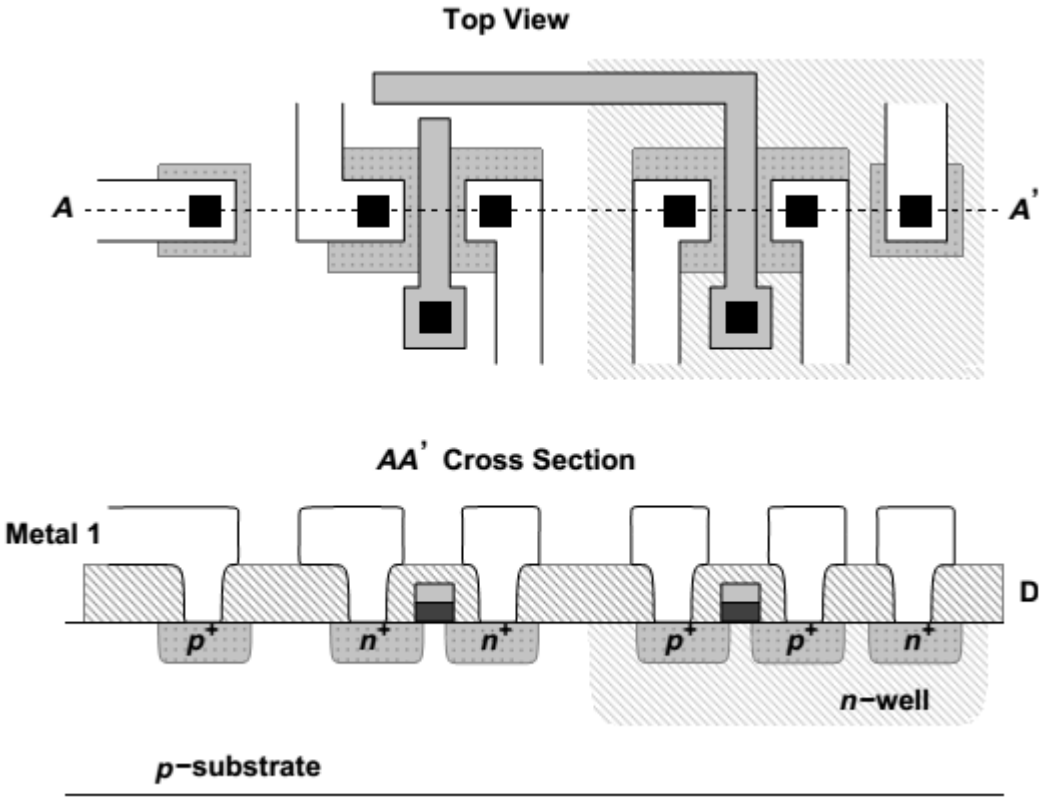
Chapter 18: CMOS Processing Technology

- ❑ **Fabrication processes**
- ❑ **CMOS fabrication**
- ❑ **Passive devices.**

General Considerations

- Before delving into a detailed study of fabrication, it is instructive to consider the basic structure of NMOS and PMOS transistors and predict the required processing steps.



Side view and top view of MOS devices.

- A p-type substrate (wafer) serves as the foundation upon which n-wells, source/drain regions, gate dielectric, polysilicon, n-well, substrate ties, and metal interconnects are built.

CMOS processes

- Considering both the side view and the top view, we may raise the following questions:

(1)How are various regions defined so accurately.

(2) How are then-wells and S/D regions built.

(3)How are the gate oxide and polysilicon fabricated.

(4) How are the gate oxide and polysilicon aligned with the S/D regions.

(5) How are the contact windows created.

(6) How are the metal interconnect layers deposited.

CMOS process flow

- **Modern CMOS technologies involve more than 200 processing steps, but for our purposes, we can view the sequence as a combination of the following operations:**

(1) wafer processing to produce the proper type of substrate

(2) photolithography to precisely define each region

(3) oxidation, deposition, and ion implantation to add materials to the wafer.

(4) etching to remove materials from the wafer.

CMOS process flow

- In semiconductor processing and characterization, we often refer to the "sheet resistance" of a layer.

- The total resistance of a rectangular bar is R= ρ.L=(W.t), where ρ is the resistivity of the material, and L, W and t, denote the length, width, and thickness of the bar, respectively.
- The quantity $R_{\square}$= ρ /t is thus defined as the sheet resistance.

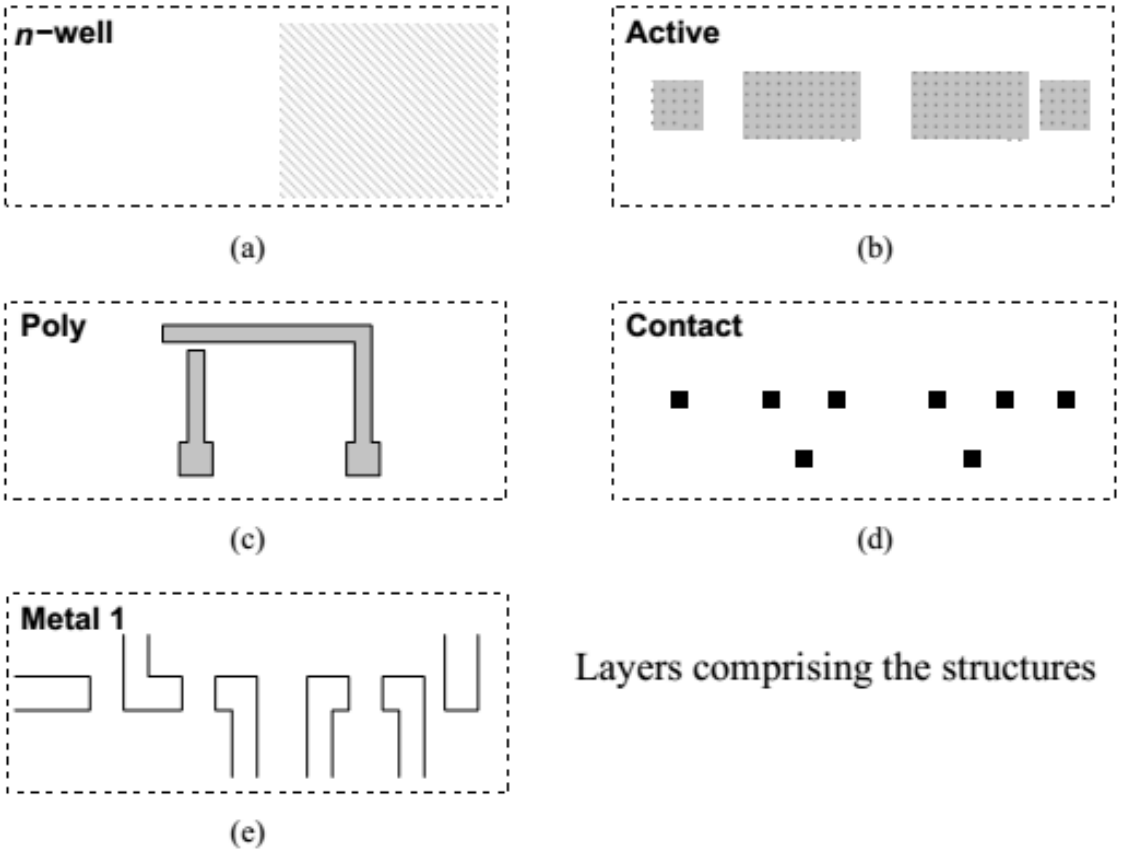
- In integrated circuits, the resistivity and thickness of the layers are set by fabrication materials
- and processing steps and cannot be changed in the layout

Wafer Processing

- The starting wafer in a CMOS technology must be created with a very high quality. That is, the wafer must be grown as a single-crystal silicon body having a very small number of "defects," e.g., dislocations in the crystal or unwanted impurities.
- This is accomplished by the "Czochralski method," whereby a seed of crystalline silicon is immersed in molten silicon and gradually pulled out while rotating. As a result, a large single-crystal cylindrical "ingot" is formed that can be sliced thin into wafers.

- Note that dopants are added to the molten silicon to obtain the desired resistivity.
- The wafers are then polished and chemically etched, thereby removing damages on the surface that are created during slicing.

