Technology Mapping with Choices, Priority Cuts, and Placement-Aware Heuristics

Alan Mishchenko

UC Berkeley

Overview

(1) Introduction
 (2) Technology mapping
 (3) Priority cuts
 (4) Structural choices
 (5) Tuning mapping for placement
 (6) Other applications

(1) Introduction

- Terminology
- And-Inverter Graphs
- Technology mapping in a nutshell

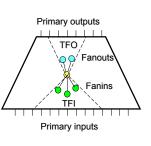
Terminology

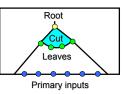
Logic network

- Primary inputs/outputs (PIs/POs)
- Logic nodes
- Fanins/fanouts
- Transitive fanin/fanout cone (TFI/TFO)

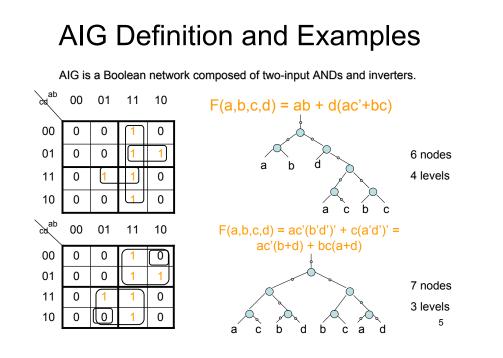
Structural cut of a node

- Cut is a boundary in the network separating the node from the PIs
- Boundary nodes are the leaves
- The node is the root
- K-feasible cut has K or less leaves
- Function of the cut is function of the root in terms of the leaves





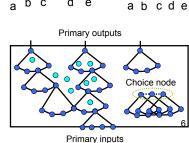
4



Mapping in a Nutshell

 AIGs reprsent logic functions Mapped network AIG - A good subject graph for mapping Technology mapping expresses logic functions to be implemented Uses a description of a technology Technology - Primitives with delay, area, etc \Rightarrow Structural mapping LUT Computes a cover of AIG using primitives of the technology Cut-based structural mapping С d е - Computes cuts for each AIG node - Associates each cut with a primitive Primary outputs Selects a cover with a minimum cost Structural bias Good mapping cannot be found because of the poor AIG structure _ Overcoming structural bias - Need to map over a number of AIG

structures (leads to choice nodes)



LUT

LUT

(2) Technology Mapping

- Traditional LUT mapping
 - Delay-optimal mapping
 - Area recovery
- Drawbacks of the traditional mapping
 - Excessive memory and runtime
 - Structural bias
- Ways to mitigate the drawbacks
 - Priority cuts
 - Structural choices

Traditional LUT Mapping Algorithm

Input: And-Inverter Graph 1. Compute K-feasible cuts for each node 2. Compute best arrival time at each node In topological order (from PI to PO) · Compute the depth of all cuts and choose the best one 3. Perform area recovery · Using area flow · Using exact local area 4. Chose the best cover In reverse topological order (from PO to PI) **Output:** Mapped Netlist

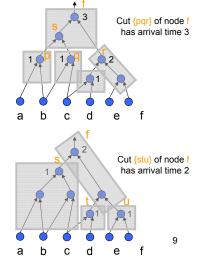
Delay-Optimal Mapping

• Input:

- AIG and K-cuts computed for all nodes
- Algorithm:
 - For all nodes in a topological order
 - Compute arrival time of each cut using fanin arrival times
 - · Select one cut with min arrival time
 - Set the arrival time of the node to be the arrival time of this cut

• Output:

