

interconnection processes commonly used in multilevel IC chips. The figure describes the processes between metal 1 and metal 2; the steps for metal 3, metal 4, and more are almost identical to these processing steps. Of course, the feature size or pattern CD increases with the number of metal layers.

13.7.4 Copper interconnections

Copper has lower resistivity and higher EMR than aluminum alloy. Copper is very hard to dry etch due to the lack of simple gaseous copper chemical compounds. This disadvantage has delayed its application in IC interconnections, since traditional interconnection processes require a metal etching process.

In the 1990s, CMP technology developed and matured quickly. It is still widely used in bulk tungsten removal processes to form tungsten plugs. The copper interconnection process is very similar to a tungsten plug formation process. Instead of via holes, trenches are etched on the dielectric surface. Then copper is deposited into the trenches, and a subsequent copper CMP process removes the bulk copper layer on the wafer surface, leaving the copper lines embedded in the dielectric layers. By using this damascene process, a metal etch process becomes unnecessary. The dual damascene process, which combines via and trench etching before metal deposition, is the most commonly used method for copper metallization. The dual damascene process can eliminate some metal deposition and CMP processes, unlike two single damascene processes.

The fundamental difference between Al-Cu and copper interconnection processes is that the Al-Cu process requires one dielectric etch and one metal etch, while the dual damascene copper process requires two dielectric etches and no metal etch. The challenges in Al-Cu interconnection processes are void-free dielectric CVD, dielectric planarization, via hole etch, and metal etch. The main challenges of dual damascene copper processes are dielectric etch, metal deposition, and metal CMP. Figure 13.29 lists and illustrates the single damascene copper interconnection processing steps of metal 1 for CMOS IC, and Fig. 13.30 illustrates a via-first dual damascene copper low- κ interconnection process.

A thin nitride cap layer (100 to 500 Å, depending on the technology node) can be deposited by a PECVD process with silane (SiH_4), ammonia (NH_3), and nitrogen (N_2). This cap layer is important for preventing copper from diffusing into the oxide layer and eventually diffusing into the silicon substrate, causing unstable performance in the transistor. A cap nitride layer can also prevent the oxidation of copper during oxide deposition. Unlike aluminum oxide, copper oxide is a loose layer that cannot prevent the further oxidation of copper. PECVD silicon-rich nitride or oxynitride (SiON) can be used for this layer, which can also serve as an ARC. ILD0 can be USG, fluorinated silicate glass FSG, or a low- κ dielectric such as SiCOH , and ULK such as porous SiCOH , depending on the technology node. Nitride can also serve as an etch-stop nitride layer for metal trench etching; this layer can provide nitrogen light emission for indication of the etch endpoint.

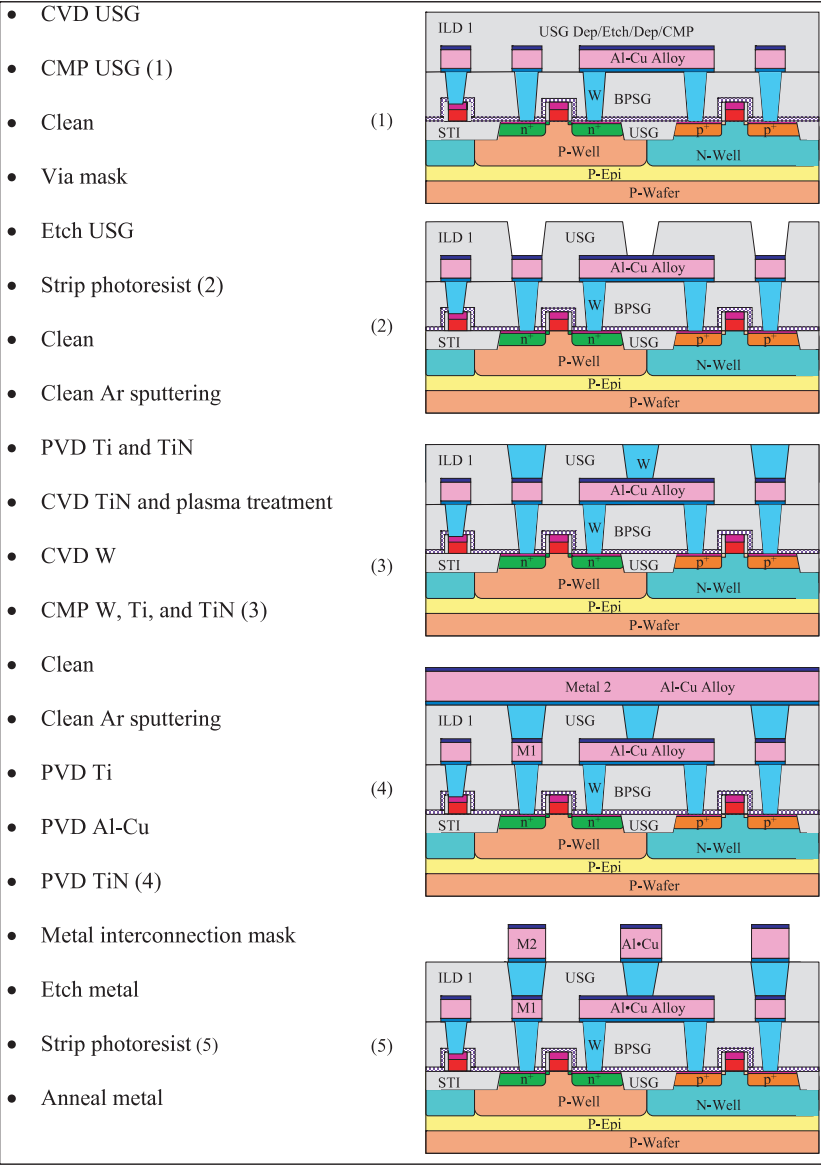


Figure 13.28 Aluminum alloy interconnection processing steps.

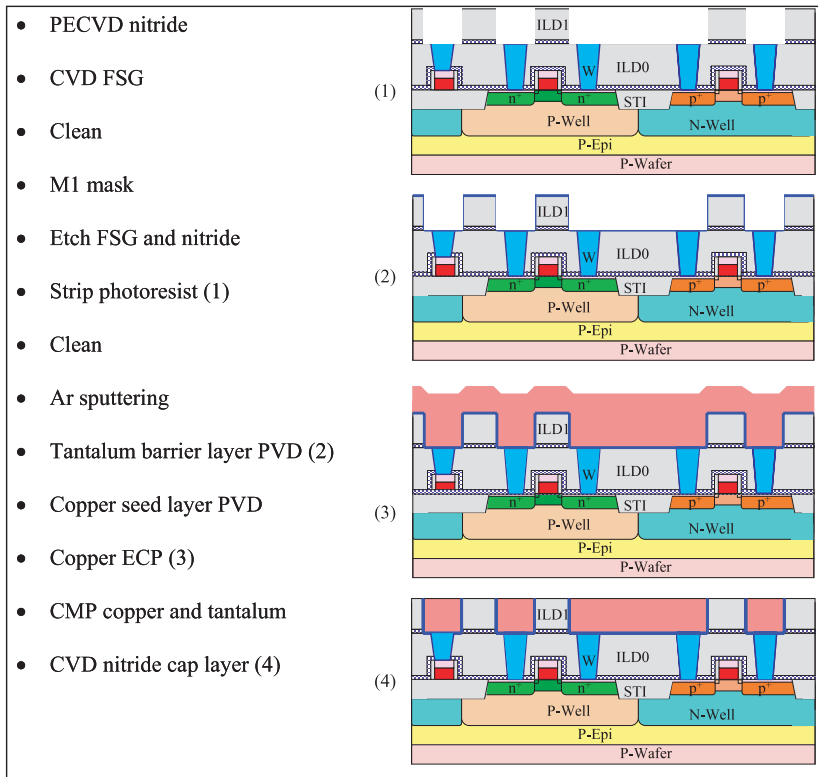


Figure 13.29 CMOS metal-1 copper interconnection processing steps.

13.7.5 Copper and low- κ dielectrics

The combination of copper and low- κ dielectrics can further increase IC chip speed. Several silicon-based low- κ dielectric materials have been developed. They can be etched with fluorine-based chemistry, similar to silicate glass etching. The most commonly used low- κ dielectric in IC manufacturing is SiCOH, or carbon-doped oxide and porous SiCOH, which has an even lower κ value due to pores inside the film. Figure 13.30 lists and illustrates the via-first copper and low- κ interconnection processing steps.

For a via-first dual damascene process, each via needs to stop on the cap layer, which can be broken through using a highly selective etching process after trench formation. It is very important to have the via holes stop on the thin cap layer. Underetching will cause open circuits because breakthrough etching is highly selective of the cap layer, not the low- κ film, while overetching will cause copper corrosion during breakthrough and lead to high contact resistance.

Because the oxygen plasma used to strip the photoresist can oxidize carbon in low- κ film and increase the κ value of the dielectric, a TiN hard mask is used

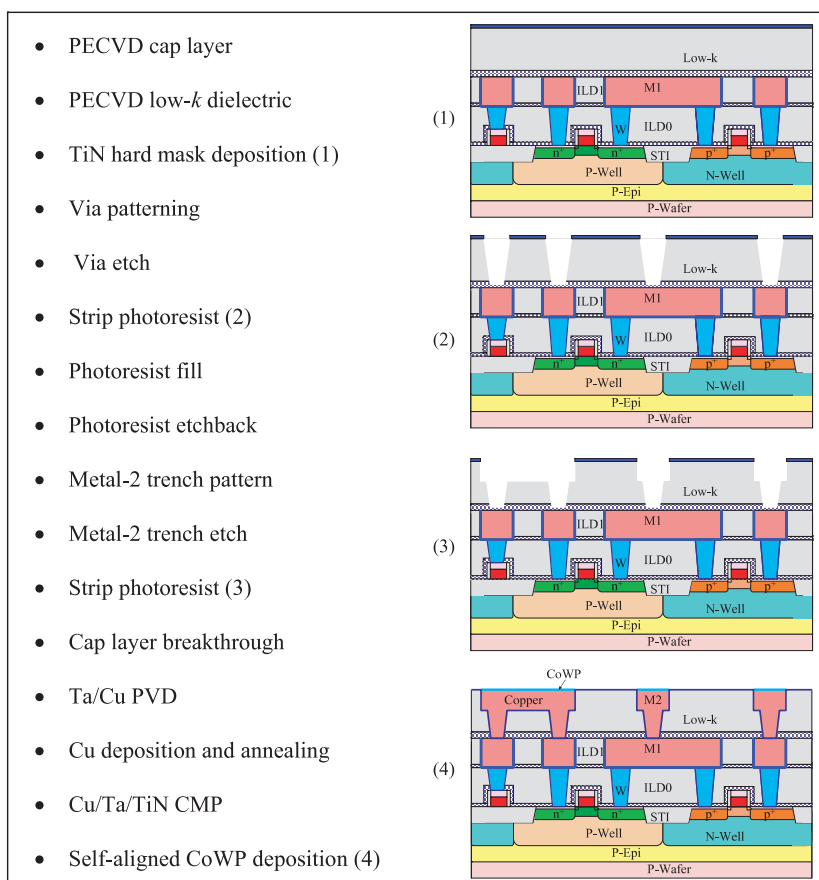


Figure 13.30 Via-first copper interconnection processing steps.

to cover the surface to protect the film. A CoWP cap layer helps prevent copper diffusion and reduces electromigration to improve device reliability. It can be deposited only on the copper surface with an electroless plating process.

Figure 13.31 illustrates a trench-first copper and ULK interconnection process with SEG SiGe pMOS, gate-last HKMG, and trench silicide, which only forms at the bottom of a trench-style contact. Because it is a metal gate, there is no need to form silicide on top of the gate.

13.8 Passivation

After the last metal layer formation, a passivation layer is deposited to protect the IC chip from moisture and other contaminants such as sodium. Silicon nitride is the most commonly used final passivation material in the IC industry. A silicon oxide layer is deposited as a stress buffer before silicon nitride deposition. A silane-based PECVD chamber is used to deposit both oxide and nitride in situ with one

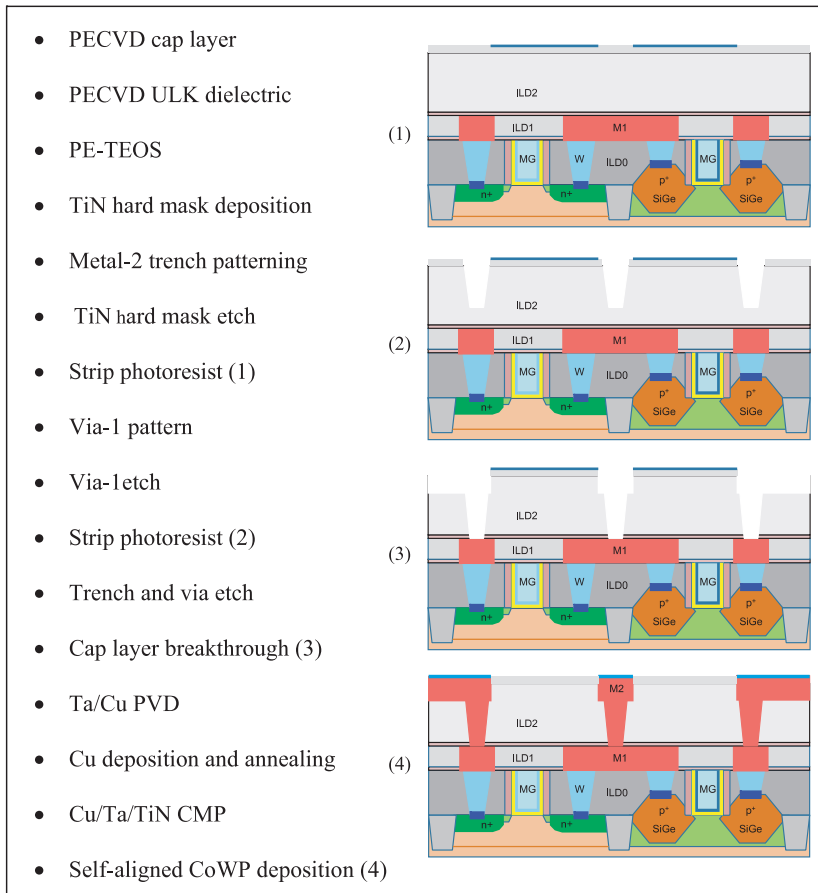


Figure 13.31 Copper/ULK interconnection with trench-first integration.

processing sequence. After nitride deposition, the last photolithography process defines the openings for the bonding pads or connecting bumps. A fluorine-based nitride/oxide etch and photoresist stripping finishes the silicon wafer processing. Figure 13.32 illustrates the passivation processing steps for bonding pad packaging.

13.9 Summary

- Blanket field oxide, LOCOS, and STI are three isolation methods used for IC chips.
- The three-well formation processes of CMOS IC chips are single well, self-aligned twin well, and double-photomask twin well.
- Sidewall spacers are formed by deposition and etchback of a conformal dielectric thin film. They are used to form S/D extensions, selective epitaxy growth, and self-aligned silicide.

- The threshold V_T adjustment implantations control the threshold voltage of MOS transistors.
- Al, polysilicon-silicide, and TiN have been used for gate and local interconnections.
- W, Ti, TiN, and Al–Cu are commonly used in Al–Cu interconnection processes.
- Basic processing steps for copper metallization are dielectric deposition, dielectric etches, metal deposition, and metal CMP.
- Ta and TaN are frequently used barrier layers for Cu metallization.
- SiCOH ($\kappa \sim 2.5$ to 2.8) and porous SiCOH (κ to 2) are commonly used low- κ dielectrics.
- CVD SiON, CVD SiCN, and self-aligned CoWP are used as cap layers to prevent Cu diffusion.

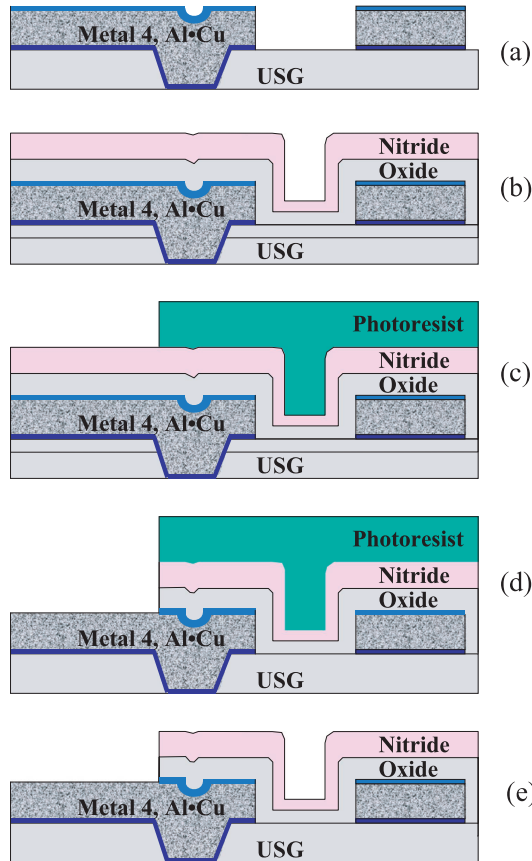


Figure 13.32 Passivation process: (a) metal annealing, (b) PECVD oxide and nitride, (c) bonding pad mask exposure and development, (d) oxide, nitride, and TiN etch, and (e) photoresist strip.

- Silicon nitride is the most commonly used passivation materials in IC processing.

13.10 Bibliography

- M. Fukasawa et al., “BEOL process integration with Cu/SiCOH ($k = 2.8$) low- k interconnects at 65-nm groundrules,” *Proc. IEEE 2005 Intl. Interconnect Technol. Conf.*, 9–11 (2005).
- A. Grill, “Plasma enhanced chemical vapor deposited SiCOH dielectrics: from low- k to extreme low- k interconnect materials,” *J. Appl. Phys.* **93**, 1785–1790 (2003).
- L. Shon-Roy, A. Wiesnoski, and R. Zorich, *Advanced Semiconductor Fabrication Handbook*, Integrated Circuit Engineering Corp., Scottsdale, AZ (1998).
- S. Wolf, *Silicon Processing for the VLSI Era, Vol. 2, Process Integration*, Lattice Press, Sunset Beach, CA (1990).

13.11 Review Questions

1. What is the main advantage of epitaxy layers in an CMOS IC chip?
2. What are the main differences between self-aligned twin-well and double-photomask twin-well processes?
3. Explain the advantages of STI over LOCOS.
4. What is the difference between LDD (or SDE) and S/D implantation?
5. Explain the self-aligned silicide process. Why did nickel silicide replace cobalt silicide when technology nodes grew smaller than 65 nm?
6. Why is tungsten used for interconnection processes?
7. What are the applications of titanium and titanium nitride?
8. List the basic processing steps for both copper and traditional interconnections. Explain the differences.
9. Explain the application of high- κ dielectrics.
10. Explain the application of low- κ dielectrics. What major impact occurs in the process when SiCOH replaces FSG?

Chapter 14

Integrated Circuit Processing Technologies

Objectives

After finishing this chapter, the reader will be able to:

- list the major processing technology changes from the 1980s to the 2010s
- explain the differences between DRAM and NAND flash memory devices.

14.1 Introduction

CMOS-based chips are the most common IC chips in the electronics industry. The digital revolution that produced personal computers, Internet networks, and telecommunications strongly drives the demands of CMOS IC chips. In this chapter, four complete CMOS processes are examined. First is the CMOS process from the early 1980s, with only one layer of aluminum alloy interconnection. Next is the CMOS with a four-layer aluminum alloy interconnection used in the 1990s. After that is an advanced CMOS process with copper and low- κ interconnections, used in the first decade of the 2000s, and the last is the state-of-the-art CMOS technology with high- κ metal gates (HKMGs), stress engineering, and copper low- κ interconnections.

Memory chips are one of most important parts of IC products. They are also important drivers of IC technological developments. The manufacturing processes of array cells of DRAM and NAND flash chips are quite different from CMOS processes. Thus, two sections of this chapter are dedicated to describing their processes. The periphery devices of both DRAM and NAND flash chips are very similar to normal CMOS processes.

14.2 Complementary Metal-Oxide-Semiconductor Process Flow of the Early 1980s

In the mid-1970s, IC processing technology leapt forward after the introduction of ion implantation, which replaced diffusion for semiconductor doping. Self-aligned S/D formation became a standard MOS transistor-making process.

Polysilicon gates replaced metal gates due to the high temperature requirement of postimplantation annealing. Because electrons move much faster than holes, nMOS is much faster than pMOS at the same size and dopant concentration. nMOS replaced pMOS shortly after the introduction of ion implantation, which, unlike diffusion, can easily dope n-types at high concentrations.

In the 1980s, CMOS IC processing technology developed rapidly, driven by the demands of digital logic electronic devices such as watches, calculators, personal computers, mainframe computers, etc. Applications of liquid crystal displays (LCDs) also accelerated the transition from nMOS ICs to lower-power-consumption CMOS ICs. Minimum feature size shrank from 3 μm to 0.8 μm , while wafer size increased from 100 mm (4 in.) to 150 mm (6 in.).

In the early 1980s, LOCOS was used with isolation ion implantation to electronically isolate neighboring transistors. PSG was used as ILD0, which was reflowed at about 1100 °C. Al–Si alloy thin films deposited with thermal or e-beam evaporators were used on tapered contact holes to form metal interconnections. Horizontal furnaces were used for oxidation, LPCVD, postimplantation annealing and drive-in, and PSG reflow. Plasma etchers were used for some patterns such as gate etching, while other larger patterns still used wet etching. Projection alignment and exposure systems were used in photolithography processes. Positive photoresist started to replace negative photoresist to meet photolithography resolution requirements. Most processing tools were batch systems. Figure 14.1 shows a CMOS process flow with early 1980s technology; the minimum feature size is about 3 μm . The figure includes steps for isolation and well formation, transistor formation, and interconnections, as well as 17 illustrations.

Figure 14.2 illustrates a cross-section of a CMOS device using LOCOS isolation, reflowed PSG as ILD0, tapered contacts, and Al–Si alloy as interconnects.

14.3 Complementary Metal-Oxide-Semiconductor Process Flow with 1990s Technology

In the 1990s, CMOS IC processing technology developed rapidly, driven by the continued demands of digital logic electronics such as personal computers, telecommunication devices, and the Internet. Minimum feature size shrank from 0.8 to 0.18 μm , while wafer size increased from 150 mm (6 in.) to 300 mm (12 in.).

A complete CMOS process flow with mid-1990s technology is provided in this section. Minimum feature size of the CMOS chip is about 0.25 μm . Figure 14.3 illustrates epitaxial-layer growth on a p-type wafer. This kind of wafer preparation is usually performed by wafer manufacturers before the wafers are sold and shipped to IC chip manufacturers.

Figure 14.4 illustrates the rest of the CMOS process flow, including STI, well formation, transistor formation, global interconnections, and passivation and bonding pad formation. The figure lists the processing steps and includes 57 illustrations. This ends the wafer process. The wafer is ready for testing, die sawing, sorting, packaging, and final testing.

Isolation and Well Formation

- Bare wafer, p-type
- Clean
- Grow pad oxide
- LPCVD nitride (1)
- Mask 1, LOCOS
- Etch nitride and pad oxide
- Strip photoresist
- Clean
- Isolation implantation (boron) (2)
- Wet oxidation, LOCOS formation (3)
- Strip nitride and pad oxide (4)
- Clean
- Grow screen oxide
- Mask 2, n-well
- n-well implantation (5)
- Strip photoresist
- n-well anneal and drive-in (6)
- Strip screen oxide
- Clean

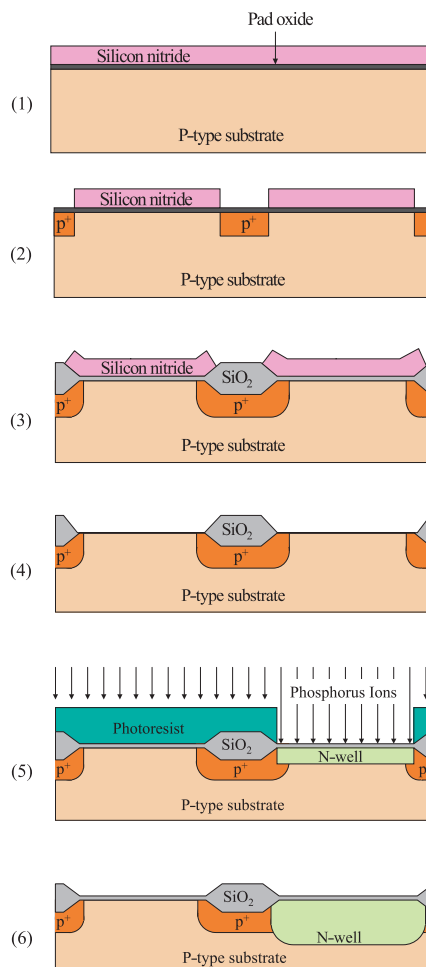


Figure 14.1 CMOS process flow of the early 1980s (*continued on following two pages*).

14.3.1 Comments

Several major developments in CMOS IC chip processing technology occurred during the 1990s. Silicon wafers sliced from single-crystal ingots made by the CZ method always have trace amounts of oxygen and carbon from the crucible materials. To eliminate these impurities and to improve chip performance, epitaxy silicon began to be used for advanced CMOS IC chips, as shown in Fig. 14.3. Shallow trench isolation, shown in Fig. 14.4, replaced LOCOS as an isolation block for preventing crosstalk between neighboring transistors. The sidewall spacer

Transistor Formation

- Grow sacrificial oxide
- Strip sacrificial oxide
- Clean
- Grow gate oxide
- LPCVD polysilicon (7)
- Clean
- Mask 3, gate, and local interconnection
- Etch polysilicon (8)
- Strip photoresist
- Polysilicon annealing and oxidation
- Mask 4, n-channel S/D
- n-channel implantation (phosphorus) (9)
- Strip photoresist
- Mask 5, p-channel S/D
- p-channel implantation (boron) (10)
- Strip photoresist
- Postimplantation annealing, furnace
- LPCVD nitride (11)

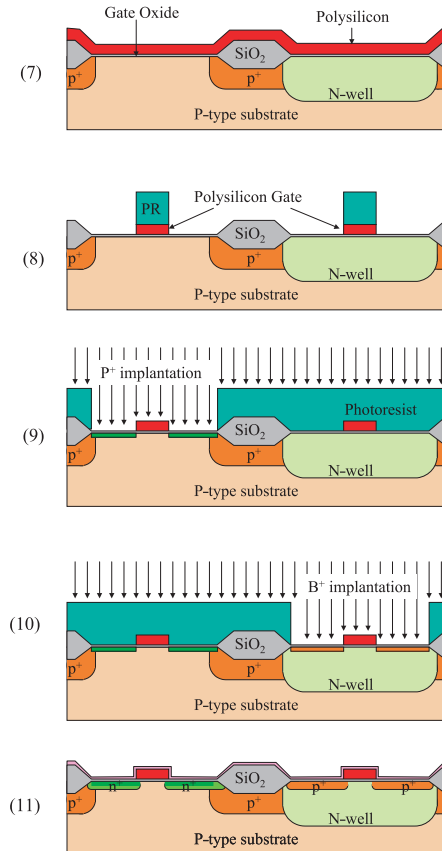
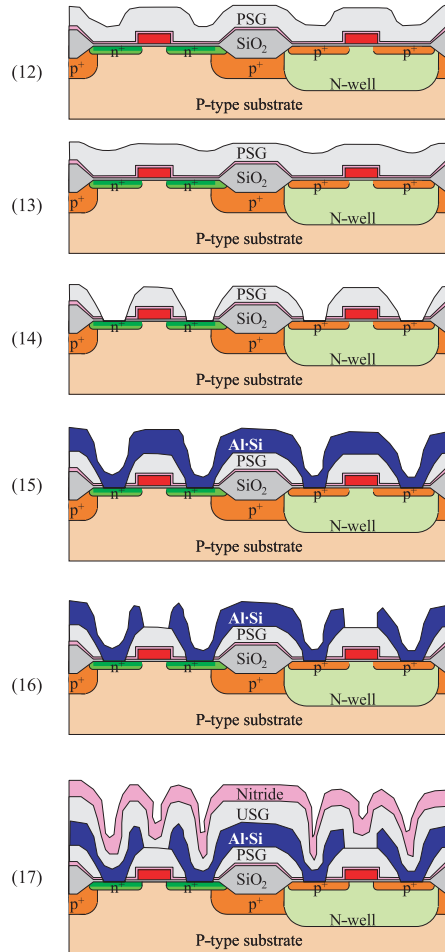


Figure 14.1 (continued)

was introduced to form LDD to suppress the hot electron effect for submicron devices and to form self-aligned silicide (salicide) to reduce resistance of gates and local interconnections. Silicide became widely used with polysilicon to form gates and local interconnections. Because silicide has a much lower resistivity than polysilicon, it improved device speed and reduced power consumption. The most commonly used silicides in the 1990s were tungsten and titanium. In that period, IC chip power supply voltage was gradually reduced from 12 to 3.3 V; therefore, threshold voltage V_T adjustment implantation processes were required to make sure that a normal-off nMOS could be turned on, and a normal-on pMOS could be turned off at the operation voltage. The process flow in Fig. 14.4

Interconnections

- CVD PSG (12)
- PSG reflow, furnace (13)
- Clean
- Mask 6, contact holes
- Etch PSG
- Strip photoresist (14)
- Clean
- Deposit Al-Si alloy (15)
- Mask 7, metal interconnection
- Etch metal
- Strip photoresist
- Metal annealing (16)
- CVD oxide
- CVD nitride (17)
- Clean
- Mask 9, bonding pad
- Etch nitride and oxide
- Strip photoresist
- Clean

**Figure 14.1** (continued)

shows a salicide process that forms titanium silicide on top of the polysilicon gate and S/D at the same time. The silicide at the S/D helps to reduce contact resistance.

Before the 1990s, most IC manufacturers made their own processing tools and developed IC processes themselves. In the 1990s, semiconductor equipment companies rapidly developed. They not only provided processing tools, they also provided integrated processes to the IC fabs. Cluster tools that were capable of running different processes in the same mainframe became very popular in the IC industry. Single-wafer processing systems became widely used because they have

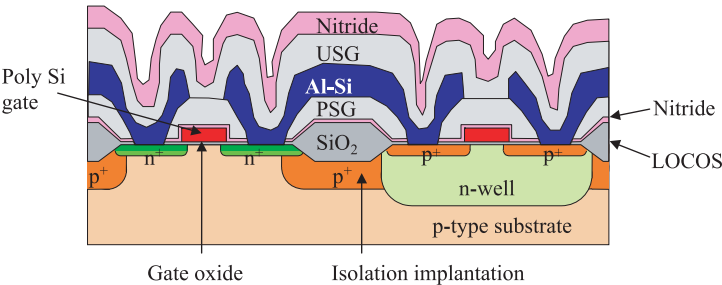


Figure 14.2 Cross-sectional view of a CMOS chip with early 1980s technology.

