Design and implementation of a delay-optimized universal programmable routing circuit for FPGAs*

Wu Fang(吴方), Zhang Huowen(张火文), Lai Jinmei(来金梅)†, Wang Yuan(王元), Chen Liguang(陈利光), Duan Lei(段磊), and Tong Jiarong (童家榕)

(State Key Laboratory of ASIC & System, Fudan University, Shanghai 201203, China)

Abstract: This paper presents a universal field programmable gate array (FPGA) programmable routing circuit, focusing primarily on a delay optimization. Under the precondition of the routing resource's flexibility and routability, the number of programmable interconnect points (PIP) is reduced, and a multiplexer (MUX) plus a BUFFER structure is adopted as the programmable switch. Also, the method of offset lines and the method of complementary hanged end-lines are applied to the TILE routing circuit and the I/O routing circuit, respectively. All of the above features ensure that the whole FPGA chip is highly repeatable, and the signal delay is uniform and predictable over the total chip. Meanwhile, the BUFFER driver is optimized to decrease the signal delay by up to 5%. The proposed routing circuit is applied to the Fudan programmable device (FDP) FPGA, which has been taped out with an SMIC 0.18-μm logic 1P6M process. The test result shows that the programmable routing resource works correctly, and the signal delay over the chip is highly uniform and predictable.

Key words: FPGA; programmable routing resource; delay; MUX; BUFFER

DOI: 10.1088/1674-4926/30/6/065010 **EEACC:** 1130; 1130B

1. Introduction

Field programmable gate arrays (FPGAs) are widely used in communication, multimedia, industrial control, numerical computation,etc. Among the research and development of FPGAs,the design of the programmable routing resource is the most important,because it costs approximately 60%–70% of the chip area and 50%–60% of the signal delay[1].

Concerning the FPGA programmable routing resource design, there are some research results, most of which are focusing on the segmentation of the interconnect lines, the switch block issue, the cluster size, low power, etc.[2–4]. But still little research has been done on designing a universal FPGA routing circuit, which makes the total FPGA chip to be highly repeatable and allows a prediction of the signal delay over the total chip.

This paper will focus on designing a delay-optimized universal FPGA routing circuit. By using the offset line and return line strategies, every module in the FPGA, including CLBs and IOBs, has a highly repeatable and uniform programmable routing architecture. These strategies also ensure that the load of the same kind of interconnect lines is equally distributed, and, furthermore, that the signal delay is predictable and regular. The realization of the PIP and the BUFFER driver are also optimized to benefit the signal delay.

2. Programmable routing circuit design

2.1. Top level routing resource design

The proposed FPGA programmable routing circuit architecture is shown in Fig. 1: CLBs are connected to the global routing resource with a general routing box (GRB), and an I/O interconnection forms the interface between the internal logic and the external I/O signals of IOBs[5]. The whole routing resource is divided into a global routing resource and a local routing resource.

The local routing resource includes:

(1) The feedback interconnect lines in the CLB, which provide a fast connection of the LUTs in the same CLB.

(2) The direct interconnect lines between vertically and horizontally neighboring CLBs, which eliminate the delay of

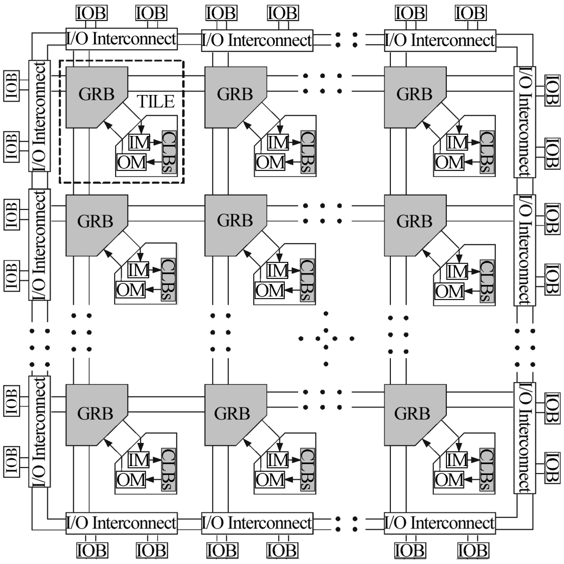

Fig. 1. Top level of the FPGA routing circuit.

