Comparative Study of Switching Techniques for Network-on-Chip Architecture

Yogita A. Sadawarte
Bapurao Deshmukh College of Engineering,Sevagram
Wardha, Maharashtra, India
y_sadawarte@rediffmail.com

Mahendra A.Gaikwad
Bapurao Deshmukh College of Engineering,Sevagram
Wardha, Maharashtra, India
gaikma@rediffmail.com

Rajendra M.Patrikar
Visvesvaraya National Institute of Technology (VNIT) Nagpur
Maharashtra, India
Rajendra@computer.org

ABSTRACT
In Network-on-Chip architecture (NoC) energy consumption is broadly classified as communication energy & computational energy. Communication energy is highly dominant at higher semiconductor technology due to delay in communication channel. Whereas computational energy decreases with recent trends in semiconductor technology. In communication among IP cores power dissipated in switching process is highly dominant, known as dynamic power. In this paper we compare different switching techniques used in Network on chip architecture. In this paper we discuss the optimization of the dynamic power due to switching activity in the Network–on-Chip architecture that is possible by applying switching algorithms. In this paper we study different switching techniques which are used in Network-on-Chip architecture.

Categories and Subject Descriptors
C.1.2 [**Processor Architectures**]: Multiple Data Stream Architectures (Multiprocessors)- *Interconnection Architectures*; C.2.1 [**Computer-Communication Networks**]: Network Architecture and Design- *Packet-switching networks*.

General Terms
Measurement, Performance, Verification.

Keywords
NoC architecture, Switching techniques.

1. INTRODUCTION
Network-on-Chip is nano scale packet switched networks. It provides attractive alternative solution to traditional bus based interconnection scheme. Network-on-Chip architectures have been proposed to overcome the future communication bottleneck.

ICCCS'11, February 12–14, 2011, Rourkela, Odisha, India.

The NoC is a universal network structure to connect all functional units on the chip. Chip designing is more complex because of Deep Sub-Micron problems. A New design can re use previously designed modules. Without having to re design them because the neighboring environment is always the NoC. This will cut down the design time. The advantages of using NoC structure can be summarized as (a) Improved (b) energy dissipation is less,(c) flexible scalability and (d)design reusability. NoC based designs will provide good solutions for flexible products that should be reconfigurable and programmable; for designs which are the basis for several product variants; for applications with a heterogeneous task mix.

In general switching techniques decides when & how internal switches should connect their inputs to the outputs & the time at which message components may be transferred along these paths.
Energies dissipated in NoC is mainly classified in two categories such as average leakage energy and average dynamic energy.
Average dynamic energy is again classified as: Energy dissipated in Links and Energy dissipated in router. Router energy can be optimized by using energy aware hardware components (i.e. by designing) or energy aware routing algorithms by assuming the energy efficient link

2. NOC ARCHITECTURE
NoC architecture consists of router & physical links. Router consists of (a) the ingress packet process unit (b) the egress packet process unit (c) the arbitration unit (arbiter) and (d) the switch fabrics. [2]. The ingress packet process unit checks the header content of incoming packet & parallelizes the serial dataflow on the transmission line into bus dataflow. The egress process unit collects the processed packets & sends the packets to their destination ports. Arbiter decides when & where a packet should be routed from ingress ports to the egress ports. It makes the decision based on arbitration algorithm. The switch fabric is an interconnect network that joins the ingress ports to the egress ports. The switch fabric circuit is the basic building block inside network router. [2]

2.1 Types of NoC architecture
There are main six types of NoC architectures classified as [3].

2.1.1 SPIN (Scalable, programmable, Integrated Network)
Guerrier and Greiner have proposed SPIN for on-chip packet switched interconnections. In this tree every node has four

children and the parent is replicated four times at any level as shown in figure1.

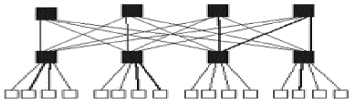

Figure 1. SPIN Architecture.

