Network-on-Chip: A New System-on-Chip Paradigm

Parag Parandkar¹, Shafi Patel² and Sumant Katiyal³
¹Chameli Devi School of Engineering, Indore, ²School of Electronics, DAVV, Indore
³p_paragp@yahoo.com, ²shafi.techoz@gmail.com and ³sumant578@yahoo.com

ABSTRACT
The aim of this paper is to expose the concept of Network-on-Chip (NoC), within the VLSI realm, in which networking principles play a significant role. It is an emerging paradigm for communications within large VLSI systems implemented on a single chip, also termed as System-on-Chip (SoC). It offers paradigm shift in various aspects, such as, in terms of architecture by replacing the earlier existence of global wires and point to point interconnects by a customized network, in terms of usage by offering data transportation through packets and flow control units (flits), in terms of organization by separating design of computation from communication within SoC. NoCs also acts as best test cases for SoCs. Networking researchers will find new challenges in exploring solutions to familiar problems such as network design, routing, Quality-of-Service, power management, fault tolerance in unfamiliar settings under new constraints. NoCs offer a wide, complex, multi-disciplinary area to the design fraternity.

KEYWORDS
CMP, IP, SoB, SoC, VC

1. INTRODUCTION
In the past decades of emergence of VLSI era, System on Board (SoB) has been the dominant methodology for designing complex digital systems. With the ever increasing complexity of applications and their required algorithms and with the facilitation provided by IC processing technology to manufacture minimal size of transistors, SoB has been replaced by System on chip (SoC) methodology.

SoC consist of a number of pre-designed intellectual property (IP) assembled together using electrical bus to form large chips with very complex functionality. Future SoC architectures will consist of hundreds of pre-designed IPs assembled together to form large chips with very complex functionality. As technology scales and chip integrity grows, on-chip communication is playing an increasingly dominant role in SoC design. To deal with the increasingly difficult problem of on-Chip communication, it has been recently proposed to connect the IPs using a Network-on-Chip (NoC) architecture.

Network-on-Chip is a new approach to the design of communication sub system of System-on-Chip (SoC). It is an emerging paradigm for communications within large VLSI systems implemented on a single chip. NoC based systems can accommodate multiple asynchronous clocking that many of the today’s complex SoC designs use. NoC is merging of data networking, System-on-chip/Chip multi processors architectures and VLSI disciplines.

In NoC each core is connected to a switch by a network interface. Cores communicate with each other by sending packets via a path consisting of a series of switches and inter-switch links.

Fig 1 Network on Chip as SoC Communication Infrastructure

As shown in Fig 1, each tile is composed of a resource(R) and a switch or router(S). The router is connected to the four neighboring tiles and its local resource via channels. Each channel consists of two directional point-to-point links between two routers or a router and a local resource.

As NoC acts as the most efficient SoC communication infrastructure. NoCs are generally targeted for large SoCs. Large SoCs have multiple clock domains. So, NoCs in large SoCs are asynchronous.

NoCs are also used to test SoCs. Testing of SoCs seek repeatable cycle accurate patterns on chip I/O pins. But systems are not cycle accurate due to the presence of multiple clock domains, synchronizers and statistical behavior. NoC facilitate cycle accurate testing of each component inside the SoC, by enabling controllability and observability on module pins, instead of chip pins.

2. MOTIVATION
With the ever diminishing feature size in the VLSI domain from 180 nm to 130 nm to 90 nm to 65 nm to 45 nm and further down to 22 nm, on-chip modules are getting multiplied and new on-chip communication solutions are evolving in order to support new inter-module communication demands.
Traditional solutions, which were based on a combination of shared-buses and dedicated module-to-module wires, are finding limitations with the increasing scalability of features sizes of the transistors, and are no longer adequate for deep sub-micron technologies. The elemental reason for the transition to NoC-based solutions is the failure of current-day VLSI inter-chip communication design methodology for the deep sub-micron chip manufacturing technology. Chip design in the last decade consisted of designing separate functional modules, and interconnecting those using standard shared buses as well as a huge bunch of inter-module wires for critical or long connections. The technological progress of the silicon industry follows Moore’s Law, doubling the number of gates every 18 months by shrinking the technology dimensions proportionally. However, whereas shrinking gates reduces their intrinsic latencies and dynamic power dissipation, the same is not true for interconnection wires. The decrease in wire dimensions increases their resistance, while decreased inter-wire spacing increases their capacitance, wire coupling delays [1], and crosstalk noises. This increase in line capacitance, resistance, and inter-wire cross-talk increases the latency of signal propagation with the advancement of the technology. In order to preserve wire speeds and voltage noise margins, today’s long wires (including bus lines) require the periodic insertion of repeaters [2], which in turn consume more dynamic and leakage power.

