Chapter 1

Introduction

1.1 General Description for Flash Memory

Flash memory is a form of EEPROM (***Electrically-Erasable Programmable Read-Only Memory***) [1] that allows multiple memory locations to be erased or written in one programming operation. In layman's terms, it is a form of rewritable memory chip that, unlike a Random Access Memory chip, holds its content without the need of a power supply. It is also an example of an NVRWM (***Non-Volatile Read Write Memory)***[1]. The memory is commonly used in memory cards, USB (***Universal Serial Bus***) [1] flash drivers, MP3 (***MPEG-1 Audio Layer 3***) players, digital cameras and mobile phones.

Flash memory is non-volatile, which means that it does not need power to maintain the information stored in the chip. In addition, flash memory offers fast read access times (though not as fast as volatile DRAM (***Dynamic Randon Access Momory***)[2]memory used for main memory in PCs) and better shock resistance than hard disks. These characteristics explain the popularity of flash memory for applications such as storage on battery-powered devices.

Though flash memory was originally used inside computers, it has invaded many

other areas outside the box. Flash memory cards used for digital cameras, cellular phones, networking hardware, and PC cards. Though the memory's read/write speed is not lightning fast, it is nice to be able to tote around a little card rather than a cumbersome hard drive.

In NOR flash, each cell looks similar to a standard MOSFET transistor, except that it has two gates instead of just one. One gate is the CG (***control gate)*** [2]like in other MOS transistors, but the second is an FG (***floating gate)*** [2] that is insulated all around by an oxide layer. The FG is between the CG and the substrate. Because the FG is isolated by its insulating oxide layer, any electrons placed on it get trapped there and thus store the information. When electrons are on the FG, they modify (partially cancel out) the electric field coming from the CG, which modifies the Vt (***threshold voltage***)[2] of the cell. Thus, when the cell is "read" by placing a specific voltage on the CG, electrical current will either flow or not flow, depending on the Vt of the cell, which is controlled by the number of electrons on the FG. This presence or absence of current is sensed and translated into 1's and 0's, reproducing the stored data. In a multi-level cell device, which stores more than 1 bit of information per cell, the amount of current flow will be sensed, rather than simply detecting presence or absence of current, in order to determine the number of electrons stored on the FG.

A NOR flash cell is programmed [2] (set to a specified data value) by starting up electrons flowing from the source to the drain, then a large voltage placed on the CG provides a strong enough electric field to suck them up onto the FG, a process called hot-electron injection. To erase (reset to all 1's, in preparation for reprogramming) a NOR flash cell, a large voltage differential is placed between the CG and source, which pulls the electrons off through quantum tunneling. In single-voltage devices (virtually all chips available today), this high voltage is generated by an on-chip charge pump. Most modern NOR flash memory components are divided into erase segments, usually called either

blocks or sectors. All of the memory cells in a block must be erased at the same time. NOR programming, however, can generally be performed one byte or word at a time.

One limitation of flash memory is that although it can be read or programmed a byte or a word at a time in a random access fashion, it must be erased a "block" at a time. Starting with a freshly erased block, any byte within that block can be programmed. However, once a byte has been programmed, it cannot be changed again until the entire block is erased. In other words, flash memory (specifically NOR flash) offers random-access read and programming operations, but cannot offer random-access rewrite or erase operations. When compared to a hard disk drive, a further limitation is the fact that flash memory has a finite number of erase-write cycles (most commercially available EEPROM products are guaranteed to withstand 10^6 programming cycles,) so that care has to be taken when moving hard-drive based applications, such as operating systems, to flash-memory based devices such as CompactFlash. This effect is partially offset by some chip firmware or filesystem drivers by counting the writes and dynamically remapping the blocks in order to spread the write operations between the sectors, or by write verification and remapping to spare sectors in case of write failure.

1.2 Flash Memory Application

Because of the combination of non-volatility, in-system re-writability, and high density, Flash Memory has already found in many applications that can broadly be classified as:

1.2.1 Communication Applications

Generally, Flash memory in communication applications is used to store important routines. Not like other applications for Flash memory, the communication gears need to

update the Flash Memory frequently for protocol development and security requirements, as well as operating software updates and upgrades. Based on the requirements (frequent updates and upgrades), the communication equipments will increasingly use Flash Memory instead of mask ROM (***Read Only Memory***) [2] to store application software.

The major applications in communication market could be listed as the followings: Cell phone, Pagers, Modems, …… The major function for Flash Memory in the communication applications is to store the program routines, parameters, phone book, etc.

It's worth to mention that there exists complex function in cell phone, such as blue tooth, digital camera function, E-mail…. So the Flash memory is used widely in communication segment and growing quickly.

1.2.2 Consumer Market

The consumer market represents a good size of total electronics market. The followings are common heard in our daily life:

a. MP3: Most of the Flash in an MP3 is used to store high quality voice, mainly in form of removable memory cards. And the share of Flash is growing constantly. The other propose of Flash is used to store program routines and security parameters.

b. STB (***Set Top Boxes***): The main function of STB is to upgrade analog TV sets with digital functionality. STBs use Flash Memory as CPU work memory to store program routines and security keys. Many STBs have the capability to store the broadcast content to be replayed later.

c. Cameras: Digital cameras use Flash memory for the two main purposes: 1) to store the pictures. 2) To store the program routines and parameters

1.2.3 Smart Cards

Smart cards consist of re-writable, nonvolatile memory and a microprocessor. The microprocessor makes the card “smart”. The microprocessor controls the communication between the system and the memory.

Currently, smart cards are widely used in many applications, suck like: SIM card used in mobile phone, pay phones, employee ID cards, authentication, and storage of personal information. The potential applications for smart cards could be extended to supermarket card, credit cards, driver’s license, access control device (E-key), and car maintenance history. The memory in smart card, Flash memory is widely used instead of EEPROM or Mask ROM for manufacture cost.[3]

1.2.4 Automotive Applications

In-system re-programmability makes Flash memory well suited to many automotive applications. Such like:

1) In-system code change for development and optimization without hardware changes.
2) In-vehicle selections of features- engine, transmission, ride options.
3) Continues system optimization through the vehicle life cycle by real time system programmability; in-field repair/alteration maintainability and recall avoidance by in-vehicle alteration at the dealer. [3]

1.3 NAND or NOR Flash

Consumer electronics and embedded software devices are using larger amounts of flash memory for nonvolatile storage than ever before. One important decision in such devices is what kind of flash memory to use: NAND or NOR?

NOR flash memory has traditionally been used to store relatively small amounts of

executable code for embedded computing devices such as PDAs and cell phones. NOR is well suited to use for code storage because of its reliability, fast read operations, and random access capabilities. Because code can be directly executed in place, NOR is ideal for storing firmware, boot code, operating systems, and other data that changes infrequently.

NAND flash memory has become the preferred format for storing larger quantities of data on devices such as USB Flash drives, digital cameras and MP3 players. Higher density, lower cost, and faster write and erase times, and a longer re-write life expectancy make NAND especially well suited for consumer media applications in which large files of sequential data need to be loaded into memory quickly and replaced with new files repeatedly.

The choice between using NAND and NOR Flash may not be a simple one for the complex embedded devices being developed today. While ever-larger media files are driving increased demand for inexpensive NAND, powerful new operating systems and intricate applications running on fast processors call for the kind of fast-executing code NOR can support. An important example is a smart phone or PDA that combines a tremendous need for storage with a demanding set of application performance requirements. In some cases an optimal design might call for both types of flash memory in the same device.

Whichever type of flash is used in a device, there are certain negative performance characteristics that need to be mitigated. NOR is fast to read current data but markedly slower to erase it and write new data. NAND is fast to erase and write, but slow to read non-sequential data through its serial interface. NAND is also prone to single-bit errors, requiring rigorous algorithms for error detection and correction.

Well-designed software strategies can be very effective in increasing the performance and reliability of Flash hardware. The goals of flash memory management

software include:

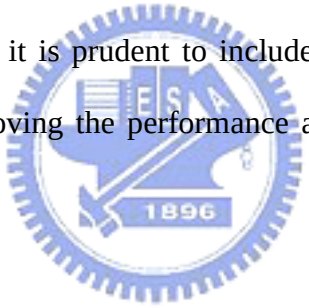
Avoid loss of data -- Perhaps the most important goal in managing flash memory is to assure that no data is ever lost as a result of an interrupted operation or the failure of a memory block. There are several ways that flash management software can achieve this goal. Rewrite operations, for example, can be managed in such a way that new data is written and verified before the old data is deleted, so that no power loss or other interruption can result in the loss of both old and new data. Bad block management is another important safeguard to prevent data being written to memory blocks that have failed. Software can check for bad blocks shipped from the factory, as is typical with NAND, and avoid writing to those blocks from the beginning. When blocks go bad over time they can be identified and managed so that they are no longer used. Finally, as the end of media life nears, good memory management software can implement a graceful strategy such as placing the entire flash unit in a read-only state, thereby avoiding data loss when the number of block errors exceeds a predefined number.

Improve effective performance -- Two ways media management software can improve performance are background compaction and multithreading. Compaction reclaims space by identifying blocks that have obsolete data that can be erased, copying any valid data to a new location, then erasing the blocks to make them available for reuse. Such compaction increases the amount of usable space on the media and improves write performance. Compaction may also help to defragment noncontiguous data for improved performance on read operations. The space recovery is particularly valuable for the more costly NOR memory and the defragmentation benefits the slower-reading NAND. Compaction is best performed in the background during idle time, however, or it can interfere with critical operations and degrade performance. This is where a multithreading system becomes important. By allowing high-priority read requests to interrupt low-priority maintenance operations, a multithreading system can reduce read latency by

orders of magnitude compared to a single-thread solution.

Maximize media lifespan -- When some blocks of memory contain fixed content, such as binary code, the remaining blocks will experience increased demand for erase and write operations, leading to earlier failure. Wear-leveling algorithms can prevent overuse of memory blocks and prevent a "stalemate" scenario in which a small region of memory becomes locked in a pattern of repeated writing and compaction. Wear leveling software can monitor block usage to identify high-use areas and low-use areas containing static data, then swap the static data into the high use areas. It can also balance write operations across all available blocks by choosing the optimal location for each write operation.

The decision between NAND and NOR memory will ultimately depend on both technical and pricing requirements of the device being built. Whatever type or combination of flash is used, it is prudent to include memory management software to prevent data loss while improving the performance and maximizing the lifespan of the memory.[1]



1.4 The Trend of Future Application for Flash Memory

Flash Memory is expected to be the optimum choice for mass storage in personal mobile systems. In mobile and handheld computers, Flash memory will replace DRAM in order to contain operating systems, BIOS, and applications software.

At the present time, Flash Memory has already displaced EPROM in many applications, and Flash Memory will be the technology of choice in many portable applications in the future. The Flash Memory will replace hard disks in low-end applications. [2][3]

The followings are the Flash market development trends:

1) PC BIOS

2) PC operating system and application storage

3) Portable computer main memory and mass storage

4) Dominance in mobile disk drive applications

5) Faster read/write, lower power dissipation

6) Application specific Flash

7) Commodity memory

1.5 Low power consumption approach [4]

Low power dissipation and low power supply NVM is getting important for potable application. Novel sensing scheme is proposed in this paper. The main advantages compared to other approached are as the followings:

1) Make the chip enters standby mode after sensing operation is completed to save power and suite for low frequency application.

2) Better flash cell grounding by pulling the sensed bit line to ground for this memory architecture to improve access speed and reliability.

3) The proposed sensing scheme could operate under very low power supply (less than 1.0v) if bit line pre-charge be well taken care.

4) Bit line is pre-charge to the trip point of sensing device to save power and speed up access “1”.

5) Each sense amplifier generates the control signals, like pre-chare pulse width and pull up/ pull down signals, automatically

The following 2 tables are the comparison between the proposed sense amplifier and conventional design and prior art for low power supply sense amplifier. Based on the comparison, the power consumption is improved much, and the access speed is not degraded.

Table 1-1: Comparison between the proposed sense amplifier and conventional design

Architecture	Active current	Static current	Access speed
Conventional SA[5]	6.5mA	6 mA	40ns
Proposed SA	0.12mA	0 uA	40ns

Table 1-2: Comparison between the proposed sense amplifier and prior art

VDD/temperature	Prior art [6]	Proposed SA (5 corners*)
2.0v/ -40C	15ns	13ns (13 ~ 19ns)
1.6v/25C	25ns	20ns (16 ~ 36ns)
1.0v/25C	27ns	29ns (22 ~ 37ns)

Chapter 2

Design Concept and Design Approach

2.1 Design Background

Due to process technology continues scale down, and application requirement, lower power supply and lower power dissipation are getting important in current/future application. If Flash memory has the features that can work under lower power supply and consume smaller power without performance degradation, it's a great benefit to extend the life time of battery for consumer application, such like MP3, mobile phone, pagers, digital camera and so on. [2]

As for power dissipation in Flash access, there are three parts contribute the power consumption, one is address buffer transition, one is output buffer transition, the other one comes from sense amplifier operating current. In this work, it provides a novel sensing scheme to achieve low power dissipation in data sensing and keep the same access performance. And providing a low power supply solution while power supply is as low as 1v, the sense amplifier can still work less than 20M Hz (access speed is less than 50ns) within whole process corner (SS, TT, FF, SF, FS. —N/P MOS process corner) and the temperature ranges from –45^{o}C to 125^{o}C.

2.2 Other Approach

Table 2.1 refers from [6], which is published on IEEE 2005. It shows the Vdd minimum could be as low as 1v (based on simulation results). But the power consumption seems to be high because many DC current paths exist for bit line bias setting and data sensing. The following table is the simulation result.

Table 2-1 Simulation result of prior art [6]

Vdd	Temperature	Access time
1.65v	27°C.	25 ns
1.35v	125°C.	43 ns
1.95v	-40°C.	15 ns
1v	27°C.	38ns

The parameters in the table are based on simulation results. And the minimum power source is 1.0V. Process condition for simulation model is not mentioned

2.3 Reading Operation of Flash Memory:

The reading operation of an EEPROM or Flash memory is performed by sensing the flash cell's current from the selected flash cell, and compared with a reference current. The reference current must be well defined for flash access performance concern. If the read sensing ratio (the ratio of reference divided by flash cell current, SR) is too high or too low the access speed will be impact. So the SR should be defined after deep considerations. The SR (sensing ratio) in this test chip is around 33%. In addition to "access speed" consideration, endurance reliability and data retention performance must also be taken into considerations in Flash memory. Because the flash memory endurance and data retention are the main features in Flash Memory.

The flash cell could be programmed and erased to generate two different Vt (***threshold voltages***) to store 2 different states of data. In this paper, the flash cell is high Vt (low cell current, the cell data is defined as "0" after sensing) after program operation and low Vt (high cell current, the cell data is defined as "1" after sensing) after erase operation. The high voltage used for program/erase operation is generated by built-in

charge pump.

For programmed state, the flash cell is higher threshold voltage and lower cell current, the cell current is close to 0uA and much less than the reference current. The data out (SO) after sensing is “0”. For erased state cell, the flash cell is lower threshold voltage and higher cell current, the cell current is close to 20uA (one shut boost scheme is implement to higher cell current to improve access performance, product yield and reliability). The sensing ratio is selected carefully to achieve best access performance and product yield and reliability; basically, the SR (***sensing ratio***) is defined based on the process consideration (or product characterization). As the previous description, the reference is around 7uA under the sensing ratio is 33%, so the current difference between reference current and cell current is around 13uA for read “1” operation.

2.4 New Design Approach [4]

In flash cell data sensing, we can separate the sensing operation into 3 stages. They are: 1) Selected bit line pre-charge. 2) Data sensing. 3) Data ready. (Sensing operation is completed).

2.4.1 Bit Line Pre-charge Scheme

As for bit line pre-charge stage, the scheme of conventional design is to pre-charge the bit line bias to a specific value that is defined by bit line clamp circuitry. The sensing scheme costs more power dissipation due to the bit line clamp circuitry that results from that there are some DC current paths in the circuitry to generate the bias. So, for power dissipation requirement, the conventional design approach is not the right solution for low power application. Due to the power dissipated more in bit line pre-charge stage in conventional design, so the improvement is significant if the better bit line pre-charge scheme is used and without access speed degradation.

