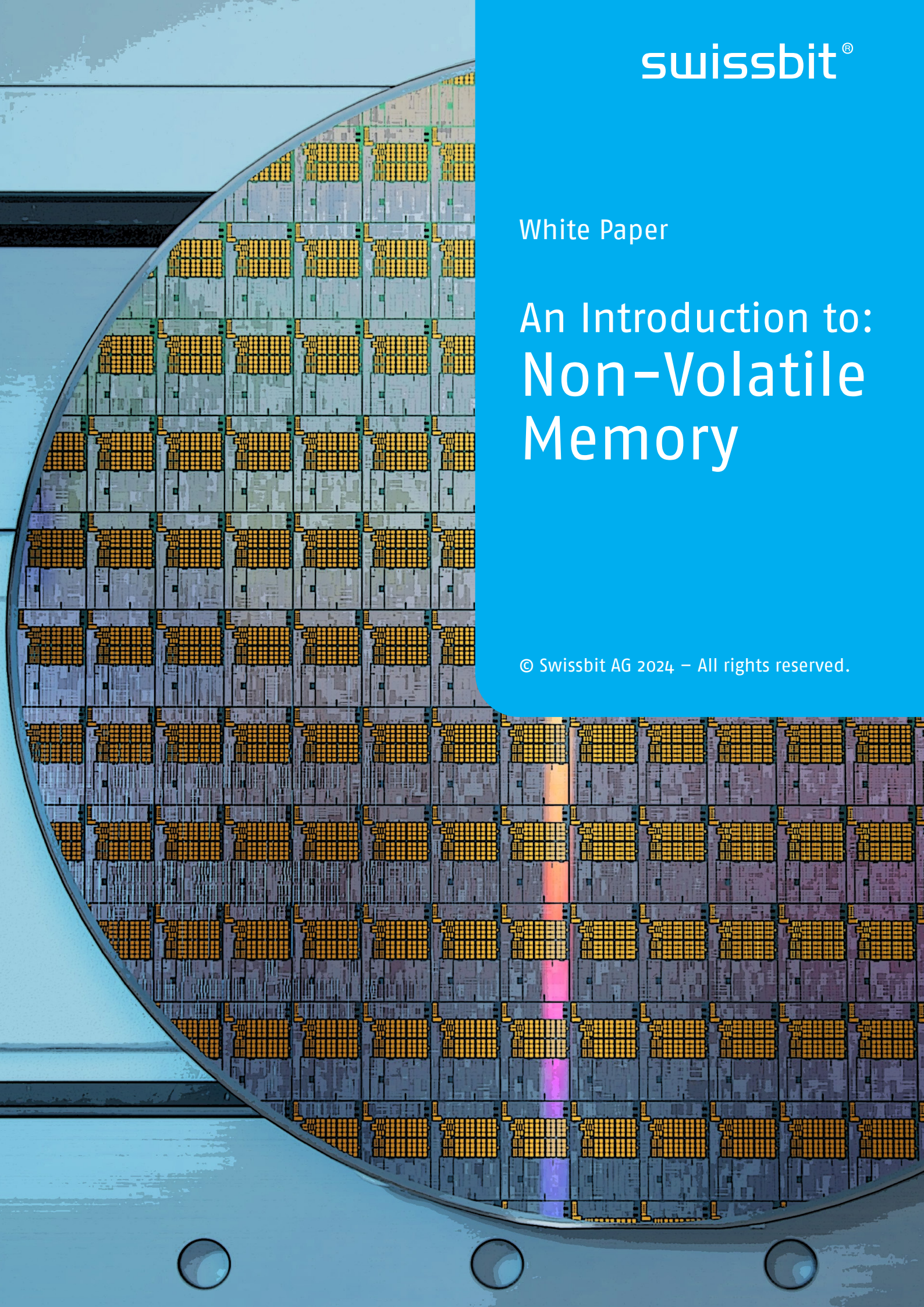


swissbit®

White Paper

An Introduction to: Non-Volatile Memory

swissbit®

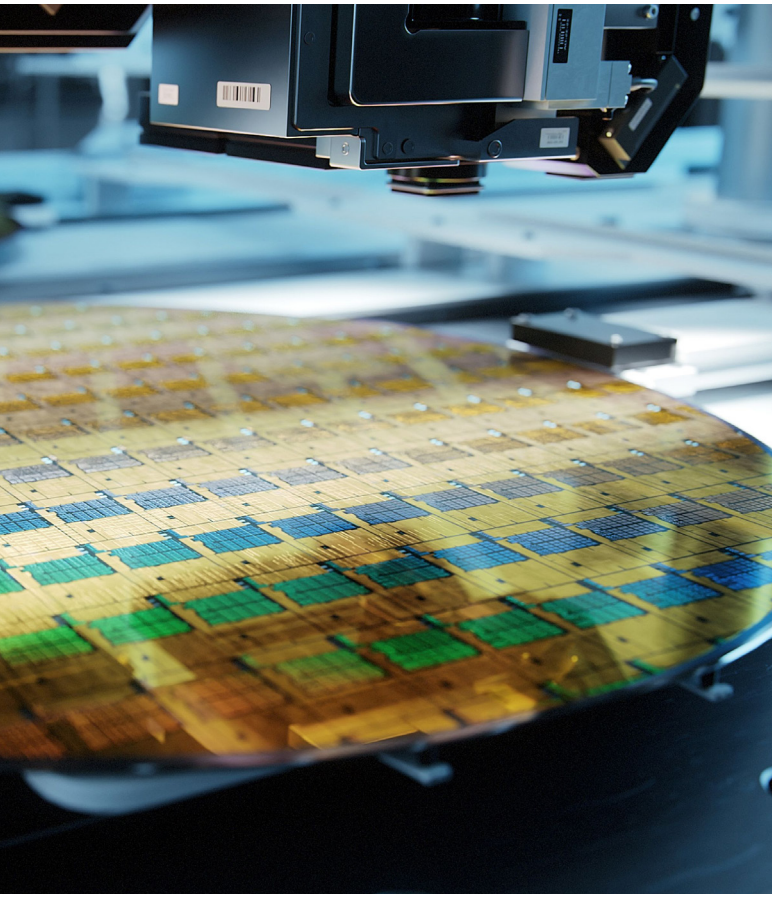
1. Abstract

NAND flash is a very cost competitive non-volatile memory technology. With the introduction of 3D-NAND, NAND based storage media have become serious contenders to replace magnetic rotating storage media in terms of system cost.