- Delay-optimal mapping for all nodes

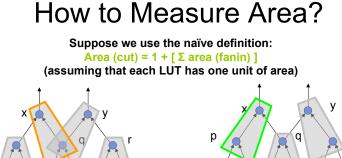


Cut size K = 3

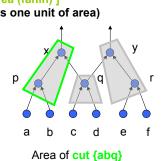
Area Recovery During Mapping

- Delay-optimal mapping is performed first - Best match is assigned at each node - Some nodes are used in the mapping; others are not used Arrival and required times are computed for all AIG nodes - Required time for all used nodes is determined If a node is not used, its required time is set to +infinity Slack is a difference between required time and arrival time If a node has positive slack, its current best match can be updated to reduce the total area of mapping - This process is called area recovery
 - Exact area recovery is exponential in the circuit size
 - A number of area recovery heuristics can be used
 - Heuristic area recovery is iterative Typically involved 3-5 iterations
 - Next, we discuss cost functions used during area recovery
 - They are used to decide what is the best match at each node

10



Ч е Area of cut {pcd} = 1 + [1 + 0 + 0]= 2



= 1 + [0 + 0 + 1]

= 2

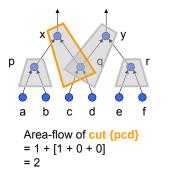
Naïve definition says both cuts are equally good in area

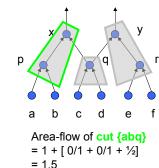
Naïve definition ignores sharing due to multiple fanouts

11

Area-flow

area-flow (cut) = 1 + [Σ (area-flow (fanin) / fanout num(fanin))]



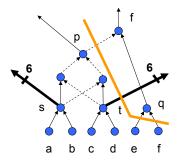


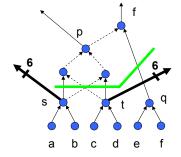
Area-flow recognizes that cut {abg} is better

Area-flow "correctly" accounts for sharing

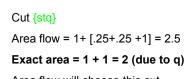
Exact Local Area

Exact-local-area (cut) = 1 + [Σ exact-local-area (fanin with no other fanout)]





Cut {pef} Area flow = 1+ [(.25+.25+3)/2] = 2.75 Exact area = 1 + 0 (p is used elsewhere) Exact area will choose this cut.



Area flow will choose this cut.

Area Recovery Summary

- Area recovery heuristics
 - Area-flow (global view)
 - · Chooses cuts with better logic sharing
 - Exact local area (local view)
 - · Minimizes the number of LUTs by looking one node at a time

• The results of area recovery depends on

- The order of processing nodes
- The order of applying two passes
- The number of iterations
- Implementation details

• This scheme works for the constant-delay model

- Any change off the critical path does not affect critical path

14

Drawbacks of Traditional Mapping

- Excessive memory and runtime requirements
 - Exhaustive cut enumeration leads to many cuts (especially when $K \geq 6)$
- Structural bias
 - The structure of the object graph does not allow good mapping to be found

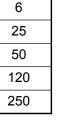
Excessive Memory and Runtime

• For large designs, there may be too many *K*-feasible cuts

- 1M node AIG has ~50M 6-cuts
- Requires ~2GB of storage memory and takes ~30 sec to compute

• Past ways of tackling the problem

- Detect and remove dominated cuts
 Does not help much
- Perform cut pruning (store N cuts/node)
 - Throws away useful cuts even if N = 1000
- Store only cuts on the frontier
 - · Reduces memory but increases runtime



Average

number

of cuts

per node

k

4

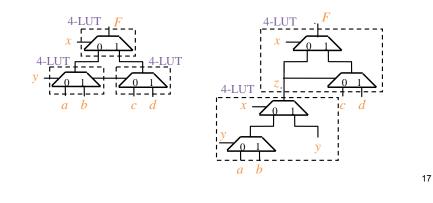
5

6

7

Structural Bias

- Consider mapping 4:1 MUX into 4-LUTs
 - The naïve approach results in 3 LUTs
 - After logic structuring, mapping with 2 LUTs can be found



Ways to Mitigate the Drawbacks

- Excessive memory and runtime requirements
 - Compute only a small number of "useful" cuts
 - · Leads to mapping with priority cuts
- Structural bias
 - Perform mapping over multiple circuit structures
 - Leads to mapping with structural choices

