External Conditions which can affect delay

a) Operating Temperature
b) Supply Voltage
c) Process Variation

Drain current is proportional to $T^{-1.5} \Rightarrow$ As temperature is increased, drain current is reduced for a given set of operating conditions, delay increases $\uparrow$

The temperature of the die is what counts, this is expressed as

$$T_j = T_a + \theta_{ja} \times P_d$$

where

$T_a \equiv$ ambient Temperature (˚C)
$\theta_{ja} \equiv$ package thermal impedance (˚C/watt)
$P_d \equiv$ power dissipation

Typical values for θ_{ja} range from 35 to 45 (˚C/watt), depending on chip package

Package Type	Pin Count	θ_{ja} still air	θ_{ja} 300 ft/min.	Units
Plastic J-Leaded Chip Carrier	44	45	35	˚C/W
	68	38	29	˚C/W
	84	37	28	˚C/W
Plastic Quad Flatpack	100	48	40	˚C/W
Very Thin (1.0mm) Quad Flatpack	80	43	35	˚C/W
Ceramic Pin Grid Array	84	33	20	˚C/W
Ceramic Quad Flatpack	84	40	30	˚C/W

Parts usually characterized for different temperature ranges:

Commercial:	0˚ to 70˚ C
Industrial	-40˚ to 85˚ C
Military	-55˚ to 125˚ C

Voltage also affects device speed:
voltage increases ↑, drain current increases, delay decreases ↓

Typically characterize device around a power supply tolerance

	Power Supply Voltage Tolerance
Commercial	± 5%
Industrial	± 10%
Military	± 10%

Process Variations also affect delay - wafer fabrication is a long series of chemical operations, variations in diffusion depth, dopant densities, oxide/diffusion geometry variations can cause transistor switching speeds to vary from wafer batch to wafer batch, wafer to wafer and even on the same wafer.

Transistors typically characterized as "fast", "nominal", and "slow". Need SPICE transistor models for these cases.

However, variations between *n*-speeds and *p*-speeds can be independent so one can obtain "four corners" model

slow *n*MOS fast *p*MOS	fast *n*MOS fast *p*MOS
slow *n*MOS slow *p*MOS	fast *n*MOS slow *p*MOS

When characterizing for high speed, also want to use lowest temperature, highest voltage.

When characterizing for "slow" case, want highest temperature, lowest voltage.

CMOS Digital Systems Checks (Commercial)

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

Power Dissipation

Power Dissipation has three components:

1. Static
2. Dynamic
3. Short Circuit

For traditional CMOS design, static dissipation is limited to the leakage currents in the reversed-biased diodes formed between the substrate (or well) and source/drain regions. But in some DSM CMOS technology subthreshold leakage tends to also contribute significant static dissipation. Subthreshold leakage increases *exponentially* as threshold voltage decreases; i.e., lower V_T (V_{Tn} and $|V_{Tp}|$) CMOS technology has more static power dissipation (due to subthreshold leakage) than higher V_T technology.

Static power dissipation can be extremely small:
1 inverter @ 5V ⇒ 1 to 2 nanowatts static power

Dynamic Power is governed by

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

This is the amount of power dissipated by charging/discharging internal capacitance and load capacitance.

Note the relations:

Higher the switching speed $\Rightarrow P_d \uparrow$

Lower the voltage $\Rightarrow P_d \downarrow\downarrow$!

the Bigger the gates $\Rightarrow P_d \uparrow$

To estimate P_d, need to know the switching frequencies of the internal signals

Typically break this into two parts:

$$P_d = (P_d)\big|_{\text{clock network}} + (P_d)\big|_{\text{all the rest}}$$

The power dissipation in the clock network tends to dominate in most designs. Usually assume the switching frequency of logic signals as some fraction of the clock frequency, can estimate by running some sample simulations and keeping switching statistics on internal nodes to build a probabilistic model of switching activity.

Logic synthesis techniques can be used to do the following:

a. minimize # of gates

or b. maximize speed

and/or c. minimize switching activity

Also, have "short-circuit" power dissipation - proportional to the amount of time when both *p*- and *n*-trees are conducting.

