Design and Analysis of VLSI Subsystems
Dr. Madhav Rao
Department of Electronics and Communication Engineering
International Institute of Information Technology, Bangalore

Lecture - 67
Voltage Scaling

Hello students, welcome to this lecture on the Power module, where we are going to talk on the voltage variations or choosing or selecting a different rail voltage. In a digital design we generally come across a different rails or we have an allowance to choose different rails and then subsequently design our digital circuit by using different rails.

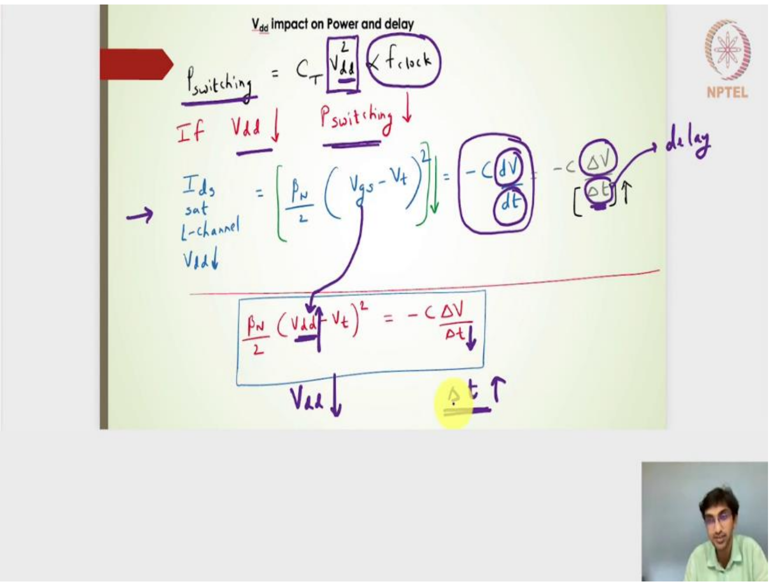
What generally goes in a digital designer mind is if I want to save a lot of power let us use a lower voltage rail, but at the same time if I want to get a higher performance or a lower amount of delay, then we will use or we will choose a higher rail, or a higher V_{dd} value.

What happens is, in this particular catering to all this features for a particular digital design, if you want to specify or cater to almost all the features we need to cater towards the power benefits or the power savings. We also need to cater towards the performance you know maximize the performance benefits.

In that sense the digital design the associate design or the chip design is kind of divided into several components and then one part of the chip area or the chip layout kind of will be designed by the lower rail voltage. The other portion or the other footprint of the chip design will be designed using the another rail.

It could be a lower rail design on one part of the layout, the another using the entirely a different set of rail on the other side of the chip. In this particular lecture we will have a look at what are the advantages of using a lower rail, what are the advantages of using a higher rail voltage and also try to see how do we interface between a lower rail design to that of a higher rail voltage design. Let us move ahead.

(Refer Slide Time: 02:25)



V_{dd} impact on Power and delay

$P_{switching} = C_T \boxed{V_{dd}^2} \alpha f_{clock}$

If $V_{dd} \downarrow$ $\quad P_{switching} \downarrow$

$I_{ds\ sat\ L\text{-}channel} = \left[\frac{\beta_N}{2}\left(V_{gs} - V_t\right)^2\right] = -C\frac{dV}{dt} \quad -C\frac{\Delta V}{\Delta t} \uparrow$ delay

$V_{dd} \downarrow$

$\frac{\beta_N}{2}(V_{dd} - V_t)^2 = -C\frac{\Delta V}{\Delta t}\downarrow$

$V_{dd}\downarrow \quad \Delta t \uparrow$

In this particular slide what I have shown is the impact of the V_{dd} rail voltage impact on the power and delay. What I mean in this particular case is,

$$P_{switching} = C_T V_{dd}^2 \alpha f_{clock}$$

Its called as the switching power, also called as or also referred to as the dynamic power. Now, if in this particular case the component of V_{dd}^2, if the V_{dd} increases the power switching also increases. Infact, it is having a relation of the square of V_{dd} and if the V_{dd} decreases the power switching actually decreases in the nature of the square profile. It may actually, if I draw the power switching versus the V_{dd} rail, you will see a parabolic profile.

A lower value of V_{dd} will definitely help in reducing the power, will definitely help in reducing the energy that will be delivered by the rail voltage. Now, but what are the causes, what are the compromises one has to do if I use a lower rail voltage?

