LOVELY
PROFESSIONAL
UNIVERSITY

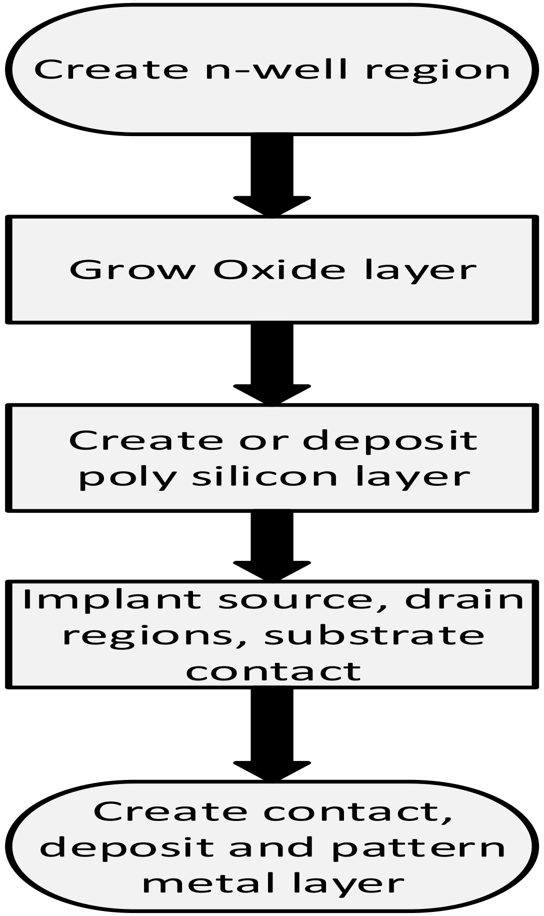
ECE318:CMOS VLSI Design

Unit 2

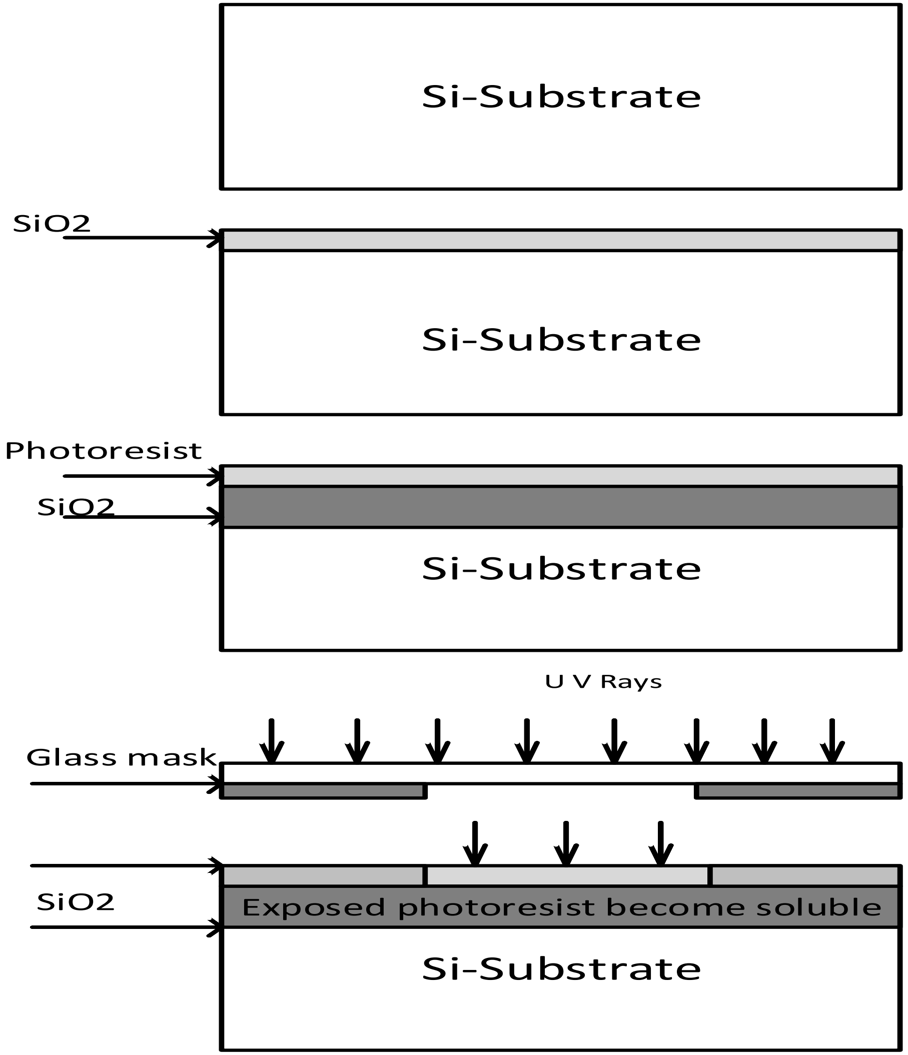
Fabrication of MOSFET and Scaling

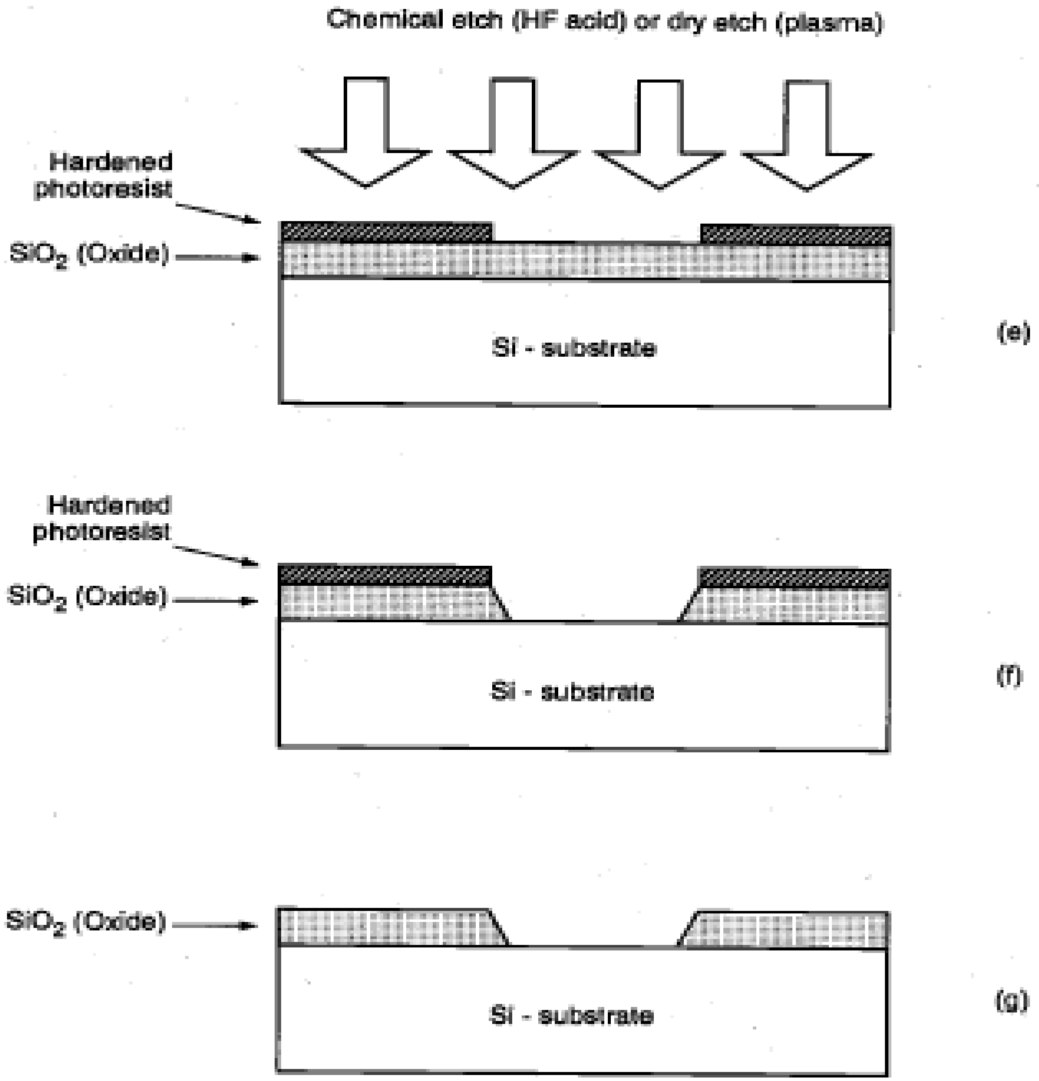
Fabrication process flow

- Creation of the n-well regions for pMOS transistors, by impurity implantation into the substrate.
- Then, a thick oxide is grown in the regions surrounding the nMOS and pMOS *active regions.*
- The thin gate oxide is subsequently grown on the surface through thermal oxidation.
- These steps are followed by the creation of n+ and p+ regions (source, drain, and channel stop implants)
- final metallization (creation of metal interconnects).


Create n-well region
Grow Oxide layer
Create or deposit poly silicon layer
Implant source, drain regions, substrate contact
Create contact, deposit and pattern metal layer

Process steps required for patterning of silicon dioxide





Fabrication of nMOS transistor

- Oxidation of Silicon Substrate (Fig.b)
- Field oxide is selectively etched to expose the silicon surface on which the MOS transistor will be created (Fig.c)
- On top of the thin oxide layer, a layer of polysilicon (polycrystalline silicon) is deposited (Fig.d)
- Polysilicon is used both as gate electrode material for MOS transistors and also as an interconnect medium in silicon integrated circuits.
- Undoped polysilicon has relatively high resistivity.
- The-resistivity of polysilicon can be reduced, however, by doping it with impurity atoms.

