

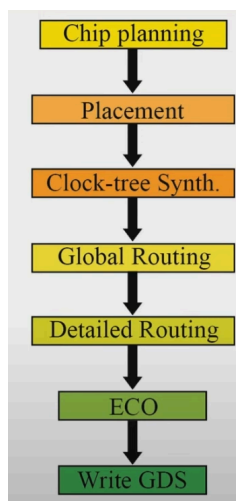
VLSI Design Flow: RTL to GDS

Dr. Sneh Saurabh

**Department of Electronics and Communication Engineering
IIIT-Delhi**

Lecture 49 Placement

Hello everybody. Welcome to the course VLSI design flow RTL to GDS. This is the 39th lecture. In this lecture, we will be discussing placement. In the earlier lectures, we have seen that physical design is broken into multiple tasks which are shown in this figure. And in the last two lectures, we had looked into chip plans.



And after chip planning, the next task is placement. So, in this lecture, we will be discussing placement and specifically we will be discussing these topics. Global placement, legalization and detailed placement, scan cell reordering and spare cell placement. So, let us first look into what a placement does.

So, the placement basically decides the location of standard cells. In the last lecture, we had looked into chip planning and there also we had looked into the placement of larger entities. So, it is important to distinguish what the placement does. So, placement is basically deciding the locations of standard cells and not of larger entities. So, the larger entities such as the macros or the blocks that we have created by partitioning or the and

the IO cells, those are actually placed during the chip planning stage.

And in this stage, we are actually placing or deciding the location of locations of the standard cells and where do we place these standard cells. So, the region for the standard cells was also decided during chip planning. Now, for each individual standard cell, we are deciding the locations in the region that were already allocated during chip planning. Now, what is the goal of placement? So, the goal of placement is to ensure the routability of a design. So, it is important to understand that the most difficult task in physical design is routing the routing of the various entities in our design.

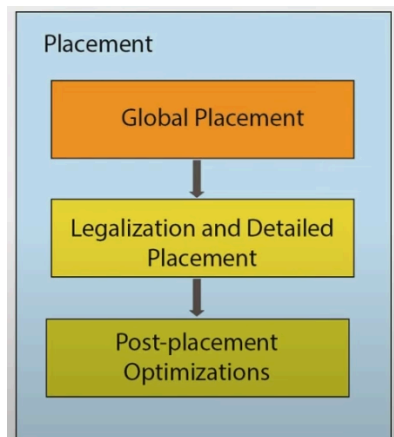
And standard cells are the most numerous entities, they are in millions and their connections are also very complicated. So, it will be in millions or billions of connections or nets that we need to make connections in physical design. And making those connections using the wires or interconnect on using various layers of interconnects is a really challenging task. And what does placement do? Placement tries to find good locations of the standard cell such that routing becomes possible. If the standard cell locations are not good then there may not be any feasible solution and then it becomes challenging to route our design and close our design.

So, the primary goal of placement is to make the or decide the placement of the standard cells such that routing is possible or routing is easy for the subsequent task that is the routing that will be performed later in the physical design flow. Now, how do we assess the routing of a design? So, the placement when it decides the location of the standard cells it cannot do the routing why it cannot do the routing? Because routing itself is a very complicated task. For example, if we have one standard cell and another standard cell there can be multiple ways in which this connection can be made. For example, one path is shown here if and there can be multiple other options which are available. And once that is located once the wire is laid out for say between two entities then the routing of the next wire becomes difficult.

For example, now if it can suppose these two are the cells this cell and this cell and we want to connect it. Now we cannot connect it through the path because it can lead to shorting. So, routing of one net affects the routing of other nets also and therefore, deciding or assessing the routability of a design is a challenging problem. So, as an alternative what does placement do is that it tries to minimize the total wire length of the design. Total wire length means that suppose there are n nets in our design then for the sum of the wire length for each of the nets is the total wire length of our design.

So, in placement what does placement do? It tries to minimize the total wire length of our design. And intuitively what it means is that if we decide we minimize the wire

length of our of the nets then what it means is that if two cells are say connected through some pin then to minimize it will keep them these two cells close together . And if we keep them together then in some sense routing becomes easier and that is why the total wire length of the design is tried to be meaningful or that is the metric that the on which the placement works is primarily the total wire length of our design . And there are other metrics also that are considered in placement for example, timing and congestion . Now placement is a complicated task and therefore, it is done in various stages .



