research to make efficient AI ubiquitous

Power efficiency

Model design, compression, quantization, algorithms, efficient hardware, software tool

Personalization

Continuous learning, contextual, always-on, privacy-preserved, distributed learning

Efficient learning

Robust learning through minimal data, unsupervised learning, on-device learning

A platform to scale AI across the industry



Perception

Object detection, speech recognition, contextual fusion



Reasoning

Scene understanding, language understanding, behavior prediction



Action

Reinforcement learning for decision making



Edge cloud



Cloud



IoT/IIoT



Automotive



Mobile



Accelerate Your Edge Compute

SYNTIANT

Making Edge AI A Reality

www.syntiant.com

Platinum Strategic Partners

Renesas is enabling the next generation of AI-powered solutions that will revolutionize every industry sector.



[renesas.com](https://www.renesas.com)



**DEPLOY VISION AI
AT THE EDGE AT SCALE**

SONY

Gold Strategic Partners



AHEAD OF WHAT'S POSSIBLE™



AHEAD OF WHAT'S POSSIBLE™

Where what if
becomes what is.

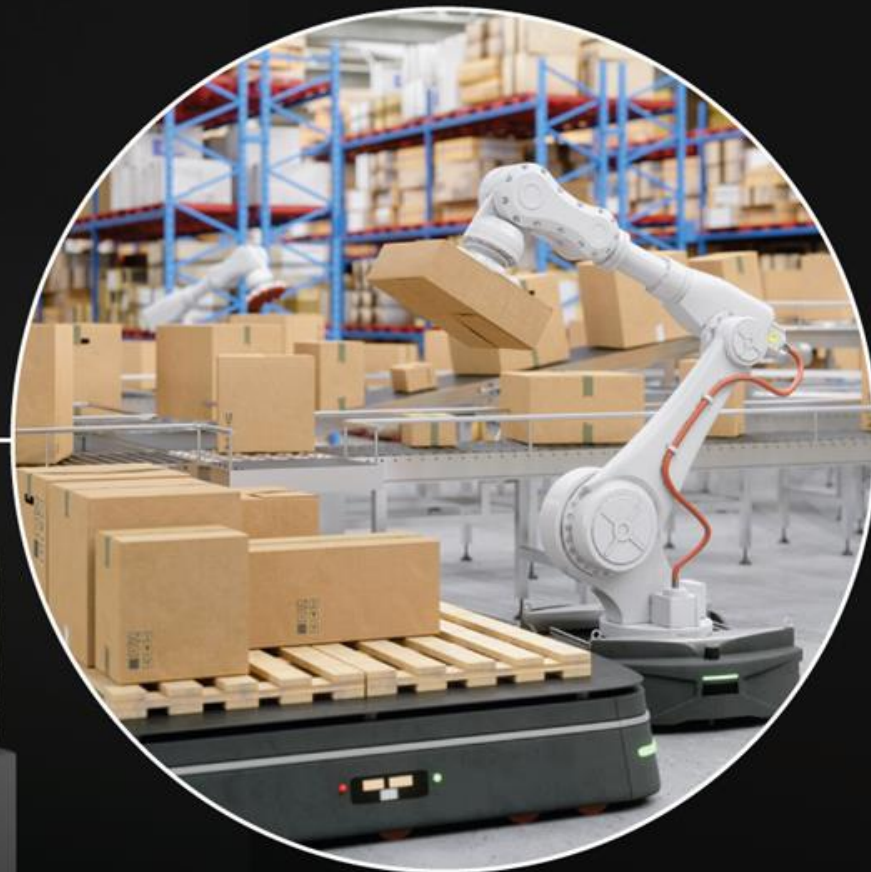
Witness potential made possible at analog.com.



PRO™

Easily deploy your
tinyML solutions with
Arduino Pro

arduino.cc/pro



Made In Italy

arm AI



Powering tinyML Innovation

Arm AI Virtual Tech Talks

The latest in AI trends, technologies & best practices from Arm and our Ecosystem Partners.

Demos, code examples, workshops, panel sessions and much more!

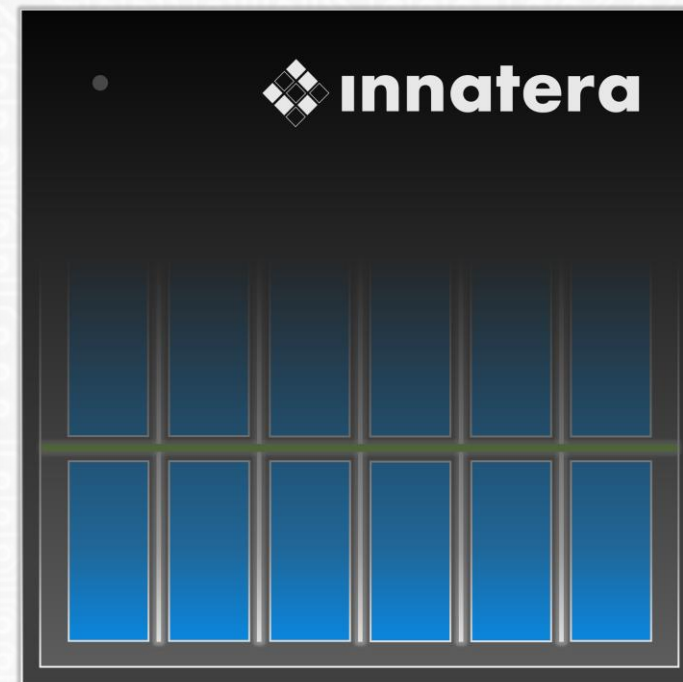
Fortnightly Tuesday @ 4pm GMT/8am PT

Find out more:

www.arm.com/techtalks



NEUROMORPHIC INTELLIGENCE FOR THE SENSOR-EDGE



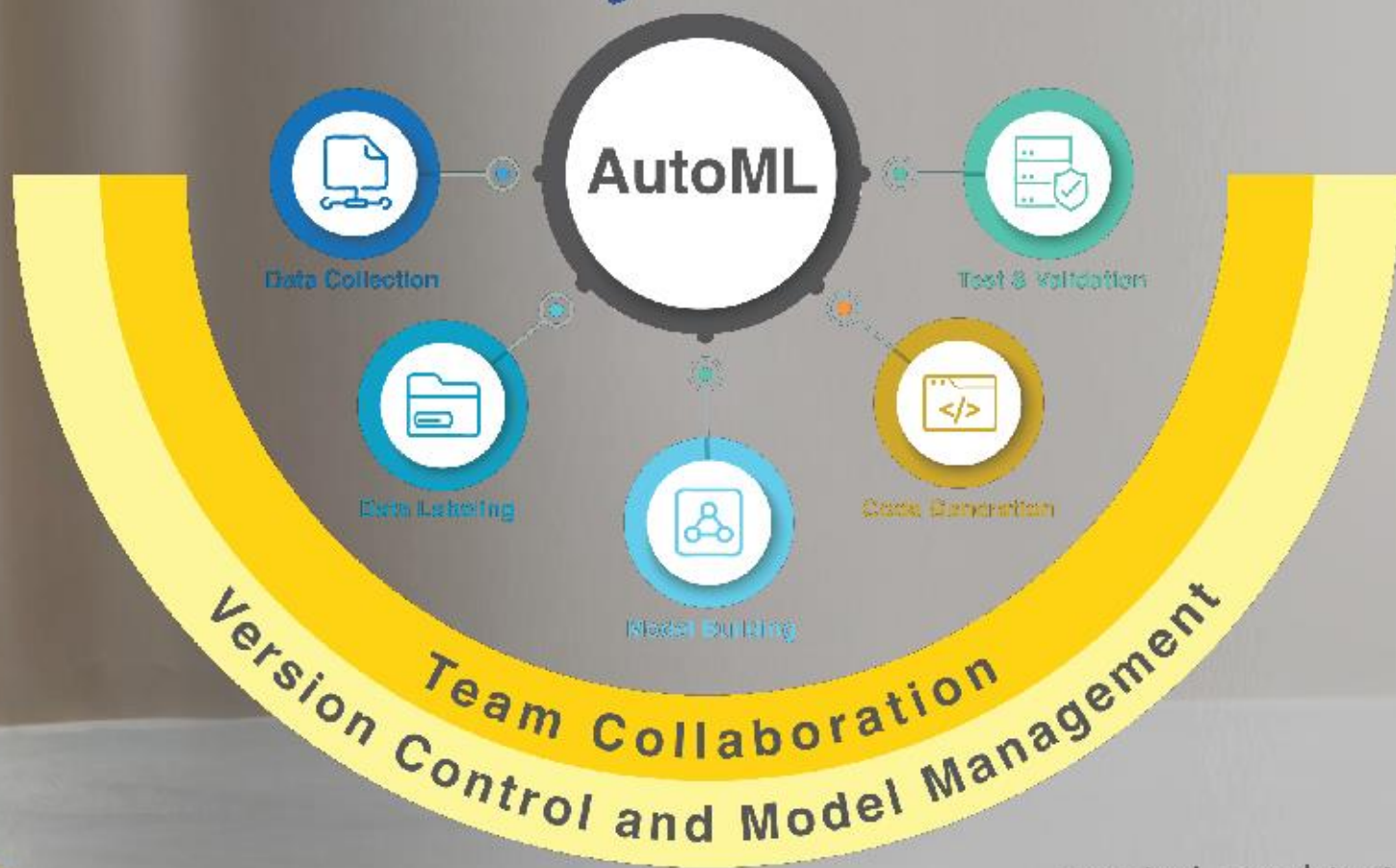


Microsoft

The Right Edge AI Tools Can Make or Break Your Next Smart IoT Product



Analytics Toolkit Suite





life.augmented

STMicroelectronics provides extensive solutions to make tiny Machine Learning easy



ENGINEERING EXCEPTIONAL EXPERIENCES

We engineer exceptional experiences for consumers in the home, at work, in the car, or on the go.

