OBJECTIVE ANALYSIS
Semiconductor Market Research

NEW MEMORIES FOR EFFICIENT COMPUTING

Reducing Energy Consumption in Battery and Large-Scale Systems

OBJECTIVE ANALYSIS WHITE PAPER

The semiconductor industry is at a turning point. Both the embedded memories in microcontrollers (MCUs) and ASICs and the external stand-alone memory chips that are used in everything from handheld devices to supercomputers are being considered for replacement.

In many cases this replacement will help the system designer reduce power consumption to extend battery life or reduce cooling requirements. In other cases replacing the conventional memory types will provide system cost savings, perhaps by allowing a more aggressive process technology to be used, or by improving the system's overall cost/performance.

In this white paper Objective Analysis reviews both the embedded memories used in battery-operated systems and the discrete memories used in systems ranging up to hyperscale data centers, and will illustrate why new technologies can benefit both types of systems. After this we will review the leading emerging memory technologies and provide the reader with an understanding of the critical success factors that may lead to the selection of one technology over the other.

Although several emerging memory technologies have been developed to address this need, few of them will succeed in this competitive market. A list of a number of these technologies is shown in Figure 1.

Figure 1.) New Memory Types

Acronyms	Name
PCM or PRAM	Phase-Change Memory
MRAM	Magnetic Random-Access Memory
STT	Spin-Tunnel Torque, a type of MRAM
FRAM or FeRAM	Ferroelectric Memory
RRAM or ReRAM	Resistive RAM, an umbrella category
CBRAM	Conductive Bridge ReRAM
OxRAM	Oxygen Vacancy ReRAM
Crossbar	Crossbar metal-filament ReRAM

Source: Objective Analysis, 2018

No matter which one wins, though, it is certain that power consumption will be lower in systems based on one of these emerging nonvolatile technologies than it is in today's systems that are based on embedded NOR flash and SRAM, or discrete DRAM and NAND flash.

Issues with Embedded Memories

Leading-edge logic processes have moved beyond 14nm, migrating to FinFET structures in the process, and the embedded NOR flash that has been used as on-chip storage for the past decade or more has lost the ability to keep pace with these process shrinks. This issue is referred to as flash's "Scaling Limit" – no matter how much the rest of the CMOS on the chip is able to shrink, the flash can't keep pace. A new embedded memory technology must be used to store firmware code and data on ASICs and MCUs that are produced on these advanced process nodes.

Embedded NOR flash is not alone. Its counterpart, embedded SRAM, is facing a related issue. As processes shrink into the tens of nanometers and smaller, the size of an SRAM bit does not keep pace. Unlike NOR, SRAM's

size does not remain the same for shrinking processes, but SRAM does not shrink in proportion to the process. It may shrink only 25% when a process shrinks by 50%.

This paves the path for both embedded NOR and embedded SRAM to be replaced by a technology that will continue to shrink in proportion to the process. Fortunately, such technologies exist, and have been in development for a number of years.

Another issue provides a strong argument to move to new memory technologies. Memory consumes energy, and mobile device designers are keenly aware of the energy it consumes. At the Edge, the Internet of Things and consumer mobile devices run on batteries and memory systems must be carefully chosen since they consume most of the battery's energy, limiting battery life. Power consumption can be reduced by converting designs to a new embedded memory technology.

Next-generation mobile architectures will integrate higher computing requirements for Artificial Intelligence at the Edge while demanding lower energy consumption to satisfy end-user expectations and win over competition. All of this must be achieved at a low cost, which is often a challenge with existing memory technologies. The MCUs that run most of today's battery-operated devices and ASICs that are used in a wide range of applications are built using CMOS processes that support two memory technologies: NOR flash and SRAM. While these technologies are readily available in CMOS logic processes, they often consume more power than desired.

When larger memories are required, designers will usually add external memory chips like SPI NOR flash, NAND flash, DRAM, or some mix of these technologies. These external-memory solutions impact power consumption even more. Designers are starting to evaluate emerging memory technologies to try and solve this problem.

Power Issues in Larger Systems

At the other end of the Internet of Things, in the Cloud, the memory and data storage architectures of the servers in data centers are also extremely important, since power consumption is often one of the highest cost elements in a data center, especially when cooling is included.

