Overview: 3D and Heterogeneous Integration

UT Austin Circuit Research Lab (CRL)

PI: Jaydeep P. Kulkarni

Associate Professor of Electrical and Computer Engineering

Fellow of Silicon Labs Chair in Electrical Engineering

University of Texas at Austin, TX, 78712

Email; jaydeep@austin.utexas.edu

Group: Circuit Research Lab

PI Quick Bio:

2017-Present, ECE faculty, UT Austin, TX

2009-2017 Intel Circuit Research Lab, OR

2009 Ph.D., Purdue Uinversity

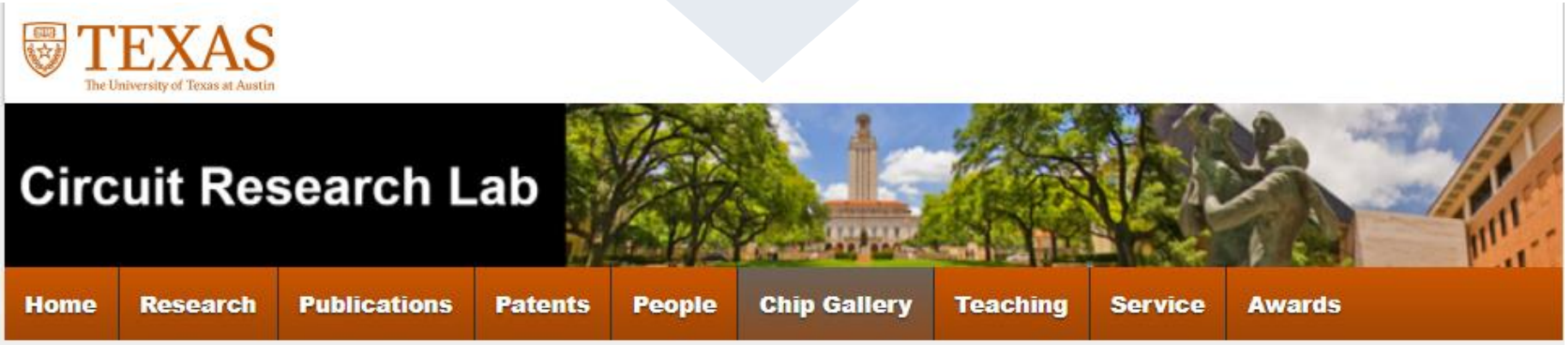
2004 M. Tech, IISc Bengaluru, India

2002 B.E., Pune University, India

Research Support:

Research Thrusts @ CRL

DTCO for Emerging Devices	ML / AI Accelerators, Memory Designs	Hardware Security
1. 2D RRAMs, Selectors	1. Accelerators (DNN, TinyML, Graph)	1. Side Channel Attack Resilience (Power, EM, Photonic)
2. Ferroelectric FETs	2. Compute-in-Memory (SRAM, ROM, CAM, FPGA, eDRAM, RRAM, MRAM)	2. Secure Supply chain with heterogeneous integrated designs
3. JJFET (quantum comp.)	3. COP accelerators (Ising, SAT, ILP)	3. Fully Homomorphic Encryption (FHE) accelerators
4. Backend TFTs (IGZO)	4. Neuromorphic Computing (SNN)	
5. **3DHI (BPR,HWB,Chiplet)**	5. ML Circuit/Arch. Aspects	
	6. Cryogenic CMOS	
	7. Radiation hardened designs	



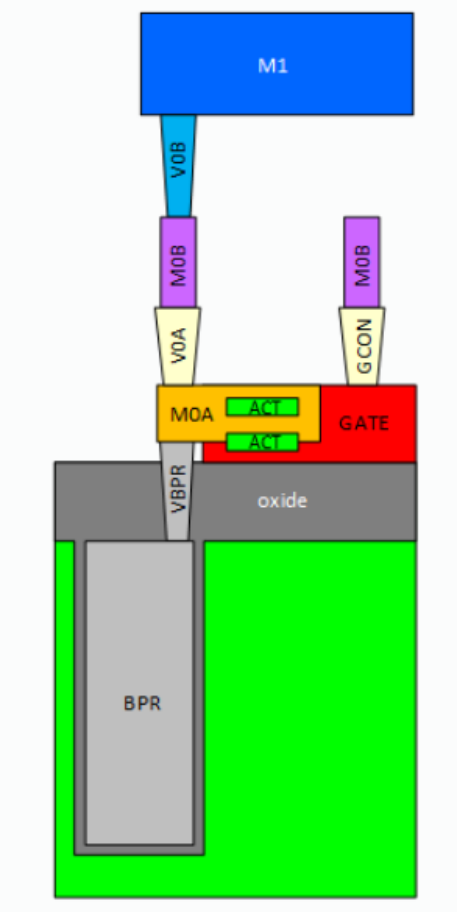
3D interconnects:

Buried Power Rail (BPR) DTCO

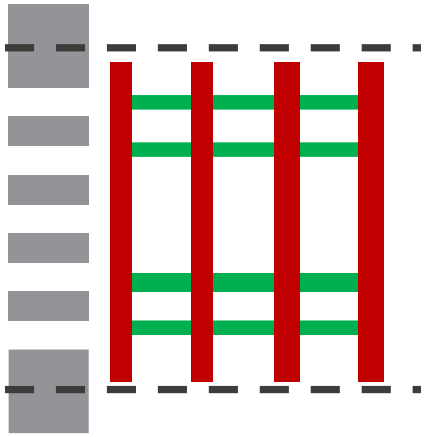
Reference: S. S. Teja Nibhanupudi, Divya Prasad, Shidhartha Das, Odysseas Zografos, Bilal Chehab, Satadru Sarkar, Alex Robinson, Anshul Gupta, Alessio Spessot, Peter Debacker, Diederik Verkest, Julien Ryckaert, Geert Hellings, James Myers, Brian Cline, and Jaydeep P. Kulkarni, “A Holistic Evaluation of Buried Power Rails and Back-side Power Grids for sub-5nm CMOS technology nodes” IEEE Transactions on Electron Devices (TED)

What are Buried Power Rails

- Burying the VDD and VSS lines beneath the substrate
- Trench etched in Silicon thru STI and filled with metal
- Unlike BEOL metals, buried rails can have high aspect ratio

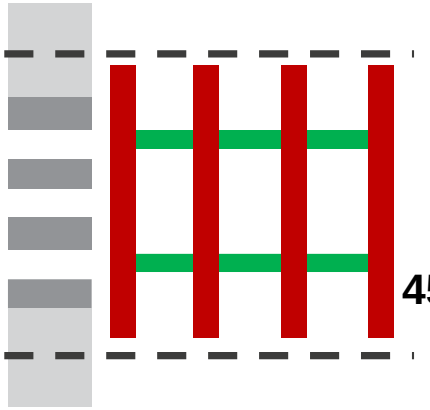


Why Buried Power Rails?

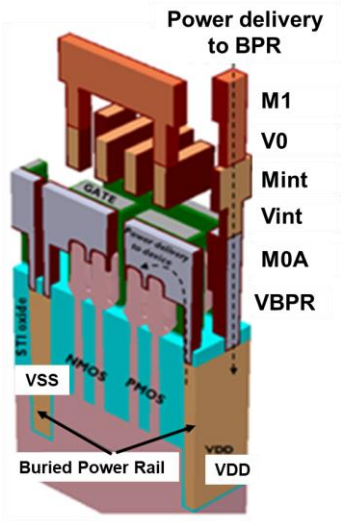


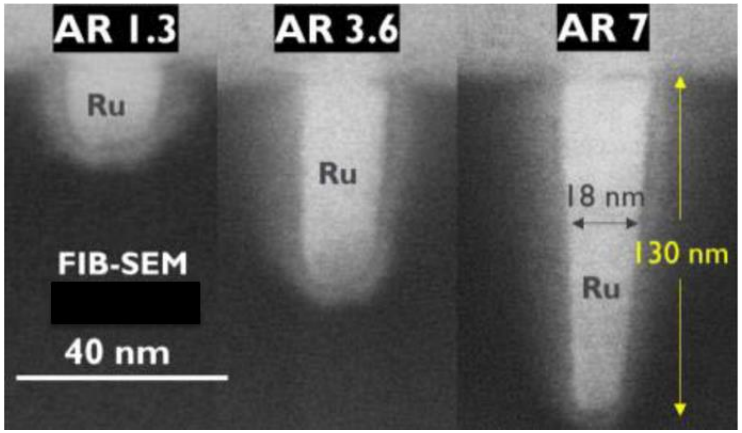
-17%

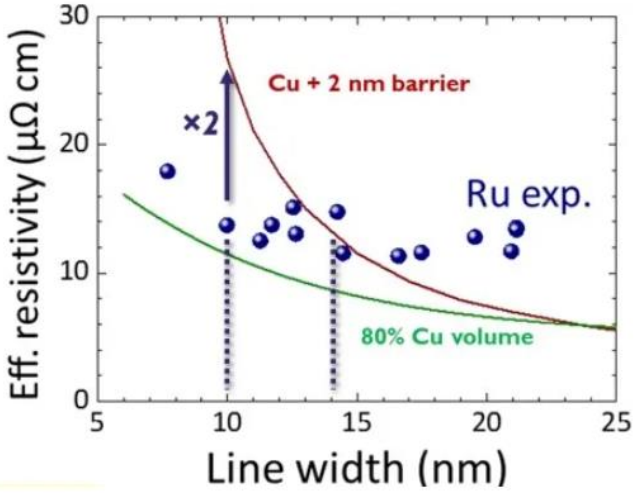
6T 2fins
45CPP22MP



5T 1fin
45CPP22MP







Ru BPR schematic

Ru BPR cross-section SEM

Image courtesy: Imec

- Standard cell track height can be scaled with buried power rails
- Ruthenium has lower resistivity and rails can withstand FEOL temperature

BPR DTCO Aim

- Study system level impact of Buried rail technology on CPU cores
  1. Power, Performance, Area
  2. On-chip IR drop
  3. Transient voltage droop

IMEC

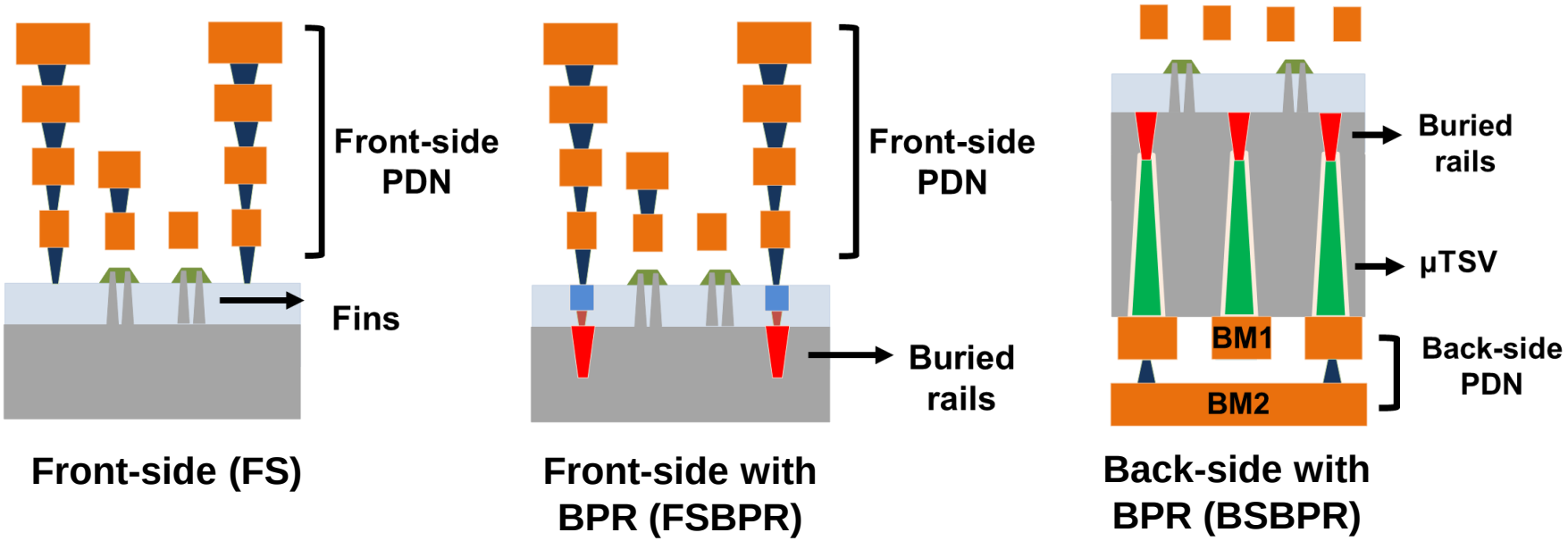
1. **Test vehicle development**
2. **Characterizing Buried Rails**