Photolithography

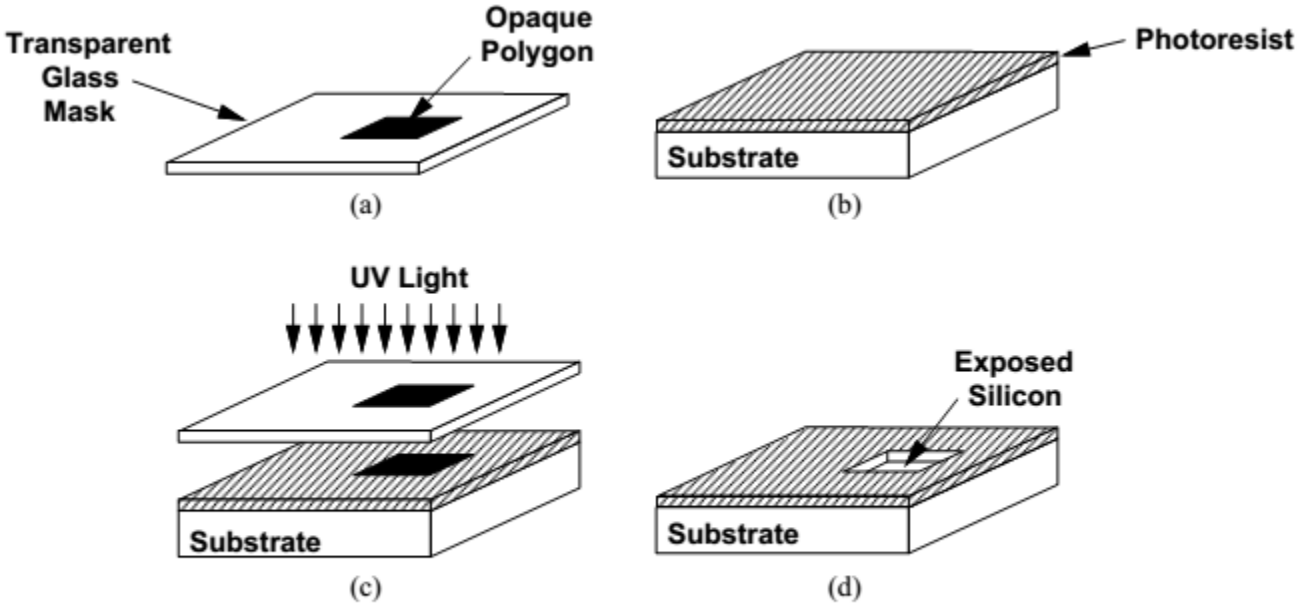


Layers comprising the structures

- **Photolithography, or simply lithography, is the first step in transferring the circuit layout information to the wafer.**

- **The figure shows the various structures comprising the layers namely, n-wells, S/D regions, contacts, polysilicon and metal interconnects. For fabrication purposes, we decompose the layout into these layers.**

Lithography Sequences



(a) Glass mask used in lithography, (b) coverage of wafer by photoresist, (c) selective exposure of photoresist to UV light, (c) exposed silicon after etching.

- **The sequence associated with the lithography of each layer involves one mask and three processing steps:**

  (1) cover wafer with photoresist

  (2) align mask on top and expose to light

  (3) etch exposed photoresist.

Lithography Sequences

- Two types of photoresists are used in processing.

1. A "negative" photoresist hardens in the areas exposed to light and
2. A "positive" photoresist hardens in the areas not exposed to light.

- The number of masks in a process heavily impacts the overall cost of fabrication, eventually influencing the unit price of the chip

- In modern CMOS processes this number is around 30, the cost of each IC has nonetheless remained low because both the number of transistors per unit area and the size of the wafer have steadily increased.

Oxidation

- A unique property of silicon is that it can produce a very uniform oxide layer on the surface with little strain in the lattice, allowing the fabrication of gate oxide layers as thin as a few tens of angstroms (only several atomic layers).

- In areas between the devices, a thick layer of SiO2, called the “field oxide” (FOX) is grown, providing the foundation for interconnect lines that are formed in subsequent steps

Field oxide.

Oxidation

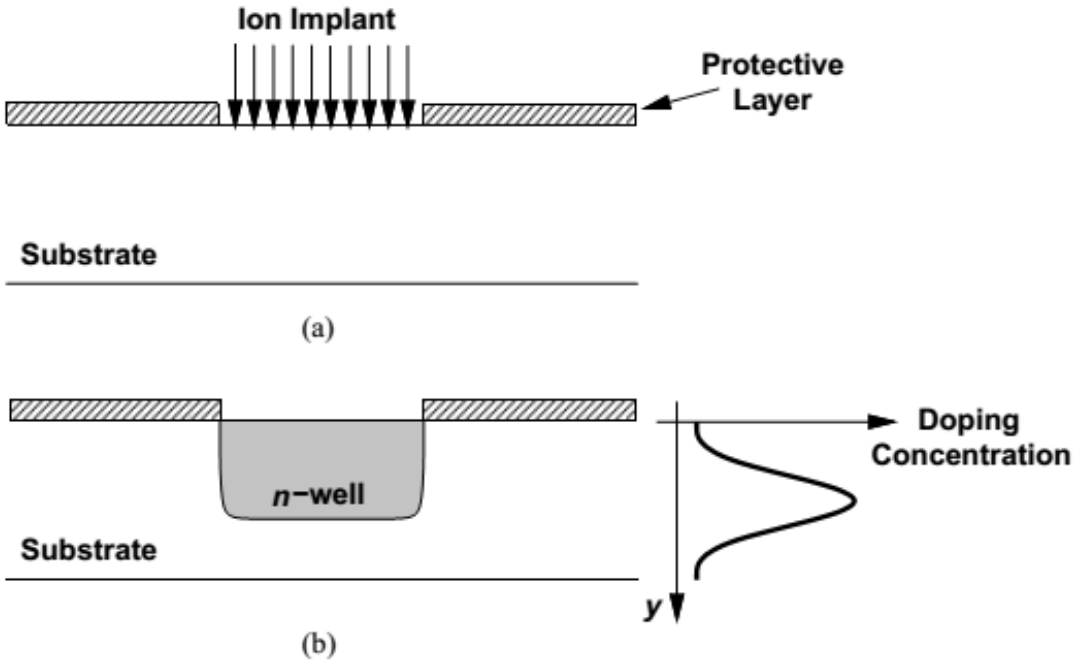
- Silicon dioxide is “grown” by placing the exposed silicon in an oxidizing atmosphere such as oxygen at a temperature around 1000^0C.

- Since the oxide thickness, t_{ox}, determines both the current handling and reliability of the transistors, it must be controlled to within a few percent.

- The “cleanness” of the silicon surface under the oxide affects the mobility of the charge carriers and thus the current drive, transconductance, and noise of the transistors.

Ion Implantation

- In many steps of fabrication, dopants must be selectively introduced into the wafer.
- The most common method of introducing dopants is "ion implantation," whereby the doping atoms are accelerated as a high-energy focused beam, hitting the surface of the wafer and penetrating the exposed areas



(a) Ion implantation, (b) retrograde profile.

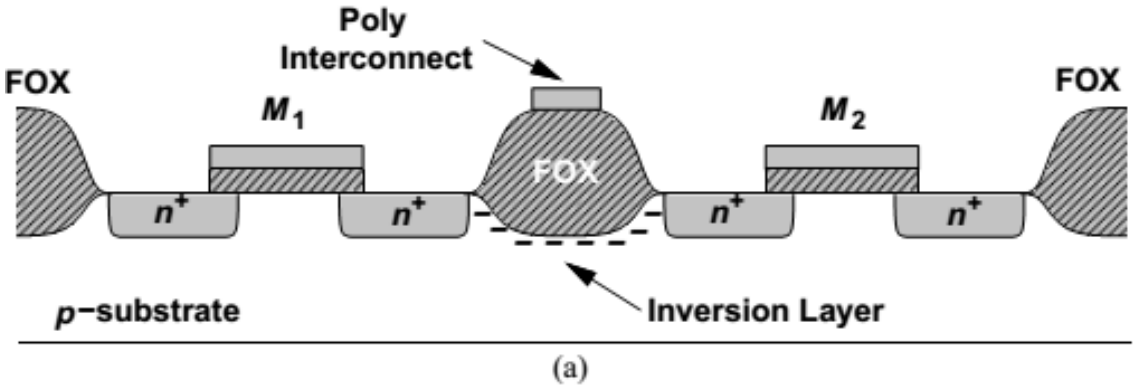
Ion Implantation

- The doping level (dosage) is determined by the intensity and duration of the implantation, and the depth of the doped region is set by the energy of the beam.

