520.492 Mixed-Signal VLSI Systems

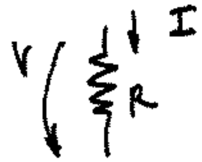
Week 10

Low Power Techniques

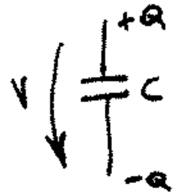
References

1. Franca and Tsividis: Chapter 3.
2. C. Mead and L. Conway, *Introduction to VLSI Systems*, Addison-Wesley, 1980.

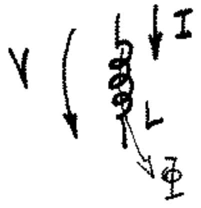
ENERGY CONSUMPTION AND POWER DISSIPATION



$$P_{diss.} = I \cdot V = R I^2 = \frac{V^2}{R}$$

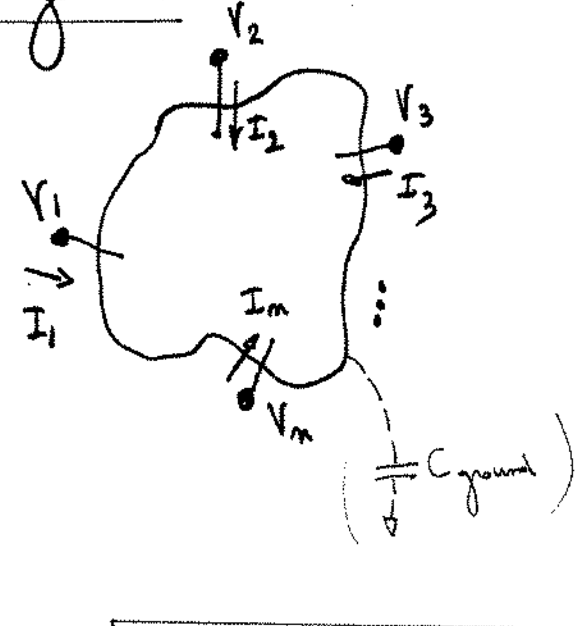


$$\begin{cases} E_{pot.} = \dfrac{CV^2}{2} = \dfrac{QV}{2} = \dfrac{Q^2}{2C} \quad \text{ELECTROSTATIC ENERGY} \\ P_{transient} = \dfrac{dE_{pot}}{dt} = C\,V\,\dfrac{dV}{dt} = V\,\dfrac{dQ}{dt} = V \cdot I \end{cases}$$



$$\begin{cases} E_{pot.} = \dfrac{LI^2}{2} = \dfrac{\Phi \cdot I}{2} = \dfrac{\Phi^2}{2L} \quad \text{MAGNETOSTATIC ENERGY} \\ P_{transient} = \dfrac{dE_{pot.}}{dt} = L \cdot I\,\dfrac{dI}{dt} = V \cdot I \end{cases}$$

In general:



$$P_{injected} = \sum_{i=1}^{m} I_i V_i$$

$$= \underbrace{P_{diss.}}_{\text{resistive}} + \underbrace{P_{transient}}_{\text{blind transient}}$$

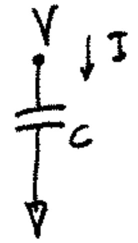
$$\Rightarrow \boxed{P_{diss.} = \langle P_{injected} \rangle = \left\langle \sum_{i=1}^{m} I_i V_i \right\rangle}$$

$\langle P_{transient} \equiv 0 \rangle$

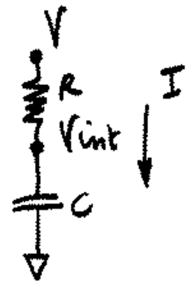
or: $\displaystyle\int_{state_1}^{state_2} P_{transient}(t)\,dt = 0$

where $state_2$ and $state_1$ equal but different times

Caveat: **DYNAMIC POWER DISSIPATION**



$$I = C\frac{dV}{dt} \Rightarrow P_{injected} = \frac{d}{dt}\left(\frac{CV^2}{2}\right) \Rightarrow P_{diss.} = \langle P_{injected}\rangle \; \underline{\underline{= 0}}$$



$$I = C\frac{dV_{int}}{dt} \quad \text{and} \quad V = V_{int} + R.I$$

$$\Rightarrow P_{injected} = I.V = R.I^2 + \frac{d}{dt}\left(\frac{C V_{int}^2}{2}\right)$$

$$\Rightarrow P_{diss.} = \langle P_{injected}\rangle = \underline{\underline{R.\langle I^2\rangle \neq 0}}$$

