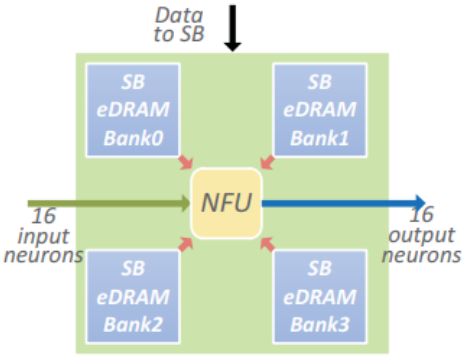
MODSIM 2025

SIMULATION OF BIT-SERIAL DIGITAL PROCESSING-IN-MEMORY SYSTEMS

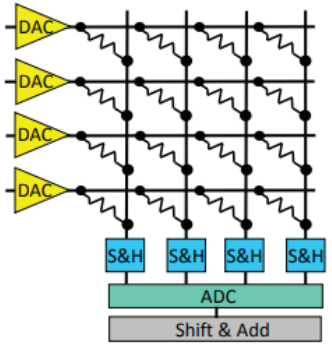
Siyuan Ma (UT Austin), Jiajun Hu (ASU), Lizy John (UT Austin), Aman Arora (ASU)

Processing-in-Memory Background



Near Memory, Digital

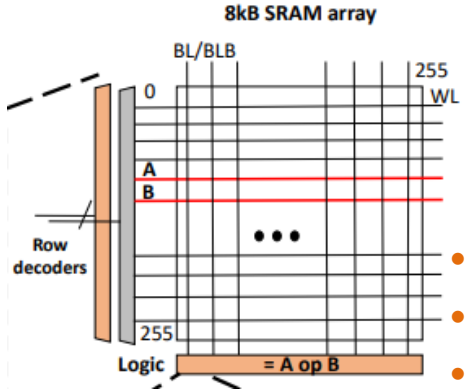
[Chen et al., MICRO 2014]



In Memory, Analog

[Shafiee et al., ISCA 2016]

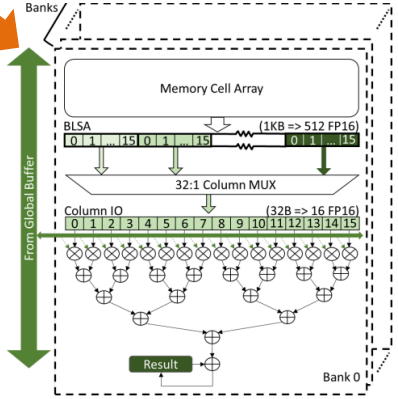
PIM



In Memory, Digital, Bit-Serial

- Flexible data precision
- Low area overhead
- Massive parallelism

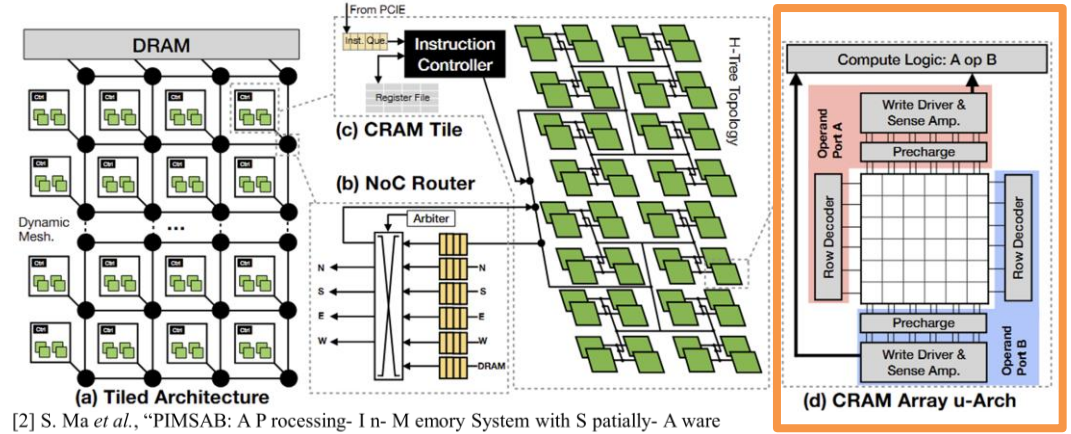
[Eckert et al., ISCA 2018]



