Fully Functional Perpendicular STT-MRAM Macro Embedded in 40 nm Logic for Energy-efficient IOT Applications

Yu Lu[a,1], Tom Zhong[b,3], W. Hsu[a], S. Kim[a], X. Lu[a], J.J. Kan[a], C. Park[a], W.C. Chen[a], X. Li[a], X. Zhu[a], P. Wang[a], M. Gottwald[a], J. Fatehi[a], L. Seward[a], J.P. Kim[a], N. Yu[a], G. Jan[b], J. Haq[b], S. Le[b], Y.J. Wang[b], L. Thomas[b], J. Zhu[b], H. Liu[b], Y.J. Lee[b], R.Y. Tong[b], K. Pi[b], D. Shen[b], R. He[b], Z. Teng[b], V. Lam[b], R. Annapragada[b], T. Torng[b], Po-Kang Wang[b,4], S.H. Kang[a,2]

[a] Qualcomm Technologies, Inc., San Diego, CA 92121. [1]E-mail: yulu@qti.qualcomm.com, Tel: (858) 845-6751, [2]E-mail: kang@qti.qualcomm.com

[b] TDK-Headway Technologies, Inc., Milpitas, CA 95035. [3]E-mail: tom.zhong@headway.com, [4]E-mail: pokang.wang@headway.com

Abstract

We present for the first time a fully functional 40 nm perpendicular STT-MRAM macro (1 Mb, ×32/×64 IO) embedded into a foundry standard CMOS logic platform. We achieved target design specifications of 20 ns read access time and 20-100 ns write cycle time without redundancy repair at standard core and IO voltages. The full 1 Mb macro can be switched reliably with write pulse as short as 6 ns, which results in full-chip write power of ~ 3.2 µW/Mbps at ×64. This is the lowest eNVM write power reported at a full-chip level and about three orders of magnitude smaller than that of eFLASH. The 0.5 Mbit high-density bitcell array also demonstrates good R_p distribution and 100 % STT switching. Our results demonstrate superior power-area-feature attributes of perpendicular STT-MRAM as a best-in-class unified eNVM solution for Internet-of-Things (IOT) applications at 40 nm as well as the scalability of these advantages to 28 nm and beyond.

Introduction

The Moore's law scaling has brought down the cost of digital information processing to such low level that, in the near future, a huge number of small, intelligent, and connected devices (IOT) will power the next wave of technology. However, a conventional hierarchical memory sub-system is too cumbersome for IOT devices, and becomes a bottleneck for the minimization of system cost and power (Fig. 1). Gaps in access speed and data granularity in the hierarchy can be several orders of magnitude and make overall hardware (HW) and software (SW) architectures complex and costly. In addition, dissimilar technologies used for different memory tiers and different operating voltage levels make heterogeneous integration with CMOS logic circuitry challenging. This adds manufacturing cost and becomes a barrier to power reduction and miniaturization.

STT-MRAM has been proposed as a unified embedded non-volatile memory for IOT applications [1,4] that can simultaneously serve as a data memory, a code storage, a one-time-programmable (OTP) memory [5,6], and a non-volatile retention flip-flop [5]. Embedded MRAM also promises energy savings through zero standby power, no refresh, reduced I/O traffic, and true random access [1-5]. In this paper, we report the first fully-functional (with zero bit error) embedded MRAM macro at the 40 nm logic node. Our results confirm many of the potentials of embedded MRAM.

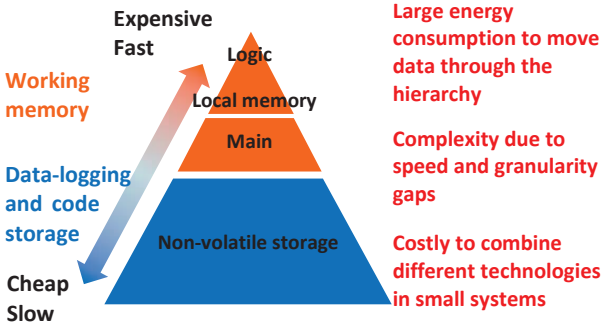

Figure 1: Conventional memory hierarchy combines several technologies to meet overall performance and capacity targets. This is not energy- and cost-effective for small systems such as most IOT devices.