Let me have a look at the current, if I actually just look at the saturation using a long channel current model, if I actually express the saturation current expression.

$$I_{\substack{ds\\ sat\\ L-ch}} = \left[\frac{\beta_N}{2}(V_{gs} - V_t)^2\right] = -C\frac{dV}{dt} = -C\frac{\Delta V}{\Delta t}$$

Which is used for the discharging. If I choose an inverted design and this particular equation will be valid for the discharging of the capacitor of the load capacitor. If I want to approximate this $\frac{dV}{dt}$ as $\Delta V/\Delta t$, what I am going to say here is, if this current decreases this Δt component increases. Because the current is inversely proportional to the Δt component which is nothing but on an you know in a very approximate manner we can say that the delay.

This is the delta t what it means is for charging or discharging the capacitor for a ΔV amount of time, ΔV amount of voltage we require this delta t amount of time. Based on this current whether it is higher or lower this Δt is going to vary. The Δt is going to vary to fill this or to charge this ΔV voltage or to discharge this ΔV voltage. If my V_{gs} value is nothing but the input is coming from the logic, the logic of the previous gate and if it is either having a V_{dd} voltage.

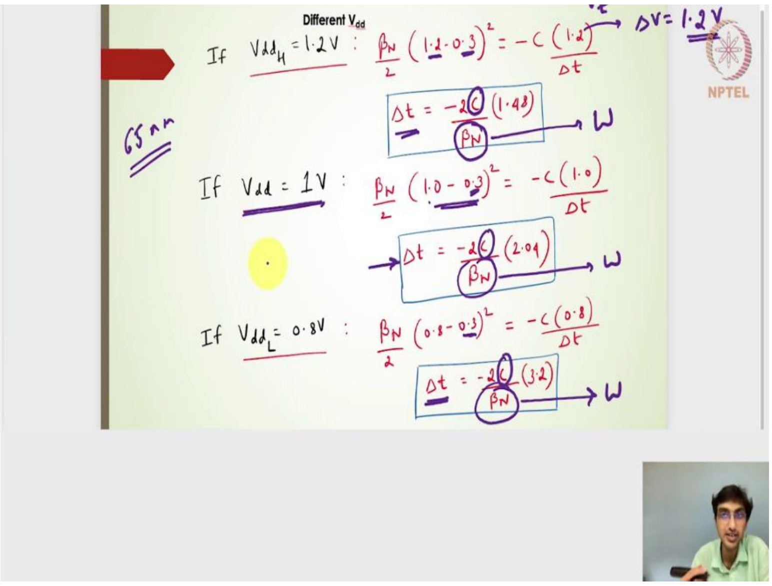
In that sense I will have,

$$\frac{\beta_N}{2}(V_{dd} - V_t)^2 = -C\frac{\Delta V}{\Delta t}$$

In that sense if this V_{dd} is going to increase, if I will have an increased value of the rail voltage this Δt is going to be very very small. That means, that the delay is going to be very very small, it is going to be very very highly performance efficient, but on the other side if the V_{dd} is going to be very very low, in that sense I will have a Δt to be very very high.

That means the output for the output node voltage whether it is charging or discharging it is going to take a lot amount of time to reach to the 50 percent of the final output voltage. The 50 percent because that is where the propagation delay rising and falling are is kind of defined, hope this is clear.

(Refer Slide Time: 06:15)



Different V_{dd}

65nm

If $V_{dd_H} = 1.2V$: $\frac{\beta_N}{2}(1.2 - 0.3)^2 = -C\frac{(1.2)}{\Delta t} \rightarrow \Delta V = 1.2V$

$\boxed{\Delta t = -\frac{2C}{\beta_N}(1.48)} \rightarrow W$

If $V_{dd} = 1V$: $\frac{\beta_N}{2}(1.0 - 0.3)^2 = -C\frac{(1.0)}{\Delta t}$

$\boxed{\Delta t = -\frac{2C}{\beta_N}(2.04)} \rightarrow W$

If $V_{dd_L} = 0.8V$: $\frac{\beta_N}{2}(0.8 - 0.3)^2 = -C\frac{(0.8)}{\Delta t}$

$\boxed{\Delta t = -\frac{2C}{\beta_N}(3.2)} \rightarrow W$

Moving ahead let us take an example, if I use a V_{ddH} I am actually annotating the different rail voltages, especially for a 65nm technology node, we will have in the technology node libraries we will see typically these three voltage rails. The V_{ddH}, low and the typical V_{dd}, which is around 1.2, 1 and 0.8 volts and if that is the case my current expression should be equal to that of the load capacitance, the discharging equation.

