Asian Journal of Information Technology 13 (10): 575-581, 2014
ISSN: 1682-3915

Performance Evaluation of Low-Power, High-Performance Serial On-Chip Communication Link Router

R. Anitha and P. Renuga
[1]Faculty of ECE, PERI Institute of Technology, Chennai, Tamilnadu, India
[2]Faculty of ECE, Thiyagarajar College of Engineering, Madurai, Tamilnadu, India

Abstract: In this research, researchers introduce a low power and high performance serial on-chip communication link based on innovative design techniques and its design methodologies are presented in this research work. The proposed semi-serial link is designed using high speed serialization/deserialization and multi-orthogonal encoding techniques. The link also employs acknowledgement scheme to maintain the high speed data intake from the serializer. The proposed semi-serial link is analyzed and compared with bit-serial and fully bit-parallel links for 64 bit data communications. The results show that the proposed semi-serial link dissipates the lowest energy per bit compared to fully bit-parallel links at the same time achieving the same performance. The proposed semi-serial on-chip is designed and simulated in Xilinx Project navigator and tested on various FPGA devices using 90 nm CMOS technology.

Key words: NoC, on-chip topology, packetization, router architecture, routing techniques

INTRODUCTION

Network on Chip (NoC) technology is rapidly displacing traditional bus and crossbar approaches for System on Chip (SoC) interconnect. NoC has been proposed as a new design paradigm to solve the communication restrictions. NoCs provide a scalable and modular architecture and help independent design of IP cores and its re-use. In an FPGA, area is available at a premium and hence the on-chip communication network should be as small as possible. This ensures that the maximum area can be utilized by the logic while maintaining the performance of the on-chip network. Also, reduction in the logic blocks used in FPGAs has a direct impact on the power consumption and the timing. The central component of NoC architecture is a router and hence, it is prudent to make its area smaller. The network area can be reduced by using a simple router supporting complete functionality, without sacrificing the performance and by reducing the number of routers, without reducing the number of communicating logic cores.

The increasing interconnection complexity and the known scalability deficiency of buses require another model of interconnection. The communication among cores of a SoC having reusable and scalable interconnections is being provided by Networks on Chip (NoC). NoCs have been proposed to integrate several IP (Intellectual Property) cores providing high communication bandwidth and parallelism. NoC are basely composed of routers, network interfaces and links. The router contains buffers, flow controllers and crossbar. Buffers are the active elements in router input channels. They consume about 64% of the total power, making them the largest source of leakage energy consumed in NoCs. A router consumes more energy to store the data rather than to transmit them. However, the effective and resourceful management of buffers in NoC routers has a significant impact in performance and efficiency of interconnection networks.

The design of a NoC can be done in a huge design space where the number of interconnects and the buffer size are the two main parameters. Interconnect is a very expensive resource in large Systems on Chip (SoC) in which more area and power are consumed and high delays relative to gate delays and clock cycles occur. Networks on Chip (NoC) were developed as a solution for the SoC interconnect problem. Since, the physical properties of each link are configured solely based on the specific requirements of the link and independently from other links, the total number of wires in the interconnect is reduced up to 40% resulting in smaller design area and considerably faster timing convergence.

Literature review: An adaptive architecture with runtime observability has been proposed to avoid faults in NoCs, providing adaptability both at the system-level and at the architecture level (Faruque *et al*., 2008). At

Corresponding Author: R. Anitha, Faculty of ECE, PERI Institute of Technology, Chennai, Tamilnadu, India

system level the architecture can re-map the system tasks and at architecture level it can re-route the packets and re-allocate the Virtual Channel Buffers (VCB). The presence of faults triggers the need of NoC adaptation at architecture level. The changes at architecture level are based on the occurrence of faults and these events occur when packets do not reach the destination or when the VCB is full. This adaptive process occurs only when a fault occurs and hence no performance or power advantage can be obtained during the normal operation of the system.

In a serial on-chip link has been designed by adopting circuits that had originally been used for off-chip communications. It uses output multiplexed transmitter architecture due to its ability to deliver better performance than input multiplexing. However, this comes at the expense of a much higher output capacitance that grows linearly with the bit-width. Both transmitter and receiver use multi-phase DLL circuits and clock calibration is required at the receiver side. They have fabricated a prototype chip in 180 nm CMOS technology and a 3 mm link has achieved a throughput of 8 Gbp (Jose *et al*., 2005).

