2019 32nd International Conference on VLSI Design and 2019 18th International Conference on Embedded Systems (VLSID)

Area efficient & High Performance Word line Segmented architecture in 7nm FinFET SRAM compilers

[1]Vinay kumar , [1]Neeraj kapoor , [1]Sudhir kumar , [1]Monila Juneja , [1]Amit Khanuja

[1]Synopsys India Pvt. Ltd. *vikumar@synopsys.com*

Abstract - **In 7nm TSMC FinFET nodes, metal resistance plays a critical role in achieving the best performance in SRAM design. Random variations and interconnect RC delay is increased due to the continual scaling of physical dimensions, which seriously degrades SRAM performance. In 7nm, it is being observed that Word line (WL) and Bit line resistance limits SRAMs to achieve the speed scaling that technology offers at SoC and Standard Cell library design level. Word line resistance plays a vital role in achieving better access time and operating frequency specifications.**

This paper proposes a High-Performance Word line segmented architecture that segments a wider memory array with minimum area impact. Based on the simulation results for a 7nm High Density, Single Port SRAM compiler, it was observed that the proposed architecture improved access time by 20% and operating frequency by 15%. Improved Word line RC also improves performance at high voltages when Read Assist (Word line underdrive) is enabled.

Keywords — 16nm, 7nm FinFET, SRAM, write margin, bitline differential, wire resistance, Metal strapping, Flip time

I. INTRODUCTION

The growth of battery-powered mobile and wearable devices has increased the importance of low power operation and cost in system-on-a-chip (SoC) design. The continuous scaling of CMOS technology has taken embedded SRAM into the nanoscale regime whereas performance of Memory IP limits by interconnect RC delay instead of device delay. In advanced technology nodes, there is constant stress on memory IP should follow standard cell scaling in both area and performance. In memory compiler design, the array size can be up to 256x640, i.e., 256 memory cells in bitline direction, defined as physical rows and 640 bitcell in word line (WL) direction, defined as physical columns. In SRAMs, word line is connected to the pass gate of the bitcell and routed over a wide bitcell array. This causes high wire RC delay for wider memories. Resistance of lower metal layers has the worst impact on both signal delay and integrity.

To compensate higher RC delay of word line in wider memory, it is routed in higher metal track (usually MX+2) in parallel with foundry provided lower metal track (referred as MX), is known as word line strapping scheme [1], [2].

Table 1 shows effective Word line wire RC across different FinFET nodes.

Normalized SRAM RC comparison - 16nm vs. 7nm FinFET

Node	WL Cap	WL Resistance	BL Cap	BL resistance	WL RC Delay
16nm	1.14*CWL	0.7*RWL	1.14*CBL	0.7*RBL	0.8*CWL*RWL
7nm	CWL	RWL	1.14*CBL	0.7*RBL	CWL*RWL

** RC delay increased by 20% but 7nm must be better in performance by ~20%

Table 1 WL RC delay data for 16nm and 7nm FinFET nodes

With shrinking generation of FinFET process technology, strapping word line in higher metal layers becoming ineffective beyond a certain width of memory. To mitigate or compensate this increased RC delay, buffers has been inserted on regular interval to split the array into smaller segments, is known as rebuffer scheme. Rebuffer scheme improves performance at the cost of area efficiency [1], [2]. Generally, wider memories with physical columns more than 256 columns, array is divided into two segments and this approach works fine till physical columns equals 512. For physical columns greater than 512, memory array is divided into three segments. Array efficiency is becoming worst at break points such as when physical columns > 256 (when array is divided into two segments) and physical columns > 512 (when array is divided into three segments).

In this paper, we propose novel word line segmentation topology for wider memory and compared with existing schemes, i.e., word line-strapping and rebuffer schemes. The new proposal has been implemented and simulated for a Single Port High-Density SRAM compiler in 7nm FinFET technology. Proposed design improves memory operating frequency up to 15% and access time up to 20% without any area impact versus conventional segmentation schemes.

This paper is organized as follows. In Section II, overview of RC delay impact in FinFET is presented. Section III describes the proposed word line segmentation concept and discuss the limitations of existing schemes to improve the performance. Next, we present the simulation results supporting our claim on 7-nm FinFET technology in Section IV and V. Finally, in Section VI, we summarized the results and highlight the conclusions.

