Optimal FPGA Mapping and Retiming with Efficient Initial State Computation

Jason Cong and Chang Wu
Department of Computer Science
University of California, Los Angeles, CA 90095

March 15, 1998

Abstract

Retiming is a well known technique for sequential circuit optimization originally proposed by Leiserson and Saxe [LeRS83, LeSa91]. For designs with given initial states, however, new equivalent initial states must be computed for retiming, which unfortunately is NP-hard. In this paper we propose a novel *polynomial time algorithm for optimal FPGA mapping with forward retiming* to minimize the clock period with guaranteed initial state computation. It enables a new methodology of separating forward retiming from backward retiming to avoid time-consuming iterations between retiming and initial state computation. Comparing with the clock period computed by our algorithm, the clock period by conventional approaches of separate mapping followed by retiming [CoDi94, LeSa91] is 20.2% larger, but the clock period by recent approaches of optimal mapping with retiming [PaLi96, PaLi99, CoWu96a] is 2.8% smaller. However, many of the optimal mapping with retiming solutions by [PaLi96, PaLi99, CoWu96a] cannot compute an equivalent initial state by SIS [SeSL92] based on the state-of-the-art algorithm of equivalent initial state computation for retiming in [ToBr93]. Our approach is also applicable to circuits with partial initial state assignment.

1 Introduction

Retiming is a well known technique to reduce the clock period by repositioning flipflops (FFs) in sequential circuits originally proposed by Leiserson and Saxe [LeRS83, LeSa91]. Many studies have been done on combining retiming with logic optimization, circuit partitioning and technology mapping to get better designs for sequential circuits [Mi91, DePR92, ChDP93, MaSB91, ToSS92, MaSB90, MaSB93]. For lookup-table (LUT) based FPGA designs, a significant advancement was made by Pan and Liu [PaLi96, PaLi99] on simultaneous optimal technology mapping with retiming for clock period minimization. Later on, Cong and Wu [CoWu96a] proposed a set of speedup techniques of their algorithm, which led to over 1000 times reduction of runtime while maintaining the optimality of the solution. However, not enough progress has been made on equivalent initial state computation for retiming. For designs with initial states, equivalent initial states need to be computed for retiming, which unfortunately is NP-hard [StSE95].

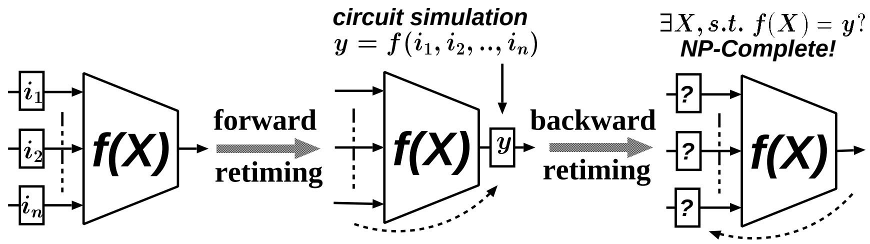

Figure 1: Retiming and initial state computation. Each small rectangle represents a FF.

A general retiming procedure usually involves two kinds of FF movements. *Forward retiming* moves FFs from the inputs of some gates to their outputs, while *backward retiming* moves FFs in the opposite direction. As shown in Figure 1, the equivalent initial state of forward retiming can be computed easily using circuit simulation. The initial state for backward retiming, however, needs to solve a satisfiability problem which is in general NP-complete.

Several heuristics for initial state computation based on ATPG (automatic test pattern generation) techniques [YoKK95, EvSS96, MaSa97] or STG (state transition graph) traversal [ToBr93] have been proposed in recent years. However, none of them guarantees to find an equivalent initial state for a given retiming.

L. Stok, et. al. [EvSS96] proposed to try to exclude backward retiming by minimizing the maximum retiming value. If backward retiming is still needed for a target clock period, ATPG based justification was proposed to compute equivalent initial states with possible iterations of retiming and initial state computation. However, ATPG itself is also NP-complete [FiTo82, KuHN92]. We also observe that minimizing the maximum retiming value as formulated in [EvSS96] does not always lead to a simpler initial state computation problem. Our experience shows that unless a complete forward retiming solution can be found, the effort of initial state computation depends on the number of nodes with backward retiming, instead of the maximum retiming value.

Maheshwari and Sapatnekar [MaSa97] proposed to compute an upper-bound on each node's retiming value based on ATPG techniques such that the equivalent initial state can be computed for a retiming within this bound., Generally the larger values of the upper-bounds, the better the retiming solution. However, to increase one node's retiming upper-bound value may need to reduce another node's retiming upper-bound value and it is hard to trade off the values of different nodes to achieve the best possible solution. As a result, the authors proposed to perform retiming for many sets of upper-bounds and choose the best possible results. Clearly this can still be time-consuming in practice.

Alternatively, Touati and Brayton [ToBr93] proposed an initial state computation algorithm based on STG traversal. Their algorithm can compute equivalent initial states under the condition that there exists an input sequence for the original initial state going back to itself in the STG of the original circuit [ToBr93]. However, unless provided externally, to determine the existence and compute such an input sequence needs to search the STG which may have an exponential number of nodes. Although representing the reset logic explicitly can avoid the STG traversal [ToBr93], the extra cost of the reset logic separated from FFs after retiming can be high. Furthermore, such kind of reset logic may restrict possible retiming of the original circuit, thus reduce the potential of retiming optimization. A recent paper by Singhal et. al. [SiMB96] showed that by retiming the reset logic together with FFs will not affect the possibility of retiming and no extra cost on the area of the circuit. But this approach is valid only for forward retiming, not general retiming.

To better understand the difficulty of the initial state computation problem and the capability of existing initial state computation algorithms, we tested SIS [SeSL92] which computed initial states for retiming based on the algorithm in [ToBr93]. For s9234.1 with 135 FFs and 1293 two-input gates, SIS could not finish the initial state computation for an optimal retiming after running 4 hours on a Sun Ultra2 with 256MB memory. We see that the initial state computation problem being a major obstacle of preventing wide-spread use of retiming technique in practice.

In order to avoid the high complexity of initial state computation for *general* retiming, we propose to perform LUT-based FPGA mapping with forward retiming. Our main contribution is the development of the first polynomial time optimal algorithm for FPGA mapping with forward retiming, which has immediate benefit of guaranteed equivalent initial state computation in linear time. With this algorithm, we can try to push FFs back to PIs as much as possible to enlarge the solution space of mapping with forward retiming, thus achieve better results. One important advantage of our approach is that during the preprocessing of backward retiming we only need to focus on the initial state computation without considering the effect on the clock period. As a result we can still achieve effective retiming for circuit optimization without time-consuming iterations between retiming and initial state computation, avoiding the slow STG traversal procedure or any extra reset logic in existing approaches [MaSa97, YoKK95, StSE95, ToBr93]. Comparing with the clock period computed by our algorithm, the clock period by conventional approaches of separate mapping followed by retiming [CoDi94, LeSa91] is 20.2% larger, but the clock period by recent approaches of optimal mapping with retiming [PaLi96, PaLi99, CoWu96a] is 2.8% smaller. However, many of the optimal mapping with retiming solutions by [PaLi96, PaLi99, CoWu96a] cannot compute an equivalent initial state by SIS [SeSL92] based on the state-of-the-art algorithm of equivalent initial state computation for retiming in [ToBr93]. Our algorithm is also applicable to circuits with partial initial state assignment.

The rest of the paper is organized as follows. Section 2 is the problem formulation and definitions and an overview of our algorithm. Our algorithm is presented in Section 3. The experimental results are presented in Section 4 and the conclusions and our future work in Section 5. All the proofs are in Appendix A.

2 Problem Formulation and Definitions

Given a *sequential* circuit, the technology mapping problem for K-LUT based FPGAs is to construct an *equivalent* circuit consisting of K-LUTs and flipflops (FFs), such that both of the circuits generate the same output sequence for any input sequence, starting from their individual initial states, respectively. We propose to study the following problem in this paper.

