Flash Memory at The Transistor Level

Nethania Morales

Advisor: Professor Brad Jackson

California State University, Northridge

Abstract—**Flash memory is a very common consumer product used to store a variety of data types. Flash memory has had a significant impact in today's digital era. There are many components that go into flash memory to have hardware that can store data, whether that be one byte or 100 Gigabytes. Flash memory is split into two types of memory: NAND and NOR memory. Each type of flash memory is structured differently, in return providing multiple advantages and disadvantages. Both types of flash memory also have different methods on how to read, write and erase data into their memory cells. This study analyses how flash memory works at the transistor level, specifically examining how a floating gate transistor can store data without power. This study will also examine the advantages and disadvantages in both NAND and NOR flash memory, as well as the advantages of using N-type floating gate transistors versus P-type floating gate transistors.**

I. Introduction

Throughout the course ECE 442 Digital Electronics, there has been a strong focus on Complimentary MOSFETs. The course specifically went over what a CMOS transistor is, the voltage transfer characteristics of this type of transistor, real-world applications using CMOS technology, etc. CMOS transistors can be used to design all types of logic gates, latches, SRAM, DRAM, etc. Transistors are seen in almost every digital electronic application; furthermore, showcasing the significance these components have in the digital electronics industry. Both SRAM and DRAM are volatile memory, meaning data is lost when power is lost. The main structure of SRAM and DRAM are composed of CMOS transistors. Because both SRAM and DRAM are volatile/temporary memory, how or what types of transistors are used in non-volatile flash memory? How can a transistor change to have the ability to store data without the need for power?

II. Basics Of Flash Memory

Flash memory is a very popular consumer grade product that comes in all shapes and sizes. There has and continues to be a high demand for fast, easily transferable, portable, and reliable data storage. Flash memory has become one of the solutions to this consumer need. Unlike Random Access Memory (temporary memory), flash memory is non-volatile. This means that it can store all types of data without the constant need for power. Some few examples of flash memory are USB sticks, flash drives, memory cards, and the list goes on. The reason why this type of memory is very popular is because users can write and erase data as many times needed. Although there is no real "cap" to how many times a memory cell can be written and rewritten to, there are some drawbacks to this over time. However, the positives out way the negatives. [1][2][3]

Flash memory is composed of multiple Field Effect Transistor FETs, also known as a floating gate transistor. A single floating gate transistor is made up of a standard MOSFET except containing two gates instead of one. The names of the two gates are the Control Gate and the Floating Gate. Fig. 1 shows the circuit symbol representation of a standard PMOS and NMOS transistor compared to that of the Field Effect Transistor FET. [1][2][3]

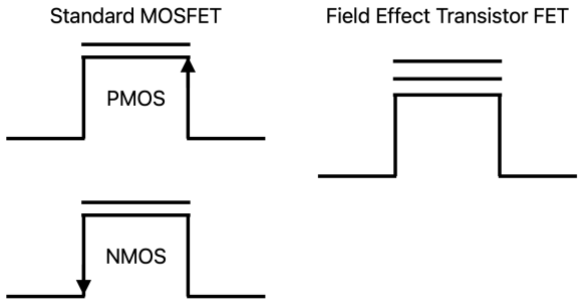

Fig. 1. The symbols used in circuit schematics to represent a standard PMOSFET, NMOSFET, and Field Effect Transistor FET. [10]

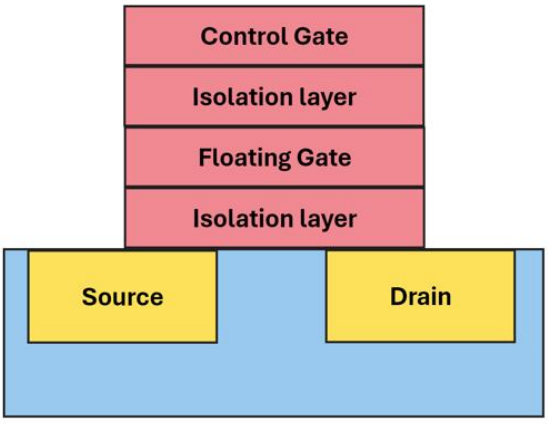

Fig. 2. A simple diagram of the structure of a Field Effect Transistor FET, floating gate transistor. [3[

As seen in Fig. 2, the Floating Gate is in between two isolation layers. These isolation layers are what allows for each individual transistor to store data (electrons). Once the electrons are past the isolation layer into the floating gate layer, there is no way for the electrons to exit unless "provoked". The isolation layers of the transistor allow for data to be stored without having power being provided to the flash memory device.

