TK
7871.99
M44
P38
2008

WENDT

FRET 40
FRONTIERS IN ELECTRONIC TESTING

Andrei Pavlov
Manoj Sachdev

CMOS SRAM

Circuit Design and Parametric Test in Nano-Scaled Technologies

Process-Aware SRAM Design and Test

Springer

Andrei Pavlov · Manoj Sachdev

CMOS SRAM Circuit Design and Parametric Test in Nano-Scaled Technologies

Process-Aware SRAM Design and Test

Springer

TSMC Exhibit 1046, Page 2 of 21

Andrei Pavlov
Intel Corporation
2501 NW 229th Street
Hillsboro, OR 97124

Manoj Sachdev
University of Waterloo
Dept. Electrical & Computer Engineering
200 University Ave.
Waterloo ON N2L 3G1
Canada

Series Editor

Vishwani Agrawal
Department of Electrical and Computer Engineering
Auburn University
Auburn, AL 36849
USA

General Library System
University of Wisconsin - Madison
728 State Street
Madison, WI 53706-1494
U.S.A.

ISBN 978-1-4020-8362-4 e-ISBN 978-1-4020-8363-1

Library of Congress Control Number: 2008924192

© 2008 Springer Science + Business Media B.V.
No part of this work may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, microfilming, recording or otherwise, without written permission from the Publisher, with the exception of any material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work.

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Contents

1 Introduction and Motivation 1
1.1 Motivation 1
1.2 SRAM in the Computer Memory Hierarchy 3
1.3 Technology Scaling and SRAM Design and Test 4
1.3.1 Moore's Law 5
1.3.2 Obstacles in SRAM Scaling 6
1.4 SRAM Test Economics 7
1.5 SRAM Design and Test Tradeoffs 8
1.5.1 Area and Stability 9
1.5.2 Quality and Yield 9
1.5.3 Test Coverage and Test Time 10
1.6 Redundancy 11

2 SRAM Circuit Design and Operation 13
2.1 Introduction 13
2.2 SRAM Block Structure 13
2.3 SRAM Cell Design 15
2.3.1 Four-Transistor (4T) SRAM Cell with Polysilicon Resistor Load 15
2.3.2 Six-Transistor (6T) CMOS SRAM Cell 16
2.3.2.1 Read Operation 17
2.3.2.2 Write Operation 19
2.3.3 Four-Transistor (4T) Loadless SRAM Cell 21
2.4 Cell Layout Considerations 22
2.5 Sense Amplifier and Bit Line Precharge-Equalization 26
2.6 Write Driver 30
2.7 Row Address Decoder and Column MUX 31
2.8 Address Transition Detector 33
2.9 Timing Control Schemes 34
2.9.1 Delay-Line Based Timing Control 35

TSMC Exhibit 1046, Page 4 of 21

2.9.2 Replica-Loop Based Timing Control 35
2.9.3 Pipelined Timing Control 38
2.10 Summary 38

3 SRAM Cell Stability: Definition, Modeling and Testing 39
3.1 Introduction 39
3.2 Static Noise Margin of SRAM Cells 40
3.3 SNM Definitions 41
3.3.1 Inverter V_{IL}, V_{IH}, V_{OL} and V_{OH} 41
3.3.2 Noise Margins NM_H and NM_L with V_{OL} and V_{OH} Defined as Stable Logic Points 42
3.3.3 Noise Margins NM_H and NM_L with V_{OL} and V_{OH} Defined as -1 Slope Points 42
3.3.4 SNM as a Side of the Maximum Square Drawn Between the Inverter Characteristics 44
3.4 Analytical Expressions for SNM Calculation 46
3.4.1 Analytical SNM Expression for a 4T SRAM Cell with Polysilicon Resistor Load 47
3.4.2 Analytical SNM Expression for a 6T SRAM Cell 49
3.4.3 Conclusions from SNM Analytical Expressions 51
3.4.4 Analytical SNM Expression for a Loadless 4T SRAM Cell 51
3.4.4.1 Alpha-Power Law Model 51
3.4.4.2 Analytical SNM Expression Derivation 53
3.4.4.3 Finding V_{OH} and V_{IL} 53
3.4.4.4 Finding V_{OL} and V_{IH} 54
3.4.4.5 SNM Expression for 4T Loadless SRAM Cell 55
3.4.4.6 Simulation Results vs. the Analytical Expression 56
3.5 SRAM Cell Stability Sensitivity Factors 58
3.5.1 SRAM SNM and Process Parameter Variations 59
3.5.2 SRAM SNM and Non-catastrophic Defects 62
3.5.2.1 SNM vs. Non-catastrophic Breaks and Bridges 62
3.5.3 SRAM SNM and Operating Voltages Variation 64
3.6 SRAM Cell Stability Fault Model 68
3.7 SRAM Cell Stability Detection Concept 69
3.8 March Tests and Stability Fault Detection in SRAMs 72
3.8.1 March 11N 72
3.8.2 Hammer Test 73
3.8.3 Coupling Fault Detection 74
3.9 Summary 77

4 Traditional SRAM Fault Models and Test Practices 79
4.1 Introduction 79
4.2 Traditional Fault Models 80
4.3 Traditional SRAM Test Practices 85

4.3.1 March Tests 87
4.3.1.1 March C 91
4.3.1.2 March C- 91
4.3.1.3 Hammer Test 91
4.3.2 Parametric Testing 92
4.3.2.1 Burn-In Test (BI) 93
4.3.2.2 I_{DDQ} Test 97
4.3.2.3 Limitations of I_{DDQ} Testing in Scaled Technologies 98
4.3.3 Design For Test Techniques 99
4.3.3.1 Built-In Self Test 100
4.4 Summary 101

