tinyML® Research Symposium

Enabling Ultra-low Power Machine Learning at the Edge

March 27, 2023

www.tinyML.org

Memory-Oriented Design-Space Exploration of Edge-AI Hardware for XR Applications

Vivek Parmar[1], Syed Shakib Sarwar[2], Ziyun Li[2], Hsien-Hsin S. Lee[2], Barbara De Salvo[2†], and **Manan Suri[1]**

manansuri@ee.iitd.ac.in

[1]Indian Institute of Technology Delhi

[2]Meta Reality Labs Research

© March 2023, NVM and Neuromorphic Research Group-IITD
and Meta Reality Labs, tinyML Research Symposium 2023

Motivation & Scope

Demonstrate benefits of memory-centric computing utilizing advanced NVM technology for XR-EAI applications

- Exploit normally-off computing due to nature of workload
- Analyze memory & power budgets for hybrid architectures through DTCO
- Estimates/Projections at multiple nodes and type of NVM devices
- Relevant Metrics

TABLE I
PROJECTED SPECS OF STATE-OF-THE-ART XR DEVICES [1].

Metric	HTC Vive Pro	Ideal VR	Microsoft HoloLens2	Ideal AR
Resolution (MP)	4.6	200	4.4	200
Refresh rate (Hz)	90	90-144	120	90-144
Motion-to-photon latency (ms)	<20	<20	<9	<5
Power (W)	N/A	1-2	>7	0.1-0.2

1. M. Huzaifa, et.al., arXiv preprint arXiv:2004.04643 (2020).

© March 2023, NVM and Neuromorphic Research Group-IITD and Meta Reality Labs, tinyML Research Symposium 2023

XR-EAI Workloads Investigated

Network	#Params	Size (kB)
EDSNet	6.63 M	6474
Detnet	1.45 M	1414

1. Eye Segmentation

Dataset: OpenEDS 2019

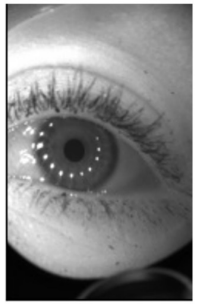
- Network: Unet (backbones: MobileNetv2)
- Framework: Tensorflow

