Circuit Characterization and Performance Estimation

(Teaching Material is from chapter4)

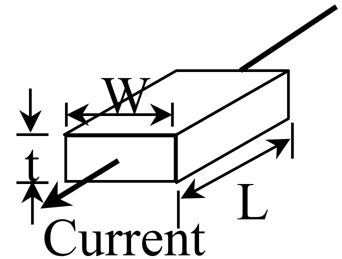
- Resistance Estimation
- Capacitance Estimation
- Switching Characteristics
- Inverter-pair delay
- Driving Large Capacitive Loads
- Dynamic Power Dissipation
- Scaling of MOS Transistor Dimensions

Resistance Estimation

- Sheet Resistance

$$R = \frac{\rho}{t}\left(\frac{L}{W}\right) = R_{\square}\left(\frac{L}{W}\right)$$

$$R_{\square} = \frac{\rho}{t}$$



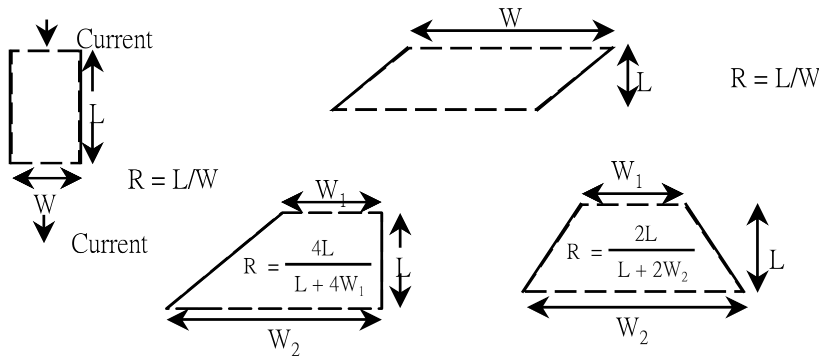
ρ : resistivity
t : thickness
L : conductor length
W: conductor width
$R_{\square}$: sheet resistance(ohm/square , Ω/□)

- Typical sheet resistance for conductors

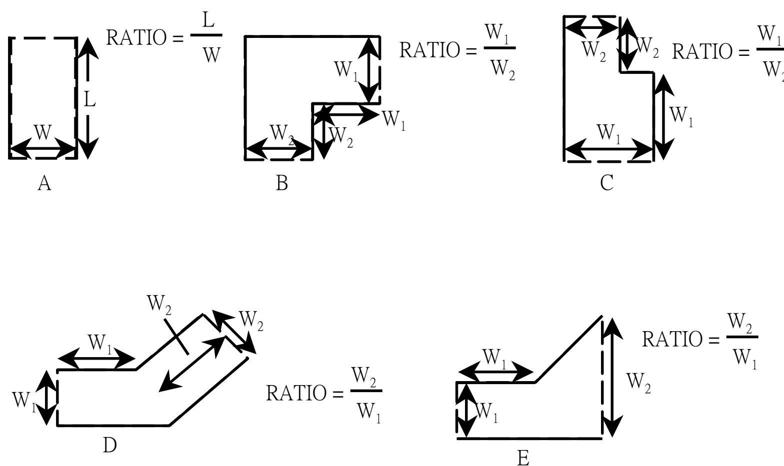
Material	Min	Typical	Max.
Intermetal (metal1-metal2)	0.05	0.07	0.1
Top-metal(metal3)	0.03	0.04	0.05
Polysilicon	15	20	30
Silicide	2	3	6
Diffusion(n^+, p^+)	10	25	100
Silicided diffusion	2	4	10
n-well	1K	2K	5K

Resistance Estimation of Nonrectangular Shapes

- Direct estimation



- Table-assisted estimation



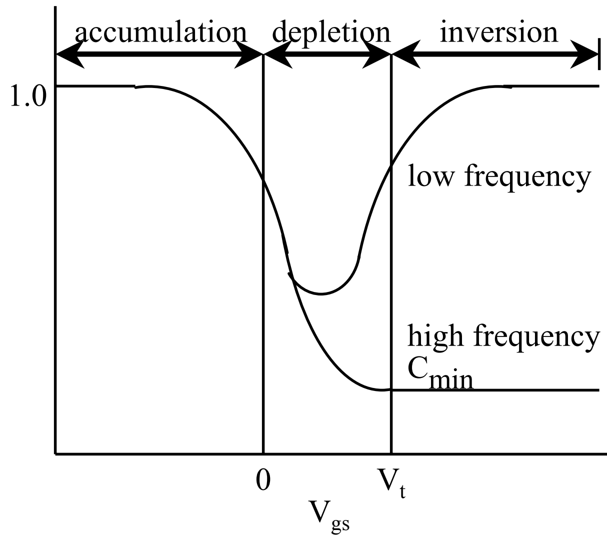
SHAPE	RATIO	RESIS-TANCE
A	1	1
A	5	5
B	5	5
B	1	2.5
B	2	2.55
B	3	2.6
C	1.5	2.1
C	2	2.25
C	3	2.5
C	4	2.65
D	1	2.2
D	1.5	2.3
D	2	2.3
D	3	2.6
E	1.5	1.45
E	2	1.8
E	3	2.3
E	4	2.65

Contact and Via Resistance

- propotional to the area of the contact, e.g. feature size↓=> Rcontact↑
- 0.25Ω ~ a few tens of Ωs
- Multiple contacts to obtain low-resistance interlayer connections

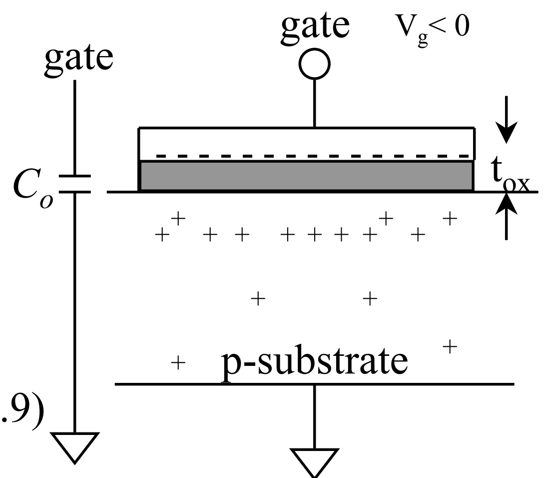
MOS－Capacitor Characteristics

- C －V plot



－ Three regions in the plot
(i) accumulation region
(ii) depletion region
(iii) inversion region

- Accumulation region



$$C_o = \frac{\varepsilon_{SiO2}\varepsilon_o}{t_{OX}} A = C_{OX} \cdot A$$

C_o ：gate capacitance

ε_{SiO_2}: dielectric constant of SiO_2(=3.9)

ε_o ：permittivity of free space

A ：gate area

$$C_{OX} = \frac{\varepsilon_{SiO2}\varepsilon_o}{t_{OX}}$$; gate capacitance per unit area

MOS – Capacitor Characteristics (Cont.)

- **Depletion region**

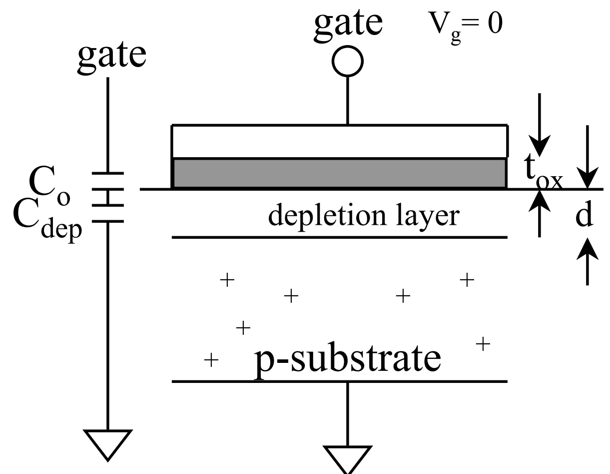
$C_{dep} = \dfrac{\varepsilon_{Si}\varepsilon_o}{d} A$

d ：depletion layer depth

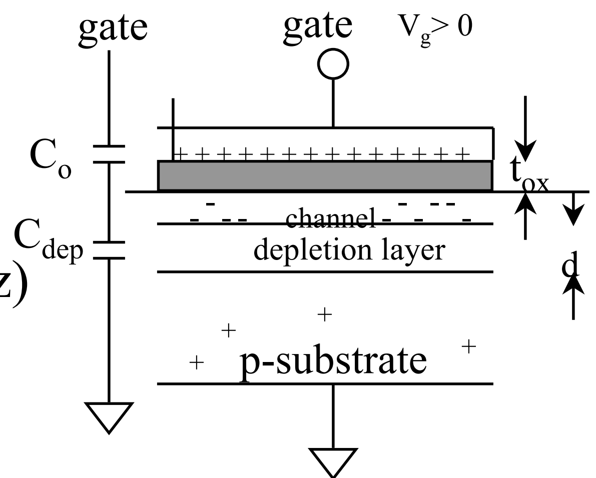
ε_{si} ：dielectric constant of Si(=12)

$C_{gb} = \dfrac{C_o C_{dep}}{C_o + C_{dep}}$

Where Co is low frequency capacitance between gate and surface



- **Inversion region**



$C_{gb} = \begin{cases} C_o \text{ ; static (i.e. low frequency,<100Hz)} \\ \dfrac{C_o C_{dep}}{C_o + C_{dep}} = C_{min} \text{ ; dynamic (i.e. high frequency)} \end{cases}$

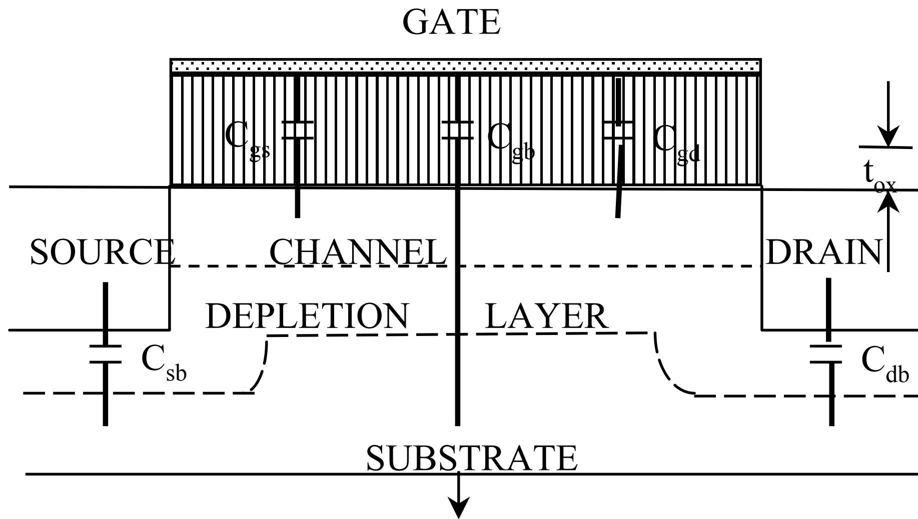
⚐ C_{min}

— C_{dep} depends on the depth of the depletion region, i.e. depends on substrate doping density.

