## VLSI System Design Part III : Technology Mapping (3) Oct.2006 - Feb.2007

### Lecturer : Tsuyoshi Isshiki

Dept. Communications and Integrated Systems,

Tokyo Institute of Technology

isshiki@vlsi.ss.titech.ac.jp

http://www.vlsi.ss.titech.ac.jp/~isshiki/VLSISystemDesign/top.html

# Timing-Driven Technology Mapping (1)

- 1. Input :
  - Circuit description : Boolean Network (DAG)
  - Timing Constraints : Maximum arrival time from primary inputs to primary outputs (often corresponding to maximum clock period)
- 2. Output :
  - Technology-mapped gate-level netlist satisfying the specified timing constraints with minimum circuit area
- 3. Computation flow
  - Partition the target DAG into trees, leaf-DAGs or cones (call this the *circuit blocks*)
  - At each circuit block : delay-optimal tree covering
  - At each connections between circuit blocks (multiple fan-out nets) : fan-out optimization
  - Nodes on non-critical paths : area recovery
  - Paths with timing violations : Boolean Network restructuring

## Timing-Driven Technology Mapping (2)



## Timing-Driven Technology Mapping (3)



### Timing-Driven Technology Mapping (4)



### Timing-Driven Technology Mapping (5)



### Timing-Driven Technology Mapping (6)



### Timing-Driven Technology Mapping (7)



## Timing-Driven Technology Mapping (8)



# **Buffer Cells for Fan-out Trees**

symbol	cell name	area	gate load	switching delay	output transition coef
	INV0	2	2	12	6
	INV	2	3	12	4
	INVP	3	6	12	2



### **Fan-Out Optimization**

- Fan-out optimization: Construct a fan-out tree which maximize the required time *R<sub>r</sub>* at the net source *r* on following conditions
  - Net source r:
    - Output transition coefficient : T<sub>r</sub> (Switching delay S<sub>r</sub> is not really needed in the optimization)
  - Buffer cell  $b_i$ :
    - Gate load :  $L_{bj}$
    - Switching delay : S<sub>bi</sub>
    - Output transition coefficient : T<sub>bj</sub>
  - Sink i (i = 1, 2, ..., n)
    - Gate load :  $L_i$
    - Required time  $R_i$





 $b_{1} \quad L_{b1} = 2 \\ S_{b1} = 42 \\ T_{b1} = 4$ 

 $b_{2} \qquad \begin{array}{c} L_{b2} = 3 \\ S_{b2} = 48 \\ T_{b2} = 2 \end{array}$ 

# Balanced Fan-Out Tree (1)

- Assumptions :
  - Required times at all sinks are identical. (Ex. clock signals)
  - Fan-out tree is balanced with a height of *M*.
  - Use only one type of buffer cell *b*.
- Compute the optimal number of buffers at each level of the tree. ( $n_k$ : number of buffers at  $k^{\text{th}}$  level,  $n_k > 0$ )
  - The  $n_k$  buffers at  $k^{\text{th}}$  level can be modeled as a single buffer with the gate load  $n_k L_b$ , switching delay  $S_k$  and output transition  $T_b/n_k$



## Balanced Fan-Out Tree (2)



➤ Delay from source to each sink (L<sub>all</sub>: sum of all sink loads):  $D_M = T_r(n_1L_b) + S_b + (T_b/n_1)(n_2L_b) + S_b + (T_b/n_2)(n_3L_b) + \dots + (T_b/n_M) L_{all}$   $= T_r(n_1L_b) + (T_b/n_1)(n_2L_b) + (T_b/n_2)(n_3L_b) + \dots + (T_b/n_M) L_{all} + M \cdot S_b$ 

> Partial derivative on  $D_M$  with respect to each  $n_k$ :

## Balanced Fan-Out Tree (3)

> The number of buffers at *k*-th level  $n_k$  when  $D_M$  is minimum:  $n_1^2 = n_2 T_b / T_r$   $n_2/n_1 = n_3/n_2 = n_4/n_3 = \dots n_M/n_{M-1} = r$  $n_M^2 = n_{M-1}L_{all} / L_b$ 

$$\rightarrow n_{1} = r (T_{b}/T_{r}), n_{M} = (1/r) (L_{all}/L_{b}) \therefore n_{1} n_{M} = (T_{b}/T_{r}) (L_{all}/L_{b}) 
n_{1} = n_{2}/r = n_{3}/r^{2} = \dots = n_{M}/r^{M-1} \therefore n_{M} = n_{1}r^{M-1} 
\rightarrow n_{1} n_{M} = n_{1}^{2} r^{M-1} = (T_{b}/T_{r})^{2} r^{M+1} = (T_{b}/T_{r}) (L_{all}/L_{b}) 
\rightarrow r^{M+1} = (T_{r}/T_{b}) (L_{all}/L_{b}) 
\rightarrow r = (T_{r}L_{all}/T_{b}L_{b})^{1/(M+1)} 
\rightarrow n_{k} = (T_{b}/T_{r}) (T_{r}L_{all}/T_{b}L_{b})^{k/(M+1)}$$

