TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Technische Elektronik

Robust Design of DRAM Core Circuits
- Yield Estimation and Analysis
by A Statistical Design Approach

Yan Li

Vollständiger Abdruck der von der Fakultät für Elektrotechnik und Informationstechnik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs

genehmigten Dissertation.

Vorsitzender: Univ.-Prof. Paolo Lugli, Ph.D.
Prüfer der Dissertation:
1. Univ.-Prof. Dr. rer. nat. Doris Schmitt-Landsiedel
2. Univ.-Prof. Dr.-Ing. Roland Thewes,
   Technische Universität Berlin

Die Dissertation wurde am 20.05.2009 bei der Technischen Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am 29.01.2010 angenommen.

Preface

The information technology (IT) has been developing at an unimaginable speed in the recent two decades and we are witnessing great changes in electronic devices and related industries such as automobile, telecommunication and computer. As the semiconductor device scaled down from several micrometers to tens of nanometers and more performance circuitries are expected to appear so as to increase the volume/bandwidth/speed of fundamental hardwares that IT industry demands, traditional circuit design methodology is no longer capable of covering all aspects required to meet the ever increasing requirements, especially some coming from statistical views. As an example, only with the aid of modern noise modeling and frequency domain system analysis can the communication circuits and systems keep up with the fast changing pace anticipated.

For high volume memories, the production yield is facing a similar challenge. On one hand, the ever increasing bit density emphasizes the significance of circuit yield. On the other hand, design for yield of memory circuits is in an embarrassing situation since until now there is no widely accepted methodology. As a consequence, more efforts have to be taken in back-end testing and measurements. In this dissertation, analytical yield analysis of DRAM core is carried out based on traditional small signal circuit analysis, probability theories and Gaussian approximations. It has superior computation speed and accuracy. In order to verify the model, it is applied to several designs and shows good agreements with silicon measurements. It is promising to pave the way for yield design, testing and optimization of other memory devices and circuits.

The text comprises 7 chapters.

Chapter 1 provides some fundamental aspects on DRAM.

Chapter 2 discusses in detail the signal amplitude loss and several interferences caused by array parasitics during sensing for different array structures. It is necessary to obtain the signal amplitude since the mean value and variance determine the yield together.

Chapter 3 introduces different sensing schemes and sense amplifiers. As the focus of this work, CMOS latched sense amplifier is highlighted and its sensing behavior, speed, power consumption and load sensitivity are studied.

Chapter 4 introduces the basics of the analytical yield analysis method with Gaussian approximations. As part of DRAM core yield design and an example of using the analytical method, the mismatch of CMOS latched sense amplifiers is modeled statistically.

Chapter 5 describes the leakage effects in DRAM cells. Some mathematic and probability transformations are made here, in order to obtain the leakage induced yield degradation.

Chapter 6 presents a new overall statistical model and an analytical expression for DRAM core yield. Based on design examples, the validity and applicability of the model are demonstrated.

Chapter 7 summarizes the work done in this dissertation.

Acknowledgment

Beginning a Ph.D career is a brand-new life experience, always full of happiness, excitement, also unexpected frustrations and disappointments. It all started with a blank page, which is eventually filled with all kinds of colorful prints with the elapse of time. For many times I wondered and puzzled in the face of setbacks and difficulties. Thanks to the help of these people, I can finally reach the goal.

The first person I appreciate is my professor, Dr. rer. nat. Doris Schmitt-Landsiedel. It is her that offered me the opportunity to study my favorite microelectronics in Germany and she took care of my research progress whenever possible. I also feel very lucky to be guided by Dr. Roland Thewes (Qimonda AG) who gave me complete freedom in planning and scheduling my topic, while providing many suggestions to my work. My advisors, Helmut Schneider (Qimonda AG) and Florian Schnabel (Qimonda AG) firstly led me into the DRAM world. They provided me DRAM knowledge and practical design considerations that are invaluable to the progress of this work. The dissertation also benefited from many individuals: Harald Roth (Qimonda AG), Michael Specht (Qimonda AG), Jürgen Lindolf (Qimonda AG), Yipin Zhang (Qimonda AG), and those who gave me opinions and suggestions. Thank you for your help.

In addition, I want to express my thanks to my colleges Marcus Weis, Philip Teichmann, Jürgen Fischer, Markus Becherer, Mohamed Abdallah, Thomas Fischer, Michael Flude, Rainer Emling, Florian Chouard, Matthias Eireiner, Andrea Merkle, Agnese Bargagli-Stoffi and those in the Lehrstuhl Technische Elektronik (LTE) who assisted me in my daily life and institute work.

Finally, I am grateful to my wife Lily and our parents. Their supports and expectations provide the most powerful strength and courage to overcome all the difficulties on the way. Without them, I will never succeed.

I thank you all.

Contents

List of Symbols

α, β, γ	constants used to define capacitance matrix $\mathbf{C}$
Δ_{Vth}	threshold voltage difference in a pair of transistors
Δ_{Vthn}, Δ_{Vthp}	threshold voltage difference of a pair of n- or p- transistors
λ	coefficient used to evaluate V_{sign} for special array patterns
$\mathbf{C}$	capacitance matrix used to evaluate bitline voltage of an array
$\mathbf{Q}$	charge matrix used to evaluate bitline voltage of an array
F	failure probability
Y, Y’	theoretical and experimental yield probability
μ_{va}, μ_{vb}, $\mu_{v'a}$	mean value of V_a, V_b and V_a'
μ	mean value of a statistical variable
σ_{Eq}, $\sigma_{Eq,m}$	standard deviation of band to band energy gap
σ_{vos}	standard deviation of V_{os}
σ_{vth}	standard deviation of threshold voltage
σ_{vthn}, σ_{vthp}	standard deviation of threshold voltage of n- or p- transistors
σ, σ_1, σ_2, σ_3	standard deviations of statistical variables
ξ	constant used to caculate sub-Vt leakage current of a transistor
a_0, a_1	linear coefficients related to a DRAM leakage source
A, A_0	small signal voltage gain
C_{bl}	bitline to ground parasitic capacitance (w/o C_{bl2bl} and C_{bl2wl} etc.)
C_{bl2bl}	bitline to bitline capacitance
C_{cpl}	coupling capacitor between bitlines
C_{dum}	dummy bitline capacitance
C_i	parasitic capacitance
C_l	load capacitance of a sense amplifier
C_s	DRAM cell capacitor
C_{bl}'	bitline capacitance per unit length
C_{bl2bl}'	bitline to bitline coupling capacitance per unit length
C_{bl2wl}'	bitline to bitline capacitance per cell
E_q, $E_{q,m}$	band to band energy gap
F	minimum feature size for a DRAM technology
g_{ds}	small signal drain source transconductance of a transistor
g_m	input transconductance of a transistor
g_{mn}, g_{mp}	input transconductances of n- and p-transistors
I_{GIDL}	gate induced drain leakage current
I_j, I_{j1}, I_{j2}	reverse biased pn junction leakage in DRAM cell
$I_{leak,m}$	average of I_{leak}
I_{leak}	total leakage current of a DRAM cell storage node
I_{sub}	sub threshold leakage current of cell transistor
k_0, k_1	coefficients to calculate V_{sign} in multiple twisted arrays
K_{cpl}	post-sensing coupling coefficient used in signal margin analysis

K_n, K_p	technology parameter for n- and p-transistors
K_t	DRAM array transfer ratio for pre-sensing
k	Boltzmann constant
L, L_n, L_p	channel length of n- or p-transistor
m	total number of bitlines in a DRAM array
n	total number of word lines in a DRAM array
Q_{array}	array consumed total charge during sensing
Q_{bl2bl}, Q_{bl2wl}	charge on bitline to bitline and bitline to wordline capacitance
Q_{cell}, Q_{bl}	charge on cell capacitor and bitline to ground capacitance
Q_{sa}	sense amplifiers consumed extra charge during sensing
q	elementary charge $1.602176487\times10-16$ coulombs
R_{bl}	parasitic bitline resistance
r_{in}	Small signal input resistance
R_{iso}	on resistance of isolation (MUX) device
r_o	small signal output resistance
R_{on}	on resistance of a transistor
R'_{bl}	parasitic bitline resistance per unit length
t_{cpl}	coupling time length
T	absolute temperature
t, Δt	time and time interval
V_a, V_b	developed signal amplitude of bitline pairs
V_{BL}	bitline voltage
V_{BLb}	complementary bitline voltage
V_{BLbn}, V_{BLbm}	n_{th} and m_{th} complementary bitline voltage
V_{BLn}, V_{BLm}	n_{th} and m_{th} bitline voltage
V_{cell}	initial cell voltage before pre-sensing
V_{dd}, $V_{dd,h}$	supply voltage
V_{ds}	drain source voltage of a transistor
V_{dum}	dummy bitline voltage
V_{eq}	equalization voltage before pre-sensing
V_{err}, V_{err1}	Voltage settling error in a charge redistribution process
V_{gd}	gate drain voltage of a transistor
V_{gs}	gate source voltage of a transistor
V_{os}	mismatch equivalent input offset voltage of a sense amplifier
V_{sign}	developed signal difference during pre-sensing
V_T	thermal voltage potential equal to kT/q
V_{th}	threshold voltage of transistors
V_{thn}, V_{thp}	threshold voltage of n- and p-transistor
V'_a	effective input voltage of a SA in consideration of post-sensing coupling
V'_{sign}	effective input voltage difference for sense amplifier
W, W_n, W_p	channel width of n- and p- transistors

Conventions

In general, symbols that begin with capital letters refer to a DC value while small letters indicate variables as small-signal values.

Since DRAM is fabricated in Metal-Oxide-Seminconductor Field-Effect Transistor (MOSFET) technology, if not specified in this thesis 'transistor' is designated to such device. As process technologies approach feature sizes below 100nm, precise high-order models of the transistor's I-V characteristics are mandatory [1, 2, 3, 4]. Unfortunately, they are too complex for hand calculations and circuit modeling. To find an analytical solution which predicts the circuit behavior to first-order, this thesis assumes the well-known basic MOS transistor model as follows.

n-channel transistor:

$$I_{ds} = \begin{cases} \dfrac{\beta_n}{2}\left(V_{gs} - V_{thn}\right)^2 & V_{ds} < V_{gs} - V_{thn} \quad \text{(saturation region)} \\ \beta_n \left(V_{gs} - V_{thn} - \dfrac{V_{ds}}{2}\right) V_{ds} & V_{ds} > V_{gs} - V_{thn} \quad \text{(triode, linear region)} \end{cases} \tag{0.1}$$

p-channel transistor:

$$I_{D} = \begin{cases} \dfrac{\beta_p}{2}\left(V_{sg} - V_{thp}\right)^2 & V_{sd} < V_{sg} - V_{thp} \quad \text{(saturation region)} \\ \beta_p \left(V_{sg} - V_{thp} - \dfrac{V_{sd}}{2}\right) V_{sd} & V_{sd} > V_{sg} - V_{thp} \quad \text{(triode, linear region)} \end{cases} \tag{0.2}$$

It has to be noted that in any case the threshold voltages are considered to be positive. The transconductance parameters β_n, β_p in above equations have the form

$$\beta_n = \frac{W_n}{L_n} K_n = \frac{W_n}{L_n}\mu_n C'_{ox}, \text{ and } \beta_p = \frac{W_p}{L_p} K_p = \frac{W_p}{L_p}\mu_p C'_{ox} \tag{0.3}$$

where W_n/L_n, W_p/L_p are design dependent width over length ratios of transistors and K_n, K_p are technology parameters, consisting of the product of the gate-channel capacitance C'_{ox} per unit area and the carrier mobilities μ_n, μ_p.

Chapter 1

DRAM Fundamentals

1.1 Dynamic Random Access Memory (DRAM)

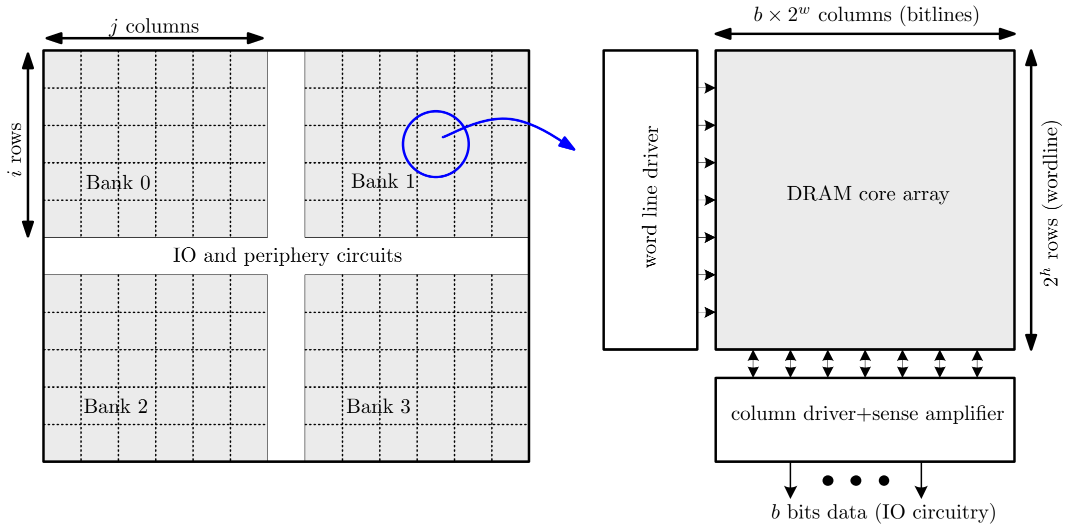
Dynamic Random Access Memories (DRAMs) are widely used in all kinds of electronic devices due to their low cost and fast operating speed. Besides its extensive applications, mass production of DRAMs usually marks the maturity of the corresponding semiconductor technology that is continuously driven to smaller dimensions and high yield by market requirements and competitions. In general, mass production of a novel DRAM generation is usually regarded as the milestone of a new semiconductor technology era.

As the term 'DRAM' implies, it is a kind of volatile memory, i.e. the data stored has to be 'dynamically' refreshed to guarantee correct memory function; Each bit in the DRAM can be accessed 'randomly' in comparison with a conventional tape recorder. References [5, 6] give the history and evolution of DRAM products in detail.

1.2 DRAM Chip Overview

A generic 512Mb DRAM block diagram comprising four banks, IO circuitry and other periphery circuits is exhibited in **Fig. 1.1** (left). Banks in DRAM are memory blocks that can accomplish most basic operations. They are usually operated individually but share the same IOs and peripheries. A DRAM chip may contain 2, 4 or even 8 banks, depending on specification, technology and design. A bank usually comprises $i \times j$ sub-arrays, and these sub-arrays are the most basic and important building blocks for a DRAM chip.

A sub-array diagram is also drawn in **Fig. 1.1** (right). It consists of three different function regions: The core array, which is built up by tens of thousands



j columns
i rows
Bank 0
Bank 1
IO and periphery circuits
Bank 2
Bank 3
$b \times 2^w$ columns (bitlines)
word line driver
DRAM core array
2^h rows (wordline)
column driver+sense amplifier
b bits data (IO circuitry)

Fig. 1.1: DRAM chip (left) and organization of core array (right)

of DRAM unit cells and numerous wordlines, bitlines, occupies the largest area; To the left of the core array are the word line driver, which control the wordlines in the core array; The sense amplifiers and columne selector drivers are placed to the bottom of the core array so as to connect bitlines that are perpendicular to the wordlines, to the following data bus and select the desired data for IOs.

When an active command comes, address bits are fed into the row column decoders, being turned into 2^h wordline driving signals and 2^w column select signals. As soon as one of the wordline goes high, its connected DRAM cells are switched on. Then sense amplifiers are turned on, pushing or pulling bitlines in the core array to V_{dd} or V_{ss} according to the cell data. Usually b sense amplifiers are under the control of one column select signal. When the column select is enabled, b bits are read out simultaneously to the succeeding circuits.

The DRAM core array is made up of tens of thousands of basic cells. A small piece of the core array is sketched in **Fig. 1.2**. Since the cells are placed regularly, the cell size can be explicitly expressed by the minimum feature size of the corresponding DRAM technology. As an example, in **Fig. 1.2** the width, spacing of the bitlines(BL) and the wordlines(WL) are all $1F$. By observation a single DRAM cell in the array occupies $2F \times 4F = 8F^2$ unit area, and therefore the array is called $8F^2$ DRAM cell array. Obviously when the semiconductor technology shrinks, the size of the cell goes down in proportional to the square of feature size and the die cost per bit drops more rapidly than the cost of the conventional production circuits, in which some circuits can not be redimensioned due to performance, noise or mismatch limitations.

As shown on the right of **Fig. 1.2**, a transistor and a capacitor in series connection form a DRAM cell. One plate of the capacitor is biased with a fixed

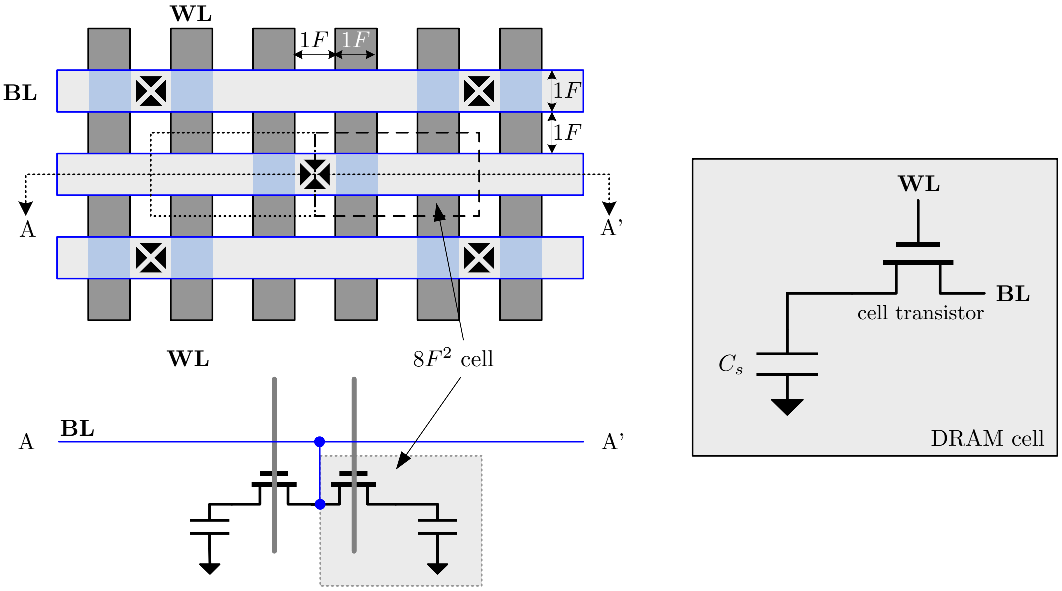


Fig. 1.2: Principle structure of a $8F^2$ DRAM core array

voltage potential that is around $(V_{dd} - V_{ss})/2$ and another plate connecting the cell transistor becomes the 'storage node'. When the cell transistor is 'on', the cell capacitor is charged through the bitline to V_{dd} or V_{ss}. Then the cell transistor is turned off to isolate the storage node from the bitline. The memory function is realized by storing positive or negative charge in the cell capacitor. Next time when the cell is accessed, its charge will be released and transferred to the bitline linked to the cell. The change in the charge of the bitline capacitance will give rise to a change of the bitline voltage that can be utilized by sense amplifiers. The amplified voltage from the sense amplifier is then led to the IO circuitry. This process is called 'voltage sensing' and will be discussed later.

In **Fig. 1.1** the data from the core array is supposed to be transferred to the IO circuitry directly. However, this is impossible in a practical design because of the long wire length from the accessed core array to the IO circuitry. This long wire introduces a large delay that may not be tolerable in meeting timing specifications. For example, the extreme case comes from the core array located at the top-left corner of the chip in **Fig. 1.1**: the minimum distance from the border of this array to the chip center where the IO circuitry is positioned is $(i-1)$ times of the height of the core array, and therefore its wire load is relatively high. To solve such problem, hierarchical array structures with multi-stage sense amplifiers can be used especially for high volume DRAM products [7, 8, 9]. **Fig. 1.3** shows an exemplary structure of a two sense amplifier stages design. In this structure, the first sense amplifier stage is shared by the left and right side core arrays by alternately switching on sw_0 and sw_1. Switches $cs_0 \ldots cs_k$ are controlled by column select wires to connect one local bitline pair to the global bitline pair.

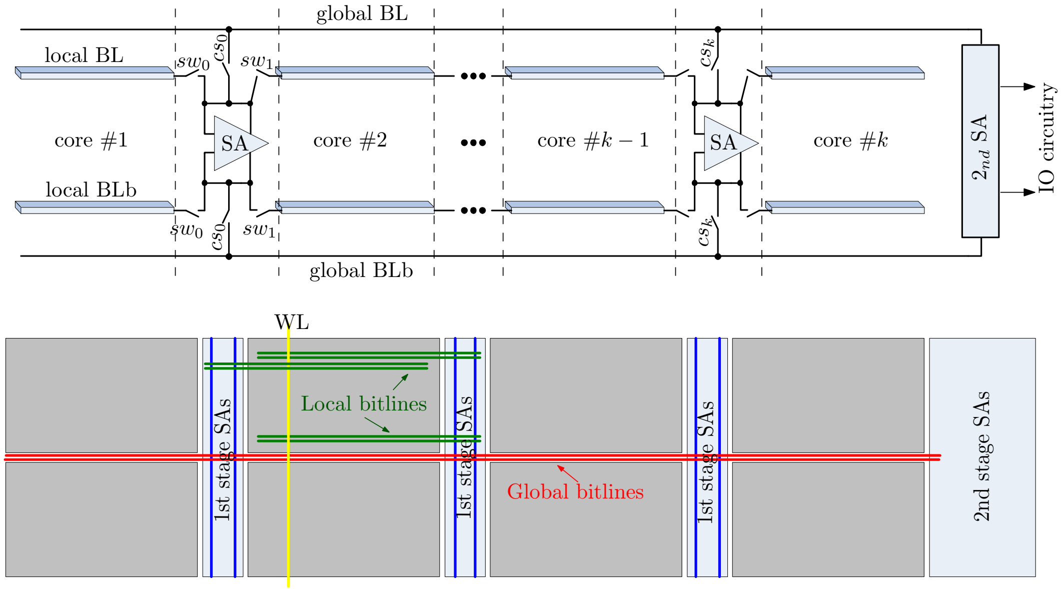


Fig. 1.3: A hierarchical array structure [7]: schematic (top) and physical floor-plan (bottom)

When the second stage sense amplifier is switched on, the voltage difference of the global bitline pair will be amplified and driven to the succeeding IO circuitry. As the second stage can occupy much larger area, they can be designed much stronger and faster in contrast with the area limited first sense amplifier stages. As a consequence, the heavely loaded global bitline can be driven rapidly, and even the core far from the IO circuitry will show similar sensing delay to that of the core array closer to the IOs. Besides the multi-stage data bus, hierarchical architecture is also applied to row and column select drivers coming from another side of the bank so as to improve array access speed.

1.3 Basic Operations

Like any other storage devices, read and write operations are necessary for DRAM. In a typical read operation, a row address is first sent to the row decoder, generating driving signals for the wordline drivers. When one of the wordlines is driven high and turns on the DRAM cells it connected, the cell charge will flow to the bitline capacitances, forming small voltage changes ranging typically from 100mV to 200mV. Then the sense amplifiers are switched on to amplify the bitline voltage change and recover the stored data. Since one wordline controls multiple cells linked to the corresponding sense amplifiers, there will be hundreds of data available at the same time. These data form one 'page' that can be fast accessed by the succeeding column selections. The column select signals turn on the switches between the desired local bitline pairs and the global bitline pairs. As soon as the global bitline pairs have enough voltage swing the secondary sense amplifiers will be activated, amplifying the voltage on global bitlines and delivering the outcomes to the IO circuits. At the end of the read operation the wordline is switched off to isolate the DRAM cells from bitline again.

Similarly, a write operation also starts by giving a row address to the row decoder and driving the corresponding wordline. After that, the column decoder connects the global bitlines to the desired local bitlines. The "secondary sense amplifers"are then switched on to drive the global bitline pairs. With the help of the first stage sense amplifier, V_{dd} or 0 corresponding to solid '1' or '0' data will be written into the cells.

Besides read and write operations, a refresh operation is required to compensate for the leakage current that reduces the cell voltages gradually with the elapse of time. The difference of a refresh process from read operation is that the column access is not necessary in refresh process because activations of the row selection and the first stage sense amplifiers are enough to accomplish the cell voltage recovery task. For a more detailed description of DRAM operations please refer to [10].

To ensure that the data stored in the DRAM cells can be accessed correctly, proper sensing is necessary and for DRAM core it is the most important procedure in consideration of production yield.

1.4 Sensing Methods

Sensing and amplification are mandatory for all kinds of memory circuits and sensors because of the tiny acquirable signal amplitude caused by lossy transfer, leakage, parasitics, circuit noise and device variations. Generally, sensing circuits can fully recover the deteriorated signals and transmit them to succeeding circuits that usually handle large amplitude digital signals. According to the properties of input and output signals during sensing, sensing can be categorized into voltage sensing, current sensing and charge sensing.

1.4.1 Voltage sensing

Evidently, the input and output of voltage sensing are voltage signals. The circuits for implementing voltage amplification are actually voltage amplifiers or comparators. Ideally, the voltage sensing output, which could be '0' or '1', is only determined by the polarity of the input voltage.

A voltage sensing process in DRAM core is illustrated in **Fig. 1.4**. It includes

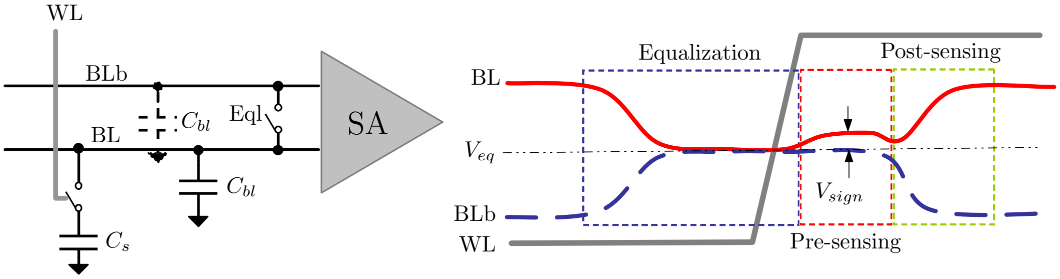


Fig. 1.4: A DRAM voltage sensing scheme can be divided into three procedures: equalization, passive pre-sensing and post-sensing.

three important time segments: equalization, pre-sensing and post-sensing. Equalization equals the voltage of the bitline pair to V_{eq} by connecting Eql to high. A row access is started by driving Eql to low again to separate the two bitlines in a bitline pair. A wordline is then turned on to access the connected cell. This behavior triggers the pre-sensing: due to the voltage difference between the cell storage node and its corresponding bitline voltage, the cell charge is shared with the bitline. This process makes the bitline voltage either lower or higher than another bitline staying at V_{eq}, forming a voltage difference V_{sign} in the bitline pair as shown in **Fig. 1.4**. By charge conservation it is known

$$V_{sign} = \frac{C_s}{C_s + C_{bl}} \cdot (V_{cell} - V_{eq}) \tag{1.1}$$

where V_{cell}, C_s, C_{bl} and V_{eq} are the initial cell voltage, cell capacitor, bitline parasitic capacitance and equalization voltage, respectively, and V_{cell} ideally exhibits two levels - V_{dd} for solid '1' and 0 for solid '0'.

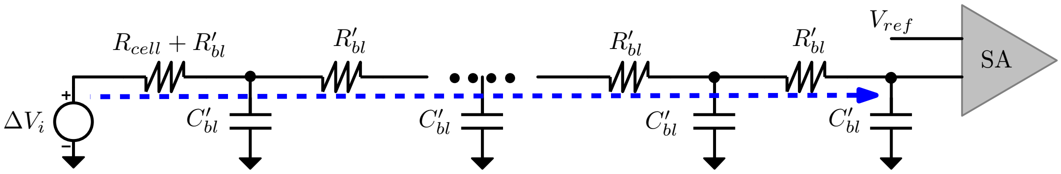


Fig. 1.5: The output voltage is delayed by the parasitic RC network in a voltage sensing process.

Voltage sensing is widely accepted in DRAM circuits not only for the first stage sense amplifiers but also the secondary sense amplifiers as shown in the hierarchical data bus design in **Fig. 1.3**. However, the passive pre-sensing process gives rise to speed penalty when the bitlines are heavily loaded as demonstrated in **Fig. 1.5**. Here, a long bitline is modeled as parasitic RC network where R'_{bl}, C'_{bl} represent bitline unit resistance and capacitance. The DRAM cell is modeled as a voltage source with series resistance. In case the cell suffers from larger cell resistance R_{cell}, e.g., the on resistance of the cell transistor or technology related parasitic series resistance, a large voltage propagation delay will appear. This effect is more especially severe for the cells located at the far-end of the bitline. Therefore, the speed of voltage sensing will be greatly limited by the large cell resistance R_{cell}, the bitline total parasitic resistance R_{bl} and capacitance C_{bl} because of the parameters related charge transfer process needed to set up the necessary input voltage difference V_{sign} for sense amplifier. By detailed analysis in Section 2.1 the time to set up V_{sign} follows

$$t_{v,sens} \propto (R_{cell} + R_{bl}) \cdot \frac{C_{bl}C_s}{C_{bl} + C_s} \tag{1.2}$$

It confirms the slow response for voltage sensing when parasitic parameters are relatively large.

1.4.2 Current sensing

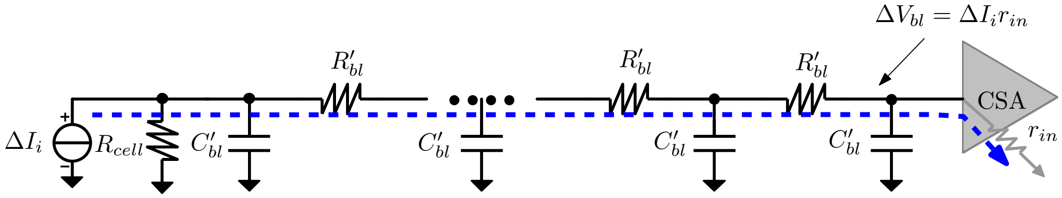


Fig. 1.6: The RC network has no delay effect in current sensing.

When the bitline voltage experiences little change during sensing, the delay caused by charging/discharging the bitline parasitic capacitance will be significantly smaller. Current sensing is an implementation of such an idea. Assume a cell can be replaced by a current source and a resistor in parallel as shown in **Fig. 1.6**. When an impedance that is much smaller than R_{cell} appears near the sense amplifier and R_{bl} is much smaller than r_{in}, the input current step induced bitline voltage change will be small because $\Delta V_{bl} \approx I \cdot r_{in}$, where r_{in} is the low impedance at far-end. A current sense amplifier (CSA) can provide such low input impedance. It converts the input current into a voltage at the output. Since the bitline voltage doesn't change much during current sensing, current sensing has speed advantage over voltage sensing in DRAM cores with large bitline parasitic capacitance and cell resistance. An example of building a current sense amplifier is shown in **Fig. 1.7**. The gain block in the figure is an ideal voltage amplifier with gain A. The input impedance in such topology will drop to $R_p/(1 + A)$. If R_p is equal to R_{cell}, the time delay of the current sensing approximates to

$$t_{c,sens} \approx \frac{R_{cell}}{A + 1} C_{bl} \tag{1.3}$$

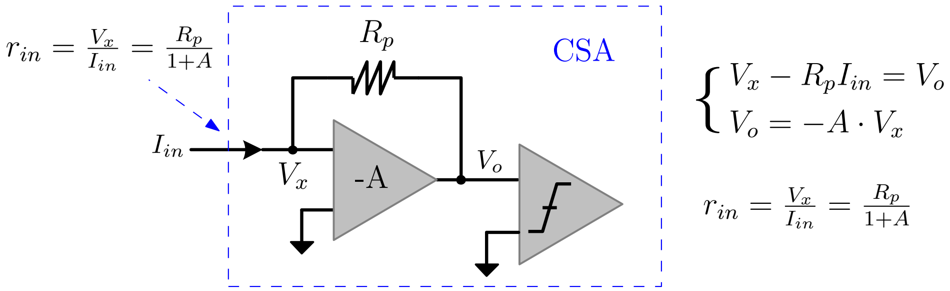

$r_{in} = \frac{V_x}{I_{in}} = \frac{R_p}{1+A}$
R_p
CSA
I_{in}
V_x
-A
V_o
$\begin{cases} V_x - R_p I_{in} = V_o \\ V_o = -A \cdot V_x \end{cases}$
$r_{in} = \frac{V_x}{I_{in}} = \frac{R_p}{1+A}$

Fig. 1.7: Implementation of a current sense amplifier by using a voltage amplifier with gain -A.

which is nearly hundred times smaller than voltage sensing when the delay from I_{in} to V_o is negligible.

However, the presumption that a cell can provide constant current is not true for DRAM. In addition, R_{bl}, which becomes more significant as minimum feature size shrinks, was assumed to be very small in the above evaluation. Consequently, current sensing can never be applied to the first stage sense amplifiers, but they are good candidates for secondary sense amplifiers with the hierarchical data bus as shown in **Fig. 1.3** because first stage sense amplifiers can provide continuous current when they are enabled. It is advantageous to use current sensing for the secondary sensing stage in particular when global bitlines confront very large parasitic capacitance.

The disadvantage of current sensing resides in two aspects: Firstly, the complex current sense amplifier and bias circuitry occupy large silicon area; Secondly, power consumption is an issue because of their static bias currents.

1.4.3 Charge sensing

In analogy with current sensing, charge sensing is supposed to be able to provide less bitline voltage fluctuation. The first charge sensing method comes from the idea that if the charge in the cell capacitor can be completely transferred to another capacitor that is located at the sense amplifier, the bitline parasitic capacitance will have no effect.

Fig. 1.8 demonstrates the implementation of the proposed charge sensing scheme. When C_p is A times larger than C_{bl}, the charge deposited in C_p will be A times of the charge in C_{bl}. When A is large enough, C_{bl} can be neglected and almost all input charge ΔQ_{in} goes into C_p. The charge sense amplifier (QSA) can be implemented by using an ideal voltage amplifier. With the aid of the negative feedback, the input equivalent capacitor is $(1+A)C_p$ instead of C_p. For example, if $C_p = C_s = C_{bl}/10$ and $A = 100$, 9/10 of ΔQ_{in} will flow to C_p, resulting in a voltage change on C_p being $9(V_{cell} - V_{eq})/10$, which is approximately 9 times

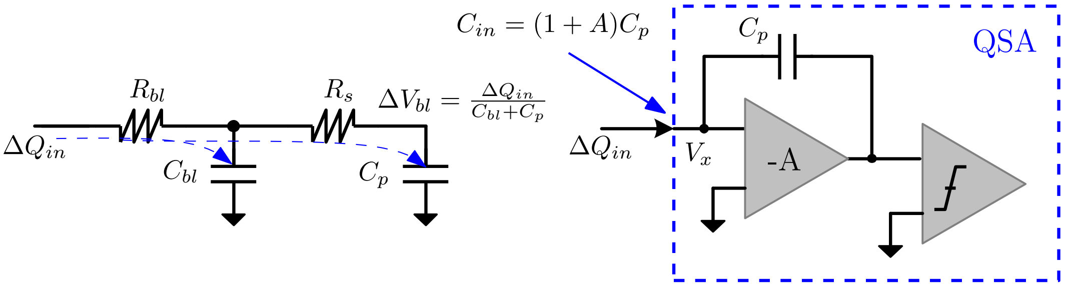


Fig. 1.8: Implementation of a charge sense amplifier by using a voltage amplifier with gain -A.

larger than V_{sign} generated by voltage sensing.

The large input capacitance, however, also generates a disadvantage for this charge sensing scheme. Due to its presence the sensing speed becomes even worse in contrast with voltage sensing. In addition, the charge sense amplifier can only be designed in Switched Capacitor (SC) circuit style that is difficult to be implemented within tight space as in the case of DRAM core.

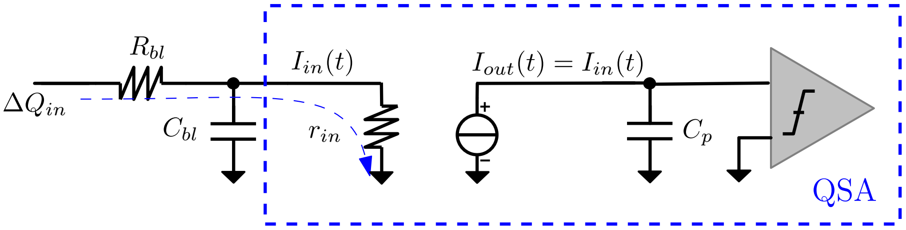


Fig. 1.9: Current sensing like charge sensing

Another charge sensing concept also called charge transfer sensing, is actually evolved from current sensing. Differently, it provides both low input impedance and zero static bias current as shown in **Fig. 1.9**. Here, the charge sense amplifier (QSA) consists of two functions: a). Its small input impedance prevents the bitline voltage fluctuation when the cell charge is released; b). The sense amplifier can copy the current through r_{in} to the output branch and as a result, the total charge in capacitor C_p will be increased or decreased accordingly. The second function actually keeps the quantity of output charge equal to the quantity of input charge. Clearly, when the total input charge coming from cell capacitor C_s goes into the output capacitor C_p and $C_s > C_p$, a much larger voltage step can be generated at the output node for the succeeding comparator. This charge sensing provides both higher speed and moderate power consumption, and thus is popular in high-performance DRAM circuits. The difficulties reside in designing and biasing such sense amplifiers as will be discussed later in Section 3.4.

Tab.1.1 summarizes the four different sensing methods and their advantages, disadvantages.

Table 1.1: Comparisons between voltage, current and charge sensing

Sensing method	Area	Power	Speed	Precision[3]	Circuit complexity
Voltage	smallest	smallest	moderate	moderate	simple
Current	large	largest	fastest	high	complex
Charge A[1]	largest	large	slowest	highest	complex
Charge B[2]	larger	moderate	fast	high	complex

[1] Charge sensing with capacitive input impedance

[2] Charge sensing with resistive input impedance or charge transfer sensing

[3] Precision is determined by the achievable voltage amplitude in pre-sensing

1.5 Sensing Failures and Yield

Simply speaking, a sensing fails when the sensing outcome is different from the data originally written into the cell. In DRAM the most significant failure sources include large cell leakage, inapropriate cell to bitline capacitance ratio C_s/C_{bl}, severe array post-sensing coupling and large sensing transistor threshold mismatch. Since data is stored in the form of charge in a cell capacitor, leakage currents at the storage node will reduce the charge quantity within a certain period of time, and therefore the available voltage difference V_{sign} when the cell is accessed. As V_{sign} is also a function of the capacitor ratio C_s/C_{bl} as shown in Eqn. (1.1), a smaller capacitor ratio results in a decreased V_{sign} as well. When sense amplifiers are switched on, the inter-bitline crosstalk coupling can destroy a sensing if V_{sign} to be sensed is much smaller compared to V_{sign} of the neighboring bitline pairs. Nonetheless, even if there is no coupling, threshold mismatch of sensing transistors can completely flip the sensing outcome when the mismatch equivalent input offset V_{os} is a little larger than V_{sign}. These sources altogether degrade the yield performance of a DRAM core by a certain degree.

1.5.1 Yield experiments

Since above failure sources are random variables, it is impossible to know exactly when and where a failure takes place. However, statistically the failure probability can be obtained by carrying out a large number of experiments and taking the percentage of fail/pass outcomes from the total number of experiments as a representative of the theoretical yield / failure probability. Since in a yield experiment the outcome is either 1 (pass) or 0 (fail), such an experiment is actually one *Bernoulli Trial* [11] with probability of Y for pass and 1-Y for fail in statistics. Y is the theoretical yield in the experiments. When the total number of experiments is n, the Failure Count (FC) is obtained by

$$\text{FC} = n - \sum_{i=1}^{n} X_i, \tag{1.4}$$

where X_i is either 1 or 0, representing the outcome of each Bernoulli trial. Usually the experimental yield Y' and failure F' is obtained by the equation

$$\text{Y'} = \frac{\sum_{i=1}^{n} X_i}{n} = 1 - \text{F'} \tag{1.5}$$

By *Law of Large Numbers* [11] with sufficient samples the experimental yield Y' and theoretical yield Y have the relationship

$$P(|\text{Y'} - \text{Y}| \geq \epsilon) \to 0, \text{ when } n \to \infty \tag{1.6}$$

where ϵ is any positive real numbers. In another words, (1.6) assures that the theoretical yield Y can be replaced by the experimental yield Y' only when the total number of experiments n approaches infinity.

1.5.2 Number of experiments

The number of experiments n is very critical in estimating yield from experiments. Because the experimental efforts are directly proportional to the number of trials n, a smaller n is always desired. The first requirement of n is the yield resolution. If n experiments are carried out, the minimum achievable yield resolution Y_{res} is

$$\text{Y}_{res} = \frac{1}{n} \tag{1.7}$$

Eqn. (1.7) implies that if $\text{Y}_{res} = 1\text{ppm}$ is expected to be discriminated, the minimum required $n = 10^6$. According to Eqn. (1.7),

$$n > \text{Ceiling}(\frac{1}{\text{Y}_{res}}), \tag{1.8}$$

where ceiling is a function used to map a real number to its next higher integer.

The second requirement comes from the confidence level of the experimental yield Y'. By *Chebyshev Inequality* [11], the experimental yield Y' and theoretical yield Y have to follow

$$P(|\text{Y'} - \text{Y}| \geq \epsilon) \leq \frac{\sigma^2_{\text{Y'}}}{\epsilon^2} \tag{1.9}$$

where $\sigma^2_{\text{Y'}}$ is the variance of experimental yield Y'. As Y' $= (\sum_{i=1}^{n} X_i)/n$ and X_i follows independent *Bernoulli Trial* with probability Y for pass and 1-Y for fail, Y' will have expectation Y and variance $\mathrm{Y}(1-\mathrm{Y})/n$. By putting the variance into Eqn. (1.9), the following equation is obtained

$$P(|\mathrm{Y'}-\mathrm{Y}|<\epsilon)=1-P(|\mathrm{Y'}-\mathrm{Y}|\geq\epsilon)\geq 1-\frac{\mathrm{Y}(1-\mathrm{Y})}{n\epsilon^2} \tag{1.10}$$

Eqn. (1.10) implies that the probability of the difference between experimental yield Y' and theoretical yield Y staying within distance ϵ is greater than $1-\mathrm{Y}(1-\mathrm{Y})/(n\epsilon^2)$. Therefore, ϵ is a number expressing the similarity between Y' and Y. If ϵ is small enough and the probability in Eqn. (1.10) is greater than L, Y' is regarded as substitute of Y. L is called *confidence level* with typical value over 90%. From above descriptions, the number of experiments needed to meet the confidence level L within ϵ is

$$n>\frac{\mathrm{Y}(1-\mathrm{Y})}{(1-L)\epsilon^2} \tag{1.11}$$

Example
If the experimental yield Y' needs to be as precise as $\mathrm{Y}\pm 10^{-m}$, the required ϵ should be smaller than $10^{-m}/2$. Suppose the confidence level $L=90\%$. The required number of experiments n from Eqn. (1.11) must be greater than $4\mathrm{Y}(1-\mathrm{Y})10^{2m}$. Given at least one digital precision is required ($m=1$) and Y is expected to be in the range from 10% to 40%. Because $Y(1-Y)$ has its maximum at Y$=0.5$ and $\mathrm{Y}(1-\mathrm{Y})$ is a monotonic decreasing function when $0.1<\mathrm{Y}<0.4$, Y$=50\%$ is used to calculate the required n, which is $4\times 50\%\times(1-50\%)\times 10^2=100$. As a result, with $n>100$ the theoretical yield Y is with certainty of 90% in the region $[\mathrm{Y'}-0.05,\mathrm{Y'}+0.05]$.

The above example shows that the confidence level condition is more strict than the yield resolution requirement that needs $n>10$.

1.5.3 Confidence region

In most yield experiments, the number of experiments n is a constant for convenience. For a constant confidence level L, this leads to the change of confidence region ϵ when the theoretical yield Y varies. From Eqn. (1.11) the confidence region

$$\epsilon>\sqrt{\frac{\mathrm{Y}(1-\mathrm{Y})}{(1-L)n}} \tag{1.12}$$

Since $\mathrm{Y}(1-\mathrm{Y})$ shows a peak at Y$=0.5$, the confidence region will show a trend as shown in **Fig. 1.10**. As the confidence region becomes wider around Y$=0.5$,

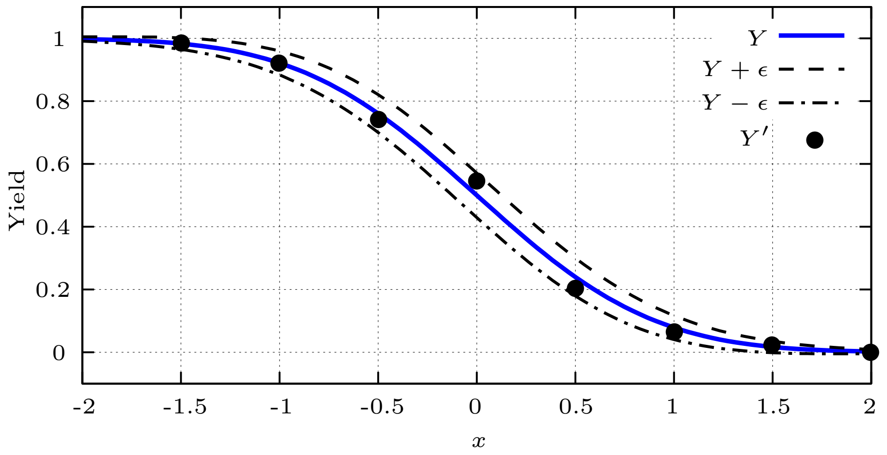


Fig. 1.10: The relationship between experimental yield Y', theoretical yield Y and confidence region ϵ (x is a random variable and yield changes with x)

the experimental yield Y' can show larger fluctuations due to more deviations from the theoretical yield Y. Sometimes, people think law of large number is no longer valid due to frequently observed experimental yield fluctuations. However, there is actually no contradiction due to the wider confidence region around 50% yield region.

As study of DRAM core yield is the main issue in this thesis, the statistical constraints and theories mentioned above will play a very important role in verifying the validity of experimental yield obtained from either silicon measurements or *Monte Carlo* (MC) [12] simulations.

1.6 Motivation and Challenges

For many years, the semiconductor technology has been following *Moore's Law* [13, 14, 15], which stated in 1975 that MOS device dimensions would continue to scale down by a factor of two every three years and the number of transistors per chip would double every one to two years. As a concrete example of this law, the volume of DRAM products increases two times and the chip size increases 1.4 times per year [16]. The consequences of downscaling technology for DRAM products give rise to a number of yield related issues coming from:

1. **Supply voltage**: To maintain device reliability, the supply voltage has to be scaled down accordingly. This reduces the available voltage amplitude stored in cell capacitors from generation to generation.

2. **Array parasitics**: They become more significant for deep sub-micrometer structures and devices. Both sensing speed and achievable signal amplitude

are greatly affected by array parasitics. In particular, the parasitic bitline to bitline capacitance introduces additional crosstalk noise during post-sensing, generating more failures.

3. **Cell leakage**: As the cell density becomes higher and junctions become smaller while supply voltage can not drop more, cell leakage is increasingly difficult to control. Thus the sensing yield within data retention time is degraded.

4. **Sense amplifier mismatch**: As device dimensions are scaled down, the corresponding mismatch can be even larger than in previous technologies.

Altogether, the production yield of DRAM becomes quite poor at the beginning of every generation for new technologies and it usually takes time for the yield to ramp up. In order to speed up the yield ramp-up process, an effective and reliable yield model for DRAM core is mandatory. In this thesis, the main contributions include:

1. Array structures and the parasitic effects are thoroughly investigated in Chapter 2. As voltage sensing is the most practical sensing method for DRAM core, voltage difference V_{sign} during pre-sensing in the voltage sensing scheme is precisely formulated by a charge conservation model considering array parasitics for a variety of arrays in Section 2.2. Additionally, post-sensing crosstalk coupling is modeled and analyzed with small signal circuits and differential equations, and a worst case design methodology is proposed based on the outcomes of the crosstalk model in consideration of both sense amplifier mismatch and post-sensing coupling in Section 2.3.

2. DRAM sense amplifiers and sensing techniques are compared in Chapter 3. A detailed analysis on operations of CMOS latched sense amplifiers in voltage sensing scheme is conducted in Section 3.3. It will be exhibited that simultaneously latched CMOS sensing has more advantages over other sense amplifiers and provides the best balance between yield and other design specifications.

3. Mismatch inside latched sense amplifiers is modeled as an input offset V_{os} by analytical yield analysis and a small signal circuits model in Chapter 4. The statistical characteristics of V_{os} are discussed and evaluated under different conditions. Yield optimization process for a CMOS latched sense amplifier is finally proposed in Section 4.3. The results show that simultaneously latched CMOS sense amplifiers with mid-level sensing are superior to other voltage sense amplifiers in yield performance.

4. A nonlinear yield analysis is applied to leakage induced yield degradation in DRAM core in Chapter 5. Both yield of signal margin plots and of

data retention time plots are obtained from the model and compared to Monte-Carlo simulations and measures. They approve the presumption in the model that the original variabilities causing the leakage fluctuations are Gaussian distributed and the leakages in DRAM cells follow Log-Norm distributions. The model provides a guide for design of DRAM core from both technology and economic aspects.

