ELE617E
Lectures

Prof. Dr. Müştak E. Yalçın

Istanbul Technical University

mustak.yalcin@itu.edu.tr

Pipelining and Parallel Processing for Low Power

- Two main advantages of using pipelining and parallel processing: – Higher speed and Lower power consumption
- When sample speed does not need to be increased, these techniques can be used for lowering the power consumption

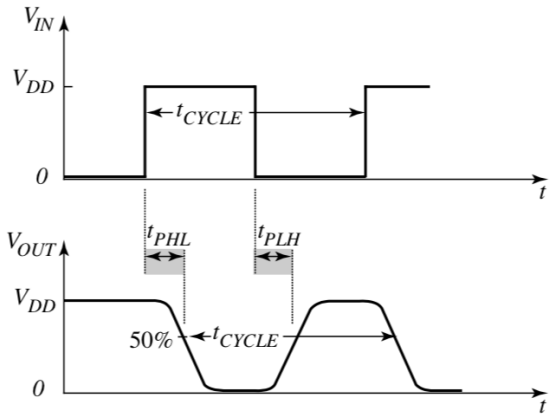
2. CMOS inverter: Propagation delay

Inverter propagation delay: time delay between input and output signals; figure of merit of logic speed.

Typical propagation delays: < 100 ps.

Complex logic system has 10-50 propagation delays per clock cycle.

Estimation of t_p: use square-wave at input

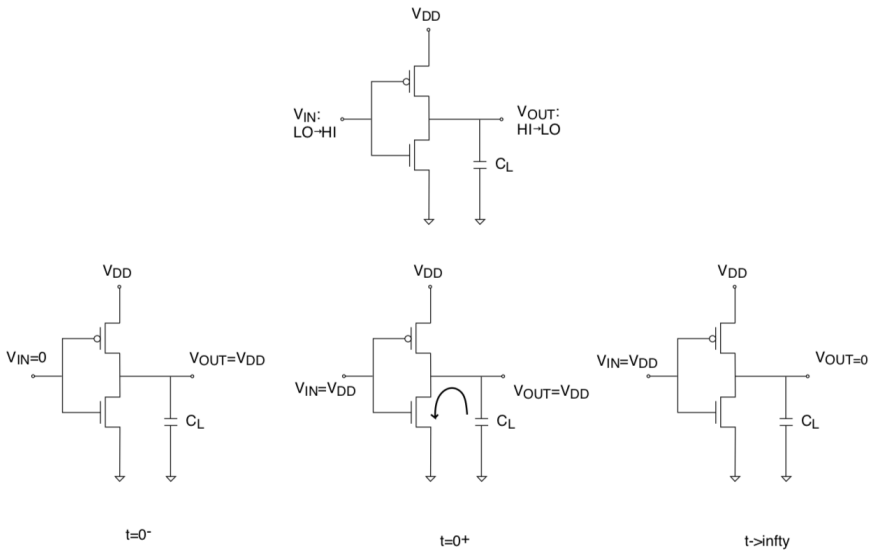


Average propagation delay:

$$t_p = \frac{1}{2}(t_{PHL} + t_{PLH})$$

http://web.mit.edu/6.012/www/SP07-L13.pdf

CMOS inverter: Propagation delay high-to-low



During early phases of discharge, NMOS is saturated and PMOS is cut-off.

Time to discharge ***half*** of charge stored in C_L:.

$$t_{pHL} \approx \frac{\frac{1}{2}\text{charge on } C_L \ @t = 0^-}{\text{NMOS discharge current}}$$

CMOS inverter:
Propagation delay high-to-low (contd.)

