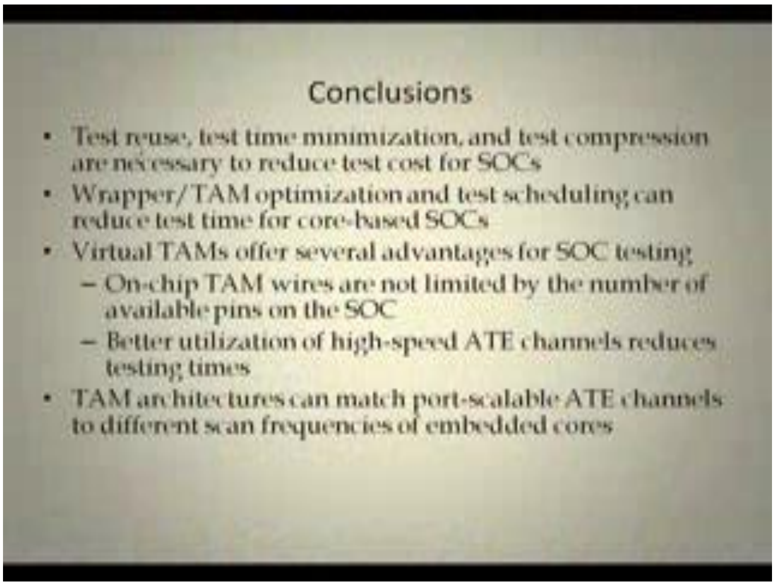
Digital VLSI Testing
Prof. Santanu Chattopadhyay
Department of Electronics and EC Engineering
Indian Institute of Technology, Kharagpur

Lecture - 51
System/Network - On - Chip Test (Contd.)

You have seen several techniques for SOC testing, and there are various problems like designing the wrapper itself and then designing the TAM part and then doing the scheduling. And if you try to exploit different features of the ATE test equipment then again there are variations in terms of the operations, the schedule, the architecture that we can get then the operational flexibility that we can get, and accordingly we can use different techniques that we have seen for the SOC testing.

(Refer Slide Time: 00:58)



Conclusions

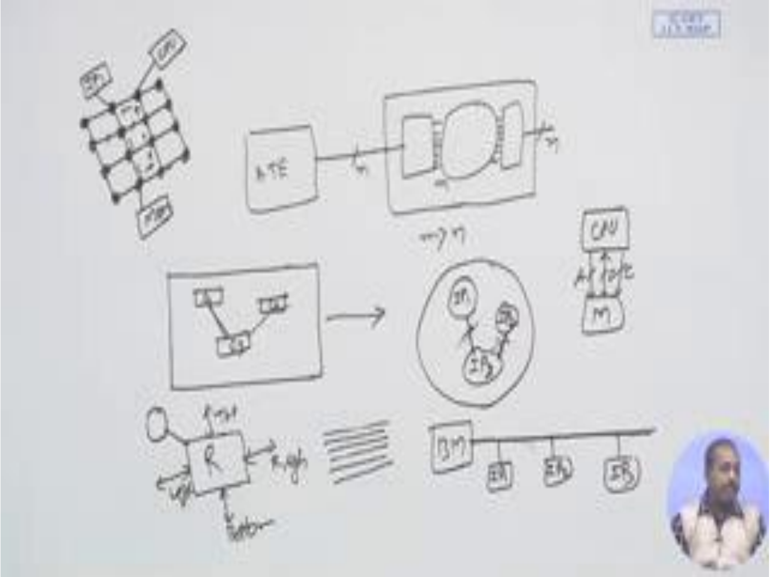
- Test reuse, test time minimization, and test compression are necessary to reduce test cost for SOCs
- Wrapper/TAM optimization and test scheduling can reduce test time for core-based SOCs
- Virtual TAMs offer several advantages for SOC testing
  - On-chip TAM wires are not limited by the number of available pins on the SOC
  - Better utilization of high-speed ATE channels reduces testing times
- TAM architectures can match port-scalable ATE channels to different scan frequencies of embedded cores

So, to conclude this SOC testing part; so we can say the test reuse test time minimization and test compression are necessary to reduce test cost for SOCs. This is one very important issue. Test reuse that same test if you can use for multiple cores, then this may be done there are some works that talk in that direction combining test patterns for different cores and then sending those test patterns, so that way that test time can be minimized. Then the compression; sp compression will be useful because this tester memory requirement will become low. And this transfer time from of patterns from tester to the chip so that will also get reduced. So, that way we can have this thing, this compression is going also going to help us in reducing the test time.

