3D IC Devices, Technologies, and Manufacturing

3D IC Devices, Technologies, and Manufacturing

Hong Xiao

SPIE PRESS
Bellingham, Washington USA

Library of Congress Cataloging-in-Publication Data

Names: Xiao, Hong, 1961- author
Title: 3D IC devices, technologies, and manufacturing / Hong Xiao.
Description: Bellingham, Washington, USA : SPIE, [2016] | Includes bibliographical references and index.
Identifiers: LCCN 2016006975 (print) | LCCN 2016014372 (ebook) | ISBN 9781510601468 (softcover) | ISBN 9781510601475 (pdf) | ISBN 9781510601482 (epub) | ISBN 9781510601499 (mobi)
Subjects: LCSH: Three-dimensional integrated circuits.
Classification: LCC TK7874.893 .X53 2016 (print) | LCC TK7874.893 (ebook) | DDC 621.3815--dc23
LC record available at http://lccn.loc.gov/2016006975

Published by

SPIE
P.O. Box 10
Bellingham, Washington 98227-0010 USA
Phone: + 1 360.676.3290
Fax: + 1 360.647.1445
Email: books@spie.org
Web: http://spie.org

Printed in the United States of America.

The image of the 3D V-NAND memory chips on the cover has been provided by Samsung Electronics.

SPIE.

This book is dedicated to my wife, Liu Huang; my sons, Jarry and Colin; and my parents, Xian-ci Xiao and Hong-ting Zhou.

此书献给

父母: 萧先赐，周宏廷

妻子: 黄柳

儿子: 萧嘉瑞，萧凯瑞

Table of Contents

Preface

My first exposure to the semiconductor industry was in 1975 at the Microwave Diode Department of Chengdu Guoguang Electric Co. with my middle-school classmates during a month-long "learn from workers" program [very common for Chinese children during the chaotic period of the "Cultural Revolution" (1966–1976)]. The silicon wafer size was 1 inch, and we were making crystal diodes used in radar as a microwave detector. The factory had several process bays, such as diffusion, wet clean, wafer dicing, assembly, and final test. I watched workers push the wafers into the pyrogenic oxidation furnace and was amazed by the barely visible bluish hydrogen flame in it. I still remember the story of a hydrogen-leakage-induced explosion told during safety training. I worked in final test, using a special instrument to test the diode and determine whether it passed or needed to be thrown into the trash bin underneath the tester.

Twenty years later, I started my career in the semiconductor industry. The wafer size at that time was 200 mm, and the technology node was 350 nm. When the first edition of my textbook *Introduction to Semiconductor Manufacturing Technology* was published by Prentice Hall in 2000, the technology had scaled down to 180 nm, and copper metallization was the state-of-the-art technology.

Ten years after publication of the first edition, the wafer size increased to 300 mm, and the technology node migrated to 32 nm. New technologies that were not mentioned in the first edition, such as immersion lithography, double patterning, selective epitaxial growth (SEG), and atomic layer deposition (ALD), were widely used to manufacture IC chips with high-k, metal-gate front-end and copper, ultra-low-k back-end. It became the main driving force for me to write the second edition of the book.

There are many new developments since I have summited the final manuscript of the second edition in the spring of 2012. Because simply scaling down the feature size of the planar MOSFET can no longer improve the device performance while reducing its power consumption, scientists and engineers have worked on scaling nanometer-scale electronic devices in the third dimension. FinFET technology is one such proposed device architecture that has been used to replace planar MOSFET technology. At the same

feature size, FinFETs can improve the drive current by increasing the effective gate width at on-state while reducing standby leakage by operating with a fully depleted regime at off-state. Theoretically, FinFETs can migrate to the next-generation-technology node by just increasing fin height without shrinking the feature size. Because it is easier to control the fin height of the FinFET with the silicon thickness of the silicon-on-insulator (SOI) substrate, it was thought that FinFETs needed a SOI wafer. Due to the high cost of SOI wafers and the difficulties of fin height control with low-cost bulk-silicon wafers, many people regarded FinFET technology as a high-risk approach for the 22-nm or 20-nm technology node, even as late as 2009. In the summer of 2012, Intel announced its 22-nm FinFETs at the Symposium on VLSI Technology. Although FinFET technology was mentioned in the second edition published by SPIE Press at the end of 2012, it was not elaborated upon due to the lack of credible information about its manufacturing processes.

In recent years, the manufacturing technology of non-volatile memory (NVM), especially NAND flash memory, has developed rapidly, driven by the demands of data storage for mobile electronics devices, such as smartphones, tablet PCs, digital cameras/camcorders, etc. Multiple patterning is required to manufacture the planar NAND flash-memory chips due to the limitation of 193-nm immersion lithography. The cost of triple patterning or quadruple patterning required by the low-teen-nm planar NAND flash will become too high, and scientists and engineers have proposed and developed an alternative vertical NAND or 3D-NAND technology that utilizes the gate-all-around vertical transistors to stack multiple memory cells in the vertical direction. In 2014, Samsung released a solid state drive (SSD) based on 3D-NAND with 32 stacks of NVM—only seven years after Toshiba published the concept. A SSD with 48-stack 3D-NAND is also available on the market. With 3D-NAND architecture, one can scale to a next-generation-technology node by increasing the number of stacks without shrinking the feature size. The second edition of *Introduction to Semiconductor Manufacturing Technology* mentions 3D-NAND in the last chapter, which discusses future trends. The future becomes reality in a very short time.

Another technology mentioned in the second edition but not described in detail is 3D packaging with through silicon via (TSV). By stacking multiple chips with TSV, one can increase the device density without shrinking the feature size, which has been limited by the capability of 193-nm immersion lithography technology and the delayed implementation of extreme ultraviolet (EUV) lithography. TSV has long been applied in CMOS image sensor packaging, which forms the tiny camera assembly used in mobile phones, tablets, and laptops. TSV wafer stacking requires very high yield for every wafer that is to be stacked; otherwise, the combined final yield will suffer. Although foundries are still proposing 2.5D packaging with an interposer due

to the high cost of TSV 3D packaging, Samsung released the first 3D TSV technology based on DDR4 modules for enterprise servers in 2014.

Many people helped me acquire the information and knowledge needed to write this book; many of them helped me by answering my questions, and some of them helped by asking me questions to which I had no clear answer at that moment, which motivated me to further study and research: Dick James, Oliver Paterson, Hanming Wu, Jong (John) Chen, Chih-Ming Ke, David Fried, Sandy Wen, Jay Guan, Xiaodong Wang, Victor Lim, Byoung-Ho Lee, Ming Lei, Qiang Zhao, Kevin Huang, Jeff Zhang, Wee Teck Chia, Takuji Tada, Jeff Barnum, Christina Wang, Paul MacDonald, Chris Mahr, Brian Duffy, Harsh Shiha, Rohan Gosain, Arun Lobo, Neeraj Khanna, Amir Azordegen, and Cecelia Campochiaro, just to name a few.

The image of the 3D V-NAND memory chips on the cover has been provided by Samsung Electronics. Figures 1.10(b)–(c) and 2.48 are provided by Coventor. Figures 3.9–3.17, used to describe the HKMG FinFET processes, were previously published in TechDesign Forums.* These images were generated using Coventor's SEMulator3D virtual-fabrication software platform.

Colin Xiao, Jarry Xiao, Sameet Shriyan, and Shishir Ramprasad helped me proofread the draft and corrected many English errors. Without the support of my wife, Liu (Lucy) Huang, and sons, Jarry and Colin, it would have been impossible to write and finish this book on time.

My generation grew up in China without television. Thanks to the "Cultural Revolution," there were very few movies for kids in China at that time. So, hungry for movies, I watched anything that projected on the screen. One such film I watched many times was "Mechanical Drawing," an educational film for college students at the Chengdu Institute of Radio Engineering (currently the University of Electronics Science and Technology of China), where my parents worked as professors. Even today, I can still vividly remember this film taught me how a 3D object can be presented by a top view, side view, and face view. The 3D concept and its presentation with a 2D drawing helped me tremendously when I took an IC design class in graduate school. [IC layout is essentially the top view of mechanical drawing with a microscopic scale (maybe it should be called nanoscopic scale now)]. This knowledge was really useful for me to reconstruct the 3D structures of IC devices and figure out the manufacturing processes by correlating the top-view images and cross-section images. I really appreciate the person who showed the film and allowed me, an elementary schooler, to watch it with college students.

Hong Xiao
March 2016

*http://www.techdesignforums.com/practice/technique/finfet-iedm-tipsheet

3D IC Devices, Technologies, and Manufacturing

Chapter 1
Manufacturing Processes of 3D IC Devices

1.1 Introduction

The scaling of integrated circuit (IC) chips becomes more and more challenging as IC technology pushes the feature size deep into the nanometer (nm) technology nodes. To extend the scaling, engineers and scientists tried to not only shrink the feature size in the x and y directions but also push IC devices into the third dimension. It took 14 years from the first publication of fin-shaped field effect transistors (FinFETs)[1] to high-volume manufacturing (HVM) of 22-nm FinFET IC chips in 2012.[2] In 2014, the first 3D-NAND-based solid state drive (SSD) was introduced to the market,[3] only seven years after the first publication.[4,5]

The same electrical performance and a significantly smaller footprint (equivalent to feature-size scaling) can be achieved by changing the IC device from a two-dimensional (2D) planar structure to a 3D structure. Figure 1.1 shows this effect for capacitors: Fig. 1.1(a) is a planar capacitor, and Fig. 1.1(b) is a cylindrical capacitor with the same capacitance. It is well-known that capacitance $C = kA/d$. Here, k is the constant and d is the thickness of the dielectric between the two electrodes, and A is the area of the electrode. The top-down area of the cylindrical capacitor is much smaller on the wafer surface than that of the planar version. Increasing the height of the cylinder can further reduce the top-down area of the capacitor while keeping the capacitance unchanged. This is the main reason why DRAM chips have used cylindrical capacitors, either stacked or deep-trench, for a long time.

Figure 1.2 illustrates three types of metal-oxide semiconductor (MOS) field effect transistors (FET, or MOSFET). Figure 1.2(a) is a 2D planar MOSFET, Fig. 1.2(b) is a FinFET, which is a 3D device, and Fig. 1.2(c) is another 3D device, a vertical gate-all-around (GAA) MOSFET, or a silicon nano-wire device. The three devices have a similar gate critical dimension (CD) with a similar footprint; however, the channel width of the planar MOSFET is the narrowest and thus has the lowest drive current. The FinFET

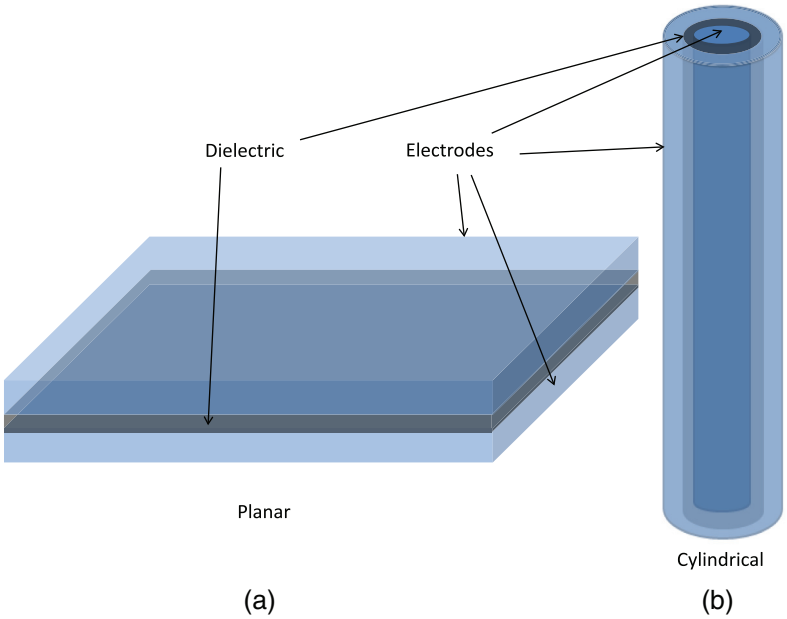


Figure 1.1 (a) Planar capacitor and (b) 3D cylindrical capacitor.

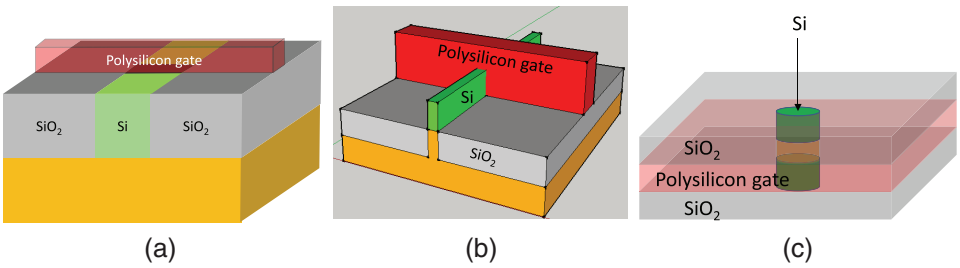


Figure 1.2 (a) Planar MOSFET, (b) FinFET, (c) and vertical gate-all-around device.

channel width is two times the fin height plus the fin top CD, whereas the GAA channel width is the circumference of the channel pillar.

This book explains the advantages of the 3D devices and their applications in dynamic random access memory (DRAM), 3D-NAND flash, and advanced-technology-node complementary MOS (CMOS) IC. The development of DRAM cell transistors and storage node (SN) capacitors of the DRAM, as well as the detail-manufacturing process flow of the most-advanced buried word line (BWL) DRAM, are discussed in Chapter 1. The 3D-NAND flash process flow is described in detail in Chapter 2, and step-by-step 3D FinFET CMOS IC devices are discussed in Chapter 3. Chapter 4 addresses the scaling trends of CMOS logic and memory IC before providing a brief summary. Devices that may be used in the "post-CMOS" era are also discussed. Other topics briefly mentioned include 3D technologies such as 3D wafer process integration of silicon-on-inter-layer dielectric (ILD) and TSV-based 3D packaging.

1.2 3D Devices in the DRAM and BWL DRAM Process

After reading this chapter, you should be able to

- List the two devices in a DRAM cell and draw the circuit;
- List the three types of cell transistor that have been used in DRAM manufacturing;
- Name the two types of storage node (SN) capacitor used in DRAM chips;
- Explain why the planar capacitor was phased out in DRAM chips long ago;
- Explain how DRAM capacitor technology has evolved;
- List at least two benefits of buried word line (BWL) DRAM; and
- List at least two types of high aspect ratio (HAR) structures in a DRAM chip.

DRAM is widely used in smartphones, tablets, laptops, desktop computers, data servers, and all computational devices. One of the major driving forces of IC technology developments involves scaling down the DRAM feature size to increase storage and speed while reducing power consumption. DRAM is the first IC product that utilized 3D devices in HVM. A 3D capacitor was implemented in DRAM HVM long before 22-nm FinFET CMOS and 3D-NAND flash.

1.2.1 DRAM introduction

A DRAM cell has two devices: one n-channel MOSFET (NMOS) as an access transistor, and one capacitor for data storage, as shown in Fig. 1.3(a). DRAM was invented by Robert H. Dennard in 1967. Figure 1.3(b) portrays the DRAM layout, and Fig. 1.3(c) shows the cross-section in Dennard's patent.[6] For this DRAM, both the access transistor and SN capacitor are planar devices with silicon dioxide as the gate dielectric and capacitor dielectric.

For a DRAM cell, the capacitance C of the SN capacitor must be large enough to hold enough charge to maintain the memory. For a planar-cell transistor DRAM, C usually is about 35 femtofarads (fF). In order to scale down the feature size of the capacitor while keeping the C value unchanged, the DRAM developers initially reduced d to the reliability required leakage limit and used silicon nitride ($k \sim 7$) to replace silicon dioxide ($k = 3.9$) for the capacitor dielectric. Due to the stress issue of nitride, the capacitor dielectric used oxide/nitride/oxide (ONO) for a long time.

1.2.2 3D devices in DRAM

An unchanged C and k made it very challenging to scale the feature size of a DRAM SN capacitor while maintaining its capacitance because the electrode surface area must be kept constant. The capacitor of the DRAM became the

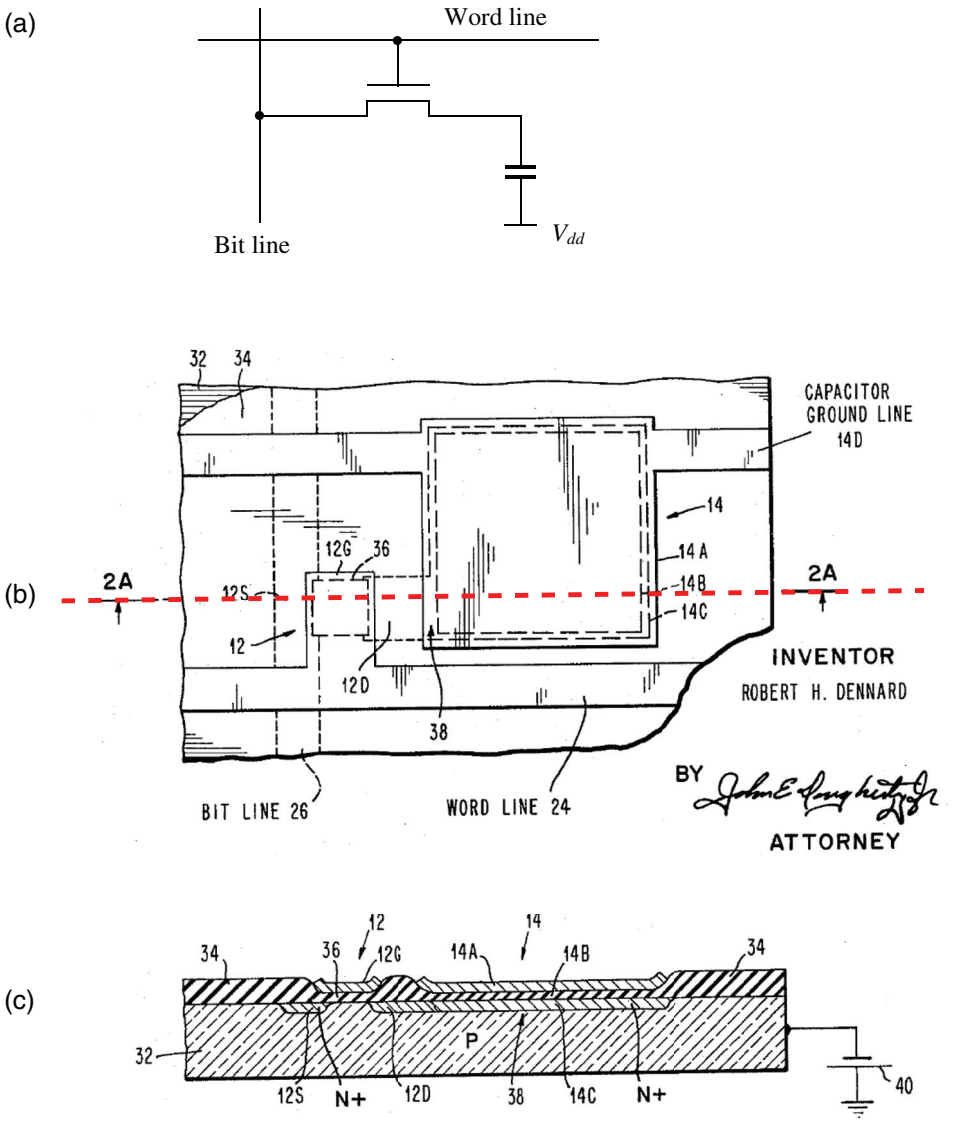


Figure 1.3 (a) DRAM circuit, (b) first DRAM layout, and (c) its cross-section. Figures (b) and (c) reprinted from IBM patent (1968, patentimages.storage.googleapis.com/pdfs/US3387286.pdf).

first IC device that became 3D after further scaling of the technology node, which helped reduce the footprint of the device.

Figure 1.4 shows the evolution of the DRAM capacitor. Engineers and scientists demonstrated a lot of creativity to scale the geometry of the SN capacitor while keeping its capacitance unchanged. HSG in the figure stands for "hemispherical grain," which is polysilicon that features rough, hemispherical,

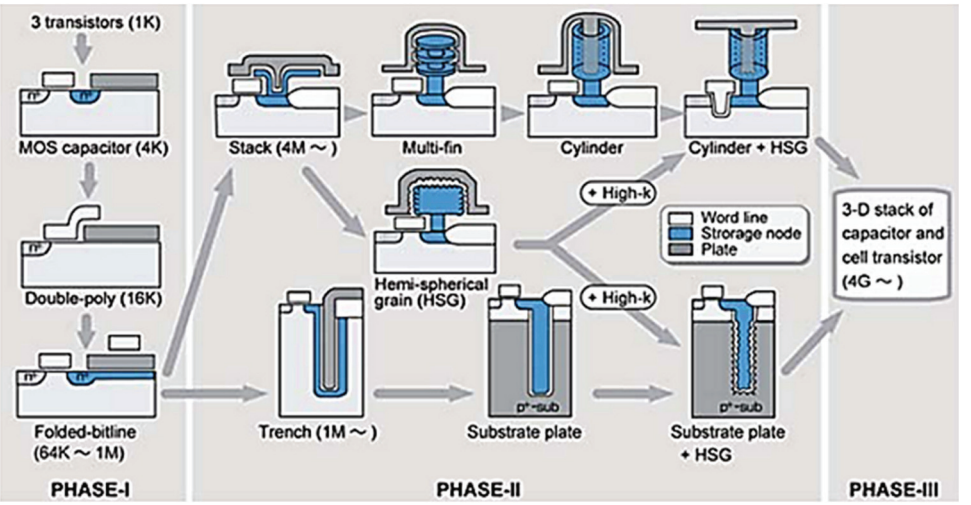


Figure 1.4 Technology evolution of the DRAM capacitor. Image reprinted from Ref. A with permission from InTech.

and grainy surfaces. It had been widely used in DRAM manufacturing to increase the surface area of capacitor electrodes. When DRAM technology scales to PHASE-III, the CD of the cylinder becomes too small to grow HSG silicon; it has been replaced by a thinner metal-electrode (mostly TiN) layer.

Figure 1.5 illustrates the advanced SN capacitors indicated in phase III of Fig. 1.4. Figure 1.5(a) shows a deep-trench capacitor DRAM; the dashed line indicates that a lot of substrate depth with the same structure has been omitted from the figure. The aspect ratio (trench depth versus trench CD) could be higher than 30:1. Currently, only a handful of silicon-on-insulator (SOI) IC chips use deep-trench capacitor DRAM as embedded DRAM, taking advantage of its compatibility in the back-end-of-line (BEoL) processes with normal CMOS processes and a lower capacitance requirement due to the SOI substrate. Most DRAM chips in the commodity market are DRAM with a cylindrical stacked capacitor, as shown in Fig. 1.5(b), which is described in detail in this chapter.

Question: For a cylindrical capacitor, how does the aspect ratio of a SN cylinder change when the DRAM feature size scales down by a factor of $1/\sqrt{2}$ while C, k, and d remain unchanged?
Answer: In order to keep C unchanged in this scaling, the height of the cylinder must be increased to scale down its CD:

$$A_1 = \pi \times CD_1 \times h_1 = \pi \times CD_2 \times h_2 = A_2;$$
$$CD_2 = CD_1/\sqrt{2}, \text{ and thus } h_2 = \sqrt{2} \times h_1;$$
$$AR_2 = h_2/CD_2 = \sqrt{2} \times h_1/(CD_1/\sqrt{2}) = 2 \times h_1/CD_1 = 2 \times AR_1.$$

If C, k, and d remain unchanged, then the aspect ratio of the SN cylinder will be **doubled** after scaling down the SN cylinder CD by a factor of $1/\sqrt{2}$!

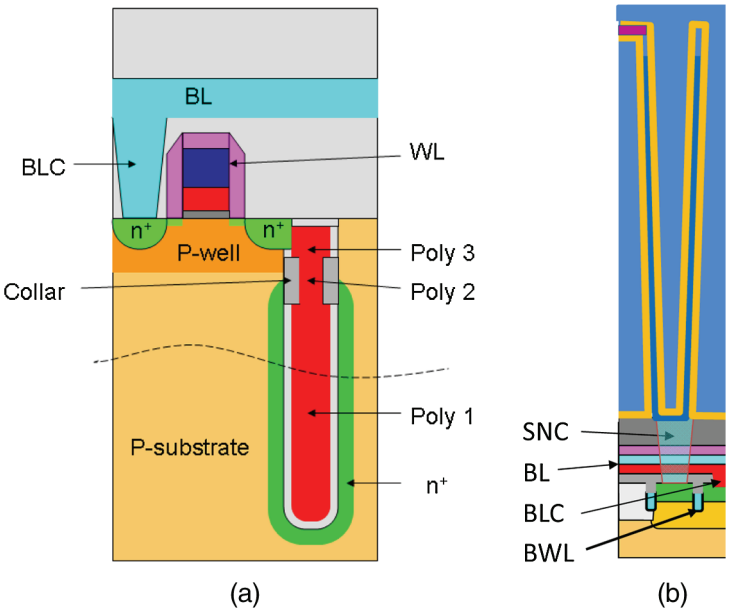


Figure 1.5 (a) Deep-trench capacitor DRAM and (b) recessed cylinder-stacked capacitor DRAM.

Shrinking the footprint of the SN capacitor is the biggest challenge of DRAM scaling. Engineers and scientists are working to find new dielectric materials with higher k values and to reduce leakage of the off-state access NMOS to reduce the required capacitance C. When the feature size scales to the next generation, the scaling factor is no longer as aggressive as $1/\sqrt{2}$, which means that the increased aspect ratio of the SN cylinder is still manageable for each technology node.

When DRAM technology scaled toward the 80-nm node, short channel effects caused higher source/drain (S/D) leakage in the access transistor and shorter data-retention time. To overcome those issues, the DRAM access transistor transitioned from a planar structure to a 3D structure, and the recess gate (RG) transistor was introduced.[7] Further development of the DRAM access transistor created the buried word-line (BWL) technology, which uses metal to form the gate electrode of the access transistor and the word line (WL), both of which are buried under the original silicon wafer surface.[8]

Figure 1.6 illustrates the evolution of the DRAM access transistor, where L is the channel length. Figure 1.6(a) shows a planar MOSFET, and Fig. 1.6(b) illustrates a RG transistor; both use polysilicon as the gate electrodes. A comparison of Figs. 1.6(a) and 1.6(b) reveals that with the same gate critical dimension (CD), the RG transistor has a longer channel length than a planar transistor. Figure 1.6(c) shows a buried access NMOS with a titanium nitride (TiN) metal gate and a tungsten (W) WL buried underneath the silicon surface.

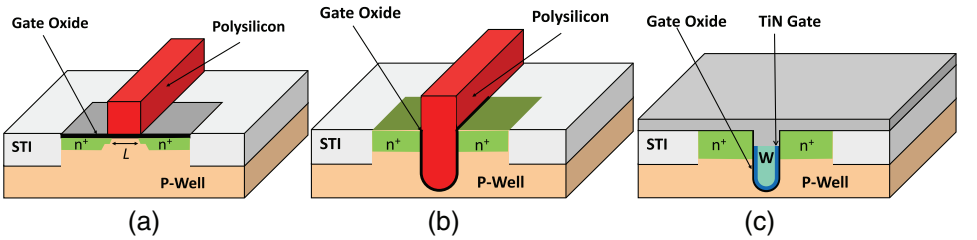


Figure 1.6 Evolution of the DRAM array cell NMOS: (a) planar, (b) RG, and (c) BWL.

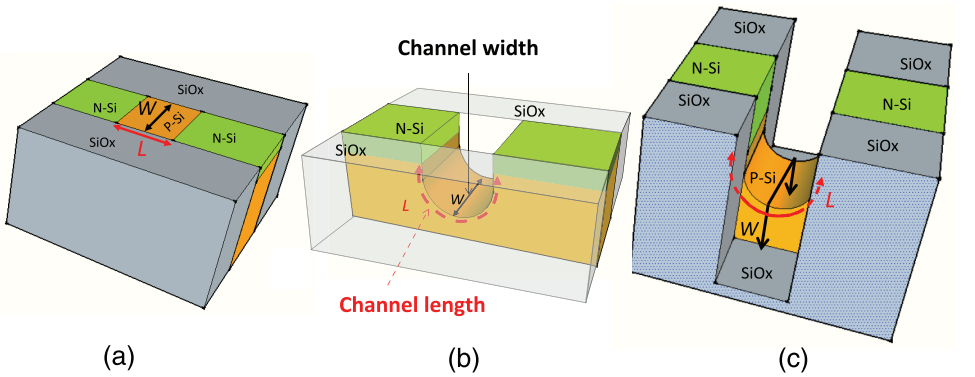


Figure 1.7 3D illustration of the channel width and channel length of DRAM cell transistors: (a) planar, (b) RG, and (c) BWL.

The channel length L and channel width W of the planar cell transistor, RG transistor, and BWL transistor are shown in Figs. 1.7(a)–(c), respectively. Note that L is the distance along the silicon surface between the source and drain, the heavily n-type-doped silicon (marked N-Si in the figure). At the same feature size, the RG transistor has the same channel width as the planar transistor, which is the CD of the AA; however, it has a much longer channel length. Furthermore, at the same geometry, the BWL transistor has the same L as the RG transistor, with a larger W, because there is more silicon surface exposed at the sides due to the deeper oxide etch in the BWL trench etch. The edge of the P-Si channel side and its curved top in real BWL devices is not that sharp—it is more like the shape of a saddle. Increasing the channel length can help to reduce leakage and maintain the retention time. Increasing the channel width can help to increase the drive current, and the BWL structure allows increasing the channel width without increasing the feature size of the device.

Figure 1.8 shows a DRAM with a planar cell NMOS and stacked cylinder capacitor. It represents the DRAM technology at ~110 nm or earlier technology nodes (described in detail in Chapter 14 of Xiao[11]), which used three layers of aluminum metal interconnection. The so-called aluminum metal layer in fact is a metal stack of Al-Cu (~0.5% Cu) alloy bulk layers with a Ti/TiN barrier layer underneath and a TiN anti-reflective coating (ARC) on

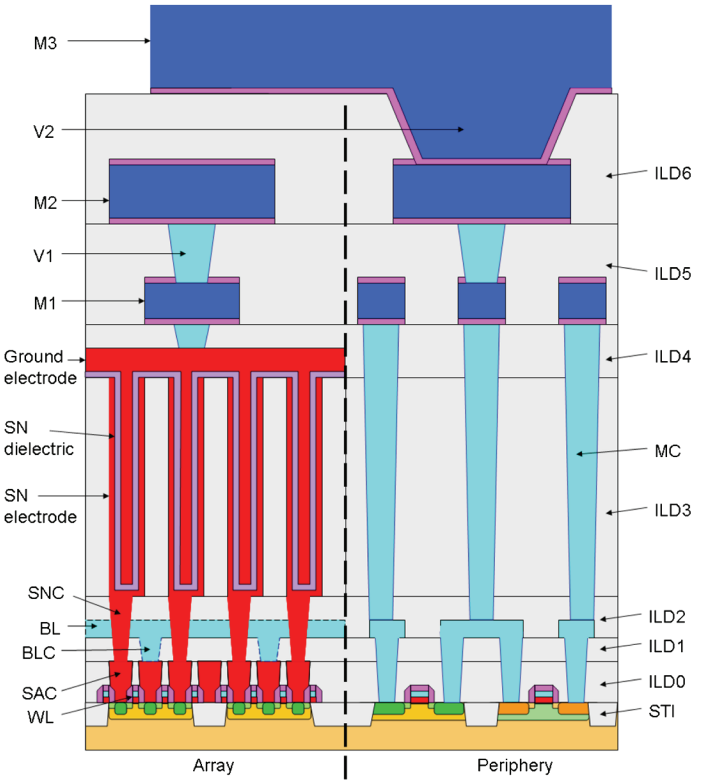


Figure 1.8 (a) Planar cell transistor DRAM and (b) 3D illustration of its memory cell structures. BL: bit line, BLC: bit-line contact, ILD: inter-layer dielectrics, MC: metal contact, Mx: metal *x* (where $x = 1$ to 3), SAC: self-aligned contact, SN: storage node, SNC: storage-node contact, and STI: shallow-trench isolation. (b) is modified from Ref. B.

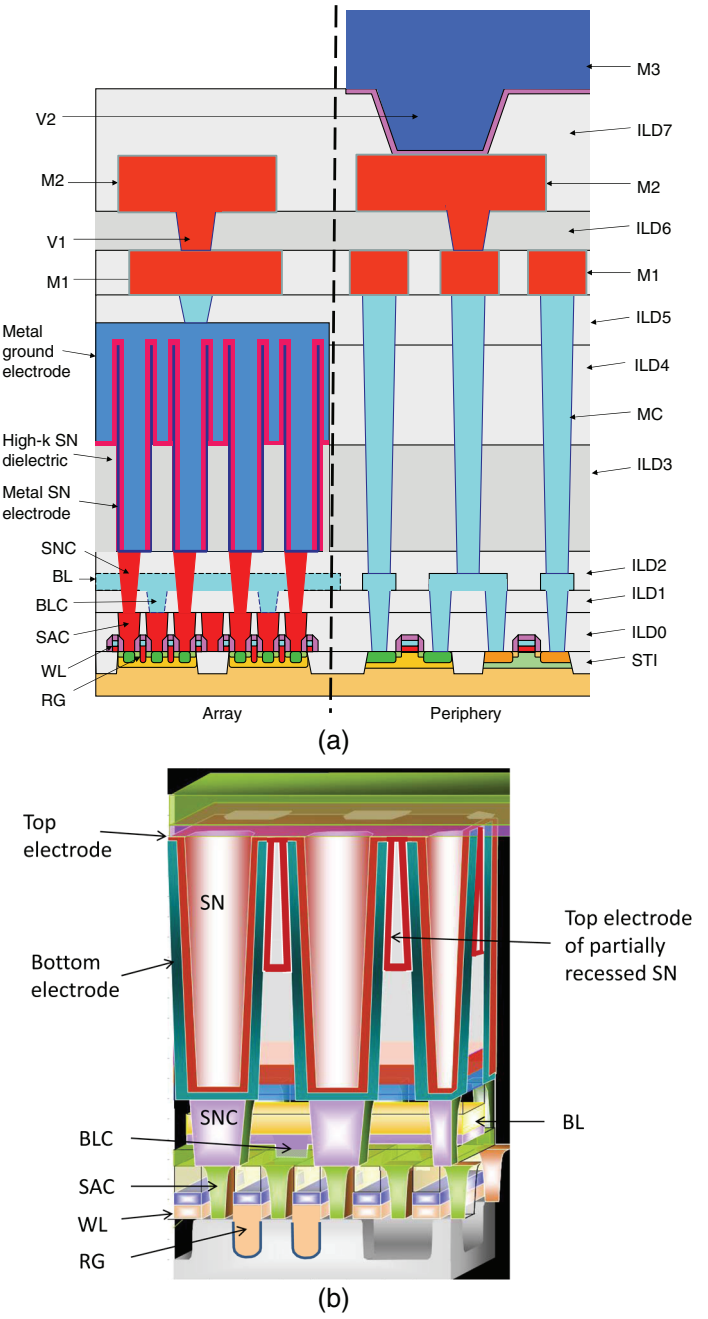


Figure 1.9 (a) Recessed-gate cell transistor DRAM, and (b) 3D illustration of its memory cell structures, which is modified from Ref B.

top. The W with the Ti/TiN barrier layer was used for the conducting plugs to connect the different layers.

Figure 1.9(a) shows a DRAM with a RG transistor and partially recessed cylinder capacitor. These images represent DRAM technology after 90-nm

and before 3x-nm (i.e., 39-nm to 30-nm) technology nodes. Besides cell transistors and capacitors, another major development of this period was the introduction of copper metallization in 5x-nm DRAM. In comparison, logic CMOS IC chips used copper metallization in interconnects since the 180-nm technology node. There were two copper layers and one aluminum layer. The top aluminum-alloy layer allowed the use of Al-Cu bond pads for the standard gold wire bonding to keep the cost down. DRAM manufacturers are very cost-sensitive, and DRAM fabrication-process technologies are cost driven: two layers of copper can replace three layers of aluminum alloy, which reduces the overall cost.

Figure 1.10(a) illustrates a BWL DRAM with a fully recessed cylinder capacitor, and Fig. 1.10(b) is the 3D illustration of its memory cell structures. Because the latter is based on the $6F^2$ layout, it is quite different from Figs. 1.8(b) and 1.9(b), which are based on the $8F^2$ layout (these layouts are discussed in Section 1.2.3). The BWL structure and $6F^2$ layout have been widely used in 3x-nm and 2x-nm DRAM technology nodes. There are three metal interconnect layers: two copper layers and one aluminum-alloy layer.

If DRAM can be made with these three types of cell transistors with the same single-patterning photolithography technology and the same capacitor materials and structure, then the planar-cell transistor DRAM requires nine masks, as shown in Fig. 1.11. The RG transistor DRAM adds a reversed WL mask for the silicon etch to form a RG cell transistor; thus, it needs ten masks in the array area, as illustrated in Fig. 1.12. The BWL DRAM uses one mask for both well implantation and S/D implantation, and one mask to form the RG and WL at the same time, while eliminating the SAC layer; thus, it only needs seven masks (Fig. 1.13). Fewer masks means lower cost, which is one of the most important reasons why BWL DRAM technology is used by all major DRAM manufacturers. Note that Figs. 1.11–1.13 only show the mask layers for the cell transistors and capacitors of the DRAM array area; they do not include the masks required to make CMOS devices in the peripheral areas.

Ion implantation masks in cell transistor formation and well implantation after active-area (AA) and S/D implantation after WL have a very large CD and require less strict control of the CD and overlay. In a DRAM chip, the feature sizes of peripheral circuits are much larger than the feature size of cell devices. The feature size of peripheral devices is usually two generations behind cell transistors. Advanced-technology nodes are processed separately to avoid loading effects caused by pattern-density and pattern-size variations on the processes such as photolithography, etch, CMP, etc. Although access transistors used 3D devices for multiple generations, peripheral devices still use planar MOSFET technology. Therefore, this book focuses on the manufacturing processes in 3D devices in array areas.

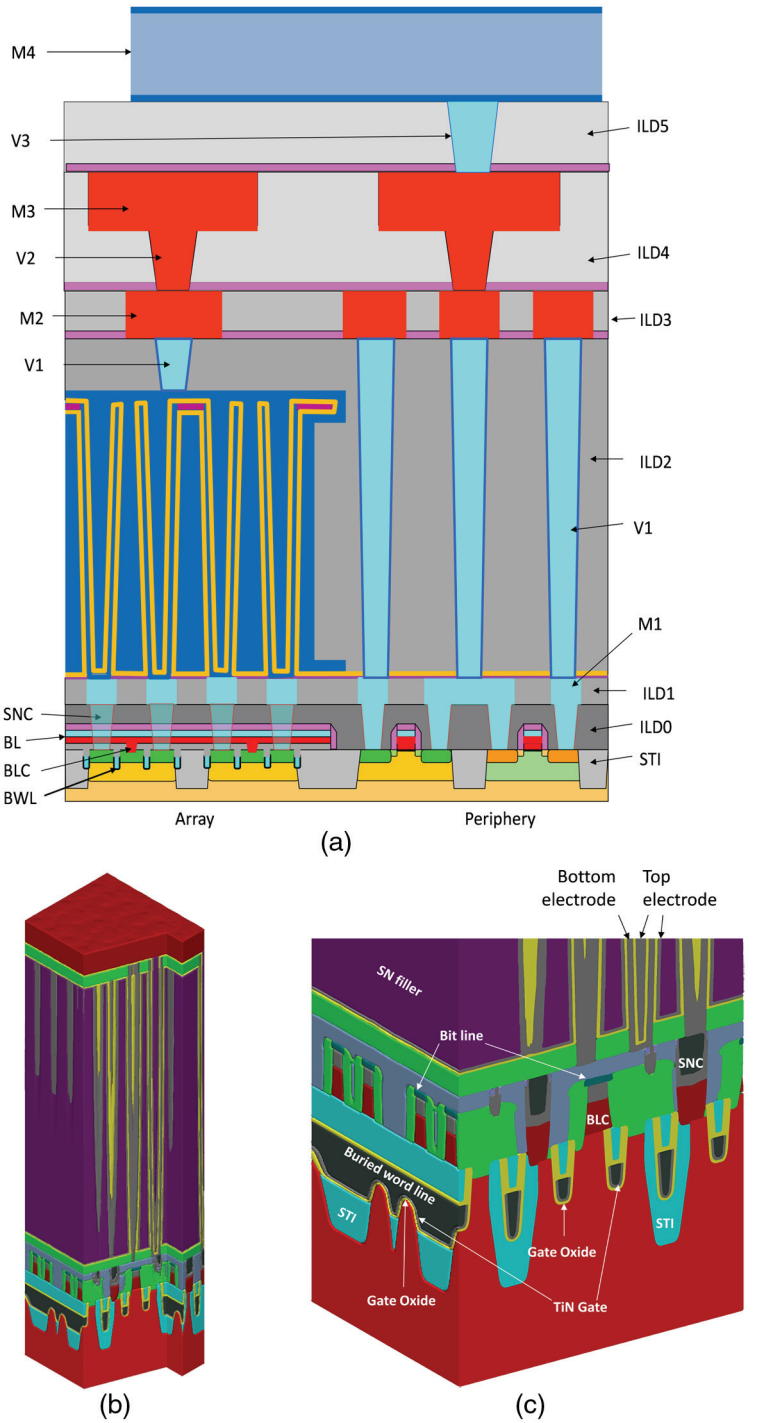


Figure 1.10 (a) BWL DRAM with a fully recessed cylinder capacitor, two layers of copper, and one layer of aluminum interconnect, (b) 3D illustration of its memory cell structures, and (c) close-up of the memory cell. Both (b) and (c) are reprinted with permission from Coventor.

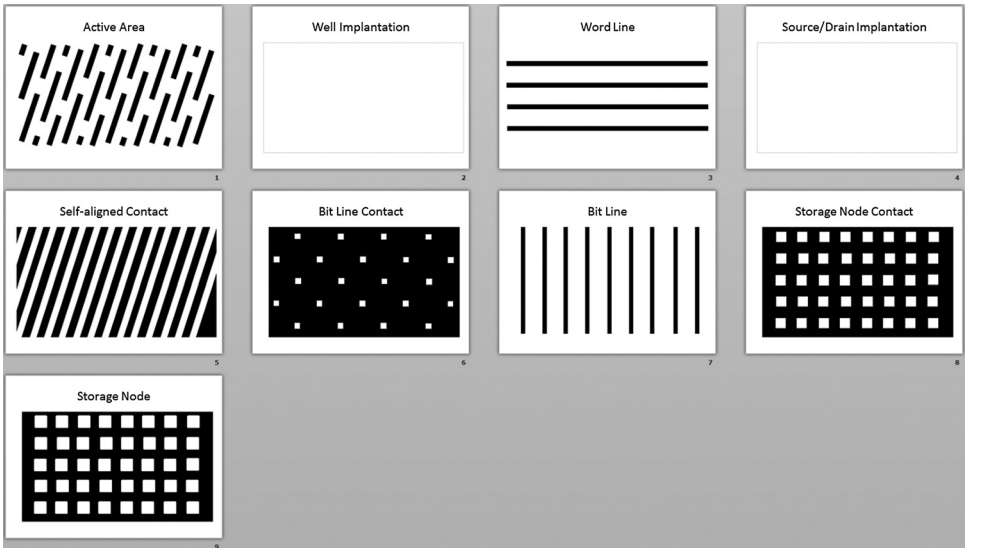


Figure 1.11 Masks for planar-access transistor DRAM in the array area.

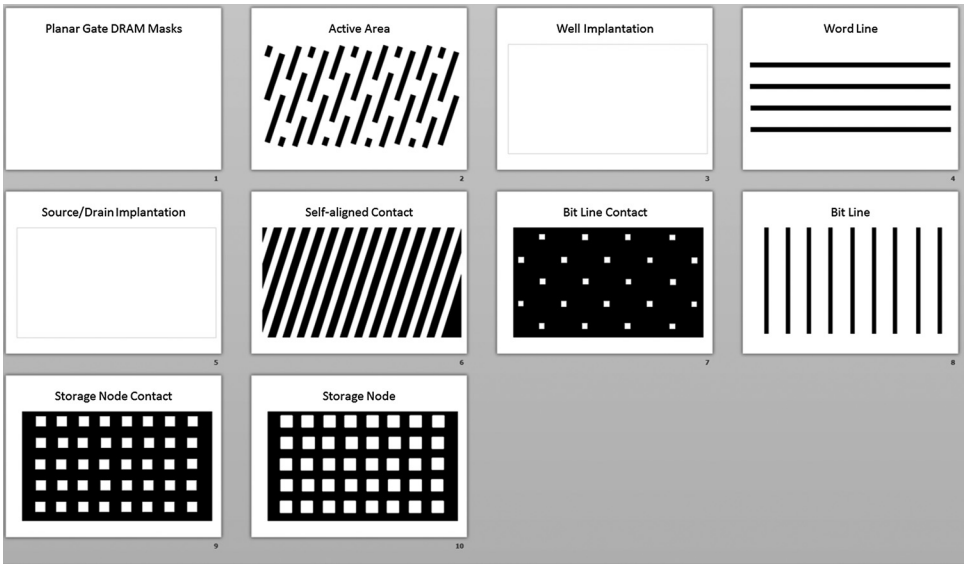


Figure 1.12 Masks for RG transistor DRAM in the array area.

1.2.3 BWL DRAM manufacturing processes

This section discusses a generic manufacturing process of 2x-nm BWL DRAM technology. Although all DRAM manufacturers use BWL technology, each of them has a different design layout and different fabrication-process

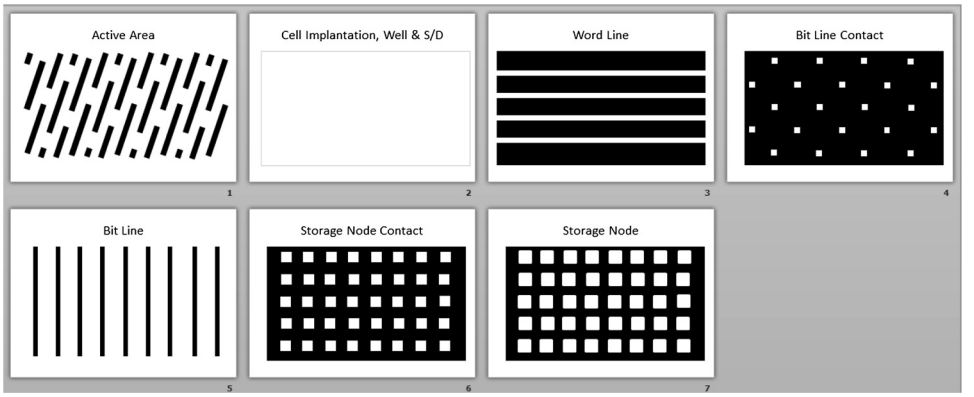


Figure 1.13 Masks for BWL DRAM in the array area.

Figure 1.14 AA (a) mask 1 and (b) mask 2.

steps. Even for the same company, 3x-nm and 2x-nm BWL DRAM processes could be quite different.

Just like CMOS IC chip manufacturing, BWL DRAM also starts with wafer cleaning. A pad oxide layer is then grown in a thermal furnace, and silicon nitride is deposited, usually with a low-pressure chemical vapor deposition (LPCVD) process. The SiN layer can be used as a hard mask for shallow-trench isolation (STI) etch. It is also used as a STI oxide CMP stop layer. Another hard mask layer is needed for a double-patterning process.

Double patterning is required for a 2x-nm DRAM, and the AA can be defined with two masks. Figure 1.14 shows the two masks of AA in a so-called "litho–etch–litho–etch" (LELE) double-patterning process. We only illustrate the masks in array area. Figure 1.15(a) illustrates the final AA pattern that is formed by combining AA mask 1 and AA mask 2. Overlay of the two masks becomes very critical. For example, the overlay shift in the y direction can affect the overlay budget of the contact layers later. The effects of the overlay shift in the x direction are shown in

Figure 1.15 (a) The final AA pattern and (b) AA pattern with the AA2-to-AA1 overlay shifted in the *x* direction.

Fig. 1.15(b); they cause alternating wider and narrower STI gaps between AAs. The etch profile of narrower gaps is different from that of the wider gaps. The narrow gaps could have trouble performing void-free gap filling during oxide CVD. The gap aspect ratio and quality of the oxide-filled inside will affect the BWL trench etch process later.

The AA pattern can also be formed with self-aligned double patterning (SADP), which uses spacers on the sidewalls of the mandrel, to double the pitch of the original patterning of the dummy layer. Figure 1.16(a) shows the mask of the mandrel pattern. Figure 1.16(b) illustrates the mandrel formed by etching the dummy layer with the Fig. 1.16(a) pattern. Figure 1.16(c) shows the mandrel with a sidewall spacer, and Fig. 1.16(d) illustrates the spacer pattern after the mandrel is removed, which doubles the pitch density of the mandrel pattern. The cut mask and the final AA pattern are illustrated in Figs. 1.16(e) and 1.16(f), respectively.

The AA SADP process starts with pad oxidation, a nitride hard mask, and dummy-mandrel film deposition. The mandrel patterns are formed after patterning and dummy film etch, as shown in Fig. 1.16(b). After photoresist (PR) strip and clean, a conformal film is deposited on the wafer surface, and a vertical etch back is performed to remove the film from the top of the dummy pattern and the bottom of the gap between the dummy patterns to form spacers on the sidewall of the mandrel, as shown in Fig. 1.16(c). After a highly selective etch process that removes the mandrel, the spacer pattern on top of the nitride hard mask, as shown in Fig. 1.16(d), can be used to etch the SiN hard mask. The cut mask illustrated in Fig. 1.16(e) can be used to cut the line-space hard mask pattern and form the final AA pattern on the SiN hard mask. This SiN hard mask can be used to etch the pad oxide and single-crystalline silicon substrate to form the final AA pattern.

In comparison with LELE double patterning, SADP requires significantly more process steps and thus has a higher cost. However, it has better CD control and less line-edge roughness. It also significantly reduces the requirement of a second mask overlaying the first mask and thus will not

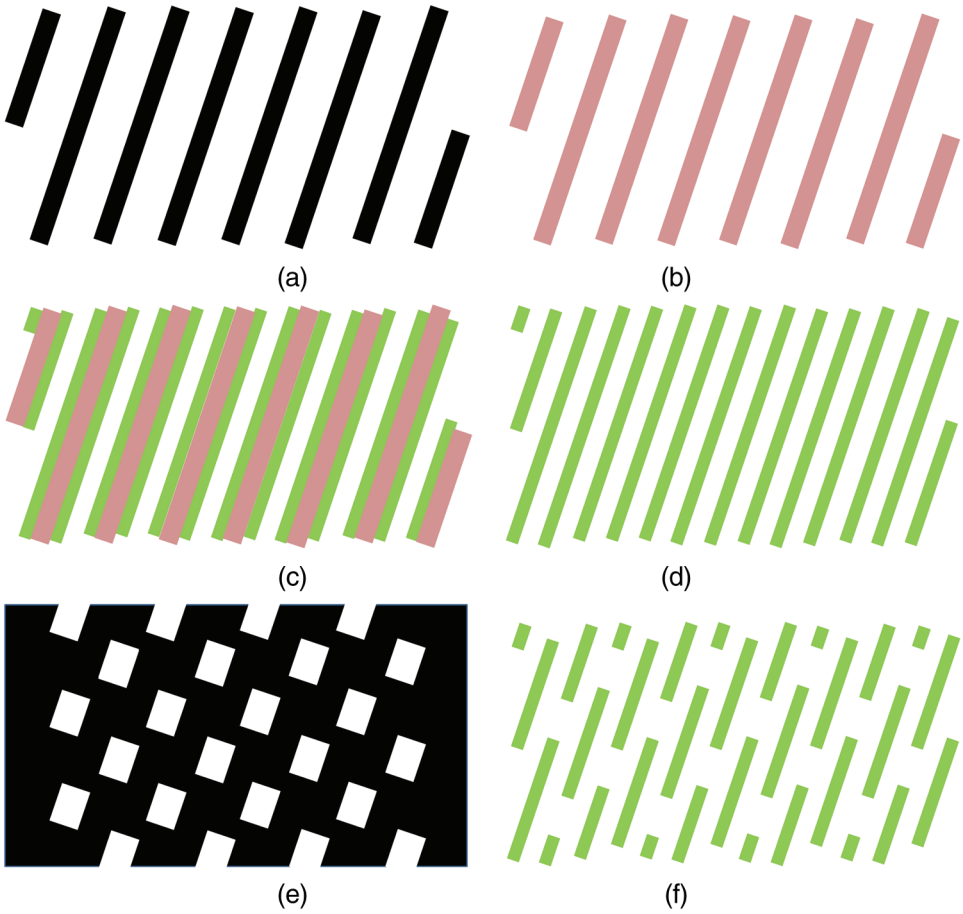


Figure 1.16 AA pattern formation with SADP: (a) mandrel pattern mask, (b) mandrel pattern, (c) spacers formed on mandrel pattern, (d) spacer pattern after dummy-pattern removal, (e) cut mask, and (f) the final AA pattern.

have the STI oxide void issue caused by the overlay error-induced gap narrowing displayed in Fig. 1.15(b).

The patterned SiN hard mask is then used to etch trenches into silicon to form STI. After STI etch, clean, metrology, inspection, and review, a thin layer of silicon dioxide is thermally grown on the silicon surface in an oxidation furnace, and an oxide layer is deposited to fill the high-aspect-ratio (HAR) trenches. Void-free trench fill is very critical at this step because any void between AAs will cause BWL trench etch-profile issues and kill the access transistor in the array. Chemical mechanical polish (CMP) of silicon oxide is performed, and the CMP process stops on the SiN hard mask layer. Hot-phosphoric-acid wet etch is commonly used to strip the SiN hard mask, and diluted hydrofluoric acid (HF) is commonly used to strip the pad oxide.

Table 1.1 AA module.

Wafer clean [Fig. 1.17(a)]	PR strip and clean [Fig. 1.17(d)]
Pad oxidation	Etch pad oxide
SiN hard mask deposition	Etch silicon trench
Amorphous silicon hard mask deposition [Fig. 1.17(b)]	Wafer clean [Fig. 1.17(e)]
AA mask 1	Oxidation
Etch top hard mask	STI oxide CVD to fill the trench [Fig. 1.17(f)]
PR strip/clean [Fig. 1.17(c)]	STI oxide CMP, stop on nitride [Fig. 1.17(g)]
AA mask 2	Wet strip nitride and pad oxide
Etch bottom hard mask	Wafer clean [Fig. 1.17(h)]

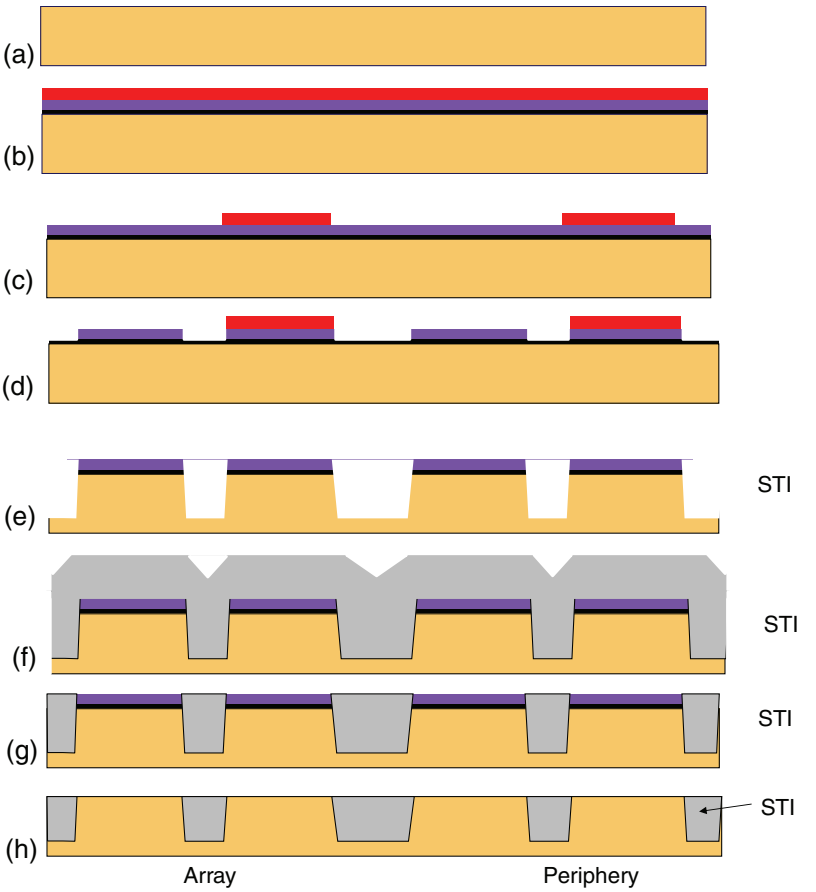


Figure 1.17 The AA process module from (a) start to (b) finish.

The AA-formation process steps are summarized in Table 1.1. Figure 1.17 illustrates the AA module process steps with LELE double-patterning techniques. The cross-section of the beginning of the wafer process and the end of the AA module are illustrated in Figs. 1.17(a) and 1.17(h), respectively.

Table 1.2 Well formation.

Sacrificial oxide growth	PR removal and clean [Fig. 1.18(b)]
Cell p-well mask	**Peripheral p-well mask**
P-well implantation	P-well implantation
N+ S/D implantation	NMOS VT adjust implantation
PR removal and clean [Fig. 1.18(a)]	Photoresist removal and clean
Peripheral n-well mask	Sacrificial oxide removal and clean
N-well implantation	Rapid thermal anneal [Fig. 1.18(c)]
PMOS VT adjust implantation	

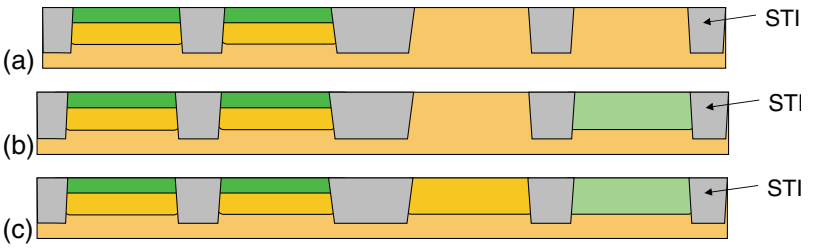


Figure 1.18 Cross-section view of well formation.

There are two mask steps in Table 1.1 because this AA process is a LELE double-patterning process. Here, a mask step indicates a photolithography process that includes multiple steps, such as pre-bake, primer coating, wafer cooling, PR coating, soft bake, wafer cooling, alignment and exposure, post-bake, wafer-edge exposure, wafer cooling, development, hard bake, wafer cooling, metrology, and inspection. If metrology and inspection find that the process is out of specification, then the wafer can be reworked by stripping PR/clean and then going through the entire mask procedure again.

After wafer clean, a thin layer of sacrificial oxide is grown, p-well masks are applied, and high-energy ion implantations are used to form the p-well for access NMOS in array areas and both p-well and n-well in the peripheral area. Well-implantation photolithography processes are not critical mask layers; they usually have large CDs, especially for access NMOS in the array area. After photoresist ashing and clean, the wafer is annealed. The process steps are listed in Table 1.2, and cross-section views of this stage are shown in Fig. 1.18.

Figure 1.18 shows the cross-section of the BWL DRAM well and channel ion-implantation processes. Figure 1.18(a) shows the cross-section illustration after ion implantations in the array area, which form both well junction and S/D junction of the access transistor of BWL DRAM. The well junctions are formed by high-energy p-type ion implantation. The access NMOS S/D junction is formed by high-current, low-energy n-type ion implantation. Figure 1.18(b) shows the cross-section view after n-well ion implantation in the peripheral area, and Fig. 1.18(c) illustrates the cross-section view after

Figure 1.19 Buried WL mask.

p-well ion implantation. All access transistors in the array area are NMOS; therefore, the array area only has p-well. There are both n-well and p-well in the peripheral area. Because peripheral transistors are planar transistors, their S/D implantations must wait until after gate formation.

After well formation, the next step forms the access transistors. For a planar access transistor, the WL mask shown in image 3 of Fig. 1.11 is used to pattern PR, which is then used to etch a line-space pattern on a SiN hard mask, tungsten silicide, and polysilicon stack on the gate oxide. The RG transistor needs two masks, the first of which is a mask in the array area, i.e., a reversed WL mask similar to image 3 in Fig. 1.12, before the gate oxide growth and WL film stack deposition. It only etches silicon in the AA with little loss of silicon oxide in the STI area. After PR strip and wafer clean, gate oxidation, and gate/WL stack deposition, a second WL mask is used to etch line-space patterns of the gate and WL. The BWL cell transistor formation is the most unique process of this type of DRAM, compared to the other two types of access transistor formation of previous DRAM generations. The process requires the BWL mask shown in Fig. 1.19 to etch trenches on both single-crystal silicon and silicon oxide simultaneously. This etch process requires good control of the etch rate and etch-rate uniformity in both materials and good control of the silicon profiles inside trenches. The access-transistor formation processes include

- BWL trench etch,
- clean,
- gate oxidation,
- TiN gate deposition,
- W CVD,
- W etch back,
- oxide CVD, and
- oxide CMP.

Figure 1.20 overlaps the BWL mask with the AA pattern. Each AA has two word lines pass through it to form two cell transistors. The middle section

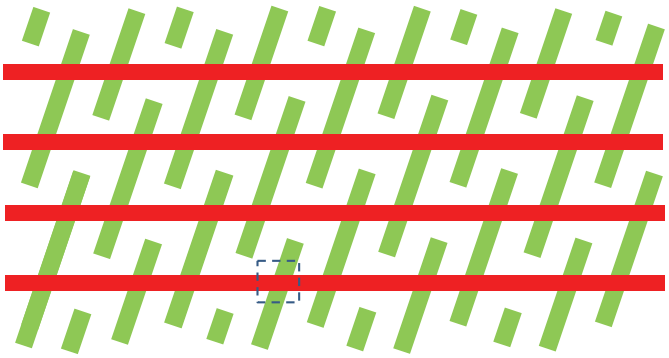


Figure 1.20 Overlap of BLW mask and AA patterns.

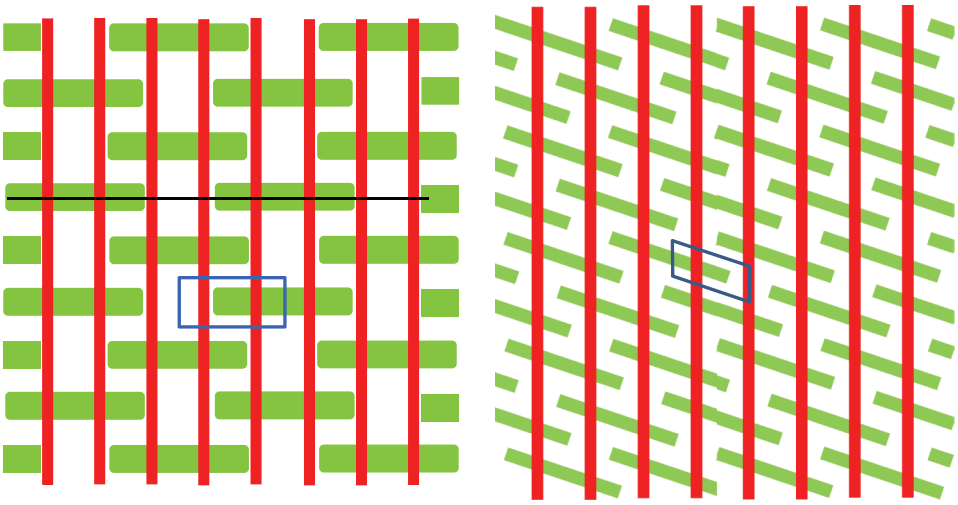


Figure 1.21 (a) $8F^2$ DRAM layout and (b) $6F^2$ DRAM layout.

is the shared S/D, which will connect to the bit line (BL) and the sections at the two ends of the AA will connect to SN capacitors.

Figure 1.21(a) illustrates a DRAM layout that was widely used during planar transistor and RG transistor DRAM eras. It has unit array area of $8F^2$, shown as the rectangular box. Here, F is the half pitch of the densest cell pattern. Figure 1.21(b) is the same layout as Fig. 1.20 with a rotation of 90 deg. It is a DRAM layout with unit array area of $6F^2$, shown in the parallelogram box. This layout obviously has a higher cell density than the $8F^2$ layout and thus has been more widely used in DRAM chip manufacturing in recent years. The cross-section of the BWL DRAM process in this section follows the dashed line in Fig. 1.21(a). Table 1.3 lists the process steps of the BWL module that forms access NMOS and WL in the array area.

Table 1.3 BWL module.

Wafer clean	W and TiN etch back [Fig. 1.23]
Oxidation	Wafer clean
Hard-mask deposition	Oxide deposition
BWL mask [Fig. 1.19]	Oxide CMP
Etch hard mask	Strip hard-mask
BWL trench etch	Oxide deposition [Fig. 1.24]
PR strip & clean [Fig. 1.22]	TiN gate electrode deposition
Oxidation	W deposition

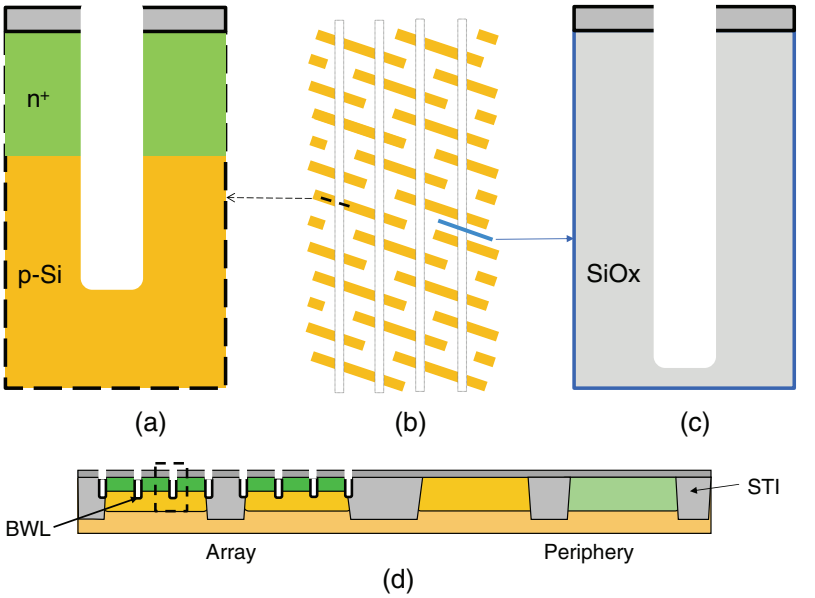


Figure 1.22 Zoomed-in cross-section of (a) BWL etch in AA, (b) a BWL DRAM layout with BWL on top of AA layer, (c) zoomed-in BWL etch profile in STI oxide, and (d) illustration of cross-section of BWL DRAM after BWL trench etch.

Figure 1.22(b) is a layout of BWL DRAM with BWL layer overlaps on top of the AA layer. Figure 1.22(a) is the zoomed-in detail of the dashed box in Fig. 1.22(d), which shows the BWL trench in AA silicon in the array area but not in the peripheral area. Figure 1.22(c) shows a close-up of the BWL trench-etch profile in STI silicon oxide. The etch rate in STI oxide is higher than the etch rate in AA silicon, which helps create silicon fins inside the BWL trenches, as shown in Fig. 1.7(c).

Figure 1.23(d) shows the cross-section of the BWL DRAM after TiN and W deposit into the BWL trenches, and an etch-back process removes W and TiN from the wafer surface, leaving W only inside the trenches from the WL.

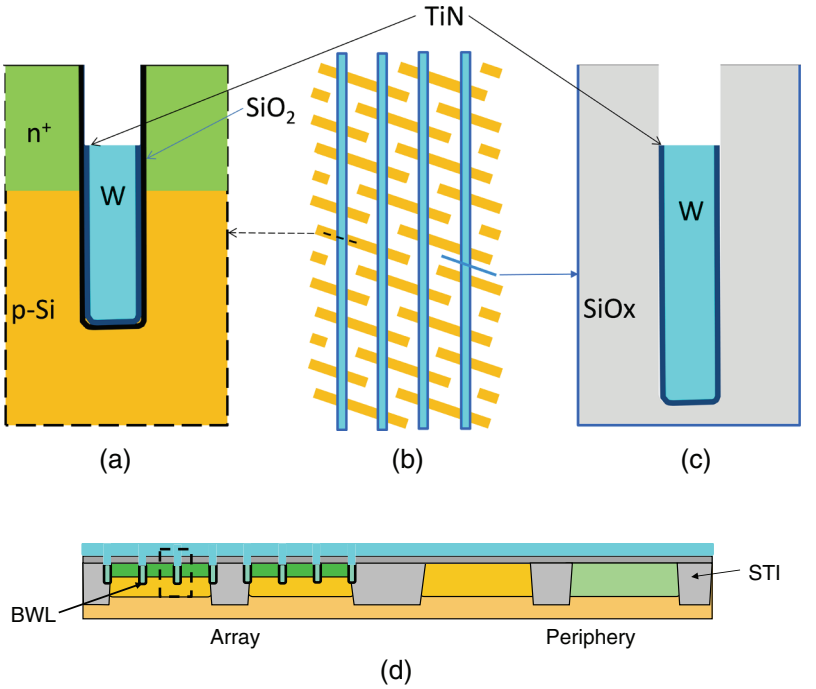


Figure 1.23 (a) Zoomed-in cross-section of a (a) BWL cell transistor, (b) BWL DRAM layout with BWL overlaps the AA, (c) zoomed-in BWL profile in the STI oxide, and (d) illustration of cross-section BWL DRAM after W and TiN etch back.

Figure 1.23(a) shows the detail of the cross-section of the cell transistor, indicated in the dashed box in Fig. 1.23(d). Figure 1.23(b) shows the BWL DRAM layout with a BWL layer on top of the AA layer. The dashed line indicates the cross-section line in the AA silicon, and the solid line indicates the cross-section line in the STI oxide, shown in Fig. 1.23(c).

Figure 1.24(d) shows the cross-section of BWL DRAM after ILD0 deposition. Figure 1.24(a) is the zoomed-in access transistor shown in the dashed box of Fig. 1.24(d). The TiN forms the gate electrode of the DRAM access NMOS, and the W in the trenches forms the WL that is buried underneath the wafer surface.

Figure 1.25 shows the cross-section along the word line, illustrated in Fig. 1.24(b). By allowing a higher oxide etch rate in STI than the silicon etch rate in the AA during BWL trench etch, a device structure is created that the TiN gate wraps around the silicon channel with gate oxide in between along three sides, which forms a device similar to a tri-gate FinFET for the access transistor. This process can further reduce the off-state leakage current of the access transistor while increasing the drive current in its on-state.