5. In Chapter 6 a linear yield model is proposed based on knowledge and outcomes gained from Chapters 2-5. By taking most important random variabilities and post-sensing crosstalk coupling into consideration, it becomes a powerful yield analysis and optimization tool for development of any new DRAM technology, core array structure and sensing circuits with less computation and time efforts. Needless to say, with the help of the model DRAM core design will be faster, cost effective and yield oriented.

Chapter 2

DRAM Core Array

2.1 DRAM Cell and Pre-sensing

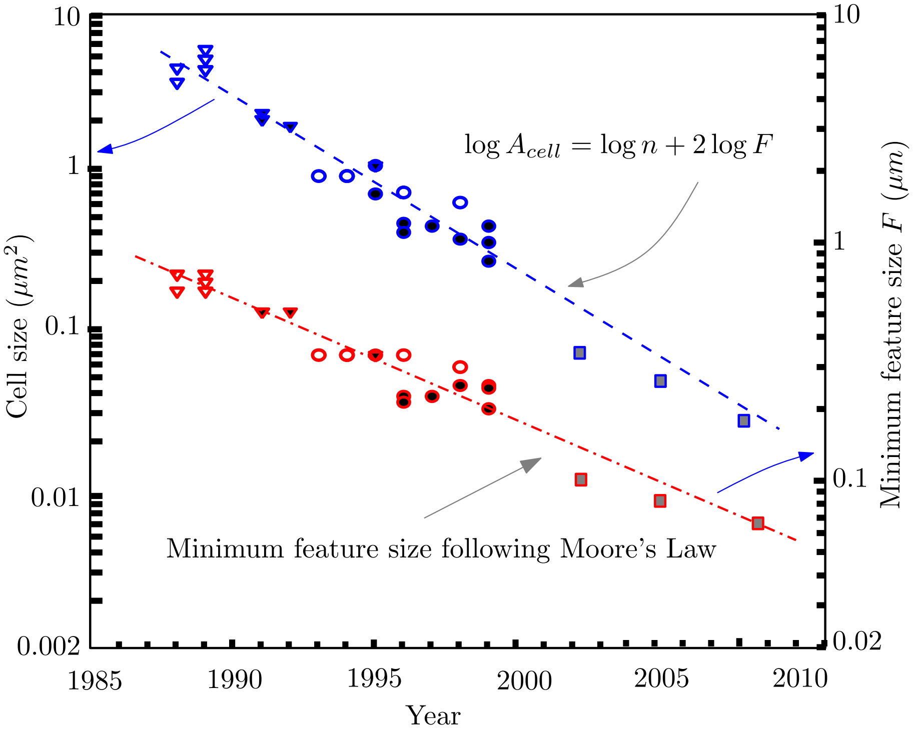


Fig. 2.1: Cell size and minimum feature size trends of different manufacturers' DRAM products (modified from [16])

As described in Section 1.2, a DRAM sub-array contains several regions, of which the core array is the most important. Core array design greatly affects an array's access speed, yield and production cost. As the basic but most crucial building element of core array, DRAM cells have been experiencing great revolutions ever since the DRAM's invention. Since chip cost is approximately proportional to its die area, smaller cells are more profitable. **Fig. 2.1** demonstrates the downscaling trend of the cell area together with the corresponding minimum feature size for different manufacturers' DRAM products. Following *Moore's Law* [13, 14, 15], the semiconductor feature size drops exponentially with the elapse of time. Since cell size is a square function of its minimum feature size F, it can be expressed as

$$A_{cell} = nF^2, \text{ or } \log A_{cell} = \log n + 2\log F \tag{2.1}$$

n in Eqn. (2.1) is an integer because cell width and length are usually multiple times of the minimum feature size as illustrated in **Fig. 1.2**. In addition, n itself reduces as well. As an example, the cell size of mainstream products transits from $12F^2$ to $8F^2$ around the year 2000, and then to $6F^2$ around the year 2007. It was predicted that the ultimate $4F^2$ cell will come into play around the year 2010 and marks the end of planar DRAM technology. However, as $\log n$ changes only a little for n =12, 8, 6, 4, all the products from different DRAM makers are placed close to the straight line, which is $2\log F$, as shown in **Fig. 2.1**.

2.1.1 Pre-sensing speed

No matter how the DRAM cell downscales, its electric characteristic and structure remain basically the same. As shown on the left of **Fig. 2.2**, a cell is made up of a transistor and a capacitor. Here, the related bitline is modeled by a lumped RC circuit.

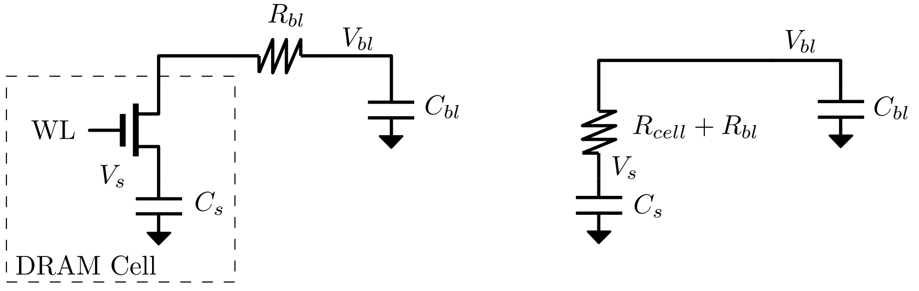


Fig. 2.2: DRAM cell structure (left) and its corresponding RC model (right) together with a lumped bitline model.

In a voltage sensing process, when the cell in **Fig. 2.2** is accessed, the transistor is first switched on to share cell charge with bitline capacitance. Consequently, the bitline voltage rises or falls according to the cell charge polarity, and a small voltage step V_{sign} is generated as shown in **Fig. 2.3**. By Kirchhoff's Laws the

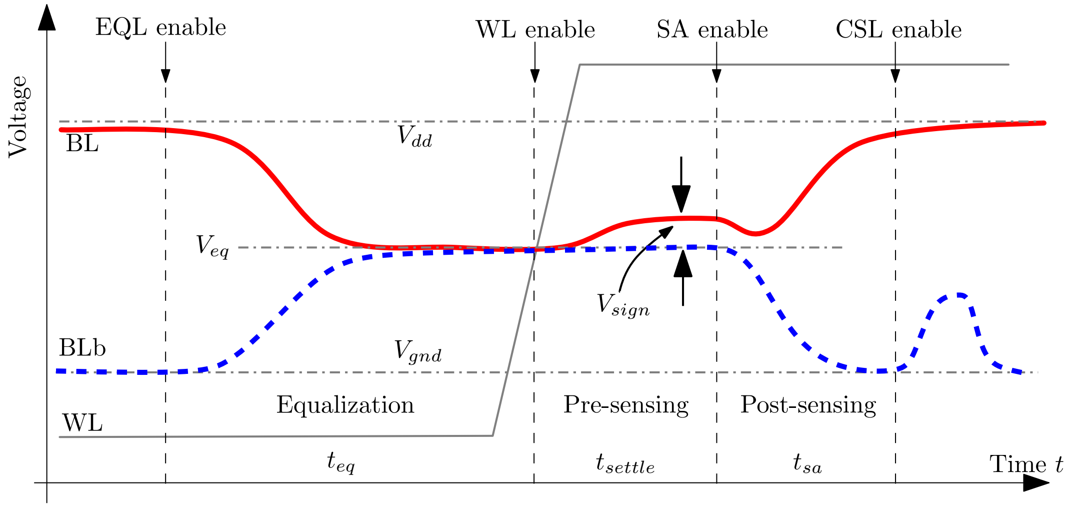


Fig. 2.3: The voltage sensing process in DRAM core

differential equations describing the pre-sensing process can be formulated as follows:

$$\begin{cases} C_s \cdot dV_s/dt + C_{bl} \cdot dV_{bl}/dt = 0 \\ |V_{bl} - V_s| = (R_{cell} + R_{bl}) \cdot C_s \cdot dV_s/dt \\ V_s(0) = V_{cell},\, V_{bl}(0) = V_{eq} \end{cases} \tag{2.2}$$

where V_s, V_{bl}, V_{cell}, $V_{bl}(0)$,V_{eq} are the voltages of the cell storage node, bitline, initial cell storage node, initial bitline and equalization, respectively. R_{cell} is the cell series resistance comprising cell transistor on resistance R_{on} and cell parasitic resistance R_{par}. Since the cell transistor in 'on' state works in linear region,

$$R_{on} = \frac{1}{K_n \frac{W_n}{L_n}(V_{gs} - V_{th})} \tag{2.3}$$

For most planar DRAM cells with $W_n/L_n \approx 1$ and $K_n = 50 - 200\mu A/V^2$, R_{on} ranges from $10k\Omega$ to $2.5k\Omega$ when $(V_{gs} - V_{th}) = 2\text{V}$. Additionally, due to the different cell structures R_{par} varies in a wide range, and is even larger than R_{on} for a trench capacitor cell due to its collar region [17].

By solving Eqn. (2.2), the storage node to bitline voltage difference $V_s - V_{bl}$ at a given time t gives

$$V_s(t) - V_{bl}(t) = [V_{cell} - V_{eq}] \cdot e^{-\frac{1/C_s + 1/C_{bl}}{R_{cell} + R_{bl}} t} \tag{2.4}$$

It is instructive that Eqn. (2.4) implies V_s is equal to V_{bl} only if time t goes to infinity. The time t corresponding to a given settling error V_{err} from Eqn. (2.4) is

$$t_{settle} = \ln(\frac{|V_{cell} - V_{eq}|}{V_{err}}) \cdot \frac{R_{cell} + R_{bl}}{(1/C_s + 1/C_{bl})} \tag{2.5}$$

Eqn. (2.5) suggests that the pre-sensing speed is proportional to the time constant determined by placing all the resistances and capacitances in series.

Example
Assume $R_{bl} = 1k\Omega$, $R_{on} = 2.5k\Omega - 10k\Omega$, $C_s = 30$fF, $V_{err} = 1$mV, $R_{bl} = 1k\Omega$, $C_{bl} = 100$fF, $V_{cell} - V_{eq} = 0.6$V. The corresponding pre-sensing delay $t_{settle} = 664$ps-1.77ns from Eqn. (2.5).

To gain faster settling speed for the voltage sensing, the time constant $(R_{cell} + R_{bl})/(1/C_{cell} + 1/C_{bl})$ should be as small as possible. From the technology aspect, a smaller R_{on} can be obtained by increasing $V_{gs} - V_{th}$, the unit gate oxide capacitance C'_{ox} and the carrier mobility μ_n. Unfortunately, due to the constant electric field scaling rule [18], increase of $V_{gs} - V_{th}$ always trades off with decreasing the thickness of the gate oxide, making R_{on} change little from one technology to another. In addition, the retention time requirement, which claims the cell leakage current is expected to be much less than 1fA in average [16], leads to higher threshold voltage V_{th} for cell transistors. Consequently, all these conditions form a tight technology constraint on fabricating the cell transistor, keeping R_{on} relatively constant for technology generations.

2.1.2 Developed bitline voltage amplitude

Clearly from the above discussion, R_{on} of the cell transistor can not be reduced further to speed up the pre-sensing process. How about C_s, C_{bl}, R_{bl} and cell series parasitic resistance? Because of charge conservation it is known that in **Fig. 2.2** at any given time t

$$C_s \Delta V_s(t) + C_{bl} \Delta V_{bl}(t) = 0 \tag{2.6}$$

and

$$V_s(t) = V_{cell} + \Delta V_s(t),\ V_{bl}(t) = V_{eq} + \Delta V_{bl}(t) \tag{2.7}$$

By taking the charge conservation equations into Eqn. (2.4), the transient response of $V_{bl}(t)$ is obtained as

$$V_{bl}(t) = V_{eq} + \frac{C_s}{C_s + C_{bl}}(V_{cell} - V_{eq})[1 - e^{-\frac{1/C_s+1/C_{bl}}{R_{cell}+R_{bl}}t}] \tag{2.8}$$

From Eqn. (2.8) the maximum voltage amplitude that the bitline can develop when t goes to infinity is

$$V_{sign,max} = V_{bl} - V_{eq} = \frac{C_s}{C_s + C_{bl}}(V_{cell} - V_{eq}) = K_t(V_{cell} - V_{eq}) \tag{2.9}$$

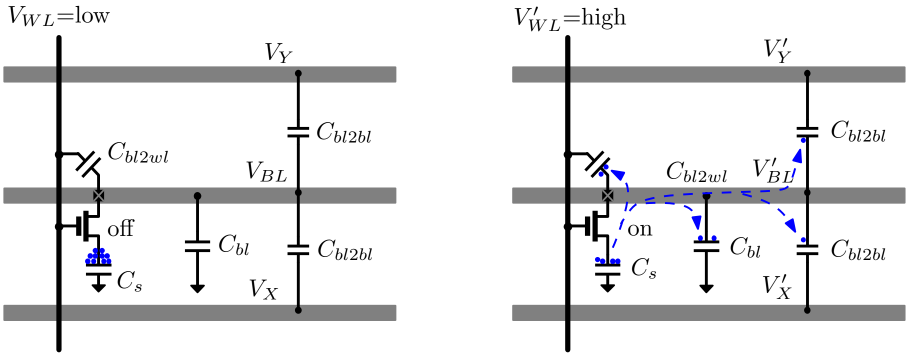
$V_{sign,max}$ is the developed maximum bitline voltage amplitude as shown in **Fig. 2.3**. K_t here is called transfer ratio, denoting the ability of the array to generate V_{sign} from the initial cell voltage V_{cell}. Eqn. (2.9) suggests that the maximum voltage difference between the pair of bitlines is completely determined by the headroom $V_{cell} - V_{eq}$ and cell to bitline capacitance ratio. V_{sign} is usually expected in the range from 100mV to 200mV considering the cell leakage, wordline to bitline coupling noise, inter-bitline interference and finite pre-sensing timing window. As technology advances, $V_{cell} - V_{eq}$ can never increase, and therefore the only way to maintain a large enough V_{sign} is to raise the capacitor ratio C_s/C_{bl}.

2.2 Array Structures and Array Parasitics

In the previous section, the developed bitline signal amplitude V_{sign} is obtained for the case with only bitline parasitic resistance and capacitance. Obviously, some other parasitic parameters are neglected in above evaluation. They can be very important design parameters in nowadays DRAM technology in which the metal wires are narrow and long with tighter pitch. For example, wordline to bitline capacitance not only couples the active wordline to floating bitlines, reducing the available voltage amplitude V_{sign} in open bitline arrays during pre-sensing, but also introduces active bitline voltage change to non-active wordlines, causing more sub-threshold leakage in non-active cells. In contrast with wordline to bitline capacitance, bitline to bitline capacitance is more harmful - it deteriorates V_{sign} during pre-sensing by absorbing part of the cell charge, and can eventually result in failing of weak pairs caused by post-sensing coupling. Several publications addressed the parasitic capacitance effect and tried to estimate V_{sign} in the presence of these parasitics from 1980 to 1990 [19, 20, 21, 22, 23]. Unfortunately, with the advances of DRAM technology, more parasitic capacitances appear to be significant and they are not considered in the earlier publications. Therefore, the previous outcomes can not meet the required accuracy for yield analysis and estimation. In this section, the charge conservation method is applied as before to the entire core array, in order to analyze the developed voltage amplitude V_{sign} precisely in pre-sensing. Based on the outcomes from this section, the yield of DRAM core can be estimated with good acuracy as will be demonstrated in Chapter 6.

2.2.1 Array parasitic effect

First think of three bitlines placed in parallel as shown in **Fig. 2.4**. The middle one with an active DRAM cell is our focus. When the cell is switched on, the bitline voltage begins to change. Ideally, a '1' cell raises the bitline voltage, and conversely a '0' cell lowers the bitline voltage. However, the ideal case comes from the assumption that its neighboring bitlines are staying quietly. If they are also changing simultaneously after wordline activation, the charge from the cell capacitor of the middle bitline will be partially attracted to the capacitors between the middle bitline and its neighbors, which may result in a smaller or even reversed voltage difference compared to the ideal case. A charge conservation equation for the middle bitline with the active cell can be setup to describe this process.



(a) Before wordline turns on

(b) After wordline turns on

Fig. 2.4: Charge conserves before and after wordline activation.

Before wordline activation:

$$\begin{aligned} Q_{tot} = {} & \underbrace{C_s \cdot V_{cell}}_{Q_{cell}} + \underbrace{nC'_{bl} \cdot V_{BL}}_{Q_{bl}} + \underbrace{nC'_{bl2bl} \cdot (V_{BL} - V_X) + nC'_{bl2bl} \cdot (V_{BL} - V_Y)}_{Q_{bl2bl}} \\ & + \underbrace{nC'_{bl2wl} \cdot (V_{BL} - V_{WL})}_{Q_{bl2wl}} \end{aligned} \tag{2.10}$$

After wordline activation:

$$\begin{aligned} Q'_{tot} = {} & \underbrace{C_s \cdot V'_{BL}}_{Q'_{cell}} + \underbrace{nC'_{bl} \cdot V'_{BL}}_{Q'_{bl}} + \underbrace{nC'_{bl2bl} \cdot (V'_{BL} - V'_X) + nC'_{bl2bl} \cdot (V'_{BL} - V'_Y)}_{Q'_{bl2bl}} \\ & + \underbrace{(n-1) \cdot C'_{bl2wl} \cdot (V'_{BL} - V_{WL}) + C'_{bl2wl} \cdot (V'_{BL} - V'_{WL})}_{Q'_{bl2wl}} \end{aligned} \tag{2.11}$$

In the above equations, C'_{bl2bl}, C'_{bl} are unit parasitic capacitance per cell between bitlines and from bitline to ground, respectively. n is the number of cells per bitline. C'_{bl2wl} is the parasitic capacitor between wordline and bitline. Assume the leakage current during the process is negligible. The charge conservation before and after wordline activation gives:

$$Q_{tot} = Q'_{tot} \tag{2.12}$$

Solving the equation, the middle bitline voltage after wordline activation is found to be

$$V'_{BL} = \frac{C_s \cdot V_{cell} + n(C'_{bl} + C'_{bl2bl}) \cdot V_{BL} + nC'_{bl2bl} \cdot (2V_{BL} + \Delta V_X + \Delta V_Y)}{C_s + n(C'_{bl} + C'_{bl2wl} + 2C'_{bl2bl})} + \frac{C'_{bl2wl} \cdot \Delta V_{WL}}{C_s + n(C'_{bl} + C'_{bl2wl} + 2C'_{bl2bl})}, \tag{2.13}$$

where $\Delta V_X = V'_X - V_X$, $\Delta V_Y = V'_Y - V_Y$, $\Delta V'_{WL} = V'_{WL} - V_{WL}$. Suppose V_{BL} is equal to V_{eq} in equalization as shown in **Fig. 2.3**. The middle bitline voltage step becomes

$$\begin{aligned} V_{sign} &= V'_{BL} - V_{BL} \\ &= \frac{C_s \cdot (V_{cell} - V_{eq}) + nC'_{bl2bl} \cdot (\Delta V_X + \Delta V_Y) + C'_{bl2wl} \cdot \Delta V_{WL}}{C_s + n(C'_{bl} + C'_{bl2wl} + 2C'_{bl2bl})} \\ &= K_t[\underbrace{(V_{cell} - V_{eq})}_{\text{signal}} + \underbrace{\frac{nC'_{bl2bl}}{C_s}(\Delta V_X + \Delta V_Y)}_{\text{bitline noise}} + \underbrace{\frac{C'_{bl2wl}}{C_s}\Delta V_{wl}}_{\text{wordline noise}}\ ] \end{aligned} \tag{2.14}$$

Eqn. (2.14) gives the general form of the pre-sensing voltage amplitude V_{sign} regarding different parasitic capacitances. Clearly, in a V_{sign} versus the cell signal $|V_{cell} - V_{eq}|$ plot the wordline to bitline and the bitline to bitline capacitances will introduce two effects as shown in **Fig. 2.5**:

1. Reduction of the array transfer ratio K_t;
2. Uncertain voltage fluctuations resulting from neighboring bitlines and active wordline.

Fig. 2.5 compares the developed voltage difference V_{sign} vs. $(V_{cell} - V_{eq})$ for cases with and without parasitic capacitance. Evidently, due to the drop of K_t, V_{sign2} becomes smaller than V_{sign1} at a given cell voltage. The voltage changes of wordline and the neighboring bitlines shift the line up or down, making V_{sign2} vary in a range. Because this shift is dependent on neighboring bitline voltage changes, it exhibits noise like characteristics when the neighboring bitline voltage fluctuations are random, and is thus called bitline noise voltage. In contrast with bitline noise, the interference introduced by wordline parasitic capacitance is always deterministic, presenting an offset when V_{sign2} is measured.

2.2.2 Array structures

Open bitline array

Eqn. (2.14) is obtained by comparing the bitline voltages before and after wordline activation. In practical circuit implementation, a complementary bitline is

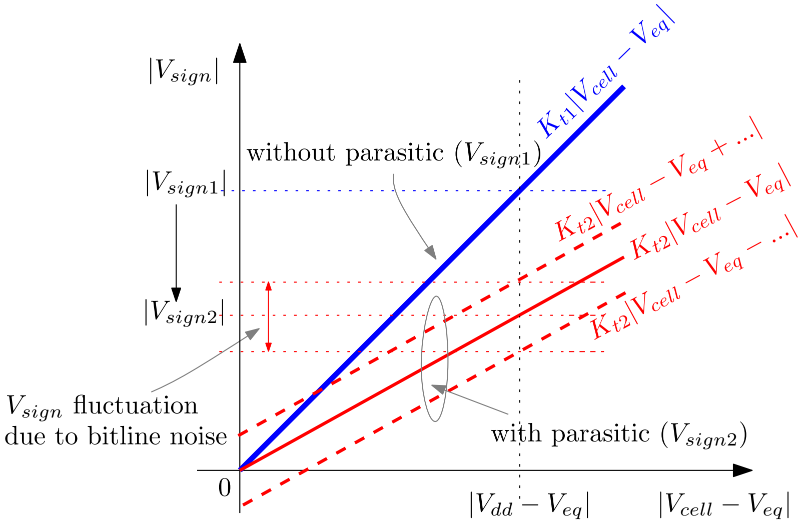


Fig. 2.5: V_{sign} vs. $(V_{cell} - V_{eq})$ with and without parasitic capacitance.

provided to 'remember' the bitline voltage before wordline activation, forming a bitline pair. In terms of bitline placement, different array structures exist. Open bitline array is one of these structures as shown in **Fig. 2.6**.

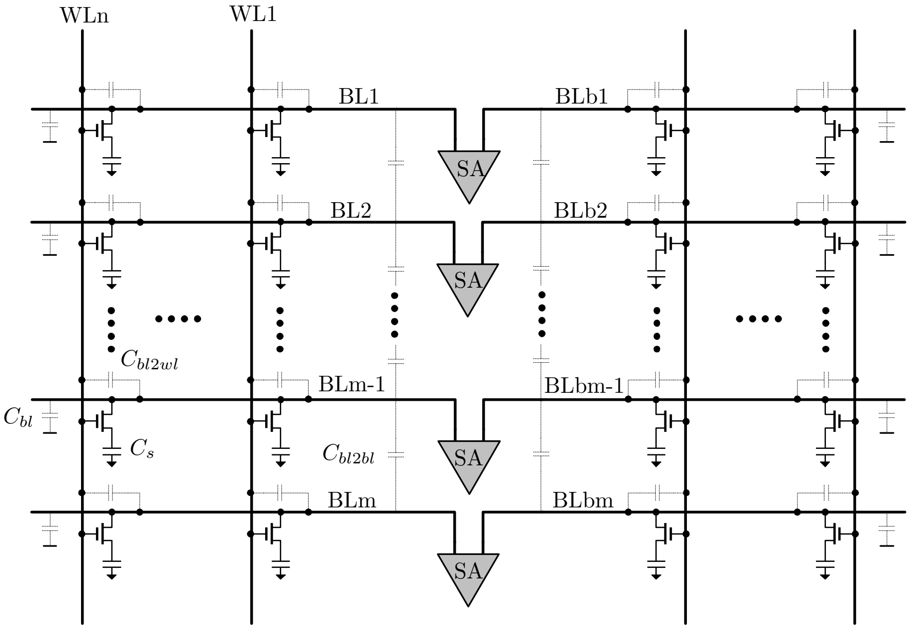


Fig. 2.6: An open bitline array with parasitic capacitances

In an open bitline array, complementary bitlines are completely separated from true bitlines that are under the control of the active wordline, and as a result, their voltage will not be disturbed during pre-sensing. From Eqn. (2.14), open bitline array with equalization voltage V_{eq} gives a developed voltage difference

$$V_{sign} = \frac{\overbrace{C_s \cdot (V_{cell} - V_{eq})}^{\text{signal}} + \overbrace{nC'_{bl2bl} \cdot (\Delta V_X + \Delta V_Y)}^{\text{bitline noise}} + \overbrace{C'_{bl2wl} \cdot \Delta V_{WL}}^{\text{wordline offset}}}{C_s + n(C'_{bl} + C'_{bl2wl} + \underline{2}C'_{bl2bl})} \tag{2.15}$$

ΔV_X, ΔV_Y in Eqn. (2.15) are independent. Presumably, since they are also generated from other bitlines, they will range from $-\Delta V$ to ΔV. The resulting maximum bitline noise amplitude is therefore $nC'_{bl2bl}/C_s \cdot 2\Delta V$ when $\Delta V_X = \Delta V_Y = \pm\Delta V$.

Example
Suppose an open bitline array provides the following parameters: $nC'_{bl2bl} = C_s/3$, $C'_{bl2wl} = C_s/200$, $nC'_{bl} = 2Cs$, $n = 512$, $\Delta V = 100\text{mV}$, $V_{cell} - V_{eq} = 0.5\text{V}$ and $\Delta V_{WL} = 2\text{V}$. From Eqn. (2.15), $K_t \approx 0.20$, and the maximum bitline noise amplitude is 20mV with average developed voltage difference $V_{sign} = 100\text{mV}$ and 2mV wordline offset. When wordline offset is ignored, the bitline noise in the open bitline array introduces 20% V_{sign} fluctuation as shown in **Fig. 2.5**.

Folded bitline array

It is observed that in open bitline arrays the voltages of two neighboring bitlines can fluctuate independently, greatly affecting the voltage of each other and introducing significant bitline disturbances. If the voltages of the neighboring bitlines change as little as possible, the bitline noise during pre-sensing can be suppressed to a certain degree with respect to open bitline array. Folded bitline array drawn in **Fig. 2.7** provides such benefit.

In folded bitline array the true and complementary bitlines are placed one after the other on the same side of sense amplifiers, and there is only one active cell for a bitline pair. Therefore, during pre-sensing each active bitline sees two quiet neighbors who are acting as reference. From Eqn. (2.14) obviously ΔV_X, ΔV_Y will be ideally zero, and thus the bitline noise can be eliminated. Besides bitline noise reduction, wordline offset can also be removed since each wordline is seen by both true and complementary bitlines in a folded bitline array. By Eqn. (2.13),

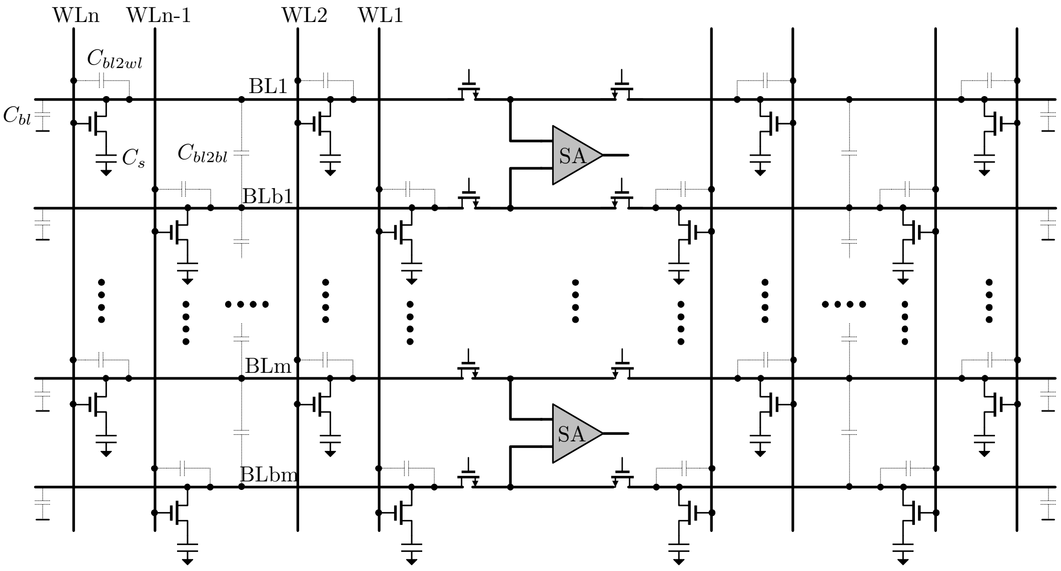


Fig. 2.7: A folded bitline array with parasitic capacitance.

the true and complementary bitline voltages after wordline activation give

$$\begin{aligned} V'_{BL} = {} & \frac{C_s \cdot V_{cell} + n(C'_{bl} + C'_{bl2bl}) \cdot V_{eq} + nC'_{bl2bl} \cdot (V_{eq} + \Delta V_X + V'_{BLb})}{C_s + n(C'_{bl} + C'_{bl2wl} + 2C'_{bl2bl})} \\ & + \frac{C'_{bl2wl} \cdot \Delta V_{WL}}{C_s + n(C'_{bl} + C'_{bl2wl} + 2C'_{bl2bl})} \end{aligned} \tag{2.16}$$

$$\begin{aligned} V'_{BLb} = {} & \frac{n(C'_{bl} + C'_{bl2bl}) \cdot V_{eq} + nC'_{bl2bl} \cdot (V_{eq} + V'_{BL} + \Delta V_Y)}{n(C'_{bl} + C'_{bl2wl} + 2C'_{bl2bl})} \\ & + \frac{C'_{bl2wl} \cdot \Delta V_{WL}}{n(C'_{bl} + C'_{bl2wl} + 2C'_{bl2bl})} \end{aligned} \tag{2.17}$$

The difference of the true and complementary bitline voltage becomes

$$\begin{aligned} V_{sign} &= \frac{C_s \cdot (V_{cell} - V'_{BLb}) + nC'_{bl2bl} \cdot (\Delta V_X - \Delta V_Y)}{C_s + n(C'_{bl} + C'_{bl2bl} + \underline{3}C'_{bl2bl})} \\ &\approx \frac{\overbrace{C_s \cdot (V_{cell} - V_{eq})}^{\text{signal}} + \overbrace{nC'_{bl2bl} \cdot (\Delta V_X - \Delta V_Y)}^{\text{bitline noise}}}{C_s + n(C'_{bl} + C'_{bl2wl} + \underline{3}C'_{bl2bl})} \end{aligned} \tag{2.18}$$

Although wordline offset disappears in Eqn. (2.18), bitline noise still exists in the folded bitline array. The reason is that reference bitlines in folded bitline array are also disturbed by their closest active bitlines, and therefore can not stay quiet during pre-sensing. Compared with V_{sign} from open bitline array, ΔV_X, ΔV_Y in folded bitline array come from reference bitline and active bitline, respectively, instead of two active bitlines, and therefore the maximum bitline noise amplitude is only half of that in open bitline array. However, the transfer ratio K_t of the folded bitline array is lower. In Eqn. (2.18), since the complementary bitline voltage V'_{BLb} does not change much during pre-sensing, it is regarded to be equal to the pre-charge voltage V_{eq}.

Example
Suppose a folded bitline array provides the following parameters: $n = 512$, $nC'_{bl2bl} = C_s/3$, $C'_{bl2wl} = C_s/200$, $nC'_{bl} = 2Cs$, $\Delta V_X - \Delta V_Y = 150\text{mV}$, $V_{cell} - V_{eq} = 0.5\text{V}$ and $\Delta V_{WL} = 2\text{V}$. From Eqn. (2.18), $K_t \approx 0.19$, the maximum bitline noise amplitude is 9.56mV with average developed voltage difference $V_{sign} = 96.67\text{mV}$. The bitline noise in the folded bitline array introduces 10% V_{sign} fluctuation that is only half of it in a corresponding open bitline array as calculated in the previous example.

Multiple twisted folded bitline array

In a folded bitline array bitline noise is reduced by a factor of two compared to its open bitline counterpart due to the 'shielding effect' of reference bitlines. To better suppress bitline noise, the term $nC'_{bl2bl}(\Delta V_X - \Delta V_Y)$ in Eqn. (2.18) is expected to be further diminished. It can be accomplished with the aid of multiple twisted bitline arrays [19, 24, 25, 26, 27, 28, 29, 30].

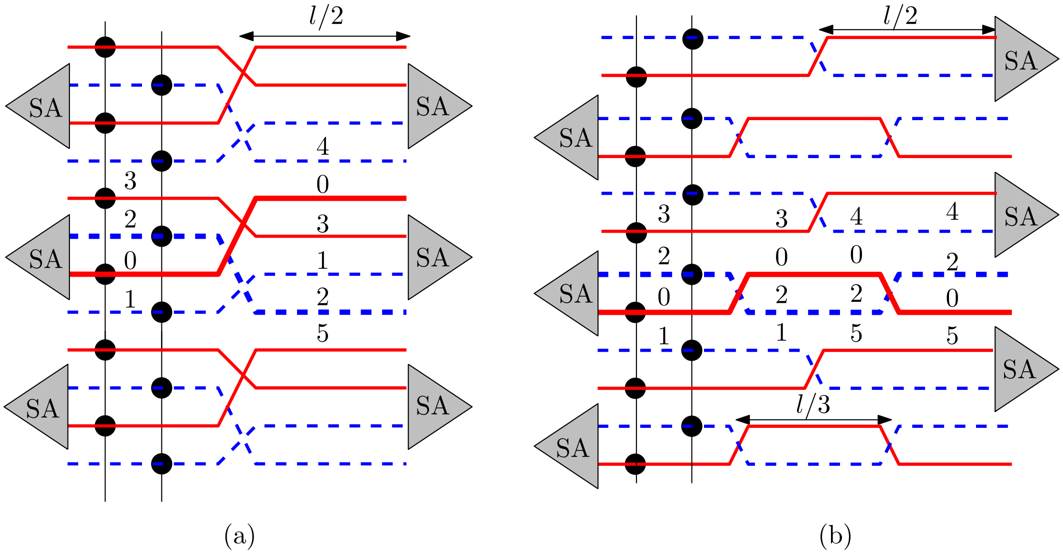


Fig. 2.8: (a) Quadruple style multiple twisted folded bitline array [26]. (b). Fully symmetric multiple twisted folded bitline array [24].

Fig. 2.8 exhibits two multiple twisted folded arrays. In (a) the solid lines are active bitlines and the dotted lines are reference bitlines. By multiple twisted bitline structure the active bitline 0 has four neighbors 1, 2, 3, 4, while its reference bitline 2 has four neighbors 0, 1, 3, 5 as well. By Eqn. (2.13), the true and complementary bitline voltages after wordline activation in such array are

$$\begin{aligned} V_0' = &\frac{C_s \cdot V_{cell} + n(C_{bl}' + C_{bl2bl}') \cdot V_{eq} + nC_{bl2bl}'/2 \cdot (4V_{eq} + \Delta V_1 + \Delta V_2 + \Delta V_3 + \Delta V_4)}{C_s + n(C_{bl}' + C_{bl2wl}' + 2C_{bl2bl}')} \\ &+ \frac{C_{bl2wl}' \cdot \Delta V_{WL}}{C_s + n(C_{bl}' + C_{bl2wl}' + 2C_{bl2bl}')}, \end{aligned} \tag{2.19}$$

$$\begin{aligned} V_2' = &\frac{n(C_{bl}' + C_{bl2bl}') \cdot V_{eq} + nC_{bl2bl}'/2 \cdot (4V_{eq} + \Delta V_0 + \Delta V_1 + \Delta V_3 + \Delta V_5)}{n(C_{bl}' + C_{bl2wl}' + 2C_{bl2bl}')} \\ &+ \frac{C_{bl2wl}' \cdot \Delta V_{WL}}{n(C_{bl}' + C_{bl2wl}' + 2C_{bl2bl}')} \end{aligned} \tag{2.20}$$

Because $V_{sign} = V_0' - V_2'$, the developed voltage difference becomes

$$\begin{aligned} V_{sign} &= \frac{C_s \cdot (V_{cell} - V_2') + nC_{bl2bl}'/2 \cdot (\Delta V_4 - \Delta V_5)}{C_s + n[C_{bl}' + C_{bl2wl}' + (2 + 1/2) \cdot C_{bl2bl}')]} \\ &\approx \frac{C_s \cdot (V_{cell} - V_{eq}) + \overbrace{nC_{bl2bl}'/2 \cdot (\Delta V_4 - \Delta V_5)}^{\text{bitline noise}}}{C_s + n[C_{bl}' + C_{bl2wl}' + (2 + 1/2) \cdot C_{bl2bl}']} \end{aligned} \tag{2.21}$$

K_t from Eqn. (2.21) for a multiple twisted array is larger than for a folded bitline array but still smaller than for an open bitline array. Meanwhile, bitline noise amplitude is further decreased by a factor of 2 from the folded bitline array. The bitline noise reduction results from the mechanism that part of the neighbors' voltage fluctuations are converted into common mode components that can be eliminated by the differential operation of sense amplifiers. In addition, Eqn. (2.21) also suggests the general form of V_{sign} in a multiple twisted array

$$V_{sign} \approx \frac{C_s \cdot (V_{cell} - V_{eq}) + nC_{bl2bl}'/k_0 \cdot (\sum \Delta V)}{C_s + n[C_{bl}' + C_{bl2wl}' + (2 + k_1/k_0) \cdot C_{bl2bl}']} \tag{2.22}$$

Here k_0 is determined by the twisted bitline to twisted unit segment length ratio, e.g., $k_0 = 2$ for **Fig. 2.8**(a) because each bitline is composed of two twisted segments. k_1 is the number of unit segments of the true and complementary bitlines being placed together. In (a) bitline 0 and 2 stay together for one twisted bitline segment, resulting in $k_1 = 1$. $\sum \Delta V$ is the total sum of neighboring non-common mode bitline voltage steps for a bitline pair. As shown in (a), bitline 4 and 5 is the kind of bitline seen only by bitline 0 and 2, respectively. Eqn. (2.22) is also valid for normal folded bitline arrays with $k_0 = k_1 = 1$.

Example

Evaluate V_{sign} in **Fig. 2.8**(b) [24]. Since there are two different twist length and their least common multiple is $l/6$, k_0 becomes 6. The neighbors of bitline 0 are 1, 2, 3, 4, 5, which are exactly the same as its reference bitline 2 that has 0, 1, 3, 4, 5 as neighbors when bitline 0, 2 are excluded. As a result, all bitline interferences in the array are common mode noise, resulting in $\sum \Delta V = 0$. Because the bitline length that 0 and 2 are routed together is 6, K_1 also becomes 6. The developed voltage difference in (b) is

$$V_{sign} \approx \frac{C_s \cdot (V_{cell} - V_{eq})}{C_s + n[C'_{bl} + C'_{bl2wl} + (2 + 6/6) \cdot C'_{bl2bl}]} \tag{2.23}$$

As demonstrated, multiple twisted folded bitline arrays can greatly suppress or even completely eliminate bitline noise. However, they suffer from the area penalty introduced by the twist area and the contact congestion problem induced yield degradation. As this problem becomes much severer with continuously shrinking feature size, multiple twisted folded bitline arrays lose their charms gradually.

Mixed array structure

Since open bitline arrays have the ability to accommodate more cells while folded bitline arrays can suppress bitline noise, a mixture of both bitline arrangements comes into play when cells become even smaller. In [31] the cells are arranged in a manner that open and folded bitline pairs can be placed in an interleaved style. When it is accessed half of the bitline pairs will work in folded bitline mode while the others in open bitline mode. In [32] a folded mixed array is proposed for $6F^2$ cell array, extra switches are placed in between two sub-arrays and a special cell arrangement has to be made. As a benefit, these arrays can reduce bitline to bitline disturbance while increasing cell density to a higher degree that folded bitline array can not achieve. However, the irregular cell arrangement complicates the layout and technology. These mixed array structures provide only temporary transitions from $8F^2$ cells to $6F^2$.

Arrays with bitline shielding

Folded bitline arrays have long been popular because of their better bitline noise suppression ability. However, with the cell size scaling down to $6F^2$ they are suddenly given up because the cells are becoming too compact to be arranged in an interleaved style required by folded bitline arrays. Mixed array structures can temporarily accommodate $6F^2$ cells with tolerable bitline noise but still confront the same problem - what if cells evolve to $4F^2$? One solution can be the bitline shielding technique.

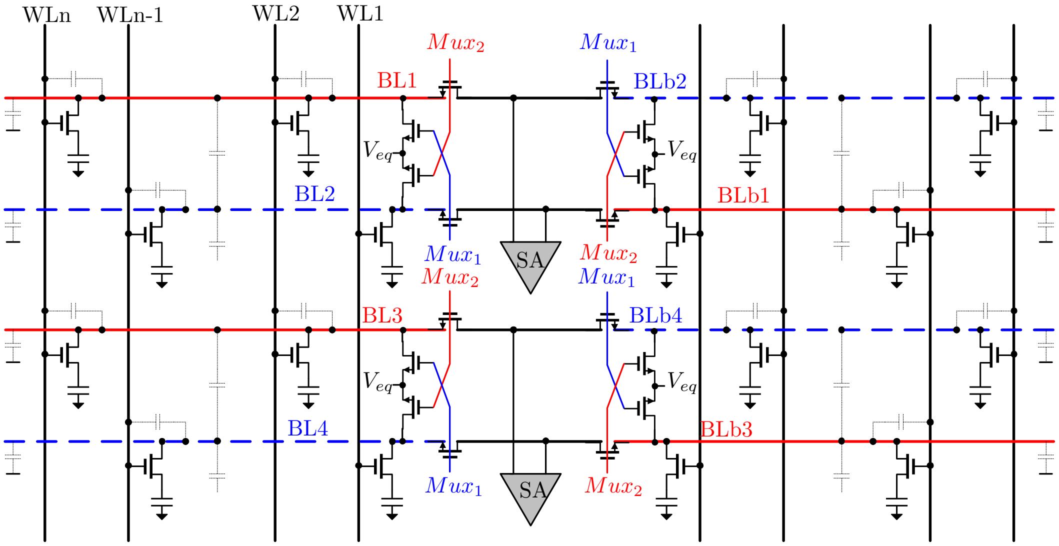


Fig. 2.9: An open bit line array with self isolation

In fact, bitline shielding has been proposed for many years. An open bitline array with self bitline shielding was proposed [33] during the time when folded bitline with multiple twists were still popular as shown in **Fig. 2.9**. The shielding function is accomplished as follows: Before the wordline WLn is accessed, Mux_2 is selected and sense amplifiers (SA) are connected to bitlines BL1, BLb1 and BL3, BLb3 respectively. At the same time, bitlines BL2, BLb2, BL4, BLb4 are tied to equalization voltage V_{eq}. When WLn is switched on, bitlines BL1, BL3 will rise or fall because of the charge redistribution process while their neighbors stay quiet. As explained in the previous sections, in this scenario the active bitlines will not be disturbed and thus the array provides a developed voltage amplitude almost identical to the multiple twisted folded bitline array in **Fig. 2.8**(b) but without extra die cost as needed for the twist region in multiple twisted array. The limitation of the isolated open bitline structure resides in applicable cell size. When cells are squeezed to below $6F^2$, it becomes infeasible.

The technology solution for bitline shielding was first proposed by Mashiko [34] in an open bitline 256K DRAM design in 1984. In his approach, the open bitlines are isolated from each other by MOS type cell capacitors and wordline contacts. However, three poly-silicon layers are used in the design and the resulting cost is relatively high.

Stacked capacitor [35, 36, 37] DRAM cell inherently has bitline shielding ability because the cell transistors, which are fabricated on the silicon surface, have to be connected to the stacked capacitors that usually stay in the top layer above the interconnections for the bitline formation. Since the capacitor contacts are made by metallic connections, a shielding structure is naturally formed between bitlines.

2.2.3 Evaluation of developed bitline voltage in pre-sensing

In the previous section, V_{sign} is estimated for different array structures. Unfortunately, the neighboring bitline voltage change ΔV_X, ΔV_Y in these V_{sign} equations are actually uncertain. Since V_{sign} is very important especially for yield estimation in succeeding chapters, a set of metrics will be developed here to help the calculation of bitline voltage in an array.

Eqn. (2.11), Eqn. (2.10) and charge equilibrium is always valid for all array structures. Once the charge conservation equations describing the voltage change of each bitline in an array are considered, a linear matrix to describe the charge balance in an $n \times m$ array can be obtained.

$$\mathbf{Q} = \mathbf{C} \cdot \mathbf{V'_{BL}}, \text{ and } \mathbf{V'_{BL}} = \mathbf{C^{-1}} \cdot \mathbf{Q}, \tag{2.24}$$

where

$$\mathbf{V'_{BL}} = \left(V'_{BLdum}, V'_{BLm}, V'_{BLm-1}, \ldots, V'_{BL1}, V'_{BL0}, V'_{BLdum}\right)^T_{m+2}, \tag{2.25}$$

$$\mathbf{Q} = \begin{pmatrix} V_{dum}C_d \\ V_{celln}C_s + nC'_{bl}V_{eq} + nC'_{bl2wl}\Delta V_{WL} \\ V_{celln-1}C_s + nC'_{bl}V_{eq} + nC'_{bl2wl}\Delta V_{WL} \\ \vdots \\ V_{cell2}C_s + nC'_{bl}V_{eq} + nC'_{bl2wl}\Delta V_{WL} \\ V_{cell1}C_s + nC'_{bl}V_{eq} + nC'_{bl2wl}\Delta V_{WL} \\ V_{dum}C_d \end{pmatrix}_{m+2}, \tag{2.26}$$

$$\mathbf{C} = \begin{pmatrix} C_d & 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ \gamma & \beta & \alpha & 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & \gamma & \beta & \alpha & 0 & \ldots & 0 & 0 & 0 \\ & & & & \ddots & & & & \\ 0 & 0 & 0 & \ldots & 0 & \gamma & \beta & \alpha & 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 & \gamma & \beta & \alpha \\ 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & C_d \end{pmatrix}_{m+2,m+2}. \tag{2.27}$$

$\mathbf{V'_{BL}}$ is the vector containing all bitline voltages after pre-sensing. m, n are the total number of bitlines in the sub-array and total bits per bitline, respectively. V_{dum}, C_d are the voltage and capacitance for dummy bitlines on the boundaries of the array. Since dummy bitlines are tied to fixed potentials, V_{dum} is a constant. α, β, γ are parameters determined by array structure and array capacitive parasitics, e.g., for a folded bitline array as shown in **Fig. 2.7**, $\alpha = \gamma = -nC'_{bl2bl}$, $\beta = (C_s + nC'_{bl} + nC'_{bl2wl} + 2nC'_{bl2bl})$.

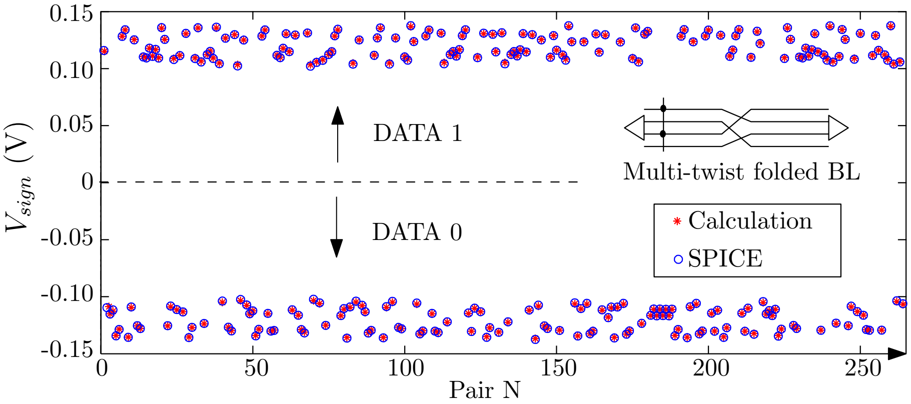


Fig. 2.10: Comparison of calculated $V_{sign}(N)$ with SPICE simulation for a multiple twisted folded array with 512 bitlines and random data pattern (256 generated V_{sign}) for a 75nm DRAM core array

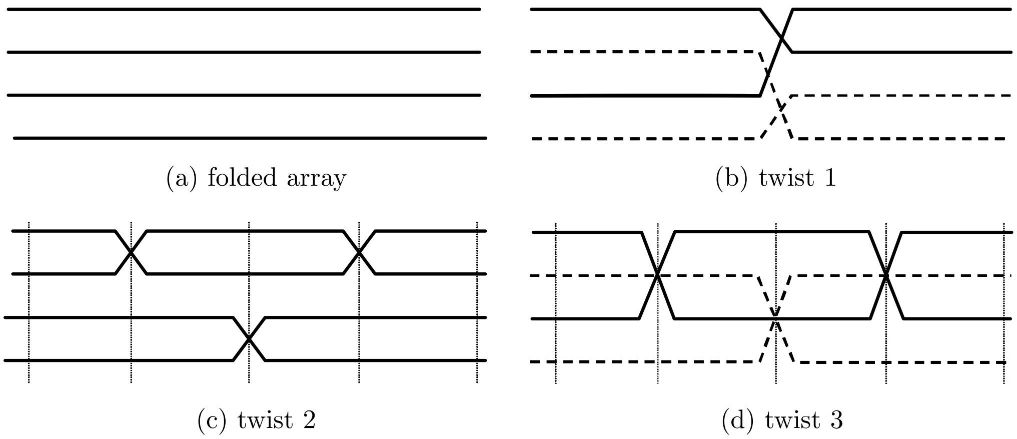


Fig. 2.11: Arrays with and without twists

Random data patterns

V_{sign} for each bitline pair is then obtained from the difference between the true and complementary bitline voltage. By means of mathematic tools, Eqn. (2.24) can be easily solved in short time. **Fig. 2.10** demonstrates an example where the theoretical calculated V_{sign} for each bitline pair is compared to the outcome of SPICE simulation with a random solid '0', '1' pattern in the multiple twisted folded array of **Fig. 2.8**. Evidently, the results prove the good accuracy of the charge conservation model.