Problem 1 *Given a sequential circuit with initial state s_0, find an equivalent LUT circuit with initial state s_0^r and its clock period is minimum under forward retiming.*

As in [PaLi99, CoWu96a], instead of solving the optimization Problem 1 directly, we shall solve its decision version and use binary search to get the minimum clock period.

Problem 2 *Given a sequential circuit with initial state s_0 and target clock period Φ, find an equivalent LUT circuit with initial state s_0^r and the clock period of no more than Φ under forward retiming.*

In this paper, sequential circuits are represented as retiming graphs. The retiming graph $G(V, E, W)$ of a sequential circuit is a directed graph, where V is the set of nodes, E is the set of edges and W is the set of edge weights [LeSa91]. Each node in V represents a gate, a primary input (PI) or a primary output (PO) in the original circuit. Each edge $e(u, v)$ in E represents a directed connection from node u to node v. The *weight* $w(e)$ of an edge e is the number of FFs on the connection represented by the edge. The *path weight* $w(p)$ of a path p is the sum of all edge weights on the path. Under the unit-delay mode[1], *delay* $d(v) = 1$ for every internal node v, or 0 for every PI or PO. *Path delay* $d(p)$ of a path p is the sum of all node delays on the path.

For a target clock period Φ, we define the edge length $length(e)$ to be $d(v) - \Phi \cdot w(e)$ for an edge $e(u, v)$. The path length $length(p)$ is $\sum_{e \in p} length(e)$. In an LUT network M, a node's ***l*-value** is defined to be $l_M(v) = max\{length(p)\}$ for all paths p from PIs to v. It can be proved that,

Theorem 1 *Given an LUT network M and a target clock period Φ, the clock period of M under forward retiming is no more than Φ if and only if the l-values $l_M(v) \leq \Phi$, for every LUT $v \in M$.*

Let M be one solution of mapping with forward retiming, thereafter called *FRT mapping solution.* In M, the output of an LUT rooted at v is generally $r(v)(\geq 0)$ cycles ahead the output of v in the original circuit due to forward retiming. A mapping solution with

[1]We use the unit-delay mode and assume synchronous sequential circuits with single-phase clock and edge-triggered flipflops in this work.

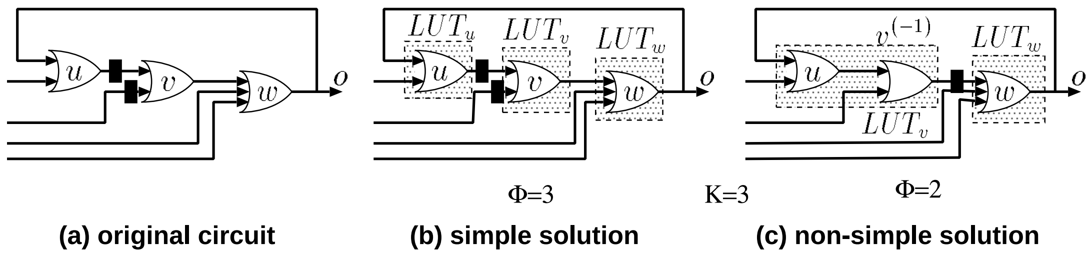

Figure 2: Simple vs. non-simple mapping solutions. Each dotted block represents a 3-LUT.

$r(v) \le 0$ for every LUT root v is called a *simple mapping solution* [PaLi99]. One important property of simple mapping solution is that the path weight of any path from PIs to a node is the same as that in the original circuit. Thus, we only need to focus on reducing the path delay for path length and clock period minimization. It was shown in [PaLi99] that, there always exists a *simple* mapping solution whose clock period will be the minimum after a second step of *general* retiming. However, there may not exist any simple FRT mapping solutions with the minimum clock period after optimal *forward* retiming. For the example shown in Figure 2(a) with $K = 3$, it is not difficult to verify that there does not exist a simple FRT mapping solution with $\Phi = 2$ under forward retiming, while there does exist a non-simple FRT mapping solution with $\Phi = 2$ as shown in Figure 2(c). As a result, the minimum l-values among all possible FRT mapping solutions are much more difficult to compute, because the path weight from PIs to a node will also be changed which makes the monotone property of l-values shown in [CoWu96a] no longer holds. To overcome these difficulties we separate the l-value into two values such that the first one has the monotone property defined in [CoWu96a] and the second one can be computed according to the first one.

Definition 1 *Given a mapping solution M and a target clock period Φ, the number of FFs moved forward across LUT v, denoted $r_M(v)$, is called the* **forward retiming value** *of v in M.*[2] *The* **s-value** $l^s_M(v) = l_M(v) - \Phi \cdot r_M(v)$ *represents the l-value in the corresponding simple mapping solution of M after pushing back the $r_M(v)$ FFs to the inputs of v, where $l_M(v)$ is the l-value of v in M.*

[2]In [LeSa91], the retiming value $\Re(v) \le 0$ corresponds to forward retiming at node v. We define $r_M(v) \ge 0$ for FRT mapping.

According to Theorem 1 we have that,

Corollary 1 *A mapping solution M has a clock period of no more than a given Φ under forward retiming, if and only if $l^s_M(v) + \Phi \cdot r_M(v) \leq \Phi$ for all the LUT roots v in M.*

Definition 2 *Given a circuit and a target clock period Φ, if there exists a FRT mapping solution with clock period of no more than Φ under forward retiming (called a* feasible solution*), the* ***s*-label** *$L^s(v)$ is defined to be the minimum $l^s_M(v)$ among* all feasible FRT mapping solutions *and, the* ***r*-label** *$R(v)$ is defined to be the minimum $r_M(v)$* among all feasible FRT mapping solutions *with $l^s_M(v) = L^s(v)$. The* **node label pair** *is defined to be a 2-tuple $(L^s(v), R(v))$.*

We shall show that $L^s(v)$ has the important monotone property in § 3.2. According to Corollary 1, we can solve the decision Problem 2 by computing the node label pairs and checking if $L^s(v) + \Phi \cdot R(v) \leq \Phi$ holds for every node. After getting the minimum clock period Φ_{min} and the corresponding label pairs, we construct a mapping solution and perform a separate forward retiming step to achieve Φ_{min}.[3] Minimizing $R(v)$ means that we want to push forward as few FFs as possible for each node v to leave the maximum freedom to the subsequent forward retiming step, because those FFs pushed forward can no longer be pushed back. On the other side, $R(v)$ cannot be *too* small because we also want to construct larger LUTs to minimize $L^s(v)$ by pushing forward some FFs. Our algorithm can simultaneously minimize both $L^s(v)$ and $R(v)$ to compute optimal solutions.

3 TurboMap-frt Algorithm

Our algorithm, named TurboMap-frt, computes optimal mapping with forward retiming solutions for synchronous sequential circuits with given initial states to minimize the clock period. It performs binary search on the clock period from 1 to $|V|$.[4] For a given clock period, a procedure named FRTcheck is used to decide whether there exists a feasible

[3]Notice that an optimal forward retiming after mapping can be computed easily based on [LeSa91] by adding linear constraints $\Re(v) \leq 0$ for every node v. In fact, we can compute a forward retiming directly from the node label pairs as shown in § 3.3.

[4]Much smaller upper-bound can be computed by performing depth-optimal mapping (FlowMap [CoDi94]) on combinational subcircuits following by forward retiming.

solution through label computation. An overview of the label computation procedure can be described as follows. First, we assign an initial lower-bound on the value of each node label pair. Then we iteratively update those lower-bounds until all of them converge to the values of node label pairs if there exists a feasible solution, or stop if we conclude there is no feasible solution. Expanded circuits are used to represent all possible LUTs under node duplication and forward retiming for label pair update.

