

# Assessment of MRAM Technology Characteristics and Architectures

Rajagopalan Desikan\* Stephen W. Keckler Doug Burger

Computer Architecture and Technology Laboratory  
Department of Computer Sciences  
\*Department of Electrical and Computer Engineering  
The University of Texas at Austin  
Austin, TX 78712-1188

Department of Computer Sciences  
Tech Report TR-01-36  
The University of Texas at Austin

## ABSTRACT

High memory latency and limited memory bandwidth have proven to be one of the major hurdles restricting the performance of computer systems on high-performance computing applications. In an effort to alleviate this problem, microprocessor designers have continued to increase the amount of on-chip memory. Until recently, fast memory has been in the form of SRAM caches and performance on some applications has been enhanced by increasing cache capacities and the number of levels in the cache hierarchy. Alternate technologies, such as embedded DRAMs can enhance on-chip memory capacity at some cost of higher latency. An emerging memory technology is magnetic RAM (MRAM) which can be fabricated on top of a conventional processor core. MRAMs offer potential advantages in density and non-volatility over conventional MOSFET memory structures, and may motivate alternate memory system architectures to exploit these characteristics for higher performance and reliability. In this report, we analyze the characteristics of SRAMs, Embedded DRAMs, and MRAMs, and examine how MRAM technology may be employed in high performance computer systems.

# 1 Introduction

Until recently, the lowest levels of microprocessor memory hierarchies have been implemented using Static RAM (SRAM). Located on-chip, this memory supplies high-bandwidth and low-latency access to computation elements, and with shrinking feature sizes, larger level-1 caches and now level-2 caches are appearing on the same dice as the processors. Still, limited SRAM capacity combined with an asymmetric scaling of processing power and pin speed has exacerbated and increased latencies to off-chip memories. Furthermore, SRAM is susceptible to failures from external radiation sources and is likely to consume significant power from subthreshold leakage in future technologies. Capacity can be increased dramatically using embedded DRAM (eDRAM) technologies, but this technology does not help reduce bandwidth limitations to large capacity off-chip memories and static power consumption.

This document examines a newer magnetic RAM (MRAM) technology as an alternate or supplement to SRAM and eDRAM. MRAM has the potential advantages of small dimensions, high bandwidth, and non-volatility. Furthermore, current research is examining methods of stacking multiple layers of MRAM above a silicon substrate to achieve substantial densities in a single package. A number of MRAM memory designs have been proposed recently [12] [15] [13] [8], but it is presently unclear which design will prove the most feasible. The purpose of this document is to compare the features and scalability of MRAM memories to evolutionary SRAM and eDRAM technologies, and to present preliminary memory system concepts constructed from MRAM technologies.

## 2 Competing Memory Technologies

Advances in fabrication technology have resulted in a number of memory technologies emerging as viable alternatives to on-chip SRAM, including embedded DRAM and MRAM. In this section, we compare the various device-level and operating characteristics of these technologies in an effort to understand each technologies advantages and disadvantages. The features for comparison are taken from published technical papers. As such, an apples-to-apples comparison is impossible and the results of this section are intended to provide a qualitative, rather than a quantitative comparison of the technologies.

### 2.1 Device-level Characteristics

The device-level characteristics of memory technologies include cell area, power, and reliability. These characteristics are heavily influenced by the physical characteristics of the device and the process technology.

#### 2.1.1 Cell Area

The cell area plays an important role in determining the viability and use of a given memory technology. Historically, DRAM designs have concentrated on achieving minimum cell area, while SRAM designs have concentrated more on achieving high speed. The cell area can be specified using the process technology independent metric  $\lambda$ , which is equal to half the minimum feature size at a particular technology. The standard 6T SRAM cell occupies approximately  $550\lambda^2$  [9], while recent embedded DRAM designs have demonstrated  $80\lambda^2$  cells [14]. Several different MRAM cell designs have been proposed with the cell area ranging from  $12\lambda^2$  for the GMR MRAM architecture, to as

small as  $6\lambda^2$  for the cross point MTJ MRAM architecture [12]. However, demonstrated MRAM cell areas have been  $80\lambda^2$  [8].