However, there are still several challenges to be met, as cost-efficient NAND flashes with high densities based on MLC, TLC or even QLC technologies are increasingly prone to causing failures during field operation, especially in harsh environments. Continuously decreasing reliability and endurance in comparison to SLC flash is the result.

Endurance, lifetime and reliability of NAND flash based storage media depends very much on the quality of the flash memory controller and its firmware, which manages the intrinsic physical properties and weaknesses.

Table of Contents



2. Introduction

This paper provides an overview of NAND flash technology, which is the most widely used storage technology in today's non-volatile memory based storage systems.

NAND flash memory is explained in detail, starting with the transistor, and then covering its structure and functionality. The handling of program, read and erase processes is considered with regard to different technologies, including 3D flash. The shortcomings and challenges of NAND flash are also explained, in conjunction with the solutions necessary to overcome them.

3. NAND Flash Memory

NAND flash is based on standard transistors with additional parts to store data instead of only switching voltage.

Therefore this paper starts with an introduction to transistors – the most important components in the semiconductor industry.

Transistor

A MOSFET transistor (Metal Oxide Semiconductor Field Effect transistor) can be seen as an electrical switch in which the current flow between two terminals (called source and drain) is controlled by a third terminal (gate).

A MOSFET transistor can be structured in two different ways: a p-type substrate and n-type source and drain regions, or an n-type substrate and p-type source and drain regions. The first type is an n-channel transistor, while the second type is a p-channel transistor. A process technology in which the two types of transistors are present is CMOS (Complementary Metal Oxide Transistor).

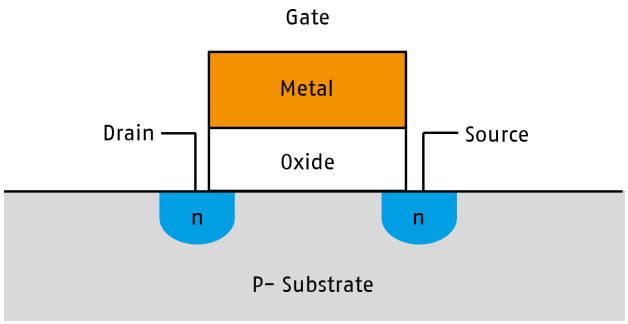


Figure 1: n-type MOSFET transistor

An n-type region is an area of the silicon that has a surplus of electrons, and a p-type region has a lack of electrons. This can be achieved by injecting elements with a surplus of electrons or inducing a lack of them using elements such as Arsenic or Boron in a process called "doping implantation".

For an n-channel transistor to allow current to flow between drain and source, a positive voltage V > Vth must be applied to the gate. Vth is the "threshold voltage" of the transistor. When V > Vth is applied, the electric field generated attracts the electrons toward the gate terminal and a conductive channel is formed between source and drain.

If a positive voltage is applied between drain and source, a current can flow between the two terminals. However, when V < Vth no current can flow, even if a positive voltage is applied between drain and source. The transistor therefore acts as an electrical switch that can be turned on and off depending on the voltage.

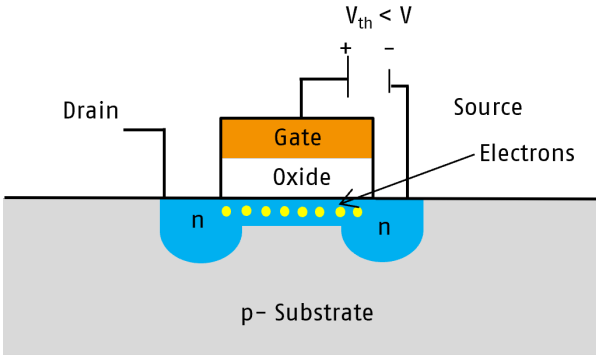


Figure 2: MOSFET where voltage is applied to the gate > threshold. Transistor is switched on and current can flow between source and drain when Vds > 0V

Floating Gate and NAND Flash Cell

A NAND flash cell is an n-channel MOSFET with an additional programming layer inserted between the control gate (CG) and the silicon substrate. This layer is electrically isolated from all the other terminals of the cell and is called the Floating Gate (FG).

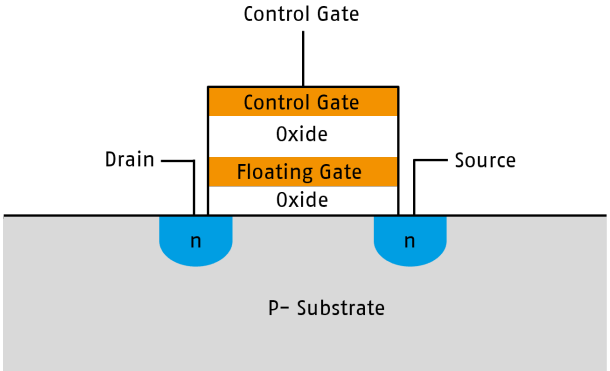


Figure 3: NAND flash cell with floating gate

Programming a NAND Flash bit

By applying a high voltage to the CG (about +20V) while source and drain are grounded, the electrons in the channel can gain enough energy to overcome the oxide barrier and move from the channel into the floating gate.

