Modern Complementary Metal Oxide Semiconductor (CMOS) Technology

April 11, 2024

1 Basic CMOS Technology and Device Structures

#Slide:3#

1. The basic building blocks of CMOS technology, inverter and NAND gate, are shown in Fig.1
2. The basic advantage of CMOS technology is that it doesn't dissipate static power.

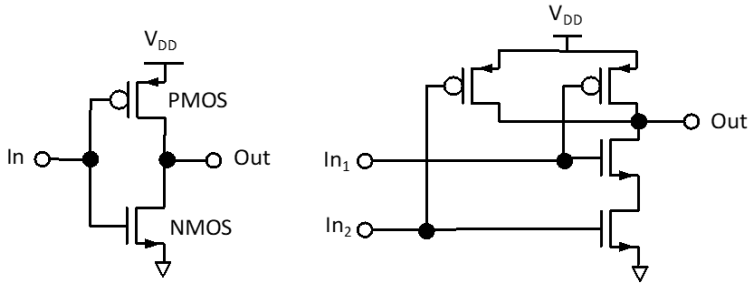


Figure 1: (a) Simple Inverter (b) NAND gate using CMOS technology

3. There are billions of transistors in a modern chip. No static power dissipation is a huge advantage as it results in power savings that no other technology such as bipolar technology can give.
4. Fig.2 shows the top view of the layout of a simple CMOS inverter.The black regions are the contact holes allowing the metal to make contact to the device source and drain regions. The blue band around and between the transistors is the shallow trench isolation (STI). The STI region provides lateral electrical isolation between transistors on the chio, The red line is the polysilicon gate.

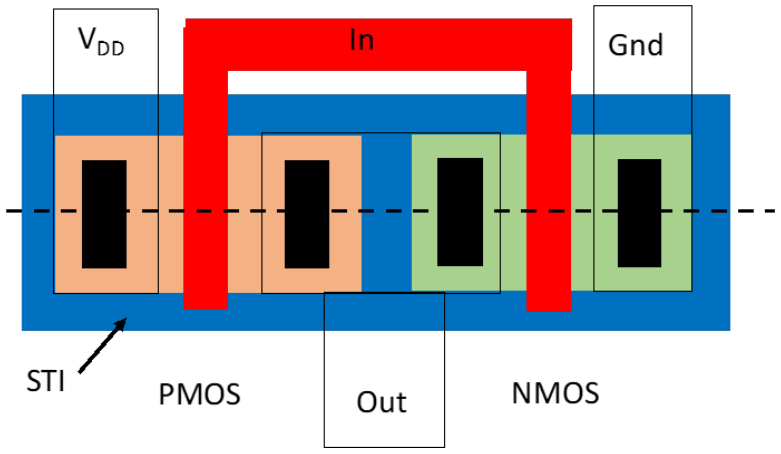


Figure 2: (a) Simple Inverter (b) NAND gate using CMOS technology

#Slide:4#

5. Logically, the way to fabricate a functioning chip is to do the steps that forms the transistors first, such as isolating them from each other, doping the silicon to make it more conductive in places and depositing the gate material to control the switching, before the structure is covered by the metal layers that interconnect the transistors together

6. All the doping, deposition and etching steps used to form the transistors are called the "front-end-process" steps and the interconnect layers on top of the devices are called the "back-end" steps.

7. The starting material is a bare silicon wafer. The key parameters for selecting this wafer are

   - the wafer diameter
   - the crystal orientation
   - the wafer doping type
   - the doping concentration in the wafer

8. Silicon wafers are crystalline in nature. The orientation of the surface defines the orientation of the wafer. The (100) wafer orientation provides the best electrical interface between the silicon substrate and insulator silicon dioxide. Si/SiO_2 interface is the best in terms of lack of defects, like unbonded atoms, charges etc. It is almost universally used for building CMOS devices

9. The doping type (P or N) in the starting wafer is somewhat flexible since the transistors are built in separate N-type or P-type regions or wells and in some sense the doping type of starting substrate does not appear critical

10. P-type is more often used in practise because it is somewhat easier to grow uniformly

11. The N-MOS transistors are typically not built on the substrate directly. The twin-well process is common because the doping method used to produce the P-well is much better controlled in manufacturing than in the substrate doping

12. Also, since the P-well and N-well doping concentrations are similar, it is easier to start with a much more lightly doped substrate and then tailor the wells for NMOS and PMOS devices individually

13. The doping concentration of the substrate, expressed in atoms per cubic centimeter (atoms cm^{-3}) or alternative in resistivity ($\Omega\ cm$) typically corresponds to a very light doping levels of parts per million of doping atoms in silicon wafer

14. Numbers in the range of 10^{15} atoms cm^{-3} correspond to a substrate resistivity of 5-20 $\Omega\ cm$

15. Boron is typically is used as the dopant for P-type substrate

16. The first step in CMOS processing is to make sure the indiviual devices are electrically isolated from each other and it is done by a building a shallow trench between the devices

17. The first step is to deposit a thin SiO_2 and a Si_3N_4 layer and spin photoresist that is used to pattern the underlying layers. The cross sectional view is shown is Fig.3

18. This oxide known as pad oxide is thermally grown and the nitride layer is deposited using lower pressure chemical vapor deposition (LPCVD)

19. A light sensitive photoresist is spin coated onto the wafer and is baked at 100 $^{\circ}C$ to obtain plastic like consist. Ultraviolet light is shown through a glass mask with patterned light and dark areas, which changes the properties of the photoresist locally (**MASK1**)

20. Positive photoresist will develop away in the regions exposed to light

21. This photolithography process is one of the most complex and expensive step in the manufacturing chips. Several innovative tricks are used to enable features smaller than the wavelength of the light source source to be built.

