Notes of Solid-State Drive

Chapter 1 Introduction

As modern computing systems (e.g., enterprise servers, data center storage, and consumer devices) process a large amount of data at an unprecedented scale, **a storage device needs to meet high requirements on storage capacity and I/O performance**. While *electromechanical disk drives* have continuously ramped in capacity, the rotating-storage technology does not provide the access time or transfer-rate performance required in demanding enterprise applications, including online transaction processing, data mining, and cloud computing. Client applications are also in need of an alternative to electromechanical disk drives that can deliver faster response times, use less power, and fit in smaller mobile form factors.

NAND flash memory has become the de facto standard for architecting a storage device in modern computing systems. A NAND flash-based solid-state drive (SSD) is a storage device that utilizes NAND flash memory to store user data. **NAND flash memory has higher performance** (i.e., low latency and high bandwidth) than magnetic storage with one or more rigid, rapidly rotating platters. Also, its capacity has continuously increased, and its costs have continuously decreased over the decades. Given these advantages, NAND flash-based SSDs can provide orders-of-magnitude higher I/O performance (i.e., lower read & write access time and higher random-access input/output operations) compared to traditional hard-disk drives (HDDs), (Also, resilient to physical shock, a small form factor, consuming less static power) with a much lower cost-per-bit value over SSDs based on emerging non-volatile memory (NVM) technologies. Thus, SSD is now at the point where the drives can serve as rotating storage replacements.

NAND flash memory has several unique characteristics, such as the erase-before-write property (i.e., a flash cell needs to be first erased before programming it), limited lifetime (i.e., a cell cannot reliably store data after experiencing a certain number of program/erase (P/E) cycles), and large operation units (e.g., modern NAND flash memory typically reads/writes data in a page (e.g., 16 KiB) granularity). To achieve high performance and large capacity of the storage system while hiding the unique characteristics of NAND flash memory, it is critical to design efficient SSD firmware, commonly called Flash-Translation Layer (FTL). An FTL is responsible for many critical management tasks, such as address translation, garbage collection, wear leveling, and I/O scheduling, which significantly affect the performance, reliability, and lifetime of the SSD.

Modern solid-state drives can be abstracted into three levels: (1) NAND Flash physics, (2) integrated circuit architecture, and (3) SSD firmware, as shown in Figure 1.*1*. At the lowest level of abstraction is the NAND Flash physics level, which describes the motion of electrons for the basic operations (i.e., program, read, and erase) for a single NAND Flash memory cell. Above the physics level is engineering level: the integrated circuit architecture and SSD firmware levels. (mark) .

SSD Firmware	Software
Integrated Circuit Architecture	Hardware
NAND Flash Physics	Electrons

Figure 1.1 Abstraction Levels of Modern Solid-State Drive

In all SSDs, a Flash microcontroller sits between one or multiple hosts (i.e., CPUs) and NAND Flash memories, and on each side, there are a lot of challenges that designers need to overcome. Moreover, a single controller can have multiple cores, with all the complexity associated with developing a multi-threaded firmware. (mark) As usual, simulation speed and precision do not go hand in hand, so it is important to understand when to simulate what.

We will first dive into the lowest level, NAND Flash physics, to understand the characteristics of NAND Flash since its unique properties decide the upper levels' structures.

Note the NAND Flash's unique properties are including:
- Large operation units
- Erase-before-write property
- Asymmetry in operation units
- Limited endurance
- Various error sources
- Asymmetry in operation latencies

(retention loss, schematic, wear leveling->for endurance, Error Correction Code->for reliability, soft decoding, randomization, read retry)

Chapter 2 NAND Flash Physics

Benefiting from two key trends: (1) effective process technology scaling and (2) multi-level (e.g., MLC, TLC) cell data coding, NAND flash memory becomes ubiquitous in everyday life today (e.g., Flash cards, USB keys, and Solid-State Drives). Unfortunately, the reliability of raw data stored in flash memory has also become more challenging to ensure due to these two trends. Manufacturing process scaling, which has increased the number of flash memory cells within a fixed area, makes fewer electrons in the flash memory cell floating gate to represent the data. Multi-level cell data coding, which represents more than one bit of digital data in a single floating-gate transistor, results in larger cell-to-cell interference and disturbance effects. Without mitigation, worsening reliability can reduce the lifetime of NAND flash memory.

To develop mitigation mechanisms, we first need to understand the inside of NAND flash memory.

2.1. NAND Flash Cell

NAND flash memory stores data as the threshold voltage of each flash cell, which is a special type of nMOS transistor. NAND flash cell is based on the Floating Gate (FG) technology, whose cross-section and symbol are shown in *Figure 2.1*. This *floating-gate transistor* is built with two overlapping gates rather than a single one: the first one is contacted to form the gate terminal, while the second one is completely surrounded by oxide. On top of the isolated gate is an interpoly oxide layer, and at the bottom is a tunnel oxide layer. As a result, **this isolated gate constitutes an excellent "trap" for electrons**, which do not discharge even when flash memory is powered off (Without the connection of a supply voltage) over years of charge retention guarantee.