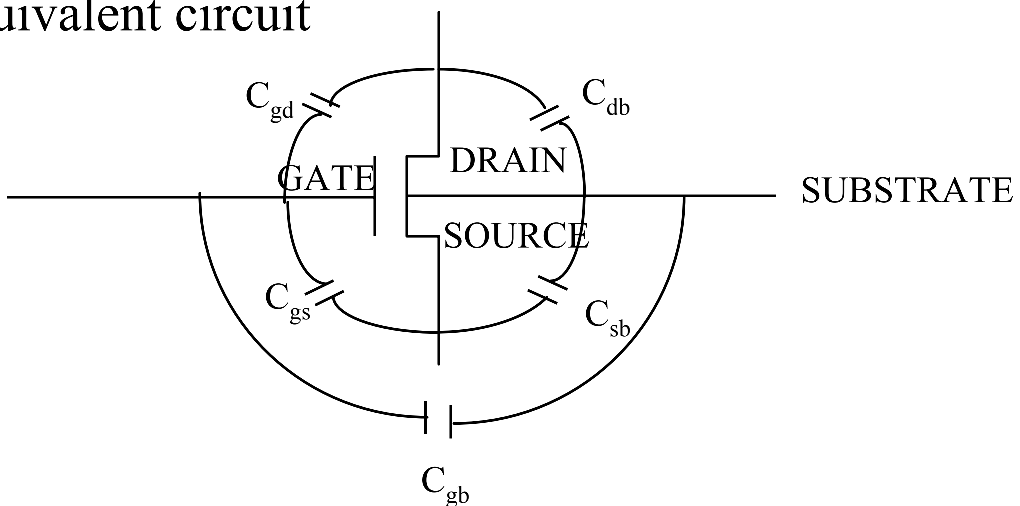
— For t_{ox}=100~200 ❹ ,C_{min}/C_o varies from 0.02~0.3 for substrate doping density varies from 1×10^{-14} cm^{-3} to 5×10^{-15} cm^{-3}

MOS Device Capacitances

- Cross section of MOS device



- Equivalent circuit



- Approximation of gate capacitance

— Self-aligned process is assumed (i.e. overlap caps. are negligible)

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)

C_{gs}, C_{gd}, and C_{ox}

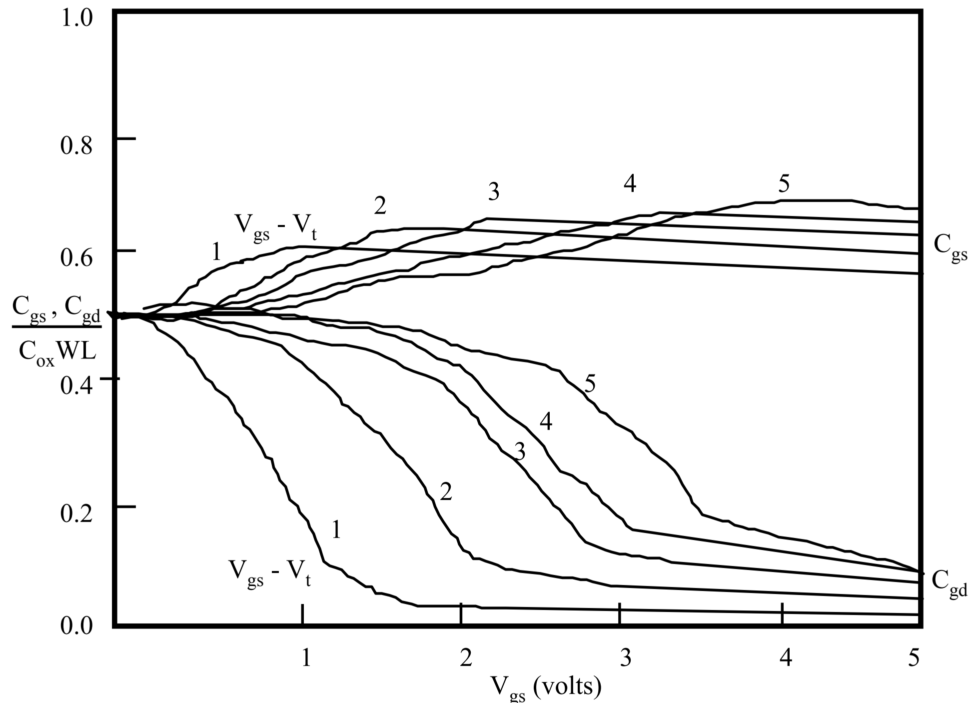
- Example 1: W=49.2μm, L=4.5μm (long channel)
  — C_{gs} and C_{gd}

large L

* large C_g & small C_{gd} (in saturation region)

$$\frac{C_{gd}}{C_g} \approx 0$$

(C_{gd} is due to channel side fringing fields between gate and drain.)



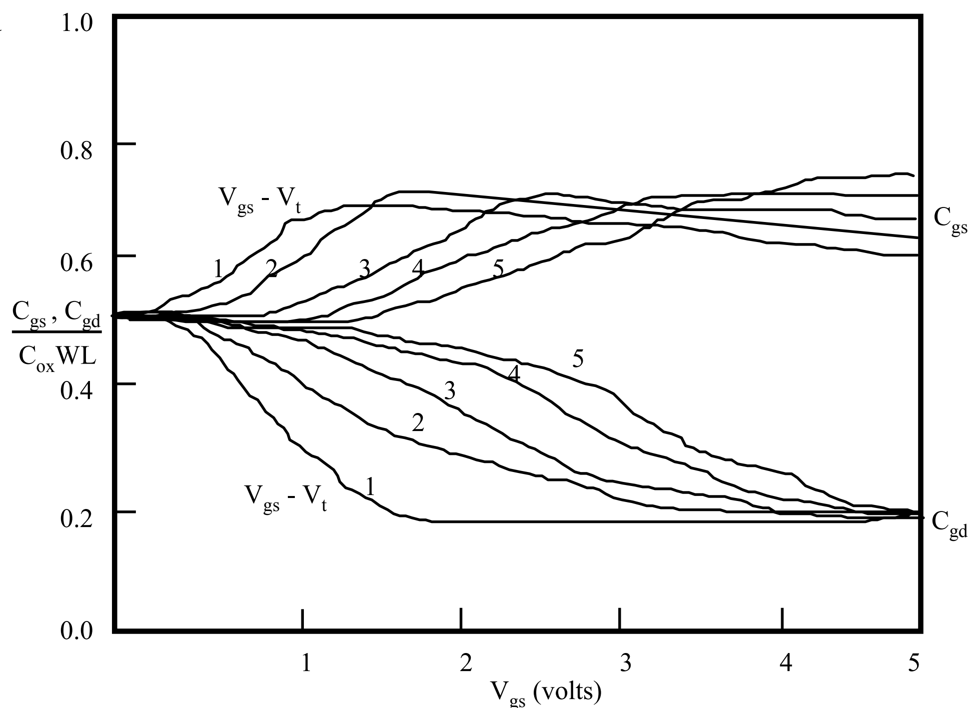
- Example 2: L=0.75μm (short channel)
  — C_{gs} and C_{gd}

small L

* small C_g & small C_{gd} (in saturation region)

$$\frac{C_{gd}}{C_g} \approx 0.2$$

(C_{gd} is due to channel side fringing fields between gate and drain.)

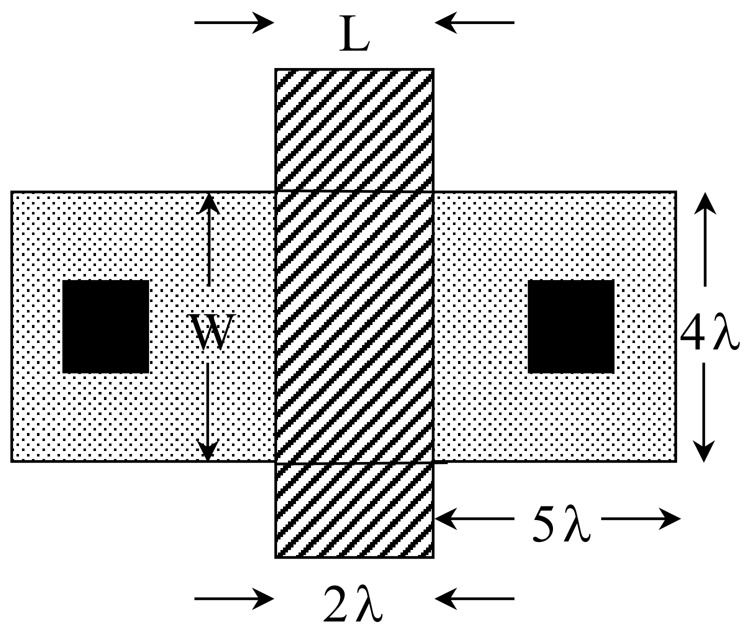


C_{ox}, gate capacitance per unit area

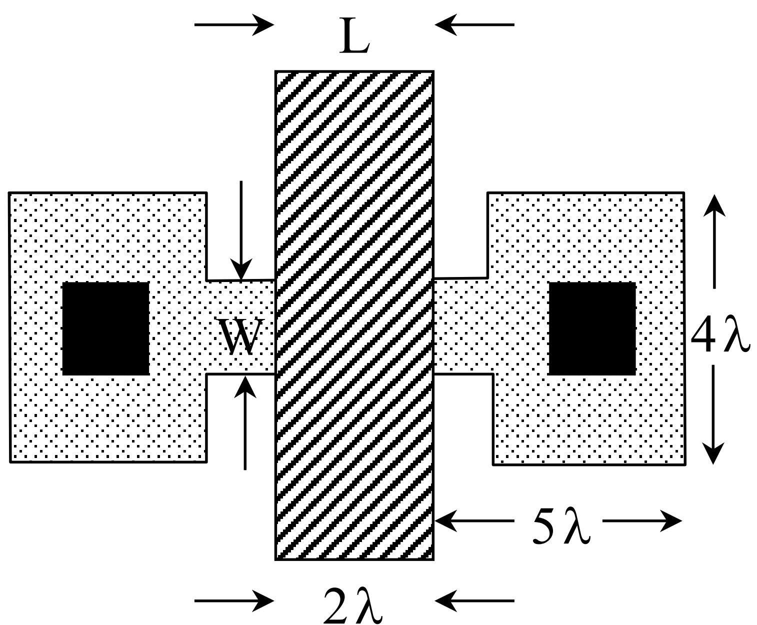
— $C_{ox} = \frac{\varepsilon_{sio_2} \ \varepsilon_0}{t_{ox}}$; where ε_{sio_2} =3.9 and ε_0=8.854×10^{-14}

— e.g. t_{ox}=350Å => C_{ox}≈1×10^{-3}pF/μm^2=1fF/μm^2

- Unit transistor
  - It is the same width as a metal-diffusion contact

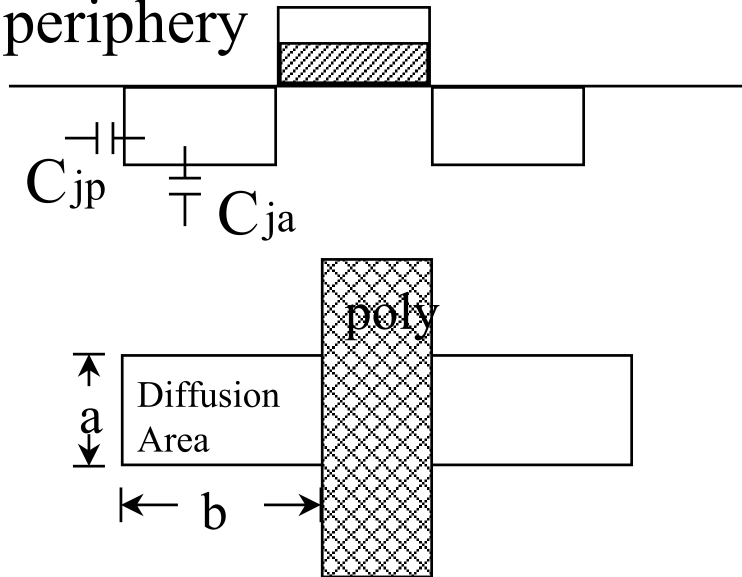


- minimum-size transistor



Diffusion Capacitance

- area and periphery



$$C_d = C_{ja} * (ab) + C_{jp} * (2a + 2b)$$

Cja: junction capacitance per μm^2

Cjp: periphery capacitance per μm

a: width of diffusion region

b: length of diffusion region

- Typical value(1μm n-well process)

$C_{ja} = 2 * 10^{-4}\ PF/\mu m^2$ (n^+ diffusion)

$5 * 10^{-4}\ PF/\mu m^2$ (p^+ diffusion)

$C_{jp} = 4 * 10^{-4}\ PF/\mu m^2$ (n^+ diffusion)

$4 * 10^{-4}\ PF/\mu m^2$ (p^+ diffusion)

- Voltage dependent

$$C_j(V_j) = C_{j0}\left(1 - \frac{V_j}{V_b}\right);$$

V_j is junction voltage

V_b is built-in junction potential~0.6V

C_{j0} is zero bias capacitance

m=0.3(graded junction)~

0.5(abrupt junction)

SPICE Modeling of MOS Capacitances

- SPICE example

```
.
M1 4 3 5 0 NFET W=4U L=1U AS=15P AD=15P PS=11.5U PD=11.5U
.
.
.MODEL NFET NMOS
+ TOX= 100E-8
+ CGBO=200P CGSO=600P CGDO=600P
+ CJ=200U CJSW=400P MJ=0.5 MJSW=0.3 PB=0.7
+ .....
.
.
```

— node4 - drain
node3 - gate
node5 - source
node0 - substrate

— channel width = 4 μm
channel length = 1 μm
TOX =100Å

— source area AS = 15 μm^2
drain area AD = 15 μm^2

— source periphery PS = 11.5 μm
drain periphery PD = 11.5 μm

— C_{gbo} occurs due to the polysilicon extension beyond the channel . (200×10^{-12} F/M)
C_{gso} and C_{gdo} represent the gate-to-source/drain capacitance due to overlap in the physical structure of the transistor . (600×10^{-12} F/M)

SPICE Modeling of MOS Capacitances (Cont.)