An obvious advantage from the model is that all kinds of data patterns and parameter alternations can be processed with much less effort compared to SPICE simulations. Here, several evaluations are carried out on different array structures as shown in **Fig. 2.11** and **Fig. 2.12**. A random data pattern with

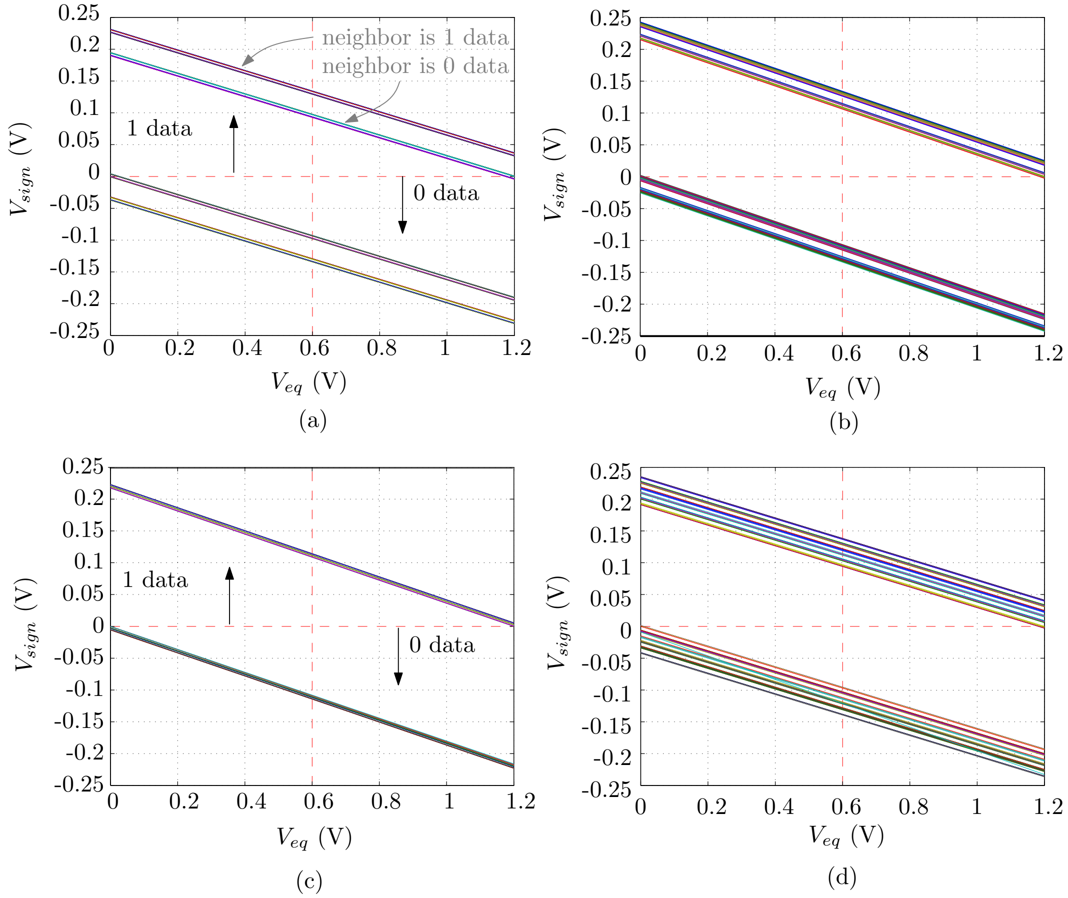


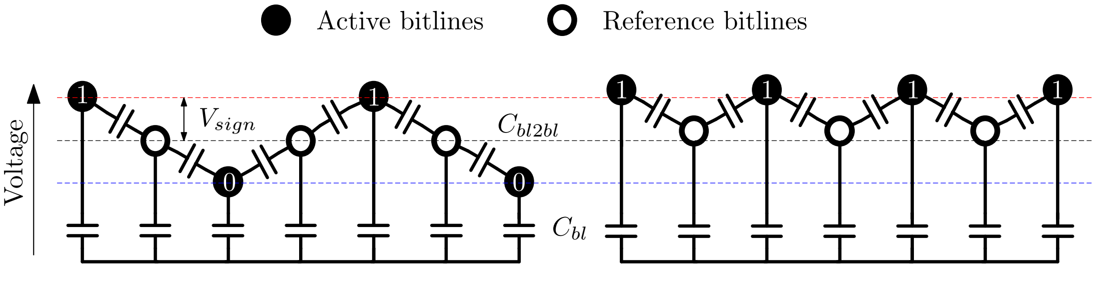
Fig. 2.12: Comparison of calculated V_{sign} for (a), (b), (c), (d) arrays in Fig. 2.11 with random data pattern

solid '0's, '1's is first fed to the arrays. When the equalization voltage V_{eq} is swept from 0 to V_{dd}, the calculated V_{sign} for all bitline pairs are recorded as shown in **Fig. 2.12**. Since V_{sign} is proportional to V_{eq} when V_{cell} is constant, these plots exhibit several straight lines. Interestingly, although 256 V_{sign} values are obtained each time from an array at a given V_{eq}, they are discretely positioned at only several levels. The reason is found in the number of different neighbor types for a given bitline pair. For the folded bitline array, each bitline pair has two neighbors - an active bitline and a reference bitline, and therefore, there are two strong and two weak trajectories caused by the bitline pair's active and reference neighbors, respectively, for both $V_{sign} > 0$ and $V_{sign} < 0$. For multiple twisted arrays, the increased number of non-common mode neighboring bitlines will give rise to more trajectories as shown in (d). Since twist (c) contributes almost no differential mode bitline noise, its outcomes are very uniform.

Some regular data patterns

When data stored in an array exhibits regular topologies such as all '0', all '1' or a '01' interleaved pattern, Eqn. (2.24) can be greatly simplified, i.e., $\mathbf{V'_{BL}}$ in the array will become uniform and amplitudes of V_{sign} for all pairs will be identical.

For an open bitline array '01' alternating pattern gives a larger pre-sensing voltage amplitude and all '0' or all '1' generates a much smaller pre-sensing voltage amplitude, whereas in a folded bitline array the data sequences give completely opposite results. However in either case, the regular pattern corresponding to the situation where the bitline interference is most significant are used in the book. In the open bitline array, ideally there will be no bitline to bitline noise when all the bitlines rise up or drop down simultaneously with all '0' or '1' pattern; Conversely the '01' alternating pattern results in completely different movements for every bitline and its neighbors, deteriorating V_{sign} for all pairs. **Fig. 2.13** illustrates the situations in folded bitline array. Due to C_{bl2bl} when the '01' alternating pattern is applied, the reference bitline voltage in the folded bitline array will barely change due to the symmetric and opposite movements of their neighboring bitlines; But they will be raised or lowered from V_{eq} in all '1' or '0' pattern. As a result, the '01' alternating pattern gives larger V_{sign} than all '0' or '1' pattern for folded bitline arrays.



(a) 01 alternating pattern

(b) all 1 pattern

Fig. 2.13: The bitline voltage for different data pattern in folded bitline array

It is also interesting to see that with $V_{eq} = V_{dd}/2$ and regular patterns, V_{sign} of bitline pairs in the array from the metrics in Eqn. (2.24) become uniform. By assuming ΔV_X, ΔV_Y in Eqn. (2.15), Eqn. (2.18) and Eqn. (2.22) have the same amplitude as the bitline pair being focused, the regular pattern related V_{sign} can be written as

$$V_{sign} = \frac{C_s \cdot (V_{cell} - V_{eq})}{C_s + n \cdot (C'_{bl} + C'_{bl2wl} + \lambda \cdot C'_{bl2bl})} \tag{2.28}$$

λ in Eqn. (2.28) depends on array structures and data patterns as summarized in **Table 2.1**. When statistical variations are applied to array parameters, V_{sign} will show a Gaussian like distribution according to central limit theorem. Therefore, these patterns are quite useful in analyzing and estimating the DRAM yield.

Table 2.1: λ for different arrays and data patterns

	Folded array	Open array	MTW$_1$	MTW$_2$
Patterns with smaller pre-sensing voltage	4	4	3	3
Patterns with larger pre-sensing voltage	2	0	2	3

MTW$_1$: Multiple twisted folded bitline array in **Fig. 2.11**(b)
MTW$_2$: Multiple twisted folded bitline array in **Fig. 2.11**(c)

2.3 Post-sensing Crosstalk Coupling

The developed bitline voltage holds its amplitude for several nanoseconds after pre-sensing until sense amplifiers are triggered. Sense amplifiers enlarge the voltage difference between a bitline pair, and recover the cell voltage to V_{dd} or V_{ss}. Theoretically, the sense amplifier's outcome should only be determined by V_{sign}, i.e., the output is '1' for $V_{sign} > 0$ and '0' for $V_{sign} < 0$. However, it can actually be opposite and hence leads to a failure due to the presence of threshold voltage mismatch of sensing transistors and cross-talk coupling between bitlines during post-sensing.

The threshold mismatch of sensing transistors will be disscussed later in Chapter 4. It can be equivalent to an input voltage source V_{os} that is directly added to V_{sign} from pre-sensing. When $|V_{sign}| < |V_{os}|$ the sensing may fail to generate a correct outcome. Taking V_{os} into consideration, the effective input differential voltage after pre-sensing is changed from V_{sign} to $V'_{sign} = V_{sign} + V_{os}$.

In addition to the impact resulting from mismatch of sensing transistors, the rapidly changing bitline voltages during post-sensing introduce high frequency disturbances through parasitic capacitances between bitlines. Furthermore, due to randomized mismatch of sensing transistors, V'_{sign} for each bitline pair differs from one another. Bitline pairs with larger V'_{sign} will be driven faster and become aggressors, attacking the weaker neighbors and turning them into sensing failures. As a result, post-sensing coupling deteriorates the situation that is already worsened by the presence of mismatched sensing transistors.

2.3.1 Cross-talk coupling with non-latched sense amplifiers

When bitlines are driven by sense amplifiers without positive feedback loop, cross-talk coupling will not generate failures since final outputs are determined by the

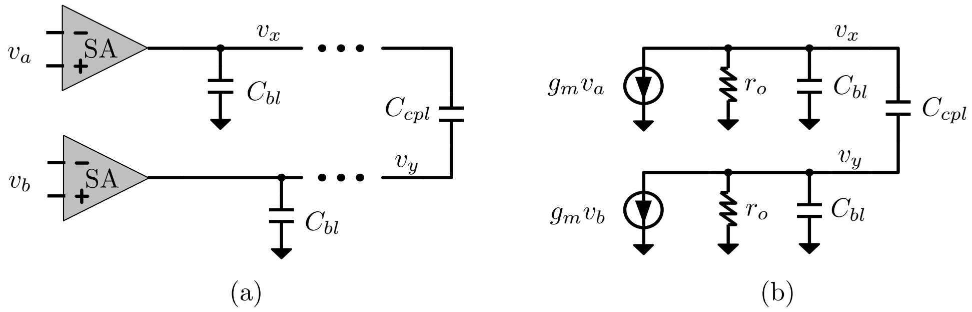


Fig. 2.14: Circuit model (a) and its small signal circuits (b) to study the coupling behavior for transconductance sense amplifiers

invariable inputs in this case. This can be accomplished by separating the sensing process from the cell refresh process. Because the output loads of sense amplifiers are capacitive in DRAM core, operational transconductance amplifiers (OTAs) can be used. Suppose the transconductance of sense amplifiers are equal to g_m and the coupling capacitor C_{cpl} exists between two outputs v_x and v_y as shown in **Fig. 2.14**(a). Differential equations can be drawn from the equivalent small signal circuits in **Fig. 2.14**(b) as

$$\begin{cases} C_{bl} \cdot \dfrac{\partial v_x}{\partial t} + C_{cpl} \cdot \dfrac{\partial (v_x - v_y)}{\partial t} = -g_m \cdot v_x(0) - v_x/r_o \\ C_{bl} \cdot \dfrac{\partial v_y}{\partial t} + C_{cpl} \cdot \dfrac{\partial (v_y - v_x)}{\partial t} = -g_m \cdot v_y(0) - v_x/r_o \end{cases} \tag{2.29}$$

In above equations, r_o is the output resistance of the OTA type sense amplifier; C_{bl} is the load capacitance; C_{cpl} is the coupling capacitor between two sense amplifier outputs. Suppose the initial outputs of v_x, v_y are 0, and initial inputs for the sense amplifiers are v_a, v_b. The above equations yield the transient output of v_x:

$$\begin{aligned} v_x(t) = &-\frac{v_a}{2} \cdot g_m r_o \cdot [2 - e^{-\frac{t}{r_o C_{bl}}} - e^{-\frac{t}{r_o(C_{bl}+2C_{cpl})}}] \\ &\underbrace{-\frac{v_b}{2} \cdot g_m r_o \cdot [e^{-\frac{t}{r_o(C_{bl}+2C_{cpl})}} - e^{-\frac{t}{r_o C_{bl}}}]}_{coupled\ voltage} \end{aligned} \tag{2.30}$$

Eqn. (2.30) suggests:

- v_x may need longer time to reach an expected amplitude when $v_b < 0$ and $v_a > 0$;
- The coupled voltage increases with larger C_{cpl};
- The coupled voltage increases with larger $g_m r_o$;

When the output from Eqn. (2.30) is referred back to the input of v_x, the equivalent initial input v_a' becomes

$$v_a' = v_a + v_b \cdot \frac{-e^{-\frac{t}{r_o C_{bl}}} + e^{-\frac{t}{r_o(C_{bl}+2C_{cpl})}}}{2 - e^{-\frac{t}{r_o C_{bl}}} - e^{-\frac{t}{r_o(C_{bl}+2C_{cpl})}}} = v_a + K_{cpl} \cdot v_b \ , \ (t \neq 0) \qquad (2.31)$$

The coefficient K_{cpl} is a function of time, coupling capacitor C_{cpl}, sense amplifier load capacitance C_{bl} and output impedance r_o. **Fig. 2.15** shows the calculated K_{cpl} vs. time delay for different C_{bl}, C_{cpl} and r_o settings. The time delay here is calculated from the on-time of sense amplifiers to the time when the outputs are sampled.

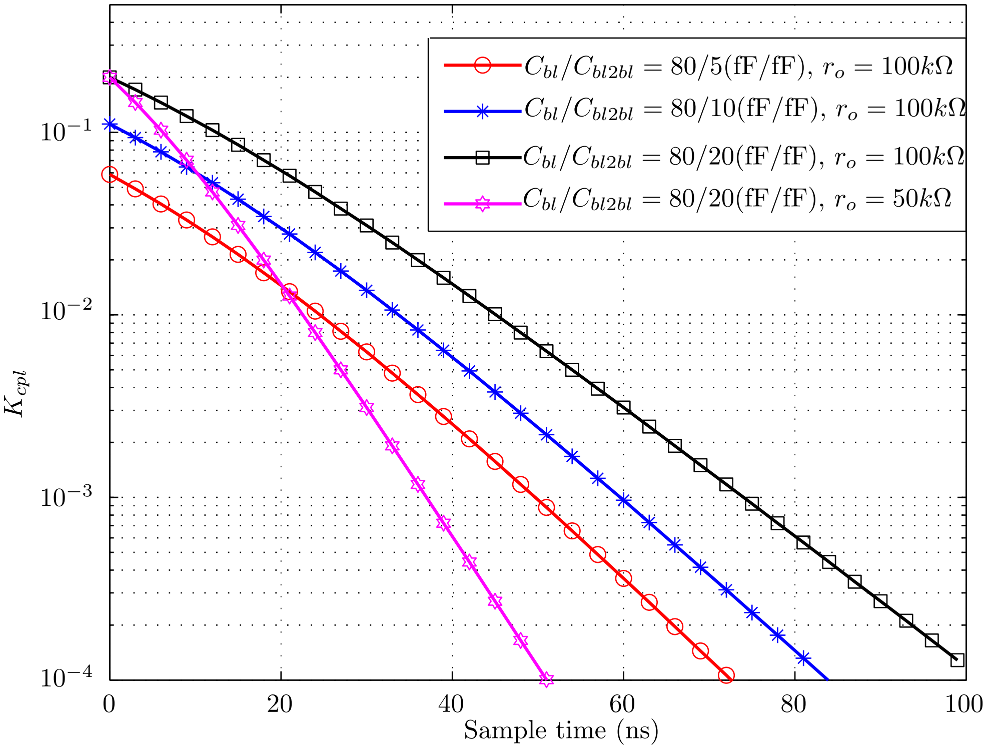


Fig. 2.15: Calculated K_{cpl} corresponding to different parameters for OTA type sense amplifiers ($t \neq 0$).

Example

Suppose in an extreme case where v_b is hundred times larger than v_a and they have different polarities, the output is sampled at time $t = 40ns$ after the sense amplifiers are enabled. From Fig.2.15 the group with $C_{bl}/C_{cpl} = 4, R = 100k\Omega$ will give rise to wrong outcomes since K_{cpl} at $t = 40$ns is still greater than 1/100. The figure also suggests the group with lower output resistance $r_o = 50k\Omega$ is favorable since their K_{cpl} attenuates much faster with the elapse of time.

Similarly, when the output is coupled on both sides by two different sources v_y and v_z that are also driven by g_m sense amplifiers, Eqn. (2.31) changes to

$$\begin{aligned} v_a'(t) &= -\frac{1}{3}g_m r_o \cdot [3 - e^{-\frac{t}{r_o C_{bl}}} - 2e^{-\frac{t}{r_o(3C_{cpl}+C_{bl})}}] \cdot v_a \\ &\quad - \frac{1}{3}g_m r_o \cdot [e^{-\frac{t}{r_o(3C_{cpl}+C_{bl})}} - e^{-\frac{t}{r_o C_{bl}}}] \cdot (v_b + v_c), \end{aligned} \tag{2.32}$$

where v_b, v_c are the initial input voltages of neighboring sense amplifiers. The equivalent input initial voltage of the victim becomes

$$v_a' = v_a + K_{cpl} \cdot (v_b + v_c) \tag{2.33}$$

which is similar to Eqn. (2.31).

2.3.2 Cross-talk coupling with latched sense amplifiers

Latched sense amplifiers as shown in **Fig. 2.16** are more popular than non-latch sense amplifiers in commercial DRAMs due to several reasons. Firstly, sensing and cell refresh can be combined into one process with the aid of the positive feedback loop. Secondly, the amplification speed can be greatly accelerated by the latch structure. Lastly, the control effort and area consumption of this sense amplifier is relatively low compared to the OTA style.

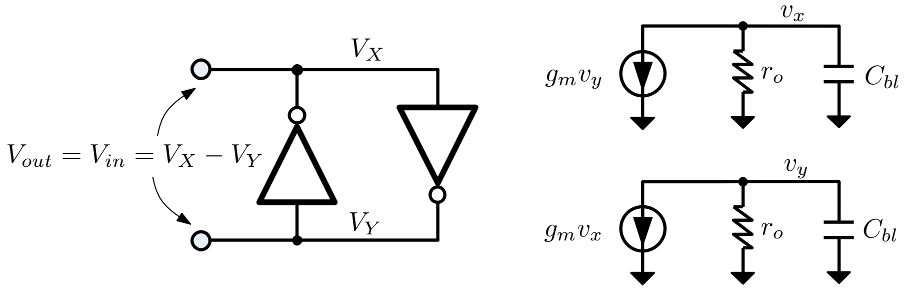
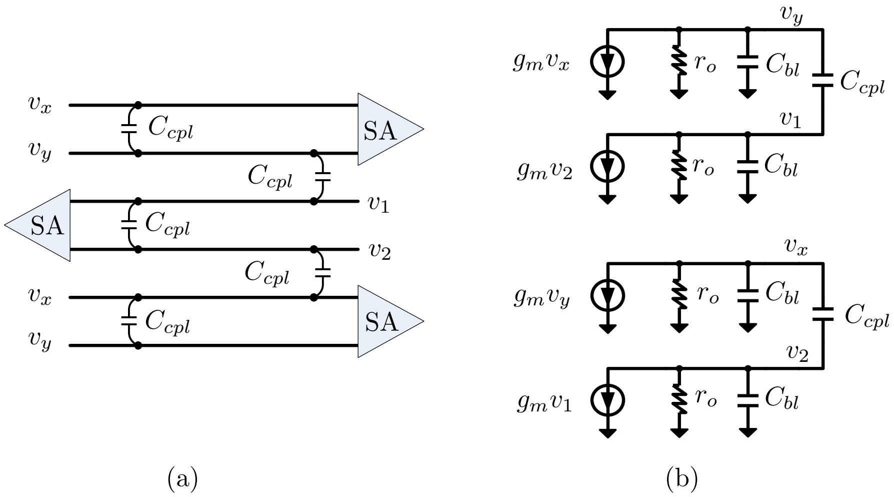


Fig. 2.16: A latched amplifier and its equivalent small signal circuits (Suppose the operating point $V_X \approx V_Y$ so that the latch can be represented as its AC counterpart consisting of transconductance g_m, output resistance r_o and load capacitance C_{bl})

The disadvantage of using latched sense amplifiers comes from the increasing failures caused by cross-talk coupling through parasitic bitline to bitline capacitance. In contrast to OTA type sense amplifiers, once the polarity of the bitline voltage difference of a weak sense amplifier is forced to flip by cross-talk interference, it will be never turned over again even with infinite sensing delay because of the internal positive feedback loop that continuously strengthens the input



(a) (b)

Fig. 2.17: A circuit set up to study the coupling effect between latched sense amplifiers.

though it may be accidentally generated under the influence of neighboring bitlines.

Suppose in a folded bitline array a weak sensing pair is placed between two strong neighbors that are barely affected by others as shown in **Fig. 2.17** (a). To make it simple, it is further assumed the two strong pairs have the same bitline voltage as v_y, v_x while the center sense amplifier's bitline voltages are v_1, v_2 as denoted. Noticeably, since the top and bottom strong pairs have the same bitline voltage their small signal circuit can be represented by the same one as shown in **Fig. 2.17**(b). When the initial inputs of the latch sense amplifiers are v_a, v_b for the weak and strong pairs, the following differential equations can be drawn as

$$\begin{cases} C_{bl} \cdot dv_1/dt + C_{cpl} \cdot d(v_1 - v_y) = -g_m v_2 - v_1/r_o \\ C_{bl} \cdot dv_2/dt + C_{cpl} \cdot d(v_2 - v_x) = -g_m v_1 - v_2/r_o \\ C_{bl} \cdot dv_y/dt + C_{cpl} \cdot d(v_y - v_1) = -g_m v_x - v_y/r_o \\ C_{bl} \cdot dv_x/dt + C_{cpl} \cdot d(v_x - v_2) = -g_m v_y - v_x/r_o \\ v_1(0) = v_a/2, v_2(0) = -v_a/2 \\ v_y(0) = v_b/2, v_x(0) = -v_b/2 \end{cases} \tag{2.34}$$

C_{bl} is not shown in **Fig. 2.17**. It can be regarded as bitline to ground parasitic capacitance. By solving the above differential equations, the output of the weak pair is obtained

$$\begin{aligned} v_o(t) &= v_1(t) - v_2(t) \\ &= \frac{v_a}{2} \cdot [e^{\frac{g_m r_o - 1}{r_o(2C_{cpl}+C_{bl})}t} + e^{\frac{g_m r_o - 1}{r_o C_{bl}}t}] + \frac{v_b}{2} \cdot [e^{\frac{g_m r_o - 1}{r_o(2C_{cpl}+C_{bl})}t} - e^{\frac{g_m r_o - 1}{r_o C_{bl}}t}] \end{aligned} \tag{2.35}$$

Suppose v_a is greater than zero. Predictably, v_o will increase with time when cross-talk coupling does not exist, i.e., $\Delta v_o/\Delta t \geq 0$. Conversely, with larger v_b

the weak sensing tends to fail because its bitline voltage difference $v_o(t)$ drops with time, i.e., $\Delta v_o(t)/\Delta t < 0$. By differentiating Eqn. (2.35), the slope of $v_o(t)$ gives

$$S(t) = \frac{1}{2} \cdot (g_m r_o - 1) \cdot [\frac{v_a + v_b}{r_o C_{bl}} \cdot e^{\frac{g_m r_o - 1}{r_o C_{bl}} t} + \frac{v_a - v_b}{r_o (2C_{cpl} + C_{bl})} \cdot e^{\frac{g_m r_o - 1}{r_o (2C_{cpl} + C_{bl})} t}] \quad (2.36)$$

In order to avoid failure, $S(t)$ has to be greater than 0. As a result, v_a and v_b should satisfy the following inequality

$$\frac{v_a}{v_b} > \frac{(2C_{cpl} + C_{bl}) \cdot e^{\frac{g_m r_o - 1}{r_o C_{bl}} t} - C_{bl} \cdot e^{\frac{g_m r_o - 1}{r_o (2C_{cpl} + C_{bl})} t}}{(2C_{cpl} + C_{bl}) \cdot e^{\frac{g_m r_o - 1}{r_o C_{bl}} t} + C_{bl} \cdot e^{\frac{g_m r_o - 1}{r_o (2C_{cpl} + C_{bl})} t}} \quad (2.37)$$

When the initial input v_a is equal to v_b, from Eqn. (2.35) the strong pair output follows

$$v_o = v_b e^{\frac{g_m r_o - 1}{r_o (C_{bl} + 2C_{cpl})}} \quad (2.38)$$

Since the coupling disappears when the outputs of the strong pairs reach maximum supply rails as shown in **Fig. 2.18**, the total coupling time approximates to the time from Eqn. (2.38) when $v_o = V_{dd}$.

$$t_{cpl} = \frac{r_o (C_{bl} + 2C_{cpl})}{g_m r_o - 1} \ln \frac{V_{dd}}{v_b} \quad (2.39)$$

By taking t_{cpl} into inequality (2.37), the pass/fail boundary for v_a is found

$$v_a > v_b \cdot \frac{(2C_{cpl} + C_{bl}) \cdot (V_{dd}/v_b)^{\frac{2C_{cpl}}{C_{bl}}} - C_{bl}}{(2C_{cpl} + C_{bl}) \cdot (V_{dd}/v_b)^{\frac{2C_{cpl}}{C_{bl}}} + C_{bl}} = v_b \cdot f(C_{bl}, C_{cpl}, v_b) \quad (2.40)$$

v_a should be larger than the boundary value to maintain a correct outcome. The outcome of function $f(C_{bl}, C_{cpl}, v_b)$ rises with increasing v_b. Eventually, when v_b approaches V_{dd}:

$$v_{a,max} > f(C_{bl}, C_{cpl}, v_b)|_{v_b = V_{dd}} \cdot v_b = \frac{C_{cpl}}{C_{bl} + C_{cpl}} \cdot V_{dd} \quad (2.41)$$

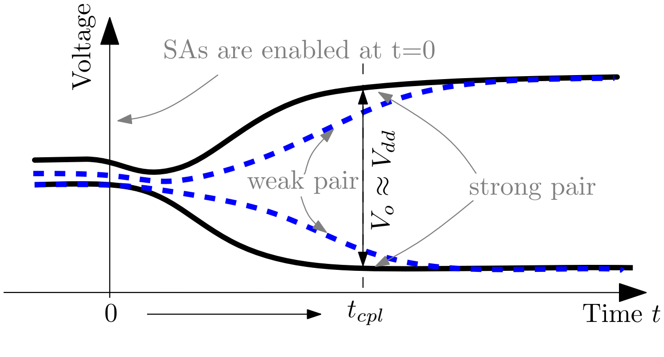


Fig. 2.18: Definition of coupling time t_{cpl}

As long as $v_a > v_{a,max}$, cross-talk coupling will not increase failures no matter what value v_b is. The result also implies that the safe margin for v_a is independent of the design parameters of the latched sense amplifier since g_m, r_o disappear in inequality (2.40).

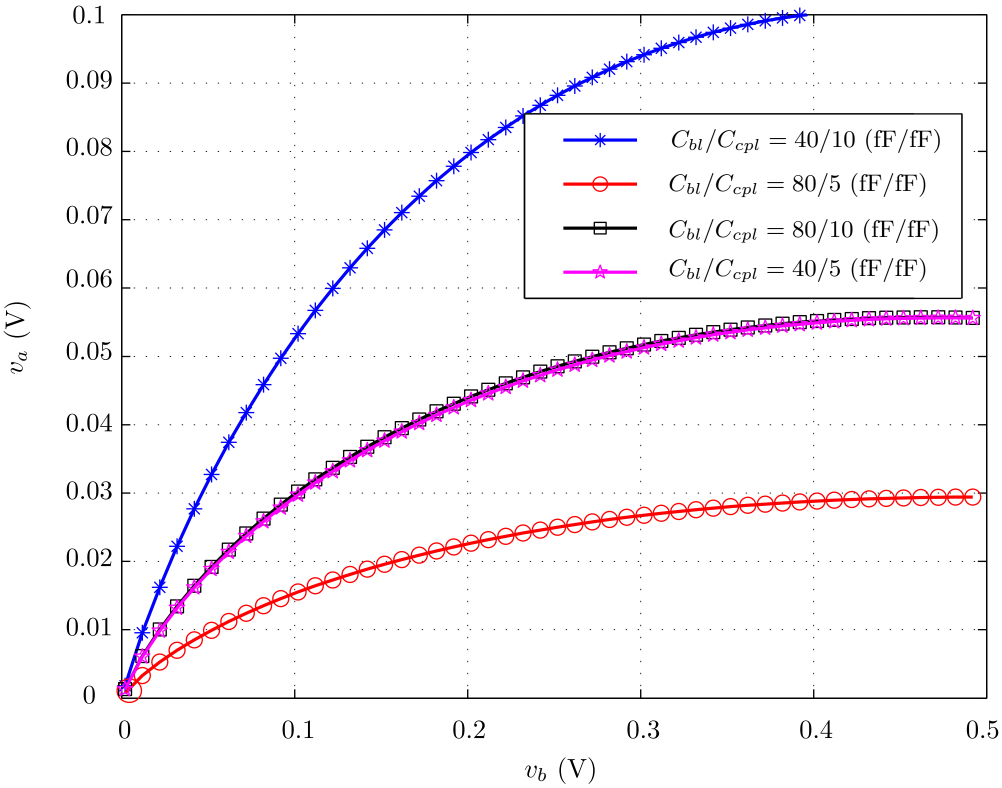


Fig. 2.19: Calculated coupling error boundary for latched sense amplifiers. When v_a is located above these curves, there will be no failure due to crosstalk coupling.

Calculated margin curves from inequality (2.40) for different bitline and sense amplifier parameters are illustrated in **Fig. 2.19**. To avoid failures by cross-talk coupling, v_a should be chosen from the region above these curves. From inequality (2.40) the safe boundary is only related to the initial voltage of the strong sensing pair v_b and capacitor ratio C_{bl}/C_{cpl}. Therefore, the third and fourth curves with the same capacitor ratio $C_{bl}/C_{cpl} = 8$ stay together while larger and smaller C_{bl}/C_{cpl} ratio move the curve down and up, respectively. The figure implies that when C_{bl}/C_{cpl} becomes smaller, post-sensing coupling will be much severer, and thus the safe margin of v_a has to be elevated.

The theoretical calculations are also compared to SPICE simulations as shown in **Fig. 2.20**. With the same circuit setup as **Fig. 2.17**(a) and by sweeping v_a the pass/fail boundary corresponding to a given v_b can be found. Evidently, the SPICE outcomes show excellent agreement to the theoretical calculations.

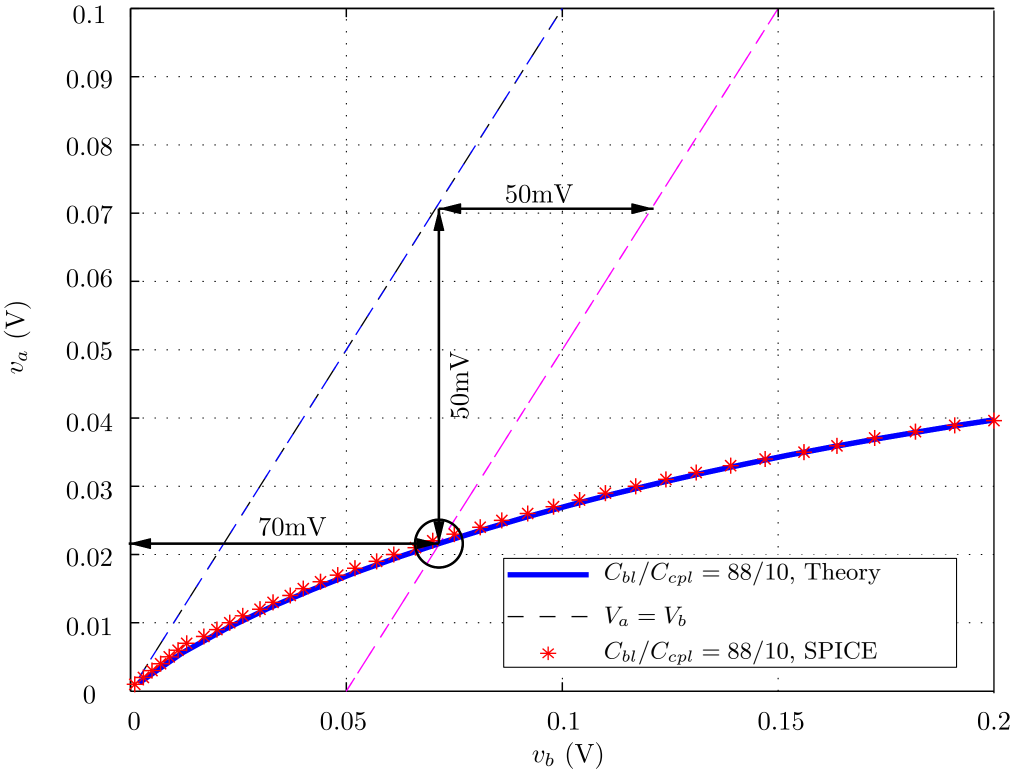


Fig. 2.20: Comparison of SPICE simulated and theoretically calculated boundary with latched sense amplifiers. An example of using this curve to facilitate the design of required V_{sign} in consideration of coupling and sense amplifier input offset V_{os} is also shown: with 50mV sense amplifier offset and worst pattern condition, the required minimum V_{sign} is 70mV.

For DRAM core array design, inequality (2.40) also provides a reliable reference on determining the minimum required voltage difference V_{sign} in consideration of both sense amplifier mismatch equivalent input offset V_{os} and post-sensing coupling.

Example

Suppose input offset V_{os} of sense amplifiers is 50mV (5σ) and the array is accessed with worst case pattern that generates smallest but uniform V_{sign} for all sensing pairs. With 50mV V_{os} the maximum effective voltage difference will be $|V_{sign}| + 50mV$ and the minimum will be $|V_{sign}| - 50$mV. Therefore, the points positioned at ($|V_{sign}| + 50$mV, $|V_{sign}| - 50$mv) correspond to (v_b, v_a). $v_a = v_b$ is shifted to the right for 50mV and the crossing point of the line and the coupling boundary curve is determined. The x coordinate of the crossing point is the required V_{sign} to avoid coupling failure. In this case, V_{sign} should be greater than 70mV, which is 1.4 times of V_{os}. Similarly, when the minimum V_{sign} is given, the required maximum V_{os} can also be estimated by this plot.

Similar to coupling analysis with OTA type sense amplifiers, from Eqn. (2.35) the equivalent initial input voltage of the weak pair can be obtained.

$$\begin{aligned} v_a' &= v_a - \frac{e^{\frac{g_m r_o - 1}{r_o C_{bl}} t} - e^{\frac{g_m r_o - 1}{r_o (2C_{cpl} + C_{bl})} t}}{e^{\frac{g_m r_o - 1}{r_o C_{bl}} t} + e^{\frac{g_m r_o - 1}{r_o (2C_{cpl} + C_{bl})} t}} \cdot v_b \\ &= v_a - K_{cpl} \cdot v_b \end{aligned} \tag{2.42}$$

K_{cpl} is a function of coupling time t_{cpl} that is determined by the strong pair from Eqn. (2.39). By taking t_{cpl} into Eqn. (2.42), K_{cpl} is transformed into a function with v_b as the input variable.

$$K_{cpl} = \frac{(V_{dd}/v_b)^{\frac{2C_{cpl}}{C_{bl}}} - 1}{(V_{dd}/v_b)^{\frac{2C_{cpl}}{C_{bl}}} + 1} \tag{2.43}$$

K_{cpl} versus v_b is plotted in **Fig. 2.21**. When v_b is close to zero, K_{cpl} becomes larger. In case the worst pattern is applied to the array, because of the sense amplifier mismatch equivalent input offset V_{os}, most pairs will have voltage amplitudes around V_{sign} while some minorities will have voltage amplitudes smaller than V_{sign}. Probably, the minorities will become the victims of the crosstalk coupling.

It is noticed that K_{cpl} versus v_b is more important for latched sense amplifiers. Since for OTA style sense amplifiers another latch circuit is used to sample the magnified voltage and memorize the valid output, for OTA sense amplifiers K_{cpl} versus sampling time t is valuable.

2.3.3 Array structures and post-sensing coupling

As has been shown in the previous analysis, array structures with OTA style sense amplifiers are insensible to cross-talk coupling since bitlines are isolated from sense amplifier outputs where coupling happens. On the contrary for array structures with latched type sense amplifiers it is crucial to suppress cross-talk coupling.

In the above calculations, a folded bitline structure was used as an example. For open bitline arrays, since each bitline has two active neighbors, the strength of cross-talk coupling will then be doubled. **Fig. 2.22** compares a folded bitline array to two different open bitline arrays with the same bitline pitch. By observation, the open bitline structures can be completely fit into the folded bitline setup, so that the equations obtained to analyze the coupling in folded bitline array can also be applied to open bitline arrays. Two different sense amplifier arrangements for open bitline array are shown in **Fig. 2.22**(b). In both open bitline structures the true and complementary bitlines are seriously coupled by their neighboring bitlines. In multiple twisted folded bitline arrays, the post-sensing coupling can

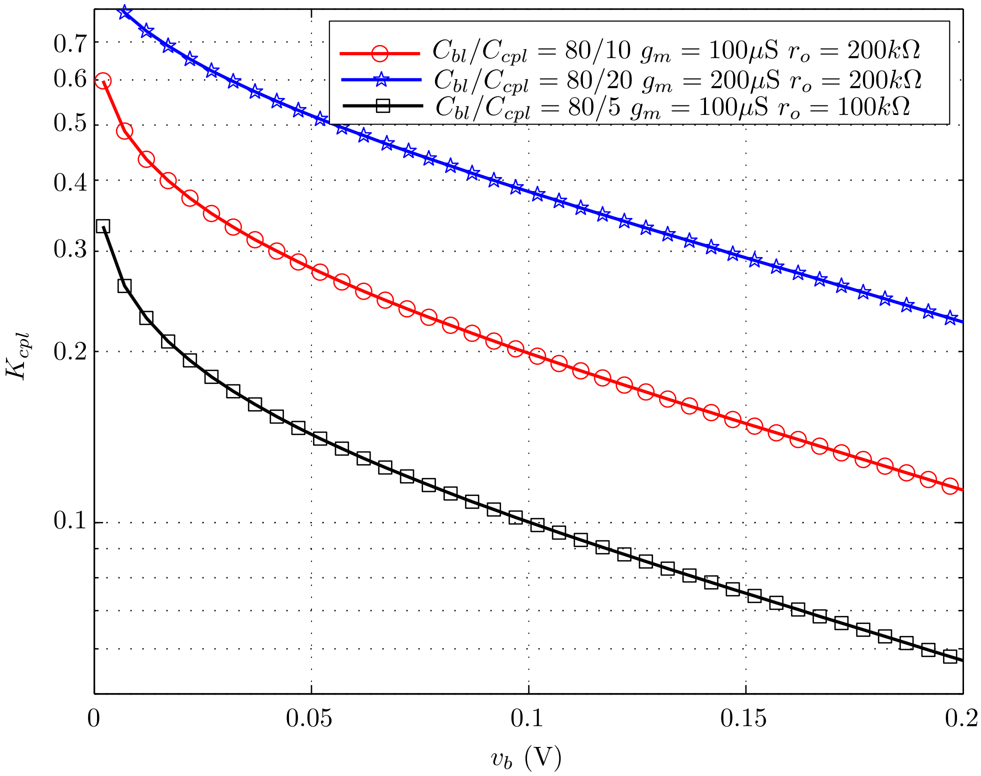


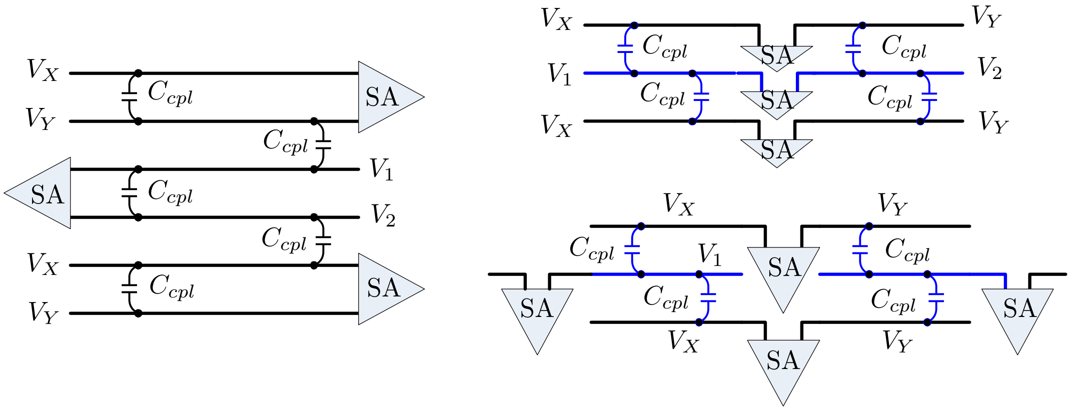
Fig. 2.21: K_{cpl} vs. v_b for the weak pair with latched sense amplifiers

be further reduced thanks to the ability of the arrays in transforming differential coupling into common mode coupling. Post-sensing coupling in different arrays exhibits the same dependency on the array structures as V_{sign}: the open bitline arrays are worse than the folded bitline arrays while both are inferior to the multiple twisted folded bitline arrays. If open bitline array has to be used for $6F^2$ or $4F^2$ cell arrays, the bitline shielding becomes the ultimate measure that has to be be taken. However, since bitline to bitline capacitance can never be completely eliminated even with bitline shielding, the above worst case model is still valid for future DRAM core design.

Suppose $C_{bl,0}$, C_{bl2bl} are bitline to ground and bitline to bitline capacitance for different arrays. **Table 2.3.3** summarizes C_{bl}, C_{cpl} in Eqn. (2.42), Eqn. (2.43) from **Fig. 2.17**.

Table 2.2: Equivalent capacitance C_{bl} and C_{cpl} for folded and open bitline arrays

Array	C_{bl}	C_{cpl}
Folded bitline array	$C_{bl,0} + 2C_{bl2bl}$	C_{bl2bl}
Open bitline array	$C_{bl,0}$	$2C_{bl2bl}$



(a) Folded bitline array

(b) Open bitline array

Fig. 2.22: Coupling capacitance comparison between folded and open bitline arrays with the worst coupling pattern

2.4 Summary

In this chapter, cell access speed and obtainable maximum bitline voltage difference V_{sign} during pre-sensing are studied in the presence of array parasitics for different kinds of arrays. By charge conservation, the metrics to acquire the bitline voltages for different arrays are formulated. With their help, V_{sign} for each bitline pair in an array with any data patterns can be evaluated with both high precision and speed. Post-sensing coupling effects between bitlines are modeled for both OTA type sense amplifiers and latched sense amplifiers. The bitline to bitline capacitance is found to be the most crucial parasitic parameter in array design. In total, the equations and models provide a guide for DRAM array/sense amplifier design in consideration of array parasitic resistance/capacitance, coupling and sense amplifier offset. Based on these important outcomes, yield estimation and analysis will be introduced and formulated in the following chapters.

Chapter 3

DRAM Sense Amplifier and Sensing Techniques

3.1 Introduction

DRAM sense amplifiers are used to 'sense' small changes caused by wordline activation. These changes can be in the form of voltage, current and charge as discussed in Section 1.4. Like other sensing circuits in SRAM, non-volatile memory (flash memory) and sensors, first of all DRAM sense amplifiers must have the ability to detect and amplify these changes within finite time. Besides that, the sense amplifiers in DRAM also function as buffers that can memorize thousands of bits obtained in a single sensing so as to improve IO throughput of a memory chip. In addition, the lost information due to the inherent destructive sensing must be restored with the help of sense amplifiers. In short, they provide three basic functions: sensing, latching and refreshing as shown in **Fig. 3.1**.

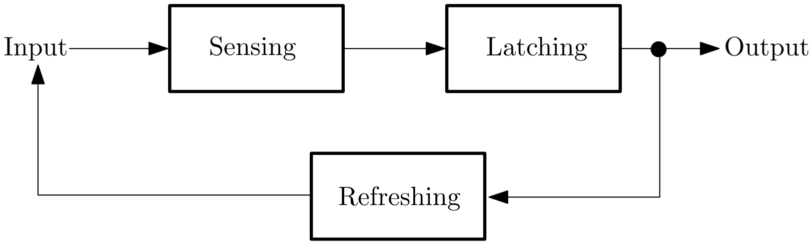


Fig. 3.1: Functions of a DRAM sense amplifier

The three functions take place successively in that latching and refreshing depend on the outcome from sensing. Generally latching occurs no later than refreshing because latching is more important in consideration of bit stream throughput. According to the listed functions, DRAM sense amplifiers can be built in complex or simple style in terms of speed, area and power consumption. However, a cross coupled transistor pair has become a necessary part for all kinds of DRAM sense amplifiers since it can implement the required latch function with least area cost.

In this chapter, different sense amplifiers and sensing schemes will be discussed. As voltage sensing is the most popular and simplest sensing style in DRAM, latched sense amplifiers in CMOS technology is the focus. Its post-sensing speed, power efficiency and other related issues will be discussed in detail. Other sensing techniques and sense amplifiers are also discussed in comparison with CMOS latched sense amplifiers with voltage sensing scheme. It will be shown that CMOS latched sense amplifier with mid-level sensing can achieve the best balances between yield, area, power and control effort.

3.2 Low-, Mid- and High-Level Sensing

3.2.1 NMOS sense amplifiers

The first sense amplifier used for a single transistor DRAM cell was demonstrated in [38] as shown in **Fig. 3.2** . It consists of equalization switch, n-latch pair and pull up transistors. NMOS technology dominated the semiconductor industry at that time and an open bitline array was used in the design to ease cell arrangement. The control signal $\overline{\phi_1}$ enables the n-latch pair to detect the input voltage difference and ϕ_{1d} enables the pull up transistors to bring one bitline to $V_{dd} - V_{thn}$. Thus the cell voltage is restored. This sense amplifier suffers from cell voltage loss, i.e., the n- pull up transistors can only raise the bitline voltage to $V_{dd} - V_{thn}$ instead of V_{dd} with non-boosted gate control voltage. Interestingly, in equalization the bitline pairs are balanced by the EQ switch to a voltage level that is not accurately defined. According to the bitline voltage that is at either $V_{dd} - V_{thn}$ or 0 when amplification is completed, the equalization voltage is around $(V_{dd} - V_{thn})/2$.

The two cross-coupled transistors in **Fig. 3.2** will be in saturation when they are enabled because their gate drain voltage differences are less than V_{thn} at the beginning. The amplification speed can be estimated by the single pole approximation [39]

$$t_{sa} \propto \frac{C_{bl}}{g_m} = \frac{C_{bl}}{K_n \mu_n C_{ox} \frac{W_n}{L_n}(V_{gs} - V_{thn})} \tag{3.1}$$

where C_{bl}, $1/g_m$ are capacitance and impedance seen from the sensing nodes. Since the bitline voltage is $(V_{dd} - V_{thn})/2$ after equalization, $V_{gs} - V_{thn}$ of the latch pair will be around $(V_{dd} - 3V_{thn})/2$. It results in slower sensing speed when V_{dd} is low or V_{thn} is high.