Fourier domain : $I = j\omega C\, V_{int} = j\omega C\,(V - RI)$

$$\Rightarrow I(\omega) = \frac{j\omega C}{1 + j\omega\tau} V(\omega) \quad \text{with} \quad \tau = RC$$

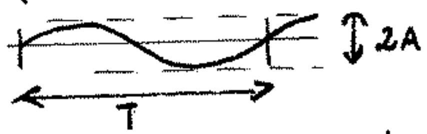
Parseval: $\langle I^2\rangle = \frac{1}{2}\int_{-\infty}^{+\infty} |I(\omega)|^2 \, d\omega$ (Wiener - Khintchine special case)

$$\Rightarrow P_{diss.} = R.\,\frac{1}{2}\int_{-\infty}^{+\infty} \frac{C^2\omega^2}{1+\tau^2\omega^2}|V(\omega)|^2 \, d\omega$$

$$= C.\frac{1}{\tau}.\underbrace{\left(\frac{1}{2}\int_{-\infty}^{+\infty} \frac{\omega'^2}{1+\omega'^2}.|V(\omega')|^2 \, d\omega'\right)}_{\text{depends on RELATIVE FREQUENCY CONTENT of } V \text{ signal}}$$

change of variable $\omega' = \tau.\omega$
normalized frequency

<u>example 1</u>: $V =$ harmonic $\qquad V(\omega') = \underset{\text{amplitude}}{A} \cdot \delta(\overbrace{2\pi \frac{\tau}{T}}^{\omega})$

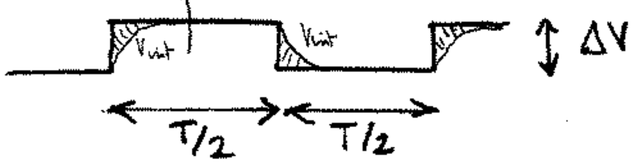


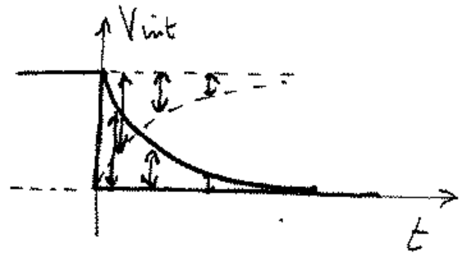
$$\Rightarrow P_{diss.} = C \cdot \frac{1}{\tau} \cdot \frac{\omega'^2}{1+\omega'^2} \cdot \underbrace{\frac{1}{2} A^2}_{V_{rms}^2}$$

$$T \gg \tau \Rightarrow P_{diss.} \approx C V_{rms}^2 \cdot \frac{4\pi^2 \tau}{T^2}$$

$$\propto \boxed{C V_{pp}^2 \cdot f^2}$$

<u>example 2</u>: $V =$ square wave





$$I = \frac{V - V_{int}}{R} = \pm \frac{1}{R} \Delta V \cdot e^{-\frac{t}{\tau}}$$

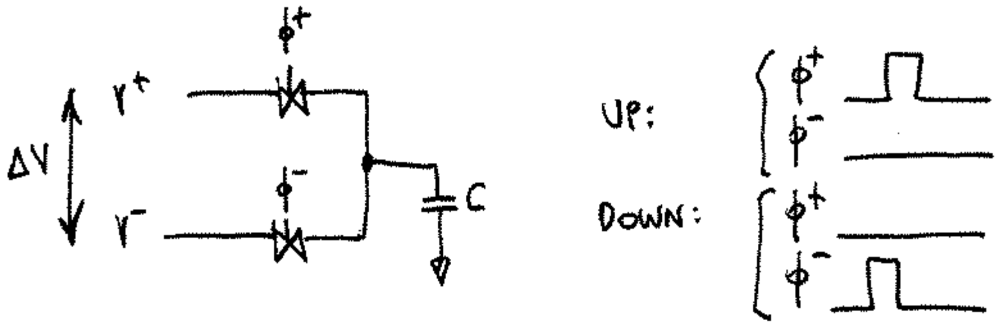
$$P = \langle R I^2 \rangle = \frac{2}{T} \frac{1}{R} \Delta V^2 \underbrace{\int_0^\infty e^{-\frac{2t}{\tau}} dt}_{\tau/2} = \frac{\cancel{2}}{T} \frac{1}{\cancel{R}} \Delta V^2 \frac{\cancel{R} C}{\cancel{2}}$$

$$= C \Delta V^2 \cdot \frac{1}{T}$$

(Why the difference?) $\qquad \propto \boxed{C V_{pp}^2 \cdot f}$

example 3: cyclic charging of a large load capacitance:

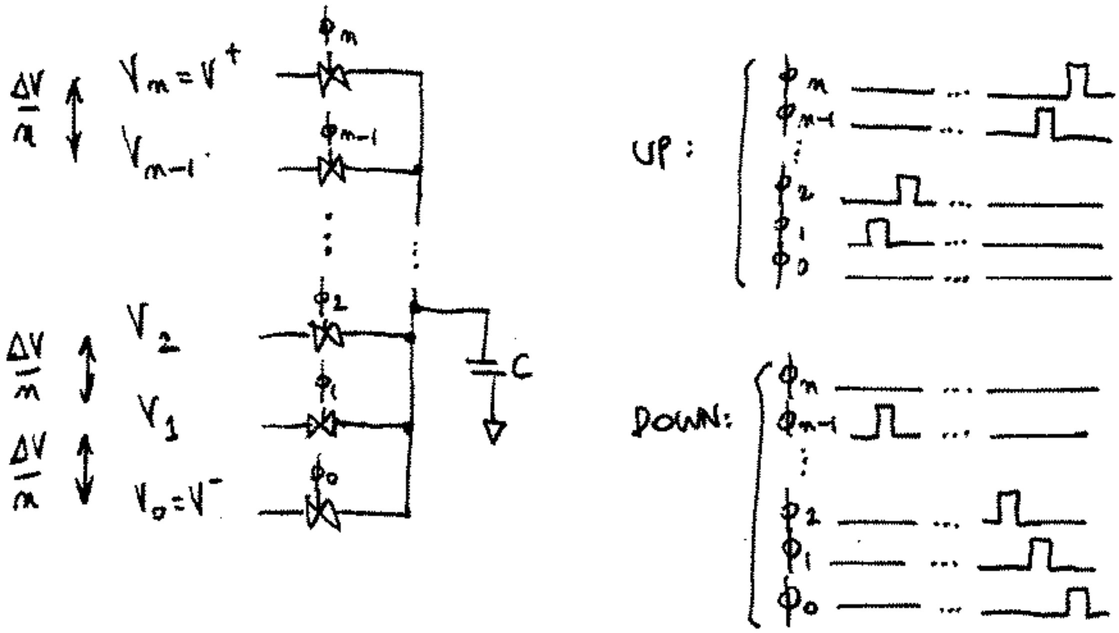
a)



Dissipated energy per UP or DOWN charging:

$$E_{diss}^{UP} = E_{diss}^{DOWN} = \frac{1}{2} C (V^+ - V^-)^2 = \frac{1}{2} C \Delta V^2$$

b)



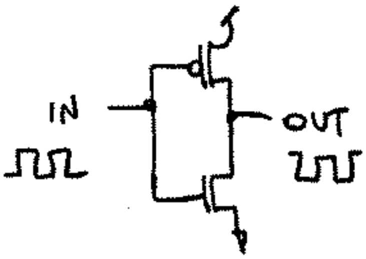
Dissipated energy per complete UP or DOWN charging:

$$E_{diss}^{UP} = E_{diss}^{DOWN} = \sum_{i=1}^{n} \frac{1}{2} C (V_i - V_{i-1})^2 = n \frac{1}{2} C \left(\frac{\Delta V}{n}\right)^2$$

$$= \frac{1}{2n} C \Delta V^2$$

which is supplied by V_n (V^+) and V_0 (V^-) only on average.

LOW POWER TECHNIQUES FOR DIGITAL CMOS



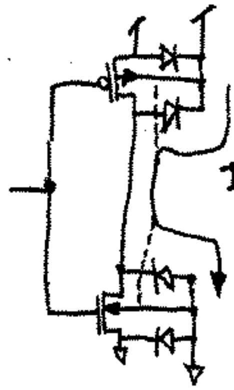
$$P_{diss.} = P_{static} + P_{transient}$$

P_{static} → LEAKAGE: junction diode + boltzmann MOS

$P_{transient}$ → DYNAMIC + SHOOT-THROUGH

- $P_{static} = I_{leak} \cdot V_{dd}$

  with $I_{leak} = I_{junction} + I_{boltzmann}$:



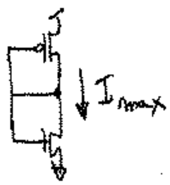
$I_{junction} \div A_D$

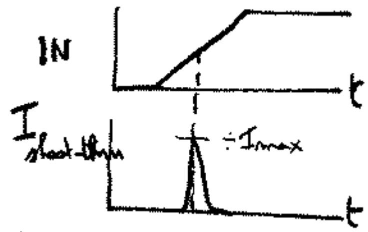
$I_{boltzmann} = I_0 \div e^{-\frac{V_{th}}{V_T}}$

