

**Digital VLSI Testing**  
**Prof. Santanu Chattopadhyay**  
**Department of Electronics and EC Engineering**  
**Indian Institute of Technology, Kharagpur**

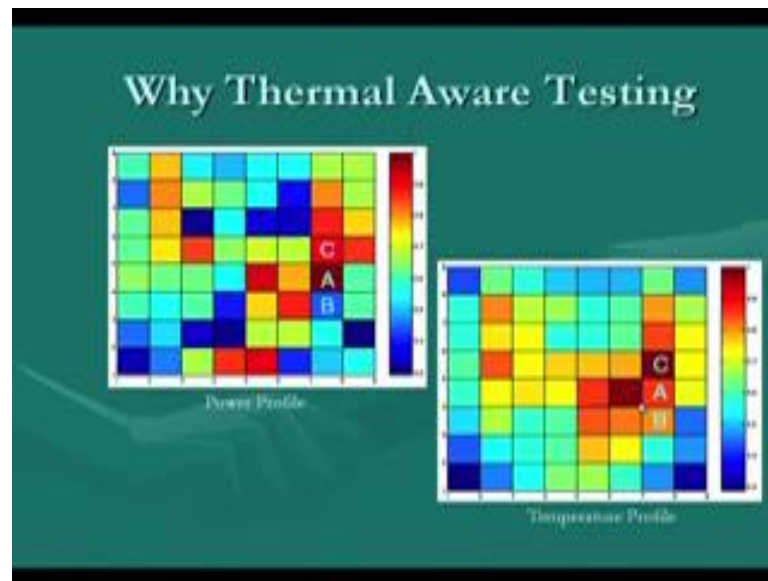
**Lecture – 37**  
**Thermal Aware Testing**

Next, we will discuss on thermal aware testing. So, power aware testing, so that was targeted to reduce the power consumption during test. Now, this power reduction is good because the circuit may have some power budgets, so we should not exceed that power budget. So, if we exceed that peak power consumption then circuit may get damaged and it may also if the average power is more, then the battery life will be poor. So, those are those who are the concerns.

Now, this power alone may not be sufficient for ensuring the goodness of the testing strategy. We also need to consider the heat that is generated in the testing process. So, if a circuit consumes lot of power, so it is expected that the heat generated will also be high because a good amount of consume power is dissipated as heat, so temperature of the chip will be high. And if the temperature goes up then there are several problems that can come, so as a result so we need to do something for that this temperature of the chip can be kept under control.

And what happens is that most of the chips that we have in today; they have got a dynamic thermal management policy are known as DTM. So, if the temperature exceeds certain frequency certain values certain threshold value then the DTM policy, so it will try to shutdown the chip or it will try to reduce the frequency of operation of that chip. So, during testing is the power consumption is high, so it is possible that we cross that temperature threshold; and if we cross that temperature threshold, the DTM policy will come into operation, and it will put the circuit in shutdown mode or it will take it to a lower frequencies affecting the efficiency of the testing process. So, that is why we need to address these thermal issues explicitly in the testing.

(Refer Slide Time: 02:24)



Now, why this thermal aware testing is so important? Suppose, we have got a circuit and we see that the power profile of the circuit is like this, so, it is divided into blocks. So, if the blocks which are that are cool, so, they are deep blue color. And as we are going to hotter and hotter block, the blocks that are consuming more and more power, so, they are turned towards red and towards brown. So, this you see that a among these blocks, so this block A is consume lot of power, so that way its power consumption is high. So, we will try to reduce the power.

On the other hand, if we look into the temperature profile, we may find that block A is not that much a concerned, but block C is more of a concern; and this block is also a concern, which are previously the block C and this one, they were not of much concern. But when you see the temperature profile C is the more heated than A, and this particular block left side of A, so that is also more heated than A. Why this thing happen, because the heat that is generated it has to that is that the heat generated depends not only on the power consumption of a single block, but it also depends the on the surrounding blocks. Because surrounding blocks through this boundary the heat will be transferred, so, they will be transported from C to A or A to C and similarly this thing.

