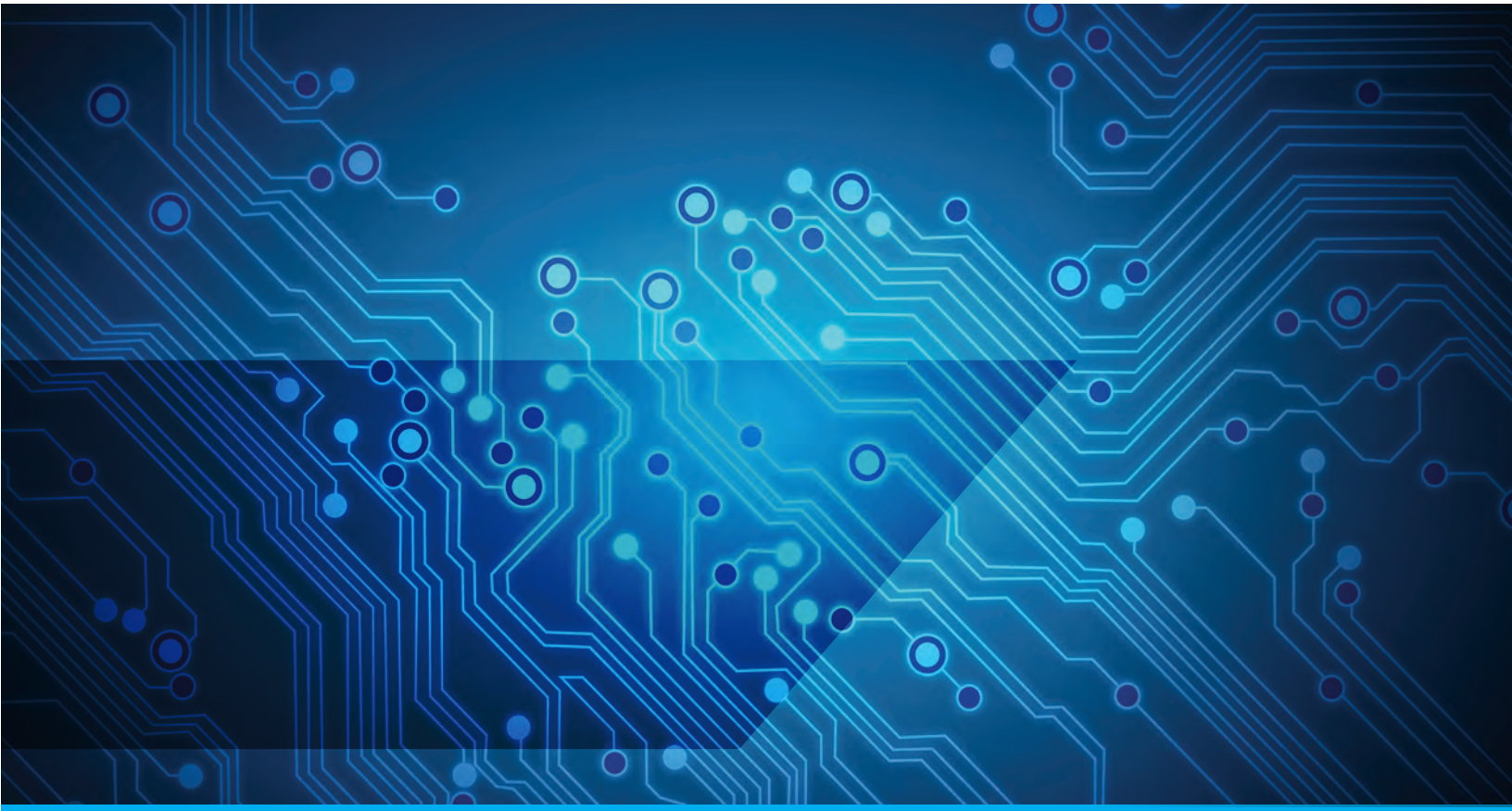
# ARTIFICIAL INTELLIGENCE HARDWARE

## THE OPPORTUNITY FOR FAST, PERSISTENT MEMORY–MRAM



SPIN MEMORY™

# MRAM, THE FASTEST NON-VOLATILE MEMORY, HOLDS THE KEY TO AI

**MRAM HAS EMERGED AS THE MEMORY OF CHOICE FOR ARTIFICIAL INTELLIGENCE (AI) DUE TO ITS UNIQUE COMBINATION OF SPEED, DENSITY, ENDURANCE AND NON-VOLATILITY. AI APPLICATIONS ARE, BY DEFINITION, DATA-DRIVEN, REQUIRING MORE AND MORE MEMORY BOTH ON- AND OFF-CHIP. TRADITIONAL MEMORY SOLUTIONS ARE FALLING SHORT OF THE DEMANDS OF AI, CREATING THE OPPORTUNITY FOR MRAM'S RAPID ADOPTION AND PROLIFERATION.**

## INDUSTRY BACKDROP

### HARDWARE IS THE KEY BATTLEFIELD FOR AI

Software has been the star of high tech over the past few decades, defining the game-changing innovations that defined the era. Although innovations in chip design have practically enabled all of these next-gen devices, semiconductor companies have only captured a small share of the total value in all this technology — about 25% of the value in PCs and a meager 15% in Mobile, according to McKinsey[1]. But the story for semiconductor companies will be different in the race for AI. Hardware is becoming the key performance bottleneck, and solutions to the bottlenecks become differentiators. That's the reason why leading internet players — such as Google, Facebook and Apple — are rushing to become silicon designers in search of a hardware competitive-edge. Even Amazon is entering this competition, developing a custom Arm-based server processor that has lowered their costs 45%[2].

The value at stake is massive. According to McKinsey[3], the AI semiconductor market should reach $65B by 2025, growing at 15% CAGR. Silicon startups, for the first time in decades, are now well-positioned to capitalize on this opportunity. The agility of small cross-functional teams is well-suited for learning fast and developing innovative hardware solutions. VCs have also realized this unique opportunity and invested a record $9.3B[4] in AI startups in 2018.

