MRAM DTCO and Compact Models

Jeehwan Song, Jian-Ping Wang, and Chris H. Kim
Dept. of ECE, University of Minnesota, Minneapolis, MN, USA, email: chriskim@umn.edu

***Abstract*—**Design-Technology Co-Optimization (DTCO) has become an important design methodology for making early decisions on technology, circuit, and system design parameters. This invited paper introduces various aspects of DTCO for MRAM development, ranging from SPICE compatible Magnetic Tunnel Junction (MTJ) device models, array level spin transfer torque magnetoresistive random access memory (STT-MRAM) power-performance-area (PPA) evaluation, scalability and variability studies of large-scale arrays, and novel read and write circuit techniques.

I. Introduction: DTCO and STT-MRAM

DTCO has become an important paradigm as the semiconductor industry grapples with the mounting challenges of device scaling [1][2]. DTCO uses an initial set of compact models and design rules to generate a preliminary standard cell library, which is used to assess the PPA metrics of digital IP blocks. The PPA results are then utilized to refine the technology and device level parameters, and the entire technology/design optimization loop is repeated until a satisfactory set of technology and design parameters are obtained. Existing DTCO methodologies have focused on digital logic but it is expected that future DTCO will cover a wide range of devices including specialty technologies such as optoelectronics, spintronic memory, high voltage devices, etc.

STT-MRAM is the leading non-volatile memory candidate for on-chip data backup and cache applications, due to the low operating voltage, good CMOS compatibility, high speed, high density, zero static power, and high endurance [3]. STT-MRAM stores data in an MTJ device, which consists of a free layer, a tunnel oxide barrier, and a fixed layer as shown in Fig. 1. The effective resistance of an MTJ depends on the relative magnetization direction between the free and fixed layers. A higher resistance (R_{AP}) is obtained with an antiparallel magnetization configuration and a low resistance (R_P) is obtained with a parallel configuration. This invited paper covers various aspects of MRAM DTCO including device, circuit, and architecture considerations. We will also introduce key MRAM circuit and layout techniques adopted in industrial STT-MRAM designs for reliable read and write operations.

II. MTJ SPICE Model

An MTJ compact model is a critical component of the overall MRAM DTCO workflow described in Fig. 2. MTJ SPICE models should capture the key physics of STT-MTJ while reproducing the detailed read and write dynamics of a real MTJ device. An important advantage of implementing the MTJ spin dynamics in SPICE is that it offers a unified simulation environment for both MTJ and CMOS devices, facilitating full array PPA evaluation.

Thermal stability factor (Δ) is a key MTJ parameter defined as the free layer's energy barrier (E_b) between the P and AP states normalized to the thermal fluctuation (Fig. 3). Intuitively, Δ determines the retention time (and hence the degree of non-volatility) of the MTJ. Anisotropy field (H_K) is another crucial parameter determining the energetic preference of the magnetization vector, also known as the easy-axis. MTJs can be classified into in-plane and perpendicular anisotropy devices depending on the orientation of the easy-axis. In this paper, we consider an interfacial perpendicular MTJ (i-PMTJ) where the magnetic anisotropy originates from the free layer interface, since it has proven to have a low switching current and good scalability. Fig. 4 illustrates the MTJ SPICE model presented in [4]. The model implements the Landau-Lifshitz-Gilbert (LLG) equation, which governs the magnetization vector movement, using sub-circuits consisting of resistors, capacitors, and voltage-/current-dependent voltage/current sources. Fig. 5 shows the detailed mapping from the LLG equation to the equivalent circuit. The impact of MTJ dimensions and material parameters on MTJ characteristics such as anisotropy, STT switching, Tunnel Magnetoresistance Ratio (TMR), and temperature effect, can be studied using this MTJ model, while the switching probability can be adjusted using the initial angle parameter.