DRAM and NAND flash are today's prevailing memory technologies for computing systems, from smart phones to data processing equipment. Neither memory type is sufficient for a system design by itself. DRAM supports fast reads and writes but consumes significant power for its refresh cycles, since its bits decay after few milliseconds. DRAM constantly uses power to refresh, even when the device is idle. Roughly 20% of the power consumed by an 8Gb DRAM chip is spent on refresh, or 25 milliwatts of the chip's total 140-milliwatt power consumption.[1] If power is removed then the contents of the DRAM are lost – bits will be in a random state once the power is restored, rendering DRAM unsuitable for code storage.

DRAM is also relatively slow, thanks to its multiplexed address bus. DRAM Row (RAS) and Column (CAS) strobes cause random reads to take from 25ns to 300ns, and this extended time results in higher overall energy consumption.

Flash bits don't decay, and maintain their contents for years after power is removed, but these bits are either more expensive than DRAM, in the case of NOR flash, or they stream out of the part sequentially in the case of NAND flash. Sequential data is a mismatch for the random-addressing needs of computer software. NAND flash must be paired with a DRAM to be used for code storage.

Like DRAM, NAND flash also has idiosyncrasies that lead it to consume more power than desired. For one, it requires high internal voltages that are generated using inefficient

[1] RAIDR: Retention-Aware Intelligent DRAM Refresh, Liu, Jaiyen, Veras, Mutlu, Carnegie Mellon University

on-chip charge pumps. NAND is also very slow to write. On top of that, a blank page must be available for a write operation – data in a flash chip cannot be overwritten. This means that flash must be erased before new data is written into a page, and an entire page (typically 8,096 bytes) must be written at a time. Flash technology doesn't use same mechanism to program or erase bits: you can't erase a single bit or a single page. Erase operations occur only on a block, which usually contains hundreds of thousands of pages. A page write is a slow and energy-intensive process, typically taking 300µs (microseconds) to achieve and consuming 80 microjoules in the process (compared to 2 microjoules for a read.) A block erase, which also requires high internal voltages, takes even longer, typically 2 milliseconds and consumes 150 microjoules of energy.[1] In return for all this complexity the system designer gets very inexpensive storage, so designers are willing to work around NAND's intricate write process and high energy consumption to take advantage of its low cost.

Most smart phones and computing systems use a mix of DRAM and NAND flash for their memory and storage needs: In a smart phone the DRAM holds copies of the programs for execution while the phone is turned on, and the NAND stores the programs when the power is off along with photos, videos, music, and other less speed-sensitive data. A server will store programs and data in its DRAM main memory, using flash-based SSDs for its long-term and backup storage. Smaller systems may use NOR flash instead of NAND and SRAM instead of DRAM, but only if their memory needs are very small; NOR costs one or two orders of magnitude more per byte than does NAND flash, and SRAM's cost is a couple of orders of magnitude larger than that of DRAM.

[1] Modeling Power Consumption of NAND Flash Memories using *FlashPower*, Mohan, Bunker, Grupp, *et al*, IEEE

How New Memories Solve the Problem

The power consumption problem with today's memories stems from issues that don't even exist in the many emerging memory technologies that are in development today. These emerging memories are all nonvolatile, so there's no need to refresh them. This automatically reduces their power consumption by 20% over that of DRAM. Since all of them can over-write old data without erasing they save the high erase energy consumption required for flash, and the delays incurred by the slow erase cycle. (This attribute is known as *in-situ* programming.) Write energy requirements for these new technologies are extraordinarily low compared to those of flash, reducing or removing the need for an inefficient charge pump. Finally, all of these new technologies provide random data access, alleviating the need for two copies to be kept – one in flash and the other in DRAM.

It need not be said that all of these attributes will lead to important power savings whenever any emerging memory technology is used to replace today's conventional DRAM + NAND flash memory architecture.

Let's review the more mainstream of these technologies to understand them better.

Some New Memory Types

Most emerging memory technologies share certain attributes:

- All of them are nonvolatile or persistent, a decided strength against DRAM with its power-hungry need to be refreshed at regular intervals
- None of them requires the high Erase/Write voltages that flash needs
- None of them uses the clumsy Block Erase / Page Write approach required by flash memory (NAND

and NOR) thereby dramatically reducing write energy requirements while improving write speed
- Some of them allow cost reductions through scaling that surpasses that of today's entrenched memory technologies: DRAM and flash

Selectors

One important difference between many of these memory types is how they are addressed, which is through a bit selector. For some the selector is a transistor, which limits how tiny a memory cell can be made. Others use a diode or other two-terminal selector device, which shrinks the size of the bit and helps allow the memory bits to be stacked into a 3D array.