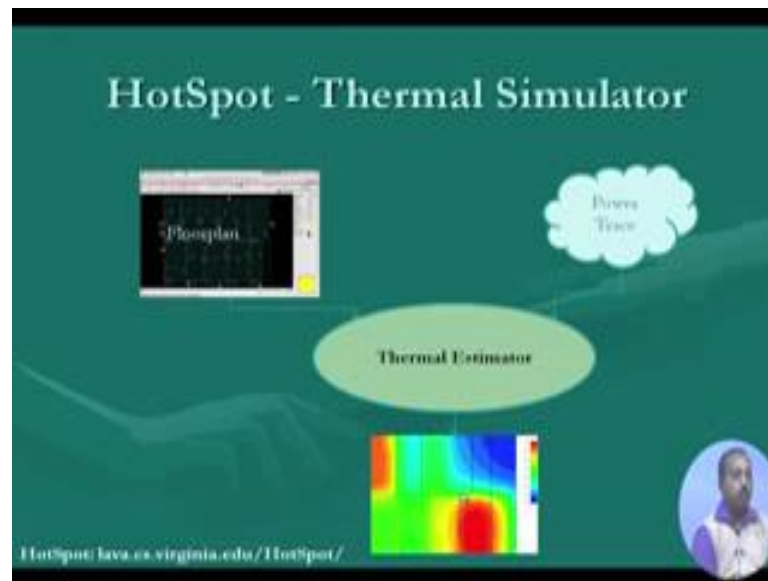
Now, if a block is surrounded by some hotter blocks then naturally it is expected that it

will be not be able to remove its heat easily. As a result its temperature will be higher than the block that is surrounded by blocks whose temperatures are not that high, so that way it will be able to transfer a good amount of its heat to its neighbors. And as a result this that will not be that that block will not be that much heated, so this is what we were looking for, so this power reduction technique or a peak power reduction technique will try to reduce the activity in the circuit, so that this block A does not see much of activities. But a thermal aware strategy should try to reduce the activities for C, so that this temperature of C is not that much affected.

So, you see what is happening is that this temperature is dependent not only on the power consumption of the block, but it also depends on layout of the chip. So, the distribution of temperature is dependent on the layout of the chips, so there is a special information that is also coming into picture apart from the power profile, so that makes this temperature handling problem more difficult compare to this power handling problem. So, and if we do not handle this, then what is going to happen, so as the temperature of our circuit or system increases, so for every 10 degree centigrade rise in temperature leakage power almost double itself. So, if we double itself then it will cause further power consumption and further heat generation, so that way that may finally lead to a runaway thermal run away may occur, so destroying the chip, so that way we need to take care of that.

And also another thing is that if different parts of the chip they are differently heated, so, their temperature are not uniform, then the delays of different blocks will also not be same. So, the delay of this block C and say this block or this block, so they are not same because the temperature may have a wide variation. And if it is so, then this delay test will fail, so delay test in some cases for good circuit it may find it as failure and for bad circuits it may find it as a pass, so that way it is also to be taken care of.

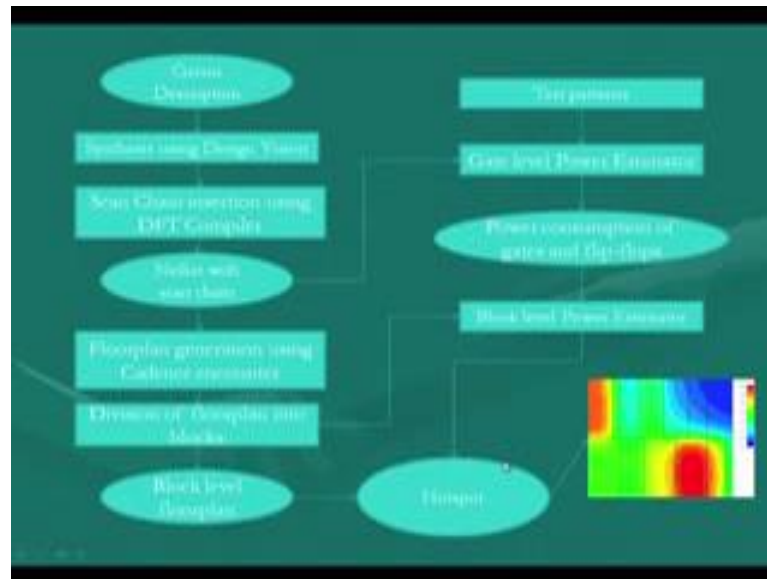
(Refer Slide Time: 06:48)



So, how do we get this temperature profile, so that is the first thing. So, what is done is that there are many thermal estimators that are available, so like the one of them is this tool hotspot. So, this is available in this particular site. So, this hotspot, what it does is that it takes the floor plan of the chip as an input, and it also takes the power trace of this individual blocks. So, and then it may be a block level estimator or it may be a gate level estimator, so whatever it is, so this floor plan and this power trace, so they are fed to this thermal estimator, it does some thermal simulation and then it comes up with the thermal modeling. So, it gives the temperature distribution of various blocks in the circuit.

So, we will not go into how this thermal simulators are designed and all that, so that is a part of this thermal estimation process, we are not going to that. So, we will be using these thermal estimators to see that temperature profile of the chip. And we will try to see how will try to see look into the techniques by which this temperature distribution can be made more uniform, and this the peak temperature of the system can be reduced.