2.1.2 CLICHÉ (Chip-Level Integration of Communicating Heterogeneous elements)

Kumar have proposed a mesh based interconnect architecture known as CLICHÉ. This architecture consists of an m*n mesh of switches interconnecting IPs placed along with the switches. This architecture is shown in figure2.

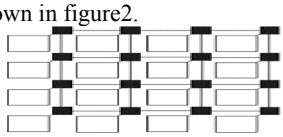

Figure 2. CLICHE Architecture.

2.1.3 Torus

Dally and Towles have proposed a 2D torus as an NoC architecture. It is same as regular mesh. Every switch has five ports, one connected to the local resource and others are connected to the closest neighboring switches as showen in figure 3.

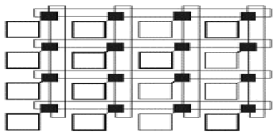

Figure 3. Torus Architecture.

2.1.4 Folded Torus

The long End around connections causes extra delays, which can be avoided by folding the torus, known as folded torus.It is shown in figure 4.

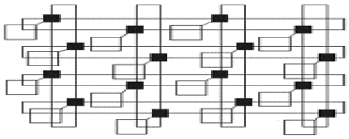

Figure 4. Folded Torus Architecture.

2.1.5 Octagon

Karim et al.have proposed the OCTAGON MP-SoC architecture. It consists of eight nodes and 12 bi-directional links. Each node is having a processing element and a switch as shown in figure 5.

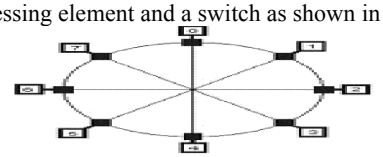

Figure 5. Octagon Architecture.

2.1.6 BFT (Butterfly Fat-Tree)

In this architecture Ips are placed at the leaves and switches are placed at the vertices. A co-ordinate pair (l,p) is used to label each node, where l is node's level and p is its position within that level. It is shown in figure 6.

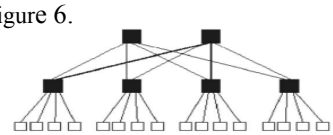

Figure 6. BFT Architecture.

Out of above types folded Torus is more suitable for VLSI implementations [3]. The common features of these kinds of architectures are that the functional IP blocks communicate with each other with the help of intelligent switches. The Switches can be considered as infrastructure IP's providing a robust data transfer medium for functional IP modules.

2.2 Performance Metrics

For comparison of different NoC architectures a standard set of performance metrics are used [3].

2.2.1 Message Throughput

It is measured as the fraction of maximum load that the network is capable of handling physically. Its unit is flit/cycle/IP. It signifies the maximum value of the accepted traffic & it is related to the peak data sustainable by the system. High throughput is desirable characteristics of NoC architecture.

2.2.2 Transport Latency

Transport latency is defined as the time that elapses between occurrence of a message header injection into the network at the source node & occurrence of a tail flit reception at the destination node. Overall latency is addition of Sender overhead, transport latency & Receiver overhead. Less latency is desirable characteristics of NoC architecture.

2.2.3 Energy

When flit travel on the interconnection network, both the inter switch wires & logic gates in the switches toggle & this will result in energy dissipation. The dynamic energy dissipation caused by the communication process in the network Low energy dissipation is desirable characteristics of NoC architecture

2.2.4 Area requirement

Switches consume relative amount of silicon area. Switches have two main components the storage buffer & logic to implement routing & flow control. Silicon area should be very low for high performance. Architecture exhibits high throughput, low latency, energy efficiency& low area overhead

NoC architecture is mainly parameterized on: 1) Size of packets 2) length & width of physical links 3) number & depth of virtual channels and 4) switching technique

3. SWITCHING TECHNIQUES

In switching message having header & data are transported to the destination through the nodes as shown in figure 7.[4]

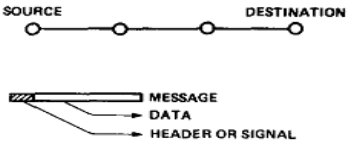

Figure 7. Switching.