Slow rise/fall times on nodes can make this significant. Usually ignored in most calculations.

Sizing Routing Calculation

The sizing of signal lines to achieve a particular RC delay was previously discussed.

For power conductors, need to worry about

1. Metal migration - too much current in too small a conductor will "blow" the conductor
2. Ground Bounce - large current spikes in V_{DD}/GND leads can occur when simultaneous outputs switch

Two components to ground bounce.

a. IR ← for on-chip conductors, R is resistance of on-chip conductor

b. $L\left(\frac{di}{dt}\right)$ ← L is the on-chip inductance and package inductance in V_{DD}/GND pins. Package inductance dominates. Note that $\frac{di}{dt}$ is affected by slew rates on input/output pins.

Example

What would be the conductor width of power and ground wires to a 50MHz clock buffer that drives 100pF of on-chip load to satisfy the metal-migration consideration (J_{AL} = 0.5mA/µm)? What is the ground bounce with chosen conductor size? The module is 500µm from both the power and ground pads and the supply voltage is 5 volts.

1. $P = CV_{DD}^2 f$
   $= 100 \times 10^{-12} \times 25 \times 50 \times 10^6$
   $= 125\text{mW}$
   $I = P/V = 25\text{mA}$
   Thus the width of the clock wires should be at least 50µm. A good choice would be 100µm.

2. $R = 500/100 \times .05$
   $= 5 \text{ squares} \times .05\ \Omega/\text{sq.}$
   $= 0.25\Omega$
   $IR = 0.25 \times 25 \times 10^{-3} = 6.25\text{mV}$

Typically, *IR* term of ground bounce very <u>small</u> compared to $L\left(\frac{di}{dt}\right)$ term.

Scaling

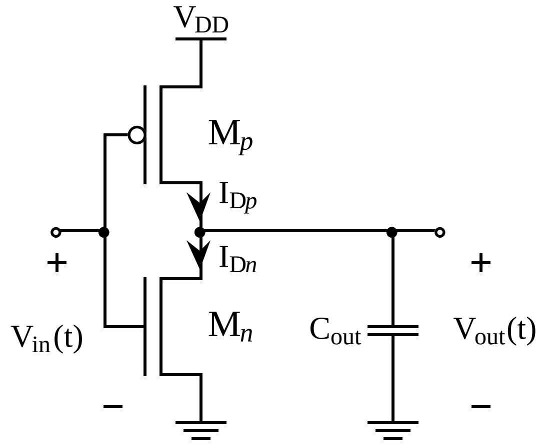
Influence of Scaling on MOS-Device Characteristics			
PARAMETER	SCALING MODEL		
	Constant field	Constant voltage	Lateral
Length (L)	$1/\alpha$	$1/\alpha$	$1/\alpha$
Width (W)	$1/\alpha$	$1/\alpha$	1
Supply voltage (V)	$1/\alpha$	1	1
Gate-oxide thickness (t_{ox})	$1/\alpha$	$1/\alpha$	1
Current ($I = (W/L)(1/t_{ox})V^2$)	$1/\alpha$	α	α
Transconductance (g_m)	1	α	α
Junction depth (X_j)	$1/\alpha$	$1/\alpha$	1
Substrate doping (N_A)	α	α	1
Electric field across gate oxide (E)	1	α	1
Depletion layer thickness (d)	$1/\alpha$	$1/\alpha$	1
Load Capacitance ($C = WL/t_{ox}$)	$1/\alpha$	$1/\alpha$	$1/\alpha$
Gate Delay (VC/I)	$1/\alpha$	$1/\alpha^2$	$1/\alpha^2$
	RESULTANT INFLUENCE		
DC power dissipation (P_s)	$1/\alpha^2$	α	α
Dynamic power dissipation (P_d)	$1/\alpha^2$	α	α
Power-delay product	$1/\alpha^3$	$1/\alpha$	$1/\alpha$
Gate area ($A = WL$)	$1/\alpha^2$	$1/\alpha^2$	$1/\alpha$
Power density (VI/A)	1	α^3	α^2
Current density	α	α^3	α^2