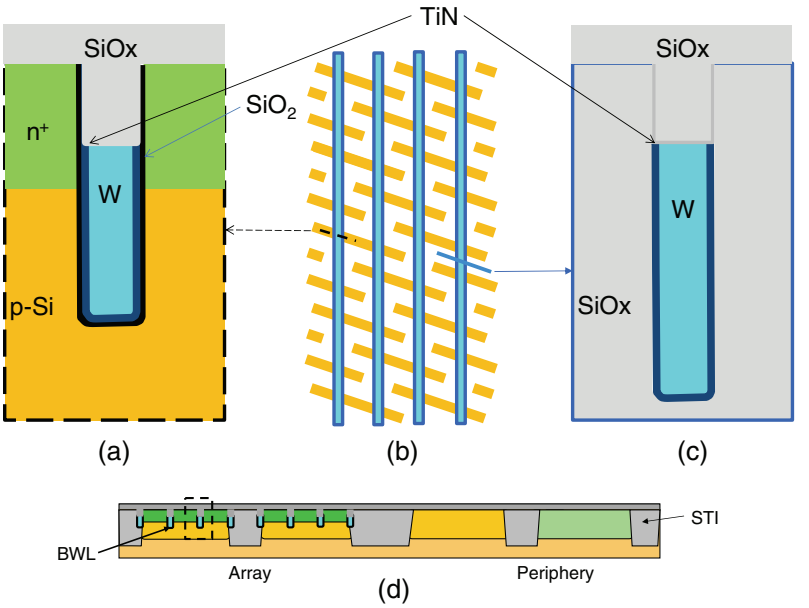


Figure 1.24 (a) Zoomed-in cross-section of cell transistor after ILD0, (b) a BWL DRAM layout with BWL on top of the AA, (c) zoomed-in cross-section in the STI oxide, and (d) illustration of cross-section BWL DRAM after W and TiN etch back.

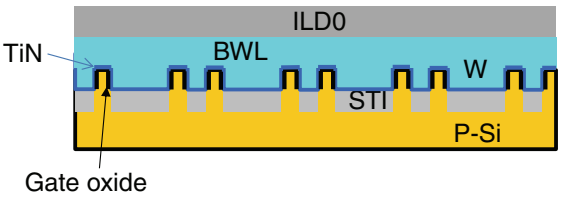


Figure 1.25 Cross-section along the word line of BWL DRAM.

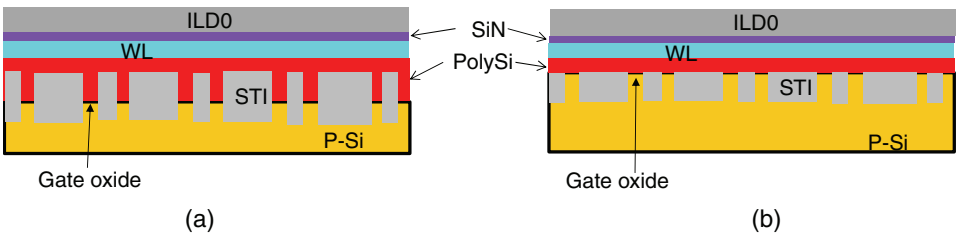


Figure 1.26 (a) Cross-section along the word line of DRAM with a RG cell transistor and (b) DRAM with planar-cell-transistor DRAM.

Figure 1.26(a) is a cross-section along the WL of a DRAM with RG access transistor, and Fig. 1.26(b) shows the same kind of cross-section of a DRAM with a planar cell transistor. These two types of transistors have the same gate width at the same feature size, whereas the BWL cell transistor has

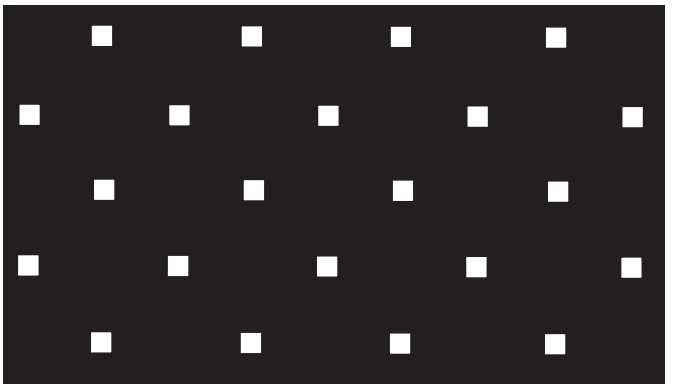


Figure 1.27 Bit-line contact mask.

the longest gate width because of its tri-gate structure. The gate electrodes of planar and RG cell transistors are polysilicon, and the WL stacks comprise a SiN hard mask, tungsten silicide, and polysilicon.

After cell transistor formation, a non-critical mask with a large CD is used to remove oxide in the peripheral area. After wafer clean, thermal oxidation and remote plasma nitridation are performed to form gate dielectric (nitridized silicon oxide, or SiON) with an enhanced k value ($\sim$5) for the peripheral CMOS devices. A heavily n-type-doped polysilicon is then deposited. A poly-dope mask that exposes the peripheral PMOS is applied. An extremely-high-dosage p-type ion implantation is performed to counter-dope peripheral PMOS polysilicon and convert it from heavily n-type to heavily p-type, which is needed for PMOS threshold-voltage control. After photoresist strip and clean, a mask is applied, and an etch process removes polysilicon in the array area. Photoresist removal and clean prepares the wafer for the bit-line contact (BLC) mask, shown in Fig. 1.27.

Figure 1.27 shows the BLC mask, which forms a contact between the BL and the S/D of the DRAM access NMOS. Figure 1.28 overlaps the BLC

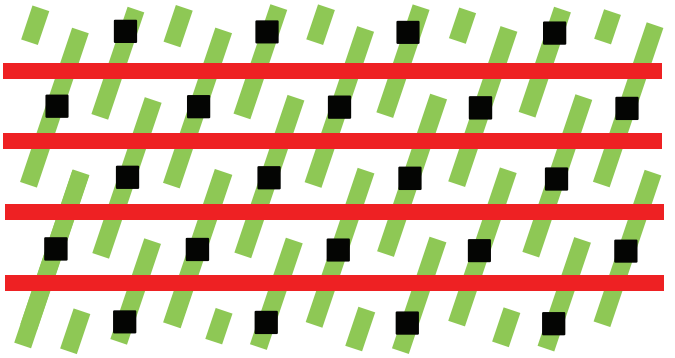


Figure 1.28 BLC mask overlaps with AA and BWL patterns.

Table 1.4 BLC module.

Wafer clean	Array area mask
Peripheral mask	Etch polysilicon
Etch oxide	PR strip/clean [Fig. 1.29(c)]
PR strip/clean [Fig. 1.29(a)]	BLC mask [Fig. 1.27]
Gate oxidation and nitridation	Etch oxide
Polysilicon deposition [Fig. 1.29(b)]	PR strip clean [Fig. 1.29(d)]
Wafer clean	Polysilicon deposition
PMOS poly-dope mask	TiN and W deposition
PMOS poly implantation	SiN deposition [Fig. 1.29(e)]
PR strip/clean	

contact mask on the AA and BWL patterns. Note that BLC always connects the S/D in the middle section of the three sections of the AA pattern. The BLC mask has the lowest hole pattern density compared to other hole pattern masks in the DRAM array area. Because the oxide film covering the AA is very thin (usually < 30 nm) and contact hole aspect ratio is very low ($\sim$1:2), the BLC etch process is not as much a challenge as other HAR contact-hole etch processes.

There is an alternative method to pattern BLC in a BWL DRAM array area. After polysilicon deposition and poly-dope implantation, the BLC mask is applied on top of the polysilicon in the array area, and BLC holes are etched through the polysilicon and silicon oxide. After wafer clean and native oxide removal, TiN/W/SiN stack is deposited, and the stack is patterned into BL and peripheral gate patterns with a BL mask.[10] The BLC process steps are listed in Table 1.4.

Figure 1.29 shows the BLC process module. Because the film stack that forms the BL in the array area is also used to form a gate stack in the peripheral CMOS, there are several process steps of film deposition and removal in the array area and peripheral area. Figure 1.29(a) shows the cross-section after peripheral oxide removal; Fig. 1.29(b) shows the cross-section after peripheral gate oxidation and polysilicon deposition; Fig. 1.29(c) shows the removal of polysilicon in the array area; Fig. 1.29(d) illustrates the cross-section of BWL DRAM after BLC etch, PR strip, and clean; and Fig. 1.29(e) shows the cross-section of the BLC holes that are filled with polysilicon and covered by TiN, W, and SiN.

After BLC formation and BL film-stack deposition, the next process is the formation of the BL and peripheral transistors. Figure 1.30 shows the BL mask, and in Fig. 1.31 the BL mask is overlapped with BLC, BWL, and AA patterns. In this BWL DRAM layout, the BL is perpendicular to the WL, and it aligns with the BLC on the middle section of the AA between two WLs. In BWL DRAM, the BL in the array area and gate in the periphery share the same film stack. It usually consists of several layers, such as polysilicon, TiN, W, or tungsten silicide (WSi_x) and SiN. Polysilicon is the gate electrode of the

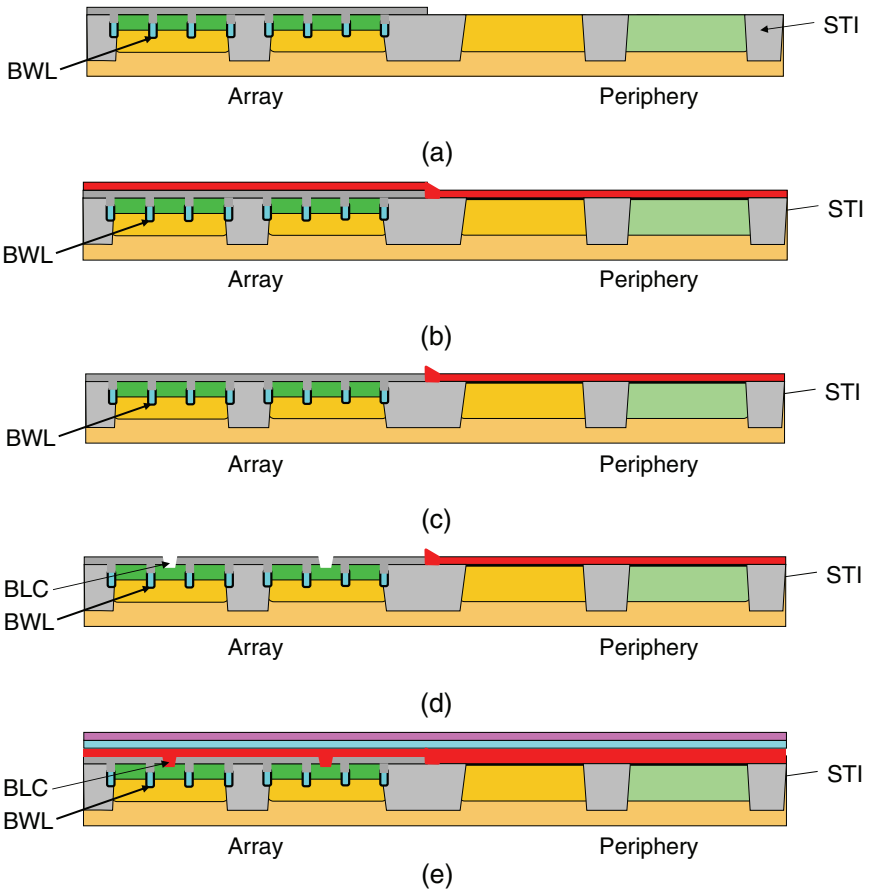


Figure 1.29 BLC processes: (a) remove oxide in the periphery, (b) grow gate oxide and deposit polysilicon, (c) remove polysilicon in the array area, (d) etch oxide and PR strip/clean, and (e) polysilicon, TiN, W, and SiN deposition.

Figure 1.30 BL mask.

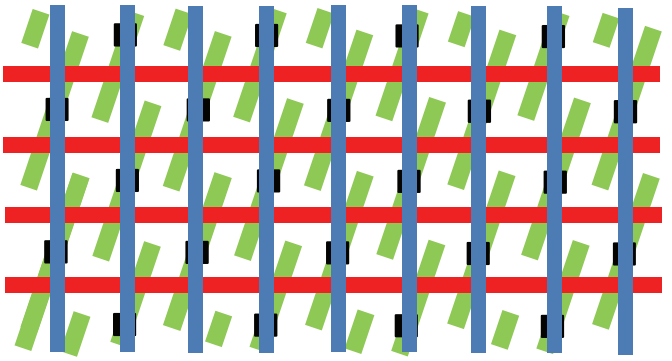


Figure 1.31 BL mask overlaps with AA, BWL, and BLC patterns.

MOSFETs in the peripheral area that forms circuits such as address decoders, sense amplifiers, and multiplexers. TiN serves barrier layer and glue layer for W or WSi_x. Either W or WSi_x is the conducting layer used to reduce the resistance of the BLs and local interconnects of the peripheral circuits. The SiN layer is usually used as a HM layer for patterning the BL and gates of the peripheral transistors.

One of significant challenges of BL patterning is overlay. Because the W or WSi_x layer is opaque to the light, it is very hard to optically measure the overlay between the PR BL pattern on the top hard mark above the W layer and the BLC pattern underneath W. After BL pattern etch, photoresist strip, and clean, the BL-to-BLC overlay can be measured by after-clean inspection (ACI). A re-oxidation process is usually used to repair etch-induced gate oxide damage and help to reduce gate leakage. There are then multiple masks in the peripheral area before the next mask in the array area. There are two masks for S/D extension (SDE) ion implantations of peripheral NMOS and PMOS. A conformal dielectric (usually SiN) CVD and a dielectric vertical etch back form spacers on the sidewall of the gates in the periphery and BLs in the array area.

The sidewall spacers and the SiN on top of the BL stack help to prevent the storage node contact (SNC) plugs shorting to the BL. They have been widely used to form SNC similar to self-aligned contact (SAC) in earlier-generation DRAM devices. Figure 1.32 illustrates the process to form a spacer on a sidewall of the BL and how it helps to form the SAC. Figure 1.32(a) shows the BL stacks, and Fig. 1.32(b) illustrates the conformal dielectric film (usually LPCVD silicon nitride) deposition. Figure 1.32(c) shows the nitride etch-back process that forms the sidewall spacer. Figure 1.32(d) shows the ILD oxide deposition and CMP; and Fig. 1.32(e) demonstrates the contact etch that is self-aligned between the sidewall spacer due to the ILD etch process, which is highly selective to oxide and etches very little on nitride.

After spacer formation, another two masks are used for high-current ion implantation to form the heavily doped S/D of the peripheral NMOS and PMOS. The peripheral devices are formed after rapid thermal anneal (RTA)

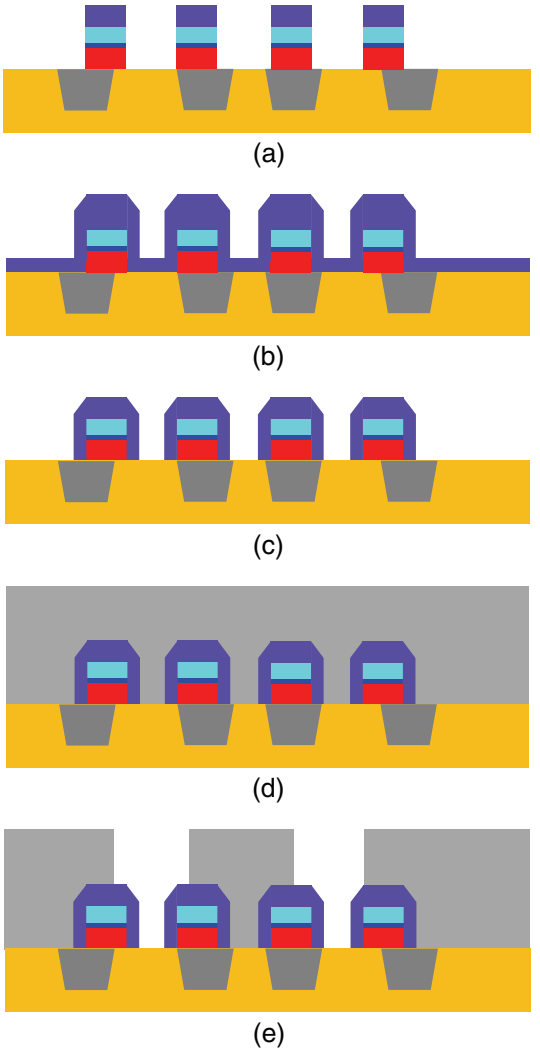


Figure 1.32 (a)–(c) Sidewall spacer formation, and (d)–(e) self-aligned contact formation.

activates the dopants. Table 1.5 lists the process steps of the cell BL and peripheral transistor formations, and Fig. 1.33 illustrates some of the steps.

After the BL and peripheral devices are built, the next mask layer in the array area is the SNC, which creates the contact plugs between the storage capacitors and the AA. At first, ILD1 oxide is deposited. The ILD1 film needs to fill the gap between the BL without voids or buried keyholes. Otherwise, after SNC etch, the SNC holes could be connected by the keyholes underneath the ILD1, and after TiN/W CVD, SNC plugs can be shorted by the metal deposited in the keyholes. After ILD1 deposition and CMP, an etch stop layer is deposited, and the SNC mask illustrated in Fig. 1.34 is applied. For a 2x-nm

Table 1.5 BL and peripheral transistor module.

Wafer clean	PR strip/clean [Fig. 1.33(b)]
BL mask [Fig. 1.30]	Spacer dielectric film deposition
BL and peripheral gate etch	Spacer film etch back [Fig. 1.33(c)]
PR strip/clean/ACI [Fig. 1.33(a)]	Peripheral NMOS SD mask
Re-oxidation	Peripheral NMOS SD implantation
Peripheral NMOS SDE mask	PR strip/clean
Peripheral NMOS SDE implantation	Peripheral PMOS SD mask
PR strip/clean	Peripheral PMOS SD implantation
Peripheral PMOS SDE mask	PR strip/clean [Fig. 1.33(d)]
Peripheral PMOS SDE implantation	RTA

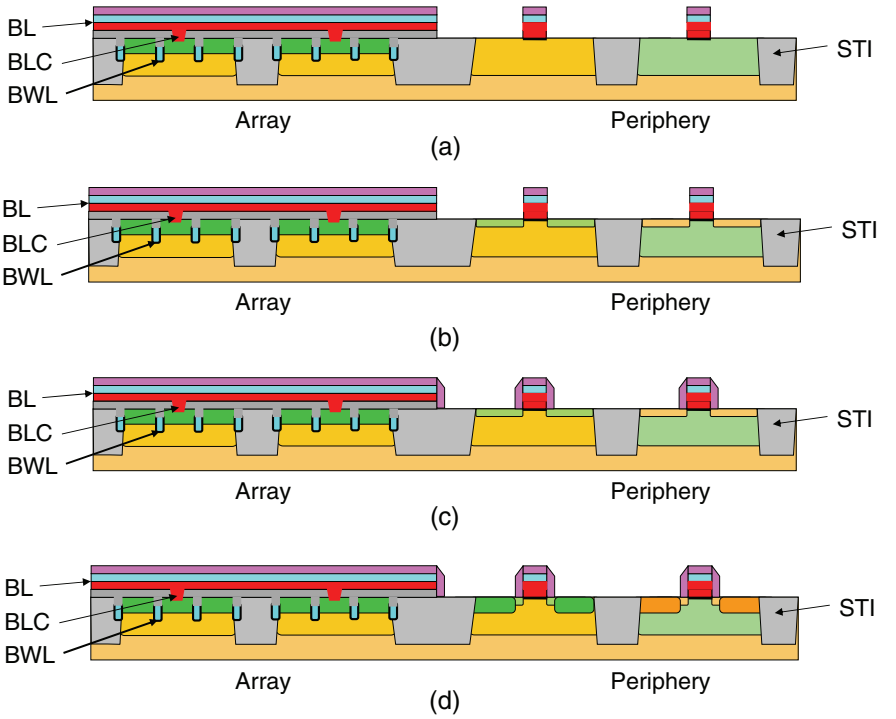


Figure 1.33 Selected process steps of BL and peripheral transistor formation.

node, the pattern density is too high for single-photolithography patterning; two masks, each with a hole pitch relaxed by a factor of $\sqrt{2}$ [as shown in Figs. 1.35(a) and 1.35(b)], are needed to perform LELE double patterning to form the SNC array.

An alternative double-patterning technique that can form this kind of array hole pattern uses two masks, as shown in Fig. 1.36. In this case, two hard mask (HM) layers are needed. The first mask shown in Fig. 1.36(a) is used to pattern the top HM layer. The etch process forms the line-space pattern in the top HM layer that stops at the bottom HM, as shown in

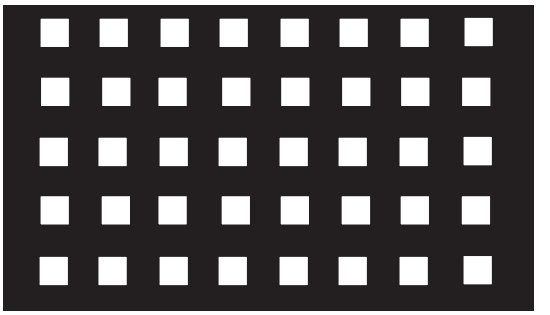
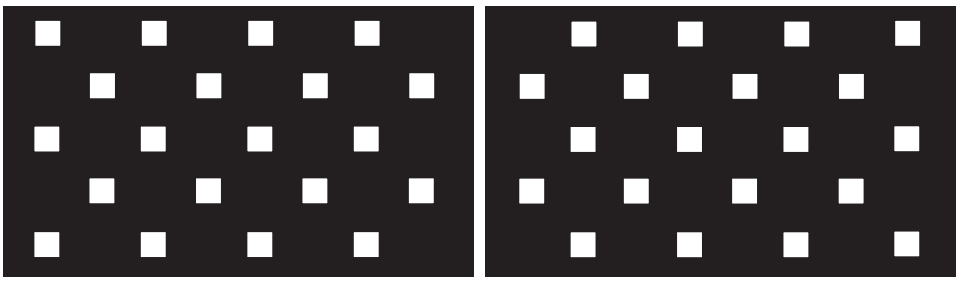
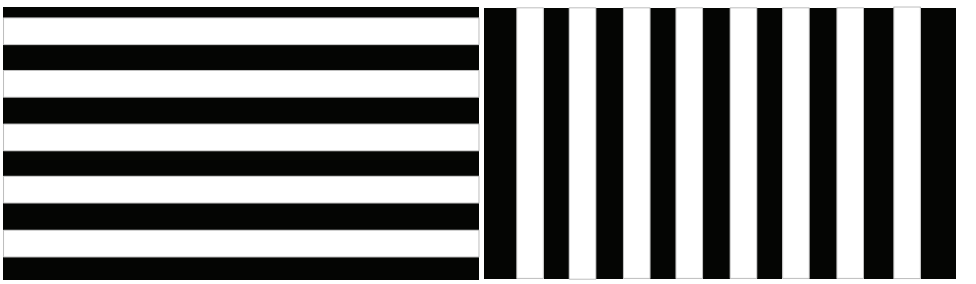


Figure 1.34 SNC mask.



(a) (b)

Figure 1.35 Two masks that relax the SNC pitch in Fig. 1.34.



(a) (b)

Figure 1.36 Two masks that can be used to form the SNC pattern shown in Fig. 1.22.

Fig. 1.37(a). After photoresist strip, clean, and another PR coating and baking, the second mask shown in Fig. 1.36(b) is applied. The second etch process is highly selective to primarily etch the bottom HM while etching very little of the top HM and the ILD1 layer; thus the bottom HM is only removed at the cross-points of the two masks to expose the ILD1 layer underneath, as shown in Fig. 1.37(b). The combined HM patterns can then be used to form the SNC hole pattern, illustrated in Fig. 1.37(c). Due to the corner-rounding effect, the final hole shape is more round than square, as shown in the

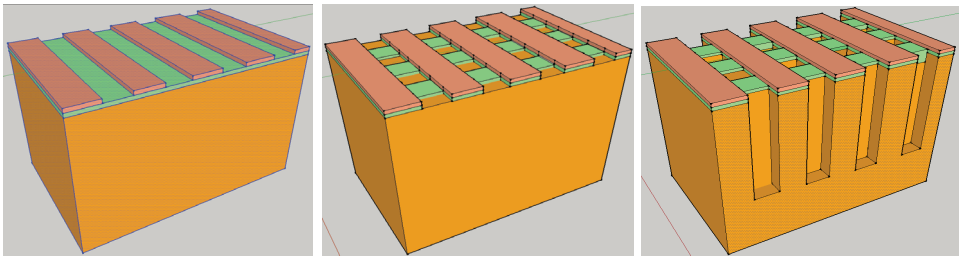


Figure 1.37 (a) Patterning of the top HM layer with the mask shown in Fig. 1.24(a). (b) Patterning the bottom HM with the mask shown in Fig. 1.24(b). (c) The combined HM is used to etch the array hole pattern.

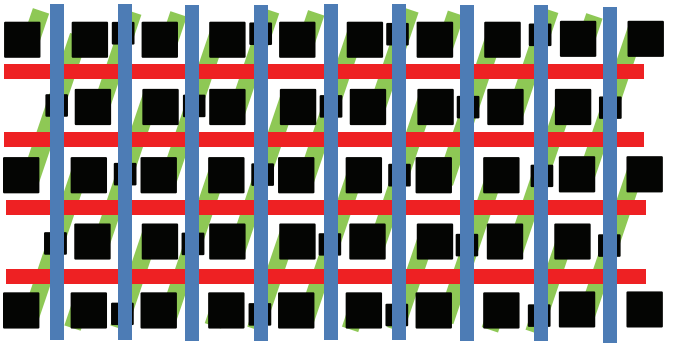


Figure 1.38 Overlapping the SNC mask with BL, BLC, BWL, and AA patterns.

Fig. 1.37(c). One of the advantages of the two-line-space mask for the SNC hole pattern is that for the first mask [Fig. 1.36(a)] only the *y*-direction overlay control is critical; for the second mask [Fig. 1.36(b)], only the *x*-direction overlay control is critical. In contrast, both of the masks shown in Fig. 1.35 need good overlay control in the *x* and *y* directions.

Figure 1.38 shows the overlapping of a SNC mask with BL, BLC, BWL, and AA patterns. The SNC holes go through the gaps between BLs and connect the two ends of the AA patterns. Dielectric spacers and cap layers on the BL allow the BLC hole to be etched in a self-aligned fashion. A thin, conformal SiN film is deposited (usually after BLC etch and clean), and an etch process removes the SiN film from the bottom of the SNC holes and wafer surface. This process is very similar to the sidewall-spacer-formation process, and the SiN film on the sidewall of the SNC hole can provide extra protection to prevent SNC contact plugs from shorting to the BL.

For some devices the same two masks can be used to form contact holes in the peripheral area. After Ti, TiN, and W deposition and CMP, the SNC module in the array area is finished. Another dielectric layer, ILD2, is deposited, and the metal 1 (M1) mask is needed to form the first

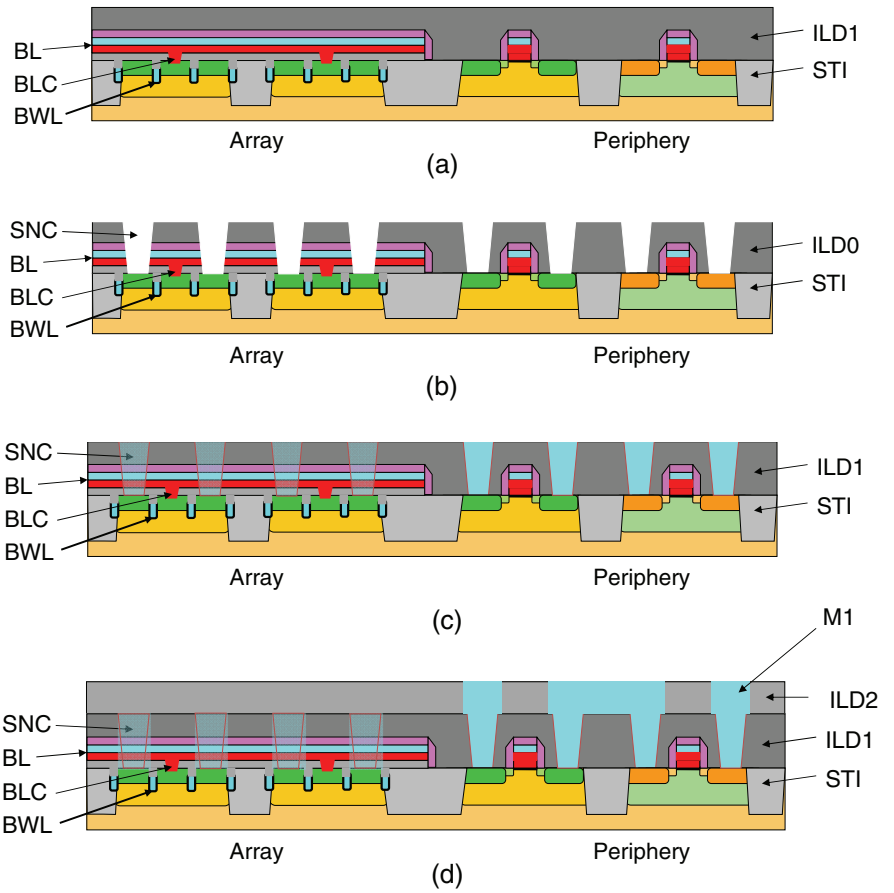


Figure 1.39 SNC, peripheral contact, and M1 process steps: (a) after ILD2 CMP, (b) after ID2 etch and clean, (c) after SNC W/TiN/Ti CMP, and (d) after M1 W/TiN/Ti CMP.

Table 1.6 SNC, peripheral contact, and M1 process steps.

ILD1 deposition	Wafer clean [Fig. 1.39(b)]
LD1 CMP [Fig. 1.39(a)]	Ti/TiN/W deposition
HM deposition	W/TiN/Ti CMP [Fig. 1.39(c)]
SNC mask 1 [Fig. 1.35(a)]	Etch stop layer deposition
Etch HM	ILD2 deposition
PR strip/clean	M1 mask
SNC mask 2 [Fig. 1.35(b)]	Etch ILD2
Etch HM	PR strip/clean
PR strip/clean	Ti/TiN/W deposition
Etch ILD1	W/TiN/Ti CMP [Fig. 1.39(d)]

metal-interconnect layer in the peripheral area. After etch, PR strip/clean, and Ti/TiN/W deposition and CMP, the wafer is ready for the next process module. Table 1.6 lists the process steps of SNC, peripheral contact, and M1, which are also illustrated in Fig. 1.39.

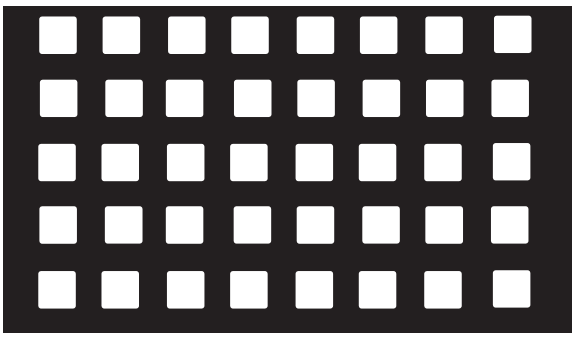


Figure 1.40 SN mask.

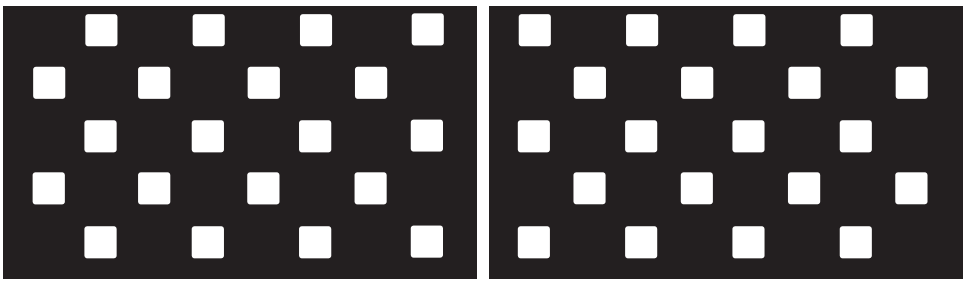


Figure 1.41 Two masks that relax the SN pitch in Fig. 1.40.

The next step deposits an etch stop layer (usually silicon nitride), a thick dielectric layer, ~1.5–2.0-μm silicon oxide, and a cap layer (SiN). A SN mask (Fig. 1.40), which is almost identical to the SNC mask in the array area, is applied. For the SNC layer, contact hole etch, metal deposition, and metal CMP are performed in both the array area and peripheral area. For the SN layer, all of the processes are performed only in the array area. Because of the high pattern density in a 2x-nm node, two masks are used to relax the resolution requirement and help pattern the dense SN holes, as illustrated in Fig. 1.41. The two-line-space pattern masks similar to that shown in Fig. 1.36 can also be used to form this SN pattern. In Fig. 1.42, the SN mask is overlapped with SNC, BL, BLC, BWL, and AA. The SN mask is aligned with the SNC mask to allow SN holes to land on the SNC plugs.

SN hole etch is one of the most challenging etch processes in IC manufacturing. The pattern density is high, the CD is small, the hole is very deep, and the aspect ratio is very high (~50:1, possibly up to ~100:1). There are serval defects of interest (DOIs) in this etch process that could affect the product yield, such as

- under-etch,
- bottom residue,

- overlay shift that could cause high contact resistance or even a short between the neighboring SN capacitors via the SNC plugs, and
- a bow-shaped etch profile that can cause a short between the neighboring SN capacitors in the middle of the cylinders, as shown in Fig. 1.43. The SN-hole aspect ratio in the figure is ~10:1, but it is much higher in real devices.

After SN hole etch and clean, a thin TiN layer (~10 nm) is deposited, which serves as the electrode of the SN capacitor that connects to the array transistor. A photoresist layer is applied to the wafer surface to fill the SN

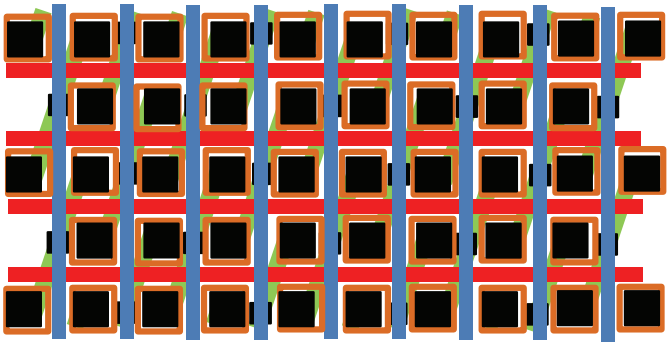


Figure 1.42 Overlapping SN mask with SNC, BL, BLC, BWL, and AA patterns.

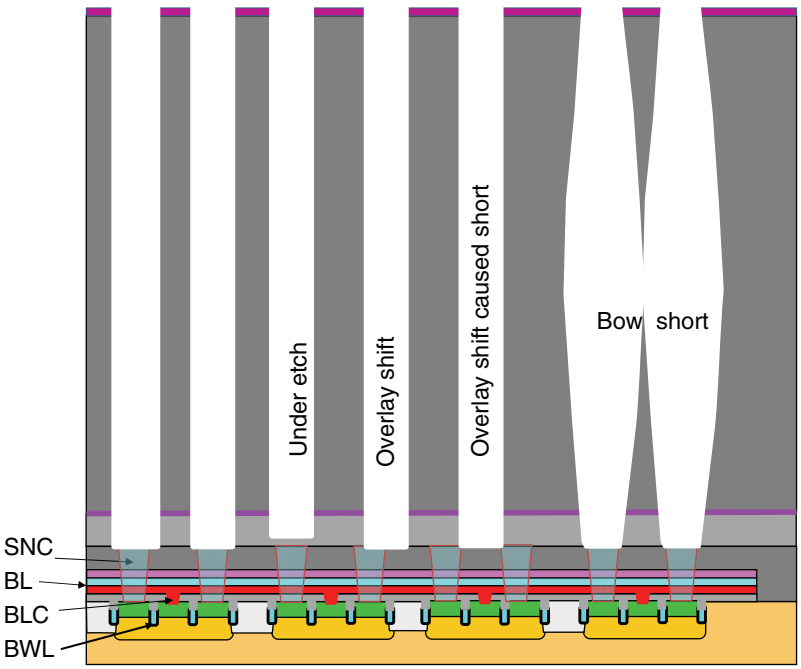


Figure 1.43 SN hole etch and main DOI.

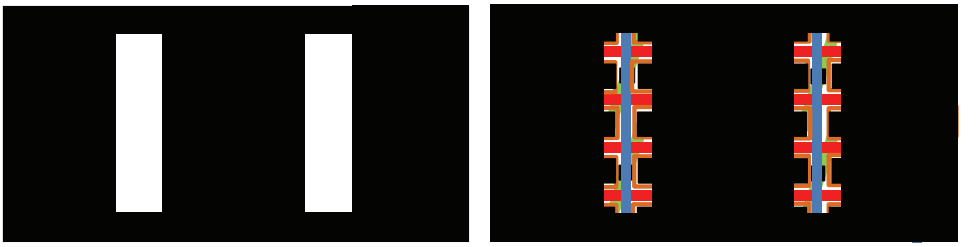


Figure 1.44 (a) SiN slot mask and (b) slot mask overlaps with previous layers.

holes, with spin coating and hard bake. A photoresist etch back is performed to expose the TiN on the top surface of the wafer. An etch process is used to remove the TiN on the wafer surface while TiN on the sidewall of the SN hole is protected by the photoresist. A mask [shown in Fig. 1.44(a)] that allows the removal of SiN from all peripheral areas and part of the array area, called slots, is applied. Figure 1.44(b) shows the SiN slot mask in array area overlapping with SN, SNC, BL, BLC, BWL, and AA layers. The top nitride layer is used to hold the tall TiN cylinders after removing the thick oxide that surrounds the TiN cylinders and prevents the cylinders from collapsing.

After SiN slot etch, hydrofluoric acid (HF) is used to remove the thick ILD silicon oxide to expose the TiN cylinder through the opening in the peripheral area and through the slots etched on the SiN in the array area. The SiN cap layer functions like the plastic ring holder of a six-pack of cans, preventing the TiN from falling down. Without the top nitride layer, the tall TiN cylinders are likely to collapse and short to each other, especially with the surface tension of liquids used in wet etch and clean. After wafer cleaning of conformal, high-k dielectric layers, such as zirconium oxide (ZrO), an aluminum oxide (AlO) stack with a total thickness <10 nm is deposited on both the inside and outside wall of the TiN cylinder. Another thin TiN layer, which is the electrode that connects to the ground (<10 nm), is deposited with full coverage of both the inside and outside of the cylinder to form the SN capacitor. Without an ILD oxide recess, the capacitor can only form inside the SN hole with electrode area $\sim\pi \times CD \times h$. Here, CD is the SN hole CD, and h is the SN hole depth. By recessing the ILD oxide and exposing the outside of the TiN cylinder, the area of the SN capacitor is doubled, which allows scaling down the SN cylinder CD by a factor of 2 without increasing the depth of the SN hole. After conformal deposition of the ground electrode (usually TiN), a conducting layer with good gap-fill capability (in many cases, SiGe) is deposited to fill the remaining SN holes and the gaps between

Table 1.7 SN module.

Wafer clean	TiN deposition
Each stop layer (ESL) deposition	PR coating
ILD deposition	RP etch back
SiN deposition [Fig. 1.45(a)]	TiN etch [Fig. 1.45(c)]
SN mask 1 [Fig. 1.41(a)]	SiN slot mask [Fig. 1.44(a)]
Etch nitride	Nitride etch
PR strip/clean	PR strip/clean
SN mask 2 [Fig. 1.41(b)]	ILD removal [Fig. 1.45(d)]
Etch nitride	Wafer clean
PR strip/clean	High-*k* film deposition
Etch oxide [Fig. 1.45(b)]	TiN and conducting filler deposition [Fig. 1.45(e)]

cylinders, which finishes the SN module. Table 1.7 lists the process steps of the recessed-stack cylinder capacitor.

Figure 1.45(a) shows the cross-section of the BWL DRAM after SN dielectric-layer (etch stop nitride, thick oxide, and nitride cap layer) deposition. Figure 1.45(b) shows the cross-section after SN hole etch, PR strip, and clean. The insert shows the top-down cross-section view of two SN holes in oxide. Figure 1.45(c) illustrates the cross-section after the top TiN is removed. The PR in the SN holes protects the TiN film on the sidewall. The inserted graph shows the hole filled with PR that protects TiN. Figure 1.45(d) shows the cross-section after the thick ILD oxide is removed; the HAR TiN cylinders stands on SNC plugs, holding on the SiN cap layer at the top and the ESL/ILD2 layers at the bottom. The insert shows the TiN cylinders stand on the wafer surface, empty both inside and outside. Some DRAM manufacturers use two top SiN layers—one on the surface and another one ~150 nm below the surface—to hold the thin and tall TiN cylinders. Figure 1.45(e) shows the cross-section after deposition of a high-*k* dielectric layer, TiN, and a conducting filler. The commonly used DRAM SN high-*k* dielectric is a zirconium oxide–aluminum oxide–zirconium oxide (ZrO_2/Al_2O_3/ZrO_2, or ZAZ) stack deposited by atomic-layer-deposition (ALD) processes. The TiN is used as the ground electrode of the SN capacitor, and a SiGe alloy is commonly used to fill the remaining SN holes and gaps between the SN cylinders.

Thus ends the front-end-of-line (FEoL) processes of the advanced BWL DRAM manufacturing process. Back-end-of-line processes occur primarily in the peripheral area, starting with a mask layer that protects the array area and etches away the metal layers in peripheral areas [Fig. 1.46(a)]. After PR strip and clean, a thick ILD3 (usually silicon oxide) is deposited, and oxide CMP planarizes the ILD3, as shown in Fig. 1.46(b). The via 1 (V1) mask is used to etch via holes through the thick oxide layer to land on M1. Although these via holes are very deep (2–3 μm)—even deeper than SN holes—their CDs are usually larger than the CD of the SN hole, and their pitches are also much

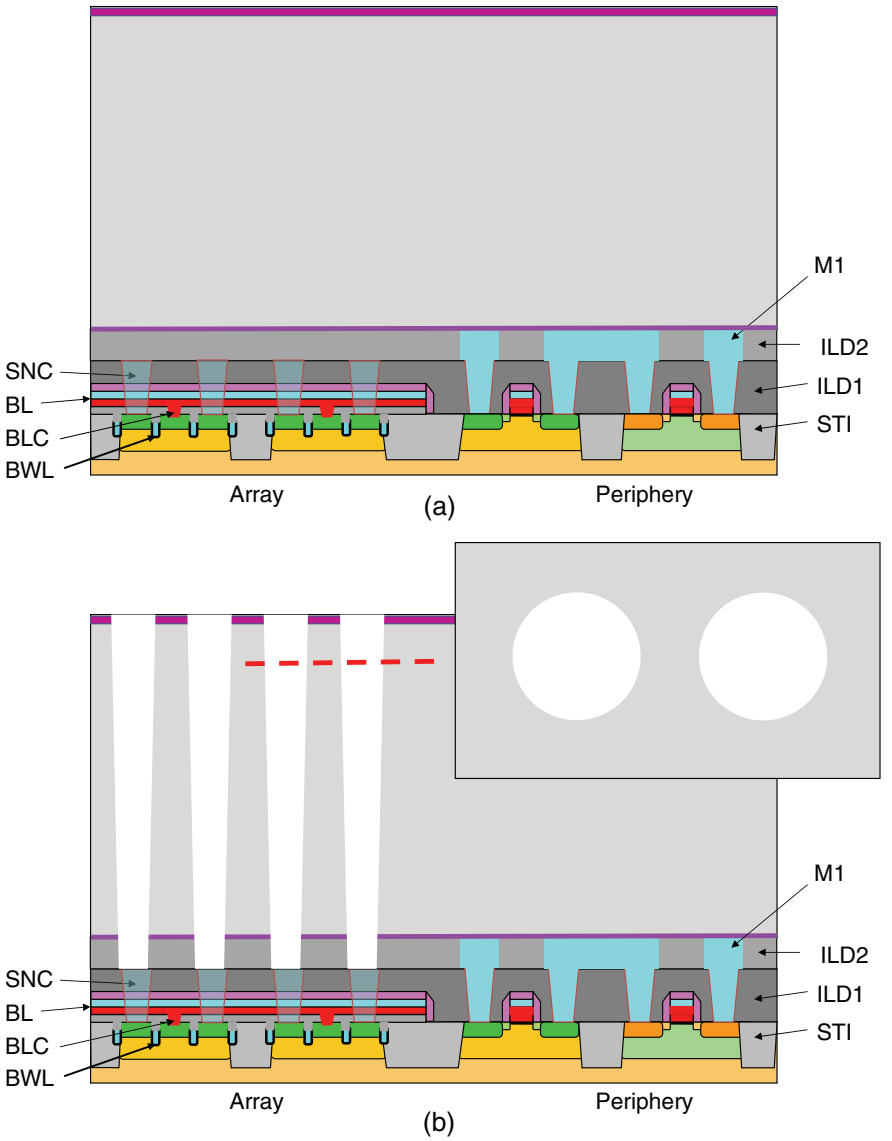


Figure 1.45 SN capacitor formation. After dielectric stack CVD (a); after SN hole etch, insert is top-down view of SN holes (b); after surface TiN removal (c); after ILD oxide wet etch (d); and after high-*k* dielectric and ground electrode deposition.

larger than that of the SN hole. V1 hole patterning has significantly relaxed control of the CD, overlay, and etch profile than that of SN hole patterning. After V1 etch and PR strip/clean, as shown in Fig. 1.46(c), Ti/TiN/W are deposited into the V1 holes, and a CMP process removes W/TiN/Ti from the wafer surface, leaving the conducting plug inside the ILD3, as shown in Fig. 1.46(d).

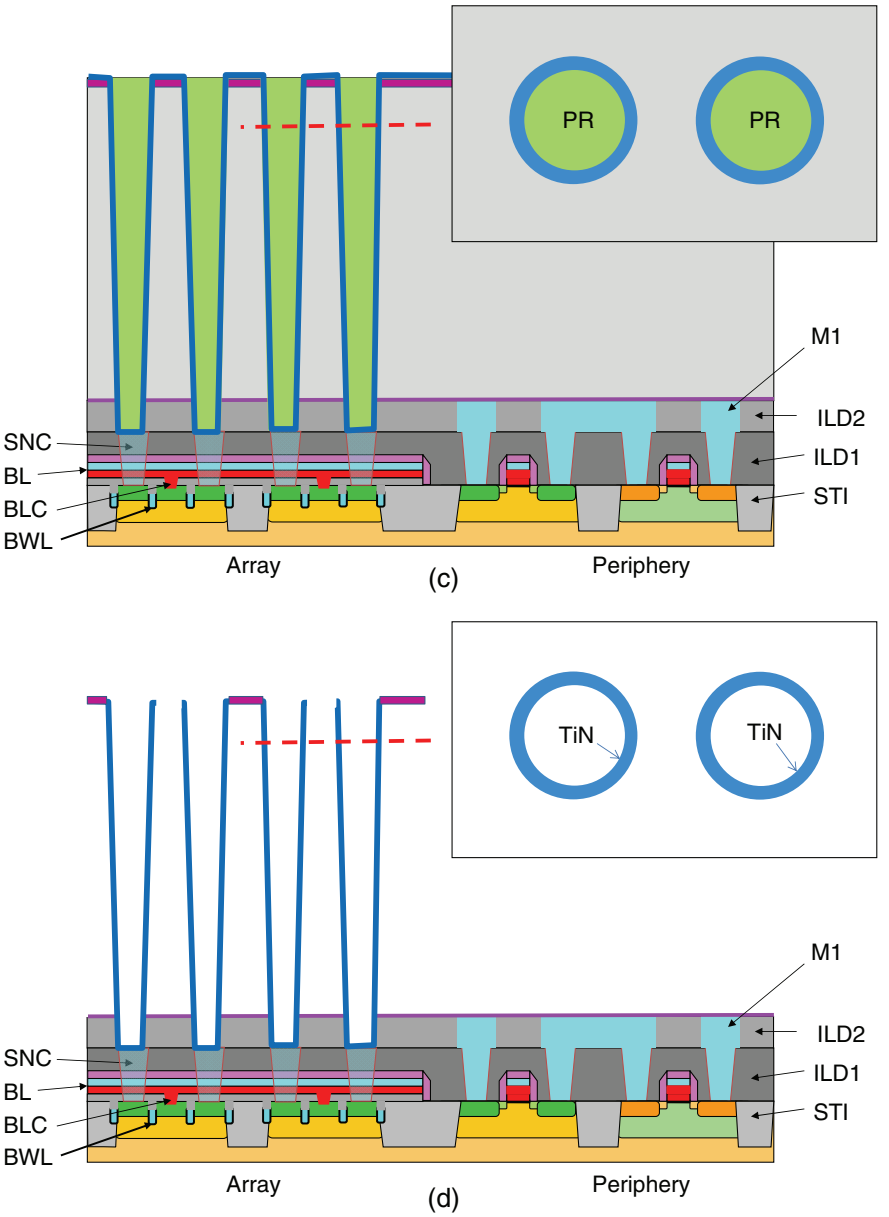


Figure 1.45 (Continued)

V1 formation is followed by metal 2 (M2), a single damascene copper (Cu) metallization process. M2 process steps include stop-layer etch and ILD4 oxide deposition [as shown in Fig. 1.47(a)], M2 mask, M2 etch [as shown in Fig. 1.47(b)], barrier and copper-seed layer deposition, copper electrochemical plating, copper anneal, and copper CMP (CuCMP) [as shown in

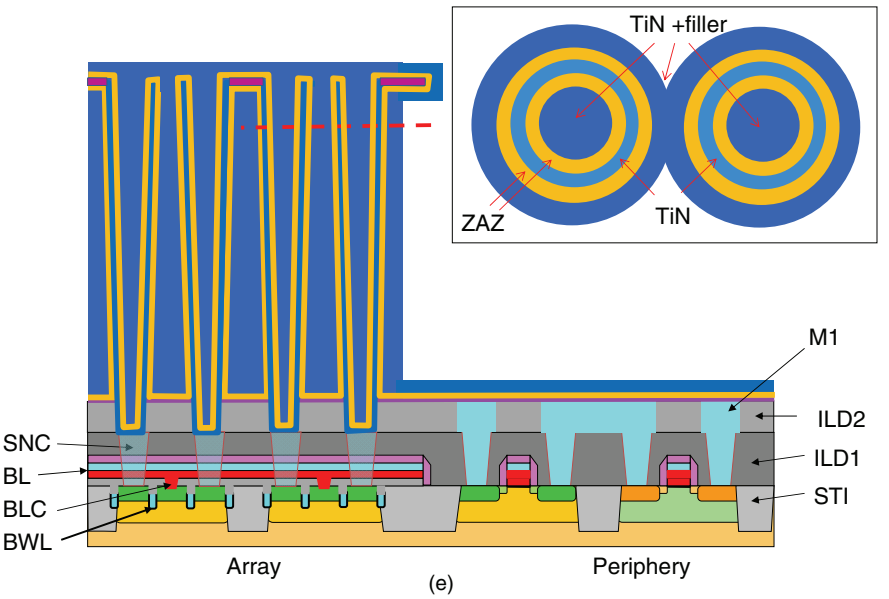


Figure 1.45 (Continued)

Fig. 1.47(c)]. For copper metallization, tantalum (Ta) or tantalum nitride (TaN) is commonly used for the barrier layer, and the seed layer is copper. Both layers are usually deposited by a physical-vapor-deposition (PVD) process. Bulk Cu plating usually uses an electrochemical process, with a wafer as the cathode and a pure-Cu plate as the anode. Copper ions from the sulfate ($CuSO_4$) in sulfuric acid will deposit on the wafer surface that is negatively biased. Additives such as an accelerator, suppressor, and leveler are added to the plating solution to accelerate Cu plating at the bottom of the hole or trench, suppress the corner deposition, and make the Cu surface flat, respectively. Table 1.8 lists the process steps of V1-M2.

After M2 CuCMP, a dual-damascene Cu metallization process is used to form V2 and M3. It starts with ESL, ILD5, dielectric cap, and metal HM deposition; a M3 mask is used to etch the TiN HM. After PR strip/clean, a V2 mask is used to etch dielectric layers and form the V2 holes and stop in ESL. After PR strip/clean, dielectric etch forms the M3 trench and allows the V2 to break though the ESL. Another clean process follows, and then the metal barrier and seed layer are deposited, and finally the bulk copper is plated, annealed, and polished. The Ta or TaN barrier and TiN HM mask are also removed during CuCMP. After wafer clean, a cap layer deposition finishes the V2-M3 module, as shown in Fig. 1.48; the process steps are listed in Table 1.9.

The V3 process starts with ESL and ILD6 deposition, as shown in Fig. 1.49(a); after the V3 mask is used to etch via holes, the PR is stripped and

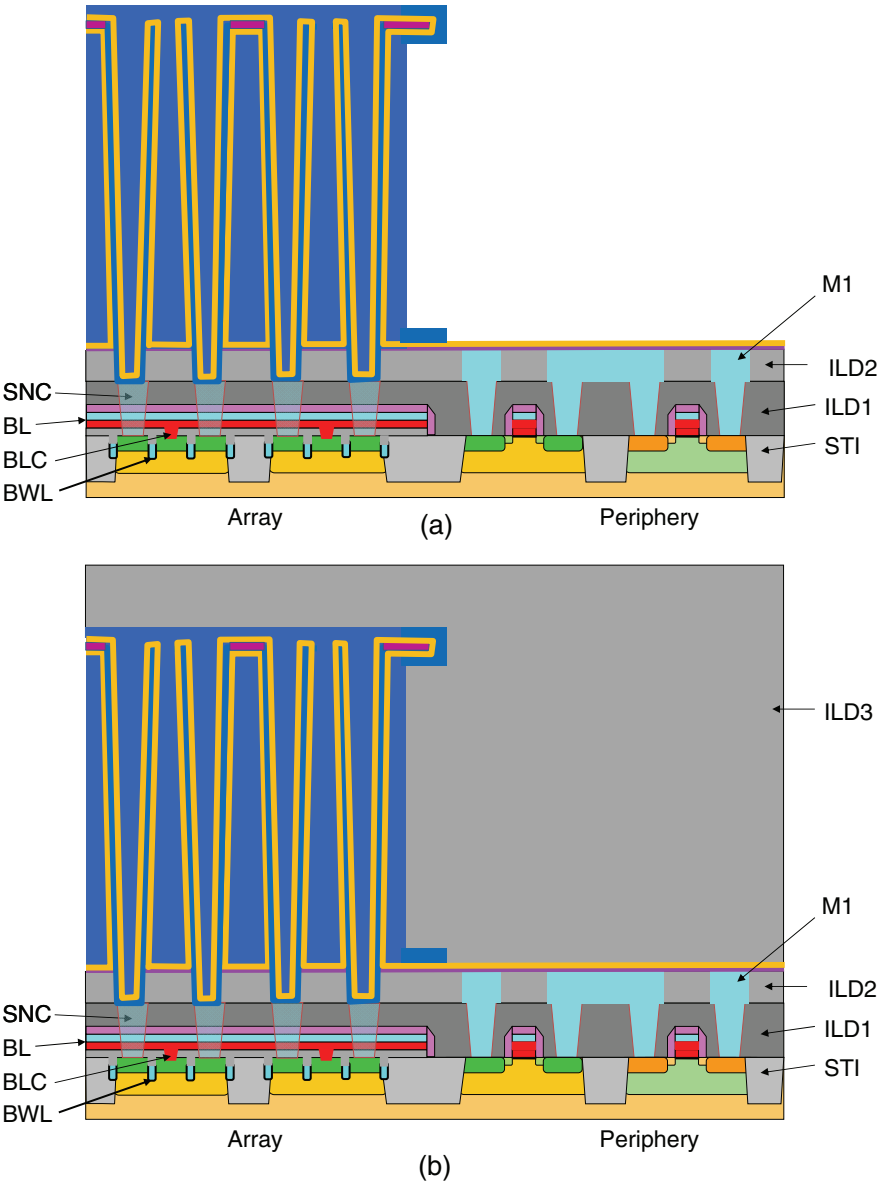


Figure 1.46 V1 process steps (a) after peripheral metal etch and clean, (b) after ILD3 CMP, (c) after V1 etch and PR strip/clean, and (d) after V1 WCMP.

wafer cleaned, as illustrated in Fig. 1.49(b). Then the Ti/TiN/W stack is deposited and the CMP process removes all metal layers on the wafer surface for W plugs that contact M4 to M3 (Fig. 1.49).

The last interconnect layer of this BWL DRAM manufacturing process is the metal 4 (M4) layer. It starts with metal stack PVD, which deposits Ti/Al-Cu/TiN, as shown in Figure 1.50(a). Here Ti is used to reduce contact

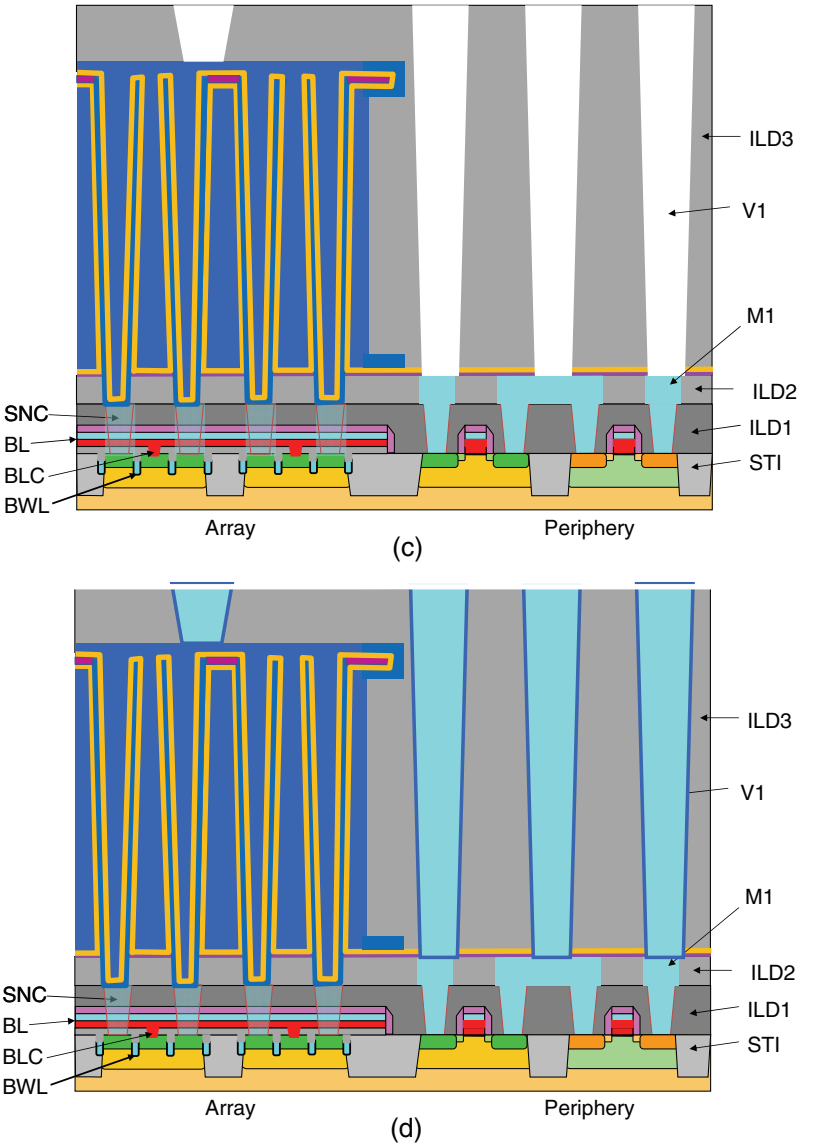


Figure 1.46 (Continued)

resistance between metal stack and W plugs. Al-Cu is an aluminum alloy with ~0.5% copper to increase resistance of electromigration and improve chip reliability. The top TiN is used as an ARC to reduce the standing wave effect during M4 patterning. The M4 mask is used to pattern and etch the M4 metal stack, which forms the last layer of interconnects and bond pads. After PR strip and clean [Fig. 1.50(b)], the passivation dielectric layers (usually silicon oxide and then silicon nitride) are deposited, as shown in

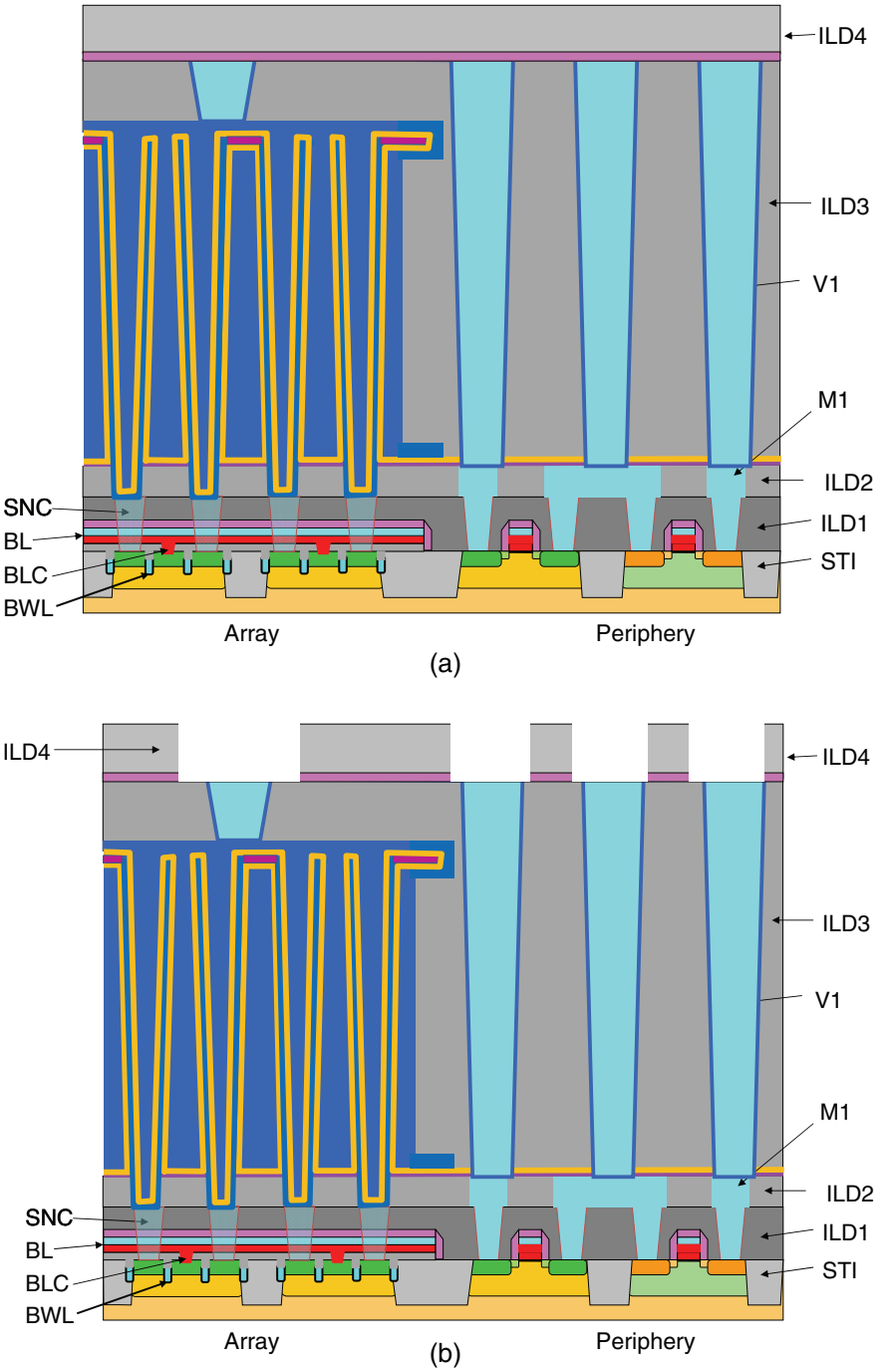


Figure 1.47 M2 processes: (a) ESL and ILD4 deposition, (b) M2 mask ILD4 and ESL etch, and (c) M2 CuCMP.

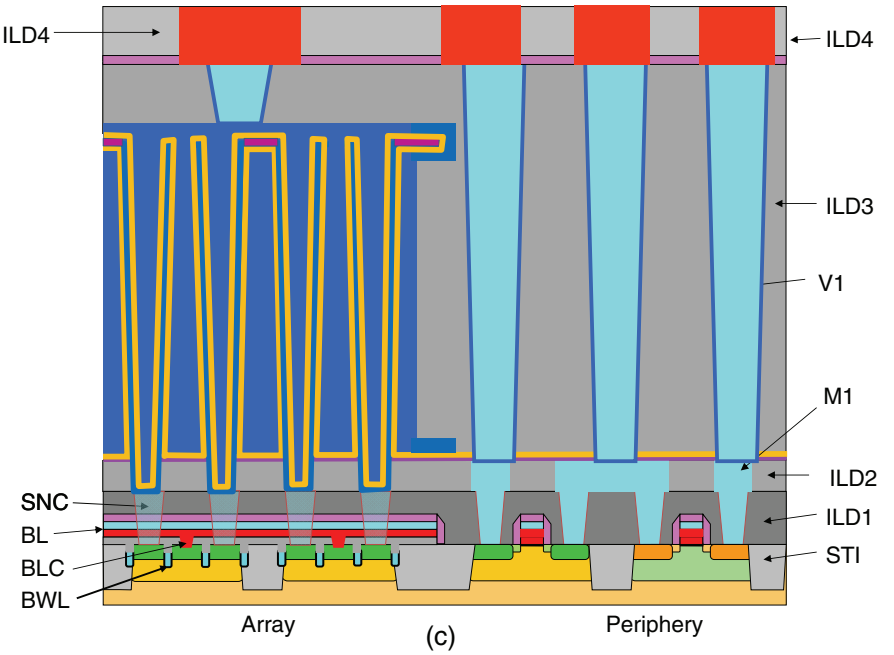


Figure 1.47 (Continued)

Table 1.8 V1 and M2 process steps.

Peripheral area mask	W/TiN/Ti CMP [Fig. 1.46(d)]
Etch SiGe/TiN/ZAZ	ESL deposition
Photoresist strip and clean [Fig. 1.46(a)]	ILD4 deposition [Fig. 1.47(a)]
ILD3 deposition	M2 mask
ILD3 CMP [Fig. 1.46(b)]	M2 etch
V1 mask	PR strip/clean [Fig. 1.47(b)]
V1 etch [Fig. 1.46(c)]	Barrier and seed-layer deposition
PR strip and clean	Bulk copper plating
Ti/TiN deposition	Cu anneal
W deposition	CuCMP [Fig. 1.47(c)]

Table 1.9 V2 and M3 process steps.

ESL, ILD5, and dielectric cap deposition	ILD 5 etch
Metal HM deposition [Fig. 1.48(a)]	ESL removal [Fig. 1.48(c)]
M3 mask	Clean
HM etch [Fig. 1.48(b)]	Barrier and seed-layer deposition
PR strip and clean	Bulk copper plating
V2 mask	Cu anneal
Dielectric cap and ILD5 etch	CuCMP [Fig. 1.48(d)]
PR strip and clean	

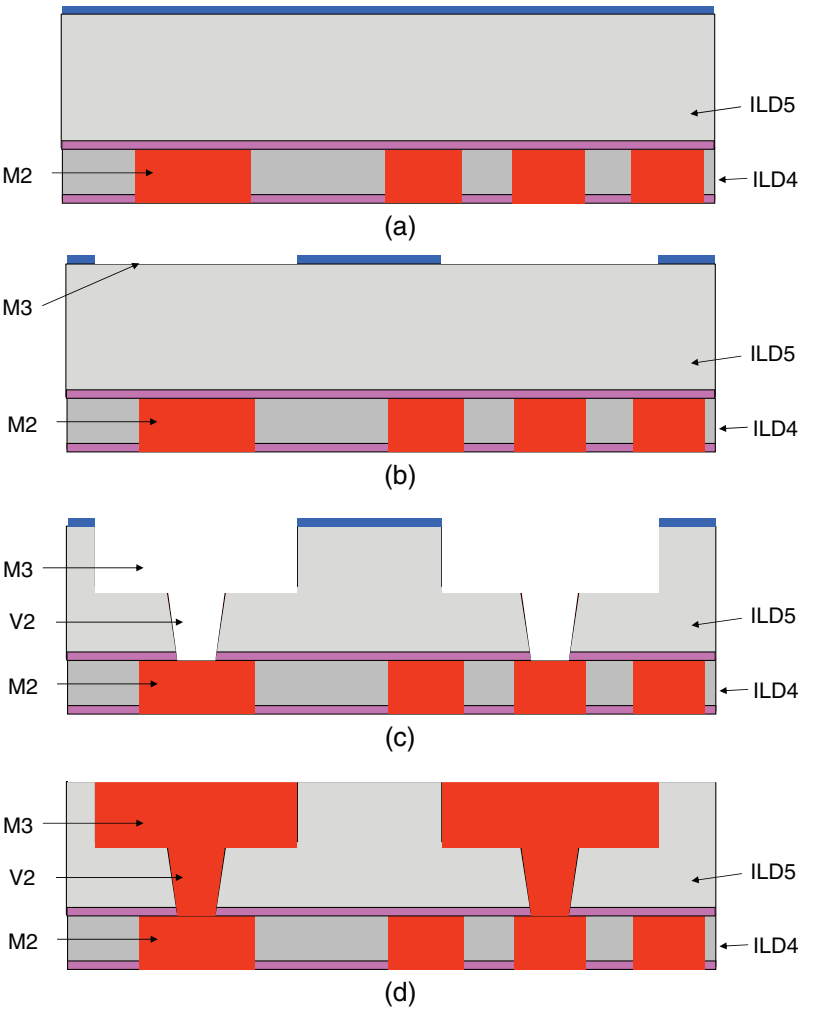


Figure 1.48 V2-M3 processes: (a) ESL, ILD4, and metal HM deposition, (b) M3 mask HM etch, (c) V2 ILD and ESL etch, and (d) M3 CuCMP.

Fig. 1.50(c); the last mask, a bond pad mask, is used to pattern and etch the passivation dielectric stack to expose the bond pads for test probing and wire bonding. After PR strip and clean, as shown in Fig. 1.50(d), the wafer is ready for the wafer-acceptance test (WAT). The wafer can be shipped for packaging if it passes the WAT. The process steps of V3-M4 and passivation are listed in Table 1.10.

Figure 1.51 illustrates the cross-section of BWL DRAM with both the array area and peripheral area. It has four metal layers: M1 is W, M2 and M3 are Cu, and M4 is an Al-Cu alloy. The wafer has finished processing, and it is ready for electrical test. The WAT will determine the yield data of

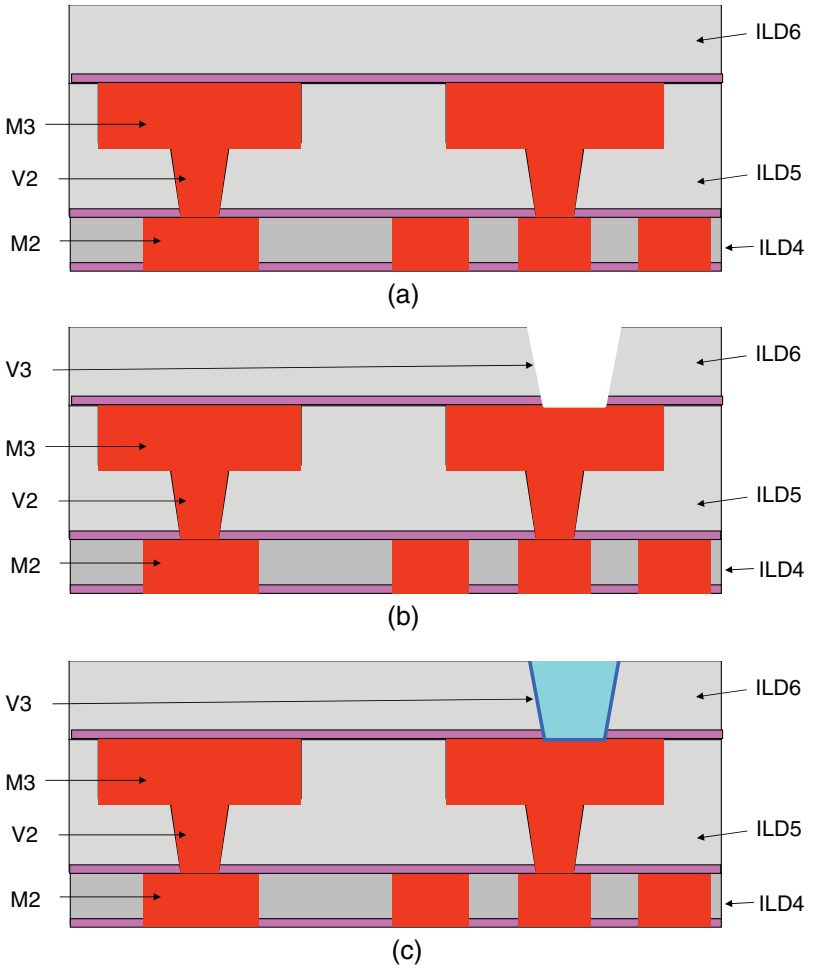


Figure 1.49 Cross-section of via 3 process steps: (a) ESL and ILD6 deposition, (b) V3 etch and PR strip/clean, and (c) V3 WCMP.

the wafer process; the yield is the key for the success of the wafer fab, which is especially important for DRAM fabs because DRAM is very cost-sensitive.

