CAL POLY
SAN LUIS OBISPO

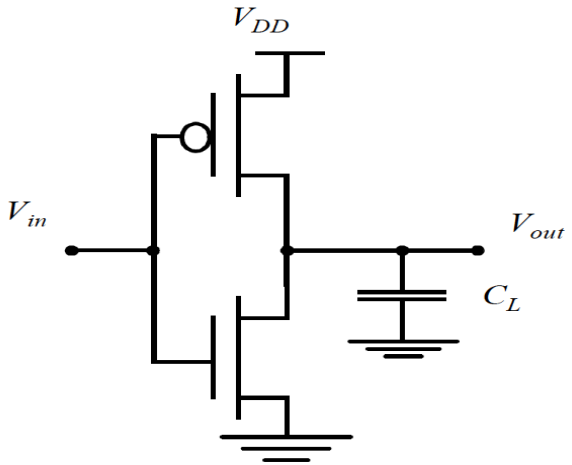
EE 431: COMPUTER-AIDED DESIGN OF VLSI DEVICES

CMOS Inverter Basics

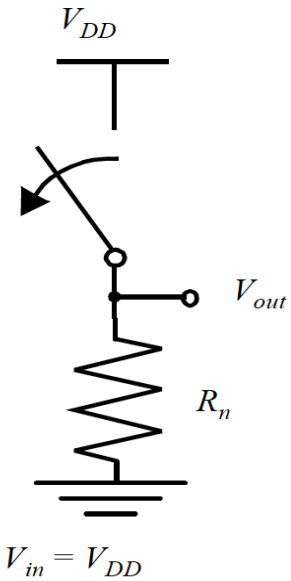
Nishith N. Chakraborty

September, 2024

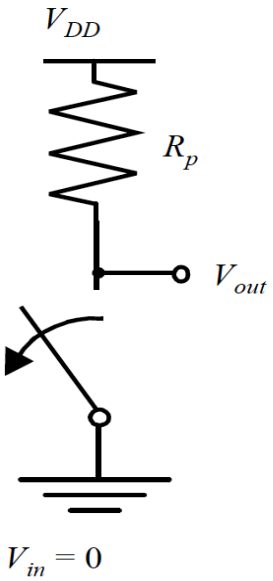
CMOS INVERTER: BASIC PRINCIPLES



IN	OUT	V_{in}	V_{out}
0	1	0	V_{DD}
1	0	V_{DD}	0

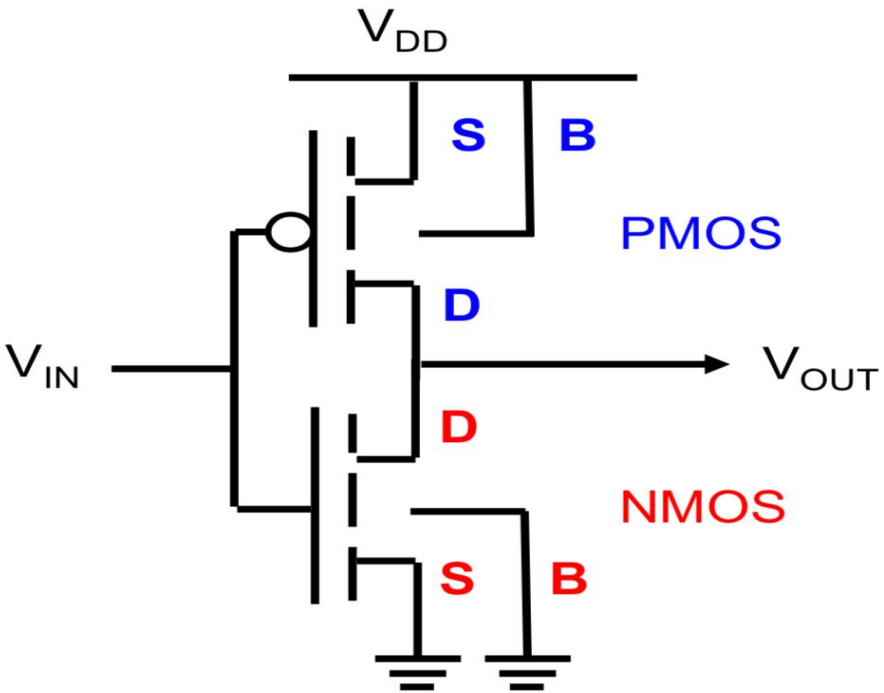


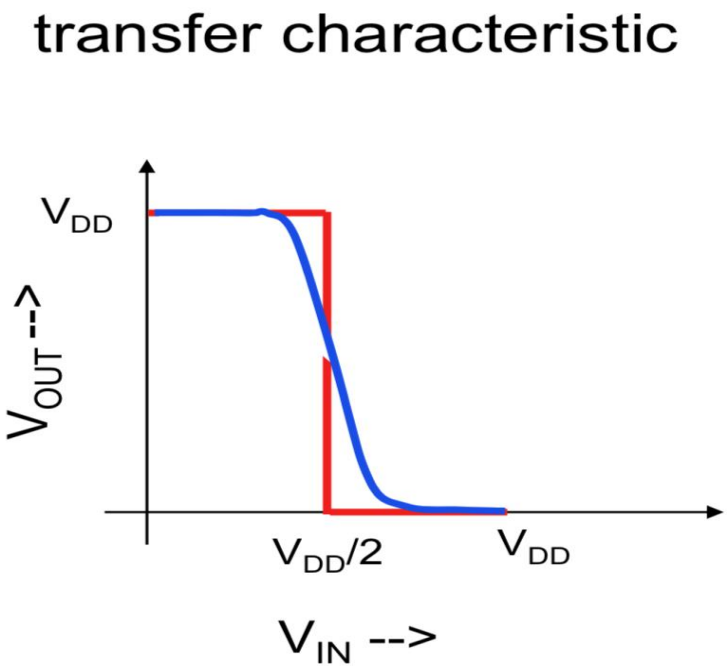
(a) Model for high input



(b) Model for low input

VOLTAGE TRANSFER CHARACTERISTICS



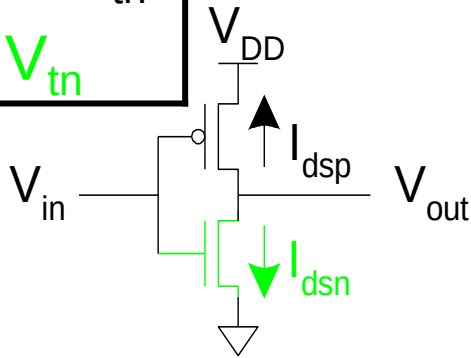


NMOS OPERATION

Cutoff	Linear	Saturated
$V_{gsn} < V_{tn}$ $V_{in} < V_{tn}$	$V_{gsn} > V_{tn}$ $V_{in} > V_{tn}$ $V_{dsn} < V_{gsn} - V_{tn}$ $V_{out} < V_{in} - V_{tn}$	$V_{gsn} > V_{tn}$ $V_{in} > V_{tn}$ $V_{dsn} > V_{gsn} - V_{tn}$ $V_{out} > V_{in} - V_{tn}$

$V_{gsn} = V_{in}$

$V_{dsn} = V_{out}$



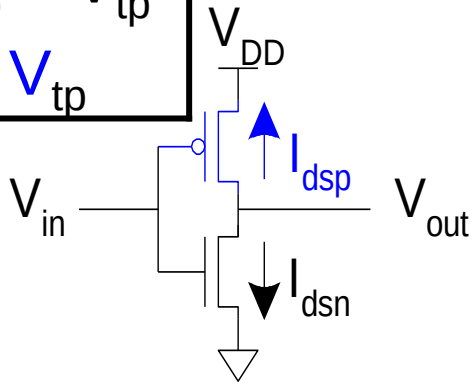
PMOS OPERATION

Cutoff	Linear	Saturated
$V_{gsp} > V_{tp}$ $V_{in} > V_{DD} + V_{tp}$	$V_{gsp} < V_{tp}$ $V_{in} < V_{DD} + V_{tp}$ $V_{dsp} > V_{gsp} - V_{tp}$ $V_{out} > V_{in} - V_{tp}$	$V_{gsp} < V_{tp}$ $V_{in} < V_{DD} + V_{tp}$ $V_{dsp} < V_{gsp} - V_{tp}$ $V_{out} < V_{in} - V_{tp}$

$V_{gsp} = V_{in} - V_{DD}$

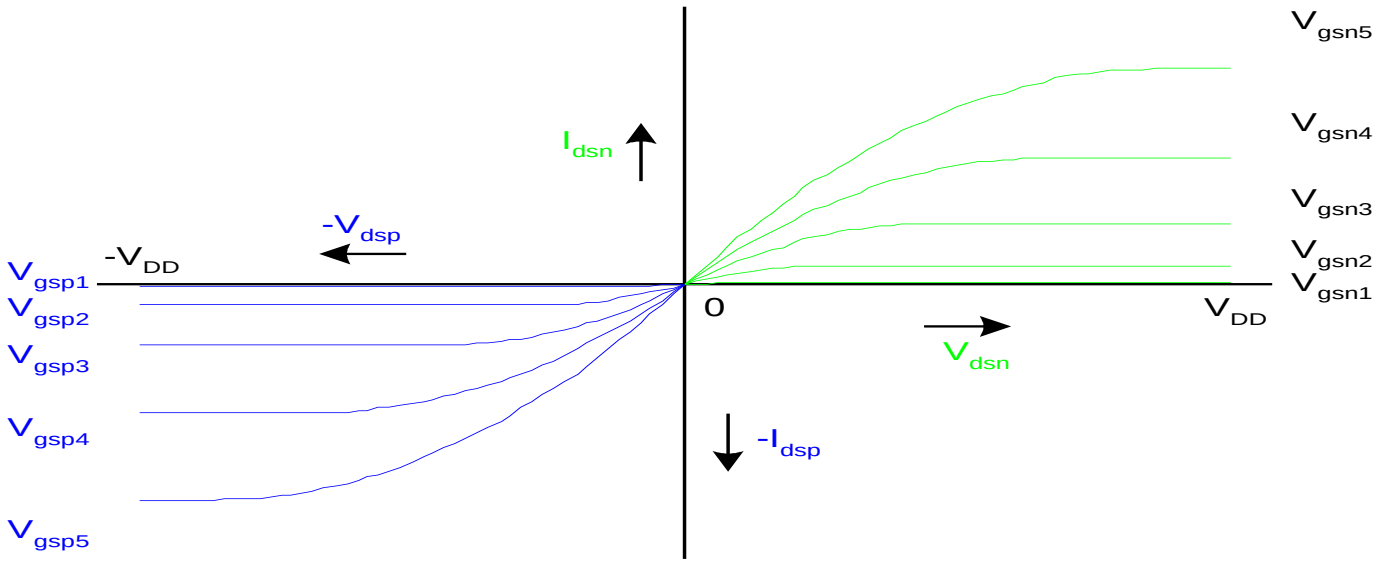
$V_{dsp} = V_{out} - V_{DD}$

$V_{tp} < 0$



I-V CHARACTERISTICS

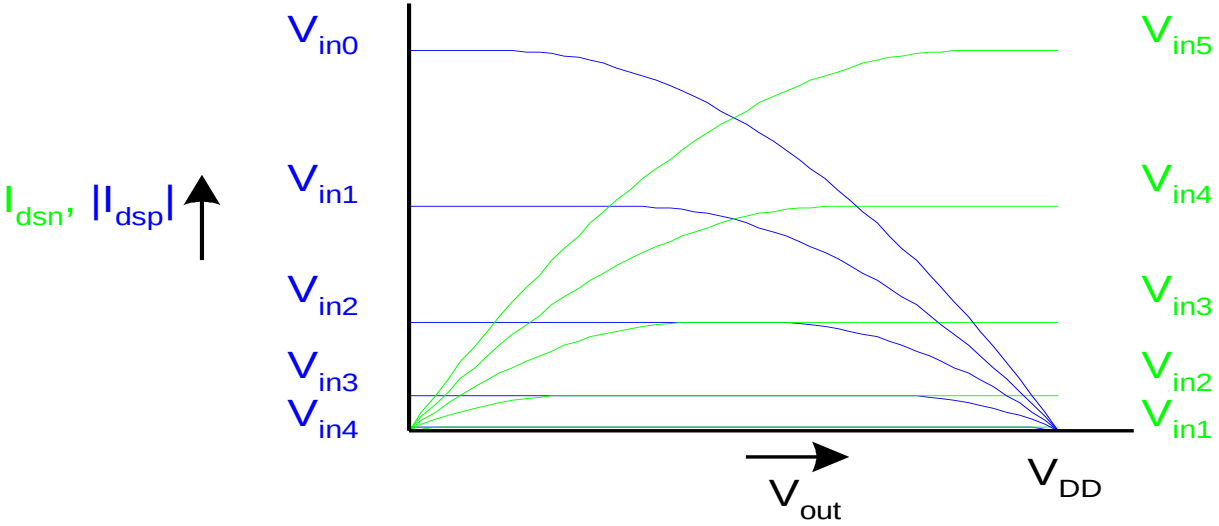
Make pMOS wider than nMOS such that $\beta_n = \beta_p$ ($\beta = \mu C_{ox}(W/L)$)