Model parameters can be tuned to match experimental data as shown in Fig. 6. The MTJ model is useful in studying variability effects at the array level considering both MTJ (W, L, t_F, RA) and CMOS (W, L, V_{th}, T_{ox}) variations (Fig. 7). Fig. 8 shows write and read delay distributions up to 6σ points. As the write voltage increases, the write delay distribution gets narrower due to the faster precession. For read, a higher TMR ratio, defined as $(R_{AP}-R_P)/R_P$, is required for improving the sensing margin. The MTJ model can also be utilized for first-order scalability studies. The required Δ for the target retention time can be met by changing the MTJ dimensions in different ways depending on the anisotropy source. Switching current requirement results in Fig. 9 suggest that i-PMTJ has a good scalability compared to other MTJs. The MTJ SPICE models introduced in the paper are available at mtj.umn.edu

III. STT-MRAM Array Level Evaluation

The scalability of perpendicular STT-MRAMs was studied in [5] based on process technology scaling trends from 65nm to 8nm. For an efficient variability analysis, the read and write operations were investigated using Monte Carlo simulations with an MTJ macromodel which includes key MTJ properties. In addition, ITRS projected transistor parameters were used for the access transistor and peripheral circuits.

A constant J_{C0}•RA/V_{DD} scaling scenario was assumed which corresponds to an iso-sensing margin across different technology generations. Here, J_{C0} is the critical current density and RA is the resistance-area product of an MTJ. The Δ for meeting a 10 years retention time was set by adjusting the free layer thickness and MTJ anisotropy. The diameter of the MTJ is fixed at F (=minimum metal width) for the smallest bit-cell size. The J_{C0}•RA and Δ play critical roles in the read/write operations and non-volatility of an STT-MRAM, and the J_{C0} scales by $1/\alpha^2$ where α is the scaling factor. The RA value is chosen for a constant J_{C0}•RA/V_{DD}. Table I shows scaling trends of an STT-MRAM based cache memory under the scaling scenario shown in Fig. 10. In this scaling study, we assumed that the number of cores double every two process generations For the initial setting, we assume a four core processor with a 4 MB per-core L3 cache in a 65nm technology [6]–[8]. The required Δ's were calculated using projected cache densities, access word size, and chip failure rate requirements. The performances of STT-MRAM and 6T-SRAM based caches were compared while considering variation effects in different technology nodes. I_{dsat} and V_{tsat} trends of core and thick T_{OX} devices for the STT-MRAM implementation are given in Fig. 11. Fig. 12 describes the sub-array architectures of the STT-MRAM cache presented in [5]. The array size of a STT-MRAM cache including all peripheral circuits is ~3 times smaller than that of a 6T-SRAM cache.

The simulation work includes process variation in the memory cells and sense amplifier (SA) circuit as well as variation of the wire resistances, capacitances, reference biases and supply levels (Table II). Based on the proposed scaling scenario and simulation methodology, simulation results in Fig. 13 shows that $J_{RD}/J_{C0}<2$ is required to avoid read disturb issues, and thus a J_{RD}/J_{C0} value of 1.5 was chosen in the work. The latencies between several embedded memories are compared as shown in Fig. 14. Critical path delay simulation results show that the normalized WL-to-BL read sensing delays of 6T SRAM, 1T1C eDRAM[9], 2T eDRAM[10], STT-MRAM are approximately 1x, 5x, 2x, and 3x, respectively. For small caches (e.g. L1), SRAM achieves the shortest cache latency because the BL sensing delay occupies a larger portion of the total latency. For larger caches (L3 or L4), however, the global interconnect delay dominates the cache latency making dense memories more desirable from a performance standpoint. Even though STT-MRAM has a 3-5x longer BL delay than that of an SRAM, it can outperform SRAMs when the cache size is greater than 64 Mb. Fig. 15 shows the 6σ sensing delay and write delay trends of STT-MRAMs under the proposed constant J_{C0}•RA/V_{DD} scaling scenario. The sensing delay is reduced with technology scaling and a higher TMR ratio. On the other hand, the write performance becomes worse with technology scaling, at least in planar CMOS technology, due to the lower drive current of the access transistor devices. These results follow basic circuit intuition where read and write operations always have conflicting requirements.

IV. Novel Read and Write Circuits, Bitcell Layout Considerations

In this section, we introduce circuit design techniques published in recent literature that can mitigate the impact of PVT variation on STT-MRAM write and read operations [11-15]. Future MRAM DTCO methodologies may have to take into the consideration the advances in write and read circuits, which can provide significant benefits over standard MRAM circuit implementations.