- Capacitance
  - gate capacitance

$Cg_{(intrinsic)} = W \cdot L \cdot C_{ox} = 4 \times 1 \times 35 \times 10^{-4}\,PF = 0.014PF$

$Cg_{(extrinsic)} = (W \cdot C_{gs0}) + (W \cdot C_{gd0}) + (2L \cdot C_{gb0})$

$= 4 \times 6 \times 10^{-4} + 4 \times 6 \times 10^{-4} + 2 \times \left(1 \times 2 \times 10^{-4}\right) PF$

$= 0.0052PF$

$Cg_{(total)} = Cg_{(intrinsic)} + Cg_{(extrinsic)} \approx 0.02PF$

  - source and drain capacitance

$$Cj = (Area \cdot Cj \cdot (1 + \frac{VJ}{PB})^{-MJ}) + (periphery \cdot CJSW \cdot (1 + \frac{VJ}{PB})^{-MJSW})$$

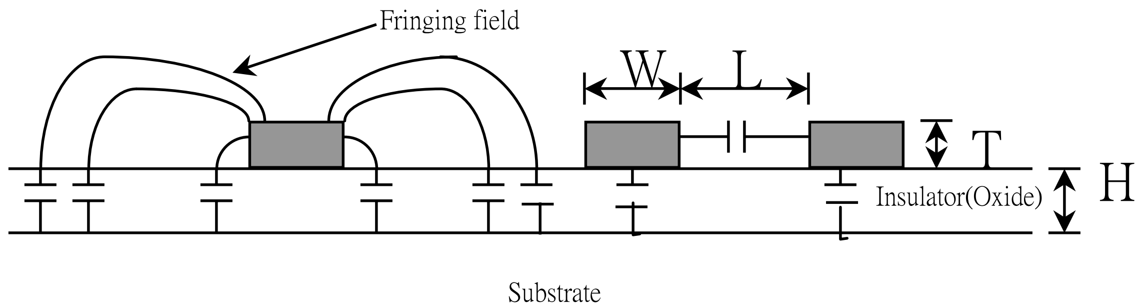
where

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 - 0.8 volts)
Area=*AS* or *AD*, the area of the source or drain
Periphery=*PS* or *PD*, the periphery of the source or drain

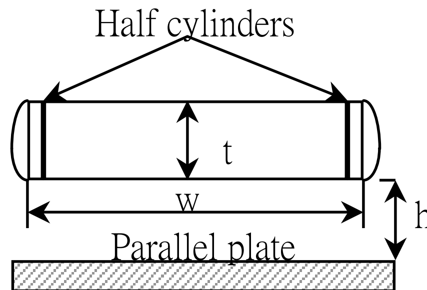
$Cj_{(drain)}$ =0.0043PF (VJ=2.5V is assumed)
$Cj_{(source)}$=0.0043PF (VJ=2.5V is assumed)

Routing Capacitance

- Single wire capacitance
  - —— parallel-plate effect and fring effect

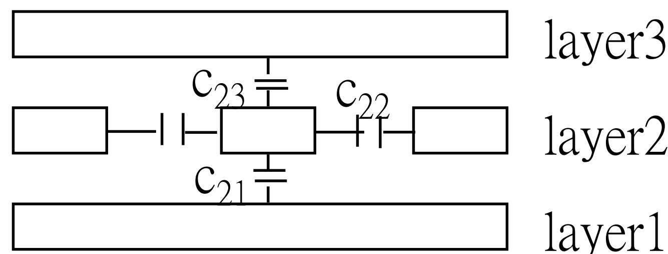


  - —— Accurate capacitance evaluation: use computer
  - —— Hand calculation: use simple model (less than 10% error)



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

- Multiple conductor capacitances
  - —— three-layer example



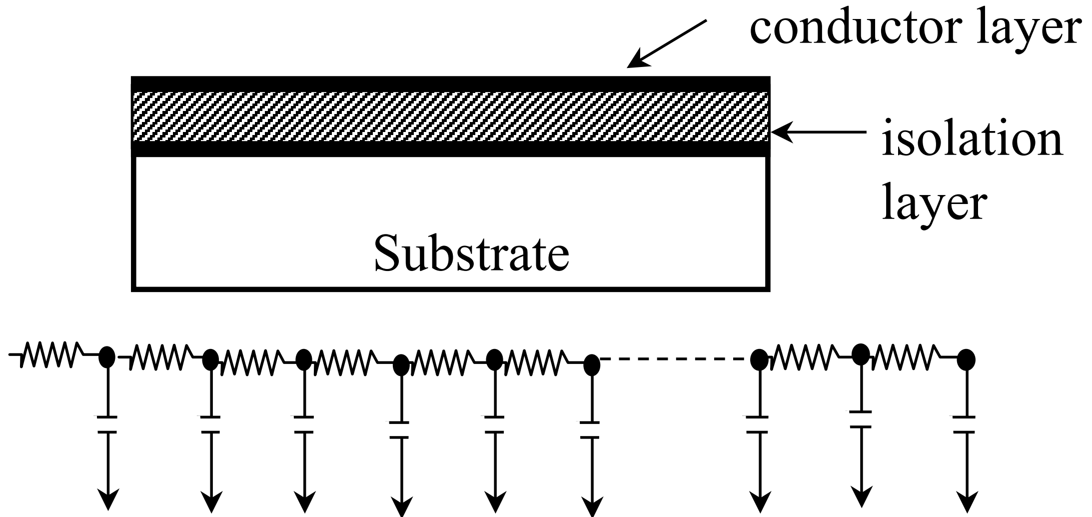
Capacitance calculation is very complex-- refer to textbook

  - —— Typical dielectric and conductor thicknesses

Thin-oxide	200Å	Metal1	6000Å
Field-oxide	6000Å	M_1-M_2oxide	6000Å
Polysilicon	3000Å	Metal2	12000Å
M_1-poly-oxide	6000Å	Passivation	20000Å

Distributed RC Effects

- Transmission line

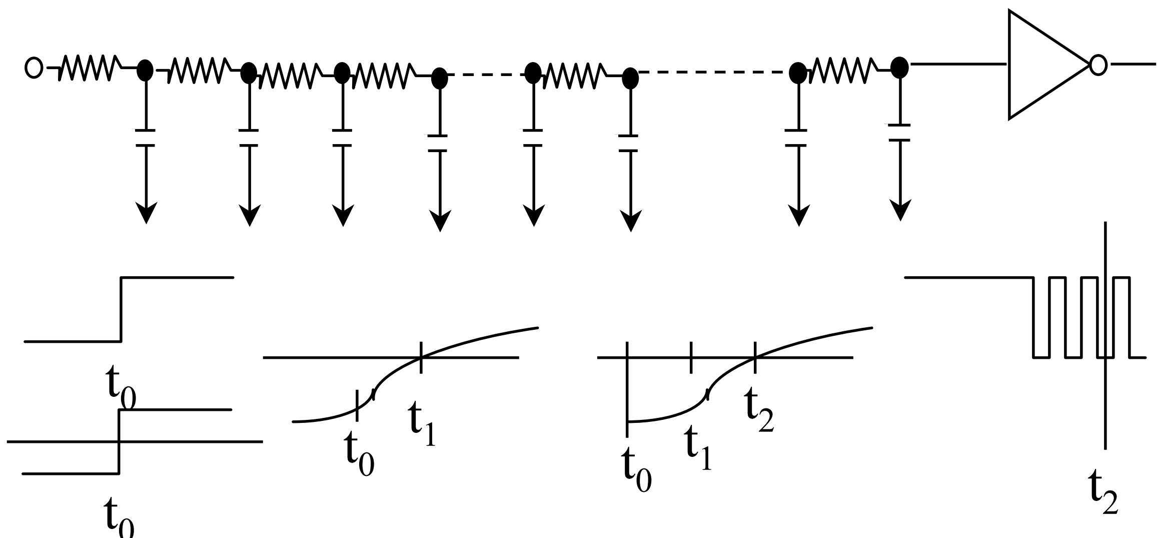


- delay time from one end to the other end

$$t \cong \frac{rc}{2} l^2$$

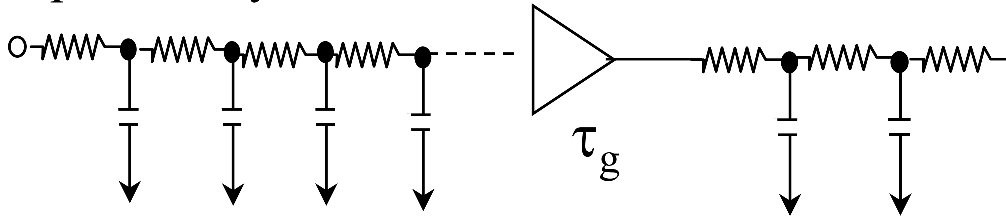
r: resistance per unit length
c: capacitance per unit length
l: length of the wire

- Disadvantages of long wire:
  (1) long long delay
  (2) reduction in sensitivity to noise



Distributed RC Effects (Cont.)

- Method to improve disadvantages mentioned previously



$$\text{delay} \approx \frac{r\,c}{2}\left(\frac{l}{2}\right)^2 + \tau_g + \frac{r\,c}{2}\left(\frac{l}{2}\right)^2$$

$$\approx \frac{rcl^2}{4} + \tau_g$$

— If $(rcl^2/4) < \tau_g$, delay time is reduced

— If $(rcl^2/2) >> \tau_g$, more buffers should be used

— In actual design, if possible,

$$\frac{rcl^2}{2} << \tau_g \quad ==> l << \sqrt{\frac{2\,\tau_g}{r\,c}}$$

- Transmission line effect is particularly severe in poly wire because of the relatively high resistance of this layer. Gate poly layer is the worst one because of its high capacitance to substrate.
- Strategies
  - use metal line: small r
  - use wider metal for signal distribution line (e.g. clock distribution line): small r, a tiny bit large C
- Design example

  refer to p.203 of textbook

Inductance

- On-chip inductance are normally small.
  Bond-wire inductance is larger.

- Inductance of bounding wires and the pins on packages

$$L = \frac{\mu}{2\pi} \ln\left(\frac{4h}{d}\right) \quad H/cm$$

μ: the magnetic permeability of the wire (typically 1.257´10^{-8}H/cm)
h: the height above the ground plane
d: the diameter of the wire

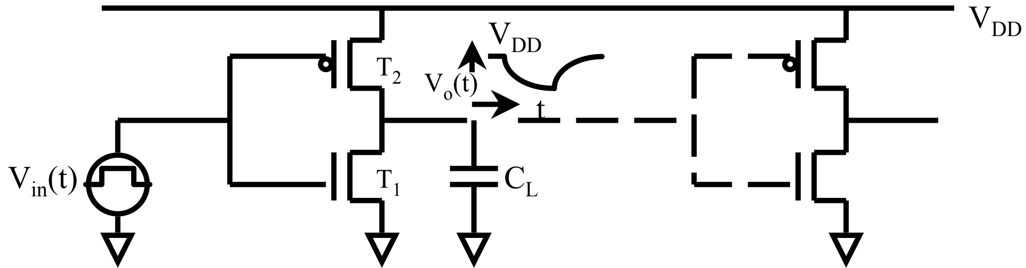
- Inductance of on-chip wires

$$L = \frac{\mu}{2\pi} \ln\left(\frac{8h}{w} + \frac{w}{4h}\right) \quad H/cm$$

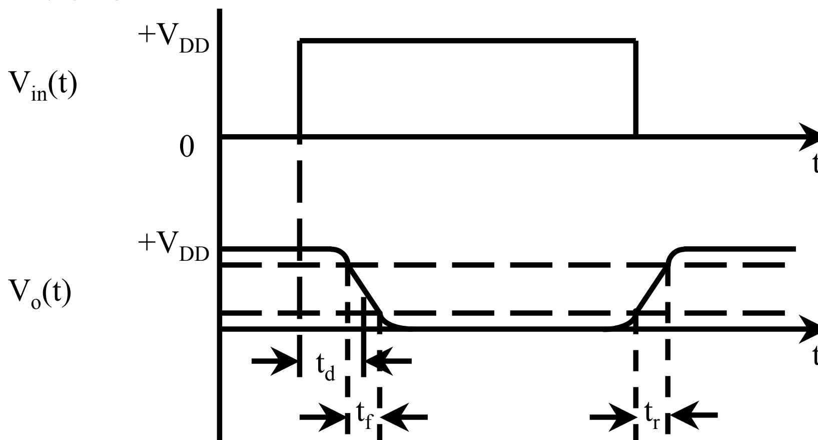
w: conductor width
h: the height above the substrate

