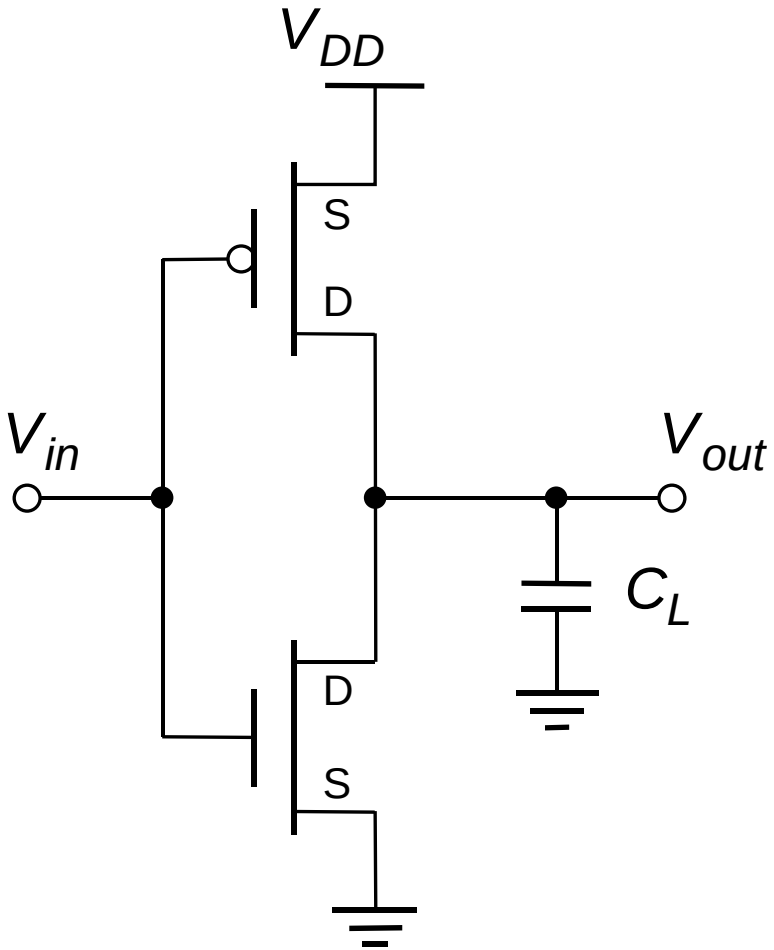
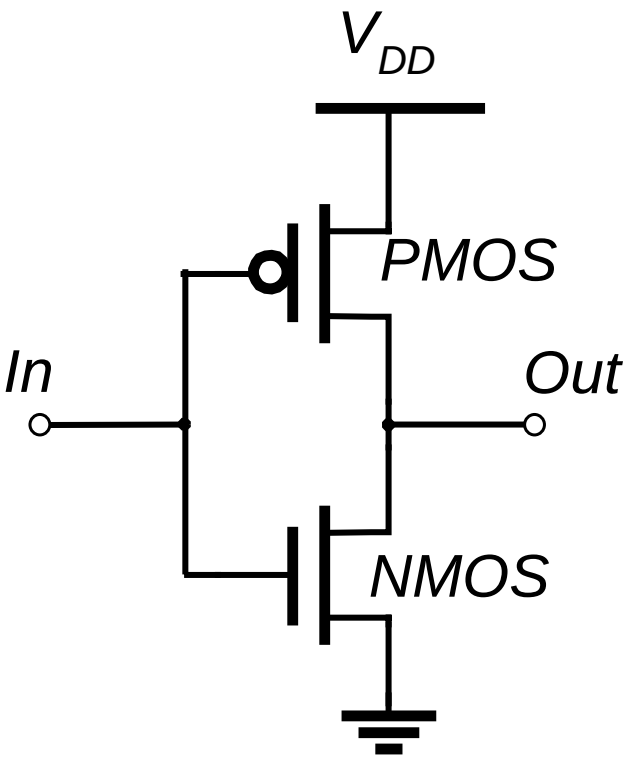
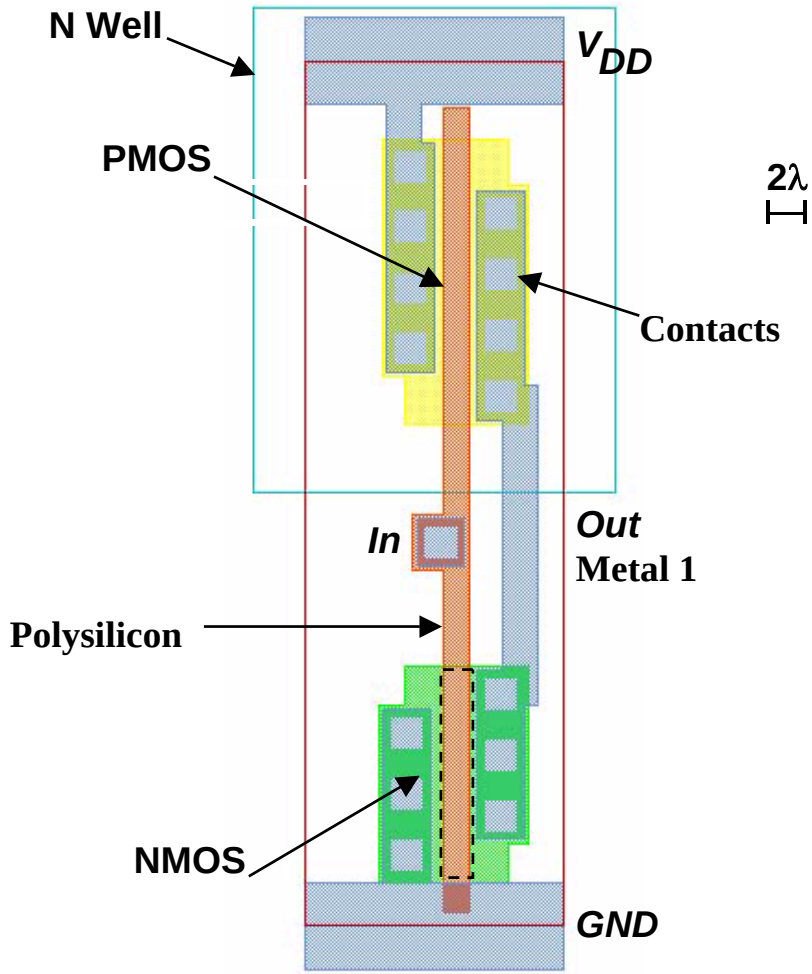
The CMOS Inverter: A First Glance



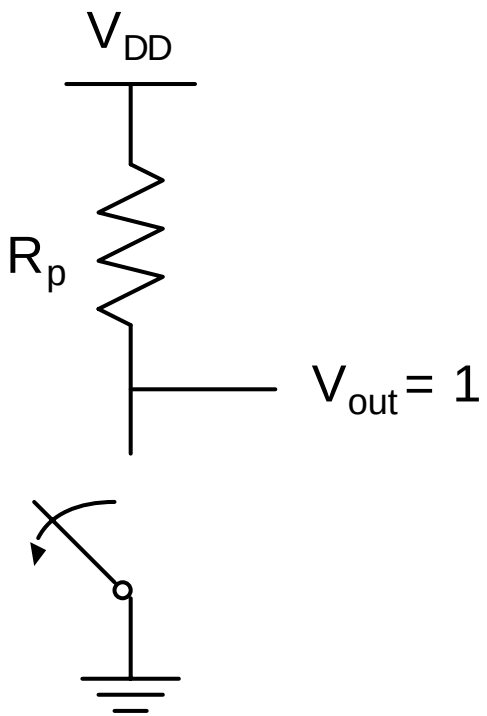
CMOS Inverter



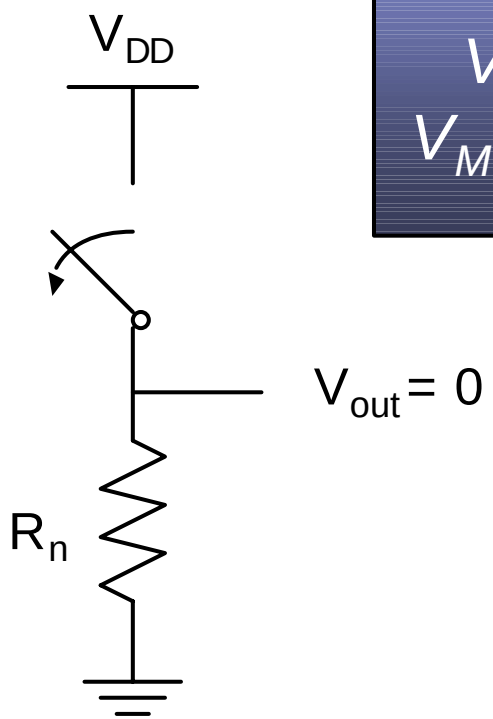


CMOS Inverter:
Steady State Response

$V_{OL} = 0$
$V_{OH} = V_{DD}$
$V_M = f(R_n, R_p)$

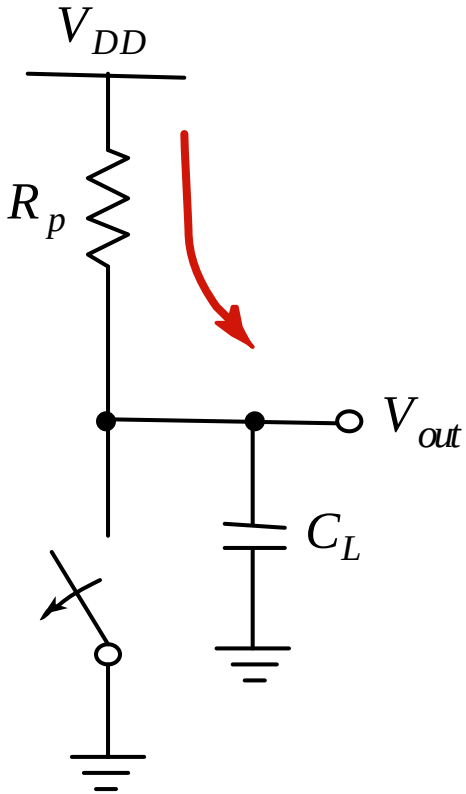


$V_{in} = 0$



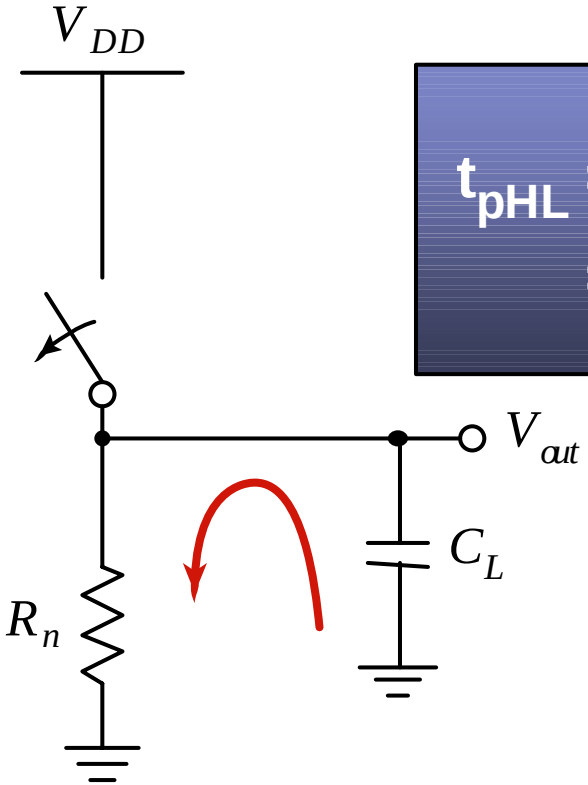
$V_{in} = V_{DD}$

CMOS Inverter: Transient Response



$V_{in} = 0$

(a) Low-to-high



$V_{in} = V_{DD}$

(b) High-to-low

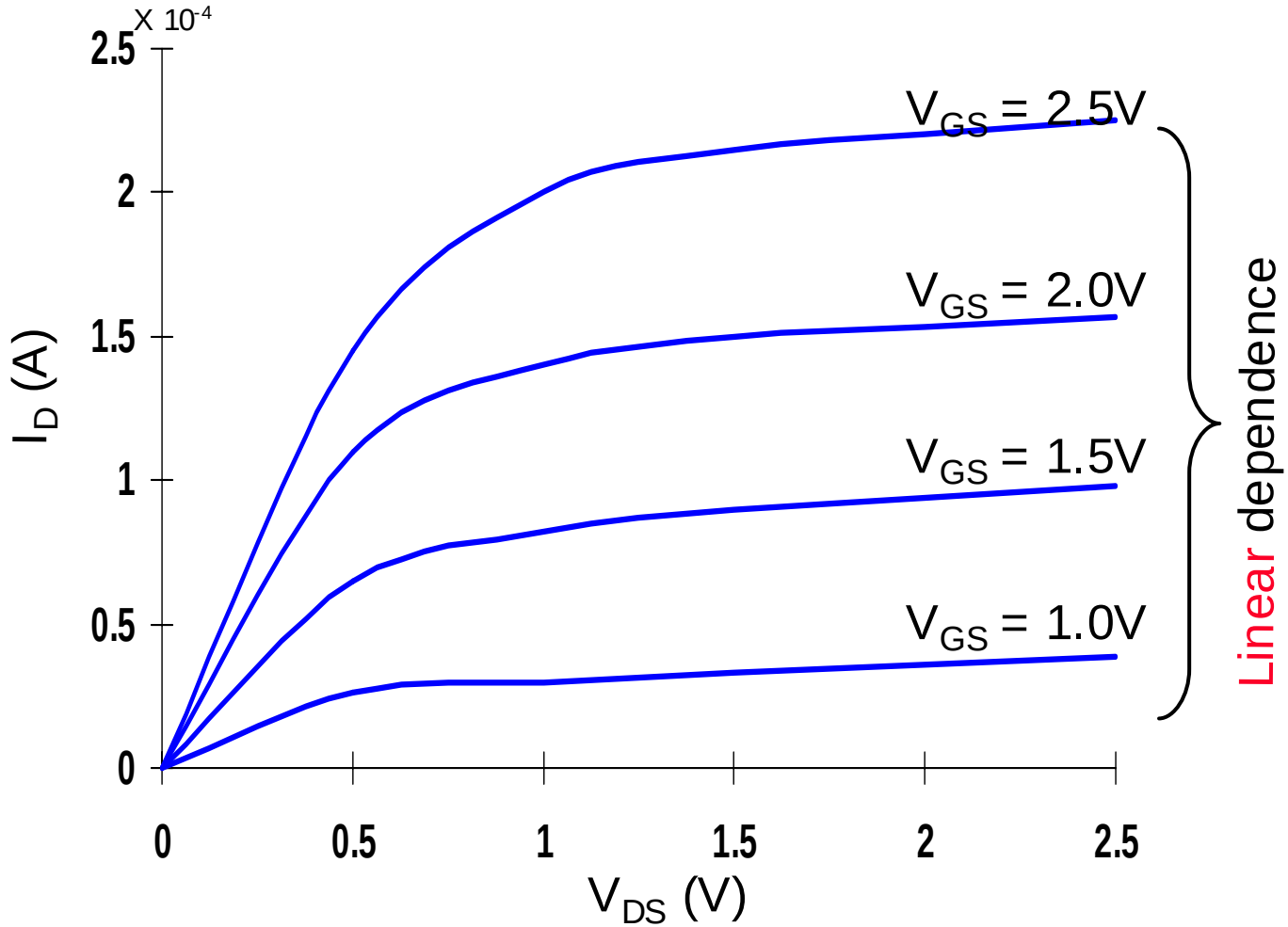
$t_{pHL} = f(R_{on}.C_L)$

$= 0.69\ R_{on}C_L$

CMOS Properties

- Full rail-to-rail swing $\Rightarrow$ high noise margins
  - Logic levels not dependent upon the relative device sizes $\Rightarrow$ transistors can be minimum size $\Rightarrow$ ratioless
- Always a path to V_{dd} or GND in steady state $\Rightarrow$ low output impedance (output resistance in kΩ range) $\Rightarrow$ large fan-out (less sensitive to noise as well)
- Extremely high input resistance (gate of MOS transistor is near perfect insulator) $\Rightarrow$ nearly zero steady-state input current
- No direct path steady-state between power and ground $\Rightarrow$ no static power dissipation
- Propagation delay function of load capacitance and resistance of transistors (limits fan-out)

