Silicon Semiconductor Technology: An Overview

Silicon, a semiconductor with intermediate electrical resistance, forms the bedrock of modern electronics. Its conductivity can be precisely controlled by introducing impurities, known as dopants, into its crystal lattice. These dopants, either donors or acceptors, dictate whether the silicon is n-type (electron-rich) or p-type (hole-rich), forming the fundamental building blocks of semiconductor devices. The arrangement of junctions between n-type and p-type regions, combined with advanced manufacturing techniques, enables the creation of sophisticated devices with specialized properties.

by CITechnocrats07

Made with Gamma

Dopants: The Key to Silicon Conductivity

The controlled introduction of dopants into the silicon crystal lattice is paramount to tailoring its electrical conductivity. Donor elements, such as phosphorus, contribute free electrons, resulting in n-type silicon. Acceptor elements, like boron, create holes by accepting electrons, leading to p-type silicon. The concentration and distribution of these dopants determine the silicon's conductivity, enabling the fabrication of transistors, diodes, and other essential components.

N-type Silicon

Excess electrons

P-type Silicon

Excess holes

Junctions

Regions where N-type and P-type silicon meet

Made with Gamma

Wafer Processing: From Ingot to Disk

The foundation of semiconductor manufacturing lies in wafer processing. Wafers, thin disks of silicon, are cut from single-crystal ingots grown from molten polycrystalline silicon using the Czochralski method. This process ensures the uniform crystal structure necessary for optimal device performance. Controlled amounts of dopants are added during ingot growth to impart the desired electrical characteristics. Wafers are then sliced, polished, and prepared for subsequent processing steps.

1 **Crystal Growth**

Czochralski method

2 **Doping**

Controlled impurity addition

3 **Wafer Slicing**

Precision cutting with diamond blades

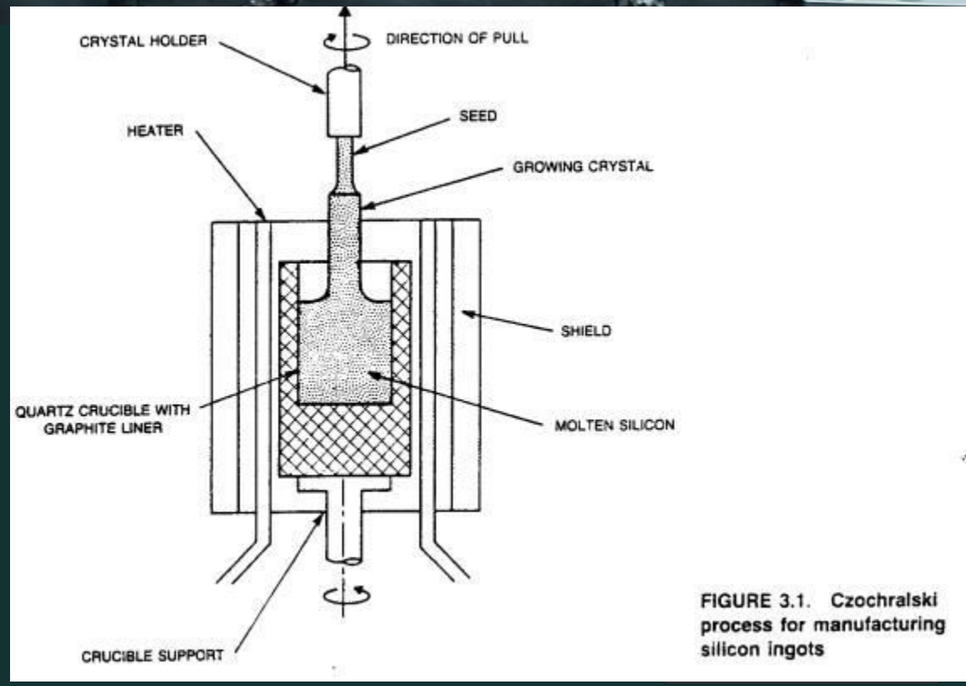


FIGURE 3.1. Czochralski process for manufacturing silicon ingots

The Czochralski Method: Growing Single Crystals

The Czochralski method is a cornerstone of silicon wafer production. A seed crystal, with a specific crystal orientation, is dipped into a crucible of molten silicon. As the seed is slowly withdrawn and rotated, the molten silicon freezes onto the seed, replicating its single-crystal structure. The diameter of the ingot is meticulously controlled by adjusting the withdrawal and rotation rates, ensuring consistent wafer dimensions and properties. Growth rates vary between 30 to 180 mm/hour.

Made with Gamma

Wafer Slicing and Polishing: Achieving Perfection

Following ingot growth, the silicon is sliced into thin wafers using internal cutting edge diamond blades. These wafers, typically ranging from 0.25 mm to 1.0 mm in thickness, undergo a meticulous polishing process to achieve a flat, scratch-free mirror finish. This polishing step is crucial for ensuring the accurate and reliable fabrication of microelectronic devices. Any imperfections on the wafer surface can negatively impact device performance.

1 Slicing

2 Grinding

3 Polishing

Made with Gamma

Oxidation: Forming the Foundation

Oxidation is a critical process in silicon integrated circuit manufacturing. It relies on the properties of silicon dioxide (SiO2), and its reliable production is essential. Silicon oxidation is achieved by heating silicon wafers in an oxidizing atmosphere, such as oxygen or water vapor. Oxidation is achieved through wet oxidation at lower temperatures or dry oxidation at higher temperatures.

1. Wet oxidation
   Uses water vapor
2. Dry oxidation
   Uses pure oxygen

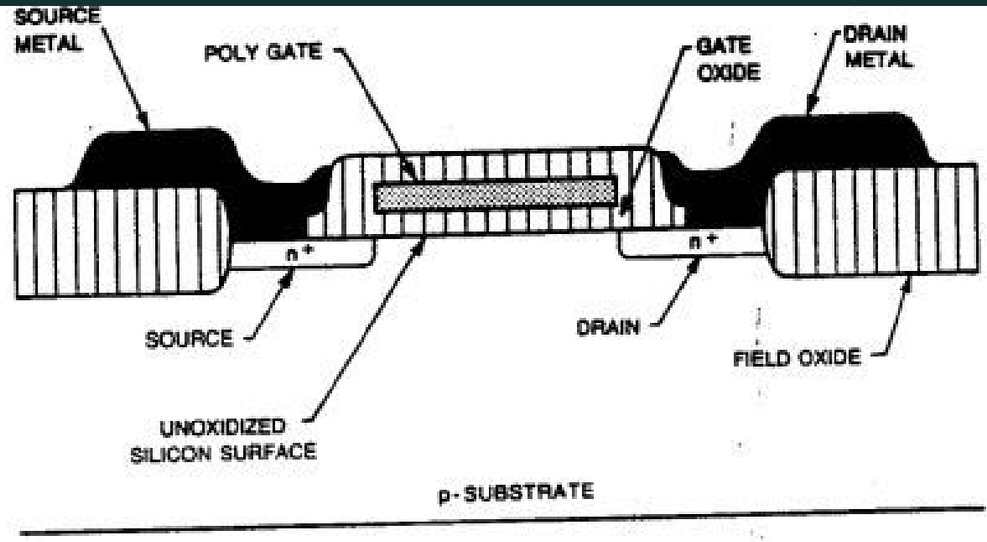


FIGURE 3.2. An nMOS transistor showing the growth of field oxide in both vertical directions

Wet Oxidation: Rapid Oxide Growth

Wet oxidation involves heating silicon wafers in an atmosphere containing water vapor, typically at temperatures between 900°C and 1000°C. This method offers a relatively rapid oxide growth rate compared to dry oxidation. However, the resulting oxide layer may exhibit slightly different properties and may be less dense than that produced by dry oxidation. The selection of wet or dry oxidation depends on the specific requirements of the fabrication process.

900-1000

Temperature

Temperature in Celsius

Fast

Growth

Growth Rate of oxide Layer

Dry Oxidation: Precise Oxide Control

Dry oxidation utilizes pure oxygen as the oxidizing atmosphere, generally at temperatures around 1200°C. While slower than wet oxidation, dry oxidation yields a high-quality, dense silicon dioxide layer with superior electrical properties. This method is favored when precise control over oxide thickness and characteristics is paramount. The silicon consumption is also less.

Oxidation Method	Oxidizing Atmosphere	Temperature	Growth Rate
Wet	Water Vapor	900-1000°C	Fast
Dry	Pure Oxygen	~1200°C	Slow

Silicon Consumption During Oxidation

The oxidation process inherently consumes silicon. The resulting silicon dioxide (SiO2) occupies approximately twice the volume of the original silicon. Consequently, the SiO2 layer grows both vertically and laterally. Understanding this consumption is critical for precisely designing device structures. In an n-channel MOS device, the SiO2 projects both above and below the original silicon surface as oxidation occurs.

Made with Gamma

Key Takeaways and Next Steps

Silicon semiconductor technology relies on precisely controlled doping, wafer processing, and oxidation techniques to create functional devices. The Czochralski method yields high-quality single-crystal silicon, while oxidation forms the essential silicon dioxide layer. Continued advancements in these areas will drive further miniaturization and performance improvements in microelectronics. Further research should focus on improved doping methods and temperature controls.

- Semiconductors
- Oxidation
- Temperature Control

Semiconductor Manufacturing: Selective Diffusion and Silicon Gate Process

This presentation covers selective diffusion, a crucial process in creating silicon with varying impurity proportions. It also explores the silicon gate process, highlighting the use of polysilicon as a gate electrode for precise source and drain electrode definition.

by CITechnocrats07

Made with Gamma

Selective Diffusion: SiO_2 as a Barrier

SiO_2 Barrier

SiO_2 acts as a barrier against doping impurities, vital for selective diffusion. Areas lacking SiO_2 allow dopant atoms to alter silicon characteristics.

Process Steps

- Open windows in SiO_2 layer.
- Remove SiO_2 with etchant.
- Expose Si to dopant source.

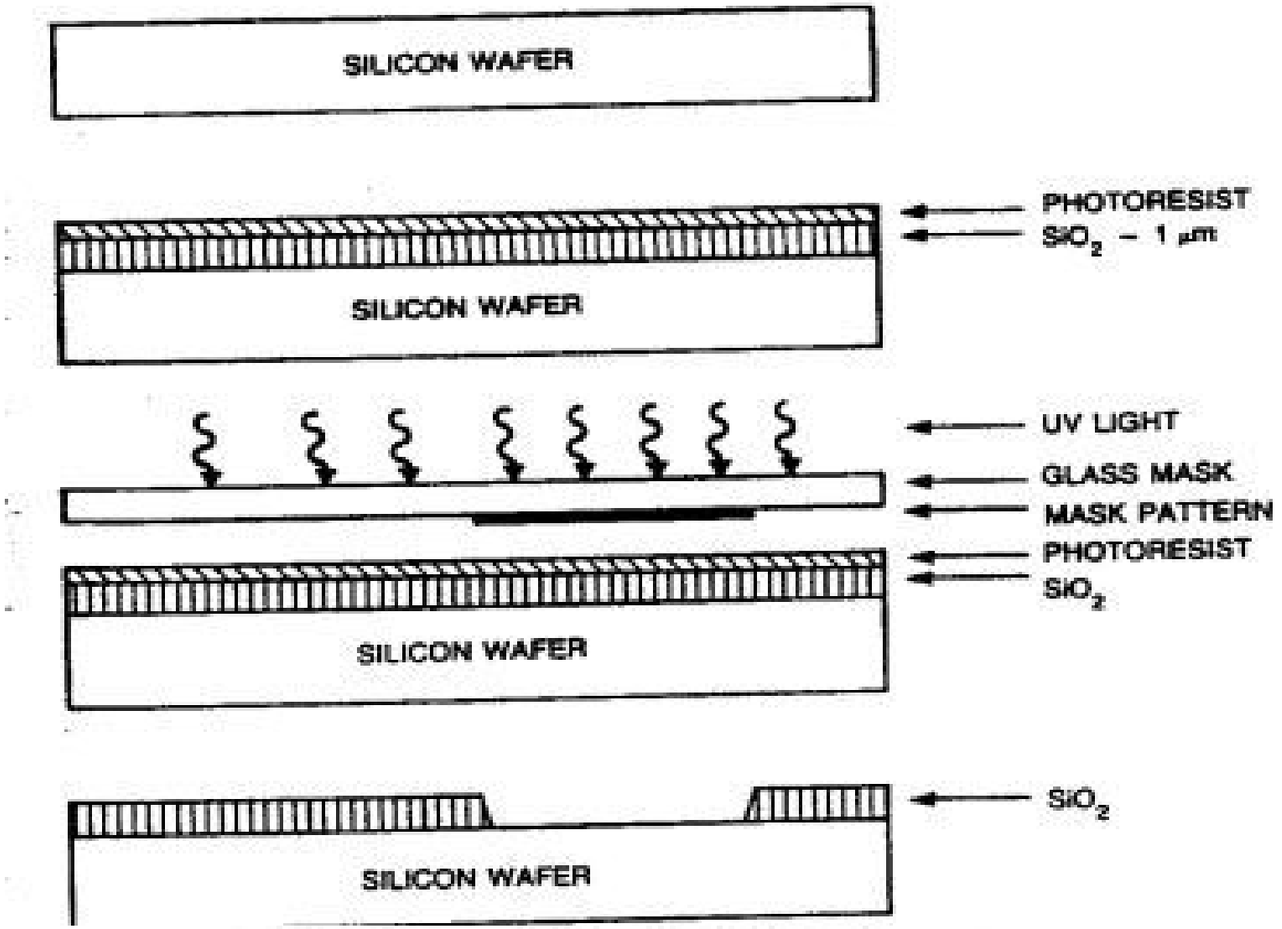


FIGURE 3.3. Simplified steps involved in the patterning of SiO_2

Selective diffusion relies on SiO_2's ability to block doping impurities. This process involves creating openings in the SiO_2 layer, exposing silicon to a dopant source, and altering its characteristics. Areas covered by SiO_2 remain unaffected.

Made with Gamma

Photoresist and Etching Techniques

Acid-Resistant Coating

Surface covered with acid-resistant coating, except where diffusion windows are needed.

Photoresist (PR)

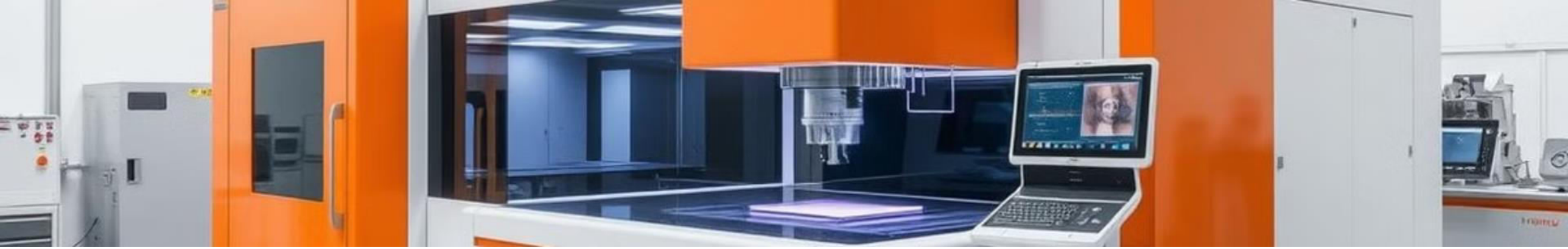
Photosensitive organic material polymerized by UV light through a mask.

Etching

Exposed SiO_2 is removed using etching techniques after unpolymerized areas are removed with organic solvent.

The process involves covering the surface with an acid-resistant coating, typically a photosensitive organic material called photoresist (PR). UV light polymerizes the coating through a mask, and etching removes exposed SiO_2.

Made with Gamma



Electron Beam Lithography (EBL)

Digital Data

Patterns derived directly from digital data.

Pattern Flexibility

Different patterns accommodated easily.

Quick Changes

Pattern changes implemented quickly.

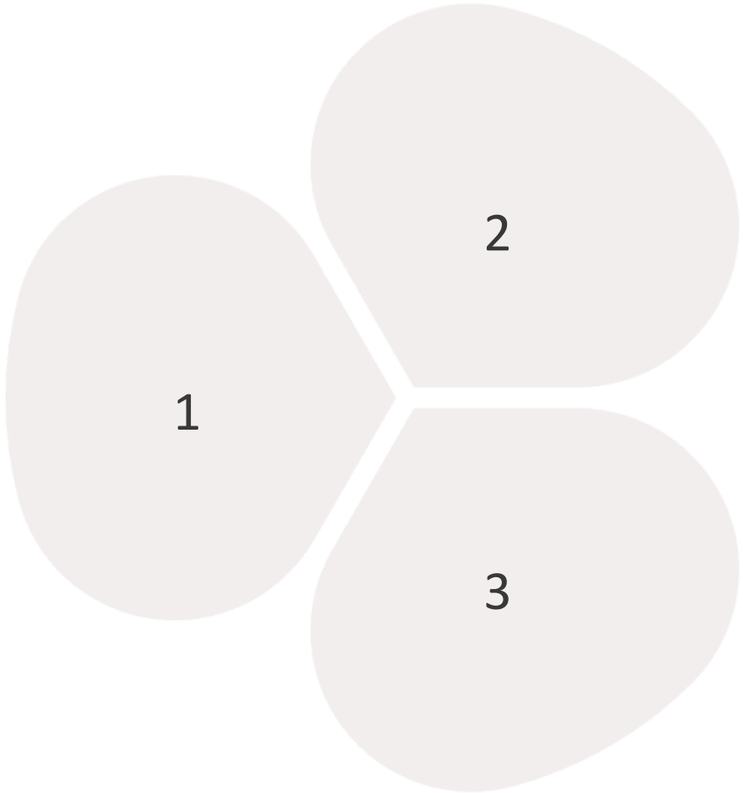
EBL offers advantages like patterns derived from digital data, accommodating different patterns easily, and quick implementation of changes. However, high equipment costs and time requirements have limited its use in commercial fabrication.

Made with Gamma

Silicon Gate Process: Polysilicon

Polysilicon Use

Used as interconnect in silicon ICs and as gate electrode on MOS transistors.



Masking Ability

Acts as further mask for precise definition of source and drain electrodes.

Formation

Formed when silicon is deposited on SiO_2 or other surfaces.

Polysilicon is used as an interconnect in silicon ICs and as the gate electrode on MOS transistors. Its ability to act as a mask allows precise definition of source and drain electrodes, improving circuit performance.

Made with Gamma

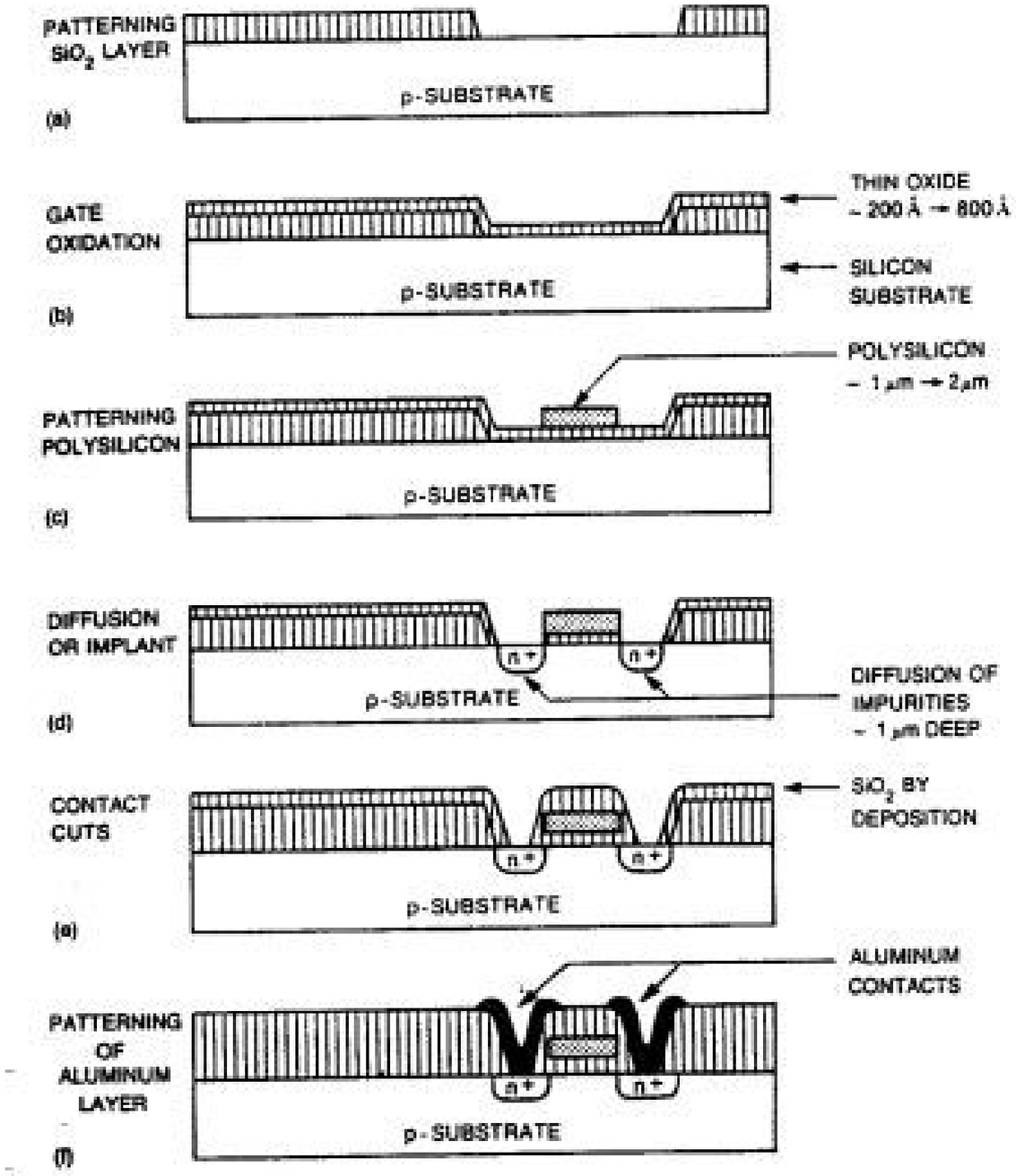


FIGURE 3.4. Fabrication steps for a silicon gate nMOS transistor

Made with Gamma

Silicon Gate Process Steps

1

Field Oxide Etching

Etch field oxide to silicon surface where transistors are placed.

2

Thin Oxide Growth

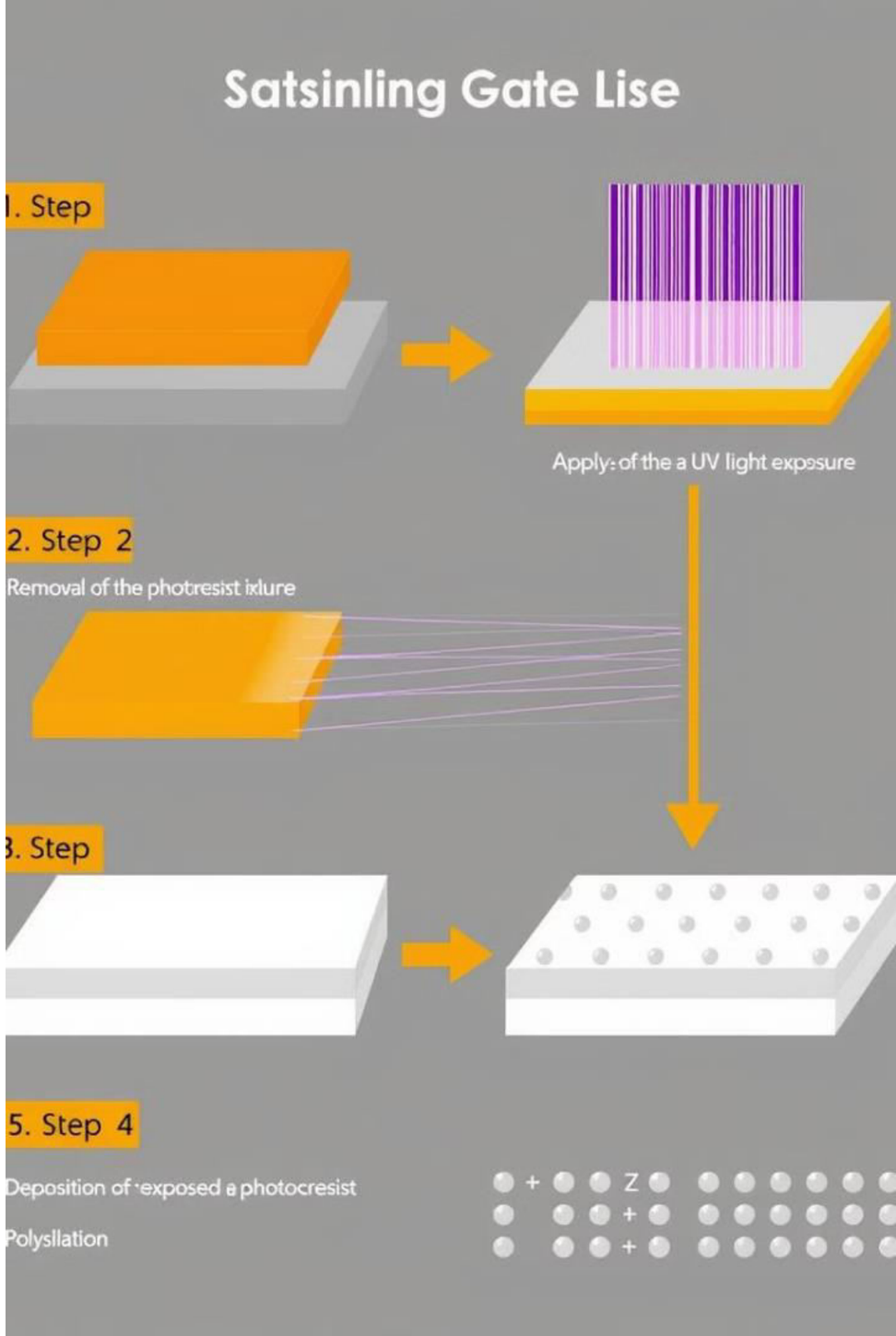
Grow thin, highly controlled SiO_2 layer on exposed silicon surface.

3

Polysilicon Deposition

Deposit polysilicon over wafer surface and etch to form gates.

The silicon gate process involves etching the field oxide, growing a thin oxide layer, depositing polysilicon, and etching to form gates. These steps are repeated during the processing sequence to achieve the desired structure.



Doping and Junction Formation

1

Dopant Exposure

Expose wafer to dopant source.

2

Diffusion Junctions

Form diffusion junctions in substrate.

3

Polysilicon Doping

Dope polysilicon to reduce resistivity.

After polysilicon etching, the wafer is exposed to a dopant source, forming diffusion junctions in the substrate and doping the polysilicon. This reduces the resistivity of the polysilicon and creates the source and drain regions.

Made with Gamma

Self-Aligned Process

0

Overlap

Minimum gate-to-source/drain overlap.

1

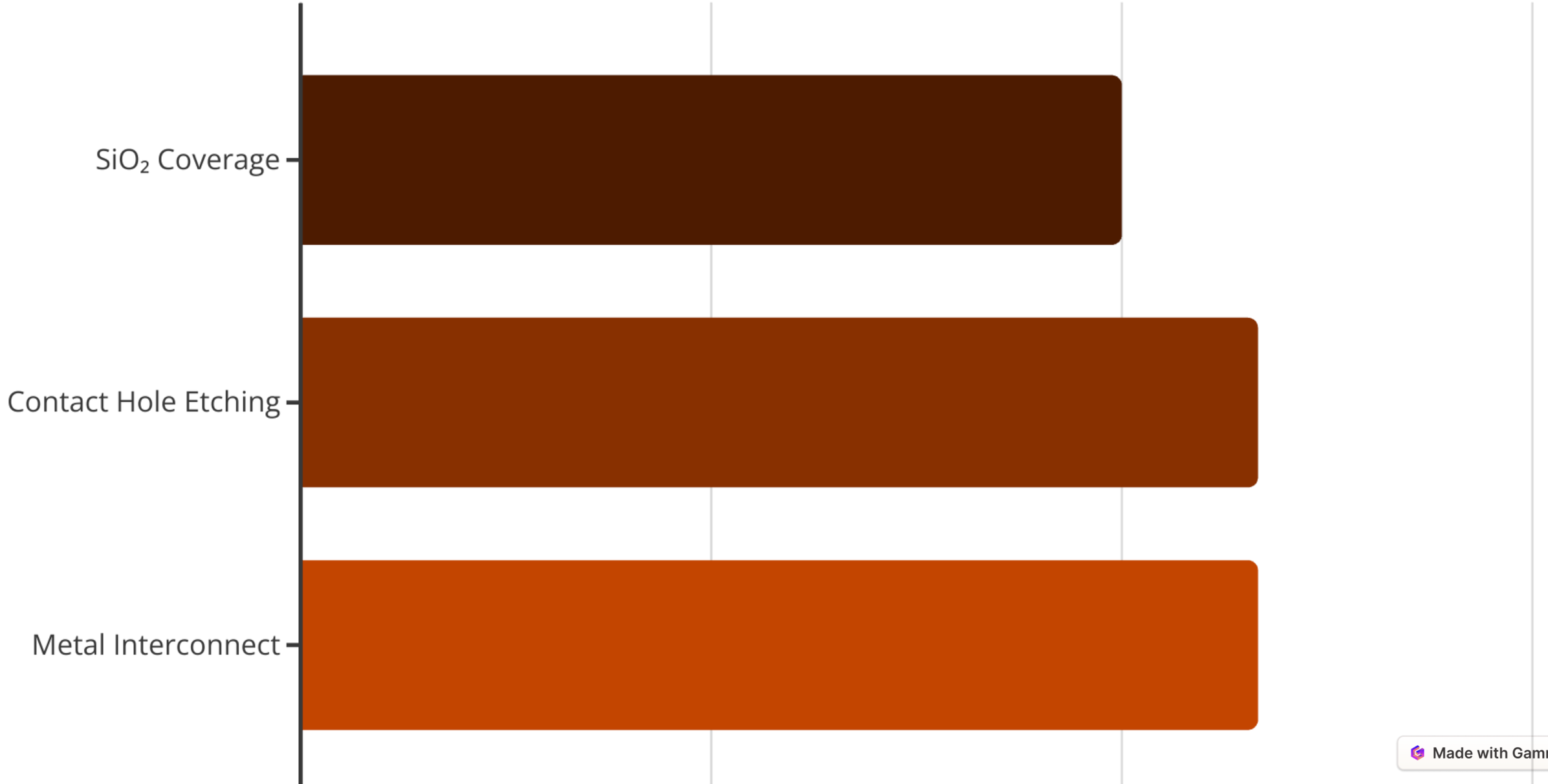
Shadowing

Diffusion junctions form where polysilicon gate does not shadow substrate.

The diffusion junctions form only in regions where the polysilicon gate does not shadow the underlying substrate. This is referred to as a self-aligned process because the source and drain do not extend under the gate.

Made with Gamma

Final Metallization



Made with Gamma

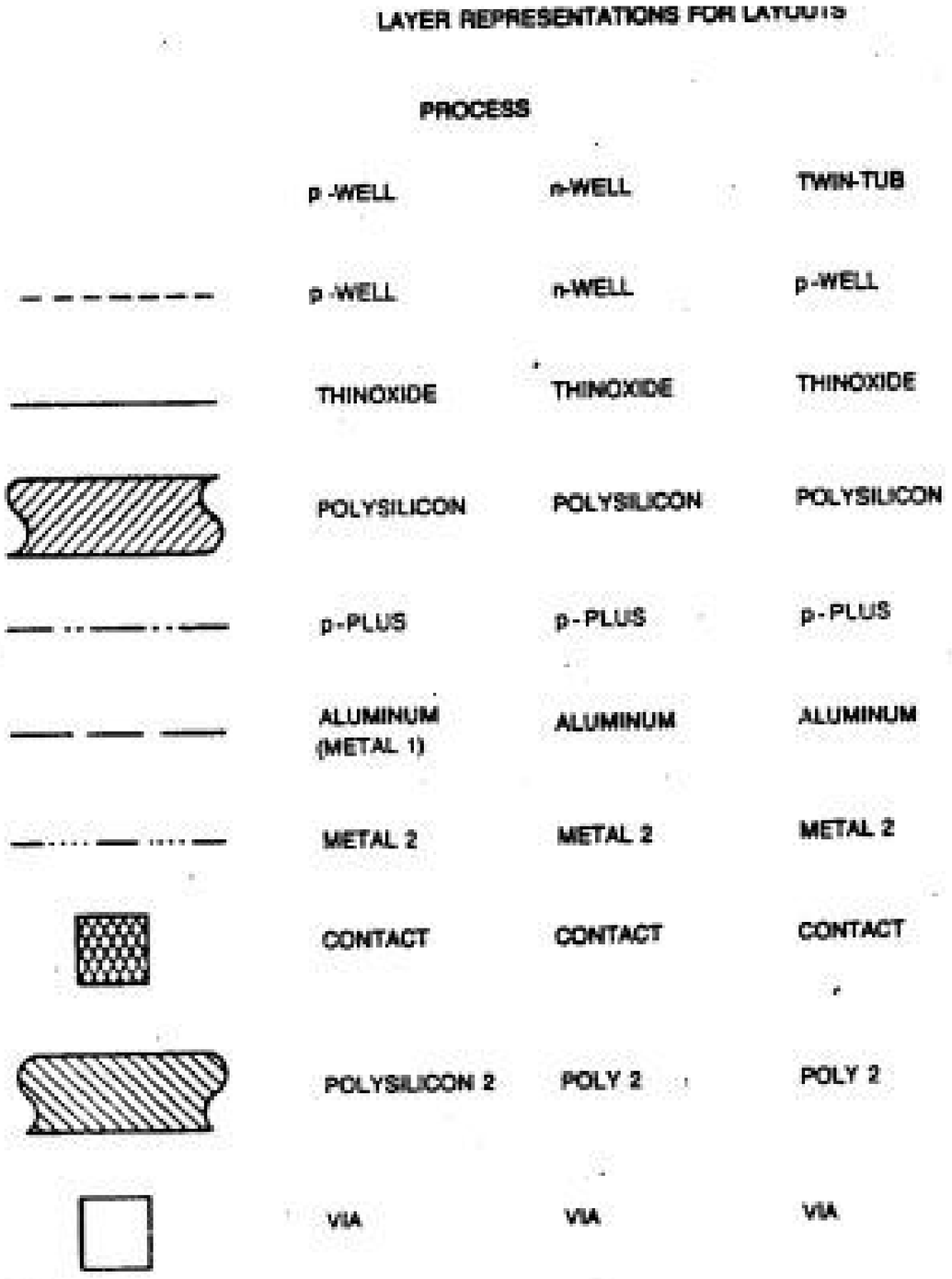


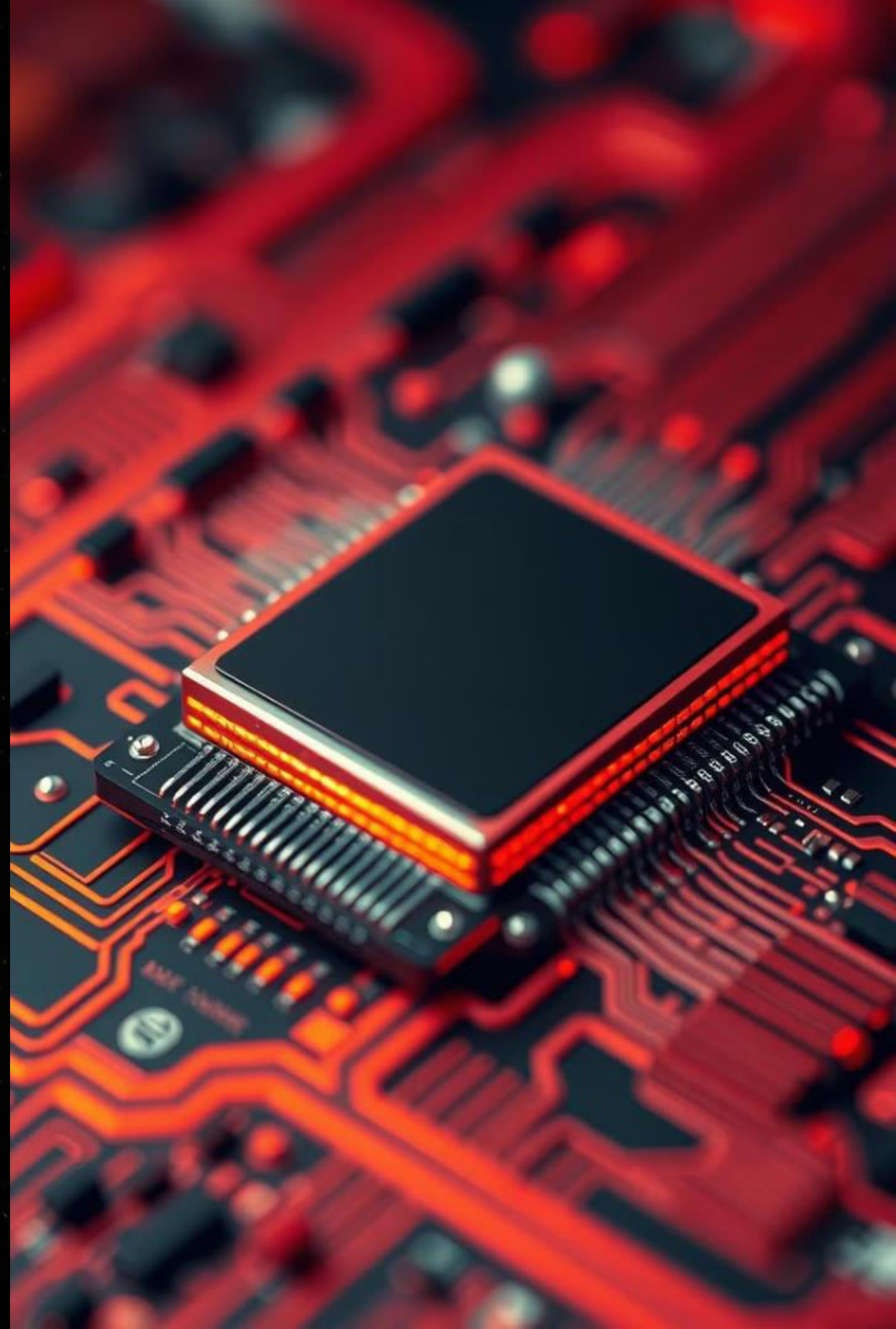
FIGURE 3.5. CMOS process cross-section and layout conventions

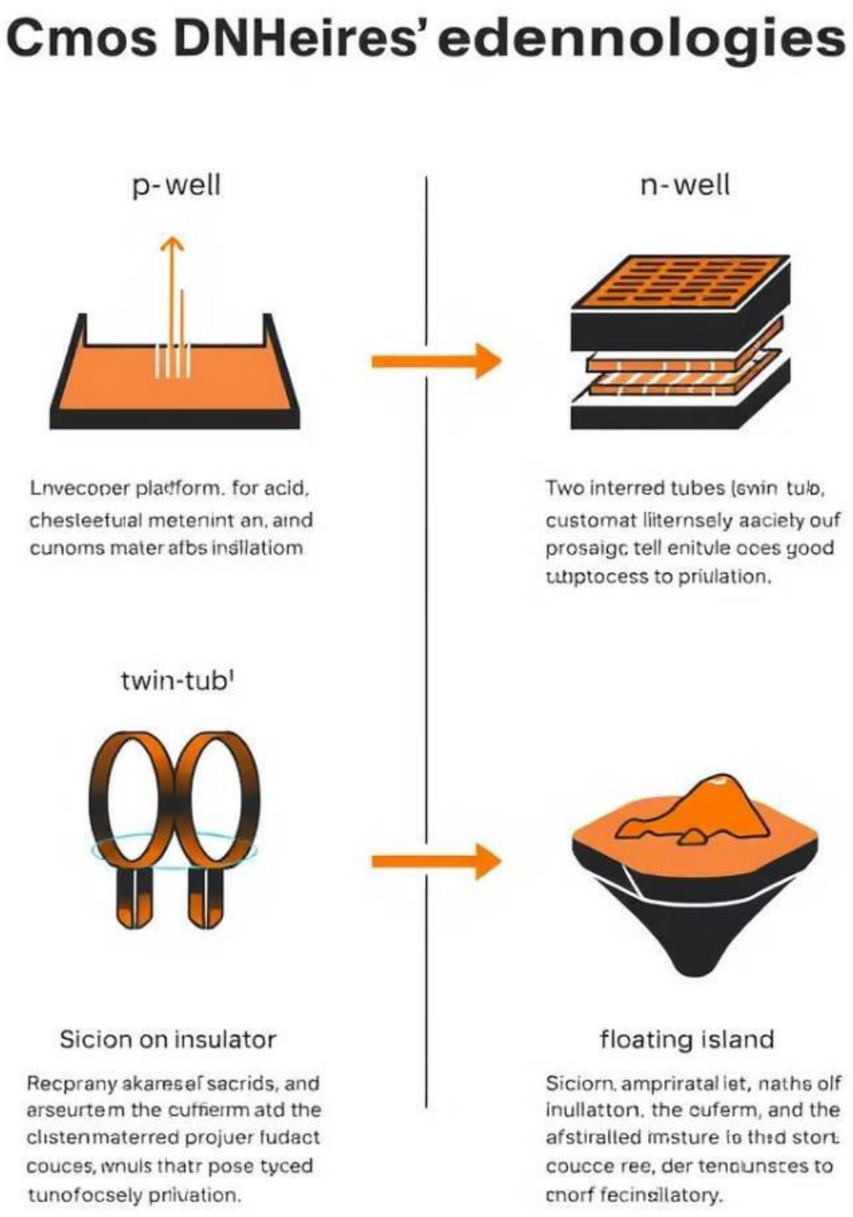
Made with Gamma

CMOS Technologies: An Overview

CMOS (Complementary Metal Oxide Semiconductor) technology is a leading contender for existing and future VLSI systems. CMOS offers inherently low power static circuit technology, providing a lower power-delay product than comparable design-rule nMOS or pMOS technologies. This presentation provides an overview of four dominant CMOS technologies, with a simplified treatment of the process steps.

by CITechnocrats07





Four Dominant CMOS Technologies

P

p-well process

N

n-well process

Twin-tub process

Silicon on insulator

The four dominant CMOS technologies are p-well process, n-well process, twin-tub process, and silicon on insulator. Process cross-sections and layouts will be presented. The drawing conventions summarize the CMOS technologies.