There are two new ideas in the pre-charge scheme of the proposed sensing amplifier: 1) Pre-charge level: the bit line is pre-charged from VSS to the trip point of the sensing device in the sense amplifier through pre-charge device. If the selected bit line is pre-charged to the trip point of the sensing device, the bit line pre-charge path is turned off automatically after the logic operation of the sense amplifier control logic. 2) The bit line pre-charge control logic: The pre-charge timing pulse is self-defined (each selected bit line) by the signal control logic in each sense amplifier. And the current path for bit line pre-charging will be turned off automatically without extra control signals. By using this approach, each sense amplifier is individual controlled to achieve the low power consumption target. The most use for the definition of pre-charge pulse is by a global signal for all sense amplifiers. And a global signal to turn off pre-charge path will result in under or over bit line pre-charge and impact the performance of sense amplifier.

So, the power dissipation in bit line pre-charge phase is reduced much compared to conventional design. The major design concept in bit line pre-charge phase is: Selected (decoded) bit line is just pre-charged from ground level to the trip point of the sensing device in the sense amplifier. And each sense amplifier has its own control logic. The on or off of the pre-charge path is controlled by each sense amplifier. Because if the bit line over pre-charge phenomenon happened, will result in more power dissipation and slow the data sensing. This is because that: [7]

$$P = I_{pre\_ch} \times \delta V \tag{2.1}$$

Bit line over pre-charge means more δV , so more power is consumed.

So, we can conclude that the features for the bit line pre-charge scheme are: 1) Bit line level is just pre-charged to the trip point of the sensing device to save power dissipation. 2) No need to generate a global signal for bit line pre-charge pulse to prevent the bit line over pre-charge or under pre-charge to achieve better access performance.

By implement the new bit line pre-charge scheme, we can know that the power dissipation is been minimized. Compared to conventional design, the DC path for bit line bias setting is existed in the whole sensing cycle. And the DC current is the major portion in the sense amplifier. Because there is no extra power consumed except the decoded bit line is pre-charge. And the power consumption is just in the period.

2.4.2 Data Sensing

When bit line has been pre-charged to the trip point of the sensing device, the sense amplifier enters the data sensing stage. The sensing scheme is the comparison between flash cell's current and reference current. Not like differential amplifier and cross latch type sense amplifiers, the data out after sensing of the proposed sense amplifier is based on the bit line level is pulled up by reference current (the data out is "0") or be pulled down by flash cell current (the data out is "1"). So the sensing speed and sensing margin depends on the following equation: [7]

$$Cbl \times \delta V = \delta I \times T \qquad (2.2)$$

1) The current difference between flash cell current and the reference current in the sense amplifier. 2) The capacitive load of the selected bit line.

In 0.18um technology, the typical power supply is 1.8v for advanced application. As for most flash cell architecture, no matter stack gate or split gate, the cell current is very small under such power supply, the worst condition for power supply is around 1.6v if 10% application margin is taken into account. The cell current is around 15uA under extreme process corner and temperature condition. (SS process condition, 1.6v and high temperature) This is because that the flash cell is composed of double poly gate. One is for control gate that is connected to word line. The other one is floating gate, the floating gate is used to store electrons to have 2 different data types one is for data "1", the other one is for data"0". So the effective gate bias is not so high as a pure MOS, this results in

low cell current and lower access speed than other memory. (Such as SRAM)[8]

$$Id = k(W/L)(Vgs - Vt)^2 \equiv Ids \quad (2.3)$$

In order to higher flash cell current to improve sensing speed, word line boost scheme (boost word line bias from a power supply to a certain level) is used widely to improve the flash cell current for lower power supply application. Due to the word line is connected to the control gate of flash cell, so higher the word line level improves cell current significantly. Not only sensing speed improvement could be achieved but also flash reliability. In the proposed sense amplifier, one shut word line boost is used to the simulation and silicon. By word line boost, the control gate of flash cell could be higher than VDD. By doing so, we could improve flash cell current significantly. And sensing speed also be improved, but the drawback is that more power consumption in word line boost circuitry.

2.4.3 Data Valid

For low power consumption requirement, it helps much to turn off the chip after data sensing. In most sense amplifiers, there exist current flow paths after data is sensed out and consume power much. So for low power application, it's necessary to turn off those current flows after data sensing to save power.

In this novel sense amplifier, a new scheme for data latch is used (the method will be explained more detail in the later section). When the cell's content is sensed, the sense amplifier latches the sensed data automatically. After the flash cell's data is latched, the sense amplifier is turned off and enters into sleep mode (standby mode). When the sense amplifier latches the cell's content, only the standby current exists in the sense amplifier.

When all sense amplifiers finish the sensing operation, all circuit blocks those are related to sensing operation are turned off. By making the sense amplifiers and its

reference circuitry into sleep mode, the whole chip (all control circuitry) is in sleep mode also. As for the read operation, the power only dissipated during the period from the sense amplifier is enabled to the flash cell's content is sensed out.

As for the sense amplifier control logic, there is no extra signal needed to turn off the sense amplifier. The signal to shut down the sense amplifier is generated by each sense amplifier. That is, each sense amplifier has its own signal to turn off itself. And a global signal to turn off the reference circuitry is generated after all sense amplifiers finish the data sensing. When the reference circuitry is turned off, then the chip enters into sleep mode and without any DC path within the whole chip till next read cycle started. The total current dissipation after flash cell's data has been latched is the standby current of the whole chip.

Chapter 3

Novel Sense Amplifier Operating Introduction

3.1 Low Power Dissipation Architecture Sense Amplifier

Sensing Scheme:

Based on the previous description, the proposed sense amplifier is not like the conventional differential amplifier, the proposed sense amplifier is more close to single end amplifier. The main sensing concept is the flash cell current pulls down the bit line level but the reference current pulls up the bit line level. The output of the sense amplifier depends on the current difference between cell current and the reference current and the bit line capacitive load. The larger current difference between flash cell current and reference current the faster sensing. For flash memory, the cell current degraded much after thousands of program/erase cycles, and the cell current degradation phenomenon must be taken into consideration when designing flash memory sense amplifier to make the flash memory in good endurance performance.

After all selected bit lines are already pre-charged, word line and word line boost

circuitry are active. In the meanwhile, the reference current path is also turned on, and then begins the data sensing.

After the cell's content sensed, the control logic makes the sense amplifier enters into sleep mode, and turns off the reference circuitry. At this moment, whole chip is in standby mode and without current consumption except standby current.

3.2 Sense Amplifier Architecture

The main architecture of the proposed sense amplifier is as Figure3-1.

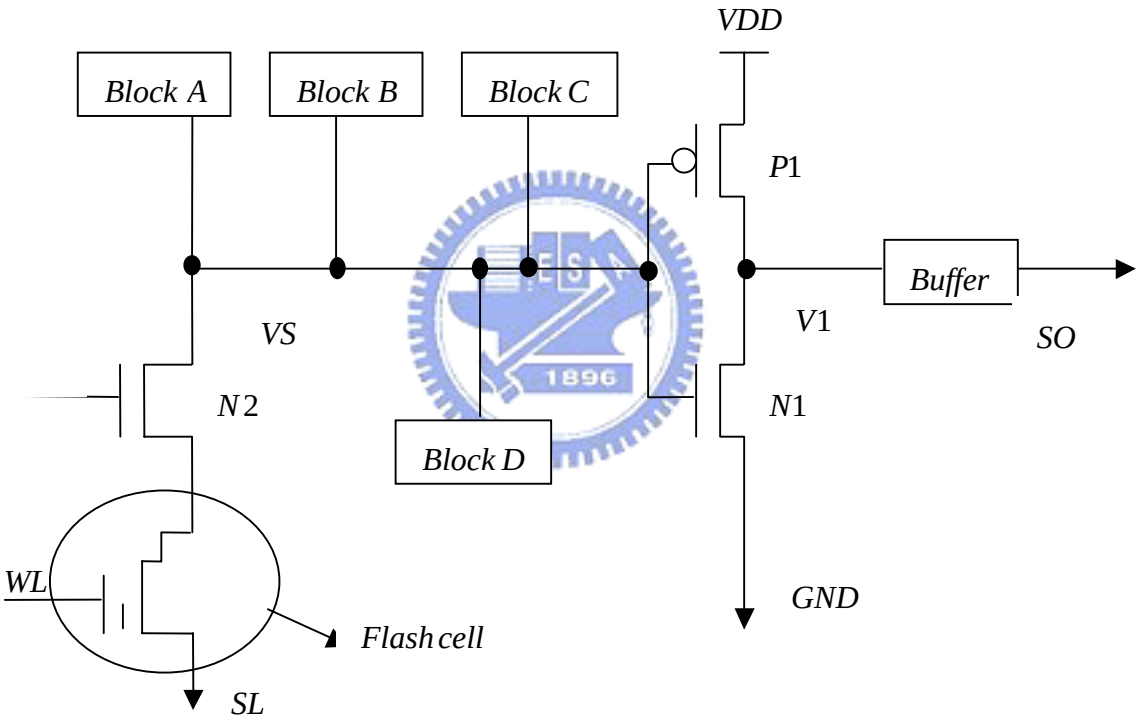


Fig 3-1 Sense Amplifier Architecture

In this sense amplifier architecture, there are 6 major portions: a) Bit line pre-charge (block A), b) Reference current path (block B), c) Buffer for the sensed data, d) Schematics for data "0" latch (block C), d) Devices for bit line pulling down (block D).

The schematic related to the sense amplifier control logic is as the following figures:

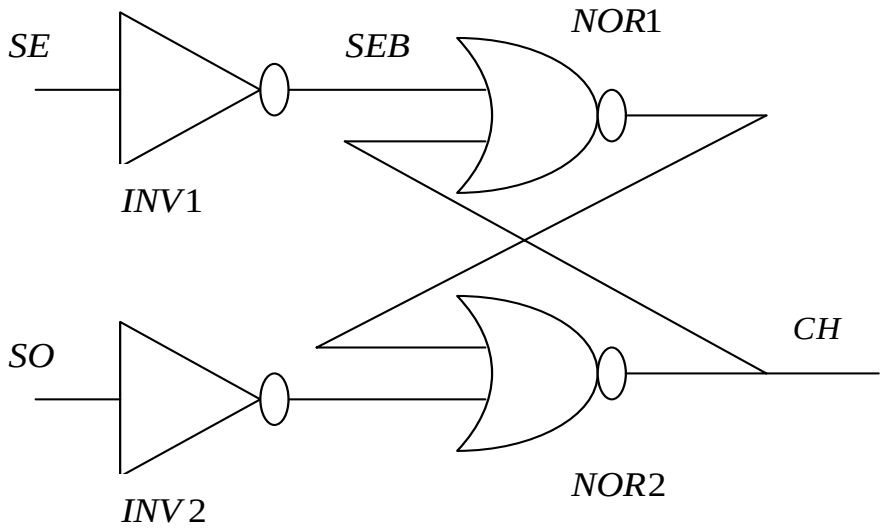


Fig 3-2 Bit line pre-charge control logic

Figure 3-2 is the bit line pre-charge control logic; it controls the bit line pre-charge timing sequence. The signal “SE” is sense amplifier enable signal; “SO” is the output of each sense amplifier; “CH” is used to control the bit line pre-charge path.

Figure3-3 is the timing waveform for the bit line pre-charge.

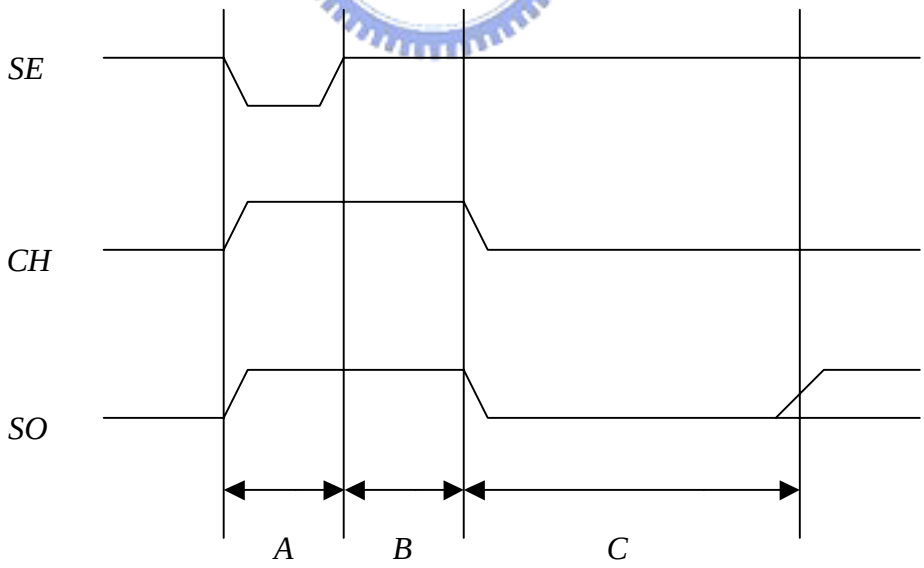


Fig 3-3 Timing waveform for bit line pre-charge logic

Figure3-3 is plotted based on the control logic as Figure3-2.

In Figure3-3, there are three portions in the waveform. Period A is for sense

amplifier initialization. In this period VS in Figure3-1 is set to GND to set “CH” to logic state “H”. Period B is for bit line pre-charge, period C is the stage for flash cell data sensing.

3.2.1Bit Line Pre-charge Path:

The bit line pre-charge path is composed of 1 PMOS (P1) and 1 NMOS (N1) in this design (as Figure 3-4).

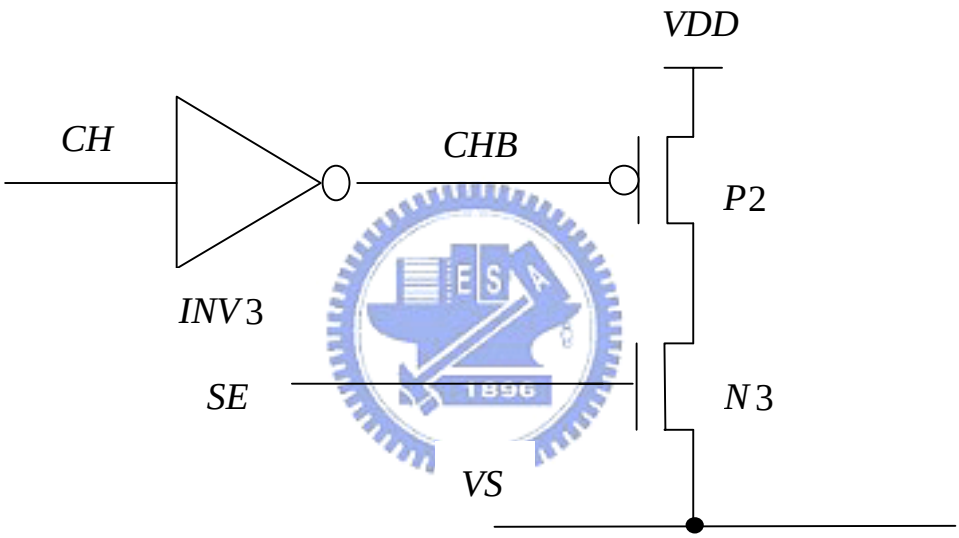


Figure 3-4 Block A: Bit line pre-charge path

The bit line pre-charge stage starts from the “SE” is active (as period B in Figure3-3), and stops after “SO” goes “L” (as Figure3-2)

Each sense amplifier controls the PMOS (P2). The bit line pre-charge path is turned off automatically by each sense amplifier when the bit line is charged to a certain level to make the output signal (SO) switched from logic state “H” to logic state “L”. Based on this pre-charge scheme, no extra signal is needed to define the bit line pre-charge period, each sense amplifier turns off its own the bit line pre-charge path. By doing so, we could make sure that each bit is just pre-charged to a target bias, and without under pre-charge

issue, but with little over pre-charge (due to the feed back speed has little timing delay for switching the bit line pre-charge path off). The advantage is important and affects the sensing function, access speed and power consumption deeply. The bit line pre-charge path is composed of 1 PMOS (P2) and 1 NMOS (N3) in this design (as Figure3-4). The PMOS is controlled by bit line pre-charge signal, CHB, the NMOS is controlled by sense amplifier active signal (SE), these two devices are serial connected, and the “source” of the NMOS is connected to the bit line. The drain of N3 and the drain of P2 are connected together. The source of the PMOS is connected to power supply, VDD. When CHB signal is at logic state “L” and the sense amplifier active signal is at logic “H”, the bit line pre-charge path is then turned on and starts the bit line pre-charge operation.