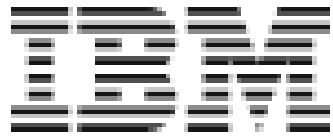
www.synaptics.com



T I N Y



Silver Strategic Partners





Join Growing tinyML Communities:



14k members in
47 Groups in 39 Countries

tinyML - Enabling ultra-low Power ML at the Edge

<https://www.meetup.com/tinyML-Enabling-ultra-low-Power-ML-at-the-Edge/>



4k members
&
11.6k followers

The tinyML Community

<https://www.linkedin.com/groups/13694488/>





Subscribe to
tinyML YouTube Channel
 for updates and notifications
(including this video)

www.youtube.com/tinyML



tinyML
4.33K subscribers

9.2k subscribers, 551 videos with 316k views

HOME VIDEOS PLAYLISTS COMMUNITY CHANNELS ABOUT

106 views · 4 days ago	138 views · 4 days ago	54 views · 4 days ago	47 views · 4 days ago	132 views · 4 days ago	137 views · 4 days ago
122 views · 4 days ago	262 views · 2 weeks ago	511 views · 3 weeks ago	229 views · 3 weeks ago	265 views · 3 weeks ago	286 views · 1 month ago
351 views · 1 month ago	462 views · 2 months ago	374 views · 2 months ago	133 views · 2 months ago	287 views · 2 months ago	336 views · 2 months ago
378 views · 2 months ago	214 views · 2 months ago	448 views · 2 months ago	159 views · 2 months ago	190 views · 2 months ago	545 views · 2 months ago



EMEA 2023

<https://www.tinyml.org/event/emea-2023>

More sponsorships are available: sponsorships@tinyML.org

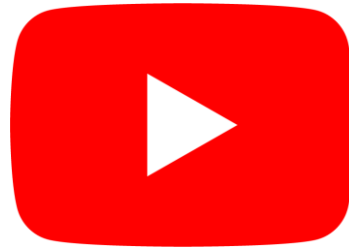


Reminders

Slides & Videos will be posted tomorrow



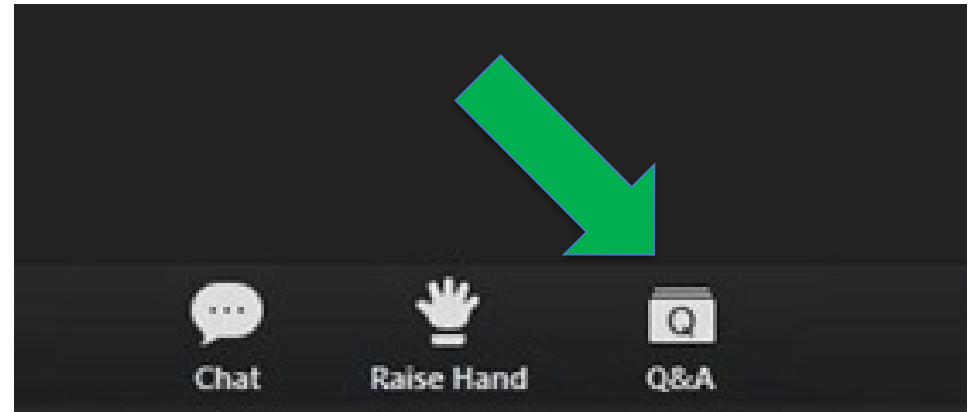
tinyml.org/forums



youtube.com/tinyml



Please use the Q&A window for your questions





Philip Leong



Philip Leong received the B.Sc., B.E. and Ph.D. degrees from the University of Sydney. In 1993 he was a consultant to ST Microelectronics in Milan, Italy working on advanced flash memory-based integrated circuit design. From 1997-2009 he was with the Chinese University of Hong Kong. He is currently Professor of Computer Systems in the School of Electrical and Information Engineering at the University of Sydney, Visiting Professor at Imperial College, and Chief Technology Officer at CruxML Pty Ltd.

Low Precision Inference and Training for Deep Neural Networks

Philip Leong
Director, Computer Engineering Laboratory
<http://phwl.org/talks>



THE UNIVERSITY OF
SYDNEY



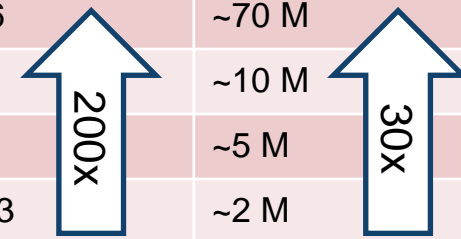
- › Focuses on how to use parallelism to solve demanding problems
 - Novel architectures, applications and design techniques using FPGAs
- › Research: reconfigurable computing, radio frequency machine learning



Tradeoff between performance and precision

- › CPUs/GPUs designed to support datatypes of fixed wordlength
 - Double, float, long, short, char
- › FPGA and ASICs can provide custom datapaths of arbitrary wordlength

Precision	Peak TOPS	On-chip weights
1b	~66	~70 M
8b	~4	~10 M
16b	~1	~5 M
32b	~0.3	~2 M



Slide: Xilinx

- › So how can we utilize low-precision for inference and training?



- › Block Minifloat
- › Time series Prediction
- › Transfer Learning

Block Minifloat

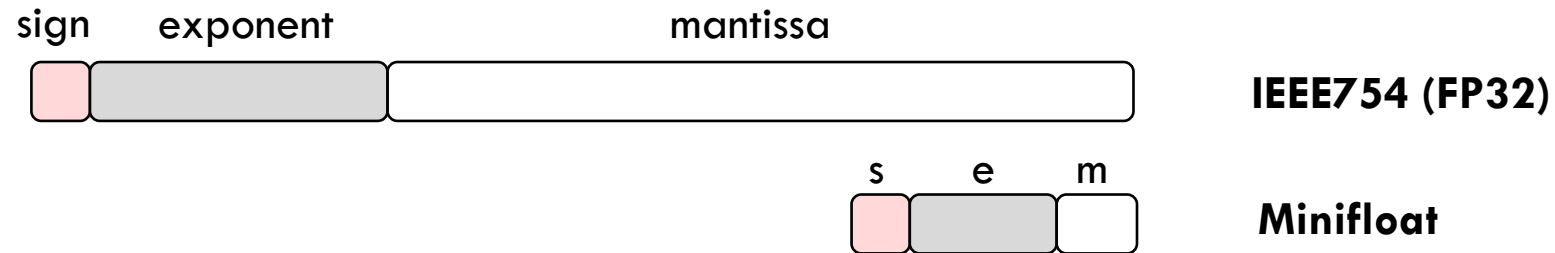
Sean Fox



THE UNIVERSITY OF
SYDNEY

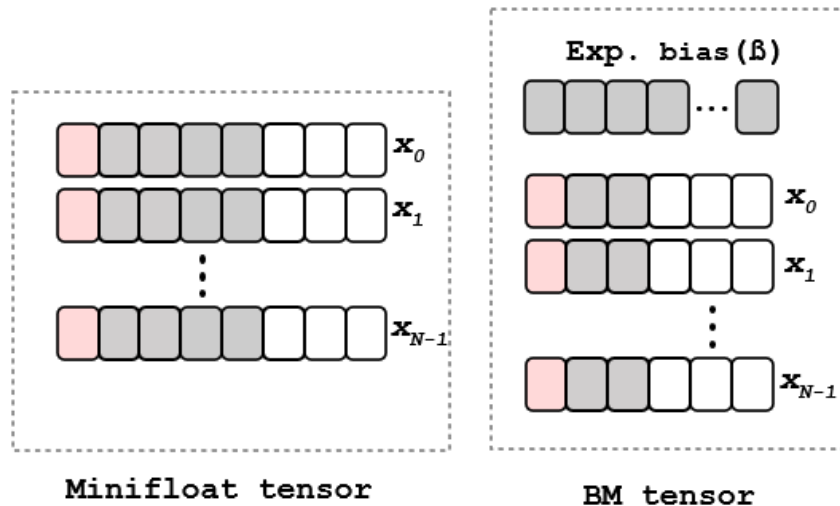
- Training has greater efficiency problem than inference!
 - E.g. 3x more MACs, much higher memory requirements
- Specialized number representations have been proposed
 - Alternatives to FP32/FP16
 - 4-8 bits for weights, activations and gradients
 - Cheaper and faster training systems
 - Focus on Edge (not sure about the Data Center)

- Narrow floating-point representation
 - Our range between 4-8 bits
 - NaN/Infinity NOT supported



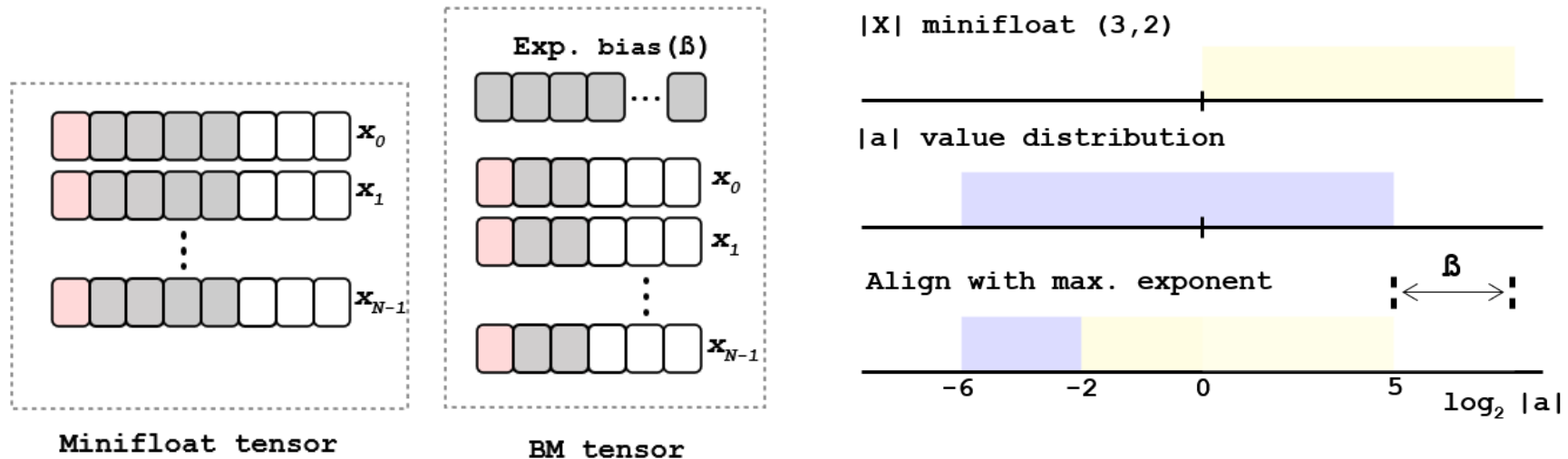
- Pros:
 - Memory (fewer bits)
 - Smaller hardware
- Cons:
 - Dynamic Range (exponent bits)

- Share exponent bias across **blocks** of NxN minifloat numbers



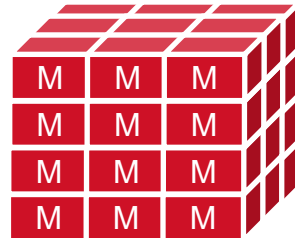
- Dynamic range (with fewer bits)
- Denser dot-products in hardware

- Share exponent bias across **blocks** of NxN minifloat numbers

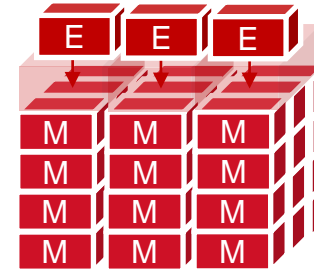


- Dynamic range (with fewer bits)
- Denser dot-products in hardware

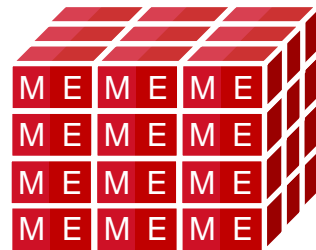
- Align with **max** exponent
- Underflow is tolerated