1.3 Brief Summary of DRAM

This chapter described an advanced-technology node DRAM manufacturing processes with a BWL, recessed cylindrical capacitor, and $6F^2$ layout. Future scaling of the DRAM has many challenges, such as capacitor formation, a new device structure, new layout, and maybe new patterning processes with EUV.

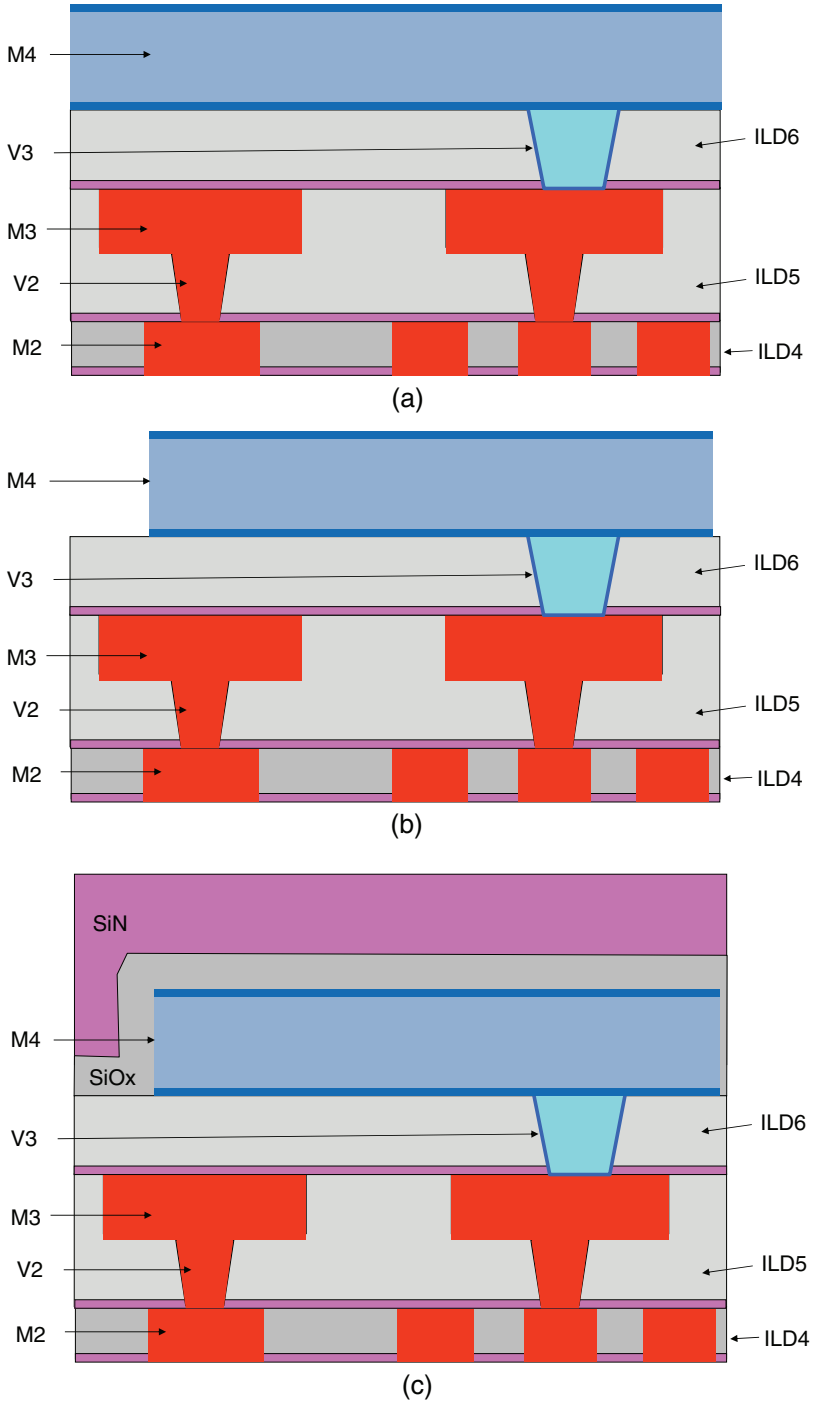


Figure 1.50 Cross-section of V3-M4 and passivation processes: (a) after M4 Ti/l-Cu/TiN PVD, (b) after M4 etch and PR strip and clean, (c) after passivation oxide and nitride CVD, and (d) after bond pad etch and PR strip and clean.

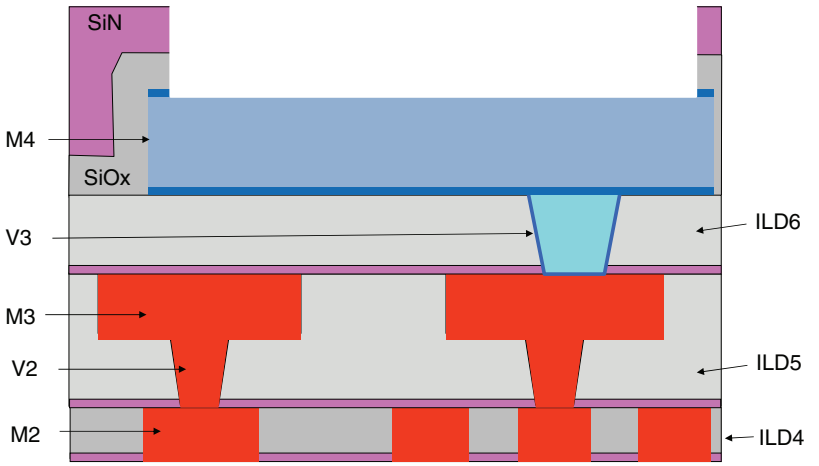


Figure 1.50 (Continued)

As addressed earlier, the capacitor is the most challenging part of the DRAM scaling. When the DRAM feature size scales down deep into a 1x-nm node, the CD and pitch of the SN cylinder holes could become too small to allow a recessed cylinder to form a double-sided capacitor; the capacitor must return to the single-side cylinder without recess, similar to the one shown in Fig. 1.8, with a much higher aspect ratio. Because the aspect ratio of the recessed cylinder capacitor is already ~50:1, the aspect ratio of a cylinder capacitor with a smaller CD and without recess could easily go higher than 200:1 if the same high-k dielectric (ZAZ) is still used. To control the aspect ratio within 100:1, a new capacitor dielectric with a k value significantly higher than ZAZ (k~20) is needed. One such candidate is strontium titanate ($SrTiO_3$ or STO, $k = 146$ in bulk material) using ALD to achieve excellent step coverage (~95%), high thickness uniformity, and good stoichiometry.[12]

Figure 1.52 illustrates the DRAM array-area factor improvement in recent years. The smallest achievable unit array area for a planar device is $4F^2$,

Table 1.10 V3-M4 and passivation process steps.

ESL and ILD6 deposition [Fig. 1.49(a)]	M4 mask
V3 mask	Etch TiN/Al-Cu/Ti stack
ILD6 and ESL etch	PR strip/clean [Fig. 1.50(b)]
PR strip and clean [Fig. 1.49(b)]	Passivation oxide and nitride deposition [Fig. 1.50(c)]
5Ti/TiN/W deposition	Bond pad mask
W/TiN/Ti CMP [Fig. 1.49(c)]	Etch nitride and oxide
Wafer clean	PR strip and clean [Fig. 1.50(d)]
Ti/Al-Cu/TiN deposition [Fig. 1.50(a)]	

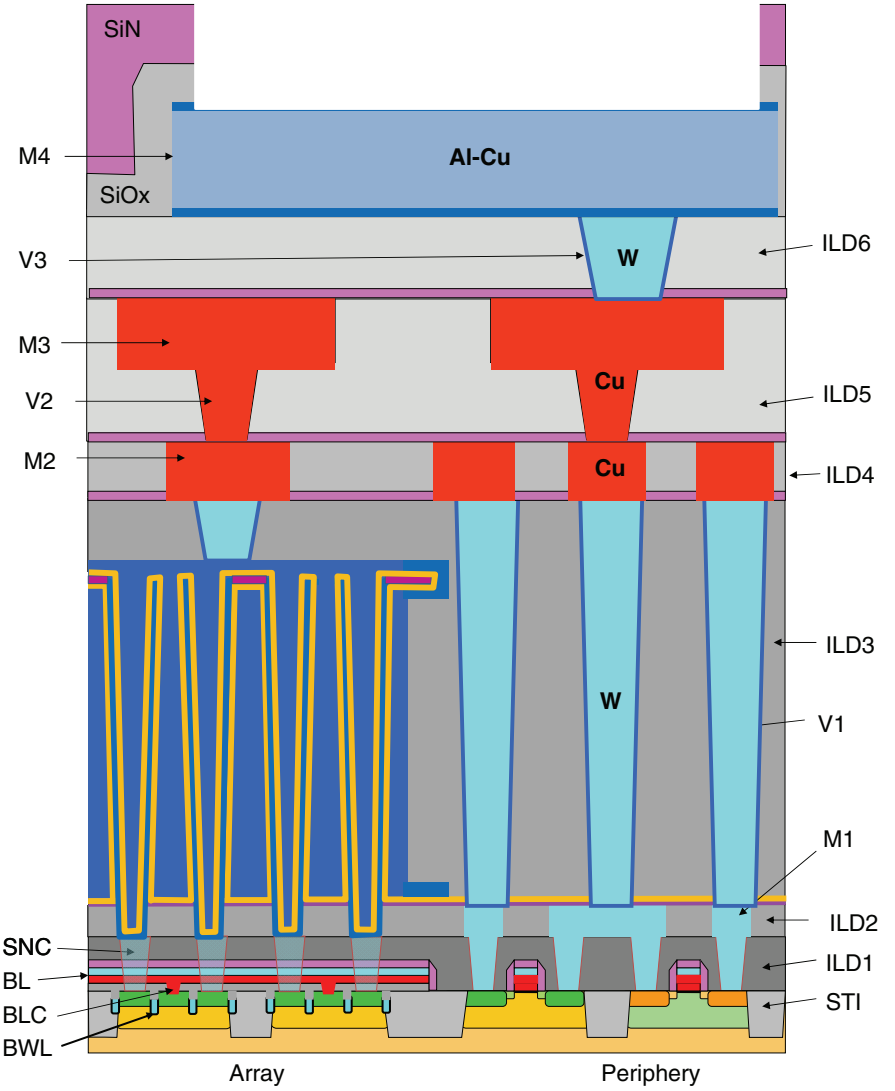


Figure 1.51 Cross-section of BWL DRAM with four metal interconnection layers.

which has been used in planar NAND flash cell design. Theoretically, DRAM can also achieve a unit array area of $4F^2$ with a vertical cell transistor and buried bit-line design, as shown in Fig. 1.53. Figure 1.53(a) shows the $4F^2$ DRAM cell layout. It can be seen from Fig. 1.53(b) that the device architecture is quite different from the previous generation. The BL is buried, and the cell transistor is a vertical gate-all-around fully depleted MOSFET. This type of DRAM has not yet been used in DRAM chip production.

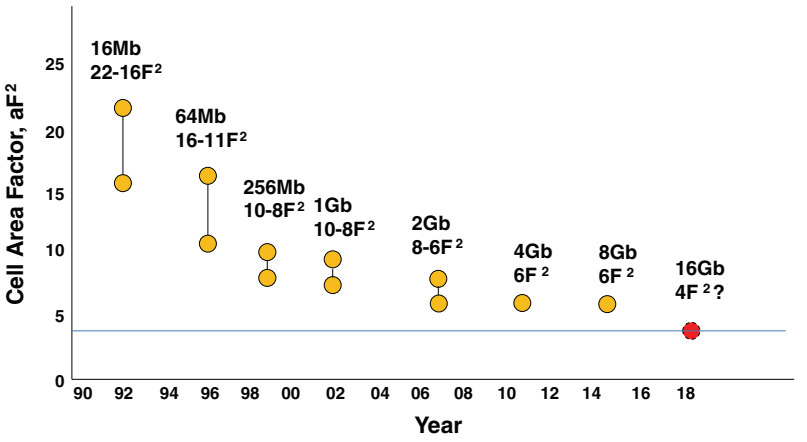


Figure 1.52 DRAM array-area factor scaling, based on Ref. C.

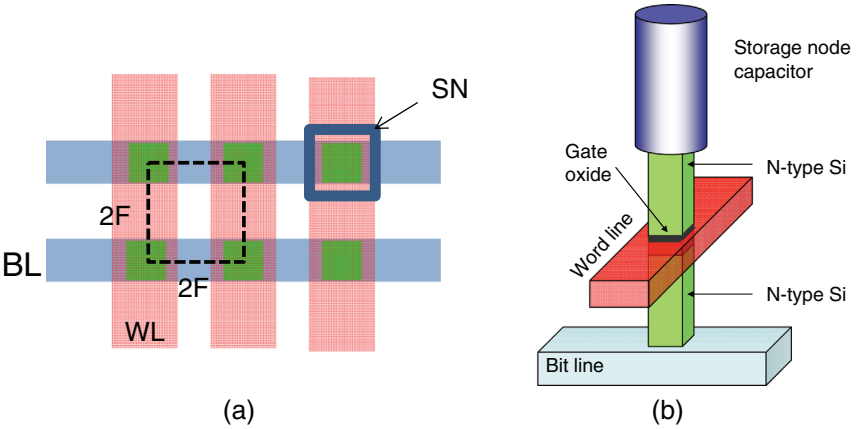


Figure 1.53 (a) Layout of a $4F^2$ DRAM, and (b) 3D illustration of $4F^2$ DRAM.

1.4 Review Questions

1. How many devices are in a DRAM unit cell? What types of devices?
2. List at least three types of DRAM cell transistors.
3. What are the advantages of a 3D capacitor compared to a 2D planar capacitor?
4. Explain how can BWL DRAM reduce three photolithography masks from RG DRAM while forming the cell transistor/WL, contacts, and SN capacitor?
5. What is the benefit of recessing a SN cylinder?
6. Describe the needs of high-k dielectric in DRAM manufacturing.
7. Explain why a SN capacitor is the most challenging part of DRAM scaling.
8. Describe the benefits and challenges of scaling a DRAM cell from $8F^2$ to $6F^2$ and then $4F^2$.

Chapter 2
3D-NAND Flash and Its Manufacturing Process

After reading this chapter, you should be able to

- Draw and describe a NAND flash memory cell;
- Give the minimum unit-cell size of planar NAND flash in relationship to the technology node;
- Explain the advantages of 3D-NAND flash compared to planar NAND flash; and
- List at least two processes in the 3D-NAND manufacturing process that need HAR etch.

2.1 Introduction

Flash memory chips are nonvolatile memory (NVM) chips, which can keep memory without power supply. In comparison, DRAM is volatile memory and needs a power supply. Flash memory chips, especially NAND flash memory chips, are commonly used in universal serial bus (USB) drives, secure digital (SD) cards, and SSDs. These NVM devices are widely used in electronics devices for data storage, especially in mobile electronics such as digital cameras, smartphones, tablets, high-end laptops, etc. Compared with a hard-drive disk (HDD), a SSD has a shorter data-access time, consumes less power, and is more reliable because it does not have any moving parts.

The basic device structure of a flash memory cell is very similar to a NMOS. It has a p-well and n + S/D, but the main difference is that it has a floating gate (FG), a control gate (CG), and inter-gate dielectrics (IGD) in between, as illustrated in Fig. 2.1. It is a charge trap device that retains memory by tunneling electrons from the drain through the gate oxide and trapping them in the FG. The memory can be erased by tunneling the stored electrons away from the FG to the CG through the IGD. Each time that

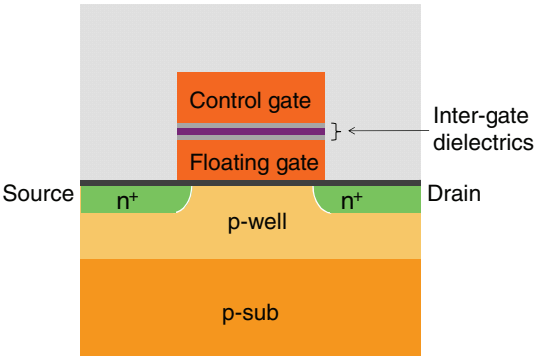


Figure 2.1 Basic structure of a floating-gate nonvolatile memory device.

electrons tunnel through the gate oxide and IGD can cause dielectric degradation, and so a flash cell has a limited number of writes and erases ($\sim 10^5$).

Figure 2.2 illustrates a variation of the charge-trap flash-memory device. Instead of a polysilicon floating gate, it uses a SiN layer to trap the charge and store the data. Most planar NAND flash memories are FG NAND flash devices. However, a majority of 3D-NAND flash devices use memory cells with a SiN charge-trap layer.

There are two types of flash memory, NOR and NAND, depending on how the flash cells are connected to the bit line and source line (ground). Figures 2.3(a) and (b) are the NOR flash circuit and its cross-section, respectively. Figures 2.3(c) and (d) are a 64-bit string NAND flash circuit and the corresponding cross-sections, respectively.

Every memory cell has a shared contact to the bit line and shared source line in a NOR flash. For a NAND flash, it shares a bit line contact and a source line contact for every string of memory cells. NOR flash is equivalent to a 1-cell string NAND flash that does not require the select gates. NOR flash can achieve random access of any memory cell while NAND flash

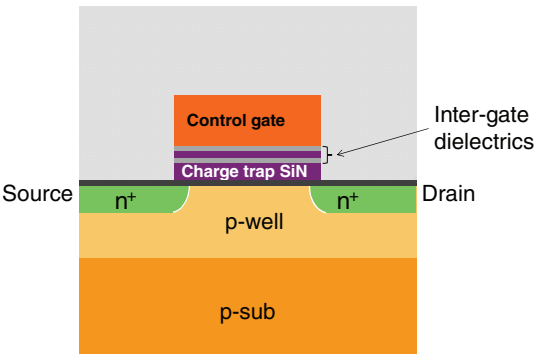


Figure 2.2 Flash memory cell with a SiN charge-trap layer.

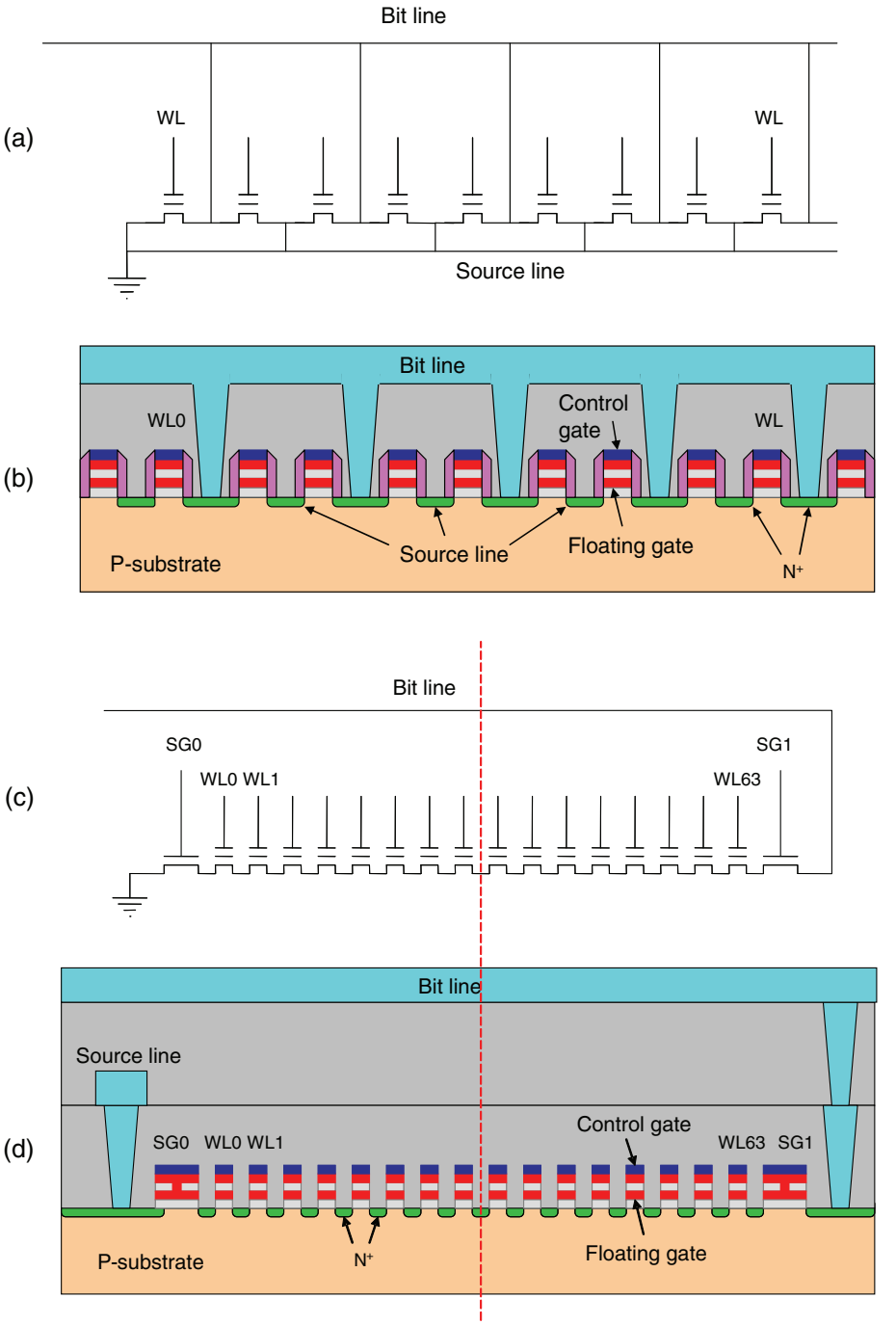


Figure 2.3 (a) NOR flash circuit, (b) NOR flash cross-section, (c) NAND flash circuit, and (d) planar NAND flash cross-section.

cannot. Although NOR flash has shorter reading time than NAND flash, it has longer write time and erase time. It also costs significantly higher than NAND flash because of its lower packing density due to the dense bit line contacts. This is the main reason why the majority of solid state storage devices are NAND flash based, and NOR flash only takes a small niche market.

Figure 2.4(a) shows the array area layout of the WL and AA of a planar NAND flash memory; Figure 2.4(b) shows its cross-section along the AA direction in both the array area and peripheral area. In the figure, CB represents the contact with the BL, SG stands for select gate, and SL stands for source line. The NAND flash-memory cell has the highest pattern density of all planar IC devices. Its technology node is the half pitch of the WL

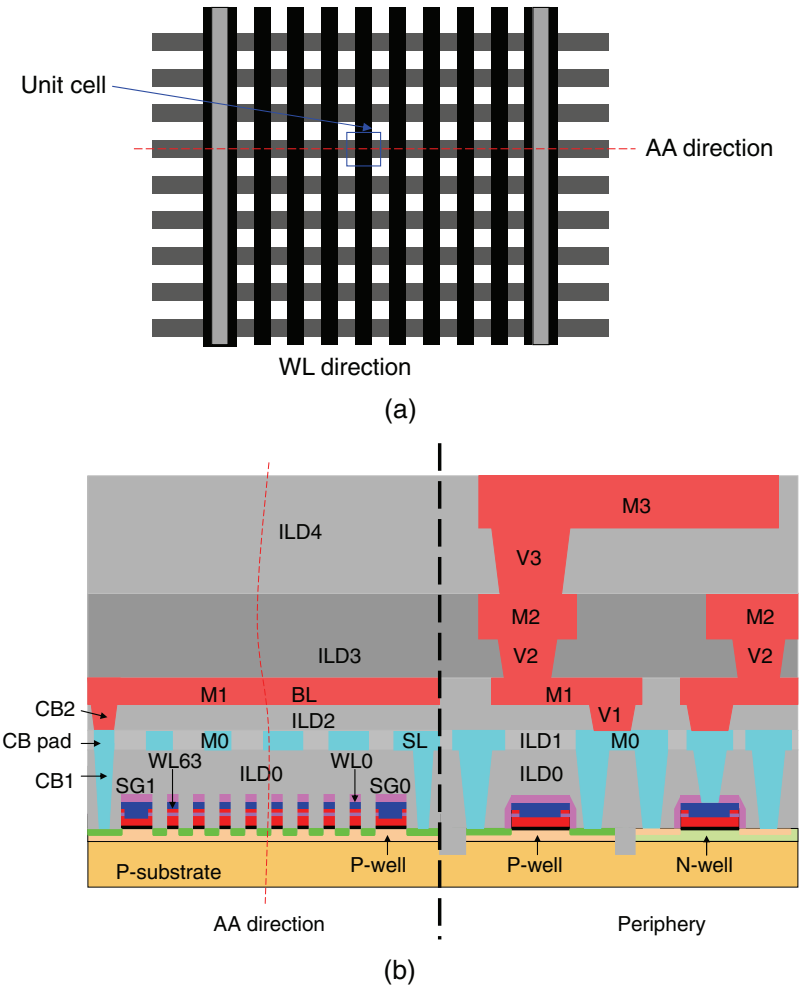


Figure 2.4 (a) Array area layout of 64-bit planar NAND flash, and (b) cross-section with both the array area and peripheral area.

pattern, and its unit array area can achieve the minimum value of $4F^2$ [see Fig. 2.5(a)]. Here, F is the technology node. For example, the WL pitch of a 16-nm planar NAND flash is 32 nm, which is far smaller than the minimum achievable pitch of a high-NA 193-nm immersion photolithography process. Without extreme-ultraviolet (EUV) lithography, 16-nm planar NAND flash requires pitch tripling or SAQP using 193-nm immersion lithography.

Scaling planar NAND flash devices beyond the 10-nm node would need more than quadrupling with a 193-nm immersion lithography process for at least three critical layers, AA, WL, and BL, which would make its manufacturing cost very expensive. To reduce the scaling cost, a different device structure based on the idea of a vertical gate-all-around MOSFET (GAA-FET) had been proposed. Making the NAND string in the vertical direction can dramatically reduce the footprint and increase the packing density. Figure 2.5 illustrates the 3D and top views of 3D-NAND flash.[5] The illustration shows the staircase contact, the lower select-gate string, four flash memory devices, and the upper select gate.

Figure 2.5(c) is the circuit of the 3D NAND array shown in Fig. 2.5(a); it basically puts the four-cell NAND string in the vertical direction and uses a vertical GAA-FET for the select gate and the flash memory cell. The layout unit area is $6F^2$ with four stacks, equivalent to $1.5F^2$. In the first high volume manufacturing of 3D-NAND, 32 stacks are used, which can achieve $0.1875F^2$ with the same layout. A quick calculation can find how much a 32-stack 3D-NAND helps relax the photolithography requirement:

$$6F_{3D}^2/32 = 4F_P^2 \quad \text{or} \quad F_{3D} = (8/\sqrt{3})F_P.$$

Here, F_{3D} is the technology node of 3D-NAND, and F_P is the equivalent planar NAND flash-technology node. With 32-stack 3D-NAND, a ~74-nm technology node can achieve the same memory density as 16-nm planar NAND flash. It is assumed here that the 16-nm planar flash uses the $4F^2$ layout, which may not always be true because the pitch of the AA and BL is sometimes relaxed so they can be formed with SADP using 193-nm immersion photolithography to reduce the manufacturing cost. Whereas a 16-nm (32-nm WL pitch) 3D-NAND pattern needs triple patterning or SAQP, the 74-nm (148-nm pitch) version can be achieved with a single-patterning process. Therefore, the photolithography requirement of 3D-NAND is dramatically relaxed, and the main challenges are the etch/clean processes that form the HAR channel holes, isolation trench, and staircase contact holes. Another challenge involves the conformal deposition of multiple thin film layers into the HAR structures. The scaling of a NAND flash memory chip occurs mainly in the z direction by stacking more layers. The trend of stacking layers is 32, 48, 64, 96, and 128, and it has room to further scale the feature size of 3D-NAND flash memory. The aspect ratio of the 3D-NAND device structure

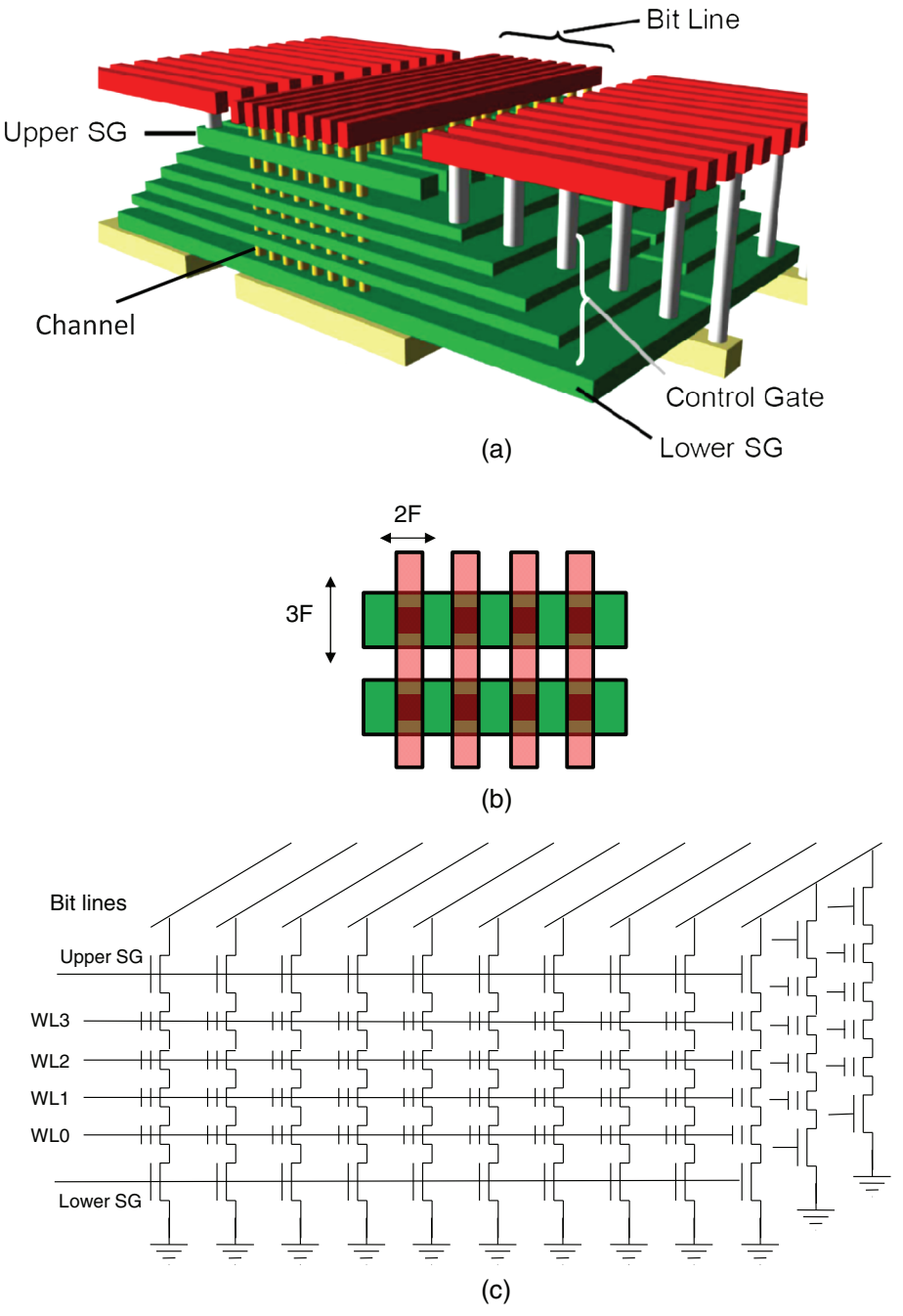


Figure 2.5 (a) 3D view and (b) top view of a 3D-NAND flash memory array (source: Ref. 5, reprinted with permission from IEEE) and (c) its circuit.

will increase further in the future, and thus etch, clean, deposition, metrology, and inspection will become even more challenging.

Question: What is the equivalent $4F^2$ planar NAND flash-memory technology node of a 74-nm, $6F^2$, 128-stack 3D-NAND flash memory chip?

Answer: $6F_{3D}^2/128 = 4F_P^2$ or $F_P = \sqrt{(3/256)}\,F_{3D} \sim 0.108\,F_{3D} \sim 8$ nm

2.2 3D-NAND Flash Memory Manufacturing Processes

The manufacturing processes of 3D-NAND flash memory are quite different from those of traditional planar NAND flash; they begin with the peripheral CMOS devices. Array-area processes, such as multi-layer deposition and staircase formation, follow with a channel hole module, which includes channel hole etch, lower SG formation, and channel layer deposition. The next module is isolation trench, which is also used to create the control gates of the 3D-NAND, followed by the contact module. Because the contact holes land on the staircase and peripheral circuits, the depth varies in a significantly larger range, which poses a large challenge to the dielectric etch processes. After that, M1 forms the local interconnect and other metal wires, M2 forms the bit line, and M3 finishes the interconnects; after passivation dielectrics are deposited, the last mask opens bond pads for wire bonding, and the process is finished. This section describes these process modules in more detail.

2.2.1 Peripheral module

For planar NAND flash memory, the peripheral devices are processes at the same time as the flash memory cells in array area. The major differences involve peripheral devices that have a larger feature size and require a mask to remove part of the inter-gate dielectric so that the control gate can short to the floating gate, just like the select gates in the array area (Fig. 2.4).

For 3D-NAND flash memory, the peripheral devices are planar MOSFETs. The feature size for peripheral devices is usually large enough to be patterned with a single 193-nm immersion-lithography process. The process flow is very similar to the front-end-of-line (FEoL) part of the CMOS process described in Chapter 14 of Xiao.[11]

The process starts with STI formation, which includes pad oxidation, nitride deposition, AA mask patterning, nitride etch, PR strip and clean, silicon etch and clean, oxidation and oxide CVD, oxide CMP, and nitride and pad oxide strip. Figure 2.6(a) shows the cross-section of the starting silicon, and Fig. 2.6(b) shows the peripheral CMOS devices after STI formation. The label SiOx represents silicon oxide deposited with the CVD process.

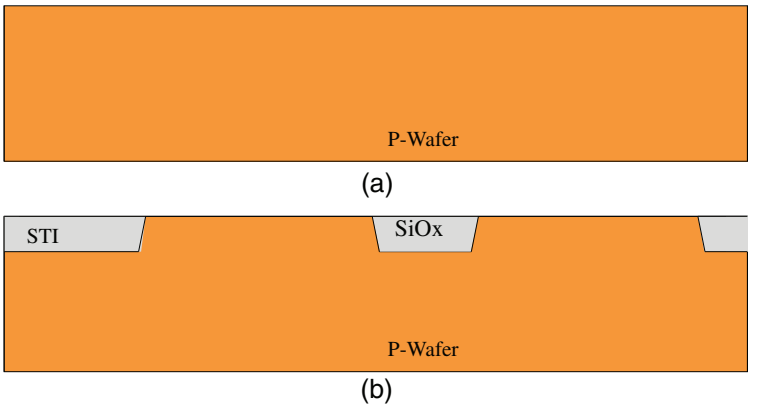


Figure 2.6 Cross-section of (a) the starting silicon and (b) the peripheral CMOS devices after STI formation.

Next are well and channel implantations for both NMOS and PMOS; Figure 2.7 shows the cross-section of peripheral CMOS after both ion implantations. There are two photomasks in these steps: one for NMOS, and one for PMOS. After wafer clean, gate oxidation, polysilicon, metal silicide, and HM deposition, the gate mask is used to pattern the hard mask, and the polysilicon/silicide gate stack is etched with the HM pattern. Figure 2.8 shows the cross-section of the peripheral CMOS devices after polysilicon/silicide gate-stack etch.

After wafer inspection and clean, an oxidation is performed, and two ion implantation masks are used to form a lightly doped drain (LDD): one mask for NMOS and another for PMOS. After wafer clean, a conformal dielectric

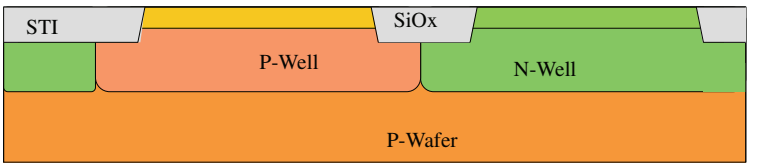


Figure 2.7 Cross-section of the peripheral CMOS after well and channel implantation.

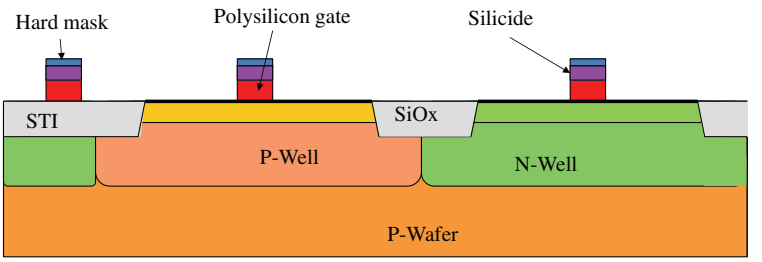


Figure 2.8 Cross-section of the peripheral CMOS after polysilicon gate etch.

layer is deposited and etched back only in the vertical direction to form the sidewall spacer. Two more masks are used for the heavily doped S/D. After wafer clean and rapid thermal annealing (RTA), the transistors of the peripheral devices are formed. Figure 2.9 shows the cross-section of the peripheral CMOS after S/D implantation and RTA. There are four photomasks and at least four ion implantations in Figs. 2.8–2.9.

The last steps of the peripheral FEoL process are SiN liner deposition and very thick (~3 μm) silicon oxide deposition (Fig. 2.10); note that PMD stands

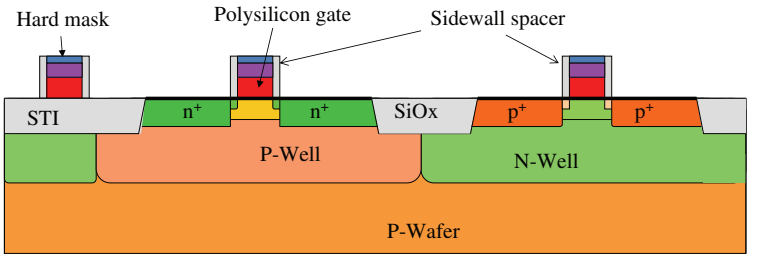


Figure 2.9 Cross-section of the peripheral CMOS after S/D formation.

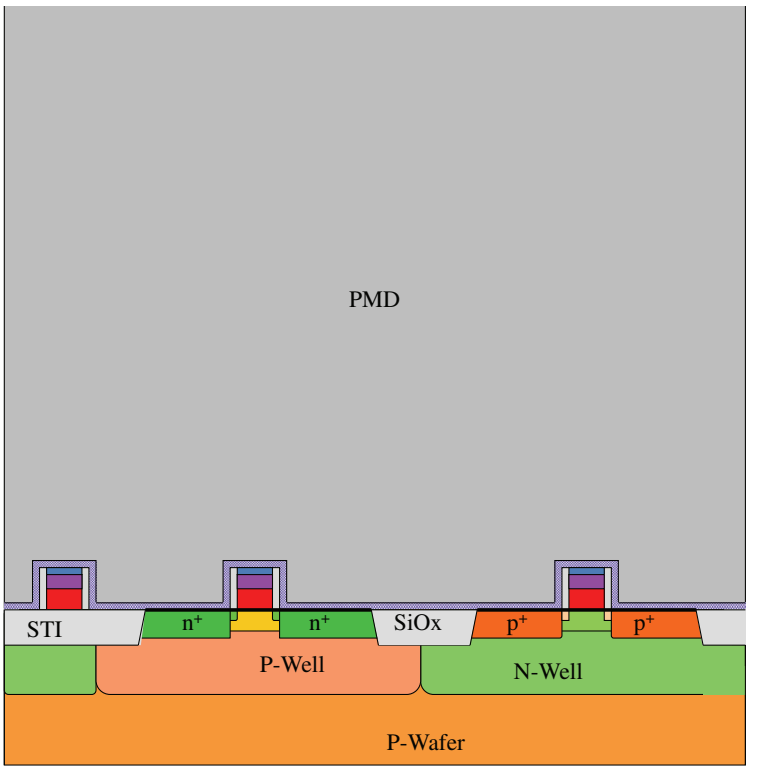


Figure 2.10 Cross-section of the peripheral CMOS after the FEoL processes are finished.

Table 2.1 Peripheral CMOS process steps.

Wafer clean [Fig. 2.6(a)]	PR strip and clean
Pad oxidation	Hard mask deposition
Nitride deposition	Gate mask
AA mask	Etch hard mask
Nitride etch	PR strip and clean
PR strip/clean	Etch silicide/polysilicon
Silicon etch	Wafer clean (Fig. 2.8)
Wafer clean	n-LDD mask
Oxidation	n-LDD ion implantation
Oxide deposition	PR strip and clean
Oxide CMP	p-LDD mask
Strip nitride and pad oxide, and wafer clean [Fig. 2.6(b)]	p-LDD ion implantation
Oxidation of sacrificial oxide	PR strip and clean
n-well mask	Spacer dielectric film CVD
n-well and p-channel ion implantation	Dielectric etch back
PR strip and clean	n-S/D mask
p-well mask	n-S/D ion implantation
p-well and n-channel ion implantation	PR strip and clean
PR strip and clean (Fig. 2.7)	p-S/D mask
Strip sacrificial oxide and wafer clean	p-S/D ion implantation
Gate oxidation	PR strip and clean
Polysilicon and silicide deposition	RTA (Fig. 2.9)
Poly-dope mask	SiN liner deposition
Poly-dope ion implantation	PMD deposition (Fig. 2.10)

for "pre-metal dielectric." The FEoL process steps of peripheral CMOS are listed in Table 2.1.

In a real peripheral process, there are more process steps than those listed in Table 2.1. For example, input/output transistors differ from sensor amplifier transistors. Their working voltages are different, and thus their gate oxide thicknesses are different, which requires additional photomask steps. It requires at least two gate-oxidation processes. Furthermore, their ion-implantation processes are different due to the different junction depths and dopant concentration requirements, which also require additional implantation mask steps.

2.2.2 Multi-layer deposition and staircase formation

The processes to build 3D-NAND memory devices in the array area start after the FEoL processes in the peripheral area are finished. A mask that defines the array area is applied, followed by an etch process that removes the thick oxide and SiN liner. Because the pattern is so large, wet etch can be used for this process; hydrofluoric acid (HF) can be used to etch silicon oxide, and hot phosphoric acid can be used to remove SiN. Figure 2.11(a) illustrates a portion of this mask that defines the array area. Figure 2.11(b) is a close-up of the tiny box on the right side of Fig. 2.11(a). Figure 2.11(c) shows the cross-section of the transition region between the array area and peripheral area

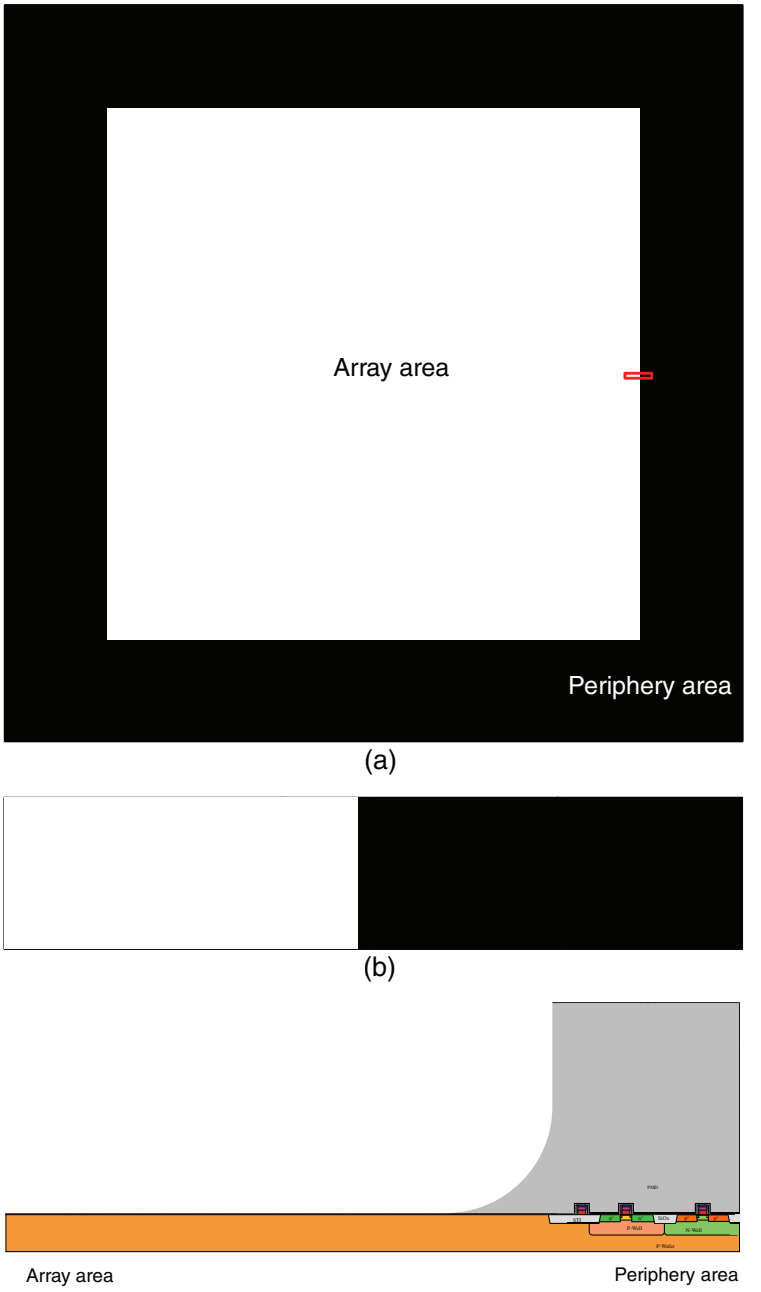


Figure 2.11 (a) Cell mask, (b) close-up, and (c) cross-section after etch. See Fig. 2.10 for a close-up of the right part of (c).

after patterning, etch, PR strip, and wafer clean correlated to Fig. 2.11(b). The features on the right of Fig. 2.11(c) are scaled from Fig. 2.10.

The wafer is then ready to deposit the multi-layers, of which there are several types, including ONON, which stands for oxide/nitride/oxide/nitride. Here, oxide refers to silicon oxide, and nitride refers to silicon nitride. Another type of multi-layer is OPOP, which stands for oxide/poly/oxide/poly; poly refers to polycrystalline silicon, or polysilicon.

This chapter discusses the ONON stack. In a CVD system, multiple pairs of ON layers are deposited. A 32-cell stack needs at least 34 pairs of ON because there are also two SGs, one at the bottom and one at the top. Multiple dummy layers are needed to isolate the SGs and the memory cells to avoid cross-talk. A reverse-engineering report shows that 32-cell 3D-NAND flash-memory products have, in fact, 39 pairs. Figure 2.12(a) illustrates the first two pairs of ONON deposition, and Fig. 2.12(b) shows after 34 pairs of ONON are deposited, in both the array area and the peripheral area. There are many challenges facing this multi-layer deposition: it needs to have uniform

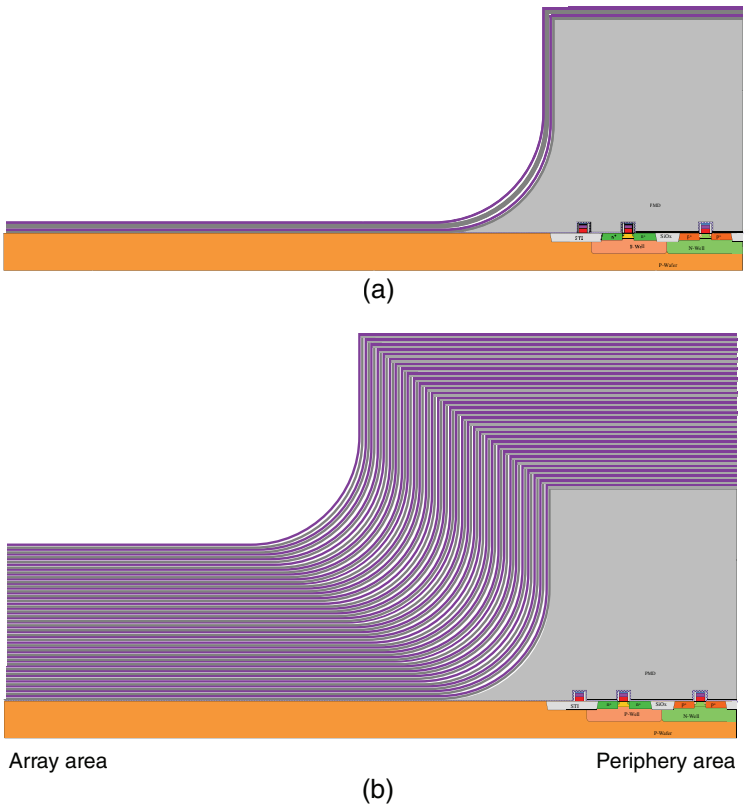


Figure 2.12 Multi-layer deposition: (a) two pairs of ONON, and (b) 34 pairs of ONON.

thickness and low defect density. It is also very challenging to measure the thickness of each layer.

Staircase formation is a unique process of 3D-NAND flash. After multi-layer deposition, a very thick photoresist layer (~5–6 μm) is coated on the wafer surface, the staircase mask is applied, and a well-controlled oxide and nitride etch is performed, which stops after the first nitride is etched away, as shown in Fig. 2.13(a). After the first pair of oxide and nitride is etched away, a controlled photoresist trimming is performed and the trimmed photoresist is used to etch the first and the second pair of oxide and nitride [Fig. 2.13(b)]. The photoresist is then trimmed again, and the first, second, and third pair of oxide/nitride are etched, which is illustrated in Fig. 2.13(c). Here, Oxide_N + 1 is the cap oxide, and Oxide_N and Nitride_N are the n^{th} pair of the oxide and nitride stack, respectively.

The number of times that photoresist can be trimmed is limited by its original thickness. After that limit, it must be stripped so that another thick photoresist can be coated, and the second staircase mask is applied to repeat the staircase etch process. Four to five staircase masks are needed for the stack of 34 pairs of oxide/nitride. This process will repeat multiple times until it reaches the silicon surface, as shown in Fig. 2.14. Here, Oxide_x and Nitride_x is the x^{th} pair of oxide/nitride. The Oxide_2 layer is thicker than other oxide layers because it is the isolation oxide used to separate the lower select gate from the cell devices.

Figure 2.15(a) shows a portion of a staircase mask. Figure 2.15(b) is a close-up of the box in Fig. 2.15(a). The staircase is located in the transition area between the array area and peripheral area, as shown in Fig. 2.15(c). The features on the right of Fig. 2.15(c) are further scaled down from Fig. 2.10. At this stage, the array area looks like a flattened, miniature ziggurat.

A thick layer of oxide is deposited, as shown in Fig. 2.16(a). After oxide CMP, the wafer surface is planarized, as shown in Fig. 2.16(b). The process steps of this module are listed in Table 2.2. Afterwards, the wafer is ready for the next process module.

2.2.3 Channel module

The channel module process is another unique process and also one of the most challenging processes of 3D-NAND. Figure 2.17 illustrates a six-stack 3D-NAND flash memory with strings that consist of a lower SG, an upper SG, and a four-cell memory string between them. It looks like a final product, with the silicon oxide between gates, around channels, and around contact plugs removed so that the polysilicon channels, contact plugs, control gate, and metal lines are exposed in the birds-eye view. The 3D flash-memory chip in high-volume manufacturing has a similar structure, but it has many more memory cells in a string. The existing products have 32 cells in a string and a total of 39 stacks; products with 48- and 64-cell strings are in development.

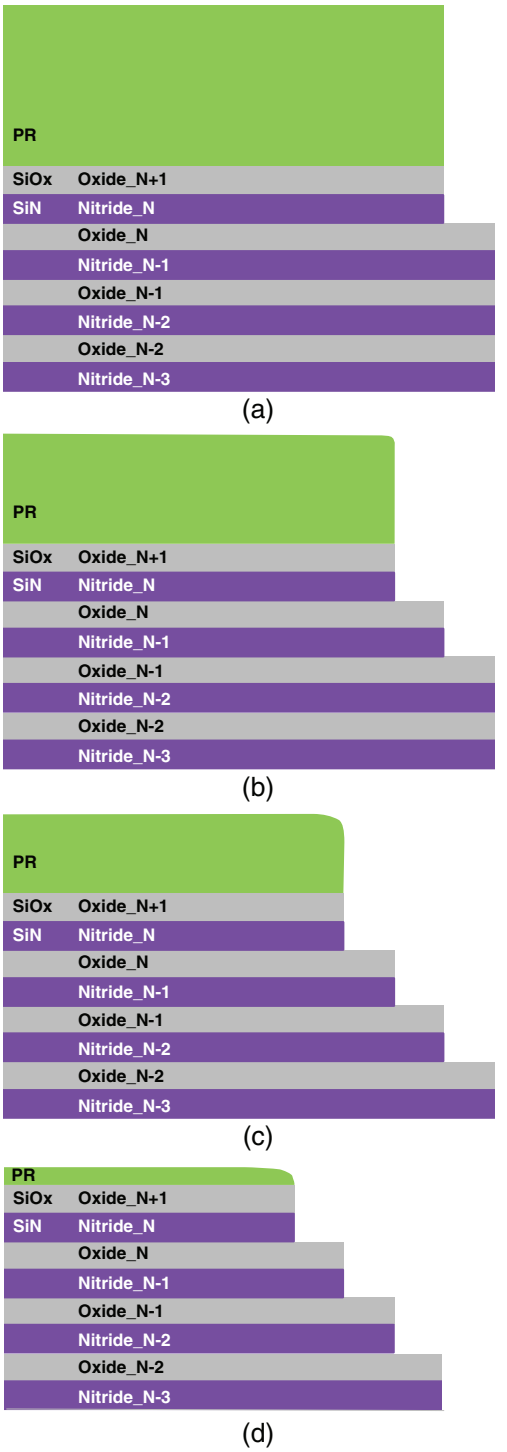


Figure 2.13 Staircase formation: (a) the first-pair, (b) the second-pair, (c) the third-pair, and (d) the fourth pair of oxide/nitride stack etch.

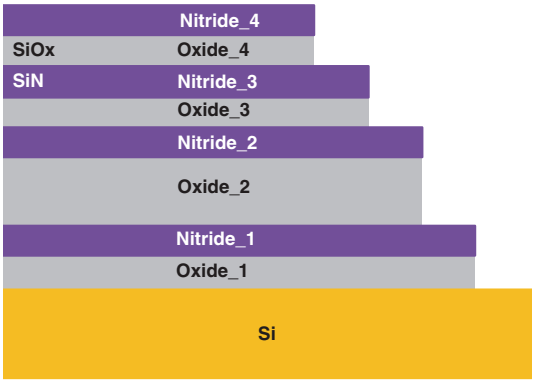


Figure 2.14 Staircase formation (note that it ends with the silicon surface).

This section explains the channel process in detail, focused on the bottom and top of the channel because the cells in the middle of the string are almost the same.

After wafer clean, HM layer deposition, and PR coating, the channel hole mask is applied, and channel holes are etched first on the HM, which is usually amorphous carbon, and then the HM pattern is used to etch through the ONON multi-layers in one etch process. The channel hole is very deep (usually > 3 μm), and the aspect ratio could be $> 30{:}1$. Figure 2.18(a) illustrates a channel hole mask, Fig. 2.18(b) is the close-up of the rectangular box in Fig. 2.18(a), and Fig. 2.18(c) shows the cross-section along the solid line in the lower part of Fig. 2.18(b) after etch, PR strip, and clean. The scale of the hole size and hole pitch in Figs. 2.18(a) and 2.18(b) is too large and not proportional to the multi-layer thickness.

Channel hole etch is a very challenging process. It needs to etch multiple silicon-oxide and silicon-nitride stacks to form a deep, straight hole with well-controlled top and bottom CDs. The channel width of the gate-all-around MOSFET and flash cells is determined by the circumference of the channel hole, and thus the CD control within the hole, within the die, within the wafer, and wafer-to-wafer are critical to keep device performance and yield consistent. The channel hole must also etch all the way to the silicon substrate; otherwise, the string would lose its connection to the source (ground) and suffer yield loss.

Figures 2.19(a)–(f) are illustrations of the cross-section along the solid line in the center of Fig. 2.18(a). Figure 2.19(a) is a top-down cross-section of the channel hole in the cap oxide, indicated by the dashed lines next to it. This channel hole is surrounded by oxide right after channel hole etch, PR strip, and clean. Figure 2.19(b) is a top-down cross-section of the channel hole in the Nitride_N-2, indicated by the dashed lines. This channel hole is surrounded by nitride at this process stage. Figure 2.19(c) is a close-up illustration of the cross-section at the top of the multi-layers, indicated with a circle in the upper

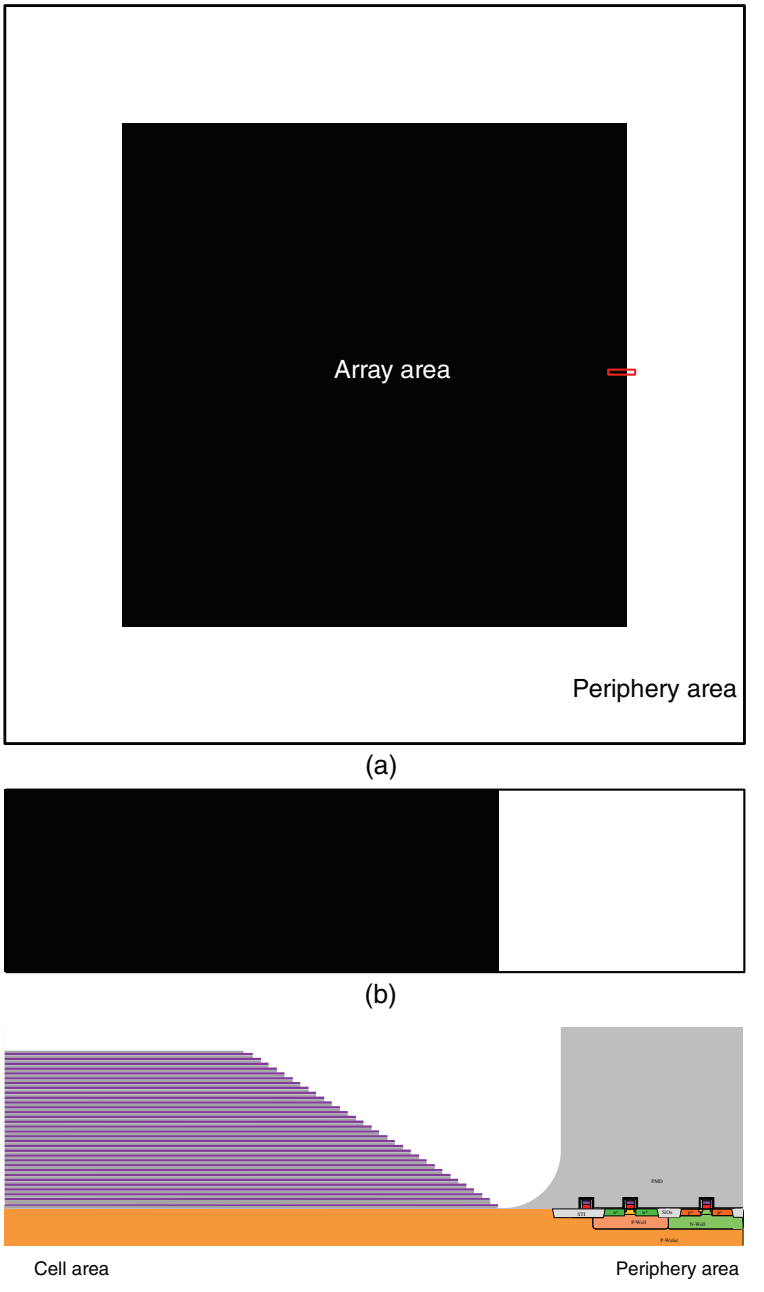


Figure 2.15 (a) Staircase mask, (b) close-up, and (c) cross-section.

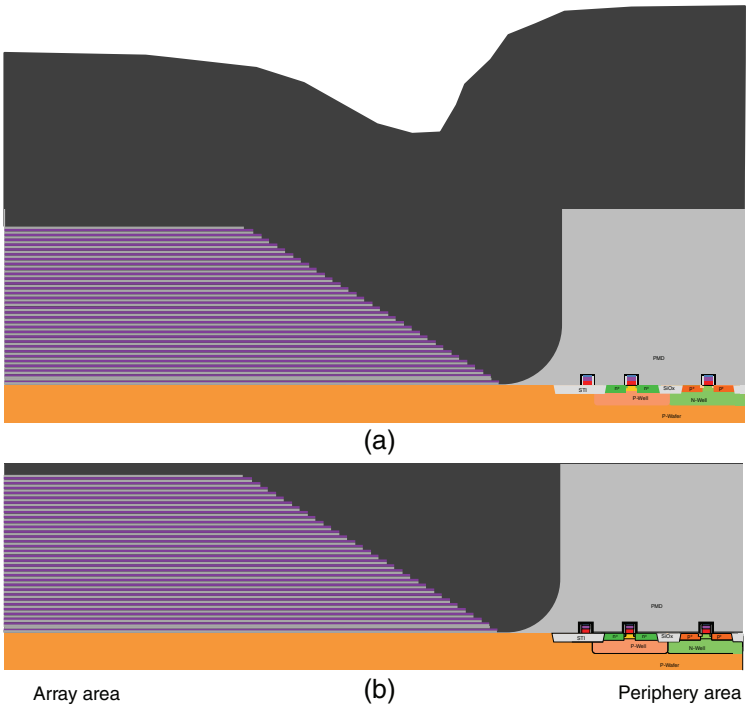


Figure 2.16 (a) Post-staircase oxide CVD, and (b) oxide CMP.

left of Fig. 2.18(c). Figure 2.19(d) is a top-down cross-section of the channel hole in the Nitride_3. This channel hole is surrounded by nitride. Figure 2.19(e) is a top-down cross-section of the channel hole in Nitride_1, which is also surrounded by nitride. This section will eventually become the lower SG, and its later processes are quite different from other devices in other nitride layers. Figure 2.19(f) is a close-up of the cross-section near the bottom of the multi-layers, indicated in the lower left of Fig. 2.18(c) as a rectangular box.

Figure 2.20 illustrates the selective epitaxial growth (SEG) of silicon at the bottom of the channel hole. Figures 2.20(a) and (b) are top-down views of the channel at this process step in Nitride_3 and Nitride_1, respectively. The epitaxial silicon only grows at the bottom of the channel hole, where the single-crystalline silicon is exposed. It is very important to have the SEG silicon within Oxide_2, which is intensively deposited thicker than other oxide layers. If the SEG is too thin, e.g., in Nitride_1, the lower SG will not function correctly. If it is too thick, or if the SEG silicon grows into Nitride_2, a short will occur between the select gate and memory cell. Either result will cause yield loss. Therefore, a pre-SEG clean process between Figs. 2.19(f) and 2.20(c) is critical because any residue at the bottom of the channel hole can cause issues in SEG Si and affect the product yield. It is obvious that during SEG Si the process only affects the bottom of channel

Table 2.2 Multi-layer-deposition and staircase-formation process step.

Array area mask [Fig. 2.11(a) and (b)]	Photoresist trimming
Etch oxide and barrier nitride	Etch Oxide_N-1/Nitride N-2 and stop on Oxide_N-2 [Fig. 2.13(c)]
PR strip and clean [Fig. 2.11(c)]	Repeating trimming and O/N pair etch
CVD Oxide_1, CVD Nitride_1, and the lower SG nitride	PR strip and clean
CVD Oxide_2, CVD Ntride_2, and cell nitride [Fig. 2.12(a)]	Second staircase mask
CVD Oxide_3/Nitride_3 pairs	Repeating trimming and O/N pair etch
Repeating the process until Oxide_N/Nitride_N [Fig. 2.12(b)]	Third staircase mask
CVD Oxide_N + 1 and cap oxide	Repeating trimming and O/N pair etch
First staircase mask	Etch Oxide_1, stop on silicon [Fig. 2.14]
Etch Oxide_N+1/Nitride_N, stop on Oxide_N [Fig. 2.13(a)]	PR strip and wafer clean [Fig. 2.15(a)]
Photoresist trimming	Oxide CVD [Fig. 2.15(b)]
Etch Oxide_N/Nitride_N-1 and stop on Oxide_N-1 [Fig. 2.13(b)]	Oxide CMP [Fig. 2.15(c)]

holes and will not affect the top layers, which is why a cross-section of the top layers is not provided here.

After SEG Si is deposited at the bottom of the channel, a thin, conformal, high-*k* dielectric layer, charge trap layer (silicon nitride), and gate oxide are also deposited. Figure 2.21(a) shows the top-down view of the channel hole in cap oxide. Three layers are drawn in the channel holes, gate oxide, nitride charge trap layer, and high-*k* dielectric. A real product may have more layers, depending on the requirements of the device performance, product yield, and produce reliability. Figures 2.21(a) and 2.21(b) are top-down views of the channel hole in cap oxide and nitride N-1, respectively. Figure 2.21(c) is the cross-section of the channel hole near the top surface of the multi-layers after the deposition of channel dielectric layers. Figures 2.21(d) and 2.21(e) are the

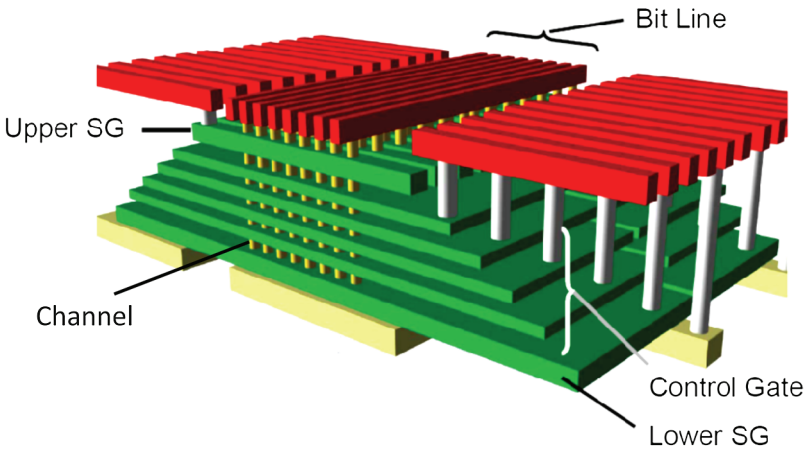


Figure 2.17 3D view of a 3D-NAND flash memory with a four-cell string and six stacks. Source: Ref. 5, reprinted with permission from IEEE.

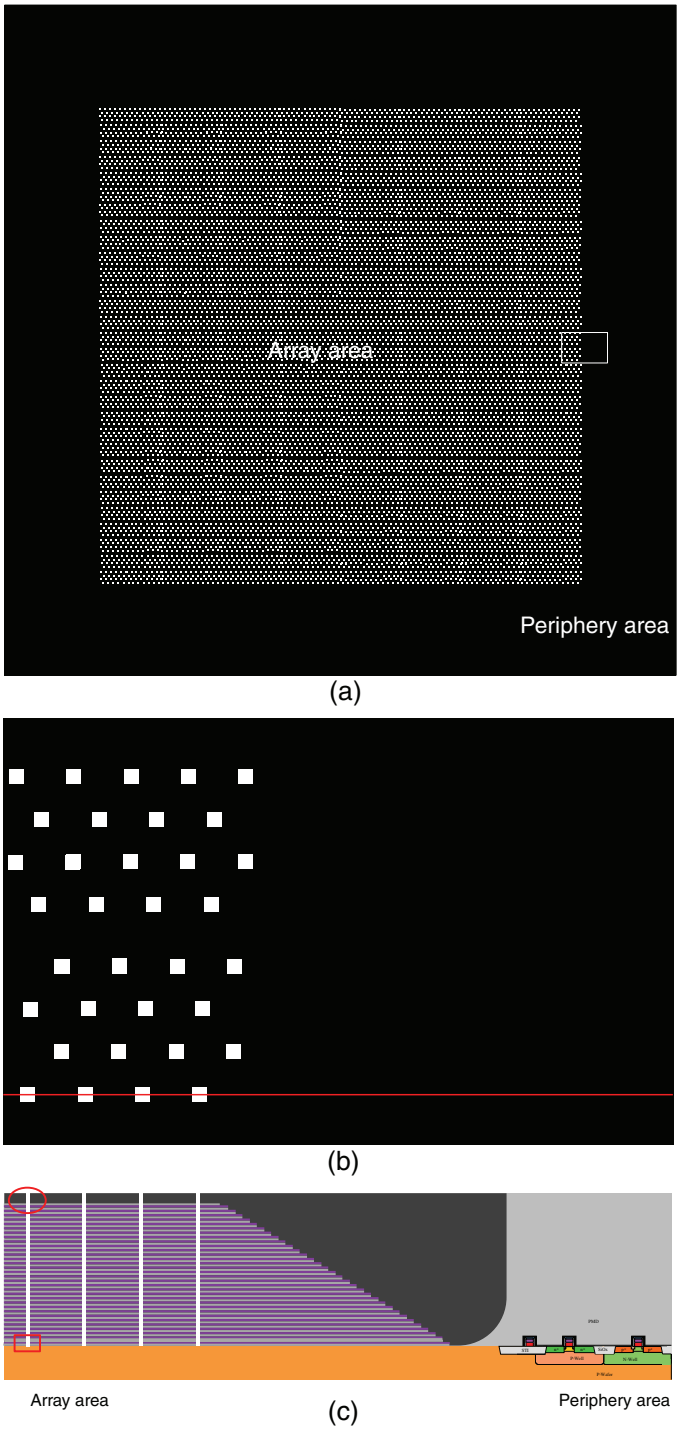


Figure 2.18 (a) Channel mask in a cell, (b) close-up of the box in (a), and (c) cross-section along the solid line.

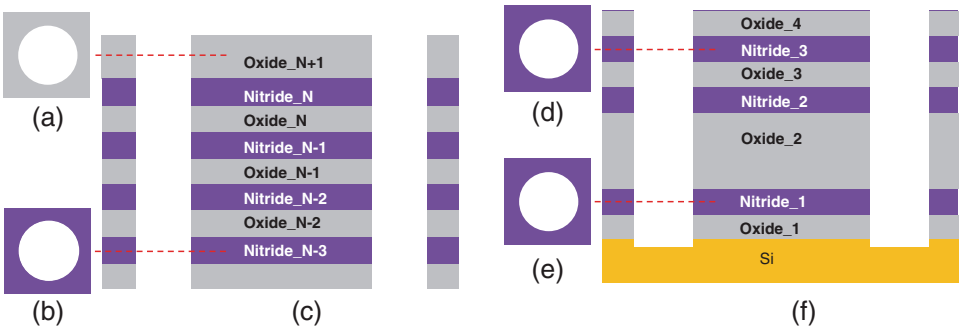


Figure 2.19 (a) Top-down cross-section of a channel hole in cap oxide and (b) in Nitride_N-2, and (c) cross-section along the channel; (d) top-down cross-section in Nitride_3 and (e) Nitride_1, and (f) the cross-section along the channel at the bottom of the channel hole.

