Analysis of Word Line Shaping Techniques for In-Memory Computing in SRAMs

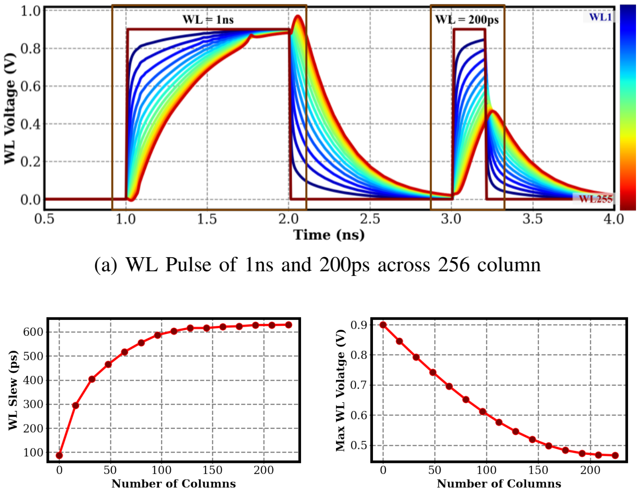
Kailash Prasad, Aditya Biswas, Joycee Mekie
Department of Electrical Engineering, Indian Institute of Technology Gandhinagar
Email: {kailash.prasad, adityab, joycee}@iitgn.ac.in

***Abstract*—In-memory computing (IMC) architectures have emerged as a promising alternative to deal with data-intensive applications. Proposals based on analog or digital IMC require multiple word lines to be activated for performing computation. Specifically, in wide SRAM IMC architectures, the word line pulse shaper circuits need to be carefully investigated as pulse-width degradation affects multiple rows, resulting in incorrect output, loss in linearity in results, or degraded performance. This paper implements compares and contrasts multiple word line shaper proposals for a wide SRAM array. Detailed post-layout simulation results of 512x256 array show that for 1-bit analog dot product, the standard deviation is improved by 0.19×, 0.21×, 0.21× 0.15× for 1,4,8, and 16bit word respectively. Further, word line shaping techniques improve the access time and compute delay by 2.86× and 3×, respectively for 128x256 array.**

I. Introduction

In-Memory Computing (IMC) is one of the promising proposals for removing the Von-Neumann bottleneck faced by data-intensive applications. IMC is targeted for applications like neural networks, machine learning, image, and video processing to achieve high computational performance with energy efficiency. SRAM-based IMC is popular because it is fast, compact, and convenient to augment with processing logic [1]–[4]. In this paper we investigate the impact of word line shaper circuits on wide memory IMC architectures where multiple word lines are simultaneously activated.

In advanced technology, the SRAM array performance is limited not only by transistors but also by interconnects. Interconnects contribute significantly to delay due to high resistances and capacitances offered by long metal lines. In SRAMs, the long bit lines and word lines adversely affect SRAM performance. To reduce the bit line delay, new sense amplifiers have been proposed, and innovative techniques like bit line boost [5] and flying bit line architecture [6] have been implemented. Along with the bit lines, word lines are also long rails running orthogonal to bit lines. Various studies have been carried out to optimize the energy of the system by adopting innovative pulse width modulation techniques on word lines. For instance, a replica bit line circuit is designed to control the word line pulse to reduce read energy in a traditional SRAM array. The inherent parasitic capacitance of access NMOS transistor and resistance and capacitance of metal interconnects and vias are the reason for WL pulse degradation. The degradation increases as the number of columns increases in the wide memories. Fig. 1(a) shows the word line pulse of 1ns and 200ps in a 128x256 memory array across columns. We observe that there is a large WL slew at the farthest column as shown in Fig. 1(b). The shorter pulse width is degraded worse as it is not able to reach the peak voltage. Fig. 1(c) shows the WL peak voltage degradation across columns.