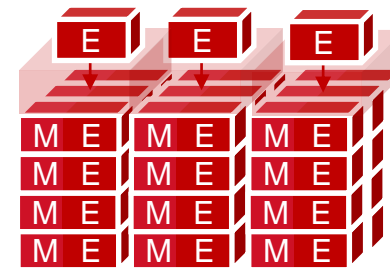
Fixed



BFP

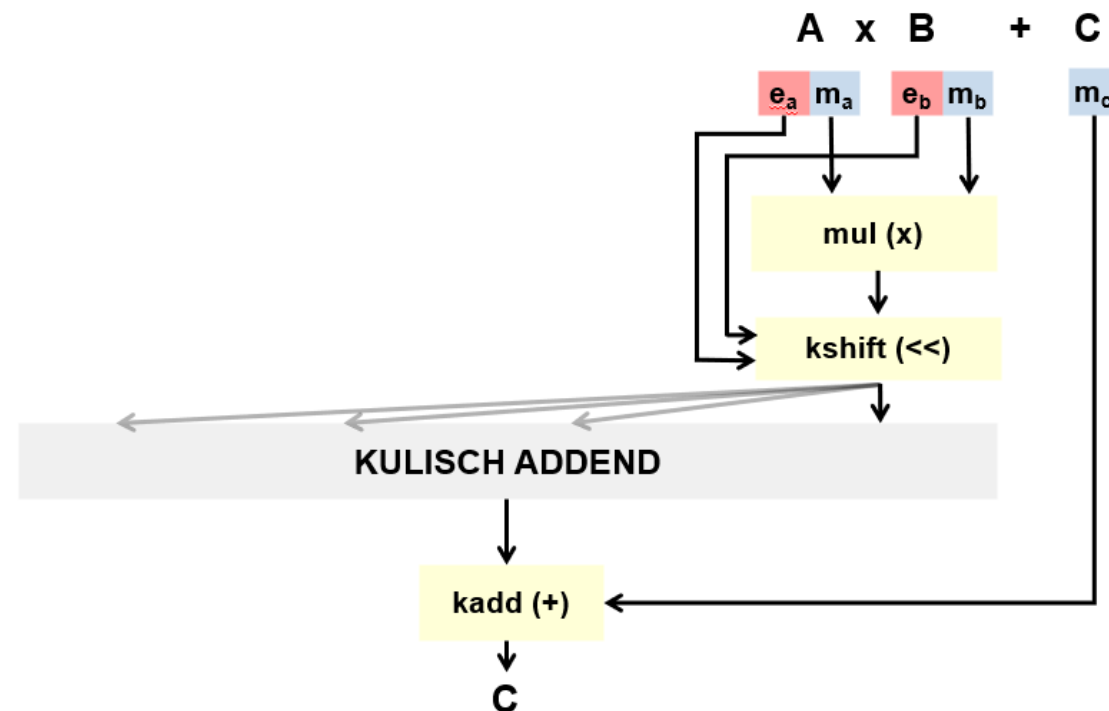


Minifloat



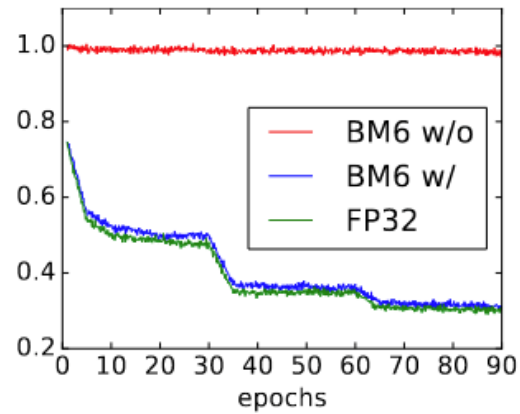
Block Minifloat

- **Kulisch Accumulator:** Fixed point accumulator wide enough to compute error-free sum of floating-point products
- Integer-like hardware complexity for **exponent ≤ 4 bits**

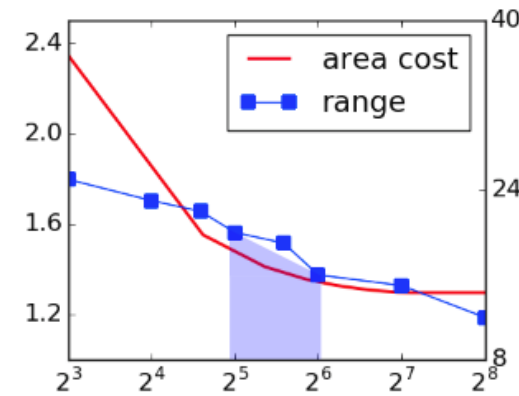


- Three techniques to reduce data loss:
 - Gradual underflow, Block Design, Hybrid Formats
- Simulate specialized BM hardware on GPU (with FP32)
 - Apply Block Minifloat to all weights, acts, grads
- Our Spectrum of Block Minifloats

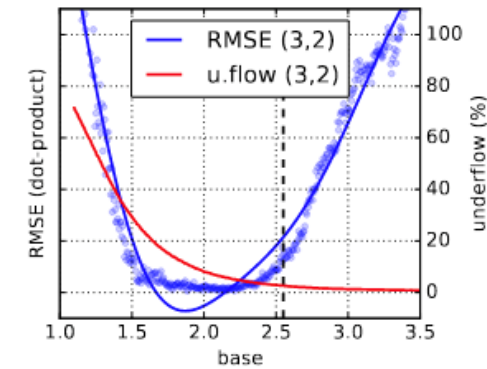
BM8 (ours)	(2,5)/(4,3)
BM7 (ours)	(2,4)/(4,2)
BM6 (ours)	(2,3)/(3,2)
BM5 (ours)	(2,2)/(3,1)
BM5-log (ours)	(4,0)/(4,0)
BM4 (ours)	(2,1)/(3,0)
BM4-log (ours)	(3,0)/(3,0)



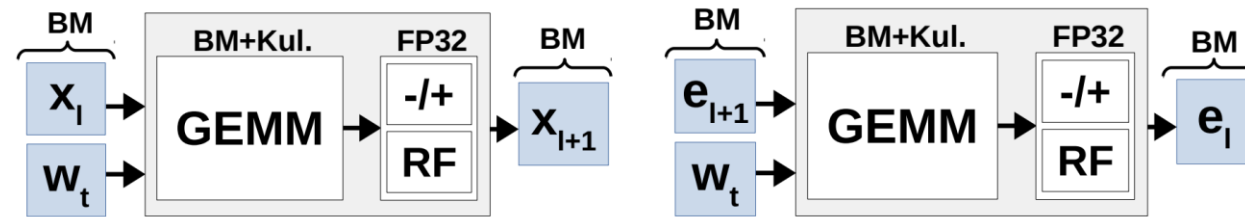
(a) Validation Accuracy: Training with denormal numbers on ImageNet



(b) HW (left axis) vs Range (right axis): Selecting the block size

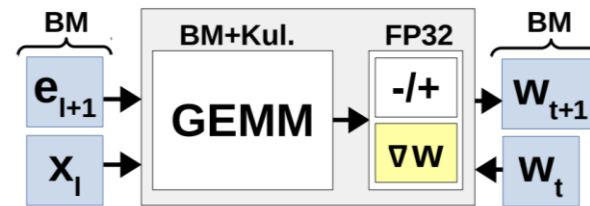


(c) Minifloat scaling by varying the exponent base



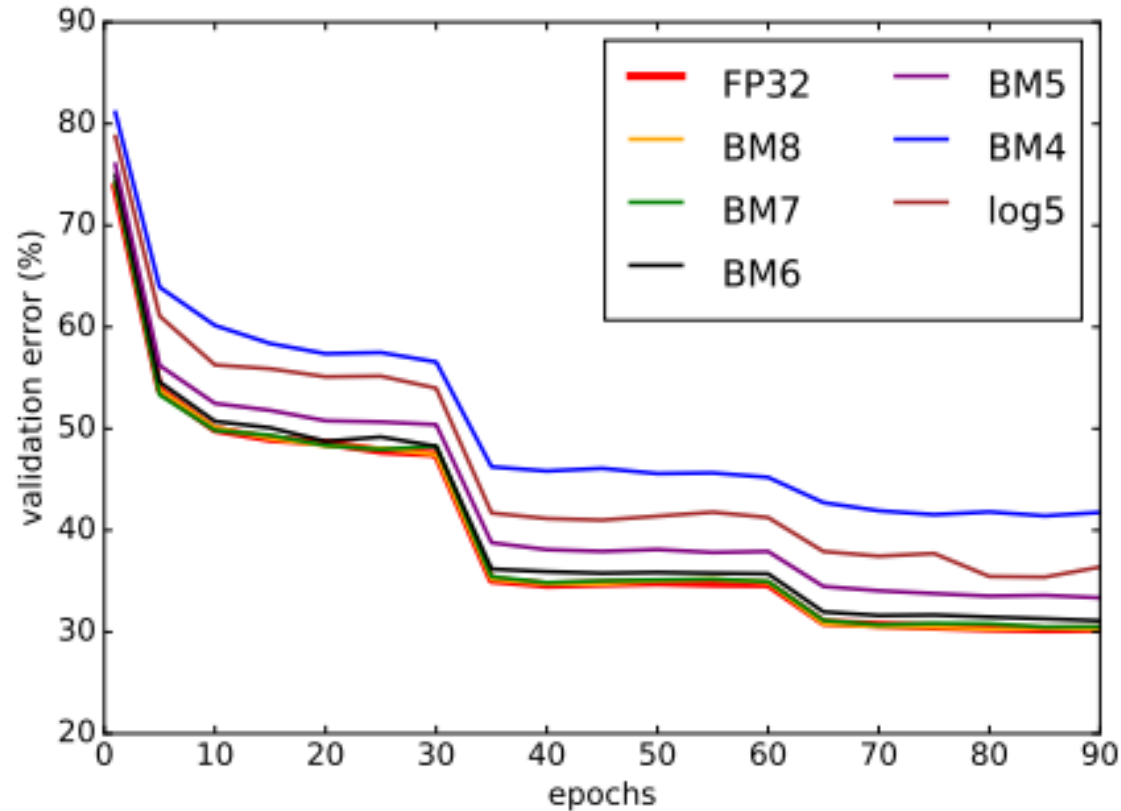
(a) Fwd activation

(b) Bwd activation grad.