2. Hand detection

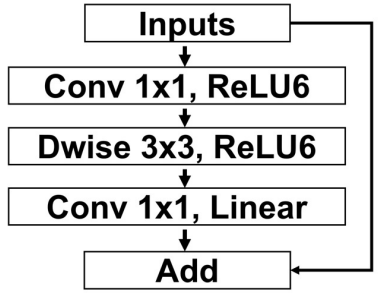
Dataset: FPHAB*

- Network: DetNet (MegaTrack)
- Framework: PyTorch

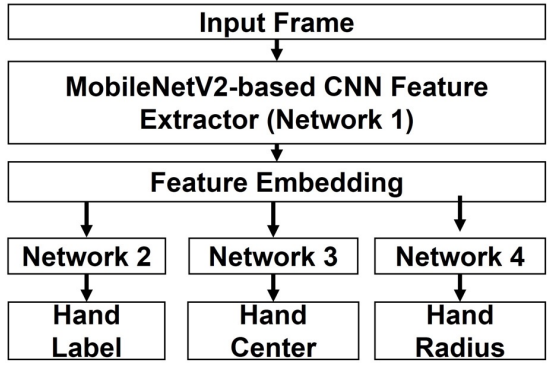
(a) FPHAB Dataset Sample



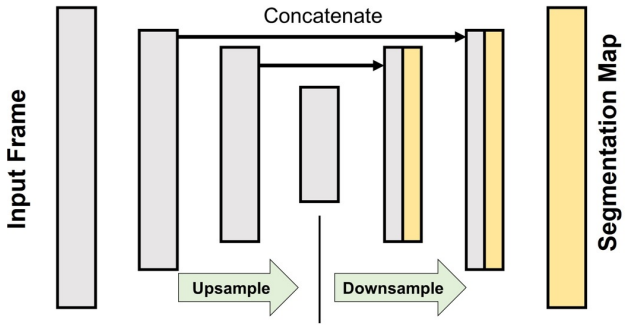
(b) OpenEDS Dataset Sample



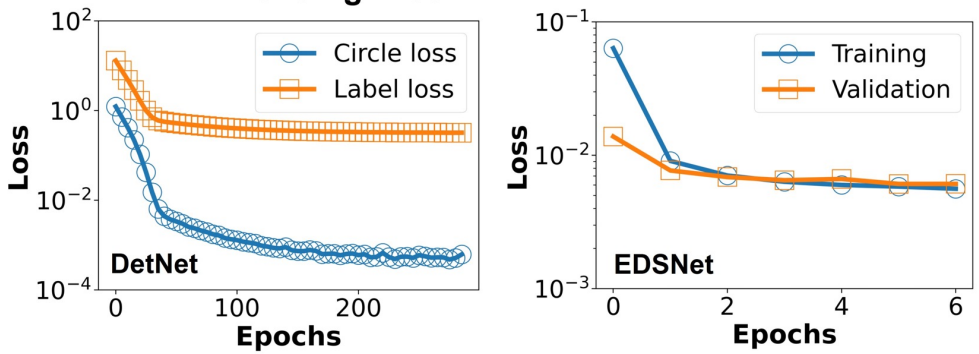
(c) MobileNetv2 Basic Building Block



(d) DetNet



(e) EDSNet (UNet + MobileNetV2)



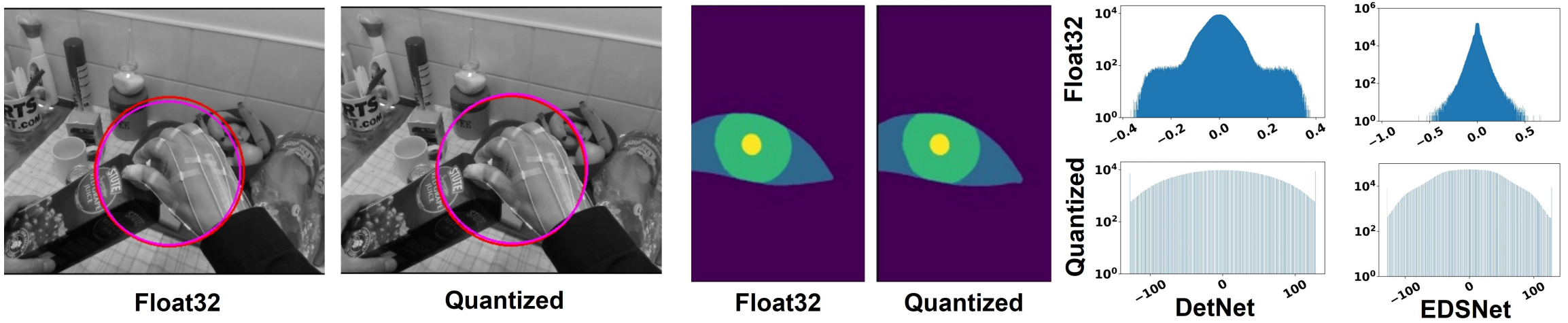
(f) Training Evolution

* Indian Institute of Technology Delhi obtained and used the FPHAB dataset

1. Garbin, Stephan J., et al. "Openeds: Open eye dataset." arXiv preprint arXiv:1905.03702 (2019).
2. Garcia-Hernando, Guillermo, et al. "First-person hand action benchmark with rgb-d videos and 3d hand pose annotations." *CVPR*. 2018.

© March 2023, NVM and Neuromorphic Research Group-IITD and Meta Reality Labs, tinyML Research Symposium 2023