II. Interconnect impact in FinFET

In Bulk CMOS technology nodes, the performance of logic circuit determined by driving strength of the global signal driver, i.e., word line and bit line driver in SRAM. The resistivity of copper wires increases rapidly at small dimensions due to increased electron scattering at the grain boundaries and surfaces [3]-[5]. This adversely impacts the scaling of the resistance, and the resultant increased delay of wires which prevents designers from fully exploiting the improvement in intrinsic device performance. As the interconnect dimensions shrink toward the 7nm technology node, metal lines resistance limits the gain offered by technology especially for wider memories as word line level at far end deteriorated and not attain the required level [3]-[5]. Figure 1 shows the minimum WL level required for Six Sigma qualification across a specified voltage range. Besides the WL level, minimum WL pulse must be met to confirm the flip of internal nodes of worst write bitcell in 100Millions Monte-Carlo write time analysis. We determined the pass-gate and pull down transistor internal parameters of bitcell that emulated the electrical behaviour

2380-6923/19/$31.00 ©2019 IEEE
DOI 10.1109/VLSID.2019.00093

of worst write bitcell in 100M MC analysis at SF/SS process corner at low temperature. In simulation, extracted netlist with worst write-time parameters to ensure flip time criterion. For design verification, this paper discusses the simulation results on this 5.6 Sigma worst bitcell.

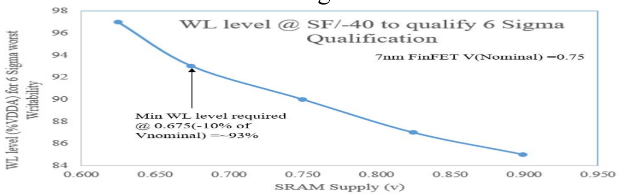


Figure 1 : WL level requirement for 6 Sigma Qualification

Fig. 2 and 3 depicts far and near end waveform of WL for existing word line-strapping and rebuffer schemes. The WL level at far end limits by RC delay and hence cannot be recover with increase in driver strength. Memory with physical column width of 512 having slope of 600ps if drives with single driver, makes design excessively slow. The slope of WL reduced to 250ps by limiting the WL RC delay into two segments using rebuffer schemes (include strapping also) as shown in Fig. 3.

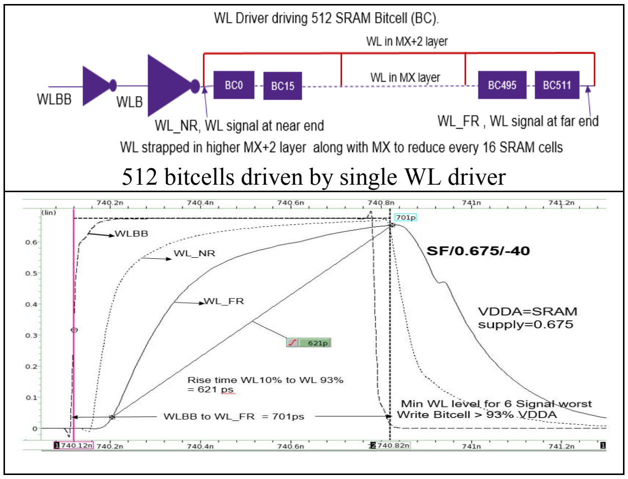


Figure 2 : WL strapping scheme for Columns=512

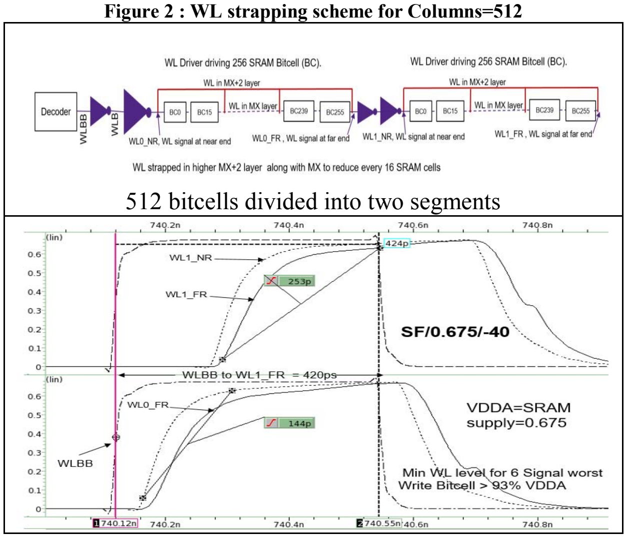


Figure 3 : WL strapping and rebuffer scheme for Columns=512

Physical columns	Re-buffer	Effective Segment	Area Efficiency Impact	Delay (WLBB to Worst WL RC) ps
32<Columns<=256	0	1	No Impact	286
256<Columns<=512	0	1	No Impact	700
256<Columns<=512	1	2	4%	424
512<Columns<=640	0	1	No Impact	1040
512<Columns<=640	1	2	2.5%	483
512<Columns<=640	2	3	5%	542

*** For worst RC delay point, WL level =~93% VDD

Table 2 : Impact of rebuffer on Performance & Area at SF/-40/0.675

Table 2 shows trade-off between area and performance depending of the rebuffer insertion point. Every inserted rebuffer improves the RC delay but effective performance gain reduced by two inverter delay and degrade the area efficiency by 2.5% (approx.). In SRAM compiler design, memory array usually splits after every 256 columns for wider memory to mitigate RC delay.

