Speed Binning Aware Design Methodology to Improve Profit under Parameter Variation

January 27, 2006

Presenter: Kunhyuk Kang

Authors: Animesh Datta, Swarup Bhunia*, Saibal Mukhopadhyay, Jung Hwan Choi, and Kaushik Roy

Dept. of ECE, Purdue University, IN, USA

* Dept. of EECS, Case Western Reserve University

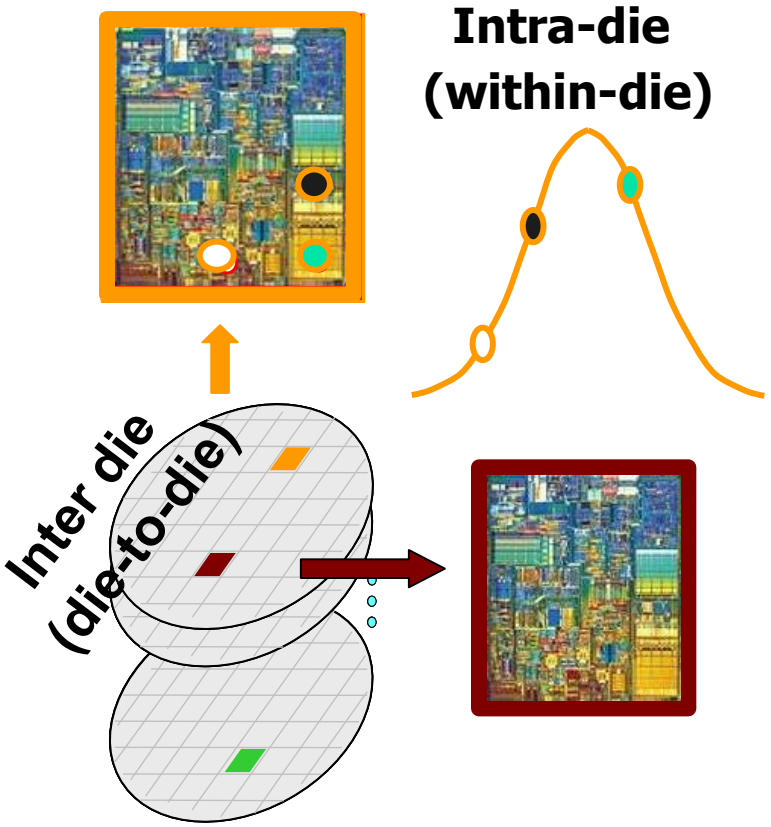
NANOELECTRONICS RESEARCH LAB

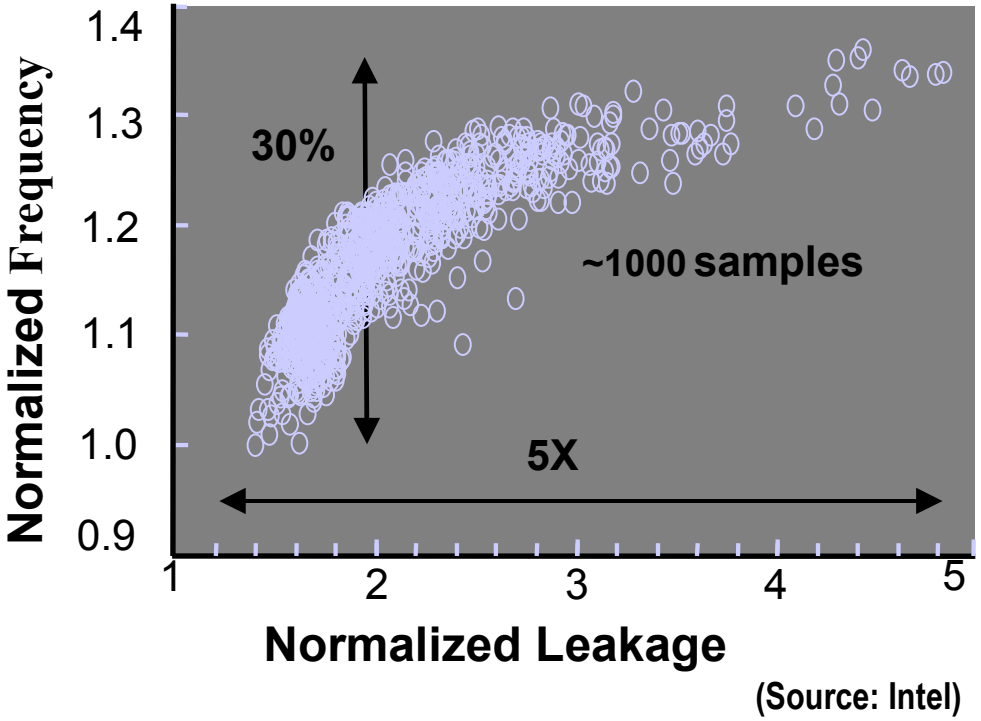
PURDUE UNIVERSITY

Outline