(Refer Slide Time: 08:14)



So, how can we do this block level simulation? So it is like this. So, start with the circuit description, then the circuit description is passed through some synthesis tool. So, maybe we can use some tool like design vision from synopsis or some other tool by which we can do the synthesis of the system. So, by the synthesis after the synthesis, we do scan chain insertion using some DFT compiler. So, this scan chains are inserted into the circuit. Then after this we have got with us the net list with scan chains. So, we have got the net list with scan chain in our hand. Now, we using some floor plan generation tool can be used. A typical example is the cadence encounter tool by which you can generate the floor plan corresponding to this scan Netlist with scan chain, the Netlist you generate the floor plan.

And then for this thermal estimation purpose, so we divide the floor plan into number of blocks. Because this thermal simulators, so if you make it to work at the granularity of individual gates, then it will take huge amount of time. So, to reduce the time requirement what is done most of the research works, so they divide the floor plan into number of blocks, so that then it does a block level simulation. So, we get after do doing this blocking, so we get a block level floor plan.

On the other hand, these test patterns have been generated by some ATPG tool for the

circuit. So, this Netlist with scan chain, so that is fed with this test pattern to a gate level power estimator. So, gate level power estimator it will find out transitions occurring at individual gates in the circuit and that way depending upon the gate type, we can have some estimation of the power consumed by individual gates. So, we get the as an output of this gate level power estimator, we can get the power consumption of gates and flip flops. And then we have already divided the floor plan into blocks, so we can sum the power consumption of individual gates and flip flops belonging to a block and sum them up and get the block level power estimation, so for individual blocks that we have in our floor plan, so we can get what is the power consumption.

Now, we have got these 2 tool; the block level floor plan is available and block level power estimator is available, so they can be fed to some thermal estimator tool such as hotspot and then it can give us the temperature profiles. So, this is the typically the flow that has to be followed, if we are trying to get the temperature profile for a circuit and a given test pattern sets. So, you see that the process is quite complex compared to say this power relay, this power optimization, this low power testing and all that, so the steps are much more complex. And in particular this at the lowest level to get the temperature values we have to go for a thermal estimations, so which is time consuming. And there is no such straightforward guideline like the scan chain transitions being more, temperature will be more there is no such guideline like this.

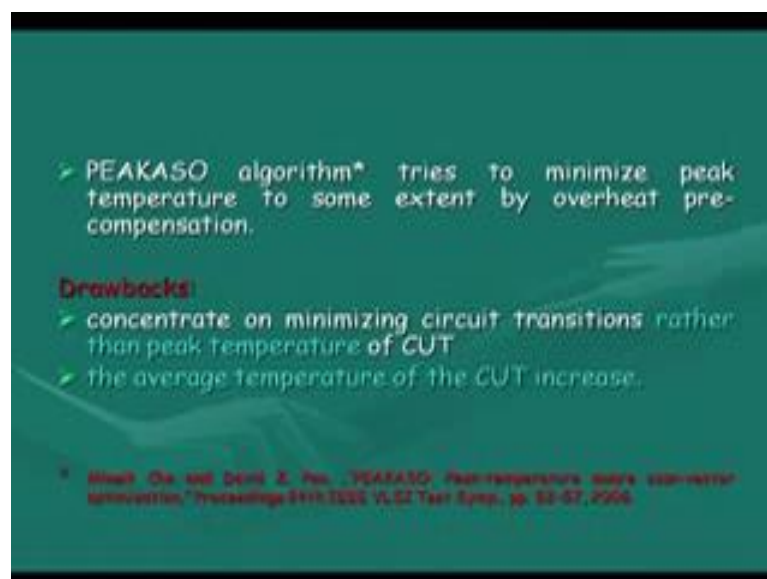
So, scan chain transitions being more, the circuit power will be more, but how this individual blocks with in a circuit, they will consume power, their consumption pattern will change, so it is not depicted by the scan chains. So, naturally the procedures that we have to use they are to be more elaborate and more exhaustive in nature. So, we will go back to the will classify this whole presentation again into 2 part that we are looking previously first part is the external testing where the test patterns are stored in some automatic test equipment; and from the automatic test equipment, the test patterns are transported to the circuit and there it is applied.

