Layout-Aware Optimization of STT MRAMs

Sumeet Kumar Gupta, Sang Phill Park, Niladri Narayan Mojumder and Kaushik Roy
School of Electrical and Computer Engineering, Purdue University, West Lafayette IN, USA. Email: guptask@purdue.edu

***Abstract*—We present a layout-aware optimization methodology for spin-transfer torque (STT) MRAMs, considering the dependence of cell area on the access transistor width (W_{FET}), number of fingers in the access transistor and the metal pitch of bit- and source-lines. It is shown that for W_{FET} less than a critical value (~7 times the minimum feature length), one-finger transistor yields minimum cell area. For large W_{FET}, minimum cell area is achieved with a two-finger transistor. We also show that for a range of W_{FET}, the cell area is limited by the metal pitch of bit- and source-lines. As a result, in the metal pitch limited (MPL) region, W_{FET} can be increased with *no* change in the cell area. We analyze the impact of increase in W_{FET} in the MPL region on the write margin and cell tunneling magneto-resistance (CTMR) of different genres of STT MRAMs. We consider conventional STT MRAM cells in the standard and reverse-connected configurations and STT MRAMs with tilted magnetic anisotropy for the analysis. By increasing W_{FET} from the minimum to the maximum value in the MPL region (at iso-cell area) and reducing read voltage to achieve iso-read disturb margin, 2X improvement in write margin and 27% improvement in CTMR is achieved for the reverse-connected STT MRAM. Similar trends are observed for other STT MRAM cells.**

Keywords- layout; MTJ;magnetic memories; optimization; STT MRAM; TMR

I. INTRODUCTION

Spin-transfer torque magnetic RAMs (STT MRAMs) have emerged as promising candidates for on-chip caches and embedded applications [1-2] because of high density, non-volatility and zero stand-by leakage. A conventional STT MRAM cell comprises of a magnetic tunnel junction (MTJ) and an access transistor (Fig. 1(a-b)). The resistance of the MTJ depends on the relative magnetization of the free layer (FL) with respect to the pinned layer (PL). Parallel magnetization of FL with respect to PL leads to a lower resistance (R_P) compared to the resistance in anti-parallel state (R_{AP}). The two resistances of the MTJ define the binary states of the memory cell. Write operation in the memory cell is performed by switching the magnetization of FL using current-induced spin-transfer torque. Let us define V_{BL}, V_{SL} and V_{WL} to be the bit-line (BL), source-line (SL) and word-line (WL) voltages. Switching from parallel to anti-parallel state is achieved by applying a voltage (V_{DD}) across the cell (i.e. $|V_{BL}-V_{SL}|=V_{DD}$) and asserting WL such that the current flows from PL to FL. Similarly, anti-parallel to parallel switching is performed by reversing the polarity of the voltage across the cell and passing a current from FL to PL. Successful write occurs if the write current (I_W) is greater than the critical switching current (I_C). Write-ability of the cell is typically expressed in terms of write margin (WM) which is defined as $(I_W-I_C)/I_C$ [2]. Read operation is performed by applying a read voltage (V_{READ}) across the cell ($V_{BL}-V_{SL}=V_{READ}$), asserting WL and sensing the cell current or the bit-line voltage. Cell tunneling magneto-resistance (CTMR), defined as $(R_{AP}-R_P)/(R_P+R_{FET})$, is a measure of the distinguish-ability of the two states of the cell. (Here, R_{FET} is the resistance of the access transistor). A large difference in R_P and R_{AP} is desired for low decision failures (defined as the inability to sense the difference in the binary states). Moreover, the read current (I_R) should be sufficiently lower than I_C to achieve low read disturb failures (defined as unexpected switching of the state during read). Read disturb margin (RDM), defined as $(I_C-I_R)/I_C$, is a measure of read stability of the cell [1].

