Real - Time Digital Signal Processing
Prof. Rathna G N
Department of Electrical Engineering
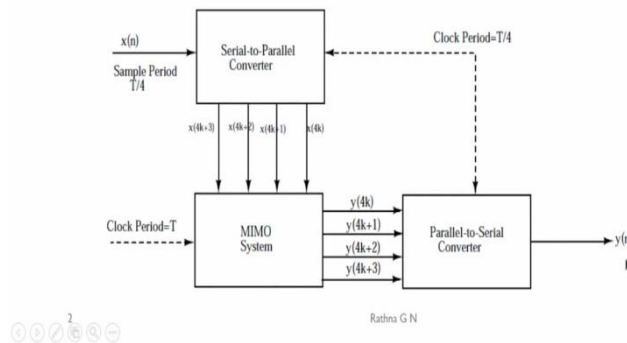**Indian Institute of Science, Bengaluru**

Lecture - 08
M2U8-Pipelining and Parallel Processing for Low Power Applications II

Welcome back. So, we are discussing about pipelining and then parallel processing for low power applications, continuing with the thing.

(Refer Slide Time: 00:30)

So, because we need for parallel processing, the data has to be fed in parallel. So, from sequential, how we; will be converting into parallel. We will see this figure. That is what we say is critical path has remained unchanged in these cases, but the interaction period is going to be reduced in this. So, what we say is $x(n)$ is the input, then if we consider in this case, as we are seeing, we will be considering 4 parallel lines.

So that it will give you a flavour of how 2 parallel lines, we had to take it, how 3 and then now we will see 4 parallel line or parallel units, if we consider, then what is the thing is going to happen. So, the sample period, what we will do is we will sub sample it to $T/4$ in this case, then we will go for serial to parallel converter, and then we will be getting $x(4k + 3), x(4k + 2), x(4k + 1)$ and $x(k)$ in this case.

And then in this case, the clock period is going to be T for multiple input and multiple output system and output all the 4 of them will be coming out parallely and then we had to convert back this parallel into serial converter, so how we can incorporate that. So, in this may clock

period is going to be $T/4$. And we will convert back this parallel into serial, so which will be coming out from the circuit as $y(n)$ in serial mode.
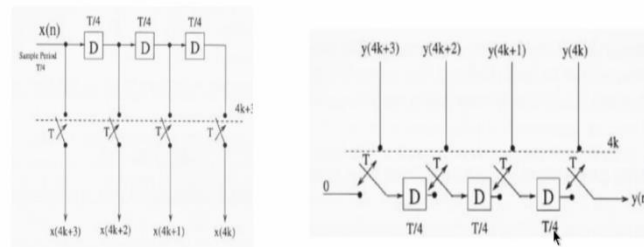
So, we will see using the switches how serial to parallel converter and then parallel to serial converter is going to work. So, this is my $x(n)$, and then input is $T/4$. So, we will provide the delay element in our path, and we will be providing the switches basically to see that all the output whatever input has been given, output will have 4 of them. So, every $T/4$ clock cycle, so, you will be switching on one of the unit.

So, all the 4 inputs are ready after my T_{clk}. So, then what will be this is how we will be feeding in all the 4 parallely. Now, this parallel has to be converted into serial basically, what we are going to do output what we have it is via $(4k + 3 + 2)$ and then +1 and then 4k. So, what we have is this is going to operate at 4k the switch and then you will be closing them every $T/4$ as you are seeing it and this $y(n)$ is the input.

And one of them is going to be opened and then you will be sending after a delay. So, for the first this thing switching all of them will be switched parallely so, you will be taking all the 4 inputs. So, this will have a 0 initially what you will be putting it and then you will be switching on this and then you will be seeing that 0 is going to be percolated first and then you will be operating on one after the other 4k will be coming out first and then next this one after one unit of time and then next this one.
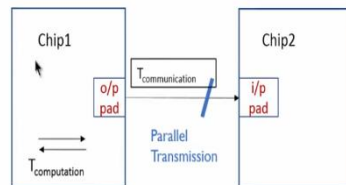
So, each one is clocked at $T/4$. So, in the 4 clock cycle, so you will be getting 4 outputs.

This is how we will be doing the serial and parallel conversion and then parallel to serial conversion. So, what happens why parallel processing is required. So, parallel leads to duplicating many copies of hardware and the cost is going to increase as well as if we are operating with the same units or same voltage we know that our speed is going to be increased and power consumption is going to be more so then why use this.

