# MOSFET Scaling

Device scaling: Simplified design goals/guidelines for shrinking device dimensions to achieve density and performance gains, and power reduction in VLSI.

Issues: Short-channel effect, Power density, Switching delay, Reliability.

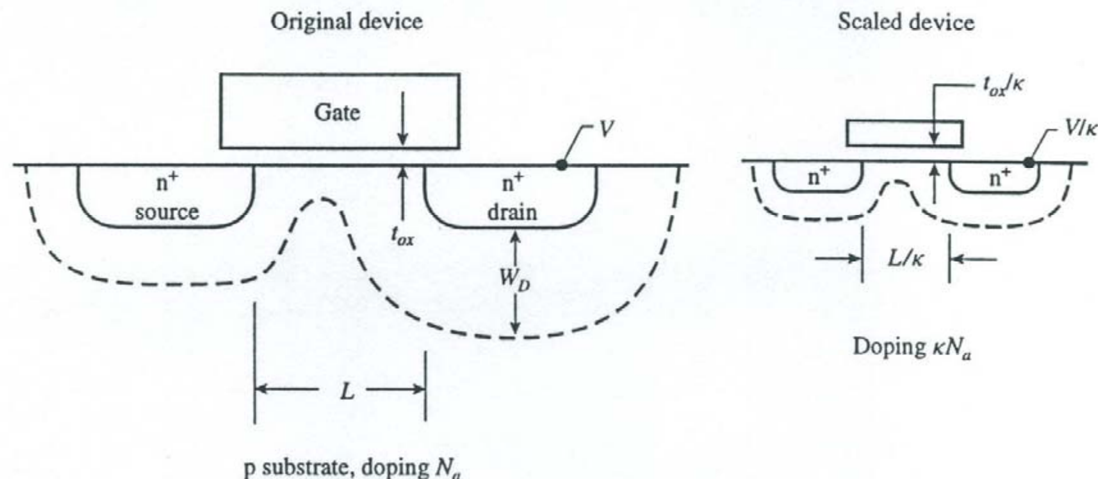



**FIGURE 4.1.** Principles of MOSFET constant-electric-field scaling. (After Dennard, 1986.)

***The principle of constant-field scaling lies in scaling the device voltages and the device dimensions (both horizontal and vertical) by the same factor, $\kappa$ (> 1), such that the electric field remains unchanged.***

# Rules of Constant Field Scaling

| | MOSFET Device and Circuit Parameters | Multiplicative Factor ($\kappa > 1$) |
|---|---|---|
| Scaling assumptions | Device dimensions ($t_{ox}$, $L$, $W$, $x_j$) | $1/\kappa$ |
| | Doping concentration ($N_a$, $N_d$) | $\kappa$ |
| | Voltage ($V$) | $1/\kappa$ |
| Derived scaling behavior of device parameters | Electric field ($\mathrm{E}$) | 1 |
| | Carrier velocity ($v$) | 1 |
| | Depletion layer width ($W_d$) | $1/\kappa$ |
| | Capacitance ($C = \varepsilon A/t$) | $1/\kappa$ |
| | Inversion layer charge density ($Q_i$) | 1 |
| | Current, drift ($I$) | $1/\kappa$ |
| | Channel resistance ($R_{ch}$) | 1 |
| Derived scaling behavior of circuit parameters | Circuit delay time ($\tau \sim CV/I$) | $1/\kappa$ |
| | Power dissipation per circuit ($P \sim VI$) | $1/\kappa^2$ |
| | Power-delay product per circuit ($P \times \tau$) | $1/\kappa^3$ |
| | Circuit density ($\propto 1/A$) | $\kappa^2$ |
| | Power density ($P/A$) | 1 |

# Scaling of Depletion Width

Maximum drain depletion width: $W_D = \sqrt{\dfrac{2\varepsilon_{si}(\psi_{bi} + V_{dd})}{qN_a}}$

For $N_a \rightarrow \kappa N_a$ and $V_{dd} \rightarrow V_{dd}/\kappa$,
$W_D \rightarrow W_D/\kappa$ if $V_{dd} >> \psi_{bi}$.

However, the source depletion width, $W_S = \sqrt{\dfrac{2\varepsilon_{si}\psi_{bi}}{qN_a}}$

is indep. of $V_{dd}$ and only scales as $W_S \rightarrow W_S/\sqrt{\kappa}$ .

Furthermore, the maximum gate depletion width,

$$W_{dm}{}^0 = \sqrt{\frac{4\varepsilon_{si} kT \ln(N_a / n_i)}{q^2 N_a}}$$

scales even less than $1/\sqrt{\kappa}$.

# Generalized Scaling

Allows electric field to scale up by $\alpha$ ($\mathcal{E} \rightarrow \alpha\mathcal{E}$) while the device dimensions scale down by $\kappa$,
i.e., voltage scales by $\alpha/\kappa$ ($V \rightarrow (\alpha/\kappa)V$).

More flexible than constant-field scaling,
but has reliability and power concerns.

To keep Poisson's equation invariant under the transformation, $(x,y) \rightarrow (x,y)/\kappa$ and $\psi \rightarrow \psi/(\kappa/\alpha)$ within the depletion region:

$$\frac{\partial^2(\alpha\psi/\kappa)}{\partial(x/\kappa)^2} + \frac{\partial^2(\alpha\psi/\kappa)}{\partial(y/\kappa)^2} = \frac{qN_a}{\varepsilon_{si}}$$

$N_a$ should be scaled to $(\alpha\kappa)N_a$.

# Constant Voltage Scaling

Special case of $\alpha=\kappa$ in generalized scaling:

The only mathematically correct scaling as far as 2D Poisson eq. and boundary conditions are concerned.

$N_a \rightarrow \kappa^2 N_a$,
therefore, the maximum depletion width, $W_{dm}{}^0 = \sqrt{\frac{4\varepsilon_{si} kT \ln(N_a / n_i)}{q^2 N_a}}$
scales down by $\kappa$.

Both the short-channel $V_t$ roll-off, $\Delta V_t = \frac{24 t_{ox}}{W_{dm}} \sqrt{\psi_{bi}(\psi_{bi} + V_{ds})} e^{-\frac{\pi L/2}{W_{dm}+3t_{ox}}}$

and the threshold voltage, $V_t = V_{fb} + 2\psi_B + \frac{\sqrt{2\varepsilon_{si} q N_a (2\psi_B + V_{bs})}}{C_{ox}}$

remain unchanged for constant-voltage scaling.

However, it is physically incorrect since
$\mathcal{E} \rightarrow \kappa\mathcal{E}$ (reliability) and $P/A \rightarrow \kappa^{2\text{-}3} P/A$ (power).

# Scaling in Practice

CMOS VLSI technology generations

| Feature Size | Power Supply | Gate Oxide | Oxide Field |
|---|---|---|---|
| 2 μm | 5 V | 350 Å | 1.4 MV/cm |
| 1.2 μm | 5 V | 250 Å | 2.0 MV/cm |
| 0.8 μm | 5 V | 180 Å | 2.8 MV/cm |
| 0.5 μm | 3.3 V | 120 Å | 2.8 MV/cm |
| 0.35 μm | 3.3 V | 100 Å | 3.3 MV/cm |
| 0.25 μm | 2.5 V | 70 Å | 3.6 MV/cm |

CMOS technology has gone through mixed steps of constant voltage and constant field scaling. As a result, field and power density have gone up, but performance gains have been maintained and power per circuit has come down. Fortunately, by physics or by learning, we managed to cope with reliability requirements at higher fields.

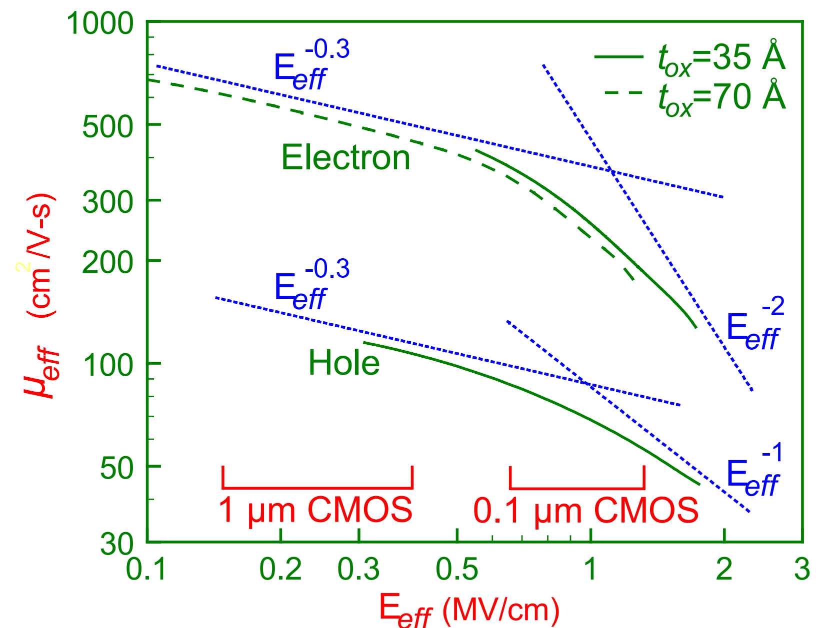
# Non-Scaling Factors

Primary nonscaling factors:

- Built-in potential $\psi_{bi}$ (***Si bandgap***)
- Subthreshold current (***thermal energy kT/q***)

Secondary nonscaling factors (due to higher $\mathcal{E}$):

- Velocity saturation
- Decreased mobility at higher fields
- Oxide reliability ($t_{ox}$ scales less, $W_{dm}$ more)



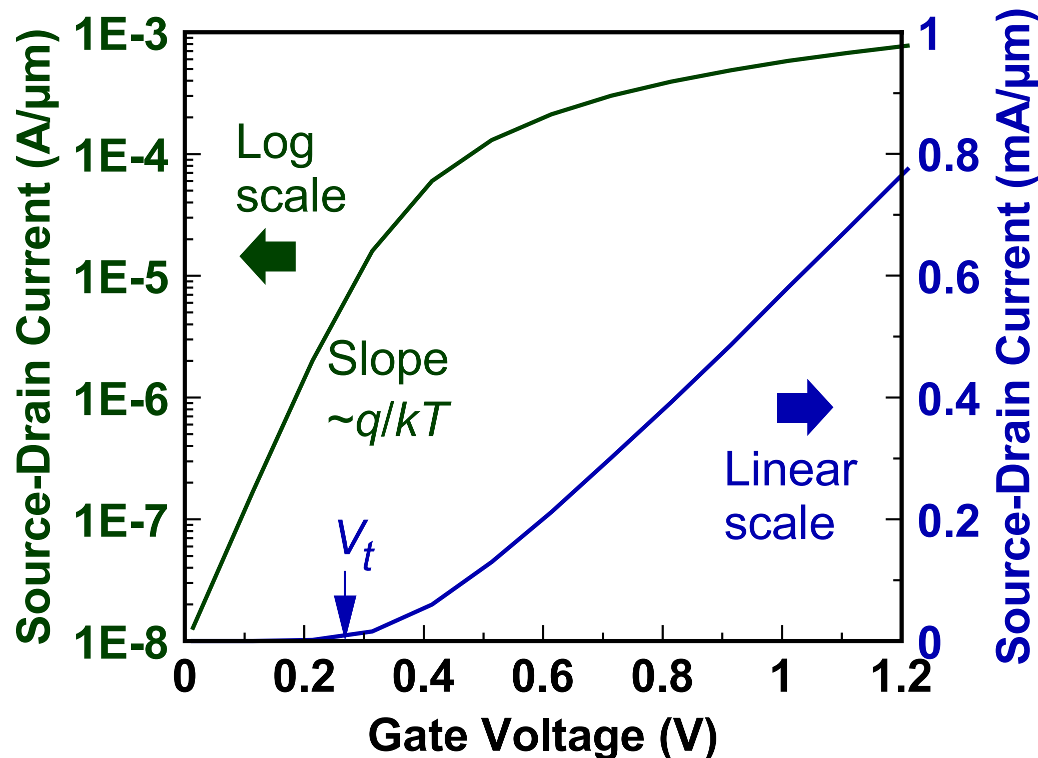
# Other Non-Scaling Factors

- Source and drain series resistance
  - Doping level limited by solid solubility and is not scalable.
  - Doping gradient or junction abruptness limited by annealing process.
- Polysilicon gate depletion
- Inversion layer depth/thickness
- Various process tolerances
  - Gate length.
  - Gate oxide thickness.
  - Dopant number fluctuations.

# MOSFET Threshold Voltage

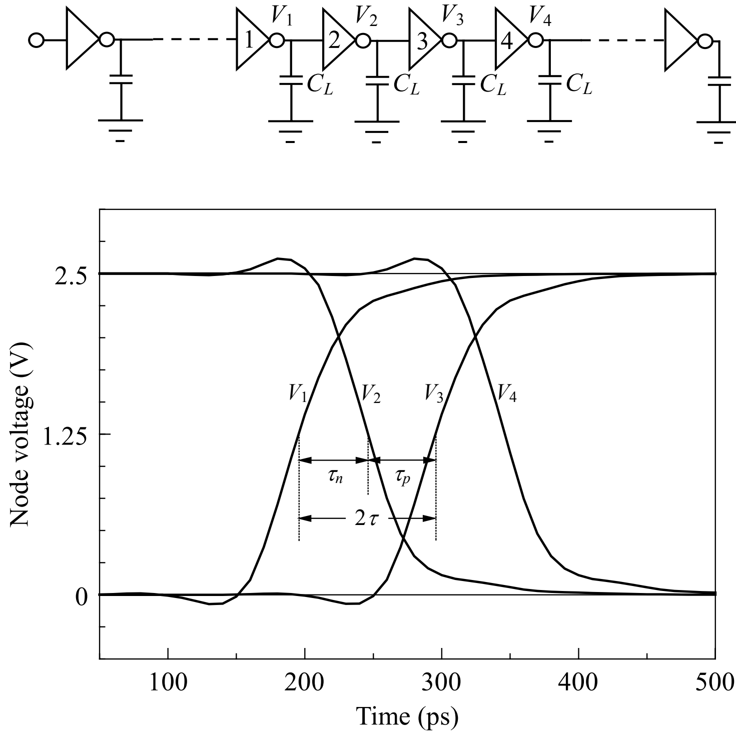
$$V_t = V_{fb} + 2\psi_B + V_{ox} = V_{fb} + 2\psi_B + \frac{Q_d}{C_{ox}}$$

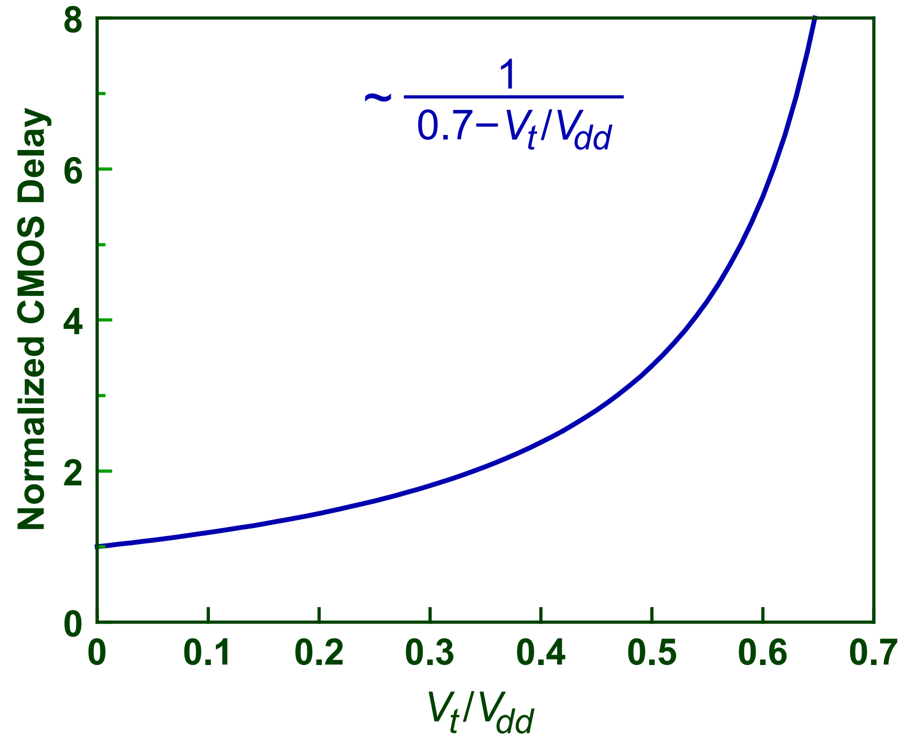
MOSFET current in linear and log scales:




Subthreshold: $I_{ds} = \mu_{eff} C_{ox} \frac{W}{L}(m-1)\left(\frac{kT}{q}\right)^2 e^{q(V_g - V_t)/mkT}\left(1 - e^{-qV_{ds}/kT}\right)$

# CMOS Circuit Delay




Delay ~ CV/I

Desirable to keep $V_t/V_{dd} < 0.3$.

