The 2nd International Workshop on Design and Performance of Networks on Chip (DPNoC 2015)

Design Trade off and Performance Analysis of Router Architectures in Network-on-Chip

Jawwad Latif\textsuperscript{a}, Hassan Nazeer Chaudhry\textsuperscript{b}, Sadia Azam\textsuperscript{c}, Naveed Khan baloch\textsuperscript{d}

\textsuperscript{a-c}University of Engineering and Technology Taxila 47080, Pakistan
\textsuperscript{b}Technische Universität Darmstadt, Germany

Abstract

On chip interconnection networks simplify the challenges of integrating large number of processing elements. Routers are backbone of networks. Buffers and crossbar in router consumes significant area and power of network. Reducing buffers could lead to degradation of network performance. Dual Xbar router architecture combines buffered and bufferless feature to reduce buffer read/write energy with dual crossbars. While Switch folding technique introduced to reduce wire density and decrease muxes in crossbar by increasing resource utilization. In this paper, we propose Folded Dual Xbar architecture by combining the Dual Xbar and Folding technique in order to get advantages of both architectures. Performance of architectures is evaluated using OMNET++ platform under different load conditions. Simulation results shows that there is slight increase in throughput and reduction in buffer read/write energy by average 46\% at high loads in proposed 2-Folded Dual Xbar as compared to conventional architecture. Proposed 3-Folded Dual Xbar results at least 16.6\% increase in throughput as compared to conventional architecture with 43-45\% reduced buffer read/write energy but slight increase in crossbar. Throughput of 3-Folded Dual Xbar decreased only by 5-7\% as compared to Dual Xbar with distributed wire density advantage.

© 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Peer-review under responsibility of the Conference Program Chairs

Keywords: Network-on-Chip; Dual Xbar; Folding technique; Router architectures

1. Introduction

Due to the advancement in deep-submicron technology, processing elements on chip are increasing with the passage of time\textsuperscript{1}. Traditional System-on-Chip (SoC) technology uses dedicated wires for communication between IP

* Corresponding author. Tel.: +92-345-515-0958 ; fax: +0-000-000-0000 .
E-mail address: jaidpk@hotmail.com
cores. The trend of increasing the processing elements on chip raises communication complexity and affects the scalability of on chip systems. The parasitic effects becomes dominant with large number of interconnects\(^2\). Cost of SoC system is highly influenced by interconnects which pays important role in power consumption, performance and overall size of the system\(^3\). To minimize the limitations of SoC technology, Network-on-Chip (NoC) is introduced to improve communication infrastructure in SoC\(^3\). In NoC, the dedicated wires for interconnection are being replaced by network with few numbers of links and routers to route the packets from source to destination. NoC offers flexibility to integrate large number of IP cores on chip\(^3\). NoC supports Globally Asynchronous Locally Synchronous (GALS) architectures which are made up of locally synchronous modules which are connected through globally asynchronous network. This concept breaks the single clock domain into multiple clock domains. Distributed nature of NoC is well suited by multiple clock domains\(^4\).