* Project supported by the National High Technology Research and Development Program of China (No. 2007AA01Z285) and the National Natural Science Foundation of China (No. 60876015).

† Corresponding author. Email: jmlai@fudan.edu.cn

Received 11 September 2008, revised manuscript received 13 December 2008

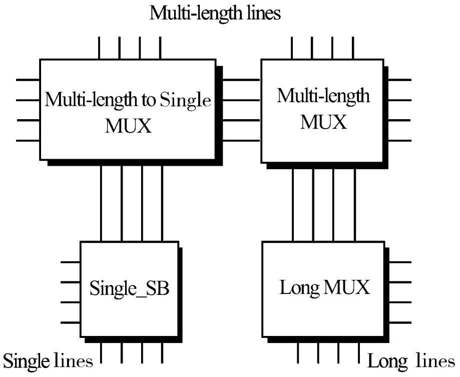

Fig. 2. Architecture of the GRB.

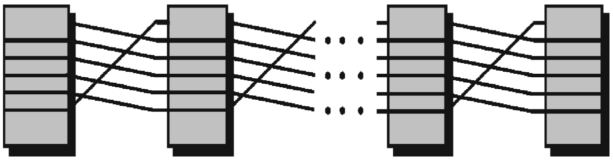

Fig. 3. Routing offset lines.

the GRB.

The global routing resource includes:

(1) Several kinds of interconnect lines in the routing channels.

(2) Repeatable TILE routing (will be described in detail in the following section).

2.2. TILE routing resource design

The TILE is comprised of CLBs and their corresponding routing resource[6], which forms the basic unit of the FPGA chip. The TILE routing resource includes:

(1) Programmable interconnect lines, including long lines, multi-length lines (length-6 lines are used in FDP) and single lines.

(2) The input multiplexer (IM), implementing the input of the signal from the interconnect lines to the CLBs

(3) The output multiplexer (OM), implementing the output of the CLBs to the interconnect lines.

(4) The GRB, implementing the connection between the interconnect lines (Fig. 2).

In this routing circuit, to facilitate the usage of software, it is prescribed that GRB long lines can drive multi-length lines, and multi-length lines can drive single lines, but not vice versa. Also, the output of the OM can drive all the three kinds of interconnect lines, but again not vice versa. Only single lines can directly connect to the input of CLBs. In addition, the multi-length line can only connect to the CLBs, which have their source point, end point, and middle point pass by. This idea is to reduce the number of PIPs and to alleviate the load pressure of the interconnect lines.

The optimization strategy of the offset by sets (Fig. 3) is used to design the top level of the routing resource and vice versa, to divide the multi-length lines and the long lines in each channel (including vertical and horizontal) into several sets, and to shift them by one position at each TILE.

The routing offset strategy guarantees that the interconnect circuit needs only one repeating unit, and the relative position of each kind of routing resource is completely identical. So, the top level distribution of the programmable routing resource is unified. This provides a post-layout design with an effective repeatable unit, which will greatly reduce the workload of the layout. Meanwhile, when the chip scales up, the physical layout can be easily extended by a copy, greatly shortening the research and development cycles of products.

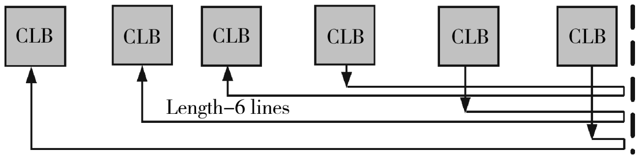

Fig. 4. Complementary hanged end-lines.

2.3. I/O routing resource design

The I/O routing resource forms the interface between the IOBs and the internal logic array; it mainly includes:

(1) The interconnect between the IOB and the internal routing resource;

(2) The interconnect between the I/O routing resource and the internal routing resource;

(3) The I/O interconnect ring between the IOB and the internal logic array.