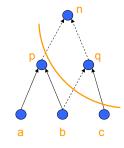
(3) Priority Cuts

- Structural cuts
- Exhaustive cut enumeration
- Prioritizing cuts
- Implementation tricks

Structural Cuts in AIG

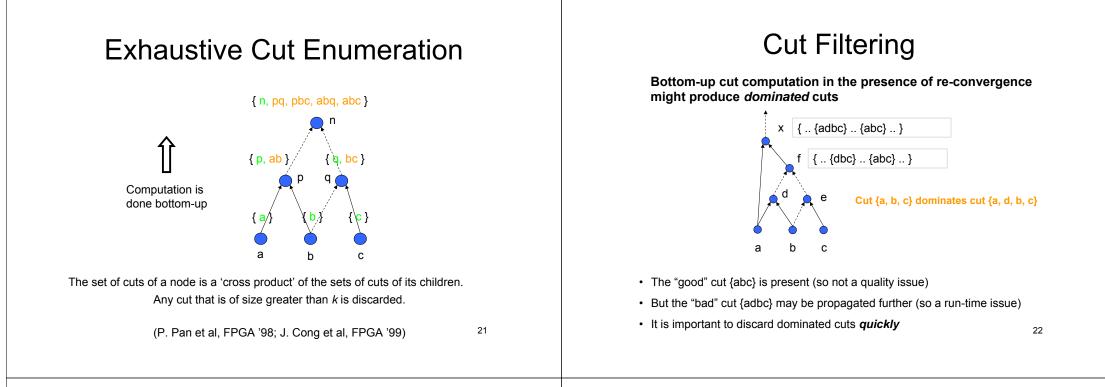
A cut of a node *n* is a set of nodes in transitive fanin such that every path from the node to PIs is blocked by nodes in the cut.

A k-feasible cut has no more than k leaves.



The set {pbc} is a 3-feasible cut of node n. (It is also a 4-feasible cut.)

k-feasible cuts are important in LUT mapping because the logic between root n and the cut leaves {pbc} can be replaced by a *k*-LUT.



Signature-Based Cut Filtering

Problem: Given two cuts, how to quickly determine whether one can be a subset of another.

Solution: Signature of a cut is a 32-bit integer defined as:

sig (c) = $\sum_{n \in C} 2^{ID(n) \mod 32}$ (Σ means bit-wise OR)

where ID(n) is the integer id of node *n*

Observation: If cut c_1 dominates cut c_2 , then

 $sig(c_1) \text{ OR } sig(c_2) = sig(c_2)$

Signature checking is a quick test for the most common case when a cut does not dominate another. Only if this check fails, an actual comparison is performed.

Example

- Let the node IDs be a = 1, b = 2, c = 3, d= 4
- Let c₁ = {a, b, c} and c₂ = {a, d, b, c}
- sig (c_1) = 2¹ OR 2² OR 2³
 - = 0001 OR 0010 OR 0100
 - = 0111
- sig $(c_2) = 2^1$ or 2^4 or 2^2 or 2^3
 - = 0001 OR 1000 OR 0010 OR 0100

= 1111

- As sig (c₁) OR sig (c₂) \neq sig (c₁), c₂ does **not** dominate c₁
- But sig (c₁) OR sig (c₂) = sig (c₂), so c₁ may dominate c₂

