Fault-Tolerant Resynthesis for Dual-Output LUTs

Roy Lee[1], Yu Hu[1], Rupak Majumdar[2], Lei He[1] and Minming Li[3]

[1]Electrical Engineering Dept., UCLA

[2]Computer Science Dept., UCLA

[3]Computer Science Dept., City University of Hong Kong

Address comments to: Dr. Lei He (lhe@ee.ucla.edu)

Outline

- Background and Problem Formulation
- Algorithms
- Experimental Results
- Conclusion and Future Work

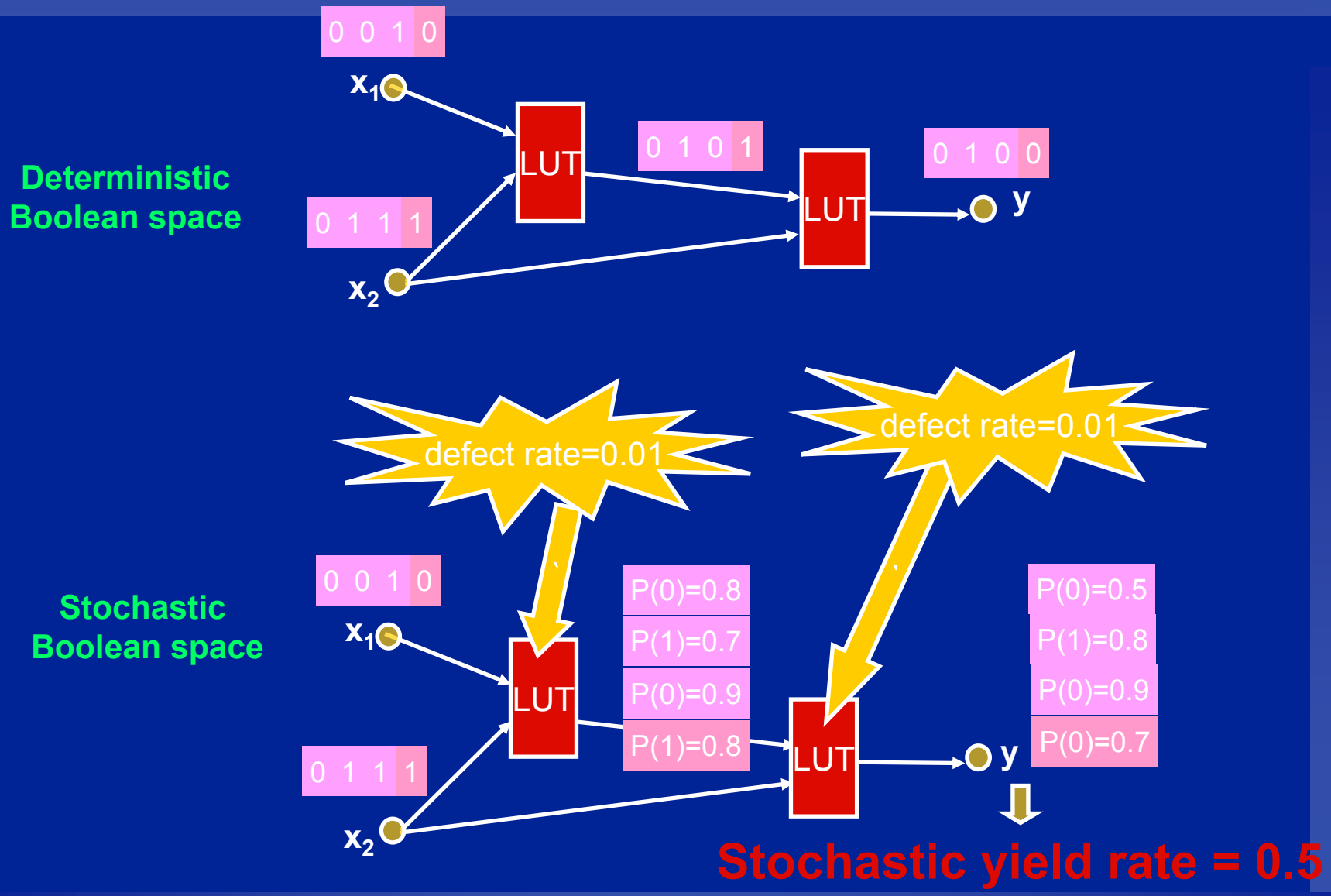
Motivation

- Same as CPU and ASIC, FPGA is susceptible to soft errors
  - Permanent ones by periodically scrubs
  - Transient ones by TMR
- TMR has 5-6x area/power overhead
  - Unbearably expensive for non mission-critical applications such as internet routers
- There does not exist a selected TMR flow to obtain desired MTBF with minimal area/power overhead

