



ELSEVIER

SCIENCE @ DIRECT®

INTEGRATION, the VLSI journal ■ (■■■■) ■■■-■■■

INTEGRATION
the VLSI journalwww.elsevier.com/locate/vlsi

An area-efficient timing closure technique for FPGAs using Shannon's expansion

Deshanand P. Singh*, Stephen D. Brown

Department of Electrical and Computer Engineering, University of Toronto, Canada M5S3G4

Abstract

This paper presents a technique to optimize the speed performance of circuits implemented in FPGAs. After synthesis, technology mapping and placement are complete, we apply Shannon's expansion to the most critical sections of the circuit. This approach allows us to precompute the values of functions that depend on late-arriving critical signals and use a multiplexer to quickly select the appropriate value when the signal arrives. Any new logic elements created by this technique are incrementally placed in a minimally disruptive fashion to ensure convergence between the circuit optimization and the netlist placement. Experimental results show that this technique can improve the performance of circuits by 11% on average, and up to 30% in some cases.

© 2006 Published by Elsevier B.V.

Keywords: ■; ■; ■

1. Introduction

FPGA designers are increasingly facing the dilemma that their designs are dominated by the connection delays routed along the programmable interconnect. Traditional solutions to this problem involve improvements to automated placement and routing algorithms; however, we can now go one step further by restructuring the circuit itself to better cope with the interconnect delays.

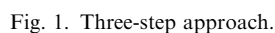
This alternative, and perhaps orthogonal, strategy involves tightly coupling timing-driven circuit optimizations with the placement step of the CAD flow. Within the placement phase there is still the freedom to add new elements to the netlist of logic elements, and the routing delays can be accurately approximated for many architectures. In this manner, critical portions of the circuit can be restructured to account for the routing delays. We know of no reported work that can simultaneously optimize a general circuit and produce a legal placement that respects the many constraints that exist in modern FPGA architectures. Thus, we have adopted a three-step approach to the coupling of placement with netlist optimizations as

shown in Fig. 1. Concepts like this three-step flow have been explored for ASIC implementations in academia [1–3] and industry [4–6].

The first step executes the conventional FPGA CAD flow of *HDL* → *synthesis* → *techmapping* → *placement*. In the second step, routing delays for every connection are estimated by calculating their fastest possible route. Any timing-driven netlist optimization technique can then be applied to perturb the circuit to reduce the critical path(s). The estimation and optimization techniques are collectively referred to as *layout-driven* optimizations since the layout of the logic elements directly affects how the circuit is perturbed. In this paper, we will focus on the use of Shannon's expansion as our netlist optimization technique, although other optimizations such as *sequential retiming* have been used in this flow [7]. Every additional logic element introduced in the circuit is given a preferred placement location. These preferred locations ignore even the most fundamental architectural constraints of the programmable device under consideration but are chosen strictly on the basis of improving timing. These preferred locations could form an illegal placement, but we rely on a subsequent fixup stage to remove the illegalities. This concept relaxes the problem of simultaneously optimizing a circuit and producing a legal placement. Our netlist optimization

*Corresponding author.

E-mail addresses: singhd@eecg.toronto.edu (D.P. Singh), brown@eecg.toronto.edu (S.D. Brown).



The application of Shannon’s expansion is performed after placement so that the circuit can cope with routing delays. One of the natural questions concerning these techniques is: how necessary is it that they be applied as a post-placement step rather than as part of LUT-level synthesis? Our implementation of Shannon’s expansion relies heavily on finding critical sections of logic. Here we show that it is extremely difficult to predict the post-placement critical connections from a LUT-level netlist. In this experiment, we refer to a connection as being *critical* if its criticality is greater than 0.9. Two different statistics are

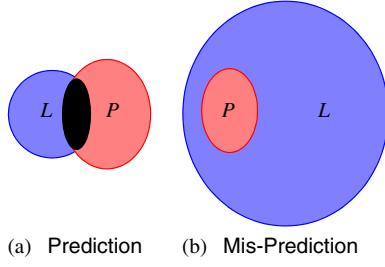


Fig. 3. Post-placement optimization; Prediction and mis-prediction rates.