The length of the interconnect lines are insufficient near the boundary of the chip, and this causes many hanged end-points. A very important function of the I/O routing resource is to deal with these hanged end-points. So, the method of complementary hanged end-lines, as shown in Fig. 4, is adopted. For example, a set of length-N lines forms length 1, 2, . . . , N – 1 hanged lines at the boundary, and connects this set of hanged lines with the length N – 1, . . . , 2, 1 hanged lines of the other set of length-N lines at the boundary. Then, it forms the return lines around the internal logic arrays.

The return lines make all the length-N lines over the chip span N + 1 CLBs from start points to end points. This not only increases the routing resource of the boundary of the chip but also cause the load of the same kind of lines at the boundary to not increase.

The complementary hanged end-lines strategy guarantees that the same kind of lines has an identical interconnect pattern, and the load of them is equally distributed. So, it makes the signal delay uniform among the same kind of interconnect lines.

3. Design and implementation of the routing circuit

3.1. Delay-prediction design

Because of the programmability of the FPGA interconnect lines, the delay of every signal transmission path is possibly uncertain under different configurations of SRAMs, and this will lead to the unpredictability of the signal delay. However, a BUFFER provides an excellent isolation effect. For example, when a BUFFER is inserted into the output of a high-fanout signal, the signal delay will not be affected by the number of fanouts, which can eliminate the signal delay uncer-

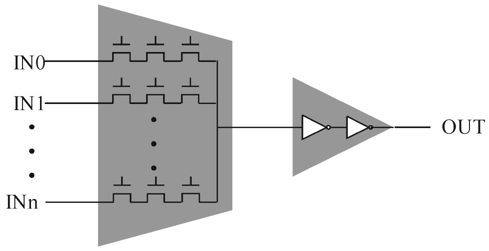

Fig. 5. MUX+BUFFER architecture as PIPs.

tainty.

In addition, the interconnect lines are usually very long, and the delay due to them cannot be omitted. However, because of the parasitic effects of long interconnect lines, the drive ability of the signal transmitted by a long interconnect line is insufficient when it arrives at the end; so, a BUFFER should be inserted into the long interconnect lines to promise the fast signal transmission.

As the characteristic size decreases, the threshold voltage in the power supply increases. When a signal is transmitted from the pass transistor, the threshold loss will be increased; so, the signal integrity becomes worse if the NMOS pass transistor alone is used as PIP.

Based on the ideas above, the MUX+BUFFER (Fig. 5) architecture is designed to build the programmable switch, and also BUFFERs are inserted into the output of high-fanout signals, such as the output of OMUX and CLBs to ensure the predictability of the signal delay through the interconnect lines.

For example, a signal transmitted from the output of a certain CLB to the input of another CLB will pass by four PIPs at most, including the PIP in OM, the PIPs connecting long lines or length-*N* lines to single lines (1 or 2 PIPs), and the PIP in the IM. Because every PIP is followed by a BUFFER, the load unpredictability caused by the configuration of the PIPs is neglected; so, the delay of each level of the transmission path can be calculated separately and the delay of the interconnect line is determined by the number of TILEs it spans. So, the delay of a signal transmitted by interconnect lines can be calculated by simply adding the delay of each level together. This promises that the delay of the signal transmitted by interconnect lines is highly predictable.

3.2. MUX implementation

In the design of the FPGA routing circuit, the speed should be maximized and the area should be minimized. As mentioned in the section above, MUXs are largely used in the proposed routing circuit; so, it is desirable to minimize the area of the MUXs and their corresponding SRAMs.

It is widely known that two memory cells can select up to four input signals, and three memory cells can select up to eight input signals. However, when two memory cells are used, two level pass transistors will be needed from input to output, and three levels of pass transistors are needed when three memory cells are used. The signal delay is directly affected by the level of the pass transistors. So, in order to balance the delay and the area, the coding pattern of the MUX circuit have to be compromised. Meanwhile, the size of the transistors used in the MUX circuit is simulated and optimized to gain the best signal delay. Furthermore, because of the programmability of the FPGA chip, the path of the MUX is selected and the input signals of the MUX are both uncertain before the configuration. So, the output of the MUX is uncertain and can possibly be floating or in conflict. In order to avoid these cases, an accessory path is added to the MUX. The input of this path is fixed at V_{DD} (power supply) when the FPGA chip is powered on and reset. This path of any MUX is always selected to promise the output is set to a high level.

