Smart Intelligent Computing and Communication Technology
V.D. Ambeth Kumar et al. (Eds.)

doi:10.3233/APC210069

Comparative Study of Hybrid Optimizations Technique for On-Chip Interconnect in Multimedia SoCs

Jayshree[a,1], Gopalakrishnan Seetharaman[b], and Debadatta Pati[c]
[a,1] *Assistant Professor, Dept of ECE, National Institute of Technology, Nagaland, India*
[b] *Professor, Dept of ECE, Indian Institute of Information Technology, Tiruchirappalli, India*
[c] *Professor, Dept of ECE, National Institute of Technology, Nagaland, India*

Abstract. This paper presents the design and analysis of on-chip interconnect architectures for real time Multimedia Systems-on-Chip (MSoC) targeting Internet of Things (IoT) applications. The interconnect architecture provides high flexibility in connection for hardware implementation of reconfigurable neural network. Due to technology's miniaturization in ultra-deep submicron technology, the on-chip interconnect performance and power consumption become a bottle-neck. In this paper, the hybrid optimization technique is proposed to address these challenges using schmitt trigger as a repeater and tapering. Here, the proposed optimization technique is incorporated with a dedicated point to point based interconnection (PTP-BI) configuration. A comparative study with others without optimization technique (Model–I) shows the effectiveness of the proposed optimization technique (Model–II). The technology node scaling impacts are also analyzed for both techniques. Finally, the percentage reduction of latency and power consumption are evaluated in two different cases to observe the impacts of varying the interconnect length.

Keywords. Global on-chip inter connect, hybrid optimizations, internet-of-things (IoT), multimedia system-on-chip (MSoC), and point-to-point-based interconnection (PTP-BI).

1. Introduction

Nowadays, customer demands are accelerating towards high quality of service (QoS), reliability, scalability and security for real-time multimedia, specifically for internet of things (IoT) application [1]–[4]. The architecture design for such applications needs a high speed hardcore processor [5]. A more significant number of memories which follow the design methodology of intellectual property (IP) based blocks instead of the generic gate level [6] on single-chip within a limited silicon area. The design accomplishment is possible with the ongoing miniaturization of the device feature size. The scaling of process technology node has various advantages for device design such as gate latency (Gl) reduces by 1/s, where s is scaling factor [7]. It provides a system

[1]Jayshree, Assistant Professor, Dept of ECE, National Institute of Technology, Nagaland, India.
Email: jay.chy@nitnagaland.ac.in

designer to embed maximum IP for performing multitask on a single chip whereas the scaling leads to large, interconnect performance deterioration such as interconnect latency (Il) increases by s2, low reliability and massive power consumption due to global core interconnection. The global core interconnection creates large wire congestion on single chip [8]. Therefore, it motivates the system designer to shift from computation focused design to an interconnect focused (or communication-focused) model. In interconnect concentric system design there are few challenges such as core placement, floor planning, IP-core mapping, and routing of packets [9]. The main objective of the work is to resolve the IP cores placement, mapping challenge in the MSoC for transmission of the message without any loss and attack. It should provide high throughput, low latency and proper security for global on chip communication. The proposed dedicated point to point based interconnection (PTP-BI) is a more advanced on-chip interconnection technique than previously used technique [10].

The salient feature of the work contribution of the paper are as follows: Initially, every IP core of the video object plane decoder (VOPD) MSoC is designed. Here, two architectures of MSoC are designed using point to point based interconnection (PTP-BI) technique is the name as Model–I (considered as without optimization technique) and hybrid optimization of Schmitt trigger as repeater and tapering is the name as Model–II (considered as with optimization technique). The latency and power consumption for both models are examined at 0.350 μm, and 0.180 μm technologies for various interconnect lengths. Later, the impact of technology scaling on global interconnect has examined and evaluated the percentage reduction of latency and power consumption for Model–I and model–II. Here, the RTL design and synthesis are performed with Quartus-II 14.0 tools and hardware implemented on the Altera FPGA development board, and mentor graphics simulators.

