Low Power VLSI Circuits and Systems

Prof. Ajit principal

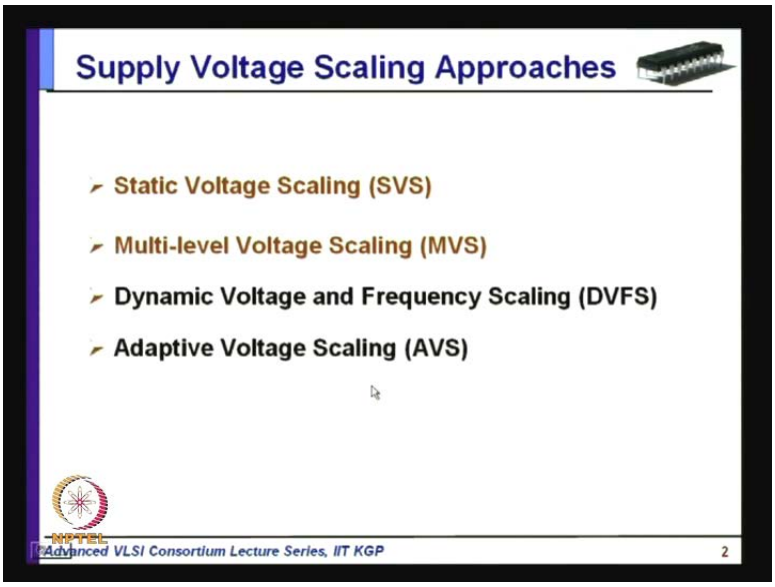
Department of Computer Science and Engineering

Indian Institute of Technology, Kharagpur

Lecture No. # 25

Supply Voltage Scaling – IV

(Refer Slide Time: 00:29)

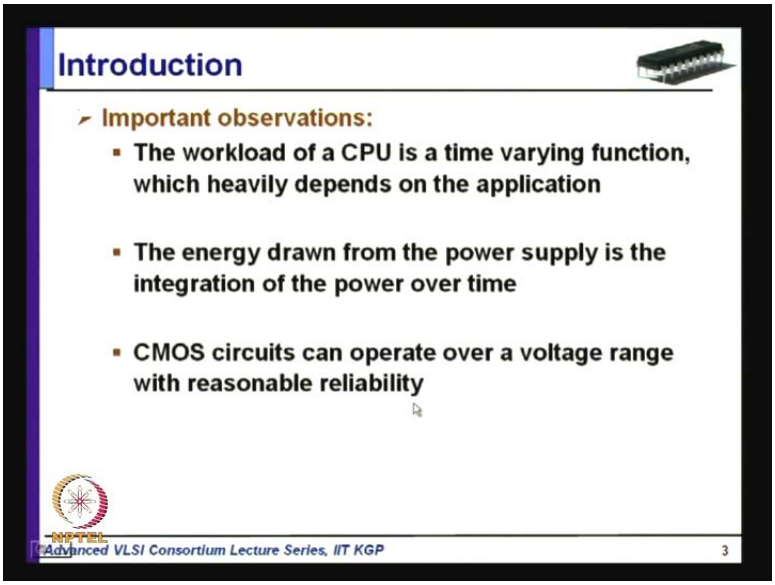


Hello and welcome to today's lecture on supply voltage scaling. This is the fourth and the last lecture on this topic. In the last lectures, we have discussed Static Voltage Scaling and Multi-level Voltage Scaling. As we have seen in both the cases, the voltages are assigned at designed time; in case of Static Voltage Scaling the supply voltage may be scaled to some lower level at design time; and then the entire circuit or module is assigned a lower level voltage and that remains fixed duration during it is normal mode of operation.

Similarly, in Multi-level Voltage Scaling, multiple voltage levels are assigned to different modules or clusters, and those voltages remain fixed during normal mode of operation. They do not change, when they are in use. Now, we shall discuss another technique in which the supply voltage is dynamically changed at run time; and today we shall discuss Dynamic Voltage and Frequency Scaling, and special case of that is known

as Adaptive Voltage Scaling, both of these two techniques, we shall consider in this lecture. Now, this Dynamic Voltage and Frequency Scaling is based on certain observation.

(Refer Slide Time: 01:45)



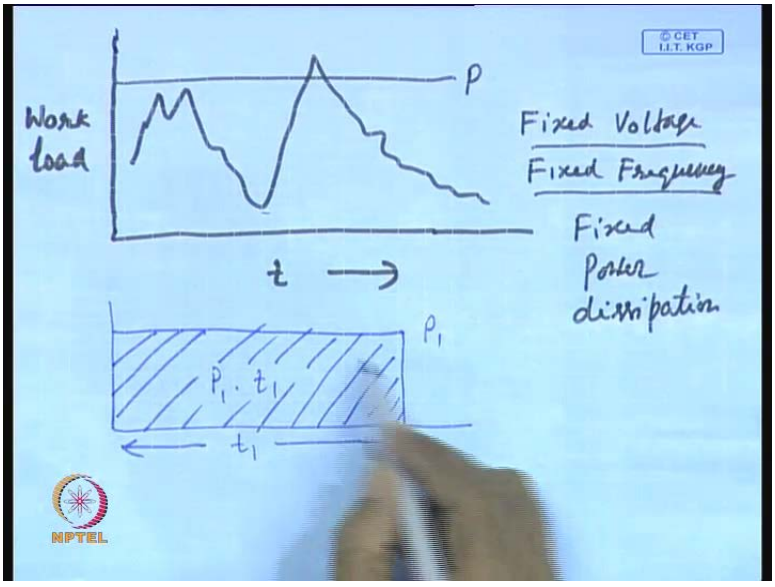
And these are the important observations; in fact the genesis of this Dynamic Voltage and Frequency Scaling will come from these observations. Number 1 is the workload of a CPU is a time varying function, which heavily depends on the application. For example, if you consider the workload of a CPU, we will find that with time it varies that means, it may somewhat like this; that means it is a time varying function and not only that, it will also depend on the class of the system.

For example, this variation of the workload will be different for a server, I mean will not be same as that of a desktop and it will not be same as that of a laptop. That means the workload is varying with time, not only it is varying with time it is depending on the type of application, server application or desktop application like that. And whenever this is a varying function of time that means this is the workload; workload is varying over time, whenever this is happening then as you see although the workload is varying over time, you are maintaining a Fixed Voltage, and also you are maintaining a Fixed Frequency in a during normal uses.

So, whenever you do so, there is a fixed power dissipation that means, the power dissipation as you know is dependent on the supply voltage, it is dependent on the frequency of operation and of course, it also depends on the switching activity and other things. That means the power dissipation may remain fixed; so this is the power dissipation p that remains fixed irrespective of the fact that the workload is changing. You may have noticed that in your desktop or laptop if you look at this CPU utilization, you will find that sometimes it is 3 percent, sometimes it is 20 percent, sometimes it is 50 percent depending on the application you are running.

So, when you are doing video processing say MPEG, JPEG and all these things, then the CPU utilization is more; on the other hand when you are doing word processing, sending e-mail, then you will find that CPU utilization is very poor. Now, whenever this is the situation, can we really exploit this particular phenomenon to reduce power dissipation? That is the basic idea behind this Dynamic Voltage and Frequency Scaling. Now let us look at other observations. The energy drawn from the power supply is the integration of the power over time. This is particularly important in the contest of battery operated system as I have already told.

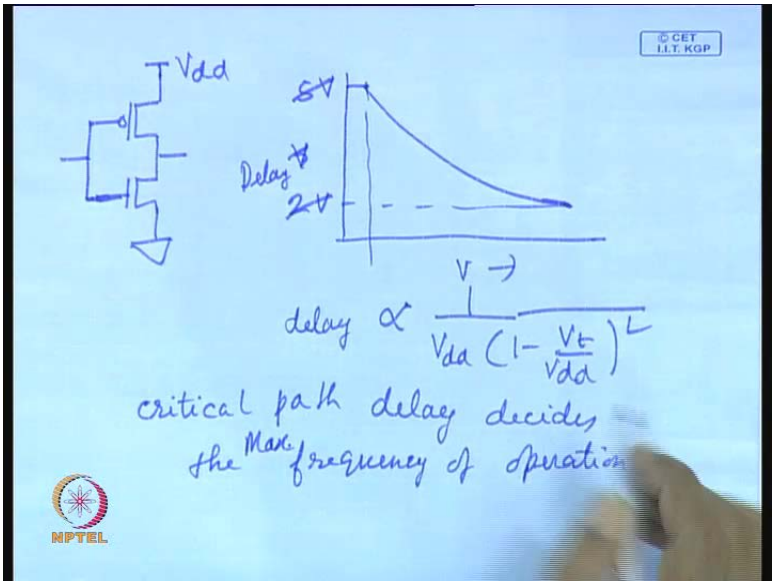
(Refer Slide Time: 02:09)



If power dissipation is say P or P1 and if the power dissipation occurs over a period of time t1 than energy that is being drawn is equal to P1 into t1 that is the area in covered by this rectangle, that means this is the energy drawn for performing this computation

say computation takes time t1 and during which the power dissipation is taking place is P1. So, for a battery operates system if we can use power dissipation over time then we can save energy that I shall discuss how it can be done.

