CICC

IEEE Custom Integrated Circuits Conference

3D-Split SRAM: Enabling Generational Gains in Advanced CMOS

R. Mathur[1,3], M. Bhargava[1], H. Perry[1], A. Cestero[2], F. Frederick[1], S. Hung[1], C. Chao[1], D. Smith[2], D. Fisher[2], N. Robson[2], X. Xu[1], P. Chandupatla[1], R. Balachandran[1], S. Sinha[1], B. Cline[1], J. P. Kulkarni[3]

[1]Arm Inc, Austin, TX, USA. [2]GLOBALFOUNDRIES, NY, USA. [3]University of Texas at Austin, TX, USA.

03/19/2021

SSCS IEEE SOLID-STATE CIRCUITS SOCIETY™

IEEE

Presenter Bio

Rahul Mathur is currently pursuing his Ph.D. at the University of Texas at Austin, TX, USA. He is pursuing a Ph.D. part-time while working at ARM Austin where he has been since 2012. At Arm, he has led multiple memory compilers at sub-10nm foundry platforms. He has filed 15 US patents and serves in the Patent Review Committee of Arm. His research interest is System-Circuit-Device Design Methodologies for 3D-IC. He is a senior member of IEEE.

Outline

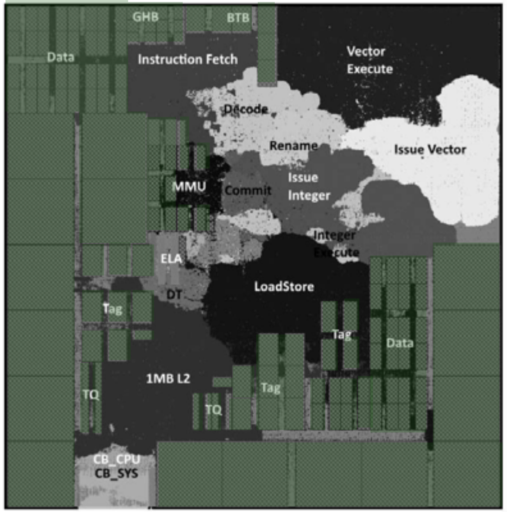
- 2D SRAM Challenges
- Motivation for 3D-Split SRAM
- 3D-BEOL
- 3D Split-SRAM Macro Design
- Measurement Results
- Summary

2D SRAM Challenges

Capacity demands

- Data deluge arising AI, IoT, automotive etc.
- Increasing demand for larger SRAM capacities.
- SRAM area dominates the floorplan of modern CPUs.

Floorplan of Arm CPU in a FinFET technology[1]. Caches occupy ~50% area (green highlight).



[1]R. Christy et al., 2020 ISSCC, San Francisco, CA, USA, 2020, pp. 148-150

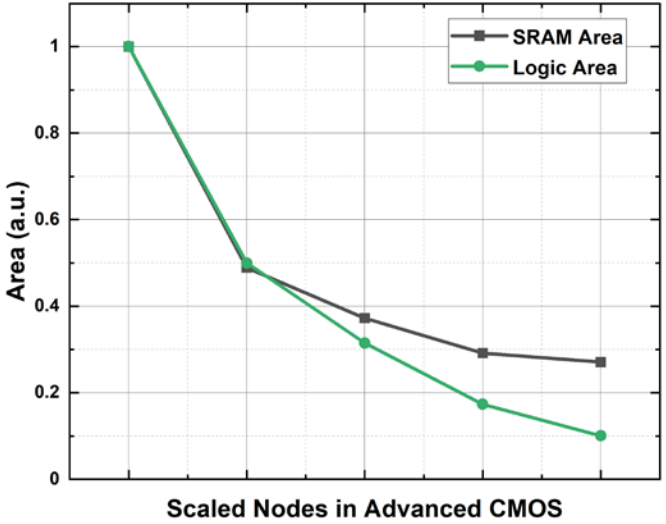
2D SRAM Challenges

Scaling trends

- SRAM scaling challenged by:
  - Gradual shrinking of critical pitches
  - High contact resistance
  - Constrained design rules
  - WL/BL resistance