III Proposed Word line Segmented Schemes

The proposed scheme has been analyzed for different physical column ranges as per the usage. The limitation of existing word line-strapping and rebuffer schemes has been countered with the proposed scheme for High density SRAM without any area impact.

a) Columns <= 256

Fig. 4 depicts an implementation of proposed scheme for columns <= 256 for improving the performance by driving WL from both edges of array. To drive the WL from far end requires to add WL driver in edge cell which increased the overall area within 0.5%. This implementation is mainly focused on high speed memory where memory with column <=256 widely used on the SOC.

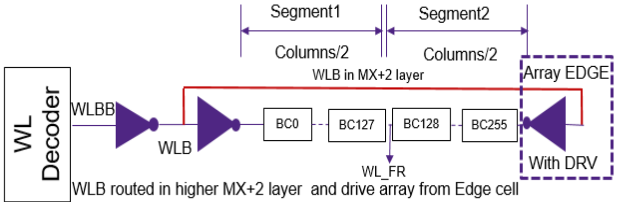


Figure 4 : Edge DRV scheme for Columns <= 256

b) Columns <= 512

Fig. 5 and 6 depicts an implementation of proposed scheme for columns <= 512 by segmenting the array (i) into three sub-arrays for better performance with similar area for high speed applications and (ii) into two sub-arrays for better area with similar performance for ultra-high density applications.

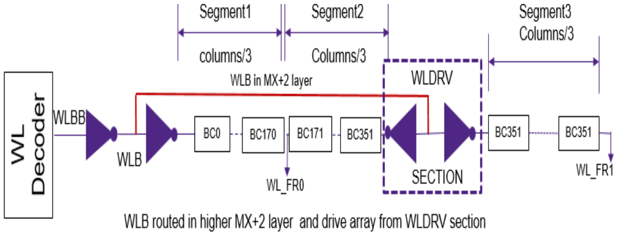


Figure 5 : High Speed WL segmentation scheme for Columns <=512

High speed can be achieved in highly dense SRAM design as shown in Fig. 5, where WLB (inverted signal of word line) signal is routed in MX+2 layer till two segments and connected with driver from both ends as well as to driver of third segment. For e.g. 512 columns SRAM, the WLB signal is routed till 352 (=2/3rd) Columns and effective RC delay of single segment will be 176 columns. In this implementation, memory array is segmented into three segments without any buffer delays because WLB RC delay is much smaller to WL RC delay.

Table 3 demonstrates the RC delay of WLB signal is about 15% of WL signal. The worst delay for WLB signal at SF/0.675/-40 is less than inverter delay without penalizing area for columns < 512. This scheme can be enhanced further for wider columns as per the requirement.

Signal	Layer	Cap(fF)	Resistance	Delay RC
WLB	MX+2	0.18*C	0.85*R	0.15*R*C
WL	MX	C	R	R*C
WLB Signal is almost 7times Faster than WL				

Table 3 : RC comparison of WL & WLB Signal

The best area can be achieved for Ultra high density SRAM design by implementing the scheme explained in section III.a scheme for columns <= 512 as shown in Fig. 6.

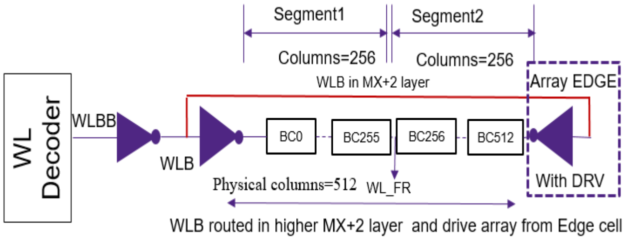


Figure 6: Area Efficient WL segmentation scheme for Columns <=512

c) Columns > 512

Fig. 7 and 8 depicts an implementation of proposed scheme for columns > 512 by segmenting the array (i) into five sub-arrays for better performance with same area for high speed applications and (ii) into three sub-arrays for area efficient design with similar performance for ultra-high density applications.