(Refer Slide Time: 12:13)



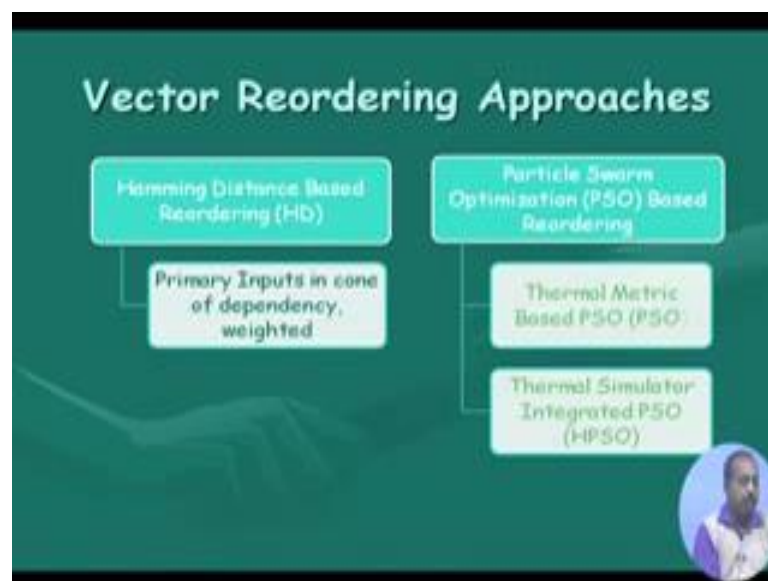
So, a very basic strategy that can be followed is the test vector reordering. So, we have seen that the test vector reordering it can reduce the test power, so similarly this test vector reordering can also reduce the temperature. So, we will see some techniques based on some techniques based on hamming distance and another technique based on a more detailed approach which is known as particle swarm optimization.

(Refer Slide Time: 12:44)



So, if we look back there are some algorithms that have been reported in the literature which is way back in 2006, PEAKASO that tries to minimize the peak temperature to some extent by overheat compensation. So, what this work does is that it takes it classifies the patterns into some patterns which are consuming high which are resulting in high temperature some of them resulting in low temperature, so between this high temperature consuming transition patterns, so they insert this low temperature computing pattern. But the problem is that temperature is not a very instantaneous phenomena, so it occurs over a sequence of patterns, so it is not very easy to identify or declared that a particular test pattern is going to a consume lot of heat, it is going to generate lot of heat. So, it is difficult to tell that way. So, they actually try to concentrate on minimizing circuit transitions rather than peak temp temperature of the circuit, so naturally they cannot reduce this temperature directly, so they try to reduce this circuit transition and average temperature may of the circuit may increase.

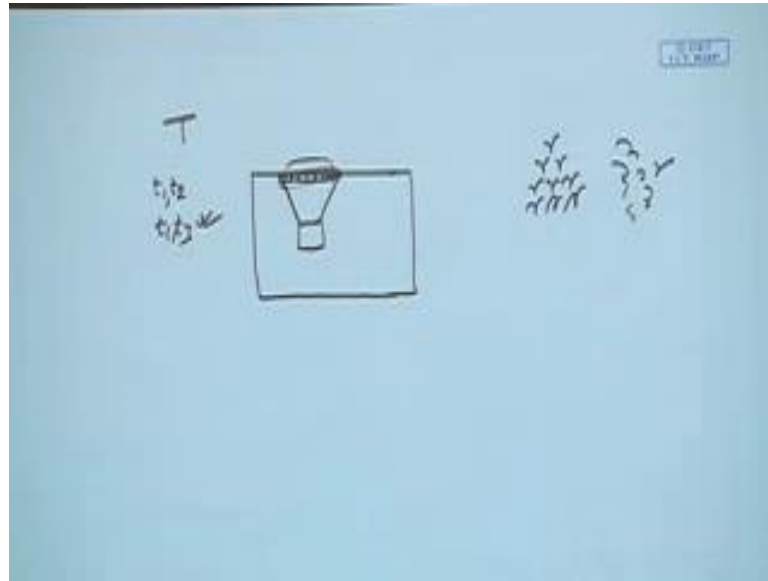
(Refer Slide Time: 13:58)



So, what we do is that we compute this in the hamming distance based approach, so will be finding out the primary inputs in cone of dependency, and the weighted part of it.