5 Techniques for Detection of SRAM Cells with Stability Faults 103
5.1 Introduction 103
5.2 Classification of SRAM Cell Stability Test Techniques 103
5.3 Passive SRAM Cell Stability Test Techniques 104
5.3.1 Data Retention Test 104
5.3.2 Low-Voltage Test 108
5.4 Active SRAM Cell Stability Test Techniques 111
5.4.1 Active Techniques with a Single Test Stress Level 112
5.4.1.1 Soft Defect Detection 112
5.4.1.2 Weak Write Test Mode 114
5.4.1.3 Integrated Weak Write Test Mode 118
5.4.1.4 Word Line Driver Underdrive During a Write Operation 119
5.4.1.5 Soft Defect Detection-II (SDD-II) Technique 120
5.4.1.6 No Write Recovery Test Mode 123
5.4.1.7 Word Line Voltage Overdrive During a Read Operation 125
5.4.2 Active Techniques with Programmable Test Stress Levels 126
5.4.2.1 Short Write Test Mode 127
5.4.2.2 Programmable Weak Write Test Mode 130
5.4.2.3 Programmable Integrated Weak Write Test Mode 133
5.4.2.4 Programmable Read Current Ratio Technique with a Pass Transistor Technique 135
5.4.2.5 Programmable Read Current Ratio Technique with Floating Bit Lines 141
5.4.2.6 Programmable Word Line Pulsing Technique 147
5.5 Summary 156

6 Soft Errors in SRAMs: Sources, Mechanisms and Mitigation Techniques 159
6.1 Introduction 159
6.2 Soft Error Mechanism 161
6.3 Sources of Soft Errors 163
6.3.1 Alpha Particles 163
6.3.2 Neutron-Induced ^{10}B Fission 164
6.3.3 High-Energy Cosmic Rays 165
6.4 Soft Errors and Defects in the Pull-Up Path of a Cell 166
6.5 Soft Error Mitigation Techniques 167
6.5.1 Error Detection and Correction 167
6.5.1.1 Parity Check 168
6.5.1.2 Error Correction Codes 168
6.5.1.3 Circuit and/or System Redundancy 174
6.5.2 Removing Radiation Sources or Reducing Their Intensity 175
6.5.3 Increasing the Capacitance of the Storage Nodes 176
6.5.4 Inserting Resistors in the Feedback Loop of an SRAM Cell 177
6.6 Leakage-Reduction Techniques and the SER 179
6.7 Summary 180

References 183

Index 191

TSMC Exhibit 1046, Page 7 of 21

Chapter 2
SRAM Circuit Design and Operation

2.1 Introduction

A significantly large segment of modern SoCs is occupied by SRAMs. For instance, SRAM-based caches occupy more than 90% of 1.72 billion transistors in the Montecito processor [19]. Similarly, SRAM content in ASIC domain is also increasing. Therefore, understanding SRAM design and operation is crucial for enhancing various aspects of chip design and manufacturing.

In subsequent sections we will discuss the salient design and operational issues of SRAMs in general and the SRAM cell in particular. SRAM cell design considerations are important for a number of reasons. Firstly, the design of an SRAM cell is key to ensure stable and robust SRAM operation. Secondly, owing to continuous drive to enhance the on-chip storage capacity, the SRAM designers are motivated to increase the packing density. Therefore, an SRAM cell must be as small as possible while meeting the stability, speed, power and yield constraints. Near minimum-size cell transistors exhibit higher susceptibility with respect to process variations. Thirdly, the cell layout largely determines the SRAM critical area, which is the chip yield limiter. Meeting the design constraints requires deeper understanding of the involved trade-offs. In scaled technologies the cell stability is of paramount significance. Static Noise Margin (SNM) of a cell is a measure of its stability. Sections 3.2, 3.3 and 3.4 present an in-depth discussion on SNM and analytical approaches for its computation.

2.2 SRAM Block Structure

Figure 2.1 shows an example of the basic SRAM block structure. A row decoder gated by appropriate timing block signal decodes X row address bits and selects one of the word lines WL_0–WL_N-1. The SRAM core consists of a number of arrays of NxM, where N is the number of rows and M is the number of bits. If an SRAM

TSMC Exhibit 1046, Page 8 of 21

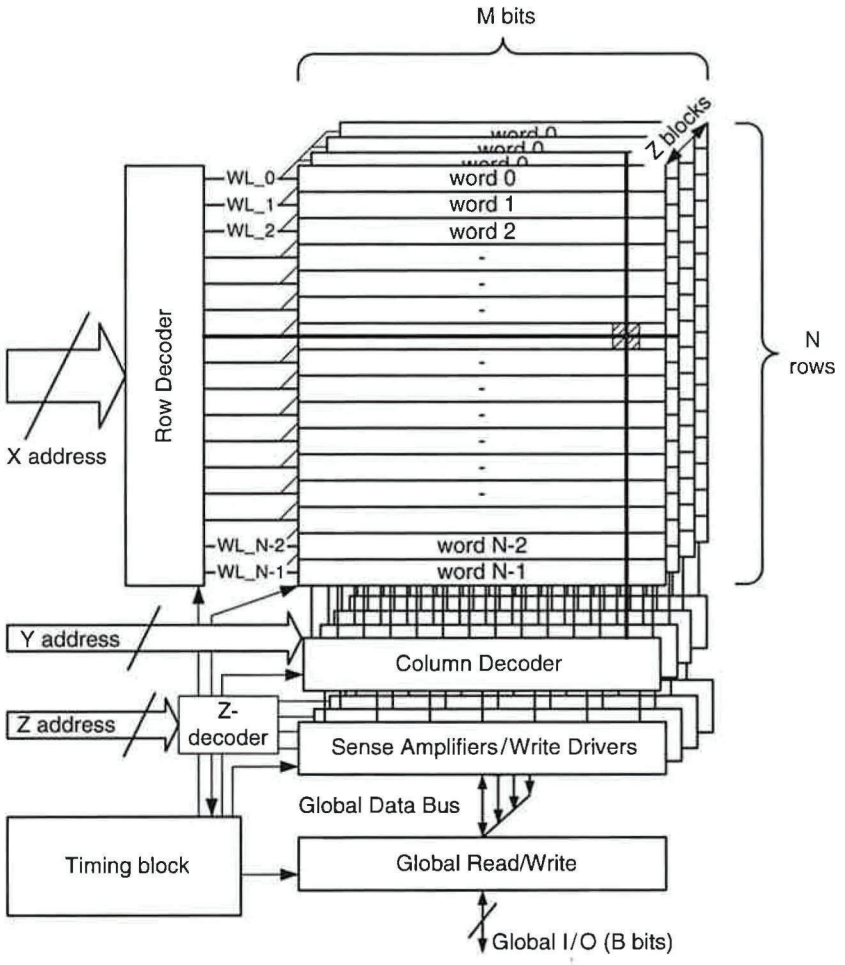


Fig. 2.1 SRAM block diagram

core is organized as a number of arrays in a page manner, an additional Z-decoder is needed to select the accessed page. Figure 2.1 shows an example of an SRAM with four pages of NxM arrays with the corresponding I/O blocks.