(Refer Slide Time: 05:58)



Another important observation is that CMOS circuits can operate over a voltage range with reasonable reliability that means, if you are already familiar with CMOS circuits we have seen that in a particular CMOS circuit, you apply supply voltage Vdd let us consider the simplest example the inverter if we if this is the circuit of an inverter, you see this Vdd if we change even in the circuit will operate over a certain range not that it will work from 0 volt or say it will work at a very high voltage that may not be true, but there is a supply voltage range over which monotonically the voltage range, it will work lively and not only that there is a region of operation for frequency increases monotonically over voltage. That means, if we consider say voltage v it can be shown that over a range voltage over a range of say may be 5 volt or may be this is 2 volt this circuit can operate and in this particular case as we can see whenever you are reducing the supply voltage the voltage is reduced, then you will find that the power dissipation will be reduced and now over this range it can work and the whatever the frequency.

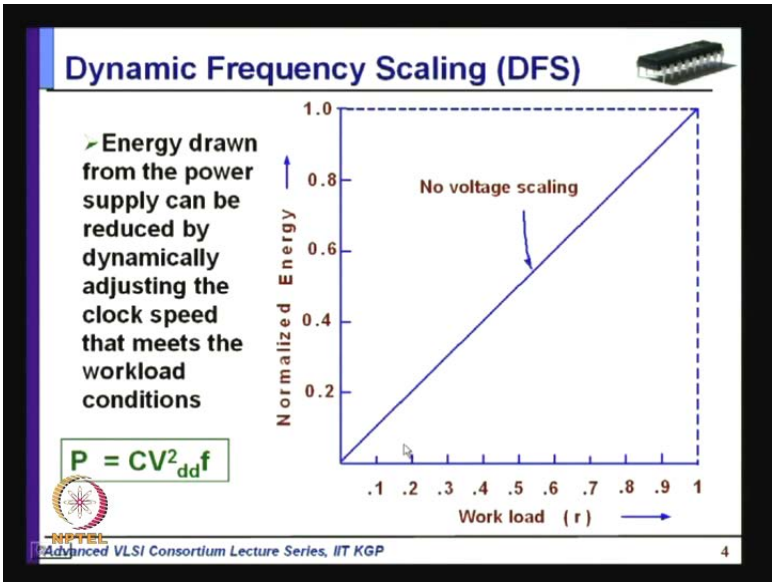
Now, as we know there is a relationship between delay of a circuit and the voltage and that relationship is represented by this 1 minus Vt by Vdd square. So, that means delay is proportional to 1 by Vdd into 1 minus Vt by Vdd, so delay increases linearly as it is

shown here that means if we plot voltage in this direction and instead of writing voltage here say delay will reduce as you increase the voltage.

So, there is a linear relationship like this, not linear means it is kind of quadratic but it is monotonic that means monotonically delay increases as supply voltage increase and the relationship is given by this expression. So, it is a monotonic means as supply voltage is reduced delay increases monotonically it increases it does not decrease after some points so there is a range over which this happens. So, in this particular range how this is or why this is happening, because as you know delay is changing so, the critical path delay decides maximum frequency of operation of this circuit that means, what does it means that you can really as you reduce the supply voltage.

Then it will operate at a lower frequency so this based on this idea we can develop circuit such that the power dissipation can be reduced by controlling the voltage and frequency dynamically as we shall see how it can be done. So, based on these observations the technique of Dynamic Voltage and Frequency Scaling has been developed.

(Refer Slide Time: 09:37)



First we shall consider a simpler situation where the frequency scaled down as the workload reduces. So, in this particular case, as you can see, you were reducing the frequency as the workload is reduced. So, there is a linear relationship between the energy and the workload; so we can show you with the help of another plot. Let us

consider a situation say this is the case where workload is 100 percent and power dissipation is P1 that is corresponds to workload is equal to 100 percent.

So, what is the… If time is t1, the energy drawn for this computation will be equal to P1 if into t1 with 100 percent workload. Now, let us consider the situation, where the workload becomes 50 percent, now if you do not change the frequency of operation, power dissipation will remain same however the computation will be complete within 50 percent of the time that means in time t2, which is 50 percent of t1 that means, t2 is equal to t1 by 2, then power dissipation will be power dissipation is remaining same, but time is decreasing. So, in this particular case as we can see energy drawn is reduced.

It is becoming half, because P1 into t2 that that is reduced however power dissipation remains same during this period t2 now you can you can in this particular case neither the voltage is scaled nor the frequency is scaled. So, during this period power dissipation is taking place and you switch off during the remaining part or you may not if you do not switch it off then power dissipation will continue to take place in the remaining part, but wiser technique is to use Frequency Scaling as it is shown in this particular case even if you do not do voltage scaling if you simply reduce the frequency of operation and make it half then what will happen?

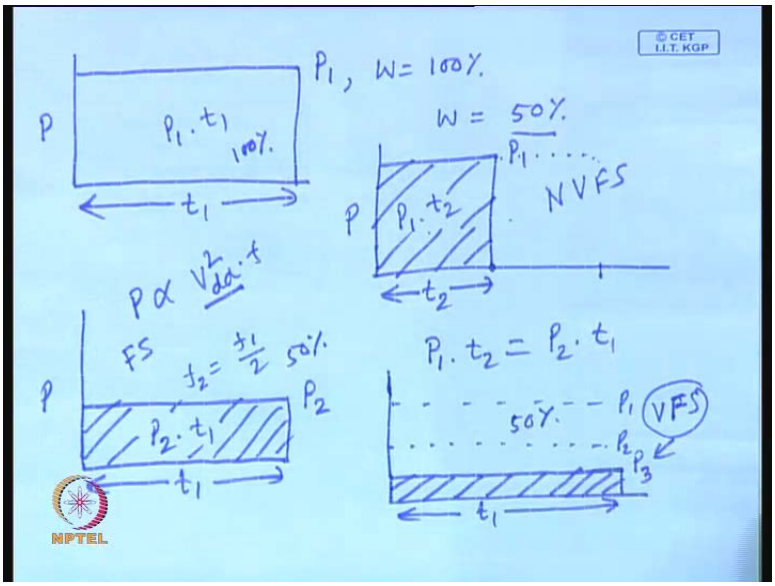
The computation time will become half, but it will continue to take place for the entire time t1, that means in this case power dissipation will reduce and this is power and this direction power and time. So, in this case you can see P2 into t1 is same as that of the previous case, that means energy drawn is reduced is same as this one the previous case that means P1 into t2 is equal to P2 into t1 energy drawn is remaining same however the power dissipation is reduced this has definitely has impact on the packaging and cooling of the system.

That means, if the power dissipation is less the lesser heat dissipation takes place and packaging and cooling will be cooling cost can be reduced. So without voltage scaling simply by Frequency Scaling, you can achieve lower power dissipation energy remaining same whenever the workload is changing. Now, let us consider the case, where you can perform voltage and Frequency Scaling that means in this case, as you can see since the frequency f2 is equal to f1 by 2 in this particular case that is the reason why it is taking longer time.

So, you have scaled down the frequency to f1 by 2 and that is why it has taken double time of that 1 and power dissipation has reduced, because P is proportional to Vdd square into f. So, here it is f2 and there it was f1. So, power dissipation has reduced because voltage was fixed, but now what you can do, as you know if the frequency is less then this circuit can operate at a lower voltage. As we have discussed few minutes back that the observation is as the supply voltage is reduced, this circuit will operate at a lower frequency or can operate at a lower frequency. So, in this case there is scope for scaling down the voltage. So, what you have done the time is remaining same, time is not changing, but instead of P1.