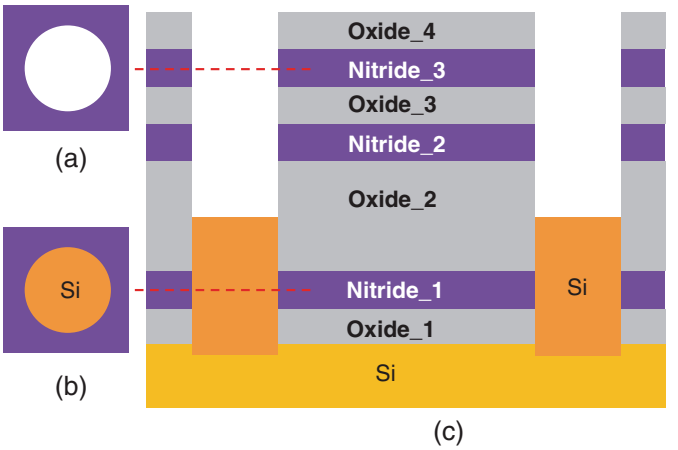


Figure 2.20 SEG Si at the channel bottom: top-down view in (a) Nitride_3 and (b) Nitride_1, and (c) cross-section along the channel.

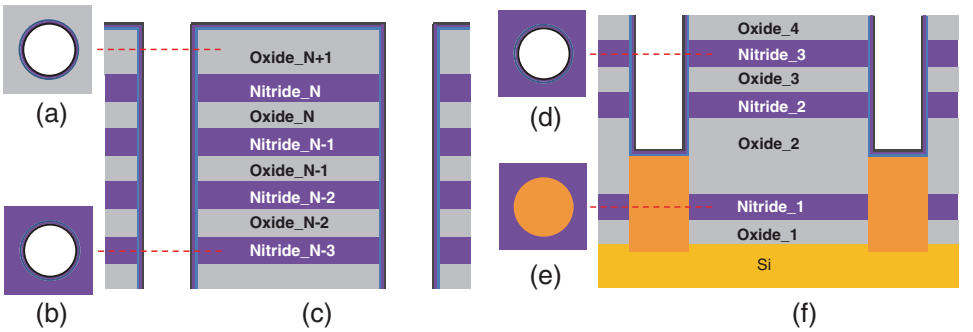


Figure 2.21 Wafer after channel dielectric deposition: top-down view (a) in cap oxide and (b) in nitride_N-2, and (c) cross-section along the channel near the top; (d) top-down view in nitride_3 and (e) nitride_1, and (f) cross-section at the bottom.

top-down view of channel hole in nitride_3 and nitride_1, respectively. Figure 2.21(f) is the cross-section along the channel near the bottom. Dielectric layers block the channel at this stage, and thus an etch process is needed to remove the layers at the bottom to ensure the electrical connection of the channel and substrate.

After the conformal deposition of channel dielectrics, a vertical etchback process removes channel dielectrics from the bottom of the channel hole and from the wafer surface. This process is very similar to the dielectric etchback process that forms the sidewall spacer, which leaves channel dielectric films on the sidewall of the channel holes from the middle of oxide_2 and above. Figure 2.22 shows the cross-sections and top-down view at different layers after etchback. A layer of polysilicon is deposited after wafer clean; this layer is the channel of the gate-all-around flash memory cells and the channel of the upper SG that connects to the bit line. The upper select gate will be formed in nitride_N, as shown in Fig. 2.23(c). Near the bottom of the channel, the polysilicon layer is connected to the SEG Si, which is the channel of the lower

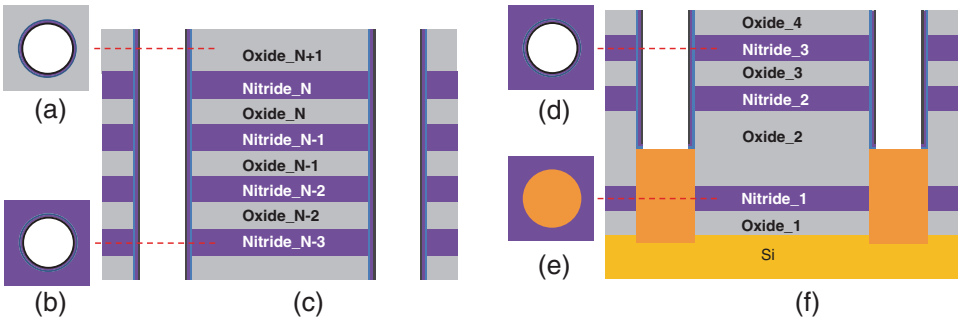


Figure 2.22 Channel dielectric etch back: (a) top-down view in cap oxide, (b) in nitride_N-2, and (c) cross-section along the channel near the top; (d) top-down view in nitride_3 and (e) nitride_1, and (f) cross-section along channel at the bottom.

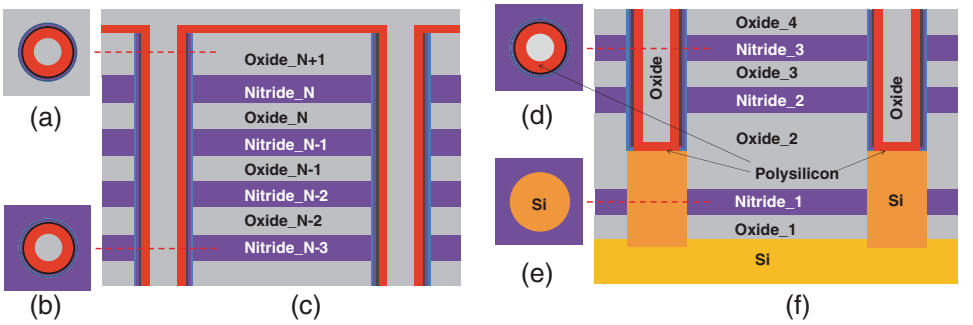


Figure 2.23 Top-down view (a) in cap oxide, (b) in nitride_N-2, and (c) cross-section along the channel near the top; (d) top-down view in nitride_3 and (e) nitride_1, and (f) the cross-section along the channel at the bottom.

SG that connects to the source line (ground), as shown in Fig. 2.23(f). After polysilicon-channel-layer deposition, the channel holes are filled with silicon oxide. After oxide and polysilicon CMP (Fig. 2.23), the channel module is finished.

After polysilicon and oxide deposition, oxide is etched back and into the channel. The oxide etch back needs to be stopped before it reaches the nitride_N; otherwise, the upper select gate will not be functional. After oxide recess, polysilicon is deposited into the channel where the oxide was recessed. After polysilicon CMP, an oxide layer is deposited to cap the channels and prepare the next process module. Figure 2.24(a) shows the top-down view of the polysilicon plug in oxide_N + 1, and Fig. 2.24(b) is a top-down view of the channel in nitride_N-2. The cross-section along the channel is shown in Fig. 2.24(c). Because the polysilicon contact plug formation happens only at the top of channel, the bottom portion of the channel is not shown in Fig. 2.24. Figure 2.25 shows the cross-section of channel in the multi-layer

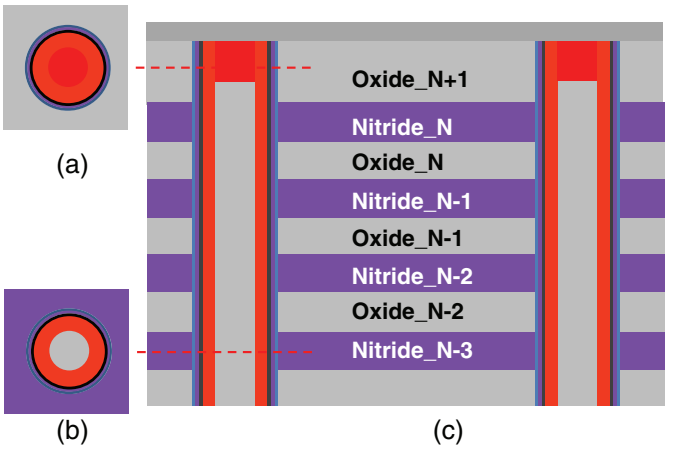


Figure 2.24 (a) Top-down view of a polysilicon plug (b) in nitride_N-2, and (c) cross-section near the top of the channel.

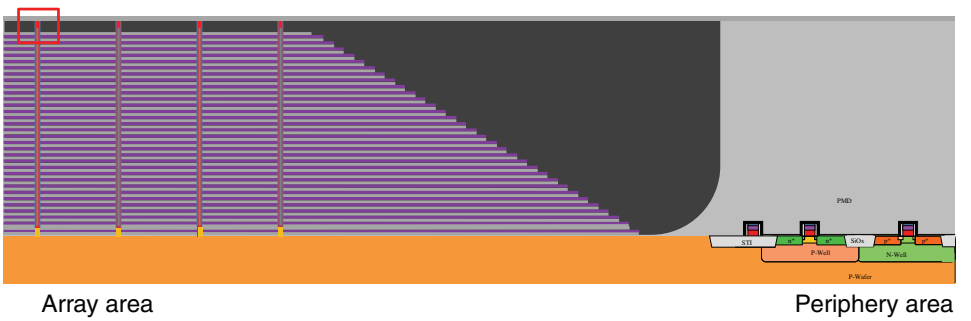


Figure 2.25 Filled-channel strings in the multi-layers in an array area.

Table 2.3 3D NAND channel formation process steps.

Channel mask [Fig. 2.18(a)]	Etch back channel dielectric layers (Fig. 2.22)
Etch hard mark	Wafer clean
Etch multi-layers	Deposit polysilicon channel
Remove hard mask and wafer clean [Figs. 2.18(b) and 2.19]	Deposit silicon oxide filler (Fig. 2.23)
SEG Si (Fig. 2.20)	Oxide recess
Deposit high-*k* dielectric	Deposit polysilicon
Deposit charge trap nitride	Polysilicon CMP
Deposit gate oxide (Fig. 2.21)	Post-CMP clean (Fig. 2.24)

stacks in the array area. Figure 2.24(c) shows a detailed drawing of the box in the upper left of Fig. 2.25. The channel module processes are listed in Table 2.3.

2.2.4 Isolation module

In this module, isolation trenches are etched into an ONON multi-layer to allow etchants to selectively remove silicon nitride and thus allow precursors to reach the channel pillar and deposit the metal gate of the select gates and flash cells. It also allows a conducting metal, such as tungsten, to fill the gaps left by the removal of nitride in the array area to form the word line and, in the staircase area, the word-line contact land areas. After forming the vertical channels in the multi-layers, a thick, hard mask layer (usually a ~1-μm amorphous carbon layer) is deposited on the wafer surface. The isolation trench mask, as shown in Fig. 2.26, is applied, and the hard mask is etched. The main etch process etches through the ONON multi-layers, only stopping when it reaches the silicon substrate to form the deep (~3 mm) and high-aspect-ratio (~30:1) trenches. After the etch process, the hard mask is removed, and the wafer is cleaned.

Figure 2.27(a) shows the isolation mask overlapping with the channel mask. Figure 2.27(b) is the cross-section along the solid line in Fig. 2.27(a) after wafer clean. Figure 2.27(c) is the cross-section along the dashed line in Fig. 2.27(a). Three locations are featured to illustrate the process steps for this module: the top portion indicated by the circle in Fig. 2.27(c), the bottom portion indicated by the rectangular box in Fig. 2.27(b), and the top portion of the staircase indicated by a rectangular box in Fig. 2.27(b).

Figure 2.28(a) is the top-down view of the channel contact plug in the cap oxide, Fig. 2.28(b) is the top-down view of a channel cell, and Fig. 2.28(c) is the cross-section near the top of the multi-layers, which is indicated by a circle in Fig. 2.27(c). Figure 2.28(d) is the top-down view of the channel in nitride_3; Fig. 2.28(e) is the top-down view of the lower SG channel, which is SEG Si, in nitride_1; and Fig. 2.28(f) is the cross-section near the bottom of the multi-layers, indicated by a rectangular box in Fig. 2.27(c). Figure 2.28(g) is the cross-section near the top of the staircase, indicated in the rectangular

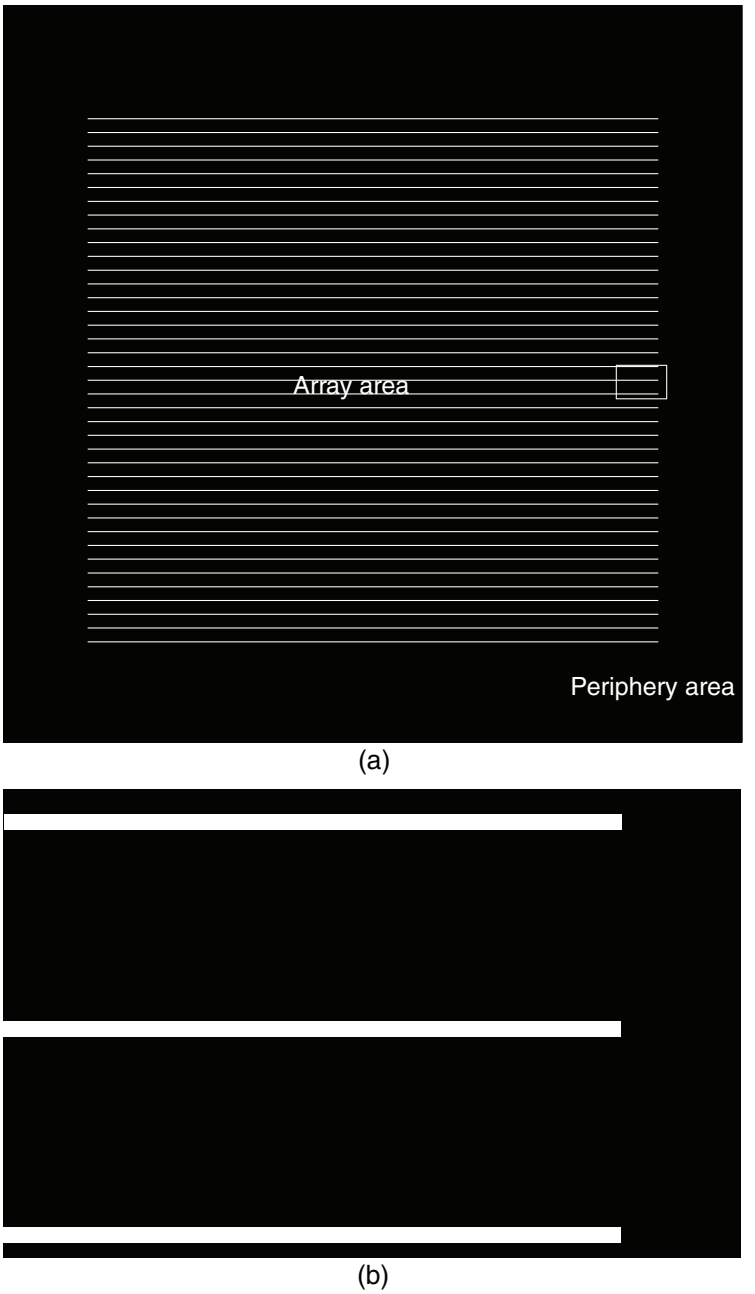


Figure 2.26 (a) An isolation mask in one cell and (b) close-up of the box.

box in Fig. 2.27(b). Figure 2.28(h) is the cross-section of the whole stack perpendicular to the isolation trench. The oval box in Fig. 2.28(h) is shown in detail in Fig. 2.28(c), and the rectangular box in Fig. 2.28(h) is shown in detail in Fig. 2.28(f).

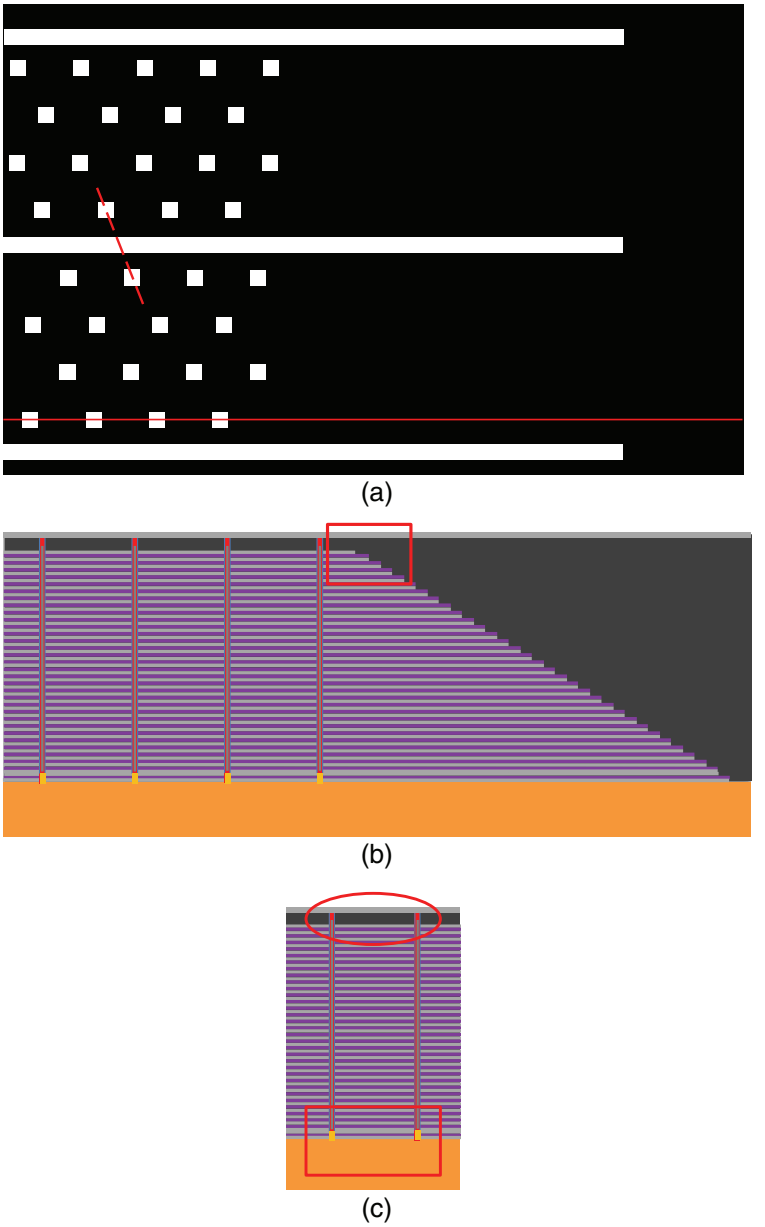


Figure 2.27 (a) Isolation mask overlapped with the channel mask; (b) cross-section along the solid line in (a) after trench etch, hard mask removal, and wafer clean; and (c) cross-section along the dashed line.

After wafer clean, a highly selective etch process is used to remove all silicon nitride layers from the multi-layers with minimum loss of silicon oxide and silicon. Etchants reach every nitride layer through the isolation trenches, and the etch by-products can be removed from these layers through those

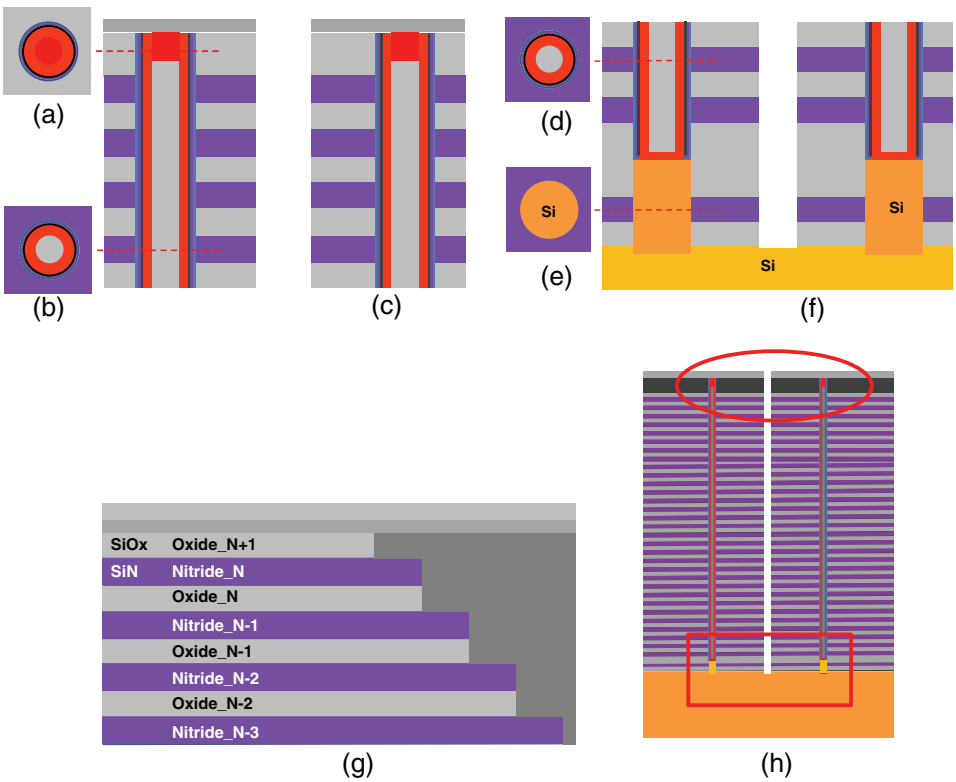


Figure 2.28 Top-down view (a) in cap oxide and (b) in nitride_N-2 (b); (c) cross-section along the dashed line in Fig. 2.27(a); top-down view (d) in nitride_3 and (e) nitride_1; (f) cross-section along the channel at the bottom; (g) close-up of the box in Fig. 2.27(b); and (h) cross-section of the whole stack.

trenches. After nitride removal, only oxide layers are left in the array area, which are supported by the pillar of filled channels. The array area at this stage looks just like an unfinished skyscraper almost 40 stories tall, with an array of pillars supporting the empty floors.

The next wafer is cleaned, and a thermal oxidation process is performed, which forms the gate oxide around the SEG Si pillars in the gap that used to be nitride_1. In other gaps formed after removal of the nitride layers, the channel polysilicon is protected by charge trap nitride and high-*k* dielectric layers; therefore, there is no poly oxidation of channel polysilicon in these gaps. Figure 2.29(a) shows the top-down view of the channel contact plugs, and Fig. 2.29(b) shows the top-down view of the channel in the gap that used to be nitride_N-2. Figure 2.29(c) is the cross-section near the top of the multi-layer with the channel pillars and isolation trench. Figure 2.29(d) is the top-down view of the channel in the gap that used to be nitride_3. In the top-down view of Fig. 2.29(e), the gate oxide is grown around the SEG Si pillar in the gap that used to be nitride_1. Figure 2.29(f) is the cross-section near

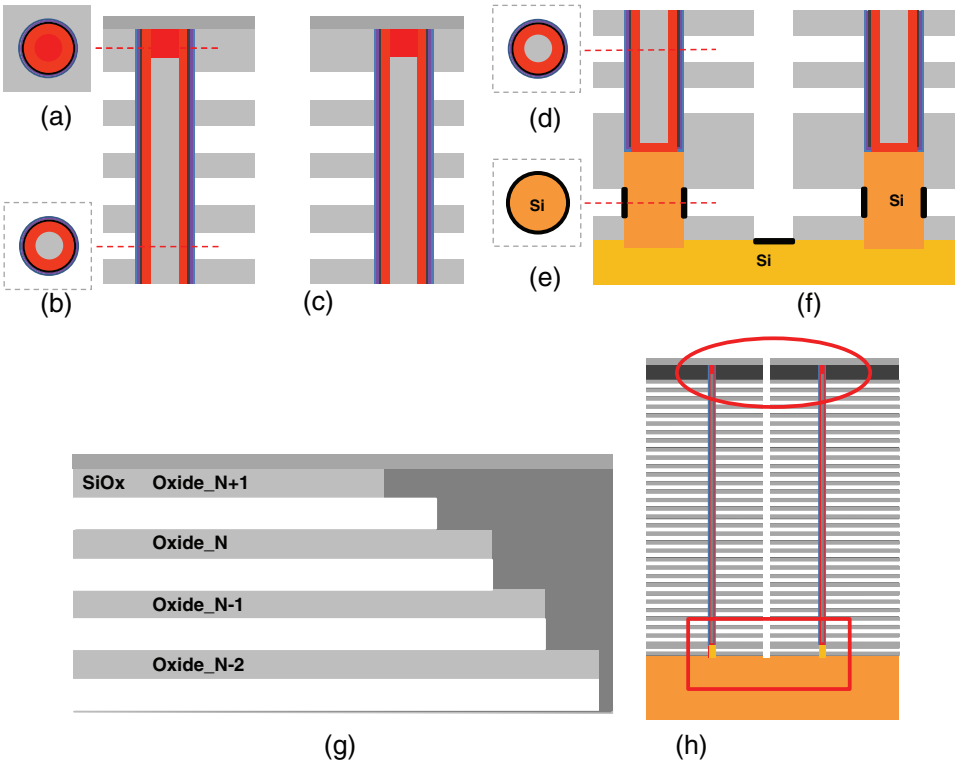


Figure 2.29 Top-down view (a) in cap oxide and (b) in nitride_N-2; (c) cross-section near the top of the channel; top-down view (d) in nitride_3 and (e) nitride_1; (f) cross-section along the channel at the bottom; (g) cross-section at the top of the staircase; and (h) cross-section of the whole stack.

the bottom of the multi-layers. Figure 2.29(g) is the cross-section along the staircase near the top surface, and the Figure 2.29(h) is the cross-section of the whole stack perpendicular to the isolation trench. The oval box in Fig. 2.29(h) is shown in detail in Fig. 2.29(c), and the rectangular box in Fig. 2.29(h) is shown in detail in Fig. 2.29(f).

After lower SG oxidation, a thin and conformal TiN film is deposited into the gaps, around the channel pillar, and covering the sidewall of the isolation trench. The atomic layer deposition (ALD) process, which deposits the film one atomic layer per circle and has excellent conformality and coverage, can be used for this application. This TiN layer will be used as the gate electrodes of the select gates and 3D-NAND cells. After TiN deposition, the W layer is deposited to fill the gaps and cover the sidewall of the isolation trenches; this layer will be used as the WL and WL pads in the staircase to allow WLC plugs to land on them. Figures 2.30(a) and 2.30(b) illustrate top-down views of the channel in cap oxide and in nitride_N-2, respectively; and Fig. 2.30(c) shows the cross-section near the

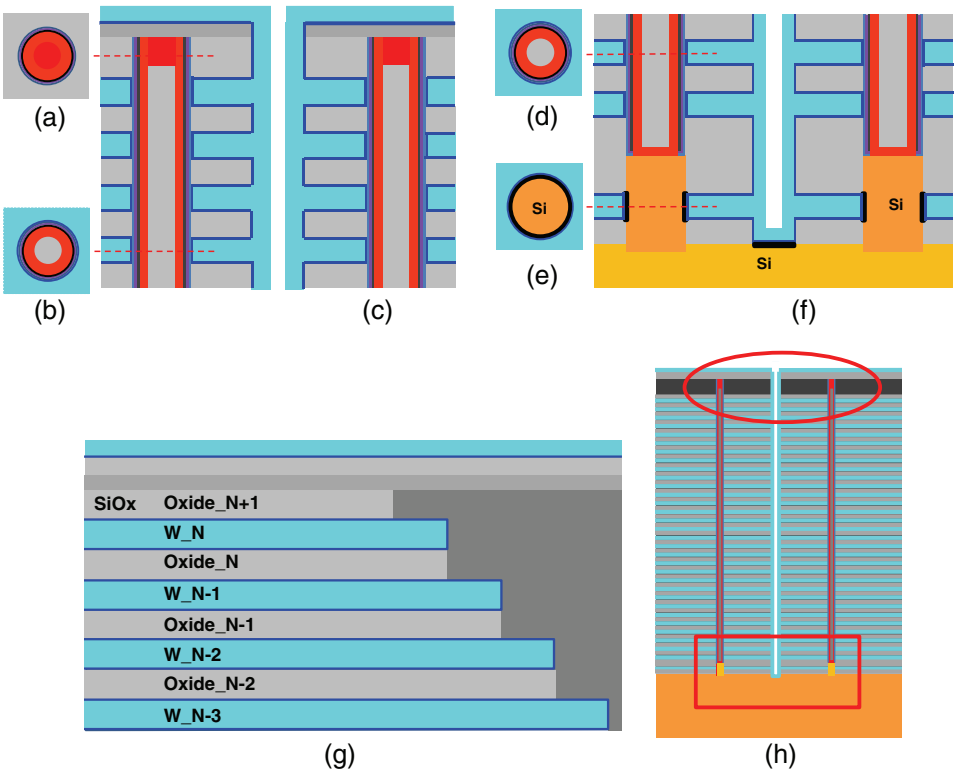


Figure 2.30 Top-down view (a) in cap oxide and (b) in nitride_N-2; (c) cross-section near the top of the channel; top-down view in (d) nitride_3 and (e) nitride_1; (f) cross-section along the channel at the bottom; (g) cross-section at the top of staircase; and (h) cross-section of the whole stack.

top of the multi-layer. Figures 2.30(d) and 2.30(e) show top-down views of the channel in nitride_3 and nitride_1, respectively; and Fig. 2.30(f) is the cross-section at the bottom of the multi-layer. Figure 2.30(g) is the cross-section at the top of the staircase. Figure 2.30(h) is the cross-section of the whole stack appendicular to the isolation trench. The oval box in Figure 2.30(h) is shown in detail in Figure 2.30(c) and the rectangular box in Figure 2.30(h) is shown in detail in Figure 2.30(f).

It is important to control the W film thickness during W deposition. If it is too thin, it will leave voids in the gap between layers. If it is too thick, it could block the next metal etch process, leave W residue on the sidewall of the isolation trench, and cause a short between the conducting layers.

In the next step, chemicals that highly select to W and TiN are used to remove W and TiN on the sidewall of the isolation trenches, with minimum loss of silicon oxide, as shown in Fig. 2.31. At this stage, the devices are formed. The channel contact plugs are formed with polysilicon, surrounded by cap oxide, as shown in Figure 2.31(a). The flash memory cell has a

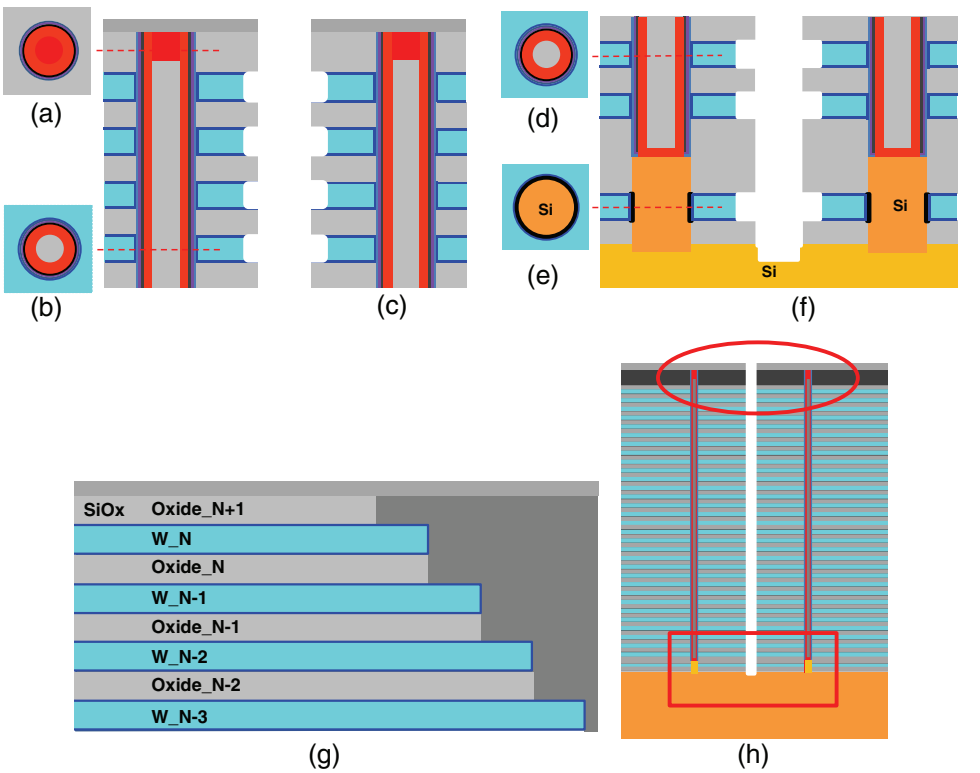


Figure 2.31 Top-down view (a) in cap oxide and (b) in nitride_N-2; (c) cross-section near the top of the channel; top-down view (d) in nitride_3 and (e) nitride_1; (f) cross-section along the channel at the bottom; (g) cross-section at the top of staircase; and (h) cross-section of the whole stack.

polysilicon channel with deposited silicon oxide as the gate dielectric, deposited silicon nitride as the charge trap layer, high-*k* dielectric as the inter-gate dielectric, and TiN as the control gate surrounded by W, as shown in Figs. 2.31(b)–(d). The lower select gate has a SEG Si channel, thermally grown SiO_2 as gate oxide, and a TiN metal gate, as shown in Figs. 2.31(e)–(f). The upper select gate is the same as the flash memory cell: it operates at a different voltage and utilizes the deposited oxide, nitride, and high-*k* dielectric as the gate dielectric, and TiN as the gate electrode. This is one of advantages of using a charge-trap SiN layer instead of a polysilicon floating gate. It is very important to completely remove the W and TiN from the sidewall of the isolation trenches; otherwise, the metals in different layers, i.e., the control gates and WL, will short to the other layer. Figure 2.31(g) is the cross-section at the top of the staircase. Figure 2.31(h) is the cross-section of the whole stack appendicular to the isolation trench. The oval box in Fig. 2.31(h) is shown in detail in Fig. 2.31(c), and the rectangular box in Fig. 2.31(h) is shown in detail in Fig. 2.31(f).

After wafer clean, an oxide layer is deposited into the isolation trenches to seal the metal gates. A vertical etch process follows to remove the oxide layer from the bottom of the isolation trenches and the wafer surface, as shown in Fig. 2.32. This will allow the W wall in the isolation trench to contact the ground and silicon substrate, and provide better electrical isolation between neighboring memory arrays. Figure 2.32(g) is the cross-section at the top of the staircase. Figure 2.32(h) is the cross-section of the whole stack appendicular to the isolation trench. The oval box in Fig. 2.30(h) is shown in detail in Fig. 2.32(c), and the rectangular box in Fig. 2.32(h) is shown in detail in Fig. 2.32(f).

After wafer clean, a TiN liner is deposited into the trenches and covering the wafer surface. A CVD W layer fills the trenches, and a WCMP process removes W and TiN from the wafer surface and forms the W isolation walls. An oxide layer is deposited on the wafer surface to cap the W isolation walls (Fig. 2.33). The process steps are summarized in Table 2.4.

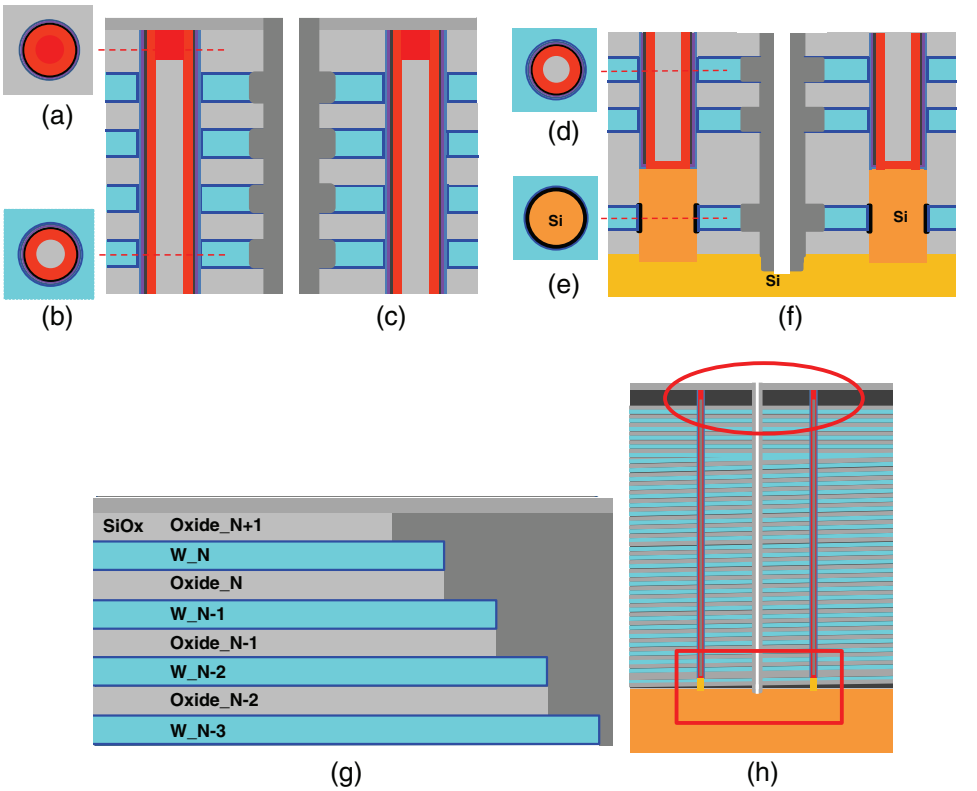


Figure 2.32 Top-down view of a channel (a) in cap oxide and (b) in nitride_N-2; (c) cross-section near the top of the channel; top-down view (d) in nitride_3 and (e) in nitride_1; (f) cross-section along the channel at the bottom; (g) cross-section at the top of the staircase; and (h) cross-section of the whole stack.

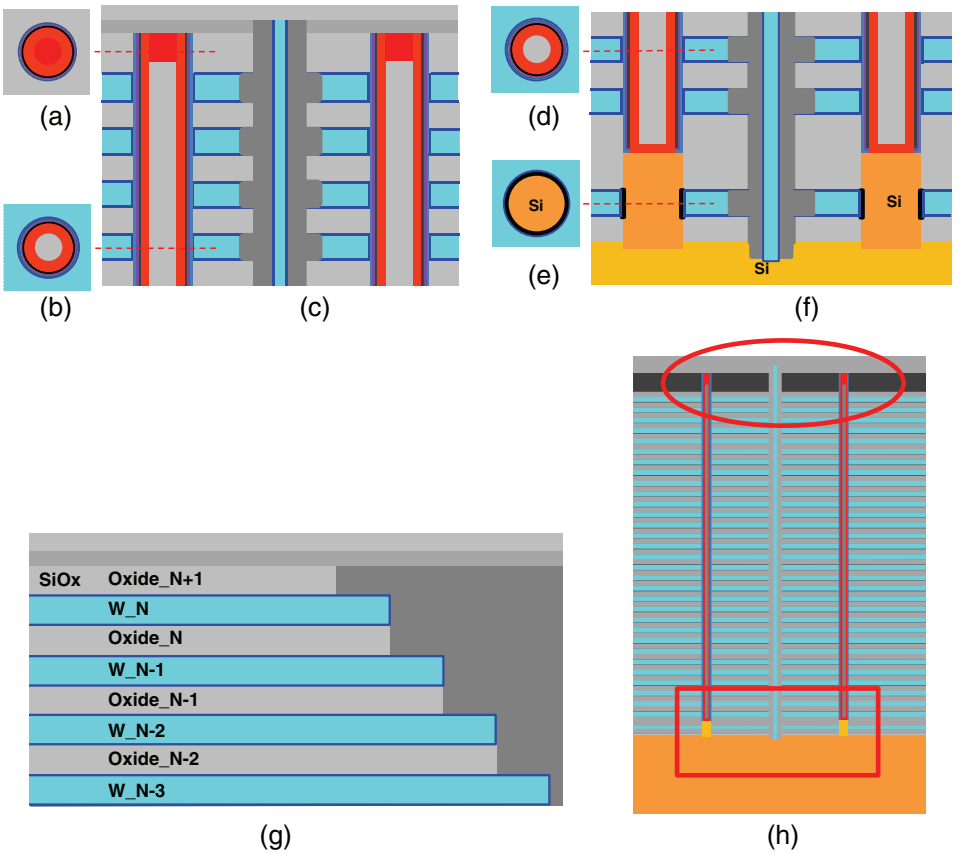


Figure 2.33 Top-down view (a) in cap oxide and (b) in nitride_N-2; (c) cross-section near the top of the channel; top-down view (d) in nitride_3 and (e) nitride_1; (f) cross-section along the channel at the bottom; (g) cross-section at the top of the staircase; and (h) cross-section of the whole stack.

2.2.5 Contact and interconnection modules

At this point, the front-end processes are finished, and all devices in both the peripheral and array areas are built. This section discusses the BEOL processes—contact and interconnect—of 3D-NAND flash. First is the contact module, in which contact holes are etched in the staircase area between the cell and periphery; they land on the W staircases and silicon substrate, creating the connection between the word lines in different layers and the source line in substrate.

After hard mask deposition and photoresist coating, the contact mask is applied [Fig. 2.34(a)]. At first the hard mask is etched and then used to etch contact holes in oxide with etch chemistries that have high selectivity to tungsten [so that the etch process stops whenever the contact holes reaches the tungsten surface in the shallower holes while continuing in the deeper holes, as

Table 2.4 Process steps for the isolation module of 3D-NAND.

Wafer clean	Trench W removal
Isolation mask (Fig. 2.26)	Trench TiN removal (Fig. 2.32)
Etch hard mask	Wafer clean
Etch trenches in ONON multi-layers and stop on silicon	Oxide deposition
Remove hard mask [Figs. 2.27(c) and 2.28]	Oxide etch back (Fig. 2.32)
Remove nitride layers (Fig. 2.29)	TiN deposition
Wafer clean	W deposition
Oxidation of SEG (Fig. 2.30)	W CMP
TiN deposition	Oxide cap deposition (Fig. 2.33)
W deposition (Fig. 2.31)	

shown in Fig. 2.34(b)]. Chapter 9 of Xiao[11] discusses this etch chemistry used for contact etches with different hole depths.

Because the depth of the contact holes vary between different layers, it would be very difficult to etch all of them with almost 40 different depths in one etch process; multiple masks would be necessary. Depending on the process, ~10 different contact-hole depths can usually be etched in one etch process, and thus four masks and four etch processes are needed to etch all of the contact holes in the staircase and periphery of a 32-cell-stack 3D-NAND flash device. Figure 2.34(c) shows the cross-section after staircase contact etch and hard-mask strip and clean.

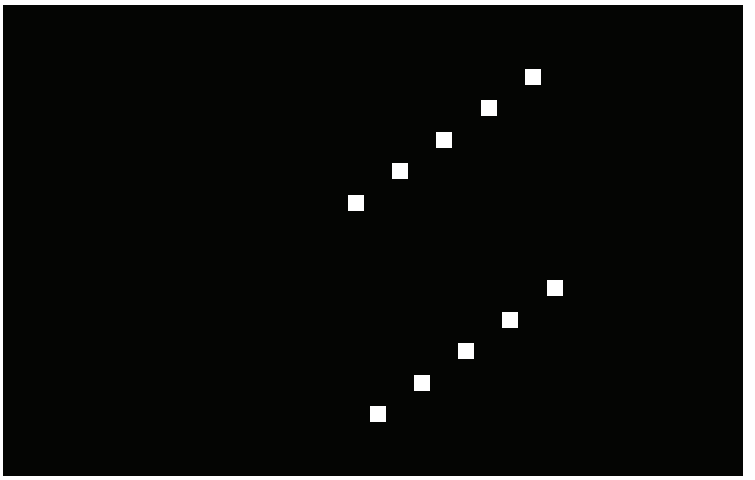
When all of the contact holes have been etched, the wafer is cleaned to remove the polymer residue at the bottom of the contact holes. After sputtering etch removes the native oxide, barrier TiN is deposited [Fig. 2.35(a)], followed by W deposition [Fig. 2.35(b)]. A WCMP process removes W and TiN from the surface [Fig. 2.35(c)], which completes the contact module.

Figures 2.36(a)–(c) illustrate the cross-section of TiN deposition, W deposition and WCMP in cell, staircase and periphery areas, respectively. Figures 2.35(a)–(c) are close-up versions of Figs. 2.36(a)–(c) at the top of the staircases, respectively.

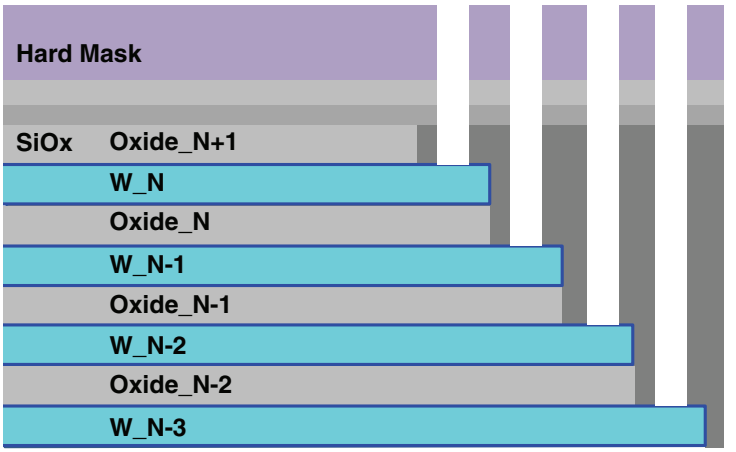
The next module is metal 1, which is a dual-damascene process that forms the local interconnect. It includes two masks, via 1 (V1) and metal 1 (M1). A layer of oxide is deposited to cap the contact W plugs. In the first via process, a V1 mask, which is like a channel mask plus a contact mask (Fig. 2.37), is applied first. Via holes are then etched to land on channel polysilicon plugs and contact tungsten plugs (Fig. 2.38).

Metal 1 (M1) forms the local interconnect in the array area and peripheral area. The pattern in the staircase area is almost the same as V1. Figure 2.39(a) shows the M1 mask, and Fig. 2.39(b) shows the M1 overlaps with the channel, isolation, contact, and V1.

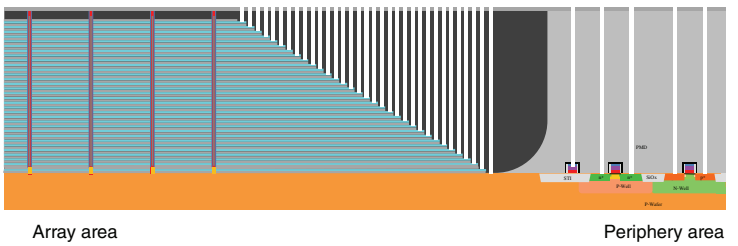
After applying the M1 mask, an oxide etch process is performed to form the trenches of the local interconnect. After photoresist strip and clean, a TiN liner and W are deposited to fill the M1 trenches and V1 holes. WCMP



(a)



(b)



(c)

Figure 2.34 (a) Staircase contact mask, (b) close-up of the top layer, and (c) cross-section of the cell and periphery contacts.

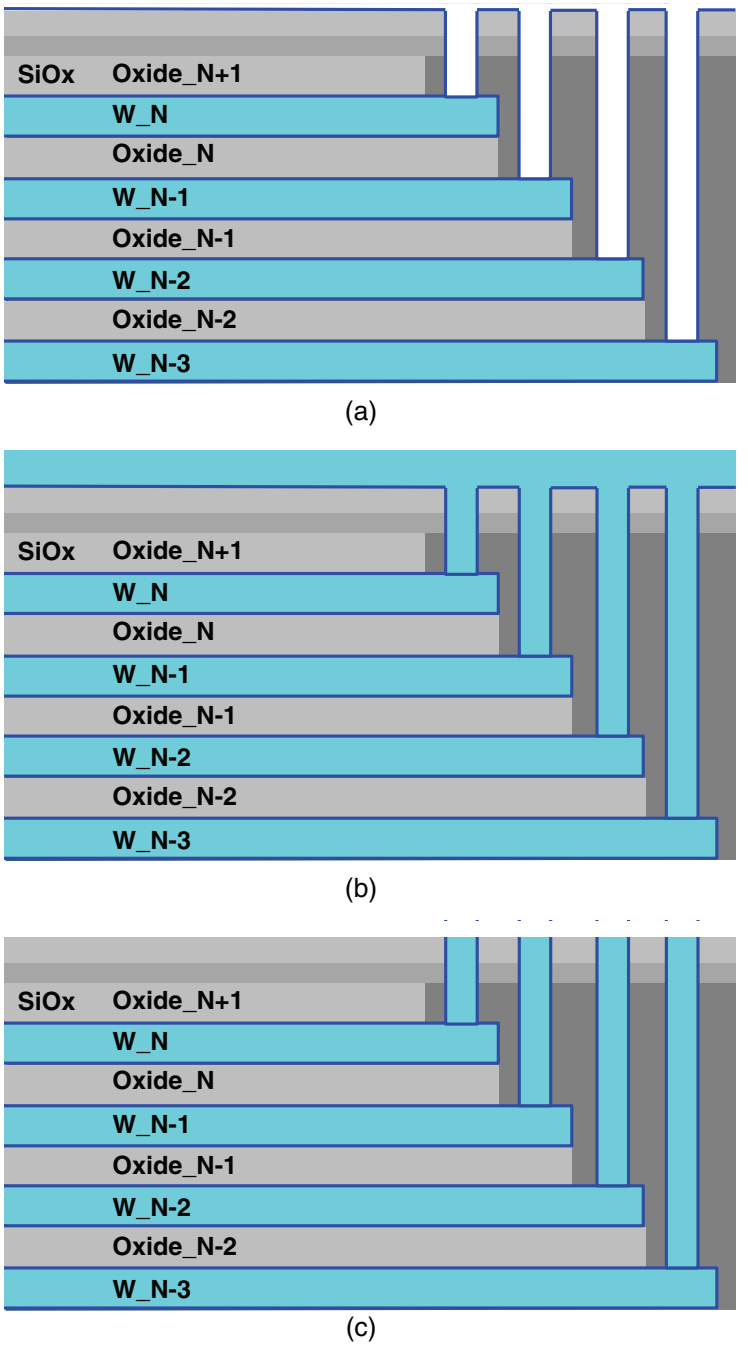


Figure 2.35 Close-up of the contact processes near the top of the staircase: (a) after TiN deposition, (b) after W deposition, and (c) after WCMP.

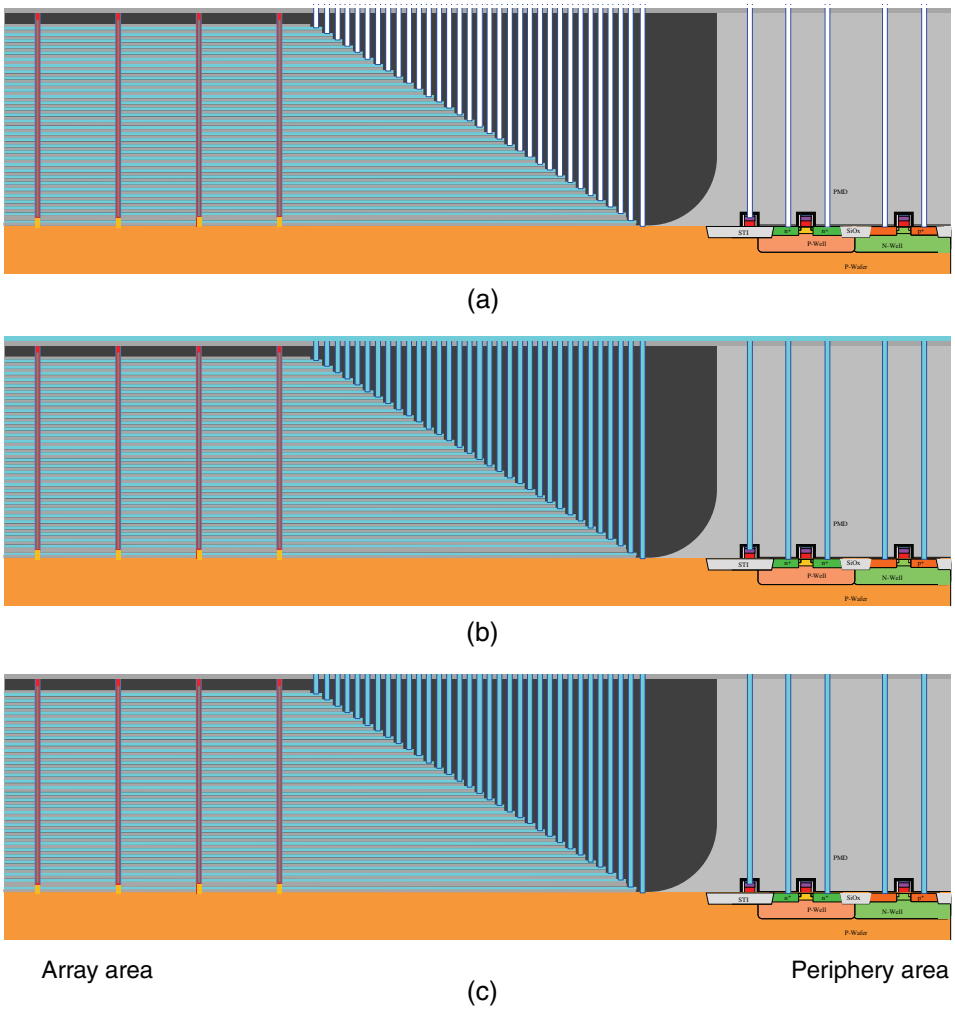


Figure 2.36 Contact process steps: (a) after TiN deposition, (b) after W deposition, and (c) after WCMP.

removes the W and TiN from the wafer surface and forms the W wires and plugs that connect the channel plugs and contact plugs, as shown in Fig. 2.40.

Metal 2 forms the bit line in the array area, source line and word line wires in the staircase area, and interconnection in the peripheral area. Because each WL between the two isolation walls has four rows of channel holes, the bit lines, which are perpendicular to the WL, must be split into a pitch density four times higher than the channel hole pitches to ensure the registration of a single cell in the string with one bit signal and one word signal. After wafer clean and ILD deposition, there is a via 2 (V2) and metal 2 (M2) process (the via 2 mask is illustrated in Fig. 2.41).

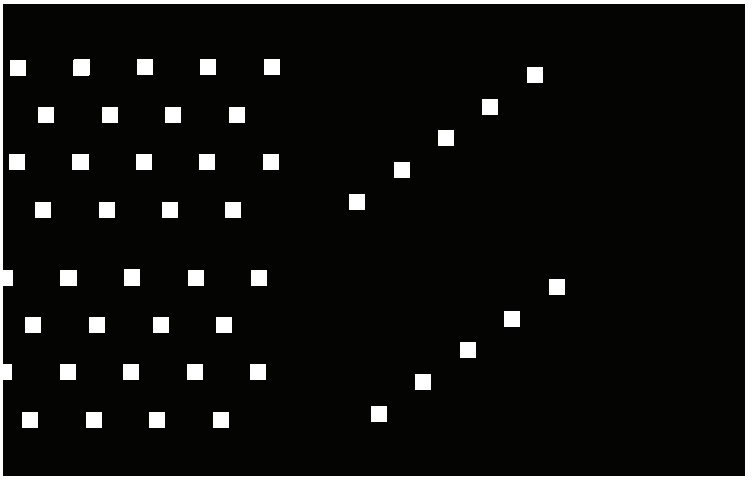
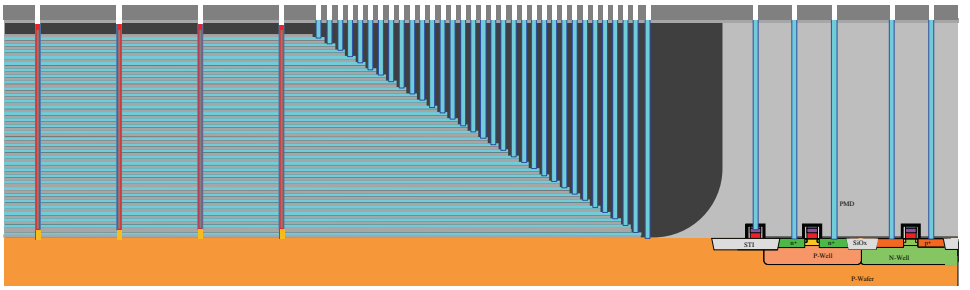


Figure 2.37 V1 mask.



Array area Periphery area

Figure 2.38 V1 etch in the cell, staircase, and peripheral areas.

After V2 etch, photoresist strip, and clean [as shown in Fig. 2.42(a)], a TiN liner is deposited, followed by W deposition [shown in Fig. 2.42(b)], and then WCMP removes the W and TiN on the surface and forms the V2 W-plugs [Fig. 2.42(c)].

After WCMP, another ILD is deposited, a M2 mask is applied, and metal trenches are etched. Figure 2.43 illustrates the M2 mask, and Fig. 2.44(a) shows the M2 etch. After photoresist strip and clean, a TaN barrier layer and Cu seed layer are deposited into the M2 trenches; after copper plating and anneal, a metal CMP process removes the Cu and TaN from the wafer surface and forms the BL in the array area and WL and SL wires in the staircase area [Fig. 2.44(b)].

Metal 3 (M3) is the last metal layer; it forms interconnect and bond pads. It is usually formed by a stack of metals: a Ti barrier layer at the bottom, an Al-Cu alloy bulk layer, and TiN ARC on top. Via 3 (V3) is a tungsten plug that connects to M2.

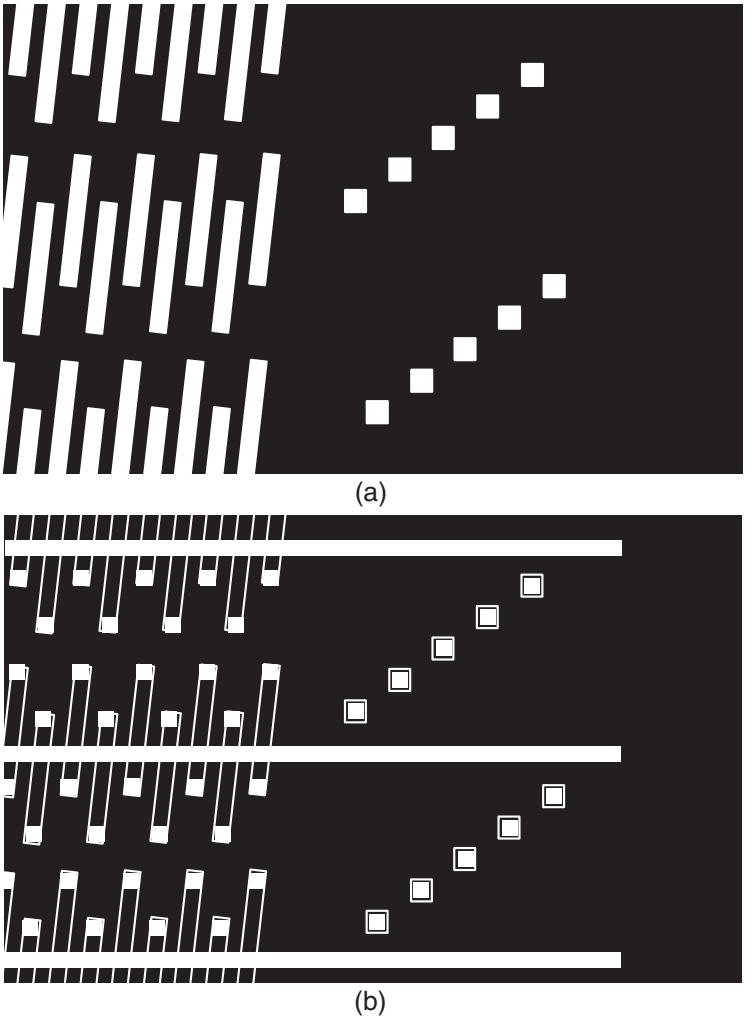


Figure 2.39 (a) M1 mask and (b) M1 overlaps (white outline) with the channel, isolation, contact, and V1 masks.

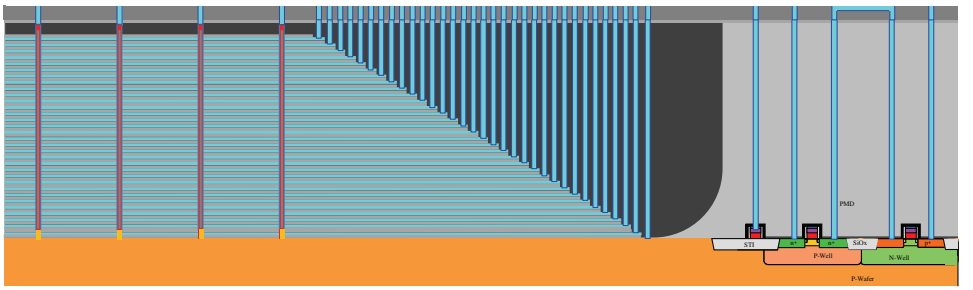


Figure 2.40 V1 etch in the cell, staircase, and peripheral areas.

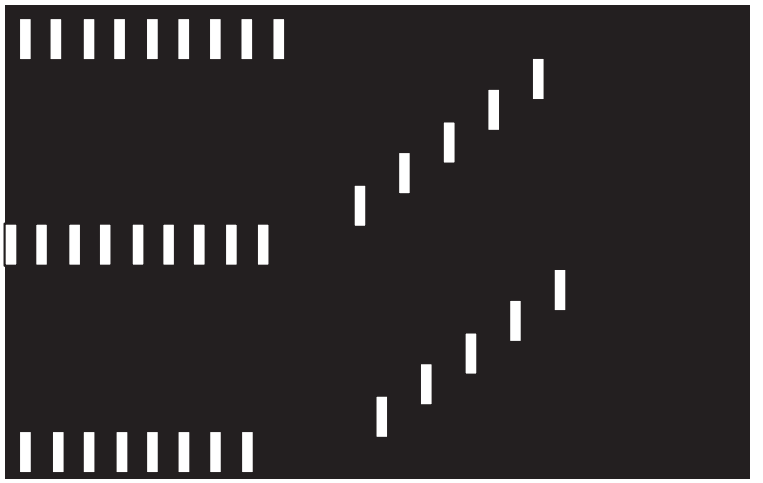


Figure 2.41 V2 mask.

The process starts with dielectric deposition, followed by V3 mask. A dielectric etch process forms via holes that stop on the M2 surface. After photoresist strip and clean, as shown in Fig. 2.45(a), a TiN barrier liner is deposited, followed by WCVD that fills via holes and covers the whole wafer surface. A WCMP process removes the tungsten and TiN from the surface and forms W-plugs of V3 [Fig. 2.45(b)].

After wafer clean, a TiN/Al-Cu/TiN metal stack is deposited on the wafer, usually with a physical vapor deposition (PVD) process, as shown in Fig. 2.46(a). After applying the M3 mask, a metal etch process is performed, and photoresist is usually stripped *in situ* to avoid metal corrosion from moisture absorption and reaction. Figure 2.46(b) shows the cross-section after M3 etch, photoresist strip, and clean that forms the M3 wires.

The last process of 3D-NAND flash manufacturing in an IC fab is the passivation and formation of bond pads. First, passivation dielectrics—usually silicon oxide layers followed by silicon nitride—are deposited with CVD processes. The bond pad mask is applied, and a dielectric etch process removes the nitride and oxide on top of the metal bond pads, which finishes the wafer process of 3D-NAND flash memory chips. Figure 2.47 illustrates the cross-section of finished 3D-NAND flash with cell string channels in the array area, WL/SL contacts plugs that land on the staircase, and peripheral CMOS logic devices with W-plugs and W local interconnect. While all three via layers are formed by TiN/W, they are all different: metal 1 is W/TiN, metal 2 is Cu/TaN, and metal 3 is TiN/Al-Cu/TiN. Because the bond pads are not located near the cell and peripheral areas, they are not shown in the figure. The wafer is now ready for test and packaging. Table 2.5 lists the process steps for the contact and interconnects.

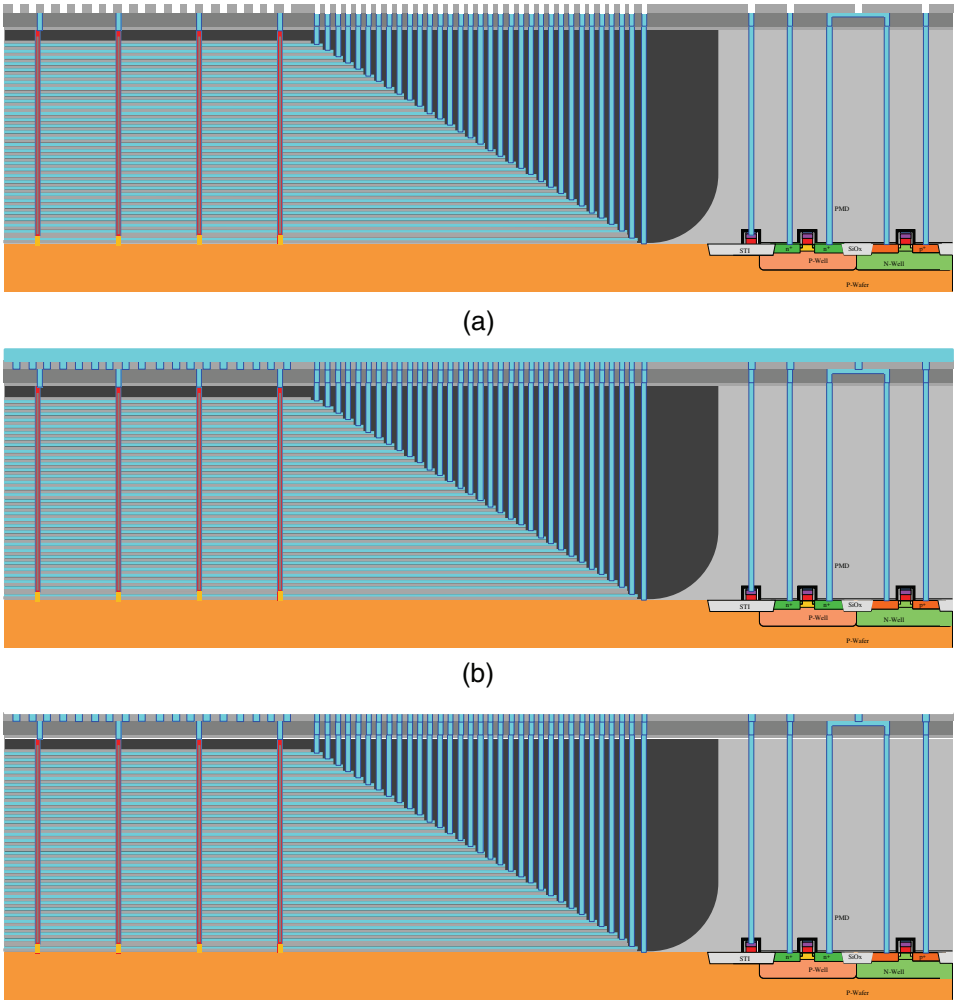


Figure 2.42 V2 process steps in the cell, staircase, and peripheral areas: (a) etch, (b) TiN and W deposition, and (c) WCMP.

2.3 3D-NAND Summary and Discussion

This chapter described the 3D-NAND process in detail. Figure 2.48 shows a 3D illustration of the 3D-NAND after channel, isolation, and contact-plug formation.[13]

The 3D-NAND device described in previous sections is very similar to that illustrated in Fig. 2.48. The cutoff shows cross-sections both parallel to and perpendicular to the isolation trenches. It can help us visualize the staircase, channels, isolation trenches, TiN/W layers that replaced the SiN layer in the multi-layer stack, and W plugs that connect to the WL (W layers)

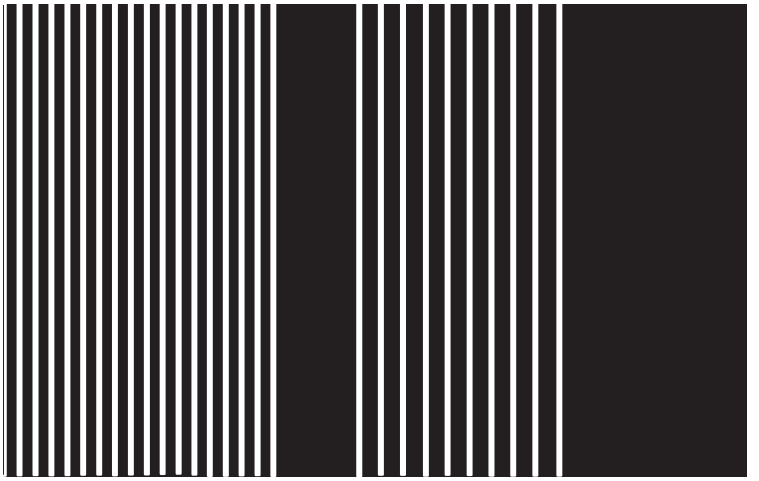
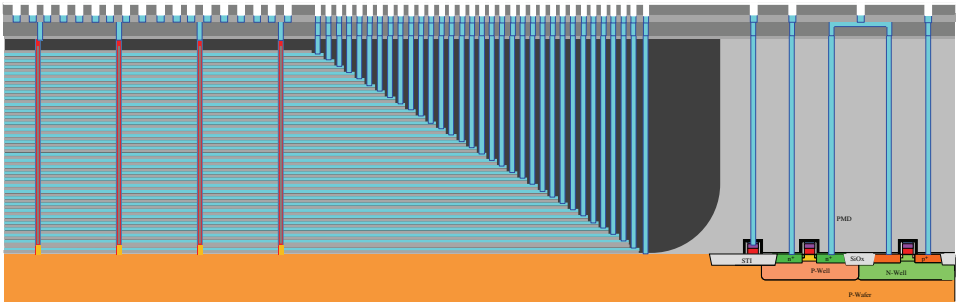
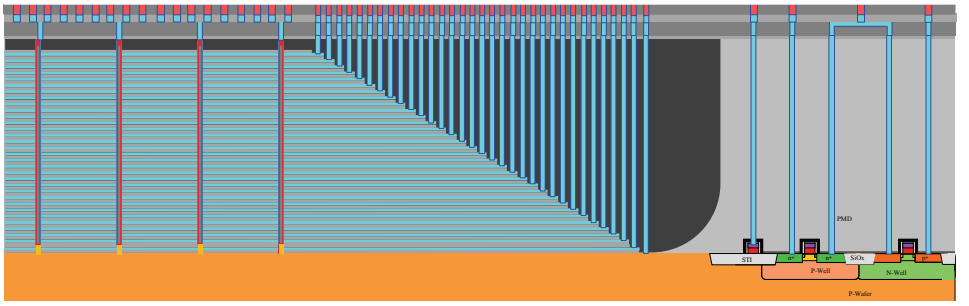


Figure 2.43 M2 mask.