considered. These are depicted graphically in Fig. 3(a). The first is the *prediction rate*. This is the ratio of connections that are critical in both the LUT-level netlist (highlighted in blue in Fig. 3) and the post-placement netlist (highlighted in red) divided by the number of critical connections in the post-placement netlist. The critical connections in the LUT-level netlist are computed by only considering the logic delay and assuming 0 delay routing connections. If the prediction rate is low, then the circuit restructuring techniques may be applied to the wrong paths and not help the true critical paths that exist post-placement. This statistic attempts to quantify how much of the timing critical region of the circuit can be predicted before the place and route process. Mathematically, this can be expressed in the following manner:

- Let L represent the critical connections in the LUT-level netlist.
- Let P represent the critical connections in the post-placement netlist.

$$\text{PredictionRate} = |P \cap L|/|P|. \quad (2)$$

We often find that the prediction rate approaches 100% for some circuits, but this is because of the situation shown in Fig. 3(b). In this case the LUT-level netlist predicts too many connections as being critical. These mis-predictions are troublesome because it means that we will apply our optimization techniques to far too many regions of logic that end up being non-critical after placement, and perhaps wasting area to improve portions of logic that are not really critical. In addition, mis-prediction also hurts our algorithm because it essentially moves delays from critical to non-critical paths. Since many non-critical paths are considered critical at the LUT-level, this tradeoff may not be possible. The mis-prediction rate can be expressed as follows:

$$\text{MisPredictionRate} = |\bar{P} \cap L|/|P|. \quad (3)$$

Table 1 shows these statistics for a cross-section of benchmark circuits. Generally the prediction rate is either extremely low, or the mis-prediction rate is extremely high. A circuit that exhibits either of these characteristics may prove challenging to apply restructuring techniques at the LUT-level. It is these results that motivate the optimization of critical paths after placement has been completed.

Table 1
Prediction rates

Circuit	#LUT	Predict (%)	Mis-predict (%)
bigkey-mcnc	1707	100	3675
dsip-mcnc	1370	100	858
diffeq-mcnc	1497	100	59.3
elliptic-mcnc	3604	0	15.2
frisc-mcnc	3556	57.8	7.32
s38417-mcnc	6406	94.2	5703
tseng-mcnc	1047	97.3	27.2
hc11-oc	3877	94.3	15
des-fip	15509	0	329
sisc8	1434	66.1	0.41

4. Definitions

The algorithms described in subsequent sections rely heavily on being able to manipulate critical regions of logic. This section presents definitions that will allow us to easily express how critical regions of logic can be manipulated.

First the most critical nodes and connections in the netlist must be established. These critical elements can be represented by an ε -graph which contains elements that are within ε of being critical. The value of ε is typically a small number that determines the threshold of what logic is considered critical. If ε is too large then Shannon's decomposition will be applied to many nodes that do not affect the path. The net effect is simply a waste of area. If ε is too small then decompositions are applied to the critical path, but are not applied to any of the near-critical paths.

Recall that a combinational circuit, or the combinational portion of a sequential circuit, can be represented by a directed acyclic graph $G(V, E)$. Each vertex in the graph corresponds to a combinational element in the corresponding circuit. The edges represent connections between the combinational elements. The ε -graph $G_\varepsilon(V_\varepsilon, E_\varepsilon)$ is a subgraph of the combinational graph $G(V, E)$. The edge set E_ε is defined as the set of edges that are within ε of being critical

$$E_\varepsilon = \{\forall e \in E \mid \text{crit}(e) \geq 1 - \varepsilon\}. \quad (4)$$

The critical vertices V_ε are the set of all vertices that have any adjacent edge which is within ε of being critical.

$$V_\varepsilon = \{\forall v \in V \mid (\exists e_{uv} \text{st. } \text{crit}(e) \geq 1 - \varepsilon)\} \cup \{\forall v \in V \mid (\exists e_{vu} \text{st. } \text{crit}(e) \geq 1 - \varepsilon)\}. \quad (5)$$

By construction, $E_\varepsilon \subseteq E$ and $V_\varepsilon \subseteq V$.

5. Applying Shannon's theorem

Consider a critical late-arriving signal x that we wish to transform such that signals affected by x incur less delay due to logic elements. Assume that $x \in V_\varepsilon$ since we are applying the decomposition techniques to critical signals. The first step is shown in Fig. 4. We find all critical vertices

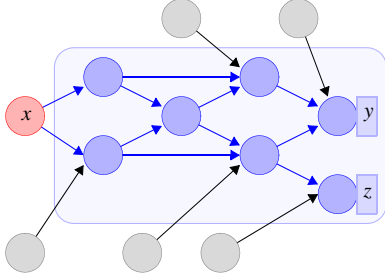


Fig. 4. Finding the transitive fanouts.