Before presenting the details of our algorithm, we review the definition of K-cuts here. In a directed graph with one sink and one source, a *cut* $(X, \overline{X})$ is a partition of the nodes in the graph such that the sink is in $\overline{X}$ and the source is in X. The *node cut-set* $V(X, \overline{X})$ is the set of nodes in X that are connected directly to nodes in $\overline{X}$. If $\mid V(X, \overline{X}) \mid \leq K$, $(X, \overline{X})$ is called a *K-feasible cut*, or *K-cut* in short, and $\overline{X}$ is called a K-feasible cone. A cut is a *min-cut*, if $\mid V(X, \overline{X}) \mid$ is minimum among all the cuts in the graph.

3.1 LUT Formation

To compute an optimal solution we need to be able to search all possible LUTs for every node under node duplication and forward retiming. Pan and Liu [PaLi99] proposed to search all possible LUTs rooted at a node on expanded circuits of the node. An *expanded circuit* $\mathcal{E}_v$ for a node v is a directed acyclic graph (DAG) rooted at v^0 and constructed from the original circuit with node duplication as shown in Figures 3b – 3e. One important property of $\mathcal{E}_v$ is that for any node, denoted u^w, every path from u^w to the root v^0 passes exactly w FFs. Pan and Liu [PaLi99] showed that *to examine all K-LUTs for a node v, it sufficed to examine all the K-LUTs that can be derived from the K-cuts in a large enough expanded circuit $\mathcal{E}_v^{Kn}$ of v* due to the one-to-one correspondence between K-cuts and K-LUTs.[5]

However, the one-to-one correspondence between K-cuts on $\mathcal{E}_v^{Kn}$ and K-LUTs rooted at v under forward retiming no longer holds. As shown in Figure 3(d), to cluster the nodes in the dotted box to form an LUT we have to move the FF backward from the fanout of b^1 to the fanins of b^1, because the path $i_1^0 \rightarrow a^0 \rightarrow c^0$ does not have any FF to push forward. Thus, such a cluster is not a valid LUT formation for FRT mapping. For FRT mapping, we construct a smaller expanded circuit for every node based on its *maximum*

[5]It was shown that the expanded circuit had no more than $O(Kn^2)$ nodes and $O(K^2n^2)$ edges for a circuit with n gates [PaLi99].

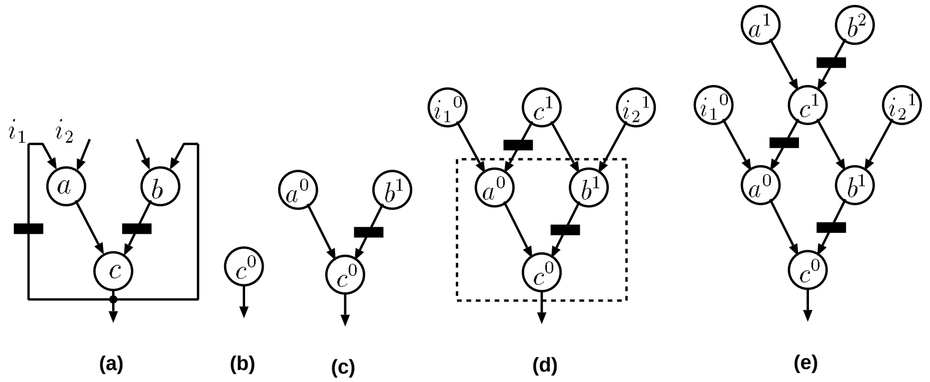

Figure 3: Expanded circuits for LUT representation.

forward retiming value so that there is a one-to-one correspondence between K-LUTs to K-cuts on the expanded circuits.

Definition 3 *The* **maximum forward retiming value** $frt(v)$ *of a node* v *in a retiming graph is the maximum number of FFs which can be moved forward from the (transitive) fanins of* v *to the output of* v.

Lemma 1 $frt(v) = min\{w(p) \mid \forall\ p : PI \rightsquigarrow v\}$.

Since the edge weight $w(e) \geq 0$, the set $\{frt(v)\}$ of all the nodes in a circuit can be computed in $O(|V|^2)$ time by Dijkstra's shortest path algorithm.

Now we define a set of expanded circuits of a node v for FRT mapping. The expanded circuit $\mathcal{F}_v^i$ for a given upper-bound i of forward retiming value of node v is a sub-DAG of $\mathcal{E}_v^{Kn}$ (defined in [PaLi99]) with root v^0 such that, u^w is an internal node of $\mathcal{F}_v^i$ if and only if u^w is an internal node of $\mathcal{E}_v^{Kn}$ and $w \leq i$, and u^w is a leaf of $\mathcal{F}_v^i$ if and only if u^w is either a leaf of $\mathcal{E}_v^{Kn}$ or a fanin of an internal node of $\mathcal{F}_v^i$ and $w > i$. Notice that if two leaves ${u_1}^{w_1}$ and ${u_2}^{w_2}$ have an edge in $\mathcal{E}_v^{Kn}$, the edge will not appear in $\mathcal{F}_v^i$ because a leaf cannot have any inputs. For example, the shaded area of Figure 4(b) represents $\mathcal{F}_c^1$ for the circuit in Figure 4(a), where the heavy shaded nodes are leaves of $\mathcal{F}_c^1$.

When i is set to $frt(v)$, we have that:

Theorem 2 *Every* K*-feasible cut* $(X, \overline{X})$ *on* $\mathcal{F}_v^{frt(v)}$ *corresponds to a* K*-LUT rooted at* v *under node duplication and forward retiming, where all the leaves are in* X *and the root*

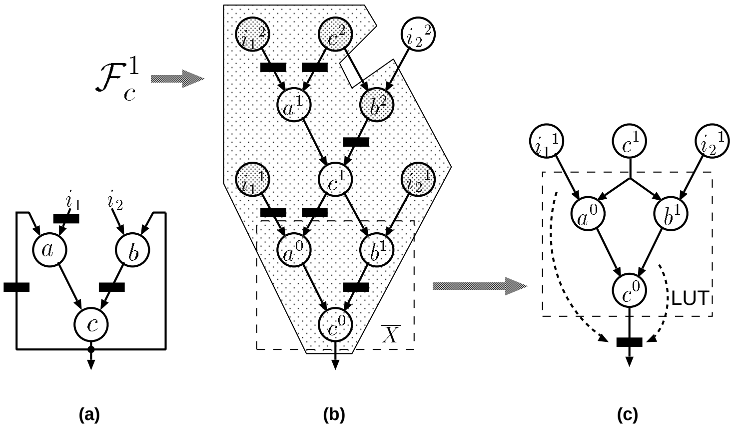

Figure 4: LUT formation for FRT mapping based on the expanded circuits.

is in $\overline{X}$. On the other hand, any K-LUT in a FRT mapping solution corresponds to a K-feasible cut on $\mathcal{F}_v^{frt(v)}$.

Note that we assume no incompatible initial states among fanout FFs of a node. Buffer nodes are needed if such case exists in a circuit. Please see Appendix B for the detail.

To further demonstrate our approach, let us look at the circuit shown in Figure 4(a). Notice that its only difference with the circuit in Figure 3(a) is one extra FF on edge (i_1, a) which makes $frt(c) = 1$. Now the 3-cut shown in Figure 4(b) can form a 3-LUT as shown in Figure 4(c).

3.2 Iterative Label Computation

For a target clock period Φ, we compute all node label pairs of a circuit to decide the existence of a feasible FRT mapping solution. First, we assign an initial lower-bound $(0,0)$ on the value of label pair for every PI and $(-\infty, 0)$ for the other nodes. Then we iteratively update the lower-bounds until they converge to the values of node label pairs. If, however, the lower-bounds cannot converge to finite values after $|V|^2$ iterations, we conclude that there is no feasible FRT mapping solution for the given Φ. The pseudo-code of the FRTcheck algorithm is shown in Figure 5.