In a modern microprocessor, 50% of the power dissipation is due to the (long) wires [3]. These phenomena intensify with time [4], as global wire delays scale worse than gates, by one to three orders of magnitude [5]. This trend favors the use of short wires and the overall reduction of the total wire length in the chip. The network-on-a-chip concept naturally embodies both ideas, by building a highly-utilized and shared interconnection network with short links connected by routers. In addition, a NoC-based interconnect achieves better scalability than traditional solutions in terms of the number of modules on a chip. Recent analysis [6] shows that the power and area of a NoC based on a grid of routers connected by short wires scales linearly with the chip density (number of modules), whereas common solutions, namely buses, segmented buses, and point to point links, exhibit super-linear growth in both dynamic power and area.

3. ON-CHIP VS OFF-CHIP COMMUNICATION
Adoption of any off-chip network feature will not prove to be as efficient for NoC. Some of the wide differences are as follows.
Guaranteed service provides a wide possibility in off-chip networks, but extremely hard to configure for NoCs. Best effort delivery service is simpler to implement.
Off-chip uses lots of buffer space but is unrealizable in NoCs due to high power dissipation. Off-chip offers complex routing techniques, but NoCs are restricted to offer fixed, single path routing to save energy and area. NoCs use priority levels and pre-emption strategies to avoid packet conflicts.
NoC topologies are fixed throughout their lifetimes. That is, they do not need to support the dynamic addition or removal of network-attached modules like in the case of off-chip communication.
NoC is synthesized anew for each design [[3],[1]] eliminating the need for standard network protocols; i.e., there is no advantage in backward compatibility of NoC protocols and architectures employed in a new chip designs with those used in previous designs.
NoC meets strict QoS requirements for distinct types of on-chip traffic, such as interrupt signals or fetching instructions and data from caches to processors. While Off-chip parallel computer network can afford to have some relaxed QoS requirements.
The NoC design process dramatically differs from that of classical networks in their ability to alter the physical layout of the network routers and links along with the chip module placements. This enables the designer to optimize the NoC geography and resource allocation according to traffic and layout constraints. Moreover, if traffic requirements are known ahead of time the NoC’s topological layout and capacity allocation can be optimized to this traffic with any desired routing protocol; this eliminates the need for a high cost load-balanced routing protocol as was prevalent in off-chip parallel computing.

4. BENEFITS OF ADOPTION OF NETWORK ON CHIP AS A COMMUNICATION INFRASTRUCTURE
Traditionally SoCs have been designed with dedicated point-to-point connections with one wire dedicated for each signal. This methodology of interconnection has got several limitations with increasing size and complexity of SoC from a physical design point of view. The wires occupy most of the chip area and in recently most prominent nanometer CMOS technology (inculcating feature sizes of less than 45 nm or so), interconnects plays a dictatating role in deciding the performance and dynamic power dissipation of the overall system lying on the chip, as compared to their logic counterparts, due to their comparable size with the transistors. The NoC links benefits from their regular structure which in turn reduces complexity of designing wires for predictable speed, power, noise and reliability. From the high abstraction system design viewpoint, with the advent of multi-core processor system, also called as chip multiprocessors (CMP), a network is a natural architectural choice.
An NoC provides separation between computation and communication within a SoC, support modularity and Intellectual Property (IP) reuse via standard interfaces, handle synchronisation issues, serves as a platform for system test, thus summarily increases system productivity.
The primarily considerations in VLSI are minimizing power dissipation and area. This has a number of implications on NoC design. First, NoC components should be extremely simple, so
as to allow implementing them with a small number of logic gates and to expend as little energy as possible. In addition, power considerations render shortest-path routes highly desirable, while area considerations dictate the use of small routing tables.