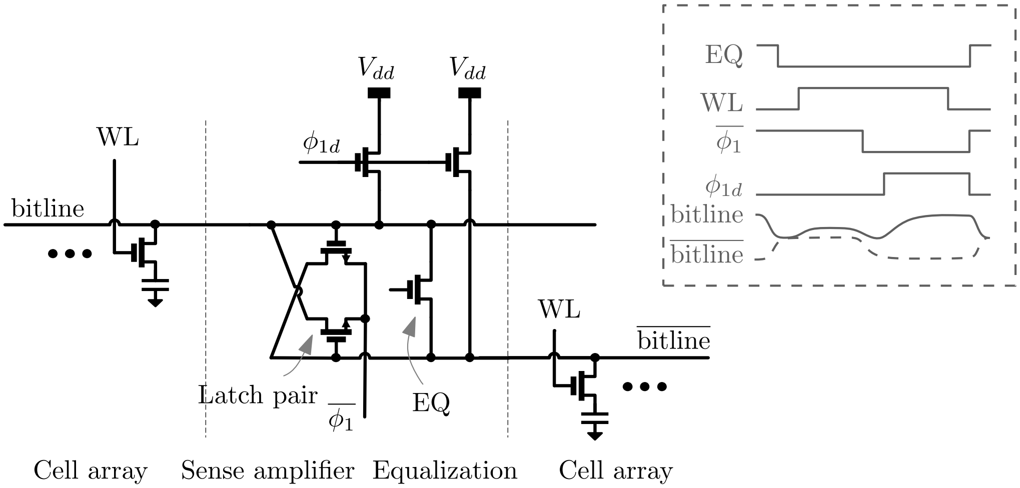


Fig. 3.2: The first sense amplifier for single transistor DRAM cell array

Because the equalization voltage is close to the middle level of bitline high and low voltages, this sensing scheme is a kind of 'pseudo' mid-level sensing. The average power consumption[1] of this sensing scheme in one access cycle is approximately

$$P_{sens,mid} = \frac{1}{T}\int VIdt = \frac{(C_{bl} + C_s/2)(V_{dd} - V_{thn})V_{dd}}{2T} \tag{3.2}$$

To improve the sensing speed, the high level-sensing scheme was introduced later in a 16-kbit DRAM design [40]. In fact there is no change in sense amplifier circuit but the bitline pair is pre-charged to $V_{dd} - V_{thn}$ instead of $(V_{dd} - V_{thn})/2$ during equalization. As a result, the higher bitline operating voltage raises the overdrive voltage $V_{gs} - V_{thn}$ in Eqn. (3.1) to $V_{dd} - 2V_{thn}$, enhancing the sensing speed.

However, since in the new scenario both bitlines are charged to around supply voltage V_{dd}, the average power consumption per sense amplifier in one access cycle gives

$$P_{sens,high} = \frac{1}{T}\int VIdt = \frac{(C_{bl} + C_s/2)(V_{dd} - V_{thn})V_{dd}}{T} \tag{3.3}$$

In contrast with the 'pseudo' mid-level sensing scheme, average power consumption of high-level sensing is doubled since it wastes part of energy in raising both bitlines to $V_{dd} - V_{thn}$. This agrees with the power speed trade-off usually seen in circuit design.

[1] In this chapter, unless specified, '1' and '0' cells are assumed to appear with 50% probability in average power consumption calculation.

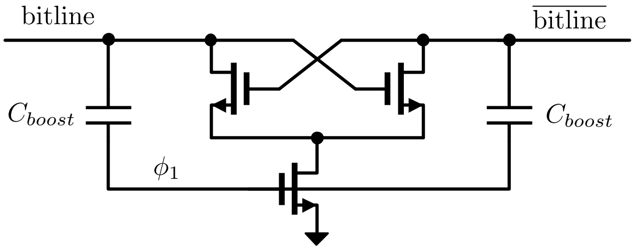


Fig. 3.3: By placing two capacitors between bitlines and ϕ_1, the bitline voltage becomes higher when the n-latch pair is enabled [41].

Further speed enhancement can be achieved by boosting bitline voltage to even higher level by placing capacitors between the bitlines and the latch enable signal as shown in **Fig. 3.3** [41]. Like high-level sensing, in this scenario bitline pair will be first charged to $V_{dd} - V_{thn}$ in equalization. As the charge in C_{boost} can not change in short time, when ϕ_1 goes high, the bitline voltage will be pumped up to a voltage higher than $V_{dd} - V_{thn}$ and with such aid the sensing speed can be further improved.

As demonstrated, in NMOS technology the sensing speed of sense amplifiers is the major concern. Moreover, these sense amplifiers suffer from great cell voltage loss due to the n-pull-up transistors. Since the on resistances of pull-up transistors in **Fig. 3.2** rise quickly with the increase of bitline voltage, the required bitline charging time is hard to determine. Though the high-level sensing scheme can improve post-sensing speed, it still takes long time to charge bitlines to $V_{dd} - V_{thn}$ in equalization. As a consequence, CMOS technology became more appealing for 64-kbits DRAM generation and beyond [42].

3.2.2 CMOS sense amplifiers

The first DRAM CMOS sense amplifier appeared in 1984 [43] as shown in **Fig. 3.4**. Compared to earlier designs, the pull-up transistors are replaced with a p-latch pair that can provide low on-resistance in pulling the bitline voltage to V_{dd}. The sense amplifier is shared by arrays on both sides by turning on left or right isolation switches. As one bitline will be at V_{dd} and another at ground when sensing is accomplished, after equalization both bitlines will be at $V_{dd}/2$ and thus this sensing scheme is a perfect mid-level sensing, which benefits from lower power consumption compared to high or low level sensing. The sensing speed of the mid-level sensing scheme with CMOS sense amplifier is found to be comparable to a high-level sensing scheme because the valid g_m in Eqn. (3.1) is the total sum of transconductance of both n- and p-sensing transistors. To trace the cell voltage fluctuation better, a pair of dummy cells are used to generate a reference cell voltage that is close to $V_{dd}/2$ when EQ is on. The reference cell voltage provides a precise comparable voltage level for active cells by compensating cell voltage

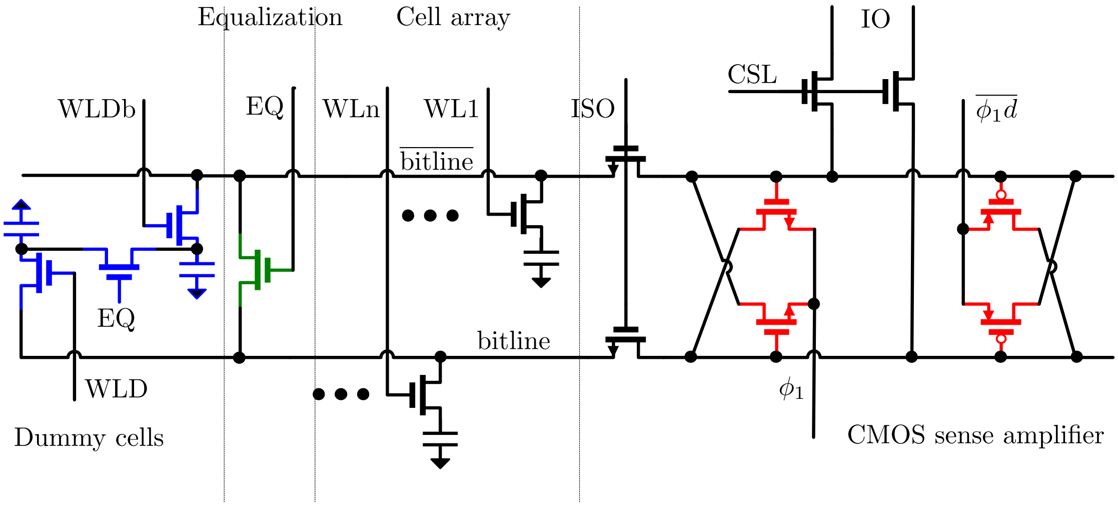


Fig. 3.4: The first CMOS DRAM sense amplifier with mid-level sensing scheme [43] in a folded bitline array

loss. However, additional power and control effort have to be put on the dummy cells.

The dummy cells in **Fig. 3.4** are not absolutely necessary when equalization voltage can be precisely defined. **Fig. 3.5** gives another CMOS sense amplifier used widely in nowadays' DRAM products. The dummy reference cells are replaced with two switches that connect the bitlines to $V_{dd}/2$ when equalization is enabled.

Extra attention should be paid to the isolation devices in **Fig. 3.5** and **Fig. 3.4**. Due to these devices the sense amplifiers can be shared by two arrays on the left and right in folded bitline arrays. To minimize the area consumption, even in CMOS technology most switching devices are designed in n-transistors that have larger carrier mobility. When a bitline is needed to be pulled to V_{dd} by the p-transistors, the isolation device will form a voltage clamp that limits the maximum bitline voltage to $V_g - V_{thn}$. To alleviate the problem, the gate voltage V_g is required to be higher than $V_{dd} + V_{thn}$ and sometimes this causes reliability problems.

Indeed, not only the switching transistors in sense amplifiers but also the cell transistors are affected by the reduced V_{gs} caused by the change of the bitline pre-charge voltage. For example, in mid-level sensing scheme, the half V_{dd} bitline equalization voltage results in a larger on resistance for '1' cells, and therefore the pre-sensing speed for these cells is slower. Typical solutions to cope with such a problem include boosting the wordline voltage above V_{dd} (boosted wordline technique [34]) or using low-level sensing scheme in which the equalization voltage is around ground potential [44].

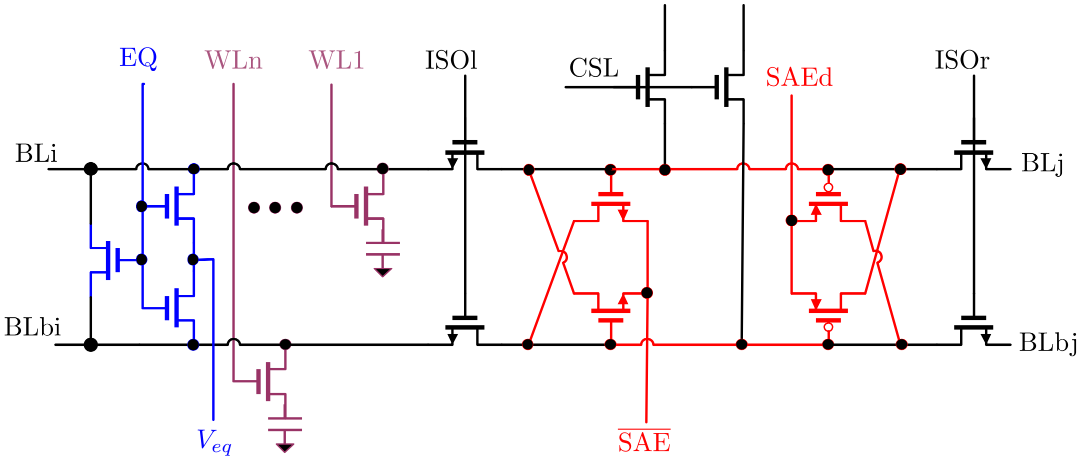


Fig. 3.5: A mature CMOS DRAM sense amplifier used in folded bitline array with isolation devices

In this section, the evolution of DRAM sense amplifiers together with low-, mid- and high-level sensing schemes is depicted. Obviously, CMOS sense amplifiers and mid-level sensing have advantages in power consumption, cell voltage restoration and sensing speed. As a consequence, they are widely used in all kinds of DRAM products up until today. It should also be noticed that the sensing scheme is flexible according to the technology and application such as available devices and voltage sources, power consumption, area, cost. For example, in publication [45] a 2/3 V_{dd} pre-charging scheme is used with CMOS sense amplifier and p-transistor cell array to achieve high speed operation. However, these designs are only for special technologies and specific products. In this thesis, since the yield analysis for normal DRAM core design is the final goal, emphasis will be put onto the CMOS sense amplifier with mid-level sensing due to their wide applications.

3.3 CMOS Latched Sense Amplifier

In this section, the CMOS latched sense amplifier with mid-level sensing scheme will be analyzed in detail with respect to its speed and power efficiency. As mentioned earlier, sensing, latching and refreshing are necessary functions for DRAM sense amplifiers. In order to make sense amplifiers power and area efficient, they are implemented by n- and p-latch pairs as shown in **Fig. 3.6**.

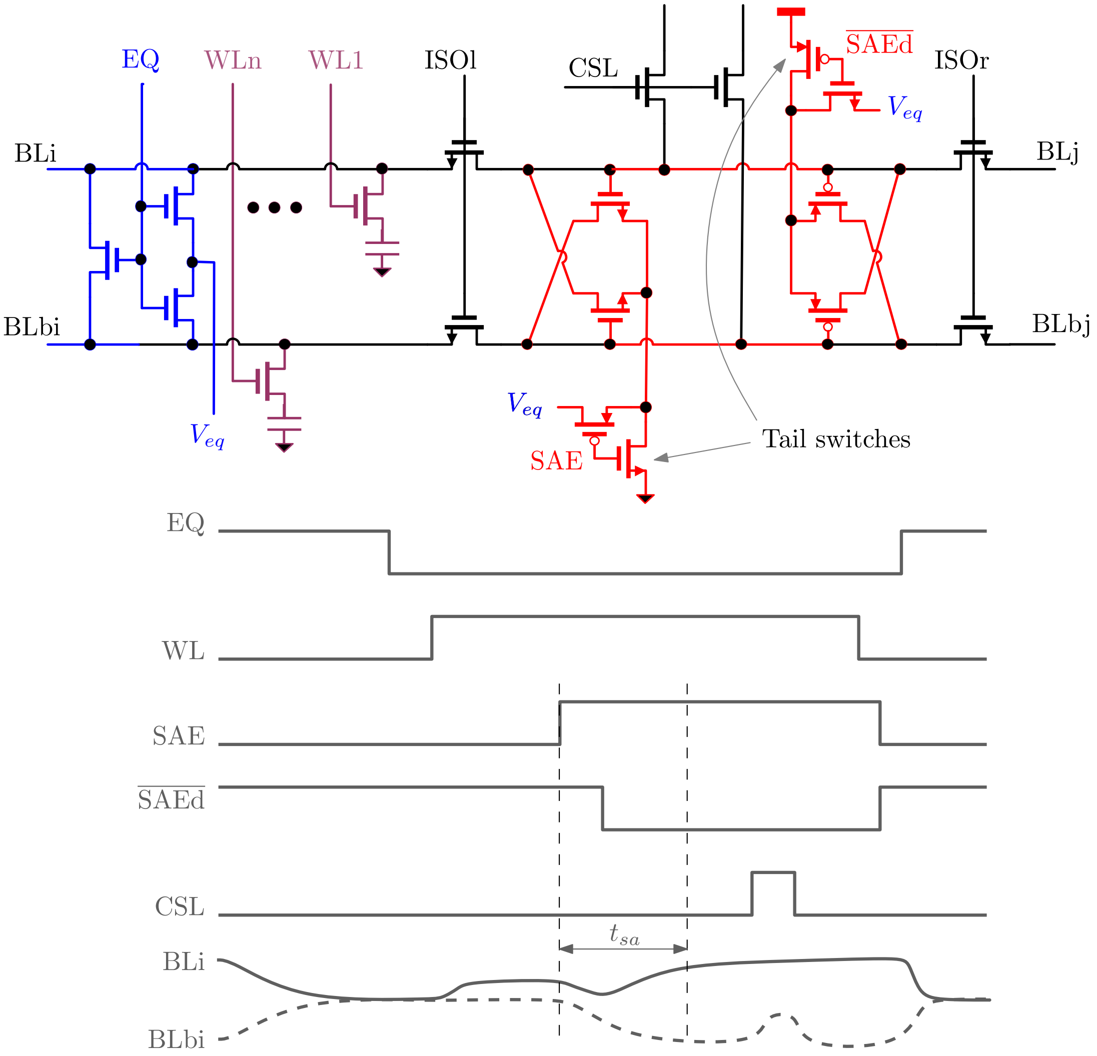


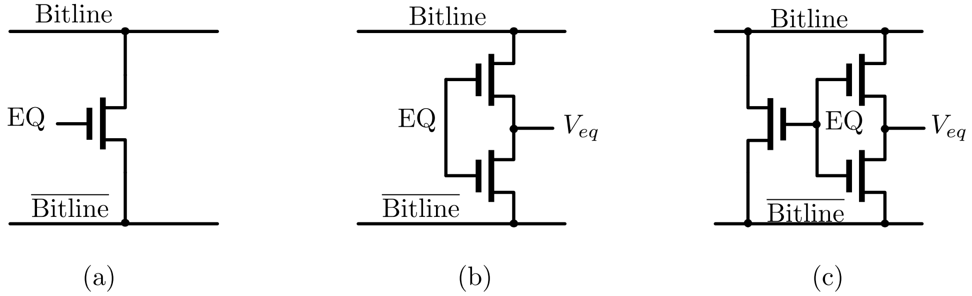
Fig. 3.6: Complete CMOS DRAM sense amplifier in mid-level sensing scheme

3.3.1 Operation and sensing speed

The entire sensing process is composed of three stages as shown in **Fig. 1.4**: equalization, pre-sensing and post-sensing. EQ is turned on to equalize and charge both bitlines to V_{eq}, which is half V_{dd} for mid-level sensing. Then a wordline is activated to release the cell charge to the bitline capacitance. Within a certain period of time, a small voltage difference V_{sign} is developed between a pair of bitlines. SAE and $\overline{\text{SAEd}}$ are enabled to switch on the CMOS latched sense amplifier. When amplification is accomplished, the column select lines (CSLs) will go high to connect the local bitlines to the global bitlines so that the sensing outcomes can be transmitted through the global bitlines to the succeeding circuits. Since the high capacitance global bitlines are pre-charged to V_{dd}, the instantaneous connection will temporarily raise the local bitline voltage to an higher voltage level depending on the sense amplifier output impedance and the capacitance ratio of the local and global bitlines.

Equalization speed

The first step of sensing is equalization. Bitline pairs can be equalized by three means: a) Connecting true and complementary bitline; b) Charging true and complementary bitlines to equalization voltage V_{eq}; c) Both a) and b) together. The corresponding circuit implementations are shown in **Fig. 3.7**.


Bitline
EQ
$\overline{\text{Bitline}}$
Bitline
EQ
V_{eq}
$\overline{\text{Bitline}}$
Bitline
EQ
V_{eq}
$\overline{\text{Bitline}}$

(a) (b) (c)

Fig. 3.7: Three different equalizer implementations

In (a) when EQ is high, positive charge will flow from the 'high' bitline to the 'low' bitline. The charge redistribution process is identical to the pre-sensing process as mentioned in Section 2.1. In analogy with Eqn. (2.5), the equalization time required for a given settling error V_{err} between true and complementary bitlines is

$$t_{eq} = \ln(\frac{|V_{dd} - V_{ss}|}{V_{err}}) \cdot \frac{R_{on}}{(1/C_{bl} + 1/C_{bl})} = \ln(\frac{V_{dd}}{V_{err}}) \cdot \frac{R_{on}C_{bl}}{2} \tag{3.4}$$

where R_{on} is the on-resistance of the equalization transistor. Since R_{on} is a function of both drain and source voltage, maximum R_{on} appears when the source voltage of the equalization transistor approaches $V_{dd}/2$.

$$R_{on,max} = \frac{1}{K_n \frac{W_n}{L_n}(V_g - V_{dd}/2 - V_{thn})} \tag{3.5}$$

V_g is the gate voltage of the equalization transistor in 'on' state. By taking the maximum R_{on} into Eqn. (3.4), the worst equalization time delay becomes

$$t_{eq,max,a} = \ln(\frac{V_{dd}}{V_{err}}) \cdot \frac{C_{bl}}{2 \cdot K_n \frac{W_n}{L_n}(V_g - V_{dd}/2 - V_{thn})} \tag{3.6}$$

In (b) bitlines are charged by the transistors in linear region to V_{eq}. As these transistors are connected to V_{eq}, currents through them follow

$$I_1 = K_n \frac{W_n}{L_n}(V_g - V_{eq} - V_{thn})(V_{bl} - V_{eq}) = C_{bl} \cdot \frac{dV_{bl}}{dt}, \text{ for 'high' bitline} \tag{3.7}$$

$$I_2 = K_n \frac{W_n}{L_n}(V_g - V_{bl} - V_{thn})(V_{eq} - V_{bl}) = C_{bl} \cdot \frac{dV_{bl}}{dt}, \text{ for 'low' bitline} \tag{3.8}$$

After sensing, the bitline voltage is either V_{dd} or V_{ss}. For the 'high' bitline, the source voltage of the equalization transistor is at V_{eq}. The time taken to settle the bitline voltage within $V_{eq} + V_{err}$ gives

$$t_{eq,high,b} = \ln \frac{V_{dd}}{2V_{err}} \cdot \frac{C_{bl}}{K_n \frac{W_n}{L_n}(V_g - V_{dd}/2 - V_{thn})} \tag{3.9}$$

Similarly, for the 'low' bitline the equalization transistor's drain voltage is fixed and its source voltage is floating. The corresponding worst settling time

$$t_{eq,low,b} = t_{eq,high,b} - \ln \frac{V_g - V_{thn}}{V_g - V_{thn} - V_{dd}/2} \cdot \frac{C_{bl}}{K_n \frac{W_n}{L_n}(V_g - V_{dd}/2 - V_{thn})} \tag{3.10}$$

In comparison with equalization time in implementation (a),

$$t_{eq,a} < t_{eq,high,b} \text{ ,and } t_{eq,low,b} < t_{eq,high,b} \tag{3.11}$$

Since in **Fig. 3.7**(b) both bitlines are needed to be charged to V_{eq}, $t_{eq,high,b}$ is taken as its equalization time. Evidently, implementation (b) is slower than (a). In fact, the speed problem of (b) also comes from the output resistance of the voltage source V_{eq}. In practical design, this resistance is rather large to prevent large static current caused by wordline to bitline electrical short failure resulting from the technology defects, which can be repaired by redundant bitline pair. Therefore, the single equalization transistor in (a) is necessary to guarantee enough equalization speed. Hence when area permits, (c) is favored for productive DRAM design as shown in **Fig. 3.5**.

Post-sensing speed

To investigate detailed operations of latched CMOS sense amplifier, **Fig. 3.8** is introduced by neglecting the on-resistance of the tail switches. **Fig. 3.9** shows the operating region of the four sensing transistors when they are enabled according to different bitline voltages. As the name suggests, in mid-level sensing since V_{BL}, V_{BLb} are around half V_{dd}, the four sensing transistors M_{na}, M_{nb}, M_{pa}, M_{pb} will be in saturation. Suppose the post-sensing delay is dominated by the time when

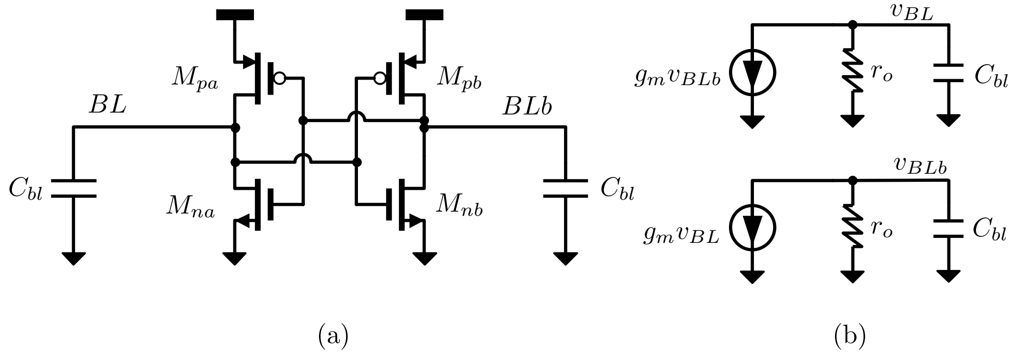


Fig. 3.8: CMOS latched sense amplifier and its small signal circuit model

sensing transistors are in saturation, according to the small signal circuits in **Fig. 3.8**(b) the following differential equations can be drawn

$$\begin{cases} C_{bl} \cdot dv_{BL}/dt + v_{BL}/r_o + g_m \cdot v_{BLb} = 0 \\ C_{bl} \cdot dv_{BLb}/dt + v_{BLb}/r_o + g_m \cdot v_{BL} = 0 \end{cases} \tag{3.12}$$

With initial conditions $V_{BL} - V_{BLb} = v_{BL} - v_{BLb} = V_{sign}$, Eqn. (3.12) gives the sensing delay

$$t_{sa} = \frac{r_o C_{bl}}{g_m r_o - 1} \ln(\frac{V_{out}}{V_{sign}}) \approx \frac{C_{bl}}{g_m} \ln(\frac{V_{out}}{V_{sign}}) \tag{3.13}$$

t_{sa} begins at the time when the sense amplifier is enabled as shown in **Fig. 3.6**. Because most of the time the sensing transistors are in saturation, the total post-sensing delay is approximated by t_{sa}. Eqn. (3.13) agrees well with the single pole approximation used in Eqn. (3.1), emphasizing that C_{bl} should be minimized while g_m and V_{sign} should be maximized to increase post-sensing speed. Because of technology shrinking and area constraint, the total sum of g_m from n- and p-sensing transistors in Eqn. (3.13) will not change much from generation to generation, and is comparable with g_m of the n-sensing transistors in a high-level sensing scheme. This explains the speed advantage of CMOS latched sense amplifiers with mid-level sensing as mentioned in reference [43]. Another interesting outcome is that when V_{sign} is expressed by the array transfer ratio K_t as in Eqn. (2.9) and the maximum output voltage is supposed to be V_{dd}, t_{sa} seems to be independent of the supply voltage V_{dd}

$$t_{sa} = \frac{C_{bl}}{g_m} \ln(\frac{2}{K_t}) \tag{3.14}$$

However, this conclusion is not true since g_m drops with the decrease of supply voltage, and thus t_{sa} will be longer.

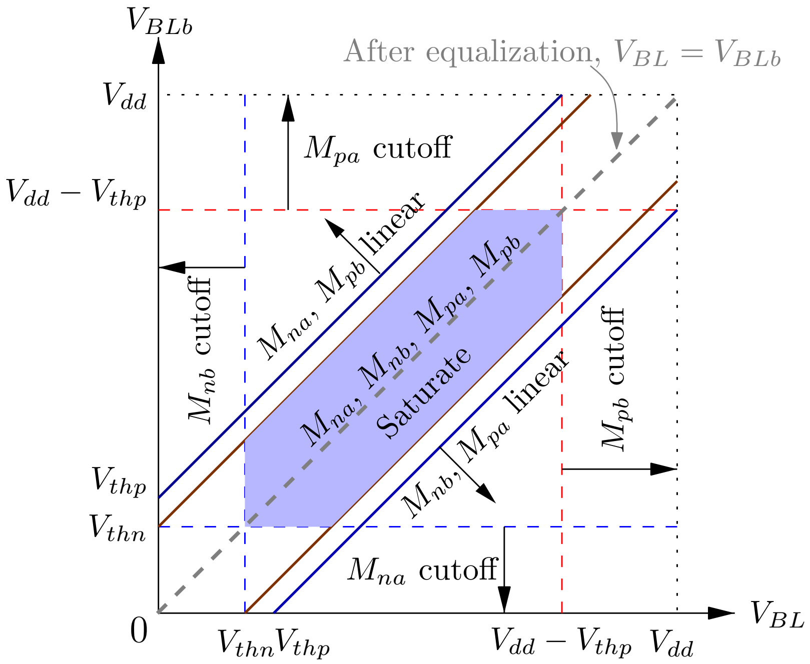


Fig. 3.9: The operating region of n- and p-transistors in Fig. 3.8 vs. bitline voltage V_{BL} and V_{BLb} (with $V_{thp} > V_{thn}$)

Sensing delay of a complete access cycle

Suppose the cell voltage V_{cell} is 0 or V_{dd}, and the post sensing maximum output is V_{dd}. The minimum required sensing delay of an access cycle can be formulated by considering the timing delay in equalization, pre-sensing and post-sensing phases. By Eqn. (3.4), Eqn. (2.5) and Eqn. (3.14) the total sensing delay is

$$\begin{aligned} t_{sens} &= t_{eq} + t_{pre} + t_{post} \\ &\approx \ln(\frac{V_{dd}}{V_{err,1}}) \cdot \frac{R_{eq}C_{bl}}{2} + \ln(\frac{V_{dd}}{2V_{err,2}}) \cdot \frac{(R_{cell} + R_{bl})}{1/C_s + 1/C_{bl}} + \ln(\frac{2}{K_t}) \cdot \frac{C_{bl}}{g_m} \end{aligned} \quad (3.15)$$

Where $V_{err,1}$ and $V_{err,2}$ are settling error for equalization and pre-sensing phases, respectively. They are usually set to a percentage of supply voltage V_{dd}. Clearly Eqn. (3.15) confirms that the most effective means to improve the sensing speed is reducing C_{bl} and R_{on} of switching transistors. Since C_{bl} here actually comprises several different parasitic capacitances, it can be written as

$$C_{bl} = n \cdot [C'_{bl} + C'_{bl2wl} + \lambda \cdot C'_{bl2bl}] \quad (3.16)$$

n is the total number of cells per bitline. λ is determined by array structures and data patterns as shown in Eqn. (2.28) and **Table 2.1**. **Table 3.1** gives some theoretical calculations as examples. The worst sensing delay happens with $\lambda = 4$ in none twist open or folded bitline arrays.

Example

Suppose $C'_{bl}+C'_{bl2wl} = 76/512\text{fF}$, $C'_{bl2bl} = 16/512\text{fF}$, $C_s = 30\text{fF}$, $N = 512$, $R_{cell} = 15k\Omega$, $R_{eq} = 2.4k\Omega$, $V_{dd} = 1.2V$, $g_m = 200\mu S$, $V_{err,1} = V_{err,2} = 1/1000V_{dd}$. Transfer ratio K_t can be obtained from Eqn. (2.28). The resulting sensing delay for each array structure is shown in **Table 3.1**.

Table 3.1: Estimated sensing speed, λ for different array structures and data pattern

Pattern 0000...0000						
Array structure	λ	t_{eq}	t_{pre}	t_{post}	t_{total}	Normalized
Folded	4	1.16ns	2.3ns	1.70ns	5.16ns	153%
MWT_1 folded	3	1.03ns	2.25ns	1.44ns	4.72ns	140%
MWT_2 folded	3	1.03ns	2.25ns	1.44ns	4.72ns	140%
Open	0	630ps	2.00ns	743ps	3.38ns	100%
Pattern 1010...1010						
Folded	2	895ps	2.19ns	1.20ns	4.28ns	127%
MWT_1 folded	2	895ps	2.19ns	1.20ns	4.28ns	127%
MWT_2 folded	3	1.03ns	2.25ns	1.44ns	4.72ns	140%
Open	4	1.16ns	2.30ns	1.70ns	5.16ns	153%
Pattern 1111...1111						
Folded	4	1.16ns	2.3ns	1.70ns	5.16ns	153%
MWT_1 folded	3	1.03ns	2.25ns	1.44ns	4.72ns	140%
MWT_2 folded	3	1.03ns	2.25ns	1.44ns	4.72ns	140%
Open	0	630ps	2.00ns	743ps	3.38ns	100%

From the table, the worst and best delay are found to be 5.16ns and 3.38ns for open bitline array when array parasitics are taken into consideration. In order to improve the worst case delay, methods to reducing bitline to bitline capacitance like bitline shielding is mandatory for open bitlines.

3.3.2 Influence of tail switches on post-sensing speed

The post-sensing delay in Eqn. (3.13) neglects two effects caused by tail switches in a CMOS latched sense amplifier as shown in **Fig. 3.6**: a) There is a switching-on time delay t_d between n- and p-sensing transistors in CMOS sense amplifiers; b) The on-resistance of the tail switches can be significant large due to their small sizes. Both effects can substantially degrade the post-sensing speed as discussed. The delay between the sensing enable signals SAEd and $\overline{\text{SAE}}$ in **Fig. 3.6** is generated due to the different path delays for the n- and p-tail switches. Since

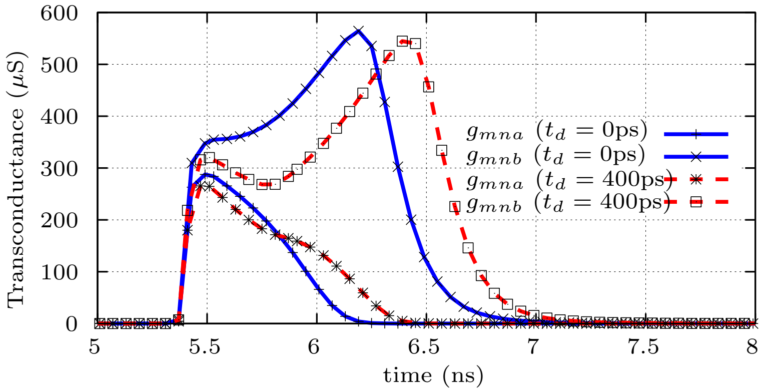


Fig. 3.10: g_m of n-sensing transistors in a CMOS latched sense amplifier with and without tail switch delay

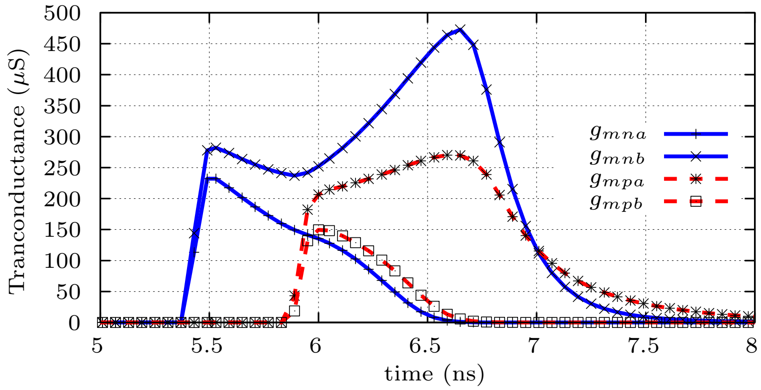


Fig. 3.11: g_m of n- and p-sensing transistors in a CMOS latched sense amplifier with 400ps delay for p-sensing transistors

from Eqn. (3.12) it is clear that the post-sensing speed depends highly on the transconductance g_m of sensing transistors, let's begin with the analysis of the g_m. By controlling the supply and ground voltage in **Fig. 3.8**, the change of transconductance of the sensing transistors corresponding to different cases can be simulated as follows.

Fig. 3.10 exhibits the simulated g_m change of the n-transistors with and without switch-on delay. g_{mna}, g_{mnb}, g_{mpa}, g_{mpb} are transconductance for the n- and p-sensing transistors M_{na}, M_{nb}, M_{pa}, M_{pb} in **Fig. 3.8**, respectively. When both n- and p-sensing transistors are triggered on simultaneously ($t_d = 0$), with the initial bitline voltage $V_{BL} < V_{BLb}$, g_{mnb} increases gradually while g_{mna} drops until one transistor steps into the cutoff region and the other into the linear region. In case the p-sensing transistors are delayed for 400ps, the common mode voltages on both sensing nodes tend to drop due to the on-current of n-sensing transistors. As a consequence, g_{mna}, g_{mnb} of the n-sensing transistors drop until

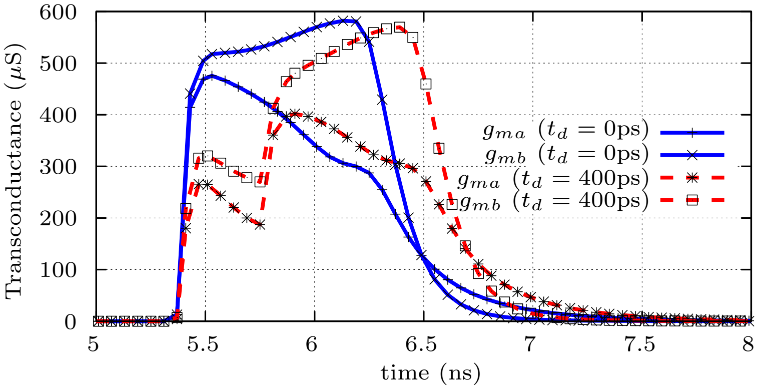


Fig. 3.12: Sum of transconductance of n- and p-sensing transistors on each side of a CMOS latched sense amplifier

the p-sensing transistors are turned on as shown in **Fig. 3.11**. At the beginning, g_m of the p-sensing transistors are zero because of the 400ps delay. As soon as the p-sensing transistors are switched on, g_m of n- and p-transistors jump higher and the sensing speeds up. **Fig. 3.12** gives the total transconductance of n- and p-sensing transistors, i.e., $g_{ma} = g_{mna} + g_{mpa}$ and $g_{mb} = g_{mnb} + g_{mpb}$. Obviously, without the switch delay g_{ma}, g_{mb} are significantly larger compared to the case with delay, and therefore the sensing speed is faster. As a conclusion, to speed up a sensing process for a CMOS latched sense amplifier, both n- and p-sensing transistors should be triggered within the achievable minimum delay.

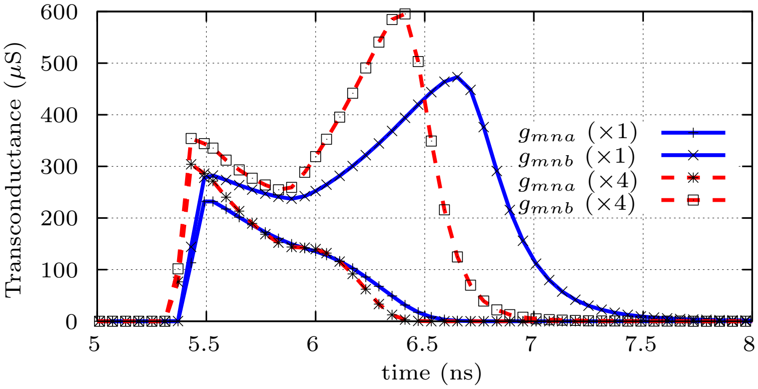


Fig. 3.13: g_m of a CMOS latched sense amplifier with $1\times$ and $4\times$ width of tail switches

Now let's take the impact of the on-resistance of the tail switches into consideration. Since they are designed in very small size to save area, their on-resistances are usually considerably large. This leads to the saturation of currents through the sensing transistors and g_m reduction as can be seen in **Fig. 3.13**. Here, only g_m of n-sensing transistors are exhibited. When the width to length ratio of the tail switch is scaled by a factor of four, g_m of n-sensing transistors will be scaled

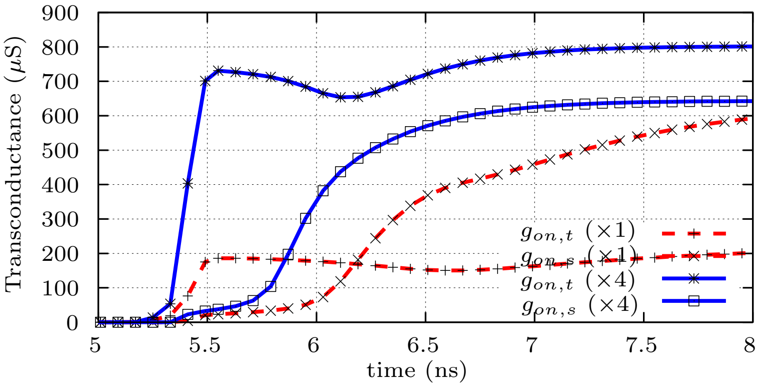


Fig. 3.14: Sensing transistor on transconductance $g_{on,s}$ and tail switch-on transconductance $g_{on,t}$ of a CMOS latched sense amplifier with 1× and 4× width of tail switches

correspondingly by a factor of 1.5 because of the square-root relationship between transconductance and transistor bias current. Both n- and p-sensing pairs are identical here in either simulation.

In addition, the overall sensing delay is actually not only determined by g_m of the sensing transistors in saturation, but also R_{on} of the tail transistors. When the sensing transistors enter linear region, the remaining time to charge the bitlines is determined by $(R_{on,t} + R_{on,s})C_{bl}$, where $R_{on,t}$, $R_{on,s}$ are the on-resistances of tail and sensing transistors in linear region. **Fig. 3.14** demonstrates the resistance change of tail switches and sensing transistors with different tail transistor widths. For convenience, transconductance is plotted with $g_{on} = 1/R_{on}$. As depicted, the sensing speed is severely degraded for the group with 1× tail switches. To minimize this effect, R_{on} of the tail switches must be several times smaller than the on-resistance of the sensing transistors in linear region.

In summary, tail switches of CMOS latched sense amplifiers have great impact on post-sensing speed and need to be carefully designed. To optimize the sensing speed of CMOS sense amplifiers, the delay between n- and p-sensing transistors should be close to 0 and R_{on} of the tail transistors should be well below one tenth of the on-resistance of the sensing transistors working in linear region.

3.3.3 Offset caused by imbalanced load capacitance

In the previous analysis the load capacitances of the true and complementary bitlines in **Fig. 3.8** are considered to be identical. However in general, this is not true. First of all, the bitline capacitance is a voltage dependent component, changing accordingly with true and complementary bitline voltages [21]; Furthermore, in recent DRAM designs dummy reference cells are removed to save area, and therefore the difference between the true and complementary bitline capacitive loads originating from the cell capacitor becomes significant.

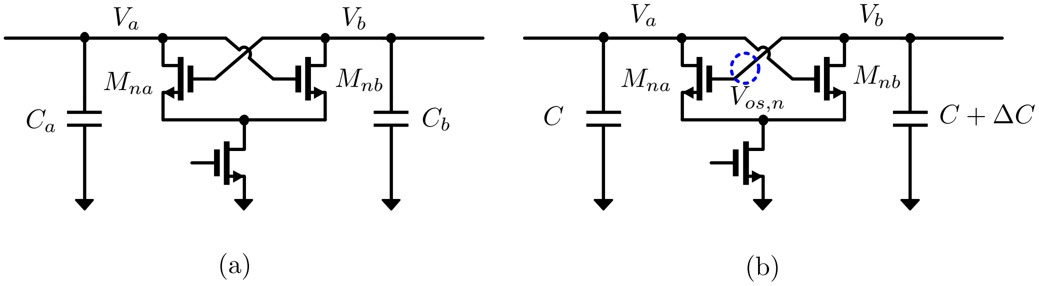


Fig. 3.15: The effect of imbalanced load capacitance (a) is equivalent to an input offset as in (b).

First consider an n-latch pair as shown in **Fig. 3.15**(a). When the latch is enabled, the currents through the load capacitances are

$$I_a = C_a \frac{dV_a}{dt} = I_{ds,mna} \text{ , and } I_b = C_b \frac{dV_b}{dt} = I_{ds,mnb} \tag{3.17}$$

Here $I_{ds,mna}$, $I_{ds,mnb}$ are currents of the saturated sensing transistors since $|V_a - V_b| < V_{thn}$. If $C_a = C_b$ and initial $V_a = V_b$, the circuit is completely symmetric and the voltages on both nodes change synchronously, i.e., $dV_a/dt = dV_b/dt$. However, once $C_a \neq C_b$, the voltage changing speed of the two output nodes will be different and the ground potential is inclined to appear on the node with smaller capacitance. In order to calibrate the capacitive imbalance, the voltage changes of either node should be identical, or

$$\frac{dV_a}{dt} = \frac{dV_b}{dt} \tag{3.18}$$

From Eqn. (3.17) the following condition has to be met

$$\frac{I_{ds,mna}}{C_a} = \frac{I_{ds,mnb}}{C_b} \tag{3.19}$$

Since the transistors are in saturation, $I_{ds,sat} \propto (V_{gs} - V_{thn})^2$. By replacing $I_{ds,mna}$, $I_{ds,mnb}$ in Eqn. (3.19), the initial voltage difference $V_a - V_b$ gives an offset

$$\begin{aligned} V_{os,n} &= \frac{\sqrt{C_a} - \sqrt{C_b}}{\sqrt{C_a}} \cdot (V_{eq} - V_{thn}) \\ &= (\sqrt{1 + \frac{\Delta C}{C}} - 1) \cdot (V_{eq} - V_{thn}) \\ &\approx \frac{\Delta C}{2C}(V_{eq} - V_{thn}), \text{ when } C \gg \Delta C \end{aligned} \tag{3.20}$$

Eqn. (3.20) suggests the imbalanced load capacitance can be equivalent to an input offset voltage $V_{os,n}$ that is applied to one input terminal of an n-latch sensing pair as shown in **Fig. 3.15**(b). With $V_{os,n}$ the two output node voltages will drop in the same speed. Since $V_{os,n}$ rises with the increase of the capacitance difference ΔC and equalization voltage V_{eq}, high-level sensing is more sensitive to load capacitance imbalance.

By the same means, the imbalanced load capacitance induced input offset in a p-latch pair can be obtained. Interestingly, it has the opposite polarity to n-latch pair

$$\begin{aligned} V_{os,p} &= -\frac{\sqrt{C_a}-\sqrt{C_b}}{\sqrt{C_a}}\cdot(V_{dd}-V_{eq}-V_{thp}) \\ &= -(\sqrt{1+\frac{\Delta C}{C}}-1)\cdot(V_{dd}-V_{eq}-V_{thp}) \\ &\approx -\frac{\Delta C}{2C}(V_{dd}-V_{eq}-V_{thp}), \text{ when } C \gg \Delta C \end{aligned} \tag{3.21}$$

As a consequence, when a CMOS latched sense amplifier is used and both n- and p-latch pairs are triggered simultaneously, the effective input offset voltage V_{os} becomes the supposition of both offsets from n- and p-latch

$$V_{os} = V_{os,n} + V_{os,p} \tag{3.22}$$

Thus V_{os} can be eliminated with $V_{eq} = V_{dd}/2$ and $V_{thn} \approx |V_{thp}|$. While it is difficult to equal threshold voltages of n- and p-transistors, the offset introduced by the imbalanced load capacitances can be minimized by proper device engineering. Therefore, simultaneously latched CMOS sense amplifiers with mid-level sensing provide not only speed advantage but also immunity against asymmetric capacitive loads.

3.3.4 Array power consumption and power efficiency

Array power consumption is an extremely important issue in DRAM design since it dominates the overall chip power budget. Ideally, from Section 3.2 the low- and mid-level sensing consume zero power during equalization and pre-sensing because there is no extra charge required from the supply while the high-level sensing consumes power only during the equalization to charge half the number of bitlines to V_{dd}. During post-sensing phase the low-level sensing needs to pull half the number of bitlines in the array from ground potential to V_{dd}, whereas the mid-level will raise them only from $V_{dd}/2$ to V_{dd}. As a result, mid-level needs less power compared to the high- or low-level sensing. As the energy consumed is

$$E = V_{dd}\cdot\int_0^{\Delta t} I(t)\cdot dt = V_{dd}\cdot\Delta Q, \tag{3.23}$$

where ΔQ is the charge used to raise bitline voltage, ideally the total power consumption can be estimated as

$$E = V_{dd} \cdot C_{array} \Delta V \tag{3.24}$$

where C_{array} is the total sum of the array capacitances being charged. The power consumption for low-, mid- and high-level are obtained from Eqn. (3.24) as shown in **Table 3.2**. Here, $C_{array} = mn \cdot (C'_{bl} + C'_{bl2wl} + \lambda C'_{bl2bl})$ where λ is obtained from **Table 3.1** for different data patterns and array structures. m is the number of bitline pairs per array. Obviously, in order to reduce power consumption for mid-level sensing, C_{bl2bl} should be as small as possible. This is in accordance with the sensing speed conclusion from the previous section.

Table 3.2: Energy consumption for low-, mid- and high-level sensing

Sensing scheme	Low-level	Mid-level	High-level
Energy consumption	$V_{dd}^2 C_{array}$	$\frac{V_{dd}^2 C_{array}}{2}$	$V_{dd}^2 C_{array}$

In fact, besides the current required to charge the bitline capacitances, current is also consumed elsewhere by the array devices during sensing. As one of the current consumers, the sense amplifiers themselves introduce parasitic capacitances to the sensing nodes that need to be charged or discharged, and in addition, a direct current path from V_{dd} to ground exists when CMOS sense amplifiers are fully switched on. Therefore, the power usage efficiency becomes lower and the total power consumption is

$$P_{tot} = P_{sa} + P_{array} \tag{3.25}$$

where P_{array} is the power consumed on array capacitances such as C_{bl}, C_{bl2bl}, C_{bl2wl} and P_{sa} is the additional power required for sense amplifiers. The power usage efficiency of a CMOS sense amplifier can be expressed as

$$\eta = \frac{Q_{array}}{Q_{array} + Q_{sa}} \tag{3.26}$$

where Q_{array}, Q_{sa} are the charge transferred from supply to array and to sense amplifiers, respectively.

Next, the power usage efficiency will be estimated from a sensing process in **Fig. 3.8**(a) with the initial bitline voltage $V_{BL} < V_{BLb} \approx V_{dd}/2$ and a CMOS latched sense amplifier. Ideally, if the sense amplifier wastes no current, M_{nb}, M_{pa} should be off at the very beginning and M_{na}, M_{pb} are fully on to discharge or charge the bitlines. Unfortunately, since it takes time to charge or discharge bitlines from $V_{dd}/2$, transistors M_{nb}, M_{pa} work for a while, forming a direct short current path from V_{dd} to ground, which results in a charge loss as implied in

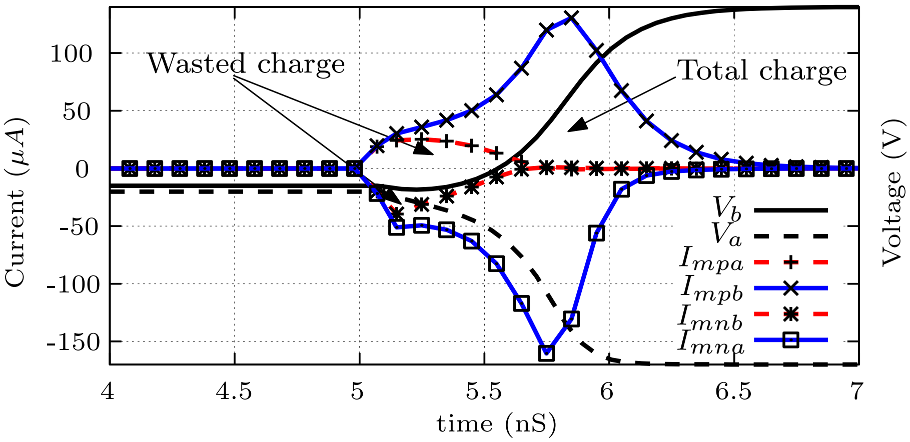


Fig. 3.16: The currents of the sensing transistors M_{na}, M_{nb}, M_{pa} and M_{pb} in a simultaneous sensing process with mid-level sensing scheme. The wasted charge is contributed by both M_{nb} and M_{pa}.