(a)



Array area (b) Periphery area

Figure 2.44 M2 process steps in the cell, staircase, and peripheral areas: (a) etch and (b) CuCMP.

at the staircase. Of course, there are some differences, such as the number of cells in a channel string; Fig. 2.48 shows 17, while the one described earlier has 32. Each WL strip has four rows of channels strings, whereas Fig. 2.48 shows only one row. The isolation trenches in the figure are filled only with oxide,

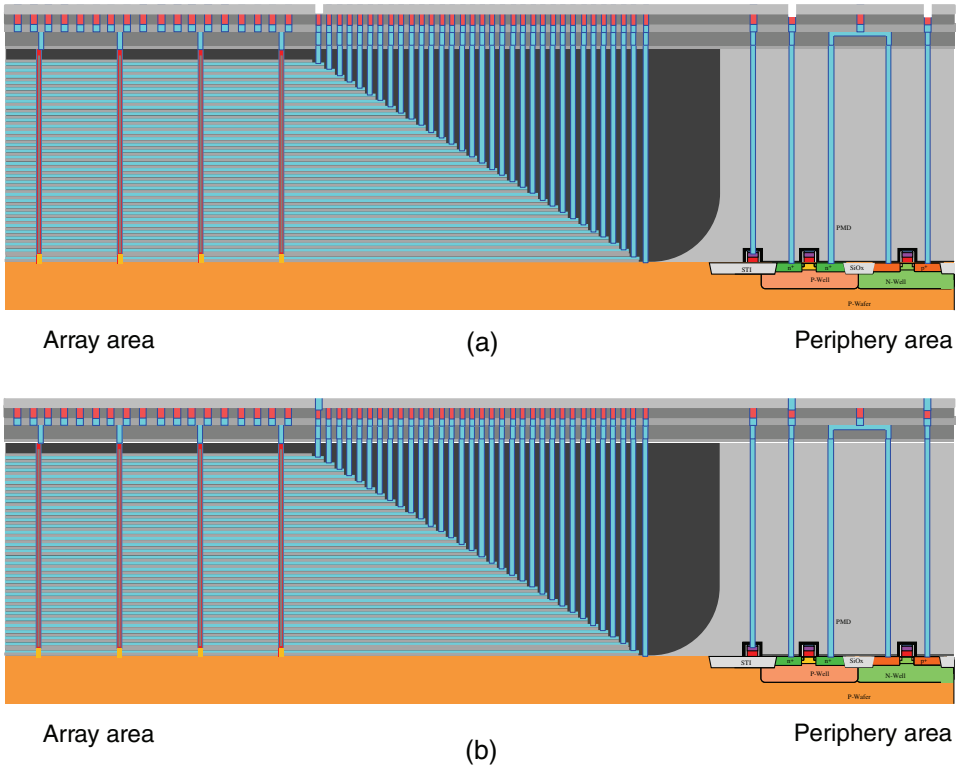


Figure 2.45 V3 process steps in the cell, staircase, and peripheral areas: (a) etch and (b) WCMP.

those in the previous example also have grounded tungsten sheets between oxide layers for better electrical isolation of neighboring word lines.

There are several 3D-NAND approaches. Besides the terabit cell array transistor (TCAT) described previously, two other approaches include pipe-shaped bit cost scalable (P-BiCS) 3D-NAND (Fig. 2.49) and vertical gate (VG) NAND (Fig. 2.50).

In Figure 2.49(a), PC stands for pipe connection; it connects two vertical channels to form a string twice as long. One of benefits of using a PC is that both the source select gate and the bit select gate are on the top of the multi-layer, making it relatively easier to form them. Of course, the pipe connection is not easy to form, and the PC formation process is very hard to monitor.

In Figure 2.50, MLx stands for metal layer x, where x ranges from 1 to 4. CSL stands for common source line, and GSL stands for ground select gate. It is basically planar NAND flash stacked in N levels, with a unit cell size of $4F^2/N$.[14]

Although the majority of 3D-NAND manufacturers use flash cells with a charge-trap layer, some still use floating-gate (FG) memory cells. A FG 3D-NAND multi-layer consists of oxide/polysilicon/oxide/polysilicon (OPOP)

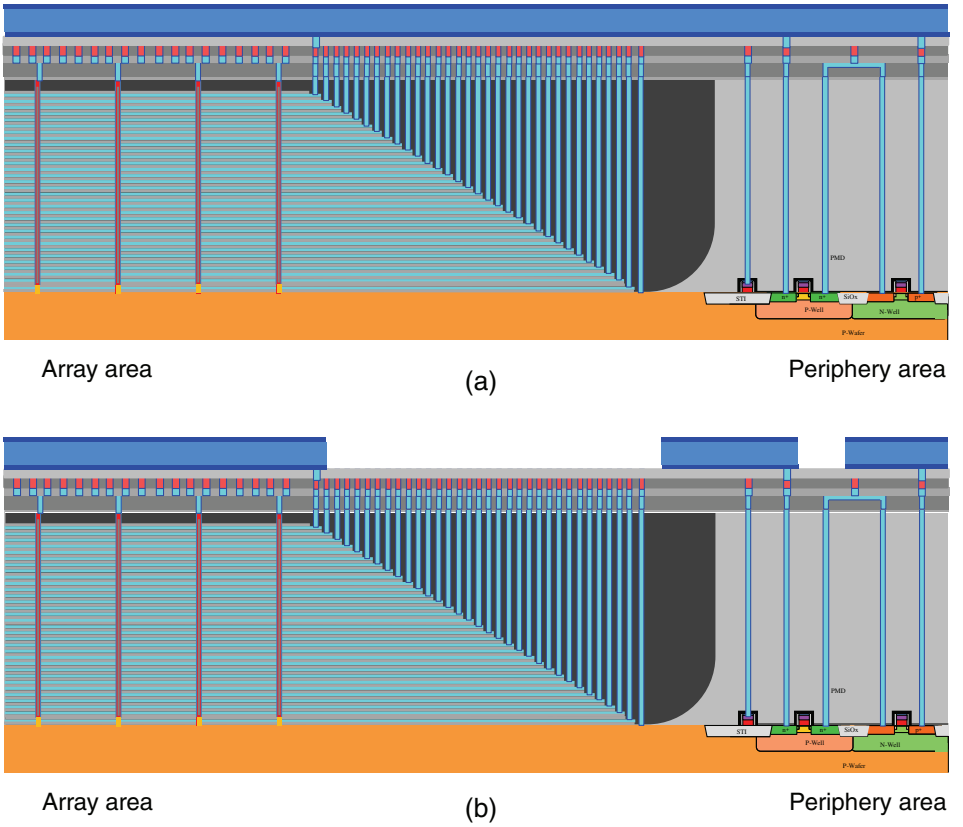


Figure 2.46 M3 process steps in the cell, staircase, and peripheral areas: (a) metal stack PVD and (b) metal etch, PR strip, and clean.

stacks. After channel hole etch and the formation and protection of the Si SEG in the bottom SG channel, the polysilicon layers in the OPOP stack are partially recessed, and an inter-gate dielectric (IGD) layer is conformally deposited with an ALD process. Polysilicon is deposited into the channel hole, filling the rest pockets formed after partial polysilicon recess. An isolation etch process removes the polysilicon on the sidewall of the channel holes and forms the ring-shaped polysilicon FGs. After wafer clean, a gate oxide layer and a channel polysilicon layer are deposited, followed by a vertical etch process to remove the polysilicon, gate oxide, and protection layer at the top of the SEG Si. CVD oxide is used to fill the channel hole. The top select gate (BL select gate) can be formed after oxide CMP and poly, oxide, poly, and IGD wet etch. After cap-oxide layer deposition, isolation trench patterns are etched, and self-aligned silicide can be formed in the trench on the side wall of the polysilicon of the OPOP stack to reduce the resistance of the word line. After the unreacted metal is removed, the trenches are filled, and the FEOL process of FG 3D-NAND is finished.

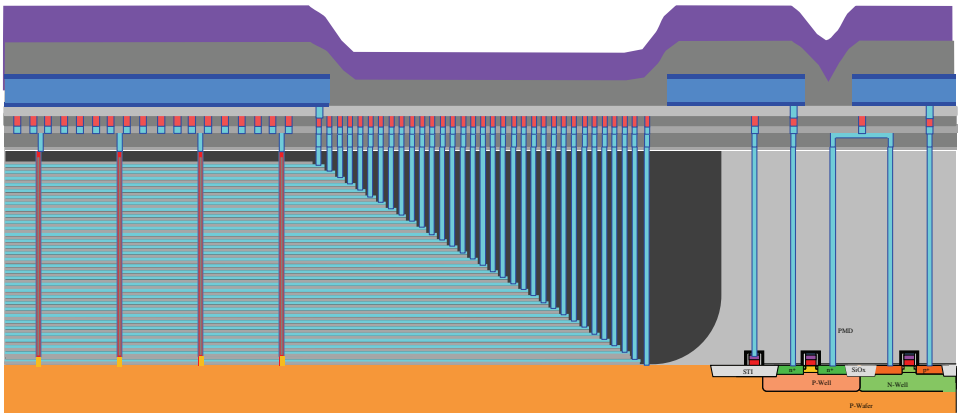


Figure 2.47 Cross-section of a finished 3D-NAND flash memory.

One of the advantages of FG 3D-NAND is that it is ready to form a multi-level cell (MLC) based on mature MLC techniques developed in planar NAND technology. All planar NAND products are FG devices. In contrast, NAND flash with a charge trap layer is usually a single-level cell (SLC). A single-level cell can only have two charging states, 0 and 1. A two-level cell has four charging states: 00, 01, 10, and 11; and a three-level cell has eight

Table 2.5 Process steps of the contact and interconnect module of 3D-NAND.

Wafer clean	TiN deposition, WCVD and WCMP (Fig. 2.40)
First contact mask [Fig. 2.34(a)]	Oxide CVD
Etch hard mask	V2 mask (Fig. 2.41)
Etch shallower staircase contacts	Etch oxide, PR strip, and clean [Fig. 2.42(a)]
Strip PR and wafer clean [Fig. 2.34(b)]	TiN deposition, WCVD [Fig. 2.42(b)], and WCMP [Fig. 2.42(c)]
Apply the second contact mask and etch staircase contacts	Oxide CVD
Strip PR and wafer clean	M2 Mask (Fig. 2.43)
Repeating staircase contact litho, etch and clean.	Etch oxide, PR strip, and clean [Fig. 2.44(a)]
Remove hard mask and wafer clean [Fig. 2.34(c)]	TaN deposition, Cu seed deposition, Cu plating, Cu anneal, and CuCMP [Fig. 2.44(b)]
TiN liner deposition [Figs. 2.35(a) and 2.36(a)]	Oxide CVD
W deposition [Figs. 2.35(b) and 2.36(b)]	V3 mask
WCMP [Figs. 2.35(c) and 2.36(c)]	Etch oxide, PR strip, and clean [Fig. 2.45(a)]
Wafer clean	TiN dep, WCVD, and WCMP [Fig. 2.45(b)]
Oxide CVD	PVD TiN, PVD Al-Cu, and PVD TiN [Fig. 2.46(a)]
V1 mask [Fig. 2.37]	M3 mask
V1 etch, PR strip, and clean [Fig. 2.38]	Etch TiN/W/TiN metal stack, PR strip, and clean [Fig. 2.46(b)]
Oxide CVD	Oxide CVD and nitride CVD
M1 mask [Fig. 2.39]	Bond pad mask
M1 oxide trench etch, PR strip and clean	Etch nitride/oxide, PR strip, and clean (Fig. 2.47)

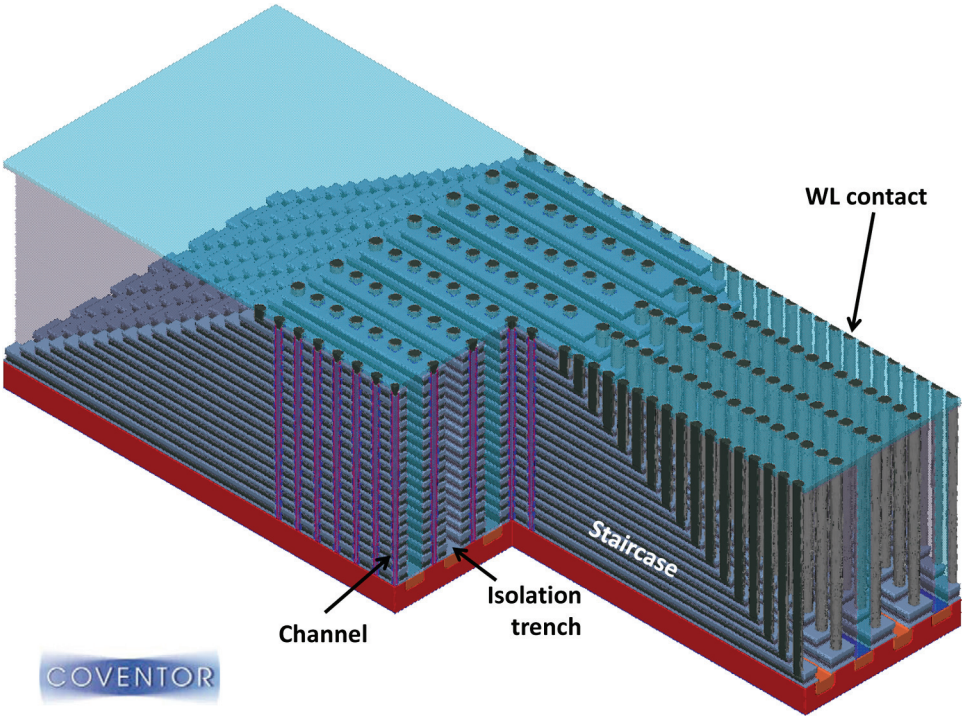


Figure 2.48 3D view of 3D-NAND after channel, isolation, and contact-plug formation. Image reprinted from Ref. D with permission from Coventor.

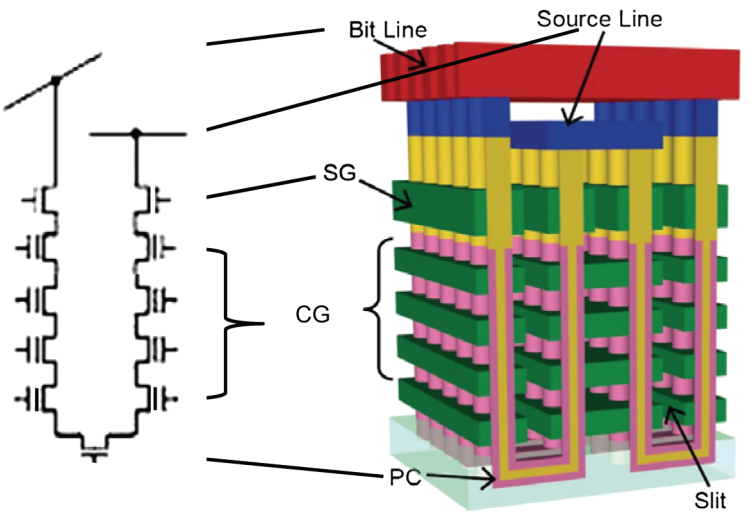


Figure 2.49 P-BiCS 3D-NAND: (a) circuit of a string and (b) 3D view. Image reprinted from Ref. E with permission from IEEE.

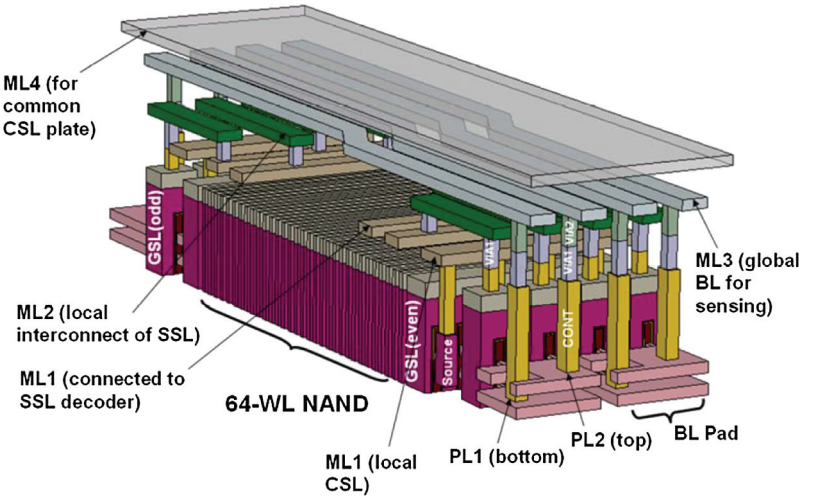


Figure 2.50 3D view of VG NAND. Image reprinted from Ref. F with permission from IEEE.

charging states and can thus store 8 bits of information. For the same cell size, a three-level cell can store four times the information of a single-level-cell NAND flash; therefore, a MLC NAND chip is more cost effective (higher bit/$) than a SLC NAND chip. Of course, there are always trade-offs. MLC NAND flash is usually slower, consumes more power, and has lower write-erase endurance than SLC NAND flash.

Figure 2.51(a) is the top-down view of a 3D-NAND cell with a polysilicon floating gate made from an OPOP stack. Figure 2.51(b) is the side-view

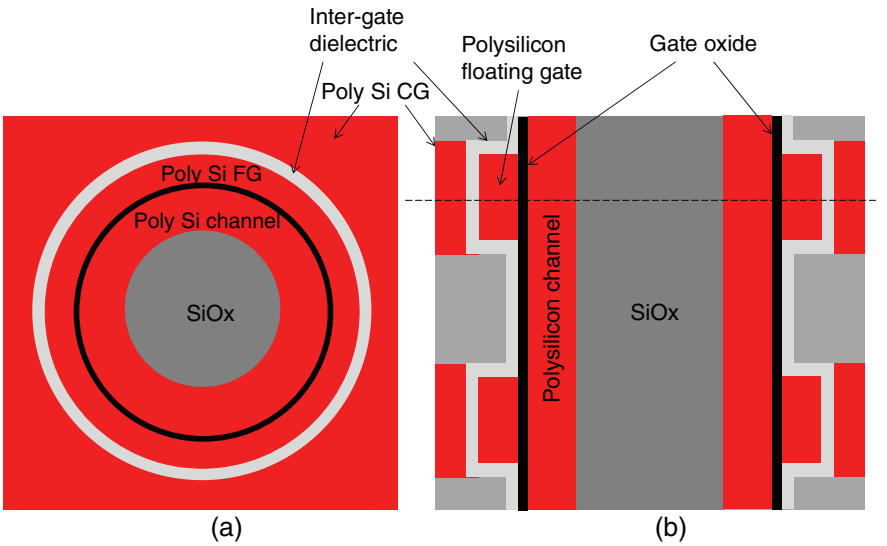


Figure 2.51 (a) Top-down view of a 3D-NAND cell with polysilicon FG and (b) its cross-section side view.

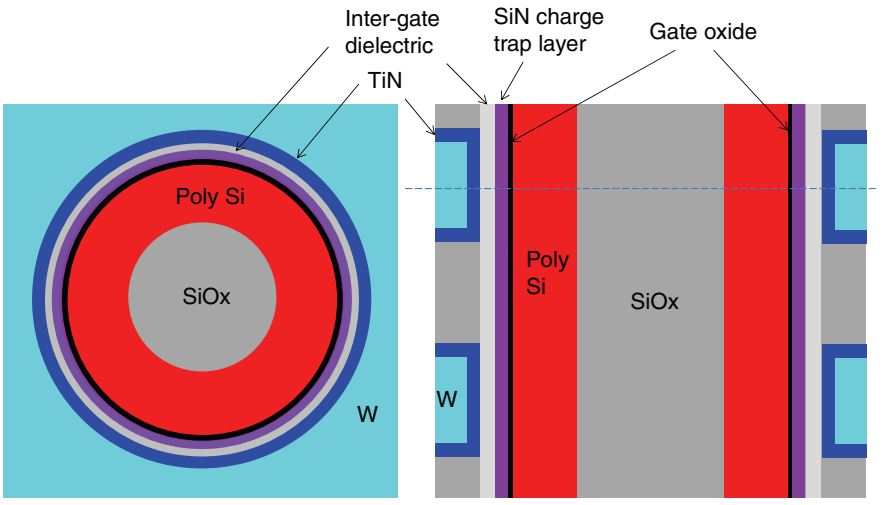


Figure 2.52 (a) Top-down view of a 3D-NAND cell with a SiN charge-trap layer and (b) its cross-section side view.

cross-section of the 3D-NAND cell with a polysilicon floating gate. In comparison, Figure 2.52(a) shows the top-down cross-section of a 3D-NAND cell with a nitride charge-trap layer that consists of an ONON stack, and Fig. 2.52(b) illustrates its cross-section from the side.

By creating the flash memory cell vertically, 3D-NAND can achieve a high bit density while relaxing the photolithography resolution requirements. It shifts the scaling difficulties from patterning to etch HAR structures and deposits a conformal thin film to cover those structures. By increasing the number of stacks, 3D-NAND flash can be scaled to the next generation without improving the photolithography resolution. 3D-NAND flash is becoming mainstream. This trend is unstoppable because alternative lithography technology that might keep scaling planar NAND flash at a cost lower than quadruple patterning of 193-nm immersion lithography, such as nano-imprint lithography or low-cost EVU lithography, cannot be implemented into HVM in time.

2.4 Review Questions

1. How many device(s) are in a flash memory cell?
2. What are the differences between NAND flash and NOR flash?
3. What is the minimum unit cell size of a 20-nm planar NAND flash?
4. What is the purpose of the SiN layer in Fig. 2.52?
5. Which one of the following processes does NOT etch a HAR pattern?
   (a) Staircase etch
   (b) Channel hole etch

(c) Isolation trench etch
(d) Staircase contact etch

6. How many pairs of ONON must be deposited to form a 64-cell 3D-NAND?
7. Can one photomask be used to form 64-cell staircase landing steps?
8. What is the equivalent $4F^2$ planar NAND flash technology node of a $6F^2$ 64-stack 74-nm half-pitch 3D-NAND device?
9. Describe the difference between ONON and OPOP stacks.
10. List at least two challenges facing 3D-NAND flash memory-chip manufacturing.
11. What is the biggest advantage of 3D-NAND flash compared to planar NAND flash?

Chapter 3
High-*k*, Metal-Gate FinFET CMOS Manufacturing Process

After reading this chapter, you should be able to

- List at least two advantages of FinFET technology compared to planar MOSFET technology;
- Explain the difference between SOI FinFET and bulk FinFETs process steps;
- Describe the differences between the gate formation of FinFETs and planar MOSFETs;
- Explain why a 14-nm-node fin needs pitch doubling and a 10-nm-node fin needs pitch quadrupling;
- Give the reason W is used as the FinFET HKMG metal filler, whereas planar HKMG MOSFET uses an Al filler; and
- Predict future FinFET scaling trends.

3.1 Introduction

In the so-called "good old days," the IC technology-node scaling of each generation always brought both higher device density and better device performance. When CMOS IC developed from the 90-nm to 65-nm node, the scaling did not improve the device performance: it only increased the device density. The main reason for this change is the thickness of the gate oxide can no longer be scaled down due to the leakage caused by the tunneling effect.

One of most important MOSFET device performance parameters, the drive current I_D, is proportional to $\mu(k/t_{ox})(W/L)$, or

$$I_D \propto \mu(k/t_{ox})(W/L), \tag{1}$$

where μ is the carrier mobility of the channel material (for NMOS, it is electron mobility, and for PMOS, it is hole mobility), k is the dielectric constant of the gate dielectric, and t_{ox} is the thickness of the gate oxide. For a

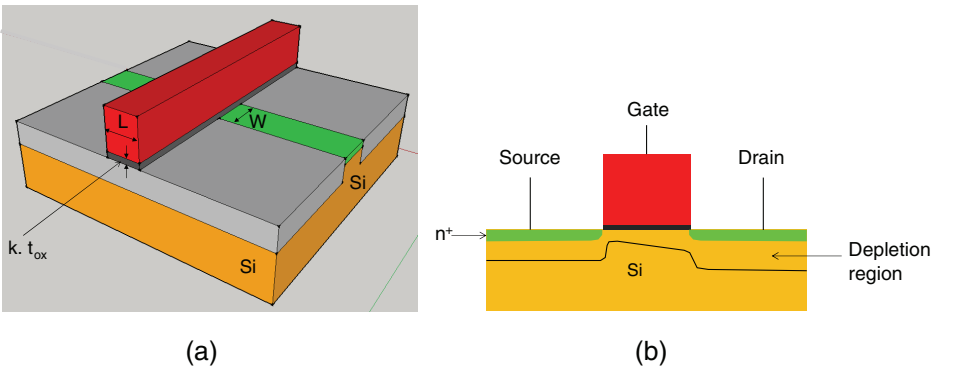


Figure 3.1 Planar MOSFET (a) in 3D and (b) 2D cross-section along the channel.

traditional gate dielectric (SiO_2), ${k_{SiO}}^2 = 3.9$. W is the channel width, and L is the channel length, as shown in Fig. 3.1.

If only the feature size of a planar MOSFET scales down, then W and L reduce at the same rate, and the drive current will not improve unless t_{ox} also scales down. The power supply voltage and threshold voltage are also scaled down occasionally with the gate oxide thickness to reduce leakage and power consumption. When t_{ox} became thin and approached the leakage and breakdown limit, scientists and engineers had to find other ways to improve I_D. One such technique uses a stress liner and selective epitaxial growth (SEG) silicon germanium (SiGe) to create channel strain, which improves carrier mobility μ and drive current. Another technique involves doping the gate silicon oxide with nitrogen to form SiON, which can slightly increase the k value (between 3.9 for SiO_2 and 7.5 for Si_3N_4, depending on the nitrogen concentration). By replacing SiON with a high-k dielectric, such as $HfSi_xO_yN_z$, the k value of the gate dielectric increased significantly. The gate dielectric could be formed with a thinner equivalent oxide thickness (EOT), which equals $(3.9/k)\ t_{ox}$, and further improves the drive current. When people talk about high-k or low-k dielectric, they are comparing the k value of SiO_2, which is 3.9.

Early MOSFETs used a metal gate (usually Al), but it was replaced by polysilicon after ion implantation and self-aligned S/D formation processes were introduced in the mid-1970s. Because polysilicon is a semiconductor, it always forms a depletion layer at the polysilicon and gate oxide interface when an external electric field is applied, which can affect channel formation, especially when the gate oxide becomes very thin. Reintroducing a metal gate solved the polysilicon depletion issue. Metal is a conductor, and it does not suffer from carrier depletion.

A high-k gate dielectric and metal gate were introduced in 45-nm technology, which helped to reduce the EOT and further improve device performance. It was a great achievement because even changing the gate

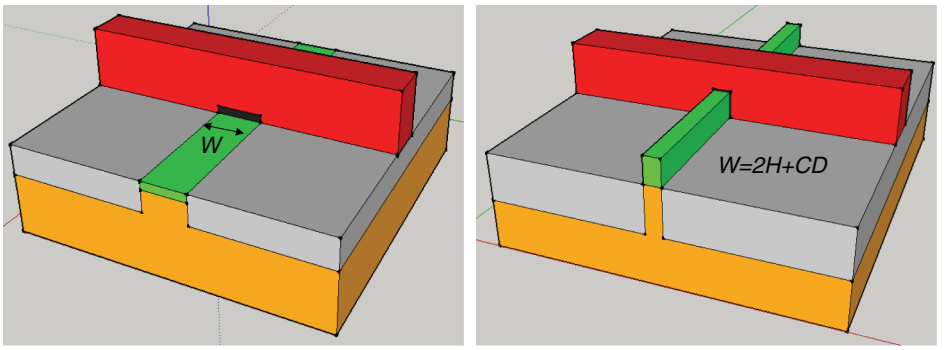


Figure 3.2 (a) Planar MOSFET and (b) FinFET.

dielectric alone presented a significant challenge due to the mismatch between the silicon and the high-*k* dielectric interface. The metal gate presented another set of challenges: NMOS and PMOS need different work-function metals to control their threshold voltages and the interface between those metals and the high-*k* dielectric. There are two ways to form a HKMG MOSFET, referred to as "gate first" and "gate last." The gate-first approach deposits the high-*k* dielectric, metal gate, and polysilicon layers first, and then gate patterning and etch form the HKMG devices. This method has only been used in the 32-nm / 28-nm technology node by a few IC manufacturers. Almost all IC manufacturers working with 22-nm / 20-nm nodes and beyond use the gate-last approach.

The next major technology development was the migration to 3D FinFETs. Figure 3.2(a) shows a planar MOSFET, and Fig. 3.2(b) shows a FinFET. The figures show that a FinFET could achieve the same channel width with a smaller silicon surface area. By increasing the fin heights, the channel width can be further increased, and therefore it can further improve device performance without scaling the device feature size. Of course, there is a limit to fin-height scaling. If the fin is too tall and the aspect ratio becomes too high, it could be very difficult to etch and clean the fins without causing them to collapse, and it could be very difficult to deposit void-free dielectric film to fill STI between the fins.

3.2 FinFET Basics

If two MOSFETs like the one shown in Fig. 3.1 were fused together, back to back (as shown in Fig. 3.3), the result can provide two conducting channels in the on-state MOSFET and double the drive current without scaling the feature size. However, the off-state leakage could also be doubled, which means a shorter standby time if this kind of IC chip is used in a mobile device.

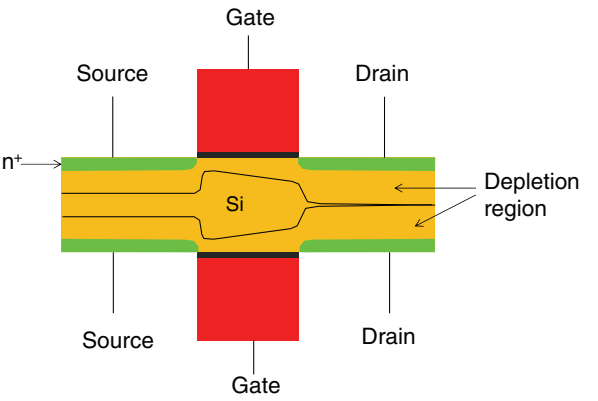


Figure 3.3 Double-gate planar MOSFET.

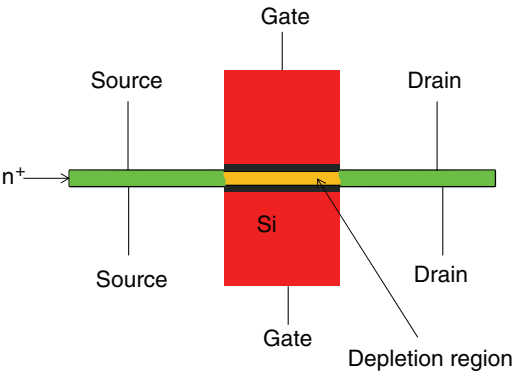


Figure 3.4 Double-gate FD planar MOSFET.

A fully depleted (FD) double-gate MOSFET can be obtained by making the silicon thinner than the depletion depth of the double-gate MOSFET, as shown in Fig. 2.4(b). The on-state device still has two conducting channels, like the partially depleted double device shown in Fig. 3.3. However, its off-state leakage can be significantly lowered because the channel is fully depleted, with very few carriers to conduct electrical current.

It is possible to make a double-sided FD MOSFET with a silicon-on-insulator (SOI) wafer; however, the processes are complicated, the cost could be high, and the bottom gate can have interference from the substrate. Figure 3.5(a) illustrates a SOI double-sided FD MOSFET with the STI oxide and the buried oxide removed to expose the active area, bottom gate, and silicon substrate.

Figure 3.5(b) illustrates a part of the double-gate FD planar MOSFET shown in Fig. 3.5(a), and Fig. 3.5(c) shows the device in Fig. 3.5(b) turned vertically and rotated 90 deg. The device in Fig. 3.5(c) is a tri-gate FinFET, a MOSFET with a fin-shaped active region and a channel surrounded on three sides by the gate.

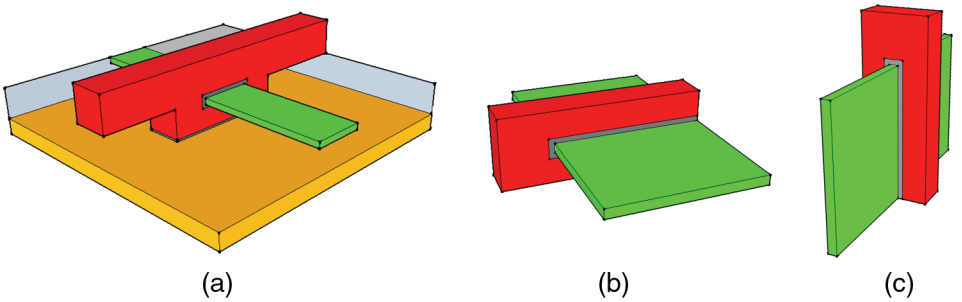


Figure 3.5 (a) Double-gate FD planar MOSFET made from SOI substrate, comprising (b)–(c) a double-gate FD planar MOSFET and a FinFET.

3.3 FinFET Process

Early versions of FinFETs were made with a SOI wafer [Fig. 3.6(a)]. The fin height is mainly determined by the thickness of the silicon on top of the buried oxide. At first, the silicon fin is etched with a process similar to the normal STI silicon etch process, using buried oxide as an etch-stop layer [Fig. 3.6(b)]; after wafer clean, the gate oxide is formed on the fin surface. Polysilicon is then deposited and planarized [Fig. 3.6(c)]. A polysilicon etch forms the gate electrode and local gate interconnect [Fig. 3.6(d)]. After that, ion implantation or SEG followed by rapid thermal anneal forms the S/D junction and finishes the FinFET device formation [Fig. 3.6(e)]. The SOI FinFET process steps are listed in Table 3.1.

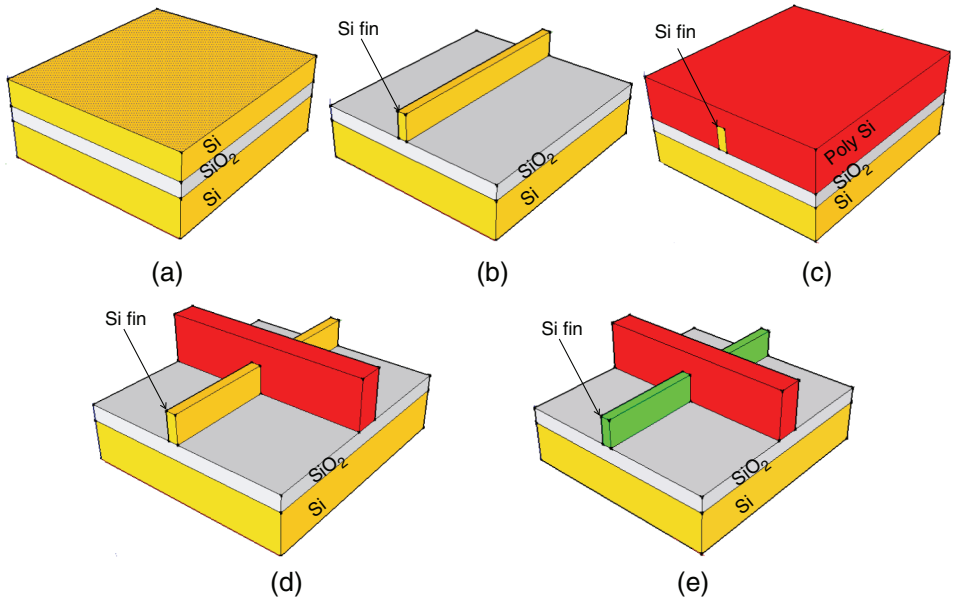


Figure 3.6 3D view of SOI FinFET formation: (a) SOI substrate, (b) fin etch, (c) gate oxidation, polysilicon deposition, and CMP, (d) polysilicon etch, and (e) fin dope.

Table 3.1 Process steps of an n-channel FinFET on a p-type SOI wafer.

Wafer clean [Fig. 3.6(a)]	Polysilicon deposition
Pad oxidation and nitride deposition	Polysilicon CMP [Fig. 3.6(c)]
Fin mask	Gate mask
Etch nitride/oxide	Polysilicon etch
Strip PR and wafer clean	Strip PR and wafer clean [Fig. 3.6(d)]
Etch silicon, stop on buried oxide	Ion implantation n-type S/D
Nitride/pad oxide strip and wafer clean [Fig. 3.6(b)]	Rapid thermal anneal [Fig. 3.6(e)]
Gate oxidation	

A manufacturing process for FinFET devices with a bulk silicon wafer was developed later. It started with silicon fin etch, which is similar to the STI process, with a much narrower active area (AA) [Fig. 3.7(a)]. After wafer clean and oxidation, an oxide layer is deposited and planarized [Fig. 3.7(b)]. The oxide is then recessed to expose the silicon fins, as shown in Fig. 3.7(c). After wafer clean and gate oxidation, polysilicon is deposited, followed by CMP [Fig. 3.7(d)]. Polysilicon etch forms the gate electrode, as shown in Fig. 3.7(e), and self-aligned S/D dope and RTA finish the bulk silicon FinFET formation, as shown in Fig. 3.7(f). The bulk wafer FinFET process steps are listed in Table 3.2.

The FinFET process is similar to the planar MOSFET process, and there are several major challenges. One such challenge is fin formation, which includes fin patterning, fin etching, and STI oxide recess. In a planar

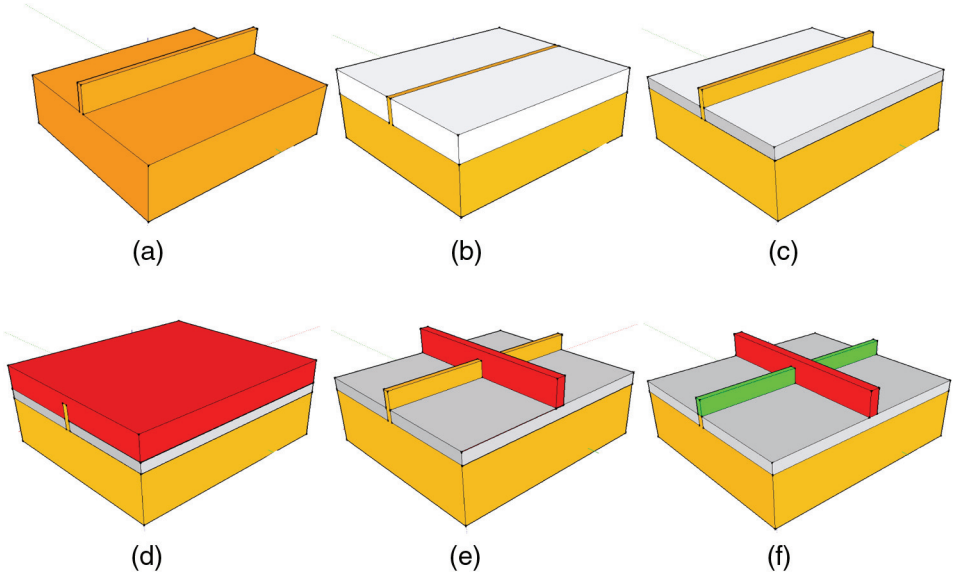


Figure 3.7 3D view of bulk silicon FinFET formation: (a) fin etch, (b) STI oxide deposition and CMP, (c) STI oxide recess, (d) gate oxidation, (d) polysilicon deposition and CMP, (e) polysilicon etch, and (f) fin dope.

Table 3.2 Process steps of an n-channel FinFET on a p-type bulk wafer.

Wafer clean	Strip nitride and pad oxide [Fig. 3.7(b)]
Pad oxidation and nitride deposition	Gate oxidation
Fin mask	Polysilicon deposition
Etch nitride/oxide	Polysilicon CMP [Fig. 3.7(c)]
Strip PR and wafer clean	Gate mask
Etch silicon	Polysilicon etch
Nitride and pad-oxide strip and wafer clean [Fig. 3.7(a)]	Strip PR and wafer clean [Fig. 3.7(d)]
Oxide deposition	Ion implantation n-type S/D
Oxide CMP and recess	RTA [Fig. 3.7(e)]

MOSFET device, the gate has the smallest pattern pitch; in a FinFET device, the fin has the smallest pattern pitch. The fin pitch for an Intel 14-nm FinFET is 42 nm, whereas the gate pitch is 70 nm. Self-aligned double patterning (SADP) is needed to pattern both layers. Self-aligned quadruple patterning (SAQP) with multiple cut masks are expected to form the fin patterns of 10-nm and 7-nm technology nodes. After hard-mask fin patterning, a silicon fin etch process is also very challenging because the gaps between the fins are deep and narrow, with a high aspect ratio. Oxide recess is also very challenging after STI oxide deposition and CMP because the STI oxide recess must stop at the right depth without an etch-stop layer. The depth of oxide recess determines the fin height of the FinFET device, which in turn affects the device channel width, $W = 2H + CD_{fin}$. Another big challenge is polysilicon etch, which determines the channel length. For planar MOSFET device, polysilicon layer is deposited on a relatively flat surface, and the etch process can use gate oxide as an endpoint. For FinFETs, polysilicon has a big step, determined by the fin height. When the etch reaches the top of the fin, the polysilicon below the fin still needs to be completely removed with an almost straight profile, shown in Fig. 3.7(e). If the gate oxides on top of the fin cracks during polysilicon etch, it will cause S/D silicon loss, or fin loss. If the etch is incomplete, it will leave a polysilicon stringer at the bottom corner of the fin, and causes an electrical short between neighboring gates, which kills the devices.

A comparison of Figs. 3.6(e) and 3.7(f) shows that a FinFET built on a SOI substrate and one built on a bulk silicon substrate are very similar. The main difference is that the fin of a SOI FinFET is completely isolated from silicon substrate, while the fin of a bulk silicon FinFET has a channel that connects to substrate, thus needing some doping to form isolation junctions to ensure good electrical isolation between the fin and substrate.

For CMOS logic devices with a 28-nm technology node and beyond, a high-k metal gate (HKMG) is needed to meet the required performance. A few more process steps are necessary to make the last HKMG FinFET. First, ILD1 oxide is deposited, and then CMP is performed to planarize the oxide surface [Fig. 3.8(a)]. Oxide CMP continues until it reaches the polysilicon gate, as shown in Fig. 3.8(b). It is very important to completely remove all of

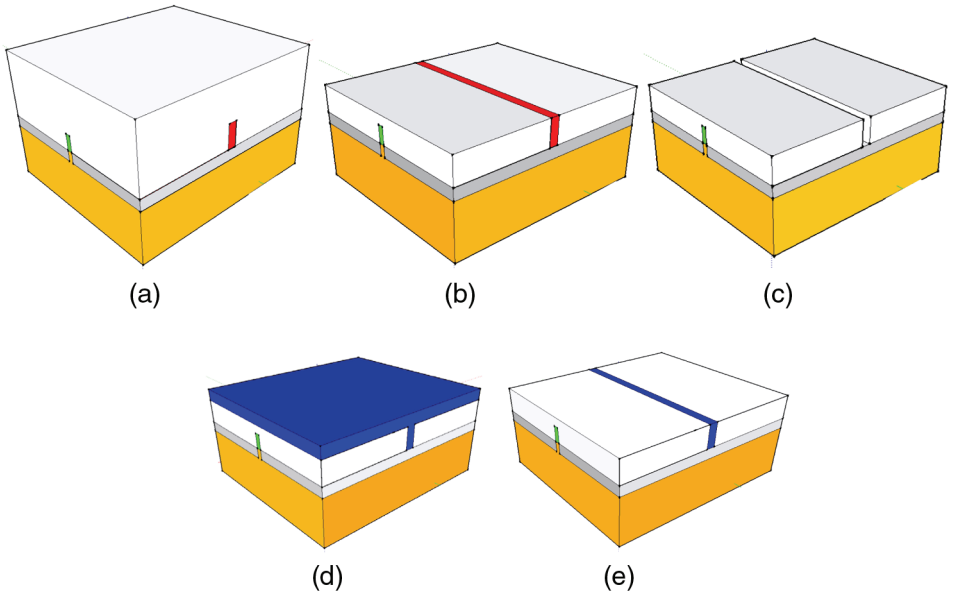


Figure 3.8 3D view of bulk silicon FinFET HKMG formation: (a) ILD deposition, (b) ILD oxide CMP, (c) polysilicon removal, (d) HKMG bulk metal deposition, and (e) metal CMP.

the oxide on top of the polysilicon gate, otherwise the remnant could affect the next process step: polysilicon removal. Dummy polysilicon removal, as shown in Fig. 3.8(c), is a critical step because all of the polysilicon must be removed from inside of the trench. Polysilicon residue here could cause device failure. Furthermore, when polysilicon is removed from the trench, the stress variation could crack the gate oxide liner near the trench wall, which could lead to fin loss underneath the ILD.

After dummy gate oxide removal and wafer clean, a thin layer of silicon dioxide is grown or deposited with an ALD process. Hafnium-oxide-based high-k dielectric is deposited with the ALD process, followed by work-function metal deposition. For a PMOS, ALD TiN is commonly used as the work-function metal; for NMOS, titanium aluminum nitride (TiAlN) is typical. After thin layers of HKMG are deposited, metal filler layers (usually an ALD TiN liner and CVD bulk W) are deposited to fill the narrow trench [Fig. 3.8(d)]. Metal gate CMP removes the metal layers and high-k dielectric from the wafer surface, leaving the metal gate and high-k gate dielectric inside the trench [Fig. 3.8(e)].

For planar HKMG MOSFETs, Al is used as the filler to fill the gate trenches after the deposition of high-k dielectric and work-function metals. For FinFETs, because the aspect ratio of the gate trench is significantly higher than that of planar MOSFETs, it requires better gap-fill capability. Therefore, W is commonly used as the filler for the narrow and deep gate trenches. The gate-last HKMG process steps are listed in Table 3.3.

Table 3.3 HKMG formation process with a gate-last approach.

ILD deposition	NMOS work-function metal deposition
ILD CMP [Fig. 3.8(a)]	TiN deposition
ILD CMP that stops on polysilicon [Fig. 3.8(b)]	W deposition
Dummy polysilicon gate removal	WCMP [Fig. 3.8(d)]
Wafer clean [Fig. 3.8(c)]	WCMP continues, followed by TiN CMP
High-*k* dielectric deposition	Wafer clean [Fig. 3.8(e)]

3.4 Advanced FinFET CMOS Process

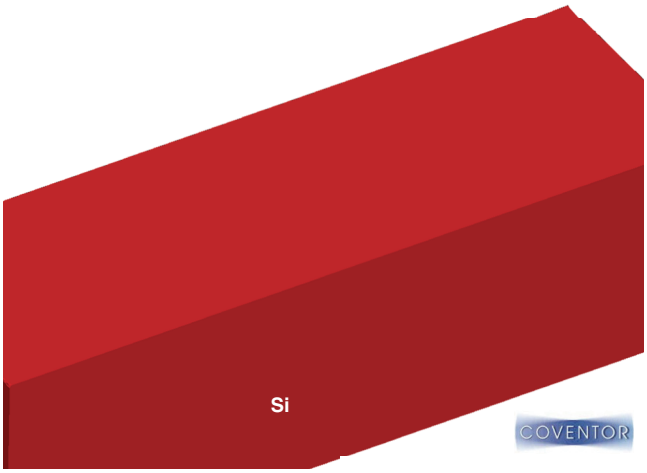
The previous section briefly explained how to build FinFET devices on a SOI wafer and on a bulk silicon wafer. It also described the gate-last HKMG process that replaces the polysilicon/SiON dummy gate with metal/high *k*. This section discusses in detail how to use advanced process technologies, such as SADP, ALD, etc., to form advanced HKMG FinFET CMOS devices on a bulk silicon substrate.

The wafer is first cleaned, as shown in Fig. 3.9(a), and then a thin layer of pad oxide is thermally grown on the silicon surface, and a silicon nitride layer is deposited on the pad oxide. This SiN layer serves as a hard mask (HM) for the silicon etch that forms the fins of the FinFET; it also serves as a CMP stop layer during STI oxide polish. The fin pitch is too small to pattern an advanced-technology node (such as 22 nm or 14 nm) with one patterning of 193-nm immersion lithography, and thus SADP is needed.

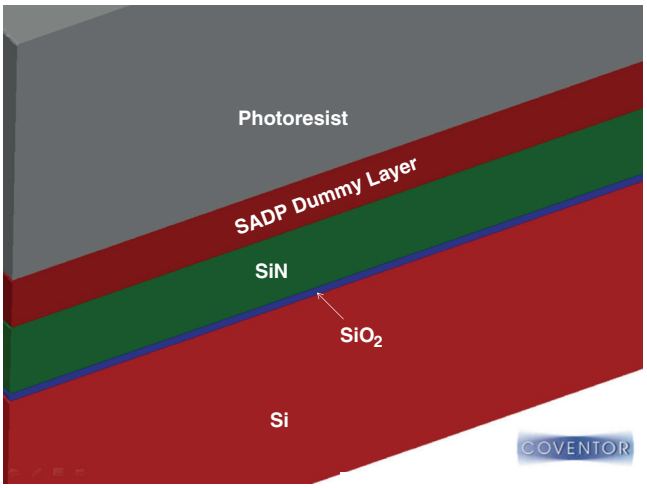
A dummy pattern layer of SADP is deposited on top of the nitride HM, and photoresist is coated on top of the entire stack [Fig. 3.9(b)]. The material of this layer needs to have high etch selectivity to the underlying silicon nitride and the spacer material that will be deposited on its sidewall to half the pitch and double the pattern density. One possible dummy/spacer-layer combination is a-Si/SiO_x (amorphous silicon and silicon oxide).

After PR patterning with a line-space fin dummy pattern mask, the dummy patterns, or mandrels, are etched with an endpoint at the SiN surface, as shown in Fig. 3.10(a). After PR strip and clean [Fig. 3.10(b)], a conformal dielectric layer is deposited [Fig. 3.10(c)], and then the dielectric layer is etched back in the vertical direction to form spacers on the sidewall of the mandrels, as shown in Fig. 3.10(d). After removal of the mandrels, the remaining spacer patterns form the line-space patterns that double the density of the original dummy pattern density [Fig. 3.10(e)]. If a-Si is used for the mandrels, potassium hydroxide (KOH) can be used to remove it with very little effect on the silicon oxide spacer and the underlying silicon nitride HM. The fin has the smallest pattern pitch of FinFET IC devices; advanced 14-nm IC chips have a fin pitch as small as 42 nm.

After wafer clean, photoresist is coated on the wafer surface, and cut mask is applied to pattern the PR [Fig. 3.11(a)]. An anisotropic etch process is applied to remove the spacer pattern in the area where no fins are designed



(a)

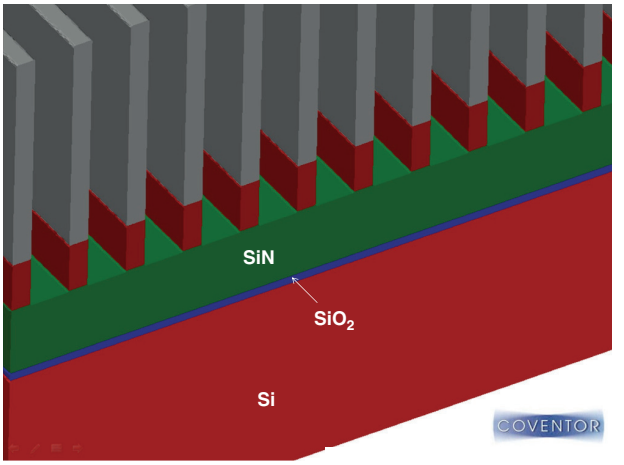


(b)

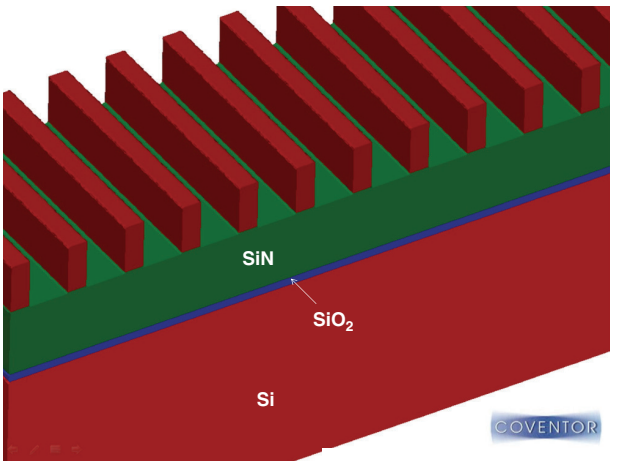
Figure 3.9 3D view of FinFET CMOS processing: (a) wafer clean and, after pad oxidation, (b) SiN deposition, SADP dummy layer deposition, and PR coating.

while keeping spacer patterns where the fins will be formed. This etch process needs high selectivity to the underlying nitride HM so that it etches away spacer patterns with minimal loss of the SiN HM. After PR strip and clean, the remaining spacer patterns can be used to etch SiN HM, as shown in Fig. 3.11(b). After etching away the pad oxide, the main etch process etches the silicon fin using SiN HM patterns [Fig. 3.11(c)].

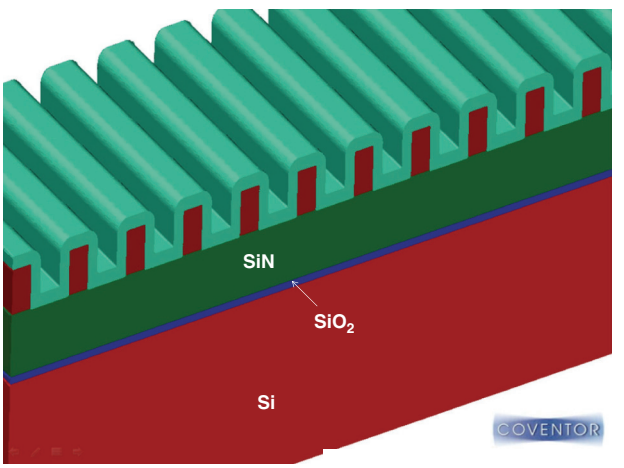
After wafer clean, an ILD layer is deposited [Fig. 3.12(a)], followed by ILD CMP with SiN as the endpoint [Fig. 3.12(b)]. After ILD recess, the SiN



(a)

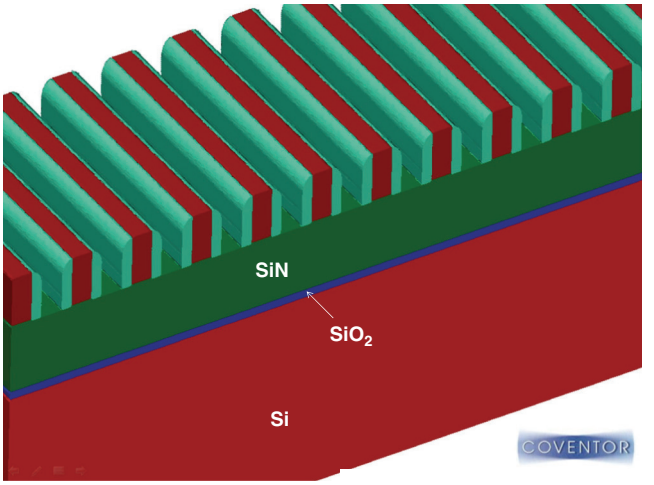


(b)

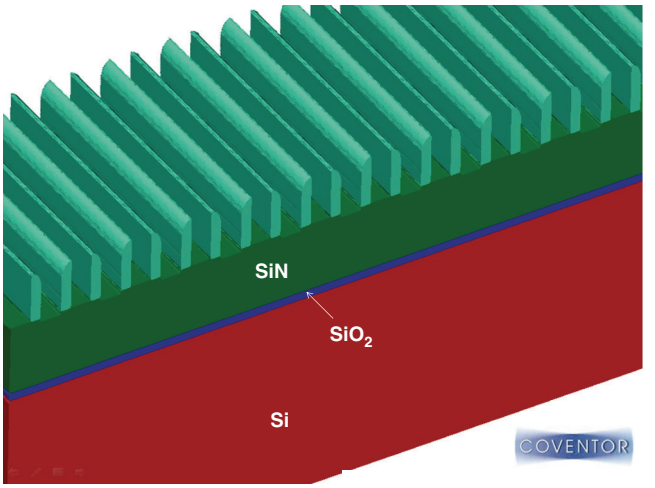


(c)

Figure 3.10 3D view of FinFET CMOS processing: (a) mandrel etch, (b) PR strip and clean, (c) spacer film deposition, (d) spacer etch, and (e) mandrel removal.



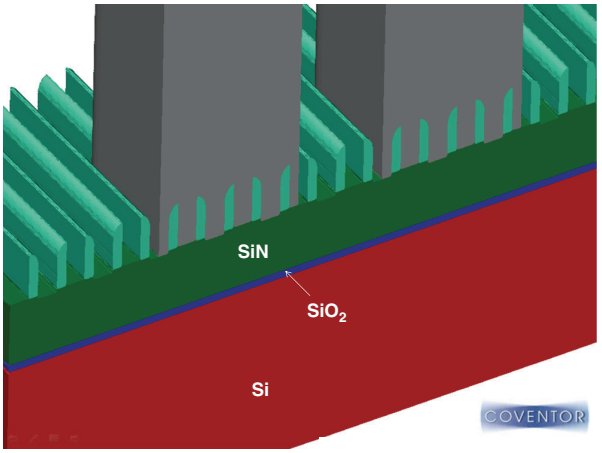
(d)



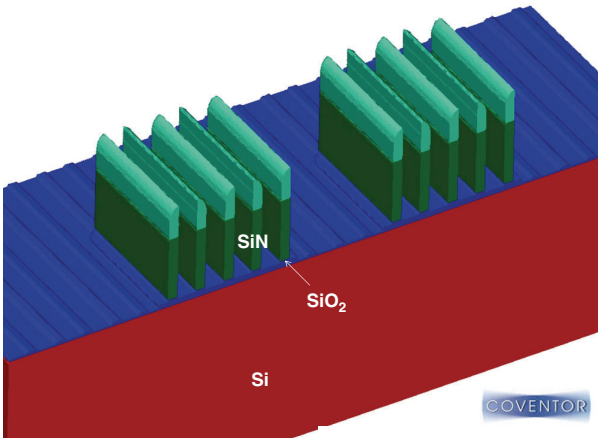
(e)

Figure 3.10 (Continued)

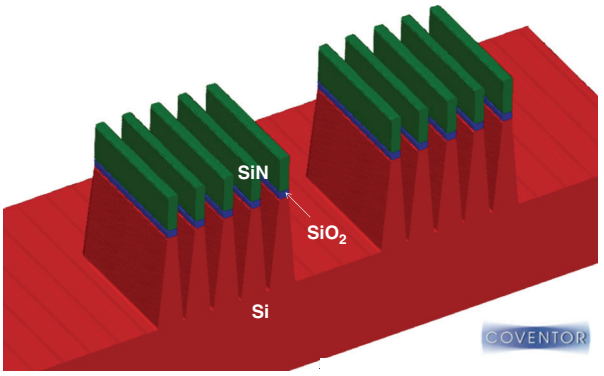
and pad oxide layer is stripped, as shown in Fig. 3.12(c). A sacrificial oxide layer is grown [Fig. 3.12(d)], a well-implantation mask is applied, ion implantation forms the isolation well between the channel and substrate, the sacrificial oxide is stripped, and wafer is cleaned, as shown in Fig. 3.12(e). This step finishes fin formation, which is somewhat similar to the STI formation of the planar CMOS process. As discussed earlier, it is critical to control the fin height because it directly affects the gate width of the FinFET devices.



(a)



(b)



(c)

Figure 3.11 3D view of FinFET CMOS processing: (a) fin cut mask patterning, (b) SiN HM etch, and (c) Si fin etch.

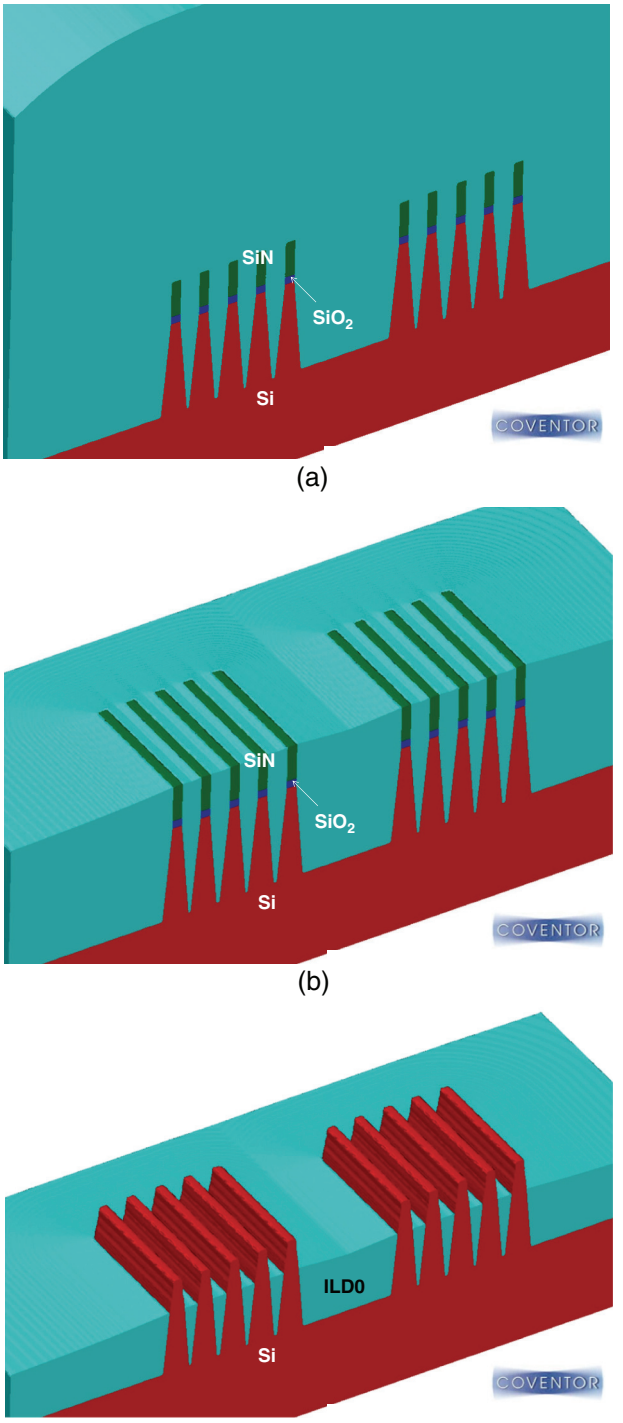
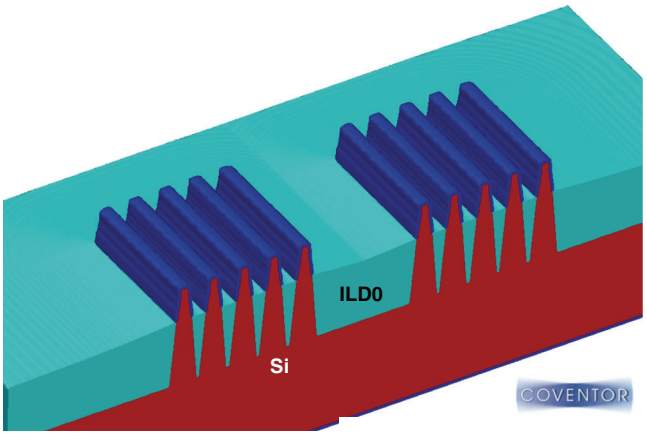
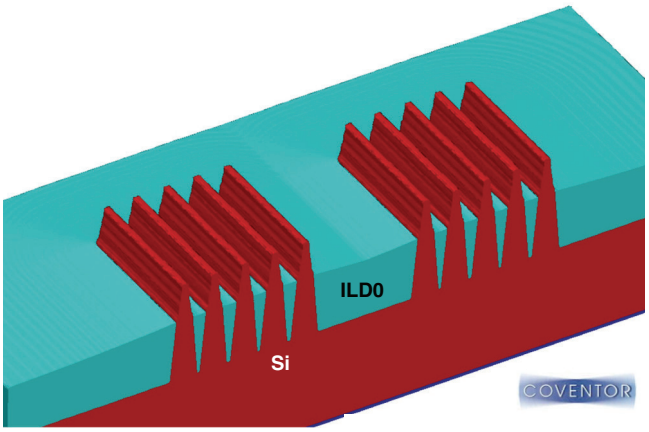


Figure 3.12 3D view of FinFET CMOS processing: (a) ILD0 deposition, (b) ILD0 CMP, (c) strip SiN and pad oxide, (d) sacrificial oxide growth, and (e) strip sacrificial oxide.



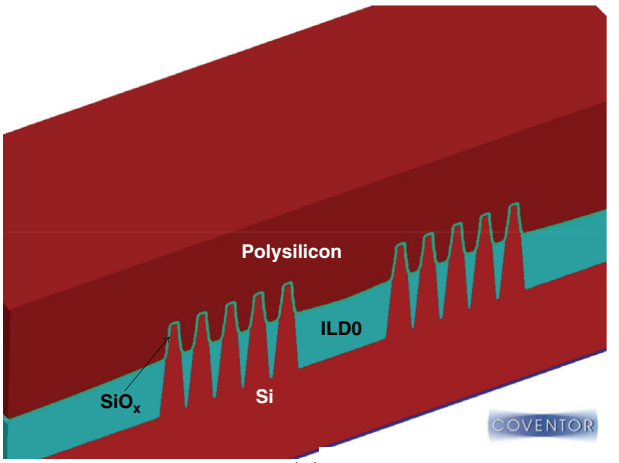
(d)



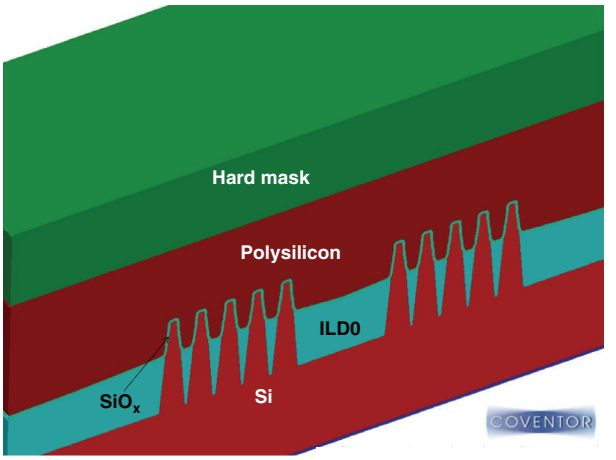
(e)

Figure 3.12 (Continued)

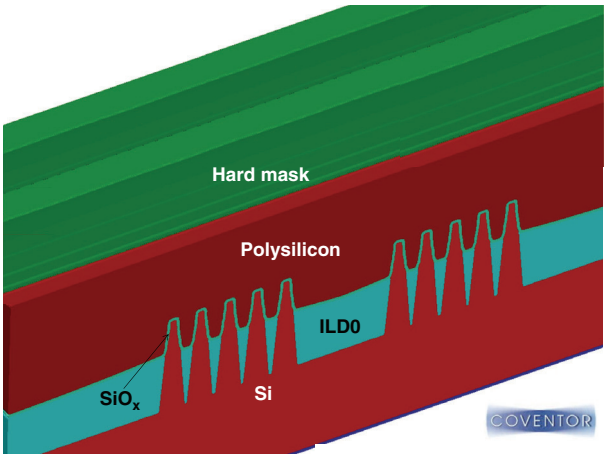
After fin formation, the wafer is cleaned, and a dummy gate oxide layer is deposited, followed by polysilicon deposition and CMP, as shown in Fig. 3.13(a). A HM layer is deposited, as shown in Fig. 3.13(b), and the gate mask is applied to form the line-space pattern on the photoresist. Depending on the technology node, if the gate pitch is larger than 80 nm, a single patterning with a 193-nm immersion lithography process can be used to form the line-space patterns. If the gate pitch is less than 80 nm, then pitch-multiplying techniques, such as SADP and SAQP, are needed. After HM etch and photoresist strip and clean, as shown in Fig. 3.13(c), a cut mask is applied, and an etch process cuts the HM line patterns. After that step, photoresist strip and clean occur, which form designed gate patterns on the HM layer [Fig. 3.13(d)]. This HM pattern is then used to etch polysilicon and form the dummy poly gate, as shown in Fig. 3.13(e).



(a)

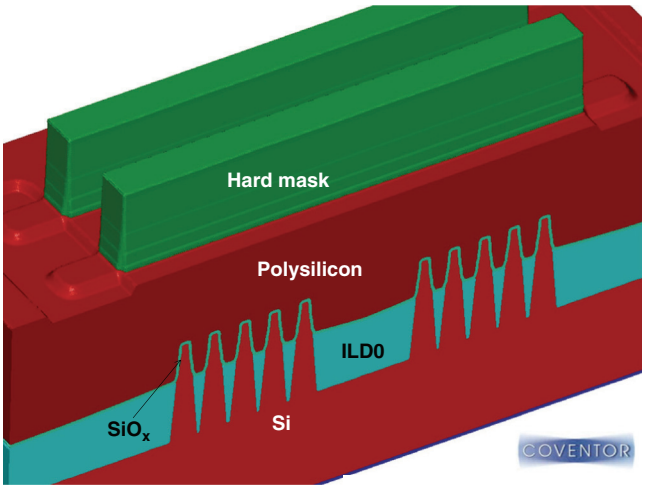


(b)

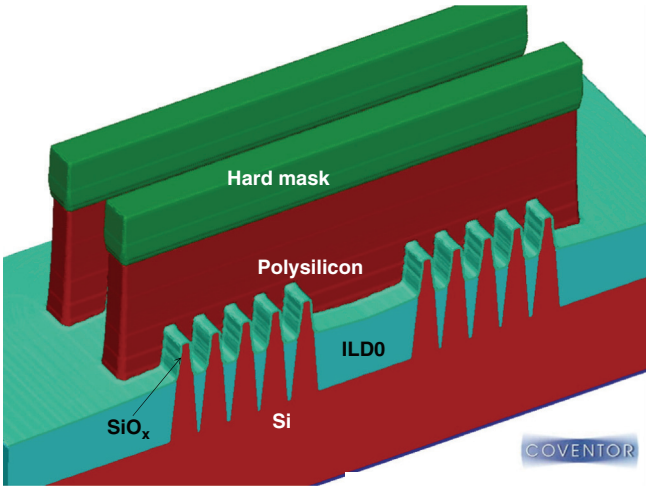


(c)

Figure 3.13 (a) Oxide deposition, poly deposition, and CMP; (b) HM layer deposition; (c) HM etch, and RP strip and clean; (d) cut mask etch, and PR strip and clean; and (e) poly etch.



(d)

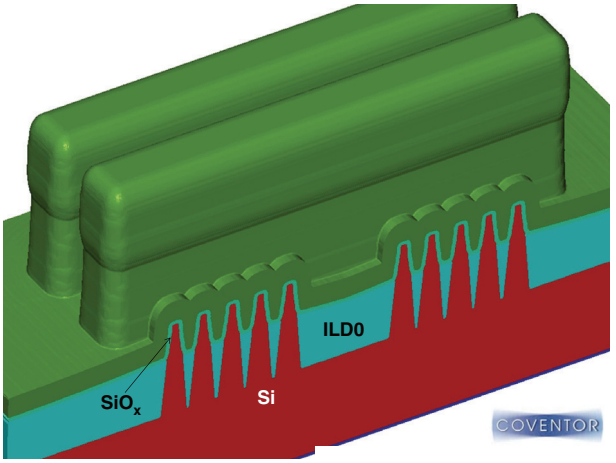


(e)

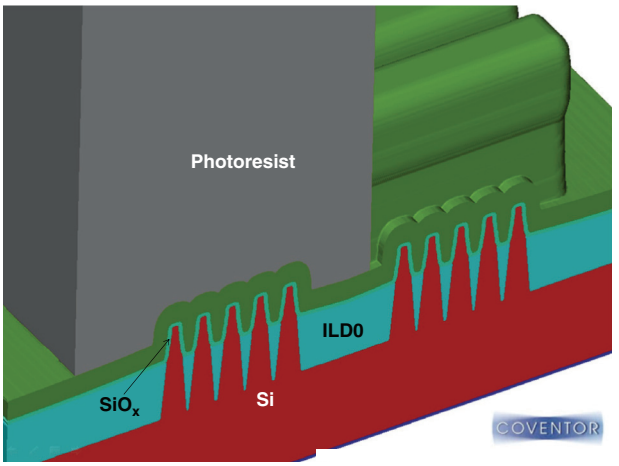
Figure 3.13 (Continued)

Now the fins and dummy polysilicon gate patterns are formed. The next processes are formation of S/D that involve spacer formation, ion implantation and selective epitaxial growth (SEG).

