30.2 A 1Mb 28nm STT-MRAM with 2.8ns Read Access Time at 1.2V VDD Using Single-Cap Offset-Cancelled Sense Amplifier and In-situ Self-Write-Termination

Qing Dong[1,2], Zhehong Wang[1], Jongyup Lim[1], Yiqun Zhang[1], Yi-Chun Shih[3], Yu-Der Chih[3], Jonathan Chang[3], David Blaauw[1], Dennis Sylvester[1]

[1]University of Michigan, Ann Arbor, MI; [2]TSMC, San Jose, CA
[3]TSMC, Hsinchu, Taiwan

1T1R spin-transfer-torque (STT) MRAM is a promising candidate for next-generation high-density embedded non-volatile memory [1-2]. However, 1T1R STT-MRAM suffers from limited sensing margin and high write power. As shown in Fig. 30.2.1(a), sense amplifier design is challenging due to the small difference (only 2×) between the high-resistance state (R_{AP}) and the low-resistance state (R_P), as well as R_{AP} degradation with increasing temperature. Moreover, R_P and R_{AP} resistance distributions shift with process variation, requiring a read reference (V_{ref}) that tracks process. To improve the sensing margin, several offset-cancellation methods have been reported to reduce sense amplifier mismatch [3]. However, these methods use multiple capacitors and hence incur significant area overheads. To address this issue, we propose an offset-cancelled sense amplifier that uses only a single capacitor to significantly improve the sensing margin by more than 60%. A second design challenge for STT-MRAM stems from the high current needed to flip a cell during a write operation. For non-volatile memory applications with a 10-year retention time requirement, the write current can be as high as several hundred μA. However, as shown in Fig. 30.2.1(b), the required write time varies with the state change required (0→1 or 1→0), process variation, and temperature. As a result, a fixed write time that ensures successful write for all conditions wastes a significant energy for typical or average conditions. We propose an *in situ* write-self-termination method to reduce write energy in most scenarios. The sense amplifier is reconfigured to continuously monitor the write operation and automatically shuts off the write drivers when the state transition is detected, without an area or timing penalty. In addition, dual dummy columns are added in each array to provide read V_{ref} tracking of row-wise PVT variation. A 1Mb STT-MRAM was fabricated in 28nm technology, and achieves a 2.8ns read-access time at 25°C and 3.6ns at 120°C, respectively. With *in-situ* self-write-termination the write power is reduced by 47% with a 20ns write-access time at 25°C and by 60% at 120°C.

Figure 30.2.1(c) shows a detailed block diagram of the 1Mb STT-MRAM, which contains 8 arrays, each with 256×514 cells. To improve read performance and sensing margins, a constant-current sensing method is used during read operations. Figure 30.2.2(a) shows a constant read current I_{read}, which can be as high as 25% of the write current, is applied to the selected BL. A BL voltage develops quickly and is sensed by the offset-cancelled amplifier. No NMOS voltage clamp is employed in the read path to avoid read disturbances; instead, a PMOS current clamping header transistor is used to limit the BL current for this purpose. Since these PMOS headers also function as write drivers, their sizes are less constrained, alleviating the mismatch of I_{read} without any additional area overhead.

Each array has two extra columns to generate the reference voltage, V_{ref}, on ref BL. A reference cell in one column is programmed to R_P and the other to R_{AP}. When WL is asserted, the two reference columns are activated and by setting the ref BL current to $2I_{read}$ an intermediate reference voltage $V_{ref} = 2I_{read}\,(R_P \| R_{AP})$ is generated and distributed to 16 two-stage sense amplifiers [4]. Row-wise PVT variations are tracked and compensated for as the reference cells are located in the same row as the selected cells.

For the second-stage sense amplifier, a single-capacitor offset cancellation method is proposed to minimize mismatch. Figure 30.2.2(c) shows simplified configurations of the design for each timing phase. During offset cancellation, the inputs of the inverters are connected to their outputs. The capacitor C_0 samples the difference between the trip voltages of the two inverters (V_L - V_R). Then, during BL sampling, this voltage is added to V_{ref} and the input of the right inverter becomes V_{ref} + V_L - V_R while that of the left inverter remains V_{ref}. This shifts the trip point during evaluation such that the mismatch of the two inverters is greatly reduced. Based on 10k Monte-Carlo simulations (Fig. 30.2.2(b)), the standard deviation of the input offset is reduced by more than 60% compared to a conventional sense amplifier with the same transistor sizes. Both offset cancellation and precharge phases are performed simultaneously with address decoding, avoiding any timing penalty for the proposed method. With only a single capacitor added, the area of the sense amplifier is reduced by approximately 15% compared to sense amplifiers using dual capacitors for offset cancellation.