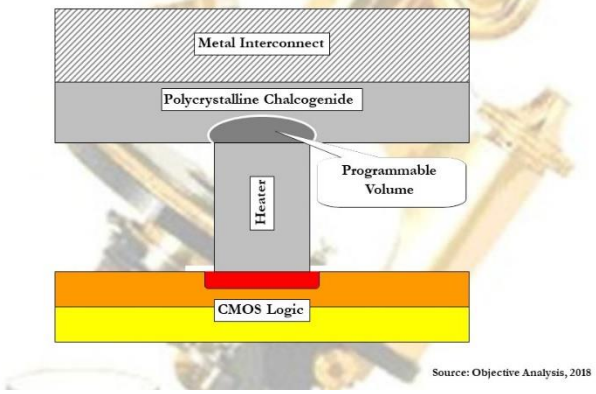
The selector type impacts the cost of the memory technology and can be a source of difficulty in producing the device.

Two-terminal selector cells can attain the ideal $4f^2$ cell area, that is, the cell is as large as the square of the twice minimum feature size "f" of the chip manufacturing process. On a 14nm process that number would be 2x14nm x 2x14nm, or $4x(14nm)^2$. A $4f^2$ cell area is the smallest that any memory cell can be made. A transistor-based cell is typically $8f^2$ but, in certain cases, can shrink to as small as $6f^2$.

Cells based on a two-terminal selector have another advantage that they can be stacked for further cost reductions. So far no company has attempted to stack transistor-based cells.

There are two types of two-terminal selectors: simple diodes, and bidirectional selectors. Of the two, a diode is significantly easier to design.

Figure 2.) Phase-Change Memory



PCM

Phase Change Memory (PCM or PRAM, illustrated in Figure 2) has been investigated for decades. Intel co-founder Gordon Moore published a paper describing an early prototype in 1970.

The technology is based on a chalcogenide glass that is melted above 200°C, then cooled either slowly, which results in a crystalline conductive state, or quickly, to provide a nonconductive amorphous state. The current moves in the same direction for both the set and reset functions, allowing a simple diode to be used for the selector device. This makes the bit cell easier to produce since diodes are better understood than bidirectional mechanisms. Since the cell is built above the CMOS logic and has been designed for stacking, new materials must be used for the selector diodes, rather than to build them in the underlying CMOS. This adds to the number of layers used to produce the bit, and increases the wafer cost accordingly.

The first commercially-available PCM chips were produced by Intel and Samsung in 2006. While the Intel chip was produced for a number of years the Samsung device appeared for less than a year in a single Samsung cell phone model prior to being discontinued.

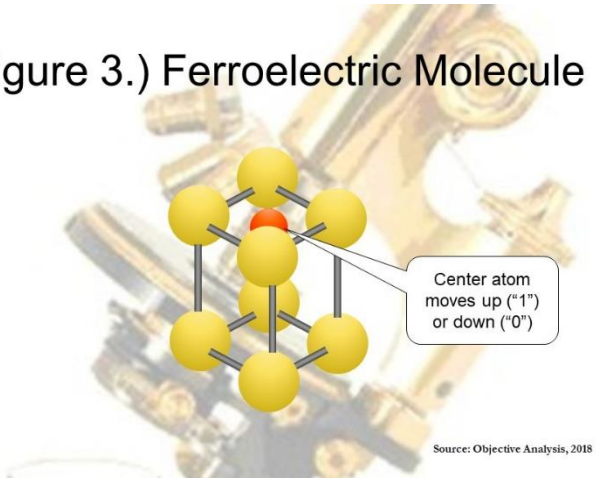
In 2015 Intel and Micron revealed plans to produce something these companies named "3D XPoint Memory," with the "XPoint" being pronounced as "Crosspoint." This PCM-based product is intended to serve as an additional memory layer fit between the DRAM main memory and NAND flash SSD in computing systems.

FRAM

Ferroelectric memories, known as FRAM or FeRAM, were introduced around 1987, but did not become commercially available until the middle 1990s.

