

CMOS Transistor Layout KungFu

Lee Eng Han
Valerio B. Perez
Mark Lambert Cayanes
Mary Grace Salaber

Copyright © 2005 by Lee Eng Han.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, mechanical, photo-copying, recording, or otherwise, without the prior written permission from Lee Eng Han, who is one of the authors of the publication.



Table of Contents

Preface

1. Introduction	1
2. MOS Transistors	2
3. Fabrication of MOS Transistor	5
4. Layout a Single Transistor	10
First Stroke The basic transistor layout	11
Second Stroke Compact the transistor layout	12
Third Stroke Speed up the transistor	16
Forth Stroke Clean up the substrate Disturbances	18
Fifth Stroke Balancing area, speed and noise	24
Sixth Stroke Relief the stress	26
Seventh Stroke Protect the gate	27
Eighth Stroke Improve yield	28

Preface

Many IC design books emphasize on circuit design theories and there is little coverage on custom circuit layout techniques. Hence this book is specially written to focus on the custom circuit layout techniques. It is an easy book for the custom layout engineer as prior knowledge on circuit design is not necessarily required to understand the content of the book.

This book is written for two groups of audiences – New & Experienced custom layout designers. The new comer to custom layout would benefit from the wider perspectives in implementing the design to a layout. The experienced custom layout designers, on the other hand, will have a better appreciation of the rationale behind the layout practices.

Feedback to the authors

Please feel free to let us have your valuable inputs for future improvement. We would also appreciate if you could let us know if the book is of value to you. We can be contacted at layout@eda-utilities.com. Your feedback is most welcome.

Acknowledgements

The authors wish to thank Hwee Ling Goh from STMicroelectronics for reviewing the technical details. The authors also like to thank Karen Phang for her effort in proof-reading the book and giving invaluable suggestions in making the book easier to read.

Chapter 1

Introduction

Welcome to *Layout KungFu*!

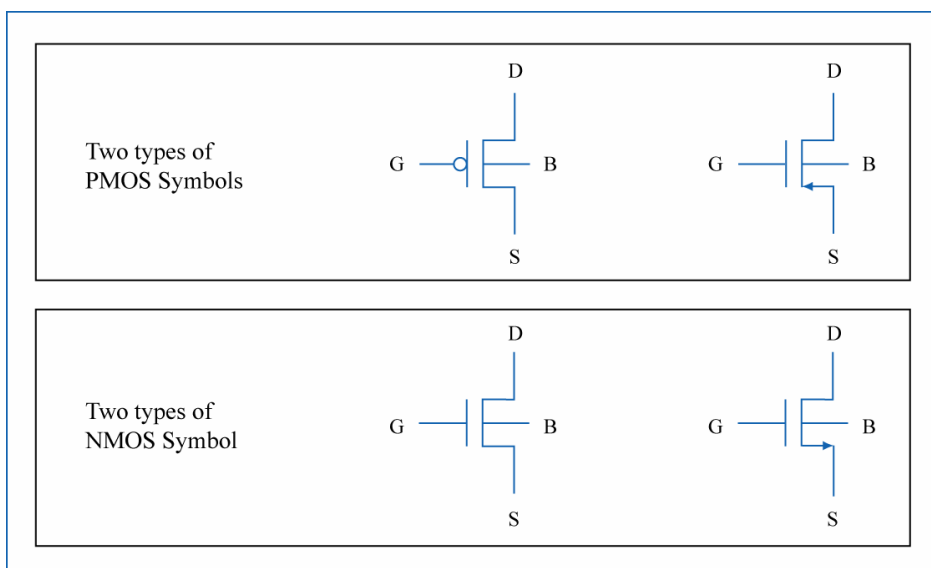
There are many levels of *Layout KungFu*, but we will focus on the fundamental and CMOS transistor layout is what you will find in this KungFu book. If you enjoy the KungFu and want to find out more, we could start a *Layout KungFu* series!

Please place your feet firmly on the floor, bend your knee at 90 degree and take a deep breath. Now, are you sitting comfortably on the chair? We shall begin the KungFu journey.

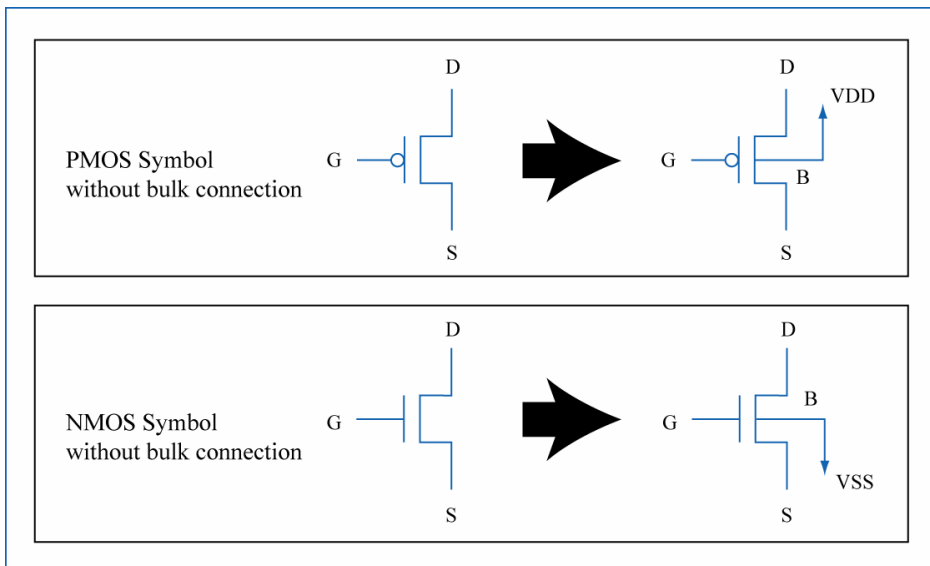
Chapter 2

MOS Transistor

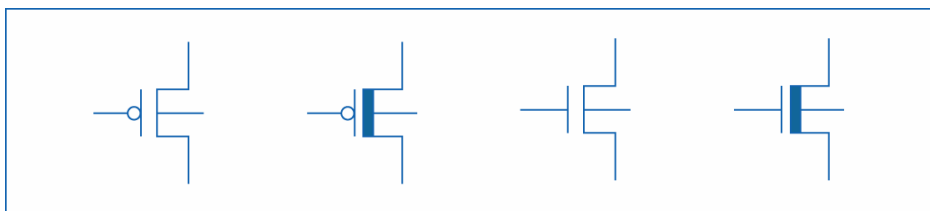
There are two types of MOS transistors. They are called n-channel MOS transistor (NMOS) and p-channel MOS transistor (PMOS). Each transistor has 4 terminals, namely drain (D), gate (G), source (S) and bulk (B) as illustrated in the transistor symbols.



The bulks of the PMOS and the NMOS are usually connected to power and ground respectively. If the bulk terminal is omitted from the schematic symbol, the connections can be assumed to be what is shown in the following diagram.



A circuit design usually uses more than one type of PMOS or NMOS to cater for different power supply voltages. For example, a design may use thicker gate-oxide transistors to operate in higher power supply at the I/O interfaces. In order to differentiate the different voltage range transistors, sometimes circuit designers conveniently make use of the “depletion-mode” transistor symbol for this purpose. Typically, the schematic symbols for the depletion-mode transistors are represented with a thicker gate drawing.



The MOS transistor’s performance varies with its channel length (L) and channel width (W). The drain current (I_D) that flows through the transistor operating in the saturation mode is shown in the following equation.

$$I_D = K * \left(\frac{W}{L}\right) * (V_{GS} - V_t)^2 * (1 + \lambda V_{DS})$$

where K and λ can be taken as process technology constants.

Note that I_D is proportional to the ratio of W over L . Typically L is kept to the minimum dimension allowed in the design rule, and is to be layout exactly as indicated in the schematic. However, this is not always the case for W . We will elaborate more in chapter 4.