XR-EAI Workloads: Impact of Quantization

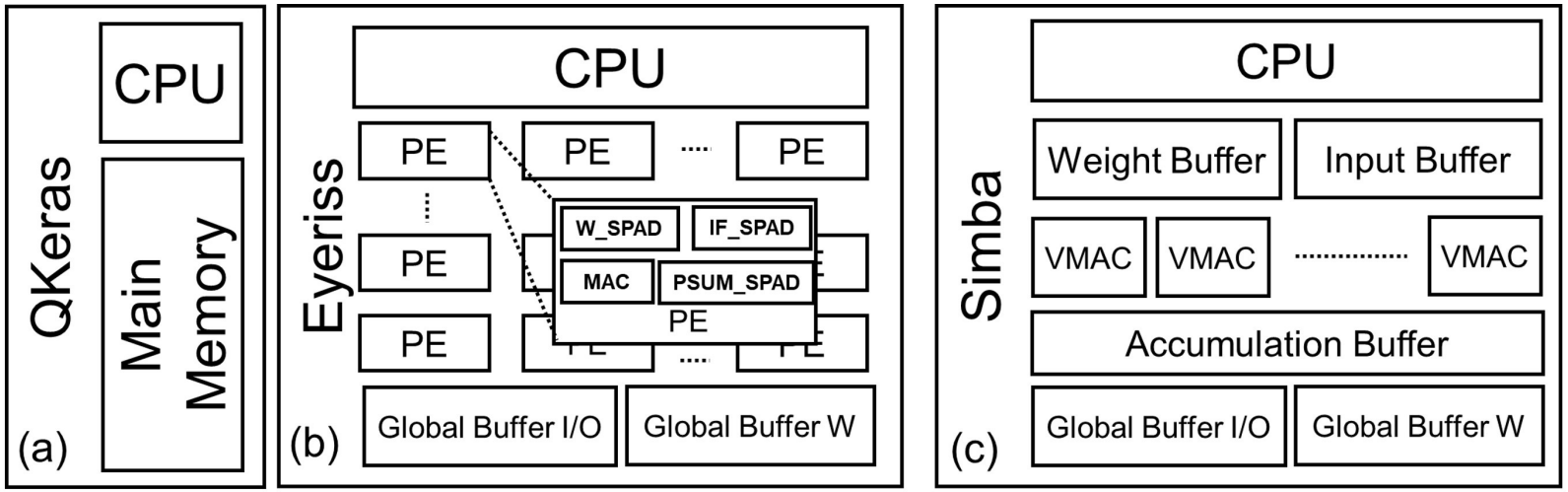


(a) DetNet Evaluation (b) EDSNet Evaluation (c) Weight Distribution

- Comparable performance between full-precision and quantized versions
- Weight distribution profile changes due to use of additional scaling factors specific to layers during quantization

© March 2023, NVM and Neuromorphic Research Group-IITD and Meta Reality Labs, tinyML Research Symposium 2023

Performance on CMOS-based Systolic Accelerators



Framework		Qkeras	Timeloop+Accelergy	
Platform		CPU	Eyeriss	Simba
PE Organization	Base	1	14×12 =168	16×16 =256
	V1		32×32 =1024	
	V2		64×64 =4096	
MAC Precision		int8		
Input buffer		16 MB × 1	12B × 168 (8)	64kB × 16 (64)
Output buffer			16B × 168 (8)	384B × 64 (24)
Weight buffer			192B × 168 (8)	4kB × 64 (64)
Global buffer (I/O)			8 MB (64)	8 MB (256)
Global buffer (W)			8 MB (64)	8 MB (256)

(d) Architectural Parameters

Workload	Platform	Energy Breakdown (%)	
		Compute	Memory
DetNet	CPU	44.90%	55.10%
	Eyeriss	3.90%	96.10%
	Simba	5.80%	94.20%
EDSNet	CPU	90.50%	9.50%
	Eyeriss	7.10%	92.90%
	Simba	9.70%	90.30%