Despite the name, FRAM uses no ferroelectric materials. The name stems from the fact that the behavior of the bit storage mechanism, a molecule depicted in Figure 3, resembles that of ferromagnetic storage: The voltage-current relationship has a characteristic hysteresis loop that can be used to store bits. A positive current will leave the bit cell in a state with a positive bias when it is removed, and a negative current changes that bit cell's state to a negative bias. The ferroelectric bit cell uses a crystal for storage that has an atom in the center. This atom rests either toward the top or the bottom of the crystal. The bit storage is a function of the position of this atom.

Figure 3.) Ferroelectric Molecule



One unfortunate fact of this is that a read is destructive – every read must be offset by a subsequent write to restore the contents of the bit to its original state. This is not only time consuming, but it also doubles the power consumed by a read cycle, a concern in power-sensitive applications.

Two companies, Ramtron and Symetrix, championed FRAMs based upon different materials. Ramtron's material, lead zirconium titanate, or "PZT" was unpopular in semiconductor fabs thanks to lead's penchant for contaminating silicon through its high ion mobility. Symetrix's proprietary material, although more complex, suffered from similar issues.

In the 1990s Ramtron's partner Fujitsu ramped an embedded FRAM into high volume production in a chip for subway fare cards. FRAM was selected because of its uniquely low write energy, allowing an interrogating radio signal to power both data reads and writes without any other power source.

FRAM continues to attract R&D investments. In 2011 the Fraunhaufer Institute's NaMLabs in Germany found that a commonly-used semiconductor material, hafnium oxide (HfO_2), could be used for ferroelectric memory. Although it is very early in the life of FRAMs based on this material, the fact that HfO_2 is well understood and accepted by the semiconductor processing community gives it promise of future acceptance.

Today's FRAMs are based on a two-transistor, two-resistor cell (2T2R) making them at least twice the size of a DRAM bit cell. A 1T1R cell is in development. Only after this is accomplished can FRAM costs be brought anywhere close to the cost of DRAM.

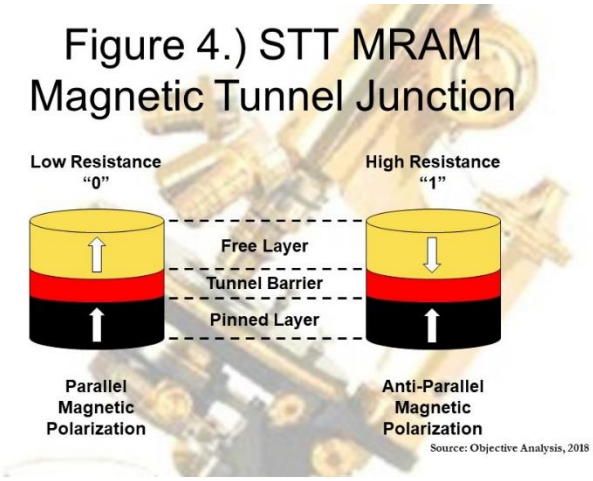
MRAM

Magnetic RAM or MRAM is a natural offshoot of magnetic recording technology. In fact, MRAM most resembles the core memory of early computers that was replaced by SRAM, then DRAM in the 1970s.

The original MRAMs, called "Toggle MRAM" or "SRAM-type" worked similarly to cores by magnetizing and demagnetizing bits and reading them by forcing them into a different state. The currents required to do this were manageable until about the 75nm process node, then became

unmanageably high, since the current remained the same as the conductors shrank, causing intolerably high current densities. This led researchers to try new approaches, starting with Spin Torque Tunneling (STT, shown in Figure 4), then perpendicular spin torque (pSST), and now migrating to Spin Orbit Tunneling (SOT).

All of these devices use the "Giant Magnetoresistive Effect" of a tunneling layer to read the bit cell: When magnets on either side of this layer are aligned the layer provides a low resistance to current, but when the magnets point in opposite directions the current flow is interrupted. This topology, shown in the figure, requires a stack of three or more layers to implement: The two magnetic layers and the tunnel layer.



There are two kinds of STT MRAM, one with a smaller, but slower, single-transistor cell and one with a larger, but faster, two-transistor cell.

