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.

Abstract— 3D integration technologies are becoming increasingly viable to mitigate the limitations and slowdown in traditional 2D transistor scaling. 3D-Split SRAMs, realized by splitting the bitlines (BL) and/or wordlines (WL) across two or more 3D-arranged tiers, promise improved power/performance due to reduced *RC* parasitics. However, their feasibility and efficacy depend on the pitch and *RC* parasitics of the inter-tier BEOL connections (*3D-BEOL*). In this work, we analyze the impact of *3D-BEOL* on the 3D-Split SRAM gains, in a face-to-face (F2F) hybrid wafer bonding 3D integration technology. Two separate approaches for reducing *3D-BEOL* parasitics viz (1) M_Z-Supervia, & (2) M_Z-less *3D-BEOL* are proposed. Measurements from 64 Kb 12 nm FinFET SRAM prototype, reconfigured to capture the BL-split and the WL-split 3D SRAM effects, show up to 107 mV lower V_{min} as well as ~15% better access-time, equivalent to the performance gains from one technology node dimensional scaling.

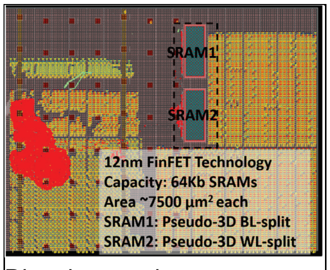
Introduction: SRAM is a foundational component of modern SoC designs. However, the pace of SRAM scaling has not kept up with its logic counterpart in advanced CMOS nodes, due to constrained front-end-of-line (FEOL) and middle-of-the-line (MOL) design rules in bitcells as well as increasing RC parasitics of the long WL and BL in lower metal layers. To extend SRAM scaling gains beyond conventional 2D scaling, multiple 3D SRAM approaches such as stacking standalone SRAMs [1], [2], Split-SRAM [3-5], and monolithic 3D SRAMs [6] have been proposed. Previous works on 3D-Split SRAM do not capture the overheads or pitch restrictions from the inter-tier *3D-BEOL*. In this work, we analyze F2F hybrid wafer bonding technology with ~1μm [7] interconnect pitches and high alignment accuracy (3σ <100nm) [8]. Key contributions of this work are (1) impact analysis of *3D-BEOL* parasitics on 3D-Split SRAMs, (2) optimized BEOL proposals for 3D-Split SRAMs: M_Z-Supervia, and M_Z-less BEOL, (3) measurements from 12nm SRAM prototypes emulating 3D-Split SRAMs and an *M_Z-less* BEOL.

2D vs. 3D-Stacked vs. 3D-Split SRAM: *Stacking* standalone SRAM arrays over other SRAM arrays/logic-blocks [1-2] can improve the access-time by the reduction in routing delay to-and-from the farthest memory (node A to B to C and back) by ~8% (Fig. 1.a). *Splitting* an SRAM macro across multiple tiers (3D-Split SRAM) can further improve the access-time due to the reduction in the device and RC parasitics of WL and/or BL by up to 20% (Fig. 1.b). Faster access-time can also be achieved by smaller capacity 2D SRAMs but at the expense of density and leakage power. Fig. 2 shows this trade-off for macros of different configurations in 2D and how 3D-Split SRAMs can extend the delay-area Pareto frontier of 2D SRAMs.

Mitigating *3D-BEOL* Parasitics: Extraction of a standard BEOL stack in 12nm reveals that the top metal layers (M_Z) constitute 67% of the total capacitance (Fig. 3.d), while the lowest (M_X) layers contribute 84% of the total resistance (Fig. 3.e). Doubling the number of M_X-vias reduces the total resistance by 31% with a 2% increase in capacitance. The high capacitance of M_Z layers can diminish the benefits of 3D-splitting. The impact will only increase at scaled nodes as the capacitance of BL and WL will scale while M_Z layers remain largely unchanged. Note, SRAMs typically only use M_X and M_Y for signal and power routing. Hence, there is an opportunity to optimize M_Z-BEOL in 3Ds-split SRAM tiers differently without limiting the BEOL needs of circuits on other 3D tiers. We present two separate approaches to optimize the M_Z layer parasitics for 3D-Split SRAMs:

(1) M_Z-Supervia: Multi-level via or Supervia technology, having an aspect ratio of 5-10, has been proposed to lower M_X metal layer *RC* parasitics compared to conventional dual-damascene via technology [9]. This approach can be extended to the M_Z layers for a direct connection between Wafer-Bond (WB) and the highest M_Y layer in a 3D-Split SRAM design (Fig 3.a). Compared to the standard *BEOL*, the M_Z -Supervia results in 28% reduced capacitance and 9% reduced resistance.

