Monolithic 3D FPGA Design and Synthesis with Back-End-of-Line Configuration Memories

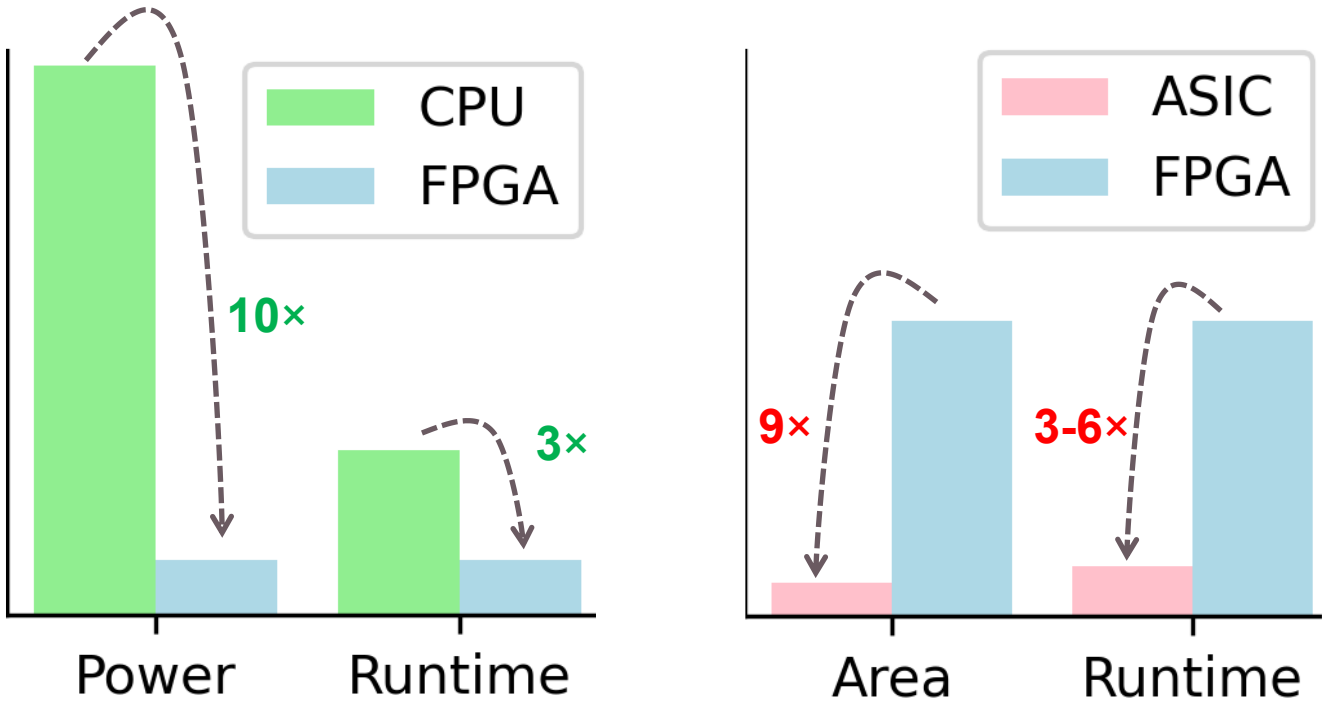
Faaiq Waqar[1], Jiahao Zhang[2], Anni Lu[1], Zifan He[2],
Jason Cong[2], Shimeng Yu[1]

[1]Georgia Institute of Technology; [2]University of California, Los Angeles

62 THE CHIPS TO SYSTEMS CONFERENCE

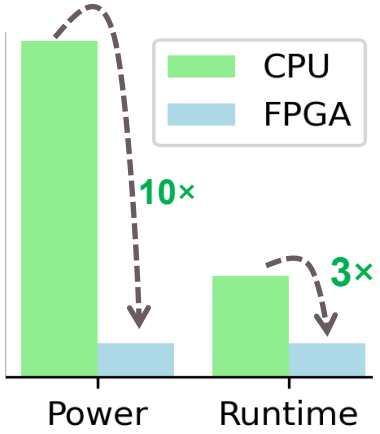
SPONSORED BY CEDA IEEE Council on Electronic Design Automation sigda acm

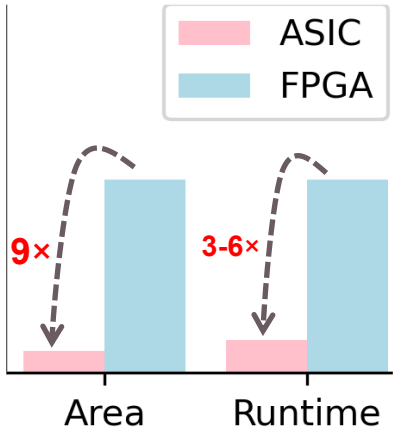
The Cost of Reconfigurability



- **FPGAs** can be desirably **reconfigured** for specialized execution
- Despite innovations, **FPGAs lag ASICs** in Power, Performance, and Area (PPA)

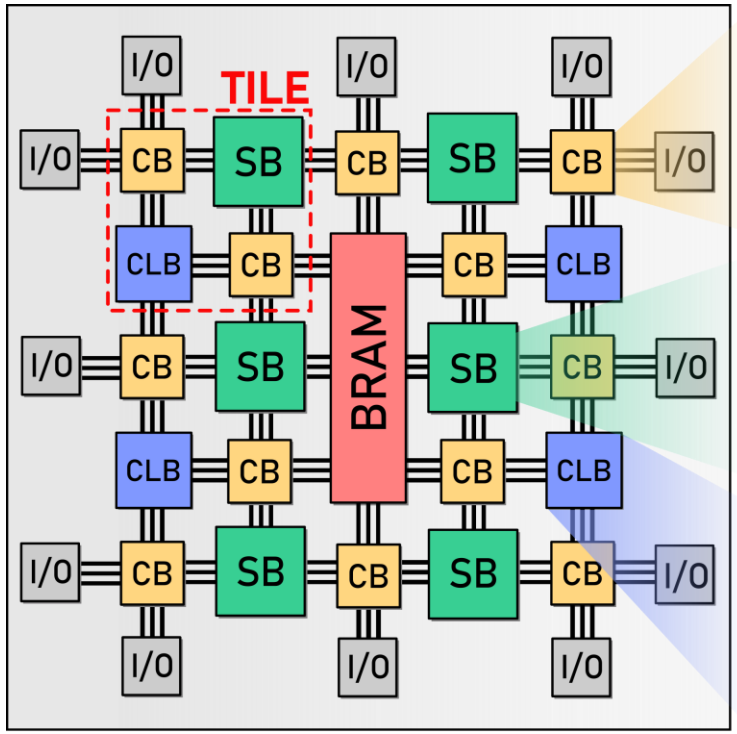
The Cost of Reconfigurability

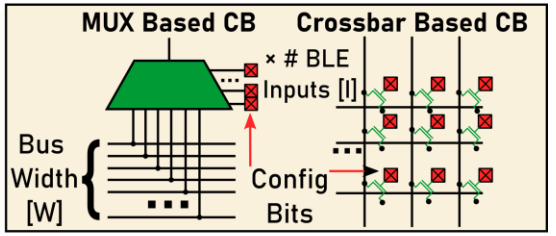


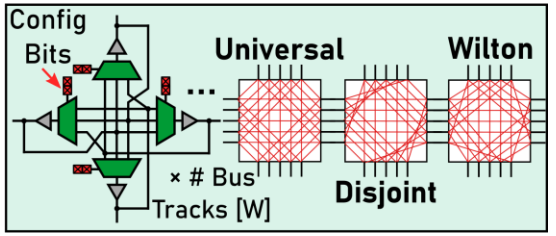


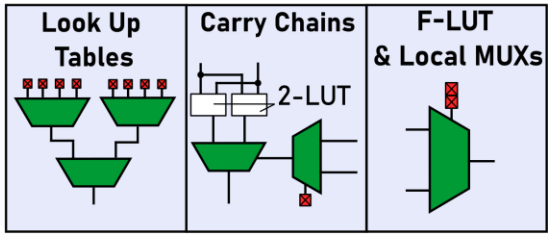
✓ Reconfigurable

✗ Large PPA Disparity









CBs connect signals on bus to logic (CLB) inputs.

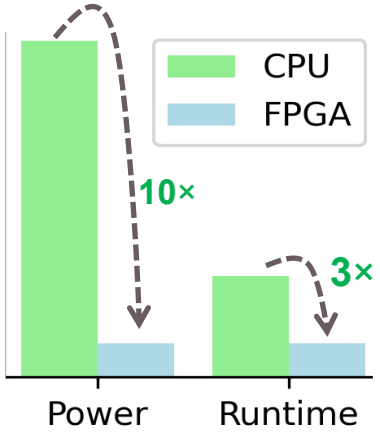
SBs steer signals at intersections to route inputs/outputs.

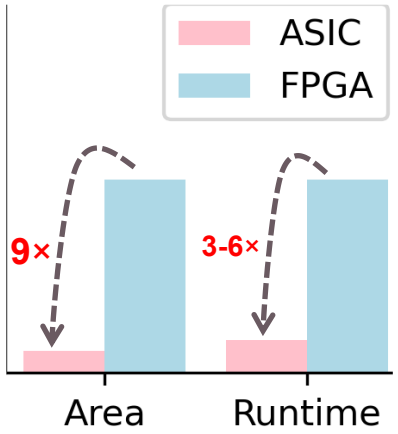
CLBs contain basic logic elements (BLEs) that emulate *K* input logical functions

Island-based FPGAs are composed of Configurable Logic Blocks (CLBs), Switch Blocks (SBs), and Connection Blocks (CBs).