Switching Characteristics

- CMOS inverter
  - Circuit

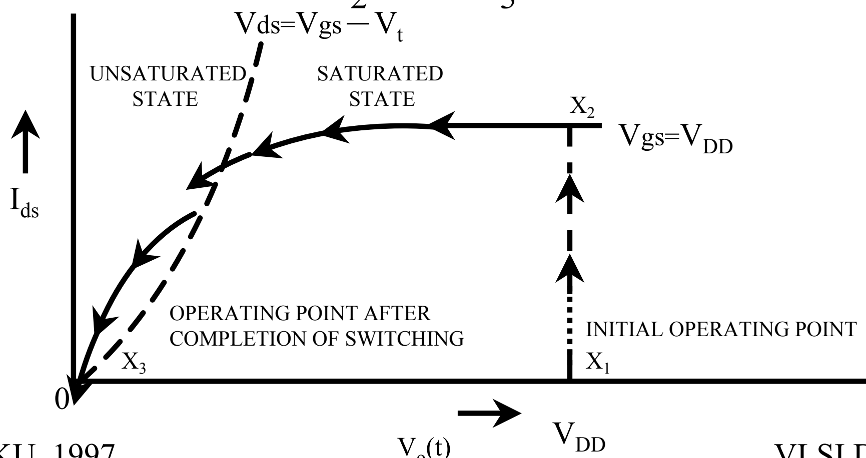


  - Waveform



  - 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

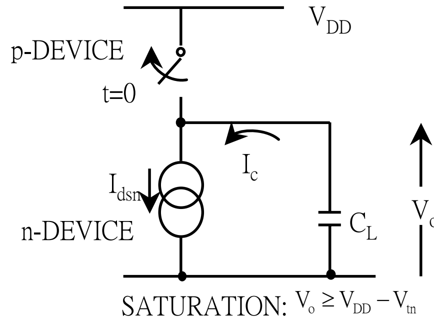
- Trajectory of n-transistor operating point during switching in CMOS inverter
  - Input transition : $X_1 \rightarrow X_2$
    Output transition: $X_2 \rightarrow X_3$

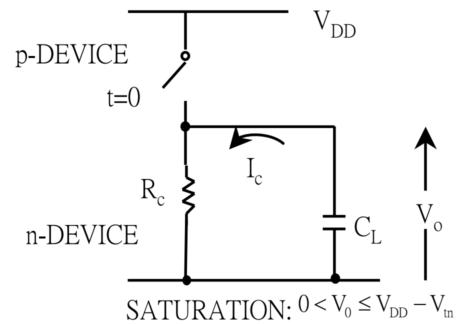


Fall Time and Rise Time

- Equivalent Circuit
- Fall

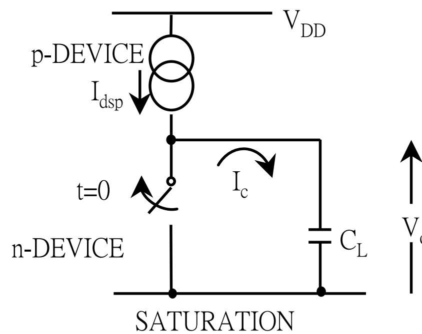
$\begin{pmatrix} V_{in} \uparrow \\ V_o \downarrow \end{pmatrix}$

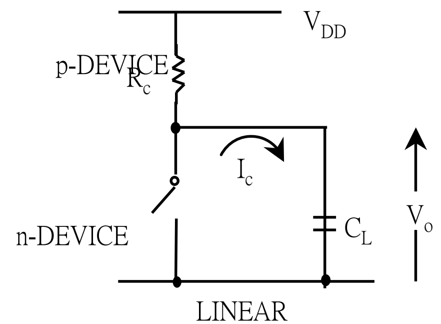




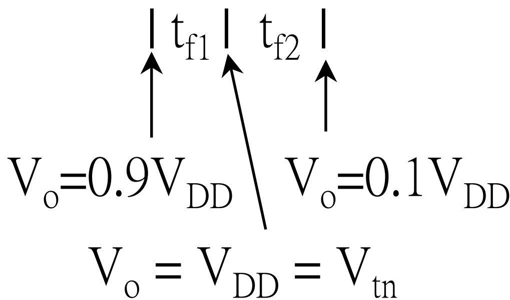
- Rise

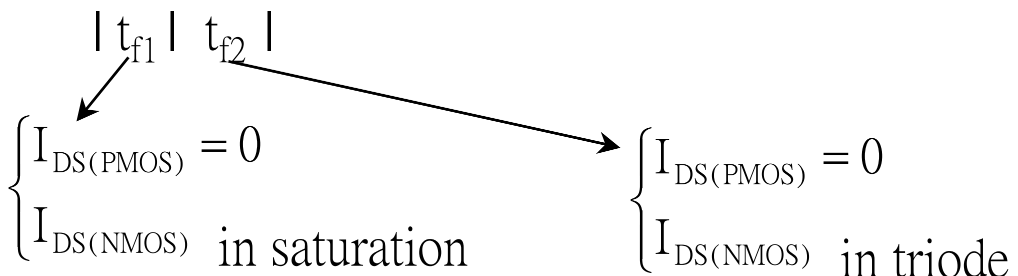
$\begin{pmatrix} V_{in} \downarrow \\ V_o \uparrow \end{pmatrix}$





- $t_f = t_{f1} + t_{f2}$; fall time





Fall Time and Rise Time (Cont.)

- t_{f1}
  - $V_o : 0.9\,V_{DD} \longrightarrow V_{DD} - V_{tn}$
  - NMOS in saturation region

$$C_L \frac{dV_o}{dt} - \frac{\beta_n}{2}(V_{DD} - V_{tn})^2 = 0$$

$$t_{f1} = \frac{2C_L}{\beta_n (V_{DD} - V_{tn})^2} \int_{V_{DD}-V_{tn}}^{0.9V_{DD}} dV_o$$

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

- t_{f2}
  - $V_o : V_{DD} - V_{tn} \rightarrow 0.1\,V_{DD}$
  - NMOS in triode region

$$C_L \frac{dV_o}{dt} = \frac{\beta_n}{2}\left[\, 2(\underset{\substack{\uparrow \\ V_{DD}}}{V_{gs}} - V_{tn})V_{ds} - V_{ds}^2 \,\right]$$

$$t_{f2} = \frac{C_L}{\beta_n (V_{DD} - V_{tn})} \int_{0.1V_{DD}}^{V_{DD}-V_{tn}} \frac{dV_o}{\frac{V_o^2}{2(V_{DD} - V_{tn})} - V_o}$$

$$= \frac{C_L}{\beta_n (V_{DD} - V_{tn})} \ln\left(\frac{19V_{DD} - 20V_{tn}}{V_{DD}}\right)$$

Fall Time and Rise Time (Cont.)

- $t_f = t_{f1} + t_{f2}$

$$= \frac{2C_L}{\beta_n (V_{DD} - V_{tn})} \times \left[\frac{V_{tn} - 0.1V_{DD}}{V_{DD} - V_{tn}} + \frac{1}{2} \ln\left(\frac{19V_{DD} - 20V_{tn}}{V_{DD}}\right) \right]$$

- e.g. $V_{DD} = 5V$, $V_{tp} = -1V$, $V_{tn} = 1V$

$$t_f \approx \frac{4C_L}{\beta_n V_{DD}}$$

- Rise time

Similarly, $t_r \approx \dfrac{4C_L}{\beta_p V_{DD}}$

- For equally sized n and p devices

$$\beta_n = 2 \sim 3\beta_p \quad (\text{i.e.} \quad \mu_n = 2 \sim 3\mu_p)$$

$$t_f = \frac{t_r}{2 \sim 3}$$

- $t_r = t_f$

$$\Rightarrow \beta_p = \beta_n \Rightarrow \omega_p \approx 2 \sim 3\omega_n$$

- Delay time

  In most CMOS circuits, the delay of a single gate is dominated by the output rise and fall time

$$t_{dr} \approx \frac{t_r}{2} \quad \text{and} \quad t_{df} \approx \frac{t_f}{2}$$

Equivalent Circuit

- Equivalent Resistance

— $$t_{df} \approx \frac{t_f}{2} \approx \frac{1}{2}\frac{4C_L}{\beta_n V_{DD}} = \frac{2C_L}{\beta_n V_{DD}} = RC_L$$

— For minimum sized nMOS

$$Req = \frac{2}{\beta_n V_{DD}} \quad \text{where} \quad \beta_n = \mu_n C_{ox} \frac{W_{(min)}}{L_{(min)}}$$

— For minimum sized PMOS

$$R(PMOS) = \frac{2}{\beta_p V_{DD}} = 2 \sim 3\,Req$$

$$\text{where} \quad \beta_p = \mu_p C_{ox} \frac{W_{(min)}}{L_{(min)}}$$

- Equivalent capacitance

$$C_{eq} = C_g + C_d$$

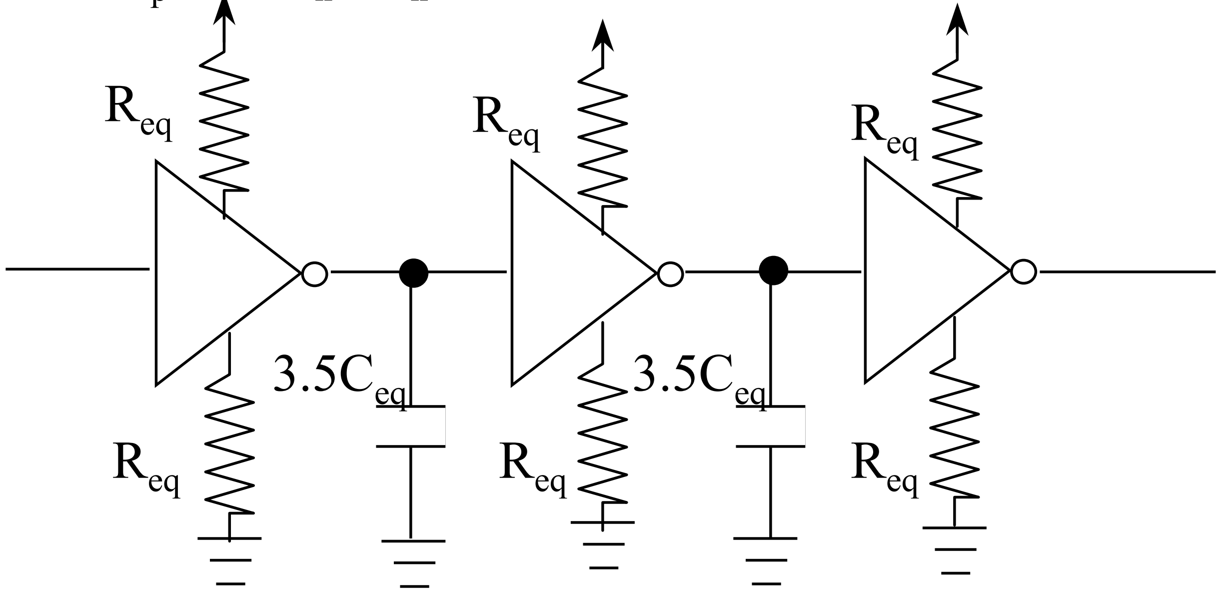
Where C_g is the gate capacitance of a minimum - sized transistor

