VLSI CMOS

Abstraction: device (spice .model) → circuit (P&R → layout) → gate (logic synthesis) → functional (RTL (VHDL), V) (high level synthesis) → behavioral model (systemC, SV)

↑ transcend

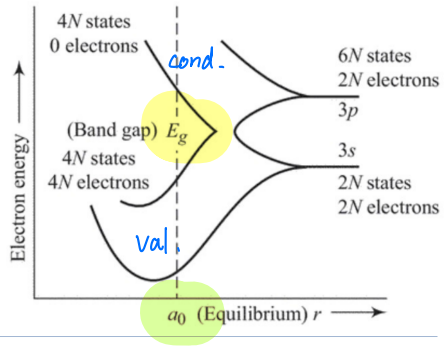
Problem: power distribution integrity (V_{DD}); reliability; noise; interference

A vs D: NM

A: more energy efficient, less reliable

Material: Si vs 3$ vs wide-bandgap (SiC, GaN)

MOS vs BJT: power

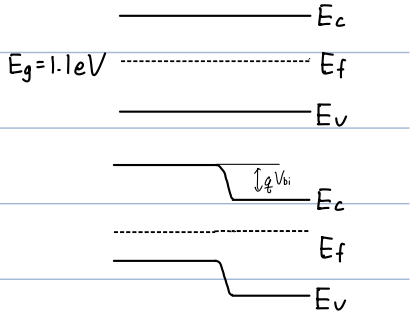


Semi: T↑, e^- jump to E_c → EHP

$np = n_i^2$

Fermi Dirac: $f(E) = \frac{1}{1-e^{\frac{E-E_f}{kT}}}$

$n = n_i e^{\frac{E_f - E_i}{kT}}$ $\quad p = n_i e^{\frac{E_i - E_f}{kT}}$



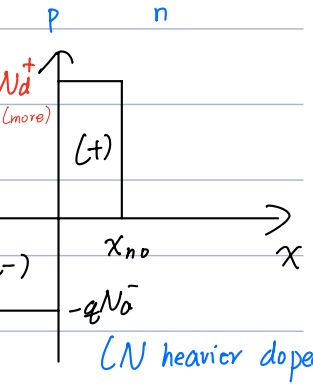
pnj: band bend ← E → ΔV

diffuse → $q^{+/-}$

$\frac{dE}{dx} = \frac{\rho}{K_s \varepsilon_0}$ $\quad K_s = 11.7$, $\varepsilon_0 = 8.85E-14$

$\int_0^{E_{max}} dE = \int_{-x_p}^{0} \frac{\rho}{K_s \varepsilon_0} dx$

lin: $E_{max} = \frac{qN_A^-}{K\varepsilon_0} x_p \overset{\cdots}{=} \frac{qN_D^+}{K\varepsilon_0} x_n$



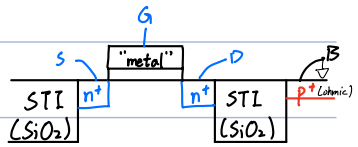
$V_{bi} = \frac{1}{q}\left(kT \ln \frac{N_A}{n_i} + kT \ln \frac{N_D}{n_i}\right) = \frac{kT}{q} \ln\left(\frac{N_A N_D}{n_i^2}\right) = \frac{1}{2} E_{max} W$

$W \overset{\cdots}{=} \left[\frac{2K_{si}\varepsilon_0 (V_{bi} - V_A)}{q}\left(\frac{1}{N_a} + \frac{1}{N_d}\right)\right]$

bias V_A: non-eq

$I \sim e^{\frac{qV_A}{kT}}$

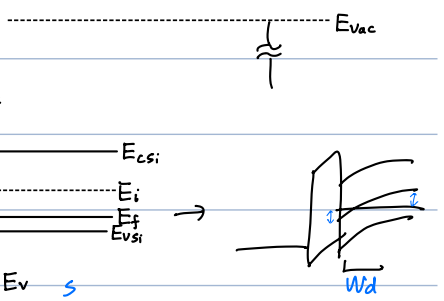
C. $C_0 = \frac{K_{si}\varepsilon_0}{w}$ (per area, ss, ↓ as RB↑, parasitic)