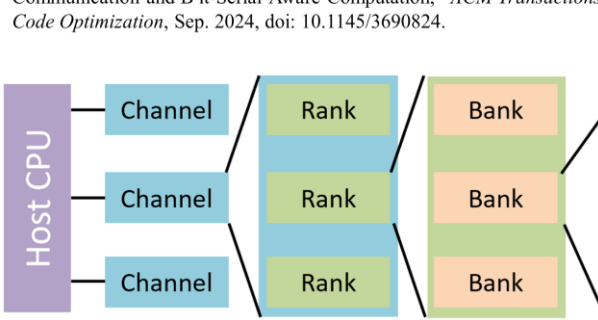
In Memory, Digital, Bit-Parallel

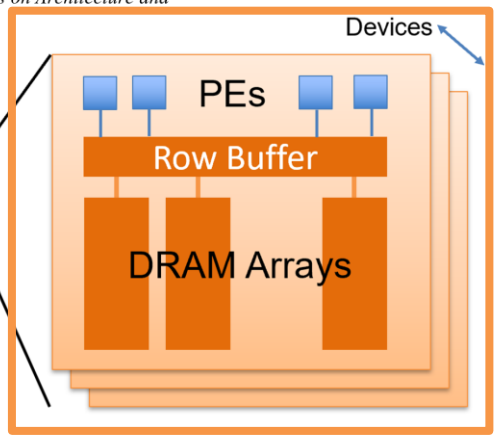
[He et al., ASPLOS 2020]

Bit-Serial PiM Architectures



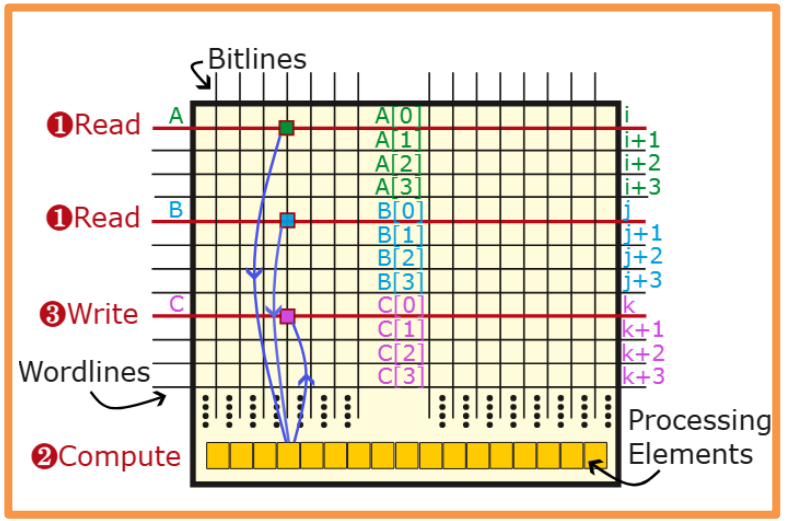
[2] S. Ma *et al.*, "PIMSAB: A P rocessing- I n- M emory System with S patially- A ware Communication and B it-Serial-Aware Computation," *ACM Transactions on Architecture and Code Optimization*, Sep. 2024, doi: 10.1145/3690824.