Then there are wrapper or TAM optimization and test scheduling. So, this can reduce test time for the SOCs. So, wrapper optimization we have seen that giving different width to the wrapper of chain, of this cores. So, they can have different types of test times. The test time may get reduced. And this TAM optimization has to be done because this TAM we can do a partition TAM so we may not go for partitioning. So, we can go for different type of scheduling policies, we can go for different TAM architectures like say virtual TAM, like then this multi high speed TAM, multi rate TAM. So, like that we can go for various types of TAM architectures that can exploit the features of the ATE.

So, in virtual TAM it offers advantages because on chip tamers are not limited by the number of available pins on the SOC. So, this we can this on chip from the ATE we have got a number of pins coming.

(Refer Slide Time: 02:53)

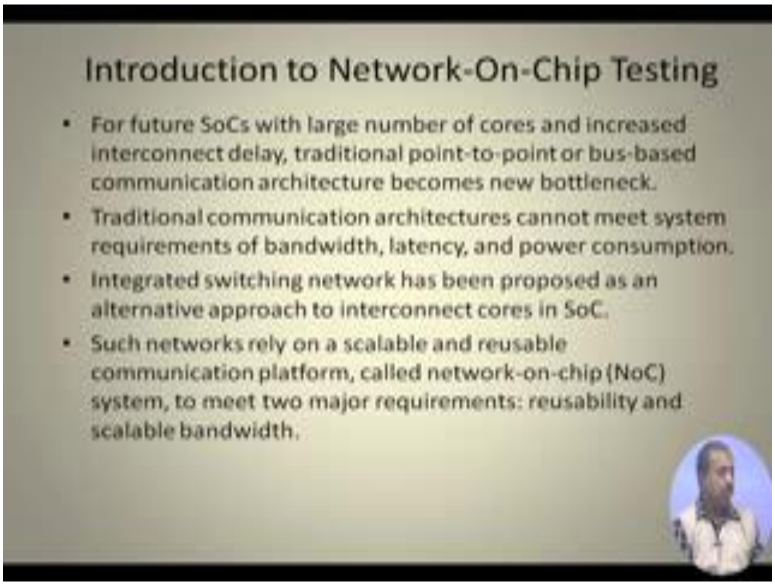


So the structure was like this; if this is the ATE from this ATE say suppose say n number of lines are coming then when it when it is reaching the SOC, so it has SOC has got n number of pins here but here I have got that expanded; that serial to parallel converter. So, the n line coming here is expanded in to a number of lines. So, that is actually used for testing the cores that are there in the system. And then all of them are again combined in to n. So, this is combined in to n and it is going back to the ATE. So, if this here the number of lines is m. So, we have seen that virtual TAM structure can make m to be larger than n.

So, from the SOC pin side SOC pin remains at n, ATE channel remains at n, but inside the for scheduling we see that it is increase to m. So, that way this test scheduling can be benefited. So, it results in better utilization of high speed ATE channels, it reduce that that is how it reduces the testing time. And this port scalability channels so that they can be made to operate; ATE channels can be made to operate at different frequencies. So, if embedded cores are of different frequency the maximum frequency at which their testing can be done are different then we can try to exploit those information and then we can try to design the TAM so that this ATE channels are utilized. So, high speed ATE channels can be utilized for applying test patterns faster to the chip.

So, after looking in to this SOC testing techniques and all that next we will look in to another topic which is known as network on chip.