Arm

1. **Standard cell library development**
2. **Physical Design flow of CPU cores**

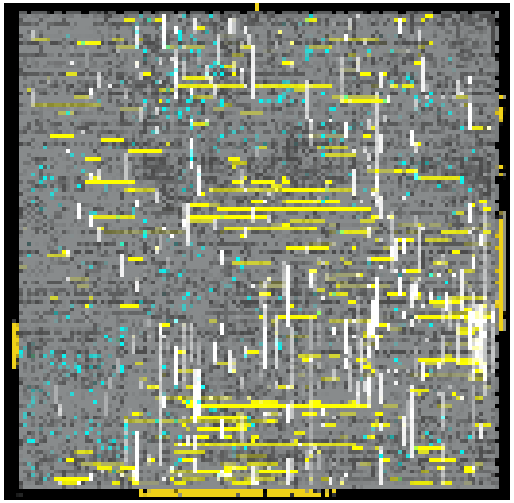
- Details of the setup
  - CPU core – Arm Cortex A-53 (LITTLE core)
  - Technology node – Imec's iN6 (equivalent to foundry 3nm)
  - 6T height standard cells (no SC height scaling assumed)

BPR– potential power delivery configurations

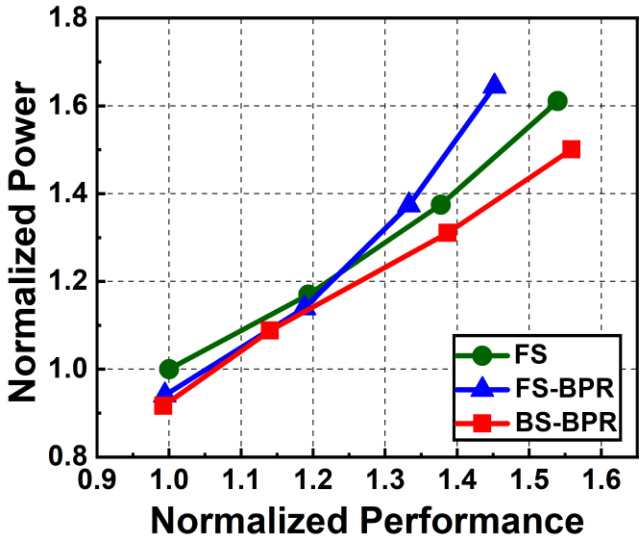


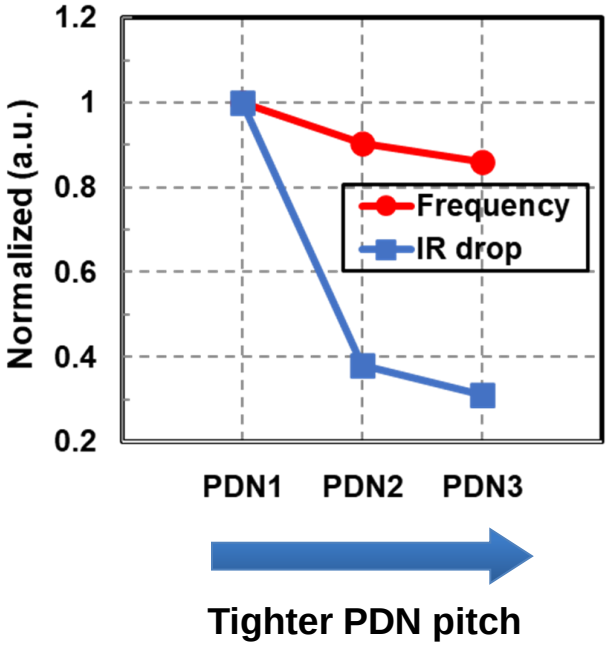
- FS – Power and signals on front-side of the chip
- FSBPR – Power and signals on front-side of the chip
- BSBPR – Power on back-side and signals on front-side of the chip

PPA comparison of PDN configurations



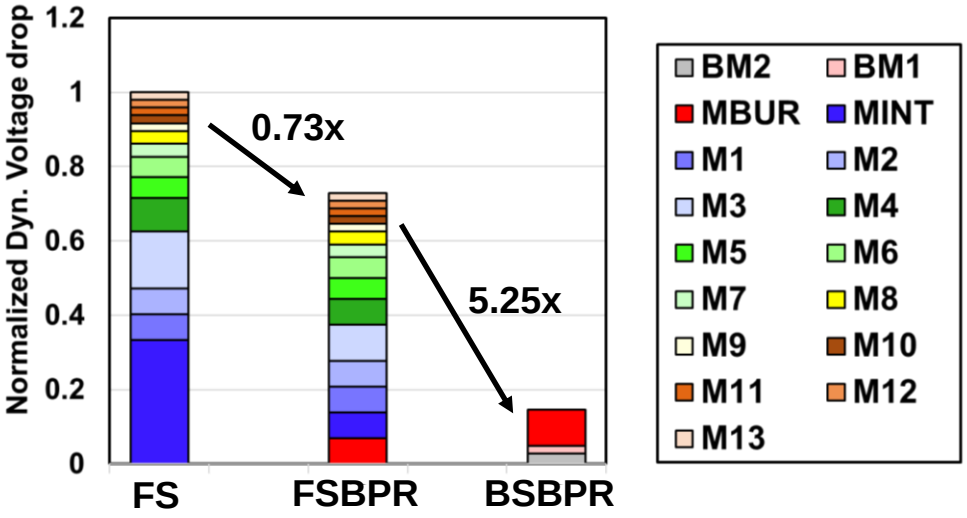
LITTLE core Cortex A-53

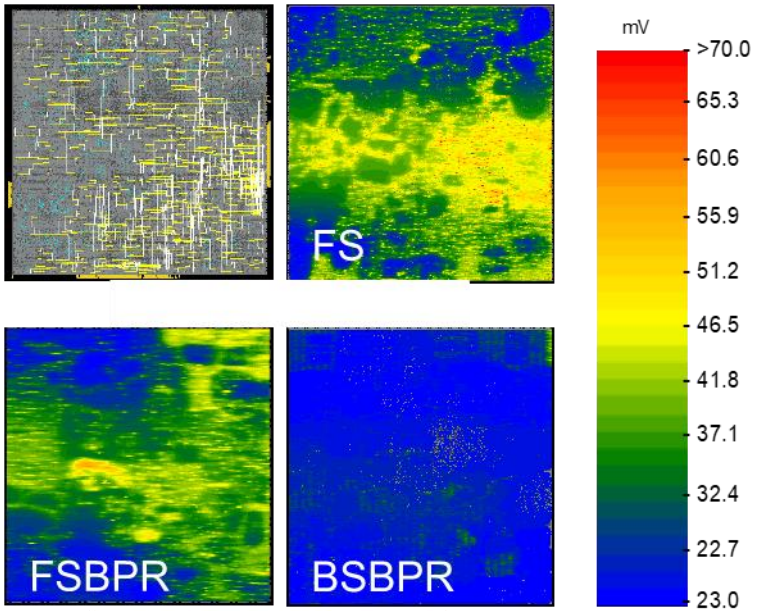




- FSBPR suffers due to routing congestion: impact of tap-cells
- BSBPR achieves higher performance at lower power
- For iso-performance of 1.4 (a.u.), FSBPR has 10% higher power and BSBPR has 8% lower power

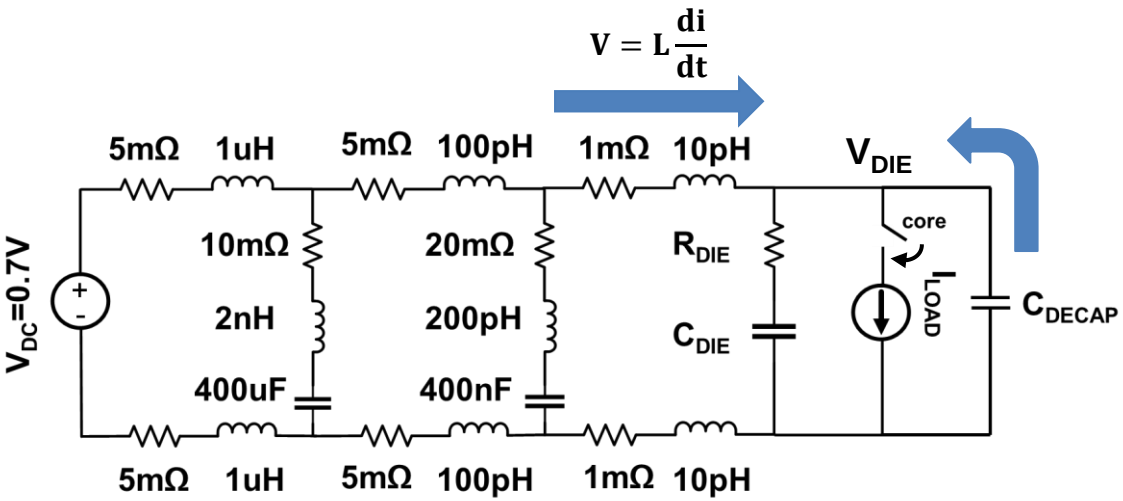
IR drop comparison of PDN configurations

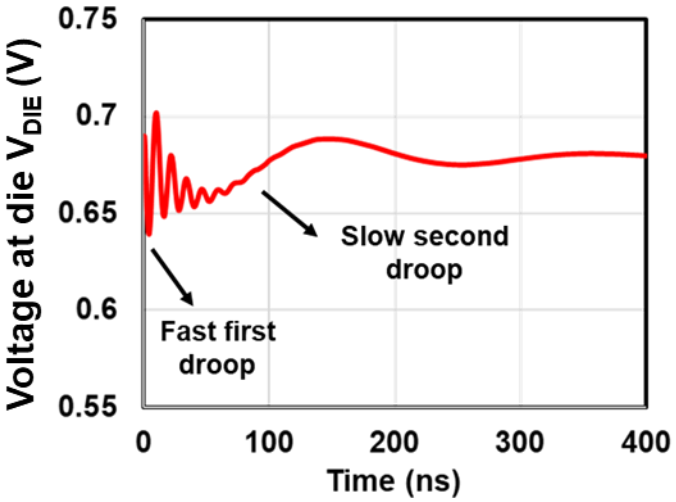




- BPR drops IR drop by 70% on lowest metal layer
- µTSV pitch crucial in determining the benefits of BSBPR

Off-chip voltage droop





- Package inductor introduces transient (Ldi/dt) voltage drop during current rush events
- Decoupling capacitance reduces the voltage droop during transient events

BSBPR – decoupling capacitance

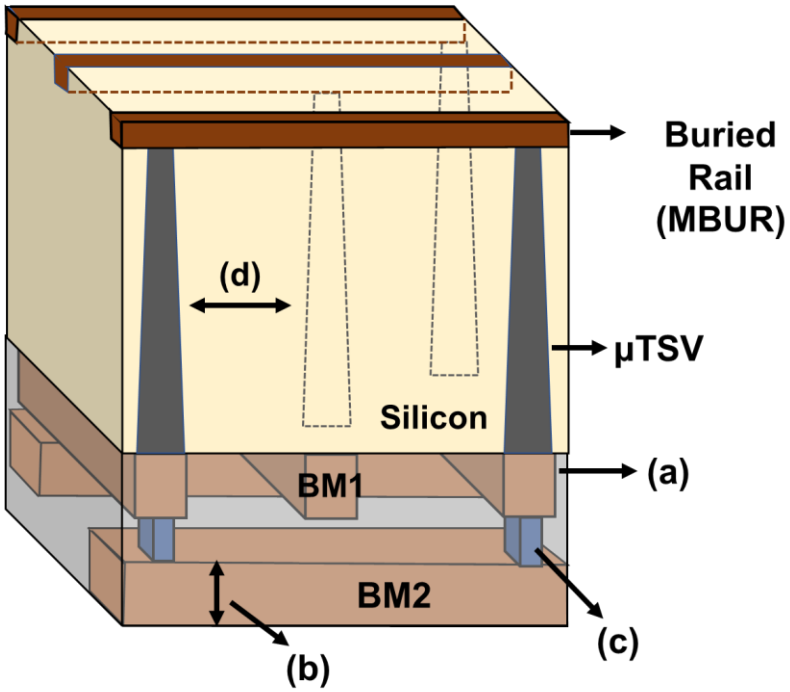
Power grid decoupling capacitance can be increased by tuning the following,