(a) WL Pulse of 1ns and 200ps across 256 column

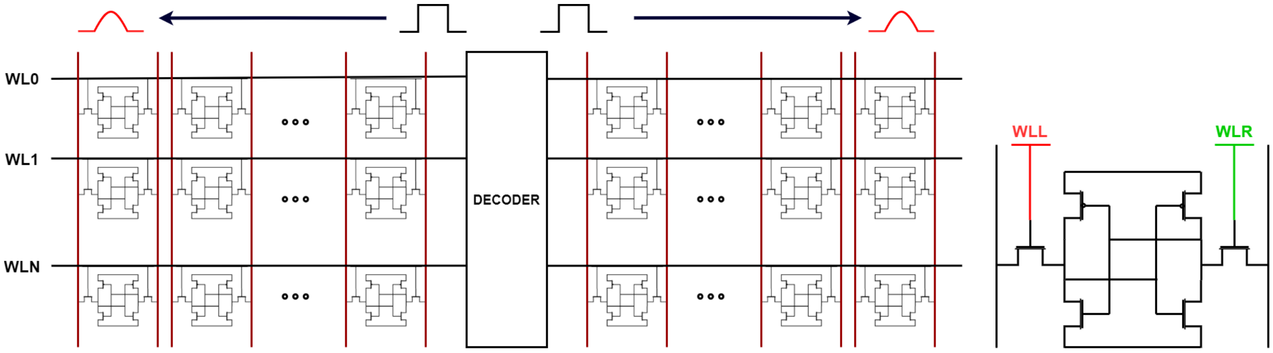
(b) WL Slew for 1ns pulse

(c) Max WL Voltage for 200ps pulse

Fig. 1: WL Pulse Degradation

a) Need for Wave shaper circuits in IMC architectures: IMC architectures - both analog and digital IMCs - generally employ wide memories to maximize throughput as wide memories achieve high parallelism. However, the long word line (WL) interconnects in wide memories causes WL pulse degradation. In addition to this, IMC architectures targeted to achieve high performance often require multiple word lines to be activated simultaneously.

IMC architectures based on digital computation generally require two word lines to be activated simultaneously. In digital IMC architecture, several techniques such as word line under-drive [1], short word line [5], split word line [7] technique, etc., are used to mitigate the problem of read disturb. However, as is clear from Fig. 1(a), short word line [5] technique cannot be used directly due to the pulse-width degradation observed in the far columns. This poses a contradicting requirement on WL pulse-width, which cannot be too large to avoid read disturb and cannot be too short to avoid pulse-width degradation. Thus, WL shaping circuits are