SRAM PiM

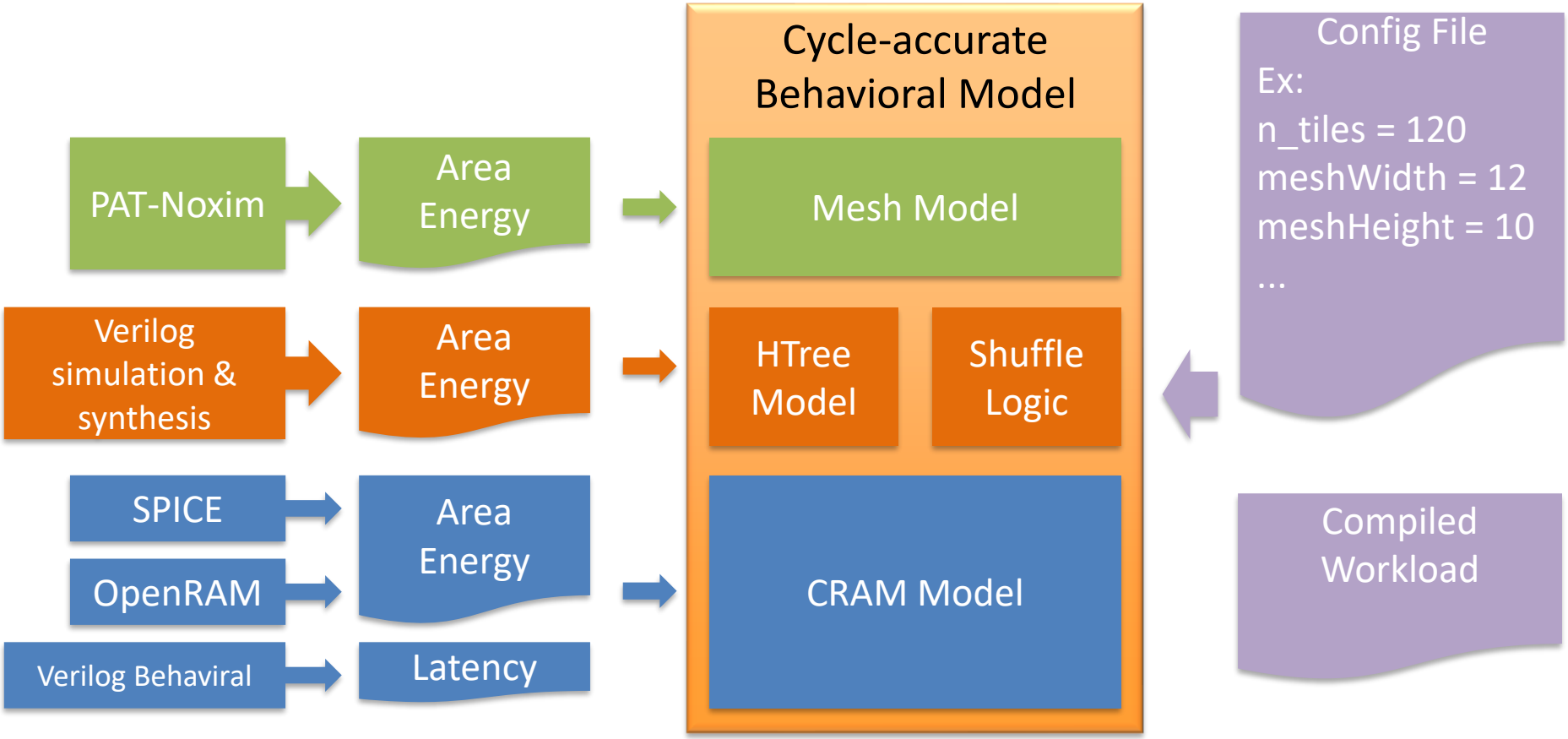
CRAM



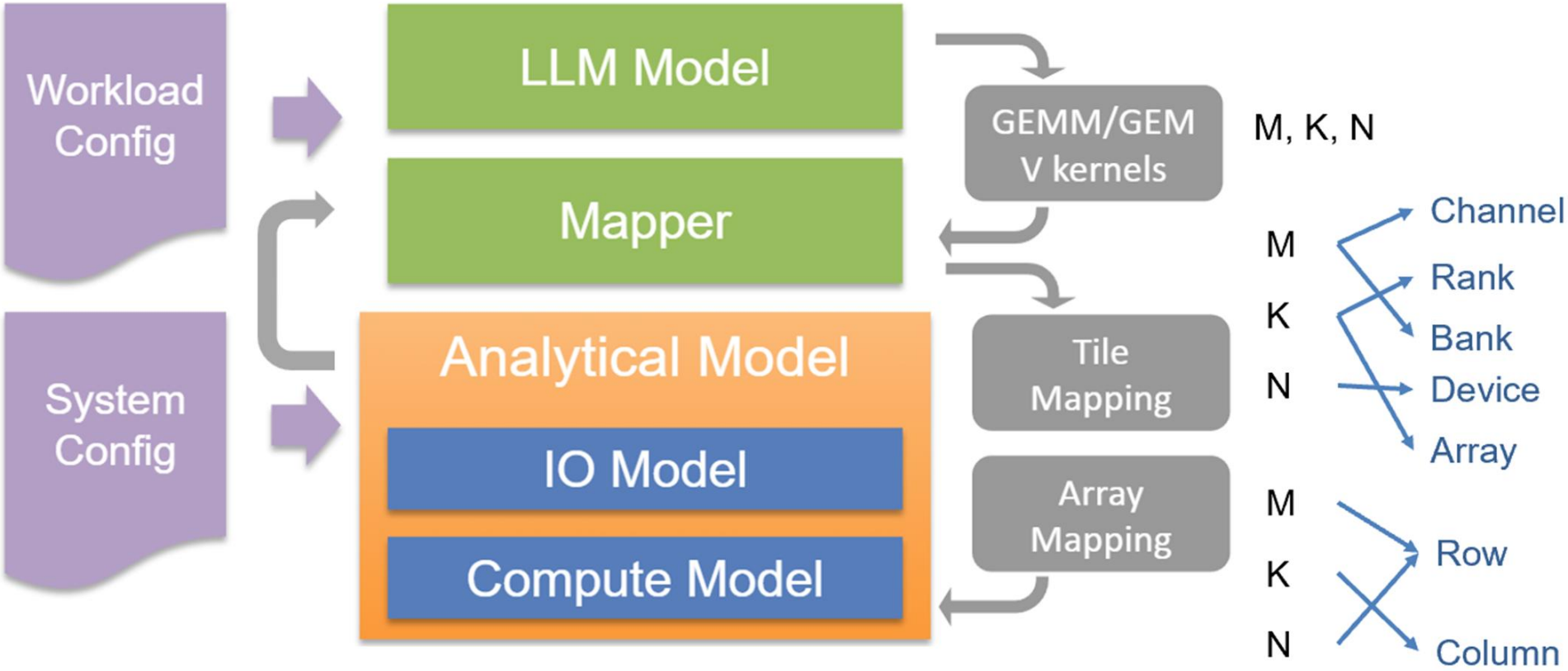
[1] A. Arora et al., "COMEFA: Deploying Compute-in-Memory on FPGAs for Deep Learning Acceleration," ACM Transactions on Reconfigurable Technology and Systems, vol. 16, no. 3, pp. 1–34, Jun. 2023, doi: 10.1145/3603504.