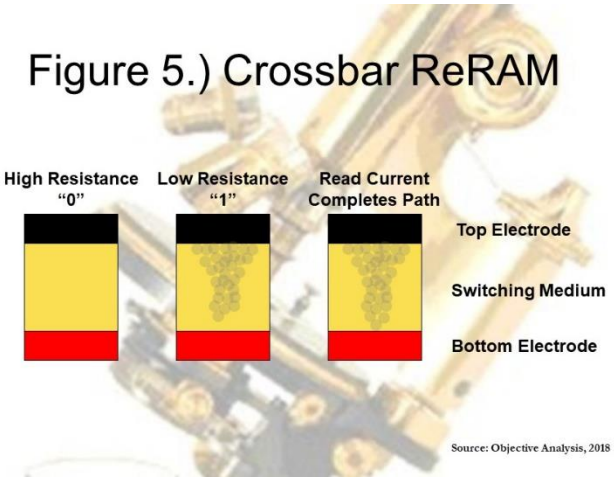
The smaller single-transistor STT MRAM cell requires one transistor and one magnetic tunnel junction per cell (called "1T1R"), allowing it to reach a die size equivalent to that of a DRAM, and has a relatively slow write cycle of 200ns.

For faster SRAM-like write speeds designers can implement a cell with two transistors (called "2T2R") to support high-speed differential sensing. This more than doubles the die size of the MRAM, though, increasing its cost significantly.

MRAM has received significant attention lately. This may stem from the fact that Everspin and Global Foundries are jointly promoting MRAM, while several other would-be producers, and most foundries are researching this technology.

ReRAM

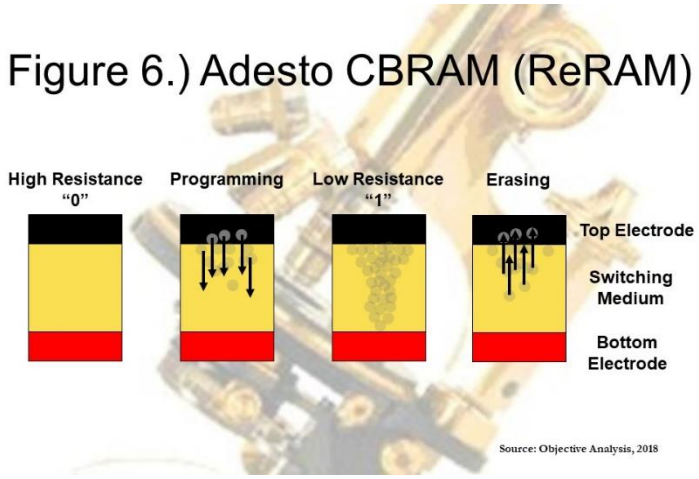
Many different technologies fall into the Resistive RAM (ReRAM or RRAM) category. Among these are Oxygen Vacancy Memories, Conductive Bridge Memories, Metal Ion Memories, Memristors, and even Carbon Nanotubes. Some even say that PCM should be included in this category. What all of these technologies have in common is that the memory mechanism consists of a resistor that is either in a high-resistance or a low-resistance state to represent a "1" or a "0." A current flows through the resistor to read it and a higher current is used to over-write it.



ReRAMs all promise to simplify and shrink the memory cell thanks to the fact that they do not necessarily use a transistor as a selector, instead employing a two-terminal selector that can be built above or below the bit cell. Not only should this reduce the cell to its theoretical minimum size of $4f^2$ but it also allows cells to be stacked vertically, greatly

increasing chip densities and possibly reducing costs.

The only ReRAMs that are currently available for production are Crossbar's 40nm ReRAM memory (shown in Figure 5), Adesto's 130nm CBRAM (Figure 6.), and the Panasonic 130nm TaOx ReRAM (Figure 7.) The Crossbar and Panasonic processes are already available in semiconductor foundries to be embedded with CMOS logic, for example: in an MCU while Adesto's is currently produced as discrete chips.



In the Crossbar memory cell, nano-metal filaments smaller than 5nm wide bridge the insulator to short circuit it.

