STT-MRAM: THE NEXT MEMORY FRONTIER

IBM RESEARCH APPROACHES A TECHNOLOGY BREAKTHROUGH

SUMMARY

The information technology industry has long relied on solid-state memory tiers of SRAM, DRAM, and flash memory, each having its benefits and limitations. An alternative approach is emerging, using magnetic memory (MRAM), that could be low-power, fast, inexpensive, high-endurance, and non-volatile. Technical limitations have constrained Spin Transfer Torque (STT) MRAM adoption to applications such as ultra-dependable solid-state drive buffers. IBM Research has been developing a more advanced version of STT-MRAM that might enjoy broader adoption. The company has labored over two decades to develop this technology. IBM scientists now believe they are closing in on the remaining limitations that have relegated MRAM to niche applications.

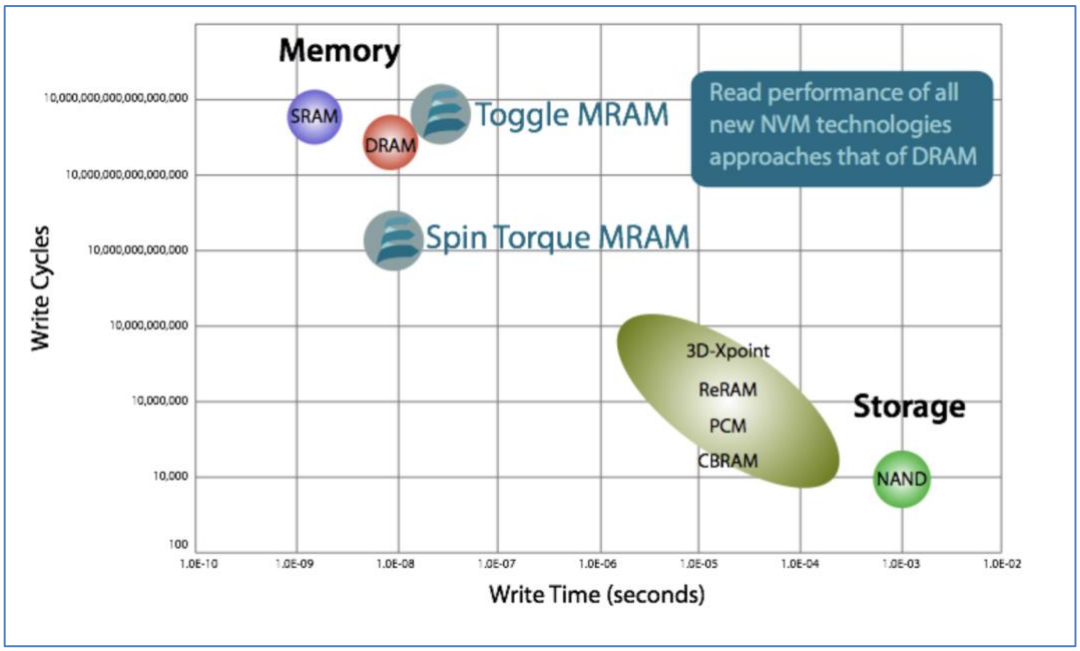

Figure 1: Memory technologies span from on-die SRAM to NAND storage. The industry seeks a memory technology that could combine speed, density, and non-volatility. Source: Everspin Technologies

This report examines the potential STT-MRAM may hold. While clearing the final hurdles will be challenging, IBM believes it is time to anticipate and prepare for the next era in memory technology.

SPIN-TRANSFER-TORQUE (STT) MRAM: MEMORY'S HOLY GRAIL?

IBM Research remains one of the industry's most vibrant research organizations, as measured by patents issued. One group in this organization has been seeking a path to a fast, inexpensive, low-power, and persistent memory technology. The team began looking into Magnetic Random-Access Memory (MRAM) over two decades ago, but that technology had significant limitations. The undertaking would require decades of hard work, ingenuity, and perhaps a little bit of luck to pull it off.

THE PROMISE OF STT-MRAM

The idea of using magnetic fields is as old as the core memory of the early IBM 360 Mainframe: set and detect a magnetic material's polarization to signify a one or a zero. However, applying these physics to a modern VLSI on-die memory device has significant challenges we will cover in the next section.