### Two-Level Tree (1)

- Two-level tree : Restrict the tree height to be 1 (M = 1)
  - Delay from source to each sink :

 $D = T_r(n_1L_b) + S_b + (T_b/n_1)L_{all}$ 

- >  $n_1$  when D is minimized :  $n_1 = (T_b L_{all} / T_r L_b)^{1/2}$
- Use only one type of buffer
- ✓ Even with this restricted tree structure, this optimization problem is NPcomplete.
- Two-level tree algorithm :
  - 1. Sort the sinks in the increasing order of their required times (in case of a tie, the decreasing order of the gate load)
  - 2. Set  $n_1 = \lceil (T_b L_{all} / T_r L_b)^{1/2} \rceil$ .
  - 3. Allocate each sink (in the sorted order) to one of the  $n_1$  buffers
    - > Choose the allocation which maximizes the required time at the source node.
  - 4. Compute the two-level tree for each buffer cell type, and choose the fastest.
  - 5. This is a greedy algorithm which do not guarantee optimality, but is a baseline algorithm for other more sophisticated methods.

### Two-Level Tree (2)



Delay :	Required time :
$D_r = T_r * L_{all} + S_r$	$R_r = R_0 - D_r$
= 4 * 28 + 12 = 124	= 132 - 124 = 8



 $D'_{r} = T_{r} * L_{b} * n_{1} + S_{r}$ = 4 \* 3 \* 3 + 12 = 48  $- \boxed{ \sum_{b=2}^{L_b = 3} S_b = 48 }$ 

Optimal number of buffer cells :  $n_{1} = \left[ (T_{b} L_{all} / T_{r} L_{b})^{1/2} \right]$   $= \left[ (56 / 12)^{1/2} \right]$   $= \left[ 2.16 \right] = 3$ 

 $\begin{array}{l} Rmin_{i}: \text{ earliest required time} \\ \text{among the child nodes of buffer } i \\ L'_{i}: \text{ total load connected at buffer } i \\ R'_{i}: \text{ required time at buffer } i \\ R'_{i} = Rmin_{i} - T_{b} * L'_{i} - S_{b} \end{array}$ 

buf	Rmin <sub>i</sub>	L'i	R' <sub>i</sub>
0	132	2	80
1	8	0	8
2	8	0	$\infty$
$MIN \{ R'_{a}, R'_{a} \} = 80$			

#### Two-Level Tree (3)



This is a better allocation

#### Two-Level Tree (4)







buf	Rmin <sub>i</sub>	L' <sub>i</sub>	$R'_i$
0	132	2	80
1	139	4	83
2	142	6	82

 $MIN \{ R'_0, R'_1, R'_2 \} = 80$ 

	buf	Rmin <sub>i</sub>	L' <sub>i</sub>	$R'_i$
	0	132	2	80
	1	139	7	77
	2	142	6	82
1	$MIN \{ R'_{0}, R'_{1}, R'_{2} \} = 77$			

#### Two-Level Tree (5)

 $-\begin{cases} L_b = 3\\ S_b = 48\\ T_b = 2 \end{cases}$ 





buf	Rmin <sub>i</sub>	L' <sub>i</sub>	$R'_i$
0	132	6	72
1	139	7	77
2	142	12	70
$MIN \{ R'_{0}, R'_{1}, R'_{0} \} = 70$			

#### Two-Level Tree (6)



IS THIS OPTIMAL???

### Two-Level Tree (7)

ACTUALLY, there are better solutions....



## Combinational Merging (1)

- Construction of general tree using multiple types of buffer cells
  - Basic idea :
    - Incrementally insert buffer cells and connect the k sink nodes with the largest required times. (This expects that the effect of load reduction due to buffer insertion is larger than the penalty of added delays for these *least critical* sink nodes)
    - k is determined by the two-level tree equation (instead of determining the optimal number of buffer cells, compute the optimal amount of loads the buffer should drive).
    - Inserted buffers become new sink nodes (sink nodes connected to the inserted buffers are no longer sinks to the net source)

## Combinational Merging (2)

- Algorithm
  - 1. Sort the sinks in the increasing order of their required times (in case of a tie, the decreasing order of the gate load)
  - 2. For each buffer cell  $b_{j}$ , compute the optimal number of sinks  $k_{bj}$  (from the tail of the sink list) to be connected to  $b_{j}$ .
    - ♦  $L_{all}$ : total gate loads in the sink list

    - ♦  $L_{all}$  /  $n_{bj}$ : Optimal load per single buffer  $b_j$
    - $\leftarrow L'_k$ : total gate loads of the last *k* nodes in the sink list
    - $\succ$   $k_{bj}$  is the smallest k which satisfies  $L'_k \ge L_{all} / n_{bj}$