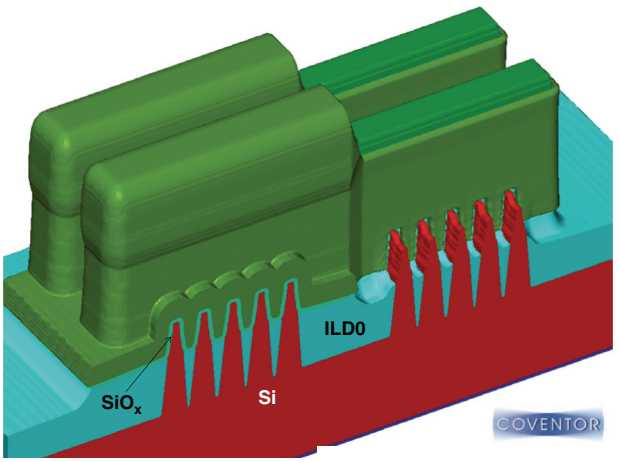
After wafer clean, a thin dielectric liner is deposited, followed by a thicker dielectric layer [Fig. 3.14(a)]. A PMOS mask is applied so that the NMOS areas are covered by PR to allow S/D formation of the PMOS [Fig. 3.14(b)]. After PMOS spacer etch and fin-spacer removal, the photoresist is stripped and the wafer is cleaned [Fig. 3.14(c)]. Then silicon is recessed and heavily p-type doped SiGe is grown in a SEG process, as shown in Fig. 3.14(d). This finishes the PMOS formation.



(a)

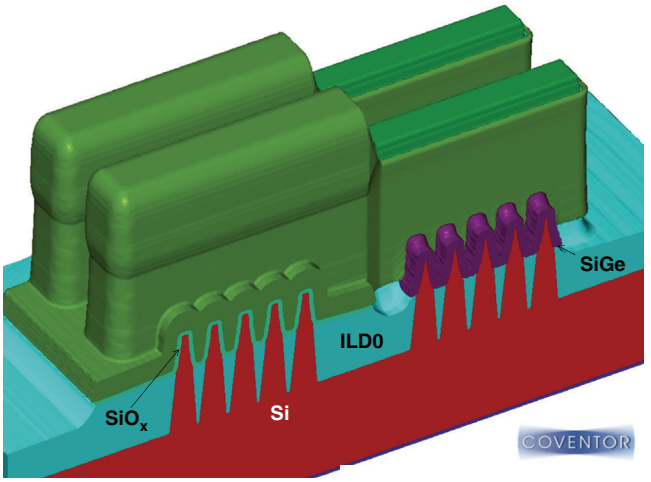


(b)



(c)

Figure 3.14 (a) Spacer dielectric deposition; (b) PMOS S/D mask; (c) PMOS spacer etch, fin spacer removal, and RP strip and clean; and (d) silicon recess and SEG SiGe.



(d)

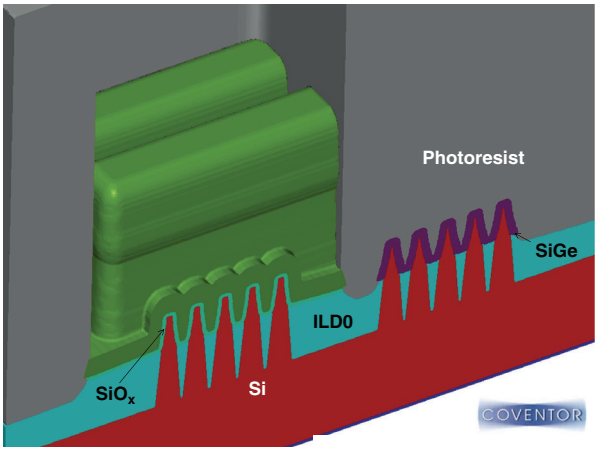
Figure 3.14 (Continued)

Next, NMOS S/D mask is applied, as shown in Fig. 3.15(a), NMOS spacer is etched and spacers on NMOS fins are removed [Fig. 3.15(b)]. N-type ion implantation is performed to heavily dope the NMOS S/D [Fig. 3.15(c)], and the photoresist is stripped and cleaned. Mini-second anneal (MSA) activated the dopant and finished the NMOS formation, as shown in Fig. 3.15(d).

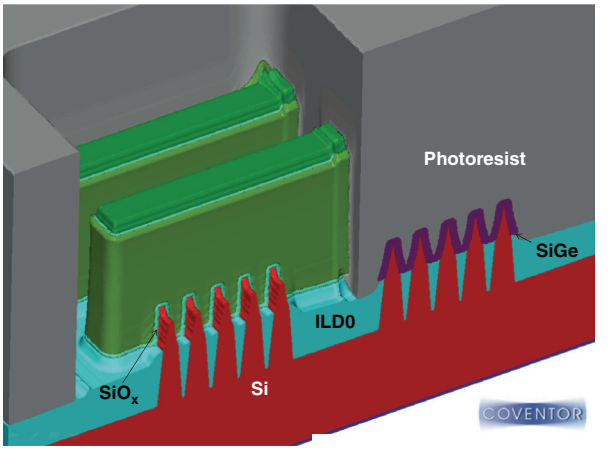
After the self-aligned silicide process forms silicide on the S/D, the wafer is ready for the replacement gate or gate-last HKMG process. First, the ILD1 is deposited, as shown in Fig. 3.16(a). Dielectric CMP removes part of the ILD1 to expose the dummy polysilicon gates [Fig. 3.16(b)]. After dummy-gate removal and oxide strip and clean, as shown in Fig. 3.16(c), hafnium-oxide-based high-*k* gate dielectric is deposited, followed by PMOS work-function metal TiN [Fig. 3.16(d)] and barrier metal TaN, all with ALD processes [Fig. 3.16(e)].

After wafer clean, photoresist is coated on the wafer, and a NMOS mask is applied to protect the PMOS and expose the NMOS, as shown in Fig. 3.17(a). An etch process removes the TaN barrier layer and TiN PMOS work-function metal, as shown in Fig. 3.17(b). After PR strip and clean, a NMOS work-function metal (such as TiAlN) is deposited [Fig. 3.17(c)]. A TiN adhesion layer is then deposited, followed by WCMP, which fills the gate trenches [Fig. 3.17(d)]. WCMP removes the bulk W. All other metal layers and the high-*k* dielectric layers are removed from the wafer surface in the over-polish step, as shown in Fig. 3.17(e).

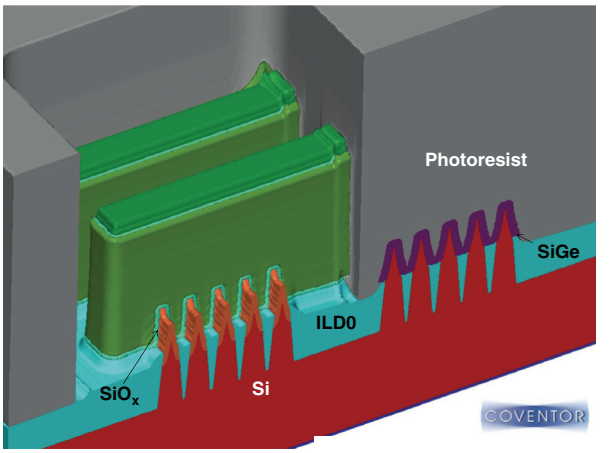
This step finishes the FEoL of HKMG FinFET CMOS devices; the following steps are the middle-end-of-line (MEoL) processes. The processes of



(a)

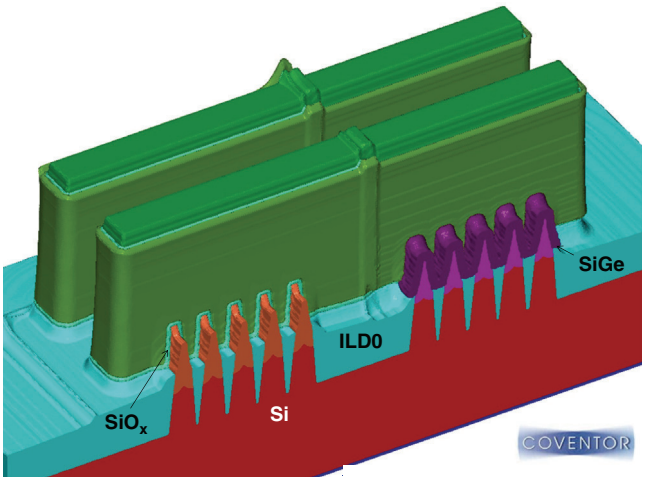


(b)



(c)

Figure 3.15 (a) NMOS S/D mask; (b) PMOS spacer etch and fin-spacer removal; (c) NMOS S/D implantation; and (d) PR strip, clean, and MSA.

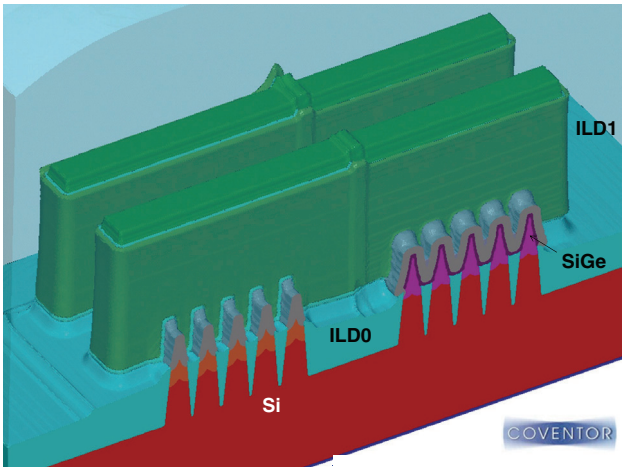


(d)

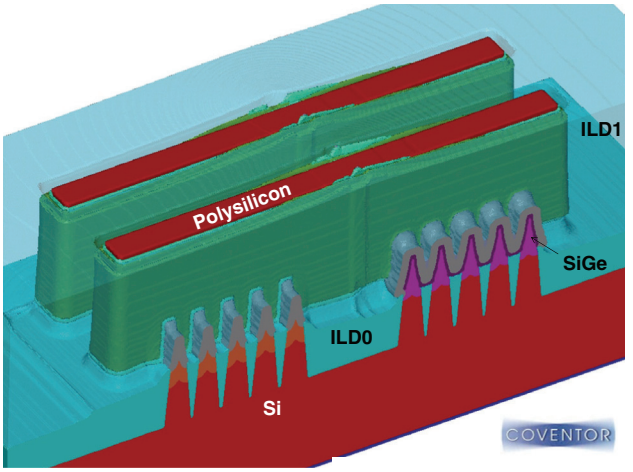
Figure 3.15 (Continued)

a 14-nm HKMG FinFET SRAM could have two layers with a total of four masks. One layer is the S/D contact (SDC) with two masks, and another layer is the gate contact (GC) with two masks. After two cycles of ILD/HM deposition, contact double patterning, contact etch, TiN and W deposition, and WCMP, the MEoL processes are finished; the contact plugs and local interconnection are formed, and the back-end-of-line (BEoL) processes could be started. Figure 3.18(a) is a transmission electron microscope (TEM) image of the cross-section of a 22-nm FinFET CMOS device with S/D contact and gate contact. Figure 3.18(b) is Fig. 3.17(e) with dashed lines to indicate the direction of the cross-section. The SDC contacts are not round holes but rather elongated trenches. The next section describes the processes of the contact module.

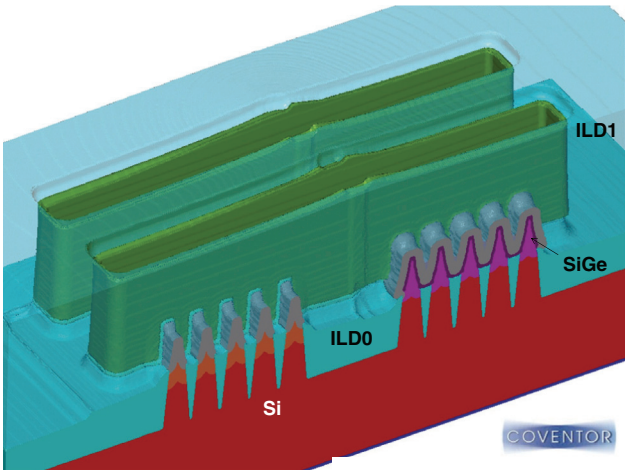
The BEoL processes start with wafer clean and etch-stop layer (ESL) deposition. They are followed by depositions of ultra-low-*k* dielectric, a cap oxide layer, TiN metal HM layer, dielectric HM layer, and the dummy layer. After inspection, review, and clean, PR is coated on the wafer, the first metal mask is applied, and mandrels are etched from the dummy layer. A conformal dielectric layer is deposited, and an etch-back process forms spacers on the sidewalls of the mandrels. After mandrel removal, the pitch of the line space is halved. The cross-section along the spacer is shown in Fig. 3.19(a). The process steps are very similar to that illustrated in Fig. 3.10. A cut mask is applied, and the looped lines formed by the spacer are cut into the designed pattern, as shown in Fig. 3.19(b). A dielectric layer is deposited to fully cover the patterns, illustrated in Fig. 3.19(c). A CMP



(a)

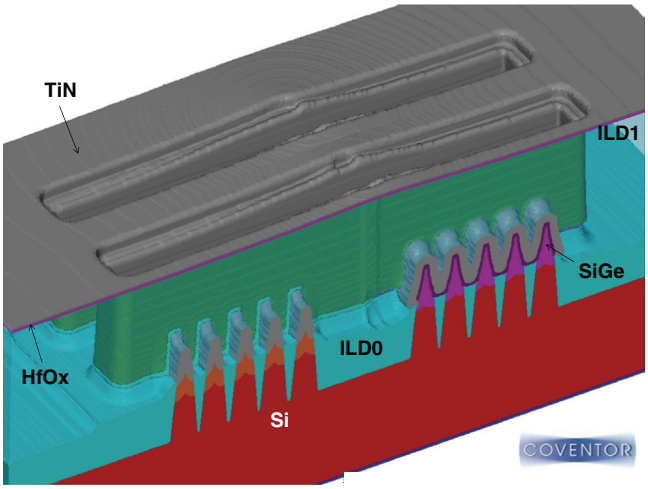


(b)

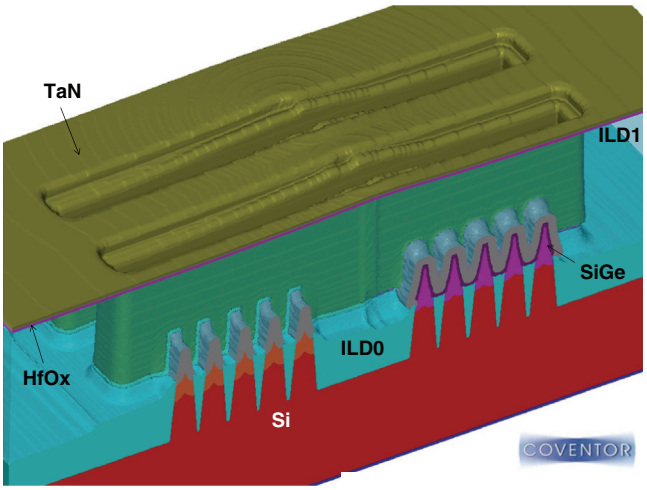


(c)

Figure 3.16 (a) ILD1 CVD, (b) ILD1 CMP (b), dummy polysilicon gate removal (c), High-k gate dielectric ALD (d), TiN and TaN ALD (e).



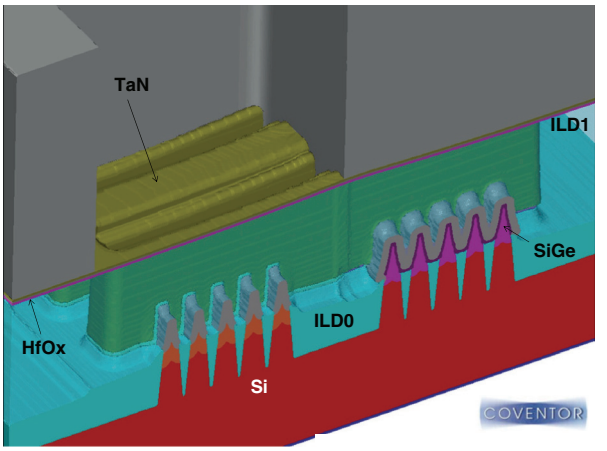
(d)



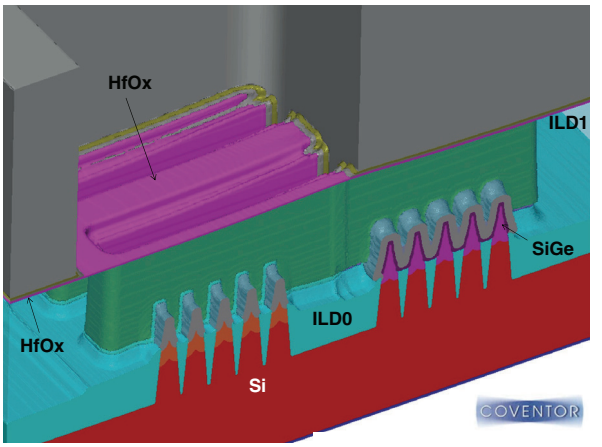
(e)

Figure 3.16 (Continued)

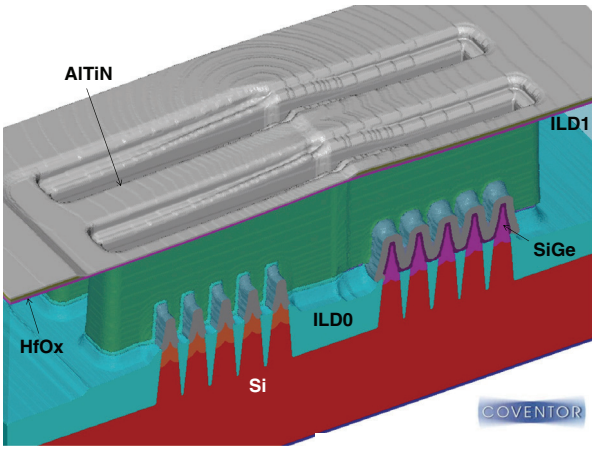
process planarizes the dielectric and exposes the patterns that are embedded in the dielectric film [see Fig. 3.19(d)]. A highly selective etch process removes the embedded patterns to form the dielectric HM etch, which can be used to etch the TiN hard mask, as shown in Fig. 3.19(e). 3D illustrations of this pattern reversal process are shown in Fig. 3.18(f). After dielectric HM strip and clean, the M1 trench patterns are transferred to TiN hard mask. Next, the V1 mask is applied, and the V1 is etched when it is aligned with M1, as shown in Fig. 3.19(i). After the via hole reaches the ESL, etch is stopped and PR is stripped. Trenches are etched using a TiN



(a)

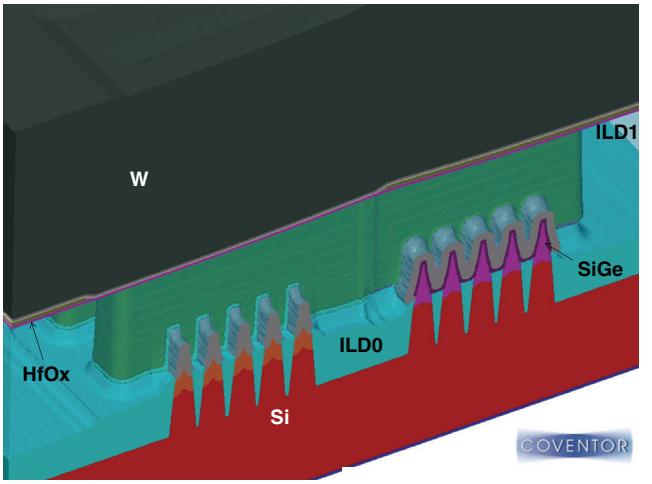


(b)

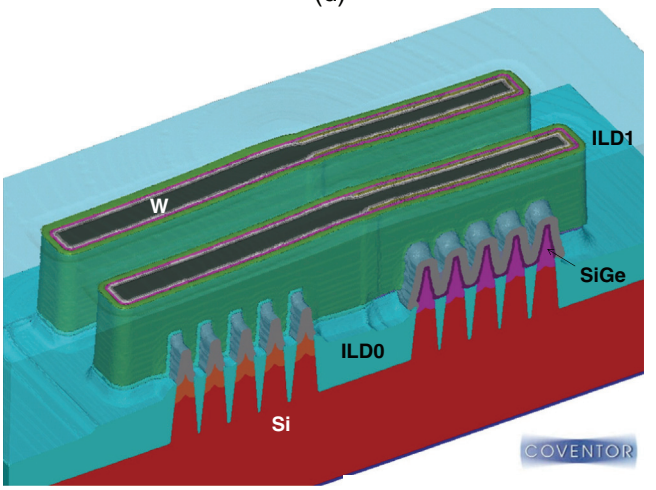


(c)

Figure 3.17 (a) NMOS mask, (b) TaN and TiN etch, (c) TiAlN ALD, (d) TiN and W deposition, and (e) WCMP.



(d)



(e)

Figure 3.17 (Continued)

HM, and the ESL is broken through at the bottom of via holes [Fig. 3.19(j)]. After barrier TaN and seed Cu deposition, bulk Cu is plated, shown in Fig. 3.19(k). The wafer is then cleaned and annealed. Metal CMP removes the Cu, TaN, and TiN metal HM from the wafer surface, and self-aligned cobalt tungsten phosphide (CoWP) is deposited on the metal surface with an electrode-less plating process, as shown in Fig. 3.19(l). This step finishes the dual-damascene M1 process module.

If V1 is not well aligned with M1, then the TiN hard mask could block the V1 etch, creating a small via hole and void in the via plug after metal deposition, as in Fig. 3.20. The small via could also be caused by the

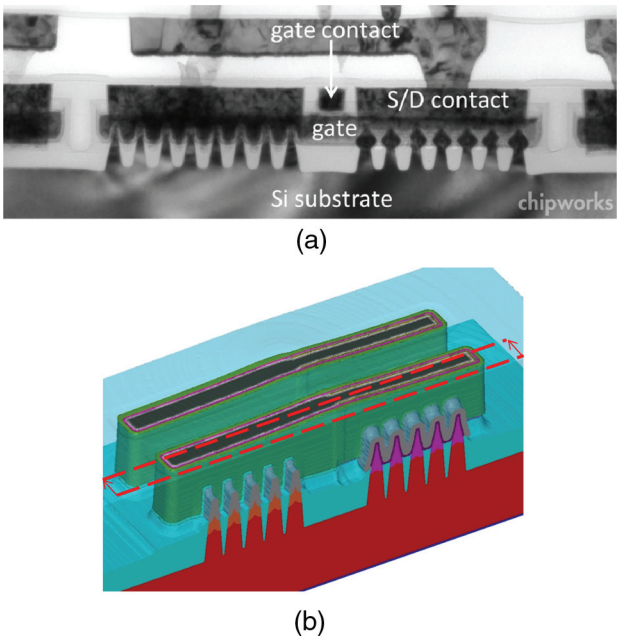


Figure 3.18 (a) Cross-section TEM image of FinFET CMOS with GC plug and SDC plugs; reprinted from Ref. 16 with permission from Chipworks. (b) Fig. 3.17(e) with an indication of the TEM sample.

pull-back of the trench pattern. Sometimes a small via cannot be etched all the way to the previous metal layer, which causes an open circuit of the interconnect.

The M2, M3, and Mx layers essentially repeat the M1 process steps (x is the number of total metal layers). There are 13 metal layers for a 14-nm FinFET CMOS chip—the most-advanced IC technology node in high-volume manufacturing at the time of publication. The main variations are the upper metal layers, which have a larger feature size and thus do not require double patterning. The feature sizes of the last few metal layers become so large that they do not require 193-nm immersion lithography; 248-nm (KrF excimer laser) or even 365-nm (i-line of a mercury lamp) lithography could be used. Figure 3.21 shows a cross-section of a 13-metal-layer HKMG FinFET chip.[17]

3.5 Advanced FinFET SRAM

Static random access memory (SRAM) is widely used in the CMOS logic IC as cache memory due to its high speed. SRAM is usually formed by two latched-up inverters and two pass NMOS, where one connects to a bit line and another connects to an inverted bit line. Figure 3.22(a) shows the SRAM

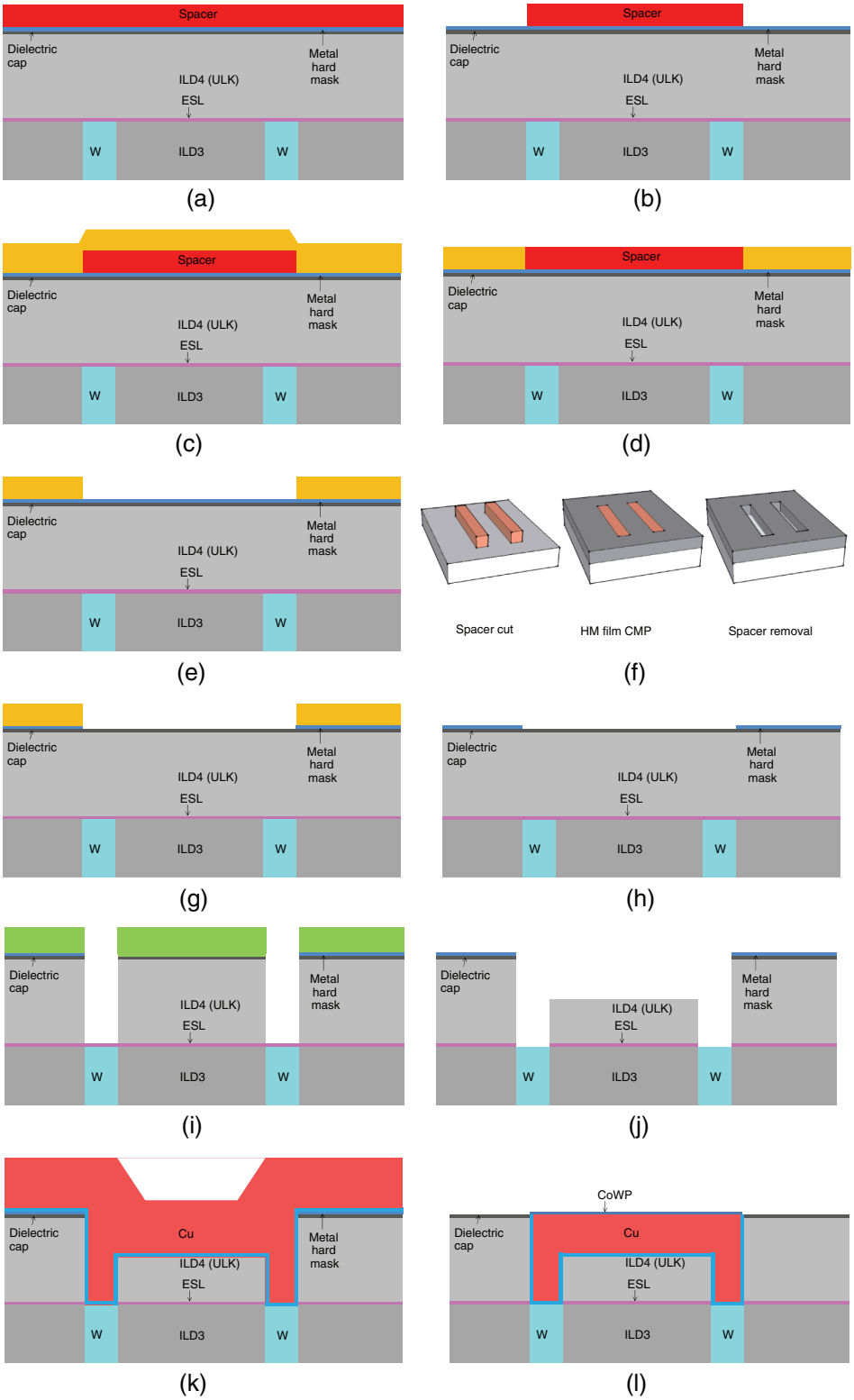


Figure 3.19 Copper metallization process: (a) SADP spacer formation; (b) spacer cut; (c) dielectric hard mask deposition; (d) dielectric hard mask CMP; (e) spacer removal; (f) 3D illustration of pattern reverse; (g) M1 pattern etch on the TiN hard mask; (h) dielectric hard mask removal; (i) via etch, (j) trench etch and ESL breakthrough; (k) copper plating, and (l) metal CMP and CoWP electrode-less plating.

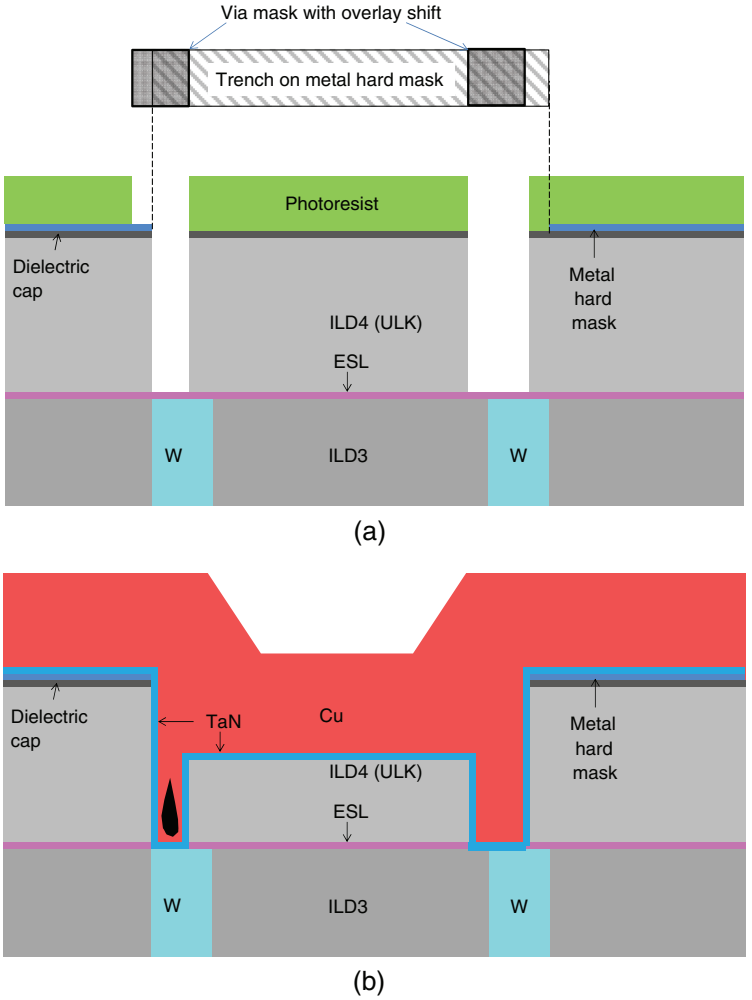


Figure 3.20 Cu void caused by via mask-overlay-induced CD variation: (a) after via etch and (b) after Cu plating with void.

circuit with two inverters and two pass gates. Figure 3.22(b) shows the detail of an inverter as one NMOS and one PMOS. Here, "bit" indicates a bit signal, and "bit′" indicates the inverted bit signal. A SRAM cell needs six transistors, four NMOS, and two PMOS—far more than a DRAM cell of one NMOS and one capacitor, or a NAND flash cell of one device.

SRAM usually has the highest pattern density in a logic CMOS chip, and it is commonly used as the test vehicle for process development of new technology nodes. Figure 3.23(a) is the top-down TEM image of Intel 14-nm SRAM.[24] The dashed rectangular box indicates the unit cell of SRAM. Each inverter consists of a NMOS and a PMOS with a shared gate. Figure 3.23(b) is

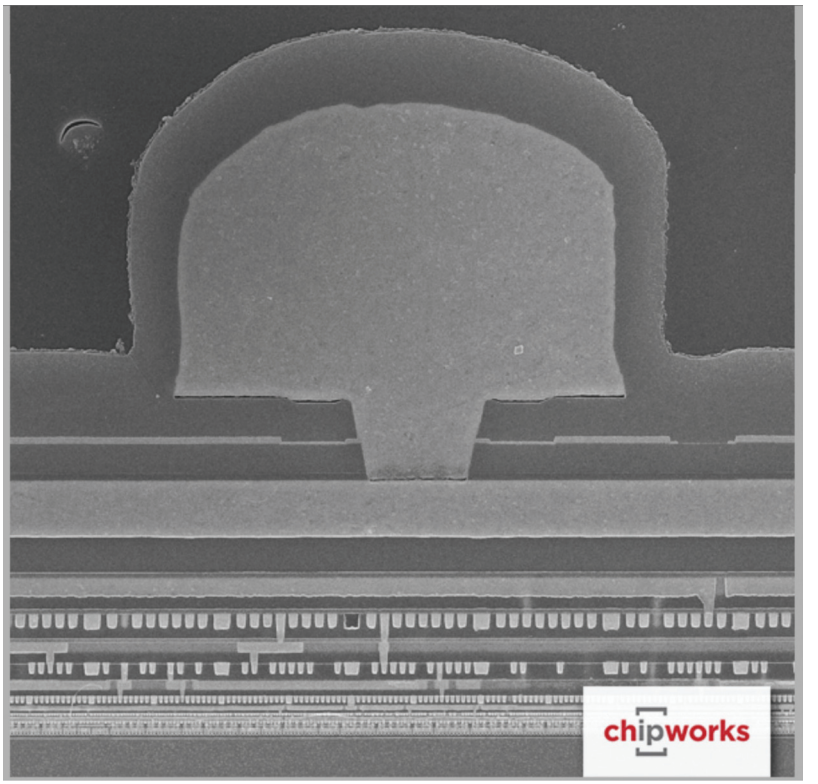


Figure 3.21 Cross-section of an Intel 14-nm chip with 13 metal layers. Image reprinted from Ref. 17 with permission from Chipworks.

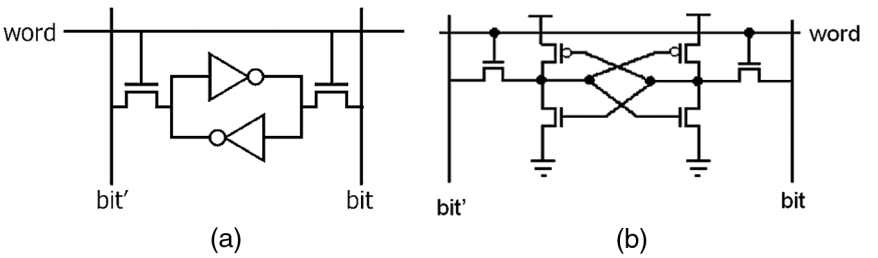


Figure 3.22 SRAM cell circuit with (a) inverters and (b) transistors.

the the SRAM design-layout drawing based on the TEM image in Fig. 3.23(a). It has four device layers: fin, gate, S/D contact and gate contact. The fins and gate contacts are in the vertical direction; the gates and S/D contacts are in the horizontal direction. A SRAM unit cell consists of two inverters and two pass gates (NMOS).

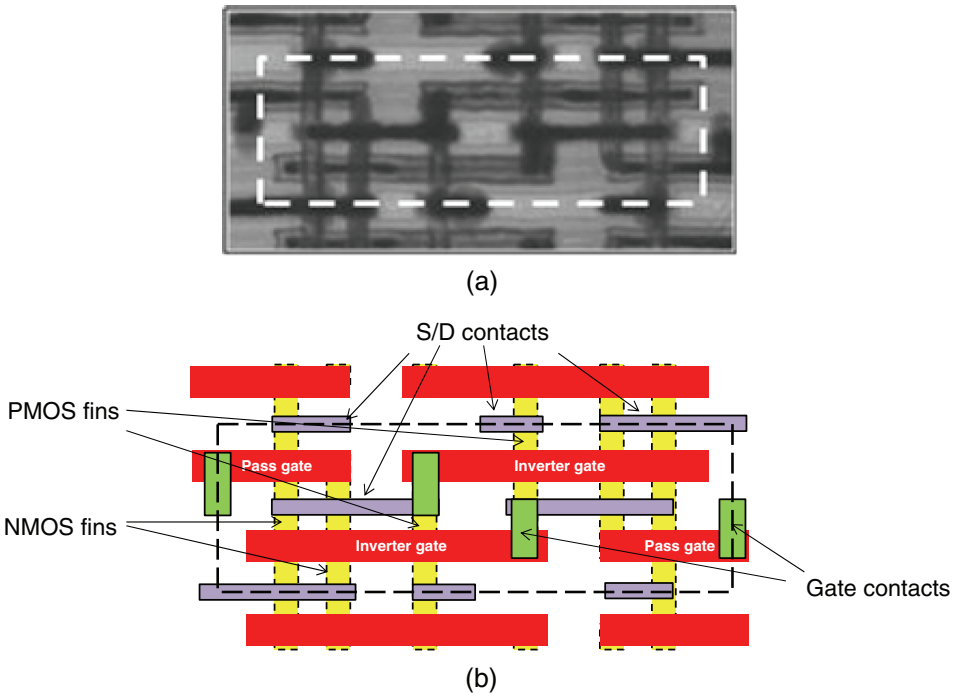


Figure 3.23 (a) Top-down TEM image of a SRAM cell. Source: Ref. 24, reprinted with permission from IEEE. (b) SRAM layout based on (a) with indications of the fins, gates, and contacts.

A SRAM array image is formed by copying, pasting, and mirroring the unit cell shown in Fig. 3.23 [see Fig. 3.24(a)]. The dashed rectangular box indicates a unit cell. An advanced CMOS logic IC chip could have multiple SRAM arrays with up to 256 million cells. In comparison, there are only 24 cells in the SRAM array shown in Fig. 3.24(a). Figure 3.24(b) is the design intent based on Fig. 3.24(a). There are four patterning layers: fin, gate, SDC, and GC. At least four doping layers are needed to form a FinFET SRAM: n-well, p-well, n-S/D, and p-S/D. For gate-last HKMG FinFETs, there is also a NMOS MG mask, which allows the formation of the NMOS work-function metal gate by either removing the PMOS work-function metal or the PMOS MG barrier layer to allow a chemical reaction that changes the PMOS work-function metal to a NMOS work-function metal. The design of each layer is one-dimensional (1D). The fin, doping, and GC patterns are in the vertical direction; the gate and SDC patterns are in the horizontal direction.

The rest of this section describes the FEoL and MEoL process steps of 14-nm FinFET SRAM from a top-down perspective, including photolithography masks, and from a cross-sectional view, mainly in the direction along the gate.

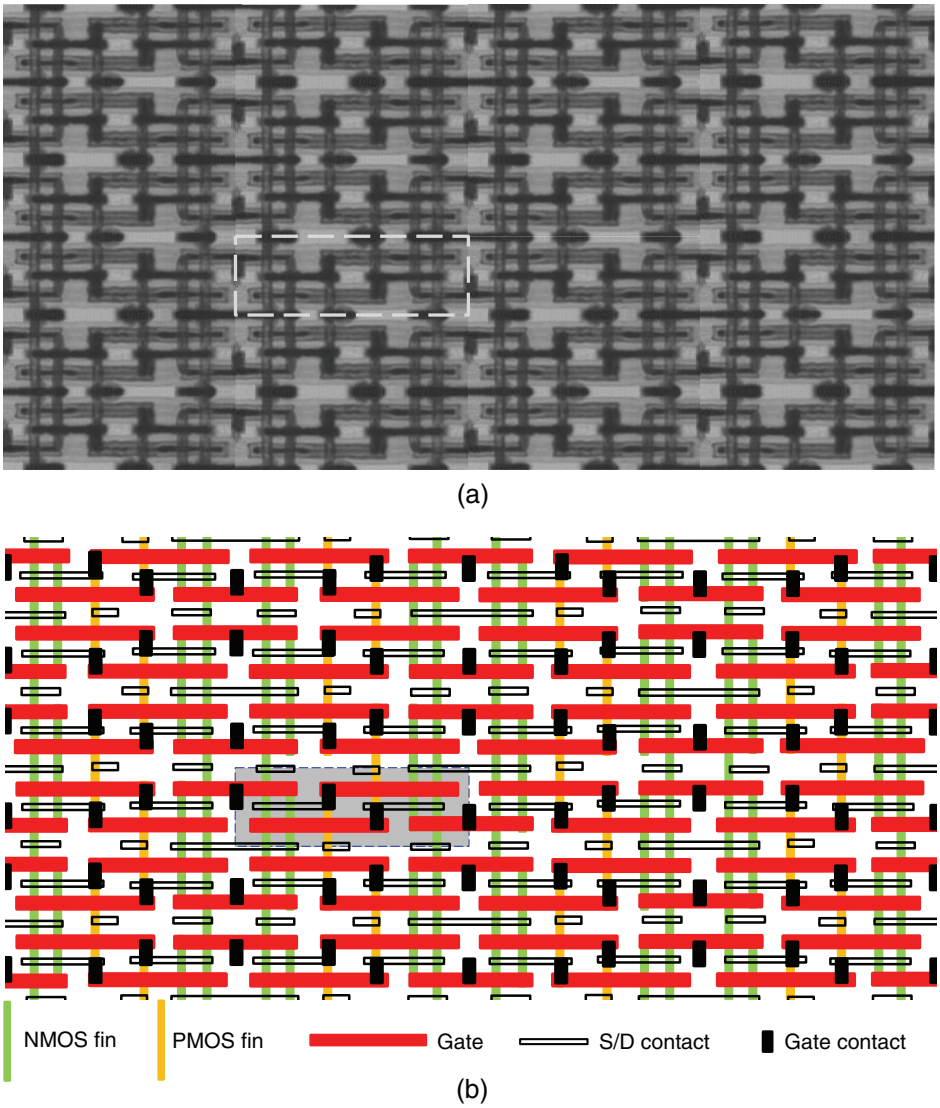


Figure 3.24 Top-down view image of a 24-cell SRAM array: (a) the dashed box is a unit cell, and (b) the design intent of the SRAM array.

After wafer clean, sacrificial oxide is grown in the diffusion furnace. An n-well mask [shown in Fig. 3.25(a)] is applied, and high-energy ion implantation forms a deep n-well junction. After photoresist strip and clean, the p-well mask shown in Fig. 3.25(b) is applied and high-energy p-type ion implantation is performed to form a deep p-well junction. After PR strip and wafer clean, an annealing process repairs implantation-induced crystalline damage and activates the dopant, which finishes the well formation.

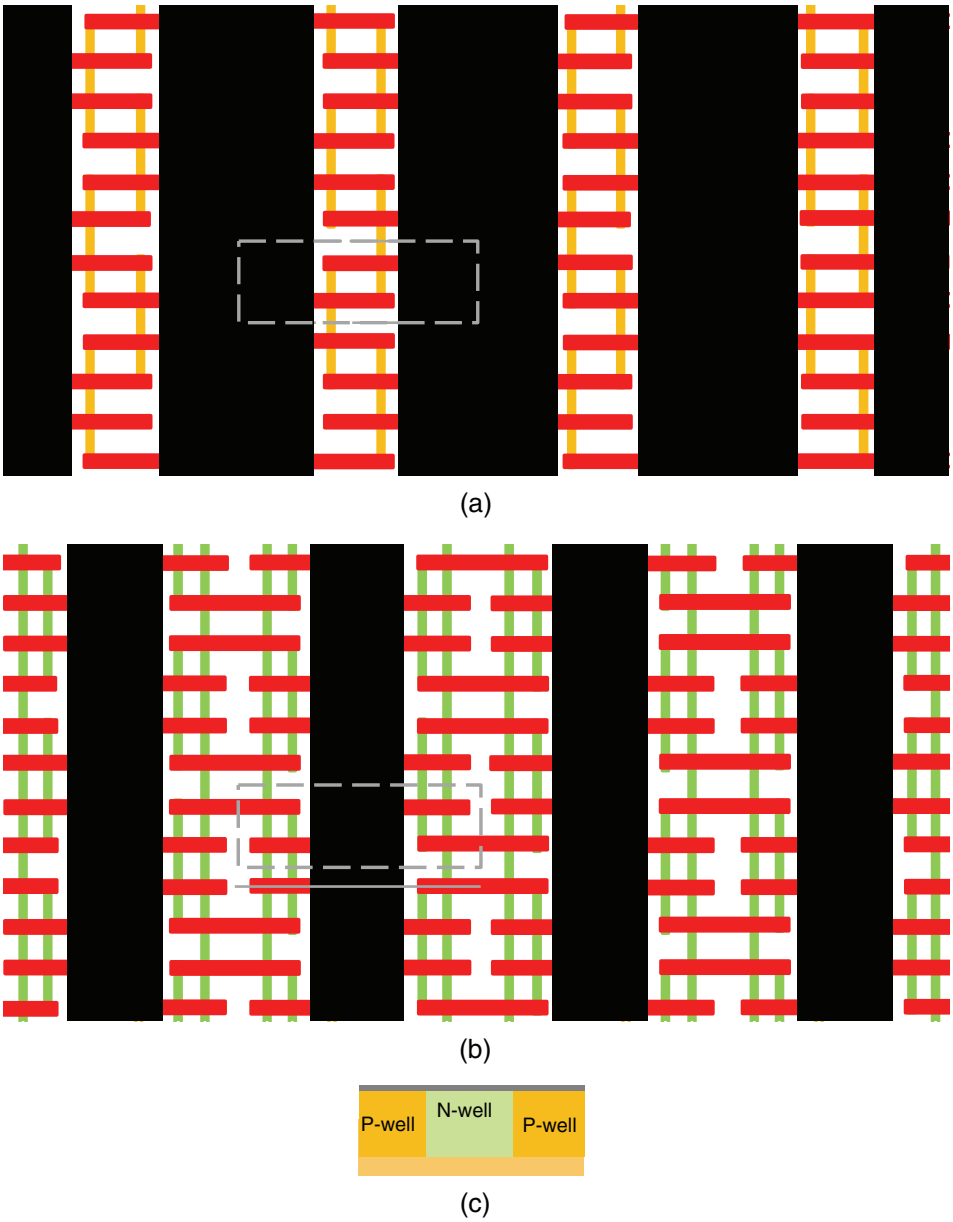


Figure 3.25 (a) N-well mask and (b) p-well overlapped with the fin and gate of a SRAM array, and (c) cross-section along the solid line after both well implantations.

A comparison of Figs. 3.25(a) and (b) shows that the n-well is for PMOS and the p-well is for NMOS; the two masks are complementary in a SRAM array. It is easier to form wells with ion implantation before the fin etch. The trade-off is the etch rates of p-type doped silicon and n-type doped silicon are different, which can cause uneven etch profile and fin heights of n-type and

p-type fins during the silicon fin etch process. Isolation junction for n-well and isolation junction p-well can be co-implanted during n-well implantation and p-well implantation, respectively. Isolation junctions are not shown in Fig. 3.25(c).

Next, the sacrificial oxide is stripped and the wafer is cleaned [Fig. 3.25(c)]. Pad oxide is grown, followed by silicon nitride deposition. A dielectric HM layer and dummy layer are also deposited for SADP. The fin dummy mask [shown in Fig. 3.26(a)] is applied, and a dummy pattern is etched. Figure 3.26(b) overlaps the fin dummy mask with the SRAM array image to show the location of spacers that will be cut into a fin in later processes.

After dummy etch and PR strip and clean, a conformal dielectric film is deposited. A vertical etch-back process forms spacers on the sidewall of the dummy patterns [Fig. 3.26(c)]. After the dummy pattern is removed, the spacer pattern doubles the pattern density of the original fin dummy pattern. The dielectric HM is then etched with the spacer pattern, which will later be cut into the final fin pattern. The top-down image at this stage of the process is shown in Fig. 3.26(d), and the cross-section along the solid line is shown in Fig. 3.26(e).

The aforementioned SADP cannot be used for 10-nm and 7-nm nodes; they need SAQP, with two spacer formations and two mandrel layers. Photolithography patterns the first mandrel layers. The first spacer doubles the pitch density of the photolithography patterns and transfers the pattern to the second mandrel layer. The second spacer doubles the pattern density again, after removal of the second mandrel pattern; the second spacer patterns can be transferred to the underlying hard mask to etch the silicon fins.

Figure 3.27(a) is a fin cut mask that overlaps with the fin dummy mask and SRAM array image. It can help us to see how the cut mask helps to form the final fin patterns. Figure 3.27(b) is the fin cut mask. If the pattern density of the cut mask is too high for single patterning with 193-nm immersion lithography, then a multi-patterning process may be needed. Figures 3.27(c) and (d) are two cut masks that are split from the cut mask in Fig. 3.27(b). Both cut mask 1 [Fig. 3.27(c)] and cut mask 2 [Fig. 3.27(d)] have half of the pattern density of the original.

The peripheral area needs another cut mask to remove the fins between the NMOS and PMOS. This cut mask has a line-space pattern that parallels the fin pattern shown in Figure 3.26(d). Some SRAM arrays are designed with a fixed fin pitch, and the fins between the NMOS and PMOS are removed using this mask. The benefit of this design is the standardization of the design of both SRAM arrays and logic gates; however, the trade-off is the pin spacing of this kind of SRAM array must be either 2 or 3 times the fin pitch and thus may not have the best power consumption and performance.

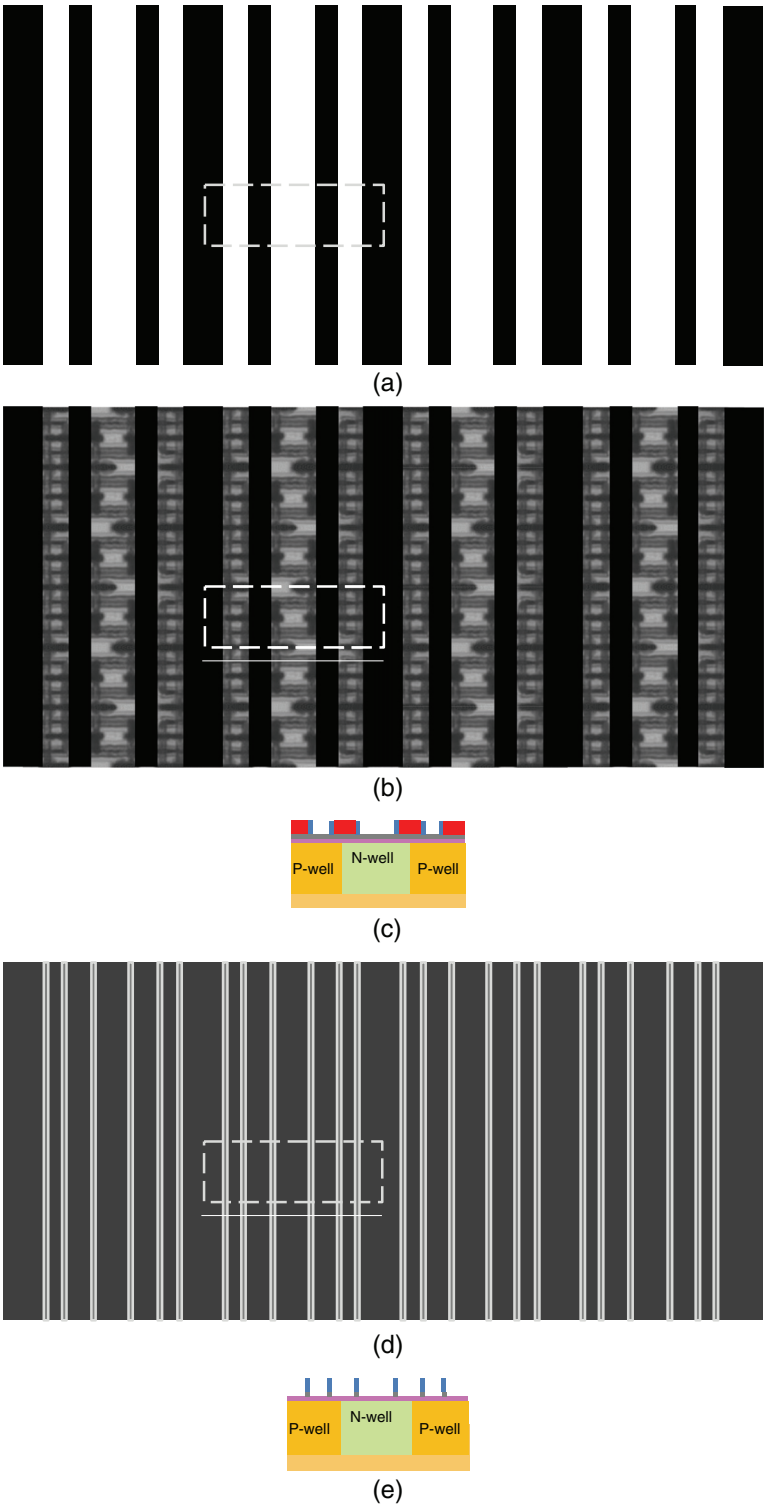


Figure 3.26 (a) Fin dummy mask, (b) fin dummy mask overlapping with a SRAM array image, (c) cross-section of a dummy pattern with a spacer, (d) top-down view of spacer patterns, and (e) cross-section of spacers.

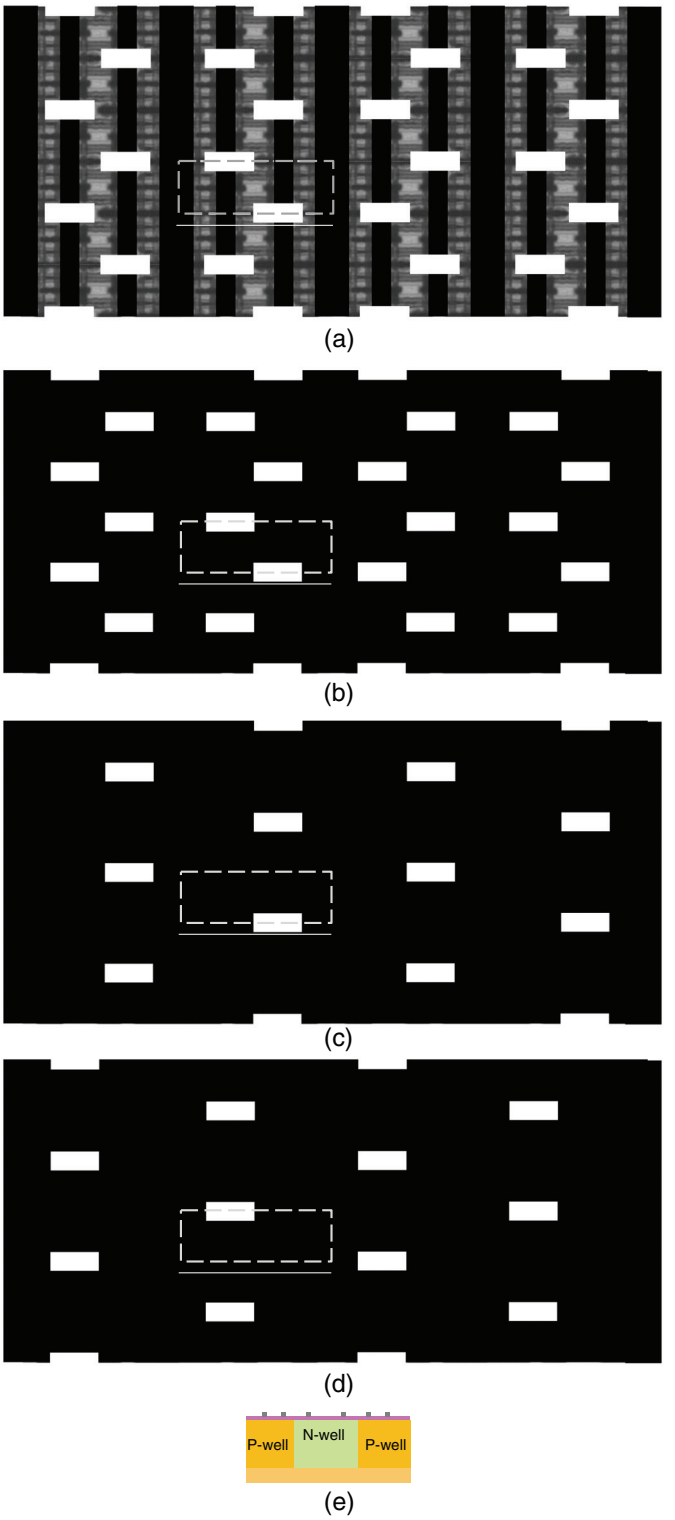


Figure 3.27 (a) Fin cut mask overlaps with fin dummy mask and SRAM array image, (b) illustration of fin cut mask, (c) fin cut mask 1, (d) fin cut mask 2, and (e) cross-section along the solid line in (a).

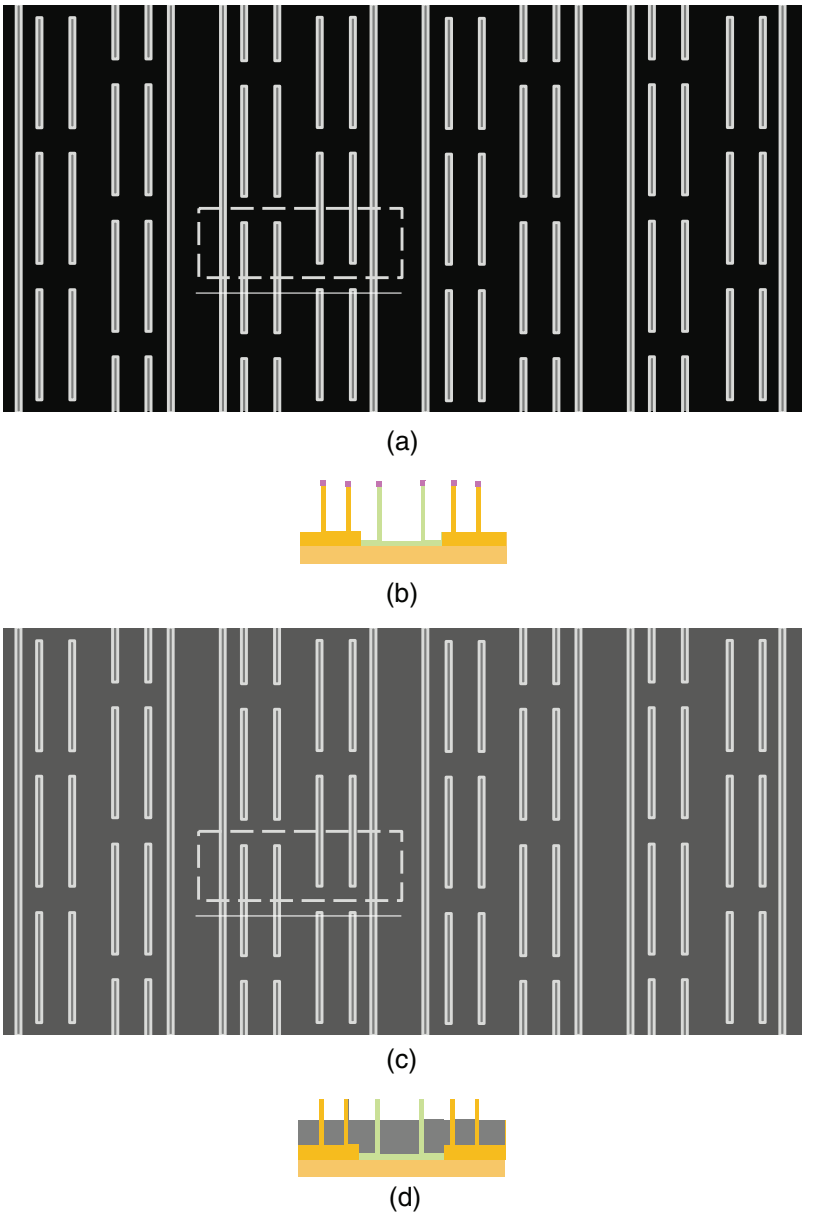


Figure 3.28 (a) Final fin pattern after silicon fin etch with a silicon nitride HM; (b) cross-section along the solid line in (a); (c) fin pattern after ILD0 recess, nitride and pad oxide strip, and wafer clean; and (d) cross-section along the solid line in (c).

After applying the cut mask(s), the final fin pattern can be etched on the silicon nitride layer, using the pattern formed on the dielectric HM. The silicon fin can be etched with the silicon nitride pattern. The top-down view of the silicon fin pattern is illustrated in Fig. 3.28(a). Figure 3.28(b) is the

cross-section along the solid line in Fig. 3.28(a). The etch rates of p-type doped silicon and n-type doped silicon are different. After wafer clean, ILD0 is deposited or spun-on and cured. ILD CMP stops on the silicon nitride surface, and after wafer clean, ILD is recessed. Silicon nitride, pad oxide strip, and wafer clean finish the fin formation process module. The top-down view at this process stage is illustrated in Figure 3.28(c). Figures 3.28(a) and (c) are very similar; the difference is that Fig. 3.28(a) shows only silicon fins with a large aspect ratio and deeper gap between the fins, whereas Fig. 3.28(c) shows the silicon with recessed ILD0, which has a lower aspect ratio and shallower gap between the fins. Figure 3.28(d) is the cross-section along the solid line in Fig. 3.28(c).

After silicon nitride and pad oxide strip and wafer clean, the dummy gate oxide and dummy gate polysilicon layers are deposited. After polysilicon CMP and wafer clean, a HM layer and dummy patterning layer is deposited on top of polysilicon. The gate dummy mask [shown in Fig. 3.29(a)] is applied to pattern the dummy layer. Figure 3.29(b) is the gate dummy mask that overlaps with the SRAM array image. The gate in the SRAM is formed by the spacer on the sidewall of the gate dummy pattern.

After PR strip and clean, a conformal film is deposited on the wafer surface, and a vertical etch process forms the sidewall spacer. The gate dummy patterns are removed, and the wafer is cleaned. The spacer pattern doubles the pattern density of the gate dummy pattern in this SADP process.

Next, the HM is etched using the spacer patterns; the top-down view of a SRAM array is illustrated in Fig. 3.29(c). The polysilicon gate is etched with the HM pattern. This etch process is very challenging because the pattern is very small and the substrate is not flat. At the top of the fin, the etch process must stop on the dummy gate oxide while polysilicon between the fins is completely removed. Any remaining polysilicon between the fins is a critical defect because it can cause an electric short between the gates. If the dummy gate oxide on top of the fins has been broken through, the chemical that etches polysilicon will also etch a single crystalline silicon fin quickly, which could cause fin loss, also a killer defect. Figure 3.29(d) illustrates the top-down view of a SRAM array after completing the gate etch. A comparison of Figs. 3.29(c) and (d) shows that Fig. 3.29(c) only shows gate patterns on the wafer surface, whereas Fig. 3.29(d) shows both surface gate patterns and subsurface fin patterns. Figure 3.29(e) shows the cross-section along the gate in a SRAM array indicated by a solid line in Fig. 3.29(d).

After wafer clean, an organic planarization layer (OPL) is applied, which planarizes the wafer surface. A HM layer is deposited on top of the OPL. Depending on the technology node, one or more gate cut masks are applied. Figure 3.30(a) shows the gate cut mask, and Fig. 3.30(b) shows the gate cut

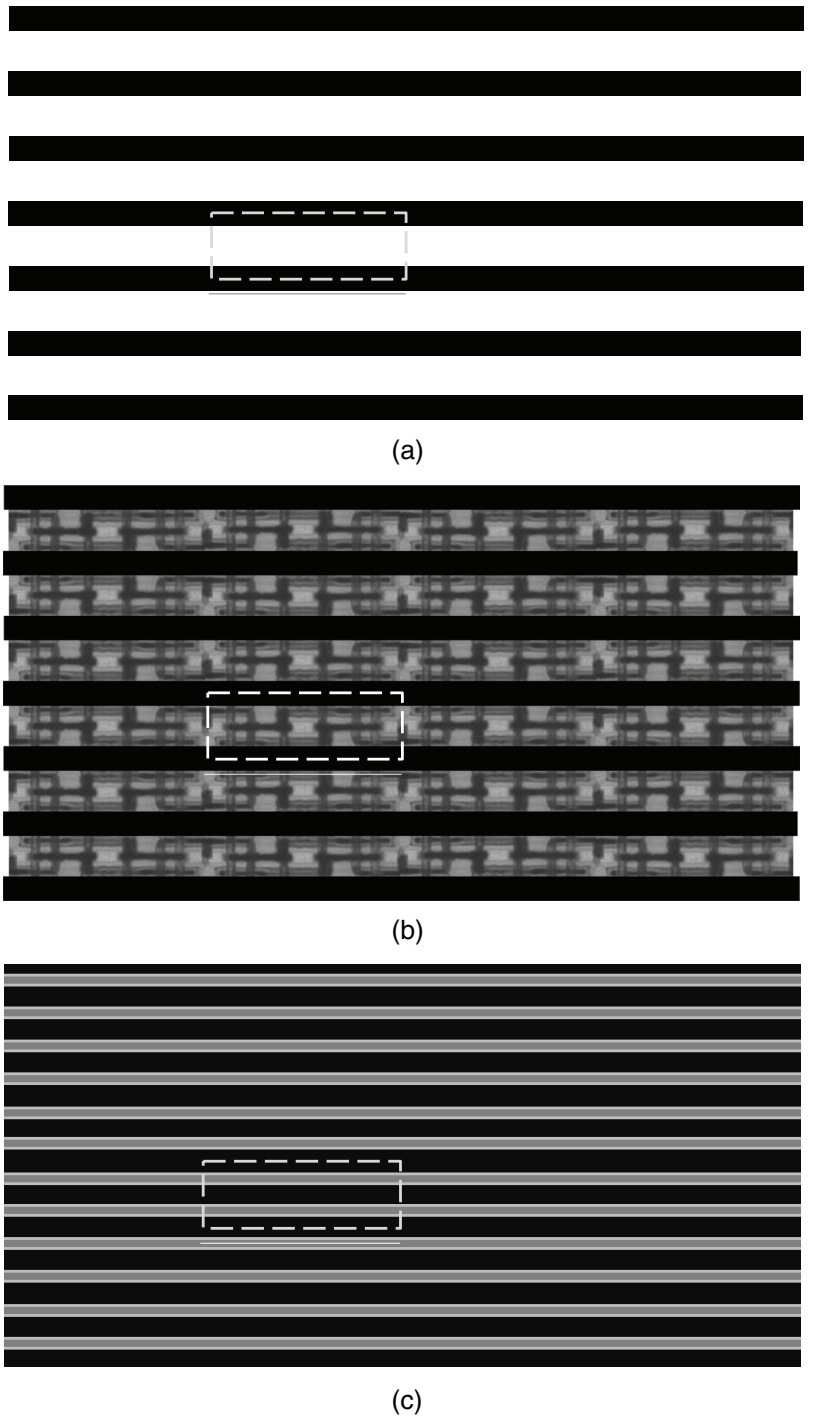


Figure 3.29 (a) Gate dummy mask, (b) gate dummy mask overlaps with SRAM array image, (c) wafer surface after HM etch, (d) wafer surface after polysilicon gate etch, and (e) cross-section along the polysilicon gate indicated by a solid line in (d).

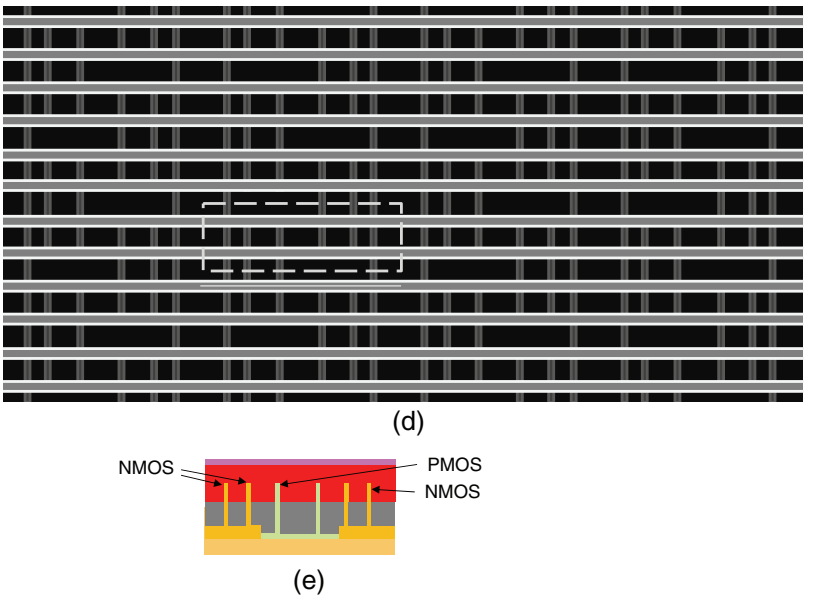


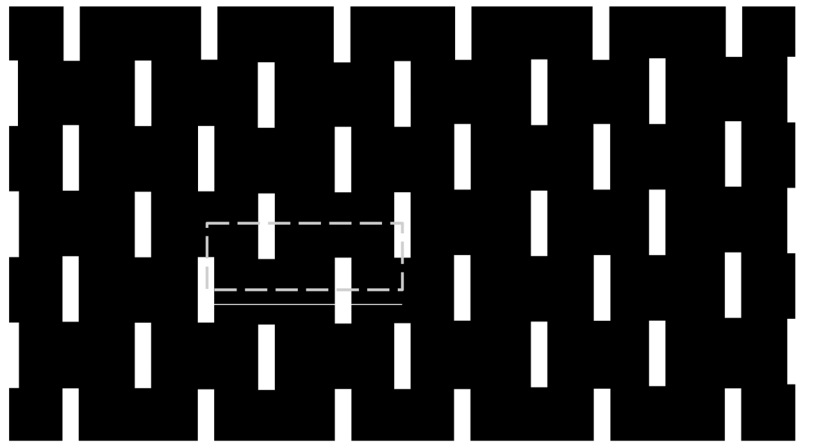
Figure 3.29 (Continued)

mask overlapping with the top-down TEM image of a 24-cell SRAM array. Figures 3.30(c) and (d) illustrate the two gate cut masks that are split from the single mask shown in Fig. 3.30(a).

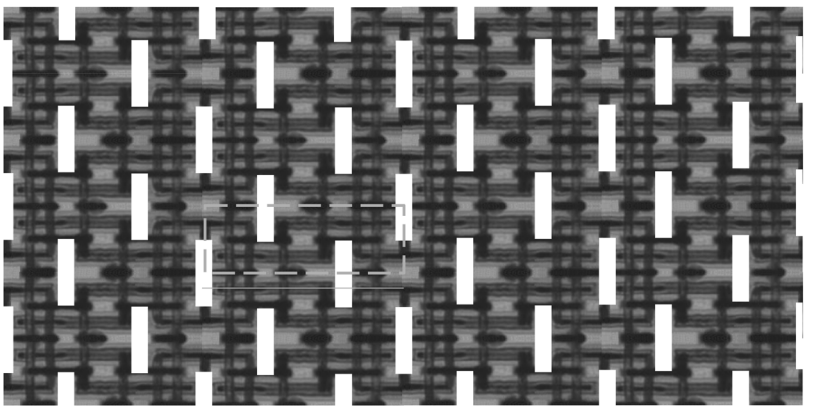
After the gate cut mask patterns etched on the hard mask, the etch process that is designed to etch polysilicon and OPL at the similar etch rate is applied, which cut the polysilicon line-space patterns shown Fig. 3.29(c) into the SRAM gate pattern shown in Figure 3.30(e). There are two types of gates in a SRAM array: one is an inverter gate, which has a NMOS and a PMOS, and the other is a pass gate, which only has one NMOS. In this case, the NMOS of the inverter is a double-fin FinFET, whereas the NMOS pass gate is a single-fin FinFET. Figures 3.31(a) and (b) illustrate the cross-sections that are indicated by the solid line in Figure 3.30(e), before gate cut and after gate cut, respectively.

After OPL strip and wafer clean, a thin, conformal, heavily-phosphorus-doped silicon oxide film (commonly called phosphosilicate glass, or PSG) is deposited. The NMOS SDE mask shown in Fig. 3.32 is applied, and the PSG on the PMOS fins and gates is removed. After PR strip and clean, a mini-second high-temperature anneal is performed that injects the phosphorus into the NMOS fins to form the NMOS SDE junction.

