A DIRECTIONAL COARSE-GRAINED POWER GATED FPGA SWITCH BOX AND POWER GATING AWARE ROUTING ALGORITHM

Chin Hau Hoo, Yajun Ha and Akash Kumar

Department of Electrical and Computer Engineering
National University of Singapore, Singapore
Email: {a0045424, elehy, eleak}@nus.edu.sg

ABSTRACT

Leakage power has become an important component of the total power consumption in FPGAs as process technology shrinks. In addition, a significant amount of leakage power in FPGAs is consumed by the routing resources. Therefore, leakage power reduction in FPGAs should begin with the routing resources. In this paper, we propose a novel directional coarse-grained power gating architecture for switch boxes. In addition, the existing VPR routing algorithm has been adapted with a new cost function to support the new power gating architecture. Results have shown that the new cost function yields an average improvement of 22% as compared to the existing VPR cost function in terms of the number of power gating regions that can be turned off.

Keywords: Coarse-grained power gating, power-aware routing algorithm.

1. INTRODUCTION

FPGA has several advantages over ASIC – faster time to market, lower NRE and post fabrication programmability. However, it has not found a widespread use in mobile applications. One of the major reasons is due to the power gap between FPGA and ASIC. It has been estimated that gap is around 7 to 14X [1]. As mobile applications are usually battery powered, this power gap deters the use of power hungry FPGA devices.

The power consumption of FPGA devices can be broken down into two main categories – dynamic power and static power. Dynamic power is consumed by parts of the FPGA that are actively switching while static power is consumed by the inactive parts of the FPGA. As transistor size shrinks, static power starts to dominate the total power consumption [2]. In addition, routing resources consumes 36% of the total static power in a 90 nm commercial FPGA [3]. Therefore, the reduction of static power of FPGA routing resources is crucial to allow FPGA to compete with ASIC.

There are various techniques for reducing static power. One of the most effective ones is power gating (PG) where high V_{th} sleep transistors are inserted between the power rails and the logic block. Inactive regions of the circuit can be turned off by the sleep transistor to reduce leakage power. The sleep transistors have to be sized properly to achieve a good balance among leakage reduction, area overhead and circuit performance. In addition, the PG granularity is an important factor as well. Fine-grained PG allows for easier sizing of sleep transistors [4] but has a larger area overhead as compared to coarse-grained PG [5].

Contributions: After considering the trade-offs of different PG designs, we have made the following contributions in this paper:

- A novel directional coarse-grained power gated FPGA switch box.
- A power-aware routing algorithm to leverage on the new PG architecture.

In our proposed architecture, multiple buffers in each direction of the switch box are power gated independently of the buffers in the other directions. In the past, the difficulty of properly sizing the sleep transistors has been associated with coarse-grained PG of complex logic block. However, due to the homogeneous structure of the switch box, properly sizing the sleep transistors is not an issue.

To maximize the leakage reduction of the coarse-grained PG architecture, the routing algorithm has to be adapted as well. We have proposed a new cost function for the VPR routing algorithm to support the new routing architecture. The new cost function has been evaluated with various benchmarks, and the result shows a significant improvement of 22% on average as compared to the existing VPR routing algorithm.

The rest of this paper is structured as follows. Section 2 provides a brief overview of the recent works on FPGA leakage power reduction. Then, the new switch box power architecture is explained in Section 3. Section 4 describes the new cost function that supports the coarse-grained PG architecture, and its performance is analyzed in Section 5. Finally, Section 6 presents the conclusion of our work.

2. RELATED WORKS

Due to the increasing contribution of leakage power towards total power consumption as process technology shrinks, several recent works on FPGAs have focused on the reduction of leakage power. One of them proposed a coarse-grained PG architecture where a cluster of logic blocks and

connection blocks are power gated by a single sleep transistor [6]. The research has shown that leakage reduction of up to 40% is possible. In addition, the area overhead of such an architecture decreases as the cluster size increases. However, the work does not consider PG of switch boxes.

In [7], a fine-grained PG architecture has been proposed where each routing switch is power gated independently. Although the author claimed that the approach achieved at leakage reduction of 38.18%, the architecture has a very high area overhead of 186%.

Anderson [8] proposed a number of FPGA routing switch designs that can be programmed to operate in three different modes high-speed, low-power and sleep mode. The different modes offer different degree of leakage power and dynamic power consumption. However, two SRAM cells are required to switch among the three modes, and this incurs extra area overhead. Although the authors suggested that one of the SRAM cells can be shared among multiple routing switches to reduce the area overhead, they did not describe how the sharing can be done. Krishnan [9] presented a novel routing switch that is based on a variation of the dual threshold CMOS. Each routing switch requires two transistors but the augmenting or limiting transistor can be shared. This work suffers from the same limitation as the one by Anderson [8]. An efficient way of sharing the augmenting or limiting transistor to reduce the area overhead was not described. We strongly believe that it is important to have a CAD algorithm that complements the proposed hardware architecture, and this issue is addressed in our work.