The most important part of our algorithm is the update of the lower-bound pair $(l^s(v), r(v))$ for every node v (procedure LabelUpdate, line 6 in Figure 5). For each node v, we update $l^s(v)$ at first and then $r(v)$ through the computation of min-height-min-weight

```
FRTcheck(G(V, E, W), Φ)
assign initial lower-bound (0, 0) for each PI and (−∞, 0) for the other nodes
converge ← FALSE, i ← 0
while (not converge and i ≤ |V|^2)
    converge ← TRUE
    for each node v ∈ V
        (l^s_new(v), r_new(v)) ←LabelUpdate(v, Φ)
        if (l^s_new(v) > l^s(v))
            l^s(v) ← l^s_new(v)
            converge ← FALSE
    i ← i + 1
return (converge? TRUE : FALSE)
```

Figure 5: Label computation for a target clock period Φ.

K-cuts on $\mathcal{F}_v^{frt(v)}$.

Definition 4 *For a K-cut $(X, \overline{X})$ in an expanded circuit of node v, the* **cut-weight** *is $w(X, \overline{X}) = max\{w \mid \forall u^w \in \overline{X}\}$.*

Definition 5 *Given a set of lower-bounds $(l^s(v), r(v))$, where $l^s(v) \leq L^s(v)$ and, $r(v) \leq R(v)$ when $l^s(v) = L^s(v)$, the* **cut-height** *of a K-cut $(X, \overline{X})$ in the expanded circuit $\mathcal{F}_v^{frt(v)}$ of node v is defined to be $h(X, \overline{X}) = \max\{l^s(u) - \Phi \cdot w + 1 \mid \forall u^w \in V(X, \overline{X})\}$.*

To compute a tighter lower-bound $l^s_{new}(v)$ for node v, we compute $\mathcal{L}^s(v) = max\{l^s(u) - \Phi \cdot w(e) \mid \forall e(u, v) \in G\}$ on the original circuit G. Then we decide whether there exists a K-cut on $\mathcal{F}_v^{frt(v)}$ with height of no more than $\mathcal{L}^s(v)$ based on the max-flow K-cut computation on a partial flow network presented in [CoWu96a].[6] If there does not exist such a K-cut, we set $l^s_{new}(v) = \mathcal{L}^s(v) + 1$ and $r_{new}(v) = 0$, because it can then be proved that $L^s(v) > \mathcal{L}^s(v)$.

If, however, there does exist such a K-cut with height of no more than $\mathcal{L}^s(v)$, we then compute a K-cut with the minimum cut-weight among those with height of no more than $\mathcal{L}^s(v)$ by binary search on the cut-weight from 0 to $frt(v)$ as follows. (Recall that $R(v)$ is the minimum $r_M(v)$ among all the feasible FRT mapping solutions M with

[6] We refer readers to [CoWu96a] for the detail.

$l^s_M(v) = L^s(v)$. To minimize $w(X, \overline{X})$ is to guarantee that $r_{new}(v) = w(X, \overline{X}) \leq R(v)$ when $l^s_{new}(v) = L^s(v)$.)

For a given $w \in [0, frt(v)]$, to decide whether there is a K-cut with height of no more than $\mathcal{L}^s(v)$ and weight of no more than w, we construct an expanded circuit $\mathcal{F}^w_v$ and decide whether there exists a K-cut on $\mathcal{F}^w_v$ with height of no more than $\mathcal{L}^s(v)$. If there does exist such a K-cut, clearly it is a K-cut with height of no more than $\mathcal{L}^s(v)$ and cut-weight of no more than w.

Let $(X, \overline{X})$ be such a K-cut with height of $\mathcal{L}^s(v)$ and weight of w_{min} computed above. If $\mathcal{L}^s(v) + \Phi \cdot w_{min} \leq \Phi$, we update the new lower-bound to be $(\mathcal{L}^s(v), w_{min})$. Otherwise, we update the new lower-bound to be $(\mathcal{L}^s(v) + 1, 0)$.

It can be proved that $l^s_{new}(v) \leq L^s(v)$ and, $r_{new}(v) \leq R(v)$ when $l^s_{new}(v) = L^s(v)$. In other words, the updated pair is still a lower-bound on the value of the label pair of v. This result is based on the monotone property of $L^s(v)$. Given a sequential circuit G which has a feasible FRT mapping solution for a target clock period Φ, a set of $L^s(v)$ is *monotone* if, $L^s(u) - \Phi \cdot w(e) \leq L^s(v)$ for any edge $e(u, v) \in G$.

Theorem 3 (Monotone Property) *Given a sequential circuit G which has a feasible FRT mapping solution for target clock period Φ, the set $\{L^s(v) \mid v \in G\}$ is monotone.*

Let one *iteration* denote the computation process where the lower-bound $(l^s(v), r(v))$ is updated once for every node v (lines $5 \sim 9$ in Figure 5). Based on the monotone property we can prove that:

Lemma 2 *Given a sequential circuit which has a feasible FRT mapping solution for a target clock period Φ, the inequality $l^s(v) \leq l^s_{new}(v) \leq L^s(v)$ holds all the time after any number of iterations.*

Lemma 3 *Given a sequential circuit G which has a feasible FRT mapping solution for a target clock period Φ, $r_{new}(v) \leq R(v)$ when $l^s_{new}(v) = L^s(v)$.*

As a result, we have the following theorem:

Theorem 4 *For a sequential circuit which has a feasible FRT mapping solution for a target clock period Φ, starting from the initial lower-bounds $(0, 0)$ for PIs and $(-\infty, 0)$ for the other nodes, the inequalities 1. $l^s(v) \leq L^s(v)$ and, 2. $r(v) \leq R(v)$ when $l^s(v) = L^s(v)$ hold all the time after any number of iterations of label update.*

Based on the above label pair update procedure, we can prove the optimality of FRTcheck algorithm as follows.

Lemma 4 *If there is a feasible FRT mapping solution for a target clock period* Φ*, the 2-tuples* $(l^s(v), r(v))$ *computed by FRTcheck will converge to the node label pairs* $(L^s(v), R(v))$ *for all nodes after no more than* $|V|^2$ *iterations.*

Lemma 5 *For a target clock period* Φ*, if FRTcheck returns TRUE, there must exist a feasible FRT mapping solution.*

Theorem 5 *For a sequential circuit with* n *gates, the optimal FRT mapping solution with the minimum clock period can be computed in the worst-case time complexity of* $O(K^3n^5\log^2 n)$ *and worst-case space requirement of* $O(K^2n^2)$*.*

Notice that the above result is based on the worst-case scenario that the expanded circuits have $O(K^2n^2)$ edges [PaLi99] and we need go through n^2 iterations (lines 5$\sim$9 in Figure 5). In practice, the expanded circuits have much fewer than Kn edges using the efficient max-flow computation on partial flow networks [CoWu96a] and the number of iterations for each Φ is around $5 \sim 15$ based on a computation order proposed in [CoWu96a]. Practically, the runtime of our algorithm is in the order of $K^2n^2\log^2 n$, the space requirement is in the order of Kn. Our experiments show that the optimal solution for s38417 with 9800 gates and 1500 FFs can be computed in 20 minutes on a Sun Ultra2 with 256MB memory.

3.3 Mapping Generation

After computing Φ_{min} based on binary search and obtaining the label pairs $(L^s(v), R(v))$ of all nodes v under Φ_{min}, the last step is to generate the mapping solution based on the K-cuts computed during the label computation and perform forward retiming with initial state computation.

First, we get all the LUT roots in the final mapping solution based on the K-cuts computed during the label computation. In the final mapping solution, all the POs of the original circuit are LUT roots and, if v is an LUT root, all the nodes in the node-cut set of the K-cut of v are LUT roots. Those LUT roots can be computed as follows: starting with a first-in-first-out (FIFO) queue with all the POs, we repeatedly extract nodes from the head of the queue until the queue is empty. For each node extracted from the queue, we mark it as an LUT root and put all the nodes in its node-cut set to the end of the queue.