- Bare wafer, p-type
- Clean
- Deposit p-type epitaxial silicon (shown)

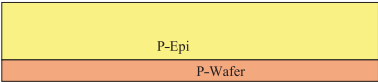


Figure 14.3 Epitaxial-layer growth on a p-type wafer.

better WTW uniformity control; however, batch systems are still currently used in many noncritical processes because of their high throughput.

In the 1990s, photolithography exposure wavelengths shifted from UV to deep ultraviolet (DUV) of 248 nm. Positive photoresists were used in photolithography. Steppers replaced other alignment and exposure systems, and integrated track-stepper systems, which perform photoresist coating, baking, alignment and exposure, and development in one processing sequence, were commonly used. All patterned etches were plasma etch processes, whereas wet etches were still widely applied in blanket film stripping and CVD film quality control. Vertical furnaces became dominant because of their smaller footprints and better contamination control. Rapid thermal processing (RTP) systems became more popular in postimplantation annealing and silicide formation because they are faster and have better processing and thermal budget control. Sputtering replaced evaporation and became the process of choice for metal deposition, and dc magnetron sputtering systems are now common in metal PVD in IC fabs.

Because the number of transistors increased so dramatically, one layer of metal was no longer enough to connect all of the microelectronic devices made on the silicon surface. Thus, multilayer metal interconnections were created to route the circuits. CVD tungsten was used to fill the narrow contact and via holes in the form of plugs to connect the different conducting layers. Titanium and titanium nitride became popular as both barrier and adhesion layers for tungsten. Titanium was also used as a welding layer for aluminum–copper alloys to reduce contact resistance, and titanium nitride was used as an ARC to help the photolithography process pattern metal wires.

Shallow Trench Isolation

- Clean
- Grow pad oxide
- Deposit silicon nitride (1)
- Clean
- Prep/spin-on photoresist/soft bake
- STI mask alignment and exposure (2)
- Postexposure bake/develop/hard bake
- Pattern inspection
- Etch nitride/pad oxide
- Strip photoresist
- Clean
- Etch silicon (3)
- Clean
- Grow barrier oxide
- HDP-CVD USG (4)
- CMP USG (5)
- Anneal USG
- Wet strip nitride and pad oxide (6)

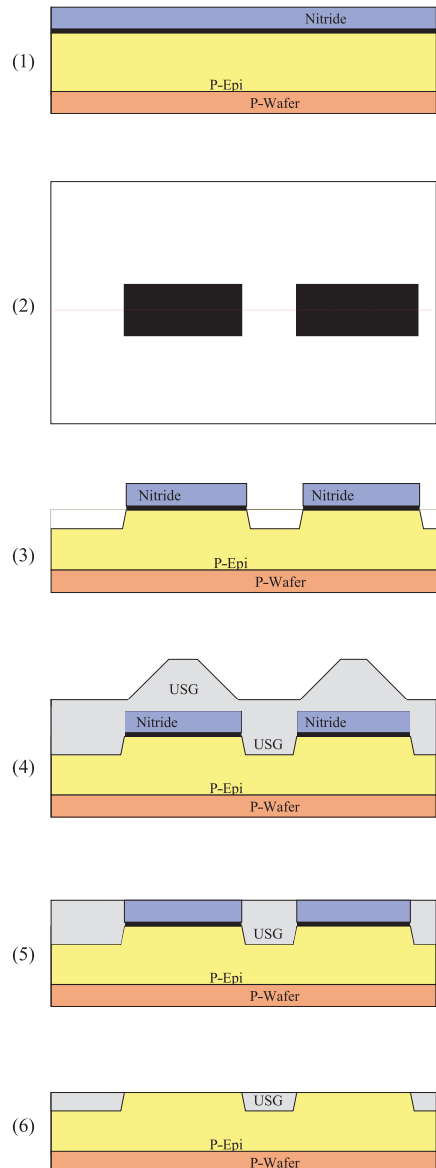


Figure 14.4 CMOS process with multilayer Al alloy interconnects (*continued on following 11 pages*).

Well Formation

- Clean
- Grow screen oxide
- Clean
- Prep/spin-on phtoresist/soft bake
- n-well mask alignment and exposure (7)
- Postexposure bake/develop/hard bake
- Inspection
- n-well ion implantation, phosphorus (8)
- Strip photoresist
- Clean
- Anneal and drive-in
- Clean
- Prep/spin-on photoresist/soft bake
- p-well mask alignment and exposure (9)
- Postexposure bake/develop/hard bake
- Inspection
- p-well ion implantation, boron (10)
- Strip photoresist
- Anneal and drive-in (11)

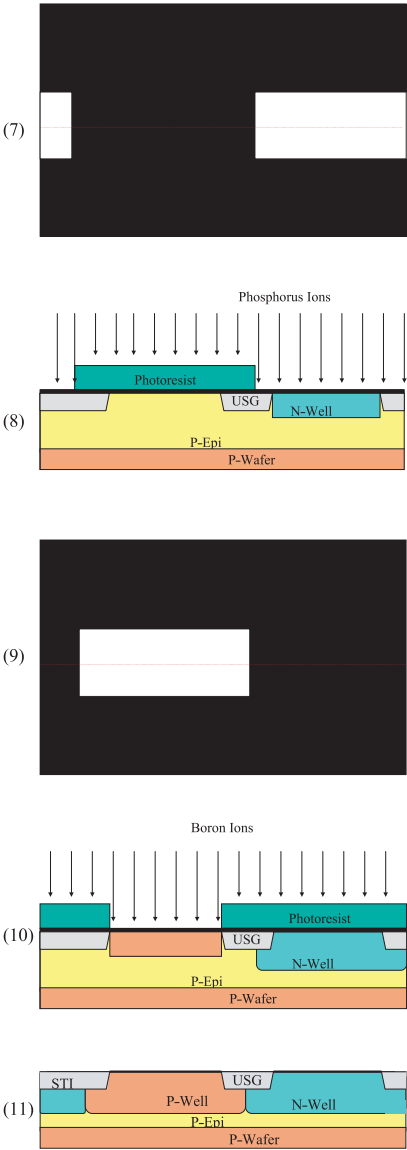
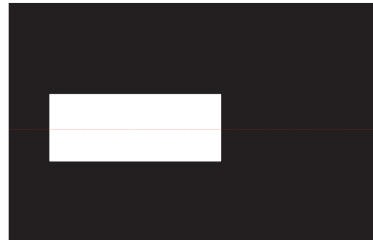


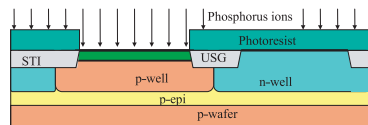
Figure 14.4 (continued)

- Clean
- Grow sacrificial oxide
- Clean
- Prep/spin-on photoresist/soft bake
- n-channel V_T mask alignment and exposure (12)
- Postexposure bake/develop/hard bake
- Pattern inspection
- n-channel V_T adjust implantation (13)
- Strip photoresist
- Clean
- Prep/spin-on photoresist/soft bake
- p-channel V_T mask alignment and exposure (14)
- Postexposure bake/develop/hard bake
- Pattern inspection
- p-channel V_T adjust implantation (15)
- Strip photoresist/clean
- Rapid thermal annealing (16)
- Strip sacrificial oxide

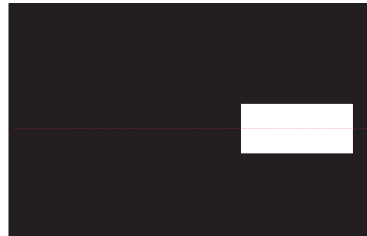
(12)



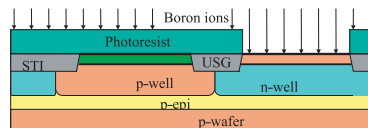
(13)



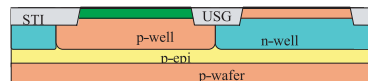
(14)



(15)



(16)



(17)

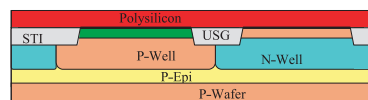


Figure 14.4 (continued)

- Clean
- Strip native oxide
- Grow gate oxide
- Deposit polysilicon (17)
- Clean
- Prep/spin-on photoresist/soft bake
- Gate mask alignment and exposure (18)
- Postexposure bake/develop/hard bake
- Inspection
- Etch polysilicon (19)
- Strip photoresist
- Clean
- Polysilicon anneal/oxidation
- Clean
- Prep/spin-on photoresist/soft bake
- n-channel LDD mask (20)
- Postexposure bake/develop/hard bake
- Inspection
- n-channel LDD implantation (21)
- Strip photoresist
- Clean
- Prep/spin-on photoresist/soft bake

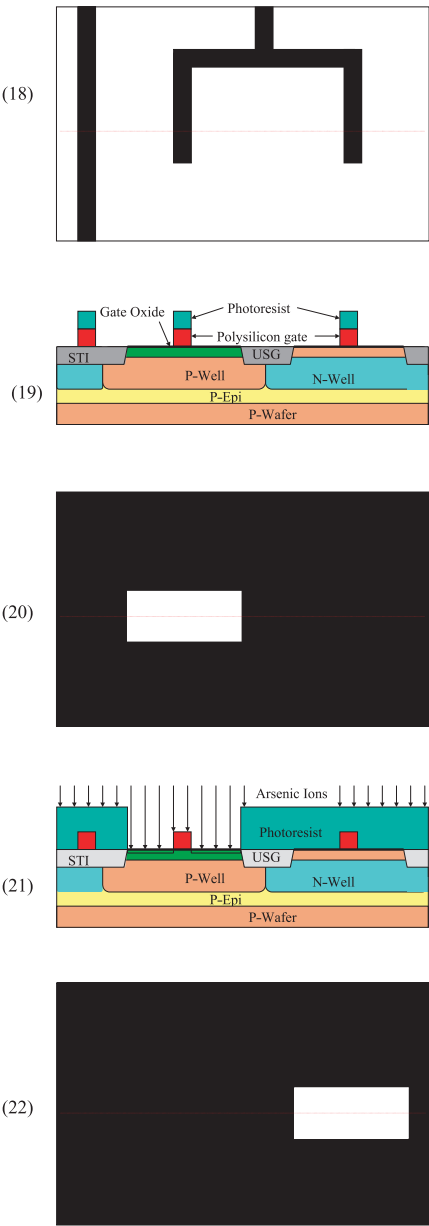


Figure 14.4 (continued)

- p-channel LDD mask (22)
- Postexposure bake/develop/hard bake
- Pattern inspection
- p-channel LDD implantation (23)
- Strip photoresist
- Clean
- LPCVD nitride
- CVD O_3 -TEOS USG (24a)
- USG etchback, stop on nitride (24b)
- Clean
- Prep/spin-on photoresist/soft bake
- n-channel S/D mask (25)
- Postexposure bake/develop/hard bake
- Inspection
- n-channel S/D implantation (26)
- Strip photoresist
- Clean
- Rapid thermal annealing
- Clean
- Prep/spin-on photoresist/soft bake
- p-channel S/D mask (27)
- Postexposure bake/develop/hard bake

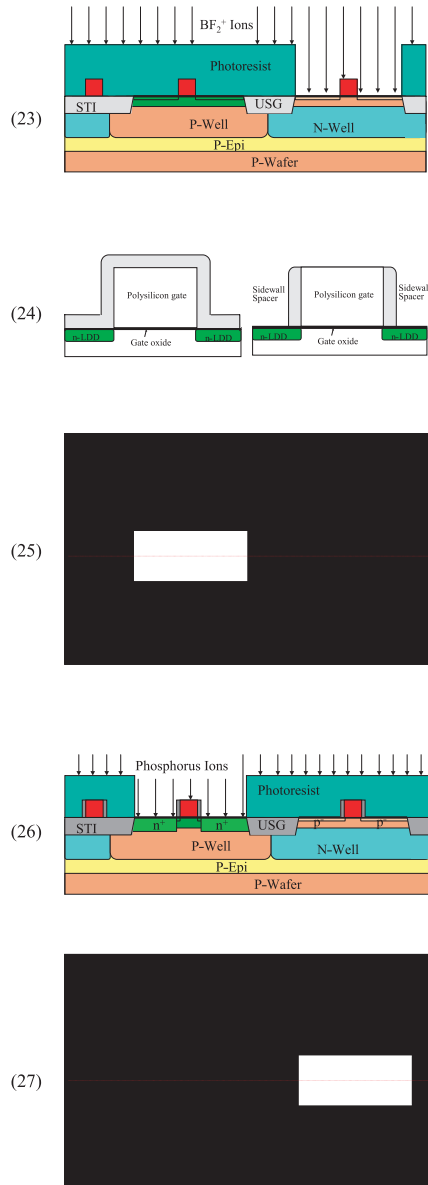


Figure 14.4 (continued)

- Inspection
- p-channel S/D implantation (28)
- Strip photoresist
- Clean
- Rapid thermal annealing

Global Interconnection

- Degas
- Argon sputtering preclean
- Sputtering deposit titanium (29a)
- Rapid thermal anneal (29b)
- Wet strip titanium (29c)
- Second silicide annealing
- Deposit nitride
- O₃-TEOS based BPSG CVD (30)
- Reflow of BPSG (31)
- BPSG CMP
- Clean
- Prep/spin-on photoresist/soft bake
- Contact mask (32)
- Postexposure bake/develop/hard bake
- Inspection

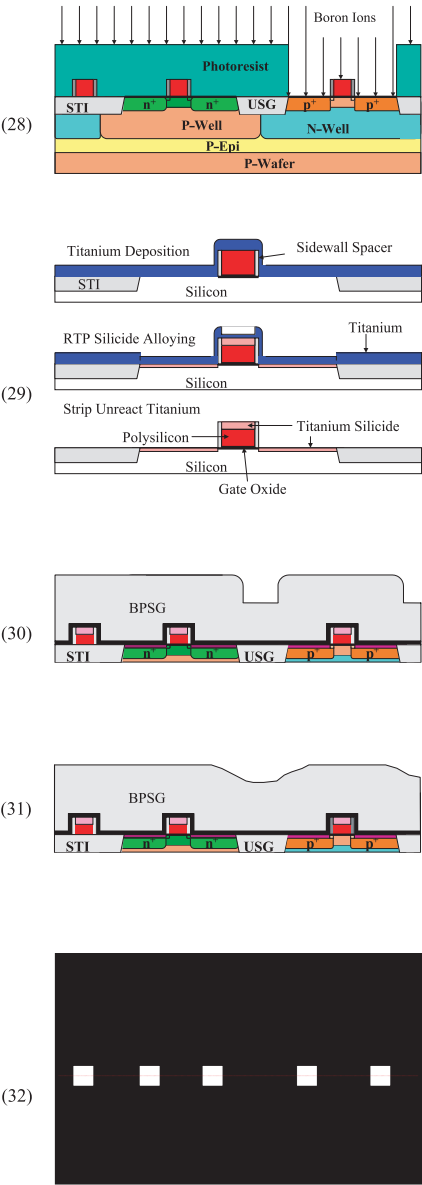


Figure 14.4 (continued)

- Etch BPSG, Stop on silicide surface

- Strip photoresist (33)

- Clean

- Degas

- Argon sputtering pre-PDV clean

- Deposit Ti/TiN by both PVD and CVD

- CVD tungsten (34)

- Polish tungsten/TiN/Ti

- Clean

- Ar sputtering pre-PDV clean

- PVD Ti

- PVD of Al-Cu alloy

- PVD of TiN ARC layer (35)

- Clean

- Prep/spin-on photoresist/soft bake

- Metal-1 mask (36)

- Postexposure bake/develop/hard bake

- Inspection

- Etch metal

- Strip photoresist (37)

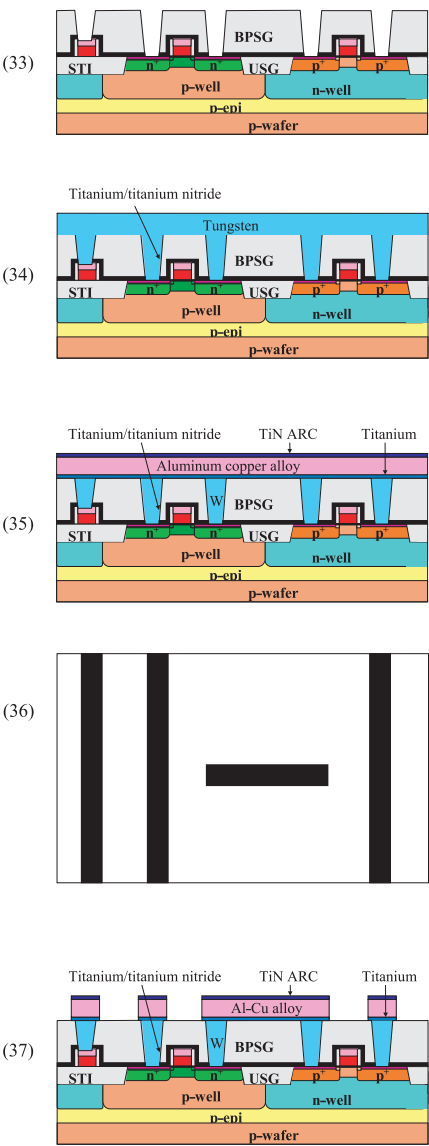


Figure 14.4 (continued)

- Metal anneal
- CVD USG
- Sputtering etchback USG
- CVD USG
- CMP USG (38)
- Clean
- Prep/spin-on photoresist/soft bake
- Via-1 mask (39)
- Postexposure bake/develop/hard bake
- Inspection
- Etch USG
- Strip photoresist (40)
- Degas
- Ar⁺ sputtering preclean
- Deposit Ti/TiN
- Deposit tungsten
- Polish tungsten/TiN/Ti
- Ar⁺ sputtering pre-clean
- Deposit Ti
- Deposit Al•Cu
- Deposit TiN (41)

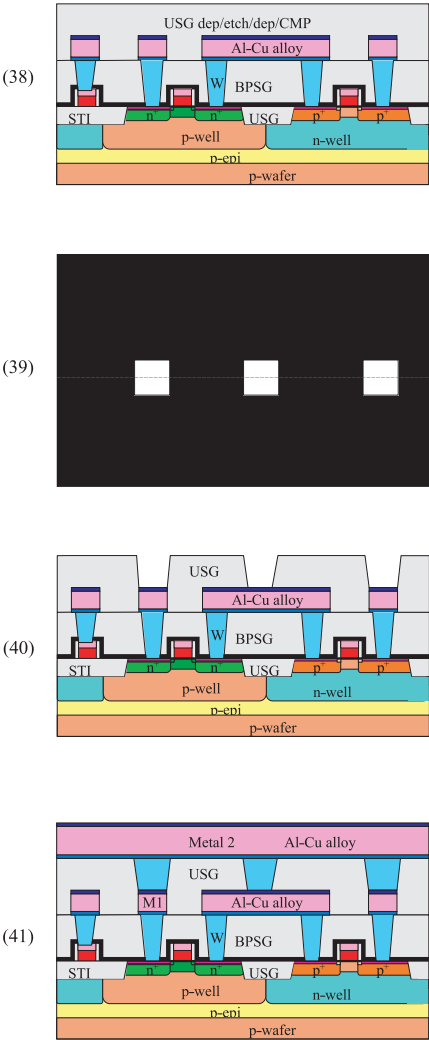
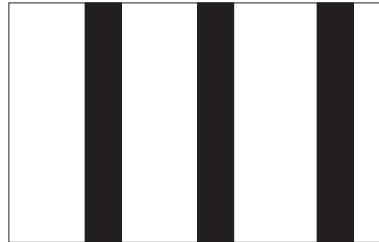


Figure 14.4 (continued)

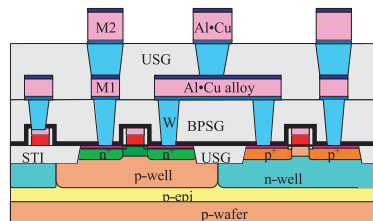
- Clean
- Prep/spin-on photoresist/soft bake
- Metal-2 mask (42)
- Postexposure bake/develop/hard bake
- Inspection

(42)



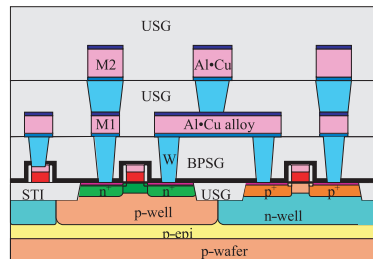
- Etch metal 2
- Strip photoresist/clean (43)
- Metal annealing
- Deposit USG, PE-TEOS
- USG sputter etchback, Ar^+

(43)



- CVD USG, PE-TEOS
- CMP USG (44)
- Prep/spin-on photoresist/soft bake
- Via-2 mask (45)
- Postexposure bake/develop/hard bake
- Inspection

(44)



- Etch USG
- Strip photoresist (46)
- Clean
- Degas
- Ar^+ sputtering preclean

(45)

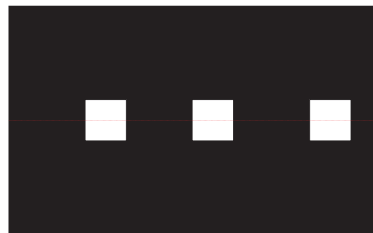
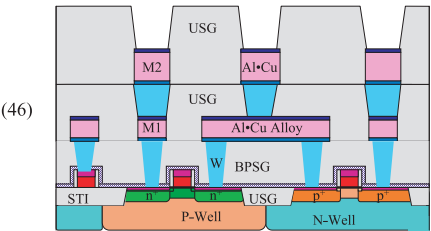
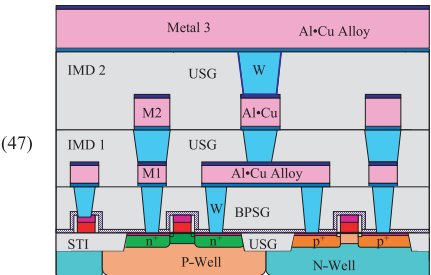


Figure 14.4 (continued)

- Deposit Ti/TiN
- Deposit tungsten
- Polish W/TiN/Ti
- Clean
- Degas



- Ar⁺ sputtering preclean
- Deposit titanium
- Deposit Al-Cu
- Deposit titanium nitride (47)
- Clean



- Prep/spin-coat photoresist/soft bake
- Metal-3 mask (48)
- Postexposure bake/develop/hard bake
- Inspection
- Etch metal 3



- Strip photoresist (49)
- Clean
- Metal annealing
- CVD USG, PE-TEOS
- Ar sputtering etchback
- CVD USG, PE-TEOS

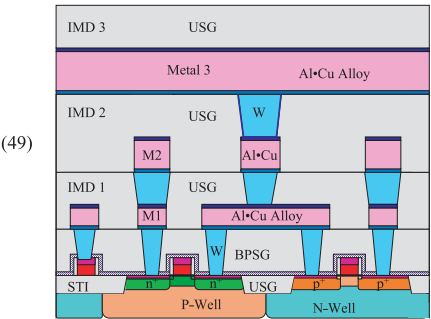


Figure 14.4 (continued)

- CMP USG
- Clean
- Prep/spin-coat photoresist/soft bake
- Via-3 mask (50)
- Postexposure bake/develop/hard bake
- Inspection
- Via-3 etch
- Strip photoresist (51)
- Clean
- Ar sputtering preclean
- Deposit titanium
- Deposit Al-Cu
- Deposit titanium nitride (52)
- Clean/prep/spin-coat photoresist/soft bake
- Metal-3 mask (53)
- Postexposure bake/develop/hard bake/inspection
- Etch metal 4
- Strip photoresist (54)
- Clean
- Metal annealing

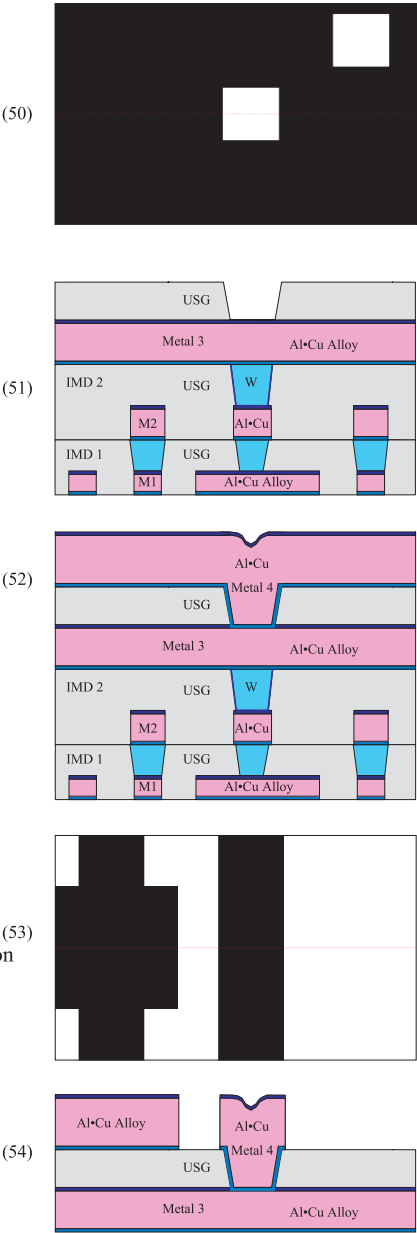


Figure 14.4 (continued)

Passivation and Bonding Pad Formation

- CVD USG, PE-TEOS

- Ar sputtering etchback

- CVD USG, PE-TEOS

- CMP USG

- Deposit silicon nitride (55)

- Clean

- Prep/spin-coat photoresist/soft bake

- Bonding pad mask alignment and exposure (56)

- Postexposure bake/develop/hard bake

- Inspection

- Nitride and oxide etch

- Strip photoresist (57)

- Clean

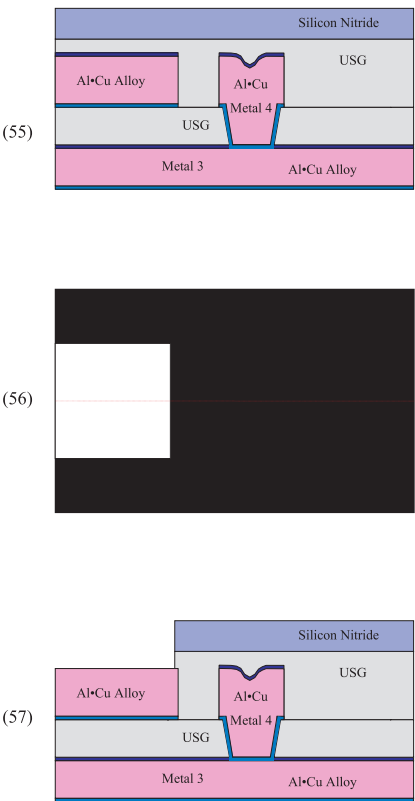


Figure 14.4 (continued)

BPSG was used as a premetal dielectric. By adding boron into the silicate glass, glass reflow temperature reduced from approximately 1100 °C for PSG to as low as 800 °C. This helped to save the precious thermal budget, which dwindled as feature size shrank. PE-TEOS and O₃-TEOS processes were widely employed in dielectric CVD processes for STI, sidewall spacers, and interconnection applications. The CMP process was frequently applied to remove bulk CVD tungsten on the wafer surface during tungsten plug formation. CMP was also used to planarize silicate glass surfaces to achieve better photolithography resolution and to make the succeeding metal deposition processes easier. Figure 14.4 shows the manufacturing process flow of a CMOS IC chip with mid-1990s processing technology. Figure 14.5 shows a cross section of a CMOS IC chip made with mid-1990s processing technology.

14.4 Complementary Metal-Oxide-Semiconductor Process Flow with Technology after 2000

Two factors affect CMOS IC speed: gate delay and interconnection delay. Gate delay is the time required to turn the MOSFET on and off, and interconnection

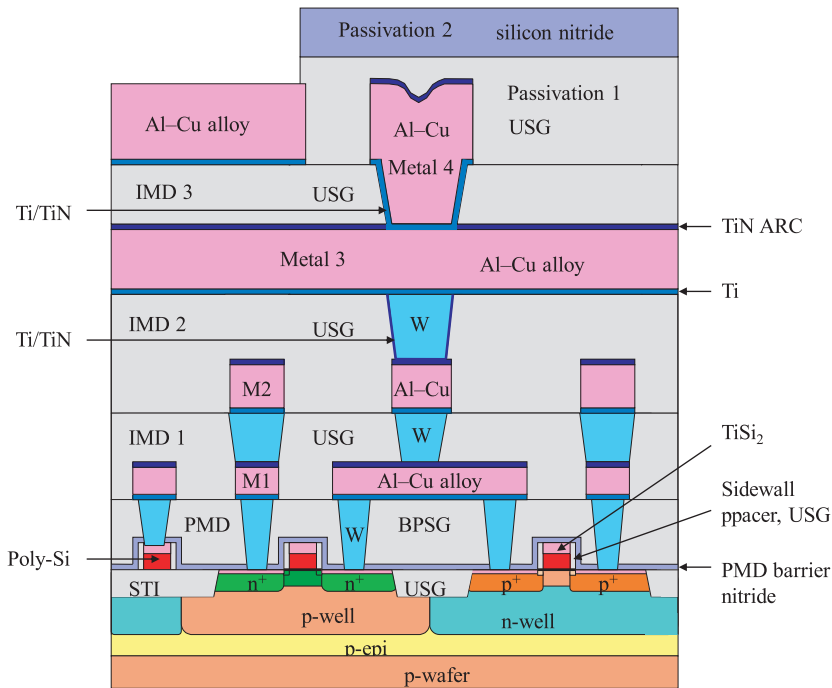


Figure 14.5 Cross section of a CMOS chip with mid-1990s technology.

delay is determined by chip design, processing technology, and the conducting and dielectric materials used for the interconnection.

Gate delay is determined by two factors: the time required to accumulate enough charge to turn on the MOS transistor, and the time required for carriers (electrons for nMOS and holes for pMOS) to pass through the channel underneath the gate between the S/D. The MOS structure of the MOSFET also forms a capacitor with the conducting gate as one electrode, the semiconductor substrate under it as a second electrode, and an insulating layer (gate oxide) between them. MOS capacitance should be large enough to hold the charges (carriers) that form the conducting channel underneath the gate between the S/D when gate voltage exceeds V_T . This allows the MOSFET to turn on. Reducing the gate capacitance can reduce the time required to form the channel and increase the switching speed. However, if the capacitance is too low, the MOSFET becomes unstable because any small noise such as background radiation can turn on or off the transistor and cause a soft error (described in Chapter 8). The distance between the source and drain of a MOSFET is called the channel length, and carriers need to pass through the channel to carry the current. Reducing gate width reduces time for carriers to pass through the channel and increases device speed. However, it also reduces gate capacitance and can cause reliability problems in the device because the MOS capacitance is already designed at the lowest possible level. To further increase

IC chip speed, a substrate with high noise resistance is required to allow smaller feature sizes. Silicon-on-insulator (SOI) is one promising approach, which isolates transistors on the silicon surface from the bulk silicon substrate; therefore, SOI almost completely eliminates the chance of radiation-induced soft errors.