SRAMs can be organized as bit-oriented or word-oriented. In a bit-oriented SRAM, each address accesses a single bit, whereas in a word-oriented memory, each address addresses a word of n bits (where the popular values of n include 8, 16, 32 or 64). Column decoders or column MUXs (YMUXs) addressed by Y address bits allow sharing of a single sense amplifier among 2, 4 or more columns. The majority of modern SRAMs are self-timed, i.e. all the internal timing is generated by the timing block within an SRAM instance. An additional Chip Select (CS) signal, introducing an extra decoding hierarchy level, is often provided in multi-SRAM chip architectures.

The main SRAM building blocks will be described in the following sections.

TSMC Exhibit 1046, Page 9 of 21

2.3 SRAM Cell Design

An SRAM cell is the key SRAM component storing binary information. A typical SRAM cell uses two cross-coupled inverters forming a latch and access transistors. Access transistors enable access to the cell during read and write operations and provide cell isolation during the not-accessed state. An SRAM cell is designed to provide non-destructive read access, write capability and data storage (or data retention) for as long as cell is powered.

We will discuss design and analysis aspects of three different SRAM cells: a resistive load four-transistor (4T) SRAM cell, a six-transistor (6T) CMOS SRAM cell and a loadless 4T SRAM cell. We will describe their advantages and disadvantages.

In general, the cell design must strike a balance between cell area, robustness, speed, leakage and yield. Cell size minimization is one of the most important design objectives. A smaller cell allows the number of bits per unit area to be increased and thus, decreases cost per bit. Reduced cell area can indirectly improve the speed and power consumption due to the reduction of the associated cell capacitances. Smaller cells result in a smaller array area and hence smaller bit line and word line capacitances, which in turn helps to improve the access speed performance. Reducing the transistor dimensions is the most effective means to achieve a smaller cell area. However, the transistor dimensions cannot be reduced indefinitely without compromising the other parameters. For instance, smaller transistors can compromise the cell stability. Often, performance and stability objectives restrict arbitrary reduction in cell transistor sizes. Similarly, cell area can be traded off for special features such as an improved radiation hardening or multi-port cell access.

2.3.1 Four-Transistor (4T) SRAM Cell with Polysilicon Resistor Load

Historically, 4T polysilicon resistor load cells are the remnants of the pre-CMOS technologies. Ratioed inverters in the cell have lower gain in the transition region and produce inherently less steep Voltage Transfer Characteristics (VTCs), which reflects on the SNM values and the time necessary to recover from the metastable state [20].

The main advantage of static 4T cells with polysilicon resistor load (PRL) (Figure 2.2) is the approximately 30% smaller area as compared to 6T CMOS SRAM cells. Due to the higher electron mobility ($\mu_n/\mu_p = 1.5-3$), all transistors in a PRL cell are normally NMOS. The load resistors serve to compensate for the off-state leakage of the pull-down devices. On one hand, the values of R_L must be as high as possible to retain a reasonable noise margin NM_L, i.e., to limit the "0" level rise and reduce the static power consumption. On the other hand, a high R_L severely increases the low-to-high propagation delay for $V_{DD}/2$ precharge and it also increases the cell size. $V_{DD}/2$ precharge reduces the read time. However, precharging

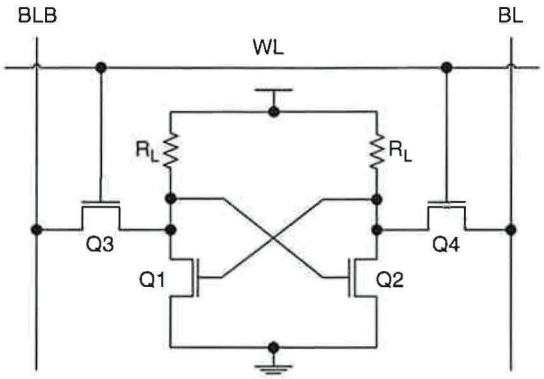


Fig. 2.2 Four-transistor (4T) SRAM cell with polysilicon resistor load

the bit lines to $V_{DD}/2$ can compromise the cell stability. Precharge of bit lines to full V_{DD} can alleviate the requirement for the low-to-high cell transition current at the cost of the additional precharge time and the associated power consumption. The upper resistance limit on R_L is set by the requirement to provide a pull-up current of at least two orders of magnitude larger than the leakage current of the pull-down transistor [21]. The lower limit on R_L is set by the required noise immunity and power consumption requirements. The technological variations of R_L caused by the limitations of doping and annealing techniques pose another constraint on the increase of R_L.

As the technology scaled into sub-micron regime (beyond 0.8 μm technology generation), the scalability of a PRL SRAM cell became an issue [22]. The polysilicon resistor in the PRL cell could not be scaled as aggressively as the cell's transistors. The switch from constant-voltage scaling to constant-field scaling to combat the short-channel effects and electric field implications on the long-term reliability, revealed non-satisfactory low-voltage power consumption vs. stability performance tradeoff of the PRL cells. Moreover, the extra technological steps of forming high-resistivity polysilicon are not a part of the standard CMOS logic technological process. Insufficient tolerance to soft errors, which can be linked to an inadequate SNM, adds to the list of disadvantages of a PRL cell. These factors prohibit using the PRL SRAM cells in Systems-on-a-Chip (SoCs) that are traditionally implemented using standard full CMOS processes. All the mentioned factors practically excluded the PRL cell from being used the current mainstream scaled-down deep-submicron technologies.