22. A technology node defines the minimum feature size that can be built in that particular processing. A 28 nm node has a minimum feature size of 28 nm

23. Shallow Trench Isolation (STI) dimensions are kept small as their role is to isolate transistors from each other. Larger dimensions waste space on the wafer

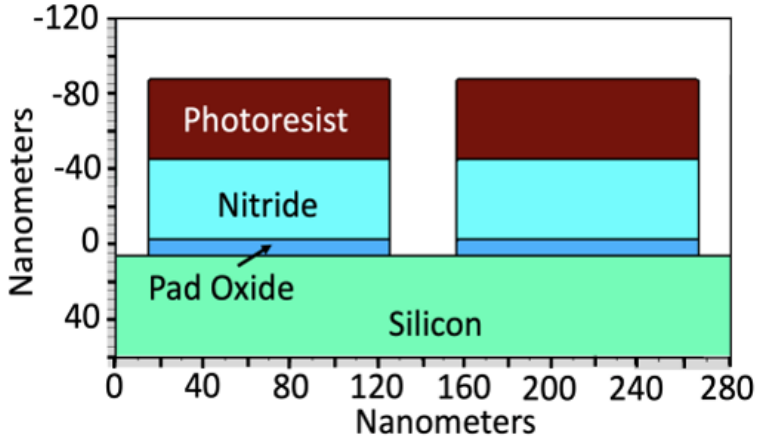


Figure 3: SiO_2, Si_3N_4 deposition; Photoresist Spin coat, etching of Si_3N_4,SiO_2

24. The features in 28 nm technology are printed with lithography tools using 193 nm light
25. The pattern in the photoresist layer is transferred to the underlying nitride, oxide and silicon by etching.
26. A gas source of flouring atoms and ions is to perfrom etching, forming volatile SiF_4. The process is called plasma etching. It occurs in a plasma similar to that in a fluorescent light bulb.
27. The three materials Si_3N_4, SiO_2 and Si must be etched sequentially. By varying the gas source and the plasma parameters etch rates can be tailored for different materials
28. By using a gas source like CF_4 a competition can be set up to modify the etch profile, resulting in a tapered sidewall Si trench. Avoiding sharp corners minimizes the electric fields and improves isolation efficiency. At this particular point the structure looks like that in Fig.4

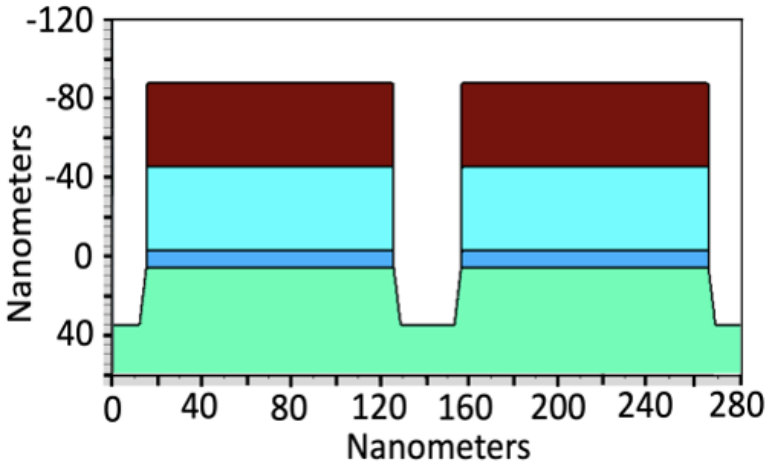


Figure 4: STI formation

29. Once the etching is completed, the job of photoresist is done and it can be chemically wet etched in piranha (H_2SO_4/H_2O_2) or dry etching in O_2 plasma, neither of which attacks the underlying nitride oxide or silicon layers
30. In the next step, the wafers are cleaned and a thin layer of silicon is grown in the trench by thermal oxidation in a high temperature furnace
31. The nitride film is very dense and acts a blocking layer for oxygen or water vapor. So the oxidation only happens only in the exposed region
32. In thermal oxidation, when one unit of silicon is oxidized, the oxide both grows into the silicon and out of the silicon, creating 2.2 units of silicon dioxide

33. If the oxidation is done locally on a planar surface, the surface is no longer flat because the oxide sticks up. More importantly, because the oxide is expanding when it grows, yet is attached to the silicon substrate, mechanical foraces are generated in the structure as the oxide expands. These mechanical stresses are particularly important at the corners of the trench structure and are the primary reason to grow only a thin liner oxide

#Slide:5#

34. The next step is to fill and indeed overfill the trench with a deposited oxide from LPCVD. Precursor gases like SiH_4 and O_2 to produce SiO_2