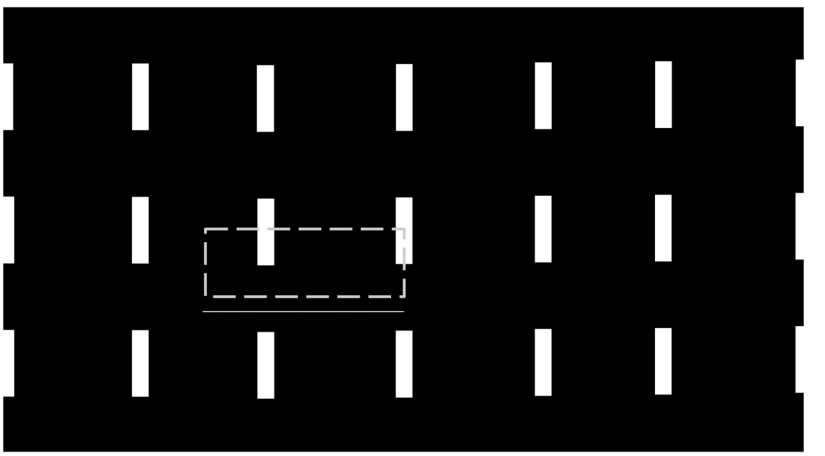
After PSG strip and clean, a thin, conformal, heavily-boron-doped silicon oxide (commonly called borosilicate glass, or BSG) is deposited. The PMOS



(a)

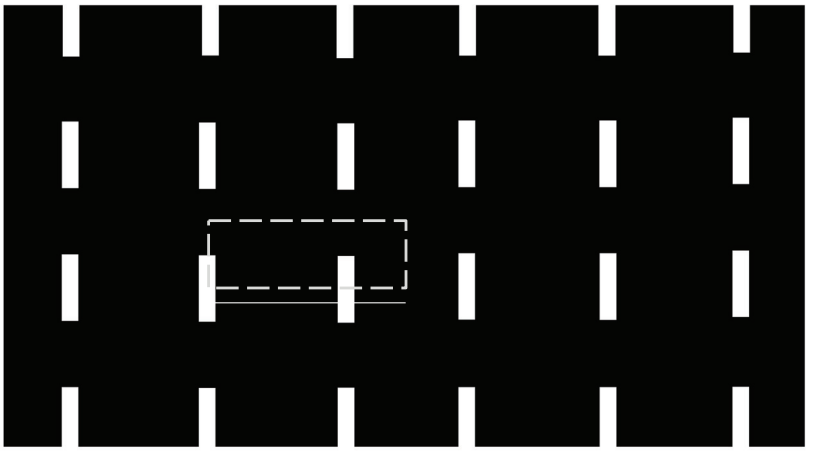


(b)

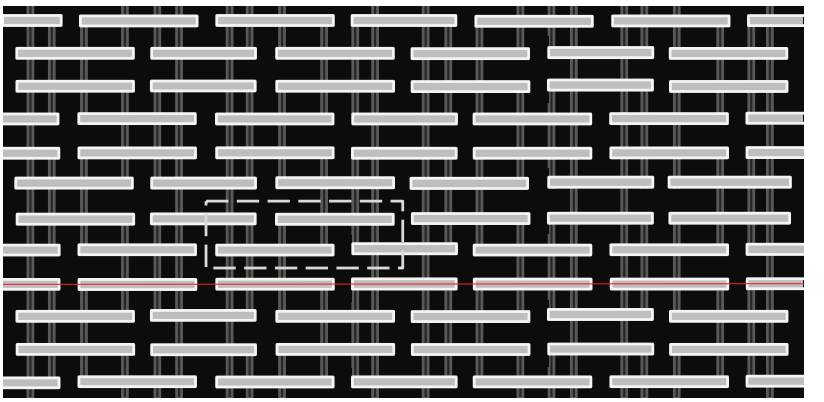


(c)

Figure 3.30 (a) Gate cut mask, (b) gate cut mask overlapping with the SRAM array image, (c) gate cut mask 1, (d) gate cut mask 2, and (e) top-down view of the SRAM after polysilicon gate cut.



(d)



(e)

Figure 3.30 (Continued)

SDE mask shown in Fig. 3.33 is applied, and the BSG on the NMOS fins and gates is removed. After PR strip and clean, another mini-second high-temperature anneal is performed that injects the boron into the PMOS fins to form the PMOS SDE junction.

After BSG strip and clean, a conformal dielectric liner is deposited on the wafer surface, and a NMOS S/D mask, as shown in Fig. 3.34(a), is applied. Figure 3.34(b) is the NMOS S/D mask overlapped with the SRAM image. The location of the dot-dash-dot line, which indicates the cross-section of the S/D formation, is different than the previous solid line, which indicates the top of the gate.

After PR patterning, as shown in Fig. 3.35(a), a vertical etch process is performed, and the dielectric liner on top of the NMOS fin is removed to

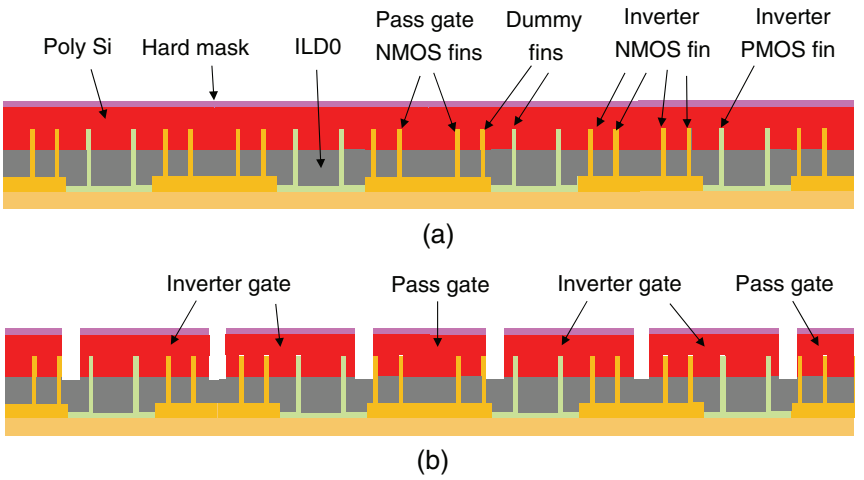


Figure 3.31 Cross-sections along the gate: (a) before gate cut and (b) after gate cut.

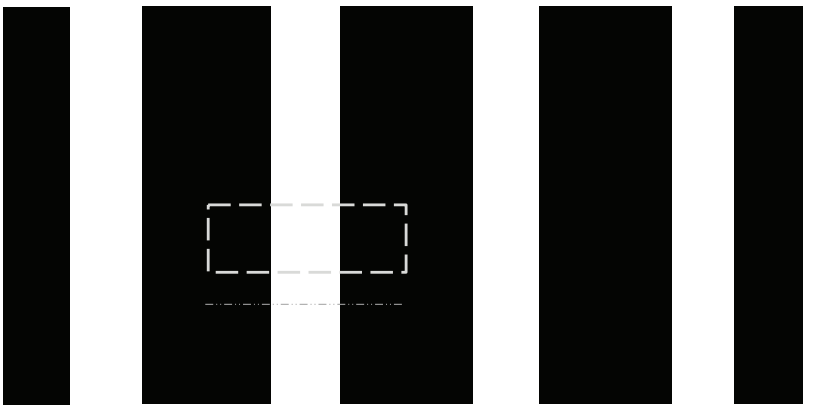


Figure 3.32 NMOS SDE mask.

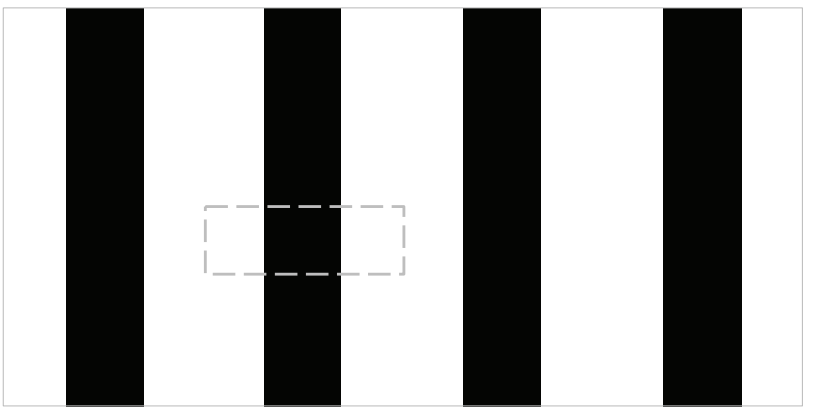


Figure 3.33 PMOS SDE mask.

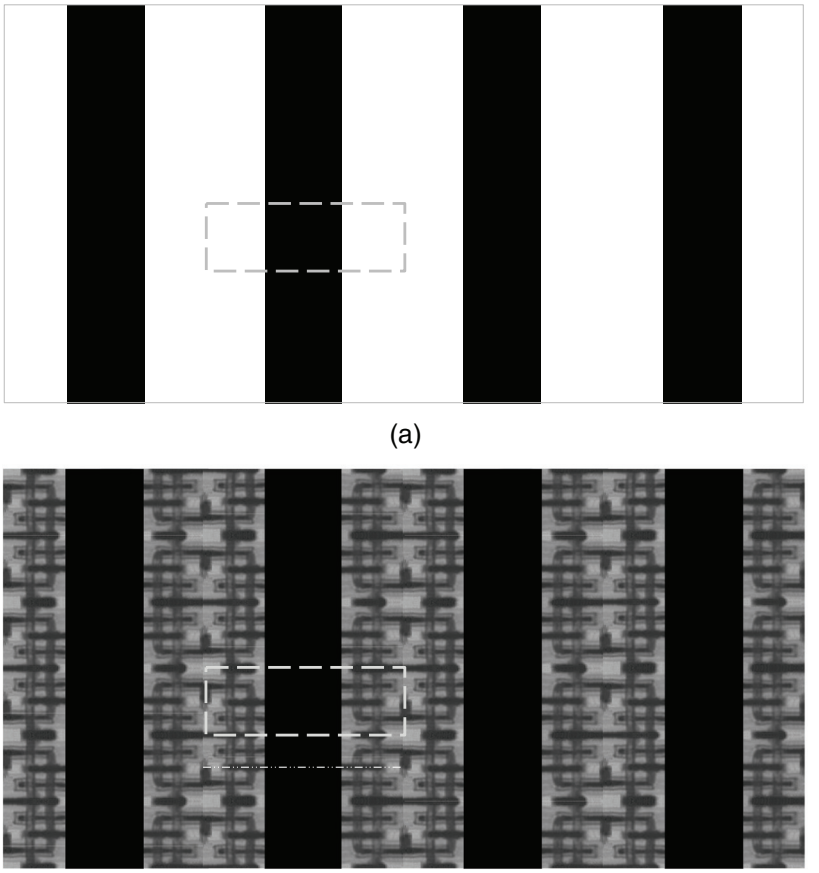


Figure 3.34 (a) NMOS S/D mask and (b) NMOS S/D mask overlapped with the SRAM array image. The dot-dash-dot line indicates the S/D cross-section drawing.

expose the silicon of the NMOS fin. Because the hard mask is still on top of the gates, the polysilicon gates are still wrapped by dielectrics, which will protect them while the fin silicon is recessed. Figure 3.35(b) shows the cross-section of the silicon recess of the NMOS fin along the dot-dash-dot line in Fig. 3.34(b). After PR strip and clean, as shown in Fig. 3.35(c), single-crystalline silicon heavily doped with phosphorus (SiP) is grown by a selective epitaxial growth process in areas of exposed silicon [Fig. 3.35(d)]. SiP may have 1–2% carbon, combined with silicon recessed into the channel, to create tensile strain of the NMOS channel, which can increase channel electron mobility. This process forms the NMOS S/D.

After SiP SEG, the dielectric liner is stripped in a wet etch process, which is highly selective to silicon, SiP, and silicon oxide. After the wafer is cleaned, another dielectric liner is deposited on the wafer surface with an ALD process.

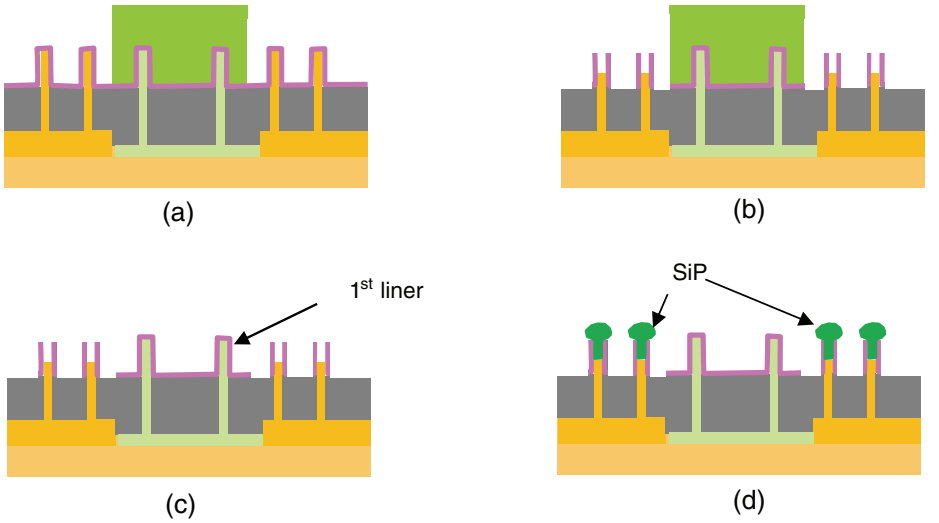


Figure 3.35 Cross-section (a) after NMOS S/D mask, (b) after liner vertical etch and Si recess, (c) after PR strip and clean, and (d) after SEG of SiP.

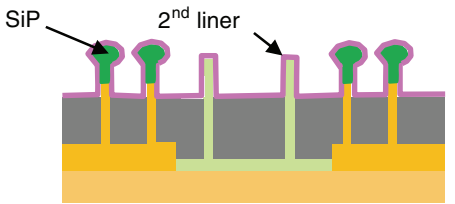


Figure 3.36 Cross-section of the SRAM fin area after the second liner deposition.

This liner will be used to protect the NMOS S/D during PMOS S/D formation. Figure 3.36 shows the cross-section after the second liner is deposited.

The PMOS S/D mask shown in Fig. 3.37(a) is then applied. Figure 3.37(b) is the PMOS S/D mask overlapping with the SRAM array image. Figure 3.38(a) illustrates the cross-section after PR patterning with a PMOS S/D mask. A vertical etch process is used to remove the second liner from the top of the PMOS fin, as shown in Fig. 3.38(b). Similar to the NMOS S/D process, the top of gate is protected by the hard mask and sidewalls are protected by the spacers formed by the second liner. The silicon in the PMOS fin is recessed [Fig. 3.38(c)], and SiGe heavily doped with boron is grown by a SEG process, as shown in Fig. 3.38(d).

After wafer clean and anneal, the FinFET SRAM has been formed. More process steps are needed to form gate-last high-*k* metal gate devices. At first, ILD1 is deposited [Fig. 3.39(b)], and a CMP process planarizes the ILD1 to

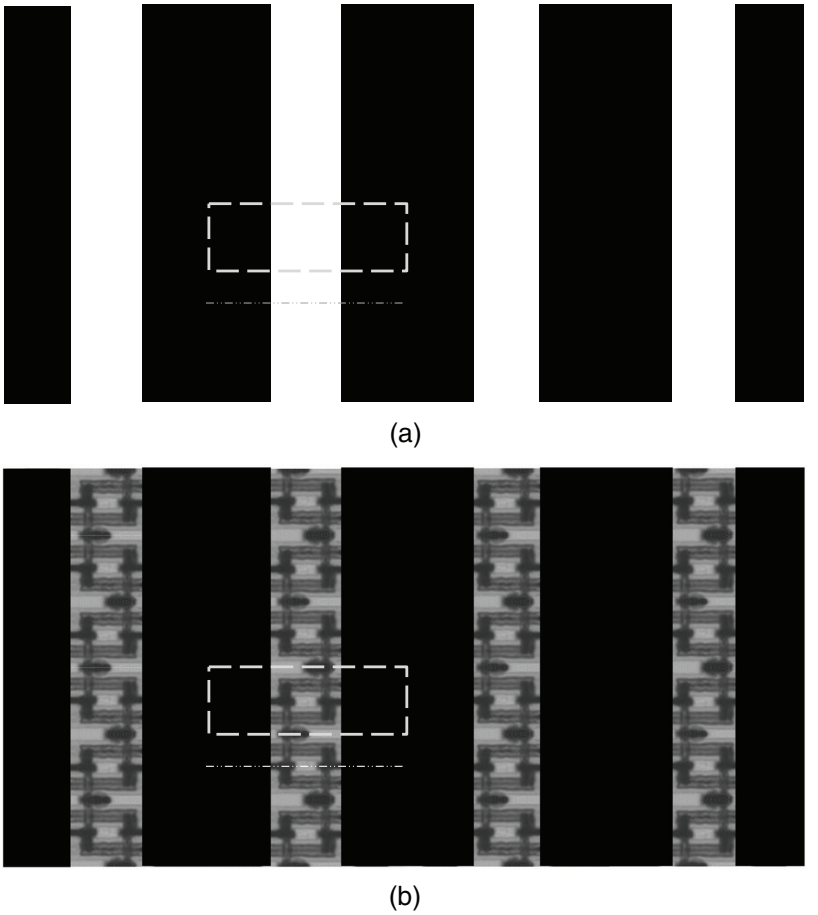


Figure 3.37 (a) PMOS S/D mask in a SRAM array and (b) PMOS S/D mask overlapping with a SRAM array image. The dot-dash-dot line indicates the location of the cross-section.

expose the polysilicon dummy gate, as shown in Fig. 3.39(c). Note that the cross-sections are along the solid line in the SRAM array image [Fig. 3.39(a)], which is on top of the gate, whereas the dot-dash-dot line used for the cross-section of S/D formation is on top of the fins and SDC.

After the removal of dummy polysilicon gates using a highly selective wet etch process, the dummy gate oxide is also stripped, and the wafer is cleaned [Fig. 3.39(d)]. It forms gate trenches with fins inside. In this SRAM array, there are two NMOS fins and one PMOS fin in the inverter gate trenches; and there are two NMOS fins in the pass gate trenches. After wafer clean and very thin silicon oxidation, hafnium-oxide-based high-*k* dielectric is deposited with ALD, followed by a PMOS work-function metal TiN and TaN barrier layer (also with an ALD process), as shown in Fig. 3.39(e).

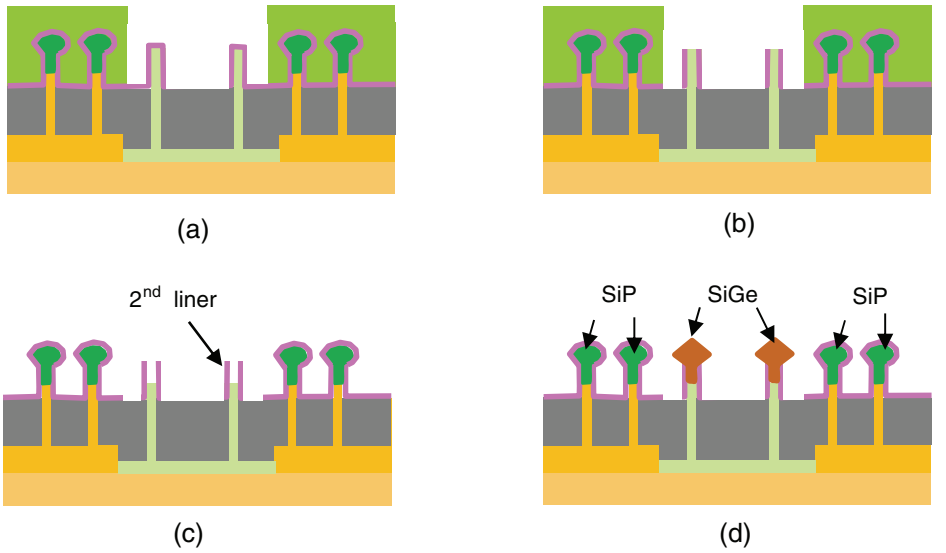


Figure 3.38 Cross-section of SRAM PMOS S/D formation: (a) PMOS S/D litho, (b) PMOS spacer etch, (c) Si recess of PMOS fin, and (d) SEG of boron-doped SiGe.

The NMOS metal gate (MG) mask, as shown in Fig. 3.40(a), is then applied. Figure 3.40(b) is the NMOS MG mask overlapping with the SRAM array image.

After PR patterning [Fig. 3.41(a)], the TaN barrier layer in the NMOS area is etched away to expose the TiN layer. After PR strip and clean [Fig. 3.41(b)], TiAl is deposited, as illustrated in Fig. 3.41(c). With an annealing process, it reacts with TiN to form the NMOS work-function metal TiAlN in the NMOS areas, whereas the PMOS work-function metal TiN is protected by a TaN barrier layer [Fig. 3.41(d)]. TiN liner ALD and bulk WCVD fill the gate trenches, as shown in Fig. 3.41(e). MG CMP removes W, TiN, TiAl, TaN, TiN, and HfOx from the wafer surface and finishes the gate-last HKMG processes [Fig. 3.41(f)].

At this point the FEoL processes are finished; they are followed by MEoL processes that create contact plugs to the source, drain and gates, and form the local interconnects. This step is followed by post-CMP clean, ILDx, and HM layers [Fig. 3.43(a)]. The term ILDx is used because it is a sacrificial layer that will be removed and not in the final device. The SDC mask shown in Fig. 3.42(a) is applied. Figure 3.42(b) is the SDC mask overlapping with the SRAM array image. Two masks may be needed for this layer for an advanced-technology node. Figures 3.42(c) and (d) illustrate two masks that are split from the SDC mask in Fig. 3.42(a). The contacts are not holes but trenches. For the SDC layers, all of the contact trenches are parallel to the gates.

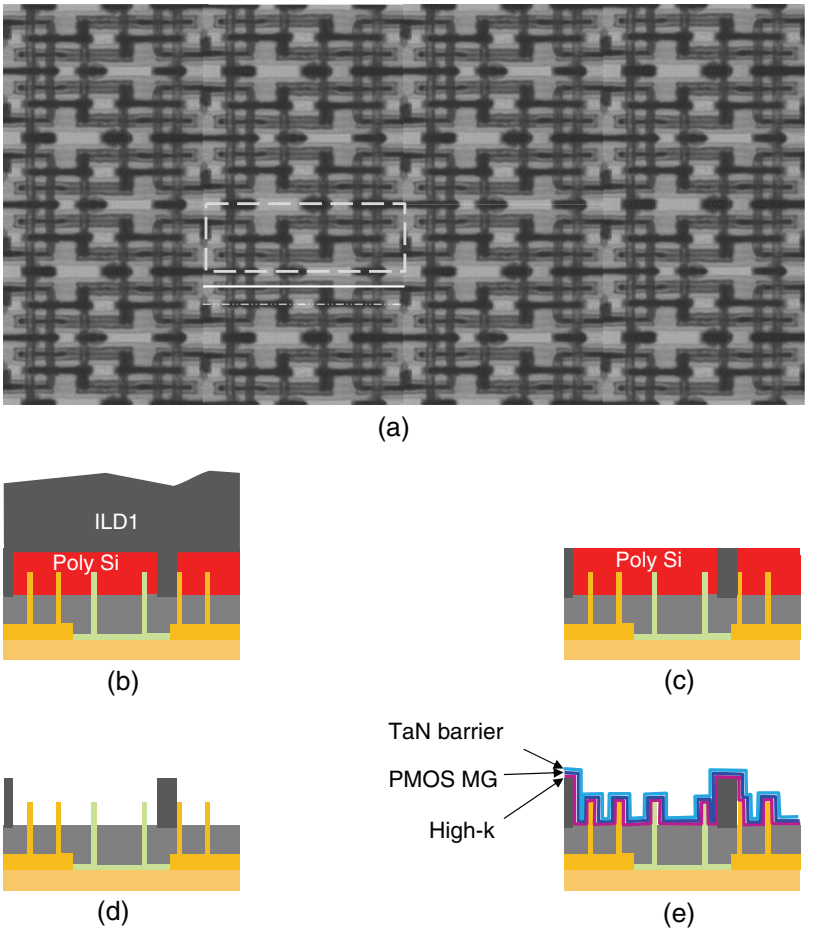


Figure 3.39 SRAM HKMG process 1: (a) SRAM array image and cross-sections are along the solid line, (b) ILD1 deposition, (c) ILD1 CMP, (d) dummy polysilicon gate removal, and (e) high-*k*, PMOS metal gate, and TaN barrier layers ALD.

The photolithography-pattern PR, as shown in Fig. 3.43(b), and the HM are etched with PR patterns [Fig. 3.40(c)]. If double patterning is needed after PR strip and clean, followed by another cycle of litho-etch, then ILDx is etched with HM patterns, and ESL is etched through during the over-etch step, as shown in Fig. 3.43(d). After wafer clean, Ti thin film is deposited at the bottom of the SDC trenches. TiN liner and W filler film are deposited, as shown in Fig. 3.43(e). RTA forms $TiSi_2$ at the bottom of the trenches and reduces contact resistance. WCMP removes bulk W, TiN, and Ti from the wafer surface, and finishes this SDC module, as shown in Fig. 3.43(f). Note that the cross-sections in Fig. 3.43 appear along the dot-dash-dot line in Fig. 3.42(b).

The wafer is now ready for the gate contact (GC) layer. After ESL, ILD2, and HM deposition [Fig. 3.43(a)], a GC mask [Fig. 3.44(a)] is applied.

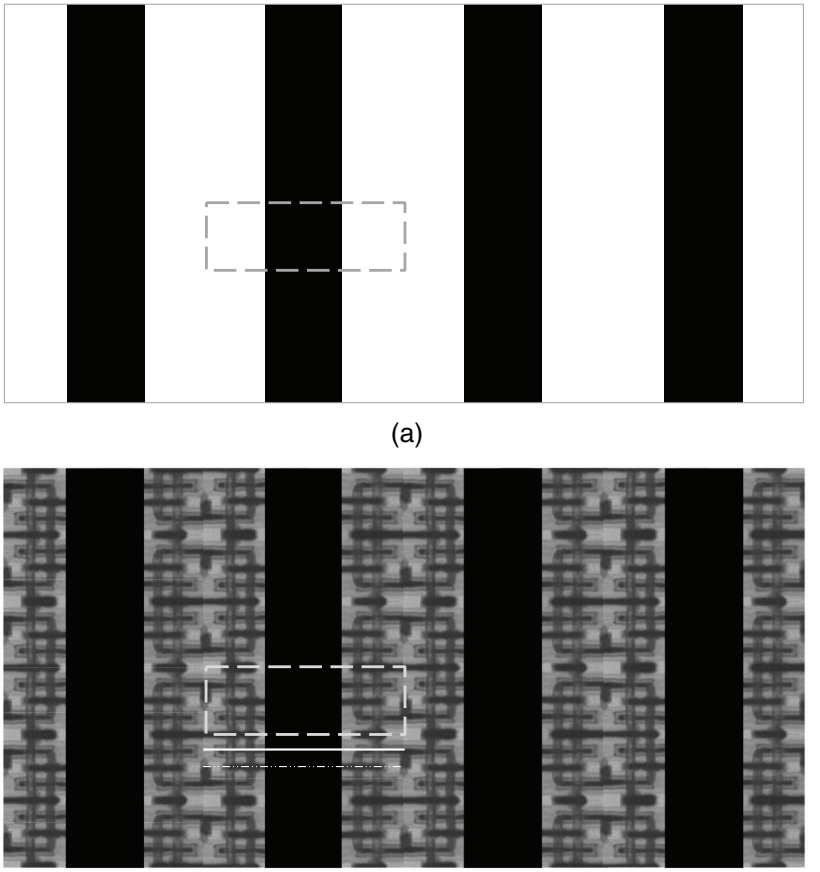


Figure 3.40 (a) NMOS MG mask and (b) NMOS MG mask overlapping with the SRAM array image.

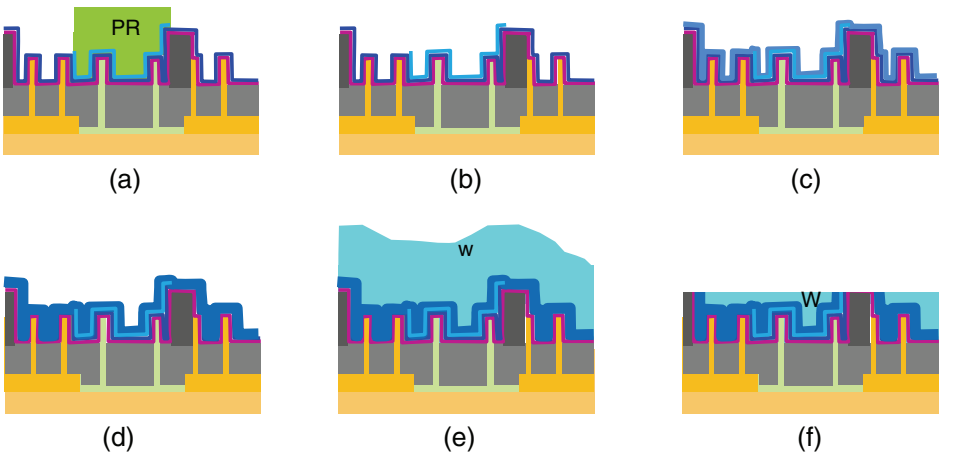


Figure 3.41 (a) NMOS MG mask PR patterning; (b) TaN etch, and PR strip and clean; (c) TiAl ALD; (d) TiAlN formation and TiN ALD; (e) WCVD; and (f) WCMP.

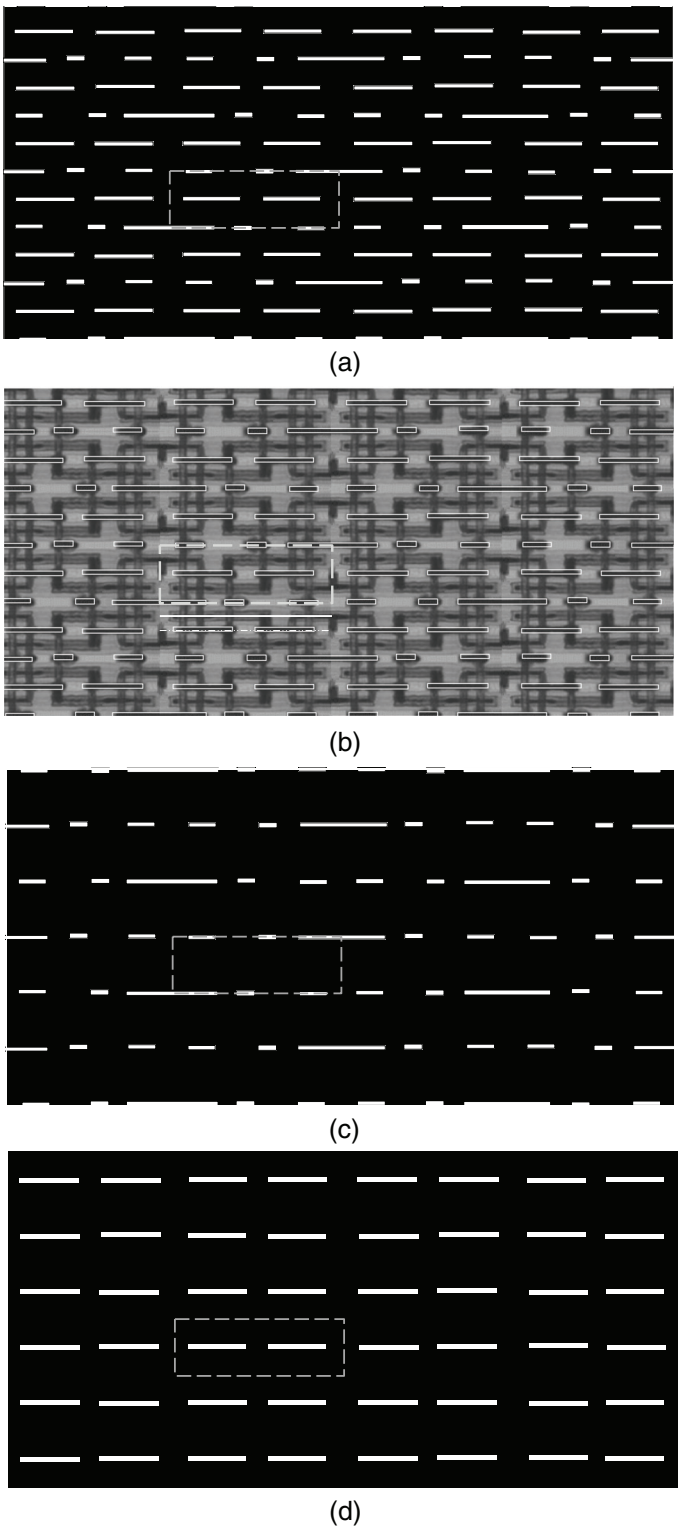


Figure 3.42 (a) SDC mask and (b) SDC mask overlapping with the SRAM array image; (c) split SDC mask 1 and (d) split SDC 2.

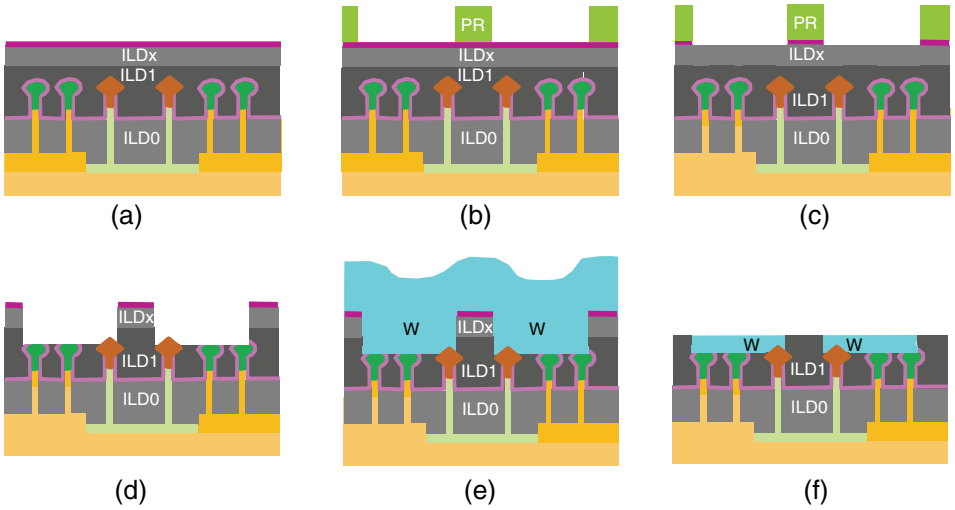
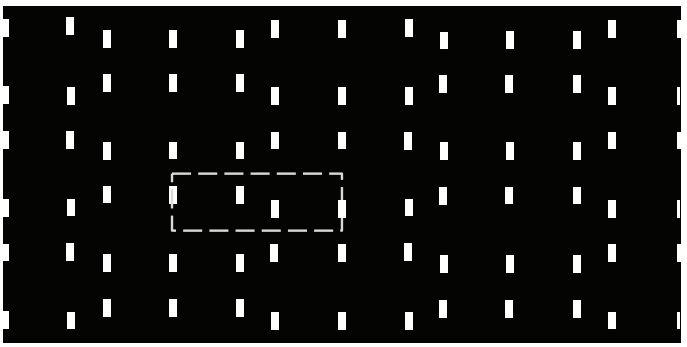


Figure 3.43 SDC module: (a) ILDx and hard mask deposition; (b) SDC PR patterning; (c) HM etch; (d) PR strip, clean, and ILD etch; (e) wafer clean, and Ti, TiN, and W CVD; and (f) WCMP.

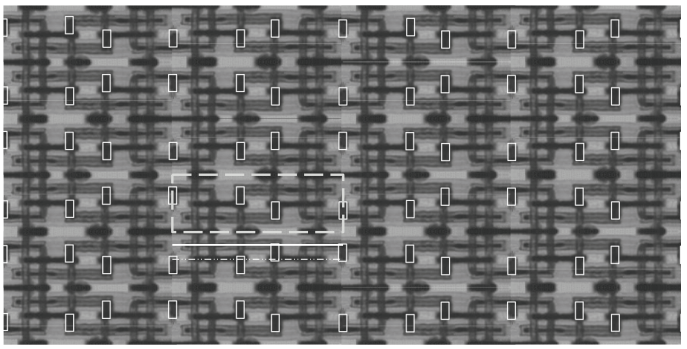
Figure 3.44(b) is the GC mask overlapping with the SRAM array image. Figures 3.44(c) and (d) illustrate the two GC masks that split from Fig. 3.44(a); they are used when a single exposure of 193-mm immersion lithography cannot pattern the dense hole pattern and thus LELE double-patterning processes are needed. The solid line in Fig. 3.44(b) indicates the cross-section on top of the gates and will be used in the cross-section of the GC process steps (Fig. 3.45).

Figure 3.45 shows the cross-sections along the solid line of Fig. 3.44(b), which lands on top of the gates in the SRAM. After GC PR patterning [Fig. 3.45(b)], the HM is etched, as shown in Fig. 3.45(c). After PR strip and clean, ILD2 and the ESL are etched using HM patterns [Fig. 3.45(d)]. TiN and W are deposited after wafer clean [Fig. 3.45(e)], and WCMP removes the bulk W and TiN from the wafer surface, thus finishing the MEoL processes. Because the BEoL processes have been covered multiple times already, this will end the section.

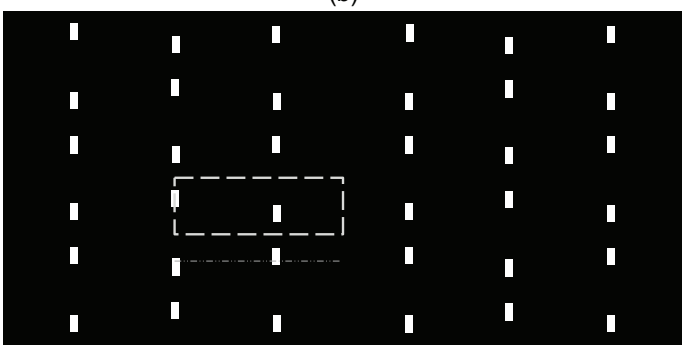
The contact to the S/D can also be formed with a self-aligned contact (SAC) process, which has been used in DRAM manufacturing for a long time. To apply a SAC process, the metal gate shown in Fig. 3.41(f) must be recessed. SiN CVD processes fill the trench formed by MG recess, and nitride CMP removes SiN on the wafer surface, leaving SiN on top of the MG. A trench-style mask [shown in Fig. 3.46(a)], where each trench aligns with the S/D, is applied. Figure 3.46(b) is the SAC mask overlapping with SRAM image, which shows that the source and drain are exposed. An etch process with high oxide-to-nitride selectivity is applied to etch away the oxide and



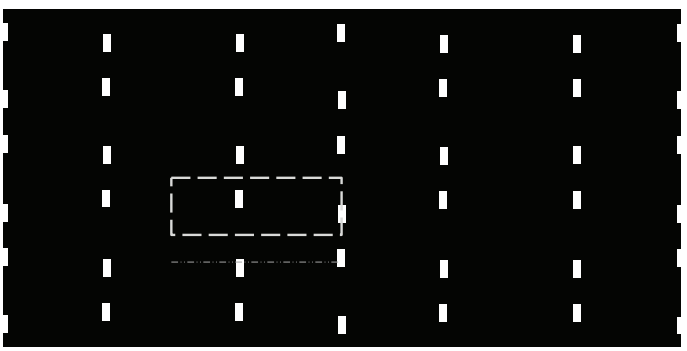
(a)



(b)



(c)



(d)

Figure 3.44 (a) GC mask, (b) GC mask overlapping with the SRAM array image, (c) GC mask 1, and (d) GC mask 2.

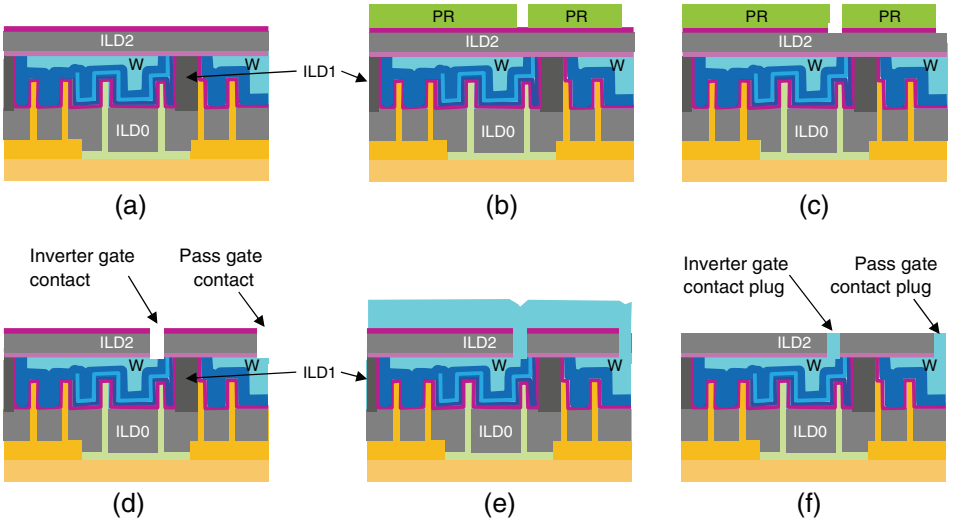


Figure 3.45 Gate contact module: (a) ESL, ILD2, and HM deposition; (b) GC PR patterning; (c) HM etch; (d) PR strip, clean, and ILD2 etch; (e) wafer clean, and TiN and W CVD; and (f) WCMP.

reach the S/D while removing very little of the SiN spacer and SiN on top of the gate. After RP strip, wafer clean, and TiN liner and W deposition, a CMP process removes W and TiN from the surface; W plugs connected to the S/D are formed [Fig. 3.46(c)].

3.6 FinFET CMOS Scaling

The development of FinFET technology allows for IC scaling into the third dimension. Previously, scaling always occurred in a planar (x or y) direction. However, the z direction becomes part of the scaling strategy. Section 3.1 established that the drive current I_D is proportional to the channel width: $I_D \propto \mu(k/t_{ox})(W/L)$. For FinFET technology, the channel width $W = 2H + CD_{fin}$. To make the FinFET a fully depleted device in the off state, the fin must be very thin ($CD_{fin} \sim < 10$ nm); thus, the channel width is primarily determined by the fin height. Increasing the fin height H can increase the channel width and improve both the drive current and device performance. Figures 3.47(a) and (b) illustrate the FinFET scaling from 22 nm to 14 nm, respectively, in Intel chips. Figures 3.47(c) and (d) are TEM images of a 22-nm and 14-nm FinFET, respectively, along the metal gate.

The FinFET structure can further scale down to 7 nm (perhaps even 5 nm) by increasing the fin height and pattern density with either EUV lithography or SAQP with multiple cuts. New materials—such as SiGe, Ge, and III-V materials such as InGaAs—could be used to replace silicon to form the fin

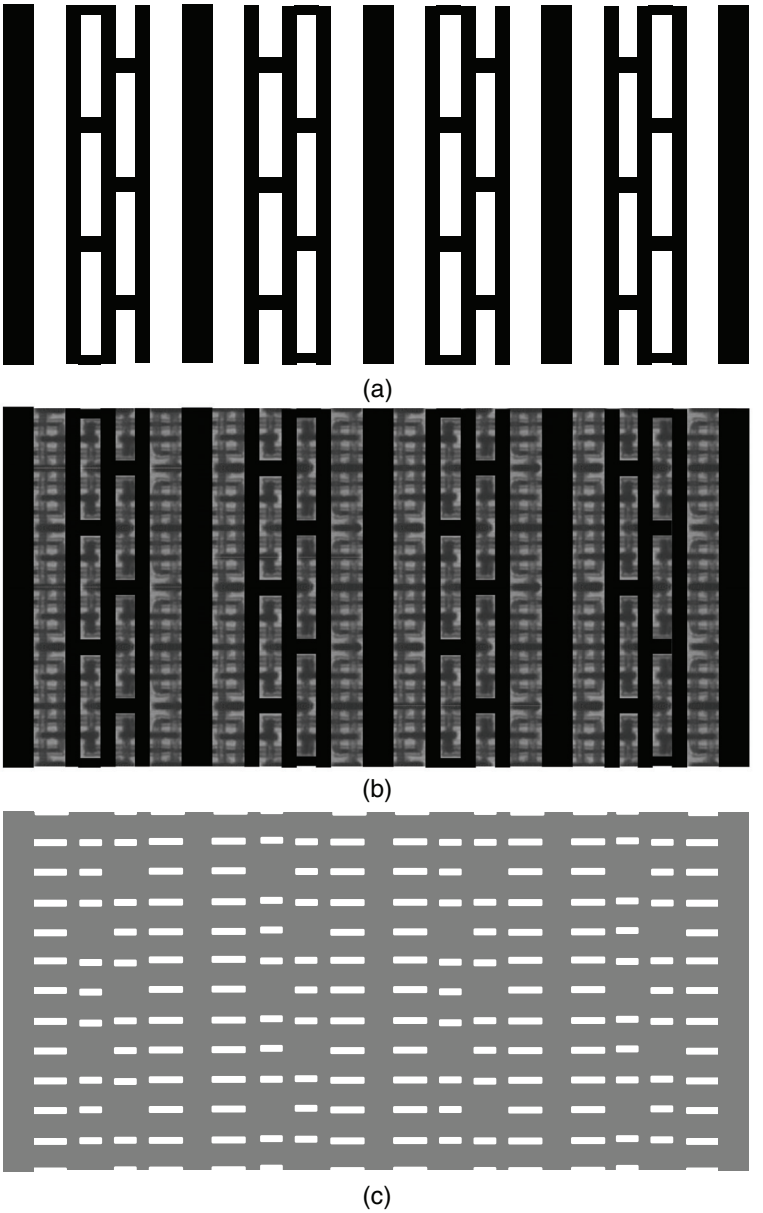


Figure 3.46 S/D self-aligned contact module: (a) SAC mask, (b) SAC mask overlap with SRAM image, and (c) S/D contact formed with the SAC process.

and increase carrier mobility of the channel. Of course, whether these scaling techniques will be developed and implemented in high-volume manufacturing will be determined by whether the scaling could bring financial profits in the long run.

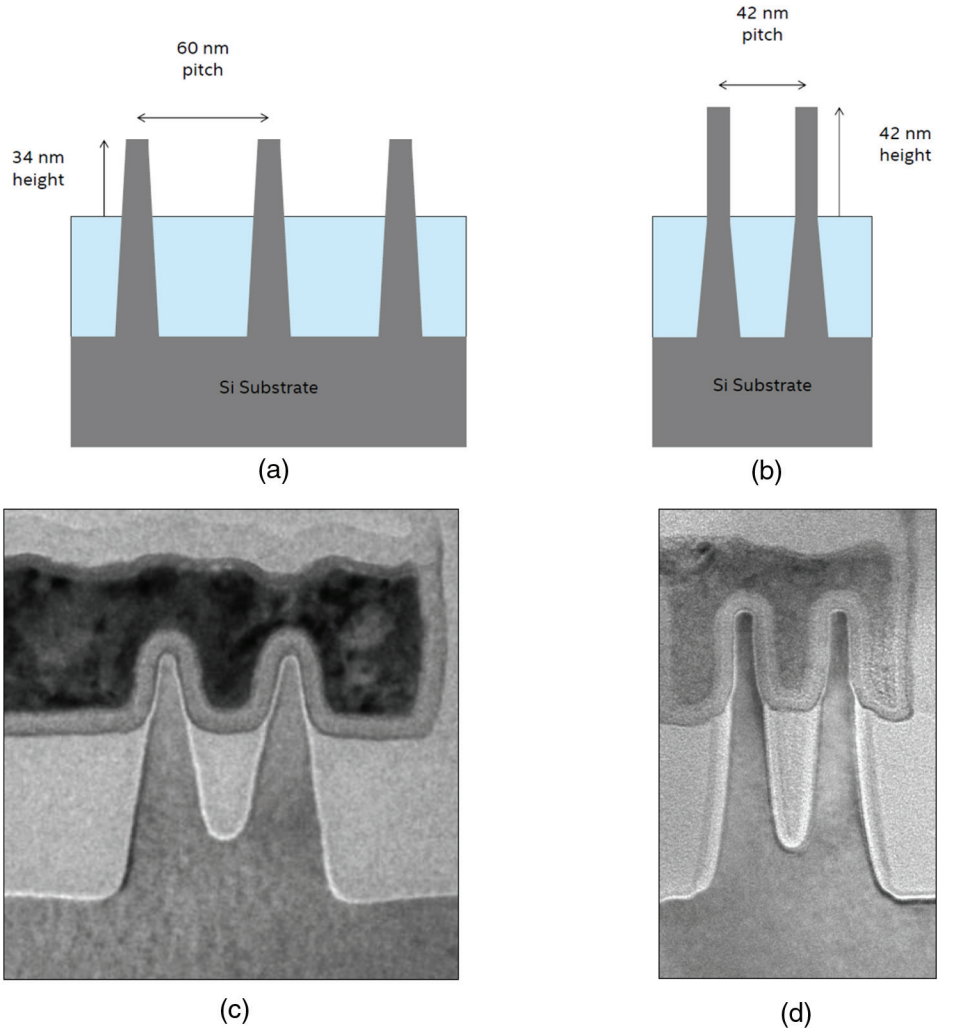


Figure 3.47 Intel's FinFET scaling from 22-nm to 14-nm technology: (a) 22-nm fin height and fin pitch, (b) 14-nm fin height and fin pitch, (c) TEM image of 22-nm FinFET, and (d) 14-nm FinFET TEM image. Image reprinted from Ref. 18 with permission from Intel.

3.7 Review Questions

1. What are the benefits of FinFET technology compared to planar MOSFET technology?
2. Describe the advantage and disadvantage of a SOI FinFET.
3. Compared to planar MOSFETs, what are the challenges of the FinFET polysilicon gate etch?
4. How many NMOS and PMOS are in a basic SRAM cell?
5. Arrange the three memories (DRAM, flash memory, and SRAM) in three ways: by speed, by cost, and by volatile/non-volatile.

6. What are the major changes that occur when a FinFET is scaled from the 22-nm to the 14-nm technology node?
7. Assuming that 14-nm to 10-nm scaling follows the same trends as 22-nm to 14-nm scaling, estimate the fin pitch and fin height of the 10-nm node.
8. For a planar MOSFET, the gate pattern has the smallest pitch and highest pattern density. Is this still true for a FinFET?
9. Why must the fin CD of a FinFET be very small ($\sim$ <10 nm)?
10. Why do mobile chip designers prefer FinFET technology over planar MOSFET technology?

Chapter 4
Summary and Future Trends of the 3D IC Process

4.1 Scaling MOSFET Technology after 14 nm

In the so-called "good old days," traditional scaling—which scaled the feature size of MOSFETs, such as gate length L, gate width W, and, more importantly, the gate oxide thickness t_{ox}—could reduce the IC manufacturing cost, improve device performance, and reduce power consumption. Table 4.1 shows the relationship between the scaling parameters and scaling factor α. The table shows that when W, L, t_{ox}, and V scale down by a factor of 2 ($\alpha = 2$), the MOSFET can run twice as fast with only one-fourth of the power consumption. If W, L, and t_{ox} scale down by a factor of 2 while the voltage V stays unchanged, the device can be four times faster with half of the power consumption.

During that era, any pattern on a photomask could be printed on the wafer surface with a high degree of fidelity, albeit with some corner rounding. Those "golden days" of scaling ended after the introduction of the 130-nm technology node, when the gate oxide thickness could be scaled down no further due to tunneling-induced leakage.

For planar MOSFET ICs, a technology node used to be defined as one-fourth of the pitch of the contacted gate (gates with contact between them), which had the smallest pitch in that chip. At 130 nm, because the pattern pitch was approaching the photolithography wavelength, the proximity effect became stronger, and what was on the mask no longer resembled what was printed on the wafer. Optical proximity correction (OPC) started being applied in that node.

The performance improvement from 90-nm nodes to 65-nm nodes primarily involved the introduction of channel strain, with limited improvement of carrier mobility μ. Doping the gate silicon dioxide with nitrogen also helped increase k value somewhat. Because the feature sizes in these technology nodes started to become significantly smaller than the photolithography wavelength, 193 nm, optical proximity correction (OPC) had to be

Table 4.1 Scaling parameters and scaling effect on performance, based on Ref G.

Parameter	Symbol	Constant Field Scaling	Constant Voltage Scaling	Constant Voltage Scaling with Velocity Saturation
Gate length	L	$1/\alpha$	$1/\alpha$	$1/\alpha$
Gate width	W	$1/\alpha$	$1/\alpha$	$1/\alpha$
Field	ϵ	1	α	α
Oxide thickness	t_{ox}	$1/\alpha$	$1/\alpha$	$1/\alpha$
Substrate doping	N_a	α^2	α^2	α^2
Gate capacitance	C_G	$1/\alpha$	$1/\alpha$	$1/\alpha$
Oxide capacitance	C_{ox}	α	α	α
Transit time	t_r	$1/\alpha^2$	$1/\alpha^2$	$1/\alpha$
Transit frequency	f_T	α	α^2	α
Voltage	V	$1/\alpha$	1	1
Current	I	$1/\alpha$	α	1
Power	P	$1/\alpha^2$	α	1
Power delay	$P\Delta t$	$1/\alpha^3$	$1/\alpha$	$1/\alpha$

applied in the mask patterns to ensure the design-intended patterns could be printed on the wafer surface. Intel was the first to introduce a high-k metal gate (HKMG) in the HMV of 45-nm chips, which helped continue performance improvement. At the same time, techniques such as double patterning and immersion lithography were developed, which helps the feature size continue scaling with the existing 193-nm ArF photolithography system without so-called "next-generation lithography (NGL)" techniques, such as EUV, e-beam direct write, nano-imprint, etc.

The 3D device era arrived when FinFET technology was introduced in 22-nm and 16-nm/14-nm technology nodes, which also kept all of the aforementioned device-performance-enhancement techniques. It continued scaling the feature sizes; however, it improved device performance mainly by increasing the channel width W, which is primarily determined by the fin height $W = 2H_{fin} + CD_{fin}$. Self-aligned double patterning (SADP) with multiple cut masks was used to pattern the fins and gates. A low-temperature ALD process could be developed to deposit conformal dielectric film directly on the PR patterns to form a spacer on the PR sidewall, which could eliminate a dummy layer for mandrel formation and thus reduce the patterning cost. Figure 4.1 shows images of the SRAM cell and MOSFET cross-section TEM images of Intel 90-nm, 65-nm, 45-nm, 32-nm, 22-nm, and 14-nm technology nodes and their release dates.[18–23]

When scaling to 10 nm and 7 nm, the scale gate width W can be scaled further by increasing the fin height. If EUV is still not ready for HVM, then SAQP with multiple cut masks will be used to form the fin patterns and the gate patterns. When the metal wire cross-section scales down to 20 nm wide by 25 nm tall, the percentage of 5-nm TaN barrier layer in the metal wire cross-section will reach 60%. In this case, the resistivity of the copper wire (which has a cross-section of 10 nm $\times$ 20 nm) will be much higher than the bulk copper resistivity due to boundary scattering. This situation occurs

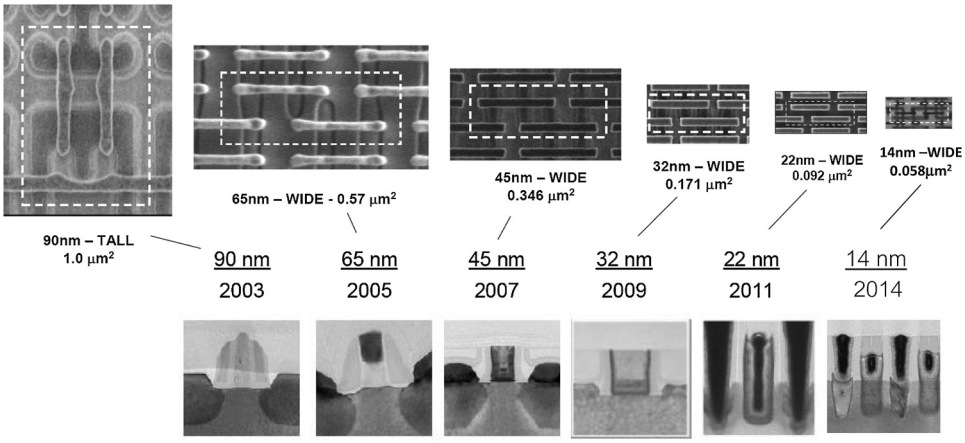


Figure 4.1 Intel IC scaling: technology nodes, release years, SRAM cell images, and device cross-section images. The upper- and lower-right images are from Ref. H, reprinted with permission from IEEE. The rest of the images are from Ref. 24, reprinted with permission from Intel.

because the electron mean free path in copper (~55 nm) is significantly larger than the copper wire width and height.[26] Therefore, a combined Cu and TaN stack is no longer the best conducting materials for a 25 nm × 25 nm^2 wire. Metals that do not require a barrier layer and that have a resistivity lower than a Cu/TaN stack could be used, especially for the first few metal layers with tight geometry. Cobalt (Co) is a possible replacement. A Co interconnect can be formed with selective grow, anneal, and CMP.[27] A shrinking feature size also means that the film thickness is shrinking, and more thin film layers are deposited by ALD processes with bulk film still using CVD, PVD, and ECP. A thin film layer (a few nanometers thick) could also be etched with an atomic layer etch (ALE) process; STT-MRAM MTJ formation is one such ALE application.

An air gap, which has the lowest dielectric constant, has already been applied in planar NAND flash[28] and a few of the upper metal layers of 14-nm logic IC. More air gap is expected in the BEoL in 10-nm and 7-nm nodes to reduce RC delay and improve device speed.

Another way to increase the channel width W is using multiple silicon nano-wires (NWs), which is called a gate-all-around (GAA) FET. In contrast, the current FinFET in HVM involves tri-date devices, in which the channel is surrounded by gate along three sides. Figure 4.2(a) shows the tri-gate FinFET with three fins, and Fig. 4.2(b) shows the GAA-FET with nine nano-wires.[29,30]

Figure 4.3 shows the simplified process steps of Si NW GAA-FET manufacturing. First, alternating SiGe and Si epitaxial layers are grown on the silicon surface of a SOI wafer, followed by silicon oxide and silicon nitride HM deposition [Fig. 4.3(a)]. After PR patterning, a fin-shaped pattern is etched, the PR is stripped, and the wafer is cleaned, as shown in Fig. 4.3(b).

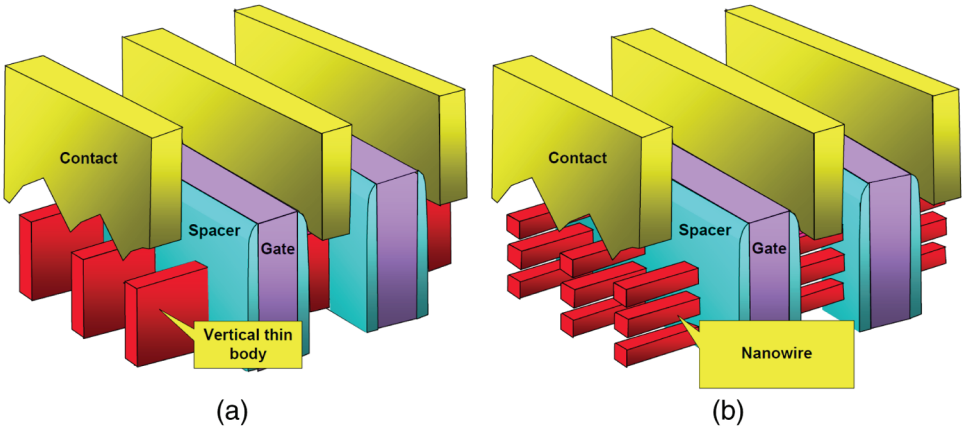


Figure 4.2 (a) Three-fin FinFET and (b) nine-wire GAA FET. Image reprinted from Ref. 29 with permission from Intel.

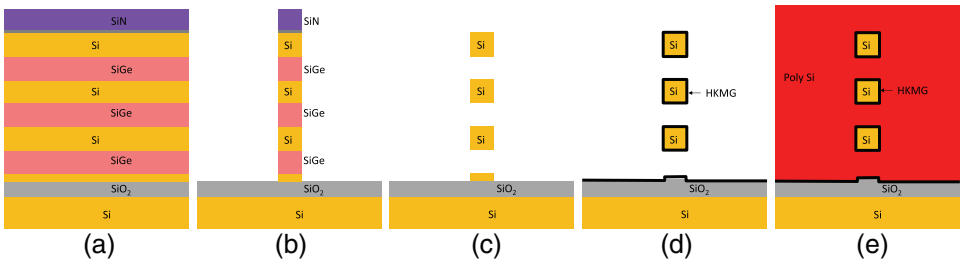


Figure 4.3 Nano-wire formation: (a) alternating SiGe and Si epitaxial growth, and pad oxide and nitride hard mask deposition; (b) photolithography, etch, PR strip, and clean; (c) remove SiGe and SiN; (d) oxidation, clean, oxidation, and HKMG deposition; and (e) polysilicon deposition.

SiGe and SiN layers are removed to form silicon nano wires, as shown in Fig. 4.3(c). After oxidation and oxide clean, which reduces the thin SOI layer to oxide, the high-k and metal gate layers are deposited, as shown in Fig. 4.3(d). Polysilicon deposition, CMP, and patterning finish the gate-first HKMG Si NW GAA-FET formation, as shown in Fig. 4.3(e). Of course, during the etch process, the silicon NW cannot be floated in mid-air, as shown in Figs. 3.16(c) and (d). The supporting structures of the NW will be patterned at the end of NW [Fig. 3.16(b)] at the locations where the contact plugs will land [Fig. 3.15(b)]. The NWs are dangling between supporting structures, and a sagging NW is a challenging defect to capture and review.

Quick drawing and calculation can help demonstrate the challenges that NW GAA-FETs will face. Figure 4.4(a) illustrates a FinFET with a 10 nm $\times$ 60 nm fin, and Fig. 4.4(b) illustrates a GAA-FET with three 10 nm $\times$ 10 nm wires with 10-nm spacing between the wires. The effective channel width of the FinFET is 130 nm, and the effective channel width of the NW GAA-FET

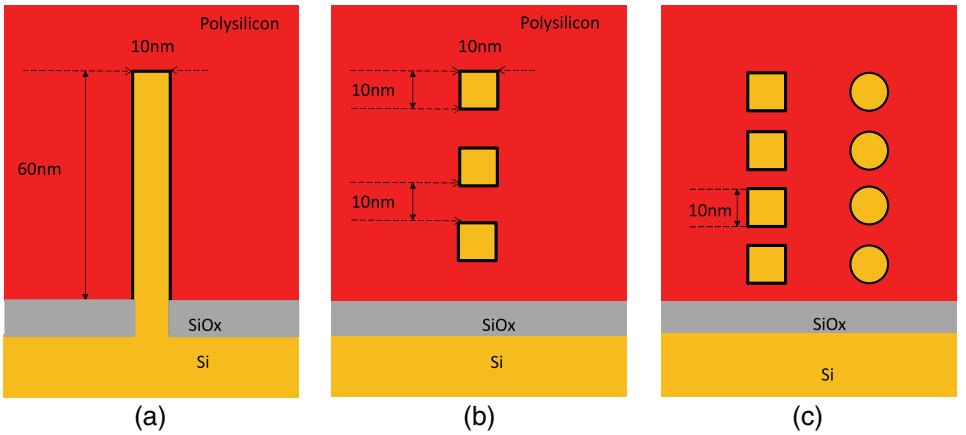


Figure 4.4 (a) 10 nm × 60 nm FinFET, (b) GAA-FET with three 10-nm Si wires, and (c) GAA-FET with four 10-nm Si wires.

is 120 nm! Thus, to get a higher drive current than a 10 nm × 60 nm FinFET, at least four 10 nm × 10 nm NWs must be stacked with 5-nm spacing between the wires, as shown in Fig. 4.4(c), which has an effective channel width of 160 nm. With corner rounding, $W = 125.7$ nm for four circular NWs with a CD of 10 nm. However, a GAA-FET [shown in Fig. 4.4(b)] has only half of the cross-section area of the fully depeleted channel at the off state of the FinFET shown in Fig. 4.4(a), which allows it to significantly reduce standby leakage while maintaining a similar drive current (an attractive trait for mobile chip designers).

Since the advent of the 90-nm technology node, channel strain has been used to improve carrier mobility μ, usually with compressive strain in the PMOS channel and tensile strain in the NMOS channel. Analysis of the drive current equation suggests that one thing that has not really changed in 50 years of IC scaling is carrier mobility μ, which is determined by the channel materials (silicon for all these years) with some minor improvement by applying strain in the channel. By changing the channel materials from silicon to high-mobility (high-μ) materials, such as Ge for the PMOS channel and III-V compound for the NMOS channel, the carrier mobility μ can be significantly increased. Table 4.2 shows the materials with different electron and hole mobilities that could be the potential channel material for next-generation IC chip scaling.

The material change of MOSFETs in IC manufacturing is very difficult and full of challenges. Many new materials had been introduced over more than 50 years of IC fabrication history, such as polysilicon in the 1970s; $TiSi_2$, WSi_2, W, and TiN in the 1980s; $CoSi_2$, Cu, and Ta in the 1990s; and NiSi in the 2000s. However, the real material change of the MOSFET device itself happened only twice. The first time was when polysilicon replaced aluminum as the gate electrode in the 1970s after the SASD formation was introduced

Table 4.2 Materials with different carrier mobility, band gap, and dielectric constant.

	Si	Ge	InP	GaAs	$In_{0.47}Ga_{0.53}As$	InSb	GaSb
Electron μ (cm^2/Vs)	1600	3900	5400	9200	14000	77000	1000
Hole μ (cm^2/Vs)	430	1900	200	400	300	850	3000
Band gap E_g (eV)	1.12	0.66	1.34	1.42	0.75	0.17	0.72
Dielectric constant k	11.8	16	12.4	13.2	13.9	16.8	15.7

using newly developed ion-implantation technology. The second time was the introduction of HKMG in the 45-nm technology node to replace nitrogen-doped silicon oxide as the gate dielectric and polysilicon gate electrode about 30 years later. For the MOS transistor, "M" and "O" were changed—perhaps it is time to change "S," the channel material.

Of course, carrier mobility is not the only thing that must be considered. For example, the band gap E_g, which determines the device's thermal stability and off-state leakage current (standby power consumption), deserves consideration. Most of the high-μ materials in Table 4.2 have a low E_g, which makes them undesirable for mobile devices. Unfortunately, mobile devices constitute the main market that drives the development of IC technology and manufacturing currently and for the foreseeable future. If Ge, SiGe, and III-V materials were used in future FinFET or NW devices, it could be the last major material change of MOSFET-based IC manufacturing. Figure 4.5 shows the process that could be used for Ge, SiGE, and III-V FinFET formation.

Serval methods can be used to form Ge, SiGe, and III-V fins. One way grows defect-free blanket Ge, SiGe, and III-V epi films on a silicon wafer and then patterns them into fins, as shown in Fig. 4.5(a). Another way forms the silicon fins first using fin-formation processes of the existing FinFET process. After ILD deposition and CMP, the silicon fins are recessed, and selective growth of a Ge, SiGe, or III-V epitaxial fin forms the silicon surface in the slot where the silicon fin used to be; the goal is to keep the lattice mismatch-induced crystalline defects, such as dislocation and stack fault, near the bottom of the slot, as shown in Fig. 4.5(b).[28]

Figure 4.6 shows another two ways to form high-μ fins proposed by the author. Basic steps shown in Figure 4.6(a) are: deposit dielectric film, then pattern the dielectric in reverse patterns of the fins; SEG high-μ fins from the silicon surface at the bottom of the trenches. After CMP removes high-μ film from the surface, recess ILD to expose the high-μ fins to the required height. These two processes basically skipped the silicon fin formation in Fig. 4.5(b), thus should have lower cost. Figure 4.6(b) modified the process steps in Fig. 4.6(a) by adding a hard mask layer on top of the ILD. After wafer clean, the ILD and HM are deposited. After fin patterning, the HM and ILD are etched, and after PR strip and clean, high-μ SEG is performed. High-μ SEG CMP stops on the HM; after HM removal, the high-μ fins are formed, with the fin height determined by the HM thickness.

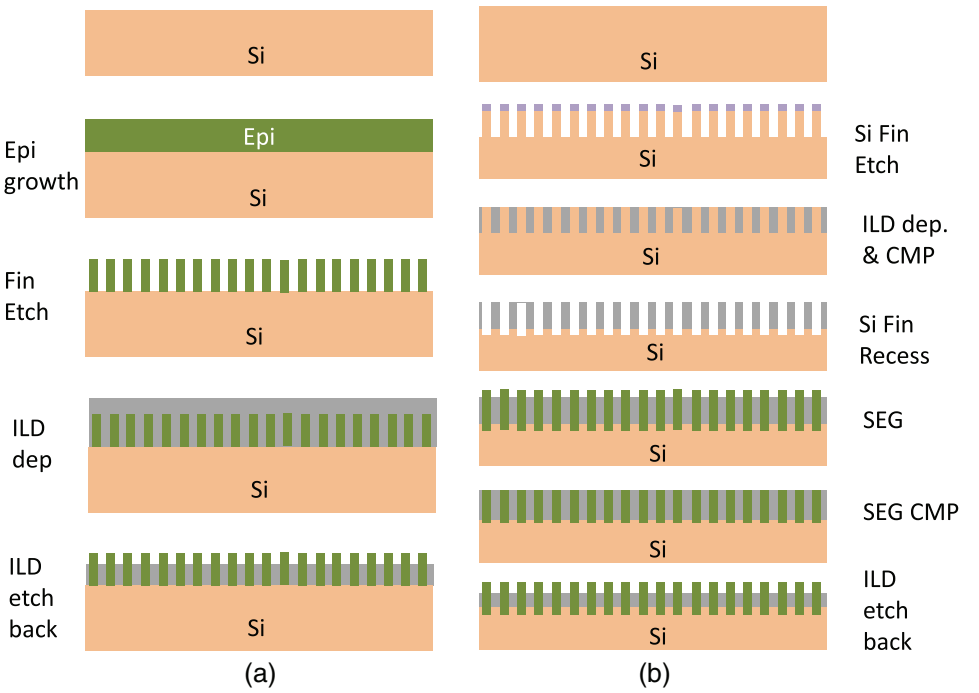


Figure 4.5 (a) Blanket high-μ epi growth and fin formation, and (b) Si fin replacement with a high-μ fin formation.

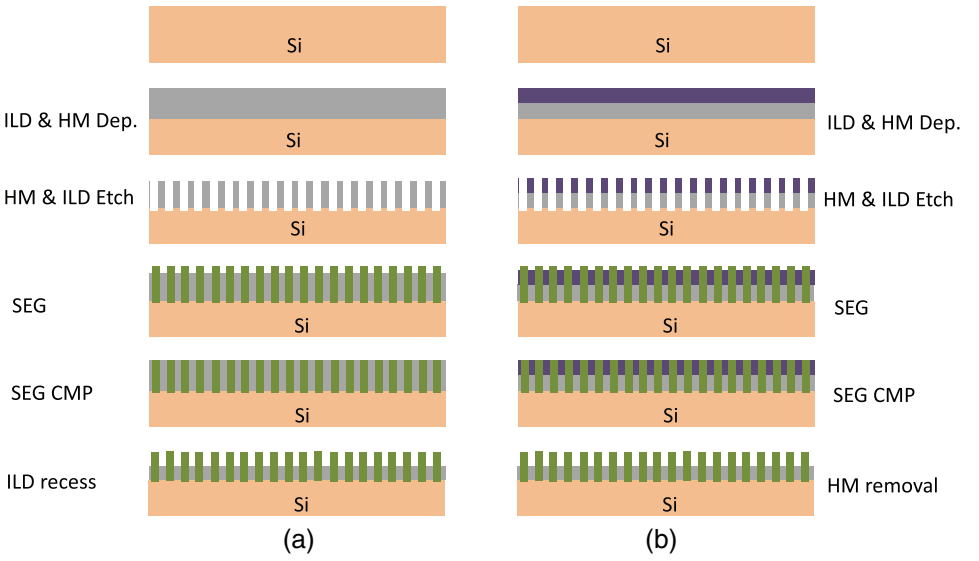


Figure 4.6 (a) Modified high-μ fin formation steps without a HM, and (b) modified high-μ fin formation with a HM.