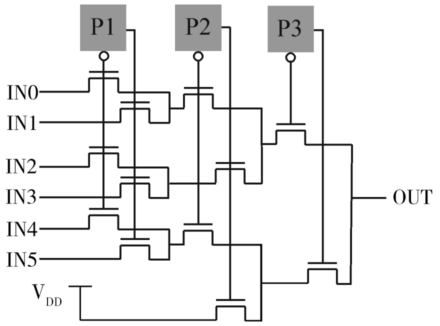

Fig. 6. 6-to-1 MUX.

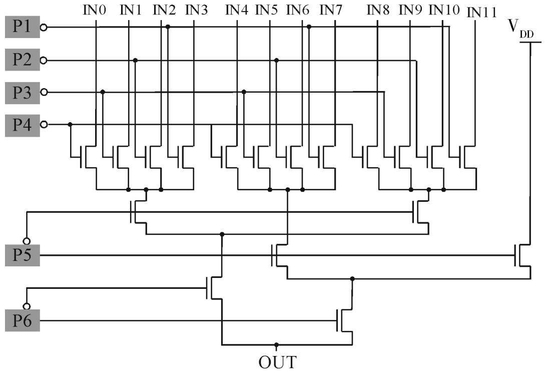

Fig. 7. 12-to-1 MUX used in the OM.

A 6-to-1 MUX (Fig. 6) is widely used in multi-length lines of the proposed routing circuit; so, the full coding pattern is adopted. The size of the NMOS transistor is set to W/L = 15, which leads to the best signal delay in this circuit.

A 12-to-1 MUX is used in the OM because all the output signals of the CLBs should pass by the OM. The delay of the 12-to-1 MUX should not be too large. As full coding pattern will need four level pass transistors, the following circuit (Fig. 7) is used, and the size of the first level NMOS transistor is W/L = 10 and for the the second and third level it is W/L = 13.

The 26-to-1 MUX (Fig. 8) is used in the IM, and it is built by a 6-to-1 MUX plus an 8-to-1 MUX. There is only one level in the 6-to-1 MUX, which is the first level in the 12-to-1 MUX shown in Fig. 7. So, it needs six memory cells, and the size of the NMOS transistors is set to W/L = 8. The 8-to-1 MUX adopts the full coding circuit. Only three memory cells are needed, and the size of the NMOS transistors is set to W/L = 10. So, the number of total memory cells is nine and the transistor level is four, which realize the delay and area tradeoff.

3.3. BUFFER optimization

The BUFFER with level-resume circuit is quite attrac-

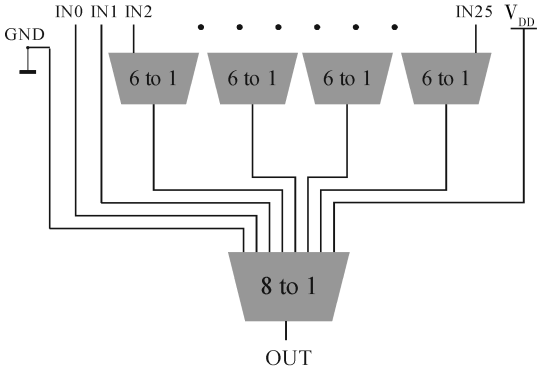

Fig. 8. 26-to-1 MUX used in the IM.

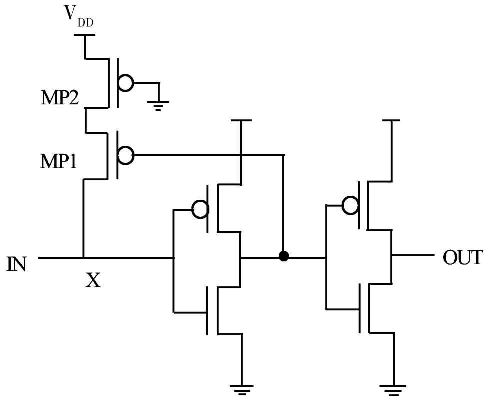

Fig. 9. BUFFER with pull-up transistors.

