NAND vs NOR Flash Memory

Content courtesy of Wikipedia.org

David Harrison,
CEO/Design Engineer for Model Sounds Inc.

Common FLASH Memory

- SD cards + mini, micro versions – serial interface – slower
- Compact Flash - parallel interface – faster
- Olympus/Fuji Xd Cards
- Sony memory stick
- USB “Thumb” DRIVES
- Novelty “Flash drives – lipsticks, other novelty cases
- SSD “Hard” Drives
- Embedded GBytes of memory in Smart Phones etc.
- These are all examples of NAND FLASH memory.

Why is it called FLASH Memory?

- Although Flash memory is technically a type of EEPROM (Electrically Erasable Programmable Read-Only Memory), the term "EEPROM" is generally used to refer to non-flash EEPROM which is erasable in very small blocks, typically bytes. EEPROM usually does not need a specific erase cycle – you can just write to a byte address with either a “1” or a “0”.

- Because erase cycles are slow, FLASH memory uses large block sizes for erasing and writing large amounts of data. I.E. the erase is FLASHED in blocks and writes are done typically in pages of 256 bytes or more.

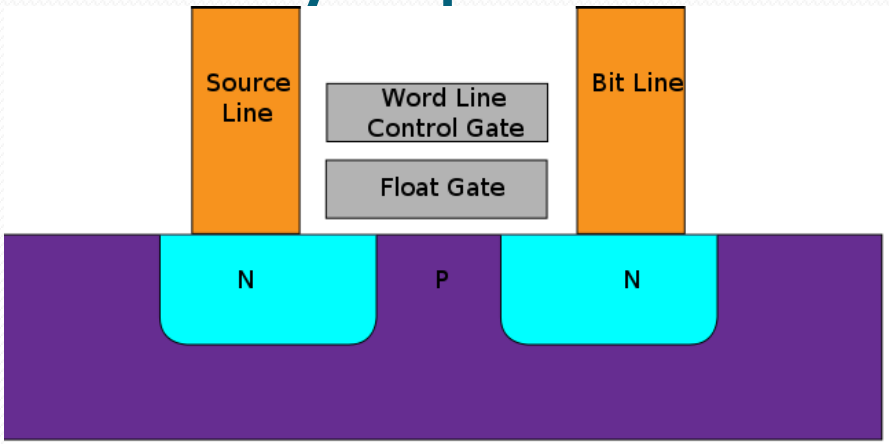
- Flash memory was developed from EEPROM. There are two main types of flash memory, which are named after the NAND and NOR logic gates.

  The individual flash memory cells use structures similar to those of NAND and NOR gates.

Principle of Flash Memory Operation

- **Floating-Gate MOSFET Transistor**

- Similar to regular N-channel MOSFET but with an additional floating gate, insulated on all sides.