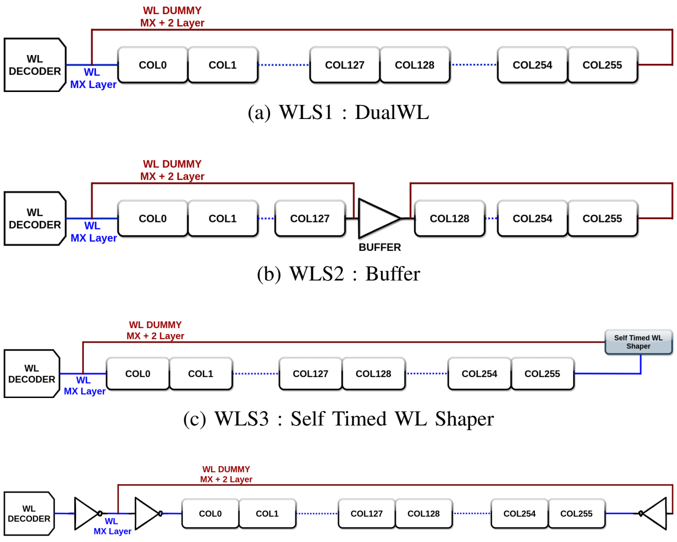


(a) Central Spine Architecture

(b) Split word line

Fig. 2: Techniques to remove word line Pulse degradation

imperative for digital IMC and need to be carefully designed. In contrast, analog IMC architectures require multiple (even up to 256) word lines [3] to be activated. Further, pulse-width modulated word lines are required in analog IMC architectures to achieve the analog dot product. Thus, WL shaping circuits are required to ensure (a) that minimum pulse-width is reached the far column without any degradation, (b) the pulses on multiple word lines are aligned. While condition (a) can be met using word line shaping circuit, condition (b) depends on process variations and may result in performance degradation even with the use of appropriate WL shaping circuits. Multi-bit multiplication is implemented by providing pulse in the word line and storing the data in columns [3].



(a) WLS1 : DualWL

(b) WLS2 : Buffer

(c) WLS3 : Self Timed WL Shaper

(d) WLS4 : Edge Driving

Fig. 3: WL Shaping Techniques

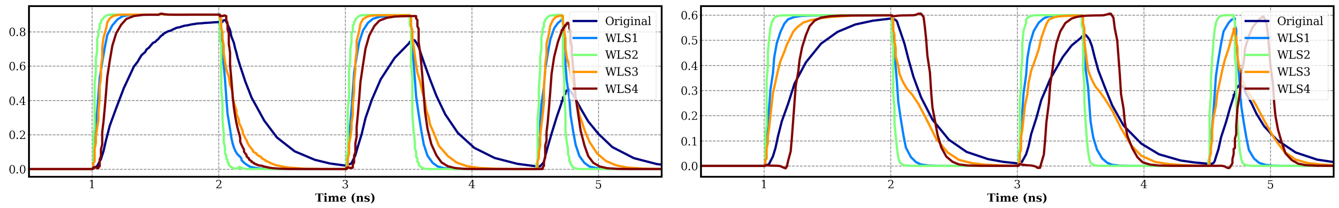
To summarize, long word lines in wide SRAM IMC architectures impact the access time and compute delay in digital IMCs. In analog IMC, WL pulse degradation leads to inconsistent bit line discharge and reduces the computational accuracy. This problem becomes severe when multiple word lines are activated. Several architectural changes and WL shaping techniques are proposed in the literature to minimize the pulse width degradation in conventional SRAM memory subsystems. However, none of these techniques have been comprehensively applied and tested for IMC architectures. This paper first shows through detailed simulations and analyzes the need for WL shaping circuits for IMC architectures and compares existing techniques comprehensively. We show that among the existing proposals, the buffer [8] technique is the most effective.

The contributions of our work are:

- We have implemented four different WL shaping techniques on 128x256 SRAM array and carried out post-layout simulations. Our comparison shows that with WL shaping, the access time improves by 2.86×.
- We have used these WL shaping techniques for both analog and digital IMC architectures. We observe 3× improvement in compute delay for digital IMC.
- Further for analog dot product on 512x256 array, we observe non-linearity in bitline discharge across the columns. The inconsistency increases with an increase in the number of activated wordlines. With WL shaping technique we observe improvement in standard deviation by 0.19×, 0.21×, 0.21×, 0.15× for 1,4,8, and 16bit word respectively.

II. Techniques to improve word line pulse width

WL pulse degradation has been identified as a concern for wide memories, and several methods have been proposed to mitigate this issue. One of the simplest solution is to use central spine architecture as shown in Fig. 2(a), where the decoding logic is placed at the center of the memory which reduces the word line length to half, thus reducing the delay by one-fourth of the original delay. Another technique is to use an SRAM cell with split WL structure [9] to reduce the delay from the first column to the last column. In this case, only one access transistor is connected to WL as shown in 2(b) to reduce the parasitic capacitance by half, which in turn reduces the overall delay. However, this method needs two identical word lines run per row. Wide memories have the word line degradation, and the WL level at the far end (column) is reduced significantly due to the RC delay and cannot even be recovered by the increased driving strength of the WL



(a) WL = 900mV

(b) WL =600mV

Fig. 4: WL Shaping Technique Comparision for different Pulsewidth at 900mV and 600mV

driver. For such wide memories, four different WL shaping techniques to improve word line degradation are proposed in the literature, as discussed in the section to follow.

