

## Propagation Delay

Several observations can be made from the analysis:

- PMOS was widened to match resistance of NMOS by **3 - 3.5**.

This was done to provide symmetrical H-to-L and L-to-H propagation delays.

This also **triples** the PMOS gate and diffusion capacitances.

It is possible to speed-up the inverter by *reducing* the width of the PMOS device (at the expense of symmetry and noise margins)!

Widening PMOS reduces  $t_{pLH}$  by increasing the charging current, but it also degrades the  $t_{pHL}$  by causing a larger parasitic capacitance.

This implies that there is an *optimal* ratio that balances the two contradictory effects.

**Propagation Delay**

Consider two identically sized CMOS inverters. The load cap of the first gate is approximated by:

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

Now assume PMOS devices are made  $\beta$  times larger than NMOS.

$$C_{dp1} = \beta C_{dn1} \quad \& \quad C_{gp1} = \beta C_{gn1}$$

$$C_L = (1 + \beta)(C_{dn1} + C_{gn2}) + C_W$$

Returning to:

$$t_p = \frac{t_{pHL} + t_{pLH}}{2} = 0.69 C_L \left( R_{eqn} + \frac{R_{eqp}}{\beta} \right)$$

$$t_p = \frac{0.69}{2} \left( (1 + \beta)(C_{dn1} + C_{gn2}) + C_W \right) \left( R_{eqn} + \frac{R_{eqp}}{\beta} \right)$$

$$t_p = 0.345 \left( (1 + \beta)(C_{dn1} + C_{gn2}) + C_W \right) R_{eqn} \left( 1 + \frac{r}{\beta} \right)$$

**Propagation Delay**

$r$  is equal to the resistance ratio of identically sized PMOS and NMOS transistors:  $R_{eqp}/R_{eqn}$ .

The optimal value of  $b$  can be found by setting

$$\frac{\partial t_p}{\partial \beta} = 0 \quad \longrightarrow \quad \beta_{opt} = \sqrt{r \left( 1 + \frac{C_w}{C_{dn1} + C_{gn1}} \right)}$$

When wiring capacitance is negligible,  $\beta_{opt}$  equals the  $\text{sqrt}(r)$ , vs.  $r$  normally used in the non-cascaded case.

If wiring cap dominates, larger values of  $b$  should be used.

This analysis indicates that smaller device sizes (and smaller area) yield a **faster** design at the expense of symmetry and noise margins.

Example in text gives  $\beta$  of 2.4 (=31 k $\Omega$ /13 k $\Omega$ ) for symmetrical response.

$\beta_{opt}$  is then 1.6 -- SPICE sims gives optimal value of  $\beta = 1.9$ .

### Sizing Inverters for Performance

Assume a symmetrical inverter (rise and fall times of inverter are identical).

Load capacitance can be divided into *intrinsic* or **self-loading** and *extrinsic* components:

$$C_L = C_{int} + C_{ext}$$

Assuming  $R_{eq}$  stands for the equivalent resistance of the gate, then propagation delay is:

$$\begin{aligned} t_p &= 0.69R_{eq}(C_{int} + C_{ext}) \\ &= 0.69R_{eq}C_{int}\left(1 + \frac{C_{ext}}{C_{int}}\right) \\ &= t_{p0}\left(1 + \frac{C_{ext}}{C_{int}}\right) \end{aligned} \quad \text{with} \quad \begin{array}{l} \text{Intrinsic or unloaded delay} \\ \downarrow \\ t_{p0} = 0.69R_{eq}C_{int} \end{array}$$

So how does transistor sizing impact the performance of the gate?

### *Sizing Inverters for Performance*

$C_{int}$  consists of the diffusion and Miller caps, both of which are proportional to the width of the transistors.

Let's use a minimum sized inverter as a *reference* gate, then:

$$C_{int} = SC_{iref} \quad \& \quad R_{eq} = \frac{R_{ref}}{S}$$

where  $S$  is the *sizing* factor.

Re-writing previous expression:

$$\begin{aligned} t_p &= 0.69 \left( \frac{R_{ref}}{S} \right) (SC_{iref}) \left( 1 + \frac{C_{ext}}{SC_{iref}} \right) \\ &= 0.69 R_{ref} C_{iref} \left( 1 + \frac{C_{ext}}{SC_{iref}} \right) \end{aligned}$$

## *Sizing Inverters for Performance*

### Conclusions

- Intrinsic delay of the inverter  $t_{p0}$  is *independent* of the sizing of the gate (determined by technology and layout only).