3.1 Circuit Switching

A complete path of communication links must be set up between two parties before real communication begins. It means that a physical path is reserved from source to destination prior to the data transmission. The path is held until data is transmitted. The path provides all the addressing information. Since Valuable resources are also tied up for the total duration & unnecessary delay will be there because of set up of an end-to-end path. It is also known as line switching. There are certain drawbacks of circuit switching first one is set up of path is very slow, which

delays transfer of message from sender to receiver. It is used in telephone network for human communication. Generally data streams are having frequent short messages with occasional long ones. Suppose circuit is released after each message then excessive signaling delay, is a serious disadvantage for short messages. Another drawback is those channels are not being used during idle periods, therefore low channel utilization in this switching technique. Figure 8.shows the circuit switched system with the number of intermediate nodes is two.

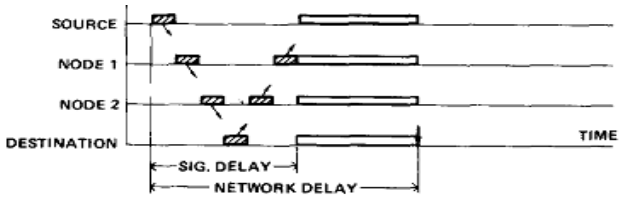

Figure 8. Circuit switching.

3.2 Message Switching

To achieve a better channel utilization idea of relinquishing of channels on the path, during the period in which both parties are silent is applied. In message switching method messages are routed toward the destination node without establishing a path earlier. The paths are assigned dynamically. There is provision of a storage facility at each node, message are stored in intermediate nodes and then forwarded to a selected adjacent node. Hence another name is store-and-forward switching technique.. The selection of adjacent node is made by a decision rule known as the routing algorithm. The process is repeated until message reaches the destination node. By joining the addressing bits the header, each message carries information regarding its destination. This method is used in pager systems. Figure 9. Shows the message switching system.

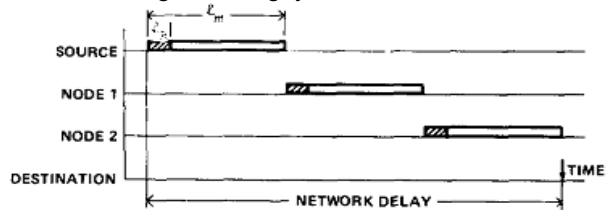

Figure 9. Message Switching.

3.3 Packet Switching

In packet switching data is portioned into fixed length blocks called packets. Each packet has its own addressing information. These causes extra overhead hence buffer requirement is high in these cases. However by dividing messages into packets, a message can use a number of links on a path simultaneously. Chacteristics of this switching are more complete resource sharing, higher channel utilization & lower network delay. It is shown in Figure 10.In message switching and packet switching extra delay is occurred since a message is not permitted to be transmitted out before it is received completely.

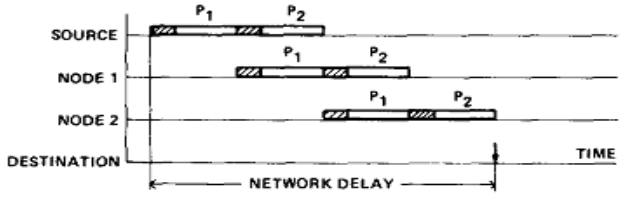

Figure 10. Packet Switching.

3.4 Wormhole Switching (WH)

In wormhole Switching packets are divided into fixed length flow control units (flits). Buffer space requirement is less since input and output buffers are required to store only few bits. Switches will be small & compact. Flits are of three types, header flit, data flit & tail flit. Header flit contains routing information. Decoding of the header flit enables the switches to establish a path and all the subsequent flits only follow the path in a pipelined fashion. The advantage of wormhole switching is no packet reordering is required at the destination. If a flit faces a busy channel, subsequent flits have to wait at their current location. One main drawback of wormhole switching is transmission of distinct messages cannot be multiplexed over a physical channel. Messages must cross the channel totally before they can be used by another message. This reduces the channel utilization if a flit is blocked in a buffer [3].