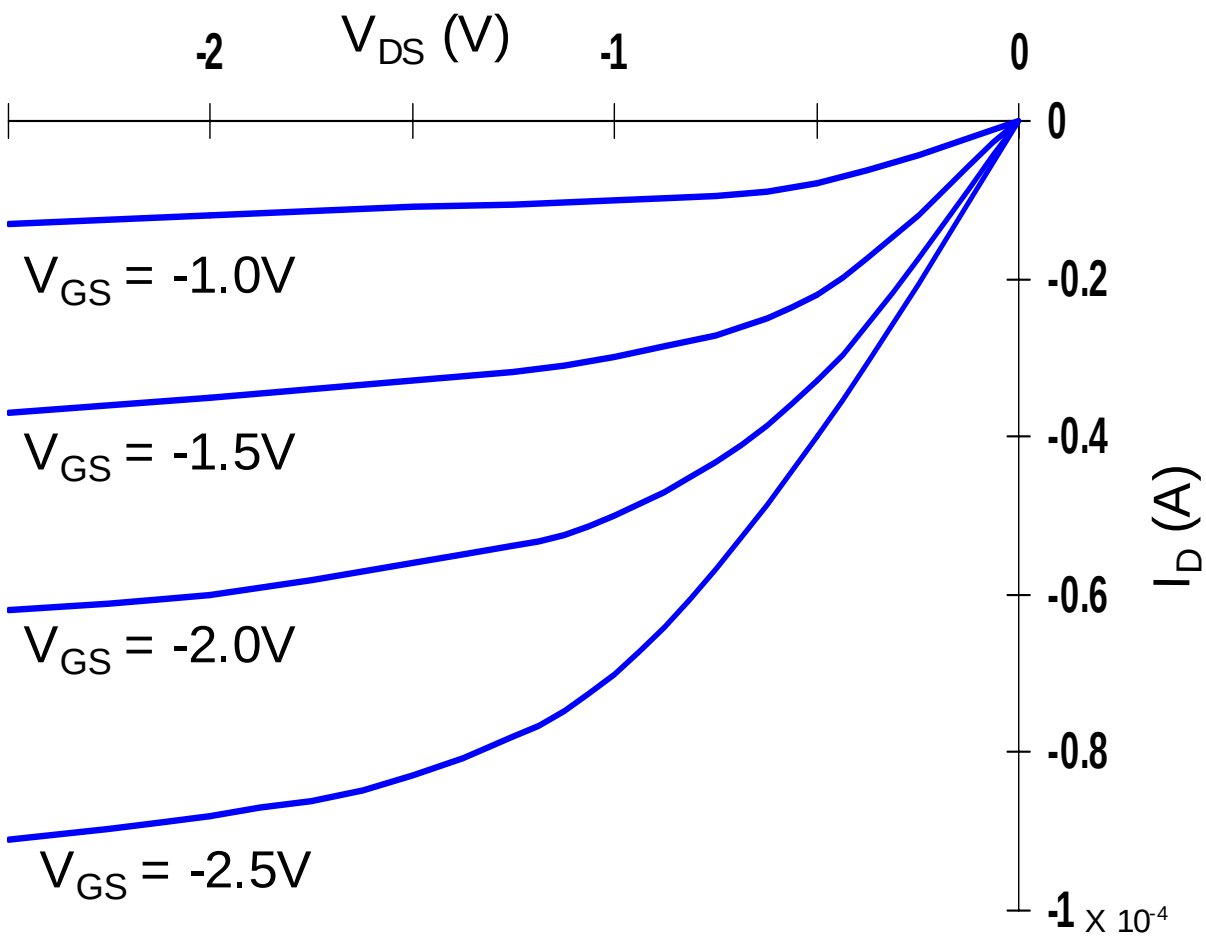
Review: Short Channel I-V Plot (NMOS)



NMOS transistor, 0.25um, L_d = 0.25um, W/L = 1.5, V_{DD} = 2.5V, V_T = 0.4V

Review: Short Channel I-V Plot (PMOS)

- All polarities of all voltages and currents are reversed

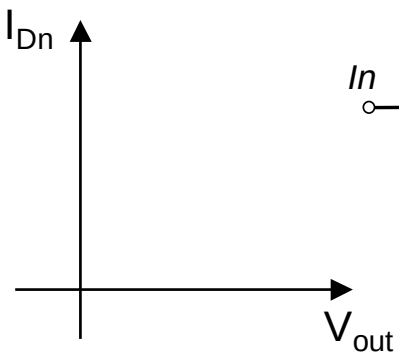


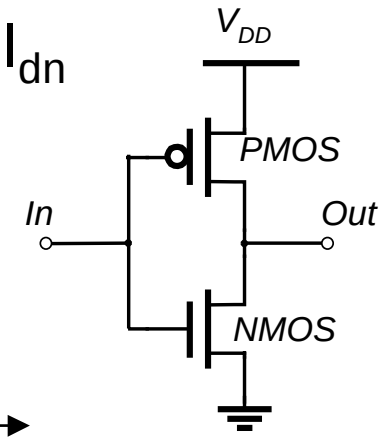
PMOS transistor, 0.25um, L_d = 0.25um, W/L = 1.5, V_{DD} = 2.5V, V_T = -0.4V

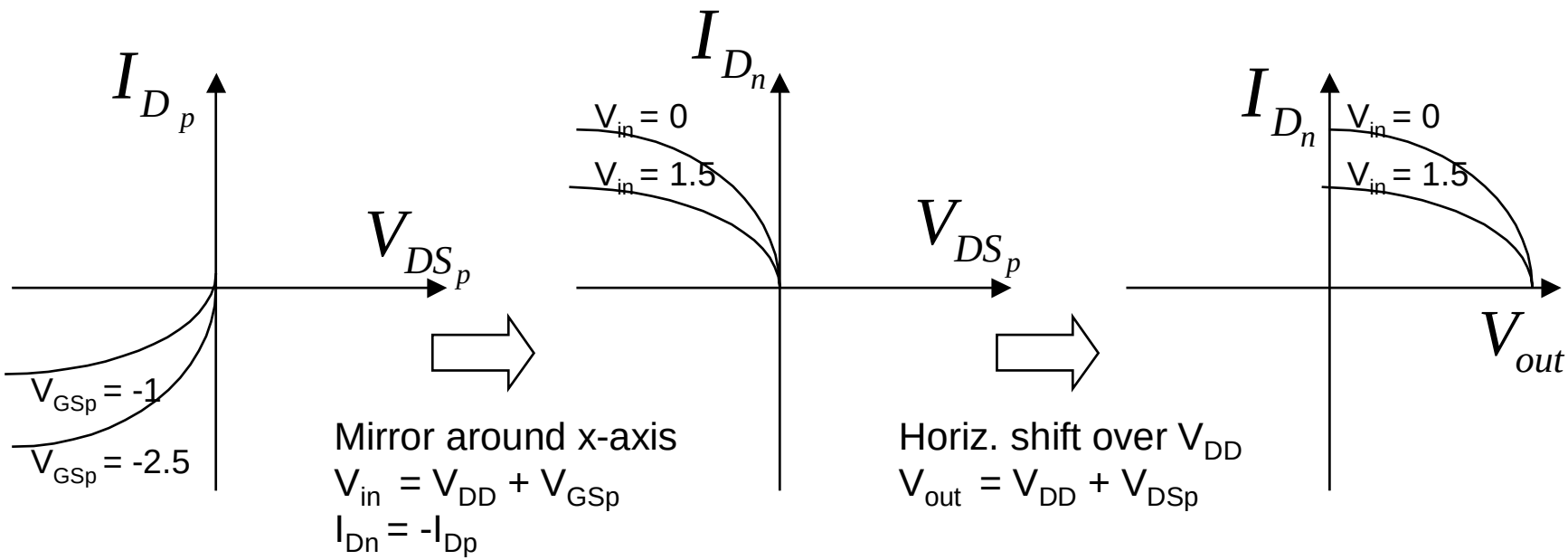
Transforming PMOS I-V Lines

- Want common coordinate set V_{in}, V_{out}, and I_{dn} (load-line plot)

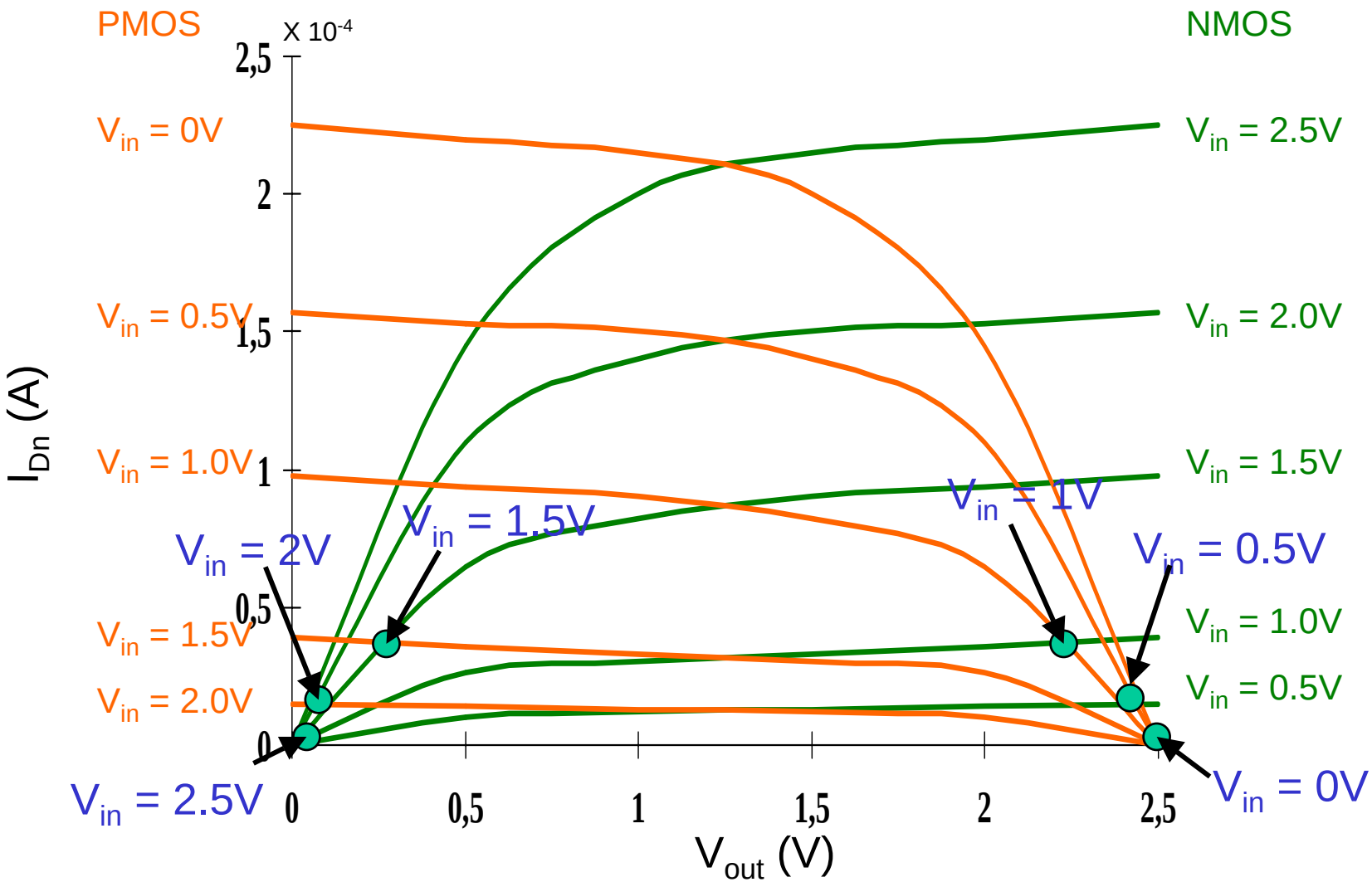
$I_{DSp} = -I_{DSn}$
$V_{GSn} = V_{in}$; $V_{GSp} = V_{in} - V_{DD}$
$V_{DSn} = V_{out}$; $V_{DSp} = V_{out} - V_{DD}$