### 2.1.2 Power

With shrinking feature sizes and increased transistor counts, power has become a first order design constraint. Dynamic power is increasing due to clock rate increases, while static power is increasing because of larger leakage currents resulting from lower threshold voltages. Leakage power is becoming a major challenge in SRAMs, which has led researchers to explore various strategies to reduce SRAM power consumption [7] [4]. Power consumption is directly related to performance, and high performance designs consume more power. Using Ectacti, we computed the power consumed by the 256 KB (2 Mbit) cache used in the Pentium 4 as 1 W [11]. An 8 Mbit eDRAM design dissipates 10 W at 1 GHz,  $0.175\mu\text{m}$  technology [14]. Initial MRAM designs have demonstrated 24 mW power consumption while running at 20 MHz in a  $0.6\mu\text{m}$  technology for a 256Kb part [8]. To compare these numbers, we divide the power consumption by the size and frequency of the part to obtain the the size and frequency independent metric, J/bit. These comparisons are shown in Table 1 and demonstrate that dynamic power for SRAM and eDRAM technologies is comparable, while it is potentially higher for MRAM technology.

We expect the power consumption of MRAMs to reduce as technology scales and as power/performance trade-offs are made. For example, SRAMs and eDRAMs can be designed for low power by sacrificing some performance [3] [6]. However, MRAMs face additional challenges due to the asymmetric power requirements of read and write operations. A recent study has shown the write power to be 8 times the read power [13].

### 2.1.3 Reliability

Integrated circuits are becoming increasingly susceptible to noise as feature sizes continue to shrink. Thus, reliability is becoming an important issue in both logic and memory designs. The reliability of a particular memory technology will depend greatly on the method the technology uses for storing information. SRAMs store the information using flip-flops, and a typical 6-T SRAM stores both the data bit and its complement using a feedback arrangement. Thus, it is more robust than eDRAMs which store their information in the charge of a capacitor node. MRAMs store their information using the direction of spin of electrons, and are thus less susceptible to strikes from external energized particles. However, this does make them susceptible to stray magnetic fields from other circuits.

## 2.2 Operational Characteristics

The operational characteristics of a memory system include the access times (read/write) and achievable bandwidth. These characteristics depend on the physical device characteristics, and the architectural organization of the memory technology.

### 2.2.1 Read and Write Times

The Read time is defined as the interval from when the address is placed on the address bus and when the data is available at the sense amplifiers. Read time is an important component of the latency of a given memory technology, particularly in caches. A number of low latency SRAM and eDRAM designs have been proposed recently. The load-use access latency of the 256 KB level-2

cache used in the Pentium 4 is 7 cycles at 1.5 GHz [5]. If we consider 4 cycles as time taken to access the level-2 cache array, then the read time is equal to 2.7 ns. A 3.7 ns access, 1 Gb eDRAM was demonstrated recently at 0.18  $\mu\text{m}$  technology [14]. MRAM speeds are nowhere close to this, with the cycle time being approximately 50ns in an initial 256Kb design at 0.6  $\mu\text{m}$  technology [8].

For DRAMs, architectural techniques like page mode access are used to reduce the access latency. However, page mode access results in the first access to DRAMs being much slower than the later accesses to the same row. Embedded DRAM designs also suffer from the same drawback, and this can significantly hurt the performance of applications which have irregular memory accesses. Embedded DRAMs also need to be refreshed periodically, and cannot be read when the refresh operation is in progress. MRAMs do not need to be refreshed and hence do not suffer from this drawback.

Write times usually do not vary greatly when compared to read times for SRAMs and eDRAMs. Writes for MRAMs on the other hand require more current than reads, and hence are likely to be slower. However, writes are not usually on the critical path, and architectural techniques like store buffers can often be used to hide any extra write latency.

### 2.2.2 Achievable Bandwidth

The achievable bandwidth from a given memory technology depends on the architectural organization of the memory system. Off-chip DRAMs have traditionally been organized as high latency high bandwidth devices, while SRAMs have been organized as low latency moderate bandwidth devices. As embedded DRAMs can be fabricated along with logic, they reduce the latency while providing high bandwidth. Recent studies have demonstrated eDRAMs with 128-Gbytes/sec throughput at 1 GHz [14]. The 256 KB level-2 cache in the Pentium 4 chip from Intel is implemented using SRAMs and achieves 48-Gbytes/sec throughput at 1.5 GHz [5]. Embedded DRAMs are likely to have higher bandwidth than SRAMs due to the smaller cell size, which results in more I/O pins. MRAMs have cell size similar to eDRAMs and can also take advantage of vertical I/O pins. Hence, they are likely to have a higher bandwidth than eDRAMs. However, MRAM designs are in their preliminary stages now, and it is not clear how advantageous the vertical interconnects will be in terms of providing higher bandwidth.