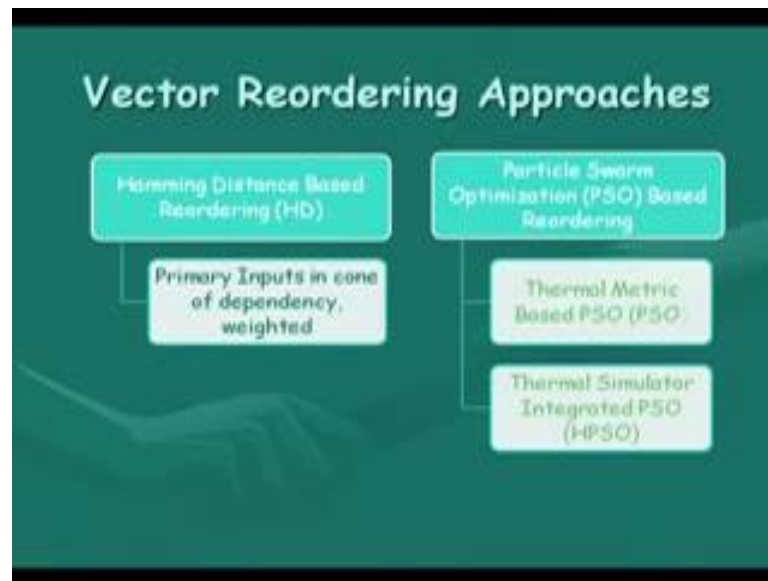


(Refer Slide Time: 14:13)



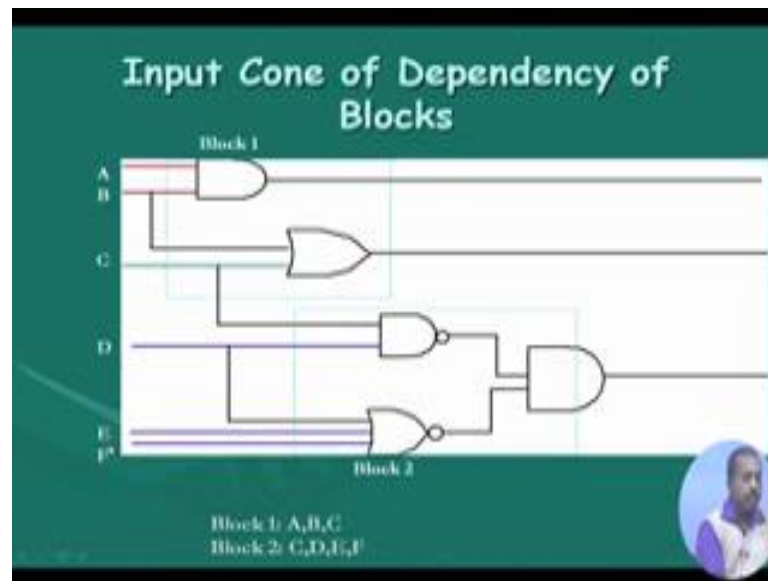
So, what happens is that, if this is my circuit and after for a given test pattern say  $T$ , we may find that this is the block which is becoming very hot. So, what we do, we figure out the cone of influence primary input cone on which this party depends. Suppose, this is the primary input cone onto which this block depends. Now, naturally if we can reduce transitions that are occurring in this block, so between 2 successive patterns  $t_1$  and  $t_2$  that we are applying to the circuit. If you find that for this part of input, the transitions are less in instead of applying  $t_1, t_2$ . So, if I apply  $t_1$  to  $t_3$ , if we find that in this case transitions in this part is less compare to  $t_1 t_2$ . Then it is expected that if I apply  $t_1, t_3$  instead of  $t_1, t_2$  in that sequence then the block that we have in our question, so that will consume less power. And it will try it may be its activity will reduce. So, the power consumption will reduce and then that is expected to reduce the temperature, so that is the hamming distance based approach.

(Refer Slide Time: 15:26)



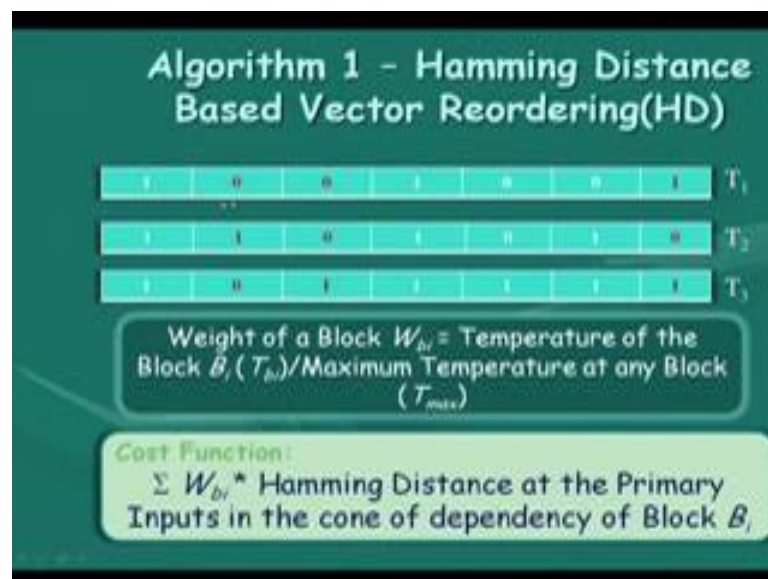
So, it reduce the we look into the hamming distance between 2 successive patterns, so that hamming distance being less means less activity. And then there are this particle swarm optimizations, so the work that is reported, so it will be doing 2 different thing, it will first of all it will use some sort of metric to model this thermal behavior of the circuit. And the next one, so it will directly integrate the thermal simulator into the circuit, so naturally the results will be best when we have got this thermal simulator integrated into the optimization process, but the competition time will be more, so that may not be feasible in many cases. So, on the other hand, this thermal metric based PSO some metric will be utilized that will reduce that is expected to reduce the temperature.