(a) Backside dielectric relative permittivity

(b) Backside metal thickness

(c) BM1-BM2 via height

(d) µTSV pitch



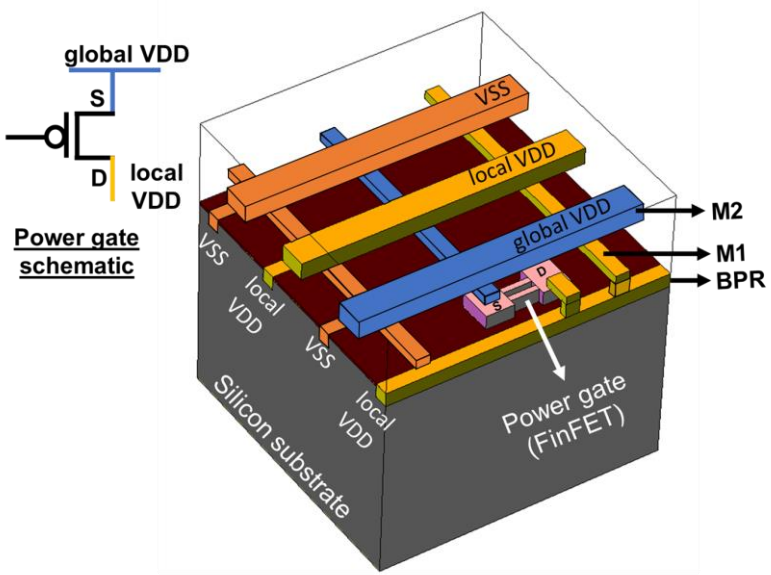
- Decoupling capacitance of power grid can be increased without worrying about signal-to-signal noise coupling
- Such modifications not possible in front-side grids

Combined IR drop, off-chip voltage droop

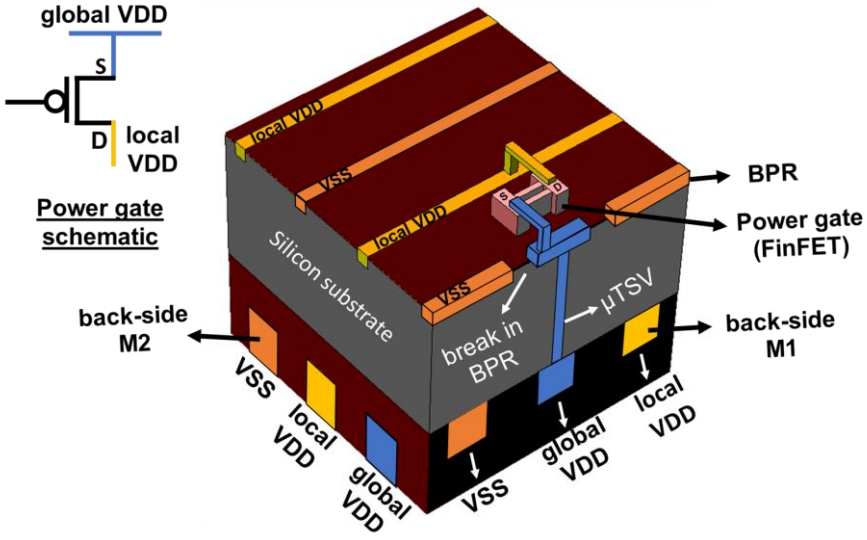
	Frequency	Power	IR drop	Off-chip voltage droop
FS	1x	1x	70mV	68mV
FSBPR	1x	1.1x	52mV	58mV
BSBPR	1x	0.92x	10mV	48mV

- Accounting for IR drop into voltage guard-band, FSBPR consumes 5% lower and BSBPR consumed 15% lower power compared to FS

Power Gating in Buried rail technology



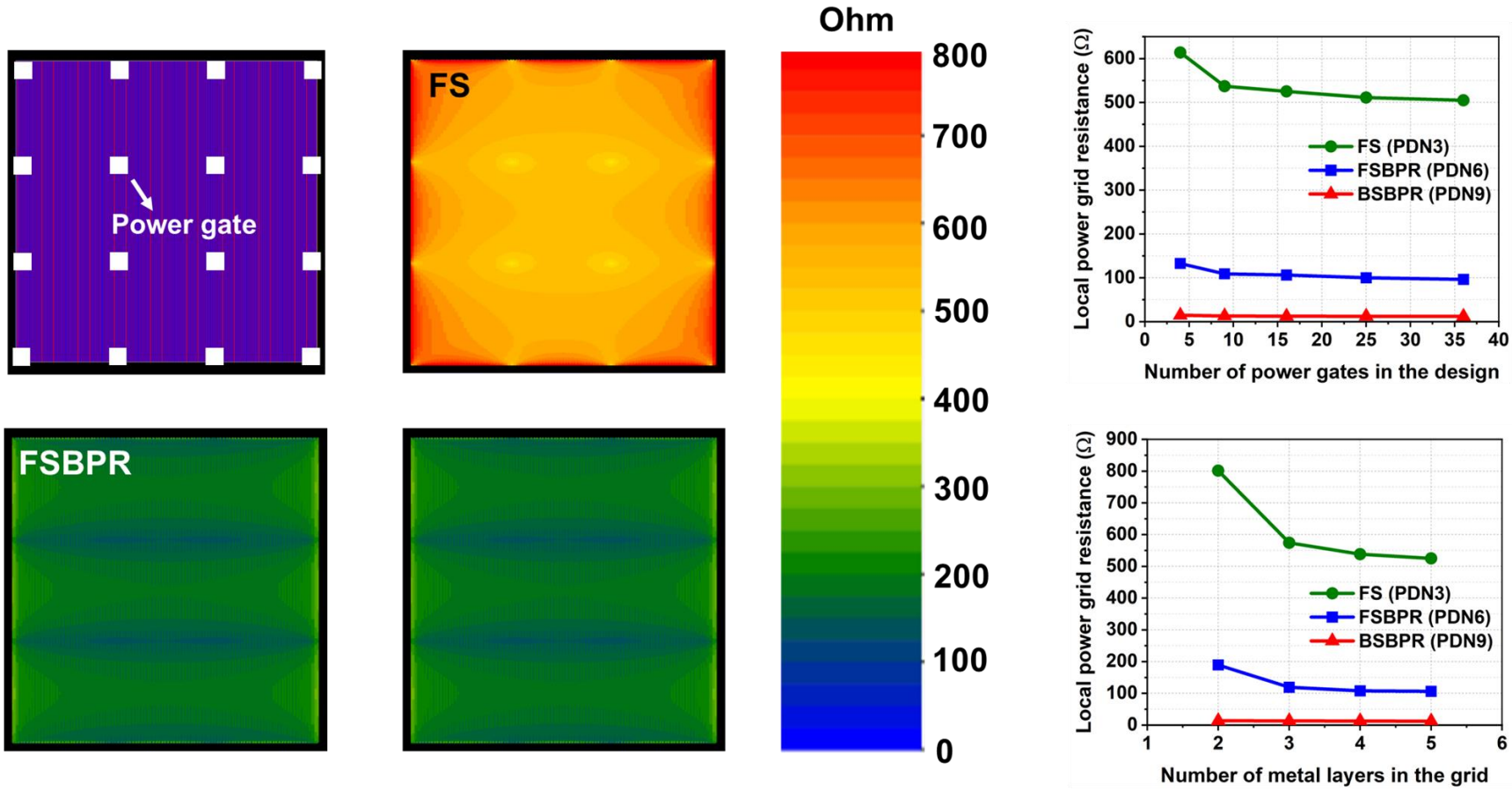
FSBPR Power Gating implementation



BSBPR Power Gating implementation

- Break in buried rail required to host the global VDD island in BSBPR technology

Power Gating in Buried rail technology

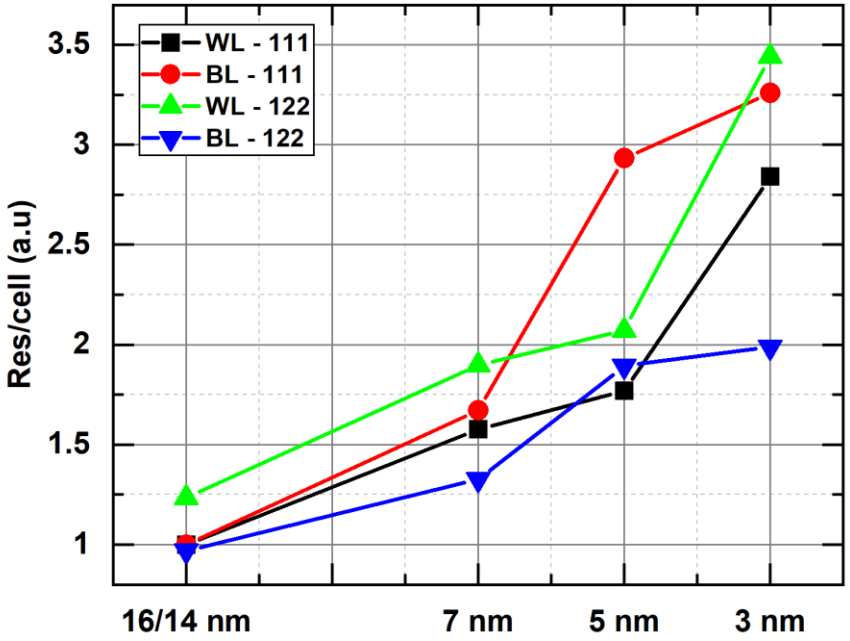


- Buried rails lowers power grid resistance of local power grid

3D Interconnects:
Buried Interconnect SRAM

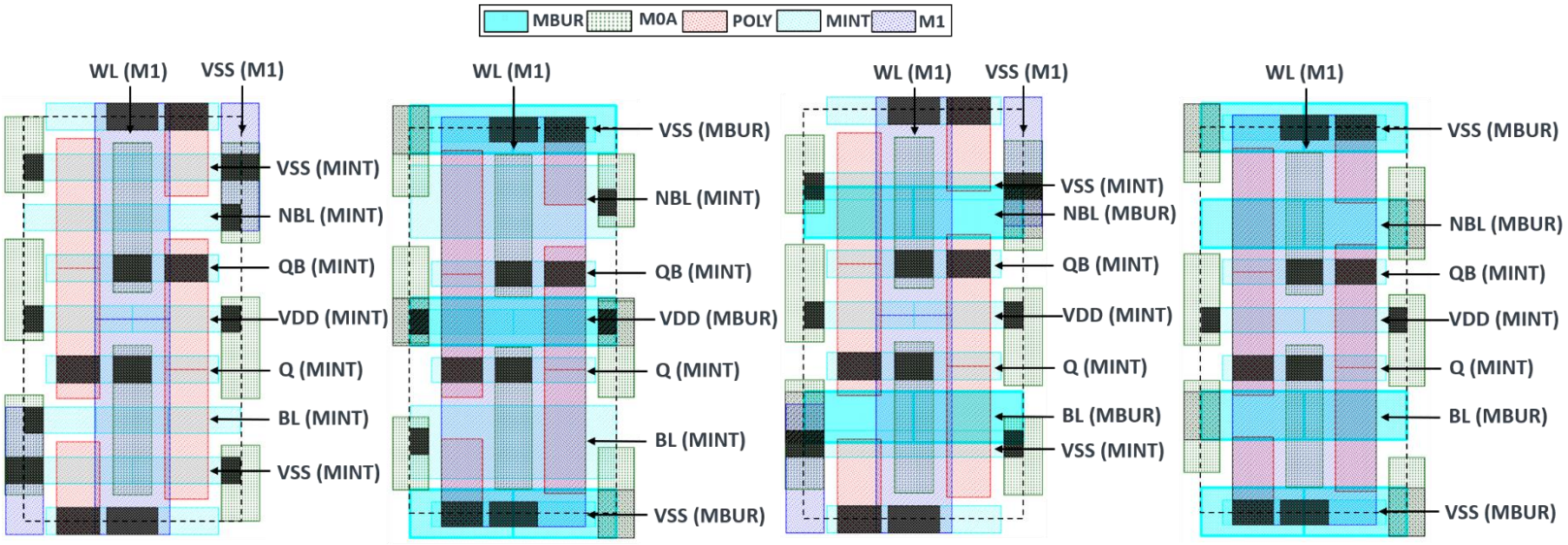
Reference: Rahul Mathur, Mudit Bhargava, Brian Cline, Shairfe Salahuddin, Anshul Gupta, Pieter Schuddinck, Julien Ryckaert, and Jaydeep P. Kulkarni, “Buried Interconnects for sub-5nm SRAM Design”, IEEE Transactions on Electron Devices (**TED**), January 2022, [**Paper**]

