Explore the Feedback Interconnects in Intra-Cluster Routing for FPGAs

Kaichuang Shi, Hao Zhou, Lingli Wang*
State Key Laboratory of ASIC and System
Fudan University, Shanghai, China
*llwang@fudan.edu.cn

***Abstract*—Routing architecture has a significant effect on the performance and area of FPGA. In academia, the global routing architecture is mainly based on the connection blocks (CBs) and switch blocks (SBs). There are input crossbars inside the logic blocks (LBs) which are used to connect the LB pins and feedbacks to the LUT inputs. Muxes in the crossbar are used to implement the intra-cluster connections. In the previous researches, there are few papers focusing on the exploration of feedback interconnects. Besides, it is hard to model the complex feedback interconnects in commercial FPGAs as the feedbacks can only connect to the muxes in the crossbar in the CB-SB architecture. In this paper, we enhance VPR to support GRB (general routing block) architecture which replaces the CBs, SBs and input crossbars to alleviate this problem and explore complex feedback interconnects. Four parameters are proposed to describe the feedback interconnects architecture. Compared to the CB-SB like feedback interconnects architecture, experimental results show that the proposed architecture can achieve 4.5% improvement on the routing area, 2.1% improvement on the critical path delay and 6.5% improvement on the area-delay product with VTR benchmarks.**

***Index Terms*—FPGA, routing architecture, feedbacks, VTR enhancement**

I. Introduction

FPGAs are widely used due to the fact that they have superiority in time to market, non-recurring engineering (NRE) cost and flexibility [1]. However, compared to ASIC, it needs more area, delay and power when the circuits are implemented in the FPGA. Researches show that the routing architecture has a large impact on the FPGA area and delay [1]. The most common global routing architecture in academia is mainly based on the CBs and SBs which is also used in VTR 8 [2]. CBs are used to connect wire segments with logic block (LB) pins while SBs provide programmable switches to connect with different wire segments. Inside the LB, there is a local crossbar to distribute the LB inputs and the local feedback signals to the LUTs of the LB. The one-level feedback network leads to big mux size and large area in the local crossbar [3].

In this paper, we explore the feedback interconnects in FPGAs. The feedbacks can not only connect to the second muxes (crossbar), but also connect to the first level muxes (CB) like routing wires. As CB-SB architecture is too restricted to model the feedback interconnects described above, we use GRB routing architecture which is proposed in [4] to carry out the architecture exploration. Our contributions include:

- We define four parameters to explore the feedback interconnects. In order to evaluate the performance of the architecture, we enhance the Routing Resource Graph (RRG) generator in the latest VTR 8 [2] to support the GRB routing architecture.
- We evaluate the performance of the feedback interconnects architecture whose area and delay parameters are extracted from COFFE 2 [5] with VTR benchmarks. Experimental results show that the proposed architecture can achieve 4.5% improvement on the routing area, 2.1% improvement on the delay and 6.5% on the area-delay product.

The rest of this paper is organized as follows. Section II introduces the academic CB-SB routing architecture and the related work. Section III presents the GRB architecture and introduces four parameters to describe the feedback interconnects. Section IV gives the enhancements in VTR 8 and COFFE 2 to support the proposed architecture. Section V presents the baseline architecture and experimental results. Section VI concludes this paper with future work.

II. Background and related work

A. Routing Architecture of Island-Style FPGAs

Island-style FPGA architecture mainly contains LBs, CBs and SBs which are interconnected by vertical and horizontal routing wires. Each LB contains several BLEs (Basic Logic Elements) which consist of LUTs and FFs. The routing wires can connect to the LUT inputs through a two-level mux topology (CB and crossbar) as shown in Fig. 1. The mux sizes in CB and crossbar are determined by Fc_{in} and Fc_{local} respectively. Fc_{in} means the fraction of wire segments in the routing channels that an LB input pin can connect to and Fc_{local} is the population density of the crossbar. As we can see, the feedbacks can only connect to the second level muxes in crossbar in the CB-SB architecture.

