Research Survey Paper On FPGA Routing Architecture

Wenxuan Zhang
Electrical and Computer Engineering
University of Toronto
Toronto, Canada
wenxuanzhang.zhang@mail.utoronto.ca

***Abstract*—This paper is a briefly research survey and analysis of several papers in the field of FPGA Routing Architecture.**
***Index Terms*—FPGA Routing Architecture**

I. Introduction

This paper is a survey and analysis of the field of FPGA routing architectures. A total of 10 papers in this area are selected and three of them are analyzed and compared in detail.

II. Briefly Survey and Summary

A.

[1] This paper examines the impact of tuning the parameters of the routing architecture on area as well as latency. The experiments in this paper are measured by adjusting a specific parameter while keeping other parameters constant. The authors conclude that there is a trade-off between critical path delay and routing area to achieve the optimal routing structure, and that among the parameters affecting these trade-offs, Fcin has no significant effect on the trade-off, increasing Fcout reduces delay but increases area, and increasing wire length helps reduce delay.

B.

[2] This paper examines the optimal use scenarios for unidirectional and bidirectional routing architectures in terms of energy and power consumption. The experiments in this paper are implemented in VPR5.0 and the results are measured by changing only the routing buffer size without changing the lut input size, while the experiments mainly measure the three directions of the two architectures, namely, critical path delay, energy consumption versus maximum operating frequency, and which architecture consumes the least energy in a given scenario. The authors conclude that single routing performs better in most scenarios. The critical path latency of single routing increases with buffer size and single routing consumes less area at 33MHz with less energy consumption, making single routing more suitable for the traditional FPGA domain. While bidirectional routing consumes less power in the low kHz to 10 MHz frequency range, which makes bidirectional routing potentially suitable for heartbeat clocks in the mobile domain.

C.

[3] This paper examines modern improvements to the design of routing architectures based on wotan, a tool for evaluating architectures based on network reliability using probabilistic analysis, with the advantage of flexibility and independence from FPGAs. The work of this paper is to compensate for the lack of support of wotan for complex architectures and for the loss of accuracy of the evaluation in some cases. The authors perform several experiments in the paper, firstly updating the route map reader and modifying the path count in order to support macroblocks, and secondly introducing new path weights in order to take into account the flexibility of the connection blocks. Finally the authors conclude that the modifications to wotan improve the routing accuracy.

D.

[4] This paper investigates the use of stochastic methods to quickly approach optimal routing architecture solutions without traversing all potential design solutions. The authors have designed a tool, TORCH, to achieve this purpose. In this paper, the optimization of the routing architecture is focused on two aspects, namely routing channel segmentation and switchbox pattern design, and the authors use five methods for routing channel segmentation and three changes to the previous optimization scheme for switchbox pattern design. The final experimental results show that stochastic methods are very effective in routing architecture design and can effectively optimize the latency and power consumption.

E.

[5] This paper investigates the difference in performance between routing architectures using a generic switch box and a traditional routing architecture with a CB/SB, and the authors have used a VPR to accomplish this experiment. The authors concluded that the reduction of switches can increase the speed of the circuit and the GSB architecture designed in the paper allows the interconnection of different LB pins and its pins can be connected to LBs in all four directions, thus substantially reducing the number of routing switches. Finally, the authors conclude that the routing architecture using GSB can reduce the number of switches while making the product of channel

width and critical path delay better than the conventional architecture.

F.

[6] This paper investigates the performance impact of using segment distribution and mixed segment distribution for line length in a unidirectional island routing architecture, and the authors have done the experiments using the VPR5.0.2 toolset. In this work the authors first tested the performance impact of different wire lengths on the performance of a single segment length, and then measured the performance impact of mixed lengths for combinations between different wires lengths. Finally, the paper concludes from the experiments that the performance of the routing architecture with a mixed distribution is only slightly improved compared to a single segment distribution, but the stable performance between different combinations can provide flexibility in selection.

G.