### AI POSES A MAJOR COMPUTING CHALLENGE

AI is rapidly building a connected world where sensing, computing and communications are incorporated into tens of billions of devices that monitor their environments, make decisions and send information to the cloud. One trillion new IoT devices will be produced by 2035 according to Arm[5]. These devices will generate an explosion of data to be processed. Cisco predicts that by 2022, global mobile data traffic will reach more than

[1] McKinsey & Co "AI hardware" January 2019
[2] AWS press release, November 11, 2018
[3] McKinsey & Co "AI hardware" January 2019
[4] CB Insights
[5] Arm press release

930 exabytes annually, 46% '17-'22 CAGR. An autonomous vehicle in 2023 will generate 4Tb per day, compared to the ~1Gb per day currently generated by an individual. For the first time in history, machines are creating more data than humans, a trend that will accelerate as AI expands.

But data is useless unless it is effectively captured, manipulated, extracted and exploited. This exponential growth of data creation is producing a massive computing and storage challenge to extract information efficiently in terms of performance, energy and cost.

At every level of the computing infrastructure — from edge to data center and back — energy consumption is becoming an ominous hurdle. At the edge, we often need ultra-low-power solutions that can run on battery or even harvested energy. In these cases, compute, memory and AI will be local. And while the power consumption of most IoT edge devices is low, the total energy consumption is staggering simply due to the massive number of them.

Consider a simple IP camera for home security. It consumes only 5 to 8 watts[6], but in 2020 all IP cameras combined will consume more power than the energy generated by a standard power plant in the United States. In the data center (projected by Applied Materials to consume 10% of all electricity worldwide by 2025), the key constraint is the amount of energy the power company can deliver to the building, rather than what the data center costs. And the problem will get worse, and get worse exponentially. It is estimated that training a single AI model can emit as much carbon as five combustion engine automobiles over their operating lifetime.

Gary Dickerson, CEO of Applied Materials, defines the AI grand challenge as the need to improve compute performance per watt by 1,000x. He highlights that the growth of AI demands immediate development of more energy-efficient computing paradigms.

## AI-SPECIFIC HARDWARE IS REQUIRED

The pending crisis is that the evolutionary balance between cost/performance and energy efficiency that fueled the computer industry for more than 40 years has stalled and can no longer match the exponential growth of data and compute:

- Transistors fundamentally are not shrinking due to the ending of Moore's Law,
- Power is limiting what can be put on a chip due to the end of Dennard scaling,
- Parallelism and the number of processors per chip are approaching their limits due to Amdahl's Law.

Thus, the rapid improvements in processing that we have enjoyed with the Moore's Law era must now come through other innovations in computer architecture, chip design and semiconductor process improvements.

**THE CURRENT PERFORMANCE BOTTLENECK IS IN A COMPONENT OF COMPUTING THAT MOST PEOPLE RARELY CONSIDER: MEMORY.**

[6] SEMICON 2019, Applied Materials

## MEMORY A CRITICAL AI BOTTLENECK

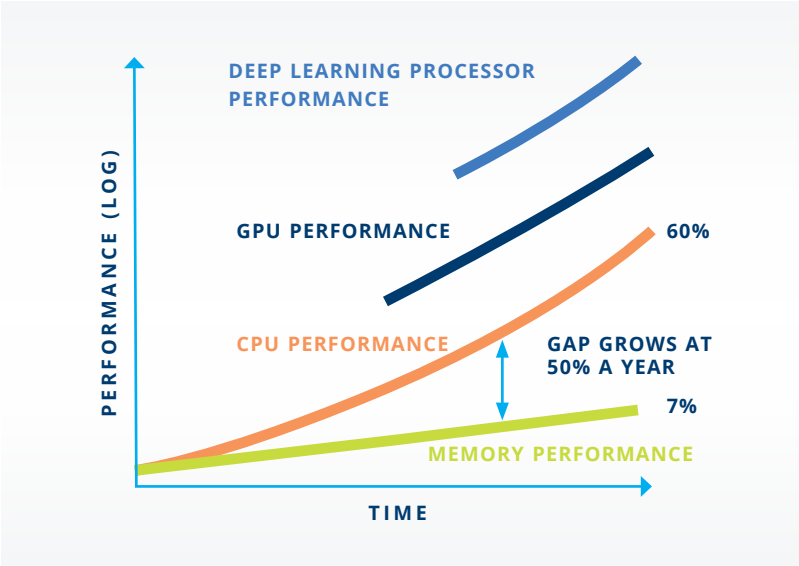



**FIGURE 1**

THE GAP BETWEEN CPU PERFORMANCE AND MEMORY PERFORMANCE GROWS AT MORE THAN 50% PER YEAR

The memory system is a fundamental performance and energy bottleneck in all computing systems and, in turn, AI performance. As processors' speeds have improved, they struggle to transfer data from the memory with enough bandwidth to keep the cores calculating and waste much of the time idling waiting for memory I/O to complete operations. This memory slowness compared with CPUs is generally known as the Memory Wall.

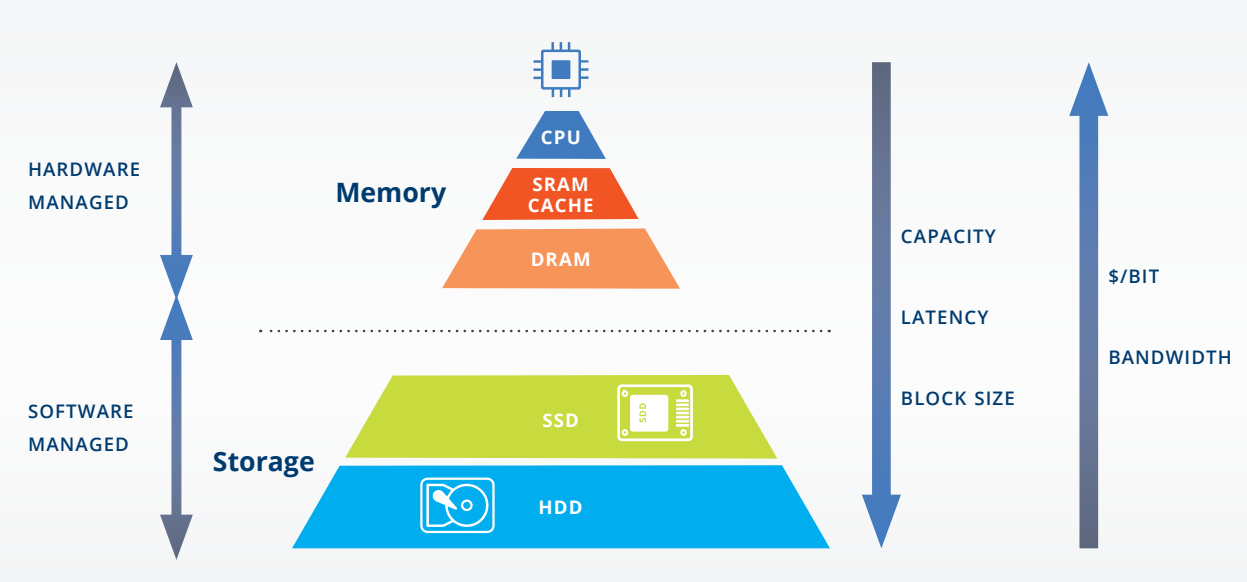
Ideally, there would be only a single memory technology — the main memory. That memory would be instantaneously accessible, allow permanent storage, require minimal energy, and be compact and cheap. Unfortunately, there is not yet a technology that offers all those features. There are two memory types:

- **Volatile Memory**, which requires power to maintain stored information, offers fast write-speeds, but is relatively expensive. DRAM accounts for a majority of the volatile memory market, while stand-alone SRAM accounts for a smaller portion. However, SRAM is the critical on-chip memory for virtually every logic device, and most of the chip's transistors are devoted to this on-chip ("embedded") SRAM block. SRAM is faster, but is more expensive than DRAM.
- **Non-volatile Memory**, which maintains stored information even in the absence of power, is cheaper, but lacks the write-speed or write-cycle endurance required by processors. NAND flash accounts for the majority of the non-volatile memory, while NOR flash makes up a smaller portion. Like SRAM, NOR is often embedded on logic chips, especially microcontrollers and other IoT devices.

To emulate the ideal memory, hardware designers created a hierarchical memory structure combining small amounts of volatile memory (fast, but expensive), and large amounts of non-volatile memory (slow, but cheap), in such a way that the combination behaves as if large amounts of fast memory were available at an affordable price.

This memory hierarchy can typically be broken down into four levels: CPU > Cache > Main Memory > Storage. If information is not present in one of the CPU registers, the CPU will request information from the memory. First, the cache will verify whether it has the requested information available. The cache, located on the same

**FIGURE 2**




chip as the CPU, is composed of a small amount of fast and expensive memory, typically embedded SRAM. Thus, if the requested data is available in the cache, it can be retrieved quickly. Otherwise the main memory, which is significantly larger and composed of slower and cheaper DRAM, must be addressed to retrieve the data. This memory is located physically further away on separate chips. If the requested data is in the main memory, it is provided to the cache, which in turn, transfers it to the CPU. If not, the storage system, which offers a vast amount of memory at the lowest price and much slower speeds is accessed. Therefore, to maintain the illusion of having lots of fast accessible memory for operations, it is critical that with high probability, the cache contains the data the CPU needs. Ideally, the main memory and the storage system would get accessed only sporadically. Modern CPUs are so fast that they spend much of their time idling at the so-called memory wall. Program execution time depends almost entirely on the speed at which memory can transfer data to the CPU. Additionally, memory and data transfer are a key driver of power consumption. All these data transfers between the different levels of the memory hierarchy often consume more energy than the energy consumed in computation. For instance, Google has found that in a mobile system, over 60% of the total system power budget is used to transfer data back and forth between on-chip and off-chip memories.

## SCALING LIMITATIONS OF LEGACY MEMORY TECHNOLOGIES

The issue is that memory access times have not been able to keep pace with increasing clock frequencies. Historically, there has been a growing disparity between CPU clock rates and off-chip memory I/O rates, (i.e., DRAM access time). Every decade gains in CPU compute performance have outpaced improvements in DRAM memory speed by 100 times. Below 20nm, the scaling of DRAM technology is leveling off. The limit is the fundamental physics of electrical charge storage, and how we're able to store that charge in such a way to maintain and retrieve bits of information. The issue with DRAM is the capacitors that sit over a memory cell to store data have had to get taller and taller as chip processes have shrunk in order to maintain the same surface area of the capacitor and therefore maintain its capacitance. This has led to increasing process complexity and cost increase.

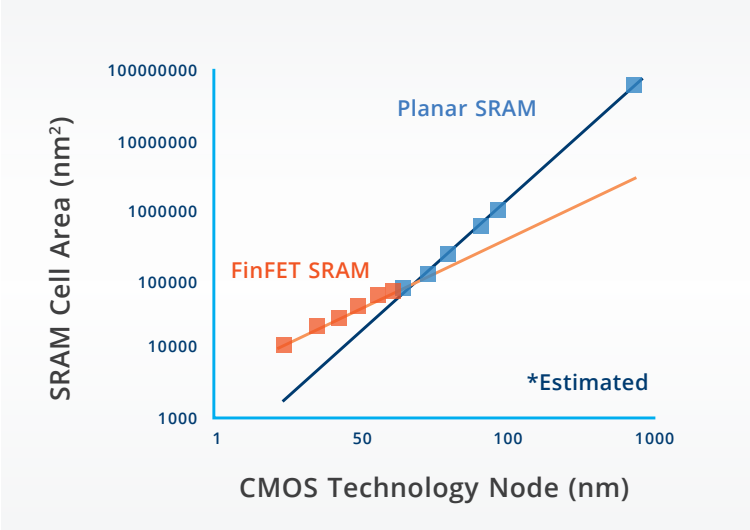



**FIGURE 3**

SRAM BITCELLS IN FINFET TECHNOLOGIES ARE NOT SCALING WITH MOORE'S LAW

On-chip cache memory, SRAM, which is 10 to 100 times faster than off-chip DRAM, was expected to knock down the memory wall. However, cache memories have their own set of problems. SRAM hit a major inflection point in scaling at around the 28-nm node. SRAM bitcells size has gone from about $150F^2$ to over $500F^2$ (F being the minimum feature size). This means SRAM is taking relatively more area as process geometries shrink. The other key mark against SRAM is its power consumption. SRAM bitcells consume power even when sitting idle ("leakage"), and this power is proportional to the number of bitcells in the memory. This creates enormous power leakage problems for AI applications that require large memories. In such applications, the majority of processor power may be consumed by cache memory, not by computations. These two SRAM challenges have conspired with the almost insatiable demand for increased on-chip cache memory speed and capacity to create real challenges in cost and wasted energy. This demand comes from both mobile and data center applications. The resulting requirement for energy efficiency is obvious in the former due to battery life limitations, but are also becoming of critical importance in the latter.

