CHAPTER 1

INTRODUCTION

1.1 NETWORKS ON-CHIP (NoC)

As the number of processors on a single chip and the computing complexity increases day by day, the interconnection and communication method between the processors became essential factors affecting the performance of chip-multiprocessor. It requires more effective interconnection and communication among the processors to improve its performance, rather than depending only on the processing speed. The system require a complete set of communication demands and characteristics of all kinds of processors, and should provide better data transmission performance in limited conditions such as chip area, power consumption, data bandwidth, etc. Therefore, it requires higher demands for on-chip communication, such as high speed, high bandwidth, high throughput while small area and low power consumption. Traditional interconnection architecture of the chip-multiprocessor, such as on-chip bus and crossbar cannot satisfy these requirements due to problems like reusability, flexibility and scalability. It needs a more perfect and effective interconnection technology.

Similarly, as the numbers of cores in the processor increases, the interconnection network between the cores play a vital role in the Chip Multiprocessor’s (CMPs) overall performance. Usually, shared buses used as communication medium for CMPs with a few numbers of cores. As
transistors improved, resulting in higher clock frequencies, the number of clock cycles required to pass from one end of the chip to the other increases due to the reason that chip size remains the same. This results in slower bus speed than that of clock frequency of cores, as the delay of cores depends on local wires, which range with transistors. As simple shared buses are unsuitable for future CMPs with hundreds to thousands of cores, the Network-on-Chips have emerged as a solution for designing interconnection fabric in CMPs. NoCs provide a low-latency, low-power, high bandwidth communication medium for CMPs by using the benefits of a modular packet switched network. Routers and links are the key components, which form NoCs. A router is responsible to receive and forward communication traffic, and a link provides point-to-point connection between two routers. Each node in NoC is connecting to a limited number of neighbouring nodes with short wires in order to achieve high speed.

The development of NoC architecture tends to replace the on-chip bus architecture and appears as a better effective architecture for on-chip communication in large-scale complex chip-multiprocessor. 2×4 NoC interconnection architecture is mainly used and it is a heterogeneous structure, i.e., main processor as a controller for system management and eight Digital Signal Processors (DSPs) as compute-intensive processors for mass data processing. NoC uses on-chip bus and interconnection architecture, where on-chip bus connects the main processor, shared memory and external devices and eight DSPs interconnect each other with NoC interconnection architecture, which communicates with the main processor through on-chip bus.

Networks-On-Chip (NoC) architectures are becoming the effective fabric for both general-purpose chip multi-processors and application-specific
systems-on-chip designs. In the design of NoCs, high throughput and low latency are both important design parameters and the router micro-architecture plays a vital role in achieving these performance goals. High throughput routers allow a NoC to satisfy the communication needs of multicore applications, or the higher achievable throughput can be traded off for power saving by using fewer resources to attain a target bandwidth. Ultimately, a router’s role lies in the efficient multiplexing of packets over the network links.

A computer network is an interconnection between general purpose programmable devices that deals with the handling of data. Computer networks can be used for a variety of purposes such as facilitating communication between devices, sharing hardware, software, files, data and information. The most popular applications include electronic mail, web-browsing, digital libraries, video-on-demand, file transfer and video/audio conferencing.

Routing is the process of selecting a path in a network to send network traffic. When a source node wants the network to deliver a message to a certain destination node, it specifies the address of the destination node. In computer networks based on the number of destinations to which the information must be transmitted, the routing is classified as unicast, multicast and broadcast transmissions.

1.1.1 Router Buffering

Router buffering was used to temporarily store the arriving flits that cannot be forwarded immediately onto output links when contention arises. This buffering can be at the inputs or the outputs of a router, corresponding to
an input buffered router (IBR) or an output buffered router (OBR). OBRs are attractive because they have higher throughput and lower queuing delays under high loads than IBRs. However, a direct implementation of an OBR would require each router to operate at $P$ times speedup, where $P$ is the number of router ports, which can realized with the router clocking at $P$ times either the link clock frequency, or the router having $P$ times more internal buffer and crossbar ports. Both are prohibitive given the aggressive design goals of most on-chip network applications, such as high-performance chip-multiprocessors. This is a key factor for the adoption of IBR micro-architectures as the effective design choice and the extensive prior effort in the computer architecture community on aggressively pipelined IBR designs.