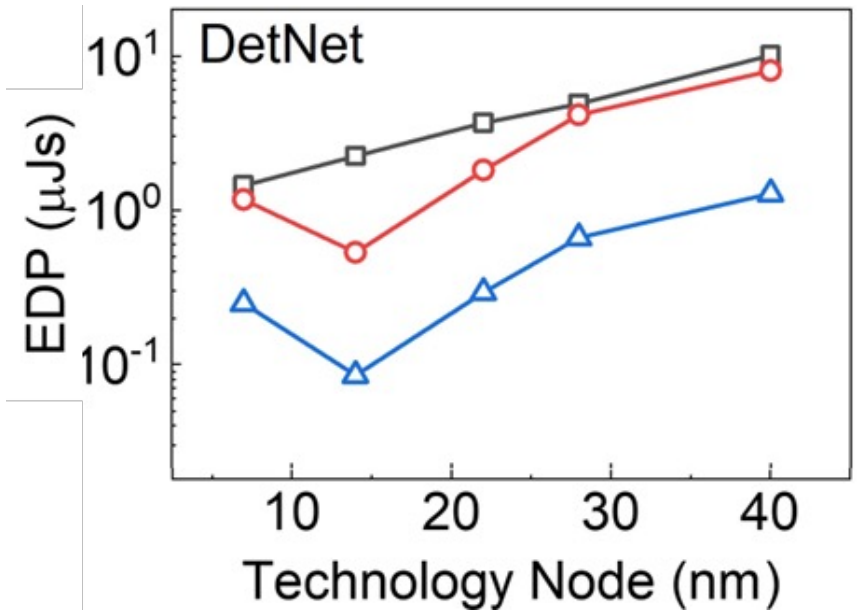
(e) Energy Contribution

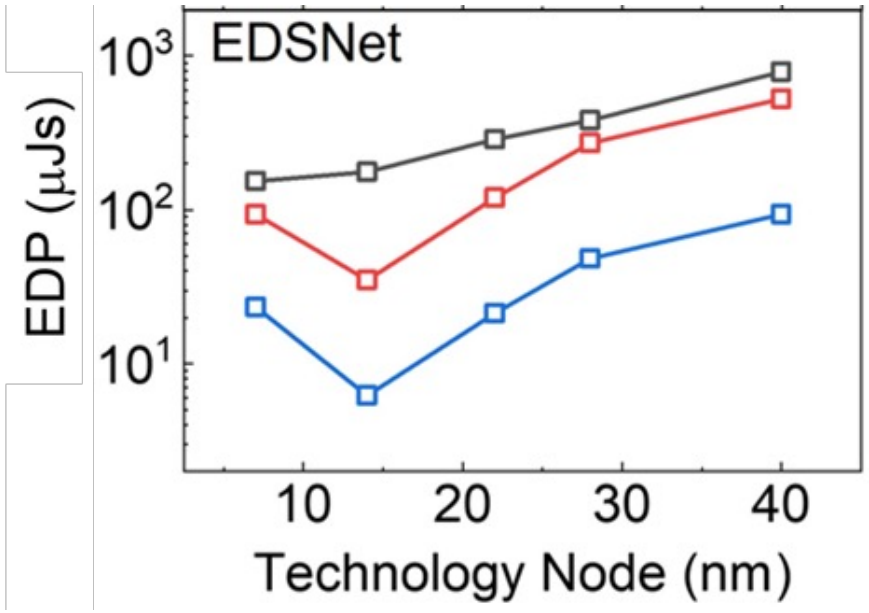
1. Coelho, Claudionor. "Google/QKeras. 2019." URL: https://github. com/google/qkeras (visited on 03/02/2022).
2. Parashar, Angshuman, et al. "Timeloop: A systematic approach to dnn accelerator evaluation." ISPASS. IEEE, 2019.
3. Wu, Yannan Nellie, et al. "Accelergy: An architecture-level energy estimation methodology for accelerator designs." ICCAD. IEEE, 2019.
4. Chen, Yu-Hsin, et al. "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks." IEEE JSSC, 52.1 (2016): 127-138.
5. Shao, Yakun Sophia, et al. "Simba: Scaling deep-learning inference with multi-chip-module-based architecture." Micro. 2019.

© March 2023, NVM and Neuromorphic Research Group-IITD and Meta Reality Labs, tinyML Research Symposium 2023

Performance on CMOS-based Systolic Accelerators

- Technology scaling based on DeepScale [1] for: 22 nm, 28 nm
- 7nm estimates based on TPUv4 [2] scaling factors
- **Benefits of scaling diminishing at 7nm**





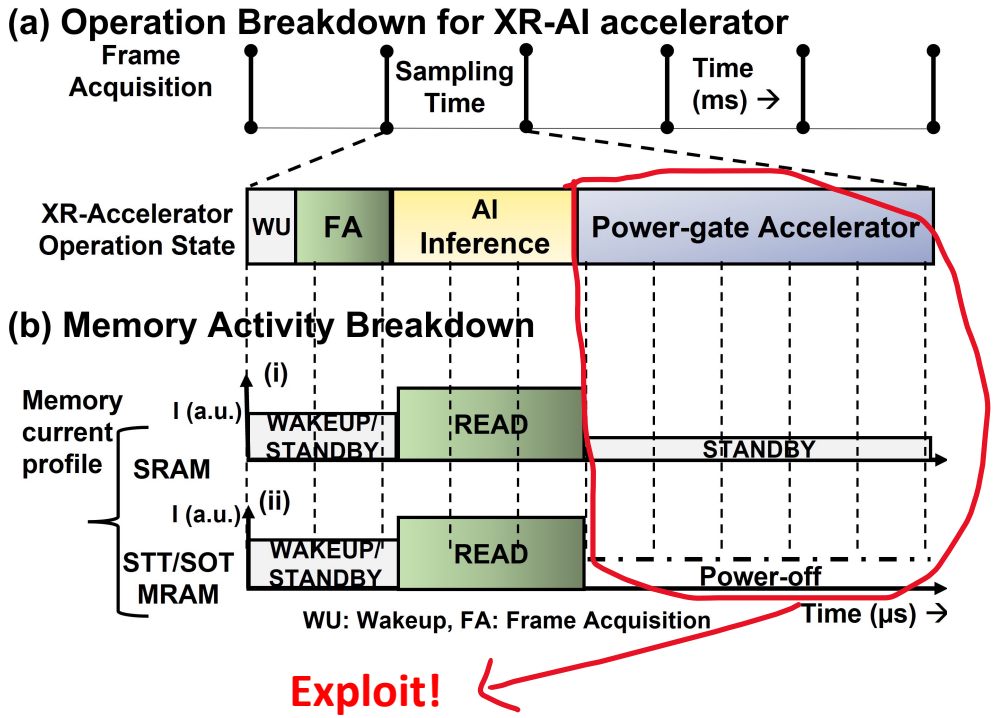
Legend

CPU: Black
Eyeriss: Red
Simba: Blue

1. S. Sarangi and B. Baas, "DeepScaleTool: A Tool for the Accurate Estimation of Technology Scaling in the Deep-Submicron Era," *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2021, pp. 1-5, doi: 10.1109/ISCAS51556.2021.9401196.
2. N. P. Jouppi *et al*., "Ten Lessons From Three Generations Shaped Google's TPUv4i : Industrial Product," *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*, 2021, pp. 1-14, doi: 10.1109/ISCA52012.2021.00010.