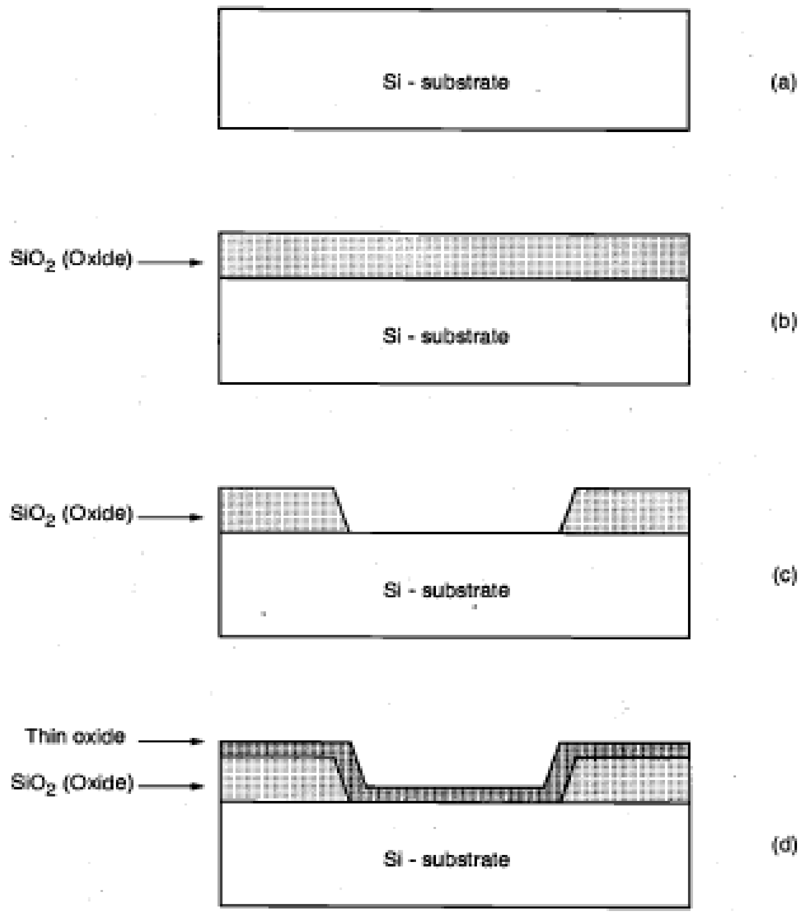
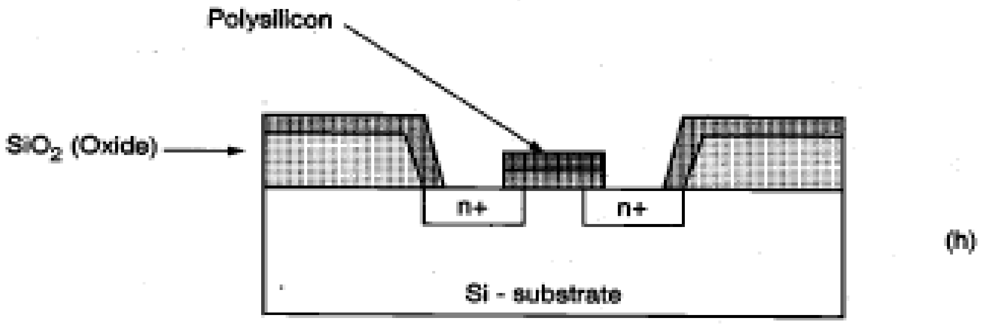
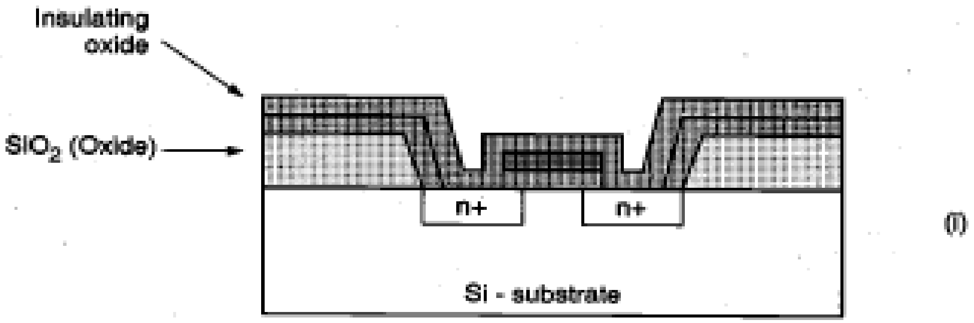


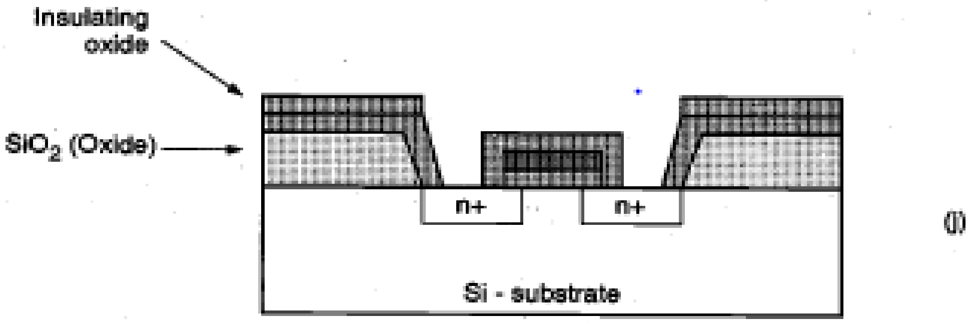
Fig. Process flow for the fabrication of an n-type MOSFET on p-type silicon.

Fabrication of nMOS transistor(Contd.)

- ➢ Polysilicon is also etched away, which exposes the bare silicon surface on which the source and drain junctions are to be formed
- ➢ The entire silicon surface is then with a high concentration of impurities, either through diffusion or ion implantation (in this case with donor atoms to produce n-type doping) (Fig. h)
- ➢ The impurity doping also penetrates the polysilicon on the surface, reducing its resistivity
- ➢ Once the source and drain regions are completed, the entire surface is again covered with an insulating layer of silicon dioxide (f)
- ➢ The insulating oxide layer is then patterned in order to provide contact windows for the drain and source junctions (Fig.j).
- ➢ The surface is covered with evaporated aluminum which will form the interconnects







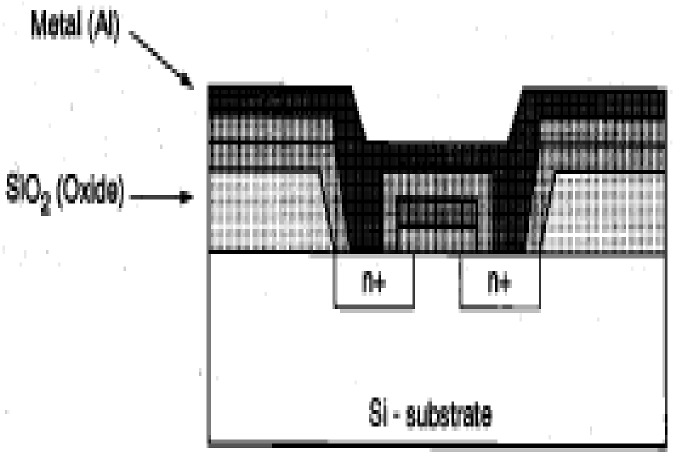
Fabrication of nMOS transistor(Contd.)

- Metal layer is patterned and etched, completing the interconnection of the MOS transistors on the surface.

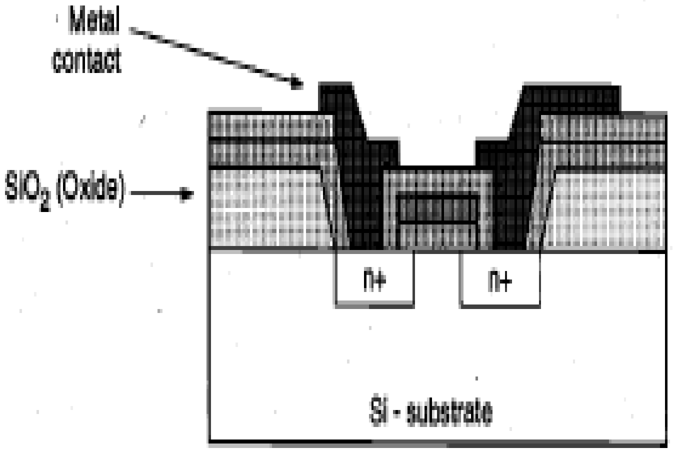
Device Isolation Techniques

The MOS transistors that comprise an integrated circuit must be electrically *isolated* from each other during fabrication.

- Isolation is required to prevent unwanted conduction paths between the devices, to avoid creation of inversion layers outside the channel regions of transistors, and to reduce leakage currents.
- To achieve a sufficient level of electrical isolation between neighboring transistors on a chip surface, the devices are typically created in dedicated regions called *active areas*
- *Active area is surrounded by a relatively thick oxide barrier called the field oxide.*
- One possible technique to create isolated active areas on silicon surface is first to grow a thick field oxide over the entire surface of the chip, and then to selectively etch the oxide in certain regions, to define the active areas. This fabrication technique, called *etched field-oxide isolation*



(k)



(l)

Device Isolation Techniques

Etched field-oxide isolation technique

- Here, the field oxide is selectively etched away to expose the silicon surface on which the MOS transistor will be created.
- Although the technique is relatively straightforward, it also has some drawbacks.
- The most significant disadvantage is that the thickness of the field oxide leads to rather large oxide steps at the boundaries between active areas and isolation Fabrication (field) regions.
- When polysilicon and metal layers are deposited over such boundaries in of MOSFETs subsequent process steps, the sheer height difference at the boundary can cause cracking of deposited layers, leading to chip failure.
- The *local oxidation of silicon (LOCOS)* technique is based on the principle of selectively *growing* the field oxide in certain regions, instead of selectively etching away the active areas after oxide growth.
- Selective oxide growth is achieved by shielding the active areas with silicon nitride (Si3N4) during oxidation, which effectively inhibits oxide growth.