B. Related Work

In the previous work, the feedback connections are included in the local crossbar. The LB feedbacks can connect to the LUT inputs through one-level muxes in the local crossbar. In [6], G. Lemieux et al. designed the sparse crossbars inside the LB, and experimental results show that it can reduce the switch densities by 50% or more with no degradation to critical path delay. G. Zgheib et al. evaluated the effect of the

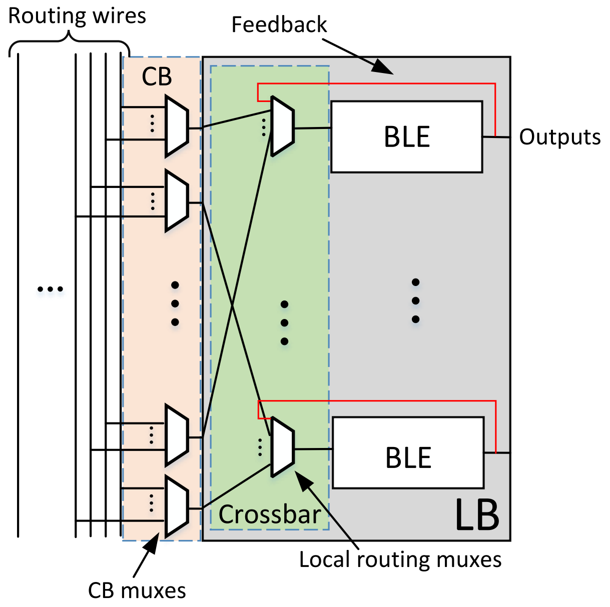


Fig. 1. The two-level mux architecture in CB-SB model.

local interconnect density in LBs on the FPGA performance and area [7]. In some researches, the feedback interconnects are not included in the local routing muxes. Instead, they are connected to the first level muxes like routing wires. W. Feng et al. proposed input interconnect blocks (IIBs) which route signals from wires and LB feedbacks to LUT inputs [8]. Experimental results show that IIB with two-level muxes can achieve great area savings with no routability decreasing. Qian et al. proposed tile-based GRB architecture to model complex commercial FPGAs [4]. In the GRB architecture, the feedbacks can connect to the LB pins through input connection block (ICB) with one or two level muxes. But it does not carry out the architecture exploration about the feedback interconnects in that paper.

In Intel's early FPGAs [9]–[12], they use a VPR-style architecture. The routing tracks can connect to the LUT pins through two-level muxes (LIMs and LEIMs). And the feedback connections are only in the LEIMs. However, in the recent Intel's FPGAs [13] [14], the feedbacks are connected to the LIMs. And there is no direct feedback (LB outputs to LB inputs through LEIMs). In Microsemi flash-based FPGAs [3], it uses a three-level cluster input routing. Feedbacks from LUTs can be feed in muxes in every level which helps route the most critical nets to the fastest LUT inputs. In Xilinx 7-series FPGA, feedbacks can not only connect to the LUT inputs directly, but also can connect to the ALT and BOUNCE highly-interconnected muxes and then pass the signals to the LUT pins [15].

This paper is largely inspired by [3] [13]. We design the feedback interconnects in FPGA routing architecture to trade off the area and delay.

III. Feedback interconnects

In this section, the GRB routing architecture [4] is introduced. Then, we design four parameters to describe the feedback interconnects based on the GRB architecture.

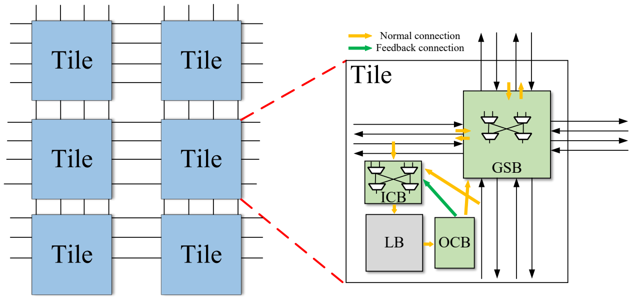


Fig. 2. The GRB architecture.

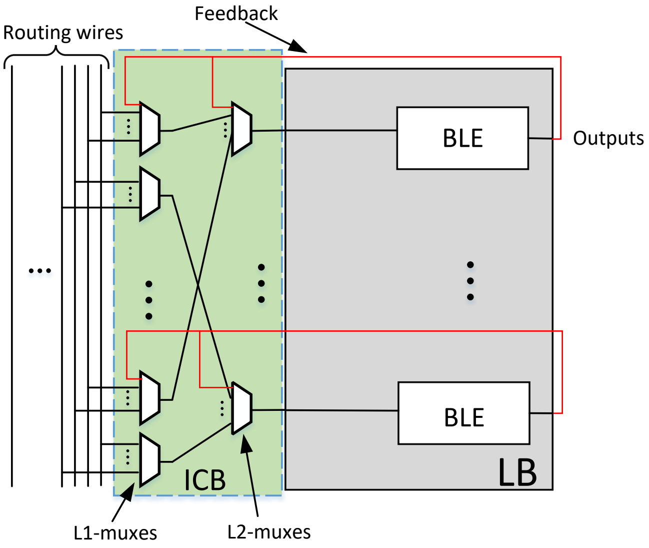


Fig. 3. The two-level mux architecture in ICB.