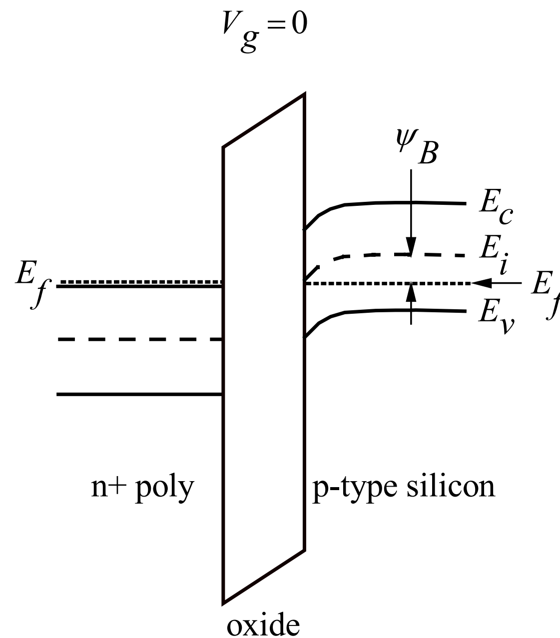
# Choice of Gate Work Function

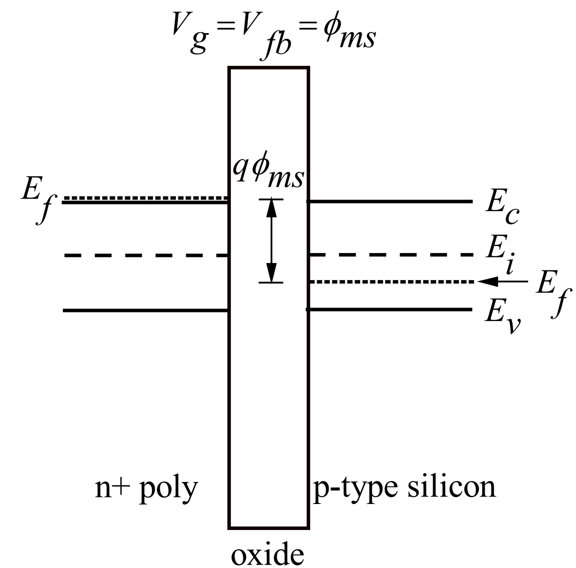
$$V_t = V_{fb} + 2\psi_B + V_{ox} = V_{fb} + 2\psi_B + \frac{Q_d}{C_{ox}}$$

$2\psi_B$ ~ 1 V, need $V_{fb}$ ~ -1 V to obtain low $V_t$;
i.e., n$^+$ poly gate on nMOSFET and vice versa.

$$\phi_{ms} = -\frac{E_g}{2q} - \psi_B$$

$$= -0.56 - \frac{kT}{q}\ln\left(\frac{N_a}{n_i}\right)$$



# Threshold Voltage Adjustment

In a uniformly doped MOSFET, the maximum gate depletion width (long-channel),

$$W_{dm}{}^{0} = \sqrt{\frac{4\varepsilon_{si}\psi_B}{qN_a}}$$

and the threshold voltage,

$$V_t = V_{fb} + 2\psi_B + \frac{\sqrt{4\varepsilon_{si}qN_a\psi_B}}{C_{ox}} = V_{fb} + (1 + \frac{6t_{ox}}{W_{dm}})2\psi_B$$

are coupled through the parameter $N_a$, and therefore cannot be varied independently (for given $V_{fb}$, $t_{ox}$).

To adjust threshold voltage, it is necessary to employ nonuniform channel doping.

# Nonuniform Channel Doping

1-D Poisson's eq.: $$\frac{d^2\psi}{dx^2} = -\frac{d\mathcal{E}}{dx} = -\frac{\rho(x)}{\varepsilon_{si}}$$

For a nonuniform p-type doping profile *N*(*x*), the electric field is obtained by integrating Poisson's equation once (neglecting mobile carriers):

$$\mathcal{E}(x) = \frac{q}{\varepsilon_{si}}\int_x^{W_d} N(x)dx$$ where $W_d$ is the depletion layer width.

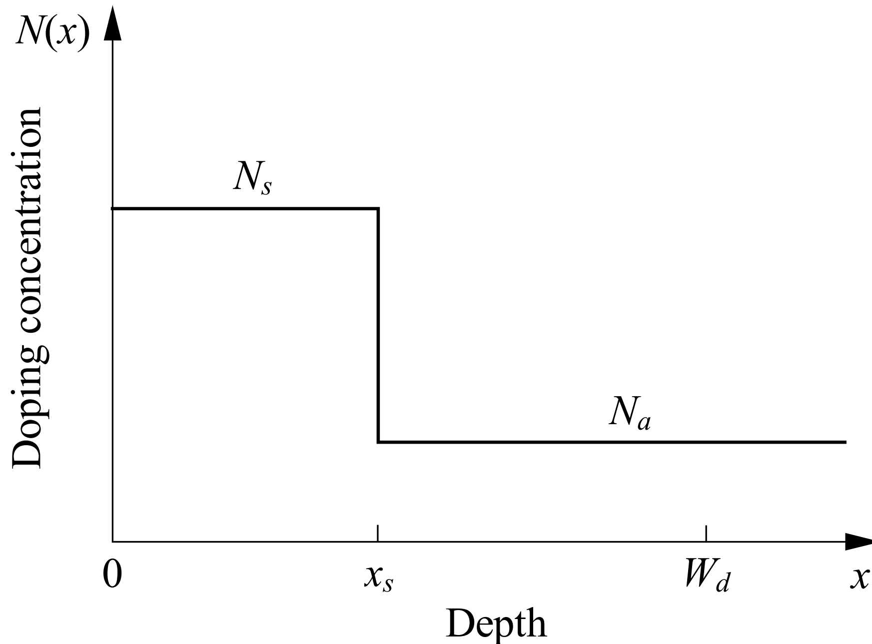
Integrating again, $$\psi_s = \frac{q}{\varepsilon_{si}}\int_0^{W_d}\int_x^{W_d} N(x')dx'dx$$

Integration by parts, $$\psi_s = \frac{q}{\varepsilon_{si}}\int_0^{W_d} xN(x)dx$$

Note that the maximum depletion layer width $W_{dm}^0$ is determined by the condition $\psi_s = 2\psi_B$ when $W_d = W_{dm}^0$.

The threshold voltage of a nonuniformly doped MOSFET is then determined by both the integral (depletion charge density) and the first moment of N(x) within (0, $W_{dm}^0$).

# High-Low Doping Profile




The maximum depletion width at threshold is:

$$W_{dm}^{0} = \sqrt{\frac{2\varepsilon_{si}}{qN_a}\left(2\psi_B - \frac{q(N_s - N_a)x_s^2}{2\varepsilon_{si}}\right)}$$

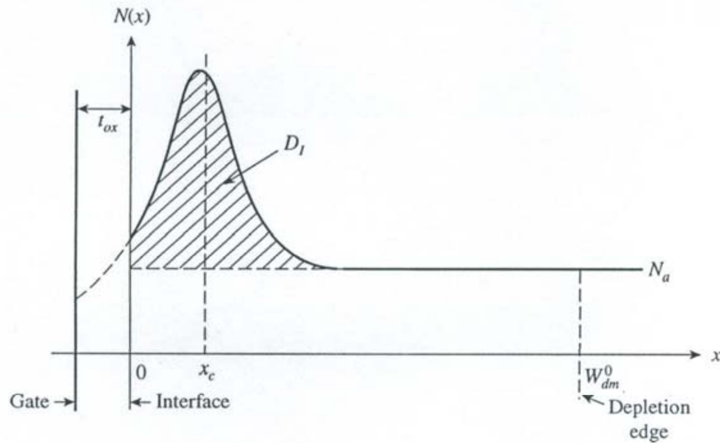
The body-effect factor takes the same form as before:

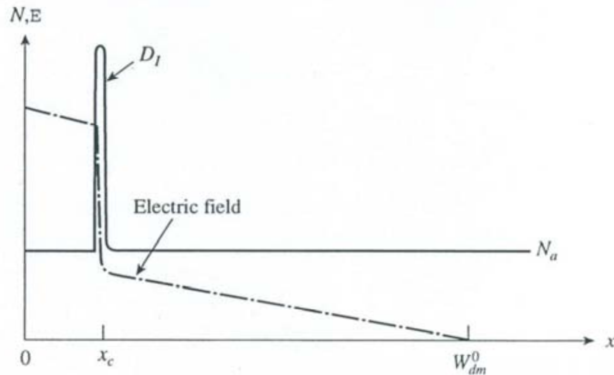
$$m = 1 + \frac{\varepsilon_{si} / W_{dm}^{0}}{C_{ox}} = 1 + \frac{C_{dm}}{C_{ox}} = 1 + \frac{3t_{ox}}{W_{dm}^{0}}$$

The threshold voltage is, again, $V_{fb}+2\psi_B+V_{ox}(=Q_d/C_{ox})$, i.e.,

$$V_t = V_{fb} + 2\psi_B + \frac{1}{C_{ox}}\sqrt{2\varepsilon_{si}qN_a\left(2\psi_B - \frac{q(N_s - N_a)x_s^2}{2\varepsilon_{si}}\right)} + \frac{q(N_s - N_a)x_s}{C_{ox}}$$

# Implanted Gaussian Profile




An implanted Gaussian profile,

$$N(x) = \frac{D_I}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - x_c)^2}{2\sigma^2}\right),$$

is equivalent to a delta-function doping at $x_c$.

Let $(N_s - N_a)x_s = D_I$ and $x_s/2 = x_c$,

Then

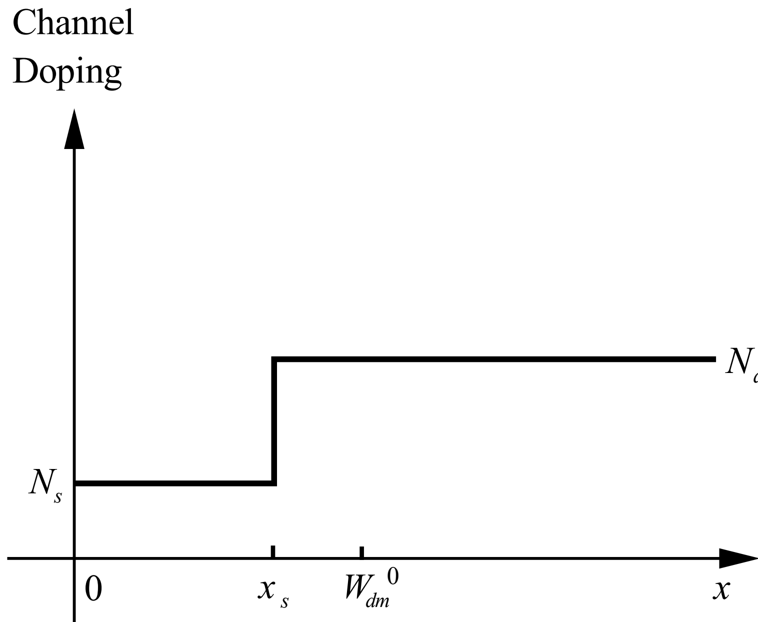
$$W_{dm}{}^0 = \sqrt{\frac{2\varepsilon_{si}}{qN_a}\left(2\psi_B - \frac{qD_I x_c}{\varepsilon_{si}}\right)}$$

and

$$V_t = V_{fb} + 2\psi_B + \frac{qN_a W_{dm}{}^0}{C_{ox}} + \frac{qD_I}{C_{ox}}$$

For shallow surface implants, $x_c = 0$, there is no change in the depletion width. The $V_t$ shift is simply given by $qD_I/C_{ox}$ like a sheet of charge at the silicon-oxide interface.

# Low-High Doping Profile



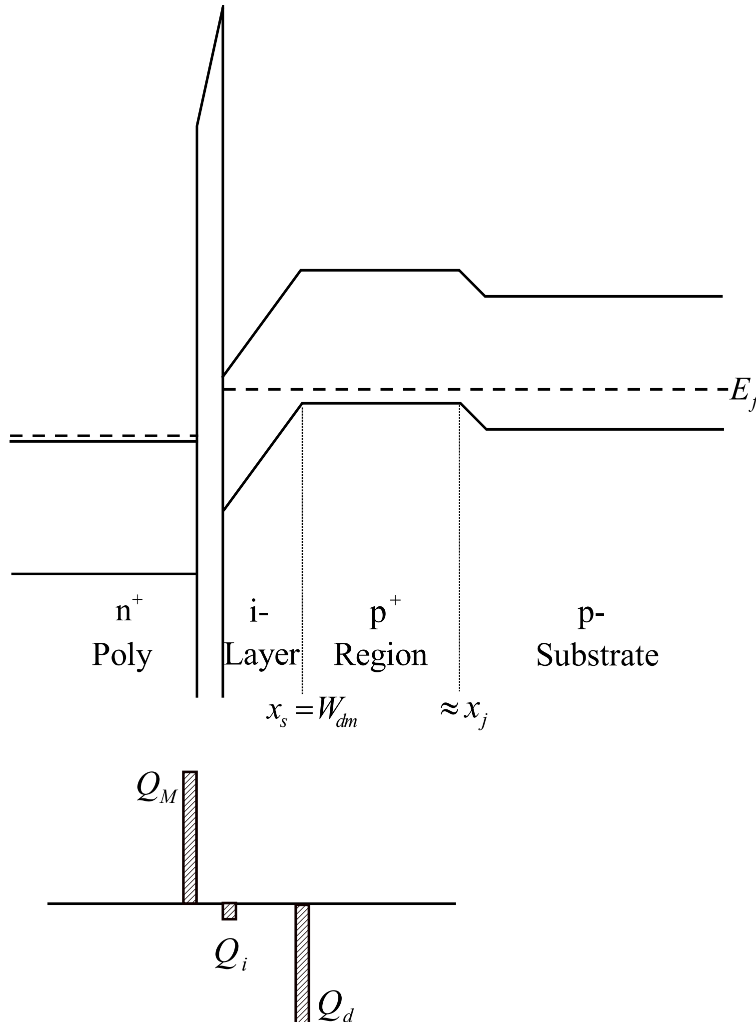

Take $N_s$ = 0, then

$$W_{dm}^{0} = \sqrt{\frac{4\varepsilon_{si}\psi_B}{qN_a} + x_s^{2}}$$

$$V_t = V_{fb} + 2\psi_B + \frac{qN_a}{C_{ox}}\sqrt{\frac{4\varepsilon_{si}\psi_B}{qN_a} + x_s^{2}} - \frac{qN_a x_s}{C_{ox}}$$

In contrary to high-low doping, low-high (retrograde) doping results in a lower $V_t$ than uniform doping for a given $W_{dm}$.