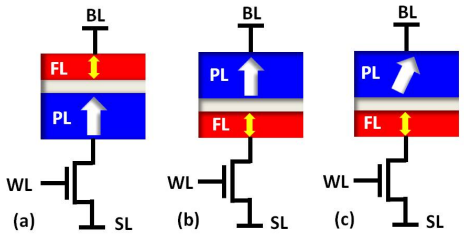

Fig. 1 Schematics of STT MRAM cells (a) in the standard-connected configuration (b) in the reverse-connected configuration and (c) with tilted magnetic anisotropy

One of the critical parameters in STT MRAM design is the width of the access transistor (W_{FET}) as the cell characteristics are strongly dependent on W_{FET} [3]. Increase in W_{FET} improves WM and CTMR, but leads to lower RDM. However, the effect of increase in W_{FET} on RDM can be mitigated by optimizing V_{READ}. Another impact of larger W_{FET} is a possible increase in the cell area (A_{CELL}). A_{CELL} of STT MRAM is believed to be limited by the transistor size since the size of MTJ is expected to be much smaller [4]. Hence, increase in W_{FET} is assumed to have a proportional impact on A_{CELL}. This restricts the sizing of the access transistor due to area constraints.

However, in this paper, we show that the estimation of A_{CELL} as a function of W_{FET} requires careful layout analysis. We perform a detailed evaluation of the dependence of A_{CELL} on W_{FET}. We show that for a range of W_{FET}, A_{CELL} is limited by the BL-SL metal pitch. Hence, in the metal pitch limited (MPL) region, W_{FET} can be increased without increasing A_{CELL}. We also analyze the dependence of A_{CELL} on the number of fingers in the access transistor (N_F).

Based on this analysis, a layout-aware optimization methodology for STT MRAMs is presented. Considering the MPL region (iso-area), we analyze the impact of increase in W_{FET} on cell characteristics. The analysis is carried out for conventional STT MRAM cells in the standard- and reverse-connected configurations (Fig. 1(a-b)) and an STT MRAM with tilted magnetic anisotropy (TMA) [5] (Fig. 1(c) – explained in Section III-C). We show that the proposed methodology leads to simultaneous increase in WM and CTMR at iso-RDM and iso-area. This expands the design space of STT MRAMs and allows the optimization of other design parameters. In particular, we investigate the joint optimization of W_{FET}-V_{READ}-V_{DD} in the MPL region. Furthermore, we analyze STT MRAM with two access transistors considering the cell layouts and perform an iso-area comparison with the conventional STT MRAM.

978-3-9810801-8-6/DATE12/©2012 EDAA

II. LAYOUT OF STT MRAM CELL

In this section, we perform a detailed analysis of the layout of STT MRAM cell. The layout dimensions and area are estimated based on general layout design rules. In order to show the trends of A_{CELL} with respect to W_{FET} and N_F, λ-based rules [6] are assumed (Here, λ is half of the minimum feature size). However, the analysis can be used for any design rules based on the general expressions obtained in this section.

Figs. 2(a-d) show the possible layouts of STT MRAM cells, drawn using the λ-based rules with the following modifications:
(a) The minimum metal width and metal spacing are assumed to be 3λ. This assumption is justified since memory layout rules are typically more relaxed than the logic design rules [7].
(b) Vias are placed on BL and SL metal tracks without a metal enclosure, as inferred from the design rules of a commercial process for long metal lines in a memory array design.

As mentioned before, the analysis provided in this section is general, and the trends and conclusions drawn from the analysis are not dependent on the aforementioned assumptions.

In an STT MRAM, the two vertical metal tracks, bit-line (BL) and source-line (SL) (Figs. 2(a-d)), are shared amongst the cells in the same column. The word-line (WL) runs horizontally and is shared amongst the cells in the same row. In general, a gate metal contact is shared by a large number of cells in a row. Hence, the average increase in area of the cell due to the gate metal contact is negligible.