35. It is important that the filling process not leave gaps or voides in the trenches especially as the trench sidewalls become more vertical to improve density. The liner deposition and the LPCVD oxide deposition steps are shown in Fig.5

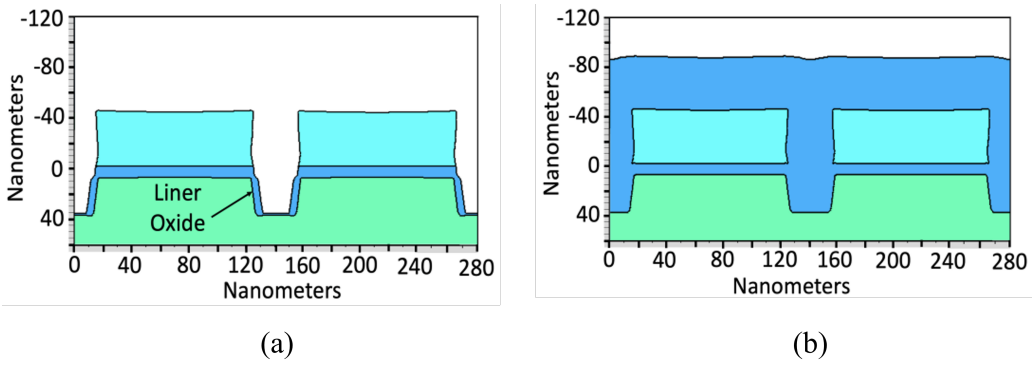


Figure 5: (a) Liner Oxide Deposition (b) LPCVD Oxide Deposition

36. The next step is a Chemical Mechanical Polishing (CMP) step that involves pressing silicon on a rotating abrasive wheel with a chemical slurry. This polishes off the excess oxide from the top surface of the wafer, leaving a planar substrate with oxide-filled trenches

37. The hard nitride layer serves a polishing step. The wafer looks like that in Fig.6

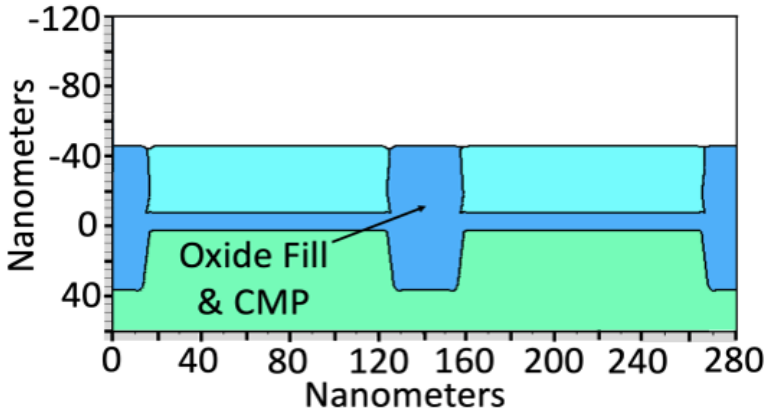
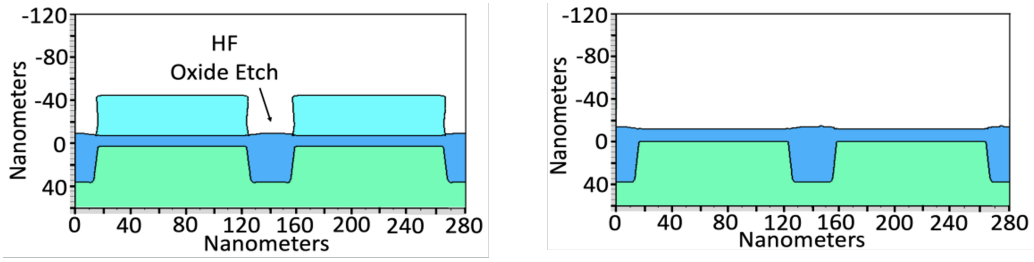


Figure 6: STI formation

38. Once the CMP operation is complete, the nitride layer is etched after the residual oxide is etched. Fig.7 shows wafer after nitride removal. At this stage, the wafers are ready for device fabrication in each of the isolated regions. The next step is to put in P- and N-type wells for the two types of transistors

39. Well formation is a two mask process, each mask is required for opening the respective well regions and doping the unprotected region

40. While the first mask for STI had features near the minimum feature size to form a narrow trench and avoid wasting space, the well mask only needs to select an entire transistor area for doping. A cheaper lithography tool could be used for these steps and these trade-offs are common and are called mix and match in the lithography tool set.