In general, the movement of electrons into the floating gate is a "write" operation. The movement of electrons out of the floating gate back into the channel is a "erase" operation. To read the data stored in flash memory, the charge is measured from the floating gate. [1][2][3]

III. Types Of Flash Memory

Flash memory can be divided into two subjections: NAND memory and NOR memory. Each type of memory can be used for different applications, and both have a list of advantages as well as disadvantages. Overall, NOR flash memory is used to store code or large applications, and NAND flash memory is typically used in auxiliary storage. [2][3]

NAND memory is composed of multiple floating gate transistors in a series connection. The reason why this type of memory is called "NAND memory" is because this connection emulates a NAND gate. This type of memory can be written to in blocks. The bit line can only go low when all the word lines are set to high. The diagram below shows a circuit diagram of such memory with a total of four memory cells. Each transistor is representing a memory cell, and each is connected to a separate word line. Fig. 3 shows what a NAND flash memory circuit schematic looks like containing four memory cells. Fig. 4 shows cross section block diagram of such memory.[1][4][8]

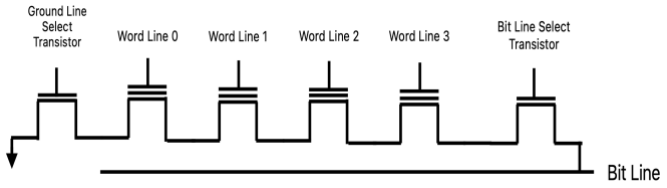

Fig. 3. The NAND flash memory schematic with a total of four memory cells. [8][10]

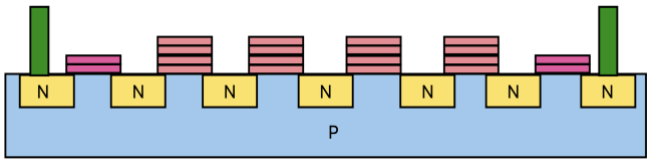

Fig. 4. A block diagram representation of a NAND flash memory cross section with a total of four memory cells. [10]

There have been many advances in how much a single NAND memory cell can store. For example, some flash memory contains SLC's. SLC stands for Single-Level Cell memory. An SLC can only store one bit in one NAND memory cell. This option is the most expensive out of all the types of NAND memory because it has the highest toleration and has the fastest programming time. The next types of NAND memory are MLC, TLC, QLC, and Vertical 3D NAND memory. MLC stands for Multi-Level Cell memory. This type of memory can store two bits of data per memory cell. Each type of memory gets cheaper in price due to an increase in the number of bits stored per each memory cell. TLC stands for Triple-Level Cell memory. It can store three bits of data per cell. QLC stands for Quad-Level Cell and can store 4 bits of data per cell. The higher the number of bits stored per cell, the longer it takes to write/program such cells. Finally, the Vertical 3D NAND memory stacks memory cell one on top of another, hence the name. [5]

The other type of flash memory is NOR memory. NOR memory is composed of multiple floating gate transistors in a NOR connection. This type of memory can be written to in bytes. Fig. 5 shows what a NOR flash memory circuit schematic looks like containing four memory cells. Fig. 6 shows cross section block diagram of such memory. Each connection for the transistors has a line to ground. For NOR memory, whenever one word line gets set to high, then the bit line goes low. [1][3][4][8]

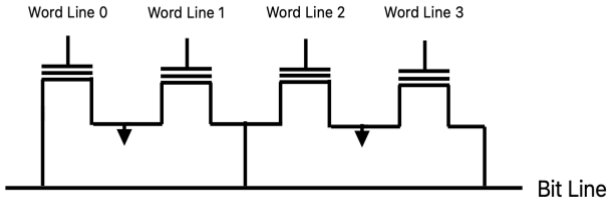

Fig. 5. The NOR flash memory schematic with a total of four memory cells. [8][11]

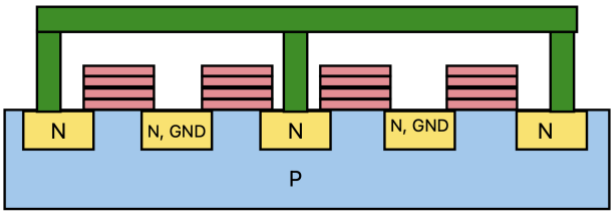

Fig. 6. A block diagram representation of a NOR flash memory cross section with a total of four memory cells. [11]

Because NAND and NOR memory are structured differently, there are many pros and cons to choosing one over the other. NAND memory is smaller in size compared to NOR memory because there is a smaller number of wires connected to the bit line. There are drawbacks to having a series connection.