A tile (2 CBs + 1 SB + 1 CLB) is a modular building block of an FPGA

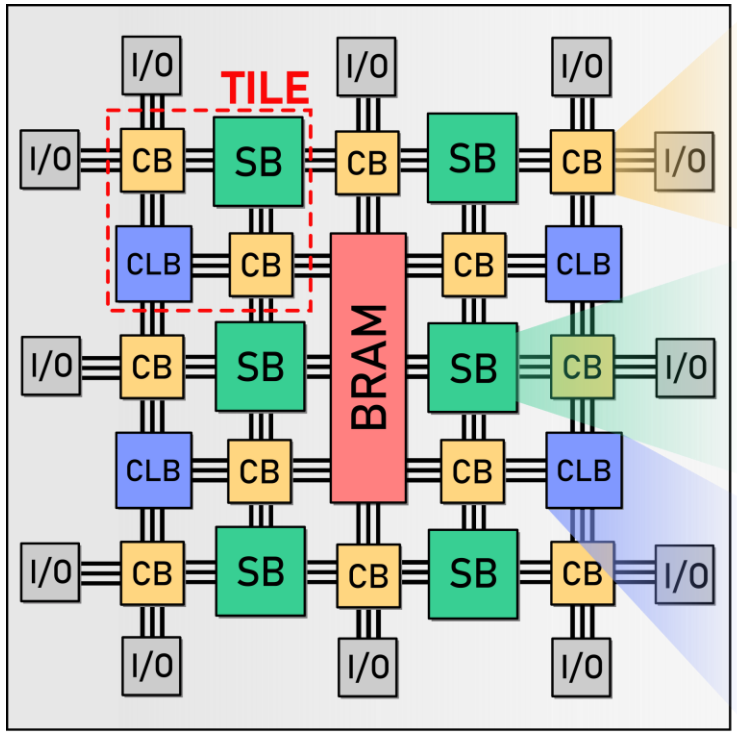
The Cost of Reconfigurability

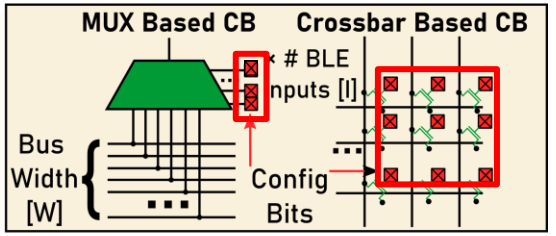


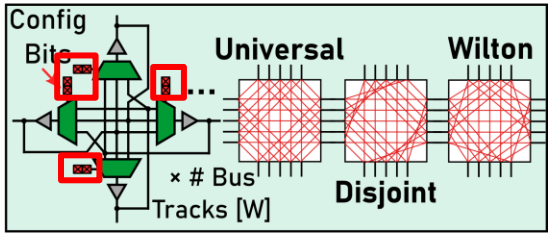


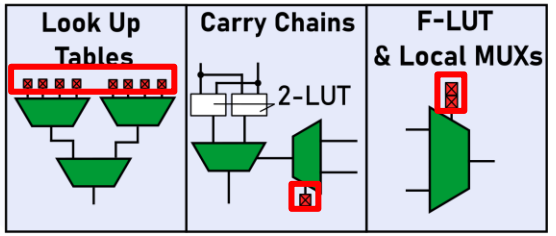
✓ Reconfigurable

✗ Large PPA Disparity









CBs connect signals on bus to logic (CLB) inputs.

SBs steer signals at intersections to route inputs/outputs.

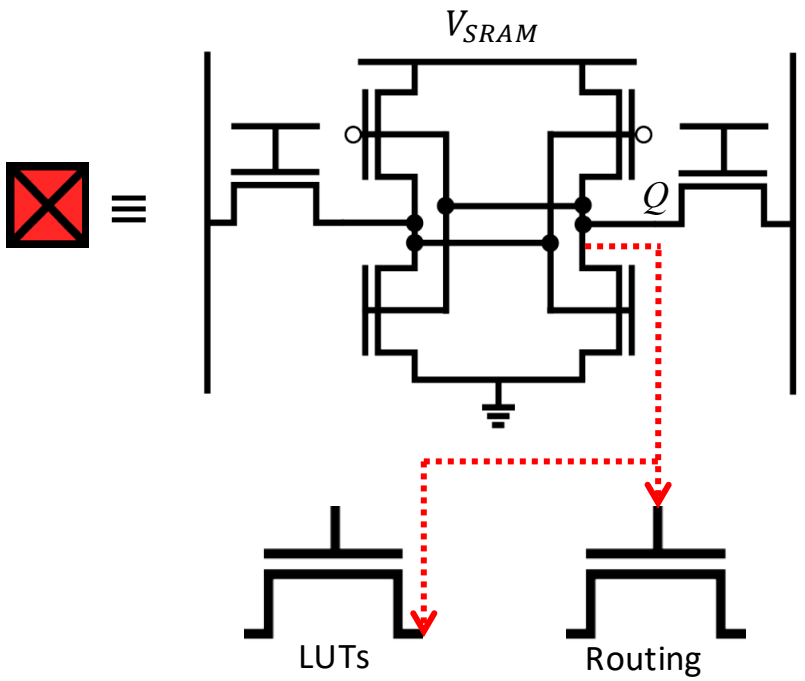
CLBs contain basic logic elements (BLEs) that emulate *K* input logical functions

Island-based FPGAs are composed of Configurable Logic Blocks (CLBs), Switch Blocks (SBs), and Connection Blocks (CBs).

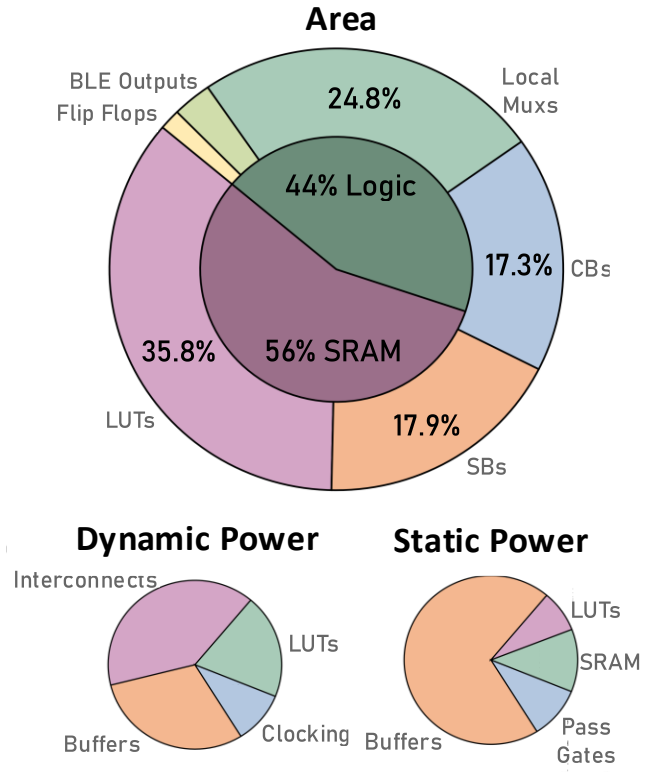
A tile (2 CBs + 1 SB + 1 CLB) is a modular building block of an FPGA

The Cost of Reconfigurability

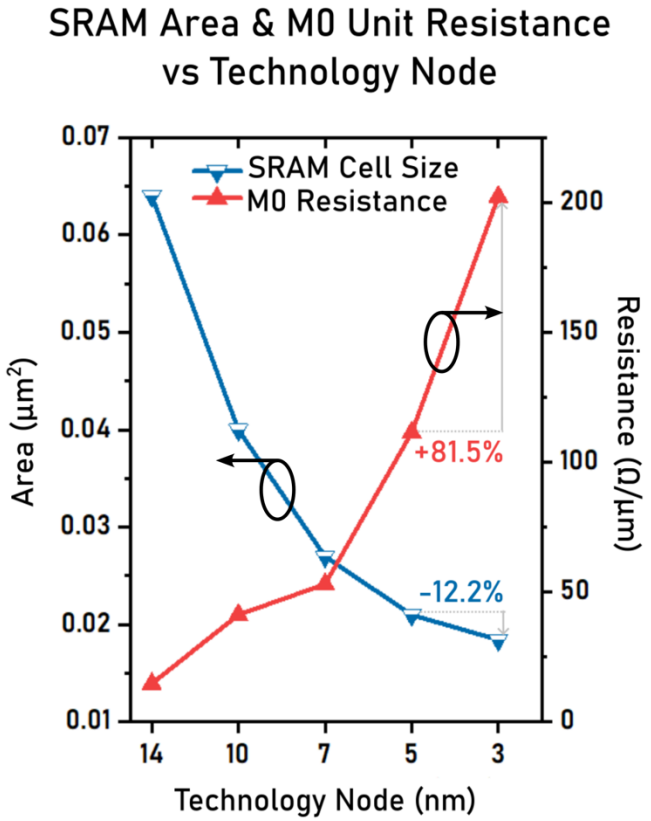
Configuration Bit (SRAM)



✗ Speed
✓ Density
✓ Stability
✓ Low Static Power

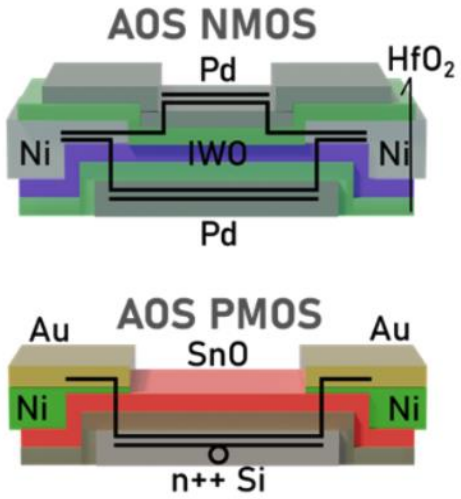


SRAM dominates area, and interconnects dominate the dynamic power



A problem **exacerbated by the limits of SRAM/RDL scalability**

Opportunity of AOS Channel Devices



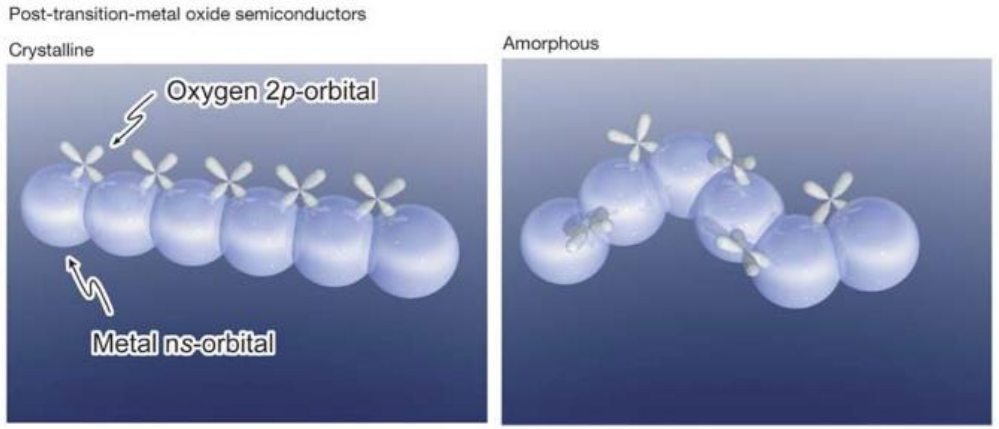
✓ Back-End-of-Line (BEOL) Compatible process (<400 ºC)

