Improving the Compatibility and Manufacturability of Digital Architectures for Processing Using Resistive Memory

Minh S. Q. Truong, Liting Shen, Alexander Glass, Alison Hoffmann, L. Richard Carley, James A. Bain, Saugata Ghose† (Carnegie Mellon Univ., †Univ. of Illinois Urbana-Champaign)

Summary of the RACER Architecture

- Resistive **processing-using-memory (PUM)** architectures
  - ➢Use electrical ***interactions between interconnected memory cells*** to perform primitive computational functions
  - ➢Can eliminate data movement b/w main memory, CPU
- Existing resistive PUM architectures' performance scales proportionally with the crossbar size, but a crossbar cannot be larger than ***200x200 cells***
- RACER is a PUM architecture and ISA designed for small crossbars, and utilizes a novel ***bit-pipelining execution model***
- ***142x speedup, 233x energy savings vs. 16-core Xeon CPU***
- ***15x speedup, 23x energy savings vs. 2304-shader NVIDIA GPU***

Bit-Pipelining Execution Model

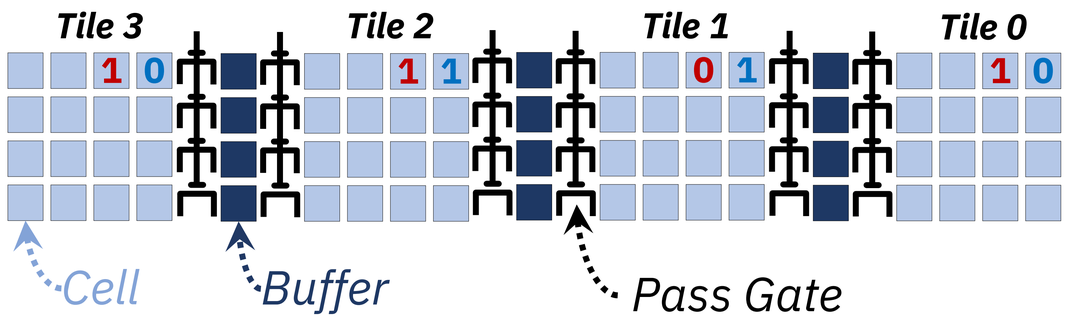
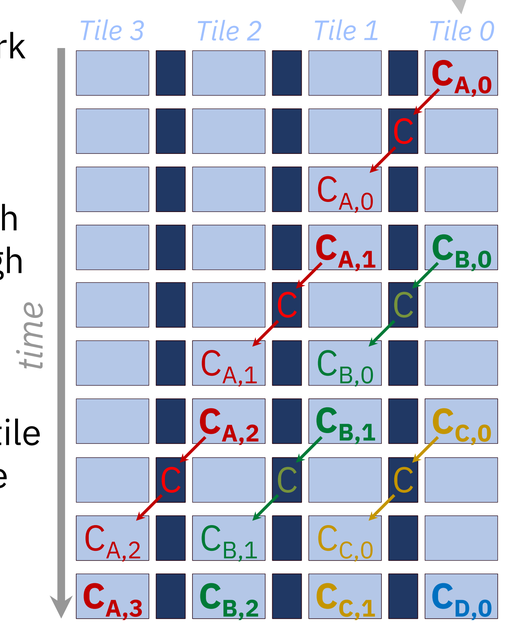


Figure 1: Four tiles and three buffers are used to store 4-bit words 1101 (in red) and 0110 (in blue)

Figure 2: Bit-pipelined ripple-carry addition, $C_{X,Y}$ is the carry bit of add operation X at bit position Y

- Multiple small 64x64 crossbars (i.e., ***tiles***) work in parallel to increase bit-serial throughput
- Each tile contains one bit of a w-bit word
- ***Buffers*** (1x64 crossbars) are added between tiles to enable ***inter-tile communication***, which can connect to or disconnect from tiles through controllable pass gates
- Bit-serial computation repeats the same operations for each bit: we can reuse instructions by propagating them from tile to tile
- In the bit-pipelining execution model: tiles are pipeline stages, buffers are pipeline registers
- ***Bit-pipelining improves the throughput of a w-bit bit-serial operation by w***



In-Crossbar Logic Families

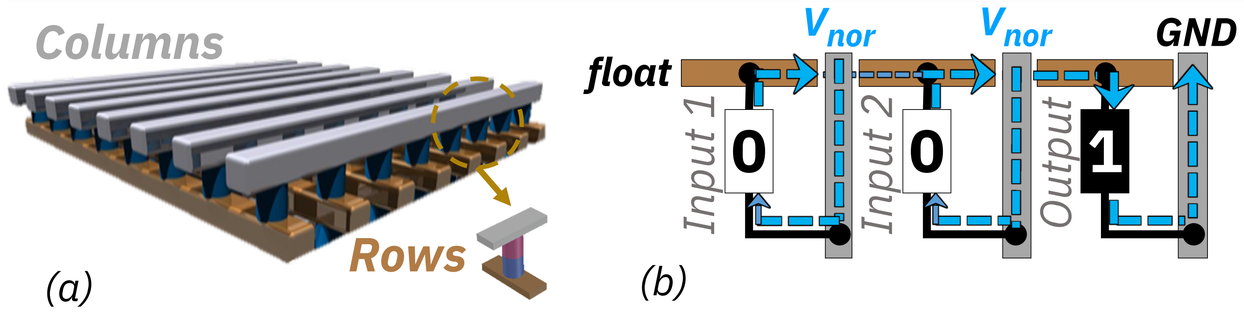


Figure 3: (a) Resistive crossbar; (b) Voltage assertions to perform NOR in memory