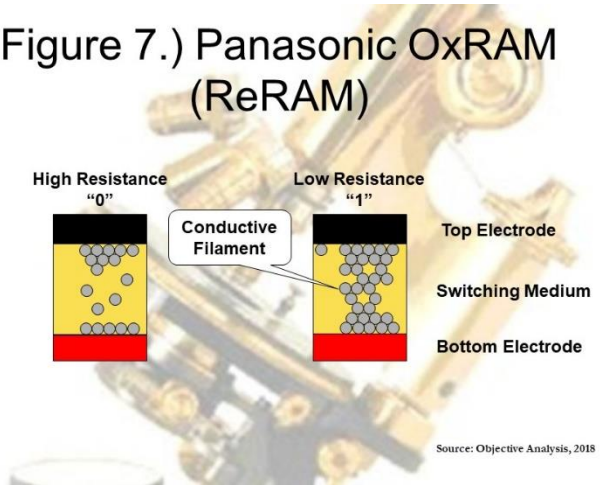
Crossbar's technology is being introduced both as a stand-alone memory and as an embedded memory process starting 40nm, and one licensee has begun development of a 1Xnm embedded design.

The Adesto CBRAM (Conductive Bridge RAM) was originally based on a metal bridge in a chalcogenide glass insulator between two electrodes, a process that has since been replaced by a different conductive bridge based on metal oxides that are easier to manage in a semiconductor production facility. A positive current creates a metal bridge across the insulator that conducts current, and a reverse current through the bit cell breaks the bridge and removes the current path. This conductive bridge is only a few atoms wide.

A difference between the Crossbar cell and the Adesto cell is that the bridge in the Crossbar cell is not fully formed until the read current passes through it. This makes the cell perform as a diode, thus eliminating the need for a select device. This vastly simplifies the production process, since the cell consists of a single combined selector + bit cell while other ReRAM technologies require a selector that is separate from the bit cell.

Panasonic introduced Oxygen Vacancy based ReRAM integrated into its 130nm MCU product line and in 2017 announced a partnership with foundry UMC to jointly develop a 40nm process.

Such memories have been based on metal oxides including those of tantalum and hafnium. These memories use high voltages to move oxygen atoms into or out of the glass bit cell to form or remove a conductive path.



Another Oxygen Vacancy ReRAM technology was developed by HPE and given the name: "Memristor." Current research has produced four-transistor cells (4T2R) and one transistor cells (1T1R) but no recent announcements have been made by HPE.

Another interesting ReRAM type that has not been sampled yet is carbon nanotube memory. This memory promises to allow a simple

slurry of carbon nanotubes (CNT) to be spun onto a wafer and processed to provide a very energy efficient memory with a small cell size. Although details of this technology are not widely available, it is certainly an intriguing approach to the problem and an interesting and unconventional manufacturing technique. The CNT memory is being championed by Nantero, and has been licensed to Fujitsu as a possible successor to FRAM.

Comparing Emerging Memory Technologies

Table 1 compares the devices reviewed in this white paper.

Perhaps the most important factor in this comparison is the cell size, since that determines cost. Cost is enormously important in the memory technology selection process – more expensive technologies are usually replaced by lower-cost technologies, even if this change requires significant work-arounds.

Selector Type

Note that the technologies that use a transistor as their select mechanism (denoted as 1T1R or 2T2R in the table) have larger bit cells than those with two-terminal bidirectional selectors (the "S" in the term "1S1R") or diodes (1D1R). The 1TnR term in the Crossbar column refers to the fact that this company's cell has a built-in selection device which behaves like an internal diode. A transistor is required for every group of "n" cells, adding only fractionally to the effective size of the cell.

Although the creation of the selector adds some complexity to wafer processing, complexity that adds marginally to the cost of the wafer, a larger cell has a more profound impact on the cost of the memory, with an $8f^2$ cell consuming twice the area of a $4f^2$ cell, and FRAM's $30f^2$ cell consuming 7½ times as much area as any of the $4f^2$ technologies.

Not only does the cell size determine the cost of the memory, but it also sets the maximum memory size that can be produced within a given area. Many embedded designs have a limited amount of die area that can be devoted to on-chip memory. Those memories that have the smallest cell size support the greatest memory densities within a given amount of space.

Persistence

When compared against existing technologies keep in mind that all of these technologies are nonvolatile, an important advantage over DRAM, and that they all support *in-situ* programming, which makes them far faster to write than either NAND or NOR flash.

Minimum Process

Another important consideration is the scalability of the technology. Certain emerging memory technologies, particularly FRAM and PCM, have proven challenging to scale. FRAM has not been successfully scaled below 90nm and PCM's "On" resistance increases as the cell size decreases, making the technology more noise sensitive as the process shrinks, although PCM researchers successfully developed a 5nm cell over a decade ago.

