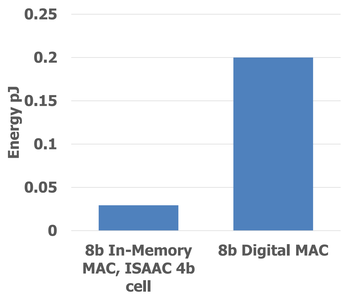
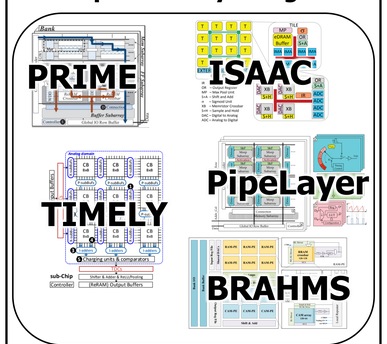
Architectural Evaluation of Processing-In-Memory Systems

Tanner Andrulis, Joel Emer, Vivienne Sze, Massachusetts Institute of Technology

Motivation



PIM efficiency has inspired many designs



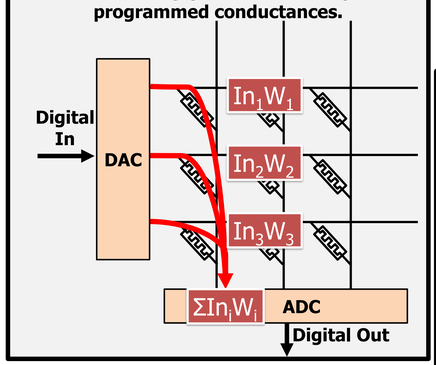
Understand the space with PIM Modeling Framework

Existing PIM Systems | New Designs with Novel Parts or Combos of Design Choices | Non-PIM Accelerator Designs

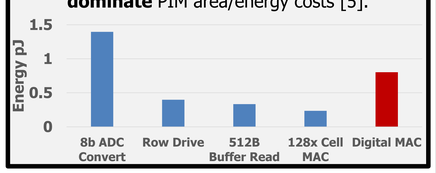
Timeloop + Accelergy + PIM Simulation Framework

- ✓ **Full Understanding of Design Space**
- ✓ **Fair Comparison of Various Systems and Design Choices**
- ✓ **Fair Comparison of PIM and Non-PIM Architectures**
- ✓ **Discovery of New Architectures**

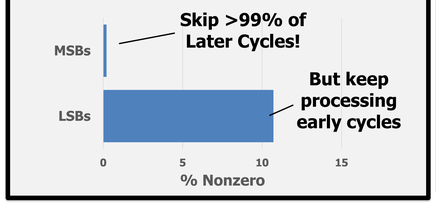
PIM Crossbar **multiply-accumulates inputs** with **programmed conductances.**



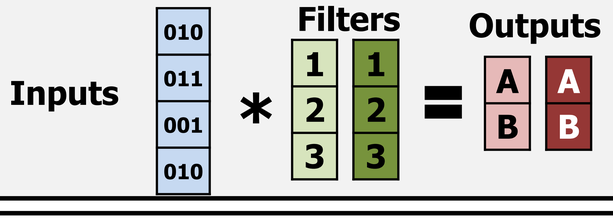
Analog MAC is very efficient; **peripherals dominate** PIM area/energy costs [5].