- To extend SRAM scaling gains:
  - Stacking standalone SRAMs
  - 3D-Split-SRAM

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



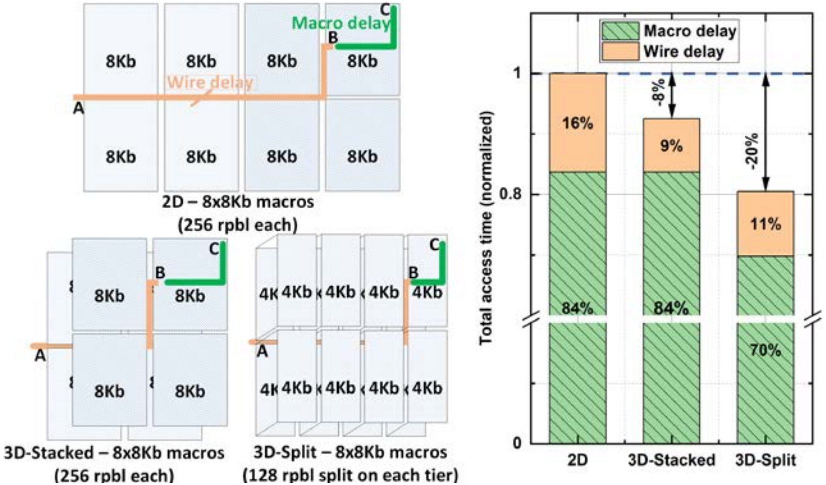
MOTIVATION

3D Stacked Vs 3D-Split

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



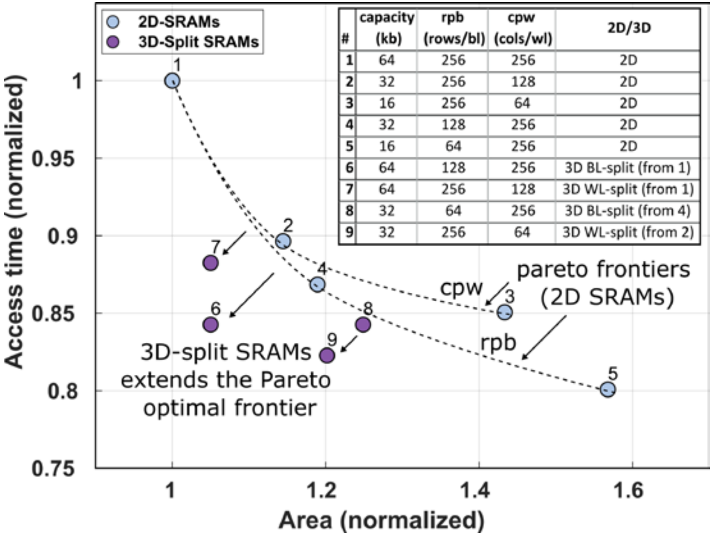
MOTIVATION

2D Vs 3D-Split

- 3D-Split SRAM Vs 2D
  - Fast access-time
  - Low area
  - Lower leakage power

- Feasibility and efficacy depend on:
  - Pitch restrictions of 3D-BEOL
  - RC parasitics of 3D-BEOL

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)

pareto frontiers (2D SRAMs)

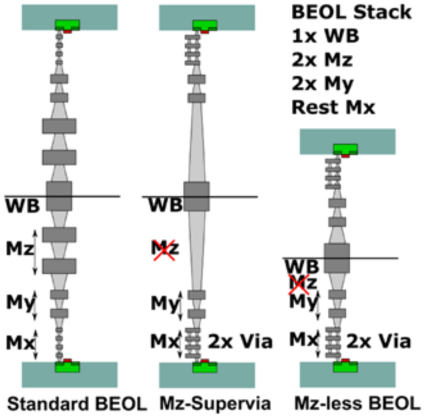
3D-split SRAMs extends the Pareto optimal frontier

3D-BEOL

Extraction study

- Goal: Analyze metal stack in 12nm FinFET
  - Assess the RC overhead of 3D-BEOL
  - identify opportunities of 3D-BEOL 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.