A combination of SOI and STI allows total isolation of the neighboring microelectronic devices and prevents crosstalk between them, a feature that enables chip designers to increase the packing density of transistors on an IC chip. IC chips designed on SOI substrate have been used in high-radiation environments, such as space shuttles, rockets, and scientific research laboratories. Another approach is using a strained silicon channel with a bulk silicon wafer.

Resistance of the interconnection wires and the parasitic capacitance between these wires determines an interconnection delay, or RC delay. To reduce RC delay, metals with lower resistivity and insulators with lower dielectric constant (low- κ) are preferred for forming the interconnection. Copper has lower resistivity than aluminum–copper alloy, so using copper can reduce power consumption and increase IC speed. Traditional aluminum–copper alloy interconnections require one dielectric etch and one metal etch, whereas a copper interconnection usually employs the dual damascene process, which needs two dielectric etches but no metal etch. Dual damascene uses metal CMP instead of metal etch to form the interconnection wires. This is the major difference between copper and aluminum–copper alloy interconnections. The main challenges for copper interconnections are dielectric etch, metal deposition, and metal CMP.

Several low- κ dielectric materials were developed with two competing approaches: CVD and spin-on dielectric (SOD). The advantages of a SiCOH-based CVD low- κ dielectric are familiar technologies, existing processing equipment, and experience. One important advantage of SOD was its extendibility to very low dielectric constants ($\kappa < 2$) with porous silica. However, reliability issues of SOD during chip packaging processes eventually determined CVD SiCOH as the best material for low- κ dielectrics used for mass production of advanced IC chips.

Bonded SOI uses two wafers: one is implanted with high-current hydrogen ions to form a hydrogen-rich layer under the silicon surface, and the other wafer grows a silicon dioxide layer on the surface (Fig. 14.6). Then, the two wafers are pressed together face to face at a high temperature to bond them, with silicon dioxide in between. At high temperatures, the hydrogen atoms in wafer A in Fig. 14.6 react with silicon atoms to form gaseous byproducts ($4\text{H} + \text{Si} \rightarrow \text{SiH}_4$), which induce voids in wafer A. This voided hydrogen-rich layer has a very high wet etch rate; therefore, wafer A can be easily separated from the bonded wafer in a wet etch process. Then, a CMP process is applied to remove defects and roughness on the silicon surface, and this also makes it highly flat and smooth (see Fig. 14.7). The thickness of the silicon layer above the buried silicon dioxide is controlled by hydrogen implantation energy and CMP time. It ranges from several hundred to approximately ten nanometers, determined by device requirements.

Another way to make SOI wafers is to use very-high-current oxygen ion implantation to form the oxygen-rich layer beneath the silicon surface. High-temperature ($>1200^\circ\text{C}$) annealing is performed, leading to the formation of a

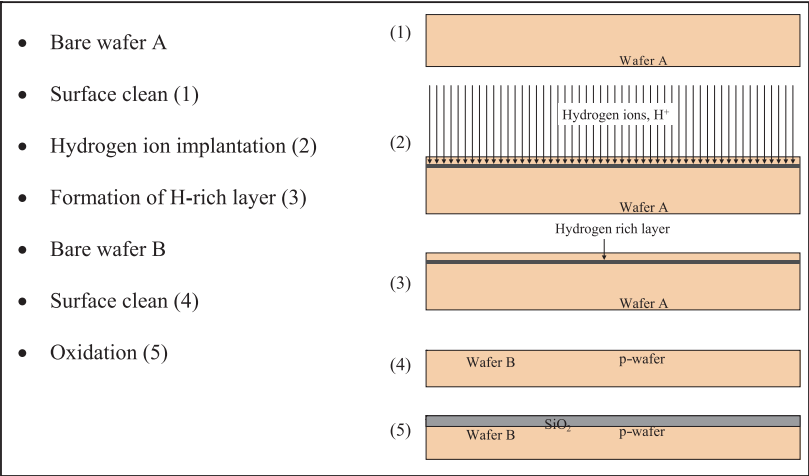


Figure 14.6 Bonded SOI (wafer A): Preparation of wafers.

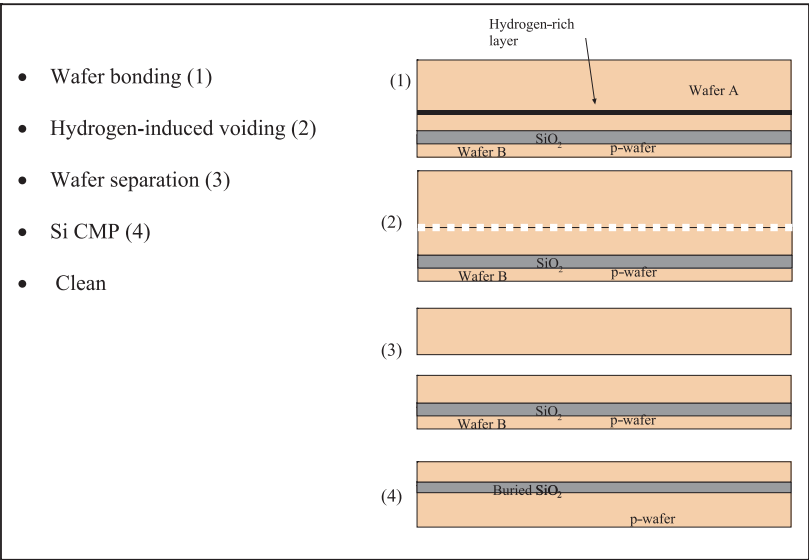


Figure 14.7 Bonded SOI (wafer B): wafer bond, wafer separation, and CMP.

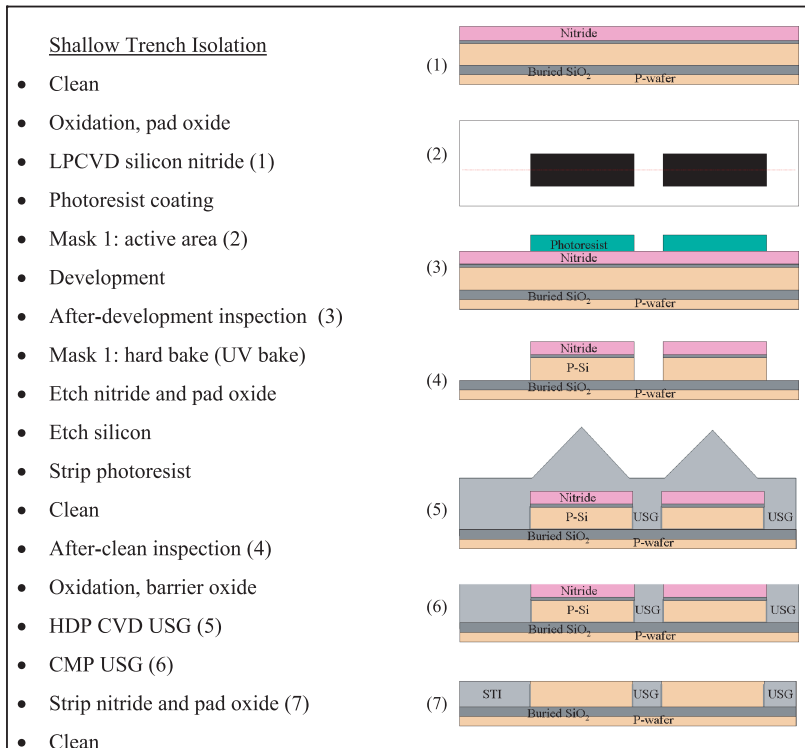


Figure 14.8 STI formation process.

buried oxide layer underneath a thin single-crystal silicon layer. Epitaxial silicon is deposited on the wafer surface, preventing oxygen in the silicon from slowing down the device.

Because activation areas are surrounded by trenches that were etched all the way to the buried silicon dioxide [see Fig. 14.8, illus. (4)], they are completely isolated by silicon dioxide after CVD trench fill and oxide CMP. This total isolation completely eliminates any possibility of crosstalk between neighboring devices and allows transistors to be densely packed. It also solved the radiation-induced soft error issue to allow further shrinking of the device.

Because feature size shrinks, the junction depths of both n- and p-wells are reduced. Therefore, existing high-energy ion implanters can directly implant dopant ions at the required depths, and a drive-in process is no longer needed for well formation. Several implantations at different energy levels are required to achieve well formation.

Without the requirement of well junction drive-in (in which thermal diffusion drives dopant atoms deep into the substrate at high temperatures), engineers can use the same photomask to perform well and V_T adjustment implantations, as shown in Fig. 14.9. Since ion implanters can precisely select the desired ion species with

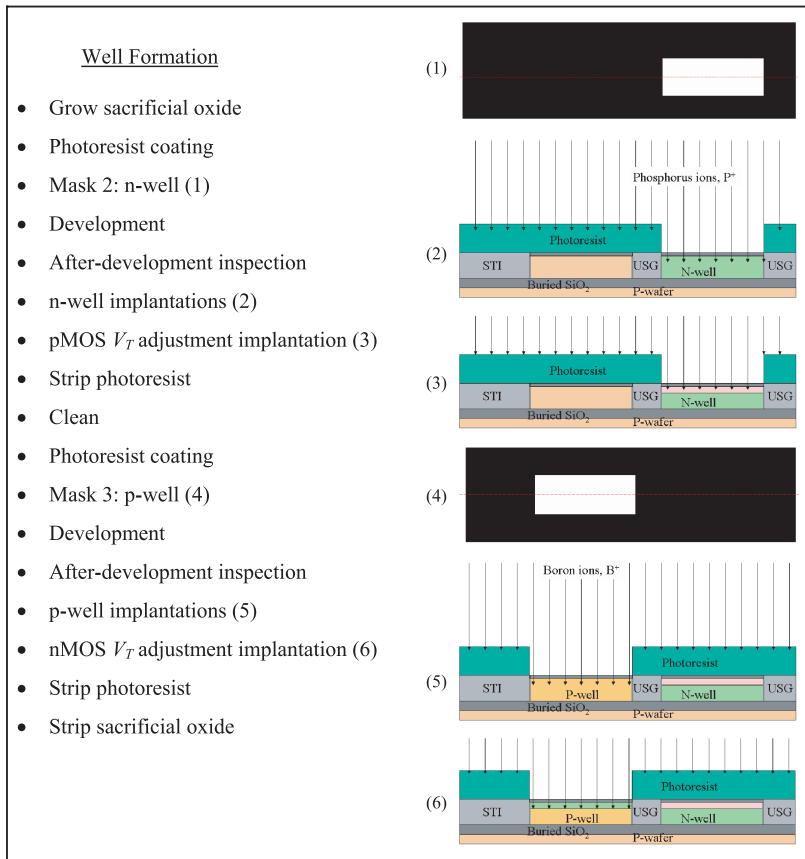


Figure 14.9 Well and threshold adjustment implantations.

their magnetic mass analyzer, all of these implantation processes can be finished with a high-energy, low-current implanter in one processing sequence.

To better control the threshold voltage of the MOSFET, an nMOS polysilicon gate needs to be heavily doped in n-type, and pMOS polysilicon needs to be heavily doped in p-type. By blanket implantation of n-type dopant and selective implantation of a heavier p-type counterdopant, doping nMOS polysilicon with n-type and pMOS polysilicon with p-type can be achieved with one photomask. This is preferred, especially for cost-sensitive DRAM manufacturers, because it helps reduce manufacturing costs and improve device yield. Figure 14.10 shows the polysilicon doping and gate formation processes.

Polysilicon consists of many single-crystal units called grains. Large grain size is preferred because it has less of a grain boundary to deflect electrons, leading to lower resistivity. However, large grain size can cause substantial surface roughness on the sidewalls of the polysilicon lines after etching. For small gates, etching amorphous silicon (α -Si), then annealing it to form polysilicon is preferred. Heavy

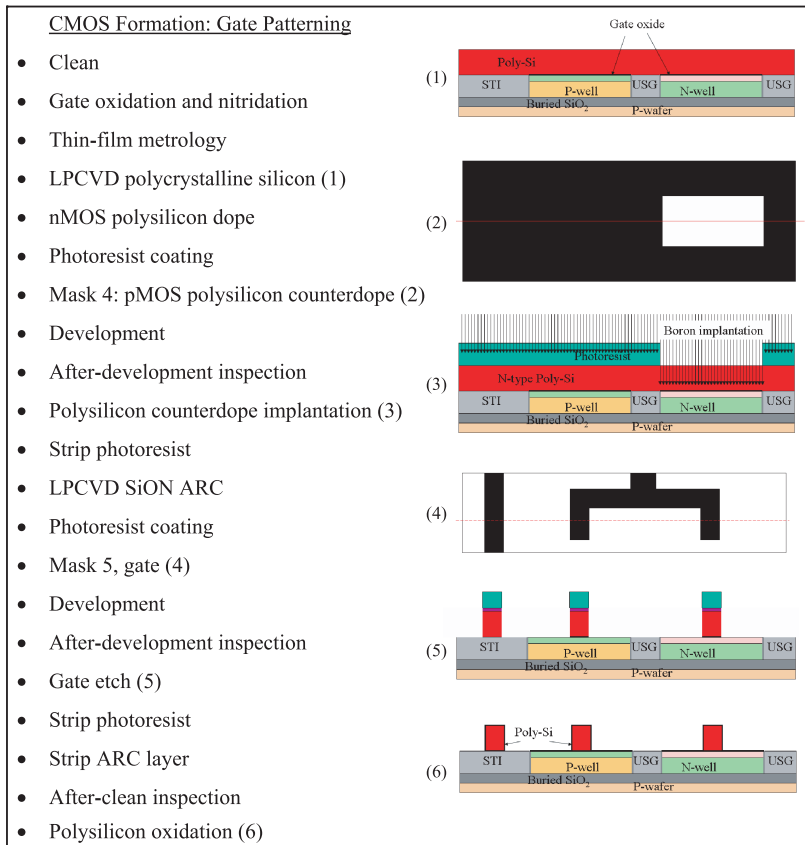


Figure 14.10 CMOS gate patterning process.

polysilicon doping implantation can create amorphous silicon, which allows better etching profile control than polysilicon. The grain size of the polysilicon annealed from α -Si is also more uniform than that of LPCVD polysilicon. After gate etching, plasma-induced gate oxide damage can be repaired by polysilicon oxidation during the annealing process.

Heavy ions are used to form the shallow junctions of the S/D extension shown in Fig. 14.11, usually BF_2^+ for pMOS SDE and As^+ or even heavier Sb^+ for nMOS SDE.

For spacer formation, both nitride and oxide are used. In Fig. 14.12, CVD oxide is used as the etchstop layer, and LPCVD nitride forms the main body of the sidewall spacer.

Figure 14.13 illustrates the halo junction and S/D junction formation of a CMOS. Halo implantation is a large tilt-angle implantation process. It usually requires two or four implantations, depending on whether the MOSFET is placed

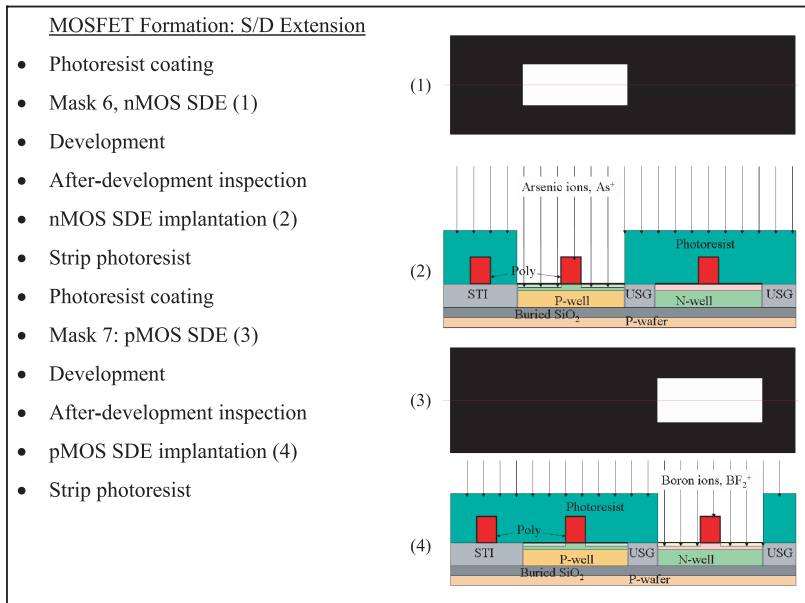


Figure 14.11 Formation of the S/D extension.

in one direction or two. Halo implantation is used to prevent punch-through of the device.

The grain size of titanium silicide must be larger than $0.2\ \mu\text{m}$ to achieve lower resistance. If the gate width is narrower than $0.2\ \mu\text{m}$, a titanium silicide application would be in trouble. Cobalt silicide has started to replace titanium silicide for gate applications after the $0.18\text{-}\mu\text{m}$ technology node. Because cobalt is very reactive and easily forms cobalt oxide when in contact with air or moisture, titanium nitride is needed to cap the cobalt to prevent contact with the atmosphere. Cobalt and titanium nitride are deposited in different PVD chambers in the same cluster tool in an integrated processing sequence.

As device dimensions shrank deep into nanometer technology nodes, the annealing temperature of CoSi_2 ($\sim 750\ ^\circ\text{C}$) became too high for the thermal budget of a tiny MOSFET. Nickel silicide (NiSi), which can form at temperatures lower than $500\ ^\circ\text{C}$, became widely used after devices reached the 65-nm technology node and smaller.

Figure 14.14 illustrates the self-aligned NiSi formation process. At first, an argon sputtering etch is required to remove native oxide from the silicon surface before nickel deposition; otherwise, contact resistance can be too high due to poor silicide formation, and this can cause electronic failure of IC chips.

Because NiSi is not thermally stable, Ni tends to further react with silicon and diffuse across the junction, causing severe device leakage. Platinum (Pt) can be alloyed with the PVD target to form NiPtSi on the wafer surface to achieve better

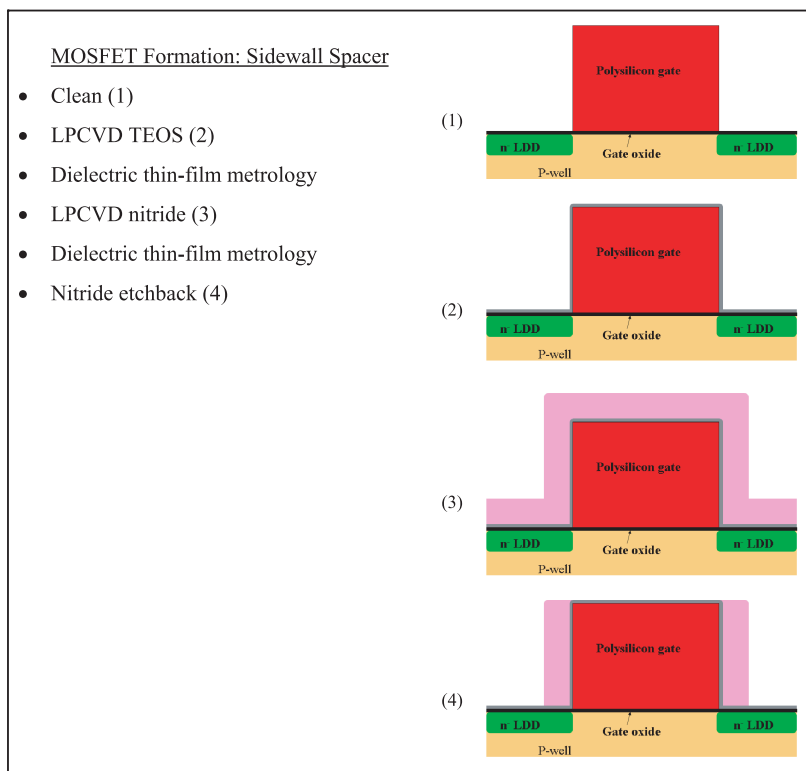


Figure 14.12 Formation of a sidewall spacer.

stability of the silicide. The yield-impacting Ni diffusion or encroachment issue can be captured and monitored with electron beam inspection (EBI) systems.

Figure 14.15 shows the contact module of a sub-90nm CMOS process. After NiSi formation, a nitride layer is necessary to prevent phosphorus from PSG diffusing into the active area. Because of thermal budget limitations, PECVD nitride deposited at a lower temperature (<580 °C) can be used to replace LPCVD nitride, which is normally deposited at 750 °C. For small devices (<0.18 μm), there is no thermal budget for ILD0 thermal reflow; therefore, boron is no longer needed in silicate glass, and PSG replaces BPSG as the material of choice for ILD0. PSG is planarized with CMP instead of reflow. Tungsten is only used for local interconnections and as plugs between the metal and S/D and silicide. Titanium and titanium nitride are still used as barrier and glue layers for tungsten.

For some advanced-technology-node CMOS processes, USG is used for ILD0; the nitride layer is used as a stress liner to strain the channel, enhance carrier mobility, and improve MOSFET performance.

The contact module is very critical because it is the connection between the devices on the wafer surface and the metal interconnect wires above. If the contact

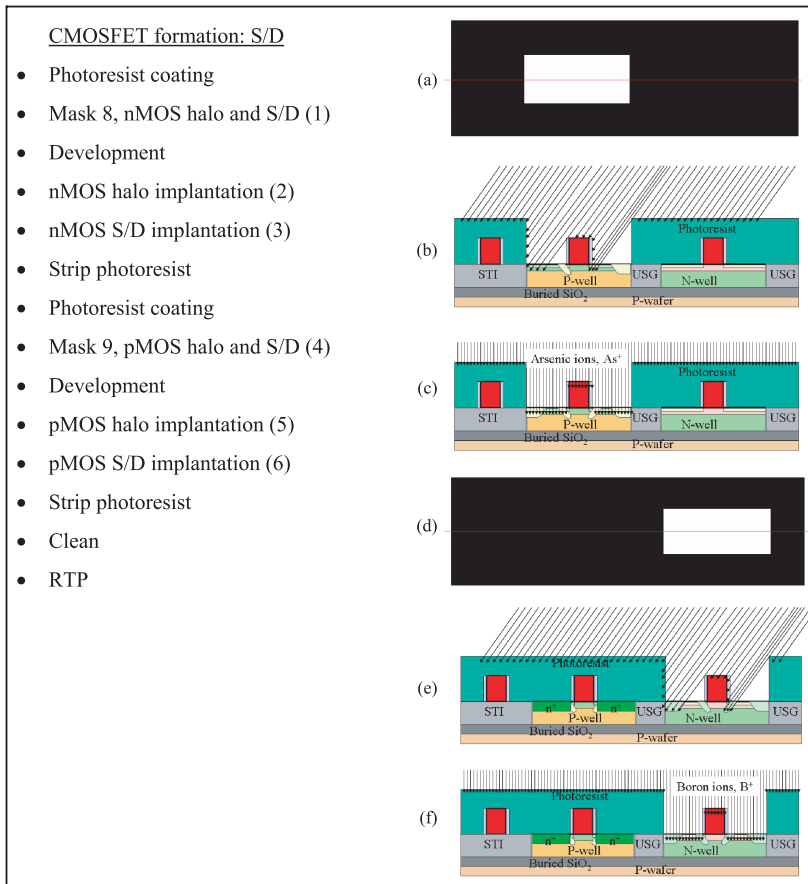


Figure 14.13 S/D implantations.

holes are not etched all the way through, the metal wires will not connect to the devices, and overall yield will suffer.

To reduce resistance, PVD titanium is widely used as a wetting layer for contacts. Titanium nitride serves as a glue layer for tungsten. Without TiN, tungsten film does not stick well to the wafer surface; it will crack and peel off, generating large numbers of particles on the wafer and along the path of the wafer. TiN can be deposited with both PVD and CVD. As device feature size shrinks, the aspect ratio of the contact hole becomes larger, and PVD processes can no longer offer sufficient step coverage; thus, CVD TiN processes have become more popular.

WCMP is one of most important layers for applications of EBI, which allows engineers to both capture device leakage issues and locate contacts that did not open. Application of EBI can help accelerate yield education, reduce the technology development cycle, and shorten yield ramp time.

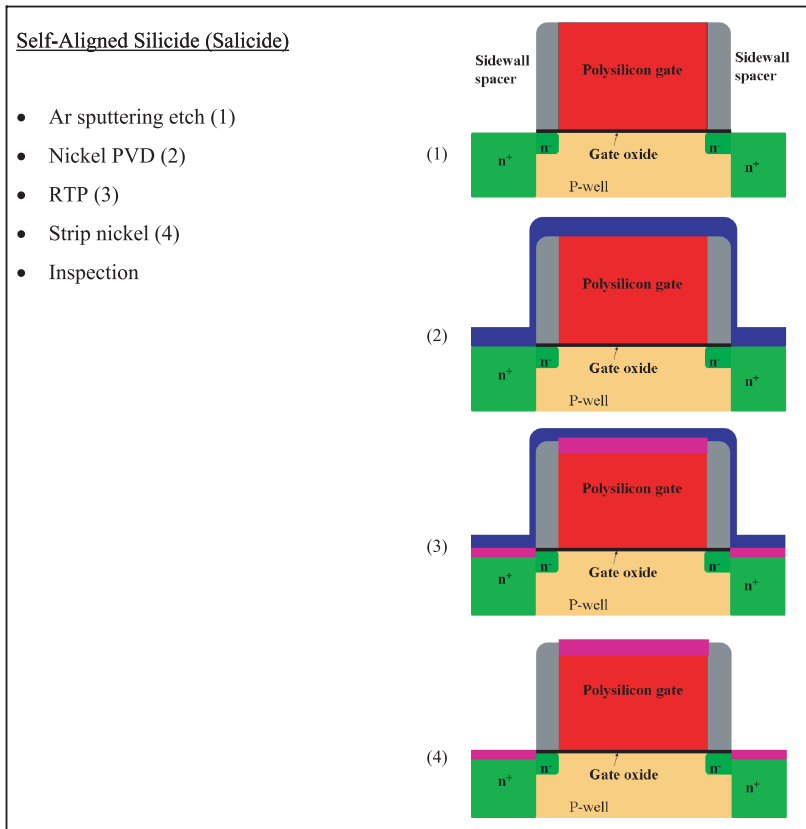


Figure 14.14 Salicide formation.

Figure 14.16 shows the metal-1 copper interconnect with a single damascene process. Silicon carbon nitride (SiCN) is a dense material that can replace silicon nitride as a barrier layer for preventing copper diffusion. It also can serve as an etch-stop layer (ESL) in copper interconnect processes. It has a lower dielectric constant ($\kappa = 4$ to 5) than silicon nitride ($\kappa = 7$ to 8); therefore, using SiCN reduces the overall dielectric constant of ILD layers. The most commonly used low- κ dielectric material is PECVD SiCOH, which is widely used in interconnect processes. Tantalum barrier and copper seed layers are deposited with a PVD process, usually with ionized metal plasma processing to enhance bottom coverage. Because a low- κ dielectric has less mechanical strength than silicate glass, the downforce of the copper CMP process with a low- κ dielectric must be lower than that of USG or FSG. After metal-1 CMP, an electric test can be performed for the first time on the test structures. Tiny probes contact the pads of the test structures and apply voltage or current to test the electrical properties of the devices. To avoid yield loss from copper oxidation, there is a time limit between copper CMP and cap layer deposition. Defect inspections with optical and electron beam inspection

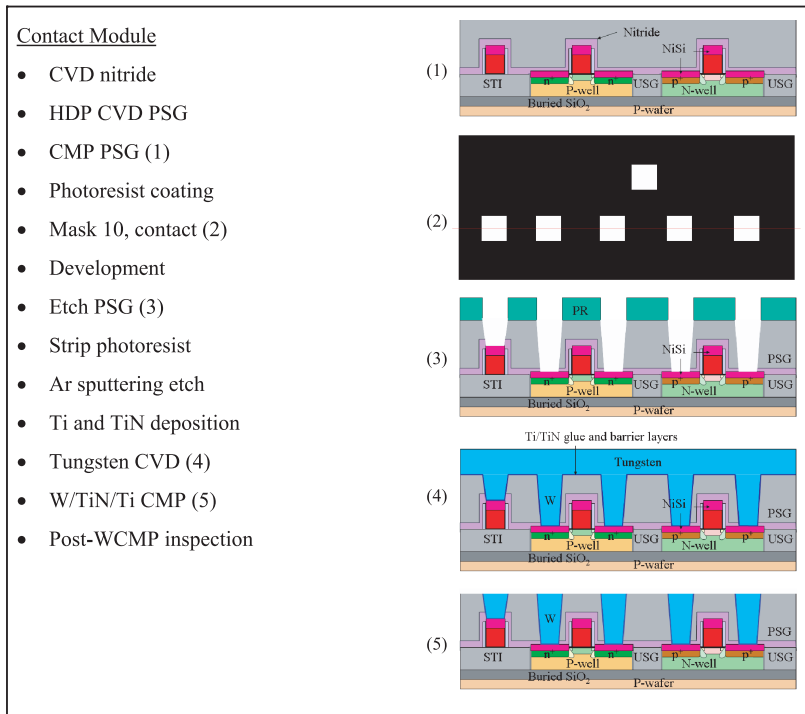


Figure 14.15 Contact module of a CMOS device.

systems are performed after cap layer deposition because of this queued time restriction.

Dual damascene processing is commonly used for copper metallization, which needs two dielectric etch processes. There are at least three different approaches for forming the trench–via structures of dual damascene copper metallization processes. One method etches the trench first, then etches the via, as shown in Fig. 14.17. Another method is the buried hard mask, which etches via holes on the ESL first, then forms both trench and via at the same time using a trench mask, as shown in Fig. 14.18. In Fig. 14.19, the via-first approach is illustrated. Both via-first and trench-first methods are used in copper metallization in IC fabrications.

The ESL helps to define the depth of trench and via holes by providing etch byproducts in the plasma to emit light signals. These signals indicate the endpoint during low- κ dielectric etch processes. PE-TEOS USG and SiCN can be etched with F/O chemistry. For buried-hard-mask dual damascene etching (see Fig. 14.18), high selectivity to SiCN of the low- κ dielectrics is required.

