A Study of Network on Chip

Trima Piedade Fernandes e Fizardo
Assistant Professor
Department of Electronics and
Telecommunication
Don Bosco College Of
Engineering, Fatorda, Margao-
Goa-403602

ABSTRACT
A brief about System On Chip, and recent development of a new approach to System On Chip, that is Network On Chip(NOC) is given. Here discussion about the NOC methodology, design, Area and performance compared to traditional architecture based on fixed connectivity is made. Finally discussion on the Nostrum Concept is done, along with the advantages of Network On Chip and future scope.

Keywords
System On Chip, Network On Chip.

1. INTRODUCTION
A System-on-a-chip[1] or system on chip (SoC or SOC) refers to integrating all components of a computer or other electronic system into a single integrated circuit (chip). It may contain digital, analog, mixed-signal, and often radio-frequency functions – all on one chip. A System-on.Chip is characterised by high degree of integration on a single integrated chip.

A typical SoC consists of:One microcontroller, microprocessor or DSP core(s),Memory blocks including selection of ROM, RAM, EEPROM and Flash,Timing Sources including oscillators and phase locked loops,Peripherals including counters- etc,External interfaces and Analog interfaces.

Network-on-a-chip (NoC) is a new approach to System-on- a-chip (SoC) design. NoC based systems can accommodate multiple asynchronous clocking that many of today's complex SoC designs use[2]. The NoC solution brings a networking method to on-chip communication and brings notable improvements over conventional bus systems. Integrated circuit technology develops fast according to Moore's law and will do so for almost another decade. In a few years from now this will lead to the possibility of having billions of transistors on one single chip.

Using current methodologies, as shared buses to design communication platforms for systems of that magnitude will introduce a number of issues to solve, both logical and physical. Some of the important ones are the utilization of on-chip communication infrastructures, the increase of wire delay, the controlling of power distribution and power dissipation and also the ability to guarantee signal integrity. To solve these issues there exists a proposed solution It's called Network-on-Chip (NoC), which suggests that future on-chip communication should work as a network[3].

Another issue in future on-chip design is the level of reusability of Intellectual Property cores (IP cores).A poor reuse of IP cores leads to long development times and makes it impossible for the system designers to meet the market requirements of short time-to market. There is a suggestion under development for how to design a platform for on-chip communication on a NoC. It is called the Nostrum concept. Nostrum suggests that a NoC should be arranged as a mesh of switches, where every switch is connected to one resource and four neighboring switches. Nostrum also declares that the network protocols should work in different layers with different responsibilities, in a similar way as the OSI reference model.NoC's have been proposed to address challenges of increasing connectivity[7]

2. NETWORK ON CHIP
2.1 Intoduction to Network on chip
Today on-chip systems use shared buses as the most common communication infrastructure. But as semiconductor technology improves, more and more subsystems can be connected to each other on a single chip. Shared buses will then no longer be a good communication solution for large designs.

2.2 Reasons for using Network On Chip
There are both physical and logical issues that prevent shared buses from being the on chip interconnection solution of the future. Amongst the physical issues we find the increase of wire delay, the control of power distribution and cross- talk between wires. A logical issue is that if more subsystems use a shared bus the usage time of the bus for each subsystem will be less. That generates a low utilization level of IP cores that have to use the bus often. Therefore it has been proposed that NoCs should be used as on-chip communication solutions instead of shared buses[4,5]. Some scientists have proposed that hybrid interconnections solutions could also be used . A hybrid solution means that the interconnection solutions would be a mixture of NoC and shared buses. The main idea with NoCs, besides the solutions to the physical issues, is the possibility for more cores to communicate simultaneously, leading to larger on-chip bandwidths.The adoption of NoC architecture is driven by several forces: from a physical design viewpoint, in nanometer CMOS technology, interconnects dominate both performance and dynamic power dissipation, as signal propagation in wires across the chip requires multiple clock cycles.

NoC links can reduce the complexity of designing wires for predictable speed, power, noise, reliability,etc., thanks to their regular, well controlled structure. From a system design viewpoint, with the advent of multi-core processor systems, a network is a natural architectural choice. An NoC can provide separation between computation and communication, support modularity and IP reuse via standard interfaces, handle synchronization issues, serve as a platform for system test, and, hence, increase engineering productivity.