1.1.2 Internet Routers

Unlike on-chip routers, Internet routers typically use the store-and-forwarding of packets, flit-level flow control is widely used, and flits can be forward immediately without waiting for the remaining flits of the same packet to arrive. These routers cannot operate at aggressive clock frequencies like on-chip routers and the router pipelines possess high complexity at each level.

Routers in the emerging Internet economy will not only forward data packets but also provide value-added services for digital goods being transported. Both customer and technology forces motivate such a trend. As network contents become priced and semantically sophisticated, network users and consumers will demand better service assurances in their transport. Security services such as copyright management and intrusion detection protect legal properties from theft. Accounting services allow network usage
to be correctly billed. Services for Quality of Service (QoS) reservation, differentiation, and adaptation allow digital goods to be delivered with user-centric performance guarantees.

Simultaneously, hardware vendors are beginning to make routing systems that interface a high speed switching fabric with software programmable communication processors. The resulting flexibility of programmable software will make it possible for value-added services to be implemented as general programs that can be deployed on demand. These router programs, implementing services according to diverse application needs and system policies, can then be invoked by traversing flows through an exported router Application Programming Interface (API).

1.1.3 Software Programmable Routers

To achieve interoperability and universal deployment of value-added services, various efforts have focused on router API standardizations. Software programmable routers pose new challenges in the design of router Operating Systems (OS). Firstly, the router programs will require access to diverse system resources, such as forwarding network bandwidth, router CPU cycles, state-store capacity, and capacity for secondary data stores which will be useful when router programs collect extensive system data for network monitoring. Scheduling algorithms must be developed that can work with different resource characteristics, and resource co-scheduling issues become interesting.

Secondly, the next generation Internet must support a large community of heterogeneous resource consumers, whose resource demands will either have to be coordinated to enable cooperation or arbitrated to resolve competition. This requires proper abstractions for resource
allocations, flexible and efficient flow differentiation for resource binding, and scheduling techniques that can support various scenarios of resource sharing and isolation. Finally, a router OS must concurrently support multiple virtual machines exporting different APIs. This gives various benefits. For example, different APIs can be provided for different classes of applications, or different Internet service providers can prefer different “router service providers”. Moreover, API versions may evolve due to bug fixes, addition of new functions or exclusion of outdated functions. It is highly advantageous if a newer version virtual machine has been introduced while an older version is still operational, so as not to disrupt services for legacy flows.

1.1.4 Latency Significant Routers

Latency is recognized as one of the most critical design characteristics for on-chip interconnection network architectures. In this work, a performance model which predicts the latency of flows in a Network-on-Chip (NoC) based system is proposed. Performance models are frequently employed by system designers for early architecture and design decisions. Typically, engineers construct a performance model, and then compare future technology options based on performance model projections. To this end, application and architecture models are first developed separately. Then, the application is mapped to the architecture and a performance model is used to evaluate the chosen application-architecture combination. Nowadays, most performance models of NoCs rely on simulations. The use of simulation experiments makes the task of searching for efficient designs computationally intensive and does not scale well with the size of networks.

Therefore, it is simply impossible to use the simulation in optimization loops. An alternative approach is an analytical model which can
estimate the desired performance metrics in a fraction of time. Analytical models can based to prune the large design space in a very short time compared to simulation. Thus, it is justified to derive accurate analytical models for performance prediction of NoCs to eliminate the need for time consuming simulations. The information provided during the performance analysis step can be used in any optimization loop for NoCs such as topology selection, application mapping, and buffer allocation. Although the use of high-level models conceals a lot of complex technological aspects, it facilitates fast exploration of the NoC design space. Accurate simulations can be setup at later steps of design process when the design space is reduced to a few practical choices.

This work optimizes the traffic regulation parameters aiming for buffer optimization. Although this approach is proper for such a system with real-time requirements, many NoC-based systems have more relaxed timing constraints. To the best of our knowledge, this work proposes the first average case analytical model for on-chip routers which takes into account the burstiness of the traffic. The proposed model can be used to develop a thorough performance analysis for arbitrary network topology with wormhole switching under arbitrary traffic pattern. The proposed model, besides providing performance metrics such as average latency and router blocking time, gives useful feedbacks about the network behavior which can be used in an optimization loop for NoCs such as topology selection, application mapping, and buffer allocation.

1.1.5 Routing and Topology

Regarding the topology, regular direct topologies, such as mesh and torus are intuitively feasible for physical layout in a 2-D chip. On the contrary,
the high wiring irregularity and the large router radix of indirect topologies such as Benes or Butterfly pose a challenge for physical implementation. However, an arbitrary permutation pattern with its intensive load on individual source-destination pairs stresses the regular topologies and that may lead to throughput degradation. In fact, indirect multistage topologies are preferred for on-chip traffic-permutation intensive applications.