A. WLS1 : Dual word line [10], [11]

This technique shown in Fig. 3(a) has a dummy WL run parallel to native WL, wherein the dummy WL is strapped to the native WL. A higher metal line is used for dummy WL, which offers an overall lower resistance and makes the dummy WL faster. Thus, the farthest end receives the proper WL signal. In this method, the worst column is reported to be at $\frac{3}{4}^{th}$ array size. To improve the WL, the dummy WL can be strapped in the middle of the array also.

B. WLS2 : Buffer [8]

The dual metal word lines method does not resolve the WL pulse degradation problem. In the WL buffer technique, a buffer is inserted in the middle of the array to regenerate the WL, as shown in Fig. 3(b). The dual word line is strapped on both sides of the buffer to improve the WL further. For wider memories, multiple such buffers can be included.

C. WLS3 : Self Timed WL Shaper [12]

Insertion of a buffer in the middle of an array increases the overall array area. Also, overall improvement will be reduced by intrinsic buffer delay. Therefore, in the self-timed WL shaper shown in Fig. 3(c), the authors use the dual metals technique with a self-timed supply clamped circuit at the last column. The idea is to improve the WL pulse of the last column by providing supply voltage through the transistor. However, the main concern with this technique is that the fall delay of the pulse is large.

D. WLS4 : Edge Driving [13]

In this technique, the performance is improved by driving WL from both the ends of the array as shown in Fig. 3(d). Driving the WL from the far end requires adding a WL driver in the edge cell. The word line is improved for wider arrays by segmenting the array into three sub-arrays for better performance with similar areas for high-speed applications. Also, it is divided into two sub-arrays for a better area with similar performance for ultra-high density applications.

We have carried out post-layout simulations for all the 4 techniques. The dual word line technique has no additional area overhead. For every 128 columns, the additional area corresponding to 8 columns of 2.5%, needs to be added for introducing buffer at mid. Driving the WL from the far end requires adding a WL driver in the edge cell, increasing the overall area by 0.5%. The self-timed WL Shaper implementation has 0.4% increase in area.

Technique	Area Overhead
WLS1	No overhead
WLS2	2.5%
WLS3	0.5%
WLS4	0.4%

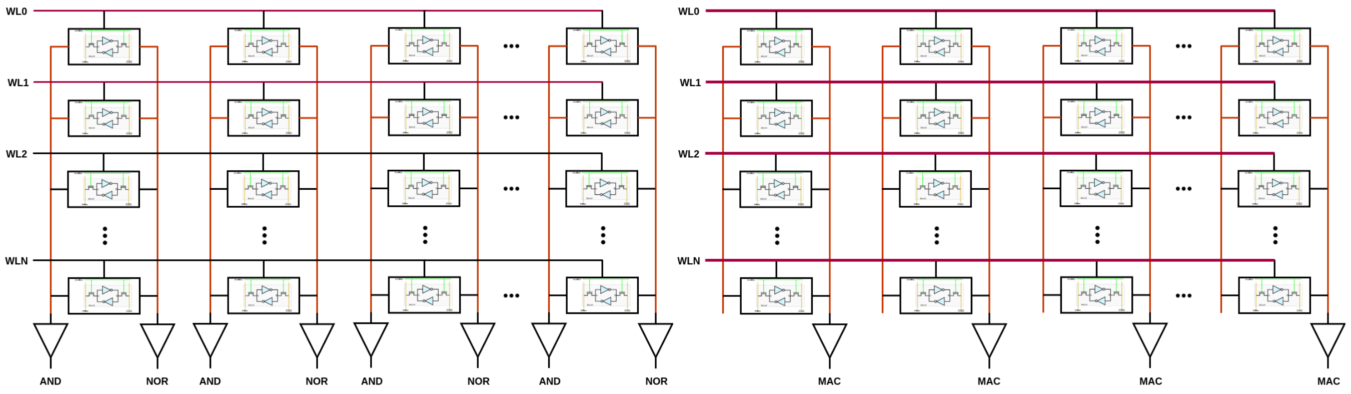
TABLE I: Area Overhead of WL Shaping Techniques