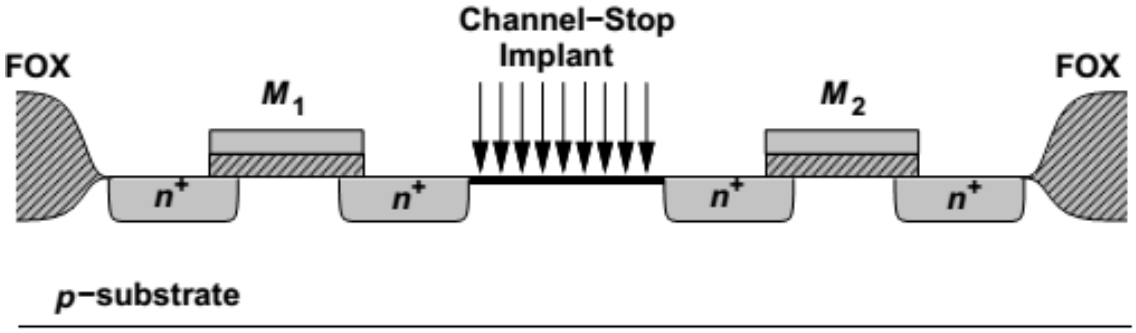
- With a high energy, the peak of the doping concentration in fact occurs well below the surface, thereby creating a "retrograde" profile.

- Such a profile is desirable for the n-well because it establishes a low resistivity near the bottom, reducing susceptibility to latch-up (to be discussed later), and a low doping level at the surface, decreasing the S/D junction capacitance of PMOS devices.

Channel-stop Implant.



(a)



(b)

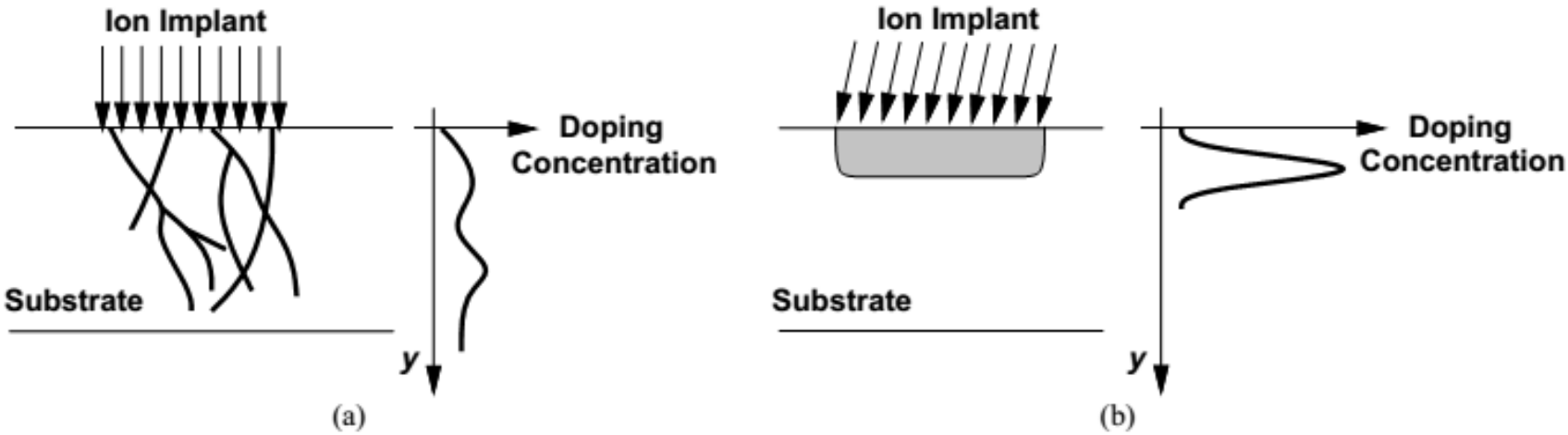
(a) Unwanted conduction due to inversion of field area, (b) channel-stop implant.

- **Another important application of implantation is to create "channel-stop" regions between transistors.**

Channel-stop Implant.

- **Interestingly, the two n+ regions and the FOX form a MOS transistor having a thick gate oxide (as in the figure)**
- **With a sufficiently positive potential on the interconnect line, this transistor may turn on slightly, creating a leakage path between M1 and M2.**
- **To resolve this issue, a channel-stop implant is performed before the field oxide deposition.**

Channeling



(a) Effect of channeling, (b) tilt in implant to avoid channeling.

- **An interesting phenomenon in ion implantation is "channeling." As shown in Figure, if the implant beam is aligned with the crystal axis, the ions penetrate the wafer to a great depth.**

- **For this reason, the implant (or the wafer) is tilted by 7–9^0, avoiding such an alignment and ensuring a predictable profile.**

Deposition

- As suggested by the structures in the layers, device fabrication requires the deposition of various materials.
- Examples include polysilicon, dielectric materials separating interconnect layers, and metal layers serving as interconnects.
- A common method of forming polysilicon on thick dielectric layers is “chemical vapor deposition” (CVD), whereby wafers are placed in a furnace filled with a gas that creates the desired material through a chemical reaction.
- In modern processes, CVD is performed at a low pressure to achieve more uniformity.

Etching

- The etching of the materials is also a crucial step. Structures with very small dimensions must be etched with high precision.

- Depending on the speed, accuracy, and selectivity required in the etching step, and the type of material to be etched, one of these methods may be used:
- (1)“wet” etching, i.e., placing the wafer in a chemical liquid (low precision)
- (2) “plasma” etching, i.e., bombarding the wafer with a plasma gas (high precision)
- (3) reactive ion etching (RIE), where ions produced in a gas bombard the wafer.

Device Fabrication

- With the processing operations described in the previous section, we now study the fabrication sequence and device structures in typical CMOS technologies.

- We consider three categories:
- Active devices –

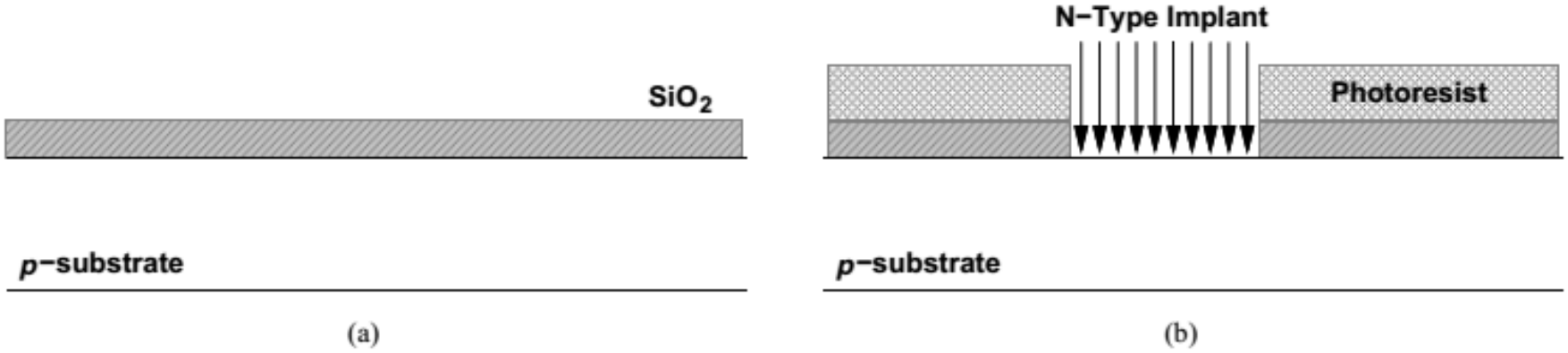
  We study the basic transistor fabrication and related back-end processing.

