RESEARCH ARTICLE OPEN ACCESS

Investigating the Performance of NoC Using Hierarchical Routing Approach

Bibin Lawrence R[1], Karthikeyan A[2]
[1]M.E-VLSI Design Vel Tech Multi Tech Dr. Rangarajan Dr. Sakunthala Engineering College, Avadi, Chennai-600 062, India
[2]Department of Electronics and Communication Vel Tech Multi Tech Dr. Rangarajan Dr. Sakunthala Engineering College, Avadi, Chennai-600 062, India

ABSTRACT
The Network-on-Chip (NoC) model has appeared as a revolutionary methodology for incorporatingmany number of intellectual property (IP) blocks in a die. As said by the International Roadmap for Semiconductors (ITRS), it is must to scale down the device size. In order to reduce the device long interconnection should be avoided. For that, new interconnect patterns are need. Three-dimensional ICs are proficient of achieving superior performance, resistance against noise and lower interconnect power consumption compared to traditional planar ICs. In this paper, network data routed by Hierarchical methodology. We are analyzing total number of logic gates and registers, power consumption and delay when different bits of data transmitted using Quartus II software.

Keywords- System-on-Chip (SoC), Network-on-Chip (NoC), Intellectual Property (IP), International Road for Semiconductors (ITRS), Integrated Circuit (IC), Hierarchical Routing, 3D-IC, Fault Tolerant, Monitoring Platform, Torus, Mesh.

I. Introduction

With technology improves, we have to scale down the device. For that we have to reduce the chip measures, but long interconnections are hinders the high performance of SoCs. These long wire connections are becoming a performance obstacles in terms of communication latency and power. The Network-on-Chip (NoC) model is emerging as a revolutionary methodology in solving the performance limitations due to the long interconnects. The NoC helps to build high numbers of intellectual property (IP) cores in a single SoC . However, limited floorplanning choices of the 2D integrated circuits (ICs) limit the performance enhancements of NoC architectures. 3D – IC, contains multiple layers of active devices. Multiple layers have the potential for boosting system performance [1. Because of having shorter interconnect length, 3D IC enhanced its performance even when we are scaling down the device [5]. Due to the shorter interconnect length, package density is enriched and so power is consumed [5]. The performance improvement arising from the architectural advantages of NoCs will be significantly enhanced if 3D ICs are adopted as the basic fabrication methodology. With a 3D degrees of freedom, we can achieve impossible works of 2D. This paper organized as follows. Section 2 explained about NoC technology and its design parameters, section 3 describes about proposed architecture is and finally section 4 describes work done and conclusion.

II. NoC Technology

Today buses are the leading technology for System –on- Chip (SoC). However, buses have relentless limitations that become obvious, if the number of components in a system large. The bus is a communication bottleneck, because it used to broadcast each transfer, thus it will consume more power. Network-on-Chip shall overcome the limitation of buses, since they provide a much large amount of communication resources and the scalable. A terminal node can be any type of component like processor, memory, hardware component, bus-based system with several components. Data packets are routed via channels and switches from source to destination. From figure.1, NoC architecture consists of Network Interfaces, Processing Elements and routers.

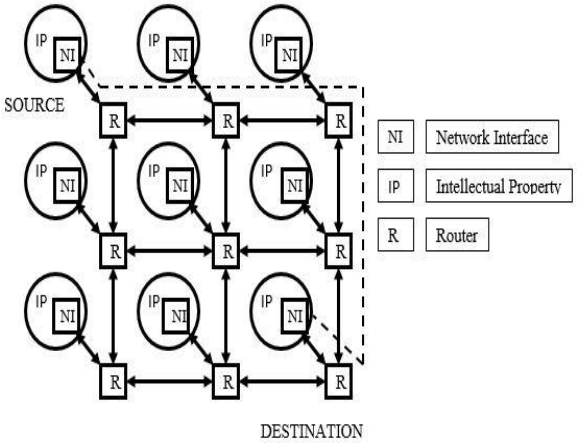

Fig1. NoC Architecture

2.1 Network Interface

Different terminals with different interfaces shall be connected to the network. The network uses a specific protocol and all traffic on the network has to comply to the format of this protocol. In order to allow for different resources to connect to the network, the network interface can be divided into a resource independent part (NI) and a resource dependent part (Resource NI).