III. Analysis of WL Shaping Techniques - WL = 256 Columns

This section analyzes the dependence of maximum frequency and V_{MIN} of SRAM on word line pulses at the worst column. We have evaluated all the techniques in the 128x256 array in CMOS 28nm technology with Post Layout simulations at the worst cases SS corner, -40°C. The ratio of RC delay of MX and MX+2 layer used in all the designs is shown in Table II. We have analyzed all the techniques and derived WL Slew (time taken for a signal to reach 10% to 90% for rising and 90% to 10% for fall) for different pulse widths at the worst column. We have plotted 1ns, 500ps, and 200ps pulse at the worst column for all the techniques as shown in Fig. 4. Table III shows the WL slew of each technique and their improvement. The best technique at the worst-case column is WLS2: Buffer Technique with the lowest WL Slew for both rise and fall, and WLS3: Self Timed WL Shaper is the worst as its fall delay is very high. We have also analyzed for WL = 600mV, and all the techniques can regenerate WL pulse even at lower voltages. Table III also shows the access time comparison at WL = 900mV and 700mV. The access time comprises of bit line discharge time for 100mV discharge after the word line is activated and the sense amplifier delay. The WLS2 : Buffer techniques show

Layer	Delay
MX	RC
MX+2	0.19*RC

TABLE II: Metal Delay in CMOS 28nm



(a) Digital IMC

(b) Analog IMC

Fig. 5: In Memory Computing architectures

Technique	WL Slew Rise (ps)	WL Slew Fall (ps)	Access Time WL=900mV (ps)	Access Time WL=700mV(ps)
Original	531	557	587	827
WLS1	112	112	205	571
WLS2	62	62	176	455
WLS3	81	273	195	540
WLS4	96	101	184	481

TABLE III: WL Slew Improvement By Techniques

2.86× and 1.45× improvement in access time at 900mV and 700mV.

Technique	Compute Delay (in ps) (WL 900mV)	Compute Delay (in ps) (WL 700mV)
Original	593	953
WLS1	230	541
WLS2	190	460
WLS3	220	567
WLS4	199	517