For wider SRAM with columns > 512 requires to rebuffer the WL twice to mitigate the RC delay that decrease the area efficiency by 5% approximately. High speed can be attained by splitting the array into five sub-arrays as shown in Fig. 7.

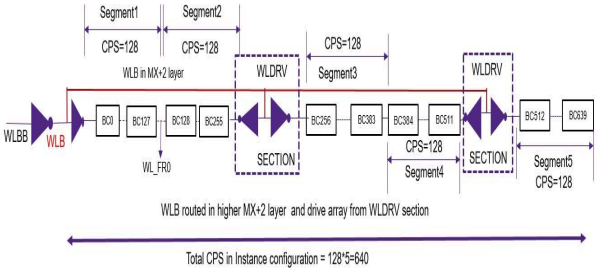


Figure 7 : Proposed WL scheme with multiple WLDRV for Columns > 512 (e.g. 640)

For e.g. 640 columns SRAM divided into five equal sub-arrays of 128 (=1/5th) columns each with having two WLDRV sections (WL driver) and each WLDRV is having two drivers for left and right sub-array. The global WLB signal routed over 4 sub-arrays and connected to each WL driver. With this scheme, single WL driver is having RC delay of 128 columns which offers substantial gain in performance as compared to two rebuffer sections without any area impact.

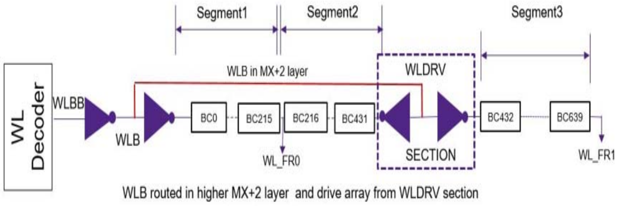


Figure 8 : Proposed WL scheme with single WLDRV for Columns > 512 (e.g. 640)

The increase of area in rebuffer scheme can be recovered through proposed implementation without any loss in performance for Ultra high density SRAM design by using III.b.(ii) scheme for columns > 512 as shown in Fig. 8. With this scheme, 640 columns SRAM divided into three equal sub-arrays of 216 (=1/3rd) columns like rebuffer scheme with two buffer sections but with better performance and area efficiency.

d) Self-time tracking for proposed Scheme

Fig. 9 presents the self-time tracking implementation for proposed different WL schemes as discussed in earlier sections. A reference word line (RWL) uses for tracking the horizontal RC of main WL and turns off the self-time circuit after covering the vertical tracking of bit line. The RWL imitates and tracks main WL rising and reset internal clock at optimum point through logic delays. To ensure that performance must not be limited by periphery global signal RC delay, replica of WL scheme is implemented in all global periphery signals as shown in Figure 9.

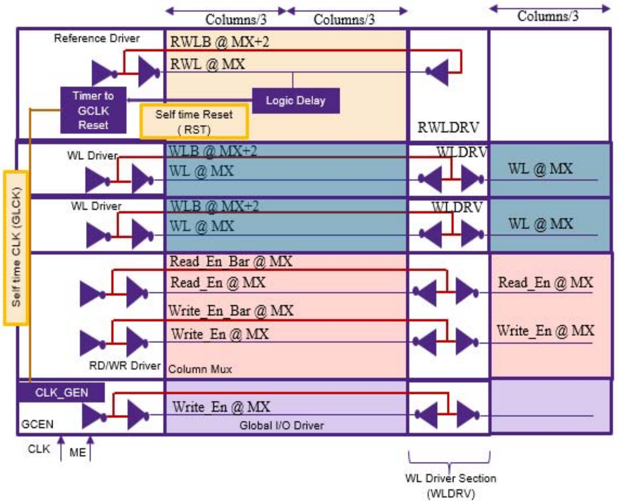


Figure 9 : Floor plan for Proposed scheme

Simlar to actual WL implementation, reference RWL is also generated to closely track WL RC. RWLB signals mimic the WLB signal of SRAM array and its delay varies with phyiscal columns and rows. When memory is enabled with ME=1, selftime clock GCLK is generated that turns on RWL and WL at the same time, RWL resets this selftime CLK at the end of Read/Write operation. The waterfall diagram for write operation is shown in Figure 10.

