MODULE 6

POWER MODELS OF CMOS GATES

Over the past several years, we can see more complex devices around us like cellphones, computers etc. As the functionality and the performance of these devices increase, so does the power dissipation. Hence, it is increasingly important for us today to understand the sources of Power Dissipation in CMOS logic and also understand what are the mechanisms for controlling some of these aspects of Power Dissipation and also minimize the Power Dissipation for the same quality of performance.

Reducing the Power Dissipation eventually leads to say in the case of mobile devices, the battery lasting longer and the device stays cooler (higher Power Dissipation means higher temperature) and in the case of laptops and desktop computers, reducing the Power Dissipation will reduce the operating costs of these devices. Hence, there is a huge push from the Industry today as well as from Research and Academia to understand the limits of Power Dissipation i.e. how low can Power Dissipation be in CMOS logic and what the mechanisms are for realizing this

Power Dissipation and challenges in Power management have worsened with Technology Scaling.

There are several reasons why Power is dissipated in CMOS logic and considering the Total Power Dissipated, we can essentially split it up into 2 different components.

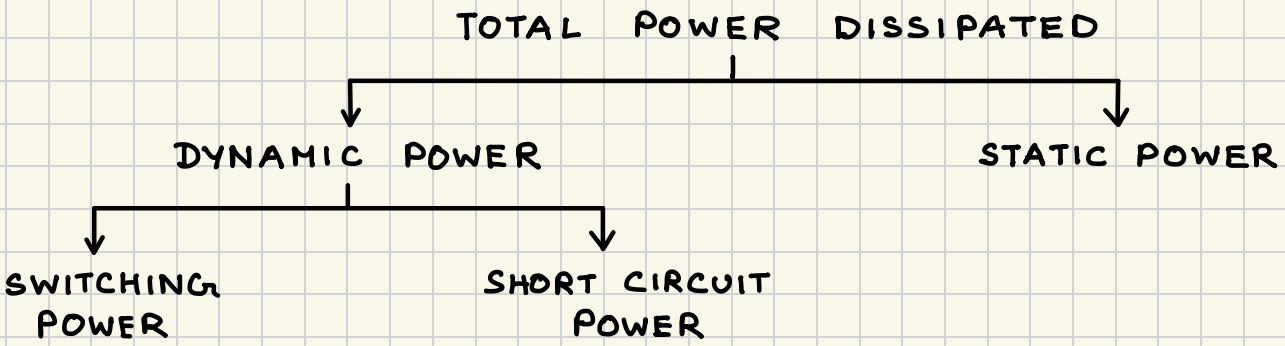
i.e. TOTAL POWER has two components :

① DYNAMIC POWER ② STATIC POWER

DYNAMIC POWER is consumed when the circuit is switching and STATIC POWER is consumed when the circuit is NOT switching. We will later look at why STATIC POWER is important as well, particularly in mobile form factors.

The DYNAMIC POWER can be further split into 2 components, SWITCHING POWER and SHORT CIRCUIT POWER.

In the earlier days of CMOS logic, the SWITCHING POWER was the most significant part of DYNAMIC POWER and it still is but with technology scaling, while, SHORT CIRCUIT POWER has increased considerably, STATIC POWER, has increased exponentially.



We need to also keep in mind the difference between Power and Energy as we proceed further with our discussions on Power Dissipation. You may remember from ECE 2040,

POWER ⟶ RATE of CHANGE of ENERGY

i.e. $\frac{d(\text{ENERGY})}{dt}$

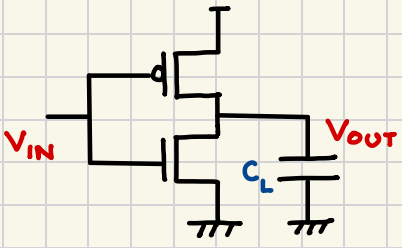
Let us start by looking at the SWITCHING component of DYNAMIC POWER.

But first let us look at SWITCHING ENERGY and from Energy, we can then find the POWER.

SWITCHING ENERGY (DYNAMIC POWER) ⟶

You will learn in details about the role of technology scaling in advanced VLSI courses (e.g. ECE 4130) but note that in most VLSI courses we always use the INVERTER as a prototypical device (logic gate) and if we can analyze an INVERTER, we can take that concept and apply it to other devices.

∴ Here too, let us start with an INVERTER.