Figure 30.2.3 shows the proposed *in-situ* self-termination write method. During write, hundreds of μA's are applied to the cell, with current flowing from BL to SL when writing a 1 and from SL to BL for 0. When the free magnetic layer in the STT bitcell flips, the resistance of the cell changes abruptly, which can be detected by observing the BL voltage. Since the resistance change and current direction both flip polarity between a write 1 and 0, the BL voltage drops in both cases when the write completes (Fig. 30.2.3(b)), simplifying write completion detection. The read sense amplifier is reconfigured as a continuous gate-connected voltage sense amplifier to detect the BL voltage drop as shown in Fig. 30.2.3(a). Since the write driver and detection circuit are shared with the read circuit, area overhead is negligible. Offset cancelation and precharge phases are overlapped with decoding, avoiding timing penalties. Once write completion is sensed, the 'stop' signal disables the write driver on that BL (Fig 30.2.3(c)). This method auto terminates the write when complete, saving write power and improving reliability. In addition, redundant writes are seamlessly detected and avoided which conventionally require a read-before-write [5] operation and incur a full read cycle overhead.

The proposed 1Mb STT-MRAM is fabricated in 28nm embedded MRAM technology. Figure 30.2.7 shows the die photo of the MRAM chip. The 1Mb MRAM macro occupies 0.214mm^2. Figure 30.2.4(a) shows the read-failure vs access time for a single 128kb array, without any error correction. The array achieves a 2.8ns read-access time with an approximate error rate of 10^{-5} at 25°C, and 3.6ns at 120°C. For an access time less than 2.8ns, higher error rates are observed with the self-generated V_{ref} because R_{BL} takes longer to stabilize due to driving 16 sense amps. However, read self-reference generation shows an error rate similar to an externally generated V_{ref} for access times greater than 2.8ns at room temperature. However, it exhibits a lower error rate at higher temperatures since it dynamically tracks the change in resistance with temperature. Figure 30.2.4(b) shows that the self-reference generation tracks array-to-array variation, improving the read failure rate from 2×10^{-5} to 9×10^{-6} compared with a fixed externally generated V_{ref} as measured across 8 arrays.

The measured Shmoo plot in Fig. 30.2.4(c) shows that the memory can be successfully read below 0.6V. Figure 30.2.4(d) shows the measured V_{DD_min} across 10 dies for the proposed sense amplifier. Due to offset cancellation, the sense amplifier is robust at low supply voltages with an average V_{DD_min} of 0.57V (σ=19mV).

Figure 30.2.5(a) shows that the required write access time must be greater than 20ns in order to achieve a 10^{-5} failure rate using a conventional, fixed write time. Since the required write time varies significantly across cells, a constant write time wastes substantial write power and the proposed self-write-termination saves 47% of write power for a 10^{-5} error rate at 20ns, and 61% for a 10^{-6} error rates at 30ns. When the temperature increases, the self-write-termination is more effective (Fig. 30.2.5(b)) because the write time decreases with temperature. Figure 30.2.6 compares this work to other MRAM work: compared with the listed references, the proposed MRAM achieves the best read-access time and power consumption within the smallest macro area.

Acknowledgements:
This work was supported by TSMC university joint development program and university shuttle program.

References:
[1] H.-C. Yu, et al., "Cycling Endurance Optimization Scheme for 1Mb STT-MRAM in 40nm Technology," *ISSCC*, pp. 224-225, 2013.
[2] K. Tsuchida, et al., "A 64Mb MRAM with Clamped-Reference and Adequate-Reference Schemes," *ISSCC*, pp. 258-259, 2010.
[3] M.-F. Chang, et al., "An Offset-Tolerant Fast-Random-Read Current-Sampling-Based Sense Amplifier for Small-Cell-Current Nonvolatile Memory," *JSSC*, vol. 48, no. 3, pp. 864-877, March 2013.
[4] D. Halupka, et al., "Negative-Resistance Read and Write Schemes for STT-MRAM in 0.13μm CMOS," *ISSCC*, pp. 256-257, 2010.
[5] H. Noguchi, et al., "4Mb STT-MRAM-Based Cache with Memory-Access-Aware Power Optimization and Write-Verify-Write/Read-Modify-Write Scheme," *ISSCC*, pp. 132-133, 2016.
[6] R. Nebashi, et al., "A 90nm 12ns 32Mb 2T1MTJ MRAM," *ISSCC*, pp. 462-463, 2009.
[7] H. Noguchi, et al., "A 3.3ns-Access-Time 71.2μW/MHz 1Mb Embedded STT-MRAM Using Physically Eliminated Read-Disturb Scheme and Normally-Off Memory Architecture," *ISSCC*, pp. 136-137, 2015.
[8] H. Noguchi, et al., "A 250-MHz 256b-I/O 1-Mb STT-MRAM with Advanced Perpendicular MTJ based Dual Cell for Nonvolatile Magnetic Caches to Reduce Active Power of Processors," *Symp. VLSI Circuits*, pp. 108-109, 2013.