Regarding the switching technique, packet switching requires an excessive amount of on-chip power and area for the queuing buffers (FIFOs) with pre-computed queuing depth at the switching nodes and/or network interfaces.

Regarding the routing algorithm, the deflection routing is not energy-efficient due to the extra hops needed for deflected data transfer compared to a minimal routing. Moreover, the deflection makes packet latency less predictable; hence, it is hard to guarantee the latency and the in-order delivery of data.

This research presents a novel silicon-proven design of an on-chip permutation network to support guaranteed throughput of permuted traffics under arbitrary permutation. Unlike conventional packet-switching approaches, network-on-chip employs a circuit-switching mechanism with a dynamic path-setup scheme under a multistage network topology. The dynamic path setup tackles the challenge of runtime path arrangement for conflict-free permuted data. The pre-configured data paths enable a throughput guarantee. By removing the excessive overhead of queuing buffers, a compact implementation is achieved and stacking multiple networks to support concurrent permutations in runtime is feasible.
1.1.6 Multi Core Architecture

This interconnect makes use of a large space in the on-chip area in a multi core architecture. A large amount of transistors may possibly have been employed to increase the complexity of the computational resources and the number of components that need to be installed for the design of communication infrastructure.

Let us consider the system which employs multiple applications running on the system at various time-domains. In such a case, adaptivity would be very advantageous over a fixed topology, i.e. with a fixed number of buffers that work more efficiently in worst case scenarios and signify an over design than various situations. Hitherto, NoCs have been analyzed only as on-chip communication architectures for static application specific multi core systems. Application specific NoCs may be defined as a design time-parameterized architecture through a fixed routing scheme, i.e. with fixed number of virtual connections at each allowed output port or a fixed application mapping illustration. Usually these are specially made for specific application domains or applications.

1.1.7 Cartesian Networks for Terrestrial Applications

The general description of wide area Cartesian networks discussed in the previous sections imposes no limits on the number of layers, and subsequently, the size of the address structure. Clearly, any implementation of a wide area Cartesian network will have limits imposed, either by the number of layers or by the size of the available address structure.
1.2 OBJECTIVE OF THE THESIS

The main objectives of this study are:

1. To evaluate the multi-port router design, its implementation and performance analysis for the proposed design.

2. To develop a performance improved multi-port linear weighted algorithm for the proposed router design.

3. To propose a generalized multi-port selection scheme for 4-port and 8-port router algorithm for performance estimation.

4. To investigate the reliability and reproducibility of developed routing algorithm scheme on various performance evaluation parameters.

5. To develop performance utilization graph based on the estimated parameters such as number of Slices (S), number of Look Up Table,(LUT) Slice Latches (SL), Occupied Slice Latches (OSL), Input-Output Block,(IOB) Latches Input Buffers, (IBUF’s) Gate Array counts (GA), Latency (L) and Power Consumption (PC).

6. To perform comparison to validate the results obtained.
1.3 ORGANISATION OF THE THESIS

This thesis comprises of eight chapters including the introduction. The general organization of the rest of the chapters is as follows.

Chapter 2 provides the detailed literature survey made in the area of router designs with low power schemes and with high scalability. Router architectures and routing table algorithm were also analyzed in this chapter.

Chapter 3 deals with the proposed router design and its architecture. The design methodology of proposed router has explained in detail and the algorithm used has described in this section. It also explains the methods of implementation of the algorithm.

Chapter 4 analyses the architecture level adaptation of the proposed router. The weighted XY Routing algorithm has modified and implemented according to proposed methodology. The run time adaptation of the router has described in detail.

Chapter 5 deals with the design of hybrid Network Interface (NI) model, which combines both master side and slave side NIs for its operation. The architecture of Priority based Router (PR) is also analyzed.

Chapter 6 explains low power and high performance serial on-chip communication link router which includes sending and receiving routers, serializer, deserializer, multi-orthogonal encoder and decoder and design methodologies were analysed.

Chapter 7 summarizes the key points of the overall design techniques and concludes the thesis.
1.4 SUMMARY

In this chapter discussed about various types of basic routers and their architecture details. This chapter also discusses the objective of the thesis work from basic level to architecture and result level, and discussed about the organisation of the thesis.