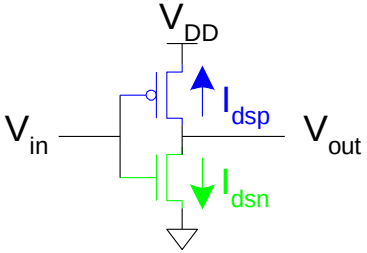


CURRENT VS V_{OUT} : LOAD LINE ANALYSIS

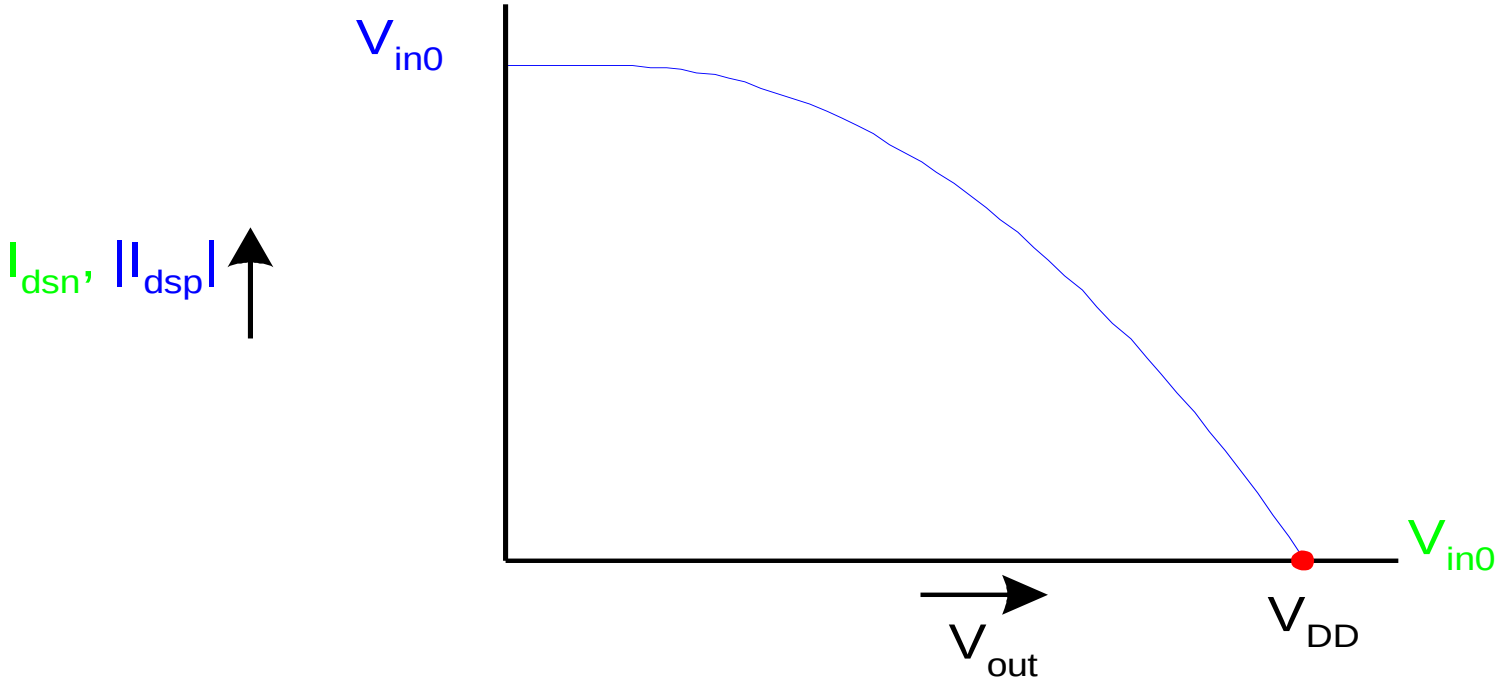
For a given V_{in}:

- Plot I_{dsn}, I_{dsp} vs. V_{out}
- V_{out} must be where |currents| are equal

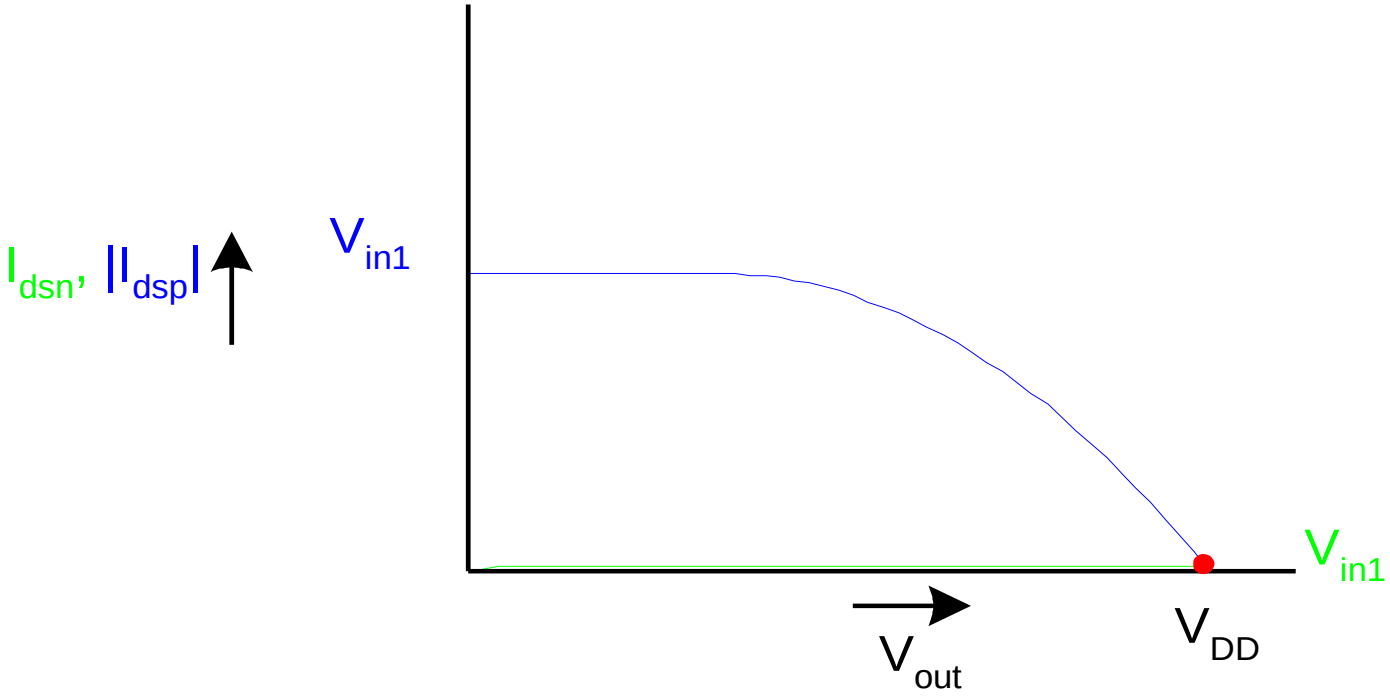




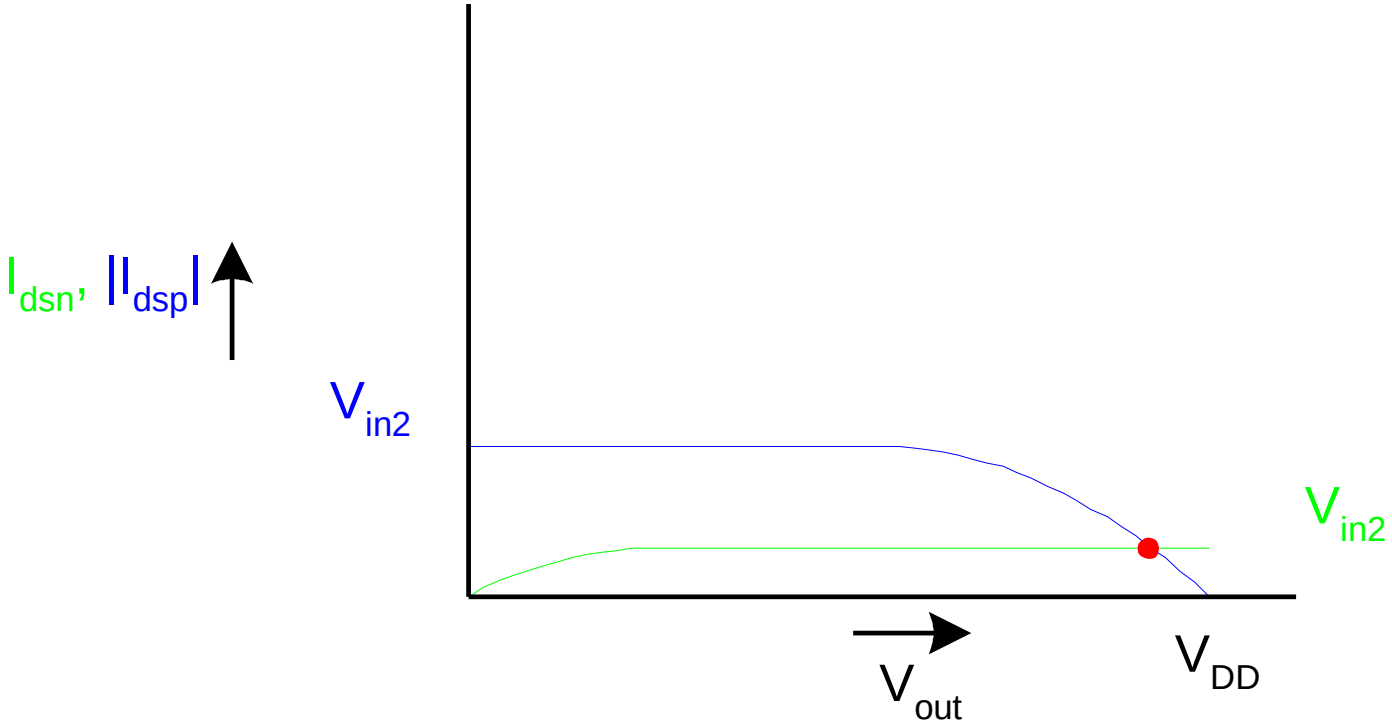
LOAD LINE ANALYSIS: $V_{IN} = 0$



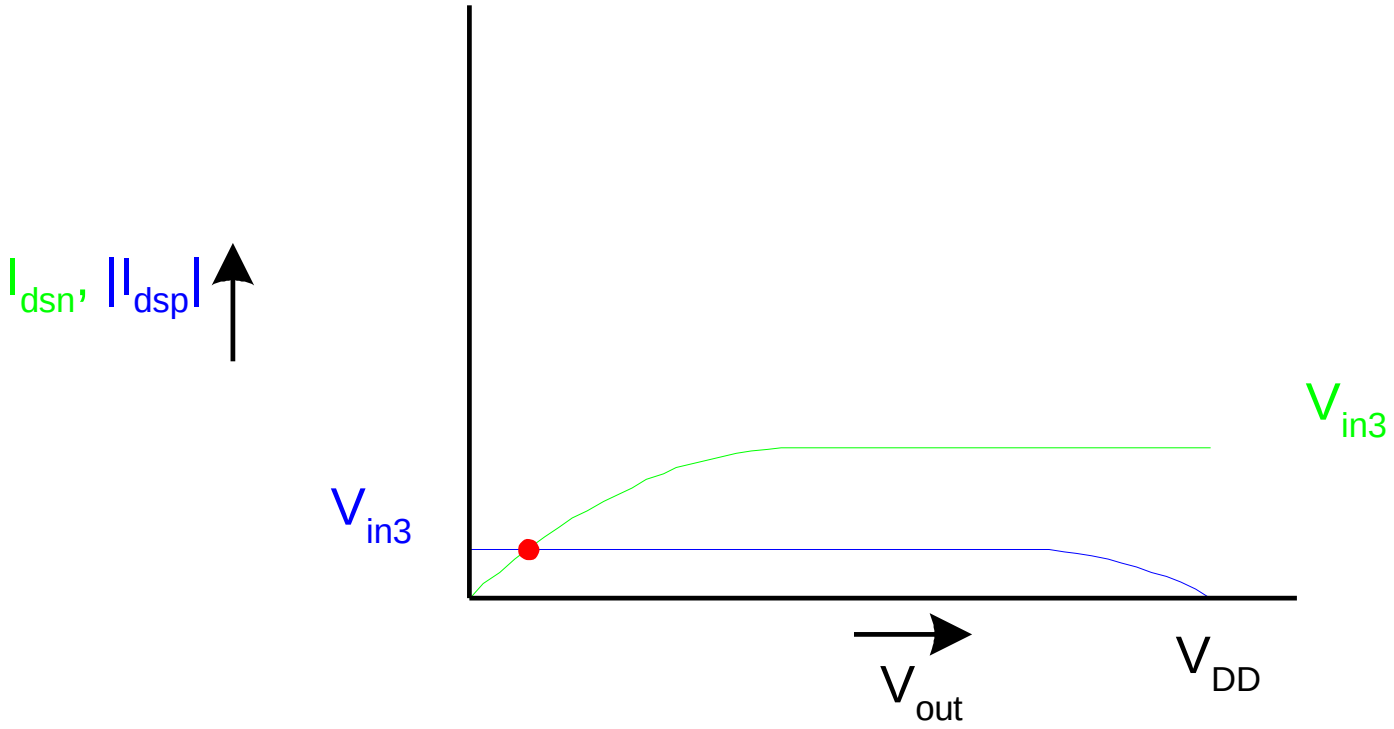
LOAD LINE ANALYSIS: $V_{IN} = 0.2V_{DD}$



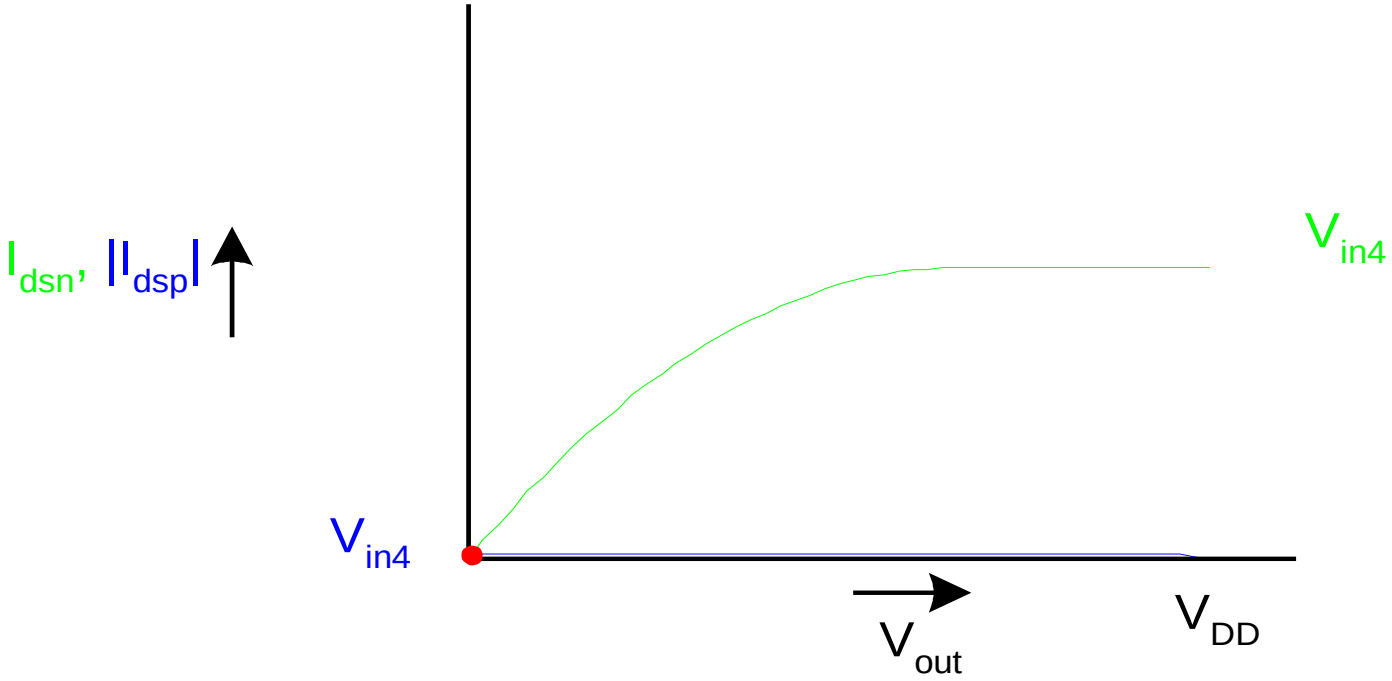
LOAD LINE ANALYSIS: $V_{IN} = 0.4V_{DD}$