A high-speed serial link was presented by Lee *et al*. (2005). The serializer/deserializer is based on chain of MUXes. The link is single-ended and employs wave-pipelining. As a timing reference, constant delay elements are used instead of a clock. Furthermore, the operation is based on the assumption that the introduced unit delays for the serializer and deserializer are the same. However, getting the same delay is almost impossible in the sub-100 nm CMOS technology due to considerable Process, Voltage and Temperature (PVT) variations which in turn affect the reliability of communication. A test chip was manufactured using 180 nm CMOS technology and the measured throughput was 3 Gbp.

In various interconnection networks were designed and evaluated in different network routing protocols (Duato *et al*., 1997). The results were analyzed and compared with existing networks.

A reconfigurable router was designed for wireless media access by Matos *et al*. (2011). But this research was entirely based on four ports only and complex architecture in its basic internal structure. It also consumed >78% hardware utilization.

A virtual channel regulator has been designed for Network on Chip Routers (Nicopoulos *et al*., 2006). Their research dealt with different configuration protocols in a linear manner and the results were compared against virtual buffer regulator of routers.

MANET's distributed architecture and changing topology and a traditional centralized monitoring technique were analyzed in MANETs acknowledgment based approach (Liu *et al*., 2007) for the detection of routing misbehaviour.

MATERIALS AND METHODS

Proposed link architecture: A basic communication link of serial on-chip link includes sending and receiving routers, serializer, deserializer, multi-orthogonal encoder and decoder. Data is sent from source router to input channel of the communication node through serializer and encoder and then it is transmitted to the corresponding output router. The block diagram of the proposed serial on-chip link router is shown in Fig. 1.

Sending side router: The router at the sending side includes buffers, switches and control units required for storing and forwarding data from the input ports to the preferred output port. The router comprises a smaller area and buffer size and is shown in Fig. 2. The router architecture has several internal modules which include Packet Counter Module (PCM), Address Counter Module (ACM), Address Differentiation Module (ADM) and Virtual Channel Buffer (VCB). The wXY routing technique is followed as the routing methodology. Consider the tuple Γ = {N, S, W, E, NE, SE, SW, NW, D_{IN}}. Each direction $d \epsilon \Gamma$ has a weight W_d and available bandwidth B_d with $B_d \leq B_{max}$ and B_{max} is the maximum link capacity. The