Tantalum (Ta), tantalum nitride (TaN), and the combination of both can be used as a barrier layer for copper to prevent diffusion through the dielectric layer into the silicon substrate, possibly ruining transistors. A copper seed layer is required for

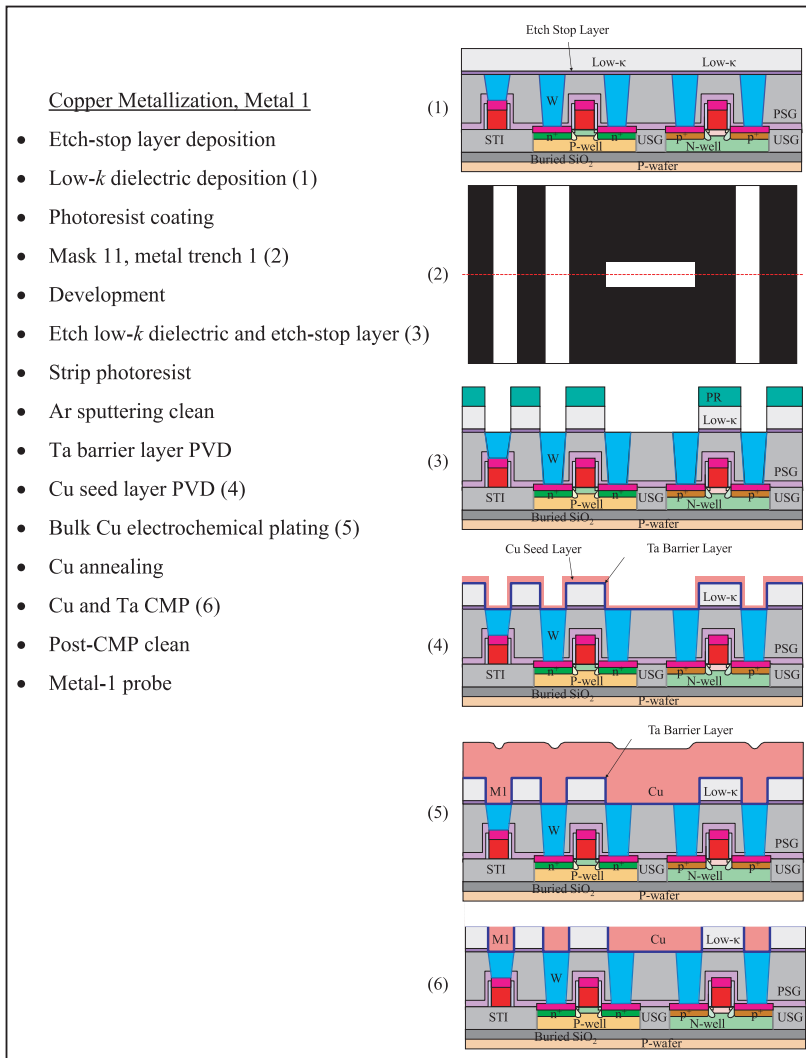


Figure 14.16 Metal-1 process.

bulk copper electrochemical plating (ECP). After the wafer is seeded, it is plated with bulk copper in the ECP process, which fills the narrow trenches and via holes. It is very important to plate the bulk copper soon after copper seed layer deposition because even at room temperature copper can quickly self-anneal. Annealed copper seeds have larger grains and rough surfaces, which can cause voids in the trench and vias during ECP and induce yield loss.

After copper plating, the wafer is annealed in a furnace at about 250 °C to increase grain size and reduce resistivity. Cu and Ta are then removed from the

Via 1 Metal 2 (Trench-First)

- Deposition of ILD stack (via etch stop, low- k , trench etch stop, low- k , and cap layers) (1)
- Photoresist coating
- Mask 12, metal trench 2 (2)
- Development
- After-development inspection
- Etch trench, stop on TESL
- Strip photoresist
- After-etch inspection (3)
- Photoresist coating
- Mask 13, Via 1 (4)
- Development
- After-development inspection
- Via-1 etch
- Strip photoresist
- After-etch inspection (5)
- Ar sputtering clean
- Ta barrier layer PVD
- Cu seed layer PVD
- Bulk Cu electrochemical plating
- Copper annealing
- Cu and Ta CMP (6)
- Post-CMP clean
- Metal-2 probe

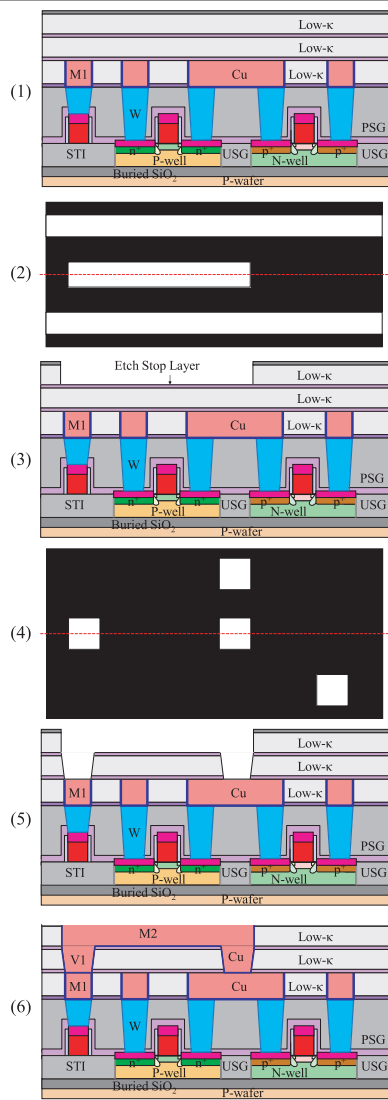


Figure 14.17 Trench-first dual damascene copper metallization.

wafer surface by a CMP process, and only metals in the trenches and via holes are left behind to form the circuit interconnections. Reflectivity of the wafer surface is used to determine the endpoint of the metal CMP because most metals are very shiny with high reflectivity. When metal is polished away from the surface, its reflectivity reduces dramatically, a reduction that indicates the endpoint.

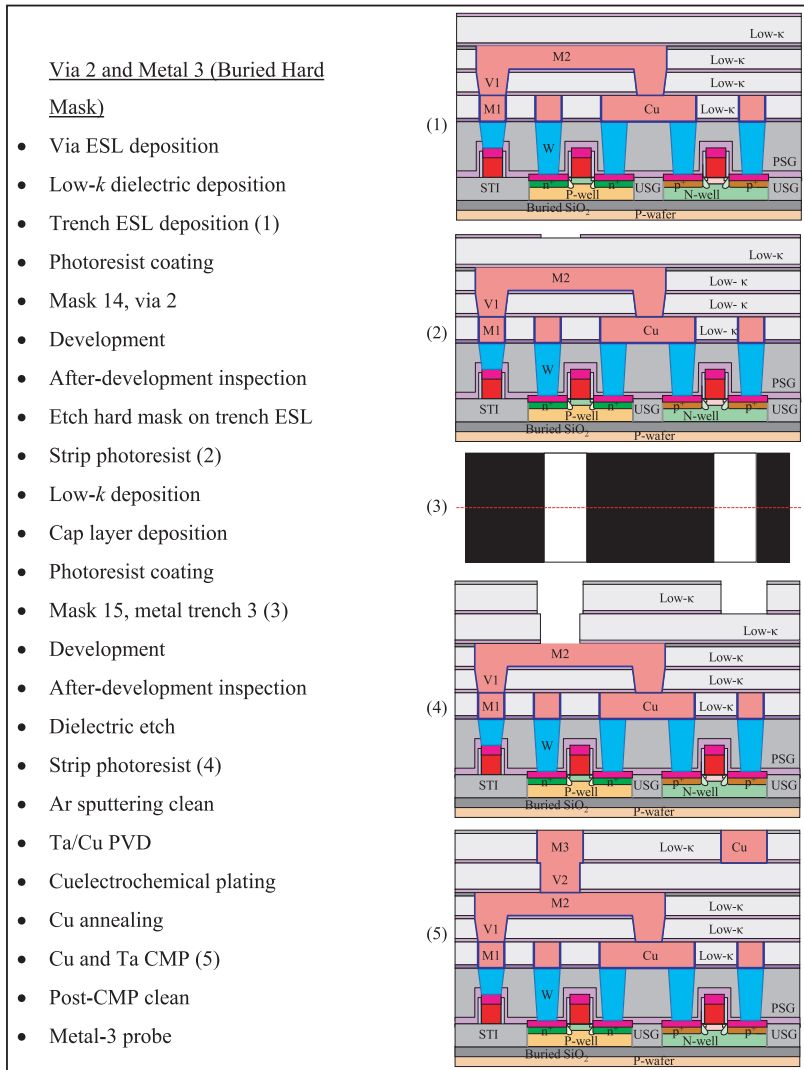


Figure 14.18 Buried-hard-mask copper metallization.

For via-first processes, the ILD of the via ESL, low- κ dielectric, trench ESL, low- κ dielectric, and cap layer are deposited, as shown in Fig. 14.19, illus. (1). The via ESL and trench ESL can be silicon nitride (SiN) or silicon carbon nitride (SiCN). The via mask [Fig. 14.19, illus. (2)] defines the pattern, and via etching stops on the via ESL [Fig. 14.19, illus. (3)]. Before trench patterning, a dummy photoresist layer is coated on the wafer surface to fill via holes and protect via ESL from the trench etching process [Fig. 14.19, illus. (4)]. After trench etching, the photoresist strip process also removes dummy photoresist in the via holes. A wet

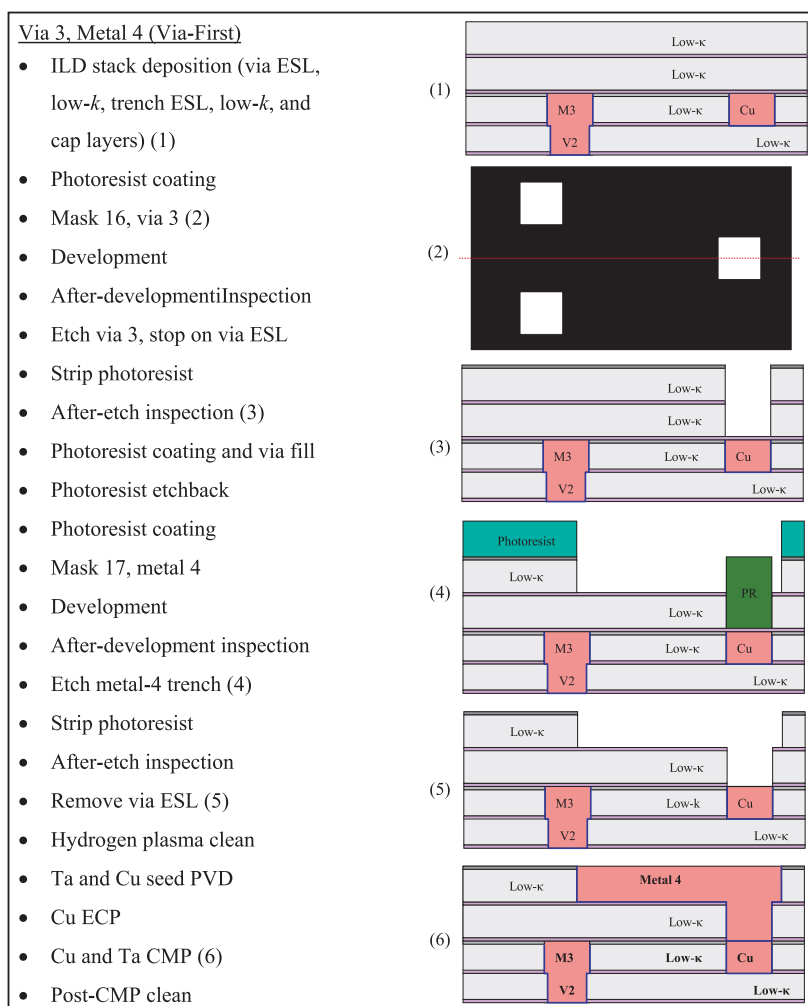


Figure 14.19 Via-first copper interconnection process.

cleaning process removes via ESL at the bottom of via holes [Fig. 14.19, illus. (5)]. After Ta/Cu PVD and Cu ECP, a metal CMP process removes Cu and Ta from the wafer surface and completes the dual damascene copper metallization of this metal layer. Figure 14.20 illustrates via first dual damascene copper metallization process. It is another approach many IC fabs used to form copper interconnection.

Heavy ion bombardment during an argon sputtering etch sometimes can sputter small amounts of copper from the bottom of the hole and deposit it on the sidewall of the via. Copper atoms diffuse very quickly in most dielectrics, such as silicate glass and porous silica. If copper atoms reach the silicon substrate, they can cause unstable performance of the microelectronic devices. Therefore, argon sputtering

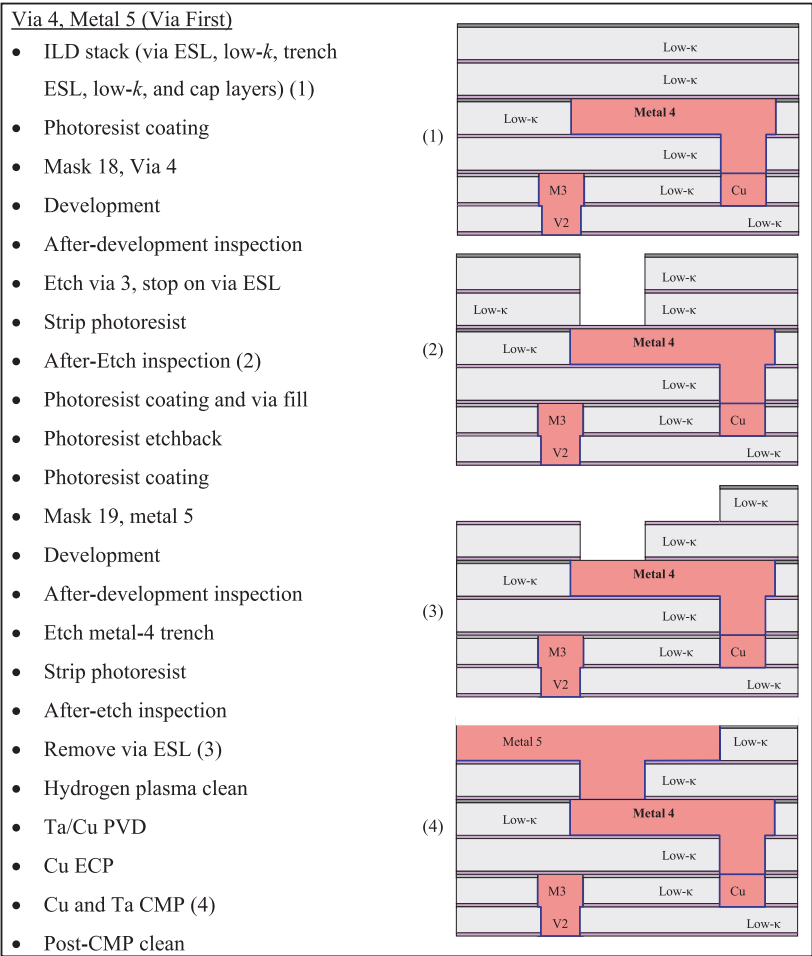


Figure 14.20 Via-first copper interconnection process.

etch prior to metal PVD can cause long-term reliability problems in IC chips. Hydrogen plasma cleaning uses plasma to generate hydrogen free radicals, which react with copper oxide to form copper and water vapor. This effectively removes the native oxide from the copper surface at the bottom of the via without ion-bombardment-induced copper sputtering.

Figure 14.21 shows the final metal layering and passivation processes. Silicon nitride is a very dense material that can prevent moisture and impurities such as sodium from diffusing into the chip and causing device damage. It is commonly used as a passivation dielectric for chips with plastic packaging to protect them from chemical contamination and mechanical damage during die testing, separation, sorting, and packaging. An oxide layer such as PSG is deposited

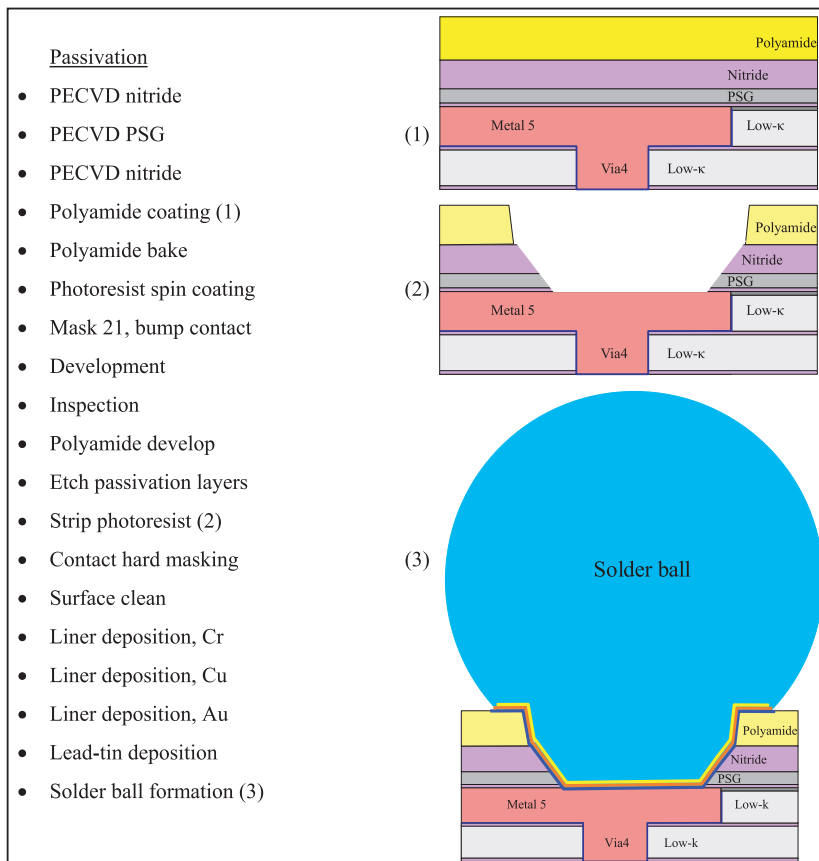


Figure 14.21 Passivation layer deposition and solder bump formation.

before thick passivation nitride deposition to provide a stress buffer. For chips using ceramic packaging (which is a better barrier than plastic packaging but more expensive), a CVD silicon dioxide or silicon oxynitride layer is commonly used as a passivation dielectric. After passivation layer deposition, a layer of polyamide is coated, followed by photoresist coating, baking, and development. Polyamide is etched during photoresist development. A polyamide coating provides protection from mechanical scratches during wafer shipping, and it can also protect microelectronic devices from damage caused by background radiation, such as alpha radiation. After photoresist stripping, the wafer processing is almost finished.

Illustration (3) in Fig. 14.21 shows the bump formation of flip-chip packaging. Bump formation is the last stage of wafer processing, usually performed in a different facility. Because its feature size is very large (about 50 to 100 μm), it does not require a high-grade cleanroom. Chrome, copper, and gold are used as liners

to achieve low-contact resistance. Chrome prevents copper, gold, and lead from diffusing into the substrate and causing heavy metal contamination, and gold helps lead–tin alloys stick to the wafer surface. A metal contact mask is mounted on the wafer surface during metal deposition, allowing only metal in the bump openings. Using masked metal deposition helps to avoid photolithography and etch, therefore reducing production cost. Bumps are formed on the wafer surface after the reflow of a lead–tin alloy. The wafer is now ready for die testing, separation, sorting, and packaging.

Copper interconnection technology is relatively new. After intensive research and development, IC chips with copper interconnections came to production in limited quantities for the first time in 1999. Copper interconnections have been applied to IC chips with minimum feature sizes smaller than 0.18 μm . They have been used in logic IC chip fabrication since the 0.13- μm technology node and have also been applied in memory chip manufacturing. Because the dual damascene process simplified the interconnection process by reducing the number of steps, the cost of copper interconnections became lower than that of traditional tungsten/aluminum-copper alloy interconnections. Copper interconnections have become mainstream in advanced IC chip fabrication.

By combining SOI substrate and copper/low- κ interconnections, fabs can manufacture faster, more powerful, and more reliable IC chips with higher noise resistance and lower power consumption. Figure 14.22 shows a cross section of a CMOS IC with SOI substrate and copper/low- κ dielectric interconnections.

14.5 Complementary Metal-Oxide-Semiconductor Process Flow with Technology after 2010

As devices shrank to 45 nm and beyond, such as 40, 32/28 nm, and 22/20 nm, gate oxide thicknesses reached a point where they could no longer be reduced, due to severe gate-to-substrate leakage. High- κ gate dielectrics have been developed to replace the commonly used silicon dioxide (SiO_2) and nitrided oxide (SiON). To further improve device performance, metal gates have been developed to replace polysilicon gates. Strain engineering has been widely used to enhance the mobility of electrons and holes, and improve device speed. Selective epitaxial growth (SEG) of silicon germanium (SiGe) and carbon-doped silicon (SiC) has been implemented in CMOS manufacturing to achieve desirable channel strain. Self-aligned electroless plating of CoWP has been developed to cap copper surfaces after copper CMP to prevent copper diffusion and reduce electromigration, thus improving the reliability of copper interconnections. Metal (TiN) hard masks are used for low- κ dielectric etching.

193-nm immersion lithography and double-pattern processing are used to pattern line-space features, and source-mask optimization (SMO) techniques are employed to pattern contact and via holes. Design teams must work closely with lithography and processing teams to optimize the design (called design for manufacturing, or DFM) to achieve a manufacturable yield.

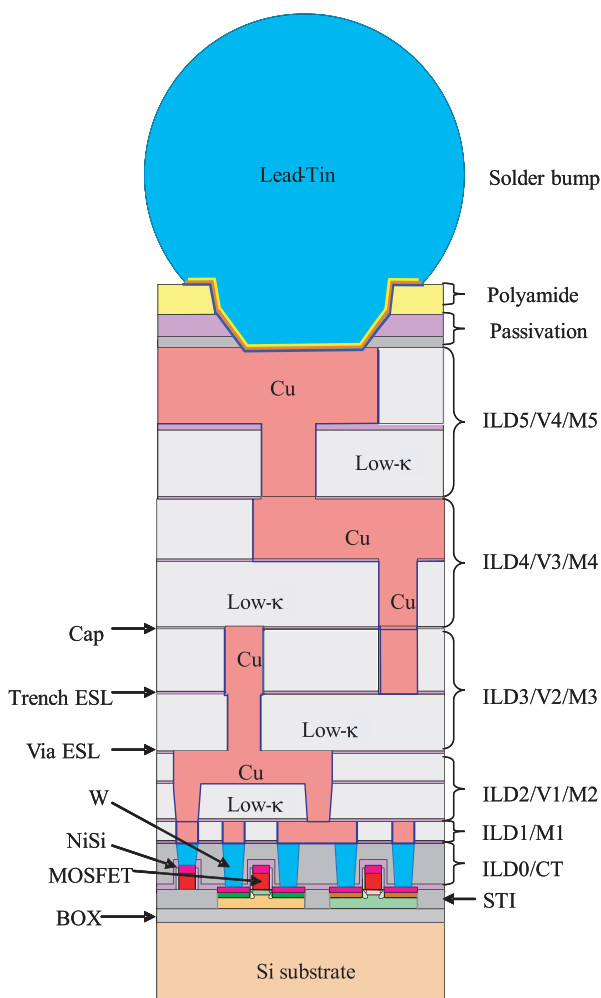


Figure 14.22 CMOS IC with SOI substrate and Cu/low- κ interconnections.

Figure 14.23 illustrates a cross section of a 32/28-nm CMOS with gate-last HKMG, SEG SiGe S/D, stress memory technique (SMT), copper/ultralow- κ interconnections, and lead-free solder bump.

Figure 14.24 illustrates the STI process. It looks almost the same as the STI process illustrated in Fig. 14.8. However, due to the shrinkage of device dimensions, many details are different. For example, the device illustrated in Fig. 14.8 requires stress-free STI fill, while the device in Fig. 14.24 can require extra stress from the STI oxide to help further strain the channel.

Figure 14.25 shows the well and V_T adjustment implantations that form the twin wells required for a CMOS. Due to diminishing device dimensions, the depth of

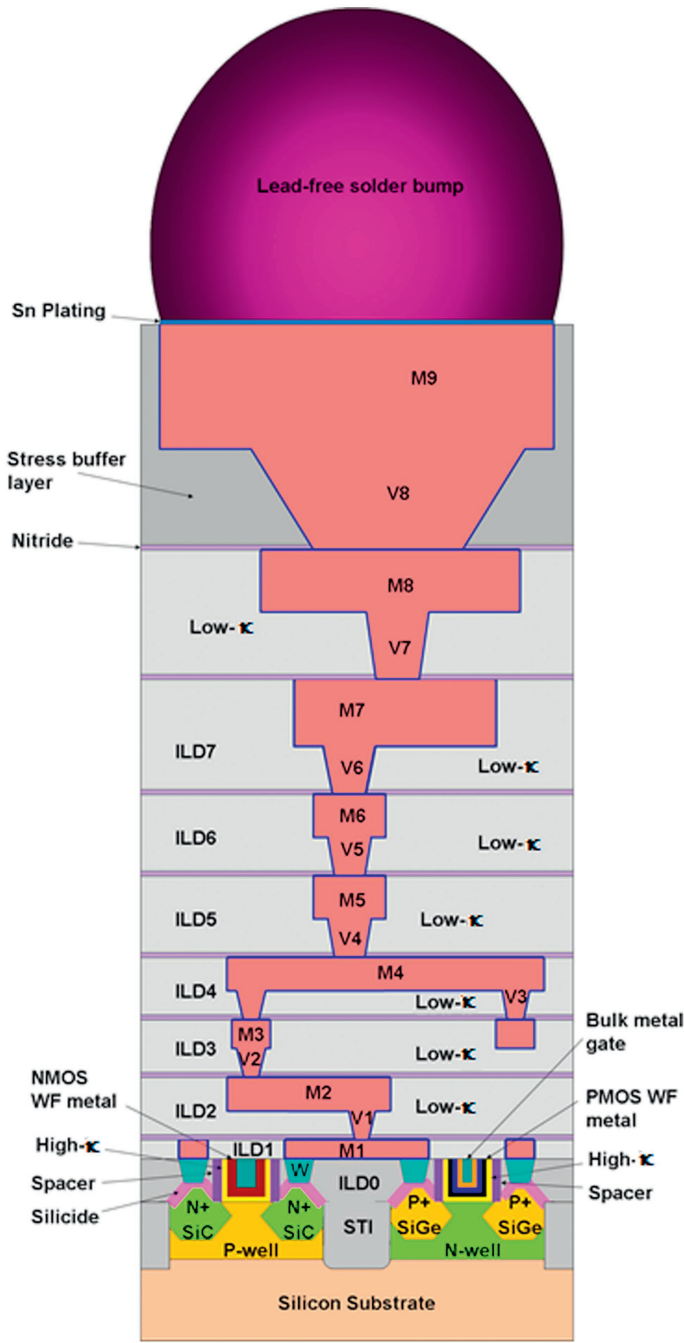


Figure 14.23 Cross section of a CMOS with HBMG, SMT, SEG SiGe S/D, copper/ultralow-κ interconnections, and lead-free solder bump.

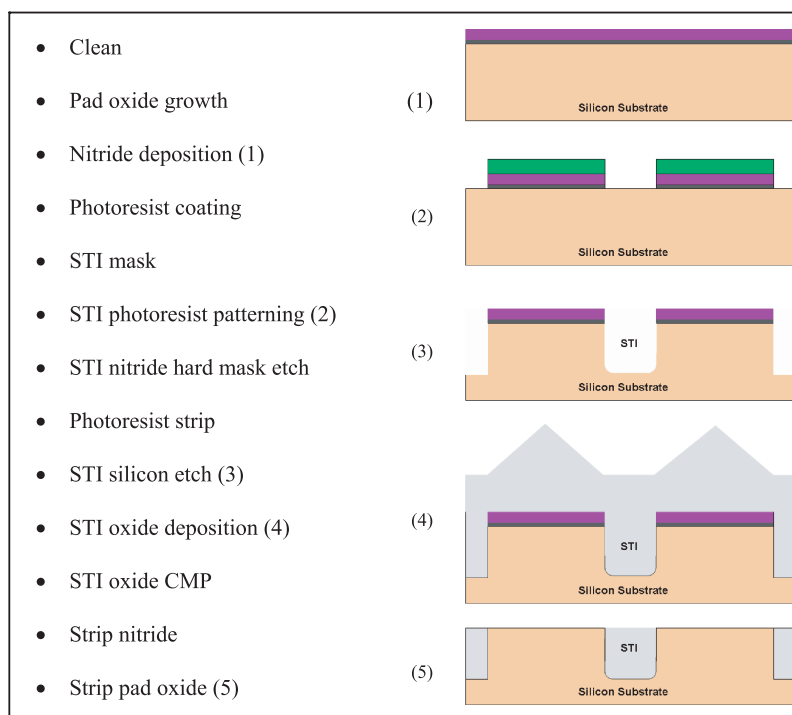


Figure 14.24 STI process.

well junctions becomes shallower compared to the well implantation illustrated in Figure 14.9.

Figure 14.26 illustrates the formation of a gate, nMOS SDE, and sidewall spacer. Because the pMOS S/D will be formed by the SEG of SiGe (which is heavily p-type doped during the SEG process), pMOS does not need SDE or SD implantations. Different from previous processes, this polysilicon gate is just a dummy gate, which will be replaced by high- κ and metal gates later on. It does not require poly-dope implantation, and the gate work function is controlled by different metal materials of the pMOS and nMOS gates.

The major differences of this advanced process are shown in Fig. 14.27. The block layer deposited in Fig. 14.27, illus. (1) is needed so that the SEG can be grown in the designated area. Figure 14.27, illus. (3) shows a KOH etch of silicon, which is highly selective to the $\langle 111 \rangle$ crystal orientation of silicon. By using a KOH etch, the silicon etch profile can be precisely controlled, and optimal device performance can be achieved due to the enhanced channel strain. pMOS S/D doping is accomplished with in-situ doping during the SEG process [Fig. 14.27, illus. (4)]. nMOS S/D doping still uses traditional low-energy, high-current ion implantation and thermal annealing [Fig. 14.27, illus. (5) and (6)]. By performing deep amorphous ion implantation and spike annealing on the nMOS

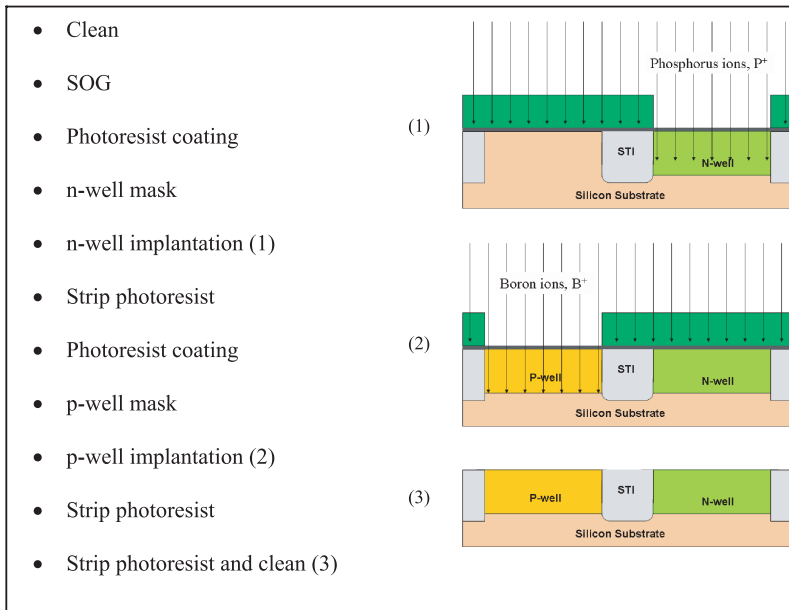


Figure 14.25 Twin-well formation.