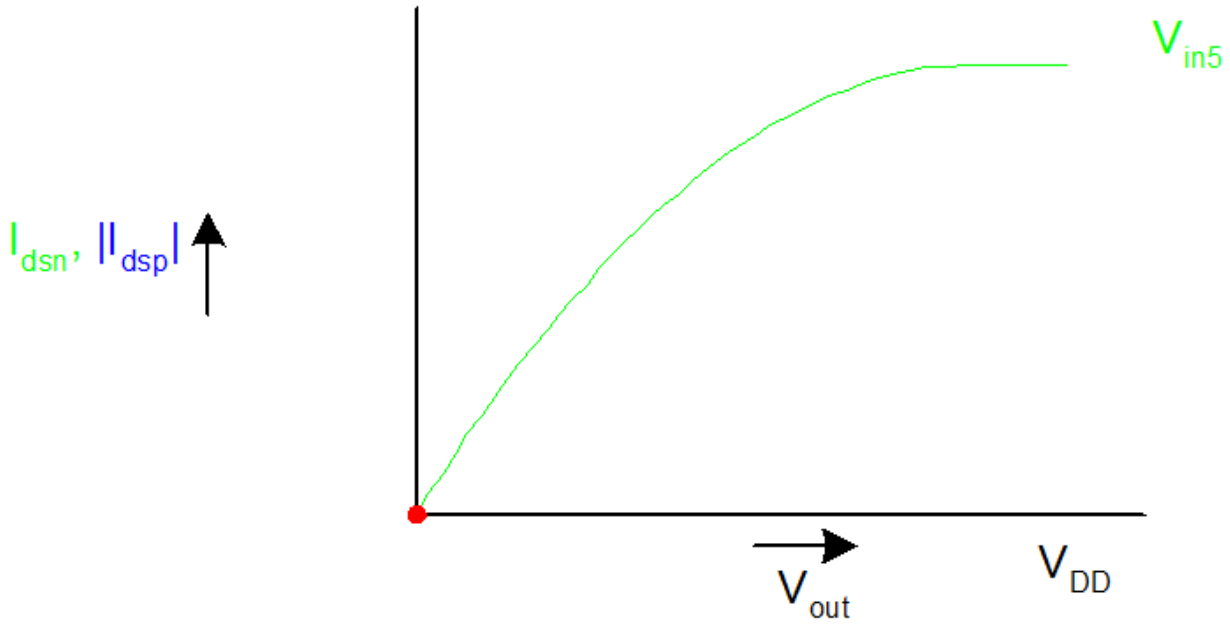
LOAD LINE ANALYSIS: $V_{IN} = 0.6V_{DD}$



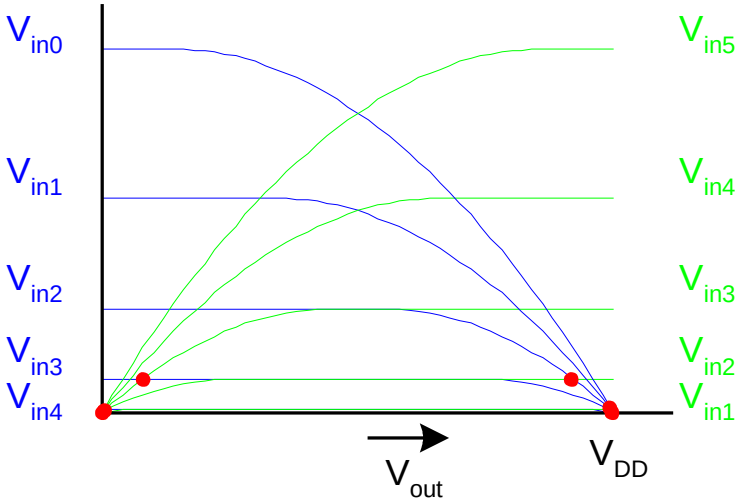
LOAD LINE ANALYSIS: $V_{IN} = 0.8V_{DD}$

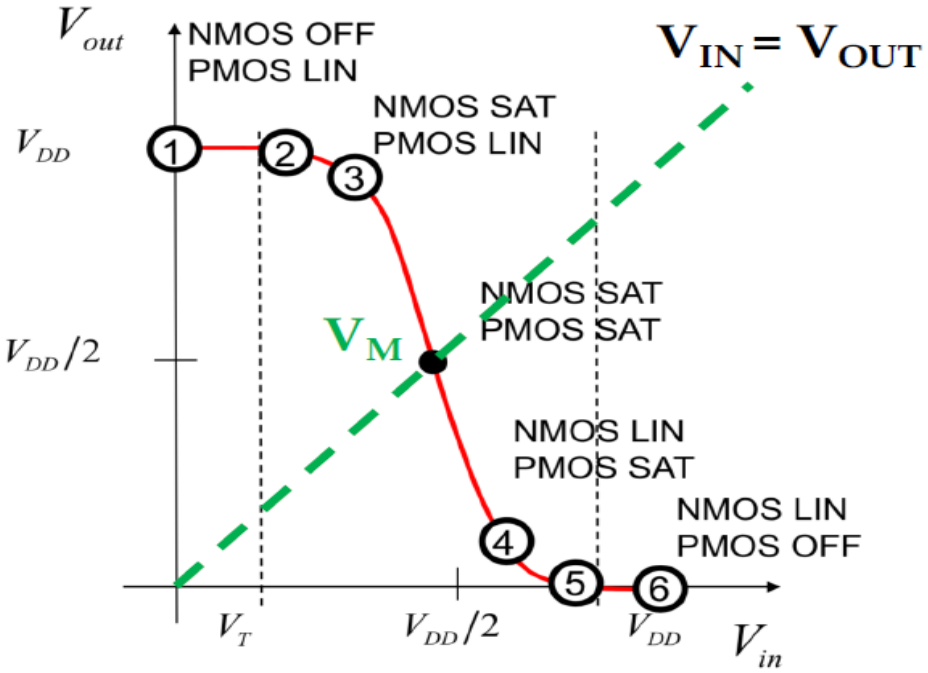


LOAD LINE ANALYSIS: $V_{IN} = V_{DD}$



DC TRANSFER CURVE





Logic Threshold Voltage: V_M

CALCULATING SWITCHING THRESHOLD

V_M : $V_{IN} = V_{OUT} = V_M$

1) $V_{GSN} = V_{IN} = V_M$, $V_{DSN}=V_{OUT}=V_M$,
$V_{GSP} = V_{IN}-V_{DD} = V_M-V_{DD}$, $V_{DSP} = V_{OUT}-V_{DD} = V_M-V_{DD}$

2) $I_N = - I_P$

3) Since $V_{GSN} = V_{DSN}$ and $V_{GSP}=V_{DSP}$, both NMOS and PMOS are in saturation (Because $|V_{DSN,P}| > |V_{GSN,P}| - |V_{THN,P}|$)

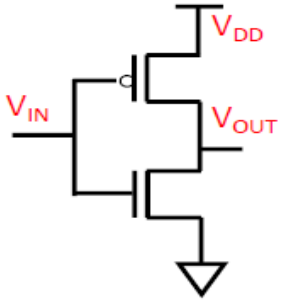
4) $\beta_N \left[\frac{(V_{GSN}-V_{THN})^2}{2}\right][1+\lambda_N V_{DSN}] = \beta_P \left[\frac{(V_{GSP}-V_{THP})^2}{2}\right][1+\lambda_P V_{DSP}]$

5) $\beta_N \left[\frac{(V_M-V_{THN})^2}{2}\right][1+\lambda_N V_M] = \beta_P \left[\frac{(V_M-V_{DD}-V_{THP})^2}{2}\right][1+\lambda_P (V_{DD}- V_M)]$

6) Solve for V_M.

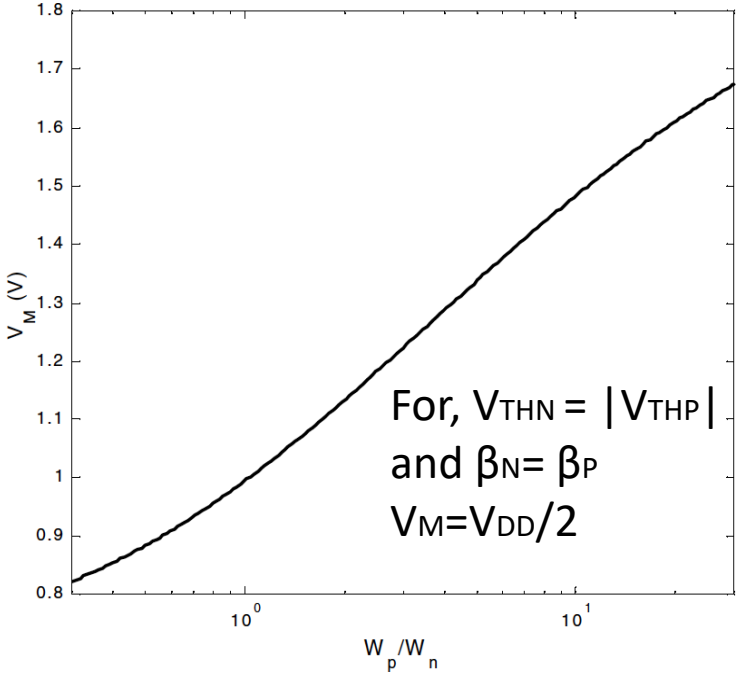
Assuming $\lambda_N=\lambda_P \sim 0$

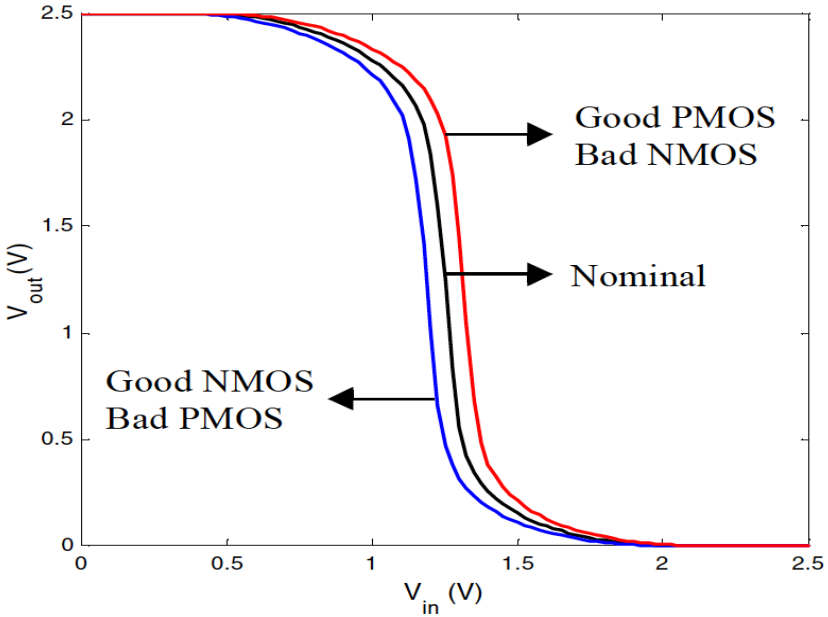
$$\sqrt{\beta_N}\,(V_M - V_{THN}) = \sqrt{\beta_P}\,(V_{DD} - V_M - |V_{THP}|)$$



$$V_M = \frac{V_{DD}-|V_{THP}|+\sqrt{\frac{\beta_N}{\beta_P}}V_{THN}}{\sqrt{\frac{\beta_N}{\beta_P}}+1}$$

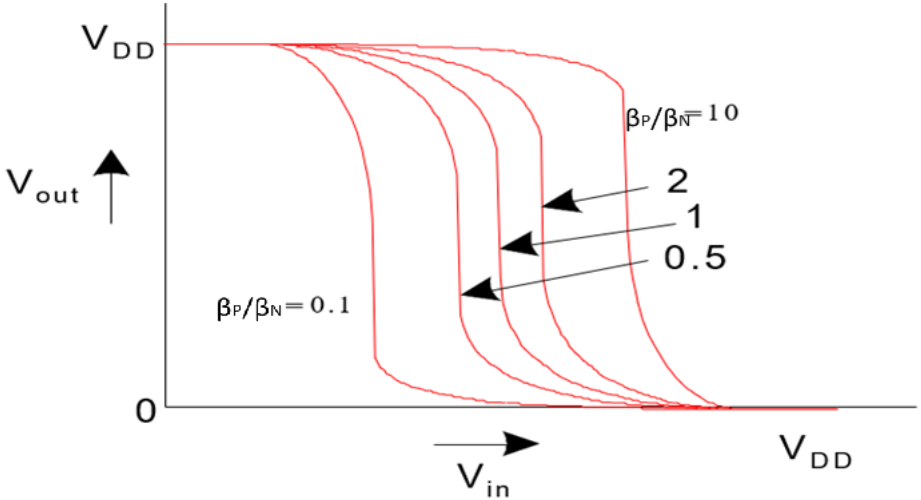
SWITCHING THRESHOLD VOLTAGE, V_M



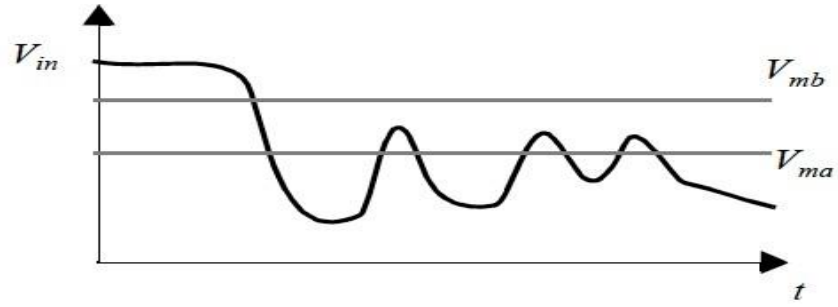


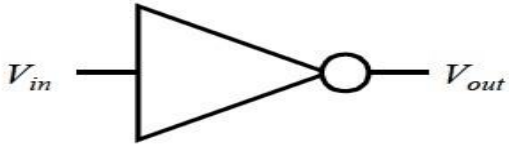
BETA RATIO

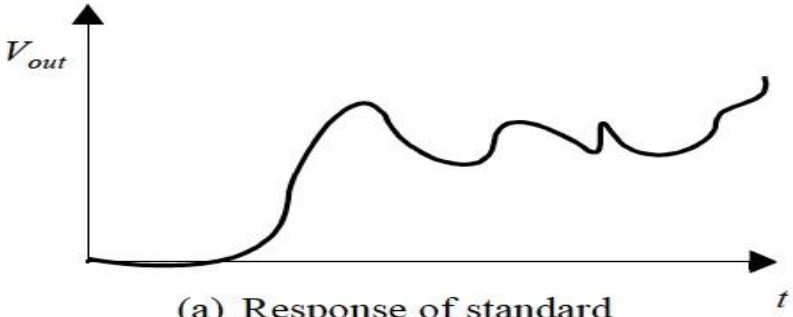
- If $\beta_P/\beta_N \neq 1$, switching point will move from $V_{DD}/2$
- Called *skewed* gate
- Other gates: collapse into equivalent inverter