✓ Ultra-low leakage (<10^{-15} A/μm)

✓ Adequate electron/hole mobility (~20 $cm^2/V{\cdot}s$, ~2 $cm^2/V{\cdot}s$)

✓ Strong V_{th} Stability (BTI)

K. Nomura et al., *Nature* 2004



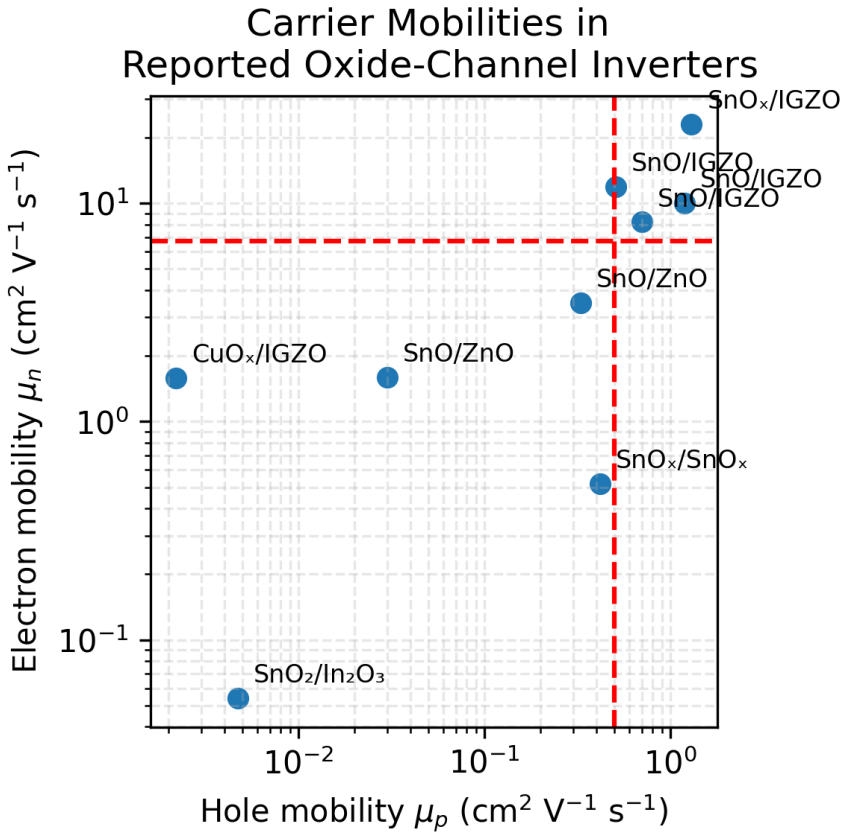
What makes it BEOL compatible?

- Equilibrium phase of solids is amorphous at low temperatures

➔ No high-temperature annealing

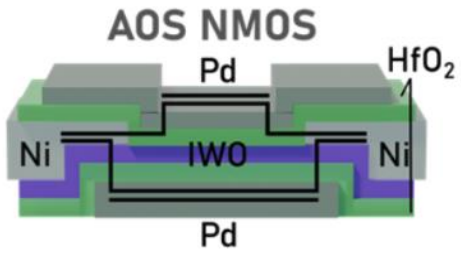
- Isotropic, large S-orbital conduction

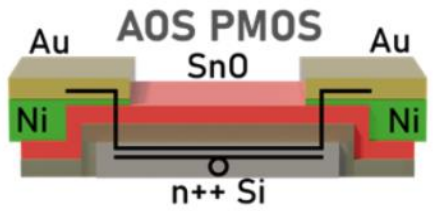
➔ Eases requirement on long-range order for conduction



Z-W. Shang et al., *Nanotechnology Reviews* 2019

Opportunity of AOS Channel Devices





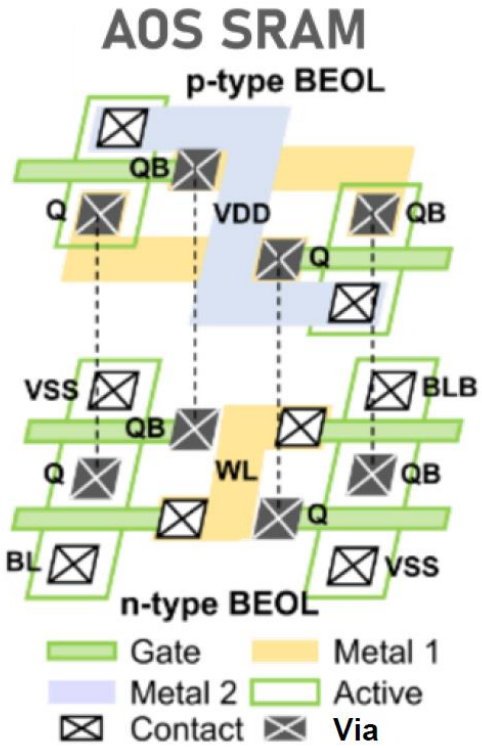
✓ Back-End-of-Line (BEOL) Compatible process (<400 ºC)

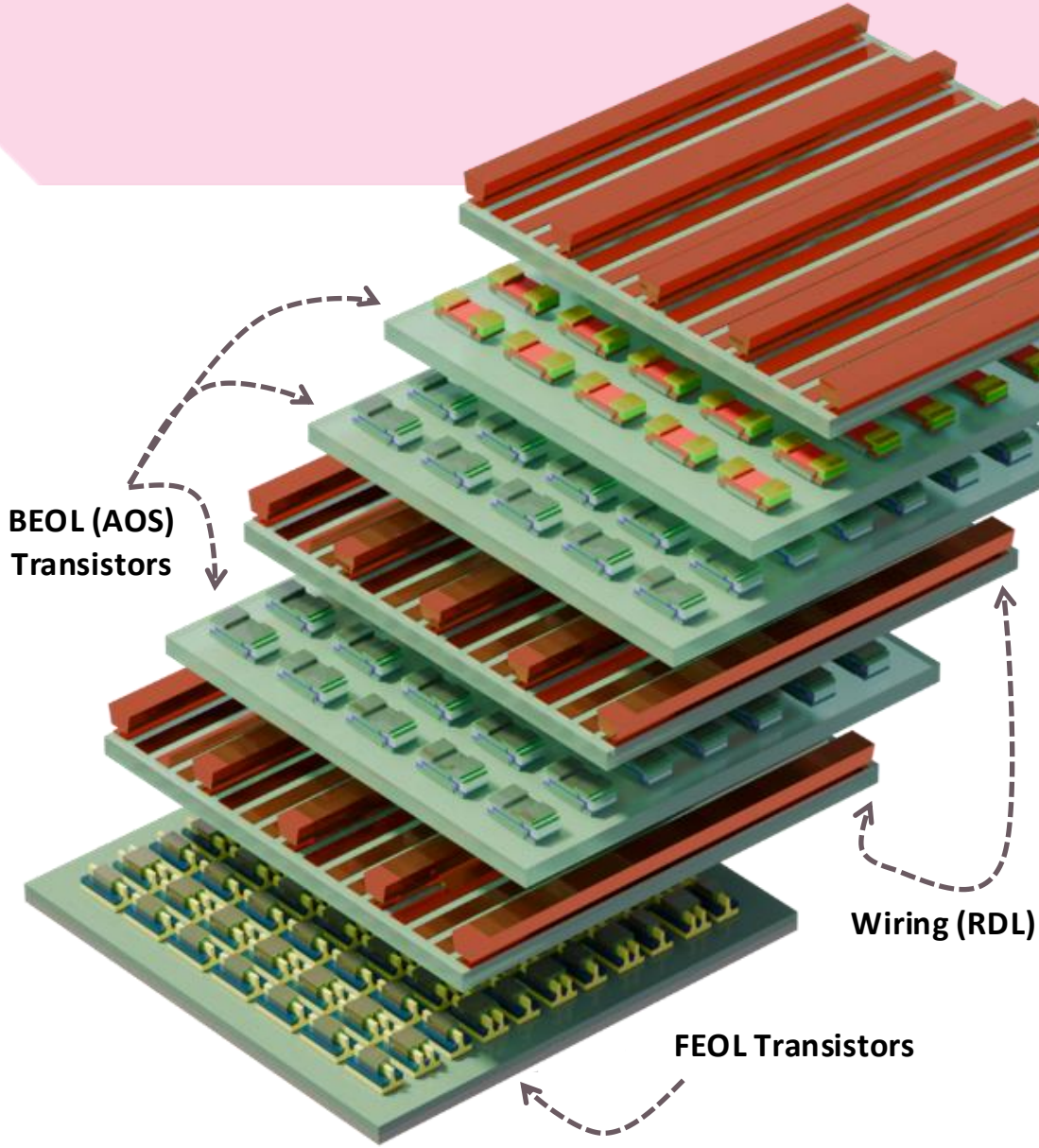
✓ Ultra-low leakage (<10^{-15} A/µm)

✓ Adequate electron/hole mobility (~20 $cm^2/V{\cdot}s$, ~2 $cm^2/V{\cdot}s$)