A write-verify-write (W-v-W) scheme was proposed to reduce the write error rate [11]. As shown in Fig. 16 (a), the MRAM cell is written repetitively until the correct data value is verified. The write error rate can be reduced significantly at the expense of a higher write energy consumption and longer write time. To optimize the write sequence, a write driver circuit with a programmable W-v-W scheme was introduced in [12] (Fig. 17). Using strong or weak write settings, the write current pulses can be carefully tuned using different write pulse count, write pulse width, and total write time (Fig. 16(b)). In [13], an array architecture with a two-column common source line (CSL) was introduced to reduce the parasitic resistance. This improves read and write margins with a modest layout area overhead (Fig. 18). To further improve the sensing margin, a negative voltage was applied to the unselected WLs to suppress the BL leakage current. A write-inhibit voltage is applied on the BL of the unselected cells to reduce the gate leakage current and voltage stress on the access transistors. The same voltage is applied to the BL and SL of the neighbor cell sharing the same CSL to prevent unwanted switching. As shown in Fig. 19, the sensing circuit includes clamp NMOS trimming, half-V_{DD} detection, and 1T4MTJ reference cell. The clamp transistors suppress the BL voltage during read operation to prevent read disturbance, and its size can be tuned by the clamp NMOS trimming circuit to remove the offset of the sensing circuit. The half-V_{DD} detection feature can improve the sensing margin by extending the signal development time. The 1T4MTJ reference cell provides a stable $(R_P+R_{AP})/2$ reference value without causing read disturbance in the reference cell. These write and sensing circuit design techniques can effectively improve the yield of large STT-MRAM arrays.

Two different STT-MRAM bit-cell layout styles have been proposed [14][15] (Fig. 20). The 1T1MTJ layout offers a smaller minimum size cell while the 2T1MTJ layout with a folded access transistor provides a higher write current per silicon area. The later layout is preferred for embedded applications where operating speed is an important design consideration. The write error rate of STT-MRAM bit-cells can be managed more efficiently with the adoption of FinFET technology due to the higher drive current with increasing fin height [15].

Acknowledgment

This work was supported in part by C-SPIN, one of the six SRC STARnet Centers, through MARCO and DARPA and in part by the NSF/SRC E2CDA Program.

References

[1] G. Yeric, et al., CICC, 2013. [2] A. Asenov, et al., TED, 2015. [3] J.-P. Wang, et al., Proceeding of DAC, 2017. [4] J. Kim, et. al., CICC, 2015. [5] K. Chun, et al., JSSC, 2013. [6] S. Rusu et al., JSSC, 2010. [7] S. Rusu et al., JSSC, 2007. [8] R. J. Riedlinger et al., ISSCC, 2011. [9] J. Barth et al., JSSC, 2011. [10] K. Chun, JSSC, 2012. [11] H. Noguchi, et al., ISSCC, 2016. [12] L. Wei, et al., ISSCC, 2019. [13] Y.-D. Chih, et al., ISSCC, 2020. [14] R. Takemura, et al., JSSC, 2010. [15] O. Golonzka, et al., IEDM, 2018.

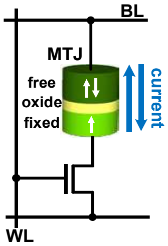


Fig. 1. STT MRAM bitcell

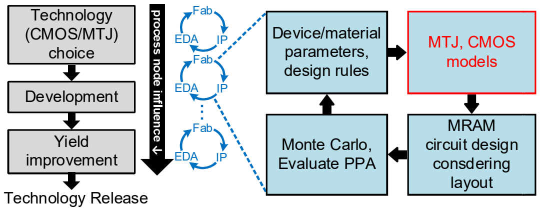

Technology (CMOS/MTJ) choice
Development
Yield improvement
Technology Release
process node influence
Fab
EDA
IP
Device/material parameters, design rules
MTJ, CMOS models
Monte Carlo, Evaluate PPA
MRAM circuit design consdering layout

Fig. 2. DTCO for MRAM development.

