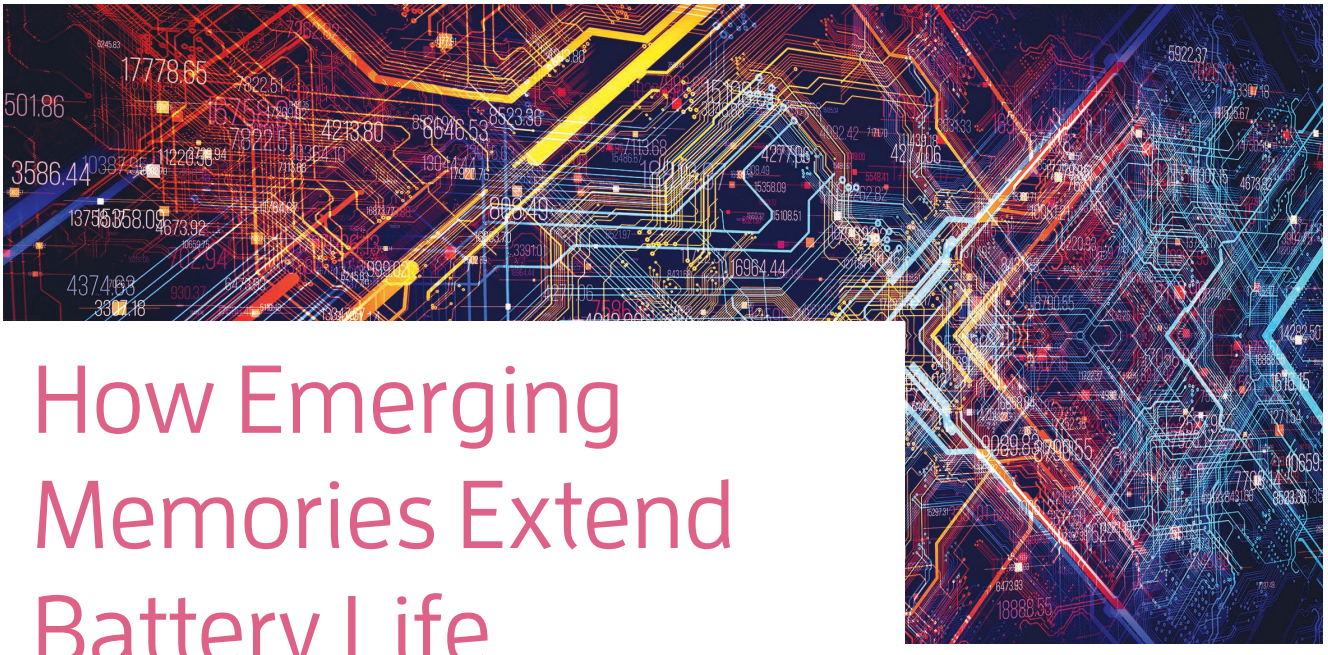
EDITOR TOM COUGHLIN
Coughlin Associates; tom@tomcoughlin.com



How Emerging Memories Extend Battery Life

Jim Handy, Objective Analysis
Tom Coughlin, Coughlin Associates, Inc.

Energy consumption is an issue with many connected digital products. Resolving energy efficiency issues and putting more memory in less die space create opportunities to use new nonvolatile memories for code storage and cache memory.

With today's explosion of battery-operated devices like industrial and consumer Internet of Things (IoT) endpoints, wearables, health monitors, and such, a growing focus is on the energy consumption of these devices.

INTRODUCTION

Designers must weigh tradeoffs among functionality, portability, and battery life since each of these play off against one another. For example, a device can have very elaborate functionality and a long battery life if a large battery is used, but that makes the device less portable. With a smaller battery the battery life will be shortened, but the product becomes more portable. If the designer strips down the feature set, then the smaller battery might do the job for a reasonable time.

Interestingly, this tradeoff is now being helped out through the use of new nonvolatile memory types, which have only become widely available in the past few years: magnetic RAM (MRAM), resistive RAM (ReRAM), phase-change memory (PCM), and ferroelectric RAM (FRAM). This article will examine those technologies and will show how their use can optimize the balance of these tradeoffs.