When there is no load, the increase in drive of the gate is **totally offset** by increased cap.

- Making  $S$  infinitely large yields the **max** performance, **eliminates** the impact of any external load and reduces the delay to the intrinsic one.

Bear in mind that any size greater than  $(C_{\text{ext}}/C_{\text{int}})$  produces similar results while increasing the silicon area -- no win beyond this size.

Bear in mind that although sizing up an inverter reduces its delay, it also *increases* its input capacitance.

So the more relevant problem is determining the optimum size of a gate when embedded in a *real environment*.

### *Sizing Inverters for Performance*

Consider a chain of inverters as the first case.

To determine input loading effect, we need to determine the relationship between the *input gate capacitance*,  $C_g$  and the *intrinsic output capacitance*.

Both are proportional to gate sizing, so the following is true:

$$C_{int} = \gamma C_g$$

The gamma factor  $\gamma$  is only a function of technology and is close to 1 for most processes.

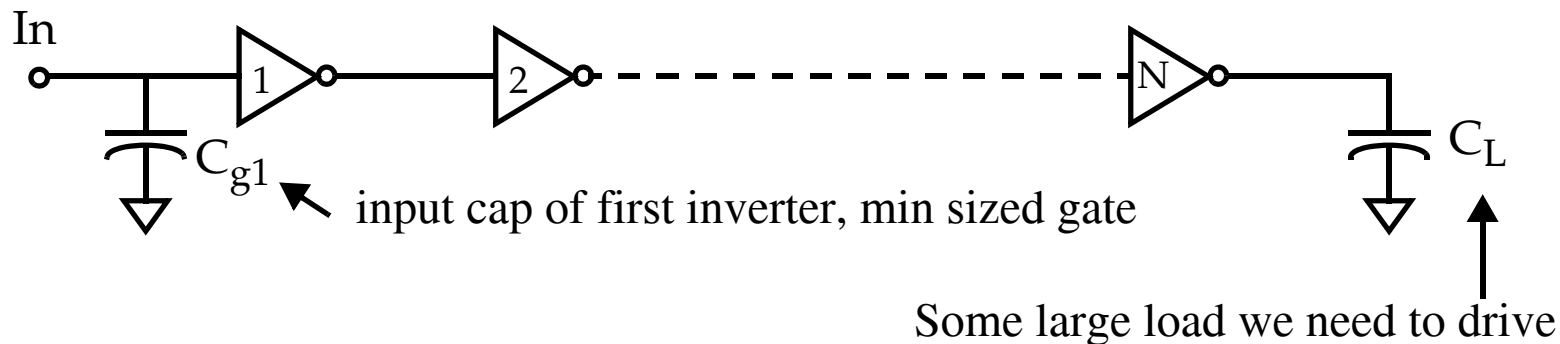
Substituting:

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

This shows the delay of an inverter is **only** a function of the ratio between its external load cap and its input cap, and is called **effective fan-out  $f$** .

### Sizing a Chain of Inverters

Goal is to minimize delay through the following inverter chain:



Delay for j-th inverter stage (ignoring wire cap):

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

The total delay of the chain is then:

$$t_{p,j} = \sum_{j=1}^N t_{p,j} = t_{p0} \sum_{j=1}^N \left( 1 + \frac{C_{g,j+1}}{\gamma C_{g,j}} \right) \quad \text{with } C_{g,N+1} = C_L$$

And we need to solve for N-1 unknowns  $C_{g,2}, C_{g,3}, C_{g,N}$ .



### *Sizing a Chain of Inverters*

Solution giving the **optimal size of each inverter** (that minimizes delay) is the geometric mean of each of the inverter's neighbors:

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

So each inverter is sized up by the same factor  $f$  (and has the same delay).

Given  $C_{g,1}$  and  $C_L$ , the sizing factor is given as:

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

where  $F$  represents the **overall effective fan-out** of the circuit and equals  $C_L / C_{g,1}$ .

The minimum delay through the chain is:

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

First component is *intrinsic delay* of the stages while second is *effective fan-out* of each stage.

### *Sizing a Chain of Inverters*

The relationship between  $t_p$  and  $F$  is a strong function of the number of stages.

The important question now is how to choose the **number of stages** so that the delay is minimized for a given value of  $F$  ( $C_L/C_{g,1}$ ).

If too many, *intrinsic delay* dominates, if too few, *effective fan-out* dominates.