So, there are three primary stages or tasks or subtasks in placement and those are global placement and legalization in detail placement and post placement optimization. So, in global placement what is done is that the cells are spread over the core area . So, suppose this is our chip and this is the allocated area for the standard cell then in global placement the cells are distributed or they are placed on the core area such that they are well spread out and also the wire length between them is tried to be minimized. So, the total wire length of the design is also tried to be minimized .

Now why do we want to spread out the cell ?

So, the over the cell core area. So, the reason is that if we spread out the cell then what are the routing resources that can be utilized better . So, understand that when we have cells over them we have say 10 or 12 or 15 layers of interconnects . Now these layers of interconnects are available throughout our chip layout . Where those are not already utilized for example, for power lines or other other signals wherever they are not utilized that are available for the routing.

So, routing resources are fairly well distributed over the layout and if we place cells well distributed over the layout then routing resources become easily utilizable and that is the motivation of the global placement to spread the cells evenly or well over the core area ok. And then during global placement the dimensions of the individual cells are not considered and therefore, there can be overlap between cells and therefore, in the next step we carry out legalization and detailed placement. So, in legalization overlap between the cells are removed by moving them further apart and cells are moved to legal positions. For example, there can be some locations where the cell cannot be there. We remove those cells and put them in legal locations. So, what are legal locations we will see in the later slides.

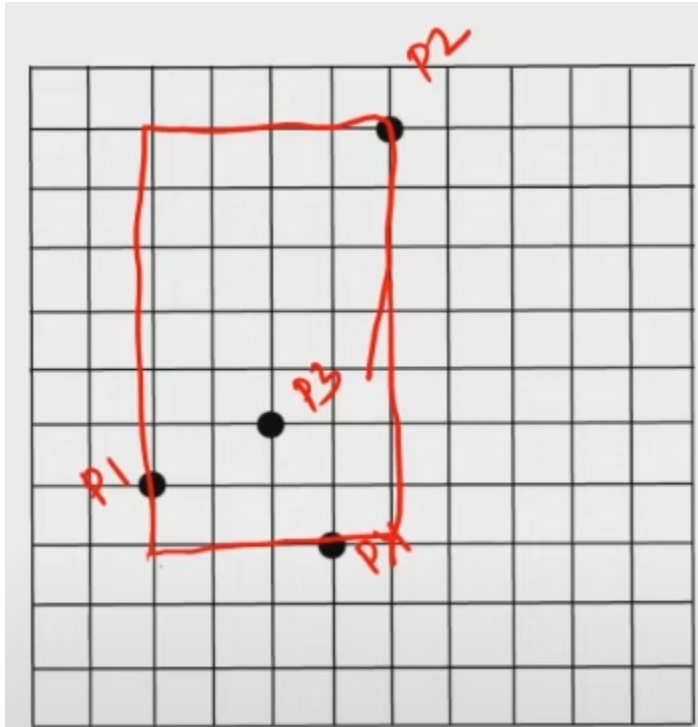
And then after we have done the legalization and placed the cells over the layout then we do some optimizations for example, buffering, resizing etcetera to improve the QR of the design. Then in placement the primary objective as I said was to minimize the total wire length of our design and therefore, for a given net each net in the design each net in the design the placement tool needs to compute its wire length or estimate its wire length. Note that it is estimating because the placement is not doing actual routing to compute to know what the wire length is, but using some estimates it has to get that whether it is a good location or not. So, when the placement tool is doing find the locations of the cell. So, during the execution of the placement algorithm it needs to determine whether a given location is good or not whether it will lead to a small wire length or a large wire length.

And therefore, wire length estimation is done many times for the nets in the in the in in in during placement. So, ideally the estimated wire length should match the post routing wire length, but this is a difficult task because we are not doing the routing during placement. So, therefore, there are many heuristics which are used to assess or estimate the wire length of the nets during placement and one of the most popular is half perimeter wire length estimation. So, in this method what is done is that it is it is a easy way to compute and it the half perimeter wire length is basically the half of the perimeter of the bounding rectangle that encloses all the pins of a net for suppose we are trying to find a wire length for a given net. So, each net will have 2 or more than 2 pins.