 978-1-5090-4940-0/18/$31.00 ©2018 IEEE

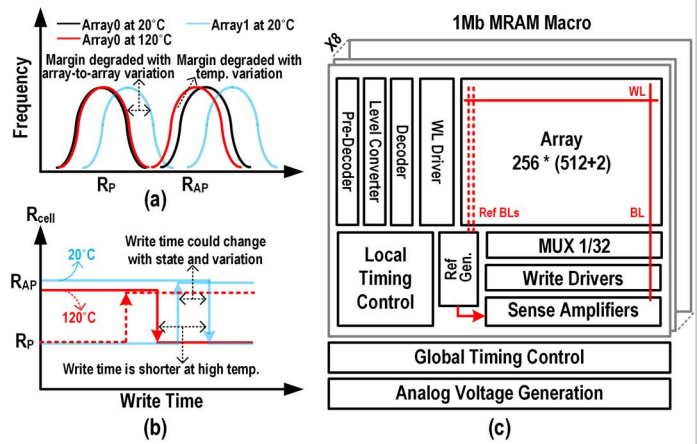

Figure 30.2.1: Required write time and read margin vary with PVT, wasting write power and posing a challenge for sensing circuit design. Row-wise reference cells are used to generate a read-reference voltage that tracks row-wise PVT variation.

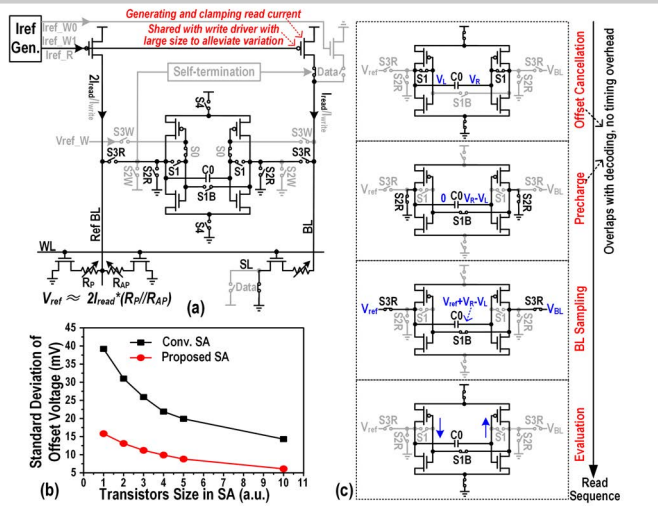

Figure 30.2.2: Constant-current based voltage sensing with a single-capacitor based offset cancellation. Compared with a conventional sense amplifier, input offset is reduced by >60%, significantly improving the sensing margin.

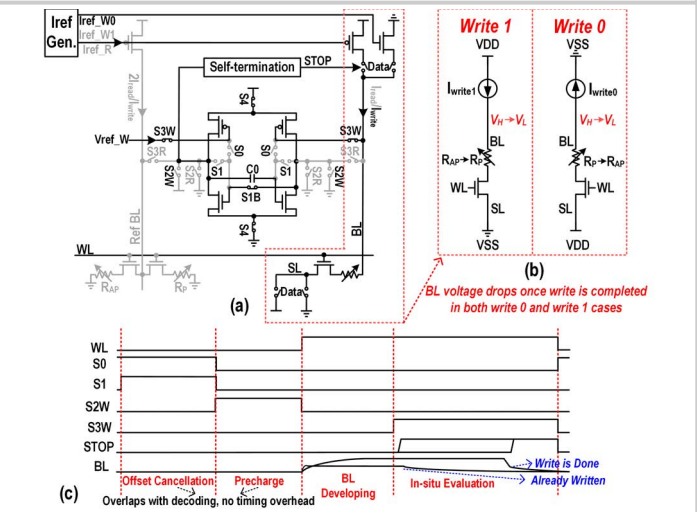

Figure 30.2.3: *In-situ* write-end detection and self-write termination using a continuous offset-cancelled sense amplifier reconfigured from the read-sense amplifier with no timing and area overhead.