Extra Punch

Transistor Spice Model

The Fab supports different types of transistor. For example, transistors can have different types of V_T to compromise between leakage power and speed. Transistors can also have different types of gate oxide thickness to allow the transistors to operate at different voltage range.

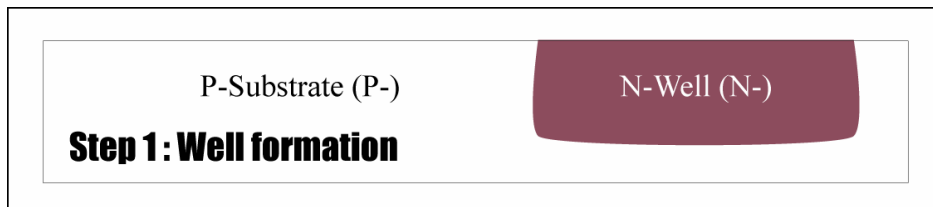
Every type of transistor is associated with its own transistor spice model. PMOS and NMOS also have different transistor model. A transistor model includes a set of parameters that define the electrical performance of the transistor. The design engineer uses the transistor spice model and the circuit netlist to simulate their design. The layout engineer has to craft the transistor layout that match the transistor spice model and the designed W and L .

Chapter 3

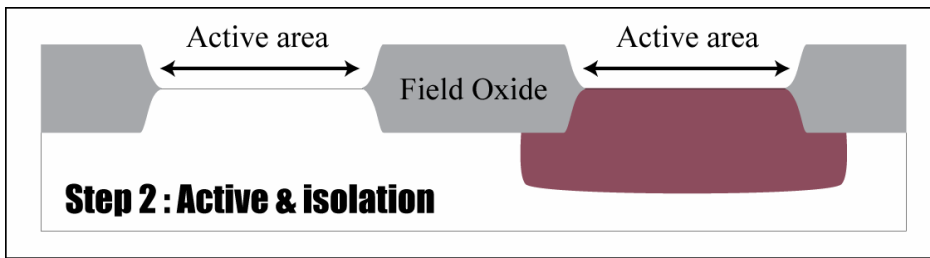
Fabrication of MOS Transistor

This chapter briefly describes a very simplified version of the fabrication of a transistor on the silicon wafer. The ability to visualize the cross-section of a layout is a basic skill that all layout designers should master.

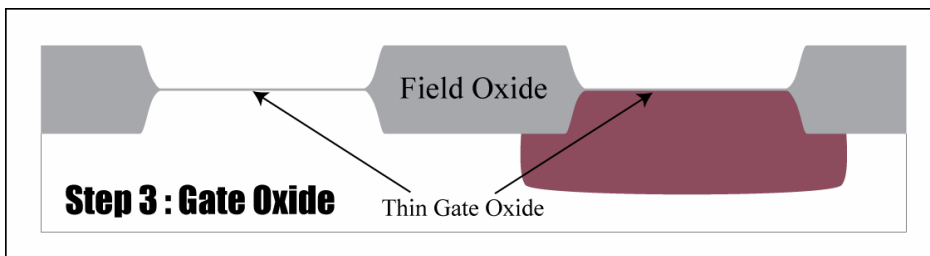
Step 1 : Well formation stage. Implants n-type impurities into the wafer followed by diffusing the impurities deep into the substrate to form the N-Wells. For CMOS process, the silicon substrate is usually p-type.



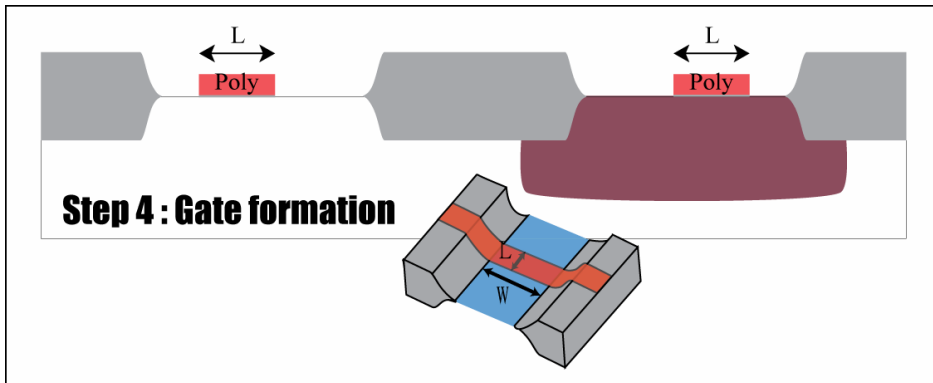
Step 2 : Active & isolation stage. Thick oxide is grown outside the active areas. Active areas are defined as areas where the CMOS transistors are fabricated. Thick oxide is also known as field oxide. Field oxides isolate the transistors from one another.



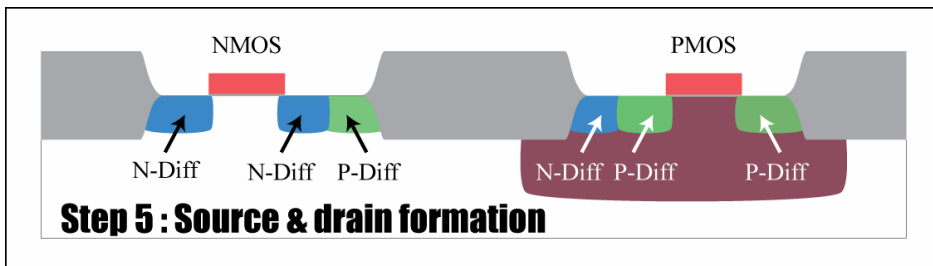
Step 3 : Gate oxide formation stage. A thin gate oxide is grown across the wafer. Gate oxide of only tens of silicon oxide atoms thick is created during the fabrication process with the current technology. Gate oxide is the insulator between the transistor's gate and its channel. Gate oxide refers to the "O" in "MOS" which stands for Metal-Oxide-Semiconductor.



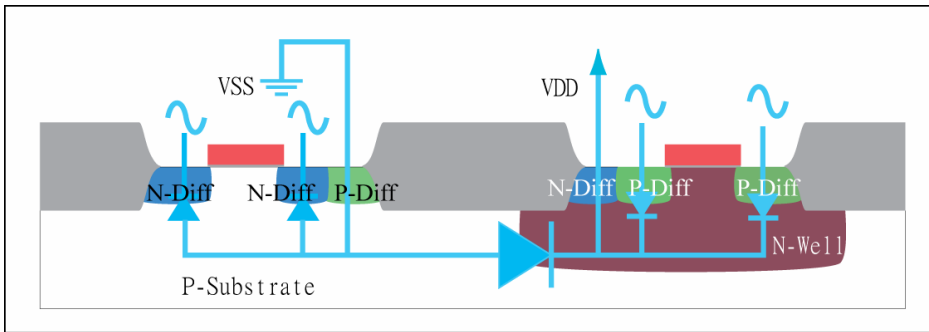
Step 4 : Gate formation stage. Poly (i.e. polysilicon) are deposited on the wafer. The poly that are deposited on the gate oxides are the gates of the transistors which are usually known as gate poly. The gate poly will incline upward when it extends over the field oxide. The gate oxide in the active area that are not covered by the gate poly will be etched away to form the source and the drain of the transistor.



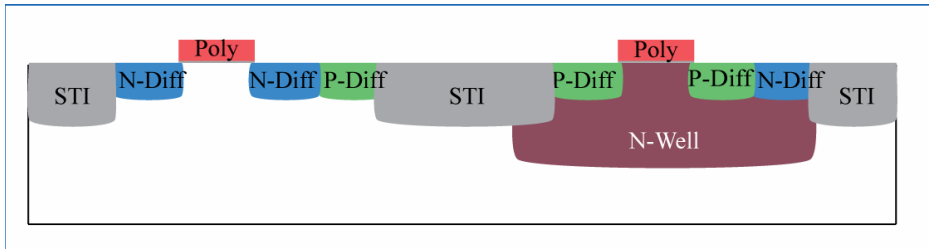
Step 5 : Source and drain formation stage. P-type and n-type impurities are implanted into the active areas. The impurities are diffused into the silicon to form the source terminals and the drain terminals. As the impurities diffuse both vertically and laterally, the gate poly will slightly overlap the sources and the drains which will result in gate overlap capacitances. The diffusions for the sources and the drains of NMOS and PMOS are N-diffusion (N-diff) and P-diffusion (P-diff) respectively.