In the proposed sense amplifier, according to the bit line pre-charge control logic as Figure3-2, SR latch is used for the bit line pre-charge control logic. And the timing waveform for SE, CH and SO is shown as Figure3-3. Before sensing operation starts (triggered by SE signal) the signal SO is preset to logic state “H”, and control the bit line pre-charge signal, CHB, at “L” from SE is “L” till SO switches from “H” to “L”. The first step of sensing is bit line pre-charge, starts from “SE” active. The 2 devices in the pre-charge path are turned on at this time. The selected bit line is then pre-charge from ground level (GND) to the trip point (VS) of the sensing device. Once the bias level of data line (VS) reaches the trip point of the sensing device, the “SO” switches from logic state “H” to logic state “L”, after the logic operation of the bit line pre-charge control logic, the “CHB” switches from logic state “L” to logic state “H” and turns the bit line pre-charge path off. In this proposed sensing scheme, each sense amplifier has its own sense amplifier control logic and bit line pre-charge control logic. So the bit line pre-charge control signals are individually.

And the bit line pre-charge pulse width is related to the bit line pre-charge current and bit line capacitive load. Regarding the pre-charge current, there are some factors those

could affect the bit line pre-charge current, such like the power supply, the size of the pre-charge devices and the trip point of the sensing device. Because the bit line is connected the source node of the N3 in Figure 3-4, so higher trip point of the sensing device means higher source level of N3 and makes lower Vgs of the NMOS, results in poor driving capability during pre-charge stage.

But the trip point of the sensing device can't be too low, because the bit line is connected to the drain node the flash cell, the lower the bias, the smaller flash cell current. Slow access speed will be happened if smaller flash cell's current. So the sensing point should be defined carefully and make some trade off between pre-charge pulse width and sensing speed. As for the bit line capacitive load, it mainly comes from the flash cell junction capacitor. The more flash cells on the bit line, the heavier capacitor load, and results in longer bit line pre-charge pulse width.

The followings are the advantages of the proposed bit line pre-charge scheme: 1) No extra signal is needed to define the bit line pre-charge pulse width; the signal for tuning off the pre-charge path is generated by each sense amplifier. 2) Each sense amplifier has its own bit line pre-charge control logic. So the bit line under pre-charged phenomenon won't be happened. 3) The pre-charge path is turned off once the bit line level is high enough and each signal for pre-charge control logic is individual, so bit line over pre-charge issue could be minimized.

3.2.2Reference Current Path

As Figure 3-5, the reference current path is composed of three PMOS (P3, P4 and P5).

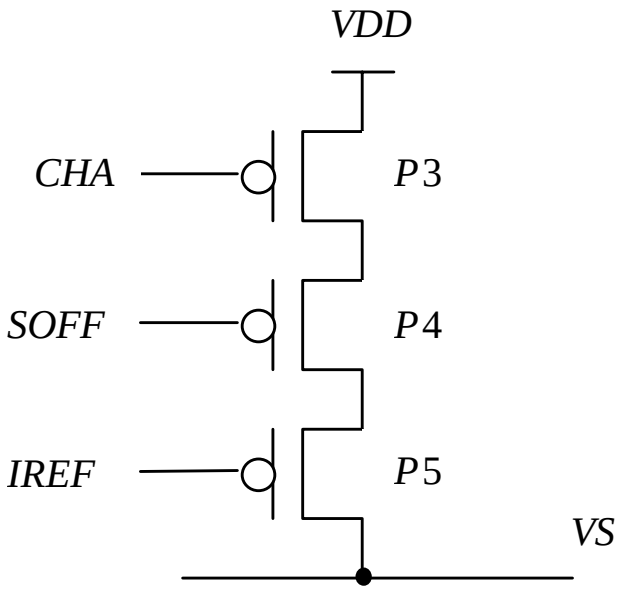
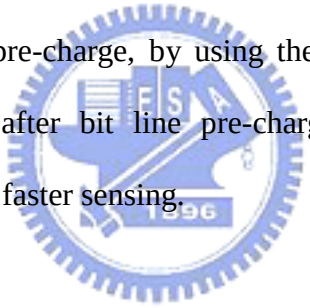


Figure 3-5: Block B, reference current path

The devices P3, P4 and P5 are serial connected, the source of P3 is connected to VDD, the drain of the P5 is connected to bit line (VS). The gate of P3 is connected to the bit line pre-charge control signal, CHA. When all the selected bit lines' pre-charge operation is finished, the CHA goes "L", and then turns the P3 on. The gate of P4 is connected by the sense amplifier disable signal "SOFF", which signal is active (disable the device) after the sensing operation is finished. The signal is generated by other logic and will be discussed later. The gate of P5 is controlled by IREF. The signal (IREF) is generated from reference cell and it's an analog bias signal. The reference current comes from the real flash cell, and we use the current mirror scheme and 33% flash cell current is chosen for sensing reference current. Each sense amplifier controls the device, P4, that is, the reference current path is turned off if each sense amplifier is finished the sensing operation.

After all selected bit lines (16 bit lines due to 16 sense amplifiers in the chip) are pre-charged, at the same time the word line boost circuitry and the reference current path are enabled. By doing so, the better sensing result will be achieved. The reasons are: 1)

Word line and word line boost circuitry enabled after bit line pre-charged is to make better bit line pre-charge efficiency. We can refer to Figure3-1, because the flash cell current sinks the bit line pre-charge current and then impact the pre-charge efficiency and consumes more current. If the word line and word line boost circuitry are enabled before bit line pre-charged, it's possible to make the pre-charge path is always on if flash current is larger than pre-charge current in the bit line pre-charge stage. It's possible to be happened because the pre-charge current is small when the bit line level is charged up, and makes lower Vgs of the N3 (as Figure3-4). When the situation happened, even the sensed data is correct, but more power is dissipated unexpected. Because current path exists from VDD to GND through flash cell. 2) Regarding the reference current path: based on the Figure3-5, we can know that the reference current pulls the bit line to a higher level, this is same effect as bit line over pre-charge, by using the novel bit line pre-charge scheme (reference current turns on after bit line pre-charged), we can minimize the over pre-charge issue and results in faster sensing.



3.2.3 Data Sensing

After selected bit lines have been pre-charged, word line boost circuitry is enabled and reference current path is turned on, then the sense amplifier enters the data sensing stage.

3.2.3.1 For Sensing Data "1"

As for the erased state flash cell data sensing, as described in previous section, the cell Vt is low; the cell current is high compared to high Vt cell. As for the Figure .3-1, the flash cell is connected to bit line through a NMOS N2, and the selected cell is turned on when word line is active and the cell current begins to sink current from bit line when word line is higher enough to turn on the flash cell and begins to pull down the bit line level. In the meantime, the reference current is to pull up the bit line level. If the cell

current (bit line pull down current) is larger than reference current (bit line pull up current) then the bit line level is getting lower. Normally, for the erased state cell, the current is larger than reference current, so the bit line level is getting lower when flash cell current is higher than reference current in sensing stage. If the bit line level is pulled down lower than the trip point of the sensing device, the bias of V1 (output node of the sensing device) is getting higher. Figure3-6 is the buffer stage for the sensing stage. The output node the sense amplifier is changed from logic state “L” to logic state “H”, this means that the “SO” is changed to “H”. At this moment, the sensing operation for erased state flash cell is finished.

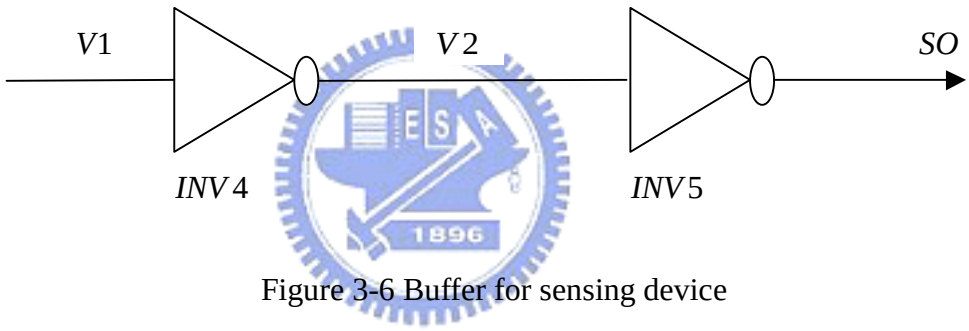


Figure 3-6 Buffer for sensing device

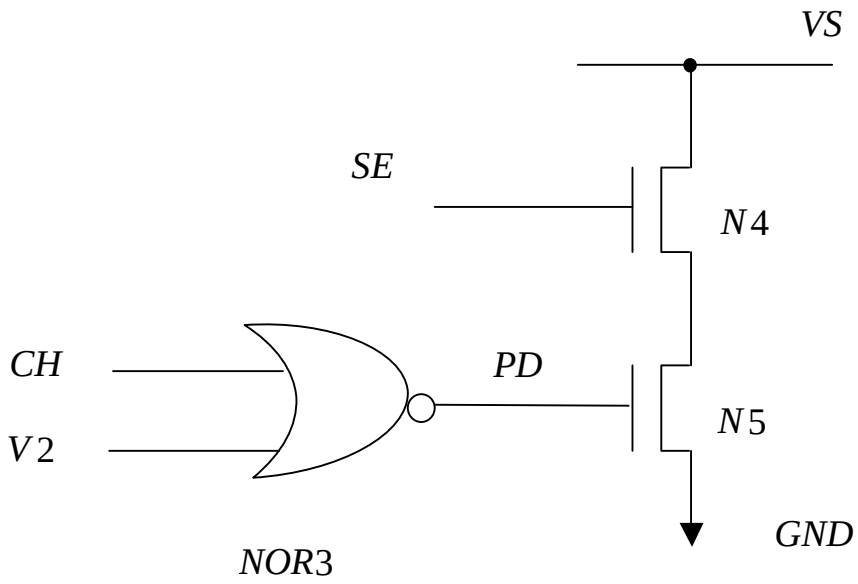


Fig 3-7 Block D: schematic for bit line pulled down

As Figure 3-7, the signal “PD” is the combination logic output of the signal “CH” and “V2” (see Figure .3-6). When the sensing operation for erased cell is completed, the logic state of “V2” is switched from “H” to “L”. This makes the “PD” switch from logic state “L” to logic state “H”. When “PD” switches from “L” to “H”, the node VS is pulled to VSS (ground level),

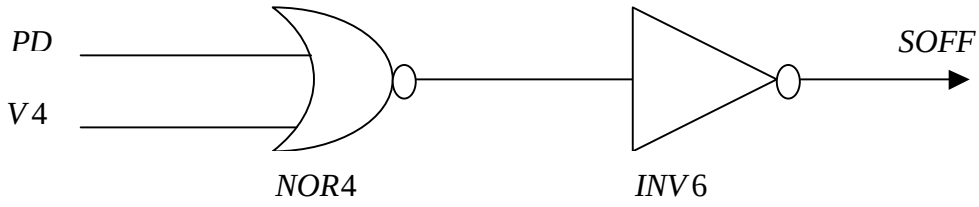


Figure 3-8: Control logic for turning off reference current

Regarding the control of reference current path, the schematic is shown as Figure3-8. When “PD” is switch from “L” to “H”, the “SOFF” switches to “H” also. Then the PMOS (P4) in Figure3-5 is turned off. At this moment, there is no current flow in the sense amplifier.

There are some benefits by pulling down the bit line level to VSS for the erased cell sensing. 1) Because the input node of the sensing device is not full swing, so there exists a DC current flow in the sensing device (P1 and N1 in Figure3-1). The signal “PD” is used to speed up the input bias of the sensing device reaches VSS and shorter the period of DC current exists in the sensing device. By doing so, make fewer power dissipation. 2) Because the bit line bias is also pulled down to ground level after the sensing operation and make the cell current without any current flow. Based on the array architecture, source line level of the flash cell will be little higher than ground level if there is current flow existed, this will make lower Vgs of the flash cell (Figure3-1) and impact flash cell current much. By pulling down the bit line bias after the sensing operation, the cell

current can be improved for those bits that sensing operation is not finished yet

From the time on, there is no current flow in the sense amplifier, and the flash cell data is latched in the sense amplifier.

Based on the sensing mechanism, the access speed for sensing data “1” from flash cell is determined by: 1) Bit line pre-charge speed, because sensing operation follows the bit line pre-charge. 2) The current difference between flash cell and reference current. So, for the access performance improvement, higher flash cell current is an approach to achieve better sensing performance.

3.2.3.2 Data “0” Sensing

As for the reading of programmed state cell, the selected flash cell is higher Vt, the decoded flash cell current is small even the word line is active. As described in previous section, the output of the sense amplifier, SO, is changed from logic state “H” to logic state “L” after the decoded bit line is pre-charged. So the state of “SO” will be kept at “L” from the sensing operation starts to the end of sensing operation.

As for managing the sense amplifier in reading “0”, we use the following schematic (Figure3-9) to handle that.

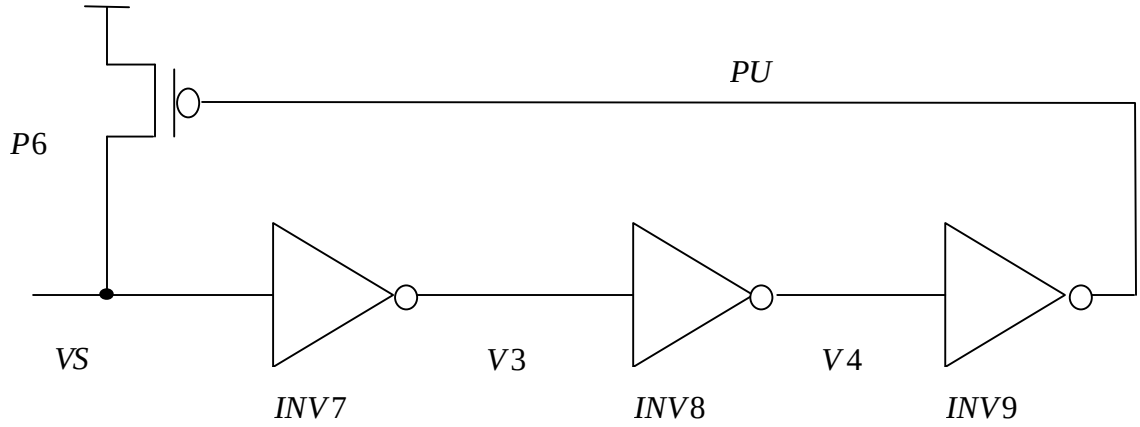


Figure 3-9 Block C: Schematics for data “0” latch

In the sensing period, the referent current exists and pulls up the bit line bias. For programmed state cell sensing, the flash cell current is lower than the reference current. According to the Figure3-1, the bit line pull up current is larger than the bit line pulled down current, this makes bit line level is getting higher. Because the reference current path is composed by PMOS, the bit line level can be as high as VDD. When the bit line level reaches the trip point of the voltage detector (the INV7 in Figure3-9), the bit line pull up signal, PU, is switched from logic state “H” to logic state “L”, then turns on the device P6 in Figure3-9. And the bit line is pulled to “H”; the flash cell data is latched in turn. After the logic operation of sense amplifier control logic, Figure3-8, the path reference current is turned off (because V4 in Figure3-9 is at “H”). From the time on, the sense amplifier enters sleep mode and without any current path exists. The purpose for the signal “PU” is same as the signal “PD”, to make the VS, input of the sensing device, to VDD earlier to save power, due to there exists a DC path in the sensing device if its input is not full swing.

Based on the sensing scheme, the access speed for read “0” from flash cell is determined by the bit line pre-charge speed. And the reference current and the voltage detector determine the timing for the sense amplifier entering sleep mode.

After all sense amplifiers enter sleep mode, the reference circuitry is then turned off by a combined signal (generated by each sense amplifier). From the time on, the chip enters standby mode. And the overall current dissipation is only standby current left.

3-3 Pre-charge Path Considerations

3.3.1 Pre-charge Path Considerations Versus Performance

According to the description in previous article, we know that we use the decoded bit line level for some purposes: 1) turned off the pre-charge path when the level is higher

than the trip point of the sensing device. 2) Enable the bit line discharge path to pull down the bit line bias when the bit line level is lower than the trip point of the sensing device in sensing stage. 3) Enable the bit line pulled up path when the bit line level is higher than the trip point of the INV7 in Figure3-9 to pull up the bit line bias for power saving. So the bit line level is very important to this novel sense amplifier, and must be handled carefully.