Figs. 2 (a-b) show layouts with a single-finger transistor. SL contact of the cell is shared with an adjacent cell. The Y-dimension of the cell can be obtained as

$$Y(1) = W_C/2 + 2W_{G2C} + W_G + W_C + W_{C2A} + W_{A2A}/2 = 11.5\lambda \quad (1)$$

(See Fig. 2(e) for the definitions of the terms in (1)). The expression for X-dimension of the cell is obtained by noting that for small W_{FET}, the X-dimension is limited by the metal pitch (Fig. 2(a)). As W_{FET} increases beyond a certain value, the X-dimension is determined by W_{FET}, (Fig. 2(b)). Considering the two cases, the X-dimension is given by

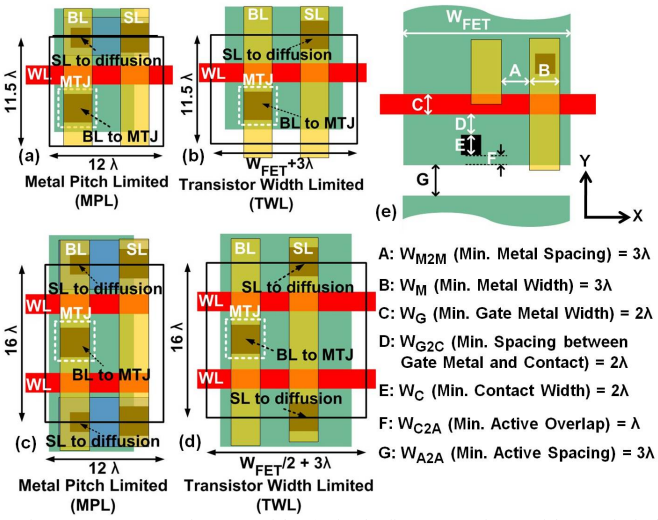

Fig. 2 STT MRAM layouts with (a) single-finger transistor with metal pitch limited (MPL) X-dimension (b) single-finger transistor with transistor width limited (TWL) X-dimension (c) two-finger transistor with metal pitch limited (MPL) X-dimension (d) two-finger transistor with transistor width limited (TWL) X-dimension (e) definition of dimensions and layout design rules.

$$X(1) = \max\left(2(W_{M2M} + W_M), (W_{A2A} + W_{FET})\right) = \max(12\lambda, W_{FET} + 3\lambda) \quad (2)$$

The first argument of max (maximum function) corresponds to the metal pitch limited (MPL) layout and the second term corresponds to the transistor width limited (TWL) layout.

Figs. 2(c-d) show the layouts with a two-finger transistor. SL contact of the cell is shared with two adjacent cells in the same column, resulting in a continuous diffusion region, unlike the one-finger transistor layout. The Y- and X-dimensions of a two-finger transistor layout are given by

$$Y(2) = 2(W_C + 2W_{G2C} + W_G) = 16\lambda \quad (3)$$

$$X(2) = \max\left(2(W_{M2M} + W_M), (W_{A2A} + W_{FET}/2)\right) = \max(12\lambda, 3\lambda + W_{FET}/2) \quad (4)$$

Note from (4), the X-dimension of a two-finger transistor layout may be limited by metal pitch or transistor width.

The expressions in (3) and (4) can be generalized for an N_F-finger transistor layout as follows.

$$Y(N_F) = \begin{cases} (N_F - 1)Y(2)/2 + Y(1) = (8N_F + 3.5)\lambda, & \text{Odd } N_F \\ N_F Y(2)/2 = 8N_F\lambda, & \text{Even } N_F \end{cases} \quad (5)$$

$$X(N_F) = \max\left(2(W_{M2M} + W_M), (W_{A2A} + W_{FET}/N_F)\right) = \max(12\lambda, 3\lambda + W_{FET}/N_F) \quad (6)$$

Cell area (A_{CELL}) is obtained as

$$A_{CELL}(N_F) = \begin{cases} \max\left((96N_F + 42), \left(\begin{matrix}(8 + 3.5/N_F)(W_{FET}/\lambda) + \\ 24N_F + 10.5\end{matrix}\right)\right)\lambda^2, & \text{Odd } N_F \\ \max\left(96N_F, 8(W_{FET}/\lambda) + 24N_F\right)\lambda^2, & \text{Even } N_F \end{cases} \quad (7)$$