(a) (b)

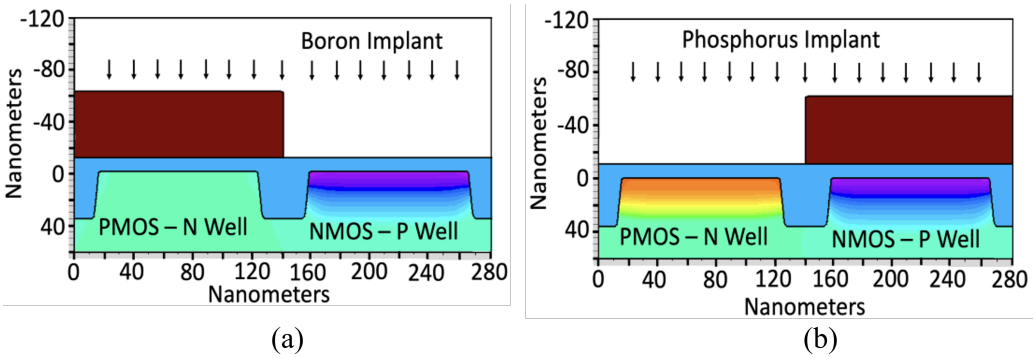
Figure 7: Nitride removal

#Slide:6#

41. In the P well regions some fraction of silicon atoms in the well region needs to be replaced with boron atoms. There are two processes to do the same. The yesteryear technique is diffusion based doping which was used for deeper junction. This is no longer used. Alternatively ion implantion is used as because it is precise, controllable and reproducible

42. The implant occurs everywhere on the wafer unless the region is protected by photoresist or any other suitable masking layer

43. The formation of P-well (Mask #2) and N-well (Mask #3) is shown in Fig.8



(a) (b)

Figure 8: (a)P-Well (b)N-Well formation

44. In the ion implantation, a small version of a linear accelerator is used to accelerate boron ions to high energies, shooting them at the silicon wafer

45. The process is precise because the accelerating voltage can be dialed in with high accuracy, determining how far the ions travel into the silicon. It is reproducible because it is easy to measure the ion current

46. Ion implantation can create a retrograde well with higher doping deep in the silicon and lower doping close to the surface. This can reduce the capacitance between the source/drain regions and the well and improve the switching spped of the transistor

47. Ion implantation comes at a cost, because the ions slow down due to two processes - electonic and nuclear stopping

48. Electronic stopping occurs because the boron is charged. Silicon nuclei surrounded by a sea of electrons create a drag force that slows the ions down

49. The nuclear stopping process occurs bbecause the boron ion can easily strike a silicon atom, knocking it off its lattice site and thus creating damage in the latticce

50. The silicon atom can recoil a signifciant distance from its original lattice site and can itself create further damage until it too comes to rest

51. This damage must be repaired somehow since the transistors need perfect crystalline substrate. This can be done repaired by a simple high temperature anneal

52. When the boron ion with hundreds or thousands of electron volts of energy hits the silicon, an atom is knocked off its lattice sit, often sent some large distance, like a elastic ball collision and the boron atom loses energy. This process gets repeated many times before coming to rest some distance into silicon

53. This is a random statistical process and can be modelled as a Gaussian distribution , where the peak occurs at the projected range, which is the average distance that a boron ion goes before it stops

54. Since it is a statistical process, there is a distribution. The peak concentration is determined by how much boron is implanted, a parameter known as dose, measured in atoms per sqcm. This is the integral or area under the Gaussian distribution

55. The choice for N-type dopants are phosphorus, arsenic or antimony.

56. Phosophorous is preferable because it gives a symmetric well. For symmetry P- and N-type wells, the dopants should diffuse at roughly the same rate, Boron and phosphorous have identifcal diffusion coefficients, thus resulting same P and N well depths

57. Arsenic and antimony diffuse a lot slower and it would have resulted in a deeper N-well and a shallower P-well

58. There is a thin oxide on the surface of the silicon. The ion energy must be high enough to penetrate that layer and get into silicon

59. It is this amorphous glass layer that helps randomize the path of the ions so that they do not go down the regular channels in the periodic silicon lattice

60. The ion energy cannot be high that it penetrates the photoresist, which acts as the blocking layer. So the ion energies must be chosen carefully

61. After the ion implantation steps, the silicon is damaged and the dopants are randomly distributed near the surface

62. The dopants have to be diffused to final junction depth while restoring the crystallinity of silicon

63. Annealing at high temperature in inert ambience is used to achieve this

64. Below the damaged silicon is perfect single crystal silicon and it provides a seed for regrowing the top part of silicon that is damaged.

65. As long as thermal energy is provided for the atoms in the top region to move around, they can find the right place to attach themselves to the substrate and layer-by-layer regrow a perfect silicon crsytal This process is called solid-phase epitaxy

66. The anneal step performs three things:

    - Provides the thermal energy for the solid-state diffusion process to drive the dopants to the correct depth
    - Activates the dopants by placing them on substitutional sites
    - Repairs the damage

67. Fig.9 shows the structure after implantation and the doping profile

68. The regions that end up being the source and drain are designed to align with the gate to minimize the overlap capacitance. Source and drain regions are formed after the polysilicon gate

#Slide:7#

69. The next steps are the implants (Mask #4, Mask #5) in the channel region under the gate to avoid implantation through polysilicon gate which would damage the fragile and critical thin gate dielectric

70. The purpose of these implants in the channel region is to fine-tune the doping concentrations under the gate of the transistor to set the threshold voltage (V_{TH}) of the transistors

71. The V_{TH} implants can be adjusted to control the threshold voltage to any desired value. Some transistors may have a high V_{TH} and very low leakage current, while others may a have low V_{TH} and a relatively higher leakage current