Recent Research for SER

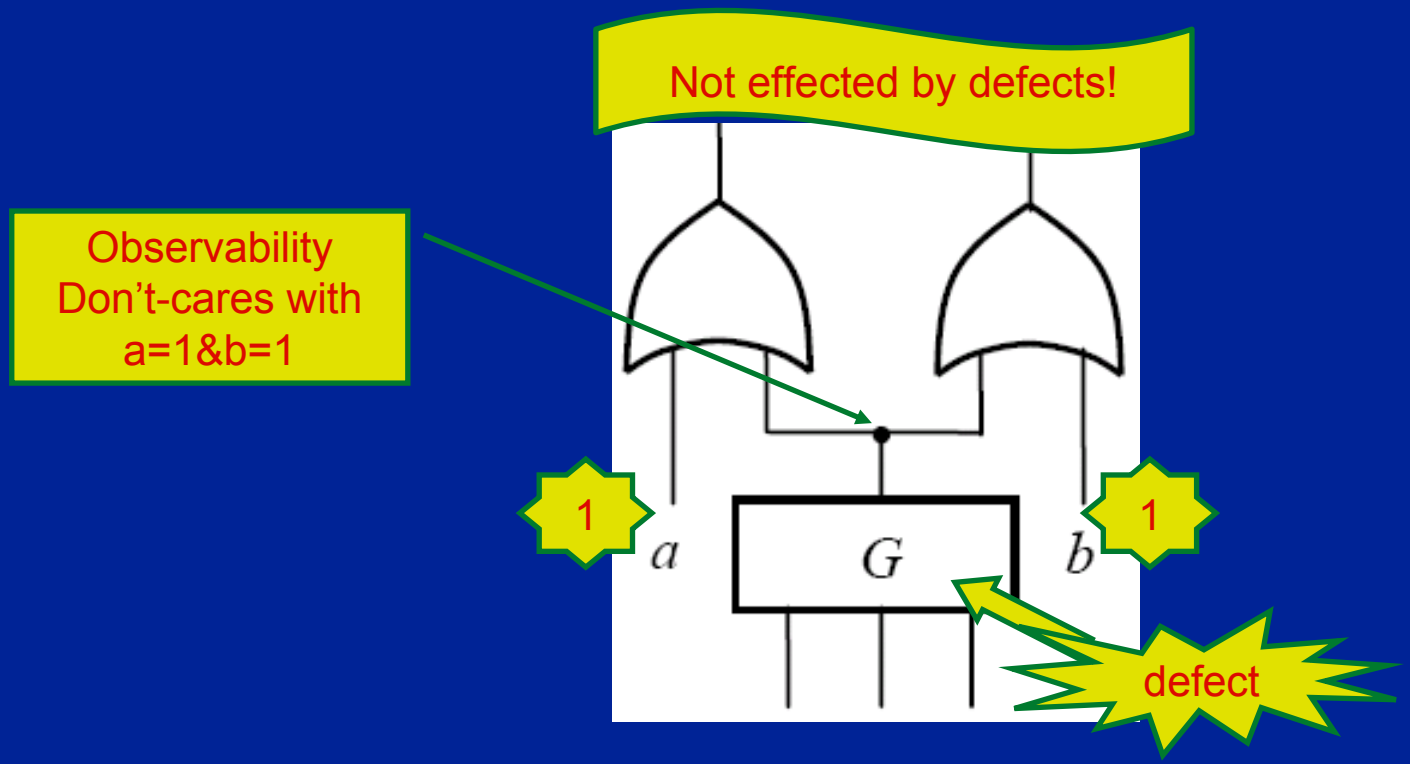
- SEU (MCU) aware FPGA routing
  - [Bozorgzadeh, DAC'07]
- Device and architecture co-optimized for SER
  - [ICCAD'07][ISFPGA'08]
- Logic synthesis for MTBF optimization
  - [Best paper nomination, ICCAD'08]

Deterministic vs. Stochastic

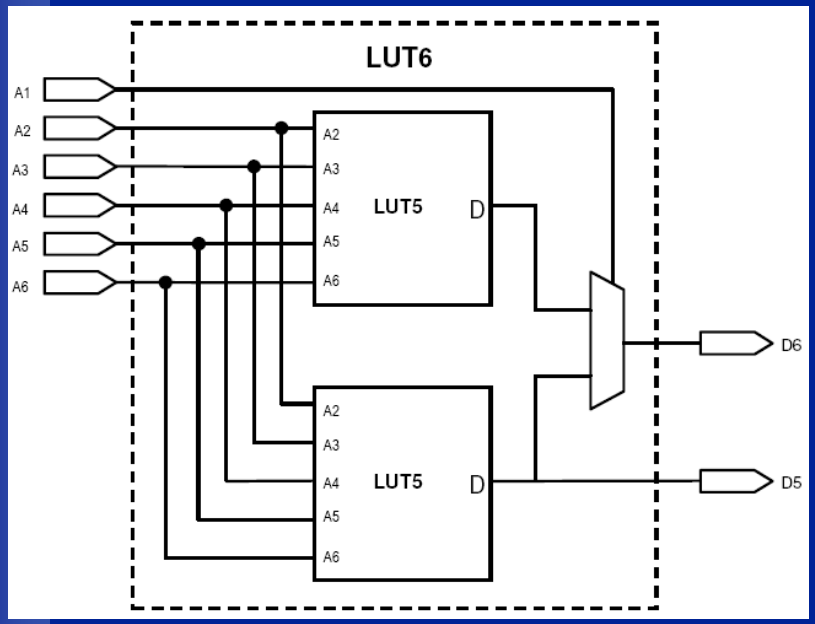


Key Design Freedom: Logic Masking

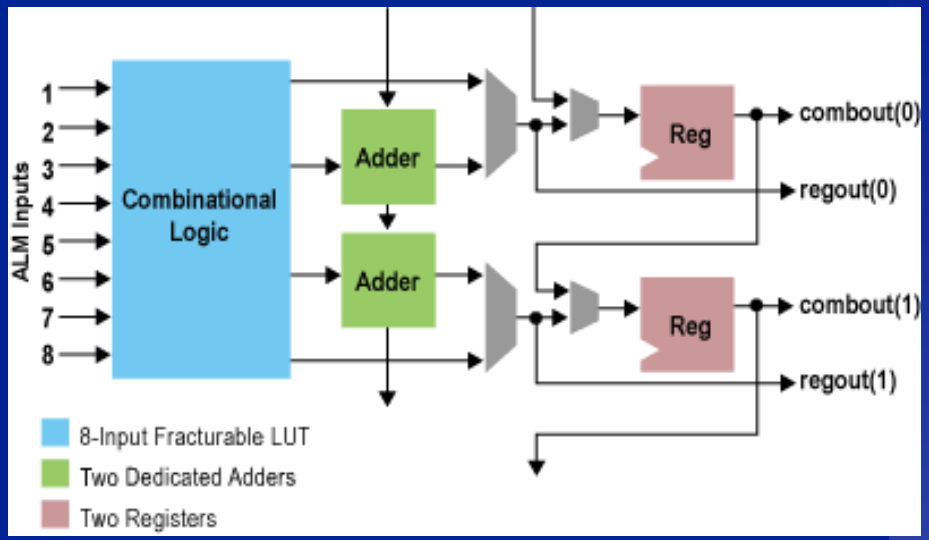
- SEUs are created equally but not propagated equally
- Stochastic logic synthesis increases MTBF by 30% without area/power/delay overhead [ICCAD'08]



