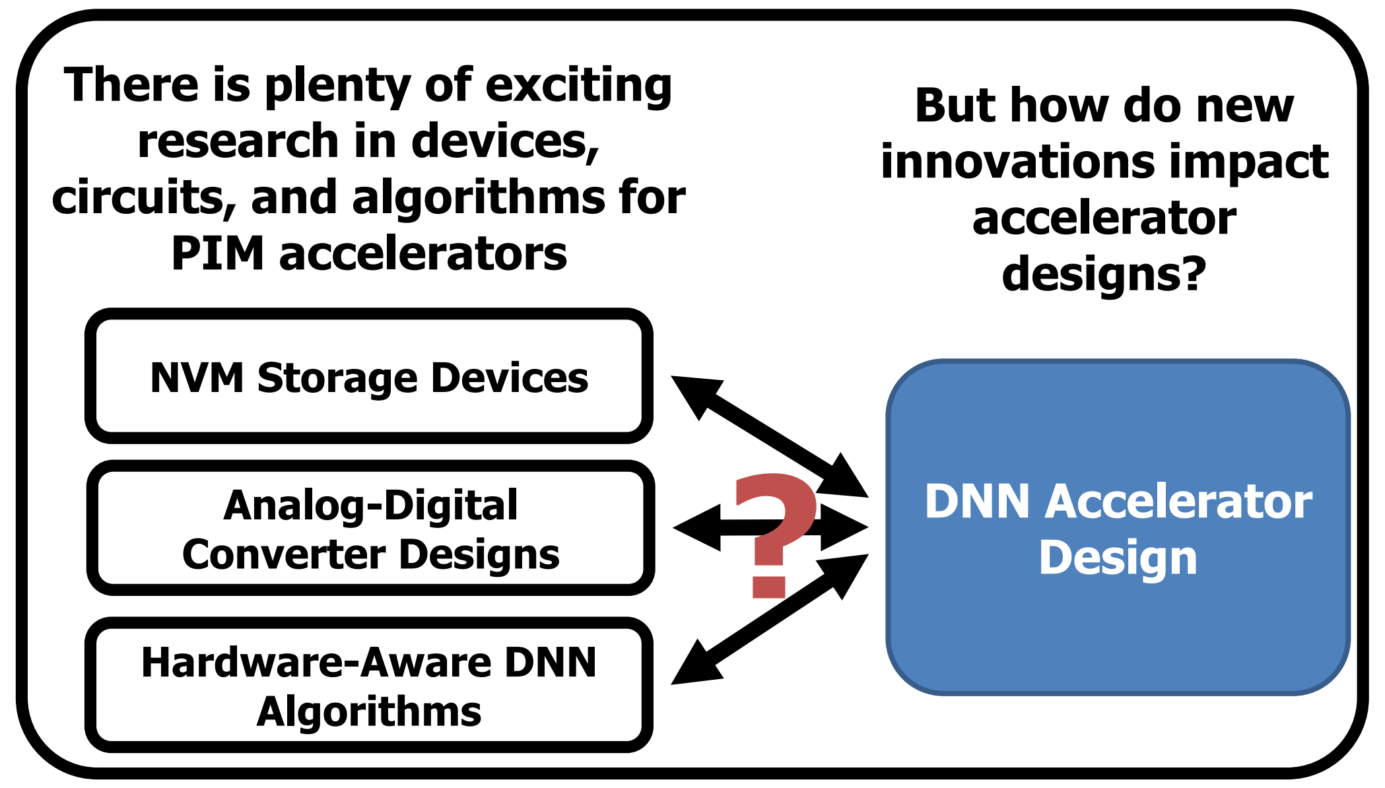
Architectural Evaluation of Processing-In-Memory Systems

Tanner Andrulis, Joel Emer, Vivienne Sze

Motivation: The Need for Fast, Flexible modeling of Processing-In-Memory (PIM) Accelerators