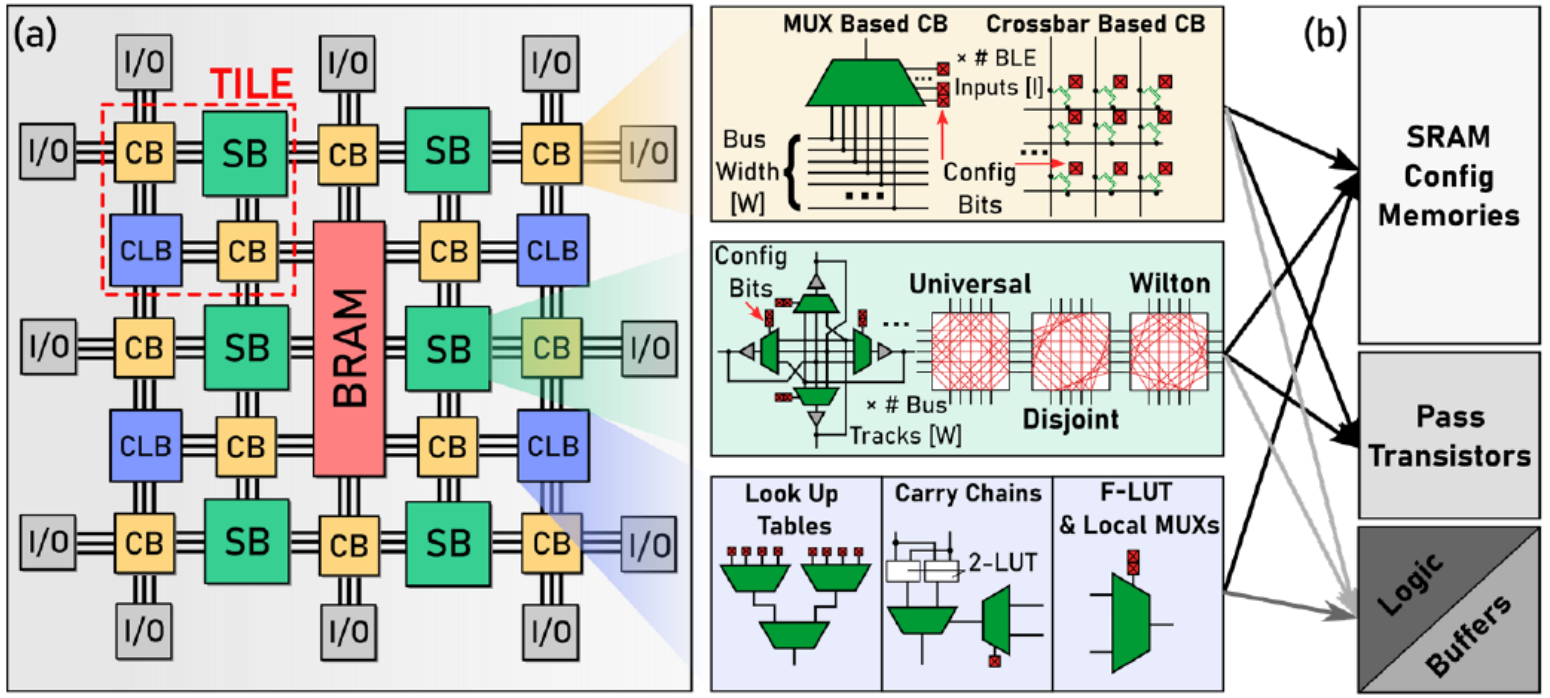
✓ Strong V_{th} Stability (BTI)

Leverage BEOL compatibility to build a **monolithic 3D (M3D)** FPGA with **tiered active devices**



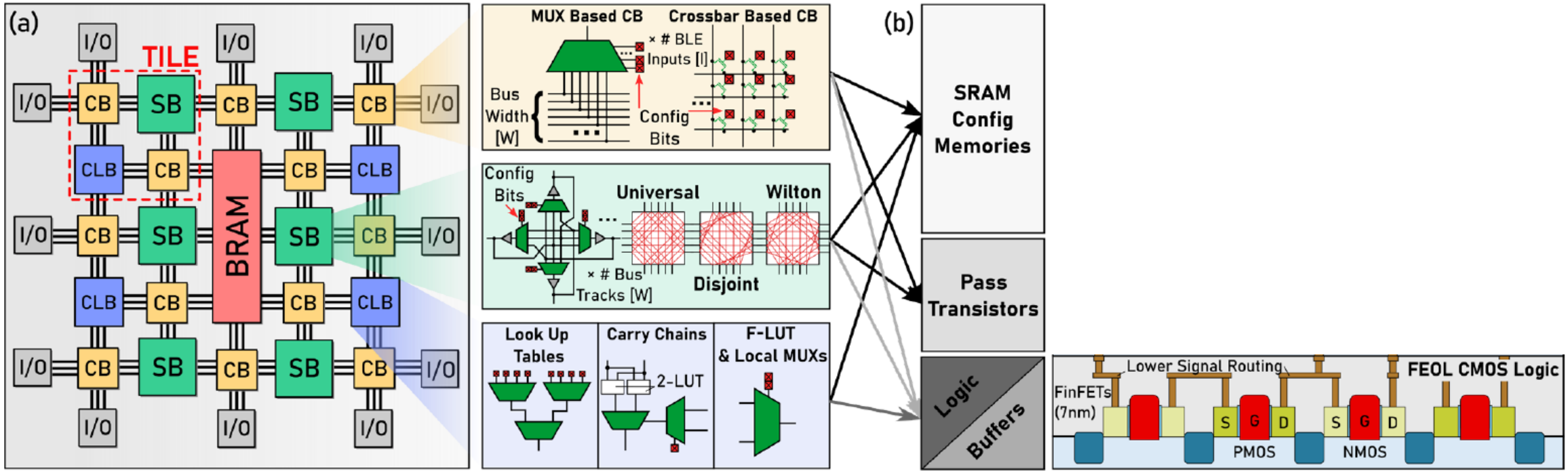


Our Proposed Design: M3D FPGA



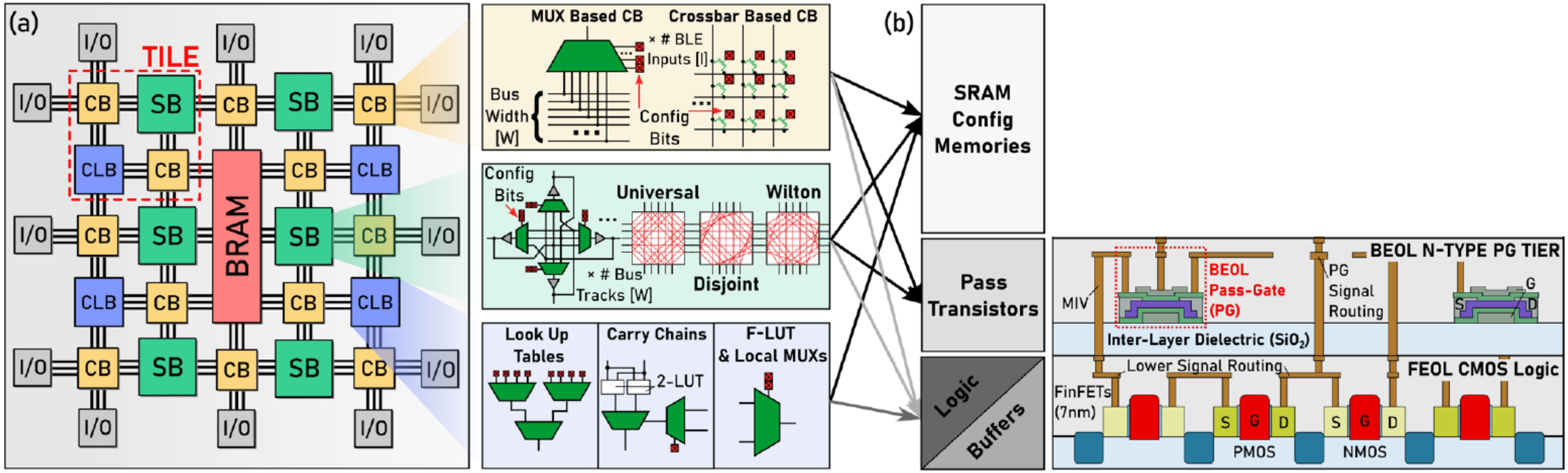
Observe the composition of each block: CBs and SBs provide connectivity to CLBs. Made mostly of configuration bits and pass transistors.

Our Proposed Design: M3D FPGA



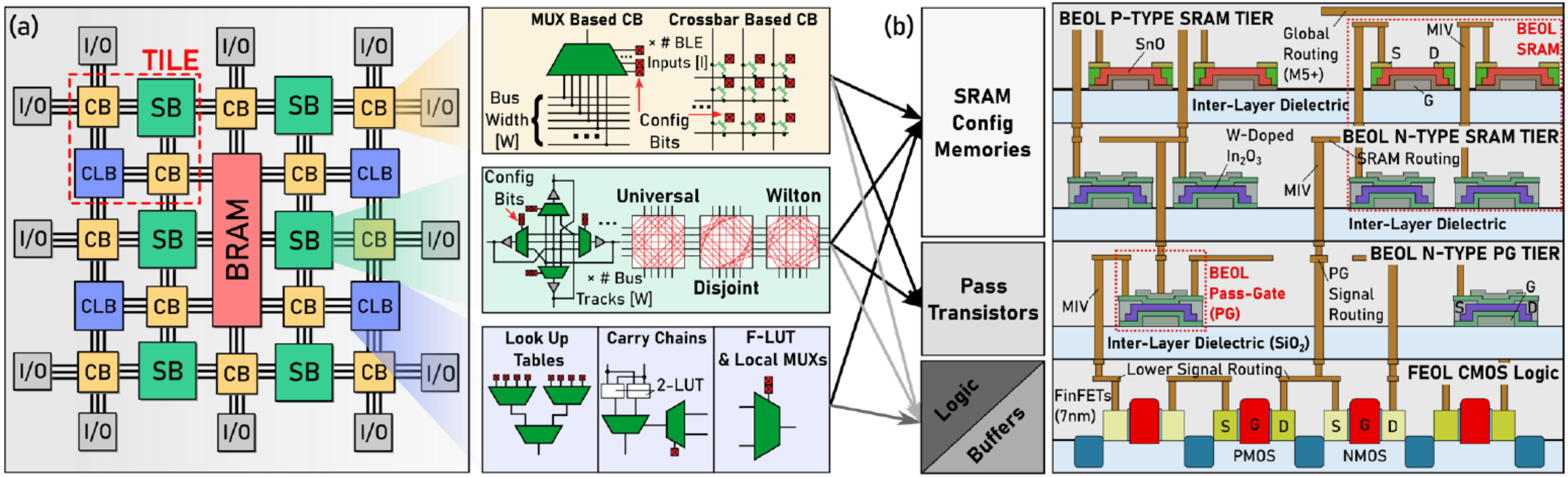
First, **implement critical path logic using silicon CMOS** in **FEOL,** such as CLBs and Buffers