Made with Gamma

The P-Well Process

Fabrication

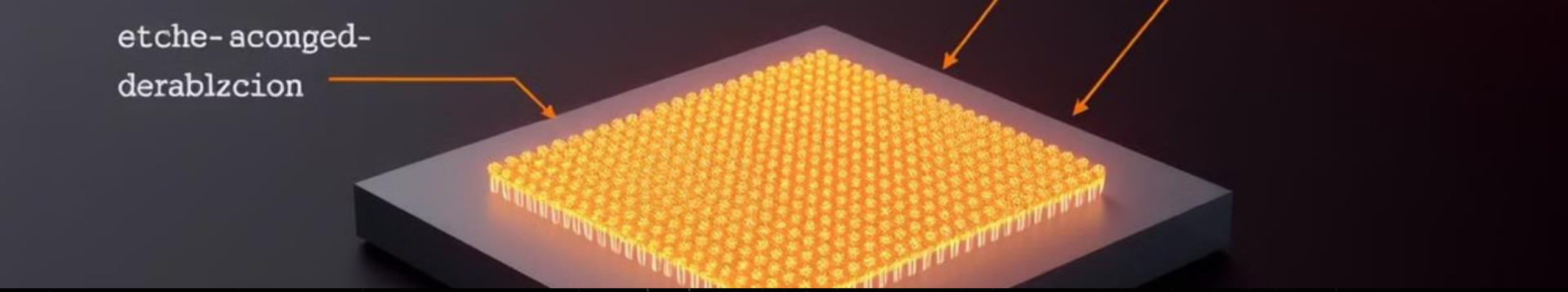
A common approach to p-well CMOS fabrication has been to start with a moderately doped n-type substrate (wafer), create the p-type well for the n-channel devices, and build the p-channel transistor in the native n-substrate.

Mask Levels

The mask levels are not organized by component function, rather they reflect the processing steps. The first mask defines the p-well (or p-tub), n-channel transistors will be fabricated in this well.

Field oxide (FOX) is etched away to flow a deep diffusion. The next mask is called the "thin oxide" or "thinox" mask, as it defines where areas of thin oxide are needed to implement transistor gates and allow implantation to form p- or n-type diffusions for transistor source/drain regions.

Made with Gamma



Polysilicon Gate Definition

Surface Covering

Involves covering the surface with polysilicon and then etching the required pattern.

Self-Aligned Regions

The "poly" gate regions lead to "self-aligned" source-drain regions.

P-Plus Mask

A p-plus (p+) mask is then used to indicate those thin-oxide areas (and polysilicon) that are to be implanted p+.

Hence a thin-oxide area exposed by the p-plus mask will become a p+ diffusion area. If the p-plus area is in the n-substrate, then a p-channel transistor or p-type wire may be constructed. If the p-plus area is in the p-well (not shown), then an ohmic contact to the p-well may be constructed.

Made with Gamma

Ohmic Contacts and N+ Diffusion

1

Ohmic Contact

One which is only resistive in nature and is not rectifying. Current can flow in both directions.

2

N+ Diffusion

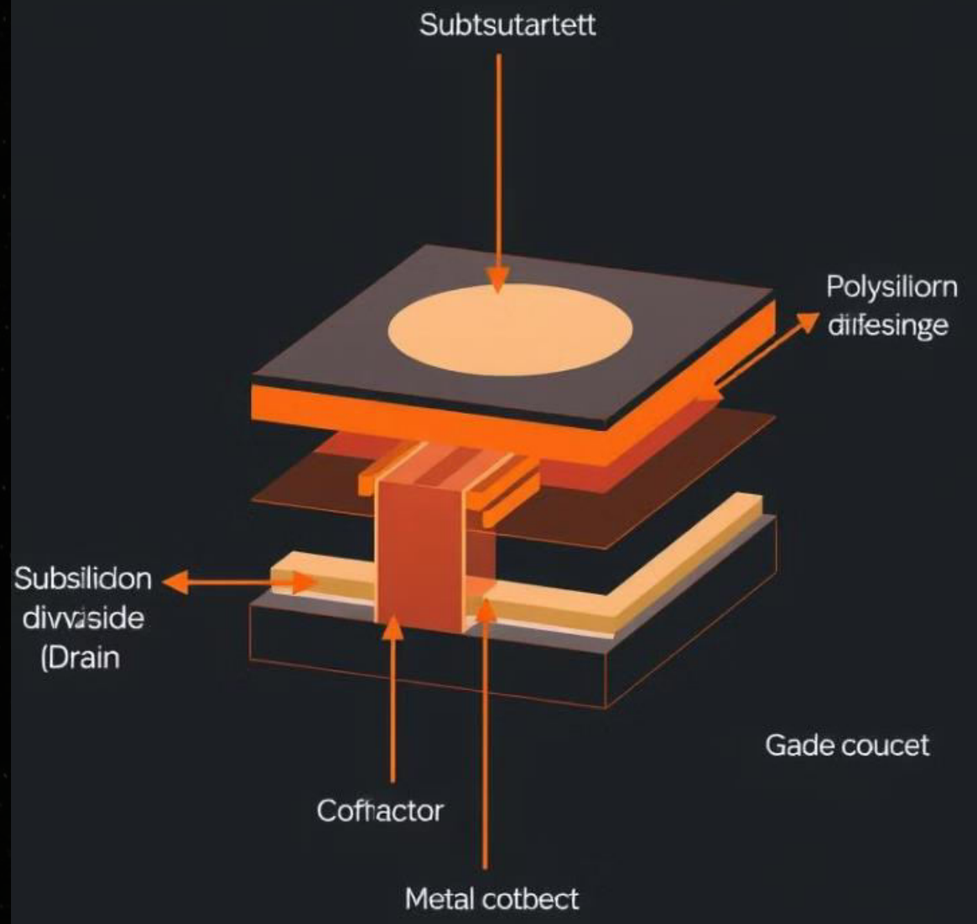
The "absence" of a p-plus region over a thin-oxide area indicates that the area will be an n+ diffusion or n-thinox.

3

N-Thinox

N-thinox in the p-well defines possible n-transistors and wires.

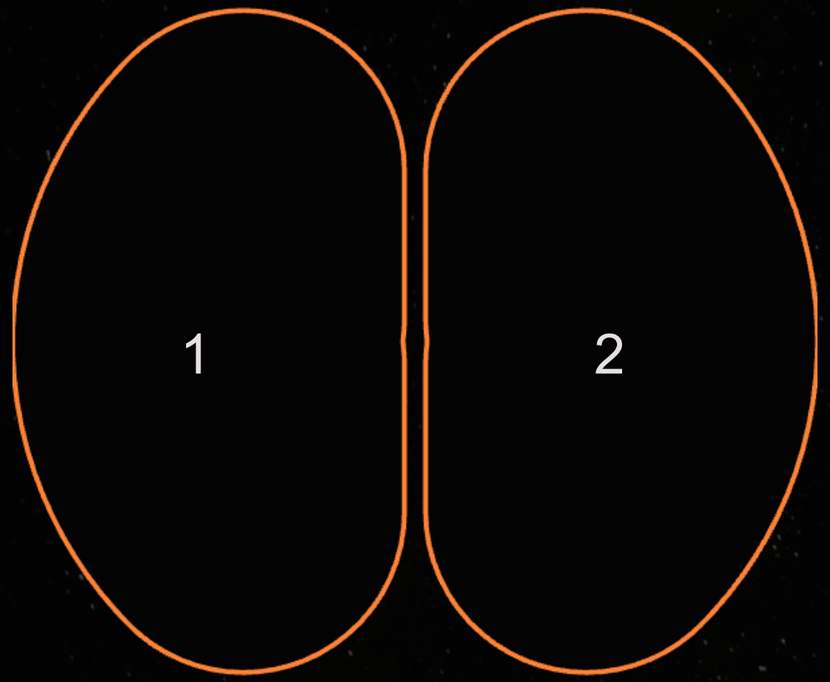
An n+ diffusion in the n-substrate allows a contact to be made. Following this step, the surface of the chip is covered with a layer of SiO2. Contact cuts are then defined. This involves etching any SiO2 down the contacted surface. These allow metal to contact diffusion regions or polysilicon regions.



Metallization and Passivation

Metallization

Applied to the surface and selectively etched.



Passivation

The wafer is passivated and openings to the bond pads are etched to allow for wire bonding.

Passivation protects the silicon surface against the ingress of contaminants that can modify circuit behavior in deleterious ways. Additional steps might include threshold adjust steps to set the threshold voltages of the n- and p-devices.

Made with Gamma

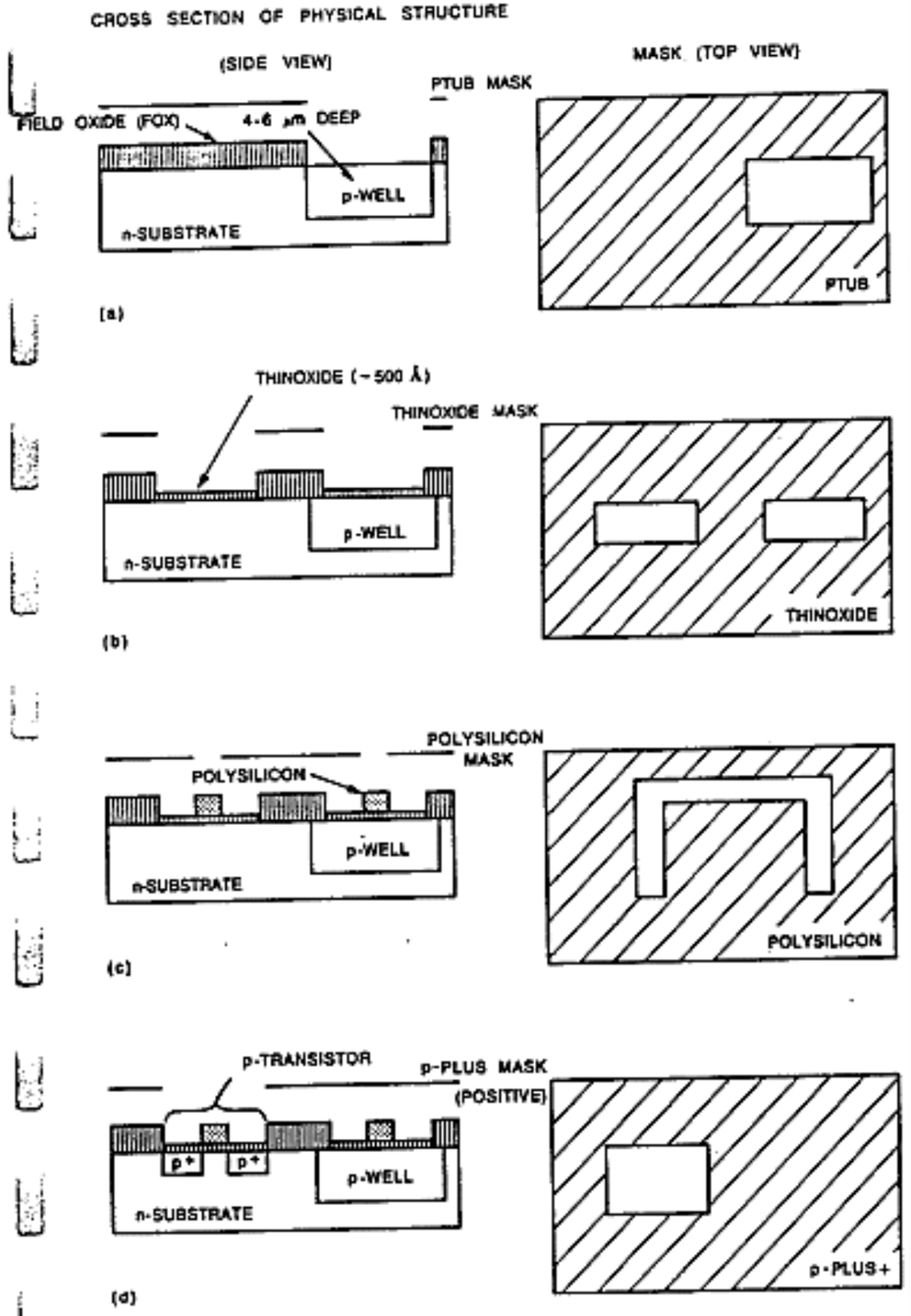


FIGURE 3.6. Typical p-well CMOS process steps with corresponding masks required

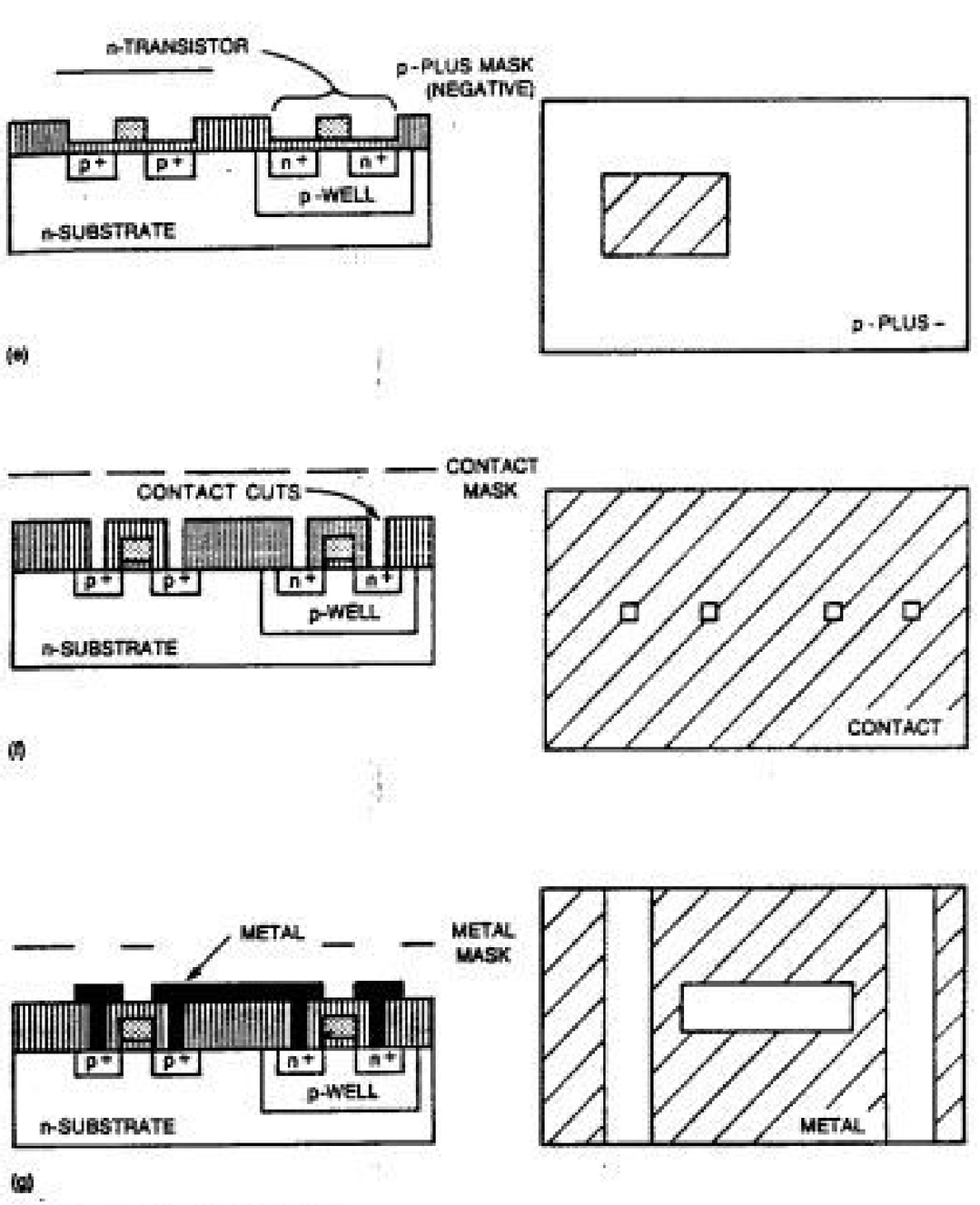


FIGURE 3.6. (Continued)

P-Well Process Layout and Schematics

N-Type Substrate

Accommodates p-channel devices.

P-Well

Accommodates n-channel devices.

The p-well diffusion must be carried out with special care since p-well doping concentration and penetration depth affect the threshold voltages as well as the breakdown voltages of the n-channel devices. To achieve low threshold voltages (0.6V-1.0V), either deep well diffusion or high well resistivity is required.

Made with Gamma

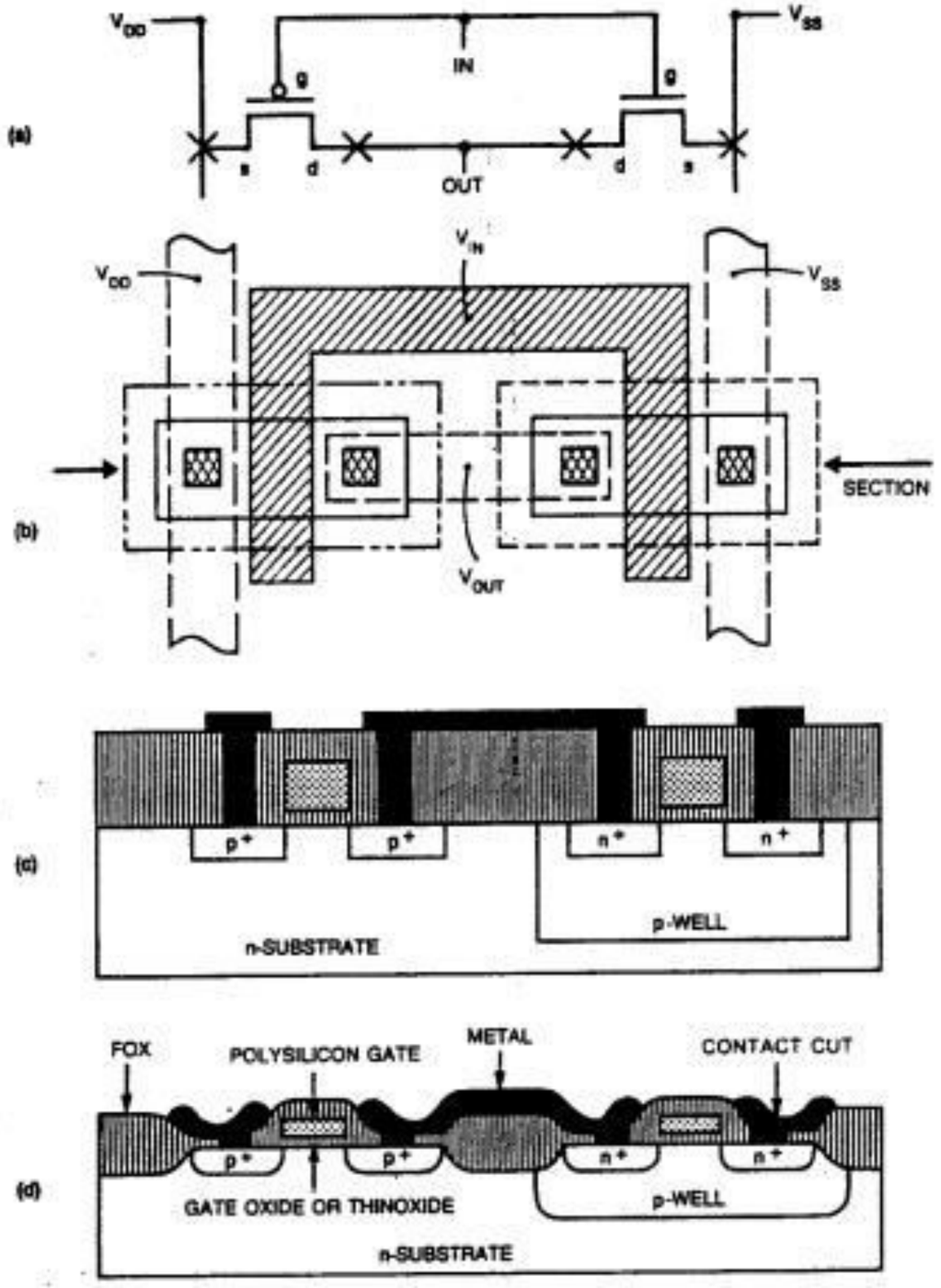


FIGURE 3.7. Layout and process cross-sections of transistors and inverter in p-well CMOS technology

Made with Gamma

Substrate Contacts

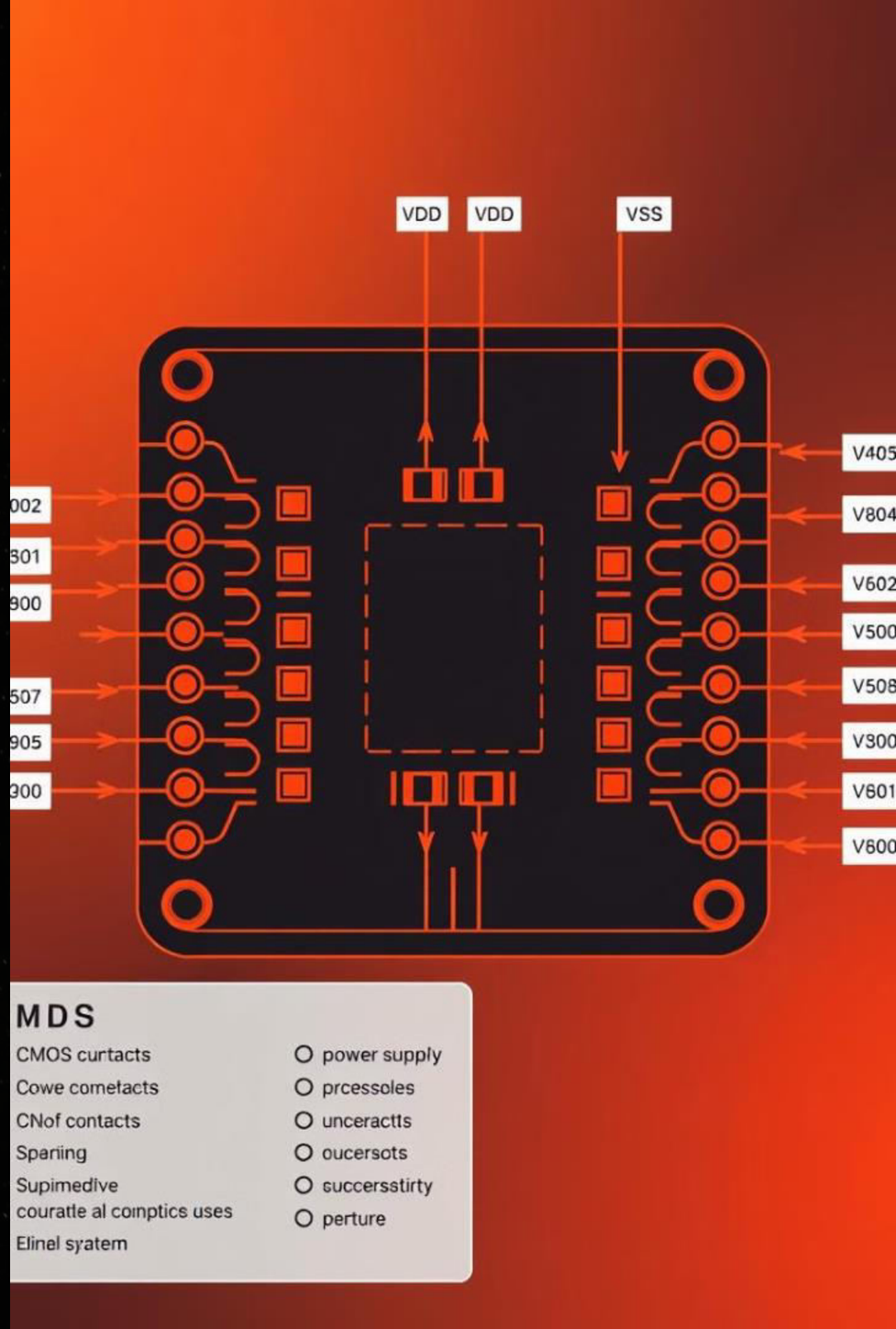
VDD Substrate Contacts

The n-type substrate may be connected to the positive supply (VDD).

VSS Substrate Contacts

The well has to be connected to the negative supply (Vss).

The interesting feature of the Vss contact is that topside connection of substrate is used. This can be compared with nMOS, where backside connection is normally used. VDD backside contact may be used but topside connection is preferred because it reduces parasitic resistances that could cause latch-up.



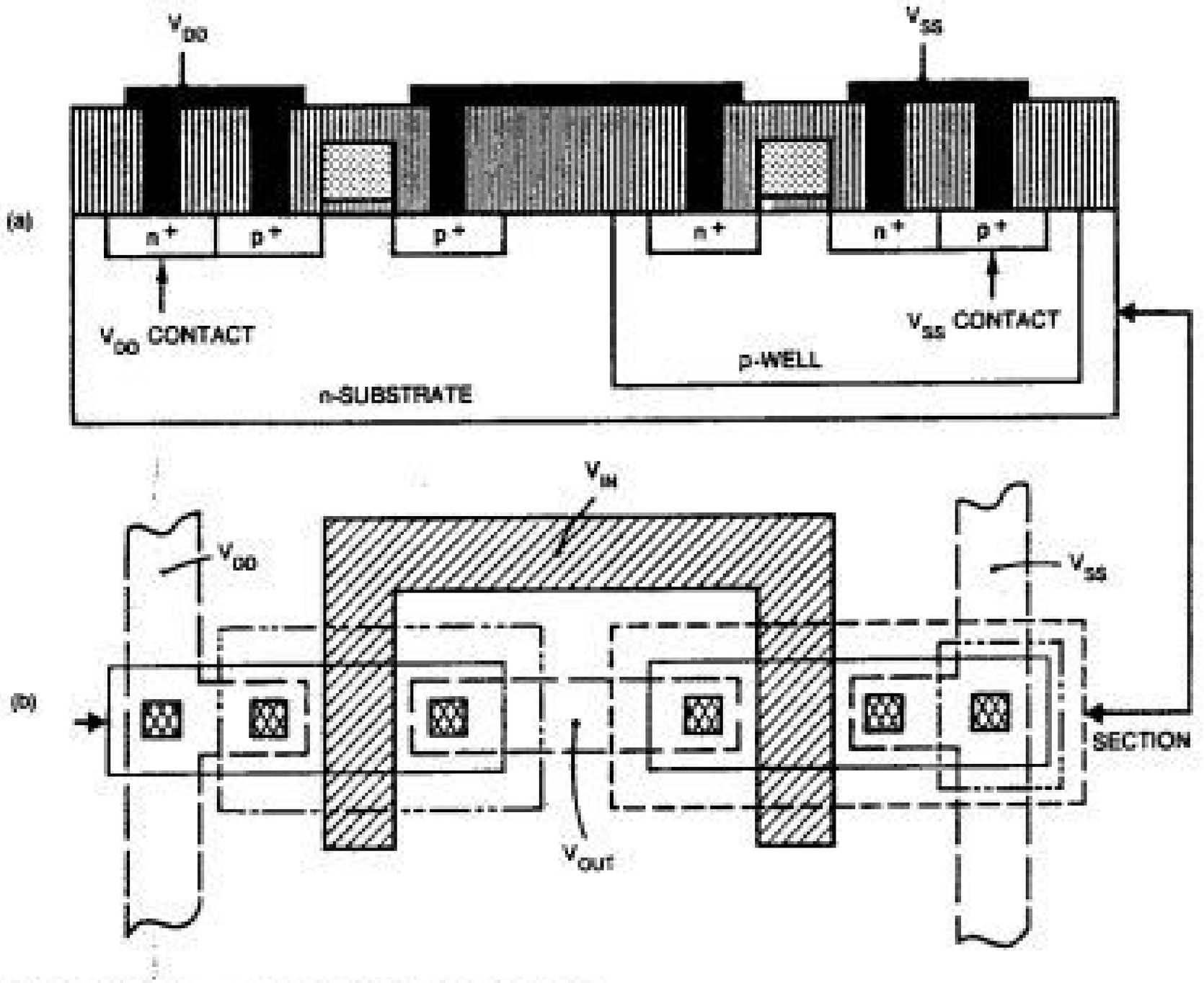


FIGURE 3.8. p-well substrate contacts

Made with Gamma

Improved P-Well CMOS Processes

1 Retrograde P-Well CMOS

2 CMOSC Process

To meet the growing need for higher packing density, improvements in latch-up, and independent threshold adjustment, a number of improved p-well CMOS processes have emerged during recent years. The "retrograde p-well CMOS" process developed by GE-Intersil, Inc., and the "CMOSC" process developed by Hewlett-Packard.

Made with Gamma

Retrograde P-Well and CMOSC Processes

Retrograde P-Well

The well is implanted with a high energy boron implant as opposed to a thermal diffusion process. As a result of this step and the fact that the implant is made after field oxide, the p-well impurities do not diffuse from their original implanted position, thus reducing the lateral diffusion of the well.

CMOSC Process

A boron implant is used to define the p-transistors and a phosphorus implant is used to define the n-transistors. Improvements in CMOSC processes have resulted in an extremely low standby leakage current.