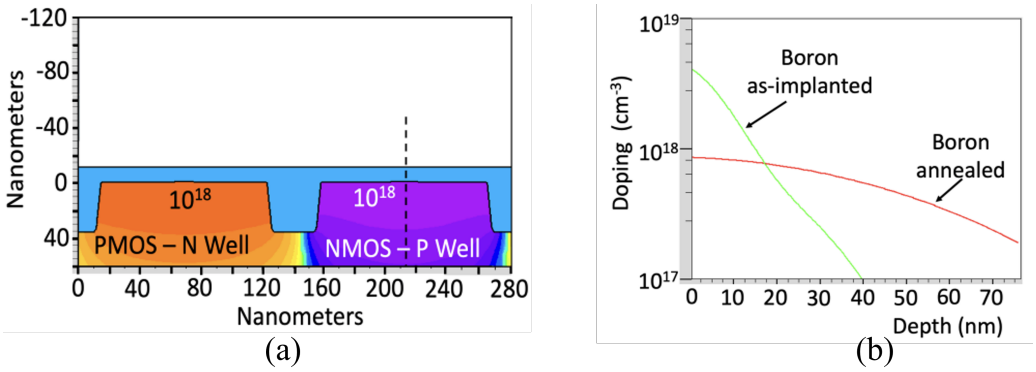


Figure 9: (a) After Implantation and anneal (b) Doping profile

72. V_{TH} implant goes everywhere in the transistor but it will get swamped out by the heavier source to drain implants in those regions later in the process

73. These are called threshold-adjust implants and there is one for each type of transistor.

74. In a modern process, there would be several threshold voltages to provide options for circuit designers

75. The change in the threshold voltage from this implant is

$$\boxed{\Delta V_{TH} = \frac{qQ_i}{C_{ox}}} \tag{1}$$

76. A modern process might use a single masking step and multiple implant at different energies to form the well structure, with a shallow V_{TH} adjust implant, a deeper implant to set the bulk well concentration and a very high energy implant to form a "punch-through" implant layer at the base of the trench. A short anneal would repair the damage and leave the implant profiles largely unchanged

77. The threshold implant step is shown in Fig.10

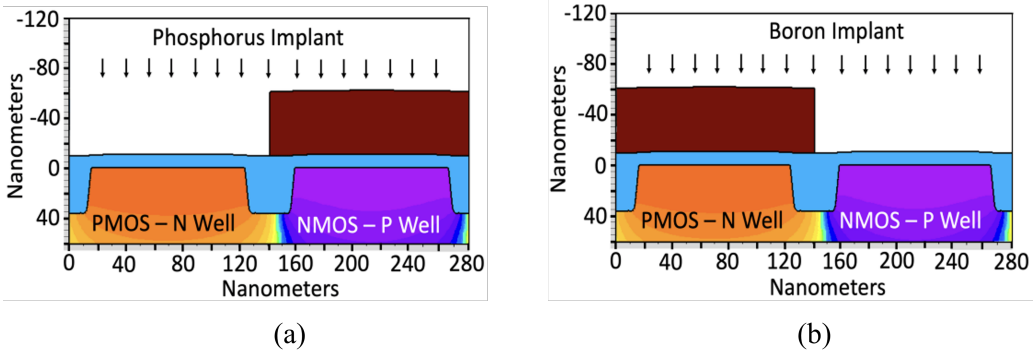


Figure 10: Threshold adjust implants(a) N-type (b) P-type

78. This kind of retrograde well profile (the doping is higher as you deeper in the substrate) may have advantages for avoiding "latch-up" in CMOS devices

79. There are small doped regions at the edge of the gate, called extensions, or tips or lightly doped drain (LDD) regions.

80. These features are much smaller than the minimum feature size in the technology

81. It is very likely that the dimension of the gate is the best that can be done lithographically, since the dimension sets the performance of the transistor and is as small as the technology can make it

82. Before the polysilicon deposition, the thin oxide that is already existing is removed and a pristine gate oxide is grown.

83. After regrowing gate oxide, polysilicon is deposited immediately to avoid contamination

84. The polysilicon is deposited from a gas source using LPCVD. The film is polycrystalline becayse it deposits on an amorphous insulator

85. The polysilicon needs to be an N-type conductor on the NMOS device and a P-type conductor on the PMOS device, which could be achieved by masking the large transistor region and implanting the N- or P-type dopnsts, respectively.