Differentiating and setting to zero yields:

$$\gamma + N\sqrt{F} - \frac{N\sqrt{F} \ln(F)}{N} = 0 \quad \text{or}$$

$$f = e^{(1+\gamma/f)}$$

Under the condition that  $\gamma$  is 0 (self-loading is ignored, load cap only consists of the fan-out), the optimal number is:

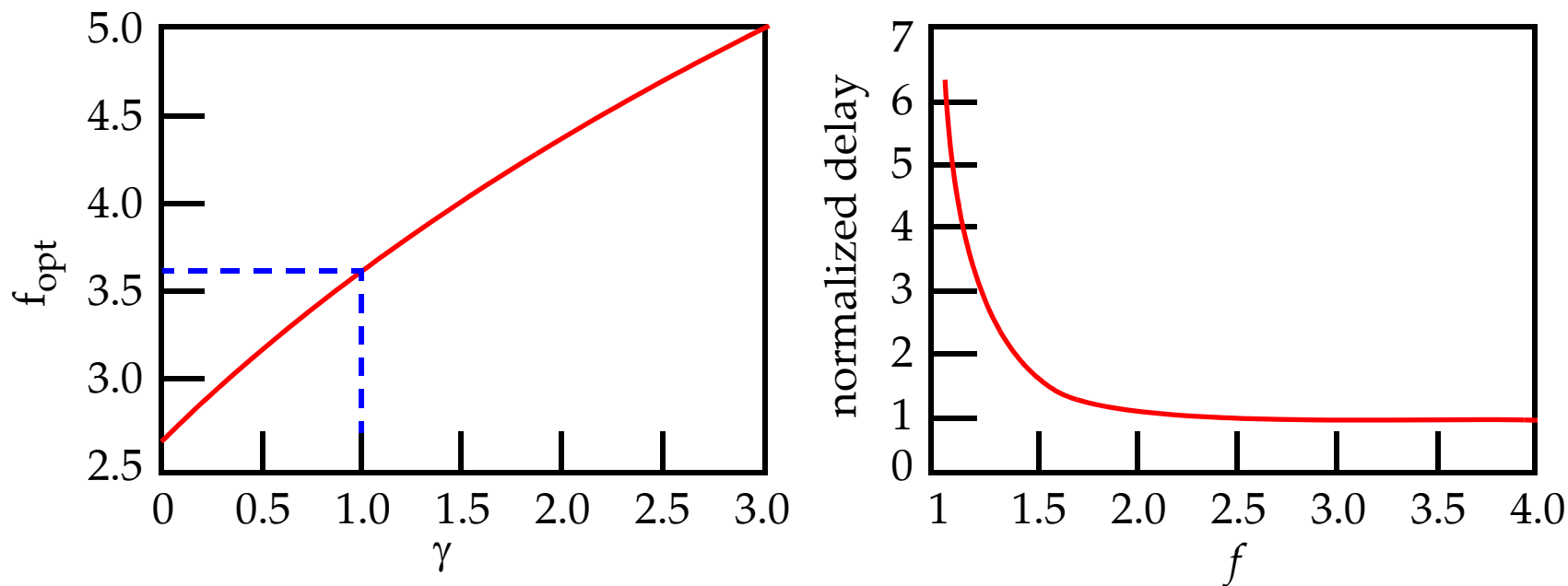
$$N = \ln(F) \quad \text{effective fan-out is set to } f = e = 2.71828$$

## Sizing a Chain of Inverters

This indicates that the optimal buffer design scales consecutive stages in an exponential fashion (*exponential horn*).

The solution when *self-loading* is included can only be computed numerically.

For a typical case with  $\gamma = 1$ , the optimum tapering factor is close to 3.6.



Right plot shows *normalized delay* ( $t_p/t_{popt}$ ) as a function of fan-out  $f$  for  $\gamma = 1$ .

### *Sizing a Chain of Inverters*

Here it is clear that choosing values for fan-out that are **higher** than the optimum does NOT effect the delay very much (and helps reduce area).

It is common to select an **optimum fan-out of 4** (FO4).

Note that the use of too few stages ( $f < f_{opt}$ ) has a significant impact on performance and should be avoided.