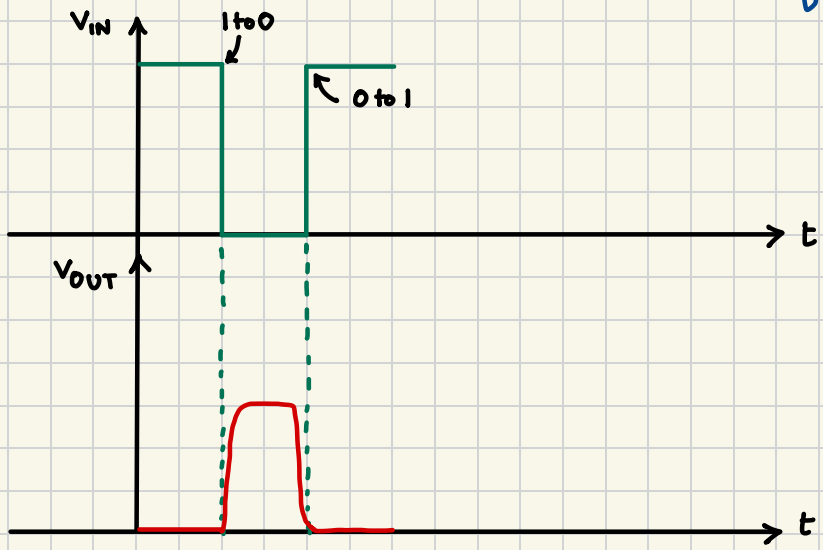
We have seen in MODULE 5, with our discussions on DELAY, that we can assume that there will be a LOAD CAPACITANCE associated with the INVERTER (C_L) and this capacitance, will be given by:

$$C_L = C_{OUT(driver)} + C_{WIRE} + C_{IN(load)}$$

Let us assume now that V_{IN} switches from 1 to 0 as shown in the figure below. There will be another switching transition of V_{IN} as the signal goes back from 0 to 1.

During the first switching transition (1 to 0), the pMOS will start charging up the output capacitance and during the second switching transition (0 to 1), the output capacitance will discharge

through the nMOS. This can be seen in the figure below.



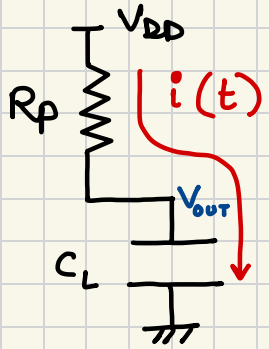
Thus, every clock cycle, i.e. when the signal goes from 1 to 0 and back to 1, there will be one charging event and one discharging event and there will be an energy dissipation component associated with each of these switching activities.

EXPRESSION FOR CHARGING ENERGY:

Power drawn from V_{DD} ($P_{VDD}(t)$) = $V_{DD}\, i(t)$

where $i(t)$ is the charging current.

(∵ We know if there is current $i(t)$ flowing into a node with potential $v(t)$, the instantaneous Power, $P(t) = v(t)\, i(t)$)



The charging current $i(t)$ flows from V_{DD} through the PMOS and charges C_L as shown in the figure.

$$\therefore\ i_{charging}(t) = C_L \frac{d}{dt}(V_{OUT}) \quad \text{(if } C_L \text{ is constant)}$$

and if C_L is NOT a constant,

$$i_{charging}(t) = \frac{d}{dt}(Q(t)) = \frac{d}{dt}(C_L V_{OUT})$$

$$\therefore \quad P_{V_{DD}} = V_{DD} C_L \frac{d}{dt}(V_{OUT}) \quad \text{(assuming } C_L \text{ is constant)}$$

$\therefore$ Energy drawn from the Supply $\mathcal{E}_{V_{DD}}$:-

$$\mathcal{E}_{V_{DD}} = \int_{t=0}^{\infty} P_{V_{DD}}\, dt = \int_{t=0}^{\infty} V_{DD} C_L \frac{d}{dt}(V_{OUT})\, dt$$

dt gets cancelled and this integration will change from an integration on time to an integration on voltage.

$\therefore$ We have to change the limits of the integration.