MOS/FET: SD sym., depends on how biased in circuit

$V_G = 0$, flat band V_{FB} (=0 for ideal) to ensure $\mathcal{E} = 0$

$V_G > 0$, band bend → depletion in S



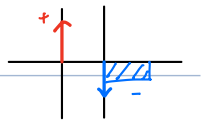
$E_i - E_f = kT \ln\left(\frac{N_A}{n_i}\right) \equiv \phi_f$

@ interface @ bulk

Once $E_f - E_i = E_i - E_f$ → inversion

past V_T ↓ $V_G - V_T$

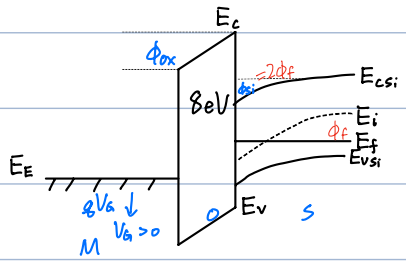
dep. region no longer grows. e^- grows in inv. to balance



$V_{ov} = V_G - V_T$ decides current

$W_d = \sqrt{\frac{2K_s\varepsilon_0\phi_s}{qN_A}}$ ϕ_s : surface vs bulk ;

$W_d^{max} = \sqrt{\frac{2K_s\varepsilon_0 2|\phi_f|}{qN_A}}$ $Q_{blk} = qN_A W_d^{max}$



$V_T = \phi_{ox} + \phi_s$

$= \frac{Q_{bulk}}{C_{ox}} + 2|\phi_f| + V_{FB}$ compensate for non-flat-band

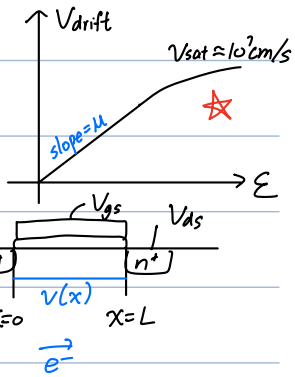
$C_{ox} = \frac{K_{ox}\varepsilon_0}{t_{ox}}$ (3.9)

Use nonuniform doping profile to control W_d and V_T

As size ↓, # dopants countable (random dopant fluctuation) → :(

sln. deplete Si → no dopant → engineer ϕ_M for a good V_{FB}

Mobility $v_{drift} = \mu\mathcal{E}$ ↗ $v_{sat} \approx 10^7$ cm/s



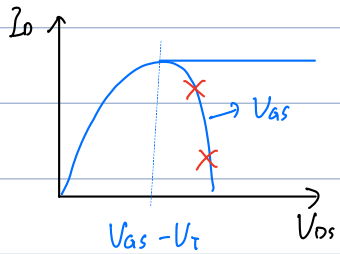
long-channel Schokley (assume ↓)

$I_D = WQ_i v_{drift}$ (Q_i: inv. layer charge density)

$= WC_{ox}(V_{gs} - V(x) - V_T)(\mu\frac{dV}{dx})$

$\int_0^L I_D dx = \int_0^{V_{DS}} WC_{ox}(V_{GS} - V_T - V(x))\,\mu\,dV$

$I_D = \frac{W}{L}\mu C_{ox}[(V_{GS} - V_T)V_{DS} - \frac{V_{DS}^2}{2}]$ (linear)



$V_{DS} > V_{ov}$ pinchoff (no free carriers, e^- swept to drain) → $I_{sat} = \frac{W}{2L}\mu C_{ox}(V_{GS} - V_T)^2$

$I_{sat} \sim V_{ov}^2$, since $Q_i^1 \cdot v_{drift}^1$

short channel $Q_i^1 \cdot v_{sat}^0 \rightarrow \sim V_{ov}^1$

$v_{drift} = \frac{\mu\mathcal{E}}{1 + \mathcal{E}/\mathcal{E}_{crit}}$ $\sim 10^4$ V/cm (short)

$I_D = WC_{ox}(V_{gs} - V(x) - V_T)\frac{\mu\mathcal{E}}{1+\mathcal{E}/\mathcal{E}_{crit}}$