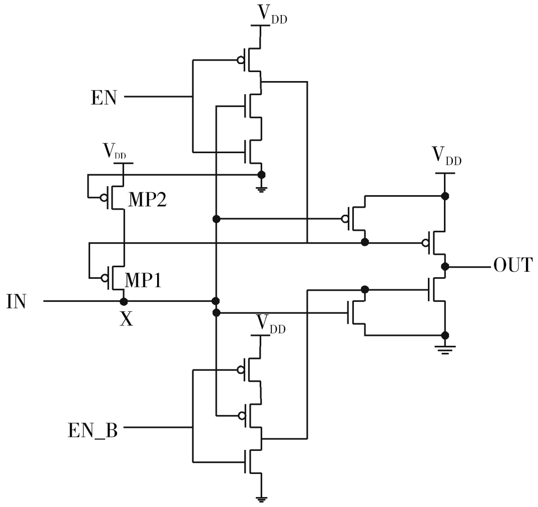

Fig. 10. Tri-BUFFER with pull-up transistors.

tive in reducing static power (Figs. 9 and 10). Before the signal transmits to the BUFFER, it passes by at least two levels of pass transistors (determined by the architecture of the MUX). This induces that the pull-down resistance of the point X becomes larger. To balance the pull-up and pull-down resistances, the PMOS transistor MP2 is inserted between V_{DD} and MP1. The size of MP2 is small, and the gate of it is connected to GND. This will decrease the parasitic capacitance of point X and speed up the pull-down process, and, thus, finally decrease the delay of the BUFFER.

Three kinds of BUFFERs, used in the implemented chip, are simulated, as shown in Table 1. Among them, BUF_BIHEX is a three-state BUFFER, and the other two are ordinary ones. The simulation results indicate that the delay of the BUFFER with a pullup is reduced by nearly 5%.

Table 1. Comparison of the delay of buffers with pullup and without pullup.

Buffer	Delay (10^{-10} s)	
	Without pullup	With pullup
BUF_UNIHEX	1.23	1.20
BUF_BIHEX	1.21	1.19
BUF_OMUX	1.45	1.40

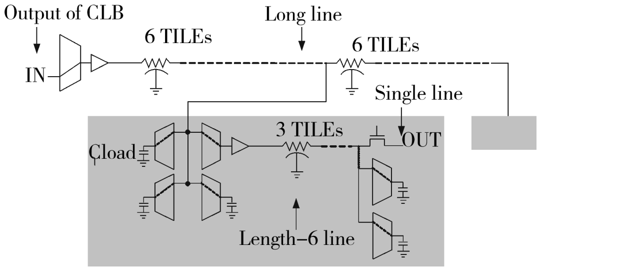

Fig. 11. Simulation module of the full path (The content of the grey boxes is the same).

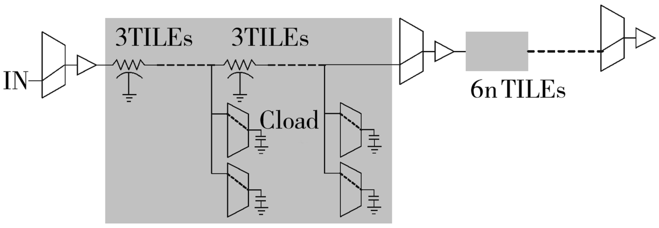

Fig. 12. Simulation module of length-6 lines (The content of the grey boxes is the same).

3.4. Modeling and simulation

As the technology develops into the deep sub-micron region, the delay on the interconnect line is comparable with that of the gate. So, the RC parasitic parameters of the metal line are drawn from the layout and taken into consideration to set up the simulation modules.

The functionality of the FPGA chip is uncertain before the configuration, so is the interconnect relationship. The simulation of the FPGA routing resource should cover all the possible connection pattern[7]. However, based on the proposed delay-predictable routing architecture, only one simulation model (Fig. 11, but not all parts are shown) is needed, viz. , the long line drives the multi-length line, the multi-length line drives the single line, and the single line to the CLB input. This module covers all the interconnect paths, and the path delay can be obtained by controlling the PIPs.