C_d is the drain capacitance of a minimum-sized transistor

Inverter-pair Delay

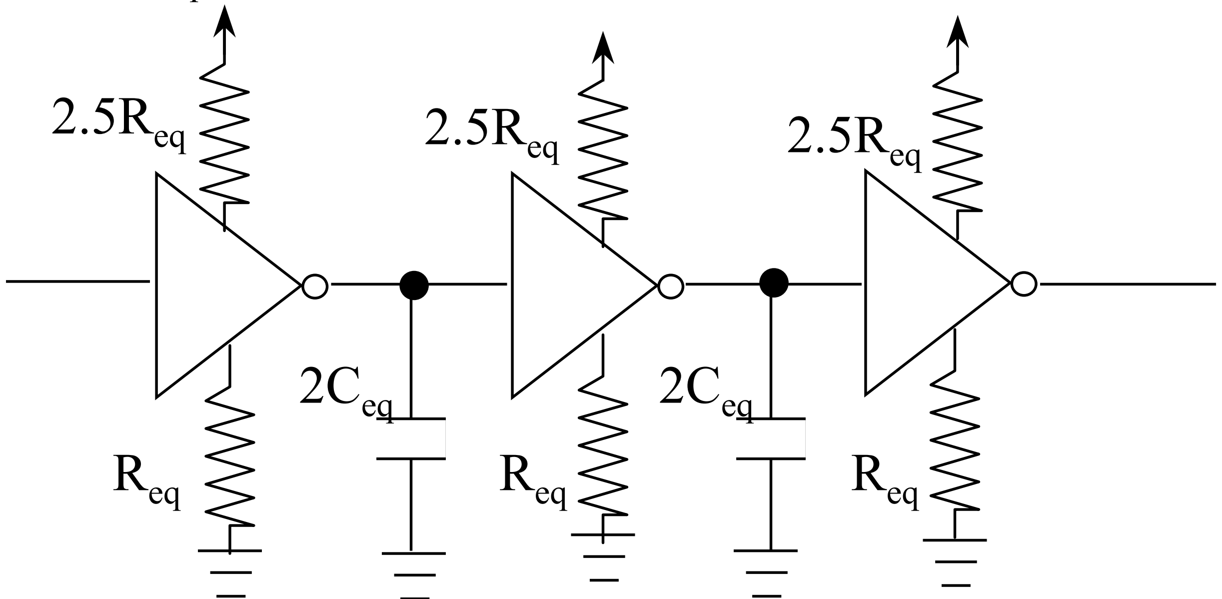
- CMOS ($\mu_n = 2.5\mu_p$ is assumed)

(a) $W_p = 2.5W_n$; W_n and L are minimum size



$$t_{inv\text{-}pair} = t_f + t_r = 3.5R_{eq}C_{eq} + 3.5R_{eq}C_{eq}$$
$$= 7R_{eq}C_{eq}$$

(b) $W_p = W_n$; All minimum sized devices



$$t_{inv\text{-}pair} = 5R_{eq}C_{eq} + 2R_{eq}C_{eq}$$
$$= 7R_{eq}C_{eq}$$

Inverter-Pair Delay (Cont.)

- Inverter threshold voltage

— $V_{inv} = \dfrac{V_{DD} + V_{tp} + V_{tn}\sqrt{\beta_n / \beta_p}}{1 + \sqrt{\beta_n / \beta_p}}$

— Variation in V_{inv} with β_n / β_p ratio

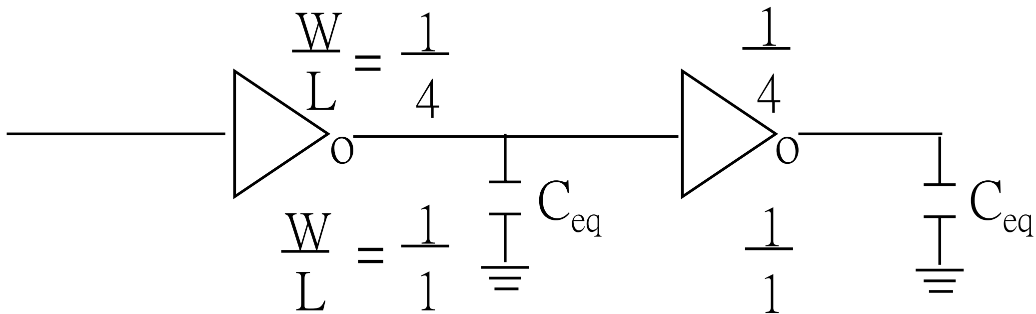
V_{DD}	V_{tn}	V_{tp}	β_n	β_p	V_{inv}
5	.7	-.7	1	1	2.5
5	.7	-.7	.5	1	2.8
5	.7	-.7	1	.5	2.2
3	.5	-.5	1	1	1.5
3	.5	-.5	.5	1	1.67
3	.5	-.5	1	.5	1.32

— Examples: V_{DD} =5V, V_{tn} = - V_{tp} =0.7V, μ_n=2.5 μ_p
(1) for W_p =2.5W_n , V_{inv} =2.5V
(2) for W_p =W_n , V_{inv} =2.1V
For (1) & (2), pair delay is the same and V_{inv} variation is 15%

- To reduce cost and power disspation, W_p =W_n is usually used. However, noise margin is reduced.
- When the circuit have to drive any significant load, n and p transistors are generally sized to yield equal rise and fall time.

NMOS Inverter-Pair Delay

- Inverter-pair delay



$$t_{inv\text{-}pair} = 4R_{eq}(C_g + 2C_d) + R_{eq}(C_g + 2C_d)$$

$$\Rightarrow 5R_{eq}C_{eq}$$

$$= 5\,\tau$$

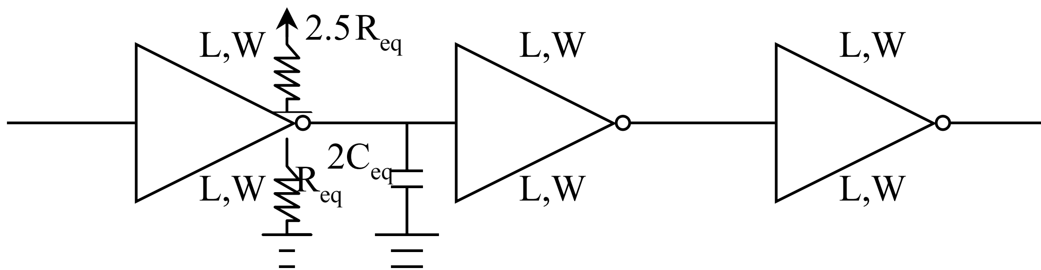
where $\tau = R_{eq}C_{eq}$ and $C_{eq} = C_g + 2\,C_d$

- NMOS inverter-pair is faster than CMOS one.
  (5 τ) (7 τ)

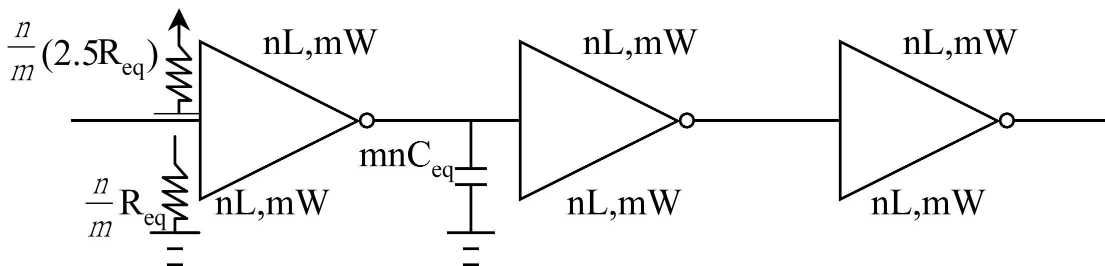
  But, NMOS inverter-pair consumes much more static power than CMOS one.

Inverter-Pair Delay (Cont.)

- Driving same size inverter



L&W are minimum size
Inverter-pair Delay$=7R_{eq}C_{eq}=7\tau$
where $\tau = R_{eq}C_{eq}$



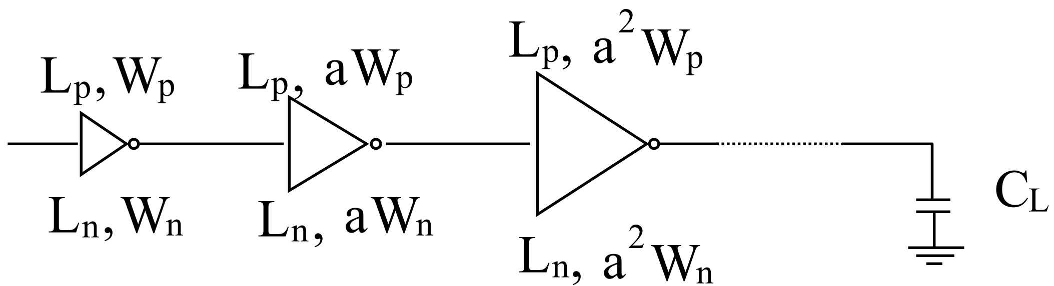
Inverter-pair Delay $=7n^2 R_{eq}C_{eq}$
$=7n^2\tau$ where $\tau = R_{eq}C_{eq}$

(Depends on channel length independent of channel width)

❄Recall that $\omega_u = \frac{g_m}{C_g} = \frac{\mu}{L^2}(V_{gs}-V_t)$

Driving Large Capacitive Loads

- CMOS



(i) If L_p & L_n are minimum channel length,

$t_{inv\text{-}pair} = 7a\tau$

(ii) If L_p & L_n are not minimum channel length, additional calculation is required to obtain $t_{inv\text{-}pair}$

－ time of the first stage (minimum L)

(i) For $\triangle V_{in}$: (a)$3.5a\tau$ $(W_p = 2.5W_n)$, (b)$2a\tau$ $(W_p = W_n)$

(ii) For $\triangledown V_{in}$: (a)$3.5a\tau$ $(W_p = 2.5W_n)$, (b)$2a\tau$ $(W_p = W_n)$

- NMOS

$t_{inv\text{-}pair} = 5a\tau$

Fall/Rise time of the first stage= $\begin{cases} a\tau & \text{for } \triangle V_{in} \\ 4a\tau & \text{for } \triangledown V_{in} \end{cases}$

Driving Large Capacitive Loads (Cont.)

- Stage ratio
  - Let $y = \frac{C_L}{\square C_g} = a^N$

    where $\square C_g$ is the gate capacitance of the first

    stage inverter

    $\Rightarrow \ln(y) = N \bullet \ln(a)$

    $\Rightarrow N = \frac{\ln(y)}{\ln(a)}$

  - For that N is even , total delay

    $t_d = \frac{N}{2} 7a\tau = 3.5Na\tau$ (CMOS)

    (or) $= \frac{N}{2} 5a\tau = 2.5Na\tau$ (NMOS)

  - For both CMOS and NMOS

    Delay α $Na\tau = \frac{\ln(y)}{\ln(a)} a\tau$

  - minimum t_d

    $\frac{d}{da}\left(\frac{\ln(y)}{\ln(a)} a\tau\right) = 0$

    $\Rightarrow a = e = 2.71828$ to have minimum value of $Na\tau$

  - number of stages for obtaining minimum t_d

    Assuming $a = e$

    $N = \ln(y)$

Driving Large Capacitive Loads (Cont.)