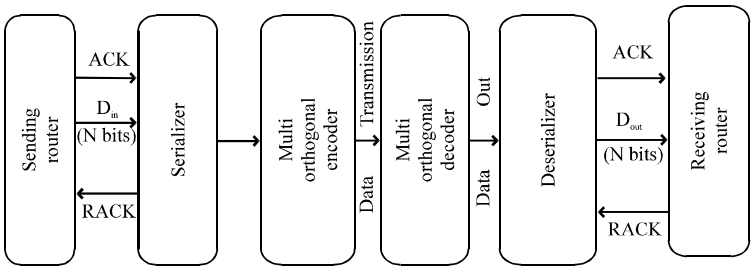

Fig. 1: Overall block diagram of proposed serial on-chip link design

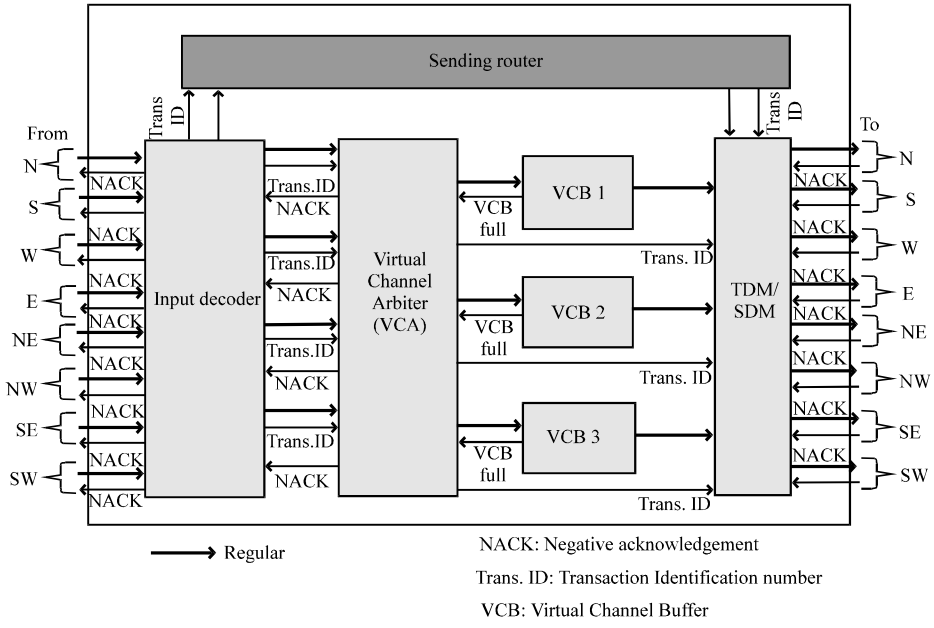

Fig. 2: Router architecture at sending side

current router coordinates are (x,y) and every packet D_{IN} has destination coordinates x_d, y_d and a required bandwidth B_p. The weights are assigned using Eq. 1-8:

$$w_N = \begin{cases} B_N \times |y_d - y| + B_{max} & y_d - y < 0 \\ 0 & B_N < B_p \\ B_N & \text{else} \end{cases} \quad (1)$$

$$w_S = \begin{cases} B_S \times (y_d - y) + B_{max} & y_d - y > 0 \\ 0 & B_S < B_p \\ B_S & \text{else} \end{cases} \quad (2)$$

$$w_W = \begin{cases} B_W \times |x_d - x| + B_{max} & x_d - x < 0 \\ 0 & B_w < B_p \\ B_w & \text{else} \end{cases} \quad (3)$$

$$w_E = \begin{cases} B_E \times (x_d - x) + B_{max} & x_d - x > 0 \\ 0 & B_E < B_p \\ B_E & \text{else} \end{cases} \quad (4)$$

$$w_{NE} = \begin{cases} (B_N + B_x) + B_{max} & x_d - x > 0 \\ 0 & B_E < B_p \\ B_E & \text{else} \end{cases} \quad (5)$$

$$w_{NW} = \begin{cases} (B_N + B_W) \times (y_d + x_d) + B_{max} & y_d + x_d > 0 \\ 0 & (B_N + B_W) > B_p \\ B_N + B_W & \text{else} \end{cases} \quad (6)$$

$$w_{SE} = \begin{cases} (B_S + B_E) \times (y_d - x_d) + B_{max} & y_d - x_d < 0 \\ 0 & (B_S + B_E) < B_p \\ B_S + B_E & \text{else} \end{cases} \quad (7)$$

$$w_{SW} = \begin{cases} (B_S + B_W) \times (y_d - x_d) + B_{max} & y_d - x_d > 0 \\ 0 & (B_S + B_W) > B_p \\ B_S + B_W & \text{else} \end{cases} \quad (8)$$

The route r is chosen with the highest weight as given by:

$$r = \begin{cases} D_{in} \;\&\; x = x_d \text{ and } y = y_d \\ i \in \{N,S,W,E,NE,SE,NW,SW\} \\ W_d = \max_d(W_d) \quad \text{else} \end{cases} \quad (9)$$

Serializer: The proposed serial on-chip link deals with the problems of collision and limited power transmission as described. Consider a link in the case of receiver side collisions as shown in Fig. 3. After node A sends packet 1 to node B, it detects if node B forwarded this

packet to node C, at the same time, node X is forwarding packet 2 to node C. In such situations, node A listens to node B whether it has successfully forwarded packet1 to node C but failed to detect that node C did not receive this packet due to a collision between packet 1 and 2 at node C.

In the case of limited transmission power as shown in Fig. 4, with the purpose of saving its own battery power, node B purposely limits its transmission power so that it is strong enough to be overheard by node A but not strong enough to be received by node C.

For false misbehaviour report, although, node A successfully overheard that node B forwarded packet 1 to node C, node A still reported node B as misbehaving as shown in Fig. 5.