2. Related Work

Buffers and crossbar are the most important components in routers architecture. They consume significant portion of area and power in a network which determines the cost of the system. Distributed-shared buffer (DSB) was proposed in which author emulated Output-buffered router (OBR) based on distributed-shared buffer architecture of router\(^8\). This mechanism uses two crossbars and memory elements are placed between them. For synthetic traffic pattern, DSB router\(^9\) has 19% increase in saturation throughput and more than 94% ideal saturation throughput. Increase in throughput gives low network latency. For SPLASH-2 benchmark, DSB router has achieved 60% reduction in network latency. Mechanism of bypassing the router to reduce per hop delay was introduced earlier using traditional router. This technique although reduces per hop latency but increases the complexity and cost of router. Low-Cost Router Micro architecture\(^10\) reduces the complexity and the cost of router while maintaining the bypass mechanism to reduce per hop latency. An iDEAL architecture\(^11\) uses dual function (three-state repeaters) links, which can be used for transmission and data storage. When there is congestion in the network, three state repeaters hold the data until congestion alleviated. With adaptive dual function links approximately 30% savings in overall network power and 40% savings in buffer power have been achieved with only 1-3% drop in performance. SCARAB router architecture\(^12\) was introduced to maintain the power and advantages of bufferless router with efficient results. Different mechanisms were proposed to support dropping protocol in order to reduce packet drops and retransmission of packets. Analysis of Blind Packet Switched (BPS), SCARAB router and Hot-Potato (HP) router is done in this paper. SCARAB has shown 12.6% more efficiency than BPS and 18.3% more than HP. The main drawback of SCARAB is the area overhead. During execution time not all routers have packets to store. Buffer utilization can be made more efficient by sharing buffers between ports. In RoShaq architecture\(^13\), buffer utilization was improved by sharing multiple buffer queues between input ports. This technique results in 14 % increase in saturation throughput over traditional router. Performance, power and energy evaluation of multiple architectures are also discussed\(^14\). QoS-aware and Congestion Aware router architecture\(^15\) provides quality-oriented transmission and improves throughput by balancing traffic load over network. DXbar\(^6\): an innovative dual-crossbar design has been proposed by combining the buffered and bufferless features in router. Time division multiplexing technique\(^7\) is introduced to decrease crossbar area and wire density by increasing the utilization of resources.

3. Proposed Architecture

In this paper, proposed architecture is Folded Dual Xbar. The idea is to combine two techniques Dual Xbar and Switch folding technique. Dual Xbar has combined buffer and bufferless feature while Switch folding gives area benefits and reduce wire density by time division multiplexing and resource utilization. Proposed architecture Folded Dual Xbar utilizes both concepts. It has buffer and bufferless feature like Dual Xbar and time division multiplexing of Switch folding technique. Therefore, this architecture gets the benefits of reducing buffer read/write energy while having benefits of Switch folding mechanism. Design of 2-Folded Dual Xbar is shown in Fig. 1(a). It has dual folded crossbars. One is primary folded crossbar which has bufferless connections with input ports. When packet arrives and corresponding output port is idle then router allows it to pass through primary folded crossbar which has bufferless feature. Second is secondary crossbar which has buffers connected to its input ports. When packet arrives and its corresponding output port is busy then the packet is stored in buffer and passes through secondary crossbar after receiving
permission from scheduler. Crossbars are also folded in which it increases resource utilization through time division multiplexing. Implementation of 2-Folded Dual Xbar is described in flowchart shown in Fig. 1(b). Router receives incoming packet and forward it to output port calculation module. XY routing is used to determine the out port. After receiving the out port value from XY routing module, packet checks whether the corresponding out port is busy or free. A single cycle plenty can be reduced by using look ahead routing for output port calculation as implemented. If the corresponding output port is idle than packet is moved through primary folded crossbar. If the output port is busy than packet is stored in FIFO input buffers and request is forwarded to the arbitration unit. Arbiter implements round robin scheduling algorithm. When the out port becomes free, scheduler sends grant to inport unit. Inport unit receives grant and moves the corresponding packet through secondary folded crossbar. If any situation arises in which primary folded crossbar continuously send packets, causing packets in buffer to wait for long duration then fairness between secondary packets and primary packets can also be applied by using appropriate value of counter. In network, crossbar failure is possible. Fault Tolerance at hardware level can be applied in Dual crossbar architecture\(^6\). If any crossbar fails, the routing could be shifted to either crossbar. Architectures are implemented using OMNeT++ platform.

Fig. 1. (a) 2-Folded Dual Xbar architecture; (b) Packet routing in Folded Dual Xbar

4. Evaluation

Our analysis includes the performance evaluation of three router architectures. 1) Conventional architecture 2) Dual Xbar and 3) Folded Dual Xbar. HNOCS\(^6\) simulator is used for conventional router. Folded Dual Xbar further consists of different number of folding switches. 2-Folded Dual Xbar contains two muxes in each crossbar and 3-Folded contains 3 muxes in each crossbar. Comparison of 2-Folded Dual Xbar, Dual Xbar, 3-Folded Dual Xbar and conventional router is being done. Throughput and latency graphs are plotted by applying Hotspot traffic pattern and Uniform Random traffic pattern. In order to estimate buffer read/power reduction, buffered and bufferless events are also counted and shown in Table. 1. Analysis is done on 4x4 2D mesh topology network using OMNeT++ platform.