Now, we first find a bounding box which contains or the smallest size bounding box which contains all the pins in that net. And then what is the half a half perimeter of that bounding box, bounding box meaning bounding rectangle. So, what is the half perimeter of that bounding rectangle that is considered as the half perimeter wire length estimates of a given net. So, this is the most popular method of estimating wire length during placement because it is very easy to compute just knowing the coordinates of the pins of the of the of a given net we can easily compute the half perimeter wire. It is very easy to compute and it is fairly accurate also it is for some type of nets it is it is accurate for example, for say 2 pin and 3 pin nets it gives very good it is accurate for other big other other types of nets also it is it also gives a fair fairly good estimate of of the wire length.

Now, let us take an example and understand how the semi half perimeter or it is also called a semi perimeter wire length of a net is estimated. Suppose that there is a cell, it is an inverter and it is driving say 3 different inverters through a net. So, this is the net n now during placement we are trying to find the wire length after we have placed the entity. So, we know the placement of this this this this and we want to now estimate what is the wire length of this net n. So, we know the location of pin p1 p2 that is the input pin of these inverters p3 and p4.

So, let these locations of p_1 p_2 p_3 p_4 be as shown in this figure. So, let this be p_1 this is p_2 this is p_3 and this is p_4 . Now, we are asked the question what is the semi half perimeter wire length estimate for this net n . So, first what we need to do is that we need to have a bounding box or rectangle which encloses all these pins. So, this will be the bounding box.



So, we have created the bounding box by looking at the minimum and the maximum coordinates of these points. So, this is the bounding box we have got. Now, for this bounding box we compute the half perimeter. So, half the perimeter means that if we count from here to here either this path or this path it will give the same length. So, we count it 1 2 3 4 5 6 7 8 9 10 11.

So, the semi perimeter or half perimeter wire length estimate for this net n is 11. So, this is how the wire length estimate can be made for a given net. Now, this is a very simple method or half perimeter wire length method is the most simple and widely used. Also there are more sophisticated methods of finding a good wire, a good wire or more accurate wire length estimates. So, now how do we do this? So, we have to do the placement. So, what is the technique of doing global placement?

So, the placement problem has been studied for a long time, maybe since 1970 and it is

known that it is a difficult problem. It is an NP complete problem . And therefore, a lot of heuristics have been developed over the last 50 years or so for the placement problems . So, various heuristics have been proposed and these heuristics work well in some designs and may not work very well in other kinds of design. So, typically these days the most widely used heuristic of placement for placement is what is known as analytical placement algorithm . So, in an analytical placement algorithm or the way of or the method of placing cells analytically what we do is that we consider each cell as a point of .

So, we can do for example, if this is a standard cell we consider for example, the left hand corner bottom corner as its coordinate or may be the center of it . Center of the cell can be considered as the coordinate of the of the of the cell and we ignore its length and height and so on. We consider the cells as point objects and assume that it is at some coordinate, say x_i, y_i . So, if this is the i th cell then its coordinate is say x comma y and then what we do is that we evaluate the cost functions such as the total wire length and the constraints mathematically . For example, we can consider the wire length as a function of the Euclidean distance. Suppose there were two cells: one cell was at say x_i, y_i and another cell was at say x_j, y_j .

So, what is the Euclidean distance between them is x_i minus x_j whole square plus y_i minus y_j whole square and a root over that is the Euclidean distance and if we take the square of Euclidean distance this root over will go up .

$$\text{Wirelength a function of } (x_i - x_j)^2 + (y_i - y_j)^2$$

So, in analytical placement algorithm we take the we consider wire length as a function of x_i , minus x_j whole square plus y_i minus y_j whole square and then based on that we compute the total wire length or find the mathematical equation or math formula for the total wire length of our of our design and then use efficient solvers to obtain the minimum cost for the mathematical formula for the simulation . Now, this is for these two cells which were connected similarly there will be other cells which are connected and then we sum it sum over sum take the sum of all these wire lengths and give it to a mathematical solver . So, formulate this in terms of say matrix equation and then give it to the matrix equation to a solver which will give us a final result that these are the x_i , and y_i coordinates for the cells . So, this is what an analytical placement algorithm will do .