2.3.2 Six-Transistor (6T) CMOS SRAM Cell

The mainstream six-transistor (6T) CMOS SRAM cell is shown in Figure 2.3. Similarly to one of the implementations of an SR latch, it consists of six transistors. Four transistors ($Q1 - Q4$) comprise cross-coupled CMOS inverters and two

TSMC Exhibit 1046, Page 11 of 21

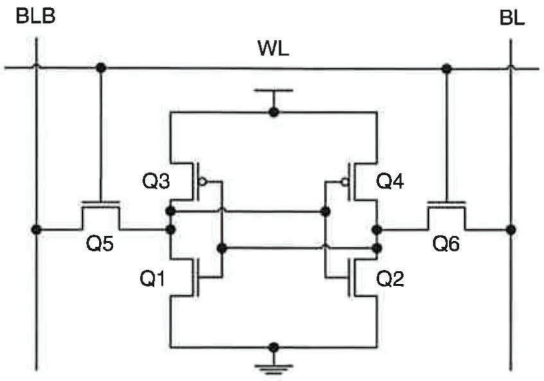


Fig. 2.3 Six-transistor (6T) CMOS SRAM cell

NMOS transistors $Q5$ and $Q6$ provide read and write access to the cell. Upon the activation of the word line, the access transistors connect the two internal nodes of the cell to the true (BL) and the complementary (BLB) bit lines.

A 6T CMOS SRAM cell is the most popular SRAM cell due to its superior robustness, low power and low-voltage operation. Therefore, we will discuss its operation and design in greater detail. An SRAM cell must be designed such that it provides a non-destructive read operation and a reliable write operation. These two requirements impose contradicting requirements on SRAM cell transistor sizing. SRAM cell transistor ratios that must be observed for successful read and write operations are discussed in the following sections.

2.3.2.1 Read Operation

Prior to initiating a read operation, the bit lines are precharged to V_{DD}. The read operation is initiated by enabling the word line (WL) and connecting the precharged bit lines, BL and BLB, to the internal nodes of the cell.

Upon read access shown in Figure 2.4, the bit line voltage V_{BL} remains at the precharge level. The complementary bit line voltage V_{BLB} is discharged through transistors $Q1$ and $Q5$ connected in series. Effectively, transistors $Q1$ and $Q5$ form a voltage divider whose output is now no longer at zero volt and is connected to the input of inverter $Q2-Q4$ (Figure 2.3). Sizing of $Q1$ and $Q5$ should ensure that inverter $Q2-Q4$ does not switch causing a destructive read. In other words, $0+\Delta V$ should be less than the switching threshold of inverter $Q2-Q4$ plus some *safety margin* or *Noise Margin*.

Ignoring the short-channel and body effects, the maximum allowed value $0+\Delta V$ of the node storing a logic "0" during read access can be expressed as [21]:

$$\Delta V = \frac{V_{DSATn}+CR(V_{DD}-V_{THn})-\sqrt{V_{DSATn}^2(1+CR)+CR^2(V_{DD}-V_{THn})^2}}{CR} \tag{2.1}$$

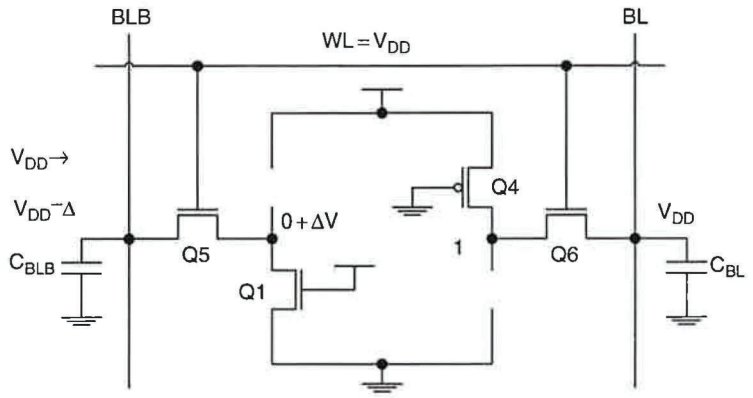


Fig. 2.4 Simplified model of a 6T CMOS SRAM cell during a read operation

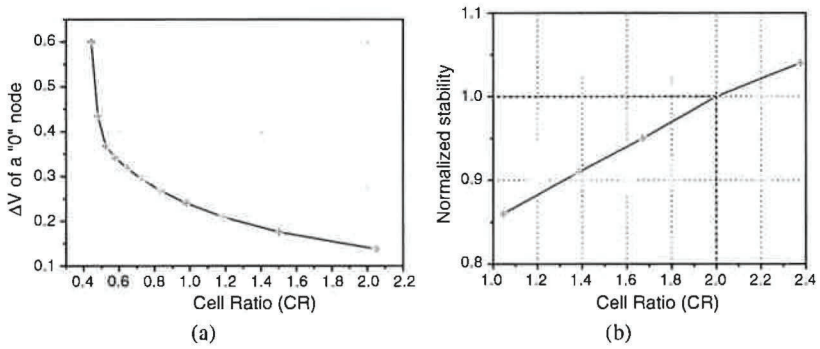


Fig. 2.5 The rise ΔV of the "0" node (a) and the SNM (b) as a function of the Cell Ratio (CR) $\left(CR = \frac{W_1}{L_1} / \frac{W_5}{L_5} = \frac{W_2}{L_2} / \frac{W_6}{L_6}$ in Figure 2.3$\right)$ in a 6T CMOS SRAM cell (simulated in CMOS 0.13 µm technology, V_{DD} = 1.2 V)

where CR (β) is the cell ratio defined as

$$\beta = CR = \frac{W_1/L_1}{W_5/L_5} \tag{2.2}$$

Since the cell is fully symmetrical, the CR is the same for $Q2$ and $Q6$.