(Refer Slide Time: 16:19)



So, the cone of dependency as we were telling that if this is block one then the cone of dependency has got inputs A, B and C; similarly if this is a block 2, so cone of influence is the pattern are the input, so D, E and F.

(Refer Slide Time: 16:34)



So, hamming distance based approach, so what is does is that suppose we have got these

3 patterns  $T_1$ ,  $T_2$  and  $T_3$ , so we see that between these 2 between these 2, so this  $T_1$ , so bit changes again here this bit changes, so these are the bits which are going to be affected. So, this may be some for some block may be these are the inputs that are effective for this block that the inputs that are marked in red that may be the in the cone of influence of the particular block. And I need to compute what is the hamming distance if we apply these patterns in this sequence  $T_1, T_2, T_3$  or  $T_1, T_3, T_2$  like that. So, what we do we find out we assign some weight to every block. What is it this is equal to temperature of the block  $B_i$ , which is given by  $T_{B_i}$  divided by maximum temperature of any block  $T_{max}$ .

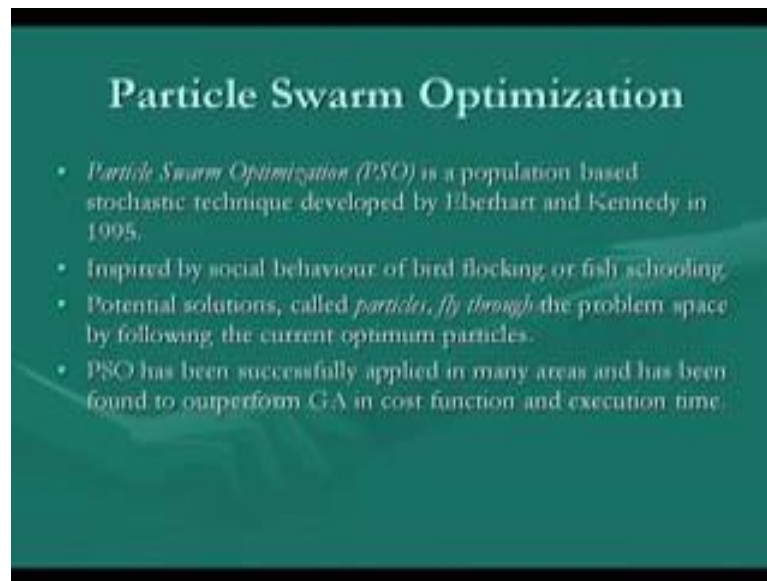
So, given a test patterns set which is unordered, so we do a thermal simulation and find out what is the temperature of the individual block. And what if for the entire chip, what is the maximum temperature, so out of that, so how critical is that blocks temperatures, so that is come that is the weight of the block, so that is computed as this temperature of the block divided by the maximum temperature value. So, this way we get the weight. Now, this cost function is computed as summation of weight of the block multiplied by hamming distance at the primary inputs in the cone of dependency of block  $B_i$ . So, for every block, so these values are the hamming distance are computed and it is multiplied by the corresponding importance of the block that is the weight factor  $W_{B_i}$  and that is summed up, so that will tell us like what will happen if I choose a particular ordering.

(Refer Slide Time: 18:37)



But the problem that we have in this case is that it may lead to an increase in the transition at primary inputs other than in the cone of the hottest blocks. So, for basically we since we are having this weight factor which is guided which is actually depicting which is up actually helping the high temperature blocks to reduce the temperature, so in effect it can increase the transition at the primary inputs which are not feeding the hottest cone, but some other cone. And as a result that new cone become now hotter than this one, so as a result this for other blocks it may increase the temperature. And average temperature of the chip may increase, because we are targeting to reduce the peak temperature, so the average temperature of the system may increase. So, the mean temperature will increase, so that may has some adverse effect on the delay and frequency of the chip.