## FAST, PERSISTENT MEMORY REQUIRED

There are two ways to attack this memory latency and power consumption problem, and they both have their proponents. One method is to keep building on the existing approaches. The industry has responded by developing several methods to increase performance by signaling at higher data rates and using stacked architectures for greater energy efficiency and performance, and by bringing compute closer to the data. All of these approaches are reaching practical limits, and they don't address the fundamental flaw of using volatile memory technologies. All applications, OS and software stacks are designed with the underlying assumption that memory is volatile and the memory contents can be lost. Complex mechanisms are used to ensure that if there is a power failure, little to no data is lost. Although this comes at the cost of performance, the industry has come to accept it as a trade-off. However, it doesn't have to be that way.

The other, more profound challenge, is the idea of finding a replacement to these legacy memory technologies. There is an incredible amount of demand for and investment in potential alternative memories. An effective replacement memory to SRAM is sorely needed, and a new main memory with even half standard DRAM latency

would give programmers an opportunity to revisit decades of assumptions about how microprocessors should be built. What's needed is a memory that is low cost, low power, non-volatile and can overcome the memory speed and power bottlenecks in the current architecture. AI needs a fast, persistent memory that can achieve large-scale energy savings and performance improvements, thus greatly extending battery life and better user experience.

## OVERVIEW OF EMERGING MEMORY TECHNOLOGIES

The scaling limit of today's technologies comes from their storage mechanism. All these mainstream memory technologies are essentially charge-based memories: SRAM stores the charges at the storage nodes of the cross-coupled inverters, DRAM stores the charges at the cell capacitor, and flash stores the charges at the floating gate of the transistor or in a dielectric such as silicon nitride or similar compounds. This charge storage mechanism creates a "scaling limit," determined by the number of electrons that can be stored on a flash gate or DRAM capacitor. As the process technology shrinks, the memory cell gets smaller and the number of electrons the cell can store declines to approach a lower limit of what can be accurately measured.

Emerging memory technologies address this issue by using innovative storage mechanisms. The candidates include PCM, ReRAM and MRAM. These innovative technologies share some common features: they are nonvolatile memories, and they differentiate their states by switching between a high resistance state (HRS, or off state) and a low resistance state (LRS, or on state). The transition between the two states can be triggered by an electrical stimulus (i.e., voltage or current pulse). However, the detailed switching physics is quite different:

- **PCM** works by changing the phase of a special kind of glass within the bit cell, switching between the crystalline phase (corresponding to LRS) and the amorphous phase (corresponding to HRS); and
- **ReRAM** is an umbrella term for any memory whose bit state is defined as a higher or lower resistance, relying on the formation (corresponding to LRS) and the rupture (corresponding to HRS) of conductive filaments or changing interfacial mechanisms in the insulator between two electrodes that affect the charge conduction across the insulator.
- **MRAM** uses magnetism to store bits. A memory cell consists of two ferromagnetic layers separated by a thin tunneling insulator layer, and alternate between parallel configuration (corresponding to LRS) and antiparallel configuration (corresponding to HRS).

Due to the different underlying physics, the device characteristics are different among emerging NVMs. **Table 1** compares the typical device characteristics. They all have different application spaces and adoption challenges from aspects of process compatibility, manufacturing yield, performance variability and reliability.

**TABLE 1**

| Technology | FeRAM | MRAM* | ReRAM | PCM | DRAM | NAND Flash |
|---|---|---|---|---|---|---|
| Nonvolatile | Yes | Yes | Yes | Yes | No | Yes |
| Endurance | $10^{12}$ | $>10^{12}$ | $10^{6}$ - $10^{11}$ | $10^{8}$ - $10^{11}$ | $10^{15}$ | $10^{2}$ - $10^{5}$ |
| Write Time | 100ns | ~10ns | ~50ns | ~75ns | 10ns | 10µs |
| Read Time | 70ns | 10ns | 10ns | 20ns | 10ns | 25ns |
| Power Consumption | Low | Medium/Low | Low | Medium | Very High | Very High |

* STT-MRAM WITH IMPROVEMENTS FROM SPIN MEMORY'S PSC AND ENDURANCE ENGINE

## PCM

PCM targets the Storage Class Memory (SCM) market with performance and cost between DRAM and NAND Flash. However the key challenge for PCRAM cell design is the relatively large write current required to melt the phase-change materials. In addition, hurdles remain in the areas of endurance (maximum reported as $10^{11}$) and reset energy. The PCRAM's switching speed (>50ns) is limited by the slow crystalline process, also 10 times longer than its STT-MRAM and ReRAM counterparts, while the PCRAM's endurance is comparable to that of the ReRAM. The PCRAM's data retention is limited by resistance drift due to the relaxation of the amorphous state. Thus, sophisticated circuit-level compensation schemes are needed.

Despite the fact that the PCRAM's cell characteristics are less competitive than ReRAM in terms of the write power and speed, today PCRAM's process and manufacturing technology is quite mature. The most successful commercial application is 3D XPoint™, a memory developed by Intel and Micron. 3D XPoint™ is positioned as either a nonvolatile cache between DRAM and 3-D NAND or as an ultra-low latency NAND-replacement. For the latter, the price for a 3D XPoint™ only SSD is several times that of a NAND-based version. No known path has been identified that would allow PCM to reach the endurance (above $10^{13}$ cycles) and speed (sub-20ns) requirements for SRAM- and DRAM-replacement. Taking Intel's Optane™ SSD as the most advanced example, its specifications for maximum "Device Writes per Day — DWPD", are between 1 and 2 million cycles.