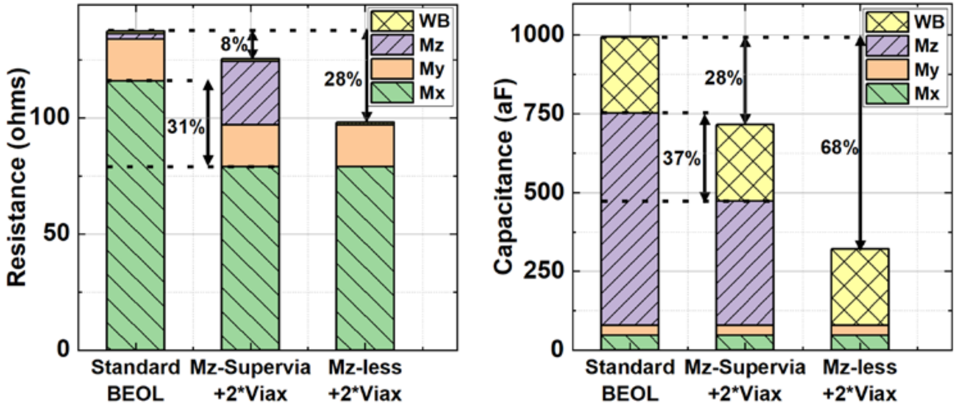


3D-BEOL

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

RC analysis with proposed DTCO of 3D-BEOL



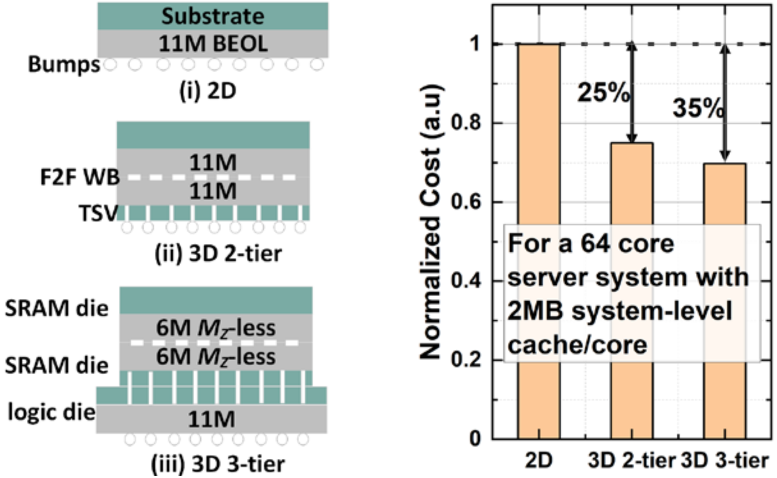
3D-BEOL

Cost Analysis

- 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 of a 3-tier system.



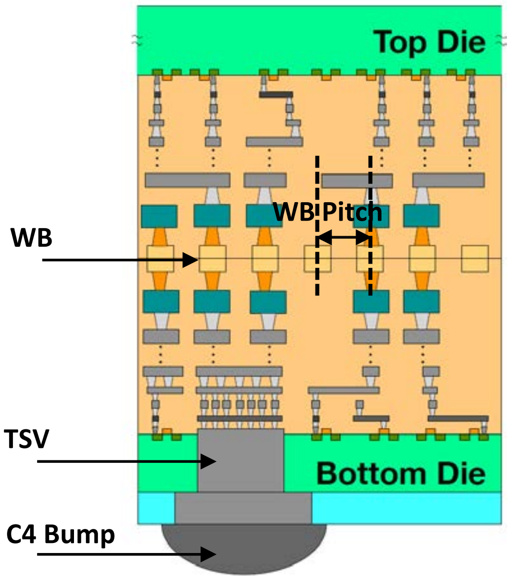
For a 64 core server system with 2MB system-level cache/core

3D-BEOL

Wafer bond (WB) Pitch recommendation

- Steady improvement in WB technology – finer WB pitches.
- Pitch limitations can be alleviated:
    - Staggering the locations of WB
    - Requires extra routing
- WB pitch requirement
    - 3D-split SRAMs must be ~1 µm
    - GF 12nm 3D test-vehicle WB pitch ~5.76 µm
    - Pitches on ~1 µm on foundry roadmap.

