Efficient Routing for CGRAs using Multi-Pole NEM Relays

ASP-DAC 2022: Session 14

Emerging Devices, Technologies, and Applications

Akash Levy*, Michael Oduoza*, Akhilesh Balasingam, Roger T. Howe, Priyanka Raina

Stanford University

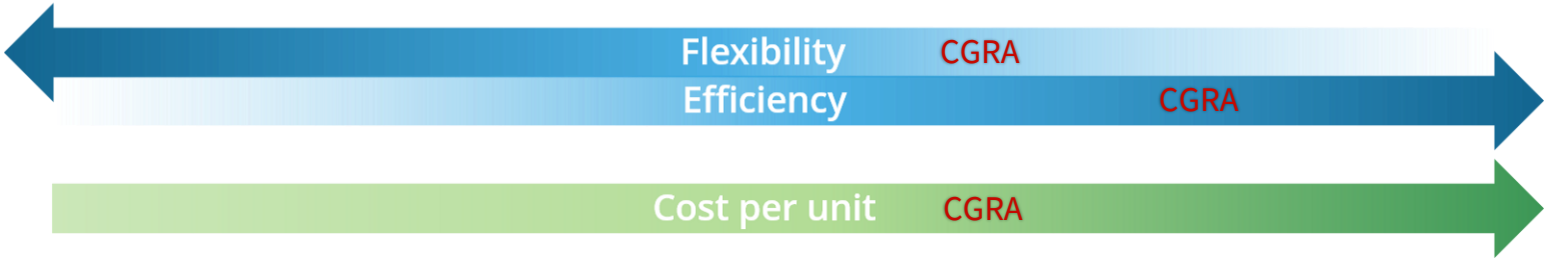
Dept. of Electrical Engineering

*Equal contribution

Designing an application today

CPU, GPU, FPGA, and ASICs

Tradeoffs



Where does reconfigurability overhead come from?

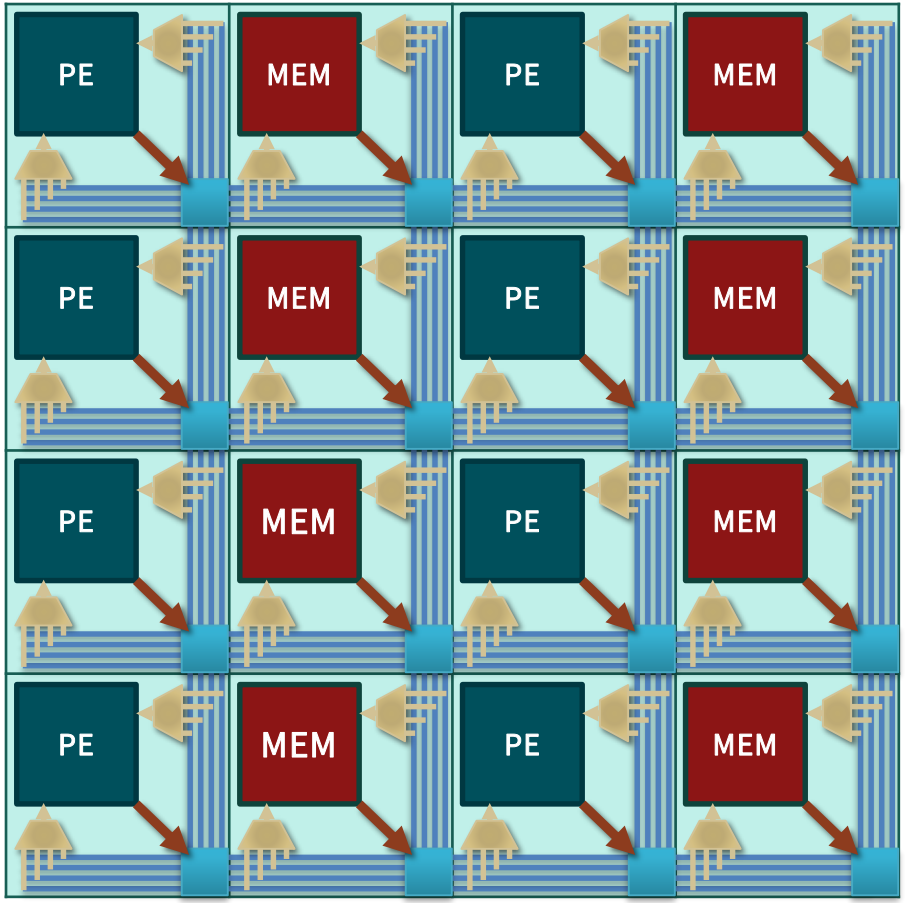
- **Reconfigurable routing**
  - Huge multiplexer cost
  - FPGA compared to ASIC:
    - Area 20-35x
    - Power 5-12x
    - Delay 3-4x

Source: Kuon, Ian, and Jonathan Rose. "Measuring the gap between FPGAs and ASICs." *IEEE Transactions on computer-aided design of integrated circuits and systems* 26, no. 2 (2007): 203-215.

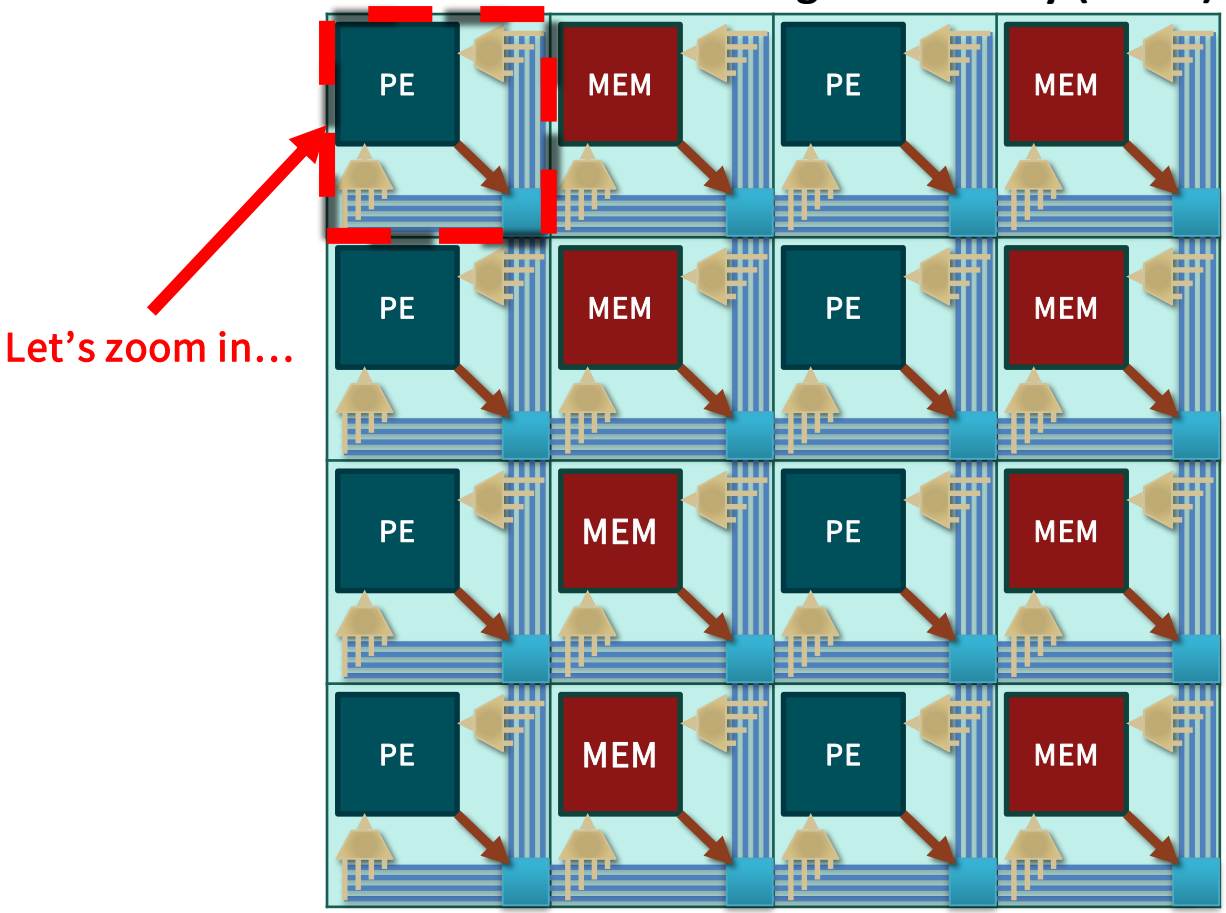
- **Configuration memory**
  - Startup time: SRAM-based FPGAs take milliseconds to seconds to copy configuration data from off-chip to on-chip
  - SRAM takes silicon area and consumes static power
  - Increases startup power/area/delay for intermittent applications

Both routing & config memory are control plane components

Coarse Grained Reconfigurable Array (CGRA)