(Refer Slide Time: 04:36)



Introduction to Network-On-Chip Testing

- For future SoCs with large number of cores and increased interconnect delay, traditional point-to-point or bus-based communication architecture becomes new bottleneck.
- Traditional communication architectures cannot meet system requirements of bandwidth, latency, and power consumption.
- Integrated switching network has been proposed as an alternative approach to interconnect cores in SoC.
- Such networks rely on a scalable and reusable communication platform, called network-on-chip (NoC) system, to meet two major requirements: reusability and scalable bandwidth.

So, what is network on chip? So, in SOC what we have done is that the basic philosophy behind de designing SOC was that; if I have a board and in this board if I have got this individual chips then the problem was that this off chip communication was problematic. So, off chip communication was much slower than the on chip communication. Now when this is translate it in to SOC design, what is happening is that we have got our silicon wafer and on to this silicon wafer all this components are fabricated; so all this chips; so this is a c 1, c 2 and c 3.

Now I get the corresponding IPs from the vendors, so this IP 1, IP 2 and IP 3 so they are fabricated on this silicon floor. And then the connection between them so they are also onto this like here I have got connection from c 1 to c 3 here IP 1 to IP 3 I have got a connection, there will be another connection from IP 2 to IP 3 where there is a board level connection from c 2 to c 3. So, the advantage that we get is this off chip communication that we had from c 1 to c 3 or c 2 to c 3 they are getting translated in to on chip communication. So, they are going at a same rate as the original one.

So, the rate at which this individual IP cores can operate this channels can also transferred data. So, this signal lines they can transfer data at the same rate. So, the delay is much much reduced compared to the off chip connection. However there are problems in the sense that you see there will be large number of lines that that are running in the system, because if say IP 1 is a CPU and IP 2 is a memory. Then what is required? From CPU the address bus will come, the data bus will come and the control bus will come. Now this address bus data bus, so for modern processors they are very wide as a result you see there will be large number of lines that needs to run parellely between this CPU and memory. So, this line that I have drawn from IP 1 to IP 2 if IP 1 is CPU and IP 2 is a memory then the width of this line is significant. So, this is the sum of the width of this address data and control lines. So, that way it is going to be large number of lines that we need.

Now in this case when there are a large number of lines running parellely. So, I have got large number of lines running parallely now we have got the problem of cross stop, we have got the problem of interference between the lines. And that way on the board level we had this problem but it was not that much of an issue because this connections. So, the system was operating at a much lower frequency. So, the delay of these channels this external channels they were deciding the delay of this communication, delay of this in to the frequency of this entire board.

And when you convert it in to SOC the basic idea was to have a better speed performance. So, I want to increase the frequency to say gigahertz range. Now the individual IPs that are designed, so they are designed very efficiently and suppose they can operate at the gigahertz range. But the communication part they cannot operate at gigahertz range due to these problems. With large number of lines running parallely, so we cannot have this type of situation.

So, what is required? As a solution there can be a standard solution is to have a bus and from this bus all this IPs will be hanging. So, this is IP 1, this is IP 2, this is IP 3, so all of them are hanging from this bus. And there must be a bus master will be there, and whenever any IP wants to communicate with another IP so it will tell the bus master that I want to use the bus. So, say that way thus bus master will generate a (Refer Time: 09:07) and these two IPs will do the communication between them through this bus.

Advantage: we definitely save on the number of this lines that are running throughout the chip, because all this lines will be replaced because now they everything will be happening over this bus. And this IPs they need to communicate some sort of coded message through this bus for this address data and control say IP 1 is CPU and IP 2 is memory so they will do this thing.

So, that makes the process regular and we can have multiple levels of buses and all that, but at the same time this bus based communication it has got limitations. So, you cannot go on hanging devices from the bus. So, that will increase the capacitance of the bus and that will also the registers of the bus, so delay of the bus will became unacceptable after we have connected a few devices to the system.

On the other hand we want the maximum parallel communication. Another problem with bus is that at one point of time only one communication can go through the bus. So, there is a communication bottleneck. So, I cannot have more than one communication going on simultaneously on a bus. If I got a hierarchical bus then to some extent we can have this parallelism, but still if the if there is a communication across this buses then one such communication will install all other communications. So, that way we have got problems with this bus based communication mechanisms. So, what we want is something like a point to point sort of communication where every module can communicate with every other module directly, but that will require large number of signal lines running through the chip.