When, $t = 0 \longrightarrow V_{OUT} = 0$

and when, $t \to \infty \longrightarrow V_{OUT} = V_{DD}$

$$\therefore \quad \mathcal{E}_{V_{DD}} = \int_{V_{OUT}=0}^{V_{DD}} V_{DD} C_L \, dV_{OUT}$$

$$\text{or,} \quad \mathcal{E}_{V_{DD}} = V_{DD} C_L \, V_{OUT} \Big|_0^{V_{DD}}$$

$$\text{or,} \quad \boxed{\mathcal{E}_{V_{DD}} = C_L V_{DD}^2} \quad \ldots (i)$$

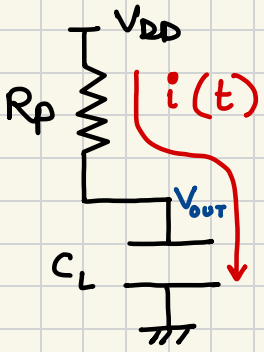
$\therefore$ During a $0 \to 1$ transition at the output, the energy drawn from the supply is $C_L V_{DD}^2$

Now, a part of this energy will be stored in the OUTPUT CAPACITOR as we have seen in ECE 2040, since the capacitor is an Energy storage device.

The rest of the energy will be dissipated.

∴ If we can find the Energy stored in the capacitor, we can find the energy dissipated.

Let us first look at the Power on the capacitance



$$P_c(t) = i_{charging}(t)\, V_{OUT}(t)$$

(∵ V_{OUT} is the voltage at the node of concern in this case)

and, $$i_{charging}(t) = C_L \frac{d\,(V_{out}(t))}{dt}$$

(assuming C_L is constant)

$$\therefore \quad P_c(t) = C_L V_{out} \frac{d\,(V_{out})}{dt}$$

which is a little different from the previous expression for Power that we had seen.

∴ ENERGY STORED in C_L after a 0→1 transition of the output

$$\mathcal{E}_c = \int_{t=0}^{\infty} P_c(t)\, dt$$

or, $$\mathcal{E}_c = \int_{t=0}^{\infty} C_L V_{out} \frac{d(V_{out})}{dt}\, dt$$

Again, dt gets canceled and this becomes an integration on Voltage.

When, $t = 0$, $V_{out} = 0$

and when, $t \to \infty$, $V_{out} = V_{DD}$

$$\therefore \quad \mathcal{E}_c = \int_{V_{out}=0}^{V_{DD}} C_L V_{out}\, d V_{out}$$

$$\text{or,} \quad \mathcal{E}_c = C_L \left[\frac{V_{out}^2}{2} \right]_0^{V_{DD}}$$

$$\text{or,} \quad \boxed{\mathcal{E}_c = \frac{C_L V_{DD}^2}{2}} \quad \ldots (ii)$$

which is what you have seen in ECE 2040 too.

$$\text{Energy} = \frac{1}{2} C V^2$$

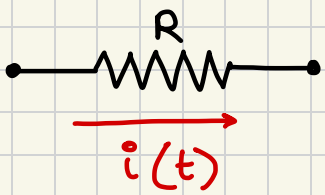
$\therefore$ During charging, $\frac{1}{2} C_L V_{DD}^2$ of energy is stored in the OUTPUT CAPACITANCE

The rest of the energy (Equation (i) - (ii))

$$C_L V_{DD}^2 - \frac{1}{2} C_L V_{DD}^2 = \frac{1}{2} C_L V_{DD}^2$$

will be the energy dissipated across R_p during charging.

We know that when current flows through a resistor,



the corresponding Power is given by,

$$P = i^2 R$$

and this dissipated Power will be radiated as heat.

Here, we have a similar condition where the Power will be dissipated as the charging current flows through the pMOS (R_p)

$\therefore$ The ENERGY DISSIPATED across R_p during $0 \rightarrow 1$ OUTPUT TRANSITION $= \frac{1}{2} C_L V_{DD}^2$

The ENERGY STORED in C_L during $0 \rightarrow 1$ OUTPUT TRANSITION $= \frac{1}{2} C_L V_{DD}^2$

and the,

TOTAL ENERGY drawn from V_{DD} to charge $C_L = C_L V_{DD}^2$

The dissipated ENERGY $\frac{1}{2} C_L V_{DD}^2$ is the component of Energy that is lost as heat and this is why, a computer feels hot to our touch. There are millions of gates that are constantly switching and when we charge up the capacitors, these are in themselves not lossy, but the process of charging a capacitor will dissipate an EQUAL AMOUNT of energy on the resistance R_p. Finally there will be an Energy transformation and the ELECTRICAL ENERGY will transform to HEAT and this will increase the temperature of the device.