TABLE IV: Compute Delay Improvement By Techniques

IV. Analysis of WL Shaping Techniques - IMC

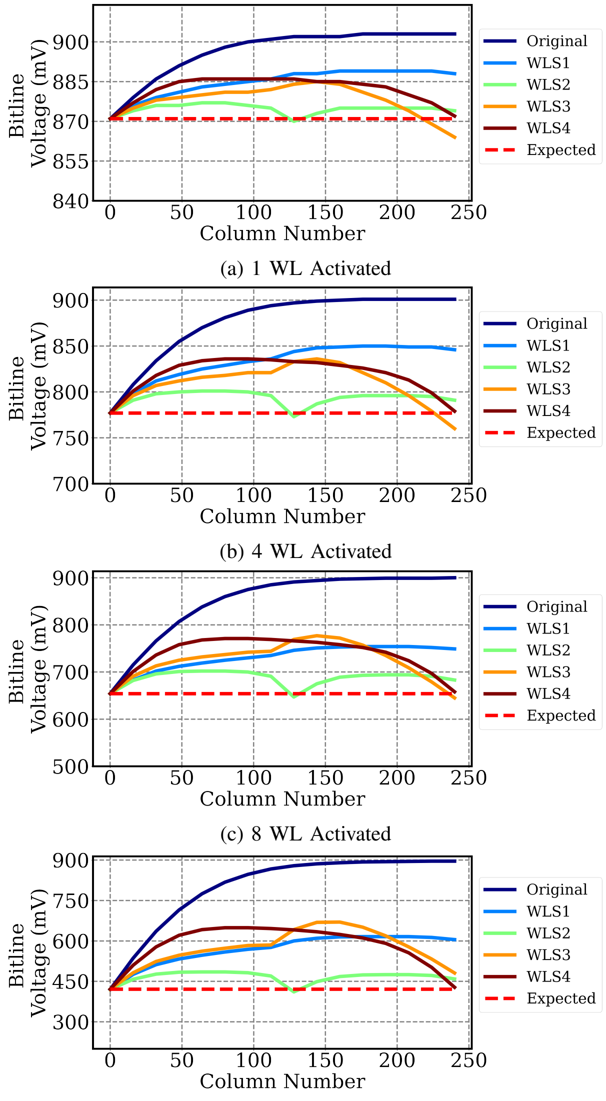
A. *IMC in Digital*

Digital in-memory computing techniques involve turning on two word lines at a given time, and logical computation of both the bits can be obtained in the bit-lines (AND in BL and NOR in BLB), which can be sensed using sense amplifiers as shown in Fig 5(a). IMC in 6T SRAM suffers from the problem of compute disturb, also alternatively referred to as read-disturb issue. To overcome the compute disturb problem, three techniques are prevalent, viz. word line underdrive, short word line pulse, and split word line. All three techniques require proper word line pulse to be generated of low width or low voltages. The efficient generation of the word line will restrict the V_{MIN} and performance of the SRAM IMC in wider memories. We have performed post-layout simulations on a 128x256 array by turning on WL0 and WL1 to perform logical AND/NOR operations. The WL pulse is 350 ps at 900mV and 1ns at 700mV. TableIV shows the improvement in Compute Access time by WL shaping techniques. The WLS2: Buffer Technique shows improvement of 3× at WL=900mV and 2.07× at WL=700mV. The WLS 3: WL shaper techniques perform well at 900mV but become worst at 700mV. The skewed inverter and pass gate used in the design slows down significantly at lower voltages. This analysis shows the improvement in the performance of digital IMC using the WL shaping technique.

B. *IMC in Analog - Inconsistency in bit line Discharge*

Analog IMC involves turning on multiple word lines to achieve high computational efficiency, as shown in Fig 5(b). One of the operands is provided in a pulse train, and the other is stored in the memory array in row-major or column-major fashion. However, the precision of analog computing is limited by the non-linearity of various sources. Many works of literature use short word line pulse to improve linearity [3]. However, generating a short WL pulse itself is the hardest challenge; it may enhance the linearity of the small-sized array. However, for a wider memory, this technique can add non-linearity. It is one of the least explored non-linearity sources. The WL pulse degradation reduces the computational consistency across the columns. Even if the SRAM array stores identical data in each column, the bit line discharge in the closest column to the decoder will differ from the farthest word line as bitline discharge is directly proportional to the wordline pulse width and amplitude. This problem becomes severe when multiple word lines are activated. The difference increases significantly. The problem can be overcome to some extent by adding any of the WL shaping techniques. However, the efficiency of these techniques will vary because of the different architecture.