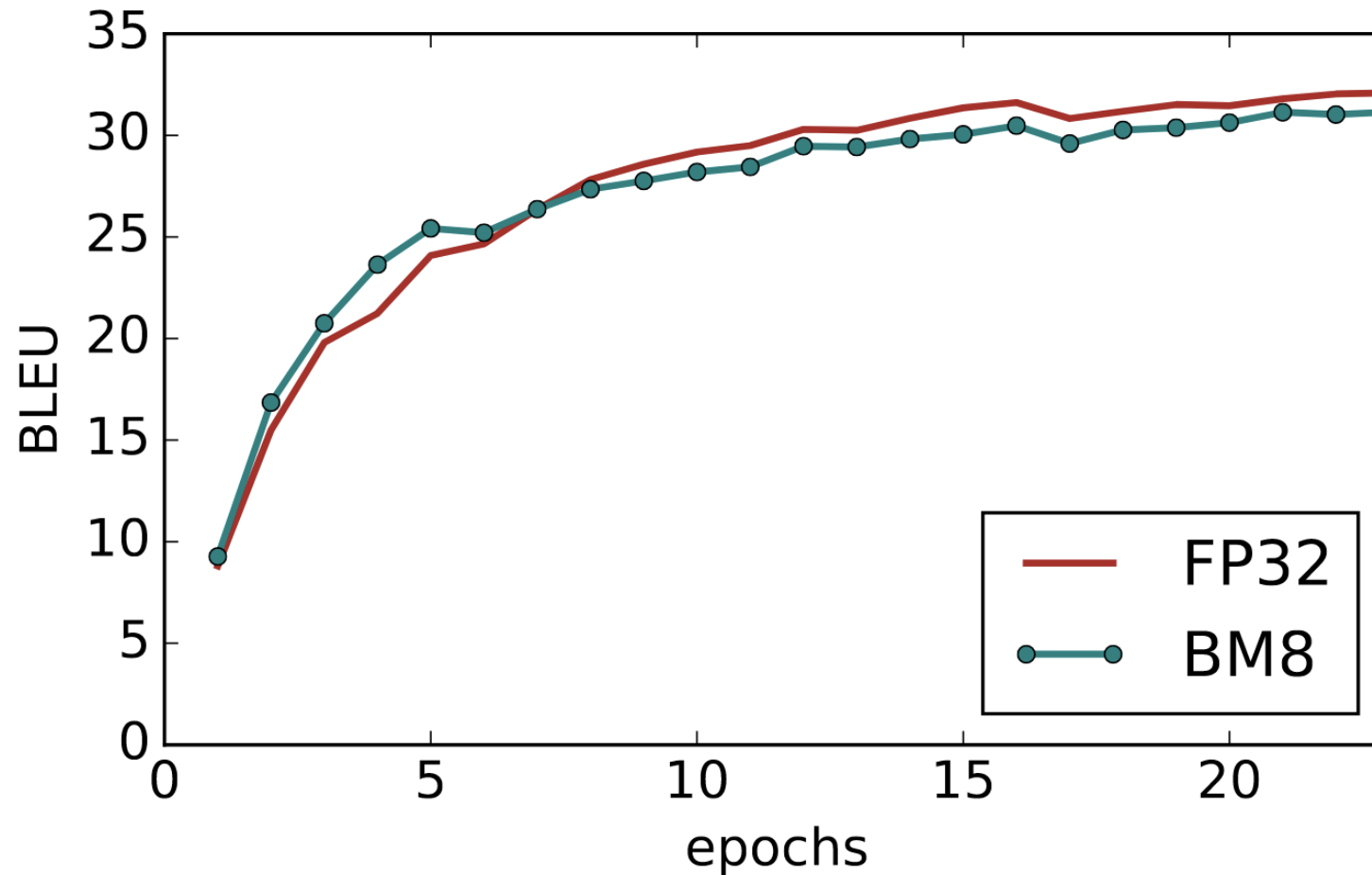
(c) Bwd weight grad. and update

- Weight, activation and gradient tensors quantized to BM with stochastic rounding
- Kulisch accumulator ensures our dot products are exact (can use FP CUDA lib directly)
- FP32 used for Kulisch to floating-point conversion, block minifloat alignments, quantization etc.
- Approx 1x floating point operation every N MACs, 5x slowdown



Scheme	BFP (ours)	BM (ours)	∇
6-bit	67.0	69.0	+2.0
8-bit	69.2	69.8	+0.6

ResNet18 on ImageNet Validation



Transformer on IWSLT'14 DE-En dataset

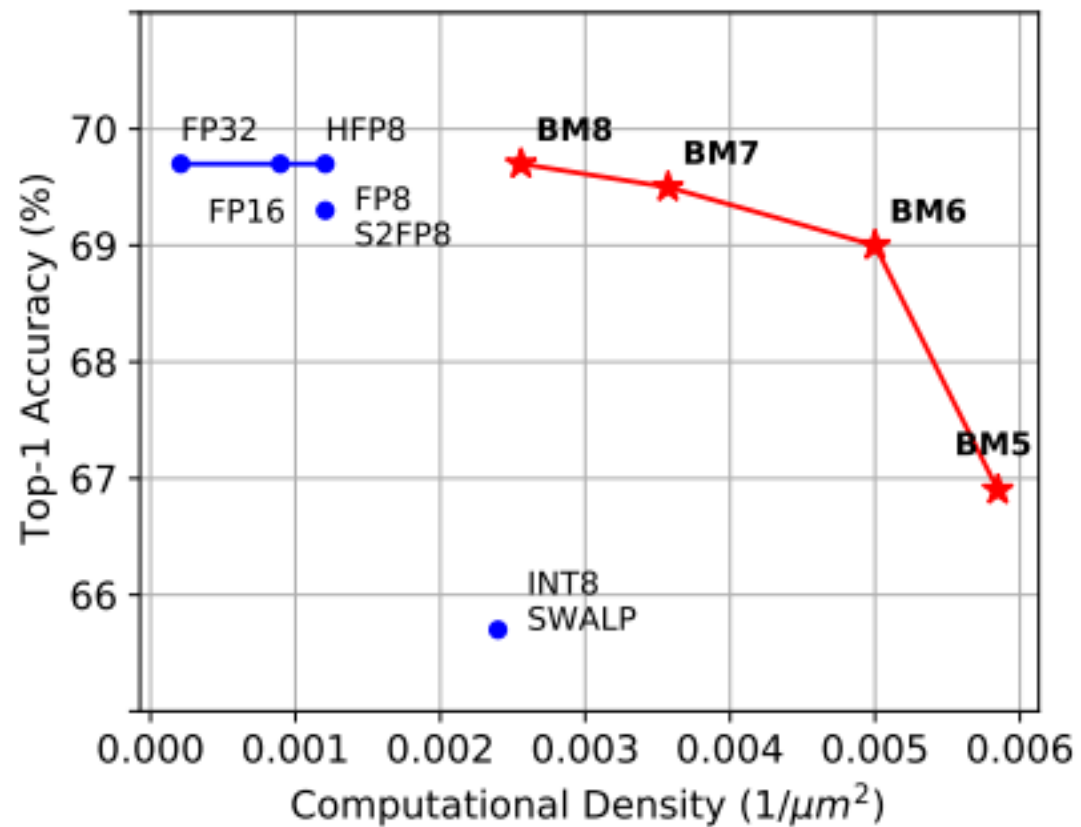
Model (Dataset) [Metric]	FP32	BM8
AlexNet (ImageNet)	56.0	56.2
EfficientNet-b0 (small ImageNet)	62.6	61.8
LSTM (PTB)[Val ppl.]	84.7	87.33
Transformer-base (IWSLT)[BLEU]	32.3	31.8
SSD-Lite (MbNetV2) (VOC)[mAP]	68.6	68.0

**Training Accuracy
with BM \approx FP32**

- Designs synthesized at 750MHz with Cadence RTL Compiler and 28nm cell library
 - Fused multiply-add (FMA)
 - 4x4 systolic matrix multipliers

Component	Area (μm^2)	Power (μW)
FP32	4782	10051
FP8 (w/ FP16 add)	829	1429
INT8 (w/ INT32 add)	417	1269
BM8	391	1141
BM6	200	624
INT8 (4x4 systolic)	7005	20253
FP8 (4x4 systolic)	18201	56202
BM8 (4x4 systolic)	6976	18765

BM8 area and power comparable to INT8



BM units are:

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

Model: ResNet-18
Dataset: ImageNet

Time Series Prediction

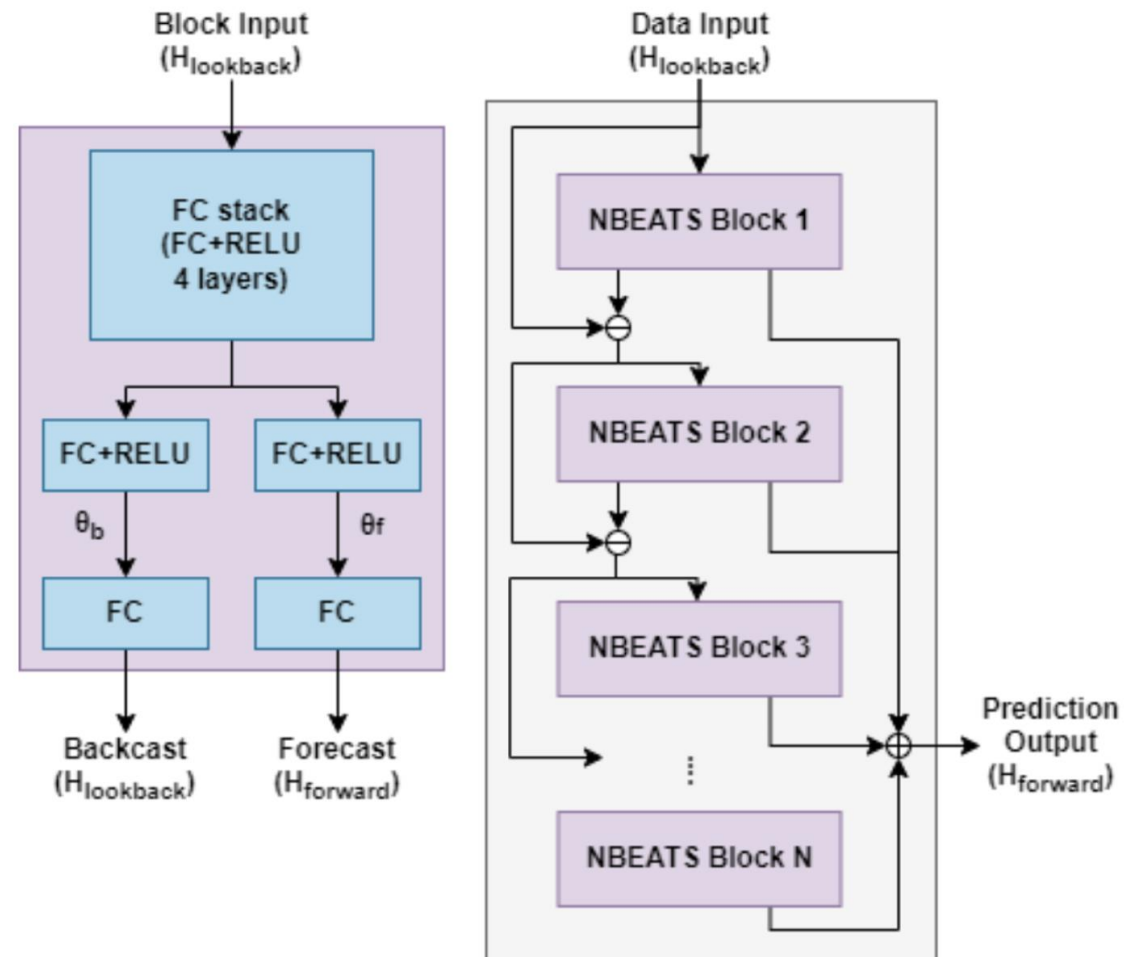
Wenjie Zhou

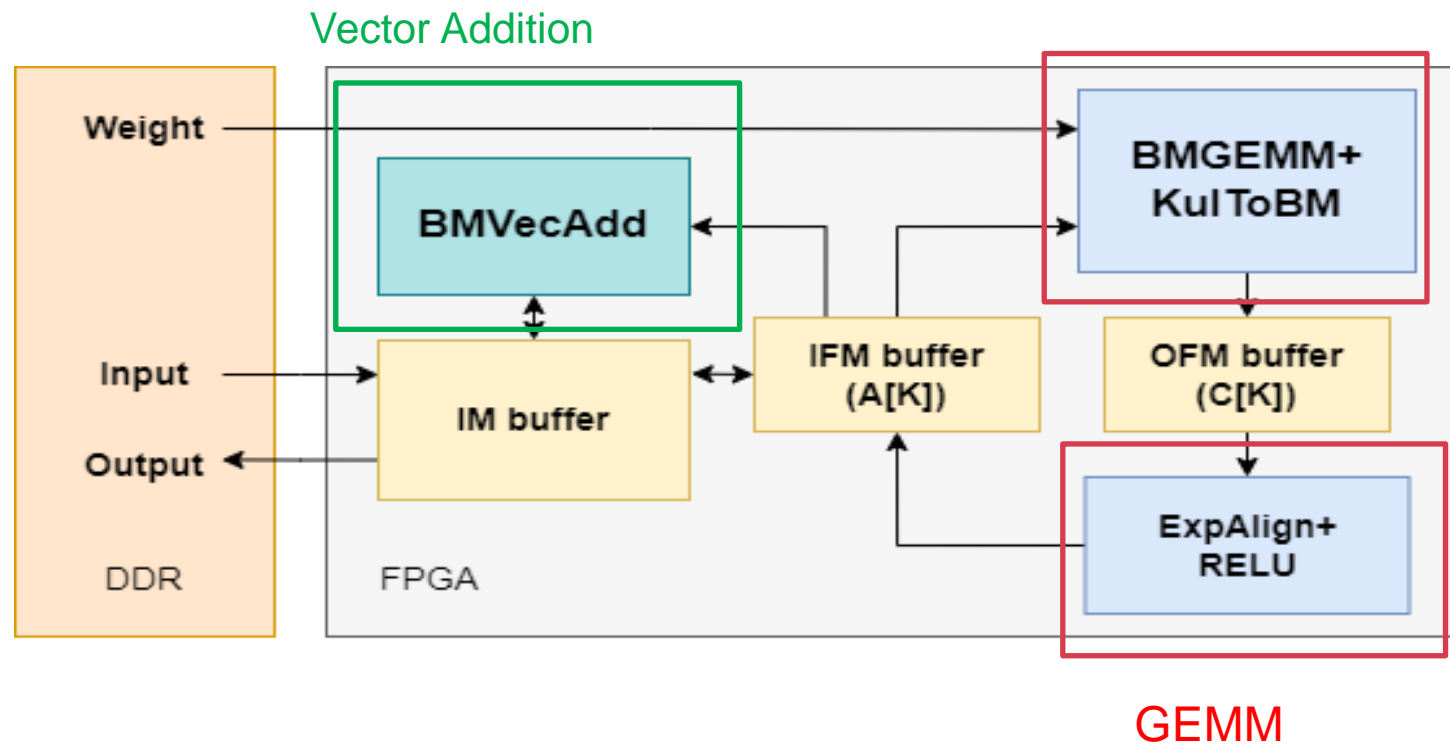


THE UNIVERSITY OF
SYDNEY

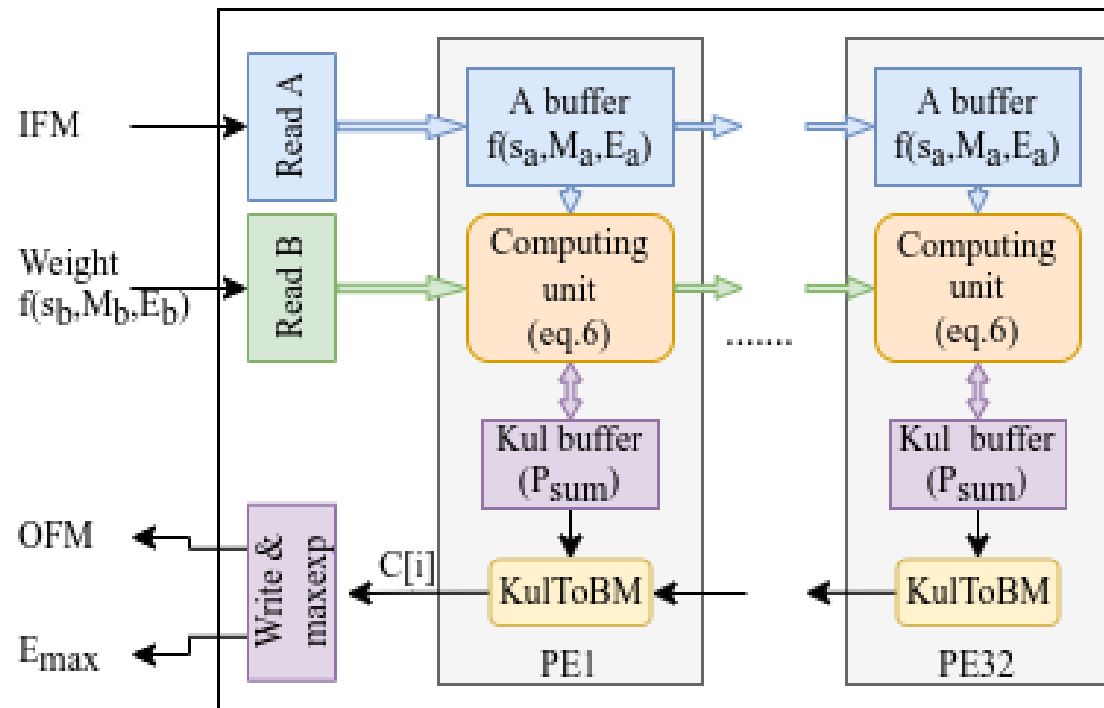
- › Previous work used GPU implementations with 28nm ASIC study
- › Here we explore FPGA implementation
 - NBEATS Inference and Training implementation using 4-bit mixed-precision BM
 - BM GEMM array and Training accelerator architecture for NBEATS

- › N-beats: Neural basis expansion analysis for interpretable time series forecasting. ICLR, 2019
- › Achieves state of the art time series prediction results
- › NN comprises mainly FC layers with shortcut connections

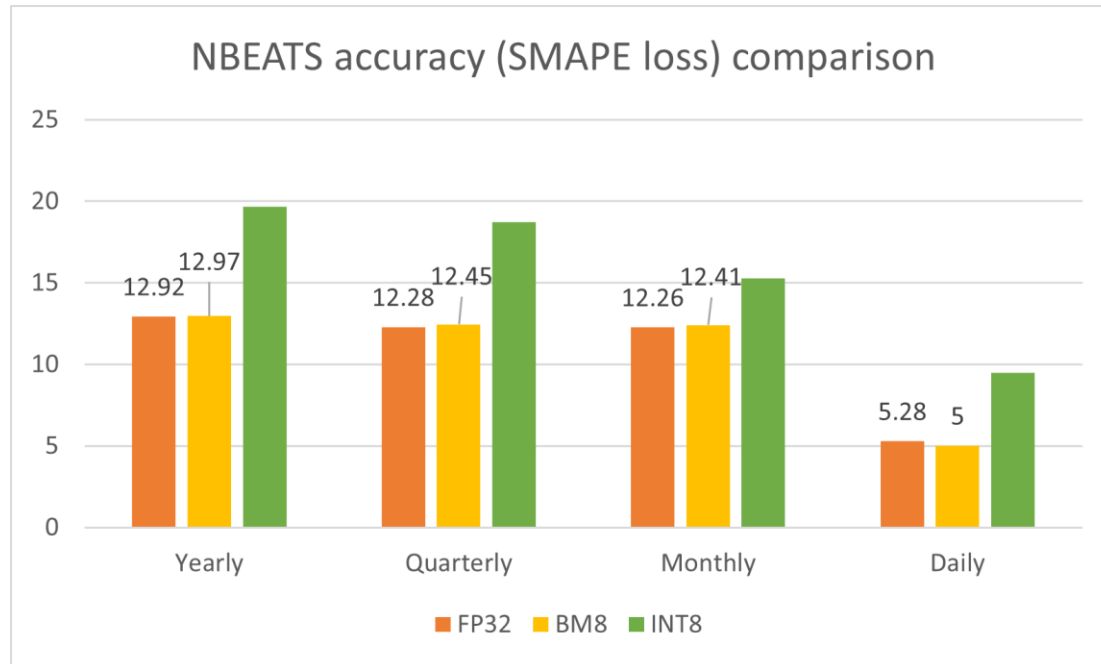




- › Each PE performs multiplication and Kulisch accumulation
- › Intermediate results are stored in the Kul buffer
- › Result transformed to a BM format



M4 competition dataset



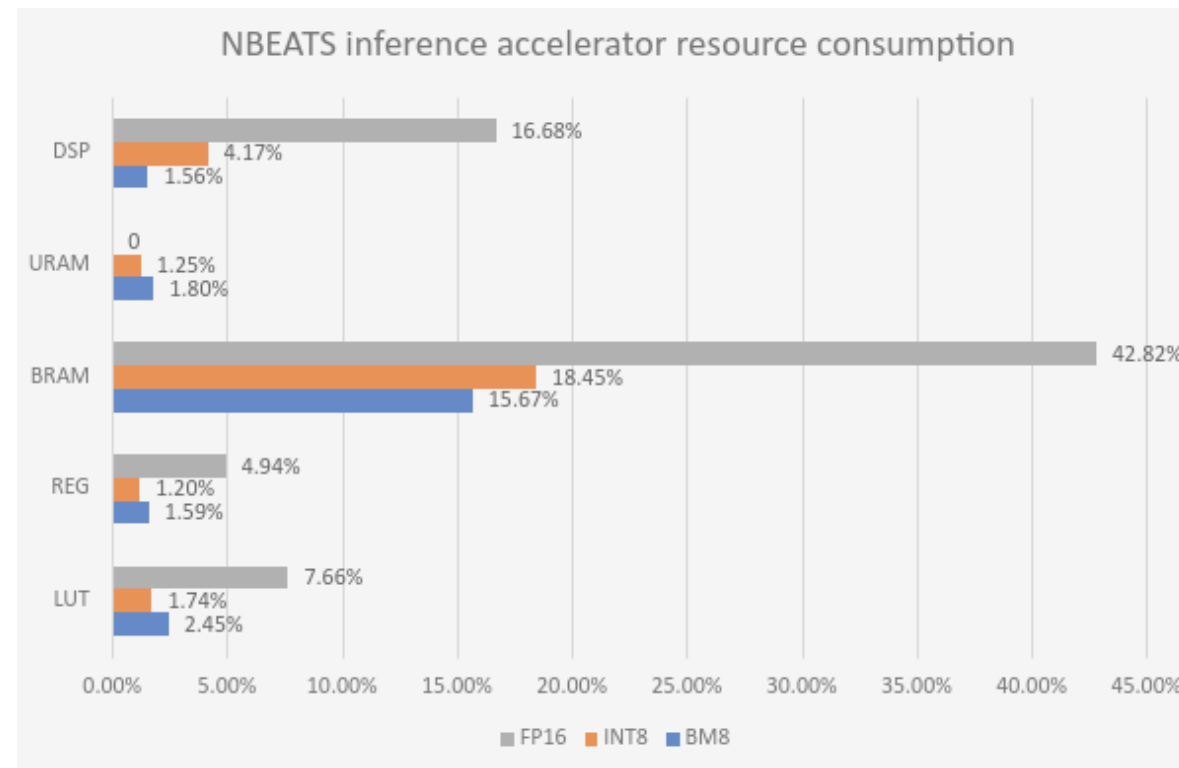
Accuracy of BM8 is similar to FP32

Benchmark	M4 dataset
Dataset	Yearly, Quarterly, Monthly, Daily
Training Loss	mean absolute percentage error(MAPE)
Validation Loss	symmetric mean absolute percentage error (sMAPE)
Batch size	1024