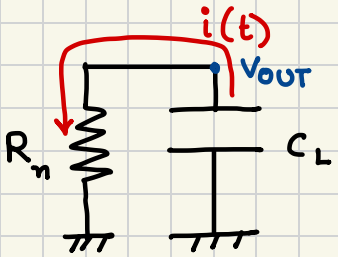
Another important observation is that:

The ENERGY DISSIPATED across R_p $\left(\frac{1}{2} C_L V_{DD}^2\right)$ is INDEPENDENT of the value of R_p.

What is the significance of this observation?

Even if we use a wide pMOS transistor, so that the R_p is low, the total Energy dissipated will still be, $\frac{1}{2} C_L V_{DD}^2$. However, the charging delay will decrease and the device will become faster. Thus, even if we use a very advanced device material, the device will become a lot faster but the energy dissipated does not change and this acts as a sort of boundary condition in CMOS logic.

Now, let us consider the second part of the SWITCHING PROCESS where the input V_{IN} transitions from 0 to 1. In this case, as we know, the capacitor will discharge through the nMOS.



V_{OUT} will now discharge from 1 to 0.

$\therefore$ TOTAL ENERGY stored in $C_L = \frac{1}{2} C_L V_{DD}^2$

(at the beginning of the discharging process)

After a 1 to 0 transition at the OUTPUT, the TOTAL ENERGY STORED in $C_L = 0$

This essentially means that $\frac{1}{2} C_L V_{DD}^2$ energy is dissipated across the nMOS during the discharging process.

NOTE → During the discharging process, the circuit is disconnected from V_{DD}. $\therefore$ We will not draw any more power from the supply.

Here, too, the TOTAL DISSIPATED ENERGY during the discharging process is independent of R_n.

$\therefore$ Again a wider nMOS will increase performance but energy dissipated will not change.

$\therefore$ TOTAL ENERGY DISSIPATED by an INVERTER

(as the OUTPUT goes from $0 \rightarrow 1 \rightarrow 0$)

$$= \underbrace{\frac{1}{2} C_L V_{DD}^2}_{\text{(charging)}} + \underbrace{\frac{1}{2} C_L V_{DD}^2}_{\text{(discharging)}} = \boxed{C_L V_{DD}^2}$$

which is equal to the energy drawn from the supply (V_{DD}) at the beginning of the clock cycle. Hence, energy is conserved.

$\therefore$ The TOTAL SWITCHING ENERGY per clock cycle

$$= \boxed{\mathcal{E}_{switching/cycle} = C_L V_{DD}^2} \quad \ldots(iii)$$

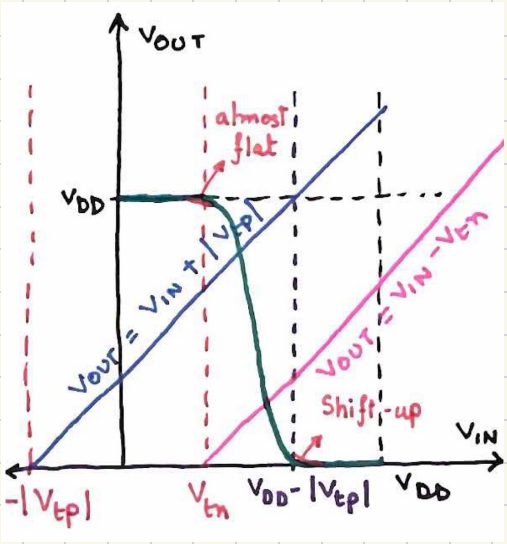
(half of which is lost during charging and the other half is lost during discharging)

Going back to our initial discussions, we need to keep in mind that this is a part of the DYNAMIC ENERGY during switching.

Now, we will look at the second component of DYNAMIC POWER / ENERGY which is the

SHORT CIRCUIT ENERGY :

During signal transitions, (both 0→1 and 1→0 transitions) the pMOS and the nMOS are both ON simultaneously for a very short time. Going back to our discussions on the VTC of an inverter, this will be the REGION 3 where both transistors are ON and in the SATURATION REGION.