# Extreme Retrograde Profile




For $x_s >> (4\varepsilon_{si}\psi_B/qN_a')^{1/2}$,

$$V_t = V_{fb} + 2\psi_B + \frac{\varepsilon_{si}/x_s}{C_{ox}} 2\psi_B$$

i.e.,

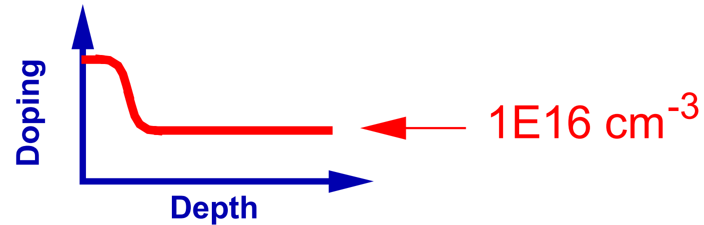
$$V_t = V_{fb} + \left(1 + \frac{3t_{ox}}{W_{dm}}\right) 2\psi_B$$

- The depletion depth is the same as the undoped layer thickness.
- All the depletion charge is located at the far edge of the depletion region.
- Magnitude of the depletion charge is one half of the uniformly doped value.

# Channel Profile Evolution

**> 1 µm CMOS:**

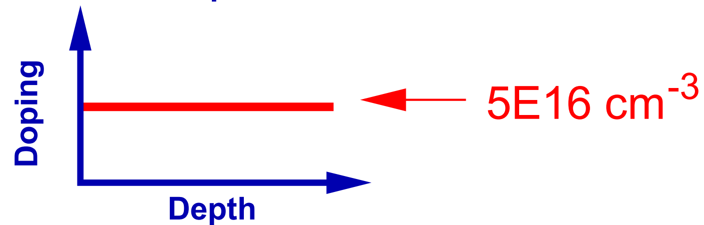
**HIGH-LOW**




1E16 $cm^{-3}$

**0.5 µm CMOS:**

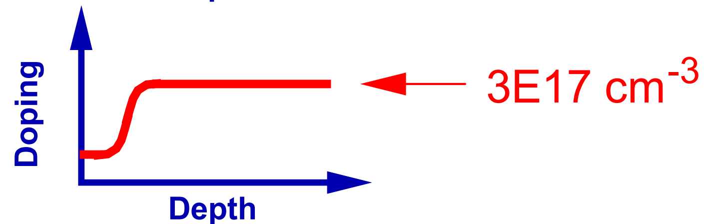
**UNIFORM**




5E16 $cm^{-3}$

**0.2 µm CMOS:**

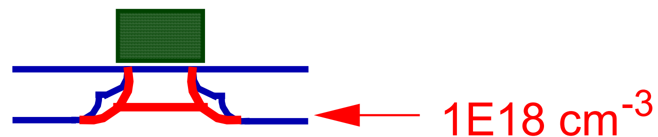
**RETROGRADE**




3E17 $cm^{-3}$

**0.1 µm CMOS:**

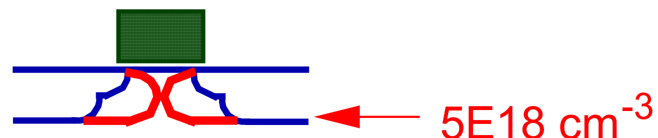
**HALO**



1E18 $cm^{-3}$

**0.05 µm CMOS:**

**SUPER-HALO**



5E18 $cm^{-3}$

# Channel Profile Trends

For uniform channel doping,

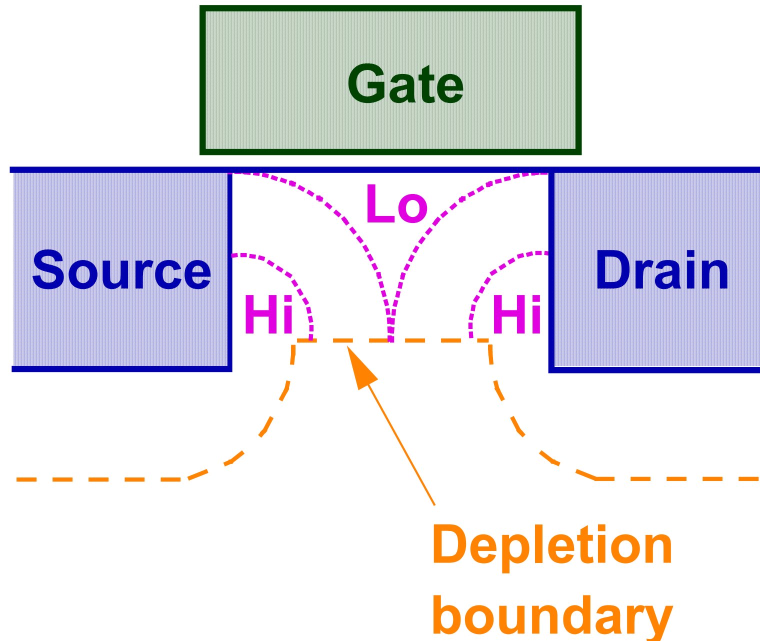
$$V_t = V_{fb} + 2\psi_B + \frac{\sqrt{4\varepsilon_{si} q N_a \psi_B}}{C_{ox}} = V_{fb} + (1 + \frac{6t_{ox}}{W_{dm}})2\psi_B$$

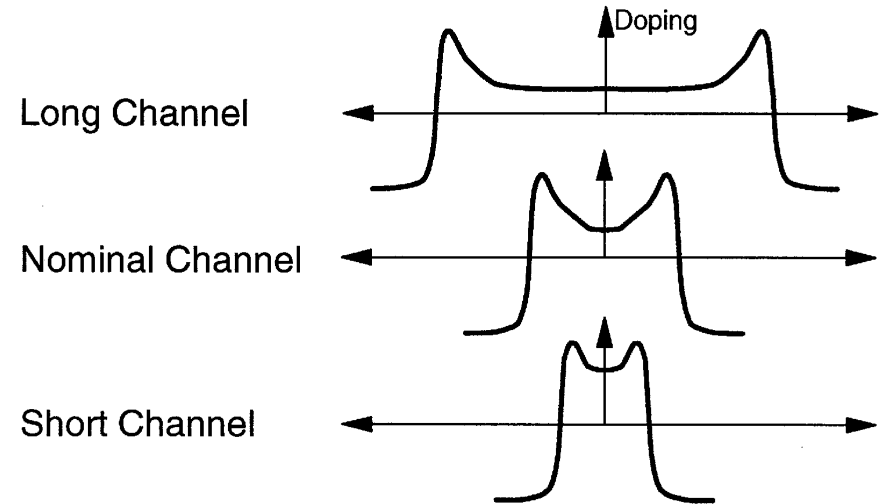
$V_t$ is increasing slightly toward shorter channel lengths and higher $N_a$. 2-D quantum effect further raises $V_t$ as the surface field increases and the electrons experience more confinement.

On the other hand, device design calls for lower $V_{dd}$ and $V_t$ as the CMOS dimensions are scaled down.

1-D vertically nonuniform doping only addresses $V_t$ of long channel devices. Laterally nonuniform doping helps control $V_t$ of short channel devices.

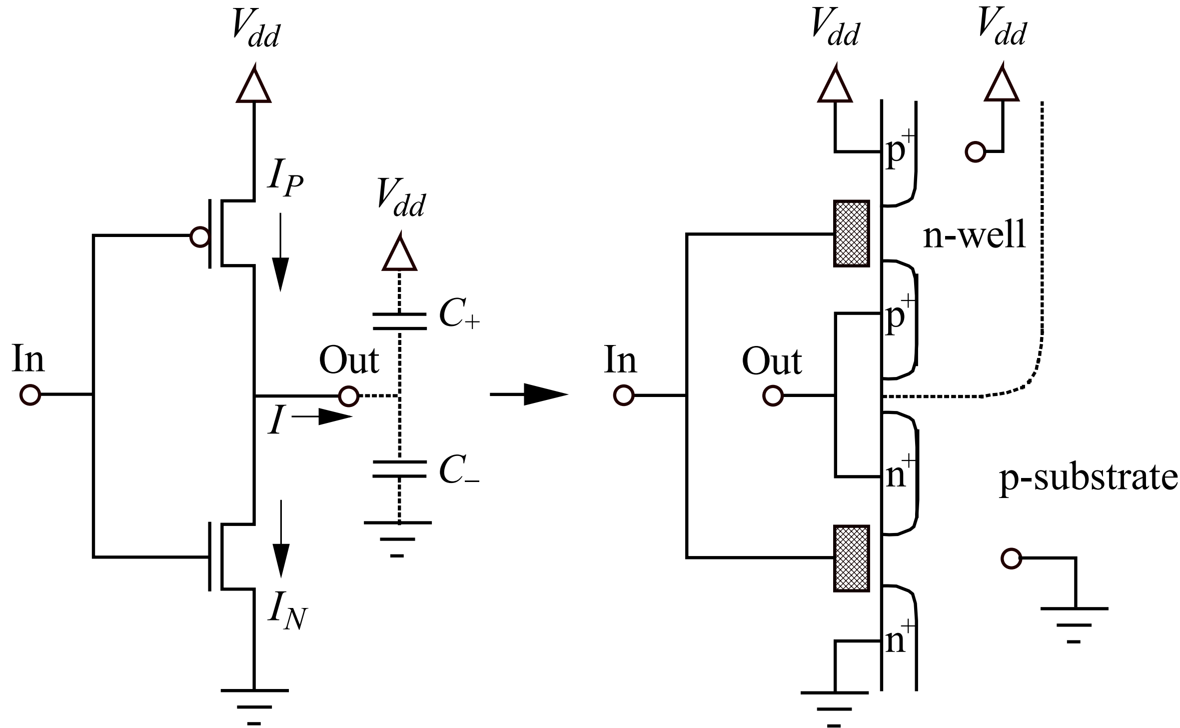
# Laterally Nonuniform Doping (Halo)



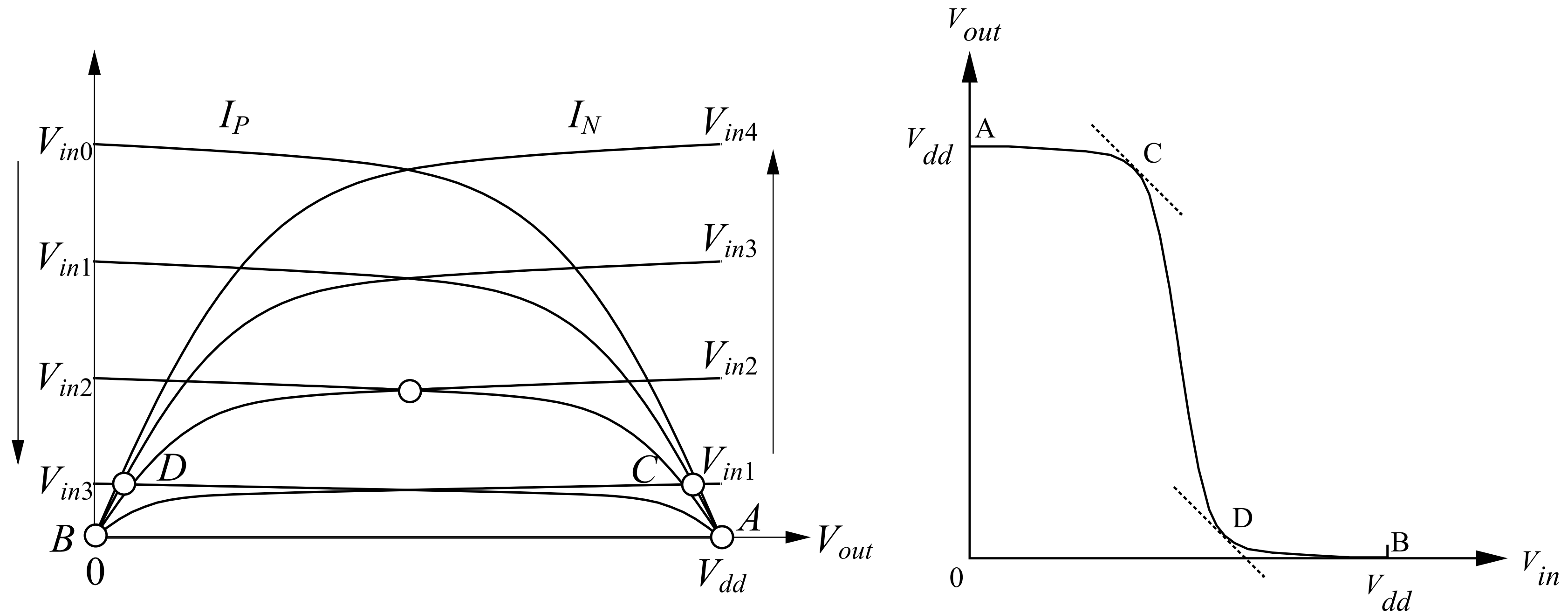

- Halo implants are made after gate patterning, therefore self-aligned to the gate like source-drain.
- Halo doped regions are farther apart for longer gates, and closer together for shorter gates.
- As a result, the “effective doping” becomes higher toward shorter devices, thus counteracting short channel effects.

# CMOS Inverter



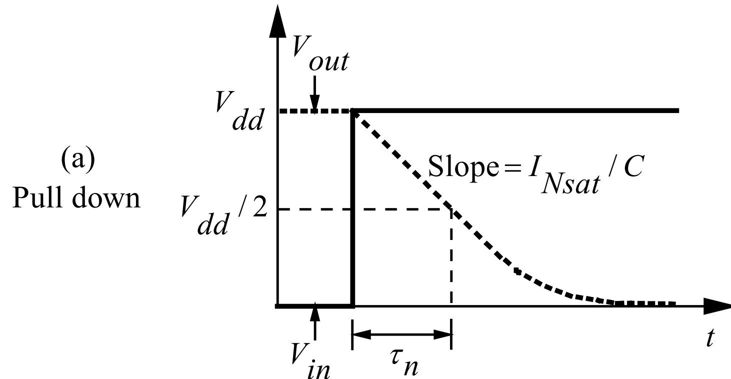

***Since only one of the transistors is on in the steady state, there is no static current or static power dissipation in a CMOS inverter.***

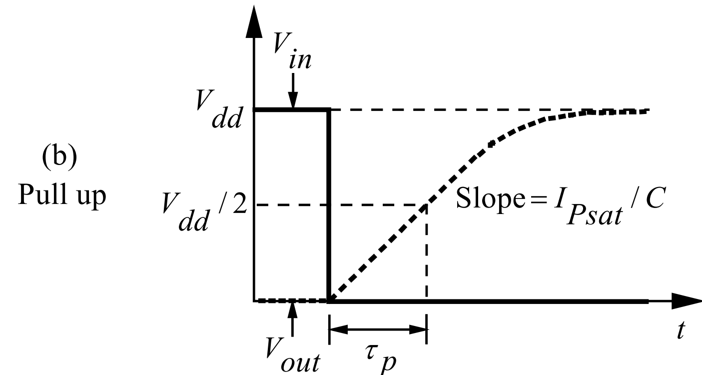
# CMOS Inverter Transfer Curve




***Qualitatively, the sharpness of the high-to-low transition of the $V_{out}$-$V_{in}$ curve is a measure of how well the circuit performs digital operations.***

# Switching Waveform for a Step Input




For nMOSFET pull down transition,

$$(C_- + C_+)\frac{dV_{out}}{dt} = C\frac{dV_{out}}{dt} = -I_N(V_{in} = V_{dd})$$

The pull down delay is

$$\tau_n = \frac{CV_{dd}}{2I_{Nsat}} = \frac{CV_{dd}}{2W_n I_{nsat}}$$

Similarly, the pMOSFET pull up delay is

$$\tau_p = \frac{CV_{dd}}{2I_{Psat}} = \frac{CV_{dd}}{2W_p I_{psat}}$$

For symmetric transfer curve and best noise margin, the width ratio should be $W_p/W_n = I_{nsat}/I_{psat} \approx 2$.