The discharging of the capacitance current equation, this is the saturation equation which is equated to the capacitive load discharging equation. My saturation current is nothing but,

$$\frac{\beta_N}{2}(1.2 - 0.3)^2 = -C\frac{1.2}{\Delta t}$$

In this case the ΔV is either going from 0 to 1.2 volts or coming from 1.2 to 0 volts. In this it is a discharging current equation, it is the difference of 1.2 volts. I will get Δt, if I bring in this Δt here and take all these things in the denominator side,

$$\Delta_t = -\frac{2C}{\beta_N}(1.48)$$

If I use a V_{dd} of 1 volts, I will get this particular expression,

$$\frac{\beta_N}{2}(1.0-0.3)^2 = -C\frac{1.0}{\Delta t}$$

$$\Delta_t = -\frac{2C}{\beta_N}(2.04)$$

This β_N which is actually a function of the width of the transistor, this β_N the width which is also a function of the width this β_N that is a function of the width of the transistor.

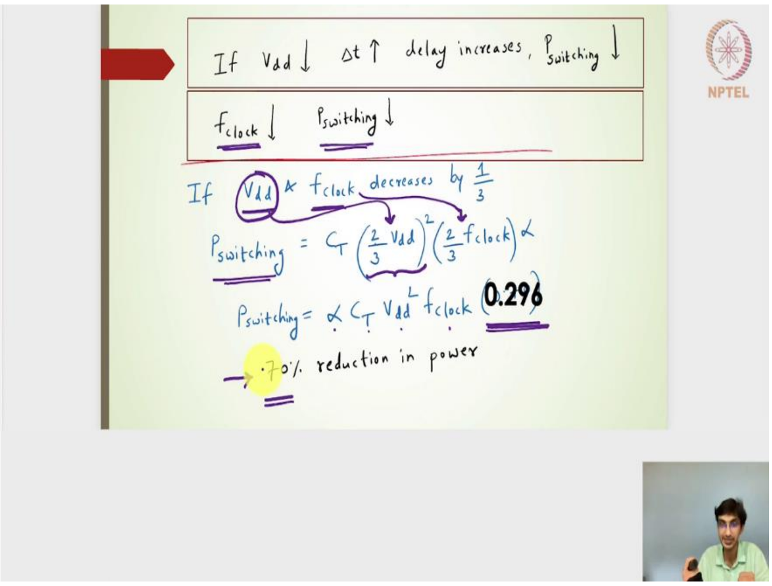
Let us say that I will have the same width of the transistor, let us say that I will have the same load capacitance connected to the output node. In that sense what we are saying is if I have a 0.8 volts I will have the largest amount of the time Δt.

$$\frac{\beta_N}{2}(0.8-0.3)^2 = -C\frac{0.8}{\Delta t}$$

$$\Delta_t = -\frac{2C}{\beta_N}(3.2)$$

The higher the V_{dd} value, the lower the Δt and if the lower the V_{dd} value, I will have the higher the Δt, but the lower the V_{dd} value will definitely reduce the power, will definitely reduce the average energy that will be delivered by the V_{dd} rail, but it will have an impact on the delay.

(Refer Slide Time: 09:04)



If $V_{dd}\downarrow$ $\Delta t\uparrow$ delay increases, $P_{switching}\downarrow$

$f_{clock}\downarrow$ $P_{switching}\downarrow$

If V_{dd} & f_{clock} decreases by $\frac{1}{3}$

$P_{switching} = C_T\left(\frac{2}{3}V_{dd}\right)^2\left(\frac{2}{3}f_{clock}\right)\alpha$

$P_{switching} = \alpha C_T V_{dd}^2 f_{clock}$ (0.296)

→ ·70% reduction in power

What it means is if $\mathrm{V_{dd}}$ decreases the Δt increases, delay increases, but the P power switching or the energy switching or the dynamic energy or the dynamic power is going to decrease. The another component of the power is nothing but the clock frequency. If the clock frequency itself is reduced instead of 1Ghz if I reduce it to say 1Mhz it is going to have a 1000 times less power that will be switched in that particular clock cycle.