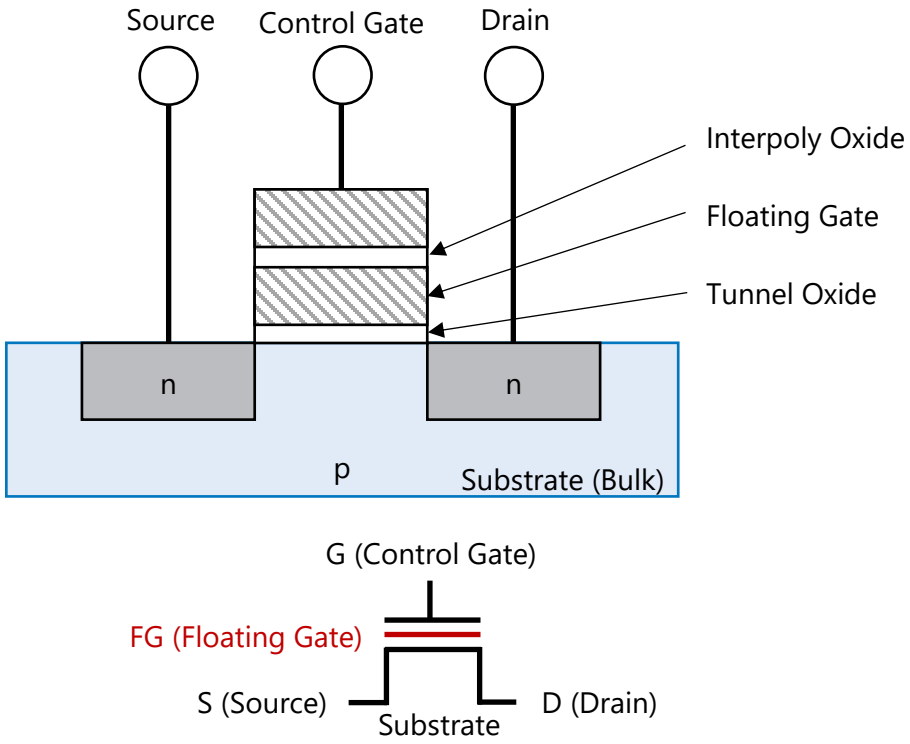
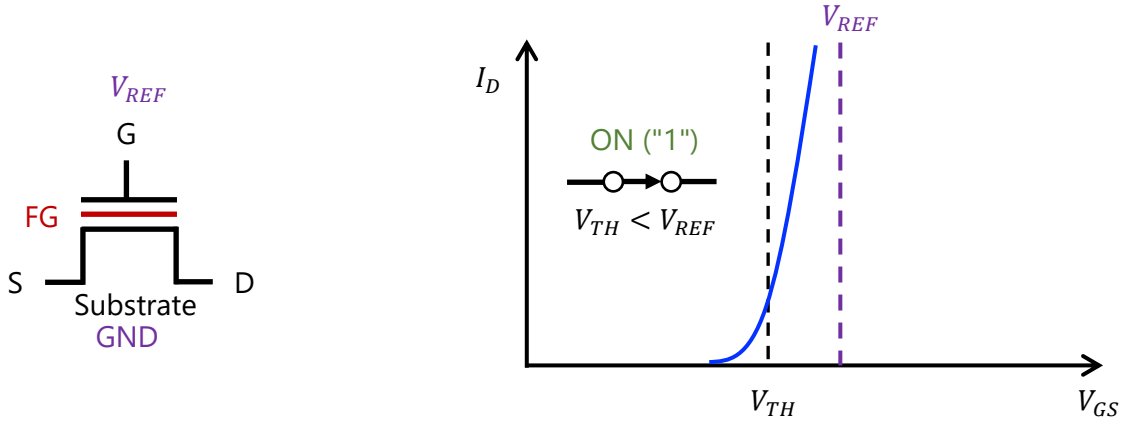


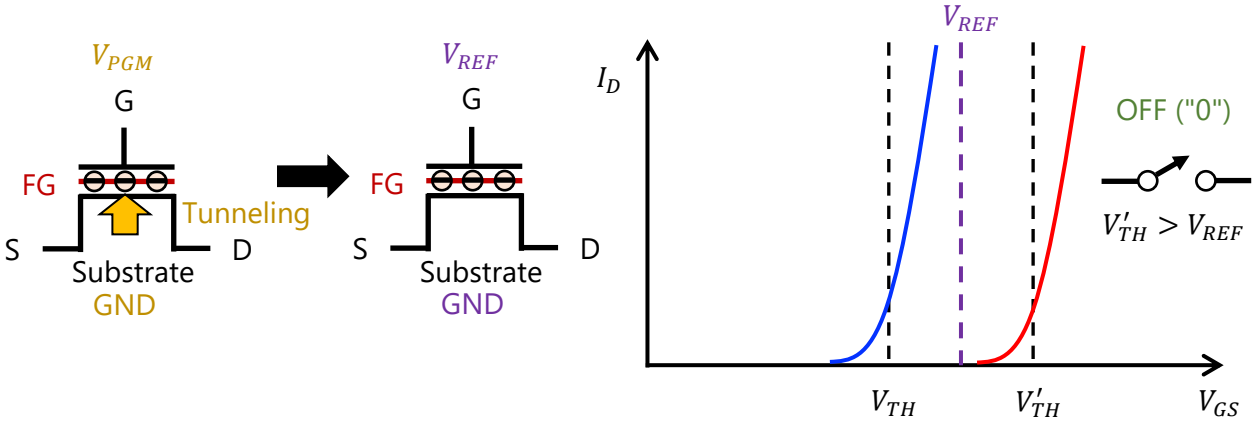
Figure 2.1 Floating Gate Memory Cell

Program, Read and Erase of a Single Floating Gate Cell

The operations performed to inject and remove electrons from the isolated gate are called program and erase, respectively. These operations modify the threshold voltage V_{TH} of the memory cell based on the *Fowler-Nordheim* (FN) *tunneling mechanism*. Earlier NAND flash chips stored a single bit of data in each cell, which was referred to as **single-level cell (SLC)** NAND flash. Applying a high program voltage (V_{PGM}, e.g., 8/18/26 V) to the cell's control gate and keeping the bulk terminal at ground potential (0 V) results in a large current transfer through the whole FG cell stack. This can set (Program operation) transistor to a specific threshold voltage within a fixed range of voltages, as shown in *Figure 2.2*. SLC NAND flash divided this fixed range into two voltage windows: one window represents the bit value 0 and the other represents the bit value 1, which we call the ER (Erase) and P1 (Program 1) states. Applying a fixed voltage (V_{REF}) to the cell's terminals allows discrimination between two storage levels (Read operation): when the gate voltage is higher than the cell's V_{TH}, the cell is ON ("1"), otherwise it is OFF ("0").