Second, we create a new equivalent network by connecting the K-feasible cones $\overline{X}_v$ of the K-cuts $(X_v, \overline{X}_v)$ of the LUT roots v. Suppose each LUT has $O(C)$ nodes for a constant $C \ll n$, where n is the number of nodes in the original network, the constructed network has $O(Cn)$ nodes. We compute a forward retiming $\{\Re(v)\}$ on the new network, where $\Re(v) = \lceil \frac{L^s(v)}{\Phi_{min}} \rceil - 1 \leq 0$ for LUT roots v and $\Re(u) = \Re(v) + w$ for $u^w \in \overline{X}_v$ or $\Re(u) = 0$ for $u^w \in V(X_v, \overline{X}_v)$ in each LUT rooted at v. After retiming, all the FFs within each $\overline{X}_v$ will be moved forward outside the cone, because $w^r(u^w \rightsquigarrow v^0) = w + \Re(v) - \Re(u) = 0$ for all $u^w \in \overline{X}$. So we can collapse each $\overline{X}_v$ into a K-LUT and the final circuit has a clock period of no more than Φ_{min} as follows.

Theorem 6 $\{\Re(v)\}$ *is a legal forward retiming to achieve clock period* Φ_{min}.

Since the retiming is a forward retiming, an equivalent initial state can be computed in linear time with circuit simulation based on the approach in [ToBr93]. For a K-bounded circuit with n gates and $O(Kn)$ edges, the first step of getting LUT roots can be done in $O(Kn)$ time. The forward retiming on the constructed circuit can be done in $O(CK^2n^2)$ time since $|\Re(v)| \leq O(Kn)$ for each v and each K-input node can be simulated in $O(K)$ time [ToBr93].[7] So the mapping generation with forward retiming and initial state computation can be finished in $O(CK^2n^2)$ time.

4 Experimental Results

The TurboMap-frt algorithm has been implemented in C language and incorporated into the SIS package [SeSL92]. Our test set consists of 14 MCNC FSMs and 4 ISCAS'89 benchmarks. The initial circuits are shown in Column Original of Table 1. Columns GATE and FF list the numbers of gates and FFs in each circuit, respectively.

Our experiments were performed on a Sun Ultra2 with 256MB memory. K was set to be 5. All the mapping results of TurboMap-frt were computed and verified by *verify_fsm* of SIS [SeSL92], except the 4 largest ones which were verified by simulations with

[7]To be precise, a node can be simulated in $O(l)$ time, where l is the number of variables to represent the function of the node, for example, the number of literals on cover table representation or the number of variables in the BDD representation.

circuit	PI	PO	GATE	FF
bbara	4	2	28	10
bbtas	2	2	15	5
dk16	2	3	162	5
dk17	2	3	42	5
kirkman	12	6	106	5
ex1	8	19	140	5
ex2	2	2	16	7
s1	8	6	107	5
sse	7	7	74	4
keyb	7	2	134	5
styr	9	10	281	5
sand	11	9	327	17
planet1	7	19	348	6
scf	27	54	516	7
s9234.1	28	39	1299	135
s5378	35	49	1503	164
s15850.1	76	148	3801	515
s38417	28	106	9817	1464

Table 1: Initial circuits of MCNC and ISCAS Benchmarks.

input sequences of 3008 random vectors due to huge amount of memory (>1GB) needed by *verify_fsm* for those cases. We compared TurboMap-frt with FlowMap-frt and TurboMap [CoWu96a]. FlowMap-frt is based on the FlowMap algorithm [CoDi94]. It first map each combinational subcircuit bounded by FFs independently with the FlowMap algorithm. Then it merges the mapped LUT subcircuits with the original FFs and performs a postprocessing of forward retiming for clock period minimization. TurboMap [CoWu96a] computes *optimal* mapping with *general* retiming solutions for synchronous sequential circuits. The initial states of TurboMap solutions were computed with SIS [SeSL92] based on the algorithm in [ToBr93]. The experimental results are shown in Table 2, where Columns LUT and FF list the numbers of LUTs and FFs, respectively, in the final mapping solutions by each approach. Columns Φ list the clock periods of the results. Columns CPU list the CPU time. Those marked with $\star$ are examples for which SIS failed to compute equivalent initial states for TurboMap solutions. Column Best lists the best valid solutions (with computed equivalent initial states) by TurboMap and FlowMap-frt.

The results show that comparing with TurboMap-frt, FlowMap-frt computed results with 20.2% larger clock period. Although TurboMap [CoWu96a] computeed results with 2.8% smaller clock period, there are 10 out of 18 solutions by TurboMap SIS concluded no equivalent initial states exist or could not find one due to huge memory requirement

circuit	FlowMap-frt				TurboMap				Best	TurboMap-frt			
	Φ	LUT	FF	CPU	Φ	LUT	FF	CPU	Φ	Φ	LUT	FF	CPU
bbara	4	13	10	0.2	3	12	7	0.4	3	3	12	12	0.2
bbtas	2	7	5	0.1	1	6	4	0.2	1	1	6	4	0.1
dk16	14	101	5	0.9	14	103	14	3.8	14	14	103	9	1.7
dk17	2	10	5	0.2	1	6	3	0.4	1	1	6	3	0.2
ex1	8	83	5	0.7	8	92	21	1.9	8	8	92	20	1.3
ex2	2	9	7	0.2	$\star$1	4	3	0.2	2	1	4	3	0.1
keyb	10	75	5	0.6	10	79	5	1.6	10	10	81	5	1.0
kirkman	6	48	5	0.7	$\star$5	57	24	1.2	6	5	57	14	0.8
planet1	19	213	6	2.0	$\star$19	201	18	12.5	19	19	199	37	5.0
s1	7	58	5	0.5	7	63	11	1.2	7	7	56	6	0.7
sand	16	176	17	1.8	$\star$15	178	30	10.6	16	15	176	12	4.3
scf	14	325	7	2.8	$\star$13	304	20	19.8	14	13	301	27	8.8
sse	7	42	4	0.4	6	45	10	0.9	6	6	44	8	0.5
styr	17	163	5	1.6	$\star$16	168	8	5.2	17	17	168	12	3.2
s5378	4	421	204	7.9	$\star$4	444	301	51.5	4	4	427	261	40.3
s9234.1	6	462	161	8.5	$\star$4	498	217	>7200	6	5	441	203	58.8
s15850.1	10	1240	504	30.3	$\star$8	1161	732	>7200	10	10	1166	621	205.6
s38417	8	3526	1464	561.5	$\star$6	3420	2264	1201.8	8	6	3301	2573	1210.6
geomean	7.0	100	15	1.4	5.6	94	24	7.4	6.3	5.8	92	23	2.8
%	+20.2	+8.4	-33.8	-47.6	-2.8	+1.9	+6.0	+167.0	+8.6	1	1	1	1

Table 2: Comparison of TurboMap-frt with FlowMap-frt and TurboMap for 5-LUT. "geomean" lists the geometric means of the results by each approach. The runtime was recorded on a Sun Ultra2 with 256MB memory. Those marked with $\star$ are circuits that SIS failed to compute initial states for TurboMap solutions. Column Original lists the original circuit size, where N and F represent the numbers of nodes and FFs, respectively, in the original circuits.

(>300MB) and long runtime (>2hours). Even if the best valid solutions by TurboMap or FlowMap-frt are used for comparison, TurboMap-frt still outperforms the two algorithms by 8.6% on the clock periods. On area part, TurboMap-frt used 8.4% and 1.9% less LUTs as compared with FlowMap-frt or TurboMap, respectively. TurboMap-frt also reduced 6% FFs vs. TurboMap. But TurboMap-frt used 33.8% more FFs as compared with FlowMap-frt. In general we think simultaneous mapping with retiming leads to smaller clock period but tends to use more FFs.

5 Conclusions and Future Work

For sequential circuits with initial states, we presented a new algorithm TurboMap-frt for FPGA mapping with forward retiming and initial state computation to minimize the clock period. Unlike previous retiming algorithms which compute retiming at first and then try to compute equivalent initial states, we compute optimal mapping with forward retiming solutions with equivalent initial states in one step. Our algorithm enables a new methodology of performing backward retiming separately before forward retiming. Since we can compute the optimal mapping with forward retiming solutions, backward retiming can be used to push back all the FFs to PIs as much as possible as long as the equivalent initial states can be computed without taking into consideration of the impact on the final clock period. This approach is also applicable to sequential circuits with partial initial state assignments by allowing FFs with unassigned initial state to be moved backward or forward freely with other FFs.