More Opportunity in Modern FPGAs



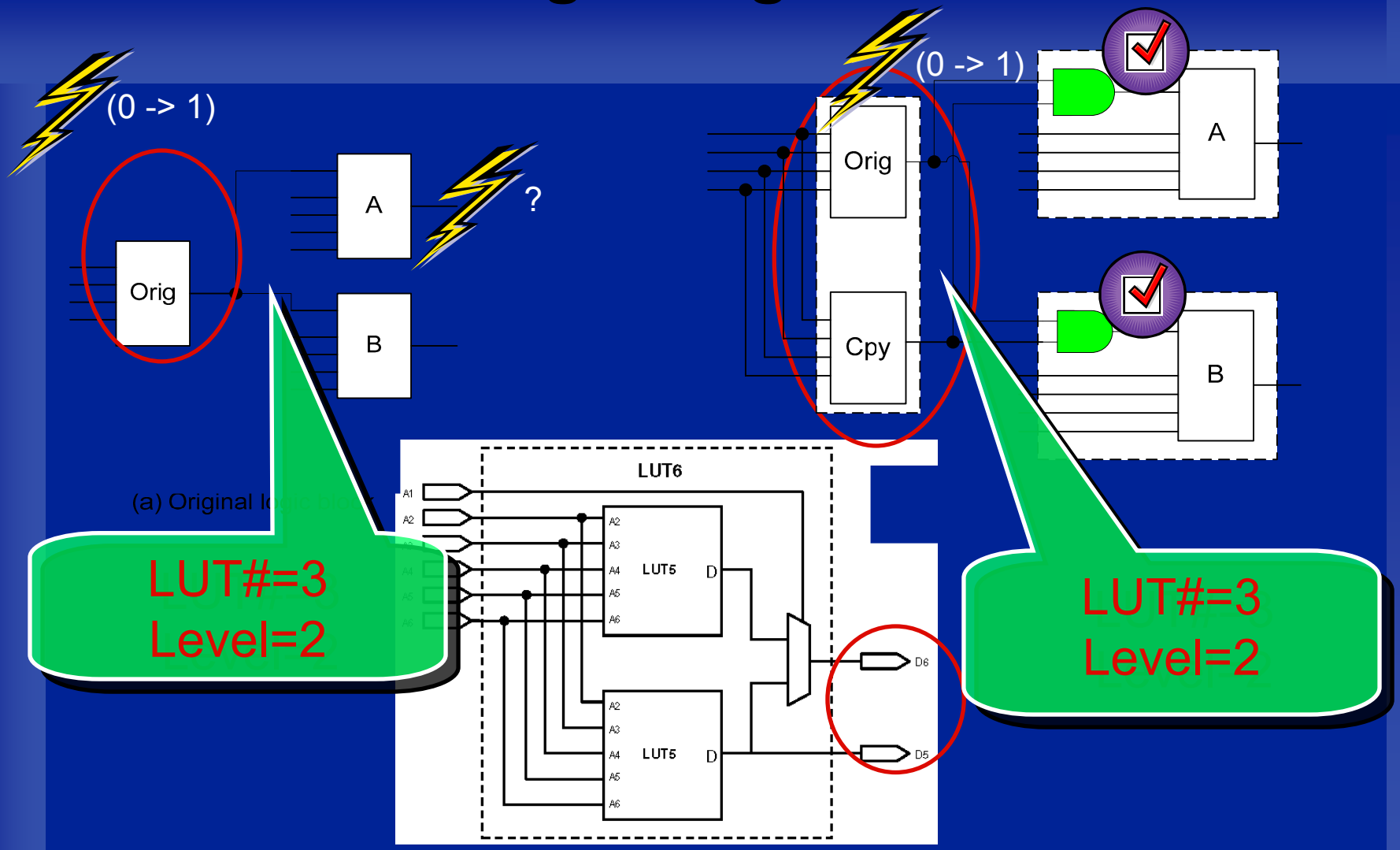
Xilinx Virtex-5 LUT



Altera Stratix III ALM

- **Dual-output (DO) LUT**
  - Merging two small LUTs into one dual-output LUT
  - **Originally designed to increase the logic density**