The dependence of ΔV on the CR is shown in Figure 2.5(a). Typically, in order to ensure a non-destructive read and an adequate noise margin, CR must be greater than one and can be varied depending on the target application of the cell from approximately 1 to 2.5. Larger CRs provide higher read current I_{read} (and hence the speed) and improved stability (Figure 2.5(b)) at the expense of larger cell area. A smaller CRs ensure a more compact cell with moderate speed and stability. Leakage through the access transistors should be minimized to ensure robust read operation and to reduce the leakage power.

TSMC Exhibit 1046, Page 13 of 21

A preferred sizing solution can be to use a minimum-width access transistors with a slightly larger than the minimal length channel and a larger than minimal width with a minimal length driver transistors.

Once the complementary bit line discharges to a certain $V_{BLB} = V_{DD} - \Delta V$ voltage level sufficient for reliable sensing by the sense amplifier, the sense amplifier is enabled and amplifies the small differential voltage between the bit lines to the full-swing CMOS level output signal.

2.3.2.2 Write Operation

The write operation is similar to a reset operation of an SR latch. One of the bit lines, BL in Figure 2.6, is driven from precharged value (V_{DD}) to the ground potential by a write driver through transistor $Q6$. If transistors $Q4$ and $Q6$ are properly sized, then the cell is flipped and its data is effectively overwritten. A statistical measure of SRAM cell writeability is defined as *write margin*. Write margin is defined as the minimum bit line voltage required to flip the state of an SRAM cell [23]. The write margin value and variation is a function of the cell design, SRAM array size and process variation. A cell is considered not writeable if the worst-case write margin becomes lower than the ground potential.

Note that the write operation is applied to the node storing a "1". This is necessitated by the non-destructive read constraint that ensures that a "0" node does not exceed the switching threshold of inverter $Q2 - Q4$. The function of the pull-up transistors is only to maintain the high level on the "1" storage node and prevent its discharge by the off-state leakage current of the driver transistor during data retention and to provide the low-to-high transition during overwriting.

Assuming that the switching will not start before "1" node is below V_{TH_Q1}, a simplified overwrite condition can be expressed as [21]:

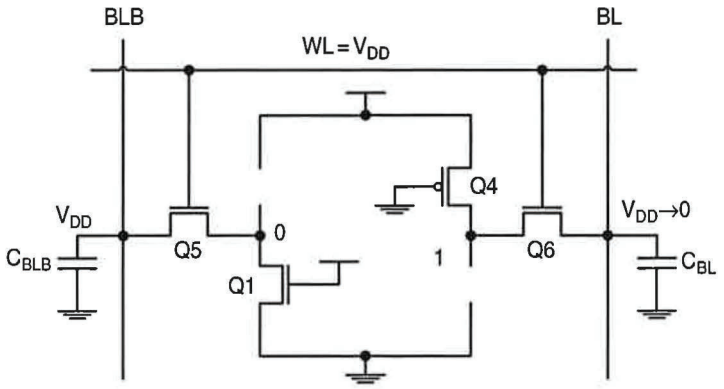


Fig. 2.6 Simplified model of a 6T CMOS SRAM cell during a write operation

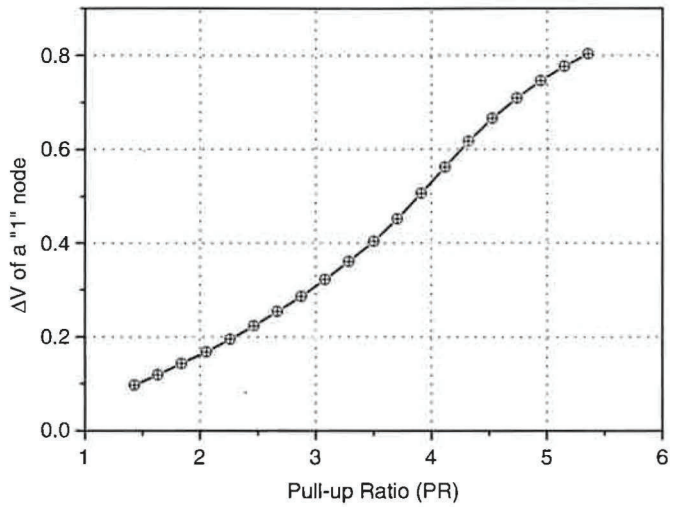


Fig. 2.7 The voltage drop at node $V_{\text{"1"}}$ during write access as a function of the Pull-Up ratio (PR) $\left(PR = \frac{W_4}{L_4}/\frac{W_6}{L_6} = \frac{W_5}{L_5}/\frac{W_3}{L_3}\right.$ in Figure 2.3$\left.\right)$ of a 6T CMOS SRAM cell (simulated in CMOS 0.13 µm technology, $V_{DD} = 1.2$ V)

$$V_{\text{"1"}} = V_{DD} - V_{THn} - \sqrt{(V_{DD} - V_{THn})^2 - 2\frac{\mu_p}{\mu_n} PR \left((V_{DD} - |V_{THp}|) V_{DSATp} - \frac{V_{DSATp}^2}{2} \right)} \tag{2.3}$$

where the pull-up ratio of the cell, PR, is defined as:

$$PR = \frac{W_4/L_4}{W_6/L_6} \tag{2.4}$$

The $V_{\text{"1"}}$ requirement is normally met using minimal-sized access and pull-up transistors only due to μ_n/μ_p ratio. Simulation results shown in Figure 2.7 demonstrate that for a normal write operation, i.e., to pull the $V_{\text{"1"}}$ node below V_{THn}, the W/L of the pull-up transistor must be less than 3-4 W/L of the access transistor. The exact maximum allowed PR is defined by the V_{THn} process option and by the switching threshold of inverter $Q1 - Q3$ in Figure 2.3.