2.3 Different NOC Topologies
There have been proposed a number of different NoC topologies. They all have in common that they connect

resources to each other through networks and that information is sent as packets over the networks[6].

2.3.1. Honeycomb Topology

The interconnect scheme is in many respects at the heart of the NOC architecture. Each resource, whether it is computational, storage or I/O, will have an address and will be interconnected by a network of switches. These resources, will communicate with each other by sending addressed packets of data, routed to their destination by the network of switches.

The overall organization is in the form of a honeycomb, as shown in Figure 2.1.The resources - computational, storage and I/O - are organized as nodes of the hexagon with a local switch at the centre that interconnects these resources. Hexagons at the periphery would be primarily for I/O, whereas the ones in the core would have storage and computational resource.

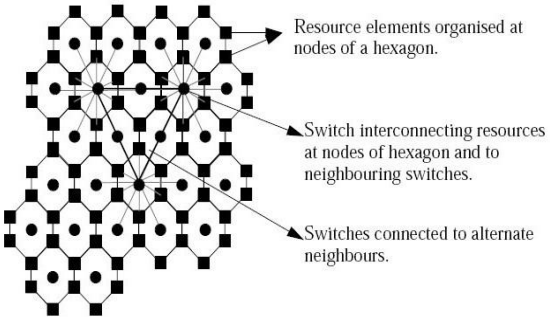

Fig 2.1: A honey Comb structure for NOC

Storage resources can be combined to create a larger virtual storage. Each resource, located on a hexagonal node being connected to three switches, can reach 12 resources with a single hop. To further improve the connectivity, switches are directly connected to their next nearest neighbors, as shown in Figure 2.1, allowing any resource to reach 27 additional resource with two hops. As a last measure to further improve connectivity, every alternate switch is directly connected making each resource element reach a lot more element with minimal number of hops.

2.3.2. Mesh Topology

NOC is a scalable packet switched communication platform for single chip systems. The NOC architecture consists of a mesh of switches together with some resources which are placed on slots formed by the switches. Figure 2.2 shows a NOC architecture with 16 resources. Each switch is connected to four neighboring switches and one resource. Resources are heterogeneous. A resource can be a processor core, a memory block, an FPGA, a custom hardware block or any other intellectual property (IP) block, which fits into the available slot and complies with the interface with the NOC switch. We assume switches in NOC have buffers to manage data traffic. Figure 2.3 shows the internal organization of a typical switch[11].

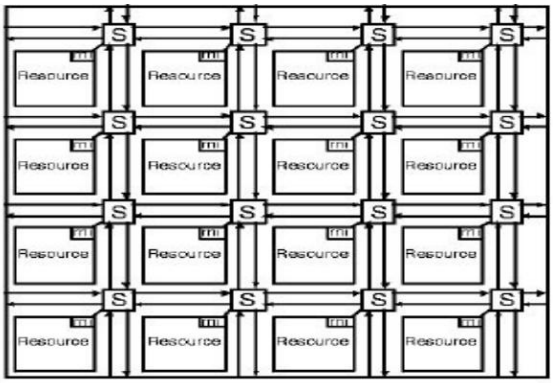

Fig 2.2: A 4X4 NOC switch

Every resource has a unique address and is connected to a switch in the network via a resource network interface (RNI).The NOC platform defines four protocol layers: the physical layer, the data link layer, the network layer, and the transport layer. The RNI implements all the four layers, whereas every switch to switch interface implements the three of four layers except physical layer.

The NOC architecture also has a concept of region. A region allows us to handle physically larger resources and can be used to provide fault tolerance.For our evaluation purposes, we assume a homogeneous NOC architecture, without any region.Network on Chip (NoC) is a new paradigm for designing such future SoCs. In the NoC paradigm a router-based network is used for packet switched on-chip communication among cores[8]

A typical NoC architecture will provide a scalable communication infrastructure for interconnecting cores.The area of multi-media is a very suitable candidate for using this high computing capacity of NoCs.NoC is a general paradigm and one needs to specialize a NoC based architecture for every application area.