## 2.3 Characteristics Summary

In this section, we have compared some of the important characteristics of the state of the art in SRAM, eDRAM, and MRAM technologies. As seen from these characteristics, each memory technology has some advantages and some disadvantages over the other technologies. It is not clear at present whether one technology will dominate the others, or whether each technology will have its own unique place in the design space. Table 1 summarizes the characteristics of the three technologies evaluated in this section, using high level published data.

As mentioned earlier, it is hard compare these numbers directly due to various factors like difference in process technology, frequency, size, and architectural organization. For example, even though the MRAM cells are roughly 7 times smaller in Table 1, actual MRAM cell size is larger than SRAM cell size because of the smaller process technology used in the SRAM cell. However, we expect the MRAM cell area to scale with technology. The comparison of power also presents similar difficulties. We have tried to normalize power by defining the J/bit metric which takes into account the part size and frequency. Using this metric, we see that the MRAM power consumption is higher than SRAM and eDRAM power consumption. However, we have not taken the process

Table 1: Preliminary estimates of SRAM, eDRAM, and MRAM

	SRAM	eDRAM	MRAM
Technology	0.18 $\mu$ m	0.18 $\mu$ m	0.6 $\mu$ m
Cell Size	550 $\lambda^2$	80 $\lambda^2$	80 $\lambda^2$
Power	1 W at 800 MHz, 2048 Kb	10 W at 1 Ghz, 8 Mbit	24 mW at 20 MHz, 256 kb
Normalized Power (J/bit)	$1.3 \times 10^{-15}$	$1.2 \times 10^{-15}$	$4.6 \times 10^{-15}$
Read Time	2.7 ns	3.7 ns	50 ns
Bandwidth	40-100 Gbytes/sec	128-1000 Gbytes/sec	N.A.

technology into consideration, and we expect MRAM power consumption to reduce when it is scaled to smaller technologies.

### 3 MRAM Design Issues

Several attributes differentiate MRAM technologies from solid-state technologies such as SRAM and eDRAM. This section outlines those features that are likely to affect memory system architectures employing MRAM.

**Feasible capacity per layer:** While the raw density of MRAM bits is a function of the minimum magnetic flux, the number of useful bits per layer of MRAM is much more complex. First, some significant fraction of the wiring layers must be reserved for power distribution in the MRAM layers. Second, routing wires must be available to connect the MRAM bits to sense amplifiers located in the active layers. Third, if MRAM layers are to be stacked, then space devoted to vias must be allocated to connect from the upper layers to the active layer. Finally, while some off-chip connections could be made from the periphery of the chip, achieving substantial off-chip bandwidth may require area bonding. Thus, vias may be required to connect the active layers to package pins and the outside world. Although many of these interconnect wires will likely be on different metal layers than the active magnetic bit and word lines, the effective density of the MRAM layers will be determined in part by how the MRAM bits can be connected into the network.

Another potentially fruitful research area is in methods of stacking multiple active silicon layers in a 3-D structure. This could enhance the utility of dense MRAM structures as the sensing circuits for each MRAM layer could be placed nearby. However, the vias for communicating between the layers as well as distributing power and off-chip bandwidth may still limit some of the useful area in the stack. Current areas of research include vertical bus structures [10] and 3-D SOI [18].

**Non-uniform access times:** Physical proximity in future systems will have a substantial impact on communication latency. For example, on a conventional silicon integrated circuit with projected clock rates and wire dimensions, the latency across the diameter of a chip in a 35nm technology is likely to exceed 30 cycles [1]. This has serious implications for on-chip memory systems; current work is examining non-uniform cache architectures (NUCA) to address the capacity and latency trade-offs [2]. MRAM memory systems are likely to be subject to the same constraints. Data located in MRAM layers close to the surface of the active silicon circuits is likely to be accessed faster than layers higher up in the stack if the vertical interconnects introduce appreciable delay. MRAM access latency must also be incurred if the lateral distance between the data and the

location in the active region where it is needed is large. In essence, the two-dimensional routes in today's integrated circuits become three-dimensional routes with MRAM stacks. It is conceivable that average distances between sources and sinks of data would decrease when migrating from a 2-D to a 3-D topology [18].