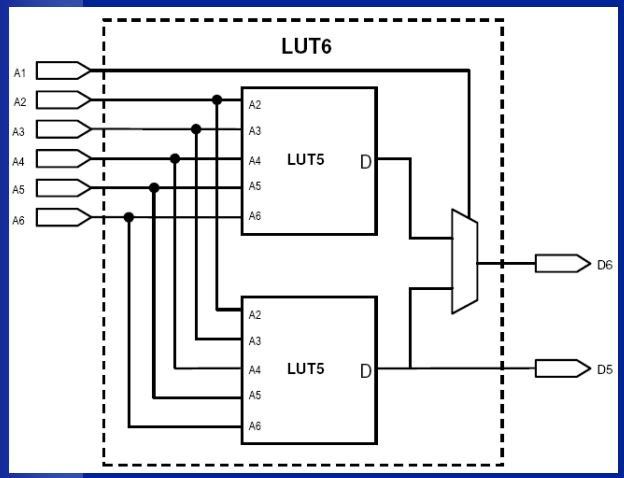
Fault Masking Using DO LUTs

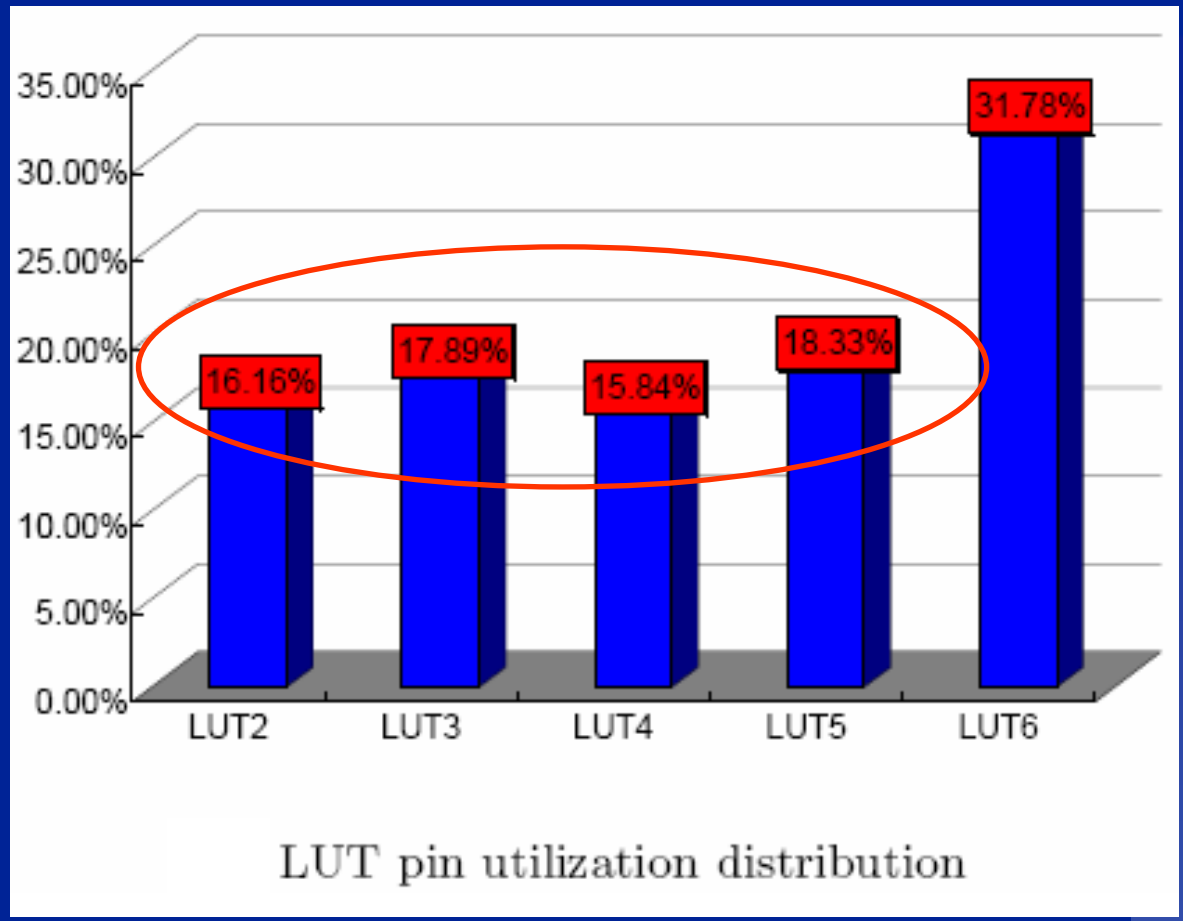


Xilinx Virtex-5 dual-output LUT

- **One spare pin is needed for duplication and encoding, respectively**

Potential of the Optimization using DO LUTs





- Pin utilization rate is low with state-of-art logic synthesis

Problem Formulation

Fault-Tolerant using Dual-output LUTs

- Given:
  - a mapped circuit for K-input 2-output LUTs,
- Design freedom:
  - perform duplication and encoding
- Optimization objective:
  - the full-chip fault rate is minimized.

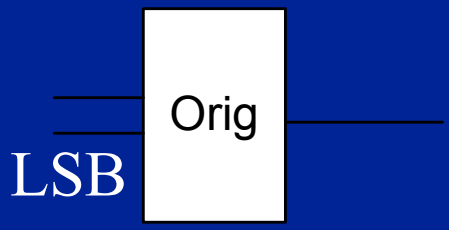
Two approaches

- Fully masking (FMD): encoding all fanouts
- Partial masking (PMD): encoding part of fanouts

Fault Modeling

- Assume a *stochastic single fault* model
  - At most one fault occurring at a time
  - A fault with identical random distribution for each SRAM bit
- The *criticality of a SRAM bit*
  - Combination of observability and signal probability
  - Measured by the percentage of input vectors that cause observable output errors if a fault occurs in this SRAM bit.
- The *fault rate of a full-chip* is the percentage of input vectors that cause observable output errors assuming the single fault.
  - Fault rate is the average criticality of all SRAM bits

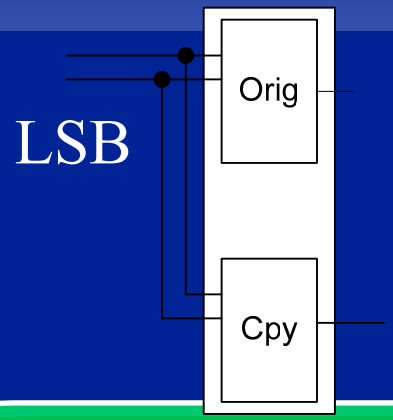
FMD Impact on SRAM Criticality



Average Crit. = 0.125

Input	Output	Crit.
000	0	0.2
001	1	0.2
010	0	0.4
011	1	0.2
100	not used	0
101	not used	0
110	not used	0
111	not used	0

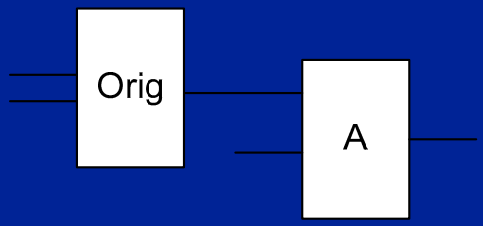
Duplication



Average Crit. = 0.25

Input	Output	Crit.
000	0	0.2
001	1	0.2
010	0	0.4
011	1	0.2
100	0	0.2
101	1	0.2
110	0	0.4
111	1	0.2