For example, if the pre-charge current is too large in bit line pre-charge stage, then the bit line level is pulled high very fast, if the bit line bias is higher than the trip point of the INV7 before the bit line pre-charge path turned off, then the sense amplifier enters sleep mode before sensing operation starts, and cause to the sense amplifier mal-function. If the pre-charge current is too small, it will take longer time to pre-charge the bit line and impact the access speed, because the sensing operation starts after bit –line pre-charge.

In order to concur the issue, a NMOS is used in the pre-charge path (N3 in Figure3-4). Because the bit line is connected to the “Source” node of the N3, the pre-charge current driving is getting weaker as bit line level is getting higher, because the Vgs (the voltage difference of the gate and source of a device) is getting smaller and the Vds (the voltage difference between the drain and source of a device) is getting smaller also. The current driving capacity, we can derive from [9]

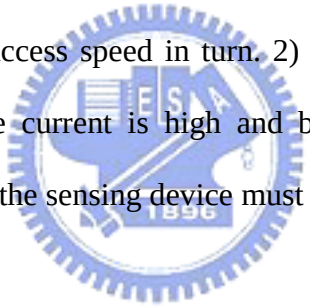
$$Id = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(Vgs - Vth)^2 \tag{3.1}$$

$$Id = \mu_n C_{ox}\frac{W}{L}\left[(Vgs - Vth)Vds - \frac{1}{2}V^2 ds\right] \tag{3.2}$$

So we can optimize the pre-charge current condition among: 1) Bit line pre-charge current, 2) The trip point of the INV7. 3) The trip point of the sensing device (inverter composed by P1 and N1). 4) Access speed. And trading off these 4 factors must be second thought. Because anyone factor of the four will impact the performance and function of the sense amplifier.

For instance, the trip point of the INV7 determines the timing to turn off the sense amplifier while reading a programmed state cell, that is, the trigger level impacts the current consumption while read “0”. If the trip point difference between the sensing device and the INV7 is too small, it’s risky to let the sense amplifier mal-function. If the difference is too wide, it costs more power. Because DC current paths are existed in the sense device (P1 and N1) and the INV7 (in Figure3-4).

Second, the relationship between pre-charge device and the sensing device should be taken into consideration, too. If the trip point of the sensing device is too high, this will cause lower Vgs and lower Vds of the N3 (As Figure3-4) in the bit line pre-charge path and causes longer bit line pre-charge time. If the trip point of the sensing device is too low, will cause the following results: 1) The drain bias of the selected cell is low and smaller cell current, and impact the access speed in turn. 2) When the trip point of the sensing device is low, the pre-charge current is high and bit line over pre-charge is easy to happened. So the trip point of the sensing device must be chosen carefully.



3.3.2 Pre-charge Path considerations Versus Process Condition

As for the bit-line pre-charge performance impacted by process condition, we can discuss the impact from circuit operation’s point of view. And the effects could be the following two: 1) Longer bit line pre-charge time. 2) Bit line over pre-charge.

First, let’s discuss the longer bit line pre-charge condition. The process worst condition for bit line pre-charge should be the SF (NMOS slow, PMOS strong). The slow condition means poor current driving or higher threshold voltage, fast process corner means strong device current driving or lower threshold voltage. This is because that:

Because there is an NMOS in the pre-charge path (N3), so the lower bit line pre-charge current and results in longer pre-charge time. Based on the following equation:

$$Cbl \times \delta V = I \times T \tag{3.3}$$

We can understand easily from the above equation that the pre-charge time will be longer in SF process condition.

The sensing device is an inverter structure, if NMOS were slow and PMOS were under fast process condition. This will result in higher trip point of the sensing devices. Based on previous description, higher the sensing device's trip point will make the lower Vgs of the pre-charge device (N3). This also makes the pre-charge poor current driving capability. We can see the result from the equation 3.1 and 3.2.

Second, let's discuss the over pre-charge condition. The worst process condition for bit line over pre-charge should be FS corner. (NMOS is under fast corner, and PMOS is under slow condition) This is because that:

1) If NMOS was in strong process condition, this means that N3 has strong current driving. That is the bit line pre-charge current is stronger than other conditions.
2) The process condition (FS) results in lower trip point of the sensing device (P1 and N1).

These 2 factors both make the stronger pre-charge current. Because there exists little time delay after the pre-charge is finished. So stronger pre-charge current makes more over pre-charge voltage. The more over pre-charge voltage will take more time to discharge the bit line to the trip point of the sensing device when reading erased cell and results in longer access speed.The other drawback is the trip point of the INV7 must be set to a higher value to prevent the path in Figure3-9 been turned on un-expect.

Based on the above description, we need to choose the trip points of the sensing device and INV4 carefully. Besides the sensing device's trip point, the driving capability of the pre-charge device is also important. Because these 2 factors are related to the performance and function of the proposed sense amplifier.

Regarding the trip points of the sensing devices and the INV7, we can adjust the

width and length of the sensing device to change the device's trip point to achieve better pre-charge and sensing performance within whole operating voltage, process corner and temperature range.

3-4 Low Power Supply Voltage Architecture Introduction

As like the description in the previous section (low power consumption sense amplifier), the sensing scheme (sense amplifier) for the low voltage supply (the power supply could be as low as 1.0v) is almost same as the low power dissipation architecture.

Figure3-10 is the sense amplifier architecture for low power supply. The only difference is the bit line pre-charge device (block A). There are 1 PMOS and 1 NMOS to compose to pre-charge path in the low power consumption sense amplifier to slower the pre-charge speed for easy controlling the latch. In low power supply version, the feature for read '0" latch function is been remove, only the read "1" latch is kept.

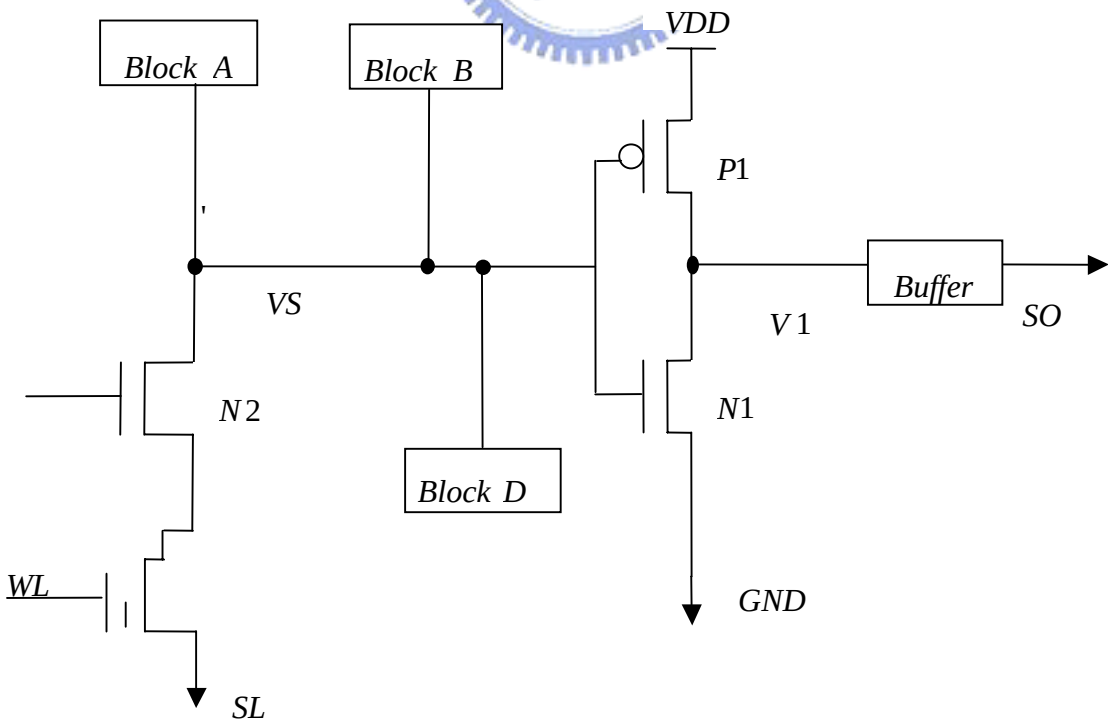


Fig 3-10 Sense Amplifier architecture for low power supply

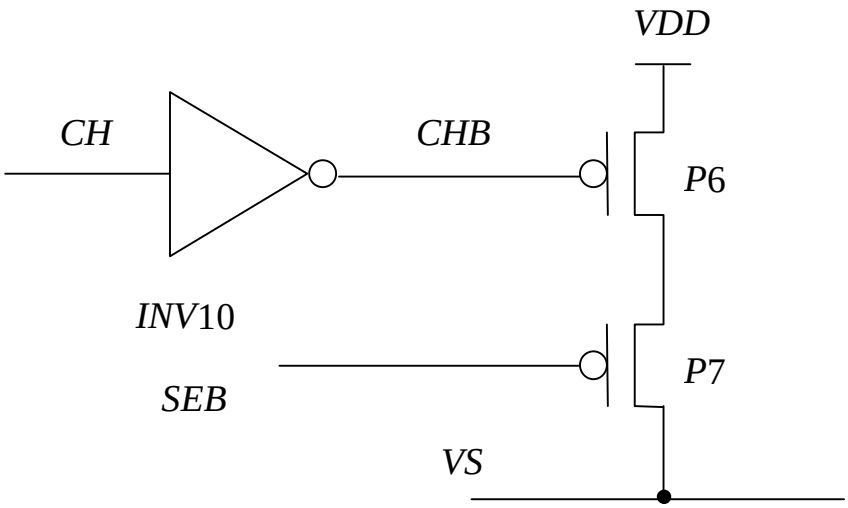


Figure 3-11 Pre-charge path for low VDD application

Due to the low voltage supply could be as low as 1 v, so the pre-charge device must be modified from 1 PMOS and 1 NOMS to 2 PMOS (as Figure 3-11), because the driving under low voltage is poor for pre-charge and can't pre-charge the bit line to the trip point of the sensing device. If only PMOS is used for bit line pre-charge, we can make sure that the bit line could be pre-charge to the trip point of the sensing device.

In this architecture, the bit line bias could be pre-charged to the level by the requirement of flash cell's operating. (The bias could be from VSS to VDD because we only use PMOS to pre-charge the bit line). And the bit line pre-charge pulse is like the low power scheme, self-defined by each sense amplifier. Due to the power supply is down to around 1v, so the word line boost scheme is also used in the simulation (one shut boost is used in the simulation).

Compared to low power dissipation architecture, due to the pre-charge device is changed to PMOS only; the bit line pre-charge speed is faster. And the overall access performance is close to 1.8v power supply. But lacks the data latch feature.

Chapter 4

Simulation and Silicon Result

4.1 Test Chip's information

Test chip's configuration is 8k bytes, composed by 512 word lines and 512 bit lines, the IO width is by sixteen, so there are 16 sense amplifiers in the test chip. The typical power supply is 1.8v, and word line boost scheme (one shut boost) is used for improving read cell current to improve access performance. Reference current is generated by real flash cell for process tracking (track the cell current for different process corner and different operating conditions)

Process and flash cell (NOR flash):

The process and spice model are based on TSMC 0.18um Embed-flash (embedded flash) process.

This paper focuses on a novel topology of sense amplifier for nonvolatile memories that suits for low power dissipation for read operation and easy to switch to low voltage application, the power supply could be as low as 1.0v, and the speed performance could meet the application requirement.

Based on HSPICE simulator and TSMC 0.18um embedded flash process model. The process variation (process corners those include SS, SF, FF, TT, FS for NMOS and PMOS process condition) is taken into consideration for simulation. Real flash cell is used to

generate the sensing reference current to tracking the main flash cell for process variation tracking. By doing so, we can make sure that the sensing ratio between flash cell current and reference current almost keep the same ratio among whole process corner. The ratio for flash cell current to the reference current is chosen to close to 3:1.

In order to gain the access speed performance and higher reliability after cycling, word line boost is needed in current technology. The simplified circuitry is as the following figure. (Figure4-1)

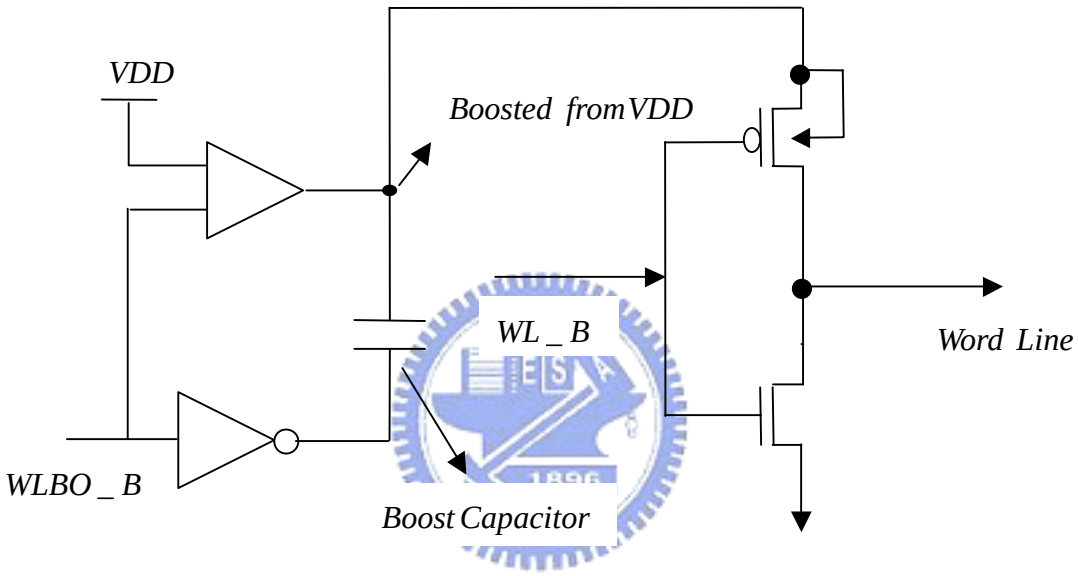


Figure4-1 Simplified circuit of word line boost

The signal “WLBO_B” is the boost enable signal, which comes from the sense amplifier. “WL_B” is word line active signal and “Word line” is connected to the control gate of flash cell. And the word line bias is boosted from VDD; the boost capacitor area is related to the number of word line driver.

Due to higher word line level results in higher cell current and improves access time much. We can see from the following simulation result (Figure4-2) and make sure that the circuitry and word line boost scheme is needed for the proposed design and application requirement. The simulation condition is under VDD=1.6v, room temperature and typical

process condition.