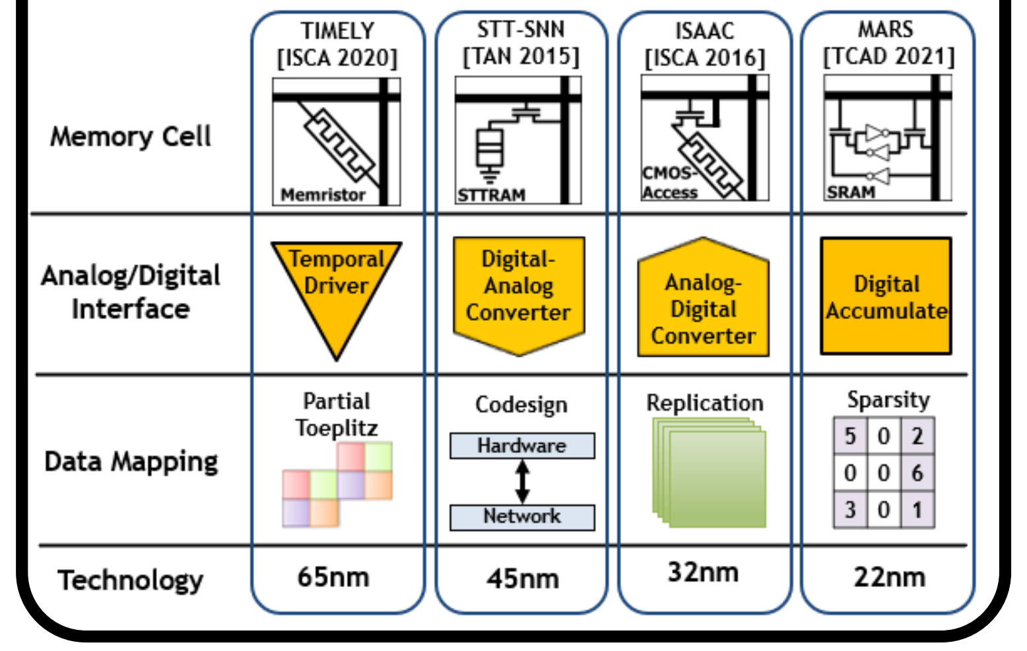
There is a large design space for digital DNN accelerators
Dataflow
Memory Hierarchy
Sparse/Dense DNN Support
Flexibility

Analog PIM makes the design space much larger!
Novel Storage Devices
Analog/Digital Peripheral Circuits
Limited-Resolution Components
Hardware/Software Codesign



There is plenty of exciting research in devices, circuits, and algorithms for PIM accelerators
NVM Storage Devices
Analog-Digital Converter Designs
Hardware-Aware DNN Algorithms
?
But how do new innovations impact accelerator designs?
DNN Accelerator Design

How do we compare design decisions across architectures?