CURRENT MODEL

For the past few decades, endpoints have tended to use the same memory types: NOR flash, SRAM, and, in some cases, dynamic RAM (DRAM), to support the central processor. Often the NOR flash and SRAM are integrated into the processor chip in the form of a microcontroller unit

Digital Object Identifier 10.1109/MC.2023.3340799
Date of current version: 1 March 2024

(MCU). Some systems increase the density of these memories by adding external discrete NOR and SRAM chips, which adds to the cost. If an external NOR is used, some of its contents will often be stored within the MCU in an SRAM cache since NOR reads are relatively slow compared with the speed of a program's execution. Furthermore, SRAM scales more slowly than CMOS logic, and that means that the relative cost of the cache increases over time to become a growing share of the MCU chip's cost.

This works well in applications where there is no need to store data, which is done either to recover cleanly from power interruptions or to allow the chip to be powered down for energy saving. Things become more difficult when data must be stored. The designer usually chooses between two options:

- Use a battery-backup SRAM.
- Write data into the NOR flash.

A battery-backed SRAM works very nicely, as long as the battery functions. During normal operation, the SRAM operates at full speed, and it consumes very little power when in standby mode. Unfortunately, batteries have a limited lifetime and so must be changed. If the device needs to maintain the SRAM's data through this battery change, then the design becomes much more elaborate. This approach is easy until maintenance is considered, and then the design becomes significantly harder.

Some MCUs include a battery-backed SRAM, and this can simplify the designer's task a bit. Still, the battery replacement issue becomes a challenge. Also, as mentioned above, SRAM is not scaling with CMOS logic, so the SRAM's cost will become an increasing part of the MCU's cost over time.

NOR flash does not need a battery to store information, making it more appealing thanks to its lower complexity, but NOR flash takes significantly more time and energy to perform a write and has to go through a block erase if space does not yet exist for that write. A memory address cannot simply be overwritten in flash.

The system is designed to find opportunities to shut down frequently, saving valuable battery energy when it is powered down.

For example, while a NOR flash chip might take a certain amount of energy to read a page, the page programming might take 15 times as much energy due to higher voltages and currents along with longer cycle times. However, that is only true if there is free space for the data to be written into. If a block must be erased to provide room for that write, then the whole erase-then-write process can consume about 20,000 times as much energy as a read.

Furthermore, embedded NOR flash stops scaling at 28 nm. The advent of the fin-shaped field-effect transistors processes at 14 nm gets in the way of producing NOR flash, so foundries that produce aggressive process geometries either are in development or have already developed other nonvolatile memory technologies to replace NOR at 14 nm and smaller process nodes.

EMERGING MEMORIES AS A SOLUTION

Those new memory types that were mentioned at the beginning of the article, MRAM, ReRAM, PCM, and FRAM, all have attributes that make them better than either battery-backed SRAM or NOR flash for data storage and are poised to become a lower-cost alternative to either SRAM or NOR flash. All offer very fast read and write, all promise to scale to process nodes beyond those supported by NOR flash and SRAM, and all can help the engineer design a lower-energy system than SRAM or NOR.

One benefit that has not been mentioned so far is the ability to power a system down at any time without needing to move data from volatile RAM (SRAM or DRAM) into a nonvolatile memory. While systems with battery-backed SRAM can simply leave the SRAM in a standby state, running off the backup battery's power, other systems must move data from RAM into NOR flash, and this consumes a lot of energy. With an emerging memory technology, the same architecture can be used as with a battery-backed SRAM: the data can remain where they are at power-down to be accessed again when power is restored. This lends itself to a power-saving approach that Intel calls "Hurry Up, Get Idle" ("HUGI"). The system is designed to find opportunities to shut down frequently, saving valuable battery energy when it is powered down.

The drawback is that none of these technologies is yet produced in the kind of volume that will drive their costs down. Current memory technologies, like DRAM and NAND flash, are produced in high enough volumes and have been produced for so very long that manufacturers understand how to drive the costs out of the production process. This is not the case with newer memory technologies, so today they are the higher-cost alternatives. From the perspective of production volume, these technologies are still very young, even if they may have been in production for a number of years.

