

# Identifying Optimal Non-Volatile Semiconductor Memory for Use in RAID Systems

Barry Hoberman, Business Development, Crocus Technology  
Steve Ciadakis, Business Development, Crocus Technology & Silicon Impact

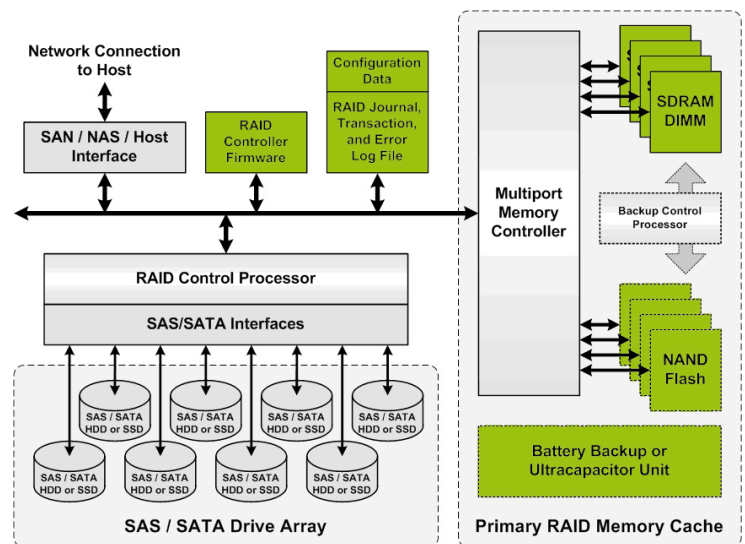
**R**edundant array of inexpensive disk (RAID) systems increase server performance and protect against data loss by exploiting disk-level parallelism. Without RAID servers, there would essentially be no commercial Internet. Because they hold mission-critical data and must provide fault-tolerant storage, enterprise customers cannot tolerate any potential for RAID server data loss in a world of less-than-perfect electrical power sources. Consequently, RAID system designers have tried using many non-volatile semiconductor memories, including NAND Flash, non-volatile static random access memory (NVS RAM), ferroelectric RAM (FeRAM or FRAM) and magnetoresistive RAM (MRAM), to retain data in the event of a power failure. Of these, MRAM comes closest to the ideal memory for RAID system server design.

The commercial behemoth known as the World Wide Web sits layered on an intricate, distributed data network called the Internet. System architects build and extend the Internet's very foundation using storage blocks called hard disk drives (HDDs). Shelves full of HDDs organized into RAID systems inhabit countless racks in worldwide data centers. The improved performance and fault tolerance of RAID systems explain their universal acceptance as the local and networked storage medium of choice for all types of Internet servers, other on-line transaction processing (OLTP) servers, and many other server types installed in large and small data centers worldwide. According to estimates and forecasts, several million of these servers ship each year.

RAID systems use sophisticated, fault-tolerant methods of data storage, including data mirroring and striping, to distribute data across multiple drives and thus protect data against loss. These methods make automated recovery possible when an individual HDD fails. Real-time data recovery, possible with more advanced hardware RAID systems, ensures that important and mission-critical data remains safe even after catastrophic hardware failure. Although a RAID system's disk drives provide much of the non-volatile data storage needed to provide fault-tolerant operation and data recovery ability, there is a real need for non-volatile semiconductor memory within the design of the RAID controller. Without non-volatile chip memory, a RAID system cannot protect data against certain types of failures such as power loss.

Figure 1 shows a block diagram of a RAID system. The RAID server's heart is the RAID control processor, which manages the attached drive array through a bank of industry-standard serial attached SCSI (SAS) and serial ATA (SATA) drive interfaces. Although largely based on HDDs, RAID systems increasingly include one or more solid-state drives (SSDs), which themselves are based on arrays of NAND Flash semiconductor memories. SSDs provide roughly 10x faster write performance and 100x faster read performance than HDDs, but they cost substantially more per gigabyte (GB) of storage. Hybrid RAID systems that combine HDD and SSD storage currently offer the best available mix of performance and capacity. Note that Figure 1 shows a superset of all possible methods used to provide non-volatile storage in the server (shown as colored boxes). Practical RAID designs use some but not all of these methods.

Figure 1. Block Diagram of a RAID Server



The RAID control processor currently requires a mix of semiconductor memory, including DRAM and non-volatile semiconductor memory. Local non-volatile memory usually serves as a repository for the hardware RAID server's firmware and as

non-volatile storage of configuration data and of a journal/transaction/ error log file. The right side of Figure 1 shows a large box labeled “Primary RAID Memory Cache.” This cache speeds disk write transactions from the host server’s perspective. The RAID controller can quickly stash write transactions in fast memory cache and then signal transaction completion to the host. Then, the RAID controller moves the transaction data from the primary cache into the disk array, which is a relatively slow process compared to saving the transaction in the cache.