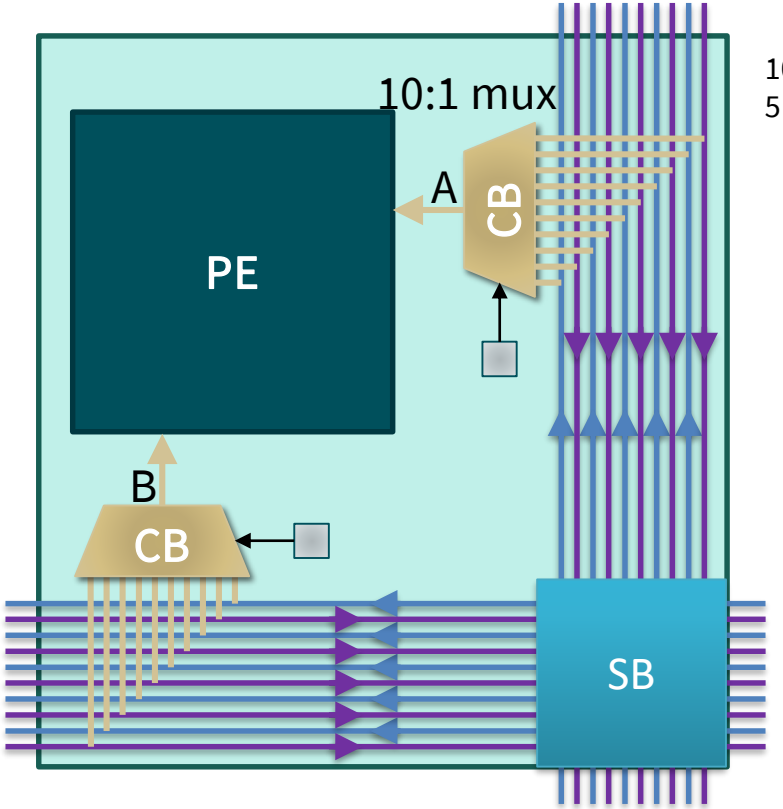


Coarse Grained Reconfigurable Array (CGRA)



Connection Box (CB)

Each PE needs two 10:1 CB muxes and control registers for them

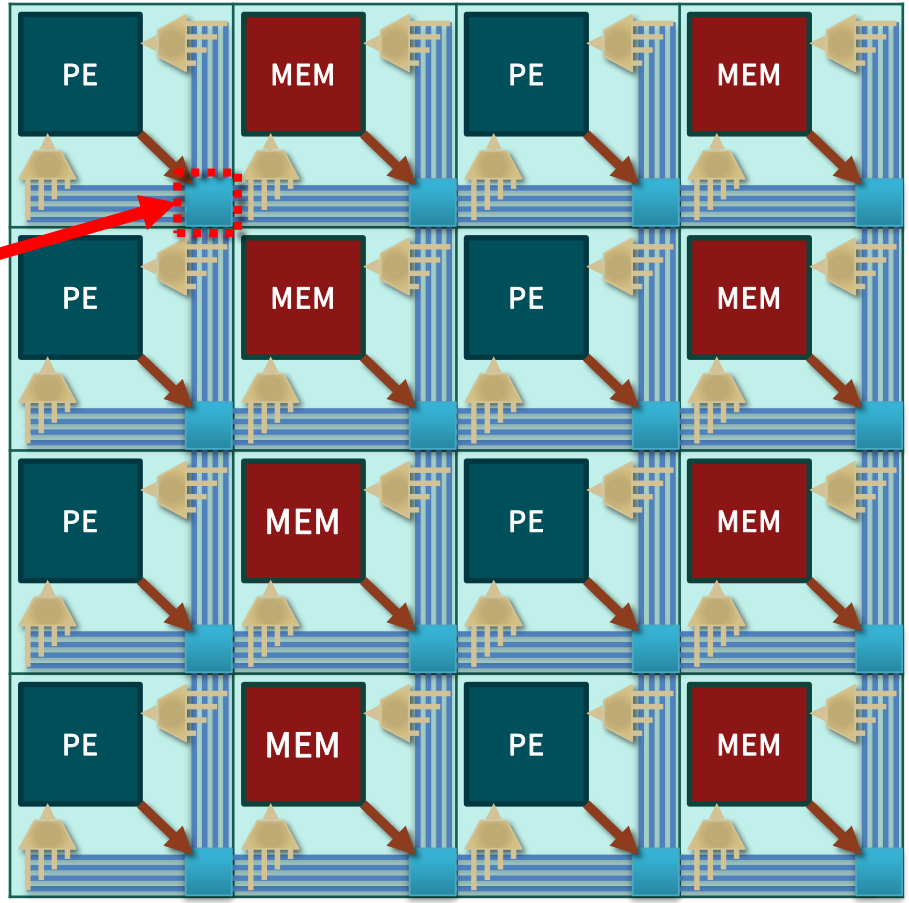


10 vertical 16-bit tracks
5 driven up, 5 driven down

10 horizontal 16-bit tracks
5 driven left, 5 driven right

There are similar 1-bit CBs for feeding 1-bit inputs from 10 horizontal and vertical 1-bit tracks

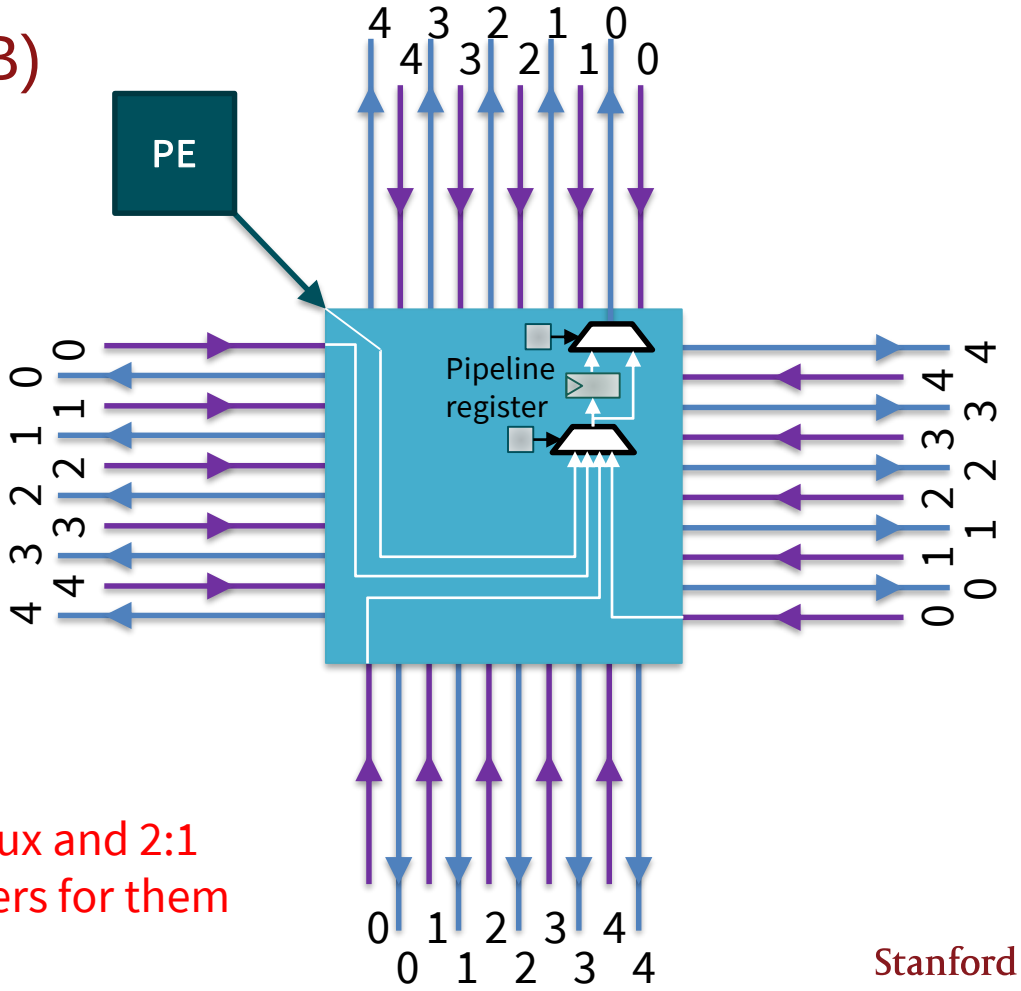
Coarse Grained Reconfigurable Array (CGRA)



Let's zoom in…

Processing Element (PE)
Memory (MEM)
Switch box (SB)
Connection box (CB)

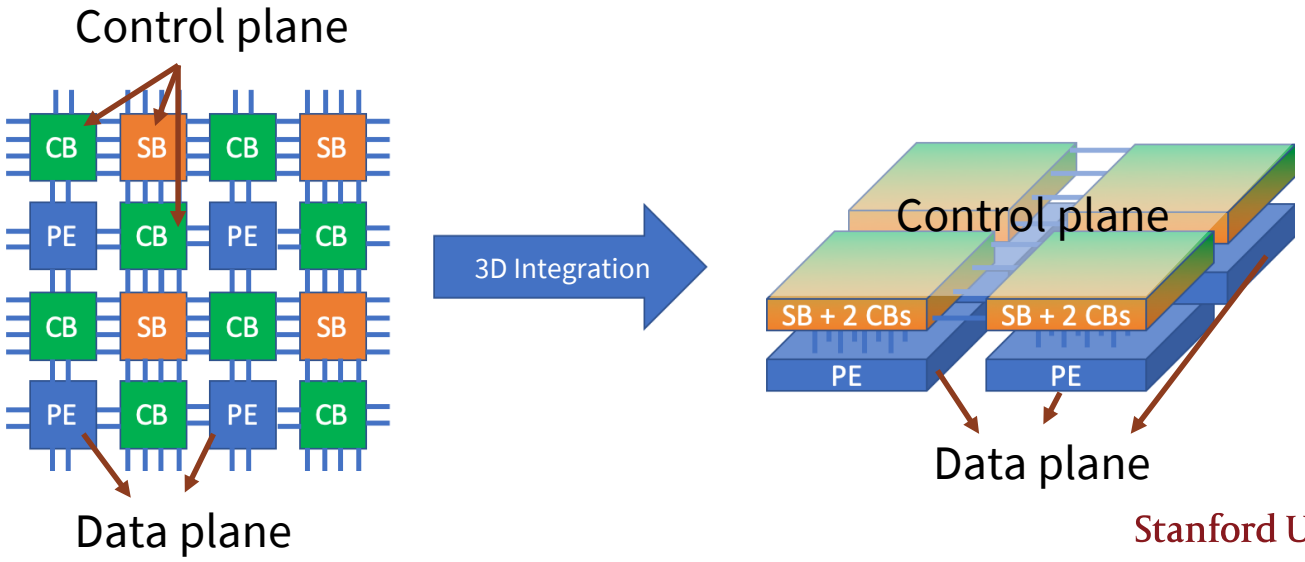
Switch Box (SB)