3.5 Virtual Cut Through Switching (VCT)

When a message comes in an intermediate node and if its outgoing channel is free then it can be transmitted out immediately. It actually does not need to be completely received in the node before being transmitted out. Buffers are required when a busy channel is encountered. Virtual cut through switching method is hybrid mixture of circuit switching & packet switching. If the packet encounters busy channels at all of the intermediate nodes, the outcome is exactly same as in packet switched network. On the other hand if all of the intermediate channels are free, the outcome is similar to a circuit switched system. When network load is low Virtual cut through switching is most advantageous.

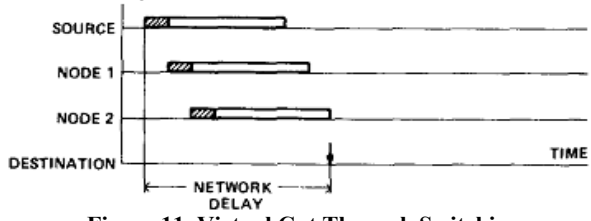

Figure 11. Virtual Cut Through Switching.

4. SELECTION OF NOC TEST CASES

Researchers were selected a small number of test cases, those were representative of a large family of network architectures [5]. The circuit switched network presents an extreme case,to compare against the other end of architectures, the speculative, single cycle, virtual channel design was selected as the second test case. After consideration of the two extremes of architecture scale, a mid way design was selected to provide comparisons to the opposing extremes. The wormhole flow control based router architecture was selected for this purpose.

4.1 Circuit Switched Router Design

In the CS router a very simple data path is provided, that is composed of crossbar with registered outputs. Every output port is 64 bits wide. In order to provide more flexibility each output port of 64 bits is divided into four, 16-bit wide lanes. The configuration memory provides the entry capacity of 20, with 5-bits per entry .Out of 5-bits, four bits are address bits used to identify an input lane and last bit is the valid bit.It is shown in figure 12.

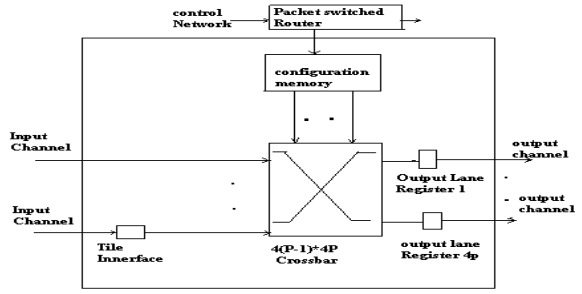

Figure 12. Circuit Switched Router

4.2 Virtual Channel Router Design

Each flit identifies VC by using a one hot encoded 4 bit VC identifier. Each flit additionally carries a 5-bit, one hot encoded next port identifier that is used by the downstream router. The architecture does not provide a separate *head* flit and every flit therefore identifies its destination X and Y address (4-bits each) and carries an additional single bit to indicate whether it's a *tail* flit or not. If all these bits combined with the 64-bit data path it will results in total flit size of 82-bits.

The Switch and VC allocators based on matrix arbiters. Its function is to allocate VCs and crossbar ports for the next clock cycle if required.Figure13 shows the structure of this router.

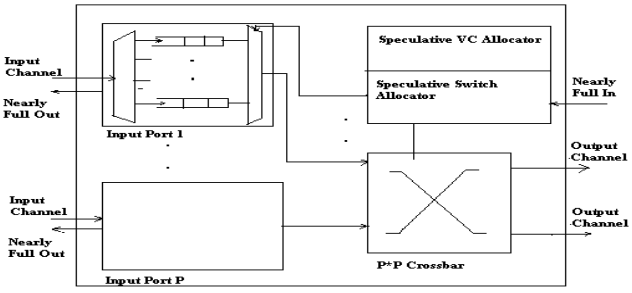

Figure 13 Virtual Channel Router.