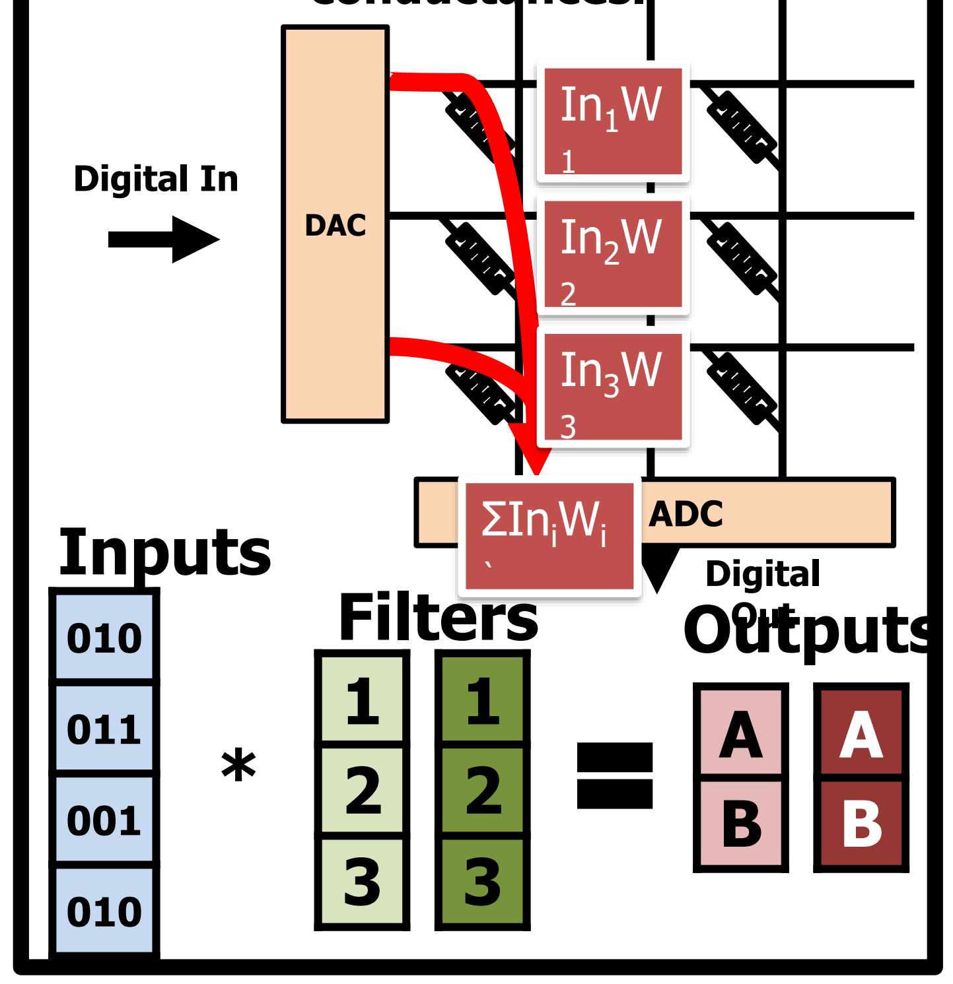
	TIMELY [ISCA 2020]	STT-SNN [TAN 2015]	ISAAC [ISCA 2016]	MARS [TCAD 2021]
Memory Cell	Memristor	STTRAM	CMOS-Access	SRAM
Analog/Digital Interface	Temporal Driver	Digital-Analog Converter	Analog-Digital Converter	Digital Accumulate
Data Mapping	Partial Toeplitz	Codesign Hardware Network	Replication	Sparsity 5 0 2 0 0 6 3 0 1
Technology	65nm	45nm	32nm	22nm

Need a fast, flexible framework to:

✓ See The Impact of New Circuits And Devices
✓ Understand the Design Space
✓ Fairly Compare Design Choices
✓ Fairly Compare PIM and Non-PIM Accelerators