It is now P3 and which is definitely less than P2 or less than P1. So, you see that there is a significant reduction in power dissipation not only power dissipation, energy is drawn from the battery also much less. So, in the previous two cases energy was same, because only Frequency Scaling was known power dissipation was less, but energy drawn from the battery was remaining same, but whenever you do voltage and Frequency Scaling Voltage and Frequency Scaling as you do as you have done here then not only the power dissipation reduces, but energy drawn from the power sources also significantly reduce that means this is very important in the contest of battery operating system.

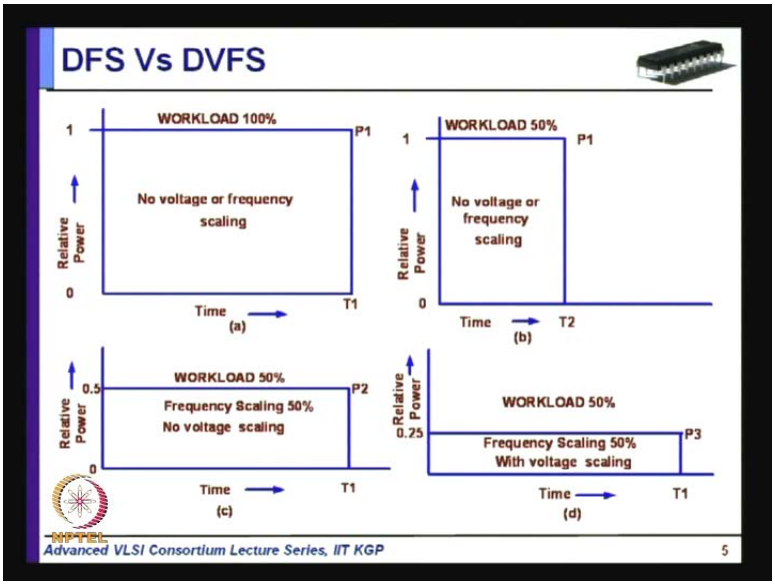
(Refer Slide Time: 10:10)



Here energy is significantly reduced and as you can see this is voltage and Frequency Scaling. This is only Frequency Scaling and in this particular cases neither no voltage

nor Frequency Scaling and here also no voltage and Frequency Scaling, but here the workload is 100 percent, here the workload is 50 percent, here the workload is 50 percent and here also the workload is 50 percent. So, this is based on this. Now we shall develop

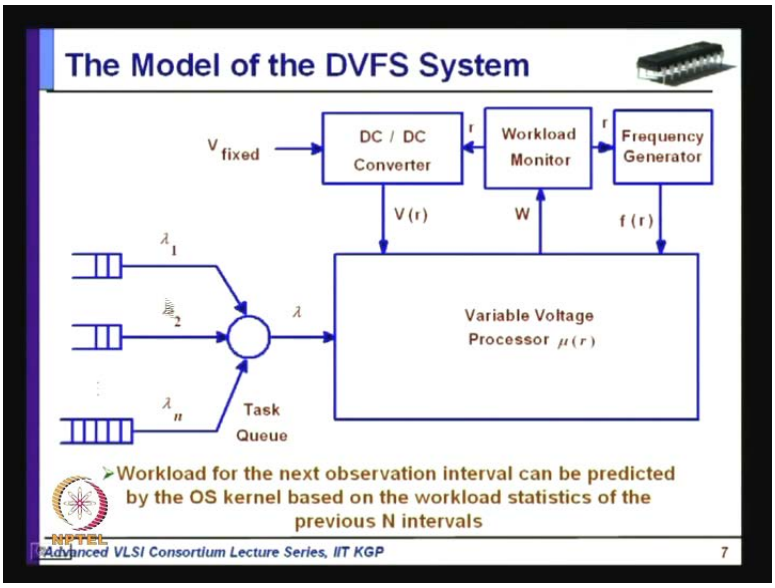
(Refer Slide Time: 16:07)



Now let us consider the realization of a circuit and as I have mentioned energy drawn can be reduced by dynamically adjusting both voltage and frequency that meets the workload condition. This is the basic idea of Dynamic Voltage and Frequency Scaling that means during run time, you will find out at what frequency this circuit should operate to meet the workload requirement. Then for that part to operate at that frequency, what supply voltage is required that you can find out. So, you can assign you can the circuit the CPU can work at a lesser voltage and later lesser frequency there by reducing the power dissipation. This curve shows how the energy reduction takes place. This is the case where there is no voltage scaling, but here this is the ideal voltage and Frequency Scaling.

So, I have used the term ideal, later on we shall see it may deviate from this ideal, because of some other reasons, because of DC to DC converter and other thing that we can consider later. But you see there is a significant reduction in energy so there is a cubic reduction in power dissipation and quadratic reduction in energy that will take place by using Dynamic Voltage and Frequency Scaling. How can it be realized?

(Refer Slide Time: 17:37)



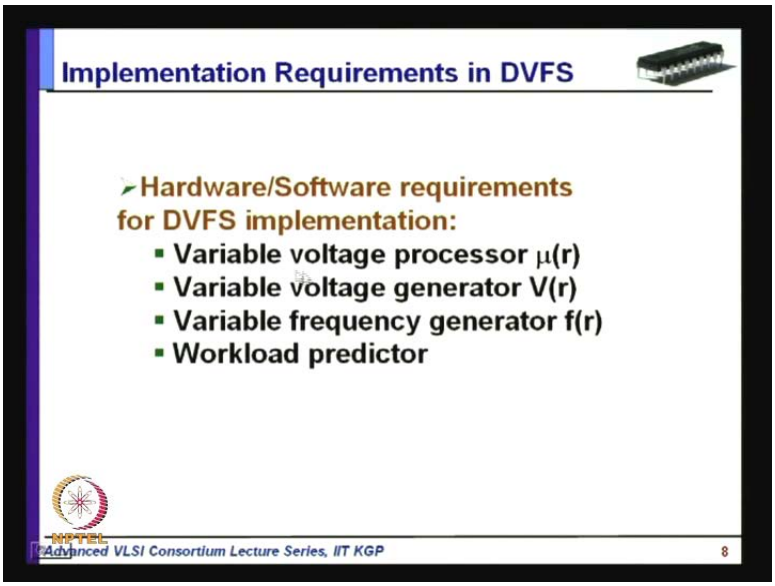
This is the model of the Dynamic Voltage and Frequency Scaling system. So, here it shows you, have got a processor CPU microprocessor mu r, which can operate over a voltage range that is why the name is variable voltage processor. So, this particular processor can operate over a voltage range. And you can see the task input is coming from a number of sources and different sources, this is the task q coming from different sources, this is the rate of arrival average rate of arrival of different I mean, the task from different sources lambda 1 is the average rate of arrival from source 1 lambda 2 is the average rate of arrival from source 2, and in this way lambda n is the average rate of arrival of task from source n. Out of which you know, although the tasks are coming to from different sources, there is a scheduler as you know, which is part of the operating system, there is a task scheduler; the task scheduler will issue tasks, which can be executed by the CPU, so lambda is the rate at which tasks are issued to this variable voltage processor.

Now the variable voltage processor has got a workload monitor. What is the job of this workload monitor? It receive, it can find out what is the present workload based on the task which have been executed by these processor and this workload monitor what it will do not only it will keep on receiving the information about the workload of the present workload it will estimate what is the workload required in the next time slot.

A kind of a prediction, workload prediction it will do and based on that prediction it will send some signal to a DC to DC converter that means, if the workload requirement is less that it this DC to DC converter will generate a lesser voltage although it will receive a fixed voltage it will generate a variable voltage V r and this variable voltage is dependent on this workload and of course, this is related to the frequency. So, there is a frequency generated variable frequency generator the workload monitor will also send information about the predicted workload or to the frequency generator and then it will generate a frequency and that frequency will be applied to the variable voltage processor. So, this voltage is applied, this frequency is applied and to sustain that frequency a particular voltage that is required that is being applied by the generated by this DC to DC converter.

So, this is the basic model, which is being shown and what is written here workload for the next observation interval can be predicted by the OS kernel based on the workload statistics of the previous N intervals. That is the job of these workload monitor. So, it can find out, it can predict the workload in the next time interval. This is the basic model.