Results

Table 1 describes the key design features of the embedded STT-MRAM macro of this work. This macro design (Fig. 2) is an advanced version of our previously reported macro [2]. We have also incorporated several 0.5 Mbit bitcell arrays that enable direct measurements of bitcell attributes and parasitic resistance for statistical characterization of devices and arrays (Fig. 3).

Perpendicular MTJ (pMTJ) devices [3] were inserted between the 4th and the 5th metal levels (Fig. 4) using the MTJ and BEOL tools at TDK-Headway. A cross-section TEM across the array is shown in Fig. 5. The completed wafers were diced and packaged by wire bonding without any special modification.

Table 1: Description of the 40 nm embedded STT-MRAM macro of this work.

CMOS Base Process	Generic 40LP with 6 metal levels
Density	1 Mbit
MTJ	Perpendicular MTJ
Read Access Time	20 ns
Write Cycle Time	20 – 100 ns
Clock Frequency	50 MHz
IO Width	×32 / ×64
ECC	SEC-DED
Redundancy	Rows & columns (not activated for this work)
Power Supply (Core/IO)	1.2 / 1.8V

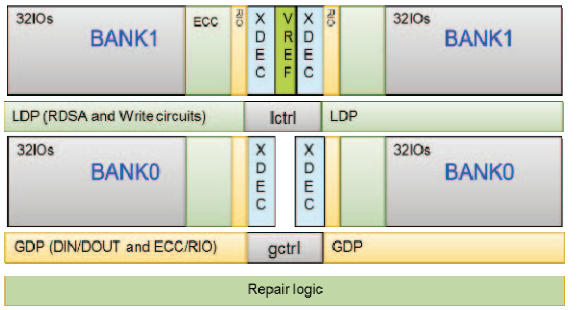

Figure 2: A block diagram of the 1 Mb macro. This design further advances the prior architecture [2] of shared sense amplifiers and merged references.

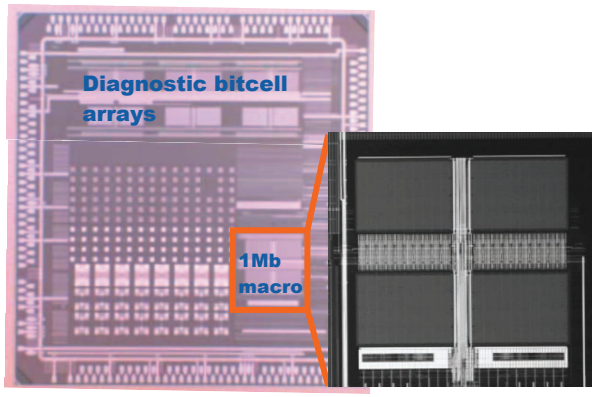

Figure 3: A die photo shows the 1 Mb macro, diagnostic bitcell arrays, and device test keys.

TMR and the parallel state resistance R_p measured on a 0.5 Mbit bitcell array (Fig. 6) show the median TMR of 110 % and median R_p of 2.01 kΩ. The read window is about 18 sigma based on parallel and anti-parallel resistance distributions shown in Fig. 7. There is zero hard defect.

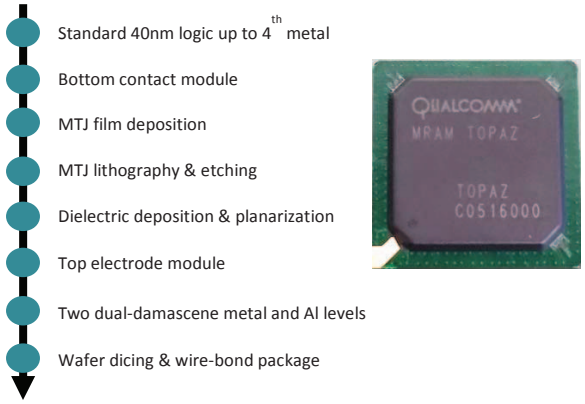

Figure 4: A simplified fabrication flow and a photo of a packaged chip.

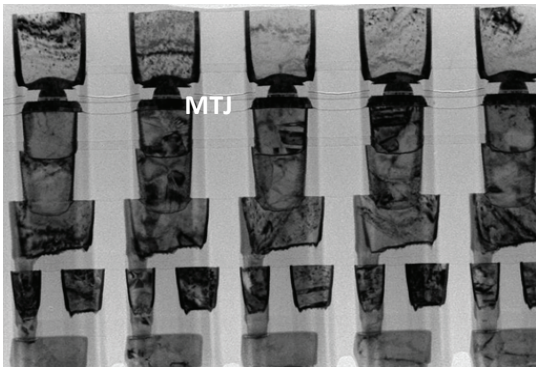

Figure 5: A TEM cross-section of the 1 Mbit macro. Perpendicular MTJs are inserted above the 4th metal level.