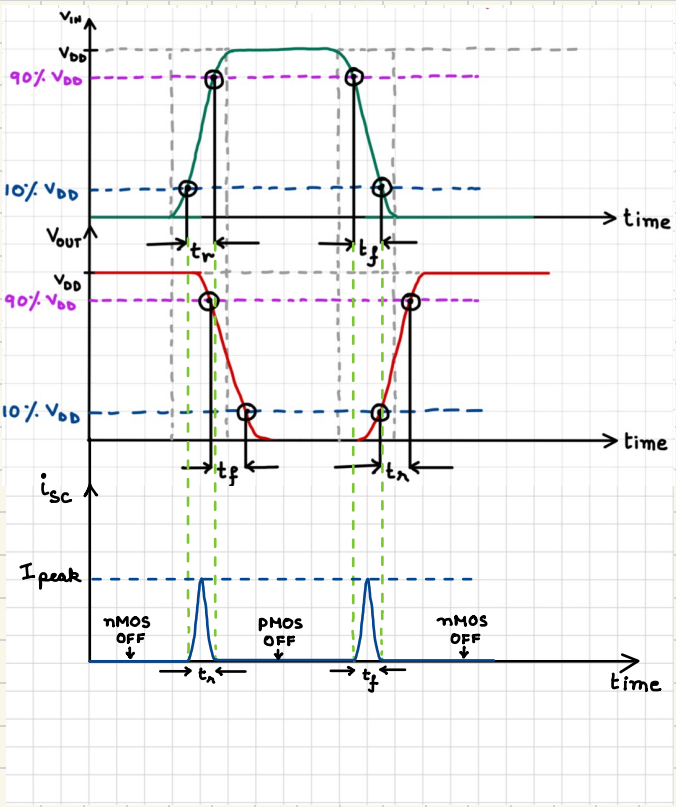


This contributes to a SHORT CIRCUIT current flowing from V_{DD} to GND, albeit for a very short time.

$\therefore$ This SHORT CIRCUIT CURRENT will contribute to a SHORT CIRCUIT ENERGY / POWER.

Now, let us go back to the RISE TIME and FALL TIME

discussions from MODULE 5, note that when the input signal transitions from 0→1 and again from 1→0 (i.e. during the RISE TIME and the FALL TIME) both transistors will be ON and there will be a SHORT CIRCUIT CURRENT. This current say i_{sc}, can be sketched as we can see in the figure. Initially, when $V_{IN} = 0$, the nMOS is OFF, the pMOS is ON and $i_{sc} = 0$. As V_{IN} increases in the t_r range, we see that i_{sc} also increases and then peaks before dropping to 0 after the pMOS turns OFF. Now, just the nMOS is ON and there is no SHORT CIRCUIT CURRENT. Similarly, i_{sc} will increase to a peak and fall to 0 again in the t_f range of V_{IN} as the nMOS is still ON and the pMOS is gradually turning ON as we can see in the figure.



Next, we will try to come up with a rough model for the ENERGY DISSIPATED due to the SHORT CIRCUIT CURRENT.

Let, $\mathcal{E}_{sc}$ be the energy dissipated due to the SHORT CIRCUIT CURRENT. Let, $\mathcal{E}_{sc}$ when V_{IN} goes from 0→1 and V_{OUT} goes from 1→0 be written as, $\mathcal{E}_{sc}^{V_{OUT}\,1\to0}$

We know, $$\mathcal{E}_{sc}^{V_{OUT}\,1\to0} = \int_{t=0}^{\infty} P_{sc}(t)\,dt$$

and, $$P_{sc}(t) = V_{DD}\, i_{sc}(t)$$

To estimate the SHORT CIRCUIT CURRENT, i_{sc}, we will assume the area under the curve (see fig.) to be a triangle. This is an approximation which will help us to roughly estimate i_{sc}.

Let the maximum or peak i_{sc} be I_{peak}.

$$\therefore \quad \mathcal{E}_{sc}^{V_{OUT}\,1\to0} = \int_0^{\infty} V_{DD}\, i_{sc}(t)\, dt$$

$$\text{or,} \quad \mathcal{E}_{sc}^{V_{OUT}\,1\to0} = \frac{1}{2} V_{DD}\, I_{peak}\, t_r$$

Assuming the area to be that of a TRIANGLE with height I_{peak} and base t_r.

Similarly, we can write,

$$\mathcal{E}_{sc}^{V_{OUT}\,0\to1} = \frac{1}{2} V_{DD}\, I_{peak}\, t_f$$

since we are looking at FALL TIME portion of the transition now.