On-processor memory, primarily Static Random-Access Memory (SRAM), offers exceptional bandwidth at low latencies, providing a fast cache between DRAM and the processor cores. However, while SRAM is fast, it is not particularly dense, limiting the size of SRAM caches to hundreds of megabytes. Meanwhile, emerging applications such as AI accelerators demand more memory capacity than SRAM can offer, resorting to expensive high bandwidth memory and DRAM. STT-MRAM could double the capacity of SRAM at low power and offer non-volatility and unlimited endurance.

Industry-standard DRAM is slower but is inexpensive and available on standardized memory cards. Non-volatile flash memory is even slower but retains data without consuming power and is quite dense. There is at least an order of magnitude difference in performance and latency across each of these domains. STT-MRAM could offer the performance of DRAM and the non-volatility of flash memory.

When fully matured, STT-MRAM could replace flash memory with 10,000 times faster performance and unlimited endurance. For processor chip cache, STT-MRAM can double the memory capacity of SRAM with the added benefit of non-volatility to reduce power consumption. ASICs such as AI accelerators could increase performance with more on-chip memory for model weights and parameters. Accelerators needing larger HBM or DDR memory could also benefit from the reduced frequency of DRAM accesses a larger cache could provide.

MRAM APPLICATIONS

IBM envisions four eventual markets for STT-MRAM. The first is what most of us think of as stand-alone memory. STT-MRAM could one day even replace DRAM in applications requiring non-volatility. The second market is for embedded non-volatile

memory in chips, where Samsung is already fabricating STT-MRAM on 28-nm Silicon on Insulator (SOI) manufacturing lines. Cache memory on slower low-power processors such as used in mobile phones is the third market opportunity. The fourth and largest market opportunity is to replace some of the processors' SRAM for high-performance computing and Artificial Intelligence.

Once IBM scientists can increase the performance, the industry will begin to envision the pervasive use of STT-MRAM. IBM is excited about the opportunity of last-level cache. IBM is not yet predicting when it could deliver technology for this market. Still, the impact could be significant for memory-intensive applications such as those found in Machine Learning and Artificial Intelligence. Inference engines would suddenly be able to access far-larger memory on-die for storing neural network models, which are doubling in size every 3.5 months, according to openAI.org.