RC Scaling Trends in SRAM



- Traditionally, metal resistance increase due to pitch scaling is offset by length scaling
  - Narrow critical dimensions (CD) of metals in deeply scaled nodes
- SRAM contain long wires routed on lower metal layers:
  - Wordlines (WL)
  - Bitlines (BL)

SRAM Cell Design with Buried Interconnect



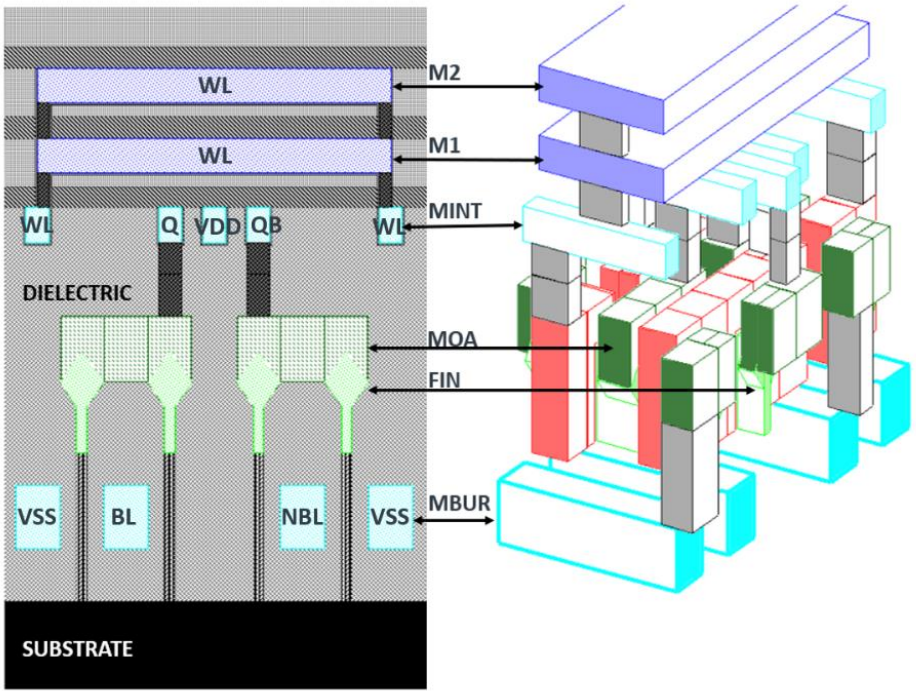
S. Salahuddin *et al.*, "Buried Power SRAM DTCO and System-Level Benchmarking in N3," 2020 *VLSI*

R. Mathur *et al.*, "Buried Bitline for sub-5nm SRAM Design," *2020 IEDM*

- Extending Buried interconnects from power routing to the signal routing
- SRAM is unique as it has long low-level high resistivity metal routing for WL and BLs

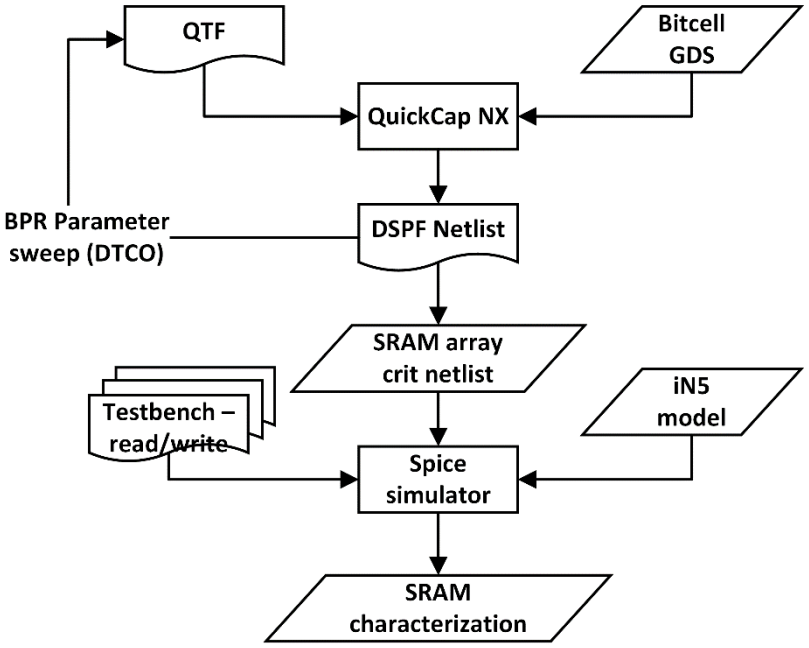
2D and 3D Cross-section of BBL-BVSS Cell

- VSS and BLs in the direction of fin orientation
  - Minimal process change from BPR approach
- Creating space in the M1 layer by buried VSS
  - Allows for wider WL wire in M1
- Skipped buried VDD



Extraction and Simulation Framework

- Quickcap: Good accuracy capturing the 3D fields without TCAD and process emulation
- QuickCap Technology File (QTF)
  - Accurate description of process geometries
  - Allows modifications to process parameters like buried metal thickness
- Industry-standard methodology for evaluation of SRAM macro-level metrics



Conclusion: Buried Interconnect SRAM

- Explored the use of buried metal for signal and power routing together
- BBL and BBL-BVSS built over BPR technology
- High accuracy parasitic extraction of SRAM cell with buried interconnect
- Buried Interconnects offers significant gain in SRAM power and performance

Metric	Cell	BPR	BBL	BBL-BVSS
Access time	1-1-1	6%	10%	11%
	1-2-2	8%	5%	10%
Write time	1-1-1	-	23%	28%
	1-2-2	-1%	13%	19%
Dynamic Power	1-1-1	-5%	1%	1%
	1-2-2	4%	4%	4%

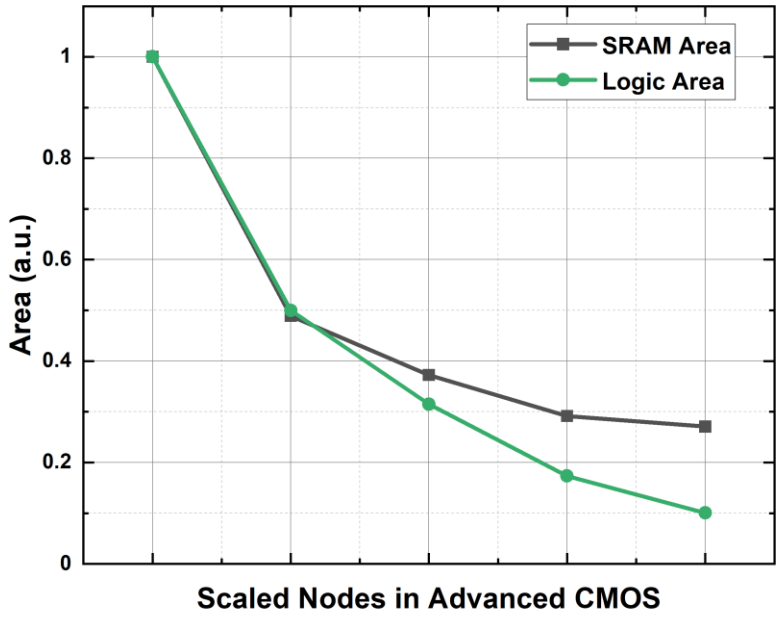
Hybrid Wafer Bonding: 3D Split-SRAM

Reference: Rahul Mathur, Mudit Bhargava, Heath Perry, Alberto Cestero, Frank Frederick, Shawn Hung, Chien-Ju Chao, Daniel Smith, Daniel Fisher, Norman Robson, Xiaoqing Xu, Pranavi Chandupatla, Raguram Balachandran, Saurabh Sinha, Brian Cline, and Jaydeep P. Kulkarni, “3D-Split SRAM: Enabling Generational Gains in Advanced CMOS” in IEEE Custom Integrated Circuits Conference (**CICC**), April 2021 [**Paper**] [**Slides**]

2D SRAM Challenges

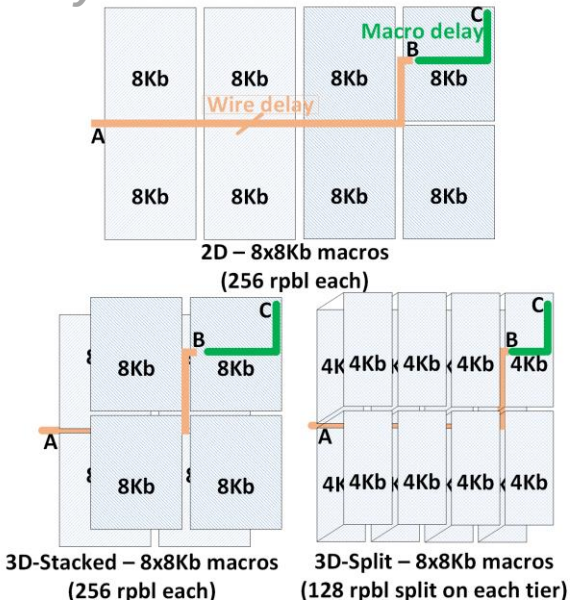
- SRAM scaling challenged by:
  - Gradual shrinking of critical pitches
  - High contact resistance
  - Specialized design with constrained rules
- To extend SRAM scaling gains:
  - Stacking standalone SRAMs
  - 3D-Split-SRAM
  - Heterogenous SRAM bitcell stacking

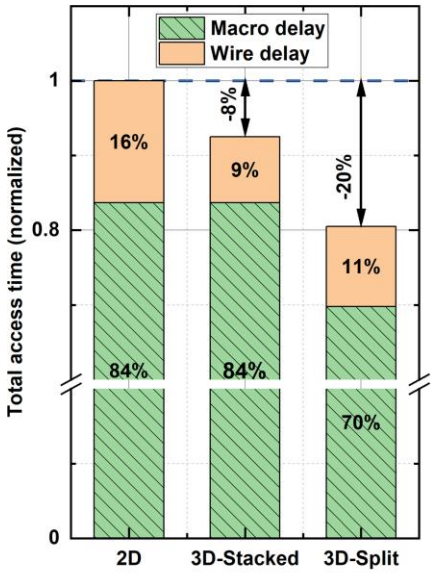
Logic still scales at ~40-45% per node, SRAM scaling lags at ~20-25%.



3D Stacked Vs 3D-Split

Simulation Study





2D & 3D configurations of 64Kb L1 cluster. Simulation in 12nm @SS/(V_{NOM}-10%)/-40°C. Wire delay ~200ps/mm.

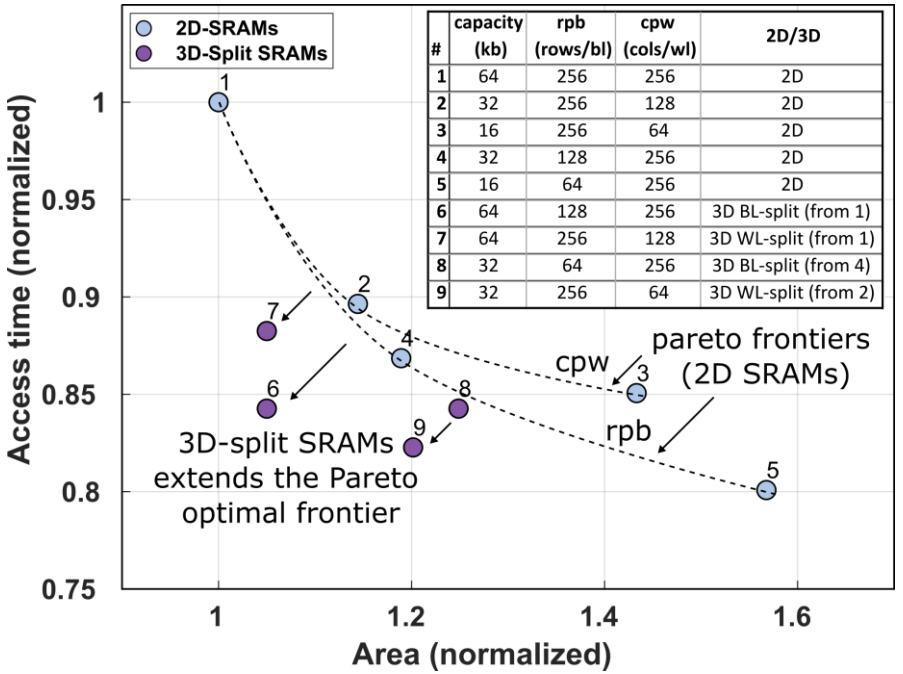
- 3D-Stacked SRAM: Memory macro on top of each other.
  - ➢ Access-time gain ~8%
- 3D-Split SRAM: splitting the WL/BL of a SRAM block across 3D tiers.
  - ➢ Access-time gain ~20%
  - ➢ Reduction in BL/WL RC