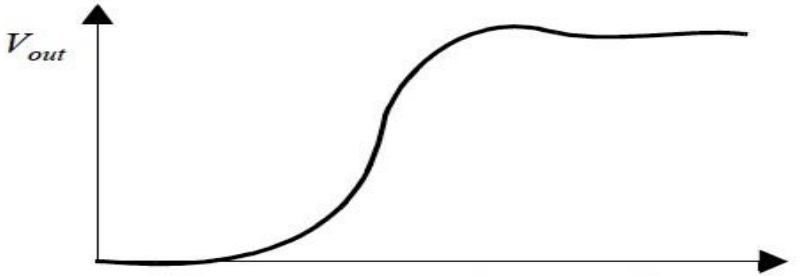
SWITCHING THRESHOLD VOLTAGE, V_M







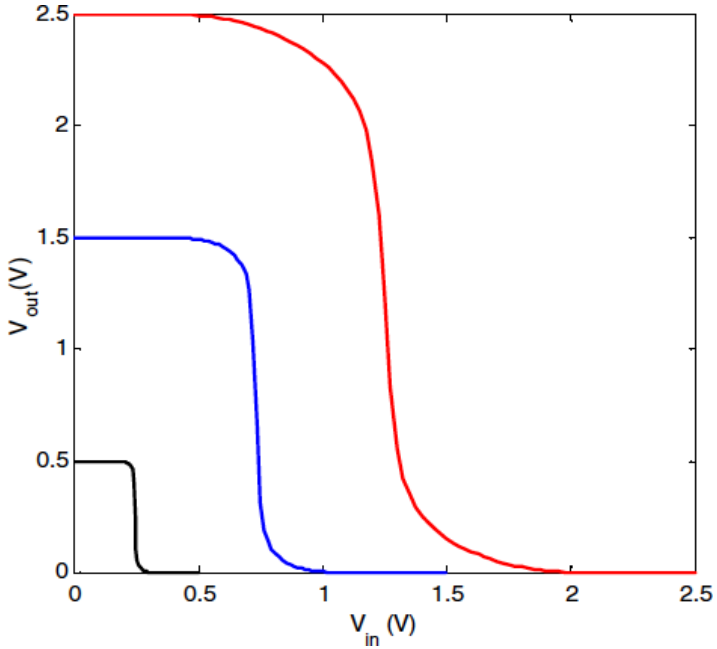
(a) Response of standard inverter



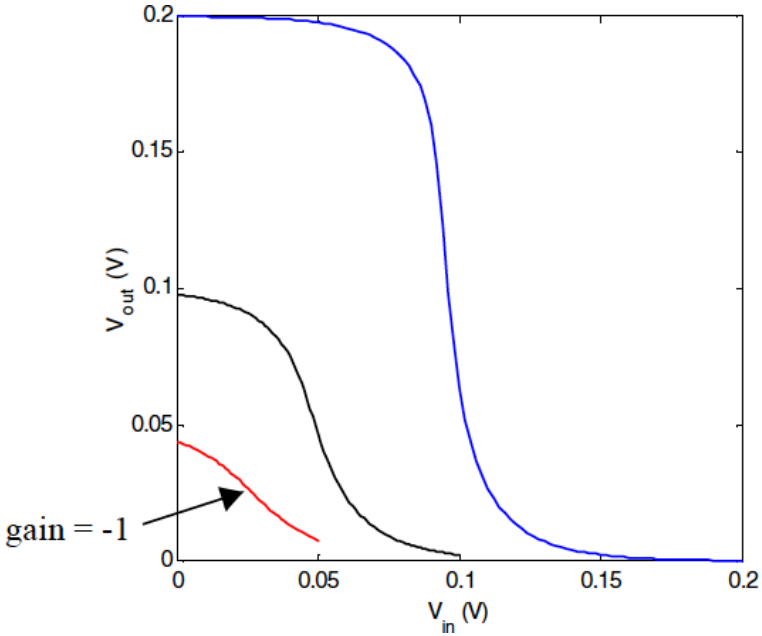
b) Response of inverter with modified threshold

SUPPLY VOLTAGE SCALING

Reducing V_{DD} improves the gain

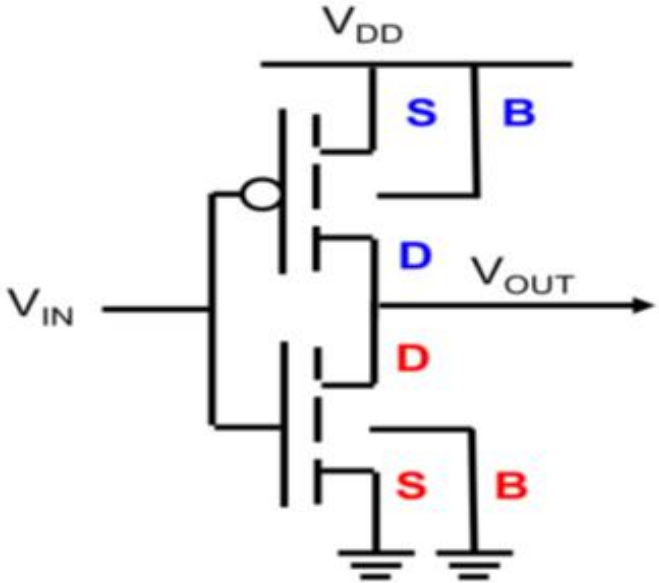


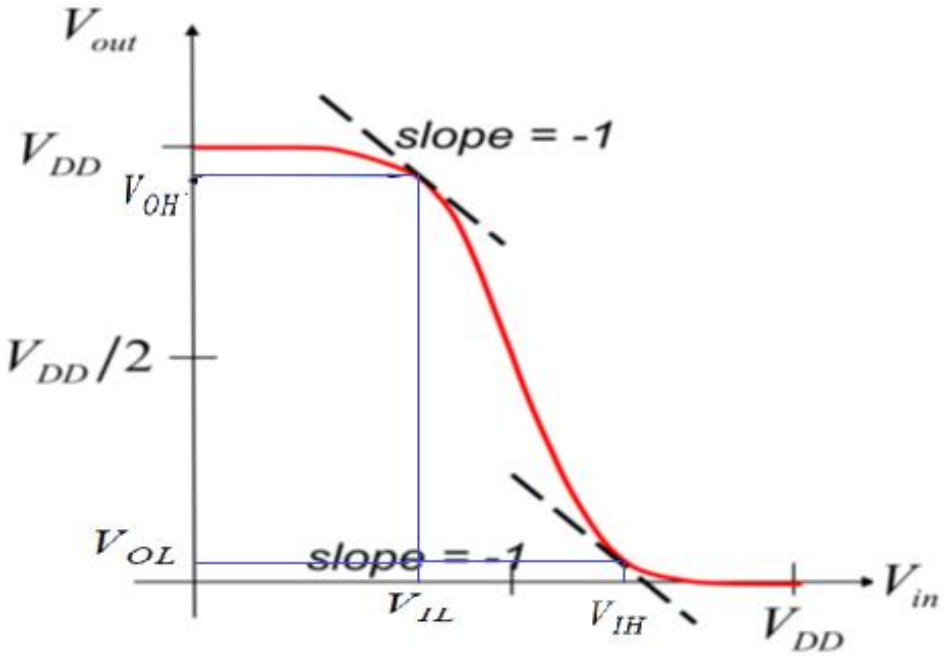
The gain decreases for very-low V_{DD}



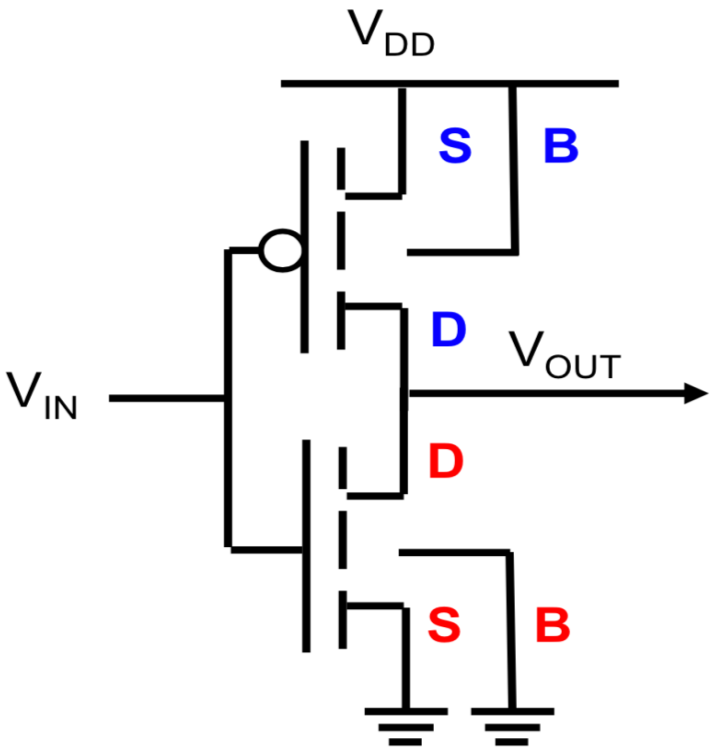
NOISE MARGIN

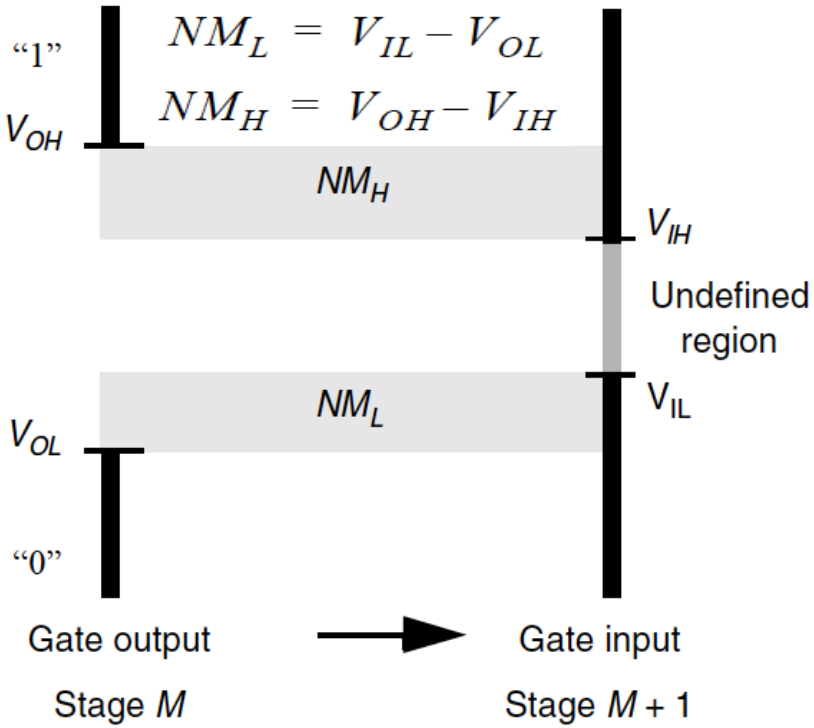
How much noise can a gate input see before it does not recognize the input?