Made with Gamma

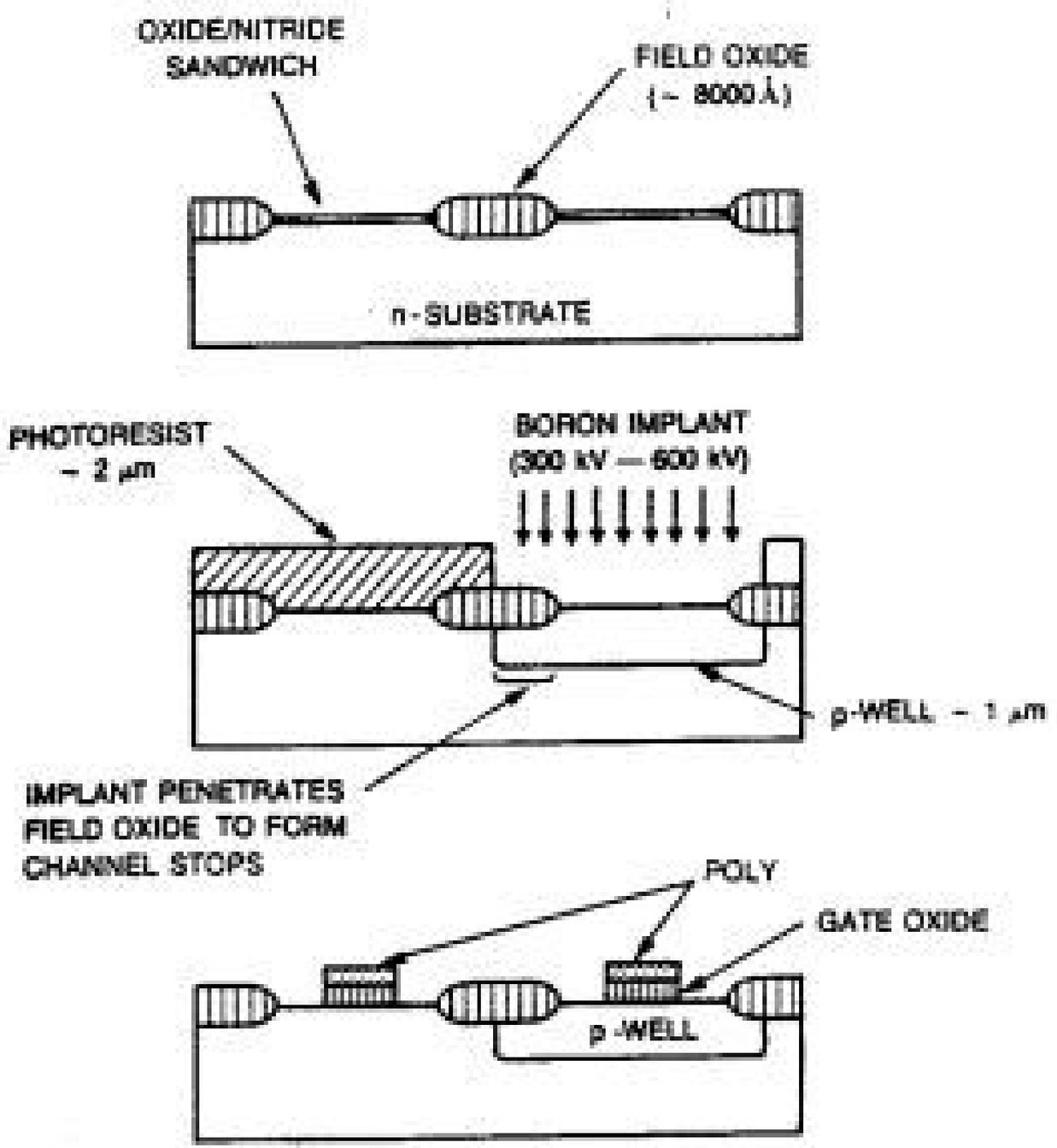


FIGURE 3.9. GE-Intersil's "retrograde p-well" process

Made with Gamma

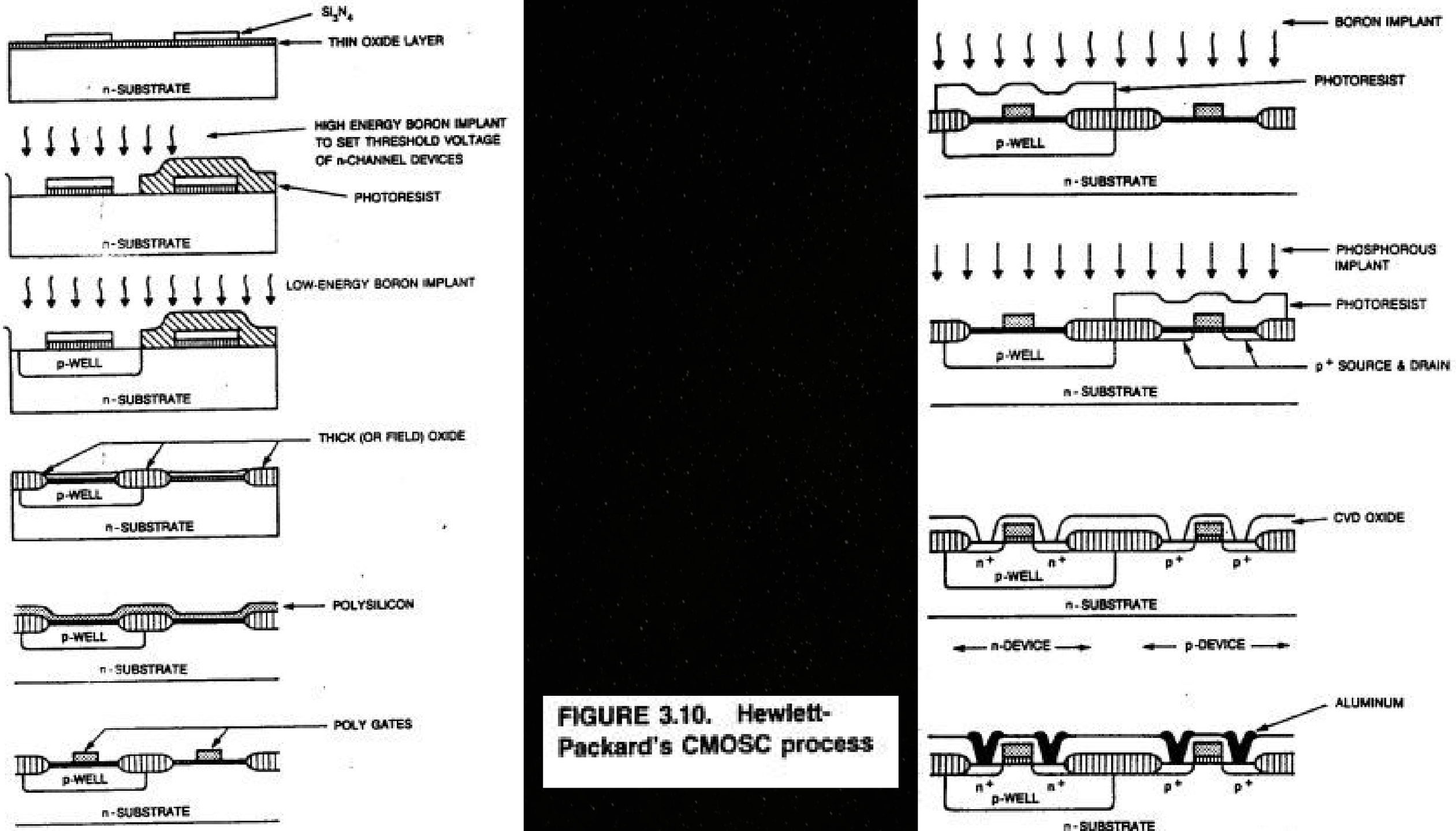
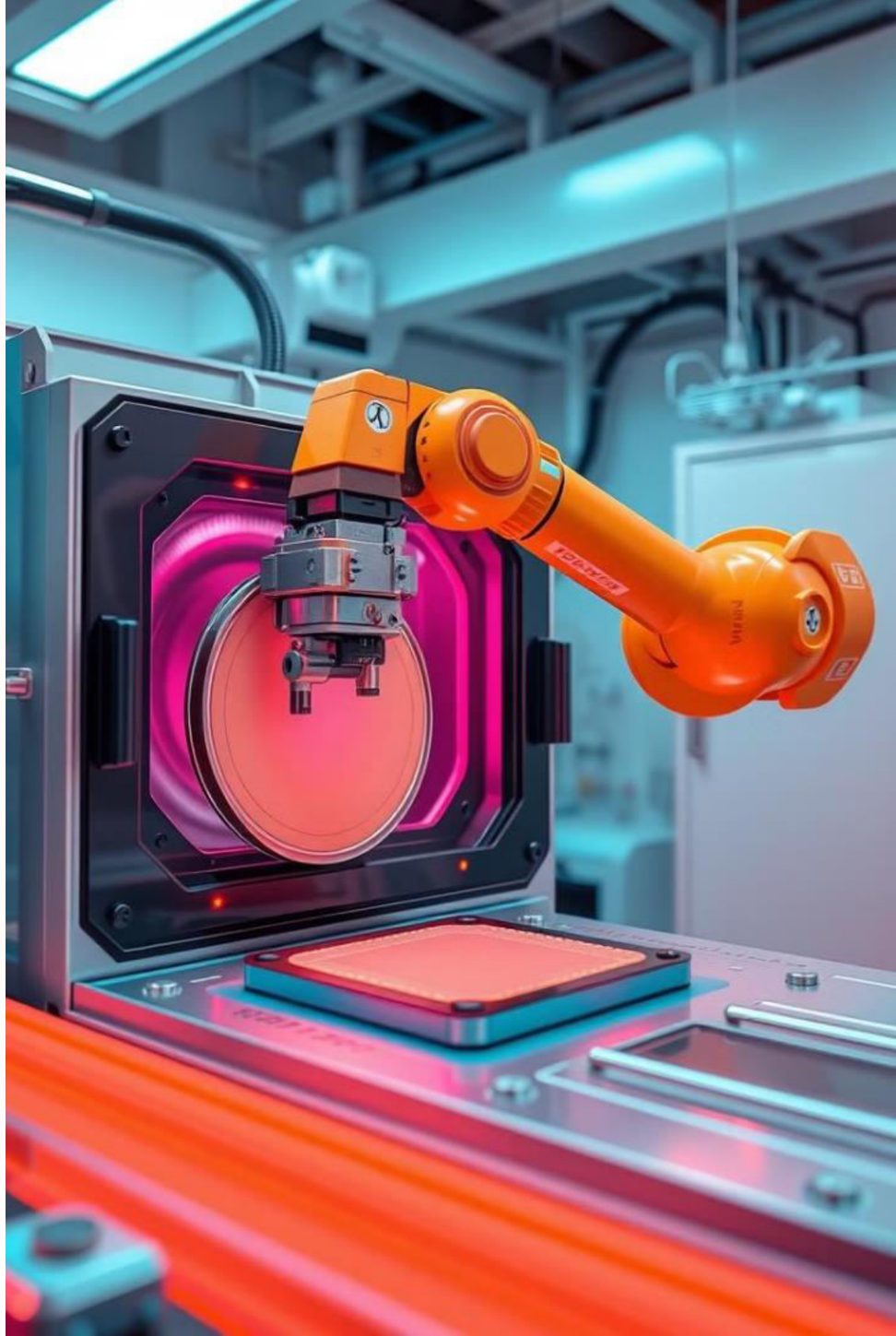


FIGURE 3.10. Hewlett-Packard's CMOSC process

N-Well CMOS Fabrication Process

This presentation explores the n-well CMOS fabrication process, a cornerstone of modern microchip manufacturing. We'll delve into its methodology, advantages, and limitations, emphasizing its historical significance and contemporary applications. We'll also touch on the process steps involved, from substrate preparation to metallization. By the end of this presentation, you'll gain a comprehensive understanding of the n-well process and its role in shaping the landscape of integrated circuit design and fabrication.



The Genesis of N-Well CMOS Technology

Initially, p-well processes dominated CMOS fabrication. However, the n-well process gained traction due to its compatibility with existing nMOS production lines. This "retrofitting" capability allowed manufacturers to integrate CMOS technology without significant infrastructure overhauls, expediting its adoption. Key advantages include optimized n-channel characteristics and high packing density. The n-well CMOS fabrication process enables manufacturers to leverage existing nMOS infrastructure while integrating CMOS technology, creating a path for innovation in integrated circuit design.

Process Compatibility

Seamless integration with nMOS production lines, reducing capital expenditure.

Performance Optimization

Superior n-channel characteristics, leading to enhanced device performance.

Made with Gamma

Fabrication Steps: A Detailed Examination

The n-well fabrication process mirrors the p-well approach, substituting an n-well for a p-well. The initial step involves defining n-well regions using a specific mask, followed by a low-dose phosphorus implant driven in by high-temperature diffusion. Precise well depth is crucial, balancing p-substrate breakdown protection with n-well separation to prevent short circuits. Subsequent steps include device definition, oxide growth, contact cuts, and metallization. This process involves precise control over implantation and diffusion parameters to achieve desired device characteristics. By carefully managing these parameters, engineers can optimize device performance for specific applications.

1

N-Well Definition

Masking and phosphorus implantation to create the n-well region.

2

Device Definition

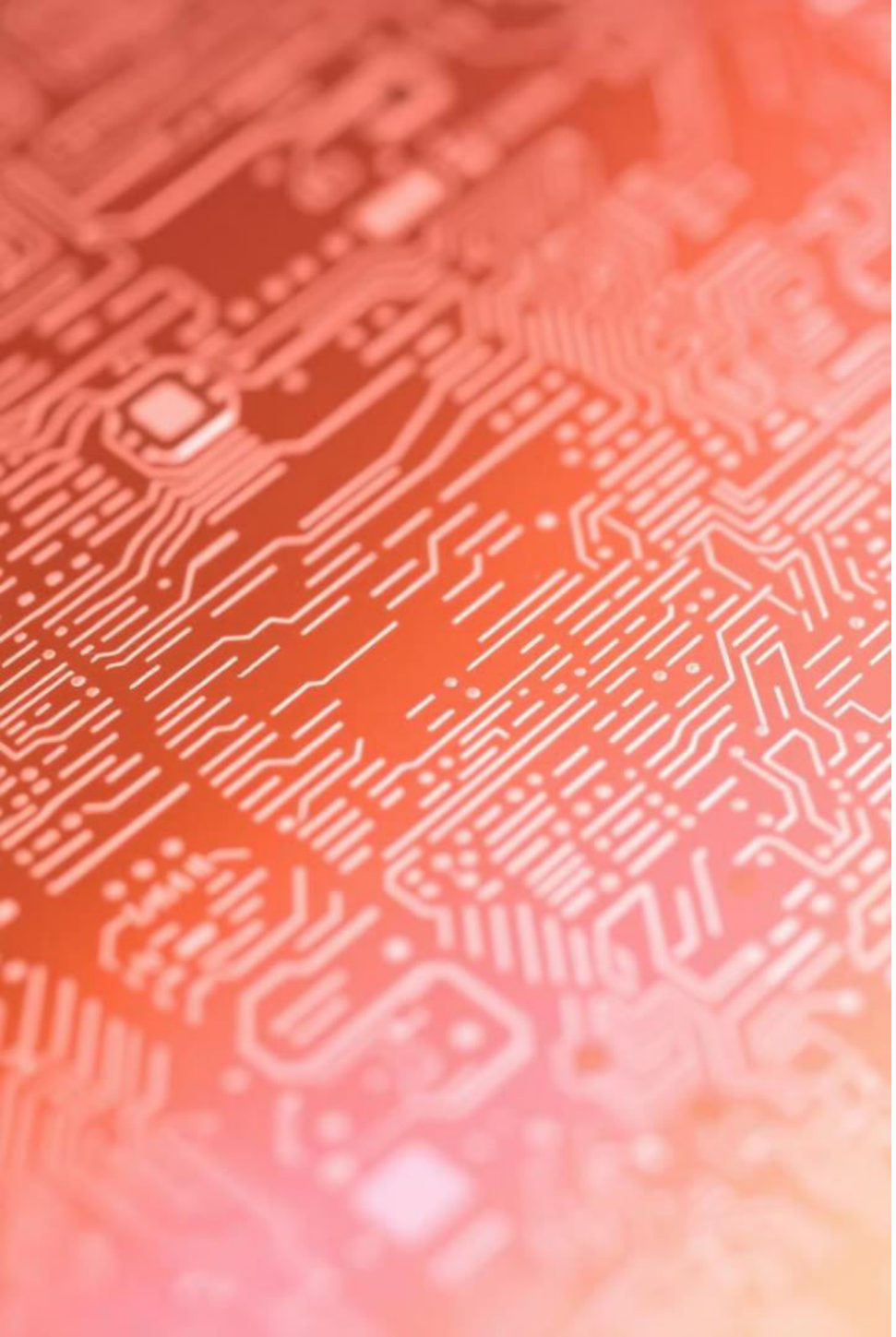
Defining transistor channels and other diffusion areas.

3

Metallization

Applying metal layers to create electrical connections.

Made with Gamma



Masking Techniques: N+ vs. P+ Masks

Masking plays a vital role in defining regions for n-channel and p-channel transistors. An n+ mask delineates n-channel transistors and VDD contacts. Alternatively, a p+ mask can define p-channel transistors, with the two masks often serving as complements. The choice between n+ and p+ masking is dictated by design requirements. In circuits requiring more p-channel transistors, the p+ mask is favored. Mask alignment and accuracy are paramount to ensure correct device performance. Precise alignment minimizes overlap and ensures proper transistor operation.

1

N+ Mask

Defines n-channel transistors and VDD contacts.

2

P+ Mask

Defines p-channel transistors (complementary to N+ mask).

Made with Gamma

Process Input Description Language (PIDL)

PIDL provides a structured way to define fabrication steps. Key commands include SUBSTRATE, OXIDE, DEPOSITION, ETCH, DOPE, and MASK. SUBSTRATE specifies substrate type and impurity level. OXIDE defines oxide layer thickness. DEPOSITION details layer deposition, including material type. ETCH defines material removal. DOPE specifies diffusion parameters such as type, peak concentration, and depth. MASK indicates resist layer information. Using PIDL enables precise control and documentation of the fabrication process. This standardization enhances reproducibility and process optimization.

SUBSTRATE

Specifies substrate type and impuri
level.

OXIDE

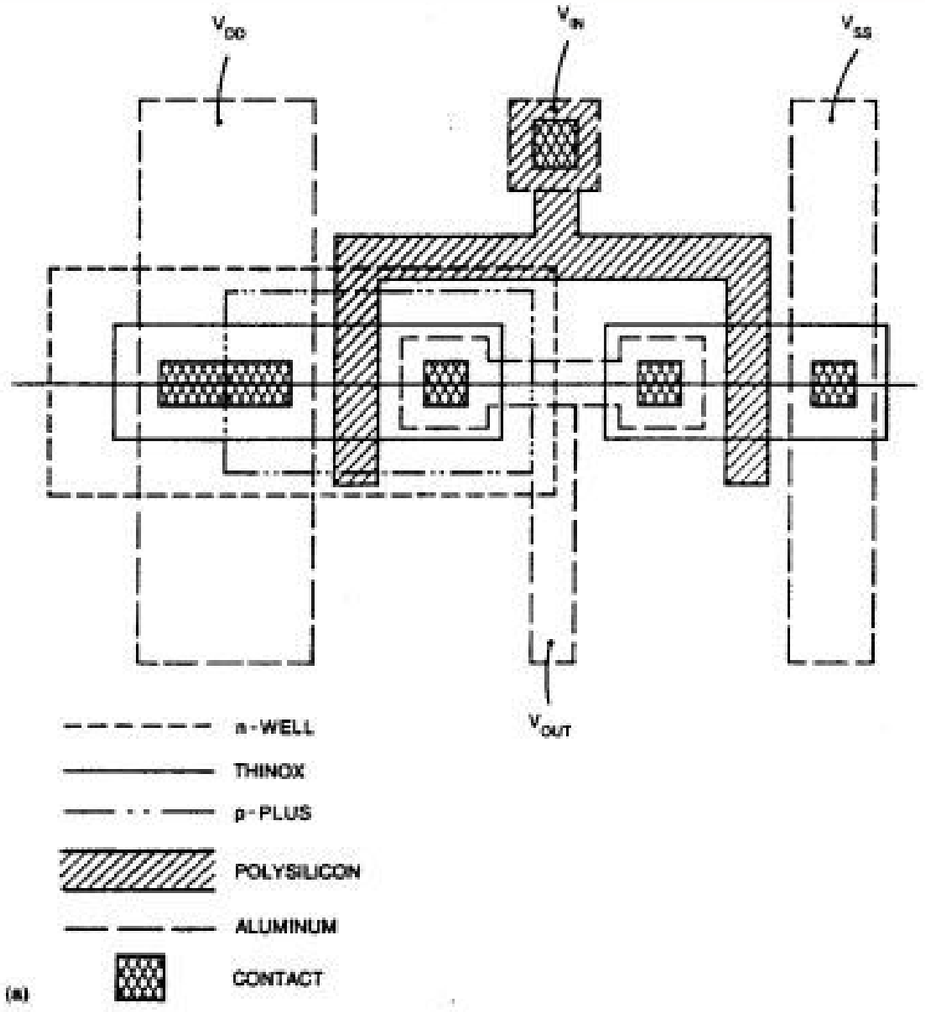
Defines oxide layer thickness.

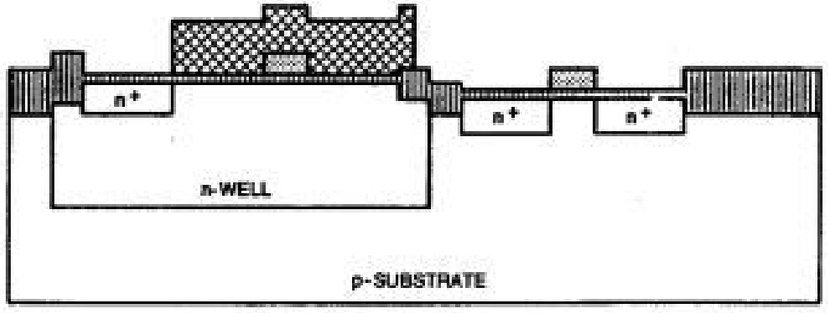
DEPOSITION

Details layer deposition.

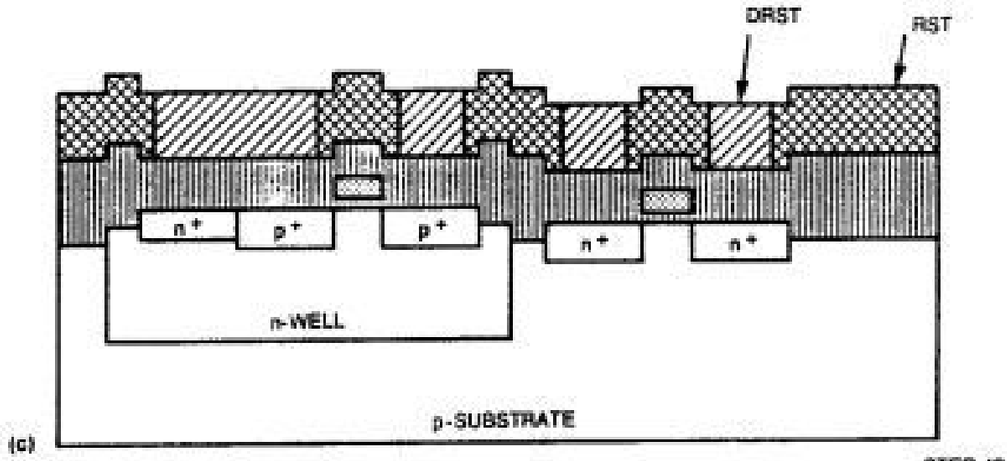
- SUBSTRATE <NAME> (*TYPE=[P,N] IMPURITY=[])
  Specifies the substrate name, type, and impurity level.
- OXIDE <NAME> THICKNESS = []
  Specifies oxide layer and thickness.
- DEPOSITION <NAME> (*) THICKNESS=[]
  Specifies a layer and thickness of a deposited layer. (*) is followed by TYPE=[] IMPURITY=[] if it is silicon.
- ETCH <NAME> DEPTH=[]
  Specifies a material and an etch depth.
- DOPE TYPE=[P,N] PEAK=[] DEPTH=[] DELTA=[] BLOCK=[]
  Specifies parameters necessary to define a diffusion step.
- MASK <RESIST NAME> <EXPOSED NAME> <MASK NAME> <POLARITY OF MASK>
  Specifies a resist layer and associated information.

Made with Gamma

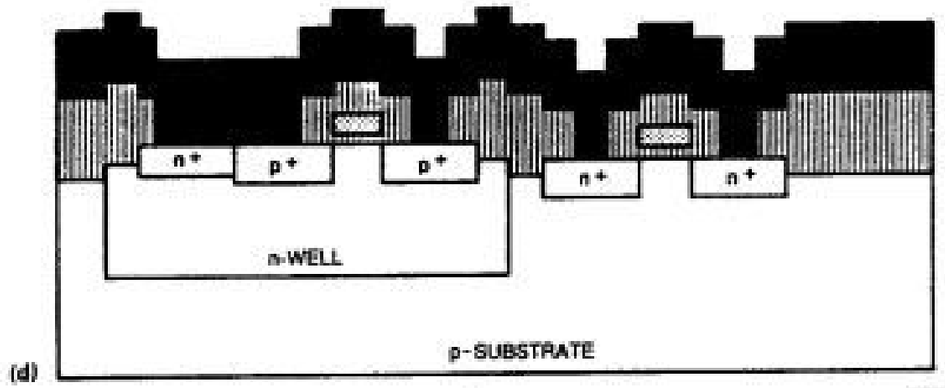




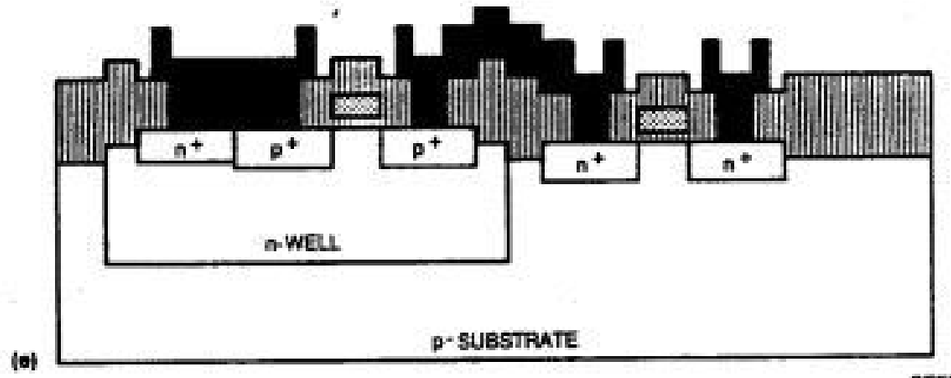
STEP 40



STEP 49



STEP 53



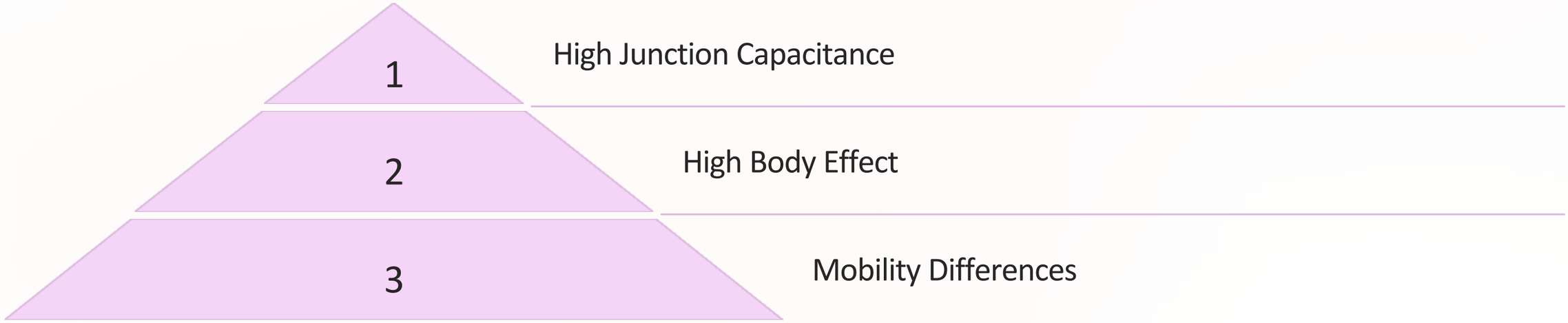
STEP 58

FIGURE 3.12. Berkeley n-well process snap-shots and layout for n-well inverter © IEEE 1983 ([GrLN83])

Made with Gamma

Challenges: Non-Optimal P-Channel Characteristics

Due to mobility differences, n-well processes yield sub-optimal p-channel characteristics, including high junction capacitance and significant body effect. This contrasts with p-well processes, where n-transistor performance may be compromised. However, careful circuit design can mitigate these effects. In designs with a higher proportion of n-channel devices, the impact of poor p-transistor performance can be minimized. Careful layout techniques and transistor sizing can reduce parasitic capacitances.



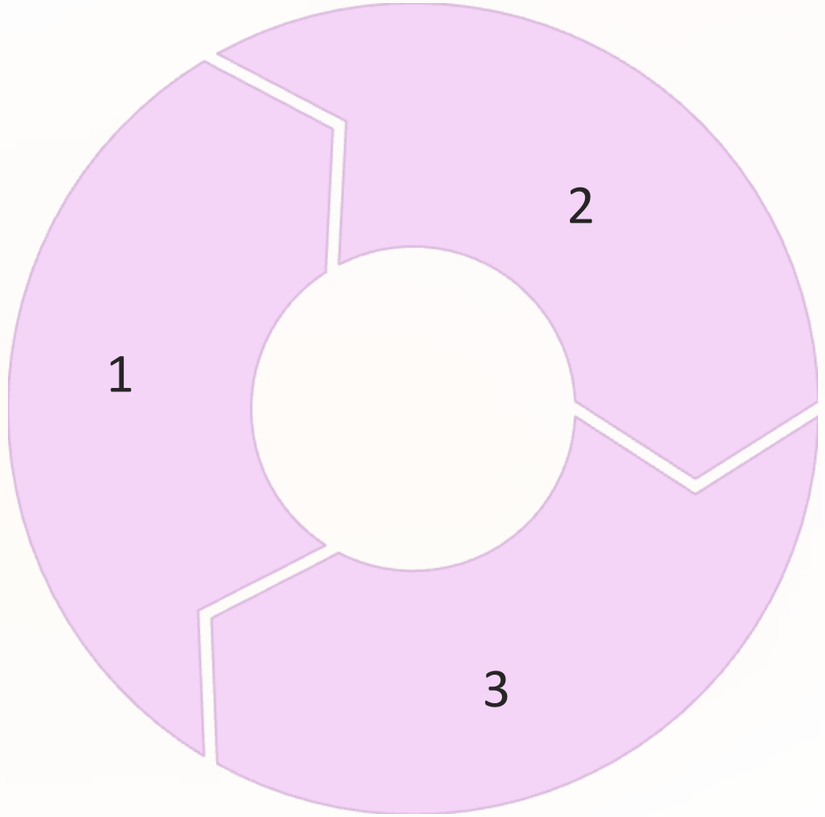
Made with Gamma

Advantages: Optimizing N-Channel Transistors

N-well technology offers a distinct advantage when optimum device characteristics are primarily needed for n-channel transistors. N-channel devices can be used to form logic elements for speed and density. P-transistors can serve mainly as pull-up devices. Fully n-type I/O circuits can also be used to advantage. This strategy optimizes performance where it matters most. By optimizing n-channel transistors, designers can achieve higher switching speeds and lower power consumption. Careful design and simulation are used to ensure performance targets are met.

Speed

Achieve faster switching speeds.



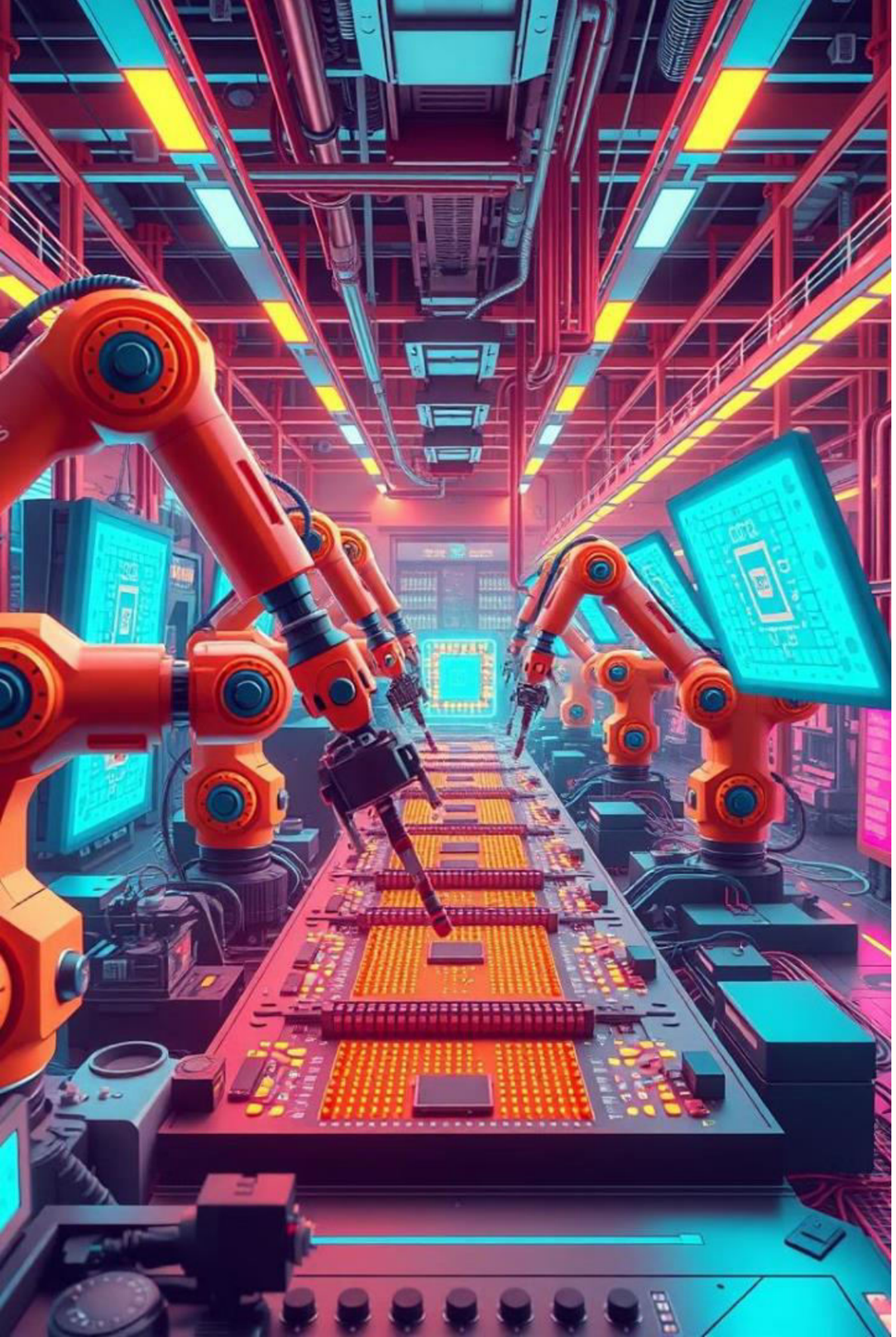
Density

Enable higher packing density.

Performance

Optimize device characteristics.

Made with Gamma



Key Takeaways and Future Trends Trends

The n-well CMOS process offered a strategic advantage by leveraging existing nMOS infrastructure. While p-channel characteristics pose challenges, careful design optimizes n-channel performance. PIDL standardizes fabrication steps, enhancing reproducibility. Modern CMOS designs often contain more n-channel than p-channel devices. N-well technology provides a distinct advantage by optimizing n-channel transistors. The future of n-well CMOS involves continuous refinements in fabrication techniques and materials. These advances will enable higher performance and density in integrated circuits.

Process Evolution

Design Optimization Optimization

Performance Gains

Made with Gamma

Twin-Tub Technology

Twin-tub CMOS technology is a fabrication technique in integrated circuit manufacturing. It enables independent optimization of P-type and N-type transistors. This leads to separately optimized wells, enabling the creation of improved N-transistors with lower capacitance and reduced body effect compared to conventional P-well processes.

Twin-Tub Process Steps

Tub Formation

Creation of both P-wells and N-wells in the substrate. This is the foundational step, differentiating twin-tub from single-well processes.

Thin Oxide Etching

Selective removal of thin oxide layers to define active areas for transistor formation. This step precisely patterns the silicon surface.

Source and Drain Implantations

Introduction of dopant ions to form the source and drain regions of the transistors. These implants define the transistor characteristics.

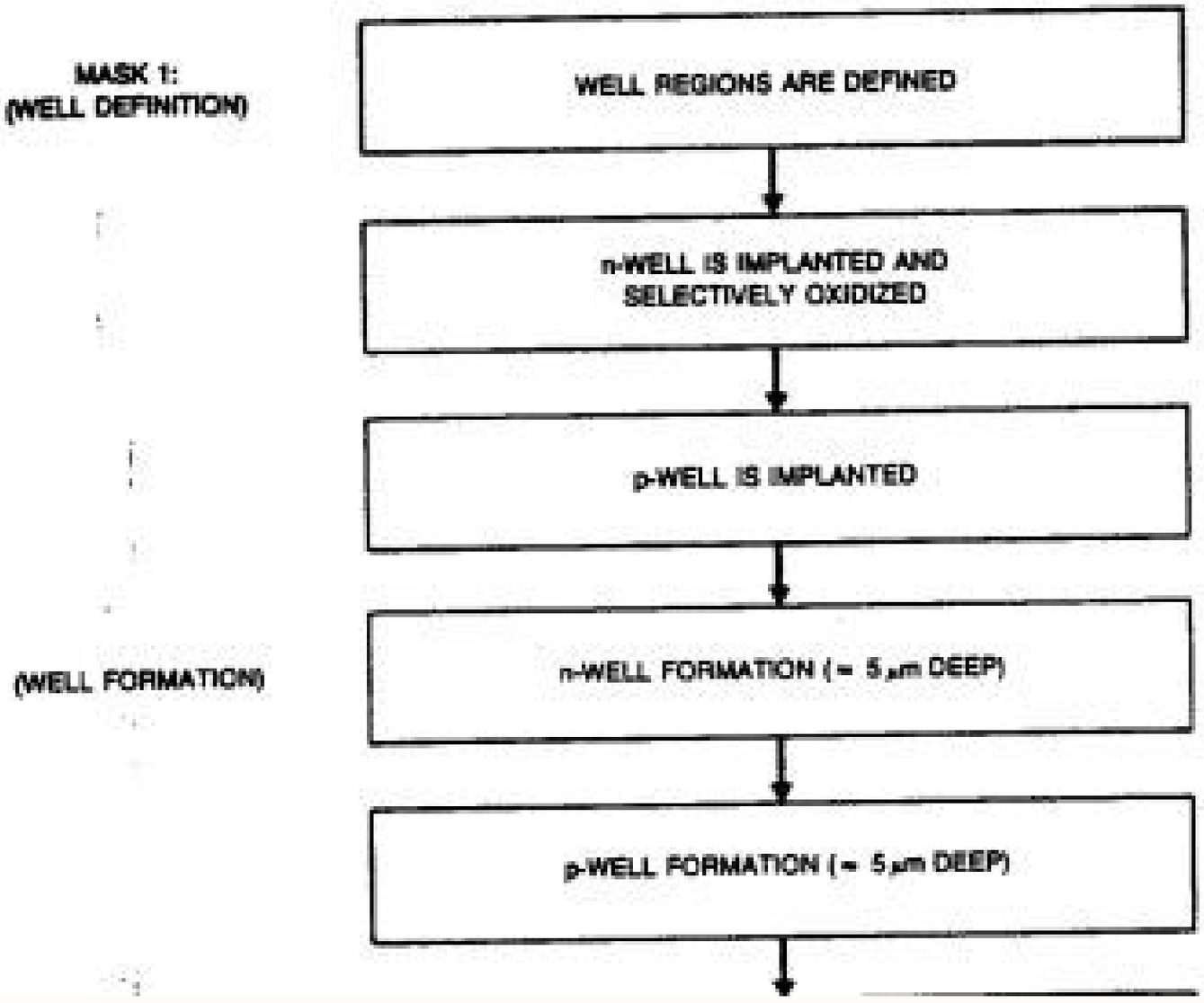
Contact Cut Definition

Etching of contact openings to allow electrical connections to the transistors. This step enables metallization.

Metallization

Deposition of metal layers to create interconnections between transistors and external circuits. This step completes the circuit.