$I_D \overset{\int\cdots}{=} \frac{W}{L}\frac{\mu C_{ox}}{1 + \frac{V_{DS}}{\mathcal{E}_{crit}L}}[(V_{GS} - V_T)V_{DS} - \frac{V_{DS}^2}{2}]$

$\frac{\partial I_D}{\partial V_{DS}} = 0 \rightarrow V_{DS} \overset{\cdots}{=} V_{Dsat} = \frac{2(V_{GS} - V_T)}{1 + \sqrt{1 + 2\mu(V_{GS} - V_T)/\mathcal{E}_{crit}L}}$

$I_{Dsat} \overset{\cdots}{=} C_{ox} W v_{sat}(V_{GS} - V_T)\frac{\sqrt{\quad} - 1}{\sqrt{\quad} + 1}$ if $\frac{V_{GS} - V_T}{L} >> \frac{\mathcal{E}_{crit}}{2}$, $I_{Dsat} \rightarrow C_{ox} W v_{sat}(V_{GS} - V_T)$ (lin)

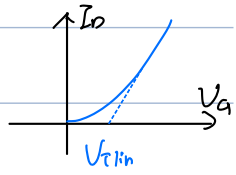
Enhancement : off when $V_G = 0$ ($V_T < 0$ for p, > 0 for n) TAOH?

Depletion on ⁀⁀⁀ (rare)

V_T def { 1. device physical: onset of strong inv.

2. triode $I_D = \sim\sim (V_{GS} - V_T) V_{DS}$ (extrapolate lin. part → $V_{T\,lin}$)

3. sat. find V_{GS} s.t. I_D reaches a specific density when $V_{DS} = V_{DD}$ (E. $\frac{1\mu A}{\mu m_{width}}$)



Scaling (Dennard) $W \to \frac{W}{K}$; $L \to \frac{L}{K}$; $t_{ox} \to \frac{t_{ox}}{K}$; $V \to \frac{V}{K}$ (const. $\mathcal{E}$ scaling)

1. Depletion region $W_{dep} \sim \sqrt{\frac{V}{N}} \overset{\text{want}}{\sim} \frac{1}{K}$ → want $N \to KN$

but $V = V_{bias} + V_{built\text{-}in}$ ~ bondgap → can't scale → short channel

2. C $C_{gate} = \frac{\varepsilon W L}{t_{ox}} \sim \frac{1}{K}$

3. $Q_i \sim \frac{\varepsilon V}{t_{ox}} \sim 1$, but V :((inv. layer charge density)

4. $I_D \sim Q_i W \cancel{v_{drift}}^{1} \sim \frac{1}{K}$

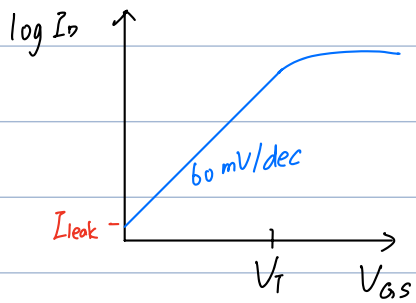
5. delay $\approx \frac{CV}{I} \sim \frac{1}{K}$ → $f \sim K$

6. power $= IV \sim \frac{1}{K^2}$ (density ~ 1)

V_T inability $I_D \sim e^{\frac{q(V_{GS}-V_T)}{mkT}}$

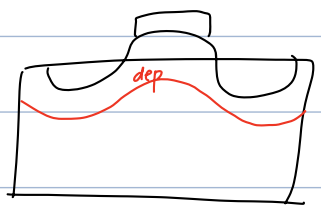
$\log_{10}$ slope $= \frac{mkT}{q} \frac{1}{\log_{10} e} \approx 60\ mV/dec$ of I_D

if $V_T \downarrow$ → $I_{leak} \uparrow$



if size ↓, but W_{dep} ✱ → D starts to control Q dep

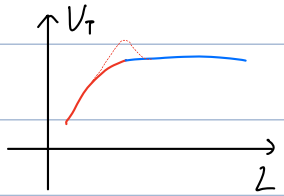
$V_T = V_{FB} + 2|\phi_F| + \frac{Q_{bulk}\downarrow}{C_{ox}}$ not much control