The paper is organized as follows: Section 1 introduces the problem for degradation of the performance and security in MSoCs due to large size of the global interconnect. Background of on chip interfaces and interconnect related work is discussed in Section 2. Section 3 elaborates the proposed architecture and methodology of three on chip interconnect techniques. The results are discussed in Section 4. Finally, Section 5 presents the concise conclusion of the proposed approach.

2. Background

An extensive related literature is studied for on-chip interfaces interconnect architecture. Primarily, it aims to rectify the limitations of existing on-chip communication architectures. Next, the purpose of the study is related to review secured, high-performance, and low energy consumption architecture. The pros and cons of the highly competent technique are studied such as the interface is open core protocol (OCP) [11] and virtual component interface (VCI) [12].

The various parameter is analyzed using Vainbrand *et al.* [14], which is discussed below. x is considered as neurons arranged in PTP-BI, then the following point are as follows:

a) The link delay is expressed as

$$T(PTP-BI) = R_o C_o \bar{l}^2 \quad (1)$$

where, R_o and C_o are resistance and capacitance of the wire per unit length, respectively, and $\bar{l}$ is the average link length and expressed as $\bar{l}\,(PTP-BI) = 2l\frac{\sqrt{x}}{3}$, total number of links is expressed as, $\bar{l}_{total}(PTP-BI) = \frac{x(x-1)}{2}$.

b) The total area is expressed as

$$\text{Area}\,(PTP-BI) = L(PTP-BI)\times \bar{w} \quad (2)$$

where, $\bar{w}$ is constant number of wires per link, $L(PTP-BI)$ is the total length of all PTP-BI connections can be calculated as $\bar{l}_{total}(PTP-BI)\times\bar{l}_{total}(PTP-BI)$, here, $L(PTP-BI) = \frac{1}{3}l(x-1)x\sqrt{x}$.

c) The maximum link architecture frequency is expressed as

$$f_{arch}(PTP-BI) = \frac{1}{T(PTP-BI)} = \frac{1}{R_o C_o \bar{l}^2} \quad (3)$$

d) The power dissipation is mainly dynamic power dissipated on the link and gate capacitances and is expressed as

$$P_D(PTP-BI) \cong f_{arch}(PTP-BI)C_T U(PTP-BI)V_{dd}^2 \quad (4)$$

where, $U(PTP-BI)$ is utilization factor, V_{dd} is supply voltage and C_T is total capacitance in link.

3. Architecture and Methodology of Proposed On-chip Interconnection Technique for MSoC

In this section, the proposed interconnect architectures are discussed for VOPD multimedia benchmarks with the proposed methodology. The VOPD architecture is elaborated here with two proposed interconnections techniques.

3.1. Dedicated PTP-BI Technique for Cores Placement and On-Chip Communication

The implementation of the PTP-BI methodology for VOPD system architecture has been discussed in this subsection. This architecture is named as Model–I and its configuration is shown in Fig. 1. This configuration provides the most rapid transmission because of the dedicated link is present with all the macros [10]. In this design, three different types of PTP-BI are considered and classified as global, intermediate and local interconnects. In the first global interconnect, the size of the architecture is considered up to 8.5 mm for rendering power supply to the on-chip. Here, the height, width and thickness are taken in the ratio of 3 : 8 : 4. In the second intermediate interconnect, the size of the architecture is considered up to 4.5 mm for linking far distance IP cores to the on-chip. Here, the height, width and thickness ratio is 6 : 9 : 4. Similarly, for the third local interconnect, size of the architecture is

considered up to 1.5 mm to link the consecutive IP cores to on-chip. In this interconnect, the widths are considered as half of thickness and height.