2.2 OSI Layers

International Standards Organization (ISO) established the Open Systems Interconnection (OSI) model to define networks. It gives an idea to organize the network components. Network-on-chip also uses the same protocol like OSI model.

- Physical: It defines connector and interface specifications, as well as the medium requirements. Electrical, mechanical, functional specifications are provided for sending a bit stream on a computer network.
- Data link: This filmaffords error detection and control across a single link (single hop).
- Network: It gives end-to-end multi hop data communication. This layer is responsible for data packets forwarding.
- Transport: Connection-oriented services over multiple links e.g., ordering of packets, error free connection.

2.3 Performance Metrices

The topology of an interconnection network is specified by a set of nodes N* connected by set of channels C. Messages originate and terminate in a set of terminal nodes N, where

$$N \in N* \quad (1)$$

Each channel $c = (x, y) \in C$ connects a source node y, where $x, y \in N*$. The channel contains number of parallel signals which is described by its width w_c or w_{xy}. The channel's source node is represented by s_c and destination node is represented by d_c.Its frequency f_c or f_{xy} is the rate at which bits are transported on each signal. Its latency t_cort_{xy}, is the time required for a bit to travel from x to y. Usually the latency is directly related to the physical length of the channel, $l_c = vt_c$ by a propagation velocity v. The bandwidth of the channel is $b_c = W_c f_c$.

- Path

A path is an well-organizedset of channels P = {c1,c2….cn}. The length or hop count of a path is |P|. A minimal path from node x to node y is a path with the smallest hop-count. Thecollection of least paths between x and y is denoted R_{xy}. The diameter H_{max} is the largest least hop count over all pairs of terminal nodes. The averagesmallest hop count H_{min} is defined as the average hop count of all sources and destinations.

$$H_{\min} = \frac{1}{N^2} \sum_{x,y \in N} H(x, y) \quad (2)$$

The physical distance of the path is

$$D(P) = \sum_{c \in P} l_c \quad (3)$$

The dealy of the path is

$$t(P) = D(P)/v \quad (4)$$

- Throughput

The throughput of a network is the data rate in bits per second that the network accepts per input port. The ideal throughput is defined as the throughput assuming a perfect routing and flow control. Maximum throughput occurs, when channel of the nework becomes saturated.

- Latency

The latency of the network is the time required for a message to traverse a network, from the time the head arrives at the input port to the time when the tail of the message departs the output port.Latency depends not only on topology, but also on routing, flow control and the design of the router.

III. Related Work

In this paper, we proposed new architecture and hierarchical routing algorithm used. And we calculated delay, power dissipation, total logic elements and total registers for 8 bit, and 32 bit data using Quartus II software and tabulated the results.

3.1 Hierarchical Routing

This paperused new logical mechanism called as Cluster Based Hierarchical Routing (CBHR) to

improve the efficiency of NoC. This algorithmembraces the following steps [8]:

- The first step is, the total network is separated by different clusters of same or different sizes.
- Second step is providing algorithm for same or different clusters for routing (local or global).
- The third step is change router's working function regarding local or global routing.

From figure.2, we can understand the concept of Hierarchical routing.