- Passive devices –

  CMOS resistors and capacitors.

- Interconnects.

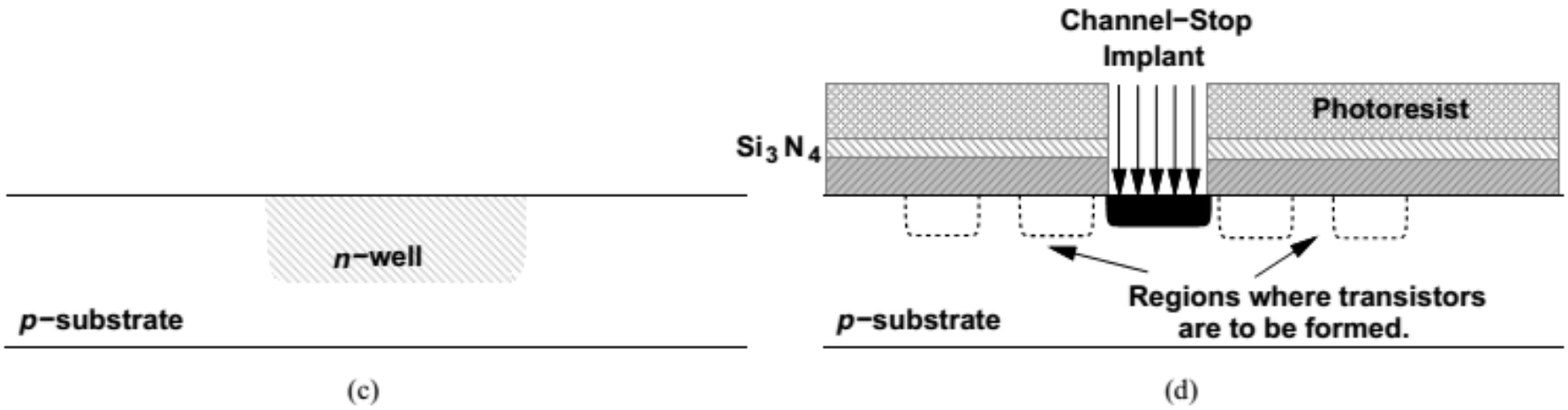
MOS Fabrication Sequence



(a) The fabrication begins with a p-type silicon wafer approximately 1 mm thick. Following cleaning and polishing steps, a thin layer of silicon dioxide is grown as a protective coating on top of the wafer.

(b) Next, to create then-wells, a lithography sequence consisting of photoresist deposition, exposure to UV light using the n-well mask, and selective etching is carried out and then-wells are implanted.

MOS Fabrication Sequence

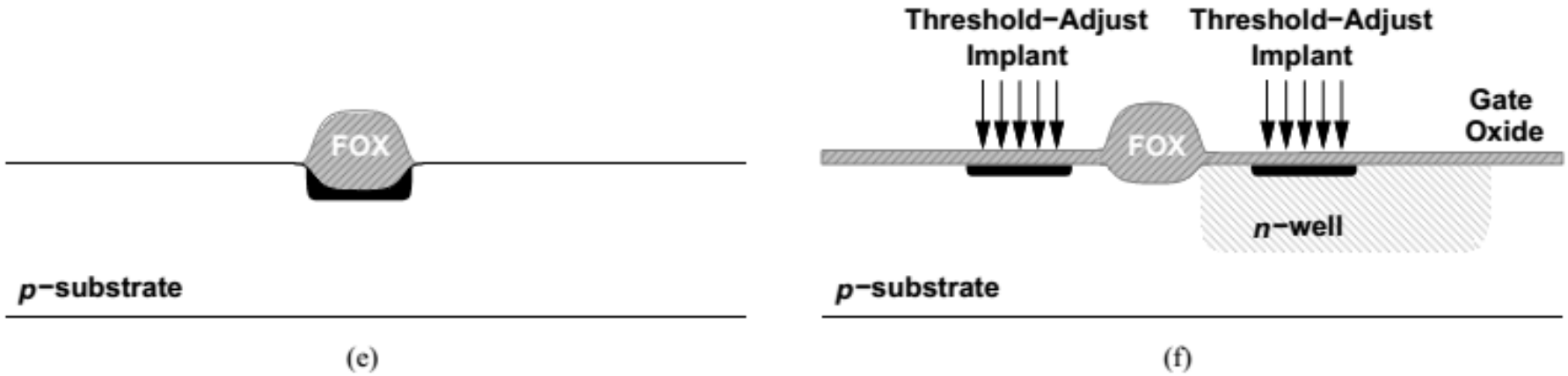


(c) The remaining photoresist and oxide layers are then removed in this step.

(d) At this point in the sequence, a stack consisting of a silicon oxide layer, a silicon nitride (Si3N4), and appositive photoresist layer is created.

Subsequently, the channel-stop implant is performed, the photoresist is removed, and a thick oxide layer is grown in the exposed silicon areas, producing the the field oxide

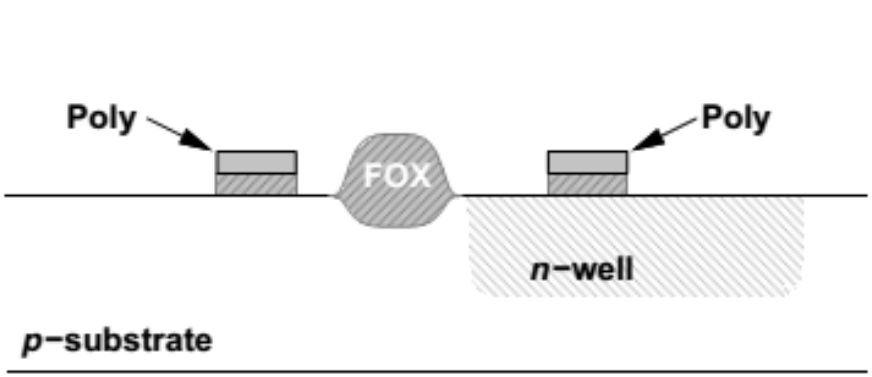
MOS Fabrication Sequence



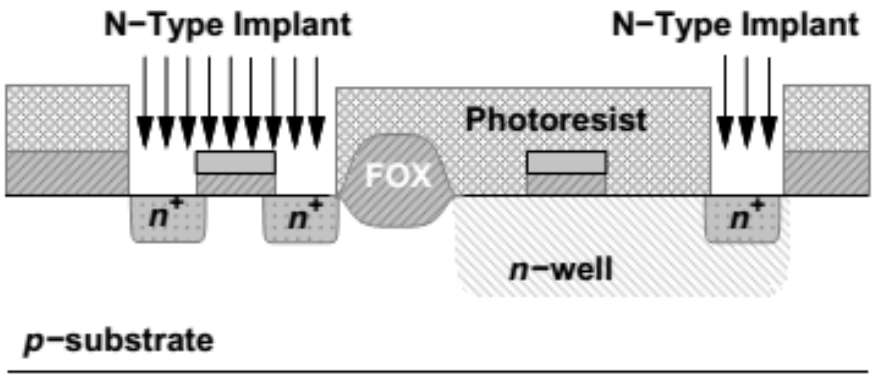
(e) The protective nitride and oxide layers are then removed thereby exposing all areas where transistors are to be formed.

(f) The next step involves the growth of the gate oxide, a critical operation requiring slow, low-pressure CVD

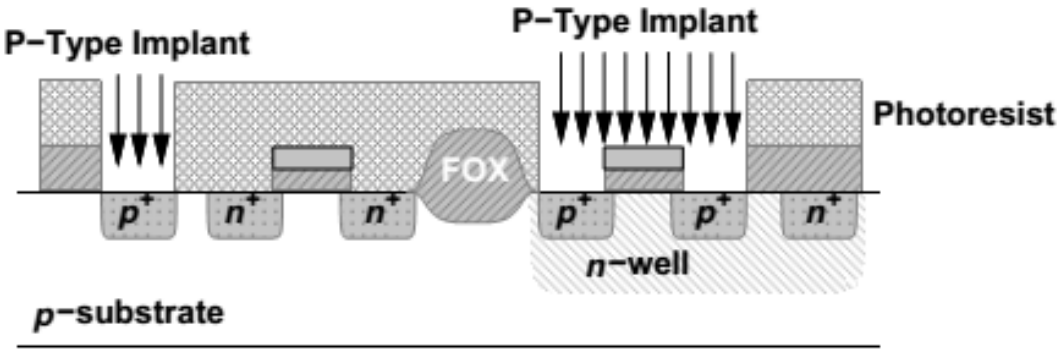
MOS Fabrication Sequence



(g)



(h)



(i)

(g) With the gate oxide in place, the polysilicon layer is deposited and the "poly mask" lithography is carried out, resulting in the structure shown in figure(g).