Each track needs 4:1 mux and 2:1 mux and control registers for them

3D integration for CGRA control plane

- Overhead comes from control plane components
- 3D advantages: less silicon area, shorter wire lengths, heterogeneous
- Control plane components are less sensitive to write speed
- Control/data plane: different physical planes of integrated 3D design



Emerging nanotechnologies for reconfigurability

CGRAs:

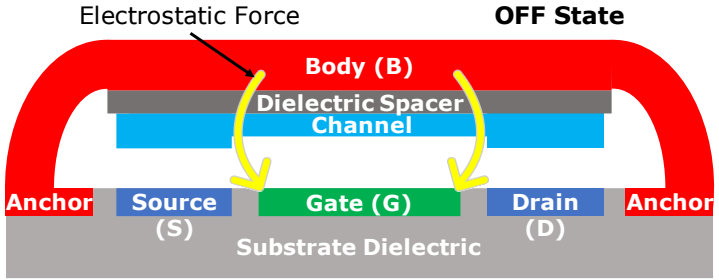
- Reconfigurable architecture with multi-bit datapaths and wide ALUs

NEM relays:

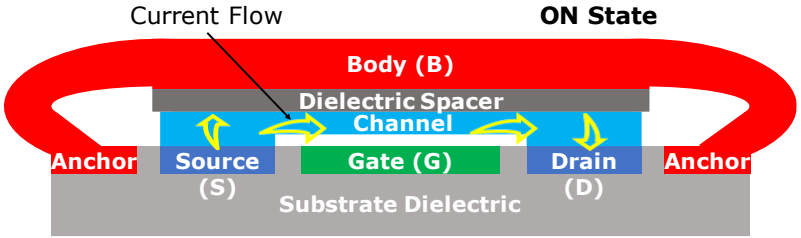
- Nano-scale mechanical switches monolithically integrated in 3D
- Reduce area/power/delay overhead from reconfigurable interconnects

Nanoelectromechanical (NEM) Relays

- Nanoscale mechanical relays
- Electrically actuated by Coulomb force between body and gate
- Can have four terminals, like MOSFET (G, D, S, B)
- Current flow between source and drain is controlled by potential across body and gate



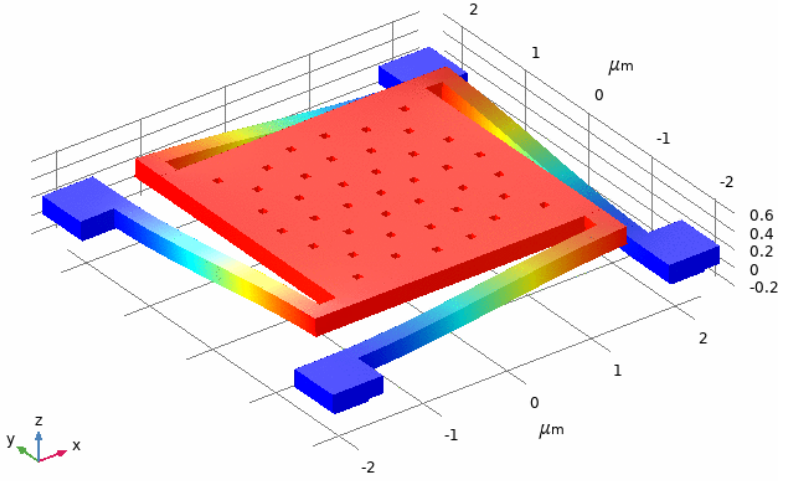
- Behaves like pass gate
  - No threshold voltage drop
- Only 5-7 mask layers are required @ below 400℃



Nanoelectromechanical (NEM) Relays

- Nanoscale mechanical relays
- Electrically actuated by Coulomb force between body and gate
- Can have four terminals, like MOSFET (G, D, S, B)
- Current flow between source and drain is controlled by potential across body and gate

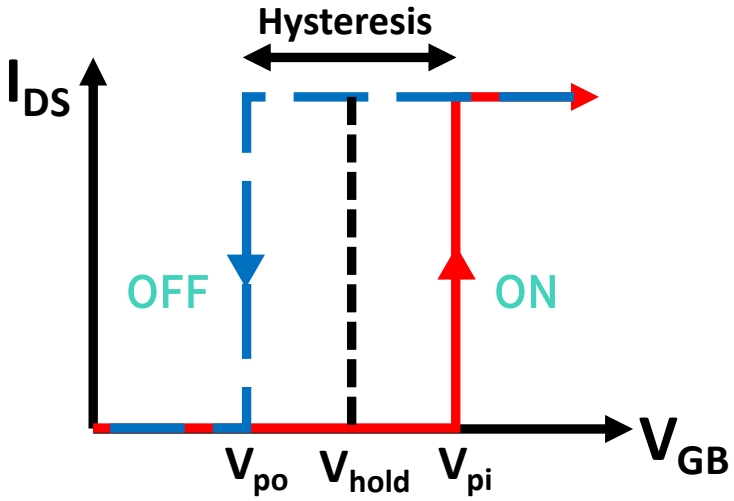
- Only 5-7 mask layers are required during fabrication
  - Temperatures below 400℃



3D view

NEM Relay I-V Characteristic

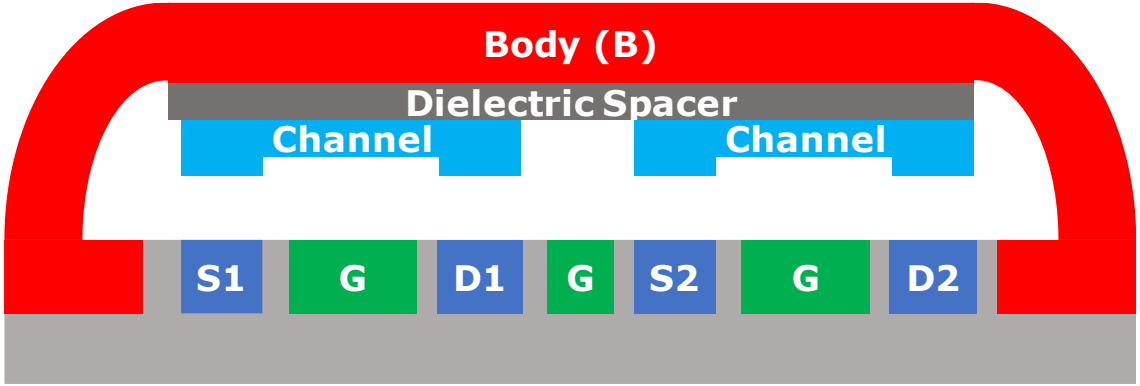
- Zero OFF-state/subthreshold leakage
- Abrupt ON/OFF behavior
- Low ON resistance
- Lower pull-out voltage due to mechanical instability & stiction
- Hysteresis
  - State memory
  - Need to ensure you have high enough voltage swing at gate to overcome hysteresis



Multi-Pole NEM Relay

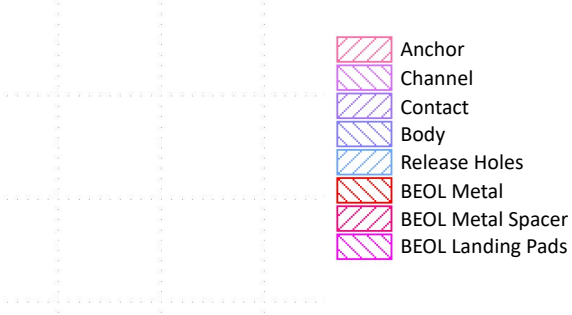
- Multiple channels connecting different source-drain pairs
- Share same gate and body