Charge in C_L at t=0$^-$:

$$\mathbf{Q_L(t=0^-)=C_L V_{DD}}$$

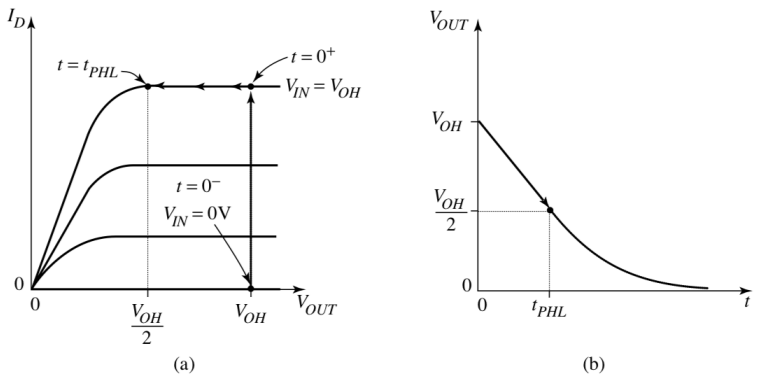
Discharge Current (NMOS in saturation):

$$\mathbf{I_{Dn}=\frac{W_n}{2L_n}\mu_n C_{ox}(V_{DD}-V_{Tn})^2}$$

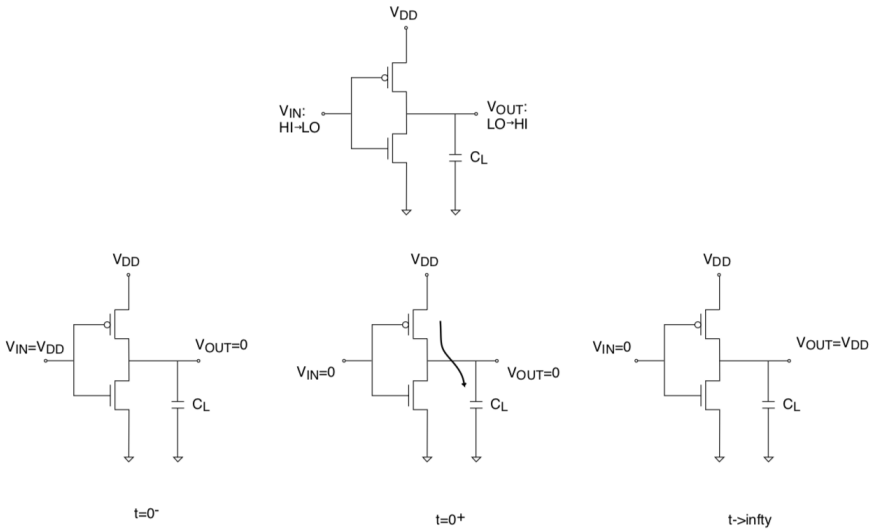
Then:

$$\mathbf{t_{PHL}\approx\frac{C_L V_{DD}}{\frac{W_n}{L_n}\mu_n C_{ox}(V_{DD}-V_{Tn})^2}}$$

Graphical Interpretation



CMOS inverter:
Propagation delay low-to-high



During early phases of discharge, PMOS is saturated and NMOS is cut-off.

Time to charge to ***half*** of final charge on C_L:.

$$\mathbf{t_{PLH}\approx\frac{\frac{1}{2}\text{charge on } C_L \ @t=\infty}{\text{PMOS charge current}}}$$

CMOS inverter:
Propagation delay high-to-low (contd.)

Charge in C_L at $t=\infty$:

$$Q_L(t=\infty) = C_L V_{DD}$$

Charge Current (PMOS in saturation):

$$-I_{Dp} = \frac{W_p}{2L_p}\mu_p C_{ox}\left(V_{DD} + V_{Tp}\right)^2$$

Then:

$$t_{PLH} \approx \frac{C_L V_{DD}}{\frac{W_p}{L_p}\mu_p C_{ox}\left(V_{DD} + V_{Tp}\right)^2}$$

Key dependencies of propagation delay:

- $V_{DD} \uparrow \Rightarrow t_p \downarrow$
  - Reason: $V_{DD} \uparrow \Rightarrow Q(C_L) \uparrow$, but I_D goes as square$\uparrow$
  - Trade-off: $V_{DD} \uparrow \Rightarrow$ more power consumed.
- $L \downarrow \Rightarrow t_p \downarrow$
  - Reason: $L \downarrow \Rightarrow I_D \uparrow$
  - Trade-off: manufacturing cost!