MOS Fabrication Sequence

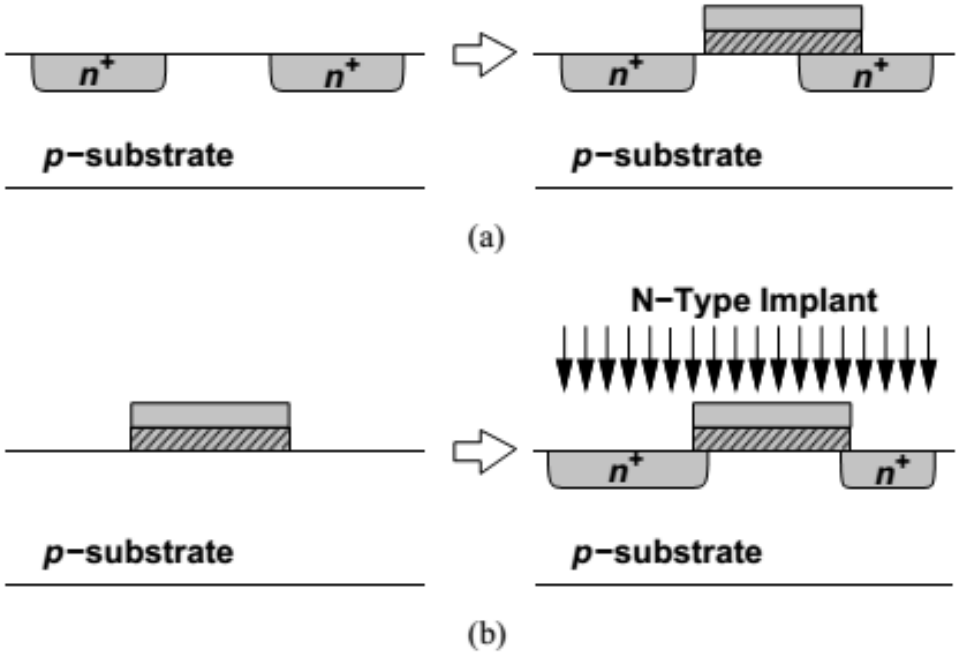
(h) and (i) In the next step, the source/drain junctions of the transistors and the substrate and n-well ties are formed by ion implantation.

This step requires a "source/drain mask" and two lithography sequences.

- **The first sequence incorporates a negative photoresist, exposing the areas to receive an n+ implant (the S/D junctions of NMOS transistors and the n-well ties).**

- **In the second sequence [Figure(i)], the same mask and a positive photoresist are used, exposing the areas to receive a p+ implant (the S/D junctions of PMOS transistors and the substrate ties)**
- **This step completes the fabrication of the basic transistors.**

MOS Fabrication Sequence



(a) Formation of n^+ regions before deposition of poly, (b) self-aligned structure.

- **The source/drain junctions are formed after the gate oxide and polysilicon for the following reason:**
- **Suppose, as depicted in Figure(a), these junctions are created first. Then, the alignment of the gate poly mask with respect to the S/D areas becomes extremely critical.**

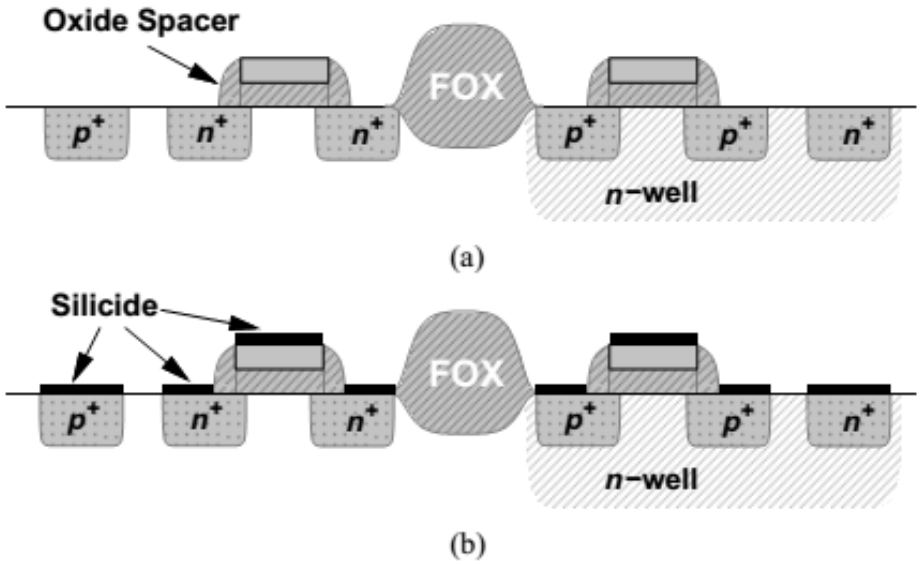
Back-End Processing

- With the basic transistors fabricated, the wafers must next undergo “back-end” processing, a sequence primarily providing various electrical connections on the chip through contacts and wires.

- The first step in this sequence is “silicidation”.

- Since the sheet resistance of doped polysilicon and S/D regions is typically several tens of ohms per square, it is desirable to reduce their resistance by about an order of magnitude.

Silicidation



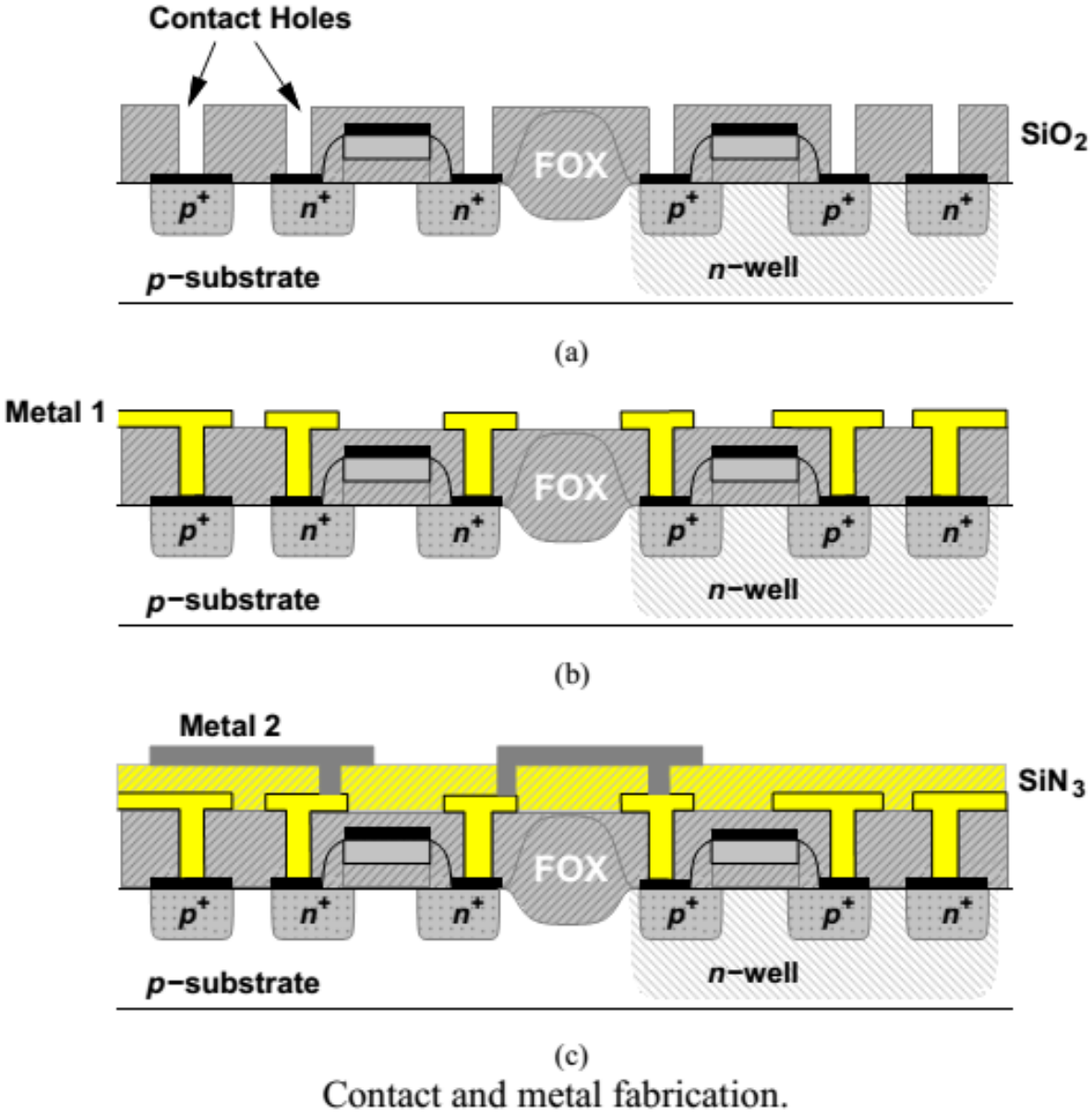
(a) Oxide spacers and (b) silicide.