Fortunately, the migration to sub-28-nm process technologies is increasing the production volume of these memory types, which will eventually lead to cost reductions. In the end, this promises to make these technologies cheaper than SRAM or NOR flash, but this is not the case today.

While the sub-28-nm problem is unique to NOR flash, SRAM's biggest problem is that each bit is very large

since it requires six transistors to implement, while a NOR flash or an emerging memory bit is much smaller, typically taking only a single transistor and, in the case of the emerging memories, some kind of bit storage element (more on those later). In some cases, the bits will be stored differentially to increase speed, but these cells still consist of only two transistors and two storage elements. This makes them necessarily cheaper than SRAM as long as the wafer costs are the same. Today the wafer costs for emerging memories are higher, but that difference will fade as the production volume increases.

EMERGING MEMORY TYPES

The following memory types are in production today.[a] All offer roughly the same attributes, and any one of them could rise above the others to become the leading memory type over the next decade. All of them provide persistence (that is, they are nonvolatile), all write in place (which is a vast improvement upon flash's block erase and page write, with erase before write), all have fast, low-energy writes, and all can scale to smaller process geometries than are currently available. They perform nearly as well as battery-backed SRAM but without the battery and with the promise of becoming much less expensive than SRAM.

MRAM

MRAM comes in several forms. Toggle mode MRAM is in the highest volume today but has trouble scaling past 120 nm, so it is being displaced by spin transfer torque (STT) MRAM. In the future, other versions, mainly spin orbit torque, with faster performance, may replace STT MRAM. Each bit of any of these technologies can be implemented with a single MRAM bit element and a single transistor. Today the transistor's size limits how small a bit can be made since the technology requires relatively high currents, but researchers are working on a solution to this problem.

All MRAM uses a special layer of material that exhibits the property of giant magnetoresistance to store the bit. This material, while produced in high unit volume to manufacture recording heads for HDDs, has a very small die size, so it is not yet manufactured in the high wafer volumes of silicon CMOS and is therefore expensive today.

MRAM is available as a foundry process from TSMC, Samsung, and GlobalFoundries. Discrete MRAM chips are available from Everspin and Avalanche.

ReRAM

ReRAM uses a resistive element to store a bit. While some manufacturers use a less-understood material to produce the bit element, certain companies, namely, Weebit Nano and Crossbar, have developed ReRAM that is based on a slight change to the same silicon dioxide insulation material that is universally used in silicon semiconductors. This should accelerate these technologies' ability to reach the economic benefits of high-volume production.

There are two basic programming mechanisms: filamentary and oxygen depletion. While this article will not explain these mechanisms, neither is as well understood as is standard silicon CMOS.

ReRAM cells consist of a single resistive element and a selector, which today is typically a transistor. This means that the bit size rivals that of MRAM and NOR flash. Future ReRAMs are expected to use a two-terminal selector, which can be built below the resistive element to cut the bit's size in half and which will facilitate layering bits in multiple "decks" to further double, triple, or quadruple the number of bits that can fit into a given area of silicon.

Today discrete ReRAM chips are produced in volume by Fujitsu and its partner Panasonic. Foundries TSMC, Samsung, Global Foundries, Winbond, Skywater, DB HiTek, SMIC, and Crocus Nano all offer an embedded ReRAM process.

PCM

PCM (or PRAM) has had its day in the sun in its 3D XPoint memory incarnation. Like a ReRAM, it stores the bit in a resistive element, but the storage mechanism is different since it involves a material change. In most PCMs, temperature ramps are used to change the storage element between crystalline (conductive) and amorphous (nonconductive) phases, but there is another method that changes the resistance through high programming currents.

PCM is based on chalcogenide glasses, which are not as well understood as is silicon. Some of these glasses also involve elements that are difficult to manage in a silicon fab.

As with ReRAMs, PCM can use either a transistor or a two-terminal selector. The most common two-terminal selector today is also based on a chalcogenide glass, so PCM is a good fit. Intel and Micron were able to use this to their advantage since it allowed multiple "decks" of 3D XPoint memory to be easily stacked, and that reduced the technology's cost for a given memory capacity.