(Refer Slide Time: 21:40)

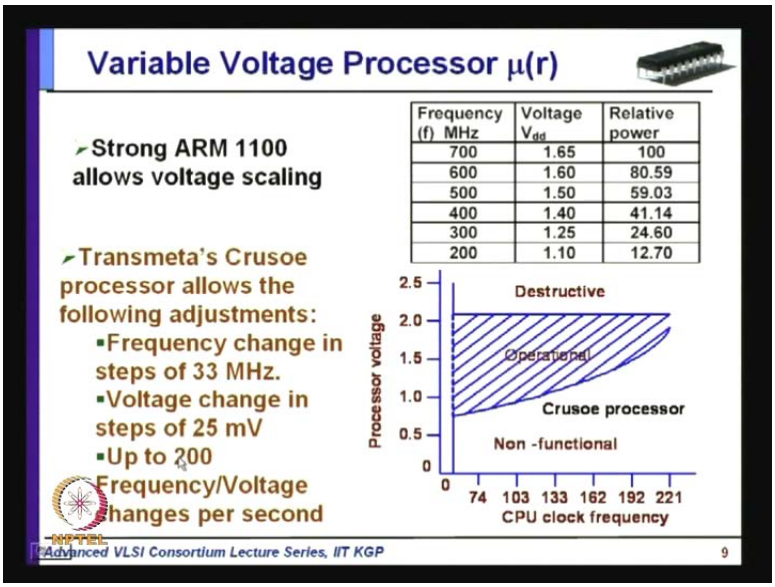


Based on this basic model we can identify what is the hardware and software requirement for implementing Dynamic Voltage and Frequency Scaling system. First of all as we have seen we require a variable voltage processor CPU, because you know there are many processors, which will operate at a fixed voltage which are not designed

to operate over a over a range of voltage, but there are processors which can operate over a voltage range so that will be the first requirement. Second requirement is variable voltage generator so V because you know, this to operate at different frequency.

You will require different voltage that has to be generated by hardware which is known as variable voltage generator you will require a variable frequency generator as you as you have seen in the basic model f r which will keep on generating variable frequencies and last, but not the least the workload predictor which is the part of the workload monitor and may be part of the operating system kernel. So, a part of the operating system kernel can perform the task of workload prediction.

(Refer Slide Time: 22:25)



Now, coming to the first requirement, that is variable voltage processor. There are processors available, one is your strong ARM 11 00. This allows voltage scaling as you can see it can operate over a voltage range 1.1 to 1.65 and accordingly it can operate at different frequency that means when the supply voltage is 1.65 volt the operating frequency is 700 megahertz and this is the relative power 100.

So, this is normalized with respect to this 1 the other power the other power dissipation and then when the frequency is 600 megahertz then the voltage optimum voltage or minimum voltage that is required is 1.60 and relative power dissipation is 80.59 of the previous one. And if the frequency is 500 megahertz it can operate at 1.50 volt and the relative power dissipation is 59 or 59 percent you can say of the first case and in this

way. It can go up to 200 mega-hertz and the voltage required is 1.10 volt and you can see the relative power is 12.87 percent.

So, this you can see one-eighth of the first case that means if the frequency is varied from 700 megahertz to 200 megahertz and accordingly the supply voltage is varied from 1.65 to 1.0 than power dissipation can vary from 100 to 12.7. So, that can be significant reduction in power dissipation and obviously the energy drawn from the battery source will be there, will be (( )) reduction drawn from the battery.

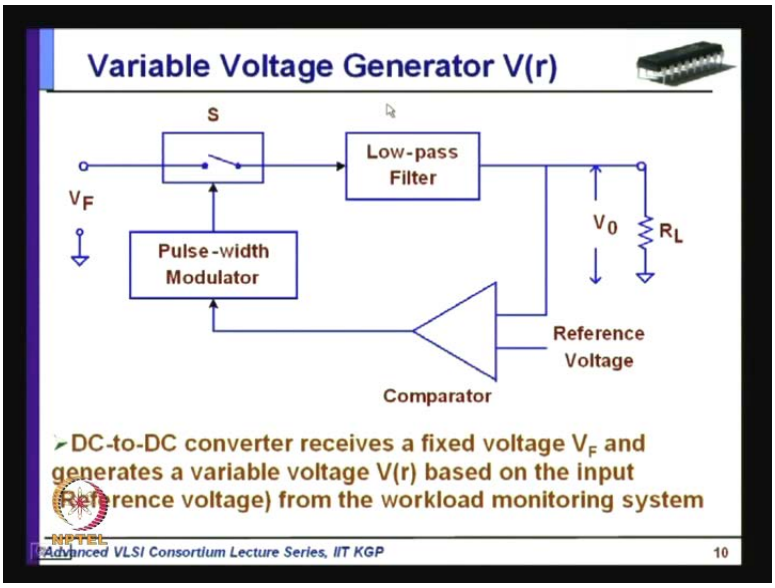
Another processor is there, that is Transmeta's Crusoe processor there is a manufacturer, transmitter; they have developed a processor which is very innovative Crusoe processor that allows the following adjustments. It allows frequency change in steps of 33 megahertz, it also allows voltage change in the steps of 25 mili-volts and there can be 200 frequency voltage changes per 200 frequency voltage changes per second.

Whenever you go from one particular frequency to another frequency model operation you cannot change very quickly, the reason for that is DC to DC converter and the frequency generator will take some time to settle down and that is the reason why the rate at which you can make this change in the frequency and voltage cannot be very fast, but you can see that 200 voltage changes per second is quite reasonable and can be suitable for many applications.

So, in this particular case as you can see this Crusoe processor can operate over the range of 2.2 to roughly 0.7, 0.7 to 2.2 volts .You can see this portion the shaded portion is operational range and if the voltage is above 2.2 it is the destructive region that means the power dissipation will be very high and that may lead to destruction of the device and similarly, as you reduce the voltage you can see the frequency of operation is also reducing. So, whenever it has to that minimum maximum frequency decided by this particular (( )).

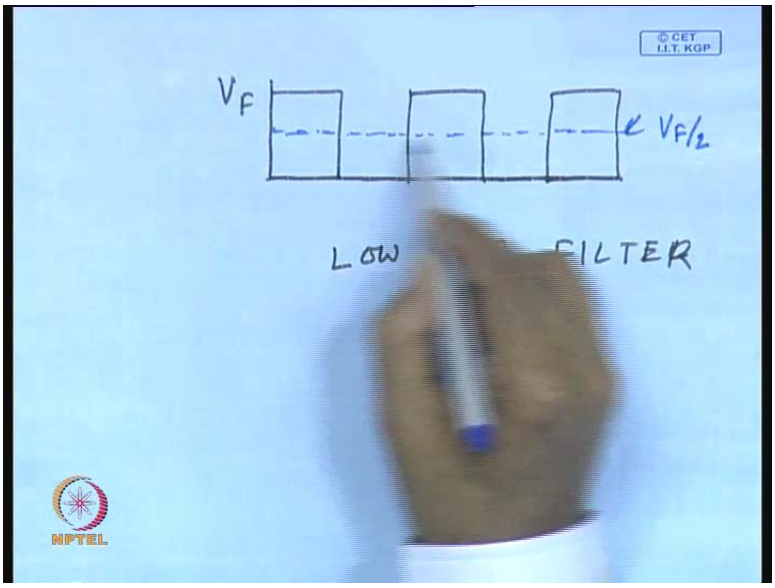
So, as the voltage is reduced from 2.2 the frequency operation will reduce from 221 megahertz to close to 70 megahertz. So, you can see this can also operate over a large voltage range and also over a large frequency range and you can have large number of voltage frequency appears. Although I mean it is shown continuous, later on we shall see, we will normally use a limited number of voltage frequency pairs. It will not change continuously. Some discrete voltage frequency pairs are used for operation of the circuit.

(Refer Slide Time: 27:43)



So, this is about the variable voltage processor so in fact this ARM processor is now available in the form of a I p core and you can use it in many embedded applications and that allows Dynamic Voltage and Frequency Scaling. Coming to variable voltage generator you can see here this variable voltage generation can be done by DC to DC converter. DC to DC converter is commonly used in on all modern digital equipments. You can see it receives a fix voltage VF is received and there is a pulse-width modulator which controls a switch S. This switch although is a shown as a mechanical switch it can be it can be an electronic switch. Normally, it is realized by device known as IGBT insulated gate by polar transistors. So, IGBT is commonly used as a switch in this DC to DC converter. And so what will happen? How does it control the voltage? Here what it does?