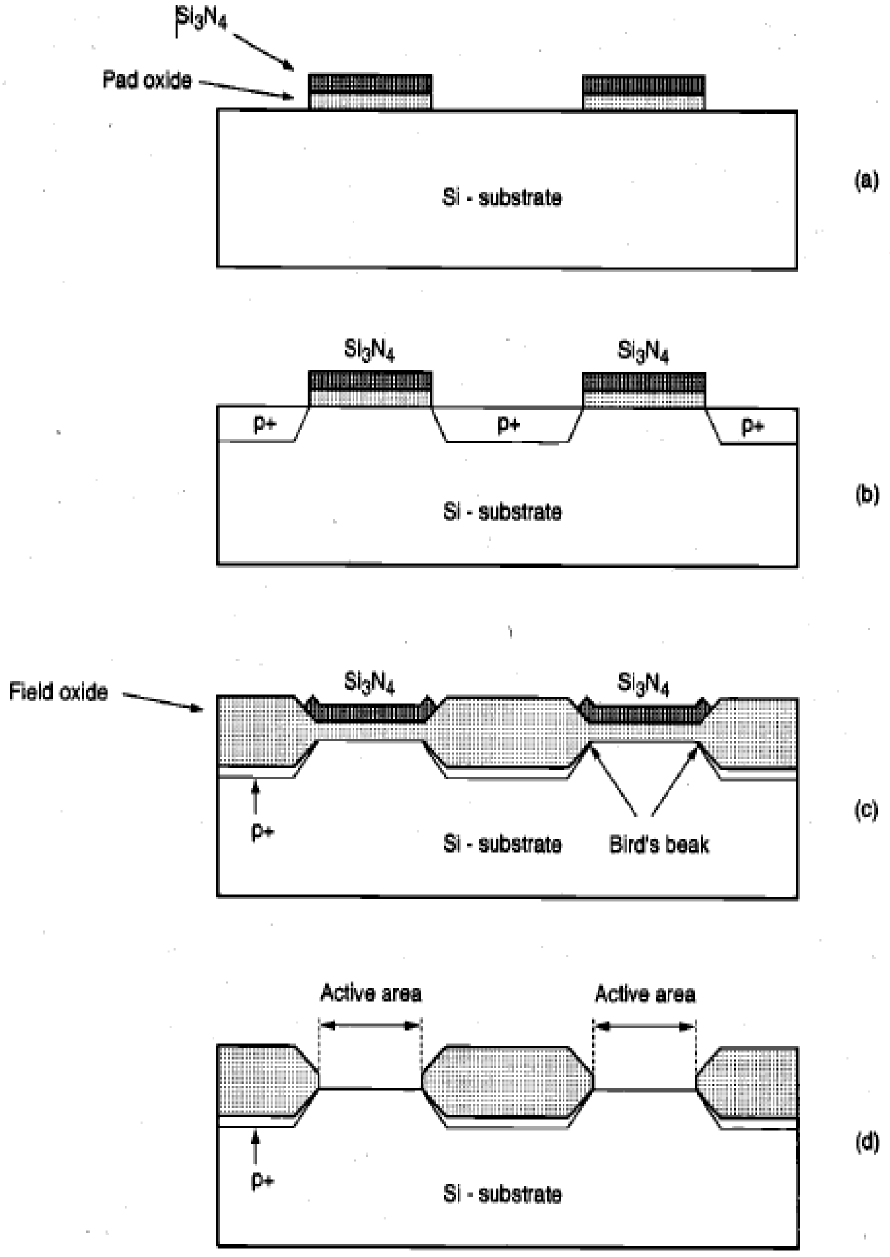
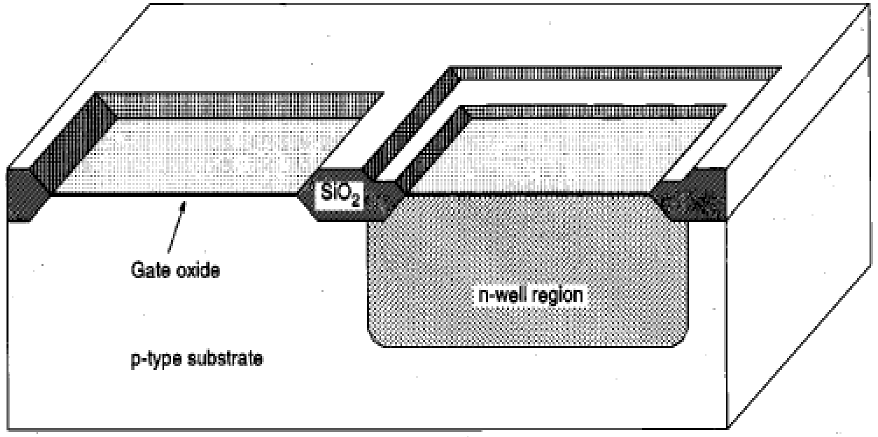
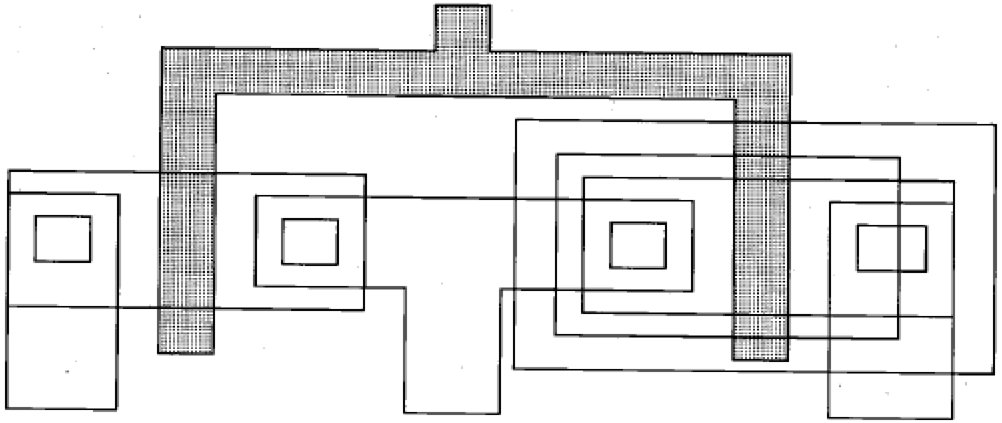


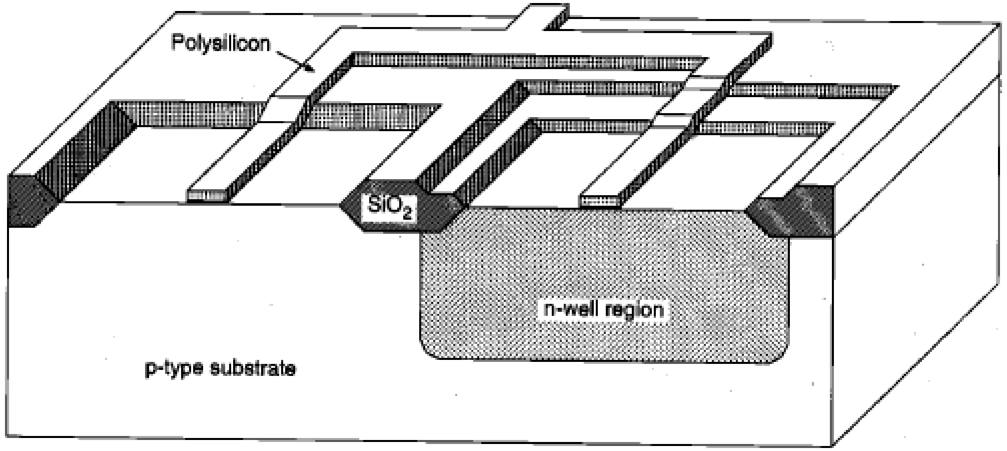
Fig. Basic steps of the LOCOS process to create oxide isolation around active areas

CMOS n-Well Process

- The n-well CMOS process starts with a moderately doped (with impurity concentration typically less than 1015 cm-3) p-type silicon substrate.
- Then, an initial oxide layer is grown on the entire surface.
- The first lithographic mask defines the n-well region. Donor atoms, usually phosphorus, are implanted through this window in the oxide.
- Once the n-well is created, the active areas of the nMOS and pMOS transistors can be defined.
- The polysilicon layer is deposited using chemical vapor deposition (CVD) and patterned by dry (plasma) etching.