$\therefore$ The TOTAL SHORT CIRCUIT ENERGY for the full cycle :-

$$\boxed{\mathcal{E}_{sc/cycle} = \frac{1}{2} V_{DD}\, I_{peak}\, (t_r + t_f)} \quad \ldots(iv)$$

$\therefore$ The TOTAL DYNAMIC ENERGY per cycle will be the SUM of the SWITCHING ENERGY (iii) and the SHORT CIRCUIT ENERGY (iv)

$$\boxed{\mathcal{E}_{DYN/cycle} = C_L V_{DD}^2 + \frac{I_{peak}\, V_{DD}}{2} (t_r + t_f)}$$

In well designed processors, the RISE TIMES and the FALL TIMES are low so that the short circuit energy dissipation stays low.

So we can conclude :

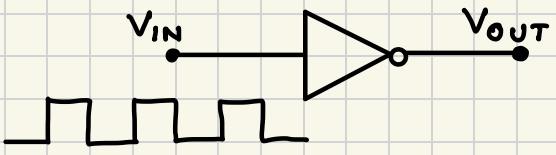
① Higher RISE/FALL TIMES implies higher Energy dissipation ∴ It is important to maintain good signal slopes and by 'good' we mean that the slopes form a small portion of the total clock cycle. If this is true, we then can say that we have a well-designed CMOS logic.

② For a well-designed CMOS logic, the SHORT CIRCUIT ENERGY is about 10% to 20% of the TOTAL DYNAMIC ENERGY.

NOTE: What we have looked at here is an approximation of the SHORT CIRCUIT ENERGY. To have a more exact understanding of the short circuit energy, we can do SPICE simulations and evaluate it numerically but that is beyond the scope of our course.

Next, we will see how we can take our understanding of the DYNAMIC ENERGY that we have seen so far and translate that to DYNAMIC POWER.

Let us consider an INVERTER and let's say the input of the inverter is switching multiple times i.e. there are multiple switching events at the input. Let the frequency of the switching be 'f'. This means that,



The number of (0→1) Transitions + the number of (1→0) transitions per second = f

e.g. if this is running at 1 GHz then the total no. of (0→1→0) transitions = 10^9 or 1 billion

$\therefore$ Energy / cycle = $\mathcal{E}_{DYN}$

and, $\therefore$ Energy / second = $\mathcal{E}_{DYN}$ × (no. of cycles / second)

or, Energy / second = $\mathcal{E}_{DYN} f$

We know, Power = $\frac{d}{dt}$ (Energy)

$\therefore$ Energy / second = $\text{Power}_{DYN} = \mathcal{E}_{DYN} f$

$\therefore$ POWER DISSIPATED →

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

where, $C_L V_{DD}^2 f$ → Switching Dynamic Power

and, $I_{peak} V_{DD} \left(\frac{t_0 + t_f}{2}\right) f$ → Short Circuit Power

This P_{DYN} holds for our INVERTER which goes through 'f' number of switching events per clock cycle.

Note that the dependance of Power to frequency is linear.

$\therefore$ If frequency increases, then power dissipation increases as 'f' (linearly)

and,

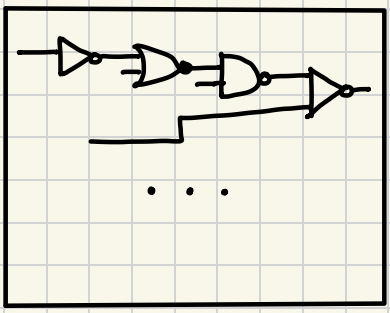
If V_{DD} increases, then the power dissipation increases as 'V_{DD}^2' (quadratically)

Note that the main component of dynamic power is the switching component which has the V_{DD}^2 term.

∴ For e.g. if we reduce the supply from 1V to 0.8V it is not a 20% reduction in power. This will be 36% which is significant.

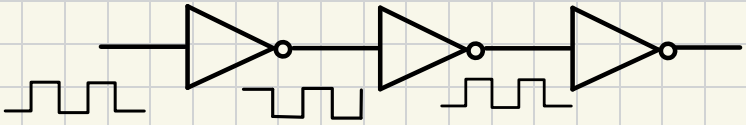
This is the reason why the Power Supply of processors has come down from about 5V to about 0.8V, in the last 40 years. This Power Supply scaling has resulted in lower power transistors and better power management techniques. This is one of the salient features of our designs today.

Say, we have a microprocessor with n-gates.



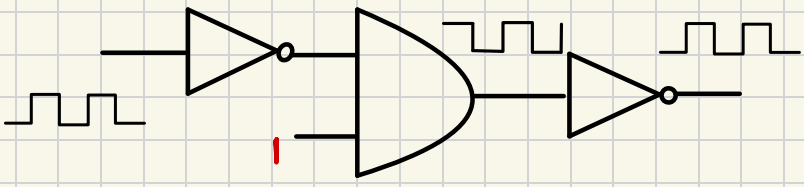
μ-processor - 'N' gates

We know that if we have a chain of inverters, if the input of one inverter



switches, then the inputs of all of the inverters will switch or toggle.