This makes it a slower process to read from a flash memory cells in NAND memory. The parallel connection in a NOR memory allows for much faster read “operations” to individual cells. However, NAND memory allows for faster writing to memory cells. NOR memory has slower write and erase “operations”. The reasoning behind this is because the configuration of how the floating gate transistors are connected (in parallel or series). [1][2][3]

As mentioned previously, programming and reading NAND versus NOR memory is different. NAND memory is read and written to in blocks, and NOR memory is read and written to in bytes. There is a total of 128 KB in one block. Because NAND reads and writes in blocks, it cannot work at the byte level. [1] [2][3]

Because NAND and NOR memory vary vastly in pros and cons, not one is better than the other. Choosing what type of flash memory to use will depend greatly on the application it will be used for.

IV. Storing Data In Flash Memory

There are a few methods on writing, reading, and erasing data in flash memory for floating gate N-type transistors and P-type transistors. The most used type of floating gate transistors in flash memory are N-type transistors. The following describes methods used for mainly N-type floating gate transistors.[9]

Reading data from flash memory is the same process for all types of transistors. In general, a specific reference voltage must be applied to both the source and the drain of the transistor. Once this is done, the charge/current can be measured. [4]

Programming a memory cell/floating gate transistor requires the addition and removal of electrons from the floating gate. Electrons need to surpass the isolation layer anytime programming occurs. The addition of electrons to the floating gate layer is the act of writing data to a memory cell. The removal of electrons to the floating gate layer is the act of erasing data from a memory cell. The two processes used to program a memory cell are electron injection and electron tunneling.[3]

Electron injection consists of channel hot-electron injection and drain avalanche hot-carrier injection. Electron tunneling refers to Fowler-Nordheim tunneling. NAND memory uses electron tunneling to program its memory cells. NOR memory uses electron injection to write to its memory cells and electron tunneling to erase from memory cells.[6]

In Fowler-Nordheim tunneling, there is a strong positive potential energy at the control gate. There is also a negative potential energy at the source and drain. This forces the electrons to go past the isolation layer into the floating gate. With the opposite effect, the electrons are moved out of the floating gate back to where they came from. When erasing in NAND memory, a high voltage is sent to the source and the drain meanwhile there is a negative voltage sent to the control gate. Because of this isolation layer, there is no need to have constant power for the transistor to keep the electrons in the floating gate layer. The diagram below shows an illustration of how Fowler-Nordheim tunneling works.[1][3][6]

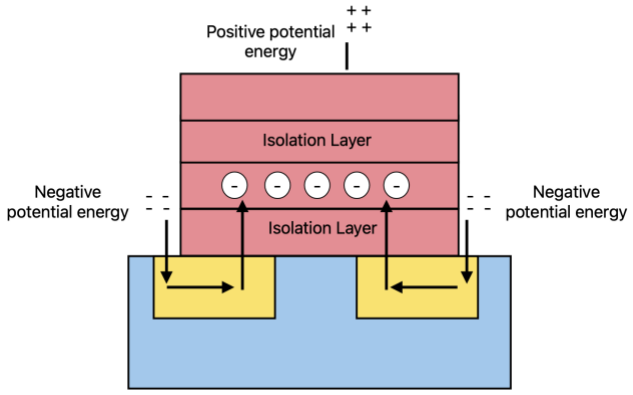

Fig. 7. A diagram showing the process of how Fowler-Nordheim tunnelling works. [3]

Hot-electron injection is the process of moving electrons to the floating gate. This happens when there is an attracting charge at the control gate of the transistor, and high current traveling through the transistor’s channel. This forces the electrons go from the source to the drain, except they break through the isolation layer and travel to the floating gate instead. Drain avalanche hot-carrier injection has a similar way of working. At the end, electrons can break through the isolation layer into the floating gate. Below is a diagram that illustrates how hot-electron injection works. [1][3][6]