Anderson et al. described a method to reduce leakage power in FPGAs without incurring any area overhead [10]. They leveraged on the fact that leakage power of multiplexers is strongly dependent on the voltage levels at the input and the output. In addition, they have identified another important property that logic signals in FPGA can be interchanged with their complements without incurring any area overhead. By carefully selecting the SRAM cell content in the LUTs, the multiplexers in the LUTs can be placed in low leakage state. The idea has been adapted to reduce leakage power in routing multiplexers in [11] and [12]. In fact, the approach is orthogonal to the idea presented in this work, and can be applied to our proposed architecture to achieve further leakage power reduction.

3. POWER GATING OF SWITCH BOXES

Figure 1(a) shows an example of our proposed architecture with two power gated regions in a switch box, indicated by the dashed lines. We define the size of power gated region as the number of buffers in the region. In this example, the power gated regions are of size 2. The switch box is based on the unidirectional single driver architecture described in [13], which has lower area and delay as compared to bidirectional multiple driver architecture. The multiplexer is the starting point of a segment, and the starting points are staggered so that there is no connection to the middle of the segments, which is required for the single driver architecture

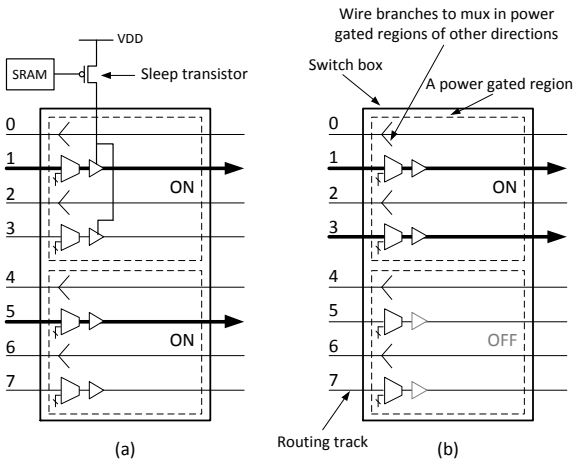

Fig. 1. (a) Existing VPR routing algorithm (b) Power aware routing algorithm

[14]. The other inputs of the multiplexer are wires coming from other directions of the switch box and outputs from the CLBs, which are not shown in the figure for simplicity purposes.

As shown in one of the PG regions in Figure 1(a), the buffers in our coarse-grained PG architecture is connected to the VDD rail of via a high V_{th} PMOS sleep transistor, and the sleep transistor is controlled by a SRAM cell. Based on the formula described in [15], the sleep transistor is sized such that the degradation in propagation delay of the buffers due to the presence of sleep transistor is limited to 5%. The coarse grained PG architecture has the advantage of smaller area overhead because only one SRAM cell is required for each PG region. In addition, as the PG region size increases, the area overhead due to the SRAM cell decreases.

To the best of our knowledge, none of the previous works have described how the buffers and multiplexers in the switch box should be grouped into a single PG region. In our architecture, the buffers in each direction of the switch box are power gated independently of the buffers in the other directions. We evaluate the advantage of this directional PG in Section 5.5.

4. POWER AWARE ROUTING ALGORITHM

4.1. Motivating Example

Figure 1 illustrates the need for the existing VPR algorithm to be adapted to take advantage of the proposed PG architecture. Consider the case where two nets needs to be routed through a switch box from the west direction to the east direction, and the nets have the lowest delay when routed through track 1 and 5. The existing VPR router will route the nets as shown in Figure 1(a). However, routing the nets in this manner results in no saving in terms of leakage power because both power gated regions have to be turned on.

The objective of adapting the existing routing algorithm is to achieve a routing result shown in Figure 1(b). In this case, only one power gated region needs to be turned on, resulting in a leakage power reduction of 50%. However, if the net on track 3 is in the critical path, the critical path delay may be increased because the net is taking a slower path. This impact on critical path delay needs to be carefully considered when routing. The goal is to maximize the number of inactive power gated regions while minimizing the effect on the critical path delay. The impact on circuit performance is evaluated in Section 5.4.

4.2. Proposed Cost Function

The routing algorithm of the latest VPR tool is based on the Pathfinder algorithm [16]. However, the cost of using a particular node in the routing resource graph is calculated based on delay and congestion only. In order to maximize the utilization of PG regions, the cost of using a PG region has to be considered.

$$\begin{aligned} cost =& crit_{net} \times t_{net} + (1 - crit_{net}) \times congestionCost \\ &+ k \times baseCost \times e^{-numNetsInPGRegion} \end{aligned} \quad (1)$$

Equation 1 shows the modified cost function, which accounts for the usage of a PG region. The first two terms of the equation are the same as the baseline VPR algorithm. The last term calculates the cost of using a PG region based on three parameters. The parameters are a scaling factor k, the base cost of using a routing resource node, and the number of nets that are routed through the PG region that the routing resource (RR) node is associated with.