Our Proposed Design: M3D FPGA



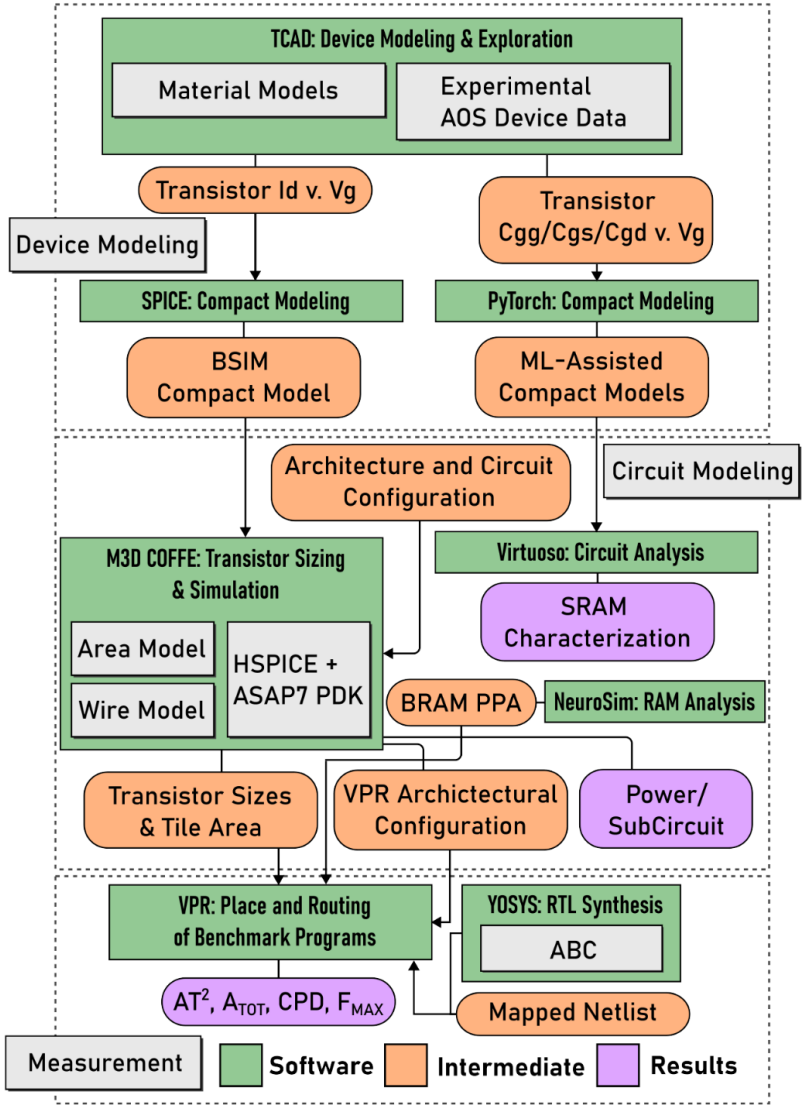
Second, build **BEOL pass transistors (N-Type, W-doped In_2O_3) above logic and buffers**. Connect using MIVs

Our Proposed Design: M3D FPGA



Finally, construct **BEOL SRAM above Pass Transistors (P-type SnO, N-type IWO)**

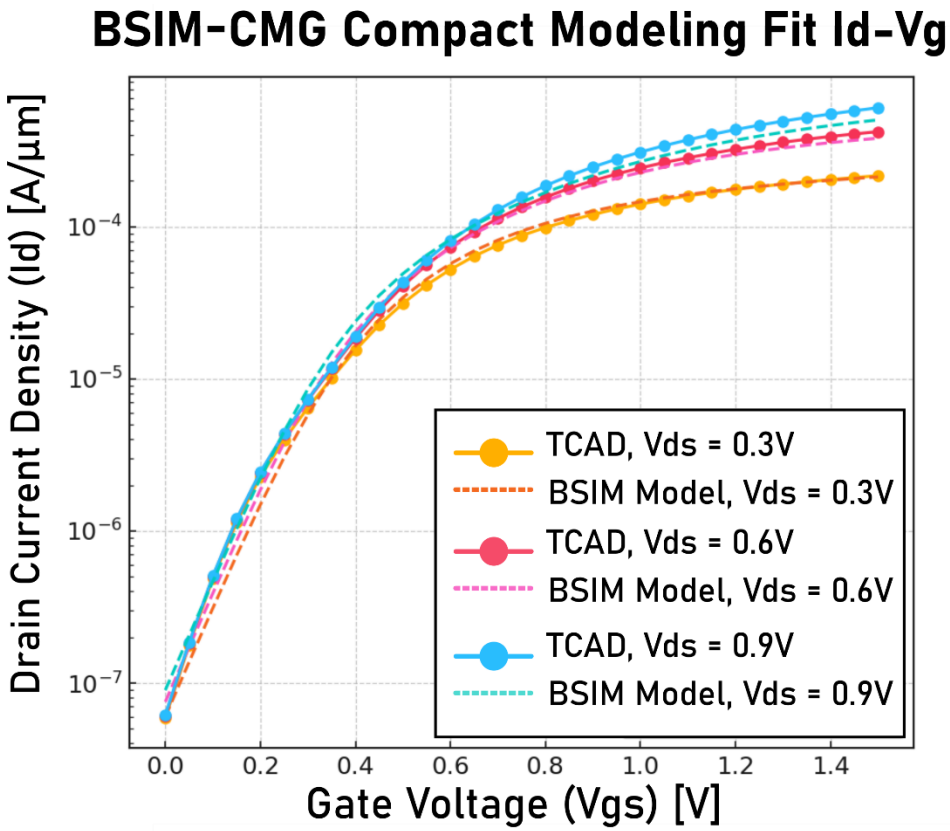
Methods: A TCAD to VTR Framework



Device Modeling: Sentaurus TCAD, transferred to SPICE using BSIM & ML-Assisted Compact Models

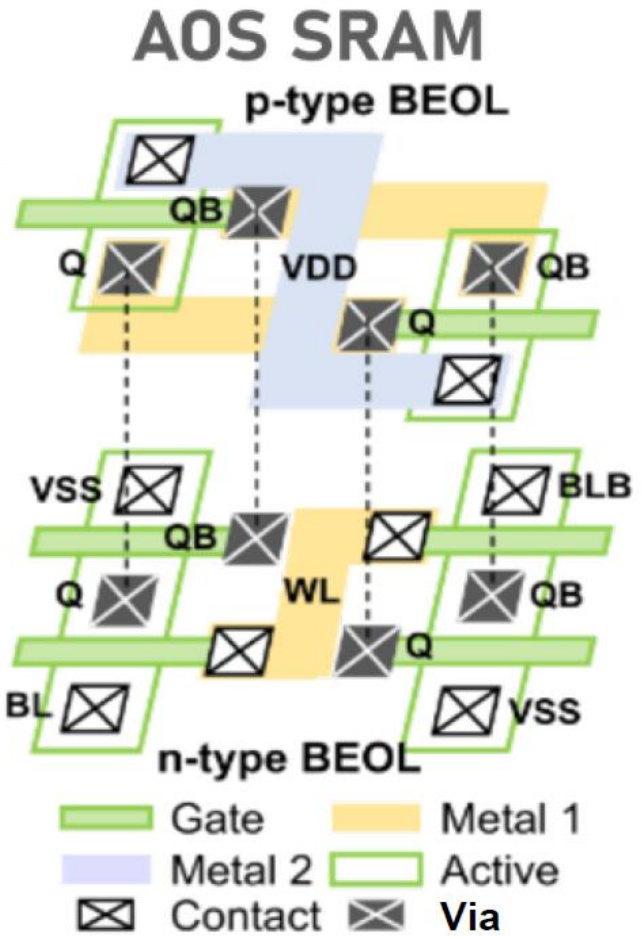
Circuit Modeling: Use modified COFFE to optimize transistors and tile sizing using ASAP7

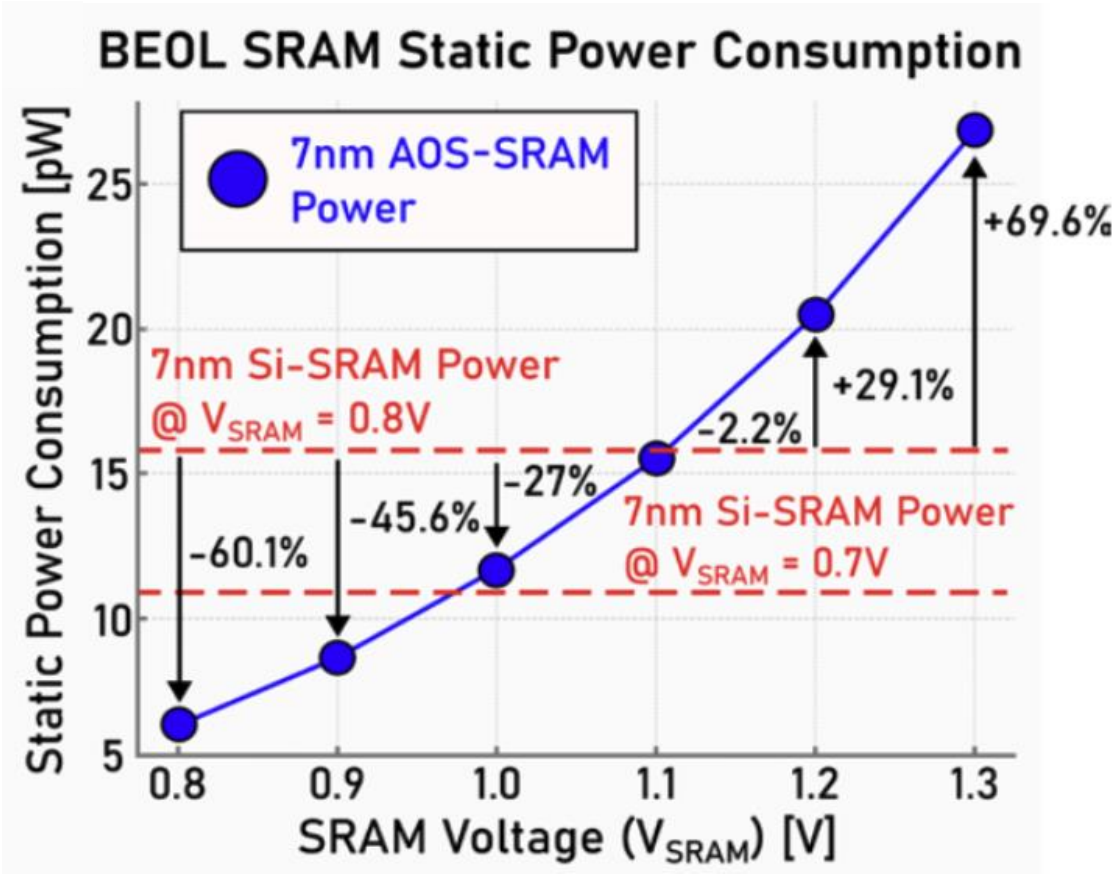
System Modeling: Use VTR tools to synthesize benchmarks and evaluate system performance



Capture C_{gs}/C_{gd} **with 95.4% accuracy**. Capture I_d-V_g/I_d-V_d **with 91.6% accuracy**