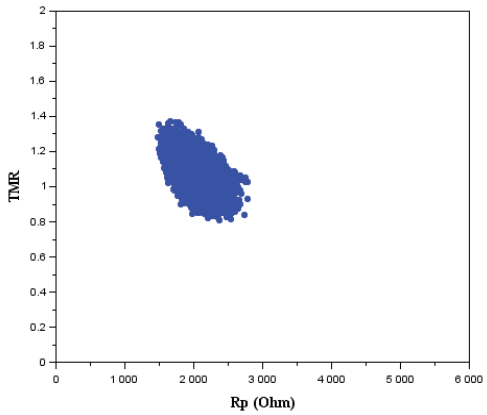

Figure 6: TMR vs. R_p of a 0.5 Mbit array. Median TMR and R_p are 110 % and 2.01 kΩ respectively. A negative TMR- R_p correlation is likely due to imperfect subtraction of parasitic series resistance. No ECC or redundancy repair was used.

The macro exceeds the read performance target (20 ns access time) with a reasonable margin (Fig. 8) over the ambient temperature range of 0 °C to 70 °C. No redundancy repair was used. The read performance can further be improved through tuning of pMTJ materials, lithography, and etch processes. Zero fail-bit-count was

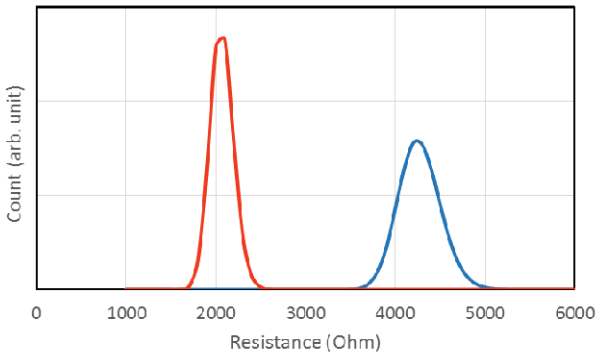

Figure 7: Resistance distributions in the parallel and the anti-parallel states show zero hard defect and ~18 sigma of read window.

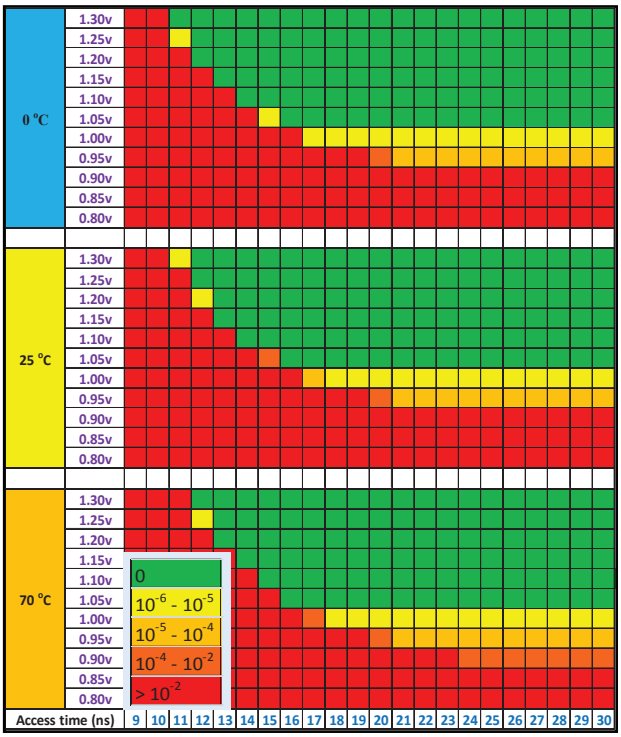

Figure 8: The 1 Mb read Shmoo shows this macro achieved 16 ns access time over temperature range of 0-70 °C with good margin. No redundancy repair was used. The inset shows color codes for error rate ranges.