4.1 Uniform Random Traffic Pattern

We compare conventional architecture with 2-Folded Dual Xbar. In our analysis, conventional router contains 2 virtual channels with a capacity to store 5 flits. Each packet is divided into 10 flits. The size of each flit is 4Bytes. On the other hand, there is no division of packets in 2-Folded Dual Xbar network. Each flit is considered as a packet and each packet is of 4Bytes. 2-Folded Dual Xbar architecture consists of one FIFO buffer with a capacity of storing 10 flits. Uniform Random traffic pattern is applied and value of throughput and latency is measured on multiple offered loads. To estimate that how much buffer read/write events are reduced in 2-Folded Dual Xbar, buffered and bufferless events are also counted and shown in Table. 1. The graph showing throughput comparison between conventional architecture and 2-Folded Dual Xbar is shown in Fig. 2. Throughput is calculated from equation 1. Offered load is applied from 0.5 to 1.3 where both networks get saturated. They have shown almost same behaviour on all offered...
\[ \text{Throughput} = \frac{\sum_{i=1}^{\text{Total cores}} \text{Received packets} \times \text{Packet length in flits} \times 4}{\text{Total cores} \times \text{Total simulation time} \times 10^6} \text{[GB/s]} \] (1)

loads except a slight increase in throughput by 2-Folded Dual Xbar at saturation. Conventional router with 2 virtual channels has saturated throughput 1.02 and 2-Folded Dual Xbar network get saturated at 1.07. The main advantage of 2-Folded Dual Xbar over conventional router is the bufferless feature to reduced buffer read/write energy. The crossbar in 2-Folded contains total 4 switch elements with slight overhead of distributor circuit at outputs of crossbar. While conventional router has 5 switch elements in its crossbar. All flits in conventional are always stored in buffer/vc while 2-Folded Dual Xbar reduces buffer read/write energy by directly sending the packet through primary crossbar when the corresponding output is idle. This feature is adopted from Dual Xbar6. Switch folding7 is applied to reduce wire density and mixes by increasing resource utilization. The percentage of bufferless events at high loads 1.06, 1.14 and 1.33 are 51%, 45.6% and 40%. An average of 46% buffer read/write energy is saved at high loads in 2-Folded Dual Xbar as compared to conventional architecture with almost same throughput. On other hand, the wire density between the input ports and crossbar also decreased in 2-Folded Dual Xbar as compare to conventional architecture which improves layout density of the circuit7. Latency of every packet is measured and then converted into average latency per byte through the equation given as:

\[ \text{Latency} = \frac{\sum_{i=1}^{\text{Total cores}} \text{Latency of received packets}}{\sum_{i=1}^{\text{Total cores}} \text{Total bytes received}} \text{[nsec/byte]} \] (2)

Latency graph is also plotted shown in Fig. 3. Graph shows the behaviour of latency in conventional and 2-Folded Dual Xbar architecture. The values of throughput are same from 0.5 to 0.94 offered load so the latency values as well. At 0.94 offered load, there is a slight increase in latency of 2-Folded Dual Xbar because its throughput is less by 0.01 from conventional router. Interesting to see that only at 1.0 offered load the latency of 2-Folded Dual Xbar is more than conventional router which could be the result of worst case or average case scenario in 2-Folded Dual Xbar in which all four input ports of router receives packet to forward them in all four different directions. After 1.0 offered load the latency of 2-Folded Dual Xbar decreases because its saturated throughput is 1.07 while conventional router has almost 1.05 throughput value. Throughput and Latency comparison of three architectures conventional, 3-Folded Dual Xbar and Dual Xbar is also shown in Fig. 4. The specifications are same as described in previous comparison. 3-Folded Dual Xbar contains 6 switches in its crossbar with slight overhead of distribution circuit and conventional router contains 5 switches in its crossbar. According to measured statistics in OMNET++, 3-Folded Dual Xbar has 16.6% increases in throughput as compared to conventional router having 2 virtual channels. Dual Xbar and 3-Folded Dual Xbar has bufferless feature while conventional router always stores flit into buffer. At saturation level, 43 – 45% events are bufferless in 3-Folded Dual Xbar which reduces significant amount of buffer read/write energy. We have achieved 16.6% increase in throughput and 43-45% saves in buffer read/write energy as compared to conventional router with 2 virtual channels. Dual Xbar has saturated throughput 1.37 which is 23.3% more than conventional router and 8% more than 3-Folded Dual Xbar. Dual Xbar also reduces buffer read/write energy by 52% at saturation level. Dual Xbar has two crossbars which consumes significant area and increases wire density between crossbar and input ports. So, we have more options for area selection through different folding technique with buffer power reduction as well. Latency graph is also shown in Fig. 5. It shows similar trend as expected from throughput graph. Traditional router’s latency with 2 virtual channels continues to increase after 0.94 offered load. Increase in throughput of Dual Xbar resulted in decrease in latency as compared to 3-Folded Dual Xbar after 1.14 offered load.