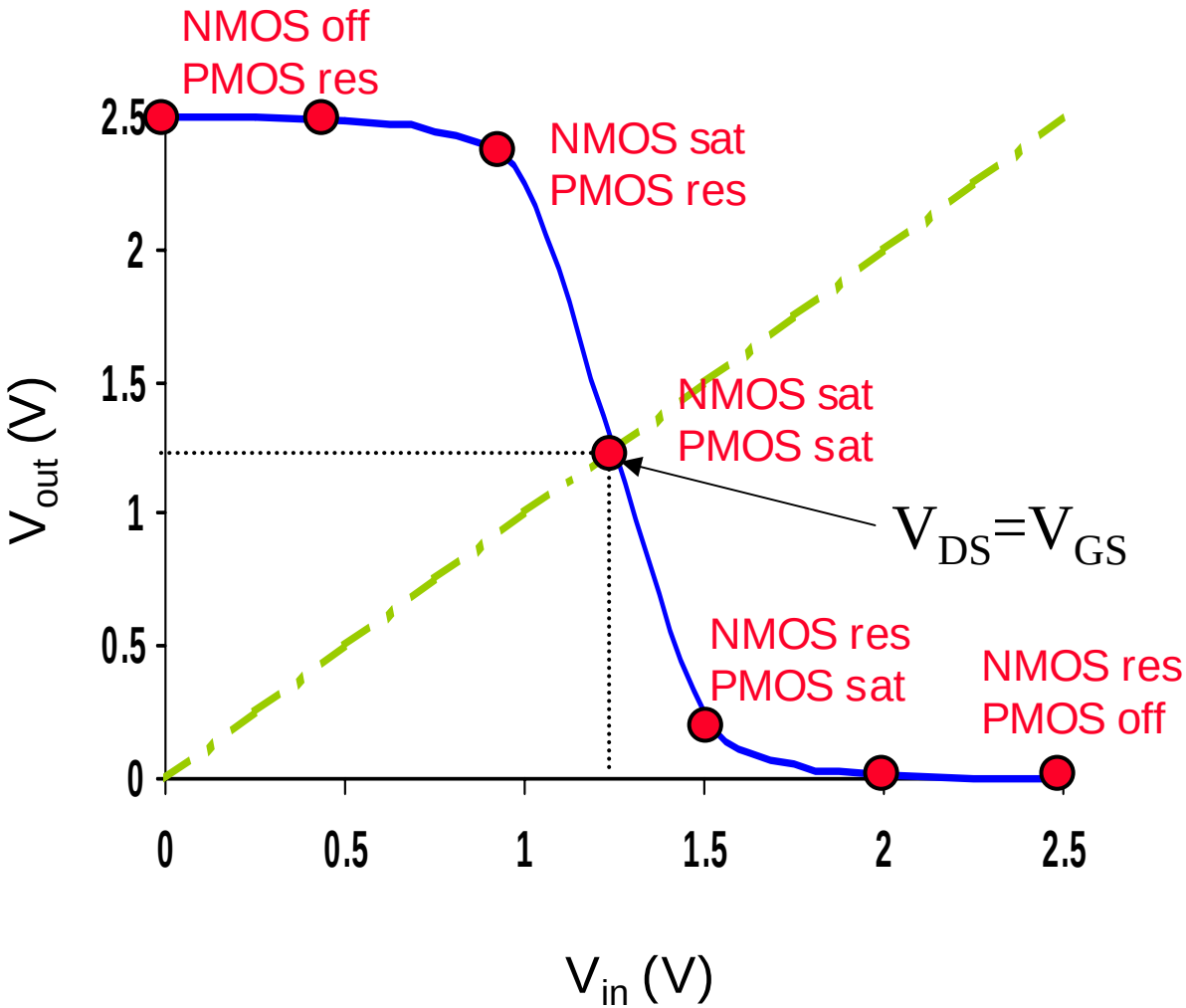


CMOS Inverter Load Lines



0.25um, W/L_n = 1.5, W/L_p = 4.5, V_{DD} = 2.5V, V_{Tn} = 0.4V, V_{Tp} = -0.4V

CMOS Inverter VTC



The response of the inverter is dominated mainly by the output capacitance of the gate $C_{L,}$ which is composed of:

- Drain diffusion capacitance of PMOS & NMOS transistors
- Capacitances of connecting wires
- Input capacitance of the fan-out gates

A fast gate is built either by keeping the output capacitance small or by decreasing the on-resistance of the transistor. The latter is achieved by increasing the W/L ratio.

Relative Transistor Sizing

- When designing static CMOS circuits, balance the driving strengths of the transistors by making the PMOS section wider than the NMOS section to
  - maximize the noise margins and
  - obtain symmetrical characteristics

Switching Threshold as a function of Transistor Ratio

In the transition region both PMOS and NMOS transistors are always saturated, since $V_{DS} = V_{GS}$. By equating the currents through the transistors (assuming velocity saturation and ignoring channel length modulation): $V_{in} = V_{out} = V_M$, $V_{GS} = V_M$ (NMOS), $V_{GS} = V_M - V_{DD}$ (PMOS)

$$k_n V_{DSATn}\left(V_M - V_{Tn} - \frac{V_{DSATn}}{2}\right) + k_p V_{DSATp}\left(V_M - V_{DD} - V_{Tp} - \frac{V_{DSATp}}{2}\right) = 0$$

Solving for V_M yields

$$V_M = \frac{\left(V_{Tn} + \frac{V_{DSATn}}{2}\right) + r\left(V_{DD} + V_{Tp} + \frac{V_{DSATp}}{2}\right)}{1 + r} \quad \text{with} \quad r = \frac{k_p V_{DSATp}}{k_n V_{DSATn}} = \frac{\upsilon_{satp} W_p}{\upsilon_{satn} W_n}$$

Switching Threshold

- V_M where $V_{in} = V_{out}$ (both PMOS and NMOS in saturation since $V_{DS} = V_{GS}$). For large values of V_{DD} (wrt treshold & saturation voltages):

$$V_M \approx rV_{DD}/(1 + r) \text{ where } r = k_p V_{DSATp}/k_n V_{DSATn}$$

- Switching threshold set by the ratio r, which compares the relative driving strengths of the PMOS and NMOS transistors

For V_M to be at $V_{DD}/2$, $r \approx 1$ which is equivalent to sizing the PMOS device so that:

$$(W/L)_p = (W/L)_n \times (V_{DSATn}\, k'_n)/(V_{DSATp}\, k'_p)$$

To move V_M upwards, a larger value of r is required, which means making PMOS wider.

The required ratio of PMOS to NMOS transistor sizes such that switching treshold is set to a desired value V_M

$$\frac{(W/L)_p}{(W/L)_n} = \frac{k'_n V_{DSATn}(V_M - V_{Tn} - V_{DSATn}/2)}{k'_p V_{DSATp}(V_{DD} - V_M + V_{Tp} + V_{DSATp}/2)}$$

For long channel devices or when supply voltage is low (no velocity saturation):

$$V_M = \frac{V_{Tn} + r(V_{DD} + V_{Tp})}{1+r}, \quad r = \sqrt{\frac{-k_p}{k_n}}$$

Switching Threshold Example

- In our generic 0.25 micron CMOS process, using the process parameters V_{DD} = 2.5V, and a minimum size NMOS device ($(W/L)_n$ of 1.5)

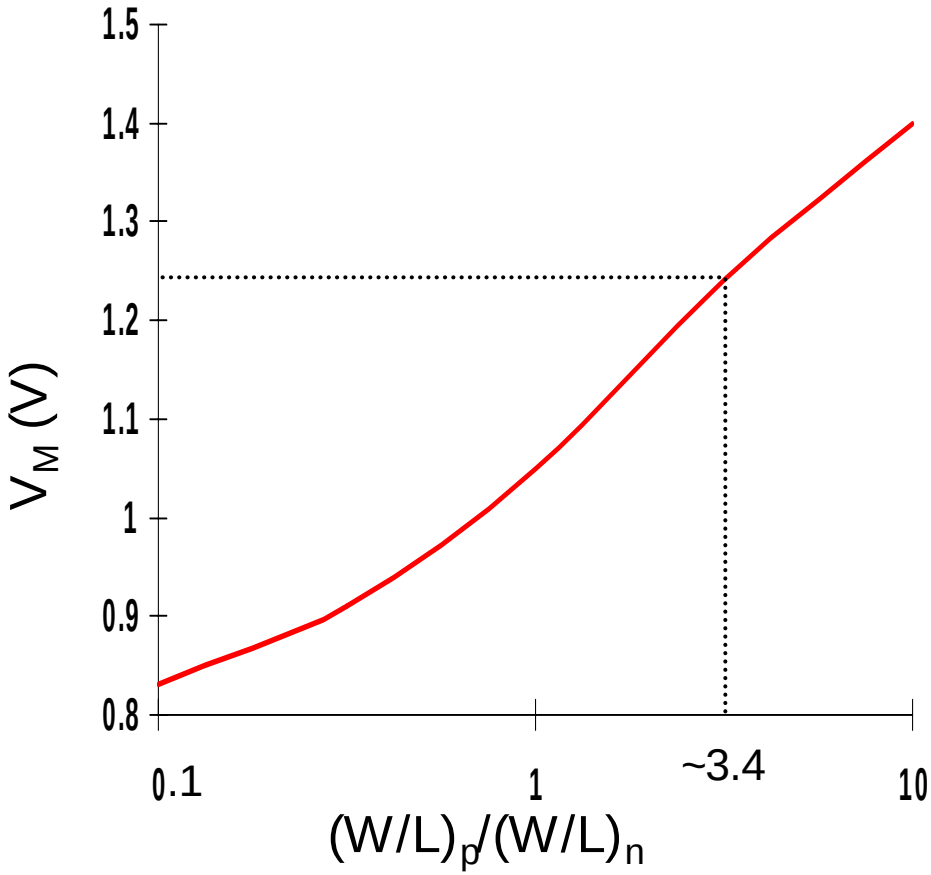
	V_{T0}(V)	$\gamma(V^{0.5})$	V_{DSAT}(V)	$k'(A/V^2)$	$\lambda(V^{-1})$
NMOS	0.43	0.4	0.63	115×10^{-6}	0.06
PMOS	-0.4	-0.4	-1	-30×10^{-6}	-0.1

For V_M = 1.25V (half of V_{DD}):

$$\frac{(W/L)_p}{(W/L)_n} = \frac{115 \times 10^{-6}}{-30 \times 10^{-6}} \times \frac{0.63}{-1.0} \times \frac{(1.25 - 0.43 - 0.63/2)}{(1.25 - 0.4 - 1.0/2)} = 3.5$$

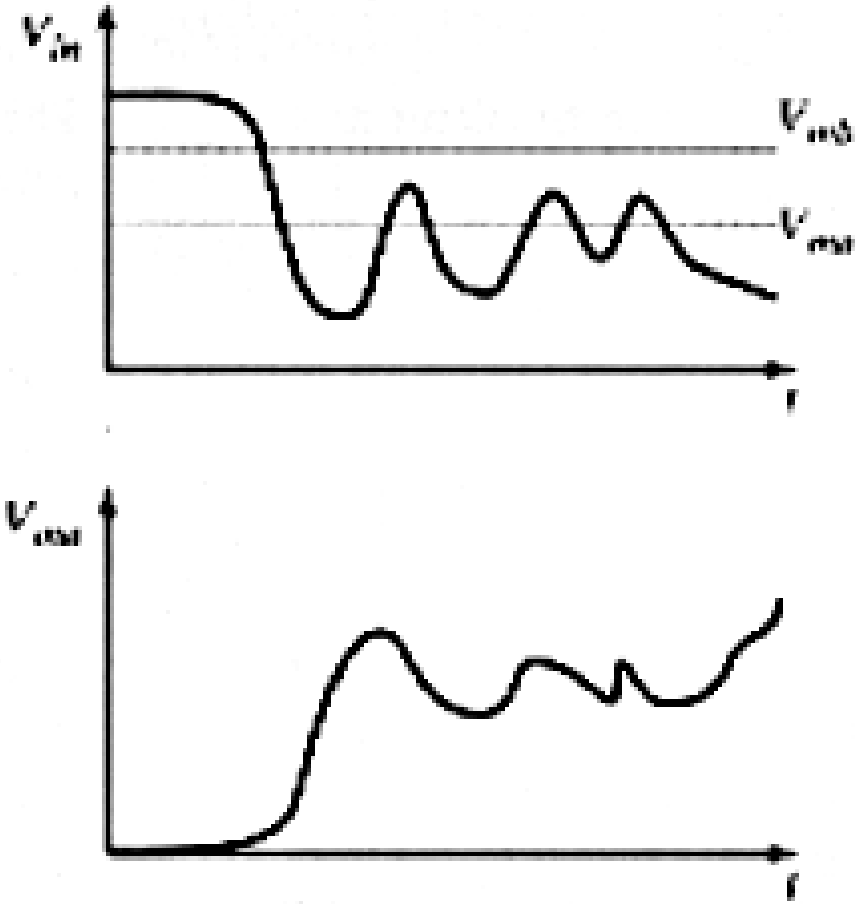
$(W/L)_p$ = 3.5 x 1.5 = 5.25 for a V_M of 1.25V

Simulated Inverter V_M

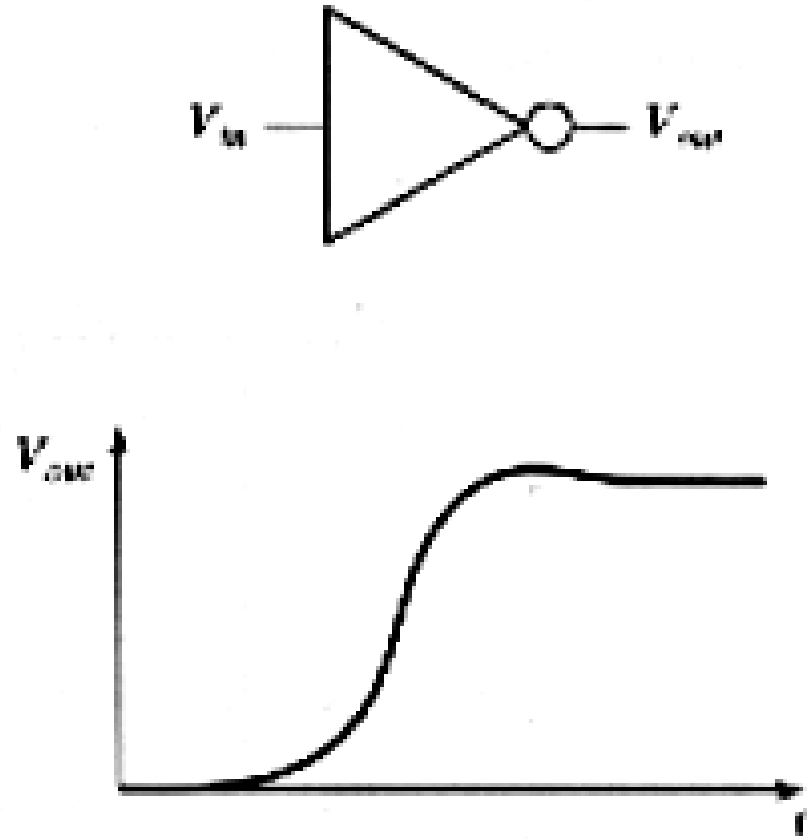


Note: x-axis is semilog

- V_M is relatively insensitive to variations in device ratio
  - setting the ratio to 3, 2.5 and 2 gives V_M's of 1.22V, 1.18V, and 1.13V
- Increasing the width of the PMOS moves V_M towards V_{DD}
- Increasing the width of the NMOS moves V_M toward GND