# Active Power Dissipation

Consider a capacitor $C$ between the output node and the ground. Initially, the output node is at the ground potential and there is no charge on $C$.

During a pull up transition, current flows from the power supply through the turned-on pMOSFET and raises the output node to $V_{dd}$. The capacitor is now charged to $Q = CV_{dd}$.

The energy flow out of the power supply is $QV_{dd} = CV_{dd}^2$, in which half or $CV_{dd}^2/2$ is energy stored in $C$; the other half is dissipated irreversibly as Joule heat.

Now the node is pulled down and the capacitor discharged by current through the turned-on nMOSFET to ground. The stored $CV_{dd}^2/2$ energy is now dissipated in the circuit.

$\Rightarrow$ A total energy of $CV_{dd}^2$ is dissipated irreversibly in an up-down switching cycle. If the clock frequency is $f$, and on the average a total capacitance $C$ undergoes an up-down cycle in a clock period $T$, then the CMOS power dissipation is

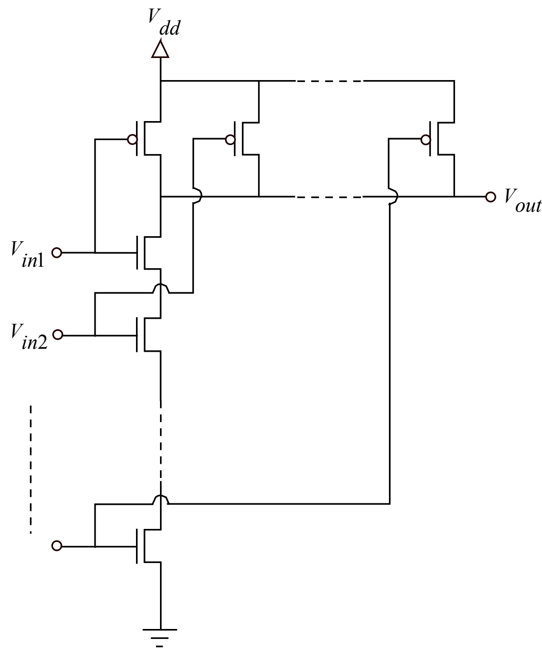
$$P = \frac{CV_{dd}^{\ 2}}{T} = CV_{dd}^{\ 2} f$$

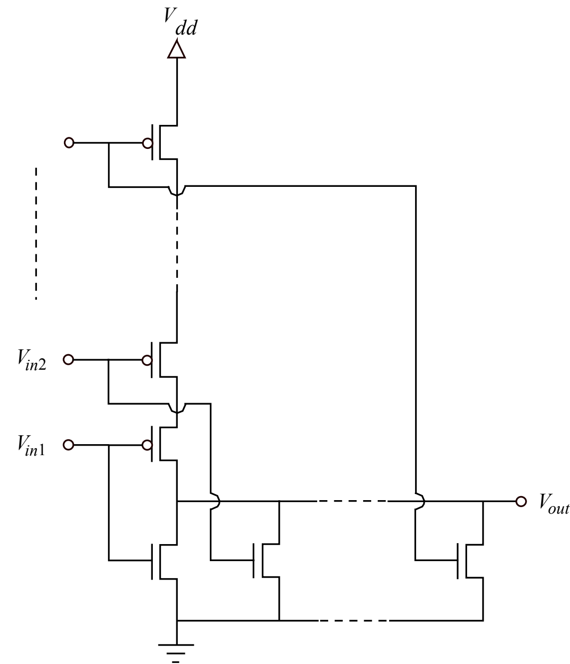
(Note that the cross-over current is neglected.)

# CMOS NAND and NOR Gates

NAND: Output is high unless all inputs are high.

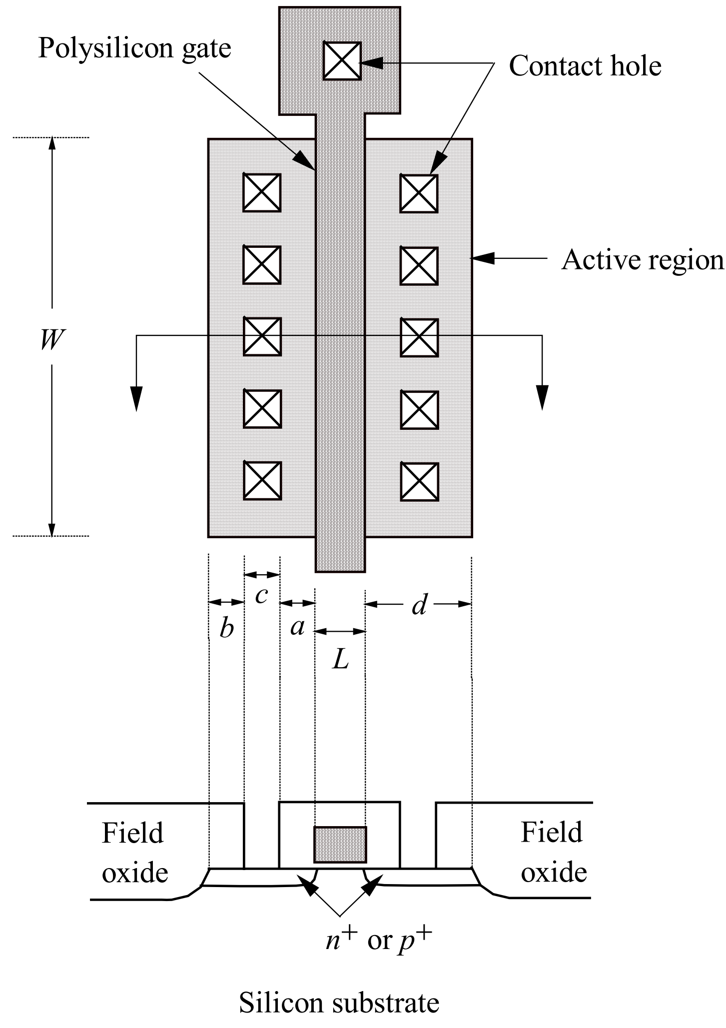
NOR: Output is low unless all inputs are low.




Like the inverter, there is no static power dissipation.

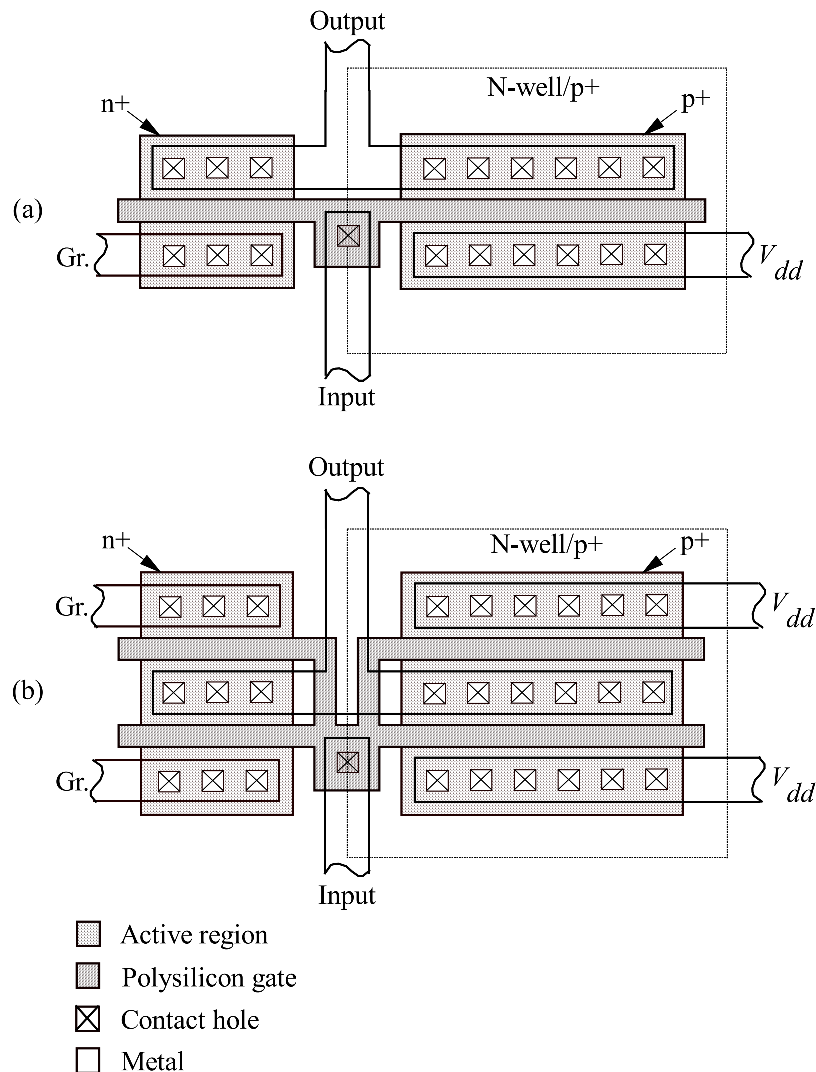
# MOSFET Layout




$d = a + b + c$ are layout groundrules dictated by lithography capability.

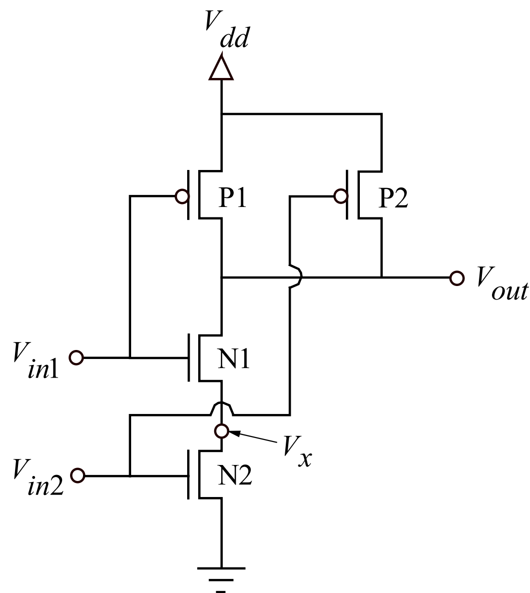
$L$ is limited by either lithography or device design.

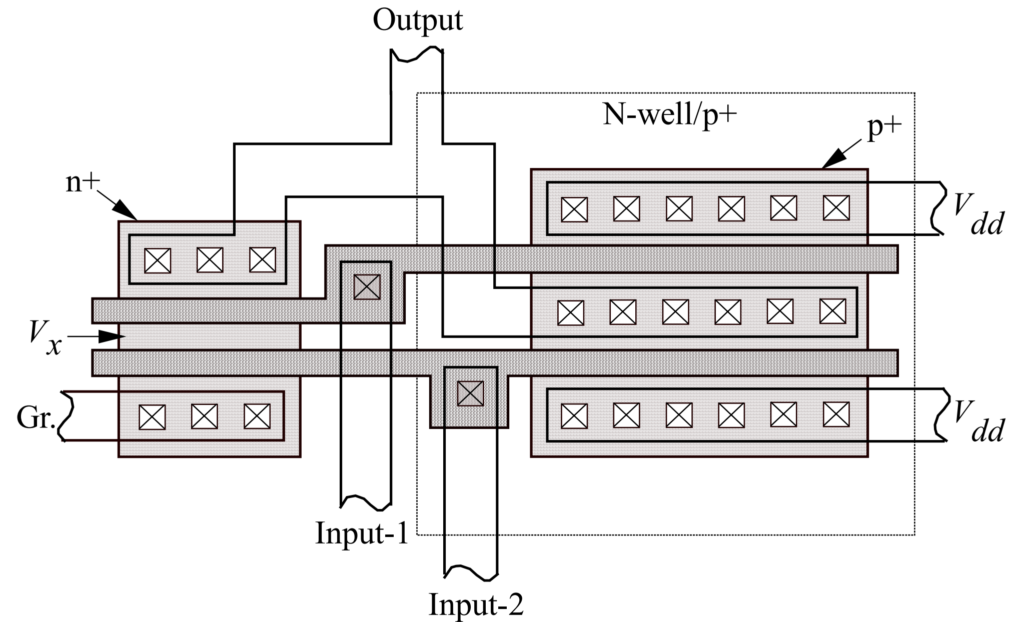
# CMOS Inverter Layout



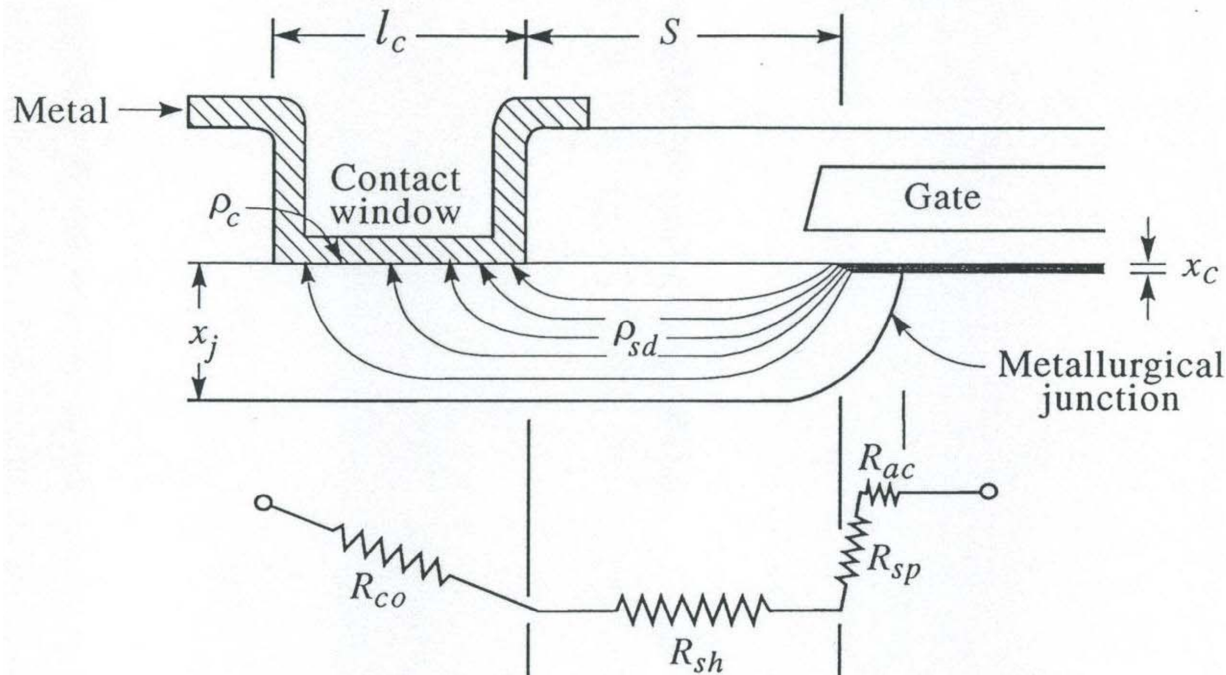

Folded (or interdigitated) layout in (b) reduces the junction capacitance contribution by a factor of 2.