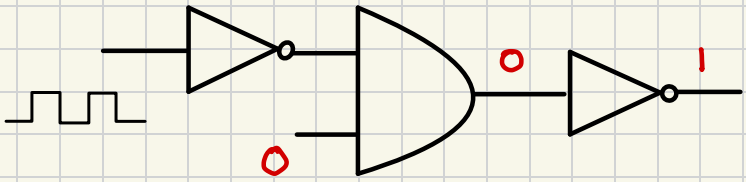
Now, say, we have an inverter connected to an AND gate



and then that is connected to another inverter.

Now, if the other input

of the AND gate is connected to '1' then the AND gate output or the input to the next inverter will also switch. However, if the AND gate input is '0' then the subsequent inputs will not switch.



What we are trying to say here is that although a microprocessor may have millions of logic gates, not all logic gates will switch all the time. Only a fraction of the logic gates will switch depending on the logic being implemented e.g. if we have an OR gate, then the output will not switch if one of the inputs is a '1'.

Let us assume that in our microprocessor with N gates, 'm' gates are switching/cycle $(m << N)$
where $m \longrightarrow$ statistical estimate
then,

$$P_{DYN/\mu p} = \left[C_L V_{DD}^2 f + I_{peak} V_{DD} \left(\frac{t_r + t_f}{2} \right) f \right] m$$

will be our total dynamic power for the microprocessor.
And the average power per logic gate can be written as,

$$\bar{P} \simeq \left[C_L V_{DD}^2 f + I_{peak} V_{DD} \left(\frac{t_r + t_f}{2} \right) f \right] \left(\frac{m}{N} \right)$$

This $\frac{m}{N}$ is an important fraction which tells us how many gates are actually switching.

$$\therefore \quad \frac{m}{N} = \frac{\text{No. of gates actually switching}}{\text{Total no. of gates}}$$

which is actually,

the probability that a gate is switching

This is called the SWITCHING ACTIVITY

There is no way to exactly determine the switching activity and will change depending on the workload, the programs being executed, etc.

The symbol α (alpha) is used to represent the switching activity $\frac{m}{N}$.

$\therefore$ If we wait for a sufficiently long period of time, then α represents the probability that a gate is switching.

$$\therefore \text{Power} = C_L V_{DD}^2 \alpha f + I_{peak} V_{DD} \left(\frac{t_r + t_f}{2}\right) \alpha f$$

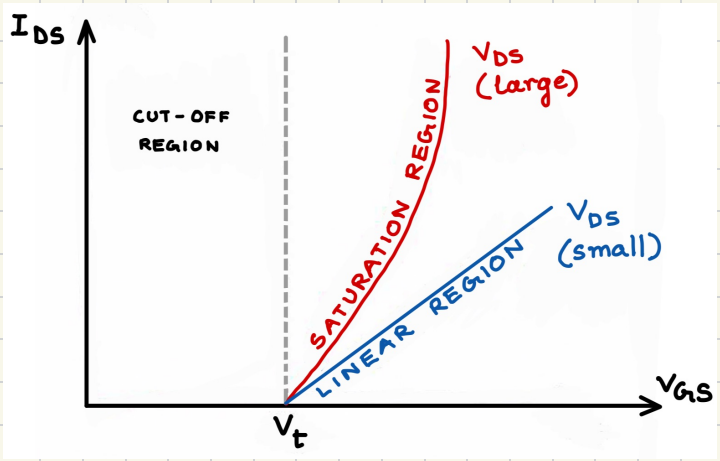
$\therefore$ For a chain of inverters $\alpha = 1$ (always switching)
and, for an AND gate with one input equal to 0, $\alpha = 0$ (const.)

We can find an approximate value of α and this is between 5% to 10% in modern microprocessors.