P-diff in p-substrate is known as p-tap, while n-diff in N-well is known as n-tap. Connections from the metal routings to the substrate and the N-wells are made through the p-tap and the n-tap. This is necessary to ensure that the wells are properly tied down and the transistors are isolated. The p-substrate should be biased to the lowest voltage potential while the N-well should be biased to the highest voltage potential. In this way, all the P-N junctions are reverse biased and hence the transistors are electrically isolated from one another as shown in the diagram below.



Isolating the transistors with thick field oxide is commonly found in 0.35 μm ¹ and larger technologies. For 0.25 μm and smaller technologies, shallow trench isolation (STI) shown in the diagram below is more commonly used to isolate the transistors. In STI fabrication, trenches are etched into the wafer and filled with silicon oxide to isolate the islands of transistor active area.



¹ A technology of 0.35 μm means that the shortest channel length (L) of a transistor is 0.35 μm .

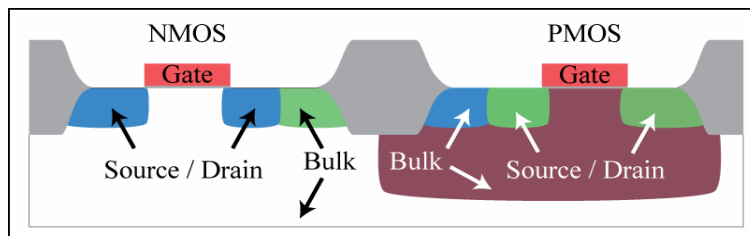


Extra Punch

Source, Drain Gate and Bulk

The drains, sources, gates and bulks of the NMOS and the PMOS are illustrated in the figure below. Observe the cross section of the layout and you will find the followings.

- The drain and the source are fabricated in the same way.
- The bulk of all the NMOS are connected together.
- The bulk of all the PMOS in the same N-well are connected together.














Chapter 4

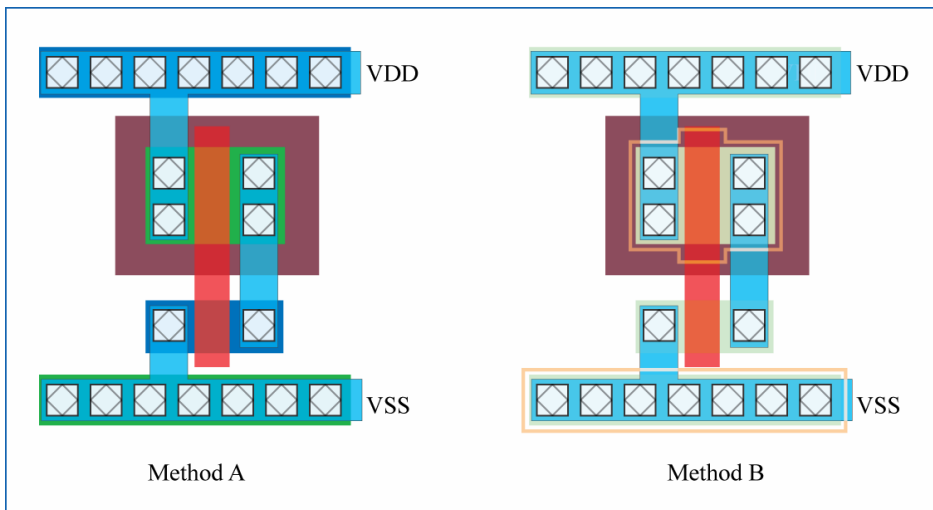
Layout a Single Transistor

Before we can start to layout the transistor, we have to settle a few logistics.

Firstly, the following legends are adopted.

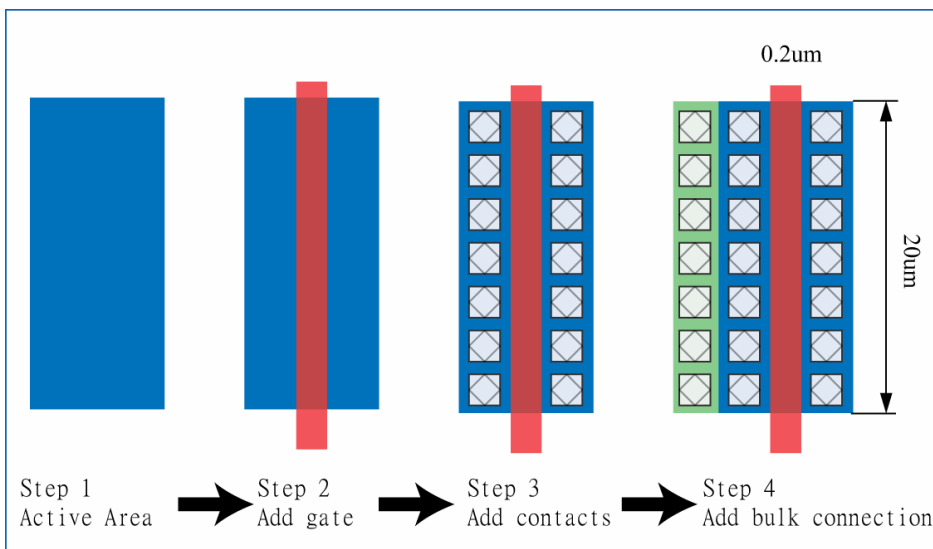
	Substrate (P-)		Poly		Diffusion
	N-Well (N-)		Metal-1		P+ Implant
	P-Diffusion (P+)		Metal-2		
	N-Diffusion (N+)		Contact		
			Via		

Secondly, there are more than one ways to draw an identical layout. For example, the two layouts in the following diagram are the same. Layout in “Method A” uses P-diffusion layer and N-diffusion layer. Layout in “Method B” uses diffusion layer and implant layer. The illustrations in this book use “Method A”.



First Stroke. The basic transistor layout

The basic transistor layout as illustrated has a channel length (L) of $0.2\mu\text{m}$ and a channel width (W) of $20\mu\text{m}$. The source diffusion and the drain diffusion should be filled with the maximum number of contacts to reduce the resistance of the connection from the metal to the diffusion, and to maximize the amount of current that can flow through the contacts.