Characterization of AOS SRAM

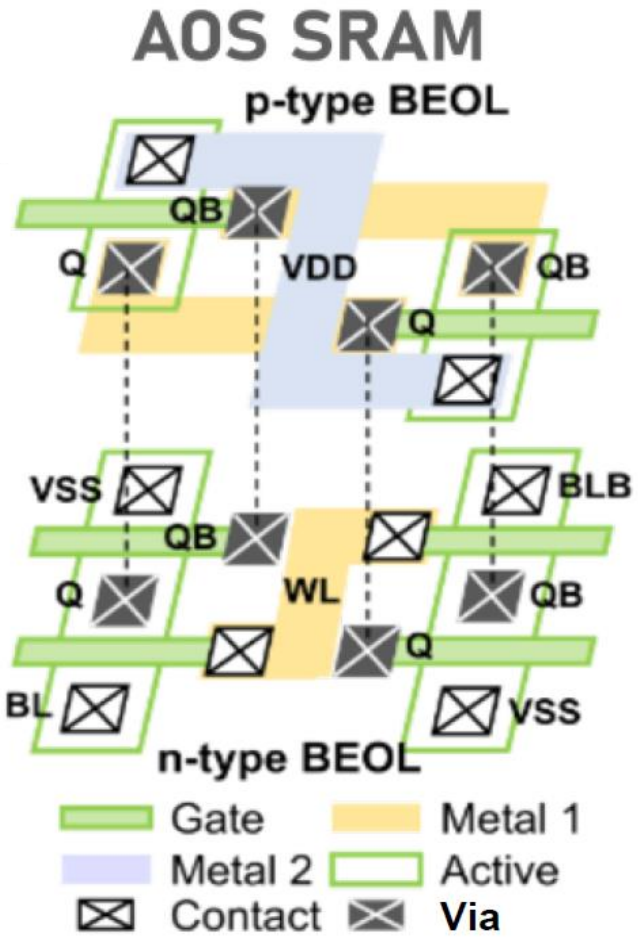


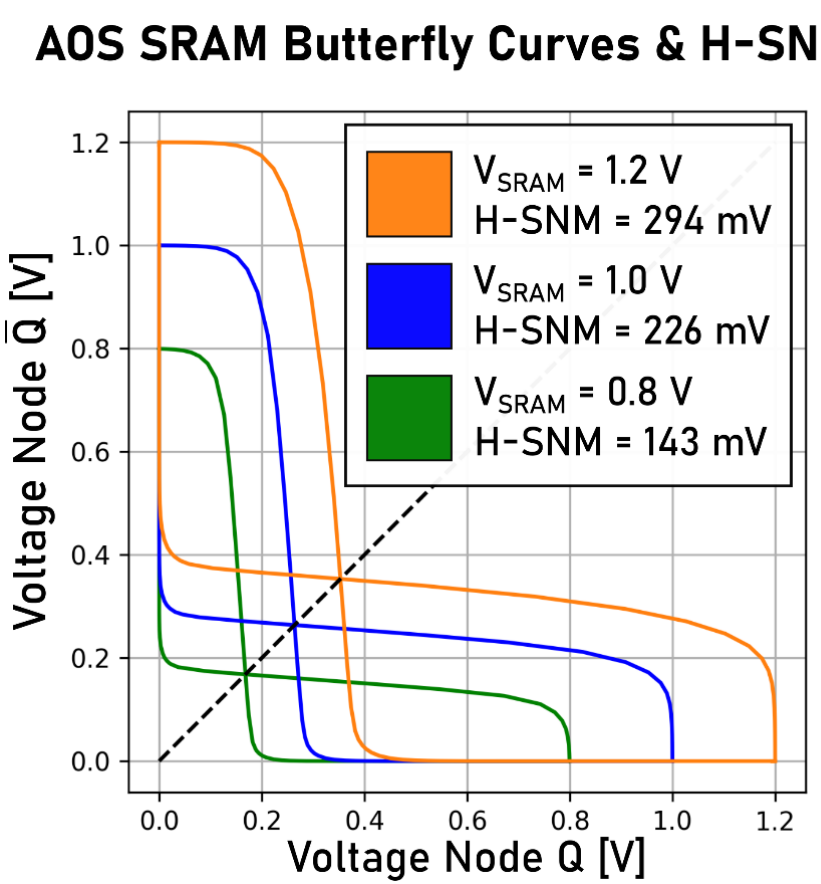


At a typical PG overdrive, static power is reduced by 60.1%

V_{SRAM} can be increased to 1.1 V w/o a static power penalty

Characterization of AOS SRAM

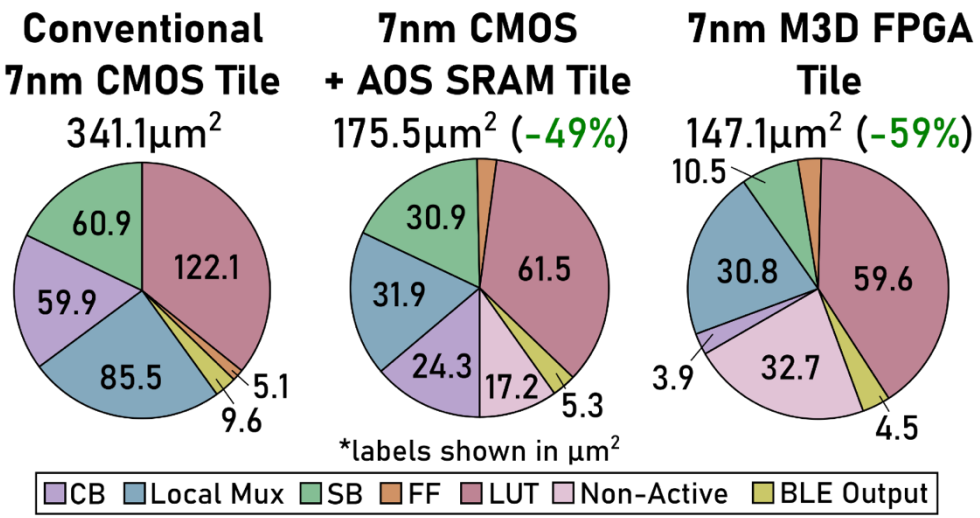




H-SNM is 226 mV at 1.0 V VSRAM, Lower than CMOS (336 mV @ 0.8 V)

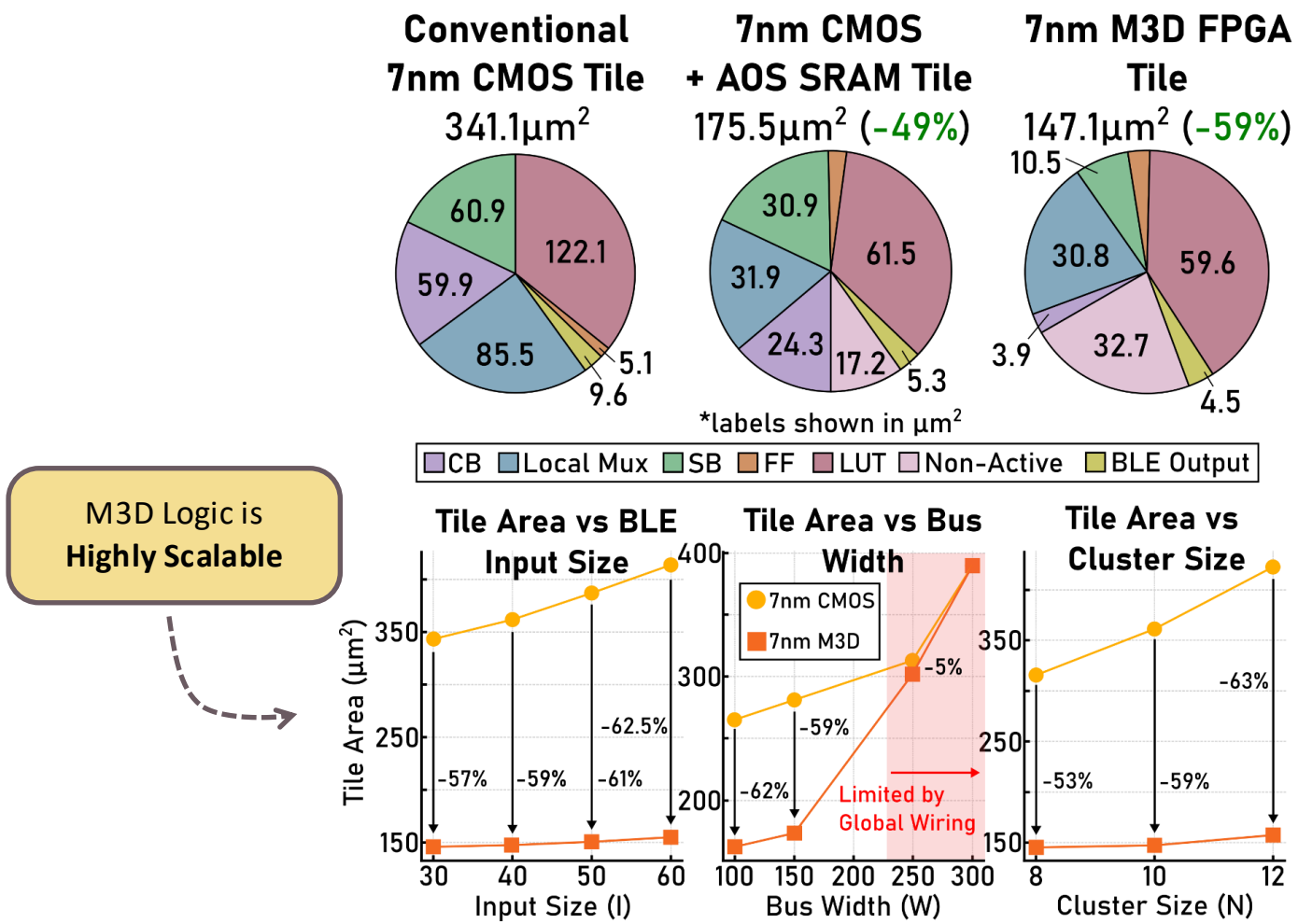
Sufficient (> 200 mV), can be remedied with better-matched NMOS (e.g. IZO)