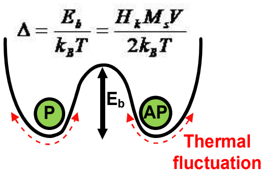


Fig. 3. Thermal stability factor (Δ) and energy diagram of P and AP states.

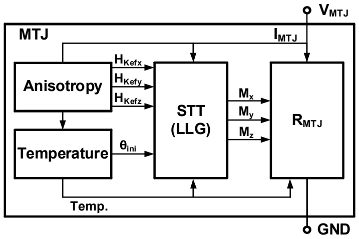
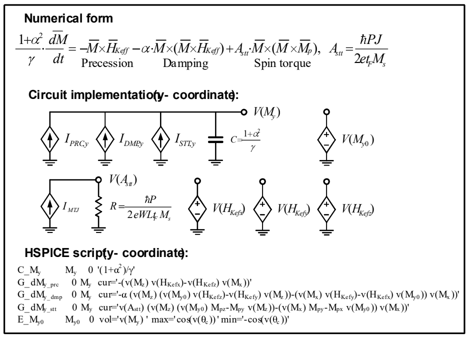


Fig. 4. Block diagram of STT-MTJ SPICE compact model.


Numerical form

$$\frac{1+\alpha^2}{\gamma}\cdot\frac{d\overline{M}}{dt} = \underbrace{-\overline{M}\times\overline{H}_{Keff}}_{\text{Precession}} \underbrace{-\alpha\cdot\overline{M}\times(\overline{M}\times\overline{H}_{Keff})}_{\text{Damping}} + \underbrace{A_{stt}\cdot\overline{M}\times(\overline{M}\times\overline{M}_p)}_{\text{Spin torque}}, \quad A_{stt} = \frac{\hbar PJ}{2et_F M_s}$$

Circuit implementation(y- coordinate):

HSPICE script(y- coordinate):

```
C_My        My   0  '(1+α^2)/γ'
G_dMy_prc    0 My  cur='-(v(Mz) v(HKefx)-v(HKefz) v(Mx))'
G_dMy_dmp    0 My  cur='-α (v(Mz) (v(My0) v(HKefz)-v(HKefy) v(Mz))-(v(Mx) v(HKefy)-v(HKefx) v(My0)) v(Mx))'
G_dMy_stt    0 My  cur='v(Astt) (v(Mz) (v(My0) Mpz-Mpy v(Mz))-(v(Mx) Mpy-Mpx v(My0)) v(Mx))'
E_My0       My0  0  vol='v(My) ' max='cos(v(θc))' min='-cos(v(θc))'
```

Fig. 5. SPICE implementation of LLG equation (y-coordinate only).

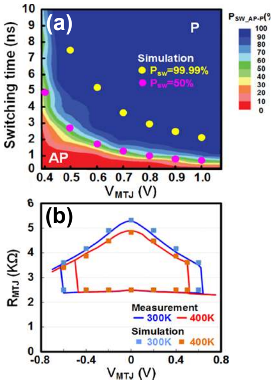


Fig. 6. Model verification: (a) switching probability, (b) R-V curves.

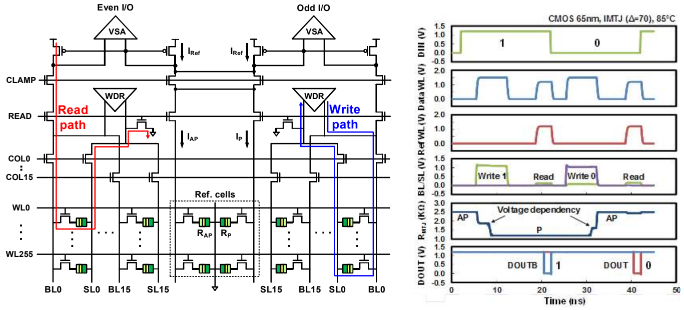


Fig. 7. STT-MRAM column circuit for read and write simulations. Simulation waveforms for write 1 → read 1 → write 0 → read 0 sequence.