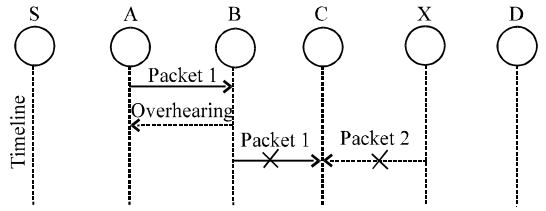

Fig. 3: Receiver collision: both nodes B and X are trying to send packet 1 and 2, respectively, to node C at the same time

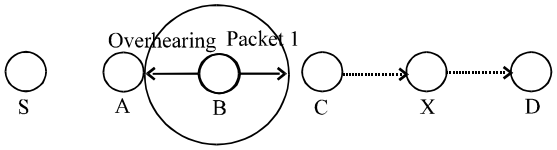

Fig. 4: Limited transmission power: node B limits its transmission power

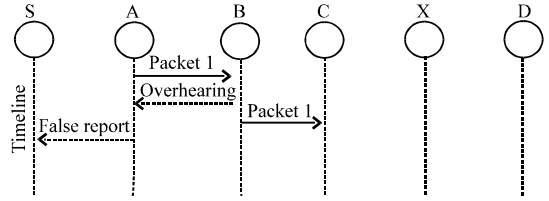

Fig. 5: False misbehaviour report: node A sends back a misbehaviour report even though node B forwarded the packet to node C

Multi-orthogonal encoder and decoder: The block diagram of the encoder is shown in Fig. 6 and consists of multi code generator, orthogonal multiplexers- i, j, k, l and orthogonal multipliers. The constant amplitude encoder possesses 3 parity bits namely, L*, L_1, L_0 generated from 3 groups of parallel bits (i*, i_1, i_0), (j*, j_1, j_0), (k*, k_1, k_0) according to the following equation:

$$\begin{aligned} L^* &= I^* \hat{} j^* \hat{} k^* \\ L_1 &= i_1 \hat{} j_1 \hat{} k_1 \\ L_0 &= i_0 \hat{} j_0 \hat{} k_0 \end{aligned} \tag{10}$$

The output of the orthogonal multiplexers i, j, k and l are given by:

$$\begin{aligned} i &= (0, 0, i_1, i_0) = (0, 0, 1, 0) \\ j &= (0, 1, j_1, j_0) = (0, 1, 0, 0) \\ k &= (1, 0, k_1, k_0) = (1, 0, 0, 0) \\ l &= (1, 1, L_1, L_0) = (1, 1, 1, 0) \end{aligned} \tag{11}$$

The multi code generator consists of two blocks, namely, serial to parallel converter and gold sequence generator. The serial to parallel converter converts the data bits in to number of branches according to the length of the gold sequence. As researchers know, the gold sequence can be generated by applying XOR operation over a few Pseudo random (PN) sequences as explained in Fig. 7.

Finally, the orthogonal multiplier multiplies the outputs of orthogonal multiplexer with orthogonal parity vector matrix 'b' as shown:

$$\begin{aligned} S = b\begin{bmatrix} c_i \\ c_j \\ c_k \\ c_l \end{bmatrix} &= \begin{bmatrix} i^* j^* k^* 1^* \end{bmatrix}\begin{bmatrix} c_i \\ c_j \\ c_k \\ c_l \end{bmatrix} \\ &= i^*.c_i + j^*.c_j + k^*.c_k + 1^*.c_l \end{aligned} \tag{12}$$

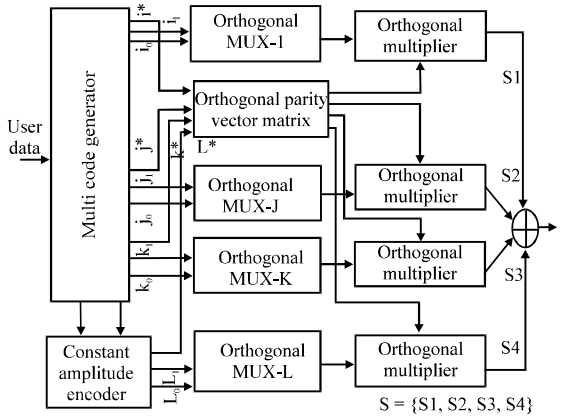

Fig. 6: Multi-orthogonal encoder block diagram

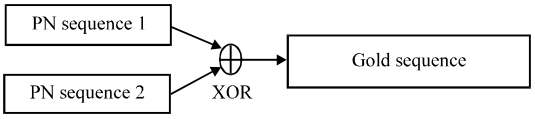

Fig. 7: Generation of gold sequence

Where:
c = The Walsh-Hadamard matrix
s = Final encoded sequence
b = Each individual path of sub branch.