S/D, a strong tensile strain on the nMOS channel can be created, increasing electron mobility and improving nMOS speed. Because tensile stress will continue after S/D formation, this technique is a SMT. Due to the limited thermal budget of this tiny device, annealing must be performed with spike or laser methods, or the combination of both.

Figure 14.28 shows a HKMG process with gate-last integration. Figure 14.28, illus. (1) shows ILD0 deposition, which has at least two layers—nitride-based liner/ESL (not shown in the figures) and an oxide layer. Figure 14.28, illus. (2) shows the poly-open CMP, which basically is overpolished ILD0 that exposes the dummy poly gate. Figure 14.28, illus. (3) illustrates the dummy gate removal process, which uses a highly selective etch process to remove polysilicon while having very little effect on ILD0 and sidewall spacers. After HF dip removes the oxide above the channel, a thin layer of silicon oxide is formed, followed by hafnium-based high- κ dielectric deposition [Fig. 14.28, illus. (4)]. HfO_2 is deposited by an atomic layer deposition (ALD) process. Titanium nitride (TiN) and tantalum (Ta) layers are deposited, also with ALD processes. TiN is commonly used for pMOS work-function metal, while nMOS requires a different work-function metal, which is formed in later process. Ta is the barrier layer that protects TiN on the pMOS. Photolithography and etching processes are needed to remove Ta from the nMOS region [Fig. 14.28, illus. (5)]. Titanium–aluminum (TiAl) alloy is deposited, and bulk aluminum (Al) is used to fill the gap. In an annealing process, TiAl reacts with TiN to form TiAlN, which serves as the nMOS work-function

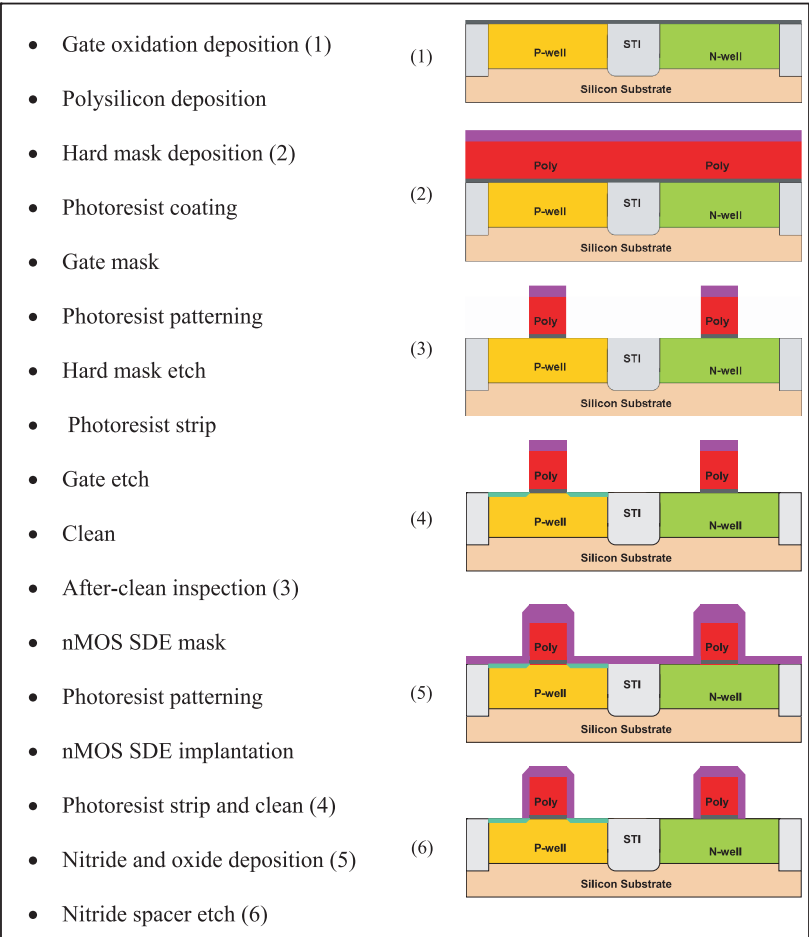


Figure 14.26 Dummy gate patterning, nMOS SDE formation, and sidewall spacer formation.

metal. The metal CMP process shown in Fig. 14.28, illus. (6) finishes the gate-last HKMG process.

While removing the dummy polysilicon gates, both pMOS and nMOS channel strains increase significantly, giving them higher carrier mobility and higher device speed. This is similar to removing the hard stop of a spring, allowing the spring to be further stressed by either compressive or tensile force. This is one advantage of gate-last HKMG integration over the gate-first approach.

Figure 14.29 shows a middle end of the line (MEoL) process. There are several differences in this contact module. One of them is the trench silicide, which is formed after trench contact etching and is not the normal salicide that was formed after S/D formation (before ILD0 deposition). The trench silicide is only formed

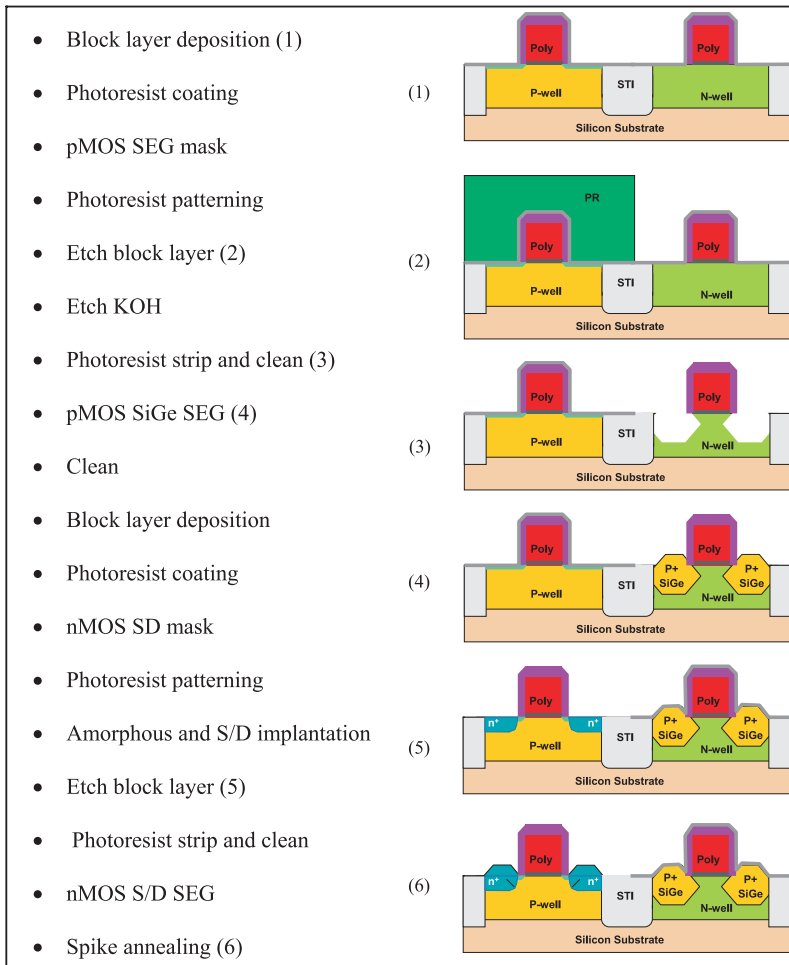


Figure 14.27 SEG processes that form the S/D.

at the bottom of a contact trench [Fig. 14.29, illus. (3)]. Because the gate is formed with metal, it does not need silicide. A small amount of plutonium (Pt) is alloyed with nickel; its silicide, NiPtSi, is more stable than NiSi. Another difference is that the tungsten (used to fill the contact trenches) is polished to the same level as the metal gate [Fig. 14.29, illus. (5)]. Because the contact trenches only connect to the raised S/Ds, there is only one depth of contact trench, simplifying overetch control. From a layout point of view, the trench style contact replaces the round and oval contact holes, simplifying the photoresist patterning process. However, this replacement can cause tungsten spiking due to overetching into the STI oxide during contact etch. Because the length of the tungsten plug is significantly shorter, resistance of the plug is also reduced.

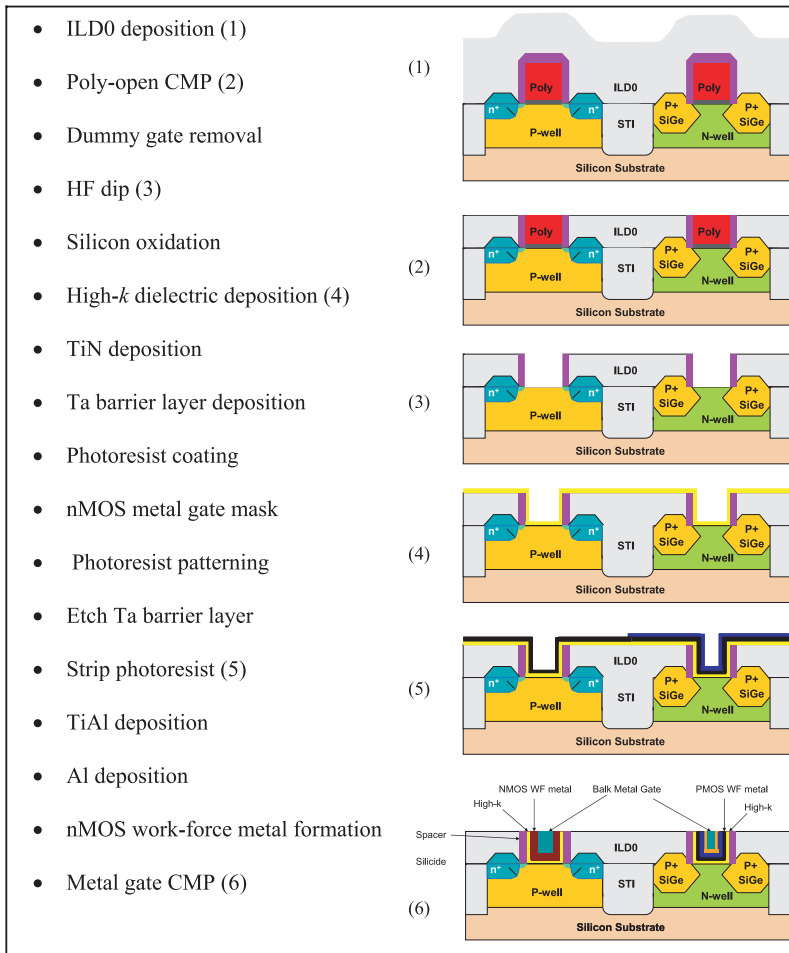


Figure 14.28 Gate-last HKMG process.

The metal-1 (M1) process is illustrated in Fig. 14.30. A hard mask (usually TiN on a TEOS cap) is used to protect the porous low- κ dielectric from damage caused by the photoresist stripping process. The porous low- κ dielectric has a lower κ -value (~ 2.2 to 2.5) than the κ -value of carbon-doped oxide or carbonated silicate glass (CSG, $\kappa \sim 2.7$ to 2.9). A porous low- κ dielectric can be formed by PECVD carbon-doped silicon oxide film with <2 -nm pores and up to 40% porosities introduced by adding porogens into the gas flow. CVD precursors can be trimethylsilanes or tetramethylsilanes, and porogen species can be norbornene or α -terpinene.

Ta barrier and copper seed layers are deposited in the PVD process with an ionized metal plasma process. Bulk copper is deposited with the ECP process.

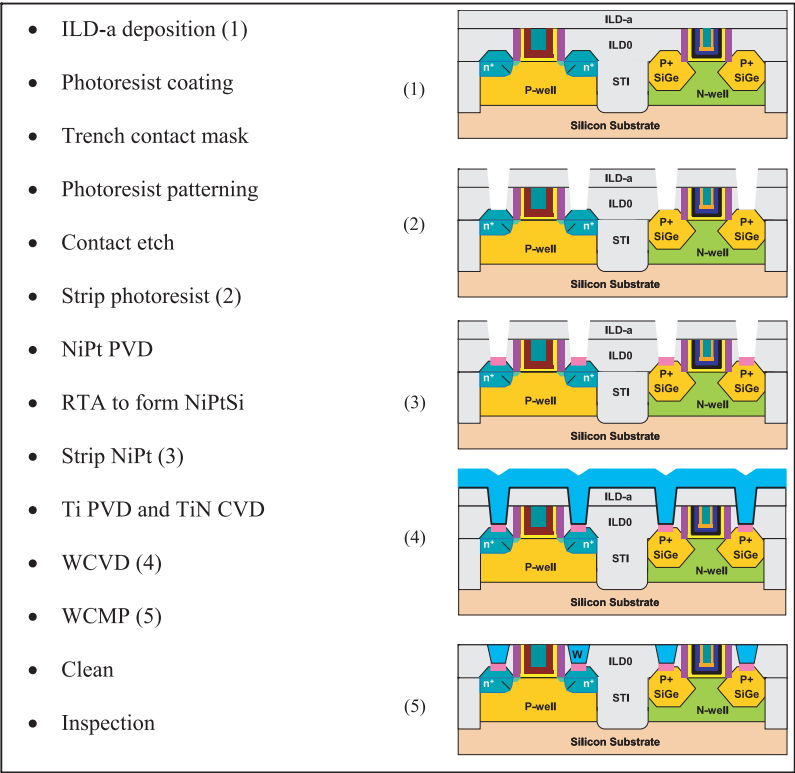


Figure 14.29 MEOL (contact module) processes: (1) sacrificial ILD-a deposition, (2) trench contact etch, (3) trench silicide formation, (4) WCVD, and (5) WCMP.

After copper annealing, the metal CMP process removes the bulk Cu layer, Ta barrier layer, and TiN hard mask. The CMP stops in the TEOS cap layer, which protects the porous low- κ dielectric from contact and damage by CMP slurry.

A dual damascene copper/low- κ dielectric interconnection process with trench-first integration is illustrated in Fig. 14.31. A metal hard mask (usually TiN) and TEOS PECVD oxide, or TEOS, cap layer are used to protect the porous low- κ dielectric from damage from photoresist stripping and metal CMP processes. The self-aligned electroless plating of CoWP can be used to prevent copper diffusion and to enhance EMR, improving reliability of the IC chip.

Figure 14.32 shows the copper/low- κ dielectric interconnections of the metal layers from M3 to M9. They are basically multiple repetitions of the trench-first process described in detail in Fig. 14.31.

Lead (Pb) was widely used for solder bumps. It is well known that lead is a contaminant that can affect the development of the heart, bones, intestines, kidneys, and nervous system, and as such it is particularly toxic to children. Every year several tons of used IC chips from outdated electronics go into landfills

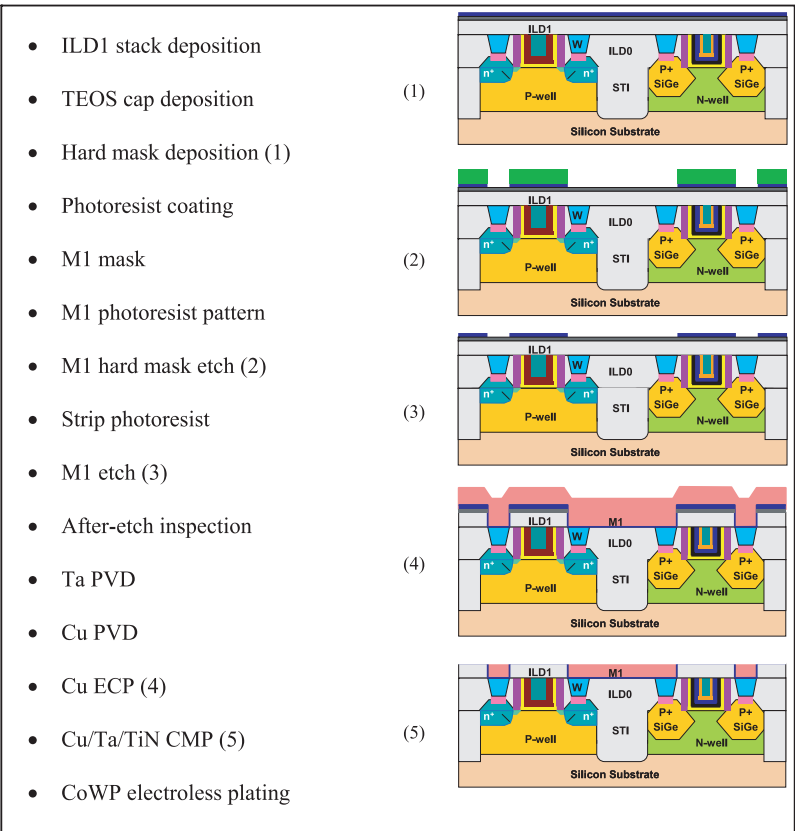


Figure 14.30 M1 process.

and create potential risks for lead contamination in the environment. Therefore, many countries such as Japan, Europe, and China have legislation or regulations restricting or eliminating the use of lead in the semiconductor industry and all electronics industries. Figure 14.33 shows a lead-free solder bump.

14.6 Memory Chip Manufacturing Processes

Memory chips play important roles in driving IC technology and its markets. Two major memory products on the market are DRAM and NAND flash. In a computer, whether it is a desk top PC, laptop, or tablet, generated data are always stored in DRAM first before being written in a nonvolatile memory (NMV) device such as a hard disk drive (HDD) or a solid-state drive (SSD) for permanent storage. The amount of memory bits in a desktop or laptop PC has multiplied many times in the last few years. In 1993, a typical PC, say a desktop 486, had 4-Mb DRAM, which could be upgraded to 8 Mb for \$100. It cost about \$30 to purchase 4-Gb DRAM for a desktop PC in 2009. With increasing demands

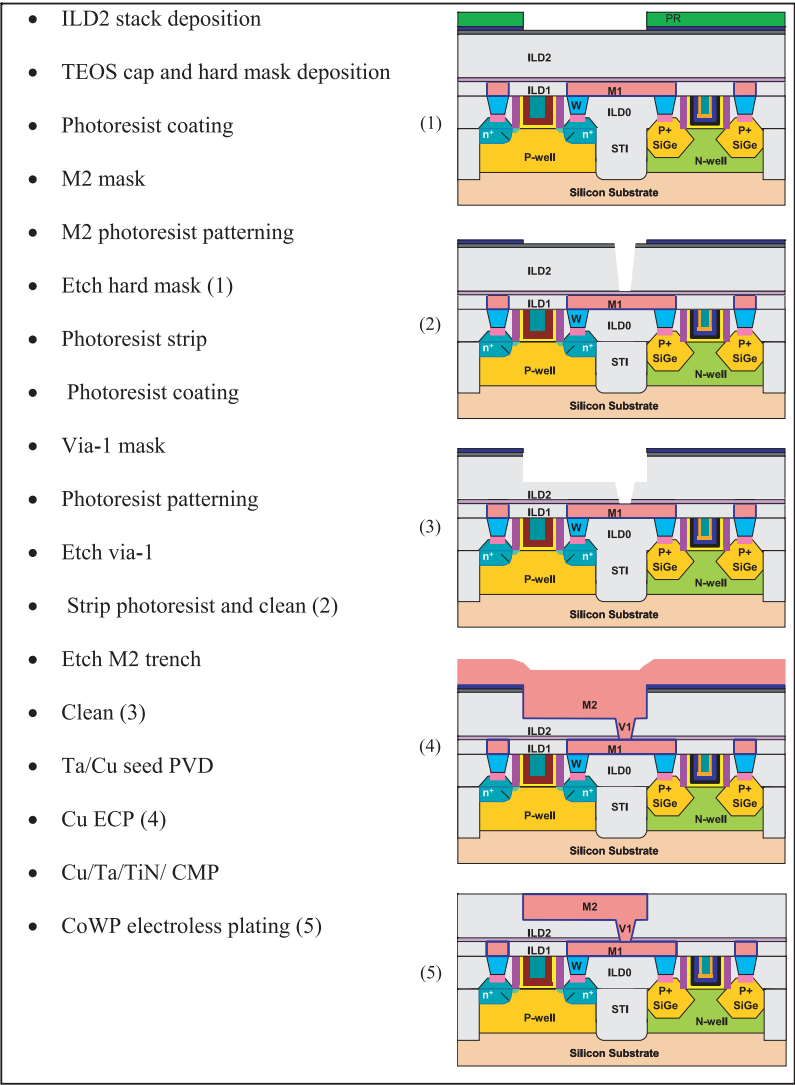


Figure 14.31 Trench-first copper/low- κ dielectric interconnection of via 1 metal 2 (V1M2).

for more graphics, especially 3D graphics, the required amount of DRAM for a computer will further increase and will continue to drive DRAM manufacturing technologies.

Different from DRAM (which requires a constant power supply to keep the stored data), a NAND flash is a nonvolatile memory that can keep stored data without a power supply for years. NAND flash is widely used for data storage in mobile digital electronic products, such as MP3 players, digital cameras, cell

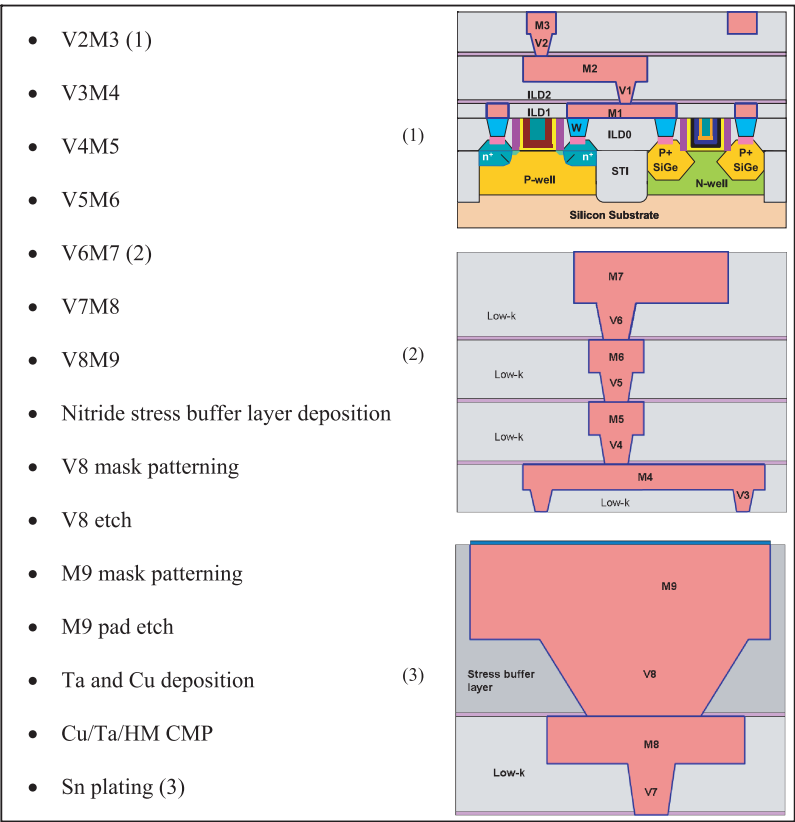


Figure 14.32 Copper/low-κ dielectric interconnections.

phones, and high-end laptop PCs. With more graphic and video applications in mobile electronic devices, the demands for NAND flash memory will further increase. NAND flash is also used in hybrid disk drives to combine the benefits of fast data access of a SSD and the low cost of a HDD.

14.6.1 Dynamic random access memory process flow

DRAM played a very important role in driving IC markets and technologies. A DRAM cell consists of an nMOS and a storage capacitor, as shown in Fig. 14.34.

There are two types of DRAM processes: one is a stacked DRAM, which stacks storage capacitors on top of select transistors (nMOS); the other is deep trench DRAM, which builds storage capacitors in a deep trench on the silicon surface alongside the nMOS. Stacked DRAM is similar to creating storage by building an attic, and deep trench DRAM is similar to creating storage by building a cellar.

Figure 14.35(a) illustrates a stacked DRAM and includes the self-aligned contact (SAC), bitline contact (BLC), wordline (WL), bitline (BL), storage node (SN, which is the storage capacitor), and storage node contact (SNC).



Figure 14.33 Lead-free solder bump formation.

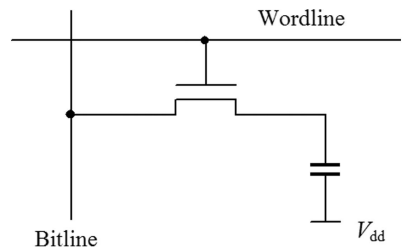


Figure 14.34 Circuit of a DRAM memory cell.

Figure 14.35(b) illustrates a deep trench DRAM. Because the trench capacitor is very deep with an aspect ratio more than 50, the depth of the drawing was reduced to fit the figure onto the page. Deep trench DRAM has fewer requirements for the metal interconnections above the silicon surface, making it more compatible with normal CMOS backend processes, and a better candidate for embedded DRAM for system-on-chip (SOC) applications. However, because it needs precious silicon real estate to build its storage capacitors, its packing density is not as scalable as the stacked DRAM, which does not require any of the silicon surface to build its storage capacitors. General-purpose DRAM chips are commodity products and thus are very sensitive to cost. Because stacked DRAM has a small advantage

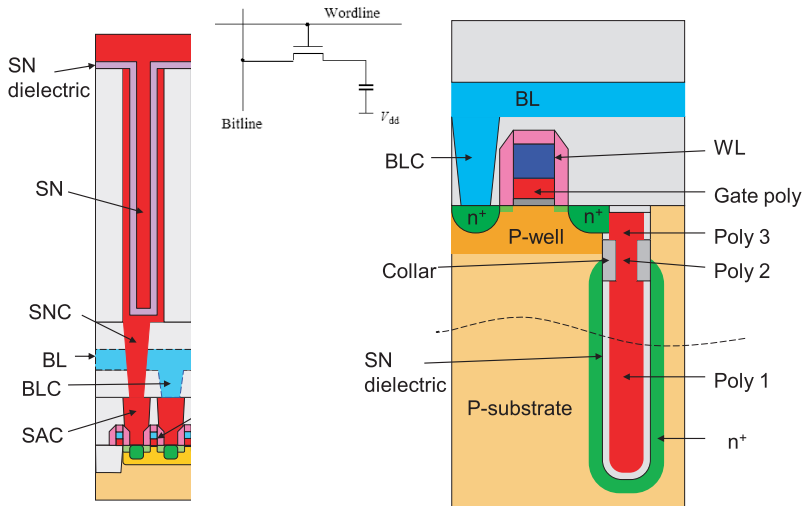


Figure 14.35 (a) Stacked DRAM and (b) deep trench DRAM.

in cost over deep trench DRAM, it dominates the DRAM market. Only stacked DRAM processes are described in this chapter.

14.6.2 Stacked dynamic random access memory process flow

Most DRAM chips in computers and other digital electronic devices are stacked DRAM chips. Figure 14.36 shows a cross section of a stacked DRAM chip. The left side of Fig. 14.36 shows a cross section of an array area with four unit memory cells. A 2-Gb DRAM made with 30-nm technology has two billion such unit cells. The periphery circuits are logic devices to control write, read, and input/output operations of the DRAM chip. Devices in the periphery usually are larger than those in the cell array area. Processes of the periphery devices are very similar to the CMOS processing technologies described earlier. This section focuses on the process flow in the cell area.

Figure 14.37 illustrates STI and well formations of a stacked DRAM cell. Figure 14.37(a) shows the layout of the AA, and the dashed line indicates the location of the cross section. Figure 14.37(b) illustrates the cross section after AA etching, Fig. 14.37(c) shows the cross section after STI formation, and Fig. 14.37(d) shows the cross section after p-well formation. The STI and p-well formation processes are simultaneously performed in the periphery area. It is important to note that there is a p-well masking step for p-well formation. It has finer patterns in the periphery area, and it is blank in the cell area. n-well formation is only performed in the periphery area, not in the cell area, because there is only nMOS in a DRAM cell.

Figure 14.38 illustrates WL, LDD, sidewall spacer, and S/D formations of a stacked DRAM cell. Figure 14.38(a) shows the layout of the WL that overlaps

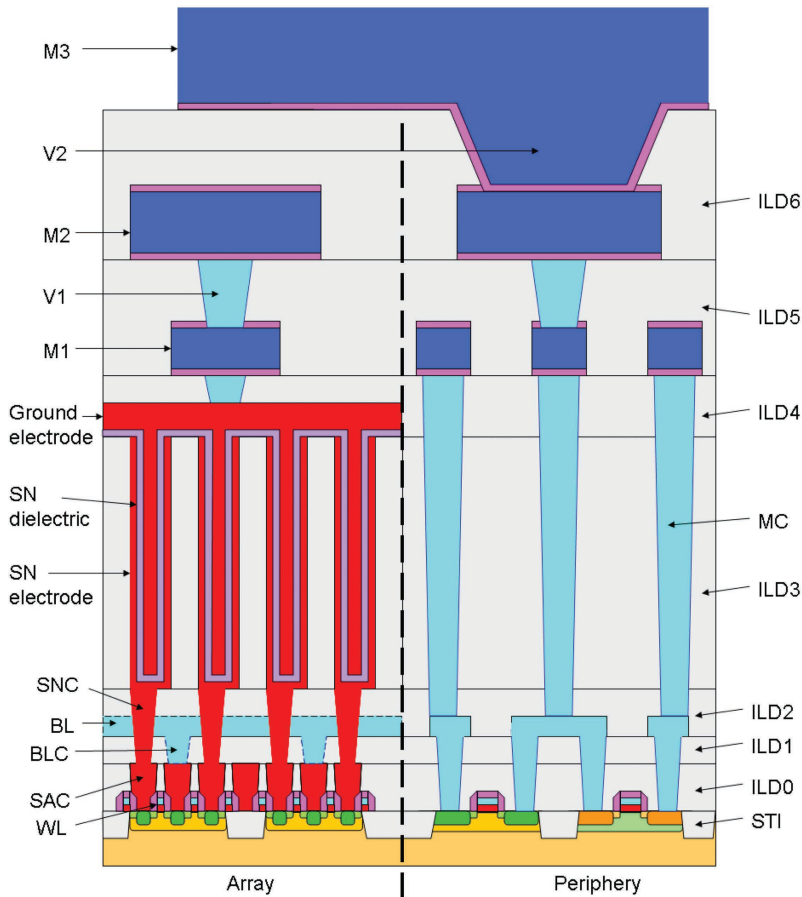


Figure 14.36 Cross section of a stacked DRAM.

with the AA layer, and the dashed line indicates the location of the cross section. Figure 14.38(b) shows the cross-sectional view of the formation of the DRAM cell nMOS gate, which is also the WL. Figure 14.38(c) shows the lightly doped drain (LDD) formation, Fig. 14.38(d) illustrates the sidewall spacer formation, and Fig. 14.38(e) shows S/D formation. There are two masking steps not shown in Fig. 14.38, the pMOS LDD and pMOS S/D formations in the periphery area. These are shown on the right side of Fig. 14.39. Cobalt silicide is also formed in periphery devices to reduce contact resistance.