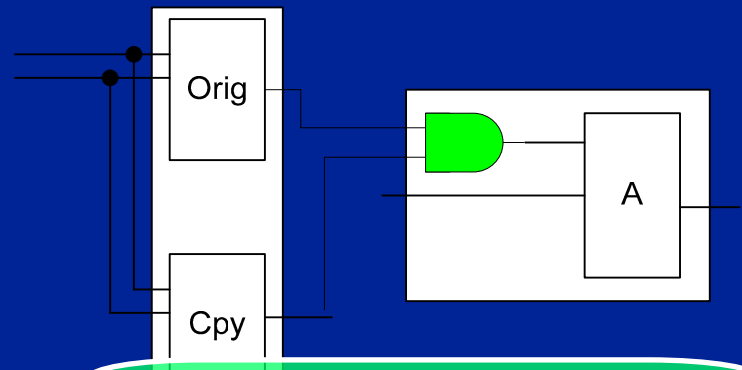
FMD Impact on SRAM Criticality (cont.)



Average Crit. = 0.25

Input	Output	Crit.
000	0	0.2
001	1	0.2
010	0	0.4
011	1	0.2
100	0	0.2
101		
110		
111	1	0.2

AND Encoding

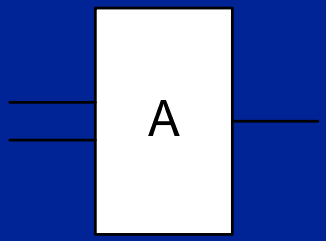


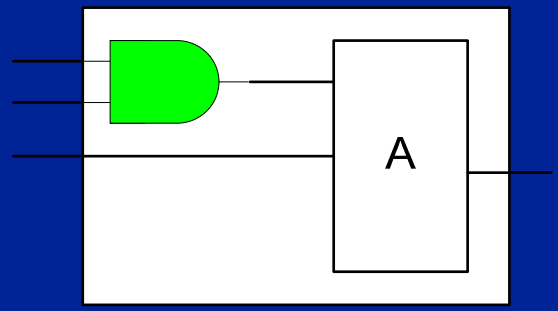
Average Crit. = 0.08

Input	Output	Crit.
000	0	0
001	1	0.2
010	0	0
011	1	0.2
100	0	0
		2
111	1	0.2

Average criticality reduces after duplication.

FMD Impact on SRAM Criticality (cont.)





AND Encoding

Average Crit. = 0.125

Input	Output	Crit.
000	0	0.2
001	1	0.2
010	0	0.4
011	1	0.2
100	not used	0
101	not	
110	not	
111	not used	0

Average Crit. = 0.125

Input	Output	Crit.
000	0	0.2
001	1	0.2
010	don't care	0
011	don't care	0
100	don't care	0
		0
		.4
111	1	0.2

Average criticality remains after encoding.

FMD ILP Formulation

- **Requirement for applying a FMD:**
  - **Each of the LUT to be duplicated and its fanout LUTs must have at least one input pin not occupied**