(Refer Slide Time: 28:48)



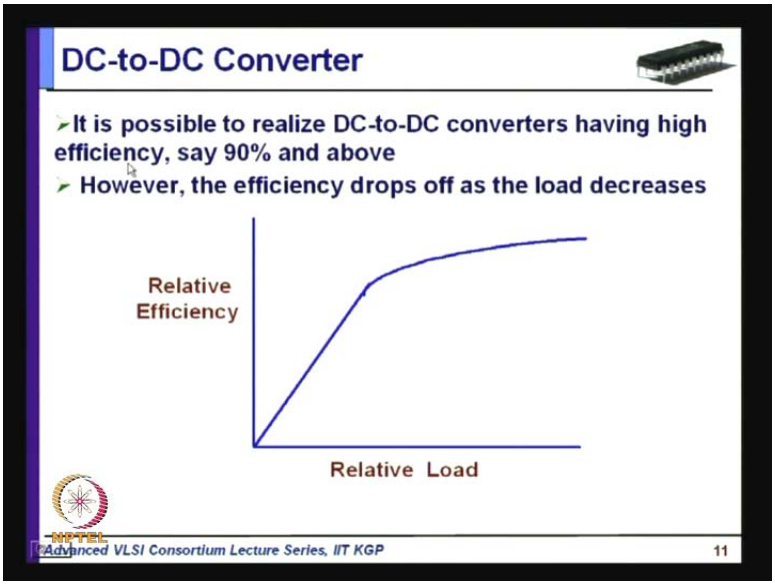
Suppose this is the fixed voltage VF and the duty cycle of this at which the switching is done by this particular device is controlled by this Pulse-width Modulator. Let us assume the duty cycle is 50 percent. So, in this case that means, 50 percent of the time this switch will be closed and 50 percent of the time the switch will be off. So, in this way the switch will operate. And this duty cycle is dependent on the control input that we shall see; so there is a control input here you can say Reference Voltage; depending on this reference voltage, this pulse-width modulator adjust the duty cycle.

So, this output of this switch is like a signal like this. So, a kind of what train of pulses; now this is applied with Low-pass Filter. So, as you apply to a low pass filter, then what will happen? The output of the low pass filter will remove the high frequency component, and you can get only the DC component, so you can see you will get a voltage, which is a DC voltage, and which will be equal to VF by 2; since the duty cycle is 50 percent, it will generate a voltage VF by 2; that is the basic principle of this DC to DC converter. And here you can see this Low-pass Filter output is applied to a comparator and this is compared with a reference voltage depending on this comparison it will generate a signal to the Pulse-width Modulator and accordingly it will increase or decrease the duty cycles depending on the output is more or less with respect to the reference voltage this is the basic principle of this variable voltage generator.

You will get a voltage VO which will vary from may be from 0 may not be exactly 0, but very close to 0 to VF and that is the Reference Voltage VF and that can be applied to the load or the circuit under operation, and this kind of DC to DC lot of research work is going on efficient DC to DC converter. Nowadays technique known as buck converter is used which provides very high efficiency may be up to 90 percent efficiency is achievable from Buck converters. Buck converters are nothing, but special type of DC to DC to converter and people are also trying to reduce the size the DC to DC converter so that it can be put on chip.

Obviously all the components cannot be put on chip, but the electronic part can be put on chip, but a capacitor and some inductor and some other components has to be connected externally. I am not going into the design of a DC to DC converter and though electric engineering students will find in their department research work is going on for the development of efficient DC to DC converter. So, but this is the basic principle I have explained. And this type of DC to DC converter can be used to generate a variable voltage variable supply voltage which is required which is one of the primary requirements for implementing Dynamic Voltage and Frequency Scaling system.

(Refer Slide Time: 32:29)

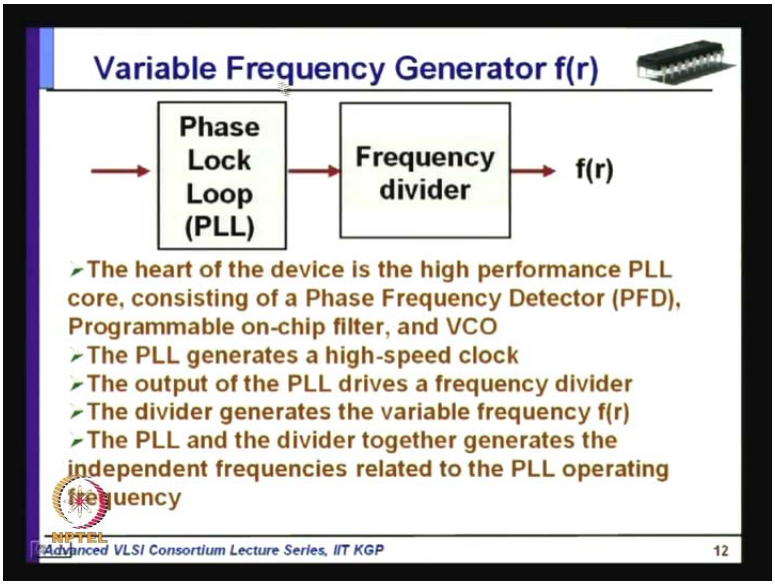


Now, coming to one very important aspect of DC to DC converter, as I have mentioned, it is possible to realize DC to DC converters having high efficiently say 90 percent and above. Unfortunately the efficiency drops off as load decreases; that means, when the

load is high 80 percent, 90 percent or 100 percent, then the efficiency of the DC converter is high, because this DC to DC converter is a electronic circuit, which has some power dissipation; and whenever the duty cycle is small, then the efficiency of the DC to DC converter reduces as you can see the relative efficiency degrades as the workload is reduced; so the efficiency drops off as load decreases.

So, you have to this I am raising this point, because whenever we will be considering the efficiency of Dynamic Voltage and Frequency Scaling system, you have to not only take into consideration the efficiently of the processor power dissipation that is taking place in the processor, but also the power dissipation that is taking place in the DC to DC converter, because DC to DC converter is now part of your system; and the power dissipation that is taking place in DC to DC converter has to be taken in to consideration. So, this point you have to keep in mind, when you are using DC to DC converter with a small workload.

(Refer Slide Time: 34:05)



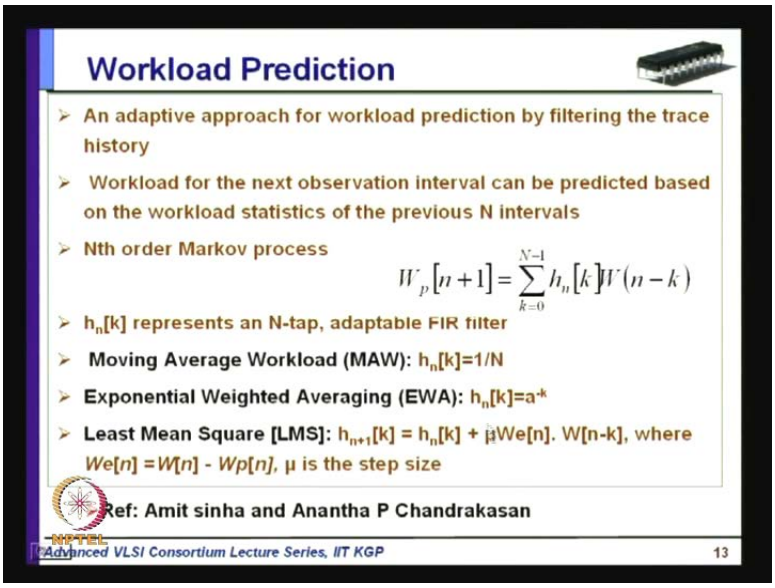
Coming to another important component, that is your Variable Frequency Generator. Variable Frequency Generator can be generated with the help of Phase Lock Loop. This is the traditional approach for generating variable frequency as you can see here there is a phase lock loop and the heart of this device is the high performance phase lock loop core consisting of a Phase Frequency Detector PFD and Programmable on-chip filter and a VCO, this is part of this part of this Phase Lock Loop.

So, inside this Phase Lock Loop, we have three important components, one is phase frequency detector, one is Programmable on-chip filter, another is VCO that is voltage control oscillator. It is here; so it generates a high frequency then that frequency is divided with the help of a frequency divider to generate the desired frequency. So you see Frequency Scaling is done in two steps; number one step is you can change the frequency of the Phase Lock Loop, and also you can change the division ratio of this frequency divider.