## ReRAM

**There are two subcategories within ReRAM:** Metal Oxide Resistive Memory (OxRAM) and conductive bridging RAM (CBRAM). The difference is that OxRAM's filament consists of oxygen vacancies in the oxide layer, while CBRAM's filament consists of metal atoms. Despite different underlying physics, these two types of ReRAMs share many common device characteristics. The only notable difference may be that OxRAM's on/off resistance ratio may be smaller (in the range of 10–100x) and offer better endurance up to $10^{12}$ cycles, while CBRAM's on/off resistance ratio can be quite large ($10^{3}$–$10^{6}$), but with limited endurance (<$10^{4}$ cycles).

The key challenge of ReRAM cell design is the variability of the switching parameters. Owing to the stochastic nature of ionic (oxygen vacancies or metal ions) migration, the filament shape varies from device to device and also from cycle to cycle (within one device). Remarkable variation in resistance distribution (which can be one or two orders of magnitude) adds challenges to the sensing circuit design and requires the write/verify techniques to program to the target states, which could be latency consuming for the MLC operations.

ReRAM is being touted as a good, embedded non-volatile solution to take the place of existing embedded EEPROM and Flash because of ReRAM's low process complexity, relatively fast write speeds of around 50ns, and an endurance at around one million cycles. No known path has been identified that would allow ReRAM to reach the endurance (above $10^{13}$ cycles) and speed (sub-20ns) requirements for SRAM- and DRAM-replacement.

## STT-MRAM

From all these emerging memories, STT-MRAM is the most promising as a direct replacement for embedded SRAM. STT-MRAM is the fastest non-volatile memory, with an endurance of over $10^{11}$ for SRAM-like applications, and the ability to embed in 5nm CMOS process nodes and the promise of cell sizes that challenge DRAM. Of all the memory technologies on the market today, none possesses the uniquely promising combination of speed, density and endurance.

Debuted as a humble 4Mb product by Freescale™ in 2006, the MRAM has grown to a 1Gb product of Everspin® in 2019. During this period, MRAM has overcome several hurdles, and we have reached a stage where the market is aggressively pursuing MRAM memory. One of the main hurdles that MRAM overcame between 2006 and 2016 is the way the information is written. The 4Mb MRAM used a magnetic field based switching technology that would be almost impossible to scale below 100 nm. The 1Gb MRAM, on the other hand, uses a different writing mechanism based on Spin Transfer Torque (STT), which is scalable to very low dimensions. In addition to the difference in the writing mechanism, there has also been a major shift in the storage material. Whereas the 4Mb MRAM used materials with in-plane magnetic anisotropy, the 1Gb MRAM uses materials with a perpendicular magnetic anisotropy (PMA). MRAM based on PMA is also scalable to much higher densities.

Today, STT-MRAM's process and manufacturing technology are relatively mature because the materials needed to manufacture MRAM are already used in very high production volumes to make silicon-based HDD heads. This is a strong advantage for MRAM technology, since the materials' interaction with silicon is already very well understood. Furthermore, Applied Materials' Endura Clover PVD system is a breakthrough in the thin films high-volume manufacturing deposition machine. This will facilitate the precise deposition of the sub-nanometer Magnetic Tunnel Junction (MTJ) films and lower the cost barrier to entry – empowering more foundries to widely adopt this technology.

As compared to SRAM, STT-MRAM has an advantage of a smaller cell area, while maintaining low programming voltage, fast write/read speed, and long endurance. Thus, STT-MRAM is attractive as a replacement for embedded memories (e.g., SRAM or NOR Flash) and even a replacement to DRAM.

# MRAM TECHNOLOGY AND CHALLENGES

## WORKING PRINCIPLE OF MRAM

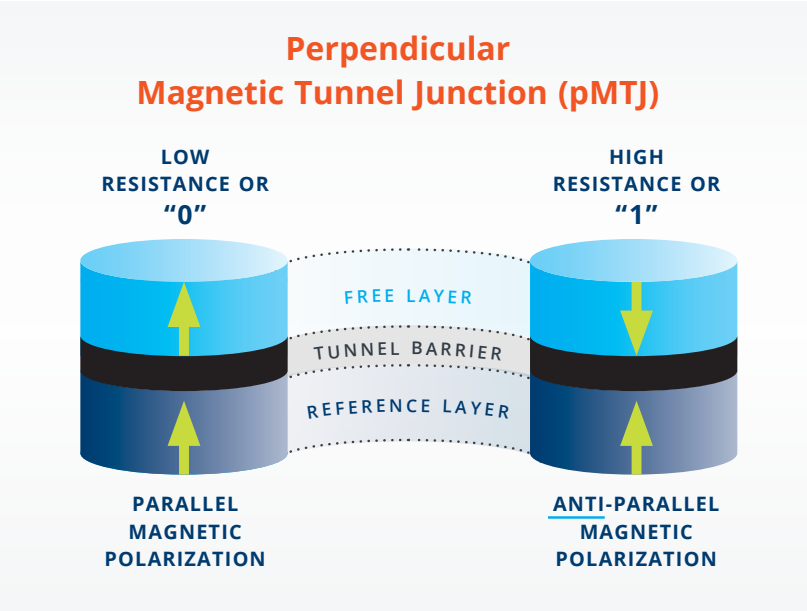



**FIGURE 4**
SIMPLE ILLUSTRATION OF A pMTJ DEVICE SHOWING THE "1" AND "0" STATE DEPENDENCE ON FREE LAYER POLARIZATION.

The most basic MTJ structure consists of two ferromagnetic layers separated by a dielectric spacer layer, the tunnel barrier. The relative orientation of the magnetizations in these two layers determines the resistance of the MTJ device. For most materials, the resistance is low when the magnetizations of the two layers are parallel, because the majority spin band electrons can tunnel into the majority spin band on the opposite side of the barrier. When the orientation is antiparallel, the resistance is high since the majority spin band electrons have to tunnel into the minority spin band of the opposite electrode.