(**a**) Applying V_{REF} to Flash Cell with No Charge, and Its Current-Voltage (I-V) Characteristics



(**b**) Applying V_{PGM} to Charge Cell, and Its Current-Voltage (I-V) Characteristics

Figure 2.2 Read (**a**) and Program (**b**) operation of SLC NAND Flash Cell

In other words, each state representing a different value is assigned to a voltage window within the range of all possible threshold voltages. Due to *variation across program operations*, the

threshold voltage of flash cells programmed to the same state is initially distributed across this voltage window. Note that programming a floating gate cell to a specific V_{TH} state is typically accomplished by the so-called "incremental step pulse programming" (ISPP) scheme.

The floating gate memory cell erase works principally in the same way, but with control gate voltages negative with respect to the cell channel region.

Multi-level cell (MLC) NAND flash divides the flash voltage range into four voltage windows that represent each possible 2-bit value (00, 01, 10, and 11), which we call ER, P1, P2, and P3 states. Each voltage window in MLC NAND flash is therefore much smaller than a voltage window in SLC NAND flash. This makes it more difficult to identify the value stored in a cell. Moreover, **triple-level cell (TLC)** NAND flash further divides the range, providing eight voltage windows to represent a 3-bit value, which we call ER, P1-P7 states. And, Quadruple-level cell (QLC) NAND flash stores a 4-bit value per cell.

Encoding more bits per cell increases the capacity of the SSD without increasing the chip size, yet it also decreases reliability by making it more difficult to correctly store and read the bits.

It is worth mentioning that, due to floating gate scalability reasons, charge trap memories are gaining more and more attention, together with their 3D evolution.

(V_{TH} distribution)
(charge trap transistors for 3D)
(retention loss, wear out effect)

2.2. NAND Flash Array

We will use schematic diagrams to delineate the NAND flash array from here. **In schematic diagrams, wires are always joined at three-way junctions**. They are joined at four-way junctions only if a dot is shown. **The slash across the input wire indicates that the gate may receive multiple inputs**.

Flash memory for non-volatile data storage was introduced commercially in the mid-1980s. Since then, common ground **NOR and NAND architecture** have become the most common memory array architectures. Traditionally, *NOR Flash is used for code storage due to faster memory cell access. NAND Flash is used for mass data storage because of its higher memory density, enabling higher storage capacities*.

The difference in memory cell area can be seen from the schematic NOR and NAND array diagrams in *Figure 2.4*. In the NOR array, two memory cells share one contact with the ground (SL: Source Line) and one contact with the bit line (BL), as shown in *Figure 2.4*(a). This results in an effective memory cell area of about $10\ \mathrm{F}^2$ for 6 cells. *Figure 2.4*(b) shows the so-called NAND string with up to $m+1$ memory cells connected in a row. For the NAND string operation, two additional select transistor devices need to be added: (1) Drain Select Line (GSL) or String Select Line (SSL), and (2) Source Select Line (SSL) or Ground Select Line (GSL). These additional structures cause the effective cell area consumption slightly higher than $4\ \mathrm{F}^2$ for 64 cells, the theoretically

smallest effective cell size. Note only NAND Flash is a viable option for SSD applications due to the required high memory capacity and bit cost structure.

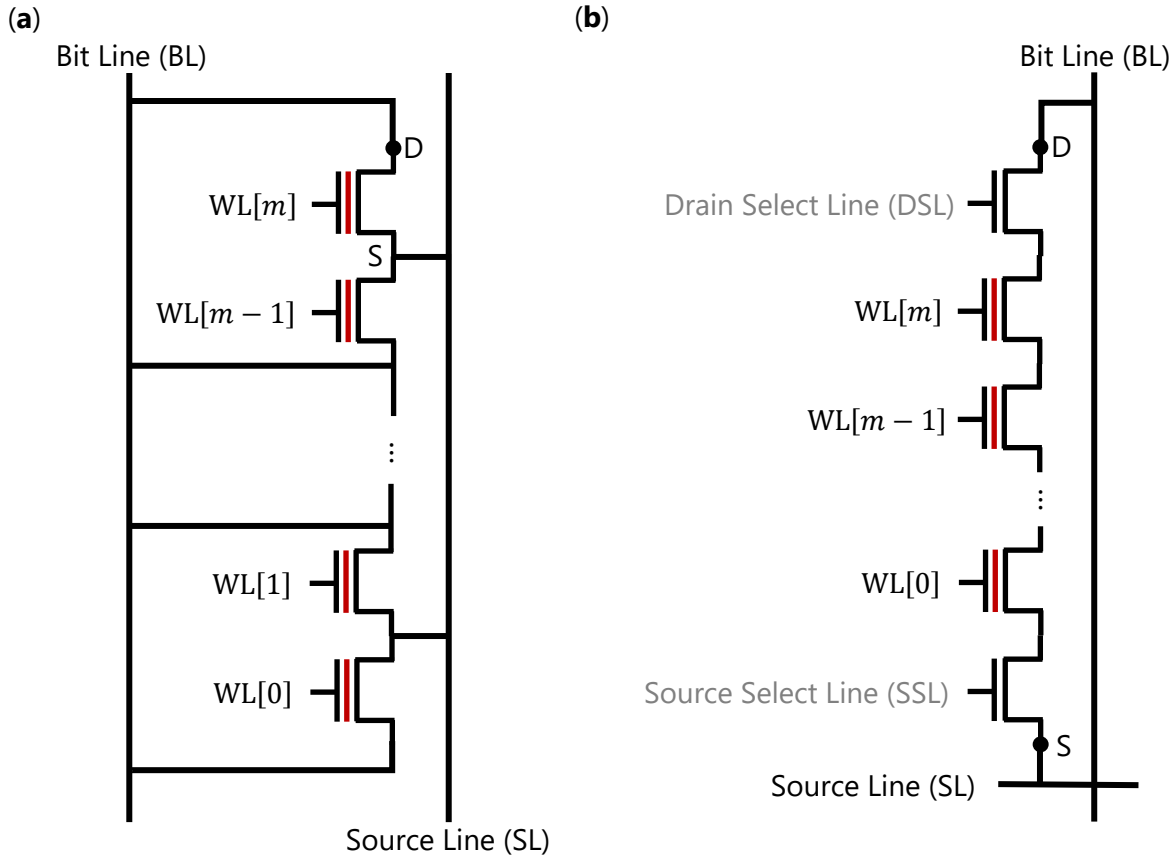


Figure 2.3 Schematic memory cell organization of the NOR array (**a**) and the NAND array (**b**). The wordlines (WL) run perpendicular to the bit lines (BL)