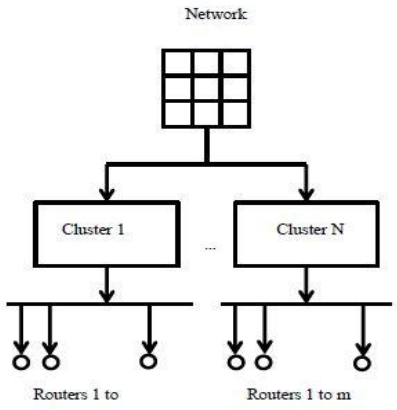

Fig2.CBHR Concept in NoC Architecture

3.2 Proposed Architecture

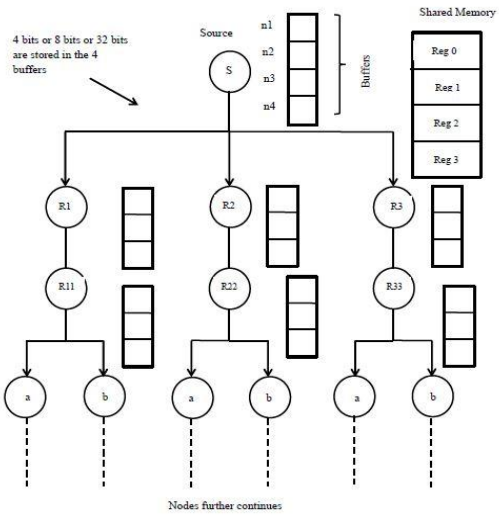

Fig3. Proposed New Architecture

Figure.3 shows the proposed architecture. Let us consider 'S' be the source of the network. Each node in the network consists of four buffers. Buffer is a temporary storage device. Four buffers are taken with 'n' length. Let n1, n2, n3, n4…are name of the buffers. Each buffer stores four bits, eight bits, sixteen bits and thirty two bit data. These 'n' length buffers are linked to the router to handle the data or store the data. Suppose for counter 1, the buffer 'n1' with 'n^{th}' length is exceeded by the data is transfer to the buffer 'n2'. 'n2' will receive the data until it got filled, once 'n2' is filled, it sends the control to the 'n3' buffer. When 'n3' is full, it will be transferred to the buffer 'n4'. For counter 2, from source 'S' it splits into three buffers. Since it's following the FIFO queuing the output data from the buffer 'n1' in counter 1 is goes to the buffer 'n1' in counter 2. The same steps are repeated here. For counter 3 also, the same steps follows. For counter 4, from three routers two other routers are created.

In all the four counter steps, the data of the fourth buffer that is 'n1' is the same (not changed). So we can stop using the additional 'n1' buffer. Instead of using '$n1^{th}$' buffer, we can use 'shared memory'. So when 'n3' filled by 90% atleast, the data is transferred to the shared memory. By following the above steps as an algorithm, a router can achieve less dropping values. So one can achieve greater through put in the NoC's since dropping value and through put have the indirect relation. Further, all the buffers connected to the router are harmonized to the clock with respect to the time along with the shared memory.We are synchronizing with shared memory to achieve smaller area and great speed in the NoC's.

3.3 FIFO Queue

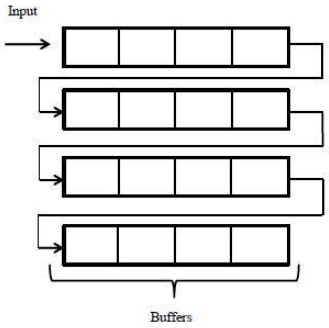

Fig4. FIFO Architecture

First in First Out (FIFO) is used as buffering element or queuing element. FIFO queuing method consists of number of buffers to store data. For example, figure. 4shows that, this FIFO architecture contains 4 bit data. So there are four buffers required. The first input is fills the first row. To give next input, the last column input is shifted to the next row's first buffer. Likewise the other remaining rows are getting filled. Thus stores the data. Size of the FIFO is depends upon the amount of data necessary to buffer. The size of the FIFO is determined by the worst case scenario for the data transfer.

For worst case scenario, difference between the data rate between write and read should be maximum. So data rate of read operation is the number of idle cycles and for read operation, there is no idle cycle. For write operation the data rate will be, Data rate = Number of data * rate of clock.

IV. Workdoneand Conclusion

4.1 8 Bit Data

Output waveforms of 8 bit data shown in figure 5(a),5(b),5(c). the clock tree synthesis, timing analysis summary, total number of registers and power analysis summary are shown in figure 6,7,8.