- Resistive memory cells store information as different levels of resistance
- Asserting correct voltages on the crossbar's column can realize whole-column Boolean primitives (e.g., NOR)
- ***Changing the assertion voltage can change the logic primitives***

Input Assertion Voltage (V)	Voltage Drop (V) Over Output When ...				
	Input 000	Input 001	Input 011	Input 111	
1	0	0.5 (switched!)	0.67 (switched!)	0.75 (switched!)	**3-Input NOR**
0.75	0	0.38	0.5 (switched!)	0.56 (switched!)	
0.67	0	0.34	0.45	0.5 (switched!)	**3-Input NAND**

Table 1: Changing the assertion voltages results in different Boolean primitives (Kvantinsky+ 2014 and Gupta+ 2018)

Extending RACER to Support Any Logic Family

- RACER's Decode & Drive units act as interface between ***technology-agnostic*** control/peripheral circuitry and ***technology-specific*** crossbars
- We formalize the assertion voltages to V_{in}, V_{out}, V_{float}, where a column in a crossbar is either
  - ➢ Asserted with V_{in} if it is either Col. A or Col. B
  - ➢ Asserted with V_{out} if it is either Col. C
  - ➢ Asserted with V_{float} if they are not involved in the current micro-op
- V_{in}, V_{out}, V_{float} can change to support different logic families

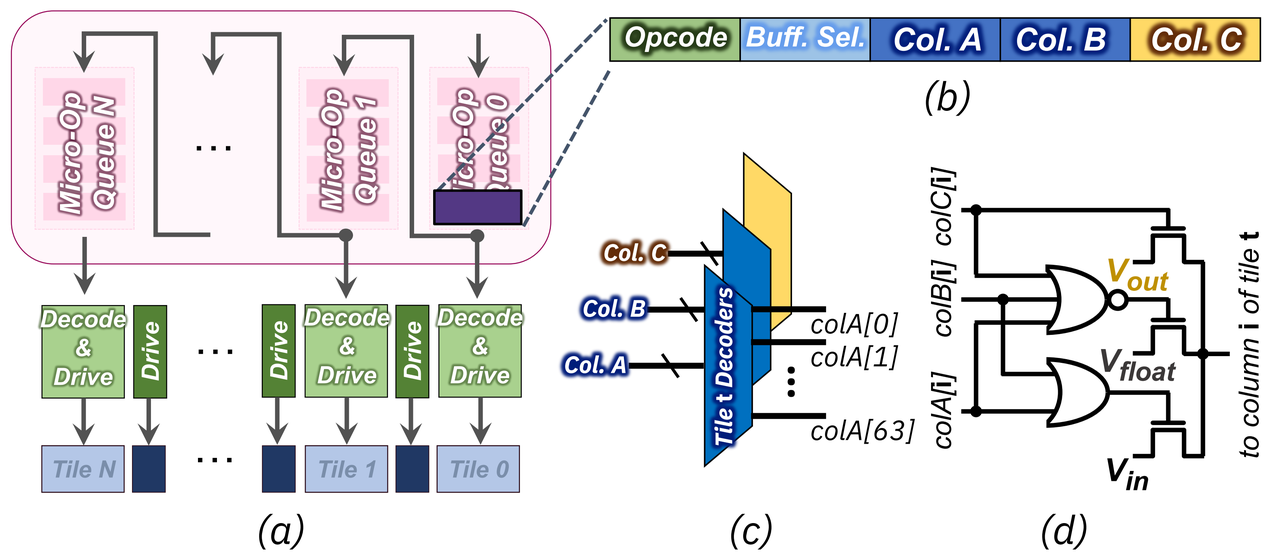
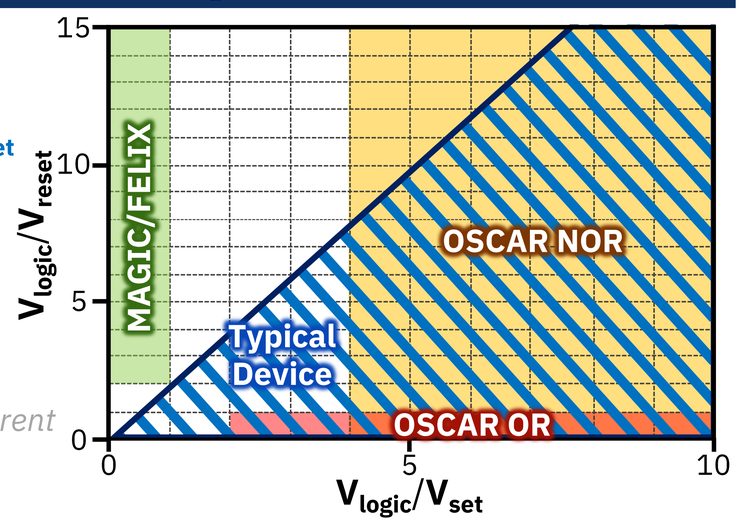