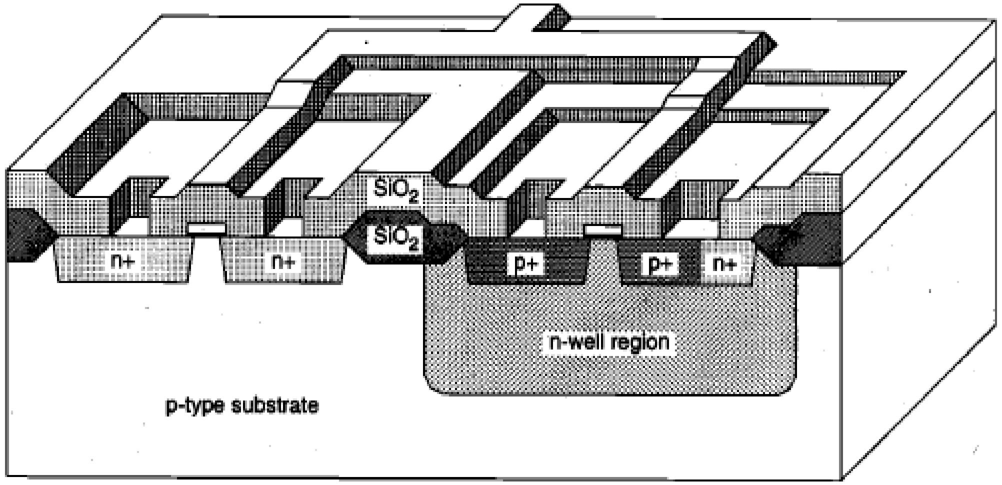


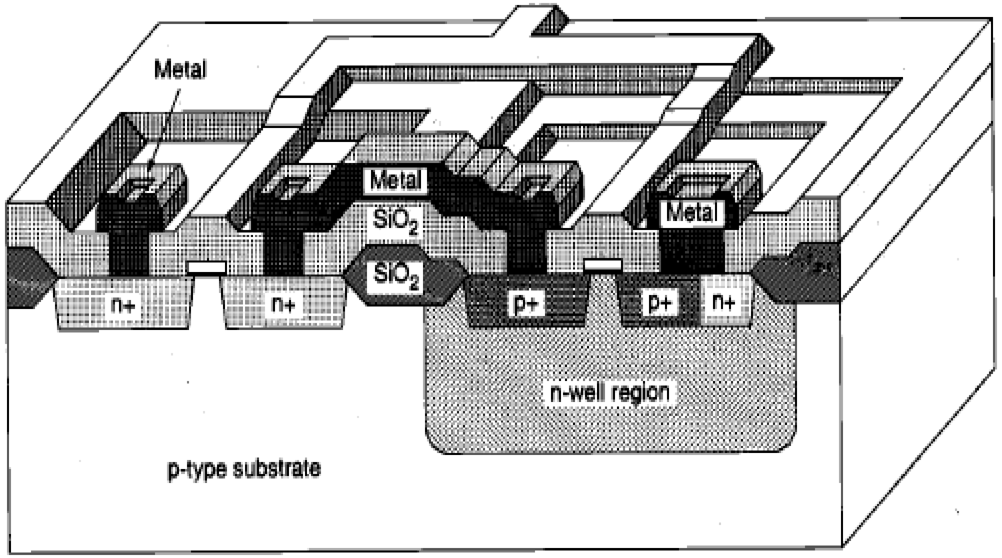
MCQ

Q1. Which layer is used for power and signal lines?
a) metal
b) polysilicon
c) n-diffusion
d) p-diffusion

CMOS n-Well Process

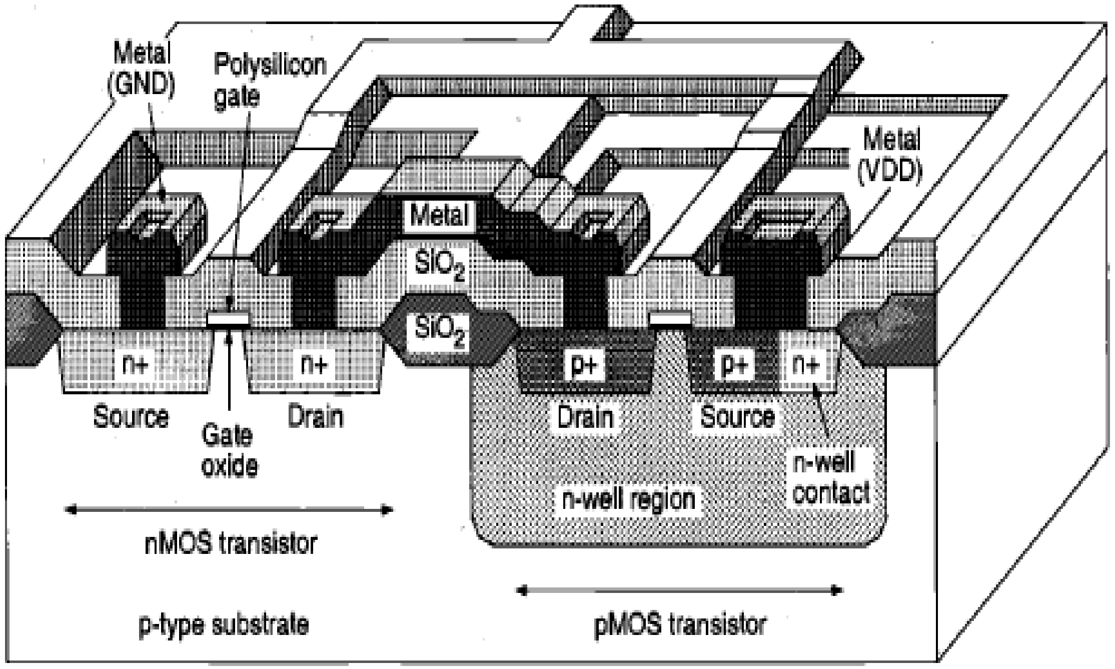
- The created polysilicon lines will function as the gate electrodes of the nMOS and the pMOS transistors and their interconnects.
- Using a set of two masks, the n+ and p+ regions are implanted into the substrate and into the n-well, respectively
- Metal (aluminum) is deposited over the entire chip surface using metal evaporation, and the metal lines are patterned through etching.
- Since the wafer surface is non-planar, the quality and the integrity of the metal lines created in this step are very critical and are ultimately essential for circuit reliability

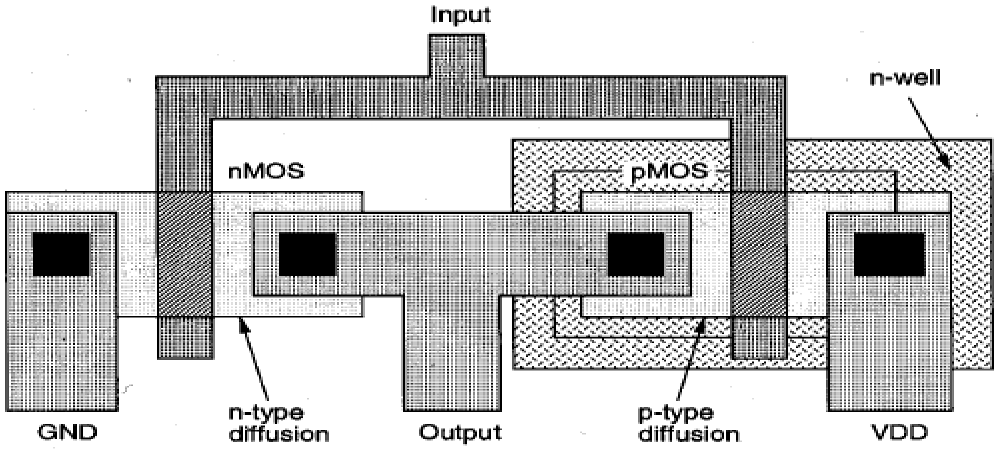




CMOS n-Well Process

- ➢ The composite layout and the resulting cross-sectional view of the chip, showing one nMOS and one pMOS transistor (in the n-well), and the polysilicon and metal interconnections.
- ➢ The final step is to deposit the passivation layer (for protection) over the chip, except over wire-bonding pad areas.