[7] In this paper, a new FPGA routing architecture with the application of time-multiplexing techniques is investigated. The only difference between the new architecture designed by the authors and the traditional one is that all the wires are multiplexable. To achieve this the authors first modify the user clock period so that when the wires are multiplexable, multiple signals can use the same connection line. Secondly, the authors also modified the wires so that when the wire segments are multiplexable, they can be multiplexed by multiple nets with different time intervals. Finally, the authors have designed a switch whose state changes with the state of the circuit and can lock the current value. In the article, the authors use the new architecture to compare with the traditional architecture and conclude that the new architecture can effectively reduce the minimum channel width and critical path delay.

H.

[8] This paper investigates the scalability of a new routing architecture, HCGP, for large FPGAs. The authors' idea of scaling the HCGP routing architecture is to increase the capacity of logic and pins in each FPGA while keeping the total number of FPGAs relatively small, which is intended to avoid the expensive costs incurred in large FPGA designs. The experiments in this paper use the statistical model of a real million gate design netlist and the routability of HCGP is evaluated by various parameters. Finally, the paper concludes that the HCGP architecture is well suited for scaling with large FPGAs.

I.

[9] This paper investigates a uniform routing architecture that makes it easier to insert more IP cores. To achieve this the authors modify the top-level routing architecture by dividing all routing resources into global and local ones and making them perform different tasks. The implementation of this routing circuit has the following three features. Firstly, the authors designed a special mux so that the signal entering the mux and then outputting it only needs to go through two transistor stages. Secondly all segmented lines in the architecture are single. Finally, all long lines in this architecture span the entire chip and a buffer is inserted before every 10th multiplexer. Finally, the authors conclude that this architecture reduces the long line latency while making it easy to insert IP cores.

J.

[10] This paper investigates and designs a new routing architecture, CRA, with the aim of improving routing resource utilization and increasing performance. This routing architecture is composed of an array of routing modules, while each module has two bypass interconnections and a staff switching core. There are two main differences of CRA compared to the traditional island architecture. The first is that its switch core can function as both a switch box and a connection box and there is no separation between the connection box and the switch box, which is intended to enable flexible resource sharing. Secondly, the CRA's switching core has a dynamic switching matrix, which allows it to dynamically expand the possible switching modes, which allows the architecture to reduce the minimum routing channel width. Finally the authors compare the performance of the CRA with Virtex-II and conclude that the CRA can route routing resources efficiently while reducing latency significantly.

K. Summary

To sum up, in the field of FPGA routing architecture, most researchers currently study how to adjust the routing architecture, such as adjusting parameters and designing new architectures, etc., to achieve the purpose of further optimizing the routing architecture.

III. Comparison and Analysis

In this section I will select [5], [10], and [7] for a more detailed analysis and comparison of these three papers. In all three papers the authors try to devise a new routing architecture to improve performance or solve the target problem. I will first describe and analyze these three architectures in detail, and then compare them.

A. Detail Analysis

The first is the GSB architecture in paper [5], the most important feature of this architecture is that the authors designed a special switch universal switch box for it to replace the traditional connection block and switch block, this new switch box can directly interconnect its four directions of different LB through its pins. The purpose of this is to design a fast path routing so that LBs can be directly interconnected by a small number of switches. In contrast to the traditional routing architecture using CB/SB, the GSB routing architecture can directly connect the input pins to the output pins to achieve fast paths. And in GSB architecture for a net want to connect two adjacent logical blocks can be directly through a routing switch and does not need to consume segments.

The second is the CRA in paper [10], which differs from the conventional interconnection architecture using LB, CB, and

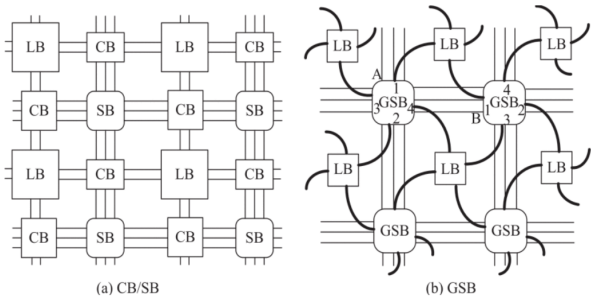

Fig. 1. Comparison between GSB and CB/SB architectures Source:[5]