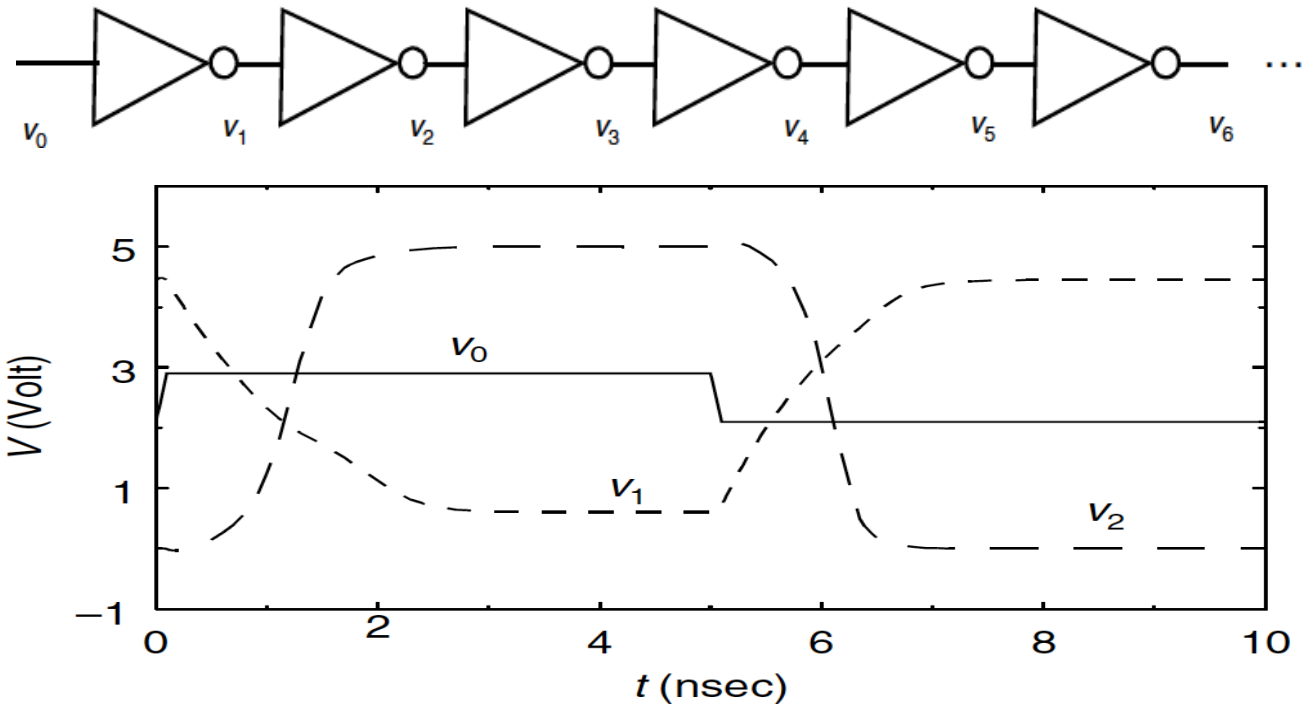


NOISE MARGIN

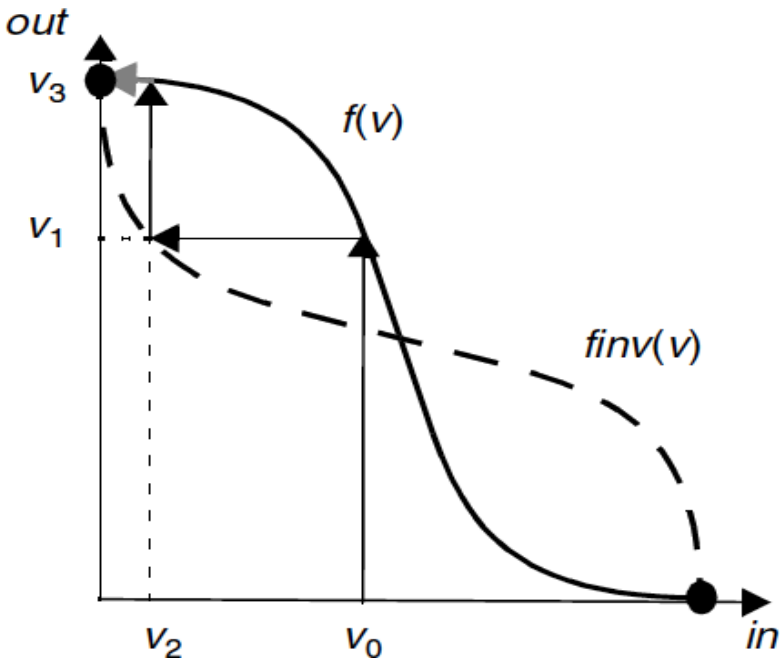




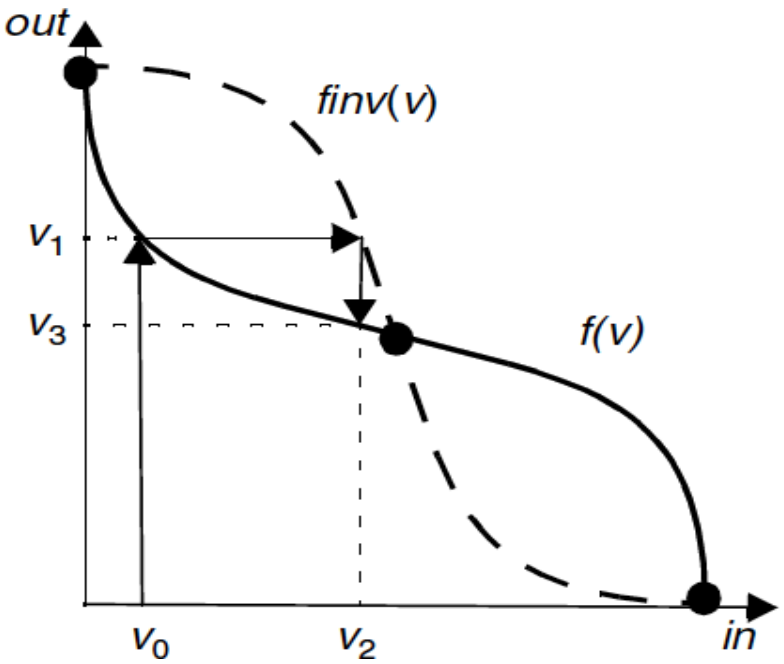
REGENERATIVE PROPERTY



REGENERATIVE PROPERTY

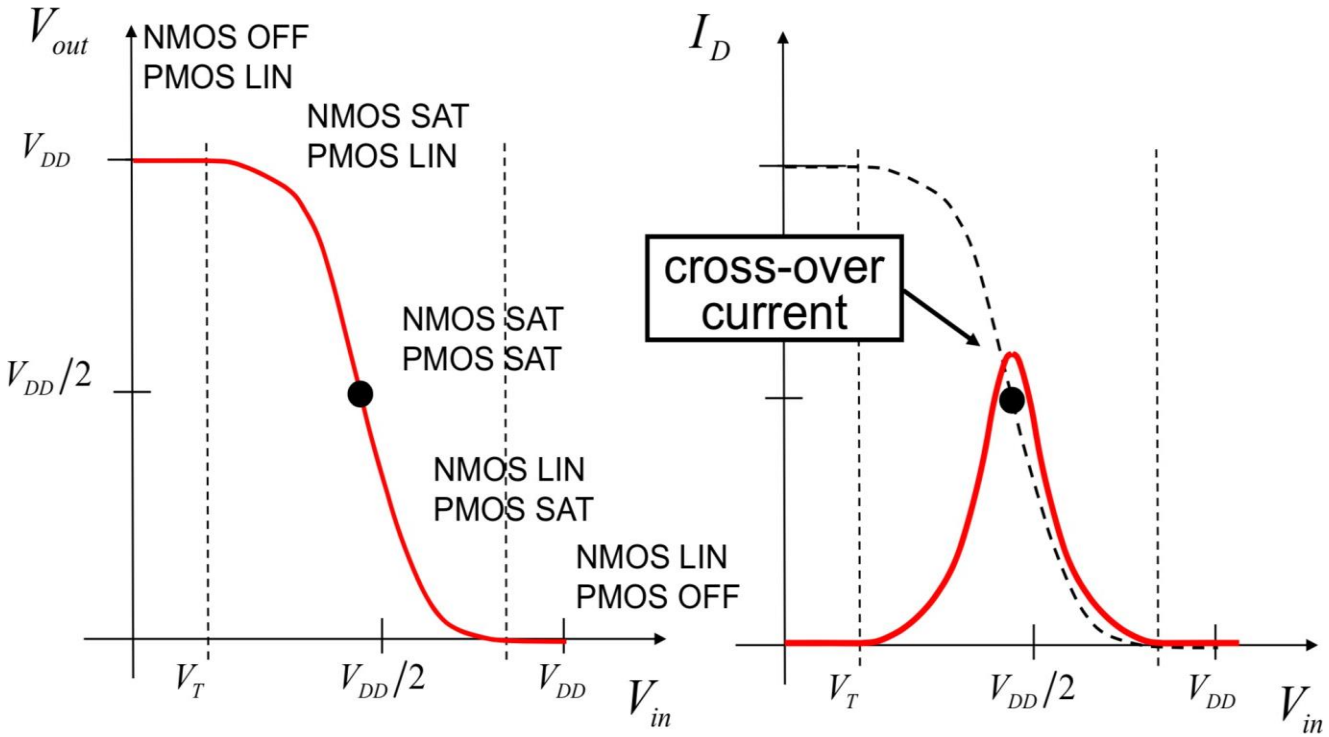


Regenerative Gate

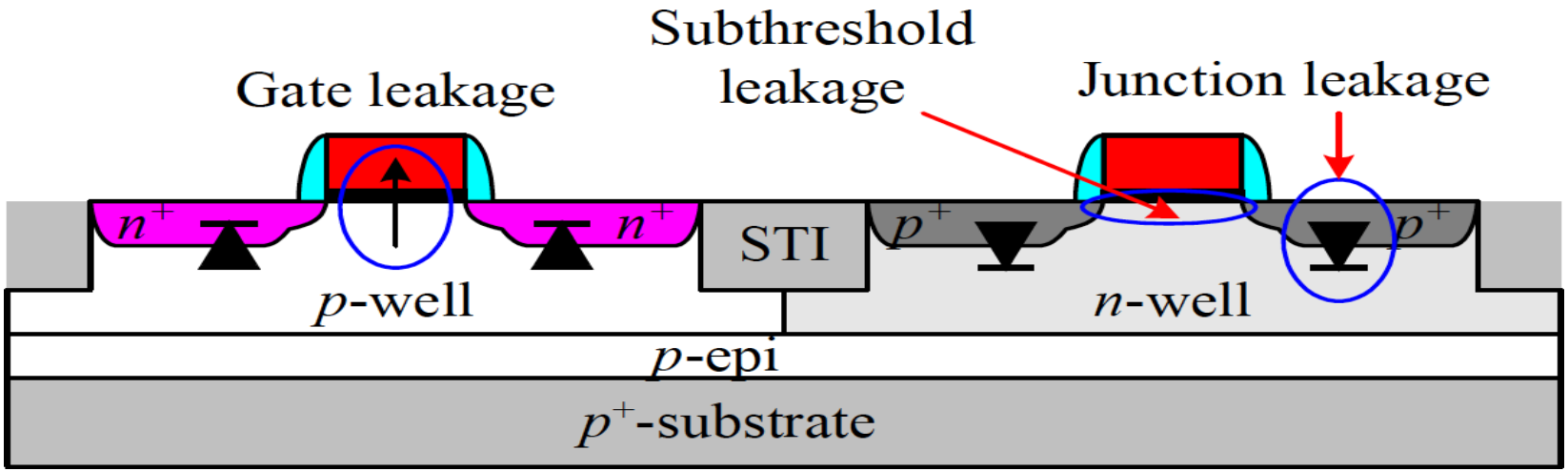


Nonregenerative Gate

SHORT CIRCUIT CURRENT



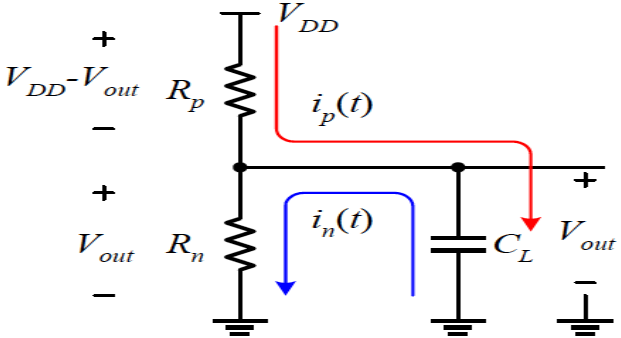
STATIC POWER DISSIPATION

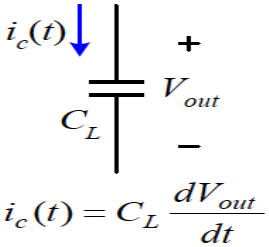


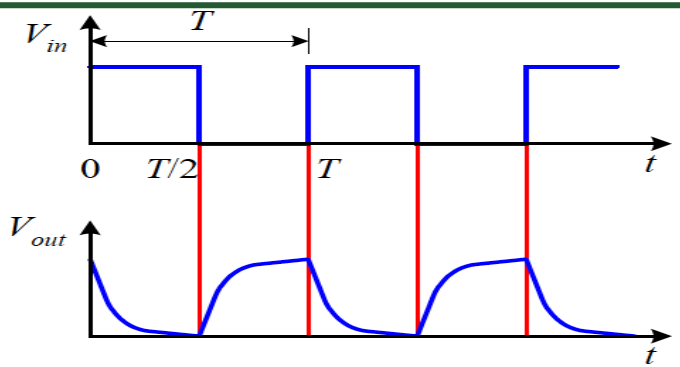
$$P_s = \sum I_{leakage} \times V_{DD}$$

$$P_{total} = P_s + P_d = P_s + P_{d(cd)} + P_{d(sc)}$$

DYNAMIC POWER DISSIPATION

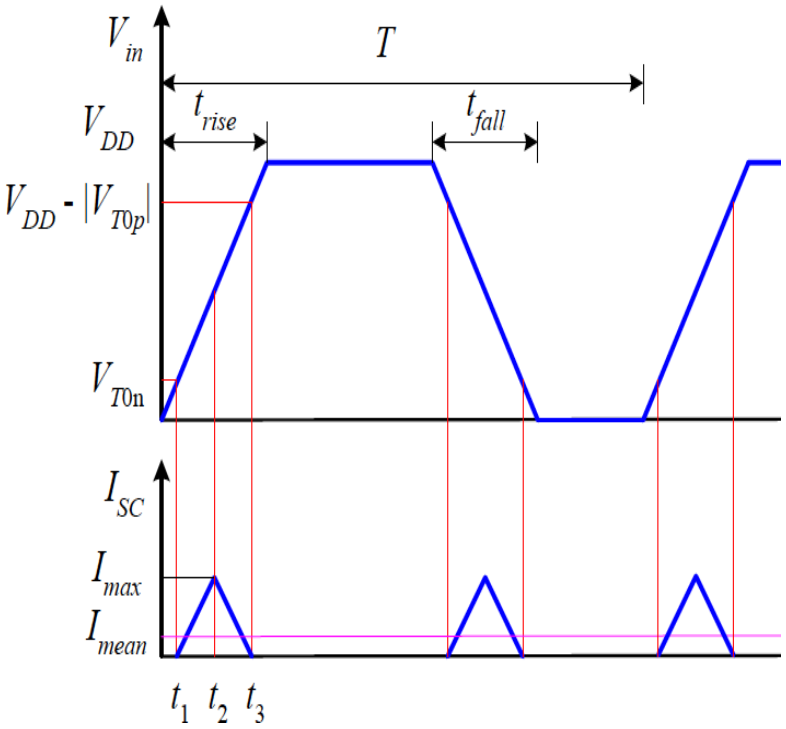






$$
\begin{aligned}
P_{d(cd)} &= \frac{1}{T}\int_0^{T/2} i_n(t)V_{out}dt + \frac{1}{T}\int_{T/2}^{T} i_p(t)(V_{DD}-V_{out})dt \\
&= \frac{C_L}{T}\int_0^{V_{DD}} V_{out}dV_{out} + \frac{C_L}{T}\int_0^{V_{DD}} (V_{DD}-V_{out})d(V_{DD}-V_{out}) \\
&= \frac{C_L}{T}V_{DD}^2 = C_L V_{DD}^2 f_p
\end{aligned}
$$

SHORT-CIRCUIT POWER DISSIPATION



$$I_{mean} = 2 \times \left[\frac{1}{T}\int_{t_1}^{t_2} I(t)dt + \frac{1}{T}\int_{t_2}^{t_3} I(t)dt\right]$$

If $V_{T0n} = |V_{T0p}| = V_T$ and $k_n = k_p = k$, then

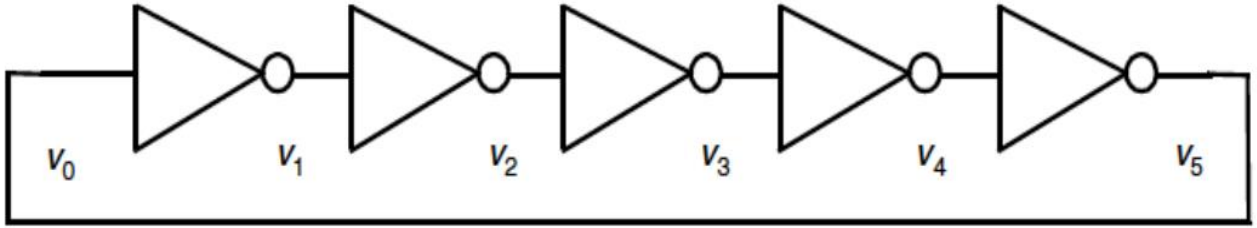
$$I_{mean} = 2 \times \left[\frac{2}{T}\int_{t_1}^{t_2} \frac{1}{2}k(V_{in} - V_T)^2 dt\right]$$

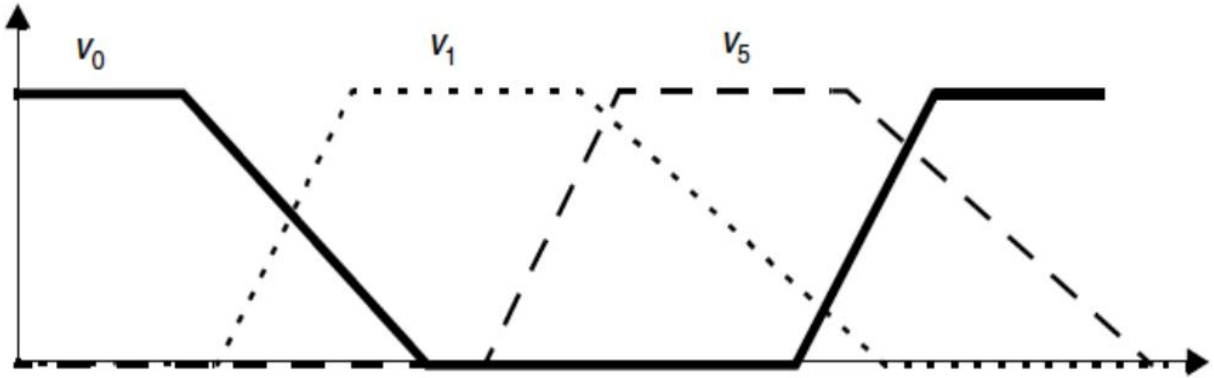
Using the facts: $V_{in} = (t/t_{rise})V_{DD}$