Experiment with K-Cut Computation

		<i>K</i> :	= 4	<i>K</i> :	= 5	<i>K</i> :	= 6	<i>K</i> =	= 7		K = 8 C/N T, s 94.7 1.02 139.7 1.90 293.5 4.49 99.7 0.84 306.3 20.99 218.0 4.80 155.9 2.80 86.7 0.44	
Name	AIG	C/N	T, s	C/N	T, s	L/N, %						
alu4	2642	6.7	0.00	12.3	0.01	23.1	0.04	45.5	0.18	94.7	1.02	0.00
apex2	2940	7.2	0.01	14.2	0.02	29.2	0.07	62.6	0.32	139.7	1.90	0.00
apex4	2017	8.5	0.00	19.5	0.03	47.0	0.10	116.3	0.62	293.5	4.49	0.10
bigkey	3080	6.6	0.01	12.1	0.02	24.2	0.05	50.1	0.20	99.7	0.84	0.00
clma	11869	8.1	0.04	18.2	0.11	44.4	0.51	114.9	3.01	306.3	20.99	1.64
des	3020	8.0	0.01	17.0	0.03	38.7	0.12	92.0	0.69	218.0	4.80	4.37
diffeq	2566	6.5	0.01	12.3	0.01	26.6	0.07	65.0	0.50	155.9	2.80	3.66
dsip	2521	6.2	0.01	10.7	0.01	20.7	0.03	42.0	0.10	86.7	0.44	0.00
elliptic	5502	6.4	0.01	10.6	0.03	18.5	0.07	36.9	0.33	83.4	2.12	0.20
ex1010	7652	9.2	0.02	23.3	0.11	61.8	0.61	165.8	4.01	438.2	30.43	1.99
ex5p	1719	9.4	0.01	24.1	0.02	66.2	0.17	188.2	1.30	514.8	10.50	14.14
frisc	5905	7.1	0.01	14.4	0.04	32.3	0.16	79.8	0.88	209.0	6.30	1.24
misex3	2441	7.7	0.01	15.7	0.02	33.3	0.08	73.7	0.38	170.7	2.48	0.00
pdc	7527	9.4	0.03	24.8	0.12	67.4	0.68	183.7	4.41	489.4	31.71	4.40
s298	2514	7.9	0.00	17.5	0.02	44.0	0.13	121.9	0.94	346.5	7.10	7.56
s38417	12867	6.6	0.03	13.5	0.10	32.0	0.46	83.1	3.24	225.9	23.72	3.38
s38584	11074	6.1	0.03	11.4	0.06	22.4	0.20	46.7	0.98	101.5	5.81	0.86
seq	2761	7.5	0.00	15.2	0.02	31.7	0.08	68.6	0.37	153.3	2.25	0.04
spla	6556	9.6	0.03	25.8	0.11	73.9	0.69	215.5	4.98	561.4	31.14	13.83
tseng	1920	6.5	0.01	11.8	0.01	23.5	0.04	50.6	0.21	112.7	1.32	1.35
Average	4954	7.56	0.01	16.22	0.05	38.05	0.22	95.15	1.38	240.0	9.61	2.94

C/N is the number of cuts per node; T is time in seconds; L/N is the ratio of nodes with the number of cuts exceeding the limit (N=1000); for K < 8, the number of cuts did not exceed 1000

Priority Cuts: A Bag of Tricks

- Compute and use priority cuts (a subset of all cuts)
- Dynamically update the cuts in each mapping pass
- Use different sorting criteria in each mapping pass
- Include the best cut from the previous pass into the set of candidate cuts of the current pass
- Consider several depth-oriented mappings to get a good starting point for area recovery
- Use complementary heuristics for area recovery
- Perform cut expansion as part of area recovery
- Use efficient memory management

Computing Priority Cuts

- Consider nodes in a topological order
 - At each node, merge two sets of fanin cuts (each containing C cuts) resulting in (C+1) * (C+1) + 1 cuts
 - Sort these cuts using a given cost function, select C best cuts, and use them for computing priority cuts of the fanouts
 - Select one best cut, and use it to map the node
- Sorting criteria

Mapping pass	Primary metric	Tie-breaker 1	Tie-breaker 2		
depth	depth	cut size	area flow		
area flow	area flow	fanin refs	depth		
exact area	exact area	fanin refs	depth		

The tie-breaking criterion denoted "fanin refs" means "prefer cuts with larger average fanin reference counters".