If V_{DS} ↑, $V_{T,eff}$ ↓, so I_D ↑ slightly → short channel effect

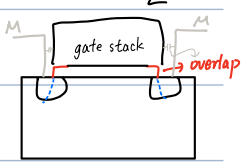
(DIBL: drain Induced Barrier Lowering ~ $\frac{mV}{V}$)

V_T rolloff if L too short, V_T ↓, same reason



want minimal variation

Cap from device / wires (0.1 fF/μm)



intrinsic : $C_{gate} \sim I_D$

extrinsic (parasitic) — overlap from V_{GS}, V_{GD}; also between M contact & gate stack

diffusion (junction): jct C across W_{dep}, controlled by layout

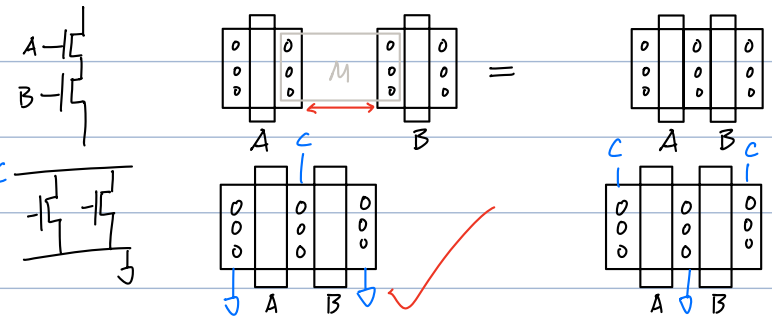
Intrinsic distribution C_{gs} C_{gd} C_{gb}

cutoff no inv. → $C_{gs} = C_{gd} = 0$, $C_{gb} = C_{ox} W L$ series w/ C_{bulk}

triode uniform inv. → $C_{gs} = C_{gd} = \frac{1}{2} C_{ox} W L$; $C_{gb} = 0$

sat. most C not at drain → $C_{gs} = \frac{2}{3}$ (narrower) $C_{ox} W L$; $C_{gd} = C_{gb} = 0$

Layout



combine S/D → $\frac{1}{2}$ diff. area & no metal spacing need

right X, since C gets more Cap

want more Cap on gnd (decoupling)

Body effect if $V_{SB} \neq 0$ (N)

$$V_T = V_{FB} + 2|\phi_F| + \frac{Q_B}{C_{ox}} \quad ; \quad Q_B = \sqrt{2 q_e \varepsilon_{Si} N_A \underbrace{2|\phi_F|}_{\downarrow}}$$

$2\phi_f + V_{SB} \rightarrow V_T \uparrow$

must apply larger V_G to create inv. May also use V_{SB} to control V_T

Transfer curve

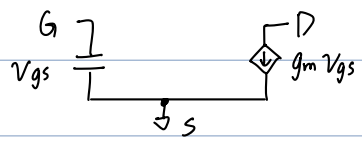
f_T: unity current gain cutoff f

FO4 (fanout of 4) } delay of D0 driving 4 D0s of same size

FO1 } ~~~~~~~~~~ | ~~~~~~~~~~

(N) sat

SS



$$g_m = \frac{\partial I_D}{\partial V_{GS}}$$

$$v_{gs} = i_g \cdot \frac{1}{sC_{gs}}$$

$$i_d = i_g \frac{g_m}{sC_{gs}}$$

$$|A_i(j\omega)| = \frac{g_m}{\omega C_{gs}} = 1$$

$$f_T = \frac{1}{2\pi} \frac{g_m}{C_{gs}} \overset{\partial}{=} \frac{1}{2\pi} \frac{W C_{ox} V_{sat}}{WL C_{ox}}$$

(transit time) $= \frac{1}{2\pi} \frac{V_{sat}}{L} \equiv \frac{1}{2\pi} \cdot \frac{1}{\tau_{tr}}$