$t_1 = (V_T/V_{DD})t_{rise}$, and $t_2 = t_{rise}/2$,

$$\begin{aligned} P_{d(sc)} &= I_{mean} \times V_{DD} = \frac{2}{T}k \times \int_{t_1}^{t_2} \left(\frac{V_{DD}}{t_{rise}}t - V_T\right)^2 dt \cdot V_{DD} \\ &= \frac{2}{T}k \times \frac{1}{3}\frac{t_{rise}}{V_{DD}}V_{DD}\left[\left(\frac{V_{DD}}{t_{rise}}t - V_T\right)^3\right]_{t_1}^{t_2} \\ &= \frac{k}{12}(V_{DD} - 2V_T)^3 \frac{t_{rise}}{T} \end{aligned}$$

PROPAGATION DELAY



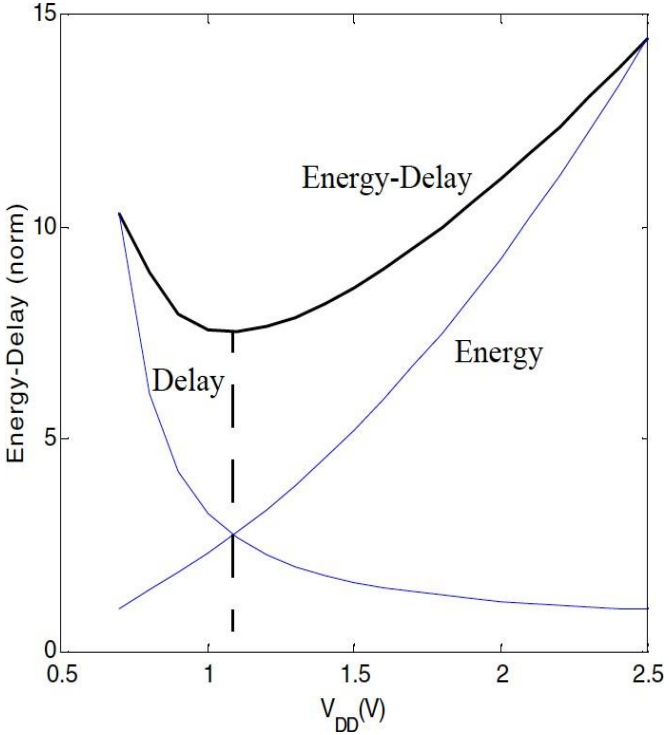


$$T = 2 \times t_p \times N$$

POWER-DELAY PRODUCT AND ENERGY-DELAY PRODUCT

$$PDP = P_{avg} \cdot t_{pd} = CV_{DD}^2 f \cdot \frac{1}{2f} = \frac{1}{2}CV_{DD}^2$$

$$\begin{aligned} EDP &= PDP \cdot t_{pd} = \frac{1}{2}CV_{DD}^2 \cdot \frac{C\Delta V}{I_{sat}} \\ &= \frac{1}{2}CV_{DD}^2 \cdot \frac{CV_{DD}}{\frac{1}{2}k\left(\frac{W}{L}\right)(V_{GS}-V_T)^2} \\ &= \frac{C^2V_{DD}^3}{k\left(\frac{W}{L}\right)(V_{GS}-V_T)^2} \end{aligned}$$

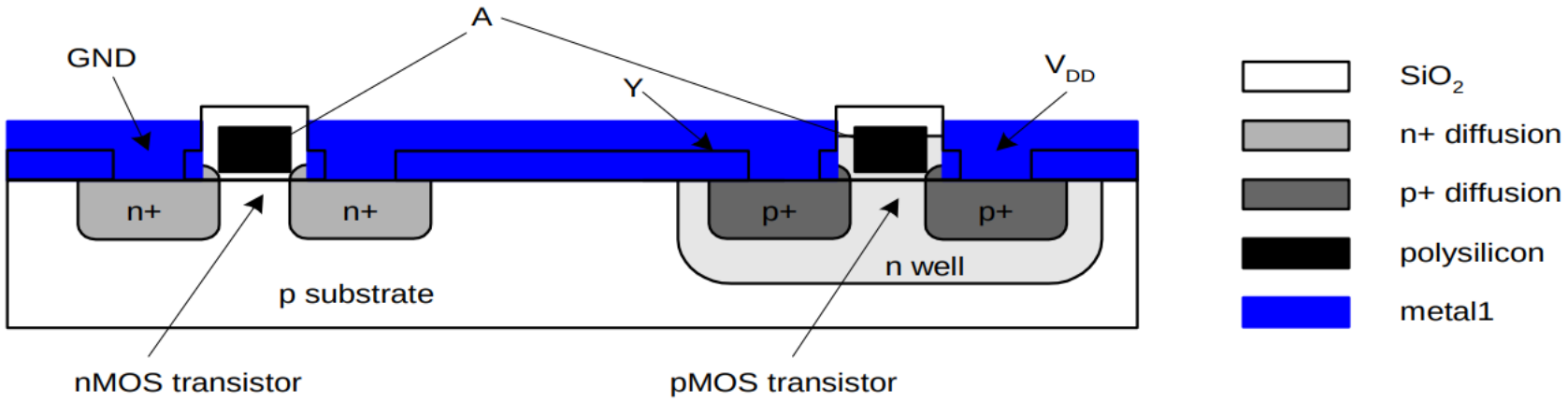


CMOS FABRICATION

- CMOS transistors are fabricated on silicon wafer
- Lithography process similar to printing press
- On each step, different materials are deposited or etched
- Easiest to understand by viewing both top and cross-section of wafer in a simplified manufacturing process

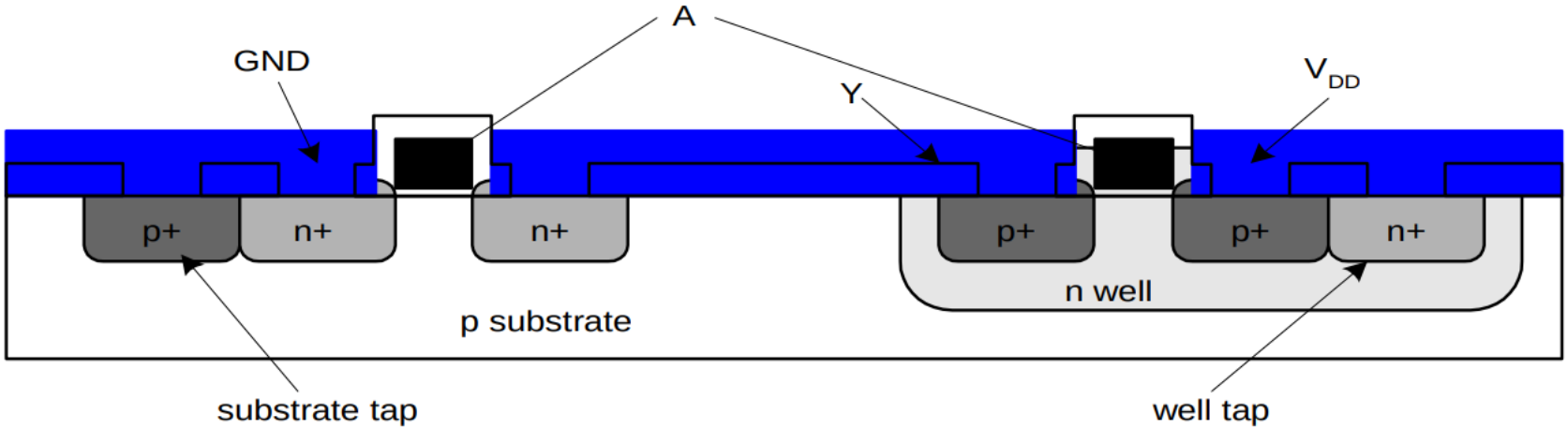
INVERTER CROSS-SECTION

- Typically use p-type substrate for NMOS transistors
- Requires n-well for body of PMOS transistors



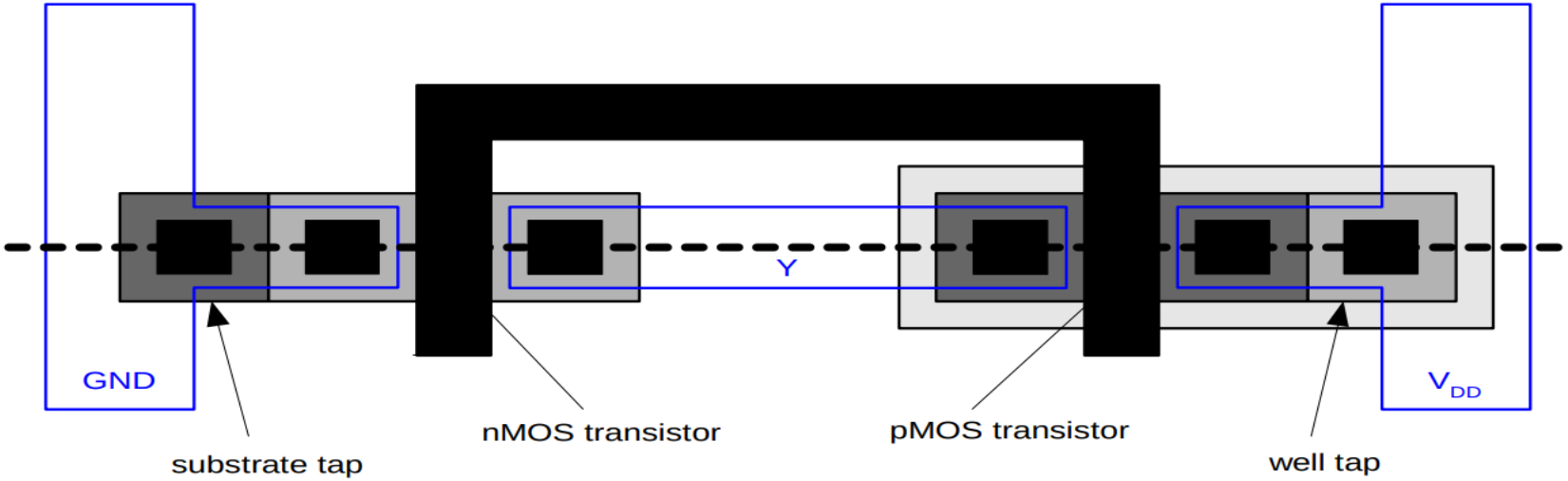
WELL AND SUBSTATE TAPS

- Substrate must be tied to GND and n-well to VDD
- Metal to lightly-doped semiconductor forms poor connection called Schottky Diode
- Use heavily doped well and substrate contacts / taps



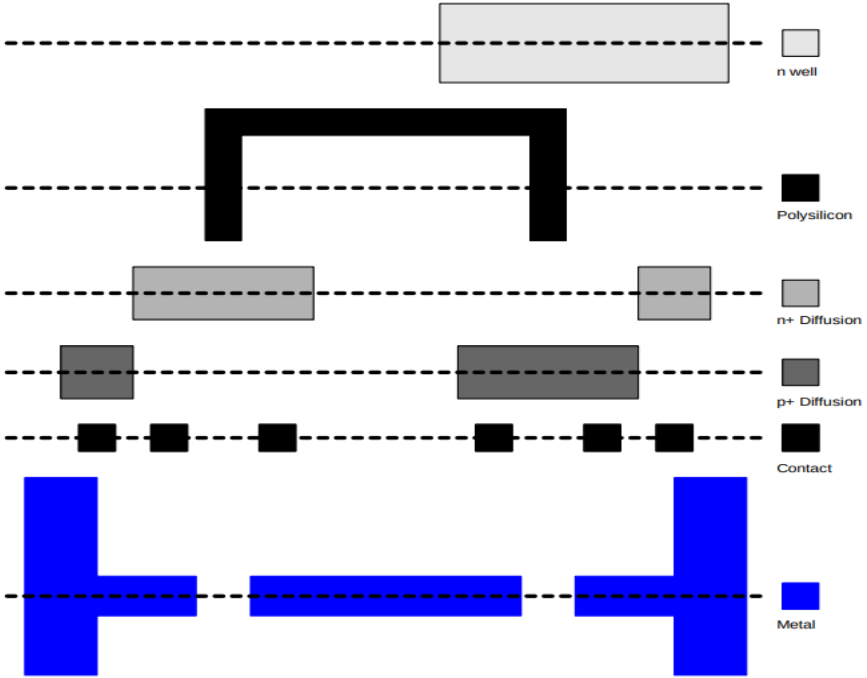
INVERTER MASK SETS

- Transistors and wires are defined by masks
- Cross-section taken along dashed line



DETAILED MASK VIEWS

- Six masks (old process)
  - n-well
  - Polysilicon
  - n+ diffusion
  - p+ diffusion
  - Contact
  - Metal

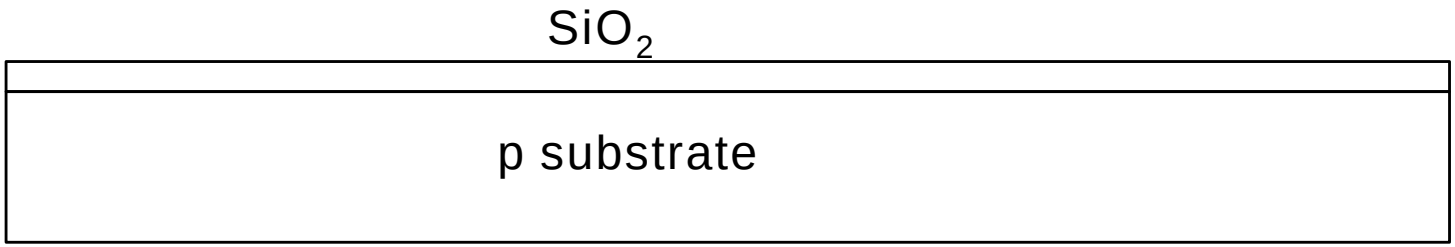


FABRICATION STEPS

- Start with blank wafer (cut from ingot of crystalline silicon)
- Build inverter from the bottom up
- First step will be to form the n-well
  - Cover wafer with protective layer of SiO_2 (oxide)
  - Remove layer where n-well should be built
  - Implant or diffuse n dopants into exposed wafer
  - Strip off SiO_2