(Refer Slide Time: 19:42)



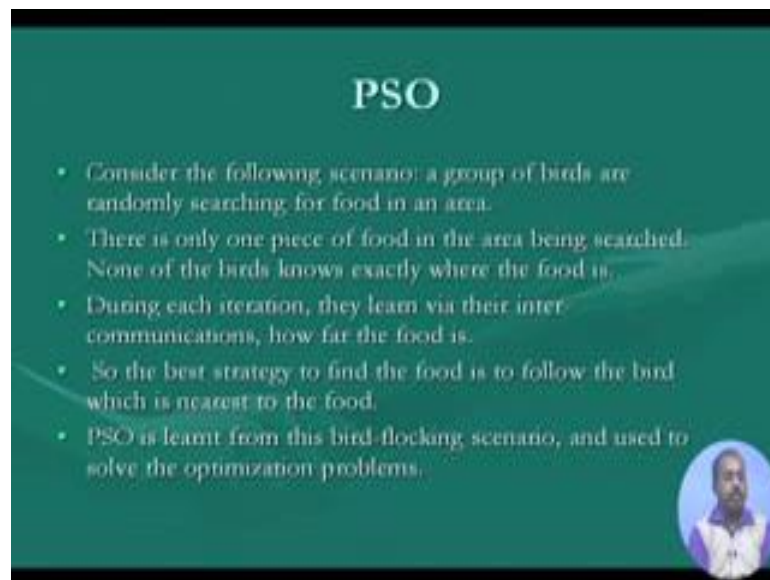
Next we will look into a particle swarm optimization based strategy. So, particle swarm optimization it is again a meta search technique similar to the genetic algorithm we had discussed some classes back while in the ATPG algorithm. So, it is similar to that that, but it is much more I should say efficient in most of the optimization problem which shows much better result compared to genetic algorithm. So, it was developed by Eberhart and Kennedy in 1995, it is inspired by social behavior of bird flocking or fish schooling.

So, if you look into group of birds flying on the sky you will find that they follow some sort of v type of pattern, so may maybe there is a bird at the at the beginning then there are some other bird which are following that bird, so it they often fly like this. And if you look it look at it then we may find that after sometime one of this bird they dive at this like this and all of them they start following this new this new direction. So, apparently this is due to may be their wish or their joy, but these 2 people they thought that it may be due to some other reason.

So, they say that these birds they are flying on the sky to search for some food and they have some estimation of food and they try to figure out the food where the food is. So in PSO particles swarm optimization is inspired by this operation, so every potential

solution that is called a particle it flies through the problem space by following some current optimum particles. So, it actually every particle is a solution like in chromo in genetic algorithm every chromosome was a solution here every particle is a solution. And these particles they fly through the search space they fly through the search space and looking for some better solution. So, PSO has been applied successfully in many areas and it has been found that it outperforms genetic algorithm in terms of cost function and execution time in many cases of course, there are many other optimization techniques, but we will discuss about it.

(Refer Slide Time: 22:06)



## PSO

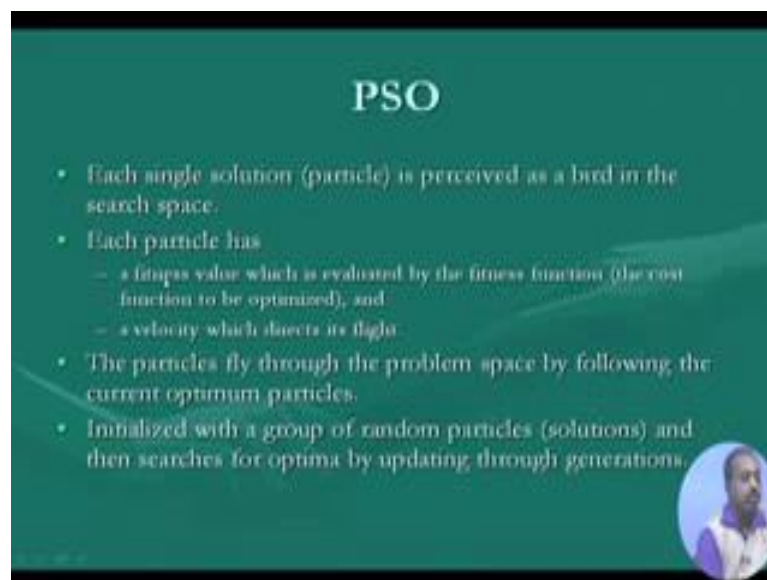
- Consider the following scenario: a group of birds are randomly searching for food in an area.
- There is only one piece of food in the area being searched. None of the birds knows exactly where the food is.
- During each iteration, they learn via their inter-communications, how far the food is.
- So the best strategy to find the food is to follow the bird which is nearest to the food.
- PSO is learnt from this bird-flocking scenario, and used to solve the optimization problems.