Power Dissipation

• Energy from power supply needed to charge up the capacitor:

$$E_{charge} = \int V_{DD} i(t) dt = V_{DD} Q = V_{DD}^2 C_L$$

• Energy stored in capacitor:

$$E_{store} = 1/2 C_L V_{DD}^2$$

• Energy lost in p-channel MOSFET during charging:

$$E_{diss} = E_{charge} - E_{store} = 1/2 C_L V_{DD}^2$$

•During discharge the n-channel MOSFET dissipates an identical amount of energy.

•If the charge/discharge cycle is repeated f times/second, where f is the clock frequency, the dynamic power dissipation is:

$$P = 2E_{diss} * f = C_L V_{DD}^2 f$$

In practice many gates do not change state every clock cycle which lowers the power dissipation.

Notation

- The propagation delay and Power of the original filter are

$$t = \frac{C_L V_o}{k(V_o - V_t)^2}$$

and

$$P = C_{\text{total}} V_o^2 f = C_L V_o^2 \frac{1}{T_{clk}}$$

PS: take t_{phl} (small one).
C_L : the cap. to be charged and discharged in a single clock cycle. C_{total} : the total cap. of the circuit. V_o : supply voltage f: clock frequency. $f = \frac{1}{T_{\text{clk}}}$ and T_{clk}: clock period.

- We will consider M-level pipeline and L-parallel. Their propagation delay and Power are:
t_{pip}, P_{pip} and t_{par}, P_{par}, respectively.

Pipelining for Low Power

- Consider an M-level pipeline system, where the critical path is reduced to $\frac{1}{M}$, then C_L is reduced to $\frac{C_L}{M}$ for a single clock cycle.
- In the same time that C_L was charge/discharge, now only a fraction of it should be charge/discharge
- Then, the supply voltage can be reduced by β, where $0 < \beta < 1$
- The power consumption of the pipeline filter will be

$$P_{pip} = C_{\text{total}} \beta^2 V_o^2 f = \beta^2 P$$

- How can the value of β be determined ?

Pipelining for Low Power

- Consider an M-level pipeline system, where the critical path is reduced to $\frac{1}{M}$, then C_L is reduced to $\frac{C_L}{M}$ for a single clock cycle.
- In the same time that C_L was charge/discharge, now only a fraction of it should be charge/discharge
- Then, the supply voltage can be reduced by β, where $0 < \beta < 1$
- The power consumption of the pipeline filter will be

$$P_{pip} = C_{\text{total}} \beta^2 V_o^2 f = \beta^2 P$$

- How can the value of β be determined ? by examining the propagation delay.

Pipelining for Low Power

- The propagation delay of the original filter is

$$t = \frac{C_L V_o}{k(V_o - V_t)^2}$$

- While the propagation delay of the pipeline filter is

$$t_{pip} = \frac{\frac{C_L}{M}\beta V_o}{k(\beta V_o - V_t)^2}$$

- The same clock speed is maintained for both filters, therefore the following equation is maintained

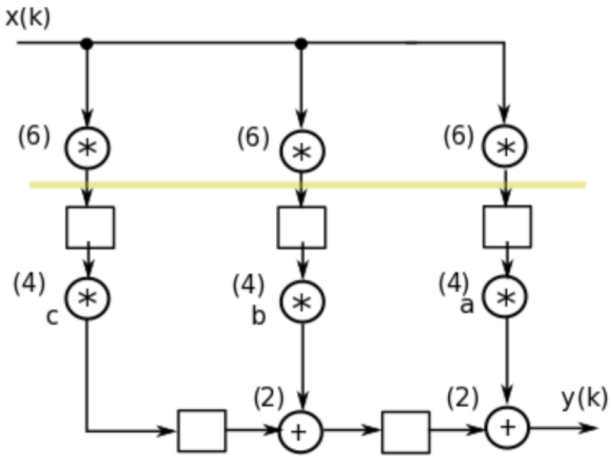
$$M(\beta V_o - V_t)^2 = \beta(V_o - V_t)^2$$