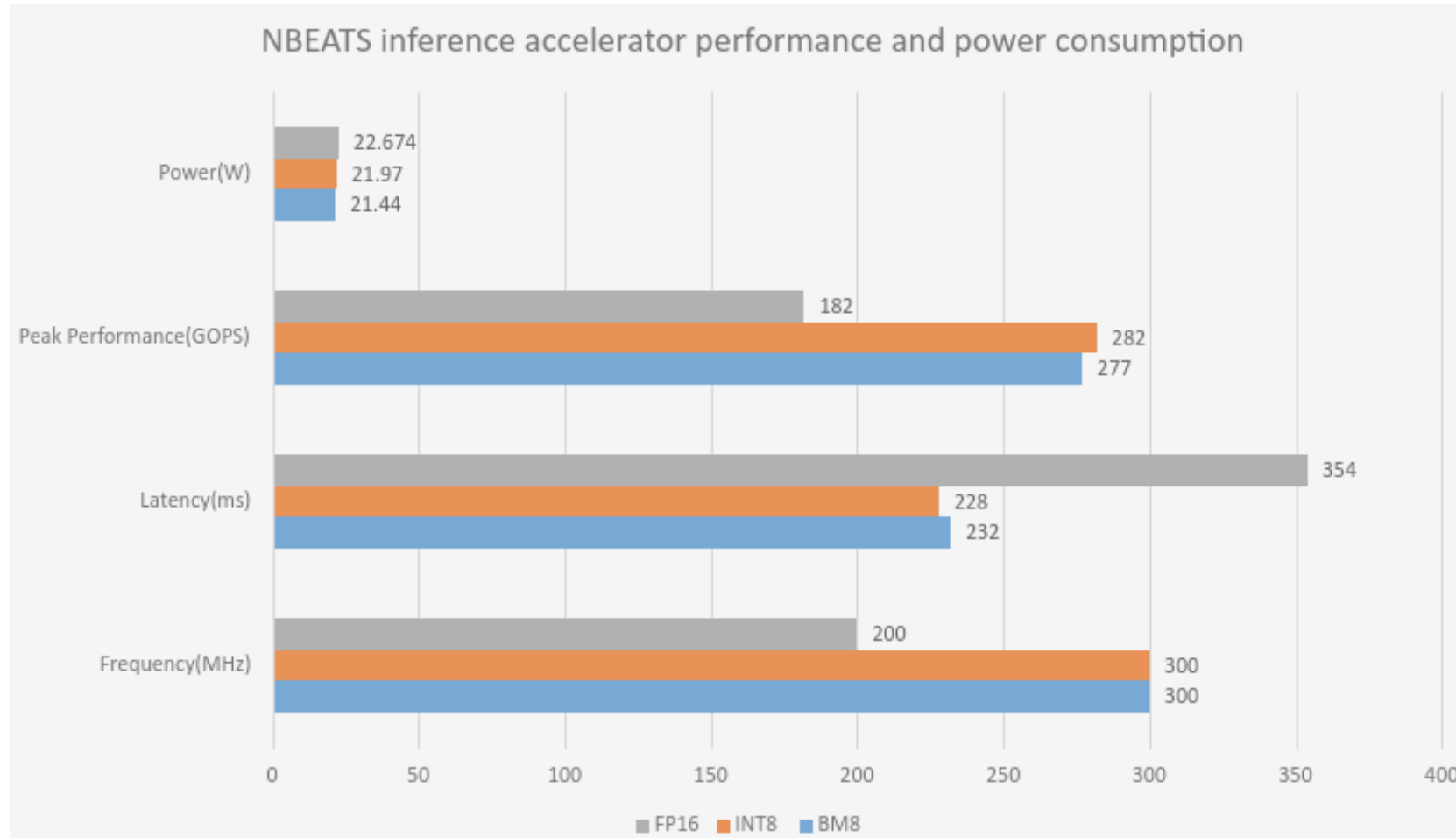
$$MAPE = \frac{1}{H} \sum_{i=1}^H \frac{|l_i - p_i|}{|l_i|} \quad (5)$$

$$SMAPE = \frac{200}{H} \sum_{i=1}^H \frac{|l_i - p_i|}{|l_i| + |p_i|}$$

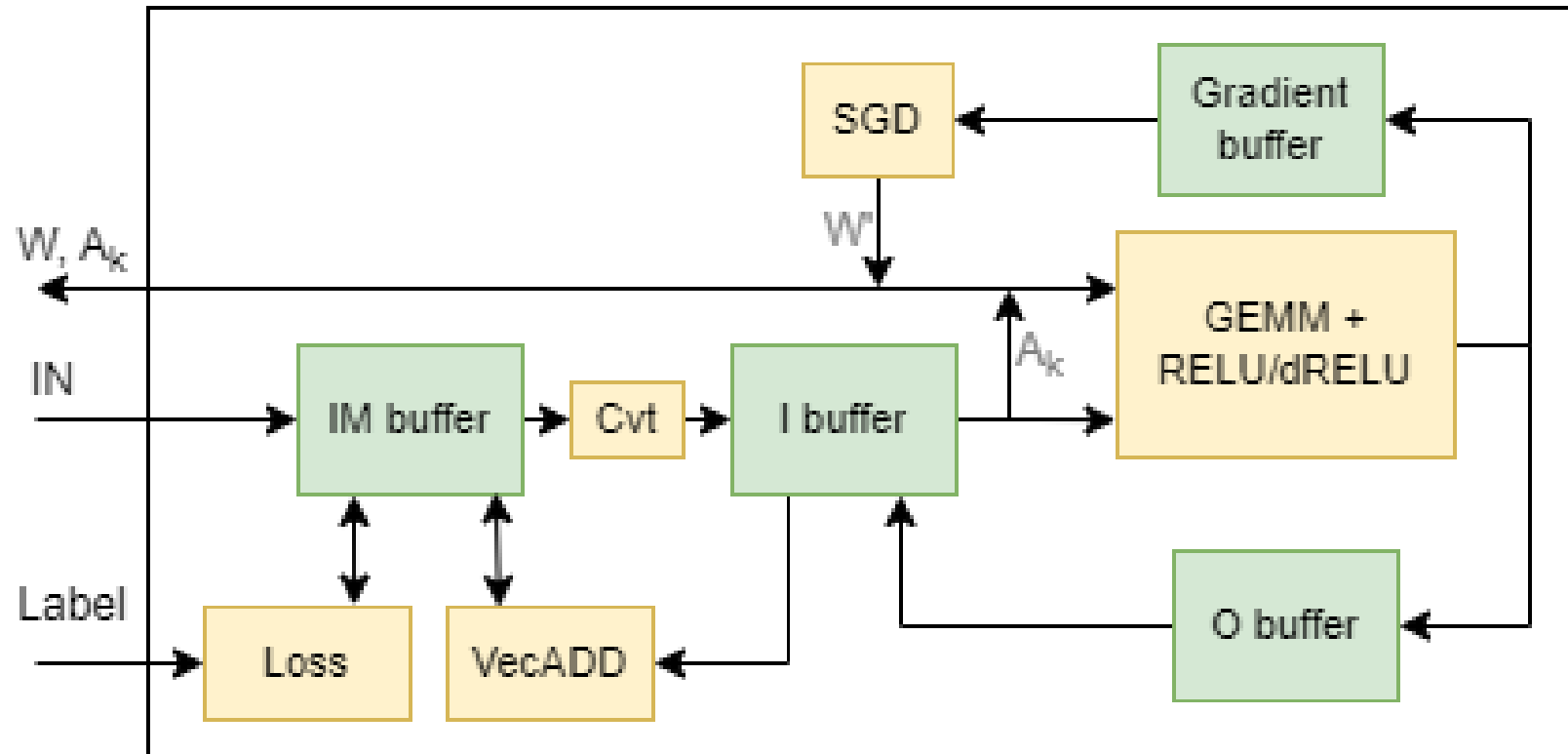
where l_i is the label in time step i , and p_i is the prediction in time step i .

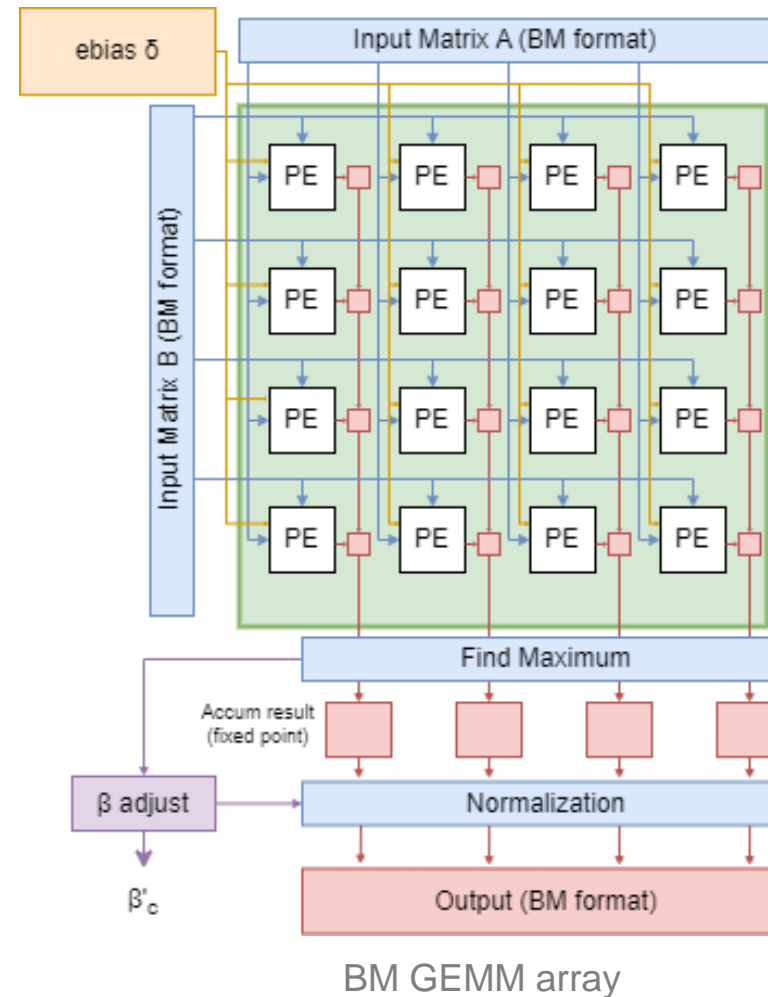
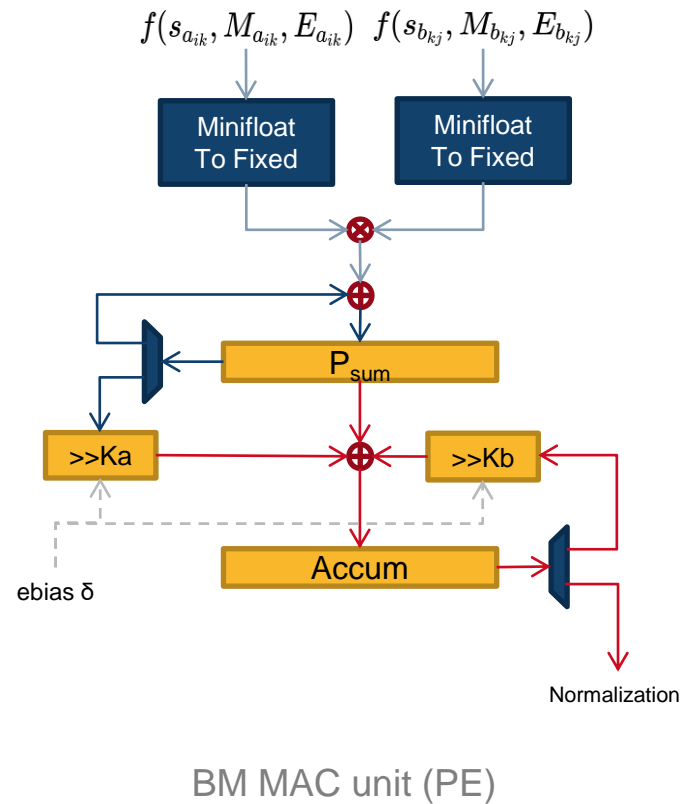


Area of BM8 is similar to INT8 but smaller than FP16



BM8 performance and power is close to INT8





› Dataset: M4-Yearly, validation loss: SMAPE loss, block size: 64

	Loss	Configuration			
		weight	activation	error	gradient
BM4(1)	14.471649	BM<2,1>	unsigned BM<0,4>	BM<0,3>	BM<0,3>
BM4(2)	14.463654	BM<2,1>	unsigned BM<0,4>	BM<0,3>	FP32
BFP8	12.914178	BM<0,7>	BM<0,7>	BM<0,7>	BM<0,7>
BM8	12.939716	BM<2,5>	BM<2,5>	BM<0,7>	BM<0,7>
FP32	12.924581				

Transfer Learning

Chuliang Guo



THE UNIVERSITY OF
SYDNEY

Why might we want to do transfer learning at the Edge?