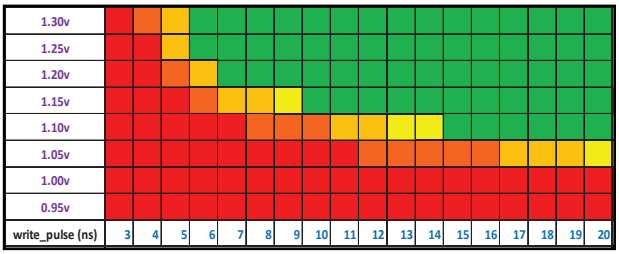

Figure 9: The 1 Mb write pulse Shmoo shows 100 % yield with a single write pulse as short as 6 ns. Full chip power consumption at this condition (50 MHz, checkerboard data, x64) was measured to be 3.2 µW/Mbps. The same color codes are used as in Fig. 8.

achieved at the write pulse width as short as 6 ns (Fig. 9), which also exceeds the write performance design target, i.e. significantly less than 20 ns access time. Such short-pulse write reduces the energy consumption since the MTJ switching current depends only weakly on the pulse duration. For the entire 1 Mb macro running continuous ×64 write (checkerboard data without ECC bits) at 50 MHz, we measured the total I_{dd} of 8.61 mA, or write power of about 3.2 µW/Mbps. In comparison, programing and reprograming in eFLASH typically take a few milliseconds and consume > 1000 µW/Mbps.

The power consumption of this macro can substantially be reduced by scaling the pMTJ diameter [7]. Further, we can extend this macro design smoothly to 28 nm node and expect the energy consumption to be reduced to < 2.0 µW/Mbps for a bitcell size of about 50 F^2 (~0.0392 μm^2) without significant MTJ materials modification or circuit revision.

The TMR vs. R_p plot for a high density bitcell (40 F^2, 0.065 μm^2) in Fig. 10 shows a tight distribution and 100 % STT switch yield, which further demonstrates the potential of high-density embedded STT-MRAM using pMTJ.

Unified low-power eNVM

Due to limitations of conventional memory technologies, present IOT systems need to have both a non-volatile storage and a working memory. In a typical IOT power cycle, the execution code must be read from the NVM either directly (execution-in-place, XIP) or indirectly (shadowing into the working memory). In addition, the temporary data being processed need to be written to and read back from the working memory (Fig. 11).

The energy consumed to read the execution code from a standalone FLASH chip and to keep at least a portion of the working memory in standby mode between power

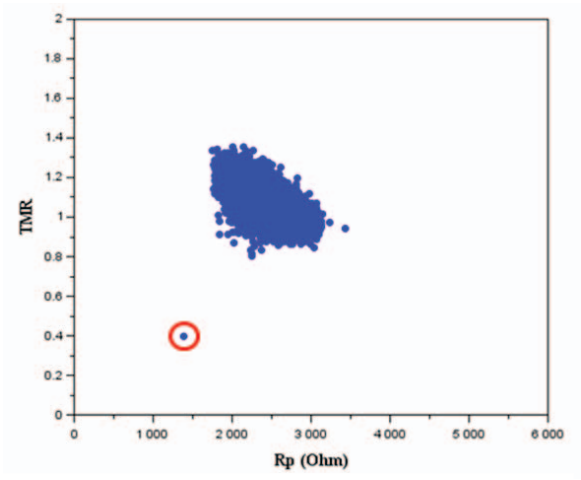

Figure 10: TMR vs. R_p plot for a high density (40 F^2, 0.065 μm^2) bitcell array shows a tight distribution and 100 % switch yield. Only one defective bit is observed. No ECC or redundancy repair was used.

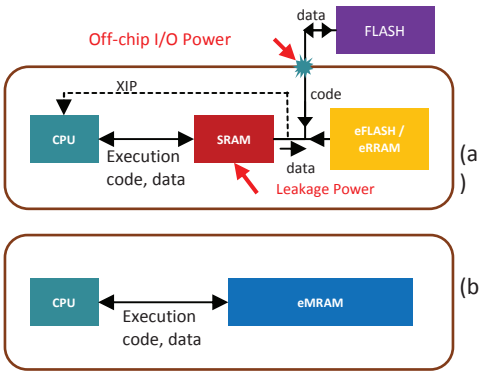

Figure 11: Illustration of (a) a conventional memory subsystem and (b) a unified memory subsystem in an IOT device. An execution code needs to be read from an NVM and some data need to be logged back to the NVM in each active period. In addition, power is wasted from leaky SRAM and from off-chip IO.

cycles can dominate the energy consumption when the duty cycle is low [1], as in the case of most IOT devices (Fig. 12). Combining NVM and working memory into a unified memory subsystem on die can improve the overall energy efficiency dramatically. This can be realized with STT-MRAM due to its high endurance (Fig. 13), high performance, and true random access.