### The simplest possible MRAM cell design has the following components:

- **The Free layer (FL)**, sometimes called the recording layer or the storage layer, is the ferromagnetic layer retaining the stored information.
- **The tunnel barrier (TB)**, a thin (around 1 nm) insulating nonmagnetic layer, provides the means to switch and read the state of the free layer with a spin-polarized tunneling current.
- **The reference layer (RL),** the other magnetic layer, provides a stable reference magnetization direction for the FL reading and switching. This layer is designed to have magnetic anisotropy much higher than the FL so that it never switches during memory operation.

In MRAM, the three basic functions of any memory device are performed as following: (1) the read operation is carried out by sensing the resistance difference between two states of a magnetoresistive device, (2) the

storage of information relies on the magnetic retention properties, arising from the magnetic anisotropy of the storage layer, and (3) the write operation is performed by changing the orientation of the storage layer magnetization.

To create a memory array from MTJ devices, each device is typically integrated with an isolation transistor that can be turned on to pass a current selectively through the MTJ devices of interest, such as during the read operation. Since each memory cell typically has one transistor and one MTJ, this particular architecture is known as the 1T-1MTJ MRAM architecture. Other architectures have been proposed and evaluated for various purposes, but the 1T-1MTJ cell is the most commonly used.

## OVERVIEW OF MRAM TECHNOLOGY GENERATIONS

MRAM technology can be classified by the switching method employed to write data. First-generation MRAM, Toggle MRAM, is generally understood to include methods using magnetic fields to program the array. This technology is in mass production for different niche applications. A strong advantage of field switching is unlimited write endurance, since reversing the free-layer magnetization with a magnetic field does not create any wear-out effects. A disadvantage is the difficulty in scaling to smaller cell sizes, due to several factors including the magnitude of the required switching currents and the somewhat complex memory cell geometry.

Second-generation MRAM uses a switching mechanism based on a physical phenomenon referred to as the spin torque transfer (STT) effect. To explain this effect, let's assume that the two magnetic layers begin the data writing process with their magnetization opposite to each other. When a current is applied to the device that injects electrons from the reference layer to the free layer, the electron spin injected from the reference layer will cause a local magnetic torque on the electrons within the free layer. Eventually enough spin will be injected that it causes the magnetic polarization of the free layer to flip into a parallel state with the reference layer. If the current polarization is reversed, the reference layer will act as a spin mirror, causing the opposite orientation spin to build up in the free layer and will flip it back into the opposite polarization. This is how 1s and 0s are written in an STT-MRAM. STT switching can be accomplished with reasonable efficiency with MTJ devices having either in-plane or out-of-plane magnetization. STT-MRAM with in-plane devices began commercial production in 2015 with a 64-Mb product, and STT-MRAM with perpendicular MTJ devices is expected to ramp to volume production in 2020 with a 1Gb product.

## HISTORICAL CHALLENGES

While MRAM is already in mass production, it has traditionally had limitations when trying to replace SRAM or DRAM:

- **Endurance:** MRAM endurance is too low. Native MRAM bits suffer from an oxide breakdown, in MRAM's case MgO, and can achieve endurance levels of $10^6$ – $10^8$ cycles before wearing out. While this is far greater than traditional Flash, it is 5-6 orders of magnitude lower than what's needed for SRAM or DRAM applications. Initial deployments of MRAM, therefore, are targeted at embedded NVM replacement and to niche buffer-memory applications in the stand-alone market. (Note: some companies incorrectly specify MRAM endurance under a presumed "usage model" that assumes that write operations are spread across the memory and there is a very low probability of a word being written per cycle [for example, $10^{-6}$], which artificially improves the memory specifications.)

- **Speed:** MRAM write speeds are too low. Fast (<20ns) MRAM write cycles require a very high write voltage or produce an unacceptably high error rate. This higher voltage has the direct effect of increasing the wear out of the MgO and further degrading endurance. This is one of the major trade-offs of traditional MRAM.
- **Stochastic Write:** MRAM is inherently stochastic. There is a non-zero probability that a write operation will fail to change the state of the bit cell. This is the first semiconductor memory that randomly "may not work" for no extrinsic reason. This is reflected in a "Write Error Rate" statistic which must be addressed by a variety of sub-optimal methods, including increasing Error Correction Code (ECC) bits and thus greatly increasing size, writing at a higher voltage and thus degrading endurance, and so on.
- **Non-symmetric Read/Write:** Due to the inherent time it takes to change the magnetic state of the MTJ, read and write times are not symmetrical in MRAM as users typically write multiple times to compensate of the stochastic write problem. While this may not cause much of an issue in NVM applications, it can significantly reduce the performance of a cache, the predominant use of embedded SRAM, and severely limit the market in stand-alone applications, especially in a DDR-based memory system.
- **Non-volatility/Data Retention:** As speeds of the MTJ grow or temperatures rise, the non-volatility of MRAM degrades. To put this in perspective, while NVM applications can achieve 50ns access times with 10-year retention at 125C, an application tuned to a 15ns access time may only achieve a few hours' retention. Data retention may not be an issue for cache applications, but it may limit the market for Storage Class Memory.
- **Density:** As sizes shrink, the switching currents decrease, but it becomes more difficult to control variations and electrical distributions. An important parameter is the energy required to switch a bit forms a key part of the total energy comparison with DRAM and SRAM. Cooperation with tool manufacturers is critical as the device formation and Encapsulation are the two main processing steps affecting distributions at high density.
- **Read Disturb:** The action of reading is similar to a write but at smaller currents. Nevertheless, read disturbs can take place and need to be controlled. The problem of read-disturb increases as the size of the MTJ decreases as the ratio between read and write currents become compressed. Smaller switching currents are also desirable for faster devices, exacerbating the problem.
- **Process Node Migration:** At each new process node, the MTJ must be re-engineered as the structure of the previous generation will not perform optimally. Simulations and modeling of the intricate quantum interactions are not as well-established as those in the semiconductor industry. Therefore, a significant number of development cycles are required to deliver a reliable MTJ at each new process node.