Made with Gamma



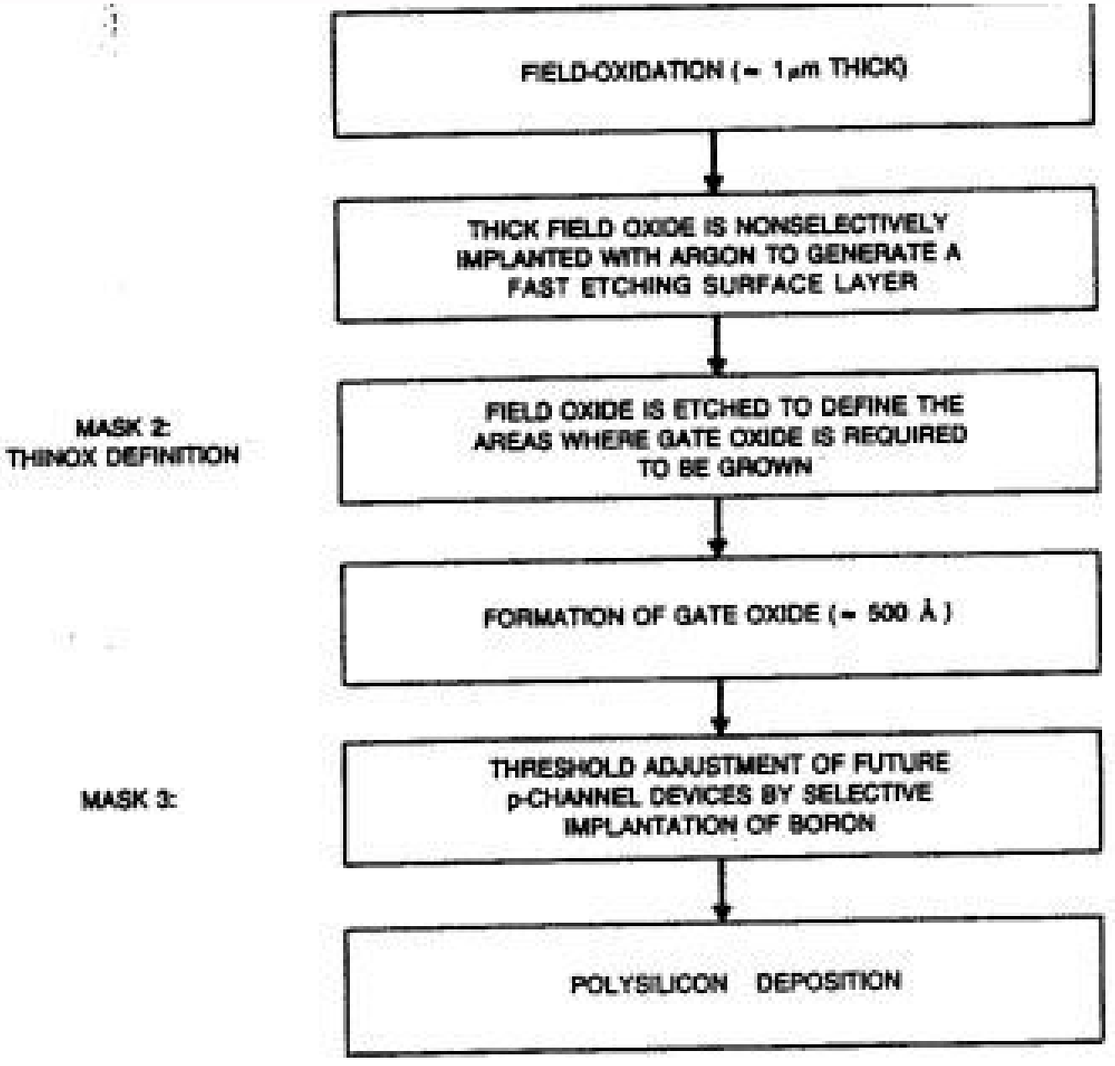


FIGURE 3.13. AT&T Bell Laboratories' twin-tub CMOS process steps

Made with Gamma

MASK 4:
(POLYSILICON PATTERNING)

POLYSILICON IS PATTERNED

BORON IMPLANT* FOR FORMATION OF p^+ REGIONS

MASK 5:

PHOSPHOROUS IS IMPLANTED FOR FORMATION OF n^+ REGIONS

PASSIVATION

MASK 6:

CONTACT CUT

ALUMINIUM DEPOSITED OVER THE WHOLE OF WAFER

MASK 7:

PATTERNING OF ALUMINIUM

MASK 8:

CUTS FOR BONDING PADS

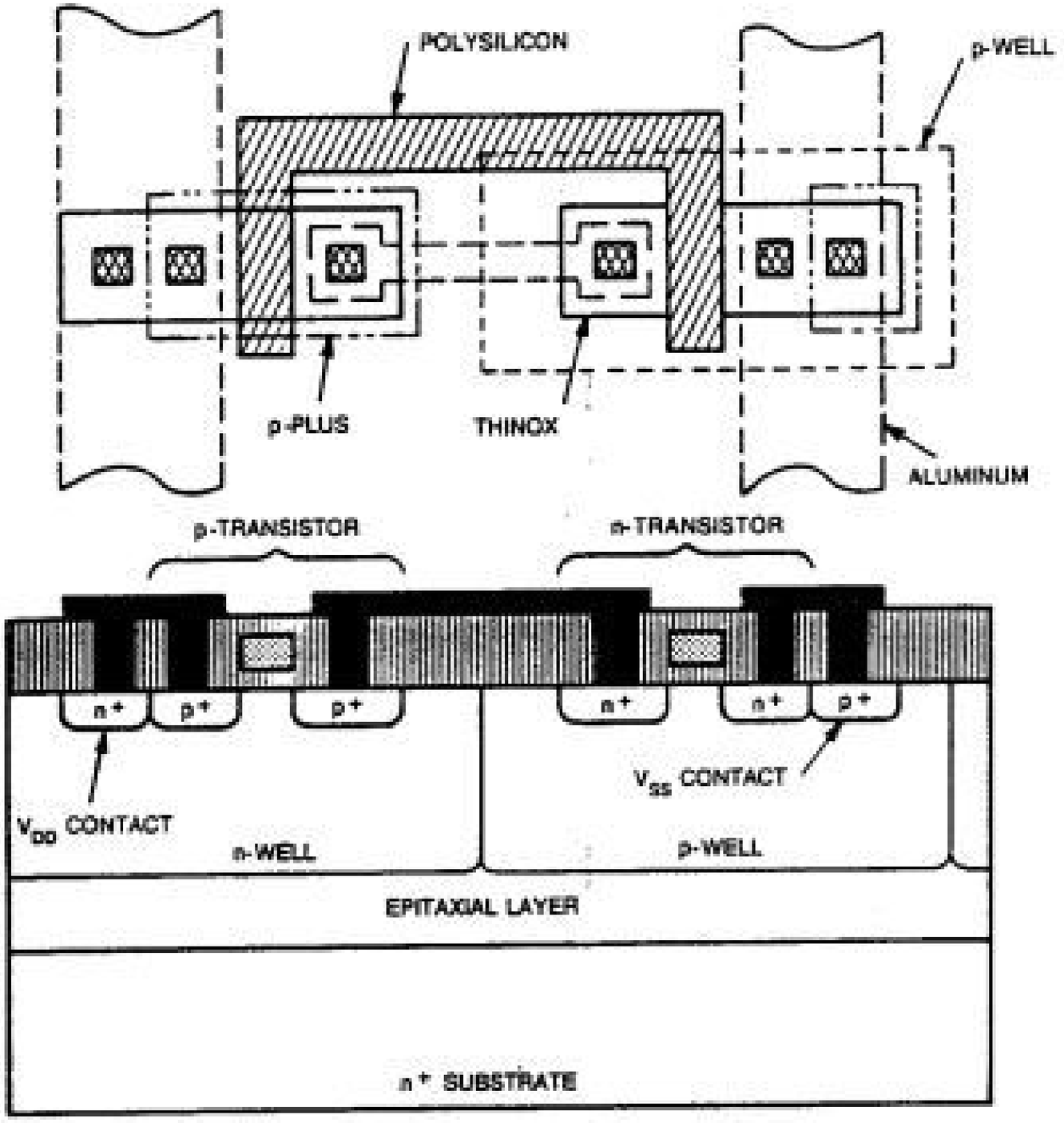
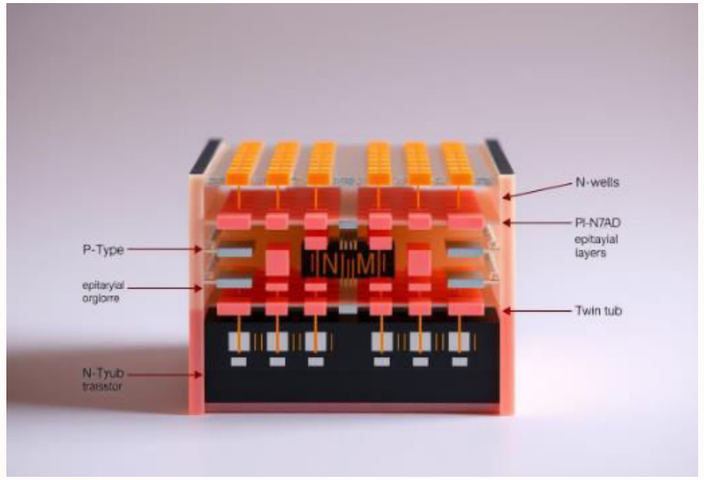


FIGURE 3.14. Twin-tub process cross-section and layout of an inverter

Epitaxial Layer and Latch-Up Prevention

Epitaxial Layer

A lightly doped epitaxial layer, grown on either an N+ or P+ substrate, provides protection against latch-up, a condition that can cause circuit failure.



High-Purity Silicon

The aim of epitaxy is to grow high-purity silicon layers of controlled thickness, with accurately determined dopant concentrations distributed homogeneously throughout the layer.

Made with Gamma

Benefits of Twin-Tub Technology

Independent Optimization

Threshold voltage, body effect, and gain can be independently optimized for N- and P-type transistors.

Improved Performance

Better performance N-transistors (lower capacitance, less body effect) may be constructed compared with a conventional P-well process.

Threshold Adjustment

The process includes threshold adjust steps, using masks derived from the thinox and n-plus masks, to fine-tune transistor characteristics.

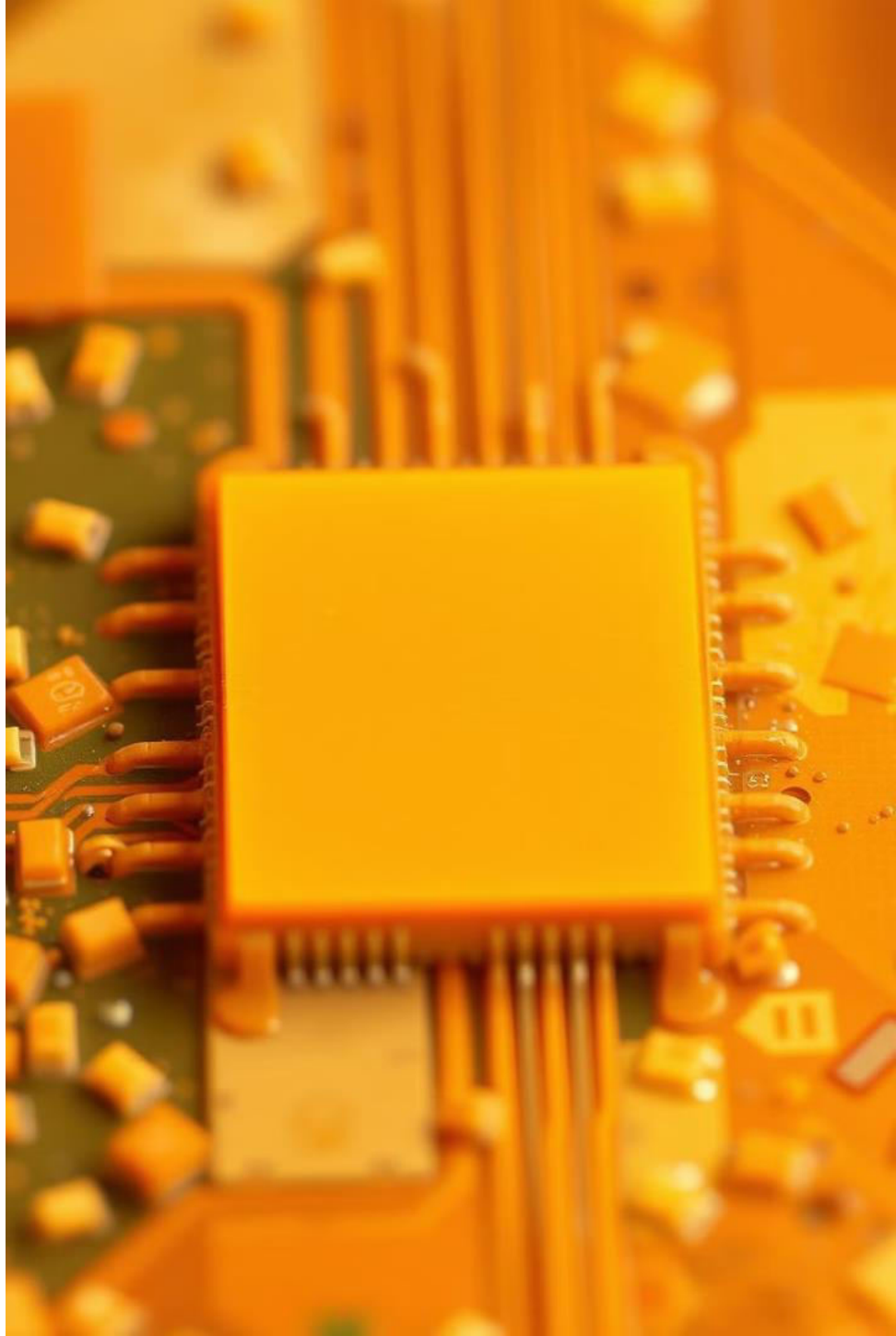
Made with Gamma

Silicon on Insulator (SOI) CMOS Technology

Silicon on Insulator (SOI) CMOS processes offer significant advantages over traditional CMOS technologies, including higher density, no latch-up issues, and lower parasitic capacitances. In SOI, a thin layer of single crystal silicon film is grown on an insulator like sapphire.

This presentation explores the SOI process, its advantages, and its limitations, providing a comprehensive overview of this advanced technology.

by CITechnocrats07



SOI CMOS Process Steps

Step1: Thin Film Growth

A thin film (7-8 µm) of lightly-doped n-type Si is grown over an insulator, commonly sapphire.

Step 2: Anisotropic Etch

An anisotropic etch removes Si except where diffusion areas (n or p) are needed, creating Si "islands."

Step3: P-Island Formation

P-islands are formed by masking n-islands with photoresist and implanting a p-type dopant like boron.

Made with Gamma

Gate Oxide and Polysilicon

1

Step 4: Thin Gate Oxide

A thin gate oxide (around 500-600 Å) is grown over all Si structures via thermal oxidation.

2

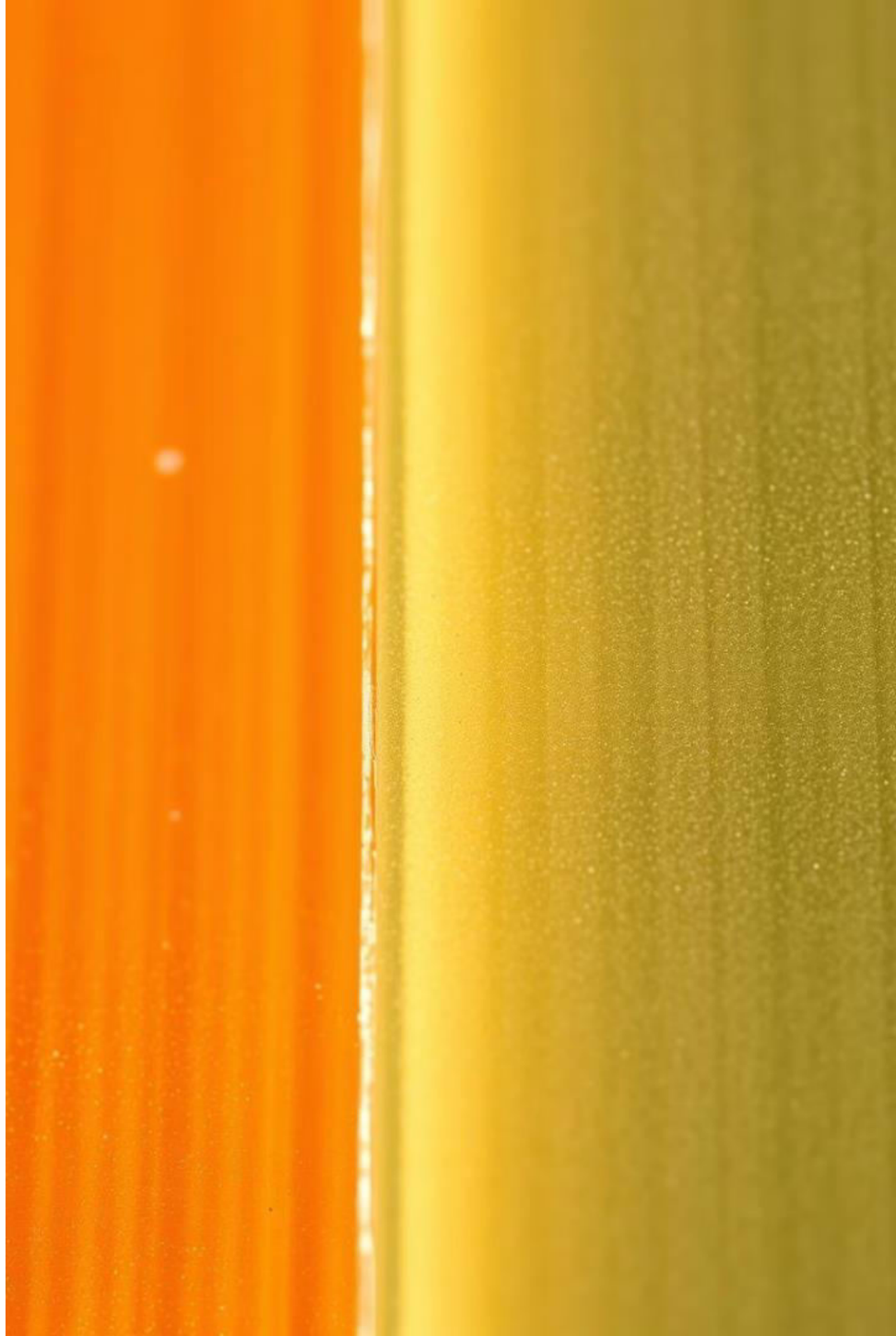
Step 5: Polysilicon Deposition

A polysilicon film is deposited over the oxide, often doped with phosphorus to reduce resistivity.

3

Step 6: Polysilicon Patterning

The polysilicon is patterned by photomasking and etching, defining the polysilicon layer.



N-Channel and P-Channel Devices Devices

Step 7: N-Doped Source and Drain

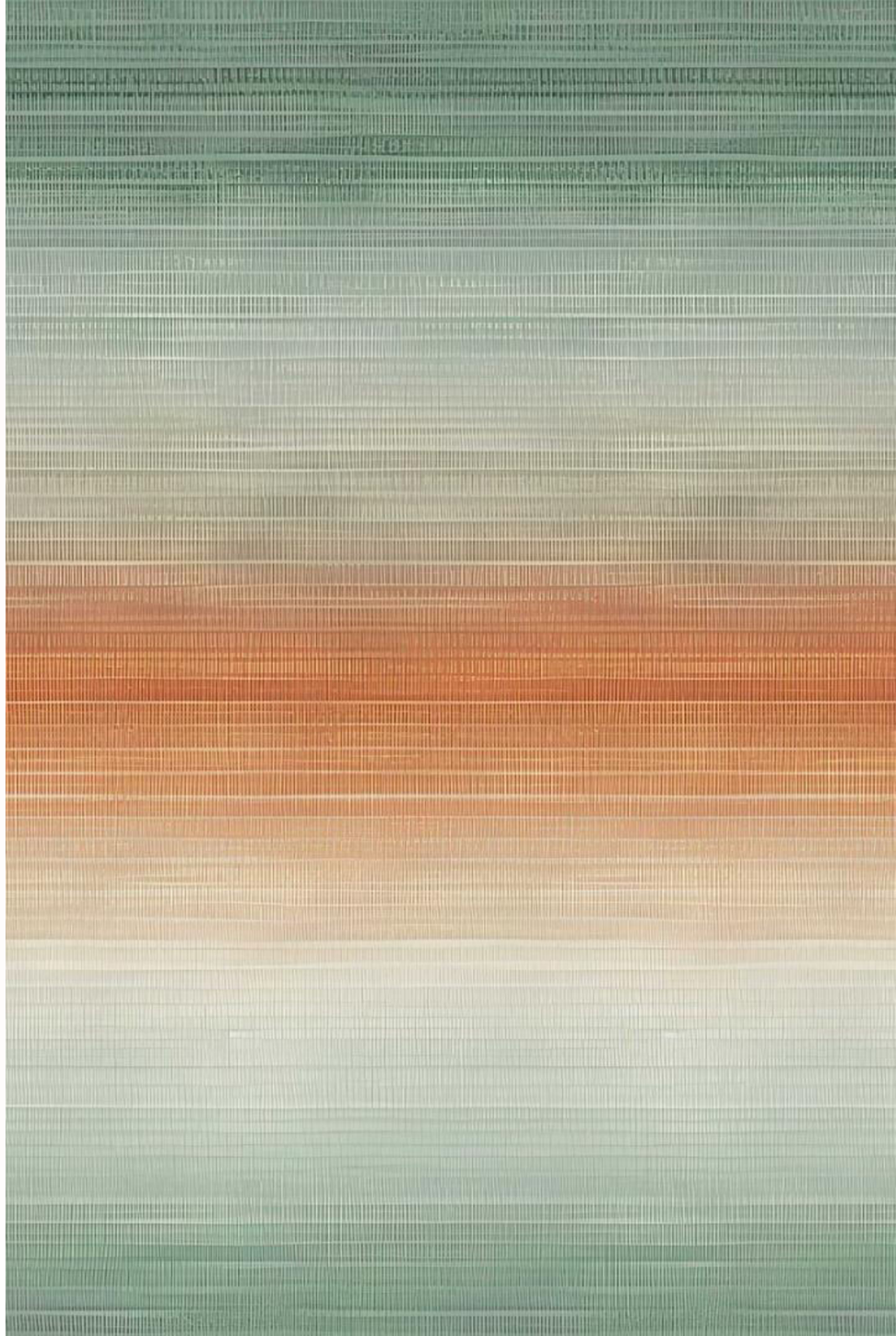
N-doped source and drain of n-channel devices are formed in p-islands by masking n-islands and implanting phosphorus.

Step 8: P-Channel Formation

P-channel devices are formed by masking p-islands and implanting a p-type dopant like boron.

Step 9: Insulator Deposition

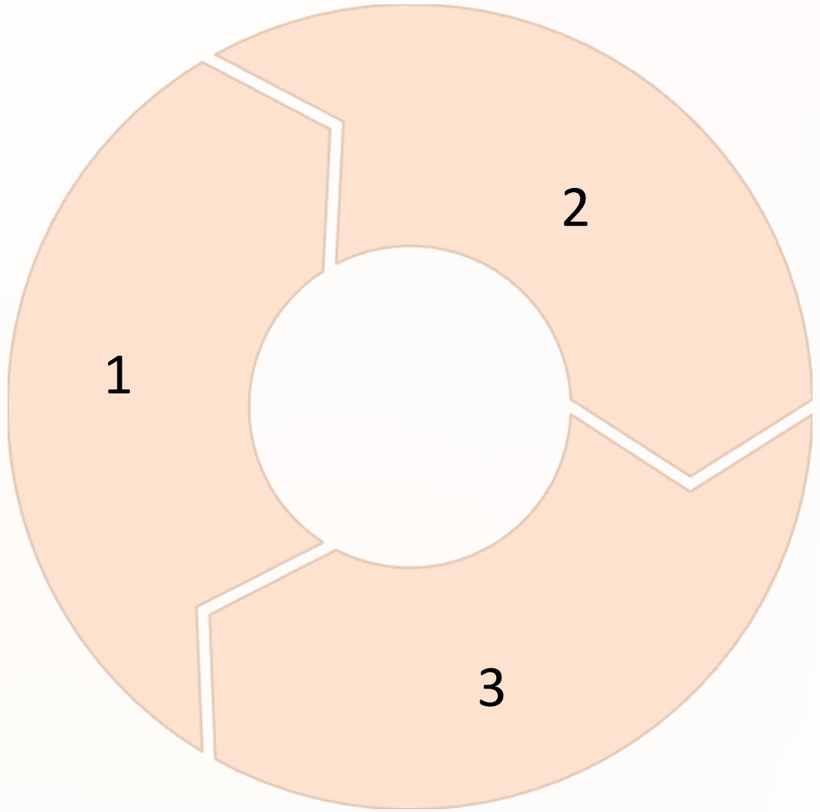
A layer of phosphorus glass or silicon dioxide is deposited over the entire structure.



Metallization and Passivation

Step 10: Contact Cuts

The glass is etched at contact cut locations to allow metal connections.



Step 11: Metallization Layer

Aluminum is evaporated over the surface and etched to form metal wires, contacting diffusion or polysilicon regions.

Step12: Passivation Layer

A final passivation layer of phosphorus glass is deposited and etched over bonding pad locations.

Made with Gamma

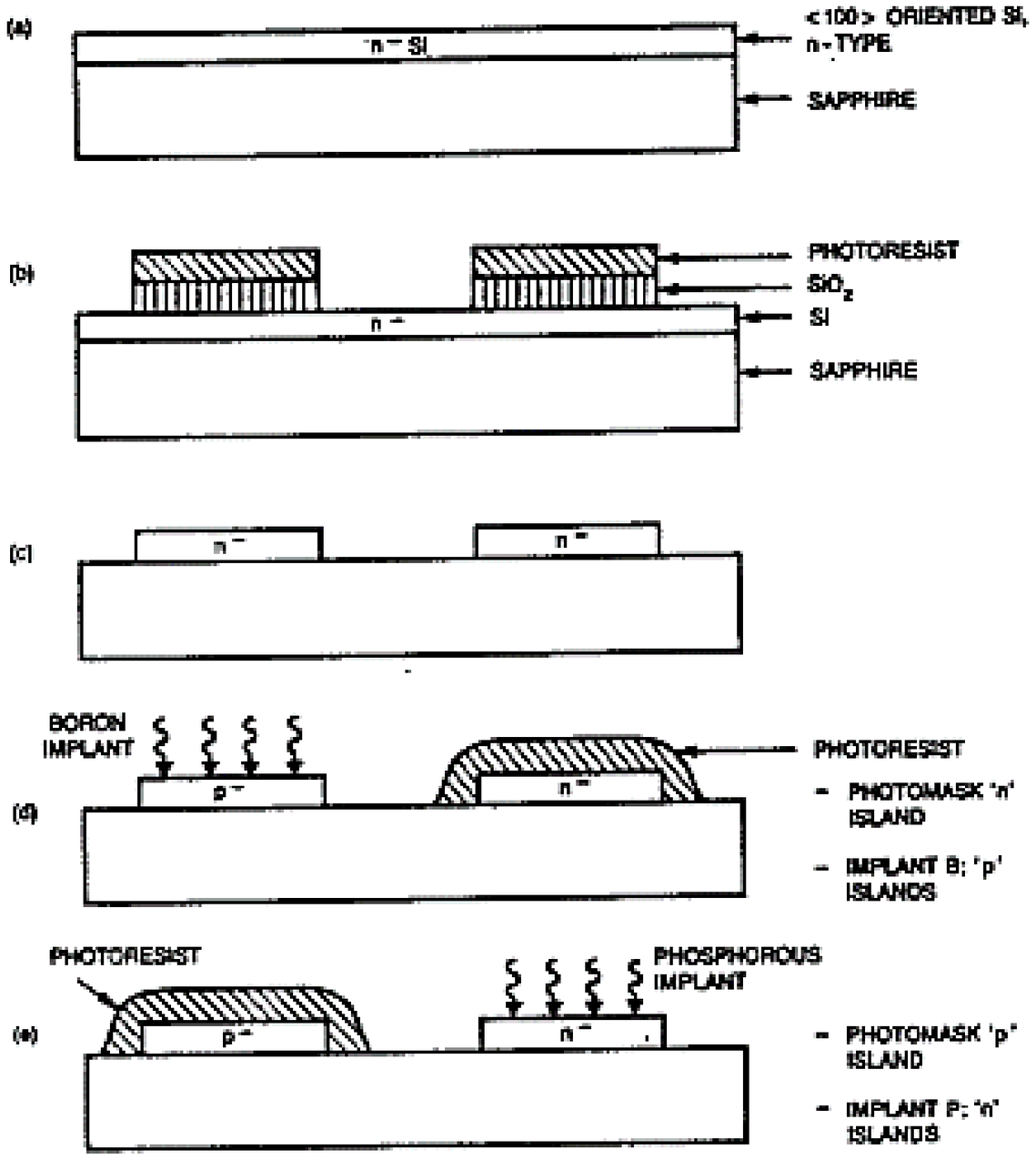
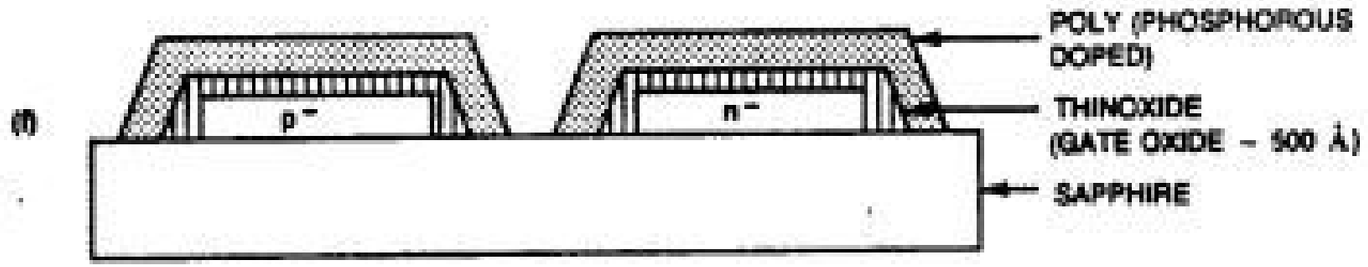
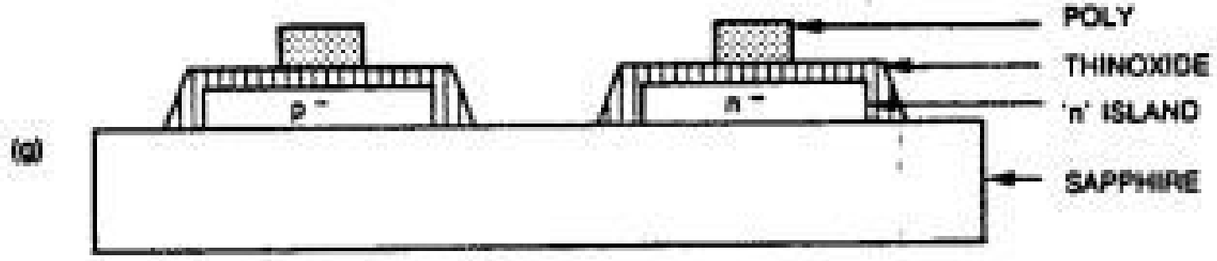
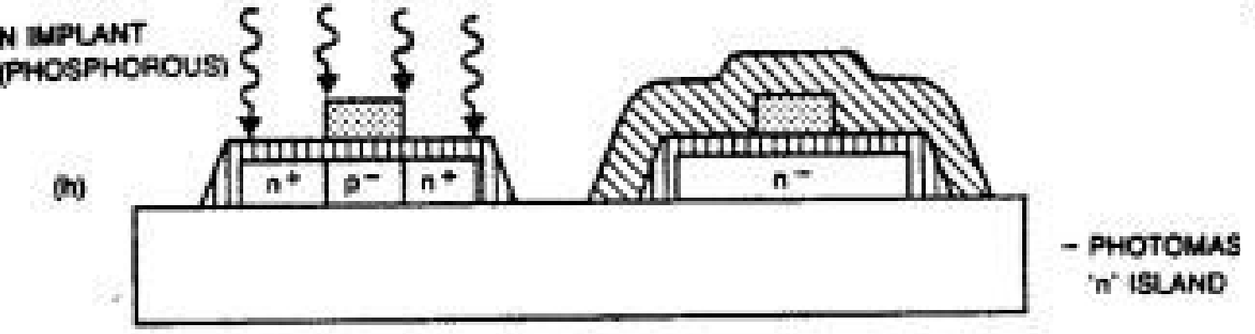


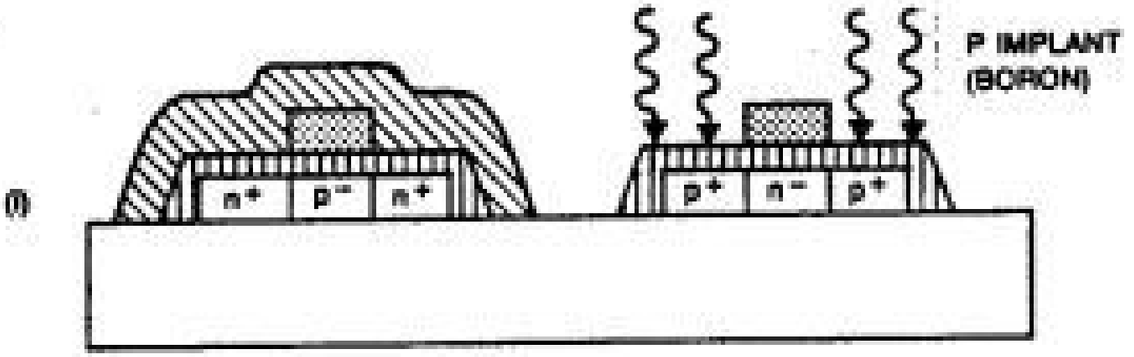
FIGURE 3.15. Typical silicon on insulator (SOI) process flow

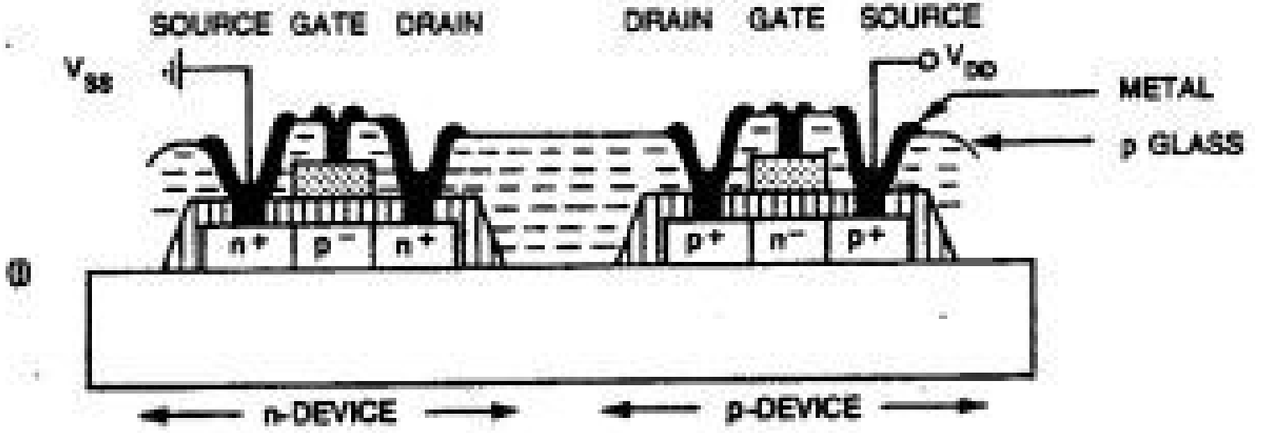
Made with Gamma


(f)
p⁻
n⁻
POLY (PHOSPHOROUS DOPED)
THINOXIDE (GATE OXIDE – 500 Å)
SAPPHIRE


(g)
p⁻
n⁻
POLY
THINOXIDE
'n' ISLAND
SAPPHIRE


N IMPLANT (PHOSPHOROUS)
(h)
n⁺
p⁻
n⁺
n⁻
– PHOTOMAS 'n' ISLAND


(i)
n⁺
p⁻
n⁺
p⁺
n⁻
p⁺
P IMPLANT (BORON)


(j)
SOURCE GATE DRAIN
DRAIN GATE SOURCE
V_{ss}
V_{DD}
METAL
p GLASS
n⁺
p⁻
n⁺
p⁺
n⁻
p⁺
n-DEVICE
p-DEVICE

Parasitic Junction Capacitance

Sidewall Areas

Only sidewall areas of source and drain diffusions contribute to parasitic junction capacitance.

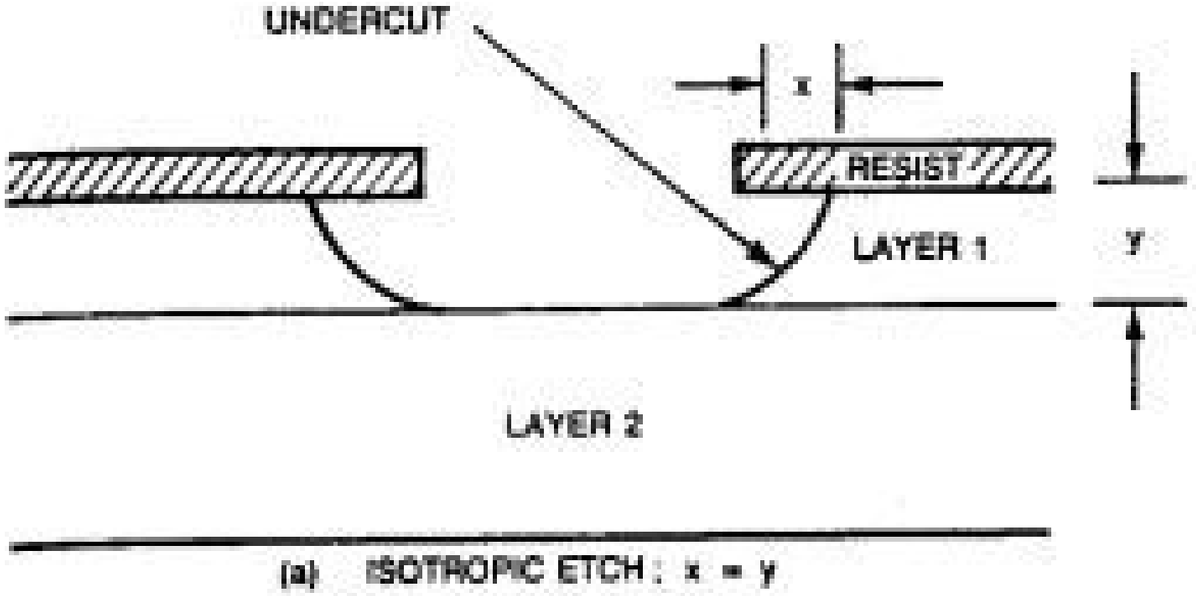
Insulating Substrate

Sapphire's excellent insulation minimizes leakage currents between transistors and the substrate.

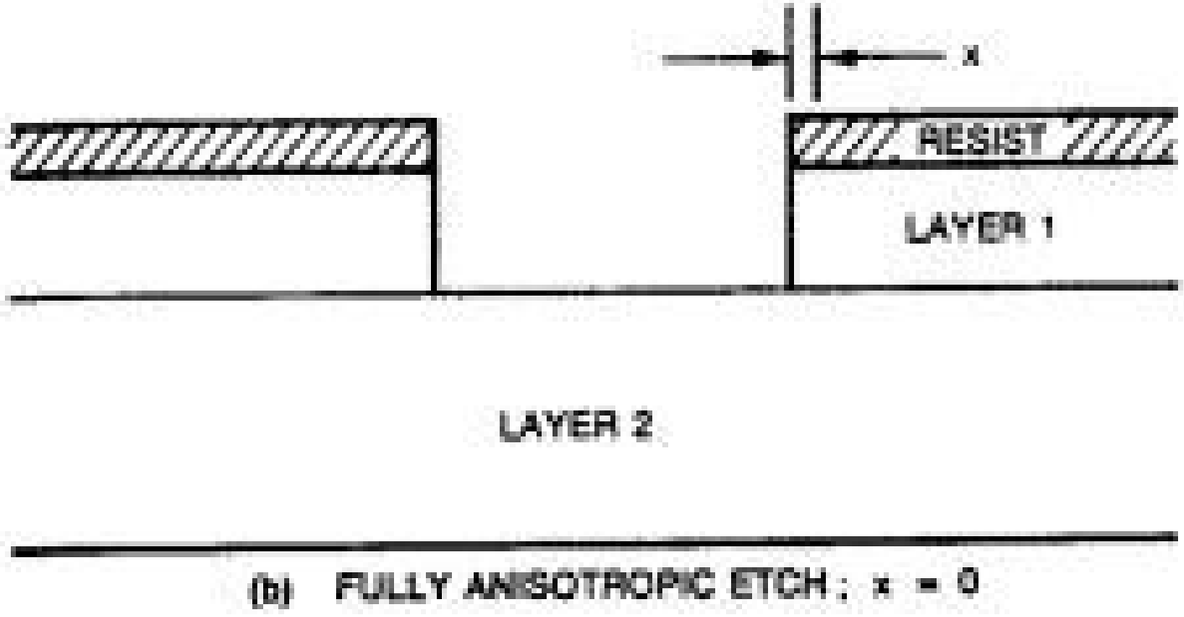
Preferential Etch

Some processes use "preferential etch" to taper island edges, improving yield.