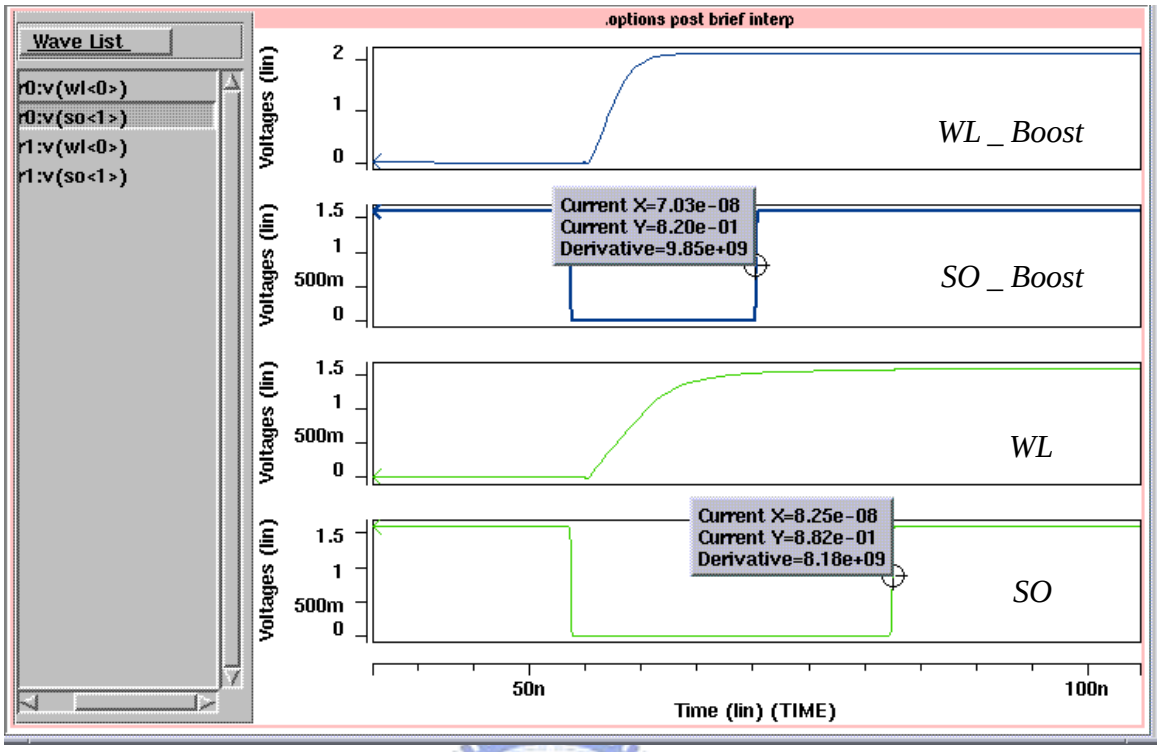
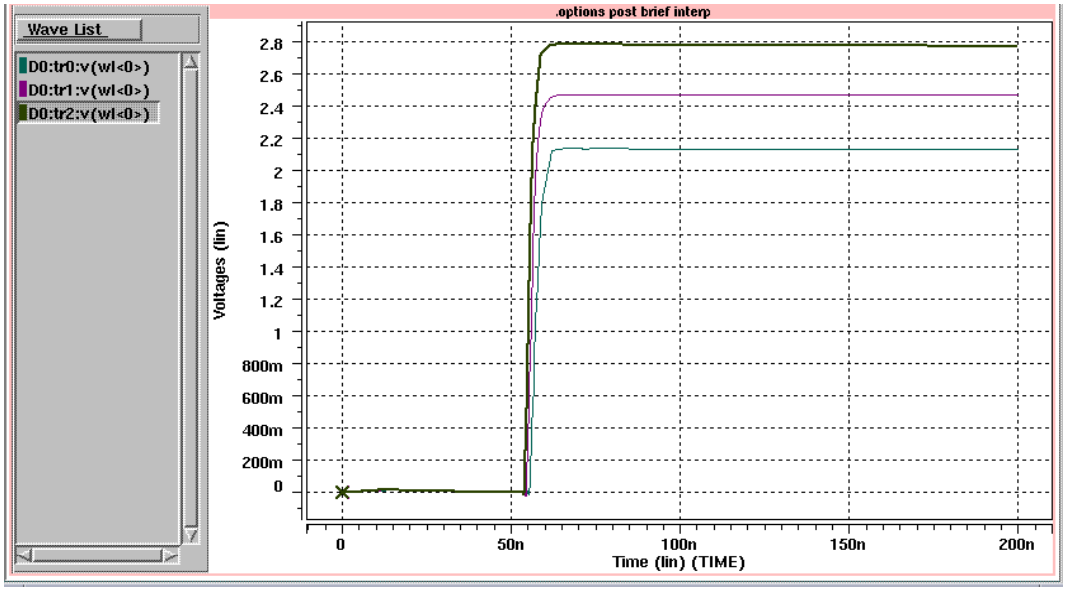


Figure4-2 The difference for speed and word line level between with word line boost and without word line boost.

The curves in Figure4-2 show the word line bias and output of the sense amplifier. The “WL_Boost” is the word line bias after boosted, and “WL” are the word line bias without word line boost. “SO_Boost” is the sense amplifier output with word line boost circuitry. “SO” is the sense amplifier output without word line boost. Based on the above simulation result, it’s easy to see that access speed is improved much with word line boost circuitry. (Around 12 ns improvement)

So the simulation result for access speed and power consumption includes the word line boost circuitry. For the word line bias after one shut boost is 2.15v (VDD=1.6v, TT), 2.45v (VDD=1.8v, TT) and 2.75v (VDD=2.0v, TT). The word line level after boosted simulation result is as the following figure. (Figure4-3)



VDD	1.6v	1.8v	2.0v
Vwl	2.15v	2.45v	2.75v

Fig 4-3: Word line bias after boosted. (VDD=1.6, 1.8 and 2.0v)

As for the array architecture of the test chip is as the following diagram. (Figure 4-4)

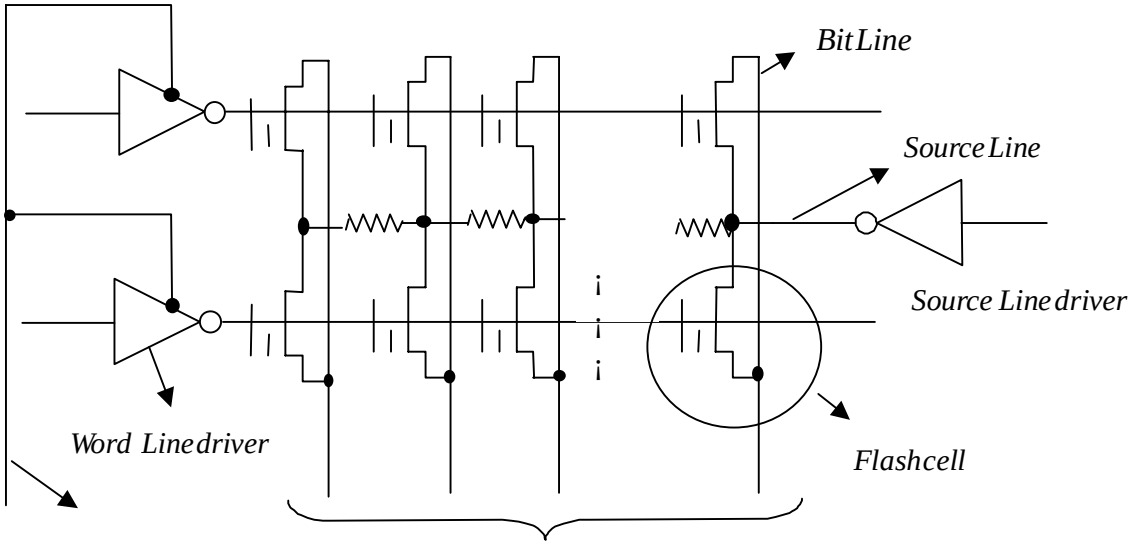


Figure4-4 Array architecture

From Figure 4-4, the gate of flash cell is connected to the output of word line drivers; the source of flash cell is connected to the output of source line driver via diffusion. The source line driver is necessary for the cell structure. (The source line driver is used to pass high voltage for program operation) The bit lines of all flash cell are connected to sense amplifier.

Regarding the read access path of the test chip is as the following figure. (Figure 4-5)

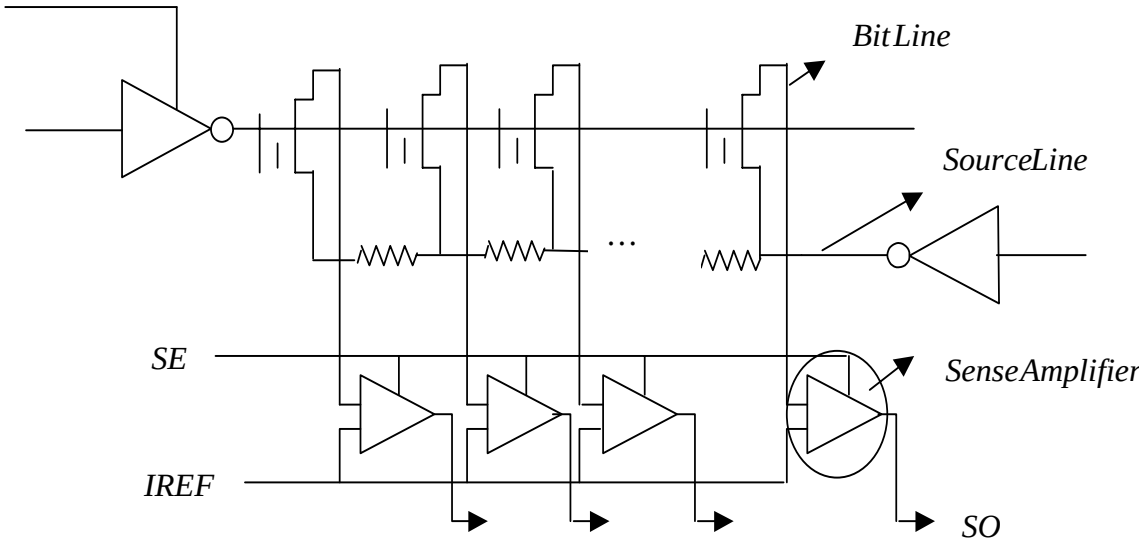


Figure4-5 Access path

From Figure4-5, we can see that there is effective resistor between the source node of flash cell and source line driver. And there is current flow existed in the path when the chip performs read operation. For example, when the “read” operation starts, there are 16 flash cells selected. From Figure4-5, the current flows through the source line for each flash cell, and the total current of the 16 cell flows through the source line driver. So the source of flash cell is not connected to real ground in real operating condition. And the effect will reduce the flash cell current. We can know the cell current degradation from the following equation.

$$Id = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(Vgs - Vth)^2 \tag{4.1}$$

$$Id = \mu_n C_{ox}\frac{W}{L}\left[(Vgs - Vth)Vds - \frac{1}{2}V^2 ds\right] \tag{4.2}$$

The flash cell current behavior is close to a pure NMOS, so the higher potential of the source node of the flash cell will cause lower Vgs and degrade the flash cell current

and result in slower speed.

4.2Simulation Result for Low Power Consumption Architecture

4.2.1 Simulation Result for Access Speed

For low power consumption version, as we discussed in the previous sections, the access speed is dominated by: 1) the timing of bit line pre-charging. 2) Flash cell current compared with reference current. As for bit-line pre-charge, the speed is dominated by a) bit line pre-charge current. b) Bit line capacitor load. Choosing a large device in the pre-charge path could enhance the current for bit line pre-charging. But if the pre-charge current is too large, it will cause the sense amplifier function fail; this is because that the bit line pre-charge control logic is not so fast as the bit line is pre-charged to the trip point of INV4. In the proposed design, the pre-charge current is chosen to cover whole process corner and the sensing function is workable within whole operating range. (VDD ranges from 1.6v to 2.0v, temperature ranges from –45℃ to 125℃). So the pre-charge current is a trade off between speed and sensing function. Figure 4-6 shows the timing waveforms of word line and bit lines for different flash cell contents.

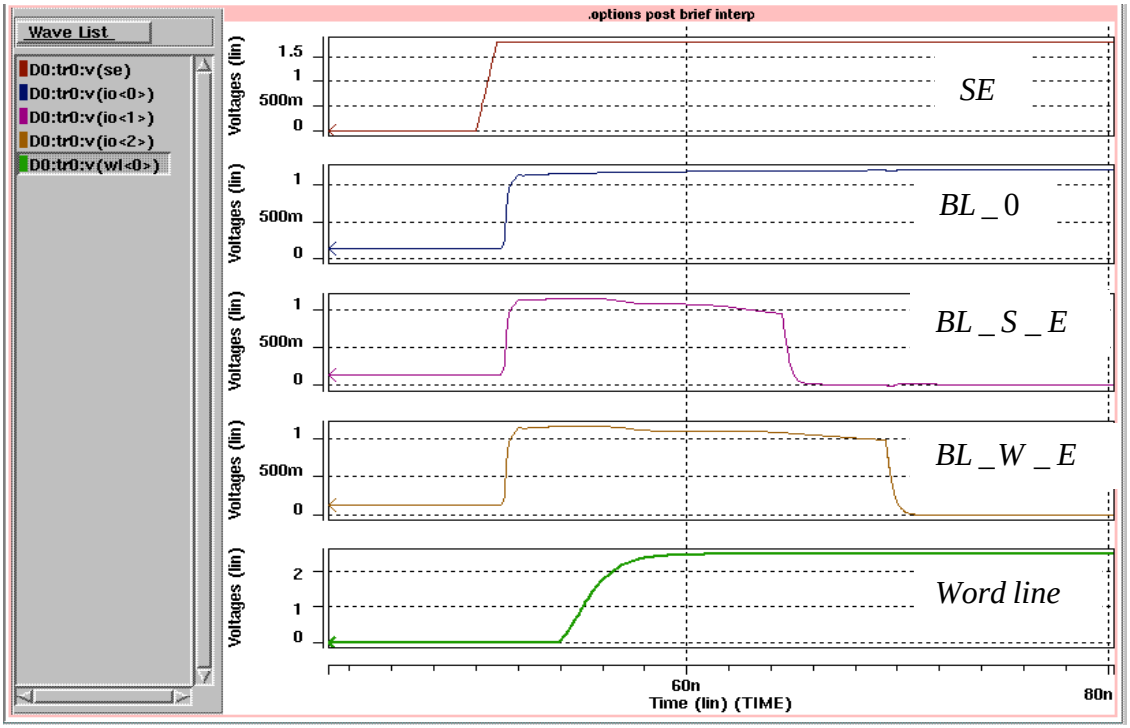


Figure 4.6 Timing waveform of word line and bit lines bias for different cell data those related to sense amplifier active signal (SE)

"SE" is the sense amplifier enable signal. The curve "BL_0" shows that the bit line that is connected to programmed state flash cell. Normally, the flash cell current is close to zero under programmed state, so the bit line pulled down current is close to zero, too. So we can see that the bit line bias is getting higher after "SE" is active. The curve "BL_S_E" shows the bit line that is connected to strong erased flash cell. Normally, most of flash cells are strong erased state in the beginning. The flash cell current for those strong erased state cells is high and the speed to pull down the bit line level is faster. The curve for "BL_S_W" shows the bit line that is connected to weak erased flash cell. The weak erased cell is used to simulate the flash cell that has been used for thousands of program/erase cycles under real application condition. After certain number of program/erase cycles, the flash cell current is degraded gradually due to charge trapped in the tunneling oxide. And the trapped charges cause the cell threshold voltage getting

higher, and results in lower flash cell current. So the speed to pull down the bit line level is slower than the strong erased cell. So it is necessary to take the cell current degradation effect into design/simulation consideration to make sure that the flash reliability is good enough to function property after thousands of program/erase cycles. The other signal is the word line waveform after word line has been boosted. In the proposed design, the word line is active after bit lines are pre-charged.

Regarding the source line potential effect as we discussion earlier in this chapter, we can understand from the following waveform.

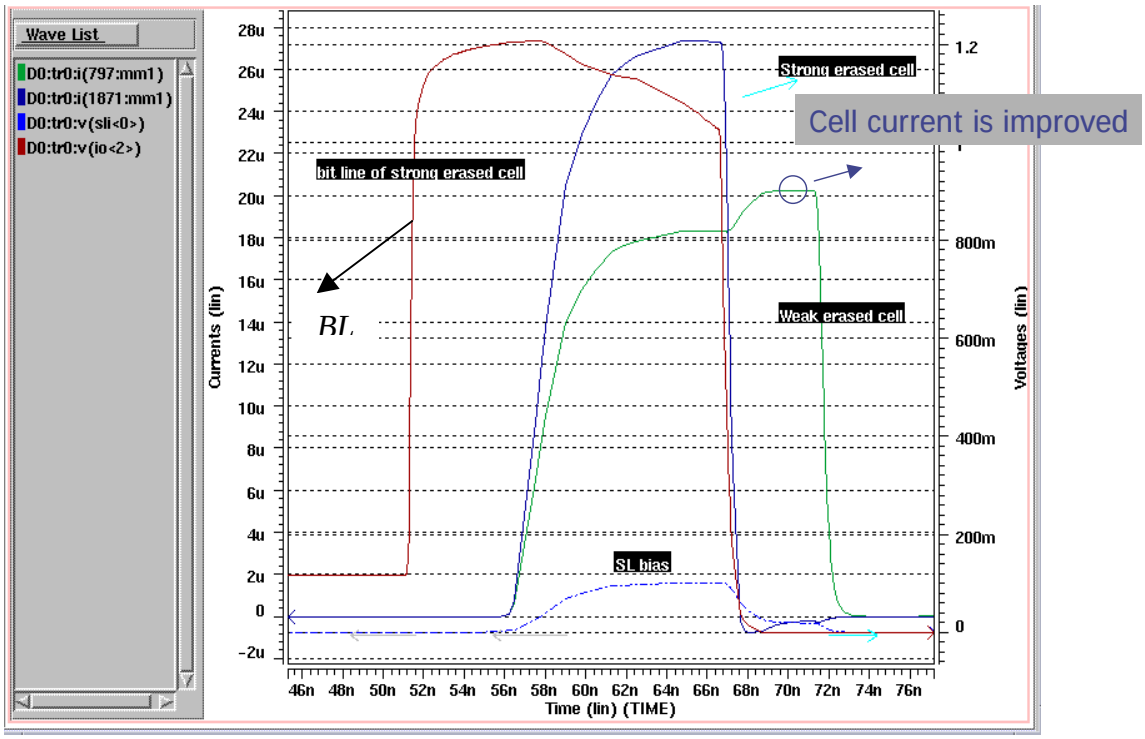


Figure 4-7: The waveforms for bit line bias, flash cell current and source line potential

In Figure 4-7, there are bit line bias of strong erased cell, flash cell current of strong erased cell, cell current of weak erased state cell current and source line potential. We can see that the timing and flash cell current and source line bias relationship. From the figure we can see that the selected bit line is pulled down to ground after the sensing operation is finished for strong erased state cell. Based on the array architecture, Figure

4-4, the source line grounding is improved after the erased state bit line is pulled down to ground. This is because that the current flow through the source line is reduced, and the source line potential is reduced also. And the from Figure 4-7, we can see the weak erased cell current is improved after the source line is better grounding, this improves the access performance much, because the current difference between flash cell and reference current is bigger after source line is lower.