# Two-way NAND Layout



# Source-Drain Series Resistance

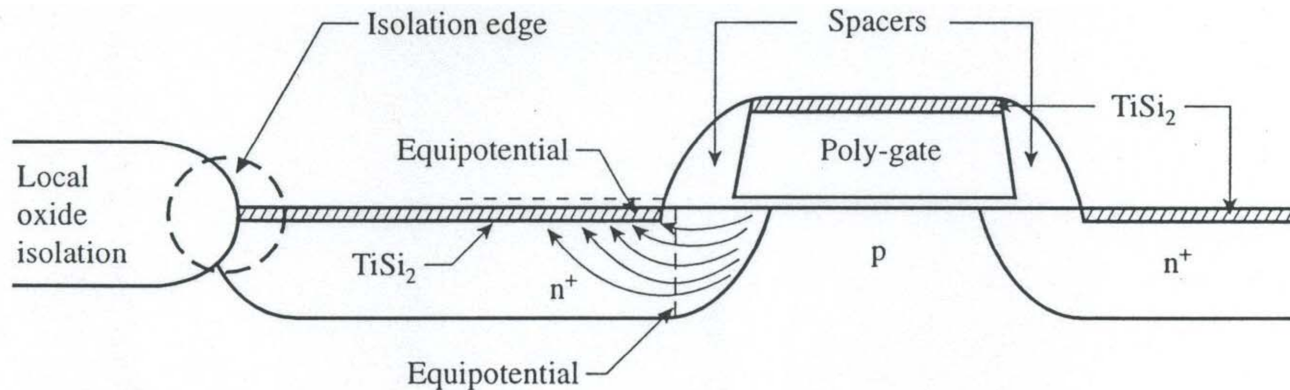



- $R_{ac}$ is the accumulation-layer resistance which is modulated by gate voltage and should be a part of the channel length.
- $R_{sp}$ is the spreading resistance associated with current injection from the surface channel into the bulk.
- $R_{sh} = \rho_{sh} S / W$
- $R_{co}$ is the contact resistance associated with current flowing from Si into metal.

# Self-Aligned Silicide Technology

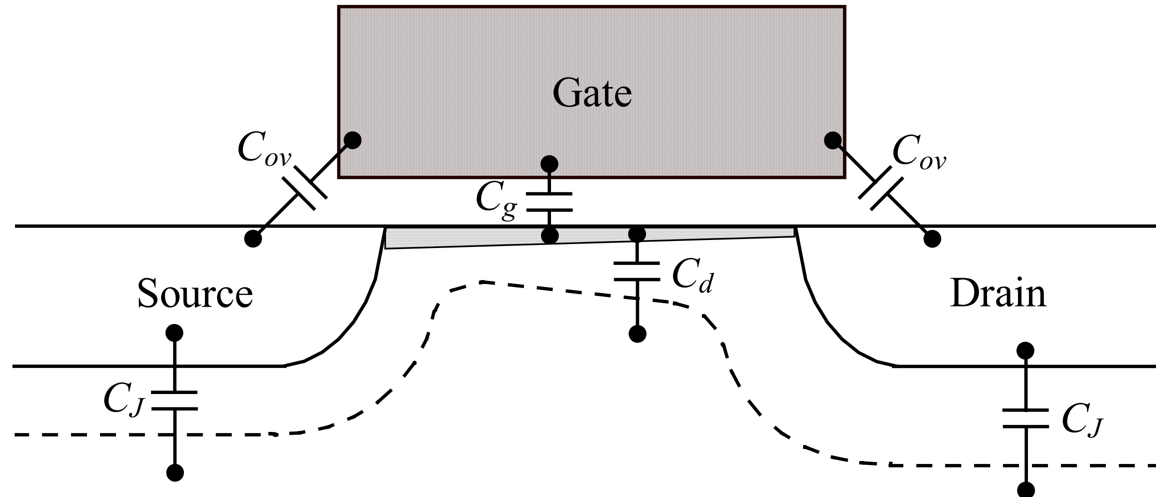
Sheet resistance:

- Metal (1 μm) — 0.05 Ω/sq
- N+, p+ diffusion (0.1 μm) — 50-500 Ω/sq
- Silicide (0.03 μm) – 5-10 Ω/sq



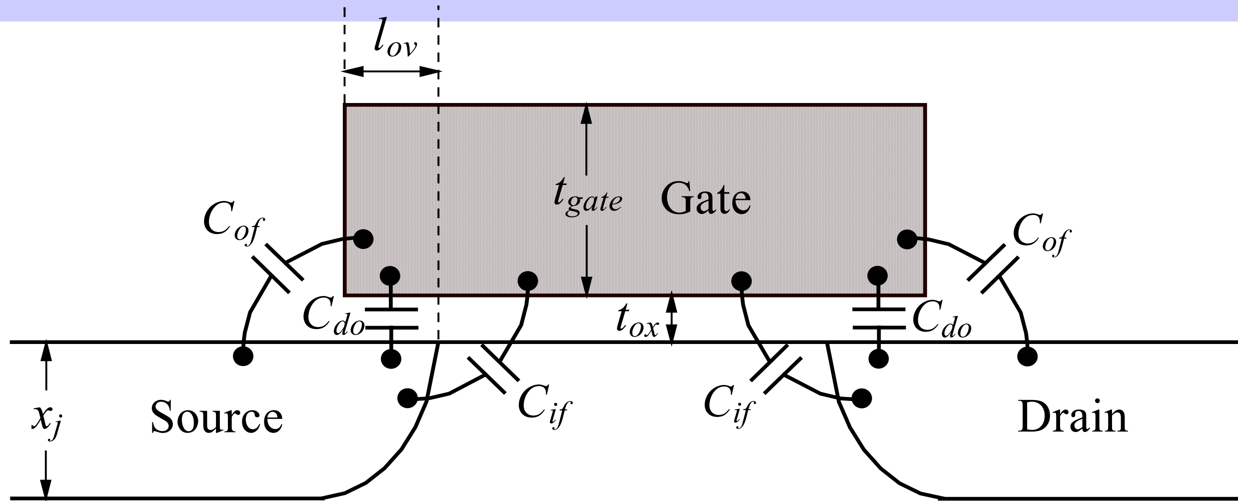

- $R_{sh}$ becomes negligible.
- $R_{co}$ between silicide and metal is negligible.
- Long contact regime between silicide and Si.

# MOSFET Capacitances




- Intrinsic capacitance: $C_g = WLC_{ox}$ $C_g = \frac{2}{3}WLC_{ox}$
- Parasitic capacitances:
  - Depletion capacitance $C_d = \varepsilon_s WL / W_{dm}$
  - Overlap capacitance
  - Junction capacitance $C_j = \varepsilon_{si} Wd / W_{dj} = Wd\sqrt{\frac{\varepsilon_{si} q N_a}{2(\psi_{bi} + V_j)}}$

# Overlap Capacitance




Direct overlap

$$C_{do} = Wl_{ov}C_{ox} = \frac{\varepsilon_{ox}Wl_{ov}}{t_{ox}}$$

Outer fringe

$$C_{of} = \frac{2\varepsilon_{ox}W}{\pi}\ln\left(1+\frac{t_{gate}}{t_{ox}}\right)$$

Inner fringe

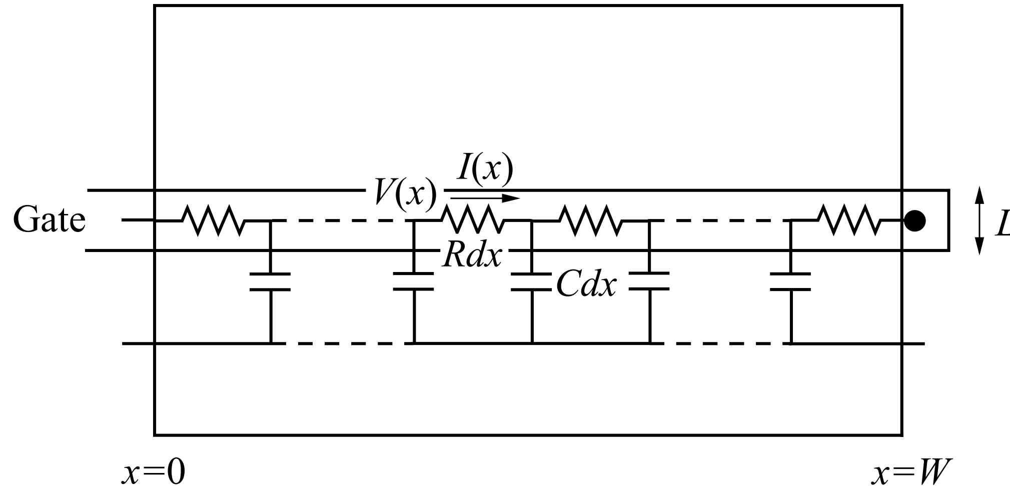
$$C_{if} = \frac{2\varepsilon_{si}W}{\pi}\ln\left(1+\frac{x_j}{2t_{ox}}\right)$$

For typical values of $t_{gate}/t_{ox} \approx 40$ and $x_j/t_{ox} \approx 20$,
$C_{of}/W \approx 2.3\varepsilon_{ox} \approx 0.08$ fF/μm, $C_{if}/W \approx 1.5\varepsilon_{si} \approx 0.16$ fF/μm (off state)

$$C_{ov}(V_g = 0) = C_{do} + C_{of} + C_{if} \approx \varepsilon_{ox}W\left(\frac{l_{ov}}{t_{ox}} + 7\right)$$

For reliability, $l_{ov} \approx (2\text{-}3)t_{ox}$, $C_{ov}/W \approx 10\varepsilon_{ox} \approx 0.3$ fF/μm

# Gate Resistance




The resistance per unit length is $R = \rho_g / L$
where $\rho_g$ is the silicide sheet resistivity (Ω/□).

The capacitance per unit length is approximately $C = C_{ox} L = \dfrac{\varepsilon_{ox} L}{t_{ox}}$

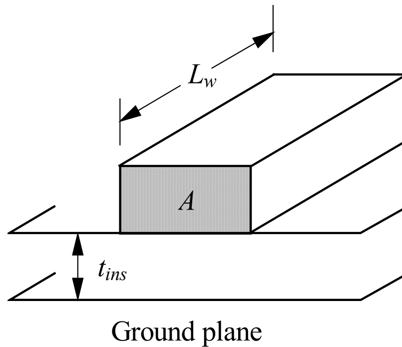
Diffusion eq.: $\dfrac{\partial^2 V}{\partial x^2} = RC \dfrac{\partial V}{\partial t}$

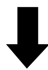
The effective *RC* delay is $RCW^2/4$ or $\tau_g = \dfrac{\rho_g C_{ox} W^2}{4}$

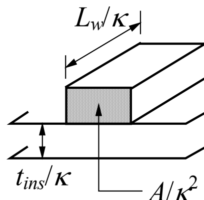
For $\rho_g$ = 10 Ω/□, $t_{ox}$ = 50 Å; $\tau_g$ < 1 ps if $W$ < 7.6 μm.

***Multiple-finger gate layouts with interdigitated source and drain regions should be used.***

# Interconnect Scaling






| | Interconnect parameters | Scaling factor ($\kappa > 1$) |
|---|---|---|
| Scaling assumptions | Interconnect dimensions ($t_w$, $L_w$, $W_w$, $t_{ins}$, $W_{sp}$) | $1/\kappa$ |
| | Resistivity of conductor ($\rho_w$) | 1 |
| | Insulator permittivity ($\varepsilon_{ins}$) | 1 |
| Derived wire scaling behavior | Wire capacitance per unit length ($C_w$) | 1 |
| | Wire resistance per unit length ($R_w$) | $\kappa^2$ |
| | Wire RC-delay ($\tau_w$) | 1 |
| | Wire current density ($I/W_w t_w$) | $\kappa$ |

# Interconnect Resistance

Interconnect RC delay is described by the same distributed RC-ckt & diffusion eq. as gate resistance.

$$\tau_w = \frac{1}{2} R_w C_w L_w^{\ 2}$$

where $R_w = \rho_w / W_w t_w$ and $C_w \approx 2\pi\varepsilon_{ins}$ (2 pF/cm).

$$\tau_w \approx \pi\varepsilon_{ins}\rho_w \frac{L_w^{\ 2}}{W_w t_w}$$

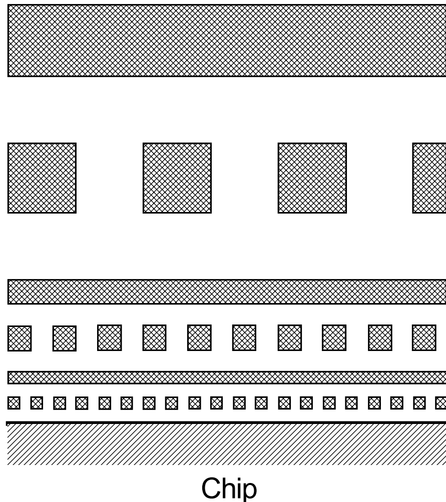
For aluminum and oxide technology,

$$\tau_w \approx (3\times10^{-18}\,\mathrm{s}) \frac{L_w^{\ 2}}{W_w t_w}$$

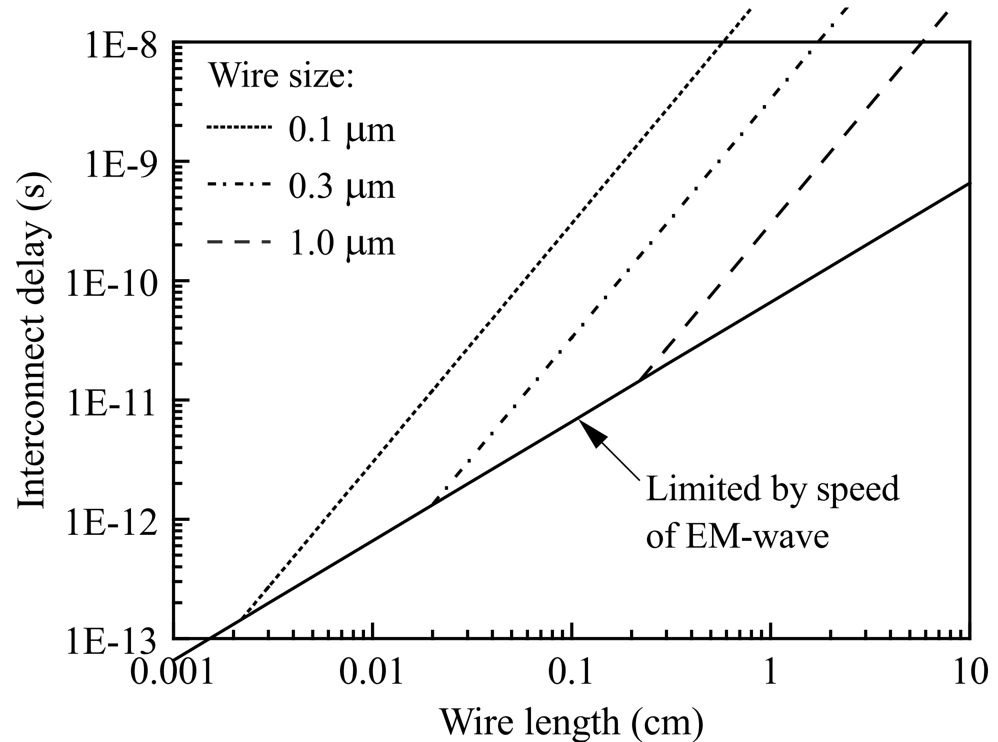
Local wire RC delay < 1 ps as long as $L_w^2/W_w t_w < 3\times10^5$ and can be scaled down without problem.

For global wires, however, $L_w$ does not scale down, e.g., $L_w^2/W_w t_w \sim 10^8$-$10^9$, and $\tau_w \sim$ 1 ns. Therefore, $W_w t_w$ cannot scale down $\Rightarrow$ Use large wires for global wiring.