Fig. 3(a) shows the plot of A_{CELL} versus W_{FET} for N_F=1 through 4. We make note of the following points from Fig. 3(a) and (5)-(7):
a) For $W_{FET}<9N_F\lambda$ (see (6)), the cell area is metal pitch limited (MPL) and A_{CELL} is independent of W_{FET}. A_{CELL} in MPL region is proportional to N_F. This is because in the MPL region, the X-dimension is constant and the Y-dimension increases with N_F.
b) For $W_{FET}>9N_F\lambda$, the cell area is transistor width limited (TWL) and A_{CELL} is proportional to W_{FET}. For even N_F, the slope of A_{CELL} with respect to W_{FET} is independent of N_F. For odd N_F, the slope is inversely proportional to N_F. The difference arises because for even N_F, the contacts of the first and last fingers are shared with the adjacent cells in the column. However, for odd N_F, only one of the contacts is shared with the adjacent cell. As a result, an N_F-independent term appears in the expression for the Y-dimension in (5) which accounts for the minimum active spacing, minimum active overlap and half the contact width for the contact that is not shared. The contribution of the N_F-independent term to the Y-dimension decreases as N_F increases and therefore, the slope of A_{CELL} with respect to W_{FET} decreases (Fig. 3(a)).
c) Amongst even N_F, N_F=2 yields minimum area. Area for all odd N_F >1 is greater than the area for N_F=2. Hence, for all W_{FET}, A_{CELL} for N_F=2 is smaller than A_{CELL} for N_F>2.
f) For small W_{FET}, N_F=1 has lower area than N_F=2. For large W_{FET}, minimum area is obtained for N_F=2.

Based on the observations (a)-(f), the plot of the minimum cell area with transistor width is obtained in Fig. 3(b). For $W_{FET}<9\lambda$, one-finger access transistor is used and the area is independent of W_{FET} (MPL region). For $9\lambda<W_{FET}<14\lambda$, one-finger transistor layout area increases with W_{FET} (TWL region). For $14\lambda<W_{FET}<18\lambda$, two-finger access transistor is used and the area is independent of W_{FET} (MPL region). For $W_{FET}>18\lambda$, two-finger transistor layout area increases with W_{FET} (TWL region).

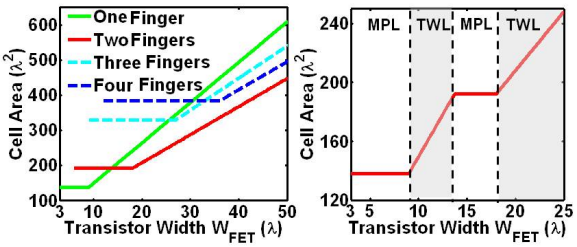

Fig. 3 (a) Cell area versus transistor width for different number of fingers in the access transistor and (b) optimal cell area versus transistor width showing metal pitch limited (MPL) and transistor width limited (TWL) regions.

The analysis presented in this section can be extended to STT MRAM cells with larger number of terminals, as in the dual-pillar STT MRAM [2] which has two bit-lines and one source-line. For such STT MRAM cells, *the consideration of MPL regions in the analysis becomes even more important due to larger number of vertical metal tracks in the layout*.

The detailed layout analysis leads to some interesting possibilities in STT MRAM cell optimization. In the MPL regions, W_{FET} can be increased without incurring any cell area penalty. In the next section, we analyze the impact of W_{FET} increase in the MPL region on the cell characteristics of STT MRAM and present a layout-aware optimization methodology.

III. Layout-Aware Analysis and Optimization of STT MRAM

W_{FET} is a critical design parameter in STT MRAM optimization. Increase in W_{FET} leads to increase in WM due to higher strength of the access transistor. Moreover, due to the decrease in R_{FET}, CTMR increases with increasing W_{FET}. However, higher read current (I_R) for larger W_{FET} leads to lower read disturb margin (RDM). Optimization of V_{READ} along with W_{FET} becomes important to achieve high RDM. In this section, we quantify the effect of increase in W_{FET} in the MPL regions (iso-area) on WM and CTMR at iso RDM. In order to achieve iso-RDM, V_{READ} is lowered with increase in W_{FET} to obtain I_R=0.5I_C. We perform the analysis for conventional STT MRAM, STT MRAM with TMA and STT MRAM with two transistors. The analysis is carried out using a simulation framework [3] based on Non-equilibrium Green's function (NEGF) models for electronic and spin transport in MTJ, Landau-Liftshitz-Gilbert (LLG) equations with an STT term for magnetization switching dynamics and 32nm predictive technology models (PTM) [8] for the access transistor.