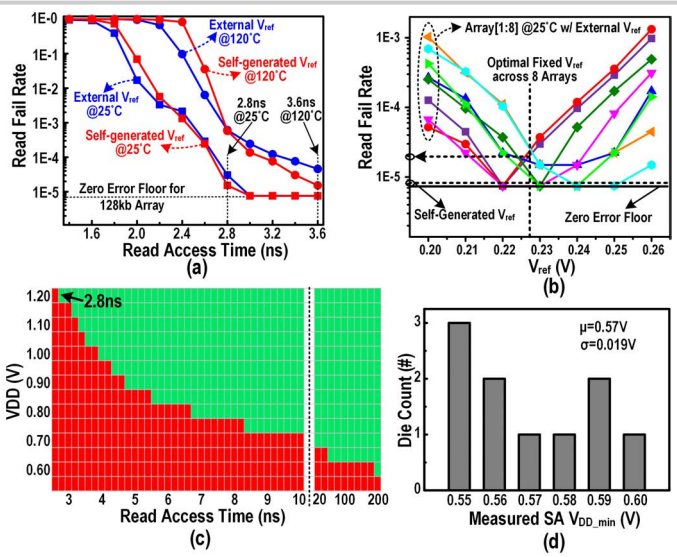

Figure 30.2.4: Read operation measured results.

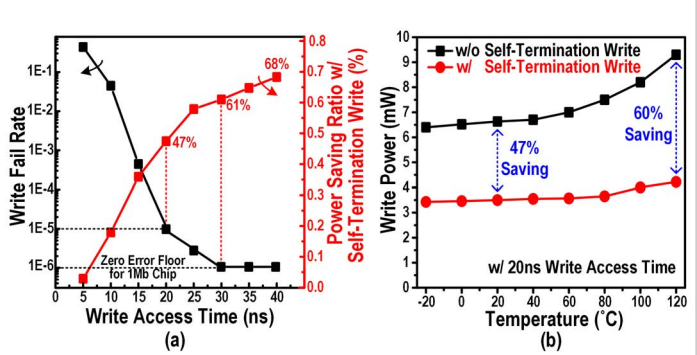

Figure 30.2.5: Write operation measured results.

	This Work	ISSCC'13 [1]	ISSCC'10 [2]	ISSCC'10 [4]	ISSCC'09 [6]	ISSCC'15 [7]	VLSI'13 [8]
Technology (nm)	**28**	40	65	130	90	65	65
Cell Type	**1T1MTJ**	1T1MTJ	1T1MTJ	2T1MTJ	2T1MTJ	2T2MTJ	2T2MTJ
Target Application	**NVM**	NVM	NVM		LLC1)	LLC1)	LLC1)
Cell Area (F^2)	**75**		85	327	169		107
Capacity	**1Mb**	1Mb	64Mb	16kb	32Mb	1Mb	1Mb
Macro Area (mm^2)	**0.214**	0.57	47.124		91.02	0.8196	0.628
Power Supply (V)	**1.2/1.8**	1.1/2.5	1.2	1.2/3.3	1.5	1.2/0.9/0.4	1.2/0.9/0.4
Word Length (bit)	**16**	32	16		32	256	256
Read Speed (ns)	**2.8**	10	30	8	12	3.3	4
Write Speed (ns)	**20**		30	10	12	3	4
Read Power (mW)	**3.9**		7.8		60	21.6	17.8
Write Power (mW)	**3.6**		9.3	20	91	55.4	46.5
Read Energy (pJ/bit)	**0.7**		14.6		22.5	0.3	0.3
Write Energy (pJ/bit)	**4.5**		17.4		34.1	0.6	0.7

1) MTJ cell in LLC applications do not require high retention, allowing low switching current MTJs

Figure 30.2.6: Comparison table to other STT-MRAM work.

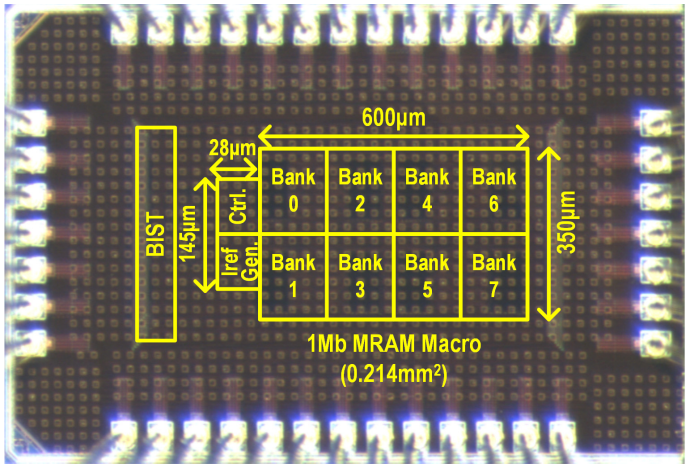

Figure 30.2.7: Die photo of 1Mb 28nm STT-MRAM.

978-1-5090-4940-0/18/$31.00 ©2018 IEEE