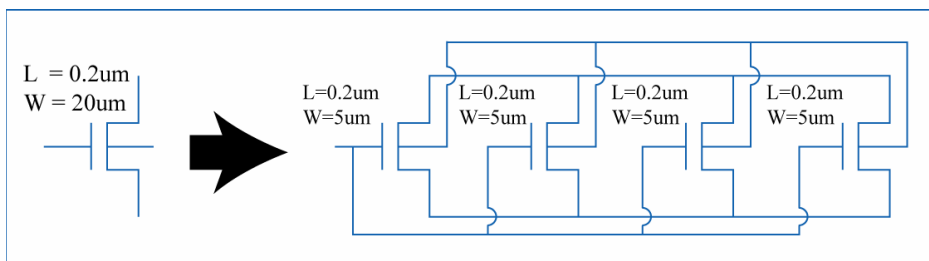
Extra Punch

Bulk Connection

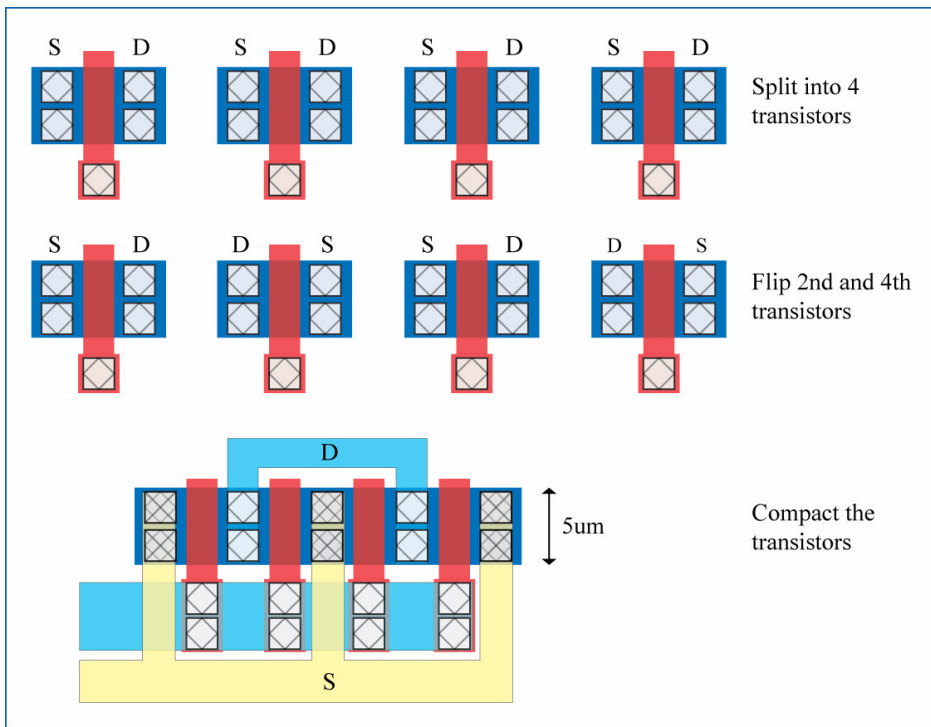
The substrate and the N-well are doped lightly. A direct connection from the metal routing layer to the bulk is not allowed. The connection should be made through a higher doped diffusion such as the p-diffusion and the n-diffusion in order to establish a good contact. In some fabrication process, two additional layers are used for these connections. These layers have much higher doping than the diffusions for the source and the drain.

Second Stroke. Compact the transistor layout

The basic transistor layout from the first stroke has a rather awkward aspect ratio. Putting transistors with fixed aspect ratio together will not give a compact layout. Fortunately, the aspect ratio of the transistor can be modified by using the transistor current equation shown in page 3. For example, the transistor with a width of 20 μm and a length of 0.2 μm is similar to having four transistors connected in parallel, each with a width of 5 μm and a length of 0.2 μm .



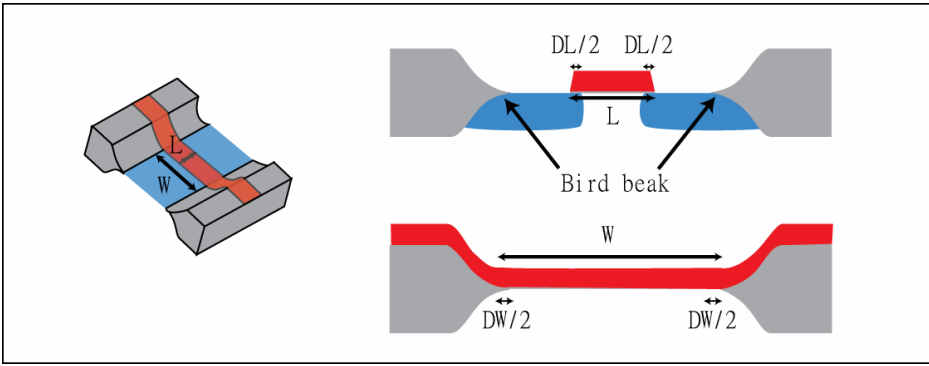
The next layout shows the transistor with four fingers. The layout has a better aspect ratio than the one in the first stroke. Note that the connection to the bulk is omitted in the layout. Bulk connection is discussed later.



So far everything looks fine. **But it is not.** The transistor with a width of 20um and a length of 0.2um is not exactly the same as four transistors that are connected in parallel, each with a width of 5um and a length of 0.2um.

Layout cannot be fabricated exactly as drawn in the layout due to the limitations in the manufacturing process, such as process tolerances and mask misalignment. Some of the manufacturing limitations are captured in the Spice transistor model. Two of the main parameters in the Spice transistor model are DW and DL. DW shows the delta difference of drawn W from effective W. DL shows the delta difference of drawn L from effective L.

The next diagram shows the transistor effective channel length can be affected by under-etching or over-etching of the poly, as well as the amount of lateral diffusion under the gate. The effective channel width of the transistor is affected by the “bird peak” of the isolation scheme. In addition, the inclination of the gate poly up to the field oxide makes it difficult to determine the exact width of the transistor.



Considering an example where DL is $0.015\mu\text{m}$ and DW is $0.045\mu\text{m}$. To simulate a typical “fast” corner transistor, the transistor is modeled to have a narrower L and a wider W . On the contrary, a “slow” corner transistor is modeled with a wider L and a smaller W . Hence, at the fast corner DL is negative (i.e. $-0.015\mu\text{m}$) and DW is positive while at the slow corner DL is positive and DW is negative (i.e. $-0.045\mu\text{m}$). The following table shows L and W of the transistors from the first stroke and the second stroke, and at slow, typical and fast corners to emulate manufacturing process tolerances.

		L (μm)	W (μm)
Transistor from first stroke	Fast corner	0.185	20.045
	Typical corner	0.200	20.000
	Slow corner	0.215	19.955
Transistor from second stroke	Fast corner	0.185	20.180
	Typical corner	0.200	20.000
	Slow corner	0.215	19.820

The width of the transistor from the first stroke differs by $0.045\mu\text{m}$ between the slow or the fast corner from the typical corner. However, the width of the transistor from the second stroke differs by $0.18\mu\text{m}$ between the slow or the fast corner from the typical corner. This could pose circuit performance deviation for the circuit designer if left unaccounted for.

The layout in the second stroke can be statistically significant. The circuit designer would perform Monte Carlo simulation to “center the design” so as to improve manufacturing yield. In Monte Carlo simulation, a small statistical variation is added to W and L of every

transistor in the circuit. Folding a transistor to four fingers means that a larger variation in the circuit performance since the variation made to the four fingers is four times larger than the same variation made to a single transistor. The layout can be challenging for the circuit designer to optimize the design for better yield.



Extra Punch

Operating Corners

Variations in fabrication process, ambient temperature and supply voltage affect the electrical performance of the transistors. For example, a higher temperature and a lower supply voltage make the transistor operate slower. Many companies require the circuit designers to verify the operation of the circuit design by simulating the design in slow (SS) corner, typical corner (TT) and fast corner (FF). Some companies also require the circuit designers to simulate the design with fast NMOS and slow PMOS corner (FS), and slow NMOS and fast PMOS corner (SF). Some companies require the circuit to work correctly within a 3 sigma spread from the typical corner (FF3, SS3).

Is folding a transistor into multiple fingers a bad idea? **Actually, it is an excellent idea.** Everyone uses it. Second stroke makes the layout compact. Third stroke is a continuation of the second stroke and it will enable the design to run at a higher speed!