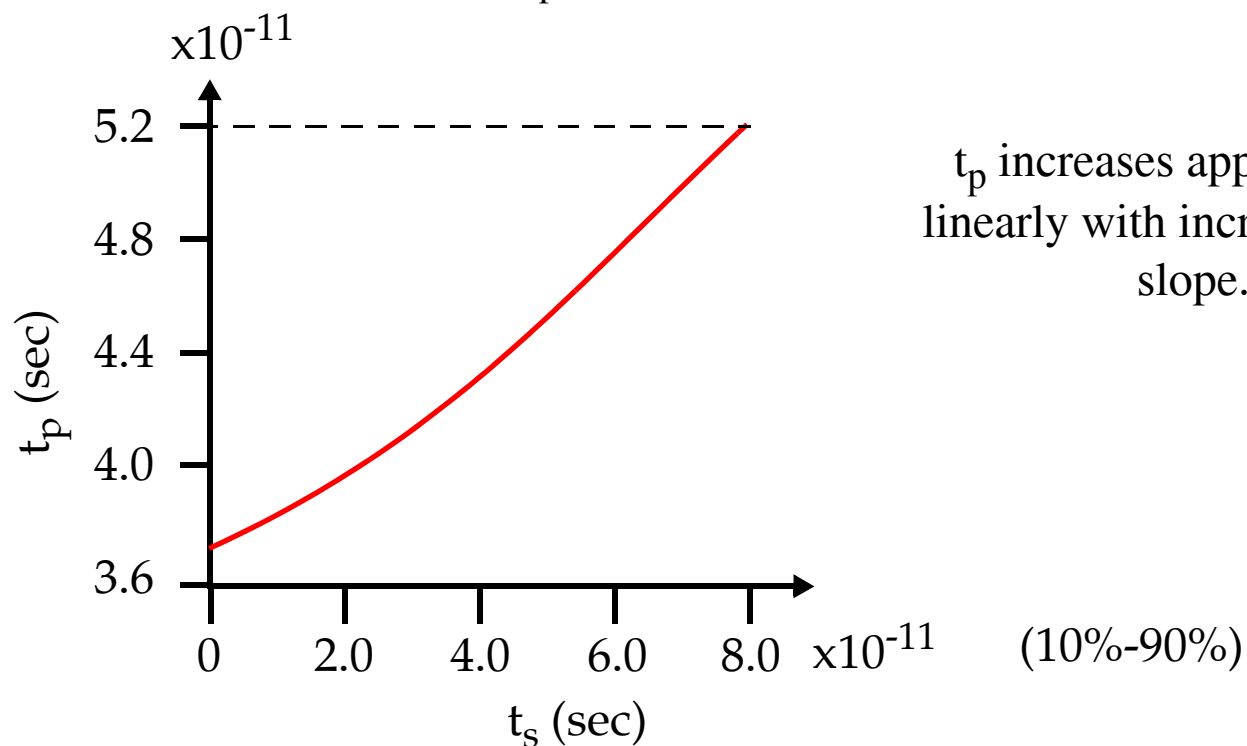
### *Rise-Fall Time of Input Signal*

It is not realistic to assume that input signal changes abruptly and only one device is on.

Reality is that both are on for some portion of time and the total charging/discharging current is directed onto/off the load caps.

### Rise-Fall Time of Input Signal

Propagation delay of a minimum sized inverter as a function of input signal slope (fan-out is a single gate), for  $t_s > t_p$ .



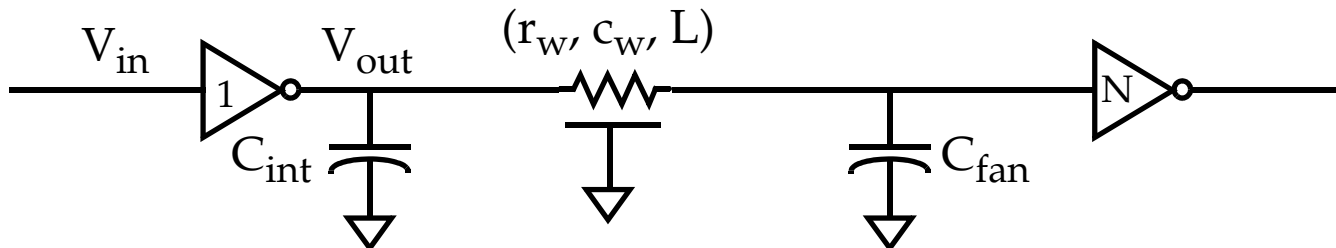
Text gives a more thorough analysis.

Key design challenge is to keep the *signal rise times*  $\leq$  the *gate propagation delay*, for speed and power consumption.