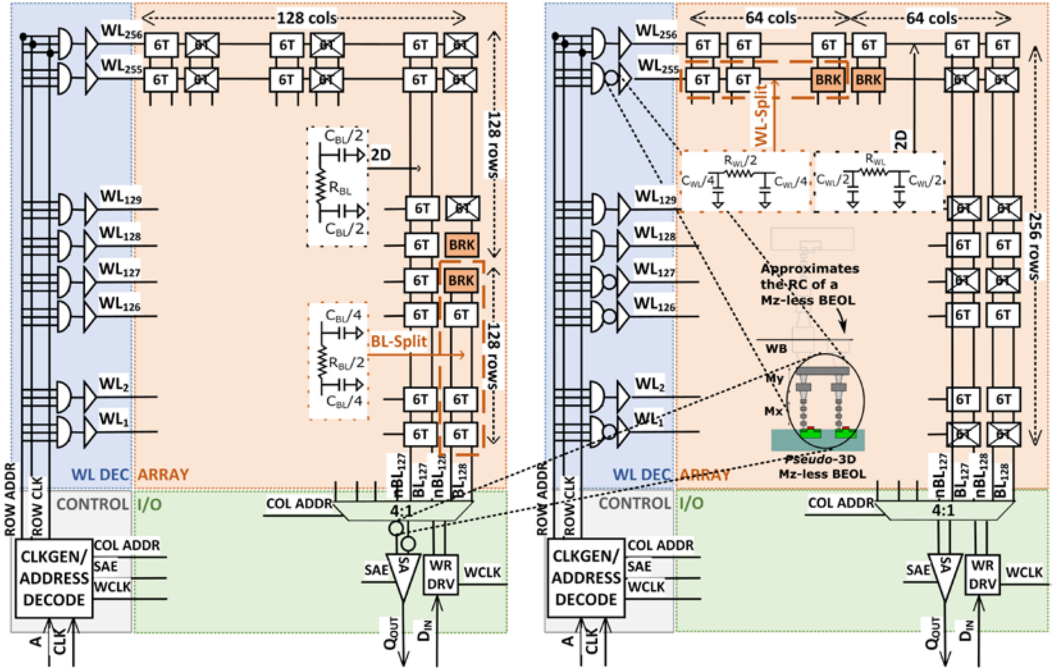
3D-stack cross-section



Integrated 2D and pseudo 3D-Split SRAM

Macro Design

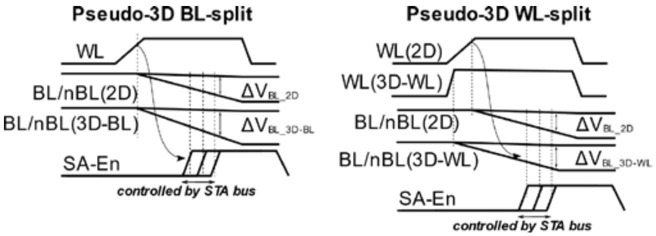
- Layout of 2D SRAM reconfigured.
  - Capture effects of BL-split and the WL-split 3D SRAM
  - A split by inserting break cells in rows or columns.
  - Effect of MZ-less BEOL by inserting a via structure and routing it back from top of M_Y.



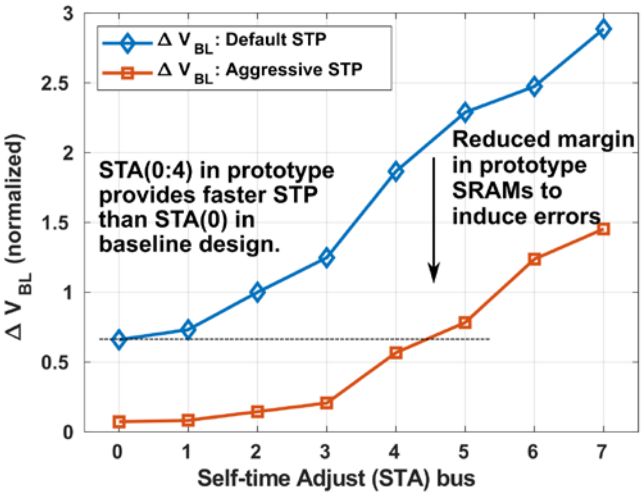
Integrated 2D and pseudo 3D-Split SRAM

Macro Design

- Margins controlled by Self-Time Path (STP).
  - STP re-tuned to push the margins.
- STP is also externally adjustable by the Self-Time Adjust (STA) bus.



Simulated read margin (ΔVBL)



STA(0:4) in prototype provides faster STP than STA(0) in baseline design.

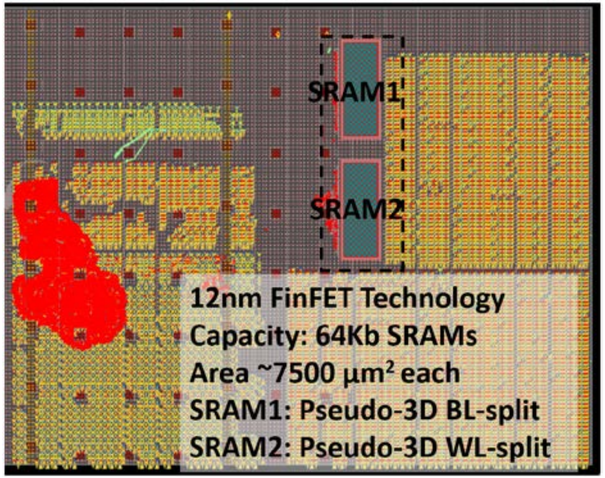
Reduced margin in prototype SRAMs to induce errors

Integrated 2D and pseudo 3D-Split SRAM

12nm Testchip with GF

- 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 fabricated in 12nm FinFET process.