Figure 14.40(a) shows the first contact, the SAC, of a stacked DRAM. This module is also called the landing pad contact (LPC) after etch and landing pad poly (LPP) after the polysilicon CMP. Both unopened contact holes and shorts to the WL are killer defects, and it is very challenging to etch contact holes through ILD0, which is BPSG in between dense WLs that must reach every S/D of the cell

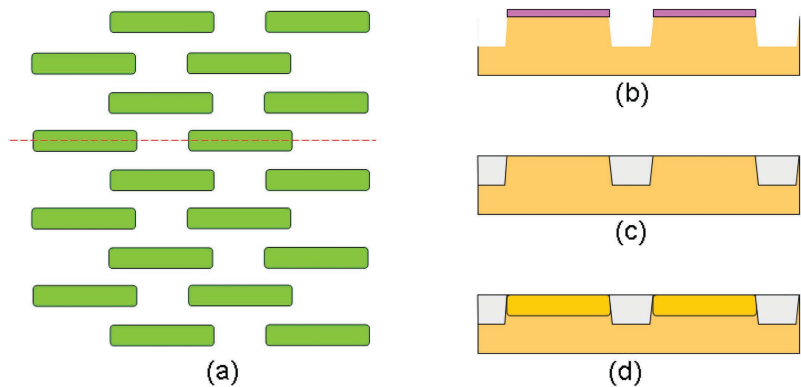


Figure 14.37 (a) Stacked DRAM AA layer layout, (b) cross section along the dashed line after AA etch, (c) after STI formation, and (d) after p-well formation.

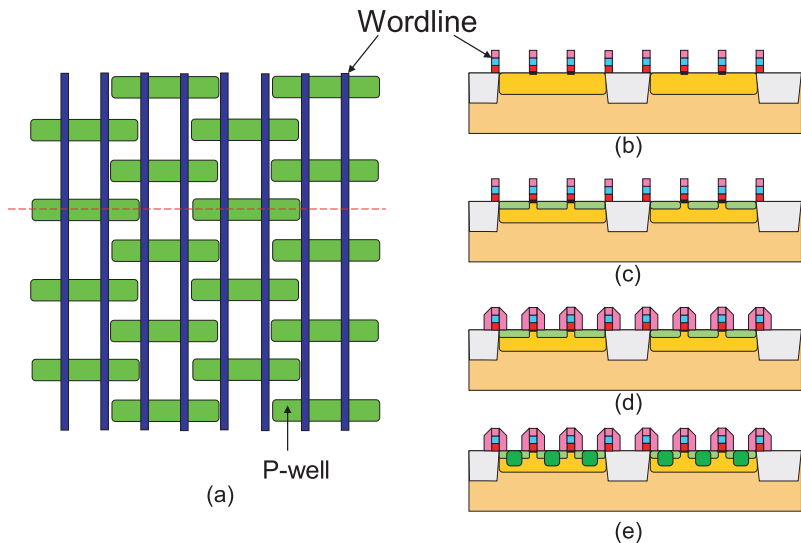


Figure 14.38 (a) Stacked DRAM WL and AA layer layouts, (b) cross section after WL etch, (c) after LDD formation, (d) after spacer formation, and (e) after S/D formation.

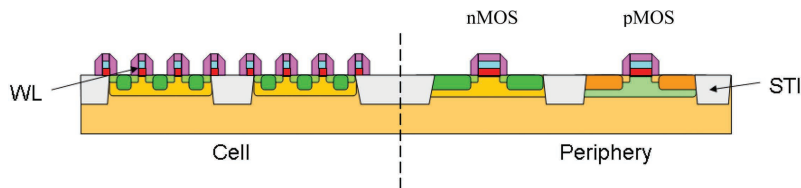


Figure 14.39 Cross section of DRAM cell and periphery devices.

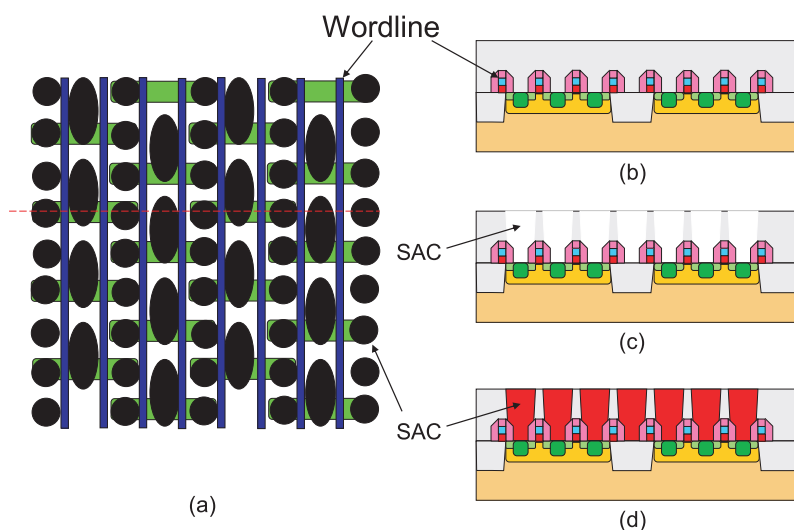


Figure 14.40 (a) Stacked DRAM SAC layout overlapped with WL and AA layers, (b) cross section after ILD0 CMP, (c) after SAC etch, and (d) after SAC poly CMP.

nMOS. Therefore, self-aligned contact processes were developed. By maintaining a nitride hard mask on top of the WLs, and with nitride spacers on both sides, WLs are surrounded with nitride. When the SAC etching process has high enough selectivity between BPSG and nitride, the etching process becomes self-aligned, allowing the contact holes to go through the dense WLs and reach the silicon surface without electrically shorting them.

Before polysilicon is deposited to fill the SAC holes, contact ion implantation (which is usually high-dose n-type ion implantation) is performed to reduce contact resistance. E-beam inspections are applied in this module, either after etch or after polysilicon CMP, to ensure the capture of defects such as unopened contacts and contact plug-to-WL shorts at the earliest stages. The SAC process only happens in the array area.

Figure 14.41 shows the BLC module of a stacked DRAM. Figure 14.41(a) shows that BLCs are landed in the SAC plugs that connect to the middle sections of the AA. Every BLC is shared by two DRAM cells. Figure 14.41(b) shows a cross section after ILD1 deposition and CMP, and Fig. 14.41(c) is the cross-sectional view after BLC etch. ILD1 is usually BPSG for a stacked DRAM.

BLC in the periphery area can be patterned with the BLC in the cell array area. Because the feature size and depth of the BLCs in the array and periphery areas are quite different, most fabs separate these two contact processes.

Figure 14.42(a) shows the BL module of a stacked DRAM. The BLs are connected to the middle sections of the array AA through BLC plugs that are landed in the SAC plugs. Tungsten is the most commonly used metal for BL formation. After Ti/TiN barrier/glue layer deposition, tungsten is deposited in a CVD process to fill BLC holes and form a thin film on the surface of the wafer.

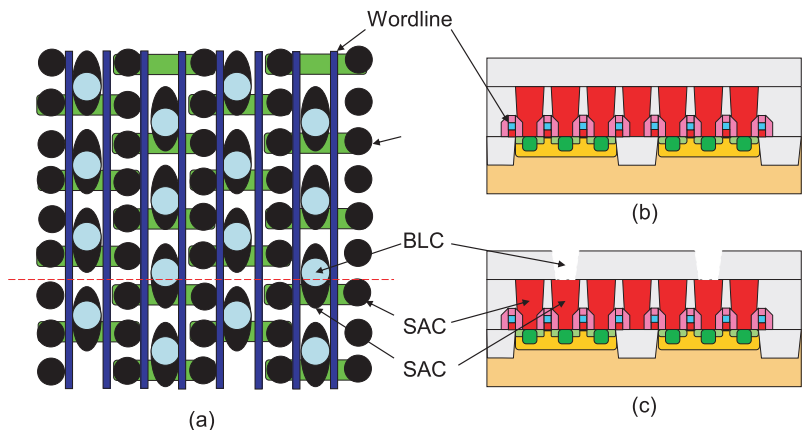


Figure 14.41 (a) BLC layout overlapped with SAC, WL, and AA layers; (b) cross section view after ILD1 deposition; and (c) after BLC etch.

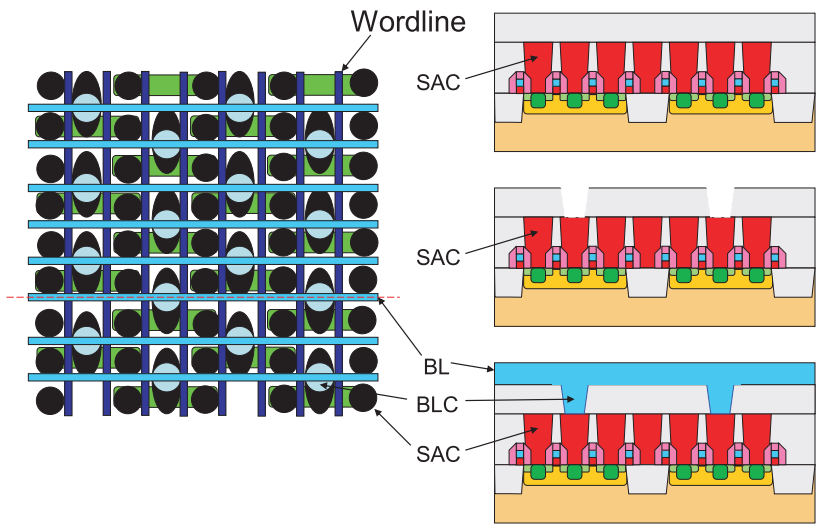


Figure 14.42 (a) BL layout overlapped with BLC, SAC, WL, and AA layers, (b) cross section view after BL metal deposition, and (c) after BL etch.

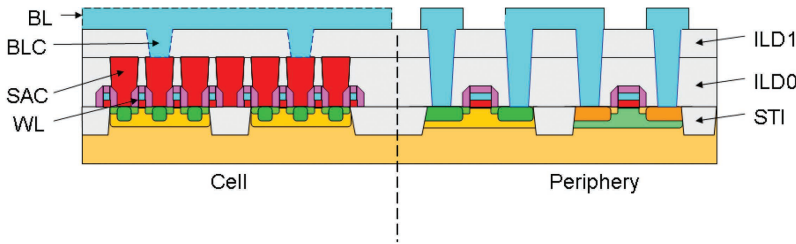


Figure 14.43 Cross section view of a stacked DRAM in both cell and periphery areas after BL formation.

A BL mask defines the metal wires of the BL in both cell and periphery areas, and a metal etching process forms the BL pattern. Figure 14.43 illustrates a cross-section of both cell and periphery areas after BL and BLC formation. To prevent the contact shorting to BL, a spacer is formed on the sidewall of the BL.

Figure 14.44(a) shows the SNC module of a stacked DRAM. The SNC holes connect to the side sections of the array AA through SAC plugs. Conducting plugs of SNC can be formed with polysilicon or tungsten, depending on the technology node. Before conductor deposition, a thin layer of silicon nitride is deposited and etched back to form a liner on the sidewall of the SNC holes to prevent the conducting plugs from shorting the BLs.

Figure 14.45(a) shows the SN layout, and Figure 14.45(b) shows the cross section of SN holes along the dashed line in Figure 14.45(a). The SN holes are connected to the SNC plugs, which are landed on the SAC plugs connected to the two side sections of the array AA.

To form the storage capacitors, two conducting layers are needed to form the two electrodes and a dielectric layer, which is sandwiched between them. Figure 14.46 shows the SN capacitor formation process. After SN hole etch and cleaning, a conductor layer such as polysilicon or titanium nitride (TiN) is deposited [Fig. 14.46(a)]. Because the aspect ratio of the SN hole is very high, this conducting layer needs to have very good sidewall and bottom step coverage. Photoresist fills in the SN holes after electrode layer deposition to protect the conducting film in the holes, and to allow an etchback process that removes the conducting film on the surface [Fig. 14.46(b)]. The SN electrodes are connected to the SNC plugs, which are connected to the two side sections of the AA via the SAC plugs. After removal of the photoresist from the SN holes, a layer of dielectric is deposited on the surface and into the SN holes [Fig. 14.46(c)]. Uniform and conformal sidewall and bottom step coverage is required for this capacitor dielectric film. Figure 14.46(d) shows the conductor deposition that forms the ground electrode of the SN capacitor. This conducting film will be removed in the periphery area in the next photomask step, which finishes the device portion of the DRAM process and starts the BEoL interconnect processes.

With the diminishing technology nodes, SN hole dimensions are becoming smaller. To keep ~ 30 -pF capacitance of the SN capacitor so it can reliably hold enough charge to maintain the memory, the aspect ratio of the SN hole will further

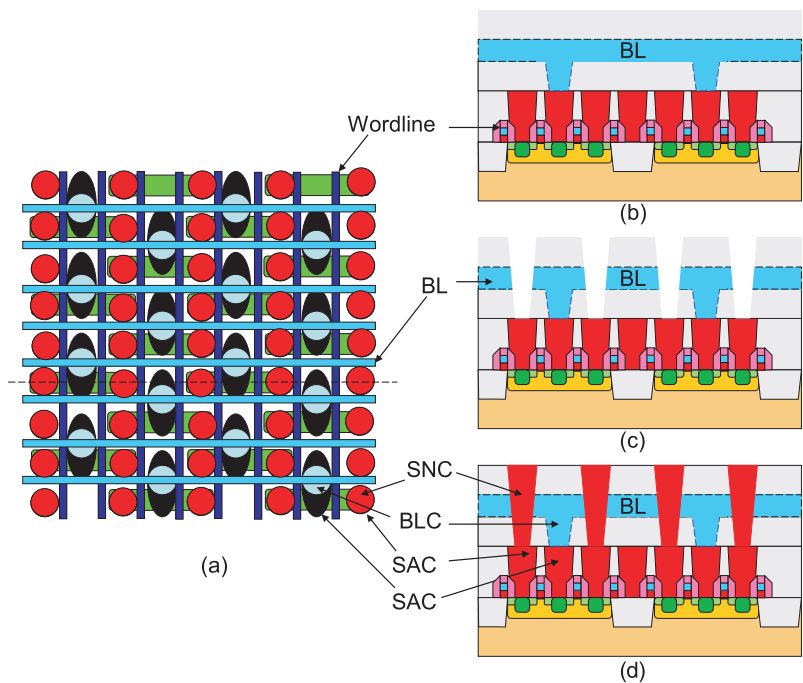


Figure 14.44 (a) SNC layout overlapped with BL, BLC, SAC, WL, and AA layers. (b) The cross section along the dashed line in (a) after ILD2 CMP, (c) after SNC etch, and (d) after SNC polysilicon CMP. The BL with the dashed outline indicates that it is behind the cross section, as shown in (a).

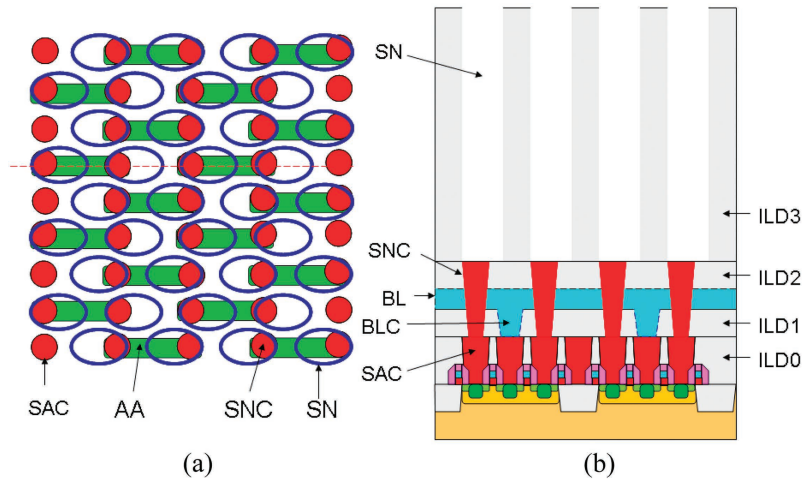


Figure 14.45 (a) SN layout overlapped with SNC and AA layers. (b) Cross section along the dashed line in (a) after SN etch.

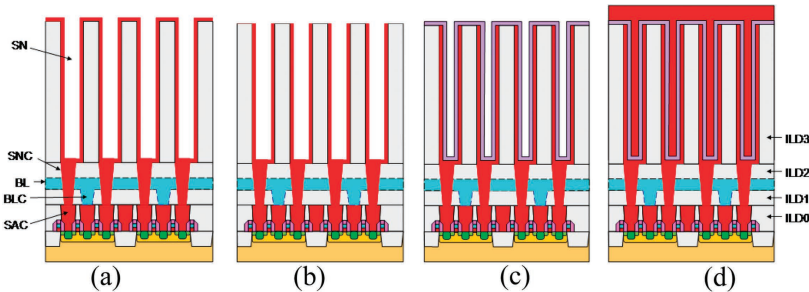


Figure 14.46 SN capacitor formation process: (a) SN electrode conductor deposition, (b) removal of the SN electrode conductor film from the surface, (c) SN dielectric deposition, and (d) ground electrode conductor deposition.

increase if the capacitor structure and capacitor dielectric materials are kept the same. Several methods have been developed to reduce the aspect ratio of SN holes. Using high- κ dielectrics can help to reduce the height of the hole and the aspect ratio. As before, silicon oxide, silicon nitride, and silicon oxide stacking have been used. High- κ materials such as aluminum oxide (Al_2O_3), hafnium dioxide (HfO_2), and zirconium dioxide (ZrO_2) have been studied and applied to SN capacitors. Other methods to reduce the height of SN holes have recessed the ILD3 after SN electrode formation; this allows the ground electrode to form at both sides of the SN electrode. Also, polysilicon was widely used early on as an electrode material. More advanced DRAM chips started to use TiN as SN electrodes. Figure 14.47 illustrates a stacked DRAM with a recessed gate (RG) array transistor, TiN SN electrodes, recessed ILD3, high- κ capacitor dielectric, and metal ground electrodes with three layers of metal interconnects. RGs are used to reduce the short channel effect (SCE) of the array nMOS that becomes stronger with smaller feature sizes.

14.6.3 NAND flash process

Flash chips are NVM chips widely used for data storage applications in electronic devices, especially mobile electronics such as digital cameras, MP3 players, cell phones, global positioning systems (GPSs), high-end notebook computers, tablet computers, etc. Compared with HDDs, flash memory has shorter data access time, consumes less power, and is more reliable because it does not have any moving parts.

Almost all flash memory chips on the market are charge trap devices based on the floating-gate structure described in Chapter 3. The basic device structure is illustrated in Fig. 14.48 and is very similar to an nMOS. Depending on the circuit architecture, there are two types of flash devices, NAND and NOR (see Fig. 14.49).

Figures 14.49(a) and 14.49(b) show the NOR flash circuit and its cross section, respectively. Figures 14.49(c) and 14.49(d) show a 64-bit string NAND flash circuit and the correspondence cross sections, respectively. A NOR flash is equivalent to a 1-bit string NAND flash that does not require select gates. NOR flash can achieve

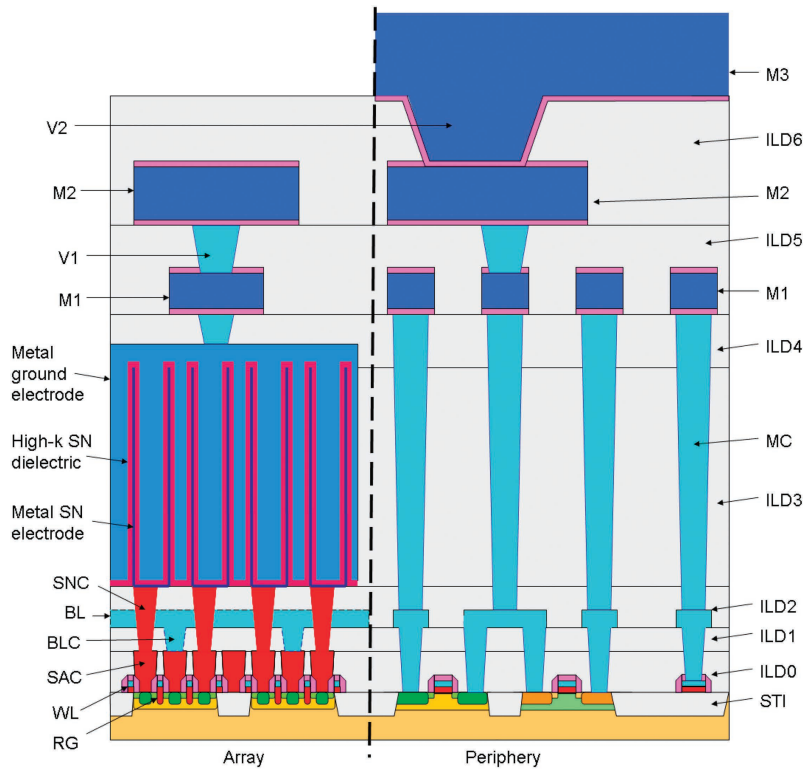


Figure 14.47 Cross-section view of a stacked DRAM with RG array transistor, TiN SN electrodes, recessed ILD3, high-κ capacitor dielectric, and metal ground electrodes.

Flash Device

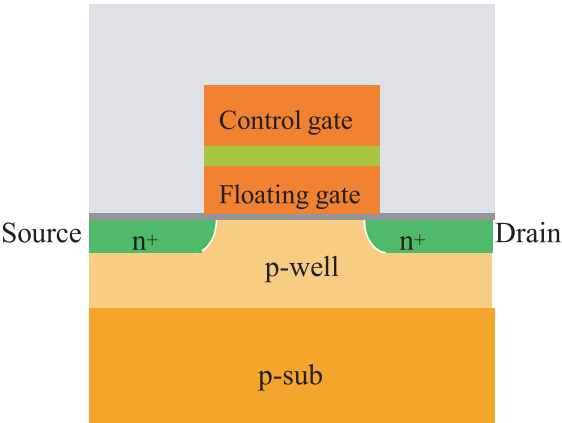


Figure 14.48 Basic structure of a floating-gate NVM device.

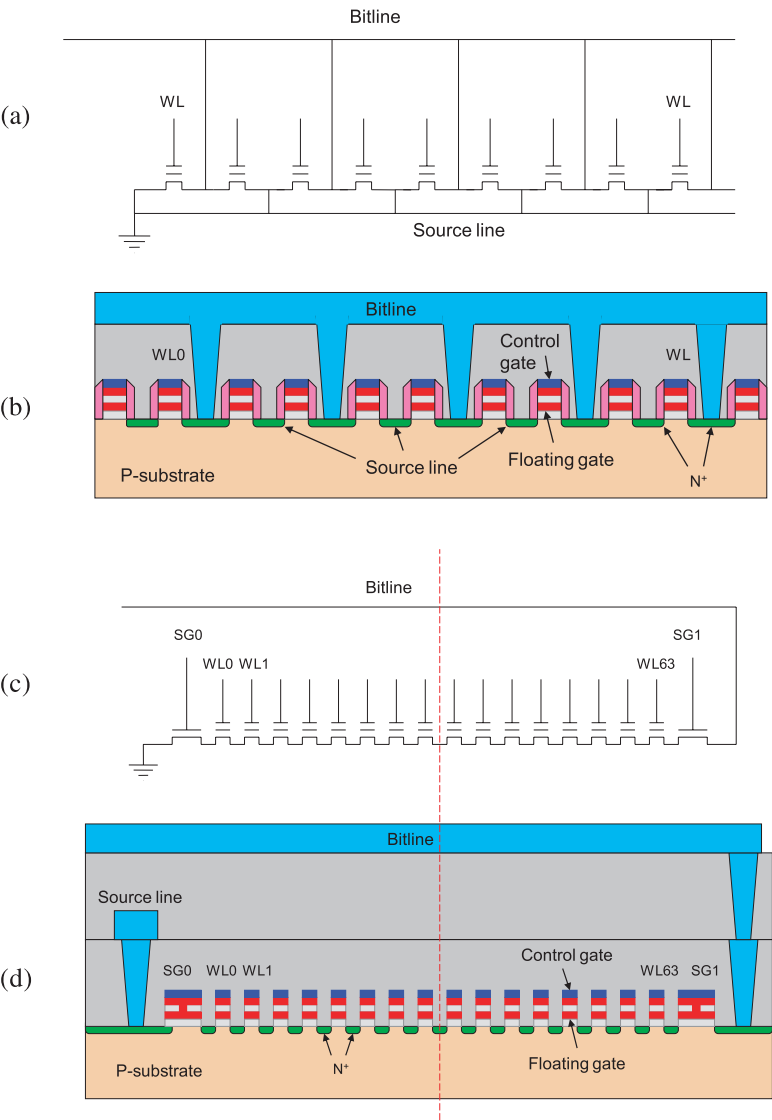


Figure 14.49 (a) NOR flash circuit, (b) NOR flash cross section, (c) NAND flash circuit, and (d) NAND flash cross section.

random access of any memory cell, while NAND flash cannot. Although NOR flash has a shorter reading time than NAND flash, it has much longer write and erase times. It also costs significantly more than NAND flash because of its lower packing density due to the dense BL contacts. Since the majority of flash memory devices are NAND flash chips, only NAND flash processes are discussed in this section.

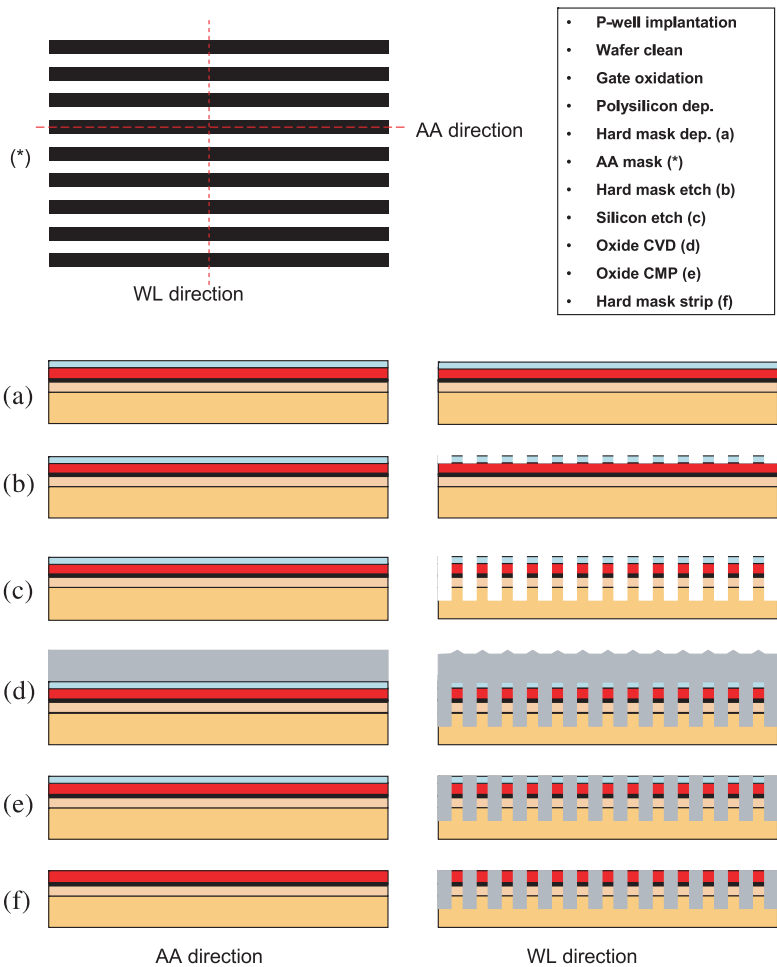


Figure 14.50 AA module of the NAND flash. The top is a schematic of an AA mask, and the left and right sides of illus. (1) through (6) are the cross sections of the SA-STI processing steps (described in the textbox) in AA and WL directions, respectively.

Figure 14.50 illustrates the self-aligned shallow trench isolation (SA-STI) processing module. After p-well implantation, gate oxide is grown and deposits a floating-gate layer with a hard mask on top. An AA mask is then used to pattern a hard mask for etching the floating gate, gate oxide, and silicon substrate, to form active areas. Silicon nitride or silicon oxynitride is the most commonly used hard mask material, and polysilicon is the most commonly used floating-gate material. After the silicon trench etch, high plasma density (HPD) CVD oxide is used to fill the trenches, and a CMP process removes the oxide and stops on the hard mask. A stripping process removes the hard mask and finishes the SA-STI process.

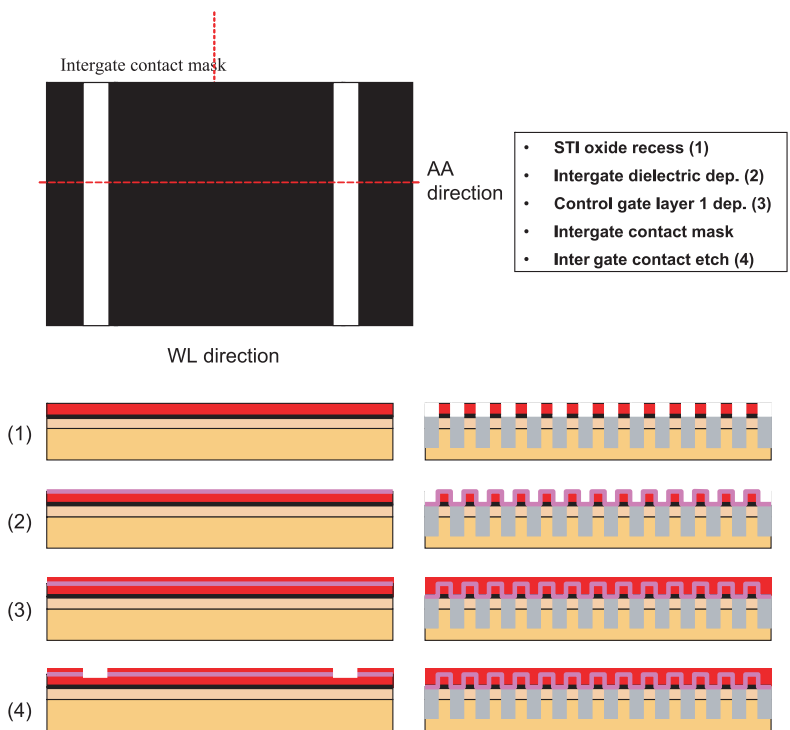


Figure 14.51 Interigate contact module. At the top is the intergate mask in the array area. The left and right sides of illus. (1) through (4) are cross sections of the intergate contact steps of the WL process (described in the textbox) in AA and WL directions, respectively.