Now what if the power switching P switching is a function of $\mathrm{V_{dd}}$, P switching is a function of $\mathrm{f_{clock}}$ as well. If I actually reduce both of them by 1/3 times, I know that the power switching is going to be 1/3 times means I will now have the $\mathrm{V_{dd}}$ as a reduced $\mathrm{V_{dd}}$,

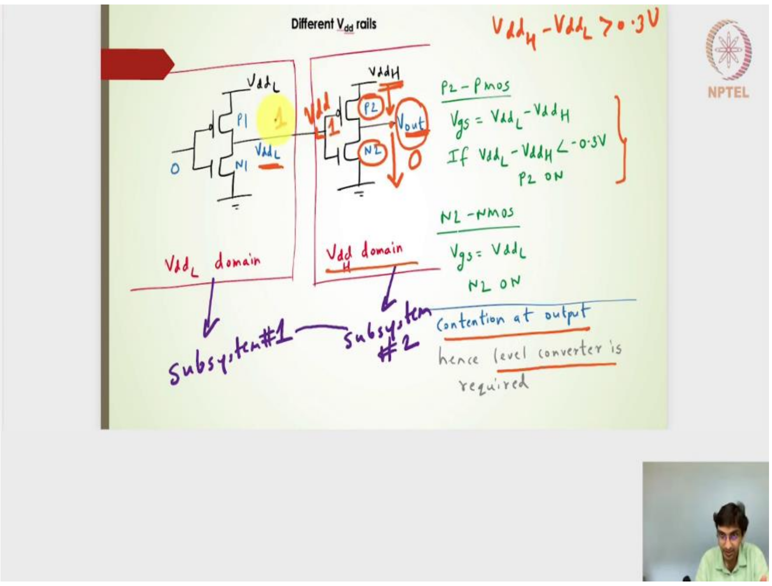
$$\mathrm{P_{switching}} = \mathrm{C_T}\left(\frac{2}{3}\mathrm{V_{dd}^2}\right)\left(\frac{2}{3}\mathrm{f_{clock}}\right)\alpha$$

$$\mathrm{P_{switching}} = \alpha \mathrm{C_T V_{dd}^2 f_{clok}}(0.296)$$

What 0.296 represents? It is 70% reduction in the power. We will have a 70% reduction in the power although this $\mathrm{V_{dd}}$ rail has been decreased to $\frac{2}{3}$. That means, my delay for the output to either charge the load capacitance or to discharge the load capacitance is going to increase, with an advantage of having a 70% reduction in the power.

Whenever we are dealing with power and delay, there is always a trade-off. If I want to design my circuit in such a way that it will have a lot of power savings, then the delay is naturally going to increase. If I want to design my circuit so that the delay has to be really fast or the delay has to be really minimum then the power will naturally increase. One has to deal with both the parameters.

(Refer Slide Time: 11:15)



Different V_{dd} rails

$V_{ddH} - V_{ddL} > 0.3V$

P2 – Pmos

$V_{gs} = V_{ddL} - V_{ddH}$

If $V_{ddL} - V_{ddH} < -0.3V$

P2 ON

N2 – NMOS

$V_{gs} = V_{ddL}$

N2 ON

Contention at output

hence level converter is required

Now let us take a look at while we design one side of the circuit using a lower rail and on the other side of the circuit we have a higher rail and there always is a point where we have to connect between the two Vdd rails design. Let us say that I have a lot of circuit here in the V_{ddL} domain, I have a subsystem 1 here, I will call it as subsystem number 1 and here I will have a subsystem number 2.

This subsystem number 2 requires the higher rail voltage this subsystem number 1 requires a lower V_{dd} voltage for whatever reasons it is. One may be the V_{dd} lower rail is just to make sure that this subsystem 1 is very power efficient. Whereas, the the subsystem 2 is need to be performance efficient. That is why we have used a higher rail, but always there needs to be an interface between them. The output of the 1st subsystem needs to be connected to the 2nd subsystem.

In this particular example I have taken a simple example, I have taken a simple circuit here an inverter connected or designed in the V_{ddL} domain and which gets connected to another inverter which is designed in the V_{ddH} domain.

The V_{ddH} domain and V_{ddL} domain means I will have the not only the rail voltages as V_{ddL}, but also the output node voltage will be the maximum voltage it can reach is V_{ddL}. Similarly, here it is not only the rail voltage which is V_{ddH}, but also the output voltage the maximum voltage it reaches is V_{ddH}.