Figure 4-8 shows the detail waveforms for word line, bit lines and source line potential the related timing.

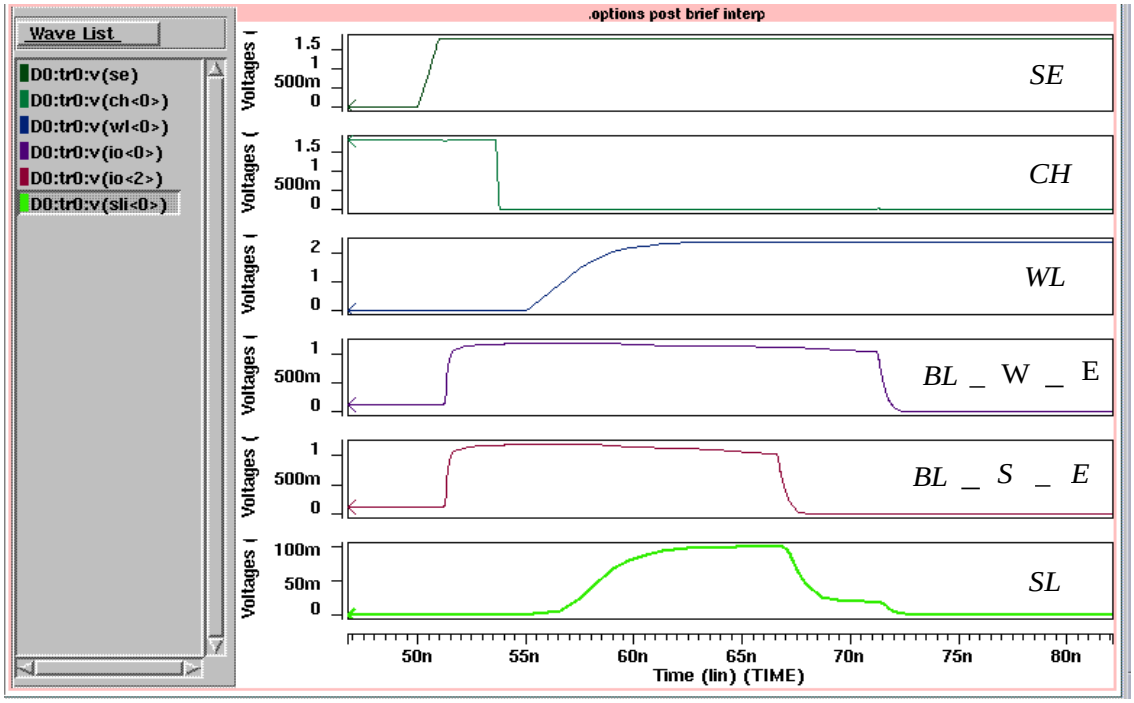


Figure4-8 Simulation waveform for word line, bit line, source line and the bit line pre-charge signal

From Figure 4-8, we can see the timing sequence and bias condition of the selected flash cell including word line, bit lines (for strong erased cell and weak erased cell) and the better grounding of source line potential. The signal “CH” is used for bit line pre-charge controlling. It is easy to explain the timing sequence from the figure; the first

step of sensing operation is bit line pre-charge. We can see that the word line and source line are at ground potential before bit lines are pre-charged. After bit lines are pre-charged, word line is then boosted and flash cell's current begins to flow through source line. So we can see the source line potential is getting higher and the timing follows word line potential. After the sensing operation of strong erased cell is finished, the bit line is pulled down to ground. At this time, no cell current of the sensed cell flow through the source line, so the source line potential is lowering. The source line is ground potential after all the selected cells are sensed.

As for data sensing, as we discussed earlier, larger flash cell current will speed up the access time, because larger cell current makes bigger current difference between flash cell current and reference current, and be supported by simulation results (Figure4-3). So the world line bias after boost (or charge pump) plays a very important role in the sensing scheme.

The access time for typical condition (The VDD=1.8v, temperature is 25℃, typical NMOS, PMOS process condition is around 18ns in simulation result. As for whole condition within process corner (SS, TT, FF, SF, FS for NMOS, PMOS process corner) and power supply (ranges from 1.6v to 2.0v) and temperature (ranges from –45℃ to 125 ℃), the access time is less than 40ns. The worst condition for access speed is under SF corner due to the slower bit line pre-charge speed, since weak N3 (as Figure .3-2) driving capability dominated the pre-charge current and slower the bit line pre-charge. For sensing erased state data, we can see from Figure4-8, the word line is boosted after bit line pre-charged, and the flash cell current begins to sink current after word line bias ramped up. So the sensing operation is after bit line pre-charge, this is because that slower pre-charge speed will cause to slower the sensing speed for reading "1".

As for the trip point of the sensing device and the voltage detector (INV4), the simulation waveform for the 2 key devices are as Figure 4-9, Figure 4-10 and Figure 4-11.

From Figure 4-11, the trip point between these 2 devices exists a voltage difference to make sure the sensing function is correct.

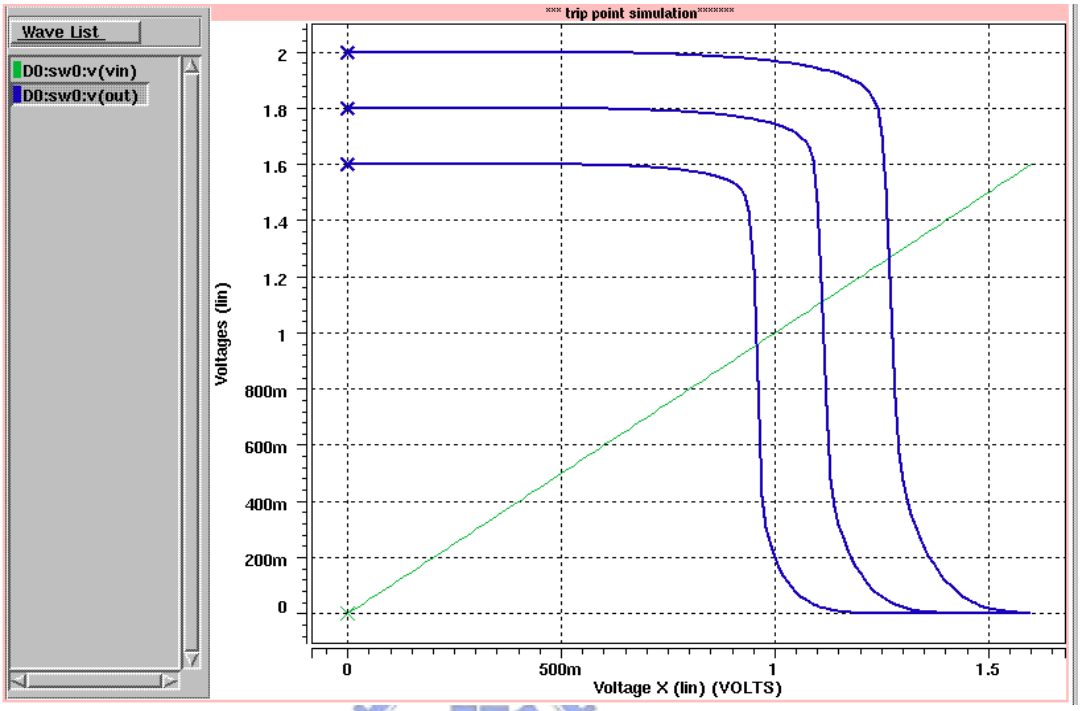


Figure4-9: Trip point of the sensing device

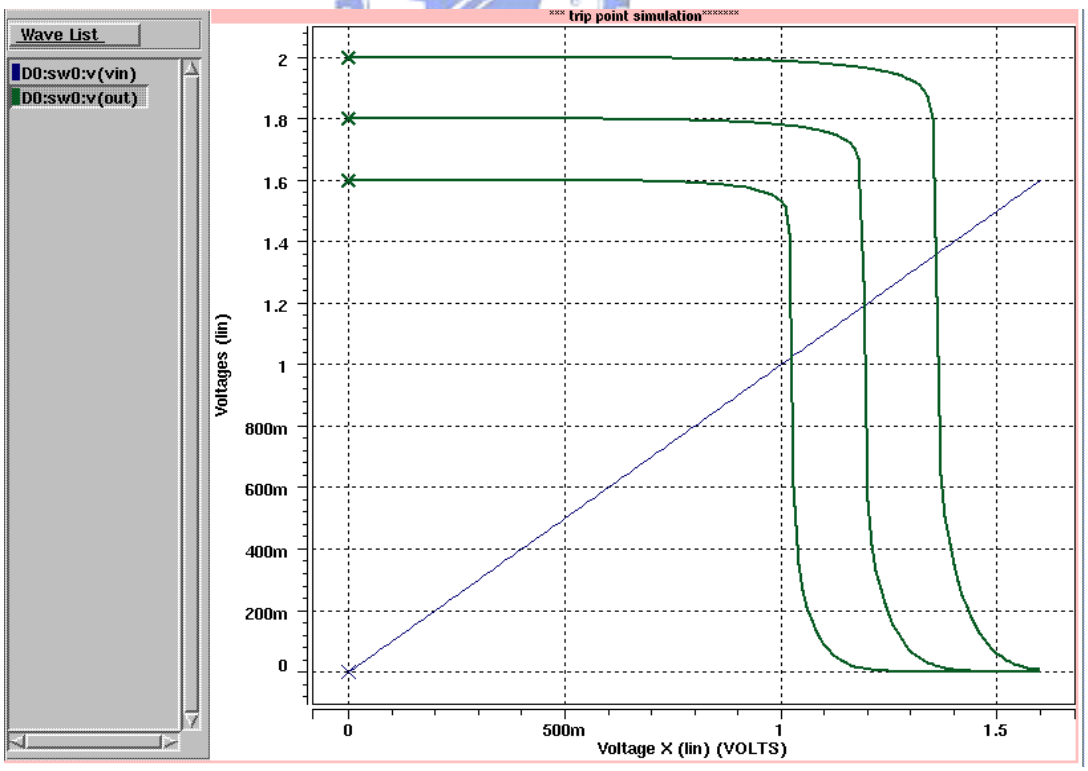


Figure 4-10: Trip point of the voltage detector (INV7)

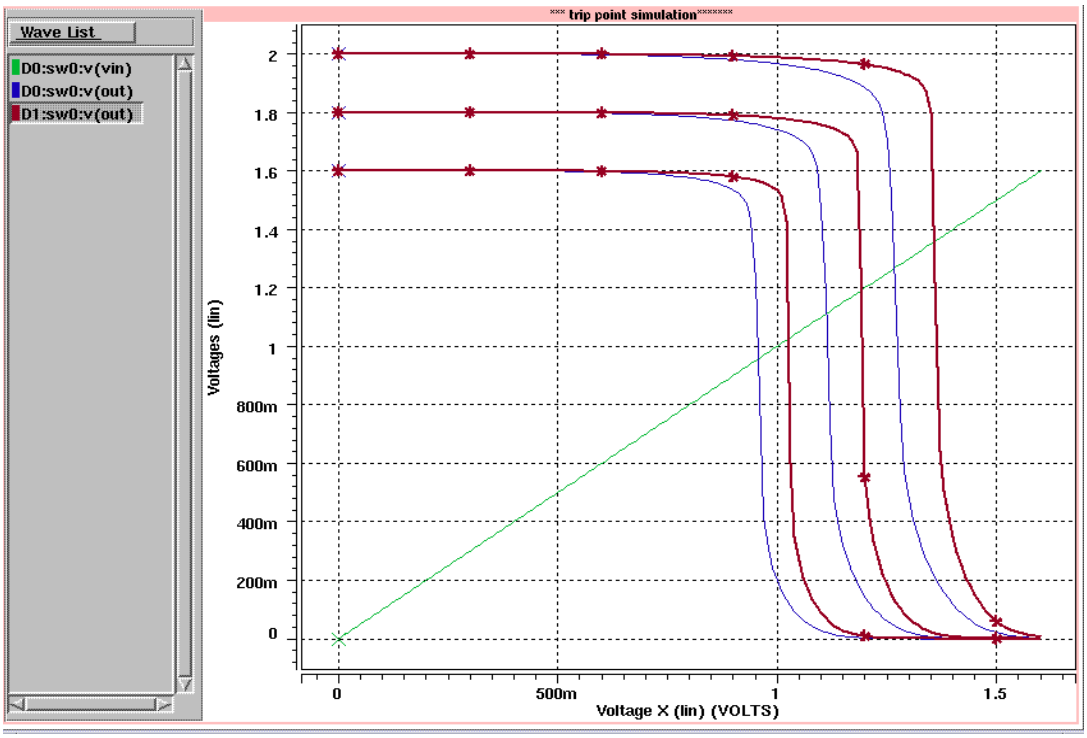


Figure 4-11: The trip point comparison between the sensing device and INV7

As we discussed, the worst process condition for bit line pre-charge is SF (N-slow, P-fast). Figure 4.12 shows the timing waveform for SF process condition, the VDD is 1.6v, temperature is -45℃.

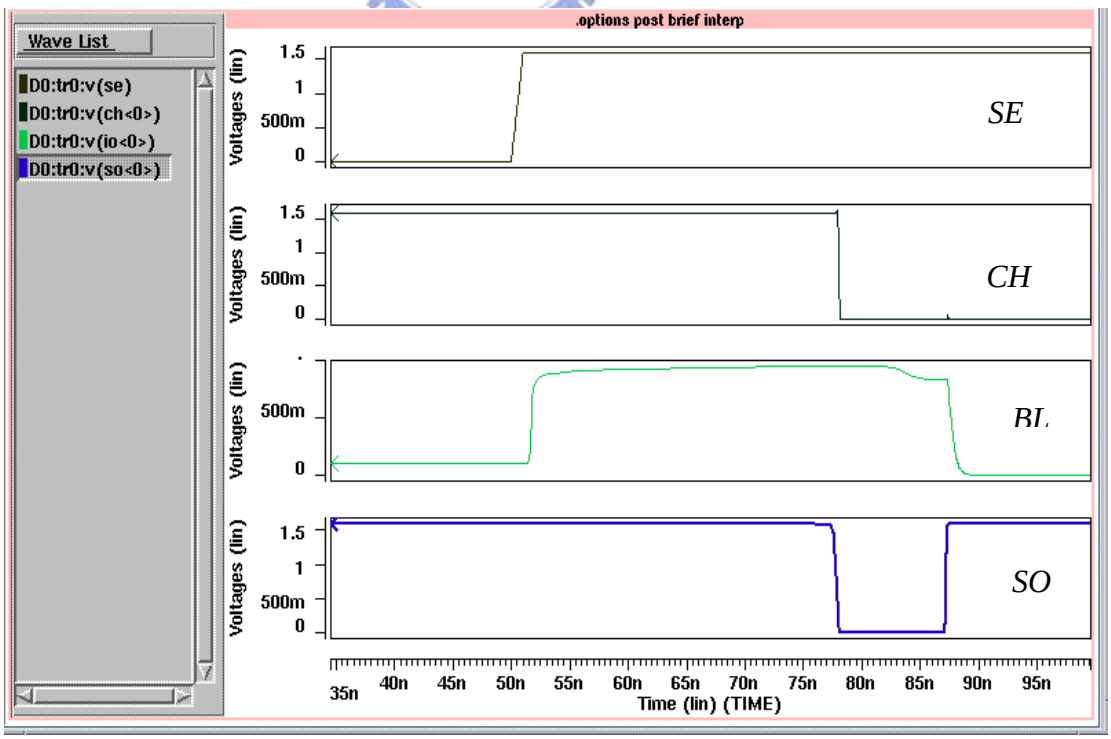


Fig4-12: Timing waveform under SF process condition, 1.6v, -45℃.