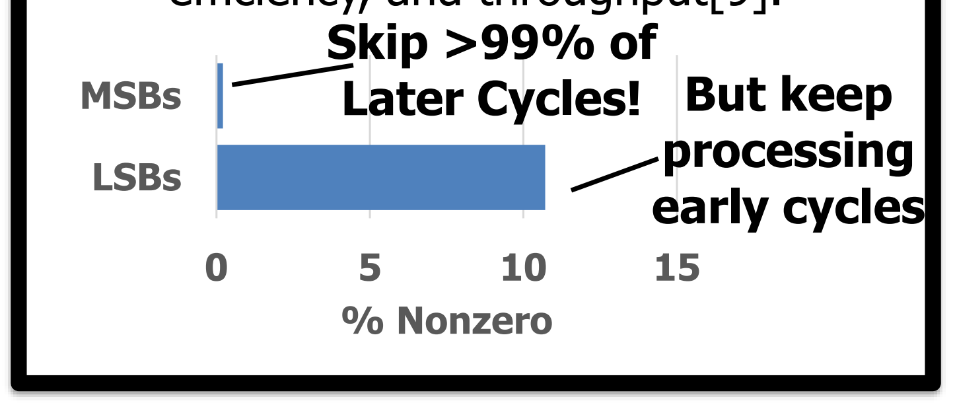
Exploring Tradeoffs in PIM

PIM Crossbar **multiply-accumulates inputs** with **programmed conductances.**



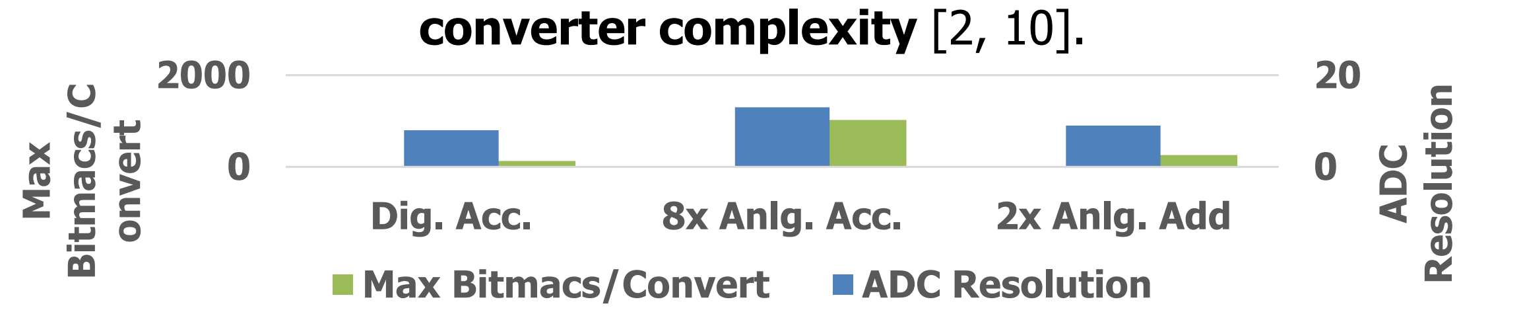
Tradeoffs in Sparsity

Digital-Analog-Converters can skip zero input bits, trading off accuracy, sparsity, efficiency, and throughput[9].