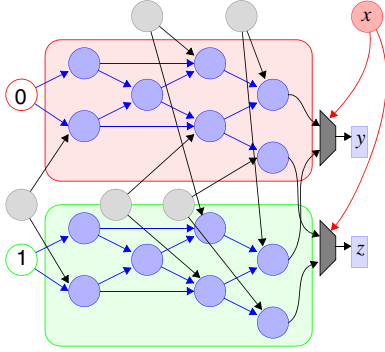


Fig. 5. Duplication of the transitive fanouts.

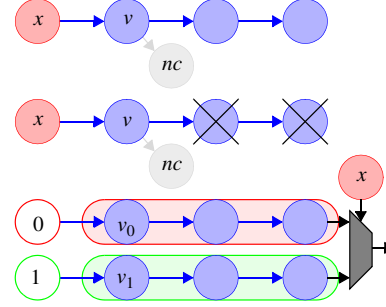


Fig. 6. Finding the vertices to delete.

The next step in the application of Shannon's decomposition is to add multiplexer logic to select from the cofactors. The multiplexer is controlled by the critical signal x on which the decomposition was based.

The final step is to remove any vertices that have become unused. Recall that each vertex $v \in TF_\varepsilon(x)$ is duplicated into two copies v_0 and v_1 . Obviously v_0 and v_1 are used in place of v so the natural question is: can v be deleted? In Fig. 5, the non-duplicated logic is no longer needed because the duplicated versions serve to produce the functions y and z . However, if any of these vertices $v \in TF_\varepsilon(x)$ was used as an input to another function $nc(\dots, v, \dots)$ such that $nc \notin TF_\varepsilon(x)$ then the vertex v cannot be deleted. This situation is depicted in Fig. 6. It occurs because we apply Shannon's decomposition to the ε -graph which is a subgraph of the entire combinational circuit. Hence non-critical fanouts of v still require the signal produced by v . This could easily be accomplished by keeping the original vertex v or by using $\bar{x}v_0 + xv_1$. Keeping the original vertex is preferable as it does not affect the number of levels of logic element delay sent to the non-critical fanouts. Thus a vertex $v \in TF_\varepsilon(x)$ is deleted if and only if $TF(v) \subseteq TF_\varepsilon(v)$.

6. Selecting logic to expand

Although it would be possible to apply the techniques described in the previous section to any critical vertex in G_ε , a method is needed to try to minimize the amount of duplication produced and to ensure that the decomposition techniques actually provide timing improvement. These goals are achieved with a cost function that evaluates the suitability for applying the decomposition and duplication techniques to groups of logic.

The first part of the cost function attempts to quantify the number of critical or near critical paths that a particular vertex affects. This quantity is denoted with a label named *cpcount* for each vertex. These identifiers are all initially set to zero $\forall v \in V_\varepsilon, cpcount(v) = 0$.

Fig. 7 graphically depicts the algorithm used. Each path highlighted in red represents a single critical path traced backward from a sink vertex. The *cpcount* value of a vertex, v , is incremented whenever the traceback goes through v . In this example there are two nodes that affect four critical

that are affected by the signal x . These vertices can be found by examining the transitive fanouts of x in the subgraph G_ε . A vertex v is a transitive fanout of x if there exists a path from $x \rightarrow v$ in the subgraph G_ε . This operation is denoted $v \in TF_\varepsilon(x)$. In Fig. 4, the transitive fanouts of x in G_ε are encapsulated in the rounded rectangle. The signal x affects two critical signals $y(\dots, x, \dots)$ and $z(\dots, x, \dots)$. Shannon's decomposition can now be applied to these two functions.

Following the identification of the transitive fanout vertices, the next step is to create two duplicate copies for each of these vertices. One set of these vertices will be used to evaluate y and z for $x = 0$, while the second set will evaluate these functions for $x = 1$. Consider a vertex $v \in TF_\varepsilon(x)$. A duplicated version is needed to evaluate for $x = 0$; it is denoted v_0 . Similarly to evaluate for $x = 1$, the duplicated vertex is denoted v_1 . This duplication operation is shown in Fig. 5. New edges must also be added to wire the appropriate combinational cells together. For each vertex $v \in TF_\varepsilon(x)$ the following procedure is used to create new edges. For every edge $e_{uv} \in FANIN(v)$:

- if $u \in TF_\varepsilon(x)$, then create two new edges from $u_0 \rightarrow v_0$ and $u_1 \rightarrow v_1$. This wires together the intermediate signals used to compute y and z for $x = 0$ and $x = 1$.
- if $u \notin TF_\varepsilon(x)$, then create two new edges from $u \rightarrow v_0$ and $u \rightarrow v_1$. This procedure is used to wire the input signals into the logic that computes $y(\dots, 0, \dots)$, $y(\dots, 1, \dots)$, $z(\dots, 0, \dots)$ and $z(\dots, 1, \dots)$.