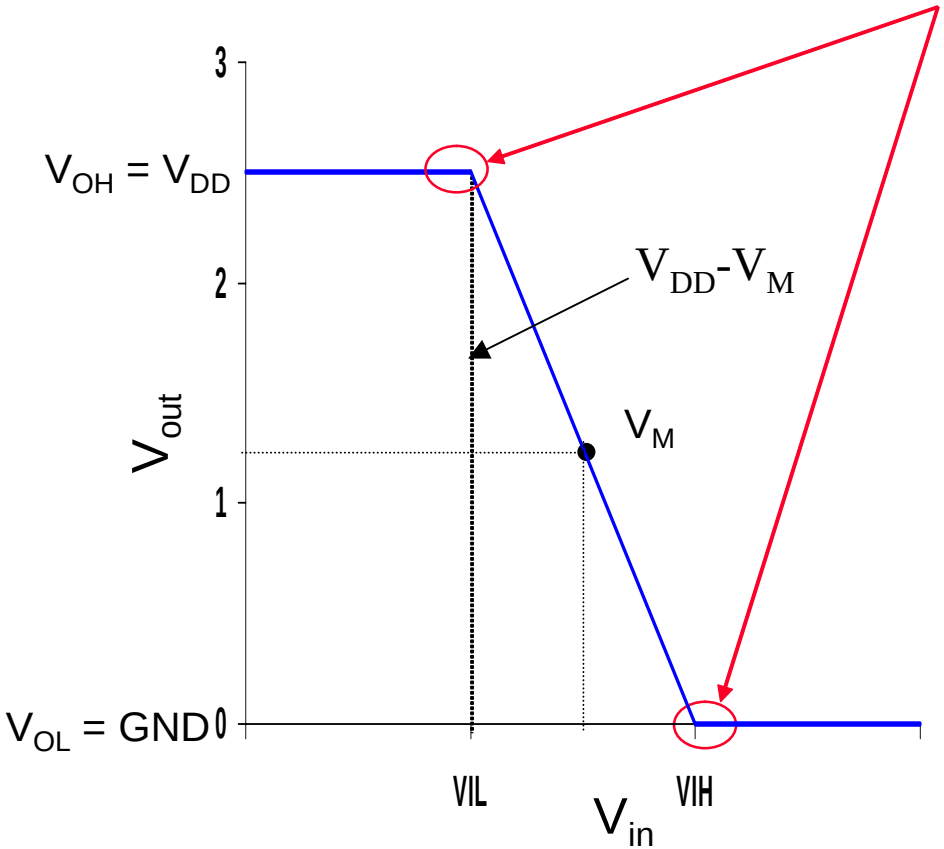


a) Response of standard inverter



b)Response of modified treshold

Noise Margins Determining V_{IH} and V_{IL}



A piece-wise linear approximation of VTC

By definition, V_{IH} and V_{IL} are where $dV_{out}/dV_{in} = -1$ (= gain)

$V_{IH}-V_{IL}=-(V_{OH}-V_{OL})/g=-V_{DD}/g$

$NM_H = V_{DD} - V_{IH}$

$NM_L = V_{IL} - GND$

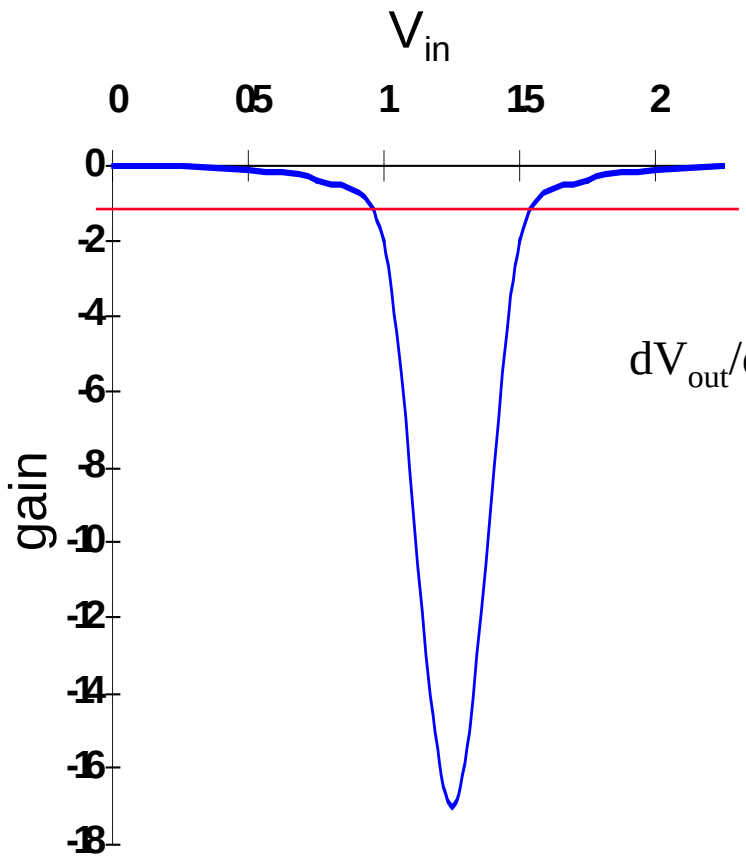
Approximating:

$V_{IH} = V_M - V_M/g$

$V_{IL} = V_M + (V_{DD} - V_M)/g$

So high gain in the transition region is very desirable

Gain Determinates



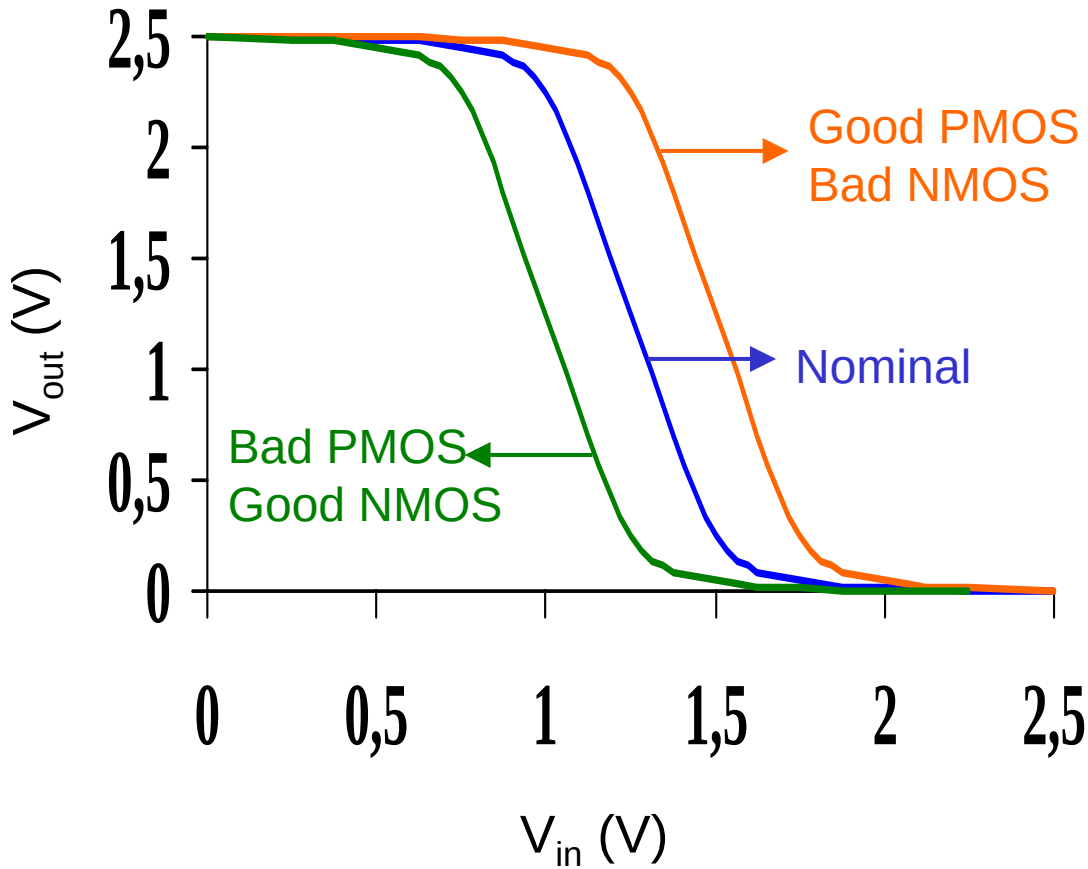
Gain is a strong function of the slopes of the currents in the saturation region, for $V_{in} = V_M$

$$dV_{out}/dV_{in} = g = -\frac{1}{I_D(V_M)}\frac{k_n V_{DSATn} + k_p V_{DSATp}}{\lambda_n - \lambda_p}$$

$$\approx \frac{1+r}{(V_M - V_{Tn} - V_{DSATn}/2)(\lambda_n - \lambda_p)}$$

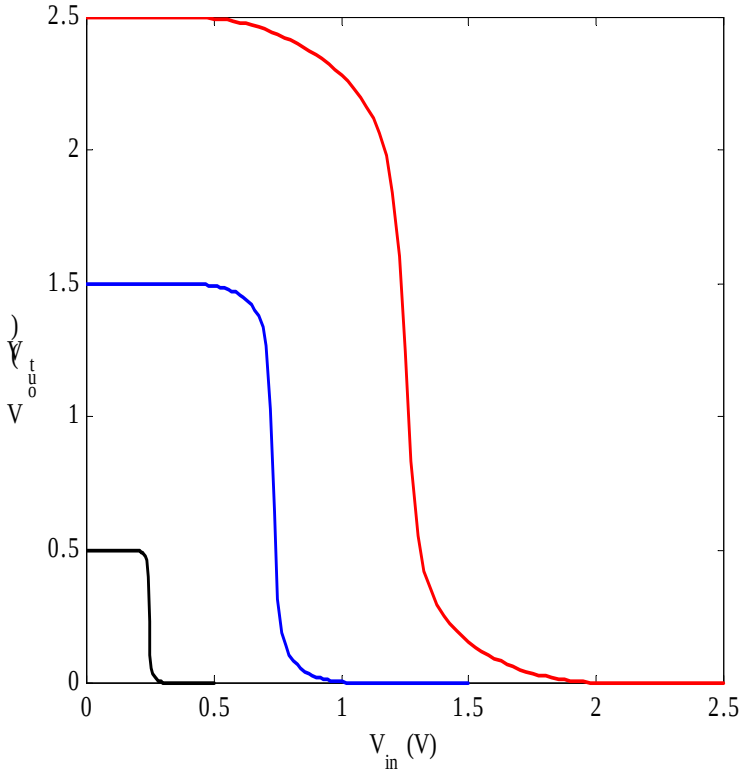
Determined by technology parameters, especially channel length modulation (λ). Only designer influence through supply voltage and V_M (transistor sizing).

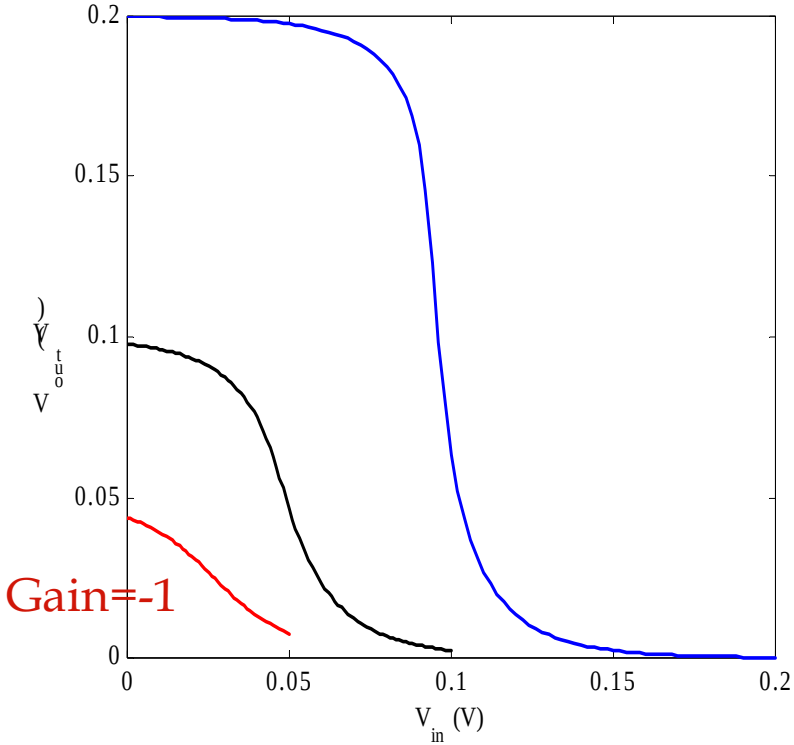
Impact of Process Variation on VTC Curve



rocess variations (mostly) cause a shift in the switching threshold

Gain as a function of VDD





Why don't we operate at low supply voltages?

- Reducing the supply voltage improves gain and energy dissipation. But severely degrades the delay of the gate
- The dc characteristic becomes increasingly sensitive to variations in the device parameters
- While it helps to reduce the internal noise (such as crosstalk), it makes the device more sensitive to external noise sources that do not scale.

