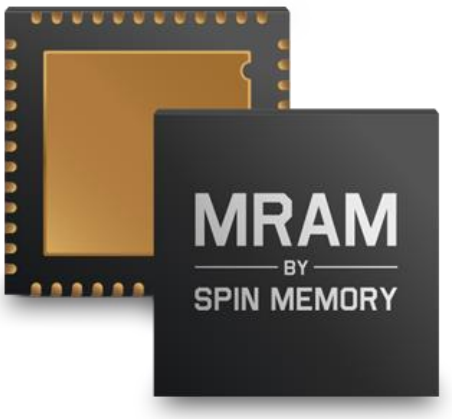
STT-MRAM
in the Fully Composable Data Center

Persistent Memory provides Retention-as-a-Service (RaaS)



A ChannelScience White Paper

Written by
Charles H. Sobey
April 5, 2019

Commissioned by

SPIN MEMORY™

ChannelScience™
Establish the State-of-the-Art™

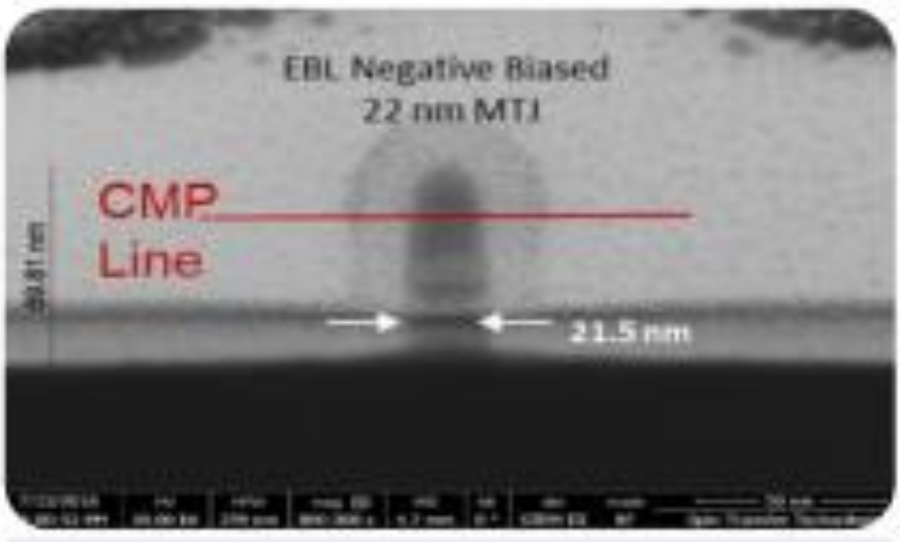
Plano, TX 75024 USA
connect@ChannelScience.com

(972) 814-3441 Voice / Text
@ChannelScience

STT-MRAM IN THE FULLY COMPOSABLE DATA CENTER

Persistent Memory provides Retention-as-a-Service (RaaS)

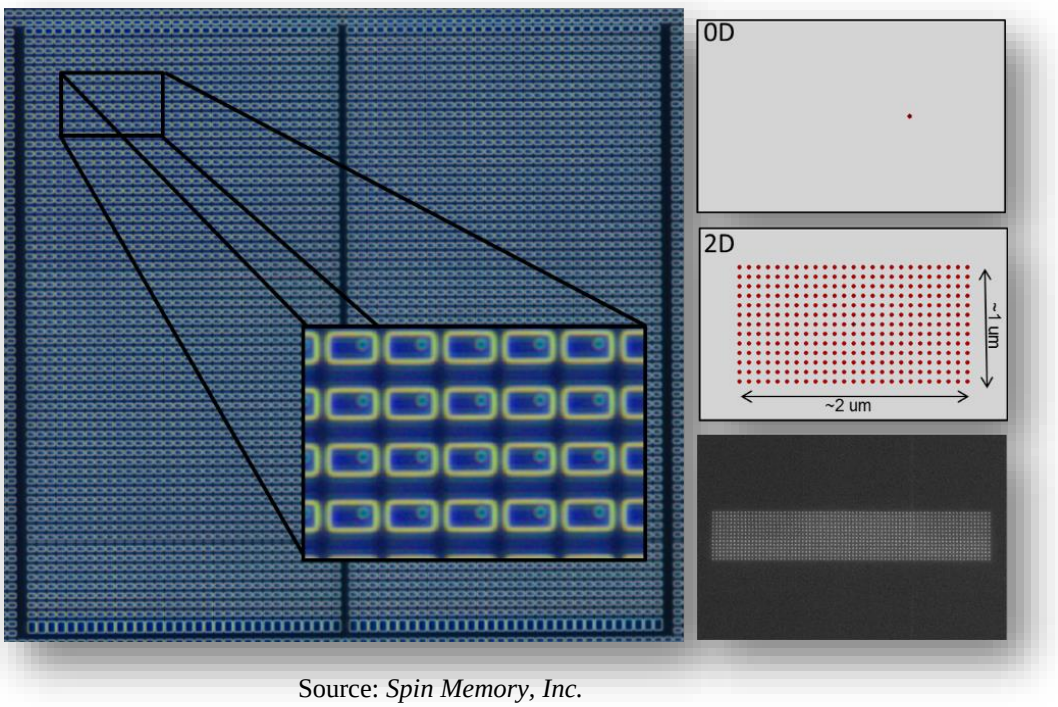
Table of Contents



Source: *Spin Memory, Inc.*

STT-MRAM IN THE FULLY COMPOSABLE DATA CENTER

Persistent Memory provides Retention-as-a-Service (RaaS)



Source: *Spin Memory, Inc.*

1. Executive Summary

The latest MRAM technology uses spin-transfer torque (STT) physics to store and retrieve information quickly at low-power. Traditional MRAM tradeoffs between speed and endurance, between write error rate (WER) and read disturb probability, and between temperature and retention are still concerns for most STT-MRAM implementations. Spin Memory, Inc. changes these tradeoffs with the introduction of two new proprietary technologies. The first is a Precessional Spin Current™ structure (aka PSC™) that improves writeability of the MTJ stack without the need for a higher, stress-causing, write voltage. The second is an on-chip MRAM management system called the Endurance Engine™, which consists of specially-designed circuitry that simultaneously improves speed, endurance, and retention.