So, we need some solution; we need some solution to this problem and this network on chip is a solution that has been proposed for designed for solving this communication problem in the SOC. So, it is otherwise a system on chip design; however the communication procedure between the chip between the cores they follow a particular paradigm. So, we will see how this can be done. So, for future SOCs with large number of cores and increased interconnect delay traditional point to point or bus based communication architecture becomes new bottleneck. So, I have said that the bus has got it cannot do parallel communication more than one communication cannot be done simultaneously, and point to point means there are large number of lines running through the chip. So, both of them are problematic.

And compare to a board a chip has got much lesser area and this with increasing device density more amount of logic will be put in to the system so there will be more amount of communication between these modules that we put on the chip. So, that way we can have this communication architecture that becomes a bottleneck that has to be resolved.

So, traditional communication architectures cannot meet system requirements of bandwidth, latency and power consumption. So, bandwidth; you cannot transfer at a very high data rate. Latency: one communication will take more time because of that delay that is increase incurred in the bus. And power consumption: so if you have got point to point communication the power consumption will be very high. Similarly bus also the power consumption is very high. And there are several techniques where the bus encoding strategies that has been proposed, but incorporating them will required extra hardware and that will add their complexity to the system.

So, we have got these problems this with the traditional communication system that do not meet the requirements of bandwidth latency and power consumption. So, integrated switching network it has been proposed as an alternative approach to interconnect cores in SOC. So, what it says; if there will be a network in the system, in the chip itself and this component that need to communicate between themselves they will do that using the basic network that is available on the chip.

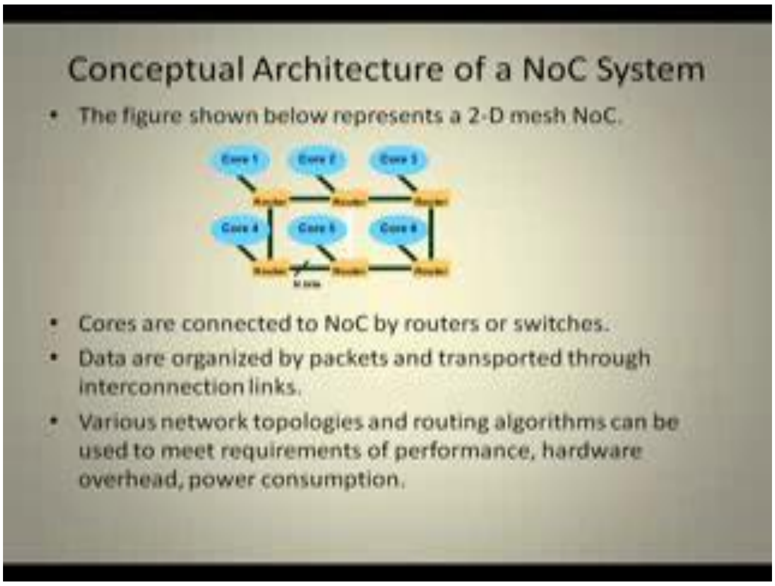
Idea is that similar to this say computer network we have got large number of computers to be connected. So, one possibility is that we can put all those computers on a bus and drop them from a bus, bus based communication. But the difficulty is that so creating such a bus is a problem. So, in a distributed processing environment what we do? We create a backbone network and from that backbone this individual computers may be dropping or it may be each backbone may be feeding sub network and from the sub network we have got the connection to the individual computers.

So, similar such situation can be done like just like distributed computing. So, this backbone design is such that they can do communication between this within the backbone very easily and very at a very high data rate and then the rest of the communications are taking place. The same thing is duplicated here in a just like in a distributed processing environment so here also it is duplicated like that.

So, there will be an interconnect integrated switching network that is incorporated in to the chip. And this networks they rely on a scalable and reusable communication platform which is known as network on chip or NOC system. To meet two major requirements: reusability and scalable bandwidth. So, reusability means the same channel can be used for doing communication between a number of such IP cores; definitely not at the same time, but over the time so they can be used for doing this communication and the bandwidth.