The orthogonal decoder consists of a decision device, correlator banks, orthogonal demultiplexer and unicode generator as shown in Fig. 8.

The decision device receives the bits and produces an output 1 when the input is >0 and produces an output 0 when the input is <0. The correlator banks perform the operation of orthogonal multiplication. The generator consists of two blocks, namely, parallel to serial converter and gold sequence despreader which works conversely to the gold sequence generator and is explained in Fig. 9. The parallel to serial converter converts the data bits in to number of branches according to the length of gold sequence.

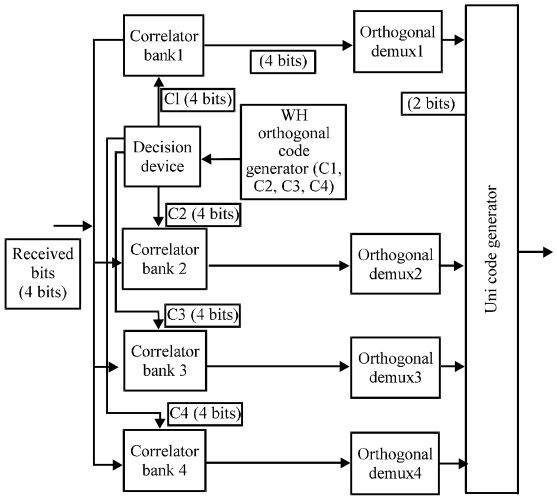

Fig. 8: Multi-orthogonal decoder block diagram

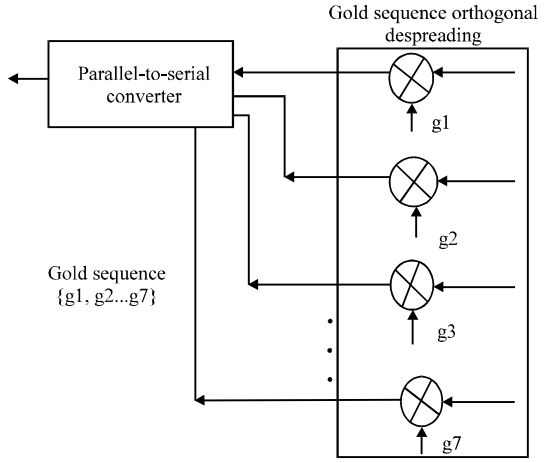

Fig. 9: Block diagram of a unicode generator

Operation of proposed link: In this study, researchers describe the working of the proposed serial on-chip link in detail. EAACK consists of three major parts, namely, ACK, Secure ACK (S-ACK) and Misbehaviour Report Authentication (MRA).

In the proposed scheme, both the source node and the destination node of a communication link are not malicious. Also researchers consider every link between the nodes to be bidirectional.

ACK: ACK is basically an end to end acknowledgment scheme. It acts as a part of the hybrid scheme in the proposed scheme, aiming to reduce network overhead when no network misbehaviour is detected (Fig. 10). In ACK mode, node S first sends out an ACK data packet P_{ad1} to the destination node D. If all the intermediate nodes along the route between nodes S and D are cooperative and node D successfully receives P_{ad1}, node D is required to send back an ACK acknowledgment packet P_{ak1} along the same route but in a reverse order. Within a predefined time period, if node S receives P_{ak1} then the packet transmission from node S to node D is successful. Otherwise, node S will switch to R-ACK mode by sending out an R-ACK data packet to detect the misbehaving nodes in the route.

R-ACK: The R-ACK scheme is an improved version of the TWOACK scheme proposed by Liu *et al*. (2007). The principle is to let every three consecutive nodes work in a group to detect misbehaving nodes. For every three consecutive nodes in the route, the third node is required to send an S-ACK acknowledgment packet to the first node. R-ACK mode detects the misbehaving nodes in the presence of receiver collision or limited transmission power and the process flow is shown in Fig. 11.