In order to validate the delay prediction design idea, a simulation module of length-6 lines and long lines are set up, and the general parts are shown in Fig. 12 (the module of long lines is similar). Figure 13 is the simulation result of length-6 lines and long lines[8]. It can be seen that, when the signal transmits through the long lines and the length-6 lines, the signal delay increases linearly, that is, a fixed signal delay is added with an increasing segment of the long lines or the length-6 lines. So, the signal delay over the chip is highly uniform and predictable.

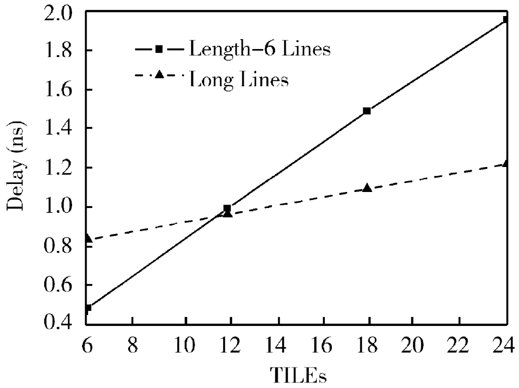

Fig. 13. Simulation result of length-6 lines and long lines.

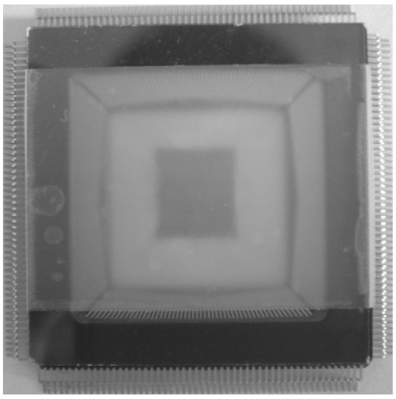

Fig. 14. FDP chip.

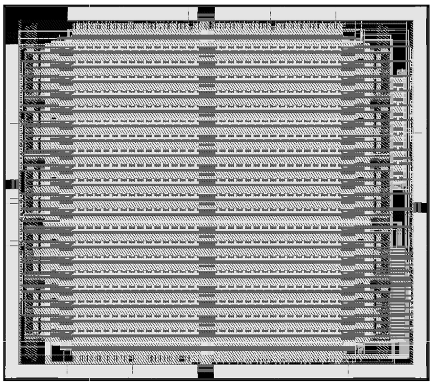

Fig. 15. Layout of the FDP.

4. Layout and test result

The proposed routing circuit is designed and applied to the FDP FPGA chip (Figs. 14 and 15), which is taped out with the SMIC 0.18-μm logic 1P6M process. The chip is designed by the full-custom method and the array of logic TILEs is 20 × 30. In this chip the multi-length lines is length-6 lines, and the long lines are buffered every 6 tiles. The die size of the FDP chip is about 6.5 × 6.8 mm^2. The routing resource costs approximately 60% of the total layout area.

The functional realization of the FPGA depends on the corresponding CAD software. An integrated CAD FPGA design environment (FDE) tool has been developed, which includes netlist conversion, partition, technology mapping, placement and routing, and a configuration bit file generation module. The FDE tool is omitted in this paper[9].

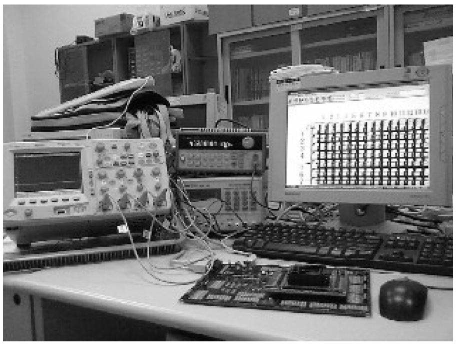

Fig. 16. Test board and the test platform.