The divider generates ultimately you are getting the variable frequency from the output of the frequency divider, but together the Phase Lock Loop and the divider together generates the independent frequencies related to the PLL operating frequency that means, PLL can operate at different frequency say f1 f2 and so on. Then f1 can be divided by a factor d1 and that ultimately is the f for the (( )) that is being generated that means, the frequency generation is done in two steps; later on again we shall discuss about it. This is how variable frequency can be generated.

Coming to the workload prediction, as I said this is also very important thing, an adaptive approach for workload prediction by filtering the trace history. So, this is a technique, which has been developed proposed by Amit Sinha and Anantha P Chandrakasan of MIT. This is the approach they have provided; so essentially, these workload prediction is somewhat similar to you know the way the filters works, digital filter works. So, workload for the next observation interval can be predicted based on the workload statistics of the previous N intervals that means, it is a n Nth order Markov process.

(Refer Slide Time: 36:00)



So, whenever you say it is a Nth order Markov process that means it present output present the presents status is dependent on the past N status so if you know the workload in the previous N time intervals you can predict the workload of the present interval that is the basic idea of this workload prediction and h and k represents an N-tap adaptable FIR filter. So, it can be generated with the help of FIR filter and there are three basic approaches. I shall discuss, one is moving average workload MAW, where this h and k is equal to 1 by M. You are finding out the workload say W p n plus 1, which is equal to 1 by N of say W 1 plus W 2 in this way W n that means, what you are doing? You are taking the average of these previous n workloads.

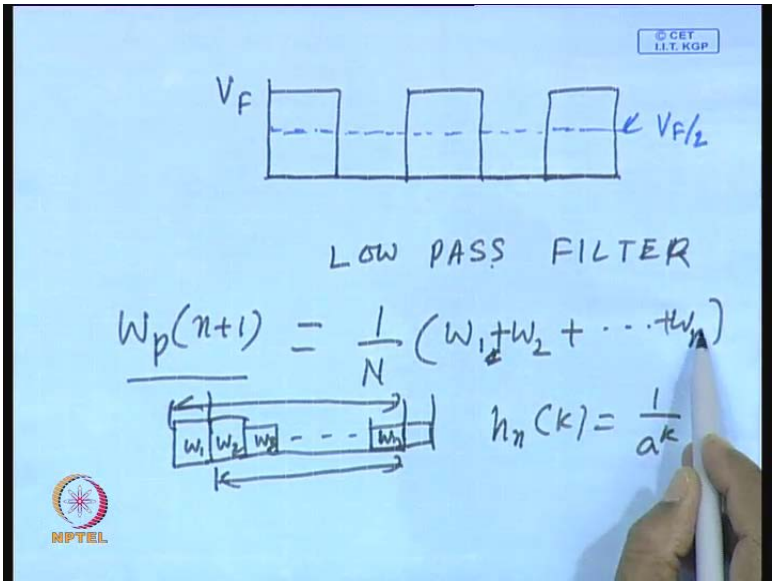
This is how you are doing the prediction that means, the workload in the n plus 1th instant is the average of the workload of the previous N time intervals, in fact that is what you are generating in this particular expression. That means in this case h n is workload prediction h n function is equal to 1 by N, and this is the summation of the w k n minus k that is n 1 W 1, W 2 up to W n. So, this is the first case moving average workload. Why it is called moving average? Because suppose this was the workload during first time interval W 1, and next time interval it was W 2, and next time interval it was W 3. So, in this way workload say W n.

The workload during this period can be predicted from this then workload for this next period will be predicted from this window that means, first you are considering this

window next you are considering this window that means, this window is moving from as the time is progressing and that is how you are keeping on taking the average of the previous n workload histories and that is you are doing the prediction. Now, this particular workload prediction works very well, when the workload changes very slowly.

So, when the workload changes very slowly, since it is changing very slowly increasing or decreasing whatever it may be if you take the average of the previous n values, it will become very close to the next workload, but whenever it changing little fast, then of course, it will not work very well. In such a situation, what you have to do? You have to give more importance to the you know that not the past workloads increasingly more importance to the workloads, which are nearer or you know closer to the present workload.

(Refer Slide Time: 37:35)



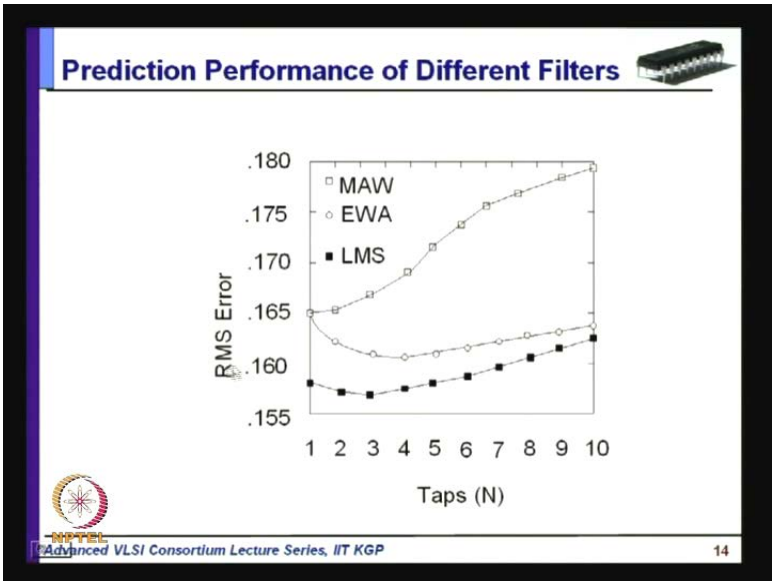
So, that is what is being done in that Exponential Weighted Averaging workload so here what you are doing here the. Here you are multiplying that h n k, this function is equal to 1 by a k. So, this you are multiplying this workload W1 with a factor 1 by a k and this value is large, this 1by a k is a k value of k is large whenever you know W is equal to 1 and for W is equal to n it is very small that means, you are dividing W1 by a large number and Wn by a small number then taking the sum of that.

And that is how you are giving more importance to the most recent history and the most past history you are giving lesser importance. So, this is the basic idea behind this

Exponential Weighted Averaging, where the workload function is equal to 1 by a to the power k. Another important workload prediction function is the Least Mean Square-LMS here h n plus 1 k is equal to h n k plus mu into W e here you were computing the error W e n is the error, W n k into W k n minus where W e n is equal to W n minus W p n so every time what you are doing the predicted workload and actual workload in a particular time slot that difference you are computing and that difference you are using for the computation of the next workload.

So, here the function is you know h this h n this function is h n plus 1 k is equal to h n k plus mu W e n where mu W e n is equal to W n minus W p n and mu is the step size of course, you can choose the step size this value constant and that constant you can adjust to, so that this work this error is less. So, this is how you can these are the three different work functions which you can use and here essentially you are using a N-tap adaptable FIR filter those N-taps are essentially storing the workload function of the previous n time slots. So, this is how workload prediction can be done and you can refer to if you want to learn about it more you can you can get the get it from the paper by Anantha P Chandrakasan.

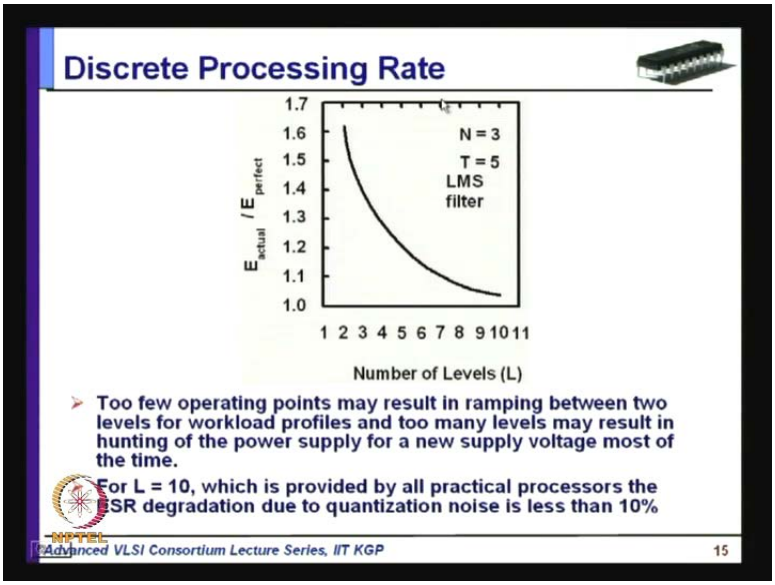
(Refer Slide Time: 43:00)



And this is the prediction performance of different filters is shown here. As you can see the RMS error root means square error is minimum for this LMS approach, this l m s approach. I have showed means least means square approach, where the error is

minimum, error is maximum for this moving average window and this exponential weighted averaging gives you lesser than moving average window, but this LMS definitely gives you much less error compare to the other two schemes.