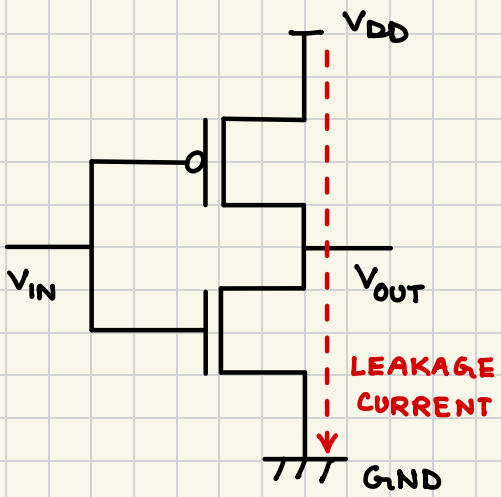
This relatively low value of α is what keeps the power dissipation in a microprocessor under control.

We were discussing so far about the DYNAMIC component of POWER/ENERGY, but as we have seen earlier, we also have a STATIC component of POWER, which we will discuss now.

So far in our course, we haven't discussed a lot on LEAKAGE CURRENT. But you know that if we draw the TRANSFER CHARACTERISTICS of a MOSFET (let's say of an nMOS), in cut-off, although we assume that there is no current flowing through the transistor, there will in reality be a LEAKAGE CURRENT even when the transistor is OFF.



This LEAKAGE or STATIC current that always flows through the transistor even when it is not switching, is responsible for the STATIC POWER component.



If we consider an INVERTER, when $V_{IN} = 0$ (and $V_{OUT} = 1$) and the pMOS is ON, although the nMOS is in cut-off, there will be some LEAKAGE current flowing through it. Similarly, when $V_{IN} = 1$ and the nMOS is ON, there will be a LEAKAGE current flowing through the pMOS. As a result, there is ALWAYS a current flowing from V_{DD} to GND, which is responsible for the STATIC POWER DISSIPATION.

NOTE :- This is the STATIC POWER component and there is no ENERGY associated with this. ENERGY is associated with an activity. Here current is constantly flowing from the supply and there is no switching activity, so we will not associate ENERGY with STATIC POWER.

If,

I_{OFF}^{n} ⟶ OFF CURRENT of the nMOS

I_{OFF}^{p} ⟶ OFF CURRENT of the pMOS

then,

TOTAL LEAKAGE/STATIC POWER = Voltage × Current

We can assume the

$$\text{TOTAL LEAKAGE CURRENT} = \frac{I_{OFF}^{n} + I_{OFF}^{p}}{2}$$

which means that we are assuming that in a logic gate,

1/2 of the time the pMOS is ON and the nMOS is leaking

and the other

1/2 of the time the nMOS is ON and the pMOS is leaking

∴ If we now write down the expression for the

TOTAL AVERAGE POWER DISSIPATED ⟶

$$P_{AVG.} = \underbrace{C_L V_{DD}^2 \alpha f}_{\text{SWITCHING POWER}} + \underbrace{I_{peak} V_{DD} \left(\frac{t_r + t_f}{2}\right) \alpha f}_{\text{SHORT CIRCUIT POWER}} + \underbrace{V_{DD} \left(\frac{I_{OFF}^{n} + I_{OFF}^{p}}{2}\right)}_{\text{LEAKAGE POWER}}$$

SWITCHING POWER + SHORT CIRCUIT POWER ⟶ DYNAMIC POWER (P_{DYN})

LEAKAGE POWER ⟶ STATIC POWER (P_{STATIC})

NOTE :- The DYNAMIC POWER component depends on the frequency as it is related to switching.

When, $f = 0$ (no switching), then $P_{DYN} = 0$

and we have just the LEAKAGE POWER / STATIC POWER P_{STATIC}

The STATIC POWER cannot be avoided even if there is no switching. This is the reason why for e.g. your mobile phone will lose battery power even when you are not using it.

From the expression of Power that we have derived, the following dependencies are important to note :-

$$P_{DYN} \propto V_{DD}^2 \text{ (SWITCHING PORTION)}$$

$$P_{STATIC} \propto V_{DD} \text{ (assuming the off currents remain constant)}$$

In MODULE 5, we had seen that the Delay,

$$t_p \propto \frac{1}{V_{DD}}$$ which means that the

$$\text{frequency} \propto V_{DD}$$

Thus, a 2x increase in V_{DD} can make the CMOS logic faster but will burn 4x higher power (since, $P_{DYN} \propto V_{DD}^2$)

Another thing to note is the Switching component is about 80-90% of P_{DYN} and thus, the Short circuit component

is about 10-20%. But the distribution between the Static Power and Dynamic Power is not easily determined. In modern microprocessors, the supply and frequency can change depending on what processes are being run. You will learn more about this in advanced VLSI courses.