To achieve sufficient gain for use in a digital circuit, it is necessary that the supply be at least $4\phi_T$ where $\phi_T = kT/q$

(the only way to operate CMOS inverters below 100mV is to reduce the ambient temperature, that is to cool the circuit)

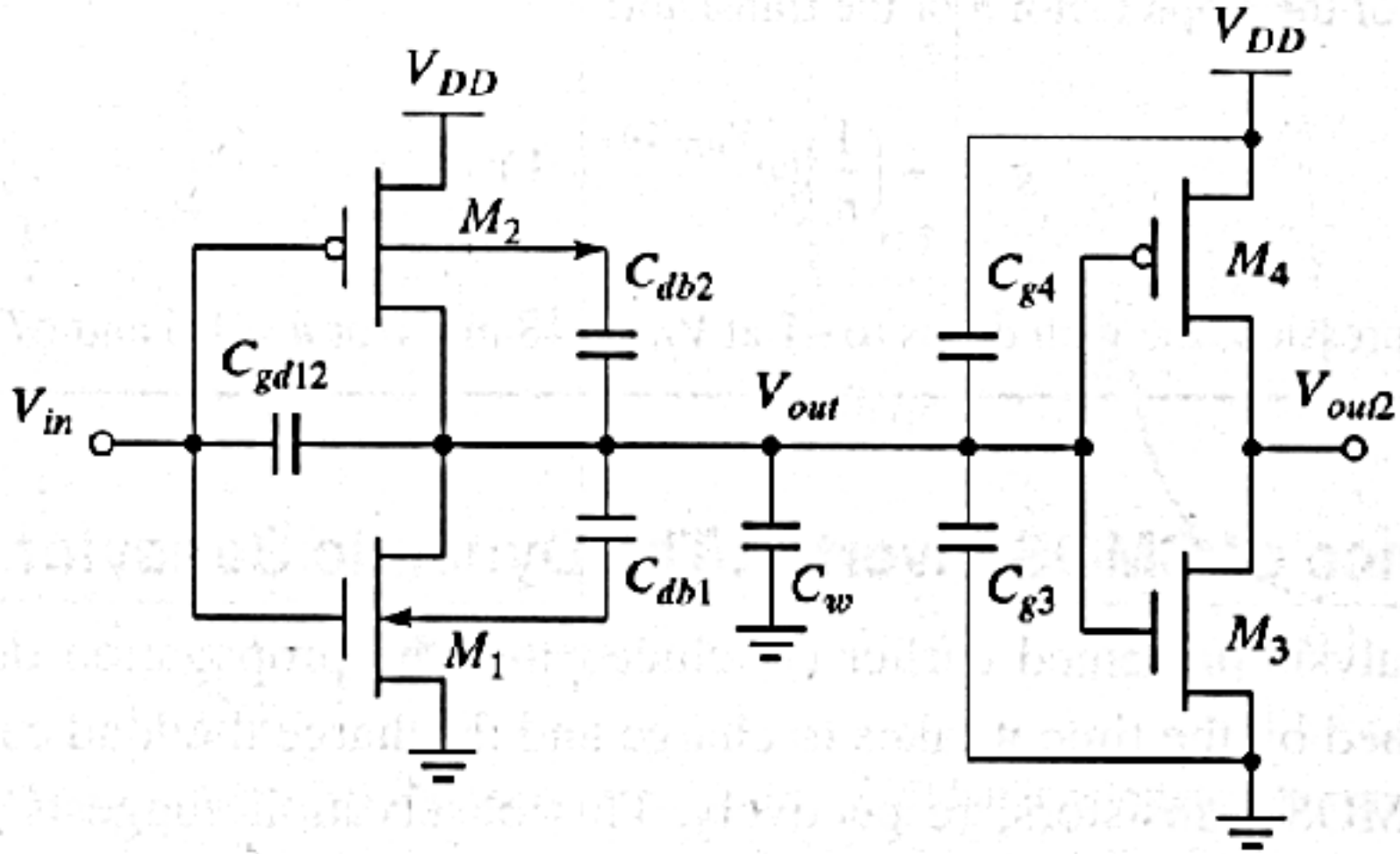


Figure 5-13 Parasitic capacitances, influencing the transient behavior of the cascaded inverter pair.

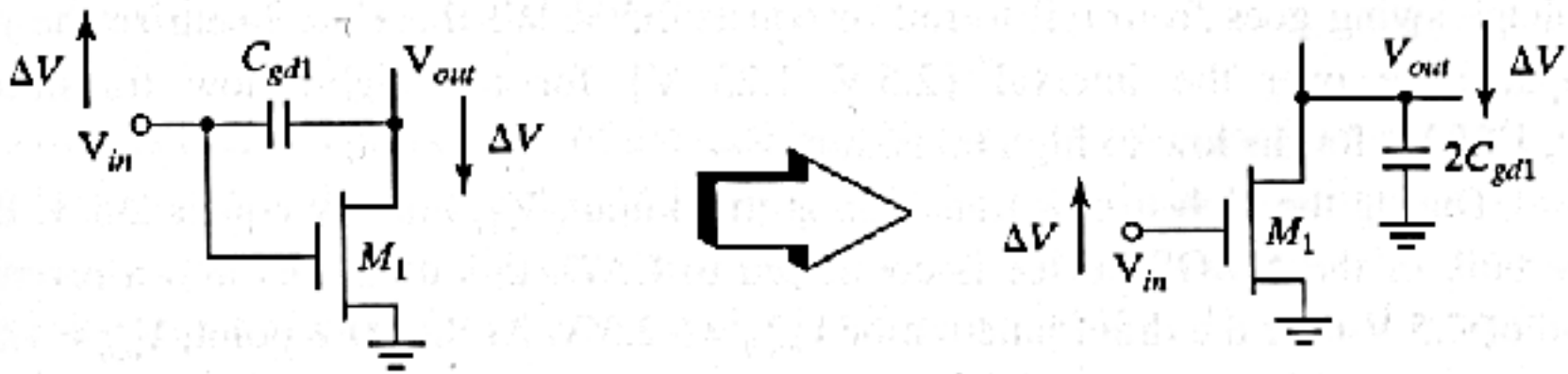


Figure 5-14 The Miller effect—A capacitor experiencing identical but opposite voltage swings at both its terminals can be replaced by a capacitor to ground, whose value is two times the original value.

Diffusion capacitances C_{db1} and C_{db2}

$$C_{eq} = K_{eq} C_{j0}$$

$$K_{eq} = \frac{-\Phi_0^m}{(V_{high} - V_{low})(1-m)} \left[(\Phi_0 - V_{high})^{1-m} - (\Phi_0 - V_{low})^{1-m} \right]$$

voltage swing goes from rail to rail or equals 2.5 V. We therefore linearize the junction capacitance over the interval {2.5 V, 1.25 V} for the high-to-low transition, and {0, 1.25 V} for the low-to-high transition.

During the high-to-low transition at the output, V_{out} initially equals 2.5 V. Because the bulk of the NMOS device is connected to *GND*, this translates into a reverse voltage of 2.5 V over the drain junction or $V_{high} = -2.5$ V. At the 50% point, $V_{out} = 1.25$ V or $V_{low} = -1.25$ V. Evaluating Eq. (5.14) for the bottom plate and sidewall components of the diffusion capacitance yields the following data:

Bottom plate: $K_{eq}\ (m = 0.5, \phi_0 = 0.9) = 0.57$

Sidewall: $K_{eqsw}\ (m = 0.44, \phi_0 = 0.9) = 0.61$

During the low-to-high transition, V_{low} and V_{high} equal 0 V and −1.25 V, respectively, resulting in higher values for K_{eq}:

Bottom plate: $K_{eq}\ (m = 0.5, \phi_0 = 0.9) = 0.79$

Sidewall: $K_{eqsw}\ (m = 0.44, \phi_0 = 0.9) = 0.81$

The PMOS transistor displays a reverse behavior, as its substrate is connected to 2.5 V. Hence, for the high-to-low transition ($V_{low} = 0$, $V_{high} = -1.25$ V), we have

Bottom plate: $K_{eq}\ (m = 0.48, \phi_0 = 0.9) = 0.79$

Sidewall: $K_{eqsw}\ (m = 0.32, \phi_0 = 0.9) = 0.86$

Finally, for the low-to-high transition ($V_{low} = -1.25$ V, $V_{high} = -2.5$ V), we have

Bottom plate: $K_{eq}\ (m = 0.48, \phi_0 = 0.9) = 0.59$

Sidewall: $K_{eqsw}\ (m = 0.32, \phi_0 = 0.9) = 0.7$

Gate Capacitances of Fan-Out C_{g3} and C_{g4}

$$C_{fan\text{-}out} = C_{gate}\,(NMOS) + C_{gate}\,(PMOS)$$

$$= (C_{GSOn} + C_{GDOn} + W_n L_n C_{ox}) + (C_{GSOp} + C_{GDOp} + W_p L_p C_{ox})$$

Table 5-1 Inverter transistor data.

	W/L	AD (μm^2)	PD (μm)	AS (μm^2)	PS (μm)
NMOS	0.375/0.25	0.3 (19 λ^2)	1.875 (15λ)	0.3 (19 λ^2)	1.875 (15λ)
PMOS	1.125/0.25	0.7 (45 λ^2)	2.375 (19λ)	0.7 (45 λ^2)	2.375 (19λ)

From Table 3-5:

Overlap capacitance: CGD0(NMOS) = 0.31 fF/μm; CGDO(PMOS) = 0.27 fF/μm

Bottom junction capacitance: CJ(NMOS) = 2 fF/μm^2; CJ(PMOS) = 1.9 fF/μm^2

Sidewall junction capacitance: CJSW(NMOS) = 0.28 fF/μm; CJSW(PMOS) = 0.22 fF/μm

Gate capacitance: C_{ox}(NMOS) = C_{ox}(PMOS) = 6 fF/μm^2

CMOS Inverters

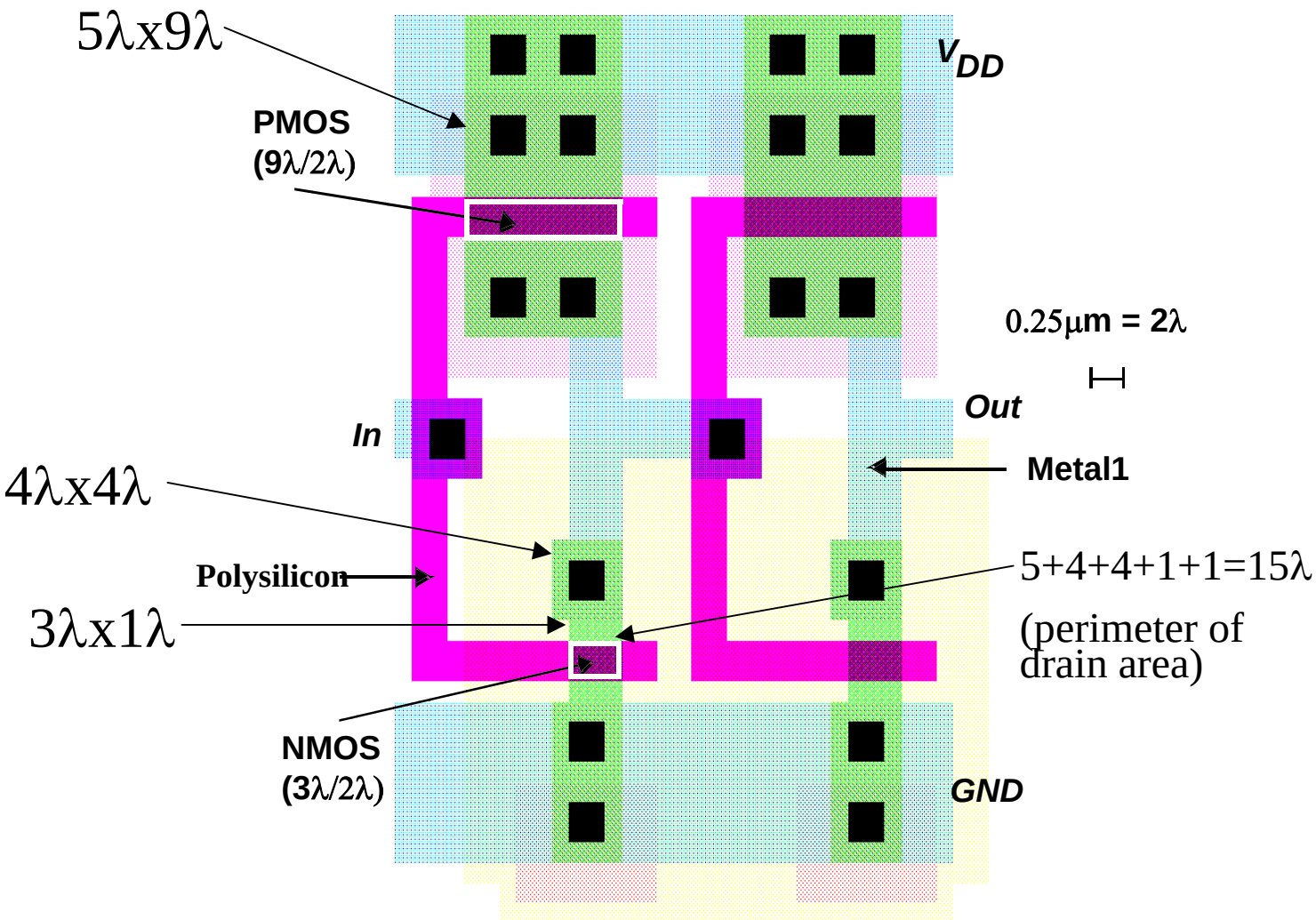


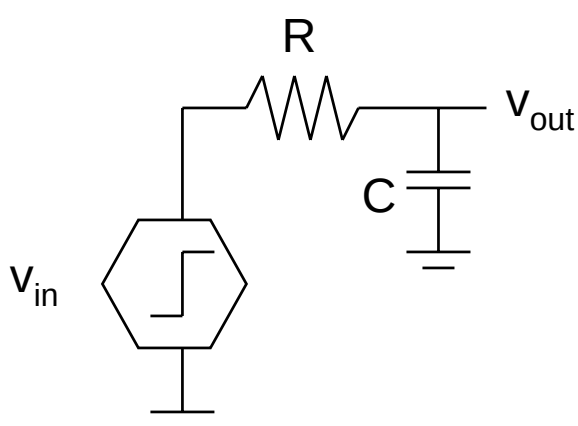
Table 5-2 Components of C_L (for high-to-low and low-to-high transitions).