Today, only STMicroelectronics provides PCM as an embedded memory in its "Stellar" microcontroller. BAE sells its PCM "C-RAM" to aerospace applications that value its immunity to radiation.

FRAM

FRAM involves no iron, despite its name. Since this technology stores a bit's state via hysteresis that resembles the ferromagnetic hysteresis loop, researchers called it FRAM. FRAM is also the first semiconductor memory, with the first multibit monolithic prototype developed in 1955, three years before Jack Kilby's 1958 invention of the integrated circuit.[b]

From the 1950s through 2010, all FRAM was produced using either strontium bismuth titanate or lead

[a]Report: Emerging Memories Branch Out, Coughlin Associates and Objective Analysis, 2023. http://Objective-Analysis.com/reports/Emerging#.

[b]FRAM Turns 68, The Memory Guy Blog, Jim Handy, 10 July, 2020. https://TheMemoryGuy.com/fram-turns-68/.

zirconium titanate, both of which include high-mobility elements that can easily contaminate a silicon fab. This limited their popularity. In 2010, NamLab in Dresden, Germany, found evidence of ferroelectric behavior in hafnium oxide (HfO), which is prevalent as a gate dielectric in very advanced silicon processes; this discovery has led to a lot of research but not yet to any actual products.

Discrete FRAM is produced by Infineon and Lapis Semiconductor, TI embeds it into a microcontroller, and Fujitsu and Panasonic embed FRAM into RFID chips for mass-transit fare cards.

LOW-POWER APPLICATIONS OF EMERGING MEMORIES

Here we will present a few of the many applications that use emerging memory technologies to save energy in low-power applications.

Mass-transit fare cards

Very early examples of such applications are the mass-transit fare cards pioneered in Japan and now used broadly in Asia. These cards have no internal power source, yet they store the value assigned to them less any transactions from the card's use. They are read via near-field communications (NFC).

The cards must store the value, allow it to be read, and then allow a new total to be written back into the card, all using only the energy provided by the NFC signal. Fujitsu and Panasonic chose to use FRAM for this application because its fast low-power write could be powered by the NFC signal.

An example of one of these cards is shown in Figure 1. They are the same size and shape of any standard charge card.

FIGURE 1. A Guangzhudong Shenzen Railway Company fare card. (Source: Wikimedia Commons, IC ticket of Guangshen Railway.jpg.)

Personal fitness monitors

There is widespread use of MRAM in personal fitness monitors, which must perform numerous sophisticated tasks for a full day or more using only the energy that will fit into a small battery within the watch-sized device. Many of these use the MRAM version of the Apollo 4 processor from Ambiq, a company that uses subthreshold logic to get the highest performance out of the absolute smallest amount of energy possible.

Figures 2 and 3 show two examples: the Fitbit Luxe from Google (Figure 2) and the Versa 4 from Garmin (Figure 3). Garmin has another device not pictured here, the Fenix 7 Solar, which adds a solar cell to an MRAM-based wearable to further extend the time between charges.

FIGURE 2. Google's Fitbit Luxe.

FIGURE 3. Garmin's Versa 4.

Medical devices and prosthetics

Various development efforts are underway to incorporate emerging memory into everything from disposable health monitoring devices, which look more like a small bandage than instrumentation, up to cardiac defibrillators and hearing aids. While the developers generally do not disclose the chips used inside their devices, we understand that MRAM, ReRAM, and FRAM are all being used in such applications.

BIG CHANGES ON THEIR WAY

Readers should expect to see significant changes leading to longer battery life in the next few years as emerging memory technologies become widespread in IoT endpoints and other battery-operated equipment. There may even be a rise in the use of scavenged power, as is already done in mass-transit fare cards and in Garmin's Fenix 7 solar wearable device.

In the end, a lot of this will be enabled through the use of new memory technologies, which drastically reduce the energy requirements of data storage. These technologies are about to ramp pretty quickly, in support of finer process geometries, so they will become common over the next five years.

JIM HANDY is the general director of Objective Analysis, Los Gatos, CA 95032 USA. Contact him at jim.handy@objective-analysis.com.

TOM COUGHLIN is the president of Coughlin Associates, Inc., San Jose, CA 95124 USA. Contact him at tom@tomcoughlin.com.