(2) M_Z-less *BEOL:* This approach eliminates the M_Z layers resulting in 68% lower capacitance and 28% lower resistance compared to the standard BEOL. Fig 3.b shows an example configuration of a F2F M_Z-less 3D-Split SRAM tiers bonded B2B with a logic die with 11M BEOL. The cost can be reduced by ~25-35% with smaller (better yielding) dies, simplified metal stack & optimized process (reduced device V_{th} options) for SRAM dies (Fig 3.c).



Die micrograph.

Integrated 2D and *pseudo*-3D SRAM macro design: Prototype SRAM macros were fabricated in a 12nm FinFET process. These macros were designed by modifying the layout of a memory-compiler-generated 64kb 2D SRAM macro with 1-2-2 6T cells, to incorporate the effects of 3D-Split SRAMs. The effects of M_Z-less BEOL and device parasitics are incorporated by designing a via pillar that goes up to the highest M_Y layer and routes back to connect to the subsequent sense amplifier inputs (in the BL-split case Fig 4.a) or WL drivers (in the WL-split case Fig 4.b). BL-Split design with ~1um (0.5um) WB pitch was made possible by a double (single) row of staggered WB connections of post-mux BLs in the I/O. The macros were designed such that the same macro can provide measurements for the baseline 2D SRAM and the *pseudo*-3D SRAMs. For the BL (WL)-split SRAMs, every even address (row) accesses a 2D SRAM bitcell (row) and every odd address (row) accesses a *pseudo*-3D BL(WL)-split SRAM bitcell. BL and WL split within the integrated macro are realized by inserting break cells in rows and columns, respectively. The physical proximity of all design variants in the integrated macro enables an accurate comparison with minimal impact of variations due to on-chip process gradients. Further, all bitcells share the same periphery circuits such as read/write control paths, sense amplifiers, and supply voltage variations thus minimizing the variations induced by the peripheral circuits. Since the baseline SRAM has large built-in design margins, the memory internal self-time path (STP) was re-designed to aggressively push the read margins to induce failures (Fig 5.b) The measurements were performed across a cumulative 2.7 Mb of SRAM (58 dies) at room temperature over a voltage range of 0.44V-1V for both write and read failures.

Measurement results: Measurements were performed on prototype 12nm FinFET SRAM macros, capturing M_Z-less *3D-BEOL* and effects of BL-split and WL-split designs. *Pseudo*-3D BL-split (WL-split) SRAM shows a Vmin improvement of up to 78mV (107mV), at iso-speed and iso-read access failure probability (Fig. 5.c). At operating voltage of 0.56V, this translates to up to 127x (777x) reduction in read access failures. The increased read access margin can be traded off to achieve performance gains. Fig. 6.b and 6.c show read access failures as a function of the Self Time Adjust (STA) setting for *pseudo*-3D BL-split and WL-split designs. Compared to baseline 2D SRAM, for iso-failure probability, the performance of the *pseudo*-3D BL-split (WL-Split) design improves by 6.7-10.9% (11-15.1%) across supply voltage of 0.44-0.56V; which is equivalent to speed-up gains obtained with one logic technology node scaling [10], [11]. Further, BL-split SRAMs offer ~14% lower power due to reduced BL capacitance. The results suggest that 3D-Split SRAM can be employed as a foundational design enabling generational performance gains in advanced CMOS.

References:

[1] https://bit.ly/3gQmpvM
[2] K. Ueyoshi, et. al., ISSCC 2018
[3] K. Puttaswamy, et. al., Trans. on Comp, vol. 58, no. 10, 2009.
[4] X. Xu, et. al., ISLPED 2019, pp. 1-6, doi: 10.1109.
[5] H. H. Nho et al., CICC 2008, pp. 201-204
[6] P. Batude et. al., 2015 Symp. on VLSI Technology, pp. T48-T49.
[7] S. Kim et al., ECTC 2020 pp. 216-222.
[8] www.tel.com/product/synapse.html
[9] A. Gupta et. al., Microelectronic Engineering, 2018.
[10] H. C. Lo et al., 2018 Symp. on VLSI Technology, pp. 215-216.
[11] S. Natarajan, 2014, IEDM, pp. 71-73.