Capacitor	Expression	Value (fF) (H → L)	Value (fF) (L → H)
C_{gd1}	$2\ CGD0_n\ W_n$	0.23	0.23
C_{gd2}	$2\ CGD0_p\ W_p$	0.61	0.61
C_{db1}	$K_{eqn}\ AD_n\ CJ + K_{eqswn}\ PD_n\ CJSW$	0.66	0.90
C_{db2}	$K_{eqp}\ AD_p\ CJ + K_{eqswp}\ PD_p\ CJSW)$	1.5	1.15
C_{g3}	$(CGD0_n+CGSO_n)\ W_n + C_{ox}\ W_n\ L_n$	0.76	0.76
C_{g4}	$(CGD0_p+CGSO_p)\ W_p + C_{ox}\ W_p\ L_p$	2.28	2.28
C_w	From Extraction	0.12	0.12
C_L	Σ	6.1	6.0

PRORAGATION DELAY

$$I = C\frac{dV}{dt} \qquad dt = \frac{C}{I}dV \qquad t_p = \int_{V_1}^{V_2} \frac{C_L}{i}\, dV$$

both C_L and i are nonlinear functions of V an exact computation of this equation is very difficult. Instead, we use the simplified switch model.



$$v_{out}(t) = (1 - e^{-t/\tau})V$$

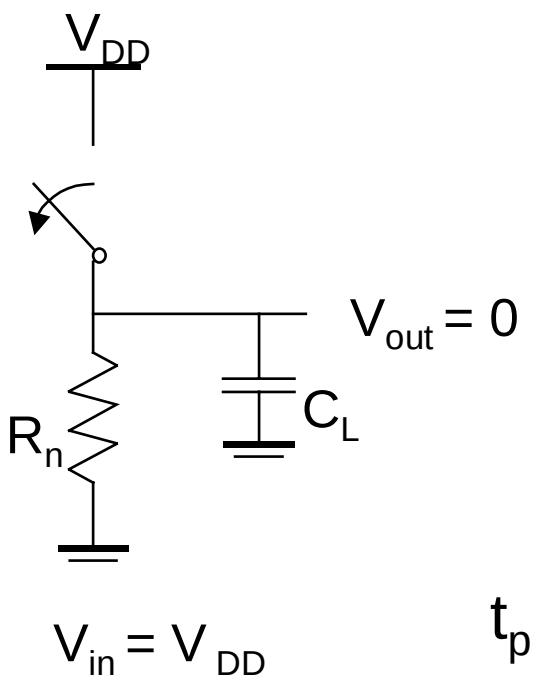
where $\tau = RC$

Time to reach 50% point is

$t = \ln(2)\,\tau = 0.69\,\tau$

Inverter Propagation Delay

- Propagation delay is proportional to the time-constant of the network formed by the pull-down resistor and the load capacitance



$$t_{pHL} = f(R_n, C_L)$$

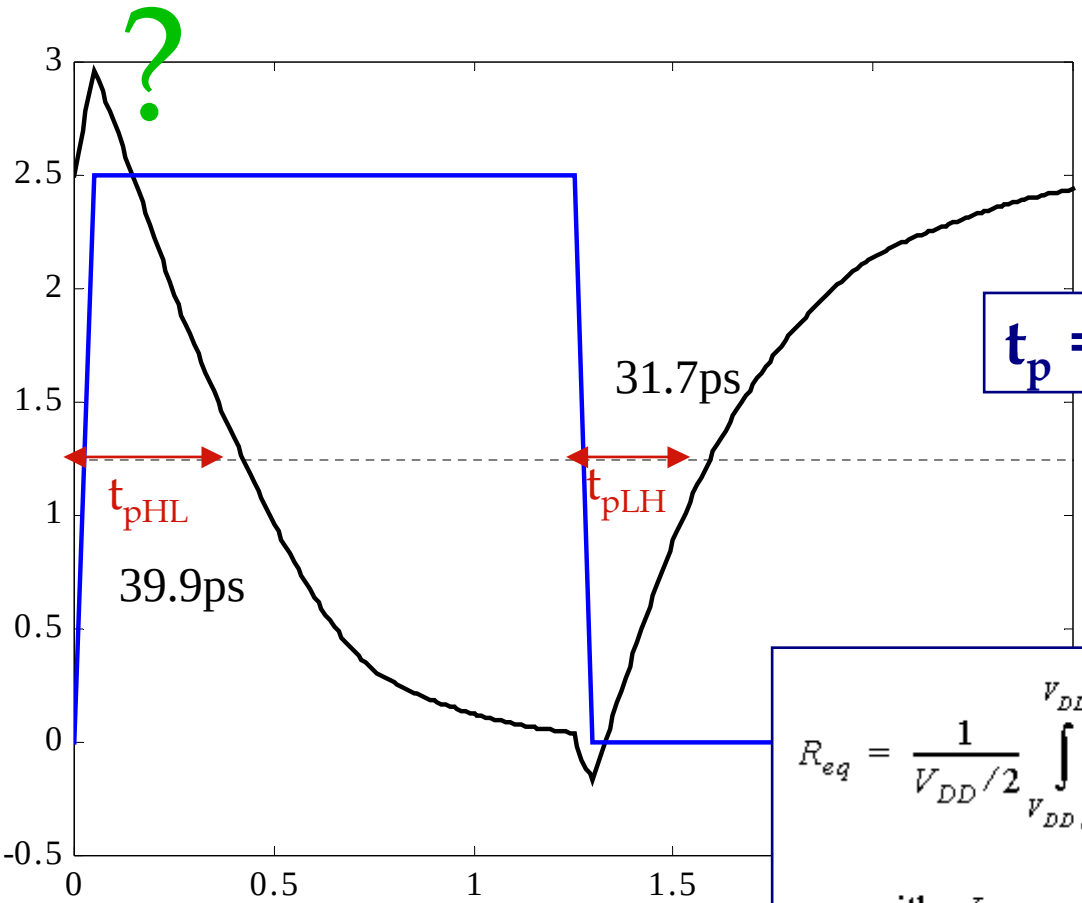
$$t_{pHL} = \ln(2)\, R_{eqn}\, C_L = 0.69\, R_{eqn}\, C_L$$

$$t_{pLH} = \ln(2)\, R_{eqp}\, C_L = 0.69\, R_{eqp}\, C_L$$

$$t_p = (t_{pHL} + t_{pLH})/2 = 0.69\, C_L(R_{eqn} + R_{eqp})/2$$

- To equalize rise and fall times make the on-resistance of the NMOS and PMOS approximately equal.

Transient Response



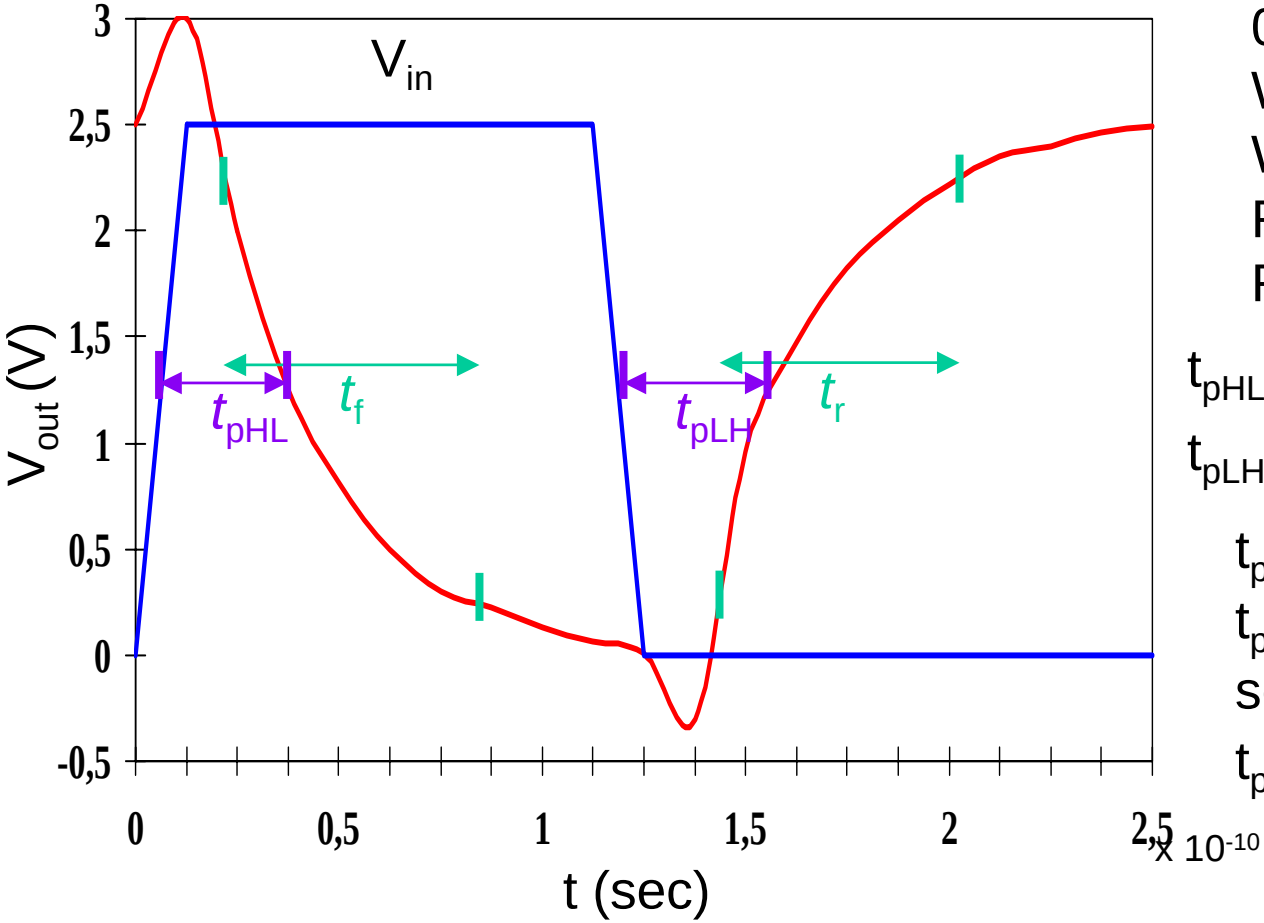
$$t_p = (t_{pHL}+t_{pLH})/2$$

$$\mathbf{t_p = 0.69\ C_L\ (R_{eqn}+R_{eqp})/2}$$

$$R_{eq} = \frac{1}{V_{DD}/2}\int_{V_{DD}/2}^{V_{DD}} \frac{V}{I_{DSAT}(1+\lambda V)}dV \approx \frac{3}{4}\frac{V_{DD}}{I_{DSAT}}\left(1-\frac{7}{9}\lambda V_{DD}\right)$$

$$\text{with}\quad I_{DSAT} = k'\frac{W}{L}\left((V_{DD}-V_T)V_{DSAT}-\frac{V_{DSAT}^2}{2}\right)$$

Inverter Transient Response



V_{DD}=2.5V
0.25μm
W/L_n = 1.5
W/L_p = 4.5
R_{eqn}= 13 kΩ (÷ 1.5)
R_{eqp}= 31 kΩ (÷ 4.5)

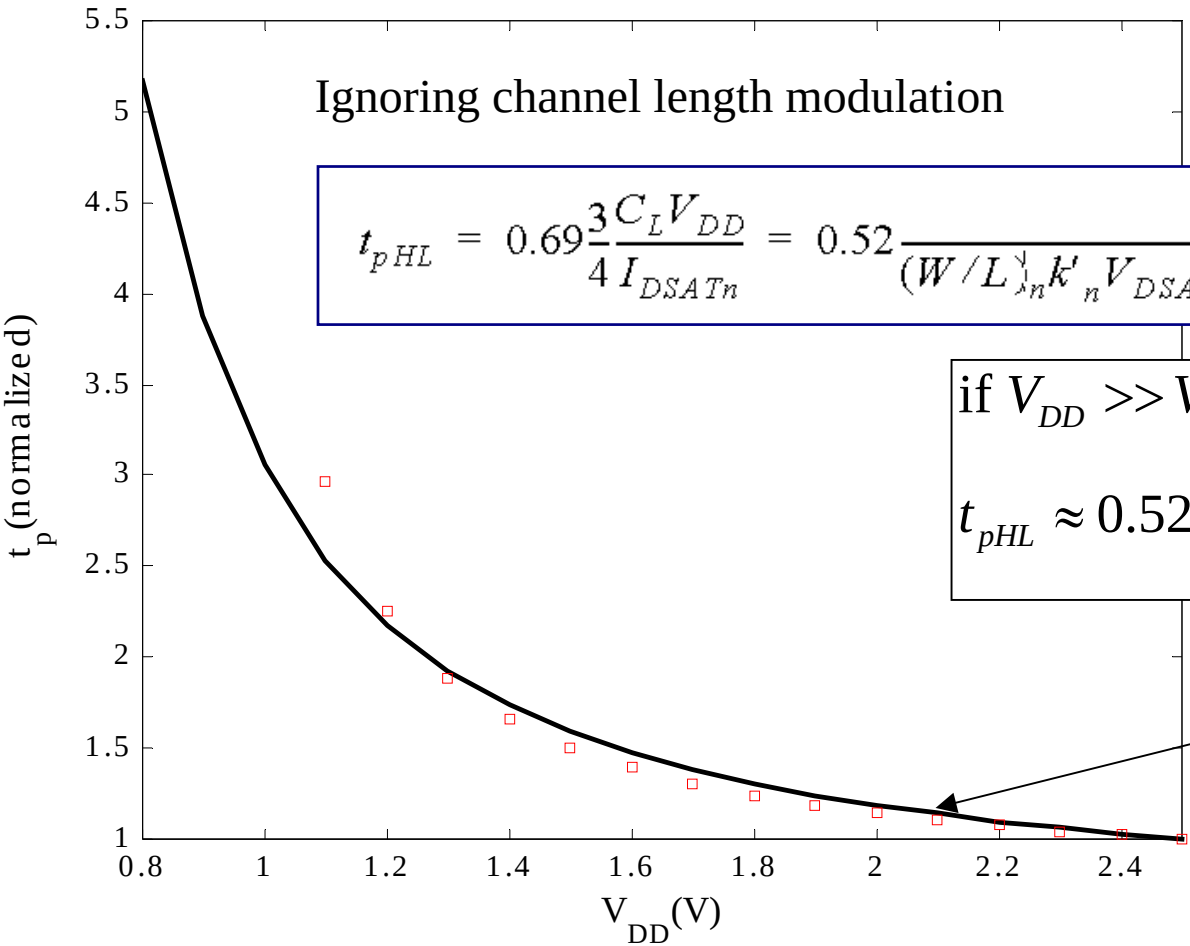
t_{pHL} =0.69×13/1.5 ×6.1

t_{pLH} =0.69 ×31/4.5 ×6

t_{pHL} = 36 psec
t_{pLH} = 29 psec
so
t_p = 32.5 psec

From simulation: t_{pHL} = 39.9 psec and t_{pLH} = 31.7 psec

Delay as a function of V_{DD}



Ignoring channel length modulation

$$t_{pHL} = 0.69\frac{3}{4}\frac{C_L V_{DD}}{I_{DSATn}} = 0.52\frac{C_L V_{DD}}{(W/L)_n k'_n V_{DSATn}(V_{DD} - V_{Tn} - V_{DSATn}/2)}$$

if $V_{DD} >> V_{Tn} + V_{DSATn}/2$

$$t_{pHL} \approx 0.52\frac{C_L}{(W/L)_n k'_n V_{DSATn}}$$

Independent of V_{DD}

Dots indicate the values calculated. Assumed velocity sat. Deviation at low supply voltages

Design for Performance

- Keep capacitances small
  Good design practice requires keeping the drain diffusion areas as small as possible
- Increase transistor sizes
  watch out for self-loading!
  intrinsic capacitance (i.e. diffusion capacitance) starts to dominate extrinsic load (wiring & f.out)
- Increase V_{DD}
  (????) trade off energy dissipation
  oxide break-down, hot carrier effects
  minimal improvement above a certain level

- So far we have considered widened PMOS so that its resistance matches that of the pull-down NMOS to yield symmetrical VTC.
- However, this does not imply that this ratio also yields the minimum overall propagation delay.
- If symmetry and reduced noise margins are not of prime concern, it is possible to speed up the inverter by reducing the width of the PMOS device.

