A MILP formulation for simultanously delay budgeting and retiming for low power FPGA Vdd assignment as a post stage optimization

Yu Hu

November 27, 2005

1 Retiming graph considering interconnect delay

The timing graph of the whole circuit is represented by a edge-weighted DAG. The nodes are the primary inputs, primary outputs, and input/out of the logic gates (i.e. LUTs in our study). Two values are asscociated with an edge $e(u, v)$:

1. $d(e)$: delay from node u to node v.
2. $w(e)$: the number of the FF's inserted in edge e. We can get the initial weight of each edge by the number of FF's in each edge after placement and routing.

The main difference between our retiming graph and the previous ones is that a gate is represented by a set of nodes, instead of a single node. The benifits we can get from such a representation are

1. We can uniformly represent all delay values in edges. Note that we need to consider both node delay and edge delay in the traditional retiming graph.
2. Both single-output and multi-output gates can be treated without difference. Note that we need to do some extension to consider multi-output gates in the traditional retiming graph with node associated with delay, such as in [1].

If we consider the particular case in LUT based FPGA designs, there is one FF in each BLE. We need to add a dump node, FF_NODE, to make sure that all inputs of a LUT share the same number of FF's in the BLE it belongs to. The delay from FF_NODE to $SUBBLK_OPIN$ is zero. Consider the following simple example (see Figure 1). If we connect the inputs of the LUT to the output of the BLE directly, the FF number on each of such edges may be different. Therefore, we add a dump node FF_NODE and constrain that FF can only be inserted at edges from FF_NODE to $SUBBLK\_OPIN$.

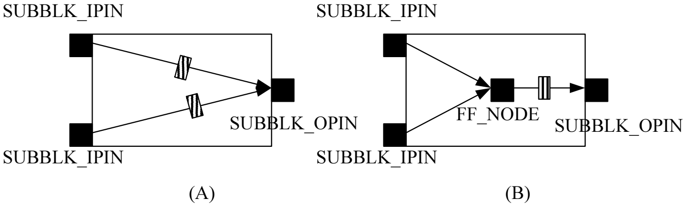

Figure 1: (a)Original representation of timing graph in VPR (b) My extension to consider FF's

2 MILP formulation

2.1 Retiming and delay constraints

The MILP formulation for retiming synchronous circuits is originally prsented in [2] to minimize clock period, and then used in [3] to minimize dynamic power by voltage scaling. Unfortunately, in both of these two works, interconnect delay is not considered in the formulation (only gate delay associated in a node is considered). In [3], the basic idea of time budgeting is used to reduce dynamic power, but the time slacks are not considered in problem formulation explcitily.

To consider the relationship between time slacks and power explicitily, as well as consider both interconnect delay and gate delay in the formulation, I extend the MILP and get the following theorem

Theorem 1 *Let $G = (V, E, d, w)$ be a synchronous circuit, and let c be a positive real number. Then there exists a retiming r of G such that $\Phi(G_r) \leq c$ if and only if there exists an assignment of a real values $s(v)$ and an integer value $r(v)$ to each vertex $v \in V$ such that the following conditions are satified:*

$$-s(v) \leq -\max_{u \in Fanin(v)} d(u,v), \qquad \forall v \in V \tag{1}$$

$$s(v) \leq c, \qquad \forall v \in V \tag{2}$$

$$r(u) - r(v) \leq w(e), \qquad \forall e(u,v) \in E \tag{3}$$

$$s(u) - s(v) \leq -d(u,v), \quad \forall e(u,v) s.t. r(u) - r(v) = w(e) \tag{4}$$

Suppose $R(v) = r(v) + s(v)/c$, then the above retiming constraints can be re-written as

$$r(v) - R(v) \leq -\max_{u \in Fanin(v)} d(u,v)/c, \qquad \forall v \in V \tag{5}$$

$$R(v) - r(v) \leq 1, \qquad \forall v \in V \tag{6}$$

$$r(u) - r(v) \leq w(e), \quad \forall e(u,v) \in E \tag{7}$$

$$R(u) - R(v) \leq w(e) - \max_{u \in Fanin(v)} d(u,v)/c, \quad \forall e(u,v) \in E \tag{8}$$

2.2 FF number constraints

Due to data structure constraint [1] in VPR, we have to the following two more constraints for FF number in edges after retiming

1. FF can only be inserted at edges from FF_NODE to $SUBBLK\_OPIN$.
2. At most one FF can be inserted in edge from FF_NODE to $SUBBLK\_OPIN$.

Based on the definition of retiming, the FF number in edge $e(u,v)$ after retiming is

$$w'(e) = w(e) + r(v) - r(u) \tag{9}$$

We add the following constraints to make sure that only edges from FF_NODE to $SUBBLK\_OPIN$ have FF's inserted after retiming. Suppose E_{ff} is the set including all FF_NODE to $SUBBLK\_OPIN$ edges.

$$-w(e) \leq r(v) - r(u) \leq 1 - w(e), \qquad \forall e(u,v) \in E_{ff} \tag{10}$$

$$r(v) - r(u) = 0, \quad \forall e(u,v) \in E/E_{ff} \tag{11}$$

2.3 Considering FF delay

In VPR, there are two modes of a BLE, i.e. combinational and sequential modes. Figure 2 shows the delay from $SUBBLK\_IPIN$ to FF_NODE and delay from FF_NODE to $SUBBLK\_OPIN$.