Insight: multi-pole NEM relay can route *multi-bit digital signals*

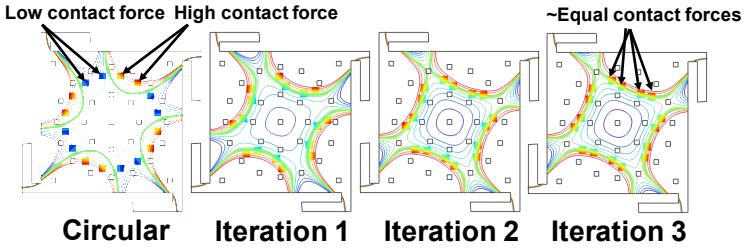


Contributions of Our Work

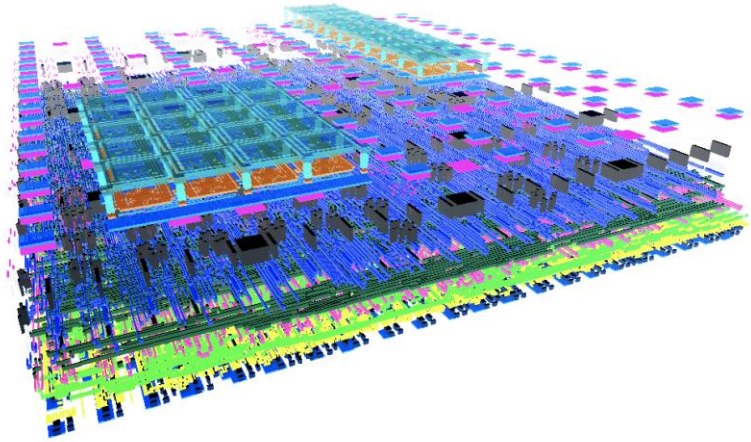
1) Improved multi-pole NEM relay design



2) Methodology for multi-pole contact placement

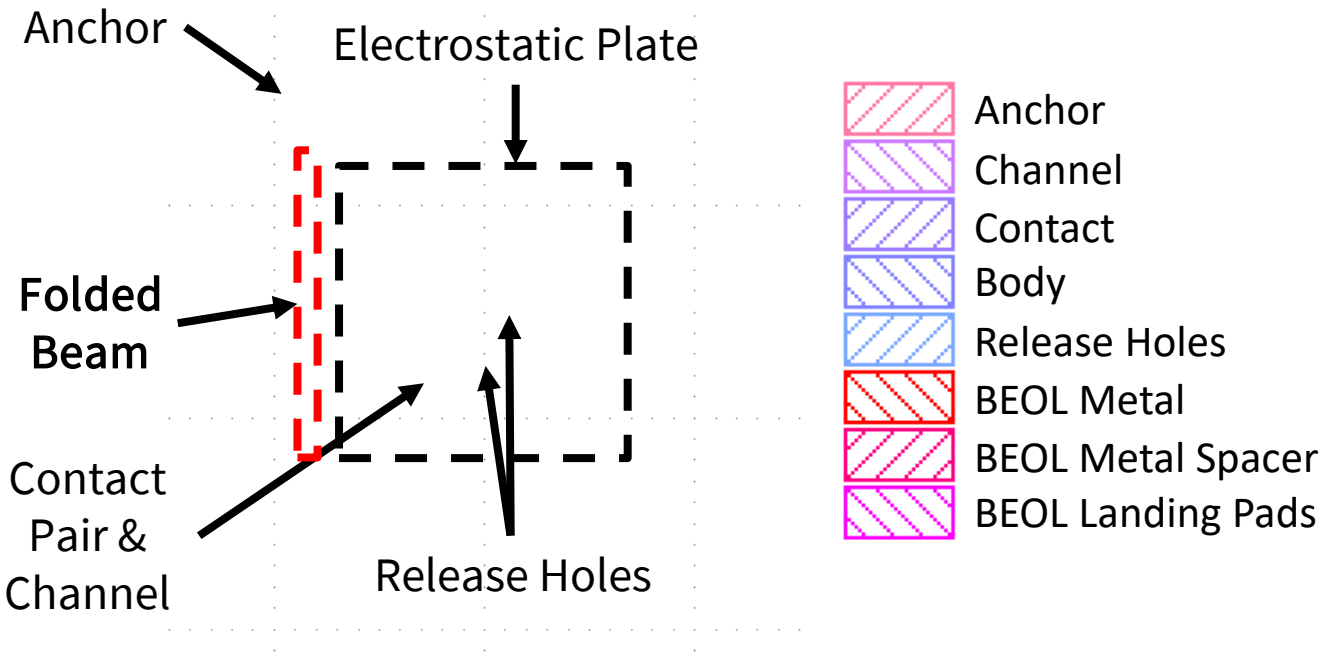


3) CMOS-NEMS integration methodology and demo of hybrid CMOS-NEMS CGRA PE tile



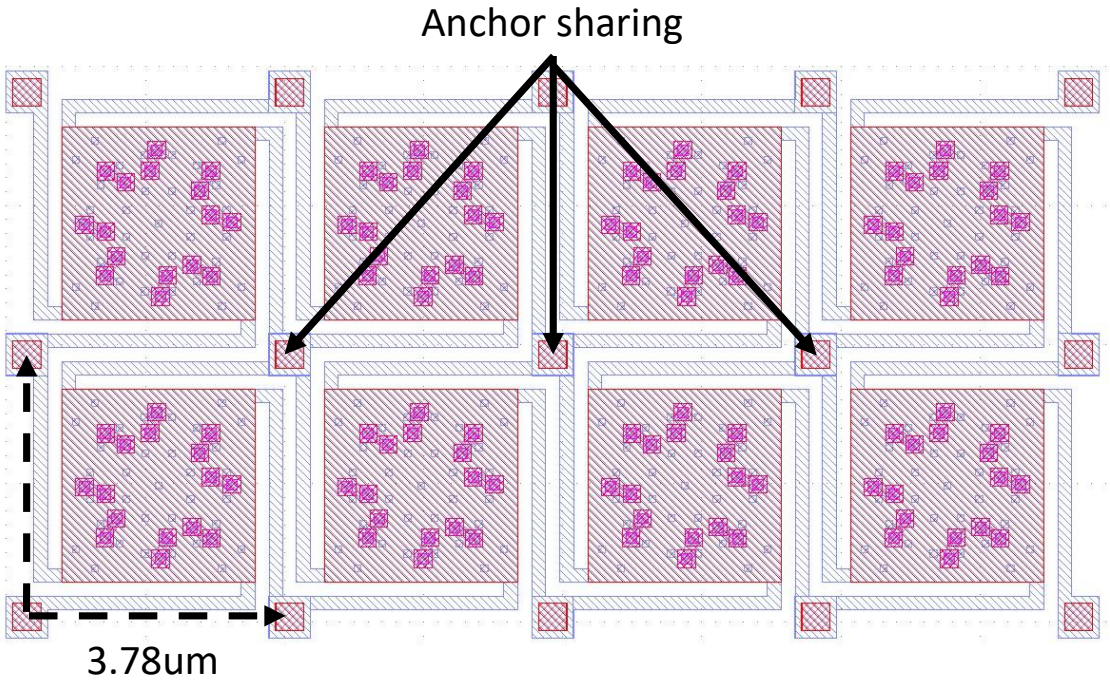
Multi-Pole NEM Relay Design

Folded beams for space saving without increasing spring constant



Multi-Pole NEM Relay Design

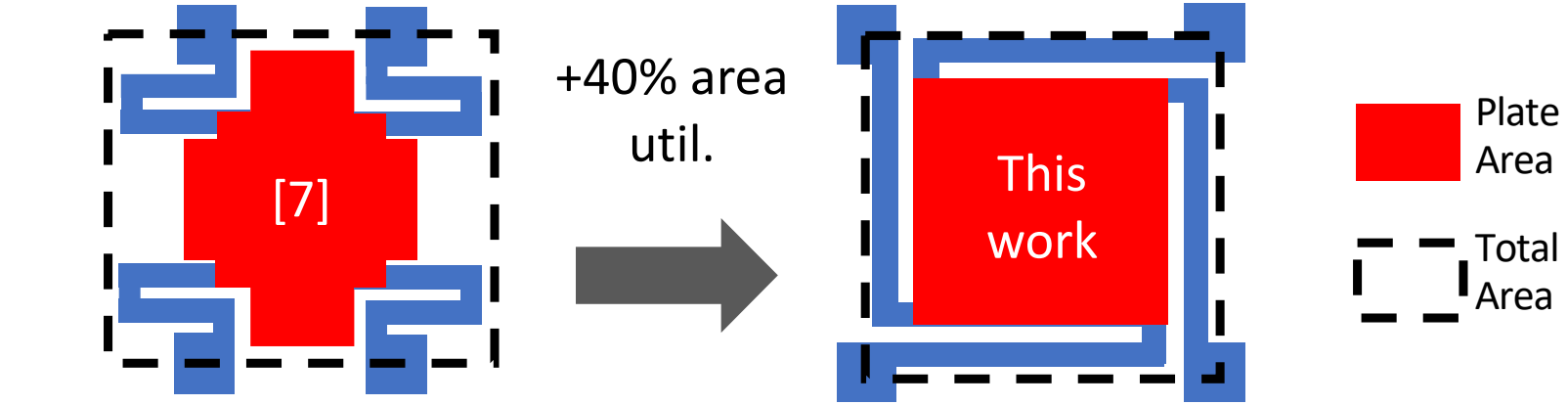
Anchor sharing for reducing area overhead over relay mosaic



Multi-Pole NEM Relay Design

Result: area utilization is improved over prior design

Footprint comparison



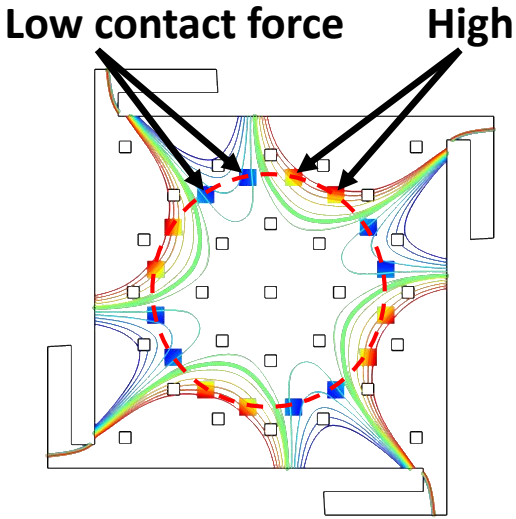
Area utilization: 0.33

Area utilization: 0.54

Multi-Pole NEM Relay Contact Placement

- **Multi-pole relays:** where to place the contacts?
  - Contacts need to touch down simultaneously
  - If misplaced, huge increase in contact resistance for some contacts

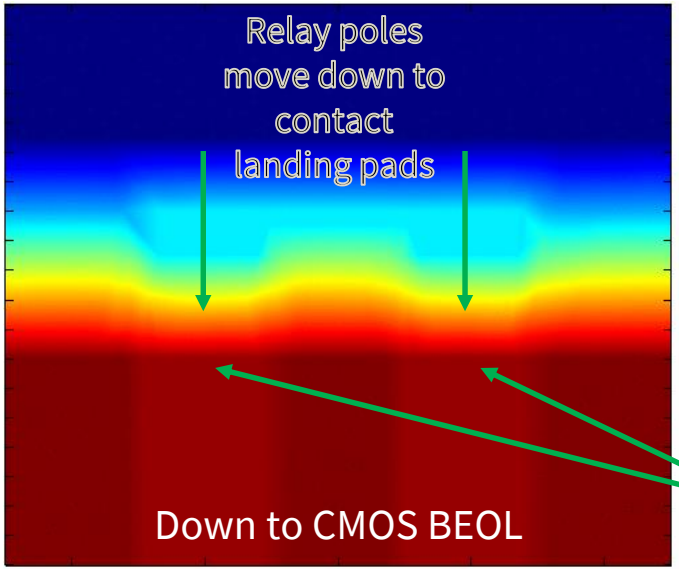
- **Example:** circular placement has huge variation in contact resistance across contacts
- *How do we reduce variation?*