In R-ACK mode, the three consecutive nodes (i.e., F1-F3) work in a group to detect misbehaving nodes in the network. Node F1 first sends out R-ACK data packet P_{sad1} to node F2. Then, node F2 forwards this packet to node F3. When node F3 receives P_{sad1} as it is the third node in this three-node group, node F3 is required to

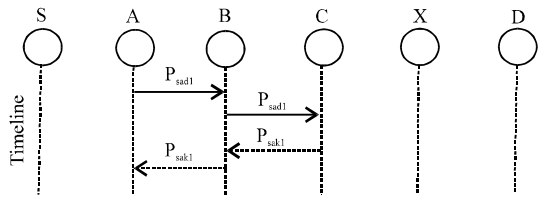

Fig. 10: ACK scheme: the destination node is required to send back an acknowledgment packet to the source node when it receives a new packet

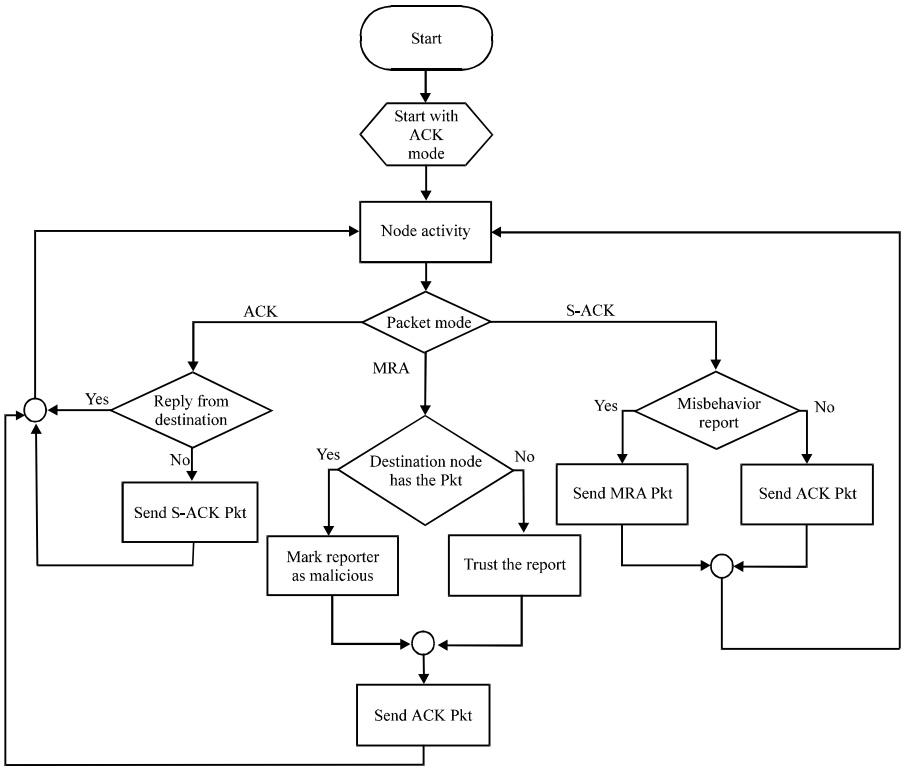

Fig. 11: R-ACK mode of operation in serial on-chip link

send back an R-ACK acknowledgment packet P_{sak1} to node F2. Node F2 forwards P_{sak1} back to node F1. If node F1 does not receive this acknowledgment packet within a predefined time period, both nodes F2 and F3 are reported as malicious. Moreover, a misbehaviour report will be generated by node F1 and sent to the source node S.

Unlike the TWOACK scheme, where the source node immediately trusts the misbehaviour report, the proposed method requires the source node to switch to MRA mode and confirm this misbehaviour report. This is a vital step to detect false misbehaviour report in the proposed scheme.

RESULTS AND DISCUSSION

With an increasing number of complex, non-uniform sized cores in a chip, high-throughput, energy and area efficient long range links become a necessity. In order to analyze the trade-off between bit-serial, semi-serial and fully bit-parallel long-range channels of a NoC, three types of fully bit-parallel links were designed. As the presented serial link is delay-insensitive, the self timed bit-parallel link was also designed as a delay-insensitive link using multi-orthogonal encoding and optimally repeated voltage-mode signalling.