The first large impact for STT-MRAM is expected to be in embedded designs. Most major foundries now offer STT-MRAM as a memory option at small process geometries. It is planned to be the denser replacement for embedded SRAM, DRAM, and/or NOR flash. Adding Spin Memory's PSC™ and Endurance Engine™ to MRAM's inherent nonvolatility (and radiation hardness) enables what we have termed "Retention-as-a-Service" (RaaS). That is, built-in reliable retention that does not need to be refreshed or backed by batteries or super-capacitors. Early adopters of this technology are expected to achieve a competitive advantage in major markets, including machine learning accelerators, IoT, inference engines, edge computing, and components and infrastructure for autonomous motion and 5G.

ChannelScience™
Establish the State-of-the-Art™

SPIN MEMORY™

2. The Hyperscale Data Center

Hundreds of thousands of square feet of secure space and megawatts of power are required for a hyperscaler's[1] data center infrastructure to provide quick access to the videos we watch, the web search results we need, and the services we rely on for research, transactions, and business analytics. Figure 1 illustrates the physical plant requirements for one such installation. The goals of the next data center design are the same as the last: Provide faster services on more data for less cost, and at lower power.

Figure 1. Example of the physical requirements of hyperscale data centers. Source: IARPA [1]

Data Center Goal
Provide faster services on more data with less cost and power

To achieve these goals, data center developers must implement cost- and power-reduction plans for every component and resource in the data center, while still satisfying quality of service (QoS) agreements and not degrading security, privacy and recoverability requirements.

This is an extremely challenging task because, throughout the day, the demands of the data center change. In the morning, the main task may be spinning up thousands of virtual machines[2]. As they start up, they cause what is creatively referred to as a "boot storm." Later, the main task may switch to serving email. Throughout the day, the needs of one part of the data center may be for online transaction processing (OLTP) to support a reservation system, inventory

[1] Hyperscaling a computing infrastructure implies that additional computing resources are rapidly provisioned to support customers' changing computing needs. Data center operators that make a business of providing such flexible computing capability on demand are often referred to as "hyperscalers." These include Facebook, Google, Amazon, Microsoft, etc.

[2] Virtual Machine (VM) – Software emulation of a PC, which enables more complete utilization of data center hardware and streamlines software maintenance and updates.

tracking, or ecommerce. Another part may be running a database application on hundreds of gigabytes of data. Still another may be supporting business analytics.

As the day continues, the mix may shift to searches and serving webpages. Later, streaming video may be the most demanded function. As one part of the world's workday ends, another begins and the data center's assets are called upon to support a different mix of needs.

The AI-ready data center includes accelerators such as FPGAs, GPUs, TPUs, and specialized network switches

Further complicating this environment is that workloads continue to evolve. For example, machine learning[3] and artificial intelligence[4] are now requirements for the modern data center. These specialized algorithms work best on hardware designed specifically for them. Accelerators such as FPGAs[5], GPUs[6], and Google's Tensor Processing Units (TPUs)[7] are proliferating throughout the data center. These accelerators can be reconfigured via software to be matched to the algorithms that need them – at any moment throughout the day.

Flexible infrastructure makes this possible and profitable. This is sometimes called a composable infrastructure, or software-defined infrastructure (SDI). Traditionally, the data center's computing, memory, storage, and networking are physically configured to match the needs of a workload. In a composable infrastructure, portions of the physical processors, memory, storage, and networking can be combined logically (with software) to support one workload, while other portions of the same hardware can be configured to support another. The workloads can even be for different customers of the data center.

This flexibility helps to keep the investment in hardware utilized at all times. Accelerators offer even more composability. They themselves can be reconfigured to match the demands of the algorithms of the workload they support. If the composability is quick, varied, and finely grained, the utilization of the hardware can be quite high, while supporting the required QoS level. In a nutshell, data centers want all of their infrastructure producing profit all of the time.

In a nutshell, data centers want all of their infrastructure producing profit all of the time.

[3] Machine Learning (ML) – Instead of fixed programmed rules, algorithms evolve statistical models based on data that are then used to make predictions about similar data.
[4] Artificial Intelligence (AI) – Machines that perceive their environment and make decisions/actions to attain goals.
[5] FPGA (Field-Programmable Gate Array) – An integrated circuit whose hardware (memory, logic, and other circuits) can be configured, and reconfigured, by the user after manufacturing.
[6] GPU (Graphics Processing Unit) – Parallel circuitry designed for rapid image manipulation and display, but has more recently gained significant popularity for accelerating parallelized algorithms for ML and AI.
[7] TPU (Tensor Processing Unit) – A proprietary integrated circuit to accelerate parallel, low-precision computations on graphs, which is particularly well-suited to many ML and AI applications.

3. A New Generation of MRAM Intersects the Path of Composable Data Centers

From magnetic core memory in the 1950s to the magnetic tunnel junctions (MTJs) of magnetoresistive random access memory (MRAM), the roadmap for magnetic nonvolatile memory (NVM) has been long and innovative. The roadmap continues with in-plane, and now perpendicular, STT-MRAM[8]. In the future, it is anticipated that the roadmap will extend to include innovations based on spin-orbit torque (SOT) and other magnetic phenomena.

"You can have high speed, low cost, or low power – pick any two."

As each new MRAM technology emerges, there are tradeoffs – as with any technology. A common trilemma in technology has long been summed up in the saying "You can have high speed, low cost, or low power – pick any two." For example, in classic (field-switched) MRAM cells the data stored on MTJ structures are written with a stray magnetic field from current flowing through the metal-layer connections. To get strong write fields, high currents are needed (at high power). This causes failures in the metal lines due to electromigration. To scale a memory array, the cells must be smaller and closer together. The smaller MTJs of field-switched MRAM are less stable (resulting in poor retention). Their closer proximity to each other causes crosstalk. STT-MRAM provides a solution to these problems, as illustrated by the dense array of MTJs in Figure 2.