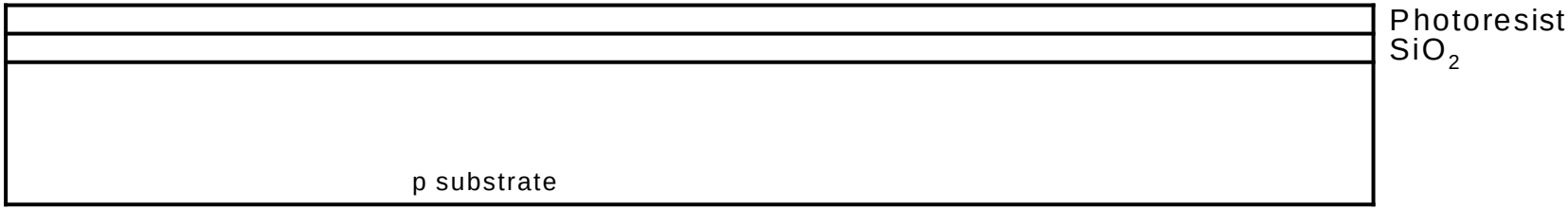
OXIDATION

- Grow SiO_2 on top of Si wafer
  - 900 – 1200 C with H_2O or O_2 in oxidation furnace



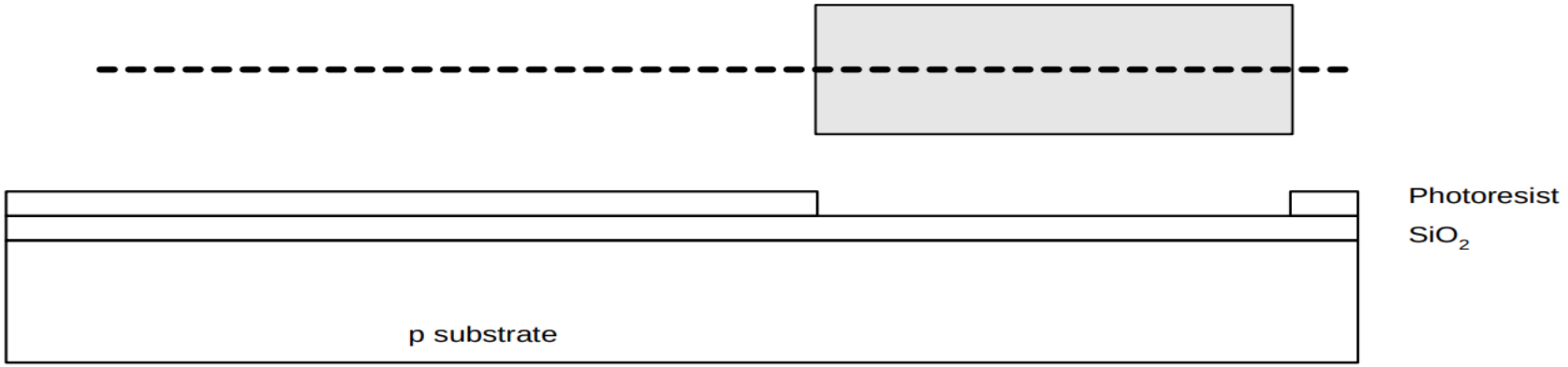
PHOTORESIST

- Spin on photoresist
  - Photoresist is a light-sensitive organic polymer
  - Softens where exposed to light



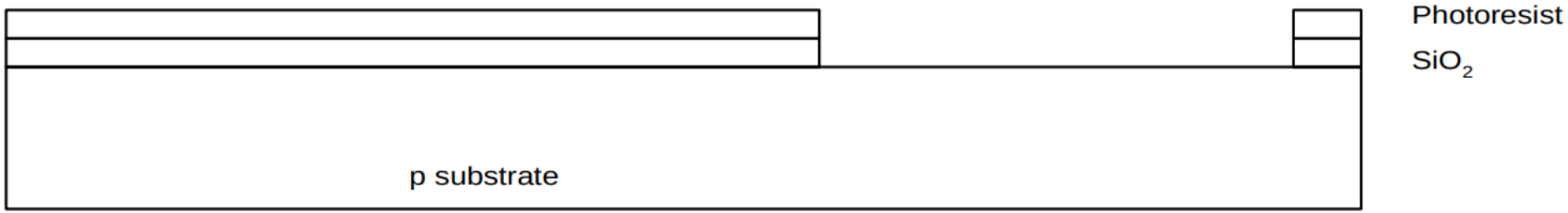
LITHOGRAPHY

- Expose photoresist through n-well mask
  - Older processes use visible light
  - For features below 14 nm, X-rays or EUV (extreme ultraviolet) used
- Strip off exposed photoresist



ETCH

- Etch oxide with something like hydrofluoric acid (HF)
  - Seeps through skin and eats bone; nasty stuff!!!
- Only attacks oxide where resist has been exposed

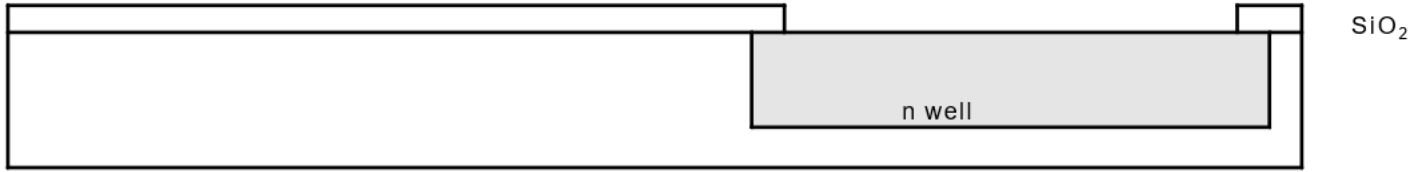


STRIP PHOTORESIST

- Strip off remaining photoresist
  - Use mixture of acids called piranha etch
- Necessary so resist doesn't melt in next step

N-WELL

- n-well is formed with diffusion or ion implantation
- Diffusion
  - Place wafer in furnace with arsenic gas
  - Heat until As atoms diffuse into exposed Si
- Ion implantation
  - Blast wafer with beam of As ions
  - Ions blocked by SiO_2, only enter exposed Si



STRIP OXIDE

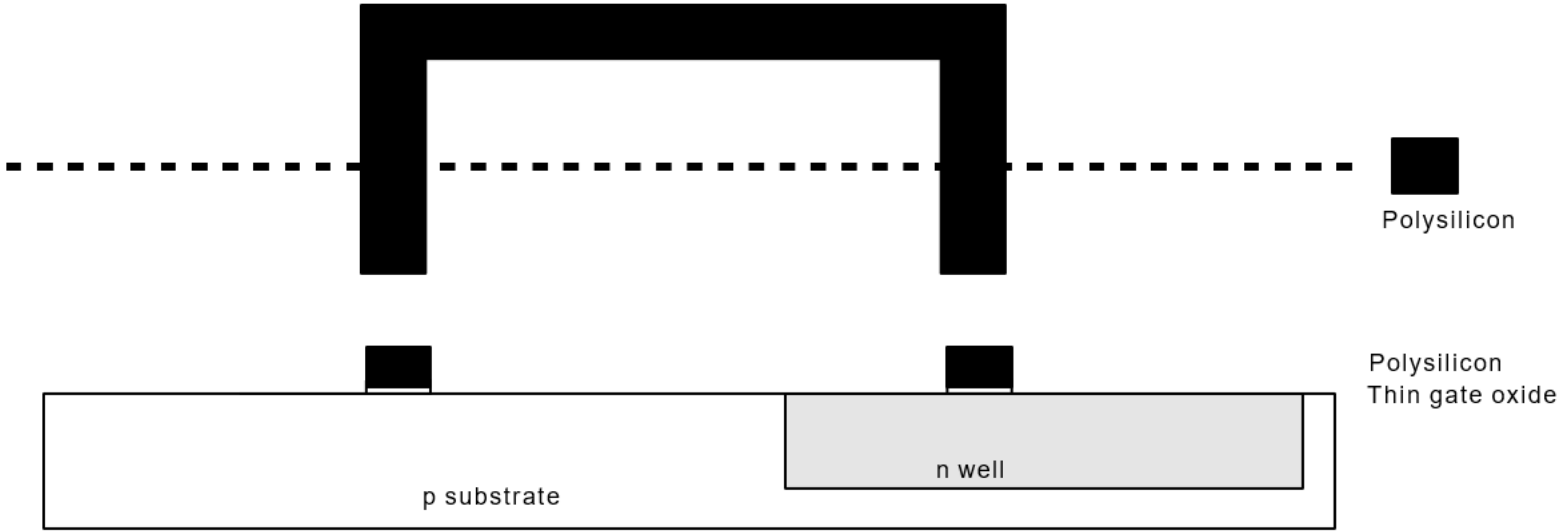
- Strip off the remaining oxide using something like HF
- Back to bare wafer with n-well
- Subsequent steps involve similar series of steps

POLYSILICON (OR OTHER METAL) GATE

- Deposit very thin layer of gate oxide
  - < 20 Å (6-7 atomic layers) for silicon dioxide
  - Thicker for hafnium oxide (high-k dielectric)
- Chemical Vapor Deposition (CVD) of silicon layer
  - Place wafer in furnace with Silane gas (SiH_4)
  - Forms many small crystals called polysilicon
  - Heavily doped to be good conductor

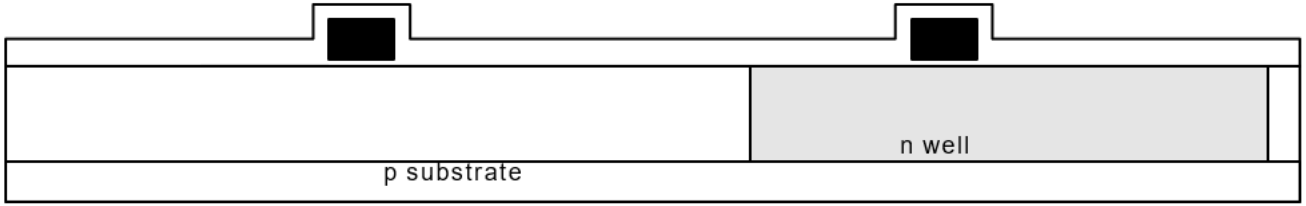
POLYSILICON PATTERNING

- Use same lithography process to pattern polysilicon



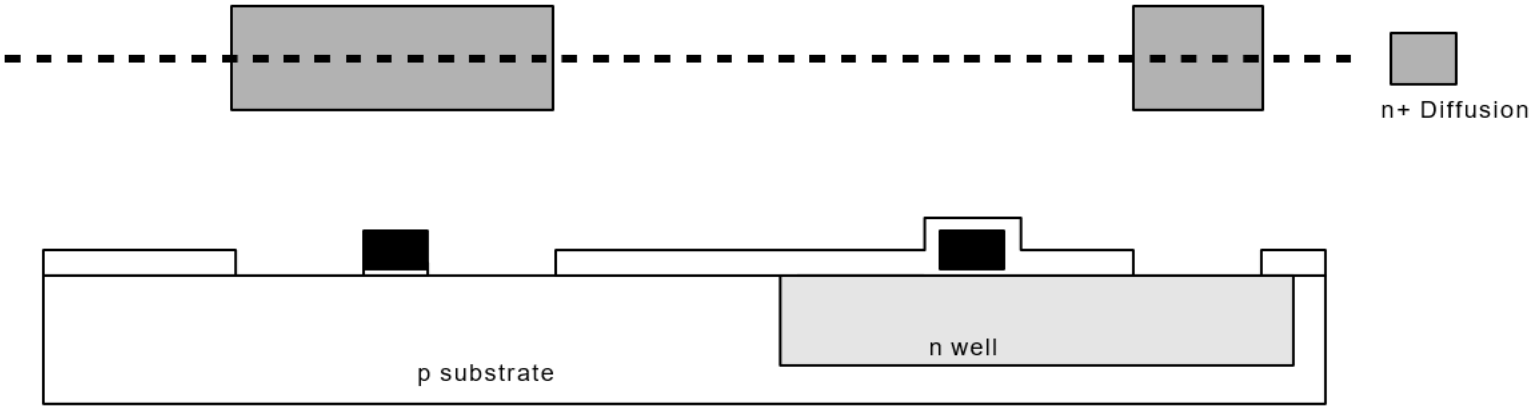
SELF-ALIGNED PROCESS

- Use oxide and masking to expose where n+ dopants should be diffused or implanted
- N-diffusion forms NMOS source, drain, and n-well contact
- Polysilicon gate structures will be used as “self-aligned” mask for drain/source



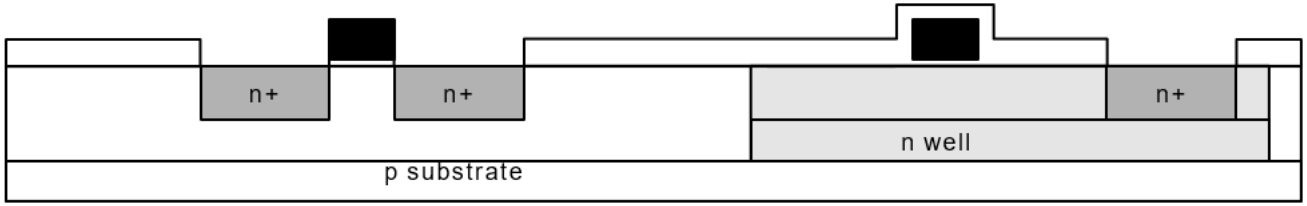
N-DIFFUSION (SOURCE, DRAIN AND N-WELL CONTACTS)

- Pattern oxide and form n+ regions
- Self-aligned process where gate blocks diffusion
- Polysilicon is better than metal for self-aligned gates because it doesn't melt during later processing



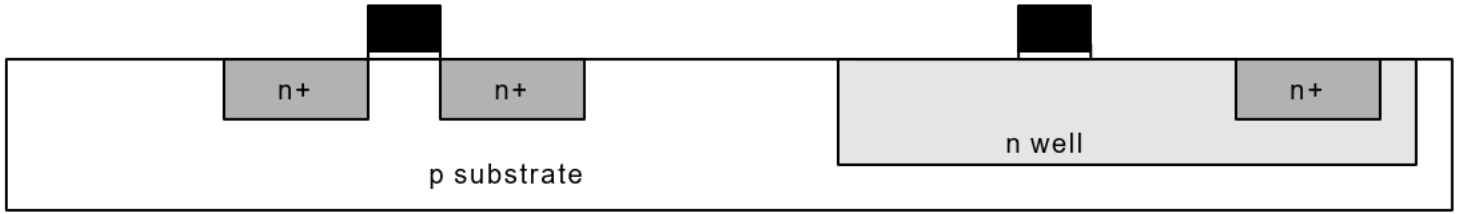
N-DIFFUSION CONTINUED

- Historically, dopants were diffused
- More often, ion implantation is used today
- But regions (source, drain, etc.) are still called “diffusion”