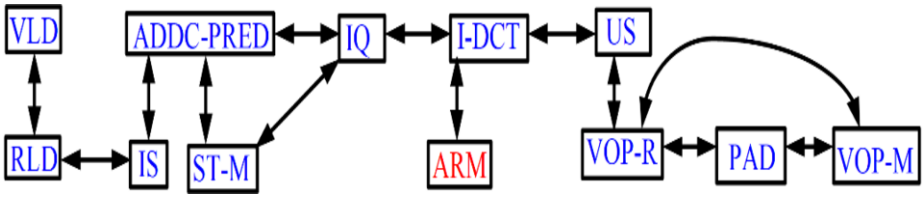

Figure 1. Model-I: The PTP-BI of VOPD without optimization.

3.2. Optimization with Schmitt Trigger as the Repeater and Tapering

In the case of without optimization technique, the effects of the resistance and capacitance are examined. Then, it is observed that the resistance and capacitance of interconnect rises linearly with an interconnecting length which leads to a quadratic increase in the latency. To reduce the latency and power consumption, a novel hybrid optimization technique is proposed. The implementation of two different optimization techniques such as the Schmitt triggers is considered a buffer with tapering in the interconnect of VOPD architecture. This architecture is named as Model–II and its configuration is shown in Fig. 2(a). The internal interconnect architecture with the optimization component is shown in Fig. 2(b). This Schmitt trigger switches to an input signal faster than the CMOS buffer repeater presented in [13]. The Schmitt trigger is designed with NMOS and PMOS transistors, and PMOS width kept thrice of that of NMOS. Here, the tapering is also performed in the interconnect.

The modified Schmitt trigger buffer is used as a repeater core to split the interconnect into equal parts. The same size is applied as repeaters to function in each section of the design, as shown in Fig. 2. The message propagates accurately from master to slave at low latency and high throughput. The power consumption is reduced by the fast switching. Therefore, the width of interconnect is considered non-uniform. The source end of the interconnect is considered wider which reduces the crucial interconnect resistance. Since this portion of the interconnects charges a maximum of the interconnect capacitance. At the receiver end, the interconnection is narrowed to reduce the total capacitance, but resistance increases due to narrow width which carries a small fraction of the charge. Therefore, the tapering technique supports reducing the interconnect latency. However, the modified buffer using the schmitt trigger increases extra the surface area.

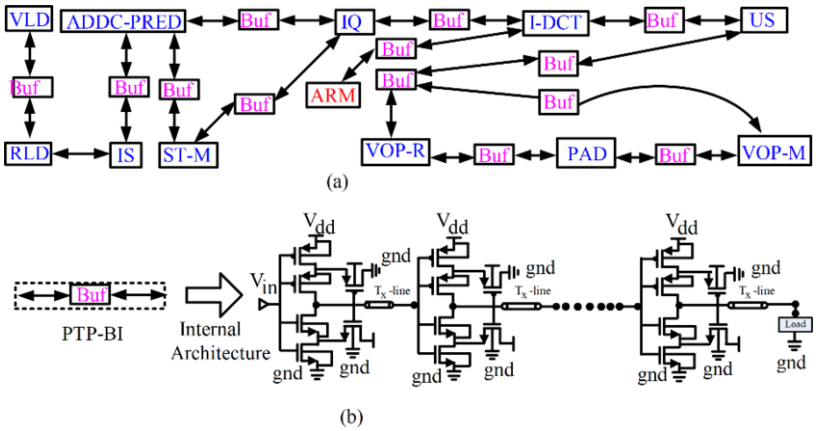

Figure 2. Model-I: The brief topology of the interconnect (a) Model-II: The PTP-BI of VOPD with optimization using Schmitt trigger as buffer with tapering; (b) Internal interconnect architecture.

4. Result and Discussion

In this section, the simulation results of the proposed interconnection techniques are presented for the VOPD multimedia benchmark. With the help of the PTP-BI technique, two different architectures of MSoCs are designed. The first technique is Model–I (Considered as without optimization) and the second technique is Model–II (Considered as with optimization). The optimization technique is known as hybrid optimization of schmitt trigger as repeater and tapering. The specification of interconnect for several variables such as interconnect length (l), thickness (t), height (h), width (w), spacing (S), and constant factor (k) are analyzed for both models as listed in Table I. To evaluate the PTP-BI technique at 0.35 μm for l = 1.5 mm, 4.5 mm, and 8.5 mm, respectively.