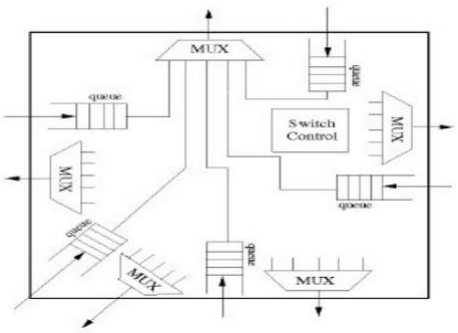

Fig 2.3: A NOC switch

2.4. Switch Architecture

Packet-switching networks place a tight upper limit on block size, allowing packets to be buffered in router's main memory. In routing algorithm, we propose using wormhole routing to reduce the size of the buffers and decrease the latency. The message is divided into smaller units called flits which is performed the flow control. The routing information is added to the first flit and then sent to the destination. As the header advances along the specified route, the remaining flits follow in a pipeline fashion. This will reduces the latency as well as the storage requirements on each node. In addition, it can make the construction of switches to be small, compact,

and fast. Fig. 2.4 shows the proposed packets format. It consists of a header flit following a data flit.

Destination ID	Source ID	Command	Packet ID

a)Header flit

Data	CRC

b)Data flit

Fig2.4: Formats of Packets

2.4.1. Switch Implementation

The switch has three ports, two children ports denoted as L, R, and one parent port denoted as P. In the bottom level of switches, local buses are connected to the children ports and parent is connected to one switch at upper level. To allow for bidirectional communication, each port has two channels, one for packets incoming switch and another for sending outgoing packets. This implies that two packets can be transmitted simultaneously in opposite directions between neighboring switches

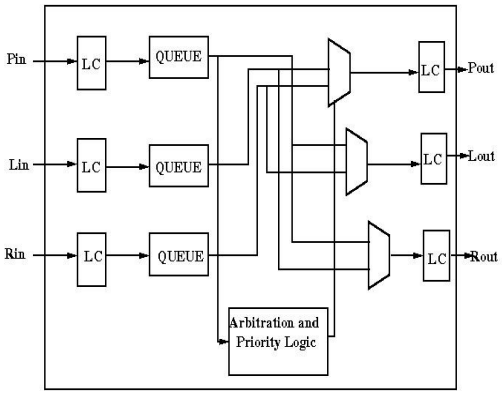

Figure 2.5: Internal architecture of switch

When data arrive on an incoming line, the switch controller must choose an outgoing line on which to forward them, so that they are able to communicate with each other by sending messages. Buffer space in on-chip switch directly impacts the area overhead of the network and thus must be kept to a minimum. Each input channel has a FIFO buffer to store messages. In a single cycle a flit is routed through the switch across a physical channel, and into the input buffer of the next switch. When a message arbitrates for the output of the switch, it simultaneously acquires the output physical channel. Each buffer also has a link controller to detect whether packets enter or not and prevent the buffer congestion. The arbitration and priority logic shown in Fig. 2.5 implements the routing algorithms and judges which packets can move out to the next node when two packets intend to forward to the same node at the same time.

2.5. Network Communication Issues

Changing from a shared bus communication platform, that uses centralized bus arbitation,to a network based on-chip interconnection platform leads to many new communication issues. Among these are: which switching technique should be chosen, how to guarantee Quality-of- Services (QoS) and how should the distributed arbitration work?

2.5.1. Communication Techniques

There are many different switching techniques used in network communication, some of them are presented briefly below. These are circuit switched, packet switched, cell switched and message switched techniques. Communication in a network can also be divided into connection-oriented or connection-less communication. Both types are presented in this section[12].

2.5.1.1. Circuit switched

A circuit switched network has dedicated paths between transmitters and receivers. When such path has been set up only the transmitter and receiver can send information over that path. A good example of a circuit switched communication network is a crossbar switched bus. In a crossbar switched bus dedicated circuits between different transmitters and receivers are set up before communication takes place. When the circuits have been set up, the information is sent as a stream through the circuits to the receivers.

2.5.1.2. Packet switched

Packet switched communication has no dedicated circuit for transmission. It just has a source address and a destination address. The packet will be delivered to the specified destination address.

2.5.1.3. Cell switched

Cell switching is a form of packet switching. Here the message is fragmented into parts. It is up to the receiver to reorder the fragments into an original message.

2.5.1.4. Message switched

Message switching is sometimes called store-and- forward switching. A message is sent out to the network where the complete message is stored at the next switch, until a new routing decision is made. The message continues to jump to the next switch in the network until it reaches its destination. This technique requires that all switches are able to store complete messages.