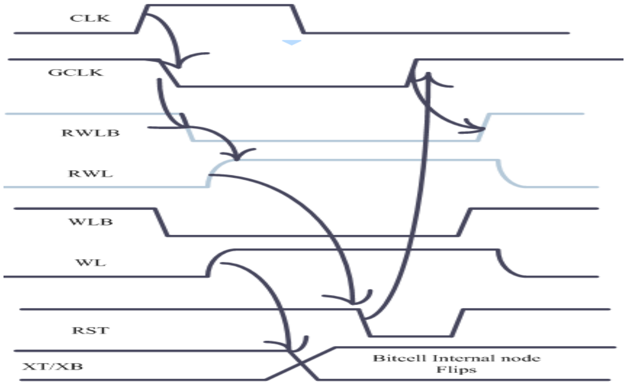


Figure 10 : Signal flow diagram for write operation

IV Simulation Results for Write Operation

Memory instances for different Columns configurations have been simulated with instance level RC extraction and the proposed scheme has been compared with existing word line-strapping and rebuffer schemes.

a) Columns <= 256

For this range of columns, word line strapping is optimum solution from the existing scheme as rebuffer introduce delay of two inverters with area overhead. The proposed scheme merges WLDRV with SRAM edge cell to gain speed with minimal impact on area as shown in Fig. 11.

Fig. 12 presents the simulated waveforms comparing word line strapping and proposed scheme.

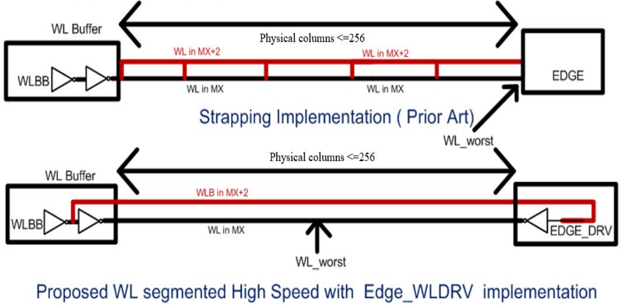


Figure 11 : Schematic representation of WL strapping vs. Proposed scheme for Columns<=256

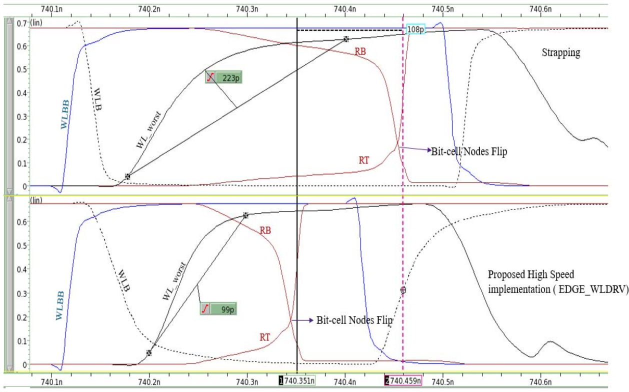


Figure 12 : Waveform comparison WL strapping vs. Proposed scheme for Columns<=256

Table 4 compares the performance and area metrics of word line strapping versus proposed scheme for Columns <= 256.

Performance & Area impact Comparison (Prior vs Proposed) @ Columns<=256			
Design Implementation	WLBB to Bit cell RT/RB Flip Time (ps)	WL Slope (ps)	Area Impact (%)
Strapping	335	220	No impact
Proposed EDGE_DRV (High speed)	224	99	0.5%

Table 4 : Comparison analysis for Columns <=256 @ SF/-40/0.675v

b) Columns <= 512

For this range of columns, two cases have been analyzed as discussed in section III.b. As shown in Fig. 13, the array is segmented (i) into three sub-arrays for better performance with similar area for high speed applications and (ii) into two sub-arrays for better area with similar performance for ultra-high density applications.

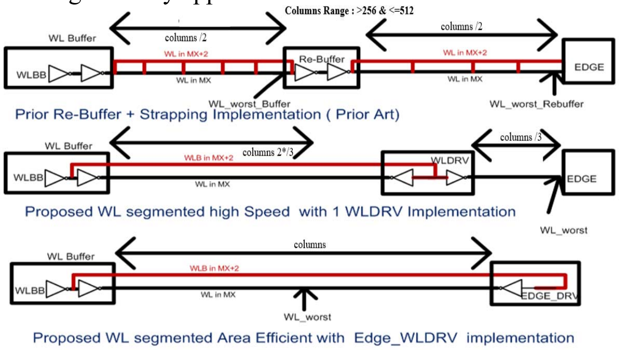
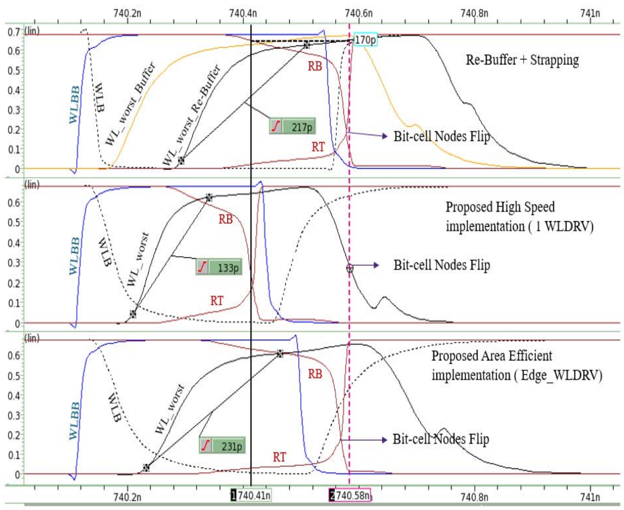