Table 1. Specification of interconnects.

		Model–I at 0.35 μm			Model–II at 0.18 μm		
S. No.	**Variables**	l = 1.5 mm	l = 4.5 mm	l = 8.5 mm	l = 1.5 mm	l = 4.5 mm	l = 8.5 mm
1.	t (μm)	0.35	0.55	1.20	0.25	0.45	1.10
2.	h (μm)	0.55	0.55	0.55	0.35	0.35	0.35
3.	K	2.50	2.50	2.50	2.20	2.20	2.20
4.	w (μm)	0.18	0.60	0.60	0.10	0.20	0.50
5.	S (μm)	0.18	0.60	0.60	0.10	0.20	0.50

The latency is examined for both the models (Method–I & Method–II) in Table II. From the Table, it is observed that the latency of Model–II is decreased as compared to Model–I at l = 1.5 mm. For further study, the technology scaled from 0.35 μm to 0.18 μm and considered l = 1.5 mm, 4.5 mm, and 8.5 mm, respectively for both the models presented in Table II. When the interconnection lengths are increase, then the latency increases. Here, the percentage reductions are evaluated in the latency for

Model–II and compared with Model–I. At 0.35 μm technology, the decrease in latency by 19.76% for l = 1.5 mm, 18.79% for l = 4.5 mm, and 42.24% for l = 8.5 mm, respectively. Similarly, at 0.18 μm technology, the reduction in latency by 8.66% for l = 1.5 mm, 43.12% for l = 4.5 mm and 13.37% for l = 8.5 mm, respectively.

Table 2. Calculation of Latency (ns) of Packet in PTP-BI Technique at Various Interconnect Length.

		Models at 0.35 μm			Models at 0.18 μm		
S. No.	***l* (mm)**	**Model–I**	**Model–II**	**Reduction (%)**	**Model–I**	**Model–II**	**Reduction (%)**
1.	1.5	33.9	27.2	19.76	53.10	48.50	8.66
2.	4.5	97.9	79.5	18.79	389.30	221.40	43.12
3.	8.5	363.8	210.1	42.24	2201.20	1906.80	13.37

Further, the percentage reductions are evaluated in Model–I and Model– II power consumption. For 0.35 μm technology, the power consumption was examined for both models at different interconnect lengths as listed in Table III. Here it has been analyzed that for the same l = 4.5 mm, the power consumption of Model–II decreases compared to Model–I. For further study, the technology is scaled from 0.35 μm to 0.18 μm and considered l = 1.5 mm, 4.5 mm, and 8.5 mm, respectively, for both the models presented in Table III. The technology scaled from 0.35 μm to 0.18 μm for the same values of l and examined the power consumption increases in the PTP-BI technique. For example, at 0.35 μm technology, the power consumption at l = 1.5 mm is 132.9 mW for Model–I. At 0.18 μm technology, the power consumption at l = 1.5 mm is 262.7 mW for Model–I. The percentage power consumption reduction of Model–II is compared with Model–I presented in Table III. At 0.35 μm technology, the decrease in power consumption by 2.40% for l = 1.5 mm, 32.44% for l = 1.5 mm, and 20.13% for l = 8.5 mm, respectively. Similarly, at 0.18 μm technology, the reduction in latency by 42.78% for l = 1.5 mm, 24.93% for l = 4.5 mm and 11.44% for l = 8.5 mm, respectively.

Table 3. Calculation of Power Consumption (mW) of the PTP-BI at Various Interconnect Length.