Fig. 3.16. From **Fig. 3.9** these two transistors will be turned off when the bitline voltage $V_{BL} < V_{thn}$ and $V_{BLb} > V_{dd} - |V_{thp}|$. By integrating the current through these two transistors over their on-time, the charge Q_{sa} wasted by the sense amplifier can be estimated.

$$Q_{sa} = \int_0^{t_{on}} (|I_{mpa}| + |I_{mnb}|)dt \tag{3.27}$$

Where I_{mpa}, I_{mnb} are the currents through transistors M_{pa}, M_{nb} and t_{on} is the time period when both of the two transistors are on. Since the two transistors enter into cutoff directly from saturation, the currents follow the equation

$$I_{ds,sat} = \frac{1}{2} K \frac{W}{L} (V_{gs} - V_{th})^2 < g_m (V_{gs} - V_{th}) \tag{3.28}$$

where g_m is transconductance of sensing transistors under initial condition. Furthermore, the node voltages V_{BL} and V_{BLb} are known for latch circuits from Eqn. (3.13)

$$V_{BL} - V_{BLb} = V_{sign} e^{\frac{g_m}{C_{bl}} t} \tag{3.29}$$

where $g_m = g_{mna} + g_{mpb}$ is the sum of both n- and p-transconductance. t_{on} of M_{pa} and M_{nb} therefore gives

$$t_{on} = \frac{C_{bl}}{g_m} \ln \frac{V_{dd} - V_{thn} - |V_{thp}|}{V_{sign}} \tag{3.30}$$

By taking t_{on} and using small signal approximations, the charge wasted by M_{nb},

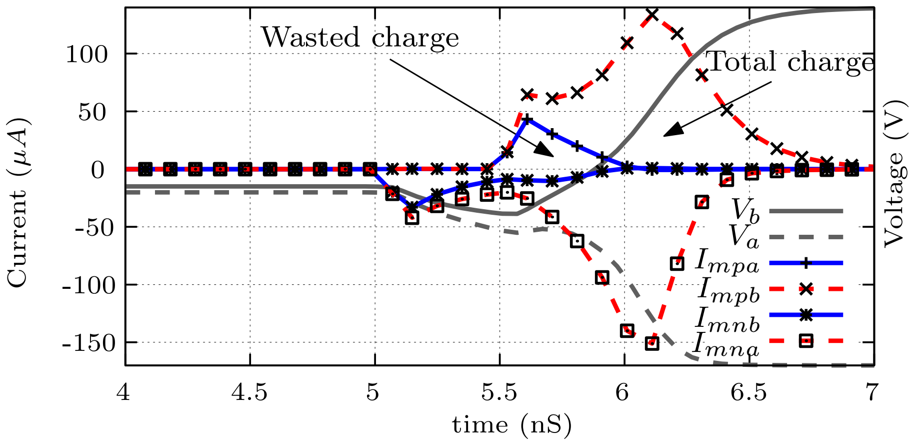


Fig. 3.17: The current waveforms of sensing transistors M_{na}, M_{pa}, M_{nb}, M_{pb} with switch delay. Because M_{nb} is almost cutoff before p-sensing transistors are switched on, the major wasted charge is contributed by M_{pa} as shown here.

M_{pa} gives

$$\begin{aligned} Q_{sa} &= \int_0^{t_{on}} (I_{mnb} + I_{mpa})dt \\ &< \int_0^{t_{on}} [g_{mnb}(V_b - V_{thn}) + g_{mpa}(V_{dd} - V_a - |V_{thp}|)]dt \\ &< \int_0^{t_{on}} [g_{mnb}(V_b - V_{thn} + V_{dd} - V_a - |V_{thp}|)]dt, \text{ suppose } g_{mnb} > g_{mpa} \\ &= \frac{g_{mnb}}{g_{mna} + g_{mpb}} C_{bl}(V_{dd} - V_{thn} - |V_{thp}|)(\ln \frac{V_{dd} - V_{thn} - |V_{thp}|}{V_{sign}} - 1) \end{aligned} \tag{3.31}$$

In mid-level sensing since the bitline voltage is raised from $V_{eq} = V_{dd}/2$ to V_{dd}, Q_{array} approximates $C_{bl}V_{dd}/2$ and the power usage efficiency of the CMOS latched sense amplifier gives

$$\eta = \frac{Q_{array}}{Q_{array} + Q_{sa}} > \frac{1}{1 + \frac{2g_{mnb}(V_{dd} - V_{thn} - |V_{thp}|)}{(g_{mna} + g_{mpb})V_{dd}}[\ln \frac{V_{dd} - V_{thn} - |V_{thp}|}{V_{sign}} - 1]} \tag{3.32}$$

Noticeably, the outcome from Eqn. (3.32) only provides a lower boundary for the power usage efficiency. It is exaggerated and usually worse than the real value.

Example

By Eqn. (3.32), in a typical DRAM design with $V_{dd} = 1.2$V, $V_{sign} = 100$mV, $V_{thn} \approx |V_{thp}| = 0.3$V, $g_{mnb} = g_{mna} = 3g_{mpa} = 3g_{mpb}$, η is greater than 62.7%.

Eqn. (3.32) suggests that η has no dependency on the size of n- or p-sensing transistors as long as g_{mn}/g_{mp} remains constant. But actually, as the size of the sensing devices grows, their parasitic capacitances C_{gd}, C_{db} will show more significant impact on total power consumption. As a consequence, η usually drops as devices are enlarged. The SPICE simulation in **Fig. 3.18** demonstrates this trend: the increase of sensing transistor width results in larger transconductance g_{mn} and g_{mp}, and therefore smaller sensing delay as indicated from Eqn. (3.13). However, the power usage efficiency drops a little from 73% to 72% due to the increase of sense amplifier related parasitic capacitances.

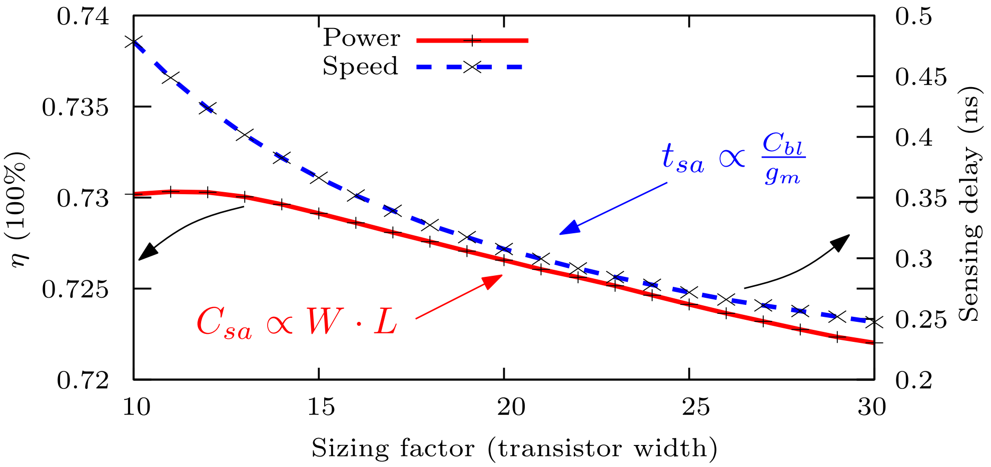


Fig. 3.18: Power usage efficiency and sensing delay vs. sizing factor of sensing transistors for a CMOS latched sense amplifier in a mid-level sensing (the horizontal axis is a multiplication factor applied to both width and length of n- and p-sensing transistors).

As can be seen from the SPICE simulations, in **Fig. 3.17** when p-sensing transistors are delayed the power usage efficiency will be higher because the n-latch pair will lower the common mode voltage in the beginning, making M_{nb} much easier enter its cutoff region when p-sensing transistors are turned on. The simulation changes for simultaneously switching as shown in **Fig. 3.16** since the charge wasted by the transistors M_{pa} and M_{nb} is increased. However, when p-sensing transistors are switched on prior to n-sensing transistors, power usage efficiency drops again.

Though simultaneously latched CMOS sense amplifier may waste some power compared to p-sensing transistors delayed sensing, it is still advantageous and favorable because of their higher sensing speed and imbalanced capacitive load immunity. Furthermore, later in Chapter 4 it will be shown that simultaneously latched CMOS sense amplifier can also provide higher yield in comparison with n- or p-latch sense amplifiers in high- or low-level sensing scheme.

3.3.5 Sensing transistors leakage control

In a low supply voltage environment the choice of the threshold voltages of sensing transistors is mainly driven by the post-sensing speed considerations. In order to fulfill the speed requirement low threshold voltage devices are necessary. On the contrary, with such low threshold voltage devices the off-current will be more significant because of the leakage in sub-threshold region [46, 47].

The leakage of sensing transistors is particularly harmful during equalization and pre-sensing. The phenomenon is shown in **Fig. 3.19**(a) (For simplicity only a n-sensing pair is drawn here). In equalization, a pair of bitlines is equalized to a common voltage V_{eq} and in order to provide enough post-sensing speed, V_{eq} is required to stay at a fixed voltage level. However, when leakage appears through the sensing and tail transistors, V_{eq} becomes difficult to maintain - the bitline voltage will drop or rise to an uncertain voltage level, making the post-sensing speed unpredictable. In addition, the sub-threshold leakage introduces additional power consumption that tends to cause a design to fail to meet specifications. During pre-sensing the effect of the leakage is subtle. Because the sub-threshold leakage relies on several parameters such as V_{th}, V_{gs} and V_{ds}, its impact on developed voltage difference is difficult to foresee. As a consequence, the best solution is to prevent the sensing transistors from generating large leakage currents in cutoff state.

The sub-threshold leakage is expressed as

$$I_{sub} = I_0(exp\frac{V_{gs}-V_{th}}{\xi V_T})(1-exp\frac{-V_{ds}}{V_T}) \tag{3.33}$$

where $V_T = kT/q$ and ξ is one plus a ratio defined by the gate oxide and channel depletion capacitances. When the source drain voltage V_{ds} exceeds a few V_T, I_{sub} becomes nearly constant and independent of V_{ds}. The slope of I_{sub} at this time gives

$$\frac{d[log_{10}(I_{sub}/I_0)]}{d(V_{gs}-V_{th})} = (log_{10}e)\frac{1}{\xi V_T} = \frac{1}{S}, \text{ where } S = 2.3V_T\xi \text{ (V/dec)} \tag{3.34}$$

Eqn. (3.34) reveals the relationship between on-current and off-current of a transistor. For example, when S = 100mV/dec, a change of 100mV in V_{th} leads to a ten-fold reduction in off-current I_{sub}. When the on-current is defined as the current at $V_{gs} - V_{th} = 0$, it will be $10^{V_{th}/S}$ times larger than off-current.

Example
When the pre-charge voltage is 0.6V and C_{bl} =40fF, to charge/discharge the bitlines to 1.2/0V in 10ns the minimum leakage current is $C_{bl}\Delta V/t = 2.4\mu$A. Suppose V_{th} = 200mV, the corresponding on-current is 200μA and sub-threshold slope S = 100mV/dec. From Eqn. (3.34) the off-current is around 2μA, which is very close

to the 2.4μA minimum leakage limit.

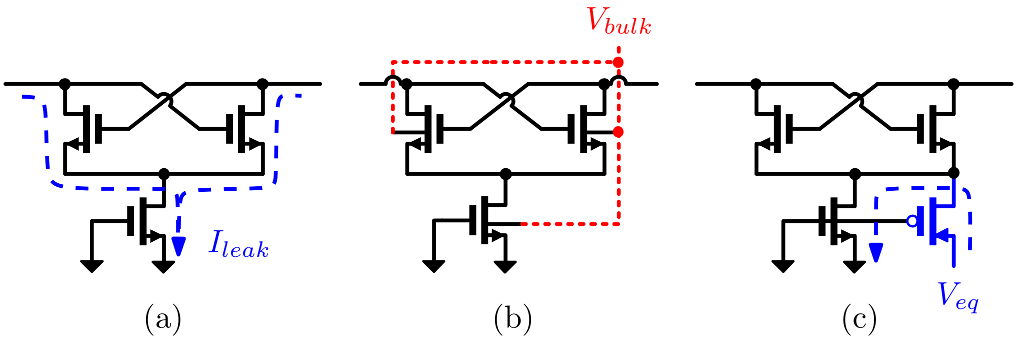


Fig. 3.19: (a) Leakage path during equalization and pre-sensing (b) Bulk control technique (c) Leakage reduction transistor

In earlier ultra low power low voltage DRAM circuits such as for battery operated products, bulk control of the sensing transistors was popular [48, 49, 50] since it can adjust both leakage and post-sensing speed as shown in **Fig. 3.19**(b). With the continuously dropping of supply voltage, this technique becomes popular again in current DRAM products. The leakage prevention and sensing speed enhancement are accomplished by raising or lowering the threshold voltage of the sensing transistors depending on the desired operation. However, when the array size becomes larger and the bulk parasitic resistance and capacitance increase substantially, it becomes too difficult to be implemented. The larger bulk parasitics introduce significant delay for sense amplifiers far from the control node and thus the method is not reliable. Besides that additional power is required to drive the capacitive bulk plate and layout effort is also a concern.

The circuit in **Fig. 3.19**(c) is more favorable. Another transistor is placed in parallel to the enable transistor, so that the source node of the sensing transistors will be clamped to V_{eq} when they are disabled. As the sub-threshold leakage can be almost neglected when V_{ds} of the sensing transistors are below 100mV [47], it has little impact on the bitline voltage.

3.3.6 Transistor sizing and layout

Unlike traditional analog circuits, DRAM sense amplifiers are strictly limited by the area they can occupy. Since in folded bitline arrays the sense amplifiers can be shared from both left and right sides as shown in **Fig. 3.5**, each sense amplifier must be accommodated in a vertical space of four $8F^2$ cells. For open bitline arrays the area limitation becomes even more severe because the double number of sense amplifiers needs to be accommodated. As a result, sense amplifiers that occupy too much area are not applicable to DRAM and the channel length of the sensing transistors must be as short as possible. On the other side, the threshold mismatch of these transistors is also important for yield. In order to cope with large threshold variation caused by channel length fluctuation, the choice of channel length is based on **Fig. 3.20** - The region where dV_{th}/dL approximates zero is the optimum position, and its location can be controlled technologically by adjusting doping concentration of the HALO region [51].

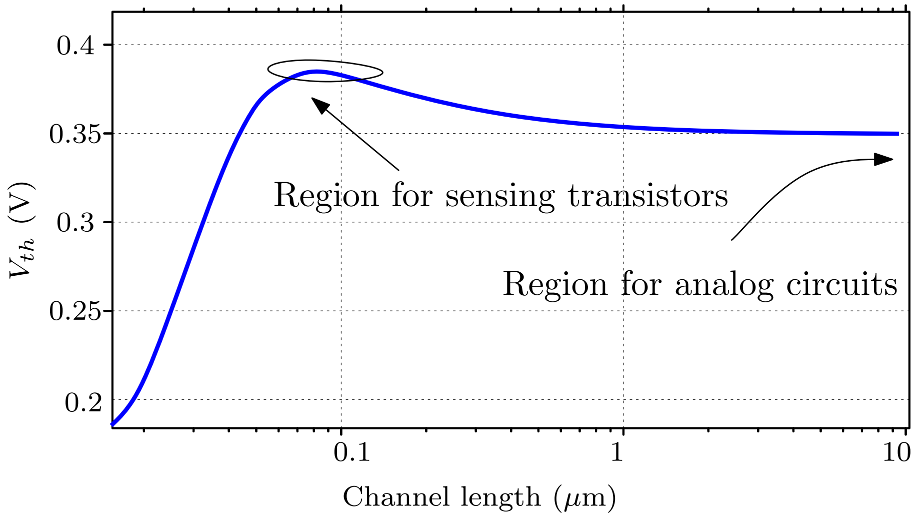


Fig. 3.20: The threshold voltage V_{th} is a function of transistor channel length.

3.4 Charge Transfer Sense Amplifier

In the above discussions, voltage sensing is mainly concerned. It is evident that DRAM sensing is based on detecting the quantity of electric charge in a cell. In the voltage sensing scheme sense amplifiers have to wait for a certain period of time, during which the bitline voltage difference V_{sign} can be established from the released charge of the cell capacitor. As mentioned in Section 2.1, since the pre-sensing process in the voltage sensing scheme is passive, the time to settle to V_{sign} is largely dependent on cell series resistance, bitline resistance, cell capacitor and bitline capacitance, which change greatly from technology to technology. On the other hand, the post-sensing speed of a latched sense amplifier also relies on the amplitude of the developed V_{sign}. If V_{sign} can be made larger during pre-sensing, it is conceivable that post-sensing speed will be greatly improved. As discussed in Section 1.4, charge transfer sense amplifiers are a sort of solution.

The first DRAM charge transfer sense amplifier is shown in **Fig. 3.21** [52]. At the first glance it seems like a normal voltage sense amplifier in NMOS technology

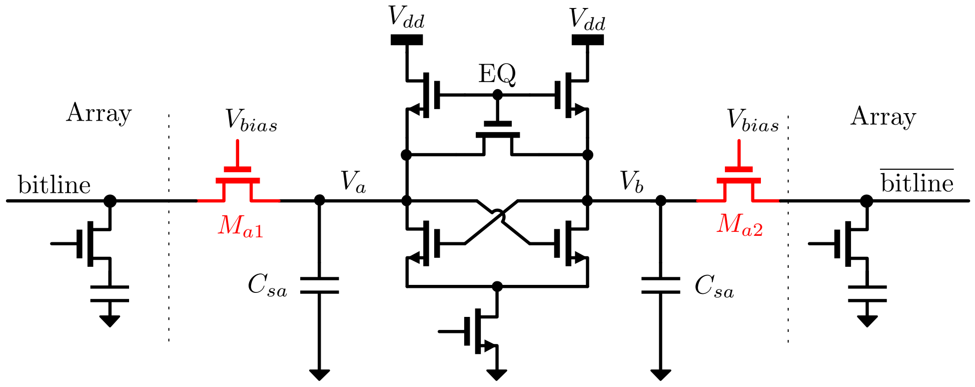


Fig. 3.21: The first DRAM charge transfer sense amplifier in NMOS technology [52]

at first glance. The difference comes from the operation of the isolation transistors M_{a1}, M_{a2}. Before wordline activation, EQ is on to equalize the bitlines. At the same time, V_{bias} is applied to the gate of the isolation devices M_{a1}, M_{a2}. Assume EQ is high enough to pre-charge the internal sensing nodes V_a, V_b to V_{dd}. With V_{bias} being lower than V_{dd}, both bitlines will be charged to around $V_{bias} - V_{thn}$ through M_{a1}, M_{a2}, where V_{thn} is the nominal threshold voltage of M_{a1}, M_{a2}. Obviously, if $V_{dd} - V_{bias} > V_{thn}$, M_{a1}, M_{a2} are biased on the edge of weak-inversion [53].

When EQ is disabled and a wordline is switched on, the charge released from the DRAM cell results in rise or drop of the bitline voltage. If the initial cell voltage is near ground, the resulting bitline voltage drop will revive M_{a1} or M_{a2} from weak-inversion into saturation, transferring part of the charge trapped at nodes V_a or V_b to the bitline until the bitline voltage is raised to $V_{bias} - V_{thn}$ again. As capacitances at the internal sensing nodes V_a, V_b are much smaller than the total bitline capacitance, the reductions of charge at these nodes produce a rapid local voltage drop.

As a result, a certain time after wordline activation the voltage difference developed at the internal sensing nodes V_a, V_b will be much larger than the originally developed V_{sign} from voltage sensing. With the larger input amplitude, a latch pair can react more quickly. In the end when post-sensing is completed, V_{bias} needs to be raised to a much higher voltage so as to restore the cell voltage to $V_{dd} - V_{th}$. Noticeably, when a cell stores a high voltage the bitline voltage will become higher and this charge transfer scheme will not function. Additional reference dummy cells are therefore needed to turn on the isolation transistor on the complementary bitline.

Actually, the isolation transistors M_{a1}, M_{a2} work as common gate amplification stage during pre-sensing. This common gate configuration provides low impedance $1/g_m$ for the bitline, and therefore similar to current sensing - the bitline capacitance shows a minor effect on pre-sensing speed as mentioned in Section 1.4. Suppose charge can be completely transferred from one internal sensing node to the bitline. It will compensate the released charge from the cell capacitor to maintain the bitline voltage. Consequently, by charge conservation the voltage change at the internal sensing node is

$$\Delta V = \frac{C_s}{C_{sa}} \cdot (V_{cell} - V_{eq}) \tag{3.35}$$

where V_{eq}, V_{cell}, C_{sa} are the bitline equalization voltage, cell initial voltage and sensing node capacitance, respectively. Eqn. (3.35) reveals the voltage margin at the sensing node between '1' and '0' cells is

$$\Delta V_1 - \Delta V_0 = \frac{C_s}{C_{sa}} \cdot (V_{cell,1} - V_{cell,0}) \tag{3.36}$$

As C_{sa} is several times smaller than C_s, Eqn. (3.36) implies that the achievable voltage margin in a charge transfer sense amplifier can be significantly larger than V_{sign} from passive pre-sensing as long as V_{dd} is much higher than V_{bias} to keep M_{a1}, M_{a2} in saturation.

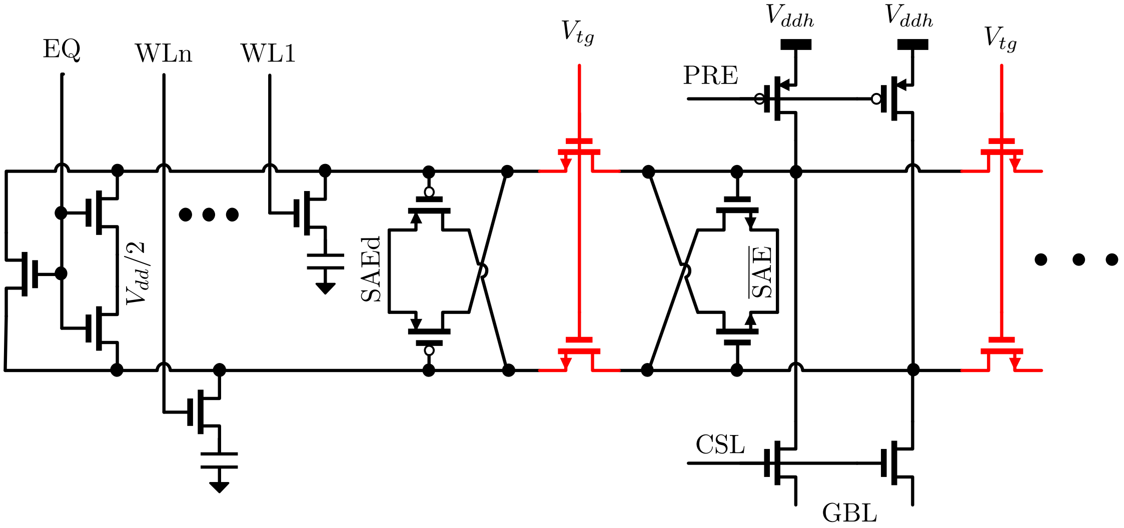


Fig. 3.22: A CMOS DRAM charge transfer sense amplifier[54]

CMOS charge transfer sense amplifiers [55, 54] can further improve the circuit performance. A CMOS charge transfer sense amplifier is shown in **Fig. 3.22** [54]. Different from the NMOS design, it incorporates another set of equalization and pre-charge switches for bitlines pairs. When the circuit is equalized, the bitlines and sensing nodes will be set to $V_{dd}/2$ (0.4V) and $V_{dd,h}$ (1.6V), respectively. This leaves a larger voltage headroom for the voltage developing at the sensing nodes during pre-sensing. The isolation device bias voltage V_{tg} is close to 0.4V+V_{thn}, so that the isolation transistors can directly go into saturation when the bitline voltage drops a little from 0.4V. The p-latch pair is placed on both sides while an n-latch pair is shared in the middle. The n-latch pair is enabled first, amplifying the voltage difference at the internal sensing node. Then the p-latch pair is used to pull one of the bitlines to the 0.8V core supply. The reported sensing speed is 1.33× faster than conventional voltage sense amplifier.

According to Eqn. (3.35), a larger C_s/C_{sa} ratio is favorable for a larger internal voltage signal. However, the available voltage headroom at the internal sensing node limits this ratio to some degree because the maximum achievable sensing node voltage change ΔV is determined by $V_{ddh} - V_{dd}/2 - V_{thn}$. This limitation is alleviated by boosting the sensing node voltage [56, 57]. Like the bitline boost technique in **Fig. 3.3** in NMOS technology, a larger voltage headroom can be obtained at the internal sensing nodes.

Though charge transfer sense amplifiers are a topic intensively discussed in the past ten years, they are actually seldomly used for commercial DRAM products. The first consideration comes from the trade-off between area and yield. The increased voltage difference at internal sensing nodes can alleviate the mismatch consideration for the latch pair, which seems beneficial compared to voltage sense amplifier design. However, this increased voltage difference results from the saturated isolation devices that contribute threshold voltage mismatch as well. Though theoretically, bitlines can be pre-charged to $V_{bias} - V_{th}$ through the isolation devices to cancel the threshold voltage variation as in the NMOS design, this takes too much time and sacrifies the speed gained by the charge transfer scheme. As a result, bitlines are equalized to a pre-defined voltage in the CMOS charge transfer sense amplifier. In this scheme their mismatch has to be taken into account. Since the isolation devices work as common gate amplifier during pre-sensing, the input offset approximates their threshold voltage mismatch. The voltage difference at the internal sensing nodes due to mismatch becomes

$$\Delta V_{os} = \frac{C_{bl}}{C_{sa}} \Delta V_{th} \tag{3.37}$$

When the mismatch is referred back to cell voltage by Eqn. (3.35), the corresponding cell voltage loss is

$$\Delta V_{cell,loss} = \frac{C_{bl}}{C_s} \Delta V_{th} \tag{3.38}$$

Compared to voltage sensing with a cell voltage loss $V_{cell,loss} = (1 + C_{bl}/C_s)\Delta V_{th}$ when $V_{os} = \Delta V_{th}$, charge transfer is still advantageous. However, the important issue is that with threshold voltage variation the internal sensing node pre-charge voltage V_{ddh} must be even higher to keep the isolation transistors well in saturation after wordline activation. Besides that, the transconductance of the isolation transistors is expected to be as large as possible so as to suppress bitline voltage fluctuations. All this imposes minimum area constraints on the isolation transistors. To meet the demands the isolation devices will be at least the same size as the sensing transistors in a voltage sense amplifier that can satisfy the yield requirement.

Secondly, the isolation bias voltage V_{tg} is hard to define. With local well technology that allows the source terminal to be connected to the bulk terminal the threshold voltage of the isolation devices can be independent of the bitline voltages but the area cost rises. If these devices are fabricated without local well, due to substrate body-bias the threshold voltage becomes a function of bitline voltage.

Thirdly, the capacitor ratio defined by the cell capacitor C_s and internal sensing node capacitance C_{sa} may vary in large range because of process variation and voltage dependent characteristics of C_{sa}.

Lastly, when the internal sensing nodes are pre-charged to a voltage that is much higher than the cell refresh voltage, the reliability and leakage controllability of the n-latch pair become even harder to implement. As a conclusion, though charge transfer sense amplifiers are fast, they still lack of favorable balances between area, yield, speed and power consumption trade-offs.

3.5 Threshold Voltage Compensation Technique

The threshold voltage mismatch of sensing transistors is problematic in that it deteriorates the effective voltage difference for sense amplifiers. In a typical DRAM design with 100-200mV V_{sign}, V_{os} is under tight control within 10-20mV so as to meet 4-5σ yield requirement. Since transistor threshold variation caused by dopant fluctuation is determined by the gate area according to [58, 59, 60], the continuously shrinking DRAM technology demands relatively larger sensing transistors because: a). The tolerable mismatch equivalent sense amplifier input offset V_{os} is expected to be even lower due to the increasing total number of bits available on a single die and thus higher yield requirement; b). The decreasing core supply voltage accompanied with technology shrinking reduces the available V_{sign}. Therefore, the ratio of sense amplifier over array tends to rise instead of going down as expected by downscaling. To better solve the problem, threshold voltage mismatch compensation techniques becomes promising.

Threshold voltage mismatch compensation sense amplifier is first reported in [61] as shown in **Fig. 3.23**. The basic idea is to self-bias the sensing transistors during equalization, so that the sensing node voltage is locally threshold voltage mismatch compensated. Here the control signals PRE, ϕ_1, ϕ_3 are on for the first step where the sensing transistors are working in diode connection with $V_{ds} = V_{gs} = V_{thn} + V_{od}$. V_{od} is called over-drive voltage, determined by voltage V_E.

Consequently, the threshold voltage of each sensing transistor is memorized by the parasitic capacitances at the drain node. With the voltage, both transistors should have the same I_{ds} independent of their threshold voltages.

Then the control signals PRE and ϕ_1 are turned off and wordline is switched on. When ϕ_2 is enabled the bitline voltage change will be transferred to the internal sensing node, since the bitline is AC coupled to the internal sensing node by a large capacitor. When ϕ_2 is on and ϕ_1 and SAE are low there will be no current through the sensing transistors. Clearly, at this point the gate voltage of the sensing transistors is the sum of their drain voltage and AC coupled bitline voltage. Eventually, ϕ_3 goes low, separating the internal sensing nodes from the large coupling capacitors and SAE enables the sensing transistors to amplify the node voltage difference. By this method the threshold voltage mismatch of the sensing transistors can be reduced to 2mV, which is more than ten times smaller than the 30mV nominal value.

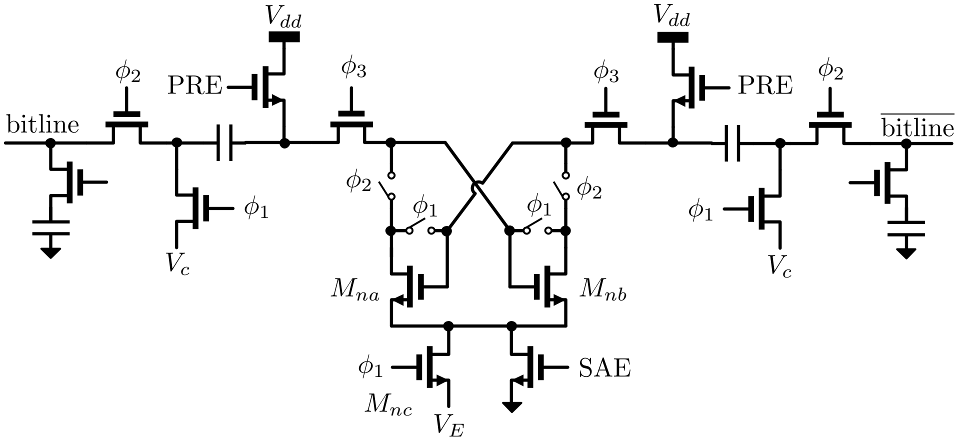


Fig. 3.23: The first proposed threshold compensation sense amplifier in NMOS technology

The disadvantage of the sense amplifier in **Fig. 3.23** is also obvious. As it comprises large AC coupling capacitors and more transistors, the area cost is comparable to a normal sense amplifier that is designed with much larger sensing transistors. Secondly, it incorporates complex control signals, degrading the sensing speed significantly.

A novel mismatch compensated sense amplifier incorporating hierarchical sensing stages is proposed in [62] as shown in **Fig. 3.24**. It is very compact due to the hierarchical sensing style, in which part of the circuits can be shared by local primary sense stages. However, in this design the cell refresh process is separated from the sensing process and the local primary sense stages lose the ability of latching sensing outcomes due to the fact that local transistors can not form a latch circuit without the help of the global sense amplifier circuitry.

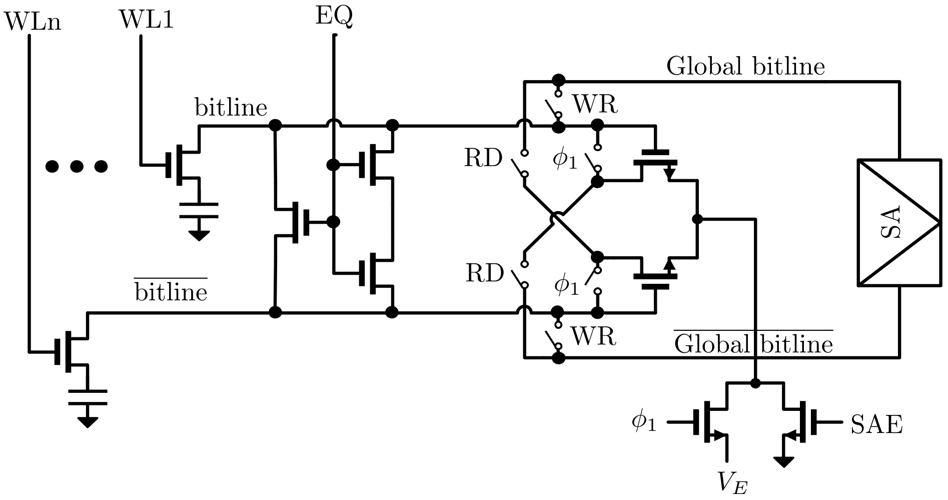

WLn
WL1
EQ
bitline
$\overline{\text{bitline}}$
Global bitline
$\overline{\text{Global bitline}}$
WR
RD
ϕ_1
SA
ϕ_1
SAE
V_E

Fig. 3.24: A hierarchical mismatch compensated DRAM sense amplifier

Table 3.3: Sensing schemes and techniques comparison

Technique	V_{eq}	Speed	Area	Control effort	Power	V_{os}*
A	V_{dd} or gnd	high	small	lowest	moderate	large
B	$V_{dd}/2$	high	moderate	low	lowest	moderate
C	V_{dd}, V_{eq}	highest	large	high	moderate	moderate
D	self-bias	slowest	largest	highest	highest	smallest

A High/Low-level sensing
B Mid-level sensing with simultaneously latched CMOS sense amplifier
C Charge transfer sense amplifier
D Threshold mismatch compensated sense amplifier
* Obtained with identical sensing transistors

Other mismatch compensated sense amplifiers have been published as well [63, 64]. However, in general they are not area efficient and more efforts are required in timing control. As a conclusion, the threshold voltage compensation technique is not suitable for a normal DRAM sense amplifier as it sacrifies area, sense amplifier functions, control simplicity and power consumption in return for yield.

3.6 Summary

In this chapter different sense amplifiers, sensing schemes and techniques are reviewed. In particular, CMOS mid-level sensing is studied in its sensing speed and power consumption. **Table 3.3** lists the characteristics of each sensing technique. In comparison, simultaneously latched CMOS sense amplifier with mid-level sensing exhibits the best balances between area, yield, speed and power consumption, and thus the yield analysis of CMOS latched sense amplifiers will be in the focuse of Chapter 4.

Chapter 4

Sense Amplifier Yield Analysis

4.1 Circuit Yield Analysis

It is well known that in the real world there are no two identical things. Semiconductor devices can not be exceptions even if there are thousands of millions of devices that are nominally designed to be identical. Device variability exists everywhere due to random dopant fluctuations [65, 66], within a die or from die to die [60]. These intrinsic dopant fluctuations manifest themselves through device parameters such as threshold voltage V_{th}. The resulting mismatch can have significant impacts on highly compact circuits like DRAM and SRAM in that yield problems emerge once the on-die device population becomes enormous. As the impact of variations continues to grow in future process generations, prediction and estimation of the related yield degradation become crucial.

Worst case modeling can be utilized to avoid yield problems by taking extreme device variations and worst environment into account. For example, in Section 2.3 the necessary V_{sign} has been calculated from a worst case coupling model with 50mV worst case threshold mismatch. It is usually sufficient to guarantee the necessary yield but it also results in the rise of production cost because actually the worst case seldom happens and it is not necessary to prepare all on-die devices and circuits for the battle. As a consequence, worst case modeling is not suitable for a large number of repeated circuit structures like DRAM.

In this chapter, statistical circuit yield analysis will be introduced in an analytical way. In spite of the availablility of computer aided yield analysis [67] or test and measurement based yield analysis [68], the yield can only be analyzed and estimated when it changes within a boundary coming from certain rules. Analytical yield analysis is a way of looking for these hidden rules inside the circuits. Due to these substantial rules variability can be handed down from random variables to final yield specifications. For example, the threshold voltage

V_{th} can be regarded as a random variable. The resulting inverter gate delay is actually a function of V_{th}. By carefully modeling the function between V_{th} and the propagation delay, the delay related yield performance can be obtained from the probability evaluations. As implied, it is important to find the variation transfer function so as to obtain the final specification related statistical distributions and performance. In most cases linear modeling of variation transfer is valid due to the fact that any non-linear function can be assumed to be linear within very small region where the crucial pass/fail is decided. With the aid of the functions that can be obtained from circuit analysis in the small linear region, yield can be predicted and its related trade-offs can be optimized.

4.1.1 Variation transfer

One of the observed phenomenon is the Gaussian distributed parameters such as measured clock timing jitter or amplifier output noise. It results from *Central Limit Theorem* [69], which states that the sum of multiple independent random variables follows Gaussian distribution[1]. Actually, Central Limit Theorem has been used for a very long time in classical noise analysis of analog circuits [71, 53], in which different noise sources are supposed to have zero mean value and constant variance. The total output noise variance of an analog block is the total sum of variances of different noise sources. According to probability theory, this addition is valid only by assuming all noise sources are following independent Gaussian distributions, though this is sometimes not directly explained in text books.

When random variables following Gaussian distributions are combined in a linear system with function $f(x, y, z \ldots) = c_1 x + c_2 y + c_3 z + \ldots$, the output will also follow Gaussian distribution with

$$\left\{ \begin{array}{lcl} \mu & = & c_1\mu_x + c_2\mu_y + c_3\mu_z + \ldots \\ \sigma^2 & = & c_1^2\sigma_x^2 + c_2^2\sigma_y^2 + c_3^2\sigma_z^2 + \ldots \end{array} \right. \tag{4.1}$$

Consequently, statistical performances in a linear system with Gaussian random variables are easy to be analyzed. When the path the random variables passing through is nonlinear, things become a little complicated because complex convolutions have to be involved in obtaining the statistical characteristic of the output. However, some linear approximations can be used as described in the following example.

By measuring the contact resistance between metal layers M_1 and M_2, it is found that the resistance from a unit via as shown in **Fig. 4.1** follows a Gaussian distribution with standard deviation σ and mean R_0. What is the statistical

[1] Classical central limit theorem deals only with sum of independent statistical variables with identical distribution while extension of central limit theorem includes sum of independent statistical variables with different distribution but *Lindeberg Condition* [70] has to be followed.

characteristic of the contact resistance between M_1 and M_2 if two unit vias are put together in parallel as shown in the figure?

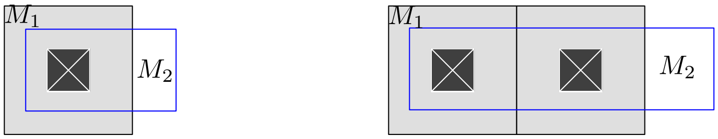


Fig. 4.1: Left, unit via between M_1, M_2; Right, unit vias in parallel between M_1, M_2

Suppose the on-resistance of the two vias in parallel are R_1 and R_2, respectively. When fringe effects are neglected, R_1 and R_2 will follow the same statistical distribution as measured for a unit via resistance. Since this time the resistance between the two metal layers is $R = R_1||R_2$, the expectation for R is

$$R = \frac{R_1 R_2}{R_1 + R_2} = \frac{R_0}{2} \tag{4.2}$$

Obviously, from Eqn. (4.2) the resistance R between M_1 and M_2 does not follow a linear function of R_1, R_2. To obtain the variance of R, some fundamental approximations have to be applied. From *Taylor Series*, it is known that any function $f(x + \Delta)$ can be written as

$$f(x + \Delta) = f(x) + f'(x)\Delta + \frac{f''}{2}(\Delta)^2 + \dots \tag{4.3}$$

If the high order terms are much smaller than the zero and first order terms, the above equation can be approximately transformed to

$$\Delta f(x) = f(x + \Delta) - f(x) = f'(x) \cdot \Delta \tag{4.4}$$

When Δ is sufficiently small, $f'(x)$ can be regarded as a constant. $\Delta f(x)$ will exhibit the same distribution as variable Δ with variance

$$\sigma^2_{\Delta f(x)} = [f'(x)]^2 \cdot \sigma^2_{\Delta} \tag{4.5}$$

In the above evaluations, the magnitude of higher than second order terms in Eqn. (4.3) are assumed to be small and can be neglected, or

$$|f(x) + f'(x)\Delta| >> |\Sigma_{n=2}^{\infty} \frac{f^n(x)}{n!} \cdot \Delta^n| \tag{4.6}$$

Fortunately, Eqn. (4.6) is valid in most cases except for some functions like the exponential function. For two random variables Eqn. (4.4) can be expanded, resulting in

$$\Delta f(x, y) \approx f'_x(x, y)\Delta x + f'_y(x, y)\Delta y \tag{4.7}$$

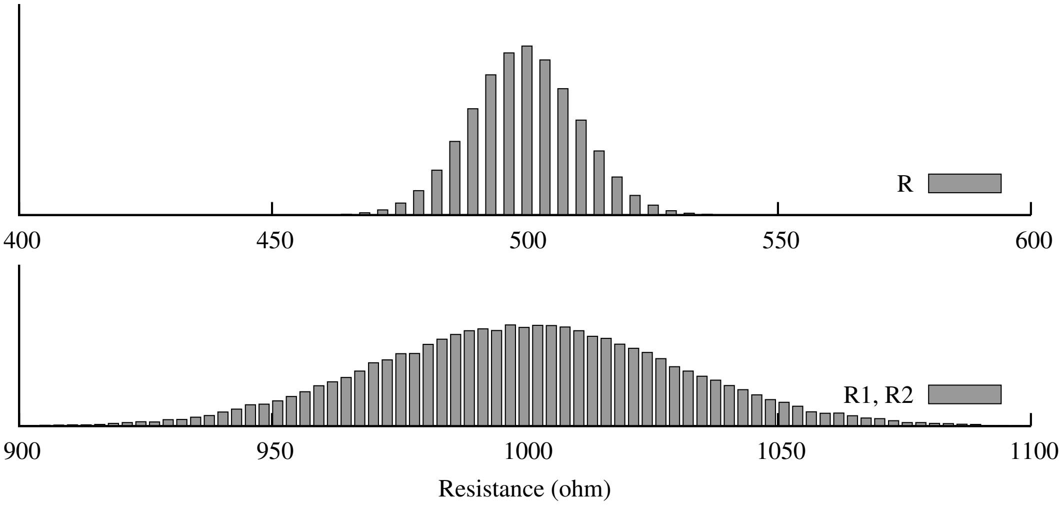


Fig. 4.2: MC simulated histogram of R and R_1

Suppose Δx and Δy follow independent Gaussian distributions. In analogy with Eqn. (4.5), the variance of $\Delta f(x, y)$ can be obtained

$$\sigma^2_{\Delta f(x,y)} \approx [f'_x(x, y)]^2 \cdot \sigma^2_{\Delta x} + [f'_y(x, y)]^2 \cdot \sigma^2_{\Delta y} \tag{4.8}$$

In addition, as $\Delta f(x) = f(x_1) - f(x_2)$ is a linear combination of function $f(x)$, the Δ sign in Eqn. (4.5) and Eqn. (4.8) can be eliminated, giving

$$\sigma^2_{f(x,y)} \approx [f'_x(x, y)]^2 \cdot \sigma^2_{x} + [f'_y(x, y)]^2 \cdot \sigma^2_{y} \tag{4.9}$$

Eqn. (4.9) implies the resulted distribution will be Gaussian with variance being a weighted combination of variance from both random variables. By Eqn. (4.2) the distribution of resistance R in the above example approximates to a Gaussian distribution with variance

$$\sigma^2_R = \frac{R_2^4}{(R_1 + R_2)^4} \cdot \sigma^2_{R1} + \frac{R_1^4}{(R_1 + R_2)^4} \cdot \sigma^2_{R2} = \frac{1}{8}\sigma^2 \tag{4.10}$$

The two vias in parallel reduce the resistance in its mean value to one half and its variance to one eighth. By this means the connection between M_1 and M_2 becomes more robust due to its narrower distribution. **Fig. 4.2** plots the histograms of R and R_1,R_2 obtained from Monte-Carlo (MC) simulations. The simulation outcomes confirm that with standard deviation of R_1, R_2 of 30Ω, the standard deviation σ_R is 10.6Ω, which agrees well with theoretical calculations from Eqn. (4.10).

In most cases, when Gaussian approximations exist for random variables, Eqn. (4.9) is valid. However, there are also cases that the simple equations can not be used and more complex methods have to be applied to obtain the system related statistical performance. One such example in DRAM core circuits is the cell leakage current, which will be addressed in Chapter 5.

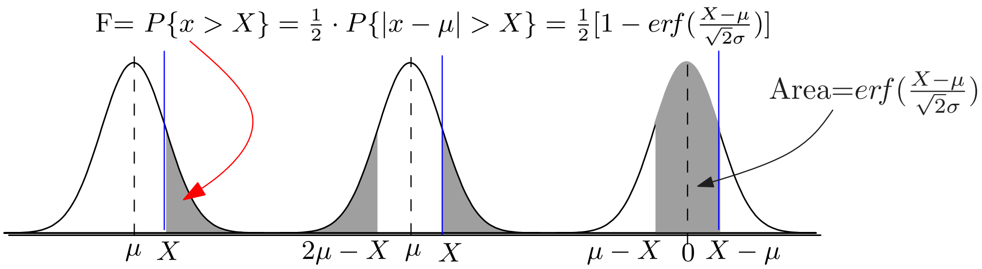

F= $P\{x > X\} = \frac{1}{2} \cdot P\{|x-\mu| > X\} = \frac{1}{2}[1 - erf(\frac{X-\mu}{\sqrt{2}\sigma})]$
Area=$erf(\frac{X-\mu}{\sqrt{2}\sigma})$
μ
X
$2\mu - X$
μ
X
$\mu - X$
0
$X - \mu$

Fig. 4.3: Process to calculate failure probability from a Gaussian distribution.

4.1.2 Analytical yield expression

The current flowing through the contact resistance generates heat. This may cause a contact malfunction when the current is too large by melting down the via material. Therefore, certain specifications demands the contact resistance to be less than $X\Omega$. Suppose the contact resistance follows a Gaussian distribution. Yield Y can be easily obtained from the distribution in **Fig. 4.3**

$$\begin{aligned} \mathrm{Y} &= 1 - \mathrm{F} \\ &= 1 - \frac{1}{\sigma\sqrt{2\pi}} \int_X^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \end{aligned} \tag{4.11}$$

Since the Error Function is defined as

$$erf(x) = \frac{2}{\pi} \int_0^x e^{-t^2} dt \tag{4.12}$$

Eqn. (4.11) can be simplified by using the error function and the yield Y becomes

$$\mathrm{Y} = 1 - \frac{1}{2}[1 - erf(\frac{X-\mu}{\sqrt{2}\sigma})] = \frac{1}{2}[1 + erf(\frac{X-\mu}{\sqrt{2}\sigma})] \tag{4.13}$$

Eqn. (4.13) implies that the ratio $(X - \mu)/\sigma$ is the most important item in a statistical design with Gaussian approximations - the yield remains constant when both numerator and denominator in Eqn. (4.13) are increased or reduced but their ratio maintains.

In typical design, X is known as failure boundary. In order to avoid failures, worst case corners are used to obtain the design goal μ that is placed far enough from X. In the above via example if the contact resistance is expected to be less than 50Ω, technology corner simulations can be carried out to verify whether the via resistances from different corners, are smaller than 50Ω. However, this ignores the statistical factor that the via resistance is in fact a random variable with fluctuations. It may happen that in a worst corner the via resistance spreads narrowly but in best corner widely. In other words, if μ in **Fig. 4.3** moves to the

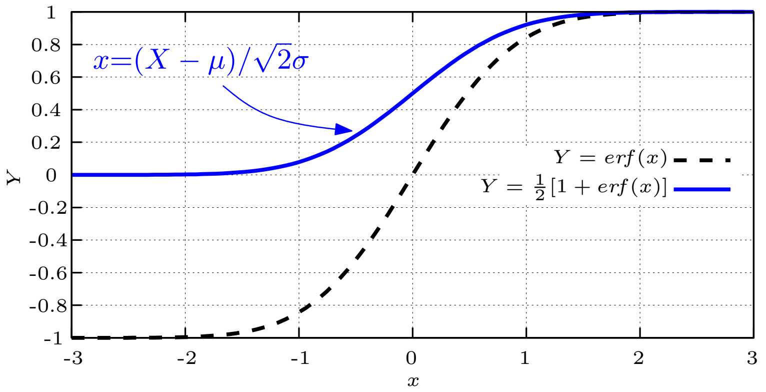


Fig. 4.4: Error function and yield function in Eqn. (4.13).