So, as you are adding more and more IP cores then this bandwidth utilization will improve till a saturation point is reached, after that if you increase that network size again the bandwidth utilization will start increasing.

(Refer Slide Time: 15:25)



Conceptual Architecture of a NoC System

- The figure shown below represents a 2-D mesh NoC.

- Cores are connected to NoC by routers or switches.
- Data are organized by packets and transported through interconnection links.
- Various network topologies and routing algorithms can be used to meet requirements of performance, hardware overhead, power consumption.

So, this is the typical structure. So, we have got this routers and the inter connection between them so they actually form the backbone network. In this particular diagram we have got six routers and they are connected in a mesh fashion. There is a two dimensional mesh fashion in to which they are connected. And from these individual routers the cores are connected. So, this is a 2-D mesh this is basically a 2 by 2 mesh now we can have a mesh of more dimension like; you can say that I will have a 4 cross 4 mesh where the situation is like this.

So, this is the 4 cross 4 mesh and then each of this corners we have got one router. So, we have got the routers at these points. Now from a router there are four channels that are going out in general. So, there are one going to the left neighbor, one to the right neighbor, one to the top and one to the bottom. Say this is a route so it has got four neighbors like that. So, these channels they are called global channels and these channels are parallel channels.

Now if I want to connect some IP core to the system, so I can do it I can connect it to the router like with this router I can have some core IP 1 connect it here, so may be my CPU is connected here, the IP which corresponds to the CPU is connected here, and the IP for the memory is connected there. So, this is the memory.

Now this CPU when it wants to get some data from memories so what will it do; it will not send this address data control lines to the memory as some signal lines rather it will prepare a message in that message it will tell the it will put the memory location address from where it needs to get the data. And it will send that message over this network. So, there has to be some routing policy, so this individual routers they will follow some routing policy. Every simple routing policy may be the x y routing so that any message that is to be transmitted is transferred in the x direction first and then in the y direction. So, it will go like this then it will come down like this and it will finally reach this memory.

So, on getting this message this memory will prepare a written message, and this the written message it will have the content of the location and then it will send it to this CPU and now may be it will go like this in the x y routing fashion; first in the x direction and then in the y direction. So, this way these cores may be communicating. So, you see that total electrical signal exchange between the IP cores it is modify to message exchange between them, and this communication is taking place via this interconnection network that we have and the routers and the links.

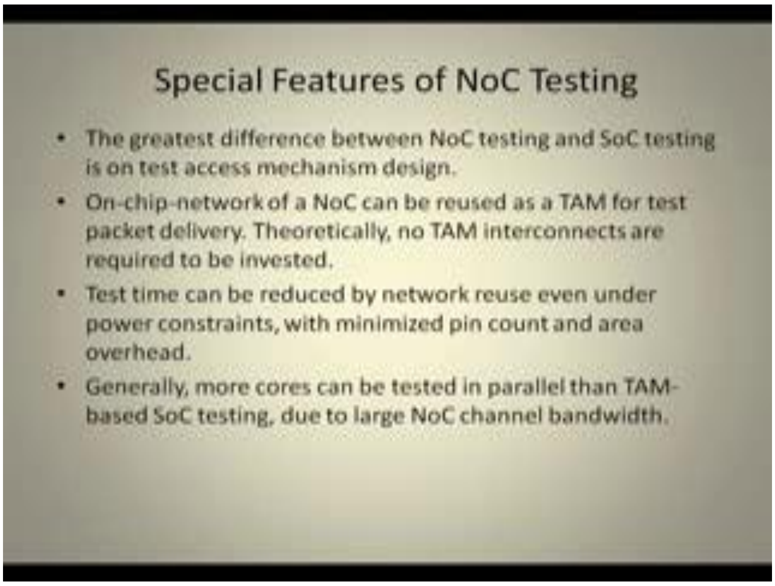
Now, if though the mesh is a just an example so you can have many other topologies; like you can have a ring topology, you can have a star topology, torus topology, then there are tree topologies, butterfly factory topologies. There are many such topologies that are been proposed, but mesh is the most popular one because this is a very regular one and this design is very simple. And it has got many other features that you can find out if you are interested you can look in to some distributed processing literature. And there you can find the benefits of having mesh; so why mesh is so popular, but what does happen is mesh has become the most popular architecture.