- Overall delay t_d for $\omega_p = \omega_n$
  - N is even

    $t_d = 3.5Ne\tau$ --- CMOS

    $t_d = 2.5Ne\tau$ --- NMOS
  - N is odd

    (i) For ΔVin (rising)

    $t_d = [3.5(N-1)+2]e\tau$ --- CMOS

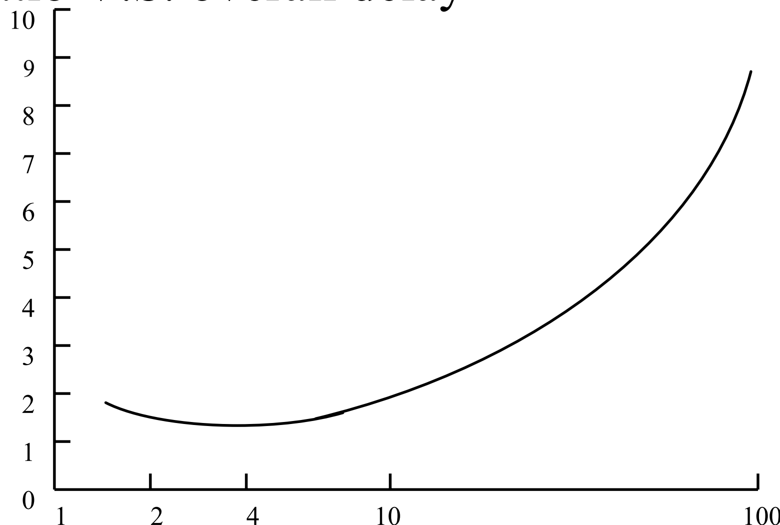
    $t_d = [2.5(N-1)+1]e\tau$ --- NMOS

    (ii) For ΔVin (falling)

    $t_d = [3.5(N-1)+5]e\tau$ --- CMOS

    $t_d = [2.5(N-1)+4]e\tau$ --- NMOS

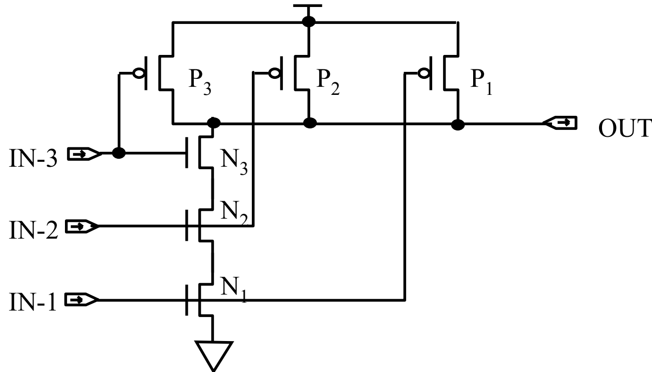
- Overall delay t_d for $\omega_p = 2.5\omega_n$ can be similarly derived.
- Stage ratio V.S. overall delay



- More detailed analysis that accounts for the contribution of the intrinsic output capacitances of inverters
  - $a = e^{\frac{k+a}{a}}$ where $k = \frac{C_{drain}}{C_{gate}}$
  - a varies from 2.3 to 5 depending on the process
  - Example : $k = 0.215 \Rightarrow a = 2.93$
- In practice, stage ratios form 2 to 10 are quite common in practical circuits depending on speed, area, and power constraints.

Gate Delays

● NAND gate explame



— $\beta_{neff} = \dfrac{1}{\dfrac{1}{\beta_{n1}} + \dfrac{1}{\beta_{n2}} + \dfrac{1}{\beta_{n3}}}$

$\beta_{neff} = \dfrac{\beta_n}{3}$ (if $\beta_{n1} = \beta_{n2} = \beta_{n3} = \beta_n$)

— $\beta_{peff} = \beta_p \sim 3\beta_p$ (if $\beta_{p1} = \beta_{p2} = \beta_{p3} = \beta_p$)

— Approximate fall time and rise time

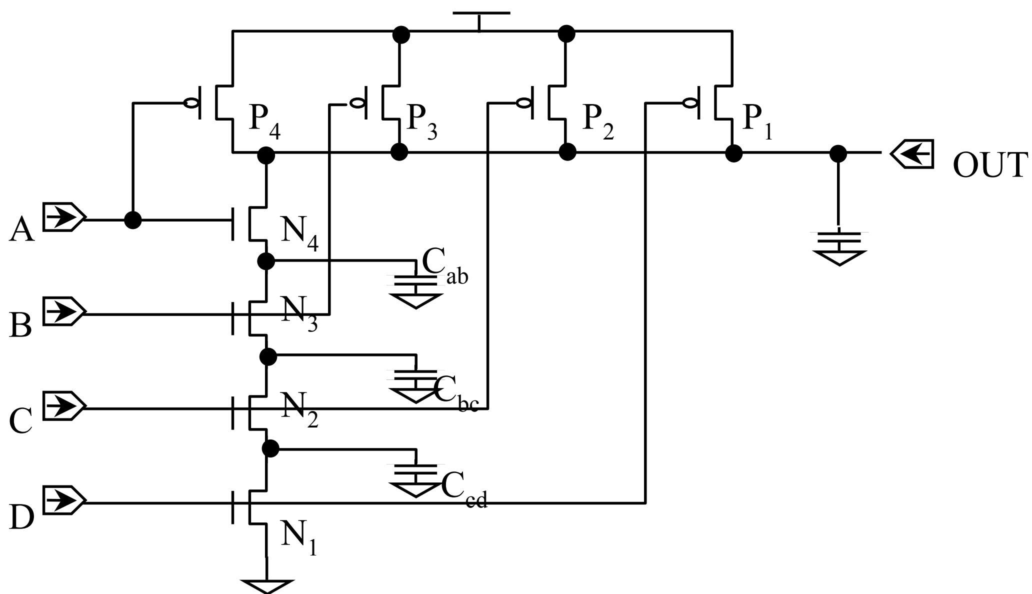
$t_{f(NAND)} \approx 3\, t_{f(inveter)}$

$t_{f(NAND)} \approx \dfrac{t_{f(inveter)}}{3} \sim t_{f(inveter)}$

(3 PMOS transisters are turned on simultaneously) (only one PMOS transister is turned on)

Switch-Level RC Models

- Example: a 4-input NAND gate with parasitic capacitances



- Simple RC delay model

— fall time

$t_{df} = \Sigma R_{pulldown} \; X \; \Sigma C_{pulldown\text{-}path}$

$= (R_{N1}+R_{N2}+R_{N3}+R_{N4}) \; X \; (C_{out}+C_{ab}+C_{bc}+C_{cd})$

where R_N is the equiualent resistance of NMOS

— rise time

$t_{dr} = R_p C_{out}/4 \quad \sim \quad R_p C_{out}$

(4 transis are turned on simultaneously) (only one transistor is turned on)

where R_p is the equiualent resistance of PMOS

- Penfield Rubenstein Model

— $t_d = \Sigma R_i C_i$

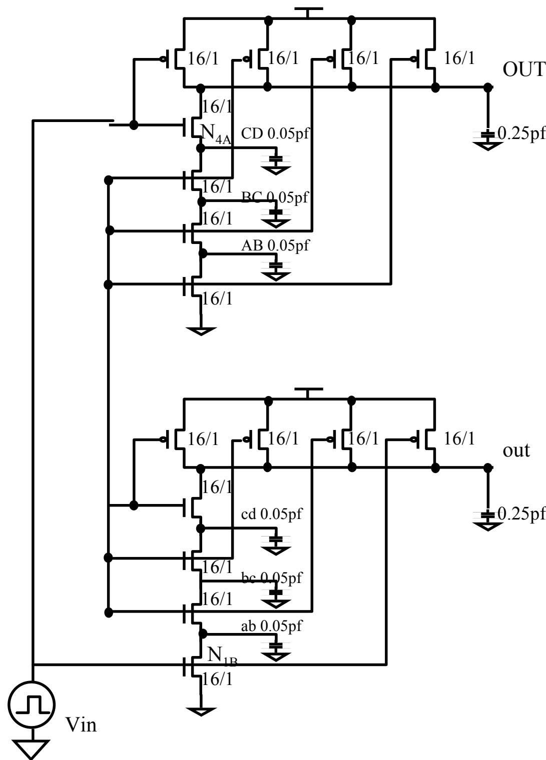
where R_i is the summed resistance from point i to power or ground and C_i is the capacitance at point i

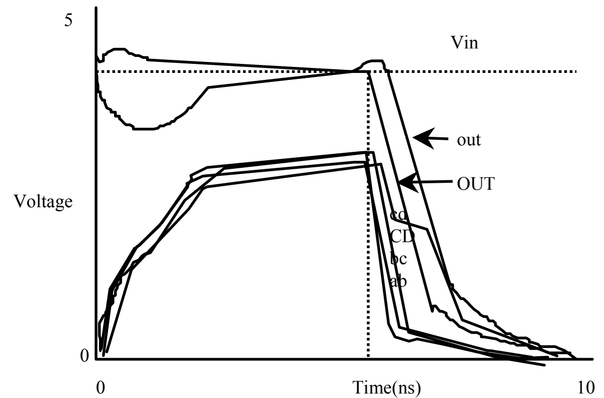
— 4 input NAND gate example

$t_{df} = R_{N1}C_{cd} + (R_{N1}+R_{N2})C_{bc} + (R_{N1}+R_{N2}+R_{N3})C_{ab} + (R_{N1}+R_{N2}+R_{N3}+R_{N4})C_{out}$

Body Effect

- NMOS(PMOS) transistors will switch slowly if their $V_{SB}>0$ ($V_{SB}<0$)
- 4-input NAND gate example
  — ciucuit diagram — SPICE simulation





— The upper circuit is faster than the lower one because the V_{SB} of the transistors are initially discharged to zero

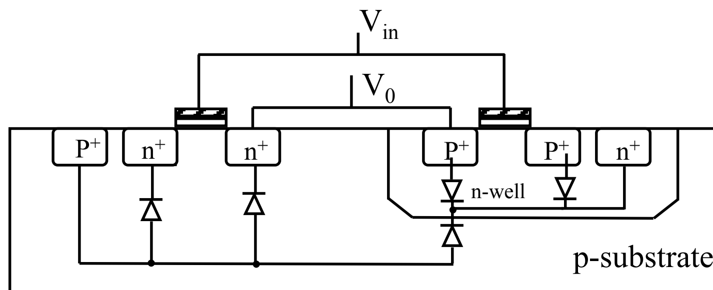
- Design strategies
  1. Place the transistors with the latest arriving signals nearest the output of a gate
     — The early signals discharge internal nodes and the late arriving signals have to switch transistors with minimum body effect
  2. Minimize the capacitance of internal nodes

Power Dissipation

- Static dissipation
- Dynamic dissipation due to
  a. switching transient current
  b. charging and discharging of load capacitances

Static Dissipation

- Due to
  a. leakage current
  b. other current drawn continuously from power supply
  e.g. NMOS circuits
  (For CMOS inverters/gates, the current in b is zero)
- Model describing parasitic diodes present in a CMOS inverter



— Junction leakages current

$$i = i_s (e^{V/V_T} - 1);\ V_T = \frac{KT}{q}$$

where i_s is reverse saturation current,
V_T is thermal voltage.

- Total static power dissipation

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

where n = number of devices

Dynamic Dissipation

- Includes

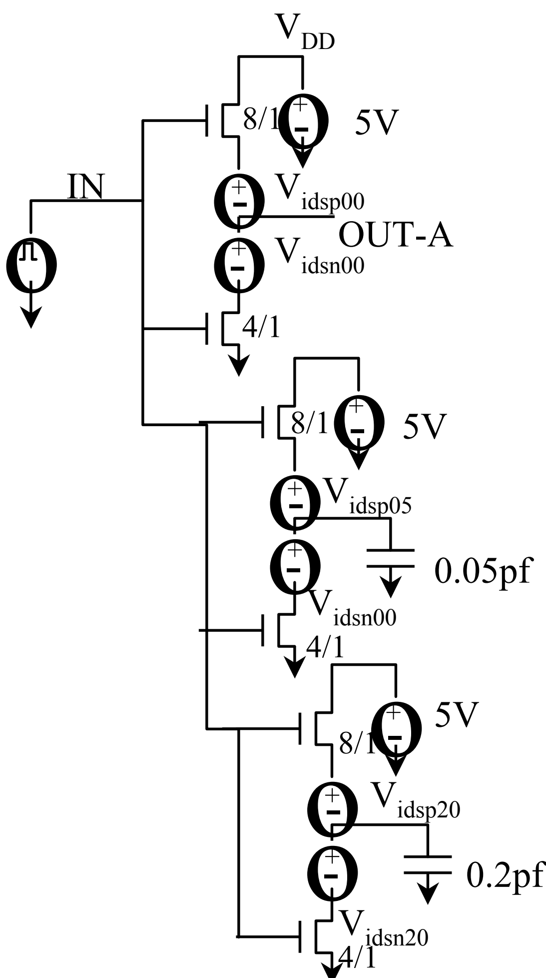
  a. switching transient current (i.e. short - circuit current)

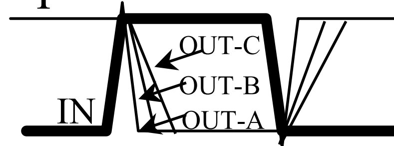
  During transition from either "0" to "1", alternatively from "1" to "0", both n-and p-transistors are on for a short period of time. This results in a short current pulse from V_{DD} to $V_{ss.}$

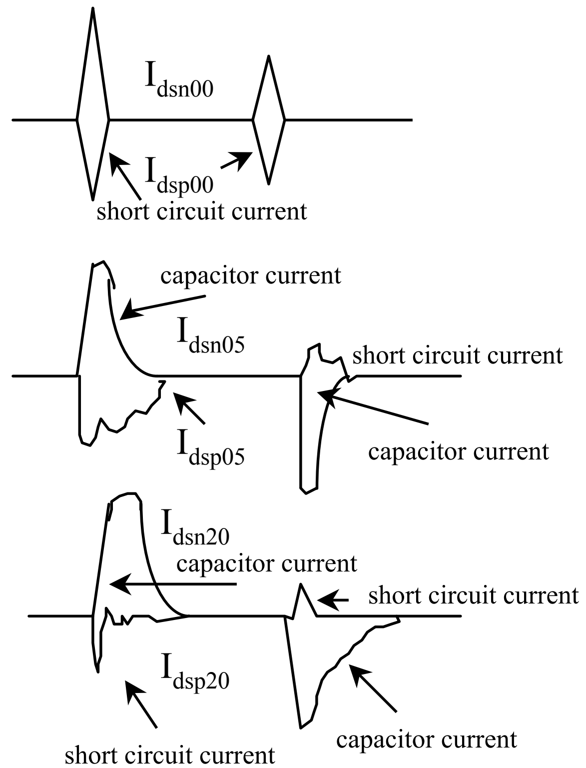
  b. Current is also required to charge and discharge the output capacitive load.