26

Priority-Cut-Based Mapping

Input: And-Inverter Graph

- 1. Compute K-feasible cuts for each node
- 2. Compute arrival time at each node
 - In topological order (from PI to PO)
 - Compute the depth of all cuts and choose the best one
 - Compute at most C good cuts and choose the best one
- 3. Perform area recovery
 - Using area flow
 - Using exact local area
 - In each iteration, re-compute at most C good cuts and choose the best one
- 4. Chose the best cover
 - In reverse topological order (from PO to PI)
- Output: Mapped Netlist

Complexity Analysis

- The worst-case complexity of traditional mapping
 - FlowMap O(Kmn) (J. Cong et al, TCAD '94)
 - CutMap O(2Kmn^{LKJ}) (J. Cong et al, FPGA '95)
 - DAOmap O(Kn^{LKJ}) (J. Cong et al, ICCAD'04)
- Mapping with priority cuts
 - O(KC²n)

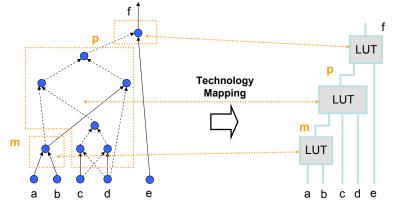
K is max cut size C is max number of cuts n is number of nodes m is number of edges

29

31

Structural Bias

The mapped netlist very closely resembles the subject graph



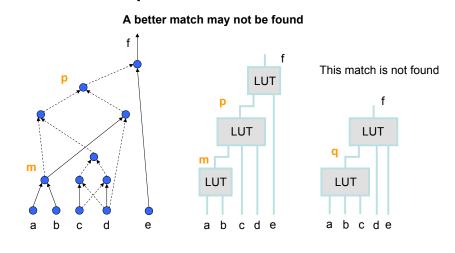
Every input of every LUT in the mapped netlist must be present in the subject graph - otherwise technology mapping will not find the match

(4) Structural Choices

- Structural bias
- · Ways to overcome structural bias
 - Need some form of (re)synthesis to get multiple circuit structures
 - · Computing and using several synthesis snapshots
 - Running several scripts and combining the resulting networks
 - Performing Boolean decomposition during mapping
- Multiple circuit structures = structural choices
- Questions:
 - How to efficiently detect and store structural choices?
 - How to perform technology mapping with structural choices?

30

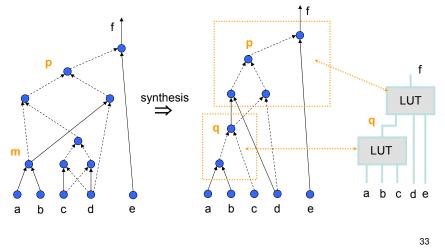
Example of Structural Bias



Since the point **q** is **not** present in the subject graph, the match on the right is **not** found

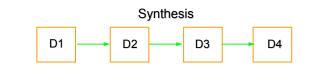
Example of Structural Bias

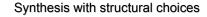
The better match can be found with a different subject graph

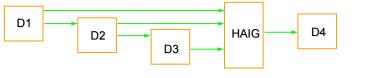


Synthesis for Structural Choices

- Traditional synthesis produces one "optimized" network
- Synthesis with choices produces several networks
 - These can be different snapshot of the same synthesis flow
 - These can be results of synthesizing the design with different options
 - · For example, area-oriented and delay-oriented scripts







Mapping with Structural Choices

- · Two questions have to be answered
 - How to store multiple circuit structures?
 - How to perform mapping with multiple circuit structures?
- Both questions can be solved due to the following:
 - The subject graph is an AIG
 - Structural hashing quickly merges isomorphic circuit structures
 - There are powerful equivalence checking methods
 - They can be used to prove equivalence
 - Cut computation can be extended to work with structural choices
 - The modification is straight-forward

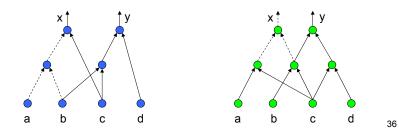
Detecting Choices

Given two Boolean networks, create a network with choices

Network 1					
x = (a + b)c					
y = bcd					

Network 2 x = ac + bc y = bcd 34

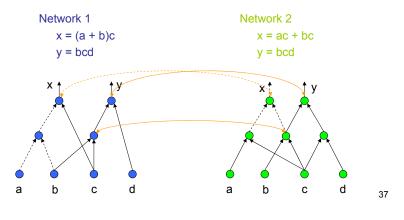
Step 1: Make And-Inverter decomposition of networks