978-1-7281-7581-2/21/$31.00 ©2021 IEEE

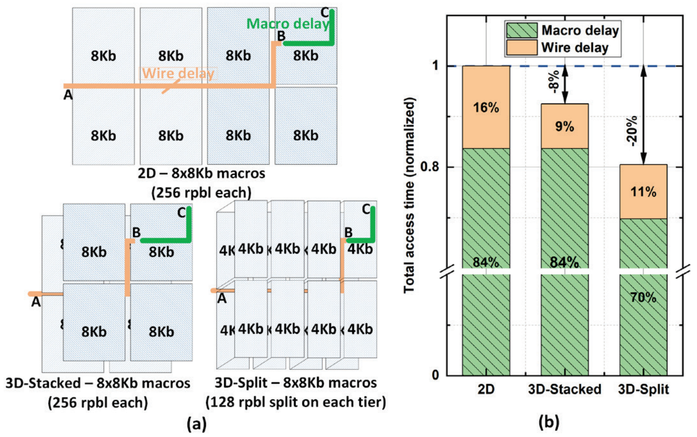
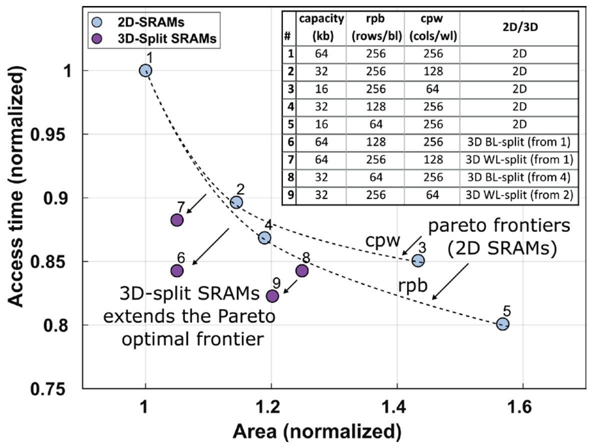


Fig. 1. (a) 2D and 3D configurations of 64Kb L1 cluster. Point A to B denotes wire delay while point B to C denotes the macro's access delay. (b) Access delay simulated in 12nm process at SS/(V_{NOM}-10%)/-40°C.



#	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)

Fig. 2. 3D-Split SRAMs extend the delay-area Pareto optimal frontier achievable by 2D-SRAMs. Details of macro configurations are provided in the table. SRAM macros #1-5 are 2D SRAMs, and #6-9 are 3D-Split SRAMs derived from 2D SRAMs.

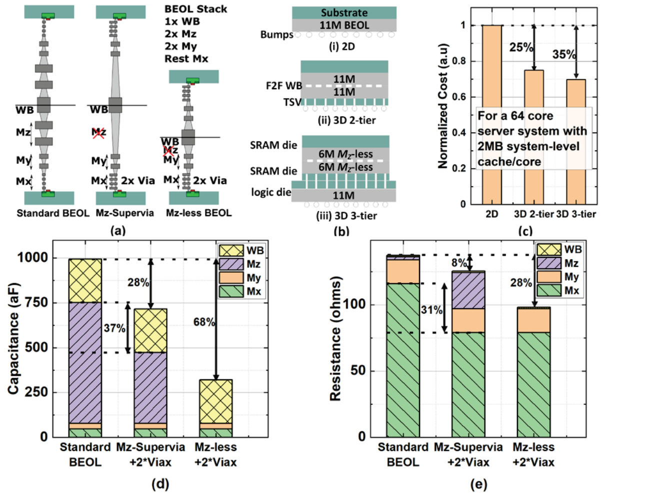


Fig. 3. (a) Proposed DTCO optimizations to 3D-BEOL for 3D-Split SRAMs. (b) Cross-section view (c) cost-comparison of 2D vs 3D system. Contribution of different metal layers to (d) capacitance and (e) resistance of the 3D-BEOL at 12nm.