4.2 Hotspot Traffic Pattern

Specifications of packets size and virtual channels are same as in Uniform Random. The throughput comparison between conventional and 2-Folded Dual Xbar is shown in Fig. 6. Offered load is applied from 0.5 to 1.3 where both networks get saturated. The final saturated throughputs of both networks are same but the interesting point is that the 2-Folded Dual Xbar reaches its saturation level too early than conventional router. At 0.94 offered load, 2-Folded Dual Xbar network reaches at its saturation level having throughput 0.91 and remains same on all other higher offered loads. It gives greater throughput than conventional router with 2 virtual channels on almost all offered loads except at saturation level where they both have same value of throughput. At saturation level, 2-Folded Dual Xbar has ability to save 40% buffer read/write energy. Latency graph is also shown in Fig. 7. Conventional router has shown significant increase in latency. Throughput performance of three architectures is shown in Fig. 8. 3-Folded Dual Xbar has 15.8%
Fig. 2: Throughput of Uniform Random traffic pattern

Fig. 3: Latency of Uniform Random traffic pattern

Fig. 4: Throughput of Uniform Random traffic pattern

Fig. 5: Latency of Uniform Random traffic pattern

Fig. 6: Throughput of Hotspot traffic pattern

Fig. 7: Latency of Hotspot traffic pattern

Fig. 8: Throughput of Hotspot traffic pattern

Fig. 9: Latency of Hotspot traffic pattern
increases in throughput as compared to conventional router. At saturation level, 47 – 50% events are bufferless in 3-Folded Dual Xbar which saves significant amount of buffer read/write energy. We have achieved 15.8% increase in throughput and 47-50% saves in buffer read/write energy as compared to conventional router. Dual Xbar reduces buffer read/write energy by 55% at saturation level. Latency graph is also shown in Fig. 9. Conventional router’s latency continues to increase after 0.57 offered load. Increase in throughput of Dual Xbar resulted in decrease in latency as compared to 3-Folded Dual Xbar. The saturation throughput in conventional router, 2-Folded Dual Xbar, 3-Folded Dual Xbar and Dual Xbar is decreased by 14.2%, 15%, 15% and 16.7%. Comparison of all three architectures is summarized in Table 1.

Table 1. Performance summary

<table>
<thead>
<tr>
<th>R.Arch</th>
<th>Uniform Random</th>
<th>Hotspot</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Max. TP</td>
<td>B.E</td>
</tr>
<tr>
<td>H-2VC</td>
<td>1.05</td>
<td>0</td>
</tr>
<tr>
<td>2-FDX</td>
<td>1.07</td>
<td>40</td>
</tr>
<tr>
<td>3-FDX</td>
<td>1.26</td>
<td>44</td>
</tr>
<tr>
<td>DX</td>
<td>1.37</td>
<td>52</td>
</tr>
</tbody>
</table>

5. Conclusion and Future Work

Our work briefly compares the performance of multiple architectures using OMNeT++ platform and also estimated possible parameters to compute energy consumption of buffers. In future, we are interested to do comprehensive analysis of area and power using ORION 3.0 model. Furthermore, these architectures can also be evaluated on different topologies and routing mechanisms to find out the better performance for specific applications.

References