This is the so-called Fowler–Nordheim tunneling effect of trapping electrons in the floating gate is the programming operation of a flash device (sometimes called the "write" operation). Since the floating gate is electrically isolated it can store electric charge permanently even if the power is removed from the flash device.

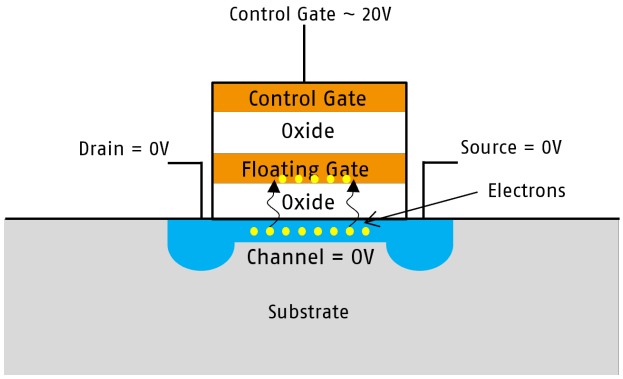
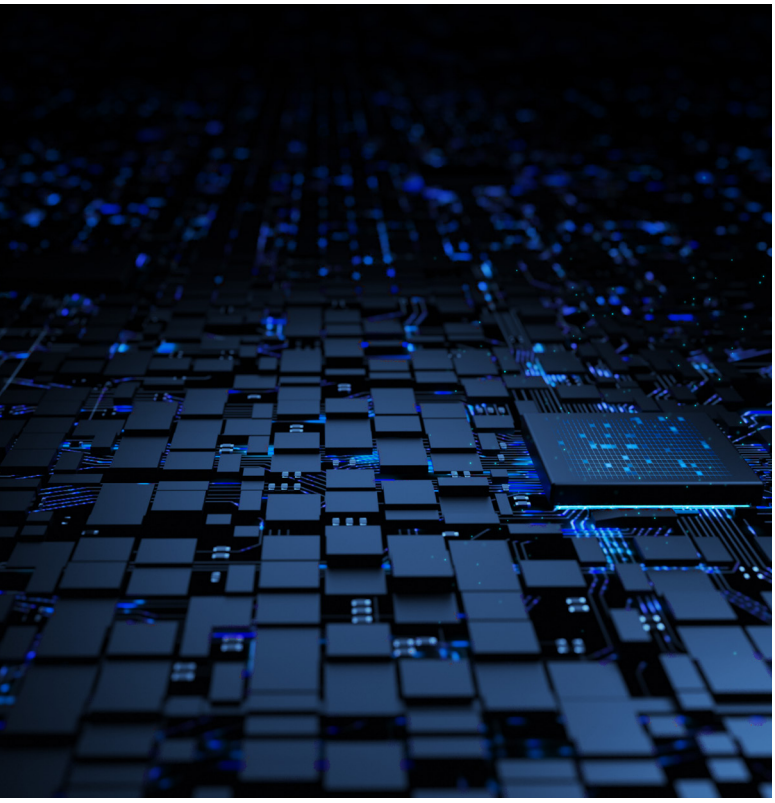


Figure 4: When programming a cell electrons move from the channel to the FG



Reading a NAND Flash bit

Whenever electric charge is stored in the floating gate, the behavior of the NAND cell is influenced. In fact, the charge trapped in the floating gate during the programming operation determines an increase in the threshold voltage, whose entity depends on the number of electrons present in the floating gate. This shift in Vth can be converted into a bit of information.

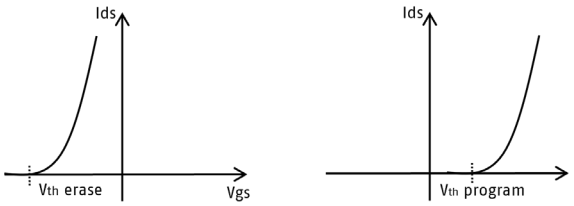


Figure 5: Charactreristic of an erased and programmed bit

Applying voltage Vgs = 0V between the control gate and source has different effects on the two bits. In the case of the programmed bit no current Ids can flow between source and drain because Vgs < Vth_program, while for the erased bit a current can flow because Vgs > Vth_erase. The presence or absence of current Ids is then converted into a bit of information.

Conventionally:
Bit = 0 for the programmed bit
Bit = 1 for the erased bit

Erasing a NAND Flash bit

The erase operation extracts the electrons trapped in the floating gate and brings the threshold voltage of a programmed cell back below zero (Vth_erase < 0V). As a result, a bit flips from 0 to 1. To enable this, a high voltage (about +20V) has to be applied to the substrate of the cell while the control gate is grounded. The erase operation also exploits the Fowler–Nordheim tunneling effect.

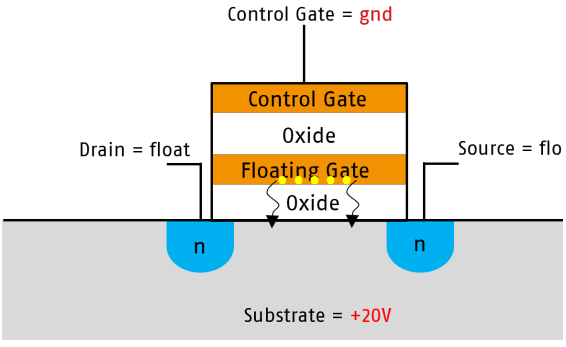


Figure 6: Erase operation of a cell

4. NAND Flash Structure

NAND cells are organized hierarchically, as shown below.