A. GRB Architecture

Fig. 2 shows the GRB routing architecture [4] which is tileable. Inside each tile, there is an LB and a GRB which contains three modules, ICB, OCB and GSB. ICB is used for the connections of the global wire segments and LB output pins to the LB input pins. OCB outputs can be used as feedbacks or feed in GSB. GSB contains the connections between global wire segments and the connections of LB output pins to the global wire segments. Inside the ICB and GSB, two-level mux topology with output sharing is used to trade off the area, delay and flexibility. Fig. 3 shows the two-level mux architecture in ICB. The feedbacks can be connected to the first (L1-muxes) or second level muxes (L2-muxes).

B. Modeling Parameters

To model the feedback interconnects, four parameters are defined. L_m is defined to represent the level of muxes that the feedbacks can connect to. The optional values contain 1, 1.5 and 2. When L_m is set to 1.5, it means the feedbacks can connect to both the L1-muxes and L2-muxes. Then, we define F_1 to stand for the flexibility of feedback interconnects in L1-muxes. It presents how many feedback interconnects in one L1-mux. Similarly, we define F_2 to stand for the flexibility of feedback interconnects in one L2-mux. It presents how many feedback interconnects in one L2-mux. For example, both the value of F_1 and F_2 are 1 in Fig. 3, because each L1-mux and L2-mux can get one feedback interconnect. And the value of L_m is 1.5 as the feedbacks can connect to both the L1-muxes and L2-muxes. Besides, we define F_p to represent the population density of the L2-muxes that can receive connections from L1-muxes. In addition to the parameters we added, Fc_{in}, Fc_{out} and *Fs* are also used to describe the global routing architecture as in CB-SB architecture. Fc_{in} and Fc_{out} mean the number of wires that an LB input and output pin can connect to respectively. *Fs* represents the number of wires that an incoming wire can connect to. As this paper focuses on the exploration of feedback interconnects, the value of Fc_{in}, Fc_{out} and *Fs* are constant in the following experiments.

IV. Tool Enhancement

In this section, the tool enhancement is introduced to implement the feedback interconnects modeling and exploring.

A. Enhanced VPR

To support the GRB architecture, we enhance the RRG generator in VTR to model the FPGA routing resources. The routing resources are presented by a directed graph *G = (V, E)*. *V* corresponds to the routing nodes which can be wires or LB pins, and *E* to the programmable switches. For two level muxes, we use the similar modeling method in [4]. An intermediate node without timing cost is used to model the connection between the first level and the second level muxes.

B. Area and Delay Modeling

We use VTR to estimate the whole FPGA area and the critical path delay which measures the area in *minimum-width transistor areas* (MWTAs) [16] and uses Elmore delay model to estimate the delay. We enhance the COFFE 2 [5] to extract the area and delay parameters needed by the VTR architecture description file. COFFE 2 is a fully automated transistor sizing tool for FPGAs which measures the area and delay by relying on HSPICE simulation. The FPGA circuitry can be constructed by the parameters in Section III-B. Then, the area and delay parameters needed by VTR can be obtained by HSPICE simulation.

V. Experimental Results

In this section, we describe the experimental methodology and the baseline architecture. Then, we use the enhanced VTR along with the provided benchmark set to evaluate the architecture in the area and delay.

A. Experimental methodology

TABLE I shows the baseline architecture parameters. The delay and area parameters which are used in VTR architecture files are extracted from COFFE 2 at the 22nm technology node. The baseline architecture is a CB-SB like architecture. The feedbacks can only connect to the L2-muxes and the population of L2-muxes is set to 0.5 which is common in the previous work [4] [17].