- Then β is obtained, the reduction of power consumption can be computed using

$$P_{pip} = C_{\text{total}}\beta^2 V_o^2 f = \beta^2 P$$

Study : Example 3.4.1

Pipelining for Low Power



$t_{seq} = T_M + T_A = 10\,u.t$ ⊕ → C_A ⊗ → $5C_A$

$t_{pip} = T_M' + T_A = 6u.t$ ⊕ → C_A m_1 → $3C_A$ m_2 → $2C_A$

$C_L = C_M + C_A = 6C_A$

$C_L = C_{M1} = C_{M2} + C_A = 3C_A$

NOTE $M = 2$

$V_0 = 5V$

$V_t = 0.6V$

$M \cdot (\beta V_0 - V_t)^2 = \beta (V_0 - V_t)^2$

$50\beta^2 - 31.6\beta + 0,72$ $\beta = 0.6$ $\beta = 2.02$

$V_0 = \beta V_0 = 3.01V$ $\beta = 36.4\%$!

Parallel Processing for Low Power

- In an L-parallel system, the charging capacitance does not change, but the total capacitance is increased by L times.
- In order to maintain the same data rate, the clock period must be increased to LT
- Then, there is more time to charge the same capacitance.
- Therefore, the supply voltage can be reduced to βV_o
- The propagation delay of the original filter is

$$t = \frac{C_L V_o}{k(V_o - V_t)^2}$$

- The propagation delay of the parallel filter is

$$t_{par} = \frac{C_L \beta V_o}{k(\beta V_o - V_t)^2}$$

(one of L)

Parallel Processing for Low Power

- The same clock speed ($t_{par} = Lt$) is maintained for both filters, therefore the following equation is maintained

$$L(\beta V_o - V_t)^2 = \beta(V_o - V_t)^2$$

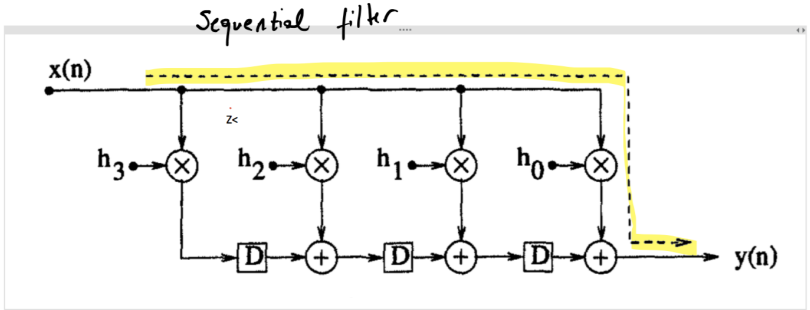
- Then β is obtained, the reduction of power consumption can be computed using

$$P_{par} = LC_L\beta^2 V_o^2 \frac{f}{L} = \beta^2 C_L V_o^2 f = \beta^2 P$$

PS: $C_{\text{total}} = LC_L$

Please read textbook for Example 3.4.2

Parallel Processing for Low Power

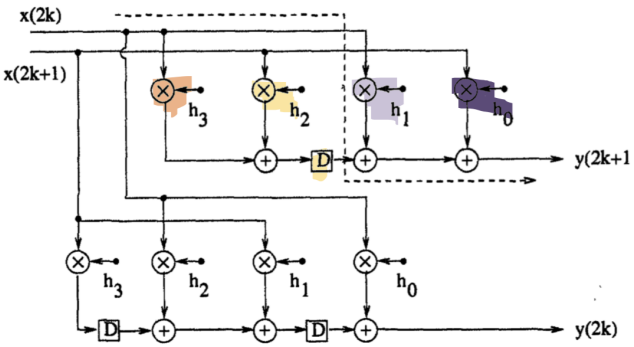


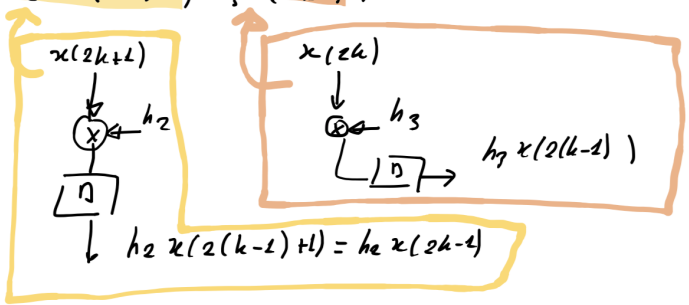
$y(n) = h_0 x(n) + h_1 x(n-1) + h_2 x(n-2) + h_3 x(n-3)$