86. An anneal at high temperature would distribute the dopant uniformly in the polysilicon. The diffusion would occur quickly, as the polysilicon grain boundaries would help redistribute the implant

87. This is done to set the work function for the transistor gates

88. In 28 nm process flow, the polysilicon acts as only a temporary or "dummy" gate and is replace with a high-K dielectric/metal stack. Thus doping of polysilicon is not an issue

89. Earlier generation of CMOS technology directly used the polysilicon material and doping was critical

90. The next step is to etch the polysilicon gate (Mask #6) as shown Fig.11(a). This dimension is probably the smallest dimension on the entire integrated circuit surface that is formed by lithography

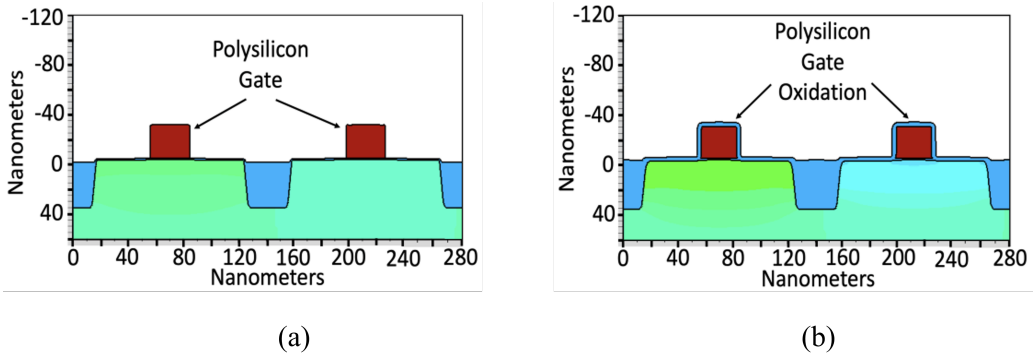


Figure 11: (a) Polysilicon Patterning (b) Passivating oxide on polysilicon

91. Various techniques are adapted to print features much smaller than the wavelength the exposing source

92. The next step is the selective anisotropic etching of polysilicon. This is a critical step and the etch has to stop at the gate oxide

93. After defining the polysilicon gates, a short oxidation is carried out to passivate the polysilicon as shown in Fig.11(b) All oxides get thicker during this step because silicon oxidizes everywhere

#Slide:8#

94. The next step is to perform the two most critical implants that determine the deivce performance.

95. The first is a halo or an angled implant that helps prevent the deep source/drain regions from punching through each other (Mask #7 and Mask #8)

96. The second is a very shallow tip or extension region that reaches to just under the gate edge

97. These complicated structures are required for the following reasons. In practise, the power supply level, the turn-on or subthreshold region of MOSFETs does not scale with the device dimensions. This results in high electric fields

98. High fields cause problems in semiconductor devices, often called "hot electron" problems because most of them are due to the high energies that electrons(or holes) can reach in high fields.

99. At high energies, carriers can cause impact ionization, which creates a multiplier effect, making additional electron hole pairs by breaking Si-Si bonds

100. These carriers can even gain sufficient energy to surmount large barriers and can be injected into the gate dielectric. They may become trapped and cause device reliability problems

101. The high fields occur on the drain side of the device because of the high gate voltage and drain voltage there, while the source is usually grounded

102. The tips are symmetrically placed for simplicity even though that may actually only be required on the drain side

103. The lightly doped drain (LDD) region grades the doping profile between the drain and the channel and allows the drain voltage to be dropped over a larger distance than would be the case if an abrupt junction were formed. (Mask #9, Mask #10)

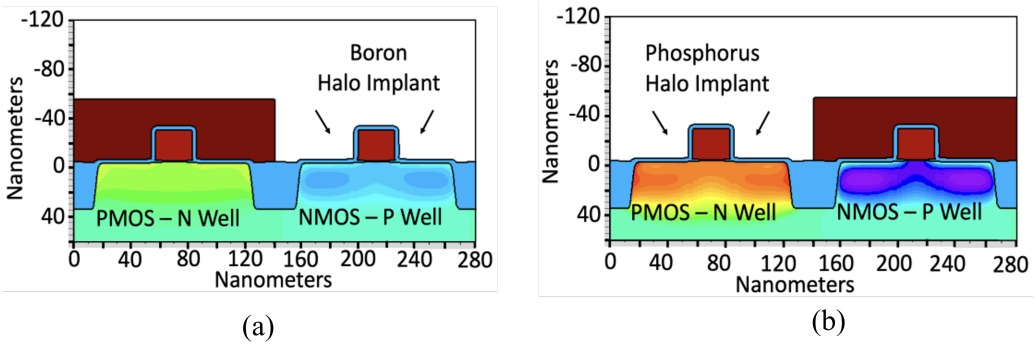


Figure 12: Halo implants(a) NMOS (b) PMOS

104. Since many of the deleterious effects of high electric fields depend exponentially on the field strength, even modest reductions in the field can make a significance difference in the reliability

105. Before LDD, a deeper implant is often performed, called as halo or angled implant. This implant is designed to prevent subsurface punch-through between the source and drain in the transistor

106. The halo implants as shown in Fig.12 use the same doping type as the channels and can be formed by implanting at a tile or an angle under the gate edge to locally raise the channel doping and help avoid short channel effects.