The experimental results show that TurboMap-frt is very efficient and effective comparing with conventional design flow of mapping each combinational subcircuits separately. It can also produce results very close to the optimal mapping with general retiming solutions in terms of both area and clock period.

In the future we plan to extend our work for library based technology mapping with (forward) retiming for high performance gate array and standard cell designs. A general framework on retiming with multiple clock designs was proposed recently by Legl, et. al. [LeVW97]. We plan to accommodate our approach into this framework as well.

6 Acknowledgements

This work is partially supported by National Science Foundation Young Investigator Award MIP9357582 and grants from Xilinx, Lucent Technologies and Quickturn Design Systems under the California MICRO program.

References

[ChDP93] S. T. Chakradhar, S. Dey, M. Potkonjak, and S. G. Rothweiler. Sequential Circuit Delay Optimization Using Global Path Delays. In *30th ACM/IEEE Design Automation Conference*, pages 408–489, 1993.

[CoDi94] J. Cong and Y. Ding. FlowMap: An Optimal Technology Mapping Algorithm for Delay Optimization in Lookup-Table Based FPGA Designs. *IEEE Trans. on Computer-Aided Design of Integrated Circuits And Systems*, 13(1):1–12, 1994.

[CoWu96a] J. Cong and C. Wu. An Improved Algorithm for Performance Optimal Technology Mapping with Retiming in LUT-Based FPGA Design. In *IEEE International Conference on Computer Design*, pages 572–578, 1996.

[DePR92] S. Dey, M. Potkonjak, and S. G. Rothweiler. Performance Optimization of Sequential Circuits by Eliminating Retiming Bottlenecks. In *IEEE International Conference on CAD*, pages 504–509, 1992.

[EvSS96] G. Even, I. Y. Spillinger, and L. Stok. Retiming Revisited and Reversed. *IEEE Trans. on Computer-Aided Design of Integrated Circuits And Systems*, 15(3):348–357, 1996.

[FiTo82] H. Fujiwara and S. Toida. The Complexity of Fault Detection: An Approach to Design for Testability. In *FTCS-12*, pages 101–108, 1982.

[KuHN92] S. Kundu, L. M. Huisman, I. Nair, and V. Iyengar. A Small Test Generator for Large Designs. In *International Test Conference*, pages 30–40, 1992.

[LeRS83] C. E. Leiserson, F. M. Rose, and J. B. Saxe. Optimizing Synchronous Circuitry by Retiming (Preliminary Version). In R. Bryant, editor, *Proceedings of the 3rd Caltech Conference on Very Large Scale Integration*, pages 87–116, Rockville, Maryland, 1983. Caltech, Computer Science Press.

[LeSa91] C. E. Leiserson and J. B. Saxe. Retiming Synchronous Circuitry. *Algorithmica*, 6:5–35, 1991.

[LeVW97] C. Legl, P. Vanbekbergen, and A. Wang. Retiming of Edge-Triggered Circuits with Multiple Clocks and Load Enables. In *International Workshop on Logic Synthesis*, 1997.

[MaSa97] N. Maheshwari and S. S. Sapatnekar. Minimum Area Retiming with Equivalent Initial States. In *IEEE International Conference on CAD*, pages 216–219, 1997.

[MaSB90] S. Malik, K. J. Singh, R. K. Brayton, and A. Sangiovanni-Vincentelli. Performance Optimization of Pipelined Circuits. In *IEEE International Conference on CAD*, pages 410–413, 1990.

[MaSB91] S. Malik, E. M. Sentovich, R. K. Brayton, and A. Sangiovanni-Vincentelli. Retiming and Resynthesis: Optimizing Sequential Networks with Combinational Techniques. *IEEE Trans. on Computer-Aided Design of Integrated Circuits And Systems*, 10(1):74–84, 1991.

[MaSB93] S. Malik, K. J. Singh, R. K. Brayton, and A. Sangiovanni-Vincentelli. Performance Optimization of Pipelined Logic Circuits Using Peripheral Retiming and Resynthesis. *IEEE Trans. on Computer-Aided Design of Integrated Circuits And Systems*, 12(5):568–578, 1993.

[Mi91] G. D. Micheli. Synchronous Logic Synthesis: Algorithms for Cycle-Time Minimization. *IEEE Trans. on Computer-Aided Design of Integrated Circuits And Systems*, 10(1):63–73, 1991.

[PaLi96] P. Pan and C. L. Liu. Optimal Clock Period FPGA Technology Mapping for Sequential Circuits. In *33th ACM/IEEE Design Automation Conference*, pages 720–725, 1996.

[PaLi99] P. Pan and C. L. Liu. Optimal Clock Period FPGA Technology Mapping for Sequential Circuits. *ACM Transactions on Design Automation of Electronic Systems*, 4(1), 1999. http://www.acm.org/todaes/V4N1/L166/paper.ps.gz.

[SeSL92] E. Sentovich, K. Singh, L. Lavagno, C. Moon, R. Murgai, A. Saldanha, H. Savoj, P. Stephan, R. Brayton, and A. Sangiovanni-Vincentelli. *SIS: A System for Sequential Circuit Synthesis.* Electronics Research Laboratory, Memorandum No. UCB/ERL M92/41, 1992.

[SiMB96] V. Singhal, S. Malik, and R. K. Brayton. The Case for Retiming with Explicit Reset Circuitry. In *IEEE International Conference on CAD*, pages 618–625, 1996.

[StSE95] L. Stok, I. Spillinger, and G. Even. Improving Initialization through Reversed Retiming. In *Proc. Euro. Design & Test Conference*, pages 150–154, 1995.

[ToBr93] H. Touati and R. K. Brayton. Computing the Initial States of Retimed Circuits. *IEEE Trans. on Computer-Aided Design of Integrated Circuits And Systems*, 12(1):157–162, 1993.

[ToSS92] H. Touati, N. Shenoy, and A. Sangiovanni-Vincentelli. Retiming for Table-Lookup Field-Programmable Gate Arrays. In *FPGA'92*, pages 89–94, 1992.

[YoKK95] H. Yotsuyanagi, S. Kajihara, and K. Kinoshita. Retiming for Sequential Circuits with a Specified Initial State and Its Application to Testability Enhancement. *IEICE Trans. INF. & SYST.*, E78-D(7):861–867, July 1995.

Appendix A. Proofs of Theorems and Lemmas

Proof of Theorem 1:

($\Rightarrow$) Let $\{\Re(v) \leq 0\}$ be such a forward retiming to achieve Φ. After retiming, every combinational path has delay of no more than Φ. Thus, for any path $p : PI \rightsquigarrow v$, the path length after retiming is

$$length^r(p) = d(p) - \Phi \cdot w^r(p) = d(p) - \Phi \cdot (w(p) + \Re(v)) \leq \Phi.$$

As a result, $l_M(v) = d(p) - \Phi \cdot w(p) \leq d(p) - \Phi \cdot w(p) - \Phi \cdot \Re(v)) \leq \Phi$.