As we have demonstrated with this embedded STT-MRAM macro, pMTJ can be seamlessly integrated into a generic CMOS logic platform without special transistor devices, a charge pump, or any modification of

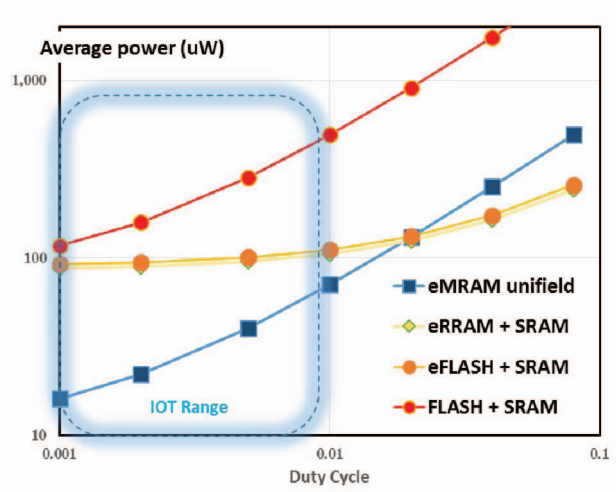

Figure 12: Average power consumption of different memory subsystem options. 50 MHz ×64 access and 4:1 read/write ratio are assumed for active cycles. Embedded STT-MRAM has clear advantage in typical IOT duty-cycle range, where the leakage from a small SRAM block sets the lower limit of average power.

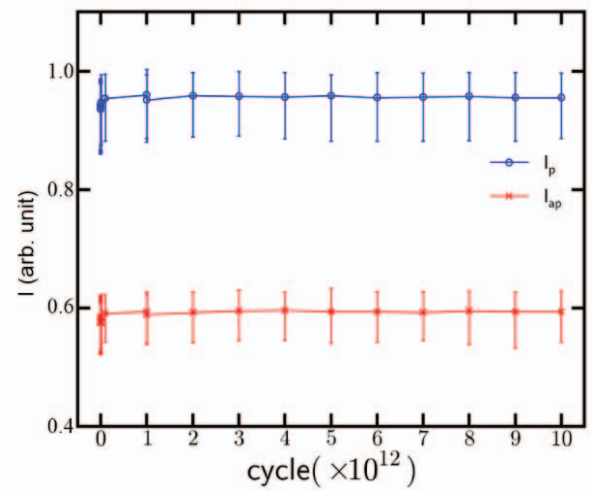

Figure 13: Read current distributions of 64 pMTJ devices through 10^{13} cycles of STT write (10 ns write pulse width) show no sign of degradation. Test was stopped due to time constraint.

conventional packaging methods. In addition, the same pMTJ device in this macro can also serve many other critical functions such as OTP and non-volatile retention flip-flop without cost adder.

Summary

The first fully functional 1 Mb embedded STT-MRAM macro in 40 nm is reported with the highest NVM energy efficiency by far. As a unified eNVM with high endurance, fast performance and true random access, eMRAM can drastically simplify HW and SW designs, improve energy efficiency, enhance form factor, and reduce costs of IOT devices simultaneously.

Acknowledgement

We thank D. Perry and M. Govindarajan for technical support, and Karim Arabi and Matt Nowak of Qualcomm Technologies, Inc. and Takuma Murai of TDK-Headway Technologies, Inc. for valuable guidance.

References

[1] K. Lee, J. J. Kan, S. H. Kang, ISLPED 2014, p.131
[2] J. P. Kim, T. Kim, W. Hao et al, Symp. VLSI Tech. 2011, p.296
[3] L. Thomas, G. Jan, J. Zhu et al, J. Appl. Phys. 115, 172615 (2014)
[4] S. H. Kang, Symp. VLSI Tech. 2014, p.36
[5] S. H. Kang, K. Lee, Acta Mater 61, p.951, 2013
[6] G. Jan, L. Thomas, S. Le et al, Symp. VLSI Tech. 2015, p.164
[7] C. Park, J. J. Kan, et al., to be published in IEDM 2015