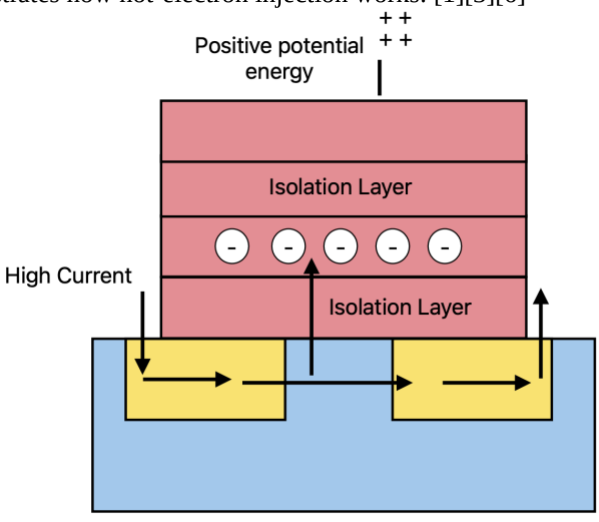

Fig. 8. A diagram showing the process of how Hot-electron injection works. [3]

V. N-Type Versus P-Type Transistors In Flash Memory

As mentioned previously, N-type floating gate transistors are mainly used in flash memory designs. However, P-type floating gate transistors do exist although not as commonly used. In a technical study titled “Floating Gate Circuits in MOSIS” by J.R. Mann, Group 23, there is a comparison on writing to N-type versus P-type floating gate transistors. Their research further proves why N-types floating gate transistors are vastly used. [9]

In their experiment/research, electron injection was used to write to the transistors, and Fowler-Nordheim tunneling was used for erasing data from the transistors. They tested three types of N/P-type transistors, but only the N-type stacked gate transistors will be discussed here. [9]

It was noted that for the electron injection to work, only the transistors lengths of two micrometers were used. The voltage across the drain to source was 17V. In around 20ms, the threshold voltage got to a maximum of -5V with the gate to source voltage equal to 10V. After this point, as the gate to source voltage increases past 10V, it is seen that the threshold voltage also decreases. [9]

For the P-type floating gate transistors, different drain to source voltages were applied: -11V, -10V, -9V, -8V. It was noted that when writing to the P-type floating gate transistor, the current in the channel increases. This makes is easier to write to the transistor yet affects the erasing capabilities. J.R. Mann writes that the erasing operation ultimately limits the ability for the transistor to easily turn on causing the write operation to not function in the transistor. J.R. Mann made it clear that a high amount of erasing among P-type floating gate transistors should exceed a certain amount. [9]

To continue, when the drain to source voltage is at -9V, there is a linear increase of the threshold voltage up to when the gate to source voltage reaches 4V. After this point, as the gate to source voltage increases, there is a rapid decrease in the threshold voltage. J.R. Mann states that this represents a “stuck bit”. However, when the drain to source voltage is -10V and -11V, this phenomenon did not occur. Mann’s research shows that the fastest writing time occurred when the drain to source voltage was -11V. [9]

This research paper provides evidence as to why most flash memory uses N-type floating gate transistors. There are more details to consider when writing to P-type floating gate transistors; therefore, it would be easier to not implement them in flash memory design. It is important to note that the research performed in the paper stated above tested individual transistors. Although this is true, this research can be applied to flash memory design. [9]

VI. Common Errors In Flash Memory

NAND flash memory is more susceptible to a higher error rate in comparison to NOR flash memory because of the series connection. NAND memory requires more programming when new bits need to be written to memory cells. This causes permanent and temporary errors in the memory cells. [7]

A permanent error that is very common in NAND memory is memory wear. Overtime, as electrons keeping passing the isolation oxide layer, it is inevitable for some electrons to get stuck in this layer of the transistor. This negatively affects the oxide layer and causes it to deteriorate. [7]

The next category of errors is temporary error. The first type of temporary error in NAND flash memory is the program disturb error. When programming a memory cell, another voltage is sent to the gate of these cells causing a change in the threshold voltage. When programming one specific transistor, the neighboring transistors can also cause a change in the threshold voltage. However, this can easily be solved. Most temporary errors can be fixed by copying the data of the affected block to another block and erasing the block with the issue. [7]