- $P_{transient} = 2 \alpha f \cdot (E_{shoot\text{-}thru} + E_{dyn.})$

  with $E_{shoot\text{-}thru} \div I_{max} \cdot V_{dd} \cdot t_{rise}$

  $E_{shoot\text{-}thru}$: depends on $\frac{V_{th}}{V_{dd}}$





and $E_{dyn} = \frac{1}{2} C V^2$

and α = activity (average # of transitions per rising or falling clock edge)

and f = clock frequency

Optimization:

$$\downarrow P_{diss.} = \big(\underbrace{I_{junction}}_{A_D \downarrow} + \underbrace{I_{Boltzmann}}_{V_{th} \uparrow}\big) \underbrace{V_{dd}}_{V_{dd} \downarrow} + \underbrace{a}_{a \downarrow}\, f \cdot \big(\underbrace{C}_{C \downarrow} \underbrace{V_{dd}^2}_{V_{dd} \downarrow} + 2 \underbrace{E_{shoot-thru}}_{\frac{t_{rise}}{T} \downarrow \text{ and } V_{th} \downarrow V_{dd} \downarrow}\big)$$

$$\downarrow E_{diss.} \text{ (per cycle)} = P_{diss.} T = \frac{P_{diss}}{f} \longrightarrow f \uparrow$$

$$\downarrow t_{prop} = f\big(\underbrace{V_{dd}}_{V_{dd}\uparrow}, \underbrace{V_{th}}_{V_{th}\downarrow}, \underbrace{C}_{C\downarrow}, \underbrace{\beta}_{\beta\uparrow} = \mu C_{ox}\tfrac{W}{L}, \kappa, \ldots\big) < T \; \left(= \tfrac{1}{f}\right)$$

for above threshold: $\frac{1}{2}\beta(\kappa V_{dd} - V_{th})^2 \cong C \frac{V_{dd}}{t_{prop}} \Rightarrow t_{prop} \cong \frac{2\,C\,V_{dd}}{\beta(\kappa V_{dd} - V_{th})^2}$

⇒ REDUCE C

method: feature size $\lambda \downarrow$ $\quad C \div \lambda$ or λ^2 — line capacitance / gate capacitance / drain/source cap.

limit: technology, yield, soft errors, quantum effects

⇒ REDUCE V_{dd} AND V_{th}

$P_{diss} \downarrow$ and $t_{prop} \uparrow$ but $E_{diss.} \approx a C V_{dd}^2 \downarrow$ as V_{dd}^2

method:
- decentralize into parallel architecture
  (to allow $f \downarrow$ in order to keep $T > t_{prop}$)
- adjust V_{th} proportional to V_{dd}, by bulk tuning

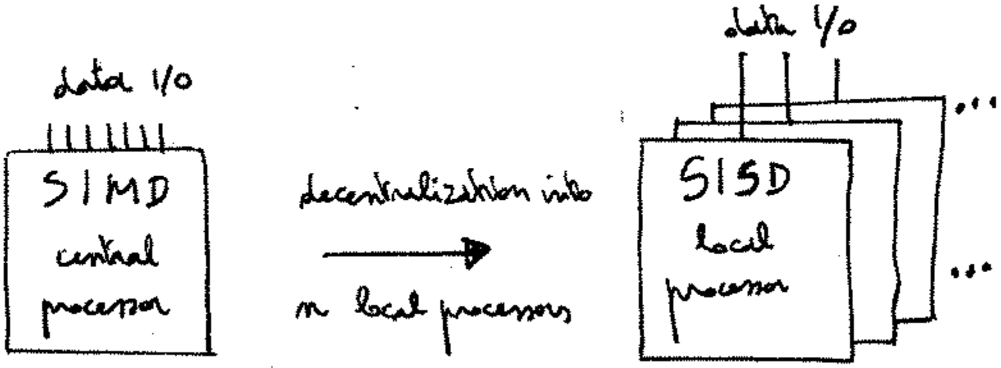
limit: relative contribution of P_{static}, because $f \downarrow$ and $V_{th} \downarrow$

⇒ REDUCE a (without compromising computational bandwidth)

method:
- minimize # transistors for a given function
- power off non-active blocks
- disable non-active branches of clock input lines

Decentralization of the computational task

PARALLEL (DISTRIBUTED) COMPUTING :



$\alpha f \quad \longrightarrow \quad \alpha f / m$

$V_{dd} \quad \longrightarrow \quad V_{dd} / m$

$V_{th} \quad \longrightarrow \quad V_{th} / m$

$$\tau_{prop} \propto \frac{V_{dd}}{(kV_{dd}-V_{th})^2}$$

$$P_{dyn} = \alpha f C V_{dd}^2 \longrightarrow P_{dyn} = m \times \frac{\alpha f}{m} C \left(\frac{V_{dd}}{m}\right)^2$$

$$= \frac{\alpha f C V_{dd}^2}{m^2}$$

$$\text{area} = A \longrightarrow \text{area} = A \times m$$

<u>limits</u>: STATIC POWER DISSIPATION :