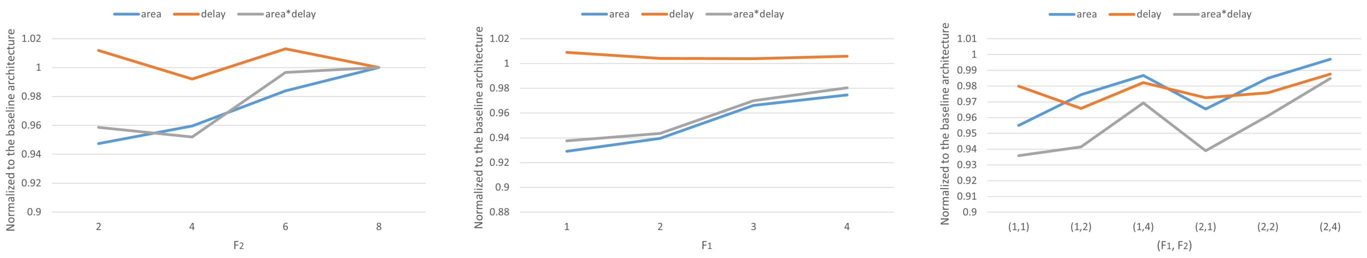
TABLE I
Baseline Architecture Parameters

LB Size	Eight 6-input LUTs
LB input pins & output pins	48 & 16
SB Pattern	Wilton
Wire Length	4
$\boldsymbol{Fc_{in}}$**,** $\boldsymbol{Fc_{out}}$ **& *Fs***	0.1, 0.1 & 3
$\boldsymbol{L_m}$**,** $\boldsymbol{F_1}$**,** $\boldsymbol{F_2}$ **&** $\boldsymbol{F_p}$	2, 0, 8 & 0.5

B. Architecture with Different Parameters

In this section, we explore the feedback interconnects with different parameters. As we focus on the exploration of feedback interconnects, the value of Fc_{in}, Fc_{out} and *Fs* are constant in TABLE I. The experimental results are normalized to the baseline architecture which can be found in Fig. 4. Firstly, we sweep the value of F_2 when the value of L_m is set to 2 which means the feedbacks can only connect to the L2-muxes. Results show that when the value of F_2 is set to 4, it can achieve the best delay (1.0%) and area-delay product (4.9%) savings. The result is similar to the conclusion in [7] which indicates that the best delay is observed at around 20% to 30% crossbar density. Although too small F_2 will bring more area savings, it also lead to routing congestion and increases the delay. Then, the value of F_1 is swept when the value of L_m is set to 1. The architecture with feedback interconnects which can only connect to the L1-muxes can achieve better routability compared to those which can only connect to the L2-muxes under the same number of feedback interconnects. So, the value of F_1 in Fig. 4 (b) is smaller than F_2 in Fig. 4 (a). Experimental results show that it achieves great area savings but worse delay. The results are predictable because the feedbacks need to go through two level muxes to connect to the LUT pins.

Then, we explore the different combinations of F_1 and F_2 when the value of L_m is set to 1.5 which means the feedbacks can only connect to both the L1-muxes and the L2-muxes. Experimental results show that it can achieve the best delay improvement by 3.5% when F_1 = 1 and F_2 = 2, and it can achieve the best area-delay product savings by 6.5% when F_1 = 1 and F_2 = 1. It can be seen that the feedback interconnects which can be feed in every level muxes can tradeoff area and delay better, as the most critical nets can achieved through the fastest one level muxes, and other nets can be achieved through the slower two level muxes. And the feedbacks to the first level muxes can achieve better area savings. Besides, we



(a) Exploration of F_2 under L_m = 2.

(b) Exploration of F_1 under L_m = 1.

(c) Exploration of F_1 and F_2 under L_m = 1.5.

Fig. 4. Exploration of feedback interconnects under different parameters.

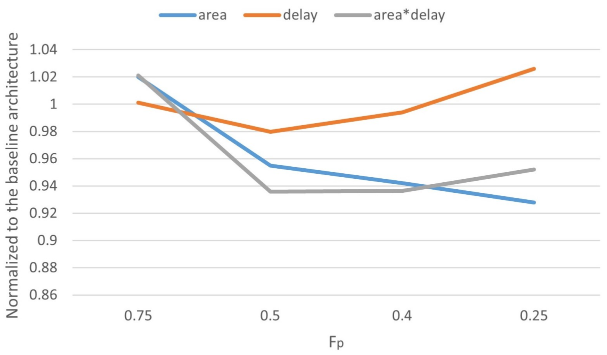


Fig. 5. Exploration of F_p under F_1 = 1 and F_2 = 1.

explore the value of F_p under $F_1 = 1$ and $F_2 = 1$. The value of F_p will also influence the global interconnects architecture. Experimental results show that it can achieve the best area delay tradeoff when $F_p = 0.5$ as shown in Fig. 5.