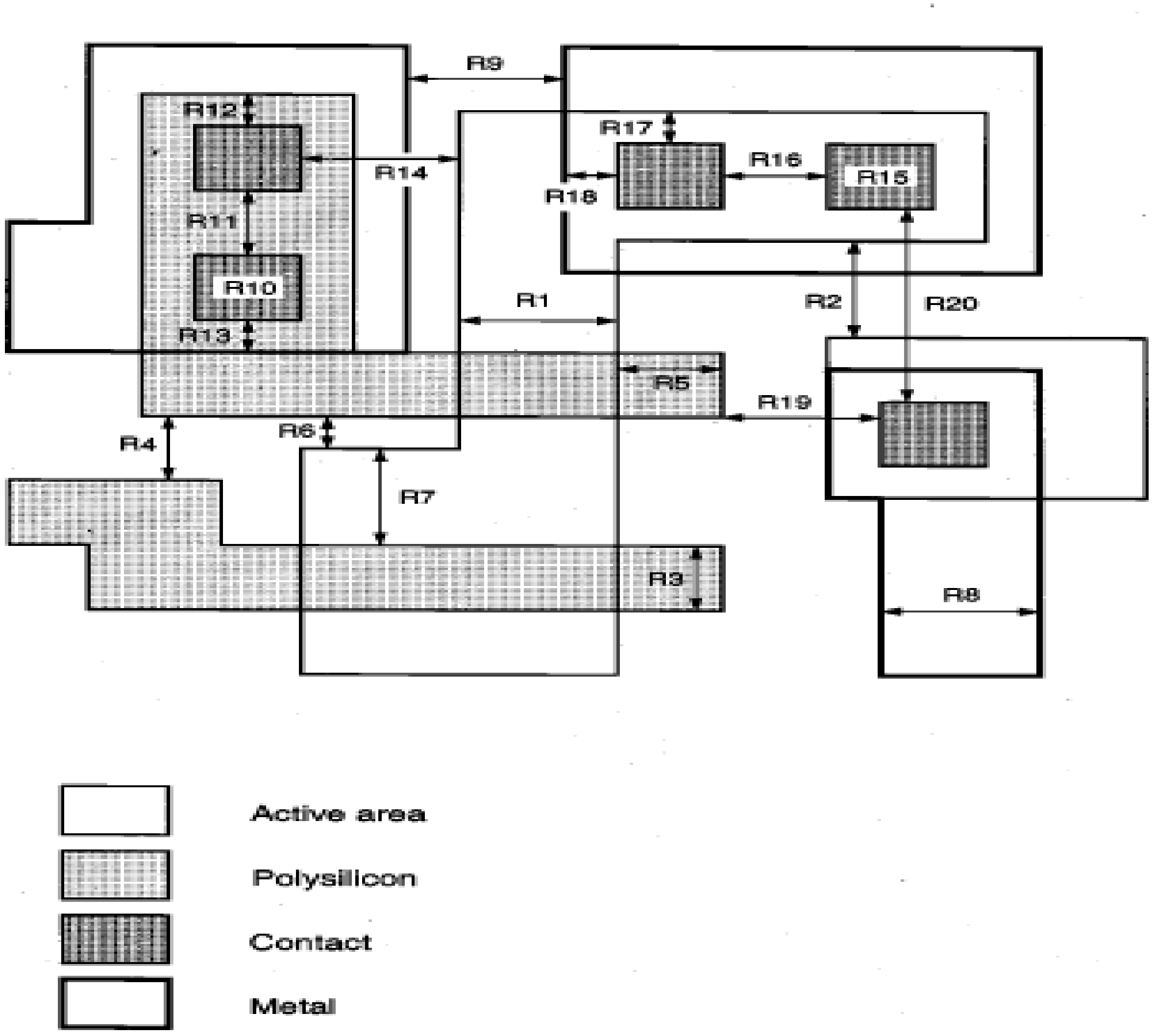
Layout Design Rules

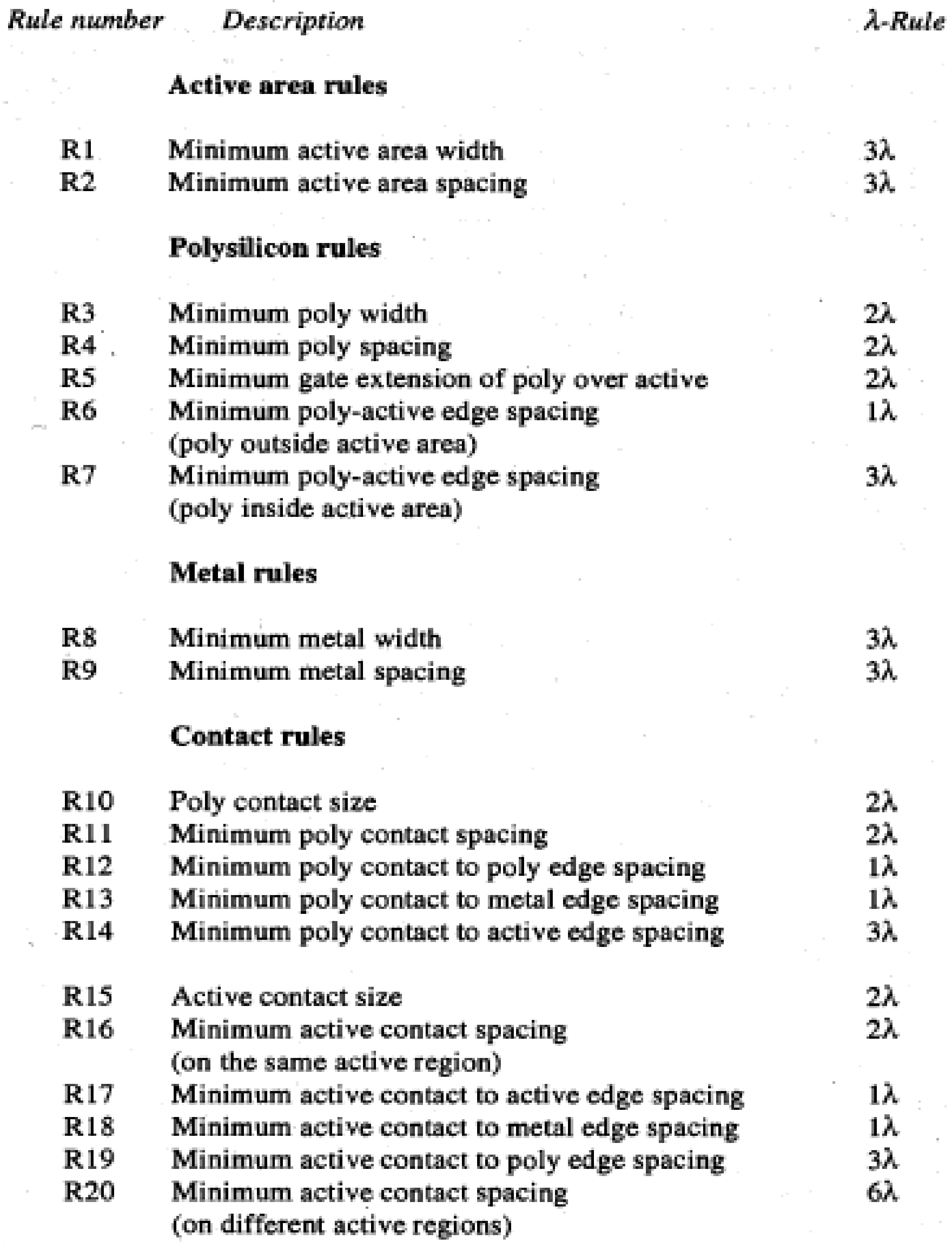
The design rules are usually described in two ways:

(i) Micron rules, in which the layout constraints such as minimum feature sizes and minimum allowable feature separations are stated in terms of absolute dimensions in micrometers, or,

(ii) Lambda rules, which specify the layout constraints in terms of a single parameter (X) and thus allow linear, proportional scaling of all geometrical constraints.

Lambda-based layout design rules devised for the MOSIS (MOS Implementation System)





Rule number	*Description*	*λ-Rule*
	Active area rules	
R1	Minimum active area width	3λ
R2	Minimum active area spacing	3λ
	Polysilicon rules	
R3	Minimum poly width	2λ
R4	Minimum poly spacing	2λ
R5	Minimum gate extension of poly over active	2λ
R6	Minimum poly-active edge spacing (poly outside active area)	1λ
R7	Minimum poly-active edge spacing (poly inside active area)	3λ
	Metal rules	
R8	Minimum metal width	3λ
R9	Minimum metal spacing	3λ
	Contact rules	
R10	Poly contact size	2λ
R11	Minimum poly contact spacing	2λ
R12	Minimum poly contact to poly edge spacing	1λ
R13	Minimum poly contact to metal edge spacing	1λ
R14	Minimum poly contact to active edge spacing	3λ
R15	Active contact size	2λ
R16	Minimum active contact spacing (on the same active region)	2λ
R17	Minimum active contact to active edge spacing	1λ
R18	Minimum active contact to metal edge spacing	1λ
R19	Minimum active contact to poly edge spacing	3λ
R20	Minimum active contact spacing (on different active regions)	6λ

MCQ’s

Q1. Design rules does not specify ___________

a) linewidths

b) separations

c) extensions

d) colours

Q2. The width of n-diffusion and p-diffusion layer should be?

a) 3λ

b) 2λ

c) λ

d) 4λ

Q3. What should be the spacing between two diffusion layers?