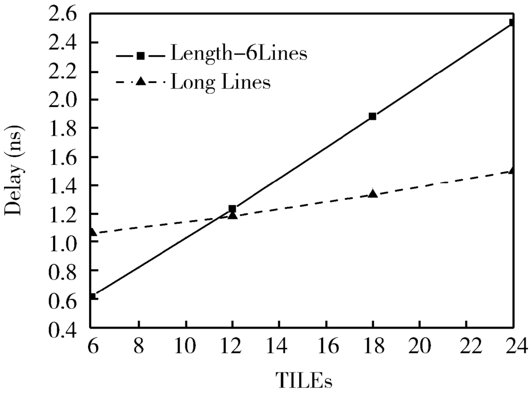

Fig. 17. Test result of length-6 lines.

For evaluating the performance of the programmable routing circuit in the FDP, a test platform is built, as shown in Fig. 16.

The FDP is systematically tested, and all the interconnect lines and the routing circuit work correctly. A series of test benchmark circuits, such as a music player and a timer, are implemented in the FDP. By using all the three kinds of interconnect lines and their corresponding routing circuits, the FDP could execute the corresponding functions correctly.

In the FDE flow, the routing process is done automatically and the usage of interconnect lines is not controlled. In order to test the delay of interconnect lines, the bit file should be created by hand.

The test result of length-6 lines and long lines are shown in Fig. 17. The linear relation between the number of TILEs, which the length-6 lines or long lines span, and the signal delay is quite obvious; and, in addition, the delay of the lines is confirmed to the ns degree, which further proves the performance and quality of the proposed design. Compared with Fig. 13, we can see that the test result is about 30% higher than the simulation result. This can be caused by the additional load in the test board and the limited precision of the oscillograph. It is also shown in Figs. 13 and 17 that long lines are better than length-6 lines for long-distance signal transmission.

Among existing FPGA chips, the Spartan2e family devices of Xilinx also adopted single line[5], length-6 lines, and long lines as their routing resource, which is similar to the proposed routing circuit. So, Table 2 compares the interconnect line delay of the two routing circuits. The line delay of the Xilinx Spartan2e device is drawn from the ISE flow. It is obvious that the proposed routing circuit improves the delay by 3%.

Table 2. Comparison of the line delay.

Routing circuit	Delay (ns)	
	Proposed routing	Xilinx Spartan2e routing
Length-6 lines (6 tiles)	0.48	0.50
Long lines (12 tiles)	0.96	0.99

5. Conclusions

A delay-optimized universal FPGA routing circuit is designed and implemented in the FDP chip, which has been taped out with the SMIC 0.18-μm CMOS technology. By replacing the CB & SB in a traditional FPGA routing architecture with GRB, IM, and OM and by adopting the offset lines and return lines strategy, the FPGA chip is highly repeatable, and the signal delay over the total chip is uniform and predictable. The novel MUX+BUFFER architecture is used as a PIP and also optimized to reduce the signal delay by up to 5%. The line delay is improved by 3% compared with the Xilinx Spartan2e devices. The test results show that all the interconnect lines and the routing circuit work correctly, and the signal delay over the chip is highly uniform and predictable.

References

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

[2] Ahmed E, Rose J. The effect of LUT and cluster size on deep-submicron FPGA performance and density. Proc ACM Int Symp FPGAs, 2000: 3

[3] Wang M, Ranjan A. Multi-million gate FPGA physical design challenges. Proc ACM ICCAD, 2003

[4] Hutton M. Interconnect prediction for programmable logic devices. Proc ACM SLIP, 2001

[5] The Programmable Logic Data Book. Xilinx Inc, 2004

[6] Young S P, Bauer T J. Input/output interconnect circuit for FPGAs. USA Patent, No. 6 204 689 B1. Mar 20, 2001

[7] Tu R. Design and implementation of 0.18 μm FPGA routing resource. Master Dissertation, Fudan University, 2007

[8] Wu F, Zhang H. A delay-optimized universal FPGA routing architecture. Proc ASPDAC, 2009

[9] Chen Liguang, Wang Yabin, Wu Fang, et al. Design and implementation of an FDP FPGA. Journal of Semiconductors, 2008, 29(4): 713