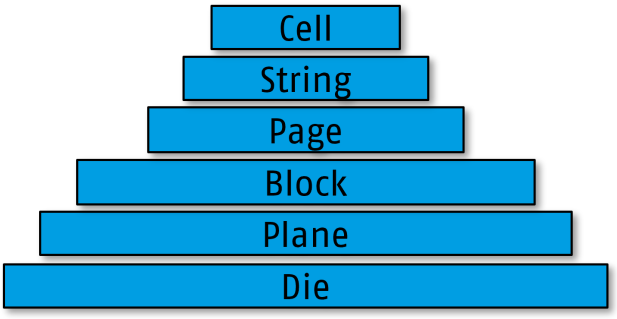


Figure 7: Physical structure of a NAND flash cell

String

A string is a series connection of NAND cells in which the source of one cell is connected to the drain of the next one. Depending on the NAND technology a string typically consists of different amounts of NAND cells (see Fig. 8). Each string has two select transistors – on the top and at the bottom, whose gates are connected to DSL and SSL lines – that allow the selection of the desired strings during different NAND operations (read, program, erase).

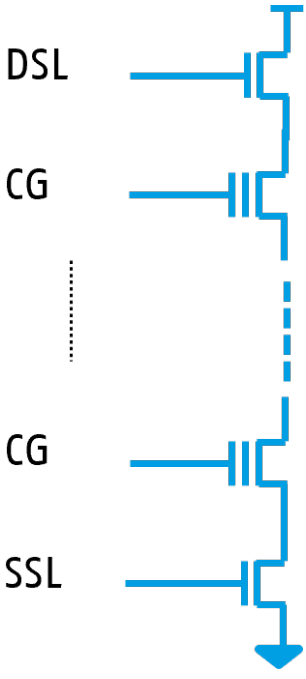


Figure 8: NAND flash string

Page and Block

Strings are organized in blocks, in which each string is connected to a separate line called a bitline (BL<i> in Fig. 9). All cells with the same position in the string are connected through the control gates by a line called wordline (WL<j> in Fig. 9). Depending on the flash device, the cells on the same wordline can be organized in one or two pages.

A page is the minimum selectable array area to be read and programmed. A block is the minimum selectable area in the erase operation.

Typical page size is between 2KB and 16KB.

Block size=(# of pages in a block)*(page size)

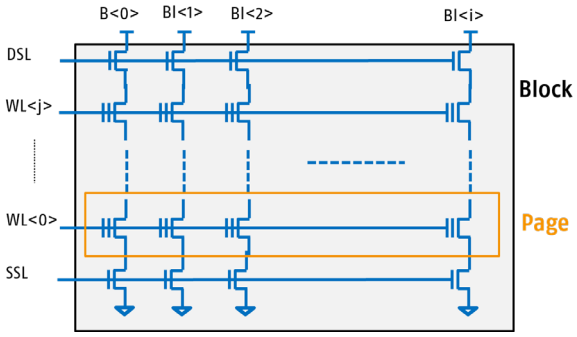
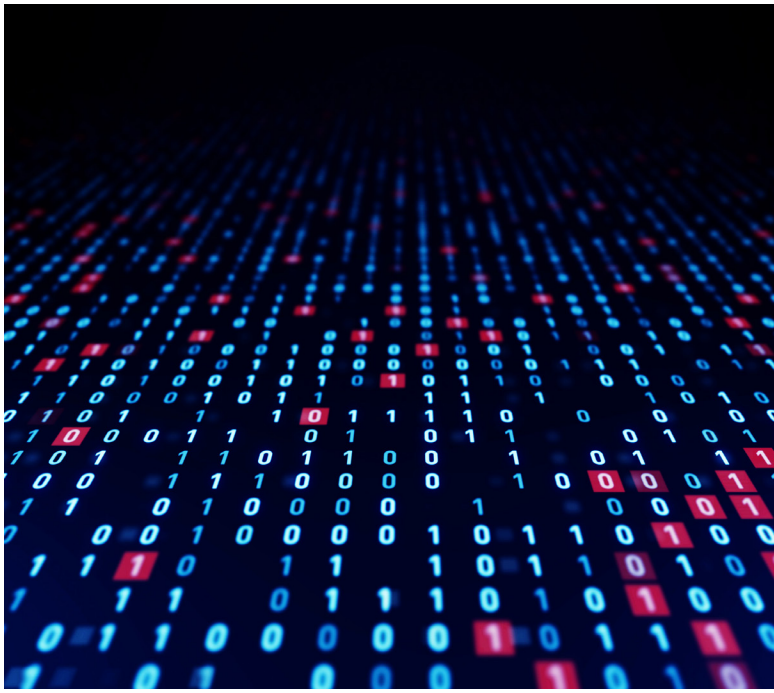


Figure 9: NAND flash block



Plane

A plane contains a certain number of blocks that are connected via the same Bit lines. Blocks have independent word lines to access all pages for individual read or write operations. All blocks and therefore all cells on one plane share the same substrate (see Fig. 10).

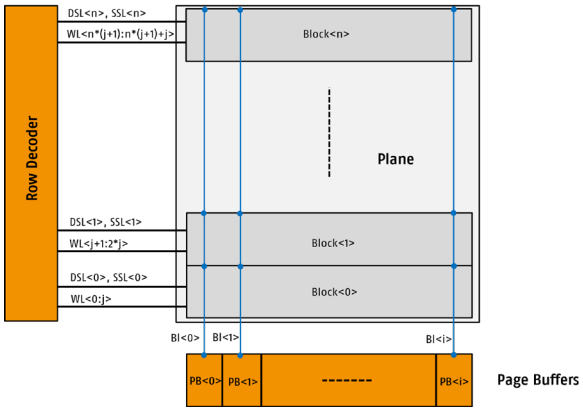


Figure 10: NAND flash plane with PBs and row decoder

Each plane has a dedicated row decoder, which is connected through DSL, SSL and word lines to each block in the plane. The row decoder selects a word line depending on the physical address information received from the flash controller. Each plane also has a dedicated page buffer, which is the dedicated circuitry to read/program data from/to the array.