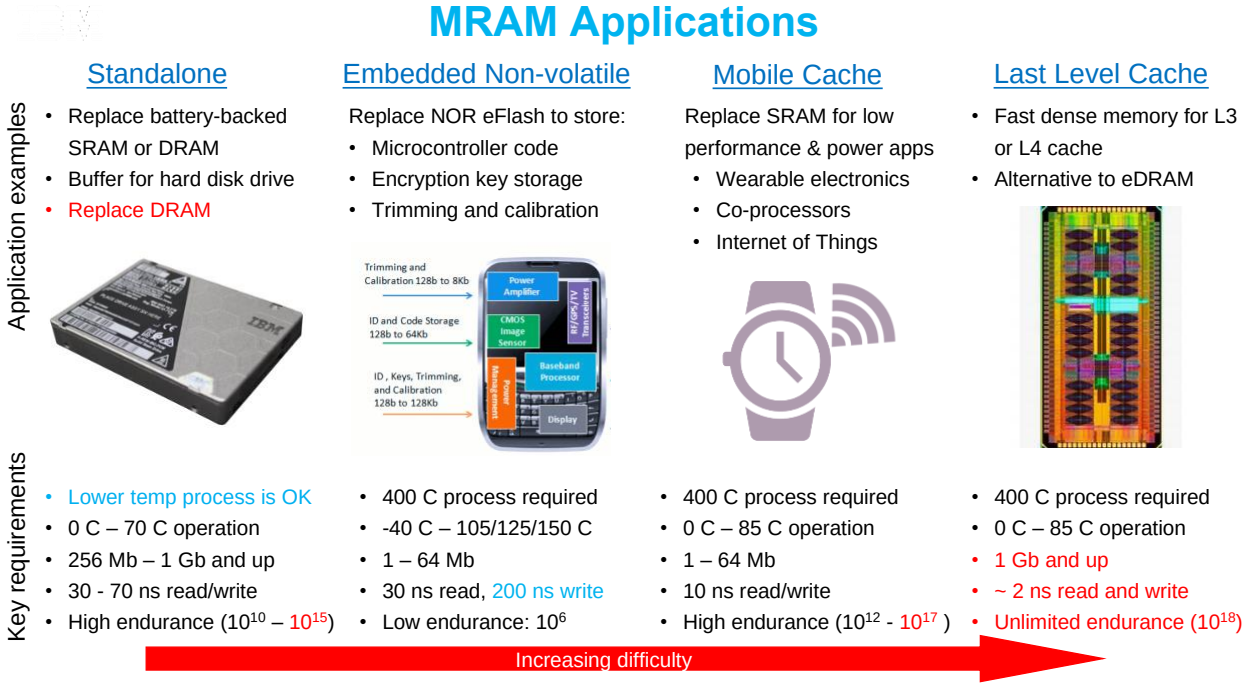

Figure 2: Use cases for STT-RAM spans a broad spectrum of applications. Text in red indicates limits that hinder the adoption of the technology. Source: IBM Research

THE CONCEPT OF STT-MRAM

The fundamental idea sounds deceptively simple: set the polarization of one layer of microscopic ferrous material by passing a current into it from an adjacent reference layer. J. C. Slonczewski hypothesized this concept in 1989, launching research into building such a device with modern VLSI technologies. IBM set out to answer some fundamental questions. How could one implement this at the scale needed to embed

such a device in a processor using advanced (e.g., 14 or 7 nm) manufacturing processes? How does one initiate the change of angular momentum, starting a wobble which would flip the polarization in nanoseconds? And how could one produce this technology reliably, at low power, and cost-effectively?

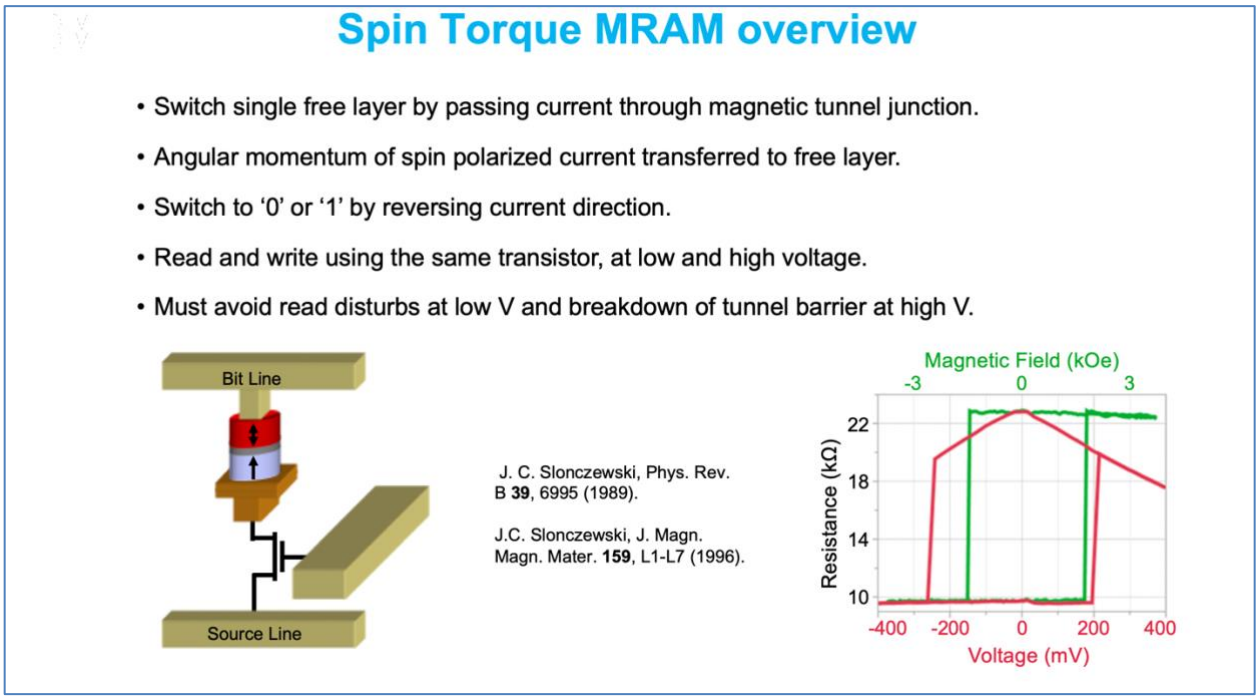

Figure 3: The concept of STT-MRAM seems quite simple, however implementing it in solid-state silicon devices has proven to be anything but simple. Source: IBM