- Silicidation accomplishes this(reduction in resistance) by covering the polysilicon layer and active areas (S/D regions and substrate and n-well ties) with a thin layer of a highly conductive material, e.g., titanium silicide or tungsten

- This step in fact begins with creating an “oxide spacer” at the edges of the polysilicon gate such that the deposition of the silicide becomes a self-aligned process as well.

- Without the spacer, the silicide layer on the gate may be shorted to that on the source/drain.

Contact and metal fabrication



Contact and metal fabrication.

- **The next step in back-end processing is to produce contact windows on top of polysilicon and active regions.**

Contact and metal fabrication

- This is carried out by first covering the wafer with a relatively thick (0.3- to 0.5-μm) layer of oxide and subsequently performing a lithography sequence using the "contact mask."
- The contact holes are then created by plasma etching [Figure(a)].

- Following contact windows, the first layer of metal interconnect (called "metal 1") (using aluminum or copper) is deposited over the entire wafer.
- A lithography sequence using the "metal 1 mask" is then carried out and the metal layer is selectively etched [Figure(b)].
- The higher levels of interconnect are fabricated using the same procedure [Figure(c)].

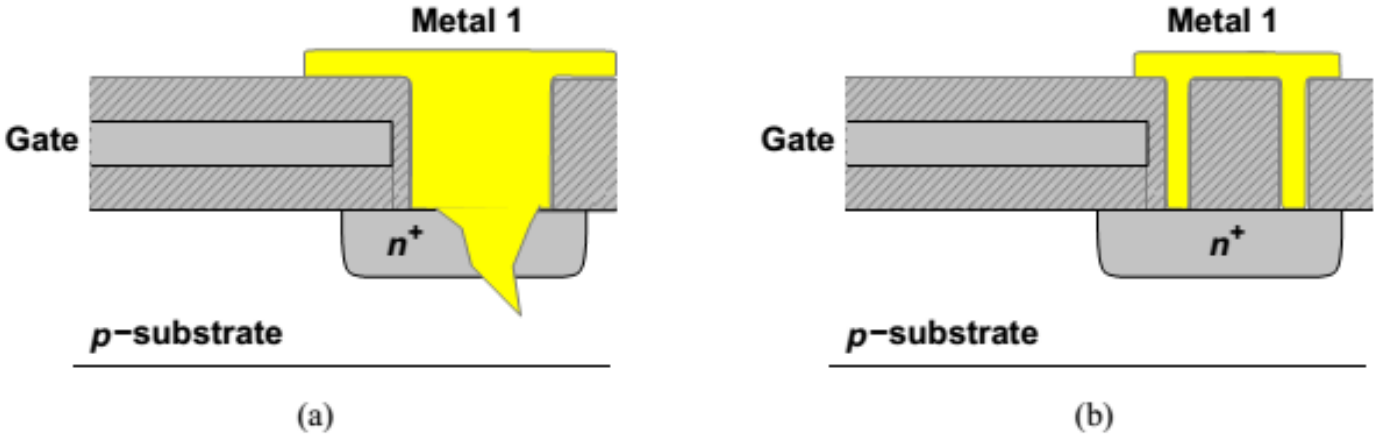
Contact and metal fabrication

- For each additional metal layer, two masks are required:
  one for the contact windows and another for the metal itself.

- Thus, a CMOS process having five layers of metal contains 10 masks for the back end.

- The contact windows between metal layers are sometimes called “vias” to distinguish them from the first level of contacts to active areas and polysilicon

Contact spiking



(a) Spiking due to large contact areas, (b) use of small contacts to avoid spiking.

- An interesting phenomenon related to large active areas is "contact spiking."
- If a large contact window allows aluminum to touch the active area, then, as depicted in figure(a), the metal may "eat" and penetrate the doped region, eventually crossing the junction to the bulk and shorting the diode.
- With small windows, on the other hand, this effect is avoided [Figure(b)].

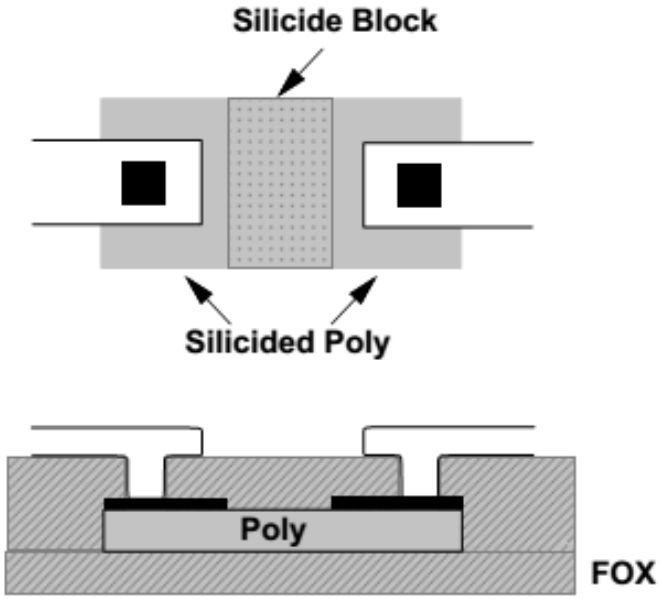
Passive Devices

- Passive components such as resistors and capacitors find wide usage in analog design, making it desirable to add these devices to standard CMOS technologies.

- In practice, however, CMOS processes target primarily digital applications and hence provide only NMOS and PMOS transistors.

- If a digital CMOS process is to be used for analog design, we must seek structures that can serve as passive components.
- The principal issue in using such structures is the variability of the component value from wafer to wafer because the process flow does not assume such structures are used in circuits.

Resistors

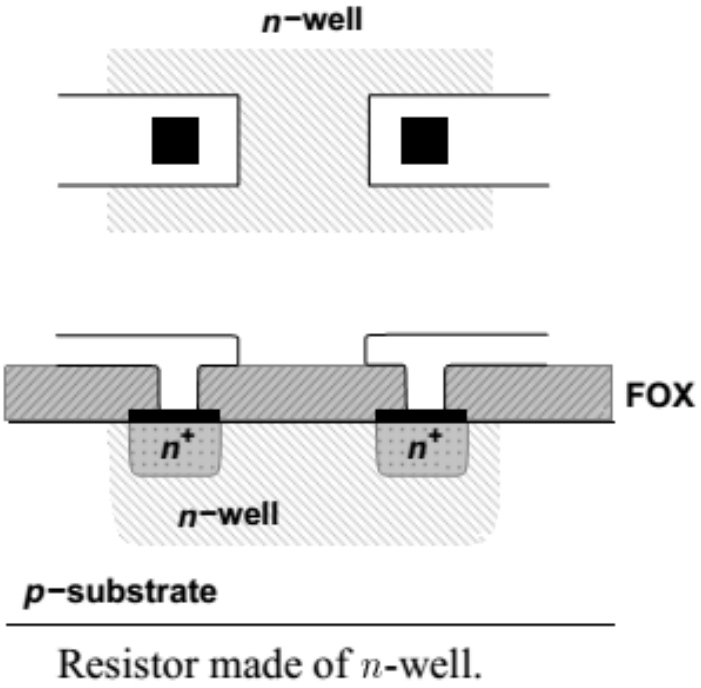


Poly resistor using silicide block.