Now, if we directly give say minimize x we give a function the total wire length ϕ which is a function of say x_i , minus x_j whole squared plus y_i minus y_j whole square and similarly other terms will be there then what will be the minima for this what will be the minima the trivial minima is of course, when we say x_i, x_j is equal sorry x_i, y_i is equal to 0

x_j, y_j is equal to 0. If we place all the cells at say one point in our design we will get the wire length as 0 and that is the mathematical minima and this is of course, not the solution why we want we do not want this kind of solution why because if we place all the cells at one point then the routing may not be possible the routing resources are not utilized over the layout and so on. So, in the mathematical formulation we need to modify it somehow and what modifications are required. The first is that there should be suitable constraints for the fixed entities and what are these fixed entities. So, remember from last lecture that we have decided the locations of the macros and the i/o cells and if we have decided the location of those entities those entities become fixed entities for the placement. And therefore, those are the locations of the fixed entities which are connected. For example, there is an i/o cell which at this point we decided its location during chip length.

Now, if the cell is connected to this i/o cell then it needs to be connected to be brought closer to it it is not it cannot be arbitrarily pulled out anyway. Similarly, suppose it is connected to another cell and this cell was connected to say another outputs i/o cell and this will pull the cell in the other direction. So, the constraints of the fixed entities will modify the solution and that must also be given to the mathematical solver or that or that that should also be taken into consideration in the mathematical formulation. And also we want that these cells be distributed fairly evenly over the layout and therefore, we add some cost terms which are related to cell density. So, we can add some term in this mathematical formulation such that if the cell density is high then that solution becomes penalized.

So, these modifications are done by the analytical placement tool while formulating the mathematical model for the placement problem and then giving it to a solver to obtain the result. Now, the cells allowed during placement sorry during global placement the cells are allowed to overlap we are considering cells as point objects we are not considering them as say their width and the height we are the length we are not considering in the in this formulation and therefore, there can be overlapped between the cells. And therefore, legalization becomes necessary after doing global placement. So, therefore, after global placement we do legalization. Now what is an illegal location and placements of a cell is illegal under a few conditions. What are those conditions? So, first is that if there are overlaps for example, if we take this is one standard cell and this another standard cell and there is an overlap between these two and therefore, this is an illegal placement for these two cells.

And sometimes what can happen is that cells occupy illegal sites for example, between the placement rows for example, if we consider this one this is not on the rows that we

have assigned during the or rows that we have allocated during the chip plan. So, this cell is out of that row and therefore, this is an illegal location and therefore, in legalization what is done is that it removes all cell overlaps and snap cells to legal sites. And during this it is tried that the minimum impact on wire length is there, timing is there and congestion is there. So, the legalization process basically tries to mix incremental changes or many or minute changes to the placements as the problem of overlap and cells being in the illegal location is solved. It does not try to improve the or or change the placement solution significantly during finding its or during legalization.

For example, in this case if these two cells overlapping and there was an empty space alongside it will simply move this cell on the and it is the cell this cell becomes legalized. Similarly, this cell was out of the row it just simply pulled it into the row and that is and the solution becomes legalized. So, legalization does not attempt a large change in the placement, only snaps the cells or moves the cell by a small amount. So, that the cells become legalized and after legalization then a detailed placement is done in whose purpose is to basically improve the QR by incremental changes to the cell location. So, during detailed placement some changes are made to the placement such that the QR of the design improves and so, it can improve the wire length and routability by making some changes.

And what are these changes? It can swap the location of say two adjacent cells or neighboring cells or it can redistribute the free size over the layout. So, that routing resources become available to the other area or other region also and therefore, routability can improve or it can move cells to unused locations. If the cell density is found to be large in some area then it will try to move away from that location and put it in some unused location. Again these changes are made in an incremental manner. So, the solution that was given by global placement was that those are only perturbed incrementally during these stages.

If we make larger changes then it can perturb the design too much and can lead to design closure problems. We can just do it and it can make the solution make the placement solution worse than what it is started with. And therefore, detailed placement does not try to make big changes in our placement solution. Now, we have discussed that the placement tries to improve the total or minimize the total wire length of our design. But for high performance designs and for industrial design the timing driven placement is also used often.