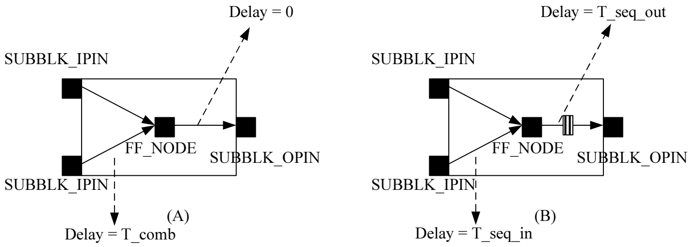

Figure 2: (a)Combinational mode of BLE (b) Sequential mode of BLE

To consider this, I further extend the constraints in subsection 2.1. Suppose set V_{ff} is the node set including all FF_NODE nodes, and V_{subblk_opin} is the node set including all $SUBBLK\_OPIN$ nodes, then we have

[1] FF's can't be accessed independently with LUT's in VPR

$$\begin{aligned}
r(u) - R(u) &\le -[T_{seq\_in}(w(e) + r(v) - r(u)) + \\
&\quad T_{comb}(w(e) + r(u) - r(v))]/c, && \forall u \in V_{ff} \\
r(u) - R(u) &\le -T_{seq\_out}(w(e) + r(v) - r(u))/c, && \forall u \in V_{subblk\_opin} \\
r(v) - R(v) &\le - \max_{u \in Fanin(v)} d(u,v)/c, && \forall v \in V/\{V_{ff}, V_{subblk\_opin}\} \\
R(v) - r(v) &\le 1, && \forall v \in V \\
r(u) - r(v) &\le w(e), && \forall e(u,v) \in E \\
R(u) - R(v) &\le w(e) - \max_{u \in Fanin(v)} d(u,v)/c, && \forall e(u,v) \in E \\
R(u) - R(v) &\le -[T_{seq\_in}(w(e) + r(v) - r(u)) + \\
&\quad T_{comb}(w(e) + r(u) - r(v))]/c, && \forall e(u,v) \in E_{subblk\_ipin->ff} \\
R(u) - R(v) &\le -T_{seq\_out}(w(e) + r(v) - r(u))/c, && \forall e(u,v) \in E_{ff->subblk\_opin}
\end{aligned} \tag{12}$$

2.4 Objective function

Observation 1 *The real value $s(v)$ assigned in node v is actually the arrival time in it.*

Note that this observation makes us enable to consider time slack explicitly in our formulation. The delay slack in edge $e(u,v)$ is

$$b(u,v) = s(v) - s(u) - d(u,v) \tag{13}$$

$$= [R(v) - R(u) + r(u) - r(v)] \cdot c - d(u,v) \tag{14}$$

Therefore the objective function can be expressed as

$$\max \quad \textstyle\sum_{u \in Src} \Delta P(u,v) \cdot b(u,v) \tag{15}$$

$$= \textstyle\sum_{u \in Src} \Delta P(u,v) \cdot [s(v) - s(u) - d(u,v)] \tag{16}$$

$$= \textstyle\sum_{u \in Src} \Delta P(u,v) \cdot \{[R(v) - R(u) + r(u) - r(v)] \cdot c - d(u,v)\} \tag{17}$$

$$\Longleftrightarrow$$

$$\max \quad \textstyle\sum_{u \in Src} \Delta P(u,v) \cdot [R(v) - R(u) + r(u) - r(v)] \tag{18}$$

2.5 MILP formulation

In summery, the MILP formulation for simultanously delay budgeting and retiming for low power FPGA Vdd assignment can be written as

$$\max \quad \sum_{u \in Src} \Delta P(u,v) \cdot [R(v) - R(u) + r(u) - r(v)] \tag{19}$$

$$s.t.$$

$$
\begin{aligned}
r(u) - R(u) \le -[T_{seq\_in}(w(e) + r(v) - r(u)) + & \\
T_{comb}(w(e) + r(u) - r(v))]/c, &\qquad \forall u \in V_{ff} \\
r(u) - R(u) \le -T_{seq\_out}(w(e) + r(v) - r(u))/c, &\qquad \forall u \in V_{subblk\_opin} \\
r(v) - R(v) \le - \max_{u \in Fanin(v)} d(u,v)/c, &\qquad \forall v \in V/\{V_{ff}, V_{subblk\_opin}\} \\
R(v) - r(v) \le 1, &\qquad \forall v \in V \\
r(u) - r(v) \le w(e), &\qquad \forall e(u,v) \in E \\
R(u) - R(v) \le w(e) - \max_{u \in Fanin(v)} d(u,v)/c, &\qquad \forall e(u,v) \in E \\
R(u) - R(v) \le -[T_{seq\_in}(w(e) + r(v) - r(u)) + & \\
T_{comb}(w(e) + r(u) - r(v))]/c, &\qquad \forall e(u,v) \in E_{subblk\_ipin->ff} \\
R(u) - R(v) \le -T_{seq\_out}(w(e) + r(v) - r(u))/c, &\qquad \forall e(u,v) \in E_{ff->subblk\_opin} \\
-w(e) \le r(v) - r(u) \le 1 - w(e), &\qquad \forall e(u,v) \in E_{ff} \\
r(v) - r(u) = 0, &\qquad \forall e(u,v) \in E/E_{ff}
\end{aligned}
$$

References

[1] J. Cong and X. Yuan, "Multilevel global placement with retiming," in *DAC*, Jun 2003.

[2] C.E.Leiserson and J.B.Saxe, "Retiming synchronous circuitry," *Algorithmica*, pp. 5–35, 1991.

[3] N. Chabini and I. Chabini, "Unification of basic retiming and supply voltage scaling to minimize dynamic power consumption for synchronous digital designs," in *GLSVLSI*, Apirl 2003.