# Global Interconnects




Must also scale up insulator spacing otherwise $C_w\uparrow$

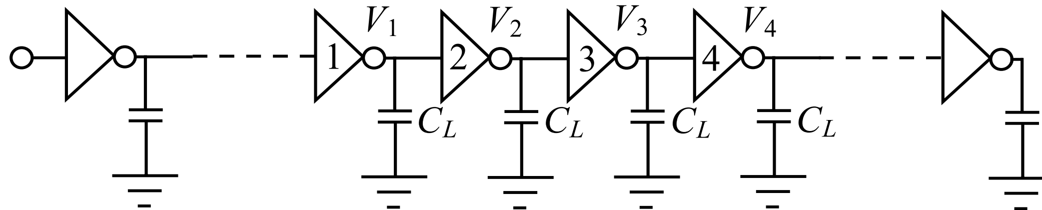


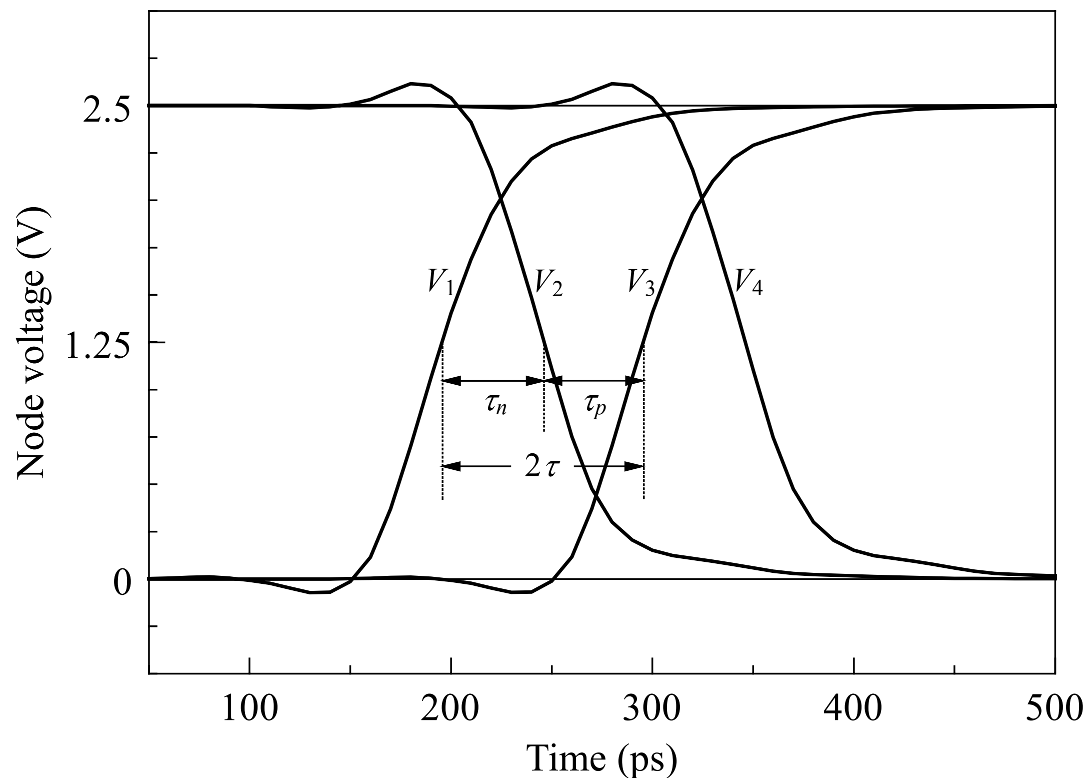

**RC Delay ~ $\ell^2/A$** **Time of Flight ~ $\ell/c$**

Ultimately, signal propagation is limited by the speed of electromagnetic wave, $c/(\varepsilon_{ins}/\varepsilon_0)^{1/2}$, (70 ps/cm for oxide), instead of by RC delay.

# Propagation Delay

For a chain of CMOS inverters or NAND/NOR gates,

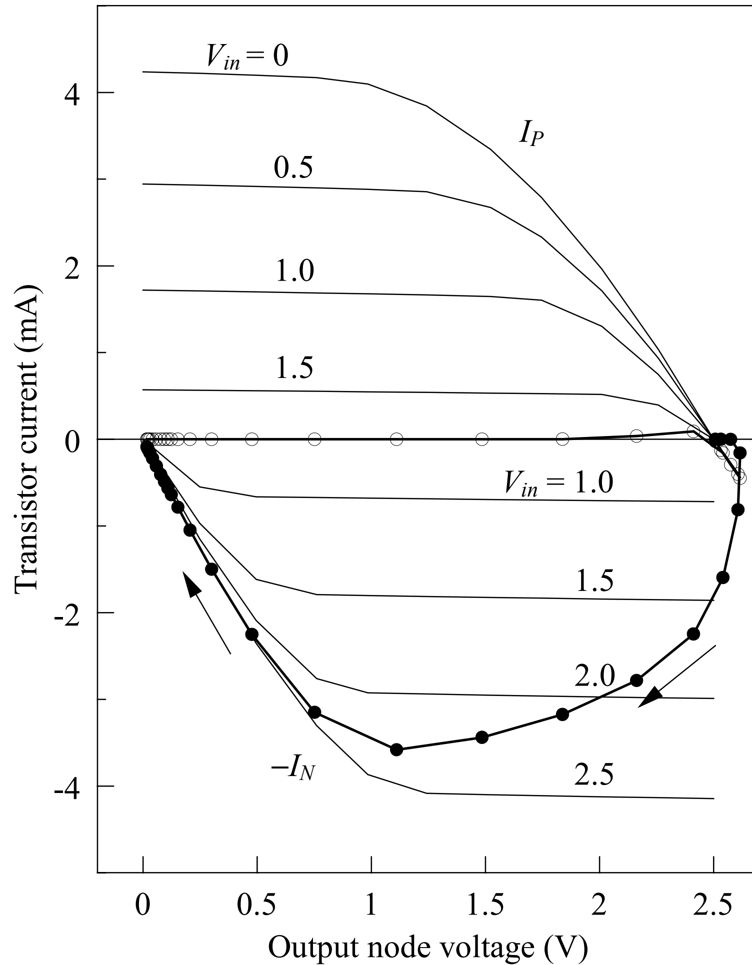


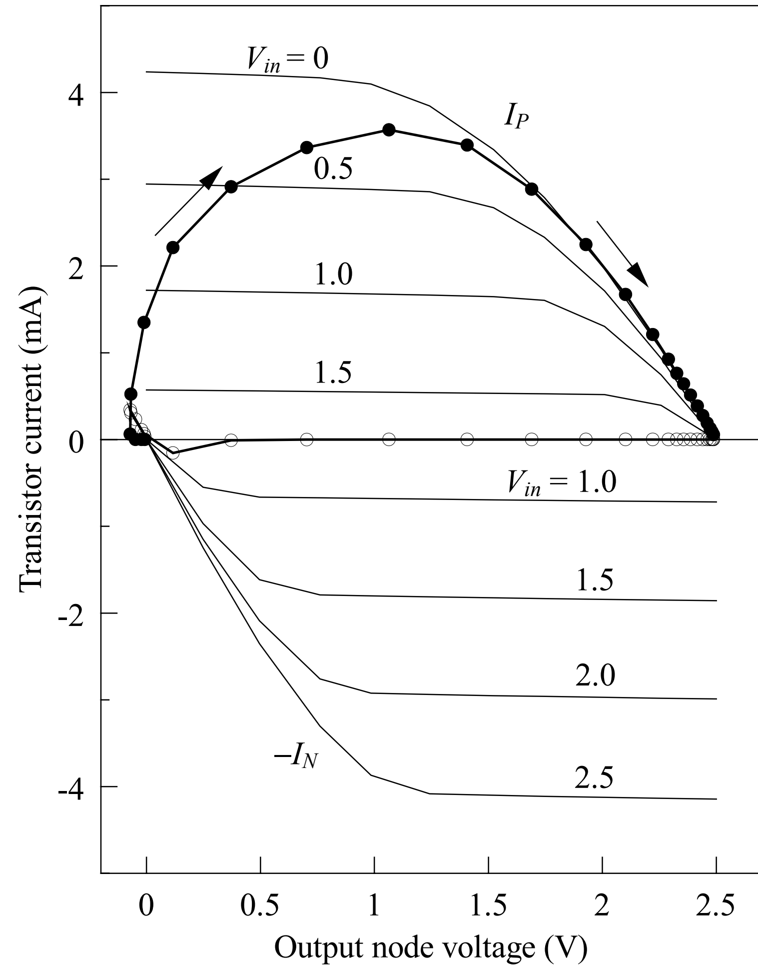

CMOS propagation delay equals $\tau = (\tau_n + \tau_p)/2$, where $\tau_n$ is the pull-down delay and $\tau_p$ is the pull-up delay.

# Bias Point Trajectories

## Pull down

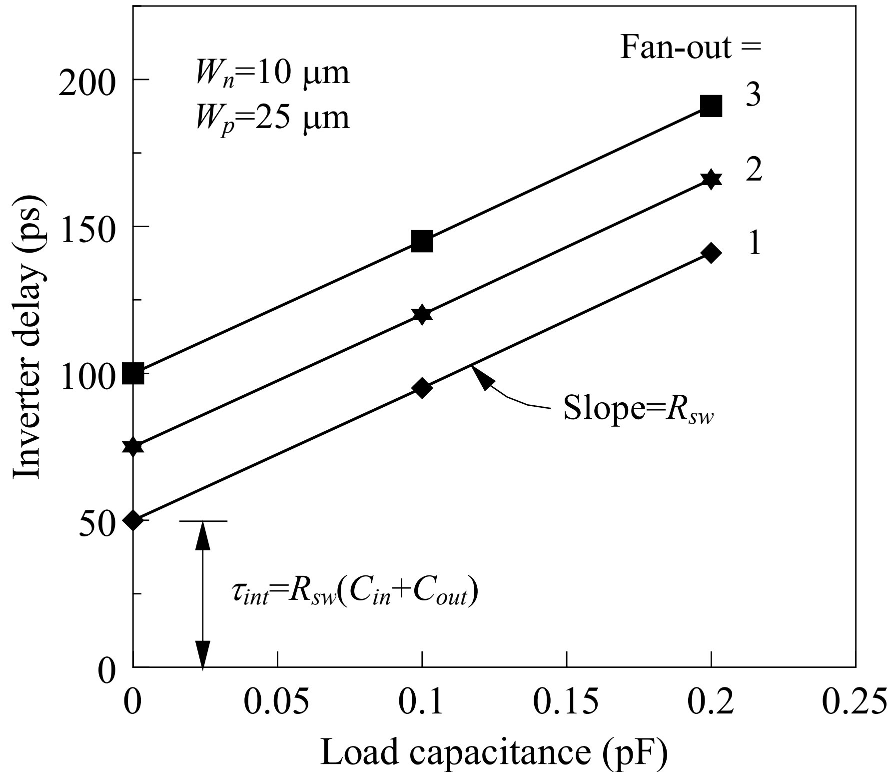



## Pull up




Equal time interval (10 ps) between dots.

# Delay Equation



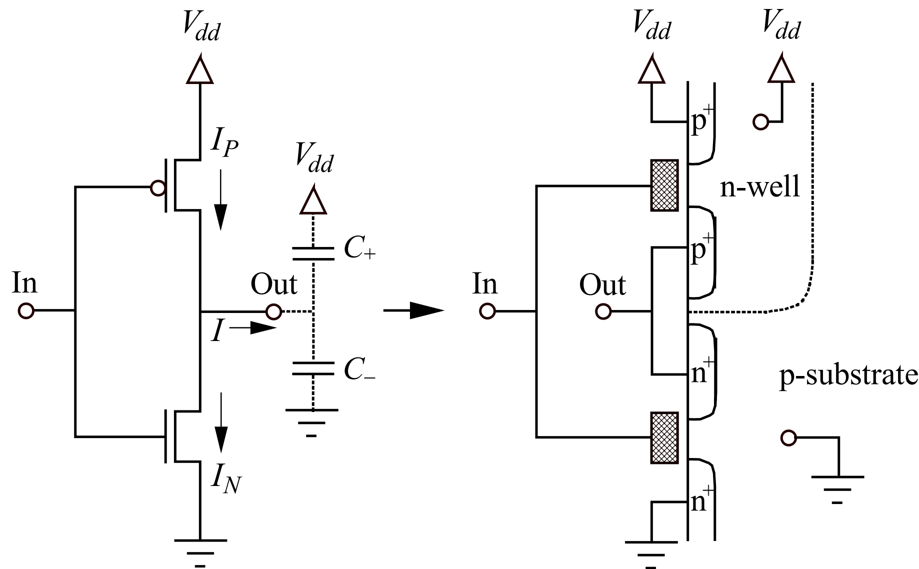

Delay equation:

$$\tau = R_{sw} \times (C_{out} + \mathrm{FO} \times C_{in} + C_L)$$

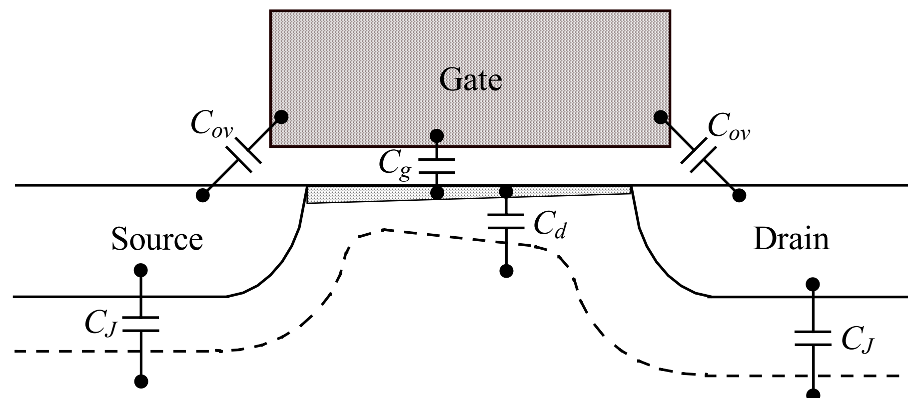
$R_{sw}$: Switching Resistance ($\equiv d\tau/dC_L$)
$C_{in}$: Input Capacitance (to next stage)
$C_{out}$: Output Capacitance
FO: Number of Fan-Out's

***The delay equation not only allows the delay to be calculated for any fan-out and loading conditions, but also decouples the two important factors that govern CMOS performance: current and capacitance.***

# Input and Output Capacitance




$C_{in}$: For switching n/p gates of the receiving stage.
$C_{out}$: For switching n/p drains of the sending stage.