Consider the 2 identical cascaded CMOS inverters, with approximate load capacitance of the 1. Gate C_L :

NMOS to PMOS Ratio

$$C_L = (C_{dp1} + C_{dn1}) + (C_{gp2} + C_{gn2}) + C_w$$

$$\beta = (W/L)_p / (W/L)_n \;\rightarrow\; C_{dp1} = \beta C_{dn1}\;,\; C_{gp2} = \beta C_{gn2}$$

$$C_L = (1+\beta)(C_{dn1} + C_{gn2}) + C_w \quad \text{since } t_p = 0.69 C_L \left(\frac{\mathrm{R}_{\mathrm{e}qn} + \mathrm{R}_{\mathrm{e}qp}}{2} \right)$$

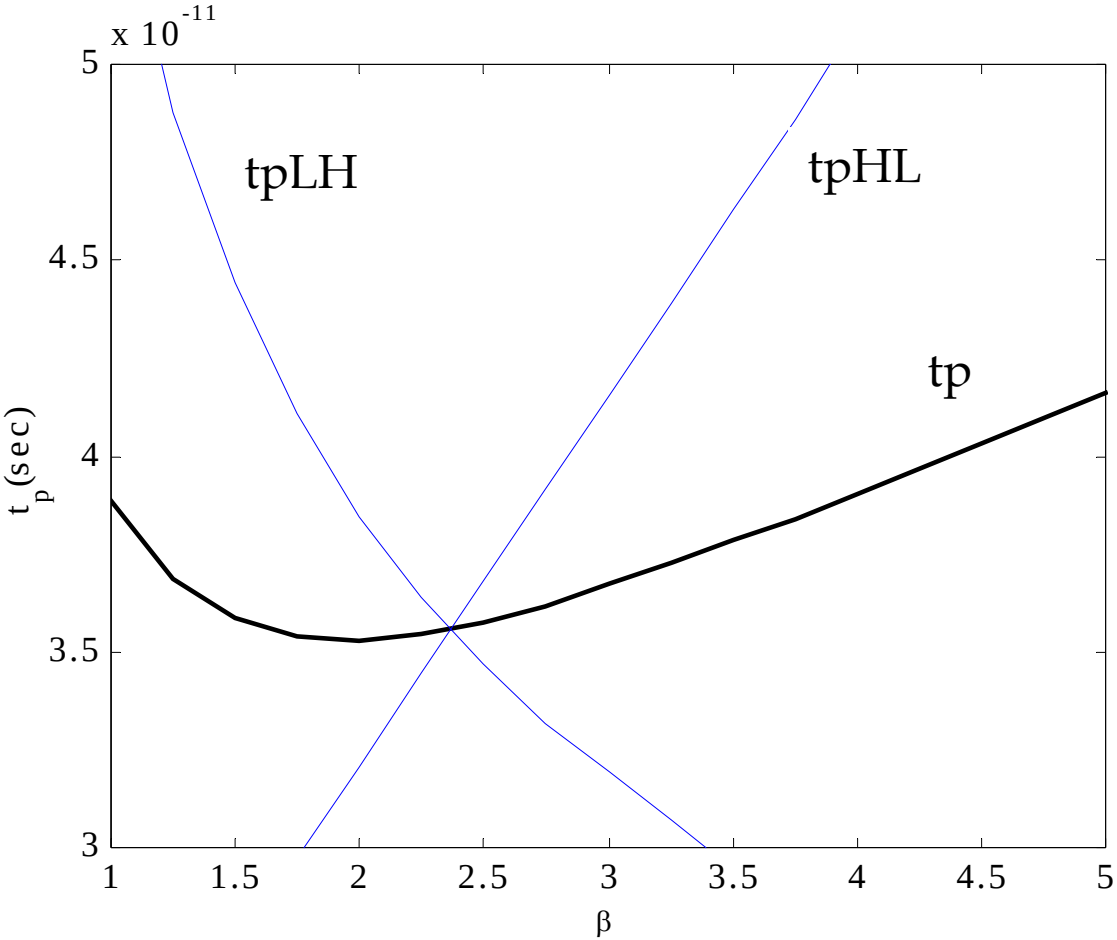
$$t_p = \frac{0.69}{2} \left[(1+\beta)(C_{dn1} + C_{gn2}) + C_w \right] \left(\mathrm{R}_{\mathrm{e}qn} + \frac{\mathrm{R}_{\mathrm{e}qp}}{\beta} \right)$$

$$= 0.345 \left[(1+\beta)(C_{dn1} + C_{gn2}) + C_w \right] \mathrm{R}_{\mathrm{e}qn} \left(1 + \frac{r}{\beta} \right) \quad \text{where, } r = \mathrm{R}_{\mathrm{e}qp} / \mathrm{R}_{\mathrm{e}qn}$$

to find the optimal value of β, setting $\partial t_p / \partial \beta = 0$ we find

$$\beta_{opt} = \sqrt{r \left(1 + \frac{C_w}{C_{dn1} + C_{gn2}} \right)} \quad \text{or } \beta_{opt} = \sqrt{r} \;\; \text{when } C_{dn1} + C_{gn2} >> C_w$$

NMOS/PMOS ratio



β of 2.4 (= 31 kΩ/13 kΩ) gives symmetrical response

β of 1.9 gives optimal performance (1.6 calculated)

NMOS/PMOS Ratio

- So far have sized the PMOS and NMOS so that the R_{eq}'s match (ratio of 3 to 3.5)
  - symmetrical VTC
  - equal high-to-low and low-to-high propagation delays

- If speed is the only concern, reduce the width of the PMOS device!
  - widening the PMOS degrades the t_{pHL} due to larger parasitic capacitance

$$\beta = (W/L_p)/(W/L_n)$$

$r = R_{eqp}/R_{eqn}$ (resistance ratio of identically-sized PMOS and NMOS)

$\beta_{opt} = \sqrt{r}$ when wiring capacitance is negligible

The next question is how transistor sizing (sizing both NMOS and PMOS with the same ratio) impacts the performance of the gate.

To answer this question, we must use a sizing factor S, which relates the transistor sizes of our inverter to a reference gate, typically a minimum sized inverter.

Device Sizing for Performance

- Divide capacitive load, C_L, into
  - C_{int} : intrinsic - diffusion and Miller effect
  - C_{ext} : extrinsic - wiring and fanout

$$t_p = 0.69\,R_{eq}(C_{int} + C_{ext})$$

$$t_p = 0.69\ R_{eq}\ C_{int}\ (1 + C_{ext}/C_{int}) = t_{p0}\ (1 + C_{ext}/C_{int})$$

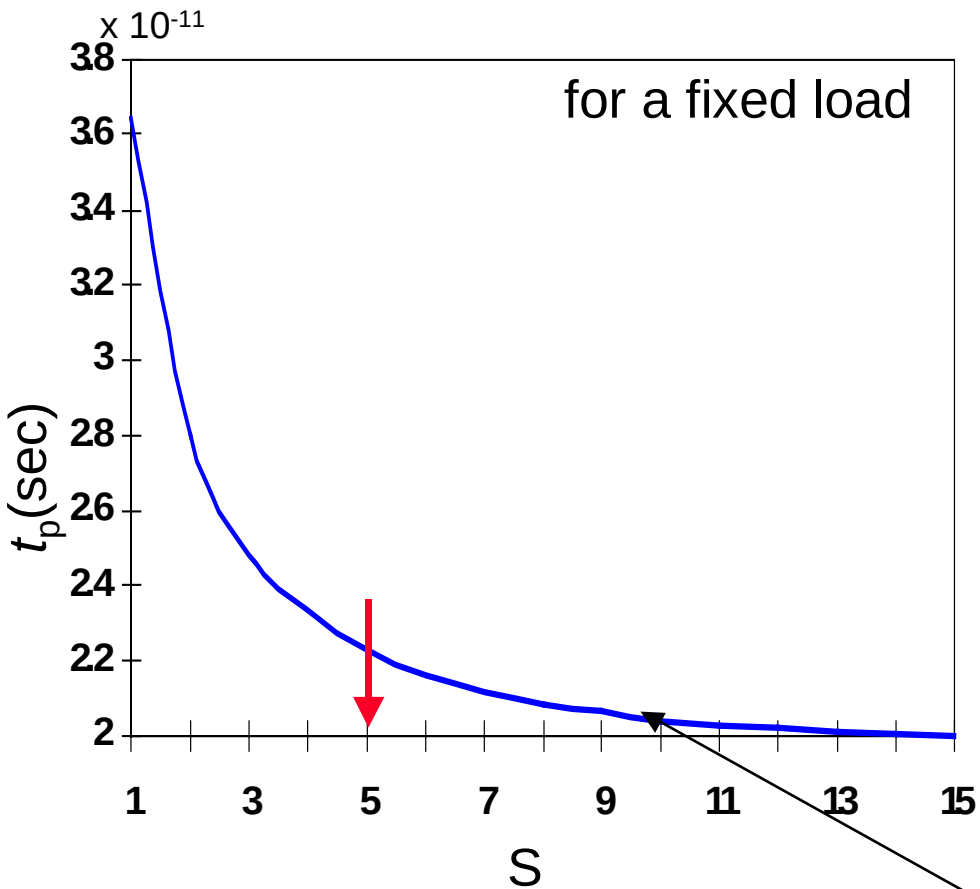
  - where $t_{p0} = 0.69\ R_{eq}\ C_{int}$ is the intrinsic (unloaded) delay of the gate ($C_{ext} = 0$)

- Widening both PMOS and NMOS by a factor S reduces R_{eq} by an identical factor ($R_{eq} = R_{ref}/S$), but raises the intrinsic capacitance by the same factor ($C_{int} = SC_{iref}$)

$$t_p = 0.69\ R_{ref}\ C_{iref}\ (1 + C_{ext}/(SC_{iref})) = t_{p0}(1 + C_{ext}/(SC_{iref}))$$

  - t_{p0} is independent of the sizing of the gate; *with no load the drive of the gate is totally offset by the increased capacitance*
  - any S sufficiently larger than (C_{ext}/C_{int}) yields the best performance gains with least silicon area impact

Sizing Impacts on Delay



The majority of the improvement is already obtained for S = 5. Sizing factors larger than 10 barely yield any extra gain (and cost significantly more area).

self-loading effect (intrinsic capacitance dominates)

Sizing a Chain of Inverters

While sizing up an inverter reduces its delay, it also increases its input capacitance. Gate sizing without taking into account its impact on the delay of the preceding gates is a purely academic enterprise. Therefore, a more relevant problem is determining the optimum sizing of a gate when **embedded in a real environment.**

Impact of Fanout on Delay

- Extrinsic capacitance, C_{ext}, is a function of the fanout of the gate - the larger the fanout, the larger the external load.

- First determine the input loading effect of the inverter. Both input gate capacitance C_g and intrinsic output capacitance C_{int} are proportional to the gate sizing, so $C_{int} = \gamma C_g$ is independent of gate sizing and

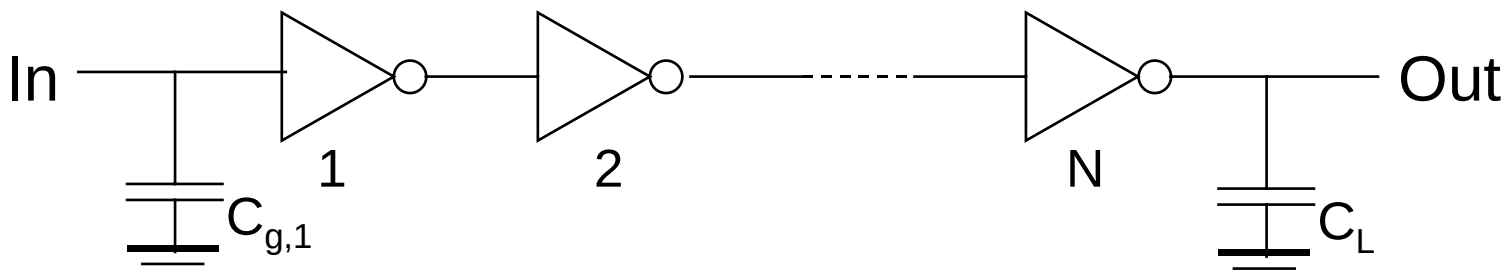
$$t_p = t_{p0} (1 + C_{ext}/ \gamma C_g) = t_{p0} (1 + f/\gamma)$$

i.e., the delay of an inverter is a function of the ratio between its external load capacitance and its input gate capacitance: the effective fan-out f

$$f = C_{ext}/C_g$$

Inverter Chain

- Real goal is to minimize the delay through an inverter chain



the delay of the j-th inverter stage is (C_W ignored):

$$t_{p,j} = t_{p0}\,(1 + C_{g,j+1}/(\gamma C_{g,j})) = t_{p0}(1 + f_j/\,\gamma)$$

and $$t_p = t_{p1} + t_{p2} + \ldots + t_{pN}$$

so $$t_p = \sum t_{p,j} = t_{p0} \sum (1 + C_{g,j+1}/(\gamma C_{g,j}))\text{ , with } C_{g,N+1}=C_L$$

- If C_L is given
  - How should the inverters be sized?
  - How many stages are needed to minimize the delay?

This equation has N-1unknowns, being $C_{g,2}$, $C_{g,3}$, ... $C_{g,N}$. The minimum delay can be found by taking N-1 partial derivatives and equating them to zero:

$$\frac{\partial t_p}{\partial C_{g,j}} = 0$$

The result is a set of constraints:

$$\frac{C_{g,j+1}}{C_{g,j}} = \frac{C_{g,j}}{C_{g,j-1}} \quad \text{with (j=2, ... N)}$$

In other words, the optimum size of each inverter is the geometric mean of its neighbors sizes:

$$C_{g,j} = \sqrt{C_{g,j-1} C_{g,j+1}}$$

Sizing the Inverters in the Chain

- The optimum size of each inverter is the geometric mean of its neighbors – meaning that if each inverter is sized up by the same factor f wrt the preceding gate, it will have the same effective fan-out and the same delay

$$f = \sqrt[N]{C_L/C_{g,1}} = \sqrt[N]{F}$$

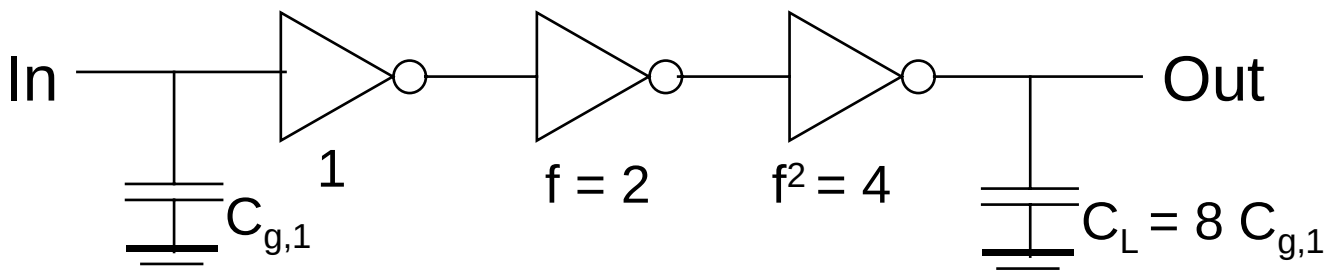
where F represents the overall effective fan-out of the circuit ($F = C_L/C_{g,1} = f^N$)

and the minimum delay through the inverter chain is

$$t_p = N\, t_{p0}\, (1 + (\sqrt[N]{F}\,) / \gamma)$$

- The relationship between t_p and F is linear for one inverter, square root for two, etc.