## Wire Delay

We've ignored the wire delay so far, even though its influence can dominate the transient response.

Consider the following circuit:



Here, inverter drives a single fan-out through a wire of length  $L$ .

Let the driver be represented by a single resistance  $R_{dr}$  (average of  $R_{eqn}$  and  $R_{eqp}$ ), and  $C_{int}$  and  $C_{fan}$  are the intrinsic cap of the driver and input cap of the fan-out gate.

Elmore delay expression yields the propagation delay of the circuit as:

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

### *Wire Delay*

Rearranging yields:

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

The 0.38 factor accounts for the fact that the wire represents a distributed delay.

$C_w$  and  $R_w$  stand for the total capacitance and resistance of the wire.

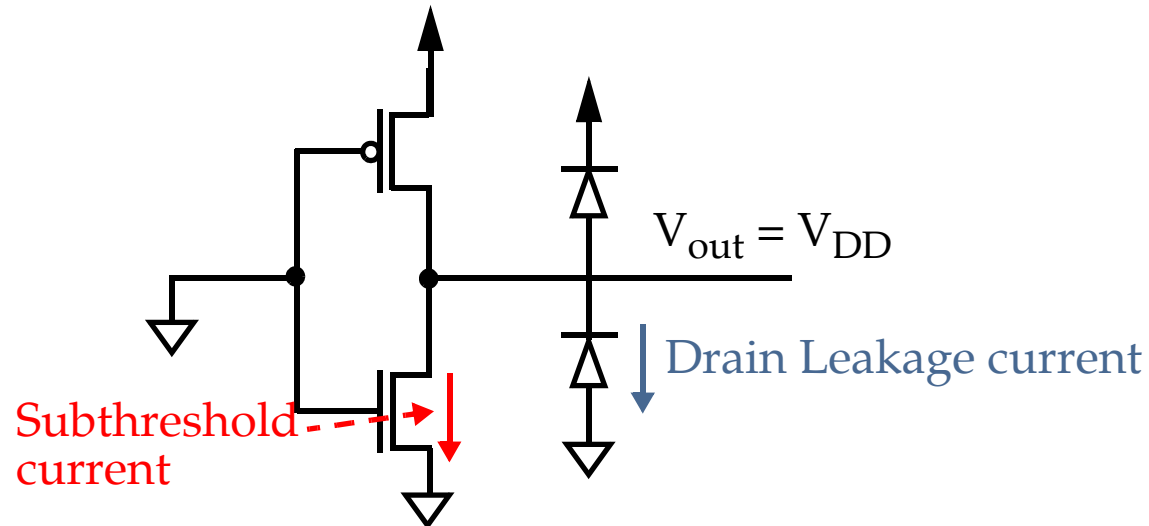
Here, the delay expression contains a component that is linear with the wire length, as well as a quadratic one.

The latter obviously becomes the dominant factor in the delay of longer wires.

### Power Consumption

The almost ideal VTC of the CMOS inverter is **not** the main reason that high-complexity designs are implemented in static CMOS.

Rather, its the almost **zero** power consumption in steady-state mode.



The reversed-bias diode current is, in general, very small.

Typical values are 0.1 to 0.5nA at room temperature.

For a device at 5V with 1 million devices, power consumption is 0.5mW.

A more serious source is the subthreshold current.

The closer  $V_T$  is to zero, the larger the leakage with  $V_{GS} = 0V$ .

This establishes a firm lower bound on  $V_T$ , which is  $> 0.5V$  today.



## Power Consumption

For both sources of leakage, the resulting static power dissipation is given by:

$$P_{static} = I_{leakage} V_{DD}$$

The junction leakage currents are caused by *thermally generated carriers*.

Their value **increases exponentially** with increasing junction temperature.

For example, 85 degrees C (a common junction temperature) results in an increase by a factor of **60** over room temperature.

Dynamic power is much larger than static power and can be broken into 2 parts.

- *Load capacitance*,  $C_L$ , power.
- Power consumed via *direct path currents* (crow-bar currents).

## Power Consumption

$C_L$  power (we derived this previously):

Charging  $C_L$  to  $V_{DD}$  draws  $C_L * V_{DD}^2$  energy from the power supply.

Half of this energy is stored on the cap ( $C_L * V_{DD}^2 / 2$ ) and later dissipated through the NMOS device.

So, an energy =  $C_L * V_{DD}^2$  is consumed for every L->H and H->L transition.