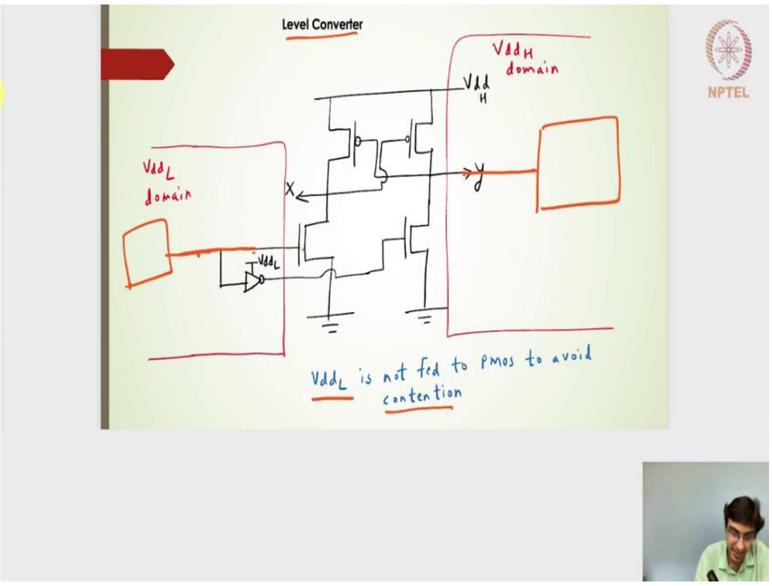
In this sense let us say that I will have a 0 at the input side of this inverter and then the output I will have a $\mathrm{V_{ddL}}$. If I have a output $\mathrm{V_{ddL}}$ here, it will be supplied as an input to the inverter which is designed in the higher rail voltage domain. The $\mathrm{V_{ddL}}$ and then this is $\mathrm{V_{ddH}}$. The $\mathrm{V_{ddL}}$ is what is supposed to or what is expected from this particular two inverters is the output. This is 0, this will be 1 here and then this one is going to be supplied as a 1 here and then this should be 0. But what happens here is because this 1 and this 1 are of different levels. This 1 is nothing but $\mathrm{V_{ddL}}$, the $\mathrm{V_{ddL}}$ value reaches here, $\mathrm{V_{ddL}}$ value reaches here this P2 transistor, actually depends on the $\mathrm{V_{gs}}$ value or the $\mathrm{V_{sd}}$ value in the PMOS side.

It is $\mathrm{V_{ddH}} - \mathrm{V_{ddL}}$ here. If the difference is more, the $\mathrm{V_{ddH}} - \mathrm{V_{ddL}}$, if the difference is more than 0.3 volts, this P2 transistor will be on and $\mathrm{V_{ddL}}$, if this is likely to be more than 0.3 volts itself, this N2 transistor will also be on.

I have P2 transistor N2 transistor both being on that means, that the current. This particular node is getting charged by the P2 transistor and then this particular node is also getting simultaneously discharged by the N2 transistor. I will have a continuous amount of current that will be flowing from the $\mathrm{V_{ddH}}$ rail to the ground rail. Unnecessarily there will be a power that is being dissipated to make this output voltage stand at a particular steady state value.

In that sense what I am saying is, the same thing has been written here. I will have a contention at the output that means, contention means the output node voltage in the $\mathrm{V_{ddH}}$ domain is going to be pulled by the $\mathrm{V_{ddH}}$ as well as pushed down by the N2 transistor as well as pulled up by the P2 transistor, and that is why it is called as a contention at the output and unnecessarily power is being dissipated in the subsystem module number 2. Hence a level converter is required to ensure that when it is passing this logic level 1 from $\mathrm{V_{ddL}}$ domain to the next one, the next one sees it as a $\mathrm{V_{ddH}}$ logic, this 1 of $\mathrm{V_{ddL}}$ should be converted as a $\mathrm{V_{ddH}}$ when it comes into the input of the circuit in the $\mathrm{V_{ddH}}$ domain, hope this is clear.

(Refer Slide Time: 15:50)

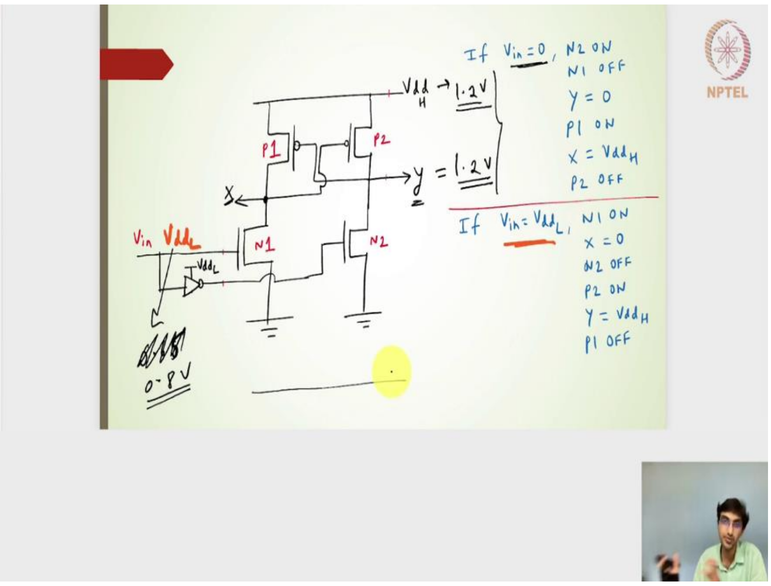


Moving ahead, this is the level converter which one will usually find in the digital design. Let us say that I have a circuit here, it could be an inverter circuit the output of this circuit is going here. This output needs to be connected or needs to be going connected from here and going to the another circuit in the V_{ddH} domain.