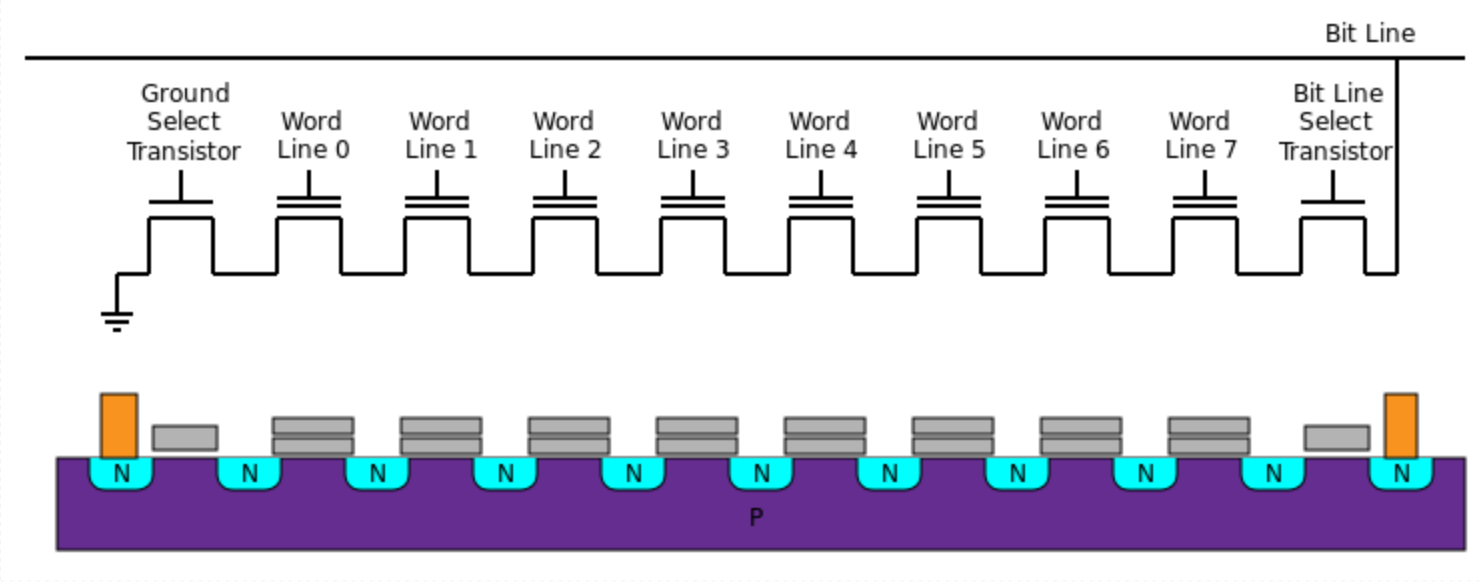


- Because the Floating Gate is electrically isolated by its insulating layers, any electrons placed on it are trapped there and, under normal conditions, will not discharge for many years.

- When the Floating Gate holds a charge, it screens (partially cancels) the electric field from the Control Gate, which modifies the threshold voltage (V_T) of the cell (more voltage has to be applied to the Control Gate to make the channel conduct).

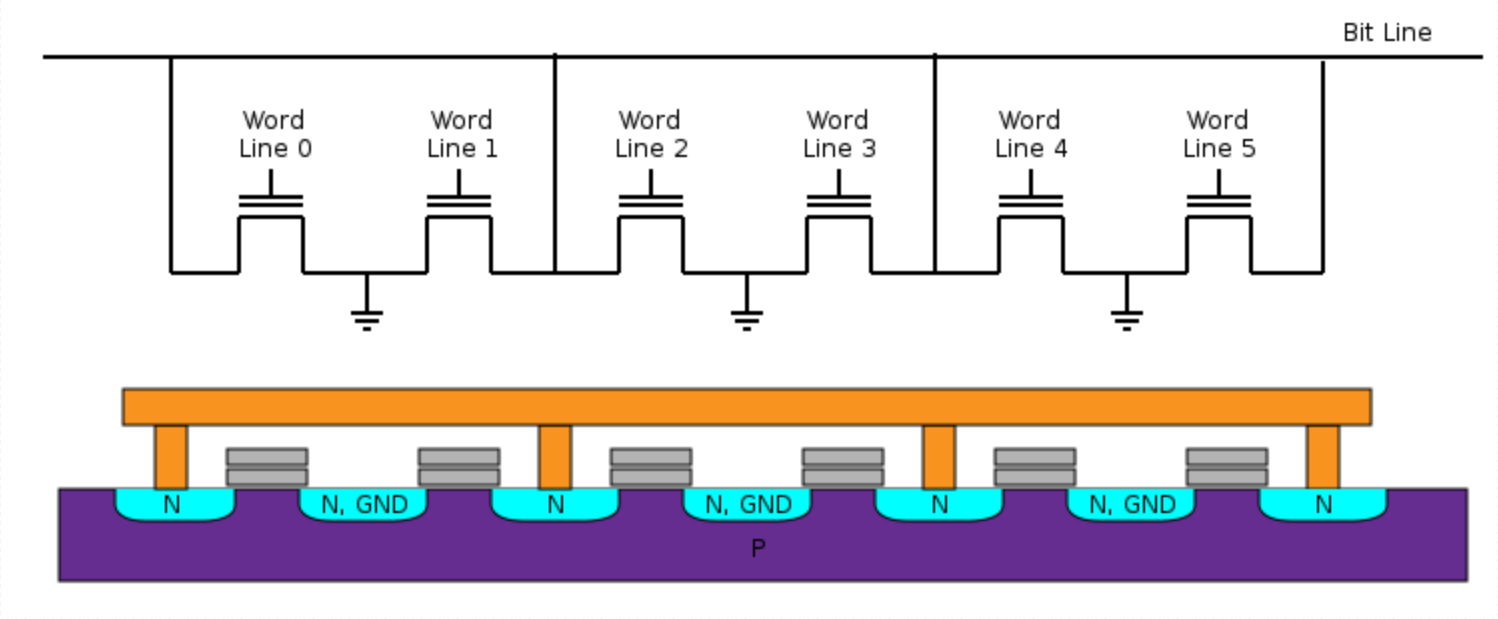
- For read-out, a voltage intermediate between the possible threshold voltages is applied to the Control Gate, and the MOSFET channel's conductivity tested (if it's conducting or insulating), which is influenced by the Floating Gate . The current flow through the MOSFET channel is sensed and forms a binary code, reproducing the stored data.