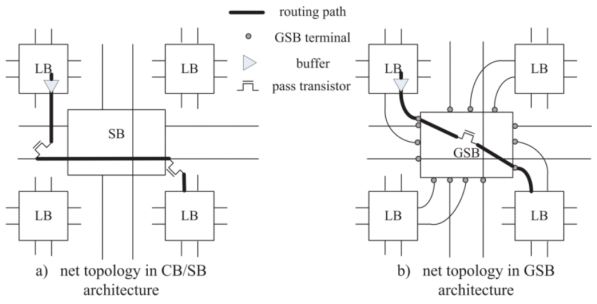

Fig. 2. Pin to pin net topologies in CB/SB and GSB architecture Source:[5]

SB in that the authors integrate these into a routing module and then connect these modules to form an entire routing array. The most special feature of the routing module is the author-designed switch core, which can fuse the connection box and the switch box together and perform both functions simultaneously to achieve resource sharing. Another feature of the switch core is that it can dynamically expand its switching mode according to the requirements, so that the switching mode can be changed even after the FPGA configuration.

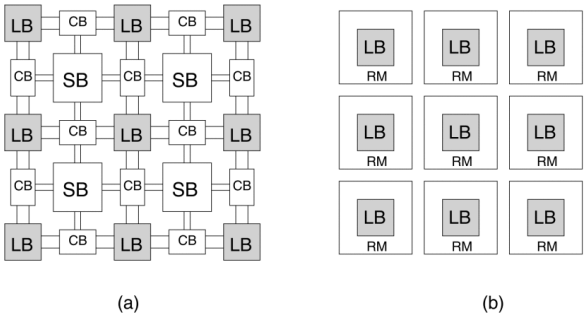

Fig. 3. Comparison between GRA and Island-Style Segmented Routing architectures Source:[10]

Finally, the TM-ARCH designed by the authors in paper [7], the only difference between TM-ARCH and the conventional architecture is that all wires are multiplexable, and its main feature is the application of the time-multiplexing technique. The only difference is that all wires are multiplexable. Secondly, since the wires are idle before the arrival and after the departure of the signals, the authors designed time-

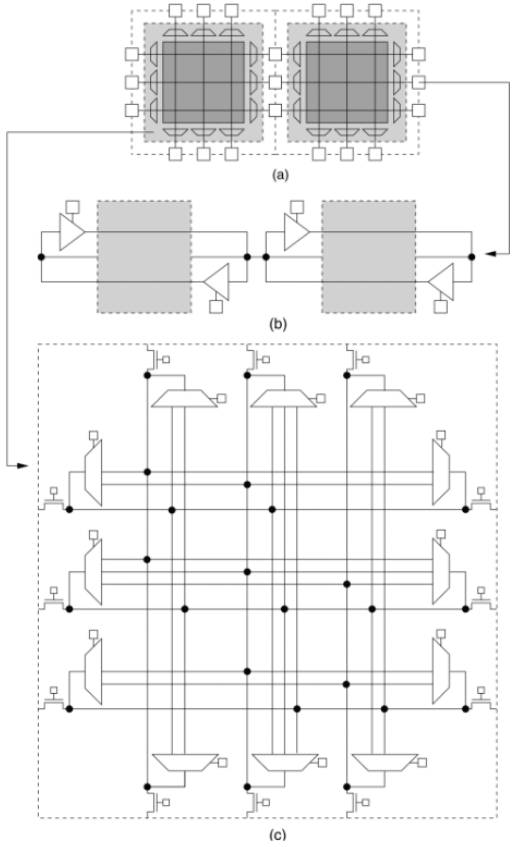

Fig. 4. Routing module of CRA Source:[10]

multiplexing wires for TM-ARCH so that multiple nets with the same clock period but different microcycles can use the same wire segments. Finally, the authors also applied the time-multiplexing technique to the switches and named them TM-switches. TM-switches can store a context at each microcycles so that they can sequentially change their switching state as the microcycles changes, and TM-switches can also be closed to lock their currently stored values when necessary.