(Refer Slide Time: 43:50)



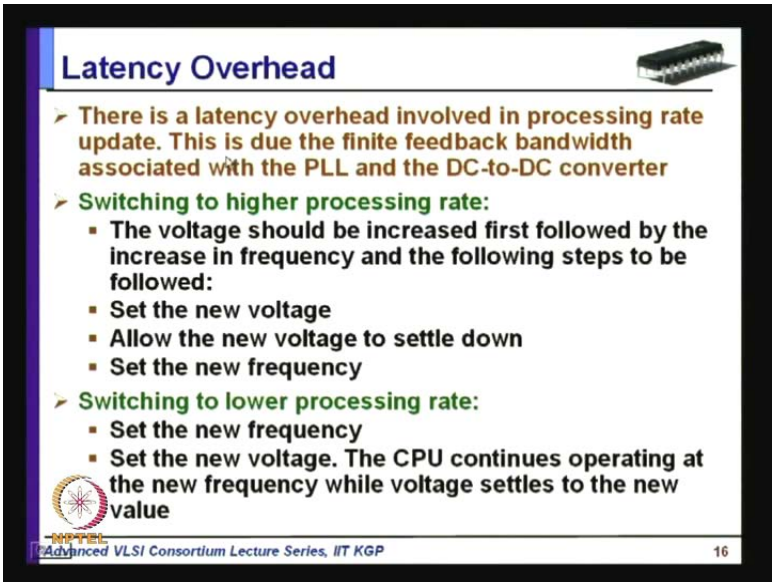
So, coming to another very important design parameter that is the discrete processing rate we have seen that we will be using a number of voltage and frequency error pairs. Question is, how many 2, 3, 4 or 100 or 200? Here the number can be large. So, depending on how many voltages you are using, there is a kind of quantization is taking place. So, you may find out that your workload requirement is W 1, but because of quantization error because we have got only two workloads, I mean two different voltages in that case you have to use either V1 or V2 so error will be more. So, in this way some simulation has been carried out to check.

How many voltage frequency pairs will give you reasonably good performance 2 or 3 or 4 or 5. So, here is the particular impact of this Discrete Processing Rate is given here you can have number of levels can vary from 2 to 10. So, in this case that means the voltage frequency pair that has that can be used is varying from 2 to 10. So, Whenever too few frequency points, I mean say you are using 2 or 3 in such a case what can happen too few frequency points may result in ramping between two levels for workload profiles and too many levels may result in hunting of the power supply for a new supply voltage most of the time. So, in these so, there are a trade up, trade up between two small and two large

you have to choose an optimal number such that these kind of problem does not arise. So, it has been found that for L is equal to 10, that means number of voltage level is voltage level 10 or in other words number of voltage frequency appear is equal to 10 you get roughly, you know I mean the accuracy is very good and almost 90 percent.

For example, for L is equal to 10 which is provided by all practical processors the ESRdegradation due to quantization, ratio quantization noise is less than 10 percent — that means if you increase beyond 10 say if you use 15, 20, 30 obviously the error will be less, but even with 10 voltage frequency pair the accuracy is, I mean the efficiency that you are getting, the reduction that you are getting is close to 90 percent than the ideal. So, this is the Discrete Processing Rates that you have discussed.

(Refer Slide Time: 47:04)



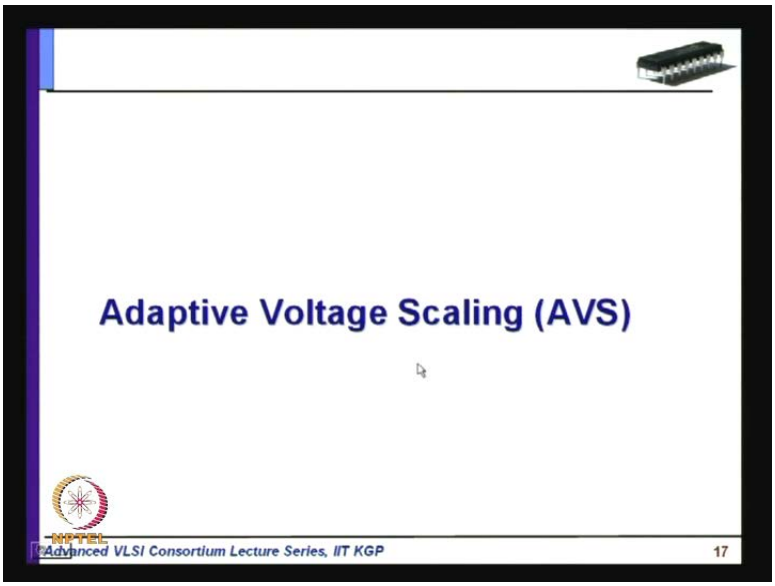
Another important parameter is latency overhead. As I have already told there is a latency overhead involved in processing rate update. This is due to finite feedback bandwidth associated with the Phase Lock Loop and the DC to DC converter, because this Phase Lock Loop and DC to DC converter these are analogue systems.

So, they will take time to settle down so even if you apply some signal for a particular voltage for a particular frequency the new voltage and new frequency will be available after certain time, that is why you will have some latency overhead and switching to higher processing rate you have to you have to be little careful and you have to follow

some steps. The voltage should be increased first followed by the increase in the frequency and the following steps will be followed. Why this is necessary?

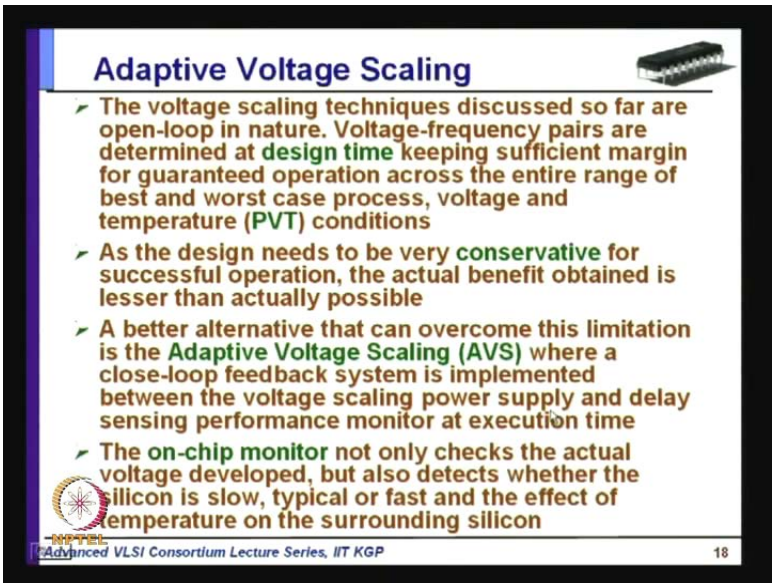
Suppose, if the frequency increases before the voltage increases what will happen chip will not operate, chip will not function properly as we know we are trying to give I mean you know, there is a relationship between voltage and frequency. So, as the voltage is increased, if the frequency is increased voltage has to be increased first. If the frequency is increased first then voltage the chip will fail to operate at that frequency that is the reason why you have to follow this step. Set the new voltage, allow the new voltage to settle down, then set the new frequency. Then secondly switching to lower processing rate in this particular case, set the new frequency, then set the new voltage the CPU continues to operate continues operating at the new frequency while voltage settles at the new voltage level. So, in this particular case restriction is less, because when the voltage is high the chip will operate at high frequency as well as low frequency, but whenever you are switching from you know you are changing from low to high frequency, there you have to be careful.

(Refer Slide Time: 49:17)



So, after discussing Dynamic Voltage and Frequency Scaling now it is time to discuss Adaptive Voltage Scaling. What is the difference between Dynamic Voltage and Frequency Scaling and Adaptive Voltage Scaling?