Normally, to minimize the cell area and hence, increase the packing density, the sizes of the pull-up and access transistors are chosen to be minimal and approximately the same. However, stronger access transistors and/or weaker pull-up transistors may be needed to ensure a robust write operation under the worst process conditions e.g., in the fast PMOS and slow NMOS process skew corner. On the other hand, a relatively strong pull-up PMOS also benefits the read stability due to the increased P/N ratio of the back-to-back inverters ($Q3 - Q1$ and $Q4 - Q2$ in Figure 2.3) of the cell's latch. The read stability of an SRAM cell on one hand and

TSMC Exhibit 1046, Page 15 of 21

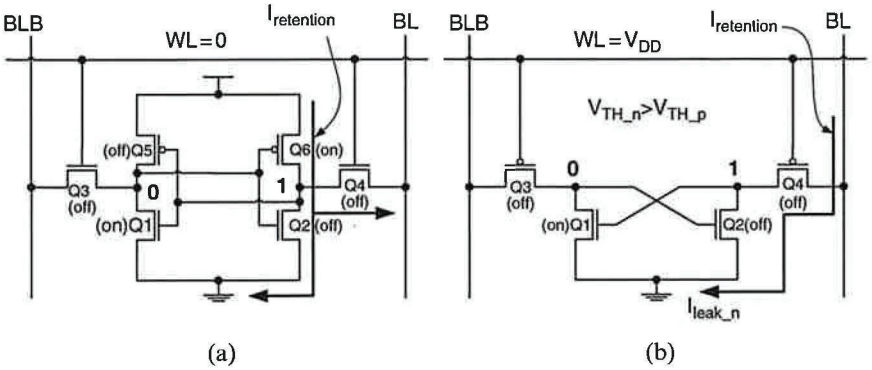
the writeability of the cell on the other hand are conflicting design requirements. It is getting increasingly more difficult to balance these requirements by conventional transistor sizing and V_{TH} optimization as the design window becomes increasingly narrower [23] with the technology scaling.

Despite the larger number of transistors compared to the other discussed cells, 6T CMOS SRAM cells offer superior stability and packing density provided the same performance and environmental tolerance.

2.3.3 Four-Transistor (4T) Loadless SRAM Cell

Recently, the new loadless 4T CMOS SRAM cell shown in Figure 2.8(b) was proposed by NEC [24, 25] for ultra-high density SRAM macros [26, 27]. A 4T SRAM cell uses minimal-size NMOS (Q1, Q2) and PMOS (Q3, Q4) transistors. Data retention without a refresh operation is ensured if the leakage current of the PMOS transistors is higher than the leakage current of the NMOS transistors. This condition is normally met by using dual-V_{TH} process with $V_{TH\_p} < V_{TH\_n}$. For stable data retention and to account for the I_{off} process distribution, the leakage current of the PMOS transistors in a 4T loadless cell must be around 100 times higher than that of the driver NMOS transistors. In a 4T SRAM cell, PMOS transistors Q3 and Q4 serve as access transistors as opposed to NMOS access transistors Q3 and Q4 in a 6T SRAM cell (Figure 2.8(a)). Due to the mobility ratio of the NMOS to the PMOS transistors μ_n/μ_p which is normally around two to three, all the transistors in a 4T SRAM cell can be of minimal size. Recall that for a 6T SRAM cell to guarantee a non-destructive read operation, NMOS driver transistors Q1 and Q2 must be 1.5–2.5 times larger than NMOS access transistors Q3 and Q4 in Figure 2.8. This fact, in addition to the larger number of transistors in a 6T SRAM cell, makes a 4T SRAM cell an area-efficient choice.

For the same access speed and comparable SNM, the area of a 4T SRAM cell is 50–65% of the area of a conventional 6T SRAM cell. Since memory blocks occupy



(a) (b)

Fig. 2.8 Six-transistor (a) and four-transistor (b) CMOS SRAM cells

considerable chip area in SoCs, SRAM cell area is a critical factor in the SoC design. The area savings offered by the 4T SRAM cells have been one of the main driving forces behind the 4T CMOS SRAM cell development.

However, 4T CMOS SRAM cells are not without drawbacks. Reliable data retention in a 4T cell can be guaranteed only if I_{leak_p} is significantly larger than I_{leak_n} under the worst case PVT conditions. This condition is increasingly more difficult to fulfill in nano-scaled processes with higher process parameter variations. The size advantage of 4T CMOS SRAM cells may be compromised by the stability degradation caused by the growing statistical spread of transistor parameters in nano-scaled technologies. Another disadvantage of 4T CMOS SRAM cells can reveal itself during a write operation. When writing to a certain column, the write driver drives the bit line from V_{DD} to $0V$. During this time, PMOS transistors in the other cells in the same column cannot provide the leakage current I_{PMOS_leak} to maintain the nodes storing "1". If the ratio of precharge time to the cycle time is not sufficient and the same memory address is being successively write accessed, there is a risk that other cells in the same column may flip.

2.4 Cell Layout Considerations

The layout of an SRAM cell defines the area density of the array and is key to manufacturing yield of the SoC containing large SRAM arrays. As the process technology continues to scale down, the limitations of lithography, etching and CMP call for modifications in the layout of SRAM cells. The general trend in the cell layout modifications is to streamline the patterns. By avoiding complex geometries, the reproducibility of the minimum feature sized patterns that define W and L of SRAM cell transistors can be greatly improved.

Four-transistor (4T) SRAM cells with a resistive load have dominated the stand-alone SRAMs since first introduced in the 1970s to around mid-1990s [28]. The stand-alone SRAM that used 4T cells often employed a special process, which allowed to shrink the cell size at the price of more process steps. However, the process for stand-alone SRAM often was not compatible with the logic process. The 6T SRAM cells, which are extensively used for on-die caches, are fabricated using the same process steps which are used for the logic, which warranted their use as embedded memory for microprocessors. The larger cell area due to two extra transistors was the main factor that initially prevented wide use of the 6T cell in the stand-alone applications.