We have evaluated all the techniques in the 512x256 array in CMOS 28nm technology with Post Layout simulations at the worst cases SS corner, -40°C. The word line pulse used is 200ps, and the bit line voltage is measured after 100ps of WL turning off. Fig. 6(a) shows the variation of bit line voltage across columns when a single word line is activated for all the WL shaping techniques. The bit line discharge decreases due



(a) 1 WL Activated

(b) 4 WL Activated

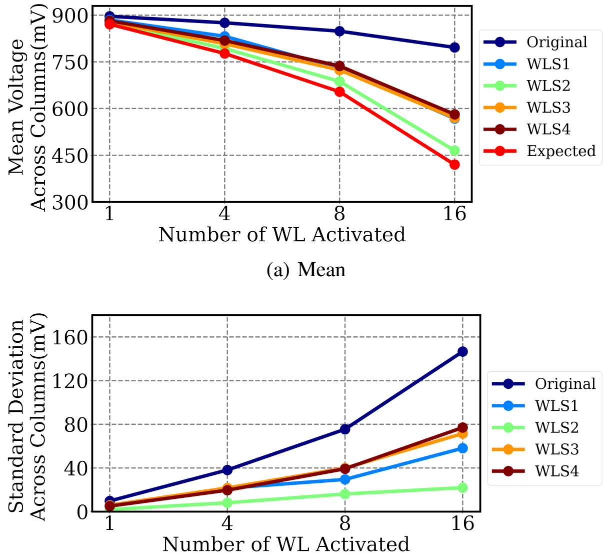
(c) 8 WL Activated

(d) 16 WL Activated

Fig. 6: Bit line discharge voltage for analog dot product when multiple word lines are activated

to degradation in pulsewidth when no WL shaping technique is applied. The WL shaping tries to linearise the bit line discharge, but all behave differently due to their architecture. WLS2: Buffer Technique is the best technique in terms of linearising the discharge, and the WLS1: Dual WL behaves the worst.

1) Multiple word lines: We have performed analog dot product by activating N (4,8, and 16) word lines as shown in Fig. 6(b-d). The bitline discharge here represents the analog dot product of N bits stored across the column. Fig. 7 shows the mean and standard deviation of the bitline discharge voltage for each technique with the same data provided across columns. Standard deviation represents the difference between average bitline discharge and the worst column bit line discharge. The WLS2: Buffer technique consistently performs better with less standard deviation. For 1-bit analog dot product it shows the improvement in standard deviation by 0.19×, 0.21×, 0.21×, 0.15× for 1,4,8, and 16bit word respectively. The WLS4: Self-timed shaper technique becomes inefficient when multiple word lines are activated because there is a direct connection in all the word lines. One of the major observations is that the standard deviation increases with the number of word lines being activated. This signifies that the efficiency of the word line shaping technique decreases with an increase in the number of activated word lines. It becomes more important to use highly efficient WL shaping techniques to minimize nonlinearity in analog IMC.



(a) Mean

(b) Standard Deviation

Fig. 7: Mean and standard deviation bit line discharge voltage when multiple word lines are activated

Acknowledgment

This work is supported through grants received from Prime Minister Reserach Fellowship (PMRF), SMDP-C2SD and YFRF Visvesvaraya Ph.D. scheme from the Ministry of Electronics and Information Technology (MEITY), and through SERB grants CRG/2018/005013, MTR/2019/001605, SPR/2020/000450 and Semiconductor Research Corporation (SRC), through contract 2020-IR-2980.

V. Conclusion

Word line degradation across columns in advanced technology nodes has a significant impact on the performance of the SRAM memory subsystem and is the limiting factor in wide memories. We have analyzed four existing techniques proposed to improve word line pulse in conventional memory for IMC architectures. For conventional memories, among all the four techniques, Buffer technique turns out to be the best in improving the word line for the worst case at the cost of 4% area overhead. We have also analyzed the impact of WL shaping techniques on digital and analog IMC performance. In digital IMC, we show the improvement in compute access time using these techniques. In analog IMC, there is high inconsistency in results across the columns due to word line degradation. Also, the word line shaping techniques improve the inconsistency but are not entirely able to remove it. The inconsistency in results further increases for the N bit dot product where multiple word lines are activated, limiting computation accuracy. In the future, we would like to study the impact of the loss in computational accuracy for error-resilient applications, such as machine learning applications.