A. Conventional STT MRAM

A conventional STT MRAM cell can be designed in two ways viz. in standard-connected configuration (free layer FL connected to BL – Fig. 1(a)) or in reverse-connected configuration (pinned layer PL connected to BL – Fig. 1(b)) [3]. The analysis for the standard-connected cell (Fig. 4(a)) shows that the range of W_{FET} required for successful write (WM>0) corresponds to the two-finger transistor layout region. On increasing W_{FET} from the minimum (W_{MIN}) to the maximum (W_{MAX}) value in the MPL region (along with reduction in V_{READ} to achieve iso-RDM), WM increases from -4.7% to 16% and CTMR increases from 85% to 107% at iso-RDM. For the reverse-connected cell, increase in WM from 18% to 36% and increase in CTMR from 83% to 106% at iso-RDM is observed in the MPL region (Fig. 4(b)). Recall, there is no area penalty in increasing W_{FET} to W_{MAX} in the MPL region.

Note, in our analysis, the reverse-connected cell yields higher WM than the standard-connected cell with comparable CTMR at iso-RDM. Hence, subsequently, we will consider the reverse-connected configuration only. Also note, there is a discontinuity in WM and CTMR at the boundary of one-finger and two-finger transistor layout regions, which is explained as follows. For the same W_{FET}, the width of each finger in a two-finger transistor is lower compared to the single-finger transistor. This leads to narrow width effects [9] in the two-finger transistor which reduce the transistor strength and result in mildly lower WM and CTMR.

B. STT MRAM with tilted magnetic anisotropy

MTJs with tilted magnetic anisotropy (TMA) (Fig. 1(c)) exhibit lower I_C compared to the standard MTJs due to relative tilt in the easy axes of PL and FL, which aids magnetization switching [5]. However, TMA in the MTJ leads to increase in read disturb and lower CTMR compared to the conventional STT MRAM. Therefore, higher T_{OX} is used in the MTJ stack to increase RDM and CTMR, while still achieving higher WM. As a result of superior write-ability, STT MRAMS with TMA have lower W_{FET} requirements. Fig. 4(c) shows that W_{FET} required for WM>0 lies in the one-finger transistor layout region. Increasing W_{FET} from W_{MIN} to W_{MAX} in the MPL region (iso-area) and reducing V_{READ} to achieve iso-RDM results in WM~ 75% along with increase in CTMR from 48% to 147%.

C. Layout-Aware Optimization

The above analysis leads to an important implication for the optimization of STT MRAM. If W_{FET} required to meet the specifications for cell stability, performance and area lies in the MPL region, W_{FET} may be increased to W_{MAX} and V_{READ} can be appropriately reduced to achieve higher WM and CTMR at iso-RDM without any area penalty. The proposed methodology