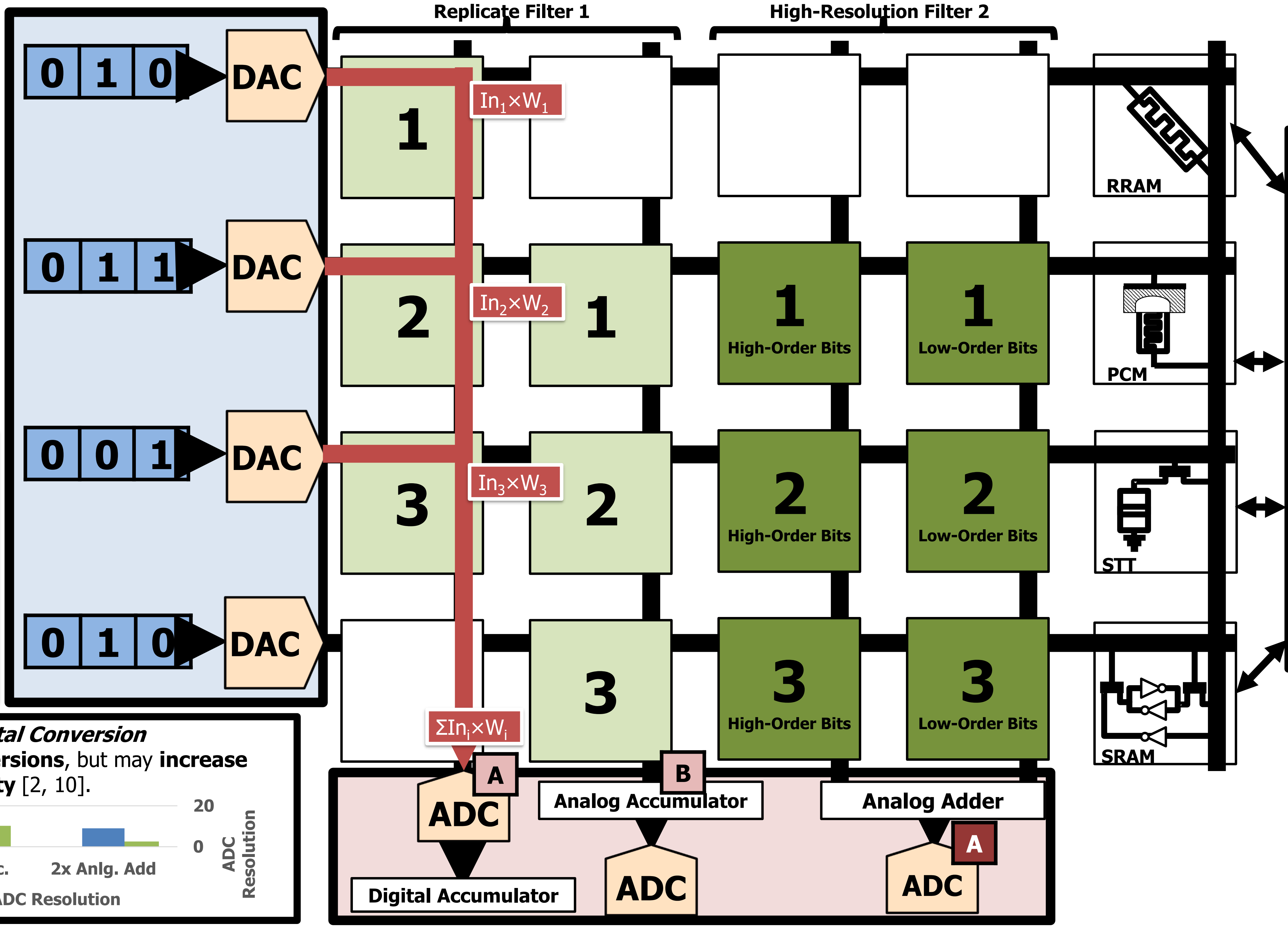
Tradeoffs in Analog-Digital Conversion

Circuits can **reduce expensive conversions**, but may **increase converter complexity** [2, 10].



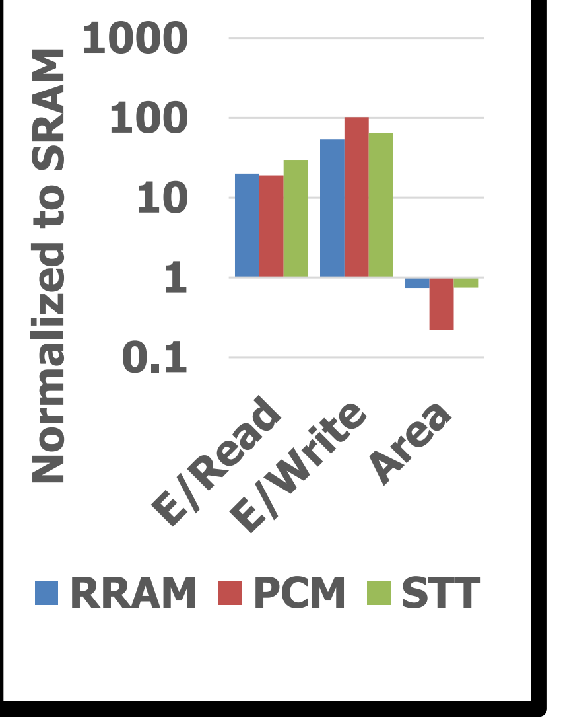
Tradeoffs in Mapping

Data in PIM-Crossbars can be **replicated** to complete multiple convolution steps / vector multiplications at once or **stored across multiple devices** to increase resolution [2].