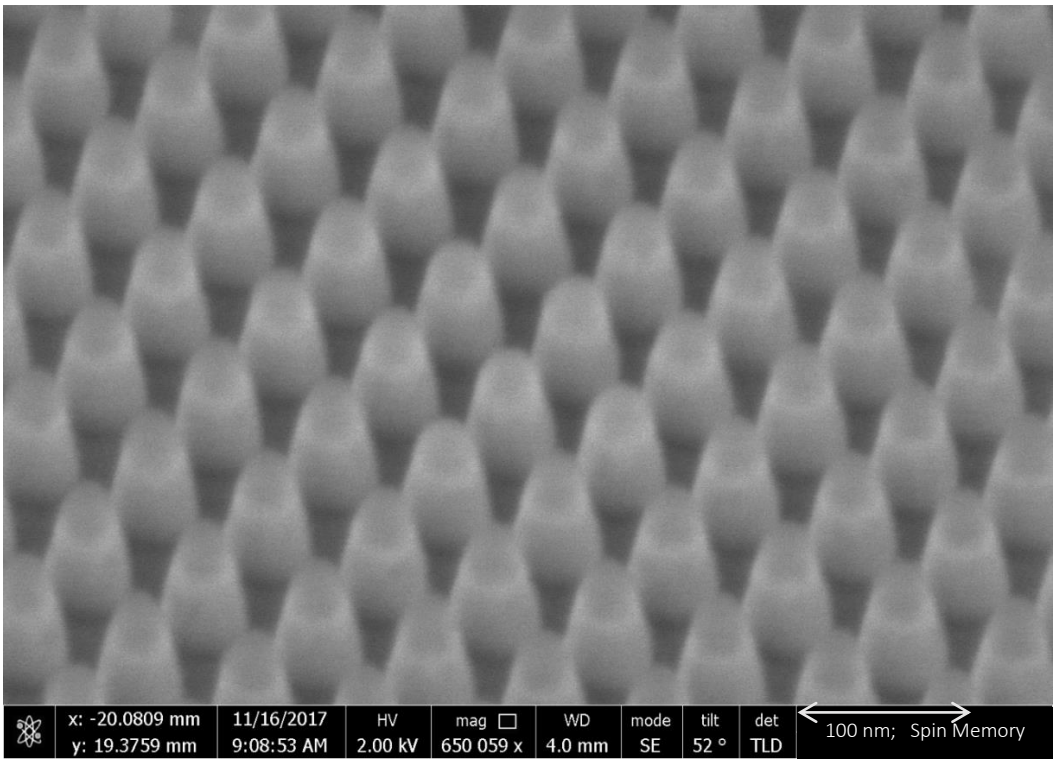


Figure 2. Array of STT-MRAM 20nm diameter pillars. Source: Spin Memory, Inc.

[8] STT-MRAM (Spin-Transfer Torque Magnetoresistive Random Access Memory) – The latest commercial MRAM technology. The transfer of electron-spin momentum between magnetic layers writes data by creating a torque on the magnetization of the free layer that establishes a parallel or antiparallel orientation between it and the magnetization of the reference layer.

3.1 Perpendicular MTJs (pMTJs)

The first STT-MRAMs used *in-plane* MTJs. That is, the magnetization of each magnetic layer lies in a plane parallel to the surface of the wafer, as shown in Figure 3a. Continued MTJ scaling reduces the volume of the magnetic material in each cell, causing retention concerns. Recently, in-plane MTJs have been replaced by perpendicular MTJs (pMTJs), whose magnetization is perpendicular to the plane of the wafer, as shown in Figure 3b. These pMTJs provide stable storage – even as process geometries shrink – because the cell's volume can increase vertically (perpendicular to the wafer's surface) to improve stability, while the cell footprint (in the plane of the wafer) continues to shrink to provide increased storage density. Of course, even for perpendicular STT-MRAMs, there are tradeoffs.

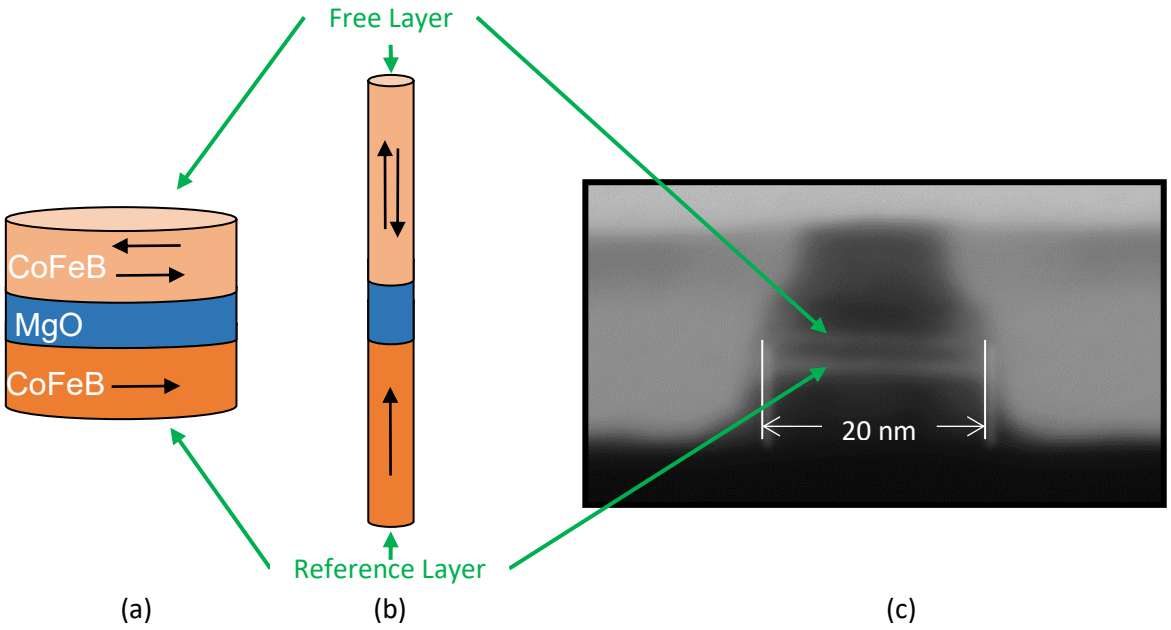


Figure 3. In-plane MTJ (a) and perpendicular MTJ (b) (Source: KnowledgeTek, Inc. [2]). (c) Photomicrograph of pMTJ stack (Source: Spin Memory, Inc.)