# Switching Resistance

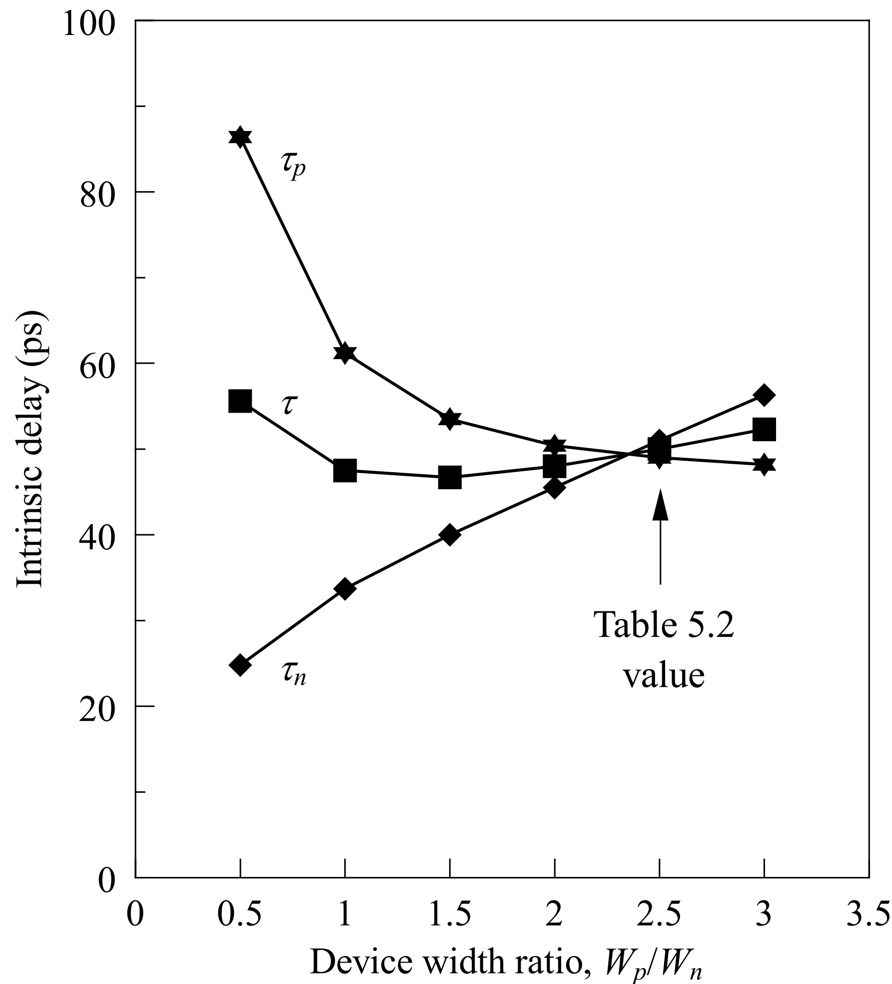
***Switching resistance*** $\boldsymbol{R_{sw}}$ is a direct indicator of the current drive capability of the logic gate.

If we define $R_{swn} \equiv d\tau_n/dC_L$ and $R_{swp} \equiv d\tau_p/dC_L$, then $R_{sw} = (R_{swn} + R_{swp})/2$ and we can write

$$R_{swn} = k_n \frac{V_{dd}}{W_n I_n} \qquad R_{swp} = k_p \frac{V_{dd}}{W_p I_p}$$

where $I_n$, $I_p$ are maximum on currents at $V_{ds} = V_g = V_{dd}$, and $k_n$, $k_p$ are numerical fitting parameters. For step inputs with zero rise time, $k_n = k_p = 0.5$. For propagation delays, $k_n$, $k_p \sim 0.75$.

# Delay Sensitivity to n/p Width Ratio




Note that for equal pull-up and pull-down delays ($\tau_n = \tau_p$) and therefore symmetric transfer curve and best noise margin, $W_p/W_n = 2.5$.

Minimum CMOS delay, $(\tau_n + \tau_p)/2$, however, occurs at $W_p/W_n = 1.5$.

# Buffer Stage for Heavy Loads

For a given $W_p/W_n$ ratio, if both widths are increased by a factor of $k$, then $R_{sw} \rightarrow R_{sw}/k$, $C_{in} \rightarrow kC_{in}$, $C_{out} \rightarrow kC_{out}$. No change in intrinsic delay,

$$\tau_{\mathrm{int}} = R_{sw} \times (C_{in} + C_{out})$$

But driving capability is improved.

Consider a CMOS inverter driving a large capacitive load:

$$\tau = R_{sw}(C_{out} + C_L)$$

If $C_L >> C_{in}$, $C_{out}$, the delay can be improved by inserting a buffer stage $k$ (> 1) times wider than the original inverter. The two-stage delay is

$$\tau_b = R_{sw}(C_{out} + kC_{in}) + \frac{R_{sw}}{k}(kC_{out} + C_L) = R_{sw}(2C_{out} + kC_{in} + \frac{C_L}{k})$$

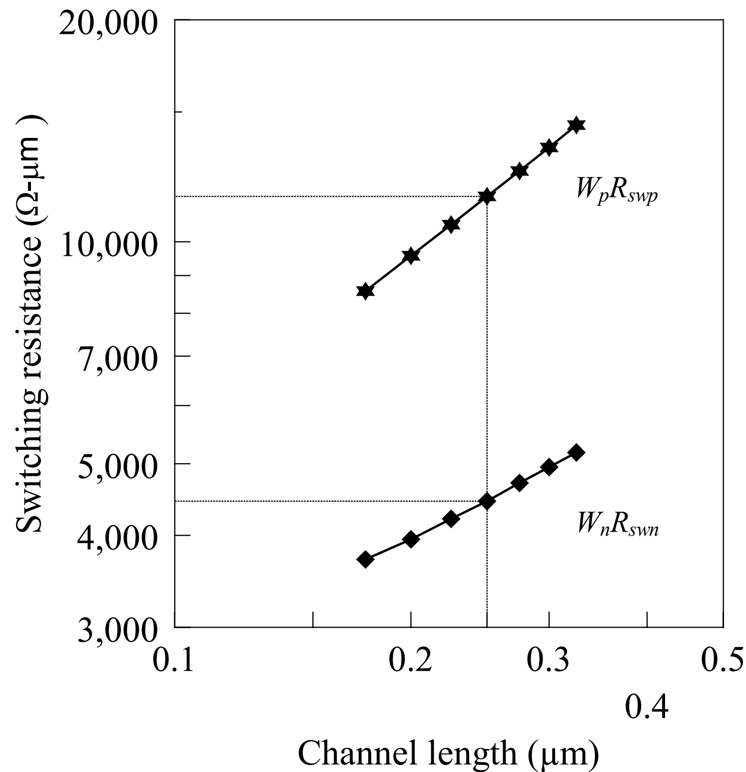
which has a minimum

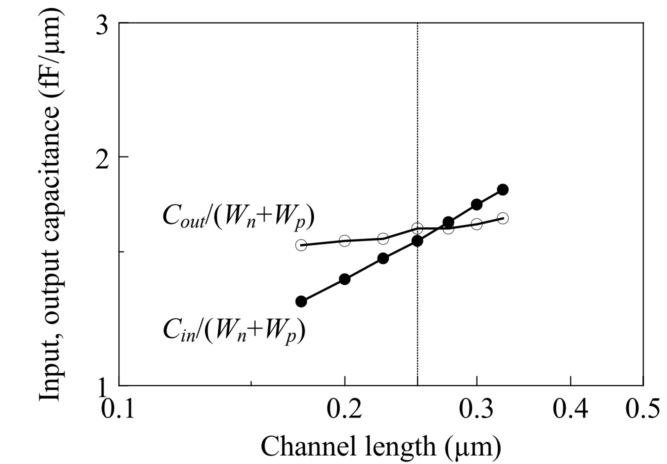
$$\tau_{b\,\mathrm{min}} = R_{sw}(2C_{out} + 2\sqrt{C_{in}C_L})$$

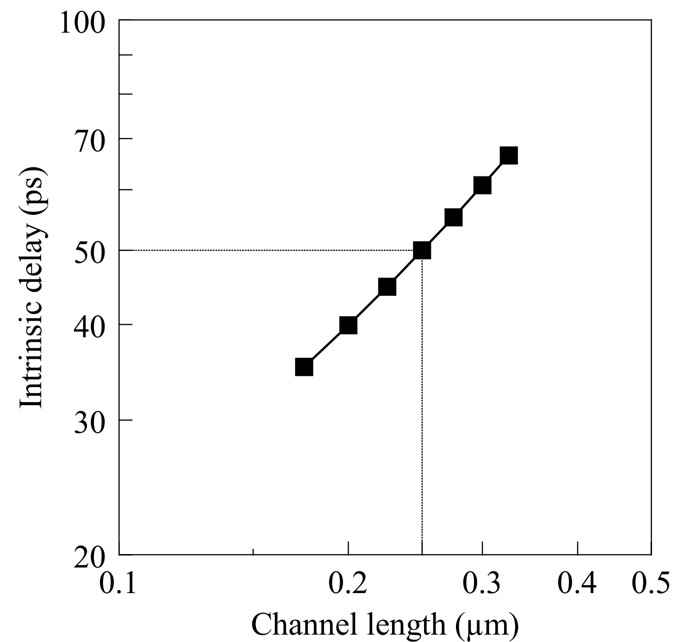
when $k = (C_L/C_{in})^{1/2}$.

Multiple-stage buffers can be used for very heavy loads (Problem 5.8-10).