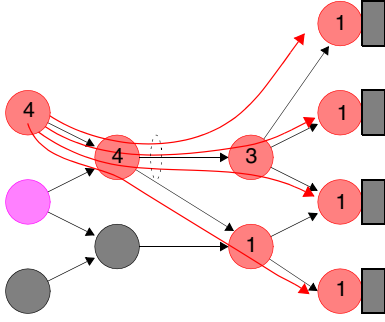


Fig. 7. Finding the most critical nodes.

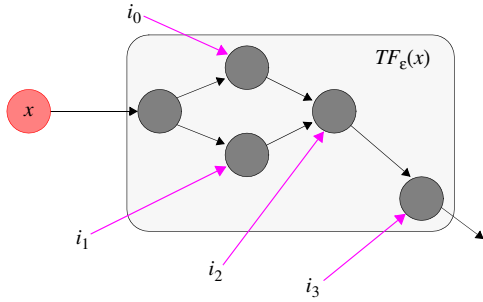


Fig. 8. Checking critical fanins.

sinks. If decomposition and duplication techniques are applied with respect to these nodes, then it may be possible to quickly reduce a number of paths by applying the techniques in the previous section. This algorithm is clearly heuristic in nature as there may be several different near critical paths that affect a sink vertex, instead of the single path that is traced backwards.

After a cost has been applied to each vertex, it is still necessary to check if the decomposition can actually provide any gain before applying it. Consider the situation shown in Fig. 8. We wish to perform the decomposition and duplication techniques with respect to the signal x . While this technique is beneficial for signals emanating from x , we have not considered the side effects of this operation. Consider the fanins labelled i_0, i_1, i_2, i_3 in the figure. After the application of decomposition and duplication, signals downstream from these connections experience an extra level of logic delay due to the multiplexer logic added at the sink nodes. This extra level of logic delay is acceptable as long as the slack on the connections $i_0 \dots i_3$ is greater than the amount of delay introduced by the multiplexer logic and the routing delay needed to wire the logic to the mux. Thus, these paths must be considered before applying our techniques.

More formally, the side effects can be checked in the following manner. Let side-input set I represent the fanin edges of $TF_e(x)$ whose source vertex is not an intermediate variable in $TF_e(x)$.

$$I = \{\forall e_{uv} \in FANIN(TF_e(x)) \mid u \notin TF_e(x)\}. \quad (6)$$

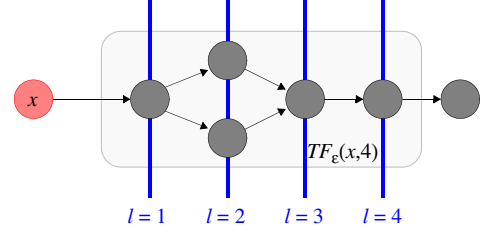


Fig. 9. Controlling vertex duplication.

Simply stated, I contains all external input edges to $TF_e(x)$. The decomposition and duplication technique can be applied as long as the following condition holds:

$$\forall i \in I, \quad \text{slack}(i) \geq \text{mux delay} + \text{routing delay to mux}. \quad (7)$$

7. Controlling duplication

The procedures described so far duplicate each vertex in $TF_e(x)$. However, this set may be arbitrarily large. The setsize is controlled by redefining the critical transitive fanout set. Each vertex in $v \in TF_e(x)$ is associated with a label $l(v)$ that is set to the maximum number of logic levels between x and v . The set $TF_e(x, D)$ represents all vertices v where there exists a path from $x \rightarrow v$ and $l(v) \leq D$. The value of D controls the tradeoff between the amount of duplication allowed and the number of levels of logic delay removed from critical paths. Note that the labels $l(v)$ can be computed using a breadth-first search from vertex x (Fig. 9).

All of the procedures described previously that use $TF_e(x)$ can easily be replaced by $TF_e(x, D)$. Experiments have shown that this strategy is quite effective as values of $D = 2$ or 3 effectively optimize the critical-path delay while adding a small number of extra logic elements. The reduction in the number of logic elements has the added benefit of making the job of the incremental placer easier as it has to resolve fewer architectural constraints.