- A CMOS process may be modified so as to provide resistors suited to analog design.
- A common method is to selectively "block" the silicide layer that is deposited on top of the polysilicon, thereby creating a region having the resistivity of the doped polysilicon.
- A The use of silicide on the two ends of the resistor in the figure results in a much lower contact resistance than that obtained by directly connecting the metal layer to doped polysilicon.

Resistors

- For a given resistance, poly resistors typically exhibit much less capacitance to the substrate than other types on the order of 90 af/μm2 for the bottom plate capacitance and 100 af/ μ m for the fringing capacitance.
- These resistors are quite linear, especially if they are long.



Resistor made of n-well.

- In a purely digital process, silicided poly, silicided p+ or n + active areas, n-well, and metal layers can be used as resistors. An n-well resistor can be formed as shown in the figure.
- But, the n-well resistivity may vary by several tens of percent with process.

Resistors

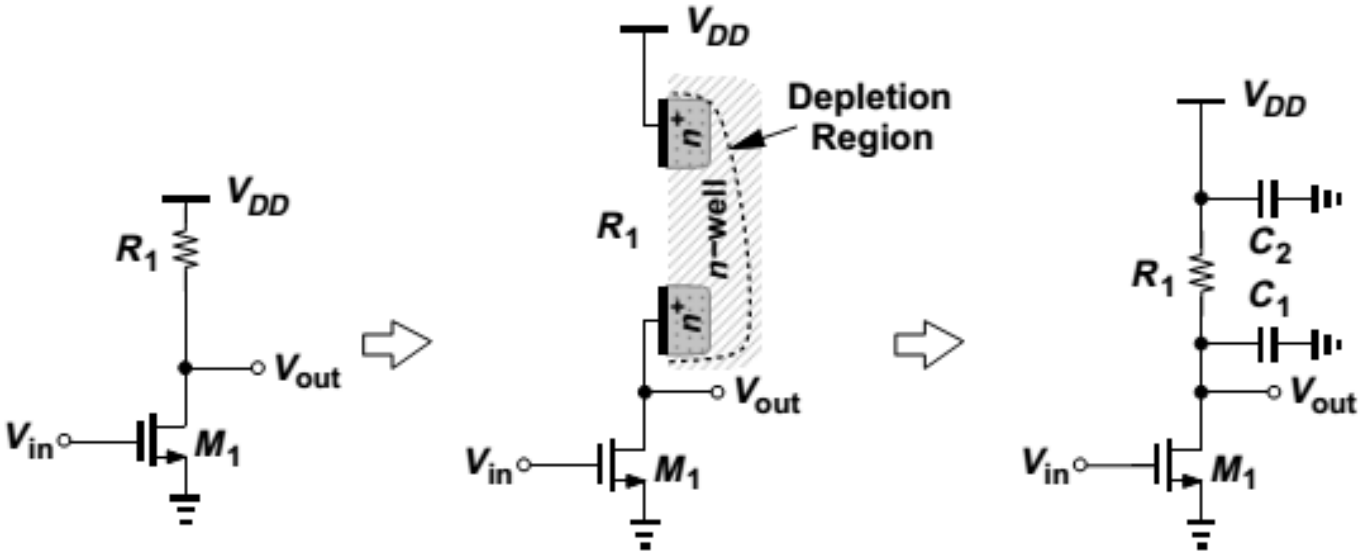


Figure 18.36. Common-source stage using n-well resistors.

- For example, the figure shows a common-source stage that is biased by means of M_0 and I_0 while employing C_1 to block the dc level of the preceding stage.
- In order to isolate the signal path from the low impedance (and the noise) introduced by M_0, resistor R_1 is inserted between X and Y. Here, the value of R_1 is not critical so long as it is sufficiently large.

Resistors

- Due to the depletion region formed between the n-well and the p-substrate, n-well resistors suffer from both a large parasitic capacitance and significant voltage dependence.

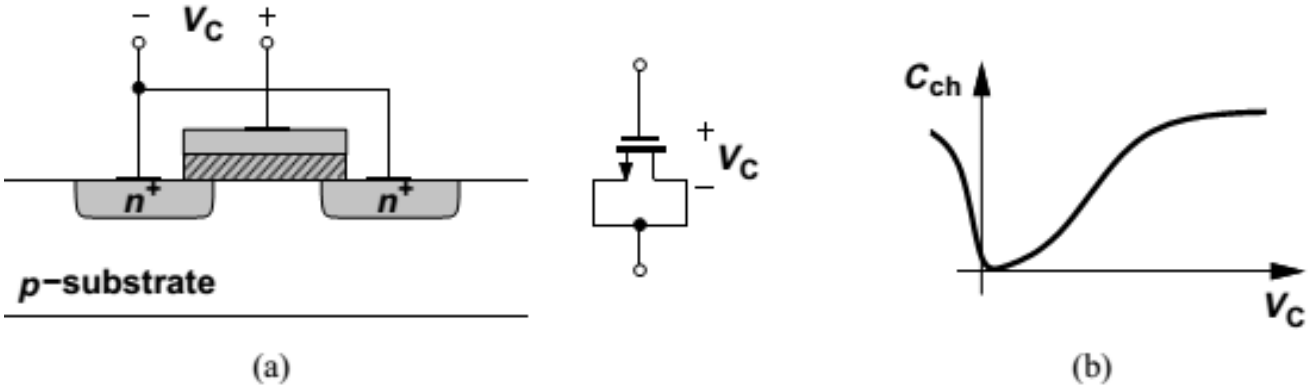
- Since the capacitance to the substrate is distributed (non uniformly) along the resistor, a lumped model may not be accurate enough, but as a rough approximation, we place half of the total capacitance on each side of the resistor.

- We also note that as V_{out} varies, so do the width of the depletion region and hence the value of the resistor.

Capacitors

- Capacitors prove indispensable in most of today's analog CMOS circuits.
- Several parameters of capacitors are critical in analog design: parasitic capacitance to the substrate, capacitance per unit area (density), and nonlinearity.

- Perhaps the simplest capacitor structure in CMOS technology is that implemented by a MOSFET, Illustrated in the figure(a).



(a) MOSFET configured as a capacitor, (b) nonlinear C/V characteristic.

Capacitors

- This device has a capacitance that varies from a small value at low voltages (where no channel exists and the equivalent capacitance is the series combination of the oxide capacitance and the depletion region capacitance)

  to a large value (C_{ox}) if the voltage difference exceeds V_{TH}.

- Since the gate oxide is typically the thinnest layer in the process, MOS capacitors biased in strong inversion are quite dense, saving substantial area if large values are required.

- For the same reason, the bottom-plate parasitic, i.e., that due to drain and source junctions is a relatively small percentage of the gate capacitance—typically 10 to 20%.

Interconnects

- The performance of today's complex integrated circuits heavily depends on the quality of the available interconnects, requiring more metal layers in new generations of the technology.

- Two properties of interconnects, namely, series resistance and parallel capacitance, impact the performance, often calling for iteration between layout and circuit design.

- The series resistance becomes especially problematic in supply and ground lines, creating dc and transient voltage drops.

- Also, for long signal lines, the distributed resistance and capacitance of the wire may result in a significant delay.