NAND Flash Memory Structure and Equivalent Schematic



Since all the MOSFETS are topologically in series the circuit looks similar to a NAND gate. But that is where the similarity ends.

NOR Flash Memory Structure and Equivalent Schematic



Since the MOSFETS are topologically in parallel pairs the circuit looks similar to a NOR gate. But that is where the similarity ends.

Why is Erasing Necessary?

- All flash memory, NAND or NOR, can only program zeroes.
- Therefore a block of memory has to be erased to set it all to 1's first.
- Once a bit has been set to 0, only by erasing the entire block can it be changed back to 1.
- A location can be rewritten as long as the new value's 0 bits are a superset of the over-written values. For example, a nibble value may be erased to 1111, then written e.g. as 1110. Successive writes to that nibble can change it to 1010, then 0010, and finally 0000.
- File systems designed for flash devices can make use of this capability, for example, to represent sector metadata.

Why Two types of FLASH Memory?

NAND FLASH

- Has a much denser layout and much greater storage capacity per chip than NOR.
- Better at reading and writing large blocks of data and are well suited to emulating "Hard Disks" accessing pages or "sectors" e.g. 512, 1024 or 4096 bytes at a time.
- Faster Writes than NOR FLASH but slower Reads.
- Manufacturers deliberately allow some bit failures during manufacture and use ECC to identify and mark bad bits and/or blocks.
- Reading and Programming is performed on a page basis, but erasing is performed on a larger block basis.

NOR FLASH

- Much faster reads than NAND Flash.
- Random access, e.g. not block oriented, allows "Execute in Place" – or XIP.
- Ideal for on-chip program flash as in microcontrollers and small, fast, random access NVM for data storage.
- XIP means code can be executed directly from the NOR flash instead of having to be copied to RAM first. Also no bit errors are allowed.

FLASH Memory Low Level Access

- The low-level interface to flash memory chips differs from those of other memory types such as DRAM, ROM, and EEPROM, which support bit-alterability (both zero to one and one to zero) and random access via externally accessible address buses.

- NOR memory has an external address bus for reading and programming. Reading and programming are random-access, and unlocking and erasing are block-wise.

- For NAND memory, reading and programming are page-wise, and unlocking and erasing are block-wise.

- Both NAND and NOR flash can be Single Level Cell (SLC) – one bit per cell or Multi Level Cell (MLC) – more than one bit based on charge level on the floating gate.

NAND vs NOR Flash Memory …

- The interface provided for reading and writing the memory is different (NOR allows random-access for reading, NAND allows only page access).