VI. Discussion and Conclusion

In this paper, we explore the feedback interconnects in intra-cluster routing for FPGAs. Four parameters are defined to describe the feedback interconnects. Besides, the RRG generator in VTR are enhanced to support the GRB architecture. Experimental results show that the proposed architecture can achieve 4.5% improvement on the routing area, 2.1% on the critical path delay and 6.5% on the area-delay product compared to the CB-SB like feedback interconnects architecture. In the future, we will explore the different combinations of Fc_{in} and Fc_{out} which are constant in this paper. Besides, we will apply the similar method to the inter-cluster routing to explore the interconnects of LB output pins to wire segments. The distribution of wire segments will also be a research direction in the future because commercial FPGAs usually contains a variety of wire segments.

Acknowledgment

This work is supported by the National Natural Science Foundation of China under grant 61971143.

References

[1] V. Betz, J. Rose, and A. Marquardt, *Architecture and CAD for deep-submicron FPGAs*. Kluwer Academic Publishers, 1999.

[2] K. E. Murray *et al.*, "VTR 8: High Performance CAD and Customizable FPGA Architecture Modelling," *ACM Transactions on Reconfigurable Technology and Systems (TRETS)*, vol. 13, no. 2, pp. 1–55, 2020.

[3] J. Greene, S. Kaptanoglu, W. Feng, V. Hecht, J. Landry, *et al.*, "A 65nm flash-based FPGA fabric optimized for low cost and power," in *Proceedings of the 19th ACM/SIGDA international symposium on Field programmable gate arrays*, pp. 87–96, 2011.

[4] J. Qian, Y. Shen, K. Shi, H. Zhou, and L. Wang, "General routing architecture modelling and exploration for modern FPGAs," in *2021 International Conference on Field-Programmable Technology (ICFPT)*, pp. 1–9, IEEE, 2021.

[5] S. Yazdanshenas and V. Betz, "COFFE 2: Automatic modelling and optimization of complex and heterogeneous FPGA architectures," *ACM Transactions on Reconfigurable Technology and Systems (TRETS)*, vol. 12, no. 1, pp. 1–27, 2019.

[6] G. Lemieux and D. Lewis, "Using sparse crossbars within LUT," in *Proceedings of the 2001 ACM/SIGDA ninth international symposium on Field programmable gate arrays*, pp. 59–68, 2001.

[7] G. Zgheib and P. Ienne, "Evaluating FPGA clusters under wide ranges of design parameters," in *2017 27th International Conference on Field Programmable Logic and Applications (FPL)*, pp. 1–8, IEEE, 2017.

[8] W. Feng and S. Kaptanoglu, "Designing efficient input interconnect blocks for LUT clusters using counting and entropy," *ACM Transactions on Reconfigurable Technology and Systems (TRETS)*, vol. 1, no. 1, pp. 1–28, 2008.

[9] D. Lewis *et al.*, "The stratix™ routing and logic architecture," in *Proceedings of the 2003 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, pp. 12–20, 2003.

[10] D. Lewis *et al.*, "The Stratix II logic and routing architecture," in *Proceedings of the 2005 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, pp. 14–20, 2005.

[11] D. Lewis, E. Ahmed, D. Cashman, T. Vanderhoek, C. Lane, *et al.*, "Architectural enhancements in Stratix-III™ and Stratix-IV™," in *Proceedings of the ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, pp. 33–42, 2009.

[12] D. Lewis, D. Cashman, M. Chan, J. Chromczak, G. Lai, *et al.*, "Architectural enhancements in Stratix V™," in *Proceedings of the ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, pp. 147–156, 2013.

[13] D. Lewis, G. Chiu, J. Chromczak, D. Galloway, B. Gamsa, *et al.*, "The Stratix™ 10 highly pipelined FPGA architecture," in *Proceedings of the 2016 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, pp. 159–168, 2016.

[14] J. Chromczak, M. Wheeler, C. Chiasson, D. How, M. Langhammer, *et al.*, "Architectural Enhancements in Intel® Agilex™ FPGAs," in *Proceedings of the 2020 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, pp. 140–149, 2020.

[15] M. B. Petersen, S. Nikolić, and M. Stojilović, "NetCracker: A peek into the routing architecture of Xilinx 7-Series FPGAs," in *The 2021 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, pp. 11–22, 2021.

[16] C. Chiasson, *Optimization and modeling of FPGA circuitry in advanced process technology*. PhD thesis, University of Toronto, 2013.

[17] K. Shi, X. Zhou, H. Zhou, and L. Wang, "An Optimized GIB Routing Architecture with Bent Wires for FPGA," *ACM Transactions on Reconfigurable Technology and Systems*, vol. 16, no. 1, pp. 1–28, 2022.