Flash die

A flash die consists of one or more planes, and the peripheral circuitry that is needed to perform all the read/write/erase operations.

Typically, in dies with multiple planes it is possible to perform read and write operations simultaneously on different planes, which increases performance.

Page program operation

To write a page into a block the data received through the flash interface is transferred into the page buffer. The page buffer transfers the data to be programmed onto the bit lines, while the row decoder selects the block and the word line addressed by the flash controller (Fig. 11). In all other blocks word lines and select transistors are deselected. The selected word line is biased to around 20V. Only the cells on the word line whose (Bit Line) bit is on 0 (i.e. gnd) experience an electric field sufficient to be programmed. However, the cells with 1 (i.e. Vcc) on the bit line do not get programmed. The unselected word lines in the selected block are biased to a Vpass voltage (~10V) that minimizes the program disturbance on all unselected cells of the block.

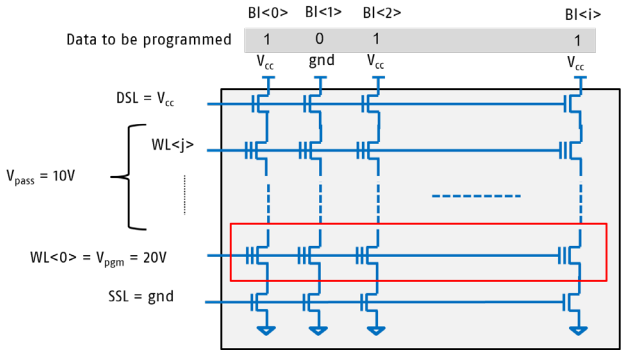


Figure 11: NAND flash plane program (only one page per WL)

Block Erase Operation

The erase operation can only be performed block by block.

To erase a selected block all the word lines in the block need to be biased at 0V and the substrate at 20V. The substrate of a block in one plane is equal to all the other blocks of that plane (see Fig. 11) and therefore the word lines of the unselected blocks that should not be erased and are left floating.

Page Read Operation

The word line of the page to be read is applied with a voltage of 0V and the other unselected word lines belonging to the same block are biased at Vpass (Fig. 12). Vpass is a voltage that has to be higher than the Vth of the programmed bits of the string (typically Vpass is around 5V) so that they behave like pass transistors and do not affect the read operation of the selected word line. All bit lines are precharged to ~1V by the page buffer and then left floating. Only the erased bit with its bias conditions can draw current and discharge the bit lines while bit lines of programmed cells remain charged (Fig. 13). The page buffer samples the status of the bit lines after a certain delay, generating a 0 for the programmed bits and a 1 for the erased ones.

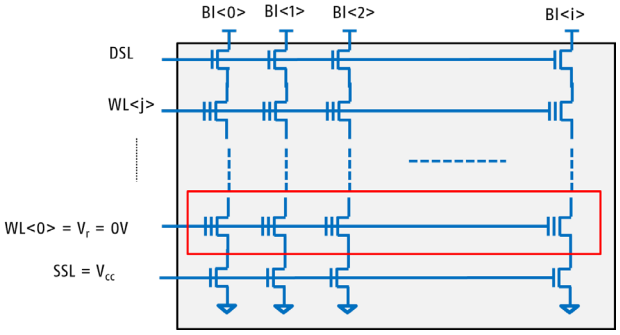


Figure 12: NAND flash page read bias

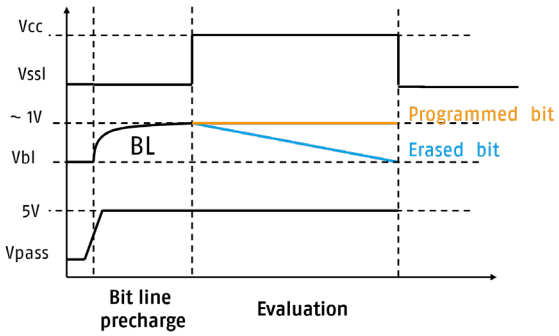


Figure 13: NAND flash page read time diagram

5. SLC, MLC, TLC, QLC

By controlling the quantity of charge trapped in the floating gate of the cell it is possible to accurately control its threshold level and store more than one bit of information in a cell.

Based on this characteristic, different families of NAND flash devices can be defined.

SLC (Single Level Cell)

SLC flash can store one bit of information in each cell. A negative threshold voltage corresponds to a 1, and a positive threshold voltage corresponds to a 0. SLC memory has the advantage of faster write speeds, lower power consumption and higher cell endurance. Programming is faster, because with only one possible threshold state for the programmed bit (Vth > 0V) less accuracy and consequently less time is needed to control the final threshold level. However, because SLC memory stores only one bit per cell, the cost per bit is high. Due to faster transfer speeds and expected higher endurance, SLC flash technology is used in high-performance memory cards. SLC is now used for applications where reliability, endurance and speed are essential. Examples are applications in the fields of industrial automation, networking and robotics.

MLC (Multi Level Cell)

MLC can store two bits of data, which means that four different Vth states are possible. The erased state has, like to SLC, a negative Vth, and the other three states have a different positive Vth level. The advantage of MLC is that it stores more data per cell and is therefore much cheaper than SLC. Conversely, MLC has a lower write performance because the differences between the thresholds are lower and therefore more accuracy is required to program the cells. Furthermore, the bit error rate (i.e. bits that fail to read correctly) is higher and endurance is lower – the smaller threshold distances make it more likely that some bits will end up in the adjacent state. To solve this issue a higher error correction capability is required for MLC devices.