(Refer Slide Time: 49:30)



Now, the voltage scaling techniques discussed so far are open-loop in nature. Voltage-frequency pairs are determined at design time keeping sufficient margin for guaranteed operation across the entire range of best and worst process, voltage and temperature conditions that means you know those voltage frequency pairs are decided at designed time, although they keep on their applied at run time, but they are decided at design time.

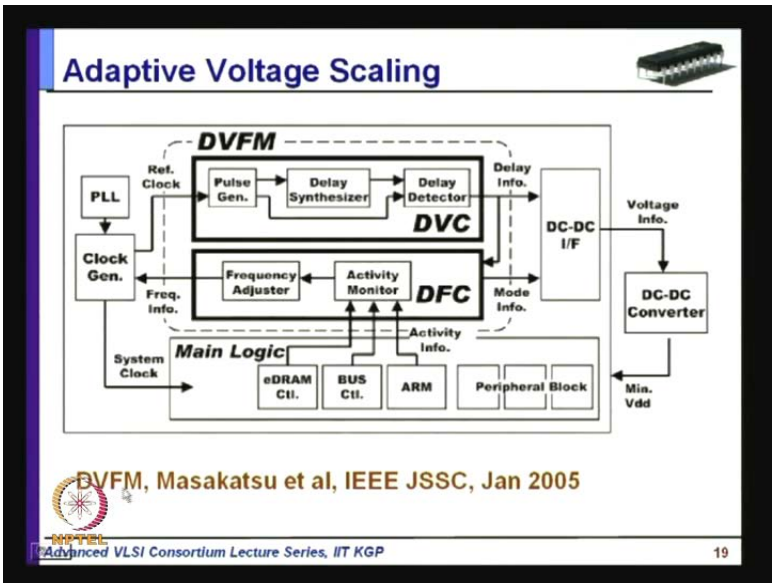
And here the design has to be very conservative, so that the chip operates under best and worst condition of process voltage and temperature variations. So, under best and worst conditions of process variation, voltage variation and temperature variation chip has to work and that is the reason why those voltage frequency pairs are chosen in a very conservative way and as a consequence, the gain or reduction in power dissipation that you really achieve is not very I mean, less than what can be achieved if the design is not that conservation, but how can it be done? It can be done if you use a close loop technique, instead of open loop technique.

So, in this case it is a open loop technique as the workload is changing, you are blindly applying a new voltage and new frequency without measuring whether the voltage and frequency requirements are appropriate or not? And that is the reason why a better alternative that can overcome this limitation is the Adaptive Voltage Scaling, where a close-loop feedback system is implemented between the voltage scaling power supply and the delay sensing performance monitor at execution time that means, as the

execution is taking place the CPU performance is monitored, it is delay is monitored, its power dissipation is monitored and by doing that you can adjust the voltage and frequency by using a close loop technique and for that purpose what you have to do? You must have an on-chip monitor.

That will not only check the actual voltage developed, but also detects whether the silicon is slowing typical or fast and the effect of temperature on the surrounding silicon. So, that means it is taking care of the PVT variation. PVT variations that take place since you are doing at run time the time of execution whatever variation has taking place that is monitored; and based on that, you are doing the design.

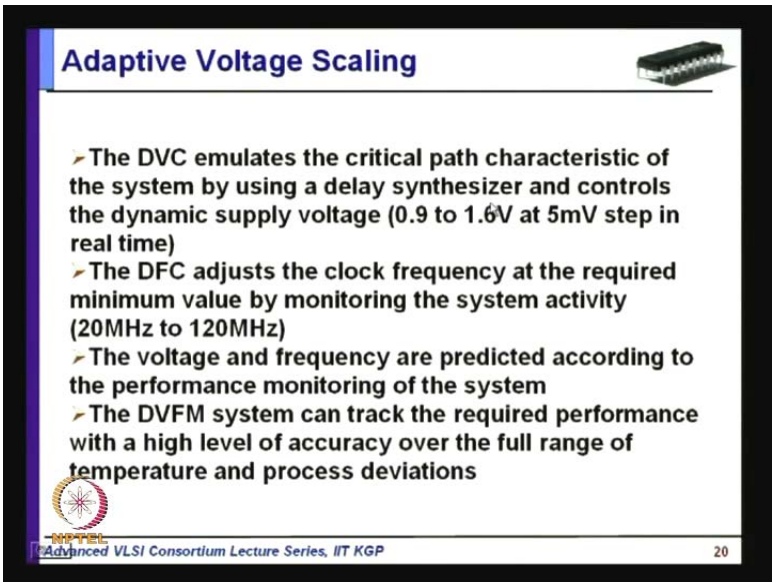
(Refer Slide Time: 52:22)



So, let us look at the system, what it does. So, here I have taken an example of Adaptive Voltage and Frequency Scaling from a paper, Dynamic Voltage and Frequency Scaling by Masakatsu published in IEEE transactions on you know Journal of Solid Set Circuits, January 2005. So, here you can see apart from these functional units the ARM processor, the RAM, the Peripheral Blocks, Blocks control circuits. This is the main block, which is shown as a small part of the circuit. Here you have got a dynamic voltage control, which does delay synthesize that means, the delay of the circuit can be the delay information is generated based on real life measurement by this. And similarly, the voltage and dynamic voltage and voltage dynamic frequency control circuit finds out at what frequency it should operate. Based on the actual activity measurement that is taking

place in the processor that means, whatever the actual performance required is based on that frequency generated and accordingly clock generator generates clocks and so on.

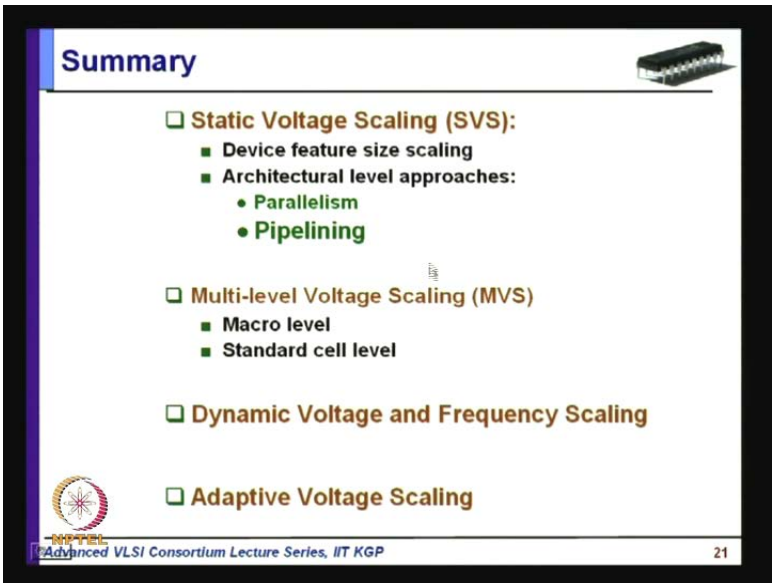
(Refer Slide Time: 53:37)



As it is explains here the DVC emulates the critical path characteristics of the system by using a delay synthesizer and controls the dynamic supply voltage in the range of 0.9 to 1.6 volt at 5millivolt step in real time. Similarly, the dynamic frequency control adjusts the clock frequency at the required minimum value by monitoring the system activity in the range of 20 megahertz to 120 megahertz. So, you see this DFC and DVC Dynamic Voltage Control and Dynamic Frequency Control. These are on-chips hardware, which does the actual measurement of delay and activity level.

And based on that, it sends information to the PLL and clock generator and then the clock generator generates the system clock and similarly, here the based on the delay information, information goes to DC to DC converter information block and then voltage is generated by the DC to DC converter and applied to the processor. So, you see the voltage and frequency are predicted according to the performance monitoring of the system the Dynamic Voltage and Frequency Scaling system can track the required performance with a high level of accuracy over the full range of temperature and process deviations. So, it is a kind of close loop control system that has been adopted, so it is a Dynamic Voltage and Frequency Scaling, where the close loop technique has been used instead of open loop technique.

(Refer Slide Time: 55:16)



So, with this we have come to the end of todays lecture and here is the summary we have discussed Static Voltage Scaling techniques, we have discussed Multilevel Voltage Scaling techniques. And today, we have discussed Dynamic Voltages and Frequency technique and special case that is Adaptive Voltage Scaling technique and in the next lecture we shall discuss about the minimization of switch capacitance. Thank you.