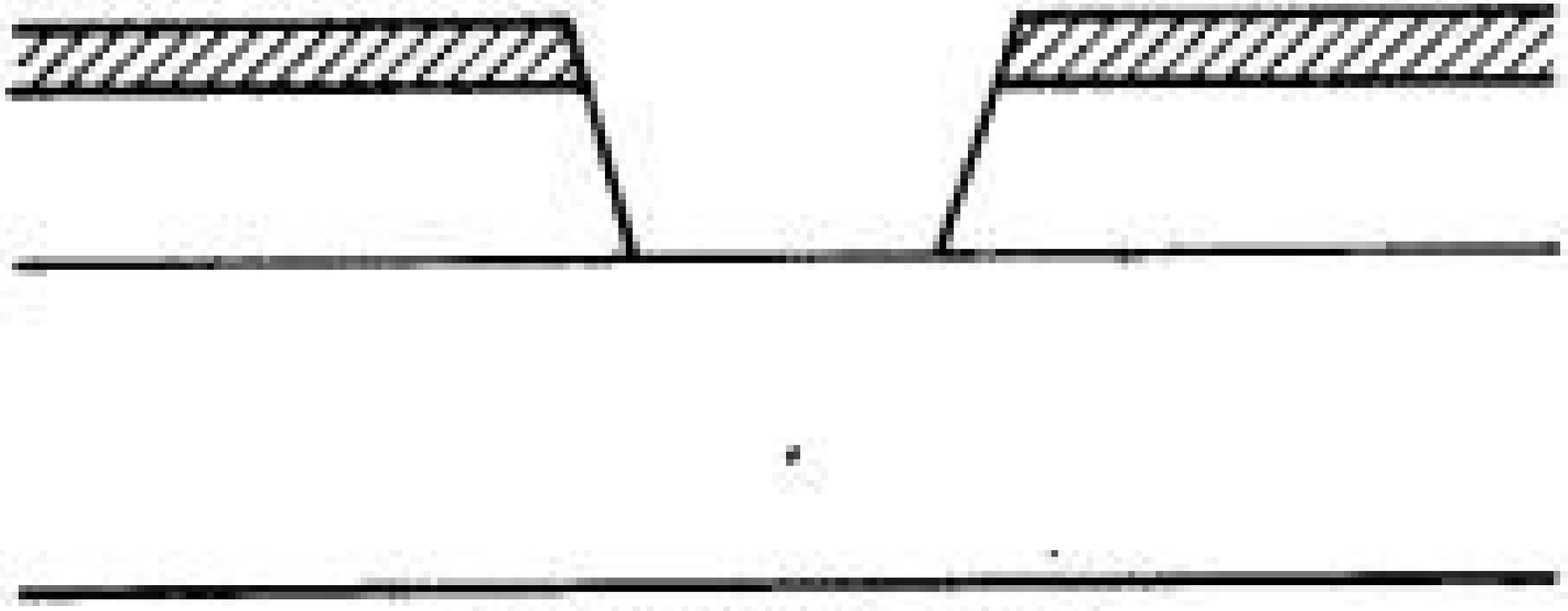
Made with Gamma



(a) ISOTROPIC ETCH ; x = y



(b) FULLY ANISOTROPIC ETCH ; x = 0



(c) PREFERENTIAL ETCH

Made with Gamma

Advantages of SOI Technology

Denser Structures

Absence of wells allows for denser structures than bulk silicon.

Faster Circuits

Low capacitances enable very fast circuits.

No Latch-Up

Isolation of n- and p-transistors by insulating substrate prevents latch-up.

SOI technology offers several advantages, including denser structures, faster circuits, and enhanced radiation tolerance.

More Advantages of SOI

No Field-Inversion

No field-inversion problems due to the insulating substrate.

No Body Effect

Absence of a conducting substrate eliminates body effect problems.

Radiation Tolerance

Enhanced radiation tolerance makes SOI suitable for harsh environments.

Made with Gamma

Disadvantages of SOI Technology

1

Input Protection

More difficult input protection due to the absence of substrate diodes.

2

Lower Device Gains

Lower device gains require larger I/O structures.

3

Expensive Substrates

Single crystal sapphire or spinel substrates are more expensive than silicon.

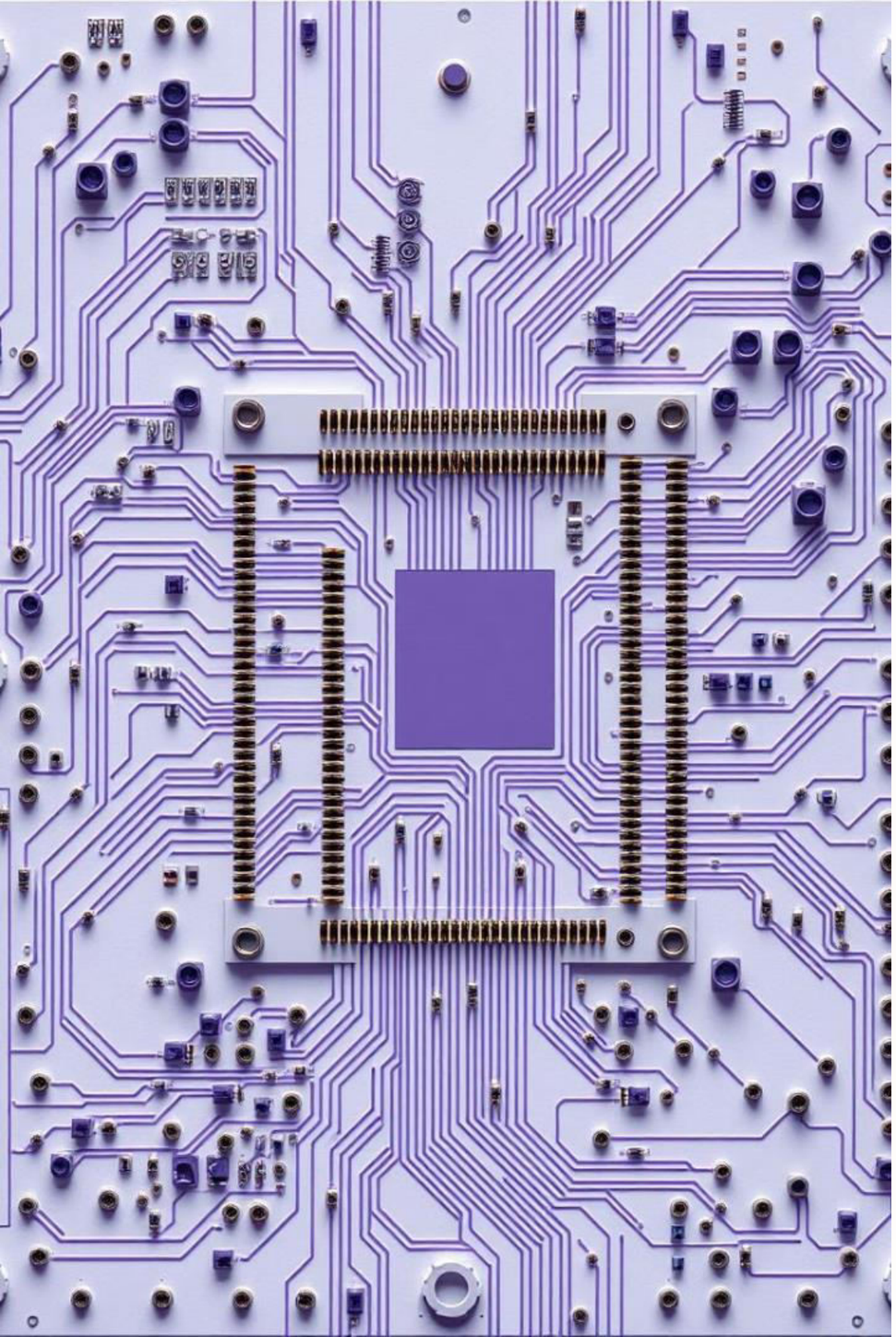
Despite its potential, SOI technology faces challenges such as higher costs and more complex processing techniques.

Made with Gamma

Conclusion

SOI technology offers significant advantages in terms of density, speed, and radiation tolerance. However, it also presents challenges related to cost and processing complexity.

While SOI has the potential to be the fastest MOS technology, it is also the most expensive. Future advancements may address these limitations, making SOI a more viable option for a wider range of applications.



Layout Design Rules in Integrated Circuit Fabrication

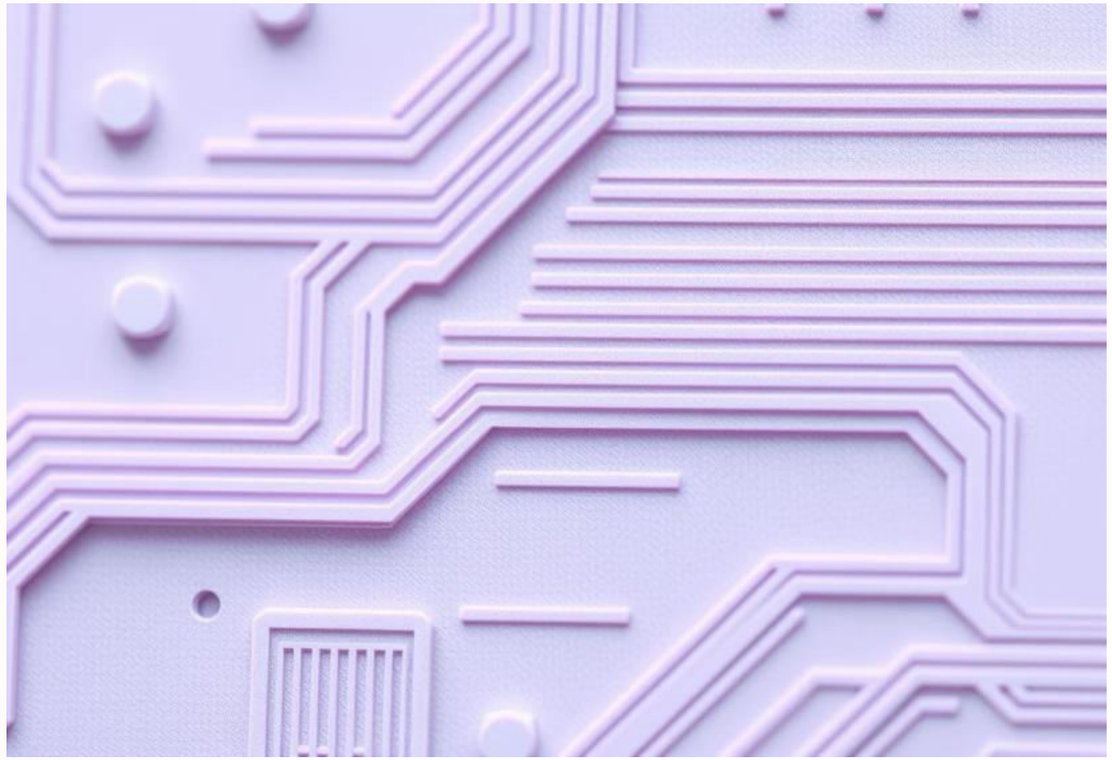
Layout design rules are crucial for creating photomasks used in integrated circuit fabrication. These rules serve as a communication bridge between circuit designers and process engineers during manufacturing. The primary goal is to produce circuits with optimal yield in the smallest possible geometry without sacrificing reliability. Design rules represent a compromise between performance and yield, where conservative rules ensure functionality, while aggressive rules enhance performance, potentially at the cost of yield.

by CITechnocrats07

Made with Gamma

Geometric Constraints and Fabrication Tolerance

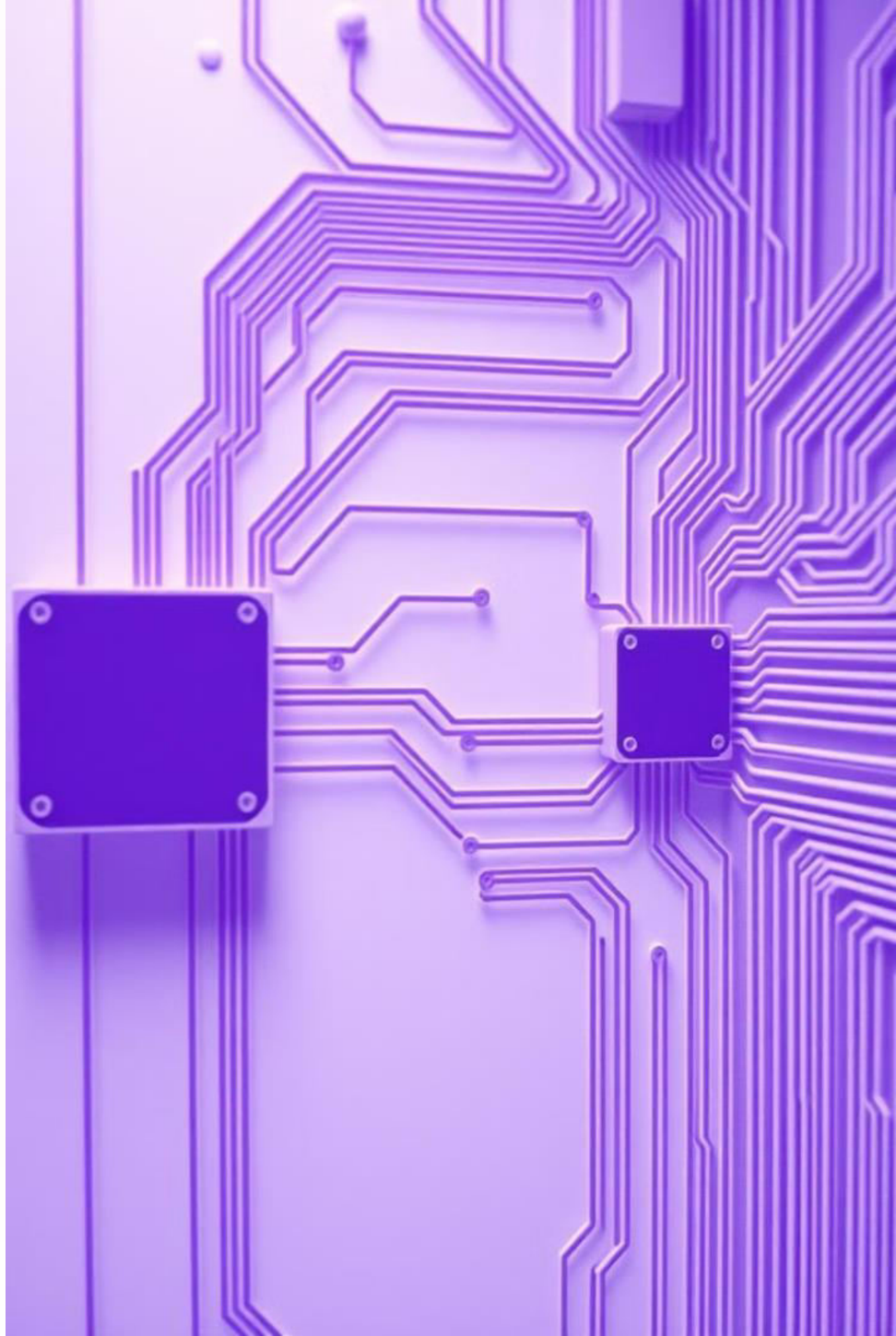
Design rules impose geometric constraints on layout artwork, ensuring that patterns on the processed wafer maintain the designs' topology and geometry. These rules aren't strict boundaries but rather tolerances that ensure a high probability of correct fabrication and operation. While layouts violating design rules may sometimes function correctly, frequent deviations can negatively impact the success of a design. Two key constraints involve line widths and interlayer registration.

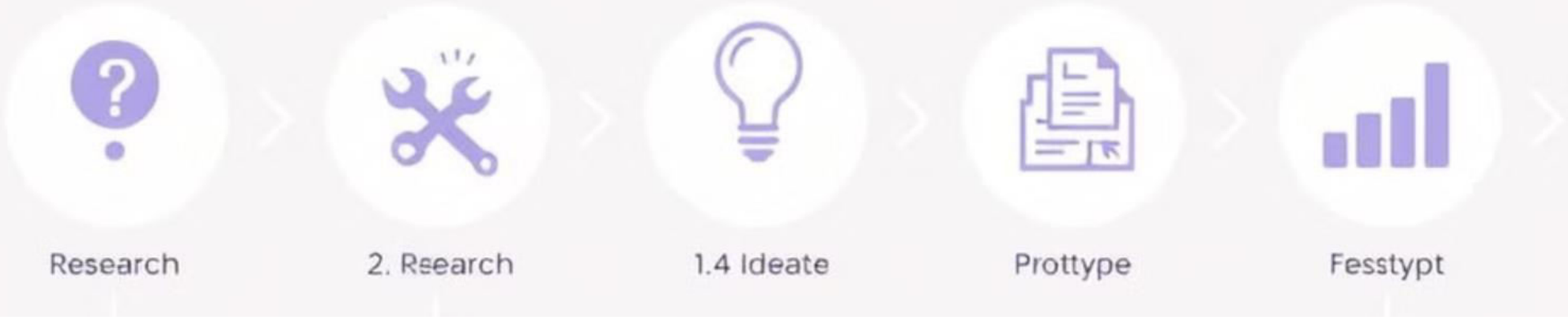


Made with Gamma

Line Widths and Interlayer Spacing

- If line widths are too small, lines may become discontinuous. Conversely, if wires are placed too close, they may merge, causing shorts between circuit nets. The spacing between layers can also be affected by the process's vertical topology.
- Design rules primarily address two issues :
  - the geometrical reproduction of features achievable through masking and lithography
  - the interactions between different layers.
- Understanding these constraints is essential for successful circuit fabrication.





Approaches to Describing Design Rules

- There are several approaches to describing design rules, including .
  - micron rules,
  - alpha (α) and beta (β) rules,
  - lambda (λ)-based rules.
- Micron design rules list minimum feature sizes and spacings for all masks in a given process.
- Alpha and beta rules define the basic feature size in terms of β, with α describing the minimum grid size.
- Lambda-based rules, popularized by Mead and Conway, use a single parameter λ to characterize the wafer implementation process's resolution.

Made with Gamma

TABLE 3.2. JPL/Mead Conway layer representation for p-well CMOS process

LAYER	JPL Color	JPL Symbolic	JPL Comments
• p-well	Brown	—	Inside brown is p-well, outside is n-type substrate.
• Thin oxide	Green	n-transistor	Thinox may not cross a well boundary.
• Poly	Red	Polysilicon	Generally n^+.
• p^-	Yellow	p-transistor	Inside is p^+.
• Metal1	Light blue	Metal1	—
• Metal2	Dark blue	Metal2	—
• Contact cut	Black	Contact	—
• Overglass	—	—	—

Made with Gamma

TABLE 3.3. Alternate layer representations for *p*-well CMOS process

LAYER	ALTERNATE COLOR	CIF CODE
• p-well	Brown	CW
• Thin oxide*	Red	CD
• Poly	Green	CP
• p^+	Purple	CS
• Metal1	Tan	CM
• Metal2	Dark blue	CN
• Contact	Black	CC
• Overglass	White stipples	CG

* This layer is referred by Mead & Conway as diffusion. n-thinox is light blue, p-thinox is purple or magenta.

Made with Gamma

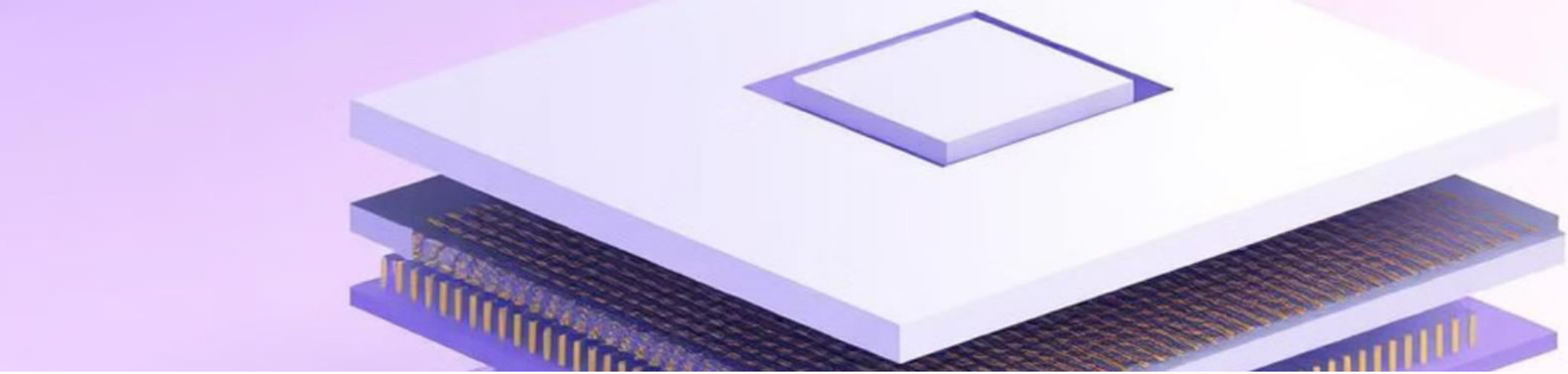
Lambda-Based Design Rules

Lambda (λ)-based design rules are based on a single parameter that characterizes the resolution of the complete wafer implementation process. These rules allow for first-order scaling, though this rarely applies in practice. They can typically be expressed on a single page. However, the degradation in circuit performance and the potential increase in silicon area may make this approach unsuitable for commercial or even experimental circuits. In this text, we will use the λ rules to illustrate principles.

TABLE 3.1. Derivation of lambda-based rules from micron rules

MASK	FEATURE	DIMENSIONS	
		Micron rule	λ rule
Thinox	Minimum thinox width	4 μm	2λ
	Minimum thinox spacing	4 μm	2λ
	Minimum p-thinox to n-thinox spacing	8 μm	4λ
Polysilicon	Minimum poly width	3.75 μm	2λ
	Minimum poly spacing	3.75 μm	2λ
	Minimum gate poly width (p)	4.5 μm	3λ
	Minimum gate poly width (n)	4.0 μm	2λ
	Minimum gate poly extension	3.5 μm	2λ
6: Aluminum	Minimum Al width	4.5 μm	3λ
	Minimum Al spacing	4.5 μm	3λ

Gamma



Layer Representations in CMOS Processes

Advances in CMOS processes are generally complex and can inhibit the visualization of all mask levels used in fabrication. However, the design process can be abstracted to a manageable number of conceptual layout levels representing the physical features observed in the final silicon wafer. At a high level, all CMOS processes use

- Two Different Substrates,
- Doped Regions,
- Transistor Gate Electrodes,
- Interconnection Paths,
- Interlayer Contacts.

Made with Gamma

The layers for typical CMOS processes are represented in various figures using a

color scheme proposed by JPL,

a modified color scheme to differentiate between nMOS and CMOS structures,

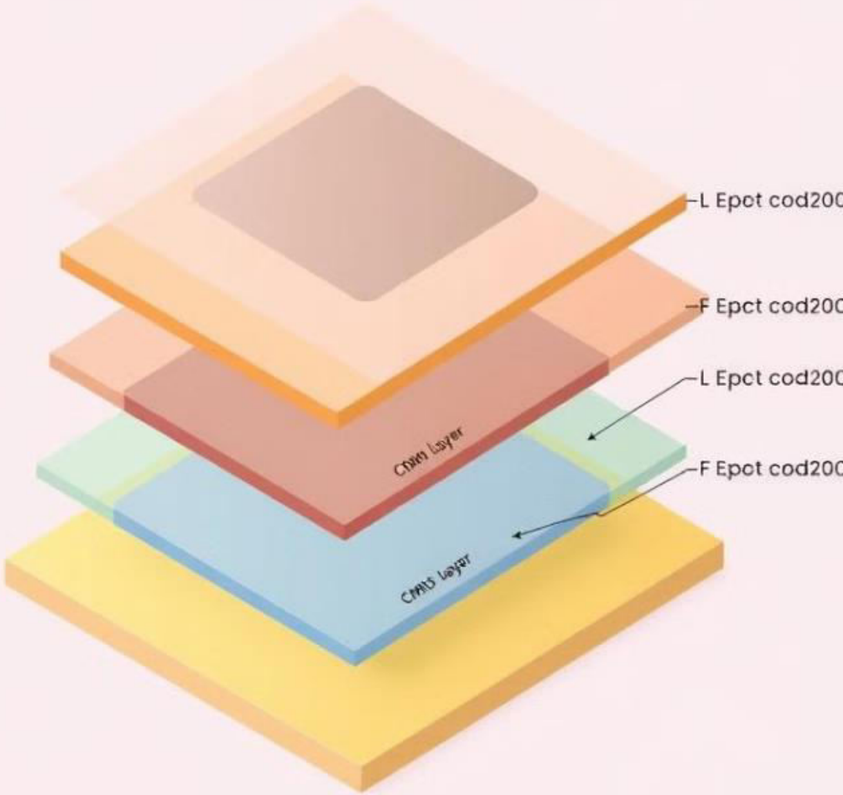
stipple patterns,

line styles,

or a mixture of these.

Diagrams will include a legend to indicate layer assignments.

At the mask level, some layers may be omitted for clarity, while at the symbolic level, only n- and p-transistors will be shown.



CIF Layer Names and Process Variations

For convenience, the CIF (Caltech Intermediate Form) layer names as used by JPL (Jet Propulsion Laboratory, California Institute of Technology, Pasadena) for bulk CMOS are also presented. CIF version 2.0 uses up to four alpha-numeric characters to describe a mask level. The first letter typically characterizes the process class (e.g., 'C' for bulk CMOS), followed by a second character to identify the layer type. Different process lines may use different combinations of masks to define the process.

PROCESS	CIF CHARACTER
n-channel MOS	N
p-channel MOS	P
bulk CMOS processes	C
silicon on insulator processes	S

Made with Gamma

Simplifying Design with Symbolic Layout

The n-well and twin-tub bulk CMOS processes, as well as the SOI process, can be represented similarly. Conceptually, the mask levels in a silicon on insulator process are the simplest. Our overall thrust is to bring the simplicity of SOI to bulk CMOS technologies by encouraging the adoption of a symbolic level of design. This allows design rules to be completely removed from layout design and paves the way for automatically performance-optimized circuits.

Made with Gamma

The rules are defined in terms of:

- Feature sizes
- Separations and overlaps.

There are a number of issues among them few are listed below

Well spacing and separation rules.

Transistor rules.

Contacts. There are several generally available contacts:

metal to p-thinox (p~diffusion]

metal to n-thinox (n-diffusion)

metal to polysilicon

VDD and V55 [Substrate contacts]

split (substrate contacts]

Poly doping

P+ and gate edges

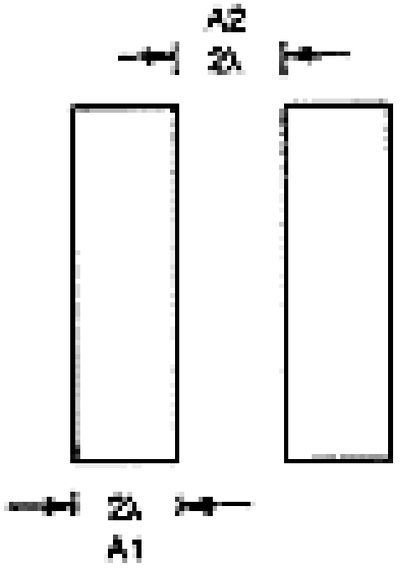
Guard rings

Made with Gamma

TABLE 3.5. Lambda-based layout rules

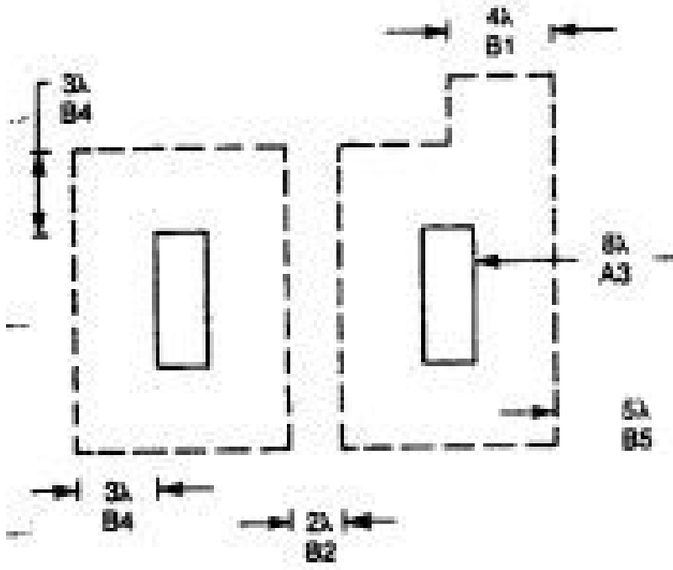
NO.	MASK	FEATURE	DIMENSION
1	Thinox	A1.Minimum thinox width	2λ
		A2.Minimum thinox spacing (n^+ to n^+, p^+ to p^-)	2λ
		A3.Minimum p-thinox to n-thinox spacing	8λ
2	p-well	B1.Minimum p-well width	4λ
		B2.Minimum p-well spacing (wells at same potential)	2λ
		B3.Minimum p-well spacing (wells at different potential)	6λ
		B4.Minimum distance to internal thinox	3λ
		B5.Minimum distance to external thinox	5λ
3	Poly	C1.Minimum poly width	2λ
		C2.Minimum poly spacing	2λ
		C3.Minimum poly to thinox spacing	λ
		C4.Minimum poly gate extension	2λ
		C5.Minimum thinox source/drain extension	2λ
4	p-plus	D1.Minimum overlap of thinox	1.5–2λ
		D2.Minimum p-plus spacing	2λ
		D3.Minimum gate overlap or distance to gate edge	1.5–2λ
		D4.Minimum spacing to unrelated thinox	1.5–2λ
5	Contact	E1.Minimum contact area	2λ × 2λ
		E2.Minimum contact to contact spacing	2λ
		E3.Minimum overlap of thinox or poly over contact	λ
		E4.Minimum spacing to gate poly	2λ
		E5.n^- source/drain contact	See Figure 3.22
		E6.p^- source/drain contact	
		E7. V_{SS} contact	
		E8. V_{DD} contact	
		E9.Split contact V_{SS}	
		E10.Split contact V_{DD}	
6	Metal	F1.Minimum metal width	2–3λ
		F2.Minimum metal spacing	3λ
		F3.Minimum metal overlap of contact	λ

Made with Gamma



MASK 1: THINOX

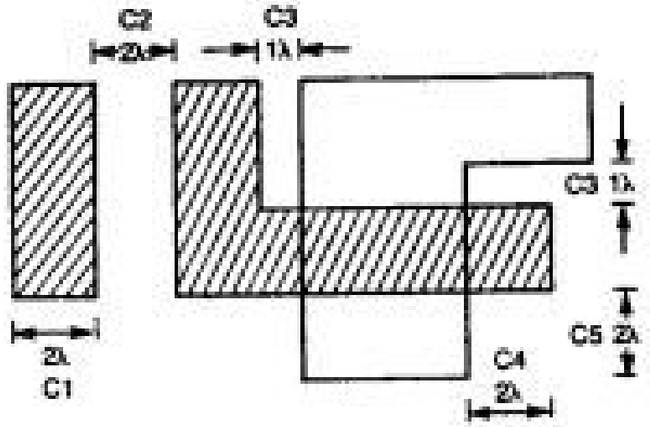
- A1. MINIMUM THINOX WIDTH 2λ
- A2. THINOX SPACING 2λ
  (n^+ to n^+ or p^+ to p^+)
- A3. p^+ to n^+ SPACING 3λ



(6λ FOR WELLS AT DIFFERENT POTENTIALS)

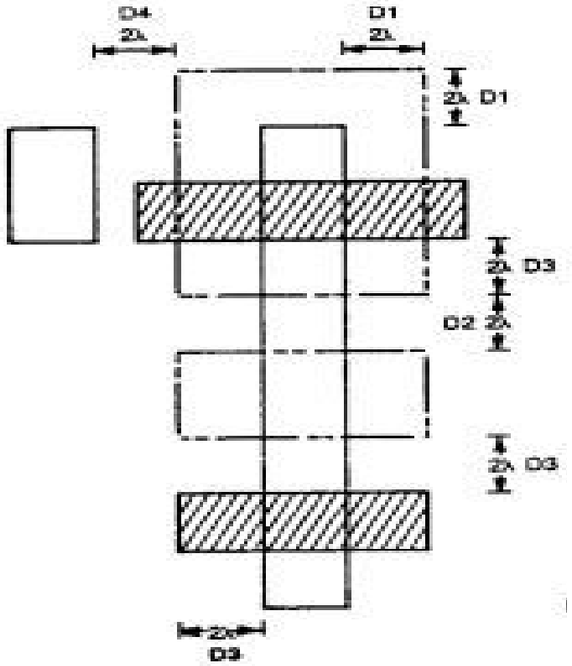
MASK 2: p-WELL

- B1. MINIMUM p-WELL WIDTH 4λ
- B2. MINIMUM p-WELL SPACING 2λ (SAME POTENTIAL)
- B3. MINIMUM p-WELL SPACING 6λ (DIFFERENT POTENTIAL)
- B4. MINIMUM OVERLAP OF INTERNAL THINOX 3λ
- B5. MINIMUM SPACING TO EXTERNAL THINOX 5λ



MASK 3: POLYSILICON

- C1. MINIMUM POLY WIDTH 2λ
- C2. MINIMUM POLY SPACING 2λ
- C3. MINIMUM POLY-THINOX SPACING λ
- C4. MINIMUM POLY GATE EXTENSION 2λ
- C5. MINIMUM THINOX SOURCE/DRAIN EXTENSION 2λ

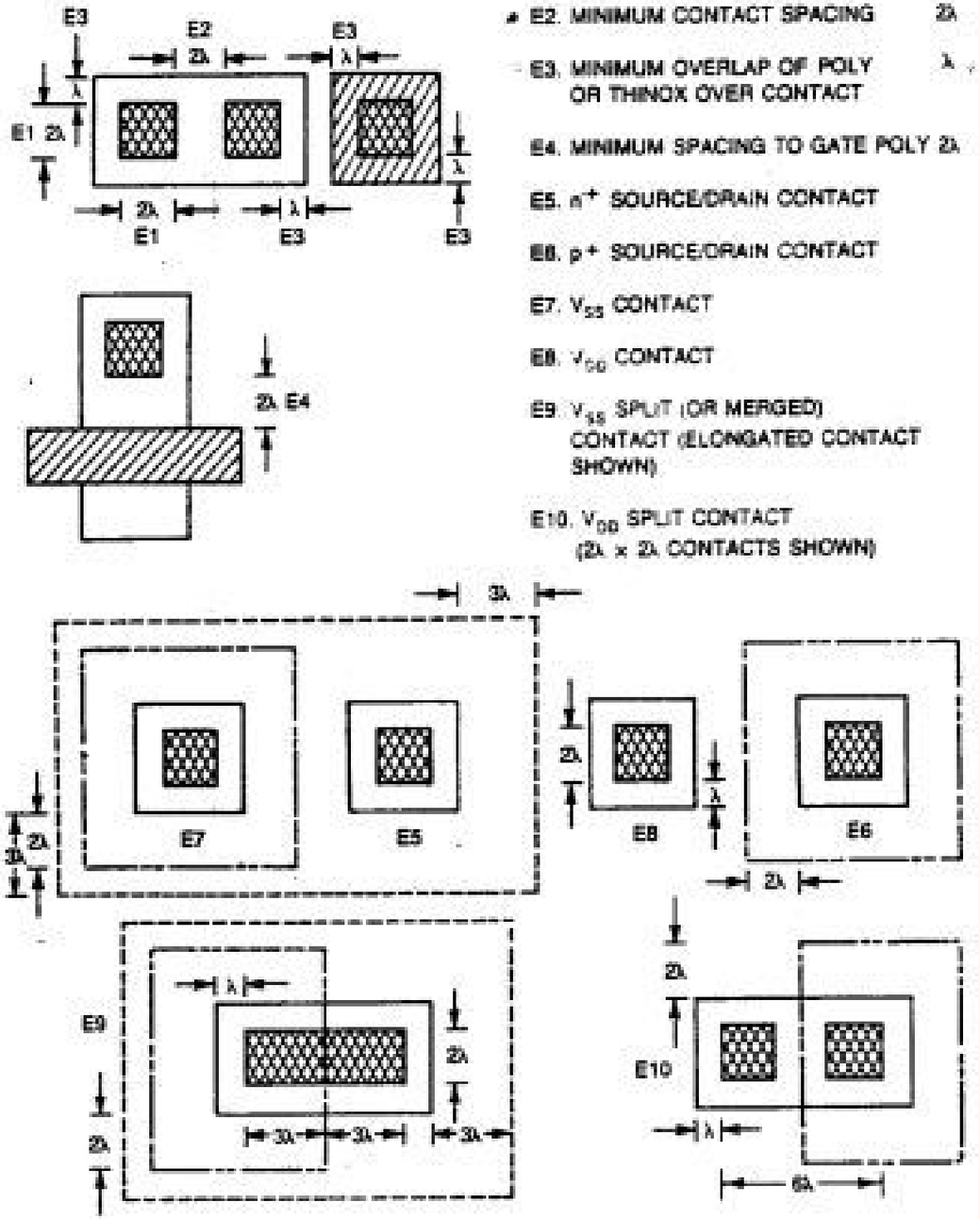


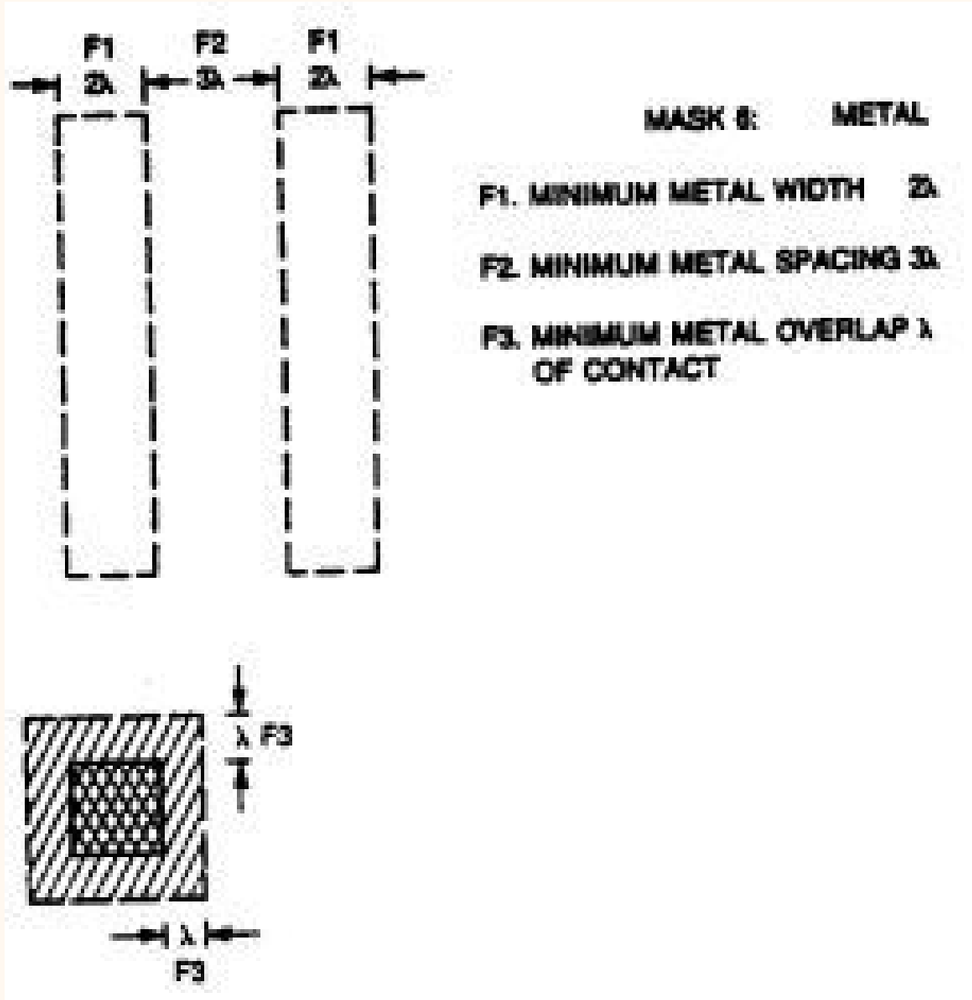
MASK 4: p-PLUS

- D1. MINIMUM OVERLAP OF THINOX 2λ
- D2. MINIMUM p-PLUS SPACING 2λ
- D3. MINIMUM GATE OVERLAP OR DISTANCE TO GATE EDGE 2λ
- D4. MINIMUM SPACING TO UNRELATED THINOX 2λ