Example of Inverter Chain Sizing



- $C_L/C_{g,1}$ has to be evenly distributed over N = 3 inverters

$$C_L/C_{g,1} = 8/1$$

$$f = \sqrt[3]{8} = 2$$

Determining N: Optimal Number of Inverters

- What is the optimal value for N given F (=f^N) ?
  - if the number of stages is too large, the intrinsic delay of the stages becomes dominant
  - if the number of stages is too small, the effective fan-out of each stage becomes dominant

$$t_p = N t_{p0}(1 + \sqrt[N]{F} / \gamma)$$

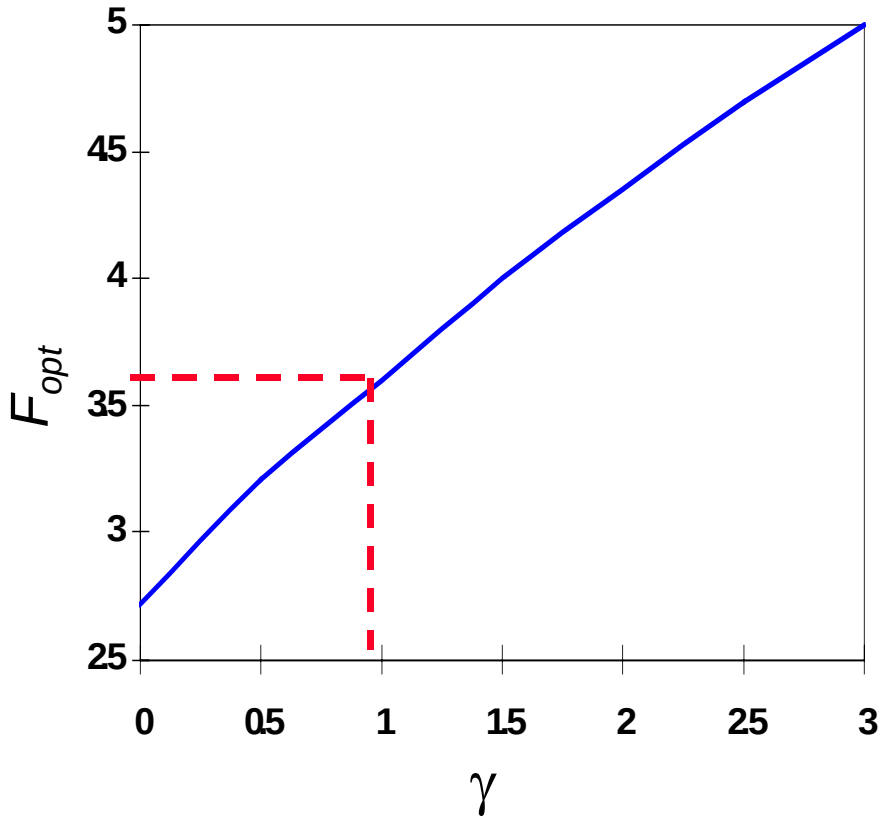
- The optimum N is found by differentiating the minimum delay expression and setting the result to 0, giving

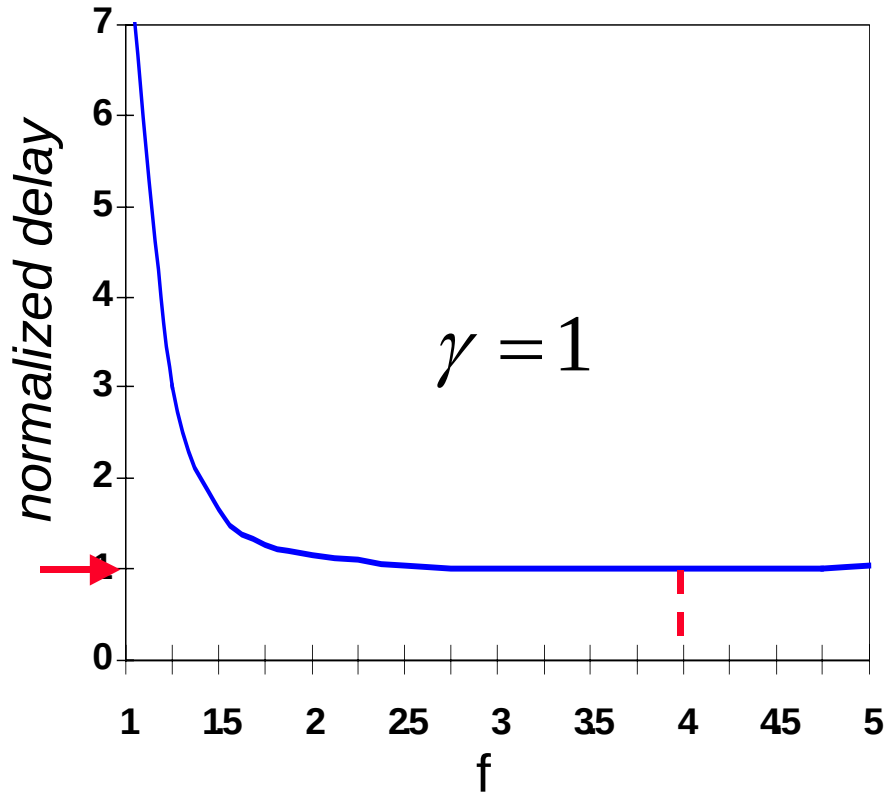
$$\gamma + \sqrt[N]{F} - \frac{\sqrt[N]{F} \ln F}{N} = 0 \qquad \text{or equivalently:} \quad f = e^{(1+\gamma/f)}$$

- For γ = 0 (self-loading is ignored) N = ln (F) and the effective-fan out becomes f = e = 2.71828

- For γ = 1 (the typical case) the optimum effective fan-out (tapering factor) turns out to be close to 3.6

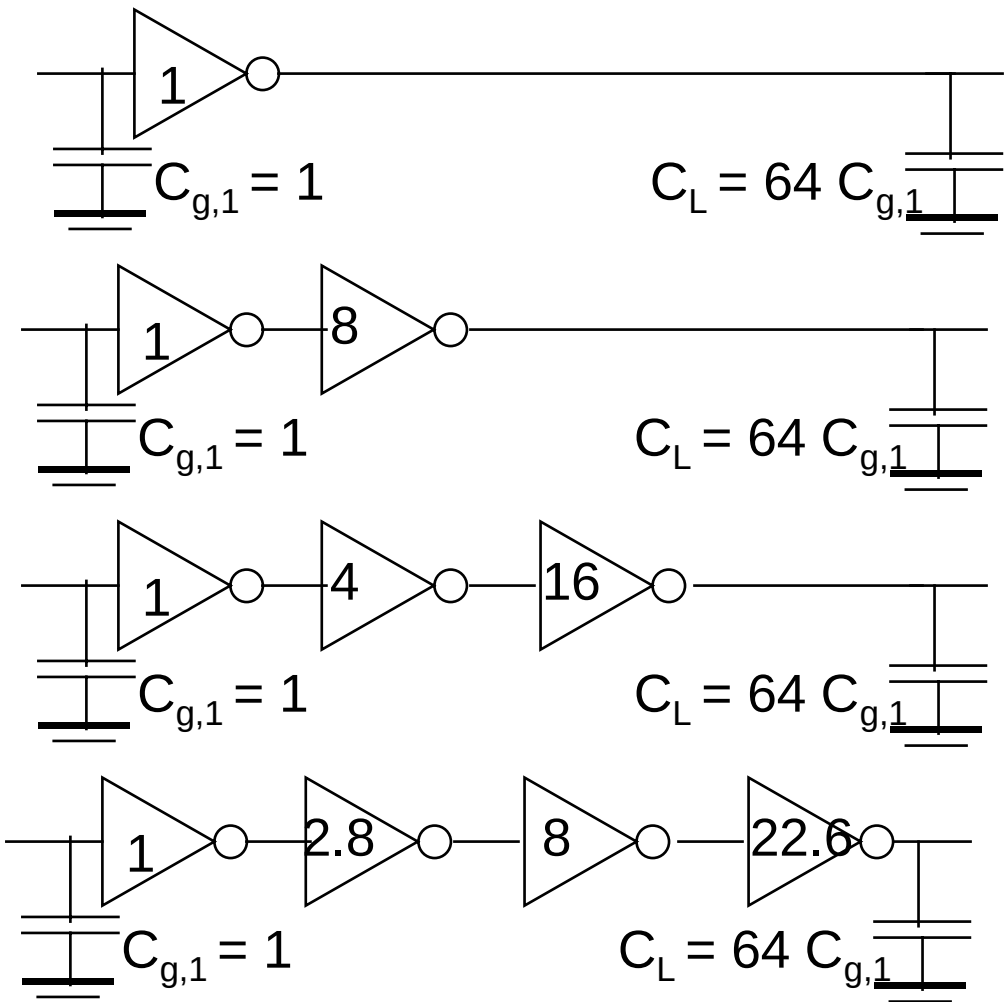
Optimum Effective Fan-Out





- Choosing f larger than optimum has little effect on delay and reduces the number of stages (and area).
  - Common practice to use f = 4 (for γ = 1)
  - But too many stages has a substantial negative impact on delay

Example of Inverter (Buffer) Staging



N	f	t_p
1	64	65
2	8	18
3	4	15
4	2.8	15.3

Impact of Buffer Staging for Large C_L

F (γ = 1)	Unbuffered	Two Stage Chain	Opt. Inverter Chain	N
10	11	8.3	8.3	2
100	101	22	16.5	4
1,000	1001	65	24.8	5
10,000	10,001	202	33.1	7

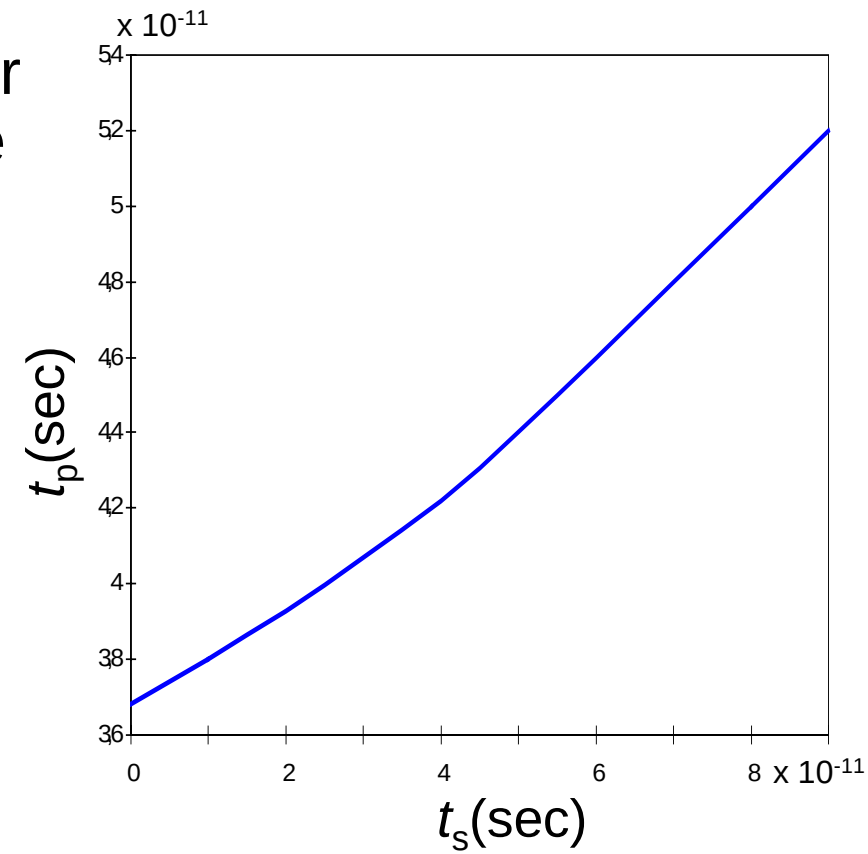
- Impressive speed-ups with optimized cascaded inverter chain for very large capacitive loads.

Input Signal Rise/Fall Time

- In reality, the input signal changes gradually (and both PMOS and NMOS conduct for a brief time). This affects the current available for charging/discharging C_L and impacts propagation delay.

- t_p increases linearly with increasing input slope, t_s, once $t_s > t_p$

- t_s is due to the limited driving capability of the preceding gate



for a minimum-size inverter with a fan-out of a single gate

Design Challenge

- A gate is never designed in isolation: its performance is affected by both the fan-out and the driving strength of the gate(s) feeding its inputs.

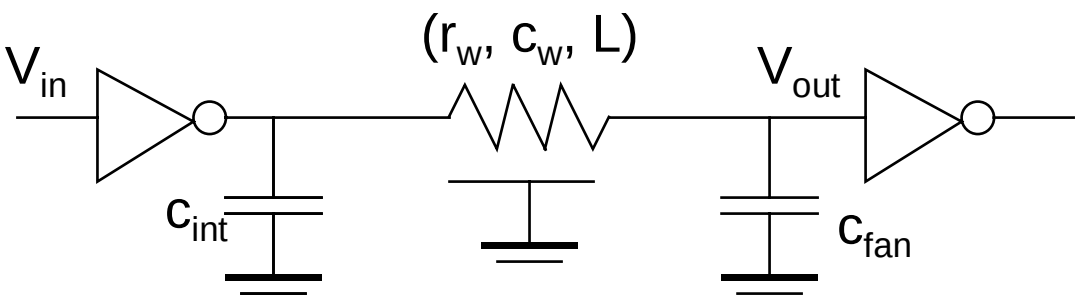
$$t^i_p = t^i_{step} + \eta\, t^{i-1}_{step} \qquad (\eta \approx 0.25)$$

- Keep signal rise times smaller than or equal to the gate propagation delays.
  - good for performance
  - good for power consumption

- Keeping rise and fall times of the signals small and of approximately equal values is one of the major challenges in high-performance designs - slope engineering.

Delay with Long Interconnects

- When gates are farther apart, wire capacitance and resistance can no longer be ignored.



$$t_p = 0.69R_{dr}C_{int} + (0.69R_{dr}+0.38R_w)C_w + 0.69(R_{dr}+R_w)C_{fan}$$

where $R_{dr} = (R_{eqn} + R_{eqp})/2$

$$= 0.69R_{dr}(C_{int}+C_{fan}) + 0.69(R_{dr}c_w+r_wC_{fan})L + 0.38r_wc_wL^2$$

- Wire delay rapidly becomes the dominate factor (due to the quadratic term) in the delay budget for longer wires.