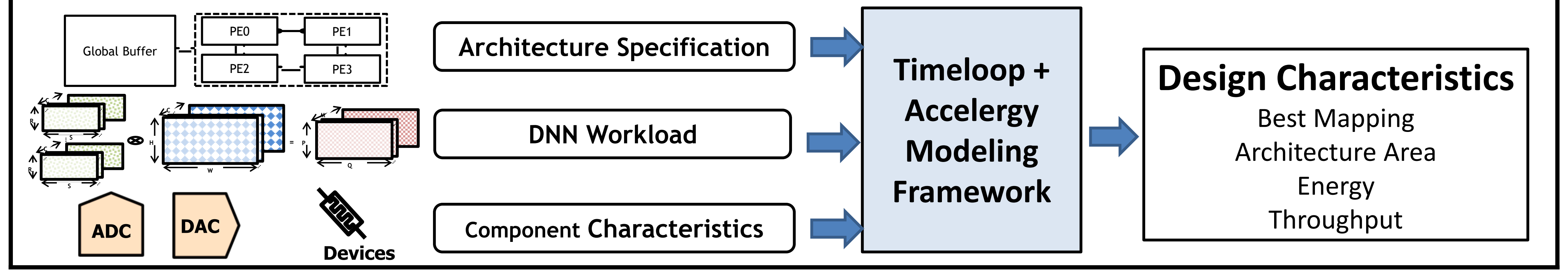
Tradeoffs in Memory Cells

Cell choice can have orders-of-magnitude effect on the read energy, write energy, area, and endurance of the system [4].