MASK 5: CONTACT

- E1. MINIMUM CONTACT AREA 2λ × 2λ
- E2. MINIMUM CONTACT SPACING 2λ
- E3. MINIMUM OVERLAP OF POLY OR THINOX OVER CONTACT λ
- E4. MINIMUM SPACING TO GATE POLY 2λ
- E5. n^+ SOURCE/DRAIN CONTACT
- E6. p^+ SOURCE/DRAIN CONTACT
- E7. V_{SS} CONTACT
- E8. V_{DD} CONTACT
- E9. V_{SS} SPLIT (OR MERGED) CONTACT (ELONGATED CONTACT SHOWN)
- E10. V_{DD} SPLIT CONTACT (2λ × 2λ CONTACTS SHOWN)





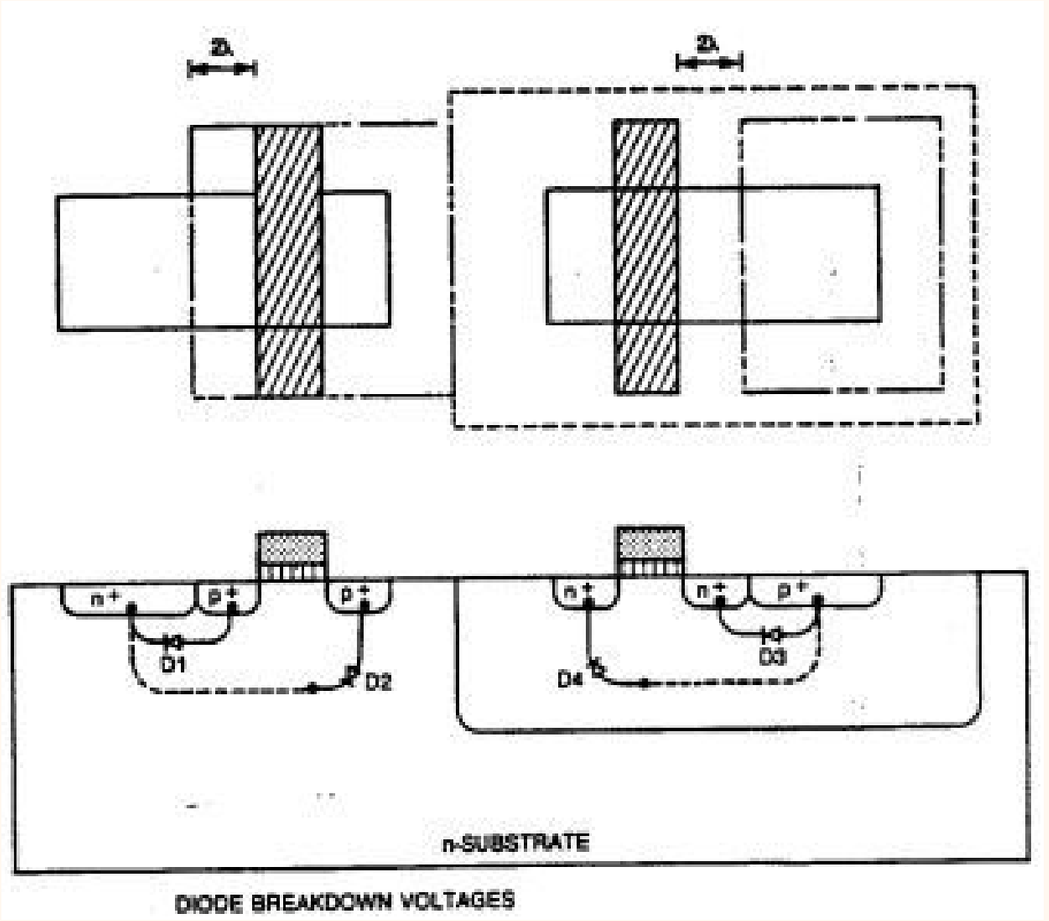
MASK 6: METAL

F1. MINIMUM METAL WIDTH 2λ

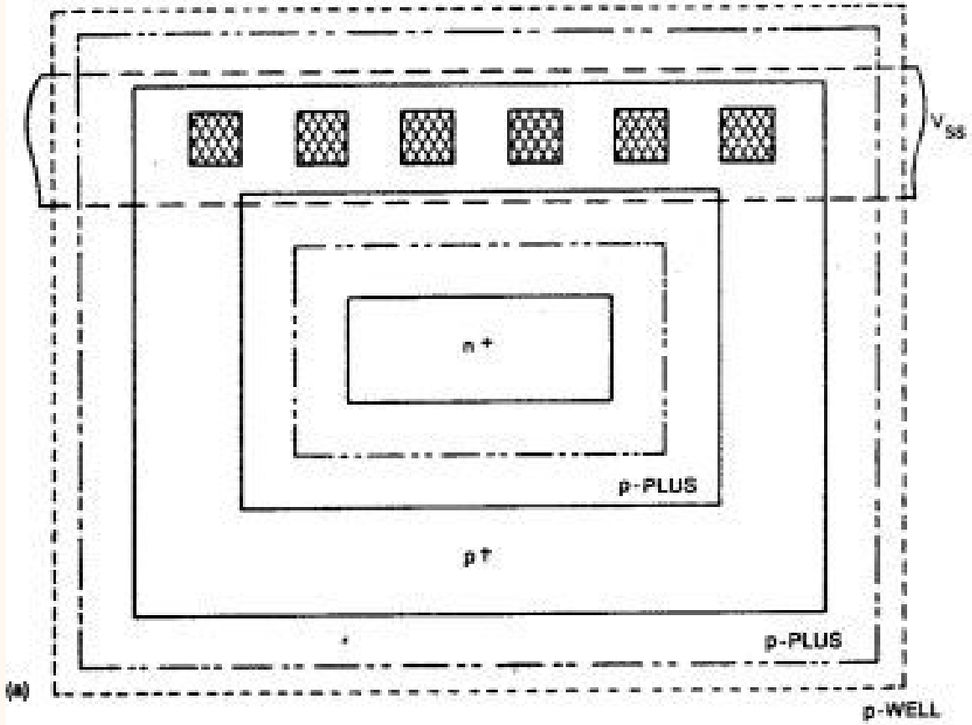
F2. MINIMUM METAL SPACING 3λ

F3. MINIMUM METAL OVERLAP λ OF CONTACT

Made with Gamma



DIODE BREAKDOWN VOLTAGES



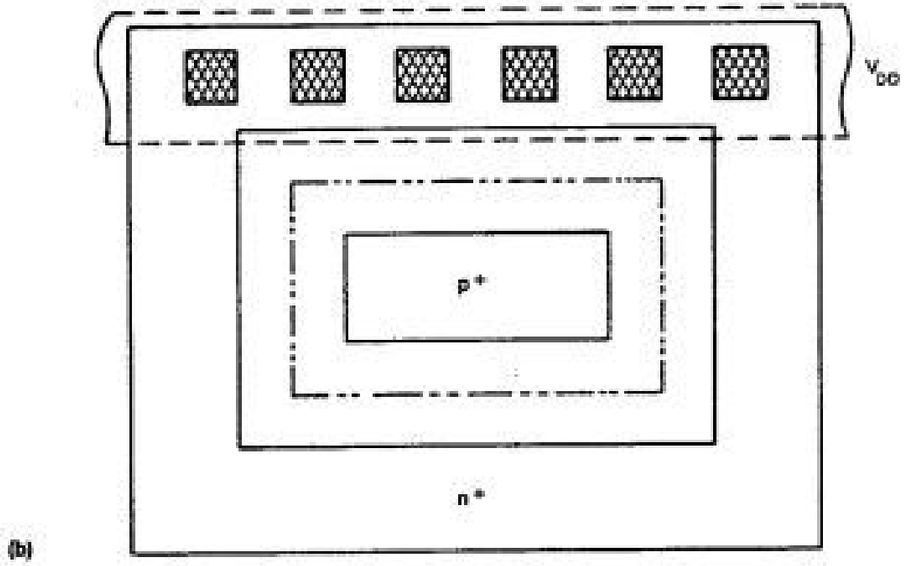


FIGURE 3.25. Realization of n^+ guard ring and p^+ guard ring

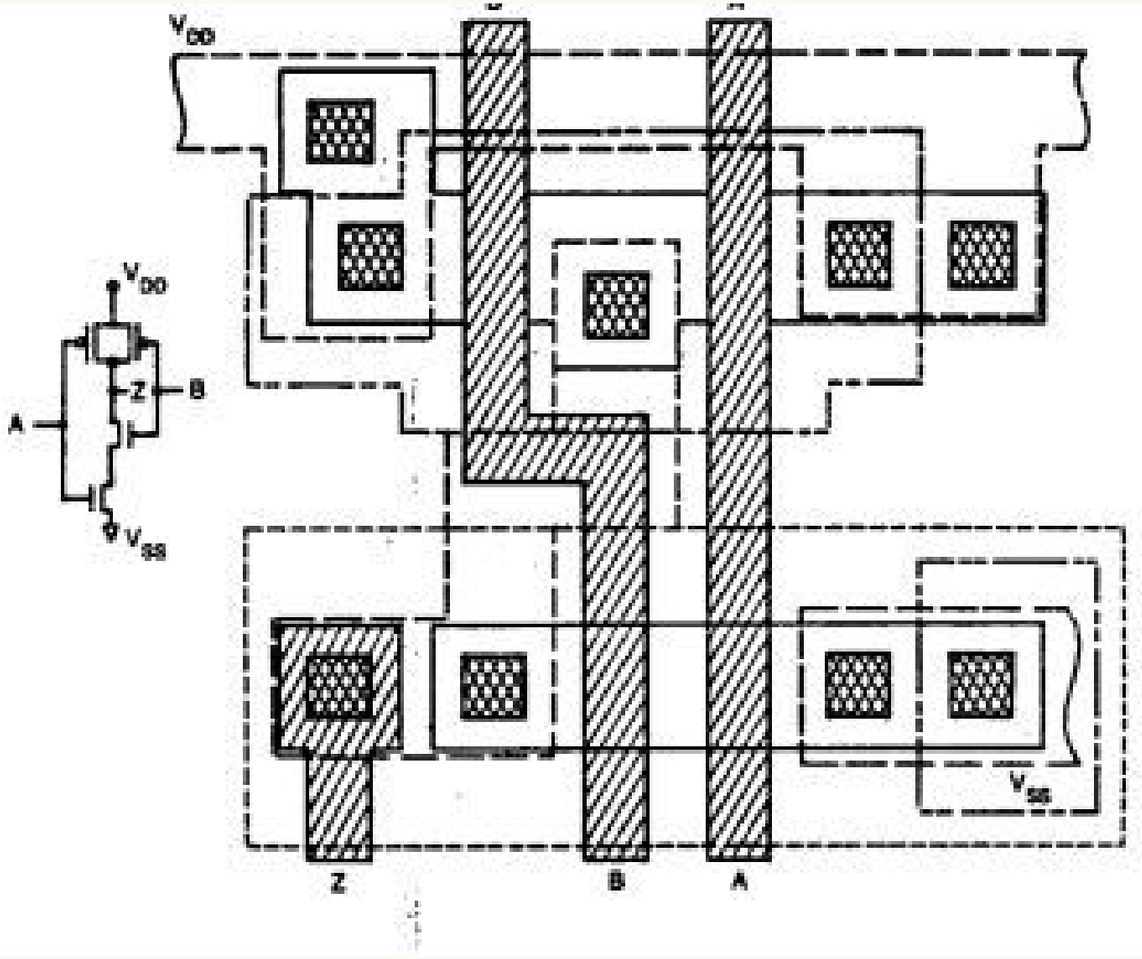


FIGURE 3.26. 2-input NAND gate layout using lambda rules for p-well CMOS

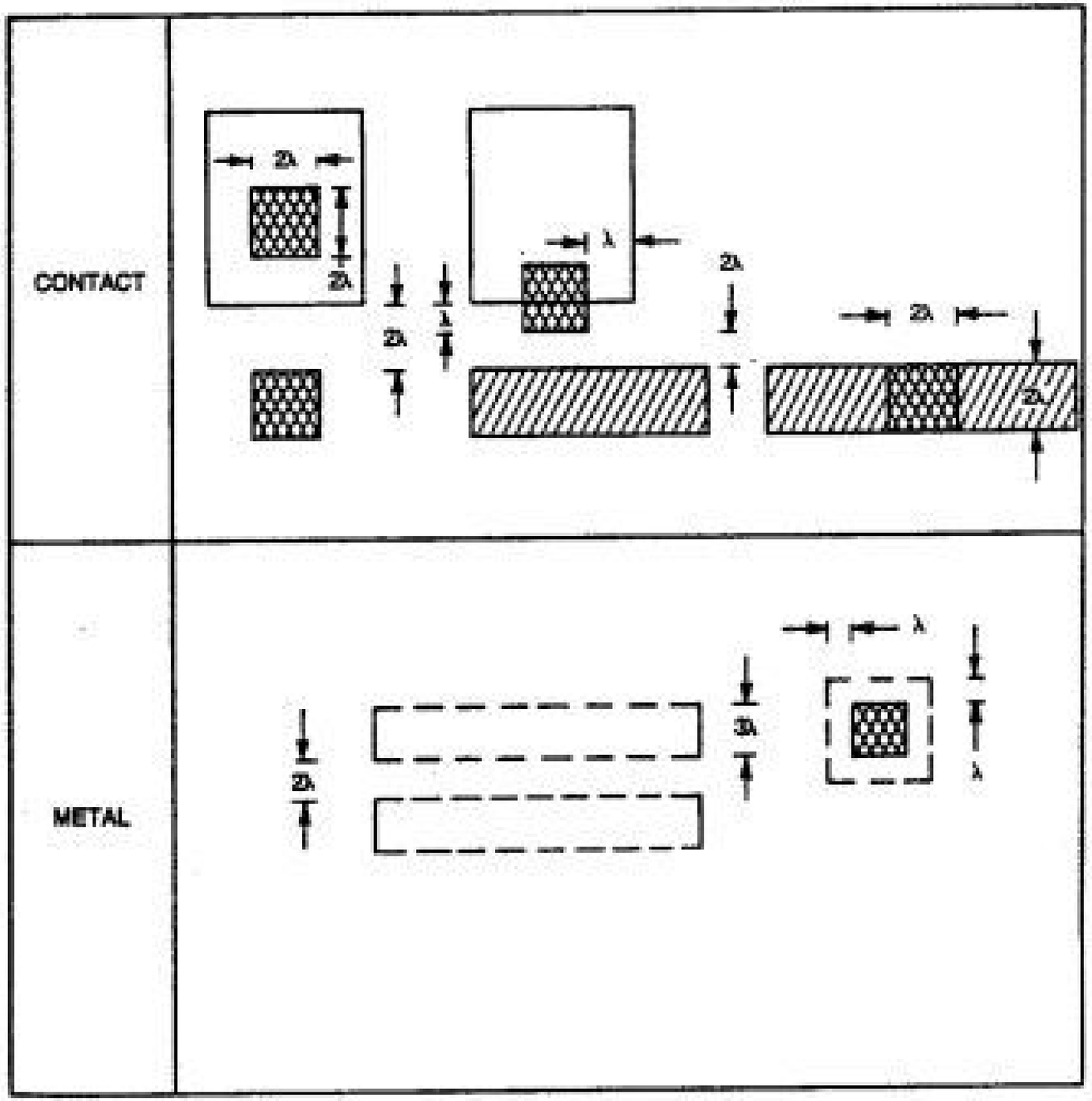
Made with Gamma

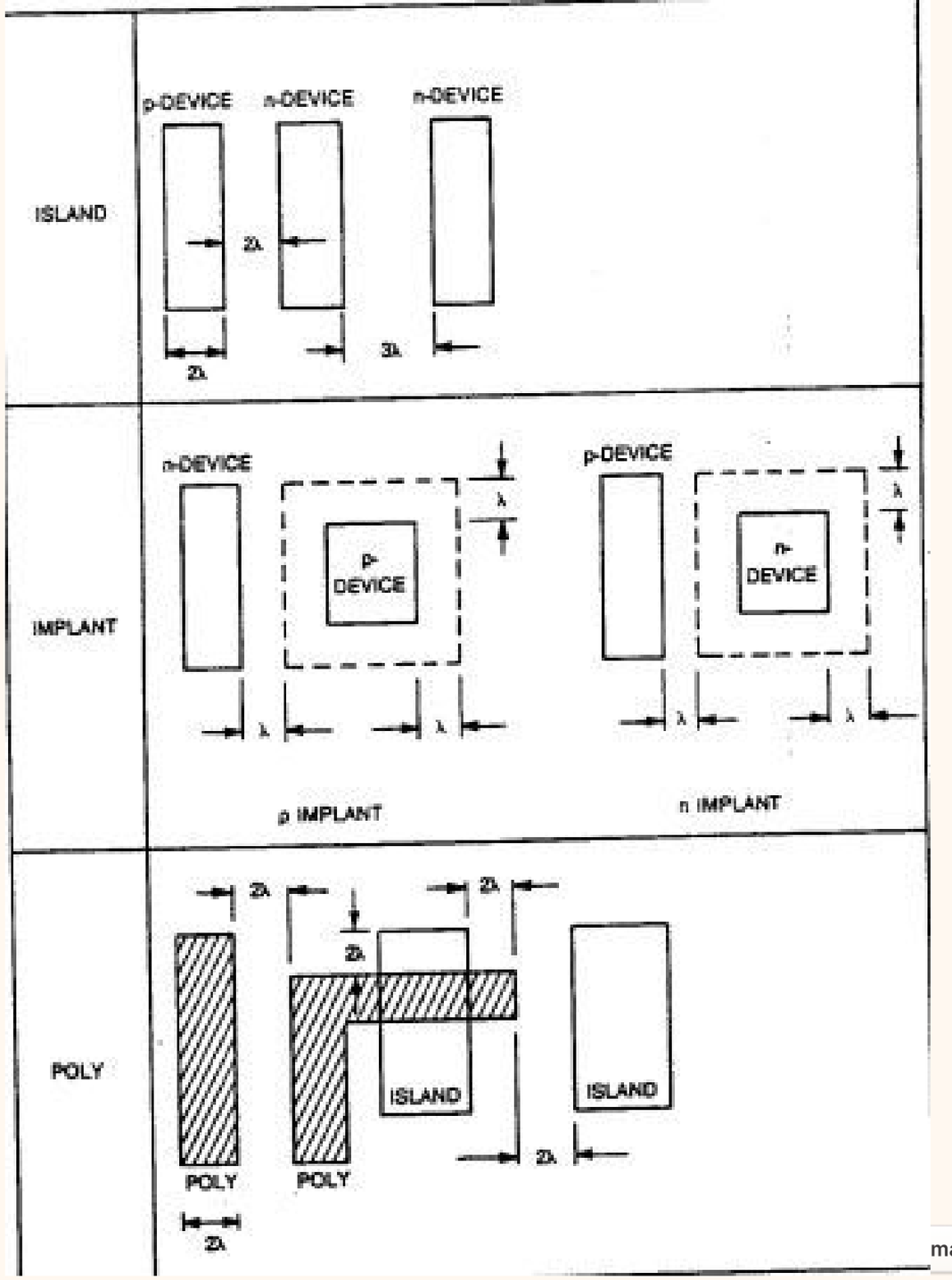
Lambda-based SOI rules

TABLE 3.6. Lambda-based layout rules for SOI

NO.	MASK	FEATURE	DIMENSION
1	Island	Minimum island width	2λ
		p-device to n-device spacing	2λ
		n-device to n-device spacing	3λ
		p-device to p-device spacing	3λ
2	Implant	Implant/island overlap	λ
		Implant/island spacing	λ
3	Poly	Minimum poly width	2λ
		Minimum poly-poly spacing	2λ
		Minimum poly to island	2λ
		Minimum island edge to poly spacing	2λ
		Minimum poly extension over island	2λ
4	Contact	Distance over poly edge	λ
		Distance over island edge	λ
		Distance from island edge	λ
		Distance from noncontacted feature	2λ
		Contact width on island	2λ
		Contact width on poly	2λ
5	Metal	Minimum metal width	3λ
		Minimum metal spacing	2λ
		Minimum metal overlap of contact	λ

Made with Gamma


CONTACT
METAL
2λ
λ
3λ


ISLAND
p-DEVICE
n-DEVICE
n-DEVICE
2λ
3λ
IMPLANT
n-DEVICE
p-DEVICE
p+ DEVICE
n+ DEVICE
λ
p IMPLANT
n IMPLANT
POLY
ISLAND
ISLAND
POLY
POLY

Made with Gamma

Chapter 4. Circuit Characterization and Performance Estimation

- 4.1 introduction.
  - Each layer of transistor-forming material has both a resistance and a capacitance that are fundamental components in estimating the performance of a circuit or a system. (It also has inductance but insignificant for most on-chip circuit.)
  - The issues considered in this chapter are
    - Resistance , capacitance, and inductance calculations.
    - Delay estimations
    - Determination of conductor size for power and clock distribution.
    - Power consumption
    - Charge sharing mechanism
    - Design margining
    - Reliability
    - Effects of scaling

- 4.2 Resistance Estimation
  - The resistance of a uniform conducting slab is $R = \left(\frac{\rho}{t}\right)\left(\frac{l}{w}\right) = R_s\left(\frac{l}{w}\right)$

    where ρ = resistivity.

    t = thickness.

    l = conductor length.

    w = conductor width.

    R_s = Sheet resistance having units of Ω/*square*.

  - According to the above formula, the two metal slabs shown in Figure 4.1 have the same resistance.

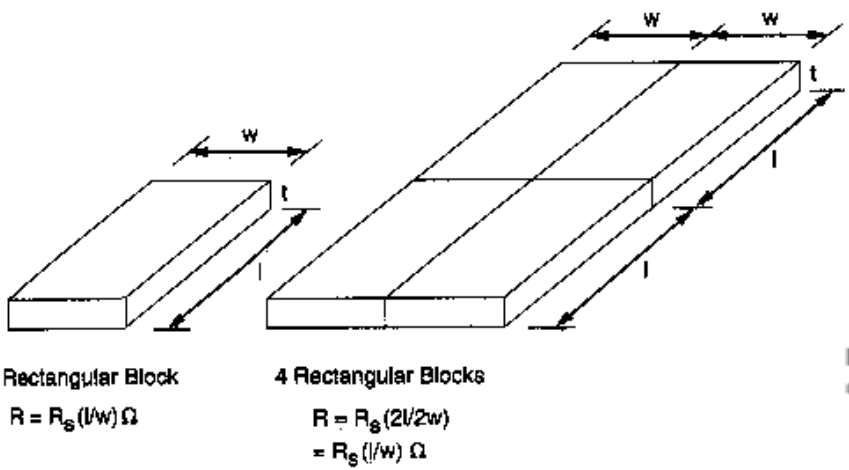


FIGURE 4.1 Determination of layer resistance

– Table 4.1 shows typical sheet resistances for *0.5@m* to *1.0@m* MOS process

TABLE 4.1 Typical Sheet Resistances for Conductors

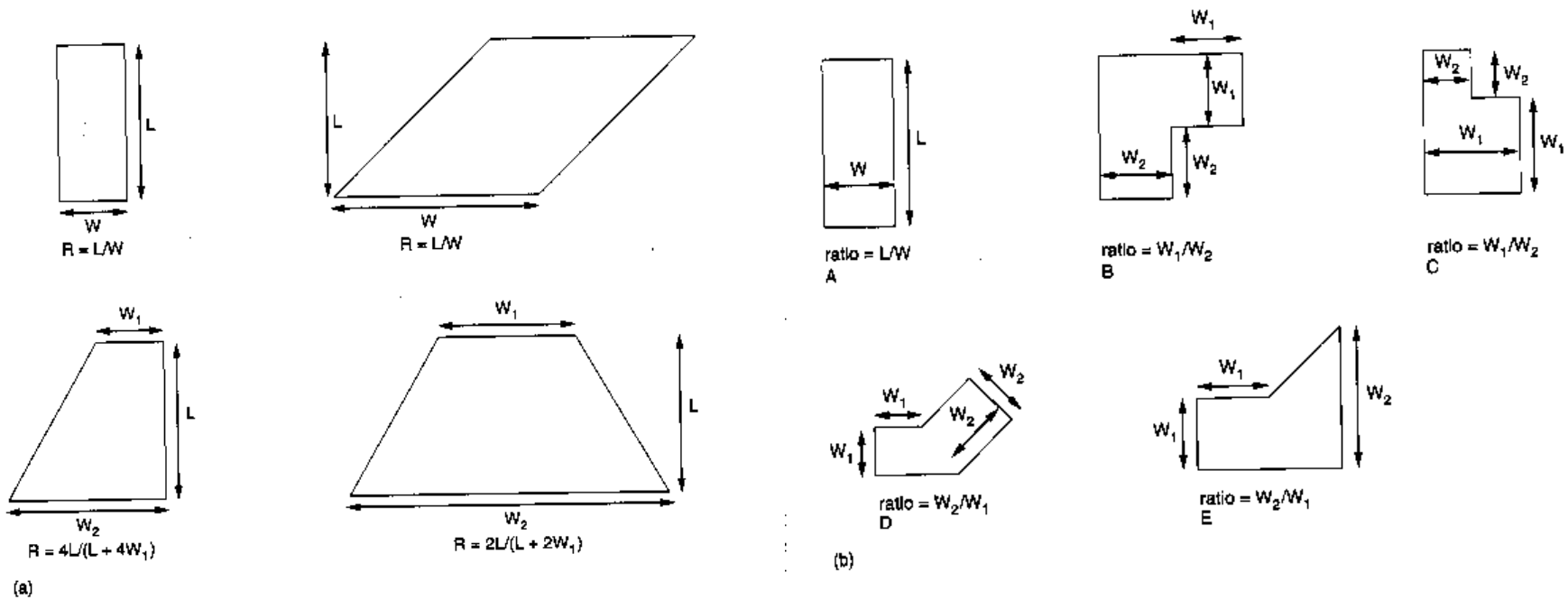
	SHEET RESISTANCE Ω/SQ		
Material	Min	Typical	Max.
Intermetal (metal1-metal2)	0.05	0.07	0.1
Top-metal (metal3)	0.03	0.04	0.05
Polysilicon	15	20	30
Silicide	2	3	6
Diffusion (n^+, p^+)	10	25	100
Silicided diffusion	2	4	10
n-well	1K	2K	5K

– Note that for metal having a given thickness, the resistivity is known, while for ploy and diffusion the resistivities are significantly influenced by the concentration density of the imparities.

– From the voltage-current characteristic of an MOS transistor, the channel resistance in the linear region can be approximated as $R_c = k(\frac{L}{W})$, where $k = \frac{1}{\mu C_{ox}(V_{gs} - V_t)}$ derived from equation (2.14), *1000* $< k <$ *30000 Ω/square* for *n*- and *p*-channel devices.

– Since the mobility of the majority carriers decreases with the increase of temperature, the channel resistance is increased by approximately *0.25% per* ^oC for temperature above *25* ^oC.

- 4.21 Resistance of Nonrectangular Regions
  - Figure 4.2(a) summarizes the resistance of a number of commonly encountered shapes.
  - Figure 4.2(b) shows some shapes that are commonly encountered in practice.
  - Table 4.2 represents the results of a study to calculate the resistance of the shapes shown in Figure 4.2(b) for different dimension ratio.



R is the resistance between the bold lines

FIGURE 4.2 Resistance of nonrectangular shapes
© IEEE 1983

TABLE 4.2 Resistance of Non-Rectangular Shapes

SHAPE	RATIO	RESISTANCE
A	1	1
A	5	5
B	1	2.5
B	1.5	2.55
B	2	2.6
B	3	2.75
C	1.5	2.1
C	2	2.25
C	3	2.5
C	4	2.65
D	1	2.2
D	1.5	2.3
D	2	2.3
D	3	2.6
E	1.5	1.45
E	2	1.8
E	3	2.3
E	4	2.65

- 4.2.2 Contact and Via Resistance
  - Typical values for processes currently in use ($0.6\mu m$) range form $0.25\ \Omega$ to a few tens of Ωs.

- 4.3 Capacitance Estimation
  - The total load capacitance on the output of a CMOS gate is the sum of
    - gate capacitance (of other inputs connected to the output of the logic gate)
    - diffusion capactance (of the drain regions connected to the output)
    - routing capacitance (of connections between the output and other inputs)

- 4.3.1 MOS - Capacitor Characteristics
  - The capacitance-voltage characteristics of an MOS capacitor (i.e. an MOS transistor without source or drain) depend on the state of the semiconductor surface.
  - The surface for a p-substrate structure shown in Figure 4.3 may be
    - *accumulation* when $V_g < 0$ (see Figure 4.3 (a)), and the gate capacitance may be approximaterd by

      $C_0 = \frac{\varepsilon_{sio_2}\varepsilon_o}{t_{ox}} A$, where

      A = area of gate.

      ε_{SiO2} = dielectric constant (or relative permittivity of SiO_2, taken as 3.9)

      ε_o = permittivity of free space.
    - *depletion* when $V_g > 0$ (See Figure 4.3 (b)), and the gate capacitance

      $C_{gb} = \frac{C_0 C_{dep}}{C_0 + C_{dep}} < C_o$, where $C_{dep} = (\frac{\varepsilon_o \varepsilon_{Si}}{d})A$ is the depletion capacitance, decreased when d increases.

      d = depletion layer depth, increased when gate to substrate voltage increases.

      ε_{Si} = dielectric constant of silicon, taken as *12*.
    - *inversion* as V_g is further increased (see Figure 4.3 (c)). In this case, minority carries (electrons for the *p*-substrate) are attracted forward the silicon surface and thus created an *n*-channel. Surface inversion yields a relative high conductivity layer under the gate. Thus for low frequency operation ($< 100Hz$), $C_{gb}=C_o$; For high frequency operation, the surface change is not able to track fast moving gate voltages

      $$C_{gb} = \frac{C_o C_{dep}}{C_0 + C_{dep}} < C_o$$

- As it can be seen, if V_g increases, the total capacitance decreases.

– Figure 4.3 (a) plots the dynamic gate capacitance as a function of gate voltage.

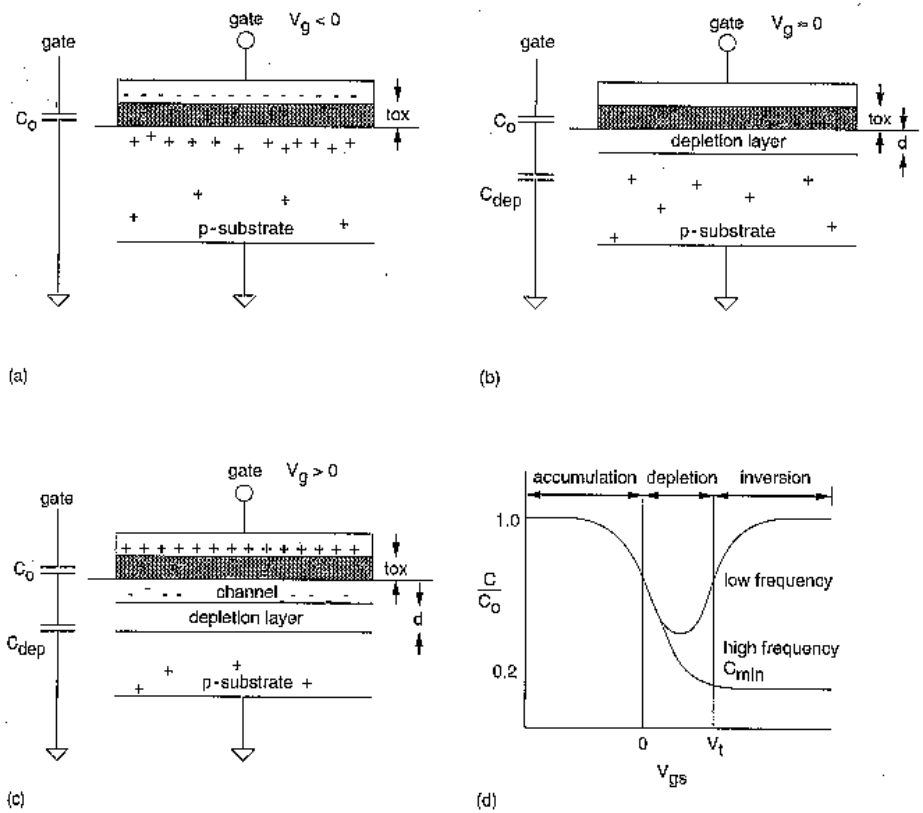


FIGURE 4.3 MOS capacitance (a) accumulation, (b) depletion, (c) inversion, (d) variation as a function of V_{gs}

- 4.3.2 MOS Device capacitances.
  - Figure 4.4 is a diagramatic representation of the parasistic capacitance fo an MOS transistor. It is assumed that the overlap of the gate over source/drain is equal to zero.
  - The following capacitance components exist:
    - C_{gs}, C_{gd} = gate to channel capacitances, lumped at the source and the drain regions of the channel, respectively.
    - C_{sb}, C_{db} = source and drain-diffusion capacitance to bulk (or substrate).
    - C_{gb} = gate to bulk capacitance.

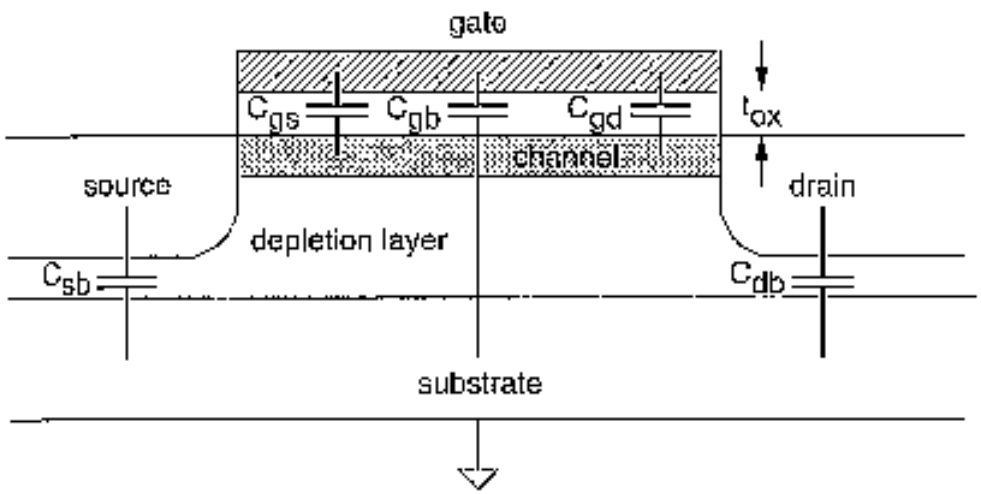


FIGURE 4.4 Process cross section showing parasitic capacitance for an MOS transistor

– Figure 4.5 shows a circuit model comprising parasitic copacitances and the transistor. The total gate capacitance is given by $C_g = C_{gb} + C_{gs} + C_{gd}$.