› Private and secure

- No personal information uploaded to cloud

› Adapt to changing conditions

- To deal with non-stationary data

› Size, weight, and power (SWaP)

- Converge to a good solution faster through pretraining

- › Back-propagation using SGD
 - 3X workload of inference

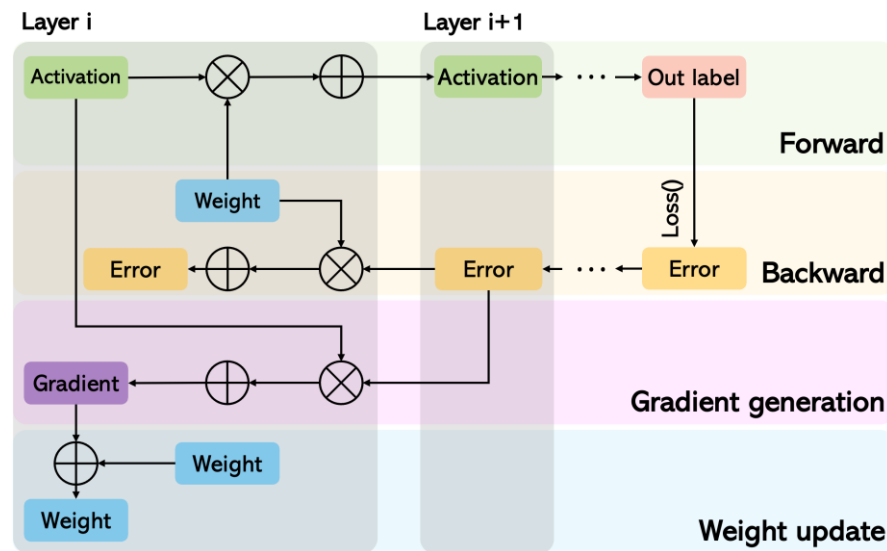


Fig. 1 CNN training workflow: (1) Conv in forward path, (2) transposed Conv in backward path, (3) dilated Conv in gradient generation, and (4) weight update.

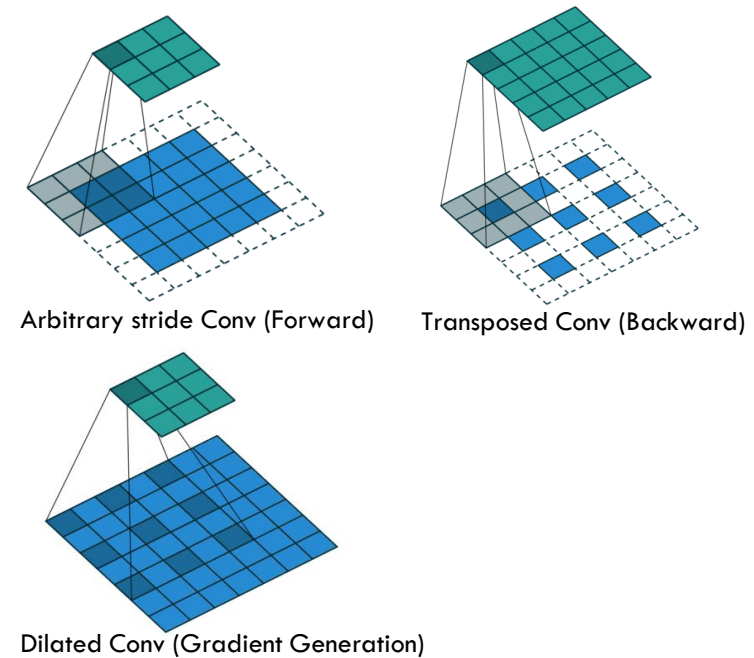


Fig. 2 Non-unit stride Conv, transposed Conv, and dilated Conv [1].

- › Layer-wise CNN blocks
 - Unified $bm(2,5)$ representation
 - Non-unit stride Conv support
 - Simplified mult/add/MAC
 - Fused BN&ReLU

- › Main blocks
 - Unified Conv
 - Conv & transposed Conv
 - Dilated Conv
 - Weight kernel partition

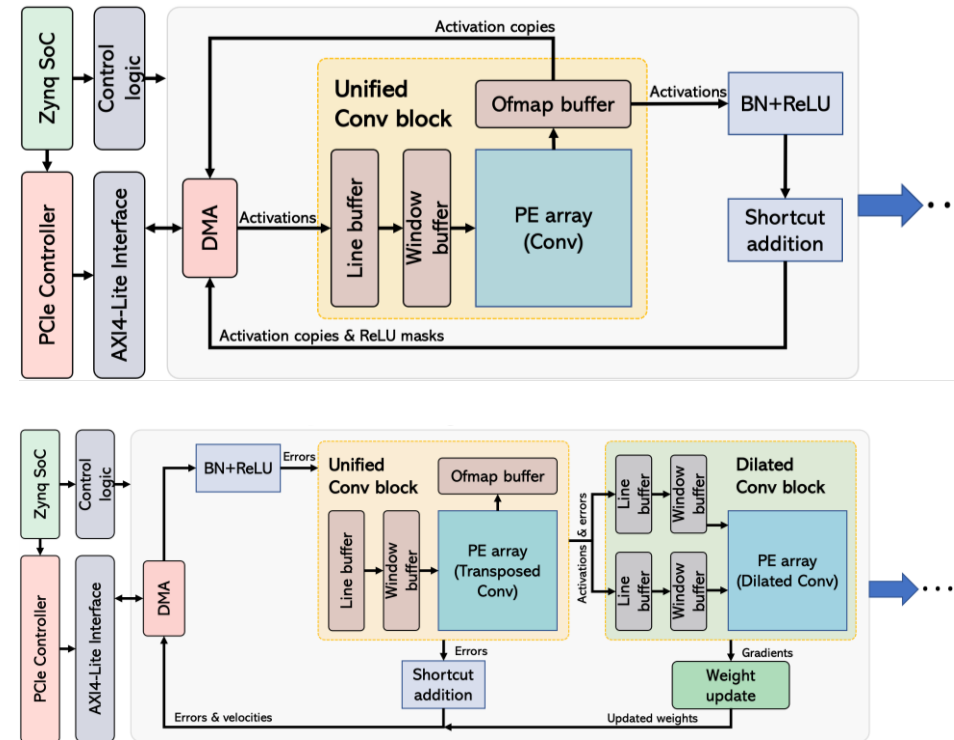
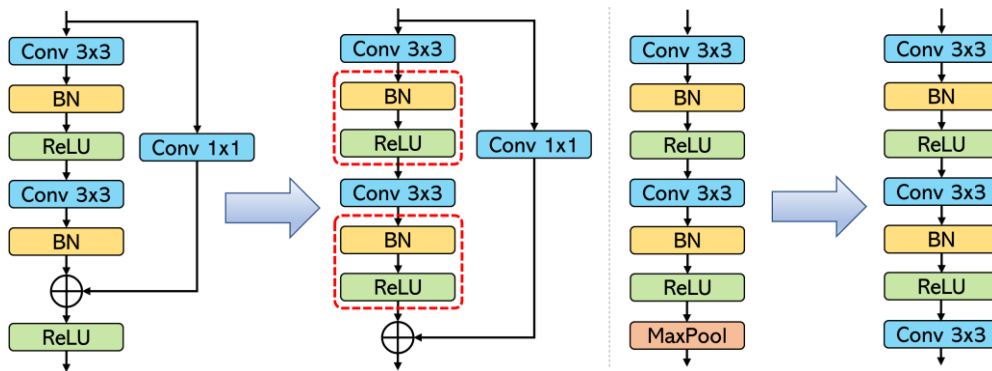


Fig. 3 Overall architecture of the generic training accelerator for layer-by-layer processing. BN and ReLU are fused.

CIFAR-10 Training from Scratch

- › Shortcut addition after BN and ReLU functions (enabling fusing)
- › Unified bm(2,5) for activations, weights, errors, and gradients (simpler HW)
- › Full precision accuracy with these changes

Fig. 4 Modifications to basic building block of ResNet20 and VGG-like.



Tab. 1 Top-1 accuracy on CIFAR-10 and SVHN.

Model	Precision (FP/BP)	CIFAR-10 Acc	SVHN Acc
VGG-like	FP32	86.64%	92.45%
	BFP8	85.65%	92.07%
	bm(2,5)/bm(4,3)	86.52%	92.51%
	bm(2,5)	86.54%	92.55%
ResNet20	FP32	90.27%	94.98%
	BFP8	87.52%	90.37%
	bm(2,5)/bm(4,3)	89.46%	95.51%
	bm(2,5)	89.87%	95.60%

– Channel tiling accelerator

– Updating last several Conv & FC

- Shortened back-propagation
- Reduced BRAM for activations
- Faster convergence

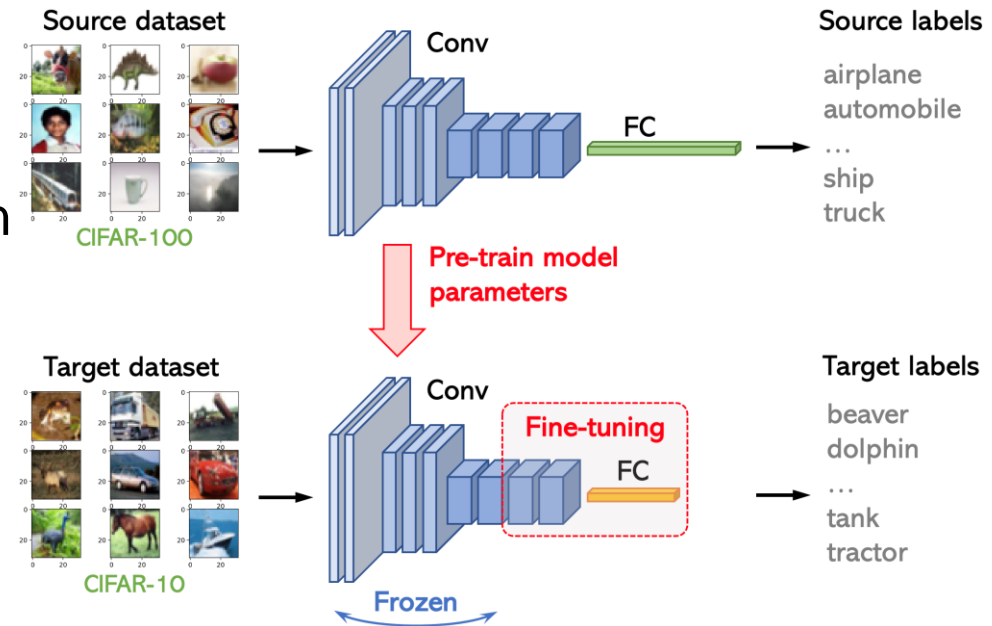
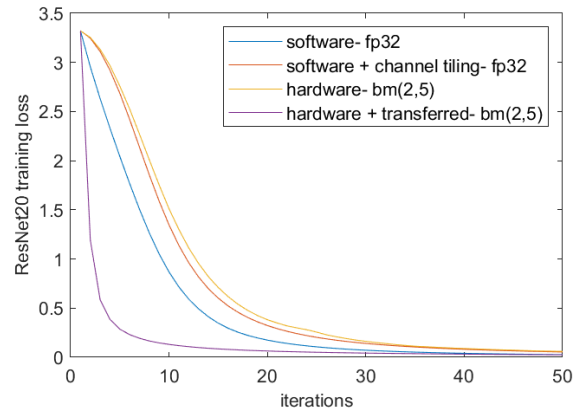


Fig. 8 Transfer learning example from CIFAR-100 to CIFAR-10.