So, in what timing driven placement is done is that the placement tool internally has got a timer or the timer basically that the tool that does static timing analysis. So, what the placement tool has, it has got an internal timer and it performs the timing of a design

incrementally . Meaning that if it changes the locations by small amounts then it can assess the impact of that change on the time . And based on that incremental and an internal timing analysis the tool can identify or can evaluate what is the slack of various paths and there can be some paths which are violated . So, therefore, during placement the paths which are violating where there is negative slack then those paths can be targeted by or those paths can the timing of those paths can be improved by changing the placement solution .

So, the timing driven placement basically targets either the worst negative slack . The worst negative slack means that given a design what is the worst, what is the slack or the the negative worst negative slack of our design . There will be many paths which may have a negative slack out of the one which is having the most negative that is the worst negative slack. And that is basically the path which must be targeted first . And then there are place timing driven placement tools which also target total negative slack.

So, there may be thousands of paths which have negative slack . In that case just solving the problem of top negative slack or the worst negative slack may not be fixing our design properly . In that case we try to improve the or or improve the total negative slack of our design. Total negative slack means that we take the sum of the negative slack of all the points and all the end points which are violating the time . And how does the timing driven placement work? So, the timing driven placement basically controls the proximity of the cells that are on the critical path.

So, if the timing driven placement finds that this is the path which is timing critical then the cells of those paths are placed close together . If the cells are placed close together then the wire length between the cells can reduce and therefore, the timing or the interconnect delay can be reduced for those that particular path . So, depending on the placement algorithm different approaches can be used to obtain timing driven placement . For example, during say analytical placement what can be done is that we put additional weights to the nets to indicate timing criticality. For example, suppose there were two cells x this has location x_i, y_i and this has got x_j, y_j .

So, the wire length between these can be estimated as the Euclidean distance between them that is x_i minus y_j sorry x_i minus x_j whole square plus y_i minus y_j whole square . Now, in addition to this if we find that these two cells were on the critical path we can say multiply by a factor of this in the cost function we take it as 10 times of what this cost is . Rather than multiplying by 1 or taking a function of the simply the Euclidean distance square we multiply a 10 times of that . So, if we add additional weight to the nets to indicate timing criticality. So, if we give more weights to nets that are timing critical then the additional weight will bias the solution of the placement solution such

that these two cells are placed together .

So, these are these two. The placement engine will automatically place these cells together because the weight of the nets corresponding to these two cells will be much higher compared to other nets in the design. And therefore, the placement solution when the solver will try to find the placement solution in which these two cells are close together will be favored and the timing of this particular path will be improved. Now, in earlier lectures when we were discussing test DFT in that we discussed that one of the most widely used methodology for testing is scan based method. So, in a scan based method just to recap what we do is that we create scan chains, we replace flip flops with scan cells and we create chains of or shift registers consisting of scan cells in our design . And when we did DFT at that time we did not have the location information .

So, we did not we do not have the information of location of any of the flip flops or scan cells in the design. And therefore, the order of the scan chain means that in the scan chain which flip flop comes first then which and so on. The order of the cells or the flip flops in the scan chain was pretty arbitrary without knowing the placement of the design. But after we know the placement of our design then we can do some optimization. What we can do is that we can form a scan chain or order such that scan cells that are nearby they form they form consecutive flow they are consecutive flip flops in the scan chain .

So, let us take an example and understand what can be done. Suppose we did a scan, we built a scan chain initially, initially we did not know that initially we did not have the information of the placement of the scan cells or flip flops . So, we made a scan chain like this . So, from the scan it is going to say FF1 then it goes to FF2 then it goes to FF3 and FF4 then FF5 and FF6 . Now, what is the problem with this scan cell ordering? The problem is that it will consume more routing resources than it is necessary. For example, there will be a net which will be going from say FF1 Q pin and into the scan input pin of FF2 which is a scan cell .

So, all these FF1, FF2, FF3 those are scan cells and they were replaced during DFT . So, if we consider the routing resource utility, this scan cell ordering is not a good one. So, what can we do? So, after we know the placement, after we have done the placement we know the location of each scan cell and therefore, we can reorder this scan scan chain such that routing resources are utilized more efficiently . So, we can reorder something like this . So, scan in from here it goes to FF1 then it goes to FF3 rather than FF2 it goes to FF3 and then FF5 then FF2, FF4 and FF6 and finally, to scan .