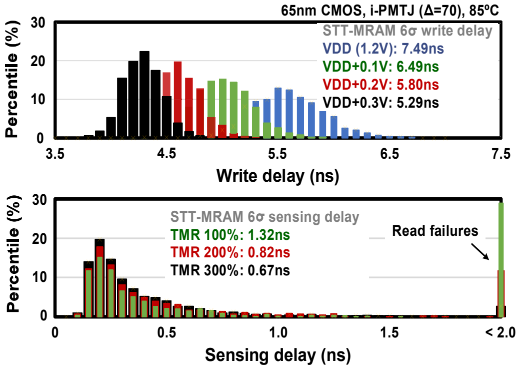
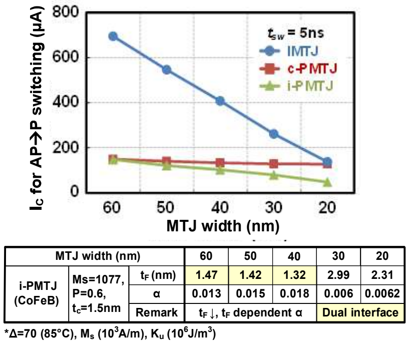


Fig. 8. Write delay and read delay distributions under different voltage and TMR conditions.



MTJ width (nm)			60	50	40	30	20
i-PMTJ (CoFeB)	Ms=1077, P=0.6, t_c=1.5nm	t_F (nm)	1.47	1.42	1.32	2.99	2.31
		α	0.013	0.015	0.018	0.006	0.0062
		Remark	t_F ↓, t_F dependent α			Dual interface	

*Δ=70 (85°C), M_s (10^3A/m), K_u (10^6J/m^3)

Fig. 9. Critical switching current (I_c) scaling trend assuming a 10 year retention failure rate target of 0.01% for a 128 MB cache.

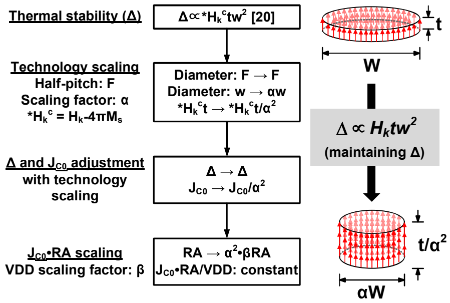


Fig. 10. STT-MTJ device scaling scenario assuming dimensional scaling and improved MTJ material parameters.

Technology node (nm)	65	45	32	22	15	11	8
Improvement of Transistor performance	Planar bulk			Multi-gate			
	Strained-Si						
	Poly gate		High-K metal gate				
VDD: Supply voltage (V)	1.2	1.1	1	0.9	0.85	0.8	0.75
On-chip cache memory size (MByte)	16	24	32	48	64	96	128
Number of cores	4	6	8	12	16	24	32
Δ: Thermal stability (for 10 yrs retention)	72	73	74	74	75	75	76
*H_k^c•t: Anisotropy and t	1.00	2.11	4.19	8.93	19.32	36.24	68.91
$J_{C0\_P\text{-}AP}$: Critical current density (MA/cm^2)	1.50	3.16	6.28	13.40	28.97	54.35	103.37
$I_{C0\_P\text{-}AP}$: Threshold write current (μA)	49.8	50.2	50.5	50.9	51.2	51.7	52.0
J_{C0}•RA/VDD (MTJ voltage headroom)	0.25 when TMR=150%						
R_{AP}A: Resistance area product (Ω•μm^2)	20.00	8.71	3.98	1.68	0.73	0.37	0.18
J_{RD}/J_{C0}: Read current density	1.50 (based on Fig. 8)						

*t and H_k^c•t are normalized to 65nm technology node.

Table I. STT-MTJ parameter values for the scaling scenario described in Fig. 10

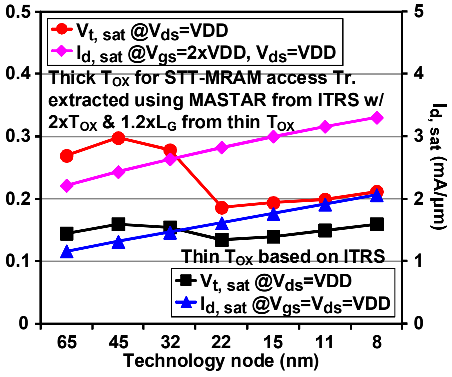


Fig. 11. High performance transistor scaling trend.