We can have an inverter here an inverter here as well. Now, following the same example which we have discussed in the last slide. If in this case notice that in the previous slide what we were doing is the output of the V_{ddL} domain was connected to both the NMOS and PMOS transistors of the V_{ddH} domain. Then the V_{ddL} level and was connected to both PMOS and NMOS transistor of the V_{ddH} domain and that was creating the contention problem because the PMOS transistor could have been on.

In this case what we are trying to do is, this particular logic, this the V_{ddL} level logic it goes to the or it gets connected only to the NMOS transistors. It does not get connected to the PMOS transistor at all, that is what I have written V_{ddL} is not fed to the PMOS to avoid the contention.

(Refer Slide Time: 17:15)

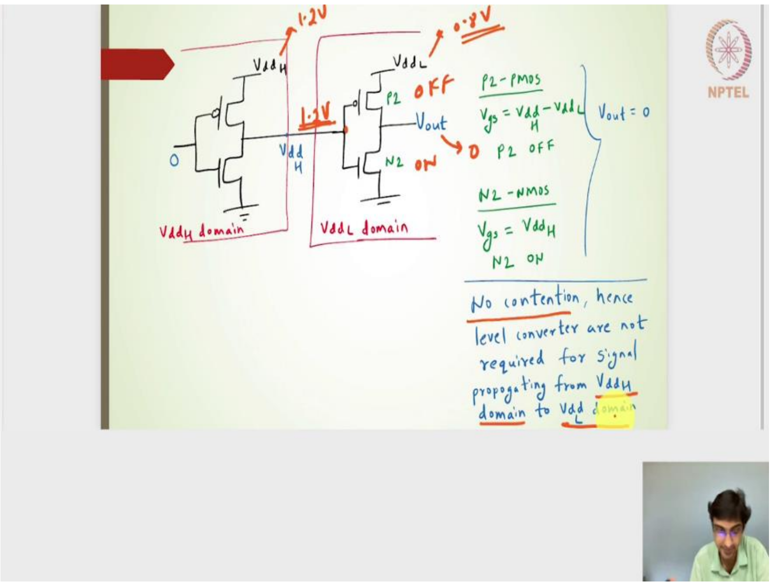


Let us take a look at it how it works. If I have a V_{ddL} line, in this case the V_{ddL} line as an input, if this is a V_{ddL} line that means, that it could be let us say I can have an example of 0.65 because it is and V_{ddH} as or instead of 0.65 let us why not use 0.8 volts and V_{ddH} as 1.2 volts. If it is 0.8 volts we know that this N1 transistor it is above the threshold voltage of 0.3 volts, this N1 transistor will be on. This output will be pulled down to ground, this output is cross coupled and connected to the P2 transistor, this is 0. This P2 transistor will be on and then ensure that the y is connected to the 1.2 volts.

Now what I am seeing is a level of 0.8 getting converted into a level of 1.2 volts and that is been applied now to the circuits in the V_{ddH} domain. I have given another example here if V input is 0, I will have this is 0, N1 will be off and it will convert, there is an inverter here for where the rail is V_{ddL}. I will have a V_{ddL} line here, V_{ddL} line will ensure that N2 will be on. The N2 will ensure that the y will be grounded, y will be 0, y0 makes this P1 transistor on and x will be now connected to V_{ddH}.

We can have y or x as 1.2 volts, depending on the logic that has been passed here and then we can use this x or y for the V_{ddH} domain, kind of a level converter is used generally for the circuits for the subsystem which is designed for design using a lower rail voltage followed by or connected to the higher rail voltage designs.