Figure 14.51 illustrates the intergate contact process, which is a unique processing module of floating-gate NVM devices. The MOSFETs of selected gates in the array and control circuits in the periphery regions require intergate contact to electrically short the floating-gate layer and control the gate layer. Floating gates are made with polysilicon, the intergate dielectric is stacked oxide-nitride-oxide (ONO), the first layer of the control gate is also polysilicon, and the second control gate layer is usually metal, such as tungsten silicide or tungsten. The recessed STI oxide shown in Fig. 14.51, illus. (1) allows an increase in coupling between the control and floating gates, an increase that is necessary when feature size shrinks.

Figure 14.52 illustrates the WL process module. Metal deposition through intergate contact shorts the floating and control gates of the selected gates in the array area and all MOSFET gates in the periphery area. Line/space patterning of the WL layer has the highest pattern density in IC production. The square box inside the WL mask (shown in the top image of Fig. 14.52) is the area of a unit cell of NAND flash memory, which is $4F^2$, where F is the minimum feature size of the structure, and $4F^2$ is the highest pattern density it can achieve. For NAND flash, the line/space ratios of AA and WL are both 1:1; therefore, F is the CD of

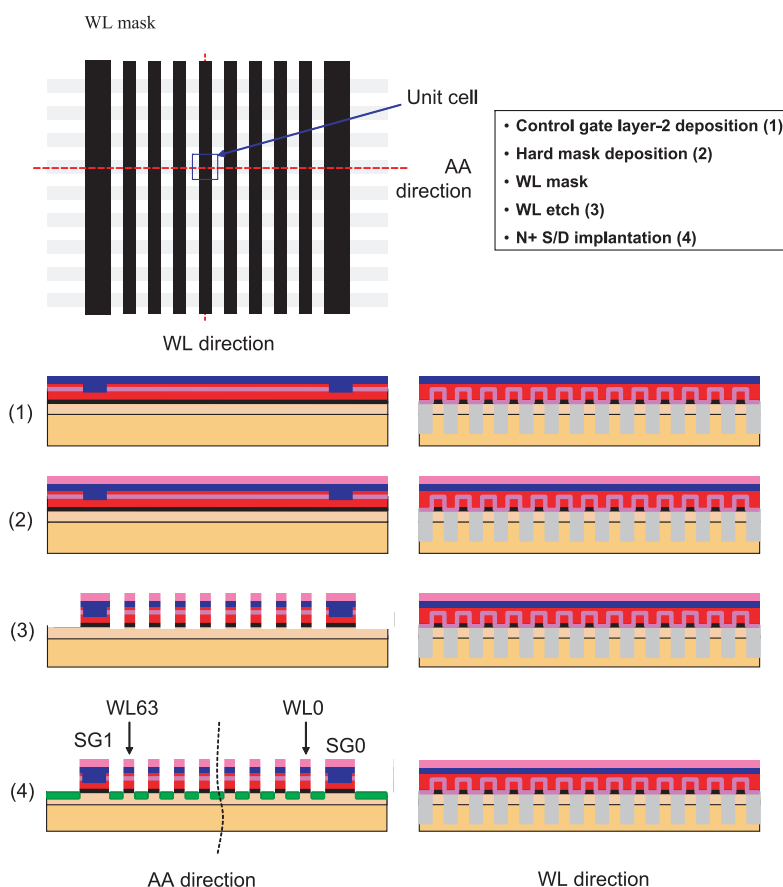


Figure 14.52 WL module. The top schematic shows the WL mask of the array area. The left and right sides of illus. (1) through (4) are the cross sections of the WL processing steps (described in the textbox) in AA and WL directions, respectively.

the AA and WL. $F = 25$ nm for a 25-nm NAND flash device, which has 25-nm CDs for both AA and WL, and the unit area of the memory cell is 2500 nm^2 , or $0.0025 \text{ } \mu\text{m}^2$. A self-aligned double-patterning technique is suitable for patterning the WL layer and also can be applied to AA and BL layers when F becomes too small for a single exposure to handle.

Figure 14.53 shows the CB1 process. In the array area, contact holes are very dense in the WL direction, but very sparse in the AA direction because there is no contact in the 32- or 64-bit strings between the two select gates (see the top schematic in Fig. 14.53). Although the CB and contact-to-source line (CS) look the same in Fig. 14.53, their requirements are quite different. For CB holes, both hole-to-hole shorts and holes that do not open are killer defects, while only nonopen

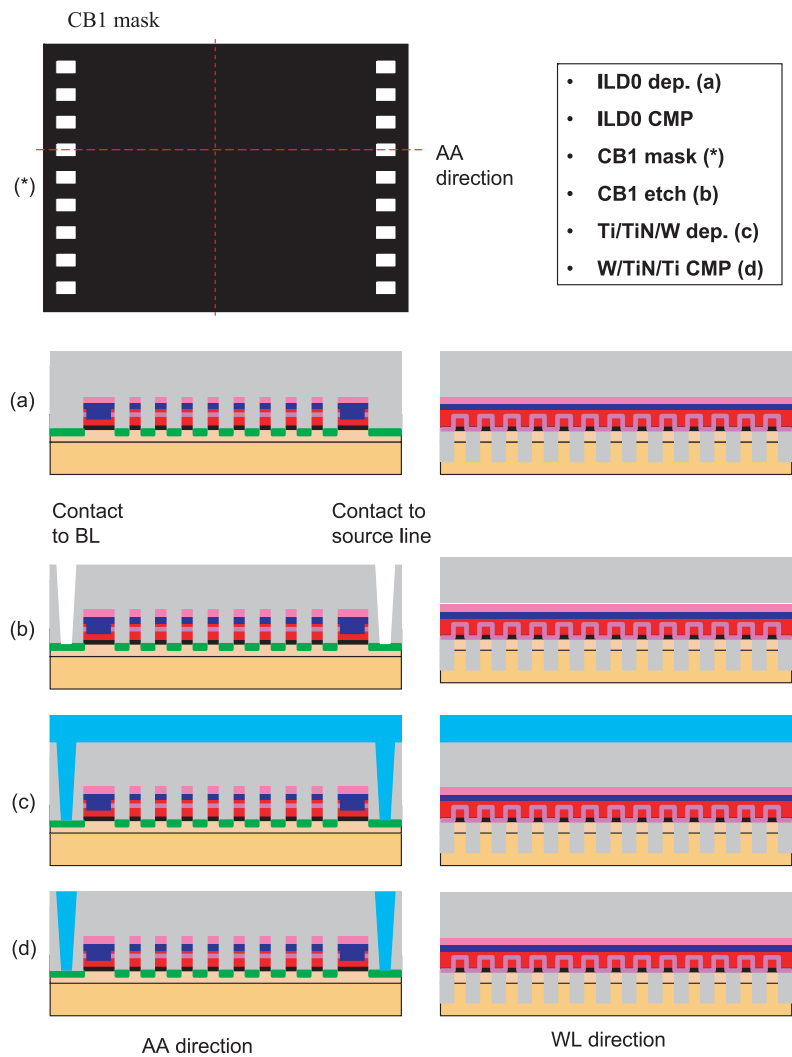


Figure 14.53 The first contact-to-BL (CB1) process. The top schematic shows the CB1 mask of the array area. The left and right sides of illus. (1) through (4) are the cross sections of the CB1 processing steps (described in the textbox) in AA and WL directions, respectively.

holes are killers for CSs. All CSs are connected to source lines that lead to a common ground; therefore, shorts between CSs are not killer defects.

Because the requirements of CB and CS holes are different, some designers use different hole patterns for CB and CS, as shown in Fig. 14.54. The curved dash line in Fig. 14.54 indicates that the distance between the CB and CS holes is far greater than what is shown in the figure.

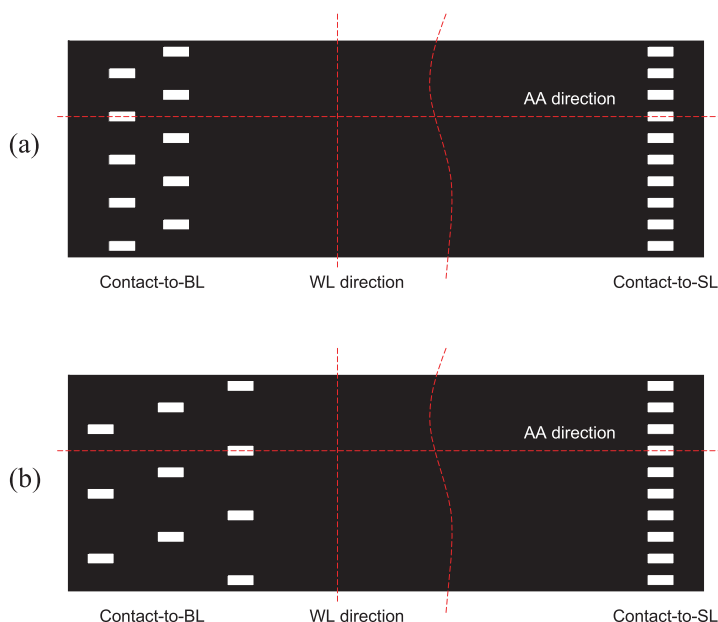


Figure 14.54 CB1 layout of an advanced NAND flash with different CB and CS hole patterns: (a) two-row and (b) three-row split of CB.

For 90-nm technology nodes, the CB hole CD is 90 nm, and the space between the CB holes is also 90 nm. 193-nm optical lithography can resolve the 90-nm hole patterns illustrated in Fig. 14.54. 193-nm high-NA immersion lithography can resolve 5 $\times$ -nm hole patterning without row splits. With further diminishment of feature sizes, CB hole patterns can be split into two rows, providing more room and allowing immersion optical lithography to resolve them. Figure 14.54(a) shows a two-row split of CB holes widely used in 3 $\times$ - to 2 $\times$ -nm nodes. At low 2 $\times$ -nm (such as 20 nm) and high 1 $\times$ -nm (18 nm) technology nodes, a two-row split can no longer be resolved by immersion optical lithography, and a three-row split of CB holes must be used, as shown in Fig. 14.54(b). Theoretically, CB holes can be split into four or even more rows to allow optical lithography to pattern low 1 $\times$ -nm contact holes. However, this would take a large amount of silicon surface from real devices. With a much shorter wavelength (13.5 nm), EUV lithography can resolve CB hole patterns without row splitting, a capability that can significantly increase device density.

The M0 process described in Fig. 14.55 is the module to form the source line and CB pads that the second contact-to-bitline (CB2) will land on. The process is a single inlay process, similar to contact and local interconnect processes for CMOS. Because the source line is very wide, oxide supports are extensively placed into it to avoid the dishing effect of metal CMP. To maintain a constant pattern density

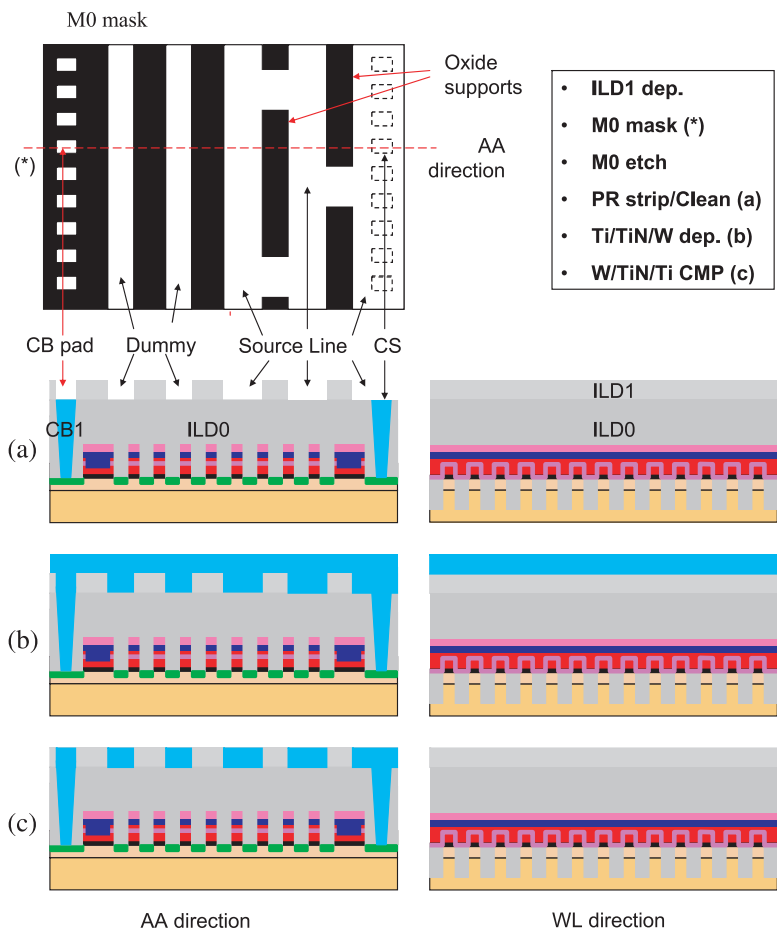


Figure 14.55 Metal 0 (M0) processes. The top schematic shows a M0 mask of the array area. The left and right sides of illus. (1), (2), and (3) are cross sections of the M0 processing steps (described in the textbox) in AA and WL directions, respectively.

to avoid erosion effects during metal CMP, dummy patterns are also designed into the M0 layer.

Figure 14.56 shows CB2 processes. ILD2 consists of an ESL and a bulk dielectric layer. The ESL is either silicon nitride, silicon oxynitride, or silicon-nitrocarbide (SiNC), and the bulk layer is silicon oxide, without or with carbon doping. After the ILD2 deposition process, a CB2 mask is used to pattern the wafer, and dielectric etching forms the CB2 holes that land on the CB pads, which connect to the CB1 plugs.

Figure 14.57 shows the M1 processes. In an array area, M1 is the BL, and in a periphery area, it is the local interconnection. Tungsten is commonly used in M1. Silicon oxide, either without or with carbon doping, is used for the ILD2.

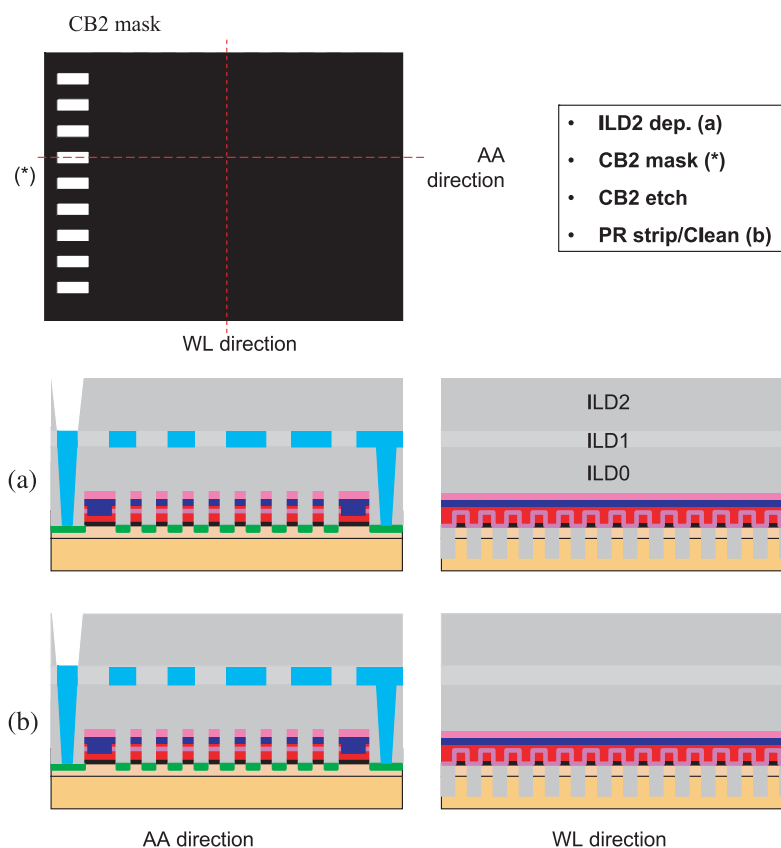


Figure 14.56 CB2 processes. The top schematic shows a CB2 mask of the array area. The left and right sides of illus. (1) and (2) are the cross sections of the CB2 processing steps (described in the textbox) in AA and WL directions, respectively.

Tantalum or tantalum nitride is widely used as a barrier layer for copper, and an ionized metal plasma process is frequently used to deposit the barrier and copper seed layers. ECP is used to deposit the bulk copper.

Figure 14.58 shows a cross section of a 64-bit string NAND flash with both array and periphery areas. There are intergate contacts in both nMOS and pMOS in the periphery area, and these devices have larger feature sizes than the devices in the array area. The CB1 in the array area is the contact in the periphery area. M0 forms the CB pad and source line in the array area and local interconnects in the periphery area. CB2 in the array area is the V1 in the periphery area, and M1 forms the BL in the array area, and the first global metal interconnect in the periphery area.

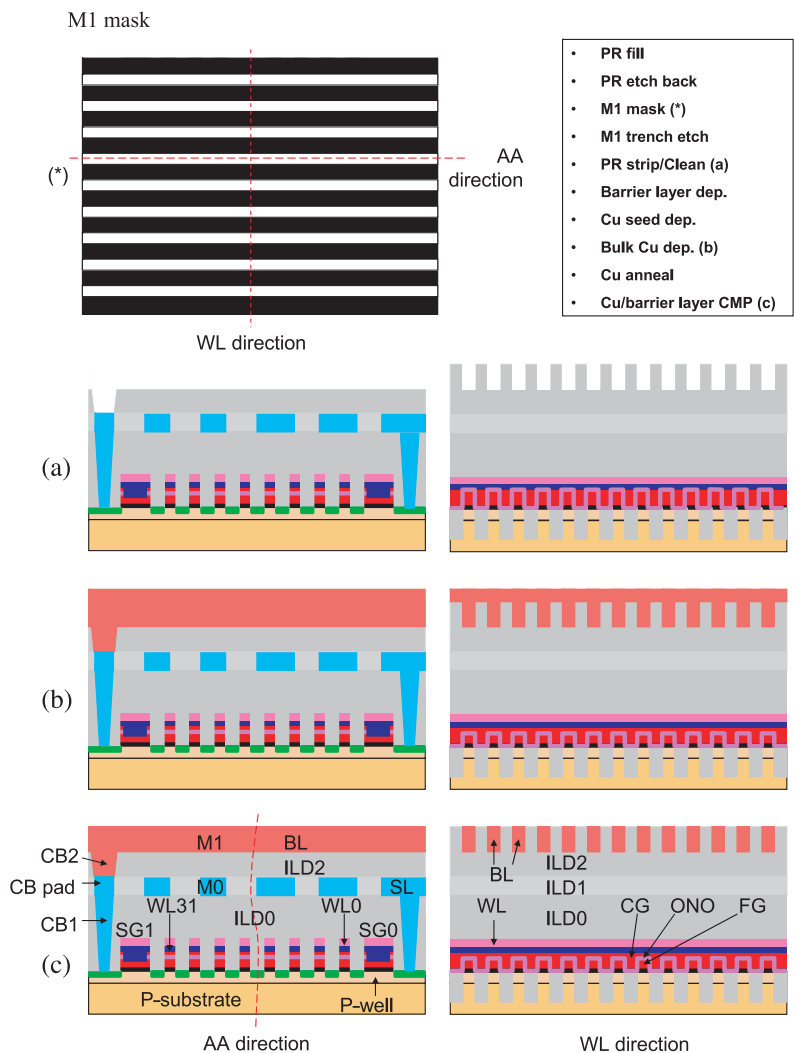


Figure 14.57 M1 processes. The top schematic shows an M1 mask of the array area. The left and right sides of illus. (1), (2), and (3) are the cross sections of the M1 processing steps (described in the textbox) in AA and WL directions, respectively.

14.7 Summary

- CMOS IC chips have dominated the semiconductor industry since the 1980s, when minimum feature size shrank from 3 μm to submicron levels.
- The 1980s saw the introduction of tungsten CVD and dielectric PECVD processes; sidewall spacers, plasma etching, steppers, and projection systems became common.
- In the 1990s, minimum feature size shrank from 0.8 to 0.18 μm .

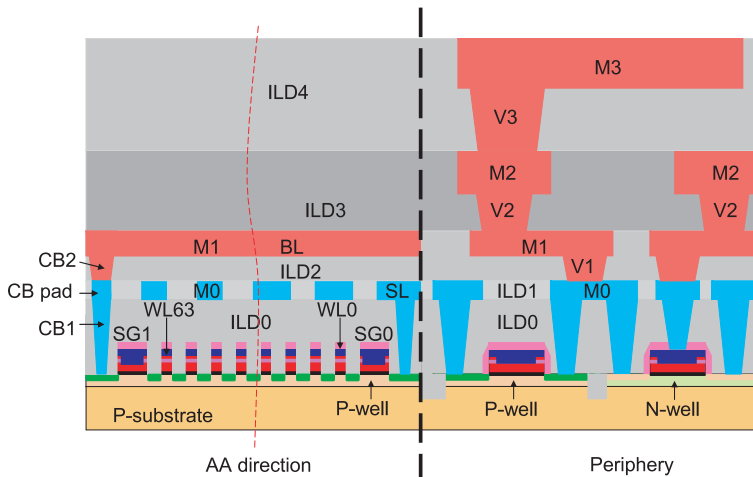


Figure 14.58 Cross section of a 64-bit string NAND flash with both array and periphery areas.

- Technologies such as RTP, silicides, CMP, HDP CVD, O_3 -TEOS-based dielectric CVD, and ECP processes were introduced or improved.
- In the first decade of 2000, IC technology nodes reduced to 25 nm in NAND flash memory devices, and to 28 nm in CMOS logic devices.
- Immersion 193-nm lithography, double patterning, nickel silicide, and SEG SiGe have all contributed to process advancements.
- High- κ dielectric and metal gates have been used in high-end microprocessors and other high-performance logic IC chips, and copper with porous low- κ dielectric interconnections have been used in high-volume production.
- DRAM and NAND flash devices are the most important memory devices in IC production. With the development of 3D graphics, stream video, and other applications, further technological developments will be seen in memory chip manufacturing.

14.8 Bibliography

- C. Y. Chang and S.M. Sze, *ULSI Technologies*, McGraw-Hill, New York (1996).
- K. Kim, "Technology for sub-50-nm DRAM and NAND flash manufacturing," *IEDM Tech. Dig.*, 323–326 (2005).
- K. Y. Lim et al., "Novel stress-memorization-technology (SMT) for high electron mobility enhancement of gate last high- κ /metal gate devices," *IEDM Tech. Dig.*, 229–232 (2010).
- K. Mistry et al., "A 45-nm logic technology with high- κ + metal gate transistors, strained silicon, 9 Cu interconnect layers, 193-nm dry patterning, and 100% Pb-free packaging," *IEDM Tech. Dig.*, 247–250 (2007).

- S. Natarajan et al., “32-nm logic technology featuring 2nd-generation high-k + metal-gate transistors, enhanced channel strain and 0.171m² SRAM cell size in a 291-Mb array,” *IEDM Tech. Dig.*, 941–943 (2008).
- K. Prall and K. Parat, “25-nm 64-Gb MLC NAND technology and scaling challenges,” *IEDM Tech. Dig.*, 102–105 (2010).
- L. Shon-Roy, A. Wiesnoski, and R. Zorich, *Advanced Semiconductor Fabrication Handbook*, Integrated Circuit Engineering Corp., Scottsdale, AZ (1998).
- H. Xiao, “Method for Forming Memory Cell Transistor,” U.S. Patent Application 12553067 - Filed on Sep 2, 2009.

14.9 Review Questions

1. What was the isolation material used in CMOS IC chips in the 1980s? In the late 1990s?
2. What were the metal materials used in CMOS IC chips in the early 1980s and mid-1990s? What were metal materials that most likely will be used in the 2010s?
3. What materials are used for the final passivation of CMOS chips?
4. List the silicon nitride layers used in copper metallization processes. Try to explain their functions.
5. Silicon nitride is a good barrier for copper diffusion. However, no one uses silicon nitride as the main dielectric in copper metallization. Why?
6. What is the main difference between titanium silicide and cobalt silicide?
7. Why has nickel silicide replaced cobalt silicide since the 65-nm technology node?
8. What is the main difference between USG and FSG?
9. Compare the ILD0 processing steps listed in Secs.14.2 and 14.3. What are the differences?
10. Discuss the advantages and disadvantages of hydrogen plasma premetal cleaning.
11. What are the advantages of a dual damascene process compared to an aluminum alloy interconnection process?
12. What is the benefit of high-κ dielectrics?
13. Why does a pMOS created with a gate-last HKMG process have a higher speed than a pMOS created with a gate-first HKMG process?
14. Why do IC manufacturers need to generate lead-free solder bumps?
15. List the two types of DRAM capacitor structures. Which type is more popular in DRAM manufacturing?
16. List at least three contact layers that are in a stacked DRAM cell array area. Which one has the highest hole density, and which one has the lowest?
17. Draw a sketch of a basic flash device. What is the main difference between it and an nMOS?
18. What are the main differences between NOR and NAND flash devices?

Chapter 15

Future Trends and Summary

In the first decade of the 21st century, IC processing technologies saw some fundamental changes from the traditional processing technologies of the 20th century.

Copper replaced aluminum–copper alloys and tungsten to form metal interconnection lines in advanced CMOS logic IC chips because of its lower resistivity and high EMR. By using copper interconnections, device speed can be increased, power consumption can be reduced, and reliability can be improved. Low- κ and ULK dielectrics replaced USG and FSG for the interlayer dielectric applications. PECVD organosilicate glass, or SiCOH, and porous OSG processes were developed and integrated with copper to form interconnections. There are very few choices for reducing RC (resistance multiplied by capacitance) delay. To change metal to reduce R, the only choice is silver (Ag), which has a bulk resistivity of $1.6\ \mu\Omega\text{-cm}$, just a little lower than the $1.7\ \mu\Omega\text{-cm}$ of copper. It is unlikely that the gain from switching to silver would justify the cost of the materials and process development, making it unattractive to IC manufacturers. New materials with even lower resistivity on a nanometer scale, such as carbon nanotubes (CNTs), have been studied to be applied to IC interconnections. To reduce C, air gap is one approach to further reduce the κ -value. In the future, interconnections that integrate CNTs and air gaps could receive some attention; however, cost will determine whether these interconnections will be implemented in mass production.

Metals made a remarkable comeback after being replaced by polysilicon as the gate electrode materials of MOSFETs in the mid-1970s. For CMOS logic devices, TiN has been used as a gate electrode for gate-first approaches. Replacement metal gate, or the gate-last process, is becoming mainstream for metal gate formation in HKMG CMOS manufacturing. Because metal has a lower resistivity than polysilicon/silicide stacks, metal gates and local interconnections can help reduce power consumption and increase device speed. TiN and W have been used as gate electrodes for array transistors in bWL DRAM chips in which TiN serves as a work-function metal, and W as a bulk metal for the gate as well as the WL.

To conduct electric current between the source and drain, a conductive channel must be formed underneath the gate by reversion of the minority carrier there. Minority carrier reversion can be achieved by charging a metal–oxide interface with applied gate voltage. Therefore, capacitance of the MOS gate capacitor must be large enough to hold the charges. With feature size reduction, the area of

the capacitor electrode decreases quickly, and so does the gate capacitance. To maintain enough capacitance, the distance between the two electrodes (which is the thickness of the gate dioxide layer) must be reduced. For 0.18- μm IC chips, the gate oxide is about 35 Å. It needs to be approximately 25 Å for 0.13- μm minimum-feature-size IC chips, and about 15 Å for 90-nm devices. The problem is, if the gate oxide thickness further decreases, gate-to-substrate leakage will increase significantly, and IC chip reliability and low power performance will be affected. Nitridated silicon oxide with $\kappa \sim 5$ has been used in 90-, 65-, 45-, and even in 32/28-nm devices. High- κ dielectrics such as hafnium dioxide (HfO_2 , $\kappa \sim 25$) have been developed and implemented in HKMG CMOS manufacturing, starting from the 45-nm technology node.

Many other high- κ dielectrics such as titanium dioxide (TiO_2 , $\kappa \sim 80$), tantalum pentoxide (Ta_2O_5 , $\kappa \sim 26$), and zirconium dioxide (ZrO_2 , $\kappa \sim 25$) have also been researched and could be implemented in IC chip manufacturing in the future.

Research and development of the deposition and etching of dielectric materials with very high- κ values, such as BST ($\text{Ba}_{1/2}\text{Sr}_{1/2}\text{TiO}_3$, κ up to 600), has continued for a long time. It could be applied in DRAM chip manufacturing as a dielectric material for storage node capacitors if a stable, integrative, and manufacturable deposition process can be developed.

Planar MOSFETs have been used as basic IC building blocks for more than half a century. Because traditional scaling is rapidly approaching its physical limit, and the cost to scale technology nodes in the deep nanometer range has skyrocketed, fabs are actively looking for other approaches to extend technology development. One promising technology is the 3D multiple-gate field effect transistor, or MuGFET. Figure 15.1 shows a MuGFET on a SOI substrate. By folding the active area (AA) of a planar MOSFET into a thin strip like a fin, device dimensions can be dramatically reduced without feature size shrinkage. This is shown in Fig. 15.1(b), which folds the AA with channel width Z from Fig. 15.1(a) into a fin, with $2h + w = Z$. Because the MuGFET shown in Fig. 15.1(b) has gate electrodes on three sides of the fin, it is called a triple-gate FET, or simply a trigate. The trigate device illustrated in Fig. 15.1(b) could perform similarly to the planar MOSFET shown in Fig. 15.1(a). By using a trigate, fabs can use the same lithographic technology to reduce transistor dimensions while maintaining the same performance, or to improve device performance without increasing device density. With trigates, IC manufacturers can further improve device performance by making the fin taller; however, that will pose great challenges for etch processing development, polysilicon deposition, polysilicon dummy gate etch, SEG, dummy gate removal, trench cleaning, HKMG deposition, etc.