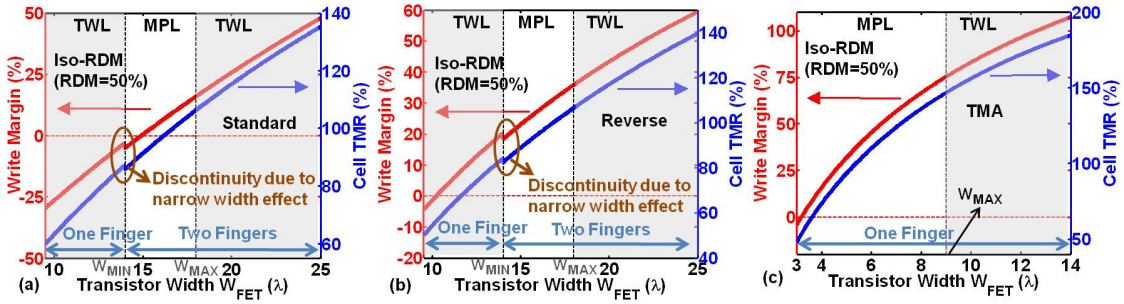

Fig. 4 Write margin and cell TMR versus transistor width at iso-RDM for (a) conventional STT MRAM with standard connection (b) conventional STT MRAM with reverse connection and (c) STT MRAM with tilted magnetic anisotropy (TMA).

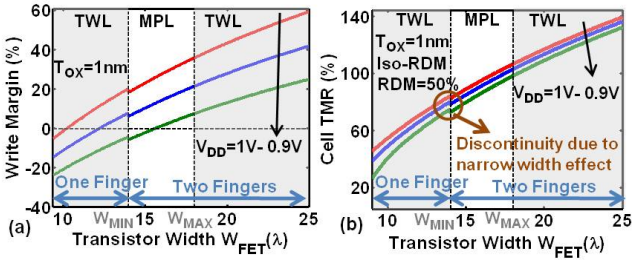

Fig. 5 (a) Write margin and (b) cell TMR versus transistor width at iso-RDM for different V_{DD} for conventional reverse-connected STT MRAM.

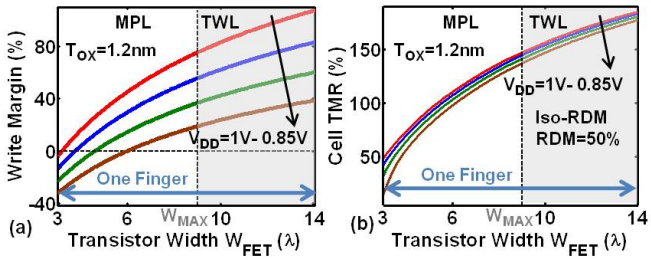

Fig. 6 (a) Write margin and (b) cell TMR versus transistor width at iso-RDM for different V_{DD} for STT MRAM with tilted magnetic anisotropy.

also allows the optimization of other cell characteristics. We discuss one such optimization technique next.

W_{FET}-V_{READ}-V_{DD} co-optimization: Increase in WM and CTMR at iso-RDM and iso-area allows reduction in V_{DD} to achieve lower write power and higher MTJ reliability [2]. As an example, it can be observed from Figs. 5(a-b) that if W_{FET} is increased from W_{MIN} to W_{MAX}, V_{DD} can be reduced by 50mV to obtain WM similar to that at the initial design point (W_{FET}=W_{MIN}, V_{DD}=1V, V_{READ}=218mV). At iso-RDM, the new design point (W_{FET}=W_{MAX}, V_{DD}=0.95V, V_{READ}=183mV) achieves CTMR of 102% compared to CTMR of 83% at the initial design point. In addition, 50mV V_{DD} reduction allows write operation at a lower power and with increased reliability. Similar optimizations can be performed for STT MRAM with TMA (Figs. 6(a-b)).

D. 2-Transistor STT MRAMs

In this sub-section, we analyze 2-Transistor STT MRAM (2T1R –Fig. 7(b)) and perform an iso-area comparison with conventional STT MRAM (1T1R-Fig. 7(a)). In a two-finger transistor layout of conventional STT MRAM (Fig. 2(c-d)), there are two WL metal tracks which are electrically connected to each other. Instead, if the two WL tracks are controlled independently using two WL signals (WL1 and WL2 – Fig. 7(b)), the design conflicts in STT MRAM optimization can potentially be mitigated [4] with no cell area penalty. During write, both the word-lines are asserted. Hence, WM of 2T1R is the same as 1T1R cell with a two-finger transistor. In order to decrease the read disturb failures, the strength of the access transistor is reduced by asserting only one word-line during the read operation. Fig. 7(c) shows significant increase in RDM at iso-V_{READ}. Hence, the conflict between write-ability and cell disturb is mitigated at iso-area. However, 2T1R cell shows a large decrease in CTMR. Therefore, this technique is useful in cases when the disturb failures are more dominant than the decision failures.