One important tradeoff is between write speed and endurance. In order to write an STT MTJ, the current density in the memory cell must be above a threshold. In order to ensure that all of the cells write quickly, the threshold must be surpassed by a sufficient margin. This corresponds to applying a higher write voltage to the cells. The higher voltage stresses the MgO (magnesium oxide) tunnel barrier and reduces endurance. A lower voltage slows writing and/or causes the write error rate (WER) to increase.

All STT-MRAM cells have a *probabilistic* write process.

Having a non-zero WER implies that data stored may not have been written correctly! This is a consequence of the quantum physics at the heart of the STT phenomenon. STT MTJs have a probabilistic write process. That is, the same conditions that successfully write a cell 999 times may not write successfully on the 1000th attempt. This example illustrates a WER of 10^{-3}, which is a high rate. WER can be reduced by using a higher write voltage, which can lower the endurance. WER can also be reduced by using a longer write pulse, which reduces

With STT-MRAM:
"You can have high speed, high endurance, long retention, or low power – Pick any two!"

the speed. Both methods increase power consumption – but a bit that is not written is not retained!

There is an additional way to lower the WER: By reducing the energy barrier, Δ,[9] of the MTJ. A given voltage has a higher probability of writing a cell if the cell is fabricated with a lower Δ. Alternatively, WER can be maintained while the endurance is increased if the lower Δ is paired with a lower write voltage. Although this latter approach does not negatively impact endurance directly, it does directly lower the retention. To paraphrase, "you can have high speed, high endurance, long retention, or low power, pick any two."

3.2 Spin Memory's Inventions Change the Tradeoffs

Spin Memory has invented two new technologies that enable STT-MRAMs to break through the traditional tradeoffs! The first is a Precessional Spin Current™ structure (aka PSC™) that improves writeability of the MTJ stack without the need for a higher, stress-causing, write voltage. The second is an on-chip MRAM management system called the Endurance Engine™, which consists of specially-designed circuitry that simultaneously improves speed, endurance, and retention. These breakthrough technologies change the traditional STT-MRAM tradeoffs, as shown in Figure 4, and enable this new generation of MRAM to intersect the path of composable data center development.

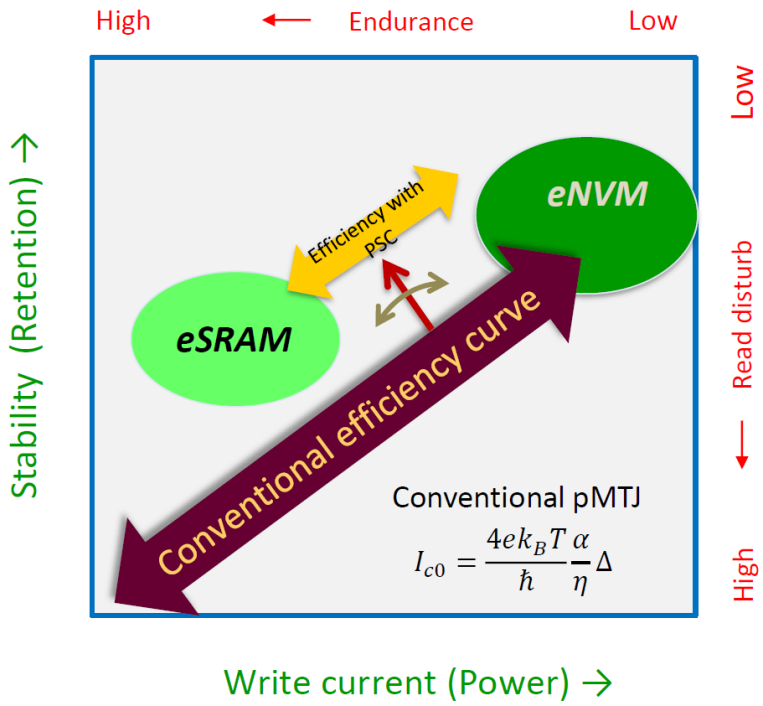


Figure 4. Traditional STT-MRAM tradeoffs, modified by Spin Memory's PSC™. Source: Spin Memory, Inc.

[9] Delta (Δ) – A unitless thermal stability factor ($\Delta=E_b/(k_BT)$). Δ is the energy barrier, scaled by the operating temperature (Kelvin) times the Boltzmann constant, which must be overcome to switch the data state of an STT structure. A higher value of Δ reflects both longer retention and more difficult writing.

What are the costs of these breakthroughs? The PSC™, as shown in Figure 5, requires a few extra seconds in the MTJ deposition tool, but no new materials or masks. Its manufacturability has been demonstrated on the latest high volume production tool from a major manufacturer. The Endurance Engine™ requires some die area in the array and power to support proprietary circuit functions. For arrays of sufficient capacity, however, the extra area is a small percentage of the memory footprint. Furthermore, because it enables the bitcell transistor's size to be reduced by 15-20%, the overall memory size will usually be reduced.

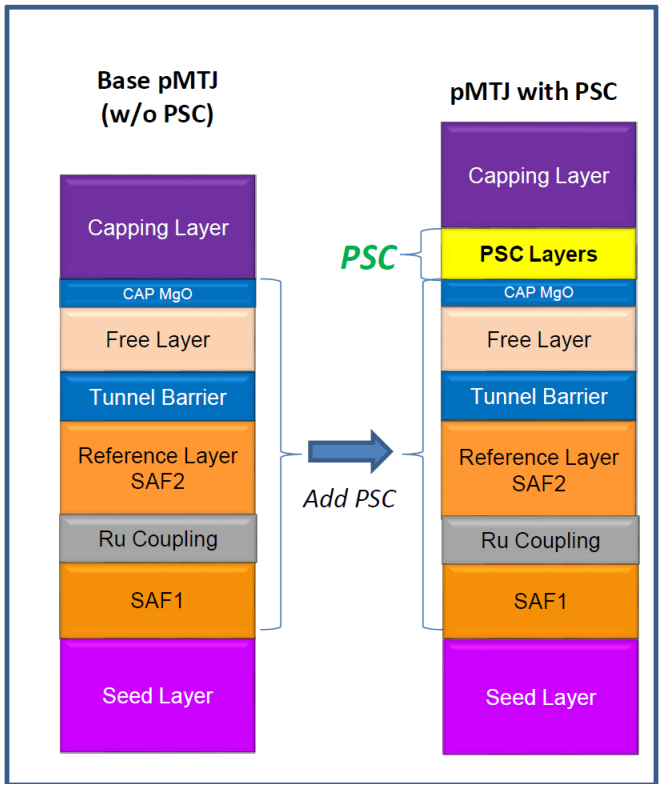


Figure 5. STT MTJ stack changes due to the PSC™. Source: Spin Memory, Inc.