© March 2023, NVM and Neuromorphic Research Group-IITD and Meta Reality Labs, tinyML Research Symposium 2023

Proposed NVM-based Enhancements



(c) AI Inference - Memory Operation Breakdown

Read inputs (R)	Read Weights R / (R+W)	Compute (R+W)	Write output (W)

(i) Traditional Memory Mapping (baseline CPU, Eyeriss, Simba)

DRAM (1st layer) / SRAM	DRAM / SRAM (load from DRAM)	Registers /SRAM	SRAM

(ii) Proposed P0 Mapping (NVM for weight matrix)

SRAM	MRAM	Registers /SRAM	SRAM

(iii) Proposed P1 Mapping (NVM in all buffers)

MRAM	MRAM	Registers /SRAM	MRAM

Two flavours explored

1. **P0: MRAM for only weights**
2. **P1: MRAM everywhere except compute registers**

© March 2023, NVM and Neuromorphic Research Group-IITD and Meta Reality Labs, tinyML Research Symposium 2023

Performance Analysis for Proposed NVM-enhanced variants

Direct Area saving in all variants

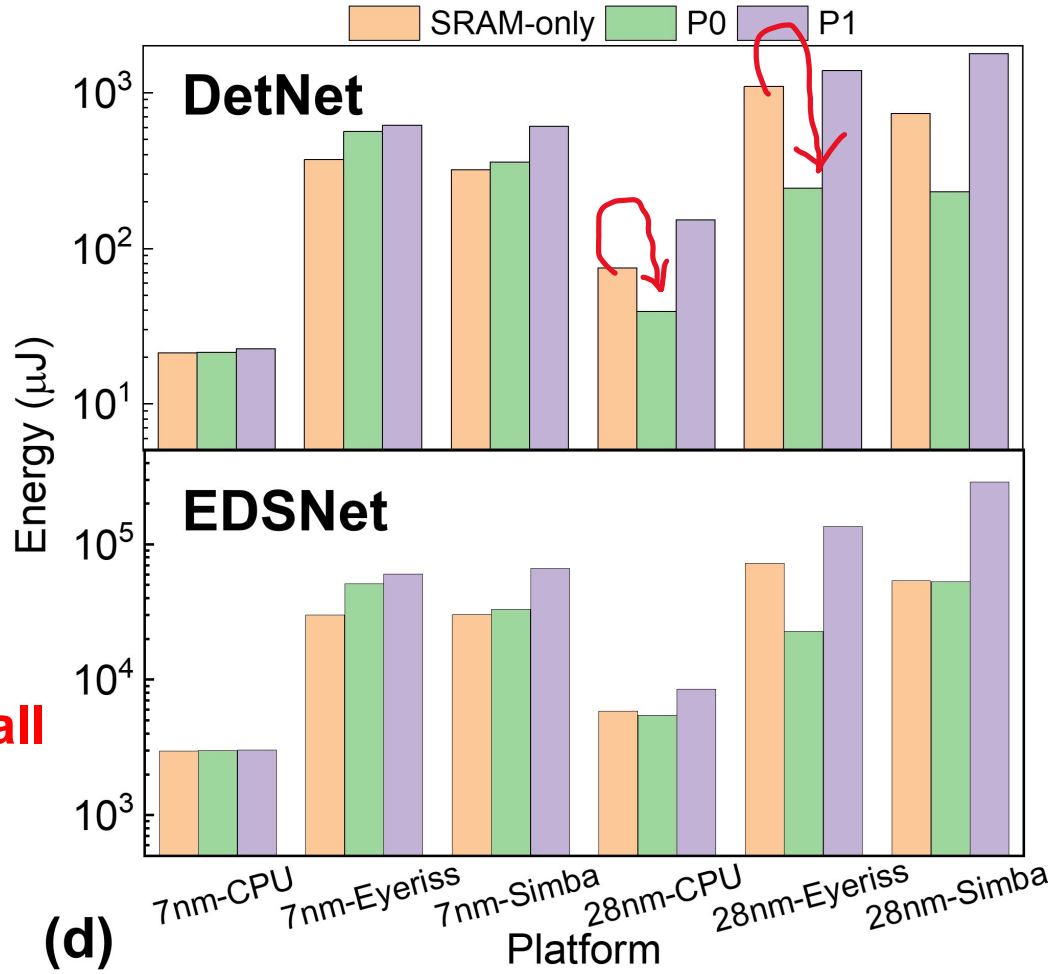
TABLE II
ESTIMATION OF AREA BENEFITS ON SYSTOLIC ACCELERATORS USING PROPOSED P0 AND P1 VARIANTS AT 7NM NODE.