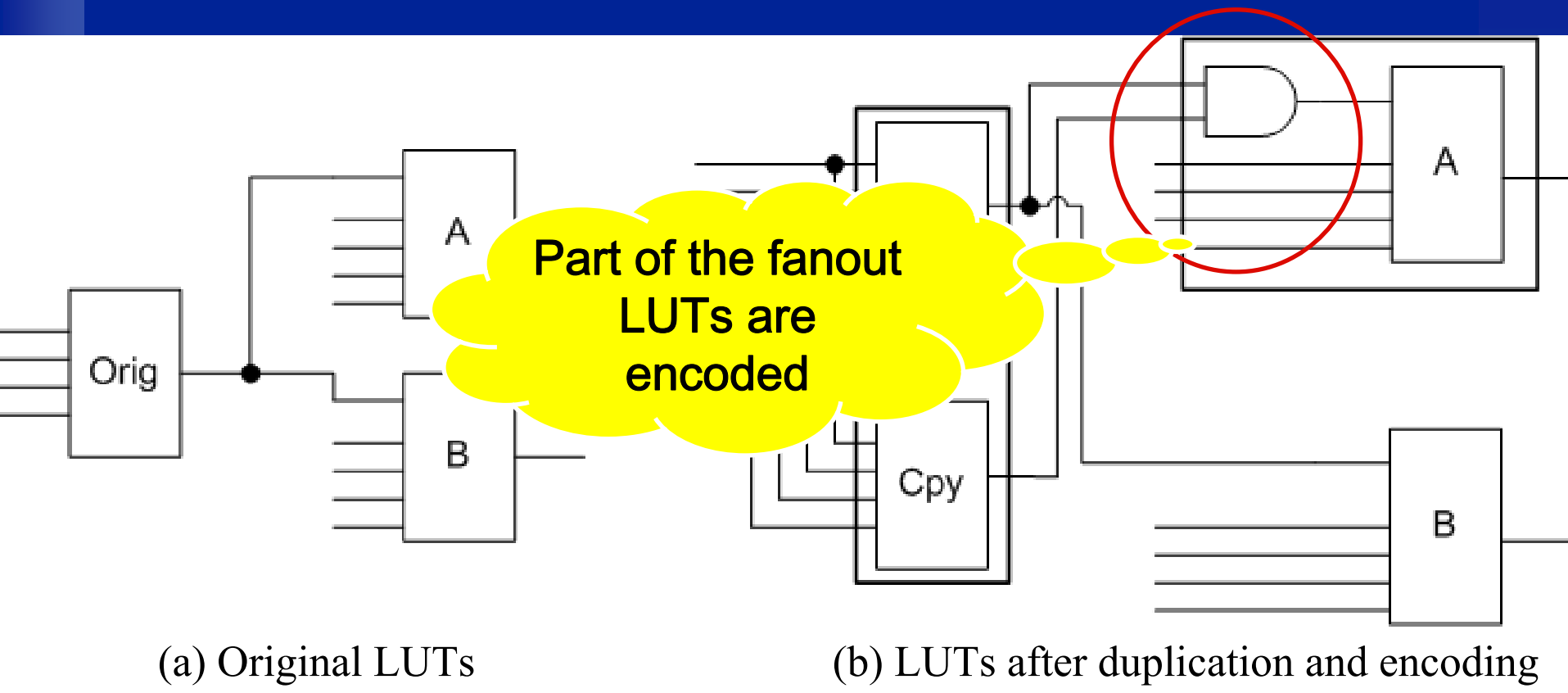
- **ILP formulation:**

Criticality reduction due to duplication of LUT L

$$\begin{aligned} \text{Maximize} \quad & \textstyle\sum_{L \in Luts(C)} w_L \cdot d_L \\ \text{Subject to} \quad & d_L + \textstyle\sum_{f \in \text{fanin}(L)} d_f \leq S_L, \forall L \in Luts(C), \\ & d_L \in \{0, 1\}, \end{aligned}$$

Decision variables: 1 indicates duplication of LUT L

Partial Masking-based Duplication (PMD)



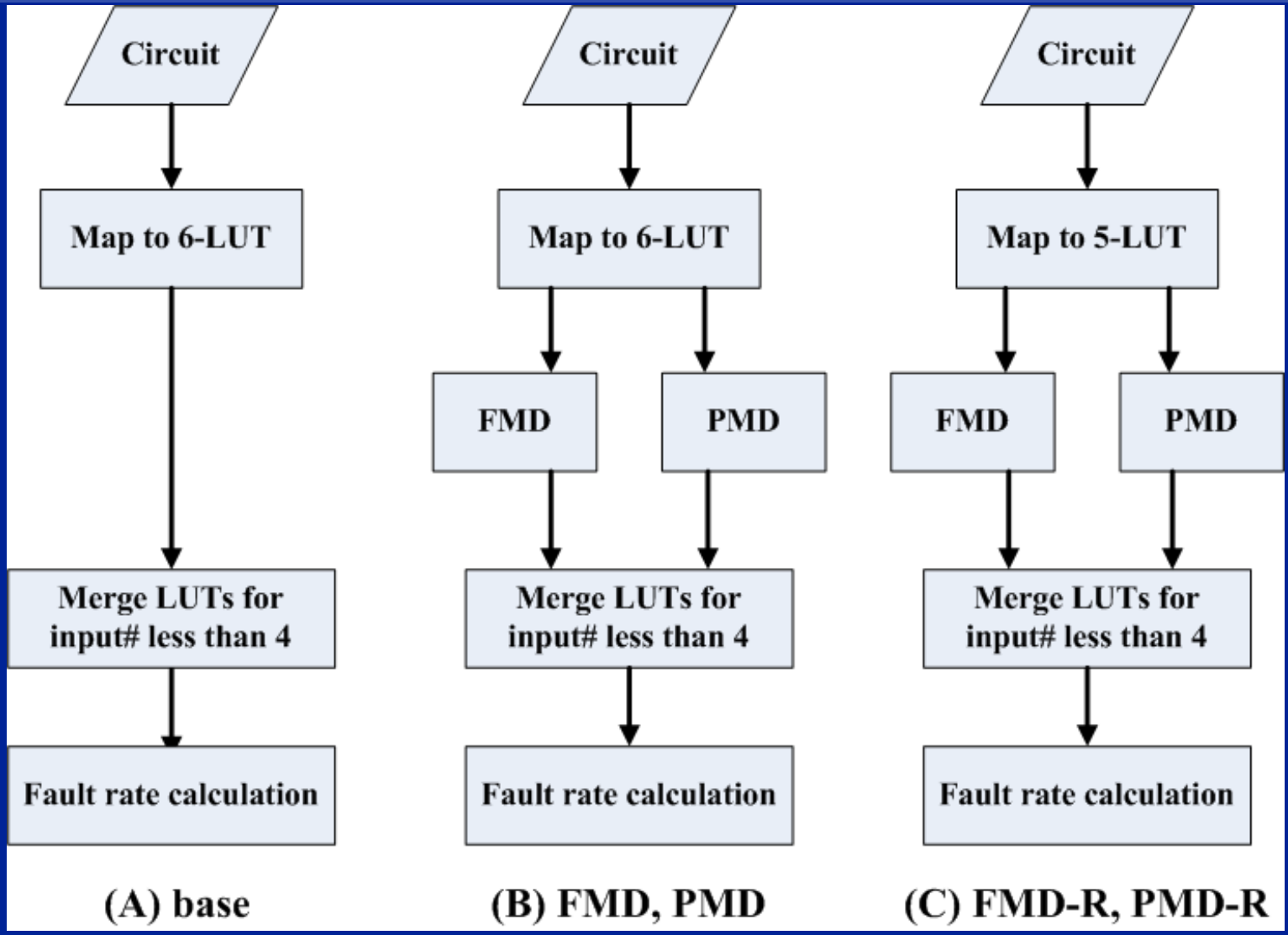
(a) Original LUTs (b) LUTs after duplication and encoding