Circular Contact Placement

Multi-Pole NEM Relay Contact Placement

Simulated pull-in process

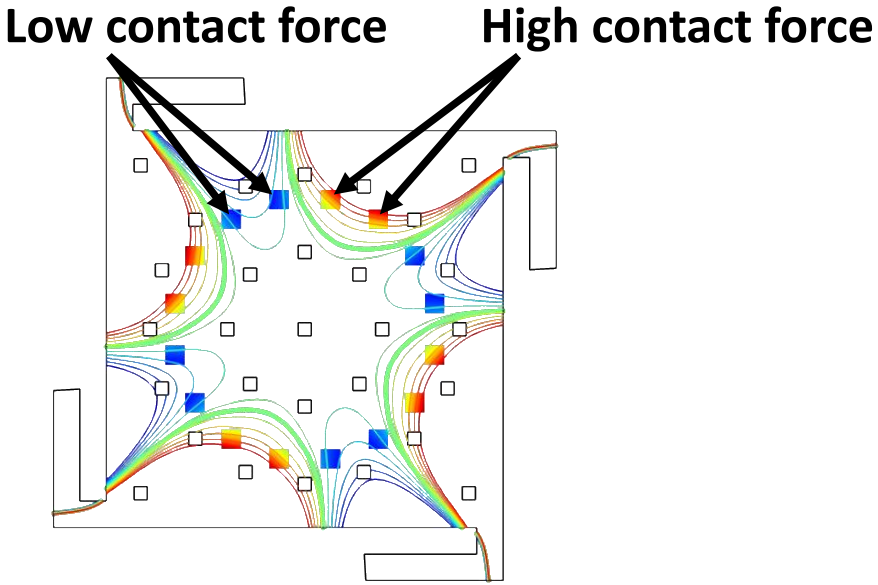


Multi-Pole NEM Relay Contact Placement

- **Insight:** contacts need to be placed on roughly the same displacement contour
- **Iterative Contact Placement (ICP)**

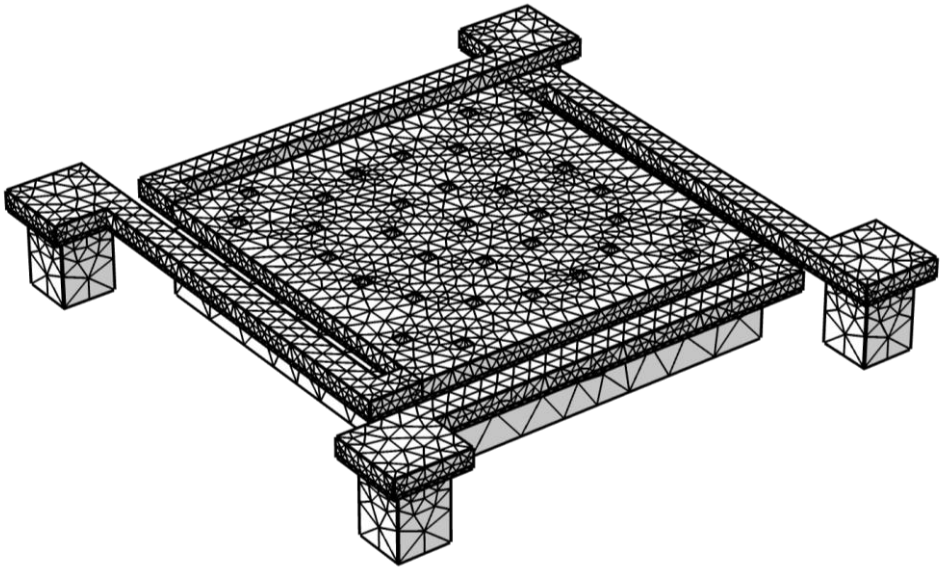
Multi-Pole NEM Relay Iterative Contact Placement

- **Step 1:** Initialize contact placement (ex: circular placement)



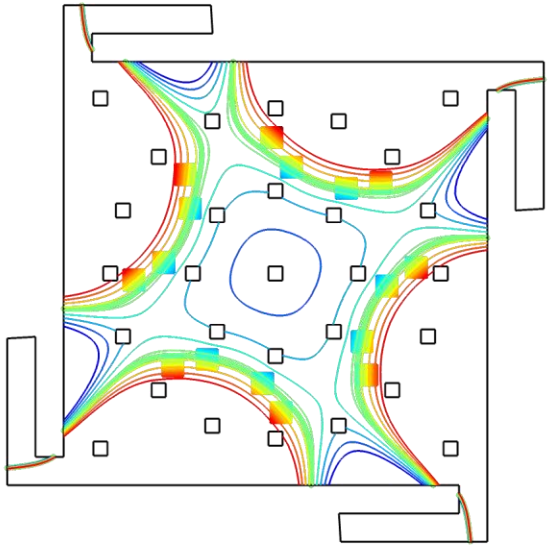
Multi-Pole NEM Relay Iterative Contact Placement

- **Step 2:** Perform finite element analysis to determine displacement contour



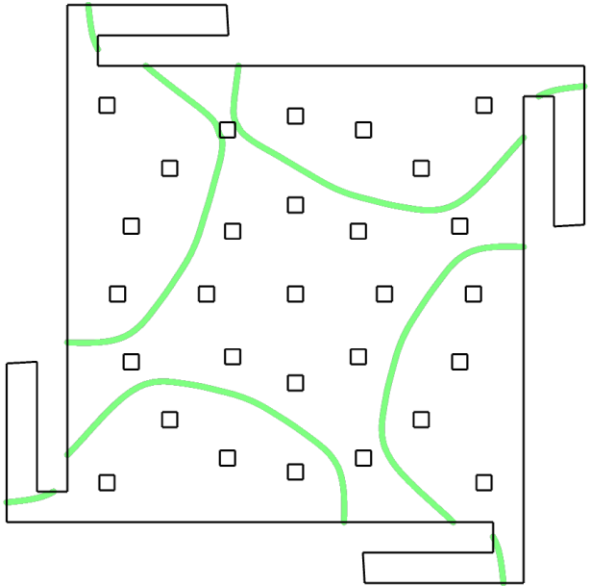
Multi-Pole NEM Relay Iterative Contact Placement

- **Step 3:** Find contact forces and resistances at operating voltage. Check if variation is below target value



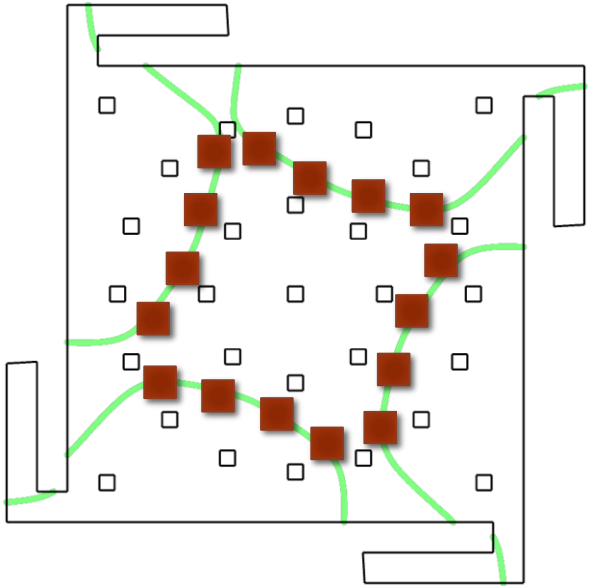
Multi-Pole NEM Relay Iterative Contact Placement

- **Step 4:** Find displacement contour at operating voltage



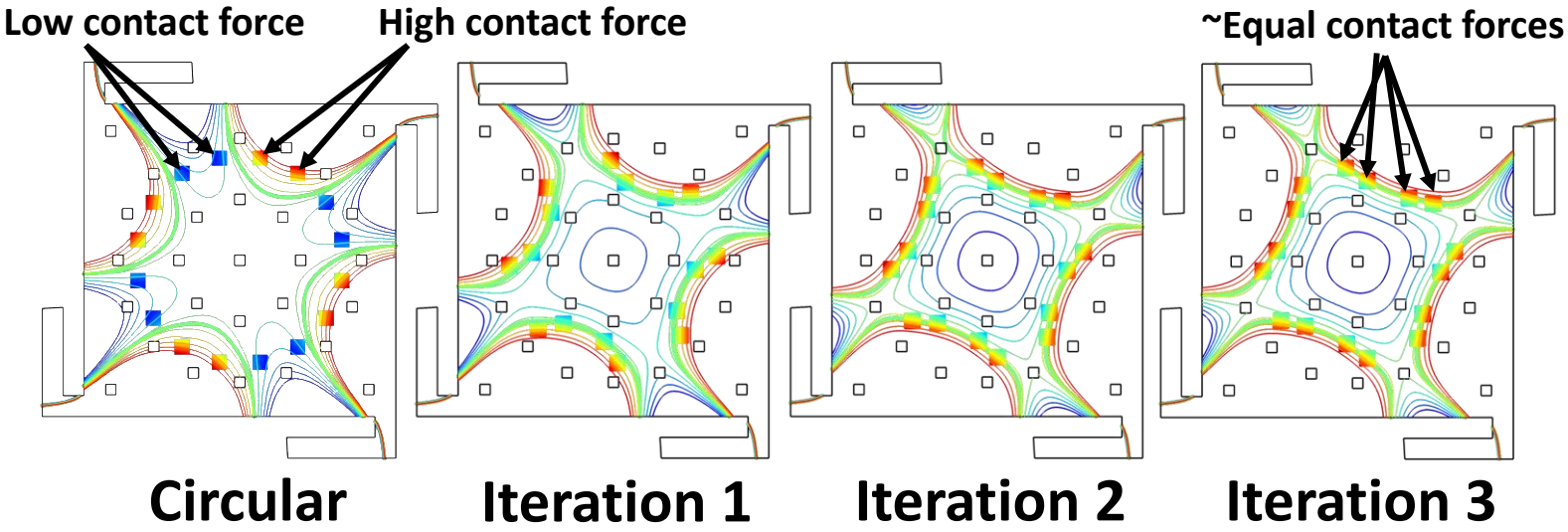
Multi-Pole NEM Relay Iterative Contact Placement