# OUR SOLUTIONS

**SPIN HAS BEEN RESEARCHING MRAM FOR THE PAST EIGHT YEARS AND HAS DEVELOPED A PORTFOLIO OF TECHNOLOGIES TO ADDRESS THE SHORTCOMINGS OF MRAM. MANY OF SPIN'S SOLUTIONS ARE SYSTEMS APPROACHES TO MRAM AND ARE MODULAR, ALLOWING THEM TO BE INDIVIDUALLY ADDED TO ANY MRAM DEPLOYMENT, SPECIFICALLY TARGETING A PARTICULAR ISSUE OF THE ARRAY. THESE TECHNOLOGIES AND INNOVATIONS ARE HEAVILY PATENTED BY OVER 150 US PATENTS AND PATENT APPLICATIONS.**

## ENDURANCE ENGINE

Spin Memory's patented Endurance Engine is the key technology that enables STT-MRAM to achieve SRAM-like performance. It is a circuit-only solution that addresses the endurance, stochastic nature and symmetric read/write problems with MRAM. The Engine improves MRAM endurance by up to six orders of magnitude while enabling 100MHz read/write operation. It converts a native $10^8$-$10^9$ write-cycle MRAM array to $10^{14}$-$10^{15}$ cycles of endurance — levels required for seamless SRAM replacement. No special processing, material changes or pMTJ changes are required to realize the benefits of the Engine. It can be added to any MRAM array from any manufacturer at any foundry of any size and will boost the endurance of the array by a staggering five-to-six orders of magnitude. This would boost MRAM from NVM applications into the highly desired SRAM, DRAM and SCM territory.

The Endurance Engine works by tracking and correcting inherent write-errors. The Engine tolerates high error rates to the point where the system cannot tell the difference between MRAM and traditional SRAM. The key to the Endurance Engine is that its operation is hidden and transparent to the system. From a system perspective, the MRAM is operating exactly as an SRAM would. The system can access the MRAM every cycle as it normally would, while the Endurance Engine is operating in the background and does not interfere with the system's requests. Further, the Engine creates symmetric read and write operations, allowing the MRAM array to run at much higher speeds, essential for cache applications. The company believes the Engine will allow MRAM to replace SRAM in process nodes less than 28nm as MRAM is one half the size and operates at zero leakage.

The Engine adds approximately 20% to the area of a 16Mb MRAM macro, but can be amortized across larger capacity macros. The overhead of a 1Gb array is less than 10%. However, as the Engine can tolerate higher error rates, the Engine allows for smaller write currents, which enables the underlying drive transistors to shrink. This results in about a 20% reduction in memory cell area which can actually shrink the overall size of the array by 10% at 1G sizes, including the overhead of the Engine itself. The Engine literally pays for itself and more.

We believe that MRAM is the technology enabler for using MRAM as a replacement to SRAM. As the Engine is circuits-based only and requires no changes to the magnetics, it can be added to the array well after the MRAM process has been established and qualified. This will speed the adoption of the Engine. The company believes this invention will be universally deployed across every MRAM array worldwide.

### PSC

The Precessional Spin Current (PSC) structure, also known as the Spin Polarizer, consists of very thin layers added to the top of an MTJ, dramatically increasing spin-transfer-torque efficiency by up to 70%. This gain in efficiency can be "spent" by lowering write currents, increasing switching speeds, decreasing MTJ size or increasing data retention. It is covered by over 30 granted and pending patents.

We have collected an extensive amount of data showing how the Spin Polarizer can be tuned to optimize for targeted parameters. For example, for a L2 Cache replacement, the gain can be applied almost exclusively to speed, leaving the current levels the same. Speeds at a given current on an identical MTJ stack with and without the Polarizer were reduced from 25ns to under 10ns. For Automotive or Military NVM applications, the gain may be spent increasing data retention at high temperatures. The company has structures that have passed 10-year levels at 150C.

The Polarizer adds 3-5nm on top of a traditional pMTJ stack, less than a 10% adder in height with no different MTJ materials and can be added during the manufacture of the MTJ itself. Spin has found that no MRAM parameters are degraded by adding the Polarizer. The Polarizer, therefore, is essentially added "free" to any MTJ, though since it is a new magnetic structure, the process itself will need to be re-qualified after the Polarizer is added. We believe the Polarizer will be deployed across all new MTJ developments as the gains are simply too compelling. Any company not using this invention will have MRAM that is larger, slower, higher power or with less data retention. In other words, highly uncompetitive.

### NEXT-GEN MRAM SELECTOR

Selectors are the underlying devices that provide current to switch the state of the MTJ. Selectors are the key to high-density arrays, and there is currently no solution in the market for densities greater than 1G. After having filed patents in the area, Spin outlined a program to develop manufacturable selectors that are compact and deliver enough current to the MTJ. Simulations and calculations have been followed by experiments, resulting in working silicon-based devices that show great promise as product level solutions for both embedded and stand-alone memory.

Spin is looking into next-generation custom structures to enable MRAM to go vertical with a second layer. The company has filed multiple patents around its work in 3D structures and selectors.

## OUR COMPETITIVE STRENGTHS

### WE APPLY OUR STRENGTHS TO ENHANCE OUR POSITION AS THE LEADING SUPPLIER OF MRAM PRODUCTS. WE CONSIDER OUR KEY STRENGTHS TO INCLUDE THE FOLLOWING:

- **Focus on Embedded Products:** We believe that our focus on the embedded market and our superior customer service distinguishes us from our competitors who generate a large portion of their sales from other markets, such as memory discrete products or logic chips. Emerging memory technologies