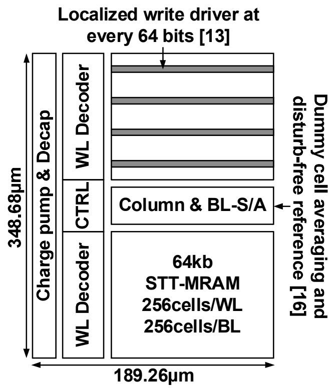

Localized write driver at every 64 bits [13]
348.68μm
Charge pump & Decap
WL Decoder
CTRL
WL Decoder
Column & BL-S/A
64kb STT-MRAM 256cells/WL 256cells/BL
Dummy cell averaging and disturb-free reference [16]
189.26μm

Fig. 12. 128 kb sub-array architecture for STT-MRAM based cache memory. The width of access transistors is set as $12F^2$ following the layout style in [3].

	STT-MRAM
Power supply noise	-10% to account for supply noise
Bit-cell	Device mismatches
Parasitic capacitance (C_{BL})	σ/μ=5%: each μ are calculated based on sub-array size
Resistance area product	σ/μ=5%
Sense Amplifier (S/A)	I-applying and V-sensing method (AP direction read) + Voltage S/A : I_{REF} σ/μ=2.5%, S/A pair mismatches
Reference cell	Reference cell averaging scheme with MTJ replica cells
Write threshold current	σ/μ=5%

* Mismatches are based on inverse square root relationship of devices' areas.
* Based on historic data, we assume σ_{Vt}/F is constant with tech. scaling
* $\mu(C_{BL})$ is assumed to be scaled proportional to scaling factor.

Table II. Simulation set-up including practical variation sources for evaluating STT-MRAM variability.

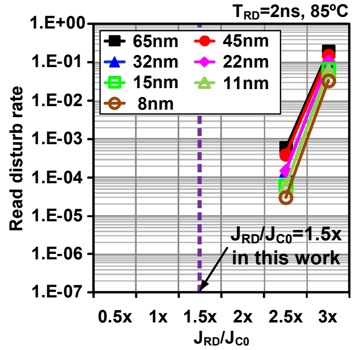


Fig. 13. Read disturb rate for different J_{RD}/J_{C0} ratios.

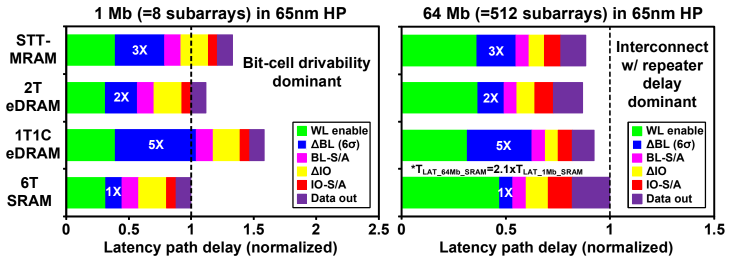


Fig. 14. Latency comparison between several embedded memories (STT-MRAM, eDRAM, and SRAM) for (a) 1 Mb cache and (b) 64 Mb cache.

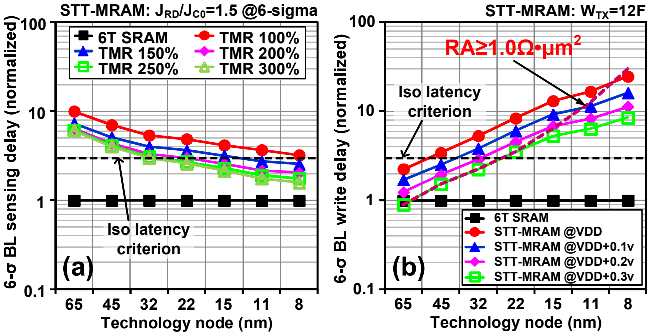


Fig. 15. STT-MRAM scaling trends: (a) Bit-line sensing delay and (b) write delay (Both normalized to SRAM delays for comparison)