Silicon fin etching has a nitride hard mask with a pad oxide underneath. By keeping this hard mask while forming the MuGFET, the gate electrode is only on two sides of the fin, not on the top or bottom. Fabs usually call this double-gate FET a finFET. For finFET devices, channel length only relates to the height of the fin ($2h$) and is no longer related to the fin CD (w), which is the smallest CD of the device. From a processing point of view, a finFET is easier to build than a trigate. Intel's 22-nm IC chips are built with trigate devices.

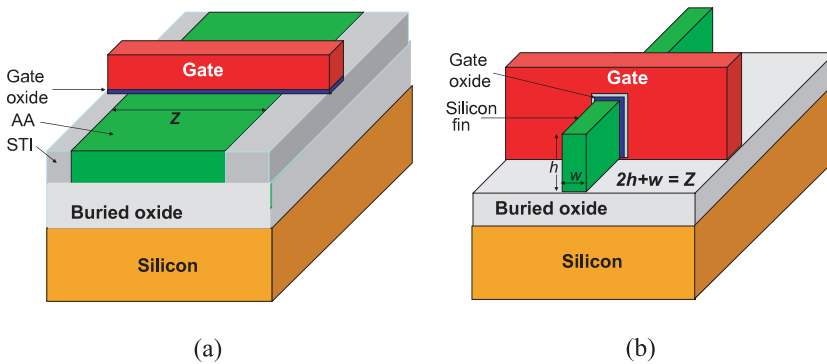


Figure 15.1 (a) A 3D illustration of a planar MOSFET with SOI substrate, and (b) an equivalent trigate FET with SOI substrate.

Using semiconductor materials with higher carrier mobilities than silicon is another way to improve device performance. For example, germanium (Ge) has much higher hole mobility than silicon, $1900 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ of Ge versus $430 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ Si, and III-V compound semiconductors such as gallium arsenide (GaAs) have much higher electron mobility than silicon ($8000 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ versus $1630 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$). If Ge is used for pMOS channels and GaAs for nMOS channels, both the hole mobility of the p-channel and electron mobility of the n-channel can be significantly improved. Because the driving current of MOSFET is proportional to carrier mobility, CMOS with epitaxial Ge pMOS channels and GaAs nMOS channels can be used in future IC manufacturing to enhance device performance without aggressively scaling the feature size. Other new materials, such as CNTs and graphene, have received a large amount of research attention for creating nanometer-scale electronic devices for silicon eras in the future.

Figure 15.2 shows a timeline of IC technology nodes and the start of new technologies and materials in IC chip manufacturing. With innovations such as double patterning and source-mask optimization (SMO), fabs can extend optical lithography to 14 nm. With quadruple patterning, optical lithography could even stretch to a 10-nm technology node. The most promising next-generation lithography technology, EUV lithography, is still in development and inching toward production readiness. It could compete with quadruple patterning of immersion 193-nm optical lithography for mass production at a 10-nm technology node.

Although the semiconductor industry has been developing for more than 60 years, it is still not a mature industry like the automobile industry. New technologies are frequently introduced, and device structure and materials change for every nanometer technology node. Manufacturing tools worth multimillions of dollars when they are brand new become obsolete in just few years and must be replaced if manufacturers want to keep up with technological developments. The skyrocketing costs of an advanced IC fab keep many IC manufacturers from pursuing the most advanced technology nodes. For example, a 300-mm fab of

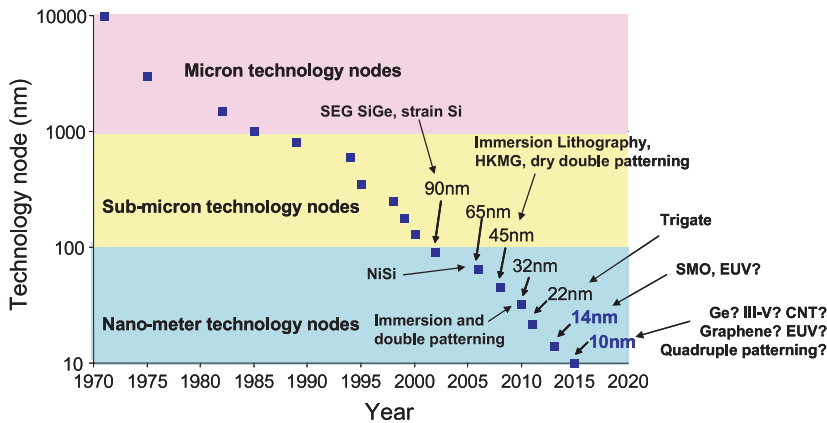


Figure 15.2 Technology nodes and introductions of new technologies and materials.

65-nm nodes with production capability of 50,000 wafers per month would cost approximately three billion dollars. It would cost approximately 4.5 billion dollars to build a 300-mm 32/28-nm fab of that capacity. The cost would be even higher for 20- and 14-nm node fabs.

As processing technology continues to develop, the cost of the semiconductor fab rapidly increases. At the same time, IC chip prices continue to decrease, which drives the manufacturers to turn to larger wafer sizes to reduce cost. The 300-mm wafer has become mainstream for advanced wafer fabs, and the first 450-mm wafer fabs could appear by the end of the second decade of the 21st century. The cost of a 450-mm fab could be more than ten billion dollars! Predictably, only few highly profitable IC manufacturers could afford this fab.

After the 10-nm technology node, silicon will run out of atoms to form a reliable MOSFET. The quantum tunneling effect between source and drain increases dramatically when the channel length is less than 10 nm, unless the drain voltage V_D drops significantly, to maybe 0.3 V. Tunneling-induced S/D leakage, I_{off} , is very sensitive to channel length, which means that IC manufacturers will need to control the gate CD and SDE junction profile very tightly, so tightly that it could become unmanufacturable.

Even before physics stops the scaling of IC devices, there is a good chance that economics could stop the process. In the good old days, shrinking the feature size made the device faster, less expensive, and less power hungry, which gave IC manufacturers enough incentive to invest in research and development for decreasing the device geometry. Now the skyrocketing research and development costs to scale nanometer devices have forced many IC manufacturers to stay away from leading edge technologies, leaving only a few financially sound companies to lead the charge to the bleeding edge.

Since planar scaling is rapidly approaching physical and economical limits, 3D integration is becoming more attractive. For a memory chip, if two dies are stacked together, the amount of storage is doubled with the same chip area, which

is equivalent to shrinking the feature size by one generation. If four dies can be stacked, the amount of storage will be quadrupled, which is equivalent to reducing the feature size by half, or two technology nodes. With wire bonding, fabs have already stacked 8 chips in mass production and 16 chips in development. Through-silicon via (TSV) technology recently has received much attention. Numerous resources have been dedicated to TSV technology development, and it will help continue the goal of “more Moore” or “more than Moore” to continue putting more devices on a chip, even though planar scaling reaches its limit.

DRAM had been an important driver of IC technology, and it still plays an important role in the IC industry. After the long competition between the two capacitor technologies—stack and deep trench—DRAM manufacturers are working on bWL technology, which can reduce a very challenging contact layer compared to a normal stacked DRAM, and allows $6F^2$ array transistor density. F is the smallest feature size of the transistor. By making a gate-all-around transistor in the vertical direction and using buried-wordline-buried-bitline (BWBB) technology (shown Fig. 15.3), the density of array transistors can further reduce to $4F^2$.

NAND flash chips have grown rapidly, driven by the demands for storage in mobile digital devices, such as cell phones, digital cameras, digital camcorders, MP3 players, USB drives, and tablet PCs. NAND-flash-based solid-state drives (SSDs) have been used in high-end laptops. They have also been used with traditional magnetic HDDs to form hybrid drives, which combine the advantages of a fast-reading SSD and the low cost of an HDD. NAND flash technology developed rapidly, driven by market demands and furious competition between the last few manufacturers. One of the leading manufacturers, Toshiba/SanDisk, makes 64-Gb (or 8-GB) NAND flash memory with 19-nm technology.

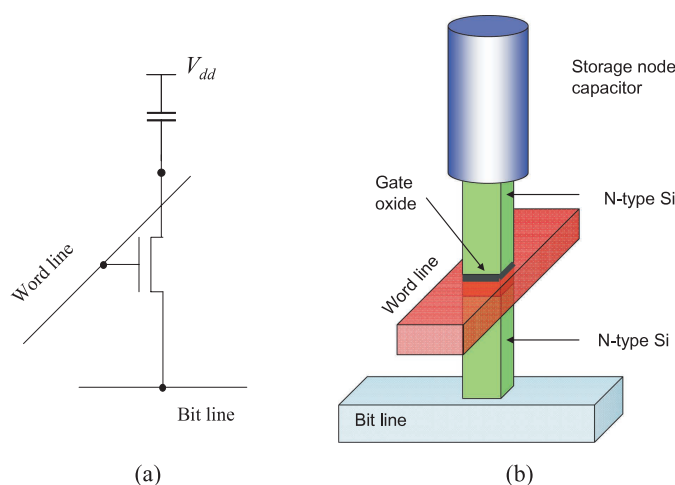


Figure 15.3 (a) Circuit of a DRAM unit cell and (b) 3D illustration of bWL, buried bitline $4F^2$ DRAM.

Traditional NAND flash is formed by strings of 32 or 64 series-connected flash devices on a silicon surface. By making the device string in the vertical direction, packing density dramatically increases. Figure 15.4 illustrates a birds-eye view and top view of a 3D NAND flash. In the illustration, the string only stacks four flash devices plus upper and lower select gates. In mass production, 16 flash devices can be stacked in a string and connected to another string from the bottom to form a 32-bit string; both select gates are located in the upper layer.

DRAM is fast and volatile, while NAND flash is nonvolatile but slow. For a long time, a fast NVM device has been sought that can be used for applications of both DRAM and NAND flash. Many types of novel memory devices have been proposed and studied. One promising candidate, phase-change memory (PCM), is already in production. Another candidate, magnetoresistive random access memory (MRAM), is still in intensive research and development, and could be integrated in system-on-chip IC manufacturing.

3D device approaches have been developed and applied in the production of CMOS logic IC, DRAM, and NAND flash chips. 3D devices will begin to replace planar devices in advanced technology nodes. Also, 3D packaging is already being used in CMOS image sensor chip manufacturing. It will eventually be applied in other IC productions.

Decreasing IC chip prices drives down the price of consumer electronics, such as flat-panel TVs, Blu-ray Disc™ players, MP3 players, smartphones, and PCs, including desktops, laptops, and tablets. The continuously dropping price of consumer electronics and the steady economic growth of developing countries, especially China and India (which together have more than one-third of the world's population), will dramatically increase the purchasing power and demand for consumer electronics.

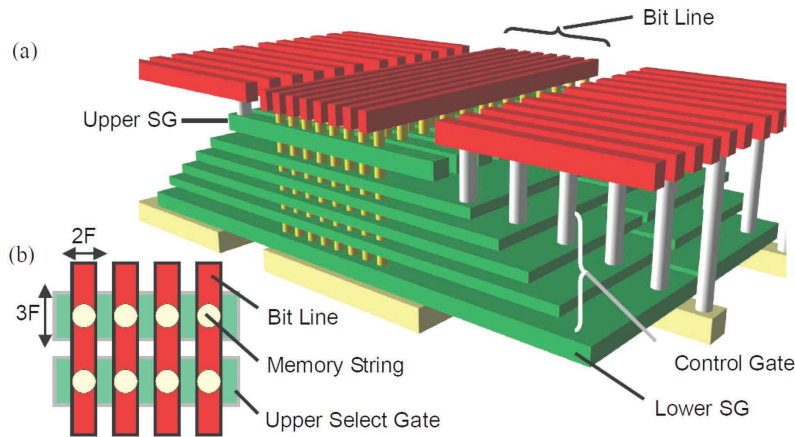


Figure 15.4 (a) Birds-eye view and (b) top view of a 3D flash memory array. (Y. Fukuzumi, et al.)

In addition to the electronics industry, demands for IC chips from other industries such as the automobile and medical/healthcare industries are also increasing rapidly. More chips will be used in motor vehicles to make them safer, more comfortable, have higher fuel efficiency, and easier to drive. Gadgets such as GPSs, smart routing and traffic management, collision warning radar, and voice-activated Internet access to audio on-demand services could become standard features in automobiles in the near future. By miniaturizing test probes and analysis circuits, medical IC chips have been developed for DNA testing, and fast, accurate diagnostics of DNA-related diseases. Medical and healthcare applications of IC chips, such as lab-on-chip (LoC) systems, will grow quickly in the future. Miniature fiber optics and photonics devices could be integrated with silicon IC chips to form interconnections between dies in a same-chip package to dramatically increase the data transmission rate. The development and popularization of 3D video in gaming and entertainment industries provide an additional driving force for more computational power and memory capability. Robotic pets have already been used in senior centers to tirelessly give comfort to lonely seniors, and robotic housemaids could serve ordinary households in the foreseeable future. Worldwide development of the wireless Internet and telecommunications combined with smartphones and tablets will continue to be an important driving force that feeds the rapid development of the IC industry in the near future.

Technologies that branched out from the semiconductor industry also have pushed into other industries, such as microelectromechanical systems (MEMS), light-emitting diodes (LEDs), and solar industries, which have helped drive down the cost of these products to make them more affordable for consumers worldwide.

It is hard to accurately forecast the demands for IC chips, a challenge that partly contributes to the periodic boom-bust nature of the IC industry. However, the trend is very clear in the foreseeable future. Global demands for IC chips will steadily grow, as will global demands for skilled, knowledgeable, innovative, creative, and hardworking technicians, engineers, and scientists in the semiconductor industry.

15.1 Bibliography

- Y. Fukuzumi et al., "Optimal integration and characteristics of vertical array devices for ultra-high density, bit-cost scalable flash memory," *IEDM Tech. Dig.*, 449–452 (2007).
- S. Lai, "Current status of the phase change memory and its future," *IEDM Tech. Dig.*, 255–258 (2003).
- S. Lai, "Non-volatile memory technologies: the quest for ever-lower cost," *IEDM Tech. Dig.*, 11–16 (2008).

Index

A

- acceptor, 62
- acoustic measurement method, 468
- adding process, 267
- alignment and exposure, 194
- aluminum, 49, 455, 471
- ammonium chloride, 166
- annealing, 156, 170, 281, 282
 - alloy, 157
- antipunch-through implantation, 295
- arithmetic-logic unit (ALU), 78
- arriving angle, 377
- Atalla, M. M., 73
- automatic diameter control (ADC)
 - system, 103

B

- bandgap, 59
- Bardeen, John, 2
- barrier materials, 499
- batch process
 - tools, 39
- batch reactors, 112
- beam line system, 285
- beam stop, 294
- Bell Laboratories, 2, 73
- bird's beak, 134, 552
- Boltzmann distribution, 248
- bonding pad, 49, 53
 - etching, 80
- borophosphosilicate glass (BPSG), 87
- Brattain, Walter, 2

C

- capacitance, 66, 67, 299

- capacitor, 66
- carbon fluoride gases, 44
- central processing unit (CPU), 78, 285
- channeling effect, 278
- chemical mechanical polishing (CMP), 44, 55, 507
 - advantages, 515
 - applications, 516
 - copper, 537
 - endpoint detection, 538
 - hardware, 519
 - oxide, 534
 - tungsten, 535
- chemical mechanical polishing (CMP) process, 107
- chemical vapor deposition (CVD), 42, 110, 159, 254, 371
 - aluminum, 457, 483
 - atmospheric pressure CVD (APCVD), 374
 - copper, 502
 - dielectrics, 370
 - high-density plasma CVD (HDP-CVD), 426
 - low-pressure CVD (LPCVD), 164, 254, 375
 - metal, 475
 - plasma etch CVD (PECVD), 242, 258
 - plasma-enhanced CVD (PECVD), 164, 254, 377
 - rapid thermal CVD (RTCVD), 173
- chemisorption, 382
- chip

packaging, 47
 cleanroom, 12, 24, 28, 29, 32, 33, 35, 45, 212
 cobalt, 464
 color chart, 402
 complementary metal-oxide semiconductor (CMOS), 85, 87, 89
 inverter, 14, 17
 process flow, 581, 582, 598, 616
 conducting band, 61
 contact printer, 194
 contamination, 29, 33, 35, 195
 control, 30
 elemental, 304
 particle, 31, 303
 contamination control, 542
 copper, 65, 462, 501
 chemical vapor deposition (CVD), 502
 deposition, 464
 electrochemical plating (ECP), 500
 interconnection, 574
 metallization, 89, 497, 498
 seed layer, 499
 slurry, 527
 copper metallization, 19
 crystal, 97, 103
 defects, 99
 orientation, 105
 crystal pulling, 104
 Czochralski (CZ) method, 102, 103

D

damaging mechanism, 333
 dangling bonds, 141
 dc bias, 238, 250
 defects, 114, 530
 degowning, 35
 deposition, 111, 188
 polysilicon, 39, 161
 rate, 467
 silicon nitride, 164
 silicon oxide, 43
 thin-film, 79
 deposition rate, 405

depth of focus (DOF), 213
 design fault, 47
 development process, 203
 diborane, 39, 153
 die
 testing, 46, 47
 yield, 25, 26, 53
 dielectric
 low- κ , 576, 612
 dies
 test, 28
 diffusion, 150, 151, 168, 268, 271, 282
 bay, 38
 doping, 38
 process, 153
 diode, 69, 70
 dioxide
 silicon, 279
 discrete device, 3
 dishing effect, 16, 530
 dissociation collision, 241
 dopant, 168, 174, 285, 296
 concentration, 154, 155, 305
 diffusion, 174
 dopants, 61
 concentration, 63
 n-type, 62, 72
 p-type, 63
 doping, 150
 drive-in, 153
 dual damascene, 439, 497, 537, 574
 process, 19
 dynamic dispense, 189
 dynamic random access memory (DRAM), 41, 66, 75, 345, 349, 479, 625, 627

E

E-chuck, 339
 edge-bead removal (EBR), 190, 191
 electric erasable programmable read-only memory (EEPROM), 76
 electrochemical plating (ECP), 500

- electrochemical plating deposition (EPD)
 - tools, 43
 - electromigration, 456
 - electron, 208, 238, 240
 - free, 238
 - electron cyclotron resonance (ECR), 264
 - electronic design automation (EDA), 16
 - electronic-grade silicon (EGS), 101, 103
 - ellipsometry, 40, 146, 399
 - epitaxial silicon deposition, 109
 - epitaxial silicon growth, 80, 110, 111
 - epitaxial silicon layer, 110
 - epitaxy
 - low-temperature (LTE), 115
 - selective, 116
 - epitaxy process, 114
 - epitaxy systems, 112
 - equipment area, 45
 - erasable programmable read-only memory (EPROM), 76
 - etch, 315
 - dielectric, 342
 - dry (plasma), 330, 340
 - main, 322
 - mechanisms, 333
 - metal, 329, 357
 - oxide, 344
 - polysilicon gate, 353
 - profile, 320
 - rate, 317
 - reactive ion (RIE), 320, 332
 - single-crystal silicon, 349, 351
 - wet, 325, 340
 - etch bay, 40
 - etch pits, 97
 - etch process, 42
 - etching, 40
 - evaporation
 - electron beam, 486
 - excitation-relaxation, 240
 - exhaust system, 129
 - exposure control, 199
- F**
- facility area, 45
 - Fairchild Semiconductor, 4
 - feature size, 6, 8, 11, 31, 53, 78, 89, 199, 201, 218
 - field effect transistor (FET), 72, 73
 - film
 - thickness, 539
 - flip-chip technology, 49
 - floating zone (FZ) method, 102
 - foundry company, 16
 - four-point probe, 305, 474
 - free radicals, 241, 242, 257, 260, 333
 - Frenkel defect, 99
 - fumed silica, 524
 - furnace, 39, 126, 167, 282
 - horizontal, 129
 - vertical, 129
- G**
- gap fill, 256, 380
 - gas delivery system, 128
 - gas system, 284
 - gases
 - processing, 128
 - gate, 10, 72, 76
 - capacitance, 445
 - germanium, 2
 - bar, 4
 - single-crystal, 3, 62, 73, 96
 - glass transition, 202
 - gowning, 33, 35
 - graphics processing unit (GPU), 78
 - gyrofrequency, 248
 - gyromotion, 247, 286
 - gyroradius, 248, 288
- H**
- hard bake, 205
 - heating process, 80, 187
 - Horni, Jean, 4
 - hot electron effect, 297

hydrofluoric acid (HF), 35, 37, 107, 325

hydrogen, 39, 114, 144

I

I-V curve, 70

implant bay, 42

implantation, 564

in-situ process, 173

insulator, 59

integrated circuit (IC)

 application-specific (ASIC), 79

 design, 16

 first, 3

 processing, 53

 technology, 1

Intel Corporation, 5

interconnection, 42, 67, 89, 570

 copper, 574

 early, 572

 local, 570

interlayer dielectric (ILD) film, 347

intermetal dielectric (IMD), 88, 370

interstitial defect, 99

ion bombardment, 248, 249, 255,

 257–259, 330, 331, 333

ion implantation, 42, 79, 82, 151, 206,

 267, 270, 279, 291, 294, 299, 305

 advantages, 270

 applications, 271

ion implanter, 247

ionization, 238, 240

J

junction depth, 150, 174

junction spiking, 456

K

Kilby, Jack, 3

killer defect density, 27

L

laminar flow, 33

light sources, 199, 215

lightly doped drain (LDD), 297, 561

line resistance, 452

lithography

 electron-beam (e-beam), 231

 extreme ultraviolet (EUV), 80, 226

 nanoimprint (NIL), 228

 next-generation (NGL), 216

 x-ray, 229

loading effect, 16, 321

loading system, 128

local oxidation of silicon (LOCOS),

 87, 89, 133, 552, 554

logic design, 14

low-pressure chemical vapor

 deposition (LPCVD), 38

M

magnetic field, 247, 288, 294, 489

mask, 16, 18, 23, 194, 196, 223, 227,

 230

mass-transport-limited regime, 111, 389

material structure, 95

 amorphous, 95

 polycrystalline, 95

 single-crystal, 95

mean free path (MFP), 244, 330

memory, 75, 76

 chip, 75, 625

 flash, 77, 78, 636

metal-insulator-semiconductor (MIS), 96

metal-oxide semiconductor (MOS), 269

metal-oxide-semiconductor field

 effect transistor (MOSFET), 42, 72,

 73, 81, 85, 120, 564

metallization, 345, 451

 copper, 497, 498

metrology tools, 40, 42, 43

microprocessor, 78

microscope

 scanning electron (SEM), 40

Moore's law, 5

Moore, Gordon, 5

N

n-channel metal-oxide semiconductor (nMOS), 72, 296
n-type region, 69
nickel, 465
nitric acid (HNO_3), 37, 107
nitrogen, 39, 46, 114
Noyce, Robert, 4
nucleation step, 478

O

orientation
 crystal, 97
overhangs, 378
oxidation, 129–131, 137
 dry, 139, 140
 high-pressure, 145
 rapid thermal (RTO), 171
 trends, 149
 wet, 138, 142
oxide
 blanket-field, 82, 88, 132, 551
 gate, 134
 growth rate, 137
oxygen, 39
ozonator, 422
ozone, 421
 concentration, 422

P

p-channel metal-oxide semiconductor (pMOS), 72, 81, 296
p–n junction, 69
p-type region, 69
p-well implantation, 288
packaging, 46, 53
 ceramic, 51
 plastic, 51
 yield, 25, 26
particle, 30
 killer, 31
passivation, 413, 577, 614
patterning process, 8, 80, 179, 301
pellicle, 17
phase shift mask (PSM), 217

phosphine, 39
phosphorus silicate glass (PSG), 82
photo bay, 40
photolithography, 80, 82, 179, 199, 215
 optical, 9, 89
 process, 19, 31, 40, 184
photomask, 230
photoresist, 23, 181, 183
 coating, 188
 negative, 31
 positive, 31
 strip, 260, 358
photoresist etchback, 514
photoresist strip, 41
photoresist stripping, 80
physical vapor deposition (PVD), 43, 374, 483
 aluminum–copper, 495
 titanium, 493
 titanium nitride, 494
physisorption, 383
planar technology, 4
planarization, 511
plasma, 237, 239, 248, 330
 definition of, 237
 high-density, 238, 262, 331, 340
 inductively coupled (ICP), 263
 potential, 250
 radio frequency (rf), 238
plasma etch, 256
 remote, 261
plasma-enhanced chemical vapor deposition (PECVD), 42
polishing head, 521
polishing pad, 519
polymer, 182
polysilicon, 65, 66, 82, 101, 296, 297, 320, 454, 603
postacceleration, 289
postexposure bake (PEB), 202
Pourbaix diagram, 526
preclean, 492
predeposition, 152

premetal dielectric (PMD), 370
Preston equation, 528
prism coupler, 401
process gases, 45
processing tools, 45
profilometer, 467
profit margin, 26, 30
projection exposure system, 195

Q

quartz sand, 96, 100

R

rapid thermal annealing (RTA), 168
rapid thermal annealing (RTA)
 systems, 42
rapid thermal processing (RTP), 80,
 157, 167, 169, 174, 283
RC time delay, 67
reflectivity, 472
reflow, 159
refractive index, 146, 398
remote plasma CVD (RPCVD), 262
removal process, 40, 44, 79
removal rate, 528
resistance, 65, 70, 154
 sheet, 155, 305, 472
resistivity, 60, 63
resistor, 64
resolution, 180, 199, 213, 215, 221
reticle, 18, 197, 201
rf power, 433

S

safety, 232, 308, 503
scanning electron microscope (SEM),
 209, 467
Schottky defect, 99
selective epitaxial growth (SEG), 116
selectivity, 320, 322, 529
self-aligned contact (SAC), 345
SEMATECH, 5
semiconductor, 59
 manufacturing, 12, 36, 80
 materials, 59

 processing, 12
sensitizer, 183
shadowing effect, 279
shallow trench isolation (STI), 41, 89,
 349, 391, 554, 617
 self-aligned (SA-STI), 557, 639
sheath potential, 249
Shockley, William, 2, 269
shrinkage, 407
sidewall spacer, 164, 391
silane, 39, 111, 383, 411, 435
silicide, 570
silicon, 24, 96, 137
 crude, 101
 dioxide, 96, 100, 125, 130, 132,
 135, 140
 epitaxial, 159
 etch process, 41
 metallurgical-grade (MGS), 100
 nitride, 166
 on isolator (SOI), 600
 oxide, 255, 390
 polycrystalline, 104
 single-crystal, 3, 13, 62, 95, 97,
 102, 103, 267, 327
 substrate, 125, 300
silicon oxide, 55
silicon-oxide surface, 72
single-crystal
 semiconductor materials, 61
 structure, 61
single-crystal structure, 281
single-wafer implantation, 293
slurry, 107, 522, 524
 aluminum, 527
 copper, 527
 metal polishing, 525
 oxide, 523
soft bake, 193
solvent, 183
spectroreflectometry, 404
spin, 204
 disk, 293
spin-on glass (SOG), 396, 424

sputtering, 418, 487
 reactive, 460
sputtering etch process, 43
standing wave effect, 201
static dispense, 189
static random access memory
 (SRAM), 75
step coverage, 377
step-and-repeat, 18
step-and-repeat system, 196
stepper, 197
sticking coefficient, 386
stopping mechanisms, 272
 electronic stopping, 272
 nuclear stopping, 272
stress, 408, 470
surface mobility, 374, 379
surface-reaction-limited regime, 111,
 388
system ground potential, 288

T

tantalum, 464
test, 52
tetraethyl-orthosilicate (TEOS), 43,
 384, 386
 plasma-enhanced (PE-TEOS), 434
tetraethylorthosilicate (TEOS), 347
tetramethyl ammonium hydroxide,
 203
Texas Instruments, 3
thermal
 evaporation, 485
thermal budget, 151, 281
thermal expansion, 409
thermal processing, 125
thermal stress, 53
thermal velocity, 245
thermal wave, 306
thickness, 402
thin films
 metal, 465
thin-film bay, 42
threshold adjustment, 563
titanium, 458, 481

nitride, 459, 481
silicide, 454, 458
track system, 211
transistor
 bipolar, 70, 71, 81
 first, 2, 70
 level design, 14
 npn, 71
 point-contact bipolar, 96
transition region, 69
trench fill, 132
trichlorosilane (TCS), 101
troubleshooting, 437
tungsten, 461, 477, 479
 silicide, 480
 slurry, 526

U

ultraviolet (UV) light, 40, 76, 232
uniformity, 407, 469, 529

V

vacancy, 99
valence shell, 59
void formation, 441

W

wafer, 97, 114, 193, 204
 charging, 302
 cleaning, 185
 finishing, 108
 handler, 292
 manufacturing, 13
 preparation, 550
 processing, 36, 47, 107
 silicon-on-insulator (SOI), 116,
 118, 272
 size, 9
 surface, 40
 temperature, 294, 339
 yield, 25, 26
wafer preparation, 88
wafer processing, 19, 28
wafer-to-wafer (WTW)
 nonuniformity, 156

wafer-to-wafer (WTW) uniformity,
112, 318

well formation, 88, 557
 self-aligned twin well, 88
 single well, 88

wet bay, 37

wet clean processing, 136

wet etch, 536
 process, 107

 rate, 406

wet processes, 37

wire bonding, 49, 50

X

xylene, 232

Y

yield, 25–28, 30



Dr. Hong Xiao is a principal engineer of KLA-Tencor Corp. Previously, he was a technical marketing specialist at Hermes-Microvision, Inc. and a technical manager at Hermes Epitek. Other past positions include a working as a consultant of semiconductor process technology, a senior process engineer at Motorola Semiconductor Production Sector, and an associate professor at Austin Community College. After receiving his Ph.D., 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 co-authored over 30 journal and conference papers. He is the author of *Introduction to Semiconductor Manufacturing Technology*, published by Prentice Hall in 2000. He has six US patents and more than ten patents in the application process. He has been a member of SPIE since 2005.

INTRODUCTION TO Semiconductor Manufacturing Technology

SECOND EDITION

Hong Xiao

IC chip manufacturing processes, such as photolithography, etch, CVD, PVD, CMP, ion implantation, RTP, inspection, and metrology, are complex methods that draw upon many disciplines. *Introduction to Semiconductor Manufacturing Technologies, Second Edition* thoroughly describes the complicated processes with minimal mathematics, chemistry, and physics; it covers advanced concepts while keeping the contents accessible to readers without advanced degrees. Designed as a textbook for college students, this book provides a realistic picture of the semiconductor industry and an in-depth discussion of IC chip fabrication technology. The text focuses on current fabrication technologies, but older technologies are discussed for historical context.



SPIE.

P.O. Box 10
Bellingham, WA 98227-0010

ISBN: 9781510616530
SPIE Vol. No.: PM220