$T_{critical} = T_M + T_A$

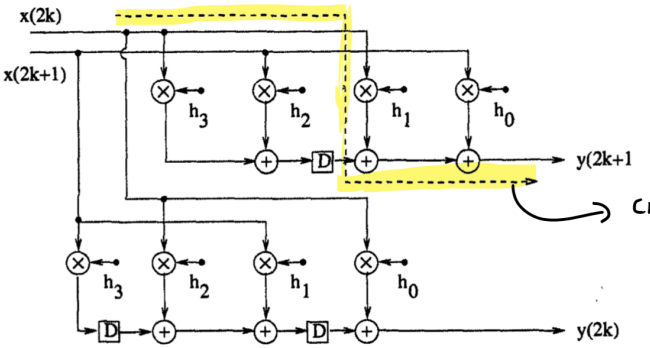
$y(2k) = h_0 x(2k) + h_1 x(2k-1) + h_2 x(2k-2) + h_3 x(2k-3) = h_0 x(2k) + h_1 x(2(k-1)+1) + h_2 x(2(k-1)) + h_3 x(2(k-1)-1)$

$y(2k+1) = h_0 x(2k+1) + h_1 (2k) + h_2 x(2k-1) + h_3 x(2k-2) = h_0 x(2k+1) + h_1 x(2k) + h_2 x(2(k-1)+1) + h_3 x(2(k-1))$





Parallel Processing for Low Power



sequential

$T_{Critical} = T_A + T_M$

$T_{clk} \geq T_A + T_M = 1+8 = 9$

$C_{total} = C_A + C_M = 9C$

critical path

parallel

$T_{Critical} = 2T_A + T_M$

$T_{clk} \geq 2T + T_M = 2+8 = 10$

$C_{total} = 2C_A + C_M = 10C$

Supply Voltage for parallel

$V_0 \longrightarrow \beta \cdot V_0$

$$t_{seq} = \frac{Q}{I} = \frac{9C \cdot V_0}{k (V_0 - V_{th})^2}$$

$$t_{par} = \frac{10C \cdot \beta V_0}{k (\beta V_0 - V_{th})^2}$$

SEQ

$T_S \geq T_{clk}$ $\quad T_{clk} \geq 9$

PAR

$\hat{T}_S \geq \frac{\hat{T}_{clk}}{2}$ $\quad \hat{T}_{clk} \geq 10$

to maintain $\quad T_{clk} = \hat{T}_{clk}$

$\hat{T}_{clk} = 2T_{clk}$

$$2 \cdot \frac{9CV_0}{k(V_0 - V_{th})^2} = \frac{10C\beta V_0}{k(\beta V_0 - V_{th})^2}$$

$$9(\beta V_0 - V_{th})^2 = 5\beta (V_0 - V_{th})^2$$

$$\beta = 0{,}65$$

$P^2 \rightarrow 43\%$

to have same data rate, clock period is increased! then there is more time to charge the cap. ! So there is more time to charge cap. and the supply voltage can be reduce!

Combining Parallel and Pipeling

- Pipeling reduces the capacitance to be charged/discharge in 1 clock period.
- Parallel processing increases the clock period for charging/discharging the original capacitance.
- The propagation delay of the original filter is

$$t = \frac{C_L V_o}{k(V_o - V_t)^2}$$

- the propagation delay of the parallel-pipelined filter is

$$t_{pip} = \frac{\frac{C_L}{M} \beta V_o}{k(\beta V_o - V_t)^2}$$

Combining Parallel and Pipeling

- The same clock speed ($t_{par} = Lt$) is maintained for both filters,

$$t_{par} = Lt = \frac{\frac{C_L}{M}\beta V_o}{k(\beta V_o - V_t)^2}$$

- therefore the following equation is maintained

$$ML(\beta V_o - V_t)^2 = \beta(V_o - V_t)^2$$