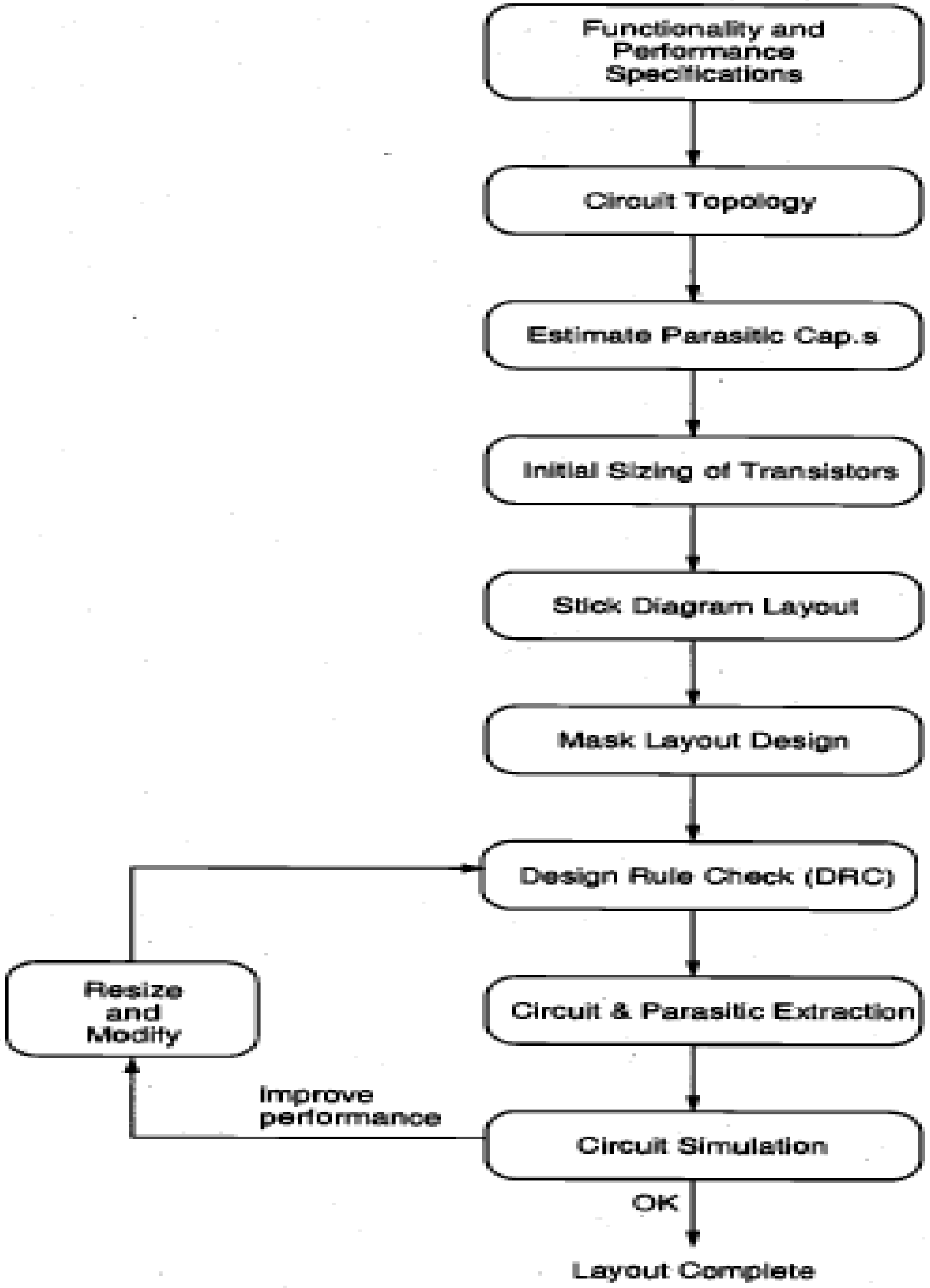
a) 4λ

b) λ

c) 3λ

d) 2λ

Typical design flow for the production of a mask layout

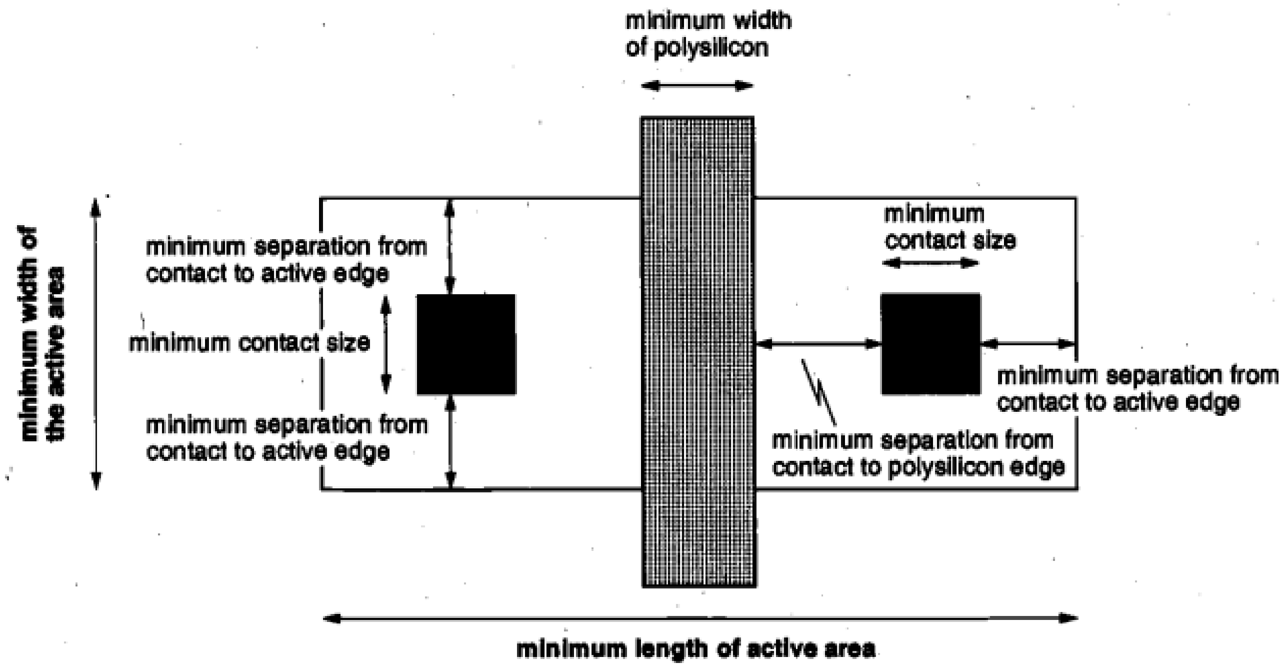


Design rules which determine the dimensions of a minimum-size transistor

For the minimum diffusion contact size (which is necessary for source and drain connections) and the minimum separation from diffusion contact to both active area edges.

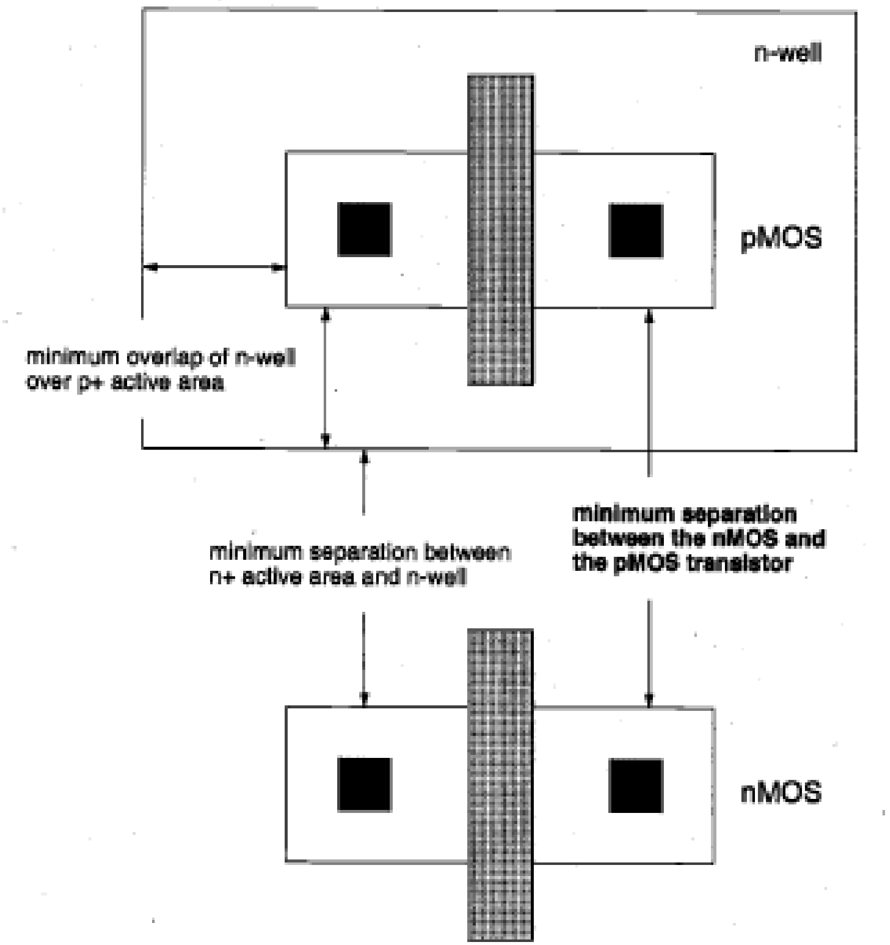
Width of the polysilicon line over the active area (which is the gate of the transistor) is typically taken as the minimum poly width (Fig. 2.14).

Minimum overall length of the active area: (minimum polysilicon width) + 2 x (minimum poly-to-contact spacing) + 2 x (minimum contact size) + 2 x (minimum spacing from contact to active area edge).


minimum width of polysilicon
minimum width of the active area
minimum separation from contact to active edge
minimum contact size
minimum separation from contact to active edge
minimum contact size
minimum separation from contact to active edge
minimum separation from contact to polysilicon edge
minimum length of active area

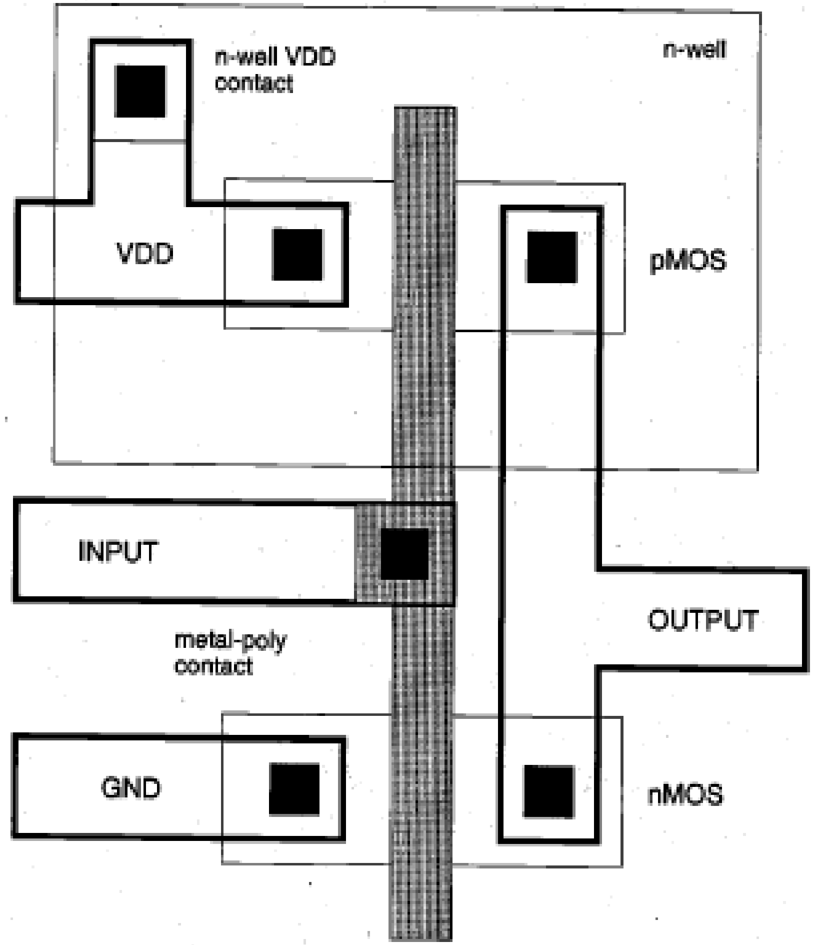
Design rules which determine the separation between the nMOS and the pMOS transistor of the CMOS inverter.