References

[1] J. Wang, X. Wang, C. Eckert, A. Subramaniyan, R. Das, D. Blaauw, and D. Sylvester, "A 28-nm compute sram with bit-serial logic/arithmetic operations for programmable in-memory vector computing," *IEEE Journal of Solid-State Circuits*, vol. 55, no. 1, pp. 76–86, 2019.

[2] W. Simon, J. Galicia, A. Levisse, M. Zapater, and D. Atienza, "A fast, reliable and wide-voltage-range in-memory computing architecture," in *2019 56th ACM/IEEE Design Automation Conference (DAC)*, pp. 1–6, IEEE, 2019.

[3] M. Kang, S. K. Gonugondla, and N. R. Shanbhag, "Deep in-memory architectures in sram: An analog approach to approximate computing," *Proceedings of the IEEE*, vol. 108, no. 12, pp. 2251–2275, 2020.

[4] N. Surana, M. Lavania, A. Barma, and J. Mekie, "Robust and high-performance 12-t interlocked sram for in-memory computing," in *2020 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 1323–1326, IEEE, 2020.

[5] K. Lee, J. Jeong, S. Cheon, W. Choi, and J. Park, "Bit parallel 6t sram in-memory computing with reconfigurable bit-precision," in *2020 57th ACM/IEEE Design Automation Conference (DAC)*, pp. 1–6, IEEE, 2020.

[6] H. Fujiwara, C.-Y. Lin, H.-Y. Pan, C.-H. Lin, P.-Y. Huang, K.-C. Lin, J.-J. Liaw, Y.-H. Chen, H.-J. Liao, and J. Chang, "24.2 a 7nm 2.1 ghz dual-port sram with wl-rc optimization and dummy-read-recovery circuitry to mitigate read-disturb-write issue," in *2019 IEEE International Solid-State Circuits Conference-(ISSCC)*, pp. 390–392, IEEE, 2019.

[7] M. Kang, S. K. Gonugondla, A. Patil, and N. R. Shanbhag, "A multi-functional in-memory inference processor using a standard 6t sram array," *IEEE Journal of Solid-State Circuits*, vol. 53, no. 2, pp. 642–655, 2018.

[8] B. Wu and M. R. Guthaus, "Bottom-up approach for high speed sram word-line buffer insertion optimization," in *2019 IFIP/IEEE 27th International Conference on Very Large Scale Integration (VLSI-SoC)*, pp. 305–310, IEEE, 2019.

[9] Z. Lin, H. Zhan, Z. Chen, C. Peng, X. Wu, W. Lu, Q. Zhao, X. Li, and J. Chen, "Cascade current mirror to improve linearity and consistency in sram in-memory computing," *IEEE Journal of Solid-State Circuits*, 2021.

[10] 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 novel architecture to counter dual patterning issues in 10nm finfet," in *2017 30th International Conference on VLSI Design and 2017 16th International Conference on Embedded Systems (VLSID)*, pp. 269–274, IEEE, 2017.

[11] J. Chang, Y.-H. Chen, W.-M. Chan, S. P. Singh, H. Cheng, H. Fujiwara, J.-Y. Lin, K.-C. Lin, J. Hung, R. Lee, *et al.*, "12.1 a 7nm 256mb sram in high-k metal-gate finfet technology with write-assist circuitry for low-v min applications," in *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 206–207, IEEE, 2017.

[12] V. Nautiyal, G. Singla, S. Dwivedi, S. Singh, I. Chang, J. Dasani, and F. A. Bohra, "Self-timed shaper circuit for wide memories in advanced cmos technologies," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5, IEEE, 2018.

[13] V. Kumar, N. Kapoor, S. Kumar, M. Juneja, and A. Khanuja, "Area efficient & high performance word line segmented architecture in 7nm finfet sram compiler," in *2019 32nd International Conference on VLSI Design and 2019 18th International Conference on Embedded Systems (VLSID)*, pp. 437–442, IEEE, 2019.