These costs are well worth the result: A scalable MRAM cell that can be the replacement for DRAM and SRAM, while providing retention. This changes architectural choices in the processor, microcontroller, and throughout the data center.

With production tools available today and MRAM being offered by the major foundries, now is the time to start designing with persistent memory in mind. Now is the time to develop tradeoff-changing architectures. The early adopters of this technology will establish a competitive advantage that can put them months or years ahead of their competition.

Get a competitive advantage by adopting STT-MRAM now.

Next, we take a closer look at speed, latency, and parallelism in the data center to more clearly identify challenges in improving data center performance. Then we examine specific opportunities to replace SRAM or DRAM with STT-MRAM and gain advantage from its inherent nonvolatility in a range of data center and edge computing applications.

4. Achieving Speed in the Data Center: Specialization vs. Commoditization

Speed is often measured as throughput, the rate at which data is moved and processed through the system. Parallelism enables higher throughput … until the costs of duplicating the lanes through the system and keeping them synchronized with each other produce slower throughput than a single high-speed path. But throughput is only one important speed metric. Another is single transaction latency. Latency measures how quickly a single action can be completed, including any delay before the requested action can begin. Poor latency is sometimes buried inside great throughput specs. But it cannot be hidden for individual transactions, where it is typically the largest factor in determining the response time.

For example, although flash memory takes a relatively long time to read any particular location, throughput from a flash SSD can be very high. This is possible because thousands of bits are read in parallel from the flash array and sequential operations can be pipelined and parallelized across multiple channels. In the data center, parallelism takes this and many other forms. One example is duplicate datasets.

Datasets are duplicated for many reasons, including recoverability in the event of a data loss, coping with multiple users in need of the same data, reduced latency through closer geographic proximity, and QoS. Multiple datasets support the latency requirements of QoS agreements, for example, by enabling the host to make redundant requests for data from multiple sources. Each source of the data has a distribution of response times. Individual response times may be delayed because of intermittent internal maintenance routines or because the data source is servicing another request. The host simply accepts the first response to its request and can ignore later responses. Clearly, this is another tradeoff: Energy and volumetric density of redundant storage in exchange for recoverability and QoS.

Hyperscalers win this tradeoff by using commodity components for their infrastructure. These are low-cost, adequate performing components that are connected to massively parallel and redundant architectures according to their proprietary configuration. This composable infrastructure is managed by the hyperscaler's proprietary software.

Data center functions supporting machine learning are becoming mainstream, and must be cost-reduced.

It is always possible to spend more and create custom hardware that improves speed over what is available from commodity products. For niche applications, this can be worth the added cost. However, what were recently niche applications, such as machine learning and artificial intelligence, are becoming mainstream functions that must be cost-reduced. Customizable hardware accelerators are the devices-of-choice to support these new and changing algorithms.

The next sections identify opportunities in the composable infrastructure, and in edge computing, for STT-MRAM to improve data center performance.

4.1 Deploying "Retention-as-a-Service" (RaaS) throughout the Composable Data Center

Buying software as a product means that the purchaser is responsible for making this software work with their existing hardware. They must select the correct drivers and keep the software updated and patched. Buying software-as-a-service (SaaS) means the software is likely hosted on a remote machine and the maintenance is done for the user by the SaaS provider. Updates are done often and without user intervention.

Some may call Spin Memory's breakthrough STT-MRAM technologies enablers of "Retention-as-a-Service" (RaaS). This term indicates that retention is being provided by the MRAM itself (as a service). This is without the user's, or the system's, intervention. In particular, the data center infrastructure does not have to support persistence through millisecond-scale refresh, or uninterruptable power supplies, or adding redundant NVM on the memory bus itself.

STT-MRAM's support for "Retention-as-a-Service" can enable a level of composability never before available in the data center.

The term retention-as-a-service hints at a developing new flexibility that STT-MRAM can provide: Enabling a finer-grained composability in the data center infrastructure. For example, it can be possible to predictably tradeoff speed, endurance, power, and retention – even temperature-dependence. The retention needed for a certain function can be tuned when the device is designed, using STT's PSC™ structure and The Endurance Engine™.

A recent improvement in data center infrastructure is the development of NVDIMMs[10], which could be considered an early form of retention-as-a-service. These devices add flash and capacitors to the DRAM DIMM on the DDR bus. They typically function as a nonvolatile backup for the contents of DRAM in the event of a power failure. The onboard capacitors supply enough power to write the contents of DRAM to the onboard flash. This is an efficient use of power and space; but is it an efficient use of flash? If the flash is used only as a backup, it is under-utilized. But if it is intended to be used as part of the DDR memory pool, its inherent slowness and large page size must be managed.

Replacing this type of NVDIMM with one populated by Spin Memory's STT-MRAM enables high speed access directly to nonvolatile media. Only enough hold-up capacitance is needed to complete in-process writes. The data are nonvolatile when written. Instead of requiring the DDR bus to be powered at all times to refresh volatile DRAM, it could be power cycled as needed. The energy savings, including reduced heat generation, could be substantial depending on the workload and implementation. Because STT-MRAM needs a much smaller hold-up capacitance than flash, the time between deliberate power cycles can be much shorter.