Everspin Technologies has effectively shipped all early STT-MRAM devices into the market thus far, targeting high-end ultra-reliable storage buffers. IBM's FlashCore module uses this technology today. However, to target the larger market of last-level cache, IBM will need to improve read-write time from 30-70ns to something like 2ns. And STT-MRAM endurance would need to improve from the current 10^{10} writes to virtually unlimited data retention, or something like 10^{18} writes.

THE CHALLENGES OF STT-MRAM

Let's examine where the industry stands in the research and development process. While the concept sounds simple, the industry has faced a long journey to realize the technology fully. IBM believes it may soon clear the final hurdle, after which the technology adoption could go from a tiny niche to a mass market.

The journey down the MRAM road began when IBM invented the magnetic tunnel junction, a sandwich of two magnetic layers separated by a thin layer of insulation. When the two magnetic polarizations both point in the same direction, the resistance is

low, but when current is applied, the spin of the electrons is transferred from one magnet to the other, switching the state from a zero to a one. In 2010, Worledge (IBM) and Ohno (Tohoku U.) demonstrated the first practical perpendicular tunnel junction, which requires much lower write currents. IBM discovered new magnetic materials to make this possible, and this invention set the stage for the next decade of refinements.

Five challenges outlined below could enable on-chip cache memory and embedded flash. IBM Research had previously solved the first four challenges, most recently in 2020. Now there remains one major issue to resolve, but it is indeed not an easy one.

- ✓ The time it takes to switch states must be fast, in the 2-3 nanosecond range.
- ✓ The switching must be reliable, down to 1e-9 write error rate.
- ✓ The switching voltage distribution must be in a tight range for consistent operation.
- ✓ The fabrication process must be possible on the advanced process nodes used in microprocessors, currently in 5 or 7nm.
- ➢ The current required to switch states must be low, about ½ what is presently possible.

The most recent advancement reduced write-time to 2 nanoseconds, making STT-MRAM competitive with DRAM, with the added advantage of non-volatility. IBM has also demonstrated that STT-MRAM is manufacturable in advanced VLSI fabrication facilities.

CONCLUSIONS

Memory technology changes have slowed dramatically over the decades. Core memory was invented in 1964. Then DRAM was invented by Bob Dennard of IBM in 1966, SRAM in 1969, and Intel produced the first DRAM chip in 1970. NAND flash memory was developed in 1980. Since then, changes have been primarily spurred through VLSI manufacturing advancements, not fundamental shifts in what constitutes a memory cell. With STT-MRAM, we are finally looking at an entirely new implementation of a one and a zero. Faster, cheaper, denser, and durable non-volatility combined in a single memory design. We aren't there yet, but this future is tantalizingly near. STT-MRAM will not replace everything, at least not anytime soon. Level 1 and level 2 cache will remain implemented in SRAM, at least for now. And NAND flash memory will remain the king of the NVM hill for low-cost and density. But STT-MRAM may soon challenge existing memory devices in Level 3-4 caches, embedded flash and DRAM where non-volatility is required.

It has been a long journey for many at IBM Research, but success is finally in sight.

IMPORTANT INFORMATION ABOUT THIS PAPER

AUTHOR: Karl Freund, Founder, and Principal Analyst at Cambrian-AI Research

INQUIRIES:

Contact us if you would like to discuss this report, and Cambrian-AI Research will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts but must be cited in-context, displaying the author's name, author's title, and "Cambrian-AI Research." Non-press and non-analysts must receive prior written permission from Cambrian-AI Research for any citations.

LICENSING

DISCLOSURES

This paper was commissioned by IBM, Inc. Cambrian-AI Research provides research, analysis, advising, and consulting to many high-tech companies mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

DISCLAIMER

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Cambrian-AI Research disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Cambrian-AI Research and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Cambrian-AI Research provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.