So, every router if you see in the in more detail; if this is a router so this router has got a number of ports. So, this is called the top port, this is the bottom port, this is the bottom, then we have got left and right; left and right so they are going to the phone neighbors of the process of the routers. So, this is the router R. And there is a local port with the routers, so this is the local port of the router to which the core will be connected. So, it is not mandatory that all routers will have cores associated with them. So, that is again another problem of this network on chip design like how can we decide which cores be assigned to which router, cut we will not go in to that.

So, basically to say finally when the designer has finished, so designer has done it like this that this cores are attached to different routers. So, cores are connected to NOC by routers or switches and data are organized by packets and transported through the interconnection network. So, message that has been formulated so it is divided in to packets and the packets will travels through this network to reach the destination.

Various network topologies and routing algorithms can be used to meet requirements of performance hardware overhead and power consumption. So, this is there are various topologies that have been proposed and then again there are innumerable number of routing algorithm that have been proposed. So, x y is just one of them, then you can have some table driven routing some fault tolerant routing, there are many such policies that have been proposed in the literature.

(Refer Slide Time: 21:06)



Special Features of NoC Testing

- The greatest difference between NoC testing and SoC testing is on test access mechanism design.
- On-chip-network of a NoC can be reused as a TAM for test packet delivery. Theoretically, no TAM interconnects are required to be invested.
- Test time can be reduced by network reuse even under power constraints, with minimized pin count and area overhead.
- Generally, more cores can be tested in parallel than TAM-based SoC testing, due to large NoC channel bandwidth.

So what is the difference, like when we are coming to the testing point then what is the special feature that this NOC testing has? So, the greatest difference between NOC testing and SOC testing is on test access mechanism design. So, you see in case of SOC the SOC testing the problem was that to send the test pattern to a particular core I had to; one possibility was that I can use the functional path passing through different cores. And then the associated difficulty that we had is we do not know the details of the individual cores so we cannot find out the excitation sequence by which the particular pattern can be send through that core.

So, in case of NOC what has happened is; so that problem is a bit resolved. In SOC we solved that problem by having dedicated test bus or test architecture embedded in to it and then from this input port we are from the ATE we are feeding the test patterns to the to the test architecture and the responses are obtained. So, that way it was solved by having some additional test architecture embedded in to the SOC.

On the other hand this NOC we can utilize the on chip network itself. So, with already a network is existing through which I can possibly reach any of the cores without going in to other one. So, by traversing through this router network I can reach any of the cores in the system. So, this on chip network of the NOC can be reused as a TAM for test packet delivery. And theoretically no TAM interconnects are required to be invested. Of course, you can say that it will go through the router network. So, that way it will require some time to go compare to our test architecture test bus architecture where it will be direct, but so TAM architecture wise it will be direct; but that is an overhead. So, if we are happy with the underline this router network then we can use it for transferring the test patterns through the router network.

So, test time can be reduced by network reuse even under power constraints when minimized in count with minimized pin count and area overhead. So, test time can be reduced because if the ATE can supply test data parallely to over two different channels then we can send we can feed that test data to two different cores and they can be tested in parallel. So, that way we can do a parallel testing of these cores that way it can reduce this testing time.

Power constraints: so individual cores their power values are there so we can follow that we can restrict ourselves within that power constraint. Pin count is minimized because what is required is that we can transfer it to any of the router and from there it can be send to the desired port which we want to test. And generally more cores can be tested in parallel than TAM based SOC testing. So, TAM based problem SOC the problem was that we cannot test two more than one core part TAM, and if the TAM lines for a partition TAM arrangement. So, if more than one TAM cannot be tested simultaneously on a TAM line. And if it is a flexible TAM architecture then at for a TAM to be for a core to be tested all the TAM lines required by this particular core must be free.

So, these types of limitations were there, so what we can do we can in case of NOC based architecture we can test more number of cores, because now none of them around TAM so it does just depends on the availability of NOC channels and the availability of this core the router path to the core for doing the test operation.