- Polysilicon gates of the nMOS and the pMOS transistors are usually aligned, so that the gate connections can be made with a single polysilicon line of least possible length.
- Reason for avoiding long polysilicon connections (as a general layout practice) is the fact that the large parasitic resistance and the parasitic capacitance of polysilicon lines may result in significant RC delays;


n-well
pMOS
minimum overlap of n-well over p+ active area
minimum separation between n+ active area and n-well
minimum separation between the nMOS and the pMOS transistor
nMOS

Complete mask layout of the CMOS inverter

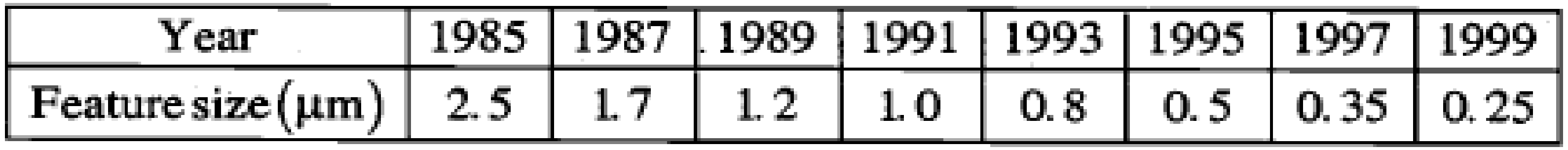
- Layout design rules dictate a set of limitations for the mask geometry
- Full-custom layout design process still allows a large number of variations in terms of device sizing, the placement of individual devices, and the routing of interconnections between the devices
- A simple circuit consisting of only two transistors.
- Depending on the dominant design criteria and design constraints (minimization of overall silicon area, minimization of delay times, placement of input/output pins, etc.), one can choose a certain mask layout design over other alternatives.



MOSFET Scaling and Small-Geometry Effects

- The design of high-density chips in MOS VLSI (Very Large Scale Integration) technology requires that the packing density of MOSFETs used in the circuits is as high as possible and, consequently, that the sizes of the transistors are as small as possible.
- The reduction of the size, i.e., the dimensions of MOSFETs, is commonly referred to as *scaling.*
- It is expected that the operational characteristics of the MOS transistor will change with the reduction of its dimensions.
- There are two basic types of size-reduction strategies: *full scaling* (also called constant-field scaling) and *constant voltage scaling.*
- Scaling of MOS transistors is concerned with systematic reduction of overall dimensions of the devices as allowed by the available technology, while preserving the geometric ratios found in the larger devices.
- To describe device scaling, a constant *scaling factor S >1 is required*
- A new generation of manufacturing technology replaces the previous one about every two or three years, and the down-scaling factor S of the minimum feature size from one generation to the next is about 1.2 to 1.5.

Table: Reduction of the minimum feature size

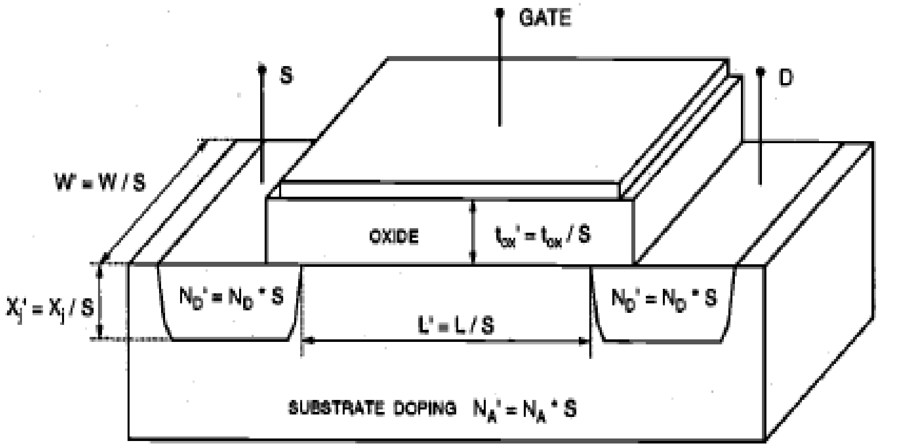


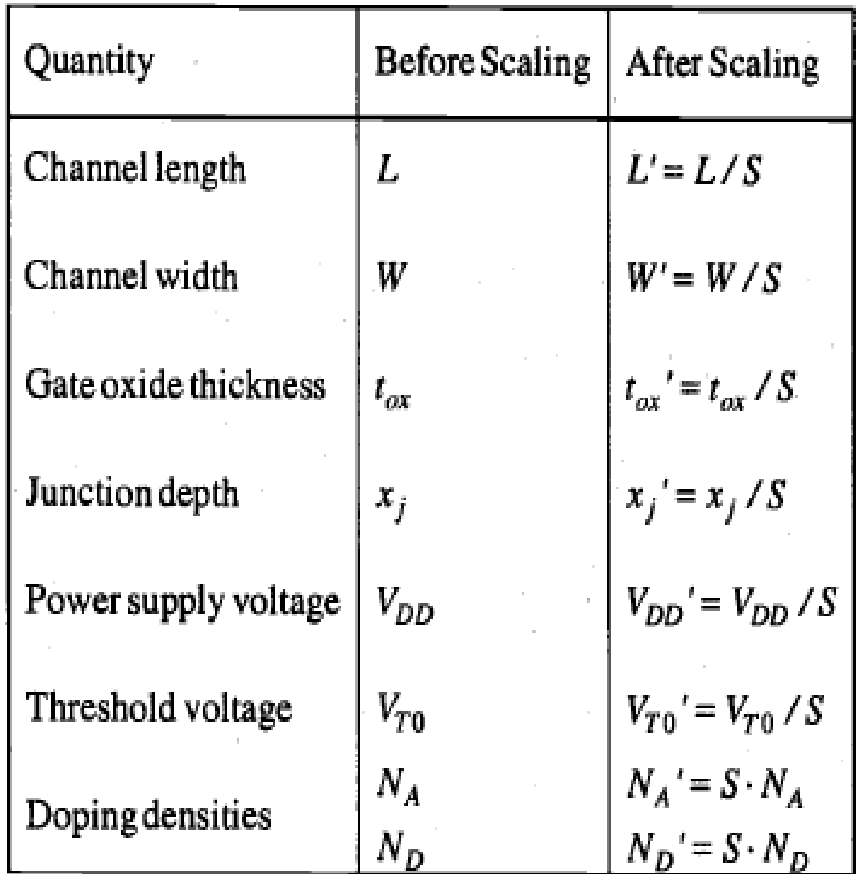
Year	1985	1987	1989	1991	1993	1995	1997	1999
Feature size (μm)	2.5	1.7	1.2	1.0	0.8	0.5	0.35	0.25

Full Scaling (Constant-Field Scaling)

- Scaling of a typical MOSFET by a scaling factor of *S*.
- This scaling option attempts to preserve the magnitude of internal electric fields in the MOSFET, while the dimensions are scaled down by a factor of S.
- To achieve this goal, all potentials must be scaled down proportionally, by the same scaling factor.
- this potential scaling also affects the threshold voltage V_{T0}
- Finally, the Poisson equation describing the relationship between charge densities and electric fields dictates that the charge densities must be *increased* by a factor of S in order to maintain the field conditions.

Table : lists the scaling factors for all significant dimensions, potentials, and doping densities of the MOS transistor





Quantity	Before Scaling	After Scaling
Channel length	L	$L' = L / S$
Channel width	W	$W' = W / S$
Gate oxide thickness	t_{ox}	$t_{ox}' = t_{ox} / S$
Junction depth	x_j	$x_j' = x_j / S$
Power supply voltage	V_{DD}	$V_{DD}' = V_{DD} / S$
Threshold voltage	V_{T0}	$V_{T0}' = V_{T0} / S$
Doping densities	N_A N_D	$N_A' = S \cdot N_A$ $N_D' = S \cdot N_D$

Full scaling effect on the current-voltage characteristics of the MOS transistor

- The gate oxide capacitance per unit area, on the other hand, is changed as follows:

$$C'_{ox} = \frac{\epsilon_{ox}}{t'_{ox}} = S\,\frac{\epsilon_{ox}}{t_{ox}} = S\,C_{ox}$$

- The aspect ratio *W/L* of the MOSFET will remain unchanged under scaling. Consequently, the transconductance parameter kn will also be scaled by a factor of S.

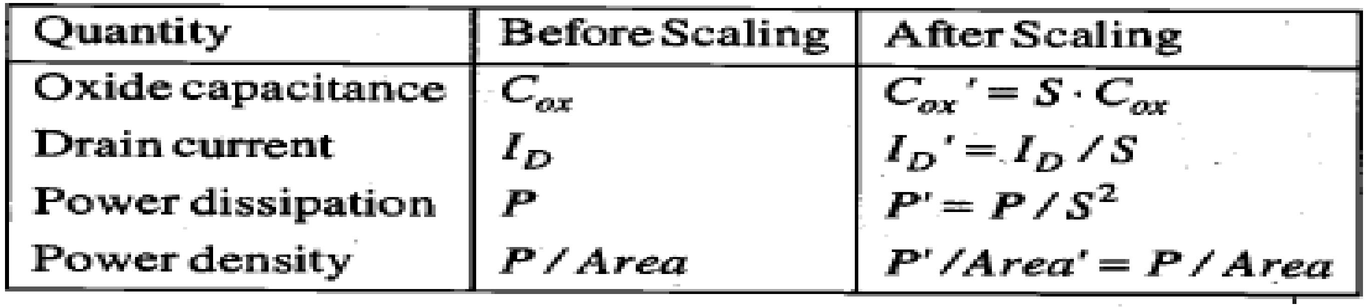
- Since all terminal voltages are scaled down by the factor S as well, the linear-mode drain current of the scaled MOSFET can now be found as:

$$I'_D(lin) = K'_n\,[2(V_{gs} - V_T)V_{ds} - V_{ds}{}^2] = \frac{K_n}{S}\,[2(V_{gs} - V_T)V_{ds} - V_{ds}{}^2] = \frac{I_D(lin)}{S}$$