The basic element of a NAND Flash memory: **NAND string** as shown in Figure 2.*4*(a), is made up of multiple cells (e.g., 128) connected serially. Two selection transistors are placed at the edges of the string: (1) the transistor connecting the drain select line ensures the connection to the bit line; (2) the transistor connecting the source select line ensures the connection to the source line. The cell's control gates are connected through the wordlines (WLs). Figure 2.*4*(b) shows how the matrix array (**NAND array**) is built starting from the basic string. In the WL direction, adjacent NAND strings share the same WL, DSL (Drain select line), SSL (Source select line), and SL (Source line). In the BL direction, two consecutive strings share the bit line contact.

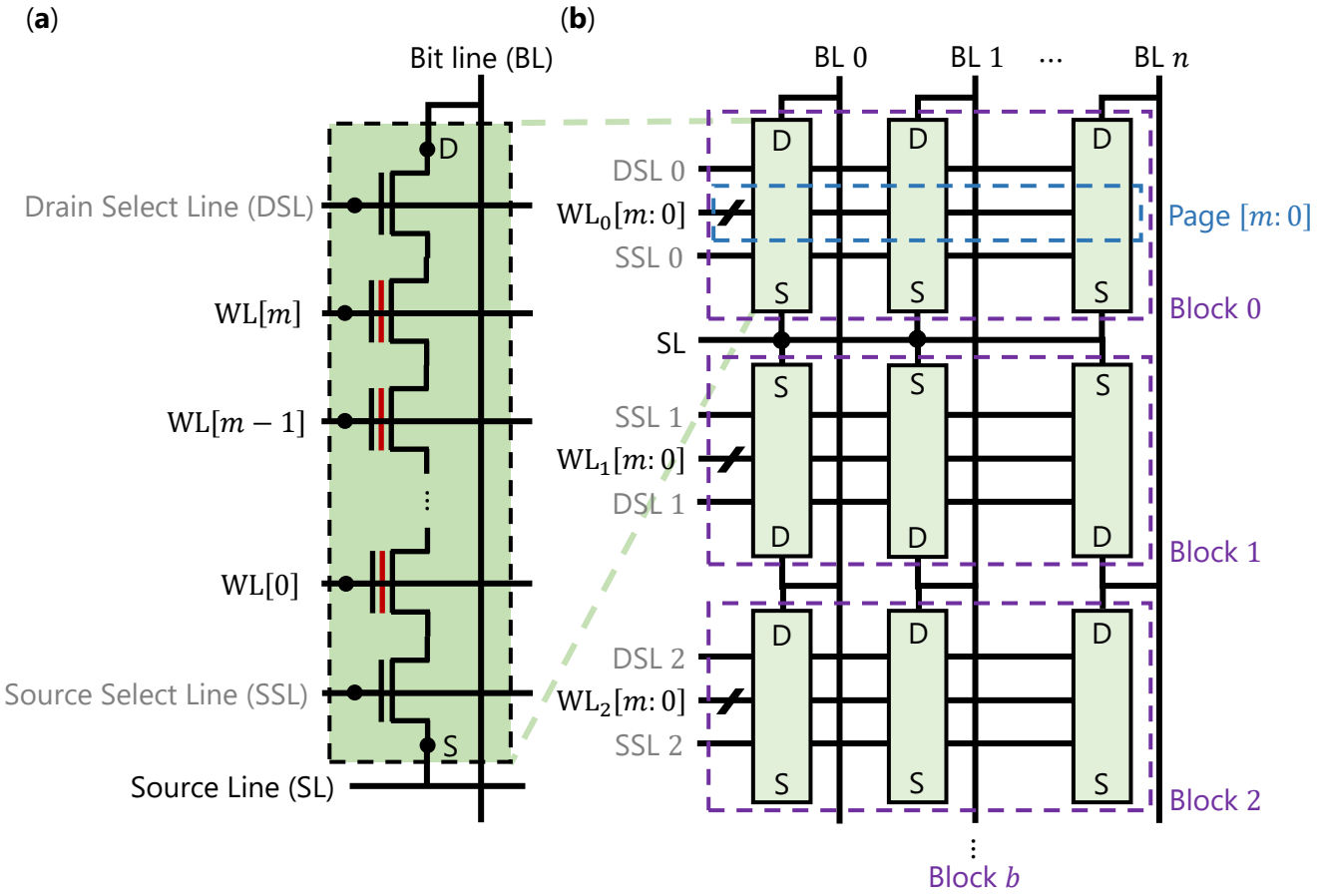


Figure 2.4 NAND String (**a**) and NAND Array (**b**)

All the NAND strings sharing the same group of WL's form **a block**. In Figure 2.*4*(b), each block is made up by WL[m: 0]. **A page** is made up of cells belonging to the same WL. *The number of pages per WL is related to the storage capabilities of the memory cell*. Depending on the number of storage levels, Flash memories are referred to in different ways:

- SLC memories stores 1 bit per cell;
- MLC memories stores 2 bits per cell;
- TLC memories stores 3 bits per cell;
- QLC memories stores 4 bits per cell.

Program, Erase, and Read of Flash Cells in the NAND String

When a large number of a floating gate cells need to be operated in the NAND array, it has to be taken into account that one floating gate cell is located at every crossing point of bit lines and word lines. Therefore, the memory cells in the NAND array cannot be operated independently of each other anymore. In the word line direction (depending on the page size), a couple of thousand FG cells are controlled by the same word line. In bit line direction, the string size (e.g., 64-66 cells) defines the number of cells that cannot be operated independently.

Consequently, it is very important to bear in mind what is happening with all neighboring cells when one cell is treated. This is even more important since the threshold voltage of each memory cell needs to be carefully adjusted as shown for SLC and MLC cells in *Figure 2.5*.