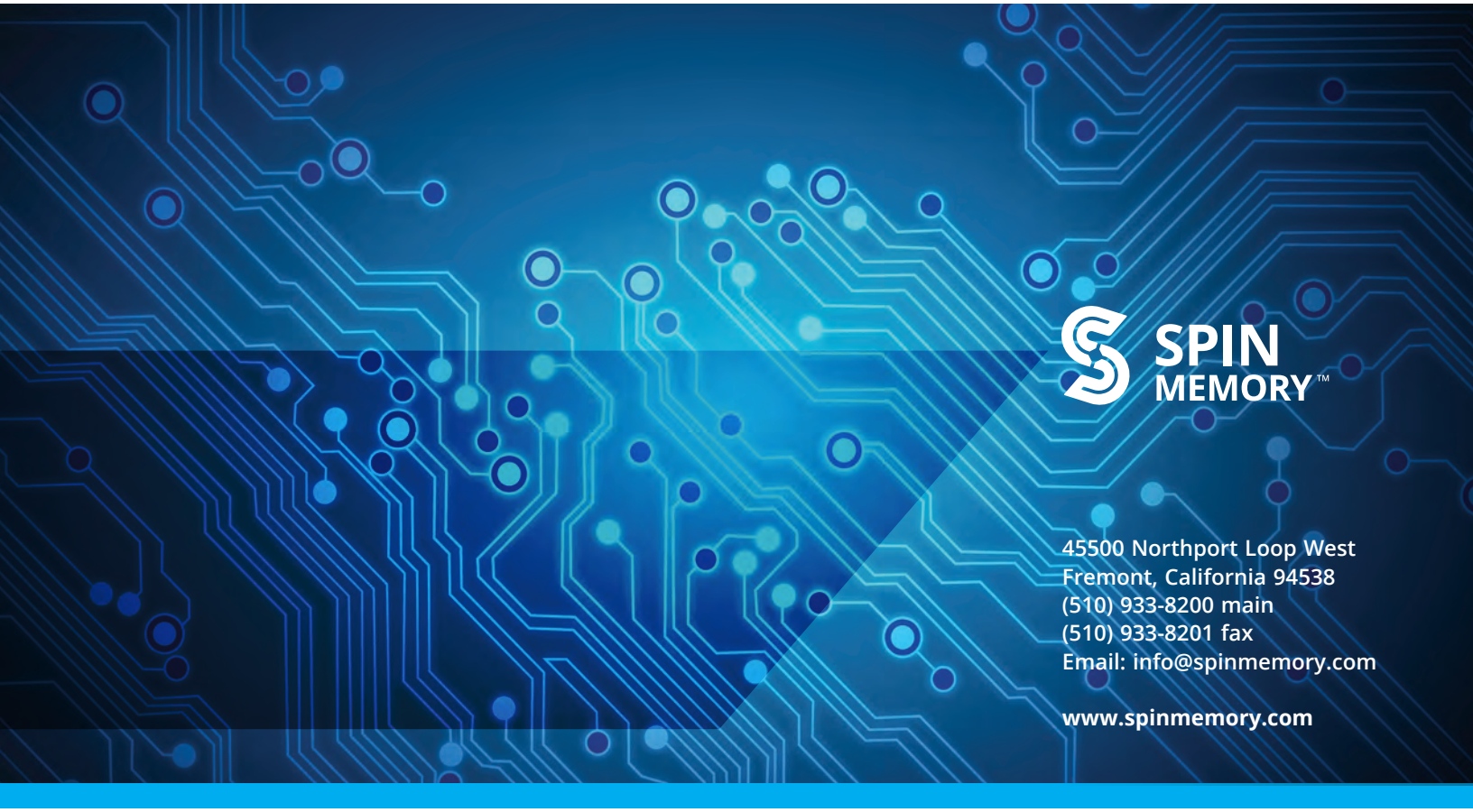
are likely to first reach high volume production as the embedded memory part of logic SOCs, and subsequently migrate to become an important part of the discrete memory market. Our process know-how from that development effort can be applied to stand-alone memory chips, and it will give us a competitive edge.

- **Differentiated IP:** We have developed fundamental IP that can challenge mainstream products like SRAM and DRAM. Industry experts, including the founder of SanDisk, have confirmed the link between Spin's Engine breakthrough and what was done previously for NAND Flash in the late 1980s.
- **MTJ Agnostic Solutions:** Many of Spin's solutions are systems approaches to MRAM and are modular, allowing them to be individually added to any MRAM deployment, specifically targeting a particular issue of the array. These technologies and innovations are heavily patented by over 200 U.S. patents and patent applications.
- **Leveraged Go-to-market Strategy:** Our commercial and licensing agreements with both Arm and Applied Materials provide us with the go-to-market infrastructure and resources for MRAM commercialization and widespread adoption.
- **Vertically Applied Technology:** The breadth and depth of our solutions enable us to cover all customer needs, from process, test, IC design and discrete products.
- **Quick Learning Cycles:** We designed and built a very fast MRAM manufacturing and test line in Fremont, California, to reduce development time of new MTJ's. Typically, it takes only five days from wafer deposition to full MTJ and Memory Array characterization. When compared to the two to three months in the industry, this is 10-20 times faster. The end result is that MTJ development and optimization can be done in months instead of years. This fast-turn line includes an Electron Beam Lithography tool that allows production of MTJ's as small as 20nm, which are suitable for 7nm process nodes. The company has also designed a test chip platform with individual MTJ slots, quick-turn 4K and large 4M complete arrays. This allows the company to quickly optimize MTJ's on any process node from 7nm to 130nm and from L2 Cache to Automotive-grade, high-temperature NVM applications.

## REFERENCES


- Gaurav Batra, Zach Jacobson, Siddarth Madhav, Andrea Queirolo, and Nick Santhanam Artificial Intelligence Hardware: Value Creation for Semiconductor Companies (McKinsey & Co. January 2, 2019)
- Brady Wang SEMICON West 2019: New Non-Volatile Memory Technologies MRAM, PCM and RRAM to Revolutionize Future AI Workloads (Counterpoint Insights. August 13, 2019)
- Introduction to Computer Engineering (UC San Diego, ECE 30)
- Sarvagya Kochak Enabling Higher System Performance with NVDIMM-N (Semiconductor Engineering. August 10, 2017)
- Jim Handy Emerging Memories Today (November 21, 2018)
- Tom Coughlin Coughlin Associates (August 2018)
- Mark LaPedus Scaling The Lowly SRAM (Semiconductor Engineering. July 5, 2013)
- Michael Byrne Memory Is Holding Up the Moore's Law Progression of Processing Power (Tech by Vice. July 1 2014)
- When Memory and Storage Converge (Rambus Press. October 15, 2015)
- Andrew Walker The Trouble with SRAM (Spin Memory. December 17, 2018)
- Tom Coughlin Artificial Intelligence Memory (July 16, 2019)
- Alex Yoon Understanding Memory (February 15, 2018)
- Mark LaPedus 1xnm DRAM Challenges (February 18, 2016)




45500 Northport Loop West
Fremont, California 94538
(510) 933-8200 main
(510) 933-8201 fax
Email: info@spinmemory.com

www.spinmemory.com