8. Integration with ICP

Each of the fundamental operations described previously must also be tightly integrated with placement. In this way, we can accurately make decisions about critical regions of logic that can be optimized by the decomposition and duplication techniques. Fig. 10 shows an example of how the preferred locations are derived. The vertices s, t, u, v, w are duplicated into two sets to precompute for $x = 0$ and 1 . The preferred location of a given vertex v_0 is simply set to the current location of v . Similarly the preferred location of v_1 is also set to the current location of v . Recall that FPGA logic blocks typically consist of clusters of several logic elements for each physical cluster location. Hence it is possible for vertices v, v_0, v_1 to exist at the same physical location. Hopefully ICP can shift non-critical logic to other locations if there is not enough space.

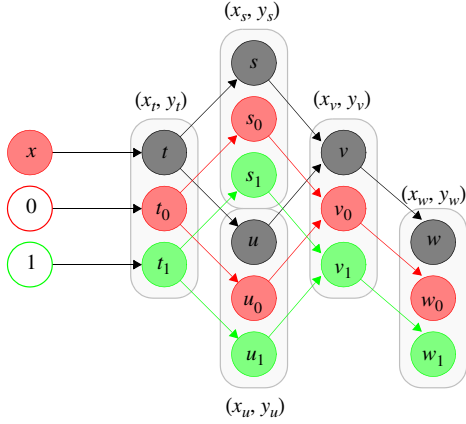


Fig. 10. Setting the preferred locations.

```

procedure optimize(  $G(V, E)$  )
begin
  do
    VerticesAdded = 0;
    do
       $G_e = \text{TimingAnalyzer}(G)$ 
       $\forall v \in \text{sinks}(V_e)$ , call  $\text{traceback}(v)$ ;
      select the "best" node  $x$  with the largest  $\text{cpcount}(x)$ 
        and large side-input slack using sliding window
      if  $x = \emptyset$  then exit-loop;
      apply decomposition and duplication w.r.t.  $x$ ;
      VerticesAdded += VerticesDuplicated - VerticesDeleted;
      if VerticesAdded > AvailableLUTs then exit-loop;
    end loop
    if VerticesAdded = 0 then exit-loop;
    call IncrementalPlacement();
  end loop
end procedure

```

Fig. 11. Iterative application.

9. Iterative application

Fig. 12 shows a graphical representation of how vertices are chosen for the application of Shannon's decomposition. The "best" nodes are chosen using a linear cost function given by

$$\text{cost}(x) = k_{cp} * \text{cpcount}(x) + \min_{i \in I} \text{slack}(i). \quad (8)$$

It simply states that we search for nodes with high cpcount along with large side-input slacks, so that the algorithm does not end up thrashing by fixing critical paths and then creating new ones. Searching for these nodes is very much like moving a small sliding window across the netlist that encompasses nodes and their transitive fanouts.

Fig. 11 shows how circuit performance can be optimized through the iterative application of the techniques described earlier. Notice that the routine contains two main loops. The inner loop controls the iterative application of Shannon's expansion. It is continued until the chip is full or

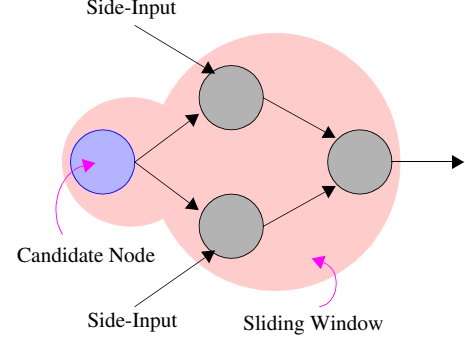


Fig. 12. Sliding window search.

there is no longer any benefit from applying Shannon's theorem.

The next step is a call to the incremental placement routine that attempts to map the preferred locations into legal physical locations on the target device. If the incremental placement routine succeeds, then the decomposition and duplication loop is attempted again because the new placement may perturb the timing characteristics of the circuit sufficiently so that decomposition may again be beneficial (Fig. 12).

Briefly, our implementation also includes area recovery steps such as constant propagation and incremental LUT mapping to collapse the multiplexer and logic that feeds it into a single lookup table.