($\Leftarrow$) Since $l_M(v) \leq \Phi$, we perform a forward retiming of $\Re(v) = 0$ if v is a PI or PO and $\Re(v) = \lceil \frac{l_M(v)}{\Phi} \rceil - 1$ for the other nodes. First, we show this is a legal retiming such that $w^r(e) = w(e) + \Re(v) - \Re(u) \geq 0$ for any edge $e(u, v)$. By definition of l-values, we have that $l_M(v) \geq l_M(u) + length(e) = l_M(u) - \Phi \cdot w(e) + 1$. Without lose of generality we assume both u and v are LUTs in the rest of this proof. So the new edge weight after retiming is $w^r(e) = w(e) + (\lceil \frac{l_M(v)}{\Phi} \rceil - 1) - (\lceil \frac{l_M(u)}{\Phi} \rceil - 1) \geq w(e) + \lceil \frac{l_M(u) - \Phi \cdot w(e) + 1}{\Phi} \rceil - \lceil \frac{l_M(u)}{\Phi} \rceil \geq 0$. Second, we show that the new path weight after retiming is $w^r(p) \geq 1$ for any path $p : u \rightsquigarrow v$ with $d(p) > \Phi$. For such a path p, $l_M(u) + length(p) = l_M(u) + d(p) - 1 - \Phi \cdot w(p) \leq l_M(v)$, therefore,

$$\begin{aligned} w^r(p) &= w(p) + \Re(v) - \Re(u) \\ &\geq w(p) + \lceil \frac{l_M(u) + d(p) - 1 - \Phi \cdot w(p)}{\Phi} \rceil - \lceil \frac{l_M(u)}{\Phi} \rceil \\ &\geq \lceil \frac{l_M(u) + d(p) - 1 - \Phi}{\Phi} \rceil + 1 - \lceil \frac{l_M(u)}{\Phi} \rceil \\ &\geq 1. \end{aligned}$$

■

Proof of Lemma 1: First, $frt(v) \leq min\{w(p) \mid \forall\ p : PI \rightsquigarrow v\}$. If otherwise, assume there is a path $p : PI \rightsquigarrow v$ and $w(p) < frt(v)$. After moving $frt(v)$ FFs forward across v, the new path weight $w^r(p) = w(p) - frt(v) < 0$. However, this is not possible.

Let $m_u = min\{w(p) \mid \forall\ p : PI \rightsquigarrow u\}$. We then show that we can push m_v FFs forward across v. We simply need to perform a forward retiming $\{\Re(u) = -m_u \mid \forall u \in G\}$ and show it is legal. For any edge $(u_1, u_2) \in G$,

$$w^r(u_1, u_2) = w(u_1, u_2) + \Re(u_2) - \Re(u_1) = w(u_1, u_2) - w(u_2) + w(u_1) \geq 0.$$

So it is a legal retiming. ∎

Proof of Theorem 2: The one-to-one correspondence between the K-cuts on the expanded circuits $\mathcal{E}_v^{Kn}$ and the K-LUTs has been proved in [PaLi99]. Obviously, each K-cut $(X,\overline{X})$ on $\mathcal{F}_v^{frt(v)}$ corresponds to a K-LUT in FRT mapping because $w(X,\overline{X}) \leq frt(v)$ so that we alway can push all the FFs within $\overline{X}$ forward to the output of v. Now we prove that the K-cut $(X,\overline{X})$ of a K-LUT rooted at v in an FRT mapping solution must be included in $\mathcal{F}_v^{frt(v)}$, or in other words, it must be that $\overline{X} \subset \mathcal{F}_v^{frt(v)}$. Let the cut-weight $w(X,\overline{X}) = max\{w \mid u^w \in \overline{X}\}$. Clearly $w(X,\overline{X}) \leq frt(v)$, because we have to push all the FFs within $\overline{X}$ forward to the fanout of v, so $w \leq frt(v)$ for any $u^w \in \overline{X}$. By definition of $\mathcal{F}_v^{frt(v)}$, any $u^w \in \overline{X}$ is an internal node of $\mathcal{F}_v^{frt(v)}$. On the other hand, any leaf $u^w \in \mathcal{F}_v^{frt(v)}$ cannot be included in $\overline{X}$ because either $w > frt(v)$ or u^w is also a leaf of $\mathcal{E}_v^{Kn}$ such that $u^w \notin \overline{X}$. This concludes our proof. ∎

Proof of Theorem 3: Given a node v of an an edge $e(u,v)$, there must exists a FRT mapping solution M such that $l_M^s(v) = L^s(v)$. The l-values $l_M(v) = l_M^s(v) + \Phi \cdot r_M(v)$ and $l_M(u) = l_M^s(u) + \Phi \cdot r_M(u)$. Let LUT_v be the K-LUT rooted at v in M. The edge weight $w^r(e)$ in M is $w^r(e) = w(e) + r_M(u) - r_M(v)$. There are two cases need to be considered.

Case 1: u is a fanin to LUT_v. Since $l_M(u) - \Phi \cdot w^r(e) + 1 \leq l_M(v)$, we have that

$$l_M^s(u) + \Phi \cdot r_M(u) - \Phi(w(e) + r_M(u) - r_M(v)) + 1 \leq l_M(v) = l_M^s(v) + \Phi \cdot r_M(v).$$

It means that $L^s(u) - \Phi \cdot w(e) \leq l_M^s(u) - \Phi \cdot w(e) < l_M^s(v) = L^s(v)$.

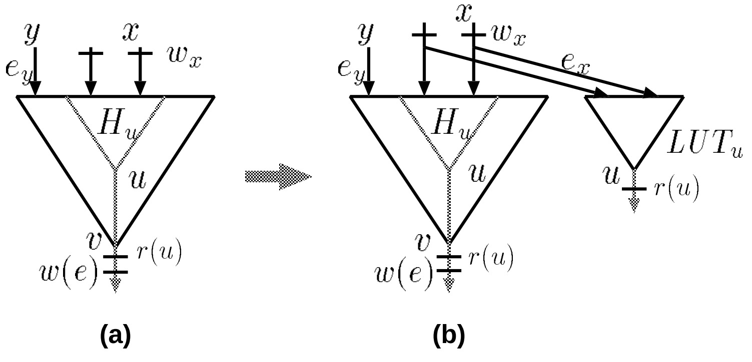

Figure 6: Proof of monotone property of node labels.

Case 2: u is covered inside LUT_v. Let H_u be the largest sub-DAG rooted at u inside LUT_v as shown in Figure 6(a). The $r(u) = R(v) - w(e)$ is the number of FFs moved

across u and $w(\overline{H_u}, H_u) \leq r(u)$. Clearly H_u can be covered by a K-LUT. By replicating H_u explicitly outside LUT_v to form LUT_u and push $r(u)$ FFs forward to u's output, we can form another FRT mapping solution M' with clock period of no more than Φ under forward retiming based on Theorem 1, because
$l_{M'}(u) = max\{l_M(x) - \Phi \cdot w_x + 1 \mid \forall x \in input(LUT_u)\} \leq l_M(v) \leq \Phi$. So

$$
\begin{aligned}
l^s_{M'}(u) = l_{M'}(u) - \Phi \cdot r(u) \leq l_{M'}(v) - \Phi \cdot r(u) &= l_M(v) - \Phi \cdot r(u) \\
&= l^s_M(v) + \Phi \cdot R(v) - \Phi \cdot r(u) \\
&= l^s_M(v) + \Phi \cdot (w(e) + r(u)) - \Phi \cdot r(u) \\
&= l^s_M(v) + \Phi \cdot w(e).
\end{aligned}
$$

As a result, $L^s(u) - \Phi \cdot w(e) \leq l^s_{M'}(u) - \Phi \cdot w(e) \leq l^s_M(v) = L^s(v)$. ■

Proof of Lemma 2: The first part $l^s(v) \leq l^s_{new}(v)$ is obvious based on our FRTcheck algorithm shown in Figure 5. We now prove by mathematical induction that the ineqality of $l^s_{new}(v) \leq L^s(v)$ holds after every label update.