- Instantaneous power dissipated by the device (before scaling) can be found as:

$$P' = I'_D V'_{DS} = \frac{P}{S^2}$$

Table: Effects of full scaling upon key device characteristics



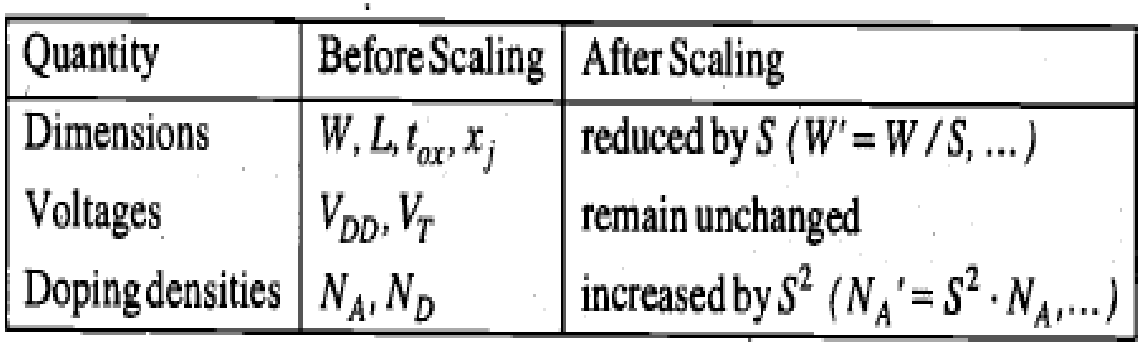
Quantity	Before Scaling	After Scaling
Oxide capacitance	C_{ox}	$C_{ox}' = S \cdot C_{ox}$
Drain current	I_D	$I_D' = I_D / S$
Power dissipation	P	$P' = P / S^2$
Power density	$P / Area$	$P' / Area' = P / Area$

Constant-Voltage Scaling

- While the full scaling strategy dictates that the power supply voltage and all terminal voltages be scaled down proportionally with the device dimensions, the scaling of voltages may not be very practical in many cases.
- In particular, the peripheral and interface circuitry may require certain voltage levels for all input and output voltages, which in turn would necessitate multiple power supply voltages and complicated levelshifter arrangements.
- For these reasons, constant-voltage scaling is usually preferred over full scaling.
- In constant-voltage scaling, all dimensions of the MOSFET are reduced by a factor of *S*, as in full scaling.
- The power supply voltage and the terminal voltages, on the other hand, remain unchanged.
- The doping densities must be increased by a factor of S^2 in order to preserve the charge-field relations.

$$I'_D(lin) = K'_n\,[2(V_{gs} - V_T)V_{ds} - V_{ds}^2] = S\,I_D(lin)$$

$$P' = I'_D V'_{DS} = S.P$$



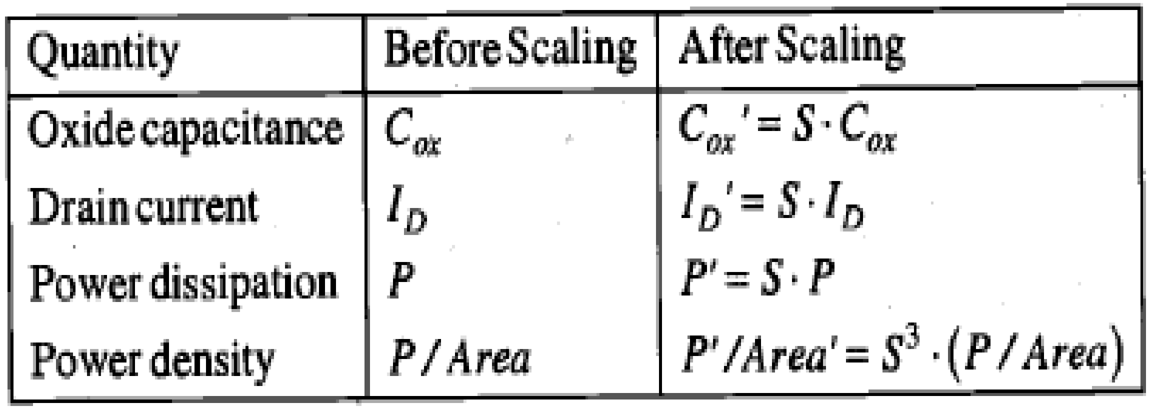
Quantity	Before Scaling	After Scaling
Dimensions	W, L, t_{ox}, x_j	reduced by S ($W' = W/S, \ldots$)
Voltages	V_{DD}, V_T	remain unchanged
Doping densities	N_A, N_D	increased by S^2 ($N_A' = S^2 \cdot N_A, \ldots$)

Table: Constant-voltage scaling of MOSFET dimensions, potentials, and doping densities.

Constant-Voltage Scaling

- Power density (power dissipation per unit area) is found to increase by a factor of *S3* after constant-voltage scaling, with possible adverse effects on device reliability.
- constant-voltage scaling may be preferred over full (constant-field) scaling in many practical cases because of the external voltage-level constraints.
- It must be recognized, however, that constant-voltage scaling increases the drain current density and the power density by a factor of *S3.*
- This large increase in current and power densities may eventually cause serious reliability problems for the scaled transistor, such as electromigration, hot-carrier degradation, oxide breakdown, and electrical over-stress.

Table: Effects of constant-voltage scaling upon key device characteristics.



Quantity	Before Scaling	After Scaling
Oxide capacitance	C_{ox}	$C_{ox}' = S \cdot C_{ox}$
Drain current	I_D	$I_D' = S \cdot I_D$
Power dissipation	P	$P' = S \cdot P$
Power density	$P / Area$	$P'/Area' = S^3 \cdot (P / Area)$

MCQ

Q1. In constant field scaling, the saturation current is scaled by the factor (S) of:
a) s
b) 1/s
c) 1/s2
d) Independent of scaling factor

Q2. In constant field scaling, the power dissipation per gate is scaled by factor (S) as:
a) s
b) 1/s
c) 1/s2
d) Independent of scaling factor

Short Channel Effect: Non-Ideal Effect

The deviation of MOSFET characteristics from ideal condition is known as non-ideal effect.

- **Subthreshold Conduction:** For $V_{gs}<V_T$ a small subthreshold current flows.
- **Channel Length Modulation:** In saturation, the depletion region at drain terminal shift towards source reducing effect channel length.
- $I'_D = \frac{L}{L-\Delta L} I_D$

Subthreshold Performance

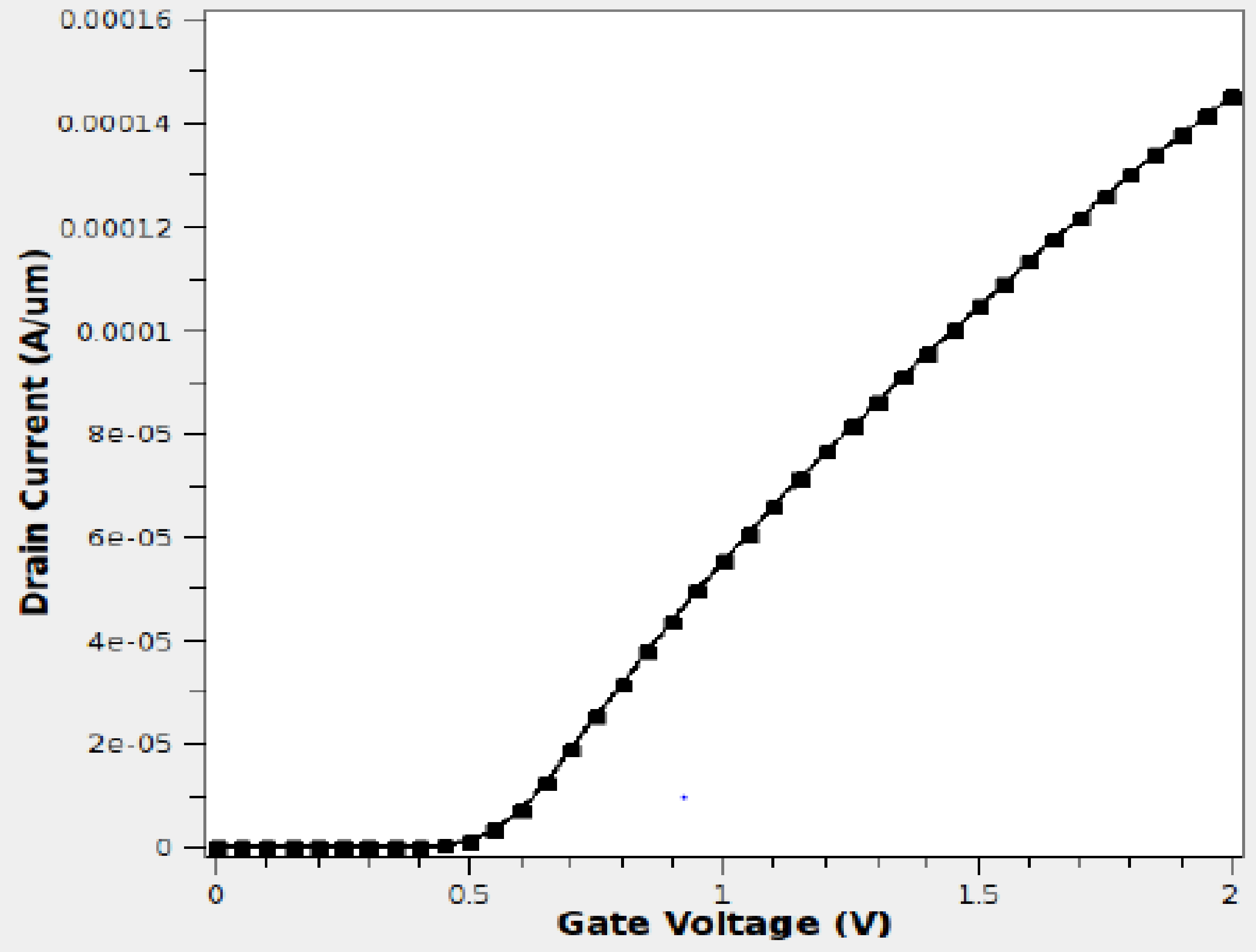
Subthreshold Performance Parameters

- A) Vth: After threshold voltage theld vs Vgs curve have shape changes

- B) Subthreshold slope
- **SS=**$\frac{dV_{gs}}{d(logI_{d})}$ mV/decade

- C)DIBL
- **DIBL=**$\frac{\partial V_{gs}}{\partial V_{ds}}$ mV/V

- D)Ioff: Subthreshold leakage current
- E) Ion/Ioff



Query??