B. Comparison

Next, I will divide these three papers into two categories and discuss them. One is the architecture that revolutionizes or redesigns the traditional architecture, which includes papers which described GSB in [5] and CRA in [10], and the other is the architecture that does not change much on the traditional architecture and combines the emerging technologies with the traditional architecture, which includes paper that designed TM-ARCH in [7].

1) Category A: The first categories of paper to be discussed is the first one that design an architecture that makes a major innovation to the traditional architecture. In this category, GSB architecture and CRA have similar design and implementation ideas, although their authors have different design intentions, as soon as possible they still have major differences in the implementation details. In both GSB architecture and CRA, the main features are realized by designing a special switch

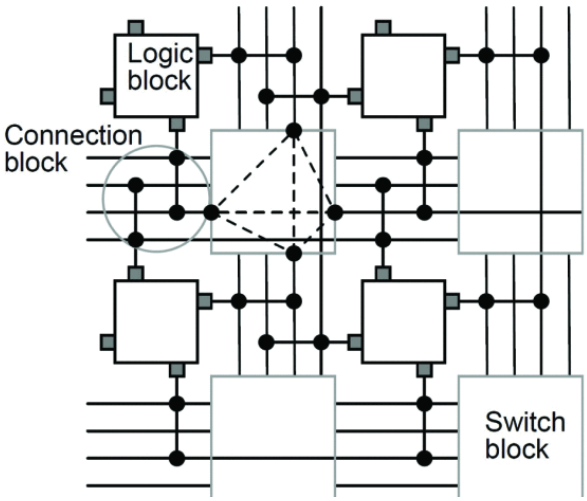

Fig. 5. TM-ARCH architecture with time-multiplexed interconnects Source:[7]

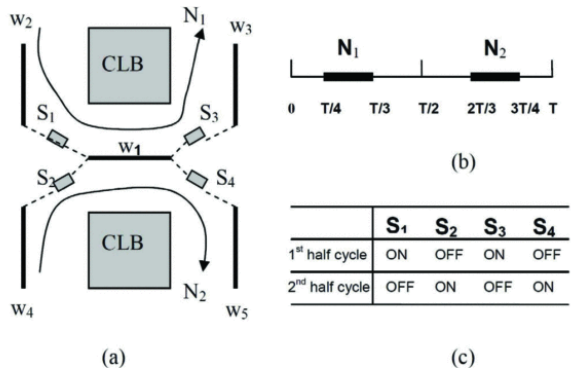

Fig. 6. (a) Signals of N1 and N2 time-multiplex a wire; (b) N1 and N2 occupy different time domains; (c) Different states of the TM switch. Source:[7]

box, which either replaces or incorporates the traditional CB and SB functions and adds new functions. The main difference between these switches is in the newly added functionality. The GSB enables the direct connection of adjacent LB pins through a single routing switch, while the CRA still requires the signal to pass through two switch cores when adjacent LBs want to interconnect. But at the same time the CRA switch cores are embedded in the switch box and connection box and therefore share routing resources such as mux. Finally, the performance improvements are also similar, with both architectures showing performance improvements mainly in the form of reduced critical path latency.

Although the architectures described in both papers appear to be effective in improving performance, both papers have their own shortcomings. First, in paper [5], there is no clear description of how the GSB is composed, and we only know its functionality and effectiveness from the text. Meanwhile, the author's initial goal and conclusion are to some extent contradictory. However, In paper [5] the authors mention that they want to use GSBs to connect the pins of adjacent LBs directly together to increase flexibility and routability and to reduce the number of routing switches. However, according to the conclusions in paper [10], in some cases the number of switches increases even though the delay of critical paths decreases, which is clearly contradictory to the authors' design idea to some extent. Secondly, regarding the shortcomings of paper [10], in paper [10] the authors hope to improve the performance of the architecture by implementing routing resource sharing. Although the article proposes several times that the resource utilization has improved, it does not clearly show how much the resource utilization has improved and the clear relationship between the final performance improvement and the resource utilization improvement, so we cannot fully determine the improvement of the capability of the new CRA architecture comes from the condition of resource sharing. So based on the above shortcomings, I believe that the authors of both articles have not fully addressed the problem they were trying to solve, even though their approach may be valid under certain premises.