Before we leave the second stroke, remember that whatever trick you used in the layout, it must be communicated to the circuit designer. The circuit designer needs to update the schematic to reflect the layout implementation. Using the same layouts from the first stroke and the second stroke as an example, the original spice netlist (i.e. the first stroke) of the transistor is

```
M1 D G S B N_model L=0.2u W=20u
```

The 'N_model' in the netlist is the name of the NMOS transistor model. The spice netlist for the layout in the second stroke is

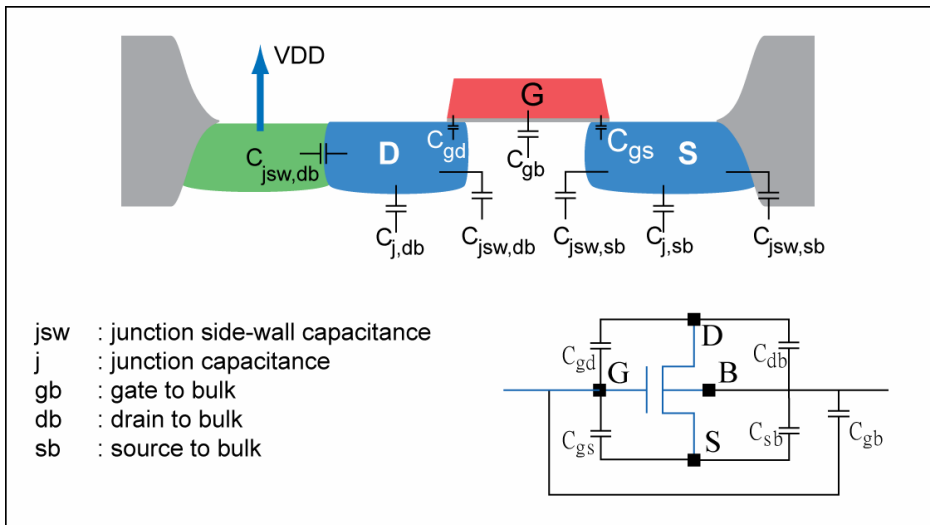
M1 D G S B N_model L=0.2u W=5u M=4

The circuit designer simulates the design using the spice netlist. Thus it is important for the spice netlist to represent the layout implementation as closely as possible.

Third Stroke. Speed up the transistor

What can you do in the layout to make the transistor operates faster?

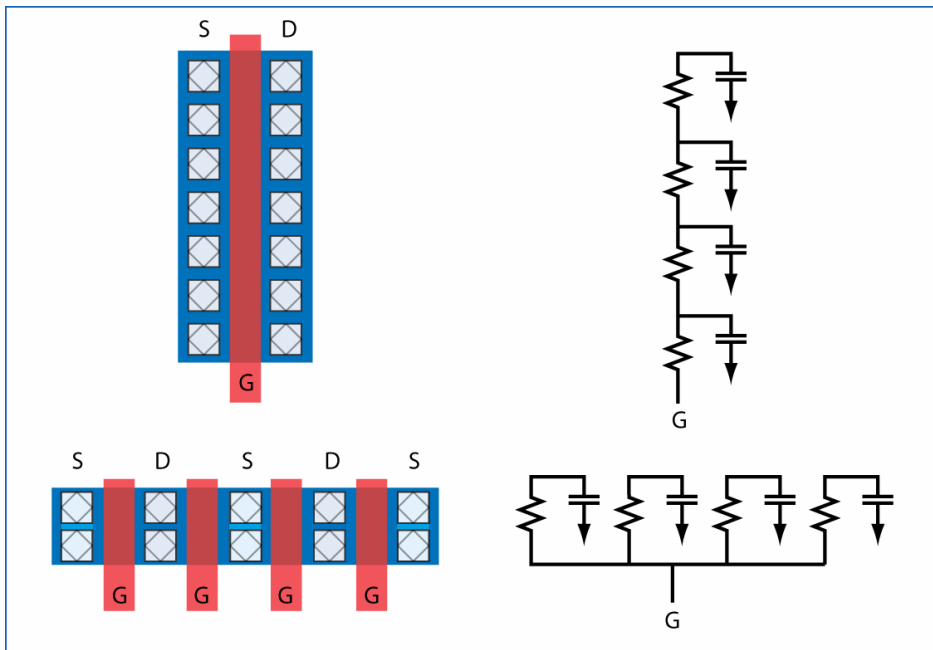
Reducing the parasitic capacitance and resistance would increase the speed of the transistor. We need to understand where the parasitic capacitances and resistances are. The following diagram shows the simplified capacitances associated with the drain, gate and source of a transistor to the bulk.



Overall C_{sb} is dependent on the area of source (AS) and perimeter of source (PS). Similarly, overall C_{db} is dependent on the area of drain (AD) and perimeter of drain (PD). Both C_{sb} and C_{db} have components that are dependent on the diffusions in the proximity. The values of AS, AD, PS and PD of a transistor can be extracted from the layout. Post-layout spice netlist should include these parameters.

The frequency response of the transistor can be improved if the source capacitance and drain capacitance are reduced. Study the transistors from the first stroke and the second stroke. Can you see that the transistor from the second stroke has the drain area reduced by half and the source area reduced by a quarter when compared to the transistor from the first stroke?

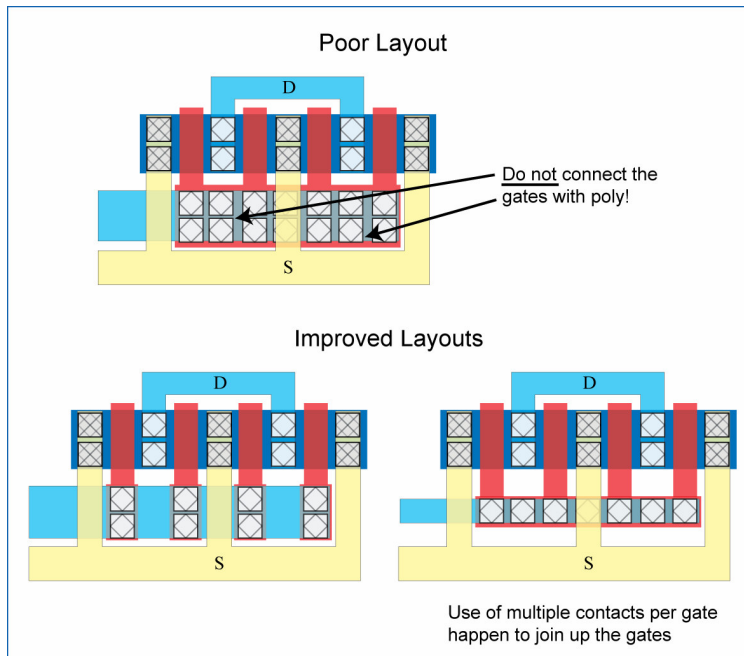
The folded transistors have smaller gate resistance as shown in the diagram below. This will make the transistors turn on and off faster than the transistor in the first stroke.



In general, try to fold a transistor to an even number of fingers. For example, the transistor from the second stroke is folded to four transistors, and is drawn to optimize the frequency response of the drain.

Resistivity of the poly is a few orders higher than the resistivity of the metal. The parasitic capacitances between the poly and the substrate, and between the metal and the poly, are also very much larger than the parasitic capacitance between the metal to the substrate. Hence, using poly for interconnect could degrade the frequency response of the transistor if the poly routing is not optimized carefully. Refer to the following layout. The layout at the top uses both poly and metal to

connect the gates. The first layout at the bottom uses only metal to connect the gates. The second layout at the bottom is a popular method to have multiple contacts per gate. To improve yield, the contacts for the gates in this layout are placed slightly further from the transistors so as to increase the distance between the diffusion and the poly that are running in parallel. The layout at the top will have poorer frequency response due to additional parasitic capacitances from the metal to the poly, and from the poly to the substrate.