So, we say answer lies in the fact that the fundamental limit to the pipelining is at IO bottlenecks. So, that is referred to as communication bound composed of IO pad delay and the wire delay. So, in this case, you know chip one and chip 2 is there, this is the between the 2 chips this is the communication delay what we are going to have it. So, in the case of pipelining if my communication cost is more than pipelining, then no point in going for the pipelining stages number of stages to be increased. Then it is better to switch over to parallel processing then I can work on independent data in these cases. So, what we call it as this is the parallel transmission what I can have it, so you will be seeing that T is my computation. So, taking in the data from chip 2 to here, and data going from chip 1 to chip 2.

Combined Fine-Gain Pipelining and Parallel Processing

- $T_{iter} = T_{sample}$
  $= \dfrac{1}{LM} T_{clk}$
  $= \dfrac{1}{6}(T_M + 2T_A)$
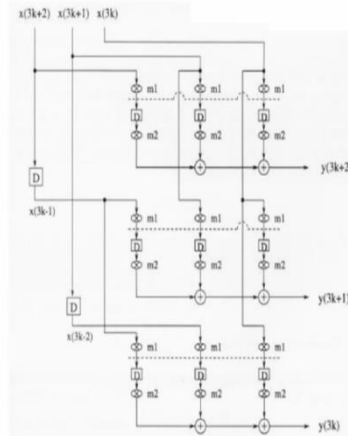
So, this defines the thing. So, we will see that how we can combine our fine gain pipelining and then parallel processing. So, you will be seeing the iteration, we said that sample period is equal to a T_{sample} in this case, then what happens if I put both of them pipelining and then parallelism, it becomes $\dfrac{1}{LM} T_{clk}$. So, that means to say $\dfrac{1}{6}(T_M - 2T_A)$, it is going to be till at iteration, every iteration, I will be getting one 6th of my what is it? I have increased my clock speed, how I can do this fine gained parallelism. So, you will be seeing that input is $x(3k + 2)$,. So, $x(3k + 1)$ and $x(3k)$. So, this is our parallel unit and then when I come to multiplier, I can bifurcate them into pipeline mode, that is m1 becomes 6 clock units and the other one I can take it as 4 clock units.

So, I have provided 2 pipelining basically, M = 2 and L = 3 then my iteration time, that is how it is going to be $\dfrac{1}{LM}, \dfrac{1}{6}$ of that original time what it will be I will be getting the output. So, you will be seeing just all the multipliers we have made it as pipeline multipliers. So that multiply clock time is made equal to the other side of the leg what we consider in the previous example.

(Refer Slide Time: 07:50)

Underlying Low Power Concept

- Propagation delay

$$T_{pd} = \frac{C_{charge} V_0}{k(V_0 - V_t)^2}$$

- Power consumption

$$P = C_{total} V_0^2 f$$

- Sequential filter

$$P_{seq} = C_{total} V_0^2 f, \quad T_{seq} = \frac{C_{charge} V_0}{k(V_0 - V_t)^2}, \quad f = \frac{1}{T_{seq}}$$

Rathna G N

So, what is the underlying low power concept we have to look in we said by, I can increase 6 times the clock speed for my both pipelining and parallelism or at the clock rate one 6th of it what I can operate, but how we are going to have a low power getting the thing. So, initially we said that this is the propagation delay equation what we had it so, that time period you will be seeing that for the sequential power consumption we call it a $C_{total} V_0^2 f$, T_{seq} is our clock period what we will be taking for every bit of data.

So, then what happens our delay is given by propagation delay pd is given by C into charge basically my capacitor charging time and then V_0 what is the input power and then k is a constant because this is proportional to we are assuming that k and V_0 is our input voltage and V_t is the threshold we know that in CMOS circuit, so, we have a threshold unit. So, in NMOS, we will be seeing that 1.5 volts whereas in CMOS it is going to be 0.5 yours about that we consider it as 1 volt.

So, that way we will see that power consumption is given by P = C total into V naught square into f, so, for the sequential filter when I applied this, this is what, what I will be getting at and then when I put the propagation delay is substituted I will be getting $\frac{C_{charge} V_0}{k(V_0 - V_t)^2}$. So, we are substituting V_0 with this and then into f. So, which is going to be $1/T_{seq}$.