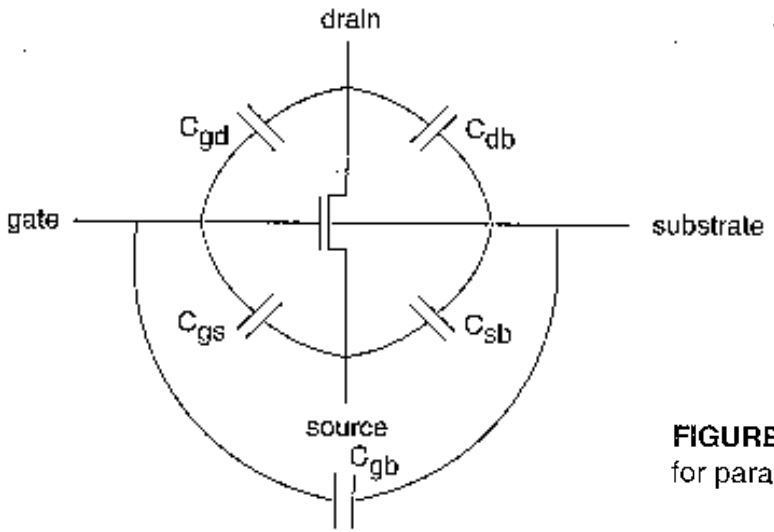


FIGURE 4.5 Circuit symbols for parasitic capacitance

– The behavior of gate capacitance of an MOS device:

1. *Off region*, where $V_{gs} < V_t$, no channel exists, hence $C_{gs} = C_{gd} = 0$. So $C_g = C_{gb}$.

2. *Non-saturated region*, where $V_{gs} - V_t > V_{ds}$. A channel exists. Thus

$$C_{gd} = C_{gs} = \frac{1}{2}\frac{\varepsilon_0 \varepsilon_{SiO2}}{t_{ox}} A \quad \text{and } C_{gb} = 0.$$

3. *Saturated region*, where $V_{gs} - V_t < V_{ds}$. The channel is pinched off at the drain end. Thus

$$C_{gd} = 0 \text{ and } \quad C_{gs} = \frac{2}{3}\frac{\varepsilon_0 \varepsilon_{SiO2}}{t_{ox}} A \ .$$

– The behavior of the gate capacitance is shown in Table 4.3 where $\varepsilon = \varepsilon_O \varepsilon_{SiO2}$

TABLE 4.3 Approximation of intrinsic MOS gate capacitance

	CAPACITANCE		
Parameter	Off	Non-saturated	Saturated
C_{gb}	$\frac{\varepsilon A}{t_{ox}}$	0	0
C_{gs}	0	$\frac{\varepsilon A}{2t_{ox}}$	$\frac{2\varepsilon A}{3t_{ox}}$
C_{gd}	0	$\frac{\varepsilon A}{2t_{ox}}$	0 (finite for short channel devices)
$C_g = C_{gb} + C_{gs} + C_{gd}$	$\frac{\varepsilon A}{t_{ox}}$	$\frac{\varepsilon A}{t_{ox}}$	$\frac{2\varepsilon A}{3t_{ox}} \rightarrow \frac{.9\varepsilon A}{t_{ox}}$ (short channel)

– Figure 4.6 (a) shows the C_{gs} and C_{gd} of a long channel n-transistor ($W = 49.2\ \mu m$, $L=4.5\ \mu m$).

– Figure 4.6 (b) shows the C_{gs} and C_{gd} of a short channel device ($L=0.75\ \mu m$). Note that C_{gd} is finite, i.e., $C_{gd} > 0$. This is due to channel side fringing fields between the gate and drain.

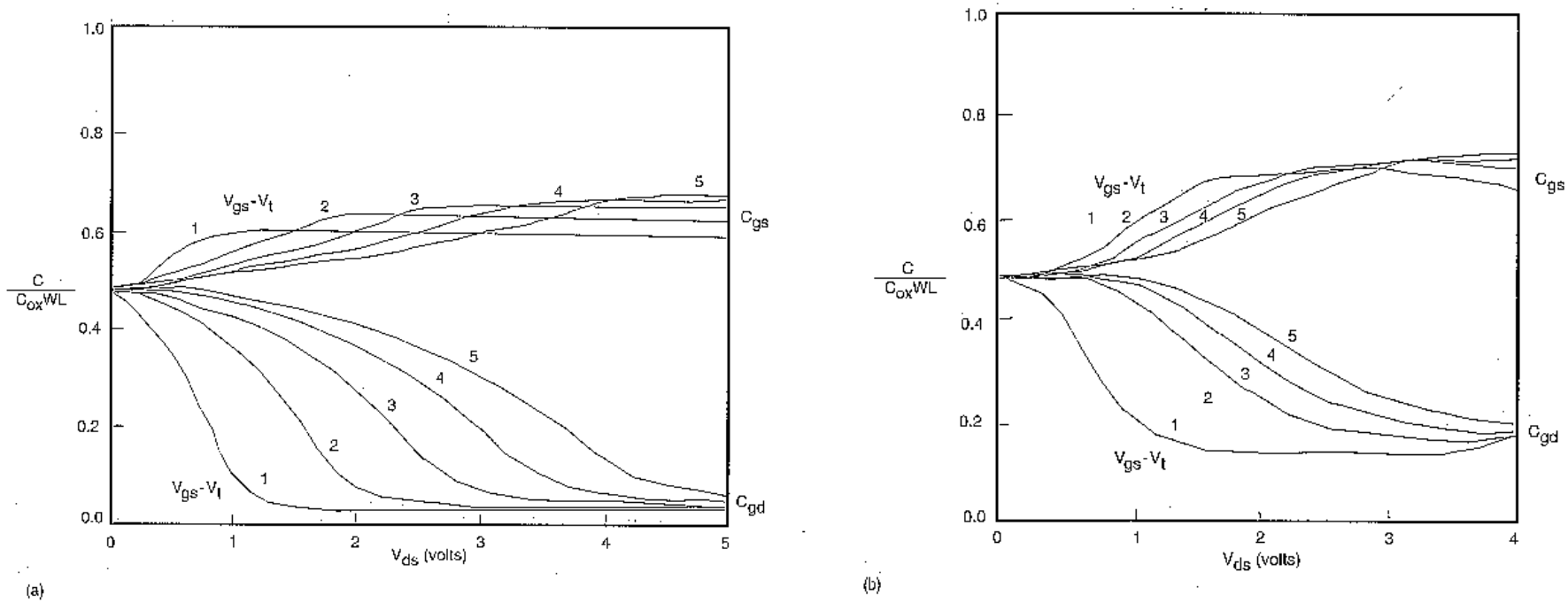


FIGURE 4.6 Total gate capacitance of an MOS transistor as a function of V_{gs} (© IEEE 1987)

– For the purpose of delay calculation, the gate capacitance can be approximated by

$$C_g = C_{ox}A = \frac{\varepsilon_o \varepsilon_{SiO2}}{t_{ox}} A, \text{ where } C_{ox} = \frac{\varepsilon_o \varepsilon_{SiO2}}{t_{ox}}$$ is the "thin-oxide" capacitance per unit area.

– With a thin-oxide thickness (t_{ox}) in the order of $100 \rightarrow 200$Å, $C_{ox} = \frac{3.9*8.854*10^{-14}}{(100 \rightarrow 200)*10^{-8}}$

$= 35 \rightarrow 17*10^{-4}\ pF/\mu m^2$.

– The gate capacitance for the case shown in Figure 4.7 for $\lambda = 0.5\ \mu m$, $W = 2\ \mu m$ and $L = 1\ \mu m$ and t_{ox}=150 Å, is $C_{g(intrinsic)} = 2*25.5*10^{-4}pF = 0.005\ pF$.

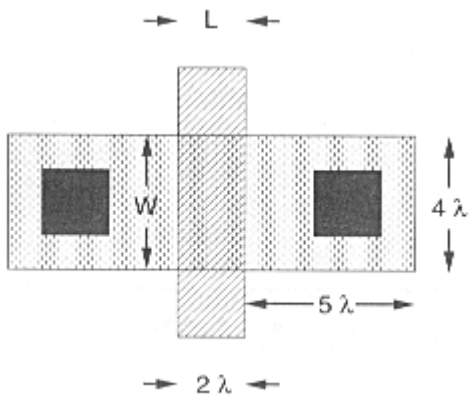


FIGURE 4.7 Physical Layout of a unit MOS Transistor for Capacitance Estimation

- 4.3.3 Diffusion (source/drain) capacitance.
  - Figure 4.8 shows a model for source/drain capacitance. Total drain capacitance C_d is given by $C_d = C_{ja} * (ab) + C_{jp} * (2a + 2b)$,
  - where C_{ja} = junction capacitance per μm^2,

    C_{jp} = periphery capacitance per μm.

    a = width of diffusion region (μm).

    b = length of diffusion region (μm).

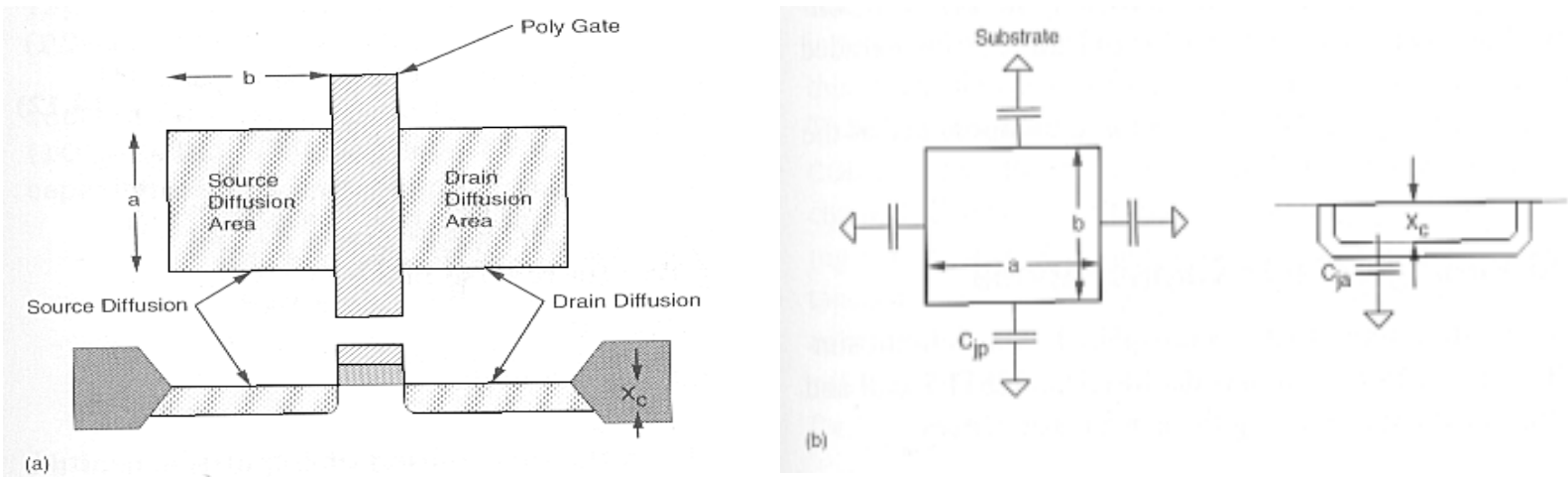


FIGURE 4.8 Area and peripheral components of diffusion capacitance

– Table 4.4 shows typical values of diffusion capacitances for both *n*- and *p*-channel devices.

TABLE 4.4 Typical Diffusion Capacitance Values (1μ n-well Process)

	n-DEVICE (OR WIRE)	p-DEVICE (OR WIRE)
C_{ja}	$3 \times 10^{-4} pF/\mu m^2$	$5 \times 10^{-4} pF/\mu m^2$
C_{jp}	$4 \times 10^{-4} pF/\mu m$	$4 \times 10^{-4} pF/\mu m$

– Note that the above simple capacitance calculations assume zero *DC* bias across the junction. However, both C_{ja} and C_{jb} are functions of junction voltage V_j due to dependence of depletion layer thickness on V_j. Thus the junction capacitance is $C_j = C_{jo}(1 - \frac{V_j}{V_b})^{-m}$,

where V_j = junction voltage (negative for reversed-bias)

C_{jo} = zero bias capacitance (i.e., $V_j = 0$)

V_b = built-in junction potential ~ *0.6* volts.

m = a constant dependent on the distribution of impurities. Effective value ranges from *0.3* ~ *0.5*.

- 4.3.4 SPICE Modeling of MOS capacitance
  - The SPICE MOSFET (and the corresponding model) call and the MOSFET MODEL statement are shown below

```
.
M1 4 3 5 0 NFET W=4U L=1U AS=15P AD=15P PS=11.5U PD=11.5U
.
.
.MODEL NFET NMOS
+ TOX=200E-8
+ CGBO=200P CGSO=600P CGDO=600P
+ CJ=200U CJSW=400P MJ=0.5 MJSW=0.3 PB=0.7
+ .....
.
.
```

- *AS* = source area,
- *AD* = drain area,
- *PS* = source periphery,
- *PD* = drain periphery.
- *.MODEL* signals the beginning of MOSFET model

- From these data, $C_{g(intrinsic)} = W * L * C_{ox} = 4*1*17*10^{-4}\ pF = 0.0068pF$. C_{ox} is derived form $TOX = 200E\text{-}8$.
- Extrinsic values of C_{gso}, C_{gdo} and C_{gbo} are added to C_{gs}, C_{gd} and C_{gb} to consider the fringing field from gate terminal . They are specified in SPICE MOSFET Model by *CGSO, CGDO* and *CGBO*.
- C_{gbo} occurs due to the polysilcon (gate) extension beyond the channel. Thus $C_{gbo} = CGSO * L * 2$.
- C_{gso} and C_{gdo} represent the gate source/drain capacitance due to overlap in the physical structure of the transistor. Thus, $C_{gso} = CGSO * W$ and $C_{gdo} = CGDO * W$.
- The extrinsic gate capacitance for a typical MOS transistor is

  $C_{g(extrinsic)} = C_{gso} + C_{gdo} + C_{gbo} = W*CGSO + W*CGDO + 2 * L * CGBO.$

  $= 4*600*10^{-12}*10^{-6}+4*600*10^{-12}*10^{-6}+2*1*200*10^{-12}*10^{-6} = 0.0052pF.$

- $C_{g(total)} = C_{g(intrinsic)} + C_{g(extrinsic)} = 0.0068 + 0.0052 = 0.012pF.$

– In SPICE the source/drain capacitance is calculated as follows:

$$C_j = (Area * CJ * (1 + \frac{VJ}{PB})^{-MJ} + (Periphery * CJSW) * (1 + \frac{VJ}{PB})^{-MJSW})$$

where CJ = the zero-bias capacitance per junction area.

$CJSW$ = the zero-bias-junction capacitance per junction periphery.

MJ = the grading coefficient of the junction bottom.

$MJSW$ = the grading coefficient of the junction sidewall.

VJ = the junction potential.

PB = the built-in voltage ($0.4\ \mu \rightarrow 0.8V$)

$Area$ = AS or AD, the area of source/drain

$Periphery$ = PS or PD, the periphery of source/drain.

– At $VJ = 2.5V$ and $PB = 0.7V$,

$$C_{j\,drain} = [15*10^{-12}*2*10^{-4}*(1+2.5/0.7)^{-0.5}] + [11.5*10^{--6}*4*10^{--10}*(1+2.5/0.7)^{-0.3}]F$$

$$= 0.0014pF + 0.0029pF = 0.0043pF = C_{j\,source}$$

- 4.3.5 Routing Capacitance
- 4.3.5.1 Single Wire Capacitance
  - Routing capacitance between metal and poly layers and the substrate is shown in Figure 4.9. It consists of three components:

    1. *A parallel-plate capacitance (C = (ε/t)*A).*

    2. *Fringing capacitance to substrate.*

    3.*Coupling capacitance between two metal line on the same layer*.

  - An empirical formula $C = \varepsilon[(\frac{w}{h}) + 0.77 + 1.06(\frac{w}{h})^{0.25} + 1.06(\frac{t}{h})^{0.5}]$

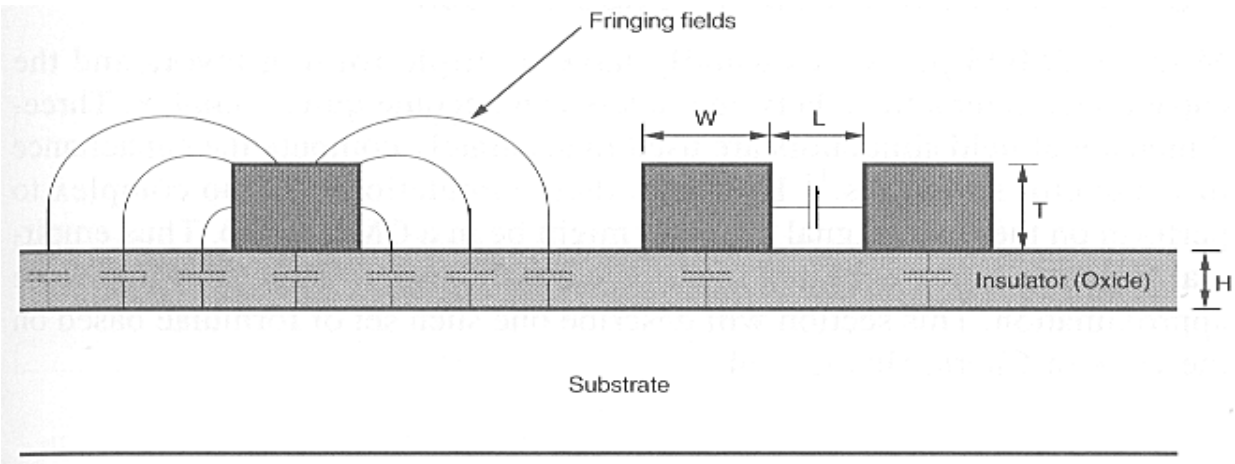


FIGURE 4.9 Effect of fringing fields on capacitance

- 4.3.5.2 Multiple Conductor Capacitance
  - Figure 4.11 shows a model for routing structure with multiple conductances.

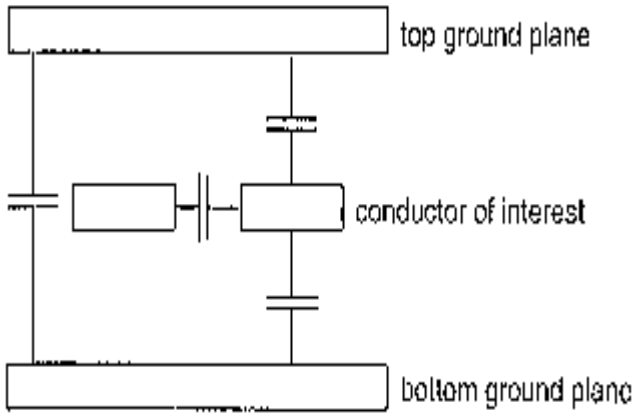


FIGURE 4.11 Multilevel-layer capacitance model

  - The capacitance of the middle layer (conductor of interest) is divided into three components.

    1. The line-to-ground capacitance.

    2. The line-to-line capacitance.

    3. The crossover capacitance.

  - Figure 4.12 shows the capacitance of middle layer 2 to ground $C_2 = C_{21} + C_{23} + C_{22}$.

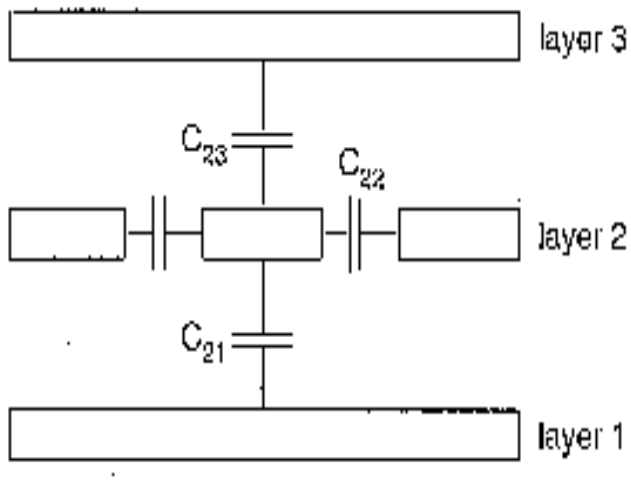
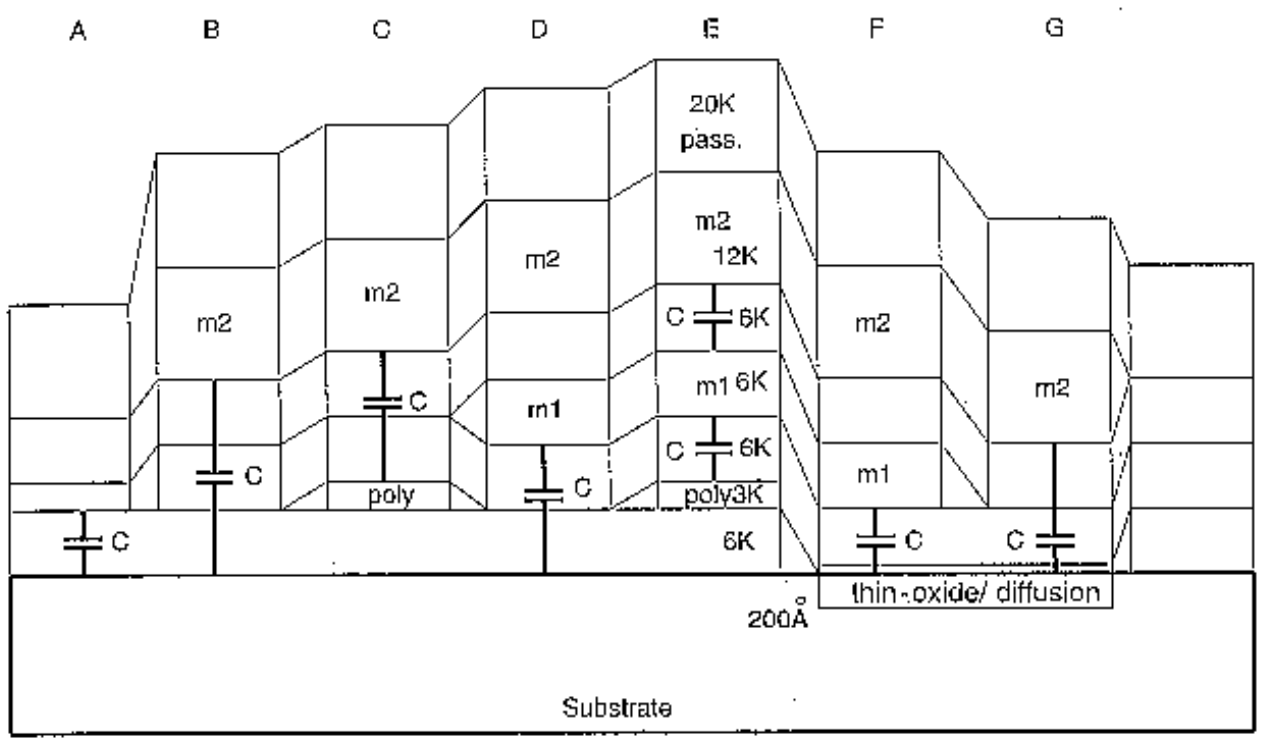


FIGURE 4.12 Specific capacitances in a three-layer-metal system

– Empirical formulas to compute these capacitance values are given in Weste text book.

– Figure 4.13 shows various capacitances for a two-level-metal process.



Metal2	12000 Å
Passivation	20000 Å

FIGURE 4.13 A process cross section showing inter-layer capacitances

- 4.3.6 Distributed RC Effects
  - The propagation of a signal along a wire depends on many factors, including the distributed resistance and capacitance of the wire, the impedence of the driving source, and the load impedence.
  - A long wire can be represented in terms of several RC sections, as shown in Figure 4.15.

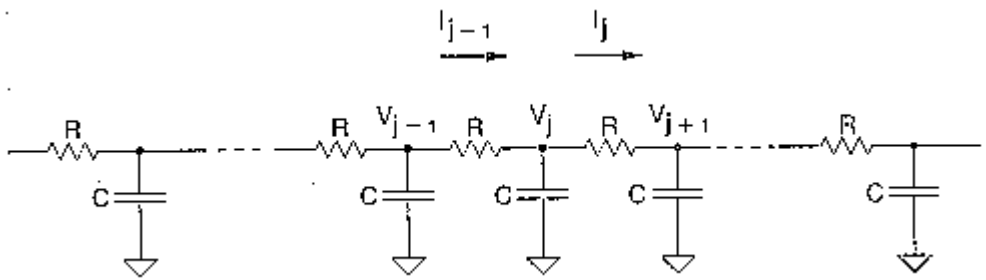


FIGURE 4.15
Representation of long wire in terms of distributed *RC* sections

  - The response a node V_j with respect to time is given by $C\frac{dV_j}{dt} = (I_{j-1} - I_j) = \frac{V_{j-1} - V_j}{R} - \frac{V_j - V_{j+1}}{R}$,

    when expressed as a differential form: $rc\frac{dV}{dt} = \frac{d^2V}{dx^2}$

    where x = distance from input

    r = resistance per unit length

    c = capacitance per unit length.

– The solution for the propagation of a voltage step along a wire of length x shows that the rise/fall delay $t_x = k\,x^2$, where k is a constatnt.

– Alternatively, a discrete analysis yields a signal delay of $t_n = 0.7 * \frac{RCn(n+1)}{2}$, where n is the number of sections. As n becomes very large, $t_l = 0.7 rcl^2 / 2$

where r = resistance per unit length.

c = capacitance per unit length.

l = length of wire.

– Figure 4.16 shows an example of using the above formula to insert a buffer on a long wire

- without a buffer, the propagation delay $t_p = 0.7 * 20 * 4 * 10^{-4} * 2000^2 / 2 = 112ns$
- with a buffer on the middle, $t_p = (2 * 0.7 * 20 * 4 * 10^{-4} * 1000^2)/2 + t_{buf} = 66ns + t_{buf}$

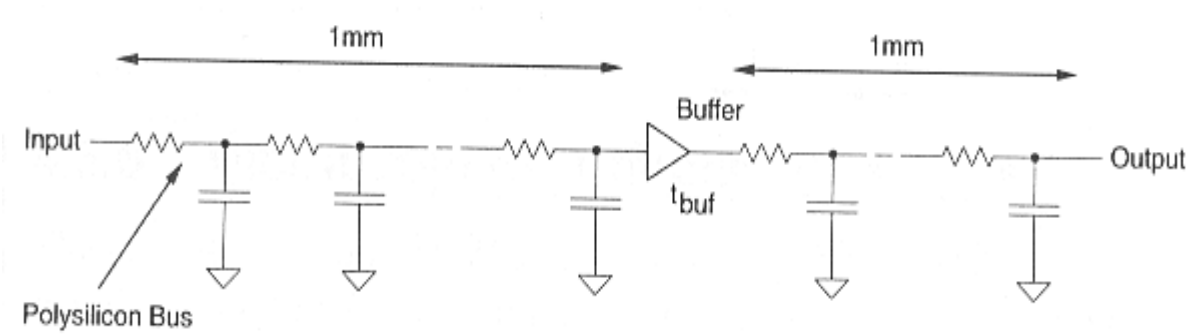


FIGURE 4.16
Segmentation of an *RC* line using a buffer

– Note that the above calculation considers only propagating a voltage step signal along the wire.

– A model for the distributed *RC* delay, which takes driver and receiver loading into account is shown in Figure 4.17, where

R_s = the output resistance of the driver.

C_i = the receiver input capacitance.

R_t = the lumped resistance of the wire.

C_t = the lumped capacitance of the wire.

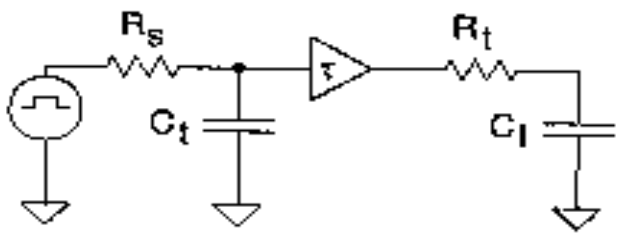


FIGURE 4.17 Simple model for RC delay calculation

- 4.3.7 Capacitance Design Guide
  - Table 4.6 gives respresentative capacitance value (no fringing) for a 1 μm n-well COMS process.

TABLE 4.6 Typical 1μm CMOS capacitances

PARAMETER	CAPACITANCE (ATTO FARADS $(10^{-18})/\mu m^2$)	COMMENTS
C_{jan}	300	n-diffusion area—varies widely with process
C_{jpn} ($aF/\mu m$)	400	n-diffusion periphery—varies widely with process
C_{jap}	500	p-diffusion area—varies widely with process
C_{jpp} ($aF/\mu m$)	400	p-diffusion periphery—varies widely with process
C_{gs}	1800	Gate capacitance—increases as t_{ox} thins
C_p	50	Poly-over field oxide
C_{m1}	30	Metal1-over field oxide
C_{m1p}	60	Metal1 to poly
C_{m1d}	60	Metal1 to diffusion
C_{m2}	20	Metal2 to substrate
C_{m2m1}	50	Metal2 to metal1
C_{m2p}	30	Metal2 to poly
C_{m2d}	30	Metal2 to diffusion
C_{m3}	10	Metal3 to substrate
C_{m3m2}	30	Metal3 to metal2
C_{m3m1}	15	Metal3 to metal1
C_{m3p}	12	Metal3 to poly
C_{m3d}	10	Metal3 to diffusion

Example

A register that fits in a data-path is 25 μm tall (the direction of repetition). A metal2 clock line runs vertically to link all registers in an n-bit register. The register has 30 μm of 1 μm metal1, 20 μm of 1 μm poly (over field), and 16 μm of 1 μm gate capacitance.

1. Calculate the per-bit clock load and the load for a 16-bit register.
2. What would be the *RC* delay to the register from a clock buffer using 5 mm of 1μ metal2 (.05 Ω/sq)?
3. How wide would the clock line have to be to keep the skew below .5 *ns* if a register file containing 32 16-bit registers was fed with the same 5 mm metal2 wire?

1. The parasitics are as follows:

$C_{m1} = 30 \times 30 = 900\ aF$

$C_p = 20 \times 50 = 1000\ aF$

$C_{gs} = 16 \times 1800 = 28{,}800\ aF$

$C_{reg1} = 900 + 1000 + 28{,}800\ aF = .030pF$

$C_{reg16} = 16 \times C_{reg1} = 0.48pF$

2. $R_{metal2} = 5000 \times .05$

$= 250$ ohms

Because the capacitance load is at the end of the wire, we can approximate the *RC* delay by adding the metal2 track capacitance to the load capacitance and performing a simple *RC* calculation.

$$\begin{aligned} C_{total} &= 0.48 + C_{metal2}\ pF \\ &= 0.48 + (5000 \times 20 \times 10^{-6})pF \\ &= 0.58pF \\ RC &= 250 \times .58 \times 10^{-12}\ S \\ &= .145\ ns \end{aligned}$$

3. We now have 32 registers, so the load capacitance of the registers is

$$\begin{aligned} C_{regfile} &= 32 \times C_{reg16} \\ &= 15.36pF. \end{aligned}$$

The *RC* for a 1 μm-wide clock feed is

$= 3.84\ ns.$

Hence the clock line has to be widened by 3.84/0.5 or 7.68. For safety one might choose a 10μ wire.

Now

$$\begin{aligned} C_{total} &= 15.36 + C_{metal2}\ pF \\ &= 15.36 + (5000 \times 10 \times 20 \times 10^{-6})pF \\ &= 16.36pF \\ RC &= 25 \times 16.36 \times 10^{-12}\ s \\ &= 0.41\ ns. \end{aligned}$$

- 4.3.8 Wire-length Design Guide
  - For sufficiently small wire length, the wire's RC delay can be ignored (i.e., the delay element τ in Figure 4.17 can be removed) and thus the wire can be modeled as a simple capacitive load.
  - To model a wire as a simple capacitive load, the wire's RC delay τ_w and gate delay τ_g must satisfy:

$$\tau_w << \tau_g \Rightarrow \frac{1}{2} rcl^2 << \tau_g \Rightarrow l << \sqrt{\frac{2\tau_g}{rc}}$$

  - For example, assume $\tau_g = 200ps$ and for a minimu-width aluminum wire

$$l << \sqrt{\frac{2*0.2*10^{-9}}{0.05/\lambda * 30*10^{-18}/\lambda}} \approx 16000\lambda$$

where $r = 0.05\ \Omega / \lambda$

$c = 30 * 10^{-18}\ F/\lambda$

λ = is the value of λ - design rule

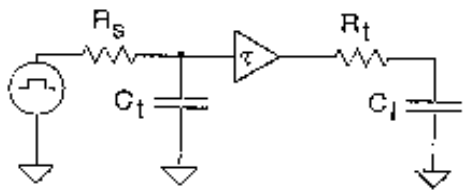


FIGURE 4.17 Simple model for RC delay calculation

– Table 4.7 shows the maximum interconect length for a typical CMOS proces in terms of λ such that a wire can be modeled as a simple capacitive load. This table assumes gate delays of the order of *100ps* to *500 ps*.

TABLE 4.7 Guidelines for Ignoring RC Wire Delays

LAYER	MAXIMUM LENGTH (λ)
Metal3	10000
Metal2	8000
Metal1	5000
Silicide	600
Polysilicon	200
Diffusion	60

- 4.4 Inductance
  - On-chip inductances are small, but bond-wire inductances are large enough to cause troubles for I/O circuits(voltage spike $dV = L\frac{dI}{dt}$).
    - For an on-chip conductor whose inductance is $L = \frac{\mu}{2\pi}\ln(\frac{8h}{w} + \frac{w}{4h})$.

      where w = the width of the conductor

      h = the height above the substrate.

    - For bond-wire inductance $L = \frac{\mu}{2\pi}\ln(\frac{4h}{d})$.

      where μ = permeability of the wire

      h = the height above the ground plane

      d = the diameter of the wire.

- 4.5 Switching Characteristics
  - Referring to Figure 4.18, some timing parameters are defined as follows:
    - *Rise time,* t_r = time for a waveform to rise from 10% to 90 % of its steady-state value.
    - *Fall time,* t_f = time for a waveform to fall from 90% to 10% of its steady-state value.
    - *Delay time,* t_d = time difference between input transition(50%) and the 50% output level.
    - *Output rising delay time,* t_{dr} = the delay time of a rising output in response to the input change.
      Output falling delay time, t_{df} = the delay time of a falling output in response to the input change.

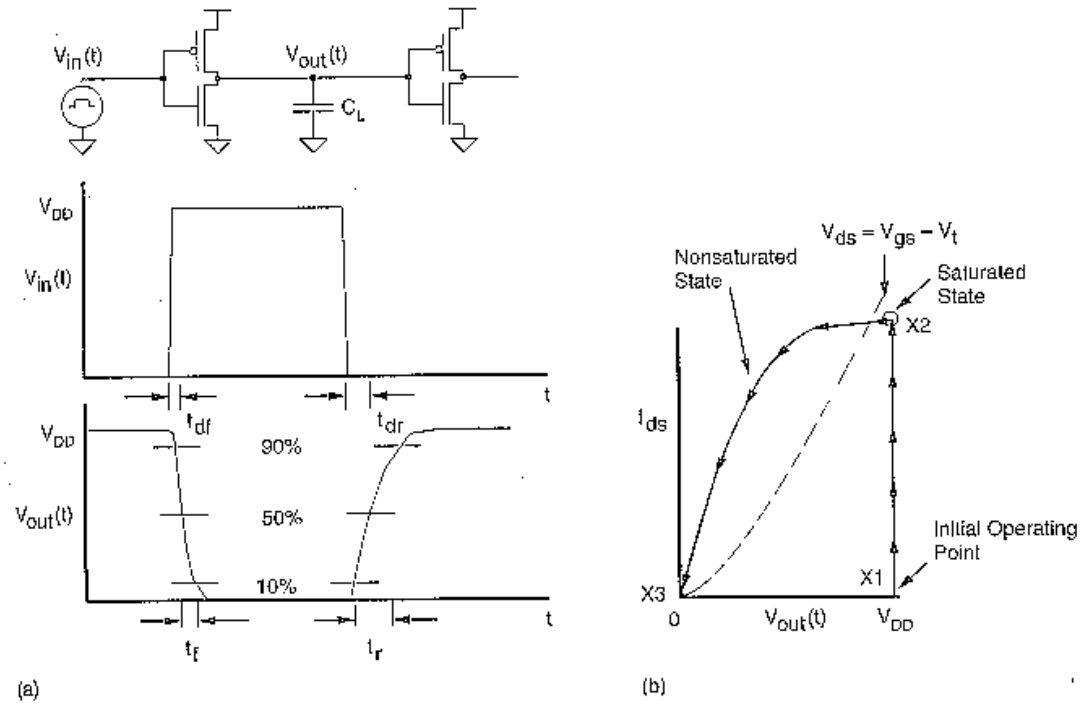


FIGURE 4.18 Switching characteristic for CMOS inverter (a) circuit and waveforms, (b) trajectory of n-transistor operating point during switching

- 4.5.1 Analytic Delay Models
- 4.5.1.1 Fall Time
  - Referring to Figure 4.18, initially the *n*-device is cut-off and the load capacitor C_L is charged to V_{DD} (at point *X1* on the characteristic curve). Application of a step voltage (i.e. $V_{gs} = V_{DD}$) at the input of the inverter changes the operating point to *X2*. From this onward, the operating point moves toward *X3*. Thus, it is evident that the fall time t_f consists of two intervals :
    1. t_{f1} = period during which the capacitor voltage V_{out} drops from *0.9 VDD* to $(V_{DD}$-$V_{tn})$. Equivalent circuit is shown in Figure 4.19(a).
    2. t_{f2} = period during which the capacitor voltage V_{out} drops from $(V_{DD}$-$V_{tn})$ to *0.1* V_{DD}.