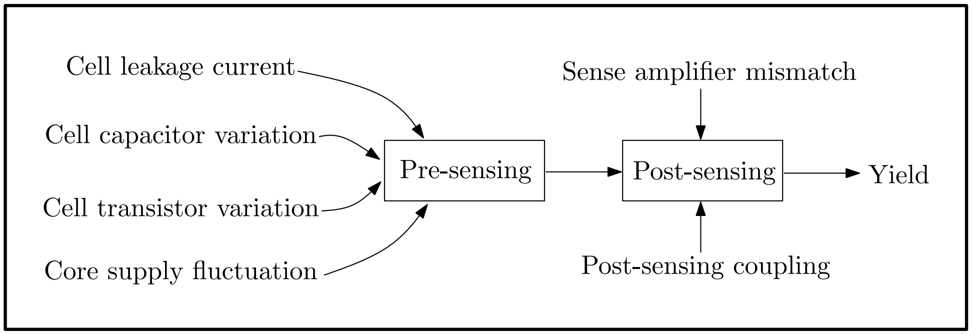


Fig. 4.5: Major random sources in DRAM core

left or right, the yield will not surely go higher or drop lower as one has expected, since it can only be determined by the ratio $(X - \mu)/\sigma$ as shown in **Fig. 4.4**. This ratio is quite similar to the Signal to Noise Ratio (SNR) in the Analog to Digital Converter (ADC) or Digital to Analog Converter (DAC) in that the noise is actually also regarded as one statistical component. Both ratios express the ability of a circuit/system to survive the real random and stochastic world.

4.2 Random Error Sources in DRAM

There are lots of different random error sources affecting the final electrical yield of DRAM core circuits as shown in **Fig. 4.5**. Before pre-sensing, cell leakage will reduce the core supply dependent charge quantity in cell capacitors. During the pre-sensing the on-resistance of the cell transistor will degrade the bitline voltage settling speed to a certain degree. For a fixed pre-sensing timing window, conversely the on-resistance will reduce the available developed voltage difference V_{sign}. Then the mismatch of sense amplifiers comes into play - it is added to V_{sign} and results in a more stochastic situation for post-sensing, in which inter-bitline coupling may deteriorate the situation by turning weak sensings into failures.

As these random variables interact with each other, the final outcome becomes uncertain for each sense amplifier. However, for a large number of sense amplifiers in a linearized sensing model the yield probability becomes nearly deterministic and can be estimated to a certain degree.

In this chapter, as the first step the threshold mismatch inside sense amplifiers will be modeled as an input offset V_{os} with Gaussian distribution. Based on the simplified input offset model and accurate array transfer functions in Chapter 2, cell leakage introduced yield degradation will be analyzed in Chapter 5, and in Chapter 6 a linear hierachical yield model is developed to analyze and optimize yield-area, yield-core supply voltage trade-offs for DRAM core arrays.

4.3 Latched Sense Amplifier Yield Analysis

4.3.1 Introduction

With deep sub-micrometer feature size the threshold mismatch[2] of sensing transistors in latched sense amplifiers becomes a crucial parameter concerning the electrical yield of DRAMs, and a deep understanding of the statistical characteristics of the threshold mismatch caused error probability of latched CMOS sense amplifiers is mandatory.

In an earlier publication [72], a complex numerical method is used to analyze the mismatch related sensitivity of CMOS latches. In [73], the mismatch is analyzed by using a set of differential equations. The state space concept is used in [74] to determine the final state of mismatched CMOS latches. However, these approaches are all based on numerical methods and are thus not suitable to provide an analytical guideline for practical optimization of latched CMOS sense amplifiers.

In this section, the mismatch of latched CMOS sense amplifiers is investigated by using small signal analysis and statistical probability theories. It is replaced by an offset voltage V_{os} with Gaussian distribution. Sense amplifier yield can be obtained and optimized by minimizing the spread of V_{os}.

4.3.2 Mismatch Equivalent offset V_{os}

DRAM sense amplifiers use devices with relatively short channel lengths to fit into the bitline pitch and save die cost. The yield of sense amplifiers is determined by mismatch of the sensing transistors. As the achievable V_{sign} from pre-sensing

[2] For simplicity, threshold mismatch is abbreviated to mismatch in this section.

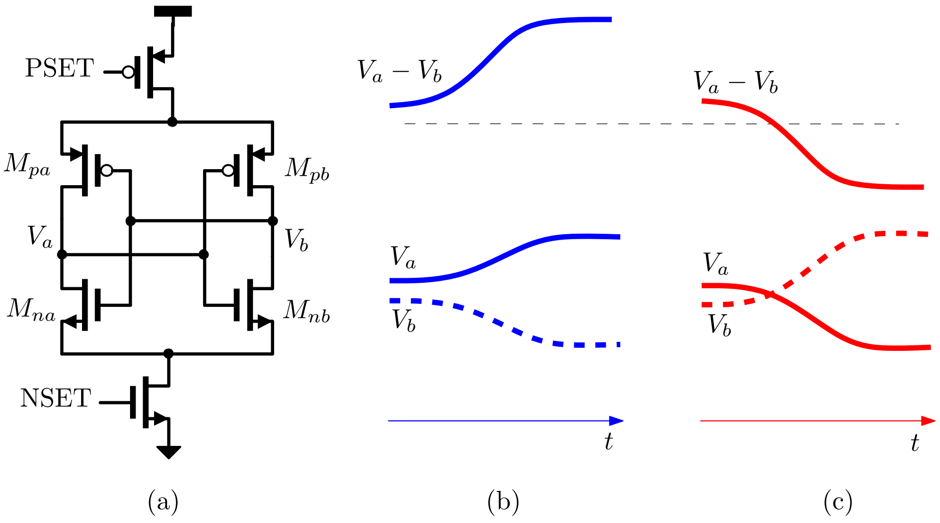


Fig. 4.6: (a) A CMOS latched sense amplifier (b),(c) Various differential output waveforms of the sense amplifier with initial voltage $v_a - v_b > 0$. When $v_a - v_b$ crosses the zero line, the sensing fails as shown in (c).

gets smaller with the continuously decreasing of supply voltage, understanding of the mismatch effects in DRAM sense amplifiers becomes increasingly important.

Generally, latched sense amplifiers consist of a complementary pair of cross-coupled n- and p-transistors as shown in **Fig. 4.6**(a). It has been suggested in Section 3.3 that simultaneously latched CMOS sense amplifiers are rather insensitive to capacitor imbalance between bitlines, and the switch-on time difference of n- and p-pairs was shown to have impacts on post-sensing speed and power usage efficiency.

Suppose v_a and v_b in **Fig. 4.6**(a) have been equalized to V_{eq}. When sensing begins, the wordline is switched on to release cell charge and generate the necessary V_{sign} for post-sensing. As NSET goes to high and PSET is dragged to low simultaneously, the output voltage difference $v_a - v_b$ will rise monotonically as shown in **Fig. 4.6**(b) due to the amplification of the sense amplifier. But sometimes the mismatch of the latched sense amplifier is rather large and the output $v_a - v_b$ will continuously drop as shown in (c). As shown in Chapter 3, the simultaneously latched CMOS sense amplifiers with initial equalization voltage V_{eq} around $V_{dd}/2$ have the following characteristics at the moment the sense amplifier is enabled: 1) the n- and p-sensing transistors are in saturation; 2) the transconductances of the sensing transistors can be treated as constants. According to previous numerical studies [72, 73, 74], the final output state is only determined by the initial states - the initial input voltage difference V_{sign} and the mismatch of sensing transistors.

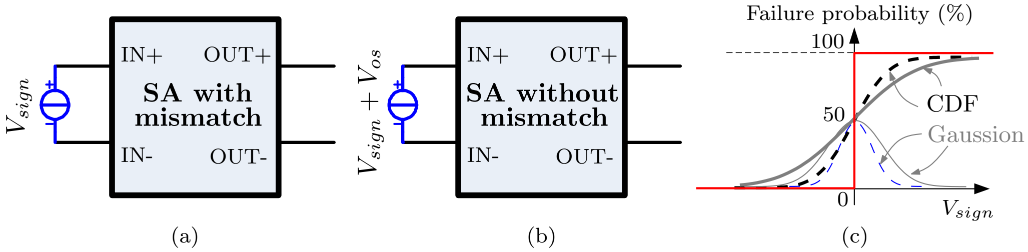


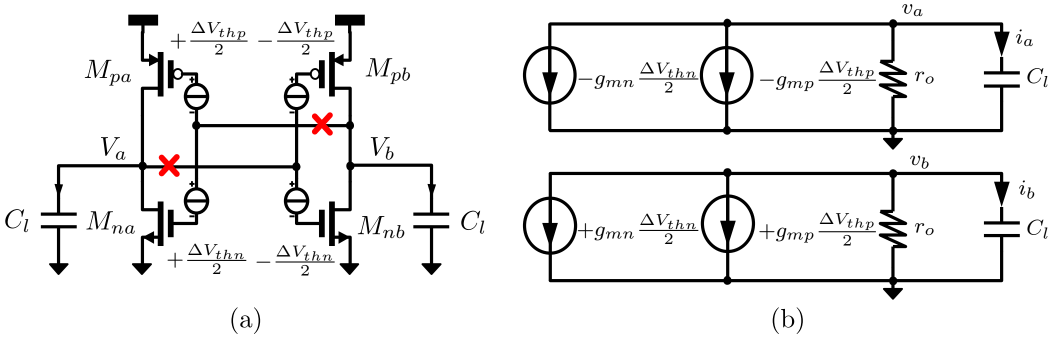
Fig. 4.7: Process to obtain V_{os}: If the sense amplifier with mismatch and input V_{sign} (a) produces the same failure probability plot as the circuit (b) containing a mismatch free sense amplifier and an input statistical voltage V_{os}, then V_{os} is capable of substituting the statistical mismatch inside the sense amplifier. (c) shows the failure probability for a full swing sweep of V_{sign} in (a) and (b).

Suppose a huge ensemble of DRAM sense amplifiers is fed with a given V_{sign} as shown in **Fig. 4.7**(a). A failure is obtained when the polarity of the output of the sense amplifier is different from the polarity of the given V_{sign}. The total number of failures from all the sense amplifiers at the given V_{sign} can be sketched into a failure count vs. V_{sign} plot as shown in **Fig. 4.7**(c). It is found that this plot is always following the cumulative density function (CDF) of a Gaussian distribution. Thus, presumably an input offset voltage V_{os} with Gaussian distribution can be used to replace the mismatch inside sense amplifiers as depicted in **Fig. 4.7**(b). In this case, the failure probability becomes easy to analyze: the relationship between the distribution of V_{os} and the final failure count is simple and deterministic, since the sense amplifier is now ideal. V_{os} with smaller variance results in a sharp transition in the failures versus V_{sign} curve and thus higher yield at a given V_{sign} as shown in (c). The question is how to obtain the statistical characteristic of V_{os} from the sense amplifiers containing both n- and p-transistor mismatch.

The analysis of mismatch in a latched sense amplifier can be simplified by studying only a pair of inverters by breaking the positive feedback loops in the latch pair as shown in **Fig. 4.8**(a), since both have the same initial conditions at all nodes. Suppose the mismatch of n- and p-sensing pairs are equal to ΔV_{thn} and ΔV_{thp}. When V_{sign} is zero and the sensing transistors stay in saturation, the small signal model shown in **Fig. 4.8**(b) can be introduced to obtain the current difference through capacitor C_l at nodes v_a, v_b when the sense amplifier is enabled

$$\Delta I_1 = i_a - i_b = g_{mn} \cdot \Delta V_{thn} + g_{mp} \cdot \Delta V_{thp} \qquad (4.14)$$

Here, g_{mn}, g_{mp} are transconductance for n- and p-sensing transistors, respectively. Since the DC bias currents and currents through the output resistance $r_o = r_{dsn}||r_{dsp}$ for both branches are identical, they are eliminated in Eqn. (4.14) and only small signal parameters are left.



M_{pa}
$+\frac{\Delta V_{thp}}{2}$
$-\frac{\Delta V_{thp}}{2}$
M_{pb}
V_a
V_b
C_l
M_{na}
M_{nb}
C_l
$+\frac{\Delta V_{thn}}{2}$
$-\frac{\Delta V_{thn}}{2}$
(a)
v_a
i_a
$-g_{mn}\frac{\Delta V_{thn}}{2}$
$-g_{mp}\frac{\Delta V_{thp}}{2}$
r_o
C_l
v_b
i_b
$+g_{mn}\frac{\Delta V_{thn}}{2}$
$+g_{mp}\frac{\Delta V_{thp}}{2}$
r_o
C_l
(b)

Fig. 4.8: (a) Translation of the initial condition of a latched sense amplifier into the corresponding inverter pair without positive feedback loop. (b) The small signal model to calculate the current difference flowing through the load capacitors C_l.

Now suppose the sensing transistors are ideal without mismatch, but the initial voltage difference V_{sign} changes from zero to V_{os}. In this case the current difference gives

$$\Delta I_2 = i_a - i_b = (g_{mn} + g_{mp}) \cdot V_{os} \tag{4.15}$$

Obviously when ΔI_1 equals ΔI_2, the two cases must have the same sensing outcome. As a result,

$$V_{os} = \frac{g_{mn}}{g_{mn} + g_{mp}} \cdot \Delta V_{thn} + \frac{g_{mp}}{g_{mn} + g_{mp}} \cdot \Delta V_{thp} \tag{4.16}$$

Because of the linear relationship between V_{os} and the mismatch values ΔV_{thp}, ΔV_{thn} in Eqn. (4.16), provided ΔV_{thn} and ΔV_{thp} are independent Gaussian distributed variables, V_{os} will also follow a Gaussian distribution [11] with

$$\begin{cases} \mu_{vos} = \mu_{\Delta vthn} \cdot g_{mn}/(g_{mn} + g_{mp}) + \mu_{\Delta vthp} \cdot g_{mp}/(g_{mn} + g_{mp}) \\ \sigma^2_{vos} = \sigma^2_{\Delta vthn} \cdot [g_{mn}/(g_{mn} + g_{mp})]^2 + \sigma^2_{\Delta vthp} \cdot [g_{mp}/(g_{mn} + g_{mp})]^2 \end{cases} \tag{4.17}$$

Eqn. (4.17) implies that the variance of V_{os} is independent of the polarities of both ΔV_{thn} and ΔV_{thp}. Furthermore, since $\mu_{\Delta vthn}$ and $\mu_{\Delta vthp}$ are zero, μ_{vos} is zero as well. As a consequence, V_{os} follows a Gaussian distribution with variance being a weighted sum of the both threshold voltage variances of n- and p- sensing transistors. The weighted variance results in a value that is smaller than the minimum mismatch of n- and p-sensing pairs. Consequently, the simultaneously latched sense amplifier provides a smaller variance than the n- or p-sensing pair with the same transistor size.

$$\sigma^2_{vos,cmos} < min[\sigma^2_{\Delta vthn}, \sigma^2_{\Delta vthp}] \tag{4.18}$$

The outcome also suggests that with CMOS latched sense amplifiers the mid-level sensing have better yield performance than high- or low-level sensing since in the latter two schemes n- or p-sensing pair dominates the sensing process, resulting in a σ_{vos} as large as $\sigma_{\Delta vthn}$ or $\sigma_{\Delta vthp}$.

4.3.3 Yield optimization for mid-level sensing

According to the inequality $a^2 + b^2 \geq 2ab$ (equality for $a = b$), the minimum σ_{vos}^2 can be obtained from Eqn. (4.17) when

$$|\frac{g_{mn}}{g_{mn} + g_{mp}} \cdot \sigma_{\Delta vthn}| = |\frac{g_{mp}}{g_{mn} + g_{mp}} \cdot \sigma_{\Delta vthp}| \tag{4.19}$$

For mid-level sensing V_{eq} equals $V_{dd}/2$. The saturated sensing transistors in the beginning of sensing gives

$$\begin{aligned} g_{mn} &\propto \mu_n \cdot W_n/L_n \cdot (V_{eq} - V_{thn}) \\ g_{mp} &\propto \mu_p \cdot W_p/L_p \cdot (V_{dd} - V_{eq} - |V_{thp}|) \end{aligned} \tag{4.20}$$

Since the threshold voltage mismatch is dominated by the doping concentration fluctuation [59], the standard deviations of mismatch of n- and p-sensing pairs are

$$\sigma_{\Delta vthn} = \sqrt{2}A_n/\sqrt{W_nL_n},\ \sigma_{\Delta vthp} = \sqrt{2}A_p/\sqrt{W_pL_p} \tag{4.21}$$

Here, A_n and A_p are constants describing the relationship between transistor's threshold voltage variance and its gate area; μ_n, μ_p are carrier mobilities of n- and p-transistors; $\sqrt{2}$ in above equations comes from the difference function of the transistor pair. When these equations are taken into Eqn. (4.19), the operating point where the minimum variance of V_{os} is found gives

$$V_{eq,m} = \frac{V_{dd} - |V_{thp} + \alpha V_{thn}|}{1 + \alpha} \tag{4.22}$$

$$\alpha = \frac{A_n \cdot \mu_n \cdot W_n/L_n \cdot \sqrt{W_pL_p}}{A_p \cdot \mu_p \cdot W_p/L_p \cdot \sqrt{W_nL_n}} = \frac{\sigma(V_{thn})\mu_n W_n/L_n}{\sigma(V_{thp})\mu_p W_p/L_p} \tag{4.23}$$

The corresponding minimum variance of V_{os} at $V_{eq,m}$ is

$$\sigma_{vos}^2 = \frac{2\sigma_{\Delta vthn}^2 \sigma_{\Delta vthp}^2}{\sigma_{\Delta vthp}^2 + m\sigma_{\Delta vthn}^2},\ m = \frac{V_{dd} - (2 + 1/\alpha) \cdot V_{thn} + 1/\alpha \cdot |V_{thp}|}{V_{dd} - V_{thn} - |V_{thp}|}. \tag{4.24}$$

It implies that when the widths and lengths of the sensing transistors are given, there is one optimum equalization voltage. Conversely, if the equalization voltage is fixed, width to length ratios of the sensing transistors can be optimized to minimize σ_{vos}^2. For mid-level sensing the initial voltage is around $V_{dd}/2$. To obtain the smallest σ_{vos}^2, from Eqn. (4.22)

$$\alpha = \frac{V_{dd}/2 - |V_{thp}|}{V_{dd}/2 - V_{thn}}. \tag{4.25}$$

With the α and Eqn. (4.23), dimensions for n- and p-sensing pairs can be optimized to achieve highest yield for mid-level sensing.

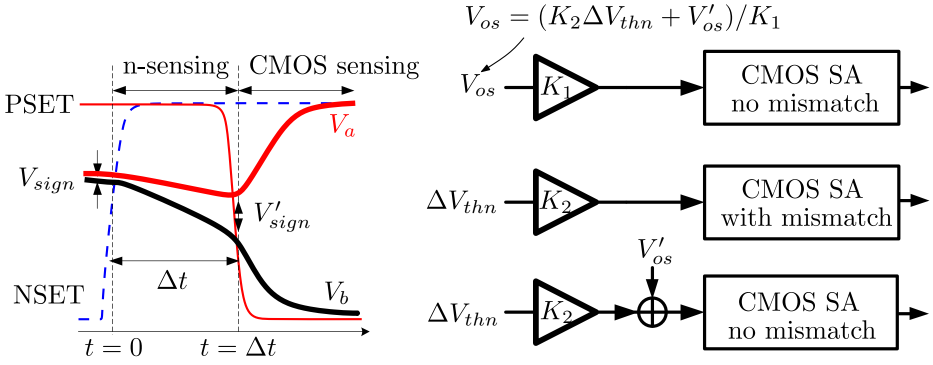


Fig. 4.9: (a) p-sensing pair delayed sensing process. (b) Corresponding linear model for the sensing process in (a).

4.3.4 Switch delay induced yield degradation

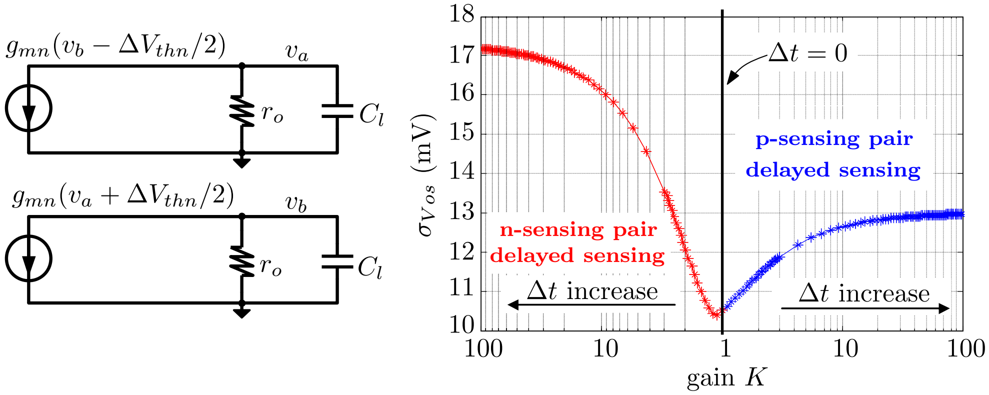
Perfectly simultaneous latching is hard to achieve as n- and p-sensing pairs are controlled by inverted enable signals as shown in **Fig. 4.6**(a). The delay Δt between n- and p-enable signals results in an increase of σ^2_{vos}.

Fig. 4.9(a) shows a p-sensing pair delayed sensing process. At first the sense amplifier works only with n-sensing pair. After Δt, the p-sensing pair is turned on and the entire CMOS sensing is activated. Evidently, the sensing process can be divided into two stages: the n-sensing stage and the CMOS sensing stage. The initial input V'_{sign} of the CMOS sensing stage comes from the outputs of the n-sensing stage at $t = \Delta t$. From $t = 0$ to Δt, the n-sensing pair acts as a gain block amplifying both V_{sign} and ΔV_{thn}. The sensing process can be regarded as equivalent to cascade of two sense amplifiers, an n-sensing pair and a simultaneously latched CMOS sense amplifier, and each one runs for only a certain period of time.

A linearized model to obtain V_{os} corresponding to the above described sensing process is given in **Fig. 4.9**(b). K_1, K_2 represent amplification factors for V_{sign} and ΔV_{thn} respectively during the n-sensing process. They will be obtained later from a small signal model. For the initial input V_{sign} assumed to be zero, the block diagram can be simplified to the middle one in (b). The mismatch of sensing transistors inside the CMOS sense amplifier can then be moved outside as shown in the bottom diagram. When the input offset ΔV_{thn} of the n-sensing pair and V'_{os} of the CMOS sense amplifier are referred back to the signal transfer path with gain K_1, the top diagram is obtained. The equivalent input offset in the CMOS sense amplifier with p-sensing pair delay gives

$$V_{os} = \frac{1}{K_1} \cdot [\Delta V_{thn} \cdot (K_2 + \frac{g_{mn}}{g_{mn} + g_{mp}}) + \Delta V_{thp} \cdot \frac{g_{mp}}{g_{mn} + g_{mp}}] \tag{4.26}$$

Since ΔV_{thn} and ΔV_{thp} follow independent Gaussian distributions, the variance



(a) (b)

Fig. 4.10: (a) Small signal model including threshold voltage mismatch to calculate K_1, K_2 during the n-sensing process (b) σ_{vos} vs. K. K is in logarithmic scale as a representation of linear change of Δt; $K = 1$ corresponds to $\Delta t = 0$. σ_{vos} rises as Δt increases for both n- and p-sensing pair delayed CMOS sensing. When Δt is quite large, σ_{vos} approaches $\sigma_{\Delta vthn}$ and $\sigma_{\Delta vthp}$, respectively.

of V_{os} of a p-sensing pair delayed sense amplifier is

$$\sigma_{vos}^2 = [\frac{g_{mp}}{K_1 \cdot (g_{mp} + g_{mn})}]^2 \cdot \sigma_{\Delta vthp}^2 + [\frac{K_2}{K_1} + \frac{g_{mn}}{K_1 \cdot (g_{mn} + g_{mp})}]^2 \cdot \sigma_{\Delta vthn}^2 \quad (4.27)$$

K_1, K_2 are obtained by small signal analysis of an n-latch pair as shown in **Fig. 4.10**(a). Suppose the initial voltage difference is $V_{sign} = v_a(0) - v_b(0)$ before the n-sensing pair is enabled, where v_a and v_b are small signal voltages of sensing nodes. The differential equations describing these small signal circuits are

$$\begin{cases} C_l \cdot dv_b/dt + v_b/r_o + (v_a + \Delta V_{thn}/2) \cdot g_{mn} = 0 \\ C_l \cdot dv_a/dt + v_a/r_o + (v_b - \Delta V_{thn}/2) \cdot g_{mn} = 0 \end{cases} \quad (4.28)$$

By solving these equations, the output gives

$$\begin{aligned} V'_{sign}(t) &= v_a(t) - v_b(t) \\ &= \Delta V_{thn} \cdot \frac{A}{A-1} \cdot (1 - e^{\frac{A-1}{\tau}t}) + [v_a(0) - v_b(0)] \cdot e^{\frac{A-1}{\tau}t} \\ &= \Delta V_{thn} \cdot K_2(t) + V_{sign} \cdot K_1(t) \end{aligned} \quad (4.29)$$

As a result K_1, K_2 can be expressed as

$$K_1(t) = e^{\frac{A-1}{\tau}t},\ K_2(t) = \frac{A}{A-1} \cdot (1 - e^{\frac{A-1}{\tau}t}) \quad (4.30)$$

Here, $\tau = C_l \cdot r_o$, $A = g_{mn} \cdot r_o \ll 1$. K_1, K_2 at $t = \Delta t$ result from Eqn. (4.30):

- $\Delta t = 0$, $K_1 = 1$, $K_2 = 0$, corresponding to simultaneous CMOS sensing;
- Δt increases, $K_2 = 1 - e^{\frac{A-1}{\tau}\Delta t} \approx -e^{\frac{A-1}{\tau}\Delta t} = -K_1$;
- $\Delta t \to \infty$, $K_2 = -K_1 \to \infty$, corresponding to n-sensing.

In general, Δt is large enough, so that a simple gain K can be used to replace K_1, K_2 in Eqn. (4.27). Since K is an exponential function of Δt as shown in Eqn. (4.30), $log(K)$ becomes a linear function of Δt. σ_{vos} versus K with a logarithmic x axis in **Fig. 4.10**(b) suggests that as the gain K or Δt increases, σ_{vos} has the tendency to reach $\sigma_{\Delta vthn}$ or $\sigma_{\Delta vthp}$ for p- or n-sensing pair delayed CMOS sensing. In the calculation the threshold mismatch of the n- and p-sensing pairs are 13mV and 17mV, respectively.

In analogy with p-sensing pair delayed CMOS sensing in the above evaluation, σ^2_{vos} of a n-sensing pair delayed CMOS sensing can also be obtained as

$$\sigma^2_{vos} = [\frac{g_{mn}}{K_1 \cdot (g_{mp} + g_{mn})}]^2 \cdot \sigma^2_{\Delta vthn} + [\frac{K_2}{K_1} + \frac{g_{mp}}{K_1 \cdot (g_{mn} + g_{mp})}]^2 \cdot \sigma^2_{\Delta vthp} \qquad (4.31)$$

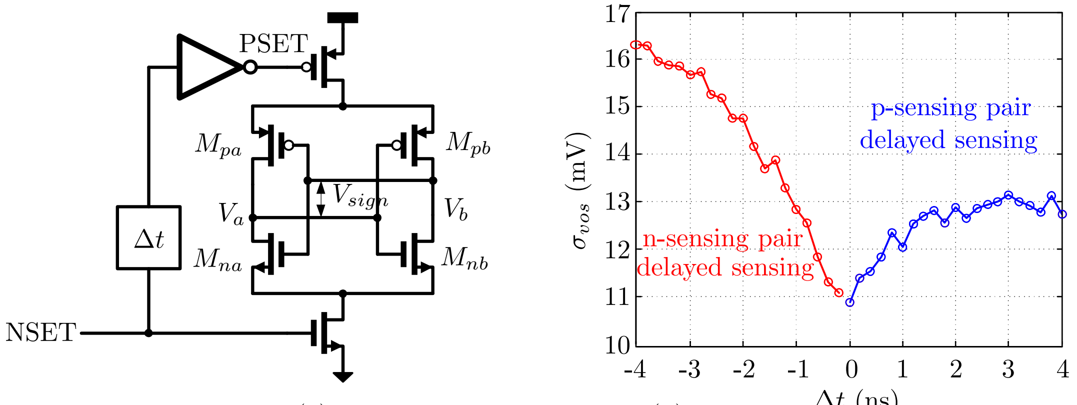
These outcomes are also valid for a high- or low-level sensing as they can be regarded as n- or p-sensing pair delayed CMOS sensing process. In high- or low-level sensing the equivalent input offset V_{os} will be identical to ΔV_{thn} or ΔV_{thp}, and thus they are inferior to mid-level sensing in yield performance.

4.3.5 Comparisons with SPICE simulations

Monte Carlo (MC) SPICE simulations are used to obtain the failure counts corresponding to a given V_{sign}. To make the simulation outcome comparable to theoretical calculations, a simulated yield probability is obtained from these failure counts and transformed into variance by inverse error function as shown in Eqn. (4.32) thanks to their known Gaussian property.

$$\mathrm{Y}(V_{sign}) = \frac{1}{2}[1 + erf(\frac{V_{sign}}{\sqrt{2}\sigma})],\ \sigma = \frac{V_{sign}}{\sqrt{2}erf^{-1}(2\mathrm{Y} - 1)} \qquad (4.32)$$

The simulation setup is shown in **Fig. 4.11**(a) with V_{sign} being swept from $-50mV$ to $50mV$. To make it simple, the correct sensing output is supposed to be logic "0". Theoretically without mismatch, $V_{sign} < 0$ gives 100% yield while $V_{sign} > 0$ results in 0% yield. With mismatch, the total number of failures at a given V_{sign} can be counted after each MC simulation and evidently the yield will deviate from 100% or 0%. In (a), a delay block is inserted between the enable signals PSET and NSET for n- and p-sensing transistors, so that the switch



(a) (b)

Fig. 4.11: (a) Test circuit setup: Δt is set to 0 for simultaneously latched CMOS sensing, to negative for n-sensing pair delayed and positive for p-sensing pair delayed sensing. (b) Simulated σ_{vos} vs. Δt shows the same curve and range as Fig. 4.10(b).

delay can be adjusted from negative to positive, in accordance to different sensing processes: from p-sensing gradually to simultaneous CMOS sensing, finally n-sensing. In the simulations V_{eq} is set to half V_{dd}, corresponding to mid-level sensing. The simulated σ_{vos} vs. Δt is plotted in (b). It verifies the theoretical calculations from **Fig. 4.10**(b): as the delay varies from negative to positive, σ_{vos} drops from $\sigma_{\Delta vthp}$ (p-sensing) to minimum (simultaneously CMOS sensing), then rises to $\sigma_{\Delta vthn}$ (n-sensing). It can also be noticed that when the switch delay Δt is within ±1ns, σ_{vos} increases less than 10% from its minimum value. As a result, the maximum Δt should be kept within 1ns to maintain the yield benefit of CMOS latched sense amplifiers in practical design.

Fig. 4.12(a) depicts the failure probability vs. V_{sign} with Δt being set to $-5ns$, 0 and $5ns$, respectively. Due to smaller σ_{vos} in simultaneous CMOS sensing its slope angle is much larger than other competitors. When Δt is set to $-5ns$, $5ns$ corresponding to p- and n-sensing, σ_{vos} is found to be almost equal to $\sigma_{\Delta vthn}$ and $\sigma_{\Delta vthp}$, which are 13mV and 17mV in **Fig. 4.11**(b). σ_{vos} of the latched CMOS sensing calculated by Eqn. (4.16) predicts a value of 10.8mV with 13mV $\sigma_{\Delta vthn}$ and 17mV $\sigma_{\Delta vthp}$, which is in very good agreement with simulations when Δt is zero.

Simulated σ_{vos} vs. equalization voltage V_{eq} is shown in **Fig. 4.12**(b) together with the calculated σ_{vos} obtained from Eqn. (4.16) and Eqn. (4.20). This time the switch delay Δt is set to zero. High or low V_{eq} corresponds to high- or low-level sensing. From the plot the minimum variance is found to occur around 0.55V, which corresponds to the estimated $V_{eq,m}$ from Eqn. (4.22). The trends are evident from the plot that when V_{eq} shifts away from $V_{eq,m}$, σ_{vos} will eventually approach $\sigma_{\Delta Vthn}$ or $\sigma_{\Delta vthp}$. Slight deviations are found near the edges of the plot.

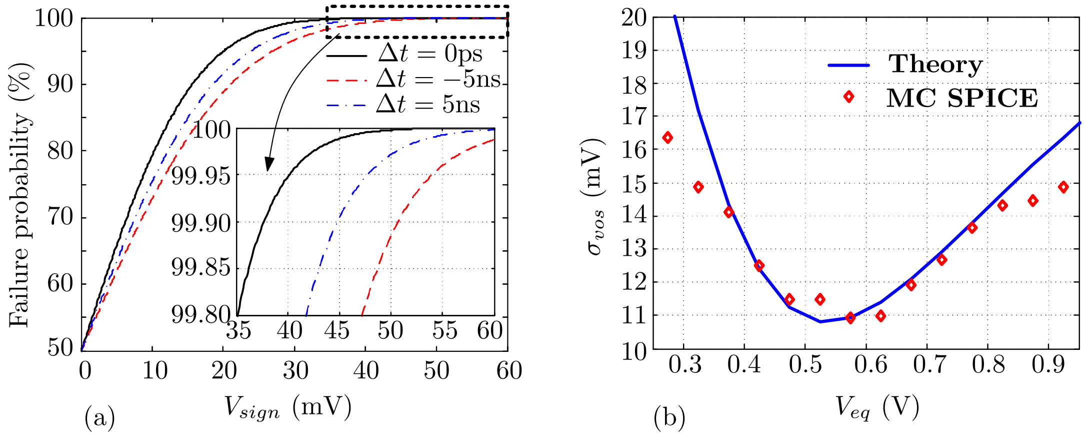


Fig. 4.12: (a) Comparison of simulated failure probability vs. V_{sign} corresponding to $\Delta t = -5ns$, $\Delta t = 0ns$ and $\Delta t = 5ns$. The inset provides the failure probability near 99.9% for the three cases. (b) Calculated and simulated σ_{vos} vs. V_{eq} show good agreements in the region from 0.4 to 0.8V with $V_{dd} = 1.2V$

4.3.6 Conclusion

The mismatch of CMOS sense amplifiers is modeled as a statistical voltage source V_{os} in front of the sense amplifier. The delay between enable signals for n- and p-sensing pairs is also analyzed by a linearized model. The results suggest that simultaneously latched CMOS sense amplifiers have superior yield performance in comparison with n- or p-latch pairs. Based on the mismatch equivalent input offset model, more complex yield analysis of a DRAM sensing process with multiple random sources shall be carried out in following chapters. The input offset model provides an effective and efficient way for yield optimization of DRAM sensing schemes and sense amplifiers.

4.4 Summary

In this chapter, circuit yield is obtained in an analytical way based on the Central Limit Theorem and Gaussian approximations. This method is quite useful for any linearized circuits with Gaussian approximated statistical parameters. Different random error sources in DRAM are also addressed but as the first step, the yield of DRAM sense amplifiers has been carefully investigated by introducing the input offset. Based on the approximations and the sense amplifier mismatch model, a more practical and universal yield model for DRAMs will be set up in the following chapters.

Chapter 5

Leakage Induced Yield Degradation

5.1 Introduction

Leakage currents in DRAM cells are always investigated from the view of device physics [75, 76] in order to prolong the charge retention time of DRAM cells. However, the relationship between cell leakage and yield is seldomly addressed due to the fact that until now there is no proper analytical yield model. With the scaling of cell feature size down to tens of nanometers, the high electric field accompanied tunneling leakage becomes more evident due to the tiny and compact cell. These effects tightly limit the yield for high volume DRAM products and raise the average cost per bit as a result of increasing redundancies.

In this chapter, a statistical analytical model including sense amplifier mismatch and cell leakage for electric yield analysis of DRAM core circuits is formulated. Different from traditional methods [68], the statistical model provides a direct relationship between the statistical parameters and sensing yield (failures), demonstrating a method of statistical DRAM core design in consideration of random parameters. It can be further applied to optimizing DRAM core circuit considering area, yield trade-off and facilitate the design of the array redundancies. The results from the model are compared to both Monte Carlo (MC) simulations and silicon measurements, showing very good agreements.

5.2 Leakage in DRAM Array

Leakage sources in DRAM cells are usually investigated by means of measurements [77, 78] and device simulations [75]. They are found to result from band to band tunneling effect and often vary accordingly with cell contents. For example, sub-threshold leakage I_{sub} and storage node junction to well leakage I_j are storage node voltage dependent, therefore '1' and '0' cell states give rise to different leakage currents. Noticeably, different leakage sources may exhibit distinct characteristics: The sub-threshold leakage I_{sub} is worse for '0' cells while the other leakage sources are significant for '1' cells; The junction to well leakage I_j is highly sensitive to temperature variation; The gate induced drain leakage (GIDL) is gate-drain voltage V_{gd} dependent. In general, when the contributions of the leakage sources to failures are considered, the leakage components are used to be put together

$$I_{leak} = I_j + I_{sub} + I_{GIDL} + I... \tag{5.1}$$

An old rule of thumb in DRAM technology design claims that the average of I_{leak} should be less than 1fA [16], which is six orders of magnitude smaller than high performance logic MOSFET circuits. For nowadays giga-bits per chip design, this rule becomes even more stringent.

5.3 Statistical Analysis of Leakage Effects

5.3.1 Sensing process with leakage

As soon as DRAM cells are loaded with '0's or '1's and isolated from active sources, the leakage components begin to take effect with the elapse of time by reducing the charge in cell capacitors. After time Δt, storage node voltage will drop from its initial value V_{cell} to

$$V'_{cell} = V_{cell} - \frac{I_{leak}\Delta t}{C_s} \tag{5.2}$$

Here, to simplify the analysis, all cells are supposed to be in '1' states and I_{leak} is greater than zero. When they are accessed, the developed signal amplitudes during pre-sensing give

$$V_{sign} = K_t \cdot (V_{cell} - V_{eq} - \frac{I_{leak}\Delta t}{C_s}) \tag{5.3}$$

where K_t is transfer ratio, meaning the array's ability of generating V_{sign} from cell charge during pre-sensing as mentioned in Chapter 2. Additionally, from Section 4.3 the mismatch of sensing transistors can be modeled as an input offset voltage V_{os} with Gaussian distribution. As a result, when mismatch of sense amplifiers is taken into consideration, the final effective voltage difference seen by an ideal sense amplifier will be

$$V'_{sign} = K_t \cdot (V_{cell} - V_{eq} - \frac{I_{leak}\Delta t}{C_s}) + V_{os} \tag{5.4}$$

Once V'_{sign} shows different polarity from the initial $V_{cell} - V_{eq}$ at a given sampling time Δt, the sensing fails. Therefore, in traditional DRAM design V_{os}, I_{leak} are made as small as possible, while a larger K_t is expected to minimize the impact of V_{os}. However, with ultra small leakage currents of less than 1fA , 30fF cell capacitance and 20ms retention time, the maximum cell voltage loss is only 0.67mV. Compared with over 400mV $V_{cell} - V_{eq}$ in the most advanced DRAM technology, the leakage introduced voltage loss is still negligible. What on earth is the real reason that the 1fA rule has to be followed up in every generation of DRAM technology?

5.3.2 Yield plot without retention time

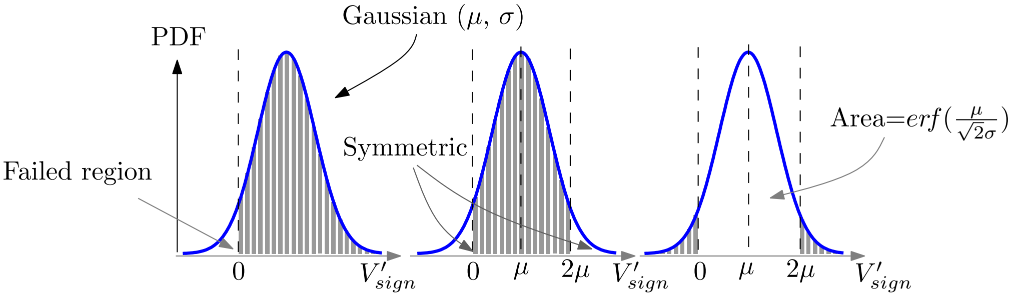
Suppose that $V_{cell} - V_{eq}$ barely fluctuates and the statistical variation of V'_{sign} comes only from I_{leak} and V_{os}. Eqn. (5.4) can be used to evaluate the distribution of V'_{sign} as follows. Clearly, in case of zero retention time, the leakage current will have no effect on the developed signal amplitude V'_{sign}. Since V_{os} follows Gaussian distribution as obtained from Chapter 4, V'_{sign} will also follow a Gaussian distribution with mean μ equal to $K_t(V_{cell} - V_{eq})$ and variance σ^2_{vos}. Similar to the yield calculation process in Section 4.1, the sensing yield probability for '1' state cells can be evaluated by using the error function as shown in Eqn. (4.12) from **Fig. 5.1**.

$$\text{Y} = P\{V'_{sign} > 0\} = \frac{1}{2}[1 + erf(\frac{\mu}{\sqrt{2}\sigma_{vos}})], \tag{5.5}$$

and the inverse error function from Eqn. (5.5) gives

$$-erf^{-1}(1 - 2\text{Y}) = \frac{\mu}{\sqrt{2}\sigma_{vos}} = \frac{K_t(V_{cell} - V_{eq})}{\sqrt{2}\sigma_{vos}}. \tag{5.6}$$

Eqn. (5.6) implies that in a $erf^{-1}(1 - 2F)$ vs. V_{cell} plot a straight line will be observed, if σ_{vos}, K_t and V_{eq} are constants independent of V_{cell}. Similarly, the



$$\text{Y}= P\{V'_{sign} > 0\} = 1 - \tfrac{1}{2}P\{|V'_{sign} - \mu| > 0\} = \tfrac{1}{2}[1 + erf(\tfrac{\mu}{\sqrt{2}\sigma})]$$

Fig. 5.1: Histogram of V'_{sign} without leakage

following yield equation can be obtained for the sensing yield of '0' cells. The relationship between the failure probability F, $erf^{-1}(1-2F)$ and corresponding μ in a Gaussian distribution is shown in **Table 5.1**.

$$-erf^{-1}(1-2\mathrm{Y}) = \frac{\mu}{\sqrt{2}\sigma_{vos}} = \frac{K_t(V_{eq}-V_{cell})}{\sqrt{2}\sigma_{vos}}. \tag{5.7}$$

Table 5.1: The relationship between Failure probability F, $erf^{-1}(1-2\mathbf{F})$, and corresponding μ of a normal distribution

F	15.87%	2.28%	134.99ppm	31.67ppm	286.65ppb	986.59ppt	1.28ppt
$erf^{-1}(1-2\mathrm{F})$	0.707	1.414	2.121	2.828	3.536	4.243	4.950
μ	1σ	2σ	3σ	4σ	5σ	6σ	7σ

(ppm: parts per million; ppb: parts per billion; ppt: parts per trillion)

5.3.3 Yield degradation with retention time

Obviously in a realistic environment retention time should be taken into account. When the retention time Δt is greater than zero, the distributions of leakage sources must be included. Due to the multiplication factor $K_t\Delta t/C_s$ in Eqn. (5.4) the leakage distribution is magnified and becomes more apparent with the increase of Δt. Here, sub-threshold leakage is taken as an example. First of all, suppose mismatch of sense amplifiers are neglected and sub-threshold leakage is the only random variability. Suppose cell transistors are n-type and follow the sub-threshold equation from [46, 47]

$$I_{sub} = I_0(exp\frac{V_{gs}-V_{th}}{\xi V_T})(1-exp\frac{-V_{ds}}{V_T}) \tag{5.8}$$

where $V_{ds} = V_{cell} - V_{eq}$, $V_{gs} = V_{wl} - V_{cell}$ for '0' cells and $V_{gs} = V_{wl} - V_{eq}$ for '1' cells. ξ is a capacitor ratio determined by the capacitances of gate oxide and depletion layer. V_T is equal to kT/q, where k, T are Boltzmann constant and absolute temperature, respectively. Obviously, in case $V_{cell} - V_{eq}$ is three to four times more than V_T, I_{sub} is approximately independent of V_{ds}

$$I_{sub,max} = I_0 exp\frac{V_{gs}-V_{th}}{\xi V_T},\ I_0 = \mu C_d\frac{W}{L}V_T^2 \tag{5.9}$$

Due to random dopant effects during fabrication, the threshold voltage V_{th} of cell transistors follows a Gaussian distribution [65, 59]. From Eqn. (5.9) I_{sub} is an exponential function of V_{th}, and therefore I_{sub} exhibits a *Log-Normal* [79] distribution with long tail on the side of larger leakage current. The V_{th} to I_{sub} probability density function transformation is demonstrated in **Fig. 5.2**. When

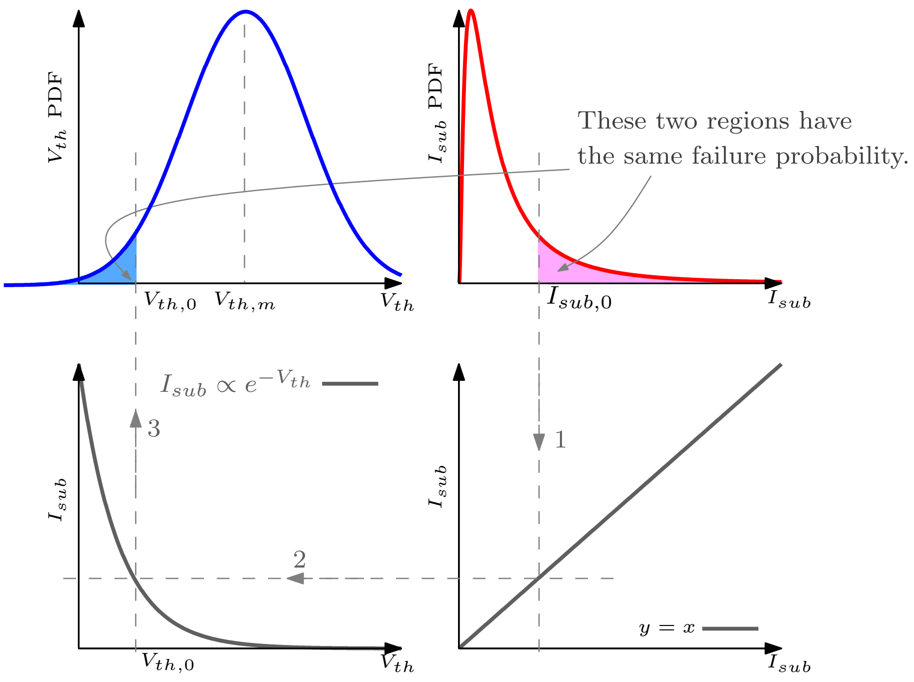


Fig. 5.2: Distribution and failure region transformation from V_{th} to I_{sub}

mismatch of sense amplifiers is neglected, from Eqn. (5.4) V'_{sign} follows an up-scaled distribution of I_{sub}.

Suppose the initial cell voltage satisfies $V_{cell} - V_{eq} > 0$ ('1' state). By probability algorithms, without sense amplifier mismatch the yield probability is known to be

$$\begin{aligned} \mathrm{Y} &= P\{V'_{sign} > 0\} \\ &= P\{K_t \cdot (V_{cell} - V_{eq} - I_{sub}\Delta t/C_s) > 0\},\ (K_t > 0) \\ &= P\{I_{sub} < (V_{cell} - V_{eq})C_s/\Delta t\} \end{aligned} \tag{5.10}$$

Let $I_{sub,0} = (V_{cell} - V_{eq})C_s/\Delta t$. $I_{sub,0}$ correlated $V_{th,0}$ can be obtained from Eqn. (5.9). As $I_{sub} \propto e^{-V_{th}}$ is a monotonically decreasing function, according to **Fig. 5.2** the yield probability in Eqn. (5.10) is equal to

$$\mathrm{Y} = P\{V_{th} > V_{th,0}\} \tag{5.11}$$

Furthermore, according to Eqn. (5.9) the threshold voltage corresponding to $I_{sub,0}$ is

$$V_{th,0} = V_{gs} - \xi V_T \ln \frac{I_{sub,0}}{I_0} = V_{gs} - \xi V_T \ln \frac{(V_{cell} - V_{eq})C_s}{I_0 \Delta t} \tag{5.12}$$

Therefore, the yield probability in Eqn. (5.11) can be evaluated by using the error

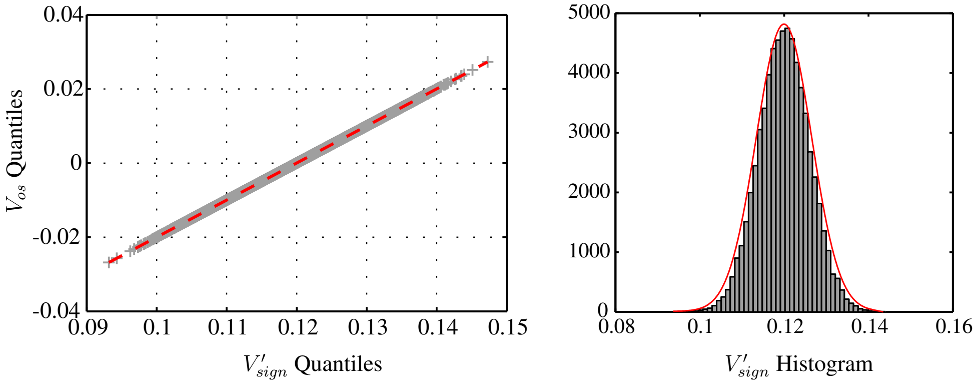


Fig. 5.3: Quantile and histogram of V'_{sign} with I_{sub} and V_{os} included.

function, giving

$$\mathrm{Y} = \frac{1}{2}[1 + erf(\frac{V_{th,m} - V_{th,0}}{\sqrt{2}\sigma_{vth}})] \tag{5.13}$$

where $V_{th,m}$, σ_{vth} are the mean and variance of the threshold voltage of cell transistors, respectively. Evidently, since the error function is a monotonically increasing function, larger threshold voltage fluctuations give rise to lower yield. When the threshold variance σ_{vth} is a constant, the inverse error function of yield Y gives

$$\begin{aligned} -erf^{-1}(1 - 2\mathrm{Y}) &= \frac{V_{th,m} - V_{th,0}}{\sqrt{2}\sigma_{vth}} \\ &= \frac{\xi V_T}{\sqrt{2}\sigma_{vth}} \ln \frac{(V_{cell} - V_{eq})C_s}{I_0 \Delta t} + \frac{V_{th,m} - V_{gs}}{\sqrt{2}\sigma_{vth}}. \end{aligned} \tag{5.14}$$

The equation suggests erf^{-1} is a linear function of $\ln \Delta t$ when other parameters are constants.