- A generalized Full Masking-based Duplication

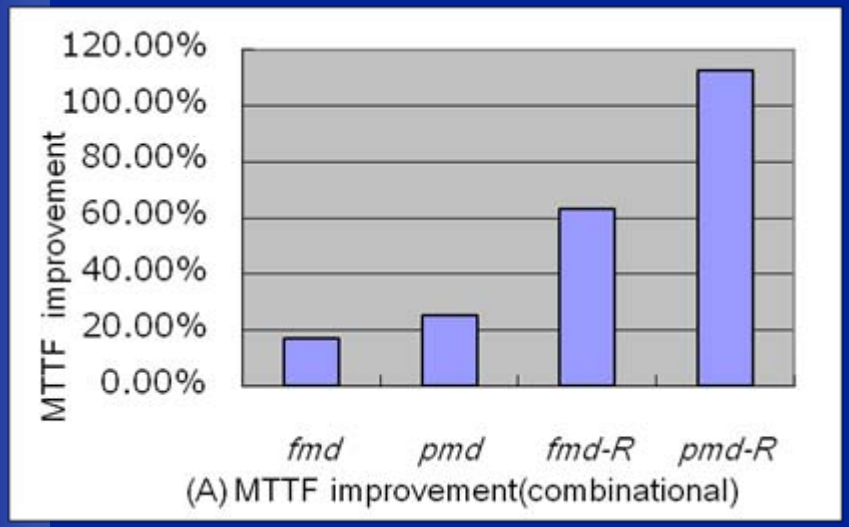
Experimental Settings

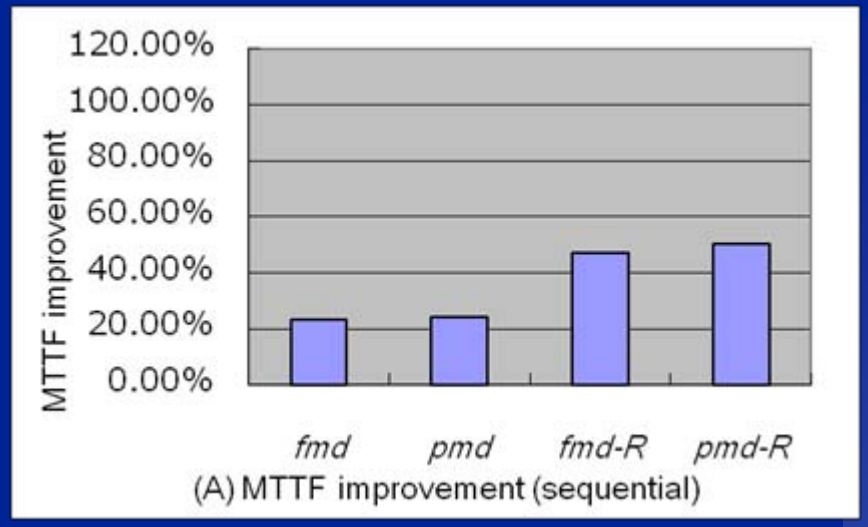
- Benchmarks
  - Biggest 20 MCNC circuits
  - Mapped to 6-LUTs by Berkeley ABC
- A LUT-merger algorithm [Ahmed et al, FPGA'07] is used for area reduction.
- Full-chip fault rate is verified by Monte Carlo simulation with 5K input vectors
- Three CAD flows are examined