So, it can easily be seen that the length of the wires in this case will be much greater than in this case and in that sense the routing resources will be utilized more efficiently

and more efficiently in this solution . So, we will do this scan to cancel ordering after we have done the after we have done the placement . Note that of course, the test pattern for this case and this case will be different, but then the test pattern can be derived later on after scan cell ordering. Now, in the VLSI design flow we have seen that we do the implementation and we also do the verification. We are very careful that the design functionality remains intact, but sometimes what happens is that a bug is found at later stages.

Later stages meaning that after we have done the tape route after we have created the mask . So, after we have created the mask and we have done the fabrication and at the post silicon validation stage we are getting some bugs . Now, if we are in that situation , what can we do? So, remember that for fabrication we need say 20 to 40 layers of mask and these masks are very costly. Now, if we find a problem after we have already made the mask then we have to do a respin, respin meaning that we need to read, make changes in our design and then do the fabrication again using a new mask and so on.

Now, the cost of this respin can be reduced if we say that out of say 20 to 40 layers we say bottom 10 layers which are related to the devices with those are not impacted. We need not change that . If we are able to do that then the cost of respin can be much lesser than what it would be if we need to make changes in all the layers of our design . So, the spare cells are basically used in our design to help us avoid the entire respin of all the layers of the mask . So, what we do is that during designing we put on the layout some spare cells, those are extra cells in the anticipation of later use. So, those are not connected, those are not doing any functional functional function or those are the functionality of spare cells at that point of time when we are doing a tape out that is not important.

But later on when we need to probably do a respin then we can use these spare cells and to fix the problem in our design . Now, what do we save if we use the spare cells? Now, spare cells are already there in the layout meaning that those cells and the bottom layers which are of the design bottom layers of the design where the cells are which are used for fabrication of transistors and the cells need not be changed because the spare cells are already there. Only what we need to do is that probably make some make changes in the interconnections of these cells our rewiring of these cells and those rewiring of the cells can be done by just making changes in the top layer topmost layers of or few layers of the interconnects. And therefore, the cost of respin can be reduced if we are able to find spare cells in our design which are not used, but in a, but later on when a bug is found we can use them and if it can, those cells can deliver the functionality ok. So, after fabrication if it is required that some cells need to be connected then spare cells can be used because they are already there in the layout and the and though the features of the

of the spare cells are already there on the device layers or the mask of the corresponding to the device layer and those mask need not be changed.

So, only the topmost top metal layers might need to be changed and therefore, the cost of respin can be saved. Now, we said that ok the spare cells we will add some spare cells in our design. Now the question is where to place those spare cells. So, ideally those spare cells should be placed where we think that those might be used, but why when we do not know the problem itself and in the first place it is very difficult to anticipate where those spare cells will actually be used. So, we can think of that if some module is doing a lot of complicated functionality then probably there may be some problem in that module or it is more likely that some problem can be found near that in that module and therefore, we can place spare cells where we anticipate that it will be required.

If we cannot anticipate where the spare cells can be required then we place them randomly over the unused placement area and then probably when some cell will be required for making interconnection then we can find some spare cells in its vicinity because we are we have we have we have put many spare cells over the layout distributed randomly. But one important thing that needs to be considered here is that these spare cells when in the layout are unconnected. If they are not used in our first layout they are unconnected and we need to do tape out with those unconnected spare cells, but what can happen is that sometimes the design tools can consider these spare cells as cells which are of no use and they can optimize them out to save the area of the cells and so on. And therefore, it is important to instruct the physical design tools and other tools where those that these are the spare cells on the layout do not remove them out. Those are those that are not used for now, but they might be used later on and therefore, they must not be optimized out.

So, we must be careful that our design tool is not optimizing these spares. Now, if you want to look further into the topics that we discussed in this lecture you can refer to this book. Now, to summarize what we have done in this lecture that we have, we have discussed placement. So, we have looked into global placement and how to show what is one of the techniques of doing global placement. Remember that global placement can be done in many ways. I have just taken one example, an analytical placement method of placing standard cells.

And then we looked into what are wire length estimates, then we looked into the detail placement and legalization and detail placement and we looked into some optimization related to say scan cell and also the placement of a spare cell. So, there are after placement there could be some other optimization task that can be done for example,

buffering and the and the resizing of cells to improve the timing and so on . In the next lecture we will be looking into clock tree synthesis. Thank you very much.