(Refer Slide Time: 09:55)

Pipelining for Low-Power

- M-level pipelined system
- Critical path → $1/M$, capacitance to be charged in a single clock cycle → $1/M$
- If the clock frequency is maintained, the power supply can be reduced to $\beta V_0 (0 < \beta < 1)$

Seq: $|\longleftarrow T_{seq} \longrightarrow|$ (V_0)

M=3: $|T_{seq}|T_{seq}|T_{seq}|$ (βV_0)
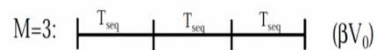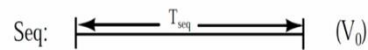
Rathna G N

So, when I do the pipelining what is the thing is going to happen we say it is M level pipeline system; the critical path we say is one $1/M$. So, that is capacitance to be charged in a single clock cycle becomes $1/M$. So, as you are seeing in the sequential, I have to charge the capacitor in T_{seq}, whereas in the pipeline if I assume M = 3, then the capacitor has to be charged as you will be seeing that time is going to be reduced basically, we call that as beta into V_0.

So, here initially it was V_0 and we will assume beta and to be not the clock frequency is if we maintain the same clock frequency, and we say the power supply can be reduced to beta times V_0, we say it is $0 < \beta < 1$.

(Refer Slide Time: 10:55)

Pipelining for Low-Power (2)

- Power consumption

$$P_{pip} = C_{total}\beta^2 V_0^2 f = \beta^2 P_{seq}$$

- Propagation delay

$$T_{seq} = \frac{C_{charge}V_0}{k(V_0-V_t)^2}; T_{pip} = \frac{\frac{C_{charge}}{M}\beta V_0}{k(\beta V_0-V_t)^2}$$

- Let $T_{seq} = T_{pip}$

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

Rathna G N

So, coming to low power continuation, we say power consumption what we said for the pipelining it is going to be $C_{total}\beta^2 V_0^2 f$. So, which we apply it as a sequential thing which is $\beta^2 P_{seq}$, then what happens to our propagation delay. So, that is T_{seq} we will be having $\frac{C_{charge} V_0}{k(V_0 - V_t)^2}$.

And for the pipeline case, we said the charging unit is reduced by M whatever in this case, we have taken it as 3. So, otherwise in normal $\frac{C_{charge}}{M}$ what it has to be charged to in pipelining into βV_0, we are substituting V_0 with β, βV_0. And here also we will be substituting $k(\beta V_0 - V_t)^2$, then we will apply that sequential is equal to period of pipelining stage.
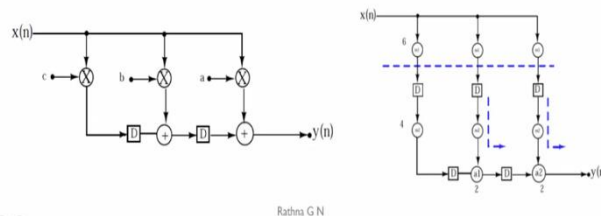
If we substitute them, then what happens by simplifying it, you will be getting it $M(\beta V_0 - V_t)^2$. So, which is equivalent to $(\beta V_0 - V_t)^2$ from this equation, we will be getting our β.

(Refer Slide Time: 12:23)

So, we will consider an example here. So, we will consider a 3-tap FIR filter and it is fine grain pipeline version what we will be assuming it so, it is shown in the following figure that this is the original a 3-tap FIR filter and this is our fine grained parallelism what we have considered. So, in this case, what the parameters have been given is my multiplier unit is going to take 10 units of time and my adder will be taking 2 units of time.

And you have been given the threshold voltage is given as 0.6 volts and then we know V_0 that is the supply voltage you are operating at 5 volts and the capacitor of multiplier equivalent to 5 times that of the capacitor of the adder. So, you will be seeing that pipeline filter the multiplier is broken into 2 parts m1 and then m2 with computation time of 6 units and then 4 units which

will be totally accounting for 10 unit of time with capacitance of 3 times and 2 times that of an adder, respectively, what we will be considering it.
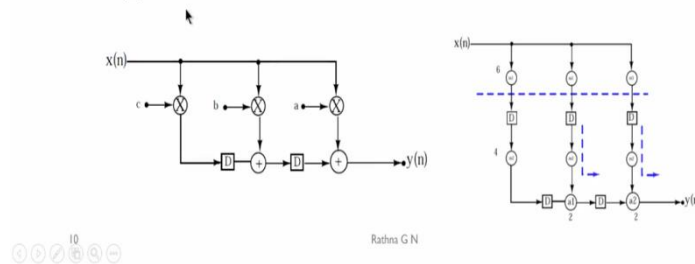
So, what happens to our equation, so, original what we had was $C_{charge} = C_M + C_A$. So, we said C_M is 5 times of C_A, which is going to be $6C_A$ in the fine grained it is going to be $C_{charge} = C_{m1} = C_{m2} + C_A$, because we have one pipeline stage there, so which is equal to 3 times that of C_A what it is going to be. So, then our equation will be seeing that m = 2 in this case.