Detecting Choices

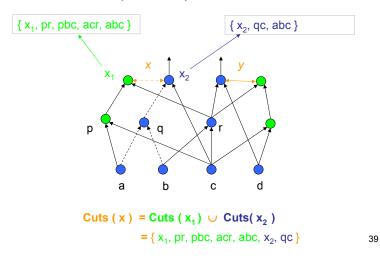
Step 2: Use combinational equivalence to detect functionally equivalent nodes up to complementation (A. Kuehlmann, TCAD'02)

- Random simulation to detect possibly equivalent nodes
- SAT-based decision procedure to prove equivalence



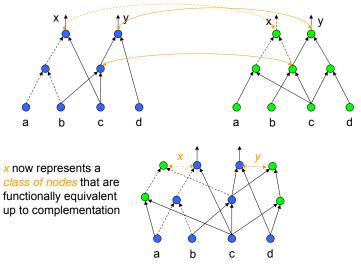
Cut Computation with Choices

Cuts are now computed for equivalence classes of nodes



Detecting Choices

Step 3: Merge equivalent nodes with choice edges



Mapping Algorithm with Choices

Only Step 1 has to be changed

Input: And-Inverter Graph with choices

- 1. Compute K-feasible cuts with choices
- 2. Compute best arrival time at each node
 - In topological order (from PI to PO)
 - · Compute the depth of all cuts and choose the best one
- 3. Perform area recovery
 - · Using area flow
 - Using exact local area
- 4. Chose the best cover
 - In reverse topological order (from PO to PI)
- Output: Mapped Netlist

(5) Tuning Mapping for Placement

- Placement-aware cost function for priority-cut computation
 - The total number of edges in a mapped network
- Advantages
 - Correlates with the total wire-length after placement
 - Easy to take into account during area recovery
- Treat "edges" as "area" resulting in
 - Edge flow (similar to area flow)
 - Exact local edges (similar to exact local area)
- WireMap
 - New placement-aware mapping algorithm

41

WireMap Algorithm

Input: And-Inverter Graph

- 1. Compute K-feasible cuts for each node
- 2. Compute best arrival time at each node
 - In topological order (from PI to PO)
 - Compute the depth of all cuts and choose the best one
- 3. Perform area recovery
 - Using area flow and edge flow
 - Using exact local area and exact local edge
- 4. Chose the best cover
 - In reverse topological order (from PO to PI)

Output: Mapped Netlist

Modified Cut Prioritization Heuristics in WireMap

- Consider nodes in a topological order
 - At each node, merge two sets of fanin cuts (each containing C cuts) getting (C+1) * (C+1) + 1 cuts
 - Sort these cuts using a given cost function, select C best cuts, and use them for computing priority cuts of the fanouts
 - Select one best cut, and use it to map the node
- Sorting criteria

Mapping pass	Primary metric	Tie-breaker 1	Tie-breaker 2		
Depth	depth	cut size	area flow		
area/edge flow	area flow	edge flow	depth		
exact area/edge	exact area	exact edge	depth		

42

Experimental Results

- Experimental comparison
 - WireMap vs. the same mapper w/o edge heuristics
- WireMap leads to the average edge reduction
 9.3% (while maintaining depth and LUT count)
- Place-and-route after WireMap leads to
 - 8.5% reduction in the total wire length
 - 6.0% reduction in minimum channel width
 - 2.3% reduction in critical path delay
- · Changes in the LUT size distribution
 - The ratio of 5- and 6-LUTs in a typical design is reduced
 - The ratio of 2-, 3-, and 4-LUTs is increased
- Changes after LUT merging
 - 9.4% reduction in dual-output LUTs