		Models at 0.35 μm			Models at 0.18 μm		
S. No.	***l* (mm)**	**Model–I**	**Model–II**	**Reduction (%)**	**Model–I**	**Model–II**	**Reduction (%)**
1.	1.5	132.90	129.70	2.40	262.7	150.3	42.78
2.	4.5	328.20	221.70	32.44	510.9	383.5	24.93
3.	8.5	798.30	637.60	20.13	1025.8	908.4	11.44

5. Conclusion

In this paper, effective on-chip interconnection techniques are presented which have been integrated with the design of MSoCs. From the results, it is observed that the rise in the latency of the data transactions and the power consumption for interconnects as technology node evolves from 0.35 µm to 0.18 µm. With the increase of interconnect length, the latency and power consumption are reducing with an optimization technique. Still, these latency and power consumption of data transactions in the interconnect values are large. It necessitates a modification in the design of interconnect architecture.

The result shows that the more promising, highly secure and high performance. But it undergoes a scalability problem due to immense complexity, expense and configuration effort. This encourages work towards scalability improvement of architecture in the future. It can be continued towards designing more promising interconnect architecture to satisfy customer requirements such as high performance and high QoS.

References

[1] M. K.Meher.P. Modifications in CMOS dynamic logic style: A review paper, J. Inst. Eng. India Ser. B, vol. 96, pp. 391–399, 2015.

[2] M. R. S. P. e. a. Biswas, B.,. An efficient VLSI architecture of the enhanced three step search algorithm, J. Inst. Eng. India Ser. B, vol. 97, pp. 303–309, 2016.

[3] Sanchez Correa.R and David.J.P .Ultra-low latency communication channels for FPGA-based HPC cluster,Integration. vol. 63, pp. 41–55, 2018.

[4] Babu.P.Reconfigurable FPGA architectures: A survey and applications. J. Inst. Eng. India Ser. B, vol. 102, p. 143U156, 2021.

[5] N. S. Parihar. A.,. High-speed high-throughput VLSI architecture for RSA montgomery modular multiplication with efficient format conversion, J. Inst. Eng. India Ser. B, vol. 100, pp. 217–222, 2019.

[6] Nurmi.J, Tenhunen.H , Isoaho.J, and Jantsch.A .Interconnect-Centric Design for Advanced SoC and NoC, 1st ed. Springer Publishing Company, Incorporated, 2010.

[7] Ebrahimi Jarihani.A, Sarafi.S, M. Köberle, J. Sturm, and A. M. Tonello, Characterization of on-chip interconnects: Case study in 28 nm CMOS technology,in Austrochip Workshop on Microelectronics (Austrochip), Oct 2019, pp. 93–99.

[8] Bhattacharjee.P and Sadhu.A, VLSI transistor and interconnect scaling overview, J. Electron. Des. Technol., vol. 5, no. 1, pp. 1–15, 01 2014.

[9] W. C. T. Sao Jie Chen, Ying Cherng Lan and Y.-H. Hu, Reconfigurable Networks on Chip. Springer, New York, NY, 2012.

[10] Jayshree, S. Verma, and A. Chatterjee, A methodology for designing LVDS interface system, in Sixth Int. Symp. Embed. Comput. Syst. Des. (ISED),, Dec 2016, pp. 284–288.

[11] I. Mackintosh, Open core protocol international partnership (OCP-IP), 2005.

[12] Rowen.C and Leibson.S, Engineering the Complex SoC: Facts, Flexible Design with Configurable Processors. Prentice Hall, 2004.

[13] Jayshree and G. Seetharaman, Design and analysis of novel interconnects with network-on-chip LVDS transmitter for low delay, in NASA/ESA Conf. Adapt. Hardw. Syst. (AHS), Aug 2018, pp. 204–209.

[14] Vainbrand.D and Ginosar.R .Network-on-Chip Architectures for Neural Networks,2010 Fourth *ACM/IEEE* Int. Symp.on Networks-on-Chip, Grenoble, France, 2010, pp. 135-144, doi: 10.1109/NOCS.2010.23.