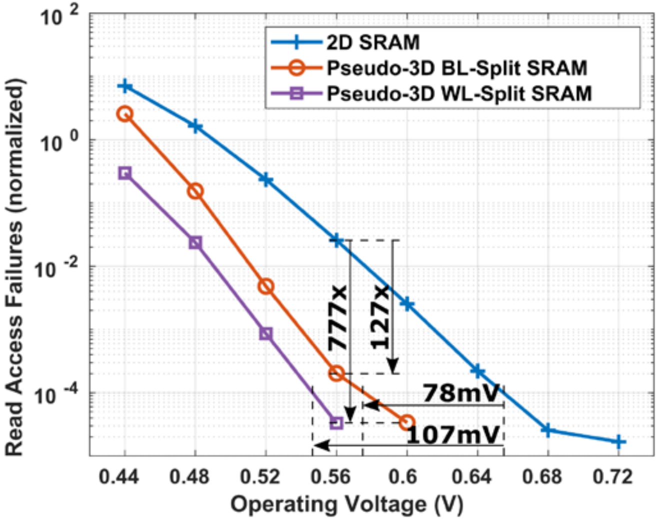
12nm FinFET Technology
Capacity: 64Kb SRAMs
Area ~7500 µm^2 each
SRAM1: Pseudo-3D BL-split
SRAM2: Pseudo-3D WL-split

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.



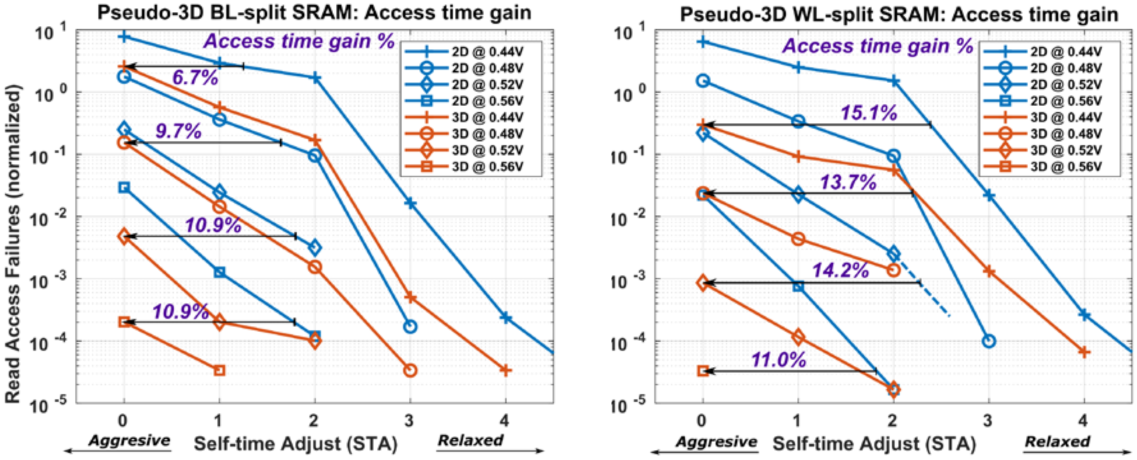
Results – Access time improvement

Measured data

- At iso-read failure probability, performance gain:
  - BL-split: 6.7-10.9%.
  - WL-split: 11-15.1%

- The measured access time gain matches simulation estimate of ~15%.

Gain in access time estimated by STA setting shift for iso-errors.



Summary

- A comprehensive analysis of 3D-Split SRAM in an advanced CMOS node.
- Two separate approaches for reducing 3D-BEOL parasitics are proposed:
  - Mz-Supervia
  - Mz-less
- WB pitch requirements to enable 3D-Split SRAM shared.
- Measurement results from prototype 12nm FinFET SRAM macros, capturing effects of BL- and WL-split designs and Mz-less 3D-BEOL are presented:
  - Vmin reduction ~107mV
  - Performance gain ~15%
  - BL-split SRAMs offer ∼14% lower power due to reduced BL capacitance.
- Gains equivalent to the performance gains from one technology node dimensional scaling.

OBRIGADO
gracias
どうも
ARIGATO
hvala
DANK U
takk
MERSI
merci
obrigado
danke schön
köszi
سپاس
PALDIES
muchas gracias
ありがとう
TEŞEKKÜR EDERIM
MOLTE GRAZIE
GO RAIBH MAITH AGAT
謝謝
danke
grazas
THANK YOU
благодаря
GRAZZI
TAK
どうも
THANKS
qujan
TAK
благодаря
asante
muchas gracias
vielen dank
PALDIES
DANK U
OBRIGADO
mesi
danke
grazie
DZIEKI
спасибо
MULTUMESC
TACK
Gràcies
多謝
شكراً
감사합니다
köszi
TEŞEKKÜR EDERIM
NA GODE
muchas gracias
obrigado