From Figure4-12, we can see the pre-charge time is pretty long under SF process condition. It takes over 25 ns to pre-charge the selected bit line (from “SE” active to “CH” goes “L”). And most of the access period is bit line pre-charge.

Based on the discussion in section 3 for the pre-charge consideration, the major concerns for the proposed sense amplifier are longer pre-charge time and over pre-charge. If we extend the operating condition, for example, we simulate the performance under VDD=1.4v

Figure4-12 shows the waveform for extending VDD to 1.4v. Form Figure4-13, the timing pulse for bit line pre-charge is over 100ns. This is because the current for bit line pre-charge is pretty low caused from lower Vgs of N3 (as figure3-4)

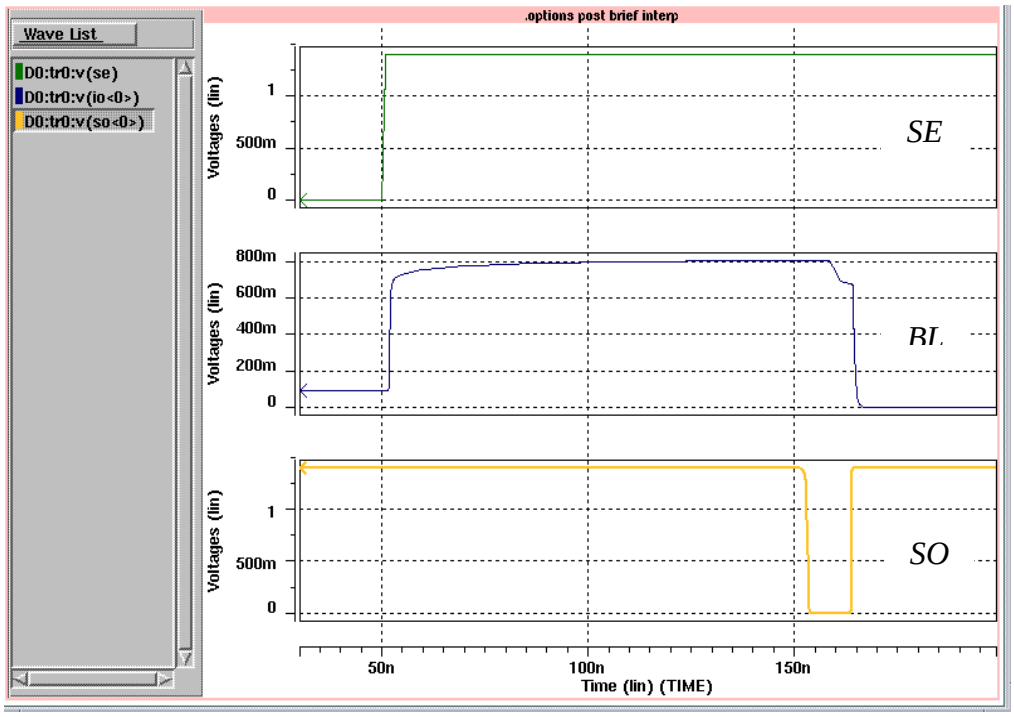


Fig4-13: Waveform under SF, 1.4v and 125℃

Why SF process corner is the worst condition for pre-charge? We can understand from the following simulation waveform, Figure4-14. Figure4-14 is the simulation result

for the sensing device under VDD=1.6v, room temperature but different process condition.

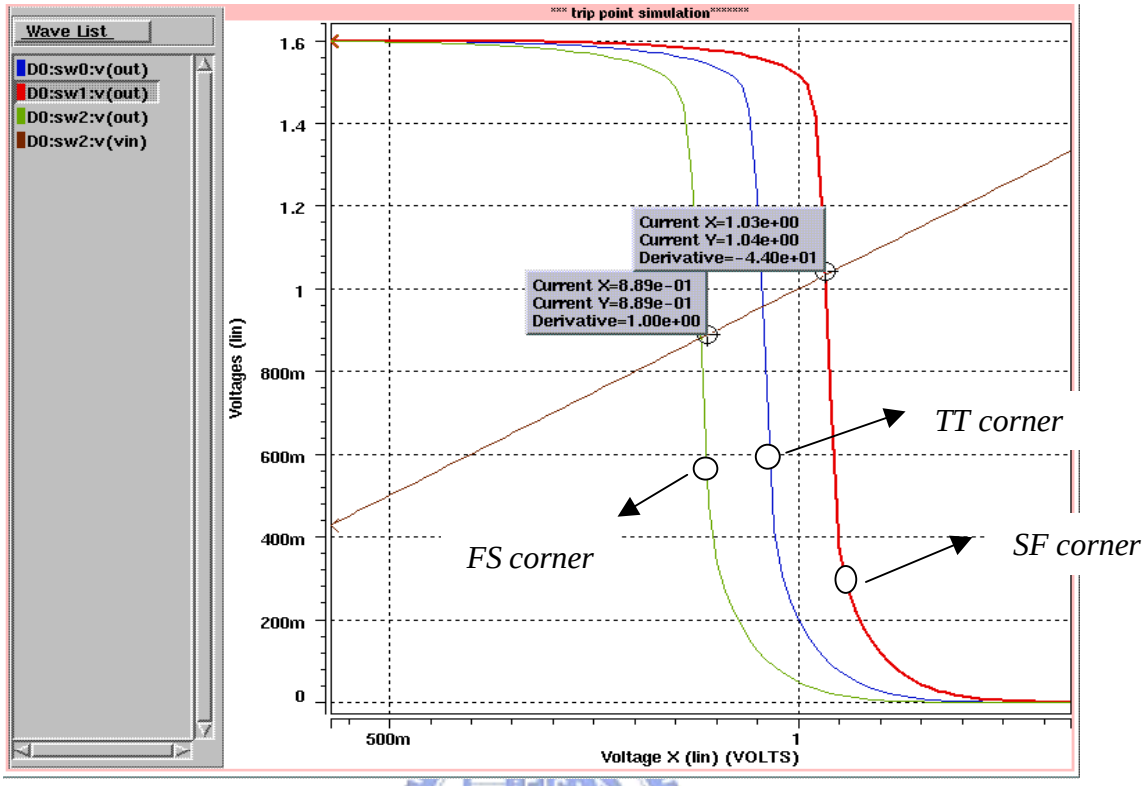


Figure4-14: Trip point under different process corner

From Figure4-14, the trip point for the sensing device exists over 0.1v difference within process variation. The variation causes to different bit line pre-charge performance. And we can realize the impact from the equations 3.1 and 3.2.

After discussion the longer bit line pre-charge issue, now we discuss the bit line over pre-charge issue. As we have discussed in the previous section, FS (NMOS is under fast and PMOS is under slow process condition) is the worst process condition and will cause bit line over pre-charge issue. The worse pre-charge the slower access speed, because it will take longer time to discharge the bit line. Figure 4-15 shows the over pre-charge cause slower access speed under FS process condition.

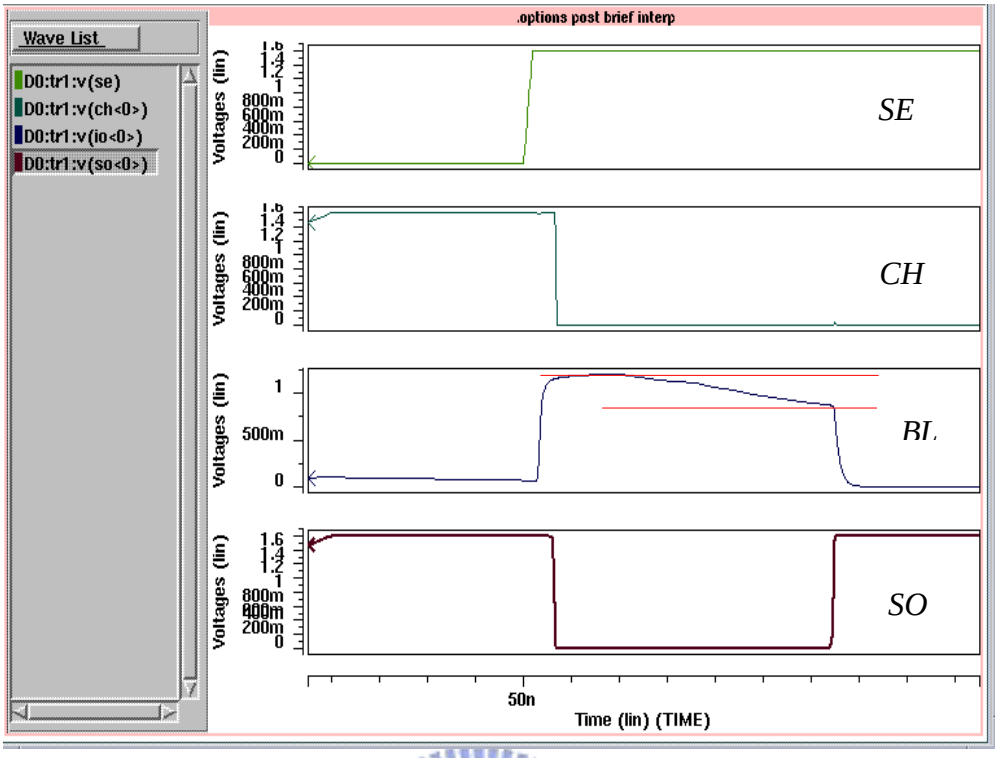


Figure4-15: Timing waveform under FS process condition, 1.6v, 125℃

Regarding the sensing ratio, the simulation result is as Figure 4-16. The sensing ratio is defined as the flash cell current divided by reference current that flows into sense amplifier. The reference current is the current flows through P5 (as Figure3-3).

The upper curve is the flash cell current and reference current flow into sense amplifier under TT process corner, room temperature and VDD=1.8v. The lower curve is same condition except VDD=1.6v. From Figure4-16, the sensing ration is close to 33% under different operation conditions.

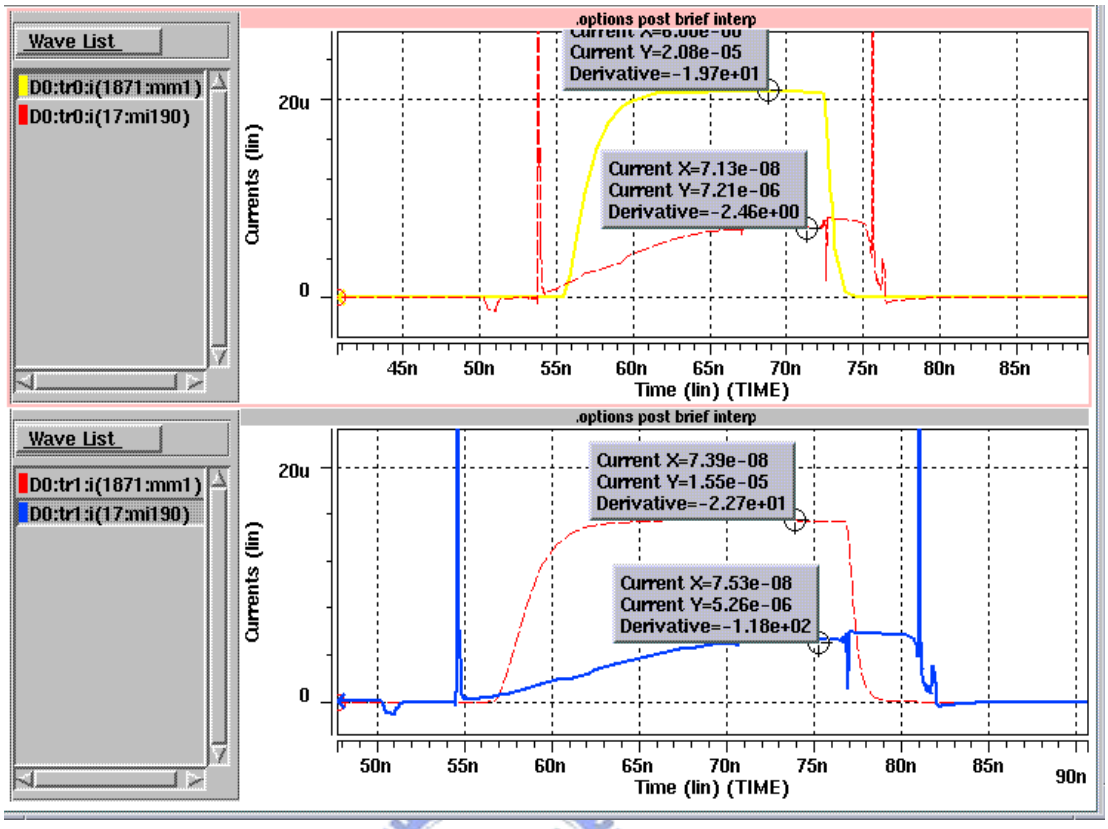


Fig4-16: Icell versus Iref under 1.8v and 1.6v

As for the signal to turn sense amplifier off, SOFF, the simulation waveform is as Figure 4-17.

Based on the previous discussion, the SOFF is active right after sensing “erased state” cell is finished. But will have time delay for sensing “programmed state” cell. This is because that the bias, VS, must be pulled to the trip point of INV7 by reference current. In Figure 4-17, the “SO_0” is the signal of the sense amplifier output for sensing “programmed state” cell. “SOFF_0” is the SOFF signal for reading “programmed state” cell.

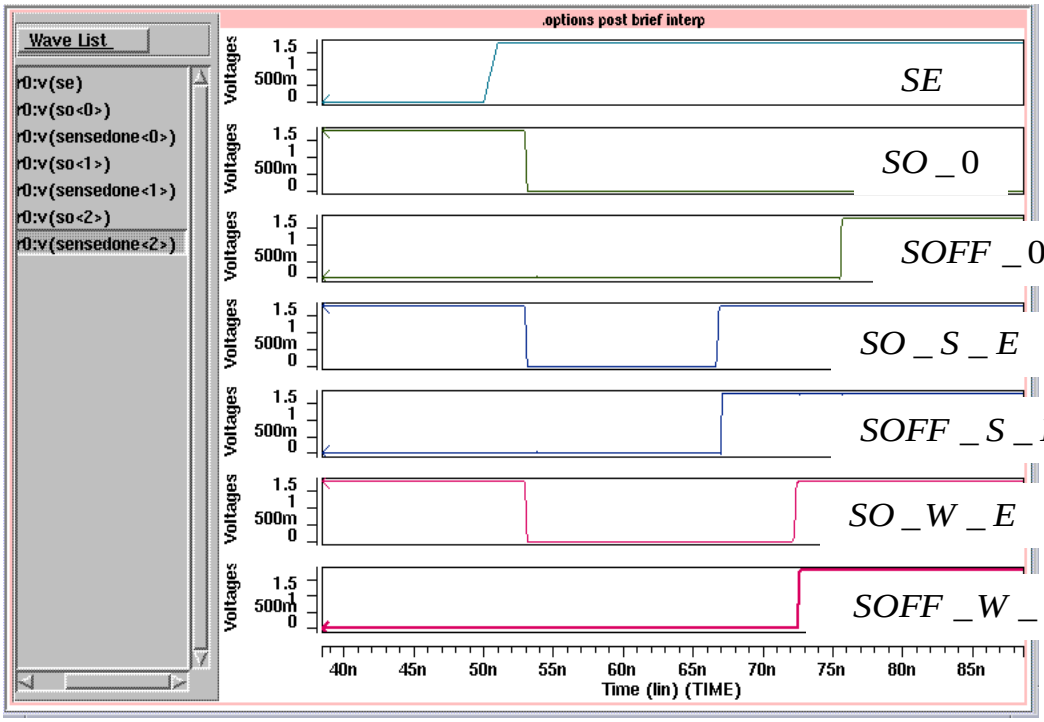


Figure 4-17: Sense amplifier turned off signals for different data patterns.

4.2.2 Simulation Result for Access Power Dissipation

The power consumption simulation result for each sense amplifier is around 3-5uA under VDD=2.0v, 1M frequency. Due to the sense amplifier enters sleep mode after sensing operation is completed; the most of the cycle is in standby mode and no power dissipation in the rest of the whole cycle. The power consumption in the sense amplifier comes from: 1) bit line pre-charge current and the power consumption is related to bit line capacitor load. 2) The sensing device, because there is a sensing period that a current path exists in the sensing device and the path won't be turned off until the sensing operation is completed. 3) The same phenomenon exists in the voltage detector device (INV4), the current path in the device is turned off while the data "1" being sensed out or the bit line level is over the voltage detector, due to the input of the device is pulled up to VDD or pulled down to VSS. 4) The summation of transient current of all logic gates in sense amplifier. Because there are more logic gates to generate signals for pre-charge and

sensing control logic in the sense amplifier, more percentage of power consumption comes from this portion. 5) Standby current. When the sensing operation is finished, the sense amplifier enters sleep mode and no current path existed in the sense amplifier, only standby current existed in this stage.