Constant field scaling - all dimensions, including vertical, scaled by α
Constant voltage scaling - constant field scaling but hold V_{DD} constant
Lateral - shrink gate length only

Influence of Scaling on Interconnect Media (Constant Field)	
PARAMETERS	SCALING FACTOR
Line resistance (*r*)	α
Line response (*rc*)	1
Voltage drop	1

Note: Scaling factor of 1 for the line response is bad! Actually the problem is even worse than this! You reduce the voltage to cope with power dissipation problems, you reduce current in gates and the gates do not drive the interconnect as well. Also, overall chip size is not decreasing, just putting more gates in same area so interconnect length is constant for long interconnects.

Transient Analysis

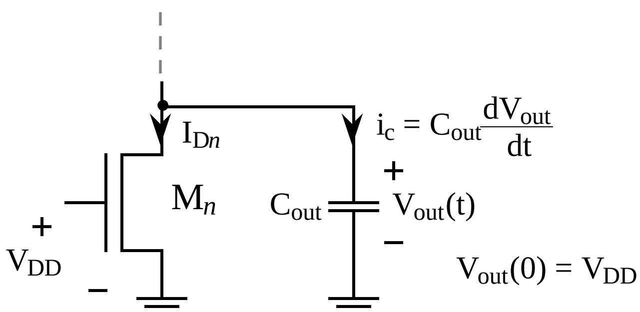


Static Inverter Response

A. Discharge

$V_{in}(t=0^-) = 0 \rightarrow V_{DD}$

$V_{out}(t=0^-) = V_{DD} \rightarrow$ decreases



(i) Initially,

M_n starts out saturated $\Rightarrow$ $$-C_{out}\left(\frac{dV_{out}}{dt}\right) = \frac{\beta_n}{2}\left(V_{DD} - V_{Tn}\right)^2$$

Integrate $\Rightarrow$ $$V_{out}(t) = V_{DD} - \frac{\beta_n}{2C_{out}}\left(V_{DD} - V_{Tn}\right)^2 t$$

This result (previous page) is valid until a time t_o such that

$$V_{out}(t_o) = V_{DD} - V_{Tn}$$
$$= V_{DD} - \frac{ß_n}{2C_{out}}\left(V_{DD} - V_{Tn}\right)^2 t_o$$

So

$$t_o = \frac{2C_{out}V_{Tn}}{ß_n\left(V_{DD} - V_{Tn}\right)^2}$$

And, in terms of t_o,

$$V_{out}(t) = V_{DD} - V_{Tn}\left(\frac{t}{t_o}\right) \qquad \text{[while } M_n \text{ sat.]}$$

(ii) For $t \geq t_o$ $\Rightarrow$ M_n is non-saturated

Here we have $\Rightarrow$ $$-C_{out}\left(\frac{dV_{out}}{dt}\right) = \frac{ß_n}{2}\left[2(V_{DD} - V_{Tn})V_{out} - V_{out}^2\right]$$

I.C. (initial condition) is $V_{out}(t_o) = V_{DD} - V_{Tn}$, so

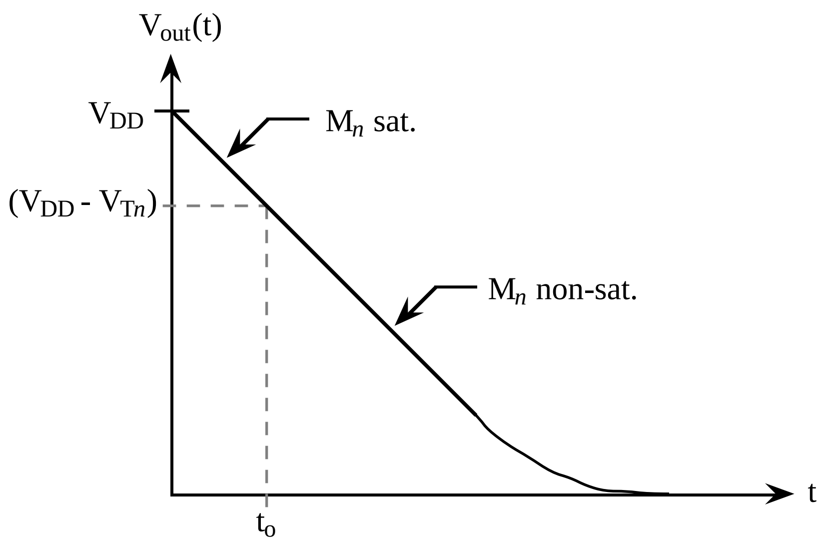
$$V_{out}(t) = (V_{DD} - V_{Tn})\left(\frac{2\exp(-t/\tau_n)\exp(t_o/\tau_n)}{1 + \exp(-t/\tau_n)\exp(t_o/\tau_n)}\right)$$