10. Results

Our experiments involved using FlowMap [10] to map circuits to a netlist of four-input lookup tables. The timing-driven VPR [11] tool was used to place and route the design. Each circuit was mapped into a device such that the chip was not more than 90% utilized. The routing architecture was also chosen such that there are 20% more tracks than the minimum needed to route the circuit. This ensures that the device is not so congested that delays cannot be predicted from netlist placement.

Table 2 shows the results of the application of the techniques described in this paper after placement. The third column shows the speedup vs. the result obtained through the conventional CAD flow. As shown, the average speedup is 10.8%. If the incremental placement is replaced with a complete re-placement, the speedup is reduced to only 1.2%. In this case, there are no preferred locations. An optimized netlist is simply passed to the VPR and another complete placement and routing step is executed. This result is not unexpected, because there is no reason to expect that the optimized netlist will correspond to the new placement. Most practical annealing algorithms produce very different results even with small differences in their initial solution. For both of these experiments, the area penalty was 2.33%.

Table 2
Decomposition and duplication results

Circuit	Size (#LUTs)	T_P (ns), % Speedup Shannon + ICP	T_P (ns), % Speedup Shannon + Replace	Area increase (%)
bigkey-mcnc	1707	7.68 (+7.8%)	8.24 (+0.5%)	0.35
dsip-mcnc	1370	8.17 (0%)	8.17 (0%)	0
diffeq-mcnc	1497	17.9 (0%)	17.9 (0%)	0
elliptic-mcnc	3604	18.5 (+24%)	22.2 (+3.1%)	9.41
frisc-mcnc	3556	29.4 (0%)	29.4 (+0%)	0
s38417-mcnc	6406	18.3 (+8.5%)	19.9 (−0.2%)	0.64
tseng-mcnc	1047	17.0 (+3%)	16.3 (+7.2%)	1.43
hc11-oc	3877	34.9 (+13%)	36.6 (+8%)	0.52
des-fip	15509	13.0 (+34.7%)	17.7 (−1%)	0.17
sisc8	1434	16.9 (+12.9%)	17.5 (+9.1%)	18.41
clma-mcnc	8383	27.7 (+12.6%)	30.7 (+1.5%)	1.06
ex1010-mcnc	4598	21.9 (+10.1%)	25.4 (−5.3%)	0.33
spla-mcnc	3690	18.7 (+6.9%)	21.9 (−8.7%)	0.24
pdc-mcnc	4575	16.9 (+12.9%)	22.9 (−0.1%)	0.07
Average (%)		+ 10.8	+ 1.2	+ 2.33

11. Summary

This paper has described a physical netlist optimization based on Shannon's decomposition theorem. This technique in concert with incremental placement algorithms described previously produce average speedups of 11% and as much as 30%. As the complete re-placement flow shows, even with complete knowledge of criticalities, the flow of netlist optimization followed by complete placement cannot compete with our incremental approach. In addition, by focusing only on the timing critical logic elements, our average area penalty is extremely small.

References

- [1] Y. Jiang, A. Krstic, K. Cheng, M. Marek-Sadowska, Post-layout logic restructuring for performance optimization, DAC, 1997.
- [2] Y. Lian, Y. Lin, Layout-based logic decomposition for timing optimization, ASPDAC, 1999.
- [3] G. Stenz, B. Riess, B. Rohfleisch, F. Johannes, Timing driven placement in interaction with netlist transformations, ISPD, 1997.
- [4] Cadence, Cadence Physically Knowledgeable Synthesis, (http://www.cadence.com/datasheets/dat_pdf/pks101801.pdf).
- [5] L. Kannan, P. Suaris, H. Fang, A methodology and algorithms for post-placement delay optimization, DAC, 1994.
- [6] Mentor Graphics, Physical Implementation with: TeraPlace, TeraOptimize, TeraCTS, (http://www.mentor.com/teraplace/teraplace_ds.pdf).
- [7] D. Singh, S. Brown, Integrated retiming and placement for FPGAs, FPGA 2002.
- [8] D. Singh, S. Brown, Incremental placement for layout-driven optimizations on FPGAs, ICCAD 2002.
- [9] C. Shannon, The synthesis of two-terminal switching circuits, Bell Syst. Tech. J. 28 (1) (1949).
- [10] J. Cong, Y. Ding, FlowMap: an optimal technology mapping algorithm for delay optimization in lookup-table based FPGA designs, IEEE Trans. CAD, 1994.
- [11] V. Betz, J. Rose, A. Marquardt, Architecture and CAD for Deep-Submicron FPGAs, Kluwer Academic Publishers, Dordrecht, 1999.