Accelerators present another opportunity to use STT-MRAM to improve data center performance and cost effectiveness. Interestingly, machine learning algorithms are inherently probabilistic. They can likely tolerate some amount of write error, and still successfully converge. A promising opportunity requiring further study is to strategically operate the STT-MRAM at a higher WER, enabling faster training through faster writing and reading of the STT-MRAM. Furthermore, contention for the same data can be handled by pausing a colliding process, switching to another thread, and resuming the paused process all without losing the persistent state of the data. This increases utilization of the hardware and the data.

5. Another Kind of Tiering

Fully integrating retention-as-a-service will make it cost-effective to have tiered processing, not just tiered storage. Today, the main data processing is done on data that resides in DRAM and has been moved into the CPU's registers. A new category of processing is emerging for analyzing data directly on storage devices themselves. This is sometimes called *in situ* processing (computing), in-storage processing (computing), or computational storage. [2]

In between these two endpoints (CPUs and computational storage) lies processing done on accelerators. These three processing locations, plus a fourth on "cold"

[10] NVDIMM (Non-Volatile (Memory), Dual Inline Memory Module) – Nonvolatile memory placed on the DRAM (DDR) bus.

data[11], can be viewed as processing tiers. This concept is facilitated and optimized by the use of STT-MRAM, and is illustrated in Figure 6.

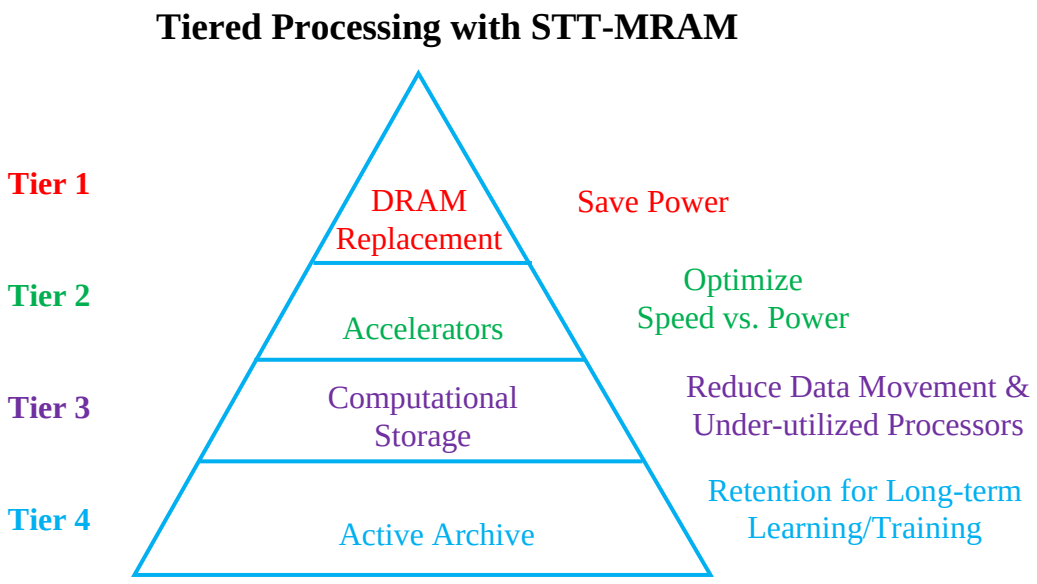


Figure 6. A new type of tiering for the data center: Processing tiers.

When STT-MRAM is in high volume production, it can support the first tier by eliminating the power typically needed for refreshing DRAM data. Some researchers estimate DRAM refresh cycles account for 35 to 45% of the energy consumed by DRAM. The percentage grows as DRAM array capacity grows. Even before STT-MRAM becomes a cost-effective replacement for standalone DRAM, expanding the amount of on-chip memory capacity with STT-MRAM will save power. Moving data to and from DRAM is estimated to consume about 60% of the system energy. Replacing on-chip SRAM with larger capacity, but smaller size, STT-MRAM allows more data to remain local. Algorithms that require repeated access to the same data will consume less power by minimizing redundant data moves from off-chip memory.

Off-chip data access can consume 30X more energy than on-chip buffer access.

The second tier of processing includes composable accelerators for machine learning algorithms and artificial intelligence implementations. These purpose-built processors will have larger on-chip memory requirements than typical CPU cores. This is because the repetitive nature of the algorithms run faster and consume less power when redundant calls for off-chip data are reduced, as also identified for tier 1 processing. For certain machine learning algorithms, off-chip data access is estimated to consume 30 times more energy than on-chip buffer access. Furthermore, because of the repetitive nature of the algorithms, larger memories enable some calculations to consume significantly less energy. One estimate shows calculations performed on an accelerator with 100 Mbit of on-chip memory take 1/10th the energy required if only 1 Mbit of on-chip memory is available.

[11] Cold data – Data stored on media that are not as quickly accessed as other tiers of storage or memory. Examples of cold storage are tape, spun-down hard disk drives, or slow or sleeping flash devices.

The third tier implements *in situ* processing of data, where the data storage device becomes a computational storage device. This emerging class of product uses its processing capability to analyze new (or existing) data that is stored on itself. The main benefit is the savings in network traffic, and power, from eliminating the need to move data from the storage device to an external processor. The computational storage device may be designed, for example, to perform searches on its data. Instead of sending its data to a processor that will search it for a match to a picture, the host can send the picture to the computational storage device. The device will search itself and will only respond over the network with the results of its extensive search. Similar to tier 2 accelerators, the processors in computational storage devices can benefit from larger on-chip buffers. In 24nm CMOS processes and below, embedded STT-MRAM arrays can be ½ to 1/5th the size of embedded SRAM arrays. The larger the capacity, the bigger the area savings.

Embedded MRAM can be 1/5th the size of embedded SRAM arrays for small process geometries.

Even data from a cold storage (fourth) tier can be scheduled for cost-effective analysis in an active archive. One way a data center can do this is to implement longer-retention, slower-writing pools of STT-MRAM and strategically place seldom accessed data on them. Portions of the pool will be analyzed only when processors would otherwise be idle. This can allow value to be extracted from low-utilization data as new algorithms are developed and deployed. Of course, the value extracted from the data must be sufficient to profitably cover the cost of this processing tier. Economies of scale do not yet exist to support a large capacity pool. Ultimately, it may still be economically necessary to have a final dark (fifth) tier for archiving only. Even this tier's value can be enhanced by its metadata being available on STT-MRAM while the rest of the storage is completely powered down.