The scaling factor k is chosen to be the PG region size. In essence, routing a net through a small unused PG region is not penalized as heavily as routing a net through a large unused PG region. The base cost of using a RR node is the same as the one used in the calculation of congestion cost in the baseline VPR routing algorithm. Finally, the cost is chosen to scale exponentially with the number of nets that passes through a PG region to discourage the use of an inactive PG region.

5. RESULTS AND ANALYSIS

5.1. Experiment Procedures

In our experiments, we targeted an island style FPGA architecture. Each configurable logic block has ten basic logic elements (BLEs), and each BLE contains a 6-input LUT and a flip-flop. The switch box topology is Wilton, and the switch box flexibility, F_s is 3. The input pin connection box flexibility, $F_{c,in}$ is 0.15 while the output pin connection box flexibility, $F_{c,out}$ is 0.1. The routing tracks are unidirectional, and the segment length is 4.

We evaluated the effectiveness of our new cost function with twenty of the largest MCNC benchmark circuits.

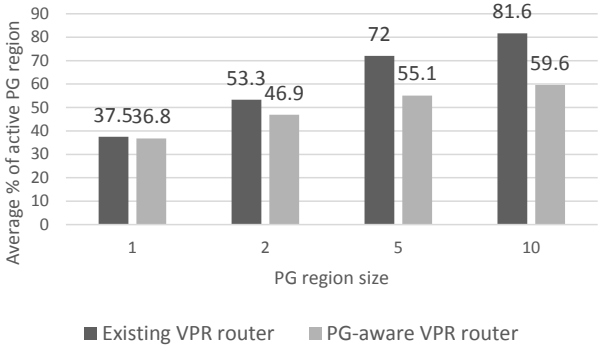

Fig. 2. Average percentage of active PG region for different PG region sizes

The circuits were provided by the VPR 6.0 package, and they were synthesized and technology mapped to the aforementioned FPGA architecture. Then, they were placed and routed with the baseline VPR 6.0 tool with the default placement and routing options to determine the minimum number of tracks required to successfully route each design. The maximum of the twenty minimum track counts was determined, and it turned out to be 68. Due to the constraints of our architecture, the number of tracks must be a muliple of eight [13]. Therefore, we chose a track count of 80 for the purpose of evaluating the performance of our PG-aware router. With 80 tracks, there are eight buffers in each direction of the switch box. Therefore, the possible PG region sizes for evaluation are the divisors of eight, which turns out to be one, two, five and ten.

5.2. Average Percentage of Active PG Region

The numbers in the Figure 2 shows the percentage of active PG region that is averaged over the twenty MCNC benchmark circuits. A PG region size of 1 is the finest PG granularity while a PG region size of 10 is coarsest PG granularity in our architecture. For a PG region size of 1, the PG-aware router yields very little improvements because the existing VPR router is implicitly routing with a PG region size of 1. However, as the PG region size increases, the gap between the existing router and the PG-aware router increases because the existing router does not penalize usage of inactive PG region.

5.3. Effect of Unused Tracks on Router Performance

The number of unused tracks after routing depends on the size of the benchmark circuit and affects the amount of improvement of the PG-aware router over the existing router. From Figure 3, the PG-aware router achieves a very small improvement over the existing router for the largest MCNC benchmark, *clma*. The reason is that the number of alternative paths that minimize the number of active PG regions decreases as the benchmark circuit size increases for a constant number of tracks.

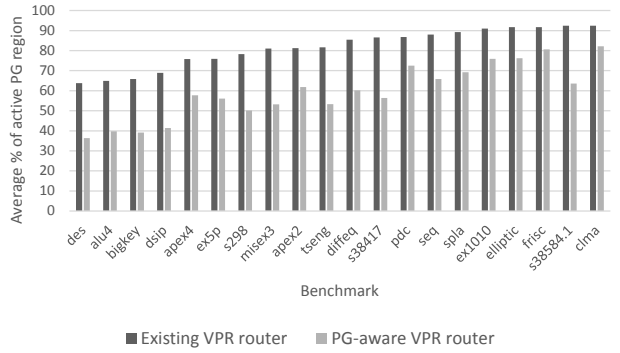

Fig. 3. Performance of routers in the twenty largest MCNC benchmarks for PG region size of 10