Forth Stroke. Clean up the Substrate Disturbances

Performances of analog designs are sensitive to electrical disturbance. Disturbance in the substrate should be minimized as much as possible. Two common types of substrate disturbance are

- Disturbance from minority carrier
- Substrate coupling noise

Disturbance from minority carrier

Minority carriers are injected into the substrate from the source diffusions and the drain diffusions when

- The source potential or the drain potential of NMOS is below the substrate potential
- The source potential or the drain potential of PMOS is above the N-well potential

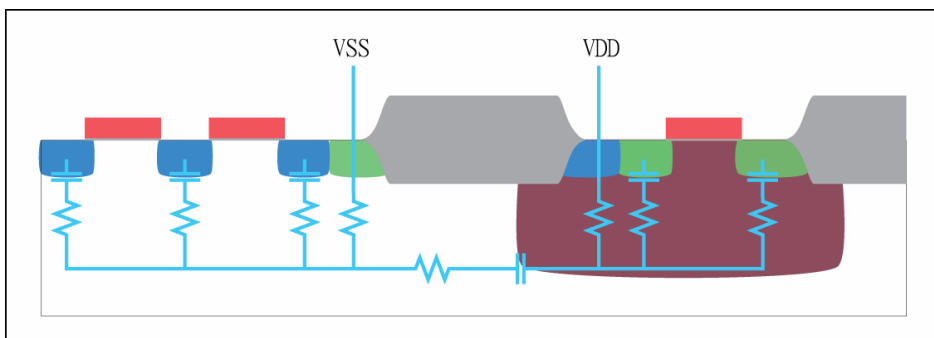
There are several possibilities for the above conditions to happen. Examples are

- Inductive ground path causes the ground in the substrate to bounce
- Resistive power and ground path from the power pins to the substrate and the N-well
- Fast switching signal with significant overshoot

The drifting of the minority carriers in the substrate and the N-well create a potential different that can affect the performance of the circuit, or trigger a latch-up.

Substrate coupling noise

A reverse biased diode has the electrical properties of a capacitor. Circuit signals can be coupled through the substrate as illustrated in the diagram below.



To reduce disturbances from minority carrier, you may use guard ring in the following configuration around noisy transistors.

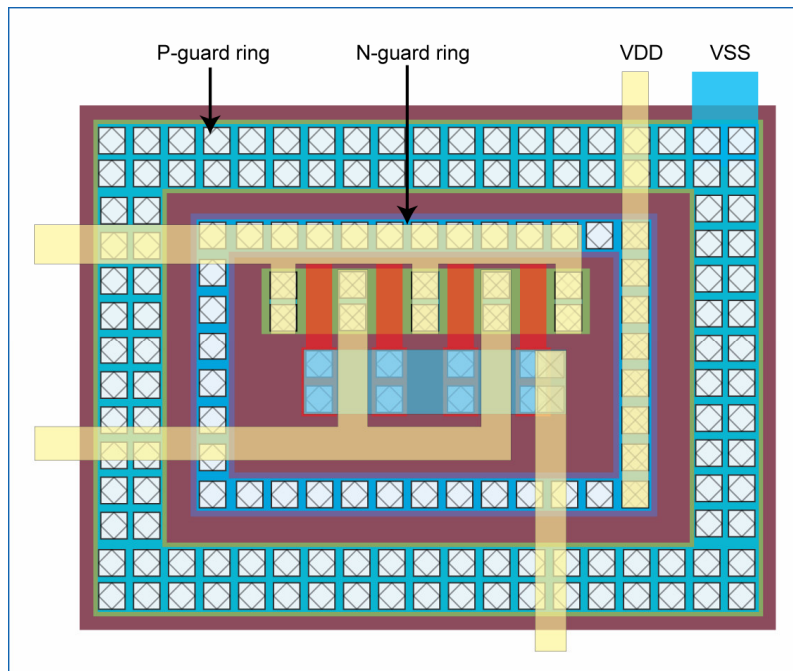
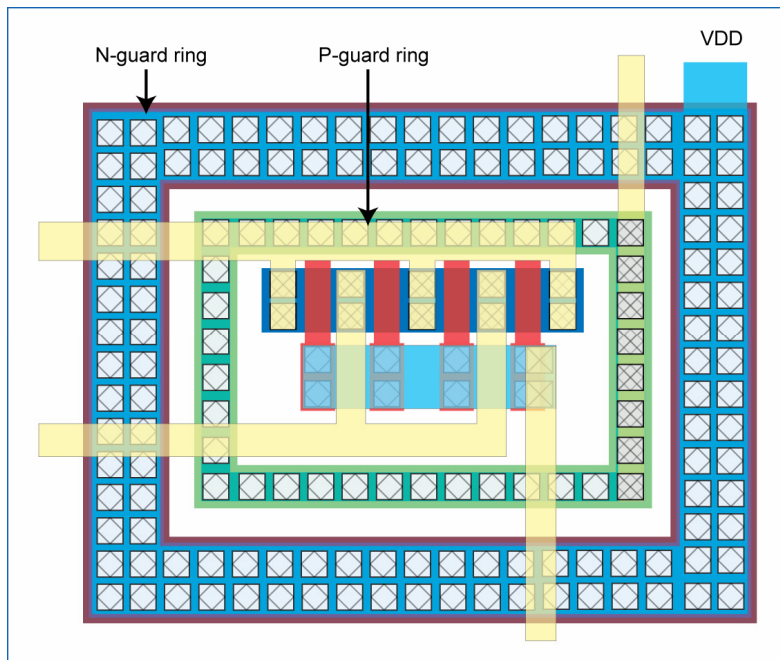
- Surround NMOS in the p-substrate with N-well guard ring. Tie the N-well guard ring to VDD. The N-diffusions from the NMOS could inject stray electrons into the substrate. These stray electrons could be collected efficiently by the N-well guard ring that is biased to VDD to attract the electrons.
- Surround the PMOS in the N-well with P-diffusion guard ring. Tie the P-diffusion guard ring to ground. P-diffusions from the PMOS inject stray holes into the N-well. These stray holes could be collected efficiently by the P-diffusion guard ring that is biased to ground to attract the holes.

For the guard rings to be effective, the resistance in the path from the straying minority carrier to the guard ring and then to the voltage source must be kept as low as possible. Hence, the minority carrier noise guard rings are made wider so as to decrease its resistance. Ideally, the guard rings should be placed as closely to the likely noise sources as possible. The guard rings are also placed around the critical transistors to minimize stray electrons and stray holes from affecting the critical transistors.

To reduce substrate coupling noise, you may use guard ring in the following configuration around critical transistors.

- Surround NMOS in the p-substrate with p-tap guard ring that is connected to ground.
- Surround PMOS in the N-well with n-tap guard ring that is connected to VDD.

The following layouts show both PMOS and NMOS surrounded with double guard rings.



It is generally believed that N-well guard ring in p-substrate and P-diffusion guard ring in N-well are not of much use. The stray electrons and holes travel deep into the substrate and are not collected by the

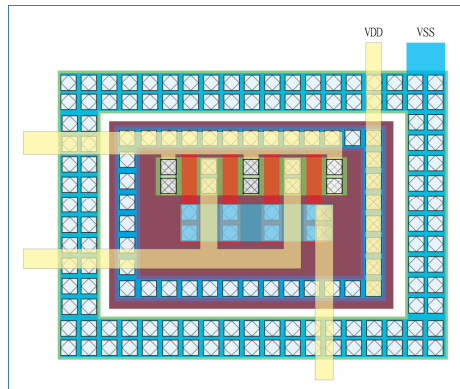
guard rings. However, putting a guard ring is better than leaving the space empty since we still have to keep the noisy transistors at a distance away from the other transistors.