Architecture	7 nm Area (mm^2)			Area savings	
	SRAM-only	P0	P1	P0	P1
Simba	2.89	2.41	1.88	16.56%	34.97%
Eyeriss	2.56	2.11	1.67	17.52%	34.98%

28nm P0 savings

- **DetNet**: ~50% with CPU, ~80% with Eyeriss, ~70% with Simba
- **EDSNet**: ~ 7% with CPU, ~70% with Eyeriss, ~1% with Simba

Energy saving evident in some variants (28nm-P0, all applications) w.r.t SRAM only variants

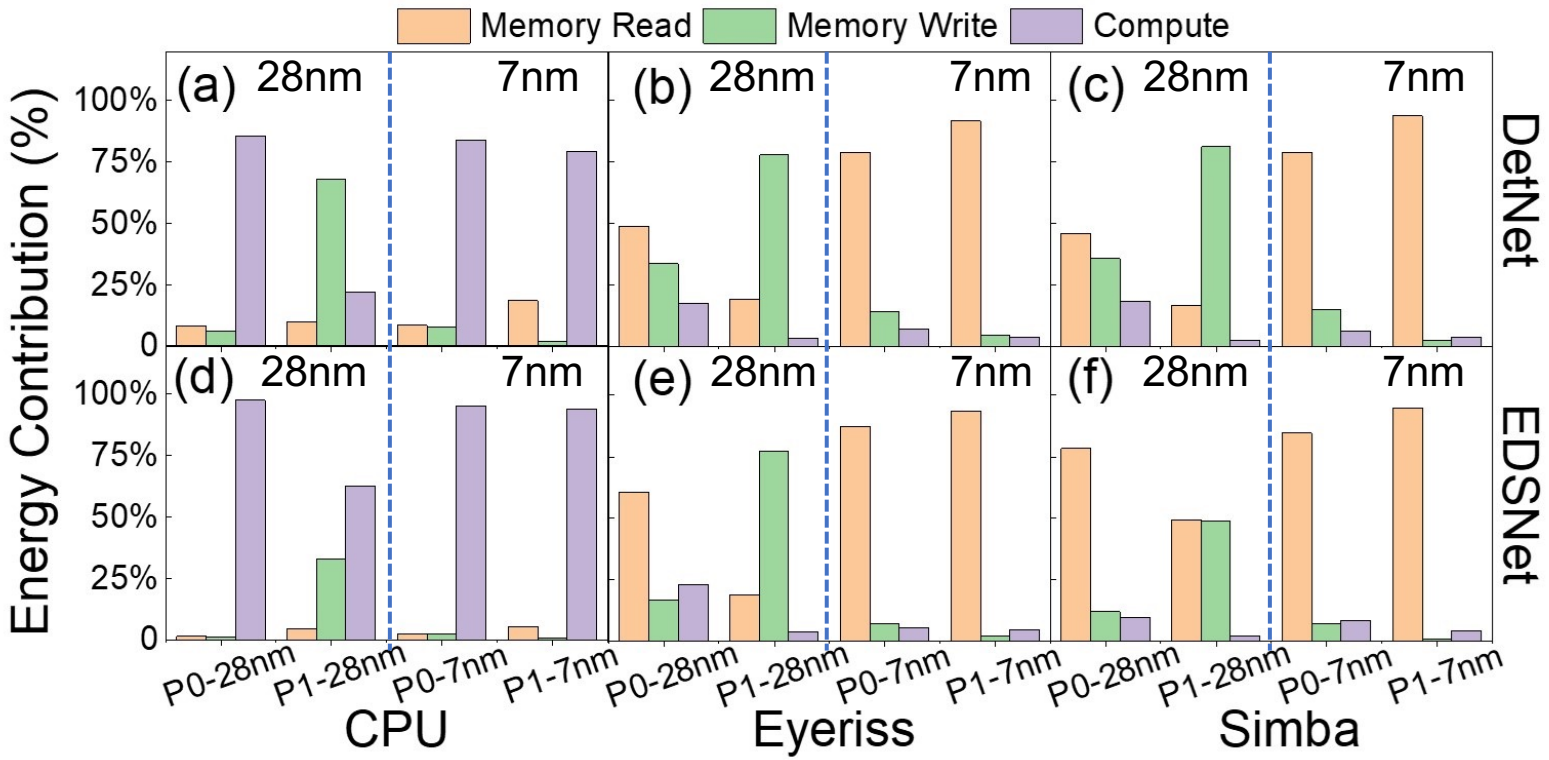


© March 2023, NVM and Neuromorphic Research Group-IITD and Meta Reality Labs, tinyML Research Symposium 2023