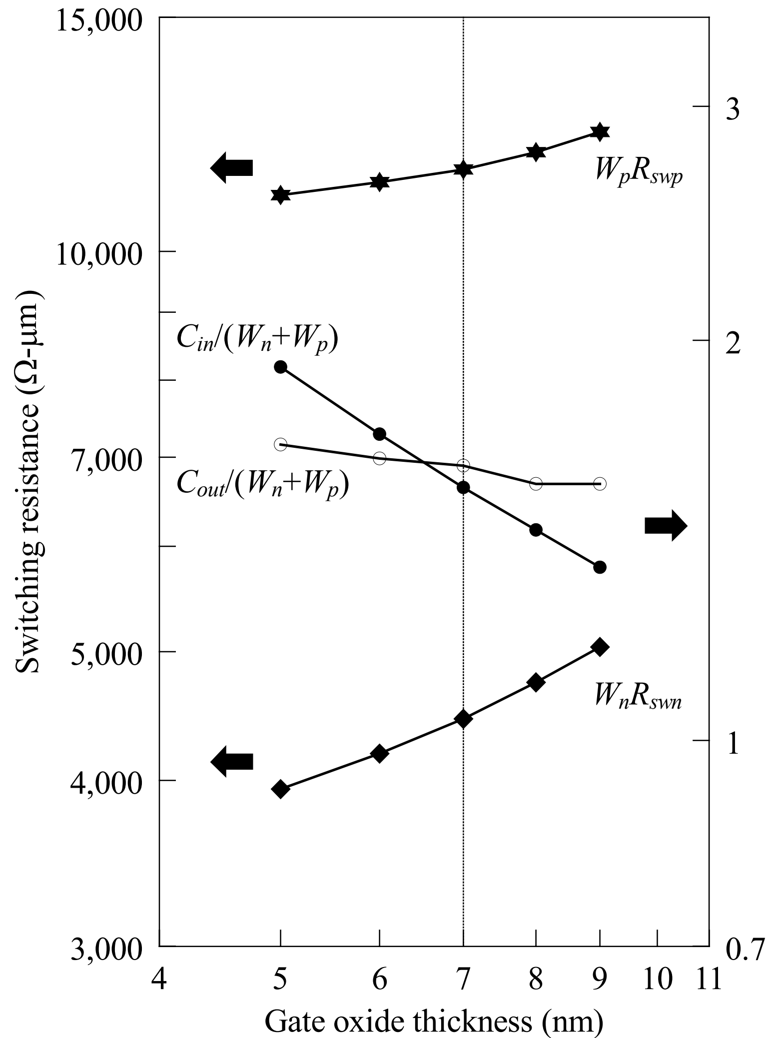
# Delay Sensitivity to Channel Length

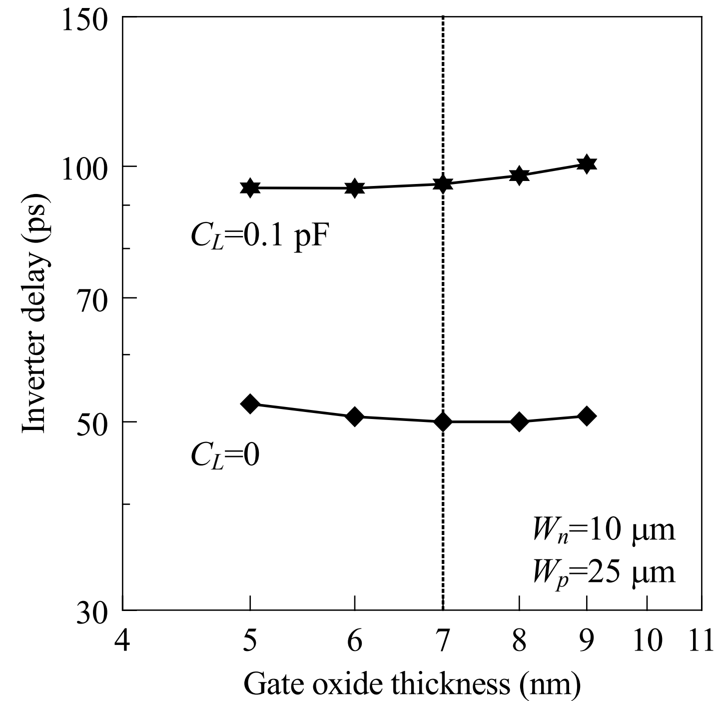


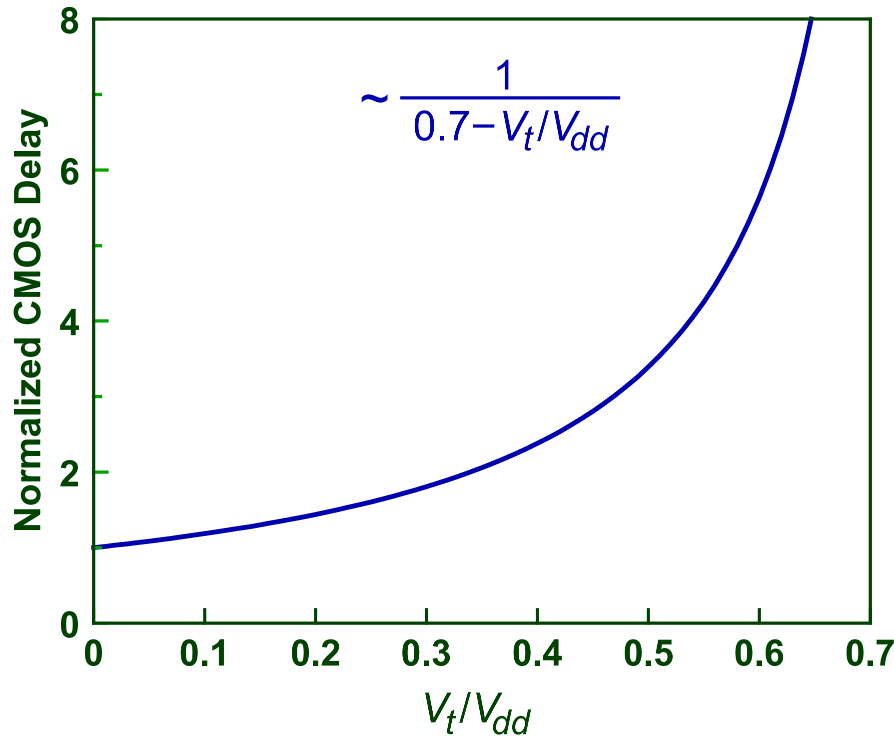



# Delay Sensitivity to Oxide Thickness



# Delay Sensitivity to $V_{dd}$ and $V_t$

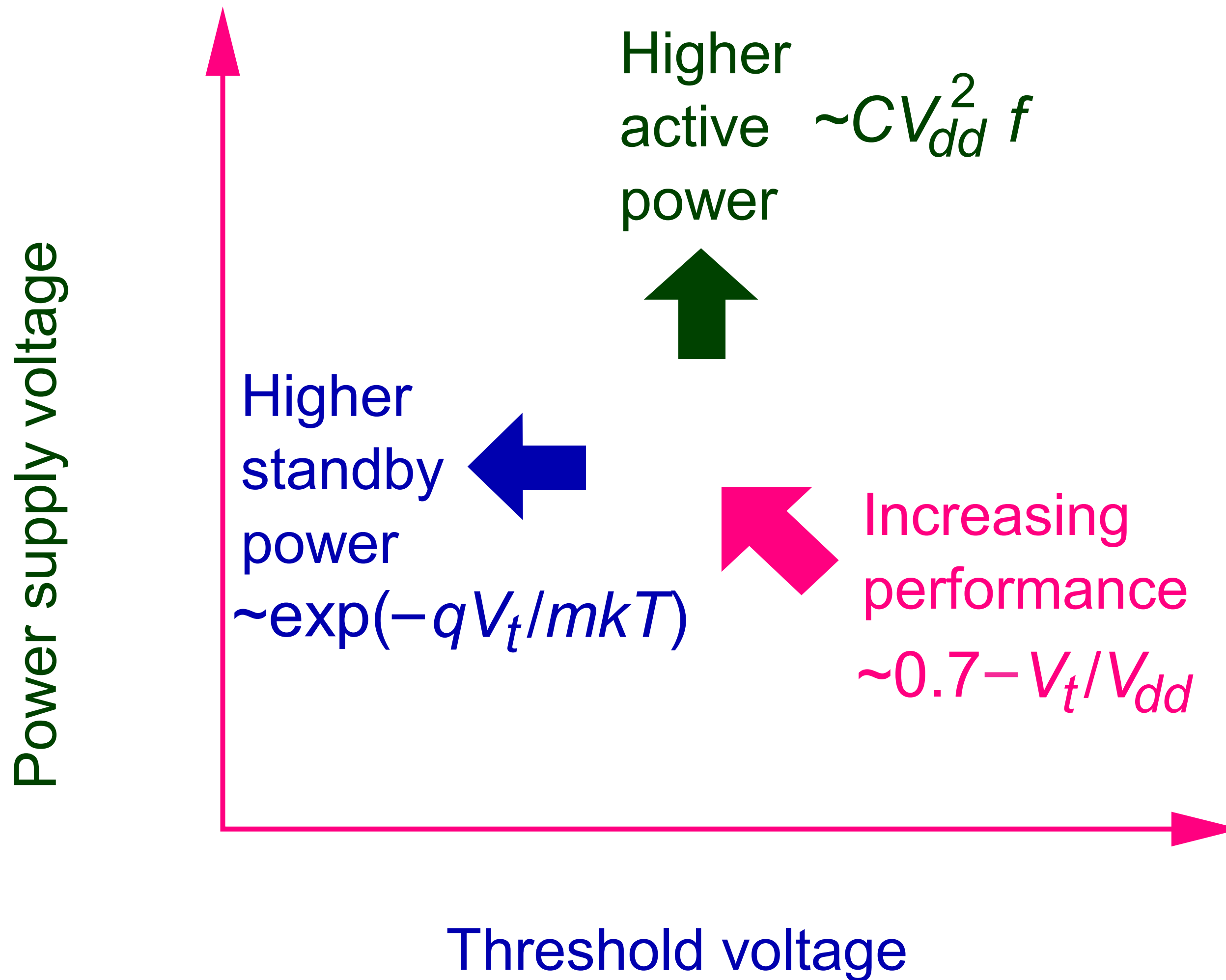



Delay sensitivity to $V_{dd}$ and $V_t$ is mainly in the $R_{sw}$ factor.

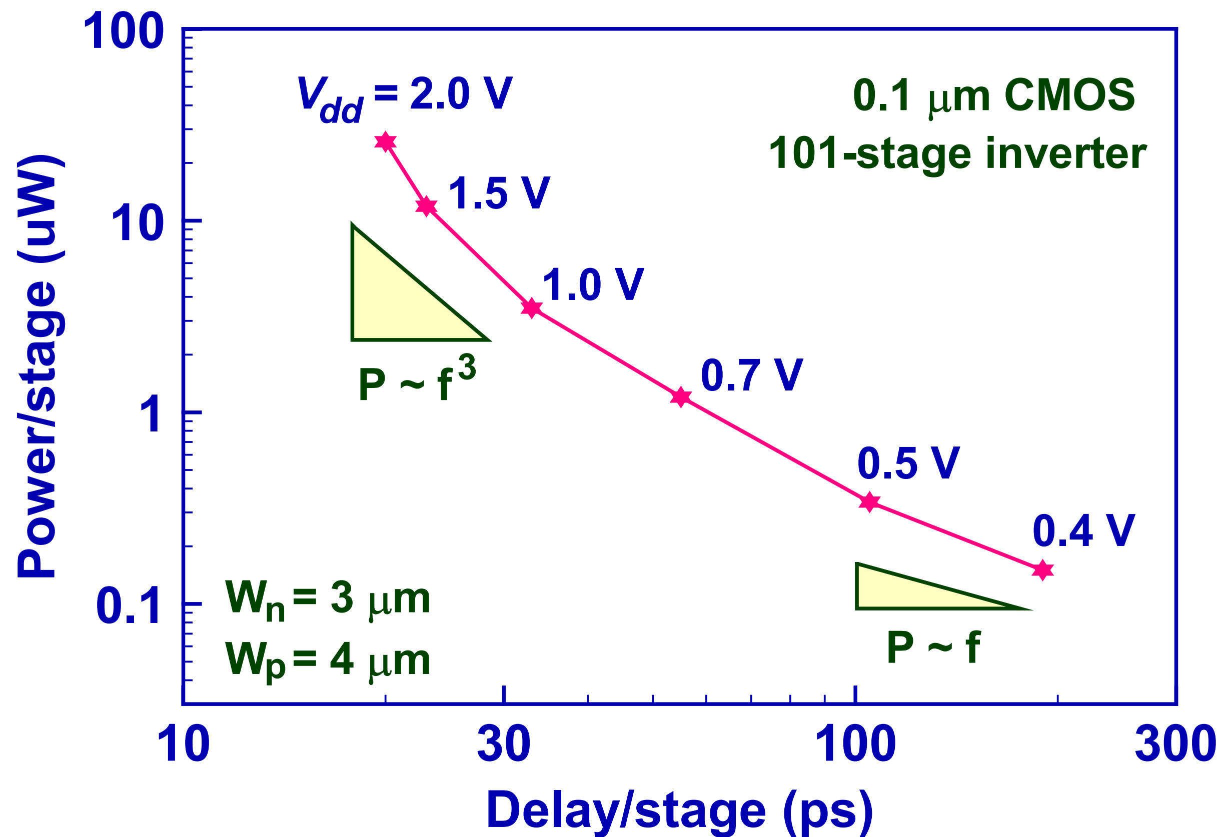
Note that the dependence is stronger than $I_{on} \propto 1 - V_t/V_{dd}$, due to the finite input rise time.

Desirable to keep $V_t/V_{dd} < 0.3$.

# CMOS Performance and Power



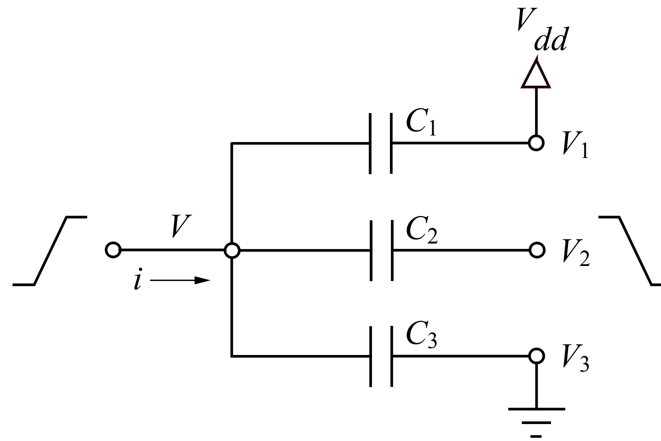
# CMOS Power vs. Delay Trade-off




Significant power savings by trading off performance and operating at low $V_{dd}$ (same $V_t$).

# Overlap Capacitance and Miller Effect

Consider driving 3 capacitors with respect to different voltages:



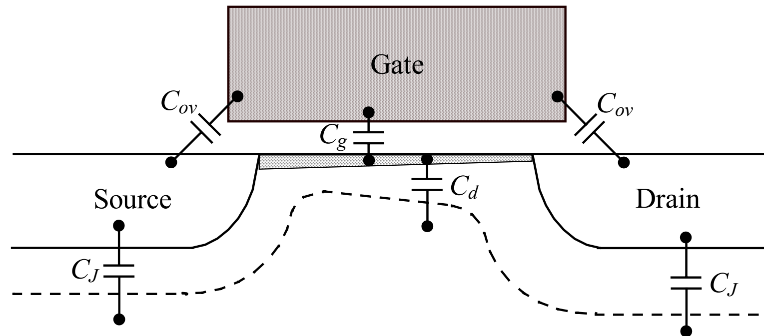

$$i = C_1 \frac{d(V - V_1)}{dt} + C_2 \frac{d(V - V_2)}{dt} + C_3 \frac{d(V - V_3)}{dt}$$

$$i = C_1 \frac{dV}{dt} + C_2 \frac{dV}{dt} - C_2 \frac{dV_2}{dt} + C_3 \frac{dV}{dt}$$

If $dV_2/dt = -dV/dt$, then $$i = (C_1 + 2C_2 + C_3)\frac{dV}{dt}$$

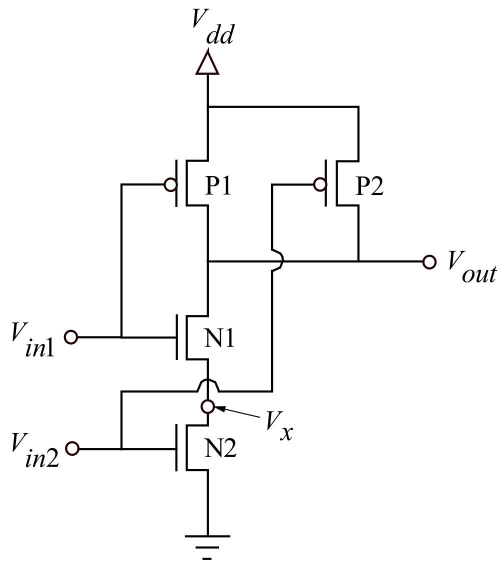
$\Rightarrow$ $C_2$ appears doubled to the driving source.

# Components of $C_{in}$ and $C_{out}$



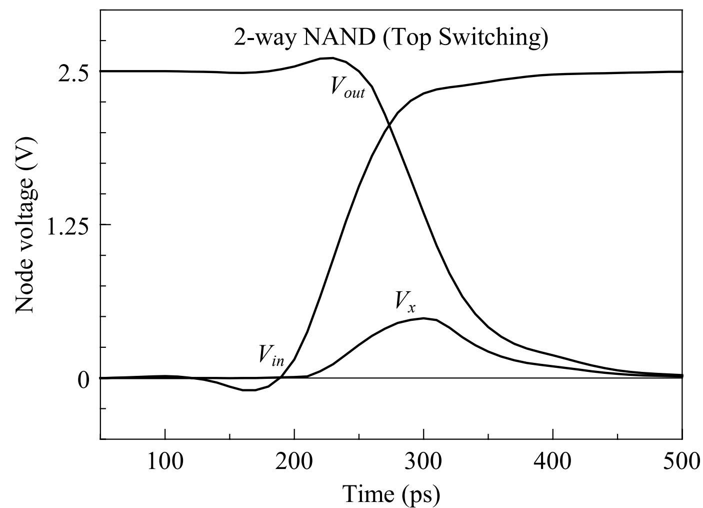

| | Input capacitance | Output capacitance |
|---|---|---|
| Intrinsic gate oxide capacitance (n & p) | 57% | 14% |
| Overlap capacitance | 43% | 35% |
| Junction capacitance (non-folded) | --- | 51% |

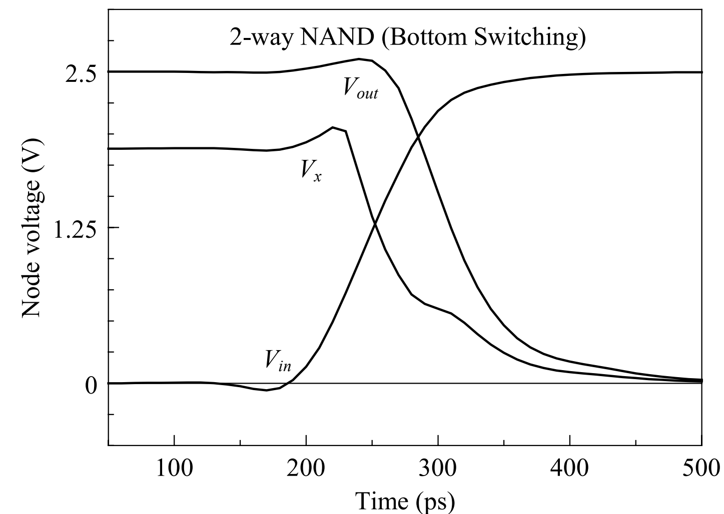
# Two Input (Two-way) NAND



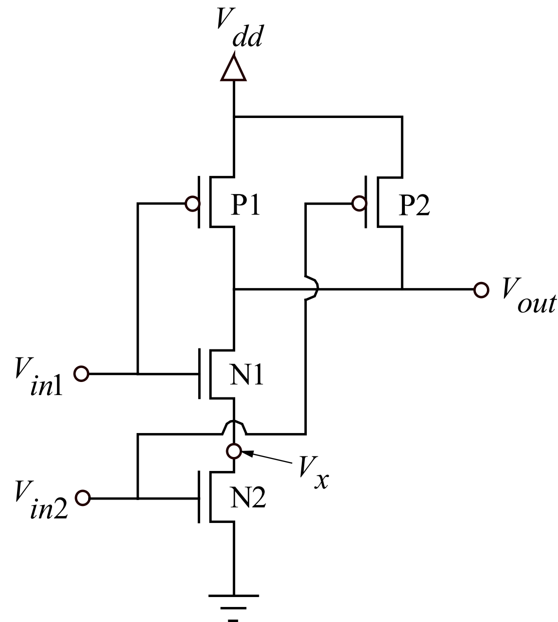

Top (N1) switching: Effective gate drive is $V_{in1} - V_x$, threshold is $V_t + (m - 1)V_x$ due to body effect.

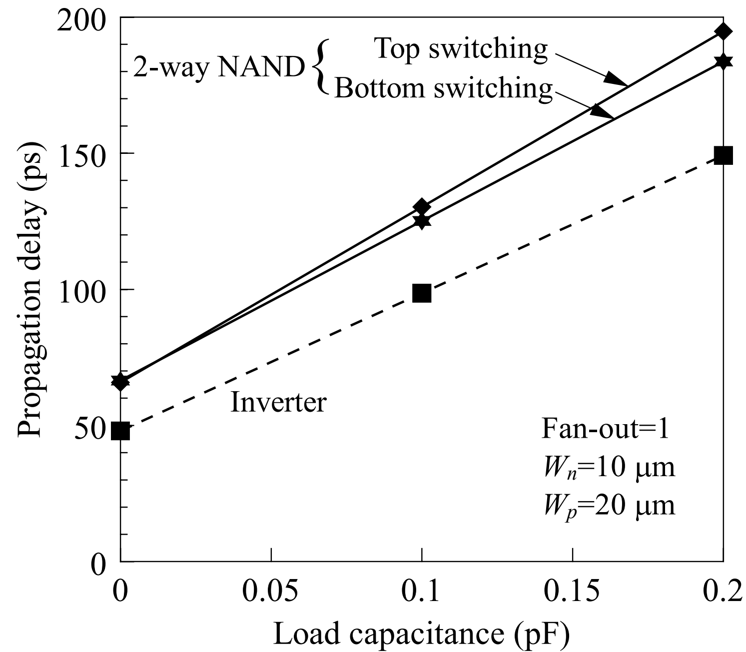
Bottom (N2) switching: Has more capacitance (of N1) to pull down.



# Two Input (Two-way) NAND




⇒ Delay is about 30% worse than inverter.
(Fan-in > 3 rarely used.)

$$R_{sw} \approx \frac{k_n V_{dd} + (\mathrm{FI}-1)V_{dsat}}{2W_n I_n} + \frac{k_p V_{dd}}{2W_p I_p}$$