A large DRAM bank serves as the primary cache in most RAID servers because DRAM currently provides the best available combination of fast write time and low cost/bit. Low per-bit cost is important because RAID memory caches are sometimes as large as 32GB. Today, this DRAM is most likely to be double data rate, second generation synchronous dynamic random access memory (DDR2 SDRAM), which will quickly transition to DDR3 SDRAM this year as sales volumes and semiconductor memory economics start to favor DDR3 SDRAM over DDR2.

However, DRAM has a severe liability when used as RAID memory cache: DRAM provides only volatile storage. If power is lost, so is the data stored in the DRAM. Because RAID systems hold mission-critical data and must provide fault-tolerant storage, enterprise customers cannot tolerate this potential for data loss in a world of less-than-perfect electrical power sources. Consequently, RAID designers employ one of several methods to add non-volatile storage to DRAM-based primary caches.

The first such method is to simply add a battery and power controller to maintain power to the primary cache when power mains fail. However, most battery systems used in RAID applications are rated for no more than 72 hours of unpowered operation. After that, data may be lost. Batteries also require maintenance. RAID back-up batteries should be replaced annually, which is both an extra expense and a potentially serious operational problem. It’s not uncommon for data center managers to be blissfully unaware that their RAID servers contain deeply embedded batteries. Consequently, many RAID back-up batteries are not serviced regularly, and mission-critical data is at risk.

Figure 1 shows an alternative design approach—adding NAND Flash memories to the primary cache—which also provides non-volatile storage for the RAID system’s primary cache. When the RAID control processor detects a loss of main power, an inexpensive back-up control processor in the primary cache independently copies the contents of the cache’s DRAM to the NAND Flash array. NAND Flash is generally rated to safely hold the data for 10 years without power. Battery power is only required for a short period while the data is copied from DRAM to NAND Flash. Some designs dispense with the battery and the associated maintenance requirements and instead use low-maintenance ultra capacitors, which provide the needed power for the short back-up interval. Using Flash as a back-up memory layer in this configuration adds the cost of the Flash memory itself, as well as the supporting back-up circuitry and hardware, to the RAID system.

Two key characteristics prevent NAND Flash from being used as the sole memory in primary RAID caches. First, NAND Flash devices have relatively long write latencies due to their long erase-write cycles. Second, NAND Flash devices deteriorate in direct proportion to the number of erase-write cycles they endure. Most NAND Flash devices are rated for only 100,000 or so erase-write cycles before the serious onset of memory cell failures. Wear-leveling techniques remediate NAND Flash wearout failures in SSD applications, but these techniques are too slow to apply to a primary RAID cache, which requires data throughput rates that are orders of magnitude faster. These two traits greatly reduce the attractiveness of NAND Flash for direct primary cache storage. For these reasons, NAND Flash can serve as a DRAM back-up in the primary RAID cache but cannot serve as the primary cache’s main memory alone. There’s a significant opportunity to replace DRAM in the primary RAID cache if a cost-competitive, non-volatile semiconductor memory with DRAM’s write speed and without NAND Flash’s write endurance problem becomes available.

Beyond the primary RAID memory cache, two other places in the RAID server block diagram require non-volatile memory—for firmware storage and for the journal/transaction/error log file. Use of non-volatile memory such as read-only memory (ROM), electronically programmable ROM (EPROM) and NOR Flash for firmware storage is pervasive in most embedded systems, including RAID servers. However, the log file is unique to storage applications.

Journaling file systems employ techniques from transaction processing database systems to maintain the structural consistency of the data stored in the RAID array by logging atomic disk input/output (I/O) transactions. Should a failure occur such as the loss of a drive in the disk array, replaying the transaction log from the last file system checkpoint restores the RAID system’s state. Depending on the checkpoint frequency, the journal/transaction/error log file need not be nearly as large as the primary RAID memory cache. A few megabytes of storage are generally sufficient for both the journal/transaction/error log file and the RAID system’s configuration data. For obvious reasons, the memory that holds these files must be non-volatile.