FO1 $V_{sat,p} \simeq \frac{1}{2} V_{sat,n}$, so $W_p = 2W_n$ (but ~ same for modern)

$$N\ t \approx \frac{V_{DD}}{I_{Dsat}} (C_{gn} + C_{dn})$$

$$= \frac{V_{DD}}{I_{Dsat}} C_{gn} \left(1 + \frac{1}{\eta}\right)$$

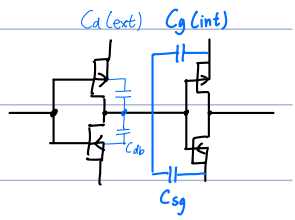
$$= \frac{V_{DD}\ WL C_{ox} \left(1 + \frac{1}{\eta}\right)}{W C_{ox} (V_{GS} - V_T) V_{sat}}$$

$$= \frac{L}{V_{sat}} \left(1 + \frac{1}{\eta}\right)$$

$$= \tau_{tr} \left(1 + \frac{1}{\eta}\right)$$

let $V_{GS} = V_{DD}$, $V_T \sim 0$

$\eta \equiv \frac{C_g}{C_d} \sim 1.5 \sim 2$



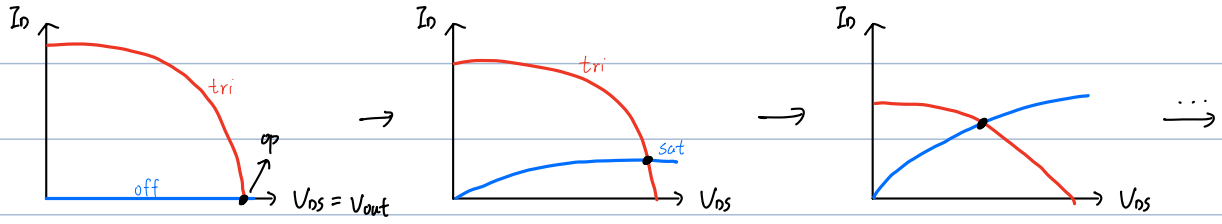
(S are not switching, irrelevant)

P+N: $t(2w) \overset{\cdots}{=} 3\tau_{tr}\left(1 + \frac{1}{\eta}\right) \equiv \tau_{inv}$ $\quad \eta = \frac{C_{gn} \times (2+1)}{C_{dn} \times (2+1)}$ (p+n)

FO4 $= \frac{4C_g + C_d}{C_g + C_d} \cdot$ FO1 (= x4 if $\eta = \infty$)

$\overset{\cdots}{=} (1+4\eta) \cdot \frac{3}{\eta} \tau_{tr} = 3\tau_{tr}\left(4 + \frac{1}{\eta}\right)$

Noise (DC)



rapid change in middle → D0 trans. curve

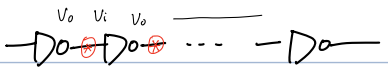
Noise at end will be restored

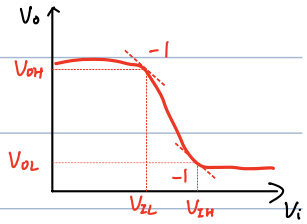
NM (DC, acting on every node)

+ worst-case DC noise (*)

if AC (NM_{AC}), will be filtered by C_g (better)

$NM_L = (V_{IL} - V_{OL})$; $NM_H = V_{OH} - V_{IH}$





Switching energy

output charges C_L

0→1 $E = \int dt\ i_{DD}(t)\ V_{DD}$

$= V_{DD} \int_0^\infty dt\ C_L \frac{dV_{out}}{dt}$

$= C_L V_{DD}^2$ but $E_C = \frac{1}{2} C_L V_{DD}^2$ → rest heat

1→0 stored $\frac{1}{2}$ dissi.

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

$I_{crowbar}$: while switching, a transient time where both Q on (~ 10-20% p)

especilly when Vin switch slow (low slew rate)

spice

Compact model : many params. Have a distribution (may be correlated)

E. channel length E. in a wafer

corner model : set 3σ from μ → ensure still work

TT, SS, FF, SF, FS → corners

(not too much σ, since same chip)