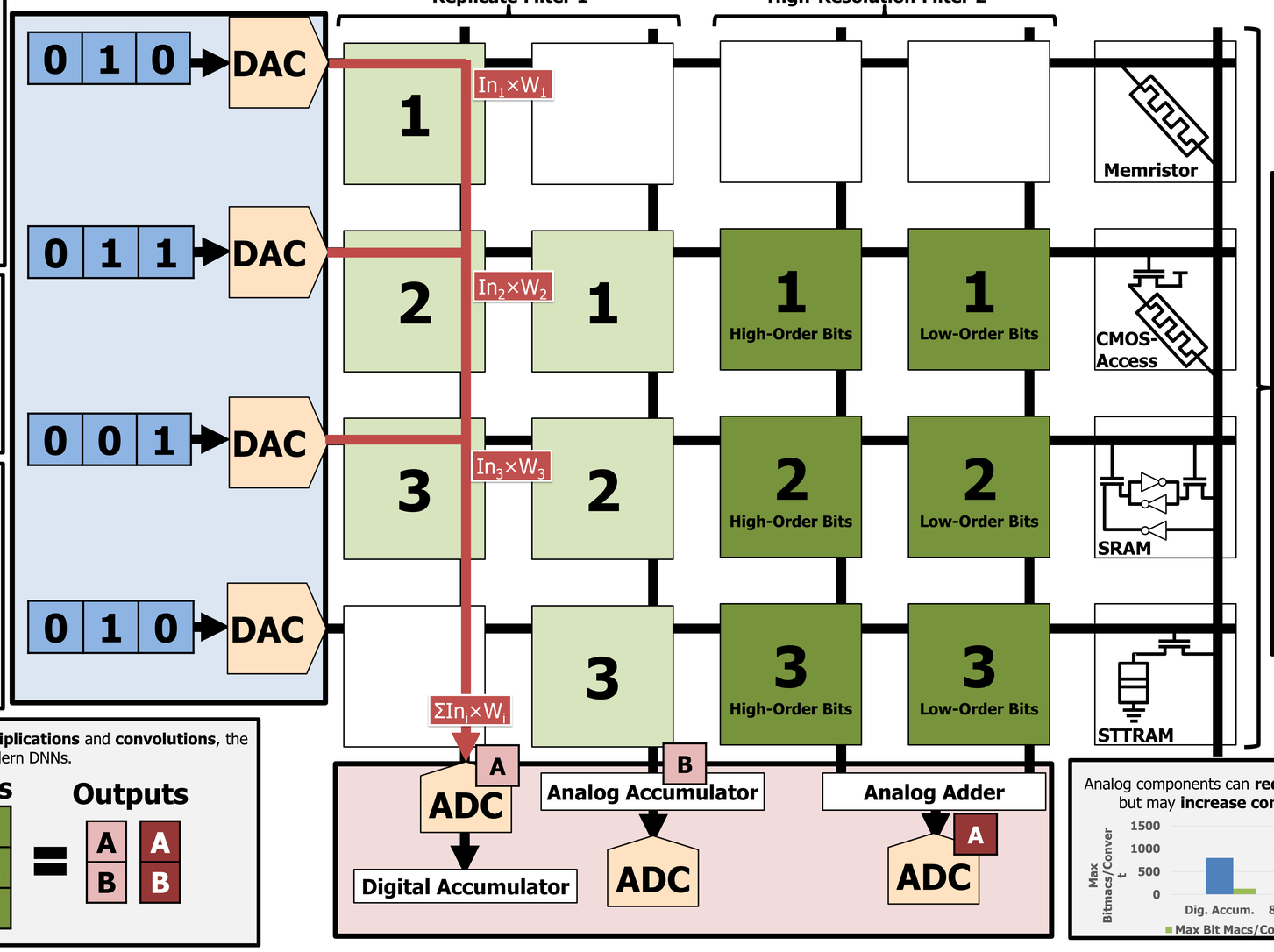
Digital-Analog-Converters feed in a few bits at a time, enabling **sparse data optimizations** [9].



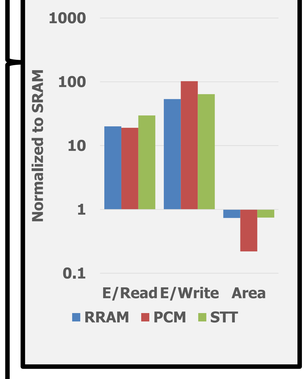
PIM architectures efficiently compute **matrix multiplications** and **convolutions**, the core computations of modern DNNs.



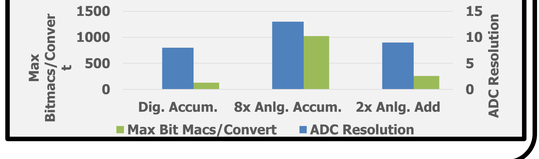
Data in PIM-Crossbars can be **replicated** to complete multiple convolution steps / vector multiplications at once or **stored across multiple devices** to increase resolution [2].



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



Analog components can **reduce expensive converter use**, but may **increase converter complexity** [2, 10].



Infrastructure

Architecture
Compact architecture spec. captures all design tradeoffs

Workload
Parameters of DNN layers used for analysis

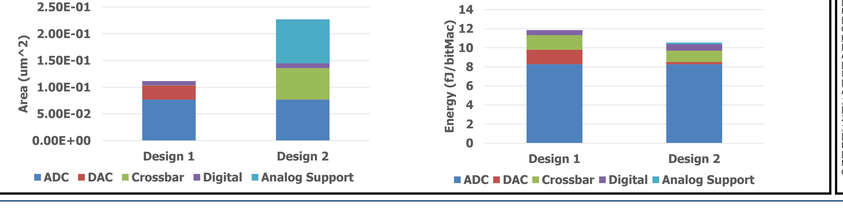
Modified Accelergy+Timeloop

- NVM Models
- Problem-To-Crossbar Mapping
- Converter Plugins
- User Custom Components

→ **Energy Estimation**, **Area Estimation**

Example Analysis

Design 1 is an ISAAC-like [5] design using memristors cells. **Design 2** uses SRAM cells [4] and 2-way analog sum devices [2]. Tested with AlexNet layer 1, large batch size.



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.