- **Step 5:** Pick points along displacement contour to place contacts in next iteration, and go back to step 2



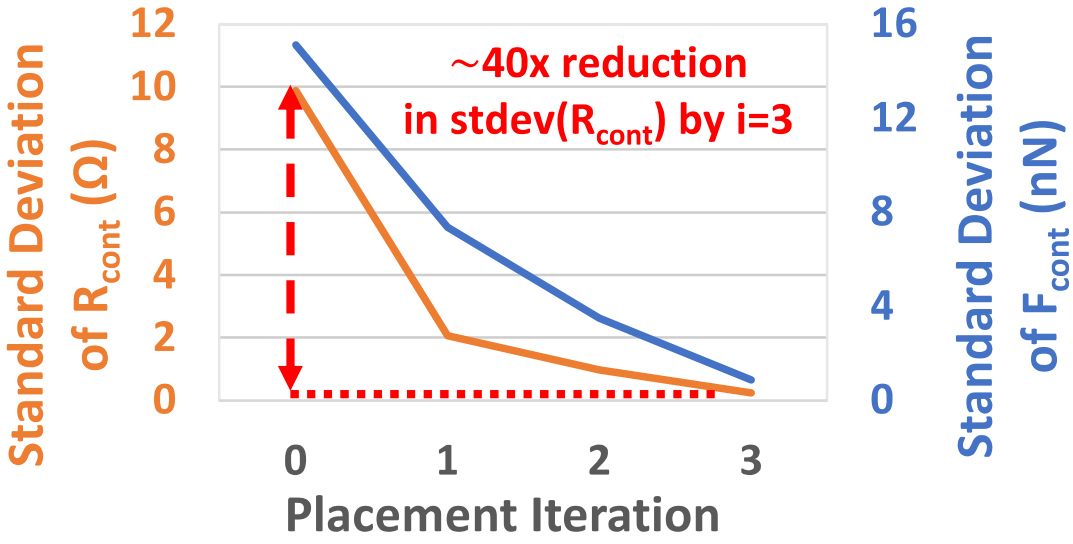
Multi-Pole NEM Relay Iterative Contact Placement

- Converges to very low contact resistance variation in just 3 iterations



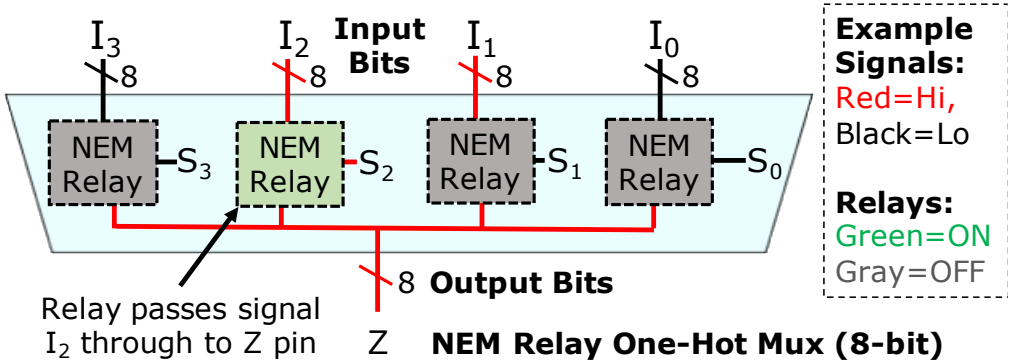
Multi-Pole NEM Relay Iterative Contact Placement

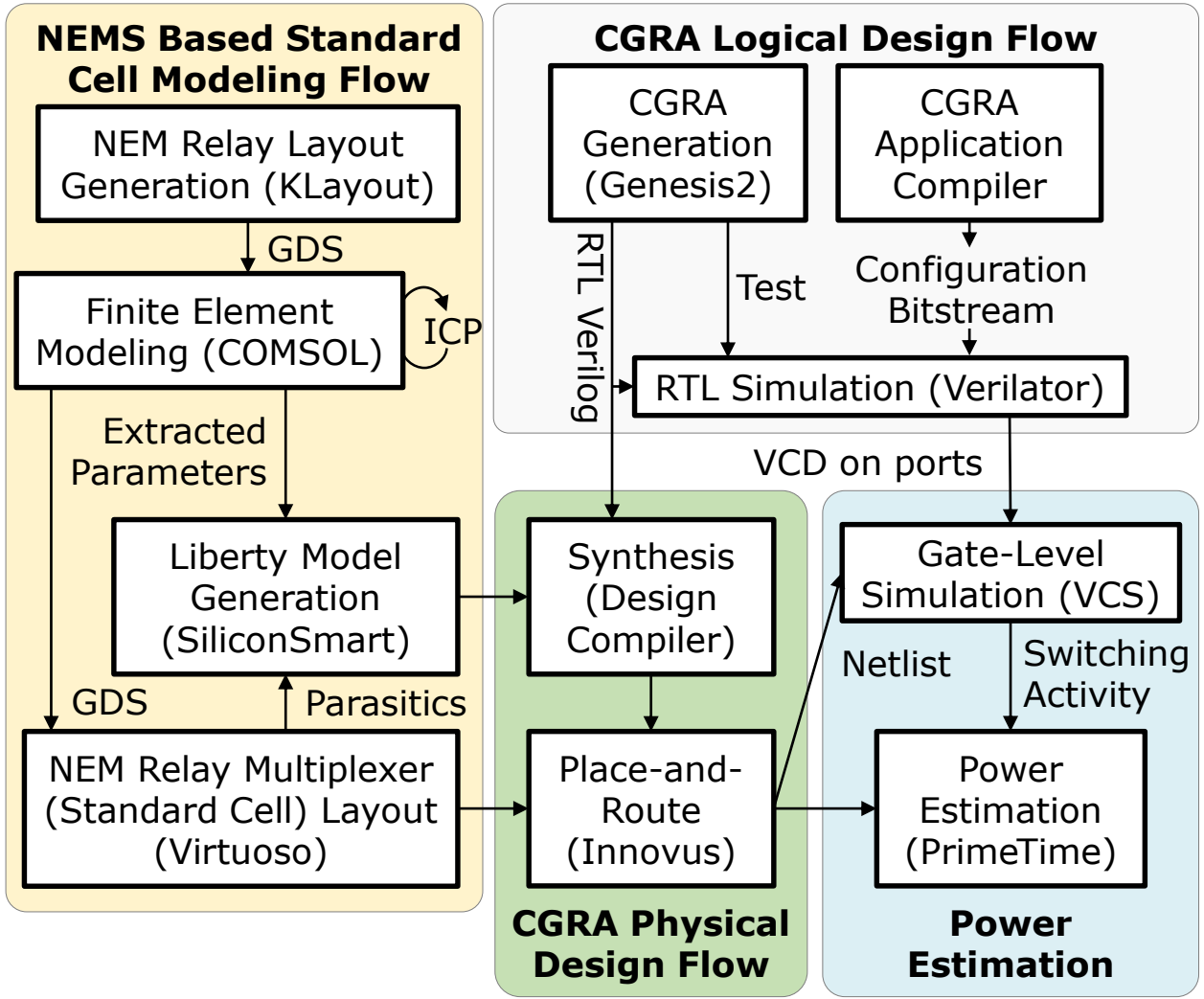
- Converges to very low contact resistance variation in just 3 iterations