## Combinational Merging (3)

- 3. For each cell type  $b_j$ , let  $R(b_j)$  be the required time at the source r where only a single cell of  $b_j$  is connected to r and cell  $b_j$  is connected to the last  $k(=k_{bj})$  nodes in the sink list.
  - $\Rightarrow R(b_j) = R'_k T_{bj}L_k S_{bj} T_r L_{bj}$
  - $\Rightarrow$   $R'_k$ : required time of the *k*-th node from the bottom of the sink list
  - Choose the cell type  $b_j$  which gives the largest  $R(b_j)$  (this will have the largest speed up effect)
- 4. Update sink list :
  - > Insert the cell  $b_i$  to the fan-out tree
  - > Delete the  $k_{bj}$  nodes from the sink list (since they are buffered by  $b_j$ )
  - > Add  $b_i$  to the sink list
    - ♦ Required time at the inserted  $b_j$  cell :  $R(b_j) = R'_k T_{bj}L_k S_{bj}$
  - > If  $k_{bj}$  is less than the total number of nodes in the sink list, go to 1.
- 5. Retrieve the best allocation during the whole process (allocation with the largest required time at the source). *End of process.*

#### Combinational Merging (4)



#### Combinational Merging (5)



_		
R	L	Ľ <sub>k</sub>
138	3	23
155	6	20
158	2	14
160	6	12
186	2	6
208	4	4

$$n_{b1} = (T_{b1} L_{all} / T_r L_{b1})^{1/2}$$
  
= (92 / 8)<sup>1/2</sup>  
= 3.39  
$$L_{all} / n_{b1} = 6.78$$
  
$$K_{b1} = 3, L'_3 = 12, R'_3 = 160$$



$$n_{b2} = (T_{b2} L_{all} / T_r L_{b2})^{1/2}$$
  
= (46 / 12)^{1/2}  
= 1.96  
$$L_{all} / n_{b2} = 11.75$$
$$K_{b2} = 3, L'_3 = 12, R'_3 = 160$$



#### **Combinational Merging (6)**



 $R(b_2) = R'_3 - T_{b2} * L'_3 - S_{b2} - T_r * L_{b2}$ 

= 56

= 138 - 2 \* 11 - 48 - 4 \* 3

$$L_{all} / n_{b2} = 9.17$$
  
 $K_{b2} = 3, L'_3 = 11, R'_3 = 138$ 

#### Combinational Merging (7)



#### Combinational Merging (8)



= -46

#### Combinational Merging (9)



## Timing-Driven Technology Mapping (9)



Primary outputs (output pins, register inputs)

. . . . . .

*If iterative technology mapping cannot resolve timing violations* 

Boolean Network restructuring on paths with timing violations (negative slack times)

Primary inputs (input pins, register outputs)

### **Boolean Network Restructuring**

• There are some cases where delay-optimal technology mapping and fan-out optimization cannot satisfy the specified timing constraints

 $\rightarrow$  <u>need to change the circuit structure</u>

- On the paths with timing violations, collapse a part of the path into a large node (internally represented in sum-of-product form).
- Apply *timing decomposition* to the collapsed node to reduce the critical path delay.



#### **Timing Analysis on Boolean Network**



For simplicity, all gate delays are assumed to be 1 (output transition coefficient is assumed as 0)



#### Node Collapsing

• Select several nodes on the path with the least slack time and merge into one node.



#### Timing Decomposition (1)

 Decompose the collapsed node into 2-input gates (AND, OR) so that the signals with smaller slack times becomes closer to the output.



#### Timing Decomposition (2)

- Bottom-up timing decomposition
  - Timing divisor extraction : Among the algebraic divisors which leads to circuit area reduction, extract the ones which is not on the critical path.
  - AND-OR decomposition : Decompose the division remainder into 2-input AND and OR gates. Put the signals on the critical path closer to the output.



#### Summary on Timing-Driven Technology Mapping and Boolean Network Restructuring

- 1. Each instance of delay-optimal technology mapping is dependent on other mapping results and fan-out tree optimizations. Therefore, several passes may be needed if timing violation occurs. Also, Boolean Network restructuring may be needed if the timing violation cannot be resolved in technology mapping phase (which will require another run of technology mapping).
- 2. There can be a number of possible strategies on the scheduling of each computation phase (technology mapping, fan-out optimization, area recovery, timing decomposition on Boolean Network) which will greatly affect the nature of the mapped circuits as well as the computation time.
- 3. In recent VLSI process technology, wiring delays are becoming a dominant factor in circuit speed compared to gate delays. Wiring delays cannot be accurately estimated until the physical layout of the circuit, and therefore very hard to anticipate during logic synthesis and technology mapping.