5. CONCERNING ISSUES:
There are three levels on which NoC can be described: at circuit level, using wires, buffers, routers and network interface; at network level using topology, routing, flow control and at architectural level using characterization of traffic by making use of packets and flits. In designing NoC systems, there are several issues to be concerned with, such as topologies, routing algorithms, performance, latency, complexity and so on. Among these factors, nothing can be independent in deciding NoC architecture.

6. TOPOLOGIES:
There are several different kinds of topologies. SPIN topology using generic interconnect template has been proposed [7], where a fat-tree architecture is used to interconnect IP blocks. In this fat-tree, every node has four children and the parent is replicated four times at any level of the tree. A mesh-based interconnect architecture called CLICHÉ (Chip-Level Integration of Communicating Heterogeneous Elements) consisting of an m x n mesh of switches interconnecting computational resources (IPs) placed along with the switches [8]. A NoC architecture using 2D Torus, which is basically same as mesh, has been proposed by Dally and Towles [9]. The only difference is that the switches at the edges are connected to the switches at the opposite edge through wrap-around channels. The OCTAGON MP-SoC architecture with a basic octagon unit consisting of eight nodes and 12 bi-directional links have been proposed [10]. Each node is associated with a processing element and a switch. Communication between any pair of nodes takes at most two hops within the basic octagonal unit. A interconnect template following a Butterfly Fat-Tree (BFT) architecture where the IPs are placed at the leaves and switches placed at the vertices are also proposed [11].

As a feasible topology in NoC systems, the mesh is getting popular for its modularity; it can be easily expandable by adding new nodes and links without any modification of the existing node structure. As the mesh nodes can be used as basic components in on-chip communication, they are potentially important components to accomplish a scalable communication model in NoC environment [12]. Another reason behind this popularity is the notion of being partitioned into smaller meshes, which is a desirable feature for parallel applications [13].

7. SWITCHING METHODS:
There are different types of switching techniques such as circuit switching, packet switching, and wormhole switching [14]. Switching techniques determine when and how internal switches connect their inputs to outputs and the time at which message components may be transferred along these paths.

In circuit switching, a physical path from source to destination is reserved prior to the transmission of the data throughout initialization processes including setup and acknowledgement. The reserved path is held until all the data has been delivered. The advantage of this approach is the reserved network bandwidth for the entire duration of the delivered data. However, with respect to the resource utilization, it ties valuable resources during the transmission of data and the initialization processes cause unnecessary delays.

In packet switching, data is divided into fixed-length packets and, instead of establishing a path before sending data in circuit switching, whenever the source has a packet to be sent, it transmits the data. In order to store entire packets in a switch, it requires large-sized buffers. Therefore, the need for storing entire packets in a switch makes the buffer requirement high, resulting in infeasible solution for a SoC environment.

In wormhole switching, the packets are further divided into fixed length flow control units or flits, resulting in smaller buffer space requirement in the switches. One drawback of this simple wormhole switching is inability of interleaving or multiplexing distinct messages over a physical channel. However, by applying virtual channels, such channel utilization can be increased.

Among several switching techniques, wormhole routing has been regarded as the most prominent method of reducing message routing latency.

8. ROUTING ALGORITHM:
Another issue in NoC environment is the routing algorithm. In terms of the way of choosing a path among the set of possible paths from source to destination, the routing algorithms are classified as deterministic/oblivious and adaptive ones [15].

The oblivious/deterministic routing algorithms choose a route without considering any information about the network’s present condition, resulting in relatively simple design complexity. DOR (dimension-ordered routing) [16], ROMM [17], and OTURN [18] are examples of deterministic or oblivious routing algorithms.