DRAM PiM

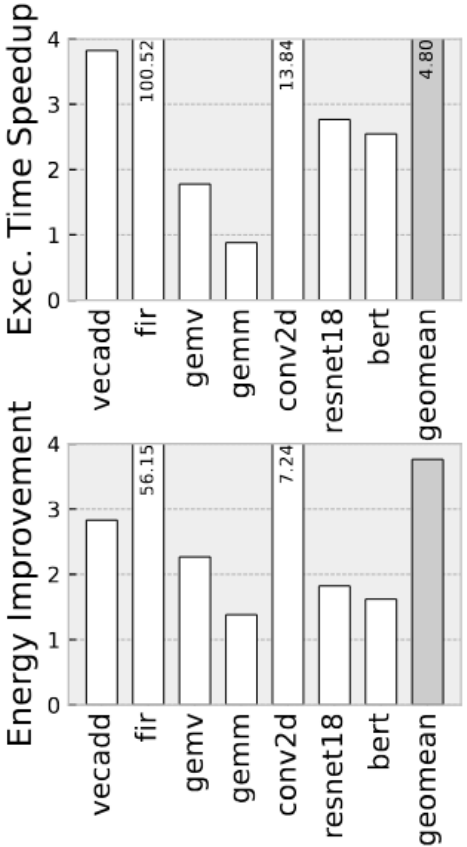
SRAM PiM Modelling Framework



DRAM PiM Modelling Framework

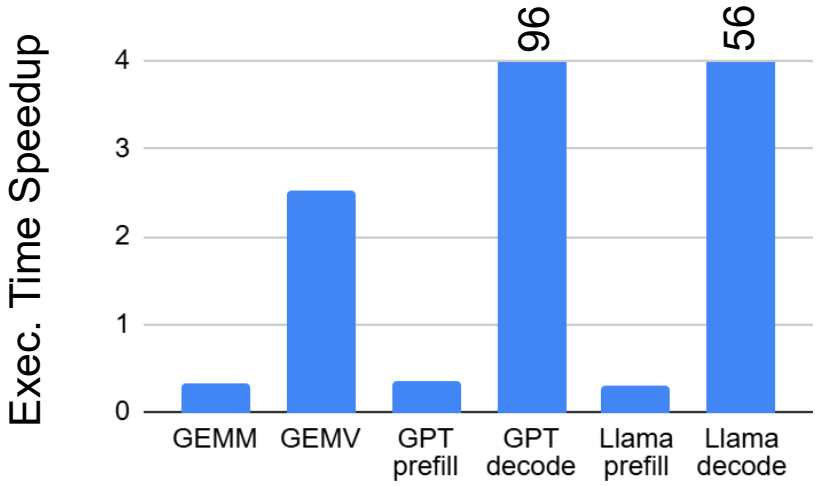


Results



SRAM PiM vs. A100 GPU

- SRAM PiM: 4.8x speedup and 3.88x energy reduction
- DRAM PiM: 2.5x speedup on GEMV, and 96x, 56x speedup on GPT and Llama decoding
- A100 better at GEMM and LLM prefill



DRAM PiM vs. A100 GPU