Now take sense amplifier mismatch into consideration as well. When V_{os} is present, the probability density function (PDF) of V'_{sign} becomes the convolution of both V_{os} and I_{sub} PDF functions:

$$PDF_{V'sign} = \int_{-\infty}^{\infty} f_{vos}(y - x) f_{Isub}(x) dx \tag{5.15}$$

where f_{vos}, f_{Isub} are PDFs of input offset V_{os} and cell leakage current, respectively. The convolution process complicates the evaluation of yield/failure probability. However, since the distribution of I_{sub} exhibits very large spread compared to the distribution of V_{os} especially in the tail part, similar to spectrum characteristic in finite time pulse sampling in signal processing, the convolution can hardly change the tail's shape, i.e., when failures appear in the tail, failure events become small probability events and the failure probability can be evaluated from

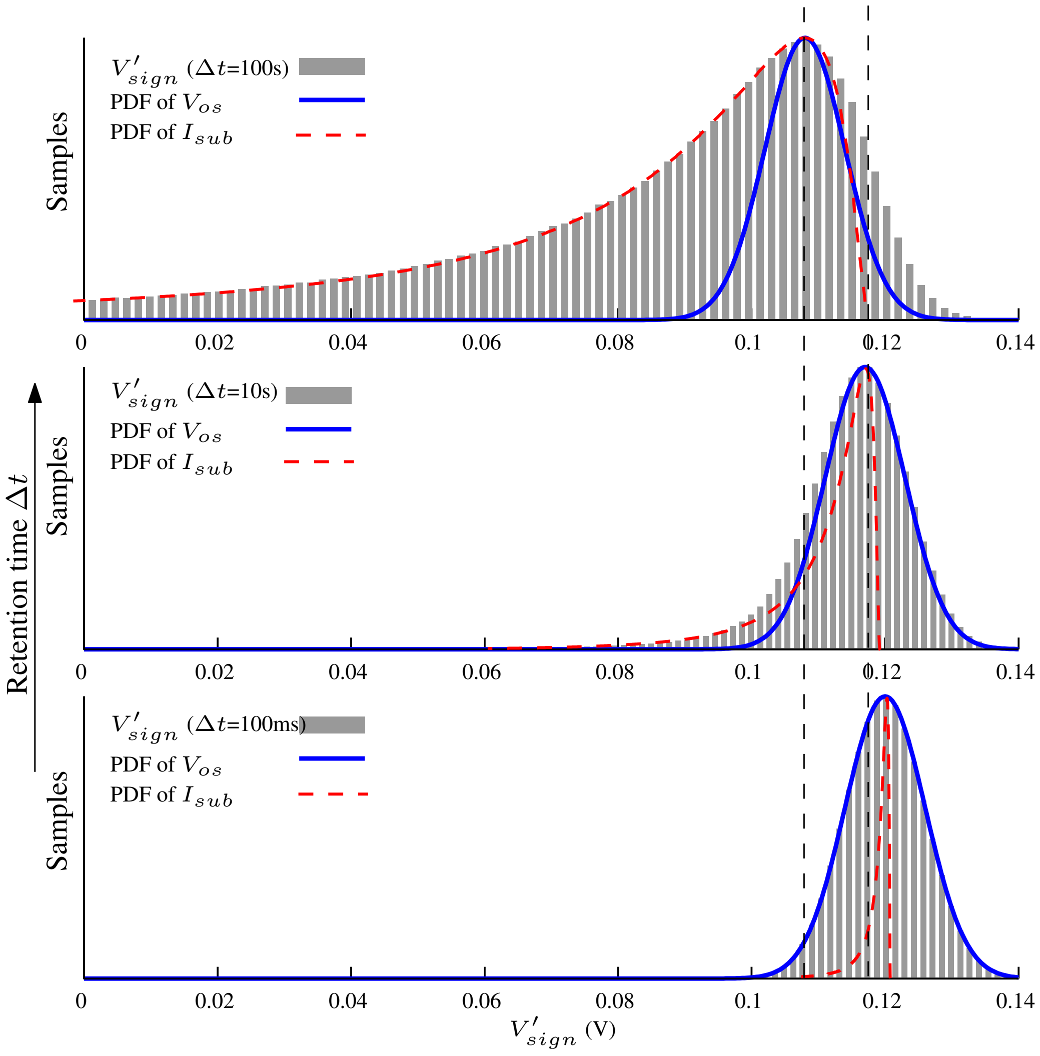


Fig. 5.4: The simulated histogram of V'_{sign} at $\Delta t = 100$ms, 10s, and 100s.

the distribution of I_{sub}. On the other side, due to the exponential function of I_{sub}, the distribution near the mean value $I_{sub,m}$ is very narrow as shown in **Fig. 5.2**, making the convolution outcome similar to the distribution of V_{os}. **Fig. 5.3** shows a 2^{16} samples Monte-Carlo (MC) simulated V'_{sign} histogram with a 0.044fA averaged sub-threshold leakage source, σ_{vos}=10mV Gaussian distributed V_{os}, 20ms retention time and $V_{cell} - V_{eq} = 0.6$V. The quantile plot proves the assumption that the body part follows a Gaussian distribution with variance close to σ^2_{vos}. Because of the limited number of samples, the tail part does not appear in the MC simulation but it surely exists as will be shown in **Fig. 5.4**.

Fig. 5.4 shows the simulated histogram of V'_{sign} corresponding to different retention time Δt (2^{24} samples). When Δt is small, e.g., less than 100ms, the distribution of V'_{sign} seems to follow a Gaussian distribution with variance approximating to σ_{vos}^2. Actually, there is a very long tail extending to -0.04V in the simulation outcome, though it can not be observed in the histogram plot due to the small quantity. With the increase of Δt, the distribution of the tail in V'_{sign} grows larger and becomes more significant as shown in the middle plot. With even larger $\Delta t = 100$s, the leakage distribution will dominate the distribution of V'_{sign} completely as shown in the top one. The shift of the average of V'_{sign} for different Δt is determined by $I_{sub,m}\Delta t/C_s$. In the demonstration, as the average of I_{sub} is 0.044fA the shifts corresponding to $\Delta t = 10$s, $100s$ are approximately 1.5mV and 14.6mV from theoretical calculations. Since the failures come from the tail distribution of V'_{sign}, their probability can be estimated from Eqn. (5.14) of the leakage distribution.

In normal operations with retention time less than 64ms and a 30fF cell capacitor, simulations found that the variance of the body part follows a Gaussian distribution with variance σ_{vos}^2. As a result, the failure probability will be determined by the distribution of the leakage source and this is the answer to the 1fA DRAM leakage rule.

Fig. 5.5 demonstrates the relationship between the distributions of V'_{sign} and the cell initial voltage V_{cell} with 100mS retention time. When V_{cell} is far from the equalization voltage V_{eq}, the failure region where $V'_{sign} < 0$ is determined by the leakage distribution. However, with the decrease of the initial cell voltage V_{cell}, the main distribution of V'_{sign} move towards y axis and eventually the main Gaussian distribution dominates the failure probability as shown in the top plot with $|V_{cell} - V_{eq}| = 100$mV.

5.3.4 Effect of multiple leakage sources

In the above discussions, the sub-threshold leakage I_{sub} is taken as an example. In fact, other leakage components coming from either band-band tunneling or traps in Si/SiO$_2$ interface will also follow distributions similar to I_{sub} [78, 77] as they can be expressed in

$$I = I_1 exp - \frac{\alpha E_q}{V_T} \tag{5.16}$$

where E_q represents band to band spacing, e.g., the spacing between trap band and conduction band $E_t - E_c$, and is affected by doping concentration fluctuations. Like the threshold voltage of transistors, generally E_q is considered to follow Gaussian distributions. Consequently, the distributions of these leakage currents will also present tail parts. By comparison with Eqn. (5.9) and Eqn. (5.14) the inverse error function of yield probability gives

$$-erf^{-1}(1 - 2\mathrm{Y}) = \frac{V_T}{\sqrt{2}\alpha\sigma_{Eq}} \ln \frac{(V_{cell} - V_{eq})C_s}{I_1 \Delta t} + \frac{E_{q,m}}{\sqrt{2}\alpha\sigma_{Eq}} \tag{5.17}$$

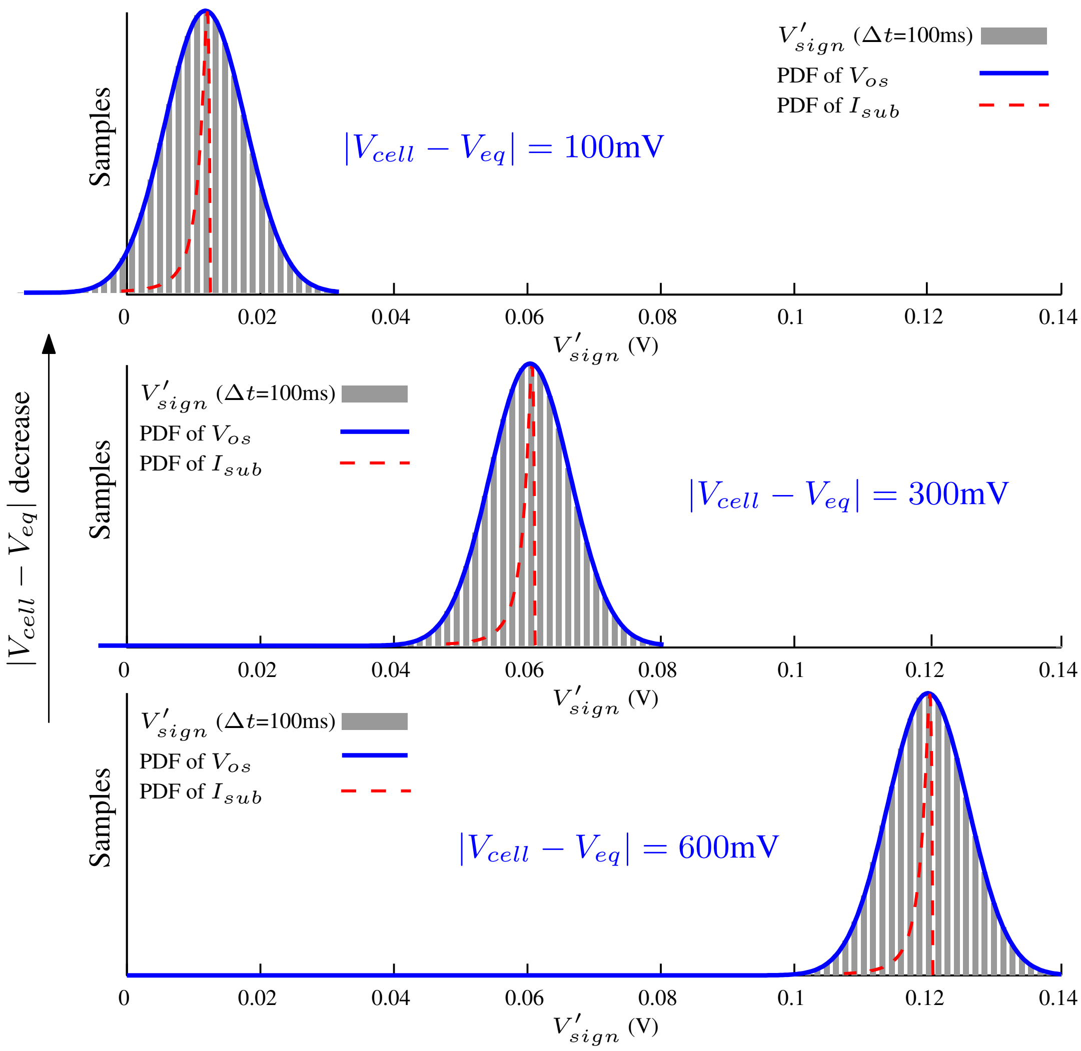


Fig. 5.5: The simulated histogram of V'_{sign} with different initial cell voltages.

As a general conclusion for leakage sources, Eqn. (5.14) and Eqn. (5.17) both show the same characteristic in an erf^{-1} plot:

$$-erf^{-1}(1-2\text{Y}) = a_0(\ln\frac{(V_{cell}-V_{eq})C_s}{I_1\Delta t}) + a_1 \tag{5.18}$$

where

$$\left\{\begin{aligned} a_0 &= \frac{\xi V_T}{\sqrt{2}\sigma_{vth}} \\ a_1 &= \frac{V_{th,m}-V_{gs}}{\sqrt{2}\sigma_{vth}} \end{aligned}\right., \text{ or } \left\{\begin{aligned} a_0 &= \frac{V_T}{\sqrt{2}\alpha\sigma_{Eq}} \\ a_1 &= \frac{E_{q,m}}{\sqrt{2}\alpha\sigma_{Eq}} \end{aligned}\right.. \tag{5.19}$$

In the presence of several leakage sources, the statistical distribution of the total leakage I_{leak} becomes more complicated. However, as these leakage compo-

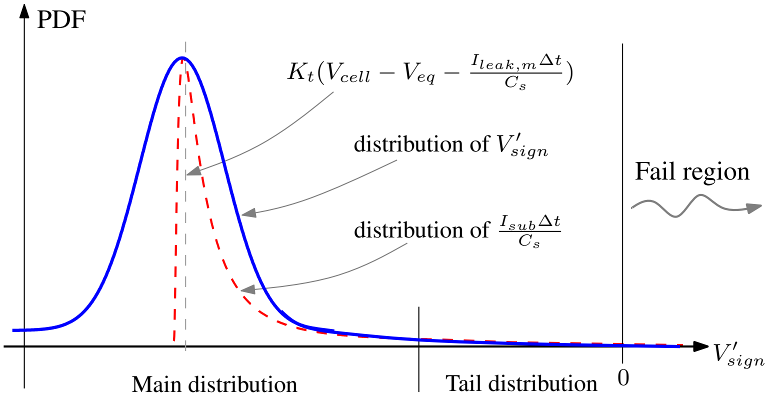


Fig. 5.6: A typical distribution of V'_{sign} within 20ms retention time (The plot is exaggerated to distinguish the differences between the distributions).

nents are distinguishable in their statistical parameters a_0 and a_1, there will be only one dominating the tail part at a given retention time Δt, and therefore the failures generated can still be approximated by using the tail distribution of the dominant leakage source. An example will be shown in Section 5.4. For the tail part, according to Eqn. (5.10) the yield probability including multiple leakage components follows the equation

$$-erf^{-1}(1-2\mathrm{Y}) = a_0[\ln[(V_{cell} - V_{eq} - \frac{I_{leak0,m}\Delta t}{C_s})\frac{C_s}{I_1\Delta t}] + a_1 \tag{5.20}$$

where $I_{leak0,m}$ is the sum of the average leakage currents without the one contributing its tail distribution. It may happen that a leakage source with smaller average value contributes more failures because of its wider tail distribution. Correspondingly, for the body part due to leakage it changes from Eqn. (5.6) to

$$-erf^{-1}(1-2\mathrm{Y}) = \frac{K_t(V_{cell} - V_{eq} - I_{leak,m}\Delta t/C_s)}{\sqrt{2}\sigma_{vos}}, \tag{5.21}$$

where $I_{leak,m}$ is the average of I_{leak} in Eqn. (5.1). Since the inverse error function erf^{-1} is a monotonically increasing function, larger $-erf^{-1}(1-2\mathrm{Y})$ means less failures or higher yield.

The final probability density function of V'_{sign} is plotted in **Fig. 5.6**. With the retention time Δt the center of the V'_{sign} distribution shifts a little from $K_t(V_{cell} - V_{eq})$ and the amount of the shift depends on the retention time and $I_{leak,m}$ which is usually smaller than 1fA.

5.4 MC Simulations vs. Measurements

In this section, Monte-Carlo (MC) simulations with a DRAM sensing behavior model written in C++ are used to verify the previous analysis. The behavior

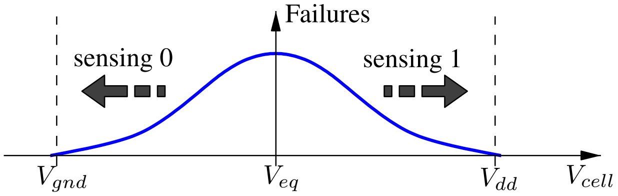


Fig. 5.7: A typical signal margin analysis plot with the cell voltage on x axis and failures on y axis.

model includes a cell capacitor with three leakage components that have different statististical distributions. The array transfer ratio is regarded as a constant independent of the cell capacitance C_s in the MC simulation. Measured yield curves from the *Signal Margin Analysis* [80] (SMA) and retention time analysis for different arrays are also demonstrated, so as to provide concrete evidence for the theoretical outcomes.

5.4.1 Sensing yield vs. Initial cell voltage

The signal margin analysis corresponds to the sensing yield vs. the initial cell voltage simulation by the MC model. As mentioned in Chapter 2, when cells in an array contain regular data sequences like all '0', all '1' or '01' alternating patterns, their developed voltage differences during pre-sensing will be identical in amplitudes due to the regular data pattern and the same cell signal amplitude $|V_{cell} - V_{eq}|$ for all '0' or '1' data. The signal margin analysis utilizes these patterns to verify whether the core circuits meet the electrical yield requirements in the presence of parasitics, crosstalk, and parameter variations.

In such analysis the cell capacitors in an array are charged to a given signal amplitude. For example, in folded bitline arrays the worst pattern is all '0' or all '1' pattern, and therefore the cells will be charged to a voltage level deviating from perfect '1's or '0's before sensing is proceeded. The sensing failures corresponding to the given cell voltage can be obtained. When the voltage is swept over the full voltage range failures can be recorded as shown in **Fig. 5.7**. Since worst case patterns are usually applied to signal margin analysis, this method is very reliable. By the same means, when the '01' alternate pattern is given to an open bitline array and by sweeping the cell signal amplitude, a plot similar to **Fig. 5.7** can be obtained. As will be discussed later, since the expected V_{sign} amplitudes developed for all bitline pairs are identical in the signal margin analysis, the Gaussian approximation can be applied to the array when different variations are involved. As a result, the failure probability from a signal margin analysis actually corresponds to the process described in **Fig. 5.5**.

Apparently, in the signal margin analysis and the MC simulation, if the data retention time is fixed within 64ms, the yield/failure probability exhibits a trend

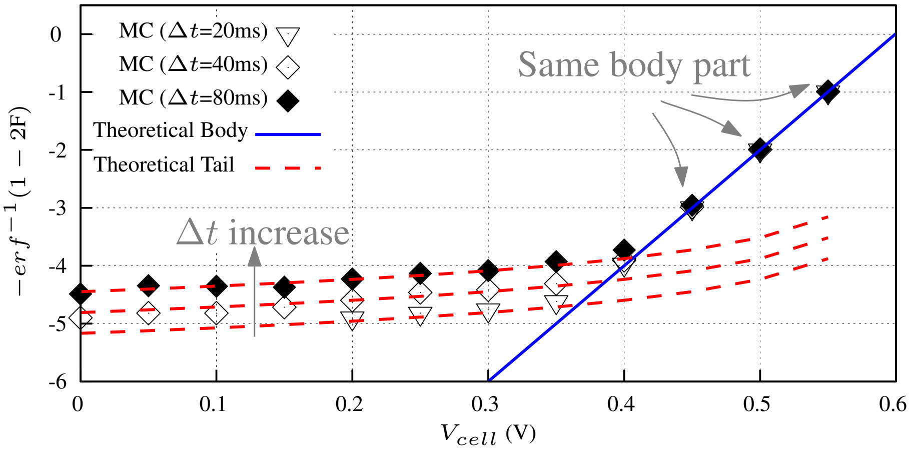


Fig. 5.8: Comparisons between MC simulation and calculations in $erf^{-1}(1-2\mathbf{F})$ vs. V_{cell} plot with different Δt and $V_{eq} = 0.6$V.

that can be observed from **Fig. 5.5**: as the cell signal $|V_{cell} - V_{eq}|$ is much greater, the yield is dominated by the leakage distribution; With the decrease of the cell signal $|V_{cell} - V_{eq}|$, the yield is eventually determined by the sense amplifier mismatch, the array parameters and structures.

In the following plots, as before in order to observe the relationship between the yield/failure and all kinds of parameters, inverse error function is used. In addition, as the yield probability of sensing '1' states is equal to the failure probability of sensing '0' states and vice versa, in following simulations only the '0' states failure probability F is presented and the mismatch equivalent input offset V_{os} is 10mV with $V_{eq} = 0.6$V for all MC simulations.

Fig. 5.8 shows the change of $-erf^{-1}(1-2\text{F})$ versus the initial cell voltage V_{cell} and the retention time Δt. The transfer ratio K_t is a constant independent of cell capacitance C_s in the simulation. From the theoretical model in the previous section, when V_{cell} is close to V_{eq} the sense amplifier mismatch dominates the failure probability; Otherwise, the leakage dominates. As shown in the picture, main and tail parts which come from Eqn. (5.21) and Eqn. (5.20) can be easily distinguished. When the retention time increases exponentially, $-erf^{-1}$ rises linearly, in full accordance with the theoretical calculations. In the MC simulations a total number of 2^{22} samples is included. Obviously, the theoretical calculations are much better than the MC simulations in both speed and accuracy - With Δt=20ms and 2^{22} samples the failure probability in the region $V_{cell} < 0.2$V is already too low for MC simulation to detect, whereas theoretical calculation can reveal the right trace within short time.

The dependency of failure probability on cell capacitance C_s suggested in Eqn. (5.20) is depicted in **Fig. 5.9** with C_s varying from 20fF to 40fF. Because of the logarithmic relationship between $-erf$ and C_s, the probability change due

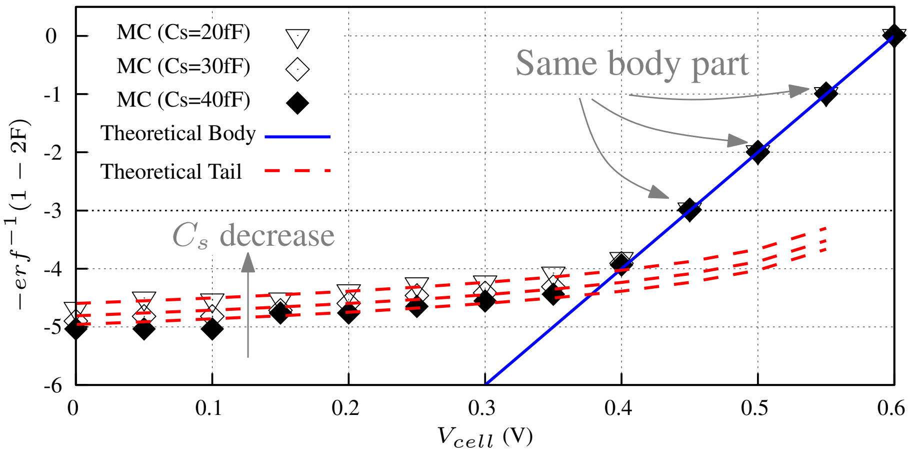


Fig. 5.9: Comparisons between MC simulation and theoretical estimation in $erf^{-1}(1-2\mathbf{F})$ vs. V_{cell} plot with different C_s and $V_{eq} = 0.6$V.

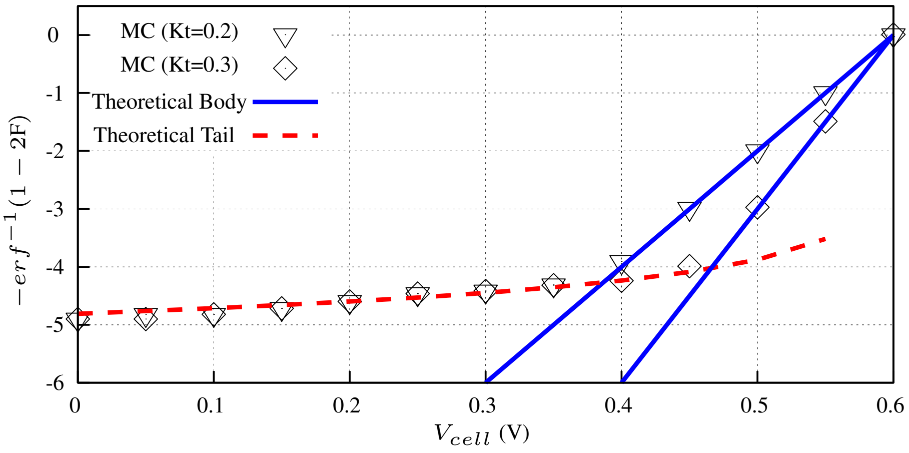


Fig. 5.10: Comparisons between MC simulation and theoretical estimation in $erf^{-1}(1-2\mathbf{F})$ vs. V_{cell} plot with different transfer ratio K_t and $V_{eq} = 0.6$V.

to C_s changing is not significant. Because the array transfer ratio K_t remains constant in the MC simulations, the body parts corresponding to different C_s values overlap each other.

Fig. 5.10 compares the failure probability for different transfer ratios. As described in Eqn. (5.3) K_t is a very important parameter in DRAM core design because it determines the available voltage difference V_{sign} that will be compared to the input offset V_{os} introduced by the sense amplifier mismatch. Since the failures in tail part result from statistical parameters of leakage components from Eqn. (5.20), the MC simulations exhibit separate body parts but overlapping tail parts. The outcome also suggests that in a mature DRAM technology, the number of failures in the tail part, or the required redundancies for a given number of cells can be estimated independently of the array structure and parasitics (e.g.,

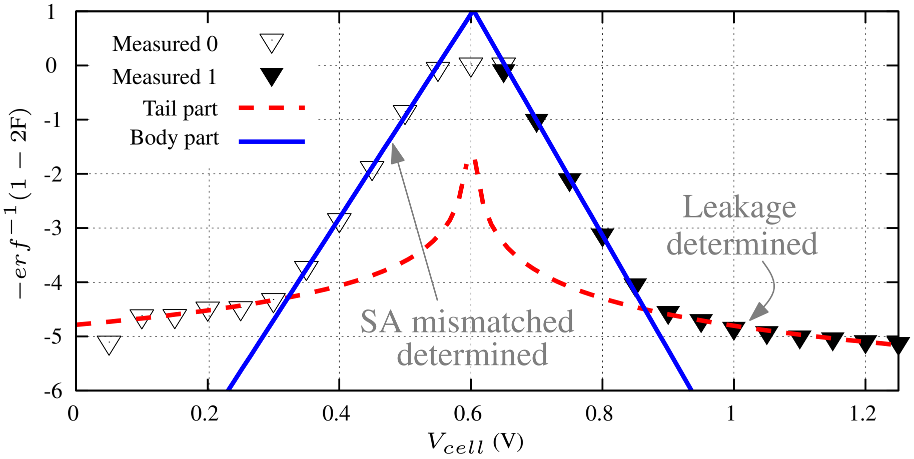


Fig. 5.11: Measured erf^{-1} vs. V_{cell} for a 512Mb DRAM volume product in 90nm technology.

open or folded bitline array, long or short bitlines, bitline parasitics capacitance C_{bl} or the mismatch of sensing transistors) since they mainly originate from cell leakage.

A erf^{-1} versus V_{cell} plot from a signal margin analysis for 90nm volume products is exhibited in **Fig. 5.11**. The plot is obtained by changing the cell storage voltage and recording the corresponding failure counts, then transforming the failure counts to probability as have been done before in the MC simulation plots. Evidently, the body part and the tail part can be fitted to theoretical estimations. σ_{vos} and the leakage properties can be easily obtained from the extracted parameters. The plot also reveals that the leakage in '0' cells is more significant than in '1' cells.

Fig. 5.12 gives another erf^{-1} vs. V_{cell} plot obtained from the signal margin analysis of a 75nm volume product. This time measurements are carried out with two array structures: 512 cells per bitline (LBL) and 256 cells per bitline (SBL) array. Because of the different bitline length and the resulted bitline capacitance C_{bl}, the two arrays have different K_t and as a result their body parts separate from each other. However, since they are based on the same technology, their tail parts show almost the same failure characteristics, in agreement with **Fig. 5.10**. Nevertheless, the asymmetrical characteristics of tails for '0' and '1' sensing suggest that the leakage distributions are different for '0' and '1' cells, which agrees well with the previous analysis. From **Fig. 5.12** the yield/area trade-off can also be noticed. Due to the doubled number of sense amplifiers the SBL array is less area efficient than LBL array. Furthermore, as leakage dominates the failures in the tail part, the electric yield of LBL is almost identical to SBL for V_{cell} close to the rail voltage. In other words, LBL shows a better yield/area trade-off than the SBL design in this region. The same method can be used to optimize the design

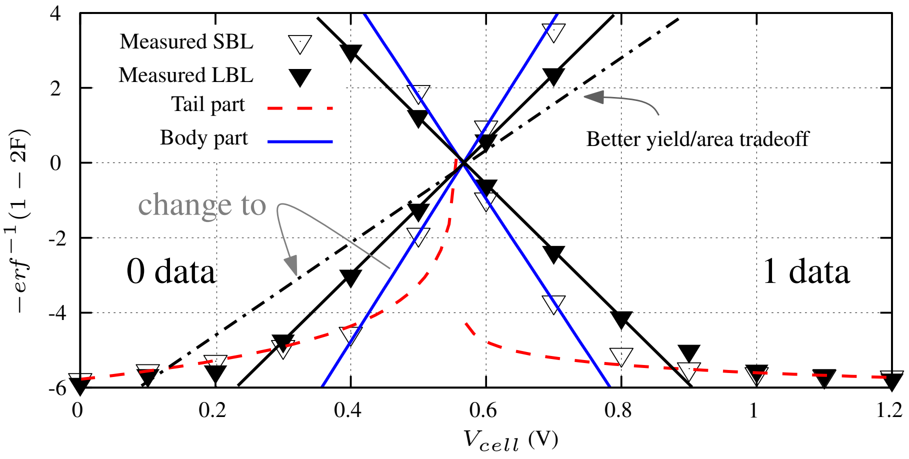


Fig. 5.12: Measured erf^{-1} vs. V_{cell} for a volume DRAM product in 75nm technology with two different arrays structures. (LBL is the long bitline array and SBL is the short bitline array.)

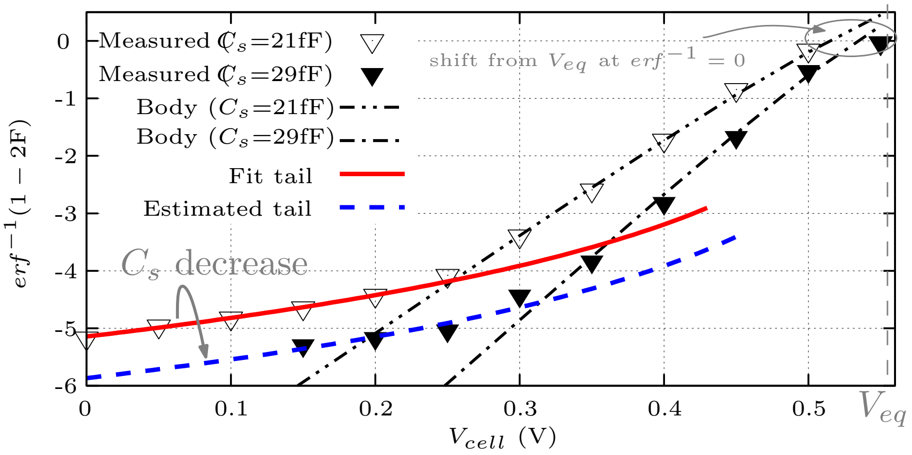


Fig. 5.13: Measured erf^{-1} vs. V_{cell} with larger delay.

of sense amplifiers, because downscaling of sense amplifiers will change the body part while keeping the tail untouched. In a yield/area optimized design, the body part curve can be pushed away from V_{eq} by reducing the width and length of the sensing transistors at the same time. As shown by the dashed lines in **Fig. 5.12** for giga-bit DRAM designs, the tail part and body part in $-erf^{-1}(1-2\text{F})$ plot are both expected to be well below -5, which corresponds to a failure probability less than 1 parts per trillion (ppb), so that the average required redundancies per die is low enough.

As previously expressed in Eqn. (5.21), when retention time Δt rises the body part will shift a little from V_{eq}. **Fig. 5.13** shows a measured erf^{-1} vs. V_{cell} plot with longer Δt. With the same array, sense amplifiers and technology but different values of cell capacitances, besides different slopes due to the change of

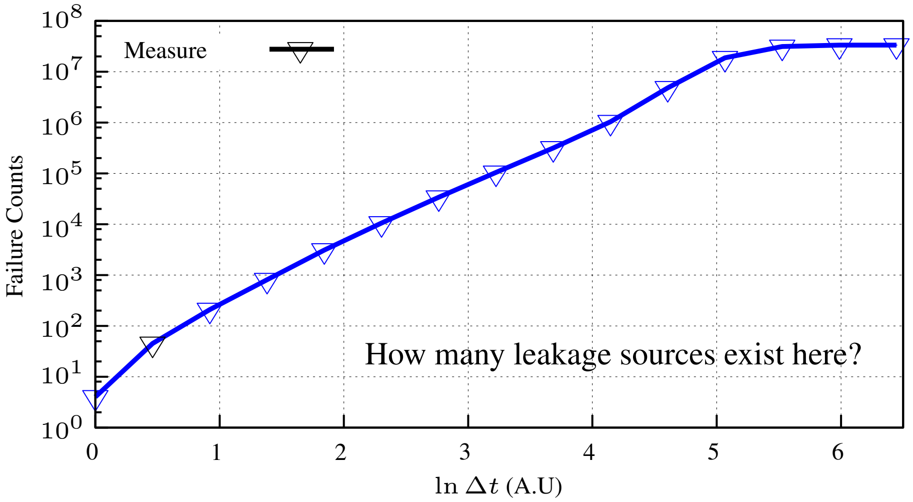


Fig. 5.14: A typical retention time plot with failure counts in y axis and $\ln \Delta t$ in x axis

transfer ratio that is dependent of C_s, the failure probability plot gives different horizontal shifts in their body parts. According to Eqn. (5.21), the shifts are inversely proportional to the values of cell capacitances, which agrees well with the measured outcomes. Due to the limited number of measurable samples, the tail part of 29fF array fails to be detected, while it can be estimated from the tail part of the array with 21fF cell capacitances. By Eqn. (5.20) the 29fF tail part can be obtained as drawn in the figure.

5.4.2 Data retention plot

The retention time analysis is different from the signal margin analysis in that the initial cell voltage V_{cell} stays at V_{dd} and the retention time Δt is swept to observe the corresponding failure probability here. The yield/failure probability is then mainly determined by the statistical characteristics of the cell leakage as implied by **Fig. 5.4**. A typical retention time plot is formed as shown in **Fig. 5.14** obtained from the measurements of a 90nm volume product.

From Eqn. (5.18) it is known that when the failure probabilities are transformed by the inverse error function and plotted against $\ln \Delta t$, several line segments will be observed. Each line segment represents a unique leakage source in cells. **Fig. 5.15** gives the transformation of **Fig. 5.14** together with analytical estimations from Eqn. (5.20) and 2^{24} samples MC simulations with leakage sources I_{s1}, I_{s2} and I_{sub} included. Obviously, the theoretical calculations from Eqn. (5.18) and Eqn. (5.20) show almost the same outcomes as drawn in the plot. It emphasizes the point addressed in the previous section that: a) In a multi-leakage environment only one leakage component dominates at a given Δt; b) All leakage

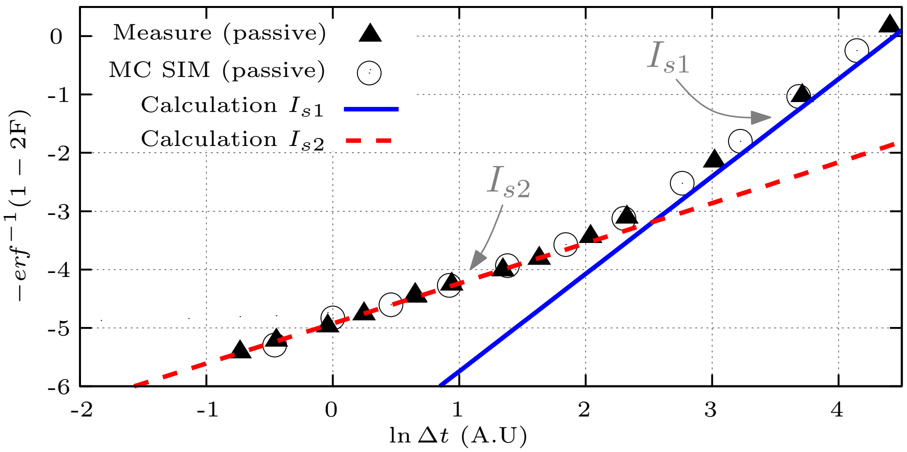


Fig. 5.15: Simulated, measured and estimated retention curves with multiple leakage sources in erf^{-1} axis.

components follow Log-Norm distributions. A small offset between the calculation and MC simulation or measures appears when $\ln \Delta t$ is greater than two. By removal of leakage source I_{s2} in MC simulation, the theoretical calculation is in agreement with MC simulation again. As a result, this offset is caused by I_{s2} that is neglected in the theoretical model in multiple leakage environment. However, this offset is small as shown in **Fig. 5.15**.

It is interesting to see that both retention time and V_{cell} plots may contain the same leakage distribution. Since **Fig. 5.15** and **Fig. 5.11** come from the same volume product, by extracting the tail distribution from **Fig. 5.11**, another dash-dotted line can be drawn in **Fig. 5.16**. Although there is a small shift due to the failure contribution of I_{s1}, it still can prove that the tail in **Fig. 5.11** comes from the leakage source I_{s2} in **Fig. 5.16** because of their almost identical slopes.

By comparing the leakage parameters to some earlier publications [78, 77, 75], neither I_{s1} nor I_{s2} in **Fig. 5.16** belongs to sub-threshold leakage. By an active retention analysis, in which the array is frequently operated a new line segment appears as shown in the unfilled triangles in **Fig. 5.16**. This is supposed to be the sub-threshold leakage I_{sub} of the cell transistors because in active mode wordlines toggle frequently and introduce coupling noise to the transistor gates of retention cells. When the wordline noise is approximated by a Gaussian distribution, it can be directly added to the variance of V_{th}, thus enlarging the distribution of I_{sub} significantly and resulting in a much smaller $|a_0|$ and $|a_1|$. Since in normal operation the retention time is limited with 64ms, I_{sub} is actually more important than others, and therefore a quiet array with less wordline coupling noise is more favored for future DRAM technologies.

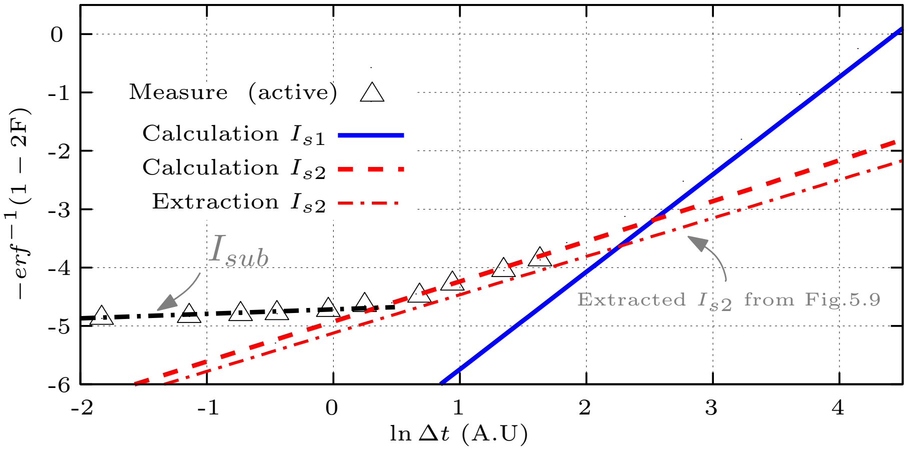


Fig. 5.16: Simulated, estimated and extracted retention curves with multiple leakage sources in erf^{-1} axis (the measured data in passive mode are removed here).

5.5 Summary

In this chapter, the leakage induced failures in DRAM core are modeled and formulated. Different from earlier publications, this model reveals that the sensing failures come from both sense amplifier mismatch and leakage sources. To optimize the yield-cost trade off, both parts need to be taken into consideration. The number of redundancies in DRAM core is largely determined by the leakage source that can be characterized in a retention plot with erf^{-1} axis. In addition, the model reveals that the leakage induced failures are strongly technology dependent, i.e., they are almost independent of sensing methods, array structures (bitline structures), signal transfer ratio K_t or sense amplifier mismatch. In a mature DRAM technology, the required number of on die redundancies can be estimated according to Eqn. (5.20).

A DRAM design-for-yield procedure is implicitly proposed. In order to optimize the yield, it is known that first of all, the DRAM technology is required to deal with leakage in cells. For designs of 1Gb or beyond, the leakage induced yield degradation within the specified retention time should be less than -5 in the $-erf^{-1}(1-2\text{F})$ format, which corresponds to less than 1ppb failure or 6σ yield probability.

Based on the conclusions, the sense amplifier and core structure related yield performance becomes less dependent on cell leakage, and thus can be discussed individually with a Gaussian approximated linear procedure. As shown in the next chapter, by a linear statistical model the yield of a DRAM core array in consideration of array structures, array parasitic, array coupling and sense amplifier mismatch can be estimated and optimized for future technologies.

Chapter 6

DRAM Core Yield Analysis and Optimization

6.1 Introduction

The compact DRAM core circuits have to face a number of variations and parasitic artifacts, being much more sensitive than the peripheral circuits with their relaxed structure size. Many publications [26, 20] are looking for new array structures and sensing methods that can give larger signal amplitude V_{sign} during pre-sensing and reduce the bitline to bitline coupling during post-sensing. The worst case analysis methodologies are usually used to analyze the robustness of the circuits. However, the outcome of the worst case outcome can never be directly applied to estimate or optimize the yield of the DRAM core design unless the time consuming Monte-Carlo (MC) simulations are performed. With very small failure probability (e.g. $< 10^{-12}$ or 1ppb) and so many replica devices, MC simulation is usually not appropriate means. Empirical methods [68] are more practical and favorable for analysis and estimation of the DRAM core yield. The disadvantage is that it can only be carried out on fabricated chips, and therefore contributes little to either discovering the sources of the failures, or guiding the design of novel DRAM technology, array structure and sense amplifiers in advance. Besides, the design-fabricate-test iteration obviously slows down the yield ramp-up speed for a new DRAM design.

In Chapter 5, the signal margin analysis is introduced. Based on the signal margin analysis, a linear statistical model will be developed here to analyze and optimize a DRAM core design. Different from earlier publications [26, 20], based on the accurate voltage difference V_{sign} for different array structures during pre-sensing obtained from a charge conservation model in Section 2.2 and the post-sensing coupling model in Section 2.3, the linear yield model can provide accurate and concrete results in the presence of parasitics. Besides that, the

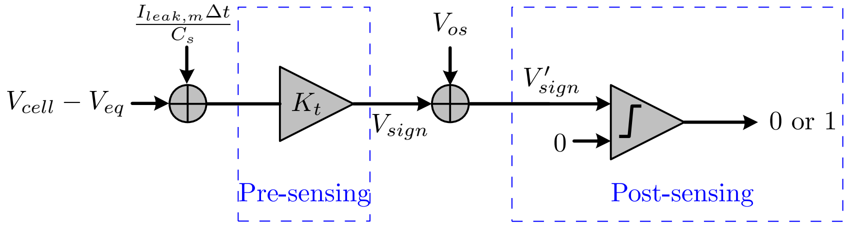


Fig. 6.1: A linearized sensing process

speed of the analytical yield model is superior to other simulation based design methodologies. In contrast with other DRAM yield analysis methods, this analytical signal margin analysis yield model exhibits a straightforward view on the core yield dependency of all kinds of effects including cell capacitance variation, sense amplifier mismatch, bitline array structure, array core voltage. DRAM core designs can thus be optimized systematically in consideration of yield related trade-offs.

6.2 Statistical Linear Sensing Process

Based on the voltage sensing process mentioned in Chapter 2, a simplified linear sensing procedure is drawn in **Fig. 6.1**. As depicted in the firgure, first of all the cell is subjected to leakage. The pre-sensing process scales the cell effective signal amplitude $|V_{cell} - V_{eq}|$ by a factor of K_t down to the bitline voltage difference V_{sign}. The sense amplifier mismatch equivalent input offset V_{os} is then added to V_{sign}, which is eventually fed to an ideal sense amplifier that behaves like a comparator according to Section 4.3. Since there are only two output states for a comparator, the sensing ends up with a '0' or '1'. The corresponding sensing process is illustrated in **Fig. 6.2**. Sensing failure appears when the bitline voltage difference changes its polarity during the sensing process as implied in (b).

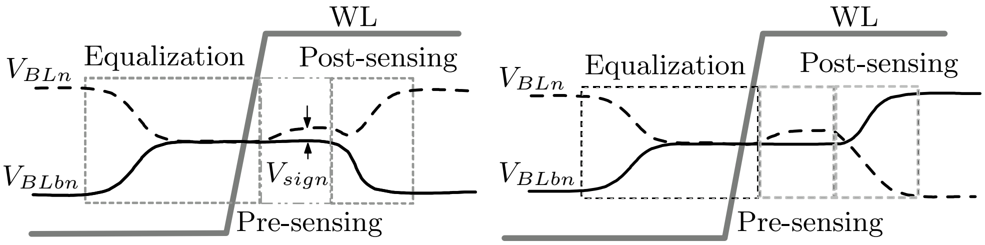


Fig. 6.2: Sensing process consisting of pre-sensing and post sensing procedures: (a) a successful sensing process (b) an erroneous sensing process.

Obviously, since the transfer ratio K_t depends on the array capacitance ratio and the amplitudes of V_{sign} for different sensing pairs are identical when the entire array is under signal margin analysis, the final yield of an array is determined by the distribution of V_{os} and leakage as discussed in Chapter 5 in this model. Furthermore, in Section 5.3 it has been stated that in a signal margin plot there will be two different parts - the leakage determined tail and the array determined body part; The Gaussian shape is preserved in the body part in normal operation with less than 64ms retention time; The tail part is found to be independent of the array structure, parameters or sense amplifier design.

Apparently, the above conclusions are valid only with reasonable statistical distribution of V_{os}, or the yield of the array will be completely determined by the main distribution and the leakage induced failure degradation can never be observed. As a consequence, in this chapter the main target is to analyze the main distribution from the linear model in **Fig. 6.1** with zero cell leakage, so that the main Gaussian distribution can be optimized with small enough failure probability, for example, 10ppb for 1Gb DRAM design. However, as a matter of fact the model in **Fig. 6.1** is too simple to be used to analyze and estimate the yield of DRAM core arrays comprising side effects such as the capacitance variations and the post-sensing cross-talk couplings.

As described in Section 2.1, the overall sensing process is divided into pre-sensing and post-sensing. In the post-sensing the bitline to bitline cross-talk coupling can seriously influence the yield of array especially in signal margin analysis with worst pattern. Since in signal margin analysis all bitline pairs are expected to develope an identical V_{sign} in pre-sensing, the coupling model in Section 2.3 fits well into the situation: When V_{os} is taken into consideration, the effective voltage difference V'_{sign} in **Fig. 6.1** becomes randomized and the majority of weak pairs is supposed to be between two pairs with nominal voltage difference obtained from Eqn. (2.28) in Section 2.2.

As modeled in Section 2.3, suppose a pair with initial voltage difference v_a is surrounded by other two sensing pairs with initial voltage difference v_b, according to Eqn. (2.42) the center pair's equivalent initial voltage changes from v_a to $v_a - v_b \cdot K_{cpl}$ due to the post-sensing coupling, where K_{cpl} in Eqn. (2.43) is the post sensing coupling coefficient. Thus the post-sensing coupling can be linearized with gain and offset. Therefore, when leakage is removed and the post-sensing coupling is comprised into **Fig. 6.1**, the linear sensing process results in a model as shown in **Fig. 6.3**, where V_{cpl} and K_p are offset and gain caused by the post-sensing coupling.