PVT : process, voltage, temperature

Delay

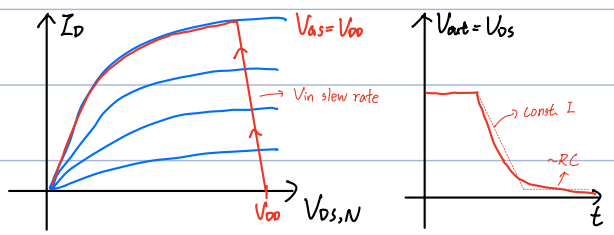
Vin : 0→1

~ saturated ramp. delay T_{pHL}^{out}, T_{pLH} = Δt for 50% V_{DD}

slew ~ ramp slope of 10% - 90%

or extrapolate to 0~100 (÷0.8), up to ur def

C_g $\left[\frac{fF}{\mu m}\right]^{\sim 2}$ @ min length



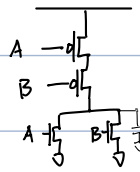
$R_{nFET} = W_n R_{pulldown}$ $[k\Omega \cdot \mu m]$ $\quad R_{pFET} = W_p R_{pullup}$

$C_g = C_{gn} = C_{gp} = \frac{C_{gate}}{W} \simeq \frac{K\varepsilon_0 L}{t_{ox}}$ $\left[\frac{fF}{\mu m}\right]$ @ min width

For p: 2W; n: W inverter : logical effort $g_{inv} \equiv 1$ (intrinsic to gate)

electrical effort $h_{inv} \equiv \frac{C_{out}}{C_{in}} = 1$ (if same size Do)

normalized delay $F \equiv \frac{\tau_{logic\ gate}}{\tau_{inv}} = g_{gate} \cdot h_{gate}$

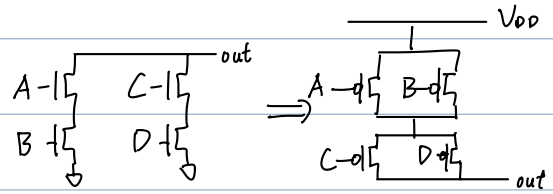
E. nor 

exactly 1 on at a time

want input w/ more delay to be closer to output

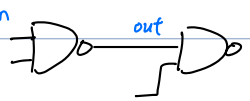
Ser. in PDN ↔ ll in PUN

want ≤ 2 Q high per network



$\tau_{inv} = FO1_{inv} = \frac{R_{pFET}}{W} \cdot 3WC_g = 3C_g R_{nFET}$ ($fF/\mu m$, $k\Omega \cdot \mu m$)

E. NAND size wrt to inv. → p: 2W ; n: 2W



$\tau_{nand} = FO1_{nand} = (4WC_g) \times \left(\frac{R_{nfet}}{2W}\right) \cdot 2 = 4C_g R_{nfet}$ (output drives 2Wp, 2Wn; C; R)

$f \equiv \frac{\tau_{nand}}{\tau_{inv}} = \frac{4}{3}$

$h = \frac{4WC_g}{4WC_g} = 1 \rightarrow g_{nand} = \frac{f}{h} = \frac{4}{3}$

E2. NOR → p: 4W, n: 1W

$\tau_{nor} = FO1_{nor} = (5WC_g) \times \left(\frac{R_{nfet}}{W}\right) = 5C_g R_{nfet}$ (4+1)

$d = \frac{5}{3}$, $g_{nor} = \frac{5}{3}$

E. if NOR driving 2 NORs w/ same 4W − W

$FO2_{nor} = (10WC_g) \times \left(\frac{R_{nfet}}{W}\right) = 10C_g R_{nfet}$

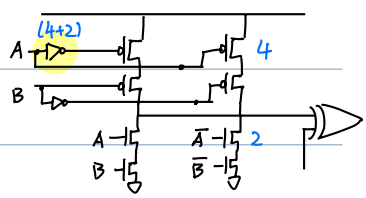
$d = \frac{10}{3}$, $h = 2$, $g = \frac{5}{3}$ (same logical, more electrical)

E3. XOR Need to size ▷

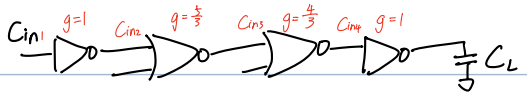
$FO1_{xor} = FO1_{main} \times FO1_{inv}$ multiply stage effort

$= (12WC_g)\left(\frac{R_{nfet}}{2W} \cdot 2\right) \times (6WC_g)\left(\frac{R_{nfet}}{2W}\right)$ (half the gates)

$= 12C_g R_{nfet} \times 3C_g R_{nfet} = 4\tau_{inv} \cdot 1\tau_{inv} \rightarrow F = 4$



E. if inv is 2W/W, $FO1_{xor} = (9WC_g) \cdots \times (6WC_g)\left(\frac{R_{nfet}}{W}\right) = 3\tau_{inv} \cdot 2\tau_{inv} \rightarrow F = 6 > 4$:(

E. 