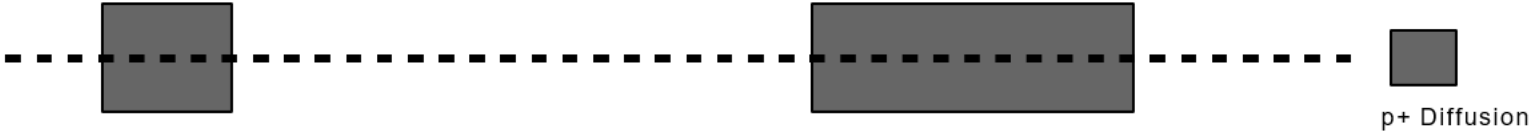
N-DIFFUSION CONTINUED

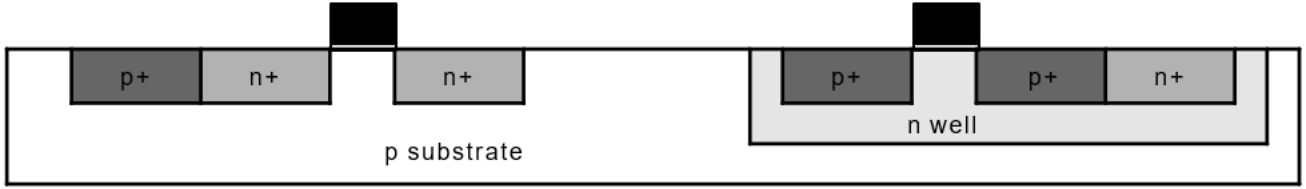
- Strip off oxide to complete patterning step



P-DIFFUSION

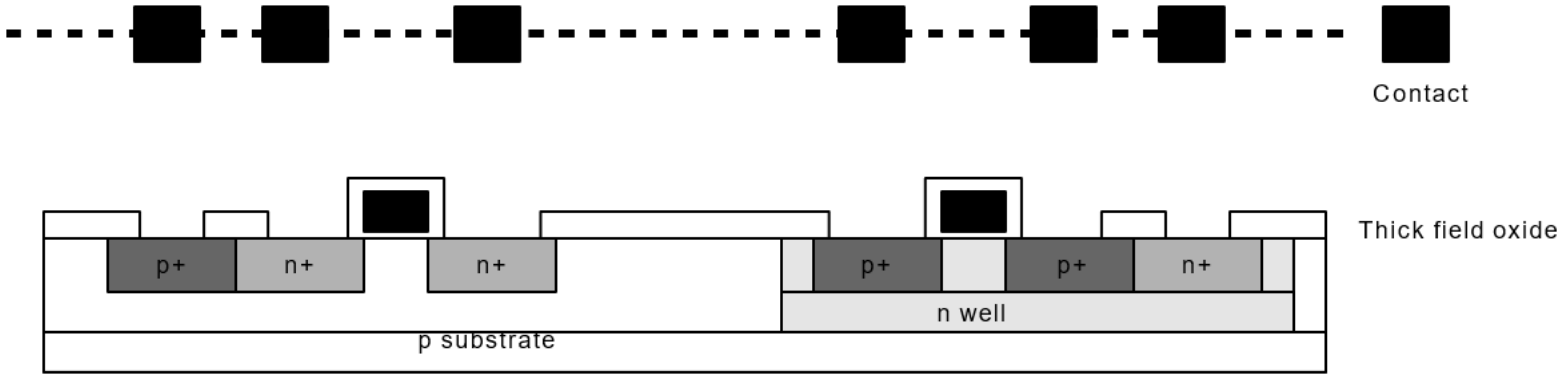
- Similar set of steps form p+ diffusion regions for PMOS source and drain and substrate contact





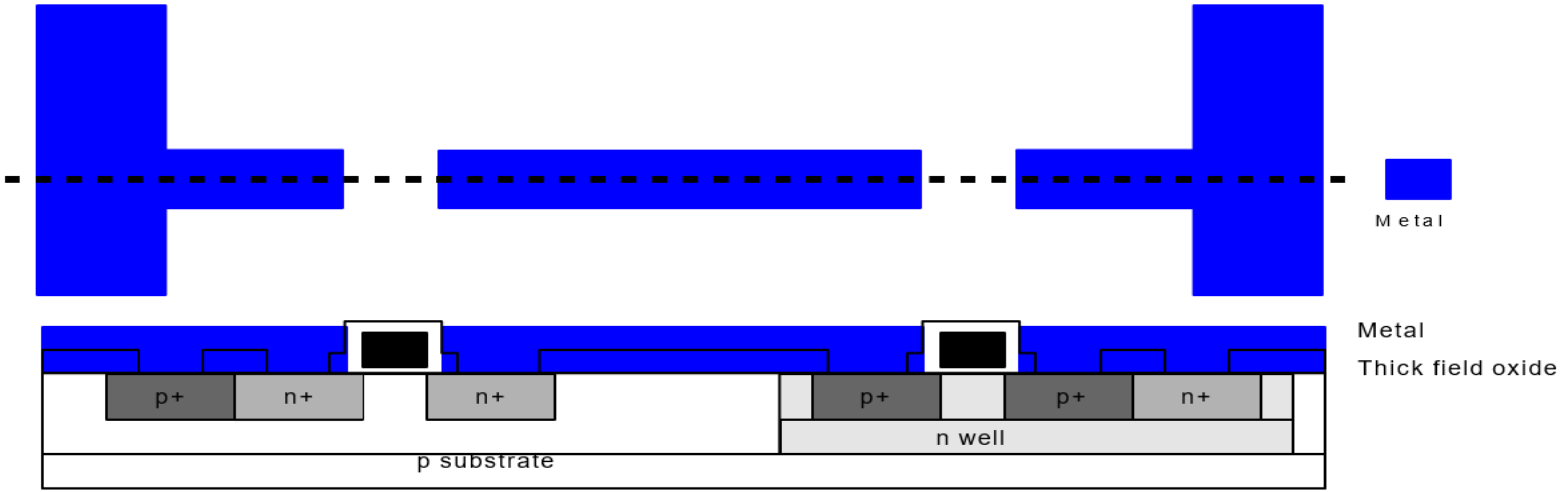
CONTACTS

- Now we need to wire together the devices
- Cover chip with thick field oxide
- Etch oxide where contact cuts are needed



METALIZATION

- Sputter on aluminum over whole wafer (may be copper, depending on process)
- Pattern to remove excess metal, leaving wires

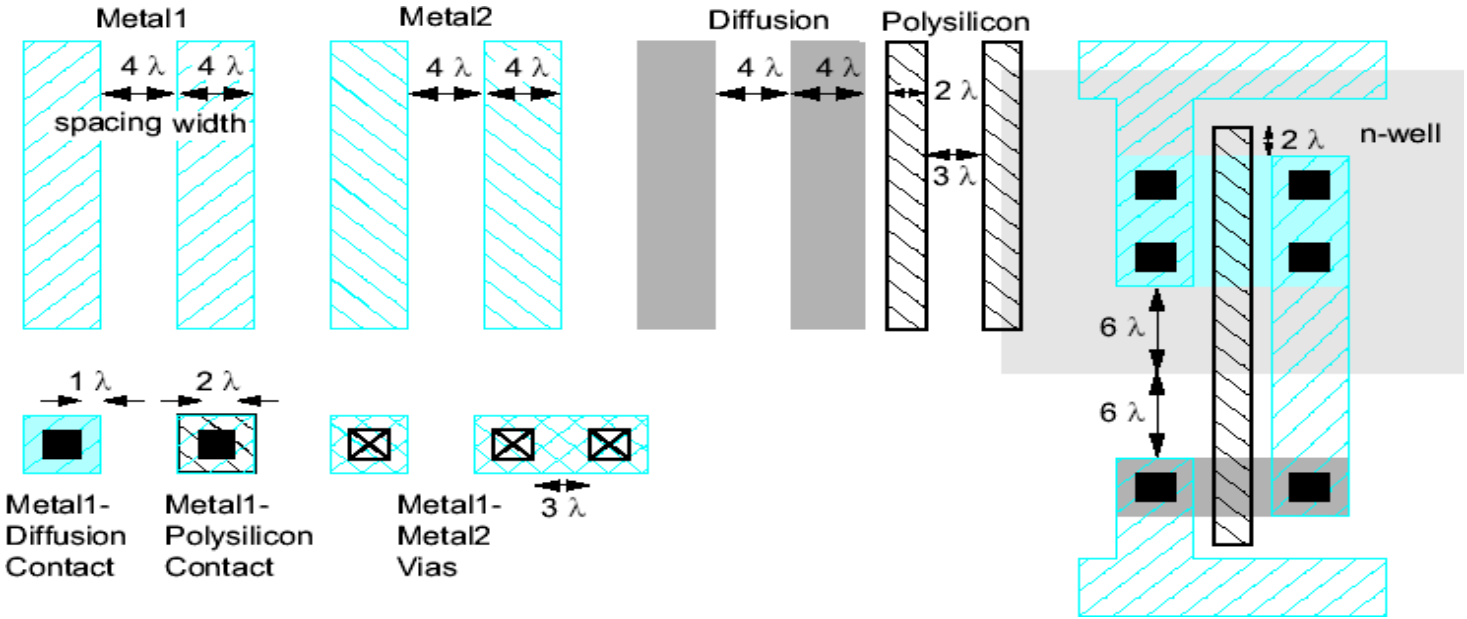


LAYOUT – CUSTOM PHYSICAL DESIGN

- Chips are specified with set of masks
- Minimum dimensions of masks determine transistor size (and hence speed, cost, and power)
- Feature size f = distance between source and drain (gate length)
  - Set by minimum width of polysilicon
  - A little different for FinFETs (fin width)
- Historically, feature size has improved 30% every 3 years or so
- Normalize for feature size when describing design rules

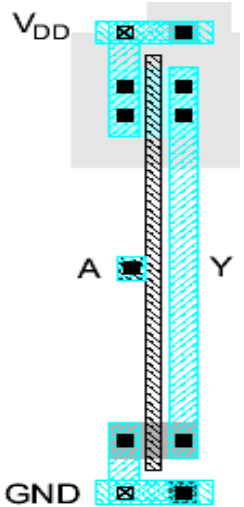
SIMPLIFIED DESIGN RULES FOR LAYOUT

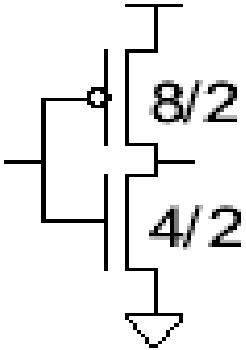
- Conservative rules are useful to get you started
- Older process use λ rules, where λ = f/2

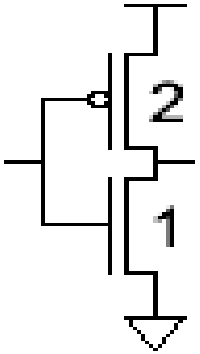


INVERTER LAYOUT

- Transistor dimensions specified as Width / Length
  - Minimum size is 4λ / 2λ, sometimes called 1 unit
  - In *f* = 0.6 µm process (old!), this is 1.2 µm wide, 0.6 µm long







COST ESTIMATION

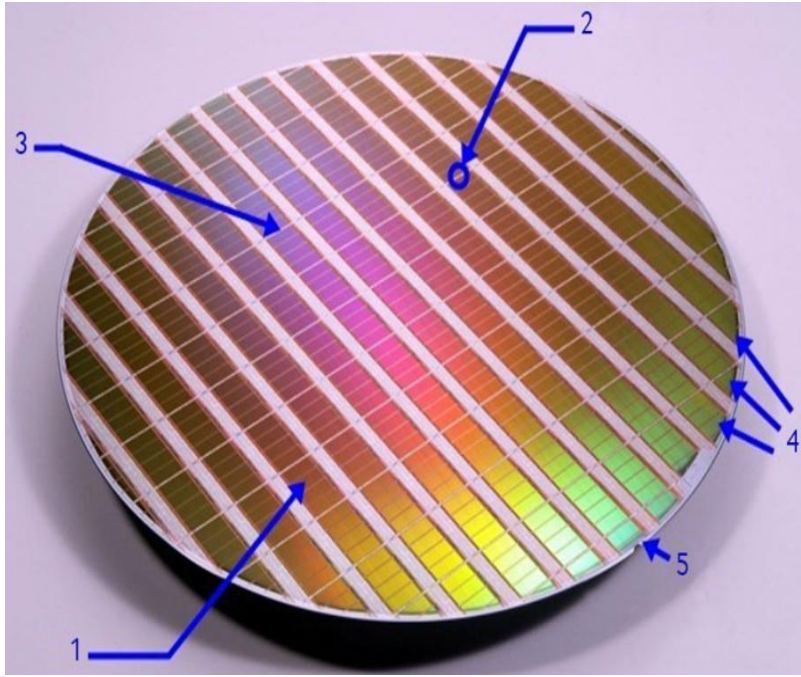
1.Chip: a tiny piece of silicon with electronic circuit patterns

2.Scribe Lines: thin, non-functional spaces between the functional pieces, where a saw can safely cut the wafer without damaging the circuits

3.TEG (Test Element Group): a prototype pattern that reveals the actual physical characteristics of a chip (transistors, capacitors, resistors, diodes and circuits) so that it can be tested to see whether it works properly

4.Edge Die: dies (chips) around the edge of a wafer considered production loss; larger wafers would relatively have less chip loss

5. Flat Zone: one edge of a wafer that is cut off flat to help identify the wafer's orientation and type



Source: https://news.samsung.com/global/eight-major-steps-to-semiconductor-fabrication-part-1-creating-the-wafer

COST ESTIMATION

$$\text{Cost per IC} = \text{Variable cost of IC} + \frac{\text{Fixed cost}}{\text{Volume}}$$

$$\text{Variable cost of IC} = \frac{\text{Cost of die} + \text{Cost of testing die} + \text{Cost of packaging and final test}}{\text{Final test yield} \times \text{Dies per wafer}}$$

The number of dies in a wafer, excluding fragmented dies on the boundary, can be approximated by:

$$\text{Cost of die} = \frac{\text{Wafer price}}{\text{Dies per wafer} \times \text{Die yield}}$$

$$\text{Dies per wafer} = \frac{3}{4}\frac{d^2}{A} - \frac{1}{2\sqrt{A}}d$$

$$\text{Die yield} = \left(1 + \frac{D_0 A}{\alpha}\right)^{-\alpha}$$

d is the diameter of the wafer A is the area of square dies.

D_0 is the defect density, i.e., the defects per unit area (in defects/cm^2); and α is a measure of manufacturing complexity. The typical values of D_0 and α are 0.3 to 1.3 and 4.0, respectively.

COST ESTIMATION

- The fixed cost, also referred to as the nonrecurring engineering (NRE) cost, is independent of the sales volume. It is mainly contributed by the cost from that a project is started until the first successful prototype is obtained.

- More precisely, the fixed cost covers direct and indirect costs. The direct cost includes the research and design (R&D) cost, manufacturing mask cost, as well as marketing and sales cost; the indirect cost comprises the investment of manufacturing equipment, the investment of CAD tools, building infrastructure cost, and so on.

- The variable cost is proportional to the product volume and is mainly the cost of manufacturing wafers, namely, wafer price, which is roughly in the range between 1,200 and 1,600 USD for a 300-mm wafer.

Thank you!