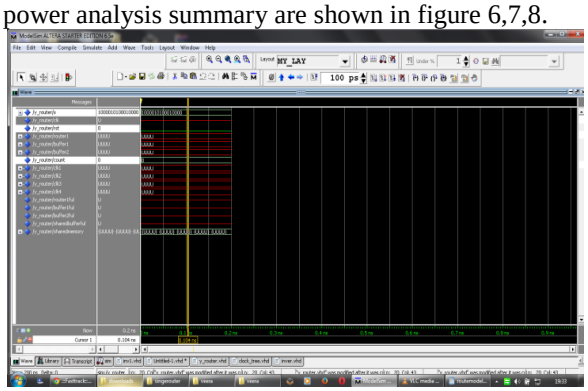

Fig 5(a).Output Waveform for 8 bit data

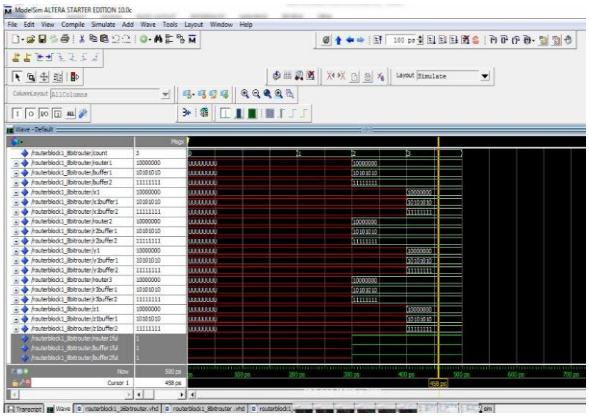

Fig 5(b).Output Waveform for 8 bit data

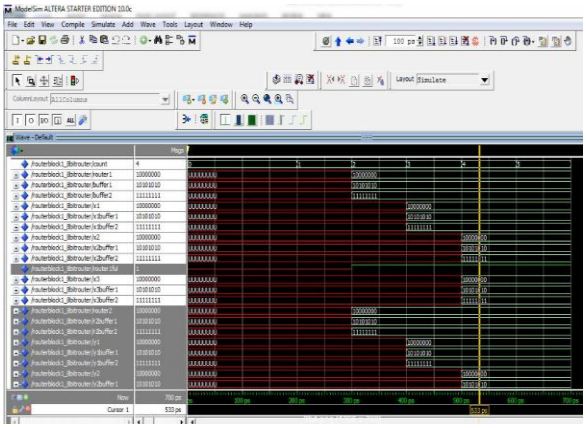

Fig 5(c).Output Waveform for 8 bit data

4.2 Clock Tree Synthesis

Standard Clock Tree Synthesis engines are driven by timing closure and, hence, are not PVT (process/voltage/temperature) variation aware. They are used to fix setup/hold violations by adjusting the clock skew, adding, removing, and swapping buffers, or exploiting different clock wire lengths and levels and so on. As a result, the skewsensitivity with respect to PVT variations cannot be kept low, since it has several contributors originating from different physical phenomena.