Programming MLC (Multi Level Cell)

In order to write two bits of information to one cell, four different threshold levels should be achieved. The two bits per cell belong to two different pages (i.e. two different physical addresses) called the "lower" and "upper" page (or the "fast" and "slow" page). As a first step, the lower page is programmed in an SLC-like mode, as only one bit per cell is present (see Fig. 14). In step two (see Fig. 15), the upper page is programmed, and the bit of the upper page combined with the current lower bit status defines the final threshold level target of each cell (see Fig. 16).

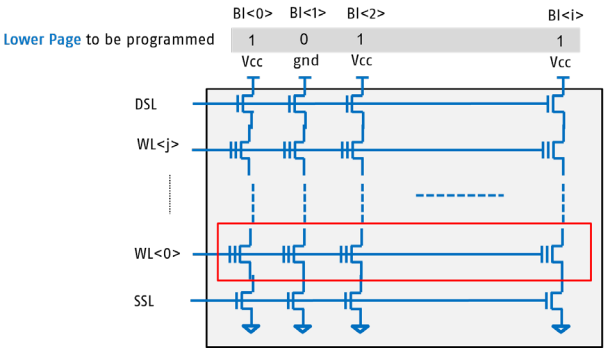


Figure 14: Programming the lower page

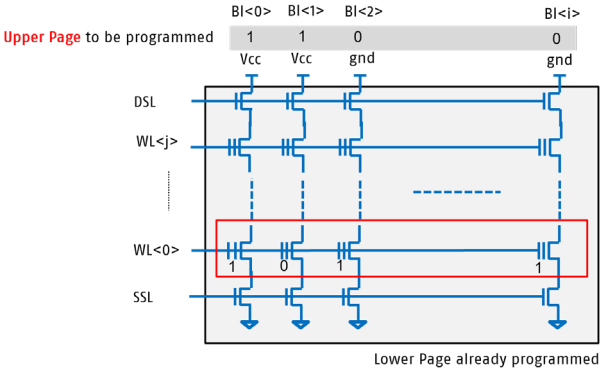


Figure 15: Programming the upper page

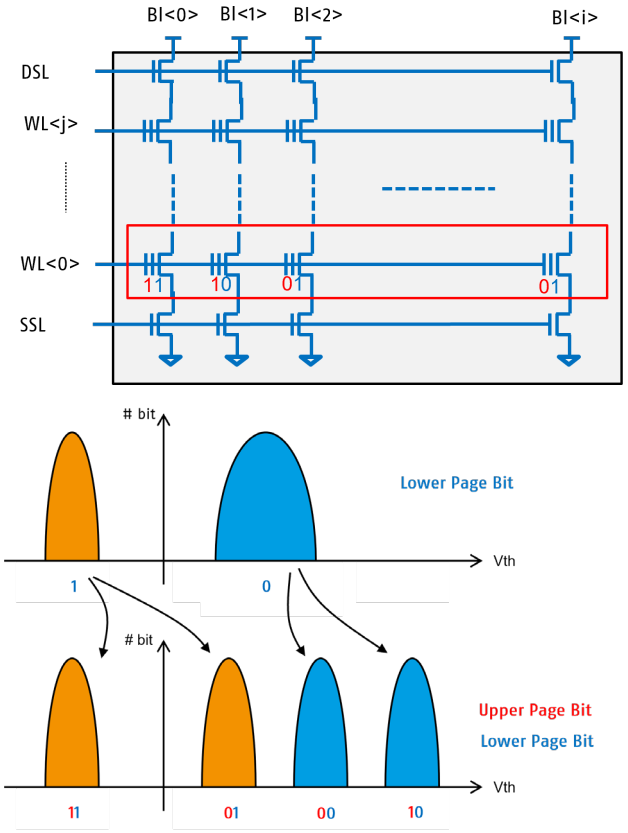


Figure 16: Lower and upper page programming states

In order to achieve the tight control of the threshold voltage of the cells required by the MLC operation, an Incremental Step Pulse Programming (ISPP) technique is used. With ISPP at each programming pulse the voltage on the control gate is increased by a Vpulse, which causes an equivalent incremental increase of the threshold voltage of the cell Vth. This means that the cells are exposed to evenly increasing voltage pulses (Vpulse), and after each increased pulse the state of the cell is verified. All the cells that have reached the desired Vth state stop getting additional programming pulses. This process continues until all cells have reached their target Vth state. SLC also uses the ISPP technique, but the Vpulse of each pulse is much larger than the pulses for MLC because the required control of the program distribution width can be less accurate. This results in faster programming operation.

TLC (Triple Level Cell)

TLC can store three bits per cell, resulting in eight different states and eight different Vth levels. The resulting higher error rate requires advanced error correction algorithms. TLC uses the ISPP technique, but with finer Vpulse steps to achieve greater accuracy. The primary benefit of TLC is its lower cost per bit of storage due to the higher data density. With improvement in technology, todays industrial grade TLC flash shows the same endurance and reliability like the outdated MLC flash. It is therefore currently the standard for applications in the automotive industry.

QLC (Quad Level Cell)

In a QLC NAND four bits can be stored in each cell, which corresponds to 16 different threshold levels. With its limitations regarding endurance and temperature range, QLC is currently only suitable for consumer devices or read extensive applications in temperature controlled environments.

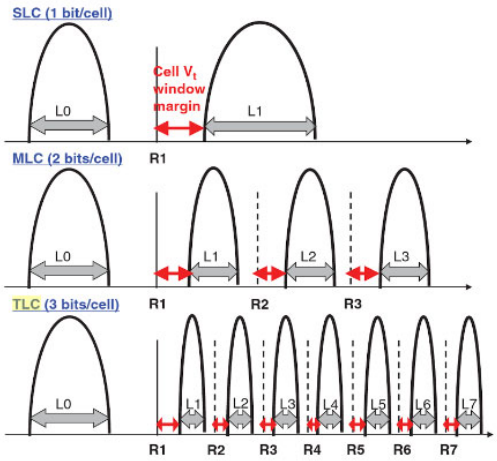


Figure 17: SLC, MLC and TLC Vth and their different states

pSLC (pseudo Single Level Cell)

A pSLC NAND is a standard MLC/ TLC QLC NAND where only one out of two bits is used. In other words, pSLC is a particular usage model of an MLC/ TLC/ QLC flash. By storing only one bit per cell, endurance, reliability and data retention are increased. The parameters described are significantly better if compared to higher density flashes and are close to true SLC while the costs are significatly lower.