Energy Breakdown for Compute & Memory

At 7nm energy estimated for NVM-based variants (P0,P1) > “SRAM-only” variant

- 7nm MRAM type considered is write-optimized (ref-IMEC). However, the XR application is **Read Dominant**.
- Gains @ 7nm can be obtained with a read optimized MRAM.
- Mem Read E > Mem Write E in P0 (all cases) → Reduced write operations in weight memory – **inference dominated workload** (not true for SRAM though)



Serial

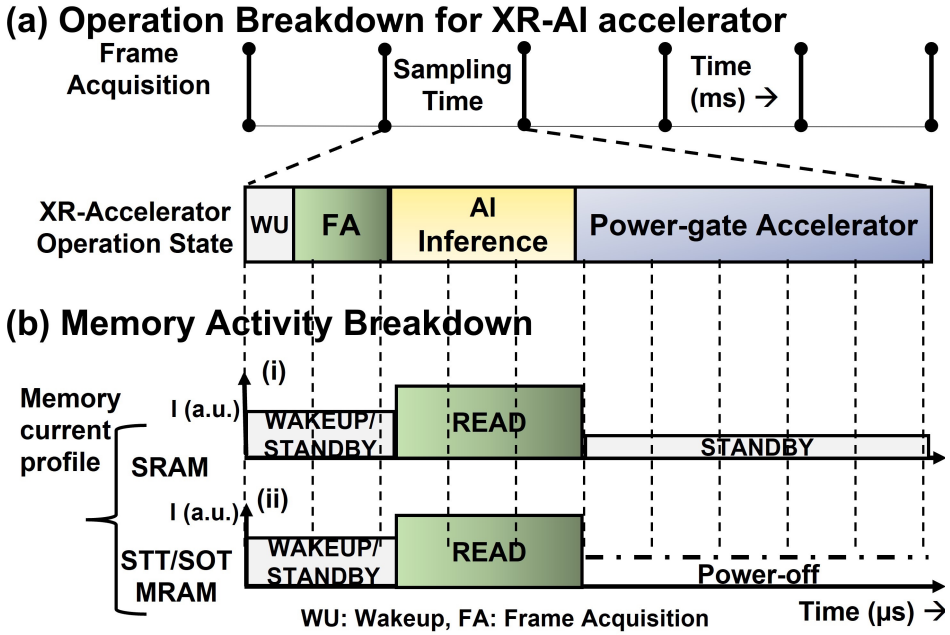
Compute Dominated

Systolic and Data Parallel

Memory-Read Dominated

© March 2023, NVM and Neuromorphic Research Group-IITD and Meta Reality Labs, tinyML Research Symposium 2023

IPS-Analysis



IPS (#Inferences Per Second over op time) / Effective Latency & not actual Inference Latency