- These two are linked by the design choices made in the development of NAND flash. A goal of NAND flash development was to reduce the chip area required to implement a given capacity of flash memory, and thereby to reduce cost per bit and increase maximum chip capacity so that flash memory could compete with magnetic storage devices like hard disks.

- In NOR flash, the parallel connection of cells resembles the parallel connection of transistors in a CMOS NOR gate. In NAND flash, cells are connected in series, resembling a NAND gate. The series connections consume less space than parallel ones, reducing the cost of NAND flash.

- When NOR flash was developed, it was envisioned as a more economical and conveniently rewritable ROM than contemporary EPROM and EEPROM memories. Thus random-access reading circuitry was necessary.

  However, it was expected that NOR flash ROM would be read much more often than written, so the write circuitry included was fairly slow and could only erase in a block-wise fashion.

  On the other hand, applications that use flash as a replacement for disk drives do not require word-level write access, which would only add to the complexity and cost unnecessarily.

Flash Erasing and Programming

NOR Flash Erasing

- A NOR flash cell can be erased, or set to a binary "1" value, by the following procedure:
  - An elevated on-voltage (typically >5 V) is applied to the Control Gate turning on the channel.
  - The source-drain current is sufficiently high to cause some high energy electrons to jump through the insulating layer onto the Floating Gate, via a process called hot-electron injection.

NOR Flash Programming

- To program a NOR flash cell (setting it to the "0" state), a large voltage *of the opposite polarity* is applied between the Control Gate and source terminal, pulling the electrons off the Floating Gate through quantum tunnelling.

Internal Charge Pumps

- Despite the need for high programming and erasing voltages, virtually all flash chips today require only a single supply voltage, and produce the high voltages via on-chip charge pumps.

Write Endurance

- SLC NAND flash is typically rated at about 100K cycles (Samsung OneNAND KFW4G16Q2M)
- MLC NAND flash used to be rated at about 5K – 10K cycles (Samsung K9G8G08U0M) but is now typically 1K – 3K cycles
- SLC floating-gate NOR flash has typical endurance rating of 100K to 1M cycles (Numonyx M58BW 100K; Spansion S29CD016J 1M)
- MLC floating-gate NOR flash has typical endurance rating of 100K cycles (Numonyx J3 flash)

Memory Wear and Wear Levelling

- One limitation of flash memory is that it has a finite number of program-erase cycles (typically written as P/E cycles). Most commercially available flash products are guaranteed to withstand around 100,000 P/E cycles before the wear begin
- This effect is partially offset in some chip firmware or file system drivers by counting the writes and dynamically remapping blocks in order to spread write operations between sectors; this technique is called wear levelling.
- Another approach is to perform write verification and remapping to spare sectors in case of write failure, a technique called bad block management (BBM).
- For portable consumer devices, these wear out management techniques typically extend the life of the flash memory beyond the life of the device itself, and some data loss may be acceptable in these applications.
- For high reliability data storage, however, it is not advisable to use flash memory that would have to go through a large number of programming cycles.
- This limitation is meaningless for 'read-only' applications such as thin clients and routers, which are programmed only once or at most a few times during their lifetimes.

Flash That Lasts … and Lasts and …

- Recently, Taiwanese engineers from Macronix announced at the 2012 IEEE International Electron Devices Meeting (IEDM) that it has figured out how to improve NAND flash storage read/write cycles from 10,000 to 100 million cycles using a “self-healing” process that uses a flash chip with “on-board heaters that could anneal small groups of memory cells.”

- The built-in thermal annealing replaces the usual erase cycle with a local high temperature process that not only erases the stored charge, but also repairs the electron induced stress in the chip, giving write cycles of at least 100 million.

- The result is a chip that can be erased and re-written over and over, even when it should theoretically break down. As promising as Macronix’s breakthrough could be for the mobile industry, however, there are no plans for a commercial product to be released any time in the near future.