The next type of temporary error is called a read disturb error. Based off the name, this error is caused from reading the memory cell/transistor. For a proper read operation, all the other memory cells must be on. Overtime, the repetition of the read operation affects all the memory cells that are not trying to be read not just the neighboring transistors. [7]

Next is the overprogramming error. As stated above, programming memory cells in NAND memory can alter/raise the threshold voltage. When the threshold voltage gets very high, it is harder to turn the transistor on. This can cause errors the data being stored. [7]

Lastly, another error is the retention error. Flash memory is not infinite. As time passes, all elements affect the durability of the storing capacities of each memory cell. Electrons can leak outside of the isolation layers, which in turn causes errors in the data being stored. [7]

All these errors continue to prove that flash memory, although a great alternative for auxiliary storage, is not permanent. Even if flash memory can last many years, it is important to keep in mind that all flash memory must eventually be replaced. [7]

VII. Advantages Of Flash Memory

There are still many advantages to flash memory. Flash memory is inexpensive compared to other types of memory, portable, reliable, and durable. Because there are no hardware components that move, flash memory lasts longer. Over the years, there have lots of advances on the size of flash memory. Flash memory as the ability to provide large amounts of byte storage. The main advantage of flash memory is that it is

composed of Field Effect Transistor FETs. These transistors, as stated throughout this research paper, do not need power to store data. [1][3]

VIII. Conclusion

In conclusion, the floating gate transistors have proven to be a very powerful tool in the world of non-volatile memory. It provides a temporary solution for fast growing industry that requires quick and easy ways to have access to permanent data storage. Even then, flash memory is only “permanent” to an extent. Eventually, all flash memory needs to be replaced because of wearing of the memory cells. So far, the positives still overpower the negatives. Depending on the type of memory composition, floating gate transistors allow for high amounts of auxiliary storage, faster read and write operations, minimum power use, etc. Floating gate transistors open the door to all types of consumers to have a easy, user friendly experience to store large quantities of data.

References

[1] "What is flash memory all about? | StoneFly," *StoneFly*. https://stonefly.com/resources/what-is-flash-memory-all-about/

[2] T. Windbacher, "Engineering Gate Stacks for Field-Effect Transistors," PhD dissertation, Vienna University of Technology, 2010.

[3] K. Yasar, "Flash Memory," *TechTarget,* Jun. 28, 2023. https://www.techtarget.com/searchstorage/definition/flash-memory

[4] Hyperstone, "Floating Gate Technology | NAND Flash Transistors (suggested #1 for training sequel)," *YouTube*. Jul. 24, 2019. [Online]. Available: https://www.youtube.com/watch?v=AO7CNNZmtTw

[5] Hyperstone, "SLC, MLC, TLC and QLC | SSD Flash Memory Explained (suggested #2 for training sequel)," *YouTube*. Jul. 16, 2019. [Online].
Available: https://www.youtube.com/watch?v=5S_gUbzVdh8

[6] J. Gray, "APPLICATION OF FLOATING-GATE TRANSISTORS IN FIELD PROGRAMMABLE ANALOG ARRAYS," MA thesis, Georgia Institute of Technology, 2005.

[7] A. Aravindan, "Flash 101: Errors in NAND Flash," *Embedded,* Dec. 04, 2018. https://www.embedded.com/flash-101-errors-in-nand-flash/ (accessed Dec. 09, 2023).

[8] A. Aravindan, "Flash 101: NAND Flash vs NOR Flash," *Embedded*, Jul. 23, 2018. https://www.embedded.com/flash-101-nand-flash-vs-nor-flash/ (accessed Dec. 09, 2023).

[9] J. R. Mann, "Floating Gate Circuits in MOSIS," Nov. 01, 1990.

[10] L. "Tech refresher: Basics of flash, NAND flash, and NOR flash," *Microcontroller Tips*, Oct. 09, 2019. https://www.microcontrollertips.com/tech-refresher-basics-of-flash-nand-flash-and-nor-flash/

[11] TECHDesign, "NAND vs NOR: Understanding the Differences in Flash Memory," *TECHDesign Blog*, Mar. 02, 2023. https://blog.techdesign.com/nand-vs-nor-understanding-differences-flash-memory