So, let us see how this PSO is formulated. So, consider this scenario a group of birds are randomly searching for food in an area; and there is only one piece of food in the area being searched that is the only piece of food. And none of the birds knows the exactly where the food is. So, similarly when we are looking for say solutions, so we do not know the exact optimum solution for the problem, so we may have some idea about what may be the optimum solution looking like or where it may be found, but we do not know the exact solution. So, none of the birds knows exactly where the food is. And during each iteration, they learn via inter communications how far the food is, so every bird has got some idea about how far am I from the food. So, it is assumed that they have got strong notion they can get a very they have got very strong idea like how far am I from

the food.

Now, birds are thought to be collaborative cooperative, so they the bird which has got the best idea, so it tells other birds that ok, see this is the current best current best estimation. And all birds all other birds, so they compare their estimate with that estimate accordingly takes some decision to fly. So, the best strategy to find the food is to follow the bird which is nearest to the food; however, there may be the case that a bird has got its own intelligence also. So, over it is flying for quite some time, so during that time also it has got some idea about what the where the food was, so that way it is guided by 2 factors. One is the information from the best information about the current best, then the history that the bird has seen during its fly through the space and some inertia. So, it was flying in some direction, so it cannot change immediately, so it has some inertia, so they will be guiding the flying direction of a bird or the in our case the change in the solution, so that will be guided by this thing. So, PSO is actually learnt from this bird flocking scenario, and it is used in solving many optimization problems.

(Refer Slide Time: 24:37)



## PSO

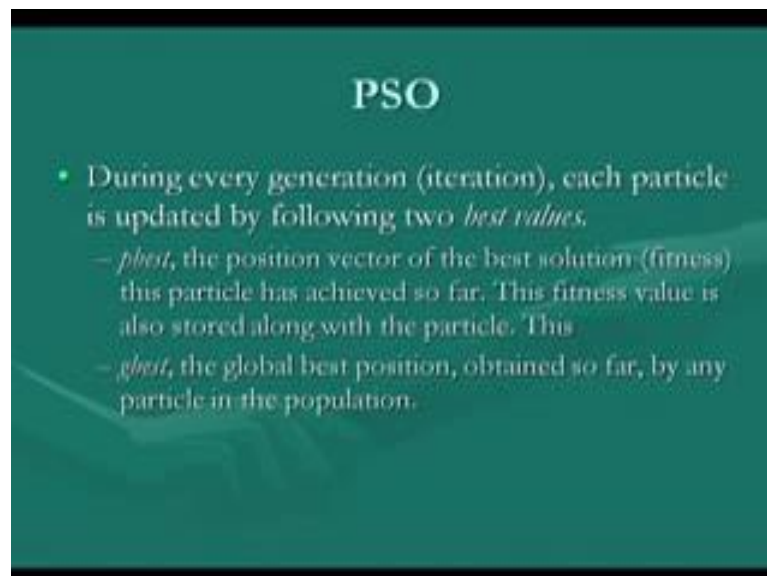
- Each single solution (particle) is perceived as a bird in the search space.
- Each particle has
  - a fitness value which is evaluated by the fitness function (the cost function to be optimized), and
  - a velocity which directs its flight
- The particles fly through the problem space by following the current optimum particles.
- Initialized with a group of random particles (solutions) and then searches for optima by updating through generations.

So, each single solution or particle is perceived as a bird in the search space. Each particle it will have a fitness value, which is evaluated by the fitness function, the cost function to be optimized and a velocity which directs its flight. So, there is a fitness every



particle has got a fitness and a velocity. So, if it is an n-dimensional space then for each dimension it has got a velocity, so that will be updated at every step and for each direction velocity will be determined. The particles that fly through the problem space by following the current optimum particle that is there and initialized with a group of random particles that is random solutions and then searches for optimum by updating through generation. So, now rest of the thing is similar to genetic algorithm, so here also we have got some initial population which may be randomly generated or may be guided by some other heuristic and then it evolves through generations by considering the global intelligence and the local intelligence.

(Refer Slide Time: 25:48)



So, during every generation the particle is updated by following 2 best values. One is the pbest that is the position vector of the best solution, this particle has achieved so far. Since, it is flying for quite some time or the particle is evolving over a number of generations, so it over the generation, so it has some idea, it has got the idea like when it saw the best solution. And since if we are looking for a say minimization problem then when the solution was at minimum for a particular particle or if you are looking for maximization problem then when was the solution; the fitness value was maximum for the particular particle, so that way, we have got this notion about this pbest. The position vector of the best solution this particle has achieved so far. And this fitness value is also

stored along with the particle. And then we have got gbest which is the global best position that is obtained so far by any particle in the population, so that is actually the global intelligence. So, this local intelligence and this global intelligence they are coming and they will be combined to determine the velocity of the particle in the near future. So, we will continue in the next class.