The power dissipation waveform is as Figure 4-18.

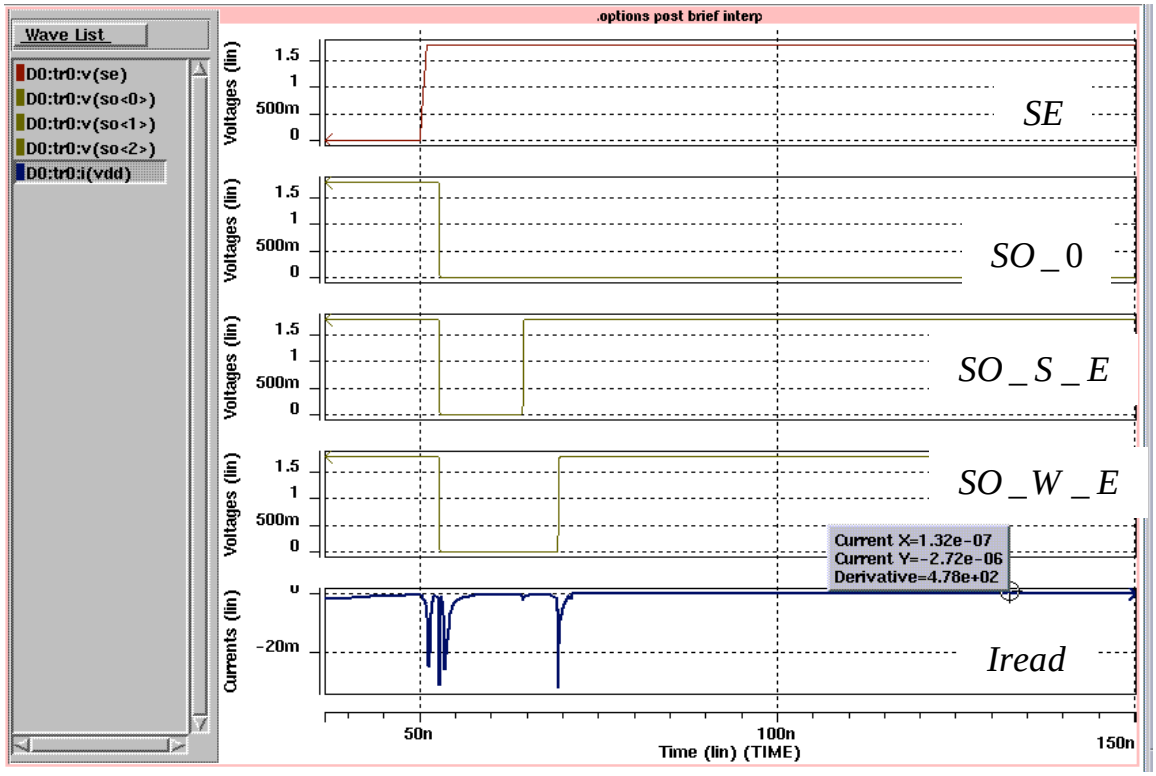


Figure 4-18: Waveform for power dissipation

The signal “SO_0” is the output signal for programmed state, “SO_S_E” is the output signal for strong erased state cell from sense amplifier, “SO_W_E” is the output signal for weak erased cell, “Iread” is waveform for the whole chip power dissipation while the chip performs “read” operation. From Figure 4-15, the power dissipation of the whole chip is only standby current left after sensing operation is completed (after weak erased state cell data is sensed), even the sense amplifier active signal (SE) is still enabled.

Table4-1 is the simulation result of whole chip average current dissipation under

1MHz frequency within whole operating range and process conditions. From the following table (Table 4-1), the current dissipation is low compared to conventional approach. Because there is no static current existed in the chip, the sensing scheme is suite for low frequency application and keeps the access speed performance.

Table 4-1: Read operation active power dissipation

125℃	FF	FS	SF	SS	TT
1.6v	97	85	71	55	65
1.8v	110	97	90	67	73
2.0v	121	109	94	76	88

25℃	FF	FS	SF	SS	TT
1.6v	61	59	64	54	57
1.8v	72	70	73	64	70
2.0v	86	77	83	73	77

-45℃	FF	FS	SF	SS	TT
1.6v	61	60	60	51	57
1.8v	74	70	69	60	68
2.0v	83	79	80	71	76

Unit: uA under 1M Hz

Table 4-2 shows the access speed for reading programmed state cell (or bit line pre-charge speed). We can see from the following table that the slowest pre-charge process condition is SF. The pre-charge even closes to 30ns under -45℃, and the pre-charge speed dominates the overall access speed.

Table 4-2: Read operation access speed for access “0”

125℃	FF	FS	SF	SS	TT
1.6v	3.5	4.5	19	10.5	5
1.8v	3.5	3.5	18	10	4
2.0v	3	4	10	9	5

25℃	FF	FS	SF	SS	TT
1.6v	3	3.5	25	11	4.5
1.8v	3	3	14	9	4.5
2.0v	3	3	10	7	3.5

-45℃	FF	FS	SF	SS	TT
1.6v	3	3.5	29	12	4
1.8v	2.5	2.5	13	7.5	4
2.0v	2.5	2.5	7	6	3

Unit: ns

Thable4-3 shows the access speed for reading erased state cell. We can see the slower conditions are SF and FS. This matches the previous discussion. The SF corner is slower bit line pre-charge, FS is easy to over pre-charge. These 2 phenomenons affect the sensing speed much.

From Table 4-3, the access speed is faster than 40ns within whole process and operating conditions.

Table 4-3: Read operation access speed for access “1”

125℃	FF	FS	SF	SS	TT
1.6v	21	36	35	27	24
1.8v	17	31	32	26	20
2.0v	15.5	27	23	21	19

25℃	FF	FS	SF	SS	TT
1.6v	15.5	28	36	24	20
1.8v	14.5	23	23	21	17.5
2.0v	15	21	19	17.5	17

-45℃	FF	FS	SF	SS	TT
1.6v	15	26	37	24	19
1.8v	15	23	20	17	15
2.0v	13	19	14	14	13

Unit: ns

4.3 Silicon Result:

The silicon result for access speed and power consumption is as table 4-4

Table 4-4: Silicon result for access speed. (read “1”)

	-45℃	25℃	125℃
1.6v	21.8	22	23.4
1.8v	15.7	16.7	18.2
2.0v	13.4	14.4	16.8

Unit: ns

Table 4-5: Silicon result for read active power dissipation.

	-45℃	25℃	125℃
1.6v	70.5	67.7	77
1.8v	83	79.5	90.5
2.0v	98	93.5	105.5

Unit: uA

The silicon is neasures by MOSAID tester, and the result is close to the simulation result.

4.4 Sense Amplifier for Low Power Supply

Based on the description for the sense amplifier architecture for lower power supply, the major difference between these two architectures is only the pre-charge path, the devices for per-charge path is changed from 1 NMOS and 1 PMOS to 2 PMOS to improve low VDD pre-charge performance. Except the pre-charge path, the latch function is removed in this architecture because the trip point for sensing device and voltage detector is hard to control under new pre-charge scheme. Table 4-6 is the simulation result

for this architecture under VDD=1.0v and 1.2v. Figure 4-19 shows the bit line pre-charge waveform for this sense amplifier including Ipre_ch (bit line pre-charge current) and CH, we can see that the pre-charge speed is fast (less than 10ns under VDD=1.0v) for the modified sense amplifier because higher bit line pre-charge current..

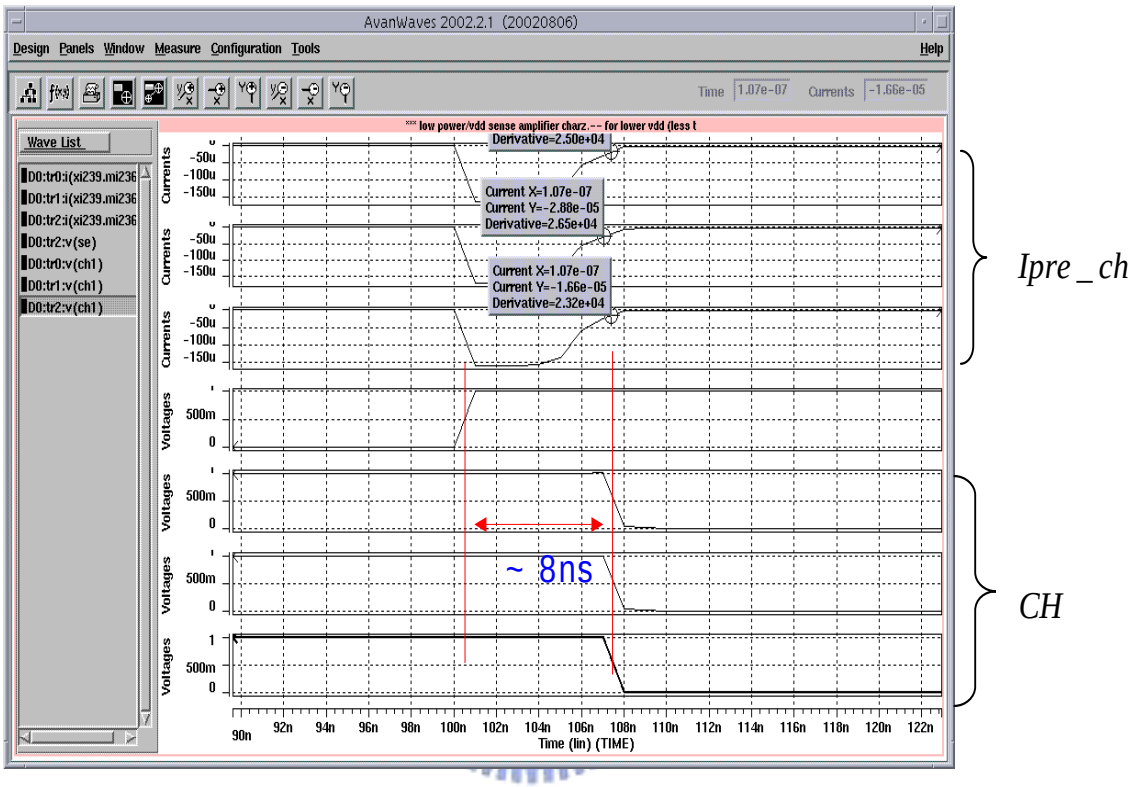


Figure 4-19: Bit line pre-charge waveform for low power supply SA

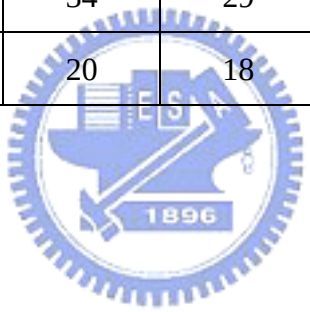
Table 4-6: Simulation result for low power supply architecture

25℃	FF	FS	SF	SS	TT
1.0v	22	33	29	37	29
1.2v	15	21	18	24	19

125℃	FF	FS	SF	SS	TT
1.0v	22	34	28	39	28
1.2v	15	21	18	25	19

-45℃	FF	FS	SF	SS	TT
1.0v	22	34	29	35	29
1.2v	14	20	18	23	18

Unit: ns



From the above table, the access speed is still within 50ns under all operating conditions. But PMOS is used for bit line pre-charge instead of 1 NMOS and 1 PMOS.

4.5 Layout

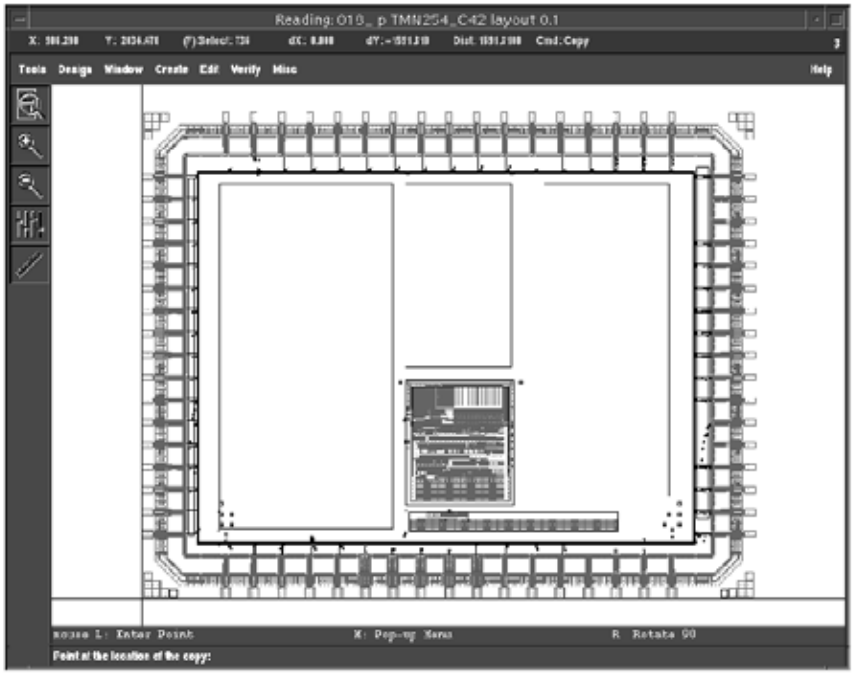


Figure 4-20: Test chip layout

Figure 4-21: Layout for the proposed sense amplifier

Chapter 5

Conclusion

The proposed sense amplifier architecture combines the advantages of low power supply and low power consumption feature in this novel-sensing scheme. The main features are as the followings:

1) Make the chip enters sleep mode after sensing operation is completed to save power and suite for low frequency application.
2) Better flash cell grounding by pulling the sensed bit line to ground for this memory architecture to improve access speed and reliability.
3) The proposed sensing scheme could operate under very low power supply (less than 1.0v) if bit line pre-charge be well taken care.
4) Bit line is pre-charge to the trip point of sensing device to save power and speed up access “1”.
5) Each sesne amplifier generates the control signals, like pre-chare pulse width and pullup/ pull down signals, automatically.

The proposed sensing scheme for low power consumption has silicon verification result, and the data shows good performance and functionality wthin operating voltage (from 1.6v to 2.0v) and temperature range (from -45℃ to 125℃). The access speed and power dissipation in “read” operation could meet application requirement in speed and power consumption.

For low power dissiaption application, the main design concept is to use the power more efficiency. First, the bit line is just pre-charge to the trip point of the

sensing device. Second, make the chip enter sleep mode after sensing operation is completed.

The proposed sense amplifier architecture could operate under very low VDD, because the architecture uses inverter as sensing device, and could use PMOS to pre-charge selected bit lines. So the limitaion for VDD is minimized.

As for low VDD sense amplifier architecture, from simulation result, it could operate under VDD=1.0v (within whole process condition and temperature range).

Reference

[1]Wikipedia,“Flash memory introduction”, http://en.wikipedia.org/wiki/Flash_memory

[2] Paolo Cappelletti et al., “Flash Memories”, Kluwer Academic Publishers, Boston, 2nd edition, Nowell, USA, p4, p242-246, p482, 2000

[3] William D. Brown,” Nonvolatile Semiconductor Memory Technology”, IEEE Press, New York, pp. 289-298, 1998

[4] Wang, ” High speed and low power sense amplifier” United States Patent: 7,057,957.June 6, 2006

[5] Datasheet of TSMC 018um Emb-flash “sfd0064_16bb”, http://ectonline.tsmc.com/j2ee/tsmcOnline”

[6] Antonino Conte et al., “A High-Performance Very Low-Voltage Current Sense Amplifier for Nonvolatile Memories”. IEEE, Journal of Solid-State Citcuit, vol.40, no. 2, 2005, pp.507-514

[7] Charles A. Deseor & Ernest S. Kuh “ Basic Circuit Theory”, McGwaw-Hill, New York, 7th edition, p35, p317, 1983

[8] Jacob Millman & Arvin Grabe “ Microelectronics”, McGwaw-Hill, New York, 2nd edition, p145, 1988

[9] Behzad Razavi “ Design of Analog CMOS Integrated Circuits”, McGwaw-Hill, New York, p17, 2001