107. The sublithography tips are formed by two step process

- Very shallow tip implants are performed with the gate edge as the mask as shown in Fig.13
- The implants are self aligned to the edges of the gates and extend slightly underneath the gate edges because of the lateral scattering of the implamnt and lateral side ways diffusion during annealing step

#Slide:9#

- Spacers are formed by putting down a thin layer of conformal oxide or nitride
- This is followed by a highly anisotropic etch to remove the film in the flat regions and leave a narrow spacer on the sidewalls with the thickness equal to the deposited film thickness as shown in Fig.14

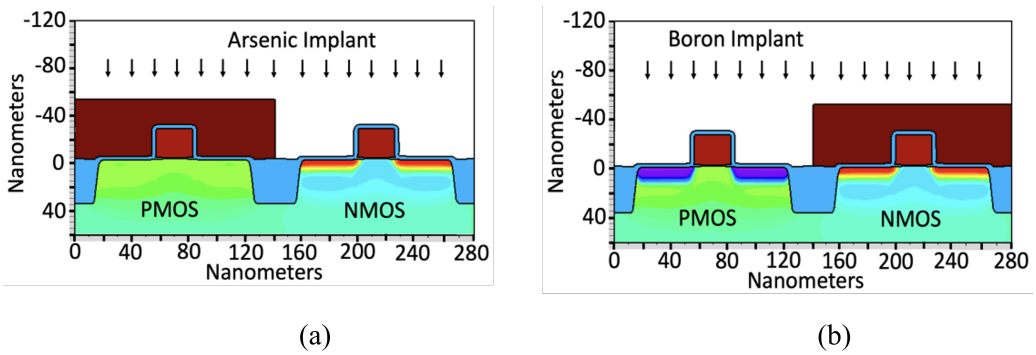


Figure 13: LDD doping (a) NMOS (b) PMOS

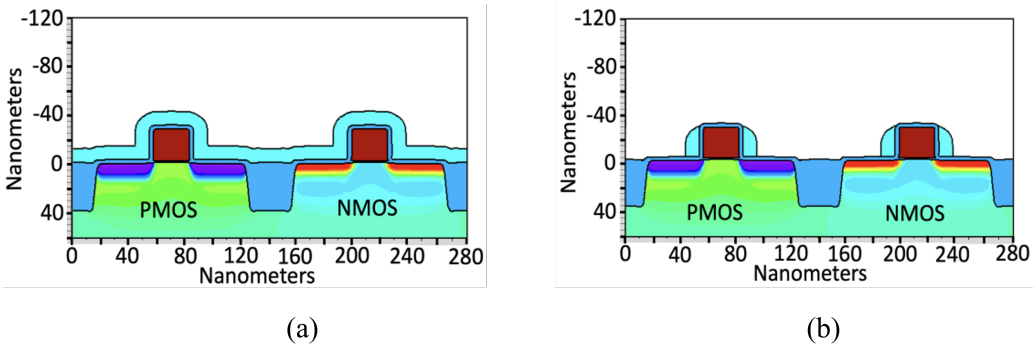


Figure 14: (a) Conformal Deposition (b) Spacer formation

108. The final step is to put in the deep source/drain regions that are self-aligned to the edge of the space region

109. During the spacer formation, the over etch required to clear the film in the flat regions would etch the underlying thin oxide by varying amounts

110. For this reason, the oxide is etched in HF down to bare siicon and a thin "screen oxide" is grown

111. Two mask steps are required to create source and drain regions of NMOS and PMOS as shown in Fig.15

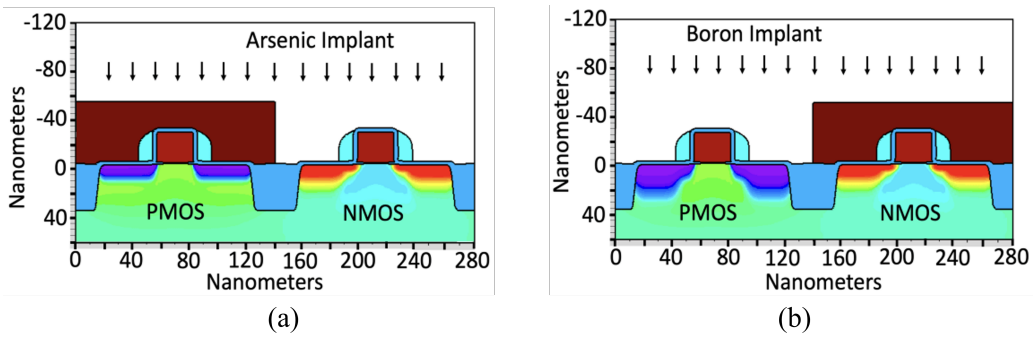


Figure 15: (Source Drain implant for (a) NMOS (b) PMOS

112. The only thing that remains is to repair the implant damage.

113. A short high temperature rapid thermal anneal (RTA) or a flash anneal is used to minimize the amount of dopant in this critical step

114. This repairs the damage in the substrate without allowing signifcant amount of dopant diffusion

115. Doing a low temperature anneal would result in an anomalous effect called transient enhanced diffusion (TED), where dopants move at very high rates because of the damage

116. The device structure shown in Fig.16 represents a fully functional self aligned polysilicon gate CMOS technology upto 90 nm node

117. At the 90 nm node, the push for higher performance required that the channel be strained to improve the hole and electron mobilities.