2D Vs 3D-Split

Simulation Study

- 3D-Split SRAM Vs 2D
  - ➢ Fast access-time
  - ➢ Low area
  - ➢ Lower leakage power
- Feasibility & efficacy depend on 3D back end of line (3D-BEOL)
  - ➢ Pitch restrictions
  - ➢ RC parasitics

Access time vs Area for 2D and 3D-split macros



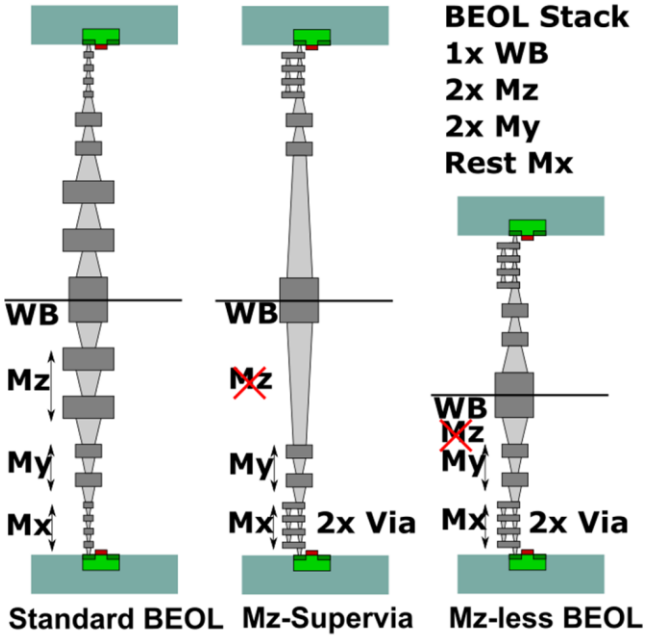
#	capacity (kb)	rpb (rows/bl)	cpw (cols/wl)	2D/3D
1	64	256	256	2D
2	32	256	128	2D
3	16	256	64	2D
4	32	128	256	2D
5	16	64	256	2D
6	64	128	256	3D BL-split (from 1)
7	64	256	128	3D WL-split (from 1)
8	32	64	256	3D BL-split (from 4)
9	32	256	64	3D WL-split (from 2)

3D BEOL

Extraction study

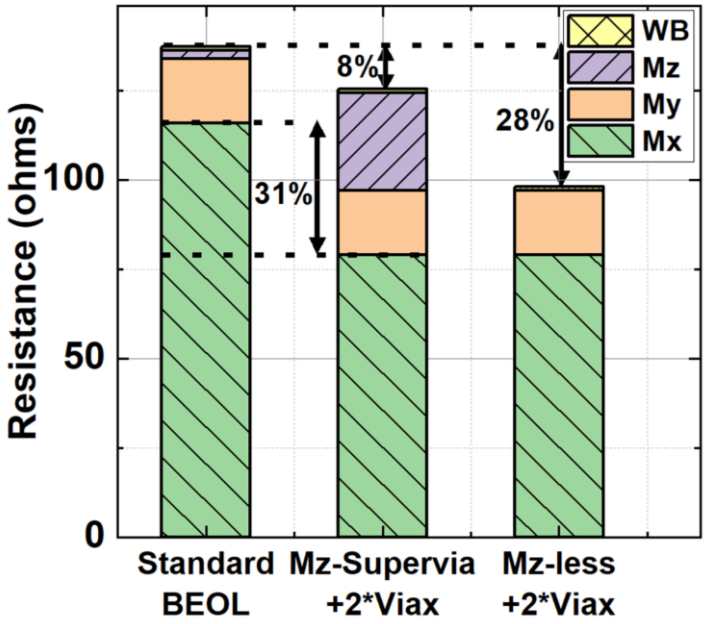
- Goal: Analyze metal stack in 12nm FinFET
  - Assess the RC overhead of 3D- BEOL
  - Identify opportunities for RC improvement.
- Two approaches to optimize BEOL RC for 3D-Split SRAMs:
  - M_Z-Supervia
  - M_Z-less BEOL

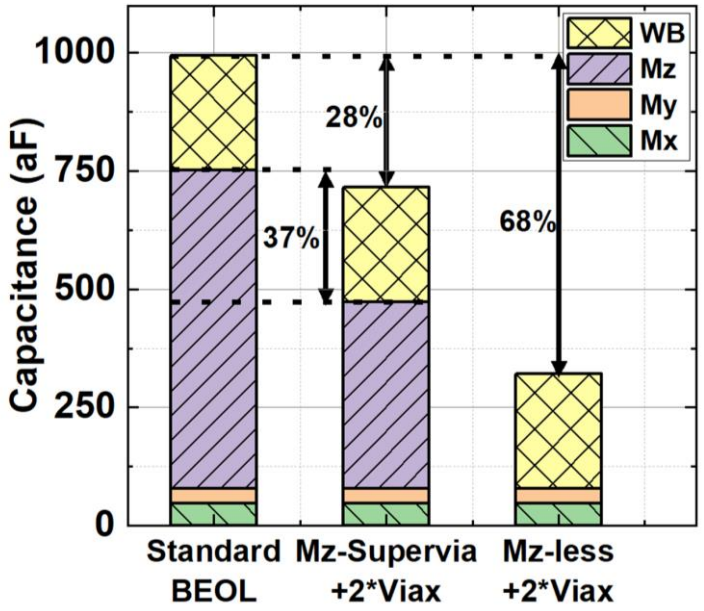
BEOL	Description
Standard	Default. Multiple M_X, two M_Y and two M_Z layers.
M_Z-Supervia	M_Z limited to 0.1 µm x 0.1 µm + 2X vias in M_X layers.
M_Z-less BEOL	M_Z layers eliminated + 2X vias in M_X layers.



3D BEOL

Extraction study





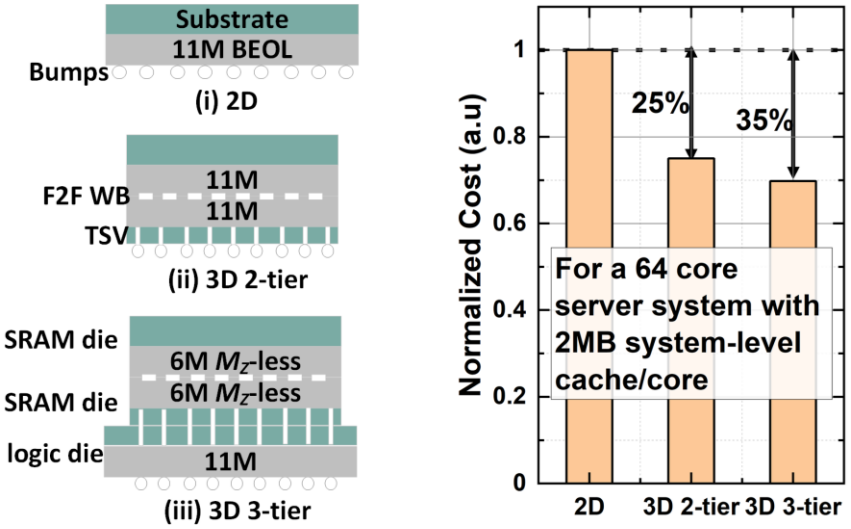
- M_X contribute 84% of the total resistance.
  - ➢ 2x VIA_X vias reduces resistance ~31%

- M_Z constitute 68% of the total capacitance
  - ➢ Not used in SRAM signal routing

3D BEOL

Cost study

- SRAMs typically only require:
  - Mx layers for signals
  - My layers for power
  - Mz-less BEOL ideal for 3D-Split SRAMs
- Cost reduction ~25-35%:
  - smaller (better yielding) dies
  - simplified metal stack
  - optimized process

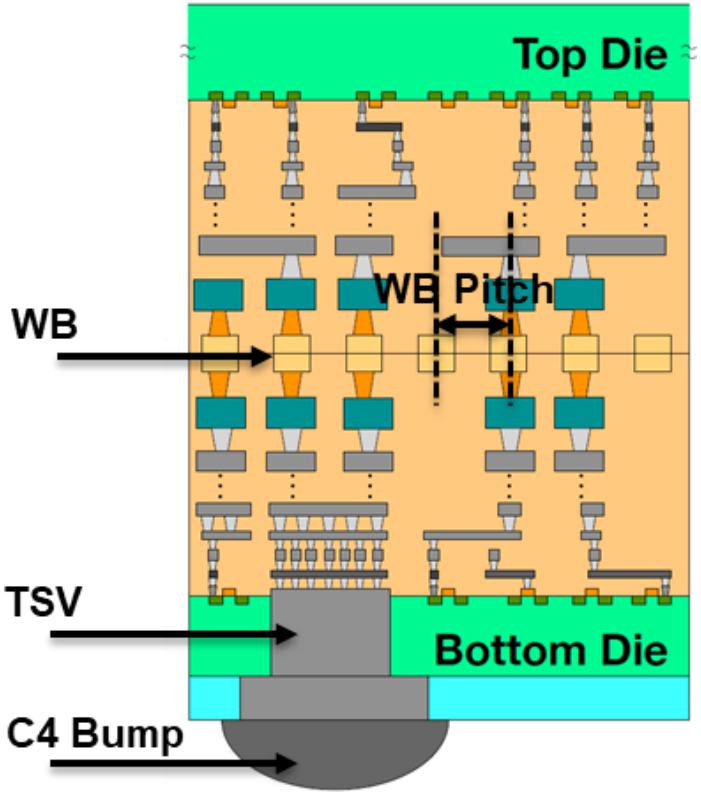


Cost-comparison at 12nm

3D BEOL

WB pitch recommendation

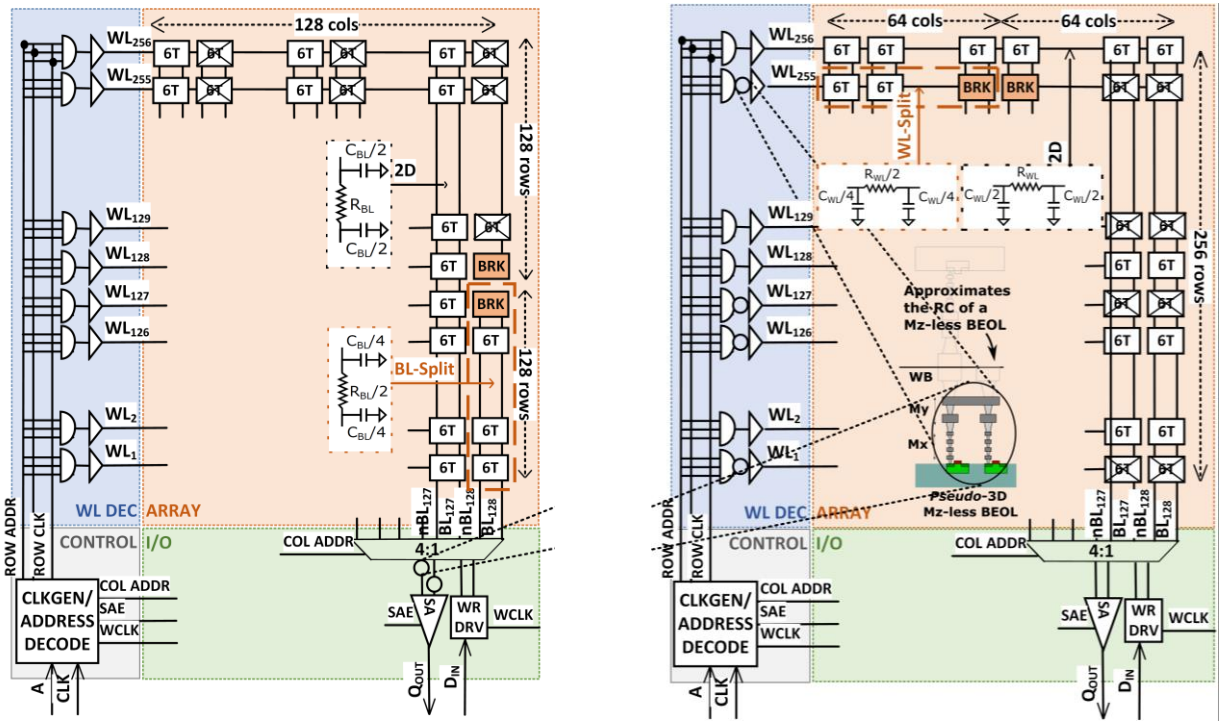
- Increasingly finer wafer bond (WB) pitches.
  - Pitches of ~1 µm on foundry roadmap.
- Pitch limitations can be alleviated:
  - Staggering the locations of WB
  - Requires extra routing
- WB pitch requirement
  - 3D-split SRAMs must be ~1 µm
  - Globalfoundries 12nm 3D test-vehicle WB pitch ~5.76 µm