A schematic and a sample layout of a 4T cell with a resistive load is shown in Figure 2.9(a). The cell uses NMOS driver ($Q1$ and $Q2$) and access ($Q3$ and $Q4$) transistors. The state of the "1" data node was preserved by a highly-resistive undoped polysilicon load resistors laid out on top of the cell (not shown in the layout). The pull-up current provided by the polysilicon resistors was chosen as a trade-off between the cell stability and the dissipated power per cell. If the resistance of the load is too large, the leakage current of the driver transistor can discharge the

TSMC Exhibit 1046, Page 17 of 21

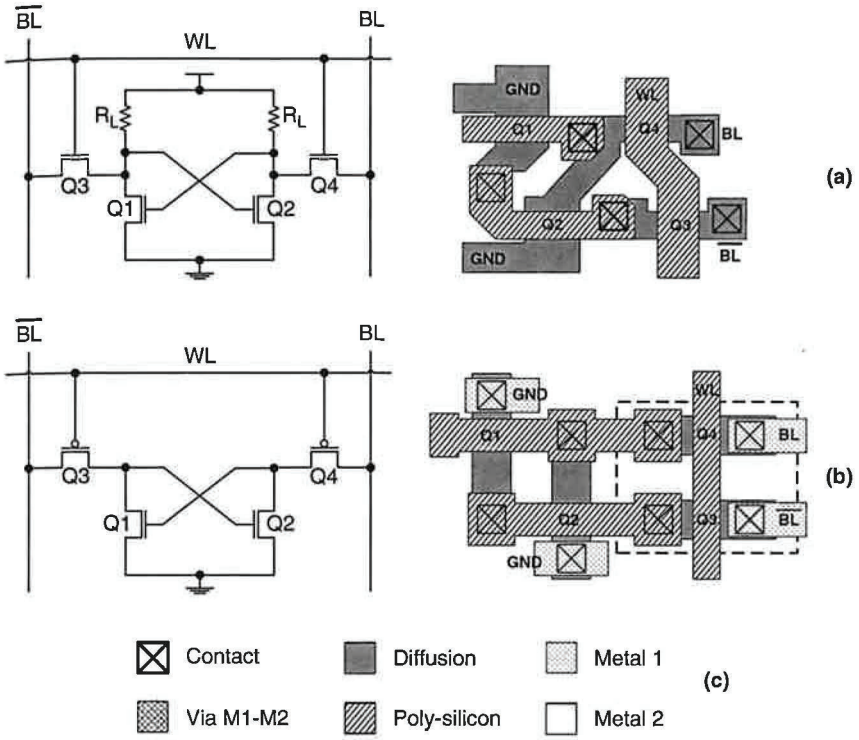


Fig. 2.9 Schematics and sample layouts of 4T SRAM cells: 4T polysilicon resistor load cell (a); 4T loadless cell (b); layout layers legend (c)

"1" data node. On the other hand, if the resistance is too small, the resulting voltage on the "1" data node caused by the voltage divider action of the load resistor and the driver transistor could partially turn the driver transistor of the other cell inverter on. Moreover, the power dissipation per cell in case of a small load resistor value multiplied by the number of cells in the array can become a significant part of the total power dissipation.

In an attempt to improve the insufficient stability margin in technology generations with the power supply voltage below 5 V, the 4T SRAM cells with polysilicon resistor load are replaced by the cells with TFT-transistor load. The "on" resistance of TFT transistors is lower than the "off" resistance, helping to reduce the cell size and the power and improve the noise immunity.

The schematic and a sample layout of another reincarnation of a 4T cell – a 4T loadless cell – is shown in Figure 2.9(b). The operation of this cell is described in more details in Section 2.3.3. The functions of the load and access transistors of this 4T cell are performed by two PMOS transistors ($Q3$ and $Q4$). SRAM designs using these cells have been reported for 180 nm [27, 29] and 130 nm [26] process technologies. Since the cell relies on the ratio of $\frac{I_{offPMOS}}{I_{offNMOS}}$, the data retention capabilities of this cell is strongly dependent on a strict process control.

By the mid-1990s and process technology generations with a feature size below 0.35 µm [28], the industry had mostly migrated from 4T cells with a resistive or TFT load to 6T CMOS cells for both the stand-alone and the on-chip storage. The constant electric field technology scaling drove the supply voltages below 5 V and became the main reason of the change in the cell type and layout. The 6T CMOS SRAM cells demonstrated the best stability in low-voltage operation while the advances in the process technology helped to improve the area efficiency of 6T cells.

A schematic and layout types of a 6T CMOS SRAM cells are shown in Figure 2.10. With a PMOS transistor as an active load, the stability margin and the power dissipation of an SRAM cell dramatically improved. Due to its form factor, the layout in Figure 2.10(a) is coined "tall". Tall layouts of SRAM cells were widely used down to the 90 nm technology generation. Due to high capital costs, the lithography equipment was updated at a slower pace than the progress of the transistor scaling, which followed the Moore's Law. Reproducible pattering of the design features that are much smaller than the wavelength of the lithography light source became problematic. The growing number of cells in the scaled SRAM arrays together with poor pattern reproducibility led to a higher number of marginal cells in the array which can reduce quality and yield and the efficiency of the power reduction techniques. Further scaling of a "tall" cell became more problematic due to poor reproducibility of the multiple pattern corners in the layout. Therefore, beyond the 90 nm technology generation node, the "tall" SRAM cell