**Bandwidth:** Two components of bandwidth are important in MRAM memory systems. First is the bandwidth to the processing elements in the active silicon substrate. This bandwidth is a function of horizontal and vertical wire dimensions (cross-sectional area) and the speed at which the MRAM cells and wires can be switched. If vertical wire dimensions from the MRAM layers can scale with the horizontal wire dimensions found in the active region, then the bisection bandwidth between the memory and the processors could exceed that of conventional 2-dimensional topologies. The second bandwidth component is the connection between MRAM and the outside world. This bandwidth will be significantly smaller than the internal bandwidth as external communication channels will likely employ transceivers implemented in the active region. Thus incoming data would go to the active region and then up into the MRAM. The bandwidth is then limited by the speed and bisectional area of the connections that can be made to the active region.

**Asymmetric reads and writes:** Because writes to the MRAM arrays require changing the magnetic polarity, power consumption and delay are likely to be much larger than for reads. Scheuerlein et. al report write power to be 8 times read power [13]. While writes are typically much less frequent than reads, particularly since caches filter both reads and writes from deeper regions of the memory hierarchy, this large asymmetry between read and write cost will likely affect the memory hierarchy design. Multiple writes might be buffered and aggregated into a single write to conserve power and delay. This approach is similar in nature to proposed non-volatile memory system architectures, such as eNVy [16].

**Non-volatility:** Using magnetic rather than electrical storage provides a number of potential benefits. Magnetic storage may be less vulnerable to atmospheric radiation sources (such as ionized neutrons and alpha particles) but its susceptibility to other noise sources, such as stray electrically induced magnetic fields from the active layers, is unknown. Clearly, magnetic non-volatility has uses for storage in mobile devices. However, it may also be useful in high performance systems that are never powered down because power consumption is reduced when the memory system is not in use. In contrast, leakage current force DRAM memories to be refreshed and SRAMs and to dissipate power even when not being accessed. Leakage currents in both of these technologies are likely to increase with shrinking features sizes [4].

**Scalability:** Transistor devices are expected to continue to scale at an aggressive rate, with feature sizes shrinking by a factor of four over the next 12 years. This will likely realize at least an order of magnitude in density of SRAM and DRAM memory arrays. MRAM is in its infancy and bit area in current prototypes is approximately  $7.1\mu\text{m}^2$  at  $0.6\mu\text{m}$  technology [8], which is 1.7 times larger than an SRAM bit in today's  $0.18\mu\text{m}$  technology [9]. If the areal density in the disk drive industry can serve as a guide, MRAM bits should shrink substantially, and wiring will likely become the limiting factor.

**Sensing:** MRAMs use the resistance of the MTJ cell for information storage. The information is retrieved by sensing the current through the MTJ cell for the two different resistance values. For

stable MRAM operation, a magneto-resistance (MR) ratio of 40% or higher is generally needed, although some recent sensing schemes work well with a lower ratio [17]. With a higher MR ratio, the sense amplifiers can be placed farther away from the cells. With a low MR ratio, if the sense amplifiers are placed far away from the MRAM cell, the signal transition might be too weak to detect. This fact should be kept in mind while designing MRAM banks, as reading large banks may prove challenging.

## 4 Possible MRAM Applications

The characteristics of MRAM that separate it from other technologies include density, bandwidth, asymmetric access times, and power. In this section, we suggest several applications for MRAM technologies for high performance computer systems and describe at a high level how the attributes of MRAM may motivate different memory architectures and organizations.