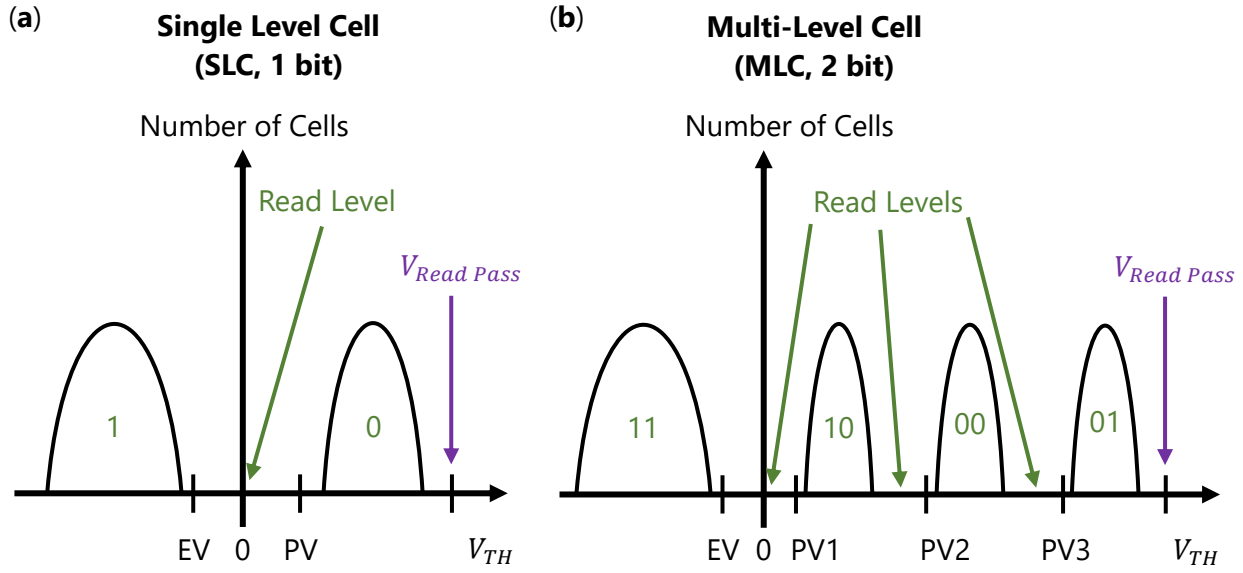


Figure 2.5 Memory cell threshold voltage distributions for one bit per cell (SLC) data storage (**a**) and two bit per cell (MLC) data storage (**b**) in a NAND flash array

The V_{TH} distribution of erased cells is placed at negative V_{TH} values. In an ISPP-like sequence, the erase voltage is increased until all cells are erased below the Erase Verify (EV) level. **The V_{TH} distributions of programmed cells** are placed in the positive V_{TH} range. For a single-level cell (SLC), the ISPP programming is continued until all cells designated for programming are above the Program Verify (PV) level. Consequently, in the case of multi-level cells (MLC), there are three program verify levels (PV1, PV2, and PV3). In addition, the margins between the different programmed V_{TH} distributions must be guaranteed to be large enough to place the read levels (V_{REF}) and have sufficient space for charge/retention loss-caused V_{TH} reductions. A specific distribution shaping algorithm with a small program step increase in certain stages of ISPP programming is necessary to obtain these kinds of narrow cell V_{TH}.

(What is read pass voltage?)