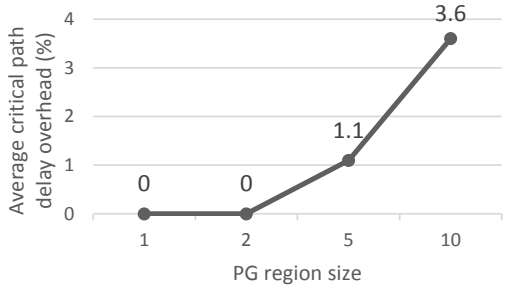

Fig. 4. Average critical path delay overhead of the new cost function for different PG region sizes

5.4. Critical Path Delay Overhead

From Figure 4, the average critical path delay overhead increases as the PG region size increases because the new cost function imposes a higher cost penalty for the use of larger PG region size. Therefore, the nets of the design might take a longer path in order to avoid being routed through a large inactive PG regions. Howerver, it is important to note that critical path delay overhead is limited to 3.6%. Therefore, the new cost function has minimal impact on the circuit performance even for the largest PG region size.

5.5. Directional PG versus non-directional PG

As shown in Figure 5, the number of active buffers after routing with the PG-aware router for non-directional PG is at least 1.58 times more than that of directional PG. Due to the Wilton switch box, the probability of the same track number being occupied in all four directions in the switch box is very low. Therefore, the tracks in the unused directions can be turned off in the directional PG architecture, resulting in a lower number of active buffers.

6. CONCLUSION

In this paper, a directional coarse-grained PG architecture for switch boxes in FPGA has been proposed. We have showed that with the new PG-aware routing algorithm, at least 40.4% of the PG regions can be switched off in this

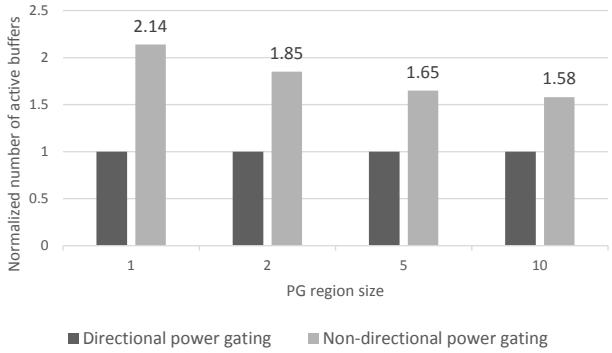

Fig. 5. Number of active buffers normalized to directional PG for different PG region sizes

coarse-grained architecture. In addition, the new routing algorithm has minimal impact on the critical path delay, and it has been showed the overhead is bounded to within 3.6%.

7. REFERENCES

[1] I. Kuon and J. Rose, "Measuring the gap between FPGAs and ASICs," *IEEE TCAD*, 2007.

[2] N. Kim, T. Austin, D. Baauw, T. Mudge, K. Flautner, J. Hu, M. Irwin, M. Kandemir, and V. Narayanan, "Leakage current: Moore's law meets static power," *IEEE Computer*, 2003.

[3] T. Tuan, S. Kao, A. Rahman, S. Das, and S. Trimberger, "A 90nm low-power FPGA for battery-powered applications," in *ACM/SIGDA FPGA*, 2006.

[4] B. Calhoun, F. Honore, and A. Chandrakasan, "A leakage reduction methodology for distributed MTCMOS," *IEEE JSSC*, 2004.

[5] A. Rahman, S. Das, T. Tuan, and S. Trimberger, "Determination of power gating granularity for FPGA fabric," in *IEEE CICC*, 2006.

[6] A. Bsoul and S. Wilton, "An FPGA architecture supporting dynamically controlled power gating," in *IEEE FPT*, 2010.

[7] Y. Lin, F. Li, and L. He, "Routing track duplication with fine-grained power-gating for FPGA interconnect power reduction," in *IEEE ASPDAC*, 2005.

[8] J. H. Anderson and F. N. Najm, "Low-power programmable FPGA routing circuitry," *IEEE TVLSI*, 2009.

[9] R. Krishnan and J. P. de Gyvez, "Low energy switch block for FPGAs," in *IEEE VLSID*, 2004.

[10] J. Anderson and F. Najm, "Active leakage power optimization for FPGAs," *IEEE TCAD*, 2006.

[11] S. Srinivasan, A. Gayasen, and N. Vijaykrishnan, "Leakage control in FPGA routing fabric," in *IEEE ASPDAC*, 2005.

[12] M. Hasan, A. Kureshi, and T. Arslan, "Leakage reduction in FPGA routing multiplexers," in *IEEE ISCAS*, 2009.

[13] G. Lemieux, E. Lee, M. Tom, and A. Yu, "Directional and single-driver wires in FPGA interconnect," in *IEEE FPT*, 2004.

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

[15] C. Long and L. He, "Distributed sleep transistor network for power reduction," *IEEE TVLSI*, 2004.

[16] L. McMurchie and C. Ebeling, "PathFinder: a negotiation-based performance-driven router for FPGAs," in *ACM/SIGDA FPGA*, 1995.