2.5.1.5. Connection-oriented communication

In connection-oriented communication a path between transmitter and receiver has to be set up prior to transmission. For example circuit switching is always connection-oriented, because the communication path always has to be set up before any information exchange. Packet switched networks can also be connection-oriented, specially when large amounts of data is to be sent. In this case there will be no dedicated path for communication. Instead a virtual circuit(VC)1is set up before transmission. The VC guarantees that a given bandwidth between the transmitter and receiver can be used.

2.5.1.6. Connection-less communication

Connection-less communication is on the other hand just what it sounds like, it's connectionless. That means that no establishment of a path has to be done prior to communication. Packet switched communication is often connectionless when small amounts of data is being sent, because the cost of setting up a path would be high compared to the amount of data.

2.5.2. Blocking & non-blocking communication

When processes communicate they can be blocking or non-blocking. For example if blocking process performs a send or receive operation it will be blocked until an acknowledged is received or some other condition is true. If a non-blocking process performs a send or receive operation it can continue to run even if there hasn't arrived an acknowledge. The non-blocking process should however have a possibility to check if the transmission was successful or not later on .

2.5.3. Blocking and non-blocking communication

A reliable data transfer means that the process gets an acknowledgement when the transfer is complete, so that it knows if the transfer was successful or not. In an unreliable data transfer the transmitting process just sends away the information with no control of the success or failure of the transmission.

2.5.4. Reliable or unreliable communication

A reliable data transfer means that the process gets an acknowledgement when the transfer is complete, so that it knows if the transfer was successful or not. In an unreliable data transfer the transmitting process just sends away the information with no control of the success or failure of the transmission.

2.6. Introduction on Nostrum Concept

The Nostrum concept is a platform for communication between processes on NoCs. Nostrum is made up of a backbone and a NoC architecture. Its main purpose is to provide a stable and reliable platform for communication on NoCs. The backbone uses a general cell switched network which can be used by many SoC designs.

2.6.1. Nostrum protocol layers

Nostrum divides the transfer protocols into layers in the same way as the Open System Interconnection (OSI) model. Nostrum doesn't use the OSI model exactly as it's used in ordinary computer networking. Instead it's used more as an aid for the designer, since there are many differences in designing an on-chip network and an ordinary computer network. That's also why Nostrum isn't strict about keeping the layers separated in hardware. For example if a fusion of two or more layers in hardware results in an optimization of the design, so can be done.

Only the three lowest layers are compulsory in Nostrum, these are physical,data link and network layer. What these three layers do together, if seen as a black-box from the transport layers perspective, is simply to deliver packets from the node where they enter the network to the node pointed out by the destination address.

Nostrum suggest the following responsibilities for the four lowest protocol layers[9,10]:

Physical layer (PL): Physical Layer has the responsibility to move a word of bits from the output of one switch to the input of another switch. PL also introduce data error at bit level.

Data link layer (DLL):Data link layer is responsible for moving a frame of data from the output of one switch to the input of another one with necessary synchronization. It's also responsible for error detection and error correction.

Network layer (NL): Network Layer has the responsibility of delivering packets from a source to a destination through the network and for how packets are routed through the network. It is also responsible of storing the physical addresses that are associated with the Process Identifiers of processes and there has to exist a possibility to manage the physical addresses either during system set-up or during run- time

Transport layer (TL): Transport Layer is responsible for setting up communication path, packetization and de-packetization of messages. The TL is also responsible to supervise the the network so that overload is avoided.

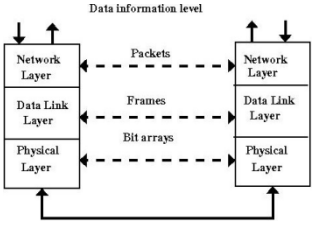

Figure 3.1: The Three compulsory layers in Nostrum

2.7. Discussion

Now let us discuss the NoC's Area and performance overhead compared to traditional architecture based on fixed interconnectivity.

2.7.1. Area overhead

In NoC area overhead is hard to quantify. Moreover, NOC being a reprogrammable architecture like FPGAs, different designs will utilize the resources to a different extent. There would be the overhead of the switch based interconnect scheme.