On Area Reduction

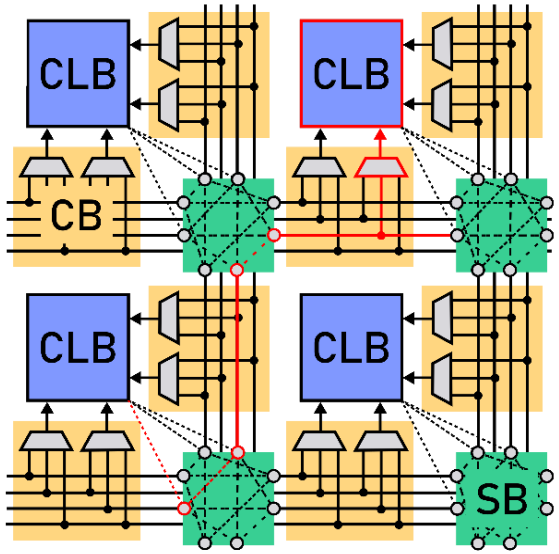


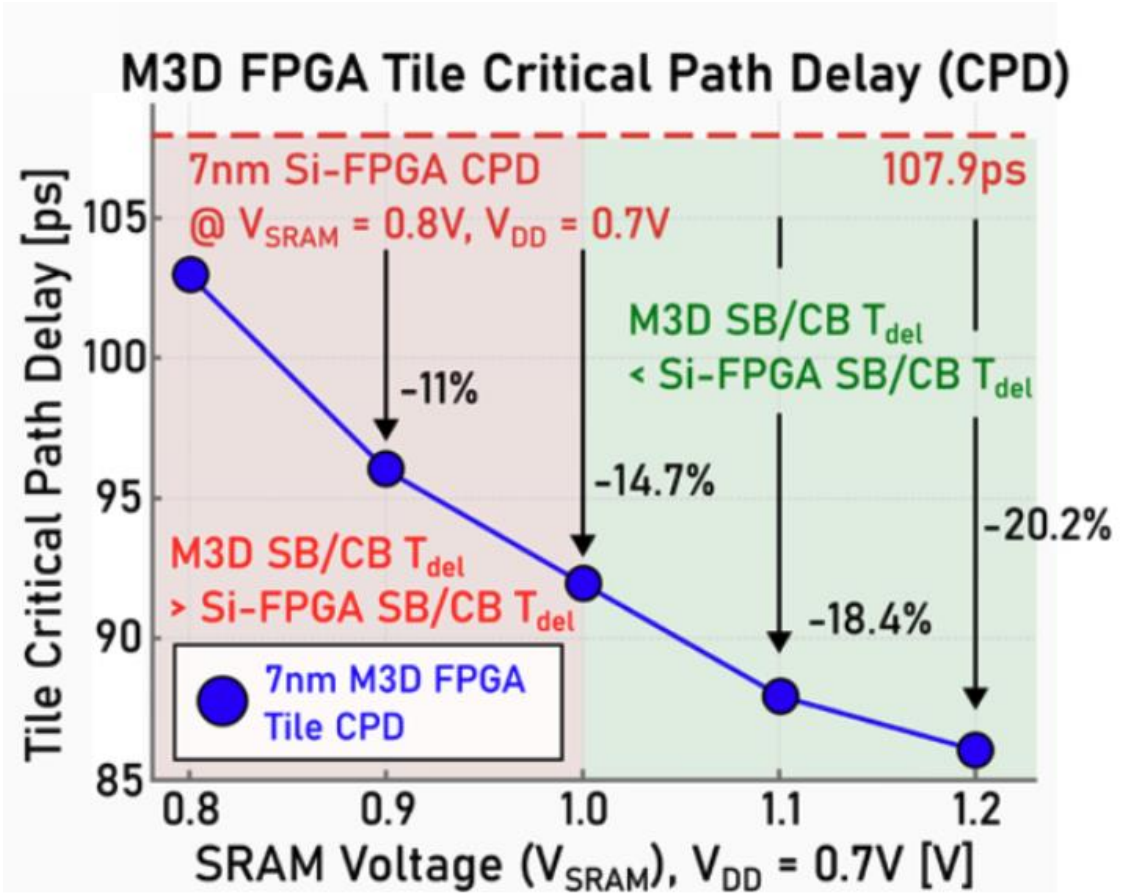
SRAM in BEOL can reduce footprint by ~49%. An additional 10% (Total **~59% reduction) with SB/CB BEOL PGs**

On Area Reduction



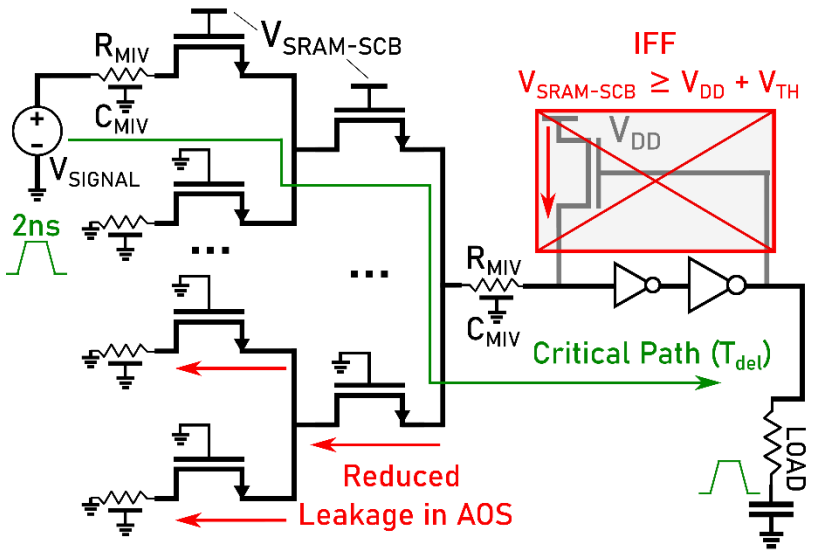
Critical Path Reduction



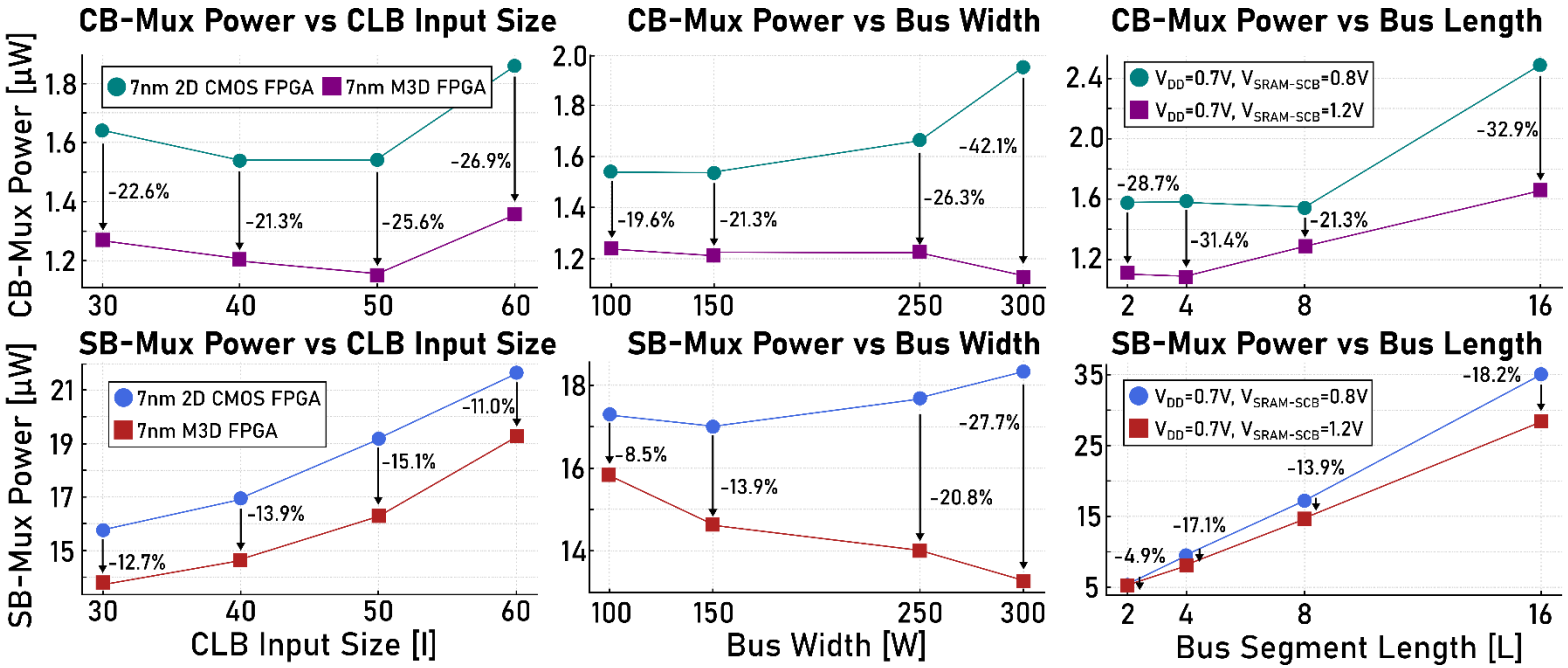


Reduced footprint = reduced wiring load (RC). Reduced tile level CPD (5.8%) at baseline, but PGs in SB/CB are slower!
Overdriving (Wide BG = Less Reliability Concern) improves this further **(>15%)**