(6) Other Applications of Priority-Cut-Based Mapping

- Sequential mapping (mapping + retiming)
- Speeding up SAT solving
- Cut sweeping
- Delay-oriented resynthesis for sequential circuits

Sequential Mapping

- That is, combinational mapping and retiming combined
 - Minimizes clock period in the combined solution space
 - Previous work:
 - Pan et al, FPGA'98
 - Cong et al, TCAD'98
- Our contribution: dividing sequential mapping into steps
 - Finding the best clock period via sequential arrival time computation (Pan et al, FPGA'98)
 - Running combinational mapping with the resulting arrival/required times of the register outputs/inputs
 - Performing final retiming to bring the circuit to the best clock period computed in Step 1

46

Sequential Mapping (continued)

- Advantages
 - Uses priority cuts (L=1) for computing sequential arrival times
 very fast
 - Reuses efficient area recovery available in combinational mapping
 almost no degradation in LUT count and register count
 - Greatly simplifies implementation
 - due to not computing sequential cuts (cuts crossing register boundary)
- Quality of results
 - Leads to ~15% better quality compared to comb. mapping + retiming
 due to searching the combined search space
 - Achieves almost the same (-1%) clock period as the general sequential mapping with sequential cuts
 - due to using transparent register boundary without sequential cuts

Speeding Up SAT Solving

- · Perform technology mapping into K-LUTs for area
 - Define area as the number of CNF clauses needed to represent the Boolean function of the cut
 - Run several iterations of area recovery
- Reduces the number of CNF clauses by ~50%
 - Compared to a good circuit-to-CNF translation (M. Velev)
- Improves SAT solver runtime by 3-10x
 - Experimental results are in the SAT'07 paper

Cut Sweeping

• Reduce the circuit by detecting and merging shallow equivalences (proposed by Niklas Een)

– By "shallow" equivalences, we mean equivalent points, A and B, for which there exists a K-cut C (K < 16) such that $F_A(C) = F_B(C)$

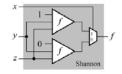
- A subset of "good" K-cuts can be computed
- The cost function is the average fanout count of cut leaves
 The more fanouts, the more likely the cut is common for two nodes
- · Cut sweeping quickly reduces the circuit
 - Typically ~50% gain of SAT sweeping (fraiging)
- Cut sweeping is much faster than SAT sweeping
 - Typically 10-100x, for large designs
- Can be used as a fast preprocessing to (or a low-cost substitute for) SAT sweeping

49

Sequential Resynthesis for Delay

- Restructure logic along the tightest sequential loops to reduce delay after retiming (Soviani/Edwards, TCAD'07)
 - Similar to sequential mapping
 - Computes seq. arrival times for the circuit
 - Uses the current logic structure, as well as logic structure, transformed using Shannon expansion w.r.t. the latest variables
 - Accepts transforms leading to delay reduction
 - In the end, retimes to the best clock period
- The improvement is 7-60% in delay with 1-12% area degradation (ISCAS circuits)
- This algorithm could benefit from the use of priority cuts

xyzUnchanged



50

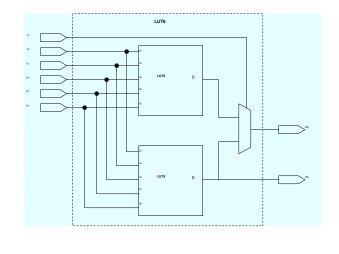
Summary

- · Reviewed traditional and novel LUT mapping
- Presented the current mapping solution
 - Starts with an optimized AIG (with choices)
 - Performs exhaustive (or priority) cut computation
 - Performs heuristic area recovery
 - Uses placement-aware heuristics
- Experimental results are promising
- Future work
 - Area- and delay-oriented resynthesis for mapped networks
 - Using delay information from preliminary placement