6. NAND Flash Weaknesses

The endurance and reliability of NAND flash devices are impacted by several different intrinsic effects that are becoming more critical as the technology nodes shrink. The following provides a brief overview of the main issues.

Program/ Erase (P/E) Cycles)

A NAND flash cell can sustain only a limited number of program and erase cycles. The high voltages that are required during program and erase operations cause a small amount of damage to the cell with every cycle. As a consequence, the cell becomes harder to erase. Furthermore, the isolation characteristics of the two oxides that isolate the floating gate worsen. This reduces the capability of the cell to retain the electrons that are stored in the floating gate, resulting in a degradation of the data retention performance. When the maximum number of P/E cycles allowed by the specification is reached, the flash is said to have reached End of Life (EOL).

Data Retention

Data retention is the ability of a cell to retain the stored information (i.e. the electrons in the FG) across time – by definition it refers only to a programmed bit. A leakage of electrons from the floating gate causes a shift toward the erase state of the programmed bits. If the threshold shift amount is so high that a bit crosses the read reference voltage, that bit will fail a subsequent read operation. Data retention performance can be significantly reduced by read/write or high temperatures. Data retention performance of MLC and TLC devices is considerably lower than SLC.

Program Disturb

Program disturb takes place during page programming, and causes some charge to be collected in the FGs of cells that should not be programmed. The effect is to soft-program bits, i.e. to increase the threshold of erased bits for SLC devices or those in the lower programmed distributions for MLC and TLC. If the amount of the disturb is sufficient to move a bit beyond the read reference voltage (see Fig. 18 for an SLC read), that bit will cause a bit read failure in a next-page read operation.

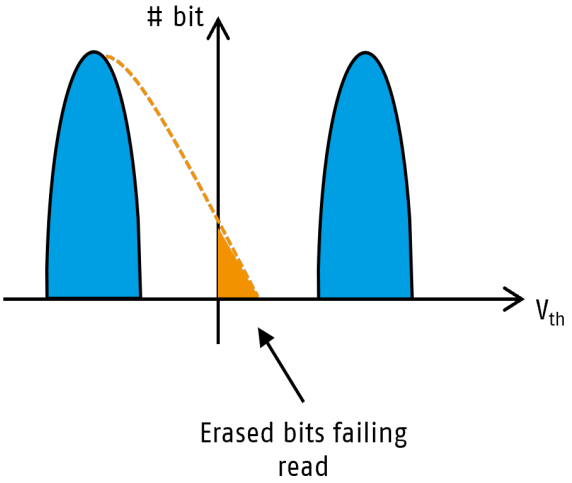


Figure 18: Program disturb effect in SLC NAND Flash

Referring to Fig. 19 below, when programming bit B, the disturb mainly affects the bits on the same word line that do not have to be programmed (bits A and C, for instance), because they are biased at the same high gate voltage (~20V). Cells on unselected wordlline of the same block (WL<1:j> in Fig. 18) are also disturbed, but less strongly than on the selected wordline because the applied voltage is lower (Vpass ~10V).

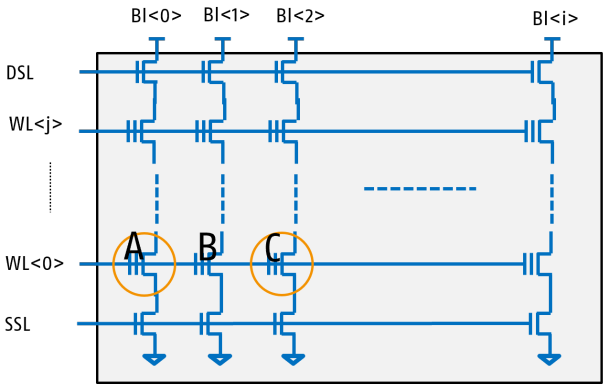


Figure 19: Program disturb

Program disturb only affects the bits in the same block of a page that is programmed. All the other blocks are not impacted. Furthermore, program disturb is not a permanent effect. After an erase cycle cells that were previously disturbed will program and erase normally.

Cell-to-cell interference

The effect by which programming one cell disturbs the adjacent ones is called cell-to-cell interference. The reason is that the electrons stored in the floating gate of the programmed bits modify the potential of the nearby floating gates. This causes a shift of the threshold voltage toward programmed values. In a NAND flash block each cell of the block is surrounded by eight adjacent cells (excluding those at the boundary of the block, as in Fig. 20). If we suppose cell A to be erased and to program all of the eight adjacent cells the disturb can be significant enough to cause bit A to be programmed in a subsequent read.