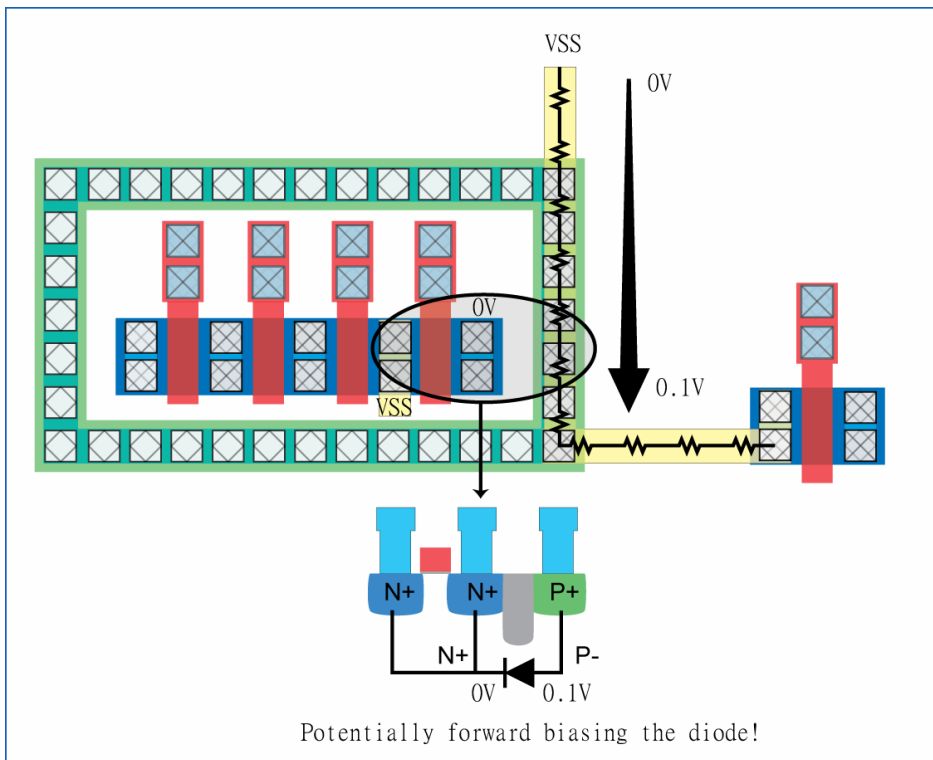
Extra Punch

Double Guard Ring for PMOS

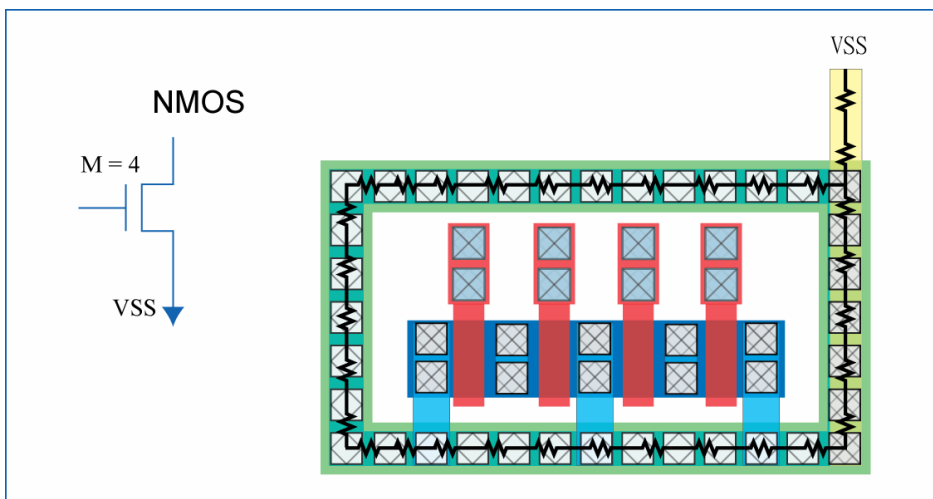
The most common way to add a double guard ring to PMOS is shown here. The outer P-guard ring acts as both a guard ring for PMOS, and as p-tap for the substrate. However, the P-guard ring might be too far from the PMOS to be effective in collecting stray hole from the PMOS.



An important layout practice is to ensure that there is no (or very little) current flowing through any part of the guard ring. Consider the layout in the diagram below. A current flowing in the p-type guard ring raises its potential above VSS. If the potential of the n-diffusion next to the p-type guard ring (shaded in the diagram) is at VSS, the PN junction potentially becomes forward biased, results in hole injected from the guard ring into the substrate!



A more subtle case is shown below. In the circuit design, the source of the transistor is connected to VSS. It is very convenient to tap the VSS from the guard ring as shown in the layout below. If the peak current going through the transistor is very small, the layout may be acceptable. Otherwise, tapping power from the guard ring is not allowed.

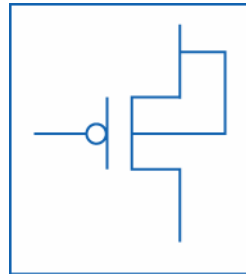




Extra Punch

Floating Well

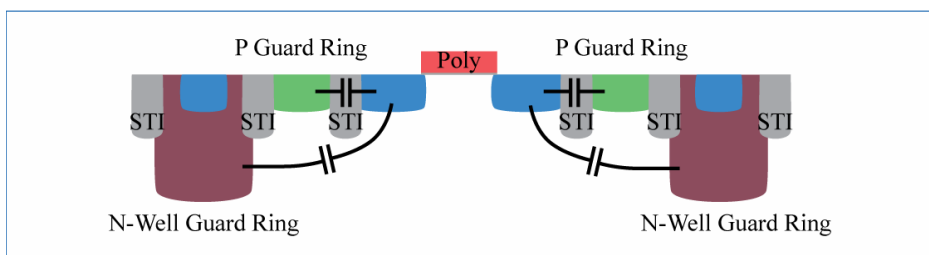
Wells that are tied to and varies with the source terminal of the transistor are called floating well. Such an example is shown on the right. Floating well is quite commonly used in analog design to reduce the “body effect” of the transistor. The PMOS



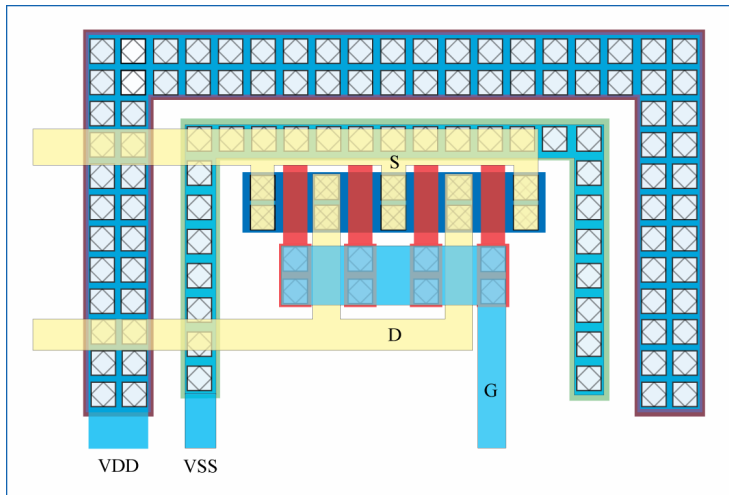
with this type of bulk connection needs an N-well on its own, and hence the layout has a huge area overhead. It is important to identify all the PMOS that sit on a floating well before floor plan the layout.

Fifth Stroke. Balancing Area, Speed and Noise

The guard rings in the forth stroke take up a lot of area. The guard rings also add capacitive load to the transistor as illustrated in the diagram below.



Instead of using a full guard ring, you may consider using a U-shape guard ring. Some designers do not favor the use of U-shape guard ring while some designers use U-shape guard ring only for p-tap and n-tap. The following layout shows a NMOS with U-shape guard rings.



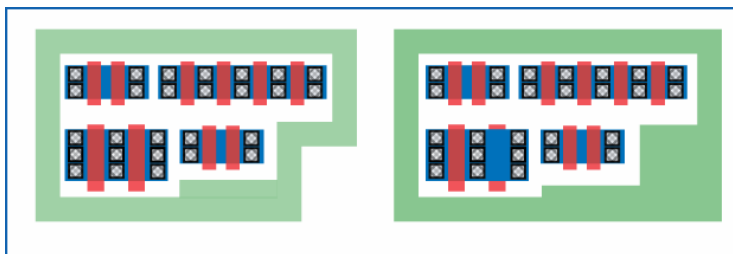
If reducing the area of the layout is of utmost priority and you do not expect much minority carrier noise in the silicon, then you may consider reducing the numbers of local N-well guard rings and P-diffusion guard rings.