Backup Slides on WireMap

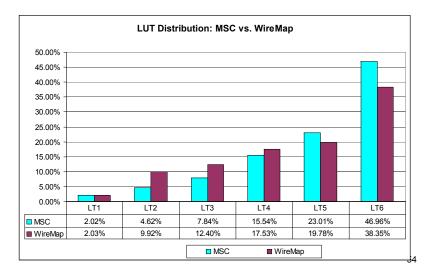
- Virtex-5 dual-output LUT
- Comparison of LUT distribution
- Comparison of area flow and edge flow mapping (K = 6)
- Wirelength, channel width, and critical path delay comparison

Virtex-5 Dual-Output LUT



53

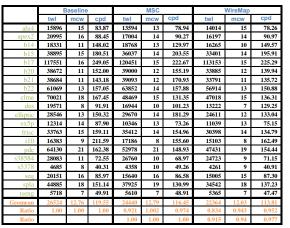
Comparison of LUT Distribution



Comparison of Area Flow and Edge Flow Mapping (K = 6)

			Baseline			Mapping with Structural Choices					WireMap				
	luts	lev	edg	t,s	clb	luts	lev	edg	t,s	clb	luts	lev	edg	t,s	clb
alu4	807	6	3927	0.6	652	742	5	3520	6.79	585	742	5	3298	7.25	550
apex2	983	6	4664	0.75	778	807	6	3850	10.73	654	805	6	3574	11.11	603
b14	1214	13	5620	1.94	976	1162	13	5578	61.36	935	1163	13	5014	53.7	823
b15	2169	15	11073	2.25	1856	2103	15	10485	61.35	1804	2056	14	9499	51.21	1626
b17	6507	21	33151	6.73	5625	6480	18	32906	191.55	5602	6419	18	29552	169.62	5090
b20	2490	15	11953	3.95	2024	2380	14	11768	138.27	1980	2312	14	10582	118.02	1760
b21	2569	15	12418	4.15	2098	2391	14	11807	135.8	1995	2399	14	10781	116.26	1815
b22	3742	15	18027	5.89	3074	3613	14	17910	187.25	3053	3618	14	16426	180.92	2787
clma	3310	10	15576	2.72	2585	2392	9	11520	44.55	1952	2478	9	10846	42.61	1833
des	681	5	3541	0.94	624	502	4	2643	14.39	473	498	3	2192	15.48	370
ex5p	624	5	3019	0.6	497	562	4	2716	10.46	450	578	4	2666	10.49	437
elliptic	1800	10	8777	1	1662	1859	10	9173	17.95	1682	1807	9	8362	18.86	1569
frisc	1750	14	8610	1.35	1621	1798	12	8753	28.02	1668	1690	12	7662	24.37	1524
i10	629	9	2863	0.59	470	603	7	2765	9.81	465	574	8	2375	8.75	404
pdc	2305	7	11307	2.4	1923	2012	6	10061	77.7	1695	1891	6	8795	73.69	1476
s38584	2740	6	11574	1.63	1996	2667	6	11219	17.32	1948	2648	6	10580	17.54	1849
s5378	392	4	1553	0.3	253	357	4	1469	2.66	248	359	4	1346	2.69	226
seq	933	5	4521	0.67	750	737	5	3577	11.69	599	732	5	3276	11.21	551
spla	1862	6	9062	2.06	1538	1588	6	8065	52.58	1350	1515	6	7013	49.14	1198
tseng	657	7	2546	0.46	455	645	7	2488	5.18	452	645	6	2343	5.41	423
Geomean	1480	8.65	6988	2.05	1194	1346	7.97	6420	54.27	1102	1329	7.78	5824	49.42	998
Ratio	1	1	1	1	1	0.909	0.921	0.919	26.473	0.923	0.898	0.899	0.833	24.107	0.836
Ratio						1	1	1	1	1	0.987	0.976	0.907	0.911	0.906

Wirelength, Channel Width, and Critical Path Delay Comparison



twl = total wire length, mcw = minimum channel width required to route in VPR, cpd = critical path delay with min channel width across the three implementations ⁵⁶