On Routing Power Reduction



DUT: Example 2-Level M3D MUX @ 250MHz



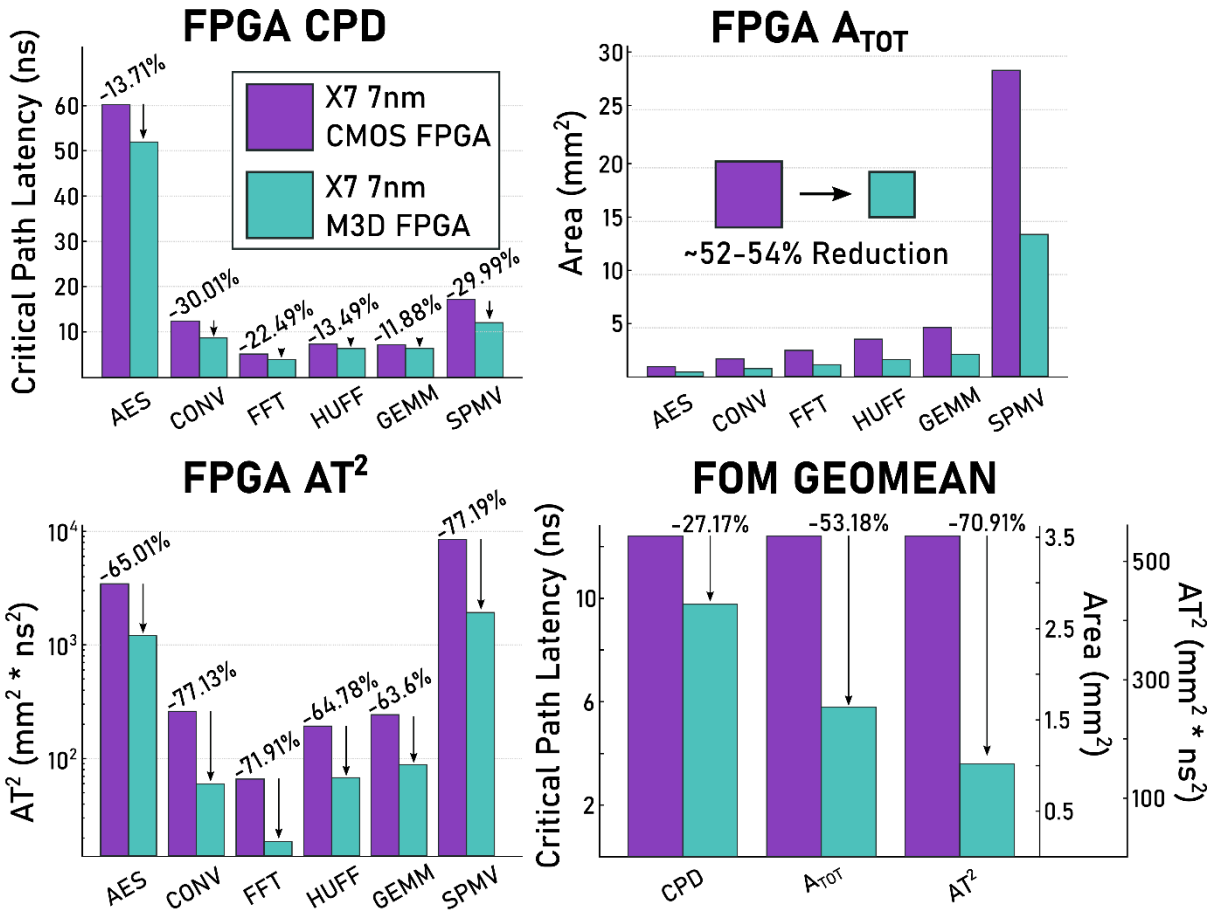
Wide BG in AOS – remove level-restorer from PG mux. Reduce total power in reconfigurable mesh

Reductions in SB-Mux power 17.6%, CB-Mux power 13.8% on average

Benchmarking on Modern Workloads

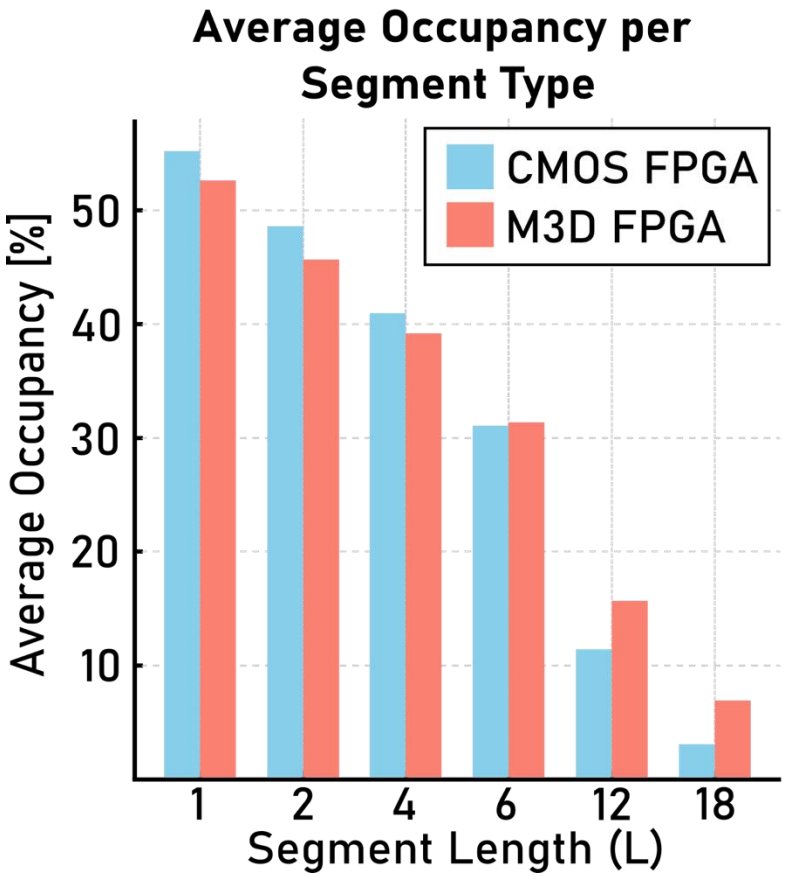
Key Kernel Benchmarking Parameters

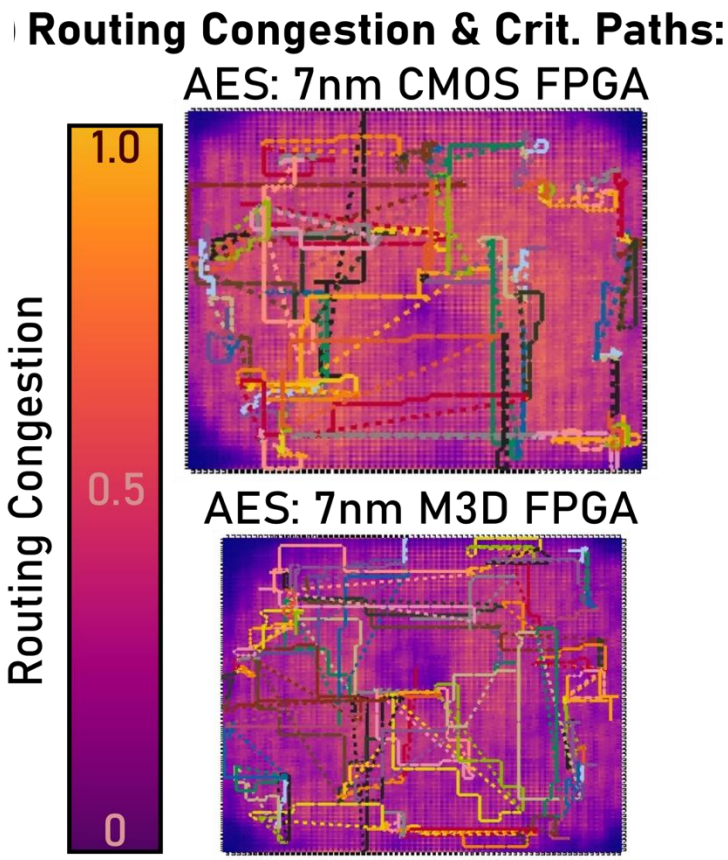
Benchmark	Parameters	Input Type
FFT	Input Data Size n = 64	32b FP
Convolution	Input Size = 128×228×228 Kernel Size = 128×128×5×5 Output Size = 128×224×224 5×5 Convolution Unit Number = 16	8b INT
GEMM	Matrix Size = 512×512 Systolic Array Size = 8×8	8b INT
SPMV	Single HiSparse [39] SPMV Cluster	32b FP
AES Enc. & Dec.	Key Size = 256b	128b INT
Huffman Encoder	Input Symbol Size = 256 Max Tree-Construction Depth = 64 Max Rebalanced Tree Depth = 27	10b UINT 32b UINT



Reductions cumulatively are 27% lower CPD, 53% lower logic area utilization, and **70% AT^2 reduction**

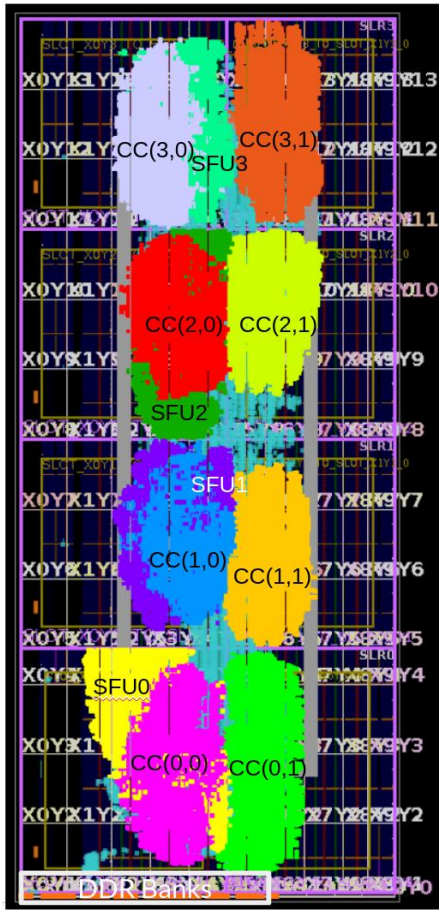
Benchmarking on Modern Workloads





Higher affinity for longer segments in M3D FPGA leading to **lower congestion, less hotspots, more opportunity for hard-block placement in local-proximity**

Benchmarking on GPT-2 Acceleration



Comparison of LLM (GPT-2) Implementation on M3D FPGA

System	Versal VPK 180	7 nm M3D FPGA
Area	736.0 mm^2	423.8 mm^2 (-42.4%)
Delay	4.17 ns	3.31 ns (-20.6%)
AT^2	12798.9 $mm^2 \times ns^2$	4643.2 $mm^2 \times ns^2$ (-63.7%)

Using Xilinx Versal VPK180 FPGA, Vivado 2023.2, and resource utilization from GEMM, translate findings onto an equivalent M3D FPGA. **Significant A, CPD and AT^2 reduction on GPT-2 Class LLM**

Conclusions

- We present **DSE of an M3D FPGA** based on **AOS configuration memories** and reconfigurable mesh

- The proposed FPGA:
  - i.) Is **programmable at logic-compatible voltages** (an issue in other eNVM designs)
  - ii.) Has up to **30% lower critical path delay, 54% lower area util, 77% lower AT^2**
  - iii.) **reduce ASIC/FPGA performance disparity to 2.05-4.37×** and **area to 4.2×**

- AOS SRAMs can **reduce static power by ~60%** while **maintaining adequate H-SNM**

- Demonstrate an **average reduction in SB/CB power by ~15%**

This Work Was Supported By:

Semiconductor Research Corporation

PRISM

NSF

Faaiq Waqar
Ph.D. Student, Georgia Tech

Jiahao Zhang
Ph.D. Student, UCLA

Anni Lu
Senior Engineer, NVIDIA

Zifan He
Ph.D. Student, UCLA

Jason Cong
Volgenau Chair, Distinguished Professor, UCLA

Shimeng Yu
Dean's Professor, Georgia Tech

Thank you for listening

Questions?

Email: faaiq.waqar@gatech.edu, shimeng.yu@ece.gatech.edu