Extra Punch

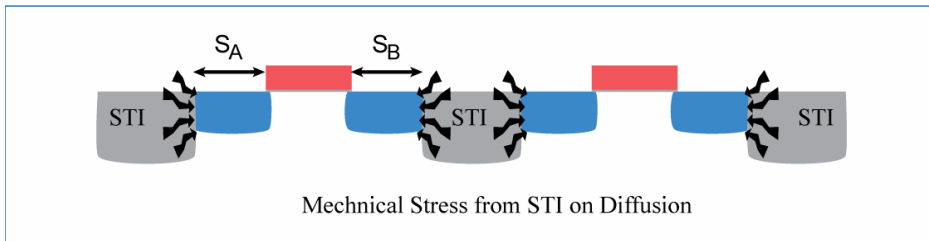
Guard Ring

The guard rings do not have to be rectangular. Two ways to add a guard ring around a group of transistors is shown here.

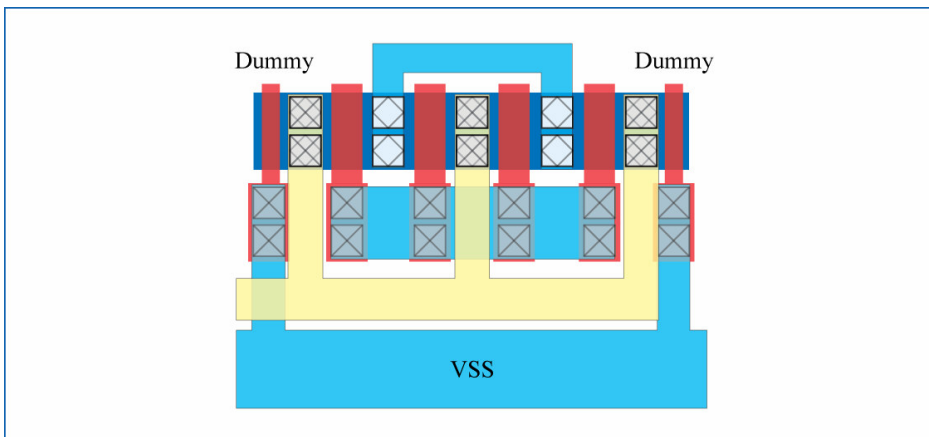


Sixth Stroke. Relief the Stress

The stress from the STI onto the drain and source has effect on the performance of NMOS and PMOS. The impact of the STI stress depends on the source and drain overhang (which are indicated as S_A and S_B in the following diagram) of the transistor active island.



To reduce the effect of STI stress, the source and drain diffusions need to be extended when they are next to the STI. However, a large diffusion also increases parasitic capacitance and layout area. A better approach is to insert one or more dummy transistors at each end of the transistor as illustrated in the diagram below. Note that the dummy transistor must share the diffusion with the non-dummy transistor.



Seventh Stroke. Protect the Gate

The gate oxide underneath the poly is incredibly thin. If the charges accumulated on the poly is sufficiently large, the charges accumulated can damage the gate oxide. This is known as process antenna effect.

The maximum amount of charges that can be accumulated on the poly is proportional to the area of the poly¹. Thus, an effective layout practice to prevent process antenna violation is to stay within the antenna ratio design rule of the respective technology. Some general guidelines are

- Minimize the use of poly for routing
- Minimize the use of poly to connect the gates together

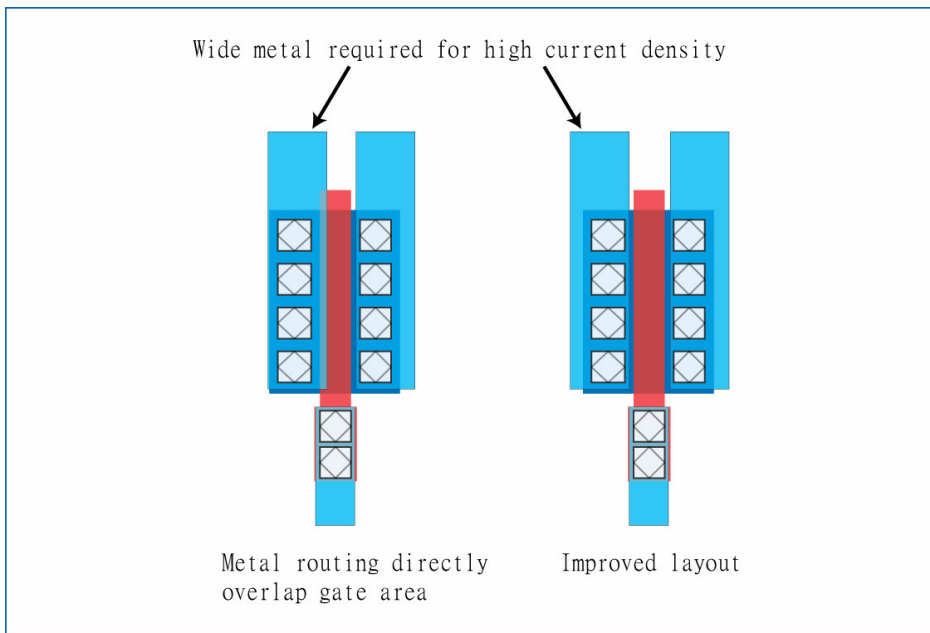
Some knowledgeable readers might suggest using diode to protect the poly from antenna ratio violation. Antenna diode is only effective in preventing antenna violation from metal routing, and does not help in antenna violation due to poly. The reason is simple. The diodes are made from diffusions, but the poly is deposited onto the wafer before the diffusions are implanted into the wafer. Hence, the diode does not exist at the time the poly is fabricated on the wafer!

Besides protecting the gates from process antenna effect, other measures to protect the gate are

- Do not place contact and via directly on top of the transistor's gate.
- Avoid routing over the gates of critical transistors. Refer to the following diagram for an example.
- Avoid routing over active areas of critical transistors.

An exception to the guidelines is to allow routing over decoupling MOS capacitors. This is done to compromise for a shorter routing.

¹ It is more correct to say that charges are accumulated on the perimeter side-wall area of the poly, which can be calculated as the perimeter of the poly multiple by the thickness of the poly.



Eighth Stroke. Improve Yield

The most compact layout does not give the best manufacturing yield. Use the layout practices discussed here to enhance yield. The practices are also illustrated in the following layout.

Avoid use of single contact or via

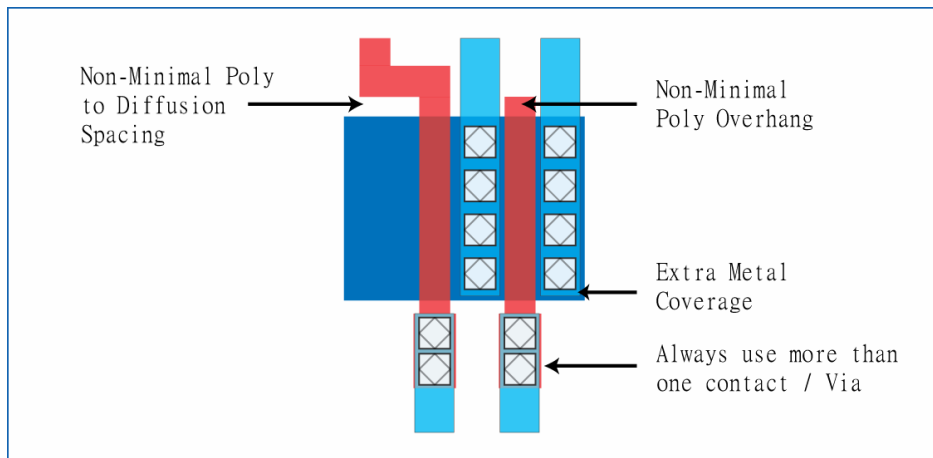
Avoid using single contact or via. Use at least double contact or double via whenever possible. A high percentage of IC manufacturing defects is related to faulty contact and via issues.

Metal coverage of contacts and vias

Always give additional metal coverage on the contacts and the vias if they are located at the end of the metal line.

Poly extension from diffusion

Exceed the DRC requirement for poly overhang rule and minimum distance from poly to diffusion rule whenever possible. In particular, do not run poly near to the diffusion edge.



Note

Chapter 5 and chapter 6 will not be included in the pdf release.