Oxygen vacancy ReRAM is said to face issues when scaling below 10nm.

Both the Adesto Conductive Bridge and the Crossbar Metal Filament technologies are expected to scale well beyond 10nm.

Process Complexity

One great challenge is to produce a bidirectional selector device that has adequate performance to prevent sneak currents from undermining the bit's integrity. Both PCM (Intel's 3D XPoint Memory) and Crossbar's Metal Filament memory have advantages in this area since their selector devices are simpler than those of other technologies. Crossbar's selector is incorporated within the bit cell itself, while PCM uses current in the same direction for set, reset, and read operations, so it requires only a simple diode.

Of these two, the simpler is the Crossbar cell, since it requires fewer deposited layers thanks to the absence of a select device.

Disadvantages

All of these emerging memory technologies do have certain disadvantages when compared to today's entrenched technologies: None are as fast as DRAM and it will be several years before any can compete on cost against NAND flash, largely due to the economies of scale.

In embedded applications, though, which normally use NOR flash to store on-chip code and data, these emerging technologies provide a realistic way to move past the scaling limit of on-chip NOR. For applications where NOR is unavailable and an alternative technology must be used, the new technology will generally be selected based on the cost it adds to the chip. These applications will gravitate towards technologies that provide the smallest cell size and the lowest increase to the wafer processing cost.

Conclusion

The industry has entered an era in which embedded memories must be evaluated at the earliest stage of new system architectures. Researchers have been working for decades on several alternative technologies which vie to be the one to replace on-chip NOR flash in embedded applications.

This paper highlights the limitations that some of these emerging memory technologies face to scale to the most advanced process nodes while preserving compelling performance at affordable manufacturing cost. Memory companies and semiconductor foundries are working in close collaboration to co-develop, and ramp embedded memory to mass production. Using standard CMOS materials and simple manufacturing processing steps and tools provides the highest chance to succeed in this competitive market.

Although this paper provides reasons why some technologies are more likely than others to succeed, in the end the early adopters and the strength of strategic collaboration between memory IP providers and manufacturing partners will determine which of these technologies will win out to become the nonvolatile memory of choice for the industry as a whole.

Jim Handy, June 2018

Comparing the Technologies

		MRAM			ReRAM		
	PCM	**Flash-like**	**SRAM-like**	**FRAM**	**CBRAM**	**OxRAM**	**Crossbar**
Source	Intel	Everspin	Globalfoundries	Cypress	Adesto	Panasonic	Crossbar
Cell Type	1D1R	1T1J	2T2J	2T2C/1T1C	1T1R	1T1R	1TnR[a]
Cell Size	$4f^2$	$8f^2$	30-40f^2	30f^2/15 f^2	$8f^2$	$8f^2$	$4f^2$
Stackable	Yes	No	No	No	Yes	No	Yes
MLC	Yes	No	No	No	Yes	Yes	Yes
Selector	Diode	Transistor	Transistor	Transistor	Transistor	Transistor	Internal
Materials	10+	Many	Many	2	4		3
Layers	11	10+	10+	3	4		3
Masks	?	4	4	2			2
Current Process	20nm	40nm	22nm	130nm	130nm	130nm	40nm
Minimum	<10nm	<10nm	<10nm	TBD	<10nm	28nm	<10nm
Status	Producti on	Production	Samples	Production	Production	Prototypes	Samples
Scaling Impact	Higher bit resis-tance	Data Retention-Endurance Ratio tightens				Worse sensing margin	Improved on/off ratio
Read		1pJ/bit	1pJ/bit	0.37pJ/bit	24pJ/bit	66pJ/bit	1pJ/bit
	115ns	105ns	12.5ns[b]	2.3ns	100ns	3µs	10ns
Write		120µA/bit	0.5pJ/bit	0.37pJ/bit	1.5nJ/bit	8.9nJ/bit	60µA/bit
	50ns	170ns	40ns	2.3ns	60µs	8,500µs	10µs
Endurance	10^6	10^{10}	10^8	10^{14}	10^5	10^6	10^6
Retention	10yr	3mo[b]	10yr[b]	100yr	10yr	10yr[b]	10yr
Max Temp	85ºC	85ºC[c]	125ºC	125ºC	105ºC	85ºC	150ºC

Notes:

a. 1T1R also possible
b. Worsens with scaling
c. 125°C with PMTJ