The encoding was chosen due to its simpler and faster completion detection and data decoding logics than the conventional two phase dual-rail encoding. To evaluate the performance of the proposed schemes, uniform and non-uniform/localized synthetic traffic patterns are considered separately. In the non-uniform mode, 70% of the traffic is local requests, where the destination memory is one hop away from the master core and the remaining 30% of the traffic is uniformly distributed to the non-local memory modules. Researchers also consider the hotspot traffic pattern where four memory nodes are chosen as hotspots receiving an extra portion of the traffic (10%) in addition to the regular uniform traffic. For the uniform and hotspot traffic profiles, researchers obtained very similar performance gains in each configuration, though they are not presented due to the lack of space. For appraising the area overhead of the proposed architectures, each scheme is synthesized by Xilinx Project navigator tool using the 90 nm CMOS technology (Table 1). The utilization of power in all the components of the link is graphically plotted in Fig. 12.

The power consumptions of the sending router with multi-orthogonal encoder and receiving router with multi-orthogonal decoder are listed in Table 2. As can be seen from the Table 2, using this architecture for the proposed router is more beneficial (in terms of power and area) than using the master-side and slave-side models when each node is composed of a dedicated processor and memory.

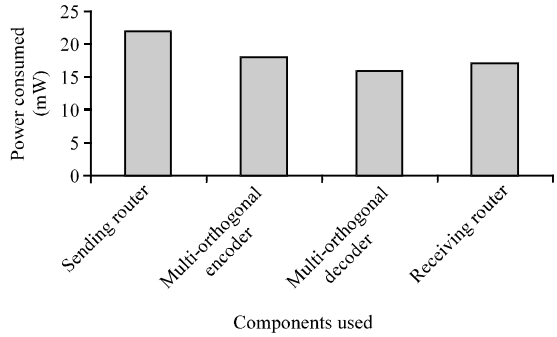

Fig. 12: Graphical plot of power utilization summary

Table 1: Utilization of power consumption for NI modules

Components	Power consumption (mW)
Sending router	22
Multi-orthogonal encoder	18
Multi-orthogonal decoder	16
Receiving router	17

Table 2: Requirements of slices for NI modules

Components	Slices requirements
Sending router with multi-orthogonal encoder	156
Receiving router with multi-orthogonal decoder	78

CONCLUSION

In this study, researchers have designed a low power and high performance serial on-chip communication link based on innovative design techniques and its design methodologies. The proposed semi-serial link is analyzed and compared with bit-serial and fully bit-parallel links for 64 bit data communications. The results show that the proposed semi-serial link dissipates the lowest energy per bit compared to fully bit parallel links at the same time achieving the same performance.

REFERENCES

Duato, J., S. Yalamanchili and L. Ni, 1997. Interconnection Networks, an Engineering Approach. IEEE Computers Society Press, Los Alamitos, CA., USA.

Faruque, A., M. Abdullah, T. Ebi and J. Henkel, 2008. ROAdNoC: Runtime observability for an adaptive network on chip architecture. Proceedings of the IEEE/ACM International Conference on Computer-Aided Design, November 13-13, 2008, San Jose, CA., USA., pp: 543-548.

Jose, A.P., G. Patounakis and K.L. Shepard, 2005. Near speed-of-light on-chip interconnects using pulsed current-mode signalling. Proceedings of the Symposium on VLSI Circuits Digest of Technical Papers, June 16-18, 2005, USA., pp: 108-111.

Lee, S.J., K. Kim, H. Kim, N. Cho and H.J. Yoo, 2005. Adaptive network-on-chip with wave-front train serialization scheme. Proceedings of the Symposium on VLSI Circuits Digest of Technical Papers, June 16-18, 2005, USA., pp: 104-107.

Liu, K., J. Deng, P.K. Varshney and K. Balakrishnan, 2007. An acknowledgment-based approach for the detection of routing misbehavior in MANETs. IEEE Trans. Mobile Comput., 6: 536-550.

Matos, D., C. Concatto, M. Kreutz, F. Kastensmidt, L. Carro and A. Susin, 2011. Reconfigurable routers for low power and high performance. IEEE Trans. Very Large Scale Integr. Syst., 19: 2045-2057.

Nicopoulos, C.A., D. Park, J. Kim, N. Vijaykrishnan, S. Yousif and C. Das, 2006. ViChaR: A dynamic virtual channel regulator for network-on-chip routers. Proceedings of the 39th Annual IEEE/ACM International Symposium on Microarchitecture, December 9-13, 2006, Orlando, FL., pp: 333-346.