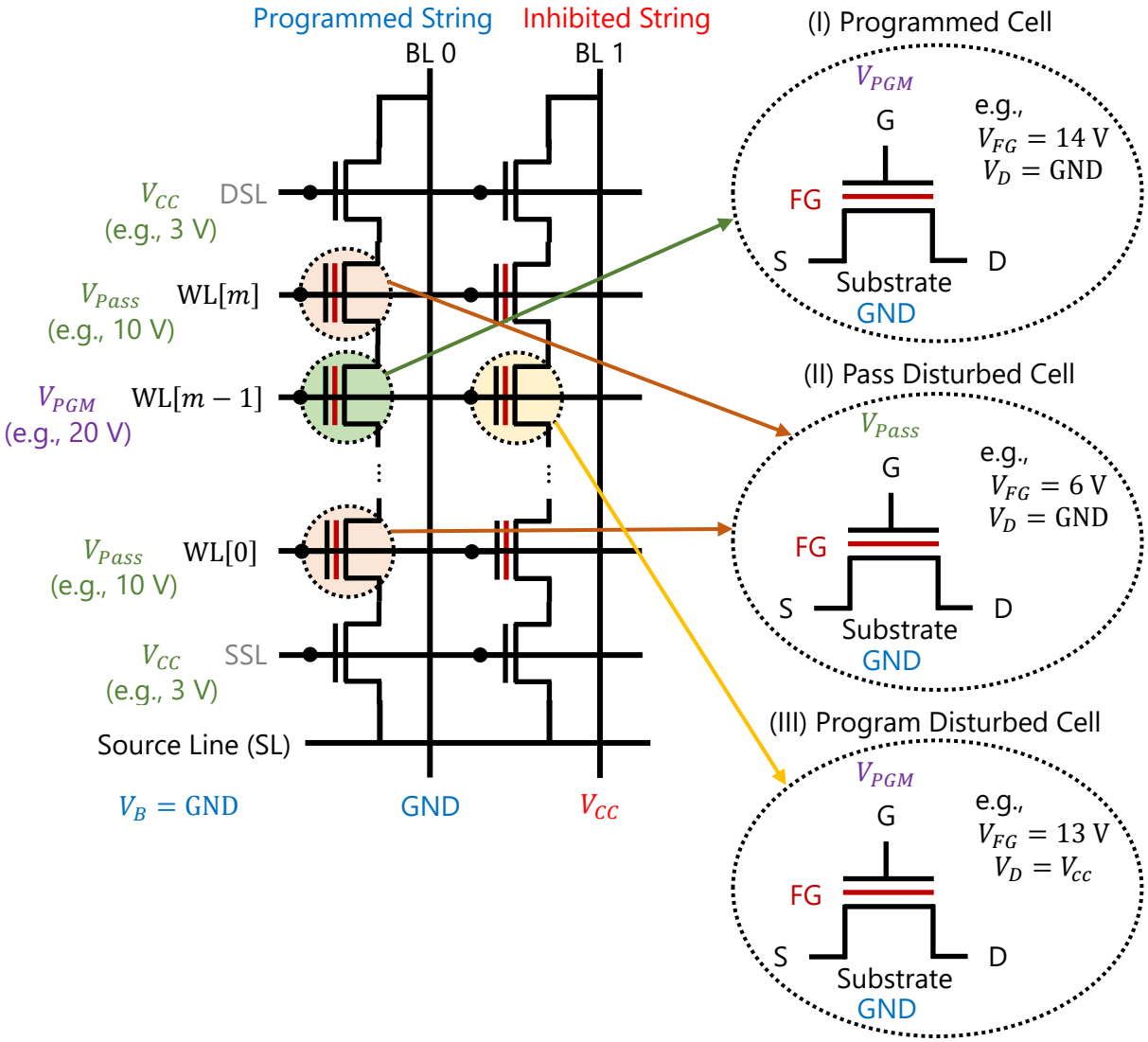


Figure 2.6 Voltage conditions during program operation in the NAND array. The memory cell at the crossing point of WL[$m-1$] and BL 0 is programmed; several other cells are disturbed by either pass disturb or program disturb.

Figure 2.6 shows the voltage condition in the NAND array when the FG cell at WL[$m-1$] and BL 0 is programmed. For this purpose, a program pulse with the pulse amplitude of V_{PGM} (e.g., 20 V) is applied to WL[$m-1$]. To conduct a successful program, it is also required to transfer 0 V (GND) to the channel region of the programmed cell as shown *Figure 2.6*(I). Consequently, the 0 V potential is applied to BL 0 and then needs to be transferred to the whole string including the programmed cell at WL[$m-1$]. This is done by applying the pass voltage V_{Pass} (e.g., 10 V) to all other word lines.

In principle, all cells addressed by the same word line could be programmed by this means at the same time. However, the programming of arbitrary information requires that (todo)

In NAND Flash memories, a physical page is the smallest addressable unit for reading and writing; a physical block is the smallest erasable unit.

Each page is made up by main area (data) and spare area as shown in . Main area can be 4, 8 or 16 KiB. Spare area can be used for ECC and is in the order of hundreds of Bytes every 4 KiB of main area.

Within each plane, flash cells are organized as multiple 2D arrays known as flash blocks, each of which contains multiple pages of data, where a page is the granularity at which the host reads and writes data.

Program of NAND flash cells:
During the program operation, the cells share the high programming voltage on the selected wordline but the program operation has to be bit selective. Therefore, a high channel potential is needed to reduce the voltage drop across the tunneling dielectric and prevents the electrons tunneling from the channel to the floating gate as indicated.

Note programming a page cannot change '0' cells to '1' cells. Thus, NAND flash cells have erase-before-write property.

Erase of NAND flash cells:
The erase operation resets the information of all the cells belonging to one block simultaneously.

Because the granularity of erase operation is usually larger the program operation (block > page), the in-place write on a page is very inefficient. As a result, SSD does out-of-place write and conducts garbage collection to recover free blocks.

NAND flash memory errors
NAND flash memory errors can be induced by a variety of sources, including flash cell wearout, disturb effects (errors introduced during programming, interference from operations performed on adjacent cells), and data retention issues due to charge leakage. (three major sources)

Disturb effects alter the memory transistors threshold voltage unintentionally during memory access operations under the influence of specific disturb conditions. Since it is essential for an efficient area consumption to arrange the storage transistors in a contact saving way, the sharing of voltage nodes can not be avoided. During the read and program operations, positive voltages are applied to the gate nodes of the memory transistors, wherein the channel is at a lower potential or even grounded. Therefore this condition is called gate disturb (also wordline disturb) and it is the most common disturb mechanism in NAND Flash memory arrays.

(Program disturb and read disturb)
Read disturbs are the most frequent source of disturbs in NAND architectures. This kind of disturb may occur when reading many times the same cell without any erase operation. All the cells belonging to the same string of the cell to be read must be driven in a ON state, independently of their stored charge. The relatively high V_{pass} bias applied on the control gate and the sequence of V_{pass} pulses applied during successive read operation may trigger the Stress Induced Leakage Current (SILC) effects in some cells that, therefore, may gain charge. Note read disturbs do not provoke permanent oxide damages: if erased and then reprogrammed, the correct charge content will be present within the floating gate.

Pass disturb is similar to the read disturbs and affects cells belonging to the same string of a cell to be programmed.

The Program disturbs, on the contrary, affect cells that are not to be programmed (inhibit) and belong to the same wordline of those that are to be programmed. In that case the program disturb is strongly related to the voltages and pulse sequences used for the self-boosting techniques. Although the program inhibit boosts the channel potential, soft programming can not be avoided especially when a high number of program pulses are applied.

The criticality of an effective program operation limiting program disturbs and/or possible successive errors is attested by the fact that in NAND memories the program operation should follow a precise and well defined "hierarchy": it is necessary to start from the cell nearest to the source selector and proceed along the string up to the cell nearest to the drain selector. This procedure is important, because the threshold voltage of a cell depends on the state of the cells placed between the considered cell and the source contact (the background pattern dependency phenomenon); the series resistance of the cells is different if they are programmed or erased.

(further) When manufacturing process scales down to a smaller technology node (i.e., the size of each flash memory cell), the amount of charge that can be trapped within the floating gate also decreases, which exacerbates reliability issues.