We have assumed and then supply voltage $V_0 = 5\beta - V_t = 0.6$ threshold, which is whole squared is equal to β, which is equal to $\beta \cdot 6$ is our C_A basically $(-0.6)^2$. So, which comes, when you solve this equation, so it is going to be β will be equal to 0.6033 or 0.0239. So, we say that because our threshold voltage is 0.6 voltage is what it has been given, this value is much below that threshold.

So, capacitor will not be switching on. So, we say that this is infeasible. So, it becomes 0.6033 is the supply voltage, what we can reduce to that is a ratio is we call it as beta square when I take it, the reduction is going to be 36.4% but we say that pipelining should have given me 50%, but in this case only the reduction can be 36.4%. So, as you can see in this case.

Comparison

System	Sequential FIR (Original)	Pipelined FIR (without reducing V_0)	Pipelined FIR (with reducing V_0)
Power (Ref)	P_{Ref}	$2P_{Ref}$	$0.364P_{Ref}$
Clock Period (u.t.)	12 ut	6 ut	12 ut
Sample period (u.t.)	12 ut	6 ut	12 ut

Rathna G N

When we do the comparison, how it is going to be that a system is power is reference and sequential FIR originally if I take it that is in terms of my power reference and pipeline FIR without reducing the V naught, what I will be getting is 2 times the original one, I am supposed to get the output and then in case of with reduction in the voltage, so, it becomes 0.364 times the whatever power consumed with respect to reference.

And the clock period unit time what we are assuming is here it is going to take 12 units of time in the original because multiplier is 10 units and then adder is 2 units which is 12 and then in the case of pipeline. So, it will be taking 6 unit of time, whereas, when I without reducing the voltage, but if we reduce the voltage, so, we know that the clock period remains as the original one which will be having 12 units of time.

And sample period we know that this is 12 units, and here it becomes 6 units, whereas in the pipeline with reducing power it will be still 12 units.

(Refer Slide Time: 17:08)

Parallel Processing for Low-Power

- L-parallel system
- Since maintaining the same sample rate, clock period is increased to (LT_{seq})
- This means that C_{change} is charged in LT_{seq} and the power supply can be reduced to βV_0

Seq: $\quad \overset{T_{seq}}{\longleftrightarrow} \quad (V_0)$

L=3: $\quad \overset{3T_{seq}}{\underset{3T_{seq}}{\overset{3T_{seq}}{\longleftrightarrow}}} \quad (\beta V_0)$

Rathna G N

So, we will see that parallel processing for low power how we are going to achieve it. So, we say that we have L parallel system, since maintaining the same sample rate clock period is increased to L times the sequential one. So, this means that your C charge is charged in L into T_{seq} and the power supply can be reduced to βV_0. So, we are seeing that in the sequential the capacity that is going to be charged with respect to T_{seq} at V_0 voltage.

Whereas in the parallel if I were assuming L = 3, so, then what happens, this becomes 3 times T_{seq} all of them and power, the voltage reduction is going to be βV_0 what we will consider.

(Refer Slide Time: 18:01)

Parallel Processing for Low-Power

- Power Consumption

$$P_{par} = (LC_{total}) (\beta V_0)^2 \frac{f}{L} = \beta^2 P_{seq}$$

- Propagation Delay

$$T_{seq} = \frac{C_{charge} V_0}{k(V_0 - V_t)^2}; T_{par} = \frac{C_{charge} \beta V_0}{k(\beta V_0 - V_t)^2}$$

- $LT_{seq} = T_{par}$
- $L(\beta V_0 - V_t)^2 = \beta(V_0 - V_t)^2 \Rightarrow$ get β