2) Category B: After this I will discuss another category of papers that describe architectures that do not make major changes to traditional architectures and combine them with emerging technologies. In this category, TM-ARCH differs dramatically from the architectures described in the previous paper categories. Compared to GSB and CRA, TM-ARCH makes changes to the traditional architecture by simply making the wires all multiplexable. The main work in paper [7] focuses on how to combine time-multiplexing techniques with traditional architectures in various aspects such as clocks and wires and design the corresponding routing algorithms for this combination. Finally the performance improvement of the architecture in [7] is similar to that in [5] and [10], which also reduces the minimum channel width and critical path delay.

Compared with papers [5] and [10] in category A, I think [7] is a more complete solution to the problem and achieves the author's purpose. First, paper [7] clearly describes how to apply the time-multiplexing technique to the routing architecture and shows the pseudocode of the corresponding algorithm. Secondly, the experimental comparison in paper [7] also shows more clearly the performance improvement of TM-ARCH for routing architectures. Therefore, in summary, I believe that the authors have successfully addressed the issue of applying time-multiplexing techniques to routing architectures and have successfully demonstrated the reliability and performance of such architectures.

IV. Future Work Suggestion

Based on the above papers and the papers I have briefly surveyed before, I believe that the future direction of work in the area of routing architectures should be similar to that described in paper [7], considering the combination of some newly discovered technologies with the most widely used architectures today. Because the final results of extensive adjustments and changes to routing architectures, as in papers [5] and [10], can be influenced by many factors, it is not always clear what causes each result and how to reproduce

it. The combination of new techniques and traditional architectures in paper [7] can help us identify the dependent and independent variables more easily, and thus solve the problem more consistently and clearly to optimize the performance of the routing architecture.

REFERENCES

[1] C. Schäfer, M. Stojilović, and L. Saranovac, "Analysis of impact of FPGA routing architecture parameters on area and delay," IEEE Xplore, Nov. 01, 2011.

[2] P. Jamieson, W. Luk, S. J. E. Wilton, and G. A. Constantinides, "An energy and power consumption analysis of FPGA routing architectures," IEEE Xplore, Dec. 01, 2009.

[3] R. Zh. Chochaev and S. V. Gavrilov, "Evaluating FPGA Routing Architectures with Complex Grid Layouts," IEEE Xplore, Jan. 01, 2021.

[4] M. Lin, J. Wawrzynek, and A. E. Gamal, "Exploring FPGA Routing Architecture Stochastically," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 29, no. 10, pp. 1509–1522, Oct. 2010, doi: 10.1109/TCAD.2010.2061530.

[5] K. Ma, L. Wang, X. Zhou, S. X.-D. . Tan, and J. Tong, "General switch box modeling and optimization for FPGA routing architectures," IEEE Xplore, Dec. 01, 2010.

[6] A. Mishra, N. Jayapalan, H. Rastogi, and T. Agrawal, "Impact of segmentation distribution on area and delay in FPGA routing architectures," IEEE Xplore, Feb. 01, 2013.

[7] R. Luo, X. Chen, and Y. Ha, "Optimization of FPGA Routing Networks with Time-Multiplexed Interconnects," IEEE Xplore, Feb. 01, 2020.

[8] M. A. S. Khalid and V. Salitrennik, "Scalability Evaluation of a Hybrid Routing Architecture for Multi-FPGA Systems," IEEE Xplore, Dec. 01, 2006.

[9] L. Wang et al., "Uniform routing architecture for FPGA with embedded IP cores," IEEE Xplore, Oct. 01, 2009.

[10] Y. Ma and M. Lin, "Collaborative Routing Architecture for FPGA," IEEE Xplore, May 01, 2007.