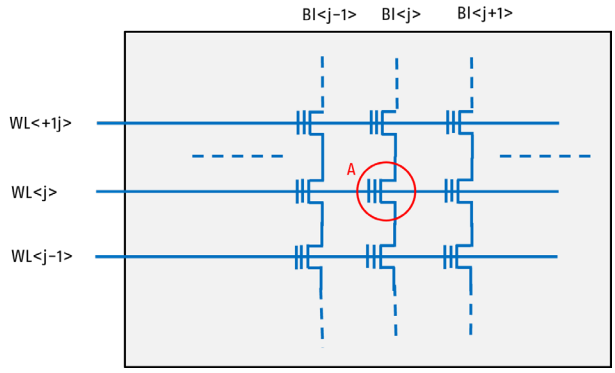


Figure 20: Cell-to-cell interference

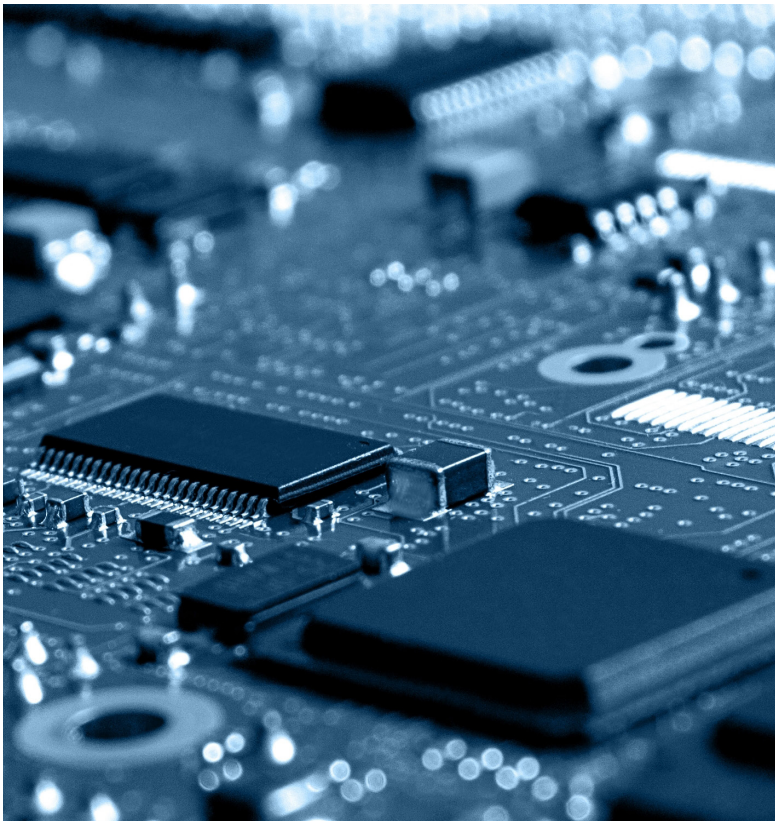
This disturb is proportional to the coupling capacitance of cell A toward the adjacent ones, and so it is more evident in recent technologies where distances between cells are very small, creating bigger coupling capacitances. MLC and TLC NAND are also more affected by this than SLC NAND, because the tolerated threshold shift is lower. To avoid programming patterns like the one described above (or at least to reduce the probability) it is necessary to use a randomizer to scramble data. Moreover, the devices require page programming in sequential order so the coupling effect can be taken into account during each program operation.

Read Disturb

As with program disturb, read disturb also causes a shift to a higher value in the threshold voltage of the cells in the block in which a page is read. Because the voltages involved in the read operations are much lower than during programming, many read cycles are needed to cause a read failure (normally more than 10K). Read disturb is not permanent, and the cells that were disturbed previously will program and erase normally after an erase cycle.

Bad Blocks

NAND flash devices are commonly delivered with some blocks (typically one to two percent of the total block number) that are not working correctly due to manufacturing defects. These blocks are called bad blocks, and should not be used. As well as these intrinsic bad blocks, some additional bad blocks may occur during the lifetime of the device, caused by failures during program or erase operations. The flash controller manages the handling of these new bad blocks, transferring the valid data to a new block and marking the failing ones as bad.



4. 3D NAND

To overcome the scaling limitations of 2D NAND flash, a new technology was introduced in which the memory cells are stacked vertically in multiple layers. This emerging technology is called 3D NAND. It has been developed to achieve higher densities at a lower cost per bit.

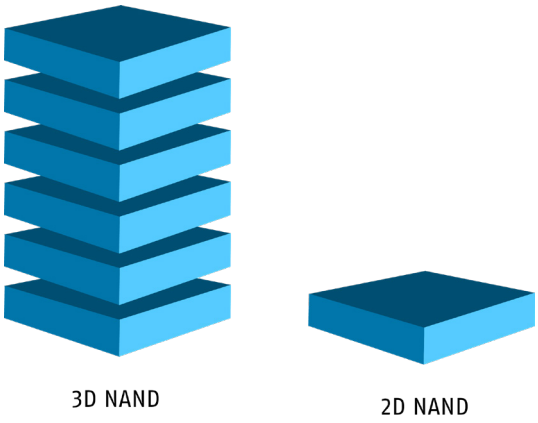


Figure 21: 2D vs 3D NAND flash

3D NAND Architecture

Meanwhile all 3D NAND devices changed from floating gate to charge trap technonoly.

In charge trap the storage layer is formed using an insulator, normally a layer of silicon nitride. Using an insulator as the means of storage enables the manufacturing process to be simplified.

3D NAND has some significant advantages over 2D NAND flash:

- It uses more relaxed technology nodes
- 3D architecture is less affected by cell-to-cell interference, allowing tighter distributions and greater endurance
- Less cell-to-cell interference enables faster programming by using a single-pass programming algorithm where two or three pages for MLC and TLC can be written simultaneously
- Single-pass programming is more energy efficient than single-page programming

However, there are some drawbacks:

- Higher endurance is achieved at a higher Error Correction Code (ECC) requirements cost
- The usage of charge trap technology can lead to lower data retention characteristics
- Longer erase time

swissbit®

Do you have any questions? Get in touch!

Europe
+49 (30) 936 954 400
sales@swissbit.com

USA
+1 (978) 490 3252
salesna@swissbit.com

About Swissbit

Swissbit AG is the leading European manufacturer of storage, security and embedded IoT solutions for demanding applications. As trusted partner, Swissbit empowers the digital and connected world by reliably storing and protecting data in industrial, security and IoT applications.

www.swissbit.com