Initially, $l^s(u) = -\infty < L^s(u)$. we prove that if $l^s(u) \leq L^s(u)$ for every node u, $l^s_{new}(v) \leq L^s(v)$ after label update for a node v. Let M be a FRT mapping solution such that $l^s_M(v) = L^s(v)$ and $\overline{X}$ is the LUT rooted at v in M. On M the cut-height $h_M(X, \overline{X}) = L^s_M(v) = L^s(v)$ and the cut-weight $w_M(X, \overline{X}) = R(v) \leq frt(v)$. So $(X, \overline{X})$ is a K-cut on $\mathcal{F}_v^{frt(v)}$ and $h_M(X, \overline{X}) + \Phi \cdot w_M(X, \overline{X}) \leq \Phi$. Since $l^s(u) \leq L^s(u) \leq l^s_M(u)$ for any node u, especially for those $u \in V(X, \overline{X})$, based on the current $(l^s(u), r(u))$, the cut-height $h(X, \overline{X}) \leq h_M(X, \overline{X}) = L^s(v)$. Also $w_M(X, \overline{X}) = w(X, \overline{X})$ is an invariant during label update, so $h(X, \overline{X}) + \Phi \cdot w(X, \overline{X}) \leq h_M(X, \overline{X}) + \Phi \cdot w_M(X, \overline{X}) \leq \Phi$. It means that $l^s_{new}(v) \leq h(X, \overline{X}) \leq L^s(v)$. ■

Proof of Lemma 3: This is obvious based on the proof of the previous lemma. Since there must exist a K-cut $(X, \overline{X})$ such that $h(X, \overline{X}) \leq L^s(v)$, $w(X, \overline{X}) = R(v)$, and $h(X, \overline{X}) + \Phi \cdot w(X, \overline{X}) \leq \Phi$ on $\mathcal{F}_v^{frt(v)}$. If $l^s_{new}(v) = h(X, \overline{X}) = L^s(v)$, it must be that $r_{new}(v) \leq w(X, \overline{X}) = R(v)$, because we always compute a min-height-min-weight K-cut. ■

Proof of Lemma 4: According Lemma 2, $l^s(v)$ will monotonically increase and converge to $L^s(v)$ finally. Furthermore, according Lemma 3, $r(v)$ will converge to $R(v)$ as well when $l^s_{new}(v) = L^s(v)$. Because the value of $r(v)$ is only used to maintain the inequality of

$l^s(v) + \Phi \cdot r(v) \leq \Phi$, it has no impact on the value of $l^s(u)$ of other nodes u. As a result, when $l^s_{new}(v) = L^s(v)$ holds for all the nodes, the equation of $r_{new}(v) = R(v)$ will hold for all the nodes as well. So we only need to prove that $l^s(v)$ will converge to $L^s(v)$ within $|V|^2$ iterations for all the nodes.

According to the theores in [CoWu96a, PaLi99], the lower-bound $l(v)$ computed by TurboMap for every node v, which equals to the height of a min-height K-cut on the expanded circuit of v, will converge to the node label within $|V|^2$ iterations. In our label computation, we have the same initial lower-bounds and a tighter upper-bounds on the node labels because we require $L^s(v) \leq \Phi$ for *every* node v, while TurboMap only requires that for those POs v. Clearly we only need to prove that $l^s_{new}(v) \geq l_{new}(v)$ for every node v, where $l^s_{new}(v)$ and $l_{new}(v)$ are the new lower-bounds computed by TurboMap-frt and TurboMap, respectively.

Since $l^s_{new}(v)$ in TurboMap-frt equals to the cut-height of a min-height-min-weight K-cut, while $l_{new}(v)$ in TurboMap equals to the cut-heigh of a min-height K-cut, it is obvious $l^s_{new}(v) \geq l_{new}(v)$. ■

Proof of Lemma 5: To prove there exists a feasible FRT mapping solution, we only need to construct one and prove it has a clock period of no more than the target Φ. This has been proved in §3.3 and Lemma 6. ■

Proof of Theorem 5: As in [PaLi96, CoWu96a], we assume each edge in the original circuit has at most one FF. For a K-bounded circuit, there are $O(Kn)$ edges and $O(Kn)$ FFs. To update one node's lower-bound, we need $O(log(Kn)) = O(logn)$ K-cut computations. Since the expanded circuits have $O(Kn^2)$ nodes and $O(K^2n^2)$ edges [PaLi96], each K-cut computation can be finished in $O(K^3n^2)$ time. Thus, update one node's lower-bound needs $O(K^3n^2logn)$ time. The entire computation needs n^2 iterations for one target clock period and we need to binary search n possible clock periods. Thus, the total computation time is $O(K^3n^5log^2n)$. All the $frt(v)$ can be computed in $O(n^2)$ time. So the time complexity is $O(K^3n^5log^2n)$ and space requirement is $O(K^2n^2)$ ■

Proof of Lemma 6: First, we show that $|\Re(v)| \geq R(v)$ because we have to push all the FFs within LUT rooted at v to the fanout of v. Since $L^s(v) + \Phi_{min} \cdot R(v) \leq \Phi_{min}$, $\lceil \frac{L^s(v)+\Phi_{min}\cdot R(v)}{\Phi_{min}} \rceil \leq 1 \Rightarrow \lceil \frac{L^s(v)}{\Phi_{min}} \rceil - 1 \leq -R(v) \leq 0 \Rightarrow |\lceil \frac{L^s(v)}{\Phi_{min}} \rceil - 1| = -(\lceil \frac{L^s(v)}{\Phi_{min}} \rceil - 1) \geq R(v)$.

After retiming, every K-feasible cone $\overline{X}$ will become a pure combinational block and can be collapsed into a K-LUT. We prove that the clock period of the LUT network is more than Φ_{min}.

1. The retiming is a legal retiming:

Suppose LUT_u rooted at u is a fanin of LUT_v rooted at v after passing w FFs, $L^s(v) \geq L^s(u) - \Phi_{min} \cdot w + 1$ and $\lceil \frac{L^s(v)}{\Phi_{min}} \rceil \geq \lceil \frac{L^s(u)}{\Phi_{min}} \rceil - w$. As a result, the new edge weight between the two LUTs is
$w^r(LUT_u, LUT_v) = w - \Re(u) + \Re(v) = w - (\lceil \frac{L^s(u)}{\Phi_{min}} \rceil - 1) + (\lceil \frac{L^s(v)}{\Phi_{min}} \rceil - 1) \geq 0$.

2. The circuit has a clock period of no more than Φ_{min}:

For any path $p : u \rightsquigarrow v$ with delay $d(p) > \Phi_{min} + 1$, we prove that $w^r(p) \geq 1$. Since
$L^s(v) \geq L^s(u) + length(p) = L^s(u) - \Phi_{min} \cdot w(p) + d(p) - 1 \Rightarrow$
$\Re(v) \geq \Re(u) - w(p) + \lceil \frac{d(p)-1}{\Phi_{min}} \rceil \geq \Re(u) - w(p) + 1$, we have that $w^r(p) = w(p) + \Re(v) - \Re(u) \geq 1$.

■

Appendix B. Buffer Insertion for Incompatible Fanout Initial States

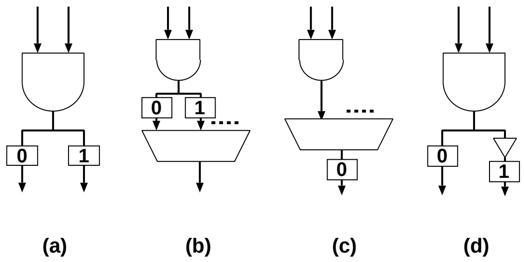

Figure 7: Incompatible initial states of fanout flipflops

In Theorem 2, we assume the FFs on different fanout edges of a node have compatible initial states in TurboMap-frt algorithm. If we number the FFs on an edge (u, v) from u to v. Node u has fanout FFs with *incompatible initial states* if there exists an i such that the initial state of the i-th FF on edge (u, v_1) differs with the initial state of the i-th FF on edge (u, v_2), where $v_1 \neq v_2$. As shown in Figure 7(a), the two fanout FFs of the gate have different initial states. If the outputs of the FFs go into one LUT as shown in Figure 7(b), we may need to count them as two inputs to the LUT. However, if we push the two FFs to the LUT's fanout as shown in Figure 7(c), the two inputs to the LUT can be combined into one. As a result, it is hard to decide the actual input number to the LUT until we get the final solution after mapping and retiming. In this case, we can simply add a buffer before one fanout FF as shown in Figure 7(d). Obviously this kind of change will not affect any path's delay after mapping, thus will not affect the minimum clock period of the FRT mapping solutions. In the worse-case, we only need to add $O(Kn)$ buffers, because each edge needs at most one. After mapping, all thost buffers will be collapse into LUTs without increasing either delay or area as a result of our label computation algorithm.