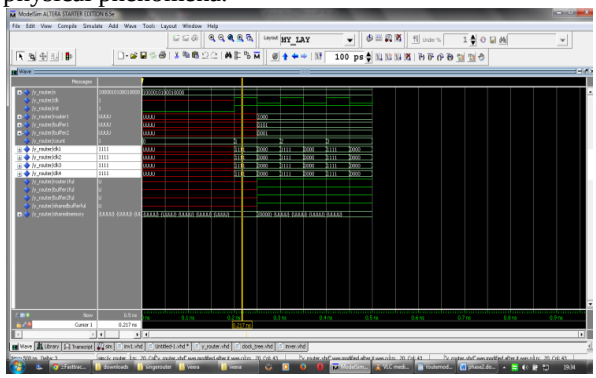

Fig 6.Output Waveform for Clock Tree Synthesis

4.3 Timing Analysis

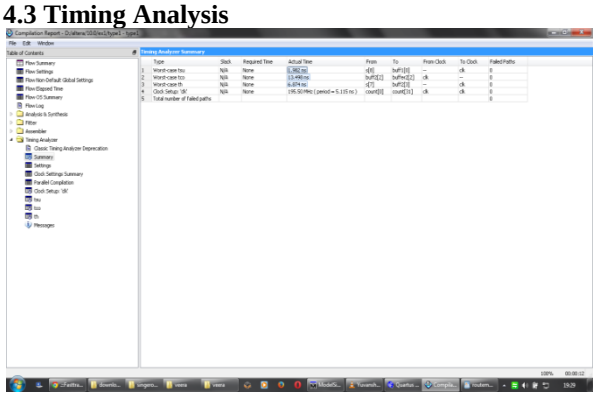

Fig 7.Simulated Result for Timing Analysis

The timing analyzer window helps to know the worst case delay. timing analysis is the methodical analysis of a digital circuit to determine if the timing constraints imposed by components or interfaces are met. Typically, this means that trying to prove that all set-up, hold, and pulse-width times are being met. The typical worst case delay is about 6.874 ns.

4.4 Power Analysis

The power dissipation of very large scale integrated (VLSI) circuits is becoming a critical concern. Accurate and efficient power estimation during the design phase is required in order to meet the power specifications without a costly redesign process. The total thermal power dissipation from the power analyzer is 70.23 mW.

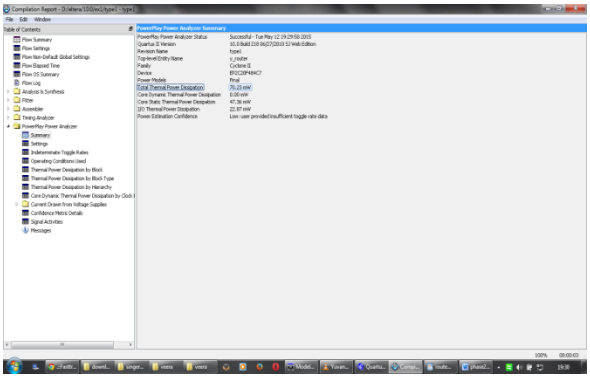

Fig 8. Simulated Result for Power Analysis

4.5 Numberof Elements

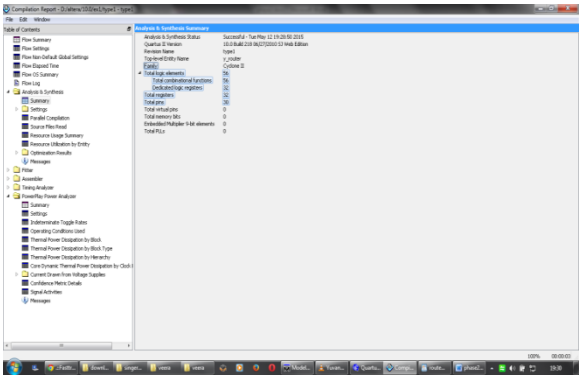

Fig 9. Simulated Result for Number of Registers

From the Analysis and Synthesis Summary, we calculated the toal number of registers is 32 and total combination functions are 56. So that we can analyse the power and delay regarding how much registers and functions are used.

4.6 Shared Memory Architecture

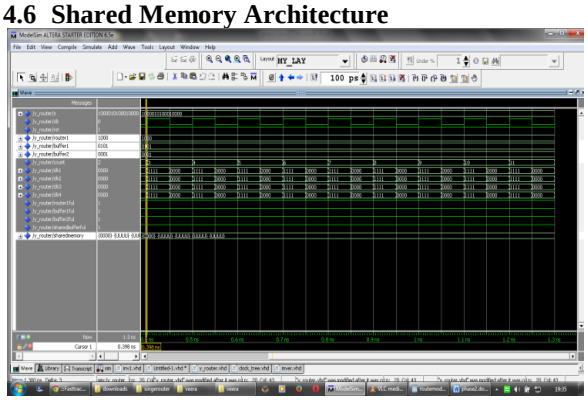

Fig 10. Simulated Result for Shared Memory Architecture

The above figure shows that the result of shared memory architecture. For all the counter, the memory was shared to reduce area and power. The usage of buffers getting reduced because we used common buffer to store same or unrepeated data.we can use limited amount of buffers. So, we can achieve the minimum buffer sizing and power dissipation.