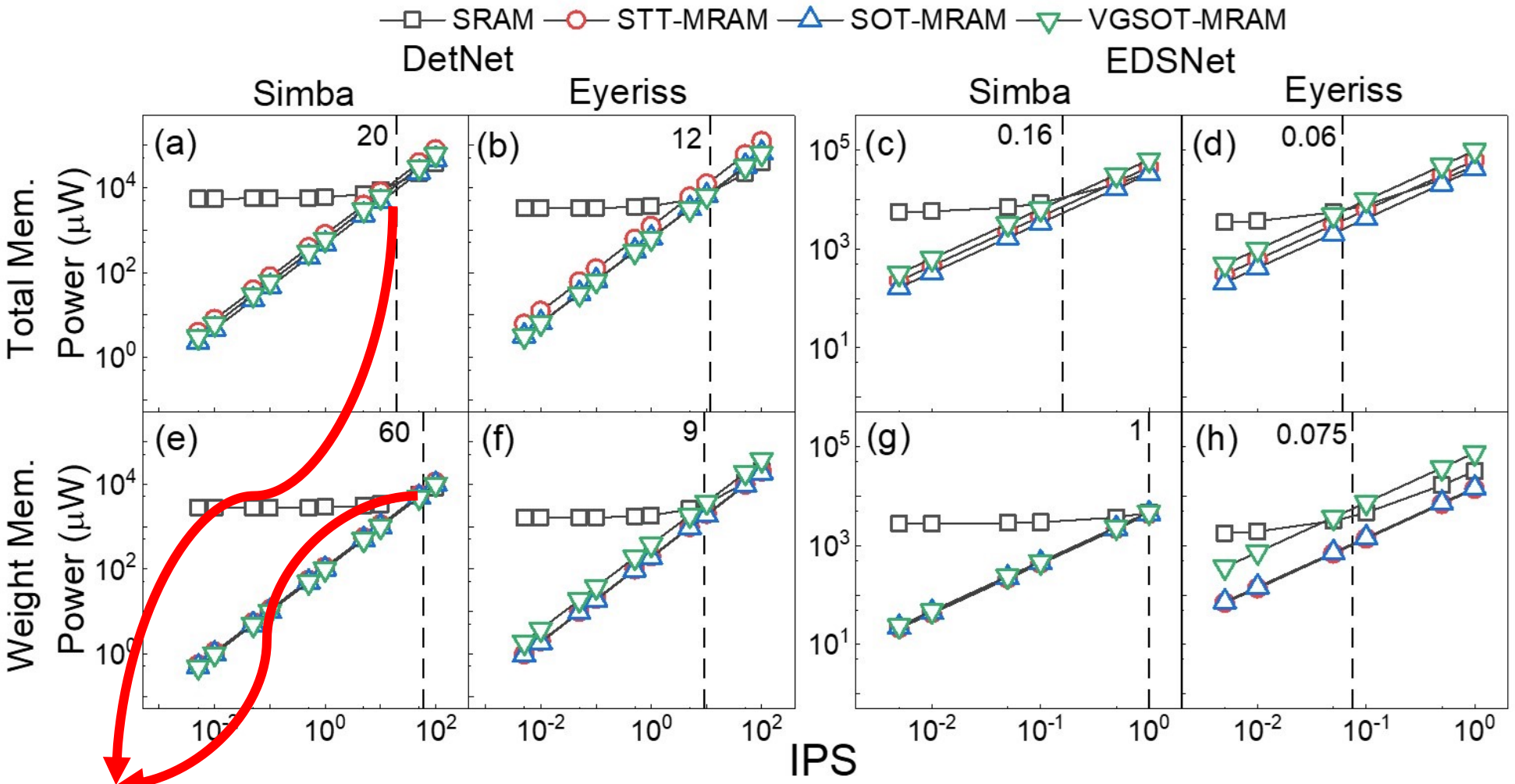
A more relevant performance metric for edge XR-AI as inference operations may:

Invoke AI for XR in Asymmetric/Infrequent manner after long/erratic intervals

Configure: Min. Hand Detection IPS ~ 10 (use)
Min. Eye segmentation IPS ~ 0.1 (use only during initiation of gaze tracking or authentication)

© March 2023, NVM and Neuromorphic Research Group-IITD and Meta Reality Labs, tinyML Research Symposium 2023

IPS-Analysis



Below IPS cross-over point → energy-saving while using advanced NVM compared to baseline SRAM only variant.

© March 2023, NVM and Neuromorphic Research Group-IITD and Meta Reality Labs, tinyML Research Symposium 2023

IPS Analysis - Summary

TABLE III
IPS ANALYSIS SUMMARY FOR PROPOSED ARCHITECTURES USING PE CONFIGURATION V2 (64×64).

XR-AI Workload	Architecture	Inference Latency (ms)		P_{Mem} Savings @ IPS_{min}	
		P0	P1	P0	P1
DetNet IPS_{min}=10	Simba	0.34	0.42	27%	31%
	Eyeriss	0.86	0.86	-4%	9%
EDSNet IPS_{min}=0.1	Simba	48.57	60.72	29%	24%
	Eyeriss	45.22	45.22	-15%	-26%

Clear power saving even with write optimized MRAM!

© March 2023, NVM and Neuromorphic Research Group-IITD and Meta Reality Labs, tinyML Research Symposium 2023

Conclusion

1. Detailed study on 2 XR-AI workloads (hand-detection and eye-segmentation).
2. Design exploration for mapping workloads on CPU and systolic accelerators (QKeras & Timeloop + Accelergy frameworks).
3. Node-scaling analysis and detailed energy breakdown analysis (compute Vs memory).
4. Memory-oriented DTCO based on the use of different types of the emerging MRAM devices.
   a) Memory-Energy Savings ≥ 24% observed for hand detection (at IPS = 10) and eye segmentation (at IPS=0.1) for Simba-like NVM accelerator variant.
   b) Substantial area reduction (≥30%) due to the high-density feature of MRAM technology.

© March 2023, NVM and Neuromorphic Research Group-IITD and Meta Reality Labs, tinyML Research Symposium 2023

Thank You

manansuri@ee.iitd.ac.in

© March 2023, NVM and Neuromorphic Research Group-IITD and Meta Reality Labs, tinyML Research Symposium 2023

Copyright Notice

This presentation in this publication was presented at the tinyML® Research Symposium (March 27,2023). 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