  (b. is usually the dominant term)

- SPICE circuits and results showing dynamic short-circuit current and capacitive current for a CMOS inverter for varying load capacitances

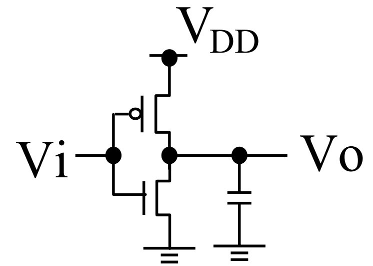






Power Dissipation due to Charging and Discharging of Load Capacitance

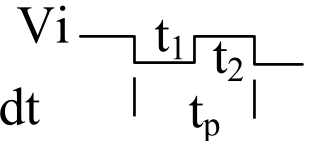
- For step-input (e.g. CMOS inverter)
  - Energy flow during t_1, i.e. Vo ↗



$Vo(t=0)=0$

$Vo(t)=V_{DD}\,(1-e^{-t/RC_L})$

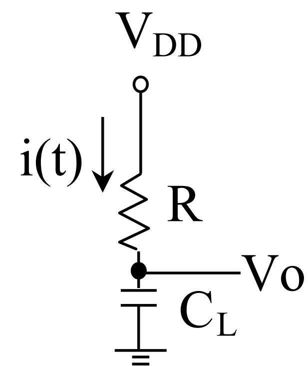
Total energy supplied by V_{DD}



$$= \int_0^{t_1} V_{DD} \bullet i(t)dt = V_{DD}\int_0^{t_1} \frac{Vo(t)}{R}\, dt$$

$$= \frac{V_{DD}}{R}\int_0^{t_1} V_{DD}\,(1-e^{-t/RC_L})\, dt$$

$$= C_L V_{DD}^2$$



Energy dissipated in R

$$= \int_0^{t_1} \frac{[\,V_{DD}-Vo(t)\,]^2}{R}\, dt$$

$$= \frac{1}{2} C_L V_{DD}^2 \text{ ——————— (A)}$$

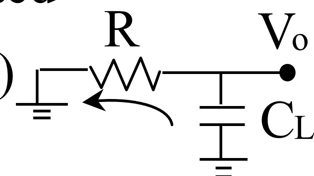
Energy stored in C_L

$$= C_L V_{DD}^2 - \frac{1}{2} C\, V_{DD}^2 = \frac{1}{2} C_L V_{DD}^2 \text{ ——————— (B)}$$

  - Energy flow during t_2 , i.e. Vo ↘

Energy stored in C_L is exhaustively dissipated

⇨ Energy dissipated $= \frac{1}{2} C_L V_{DD}^2$ —— (C)



  - During a period t_p , (B)=(C)

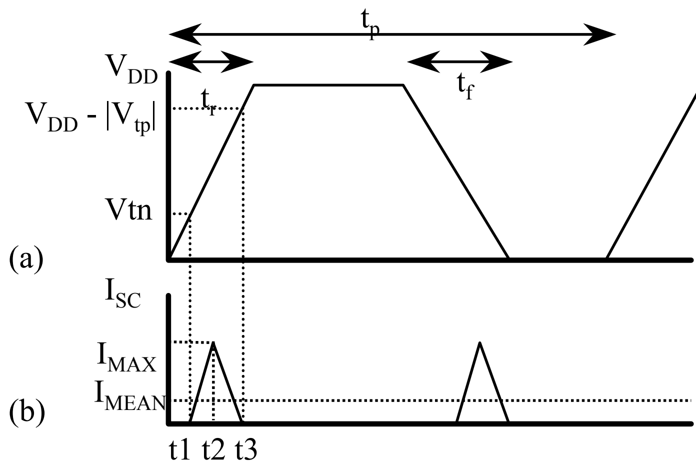
Energy dissipated = (A)+(B) = $\frac{1}{2} C_L V_{DD}^2$

⇨ Power dissipation $P_d = C_L V_{DD}^2 f_p$ where $f_p = \frac{1}{t_p}$

- Ramp V_{DD} and GND can be used in low power systems to reduce power dissipated in transistors(i.e. the above term (A) is reduced)

Short-Circuit Dissipation

- For an unloaded inverter
  - input switching waveform and model for short-circuit current



  - short-circuit dissipation

$$P_{sc} = I_{mean} \cdot V_{DD}$$

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

$$= \frac{1}{t_p}\int_{t_1}^{t_2} \frac{\beta}{2}\,[V_{in}(t) - V_{th}]^2\,dt$$

where $V_{tn} = -V_{tp}$, $\beta_n = \beta_p = \beta$ are assumed

If $V_{in}(t) = \frac{V_{DD}}{t_r} t$, $t_1 = \frac{V_{th}}{V_{DD}} t_r$, and $t_2 = \frac{t_r}{2}$,

then $P_{sc} = \frac{\beta}{12}(V_{DD} - 2V_{th})^3 \cdot t_{rf} \cdot f_p$

where $t_r = t_f = t_{rf}$ and $f_p = 1/t_p$ are assumed.

  - short-circuit dissipation is dependent on the input waveform rise and fall times
  - Keep all falling and rising edges fast if power dissipation is a concern

- As load capacitance is incresed, Psc remains almost the same. However P_d is incresed.

Total Power Dissipation

- $P_{total} = P_s + P_d + P_{sc}$
  — where P_d = activity-percentage X $(C_{total} V_{DD}^2 f_p)$
  — P_{sc} is also multiplied by an activity-percentage

Power Economy

- Power dissipation constraint must be met in a design because of low-power and thermal-dissipation concerns.
- Methods to minimize power dissipation.
  1. use minimum-sized devices to reduce.
     a. diffusion area and hence leakage current.
     b. load capacitance to reduce dynamic dissipation.
  2. Reduce supply voltage, e.g. 1.5V~3V.
  3. Manual layout techniques are used to minimize routing capacitance.
  4. Operate circuitry at the lowest possible speed
  5. Many other clever methods of manipulating architecture, circuit, and layout to achieve low-power and high-speed goals.

Size Routing Conductors

- Metal power-carrying conductors have to be sized for three reasons :

  1. Metal migration
     — current density is the dominant factor
     — for 1μm-thick Aluminum, maximum allowable current density
     $J_{Al} \approx 1 \sim 2mA/\mu\mathrm{m}$
     — V_{DD} and V_{SS} lines
  2. Power supply levels
     — IR drop during charging transients
     (Poor V_{DD} or V_{SS} levels can lead to poor logic levels which reduce noise margin and cause incorrect operation)
  3. RC delay
     has been discussed in sec. 4.3.5

Power and Ground Bounce

- Voltage spikes occur on the power and ground lines during current switching :

  1. voltage drop due to line inductance
     $$\Delta V = L\frac{\mathrm{d}i}{\mathrm{d}t}$$
  2. voltage drop due to line resistance
     $$\Delta V = \Delta i \cdot R$$

- Occur in I/O pads while driving an outside load
  — Separate power and ground buses are routed to the I/O buffers so that the ground bounce does not flow through internal circuitry.

Contact Replication

- Big contact is not suggested

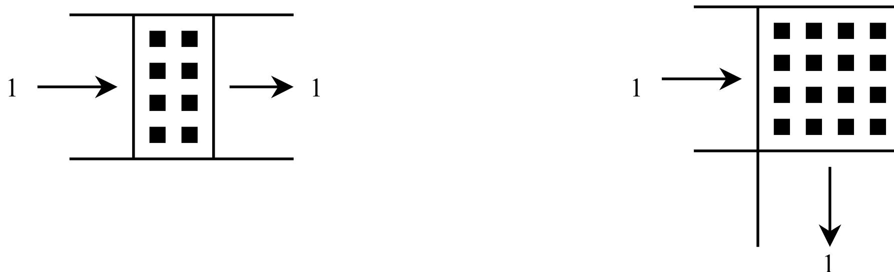
If a large contact opening were permitted, the central areas of these openings would be etched away before the oxide was completely removed for the smaller contact openings. As the etching continued in the small openings, any pinholes in the large open area could be further attacked by the etchant. Since the underlying poly layer thicknesses and diffusion depths are comparable in thickness to the layer that must be removed during contact openings, and since underlying thin oxides are much thinner, these pinholes could cause device failure. Although the probability of failure due to a single larger contact may be very low, the probability of a single failure that would render a circuit defective if a large number of these large openings were permitted may be unacceptably large. Even if the pinholes did not cause shorting, the reliability of such devices deteriorates with increased risks of premature device failures after the part is in use. For these reasons, contact openings on the gates of transistors (no contact to poly inside active) are usually not permitted either.

Contact openings to POLY II on top of POLY I are permitted. Although this type of contact is also plagued by the pinhole problem, the POLY II layer is generally used in analog applications as an upper plate of a capacitor. The total number of capacitors in these circuits is generally quite small compared to the number of transistors in a large digital circuit, thus minimizing (in the probabilistic sense) the failures due to the pinhole problem.

- Multiple-contact

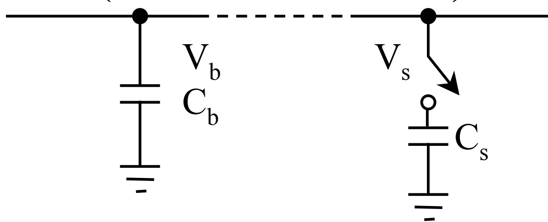
Often, a single run of a conductor can not be made to supply all circuits or modules in a design. In these cases a layer change may be necessary. Because this involves the use of interlayer contacts or vias, the resistance and current-carrying capacity of these structures must be taken into account for the effects mentioned in this section.

The current density in a contact (window, cut) periphery must be kept below about 0.1 mA/μm. We find that, due to current crowding around the perimeter of a window, a chain of small windows, suitably spaced, generally provides just as much current-carrying capacity as a single long, narrow contact. The direction of the current flow after passing through a contact can also influence the current-carrying capacity. If the current flow turns at right the flow is in the same direction, fewer contacts may be used.



Charge Sharing

- Bus(for data transfer)



— C_b is bus capacitance
— C_s is node capacitance (e.g. C_d,C_g,...,etc.)

— Total charge $Q_T = C_b V_b + C_s V_s$
— When the switch is closed, the resultant voltage V_R on C_s is ($V_b = V_{DD}$, is assumed)

$$V_R = \frac{Q_T}{C_b + C_s}$$ ===> $C_b >> C_s$ is required to ensure reliable data transfer

- Dynamic logic circuit

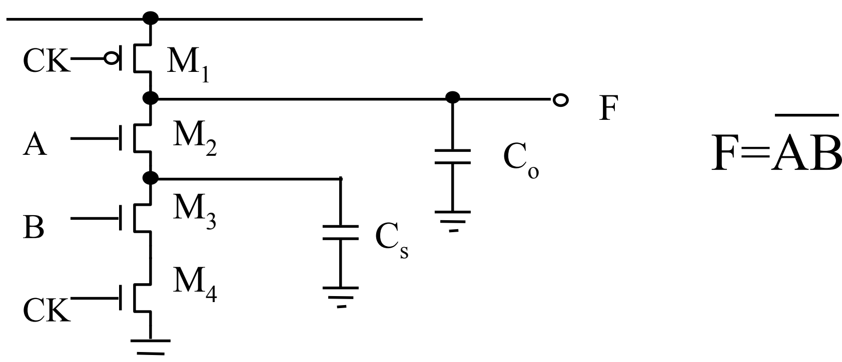
(1)When C_k=0 (precharge)

— M_1 is on, M_4 is off, $V(C_o)=V_{DD}$
M_2 off and $V(C_s)=0$ are assumed.