4.7 Comparative Table

Data	*Delay*	*Power Dissipation*	*Total Logic Gates*	*Total Registers*
8 Bit	6.874 ns	70.23mw	32	32
16 Bit	6.832 ns	72.01mw	32	25
32 Bit	6.735 ns	79.85mw	32	25

Table 1 Comparison Table for Different Bit Data

V. CONCLUSION AND FUTURE WORK

From the above table, we conclude that, when transferring 32-bit data has maximum power dissipation. But delay of 32-bit data is small, since delay and power dissipation have inversely proportional relation. The increased number of registers and logic elements will leads to experienced delay. The circuit or device is complex, when the number or logic gates and registers are increased. From the above comparison table, we know that, the hierarchical routing algorithm is better results than adaptive-z algorithm in terms of power, delay and number pf registers. Here, we calculated power dissipation, delay, number of logic gates and number of registers of 8-bit, 16-bit, 32-bit data using Hierarchical Routing and Adaptive Z algorithm and Compare the results by Quartus II (Model Sim) software. In future, we can give separate clocks for each node so that we can improve the data speed. Therefore we can avoid the unnecessary routing path so we can consume power.

Reference

[1]. "*High-performance and Fault-tolerant 3D NoC-Bus Hybrid Architecture Using ARB-NET- Based Adaptive Monitoring Platform*" by Amir-Mohammad Rahmani, Kameshwar Rao Vaddina, Khalid Latif, IEEE Transaction on computers, vol.63, No. 3, March 2014.

[2]. "*TSV Redundancy: Architecture and Design issues in 3D IC*" ,April 2012 by Ang Chih Hsieh and Ting Ting Hwang

[3]. L.Guang, E.Nigussie, P.Rantala, J.Isoaho *" Hierarchical Agent Monitoring Design Approach towards Self-Aware Parallel Systems-on-Chip*", ACM Trans. Embedded Computing Systems, vol 9, 2010.

[4]. C.Rusu, L.Anghel and D.Avresky, "*RILM: Reconfigurable Inter Layer Routing Mechanism for Three-Dimensional Multi-*

Layer Network-on-Chip", Proc, Int. On Line Testing Symp-2010.

[5]. "*Network-on-Chip in a Three-Dimensional Environment: A Performance Evaluation*", by Brett Stanley Feero and Parthapratim Pande, January 2009.

[6]. D. Park, S. Eachempati, R. Das,A. Mishra, Y. Xie, N.Vijaykrishnan, and C.Das, " *MIRA: A Multi-Layered On-Chip Interconnect Router Architecture*", Proc. Int. Symp. Computer Architecture, vol 56, 2008

[7]. H. Matsutani, M. Koibuchi, and H. Amano, " *Tightly-Coupled Multi-Layer Topologies for 3D NoCs*" Proc. Int'l Conf. Parallel Processing, 2007.

[8]. F. Li, C. Nicopoulos, T. Richardson, Y.Xie, V. Narayanan, and M. Kandemir, " *Design and management of Three-dimensional chip multiprocessors using Network-in-Memory*", Proc. Int'l Symp. Computer Architecture, 2006.

[9]. Three-Dimensional Integrated circuits. July/ September 2006 – A.W.Topol, D.C. La Tulipe, L. Shi, K. Bernstein, D.J.Frank.

[10]. M.Li,Q.A.Zeng, and W. B. Jone, "*DyXY: A Proximity Congestion-Aware Deadlock-Free Dynamic Routing Method for Network-on-Chip*", Proc. Design Automation Conf., 2006

[11]. L. Ni and P. Lilje berge, and P. Mc Kinley, "*A Survey of Wormhole Routing Techniques in Direct Networks*", Computer, vol.26, Feb 1993