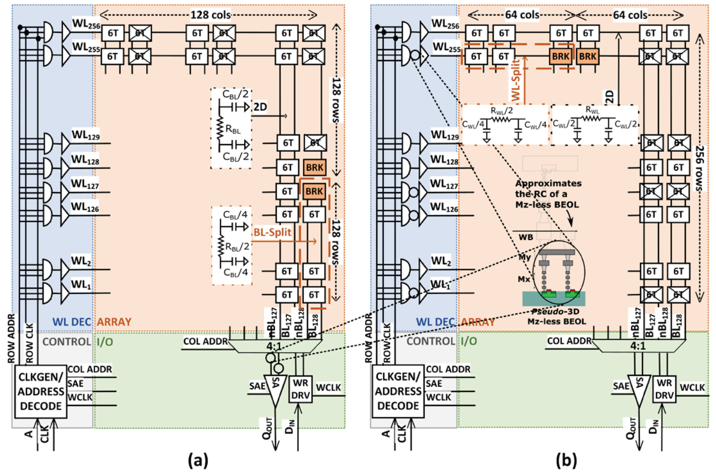


Fig. 4. Schematics of (a) *Pseudo*-3D BL-Split SRAM, (b) *Pseudo*-3D WL-Split SRAM. Both SRAMs require one WB connection per column & per row, respectively. Pitch limitations can be alleviated, to some extent, by staggering the locations of WB connections.

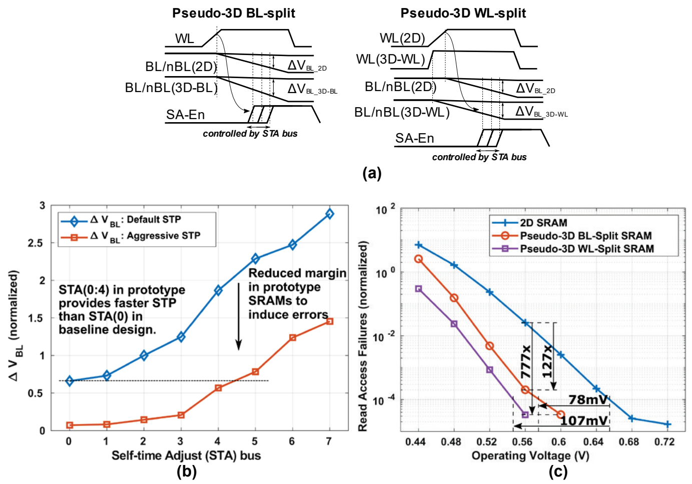


Fig. 5. (a) Read margin or bitline differential (ΔV_{BL}) (b) Simulated ΔV_{BL} for various STA settings. The self-time path (STP) was redesigned to reduce ΔV_{BL}. (c) Measured read errors (normalized) for *pseudo*-3D-Split and baseline 2D SRAMs.

Chip Details		SRAM Config		Words	Bits	Rows/BL (rpb)	Cols/WL (cpw)
Technology	12nm FinFET	Baseline 2D SRAM		1024	64	256	128
Bitcell Type	1-2-2 6T	3D WL-Split integrated macro	2D	512	64	256	128
Capacity	64kb		Pseudo-3D	512	31	256	60
	1k words x 64 bits	3D BL-Split integrated macro	2D	512	64	256	128
Area	~7500μm² each		Pseudo-3D	248	64	127	128

(a)

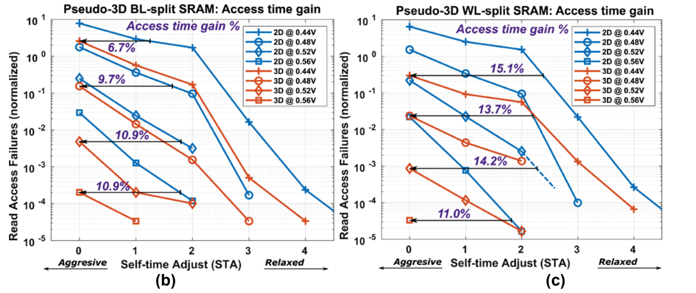


Fig. 6. (a) Details of chip and SRAMs on-chip. (b) Measured read errors (normalized) vs. Self-time adjust (STA) for *pseudo*-3D BL-split SRAM and (c) *pseudo*-3D WL-split SRAMs. Gain in access time estimated by STA setting shift for iso-errors.