4.3 Wormhole Router Design

The WH router is nothing but simplified SpecVC router. As in the SpecVC router, no separate head flit was included and thus, besides the data payload ,each flit carries a 5-bit ,one hot encoded next port identifier ,X and Y destination addresses(2-bits each) & 1-bit to identify tail flits. By combining all these with the 64-bit data path results in total flit size of 74-bits.Structure of the Wormhole router is as shown in Figure 14.

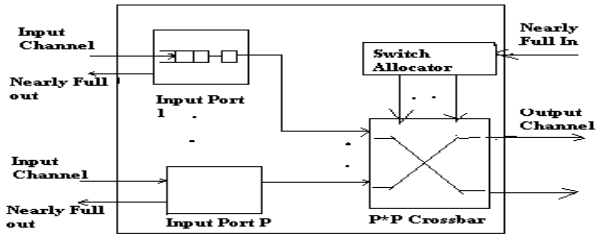

Figure 14. Wormhole Router.

5. Comparison of VCT with WH

Arnab Banerjee et al. [5] analyzed NoCs in an energy perspective view by accurately modeling a circuit switched router. a wormhole router & a speculative virtual channel router. Only limited range of architectures were used which limits the result that could be obtained. The three networks were selected. The circuit switched (CS) network, Speculative virtual-channel (SpecVC) network, & Wormole network. Observations are done:

Power at fixed throughput: SpecVC router dissipates a lot much power than the other design .However the power dissipation of WH & CS is seen to be very similar. The router power is more than the link power for all designs & is significant for SpecVC design.

Packet Energy under no congestion: Router energy is similar amount to the link energy for all the designs. The datapath components dominate over the control elements in all designs.

Packet energy under congestion: The link energy is not determined simply by the contents of flit but also by the contents of pervious flit or any idle periods before the flit.

6. Conclusion

Router power found to be significant overhead beyond the link power. Virtual cut through (VCT) switching gives lower latencies at higher acceptance rate than Wormhole (WH) switching whereas for a particular number of virtual channels VCT & WH consume almost identical power. Therefore Virtual cut through (VCT) switching gives better performance at similar power consumption in comparison with Wormhole (WH) switching.

7. REFERENCES

[1] Nilanjan Banerjee, Praveen Vellanki and Karam S Chatha." A Power and Performance Module for Network on Chip Architectures", In DATE'04: *Proceedings of the Conference on Design, automation and test in Europe,* page 21250, Washington, DC, USA, 2004.IEEE Computer Society

[2] T.T.Ye, L. Benini, G. D. Micheli "Analysis of Power Consumption on Switch Fabrics in Network routers". In Proc. Design Automation Conference, 2002

[3] Partha Pratim Pande, Michael Jones,,Andre Lanov,andResveSaleh,"Performance evaluation and Design Trade–Offs for Network –on Chip Interconnect Architectures" 15 June 2005

[4] Parviz Kermani and Leonard Kleinrock "Virtual Cut–Though: A New Computer Communication Switching Technique "p 267-269.

[5] Arnab Banerjee, Robert Mullins and Simon Moore "A Power and Energy Exploration of Network-on–Chip Architectures". Proceedings of the First International Symposium on Networks-on-Chip, 2007

[6] Orion: A power Performance Simulator for interconnection network by Hang-sheng wang , XinpingZhu , Li-Shiuan Peh Sharad Malik[7] L-S.Peh and W.J.Dally.A delay model and speculative architecture for pipelined routers.In Proc International Symposium on High Performance Computer Architecture,2001

[8] Hang Sheng Wang Xinping Zhu Li- Shiuan Pet Shard Malik, "Orion: A Power – Performance Simulator for Interconnection Network" In International symposium on Microarchitecture,Istanbul,Turkey, November 2002

[9] A.G.Wassal and M.A.Hasan, "Low Power System Level design of VLSI packet switching fabrics ", I*EEE Transactions on CAD*.20: 723-738, June2001.

[10] R.D.Mullins, A.F.West,and S.W.Moore. The design and implementation of a low- latency on-chip network. In the *Proceedings* of 11th Asia and Suoth Pacific Design Automation Conference (ASP-DAC), 2006.