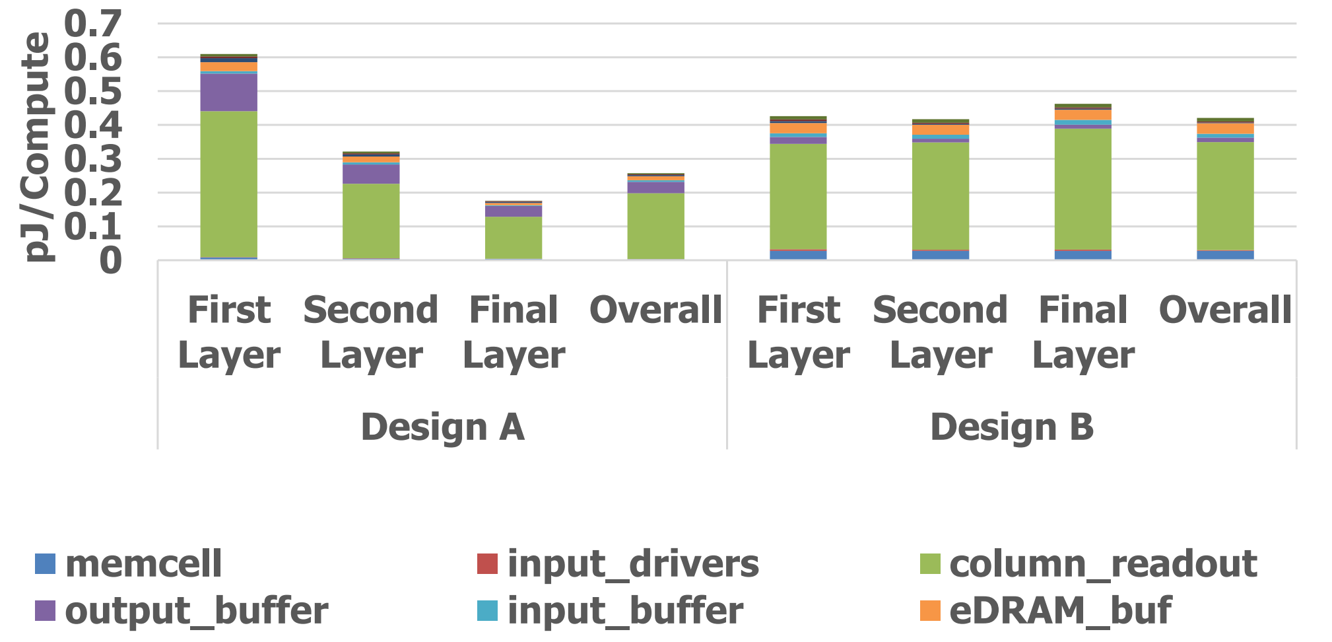
Infrastructure

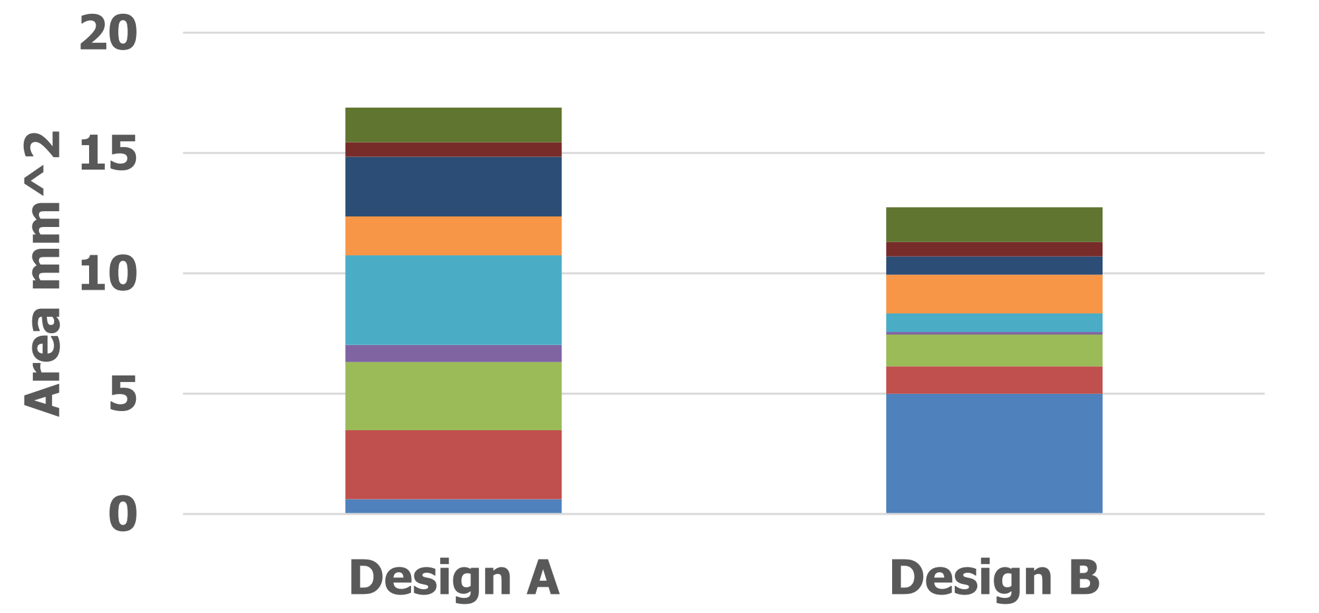
Infrastructure simulates PIM DNN accelerators up to 10,000x faster, provides flexible architecture models, and has easy-to-modify components.



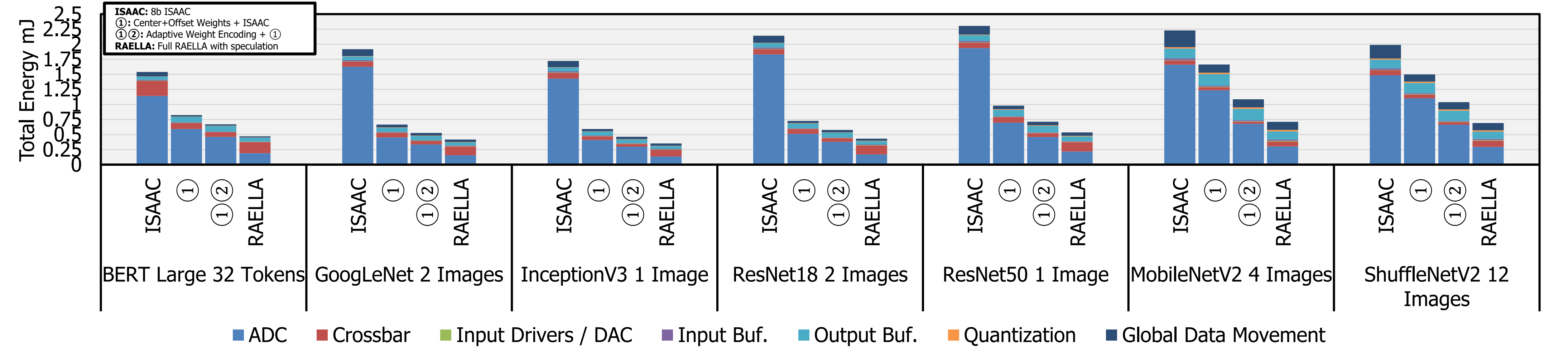
Example Experimental Results

Design A uses 512x512 2-bit ReRAM crossbars and a 1-bit digital-analog-converter. **Design B** uses 170x128 SRAM crossbars and a 2-bit temporal digital-analog-converter. Tested with ResNet18.