(2)When C_k=1 (evaluate)

If A=1 and B=0,

then $V(C_o) = \frac{V_{DD}}{C_o + C_s} < V_{DD}$
($C_o >> C_s$ is required.
Otherwise, $V(C_o) < V_{IH}$ may happen)

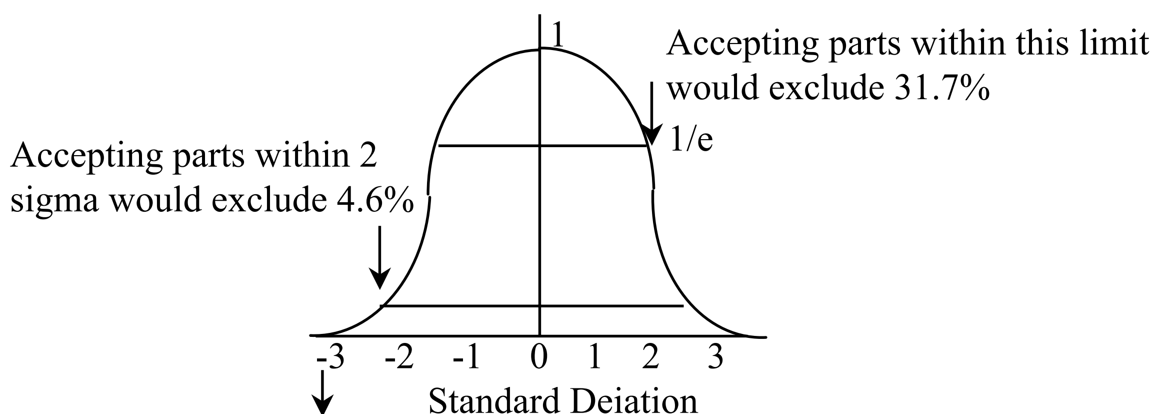


$F = \overline{AB}$

Design Margining

- Three sources of variation
  (two environmental and one manufacturing)

1. Temperature
   — ranges for normal operation
   commercial: $0^{o}C \sim 70^{o}C$
   military: $-55^{o}C \sim 125^{o}C$
   — transistors:
   T ↑ => IDS ↓ for a given set of voltage biases
   — capacitors and resistors:
   thermal coefficients
   — variations with temperature are not very important for digital circuits but are very important for analog circuits
2. Supply voltage
   — 10% variation is the general spec. for commercial product e.g. 4.5V ~ 5.5V for a 5V system
   3.0V ~ 3.6V for a 3.3V system
3. Process variation
   — The fabrication process is a long sequence of chemical reactions that result in device characteristics that follow a normal or Gaussian distribution
   — distribution of process parameters



— worst-case considerations

a. transistors:
- ✶ nominal
- ✸ fast
- ✹ slow

b. CMOS circuits
- ✶ fast-n fast-p
- ✸ fast-n slow-p
- ✹ slow-n slow-p
- ✵ slow-n fast-p

● Design corners

— one must aim to design a circuit that will reliably operate over all extremes of these three variables

CMOS Digital System Checks(Commercial)

PROCESS	TEMP	VOLTAGE	TESTS
Fast-n/fast-p	0 °C	5.5V(3.6V)	Power dissipation(DC),clock races
Slow-n/slow-p	125 °C	4.5V(3.0V)	Circuit speed,external setup and hold times
Slow-n/fast-p	0 °C	5.5V(3.6V)	Pseudo-nMOS noise margin, level shifters,memory write/ read,ratioed circuits
Fast-n/slow-p	0 °C	5.5V(3.6V)	Memories,ratioed circuits,level shifters

→ Use design-centering technique if possible

●Packaging issues

— Package selection can be very important because packages vary widely in cost and thermal impedance

— Usually the more expensive a package for a given number of pins, the better the thermal impeadance. (e.g. Ceramic is better than plastic)

— Heat sink, fan, or large cooling systems may be required

— Packages also have a wide variation in lead inductance, with ceramic pin-grid arrays having the lowest values and cheap plastic packages the highest

Yield

- Defined as $Y = \frac{\text{No. of good chips on wafer}}{\text{Total number of chips}} \times 100\%$
- Seeds's model
  — This model is used for large chips and for yields less than about 30%

  $Y = e^{-\sqrt{AD}}$

  A= chip area

  D= defect density
- Murphy's model
  — This model is used for small chips and for yields greater than 30%

  $Y = \left[\frac{1-e^{-AD}}{AD}\right]$
- A more recent generalized model

  $Y = \prod_{i=1}^{N}\left(1 + \sum_{j} \frac{A_j D_i P_{ij}}{C_j}\right)^{-C_i}$

  i= the i-th type of defect

  j= the j-th module

  P_{ij} = the probability that an i defect will cause a fault in the j-th area

  C_i = the constant relating to the density of a i-th type of defect

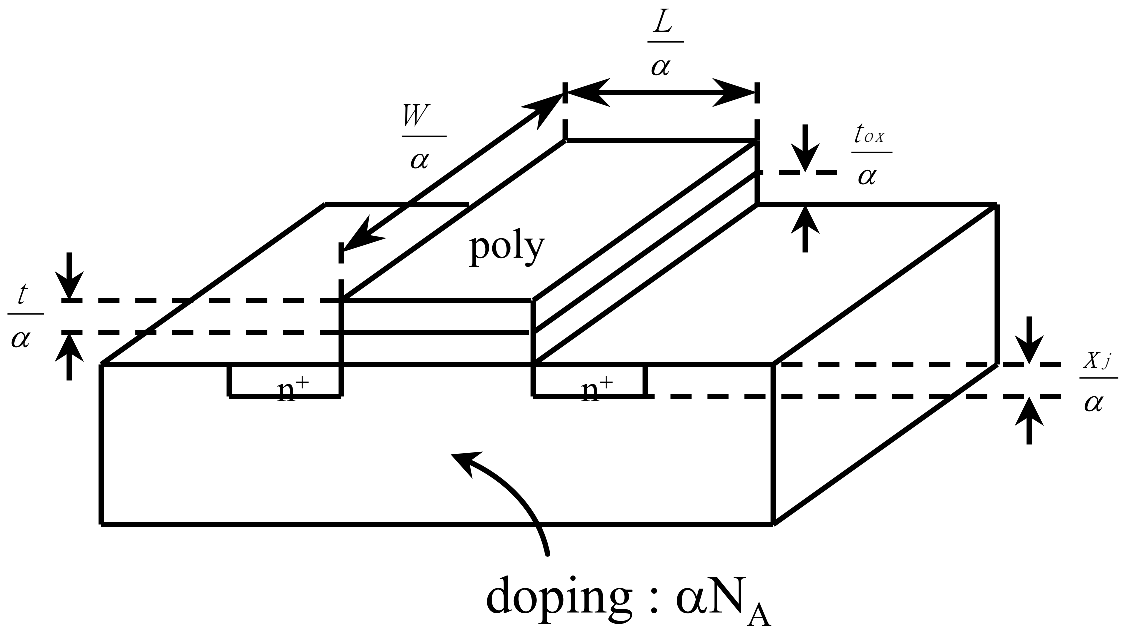
Realialility

- long lifttime of a part should be guaranteed
- potential reliability problems
  — Hot electron effects
  — Electromigration
  — Oxide failure
  — Bipolar transistor degradation
  — Package / chip power dissipation(die temperature)
  — ESD protection

Scaling of MOS transistor dimensions

Scaling factor α ⟹ (components/unit area) will increase by a factor of α^2

λ–rule : λ變大/小 ⟹ 所有rule都變大/小
⟹ layout rule變大/小



doping : αN_A

Scaling Model : constant field scaling (fist order)

Scaling factor so applied to :

- All dimensions, including those vertical to the surface
- device voltages
- the concentration densities

i.e.
- L, W, t_{ox}, X_j ....$1/\alpha$
- V_{DD} ⟹ V_{DD}/α
- N_A ⟹ αN_A

Influence of Transistor Scaling

- First-order influence

	PARAMETERS	SCALING FACTOR
DEVICE PARAMETERS	Length; L	$1/\alpha$
	Width; W	$1/\alpha$
	Gate oxide thickness; t_{ox}	$1/\alpha$
	Junction depth; X_j	$1/\alpha$
	Substrate doping; N_a	α
	Supply voltage; V_{DD}	$1/\alpha$
RESULTANT INFLUENCE	Electric field across gate oxide; E	1
	Depletion layer thickness; d	$1/\alpha$
	Parasitic capacitance; WL/t_{ox}	$1/\alpha$
	Gate delay; (VC/I)	$1/\alpha$
	DC power dissipation; P_s	$1/\alpha^2$
	Dynamic power dissipation; P_d	$1/\alpha^2$
	Power- delayproduct	$1/\alpha^2$
	Gate area	$1/\alpha^2$
	Power density; (VI/A)	1
	Current density; (I/A)	α
	Transconductance; g_m	1

- Yield

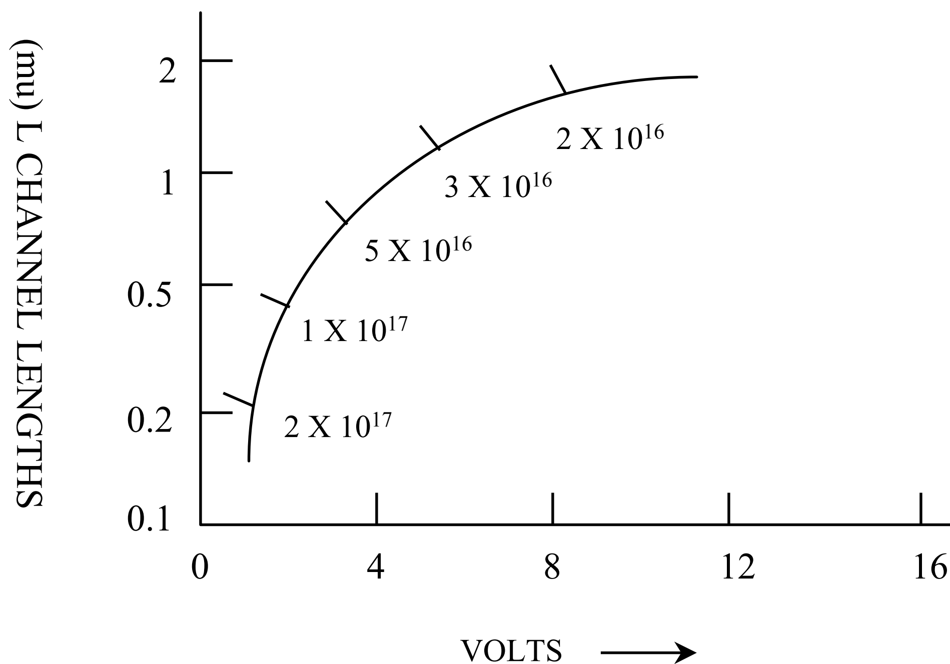
  Device dimension ↓ => chip area ↓ => yield ↑

Limitations to Scaling

- Power supply :

  — Noise margin is reduced

- Relationship among V, L , and doping concentration



- Metal Migration :

  — Current density limit on metal (conductor) and others

- Latchup

  — spacing between NMOS and PMOS can not be arbitrarily scaled down without the danger of including latchup conditions

Scaling of Wires and Interconnections

- Influence of scaling on interconnect media

PARAMETERS	SCALING FACTOR
Line resistance; r	α
Line response; rc	1
Normalized line response	α
Line voltage drop; V_d'	1
Normalized line voltage drop V_d' / V_{DD}'	α
Current density; J	α
Normalized contact voltage drop; V_C/V	α^2

- $R' = \frac{\rho}{t/\alpha} \left[\frac{L/\alpha}{W/\alpha} \right] = R$
- $R'C' = (\alpha R)\left(\frac{C}{\alpha}\right) = RC$
- $\frac{R'C'}{\tau'} = \frac{R'C'}{\tau/\alpha} = \alpha\tau$

  (Normailized line response)
- $V_d' = \frac{I}{\alpha}\,\alpha R = IR$
- Normailized line voltage drop

$$\frac{V_d'}{V_{DD}'} = \frac{IR}{V_{DD}/\alpha} = \frac{V_d}{V_{DD}/\alpha} = \alpha$$