### 4.1 High Bandwidth Caches

A conventional architecture may employ MRAM as an independent level in the cache hierarchy. This technology assessment indicates that MRAM cannot deliver the speed or bandwidth of a small SRAM array and thus is not suitable for a level-1 cache. However, the capacity and latency is likely amenable to a level-2 or perhaps a level-3 cache. In a chip-multiprocessor architecture, the MRAM stack could be partitioned so that the layers above one of the processors and its level-1 cache would constitute that processor's principal level-2 cache. This organization would enhance physical locality for each processor to its memory system and decrease communication distances between the processors, thus reducing inter-processor communication latency. It also remains to be seen if a single memory controller will be enough to orchestrate accesses to multiple MRAM banks. If multiple controllers are required, designs must consider the space and time overhead associated with them.

A second factor when organizing an MRAM stack as a cache is that distant layers will require longer access times than nearer layers. This non-uniform access time motivates an architecture that attempts to place most frequently accessed data nearest to the processor and less frequently accessed data (but still data that is useful to cache) in layers further from the processor. We are currently examining non-uniform cache architectures for a two-dimensional silicon substrate and policies for data placement and migration. The results from these experiments can be applied to MRAM architectures. SRAM row buffers in the active layer can be employed to cache data from the MRAM banks in the higher layers.

### 4.2 MRAM Main Memory

With on-chip memory increasingly occupying more real estate on a processor die, MRAMs might completely replace DRAM banks as main memory in a computer system. If multiple layers of MRAM can provide enough capacity, then external DRAM banks can be completely eliminated, with users buying MRAM processors with more memory when the need arises.

Even if we do not completely eliminate external DRAM banks, the properties of MRAM can be leveraged in some other interesting ways. As an alternative to automatic management of data through caching, MRAM could be effectively used by the operating system or application as a software controlled cache. For example, the capacity of MRAM may allow the operating system to migrate hot pages of data from disk or main memory into MRAM storage. An operating system

could also map a subsection of the file system to non-volatile storage to speed reboot or accelerate applications with frequent disk access. Either of these techniques may be applicable to programs with enormous data sets and to enhance out-of-core computational algorithms. Alternately, an application may employ MRAM storage as separate from the cache hierarchy, perhaps as high bandwidth storage to selected parallel data structures. Due to the high cost of writes to MRAM, the operating system can elect to store only clean pages in MRAM and store the dirty pages in DRAM instead.

### 4.3 Data Structures for Reliability

The non-volatile properties of MRAM may enable its use in improving reliability through more efficient logging and recovery. Today's systems maintain stable update logs (if they support logging) on disks and require substantial disk access to recover from a crash. With high bandwidth access to stable storage, logging can be more frequent and efficient. Furthermore, recovery times will decrease due to the speed and bandwidth available to restore state from MRAM-based logs and checkpoints.

## 5 Conclusion

In this document we have compared the key technological features of MRAM to conventional SRAM and embedded DRAM, and articulated distinct issues in the design of MRAM based systems. We believe that the most important challenges are as follows:

- While MRAM compares favorably in term of anticipated bit density to SRAM and eDRAM, interconnects between the magnetic cells may reduce the inherent benefits of density.
- While the cross point MTJ MRAM architecture requires only one diode per cell [12], the sensing circuits will require more active devices like MOSFETs. The speed and bandwidth of MRAM memory will depend on where the sensing circuits can be placed and the wire latencies between the MRAM bits and these circuits.
- The power and delay associated with writing into the MRAM array will likely prevent it from being used as a purely random access memory.
- The non-volatility provides a unique opportunity for scientific computations, perhaps to help reduce standby power consumption and power currently spent on disk accesses.
- Novel circuit designs will be required to exploit MRAM memory organizations.

While MRAM could become another layer in a traditional memory hierarchy, its bandwidth and volatility attributes suggest that it may be used in news ways for high-performance and reliable computer systems. We believe that there are fruitful opportunities for using MRAM as a high volume, high bandwidth memory for local and persistent data, as a staging area for persistent data in out-of-core computations, and as a storage for persistent logs to enhance recovery times and reduce logging overheads. Examination of these and other techniques is ongoing.