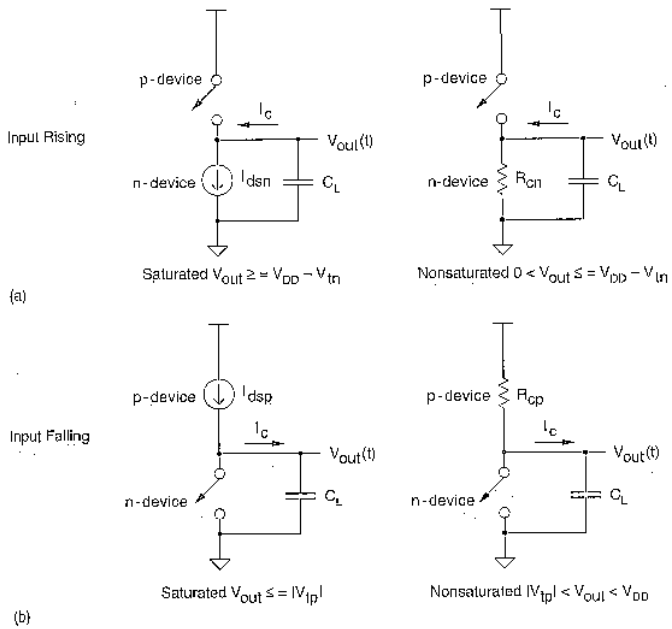


FIGURE 4.19 Equivalent circuits for fall- and rise-time determination

– During t_{f1}, the *n*-device is in saturation, thus

$$C_L \frac{dV_{out}}{dt} = -\frac{\beta_n}{2}(V_{DD} - V_{tn})^2$$ solving this equation for *t*, we obtain

$$t_{f1} = \frac{2C_L(V_{tn} - 0.1V_{DD})}{\beta_n(V_{DD} - V_{tn})^2}$$

– During t_{f2}, the *n*-device is non-saturated, thus $C_L \frac{dV_{out}}{dt} = -\beta_n((V_{DD} - V_{tn})V_{out} - \frac{1}{2}V^2{}_{out})$

Solving this equation, we obtain $t_{f2} = \frac{C_L}{\beta_n V_{DD}(1-n)}\ln(19 - 20n)$, where $n = \frac{V_{tn}}{V_{DD}}$

– $$t_f = t_{f1} + t_{f2} = \frac{2C_L}{\beta_n V_{DD}(1-n)}\left[\frac{(n-0.1)}{(1-n)} + \frac{1}{2}\ln(19 - 20n)\right] \qquad (4.37)$$

that is, $t_f \approx k \times \frac{C_L}{\beta_n V_{DD}}$ where *k = 3~4*.

- $t_f \propto C_L$, $t_f \propto 1/V_{DD}$ and $t_f \propto \frac{1}{\beta_n}$.

- 4.5.1.2 Rise Time
  - Due to the symmetry of the CMOS circuit, $t_r = \frac{2C_L}{\beta_p V_{DD}(1-p)}\left[\frac{p-0.1}{1-p} + \frac{1}{2}\ln(19-20p)\right]$ (4.39)

    where $p = \frac{|V_{tp}|}{V_{DD}}$. As before, $t_r \cong 3 \to 4\frac{C_L}{\beta_p V_{DD}}$.
  - For equally sized *n*- and *p*-transistors, where $\beta_n = 2\beta_p$, $t_f = \frac{t_r}{2}$
  - For equal rise and fall time, $\frac{\beta_n}{\beta_p} = 1$. This implies $w_p = 2 \sim 3w_n$, where w_p is the channel width of the p-device and w_n is the channel width of the *n*-device.

Prof. CHANDRA SHEKAR P (CSP)
Department of ECE
ATMECE, Mysuru

- 4.5.1.3 Delay Time
  - In most CMOS circuits, the delay of a single gate is dominated by the output rise and full times. It is approximated by $t_{dr} = \frac{t_r}{2}$ and $t_{df} = \frac{t_f}{2}$.
  - The average gate delay for rising and falling transition is $t_{av} = \frac{t_{df} + t_{dr}}{2} = \frac{t_r + t_f}{4}$.
  - Figure 4.20 illustrates a SPICE simulation of a step input applied to an inverter driving a capacitive load. With $V_{tn} = .767V$, $V_{tp} = -.938V$, $\beta_n = 4.04 \times 10^{-4}$, $\beta_p = 3.48 \times 10^{-4}$, $V_{DD} = 5.0V$, $C_L = 0.5pF$.
    - By Eq.(4.39), $t_r = 1.04ns$ (compared to *1.14ns* from SPICE)
    - By Eq.(4.37), $t_f = 0.83ns$ (compared to *0.89ns* from SPICE)
    - $t_{dr} = \frac{t_r}{2} = 0.502ns$ (compared to *0.52ns* by SPICE)
    - $t_{df} = \frac{t_f}{2} = 0.41ns$ (compared to *0.45ns* by SPICE).

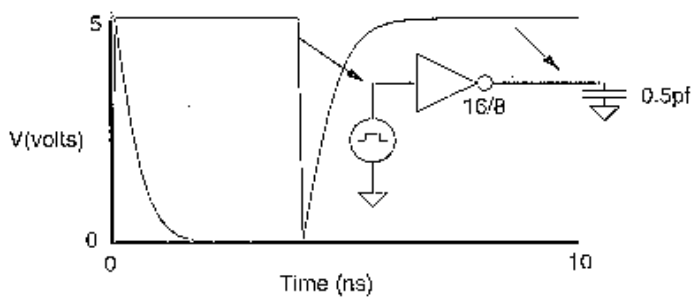


FIGURE 4.20 SPICE simulation of CMOS inverter transient response

– Alternatively, the output fall delay time can be approximated by $t_{df} = A_N \frac{C_L}{\beta_n}$ (i.e., $R_n = \frac{A_N}{\beta_n}$). A_N is a process-specific constant and derived as $A_N = \frac{1}{V_{DD}(1-n)}\left[\frac{2n}{1-n} + \ln(\frac{2(1-n)-V_O}{V_O})\right]$

where $n = \frac{V_{tn}}{V_{DD}}$ and $V_O = \frac{V_{out}}{V_{DD}}$.

Similarly, $t_{dr} = A_p \frac{C_L}{\beta_p}$ (E.Q. 4.47) and $A_p = \frac{1}{V_{DD}(1+p)}\left[\frac{-2p}{1+p} + \ln\left(\frac{2(1+p)-V_O}{V_O}\right)\right]$ where $p = \frac{V_{tp}}{V_{DD}} < 0$.

- 4.5.2 Empirical Delay Models
  - A circuit simulator is employed to find the timing parameter of interest and the coefficient of some delay equation is derived. For example, the coefficients A_P and A_N can be found by

$$A_p = t_{dr-spice}\frac{\beta_p}{C_L} = 0.52\times10^{-9}\times\frac{3.48\times10^{-4}}{0.5\times10^{-12}} = 0.36 \ (0.31\ calc)$$

$$A_N = t_{df-spice}\frac{\beta_n}{C_L} = 0.45\times10^{-9}\times\frac{4.04\times10^{-4}}{0.5\times10^{-12}} = 0.36 \ (0.29\ calc)$$

Note that the values of $t_{dr\text{-}spice}$ and $t_{df\text{-}spice}$ are obtained from SPICE simulation.

Thus, $t_{dr} = 0.36\frac{C_L}{\beta_p}$ and $t_{df} = 0.36\frac{C_L}{\beta_n}$ for the gates with $W_p = 2W_n$.

- 4.5.3 Gate delays
  - The delay of simple gate may be approximated by constructing an "equivalent" inverter. The pull-down *n*-transistor and pull-up *p*-transistor of the equivalent inverter are of a size to reflect the effective strength of the real pull-down or pull-up in the gate.
  - For example, the equivalent inverter for the 3-input NAND gate shown in Figure 4.21 has the effective β of the *n*-transistor as $\beta_{neff} = \dfrac{1}{\dfrac{1}{\beta_{n1}} + \dfrac{1}{\beta_{n2}} + \dfrac{1}{\beta_{n3}}}$
  - For $\beta_{n1} = \beta_{n2} = \beta_{n3}$, $\beta_{neff} = \dfrac{\beta_n}{3}$

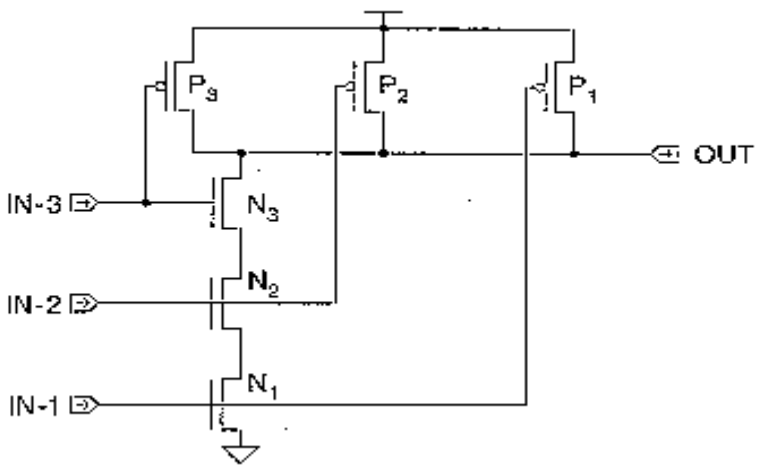


FIGURE 4.21 A 3-input NAND gate

– For the pull-up case, only one p-transistor has to turn on to raise the output. Thus, $\beta_{peff} = \beta_p$.

For $\beta_p = 0.3\beta_n$, $t_r = k\dfrac{C_L}{0.3\beta_n V_{DD}}$, $t_f = k\dfrac{C_L}{\frac{\beta_n}{3}V_{DD}}$. Thus $\dfrac{t_r}{t_f} \approx 1$.

– More clearly, for a series of 3 n-transistor, $\beta_{series} = \dfrac{\beta_n}{3}$

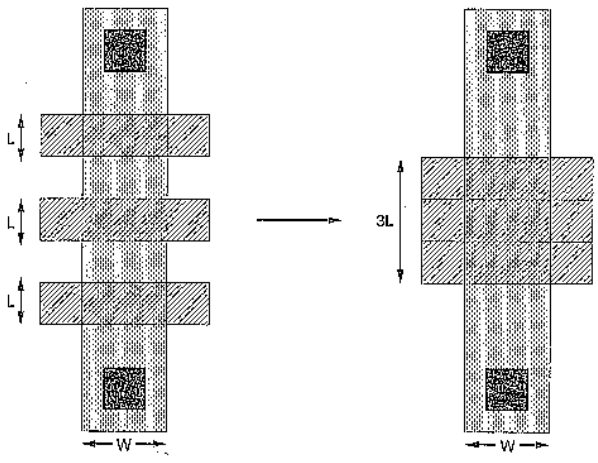


FIGURE 4.22 Graphical illustration of the effect of series transistors

– In general, the fall time is mt_f for m n-transistors in series; the rise time for m p-transistor in series is mt_r.

– The fall (rise) time for a parallel connection of $n(p)$-transistors is t_f/m (t_r/m) for m transistors in parallel, if all the transistors are turned on simultaneously.

- 4.6 CMOS-Gate Transistor Sizing
- 4.6.1 Cascaded Complementary Inverters
  - Previously discussed, we must take $W_p = (2 \rightarrow 3)\ W_n$ to approximately equalize the rise and fall times of an inverter. However, in some cascaded structures it is possible to use minimum or equal-sized devices to achieve the same result.
  - For example, as for the circuit shown in Figure 4.31(a), with $W_p = 2W_n$,

$$t_{inv-pair} \propto t_{fall} + t_{rise} \propto R3C_{eq} + 2\frac{R}{2}3C_{eq} = 3RC_{eq} + 3RC_{eq} = 6RC_{eq} \quad ,$$

where R is the effective "ON" resistance of a unit-sized n-transistor ($W=2, L=1$), and $C_{eq} = C_g + C_d$ is the capacitance of a unit-sized gate and drain region. As for the case shown in Figure 4.31(b) with $W_p = W_n$,

$$t_{inv-pair} \propto t_{fall} + t_{rise} \propto R2C_{eq} + 2R2C_{eq} = 6RC_{eq}$$

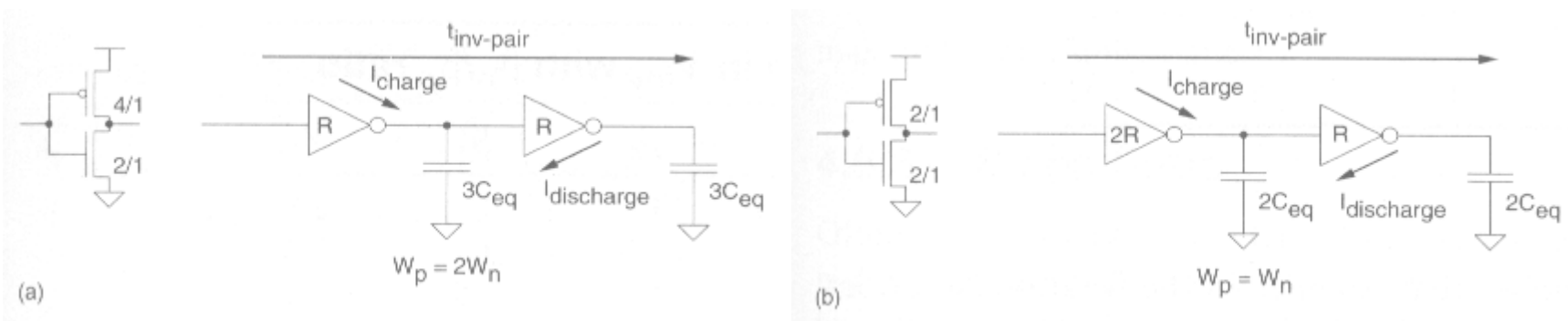


FIGURE 4.31 CMOS inverter pair timing response

– Also note that changes in the β ratio affect the inverter logic threshold voltage V_{inv}, which directly influences the delay of output response. From equation (2.2),

$$V_{inv} = \frac{V_{DD} + V_{tp} + V_{tn}\sqrt{\frac{\beta_n}{\beta_p}}}{1 + \sqrt{\frac{\beta_n}{\beta_p}}}$$

– Table 4.10 summarizes V_{inv} for a range of its constituent parameters.

TABLE 4.10 Variation in V_{inv} with β_n/β_p ratio

V_{DD}	V_{tn}	V_{tp}	β_n	β_p	V_{inv}
5	.7	–.7	1	1	2.5
5	.7	–.7	.5	1	2.8
5	.7	–.7	1	.5	2.2
3	.5	–.5	1	1	1.5
3	.5	–.5	.5	1	1.67
3	.5	–.5	1	.5	1.32

- 4.7 Power Dissipation
  - Two components of CMOS power dissipation:
    - *Static dissipation* due to leakage current.
    - *Dynamic dissipation* due to switching transient current and charging/discharging of load capacitance.
- 4.7.1 Static Dissipation
  - For the inverter shown in Figure 3.34, if the input is either at one of its steady-state logic values (0 or 1), there is no direct path from V_{DD} to V_{SS} . Thus no static power P_s dissipates, theoretically.
  - However, some small static dissipation actually exists due to:
    - Reverse biased *leakage* between diffusion regions and the substrate.
    - Subthreshold conduction (i.e. $V_{in} < |V_t|$)

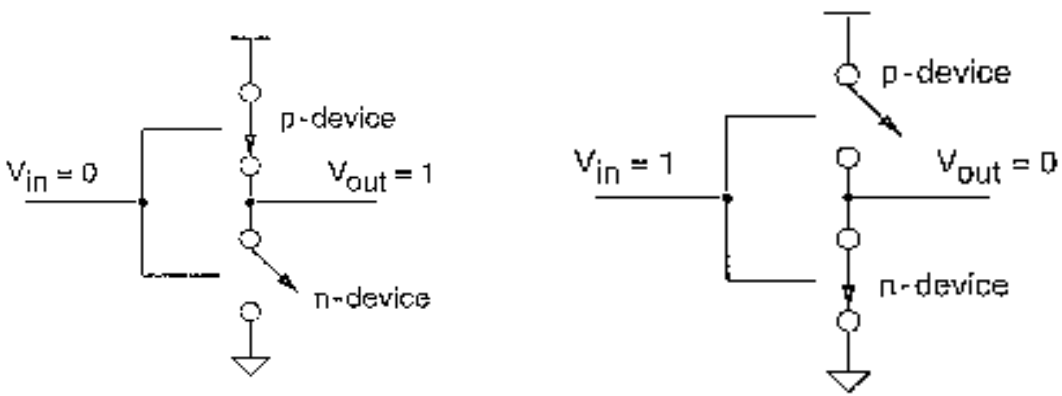


FIGURE 4.34
CMOS inverter model for static power dissipation evaluation

– The leakage current between diffusion and substrate can be explained by the model shown in Figure 3.35. The diodes in the model are reverse-biased and the leakage current is described by diode equation.

$$i_o = i_s(e^{qV/kT} - 1)$$

Where i_s = reverse saturation current

V = diode voltage

q = electronic charge ($1.602*10^{-19}$ C)

K = Boltzmann's constant ($1.38*10^{-23}$ J/K)

T = temperature

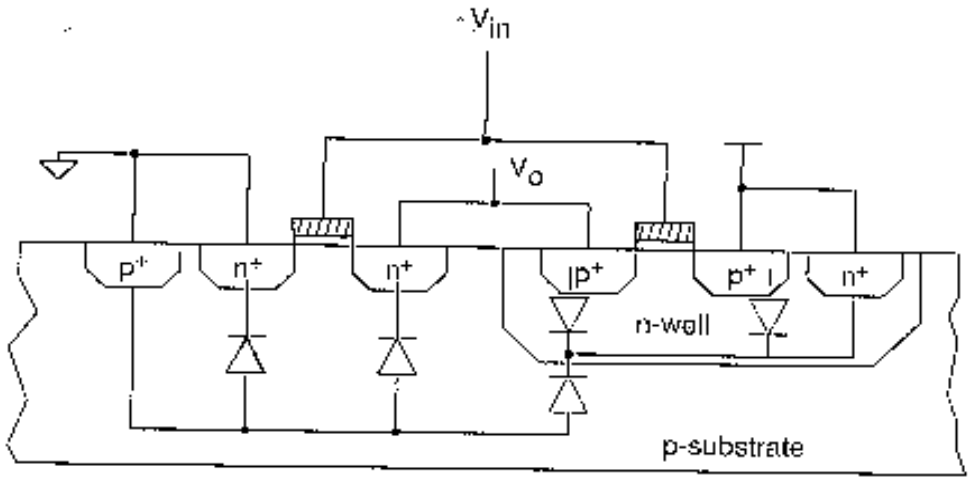


FIGURE 4.35 Model describing parasitic diodes present in a CMOS inverter

– The leakage current per device ranges from *0.1nA* to *0.5nA*.

– Total static power dissipation P_s can be obtained by

$$P_s = \sum_{1}^{n} \text{ leakage current * supply voltage}$$

where n = number of devices.

Example

For a process with β_p of 30 $\mu A/V^2$ and a β_n of 85 $\mu a/V^2$ ($V_{tn} = |V_{tp}| = 0.7V$, $V_{DD} = 5V$), calculate the static power dissipation of a 32 × 32 ROM which contains a 1:32 pseudo-nMOS row decoder and pMOS pull-ups on the 32-bit lines. The aspect ratio of all pMOS pull-ups (W/L) is 1. Each pMOS load can source $(\beta(V_{gs} - V_t)^2)/2$ of current.

$$I_{load} = \left(30\,(5 - 0.7)^{\,2} \right)/2\mu A \;=\; 277\mu A$$

$$P_{load} \;=\; 1.4\ mW \;=\; (277\mu A \times 5V)$$

Assuming that one row decoder is on and 50% of the bit lines are on at any one time yields

$$P_{total} \;=\; 17 \times 1.4\ mW$$

$$=\; 23.6\ mW.$$

- 4.7.2 Dynamic Dissipation
  - Consists of short circuit power and load capacitance charging/discharging power.
  - Short circuit dissipation is due to existence of a direct path from V_{DD} to V_{SS} when the output changes either from 1 to 0 or 0 to 1.
  - Short circuit dissipation depends on the input rise/fall time, the load capacitance and gate design.
  - Fig 4.36 depicts a scenario about how output loading could influence the short circuit current.
  - The average dynamic power P_d dissipated during switching for a square-wave input, V_{in}, having a repetition frequency of $f_p=1/t_p$ is given by

$$P_d = \frac{1}{t_p}\int_0^{\frac{t_p}{2}} i_n(t)V_{out}dt + \frac{1}{t_p}\int_{\frac{t_p}{2}}^{t_p} i_p(t)(V_{DD}-V_{out})dt$$

    For a step input and with $i_n(t) = \frac{C_L dV_{out}}{dt}$,

$$P_d = \frac{C_L}{t_p}\int_0^{V_{DD}} V_{out}dV_{out} + \frac{C_L}{t_p}\int_{V_{DD}}^{0}(V_{DD}-V_{out})d(V_{DD}-V_{out}) = \frac{C_L V_{DD}^{\ 2}}{t_p} = C_L V_{DD}^{\ 2} f_p$$

  - Dynamic power dissipation can be limited by reducing supply voltage, output capacitance and the switching frequency.

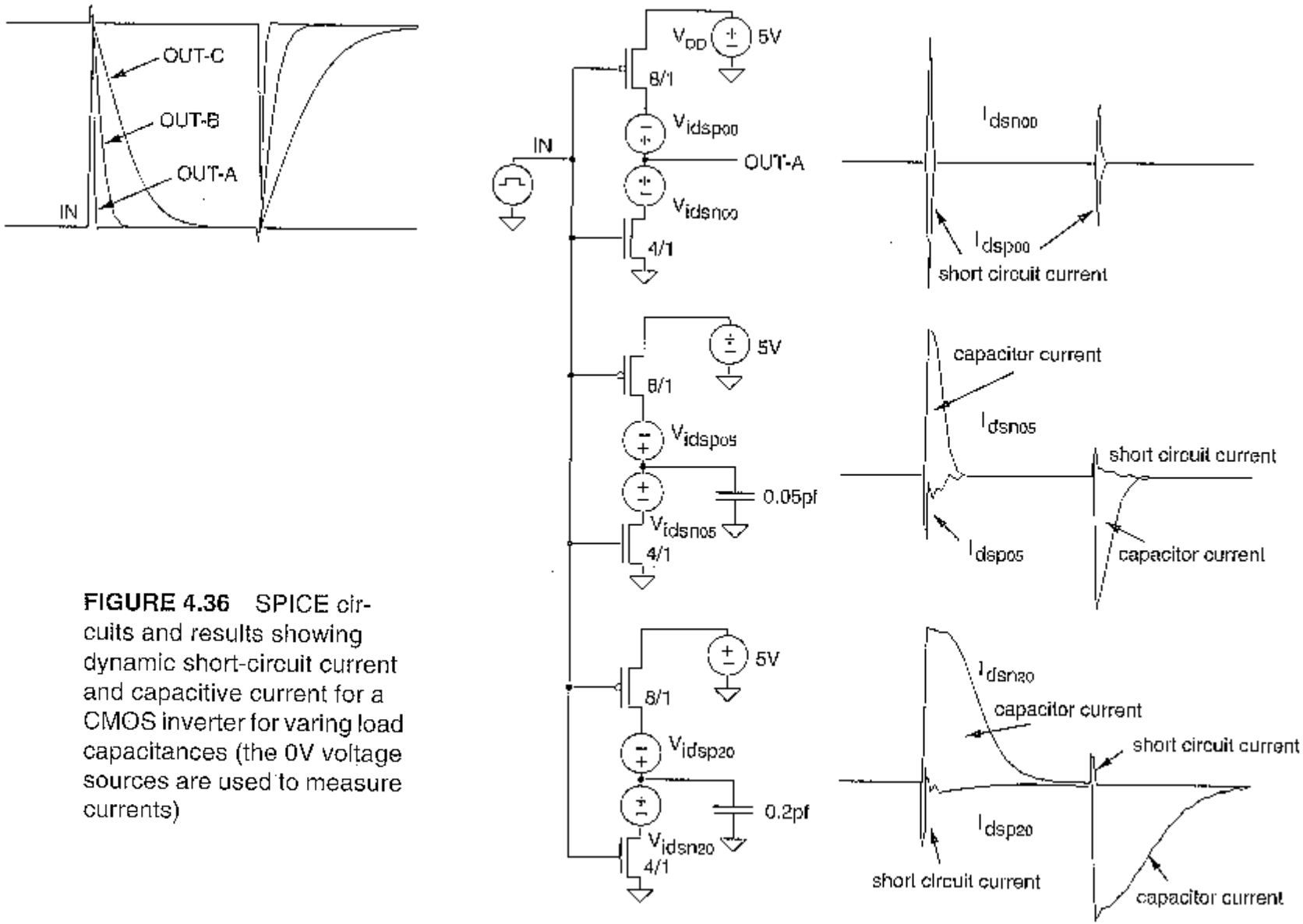


FIGURE 4.36 SPICE circuits and results showing dynamic short-circuit current and capacitive current for a CMOS inverter for varing load capacitances (the 0V voltage sources are used to measure currents)

- 4.7.3 Short-circuit Dissipation
  - For an inverter without load, assuming $t_r=t_f=t_{rf}$, the short circuit power dissipation (detailed derivation can be found in Weste textbook).

$$P_{sc} = \frac{\beta}{12}(V_{DD} - 2V_t)^3 \frac{t_{rf}}{t_p}$$

- 4.7.4 Total Power Dissipation
  - $P_{total}=P_s+P_d+P_{sc}$
  - More realistic situation, the dynamic power P_d should be calculated as follows:

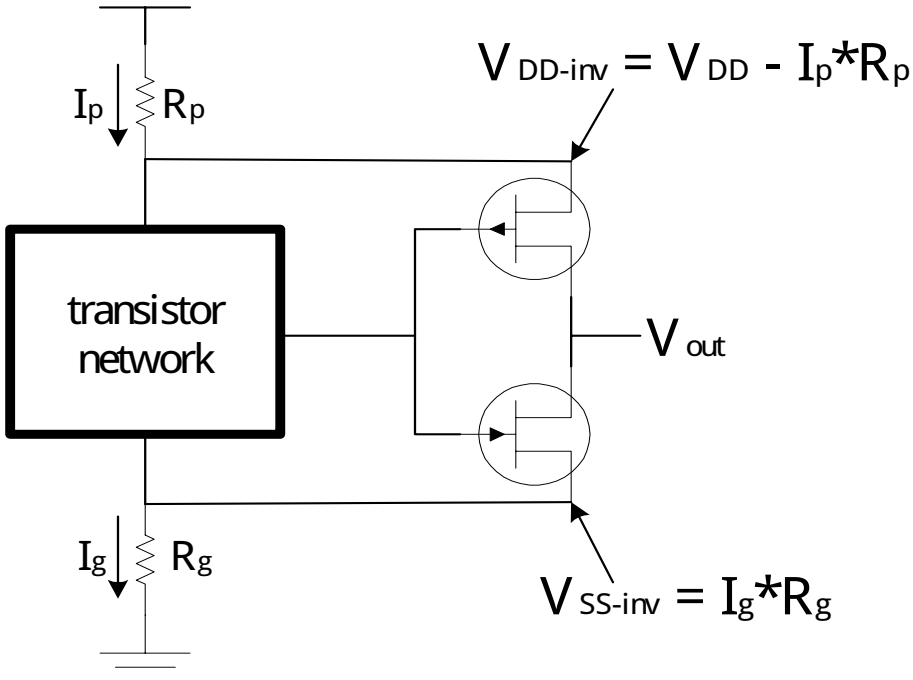
$$P_d = \frac{Percentage\text{-}activity * C_{Total} V_{DD}^2}{t_p},$$

where *percentage-activity* is a ratio between the estimated number of switchings and the number of clock cycles during a certain period of time.

Prof. CHANDRA SHEKAR P (CSP)
Department of ECE
ATMECE, Mysuru

- 4.8 Sizing Routing Conductors
  - Metal power-carrying conductors have to be sized for three reasons:
    - *Metal migration* (electro-migration)
    - *Power supply noise and integrity* (i.e., satisfactory power and signal voltage levels are presented to each gate.)
    - *RC delay*
  - Metal migration is the transport of metal ions through a conductor. It may cause breakage or deformation of conductor.
  - Factors that influence the electro-migration rate are:
    - *Current density*, as a rule of thumb, *0.4 mA/um* to *0.1 mA/um* of metal width should be used for both V_{DD} and V_{SS} lines.
    - *Temperature*
    - *Crystal structure*
  - Existence of substantial resistance on supply or ground line would cause considerable *IR* drop (power drop and ground bounce) during charging transients. This directly reduces the noise margin of gates and causes incorrect operation of gates.

- 4.8.1 Power Drop and Ground Bounce
  - For a clocked synchronous system, the output transitions of the gates are made close to the clock transition. Thus, a large current spike would appear on power or ground bus. This would effectively reduce the noise margin. (see the following figure)



Example

What would be the conductor width of power and ground wires to a 50 MHz clock buffer that drives $100pF$ of on-chip load to satisfy the metal-migration consideration (J_{AL} = 0.5 mA/μ)? What is the ground bounce with the chosen conductor size? The module is 500μ from both the power and ground pads and the supply voltage is 5 volts. The rise/fall time of the clock is 1 *ns*.

1. $P = CV_{DD}^2 f$

$= 100 \times 10^{-12} \times 25 \times 50 \times 10^6$

$= 125$ mW

$I = 25$ mA

Thus the width of the clock wires should be at least 60μ. A good choice would be 100μ.

2. $R = 500/100 \times .05$

$= 5$ squares $\times$ .05 Ω/sq.

$= .25\Omega$

$$IR = \frac{C_d V}{dt} R = \frac{100 \times 10^{-12} \times 5}{1 \times 10^{-9}} \times .25$$

$= 125$ mV (also see Section 5.5.16)

- 4.8.2 Contact Replication
  - For a very wide conductor to connect to another layer, an array of small contacts, suitably spaced, generally provides just as much current-carrying capacity as a single long, narrow contact. (see Figure 4.38)

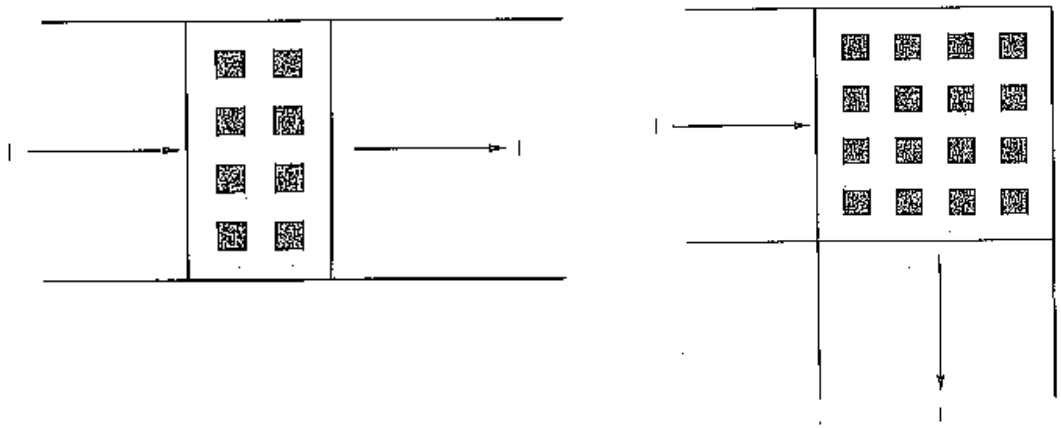


FIGURE 4.38 Contact structures for linear and orthogonal joints

- 4.9 Charge Sharing
  - In many structures a bus can be modeled as a capacitor C_b as shown in Figure 4.39. Sometimes the voltage on this bus is sampled (latched) to determine the state of a given signal. Charge sharing thus occurs.

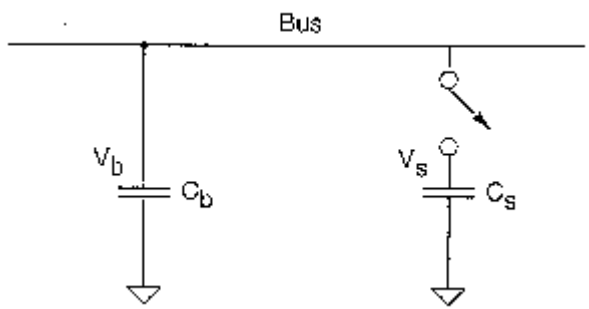


FIGURE 4.39 Charge-sharing mechanism

– For a successful sampling, the resultant voltage V_R (not shown) should be correctly reflect the state of the bus. For example,$Q_b=C_bV_b$, $Q_s=C_sV_s$, $Q_T=C_bV_b+C_sV_s$, $C_T=C_b+C_s$, Thus $V_R=Q_T/C_T=(C_bV_b+C_sV_s)/(C_b+C_s)$, if $V_b=V_{DD}$ (the bus is in state '1'), and $V_b>>V_s$, then $V_R=V_{DD}[C_b/(C_b+C_s)]$. For a successful sampling, $C_s<<C_b$. This results in $V_R \approx V_{DD}$.

– Example: A precharge bus has a loading of $10pF$. At a point in the clock cycle, 64 registers with transmission gates on their inputs turn on. The input load of each register (after the transmission gate) is $.1pF$. Calculate the change in precharge voltage.

What would be an alternative approach?

1. Here $C_b = 10pF$

$C_s = 64 \times .1pF = 6.4pF$

$V_{DD} = 5V$

Hence

$$V_R = 5 \times \frac{10}{10 + 6.4}$$

$= 3.05$ volts (change in voltage is 1.9V).

2. The most obvious approach to alleviating the problem is to use buffer inverters on the input of each register. (The above example would probably point to a very suspect design approach!)

- 4.11 Yield
  - The yield is influenced by such factors as:
    - technology
    - chip area
    - layout
  - Yield is defined as *Y= (No. of good chips on wafer) / (Total number of chips)*

- 4.12 Reliability
  - The potential reliability of a CMOS chip:
    - “Hot electron” effects
    - Electro-migration
    - Oxide failure
    - Die temperature
    - ESD protection.

- 4.13 Scaling of MOS-transistor Dimensions
  - *Scaling*: to shrink the size of physical dimensions or to scale the property values of a transistor by a dimension factor α .
  - Three types of scaling:
    - *Constant field scaling*: involve scaling of all physical dimensions (including those vertical to surface), device voltages and the concentration densities.
    - *Constant voltage scaling*: where V_{DD} voltage is kept constant, while the process is scaled.
    - *Lateral scaling*: where only gate length is scaled.
  - The resultant effect of these three types of scaling is shown in Table 4.12.

TABLE 4.12 Influence of Scaling on MOS-Device Characteristics

PARAMETER	SCALING MODEL		
	Constant field	Constant voltage	Lateral
Length (L)	$1/\alpha$	$1/\alpha$	$1/\alpha$
Width (W)	$1/\alpha$	$1/\alpha$	1
Supply voltage (V)	$1/\alpha$	1	1
Gate-oxide thickness (t_{ox})	$1/\alpha$	$1/\alpha$	1
Current ($I = (W/L)(1/t_{ox})V^2$)	$1/\alpha$	α	α
Transconductance (g_m)	1	α	α
Junction depth (X_j)	$1/\alpha$	$1/\alpha$	1
Substrate doping (N_A)	α	α	1
Electric Field across gate oxide (E)	1	α	1
Depletion layer thickness (d)	$1/\alpha$	$1/\alpha$	1
Load Capacitance ($C = WL/t_{ox}$)	$1/\alpha$	$1/\alpha$	$1/\alpha$
Gate Delay (VC/I)	$1/\alpha$	$1/\alpha^2$	$1/\alpha^2$
	RESULTANT INFLUENCE		
DC power dissipation (P_s)	$1/\alpha^2$	α	α
Dynamic power dissipation (P_d)	$1/\alpha^2$	α	α
Power-delay product	$1/\alpha^3$	$1/\alpha$	$1/\alpha$
Gate Area ($A = WL$)	$1/\alpha^2$	$1/\alpha^2$	$1/\alpha$
Power Density (VI/A)	1	α^3	α^2
Current Density	α	α^3	α^2

- 4.13.2 Interconnect-Layer scaling
  - Scaling of interconnect could result in changes of interconnect resistance and capacitance.
  - For example, scaling the thickness and width of a conductor by α ,the scaled-line resistance *R'* is given by

$$R' = R_s'[\frac{L/\alpha}{W/\alpha}] = \frac{\rho}{t/\alpha}[\frac{L/\alpha}{W/\alpha}] = \alpha R,$$

where ρ = conductivity

$R_s' = \frac{\rho}{t/\alpha}$ is sheet resistance of the scaled-line

t = conductor thickness before scaling

  - The voltage drop along the scaled-line for a constant field scaling is

$$V_d' = (\frac{I}{\alpha})(\alpha R) = IR$$

  - The line-response time is

$$t_s' = (\alpha R)(\frac{C}{\alpha}) = RC$$

– The influence of scaling on interconnect, if the interconnect is scaled by α and the current is increased by $\frac{1}{\alpha}$, is summarized in Table 4.13.

TABLE 4.13 Influence of Scaling on Interconnect Media (Constant Field)

PARAMETERS	SCALING FACTOR
Line resistance (r)	α
Line response (rc)	1
Voltage drop	1

- 4.13.3 Scaling in Practice
  – As it can be seen from Table 4.14, constant voltage scaling has been used in the past.

TABLE 4.14 A Scaling History Since 1980 (Personal)

YEAR	TECH	CHIP	SIZE (Tr)	SIZE (mm^2)	SPEED (MHz)	V_{DD}	TYPE OF SCALING
1980–1984	3.5µ	16-bit datapath and RAM	12K	25	5	5	
1985	2.0µ	Lisp µProcessor	250K	225	5	5	Constant voltage
1987	1.5µ	Lisp µProcessor	250K	144	8	5	Constant voltage
1989	1.2µ	Lisp µProcessor	250K	100	12	5	Constant voltage
1990	1.0µ	Ghost canceller	500K	54	56	3–5	Constant voltage
1992	0.8µ	Video decoder	1.2M	120	40	5	Constant voltage
1993+	0.5µ	??	>1M	100+	100+	3.3	Scaled V_{DD} and gate length

- 4.14 Summary
  - resistance, capacitance, inductance calculations
  - delay estimation
  - power estimation
  - design margining and reliability
  - effect of scaling

Prof. CHANDRA SHEKAR P (CSP)
Department of ECE
ATMECE, Mysuru