3D-stack cross-section

Pseudo 3D-Split SRAM

Macro Design

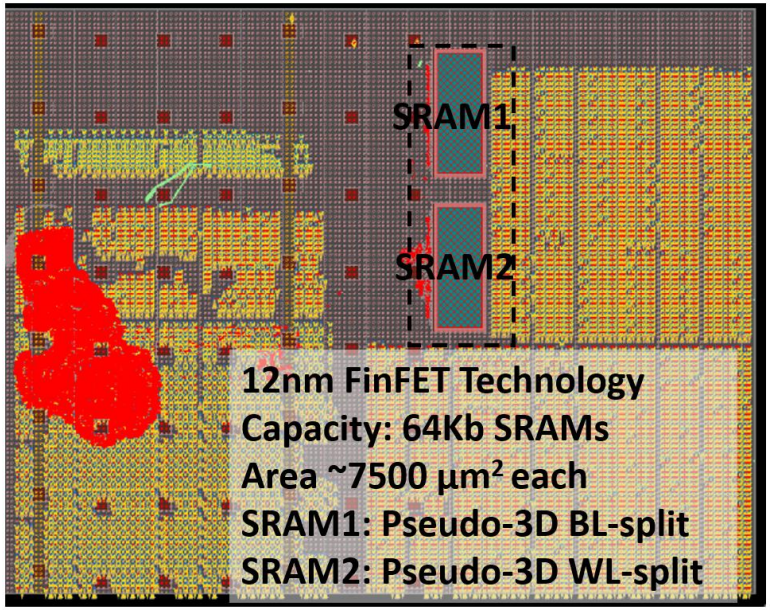


- Layout of 2D SRAM reconfigured.
  - Capture effects of BL-split and the WL-split 3D SRAM
  - A split by inserting break cells in rows or columns.
  - Effect of MZ-less BEOL by inserting a via structure and routing it back from top of M_Y.

Test chip Tape-out

GlobalFoundries 12nm FinFET process

- Integrated macro
  - Even address for 2D row.
  - Odd address for 3D-split row.
- Enables accurate comparison
  - Proximity of design points.
  - Less impact of on-chip process variation.
  - Bitcell share same peripheral circuits.

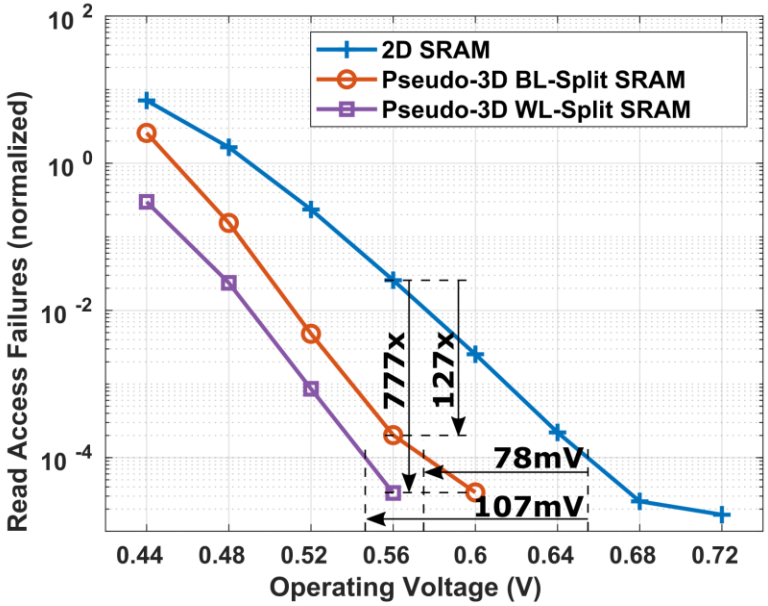


Physical layout view of prototype SRAM macros

Results – V_{MIN} improvement

Measured data

- Reduction in read access failures @0.56V
  - 127x for BL-split
  - 777x for WL-split
- Iso-read failure probability, V_{MIN} gain:
  - 78mV for BL-split
  - 107mV for WL-split
- V_{MIN} gain can be traded off for performance.



Read errors (normalized) across 58 dies (2.7 Mb of SRAM) at room temperature.

Conclusion: 3D-Split SRAM

- Comprehensive analysis of 3D-Split SRAM in an advanced CMOS node.
- Two separate approaches for reducing 3D-BEOL parasitics proposed:
  - Mz-Supervia
  - Mz-less
- WB pitch requirements to enable 3D-Split SRAM shared.
- Measurement results from 12nm Test Chip presented:
  - Vmin reduction ~107mV or equivalent performance gain ~15%
  - BL-split SRAMs offer ~14% lower power due to reduced BL capacitance.
- Gains equivalent to one technology node dimensional scaling.

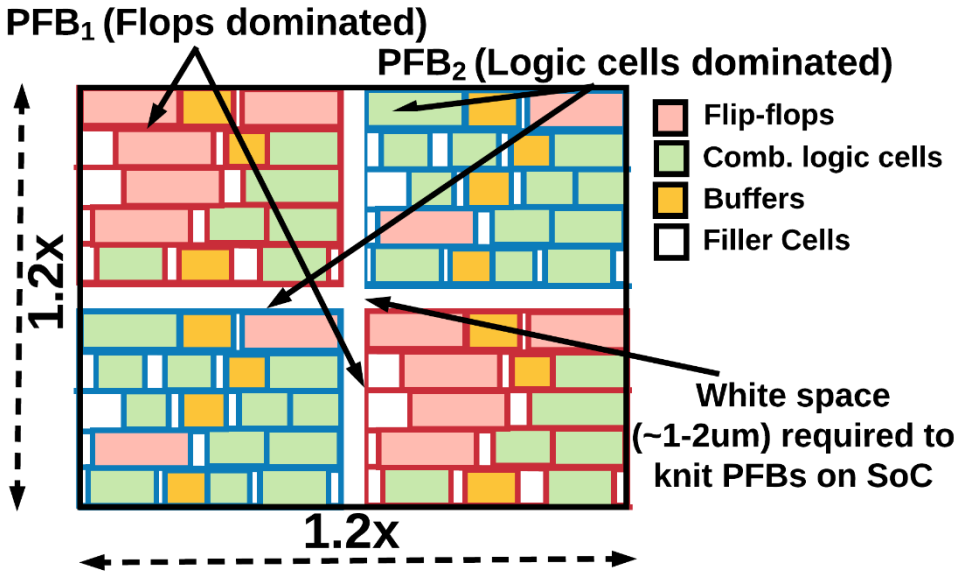
Heterogeneous Integration:

Microscale Modular Assembled ASICs (M2A2)

Reference: A. Sayal, P. Ajay, M. W. McDermott, S. V. Sreenivasan and J. P. Kulkarni, “M2A2: Microscale Modular Assembled ASICs for High-Mix, Low-Volume, Heterogeneously Integrated Designs,” IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (**TCAD**), vol. 39, no. 12, pp. 4760-4776, Dec. 2020 [**Paper**]

Proposed Idea: M2A2 Technology

- Pre-fabricated block (PFB): A micro scale circuit consisting of front-end critical mask layers (transistors, and front-end interconnects)
- Multiple types: logic, memory, IO
- PFB size: 50µm * 50µm to 1000µm * 1000µm
- M2A2 Design: PFB knitted SoC

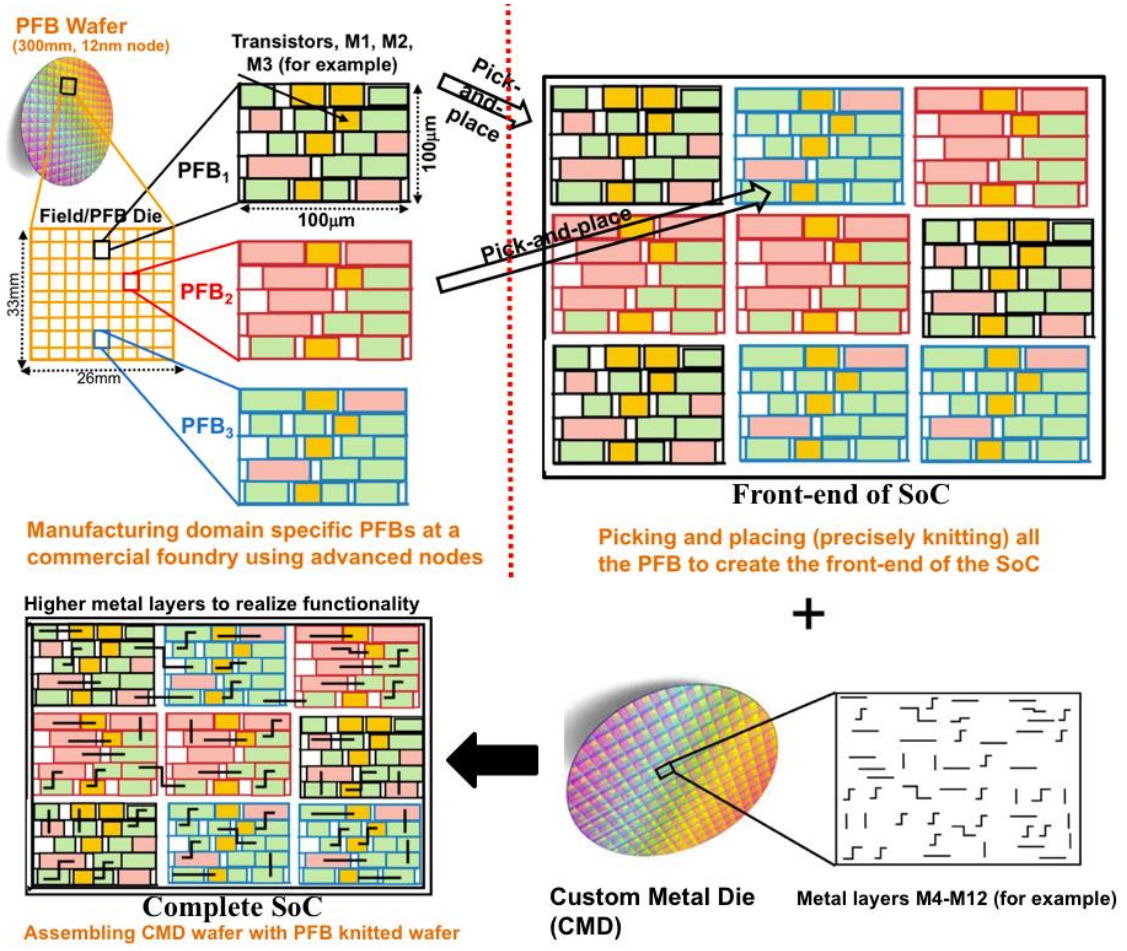


Central Idea:

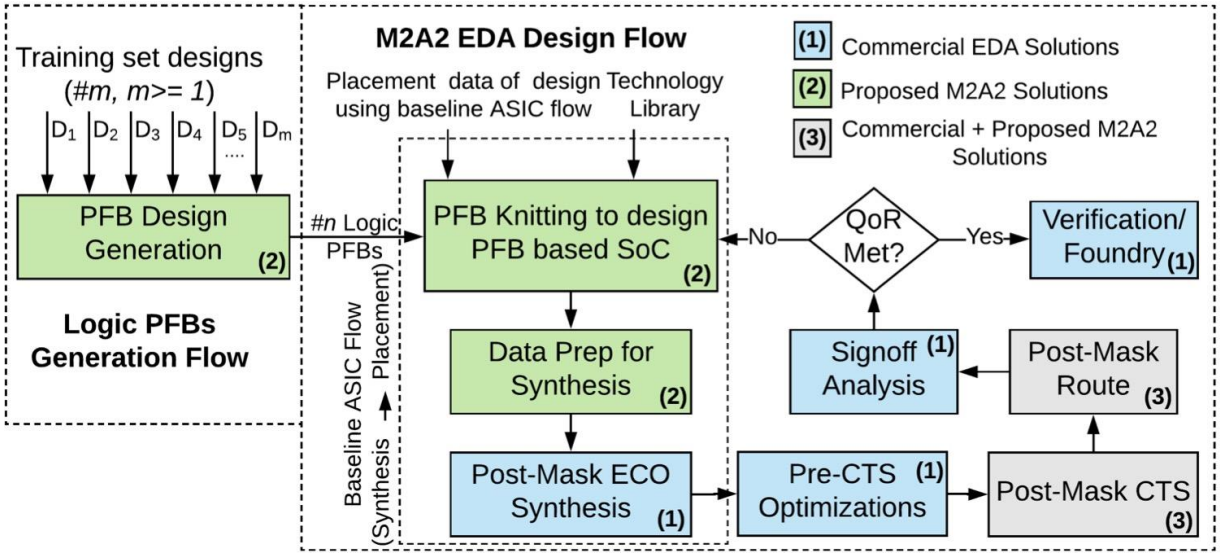
Enables sharing the mask cost for nano-scale feature sizes across many ASICs using limited number of PFBs, thus decreasing the non-recurring engineering costs for individual designs