Multi-bit per cell storage (mutil-level):

The flash memory cell can encode one or more bits of digital data, which is represented by the level of charge stored inside the transistor's floating gate. Earlier NAND flash chips stored a single bit of data in each cell (i.e., a single floating-gate transistor), which was referred to as single-level cell (SLC) NAND flash. Multi-level cell (MLC) NAND flash stores 2-bit value (00, 01, 10, and 11). Triple-level cell (TLC) flash stores 3-bit value. Quadruple-level cell (QLC) flash stores 4-bit value. The benefits of multi-bit per cell storage is the additional capacity of the SSD without increasing the chip size, while it also decreases reliability by making the cells more difficult to correctly store and read the bits.

NAND Flash Memory Organization:
The flash memory is spread across multiple flash chips (typical values: 4, 16 chips), where each chip contains one or more flash *dies*, which are individual pieces of silicon wafer that are connected together to the pins of the chip. Each chip is connected to one or more physical memory channels, and these memory channels are not shared across chips. A flash die operates independently of other flash dies, and contains between one and four *planes*. Each plane contains hundreds to thousands of flash *blocks*. Each block is a 2D array that contains hundreds of rows of flash cells (typically 256-1024 rows) where the rows store contiguous pieces of data.

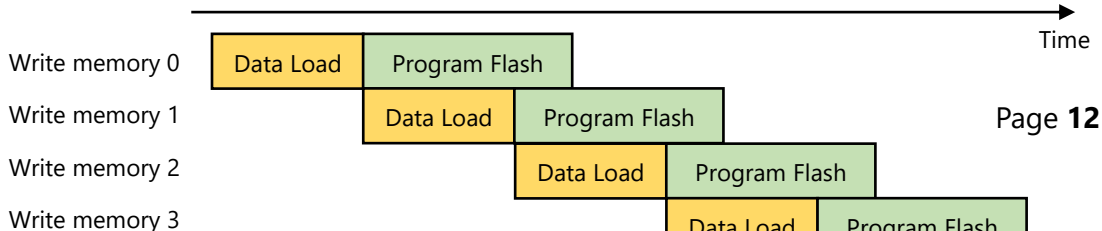
In NAND Flash memories, a logical page is the smallest addressable unit for reading and writing; a logical block is the smallest erasable unit.

If we consider the SLC case with interleaved architecture, even cells belong to the even page (BLe), while odd pages belong to the odd page (BLo). For example, a SLC device with 4 KiB page has a WL of 32,768 + 32,768 = 65,536 cells. In the MLC case we have MSB and LSB pages on even BL, and MSB and LSB pages on odd BL.

Each page is made up by main area (data) and spare area as shown in Fig. 1.5. Main area can be 4, 8 or 16 KiB. Spare area can be used for ECC (Error Correction Code) and is in the order of hundred of Bytes every 4 KiB of main area.

The **planes** can execute flash operations in parallel, but the planes within a die share a single set of data and control buses. Hence, an operation can be started in a different plane in the same die in a pipelined manner, every cycle.

Channel is the data bus (typical width: 8-bit) for connect different memories to the SSD controller (or NAND controller inside). Operations on a channel can be interleaved, which means that a second chip can be addressed while the first one is still busy. For instance, a sequence of multiple write operations can be directed to a channel, addressing different NANDs, as shown in Fig. 1.13: in this way, the channel utilization is maximized by pipelining the data load phase; in fact, while the program operation takes place within a memory chip, the corresponding Flash channel is free.



Data in a block is written at the unit of a *page*, which is typically between 8 and 16 KiB in size in NAND flash memory. All read and write operations are performed at the granularity of a page. Each block typically contains hundreds of pages.

Flash cards, USB keys and Solid State Drives are definitely the most known examples of electronic systems based on non-volatile memories.

Chapter 3 Integrated Circuit Architecture

SSDs are the prevalent application for NAND. An SSD is a complete, small system where every component is soldered on a PCB and is independently packaged.

The basic structure of a solid-state drive is shown in Figure 3.*1*. In addition to Flash memories and an SSD controller (a microcontroller), there are usually other components. For instance, an external DC-DC converter can be added in order to derive the internal power supply, or a quartz can be used for a better clock precision. Of course, reasonable filter capacitors are inserted for stabilizing the power supply. It is also very common to have a temperature sensor for power management reasons. For data caching, a fast DDR memory is frequently added to the board: during a write access, the cache is used for storing data before transfer to the Flash.

NANDs are usually available both in TSOP (Thin small outline package) and BGA (Ball grid array) packages. In order to improve performances, NANDs are organized in different Flash channels