Assume C_{in}, C_w fixed (if not, make $C_{in} \to \infty$)

$$F = \prod_{i=1}^{n} F_i = \prod_{i=1}^{n} (g_i h_i) = \prod_{i=1}^{n} g_i \cdot \prod_{i=1}^{n} h_i = \left(\frac{20}{9}\right)\left(\frac{C_{in2}}{C_{in1}} \times \cdots \times \frac{C_L}{C_{in4}}\right) = \frac{20}{9}\frac{C_L}{C_{in1}} \to \text{fixed}$$

(fixed)

$$\tau_{path} = \sum_{i=1}^{n} (g_i h_i)\tau_{inv} \to \text{minimizable}$$

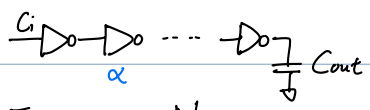
E. just one NOR, one NAND

Let norm. delay $y \equiv f_{nor} + f_{nand} = \frac{F}{f_{nand}} + f_{nand} = f_{nand}\left(1 + \frac{F}{f_{nand}^2}\right)$

$$\frac{\partial y}{\partial f_{nand}} = 1 - \frac{F}{f_{nand}^2} = 0 \to f_{nand} = \sqrt{F}$$

General for n stages: $f_i = \sqrt[n]{F}$ for min delay (if no branching (const. H))

+ stages N-stage inv. chain, $\eta = \infty$ (ignore C_d), $f_i = \alpha$



$$F = \prod_{i=1}^{N} f_i = \alpha^N = \frac{C_{out}}{C_{in}}$$

$$N = \frac{\ln(C_{out}/C_{in})}{\ln \alpha}$$

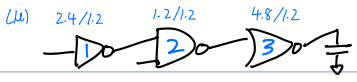
$$\tau_d = \tau_{inv} \sum_{i=0}^{N} F = \tau_{inv}\, \alpha N$$

$$= \tau_{inv}\, \alpha \frac{\ln(C_{out}/C_{in})}{\ln \alpha}$$

$$0 = \frac{\partial \tau_d}{\partial \alpha} =$$

$\alpha = e$ (closest even)

if include C_d, $\alpha = 4$

E. 

$C_g = \frac{2fF}{\mu m}$, $R_{pf} = 16\ k\Omega \cdot \mu m$, $R_{nf} = 8\ k\Omega \cdot \mu m$, $\tau_{inv} = 3R_{nf}C_g = 48$ ps

① find eff C_{out}. $f_3 = \frac{500\mu m}{3 \cdot 1.2\mu m} = \frac{5}{3} \cdot \frac{1pF}{2fF \cdot 6\mu m}$

(inv. size to match gate; actual size)

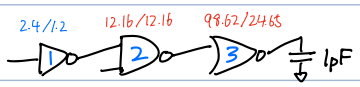
$$f_2 = \frac{6\mu m}{3 \cdot 0.6\mu m}$$

$$f_1 = \frac{2.4\mu m}{3 \cdot 1.2\mu m}$$

② $\tau_{delay} = \tau_{inv}(f_3 + f_2 + f_1) \approx 6858$ ps, too long

③ $F = f_3 f_2 f_1 = 308.6 \to f_{op} = \sqrt[3]{F} \approx 6.76 \to \tau'_{delay} = 973$ ps

④ $W_{inv,eff,3} = \frac{500\mu m}{3 \cdot 6.76} \cdots \approx 24.65\ \mu m \to 4W/W = 98.62/24.65$



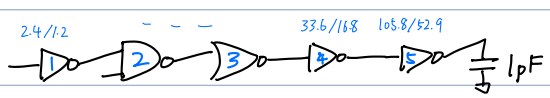
$W_{inv,eff,2} = \frac{5 \cdot 24.65\mu m}{3 \cdot 6.76} = 6.08\ \mu m \to 2W/2W = 12.16/12.16$ (output)