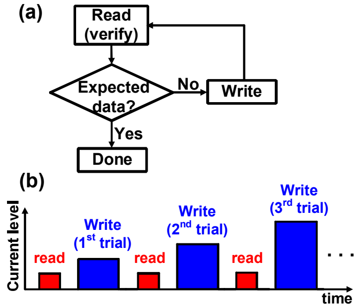


Fig. 16. Write-verify-write operation: (a) Flow chart and (b) current pulse sequence [11] (Intel 22nm).

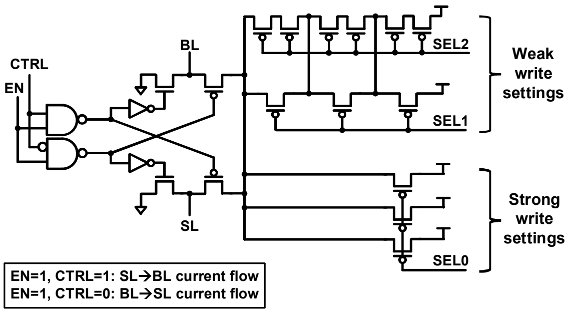
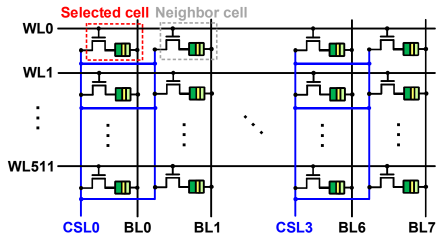

CTRL
EN
BL
SL
SEL2
SEL1
SEL0
Weak write settings
Strong write settings
EN=1, CTRL=1: SL→BL current flow
EN=1, CTRL=0: BL→SL current flow

Fig. 17. Write driver for programmable write-verify-write scheme [11] (Intel 22nm).



	READ	WRITE 0	WRITE 1
WL0	V_{WL_READ}	V_{WL_WRITE0}	V_{WL_WRITE1}
WL<1:511>	$V_{NEGATIVE}$	$V_{NEGATIVE}$	$V_{NEGATIVE}$
BL0	V_{BL_READ}	V_{PP}	0
BL1	0	0	V_{PP}
BL6	0	$V_{INHIBIT}$	$V_{INHIBIT}$
BL7	0	$V_{INHIBIT}$	$V_{INHIBIT}$
CSL0	0	0	V_{PP}
CSL3	0	$V_{INHIBIT}$	$V_{INHIBIT}$

Fig. 18. MRAM bitcell array with shared SL. Voltage bias table for read and write operations [12] (TSMC 22nm).

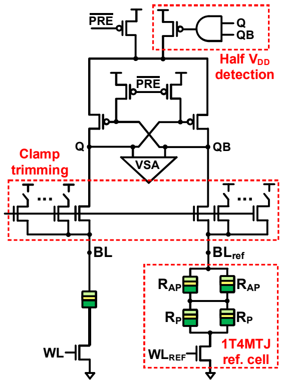


Fig. 19. Schematic of MRAM sensing circuit including half-V_{DD} detection, clamp NMOS trimming, and disturb-free reference cell [12] (TSMC 22nm).

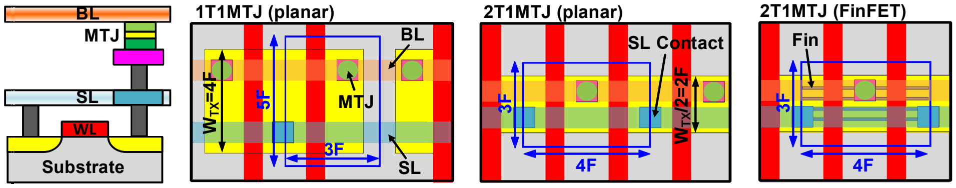


Fig. 20. STT-MRAM bit-cell layout comparison: 1T1MTJ (planar), 2T1MTJ (planar), and 2T1MTJ (FinFET).