## References

- [1] V. Agarwal, M. S. Hrishikesh, S. W. Keckler, and D. Burger. Clock rate versus IPC: The end of the road for conventional microarchitectures. In *Proceedings of the 27th Annual International Symposium on Computer Architecture*, pages 248–259, June 2000.
- [2] V. Agarwal, C. Kim, S. W. Keckler, and D. Burger. The effect of system-level interconnects on microarchitecture design and simulation. Submitted to *IEEE Transactions on VLSI Systems*, August 2001.
- [3] F. Hamzaoglu et al. Dual- $V_t$  SRAM cells with full-swing single ended bit line sensing for high-performance on-chip cache in 0.13 $\mu$ m technology generation. In *Proceedings of the 2000 International Symposium on Low Power Electronics and Design, ISLPED 2000*, pages 15–19, Jul 2000.
- [4] H. Hanson, M. S. Hrishikesh, V. Agarwal, S. W. Keckler, and D. Burger. Static energy reduction techniques for microprocessor caches. In *2001 International Conference on Computer Design*, pages 276–283, Sep 2001.
- [5] G. Hinton, D. Sager, M. Upton, D. Boggs, D. Carmean, A. Kyker, and P. Roussel. The microarchitecture of the pentium 4 processor. *Intel Technology Journal*, 1, February 2001.
- [6] K. Itoh, T. Watanabe, S. Kimura, and T. Sakata. Reviews and prospects of high-density RAM technology. In *Semiconductor Conference, 2000. CAS 2000 Proceedings. International*, volume 1, pages 13–22, June 2000.
- [7] S. Kaxiras, Z. Hu, and M. Martonosi. Cache decay: Exploiting generational behavior to reduce cache leakage power. In *Proceedings of the 28th Annual International Symposium on Computer Architecture, ISCA 2001*, pages 240–251, Jun 2001.
- [8] P. K. Naji, M. Durlam, S. Tehrani, J. Calder, and M. F. DeHerrera. A 256kb 3.0v 1T1MTJ nonvolatile magnetoresistive RAM. In *2001 IEEE International Solid-State Circuits Conference*, pages 122–123, 438, Feb 2001.
- [9] H. Pilo et al. An 833mhz 1.5w 18mb CMOS SRAM with 1.67Gb/s/pin. In *2000 IEEE International Solid-State Circuits Conference*, pages 266–267, Feb 2000.
- [10] P. Ramm, D. Bonfert, H. Gieser, J. Haufe, F. Iberl, A. Klumpp, A. Kux, and R. Wieland. Interchip via technology for vertical system integration. In *Proceedings of the International Interconnect Technology Conference*, pages 160 – 162, June 2001.
- [11] G. Reinman and N. Jouppi. Extensions to CACTI, 1999. Unpublished document.
- [12] R. E. Scheuerlein. Magneto-resistive IC memory limitations and architecture implications. In *Proceedings of the International NonVolatile Memory Technology Conference*, pages 47–50, May 1998.
- [13] R. E. Scheuerlein, W. Gallagher, S. Parkin, A. Lee, S. Ray, R. Robertazzi, and W. Reohr. A 10ns read and write non-volatile memory array using a magnetic tunnel junction and FET switch in each cell. In *2000 IEEE International Solid-State Circuits Conference*, pages 128–129, Feb 2000.
- [14] O. Takahashi et al. 1-GHz fully pipelined 3.7-ns address access time 8k x 1024 embedded synchronous DRAM macro. *International Journal of Solid-State Circuits*, 35:1673–1679, Nov 2000.
- [15] S. Tehrani, B. Engel, J. M. Slaughter, E. Chen, M. DeHerrera, M. Durlam, P. Naji, R. Whig, J. Jenesky, and J. Calder. Recent developments in magnetic tunnel junction MRAM. In *IEEE Transactions on Magnetics*, volume 36, pages 2752–2757, Sep 2000.
- [16] M. Wu and W. Zwaenepoel. eNVy: A non-volatile, main memory storage system. In *Proceedings of the 6th Symposium on Architectural Support for Programming Languages and Operating Systems*, pages 86–97, Oct 1994.
- [17] K. Yamada, N. Sakai, Y. Ishizuka, and K. Mameno. A novel sensing scheme for a MRAM with a 5% MR ratio. In *2001 Symposium on VLSI Circuits Digest of Technical Papers*, pages 123–124, Mar 2001.

- [18] R. Zhang, K. Roy, C.-K. Koh, and D. B. Jones. Exploring SOI device structures and interconnect architectures for 3-dimensional integration. In *Proceedings of the Design Automation Conference*, pages 846 – 851, Jun 2001.