Adaptive routing algorithms use the state of the network like the status of a node or link, the status of buffers for network resources, or history of channel load information. Adaptive routing algorithms are defined as minimal or fully adaptive routing ones depending on the degree of adaptivity. Even though the adaptive routing algorithms utilize the flexibility in routing paths, the design complexity should be increased.

Better performance routing algorithms using adaptive routing algorithms are proposed as [19][20][21][22][23][24].

On the other hand, the adoption of virtual channel (VC) has been prevailing because of its versatility. By adding virtual channels and proper utilization, deadlock-freedom can be easily accomplished. Network throughput can be increased by dividing the buffer storage associated with each network channel into several virtual channels [20], resulting in increase of channel utilization. By proper control of virtual channels,
network flow control can be easily implemented [25]. Also to increase the fault tolerance in network, the concept of virtual channel has been utilized [26][27]. However, in order to maximize its utilization, how to allocate virtual channels is a critical issue in designing routing algorithms [28][29].

9. SIMULATORS
Design decisions on NoC architectures are typically made on the basis of simulation before resorting to emulation or implementation since it is cheap and flexible. To make a right decision on the network architecture, a simulation tool should enable to (1) faster explore the architectural design space; (2) assess design quality regarding performance, cost, power and reliability etc.; (3) evaluate extensively with various regular traffic patterns and application-oriented traffic.

Network-on-chip research can be carried out on following simulators:
1. Noxim
2. Netmaker System Verilog on-chip interconnection library for simulation from University of Cambridge UK
3. Maia Version 3
5. NOXIM SystemC NOC simulator
6. NIRGAM SystemC NOC simulator
7. Luna: Link Utilization for Network power Analysis
8. Orion: A Power-Performance Simulator for Interconnection Networks (Princeton)
9. Nostrum NoC Simulation Environment (NNSE) System-C.
10. gpNoC (General Purpose Network on chip simulator)

Relevant Benchmark Suites and Sources
- RAW Benchmark Suite (MIT)
- VersaBench (MIT)

10. CONCLUSION
The facts and figures explored in the current survey regarding NoC imply that NoC design requires the development of new solutions. Hence, NoC is a promising research field, which can have a large practical impact on the future of VLSI, and where networking expertise will be valuable. The shift to NoC-based architectures is inevitable, and is likely to become the main trend in VLSI in years to come. The main requirements from a NoC are conserving power and area, which will be manifested in designs consisting of short wires interconnected by simple routers, and energy conserving protocols.

NoC also provides distributed nature of modern chip architectures, which introduce on chip peer-to-peer communication patterns. Unlike traditional chip designs, which had rigid master-slave or pipelined communication patterns, the next generation of chip designs will incorporate multiple autonomous intelligent modules with a rich collection of communication services among them. There are already commercial examples of Chip Multi Processors (CMP), e.g., IBM’s Cell, Xilinx’s FPGA with multiple PowerPC cores, etc.

The companies which started using NoC communication architecture include Freescale, NXP, ST Microelectronics, Infineon and Intel. The companies involved in selling the NoCs includes Sonics (USA), Arteris (France), Silitix (UK).

To date, several prototype NoCs have been designed and analyzed in both industry and academia but only few have been implemented on silicon. However, many challenging research problems remain to be solved at all levels, from the physical link level through the network level, and all the way up to the system architecture and application software.

11. FUTURE SCOPE
The NOC concept separates the concerns of computing and communication, and is expected to be ideally suited to address this increased system complexity and declining system productivity. NoC architectures and hardware-related issues are well addressed, but still an integrated approach for modeling, co-designing and co-developing Hardware-Software with a NoC architecture is missing. Future scope exists for Application mapping strategies and feasible applications for NoC. Research on low cost, area and power efficient solutions of NoC for it to be applicable in the embedded systems industry is the need of the time in future.

REFERENCES

Copy Right © INDIACOM-2011 ISSN 0973-7529 ISBN 978-93-80544-00-7
Network-on-Chip: A New System-on-Chip Paradigm