Obviously, when variabilities are Gaussian approximated the yield of an array from the linear process can be expressed by the error function as discussed in Section 4.1 and Section 5.3. **Fig. 6.4** demonstrates the evolution of the distribution from the initial cell voltage $V_{cell} - V_{eq}$ to the failure probability for 0 cells during signal margin analysis:

With the assumption that the capacitance variations have independent Gaus-

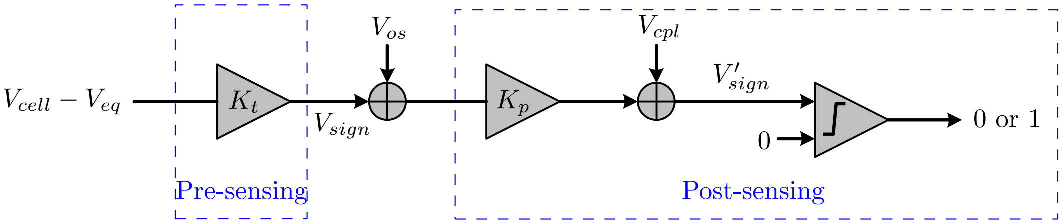


Fig. 6.3: The modified linear model including gain K_p and offset V_{cpl} caused by post-sensing coupling in SMA

sian distributions (Suppose capacitor leakage does not exist), the developed voltage difference V_{sign} will follow a Gaussian distribution with expectation μ and standard deviation σ_1 for a large number of bitline pairs. μ is array structure dependent as shown in Eqn. (2.28). From Eqn. (2.28) and Eqn. (4.9), the variance of V_{sign} is approximated by:

$$\sigma_1^2 = [\sum_i (\frac{\partial K_t}{\partial C_i})^2 \sigma_{Ci}^2](V_{cell} - V_{eq})^2 \tag{6.1}$$

where C_i represents array capacitances C_s, C_{bl} and C_{bl2bl}. As an example, **Fig. 6.5** shows a MC simulated V_{sign}. Here C_s and C_{bl} variations are included and supposed to follow Gaussian distributions with 30fF, 70fF mean and 1.5fF, 3.5fF standard deviation, respectively. By calculation, the standard deviation $\sigma_1 = 0.089$mV, which is identical to MC simulations. In fact, the variation of $V_{cell} - V_{eq}$ can also be included in σ_1. However, as V_{cell} is supposed to be rather stable, this fluctuation is neglected here.

As discussed in Section 4.3, the mismatch inside sense amplifiers can be replaced with input offset V_{os} and statistically this offset voltage follows a Gaussian distribution with zero mean and variance σ_{vos}^2 [81]. The addition of V_{os} to the developed signal V_{sign} results in a new Gaussian distribution (μ, σ_2) as illustrated

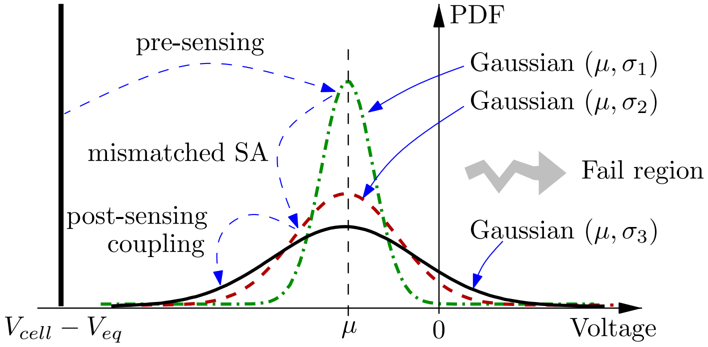


Fig. 6.4: Evaluation of the variance and probability of failures corresponding to V_{cell} in signal margin analysis

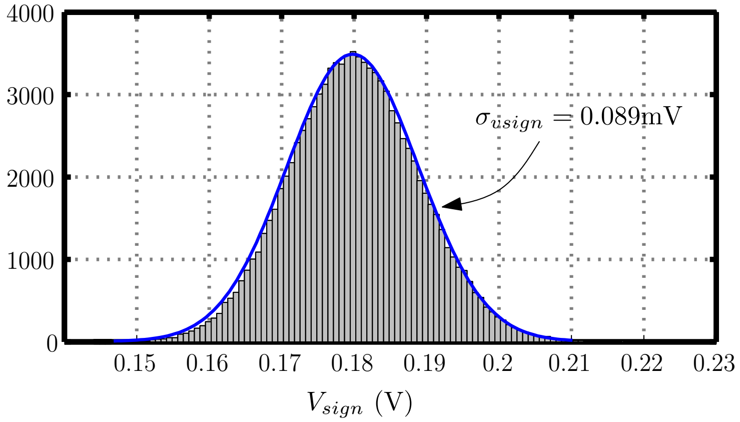


Fig. 6.5: MC simulated V_{sign} shows the same standard variation as theoretical calculation from Eqn. (6.1).

in **Fig. 6.4** with

$$\sigma_2^2 = \sigma_1^2 + \sigma_{vos}^2 \tag{6.2}$$

When ideal sense amplifiers are replaced by comparators with zero offset, the probability of failures becomes the area under each individual Gaussian curve on the right side of $x = 0$. As **Fig. 6.4** implies, the failure probability rises significantly after introduction of V_{os}.

Finally, the effect of post sensing coupling has to be considered as the coupling will further enlarge the variance of the effective voltage difference as implied by **Fig. 6.4**. Since the voltage V_{sign}+V_{os} in **Fig. 6.3** for each bitline pair in the worst pattern case is nominally equal, statistically each pair will have an independent Gaussian distribution G(μ, σ_2). Furthermore, according to Eqn. (2.42) due to the crosstalk coupling the effective input of each pair is a combination of the initial input of it and its neighbors, i.e., v'_a=v_a-$K_{cpl}v_b$. In the worst pattern case, the initial inputs of all the pairs follow the same characteristic, which gives μ_{va}=μ_{vb}=μ and σ_{va}=σ_{vb}=σ_2. Consequently, the effective signal V'_{sign} in **Fig. 6.3** including post-sensing coupling effect results in a variance

$$\begin{aligned}\sigma_{v'sign}^2 &= \sigma_{va}^2 + \sigma_{vb}^2 \cdot [K_{cpl}^2 + (v_b\frac{\partial K_{cpl}}{\partial v_b})^2] \\ &\approx (1 + K_{cpl}^2)\{\sigma_{vos}^2 + [\sum_i (\frac{\partial K_t}{\partial C_i})^2 \sigma_{Ci}^2](V_{cell} - V_{eq})^2\}\end{aligned} \tag{6.3}$$

and a mean

$$\mu_{v'sign} = \mu_{va} - \mu_{vb} \cdot K_{cpl} = \mu(1 - K_{cpl}) \tag{6.4}$$

As a result, the effect of post sensing coupling is equivalent to a process in which the Gaussian distribution G(μ,σ_2) of the input voltage difference V_{sign}+V_{os} is first scaled to the Gaussian distribution with mean $\mu_{v'sign}$ and variance $\sigma^2_{v'sign}$, and then pass through an ideal post-sensing process without crosstalk coupling. Since the failure probability will not change when both the mean and standard deviation of a Gaussian distribution are scaled, to simplify the calculation process, $\mu_{v'sign}$ is scaled down to μ, and the variance $\sigma^2_{v'sign}$ is transformed to σ_3^2 as in **Fig. 6.4** with

$$\sigma_3^2 = \frac{(1+K_{cpl}^2)}{(1-K_{cpl})^2}\{\sigma^2_{vos} + [\sum_i(\frac{\partial K_t}{\partial C_i})^2\sigma^2_{Ci}](V_{cell}-V_{eq})^2\} \tag{6.5}$$

Therefore, the post sensing coupling in signal margin analysis can be considered as a linear transformation that keeps the shape of the Gaussian distribution unchanged as implied by **Fig. 6.4**. The final failure probability for a give V_{cell} is obtained by using the error function as before:

$$\mathrm{F}(V_{cell}) = 1 - \mathrm{Y} = \frac{1}{2}[1 - erf(\frac{\mu}{\sqrt{2}\sigma_3})] \tag{6.6}$$

Conversely, the standard deviation corresponding to a given failure probability is

$$\sigma_3 = \frac{\mu}{\sqrt{2}erf^{-1}(1-2\mathrm{F})} \tag{6.7}$$

where μ is the nominal value of V_{sign}.

Fig. 6.6 compares σ_3 theoretically calculated from Eqn. (6.5) with the one transformed by Eqn. (6.7) from SPICE Monte-Carlo (MC) simulations to verify the post sensing coupling model for signal margin analysis. Obviously, the MC simulations agree with the conclusion that in signal margin analysis the post sensing coupling will preserve the Gaussian characterisitc originating from the input. Since the failure probability varies with C_{bl2bl} to C_{bl} ratio, the confidence region of a N sample MC simulation becomes wider with the increase of the failure probability as has been discussed in Section 1.5, and the simulation outcomes scatter gradually in the figure correspondingly.

Eqn. (6.5) also exhibits that σ_3 is dependent of the initial cell signal $|V_{cell} - V_{eq}|$: When V_{cell} approaches V_{eq}, σ_1 is close to zero and σ_3 approximates to a fraction of σ_{vos}, which is almost a constant if the post-sensing coupling is not severe; when V_{cell} is far from V_{eq}, σ_1 comes into play and σ_3 grows with $|V_{cell} - V_{eq}|$. However, some calculations show that the capacitor related variation σ_1 is always smaller than the sense amplifier mismatch σ_{vos} at $V_{cell} = 0$ for a DRAM core as shown in **Fig. 6.7** even with an aggressive capacitance variance, e.g., σ_{Cs}/C_s=9%. For typical DRAM designs with σ_{vos} = 10mV and V_{dd} = 1.2V, σ_1 only gives 5mV at V_{cell} = 0V with a capacitance standard deviation σ_{Cs}/C_s = 6%. The conclusion rules out the chance that the tail part of a measured signal margin analysis may come from capacitor variations and highlights the conclusion obtained in Chapter 5 that cell leakage is the major source for producing the tail part.

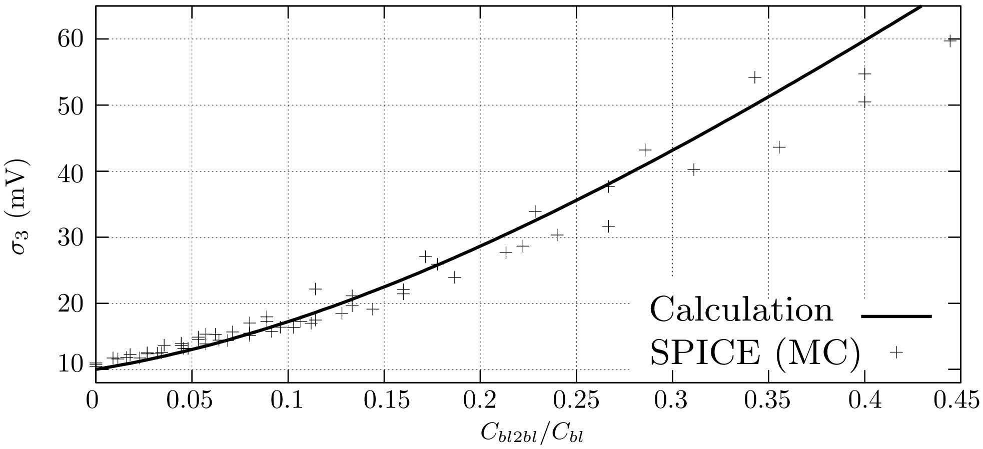


Fig. 6.6: σ_3 **vs. Capacitance ratio with** $\sigma_{vos} = 10$**mV (suppose capacitors are ideal without variations).**

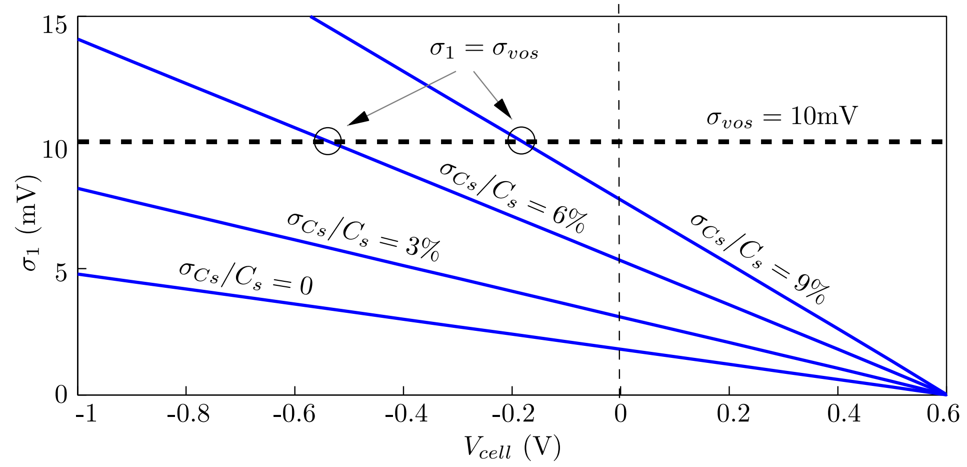


Fig. 6.7: V_{cell} **dependent** σ_1 **vs.** σ_{vos}

6.3 Yield Estimation vs. Measurements

To verify the validity of the linear Gaussian model, signal margin tests are carried out for different arrays of 75nm $8F^2$ products. To make it easier to compare the test outcomes with the analytical estimations, Eqn. (6.6) is transformed into

$$\frac{V_{sign}}{\sqrt{2}\sigma_3} = erf^{-1}(1 - 2\mathrm{F}) \qquad (6.8)$$

where the left side term can be calculated from Eqn. (2.28), Eqn. (6.5) and the right side term comes from signal margin tests. The parameters as mean and variance of the capacitances and the mismatch of transistors are obtained from other technology measurements.

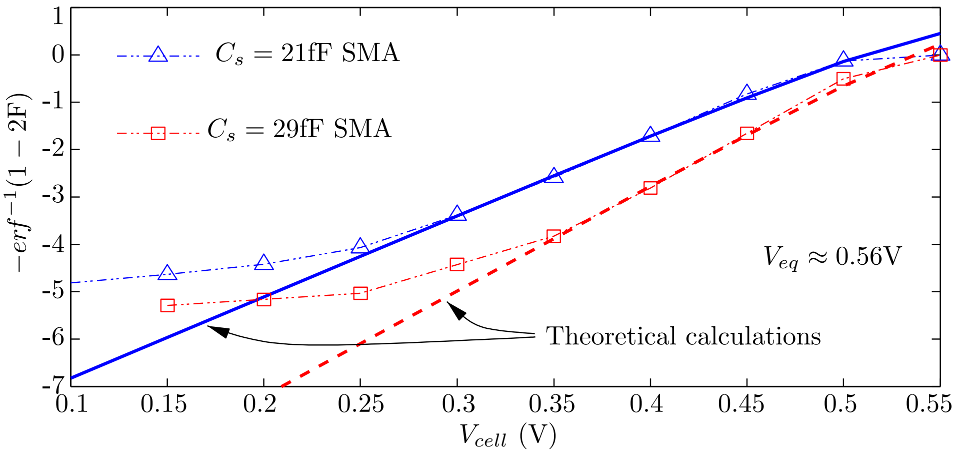


Fig. 6.8: Theoretical calculations vs. Measurements for designs in 75nm technology with different cell capacitor values.

6.3.1 Arrays with different cell capacitors

Fig. 6.8 compares the calculated values to the data obtained from a signal margin analysis for a multi-twisted folded array with multiplexers. The plot consists of two tests for two different designs with the same array and sense amplifiers but different C_s, which are 21fF and 29fF, respectively. Meanwhile, all capacitances are assumed to have 5% mismatch in calculations. Because the 29fF cell capacitor has a larger transfer ratio K_t, the angle of its slope is greater in the region near V_{eq}. Obviously, both curves fit the theoretical calculations from the linear signal margin model well when V_{cell} is close to V_{eq}. Divergence happens when $|V_{cell} - V_{eq}|$ is raised higher than 200mV. According to Chapter 5 and the previous analysis, the yield in these parts are determined by the cell leakage.

6.3.2 Long vs. Short bitline arrays

Fig. 6.9 gives comparisons between estimated and measured data for different array structures with the same cell capacitor, sense amplifiers and multiplexers. Long Bit Line (LBL) represents 512bits per BL, multi-twisted folded arrays and Short Bit Line (SBL) represents 256bits per BL, none-twist folded arrays. From calculation, K_t of SBL is more than 1.62 times larger than K_t of LBL. Like before, the calculations fit well within the critical region near V_{eq}. Although the failure drops with SBL, due to the increased total number of sense amplifiers the cost goes up. However, the area of sense amplifiers in SBL can shrink further down by pushing its curve to the position near the LBL curve. A detailed example illustrating how the yield, supply voltage and area trade-offs are optimized by the yield model will be given in the following Section 6.4.

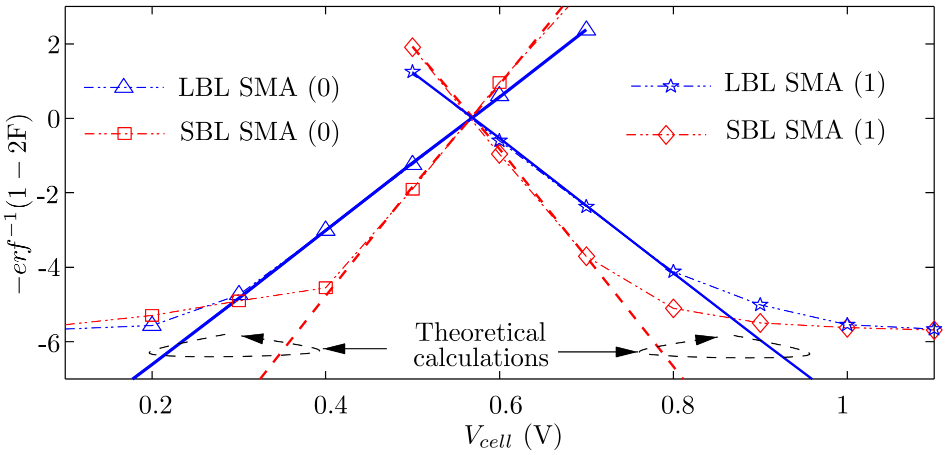


Fig. 6.9: Theoretical calculations vs. Measurements for long bitline array (512 cells/BL) and short bitline array (256 cells/BL) with the same cell capacitance and sense amplifiers in 75nm technology.

6.4 Yield Estimation for Future DRAM Core

6.4.1 $8F^2$, $6F^2$ and $4F^2$ arrays

The yield model in Section 6.2 can be applied to predicting the yields of any DRAM technology and array structure. **Fig. 6.10** demonstrates a plot comparing the yield performance of different arrays. In this comparison, the minimum wire pitch is supposed to be the same, and bitline of each technology has 512 bits with 1.2V core supply voltage. Here, $4F^2$ cells are supposed to have 1 unit bitline and wordline pitch but without any physical implementation. $8F^2$ folded bitline arrays with and without multiple twisted techniques are included. Shielded open bitline arrays are supposed to be used in both $6F^2$ and $4F^2$ cell arrays.

As shown in the figure, owing to smaller inter-bitline capacitance C_{bl2bl} in shielded open bitline arrays, large yield improvements can be observed for shielded open bitline arrays with $6F^2$ and $4F^2$. In addition, since the lengths of bitlines in open bitline arrays with $6F^2$ and $4F^2$ cells are scaled down by factors of 3/4 and 1/2 from folded bitline array with $8F^2$ cells, respectively, the open bitline arrays have smaller bitline capacitance and thus larger transfer ratio. However, without bitline shielding or the bitline to bitline coupling capacitance can not be removed completely, the open bitline arrays suffer from severer post-sensing coupling and their yields become relatively low.

6.4.2 Sense amplifier optimization

Fig. 6.11 illustrates the effect of another technology trend in DRAM core - re-

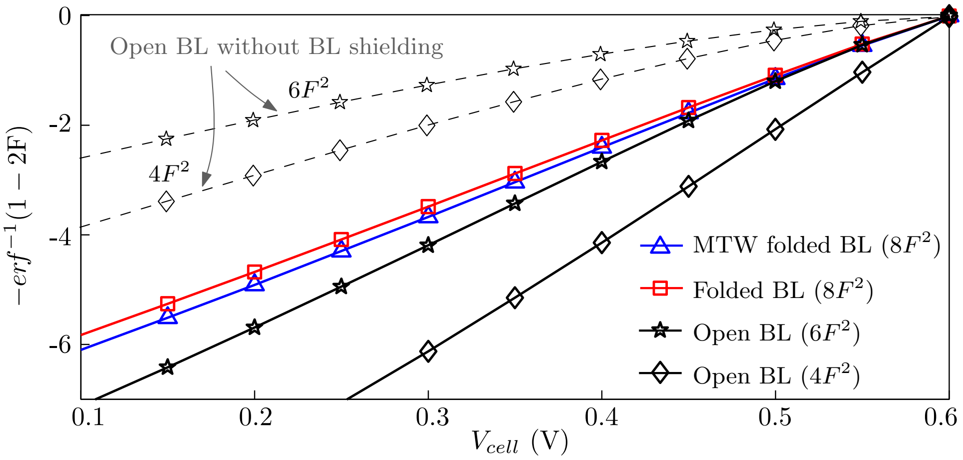


Fig. 6.10: Estimated V_{sign}/σ_3 for $8F^2$ cell multiple twisted folded, normal folded, $6F^2$ and $4F^2$ cell open bitline arrays with the same wire pitch, 512 bits per bitline and 1.2V core supply.

duction of core supply voltage. The calculations are done with the cell voltage V_{cell} fixed at 0.1V, which is reasonable in consideration of leakage and cell voltage fluctuations. Assume a qualified design should have V_{sign}/σ_3 of maximum -6. The model provides now the minimum core supply voltage V_{dd} requirements for different arrays with 10mV σ_{vos} of sense amplifiers. Due to the previously mentioned effects the arrays with $8F^2$ cells need a supply voltage of 1.2V while shielded open bitline arrays with $6F^2$ cells achieve the same goal with only 1V.

When the supply voltage further drops down to 0.9V, in order to meet the yield requirement of -6, the $6F^2$ cell arrays have to decrease the sense amplifiers mismatch σ_{vos} from 10mV to 8mV, resulting in an increase of the sensing transistors gate area by a factor of $(10/8)^2 \approx 1.56$ according to equation $\sigma_{V_{th}} = A/\sqrt{WL}$ [58, 60]. However, the use of open bitline arrays with $4F^2$ cells points out another way to reach this goal without increasing the area of the sense amplifiers. These demonstrations show the trade-offs between yield, supply voltage, area and cost in DRAM core design.

It is noticeable that the above examples are applicable only for on-die variations. When wafer to wafer or lot to lot variations are dealt with, the mean values of the used parameters can be assumed to vary in a range, e.g. $C_s = 25 \pm 2.5$fF, $\sigma_{vos} = 10 \pm 1$mV, etc. Thus, the signal margin analysis model built here can systematically optimize DRAM core design with enough safe margin.

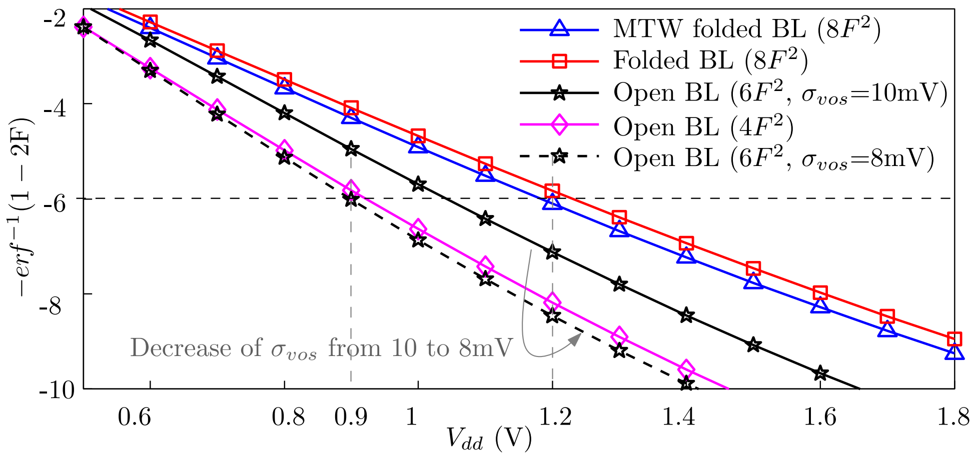


Fig. 6.11: Estimated V_{sign}/σ_3 vs. Core supply voltage with $V_{cell} = 0.1$V for different arrays.

6.5 Summary

In this chapter, by separating the leakage induced yield degradation from the overall yield, a hierarchical statistic signal margin analysis model is developed based on the linear sensing model in **Fig. 6.3** and the statistical characteristics of different blocks from the previous chapters: The cell effective voltage $|V_{cell}-V_{eq}|$ is scaled down by the array factor K_t during pre-sensing; Due to variations of array capacitances, the downscaled voltage V_{sign} follows a Gaussian distribution; The input offset voltage with a Gaussian distribution determined by the sense amplifier mismatch is added to V_{sign} before post-sensing; The bitline to bitline post-sensing coupling effects in signal margin analysis can be regarded as a gain applied to the initial input variance; and in the final the analytical yield expression is obtained for the body part that is determined by the array parameters, structure and mismatch of sense amplifiers.

It is also found that the variation in cell capacitor is found to have less impact on yield performance. It is the developed signal amplitude V_{sign} in pre-sensing phase, the mismatch of sense amplifiers and the post sensing coupling that determine the electric yield of DRAM core. The model estimated yields for different parameters and array structures are well in agreement with the measurements from different signal margin analysis, proving its accuracy and effectiveness. Finally, the impact of current DRAM technology trends on yield is evaluated by the model, and the good accuracy of the model can facilitate the optimization of core supply voltage and sense amplifier area, lowering the product costs. The analytical model provides a powerful tool for yield estimation of future DRAM technologies and novel array structures with lower cost, and a guide to design for yield of DRAM core circuits.

Chapter 7

Conclusions and Outlook

With the advances of semiconductor technology, the minimum feature size of on die devices becomes ever smaller, whereas the increase of single die size still continues. Yield and cost balance gradually comes to be the most important issue for all semiconductor companies. In order to address the increasingly important trade-off between yield and core circuit area, strong efforts have to be taken in technology optimizations, circuit simulations and device testing. Obviously, a more economic procedure would be beneficial.

In this thesis, conventional circuit design is combined with statistical probability analysis. It gives rise to a statistical circuit design methodology that can be applied to DRAM cores. In order to form the hierarchical model, the developed voltage amplitude V_{sign} for different arrays with parasitics is acquired by a charge conservation model first; Post-sensing coupling is analyzed with a circuit model, which fits in the signal margin analysis situation; the mismatch of sense amplifiers is referred back to input as V_{os} and its statistical characteristics are evaluated; The cell leakage induced yield degradation is also investigated and the outcomes discover that a) the cell leakage components are following Log-Norm distributions, b) the yield loss due to the cell leakage and core circuit design can be separated, c) however, both needs to meet the yield requirement; Based on the equations and approximations, a systematic analytical yield model is eventually produced. By using the model, the relationships between yield and other design parameters are revealed, so that different trade-offs in DRAM core can be balanced in an intuitive and explicit way.

Statistical design methodology is promising not only for DRAM cores but other large volume memories such as SRAM, Flash and Phase Change memory as well. Due to the thousands of millions of replicas of devices inside a single die for these chips, statistical design can provide the lest time-to-market cycle together with less cost. Although a new statistical model for every type of memory is required, the methods, equations and experiences obtained here are still beneficial.

Based on this dissertation, further improvements may be carried out in the future. The most significant one is the sensing timing issue that is not addressed statistically here. Actually, it is more important for SRAM and has been statistically analyzed as shown in [82, 83, 84]. Similarly, it can be included and modeled for a DRAM core, but some new statistical distributions are indeed required. Statistical analysis of power consumption [85, 86, 87] is also a topic in consideration of the increasingly stringent power budgets. Because of the complexity of the on die device variations, it is worth of more investigations.

Bibliography

[1] T. Sakurai and A. R. Newton, "Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas," *IEEE J. Solid-State Circuits*, vol. 25, pp. 584–594, April 1990.

[2] N. Arora, *MOSFET Models for VLSI Circuit Simulation.* Springer Verlag, 1993.

[3] A. I. A. Cunha, M. C. Schneider, and C. Galup-Montoro, "An MOS transistor model for analog circuit design," *IEEE J. Solid-State Circuits*, vol. 33, pp. 1510–1519, Oct. 1998.

[4] W. Sansen, M. Steyaert, V. Peluso, and E. Peeters, "Toward sub 1 v analog integrated circuits in submicron standard CMOS technologies," in *Proc. Digest of Technical Papers Solid-State Circuits Conference 45th ISSCC 1998 IEEE International*, pp. 186–187,435, 5–7 Feb. 1998.

[5] B. Prince and G. Due-Gundersen, *Semiconductor Memories - A Handbook of Design, Manufacture, and Application (Second Edition).* WILEY, 1991.

[6] K. Itoh, *VLSI Memory Chip Design.* Springer, 2001.

[7] K. Itoh, "Trends in megabit DRAM circuit design," *IEEE J. Solid-State Circuits*, vol. 25, no. 3, pp. 778–789, 1990.

[8] M. Aoki, Y. Nakagome, M. Horiguchi, H. Tanaka, S. Ikenaga, J. Etoh, Y. Kawamoto, S. Kimura, E. Takeda, H. Sunami, and K. Itoh, "A 60-ns 16-mbit CMOS DRAM with a transposed data-line structure," *IEEE J. Solid-State Circuits*, vol. 23, no. 5, pp. 1113–1119, 1988.

[9] D. Chin, C. Kim, Y. H. Choi, D. S. Min, H. S. Hwang, H. Choi, S. I. Cho, T. Y. Chung, C. J. Park, Y. S. Shin, K. Suh, and Y. E. Park, "An expermental 16Mb DRAM with reduced peak-current noise," in *Proc. Symposium on VLSI Circuits, 1989. Digest of Technical Papers*, pp. 113–114, 2002.

[10] "JESD79-2 DDR2 SDRAM specification."

[11] C. M. Grinstead, *Introduction to probability (Second revised edition).* Dartmouth College, 1997.

[12] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods.* Springer; 2nd Edition, 2005.

[13] G. E. Moore, "Progress in digital integrated electronics," in *Proc. International Electron Devices Meeting*, vol. 21, pp. 11–13, June 1975.

[14] R. R. Schaller, "Moore's law: past, present and future," *IEEE Spectr.*, vol. 34, pp. 52–59, June 1997.

[15] C.-G. Hwang, "Semiconductor memories for IT era," in *Proc. Digest of Technical Papers Solid-State Circuits Conference ISSCC. 2002 IEEE International*, vol. 1, pp. 24–27, 3–7 Feb. 2002.

[16] J. A. Mandelman, R. H. Dennard, G. B. Bronner, J. K. DeBrosse, R. Divakaruni, Y. Li, and C. J. Radens, "Challenges and future directions for the scaling of dynamic random-access memory (DRAM)," *IBM J. RES. & DEV.*, vol. 46, pp. 187–212, 2002.

[17] L. Nesbit, J. Alsmeier, B. Chen, J. DeBrosse, P. Faheyk, M. Gall, J. Gambino, S. Gernhard, H. Ishiuchi, R. Kleinhenz, J. Mandelman, T. Mii, M. Morikado, A. Nitayama, S. Parke, H. Wong, and G. Bronner, "A 0.6 &256 Mb trench DRAM cell with self-aligned BuriEd STrap (BEST)," in *Proc. International Electron Devices Meeting Technical Digest*, pp. 627–630, 5–8 Dec. 1993.

[18] R. P. Vollertsen and W. W. Abadeer, "Comprehensive gate-oxide reliability evaluation for DRAM processes," in *Proc. 7th European Symposium on Reliability of Electron Devices, Failure Physics and Analysis*, pp. 1631–1638, 8–11 October 1996.

[19] T. Yoshihara, H. Hidaka, Y. Matsuda, and K. Fujishima, "A twisted bit line technique for Multi-Mb DRAMs," in *Proc. IEEE International Solid-State Circuits Conference Digest of Technical Papers. 36th ISSCC*, pp. 238–239, 1988.

[20] Y. Nakagome, M. Aoki, S. Ikenaga, M. Horiguchi, S. Kimura, Y. Kawamoto, and K. Itoh, "The impact of data-line interference noise on DRAM scaling," *IEEE J. Solid-State Circuits*, vol. 23, no. 5, pp. 1120–1127, 1988.

[21] W. H. Henkels, "Dynamic capacitance effects in DRAM word lines," *IEE Proceedings I Communications, Speech and Vision*, vol. 135, no. 1, pp. 1–6, 1988.

[22] H. Masuda, R. Hori, Y. Kamigaki, K. Itoh, H. Kawamoto, and H. Katto, "A 5V-only 64K dynamic RAM based on high S/N design," *IEEE J. Solid-State Circuits*, vol. 15, pp. 846–854, Oct 1980.

[23] J. S. Yuan and J. J. Liou, "Interconnect noise analysis for megabit DRAMs," in *Proc. Seventh International IEEE VLSI Multilevel Interconnection Conference*, pp. 205–211, 1990.

[24] H. Hidaka, K. Fujishima, Y. Matsuda, M. Asakura, and T. Yoshihara, "Twisted bit-line architectures for multi-megabit DRAMs," *IEEE J. Solid-State Circuits*, vol. 24, no. 1, pp. 21–27, 1989.

[25] S. Watanabe, K. Tsuchida, D. Takashima, Y. Oowaki, A. Nitayama, K. Hieda, H. Takato, K. Sunouchi, F. Horiguchi, K. Ohuchi, F. Masuoka, and H. Hara, "A novel circuit technology with surrounding gate transistors (SGT's) for ultra high density DRAM's," *IEEE J. Solid-State Circuits*, vol. 30, no. 9, pp. 960–971, 1995.

[26] D.-S. Min and D. W. Langer, "Multiple twisted dataline techniques for multigigabit DRAMs," *IEEE J. Solid-State Circuits*, vol. 34, no. 6, pp. 856–865, 1999.

[27] D.-S. Min, D. W. Langer, and G.-H. Kim, "Multiple twisted data line technique for scaled DRAMs," *Electronics Letters*, vol. 34, no. 13, pp. 1296–1297, 1998.

[28] D.-S. Min and D. W. Langer, "Multiple twisted data line techniques for coupling noise reduction in embedded DRAMs," in *Proc. Custom Integrated Circuits the IEEE 1999*, pp. 231–234, 1999.

[29] Z. Al-Ars, M. Herzog, I. Schanstra, and A. J. van de Goor, "Influence of bit line twisting on the faulty behavior of DRAMs," in *Proc. Records of the 2004 International Workshop on Memory Technology, Design and Testing*, pp. 32–37, 2004.

[30] Z. Al-Ars, S. Hamdioui, A. J. van de Goor, and S. Al-Harbi, “Influence of bit-line coupling and twisting on the faulty behavior of DRAMs,” *IEEE Trans. Computer-Aided Design*, vol. 25, no. 12, pp. 2989–2996, 2006.

[31] D. Takashima, S. Watanabe, H. Nakano, Y. Oowaki, and K. Ohuchi, “Open/folded bit-line arrangement for ultra-high-density DRAM's,” *IEEE J. Solid-State Circuits*, vol. 29, pp. 539–542, April 1994.

[32] J.-S. Kim, Y.-S. Choi, H.-J. Yoo, and K.-S. Seo, “A low noise folded bit-line sensing architecture for multi-Gb DRAM with ultra high density $6F^2$ cell,” in *Proc. 23rd European Solid-State Circuits Conference ESSCIRC '97*, pp. 192–195, 1997.

[33] H. Yoon, J. Y. Sim, H. S. Lee, K. N. Lim, J. Y. Lee, N. J. Kim, K. Y. Kim, S. M. Byun, W. S. Yang, C. H. Choi, H. S. Jeong, J. H. Yoo, D. I. Seo, K. Kim, B. I. Ryu, and C. G. Hwang, “A 4 Gb DDR SDRAM with gain-controlled pre-sensing and reference bitline calibration schemes in the twisted open bitline architecture,” in *Proc. Digest of Technical Papers Solid-State Circuits Conference ISSCC. 2001 IEEE International*, pp. 378–379, 467, 2001.

[34] K. Mashiko, T. Kobayashi, W. Wakamiya, M. Hatanaka, and M. Yamada, “A 70ns 256k DRAM with bitline shielding structure,” in *Proc. Solid-State Circuits Conference. Digest of Technical Papers. 1984 IEEE International*, vol. XXVII, pp. 98–99, 1984.

[35] M. Koyanagi, H. Sunami, N. Hashimoto, and M. Ashikawa, “Novel high density, stacked capacitor MOS RAM,” in *Proc. International Electron Devices Meeting*, vol. 24, pp. 348–351, 1978.

[36] H. Watanabe, K. Kurosawa, and S. Sawada, “Stacked capacitor cells for high-density dynamic RAMs,” in *Proc. International Electron Devices Meeting IEDM '88. Technical Digest*, pp. 600–603, 1988.

[37] S. Kimura, Y. Kawamoto, T. Kure, N. Hasegawa, J. Etoh, M. Aoki, E. Takeda, H. Sunami, and K. Itoh, “A new stacked capacitor DRAM cell characterized by a storage capacitor on a bit-line structure,” in *Proc. International Electron Devices Meeting IEDM '88. Technical Digest*, pp. 596–599, 1988.

[38] K. U. Stein, A. Sihling, and E. Doering, “Storage array and sense/refresh circuit for single-transistor memory cells,” *IEEE J. Solid-State Circuits*, vol. 7, no. 5, pp. 336–340, 1972.

[39] D. Johns and K. Martin, *Analog Integrated Circuit Design.* New York: Wiley, 2000.

[40] C. N. Ahlquist, J. R. Breivogel, J. T. Koo, J. L. McCollum, W. G. Oldham, and A. L. Renninger, “A 16 384-bit dynamic ram,” *IEEE J. Solid-State Circuits*, vol. 11, no. 5, pp. 570–574, 1976.

[41] M. Kondo, T. Mano, F. Yanagawa, H. Kikuchi, T. Amazawa, K. Kiuchi, N. Ieda, and H. Yoshimura, “A high speed molybdenum gate MOS RAM,” *IEEE J. Solid-State Circuits*, vol. 13, no. 5, pp. 611–616, 1978.

[42] T. Wada, M. Takada, S. Matsue, M. KaMOShida, and S. Suzuki, “A 150 ns, 150 mw, 64k dynamic MOS RAM,” *IEEE J. Solid-State Circuits*, vol. 13, no. 5, pp. 607–611, 1978.

[43] N. C. C. Lu and H. H. Chao, “Half-Vdd bit-line sensing scheme in CMOS DRAMs,” *IEEE J. Solid-State Circuits*, vol. 19, no. 4, pp. 451–454, 1984.

[44] S. Eto, M. Matsumiya, M. Takita, Y. Ishii, T. Nakamura, K. Kawabata, H. Kano, A. Kitamoto, T. Ikeda, T. Koga, M. Higashiho, Y. Serizawa, K. Irabashi, O. Tsuboi, Y. Yokoyama, and M. Taguchi, “A 1-gb SDRAM with ground-level precharged bit line and nonboosted 2.1-v word line,” *IEEE J. Solid-State Circuits*, vol. 33, no. 11, pp. 1697–1702, 1998.

[45] S. H. Dhang, N. C. C. Lu, W. Hwang, and S. A. Parke, "High-speed sensing scheme for CMOS DRAMs," *IEEE J. Solid-State Circuits*, vol. 23, no. 1, pp. 34–40, 1988.

[46] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices.* Cambridge University Press, 1998.

[47] S. Sze, *Physics of semiconductor devices.* Wiley, 1984.

[48] M. Aoki, J. Etoh, K. Itoh, S. Kimura, and Y. Kawamoto, "A 1.5-v DRAM for battery-based applications," *IEEE J. Solid-State Circuits*, vol. 24, pp. 1206–1212, Oct 1989.

[49] T. Yamagata, S. Tomishima, M. Tsukude, T. Tsuruda, Y. Hashizume, and K. Arimoto, "Low voltage circuit design techniques for battery-operated and/or giga-scale DRAMs," *IEEE J. Solid-State Circuits*, vol. 30, no. 11, pp. 1183–1188, 1995.

[50] T. Ooishi, M. Asakura, S. Tomishima, H. Hidaka, K. Arimoto, and K. Fujishima, "A well-synchronized sensing/equalizing method for sub-1.0-V operating advanced DRAMs," *IEEE J. Solid-State Circuits*, vol. 29, pp. 432–440, April 1994.

[51] T.-H. Kim, H. Eom, J. Keane, and C. Kim, "Utilizing reverse short channel effect for optimal subthreshold circuit design," in *Proc. International Symposium on ISLPED'06 Low Power Electronics and Design*, pp. 127–130, 2006.

[52] L. G. Heller, D. P. Spampinato, and Y. L. Yao, "High sensitivity charge-transfer sense amplifier," *IEEE J. Solid-State Circuits*, vol. 11, no. 5, pp. 596–601, 1976.

[53] B. Razavi, *Design of Analog CMOS Integrated Circuits.* McGraw-Hill, 2001.

[54] M. Tsukude, S. Kuge, T. Fujino, and K. Arimoto, "A 1.2- to 3.3-V wide voltage-range/low-power DRAM with a charge-transfer presensing scheme," *IEEE J. Solid-State Circuits*, vol. 32, no. 11, pp. 1721–1727, 1997.

[55] S. Chou, T. Takano, A. Kita, F. Ichikawa, and M. Uesugi, "A 60-ns 16-Mbit DRAM with a minimized sensing delay caused by bit-line stray capacitance," *IEEE J. Solid-State Circuits*, vol. 24, pp. 1176–1183, Oct 1989.

[56] J.-S. Kim, H.-J. Yoo, and K.-S. Seo, "Boosted charge transfer preamplifier for low power gbit-scale DRAM," *Electronics Letters*, vol. 34, no. 18, pp. 1785–1787, 1998.

[57] H.-C. Chow and C.-L. Hsieh, "A 0.5V high speed DRAM charge transfer sense amplifier," in *Proc. 50th Midwest Symposium on Circuits and Systems MWSCAS 2007*, pp. 1293–1296, 2007.

[58] M. J. M. Pelgrom, A. C. J. Duinmaijer, and A. P. G. Welbers, "Matching properties of MOS transistors," *IEEE J. Solid-State Circuits*, vol. 24, no. 5, pp. 1433–1439, 1989.

[59] M. J. M. Pelgrom, H. P. Tuinhout, and M. Vertregt, "Transistor matching in analog CMOS applications," in *Proc. International Electron Devices Meeting IEDM '98 Technical Digest*, pp. 915–918, 6–9 Dec. 1998.

[60] S. R. Nassif, "Modeling and analysis of manufacturing variations," in *Proc. IEEE Conference on. Custom Integrated Circuits*, pp. 223–228, 6–9 May 2001.

[61] S. Suzuki and M. Hirata, "Threshold difference compensated sense amplifier," *IEEE J. Solid-State Circuits*, vol. 14, no. 6, pp. 1066–1070, 1979.

[62] T. Kawahara, T. Sakata, K. Itoh, Y. Kawajiri, T. Akiba, G. Kitsukawa, and M. Aoki, "A high-speed, small-area, threshold-voltage-mismatch compensation sense amplifier for gigabit-scale DRAM arrays," *IEEE J. Solid-State Circuits*, vol. 28, pp. 816–823, July 1993.

[63] Y. Watanabe, N. Nakamura, and S. Watanabe, “Offset compensating bit-line sensing scheme for high density DRAM’s,” *IEEE J. Solid-State Circuits*, vol. 29, pp. 9–13, Jan. 1994.

[64] J.-W. Sub, K.-M. Rho, C.-K. Park, and Y.-H. Koh, “Offset-trimming bit-line sensing scheme for gigabit-scale DRAM’s,” *IEEE J. Solid-State Circuits*, vol. 31, pp. 1025–1028, July 1996.

[65] S. Nassif, K. Bernstein, D. J. Frank, A. Gattiker, W. Haensch, B. L. Ji, E. Nowak, D. Pearson, and N. J. Rohrer, “High performance CMOS variability in the 65nm regime and beyond,” in *Proc. IEEE International Electron Devices Meeting IEDM 2007*, pp. 569–571, 10–12 Dec. 2007.

[66] P. A. Stolk, F. P. Widdershoven, and D. B. M. Klaassen, “Modeling statistical dopant fluctuations in MOS transistors,” *IEEE Trans. Electron Devices*, vol. 45, no. 9, pp. 1960–1971, 1998.

[67] H. E. Graeb, *Analog Design Centering and Sizing.* Springer Netherlands, 2007.

[68] S. Akiyama, T. Sekiguchi, K. Kajigaya, S. Hanzawa, R. Takemura, and T. Kawahara, “Concordant memory design: an integrated statistical design approach for multi-gigabit DRAM,” *IEEE J. Solid-State Circuits*, vol. 41, no. 1, pp. 107–112, 2006.

[69] H. Tijms, *Understanding Probability: Chance Rules in Everyday Life.* Cambridge University Press, 2004.

[70] J. W. Lindeberg, “Eine neue herleitung des exponentialgesetzes in der wahrscheinlichkeitsrechnung,” *Mathematische Zeitschrift*, vol. 15, pp. 211–225, 1922.

[71] P. R. Gray, P. J. Hurst, S. H. Lewis, and R. G. Meyer, *Analysis and Design of Analog Integrated Circuits (4th Edition).* Wiley, 2001.

[72] R. Sarpeshkar, J. Wyatt, J. L., N. C. Lu, and P. D. Gerber, “Mismatch sensitivity of a simultaneously latched CMOS sense amplifier,” *IEEE J. Solid-State Circuits*, vol. 26, no. 10, pp. 1413–1422, 1991.

[73] R. Kraus and K. Hoffmann, “Optimized sensing scheme of DRAMs,” *IEEE J. Solid-State Circuits*, vol. 24, no. 4, pp. 895–899, 1989.

[74] W. A. M. Van Noije, W. T. Liu, and J. Navarro, S. J., “Precise final state determination of mismatched CMOS latches,” *IEEE J. Solid-State Circuits*, vol. 30, no. 5, pp. 607–611, 1995.

[75] S. Jin, J.-H. Yi, J. H. Choi, D. G. Kang, Y. J. Park, and H. S. Min, “Prediction of data retention time distribution of DRAM by physics-based statistical simulation,” *IEEE Trans. Electron Devices*, vol. 52, pp. 2422–2429, Nov. 2005.

[76] S. Jin, J.-H. Yi, Y. J. Park, H. S. Min, J. H. Choi, and D. G. Kang, “Modeling of retention time distribution of DRAM cell using a Monte-Carlo method,” in *Proc. IEDM Technical Digest Electron Devices Meeting IEEE International*, pp. 399–402, 13–15 Dec. 2004.

[77] A. Hiraiwa, M. Ogasawara, N. Natsuaki, Y. Itoh, and H. Iwai, “Statistical modeling of dynamic random access memory data retention characteristics,” *Journal of Applied Physics*, vol. 80(5), p. 3091, September 1996.

[78] T. Hamamoto, S. Sugiura, and S. Sawada, “On the retention time distribution of dynamic random access memory (DRAM),” *IEEE Trans. Electron Devices*, vol. 45, pp. 1300–1309, June 1998.

[79] R. B. Leipnik, “On lognormal random variables: the characteristic function,” *Journal of the Australian Mathematical Society Series B*, vol. 32, pp. 327–347, 1991.

[80] E. Nelson, Y. Li, D. Poindexter, M. Ruprecht, E. Lim, Y. Matsubara, H. Sawazaki, Q. Ye, M. Iwatake, and W. Tonti, “Signal margin test to identify process sensitivities relevant to DRAM reliability and functionality at low temperatures,” in *Proc. IEEE International Integrated Reliability Workshop Final Report*, pp. 6–9, 18–21 Oct. 1999.

[81] Y. Li, H. Schneider, F. Schnabel, R. Thewes, and D. Schmitt-Landsiedel, “Latched CMOS DRAM sense amplifier yield analysis and optimization,” in *Integrated Circuit and System Design (Power and Timing Modeling, Optimization and Simulation)* (L. Svensson and J. Monteiro, eds.), no. ISBN 978-3-540-95947-2, Springer Verlag Berlin Heidelberg, 2009.

[82] R. M. Houle, “Simple statistical analysis techniques to determine optimum sense amp set times,” *IEEE J. Solid-State Circuits*, vol. 43, pp. 1816–1825, Aug. 2008.

[83] M. H. Abu-Rahma, K. Chowdhury, J. Wang, Z. Chen, S. S. Yoon, and M. Anis, “A methodology for statistical estimation of read access yield in srams,” in *Proc. 45th ACM/IEEE Design Automation Conference DAC 2008*, pp. 205–210, 8–13 June 2008.

[84] R. B. Brawhear, N. Menezes, C. Oh, L. T. Pillage, and M. R. Mercer, “Predicting circuit performance using circuit-level statistical timing analysis,” in *Proc. European Design and Test Conference EDAC, The European Conference on Design Automation. ETC European Test Conference. EUROASIC, The European Event in ASIC Design*, pp. 332–337, 28 Feb.–3 March 1994.

[85] Y.-W. Kim, J.-S. Kim, J.-W. Kim, and B.-S. Kong, “CMOS differential logic family with conditional operation for low-power application,” vol. 55, no. 5, pp. 437–441, 2008.

[86] B. Arts, L. Benini, N. van der Eng, M. Heijligers, A. Kenter, E. Macii, H. Munk, and F. Theeuwen, “Enhancing behavioural-level design flows with statistical power estimation capabilities,” *IEE Proceedings -Computers and Digital Techniques*, vol. 152, no. 6, pp. 731–737, 2005.

[87] Y. H. Park and E. S. Park, “Statistical power estimation of CMOS logic circuits with variable errors,” *Electronics Letters*, vol. 34, no. 11, pp. 1054–1056, 1998.