Rathna G N

So, we will same thing what we will be applying the equation. So, the parallel will be (LC_{total}) $(\beta V_0)^2 \frac{f}{L}$ because I will be getting 3 outputs in one clock cycle. So, I will be my frequency can
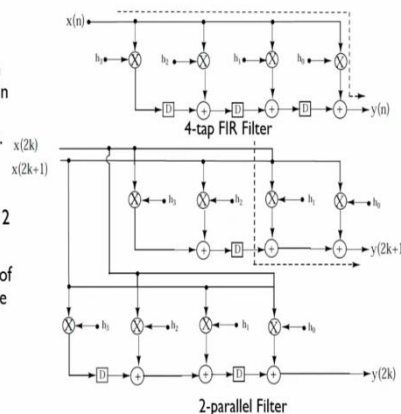
be $\frac{f}{L}$. So, which will be equating it as beta square into P_{seq} then what happens to our propagation delay. So, this is T_{seq} original $C_{charge}V_0$ by this one. In the parallel case; $\frac{C_{charge}V_0}{k(V_0-V_t)^2}$.

So, then $LT_{seq} = T_{par}$, so, we will be applying both together $L(\beta V_0 - V_t)^2$, which is equal to $\beta(V_0 - V_t)^2$. So, we will be getting a β from this equation.

(Refer Slide Time: 19:04)

As an example, we consider a 4-tap FIR filter shown in this figure, basically, and we have going to consider 2 parallel versions of this one. Here we will be considering the first version in the next slide we will consider the second version what is the thing is going to happen we will see that. The 2 architectures are operated the sample period 9 unit of time assume your multiplier is going to take 8 units of time and adder is going to take 1 unit of time.

And the threshold in this case voltage is given as 0.45 volts above that capacitor is going to be charged to 1. So, $V_0 = 3.3V$ what it is given that is supply voltage and then C_M, what we are going to have is a capacitor for the multiplier is equivalent to 8 times that of the adder. So, it is asking the question is what is the supply voltage of the 2 parallel filter and what is the power consumption of the 2 parallel filter as a percentage of the original filter. So, you are seeing the 2 parallel filter what we have considered.

(Refer Slide Time: 20:23)

Example – 3 Solution

Original: $C_{charge} = C_M + C_A$
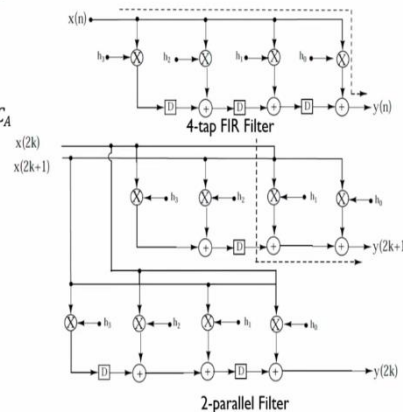
2-parallel: $C_{charge} = C_M + 2C_A = 10C_A$
$9(3.3\beta - 0.45)^2$
$= 5\beta(3.3 - 0.45)^2$

$\Rightarrow \beta = 0.6589 \ or \ 0.0282$
$V_{par} = 2.1743V$

Ratio= $(\beta)^2 = 43.41\%$

4-tap FIR Filter

2-parallel Filter

Rathna G N

So, here in this case, what is it original or capacitor charge is equal to $C_M = C_A$. And then, in the case of 2 parallel section, what happens $C_M + 2C_A$ which is nothing but $10C_A$ and then if we apply the equation, so, we will be seeing that $9(3.3\beta - 0.45)^2$ which is equal to $5\beta(3.3 - 0.45)^2$. So, in this case β becomes 0.6589 or 0.0282 as previous case, we have to ignore this because this is less than the threshold voltage of 0.45.

So, we will be considering 0.6589 as beta then for the parallel section 0.6589 into 3.3 volts which is going to come down to 2.1743 voltage. So, how much reduction we were able to get it 43.41% in this case.

(Refer Slide Time: 21:40)

Example – 4

• A more efficient structure than the previous one is depicted in the figure.

a) What is the supply voltage of the efficient 2-parallel filter?

b) What is the power consumption of the efficient 2-parallel filter as a percentage of the original filter?

Area-efficient 2 parallel filter
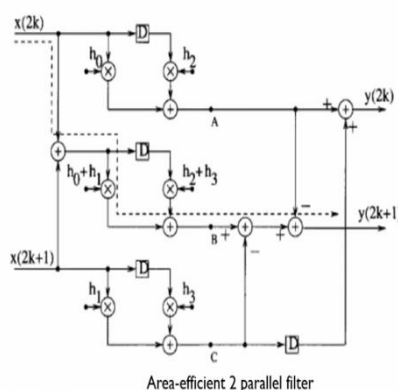
Rathna G N

So, coming to the next section that is parallel here what we have done is you have the 4 this things of multipliers here and here also 4 multipliers and you will be seeing that $x(2k)$ and $x(2k + 1)$ are the input to this structure. So, we will see that by modifying it, the structure in

this way, what we have taken is $x(2k)$ here with little arrangement that is we call it as a linear draw phase FIR filter if I consider the thing h naught and h 2 can be here.