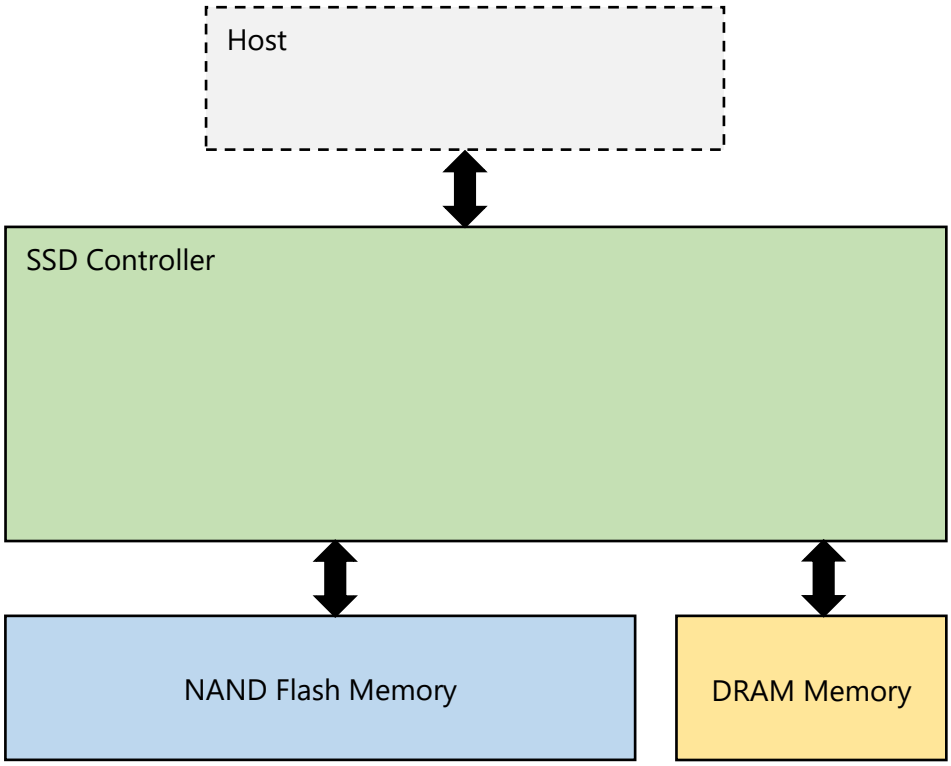


Figure 3.1 Hardware View of SSD system

For many applications the host interface to SSDs remains a bottleneck to performance. PCI Express (PCIe)-based SSDs together with flash-optimized host control interface standards address this interface bottleneck. SSDs with legacy storage interfaces are proving useful, and PCIe SSDs will

further increase performance and improve responsiveness by connecting directly to the host processor.

The SSD controller is responsible for scheduling the distributed accesses at the memory channels. And it uses dedicated engines (i.e., NAND controller) for the low-level communication protocol with the Flash.
SSD Controller: The SSD controller is responsible for (1) handling I/O requests received from the host, (2) ensuring data integrity and efficient storage, and (3) managing the underlying NAND flash memory.

Charge pumps are used to generate all the needed voltages within the chip

In multilevel storage, cell's gate biasing voltages need to be very accurate and voltage regulators become a must.

The Row Decoder is the block in charge of addressing and biasing each single wordline and it is located between the planes. Bit lines are connected to a sensing circuit. The purpose of sense amplifiers is to read the analog information stored in the memory cell.
The Row Decoder, also called Wordline Decoder or Wordline Driver.

Especially, SSDs call for a higher read and write throughputs; in other words, SSDs need to manage more NAND dies in parallel. Basically, there are a couple of options:
- The first one is to increase the number of dies per channel;
- The second option is to increase the number of channels.

Flash chip controllers (FCCs): A Flash chip controller is assigned to a flash memory channel for data and control connection.

DRAM: The on-board DRAM memory stores various controller metadata (e.g., how host memory addresses map to physical SSD addresses) and to cache relevant (e.g., frequently accessed) SSD pages.

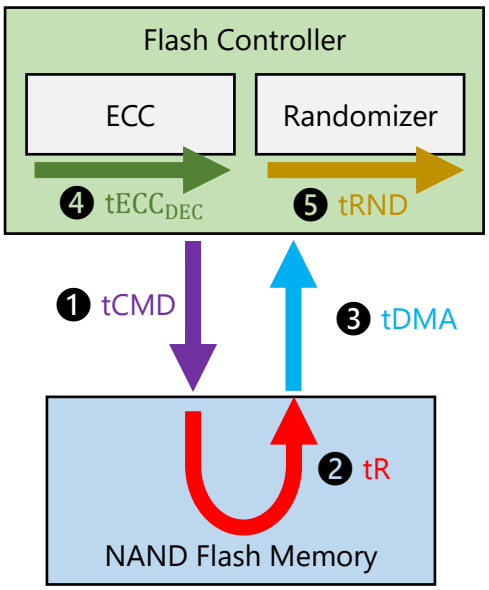
SSD Performance:
In the evaluation of SSD performance, there are three metrics and they are: (1) latency (or response time), (2) bandwidth, (3) throughput.

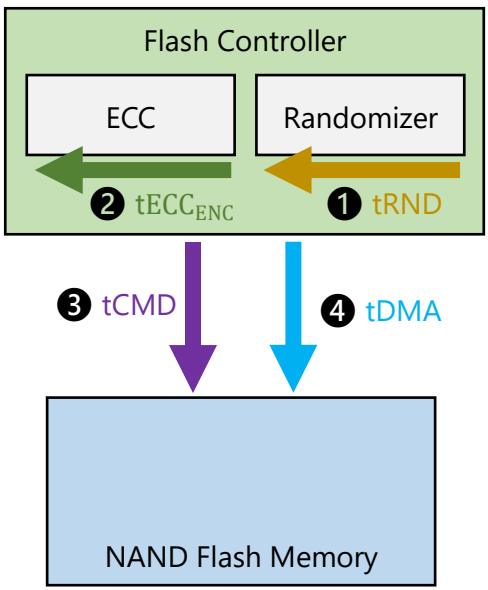
First two metrics are similar as mentioned before, latency is the time delay until the request is returned and bandwidth is the amount of data that can be accessed per unit time. Typical values are,
Latency:
Average read latency (4 KiB): 67 us

Average write latency (4 KiB): 47 us

Bandwidth:
Sequential read bandwidth: up to 3,500 MB/s
Sequential write bandwidth: up to 3,000 MB/s.

Throughput is the number of requests that can be serviced per unit time. SSDs define the measurement of the throughput as the IOPS: Input/output Operations Per Second. Typical values are,
Random read throughput: up to 500K IOPS
Random write throughput: up to 480K IOPS





SSD Firmware

Wear Leveling (endurance):
(insight) Usually, not all the information stored within the same memory location change with the same frequency: some data are often updated while others remain always the same for a very long time in the extreme case, for the whole life of the device.

(goal) In order to mitigate disturbs, it is important to keep the aging of each page/block as minimum and as uniform as possible: that is, the number of both read and program cycles applied to each page must be monitored.

(endurance definition) the maximum number of allowed program/erase cycles for a block (i.e. its endurance)

The controller firmware groups blocks with the same ID number across multiple chips and planes together into a *superblock*. Within each superblock, the pages with the same page number are considered a *superpage*. The controller opens one superblock (i.e., an empty superblock is selected for write operations) at a time, and typically writes data to the NAND flash memory one superpage at a time to improve sequential read/write performance and make error correction efficient, since some parity information is kept at superpage granularity. Having the ability to write to all of the pages in a superpage simultaneously, the SSD can fully exploit the internal parallelism offered by multiple planes/chips, which in turn maximizes write throughput.