TABLE III
RESOURCE UTILISATION OF AND POWER THE RESNET20 ACCELERATOR
(WITH THE STATIC POWER OF 30W).

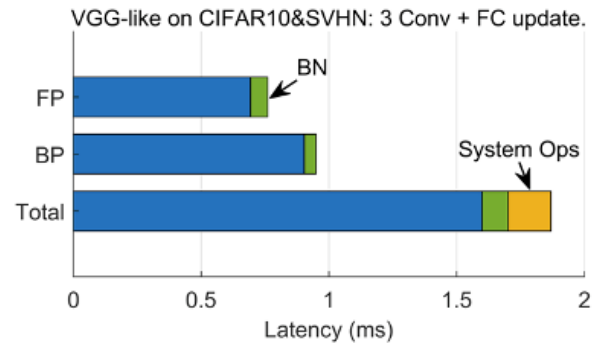
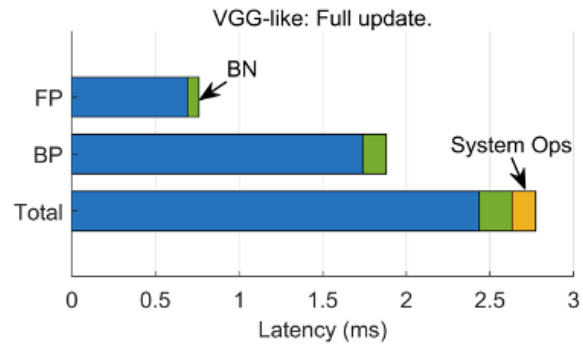
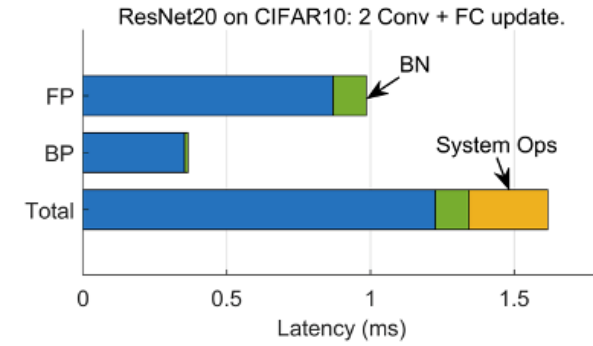
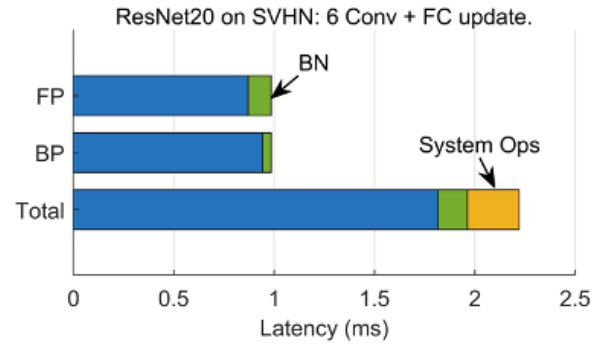
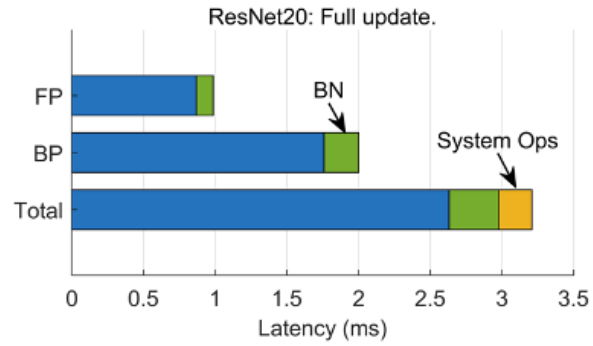
	CLB	LUT	DSP	BRAM	Vivado(W)	PPS(W)
Full update	28824	166502	686	1171	8.714	35
6 Conv+FC	25589	161129	685	671	7.725	34
2 Conv+FC	21340	129453	621	571	6.779	34

TABLE IV
RESOURCE UTILISATION AND POWER OF THE VGG-LIKE ACCELERATOR
(WITH THE STATIC POWER OF 30W).

	CLB	LUT	DSP	BRAM	Vivado(W)	PPS(W)
Full update	20688	119086	614	505	6.824	34
3 Conv+FC	20489	119740	613	325	6.499	34



Latency Breakdown



Conclusion

2

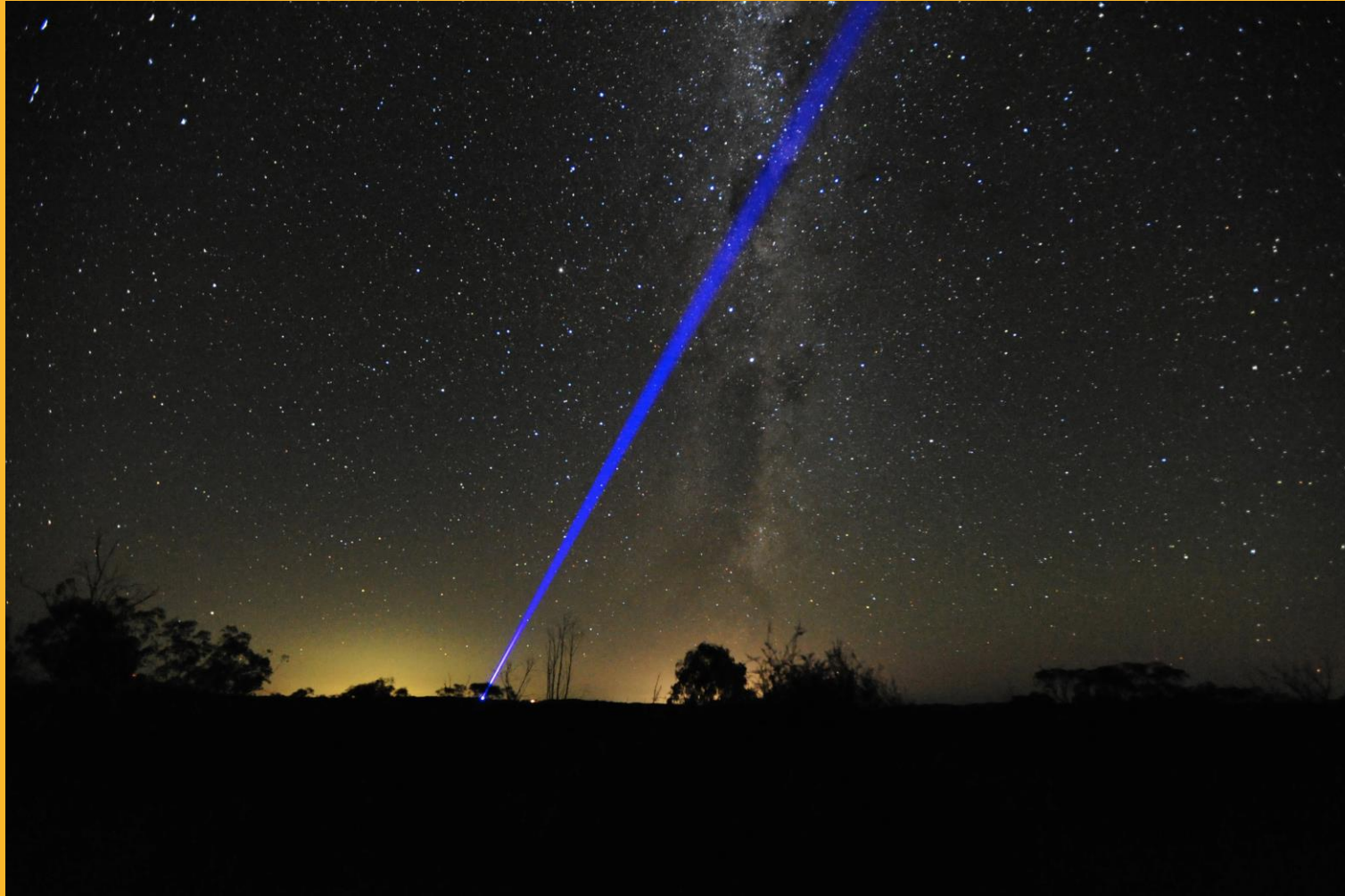


THE UNIVERSITY OF
SYDNEY

- Low-precision formats have wide applicability for inference and training in Edge applications
 - Doesn't necessitate accuracy reduction
- Faster Training is possible using BM
 - Fewer bits – important for memory-bound
 - Narrow exponents – denser MAC in compute-bound

What are the applications?

- [1] 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. URL: [bm_iclr21.pdf](#).
- [2] Wenjie Zhou, Haoyan Qi, David Boland, and Philip H.W. Leong. FPGA implementation of N-BEATS for time series forecasting using block minifloat arithmetic. In *Proc. Asia Pacific Conference on Circuits and Systems (IEEE APCCAS 2022)*. 2022. URL: [nbeats_apccas22.pdf](#).



Philip Leong (philip.leong@sydney.edu.au)
<http://phwl.org/talks>



Copyright Notice

This multimedia file is copyright © 2023 by tinyML Foundation. All rights reserved. It may not be duplicated or distributed in any form without prior written approval.

tinyML[®] is a registered trademark of the tinyML Foundation.

www.tinyml.org



Copyright Notice

This presentation in this publication was presented as a tinyML® Talks webcast. The content reflects the opinion of the author(s) and their respective companies. The inclusion of presentations in this publication does not constitute an endorsement by tinyML Foundation or the sponsors.

There is no copyright protection claimed by this publication. However, each presentation is the work of the authors and their respective companies and may contain copyrighted material. As such, it is strongly encouraged that any use reflect proper acknowledgement to the appropriate source. Any questions regarding the use of any materials presented should be directed to the author(s) or their companies.

tinyML is a registered trademark of the tinyML Foundation.

www.tinyml.org