M2A2 Technology in Action

- Manufacturing generic PFBs at a commercial foundry using advanced CMOS nodes
- Manufacturing custom metal die comprising of higher metal layers at a trusted foundry
- Knitting PFBs and CMD using a pick-and-place fabrication technique at a trusted foundry



Overview of M2A2 EDA methodology



Proposed M2A2 EDA Flow for logic PFB based SoC

- **Unsupervised machine learning** and **bi-partite graph matching** techniques for optimal PFB design, and knitting of PFBs on SoC
- Leverages Cadence Conformal ECO tool infrastructure to perform synthesis
- CTS/backend design optimizations using **graph matching** technique and commercial EDA solutions

M2A2 EDA Design Parameters and Benchmarks

- 15 IWLS'05 benchmarks are used - larger benchmarks from functional categories: encryption standards, processors, controllers, communication IPs, and peripherals
- 9 out of 15 benchmarks are used in training process to generate the PFB library
- M2A2 designs performance is compared with ASICs, sASICs and FPGAs at 130nm, 40nm and 16nm

(a)

Layer	Pitch (nm)
Gate	70
M0	56
M1	70
M2	56
M3	56
M4	80
M5	100
M6	160
M7	160
M8	160

Layers in PFB

Layers in CMD for routing

(b)

Layer	Pitch (nm)
Gate	70
M0	56
M1	70
M2*	56
M3*	56
M4	80
M4_1	80
M4_2	80
M5	100
M6	160
M7	160
M8	160

** M2-M3 layers are used to form super vias in PFBs*

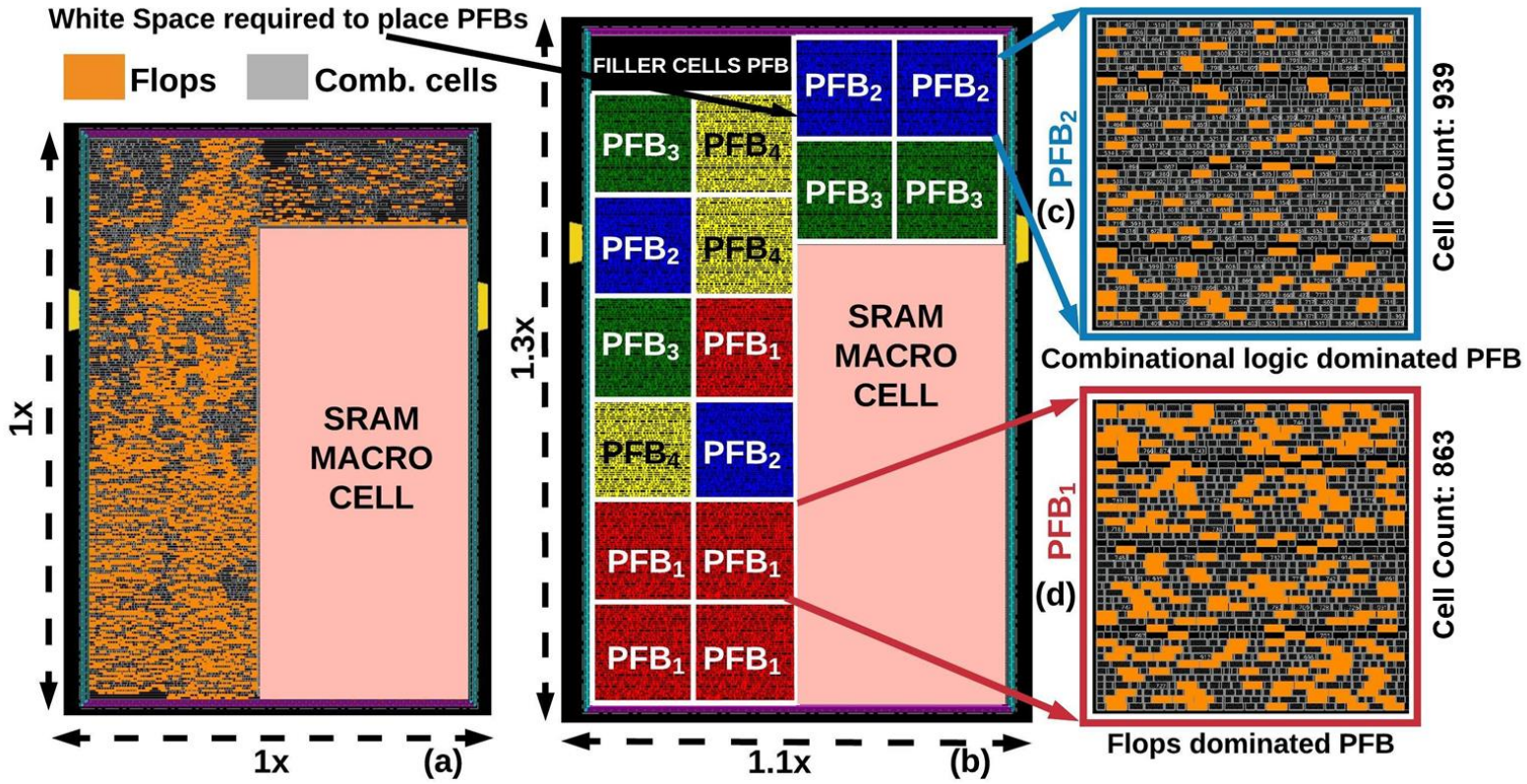
Metrics	130nm	40nm	16nm	16nm*
PFB Width	110.00um	55.00um	28.00um	31.00um
PFB Length	110.16um	55.44um	28.45 um	30.72 um
#PFB Types	5	4	4	4
PFB metal layers	M1	M1	M0, M1	M0-M3
Backend CMD metal layers	M2-M8	M2-M8	M2-M8	M4-M8

**PFB comprises of super-via (M1-M3), CMD comprises of M4-M8*

Metrics	130nm	40nm	16nm
Device Family	Virtex-II	Virtex-6	Virtex UltraScale+
Device	xc2v250	xc6vlx75t	xcvu3p
Package	fg456	ff484	ffvc1517
Speed Grade	-6	-2	-3

FPGA Design Metrics

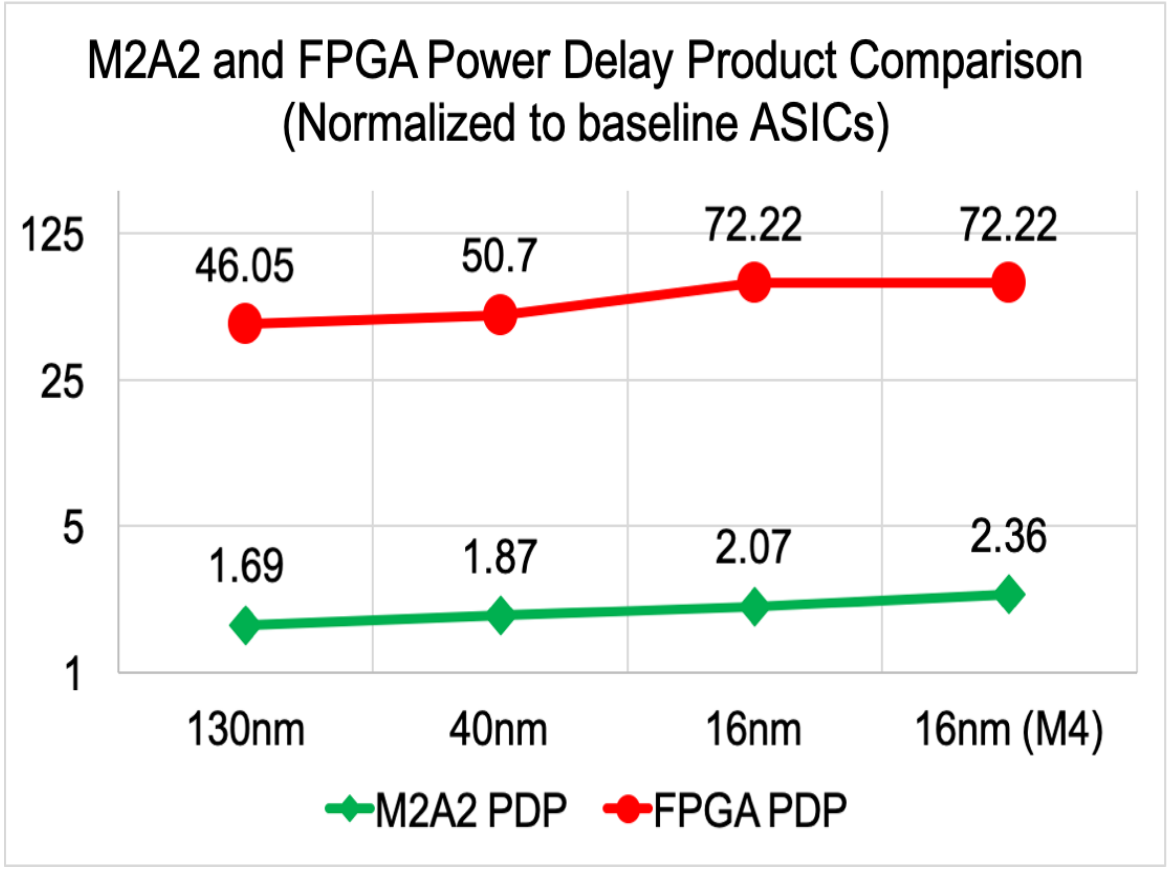
M2A2 EDA Design Parameters and Benchmarks

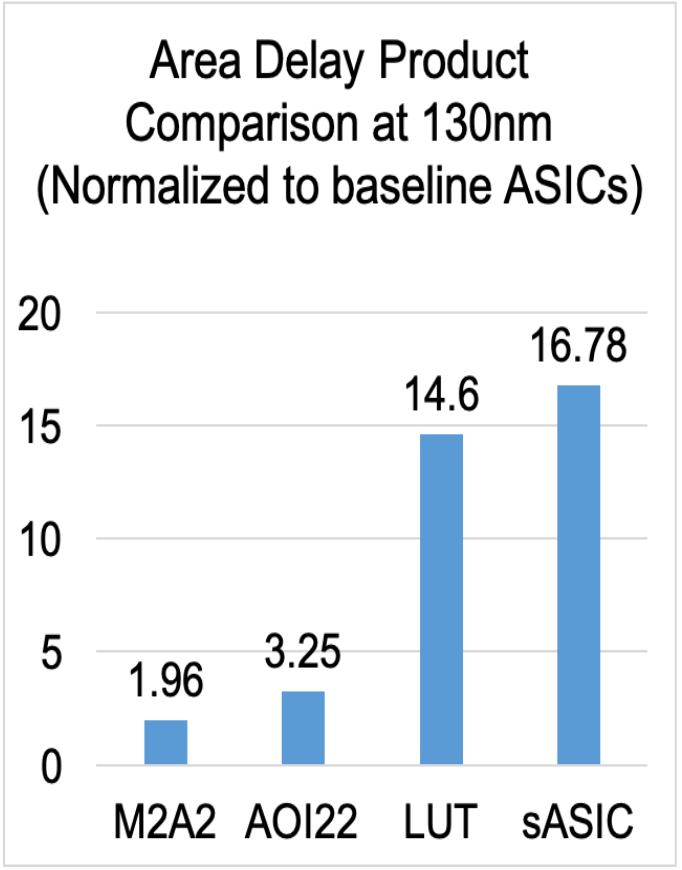


40nm Ethernet IP floorplan comparison: (a) Baseline ASIC flow (b) Proposed M2A2 flow, (c) Combinational logic dominated pre-fabricated block (PFB_2), (d) Flops dominated pre-fabricated block (PFB_1).