6. The Expanding Data Center: Data Lakes and Edge Computing

FPGAs, GPUs, and TPUs are better suited to specific classes of algorithms, such as machine learning, than are general purpose CPUs. As new algorithms are designed – by humans, and by other algorithms – the optimum hardware configuration for each is likely to be novel. A composable infrastructure that includes composable accelerators is an obvious design goal for a data center.

6.1 Data Lakes

However, the accelerators must pay for themselves by providing high-profit services at high-utilization. It is likely that multiple algorithms may want to access to the same data. For example, the concept of data lakes is developing, where a company or institution provides access to a large amount of specialized data (e.g., 3D seismic data, weather history, satellite imagery) and many algorithms from different companies may pay to mine it for new discoveries. A data lake that is just sitting provides little value. It must be accessed, shared, analyzed, and cross-referenced to provide insight, inference, and direction.

To support this and other concepts, the architecture of computing is undergoing a transformation from compute-centric to data-centric systems. In these, a large pool of interconnected memory is shared by many processors. In such an arrangement, all data are accessible to all processors. But it is likely that some

data will be of more immediate relevance, and collisions for access to it could occur. In this case, a processor must wait idle, or be able to switch to another thread[12], return when the data is ready, and pick up exactly where it left off.

STT-MRAM provides the robust way to support this environment and have composable building blocks as part of the accelerators themselves. This is in addition to the benefits of deploying STT-MRAM in a persistent memory computing environment overall.

6.2 Edge Computing

Another method for increasing the performance of the data center is to minimize the amount of data sent to it for low-value processing. An increasingly popular way to do this, while keeping the data center's infrastructure working on higher-value processing, is to make extensive use of edge computing. Edge computing aggregates local data, preprocesses it, and only sends the essential traffic to the cloud. Edge computing is seen as a key enabler for many IoT applications.

Edge computing can enhance or erode privacy.

An interesting additional feature of edge computing is that it can enhance or erode privacy and security, depending on how it is implemented. For example, a camera and microphone in an office that broadcast their information to the cloud can reveal personal activity moment-by-moment. However, an edge device that interprets the sensors' output and transmits to the cloud only that a particular meeting room is occupied preserves much more anonymity and privacy.

It is expected that edge computing will be essential for optimizing autonomous driving in dense environments. Some refer to the cars themselves as "data centers on wheels" due to the massive volume of data they generate, as shown in Figure 7. There will be edge devices in the vehicles, in the traffic control infrastructure, and for local information fusion. Only the relevant information for improving the learning of the fleet needs to be sent to the cloud for storage and analysis.

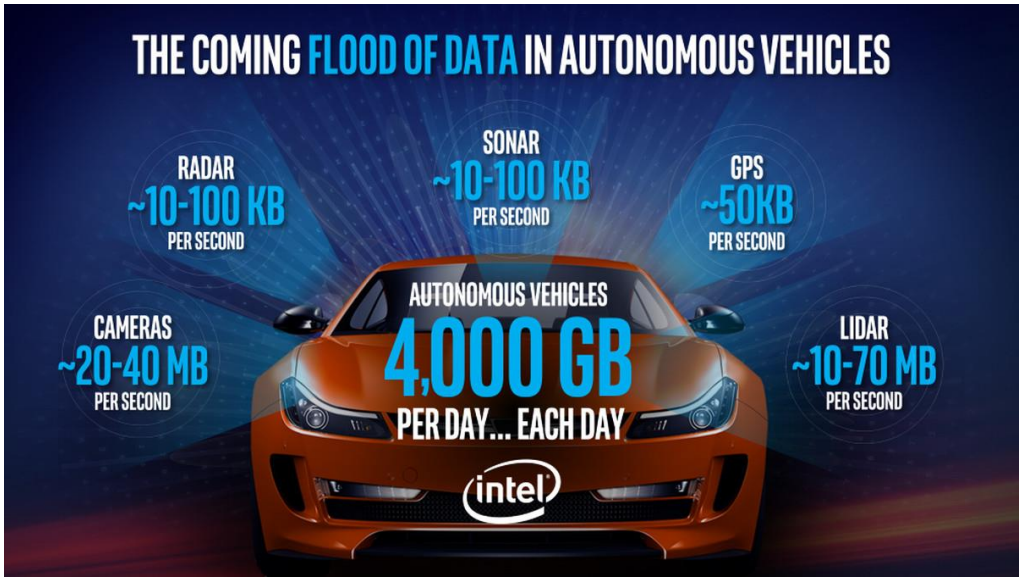


Figure 7. "Data is the new oil." Source Intel [3]

[12] Thread – A small, independent sequence of computer instructions.

Having STT-MRAM in the IoT devices that feed edge computing, and in the edge computing devices themselves, will enable low-power mobile devices and long-lasting energy harvesting sensors to be widely deployed as a distributed data center. The distribution begins at sensors and mobile devices; passes through edge computing; is stored, processed, and utilized in the cloud; and extends back down to the individual user in the street, home, field, or factory.

8. How to Start and What to Watch

Profitability requires experience and large-scale production of STT-MRAM.

To profitably enable the scenarios above, experience with STT-MRAM use and economies of scale for its manufacture must be developed. This means that early adopters can create a competitive advantage by moving quickly to benefit from STT-MRAM incrementally. There are places to innovate first to get the most advantage fast. Here are four strategies to consider:

1. Replace standalone NVM chips, such as NOR flash for booting or journaling, with STT-MRAM. This is a low risk option, but it is likely to have a higher cost if no other changes are made to the system. To mitigate this, one approach is to amortize the higher initial cost by using a larger memory. The larger memory can, for example, store different firmware-customizable versions of your product instead of having to manage them as separate physical products.

2. Embed STT-MRAM to replace embedded NOR flash in your current designs. This is more costly if used as a direct replacement, but it provides a cost reduction if the resulting design can use a smaller geometry process and hence occupies less die area.

3. Embed an array of STT-MRAM in an existing design, to replace embedded SRAM or DRAM. This has the same benefits as the strategy above, and it provides a low risk way to gain experience with the benefits of persistence in your system. The experience gained can be utilized by products already deployed in the field via firmware updates.