2.7.2 .Fault tolerance

According to ITRP99 below 100 nm, soft errors are predicted to be frequent enough to severely impact both semiconductor yields and field level product reliability.Having, a switch based interconnectivity and a regular structure can significantly alleviate this problem.

2.7.3. Parallelism and Scalability

The wires in the links of the NoC are shared by many signals. A high level of parallelism is achieved, because all links in the NoC can operate simultaneously on different data packets.

Therefore, as the complexity of integrated systems keeps growing, an NoC provides enhanced performance (such as throughput) and scalability in comparison with previous communication architectures (e.g., dedicated point- to-point signal wires, shared buses, or segmented buses with bridges). Of course, the algorithms must be designed in such a way that they offer large parallelism and can hence utilize the potential of NoC.

2.7.4. Quality of Services in Network on Chip

Network on Chip is likely to become an attractive alternative for implementing SoCs for many application areas like real time multi-media applications. This implies that the underlying on-chip communication network will be required to provide deterministic bounds on delays and throughput for communication among some pairs of cores on the chip. This is generally achieved by special designs of network routers which have capabilities of reserving channels or bandwidth for certain traffic and/or by allowing packets with different priorities. In any design there will also exist communication

traffic that will require no guarantees on delay or throughput. The communication network is required to provide best possible performance for such traffic.

These two types of communication traffic are referred to as Guaranteed Throughput (GT) and Best Effort (BE) traffic. A system running concurrent applications may also have concurrent GT and BE traffic in the network. In such mixed systems the quality expressed in latency, throughput or jitter of traffic flows between nodes that communicate.

QoS routing schemes for on-chip networks to optimize the communication performance, and the main problems that will be addressed are as follows:

• Guaranteed-Throughput for on-chip communication.

• Specialization of communication protocols for an application or a set of applications.

• Design and analysis of heterogeneous communication architectures for on-chip communication

3. CONCLUSION

Research community has recently witnessed the emergence of interconnect networks methodology based on Network-on-Chip(NoC).Network on Chip is likely to become an attractive alternative for implementing SoCs for many application areas like real time multimedia applications. In a few years from now this will lead to the possibility of having billions of transistors on one single chip. As the complexity of integrated systems keeps growing, an NoC provides enhanced performance (such as throughput) and scalability in comparison with previous communication architectures (e.g., dedicated point-to-point signal wires, shared buses, or segmented buses with bridges). Factors like the increase of wire delay, the controlling of power distribution and power dissipation and also the ability to guarantee signal integrity are comparatively solved in Network on chip design. Network-on chip technology has a great scope in the future.

4. ACKNOWLEDGEMENT

I thank my husband Mr.Patwin Fizardo for all his support.

5. REFERENCES

[1] Wael Badawy, Graham Jullien .2003. System-on-chip for real-time applications.

[2] S. Kumar A. Postula J.Oberg M. Millberg A. Hemani, A. Jantsch and D. Lindqvist. "Network on chip: An architecture for billion transistor era", November. 2000.

[3] International Technology Roadmap for Semiconductors, 2003. http://public.itrs.net.

[4] Kenneth W. Mai Ron Ho and Mark A. Horowitz. "The future of wires" in Proceedings of the IEEE. 89:490–504, April 2001.

[5] T.N. Theis. The future of interconnection technology. in ibm j. research and development. 44:379–390, May2000.

[6] NoCSim, http://codesign.cs.tamu.edu/nocsim and Exhibition (DATE), Paris, 2000, pp. 250-256.

[7] L. Benini and G. D. Micheli. Networks on chips: "A new soc paradigm. in IEEE computer", January 2002.

[8] [8] William J. Dally and Brian Towles, "Route packets, not wires: On-chip interconnection networks. in dac - design automation conference", June 2001.

[9] Erland Nilsson Mikael Millberg and Rikard Thid. "The nostrum protocol stack and suggested services provided by the nostrum backbone", November 2002

[10] Erland Nilsson, Mikael Millberg, Rikard Thid,S.Kumar,A.Jantsch. "The Nostrum backbone-a communication protocol stack for Networks on Chip" 2004

[11] Abdelhamid Helali,Adel soudani,Jamila Bhar and Salem Nasri "Study of Nework on Chip resources allocation for QoS Management"2006

[12] Behrouz A Forouzan,Data Communications and Networking.