$W_{inv,eff,1} = \frac{4 \cdot 6.08\mu m}{3 \cdot 6.76} = 1.2\ \mu m \to$ ✓

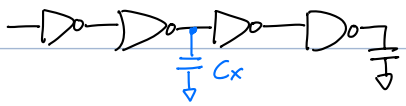
E': add 2 ▷○ → $f_{op} = \sqrt[5]{F} \approx 3.15$

$\tau'_{delay} \cdots = 756$ ps

(if other gate added, $g\uparrow \to f\uparrow$)



Branching (fanout)

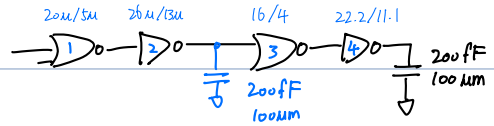


Add C_x → h no longer const. $H = \frac{C_L}{C_{in4}} \frac{C_{in4}}{C_{in3}} \frac{C_{in3}+C_x}{C_{in2}} \cdot \frac{C_{in2}}{C_{in1}}$

$= \frac{C_L}{C_{in1}} \cdot \frac{C_{in3}+C_x}{C_{in3}}$ (if $C_{in3} >> C_x$, ideal)

logical G unaffected

E.

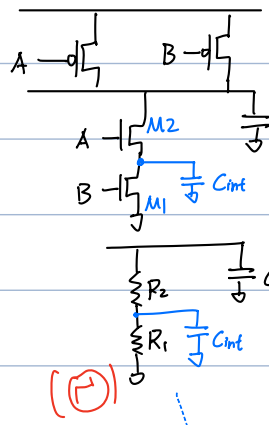


$C_g = 2 \frac{fF}{\mu m}$

size up the following gate

$f_1 \doteq 2.6$, $f_2 \doteq 3$, $f_3 \doteq 3$, $f_4 \doteq 3$ w/o C_x	11.7 τ_{inv}
if $1.5 \times f_3 \rightarrow f_2' = 3.25$, $f_3' = 1.83$, better! (want C_{in3} ↑)	10.78 τ_{inv}
↳ 4: 30/15 → $f'_{2:4} = 3.25, 2.5, 2.27$, ✓✓	10.63 τ_{inv}

Stacks



if A=1, B = 0→1

C_{int} needs to be discharged first before M2 can be on

B slower, delay ∝ (stack level)^2

C_{int} only discharges via R_1.

$T_{elmore} = R_1 C_{int} + (R_1 + R_2) C_{out}$

(add $R_i C_i$ for each dev. in stack)

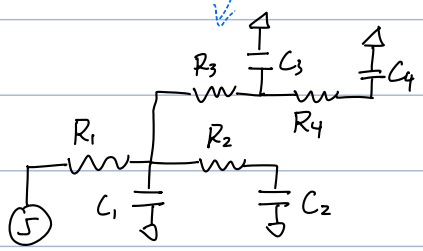
$T_D = \sum_{node\,i} R_{i\text{-to-src}} C_i$

only valid for 1. single driving node &

2. all caps are gnded (path to R gnd)

3. no resistive loops

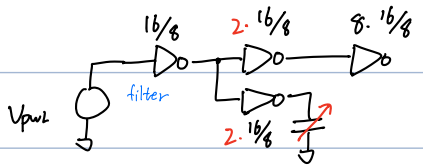
E.



$T_D = C_1 R_1 + C_2 (R_1 + R_2) + C_3 (R_1 + R_3) + C_4 (R_4 + R_3 + R_1)$

PS3 C

want FO4

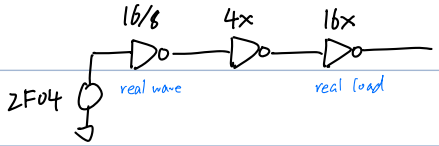


slew time = 2 FO4

vary C_L to get same delay

$C_g = \frac{C_L}{8(16+8)\mu m}$

R



$R_{nfet} = \frac{t_{pHL}}{C_L W_L} \times W_{nfet}$

$R_{pfet} = \frac{t_{pLH}}{C_L W_L} \times W_{pfet}$