(Refer Slide Time: 19:24)



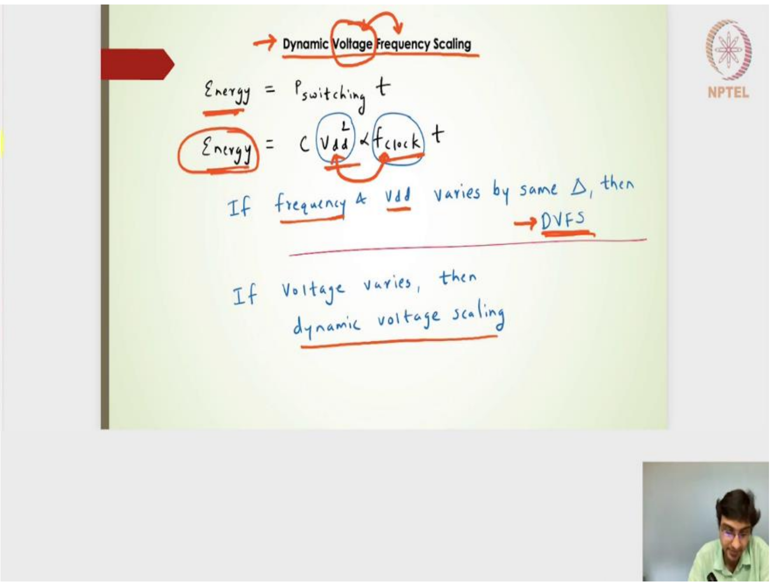
No contention, hence level converter are not required for signal propogating from Vdd_H domain to Vdd_L domain

If I have the other way, if I have a V_{ddH} the higher rail design and the design is connected to the lower rail design. I have a subsystem module 1 which required us to design the circuit in V_{ddH} rail a higher rail voltage and the subsystem module number 2, which required a lower rail voltage. Then what happens?

I mean do we really require the level converter or not. If I have an inverter, a primitive example of inverter in both the case and if I directly connect the inverter without using the level converter let us see what happens. In this case the V_{ddH} if this is 0, I will have a V_{ddH}, the V_{ddH} is passed here the V_{ddH} of let us say we can take an example of 1.2 volts. This could be 1.2 volts and V_{ddL} could be 0.8 volts. If it is 1.2 volts here at the input at the V_{ddL} is 0.8, it is anyways be on off. I will have the N2 to be on and V_{out} to be connected to ground, it will be 0. A 0 is passed here and a 0 is the output of the second inverter. There is no contention here we are getting a 1.2 volts here.

In the earlier case we had a 0.8 volts here at the input of the second inverter. Whereas, the rail was 1.2 volts, here the input is 1.2 volts and a rail is actually 0.8 volts. My V_{gs} value will always be or V_{sd} value in this case will always be less then 0.3 volts, no contention hence level converter are not required for the signals propagating from the higher domain rails to the lower domain rails. Only the rail the level converters are required only when we are going from the lower domain circuit connected to the higher domain rail circuits, hope that is clear.

(Refer Slide Time: 21:18)



Dynamic Voltage Frequency Scaling

$Energy = P_{switching}\, t$

$Energy = C V_{dd}^{2} \alpha f_{clock}\, t$

If frequency & Vdd varies by same Δ, then → DVFS

If Voltage varies, then dynamic voltage scaling

I am introducing altogether a new topic here, this is just a two slide just as to warm up to the next lecture. In this particular slide it talks about the dynamic voltage frequency scaling. This lecture was about the voltage scaling and then how we had seen having a different rail voltages, how it is going to save the power, but at the same time we will have a compromise on the delay and for interfacing between the different rail voltages we had seen how the level converters has to be designed. This is an extension of the voltage scaling and that is why these two slides are kind of included in this particular lecture, but it is not only about the voltage scaling, but it is also about the frequency scaling. If I consider the energy term it is nothing but,

$$\mathrm{Energy} = \mathrm{P}_{\mathrm{switching}}\mathrm{t}$$

In this case the energy I can rewrite it as,

$$\mathrm{Energy} = \mathrm{CV}_{\mathrm{dd}}^{2}\alpha\mathrm{f}_{\mathrm{clock}}\mathrm{t}$$