4. Replace a battery-backed or capacitor-backed system with sufficiently-persistent, strategically located, standalone STT-MRAM. The savings should be evident at the design phase.

The time is NOW to gain experience with STT-MRAM designs and persistent memory programing.

The strategies above are low-risk, incremental entry points for using persistent memory in standalone and embedded applications. But for companies starting a new design of a new product – or a new product category – STT-MRAM can be designed in from the beginning to give the product an edge over competitors who are using traditional architectures to develop a similar product.

For example, new data-centric architectures will bring processing to the data. Such *in situ* computing can take many forms. Some will simply add more processing power to an existing NVMe SSD design. Some may add more or less DRAM to their drives to increase speed. However, it is possible to employ a tactic of data centers themselves and use storage tiering within the drive by replacing or augmenting DRAM with STT-MRAM. This can provide a performance, endurance, and $/GB advantage over other architectures.

Early adopters of persistent memory will gain valuable experience on how to mitigate unique security risks.

Early adopters of persistent memory will gain crucial experience regarding possible unintended consequences of using it in a composable infrastructure. For example, when persistent memory resources are returned to the composable infrastructure pool for reuse, they may retain proprietary data from the previous user. Experience and knowhow for identifying and mitigating this novel security threat is likely to be a key competitive advantage for these early adopters.

5G and AI will change our expectations!

Another area to watch that has the potential to change our expectations of mobile connectivity is the deployment of 5G cellular systems. These are designed to have such low latency that they enable the tactile internet. The vision of the tactile internet is to deliver the effect of skilled labor as easily as today's internet delivers information, documents, sounds, and images. It is planned that a person can operate equipment from a remote location and get the tactile feedback from their actions within the typical human reaction time. This can enable tele-surgery or new levels of virtual reality experiences. To achieve this level of performance, every point in the link must provide the lowest latency possible. STT-MRAM is a great choice for fast reads and pre-loading of information for low power storage in these systems. 5G will also demand low latency cloud computing support from data centers.

The STT-MRAM benefits of power, cost, and time savings will make its early adoption attractive to hyperscale data centers. Its retention will make it a memory of choice for IoT devices, edge computing, and low latency communication systems. The combination of these broad, but related, uses can drive economies of scale for STT-MRAM sooner than any previous memory or storage technology.

It took about 30 years for flash to catch hold and begin to grow in computing. MRAM, too, has been available for many years. But Spin Memory's breakthrough combination of PSC™ and the Endurance Engine™ enable current and future generations of perpendicular STT-MRAM to achieve the previously elusive combination of high speed, low power, endurance, and retention. It is the new memory for a new era.

For more information, contact info@spinmemory.com.

9. References

[1] IARPA, 2018. [Online]. Available: https://www.iarpa.gov/images/files/programs/mist/MIST_proposers_day_briefing.pdf.

[2] SNIA, "Computational Storage Technical Working Group," [Online]. Available: https://www.snia.org/computational. [Accessed 27 March 2019].

[3] B. Krzanich, "Data is the New Oil in the Future of Automated Driving," Intel, 15 November 2016. [Online]. Available: https://simplecore.intel.com/newsroom/wp-content/uploads/sites/11/2016/11/Automobility-2-small.png.

[4] C. H. Sobey, "New NVM Technologies," KnowledgeTek, Inc., 2019. [Online]. Available: http://www.knowledgetek.com/%20new-volatile-memory-technologies-nvm-1-day-1195%20/.

Plano TX 75024 USA
connect@ChannelScience.com
(972) 814-3441 Voice / Text
@ChannelScience

About Spin Memory, Inc.

Spin Memory (previously Spin Transfer Technologies) is dedicated to solving the scaling and power problems of today's memories. To meet these goals, Spin Memory is collaborating with world leaders to transform the semiconductor industry by offering MRAM solutions to replace on-chip SRAM, stand-alone persistent memories, and a range of other non-volatile memories.

Spin Memory is developing spin-transfer torque (STT)-MRAM technologies and products that can replace SRAM (static RAM) and ultimately DRAM (dynamic RAM) in both embedded and stand-alone applications. Already, Spin Memory has developed breakthrough technologies in both magnetics and CMOS circuits and architectures that bring STT-MRAM to the next generation. The result: MRAM operating speeds matching those of SRAM cache memories or DRAM — but with far lower cost, no leakage power and without the endurance and data retention limitations of other STT-MRAM implementations.

As the pre-eminent MRAM IP supplier, Spin Memory is uniquely positioned to offer the industry's highest-performance, highest-density STT-MRAM memories. Contact us at info@spinmemory.com.

About the Author

Charles ("Chuck") H. Sobey founded ChannelScience in 1996 and has grown it into the top technical consultancy for emerging nonvolatile memory (NVM) and data storage technology innovations. He is a senior member of the IEEE and General Chairman of the world's #1 independent storage conference, **Flash Memory Summit**.

Chuck's technical expertise, industry connections, and clear communication help clients define and execute strategies that develop new lines of business and "Establish the State-of-the-Art!"™ His primary technical focus is analyzing signal, noise, distortion, and defects in new NVMs – at the bit-level and below. Using extensive probability analysis, he develops signal processing, coding, architectures, and applications that optimize performance and enable profitable mass production.

Chuck's guidance has been sought by governments on four continents to help them develop plans for the strategically important areas of memory and storage. He is also well-known for teaching storage technology to literally thousands of experienced engineers, scientists, technicians, and executives. He is currently accelerating AI/ML performance and spreading its application, including developing standards for AI agents.

You can work with Chuck through the following organizations.
www.ChannelScience.com, www.KnowledgeTek.com, or www.FlashMemorySummit.com

You can connect with him directly, or through social media.
www.LinkedIn.com/in/ChuckSobey, www.Twitter.com/ChannelScience @ChannelScience

Please Note

This white paper is for information only and is provided as is. The reader is encouraged to refer to original sources and to check the patent ownership of ideas or technology presented herein before using any of the information provided. Not every architecture will be successful, regardless of the company or method used. Neither the author nor Spin Memory, Inc. is responsible for any direct or consequential losses due to the use of the information contained herein.