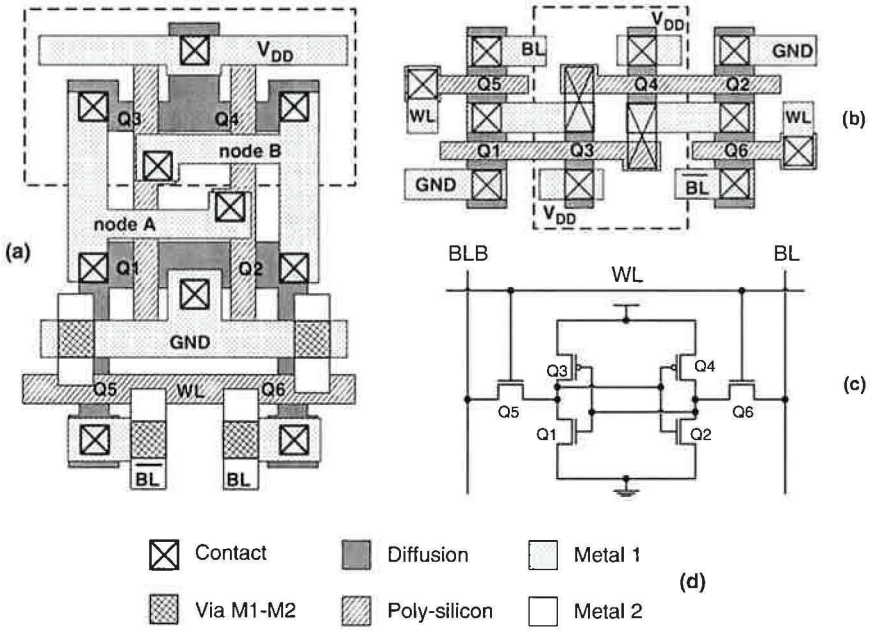


Fig. 2.10 Schematic and sample layouts of 6T SRAM cells: 6T CMOS "tall" cell (a); 6T CMOS "wide" cell (b); schematic of a 6T CMOS SRAM cell (c); layout layers legend (d)

TSMC Exhibit 1046, Page 19 of 21

layout shown in Figure 2.10(a) has been commonly replaced by the "wide" layout shown in Figure 2.10(b) [30–33]. The much streamlined layout of a "wide" cell and the uniform orientation of all the transistors in the cell provide better pattern reproducibility (W and L matching) and transistor V_{TH} matching. The impact and importance of good transistor matching in an SRAM cell is described in more detail in Chapter 3. Since the word lines in the "wide" bit layout are implemented in metal, the "wide" cell layout can offer reduced word line resistance and shorter propagation time of the WL signal to the parts of SRAM array which are further from the word line drivers. Another benefit of migrating to the "wide" SRAM cell is the shorter bit line length per cell. The shorter bit line length per cell transfers to the lower bit line capacitance C_{BL}. In turn, a lower C_{BL}/bit can be beneficial in one of the following ways. Firstly, it can be transferred to a faster developing of the bit-line differential signal and, thus, helps to achieve a faster cycle time of an array built of "wide" SRAM cells. Secondly, it increases the number of cells per bit line while keeping the same bit-line differential signal slew rate and, thus, helps to improve the array efficiency.

Diffusion, poly, contact and metal 1 are the four masking layers considered critical in the processing of SRAM cells. The tightened design rules applied to the SRAM array in these layers increase the risk of bridging/shorting. The fidelity of the patterns in these layers is, however, critical to cell stability, drive and leakage currents. To improve the pattern fidelity, the patterns in deep-submicron processes are corrected using Optical Proximity Correction (OPC). OPC of SRAM cells is especially challenging considering the tighter design rules. SRAM OPC often apply the accumulated knowledge of lithographic conditions and properties of the photoresist to be employed for each masking layer and a careful study and simulation of the resulting photoresist pattern on Si. The automated OPC algorithms though more generic, help to account for the process condition modifications [34]. Table 2.1 summarizes the issues associated with the critical layers in SRAM cell layout.

Table 2.1 Layout issues addressed during SRAM cell design

Layer	Issues addressed
Diffusion	(a) Critical Dimensions (CDs) (for I_{read}) (b) Bridging (c) Adequate area for contact (for low $R_{contact}$)
Poly	(a) Critical Dimensions (CDs) (for I_{read}) (b) Symmetry (for SNM) (c) Bridging (d) Adequate end-cap
Contact	(a) Adequate coupled contact coverage over diffusion and poly (b) Bridging with adjacent contacts (c) Adequate clearance from potential ILD voids
Metal 1	(a) Bridging (b) Adequate overlap of contact and via (for low $R_{contact}$)

TSMC Exhibit 1046, Page 20 of 21

Depending on the target application, SRAM cells may undergo layout modifications. While optimized for maximum bit density and minimum cell leakage, the cells for high density (HD) applications may have smaller area than cells for high speed (HS) applications. HS cells need larger driver transistors to provide larger cell currents and improved SNM.

2.5 Sense Amplifier and Bit Line Precharge-Equalization

Sense amplifiers (SA) are an important component in memory design. The choice and design of a SA defines the robustness of bit line sensing, impacting the read speed and power. Due to the variety of SAs in semiconductor memories and the impact they have on the final specs of the memory, the sense amplifiers have become a separate class of circuits.

The primary function of a SA in SRAMs is to amplify a small analog differential voltage developed on the bit lines by a read-accessed cell to the full swing digital output signal thus greatly reducing the time required for a read operation. Since SRAMs do not feature data refresh after sensing, the sensing operation must be nondestructive, as opposed to the destructive sensing of a DRAM cell. A SA allows the storage cells to be small, since each individual cell need not fully discharge the bit line.

Design constraints for an SA are defined by the minimum differential input signal amplitude, the minimum gain, A, and tolerance to the environmental conditions and mismatches. The gain is a function of the initial bit line voltage (precharge level). In addition, it influences the sense delay t_{sense}. However, a high A does not necessarily reduce t_{sense}. Usually, t_{sense} is traded off for reduced power consumption, layout area and for improved tolerance of environmental conditions [17].

Special attention is given to the SA area. Depending on the SRAM array column architecture, the area requirements for the SA may vary. Architectures using column multiplexing can share a single SA among the multiplexed columns such that only one column is connected to the sense amplifier at any given time. The total area available for the SA is defined by a multiple n of the bit line pitch values, where n can normally be from 1 to 16. In turn, the bit line pitch is defined by the size of a memory cell. This example illustrates the complexity of SRAM design and layout planning. The choice of the cell size, number of columns, number of cells per column, the minimum differential swing, the choice of the SA architecture and size are all the factors taken into consideration when designing an SRAM compliant with the target power, speed and reliability.

Generally, the parameters characterizing a sense amplifier include:

- Gain $A = V_{out}/V_{in}$
- Sensitivity $S = V_{in\_min}$ – minimum detectable signal
- Offsets V_{offset} and I_{offset} – the difference at the outputs with the common mode signal at the inputs

TSMC Exhibit 1046, Page 21 of 21