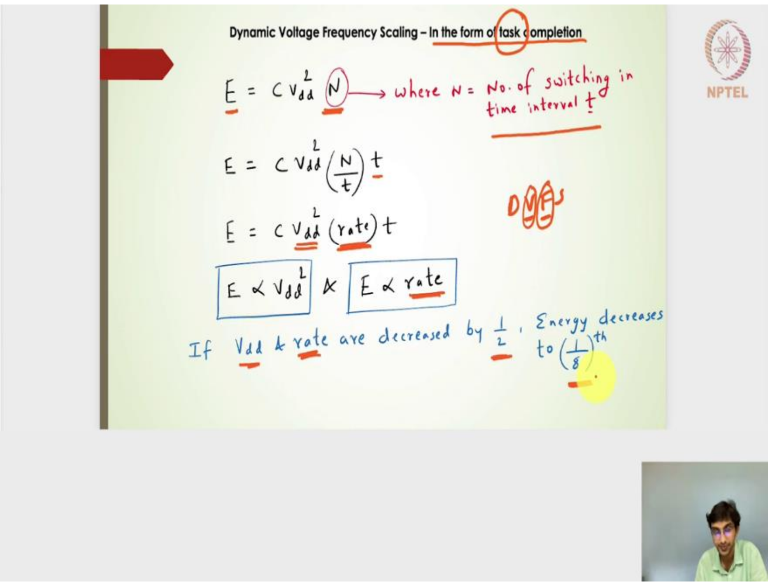
If the frequency of the clock and then the V_{dd} varies by the same Δ, then we say that it is dynamic voltage frequency scaling. What it means is if I can somehow relate this voltage and frequencies, for a particular frequency I will have a particular voltage. A reduction in the frequency will have a reduction in the voltage or an increase in the frequency will have an increase in the voltage, I have some kind of a proportional relationship between the $\mathrm{V}_{\mathrm{dd}}\mathrm{f}_{\mathrm{clock}}$, where both the parameters are decreasing or increasing.

If it is decreasing, I will have a significant savings in the energy. This particular property of frequency in V_{dd} kind of related to each other and where it is increasing and V_{dd} is also increasing, frequency is also increasing or if I have a lower rail voltage then I will also reduce my clock frequency. In that sense that particular architecture is called dynamic voltage frequency scaling.

If in this particular term of the energy term, if only the voltage varies and f_{clock} is independent of the V_{dd} then it is called as a dynamic voltage scaling. Only one voltage parameter is varying or scaling and that is why this particular architecture is called dynamic voltage scaling.

Just to summarize, if the DVFS which is nothing but Dynamic Voltage Frequency Scaling. Where both voltage and frequency are proportionally increasing or decreasing, that is called as DVFS.

(Refer Slide Time: 24:00)



Dynamic Voltage Frequency Scaling – In the form of task completion

$E = C V_{dd}^2 N \longrightarrow$ where N = No. of switching in time interval t

$E = C V_{dd}^2 \left(\frac{N}{t}\right) t$

DVFS

$E = C V_{dd}^2 (\text{rate}) t$

$\boxed{E \propto V_{dd}^2}$ & $\boxed{E \propto \text{rate}}$

If V_{dd} & rate are decreased by $\frac{1}{2}$, Energy decreases to $\left(\frac{1}{8}\right)^{th}$.

The last one, the last slide, in terms of the task completion when we talk about the task completion that what it means is if I have one process, the process or it is given to a particular circuit or as the particular circuit is been selected the output of the circuit when it switches from 0 to 1 that is when the energy is kind of delivered by the V_{dd}.

Any kind of a task I will have the output node to be switching. If I want to find out the energy delivered by the V_{dd} for that particular task, I need to understand what is the number of times, number of switching at that particular output node in the time interval of t. Here the energy delivered by the $\mathrm{V_{dd}}$ for that particular task to get completed is given as,

$$\mathrm{E = CV_{dd}^2 N}$$

What here I mean is, if there is a particular task then that has to be characterized and if it is characterized by the number of times it is switching at the output node from 0 to 1, I can use that to define my energy or to extract the energy that is to be delivered by the $\mathrm{V_{dd}}$ to complete that particular task.

$$\mathrm{E = CV_{dd}^2 \left(\frac{N}{t}\right) t}$$

This $\mathrm{\frac{N}{t}}$ I can say that the number of times it is switching divided by the time of interest. So, that will be given as that could be called or referred to as the rate of switching.

$$\mathrm{E = CV_{dd}^2 (rate) t}$$

$$\mathrm{E \propto V_{dd}^2 \text{ and } E \propto rate}$$

In the previous slide what we had seen was DVFS, where the voltage and frequency we had used the energy in terms of the voltage and in terms of the clock frequency. Here in this particular slide instead of the clock frequency, we are introducing a new term called as the rate of switching. The number of times the output node switches from 0 to 1 and divided by that time the duration of interest, that will give us the rate. The energy term is now directly proportional to the $\mathrm{V_{dd}}$ or the square of $\mathrm{V_{dd}}$ and then directly proportional to the rate of switching. If $\mathrm{V_{dd}}$ and the rate are decreased by half, then I know that the energy decreases to $\mathrm{\left(\frac{1}{8}\right)^{th}}$.