*4 types of PFBs each sized to 55μm * 55.44um are used*

Performance Summary – IWLS Benchmarks





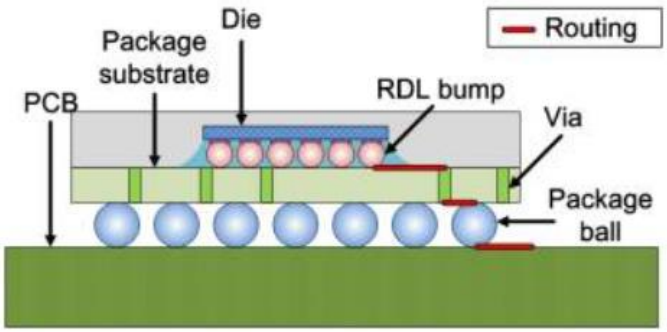
Summary: M2A2 Improvements

- Over 15 IWLS benchmarks, PDP benefit of **27.11x-34.89x** over FPGAs, and are **1.69x-2.36x** worse compared to baseline ASICs
- M2A2 designs achieve **15%-68.5%** smaller area and **8.5%-52%** higher performance compared to earlier proposed sASIC methodologies
- M2A2 Accelerators: Energy Efficiency degraded by **~40%** over baseline ASICs

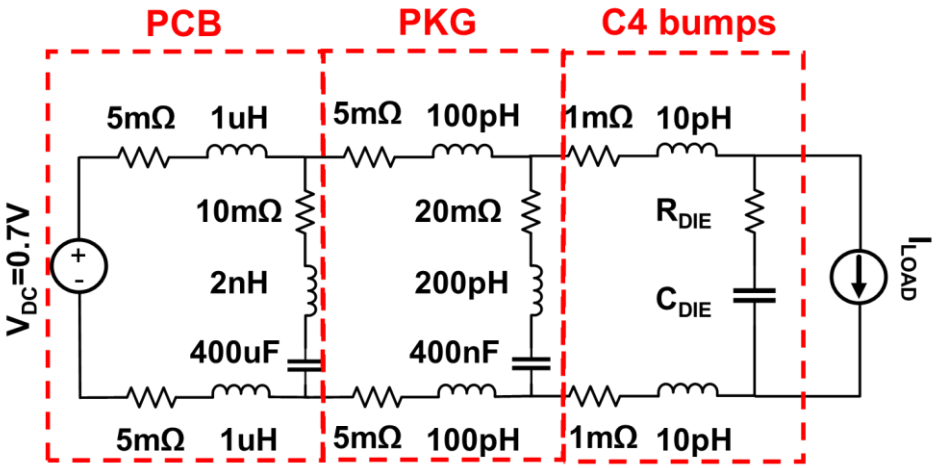
Metric	Baseline ASICs	Structured ASICs	FPGAs	M2A2 (This work)
Non-recurring Engineering (NRE) cost	Highest	Medium	Lowest	Low
Time-to-Market (TTM)	Highest	High	Lowest	Low
Power-Performance-Area (PPA)	Lowest (1x)	High (~5x)	Highest (~500x)	Medium (~2.5x)
Supports Heterogeneous Integration	No	No	No	Yes
Makes use of commercial CAD flows	Yes	No	Yes	Yes
Ability to fabricate secured chips at cutting edge nodes	No	No	No	Yes
Contribution of interconnect delays and congestion	Low	Medium	High	Low
Placement of cells (logic, buffers, flops)	Unrestricted	Restricted	Restricted	Flexible

Backup

Off-chip voltage droop



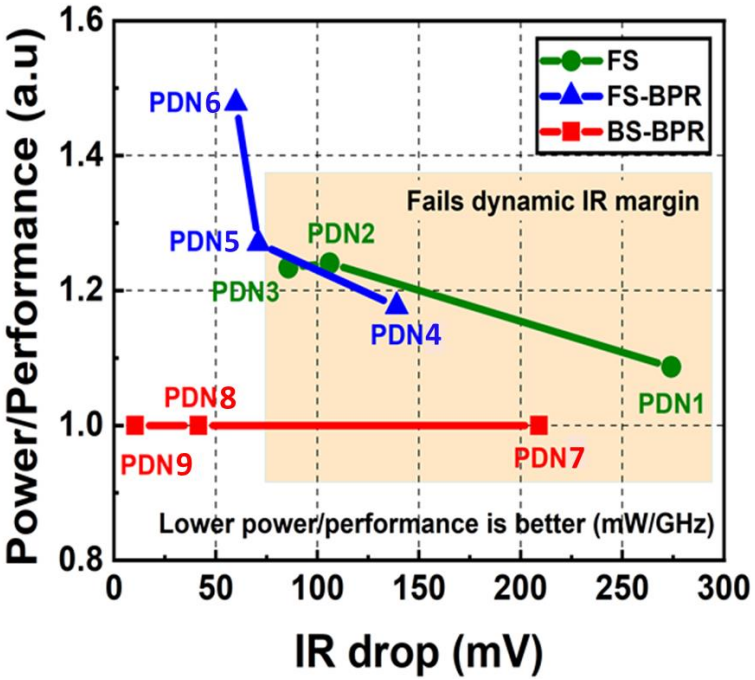
Die-package-PCB schematic [2]

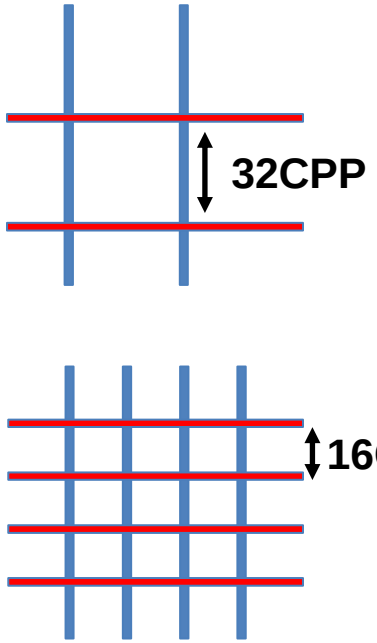


Die-package-PCB equivalent circuit

- Die/Package/PCB connections introduce parasitic inductance

IR drop comparison of PDN configurations



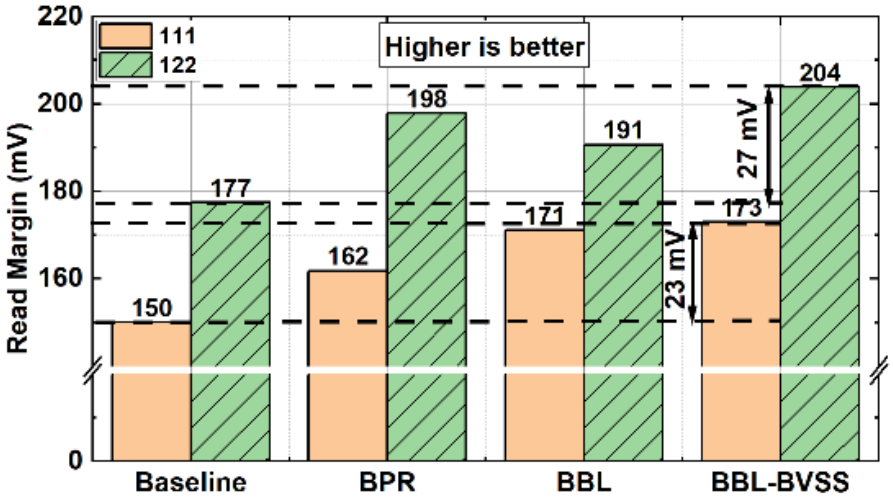


Sparse pitch PDN – high IR drop, less congestion

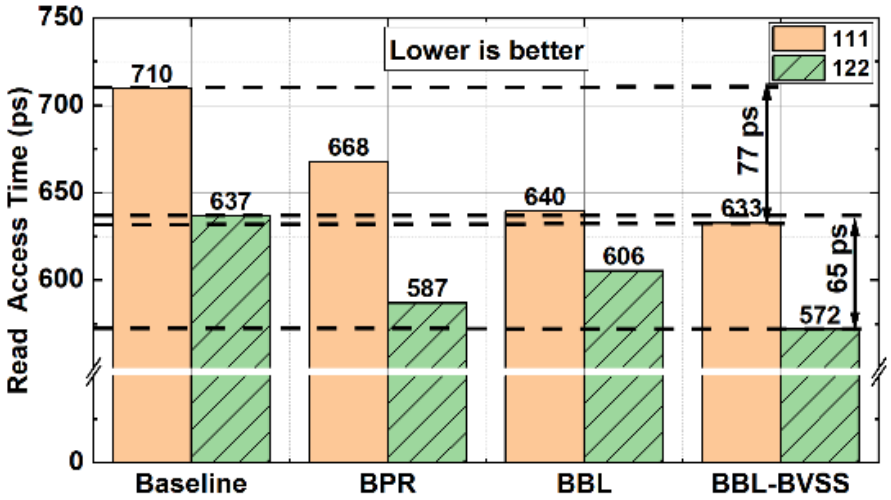
Tight pitch PDN – low IR drop, more congestion

- FSBPR has to expend more energy to meet the IR drop target
- BSBPR effectively decouples the trade-off between Power/Performance and IR drop

Macro-level Results – Read (SS/0.63V/-40C)

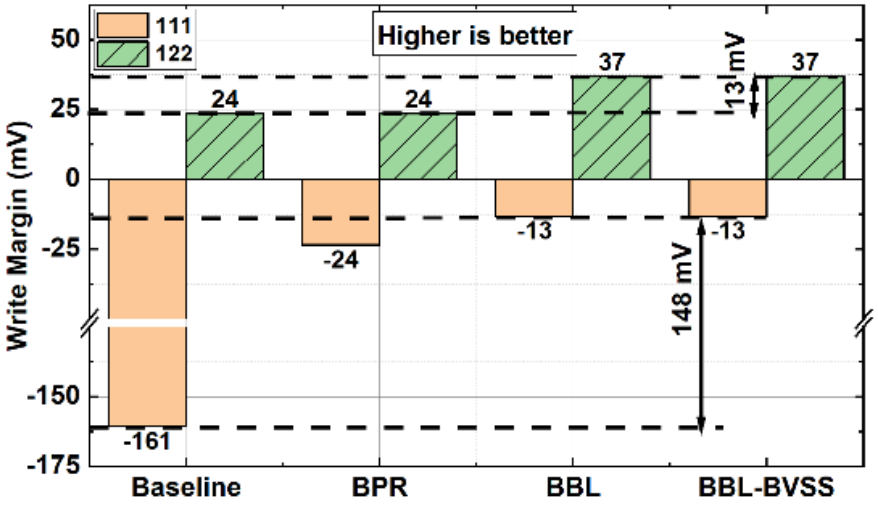


Read margin - bitline differential at sense amplifier trigger

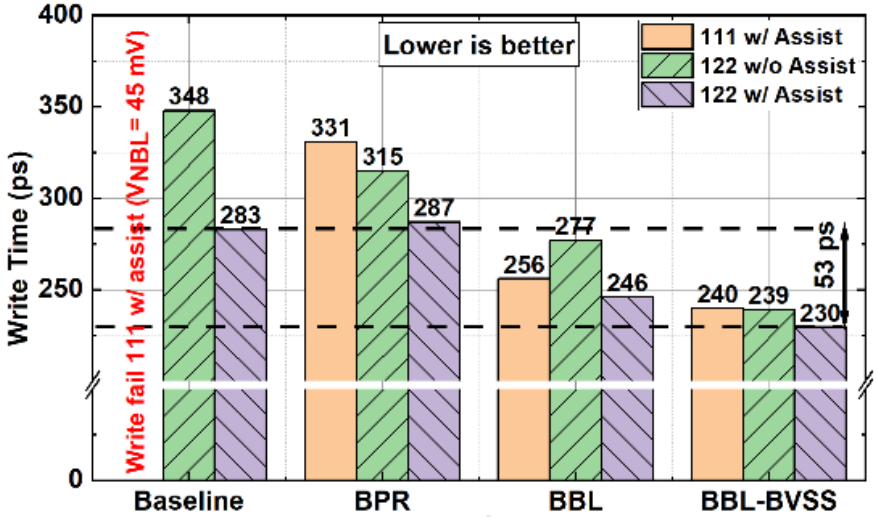


Read Access time – CLK to Q (output) timing

Macro-level Results – Write (SF/0.63V/-40C)



Static write margin - minimum V_{NBL} to flip the SRAM cell



Write time - WL rise to bit flip