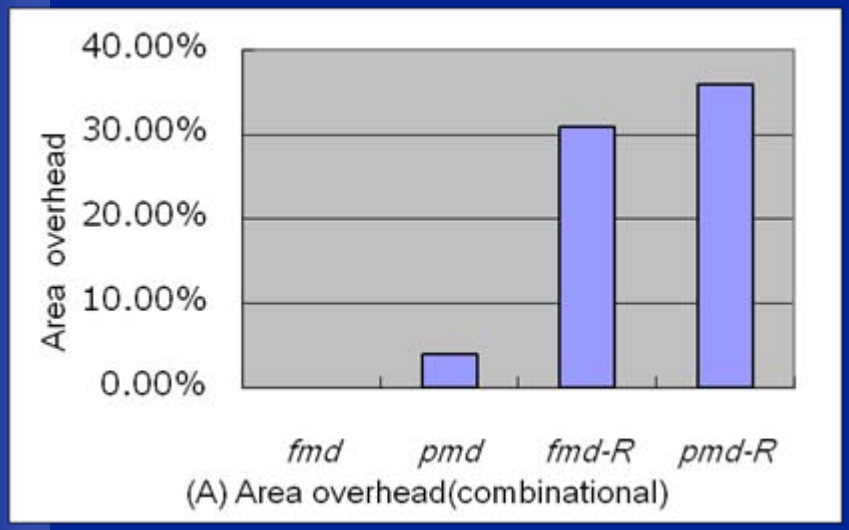
Experimental CAD Flows

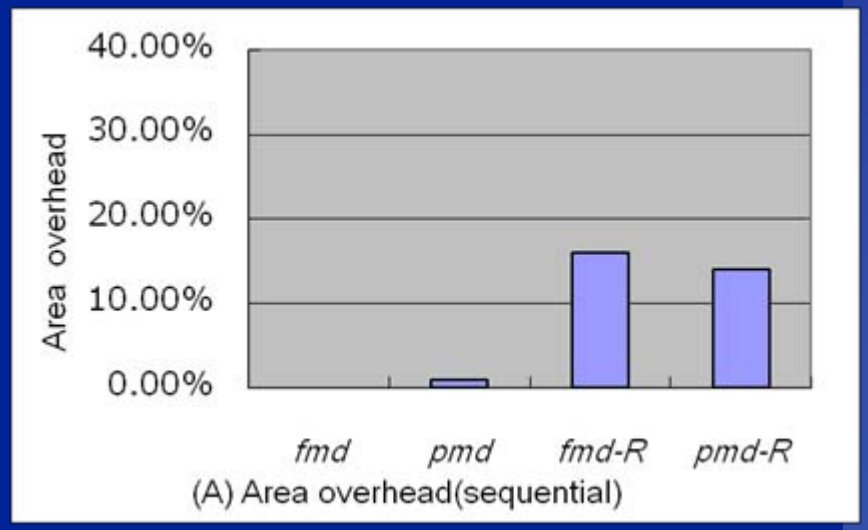


MTTF & Area Overhead

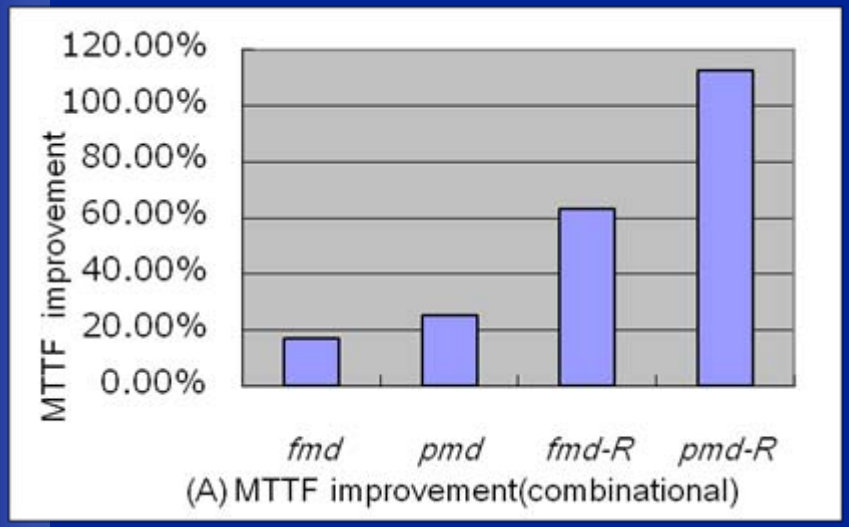


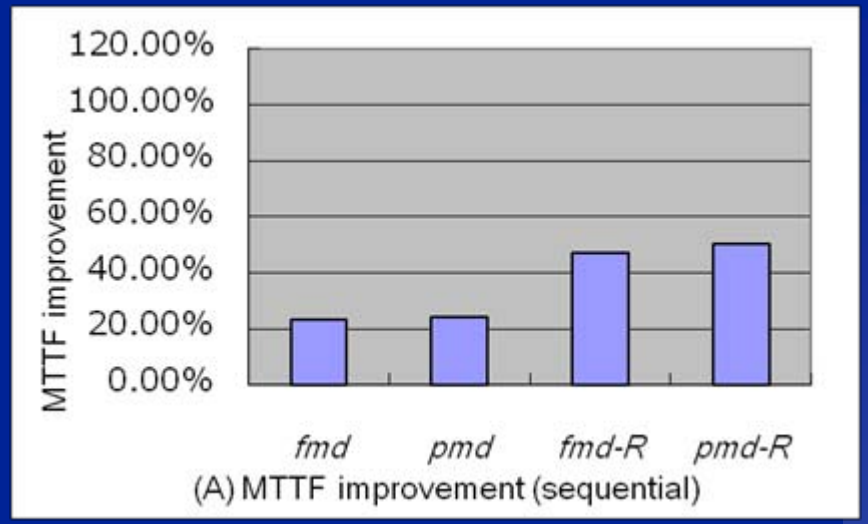


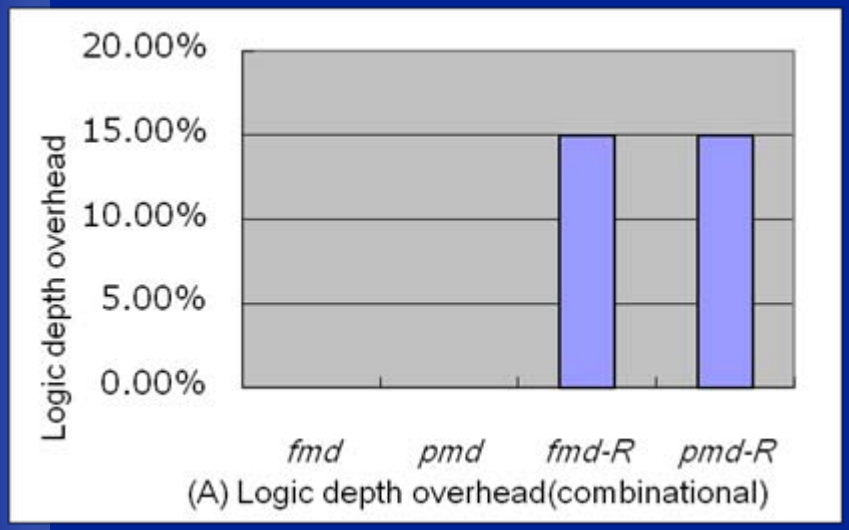


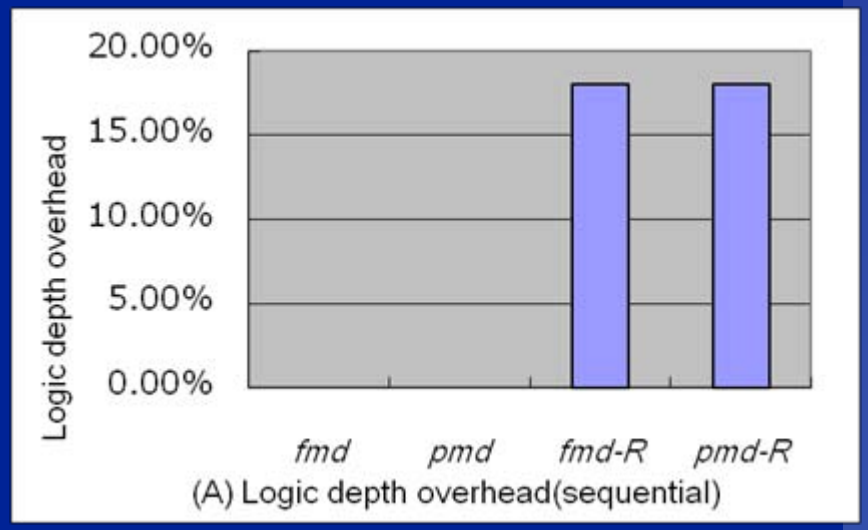


MTTF & Performance Overhead









Conclusions and Future Work

- Proposed a novel fault-tolerant technique using dual-output LUTs
  - 2X MTTF increase w/ 24% area overhead
- In the future, we will consider
  - Different type of encoding logic
  - Dual-output-aware physical synthesis which considers interconnects explicitly
  - Path-based duplication
- To build a selected TMR flow

Thank you!

Electronic Design Automation Group

Electrical Engineering, UCLA

Website: http://eda.ee.ucla.edu