Figure 13: Schematic representation of Prior schemes vs. Proposed scheme for Columns<=512

Fig. 14 presents the simulated waveforms comparing word line strapping, rebuffer and proposed scheme.

The proposed scheme provides designers the flexibility to either choose better performance with no added area loss or to improve SRAM area efficiency without penalizing performance. Results have been summarized in Table 5.



Waveform comparison Existing Schemes vs. Proposed Scheme

Figure 14: Comparison analysis for Columns<=512 @ SF/-40/0.675v

Performance & Area Impact comparison (Prior vs Proposed) @ 256<Columns<=512			
Design Implementation	WLBB to Bitcell RT/RB Flip time (ps)	WL Slope(ps)	Area Impact (%)
Rebuffer + Strapping	464	217	4%
Proposed Edge_WLDRV (Area Efficient)	445	230	0.5%
Proposed WLDRV (High Speed)	397	132	4%

Table 5 : Performance & Area impact comparsion Columns<=512

c) Columns > 512

For wider range of columns, two cases have been analyzed as discussed in section III.c. As shown in Fig. 15, the array is segmented (i) into five sub-arrays for better performance with similar area for high speed applications and (ii) into three sub-arrays for better area with similar performance for ultra-high density applications.

Fig. 16 compares the simulated waveforms between word line strapping, rebuffer and proposed scheme and results have been summarized in Table 6.

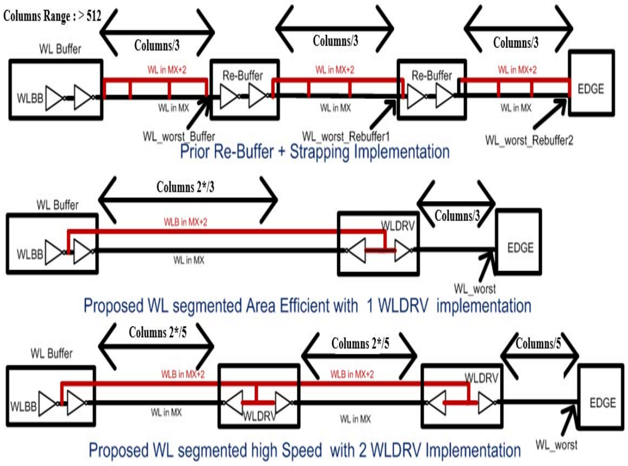
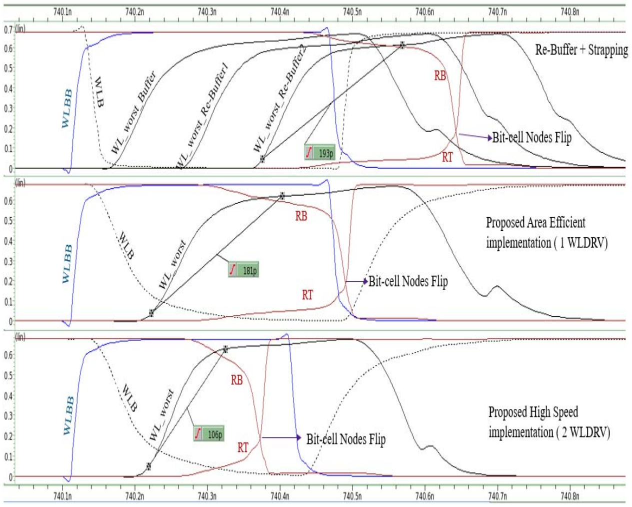


Figure 15: Schematic representation of Prior scheme vs Proposed scheme for Columns>512