The framework is used to model the RAELLA architecture in ISCA ‘23. Below, we show the energy ablation study from the RAELLA paper. The framework was used to model designs flexibly for a range of DNNs.



References

[1] P. Chi et al., “PRIME: A Novel Processing-in-Memory Architecture for Neural Network Computation in ReRAM-Based Main Memory,” in 2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA), Jun. 2016, pp. 27–39. doi: 10.1109/ISCA.2016.13.

[2] W. Li, P. Xu, Y. Zhao, H. Li, Y. Xie, and Y. Lin, “TIMELY: Pushing Data Movements and Interfaces in PIM Accelerators Towards Local and in Time Domain,” arXiv:2005.01206 [cs, eess], May 2020, Accessed: Nov. 26, 2021. [Online]. Available: http://arxiv.org/abs/2005.01206

[3] A. Parashar et al., “Timeloop: A Systematic Approach to DNN Accelerator Evaluation,” in 2019 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), Mar. 2019, pp. 304–315. doi: 10.1109/ISPASS.2019.00042.

[4] L. Pentecost, A. Hankin, M. Donato, M. Hempstead, G.-Y. Wei, and D. Brooks, “NVMExplorer: A Framework for Cross-Stack Comparisons of Embedded Non-Volatile Memories,” arXiv:2109.01188 [cs], Jan. 2022, Accessed: Feb. 24, 2022. [Online]. Available: http://arxiv.org/abs/2109.01188

[5] A. Shafiee et al., “ISAAC: A Convolutional Neural Network Accelerator with In-Situ Analog Arithmetic in Crossbars,” in 2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA), Jun. 2016, pp. 14–26. doi: 10.1109/ISCA.2016.12.

[6] L. Song, X. Qian, H. Li, and Y. Chen, “PipeLayer: A Pipelined ReRAM-Based Accelerator for Deep Learning,” in 2017 IEEE International Symposium on High Performance Computer Architecture (HPCA), Feb. 2017, pp. 541–552. doi: 10.1109/HPCA.2017.55.

[7] T. Song, X. Chen, X. Zhang, and Y. Han, “BRAHMS: Beyond Conventional RRAM-based Neural Network Accelerators Using Hybrid Analog Memory System,” in 2021 58th ACM/IEEE Design Automation Conference (DAC), Dec. 2021, pp. 1033–1038. doi: 10.1109/DAC18074.2021.9586247.

[8] Y. N. Wu, J. S. Emer, and V. Sze, “Accelergy: An Architecture-Level Energy Estimation Methodology for Accelerator Designs,” in 2019 IEEE/ACM International Conference on Computer-Aided Design (ICCAD), Nov. 2019, pp. 1–8. doi: 10.1109/ICCAD45719.2019.8942149.

[9] J. Zhang, H. Yang, F. Chen, Y. Wang, and H. Li, “Exploring Bit-Slice Sparsity in Deep Neural Networks for Efficient ReRAM-Based Deployment,” arXiv:1909.08496 [cs, stat], Nov. 2019, Accessed: Dec. 14, 2021. [Online]. Available: http://arxiv.org/abs/1909.08496

[10] T. Chou, W. Tang, J. Botimer, and Z. Zhang, “CASCADE: Connecting RRAMs to Extend Analog Dataflow In An End-To-End In-Memory Processing Paradigm,” in Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture, Columbus OH USA, Oct. 2019, pp. 114–125. doi: 10.1145/3352460.3358328.]

[11] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2016.

Sponsored By Ericsson, TSMC, MIT AI Hardware Program (ASML, Intel), MIT Quest

Energy-Efficient Multimedia Systems Group (www.rle.mit.edu/eems)