[note: $\exp(t_o/\tau_n)$ is a time shift function]

$$V_{out}(t) = (V_{DD} - V_{Tn})\left(\frac{2\exp(-(t - t_o)/\tau_n)}{1 + \exp(-(t - t_o)/\tau_n)}\right)$$

where $$\tau_n = \frac{C_{out}}{ß_n\,(V_{DD} - V_{Tn})} = R_nC_{out}$$

Discharge "picture"



If we define t_{HL} as 90% - 10% time (time to discharge from $0.9V_{DD}$ to $0.1V_{DD}$),

$$t_{HL} = \tau_n \left(\frac{2(V_{Tn} - V_o)}{(V_{DD} - V_{Tn})} + ln \left(\frac{2(V_{DD} - V_{Tn})}{V_o} - 1 \right) \right)$$

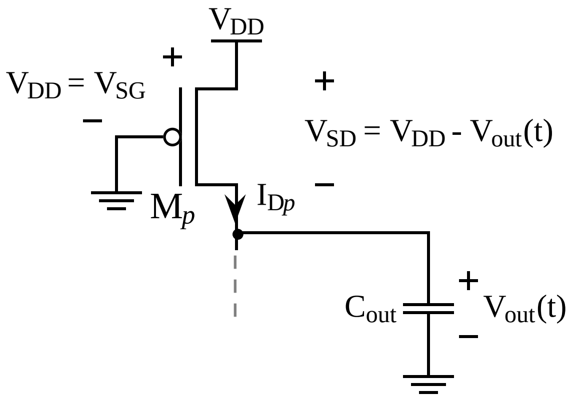
where $V_o = 0.1V_{DD}$.

Note that $t_{HL} \propto \tau_n$.

B. Charge

$V_{in}(t=0^-) = V_{DD} \rightarrow 0$

$V_{out}(t=0^-) = 0 \rightarrow$ increases



(i) Initially, M_p starts out saturated —

$$C_{out}\left(\frac{dV_{out}}{dt}\right) = \frac{ß_p}{2}\left(V_{DD} - |V_{Tp}|\right)^2$$

Integrate:

$$V_{out}(t) = \frac{ß_p}{2C_{out}}\left(V_{DD} - |V_{Tp}|\right)^2 t$$

Valid until

$$V_{out}(t_o) = |V_{Tp}| = \frac{ß_p}{2C_{out}}\left(V_{DD} - |V_{Tp}|\right)^2 t_o$$

$$\Rightarrow t_o = \frac{2C_{out}|V_{Tp}|}{ß_p\left(V_{DD} - |V_{Tp}|\right)^2}$$

So

$$V_{out}(t) = |V_{Tp}|\left(\frac{t}{t_o}\right)$$ [while M_p is sat.]

(ii) For $t \geq$ to $\Rightarrow$ M_p is non-saturated

$$C_{out}\left(\frac{dV_{out}}{dt}\right) = \frac{ß_p}{2}\left(2(V_{DD} - |V_{Tp}|)(V_{DD} - V_{out}) - (V_{DD} - V_{out})^2\right)$$

Integrating:

$$\int \frac{dV_{out}}{2(V_{DD} - |V_{Tp}|)(V_{DD} - V_{out}) - (V_{DD} - V_{out})^2} = \frac{\beta_p}{2C_{out}} \int dt$$

Helpful to define $\upsilon \equiv V_{DD} - V_{out}$

$d\upsilon \equiv -dV_{out}$

Then

$$-\int \frac{d\upsilon}{2(V_{DD} - |V_{Tp}|)\upsilon - \upsilon^2} = \frac{\beta_p}{2C_{out}} \int dt$$

This form is now similar to the discharge case.