Waveform comparison Existing Schemes vs. Proposed Scheme

Figure 16: Comparison analysis for Columns>512 @ SF/-40/0.675v

Performance & Area Impact comparison (Prior vs Proposed) @ 512 <Columns<=640			
Design Implementation	WLBB to Bitcell RT/RB Flip time (ps)	WL Slope(ps)	Area Impact (%)
Re-Buffer + Strapping	513	193	5%
Proposed 1 WLDRV (Area Efficient)	372	181	3%
Proposed 2 WLDRV (High Speed)	252	106	5%

Table 6 : Performance & Area impact comparsion Columns>512

V. Simulation Results for Read Operation

Proposed scheme shows foremost gain in read Metric, i.e., Access time. Access time has two major components, generation of WL when CLK is fired and WL ON duration which decides the amount of differential signal generated at sense-amp internal nodes. The proposed WL segmentation scheme generates WL rise edge faster as compared to existing word line strapping and rebuffer schemes. In rebuffer and word line strapping approach, rebuffer delay comes in sequential manner that limits its gain in access time arc. The proposed scheme escapes the far end word line from rebuffer delays, that gives direct gain to turn on WL at far location. The proposed scheme splits the array into more effective manner that aid worst bitcell read current to generate differential signals at a faster rate, i.e., time delay from WL ON time to sense amplifier enable time is also reduced. Fig. 17 shows the waveform for differential signal generation at sense Amplifier nodes for Columns = 512 with single WLDRV.

Proposed scheme offers improvement in Read performance. SoC designers need the access time to be almost 30% less than memory cycle time. In FinFET, especially for wider memories, this constraint is very difficult to meet on SoC and finally forces the designer to relax the cycle-time to satisfy this constraint. The improvement in Read performance by proposed implementation helps meet the constraint of access time to be within 70-75% of cycle time.

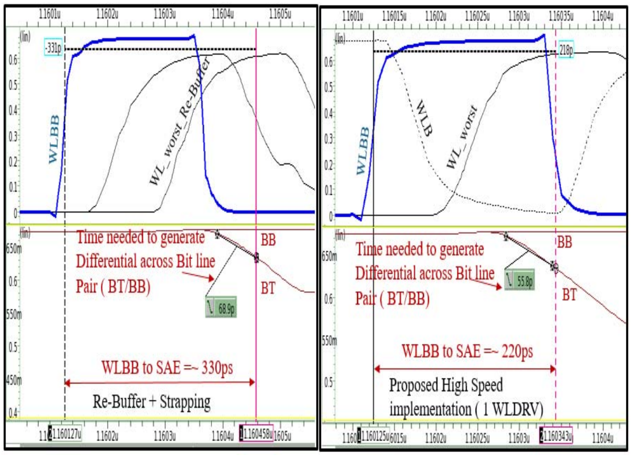


Figure 17: Waveform comparison for Read Window at Columns=512

Read performance improvement for entire Columns range has been summarized in Table 7.

Comparsion Analysis for Columns <=256		
Design Implementation	WLBB to Sense Enable (ps)	Delay for 80mv Differential Signal on BT/BB(read window) ps
Strapping	304	185
Proposed High speed EDGE_DRV	261	132
Comparsion Analysis for 256<Columns <=512		
Design Implementation	WLBB to Sense Enable (ps)	Delay for 80mv Differential Signal on BT/BB(read window) ps
Re-buffer + Strapping (1 WL Re-Buffer)	425	170
Proposed Area Efficient Edge_WLDRV	365	175
Proposed High Speed 1 WLDRV	287	135
Comparsion Analysis for 512<Columns <=640		
Design Implementation	WLBB to Sense Enable (ps)	Delay for 80mv Differential Signal on BT/BB(read window) ps
Re-buffer + Strapping (2 WL Re-Buffer)	488	156
Proposed Area efficient 1 WLDRV	326	160
Proposed high Speed 2 WLDRV	288	140

Table 7 : Read Window Comparison

VI. Conclusion

Fig. 18 depicts the layout implementation of proposed scheme. The proposed WL segmentation scheme nullifies the interconnect impact. The described solutions allow SRAM designers to improve area efficiency without compromising on performance when compared with existing Rebuffered and strapping solution.