And then the other thing I can derive from that, that is $h_0 + h_1$ will be multiplied here and the other one is $h_2 + h_3$ what I can combine and h_1 and then h_3 are coming from this parallel section. So, by doing little modification to the previous structure, what we will be getting is output is going to remain $y_2(k)$ and $y_2(k + 1)$. So, one of the assignment for you is what will be the output here a junction A.

And then the; junction B what you have to calculate as well as at C. So, we call this is area efficient 2 parallel multiplier. So, you have to count how many adders and how many multipliers are present in this case.

(Refer Slide Time: 23:17)
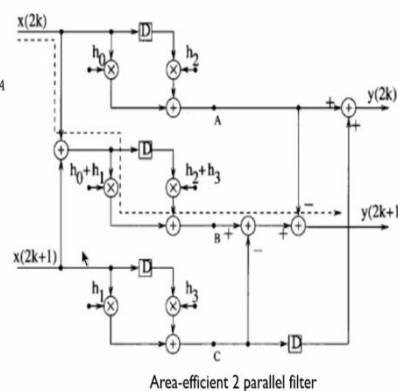
So, we assumed that we are charging $C_M + C_A$ what we are assuming it which is going to be $9C_A$ because $8C_M + 1C_A = 9C_A$ and we have new 2 parallel that is charge of the capacitor is equal to $C_M + 4C_A$ what we will be having get because we have 2 parallel section so, which is equal to $12C_A$. So, if we substitute in this because here also you are 2 parallel section 2 into 9 we will be achieving this equivalent to $12\beta(3.3 - 0.45)^2$.

So, then β turns out to be 0.745 or 0.025. So, as earlier cases this is infeasible and we consider our pipeline the parallel version of its going to have 2.45857 volts and then the ratio with respect to this you will be calculating it as 43.6% the saving in the voltage. So, in the previous

case you have saved 43.41 here you are saving 43.6% with area efficient because I have reduced my adders and multiplier by introducing the delay in a proper so please look into this.
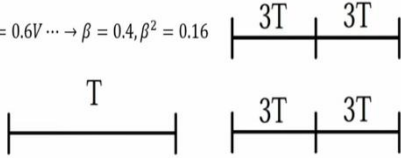
(Refer Slide Time: 24:58)

So, now you can we combine pipelining and then parallelism together and then try to achieve better voltage reduction. That is what we will look in these slides. What is it we have the sequential as usual and then pipelining we have $\frac{C_{charge}}{M}$, whereas in the case of our parallel, it becomes LT_{seq}. So, which implies that, I will be making the left hand side as the number of stages for pipeline into number of stages for parallelism.

Which is given by $ML(\beta V_0 - V_t)^2$ which is equal to this. So, if we consider both of them M = L = 2, then we are operating $V_0 = 5V$, then what threshold is given as 0.6 volts. So, if we compute the values, β becomes 0.4 then $\beta^2 = 0.16$. So, this is how we will be doing the thing this is our sequential. So, what we have done is 3 parallel units what we are going to have it so, and then one pipeline what we have considered.

(Refer Slide Time: 26:24)

- Methodologies of pipelining of 3-tap FIR filter
- Methodologies of parallel processing for 3-tap FIR filter
- Methodologies of using pipelining and parallel processing for low demonstration

19   Rathna G N

So, to conclude this pipelining for low power we discussed about 3-tap FIR filter for pipelining and then we consider for parallel processing also 3-tap and then 4-tap in 2-tap filters. And then pipelining and parallel processing together how you can achieve low power that is what demonstrated, you can work out some of the problems and then see that how it is going to improve on your power consumption.

(Refer Slide Time: 27:00)

M2U9

- IIR Filters - Low Pass, High Pass, Linear Phase

20   Rathna G N

That is what it is going to be reduced. So, in the next class, we will be discussing about IIR filters, that is low pass and a little bit on high pass filter. And then, as we know that IR filter is going to lag in the case of linear phase will be achieved only in FIR filter. So, IR filter becomes nonlinear. So, we will look into that in the next class. Thank you.