Integration Methodology: Multiplexers

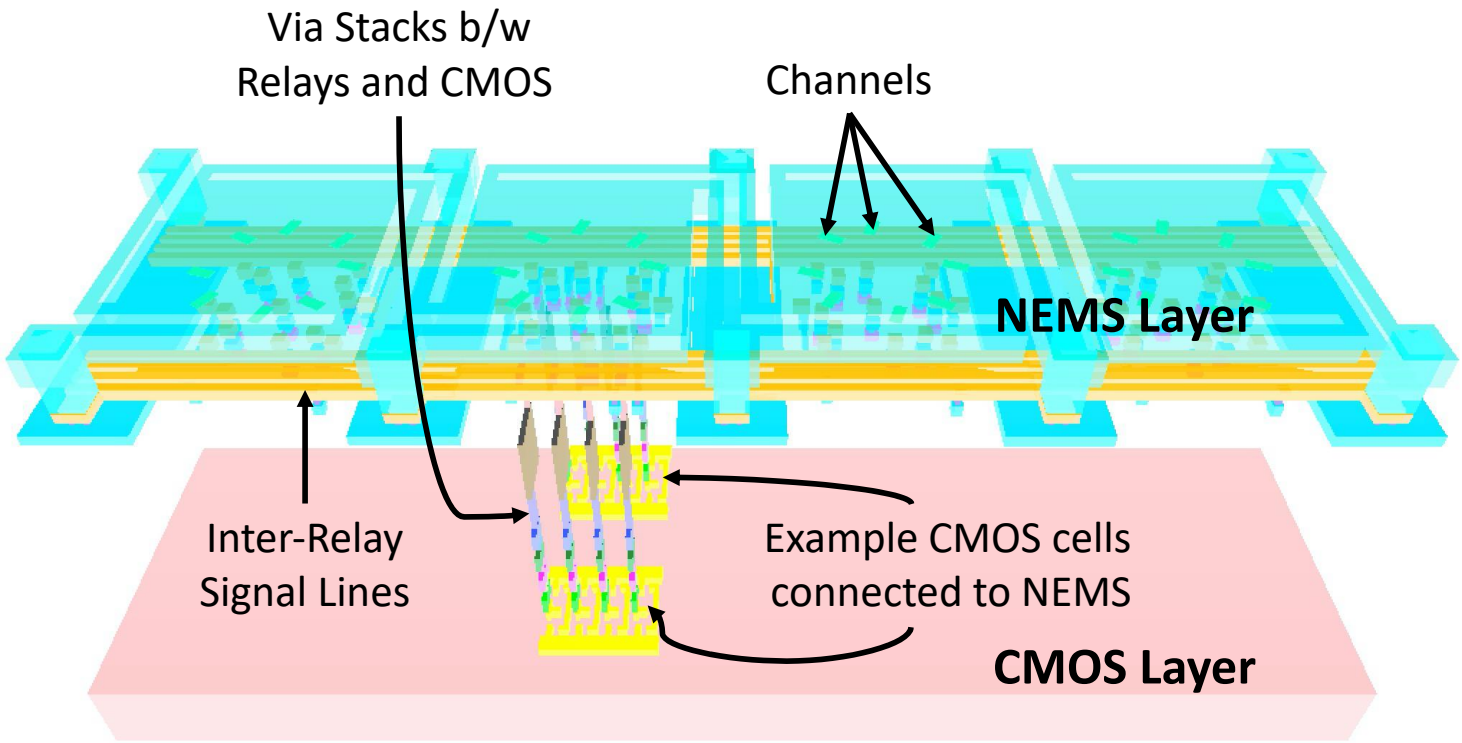
- How to compose relays into efficient multiplexers?
- **Strategy: One-Hot Multiplexing**
- Each relay takes S_0-S_{N-1} (one-hot)
- Each relay takes I_0-I_{N-1} (8 bits wide each)
- When S_i is on, $I_i[7:0]$ is electrically connected to $Z[7:0]$



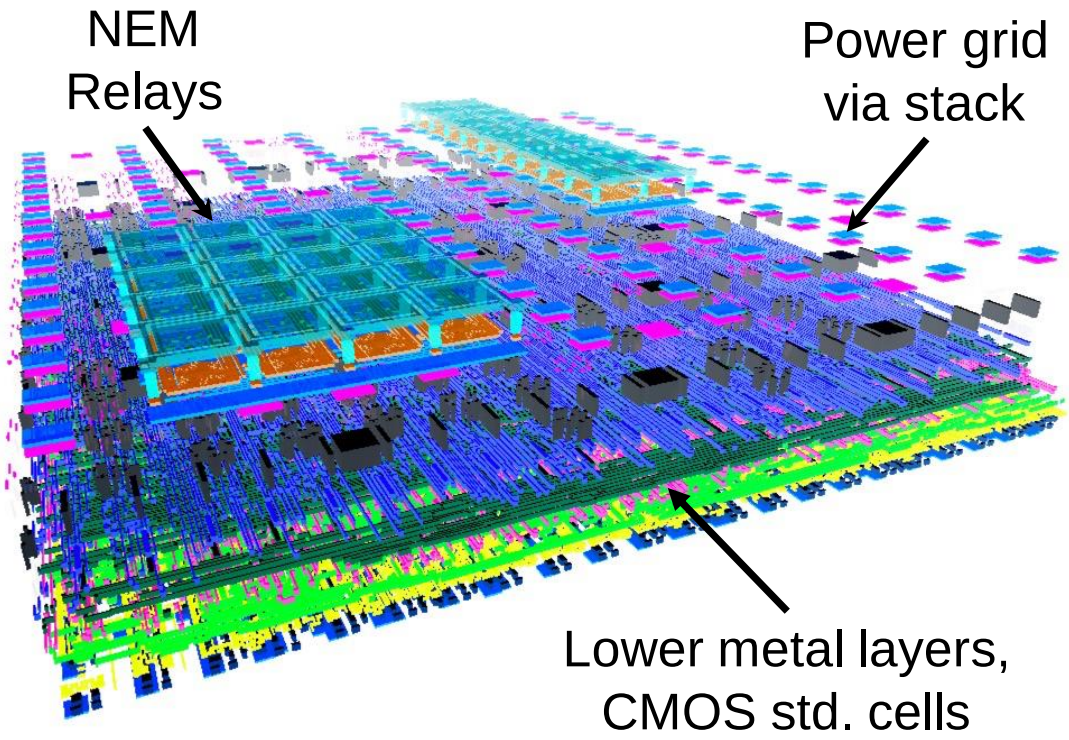


Integration Methodology: Design Flow

Hybrid CMOS-NEMS Multiplexer



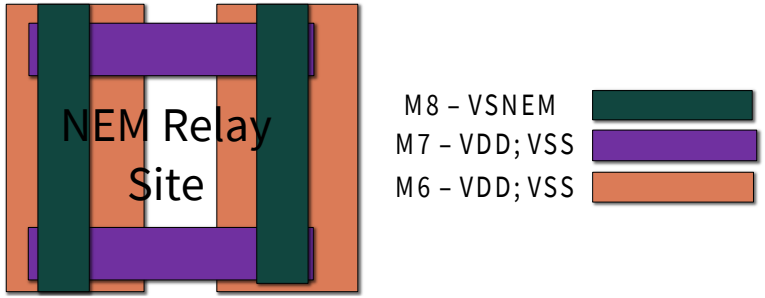
Hybrid CMOS-NEMS CGRA PE Tile



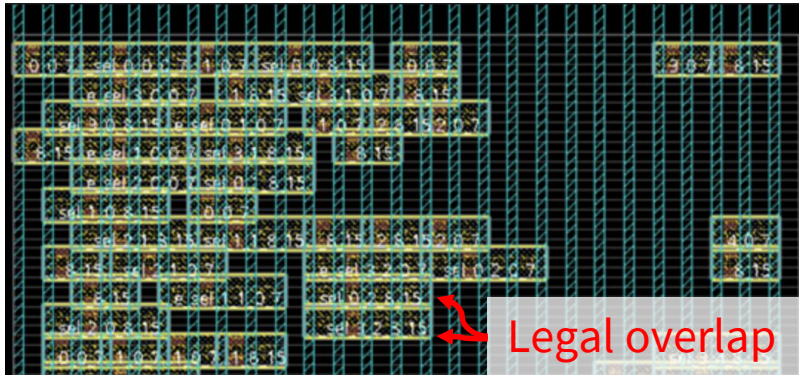
Integration Methodology: Design Flow Challenges

Challenges:

- Power grid for NEM relay body bias should not disrupt anything
- Layout and power grid must be designed concurrently to ensure compatibility (DRC)



- CMOS cells need to be placed below NEMS layer
  - Treat NEMS as cover cells
  - Custom legalization script
    - Snap to grid
    - Auto-fix overlapping cells

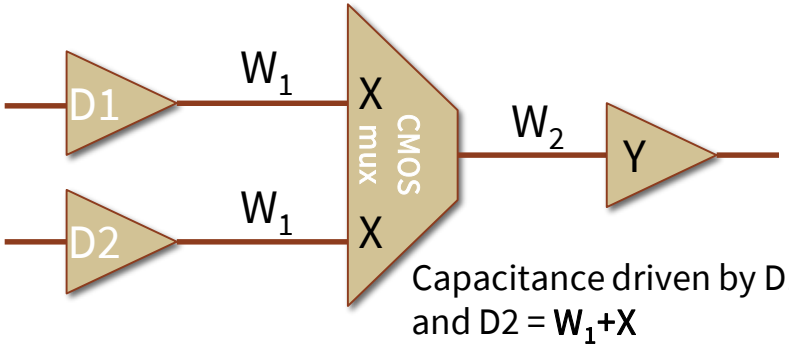


Integration Methodology: Design Flow Challenges

Challenges:

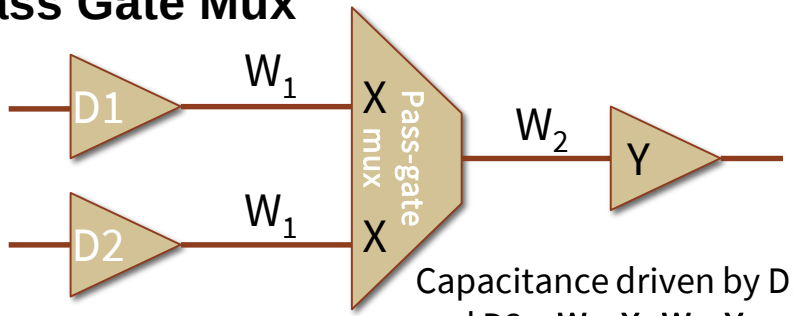
- Power/delay estimation
  - NEM relays need to be treated as *pass gates*
  - Need to propagate parasitic capacitances back to driving cells
  - Essentially introduce pass gate logic into power/delay estimation flows
  - Developed scripts to automate integration of pass gate logic