Figure 4: (a) RACER control circuitry decoupled from the pipeline by the decode & drive units; (b) micro-op fields; (c) decode units; (d) drive units

OSCAR: Relaxing Device Switching Constraints

- **MAGIC/FELIX**
  - ➢**Constraints: $2V_{reset}<V_{nor}<V_{set}$**
- **OSCAR NOR**
  - ➢**Constraints: $V_{nor}>4V_{set}$**
- **OSCAR OR**
  - ➢**Constraints: $V_{set}<V_{or}<2V_{reset}$**

Figure 5: Device constraints of different logic families



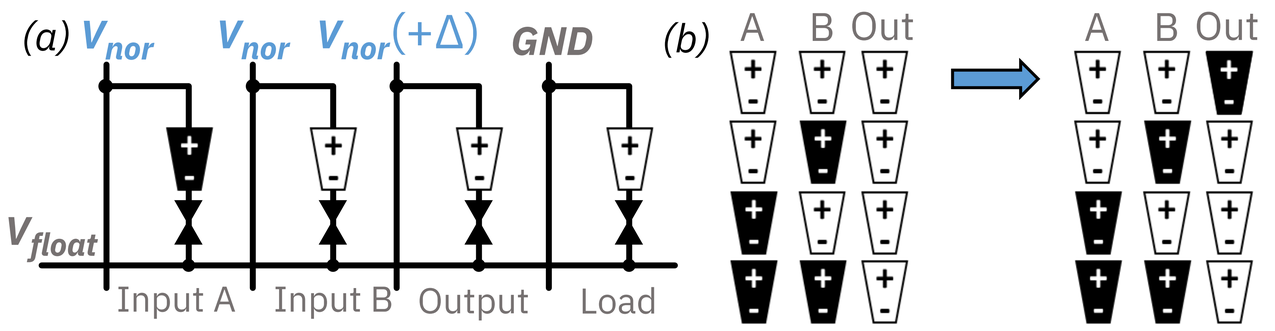


Figure 6: (a) Voltage assertions for OSCAR NOR; (b) possible output transitions for NOR, with the blue arrow indicating output cell resistance switches

Results

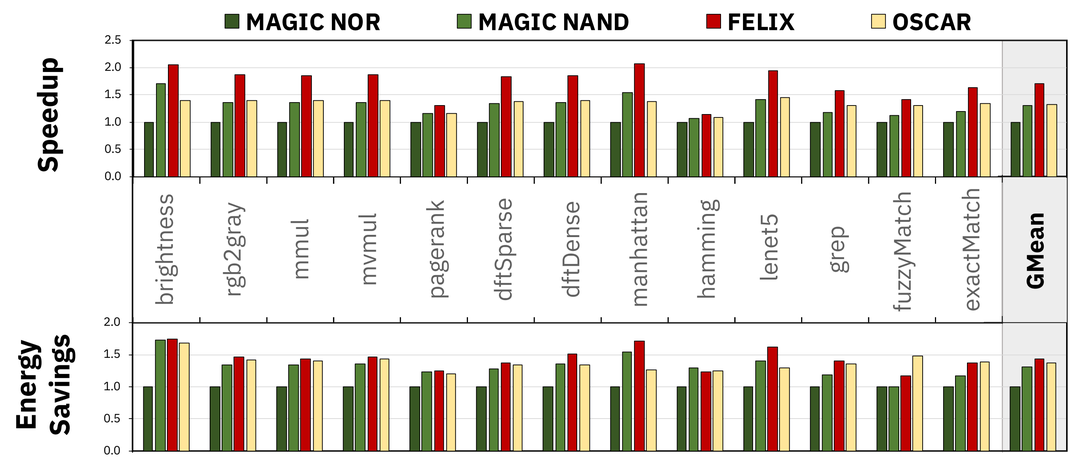


Figure 7: Speedup and energy savings normalized to MICRO'21 RACER results

- ***OSCAR increases RACER's speedup and energy savings by 30% and 37% compared to RACER + MAGIC***
- ***RACER + OSCAR achieves 142× speedup and 233× energy savings compared to a modern 16-core Xeon CPU***
- ***OSCAR's constraints are easier to realize on real resistive devices***

Video: MICRO'21 RACER Architecture

Paper: MICRO'21

Paper: JETCAS'22

University of Illinois Urbana-Champaign

Carnegie Mellon University