Table 4.3 Summary of MOSFET scaling types and performance-enhancement factors.

Scaling type	Technology node	Performance improving factor in $I_D \propto \mu(k/t_{ox})(W/L)$
Traditional	≥130 nm	Reduction of the gate oxide thickness t_{ox}
Extension of traditional	90 nm and 65 nm	Channel strain-induced higher μ and nitrogen-doped gate oxide increased k
High-k metal gate	45 nm and 32 nm	Increased k and MG-induced lower EOT
3D (tri-gate FinFET)	22 nm to ≤10 nm	Increased fin height: channel width $W = 2H_{fin} + CD_{fin}$
3D (NW GAA-FET)	≤5 nm?	Increased number of nano-wires N: channel width $W = \pi NCD_{NW}$
High-μ	≤5 nm?	Increased carrier mobility μ

From Figure 4.1 we can see that technology node scaling is slowing down. It took three years from 22 nm scaled down to 14 nm, while previous technology node scaling usually took two years. Likely 14-nm to 10-nm and 10-nm to 7-nm scaling will take at least three years or even longer. This is because the smaller the technology node, the harder it is to pattern the features, and the smaller the killer defect size is that has already approached the detection limit of HVM inspection systems, and the harder it is to develop defect-free processes for high-yield HVM. Table 4.3 summarizes the MOSFET IC scaling.

4.2 Scaling and Development of Memory Devices

The cost and speed of different memory devices are indicated in Fig. 4.7. In a personal computer or a mobile computing device—such as a smartphone, tablet, or laptop—SRAM is commonly used as the cache memory of the central processing unit (CPU) to store data that need quick access. DRAM is

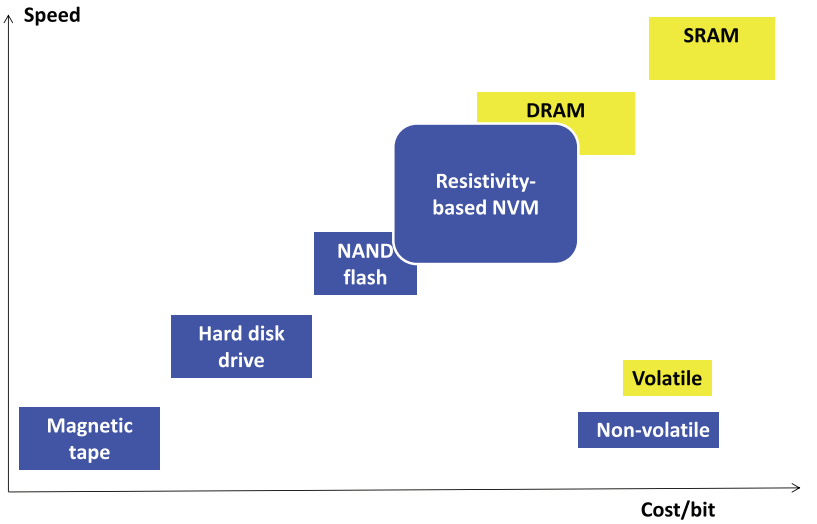


Figure 4.7 Cost and speed of different memory devices.

the main memory used to run the program and keep data for processing and editing. Both SRAM and DRAM are volatile memory, which means that they need a power supply to maintain the memory. A sudden power outage could cause data to be lost if they were not saved in a non-volatile memory, such as a NAND flash-based solid state drive (SSD) or hard disk drive (HDD). Most mobile devices use a SSD because it is smaller, consumes less power, and is more reliable because it does not have any moving parts. HDDs are widely used in PCs, and they are the backbone for data storage, especially servers and data centers, because they are cheaper than SSDs and are more cost effective for large amounts of data storage. Magnetic tape is still used for data backup in data centers due to its low ownership cost.

Buried WL architecture will continue be used for DRAM scaling; the main challenge is how to scale the storage node capacitor because its aspect ratio increases each generation of scaling. It will reach a point where it can no longer be made cost-effectively due to issues such as under-etch, bowing-profile-induced SN-to-SN short, and misalignment to the SNC plug at the bottom of the SN hole. Those issues could become a roadblock for further DRAM scaling. For NAND flash, 3D-NAND is gaining momentum, and the scaling trend involves adding more stacks. The main challenges are multilayer stack deposition, staircase formation, channel hole etch, bottom select-gate Si SEG, isolation trench etch, CG/WL deposition and isolation formation, and staircase contact etch. For multilayer deposition, not only do the deposition rate and uniformity of the multilayers need to be well-controlled but also the stress of the stacks. The final roadblock for 3D-NAND is similar to DRAM, i.e., HAR channel hole etch: under-etch and hole bowing-profile-induced device variation. Si SEG at the bottom of the deep channel hole could be even more challenging. Furthermore, contact hole etch on a 128-step (or more) staircase could become really difficult to control.

In the field of memory development, scientists and engineers have been trying to find alternatives that are as fast as DRAM and as low-cost and non-volatile as NAND flash. Both DRAM and NAND flash are charge-based memory, as well as SRAM. A capacitor-type device is needed for this kind of memory to hold the charges and keep the memory of data. It is very difficult to keep the capacitance of a planar device to hold the charge while scaling down the feature size, which is the main reason why a DRAM SN capacitor goes to such HAR 3D structures and NAND flash goes to 3D.

Almost all recently developed alternative NVMs are not charge-based devices; rather, they are based on resistivity. Although these NVMs have different working mechanisms, they all have at least two stable resistivity states: one high-resistivity state, and one low-resistivity state. By staying in different resistivity states, they can permanently store digital data of 1 or 0, and they can quickly switch between the two states.

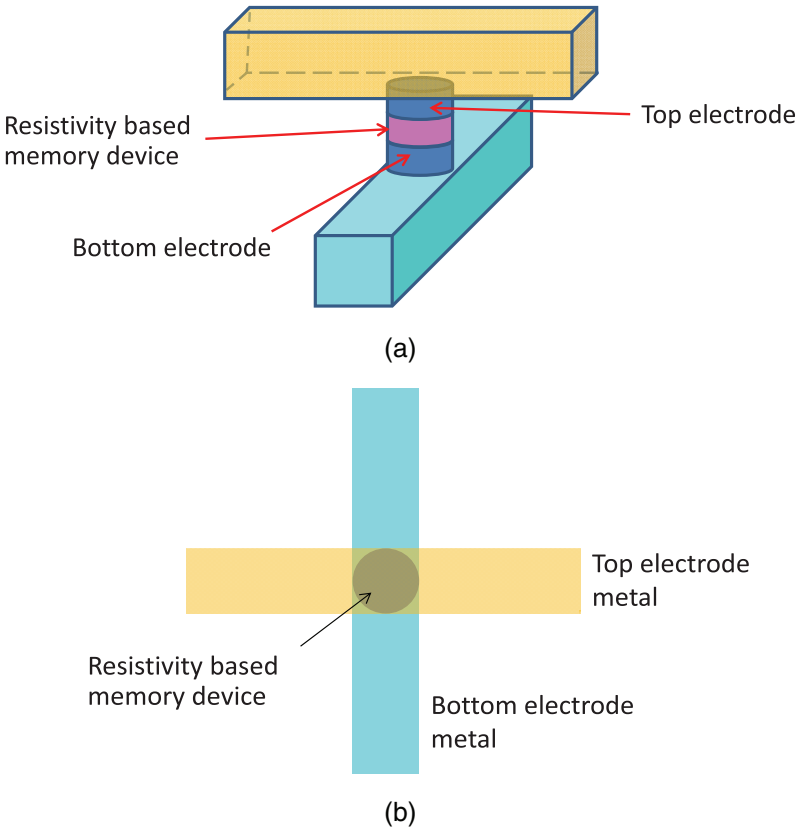


Figure 4.8 (a) Basic structure of the resistivity-based high-speed non-volatile memories and (b) its layout.

Figure 4.8(a) illustrates the base structure of the resistance-based high-speed non-volatile memories. These resistivity-based memories usually have a bottom electrode that can be connected to a select transistor. There is also a top electrode. A metal wire connects the top electrode and bottom electrode (normally perpendicular to each other), locating the memory device at the cross-point of the two metal wires in the top-down view, as shown in Fig. 4.8(b). This arrangement is why some people call this type of device "cross-point" memory.

This type of memory can be stacked, just like multi-layer interconnects in IC chips. Figure 4.9(a) shows a two-stack version, Fig. 4.9(b) shows a four-stack, and Fig. 4.9(c) illustrates an eight-stack version with a 3D structure, which can be manufactured with processes that are very similar to 3D-NAND flash. For the stacking shown in Figs. 4.9(a) and (b), each additional stack requires two masks, and thus the cost of stacking increases linearly with the number of stacks. For the 3D stacking shown in Fig. 4.9(c), the number of

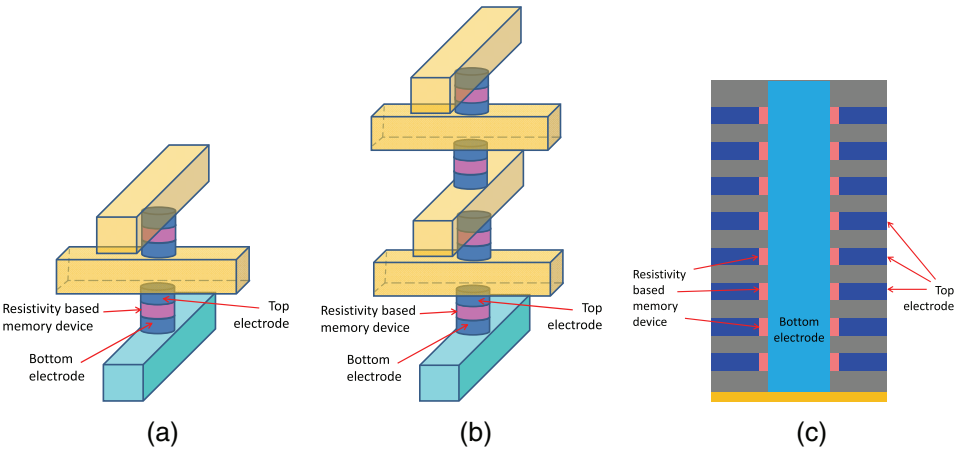


Figure 4.9 Resistivity-based memory: (a) two-stack, (b) four-stack, and (c) 3D eight-stack versions.

masks is not linearly related to the number of stacks. The main challenges are film deposition, etch, and clean, just like 3D-NAND flash.

There are currently three resistivity-based NVM devices that get a lot of attention:

- phase-change random access memory (PCRAM),
- spin-transfer-torque magnetic random access memory (STT-MRAM), and
- resistive random access memory (ReRAM).

Figure 4.10 shows a schematic of PCRAM, which changes the state of a chalcogenide-based material from amorphous to crystalline in order to store a digital signal "0" or "1." Alloys of germanium, antimony, and tellurium (GeSbTe, or GST) are commonly used chalcogenide-based materials in PCRAM development and manufacturing. GST can be switched between the

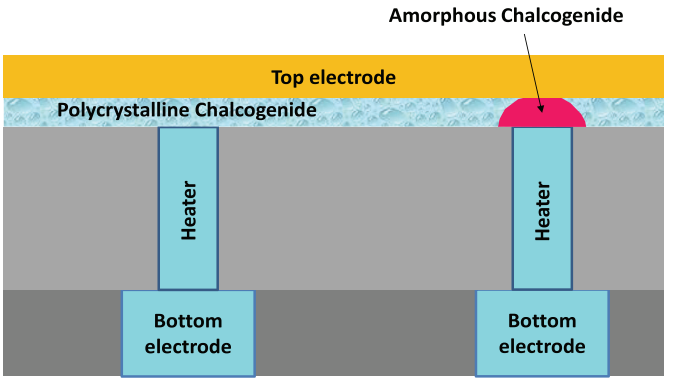


Figure 4.10 PCRAM with both memory states.

crystalline phase and the amorphous phase when GST cools down from different heating temperatures, which can be controlled by applying different current. When GST is heated to a temperature higher than its melting point (~600°C) and then cooled quickly, it forms an amorphous state. When it is heated higher than its crystalline temperature (~300°C) and then cooled, it switches to a crystalline state. The crystalline phase has low electrical resistivity, whereas the amorphous phase has high resistivity.[31,32]

The right-side chalcogenide in Fig. 4.10 is in amorphous state with high-resistance and could be registered as logic "1." The left side illustrates the crystalline state with low resistance and could be registered as logic "0." PCRAM is already being produced by several memory chip manufacturers and serves some niche markets. The commercial PCRAM chip has a unit cell with one access transistor and one GST, or 1T1R. However, it is possible to stack PCRAM in a 3D structure (Fig. 4.9) without an access transistor, which could allow people to scale PCRAM in the vertical direction to achieve HVM of low-cost, high-speed, non-volatile memory.

STT-MRAM is another resistivity-based NVM; it also uses the 1T1R structure, as shown in Fig. 4.11(a). The resistance-change device in STT-MRAM is commonly called magnetic tunnel junction (MTJ). Figures 4.11(b) and (c) show the MTJ structure. When the free-layer magnetization is parallel to that in the fix layer, as shown in Fig. 4.11(b), the resistance between the top electrode (TE) and the bottom electrode (BE) is low. When the free-layer magnetization is anti-parallel to that of the fix layer, as shown in Fig. 4.11(c), the resistance is high. Both states are stable, which gives the device the capability to permanently store digital information, and the stored information can be rewritten by switching the parallelism of the free-layer magnetization.

The fix layer has a fixed magnetization orientation and is used as the reference to identify the parallelism of the free-layer magnetization orientation. The fix layer is made with multiple thin layers of ferromagnetic

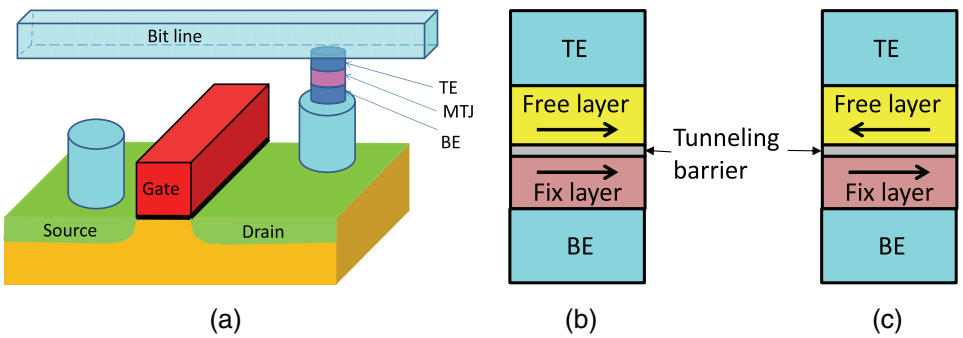


Figure 4.11 (a) 1T1MTJ STT-MRAM, (b) MTJ with parallel magnetization orientation, and (c) MTJ with anti-parallel magnetization orientation.

materials, such as multiple Co/Pd stacks, or stacks of InGa/FePt/Heusler alloy. In that multilayer, every film is very thin (~1 nm). The tunneling barrier is usually a very thin (~1–2 nm thick) dielectric layer, such as Al_2O_3 or, more recently, MgO. The free layer is also called the record layer, which consists of ferromagnetic materials so that the parallelism of its magnetization can be modified through a spin transfer operation. The free layer can be made of ferromagnetic materials, such as Co, Fe, Ni, or Mn. A commonly used material is a CoFeB alloy ~1–2 nm thick with a metal cap of Ta, Ru, and TiN. One advantage of STT-MRAM is that it can have high device density: $6F^2$, or even $4F^2$ if one can use vertical GAA-FET for the access transistor, which is very attractive to many memory IC chip manufacturers because it could potentially replace DRAM and NAND flash. It is also very attractive to logic IC manufacturers because embedded STT-MRAM is very fast and has the potential to replace SRAM and embedded DRAM. The MTJ formation is very challenging when etching a MTJ pattern with $CD < 20$ nm because it etches multiple metal layers with a very thin dielectric layer sandwiched in between and must control the MTJ profile very well. Any conductive sidewall residue across the thin barrier layer will short the free layer and the fix layer, causing yield loss.[33,34]

The third popular resistivity-based NVM is resistive random-access memory (RRAM or ReRAM). Figure 4.12 shows the ReRAM memory cell; it is formed with metal oxides or non-oxide metal compounds, such as metal sulfides and metal tellurides. An example of a ReRAM memory cell is formed by conductive tantalum oxide (TaO_x) and non-conductive tantalum pentoxide ($Ta_2O_{5-\delta}$, δ is a small amount) with a Ta filament. When the top electrode is positively biased and the bottom electrode is negatively biased, the strong electric filed causes the oxidation of Ta filament and cuts off the metal bridge. It is the high-resistance state, as shown in Fig. 4.12(a). When the bias is reversed, i.e., the top electrode bias is negative and the bottom electrode bias

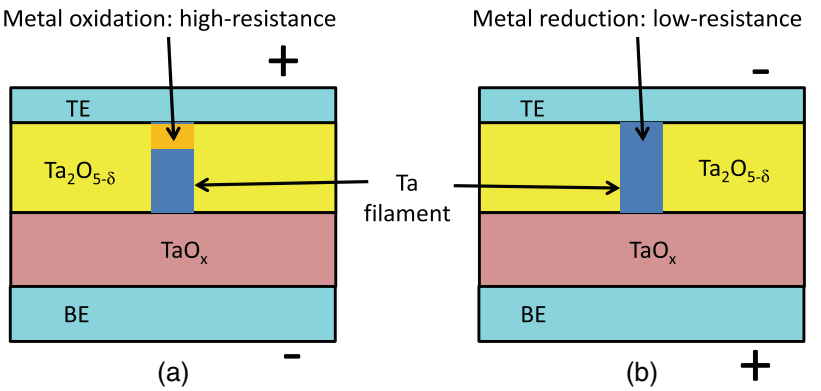


Figure 4.12 (a) High-resistance-state ReRAM device and (b) low-resistance state with a fully bridged channel.

is positive, the electric field is reversed, which causes the reversion of the aforementioned metal oxidation process; metal reduction occurs, and the greater the reduction, the stronger the electric field accelerates the process. A full metal bridge can be formed within a few nanoseconds. This is the low-resistance state of ReRAM, as shown in Fig. 4.12(b). Because a conductive bridge is formed in the low-resistance state, this type of ReRAM is also called conductive-bridge RAM (CbRAM).

A 16-Gb ReRAM with a 27-nm lithography process with a $6F^2$ pattern density has been in production since 2014. This ReRAM used the 1T1R architecture. Its access transistor is a BWL transistor, and its resistive memory cell consists of a TE, a copper telluride (CuTe) conductive layer, a Si_xO_y insulator with (0) or without (1) a fully bridged copper filament, and a BE, as shown in Fig. 4.13.[35,36]

ReRAM can be stacked in a 3D structure, as illustrated in Fig. 4.9, without the select transistor; thus, it can potentially have very high integration density in order to achieve a low-cost, high-capacity, and high-speed non-volatile memory device that can supplement or even replace DRAM and NAND flash in some mobile devices.

An interesting fact about ReRAM is that it could be the fourth basic element of a nonlinear electrical circuit, called a memristor.[37,38] The other three basic elements are well-known: resistor, capacitor, and inductor. The word "memristor" is derived from "memory" and "resistor;" its symbol is shown in Fig. 4.14(a), and its relationship with the other three elements is shown in Fig. 4.14(b).

Figure 4.14(b) shows that the relationship of resistor $R = dV/dI$ is the derivative of the famous Ohm's law. The relationship between voltage and magnetic flux $V = d\Phi/dt$ is the Faraday's law of induction. Memristor $M = d\Phi/dQ$ completes the fourth edge of the square with four corners of

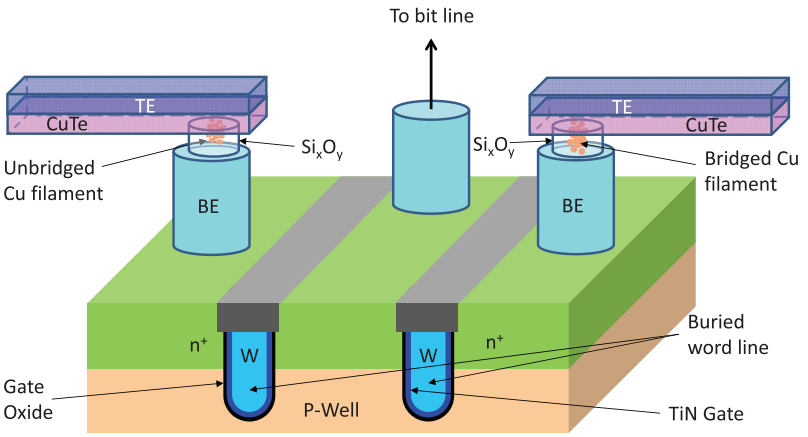


Figure 4.13 Buried-word-line ReRAM. Based on Ref. 29.

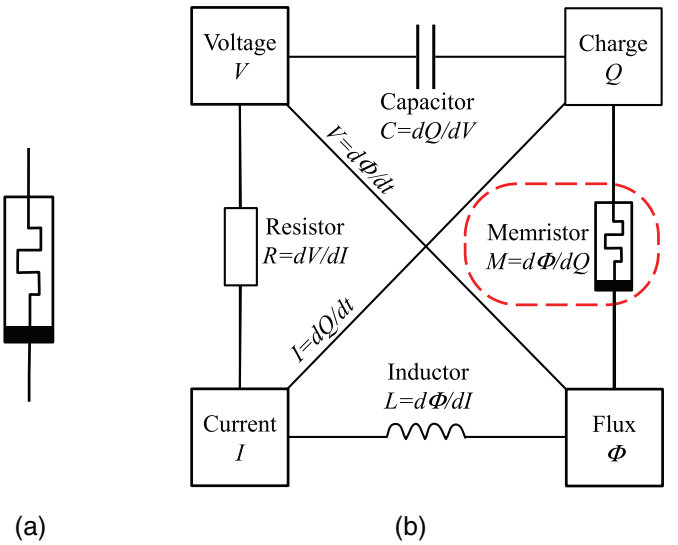


Figure 4.14 (a) Symbol for a memristor and (b) conceptual symmetries schematic of a resistor, capacitor, inductor, and memristor. Based on Ref. 39.

voltage, current, magnetic flux, and electric charge. If $d\Phi = Vdt$ and $dQ = Idt$ are substituted, then $M = V/I$, which has the unit of resistance (Ohms). It also illustrates that the memristor is not an independent element of a linear circuit; it is only for nonlinear circuits.

The term "1T1R" was used earlier in this section to describe the structure of the basic ReRAM cell, but perhaps "1T1M" would be a better description. It is very likely that in the near future, chips with all four basic elements of a nonlinear electrical circuit will be integrated in a single system-on-chip (SoC) IC.

Table 4.4 summarizes the different memory devices discussed in this book. Recently, Intel and Micron jointly announced a new type of NVM, 3D XPoint memory, that has been claimed to be 1000 times faster and more

Table 4.4 Parameters of different memory devices (based on Ref. I).

	SRAM	DRAM	NAND Flash	PCRAM	ReRAM	STT-MRAM
Non-volatile	No	No	Yes	Yes	Yes	Yes
Cell size (F^2)	50–120	6–10	5	6–12	6–10	6–20
Read time (ns)	1–100	30	50	20–50	10–50	2–20
Write/erase time (ns)	1–100	15	1.0 ms/0.1 ms	50/120	10–50	2–20
Endurance	10^{16}	10^{16}	10^5	10^8	10^8	10^{15}
Write power	Low	Low	Very high	Low	Low	Low
Other power consumption	Current leakage	Refresh current	None	None	None	None
High voltage required	No	3 V	16–20 V	1.5–3.0 V	1.5–3.0 V	<1.5 V

durable than NAND flash.[40] The table demonstrates that both PCRAM and ReRAM can fit that claim, and the latter fits better thanks to its higher write/erase speed.

4.3 3D Packaging

To increase device density without scaling feature size of the wafer processing, vertically stacked IC chips, such as DRAM or flash chips are used in large storage capacity DRAM and flash products. By stacking four DRAM chips in a package, it is equivalent to scaling feature size of a DRAM chip by half, which represents two full generations of scaling. Engineers and scientists have developed technologies to stack multiple DRAM chips in a package, where the chips on top are shifted a certain distance to expose the bond pads of the chip below so that multiple chips can be wire-bonded together. Figure 4.15(a) illustrates four-chip stacking with wire bonding, and Fig. 4.15(b) shows four four-chip stacked to form a 16-chip stacked DRAM in 3D packaging.[41]

Another chip-stacking technique, called through silicon via (TSV), recently received a lot of attention and traction. By making a connection through metal (Cu or W) plugs that directly pass through the silicon substrate, the routing distance becomes much shorter than the wire-bonding stacking, and thus it can improve the device speed and reduce power consumption. Because all of the bond pads are located near the edge of the die, wire-bonding stacking can happen only after die separation, with die-to-die stacking and wire bonding performed near the die edge. TSV stacking does not have this constraint, so it can be more flexible with the outlet locations to allow chip designers to optimize the routing. Another advantage of TSV stacking is that it can perform wafer-level stacking, whereas wire-bonding stacking can only be performed at the chip level. There are two types of TSV bonding. One is "die-to-known-good-die," which is a thinned good die with TSV bumps that bonds with a good die on the handling wafer. Another one is "wafer-to-wafer." If the wafer yield is high enough, then wafer-to-wafer stacking with

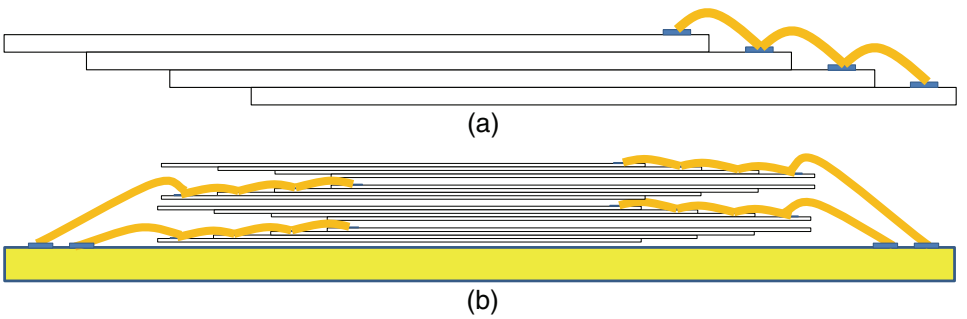


Figure 4.15 Wire-bond 3D-DRAM: (a) four-chip stacking with wire bonding, and (b) 16-chip 3D packaging with wire bonding.

STV will have a lower cost than both chip-to-chip wire-bonding stacking and die-to-known-good-die chip-to-wafer bonding.

TSV 3D chip stacking has been used in CMOS image sensors for many years.[42] DRAM chips with a four-chip TSV stack has been in HVM, and the products are available in the market.[13] A 256-Gb NAND flash chip with a 16-die stack with TSV technology has also been demonstrated recently during the Flash Summit.[43]

Question: If the DRAM wafer yield is 90%, what is the yield after four-wafer stacking?

Answer: If the yield loss is caused by random defects, then the final yield is four 90% multiplied together, which is 65.61%.

Figure 4.16 illustrates a cross-section of a TSV-ready BWL DRAM chip. The BWL DRAM illustrated in Fig. 1.10(a) is located at the upper-left corner of Fig. 4.16. TSV has many approaches; this figure shows the so-called via-middle approach, which forms the TSV after the first contact WCMP process. In this figure, it is SNC WCMP.

The TSV has a plug CD from 1–5 μm and a depth from 10–50 μm. After all of the interconnect processes have finished and the passivation dielectric is etched to expose the bond pads, the wafer is taped from the front side to

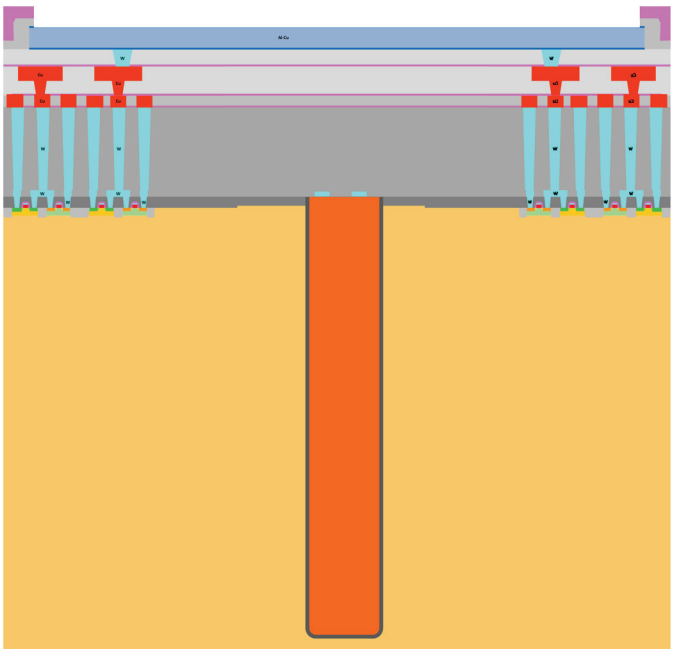


Figure 4.16 Cross-section of a TSV-ready BWL DRAM.

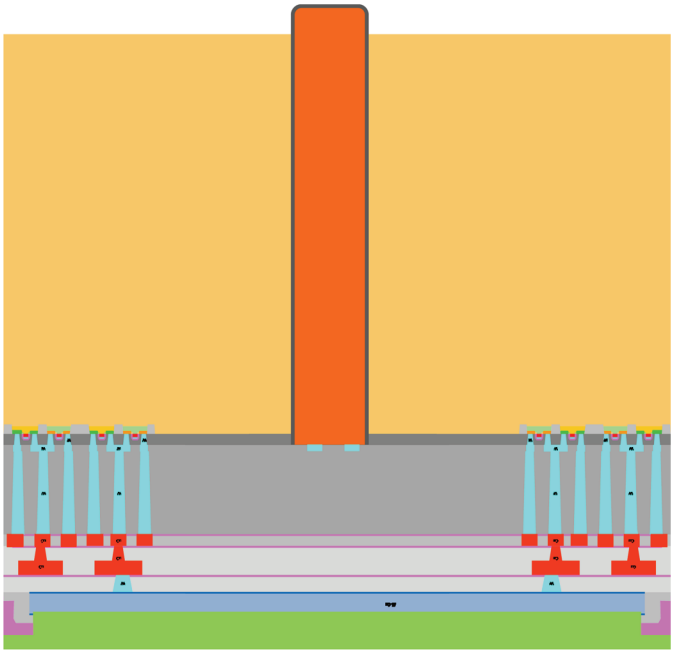


Figure 4.17 Cross-section of a TSV BWL DRAM after wafer thinning.

protect the circuitries already manufactured, and the wafer is thinned from the backside of the wafer to expose the TSV Cu plugs, as shown in Fig. 4.17. A standard 300-mm Si wafer is 775 μm thick; to expose a 50-μm TSV, at least 725 μm of Si must be removed from the backside. After dielectric-layer deposition and removal expose the TSV plugs again with dielectric on the surface, metal bumps are formed on top of the TSV plugs, as shown in Fig. 4.18.

The thinned wafer with bumps then can be bonded with a handling wafer, which is not thinned. The tape on the front surface of the thinned wafer is removed, and the wafer is cleaned, as shown in Fig. 4.19. More of the thinned wafer with bumps can be bonded with the wafer shown in Fig. 4.19 to achieve multiple-wafer stacking with TSV. Figure 4.20 illustrates four DRAM wafers stacked together with TSV.

The die layout for memory chips such as DRAM and NAND flash is the same, and so they can be designed into a TSV 3D stacking-ready format relatively easily, especially compared to CMOS logic devices. DRAM and flash products are highly cost-sensitive; thus, their die yield on the product wafer in HVM usually is very high, which is extremely important when implementing wafer-level 3D packaging with TSV. Beyond stacking memory, people are also working on CMOS chip stacking with TSV technology. Figure 4.21 shows a cross-section a SEM image of an eight-die stack CMOS

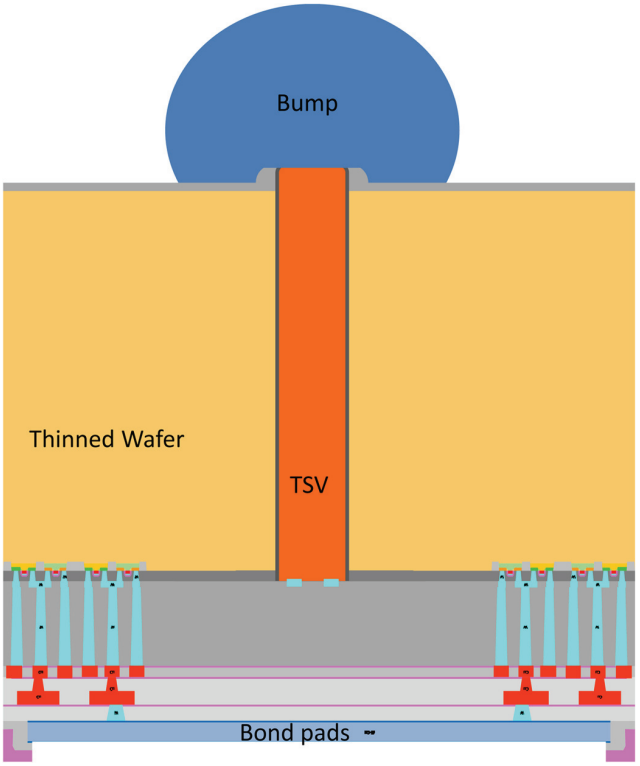


Figure 4.18 Cross-section of a TSV BWL DRAM after bump formation.

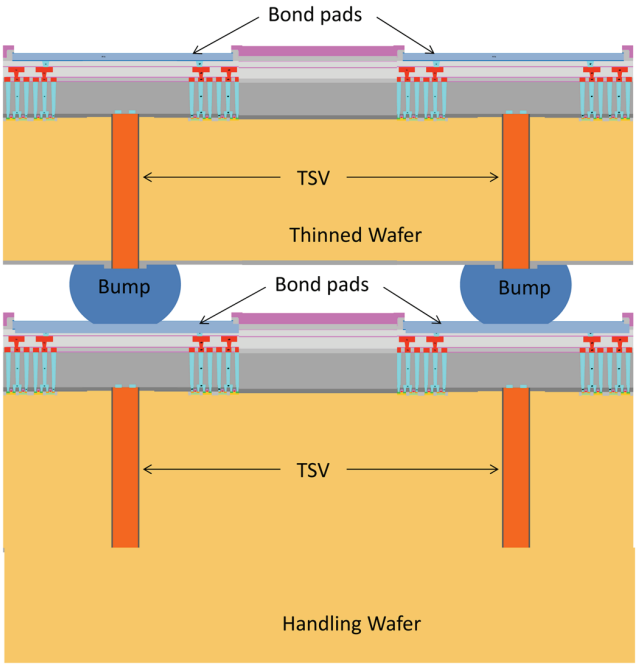


Figure 4.19 Two-wafer stacking with TSV.

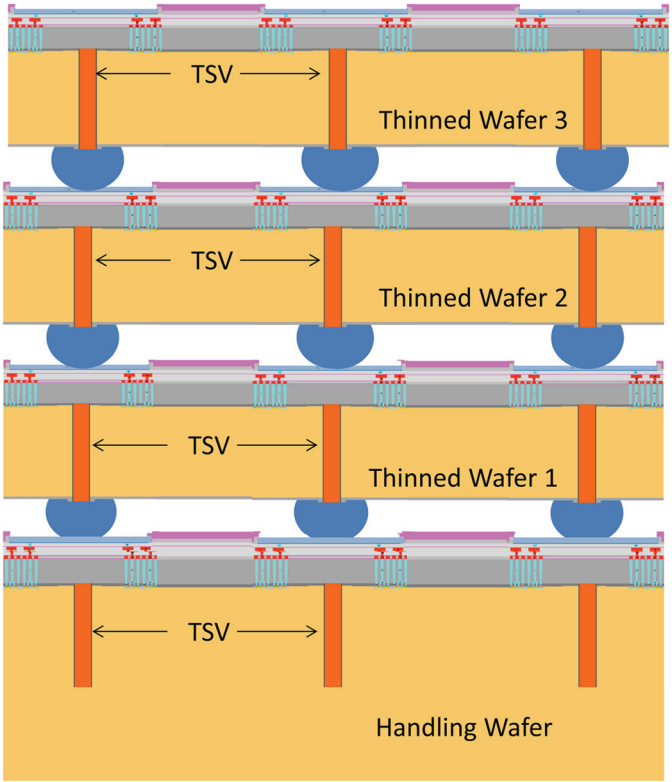


Figure 4.20 Four-wafer stacking with TSV.

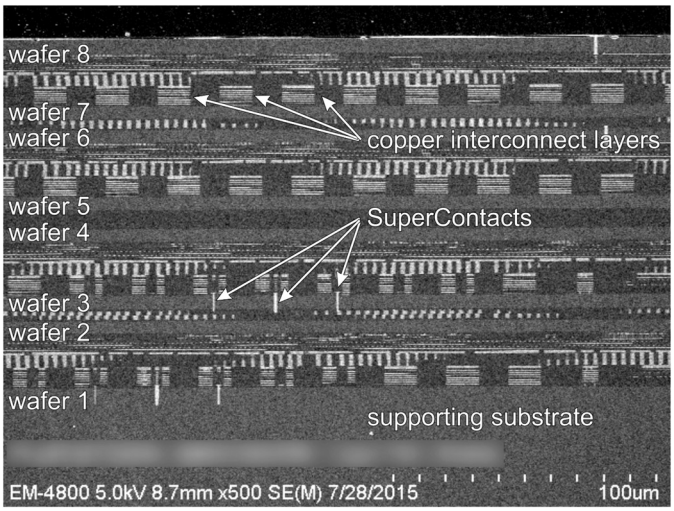


Figure 4.21 Cross-section SEM image of an eight-wafer stack that contains active CMOS and SuperContacts. From Ref. J. Reprinted with permission from Tezzaron.

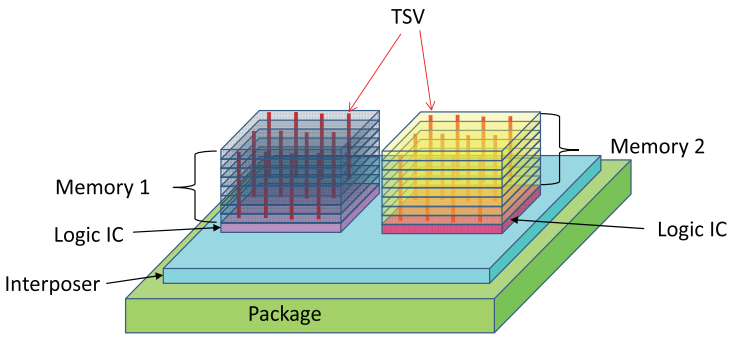


Figure 4.22 Two hybrid cubes with TSV in a package connected with an interposer.

chip.[38] (It uses the term "SuperContacts" instead of TSV.) The CD of the SuperContacts is 1.2 μm, and the depth is 6 μm. The eight dies are stacked in alternating "face-to-face" (F2F) and "back-to-back" (B2B) bonding. Wafer 1 faces up and Wafer 2 faces down for F2F bonding, Wafers 2 and 3 are B2B bonded with bumps formed at the tip of "SuperContacts" (Fig. 4.18). Wafers 3 and 4 are F2F bonded, etc. In comparison, the TSV stacking illustrated in Fig. 4.20 is back-to-face, which allows all of the stacked wafers to be exactly the same as the handling wafer.

It is also possible to stack different types of chips, such as multiple memory chips stacked together to stack on top of a logic chip. A hybrid memory cube (HMC) is one such concept that people are working on to make the chip much faster by shortening the distance between memory cells and logic circuits. It requires the memory-chip designer and logic-chip designers to collaborate and ensure that the outlets of different chips are well aligned so that they can be connected and bonded seamlessly with the TSV bumps. Multiple stacked chips can also be put into one package with an interposer to form a chip with even more functions, as shown in Fig. 4.22.

4.4 Other 3D Devices and 3D IC Processing Techniques

There are many device architectures and 3D integration techniques currently under research, and some of them are in development. A tunnel field effect transistor (TFET) is one such device being researched with the potential to replace MOSFETs in the so-called "post-CMOS" era. Figure 4.23(a) illustrates a planar TFET. It looks similar to a planar MOSFET; however, the source and drain are doped differently, and they are asymmetric. It is possible to make a vertical gate all around the TFET, even with SiGe and III-V high-μ materials, to achieve a high-speed, low-power-consumption IC circuit, such as SRAM and other logic devices, as shown in Fig. 4.23(b).[45–47]

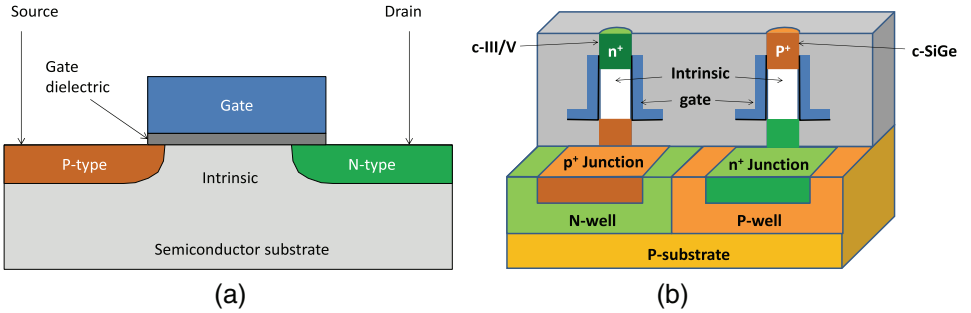


Figure 4.23 (a) Planar TFET and (b) vertical TFETs with high-μ channel materials, based on Ref. 41.

Graphene is a 2D material that has a single layer of carbon atoms arranged in a honeycomb-shape with excellent mechanical strength, high thermal conductivity, and adjustable electrical conductivity. It has attracted widespread attention from researchers of next-generation electronic devices almost immediately after its discovery.[48] Because it has zero band gap ($E_g = 0$ eV), it is unsuitable for the low-standby-power transistors that are required for the low leakage current of mobile devices. However, it sparked interest for researchers looking into other 2D materials (many of them metal chalcogenides, such as MoS_2, WS_2, etc.) as candidate materials to replace silicon for post-Si nano-electronic devices.[49] Figure 4.24 illustrates a FET with a 2D MoS_2 channel material. It is very similar to a SOI MOSFET: it uses mono-layer MoS_2 to replace Si to form the fully depleted channel, and graphene can be used to form the gate electrode and source–drain contact.

TSV 3D chip or wafer stacking is performed with fully manufactured chips or wafers, after wafer thinning and backside bump formation. A different approach of 3D IC stacking has been proposed that uses technology very similar to the smart cut of SOI wafer manufacturing, which includes a handling wafer (wafer B) with a silicon dioxide layer and a transfer wafer (wafer A) implanted with hydrogen. Bonding a hydrogen-implanted transfer wafer to

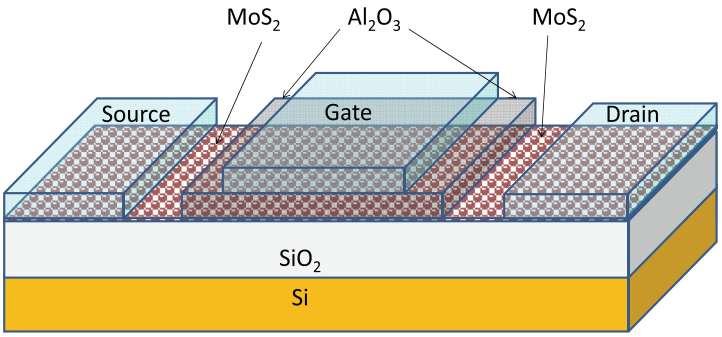


Figure 4.24 MOSFET with 2D material MoS_2.

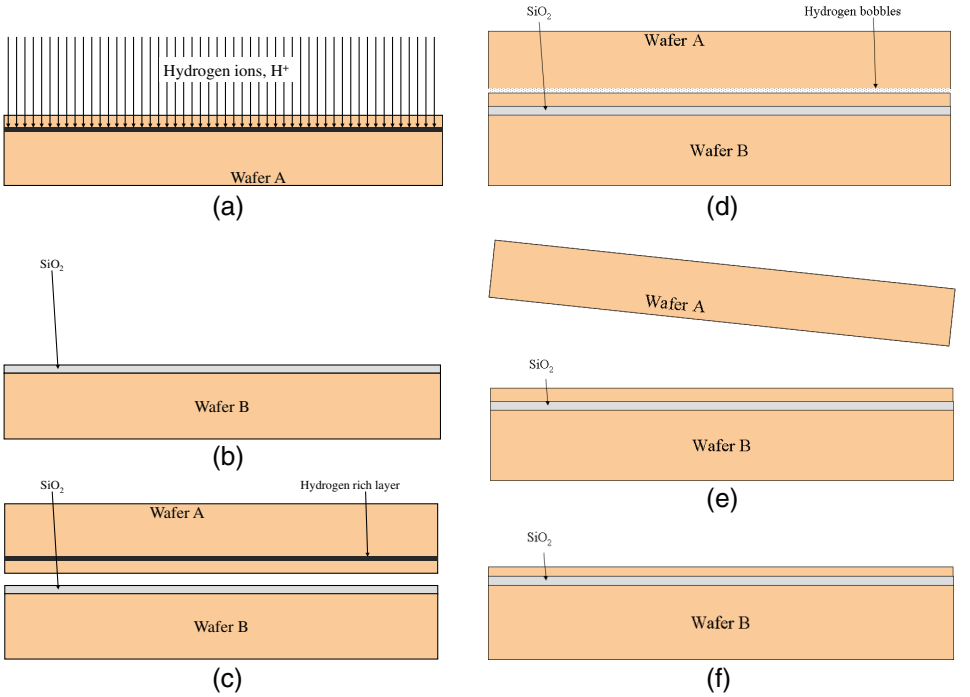


Figure 4.25 Bonded SOI formation process: (a) wafer-1 hydrogen implantation, (b) wafer-2 oxidation, (c) wafer-1 and wafer-2 F2F contact, (d) thermal bonding and hydrogen outgassing, (e) wafer splitting, and (f) Si CMP and clean.

the SiO_2 of the handling wafer can transfer a thin silicon layer to the handling wafer at the relatively low temperature of 400°C, as shown in Fig. 4.25.[50]

If a fully processed wafer with a planarized silicon oxide layer on the surface replaces the handling wafer and wafer bonding is performed with a hydrogen-implanted transfer wafer, a thin silicon layer is transferred on top of that device wafer. More IC devices can be manufactured on the thin layer of silicon, just like SOI wafer processing, and the previously processed device wafer is used as a handling wafer. A deep contact or via can be used to connect the circuits on the thin silicon layer to the circuits on the underlying carrier wafer. After low-temperature metal interconnect formation and a layer of silicon oxide is deposited and planarized with a CMP process, the wafer can be used as a handling wafer again to allow another layer of single-crystalline silicon to be transferred on top of the planarized silicon oxide. The process can be repeated multiple times.

Figure 4.26 shows a three-layer HKMG CMOS formed with two silicon-on-ILD processes. Figure 4.26(a) shows the wafer with finished HKMG CMOS devices, contacts, local interconnects, ILD (SiOx), and the first single-crystalline silicon layer already transferred on top of the ILD. Figure 4.26(b) shows the finished second layer of CMOS devices, i.e., fully depleted SOI

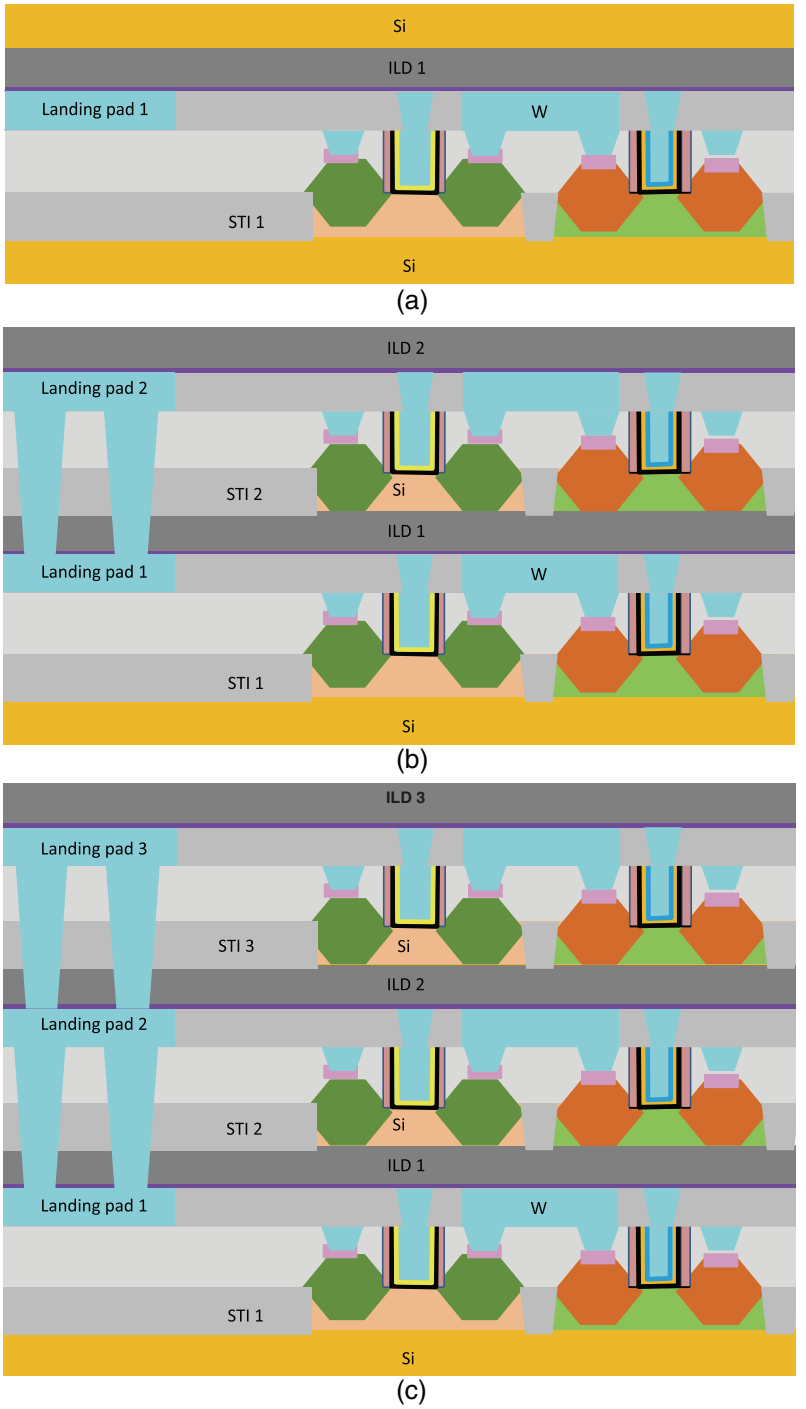


Figure 4.26 A three-layer HKMG CMOS formed with two silicon-on-ILD processes.

devices, have been finished; the contacts to the first device layer, i.e., landing pad 1; and the deposited ILD 2, ready for the transfer of the second silicon-crystalline layer. Figure 4.26(c) shows the finished third layer of CMOS devices, which are also fully depleted SOI devices; contacts to the second device layer (landing pad 2); the deposited ILD 3—the wafer is ready for the transfer of the third silicon-crystalline layer or low-*k* copper interconnect. This technique can also be applied in memory chip manufacturing and has been demonstrated in a three-layer NAND flash chip.[51,52]

4.5 The End of Moore's Law?

Since Jack Kilby created the first functional IC device in 1958, IC technology has developed rapidly, driven by consumer electronics, personal computers, the Internet, and mobile electronics. People joked that if airplane speeds followed Moore's law, they could fly around the world in less than a second today; if vehicle prices scaled like IC chips, a new car could be purchased in pennies. Computers used to weigh several tons and were owned only by government labs and large corporations, but now they are pocket-sized and have more computational power than a 20-year-old super-computer. The first generation of mobile phones needed a suitcase, and a coiled wire connected its handset to the system in the case; thus, its only practical use was to mount it in a car. When mobile phones were reduced in size to about half of a brick, people started to carry them on their belts or in their handbags, thus prompting the mobile revolution. The size of mobile phones continued to shrink until demand for larger screens increased, and the trend was reversed. The border between smartphones, tablet computers, and laptop computer has become fuzzy. It is quite possible that in 10 years one will only need a smartphone that has enough computational power, storage, apps, and easy access to networks and other accessories, such as a large screen, keyboard, mouse, and charger; all personal computers, laptops, and tablets could become obsolete. If that smartphone can project not only 3D holographic images in front of the user but also a keyboard on the desk, who needs a laptop or tablet?

Currently, self-driving cars are being test driven on the road, and they are expected to reach the market in the near future. For a person who has lost the ability to drive but still prefers to live independently, a self-driving car could be a blessing. Furthermore, self-driving vehicles can be used by the military to send supplies to frontline troops and avoid casualties due to enemy ambushes.

Robotic devices are used in households to vacuum the floor. Some hotels staff robots as front-desk clerks and luggage carriers. Likely applications of humanoid robots will be widened by intelligent ones that can not only recognize a person's face but also determine a person's moods through facial expression identification; they can not only recognize a person's voice but also

their emotion. Such a robot can perform household chores and provide care for the sick or elderly. The IC chips necessary to realize these functions must have very powerful data-processing capability with a large amount of data storage and a short data-access time, which could drive the next round of scaling.

According to Moore's law of IC scaling so far, with the IC-chip feature size scaling by a factor of $1/\sqrt{2}$ every two years, a 2-nm technology node could be reached by 2026. Even if it is assumed that after 14 nm the scaling slows down to a factor of $1/\sqrt{2}$ for every three years, the result will be a 3.5-nm technology node. For FinFET logic IC, the minimum pattern fin pitch of 3.5 nm could be 10.5 nm, which can be formed with a 13.5-nm EUV double-patterning process. The main issue will be cost. There was some argument that Moore's law ended at 28 nm, mainly because some data showed that the per transistor cost bottomed at 28 nm [Fig. 4.27(a)]. Some companies accepted this idea and left IC manufacturing. However, not everyone believes that Moore's law ended at the 28-nm node; some people believe that the law still holds, supported by the data shown in Fig. 4.27(b). Some companies are still heavily investing in development of sub-10-nm technology for future IC manufacturing, which will keep scaling IC chips for the foreseeable future. Of course, another argument is that there was never such a thing as "Moore's law;" it was just an observation and prediction. Rather, it has always been the "law of more" that drives the IC technology and overlaps with Moore's prediction.[11]

No one knows exactly what the future of IC chips looks like, although there will be chips made with 3D devices and perhaps 3D chip stacking using TSV and an interposer. One of my favorite futuristic 3D IC devices looks like

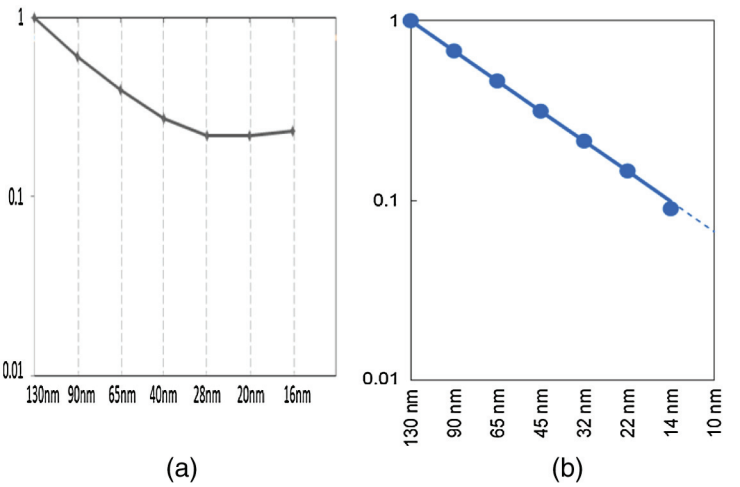


Figure 4.27 Normalized per transistor cost: (a) chart based on data from Ref. 53, and (b) chart based on Ref 24.

the one shown in Fig. 4.22: a package with multiple chip stacks connected with an interposer. Each chip stack has a logic base chip and multiple TSV stacked memory chips on top of the base. Each logic base chip has a vertical GAA-FET on the substrate level and multiple layers of silicon-on-ILD and 2D-on-ILD logic devices, depending on their functionality. Here 2D means 2D semiconductor materials such as MoS_2. Graphene can be used as gate electrodes and contact materials for the 2D devices, and carbon nano-tubes can be used for the inter-device layer connection. Memory 1 can be multiple 3D-NAND or multilayer 3D cross-point chips stacked with TSV, and memory 2 can be multiple $4F^2$ DRAM chips stacked with TSV. Hopefully, in the future, people can develop the technologies needed to manufacture this kind of chip cost-effectively.

In order to achieve this kind of within-chip and within-package 3D scaling, scientists and engineers of different disciplines must work together to research and develop novel materials, new process technologies, and innovative inspection and metrology solutions. Even after the end of Moore's law, demand for advanced IC chips will still be strong and growing, especially with the economic growth of China, India, and other developing countries. The semiconductor industry and supporting industries still need many highly trained, innovative, and hardworking scientists, engineers, technicians and supervisors to operate multi-billion-dollar IC manufacturing fabs, troubleshoot the manufacturing process, and maintain multi-million-dollar equipment.

References

1. D. Hisamoto, W.-C. Lee, J. Kedzierski, E. Anderson, H. Takeuchi, K. Asano, T.-J. King, J. Bokor, and C. Hu, "A folded-channel MOSFET for deep-sub-tenth micron era," *IEDM Tech. Digest*, 1032–1034 (1998).
2. E. Karl et al., "A 4.6 Ghz, 162 Mb SRAM Design in 22 nm Tri-Gate CMOS Technology with Integrated Active Vmin-Enhancing Assist Circuitry," *ISSCC Dig. Tech. Papers*, 230–232 (2012).
3. Samsung website, "Samsung Starts Mass Producing Industry's First 32-Layer 3D V-NAND Flash Memory, Its 2nd Generation V-NAND Offering," http://www.samsung.com/semiconductor/insights/news/13481 (May 29, 2014). Last accessed on 02/06/2016.
4. Toshiba website, "Toshiba Develops New NAND Flash Technology," http://www.toshiba.co.jp/about/press/2007_06/pr1201.htm (June 12, 2007). Last accessed on 02/06/2016.
5. Y. Fukuzumi et al., "Optimal Integration and Characteristics of Vertical Array Devices for Ultra-High Density, Bit-Cost Scalable Flash Memory," *IEDM Tech. Digest*, 449–452 (2007).
6. R. H. Dennard, "Field-effect transistor memory," U.S. Patent #3387286 (1968).
7. J. Y. Kim et al., "Transistor(RCAT) for 88-nm feature size and beyond," *Symp. on VLSI Tech.*, 11–12 (2003).
8. I.-G. Kim et al., "Overcoming DRAM Scaling Limitations by Employing Straight Recessed Channel Array Transistors with < 100 > Uni-Axial and {100} Uni-Plane Channels," *IEDM Tech. Digest*, 319–322 (2005).
9. T. Schloesser et al., "A $6F^2$ Buried Wordline DRAM Cell for 40 nm and Beyond," *IEDM Tech. Digest*, 809–812 (2008).
10. H. Xiao, "Method for forming memory cell transistor," U.S. patent #8778763 B2 (2014).
11. H. Xiao, *Introduction of Semiconductor Manufacturing Technology*, 2nd ed., SPIE Press, Bellingham, WA (2012) [doi: 10.1117/3.924283].
12. S. W. Lee et al., "Atomic Layer Deposition of SrTiO3 Thin Films with Highly Enhanced Growth Rate for Ultrahigh Density Capacitors," *Chem. Mater.* **23**(8), 2227–2236 (2011).

13. R. Courtland, "Chipmakers Push Memory into the Third Dimension, Samsung, Micron, and SK Hynix bet that transistor redesigns and chip stacking will make memory smaller and faster," http://spectrum.ieee.org/semiconductors/design/chipmakers-push-memory-into-the-third-dimension (Dec 23, 2013). Last accessed on 02/06/2016.
14. Solid State Technology, "Horizontal channels key to ultra-small 3D NAND," http://electroiq.com/blog/2012/09/horizontal-channels-key-to-ultra-small-3d-nand. Last accessed on 02/06/2016.
15. D. Fried, "FinFET tipsheet for IEDM," http://www.techdesignforums.com/practice/technique/finfet-iedm-tipsheet (December 4, 2012).
16. D. James, "Intel's 22-nm Trigate Transistors Exposed," http://www.eet.bme.hu/~mizsei/Montech/intel-s-22-nm-trigate-transistors-exposed.html (2012). Last accessed on 02/06/2016.
17. D. James, "Intel's 14-nm Parts are Finally Here!," http://www.chipworks.com/about-chipworks/overview/blog/intel%E2%80%99s-14-nm-parts-are-finally-here (2014). Last accessed on 02/06/2016.
18. T. Ghani et al., "A 90 nm High Volume Manufacturing Logic Technology Featuring Novel 45 nm Gate Length Strained Silicon CMOS Transistors," *IEDM Tech. Digest*, 197–200 (2003).
19. P. Bai et al., "A 65nm Logic Technology Featuring 35nm Gate Lengths, Enhanced Channel Strain, 8 Cu Interconnect Layers, Low-k ILD and 0.57 μm^2 SRAM Cell," *IEDM Tech. Digest*, 197–200 (2004).
20. K. Mistry et al., "A 45 nm Logic Technology with High-k + Metal Gate Transistors, Strained Silicon, 9 Cu Interconnect Layers, 193 nm Dry Patterning, and 100% Pb-free Packaging," *IEDM Tech. Digest*, 247–250 (2007).
21. S. Natarajan et al., "A 32nm Logic Technology Featuring 2nd-Generation High-k + Metal-Gate Transistors, Enhanced Channel Strain and 0.171 μm^2 SRAM Cell Size in a 291 Mb Array," *IEDM Tech. Digest*, 941–943 (2008).
22. C. Auth et al., "A 22nm high performance and low-power CMOS technology featuring fully-depleted tri-gate transistors, self-aligned contacts and high density MIM capacitors," *VLSI Symp. Tech.*, 131–132 (2012).
23. C.-H. Jan et al., "A 22nm SoC Platform Technology Featuring 3-D Tri-Gate and High-k/Metal Gate, Optimized for Ultra Low Power, High Performance and High Density SoC Applications," *IEDM Tech. Digest*, 44–47 (2012).
24. M. Bohr, "14 nm Process Technology: Opening New Horizons," http://www.intel.com/content/dam/www/public/us/en/documents/pdf/foundry/mark-bohr-2014-idf-presentation.pdf (2014). Last accessed on 02/06/2016.
25. D. Nikonov, "Beyond CMOS computing," https://nanohub.org/resources/18348/download/NikonovBeyondCMOS_1_scaling.pdf (2013). Last accessed on 02/06/2016.

26. Y. Hamaoka, K. Hinode, K. Takeda, and D. Kodama, "Increase in Electrical Resistivity of Copper and Aluminum Fine Lines," *Mater. Trans.* **43**(7), 1621–1623 (2002).
27. M. H. van der Veen et al., "Cobalt Bottom-Up Contact and Via Prefill enabling Advanced Logic and DRAM Technologies," *Proc. IEEE IITC/MAM*, 25–28 (2015).
28. K. Prall and K. Parat, "25 nm 64Gb MLC NAND Technology and Scaling Challenges," *IEDM Tech. Digest*, 102–105 (2010).
29. K. J. Kuhn, "Technology Options for 22nm and Beyond," http://download.intel.com/pressroom/pdf/kkuhn/Kuhn_International_Workshop_on_junction_technology_keynote_2010_5-1-10_slides.pdf (2010). Last accessed on 02/06/2016.
30. T. Ernst et al., "Novel Si-based nanowire devices: Will they serve ultimate MOSFETs scaling or ultimate hybrid integration?," *IEDM Tech. Digest*, 745–748 (2008).
31. S. Lai, "Current status of the phase change memory and its future," *IEDM Tech. Digest*, 10.1.1–10.1.4 (2003).
32. X. Dong, N. Muralimanohar, N. Jouppi, R. Kaufmann, and Y. Xie, "Leveraging 3D PCRAM Technologies to Reduce Checkpoint Overhead for Future Exascale Systems," http://research.hp.com/people/naveen_muralimanohar/sc09.pdf (November 2009). Last accessed on 02/06/2016.
33. W. Boullart et al., "STT MRAM patterning challenges," *Proc. SPIE* **8685**, 86850F (2013) [doi: 10.1117/12.2013602].
34. W. Kim et al., "Extended scalability of perpendicular STT-MRAM towards sub-20nm MTJ node," *IEDM Tech. Digest*, 531–534 (2011).
35. J. Zahurak et al., "Process Integration of a 27nm, 16Gb Cu ReRAM," *IEDM Tech. Digest*, 140–143 (2014).
36. D. Jana, S. Roy, R. Panja, M. Dutta, S. Z. Rahaman, R. Mahapatra, and S. Maikap, "Conductive-bridging random access memory: challenges and opportunity for 3D architecture," *Nanoscale Res. Lett.* **10**(188), doi: 10.1186/s11671-015-0880-9(2015).
37. L. O. Chua, "Memristor - The Missing Circuit Element," *IEEE Trans. Circuit Theory* **18**(5), 507–519 (1971).
38. D. B. Strukov, G. S. Snider, D. R. Stewart, and S. R. Williams, "The missing memristor found," *Nature* **453**(7191), 80–83 (2008).
39. Wikipedia, "Memristor," https://en.wikipedia.org/wiki/Memristor (2015). Last accessed on 02/06/2016.
40. L. Kelion, "3D X-point memory: Faster-than-flash storage unveiled," http://www.bbc.com/news/technology-33675734 (July 28, 2015). Last accessed on 02/06/2016.
41. D. James, "3D ICs in the real world," *Proc. ASMC*, 113–119 (2014).
42. D. Henry et al., "Through silicon vias technology for CMOS image sensors packaging," *Proc. ECTC*, 556–562 (2008).

43. Business Wire, "Toshiba Develops World's First 16-die Stacked NAND Flash Memory with TSV Technology," http://www.businesswire.com/news/home/20150805006880/en/Toshiba-Develops-Worlds-16-die-Stacked-NAND-Flash#. VfPJsxFVhBc (2015). Last accessed on 02/06/2016.
44. Business Wire, "Tezzaron Announces World's First Eight-Layer Active Wafer Stack," http://www.businesswire.com/news/home/20150830005041/en/Tezzaron-Announces-World%E2%80%99s-Eight-Layer-Active-Wafer-Stack#. VfPIiBFVhBc (2015). Last accessed on 02/06/2016.
45. X. Yang and K. Mohanram, "Robust 6T Si tunneling transistor SRAM design," *Proc. DATE*, 1–6 (2011).
46. G. Zhou et al., "Vertical InGaAs/InP Tunnel FETs With Tunneling Normal to the Gate," *IEEE Electron Device Lett.* **32**(11), 1516–1518 (2011).
47. A. Thean, "Challenges & Enablers of Logic CMOS Scaling In The Next 10 Years," *IMEC Technology Forum Taiwan*, http://www2.imec.be/content/user/File/ITF2013%20Taiwan/Aaron%20Thean.pdf (2013). Last accessed on 02/06/2016.
48. A. K. Geim and K. S. Novoselov, "The Rise of Graphene," *Nature Mater.* **6**, 183–191 (2007).
49. H. Wang, L. Yu, Y.-H. Lee, W. Fang, A. Hsu, P. Herring, M. Chin, M. Dubey, L.-J. Li, J. Kong, and T. Palacios, "Large-scale 2D Electronics based on Single-layer MoS2 Grown by Chemical Vapor Deposition," *IEDM Tech. Digest*, 88–91 (2012).
50. H. Xiao, "Wafer Manufacturing, Epitaxy, and Substrate Engineering," Chapter 4 in *Introduction to Semiconductor Manufacturing Technology*, 2nd ed., SPIE Press, Bellingham, WA (2012) [doi: 10.1117/3.924283.ch4].
51. M. Sadaka and L. Di Cioccio, "Building blocks for wafer-level 3D integration," *Solid State Technol.* **52**(10), 20 (2009).
52. S.-M. Jung et al., "Three Dimensionally Stacked NAND Flash Memory Technology Using Stacking Single Crystal Si Layers on ILD and TANOS Structure for Beyond 30nm Node," *IEDM Tech. Digest*, 37–40 (2006).
53. B. Bailey, "Moore's Law Tail No Longer Wagging The Dog," http://semiengineering.com/the-tail-of-moores-law-no-longer-wagging-the-dog (June 26, 2014).

A. P. K. Chu, "Advances in Solid State Circuit Technologies," http://www.intechopen.com/books/advances-in-solid-state-circuit-technologies
B. http://blog-imgs-36-origin.fc2.com/i/s/a/isanghan/DRAM.jpg, last accessed on 02/06/2016.
C. http://electroiq.com/chipworks_real_chips_blog/2011/01/31/samsungs-3x-ddr3-sdram-4f2-or-6f2-you-be-the-judge/, last accessed on 02/06/2016.
D. D. Fried and S. Wen, private communications, February 2016.

E. M. Ishiduki, et al., "Optimal device structure for Pipe-shaped BiCS Flash memory for ultra high density storage device with excellent performance and reliability," *IEDM Tech. Digest*, 625–628 (2009).
F. C.-H. Hung, et al., "Design innovations to optimize the 3D stackable vertical gate (VG) NAND flash," *IEDM Tech. Digest*, 227–230 (2012).
G. B. Van Zeghbroeck, "Principles of Electronic Devices," http://ecee.colorado.edu/~bart/book/book/chapter7/ch7_7.htm, last accessed on 02/06/2016.
H. S. Natarajan, et al., "A 14nm logic technology featuring 2nd-generation FinFET, air-gapped interconnects, self-aligned double patterning and a 0.0588 μm^2 SRAM cell size," *IEDM Tech. Digest*, 71–73 (2014).
I. R. Bonnecaze, S. V. Sreenivasan, and J. Ekerdt, "Nanomanufacturing Presentation for UT in Silicon Valley," 2013, http://www.slideshare.net/cockrellschool/nanomanufacturing-presentation-for-ut-in-silicon-valley-2013, last accessed on 02/06/2016.
J. P. Garrou, "IFTLE 253 China Inc Seeks to Acquire GF; Tessera Acquires Ziptronix; Tezzaron 8 layer 3DIC," http://electroiq.com/insights-from-leading-edge/2015/09/iftle-253-china-inc-seeks-to-acquire-gf-tessera-acquires-ziptronix-tezzaron-8-layer-3dic/, last accessed on 02/05/2016.

Index

Dr. Hong Xiao Xiao is an e-beam technologist at KLA-Tencor Corp., an expert of IC chip process technologies, and one of the top experts of applications of scanning electron microscopes in IC chip manufacturing processes. Previously, he was a technical marketing specialist at Hermes-Microvision, Inc. and a technical manager of Hermes Epitek Corp. He was also a consultant of semiconductor process technology, senior process engineer for the Motorola Semiconductor Production Sector, and an associate professor of Austin Community College in their Semiconductor Manufacturing Technology program.

After receiving his Ph.D. in physics from the University of Texas at Austin, Dr. Xiao worked at Applied Materials as a senior technical instructor with expertise in dielectric thin film deposition, semiconductor process integration, and plasma physics. Dr. Xiao has authored and coauthored over 30 journal and conference papers. He has 21 US patents and about 10 patents in the application process. He is the author of *Introduction to Semiconductor Manufacturing Technology*, Second Edition (SPIE Press, 2012). He has been a member of SPIE since 2005, and he is also a member of IEEE.