Typical CMOS Mux



Capacitance driven by D1 and D2 = W_1+X

Pass Gate Mux



Capacitance driven by D1 and D2 = W_1+X+W_2+Y

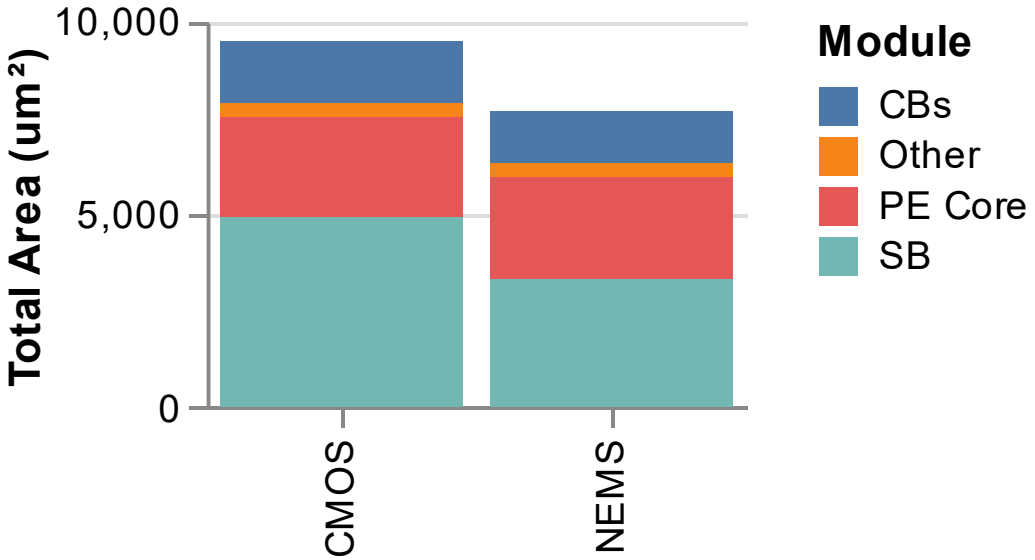
Results: Hybrid CMOS-NEMS CGRA PE Tile

- **Constrained for iso-delay:** 5 ns
- **Area Benefit:** 19.2%
- **FOM:** $\text{Area FOM} = \frac{\text{CMOS Design Stdcell Area} - \text{NEMS Design Stdcell Area}}{\text{SB/CB Mux Area}} = 73.3\%$

- **Image Processing Applications Tested (for Power):**
  - 3x3 convolution
  - Cascaded convolution
  - Harris corner detector

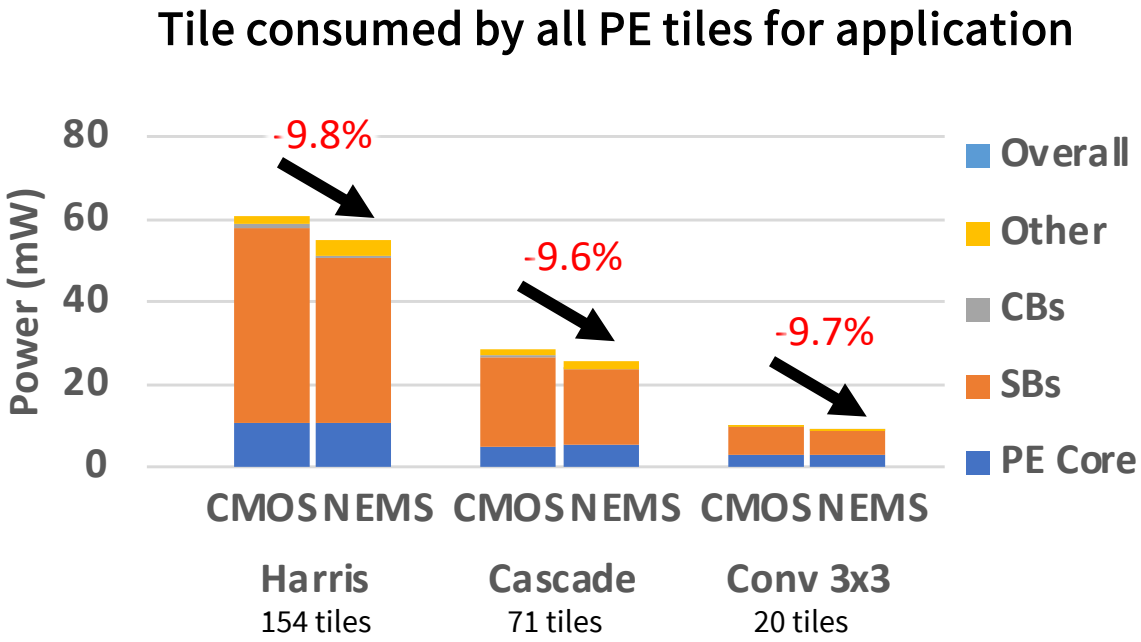
- **Average Power Benefit across Applications:** 9.7%

Results: Hybrid CMOS-NEMS CGRA PE Tile Area



Most of the area reduction is in the switch box; a little in CB too

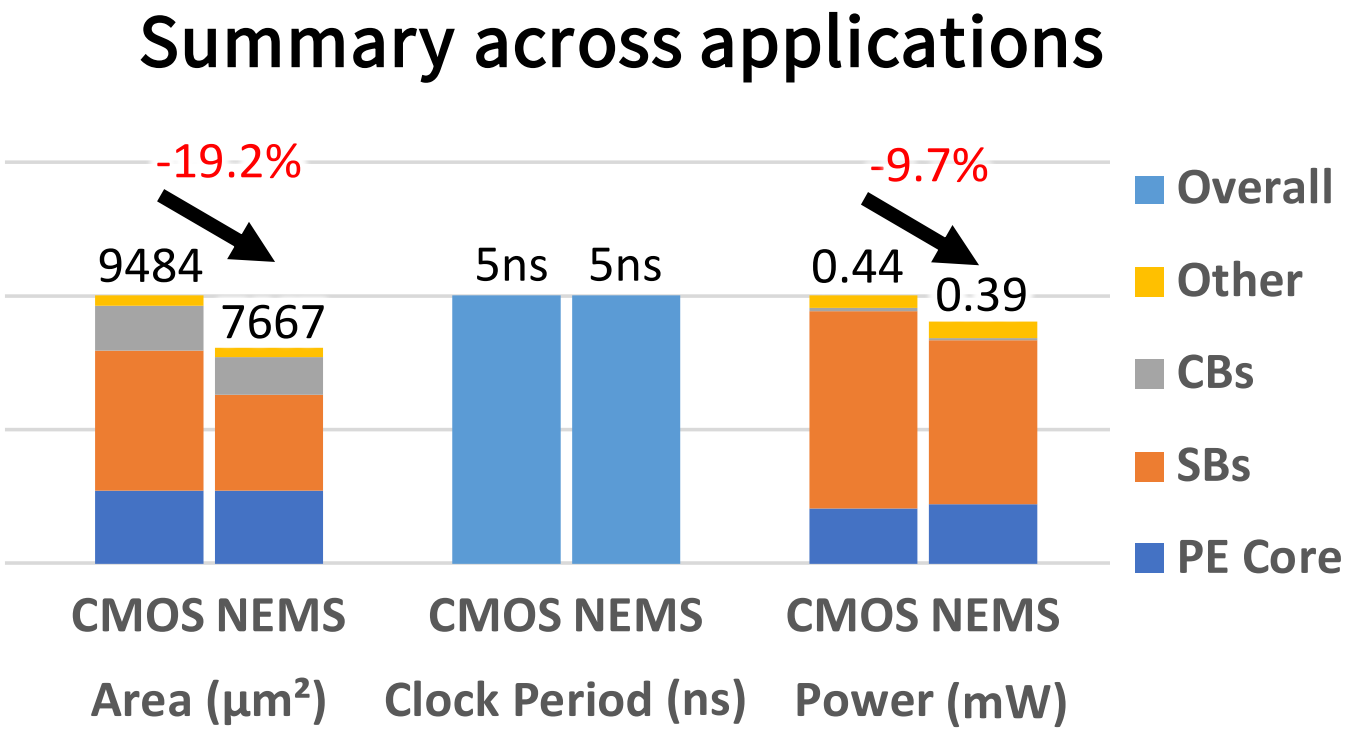
Results: Hybrid CMOS-NEMS CGRA PE Tile Power



Power benefit is similar across applications

Switchboxes dominate the PE tile power; CBs are almost zero

Results: Hybrid CMOS-NEMS CGRA PE Tile

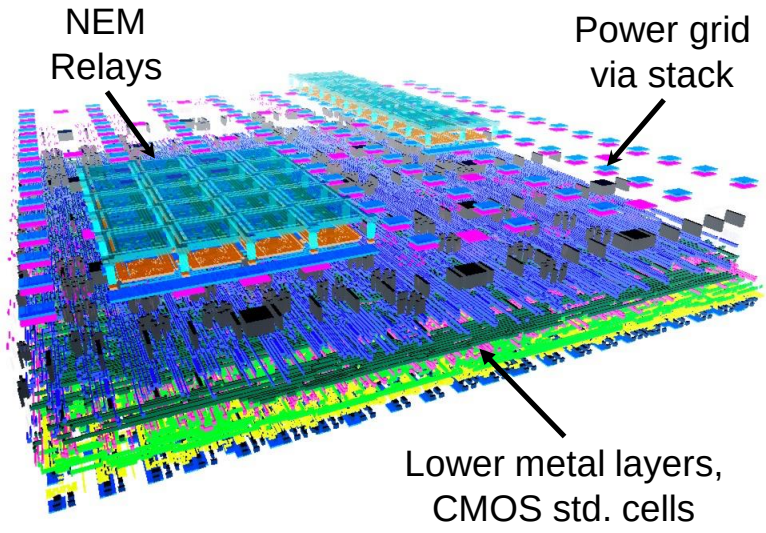


*Power is per-tile

Conclusion

- **Contributions:**
  1. Improved multi-pole NEM relay design
  2. Solving multi-pole contact placement
  3. CMOS-NEMS integration methodology & demo of hybrid CMOS-NEMS CGRA PE tile

- **Results on PE tile:**
  - Constrained for iso-delay
  - 19.2% area improvement
  - 9.7% power improvement



Thank you!

QUESTIONS?