118. This was accomplished by replacing the deep source/drain implants in the PMOS device by etching the silicon in source/drain region and regrowing the region with a boron-doped SiGe layers through a process known as epitaxy

119. The germanium atom is bigger than silicon. So this process introduces compressive strain in the channel which improves hole mobility

120. A similar process using phosophorus-doped SiC epitaxy introduces tensile strain in the NMOS device and improves electron mobility

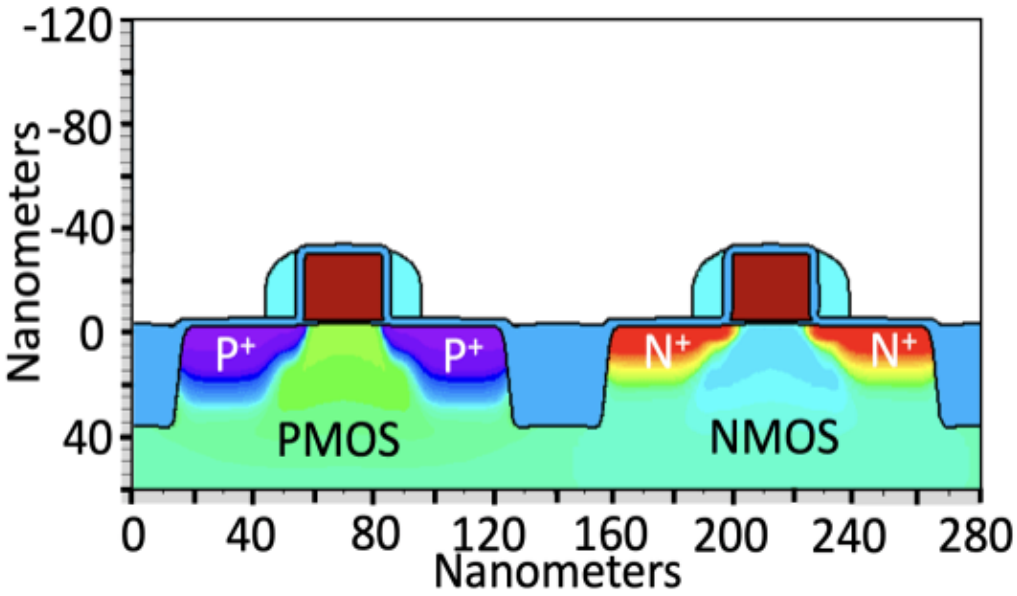


Figure 16: Transistor profile upto 90 nm technology node

121. This completes the fabrication of high performance strained channel CMOS device and the next step is the contact formation

122. The first step in the contact formation is to lower the sheet resistance of the source/drain regions by a process called silicidation

123. By depositing a reactive metal such Ti, Ni or Co using sputtering and performing a short anneal at low temperatures such as 350 or 450 $^{\circ}C$, a metal silicide layer is formed

124. The process is largely conformal. After the metal silicide is formed the unreacted metal on the dielectric regions must be etched off

125. This thin layer of metal silicide forms in the source/drain regions and gate region and lowers the sheet resistance that connects the contacts to the channel

126. This process was used upto 45 nm technology node. There after due to the introduction high-K dielectric as the gate electric, the gate contact is not silicided

127. The basic idea behind using high k dielectrics is to use a gate insulator with a higher dielectric constant than SiO_2 and trade off for a thicker layer which maintains the same gate capacitance and hence same electrical control over the channel.

128. Since the high K dielectrics are somewhat unstable, they are deposited as a last step for the FEOL process

129. A thick oxide layer is deposited and is polished all the way to the gate polysilicon using CMP

130. The polysilicon gate is removed in a very selective plasma etch process that stops on the oxide

131. The thin gate oxide is removed by a quick wet etch revealing the bare silicon channel surface

132. To stablize the interface, a very thin (, 1 nm) chemical oxide is formed on the silicon surface

133. The high K material is deposited using Atomic Layer Deposition (ALD)

134. To protect the dielectric, a thin capping layer of Titanium Nitride is deposited by ALD or sputtering

135. A thin etch stop layer of tantalum oxide is then deposited which enables a different metal stack to be formed in the NMOS and PMOS regions.

136. A thick TiN layer can be deposited everywhere and this sets the workfunction of PMOS devices

137. The PMOS devices is masked and the TiN layer etched in the NMOS gate stack. A deposited aluminium layer sets the work function of the NMOS device

138. Finally both gate stacks are filled with a final metal such as thick aluminum or Tungsten and the whole surface is planarized using CMP to the top of the gate stack

139. There may be one or more low temperature anneals performed during the stack depositions to cause some interdiffusion of the metal layers and correctly set the work function in the NMOS and PMOS devices respectively

140. The BEOL processing consists basically of a series of inuslator/metal depositions that are stacked on top of one another to produce interconnect or wiring levels

141. They rely heavily on CMO to planarize or flatten the deposited layers.

142. Inter-level dielectric (ILD) layers in the back end structure are typically deposited using plasma enhaced CVD at low temperatures

143. The oxide layers are doped with fluorine or carbon to reduce the dielectric constant. This minimzes the coupling capacitance between the layers and between the wires.

144. The contact holes are etched to the source/drain silicide regions and the gate metal regions.

145. A very thin TiN barrier layer is deposited by sputtering and a CVD tungsten fill is performed. The W is overfilled to completely cover the top of the surface

146. A CMP step polishes the W, removing it from the oxide regions, leaving the W plugs. The iterative portion of BEOL is carried out to interconnect devices