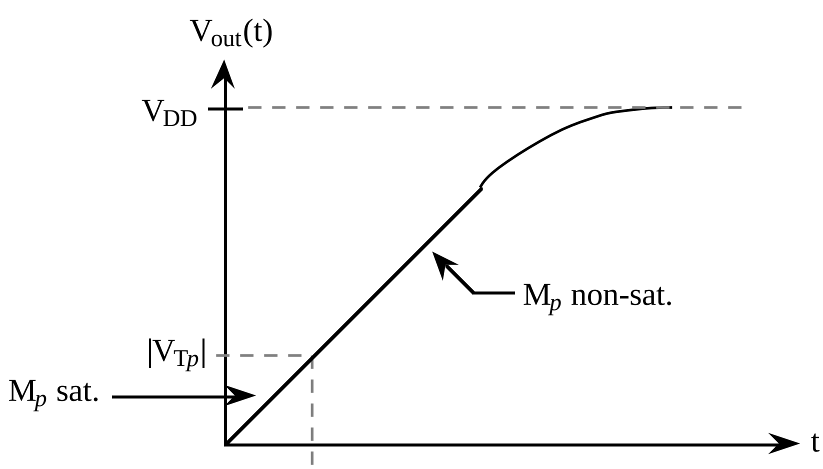
Integrate to get

$$V_{out}(t) = \left(V_{DD} - \left(V_{DD} - |V_{Tp}| \right) \frac{2\exp\left(-(t - t_o)/\tau_p\right)}{1 + \exp\left(-(t - t_o)/\tau_p\right)} \right)$$

where

$$\tau_p = \frac{C_{out}}{\beta_p\left(V_{DD} - |V_{Tp}|\right)} = R_p C_{out}$$

Charging "picture"



If we define t_{LH} as 10% - 90% time (time to charge from $0.1V_{DD}$ to $0.9V_{DD}$),

$$t_{LH} = \tau_p \left(\frac{2(|V_{Tp}| - V_o)}{(V_{DD} - |V_{Tp}|)} + ln \left(\frac{2(V_{DD} - |V_{Tp}|)}{V_o} - 1 \right) \right)$$

where $V_o = 0.1V_{DD}$.

Note that $t_{LH} \, \alpha \, \tau_p$.

Maximum Switching Frequency

A gate's minimum time requirement to undergo a complete switching cycle is $(t_{HL} + t_{LH})$. The maximum switching frequency for a gate is

$$f_{max} = \frac{1}{t_{HL} + t_{LH}} .$$

Propagation Delay

Propagation delay time, t_P, conveniently describes the logic delay through a gate.

$$t_P = \frac{1}{2} (t_{HL} + t_{LH}),$$

where

$$t_{PHL} = \tau_n \left(\frac{2V_{Tn}}{(V_{DD} - V_{Tn})} + ln \left(\frac{4(V_{DD} - V_{Tn})}{V_{DD}} - 1 \right) \right)$$

and

$$t_{PLH} = \tau_p \left(\frac{2|V_{Tp}|}{(V_{DD} - |V_{Tp}|)} + ln \left(\frac{4(V_{DD} - |V_{Tp}|)}{V_{DD}} - 1 \right) \right).$$

Here, t_{PHL} and t_{PLH}, are the propagation delays for a high-to-low and a low-to-high transition, respectively. t_{PHL} is the time required for the output to change from V_{DD} to V_{th} (for the above equations, $V_{th} = (V_{DD}/2)$ is assumed). Likewise, t_{PLH} is the time needed for a gate's output to rise from V_{OL} to V_{th}.

Physical interpretation of $t_P \Rightarrow$ average time for a gate's output to respond to a logic state change at its input .