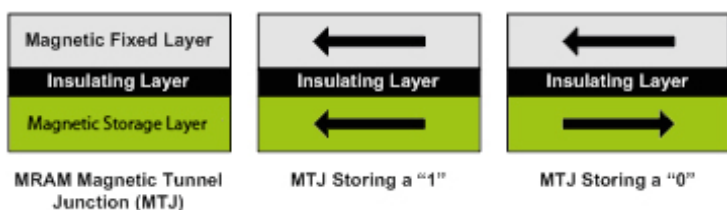
Several semiconductor memory technologies vie for this socket: NVSRAM pairs a six-transistor (6T) static RAM cell with a silicon-oxide-nitride-oxide-silicon (SONOS) electrically erasable programmable ROM (EEPROM) cell, replicating the SDRAM/NAND Flash pairing previously described but at the cell level. The result is a fast SRAM array that can be backed up in one write cycle. NVSRAM is currently the technology of choice for the non-volatile memory in RAID systems (excluding the primary memory cache). However, the NVSRAM memory cell is more than twice as large as a 6T SRAM cell, which itself is relatively large compared to DRAM or NAND Flash memory. Consequently, NVSRAM storage is relatively expensive on a cost/bit basis and will likely stay that way relative to other memory technologies.

FRAM inserts a ferroelectric material, typically lead zirconate titanate (PZT), into the semiconductor processing flow. The ferroelectric material creates a bistable bit storage element that operates at the molecular level. An electric field sets the physical position of one central atom trapped in a tetrahedron of oxygen atoms. The central atom's position represents the stored bit state within the PZT ferroelectric molecule. Commercial FRAMs have been available for at least two decades, but like NAND Flash memories, FRAMs also exhibit wearout failure. More importantly, FRAM capacities remain small, and lithographic scaling may become a severe problem because ferroelectric materials tend to lose their ferroelectric properties when the amount of ferroelectric material used drops below a threshold value.

Phase-change memory (PCM) first appeared as a cover story in Electronics magazine nearly 40 years ago, but very few commercial PCM devices have been introduced so far. PCM stores bits as physical state changes in a chalcogenide glass that can take either a crystalline or amorphous form. In crystalline form, chalcogenide glass is a good electrical conductor. In the amorphous form, it's not. The conductivity difference produces a usable memory cell. Chalcogenide glass is the active material used for making recordable CDs and DVDs, so its crystalline/amorphous properties are very well understood by now. However, writing to a PCM cell literally involves melting and annealing glass, so PCM write cycles aren't particularly fast (approximately 100-200 nanoseconds at current lithographies), and PCM storage retention drops quickly as the operating temperature rises. Retention time for one vendor's prototype PCM cells is on the order of 10 years at 85°C but only 10 seconds at 165°C and 10 microseconds at 225°C. Similar to NAND Flash and FRAM, cycling stresses cause wearout failure in PCMs which have endurance ratings of approximately  $10^6$ - $10^9$  write cycles.

MRAM stores data in magnetic material introduced into the semiconductor cell. Its advantages are density, speed, symmetrical read and write cycle times, and infinite write endurance. The MRAM storage element, called a magnetic tunnel junction (MTJ), consists of a sandwich of one fixed (or "pinned") magnetic layer and one switchable magnetic layer separated by an insulating layer to form a tunnel junction, as shown in Figure 2. Write currents switch the magnetic orientation of the switchable layer inducing a measurably different junction resistance, depending on whether the magnetic polarities of the fixed and switchable layers are aligned or opposed. The difference in resistance provides the readout of the cell's state.

**Figure 2. MRAM Cell Design**



Many startup companies and established memory IC vendors, including IBM, have invested heavily in developing MRAM process technologies that are compatible with semiconductor manufacturing. The only company with commercially available RAM products so far appears to be Everspin Technologies, which was spun out of Freescale in 2008.

Crocus Technology, an outgrowth of the CNRS and CEA national labs in Grenoble, France, has developed an improved MRAM cell design and process technology which is based on what the company calls thermal-assisted switching (TAS). The Crocus TAS MRAM cell uses thermal heating to raise the temperature of the MTJ during the write cycle. The heat softens the cell's hardness against magnetic polarity change, permitting the use of a single write line and reduced write currents. Once cooled, the Crocus MRAM cell regains its magnetic hardness. Crocus Technology's TAS innovation for MRAM improves cell retention stability by allowing Crocus' MRAM cell to use a "harder" magnetic layer that's more magnetically stable than non-TAS MRAM cell designs without making the writing currents and other characteristics unattractive.

This difference will prove to be a critical development as MRAM manufacturing lithographies scale below 65nm for two reasons:

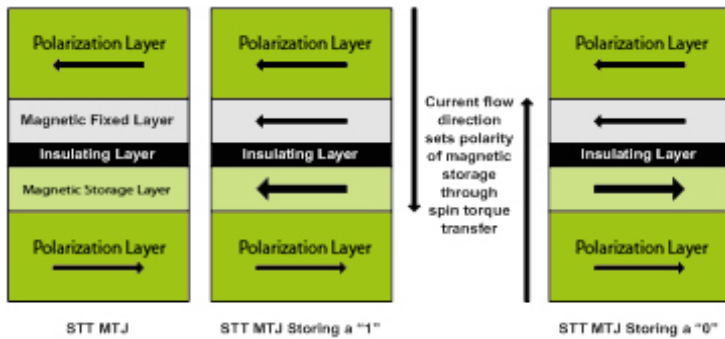
1. The currents needed to switch an MTJ are not insignificant. Magnetic switching in other vendors' non-TAS MRAM cell occurs where two energized wires intersect and when nearby conductors supply enough combined magnetic flux to switch the magnetic material's polarity. This mechanism looks a lot like the one used for magnetic-core memory in the middle of the 20th century and requires a fair amount of current to induce magnetic switching. Crocus Technology's patented TAS switching requires only one energized wire to supply the switching flux. This use of a single field-generating wire more than halves the necessary write power. At the same time, the currents necessary to heat the TAS memory bit is quite small.

2. The stability ("hardness") of the magnetic storage, or in other words, the MTJ's resistance to inadvertent switching, faces scaling challenges at feature sizes of 65nm and smaller. This obstacle creates real storage-reliability problems as the parts are scaled to densities approaching that of today's NAND Flash devices. The onset of MRAM stability loss with shrinking cell size is particularly fast below 65nm. Such MRAMs are at jeopardy of needing little or no error correction (ECC) above 65nm, to needing a lot of redundant and expensive ECC bits as device feature size shrinks below 65nm. With its hardened magnetic storage layer, Crocus Technology's TAS MRAM design offers orders of magnitude greater stability compared to competing MRAM designs.

In mid-2009, Tower Semiconductor and Crocus announced an agreement to add a Crocus Technology TAS MRAM process module to Tower's CMOS manufacturing capability. In 2011, Crocus' MRAM technology will be ready for manufacture at Tower in both discrete memory chips as well as in embedded designs.

Several memory vendors including Crocus Technology are currently developing a new and different approach to MRAM technology, dubbed spin torque transfer (STT). The STT memory cell employs special layers that uniformly polarize the spin of the electrons flowing through the MTJ, and the spin-polarized current imparts magnetic moment to the storage layer depending on the direction of electron flow through the STT memory cell, as shown in Figure 3. STT will especially be effective for MRAM at the 65nm and smaller lithography nodes.

**Figure 3. Spin Torque Transfer (STT) MTJ Memory Cell**



Significantly, an STT memory cell’s write current shrinks with the square of the linear lithographic dimension of the MTJ. This suggests that shrinking lithographies will allow STT MRAMs to achieve NOR Flash memory densities and perhaps even approach single-level cell (SLC) NAND Flash memory densities. This is critical to the success of MRAM in the non-volatile semiconductor memory market because NAND Flash popularity depends solely on the technology’s lead in cost/bit storage, which is built primarily on the NAND Flash memory cell’s small size. Note that MRAM has infinite write endurance, unlike competing non-volatile semiconductor memory technologies.

Commercial STT MRAMs should be available within the next few years.

Table 1 compares significant attributes of various volatile and non-volatile memory technologies.

**Table 1: Attributes of Volatile and Non-Volatile Memory Technologies**

Attribute	SRAM	DRAM	NAND-Flash	NOR-Flash	MRAM
Non-Volatile			✓	✓	✓
Fast Read	✓			✓	✓
Fast Write	✓	✓			✓
High Endurance	✓	✓			✓
Low Power		✓			✓

If the RAID server’s primary memory cache—currently implemented with volatile DRAM—were inherently non-volatile, there would be no need for smaller non-volatile memories to hold the RAID server’s firmware, configuration data, and the journal log file. NAND Flash could serve as the only memory a RAID server needed if it had faster read and write cycle times and if it were not susceptible to write endurance failures. Currently available and soon-to-be available MRAMs are already the best candidates for the RAID server’s journal and log files. STT MRAM, when it becomes available, will have all the required attributes needed to serve all non-volatile memory functions in the RAID server, including the primary RAID cache.

*About the Authors*

*Barry Hoberman has held management positions at several technology companies, including founder and chief executive officer of inSilicon (now part of Synopsys) and chief executive officer of Virtual Silicon. His primary focus is in strategy and business development for semiconductors, semiconductor manufacturing and semiconductor intellectual property (IP). He has 13 U.S. patents and holds two B.S. degrees from the Massachusetts Institute of Technology. You can reach Barry Hoberman at bhoberman@crocus-technology.com.*

*Steve Cliadakis has over 20 years experience in business development, marketing and product development for technology companies, with a concentration in semiconductors and IP. He is the founder of Silicon Impact, providing business development and strategy services for start-ups and well-established companies. Cliadakis holds a B.E. in electrical engineering from the State University of New York at Stony Brook and an MBA from Adelphi University in New York. You can reach Steve at steve@siliconimpact.com.*