Therefore, for a clock frequency of  $f$ ,

$$P_{\text{dyn}} = C_{\text{eff}} V_{DD}^2 f \quad \text{with } C_{\text{eff}} = \alpha C_L$$

Technology advances decrease  $t_p$  and increase  $f$  and  $C_L$  (higher integration).

For example, at 30fF/gate at 100MHz and  $V_{DD} = 5V$ , 75 $\mu$ W is dissipated per gate.

With 200K gates and  $\alpha = 20\%$ , 3W are dissipated.

1W is consumed with 100 output pins at 20pF/pin and  $f = 20$ MHz.

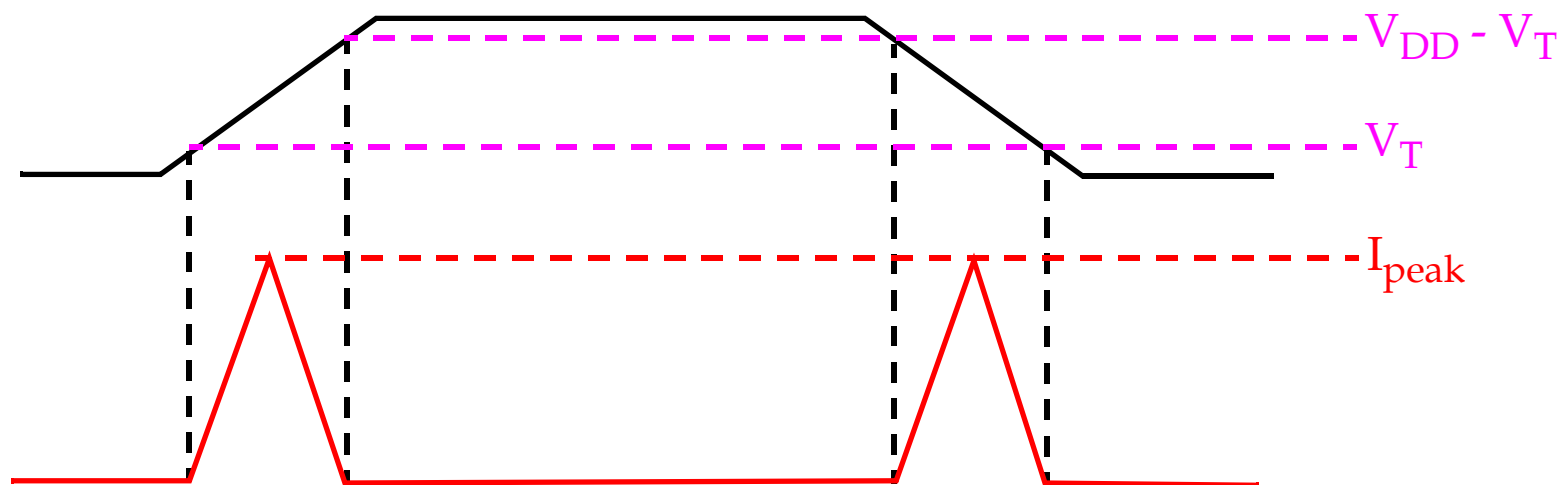
One of the driving forces for lower supply voltages (**quadratic** effect).

For example, 5V -> 3V drops 4W to 1.44W (assuming the same  $f$ ).

## Power Consumption

Direct-path currents.

Zero rise/fall times is not a realistic assumption.



Using triangles and  $V_{DD} \gg |V_T|$ , the power consumed is

$$P_{dp} = \left( V_{DD} \frac{I_{peak} t_r}{2} + V_{DD} \frac{I_{peak} t_f}{2} \right) f = \frac{t_r + t_f}{2} V_{DD} I_{peak} f$$

Avoid large values for  $t_f$  and  $t_r$  to minimize.

Direct-path power is typically only about **20%** of the dynamic power.

## Power Consumption

Total power is then:

$$P_{tot} = P_{dyn} + P_{dp} + P_{static} = C_L V_{DD}^2 f + V_{DD} I_{peak} \left( \frac{t_r + t_f}{2} \right) f + V_{DD} I_{leak}$$

The **Power-Delay** product was also defined previously.

It is the energy consumed by the gate per switching event.

We've defined a switching event to consist of a 0 -> 1 and a 1 -> 0 event.

This results in a **PDP** of

$$PDP = C_L V_{DD}^2$$

Under the condition that the static and direct-path currents are ignored.