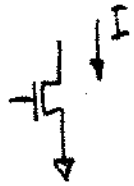
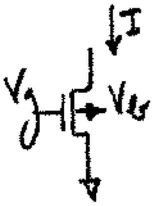
$$P_{stat} = \left(I_{junction} + I_0 e^{-\frac{V_{th}}{V_T}}\right) V_{dd} \longrightarrow P_{stat} = \cancel{m}\left(I_{junction} + I_0 e^{-\frac{V_{th}}{mV_T}}\right) \frac{V_{dd}}{\cancel{m}}$$

$$e^{-\frac{V_{th}}{mV_T}} \uparrow\uparrow$$

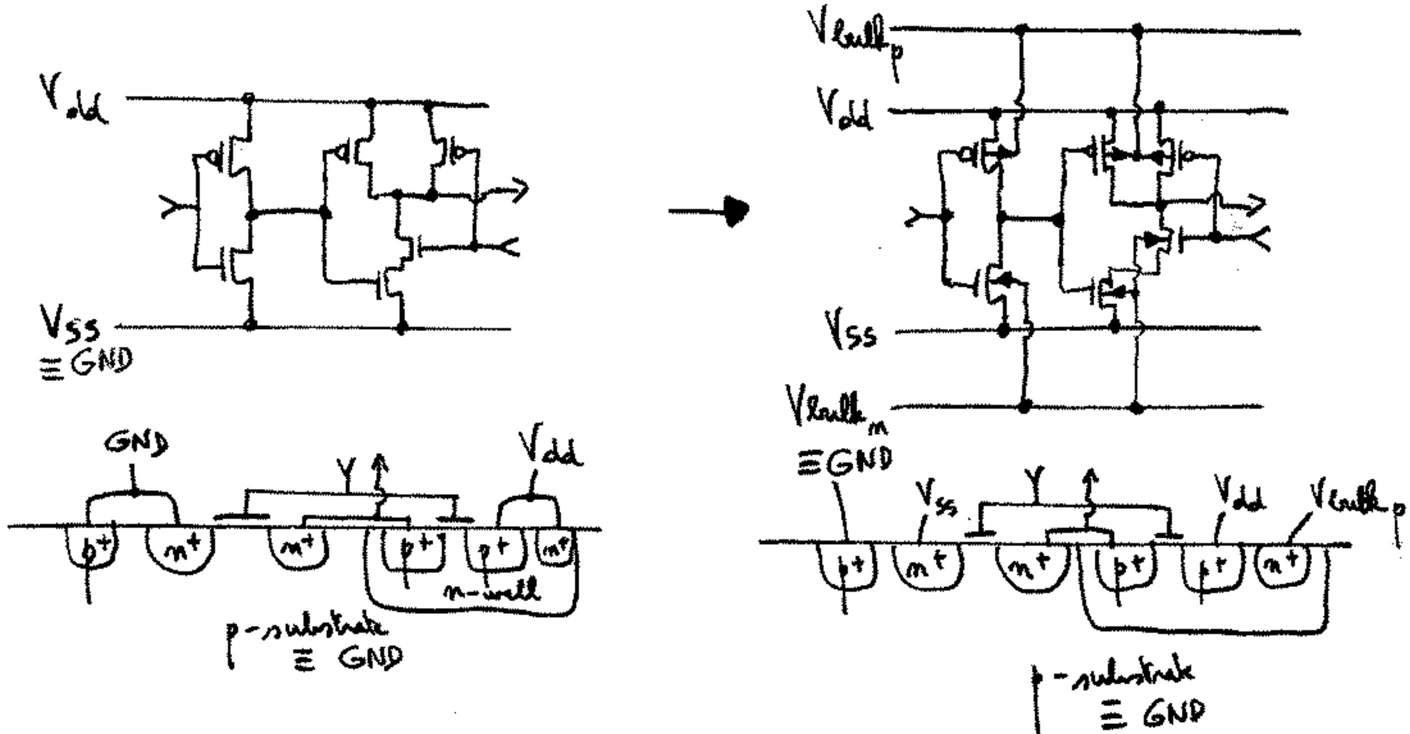
Substrate modulation for external adjustment of V_{th}:

(Stanford Ultra Low Power CMOS)

V_{th} is not adjustable after fabrication, but can effectively be modulated through the body effect, by the V_{bulk} voltage:

V_g  I $\quad I = f(\kappa V_g - V_{th}) \;\rightarrow\;$ V_g  V_b I $\quad I = f(\kappa V_g + \underbrace{(1-\kappa)V_{bulk} - V_{th}}_{-V_{th}^{eff}})$

$$\Rightarrow \quad \underline{V_{th}^{eff} = V_{th} - (1-\kappa)V_{bulk}}$$

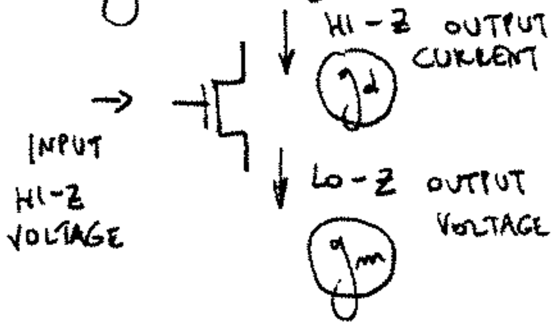


NOTE: • $V_{bulk_p} > V_{dd}$ and $V_{bulk_n} < V_{ss}$, but V_{bulk_n} and V_{bulk_p} dissipate little (no) dynamic power

• $V_{bulk_n} \equiv GND$ identical for all gates $\Rightarrow$ local adjustment is not possible, unless in a twin-tub process

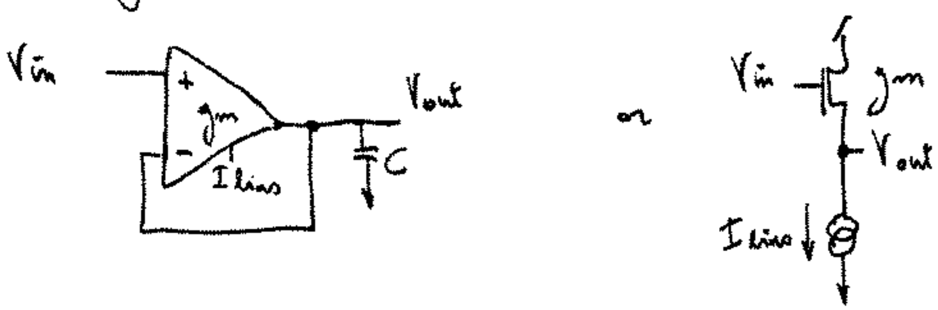
LOW-POWER TECHNIQUES FOR ANALOG CMOS

Basic function of MOS transistor:



- TRANSCONDUCTANCE (g_m)
- VOLTAGE GAIN ($A_v = \frac{g_m}{g_d}$)

example: <u>follower</u>



$$I_{out} = g_m (V_{in} - V_{out}) - g_d V_{out} = C \frac{dV_{out}}{dt}$$

time constant : $\tau = C/(g_m+g_d) \approx C/g_m$

power : $P = I_{bias} \cdot V_{dd}$

power-delay product (typical) : $\boxed{P \cdot \tau = \frac{I_{bias}}{g_m} \cdot C V_{dd}}$ $\downarrow$

$\Rightarrow$
- increase $\frac{g_m}{I_{bias}}$
- decrease V_{dd} and C

1) lower C and V_{dd}

<u>limit</u>: SNR : $\frac{\text{range}}{\text{SNR}} \gg \sqrt{\frac{kT}{C}}$ and $V_{dd} > \text{range}$

how much depends on circuit techniques

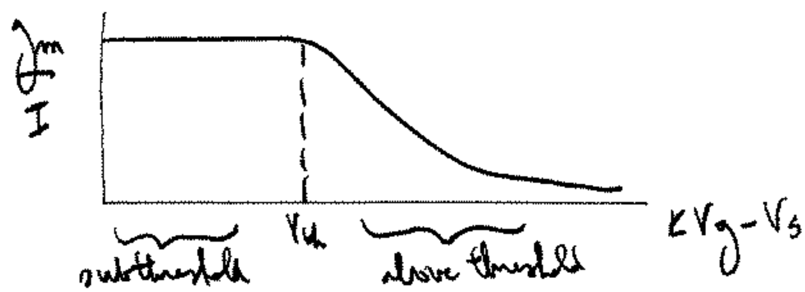
2) increase g_m / I : relative transconductance (units V^{-1})

- operational:

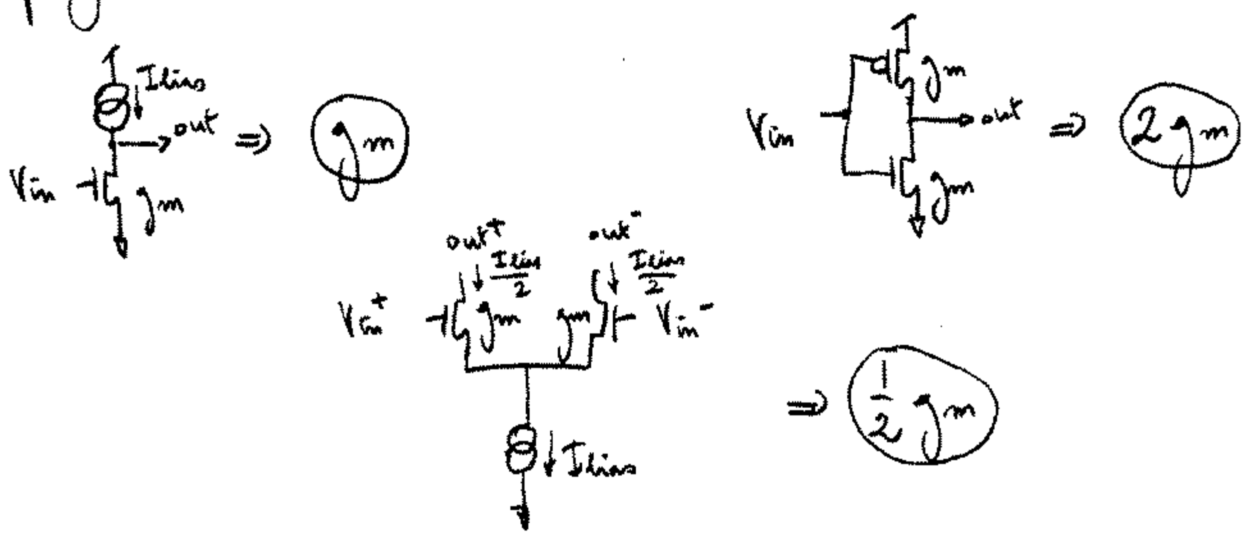
sub-threshold: $I \cong I_0 e^{\kappa V_g / V_T \cdots} \Rightarrow g_m = I_0 \frac{\kappa}{V_T} e^{\cdots} \Rightarrow \frac{g_m}{I} = \frac{\kappa}{V_T}$

above threshold: $I \cong \frac{\beta}{2} (\kappa V_g - V_S - V_{th})^2 \Rightarrow g_m = \beta \kappa (\kappa V_g - V_S - V_{th}) \Rightarrow$

$$\frac{g_m}{I} = \frac{\kappa}{\kappa V_g - V_S - V_{th}}$$

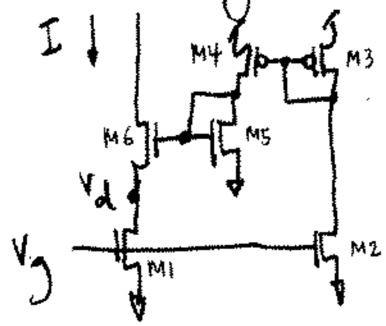


- topological:



Low-voltage techniques (a few of them)

- low-voltage cascode:



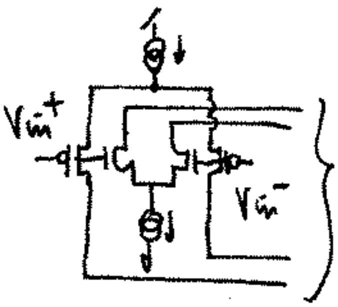
$$V_d = V_T \log\left(\frac{M_2 \cdot M_4 \cdot M_6}{M_1 \cdot M_3 \cdot M_5}\right)$$

$> 5\,V_T$ for saturation of M1

$$\Rightarrow \frac{M_2\, M_4\, M_6}{M_1\, M_3\, M_5} \gtrsim 100$$

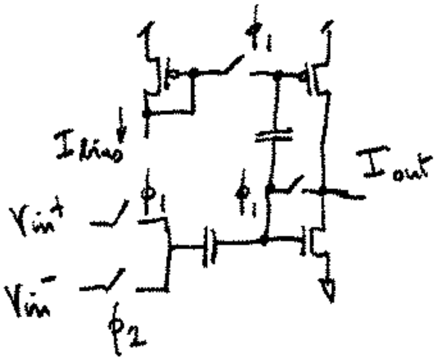
- low-voltage differential transamp:

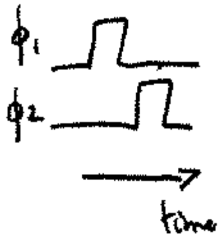
→ combine n- and p-type diff pairs:



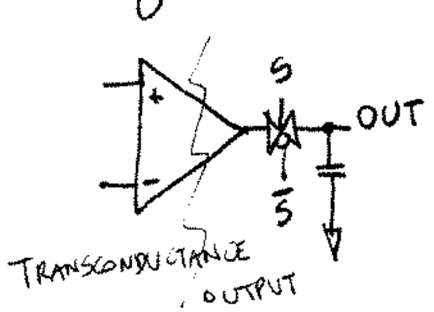
} to current mirrors

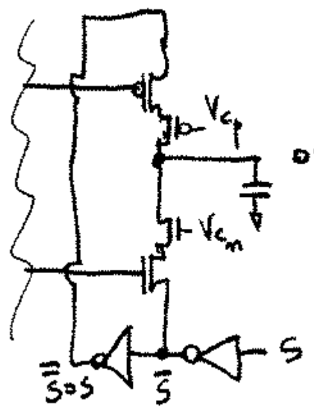
→ OR: dynamic:





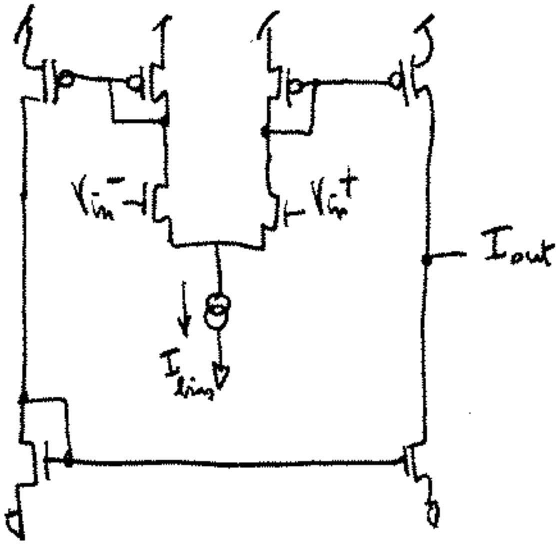
- low-voltage active switch:

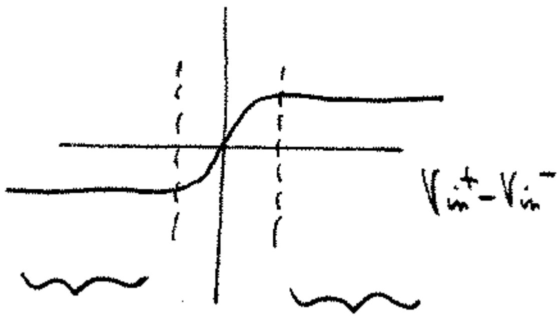




DYNAMIC BIASING TECHNIQUE :

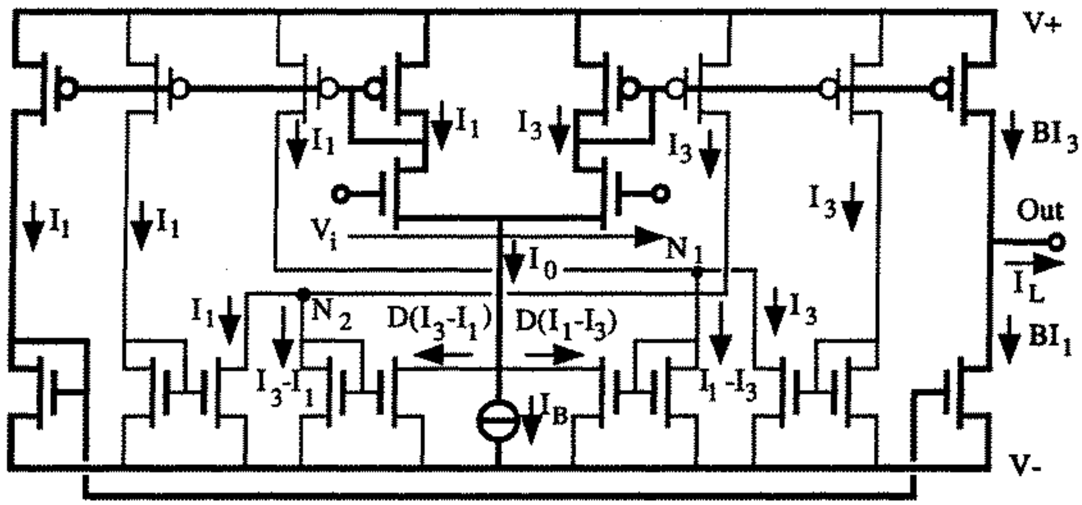
adjust I_{bias} to the needs of the load





here also

here, need to increase I_{bias}



$$-I_L = B\,(I_1 - I_3) = BI_B \frac{\tanh(V_i / 2nU_T)}{1 - D\,|\tanh(V_i / 2nU_T)|}$$