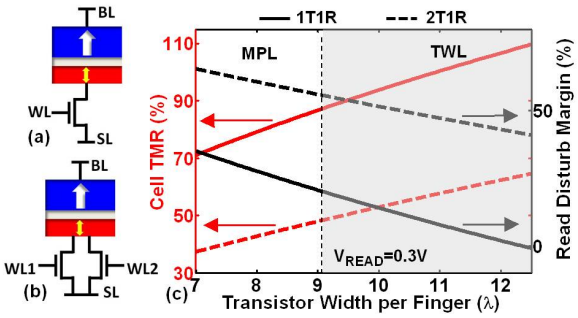

Fig. 7 (a) Schematic of 1T1R cell (b) schematic of 2T1R cell and (c) cell TMR and read disturb margin of 2T1R and 1T1R (with a two-fingered transistor).

IV. CONCLUSIONS

In this work, we presented a layout-aware design methodology for STT MRAMs. We performed a detailed analysis of the dependence of A_{CELL} on W_{FET} and N_F. We showed that for W_{FET} < 14λ, N_F=1 yields minimum A_{CELL}. For W_{FET} >14λ, N_F=2 should be used to achieve optimal A_{CELL}. Our analysis also showed that for a range of W_{FET}, the layout area is limited by the metal pitch of SL and BL. In the MPL regions, W_{FET} can be increased without increasing A_{CELL}. Based on the layout analysis, we showed that for different genres of STT MRAMs, improvement in WM and CTMR can be achieved at iso-RDM and iso-A_{CELL} by increasing W_{FET} in the MPL regions. We also discussed that the increase in WM in the MPL regions allows W_{FET}-V_{READ}-V_{DD} co-optimization to achieve lower write power and higher MTJ reliability. Finally we analyzed 2T1R STT MRAM and showed that 2T1R cell (with two transistors of equal width) does not show any A_{CELL} increase compared to 1T1R with a two-finger transistor. 2T1R cell shows improvement in RDM over 1T1R cell at iso-WM; however, this comes at the cost of degradation in CTMR.

ACKNOWLEDGEMENT

This research was funded in part by the Nano Research Initiative, INDEX center, Qualcomm and Intel Corporation.

REFERENCES

[1] A. Raychowdhury, D. Somashekhar, T. Karnik and V. De, "Design space and scalability exploration of 1T-1STT MTJ memory arrays in the presence of variability and disturbances," *IEDM*, Dec. 2009.

[2] N. N. Mojumder. S. K. Gupta and K. Roy, "Dual Pillar Spin Transfer Torque MRAM with tilted magnetic anisotropy for fast and error-free switching and near-disturb-free read operations," *DRC*, June 2011.

[3] X. Fong, S. H Choday and K. Roy, "Bit-cell Level Optimization for Non-volatile Memories Using Magnetic Tunnel Junctions and Spin-Transfer Torque Switching," *IEEE Trans. on Nanotech,* in press.

[4] J. Li, P. Ndai, A. Goel, S. Salahuddin and K, Roy, "Design Paradigm for Robust Spin-Torque Transfer Magnetic RAM (STT MRAM) From Circuit/Architecture Perspective," *TVLSI*, vol. 18, no.12, 2010.

[5] N. N. Mojumder and K. Roy,"Switching current reduction and thermally induced delay spread compression in tilted magnetic anisotropy spin-transfer torque (STT) MRAM," unpublished.

[6] Available: http://www.mosis.com/Technical/Designrules/scmos/

[7] L. Liebman, "DfM, the teenage years," *Proc of SPIE*, 692502, 2008.

[8] Available: http://ptm.asu.edu/

[9] K. Ohe, S. Odanaka, K. Moriyama, T. Hori and G. Fuse, "Narrow-width effects of shallow trench-isolated CMOS with n+ -polysilicon gate," *IEEE Trans. Electron Devices*, vol. 36, no. 6, 1989.