2.1.1 Flash Memory

Flash memory was invented by Dr. Fujio Masuoka \cite{34} in 1980 at Toshiba. Flash memory can be divided into NOR- and NAND-based memory \cite{35}. NOR-based flash memory provides high read performance and enables full address and data bus access: Thus, it supports eXecution In Place (XIP), which allows applications to run directly from the flash memory instead of reading the program into the system RAM first. The disadvantages are extremely slow write-and-erase cycles and a bigger cell size compared to NAND-based flash memory, which makes it cost effective in low-capacity data storage which rarely needs to be updated, like in computer BIOS or the firmware of set-top boxes. NAND flash memory has about half the cell size of NOR flash memory and is an ideal solution for high-capacity data storage. It offers fast read and write performance, but lacks the easy memory access of NOR flash memory. Data must be read serially in blocks. Typically block sizes range from hundreds to thousands of bits. This feature disables the use of NAND flash memory as a drop-in replacement for program Read Only Memorys (ROMs), because most microcontrollers and microprocessors need byte-level access. Therefore, NAND flash memory is used in the category of other secondary storage devices like hard disks or optical media (e.g. CD, DVD). It is utilized as mass storage such as memory cards and USB flash drives. Due to the extremely high packing density it was possible to release a new generation of memory card formats exhibiting extremely small feature size. For instance, the microSD\textsuperscript{TM} card has an area of about \(1.5 \text{ cm}^2\), with a thickness of less than \(1 \text{ mm}\) and offers presently up to \(16 \text{ GB}\)\cite{36} of storage capacity at the same time.

2.1.1.1 Working Principle

Flash memory is made out of memory cells which are placed in an array. Every memory cell contains one floating gate transistor allowing to store at least one bit (Fig. 2.2). Single-Level Cells (SLCs), are able to store one bit of information, while Multi-Level Cells (MLCs) allow to store more than one bit per cell, by choosing between multiple levels of electrical charge in the floating gate of a cell.
Figure 2.2: Basic scheme of a flash memory cell. Depending on the charge stored in the floating gate one bit SLC or multiple bits MLC can be saved.

Figure 2.3: Schemes of the basic circuits for NAND and NOR flash memory devices.

2.1.1.2 NOR

In NOR gate flash memory each cell consists of a standard MOSFET with two gates instead of one. The top gate is the so called Control Gate (CG), which is used like a normal MOSFET gate. The second gate below is called Floating Gate (FG) Fig. 2.2. The FG is insulated by a surrounding oxide. Electrons in the FG are trapped and will retain in there for many years, if there is no manipulation from outside. A charge stored in the FG will (partially) screen the electric field from the CG. Thus the threshold voltage $V_t$ of the cell is modified. During the read-out, a voltage, sufficiently low to preserve the amount of charge in the FG, but high enough to distinguish between a charged and uncharged FG is applied to the CG. Depending on the amount of charge in the FG the MOSFET will stay insulating or become conducting at the chosen CG voltage. The current through the channel is sensed as binary information and
represents in combination with the other cells the stored data. In cells which are able to store more than one bit, instead of just sensing if there is current, the amount of current is mapped to a corresponding bit pattern.

This flash type is called "NOR flash", because it behaves like a NOR gate Fig. 2.3. If one of the word lines is brought high, the related floating gate transistor pulls the output bit line low. SLC NOR flash cells are in their default state logically equivalent to binary "1", because by applying a moderate voltage to the control gate current will flow.

The following steps are needed to write or program a NOR flash cell to "0":

- a large voltage of the opposite polarity is applied between gate and source
- the generated large electrical field pulls the electrons out of the floating gate via quantum-mechanical tunneling.

Up-to-date NOR flash chips split their memory into erase segments (also known as blocks or sectors). Erasing can only be performed on a block-wise basis, while the write procedure can be performed on a single byte or word at a time basis.

### 2.1.1.3 NAND

In a NAND flash memory the transistors are connected in series (Fig. 2.4). Only if all word lines are pulled high (above the transistors $V_i$), the bit line is pulled low. This resembles a NAND gate. These groups are connected via some additional transistors to a NOR-style bit line array.

For reading, all word lines except the one to read, are set to a voltage above $V_i$ of a programmed bit, while the bit line for reading is set just over the $V_i$ of an erased bit. The series group will conduct only (and pull the bit line low), if the selected bit has not been programmed. Regardless of the additionall transistors, NAND flash allows a denser layout, due to the reduced bit lines and ground wires. Additionally, NAND flash is usually allowed to exhibit a certain number of faults, while NOR flash is expected to be fault free (e.g. for use in a BIOS ROM).

NAND flash writing operates with tunnel injection and tunnel release for erasing.

Figure 2.4: Scheme of a SONOS gate stack.
Footnotes

... memory\(^2.1\)
    due to their operation as logic NAND and NOR

_T. Windbacher: Engineering Gate Stacks for Field-Effect Transistors_