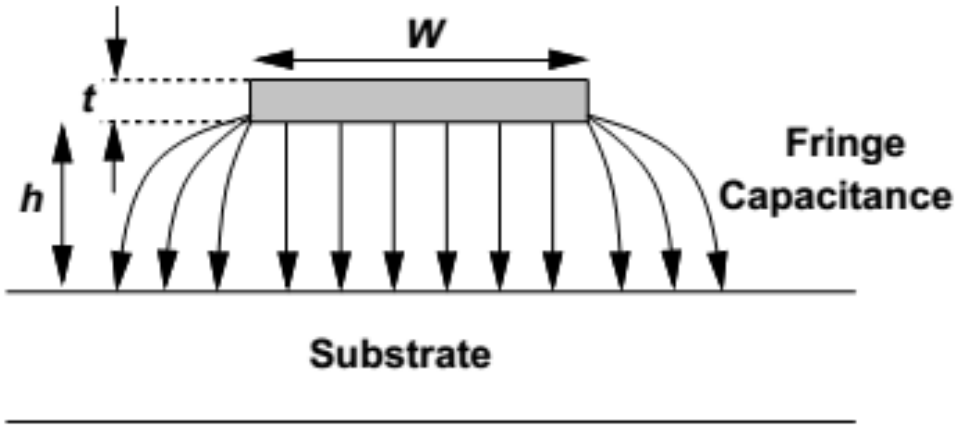
Interconnects - Resistance

- The finite resistance of wires influences the choice of line widths for high-current interconnects such as supply and ground buses.

- Another factor determining the width of interconnects is “electromigration”. At high current densities, the aluminum atoms in a wire tend to “migrate”, leaving a void that eventually (after some years of operation) grows to a discontinuity.
- For this reason, long-term reliability considerations restrict the maximum current density of interconnects.

- As a rule of thumb, a current-density of 2 mA per micron of width is acceptable, but the actual value varies according to the thickness of the metal.

Interconnects - Capacitance



Parallel-plate and fringe capacitance of an interconnect.

- **The problem of interconnect capacitance is much more complicated. We begin with a single wire on top of a substrate (as in the figure), identifying a "parallel-plate" capacitance and a "fringe" capacitance.**

- **For narrow lines, the two are comparable.**

Interconnects - Capacitance

- A simple empirical relationship for calculating the total wire capacitance per unit length on top of a conducting substrate is:

$$C = \epsilon \left[\frac{W}{h} + 0.77 + 1.06 \left(\frac{W}{h} \right)^{0.25} + 1.06 \left(\frac{t}{h} \right)^{0.5} \right],$$

Where W, h and t denote the dimensions

- For typical dimensions, this equation predicts the capacitance with a few percent of error.

- While upper levels of metal in a process exhibit less capacitance per unit width and length, their minimum allowable width is usually greater than that of the lower layers.
- Thus, the minimum capacitance for a given length may be only slightly smaller for the topmost layer(s).

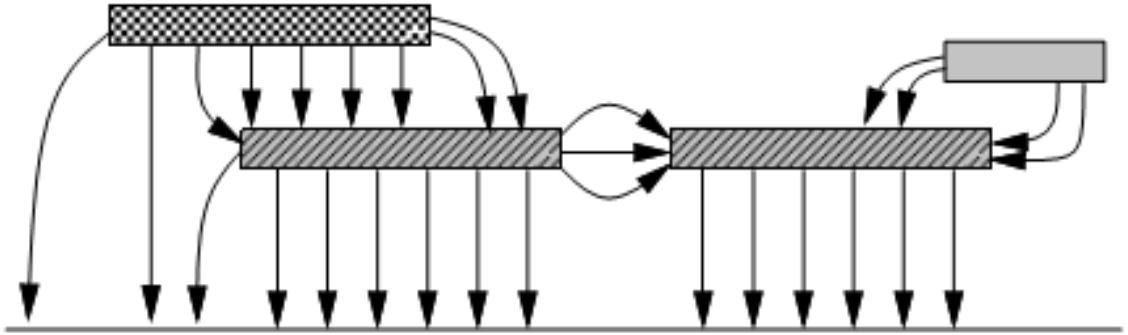
Interconnects - Capacitance

	Poly	Metal 1	Metal 2	Metal 3	Metal 4
Minimum Width (μm)	0.25	0.35	0.45	0.50	0.60
Bottom-Plate Capacitance (aF/μm^2)	90	30	15	9.0	7.0
Fringe Capacitance (Two Sides) (aF/μm)	110	80	50	40	30

Minimum widths and capacitances of interconnects in a 0.25-μm technology.

- **Table depicts typical values of minimum widths and parallel-plate and fringe capacitances (to the substrate) in a four-metal 0.25-μm process.**

Interconnects - Capacitance

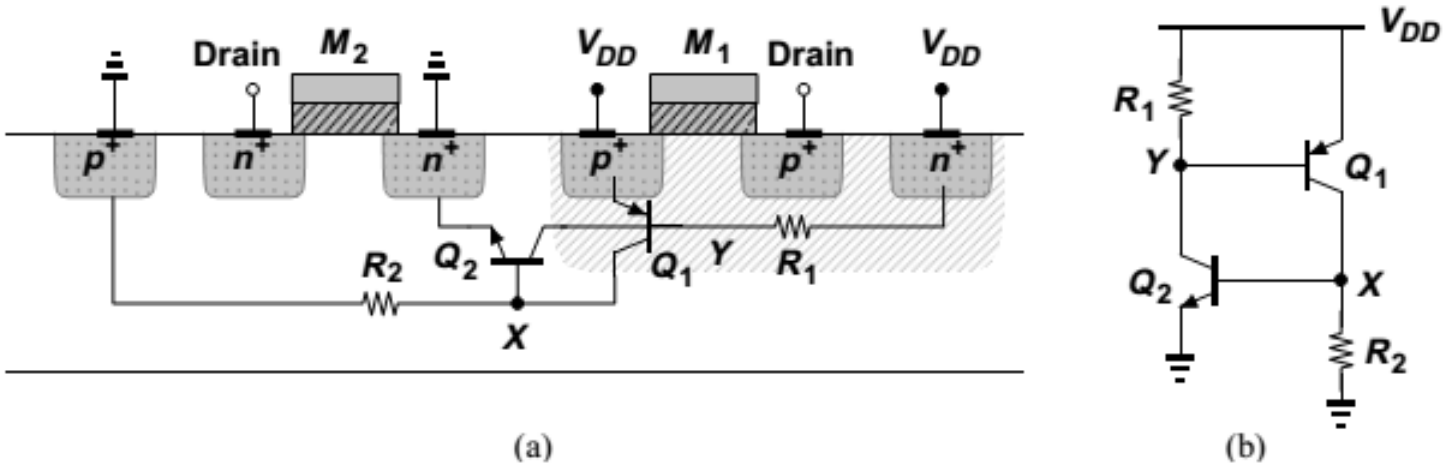


Complex interconnect structure.

- **Wires also suffer from parallel and fringe capacitances between them. Illustrated in the figure, this effect is difficult to quantify for a complex layout, often necessitating the use of computer programs.**

- **In practice, the capacitances between the layers are calculated by "electromagnetic field solvers," measured experimentally, and tabulated in the process design manual.**

Latch-Up



(a) Parasitic bipolar transistors in a CMOS process, (b) equivalent circuit.

- **Another issue that did not exist in NMOS implementations but arose in CMOS circuits was latch-up.**
- **We make two observations:**

(a) the base of each bipolar transistor is inevitably tied to the collector of the other.

(b) owing to the finite resistance of the n-well and the substrate, the bases of Q_1 and Q_2 see a nonzero resistance to V_{DD} and ground, respectively.

Latch-Up

- The parasitic circuit can therefore be drawn as in Fig(b), revealing a positive feedback loop around Q_1 and Q_2.
- In fact, if a current is injected into node X such that V_X rises, then I_{C2} increases, V_Y falls, I_{C1} increases, and V_X rises further.

- If the loop gain is greater than or equal to unity, this phenomenon continues until both transistors turn on completely, drawing an enormous current from V_{DD}. We say the circuit is latched up.
- The initial current required to trigger latch-up may be produced by various sources in an integrated circuit.

- In the circuit [fig (b)] a large voltage swing at the drains can therefore inject a significant displacement current into the n-well or the substrate, initiating latch-up.

Latch-Up

- A common case of latch-up occurs with the use of large digital output buffers (inverters).
- These circuits inject high currents into the substrate through the large drain junction capacitance of the transistors and by forward-biasing the source-bulk junction diodes.
- In order to prevent latch-up, both process engineers and circuit designers take precautions such that the loop gain of the equivalent circuit shown in the figure earlier, remains well below unity.