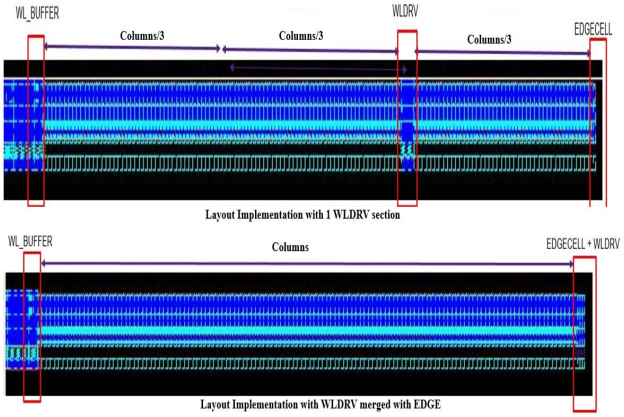
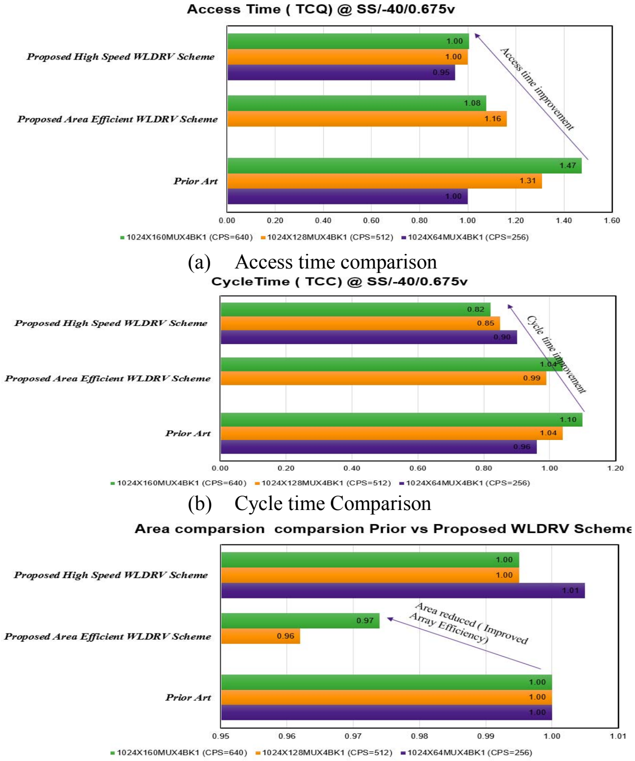

WL_BUFFER
WLDRV
EDGECELL
Columns/3
Columns/3
Columns/3
Layout Implementation with 1 WLDRV section
WL_BUFFER
Columns
EDGECELL + WLDRV
Layout Implementation with WLDRV merged with EDGE

Figure 18: Layout Implementation

The proposed implementation has overall better PPA (power, performance & area). It offers designers the opportunity to make performance almost Columns independent within same area as compare to prior schemes. The proposed scheme was evaluated for a Single Port High-Density SRAM compiler for 7nm FinFET technology. Fig. 19 shows improvement in performance up to 25% in cycle time and ~40% in access time in widest configuration. All the data in this paper is generated with instance level simulation using 7nm silicon extracted mature model with varying physical columns keeping rest of the memory parameters constant.



(a) Access time comparison

(b) Cycle time Comparison

(c) Area Comparison Analysis

Figure 19: Area/Performance comparison at Instance level

REFERENCES

[1] V. Kumar, N. Puri, S. Kumar, and S. Srivastav, “A sub-0.5 v reliability aware-negative bitline write-assisted 8t dp-sram and wl strapping novelarchitecture to counter dual patterning issues in 10nm finfet,” in VLSIDesign and 2017 16th International Conference on Embedded Systems(VLSID), 2017 30th International Conference on. IEEE, 2017, pp. 269–274.

[2] Vivek Nautiyal; Gaurav Singla; Sagar Dwivedi; Satinderjit Singh; Ingming Chang; Jitendra Dasani; Fakhruddin Ali Bohra “Self-Timed Shaper Circuit for Wide Memories in Advanced CMOS Technologies“, ISCAS 2018.

[3] M. T. Bohr, “Interconnect scaling-the real limiter to high performance ulsi,” in Electron Devices Meeting, 1995. IEDM’95., International. IEEE, 1995, pp. 241–244.

[4] D. Shah, K. Siva, G. Girishankar, and N. Nagaraj, “Optimizing interconnect for performance in standard cell library,” in Circuits and Systems, 2006. APCCAS 2006. IEEE Asia Pacific Conference on. IEEE, 2006,pp. 1280–1284.

[5] A. Ceyhan, M. Jung, S. Panth, S. K. Lim, and A. Naeemi, “Evaluating chip-level impact of cu/low- kappa performance degradation on circuit performance at future technology nodes,” IEEE Transactions on Electron Devices, vol. 62, no. 3, pp. 940–946, 2015.