- Introduction
- Speed-binning aware weighted-yield model
- Gate-sizing technique to improve design profitability
- Profit-aware bin boundary placement
- Integrated profit-aware statistical design method
- Conclusions

Process Variation Issues





(Source: Intel)

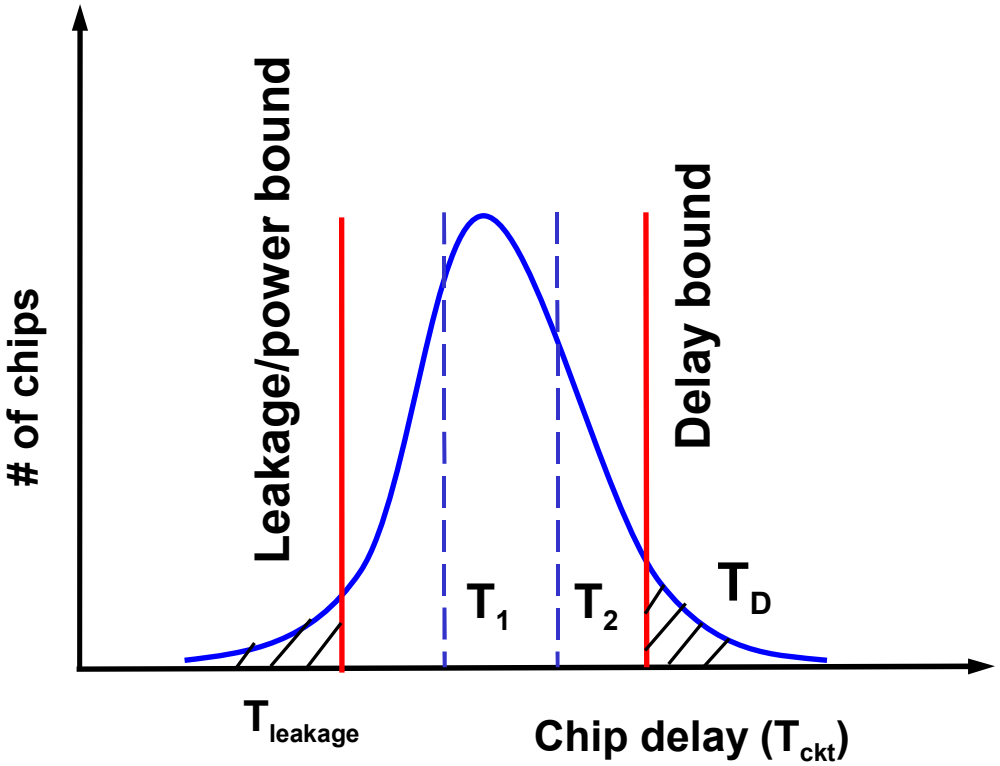
Process parameter variations

(W, L, T_{ox}, RDF) → V_{th}

- **Delay (frequency) spread**
- **Leakage current spread**
- **Statistical circuit design**

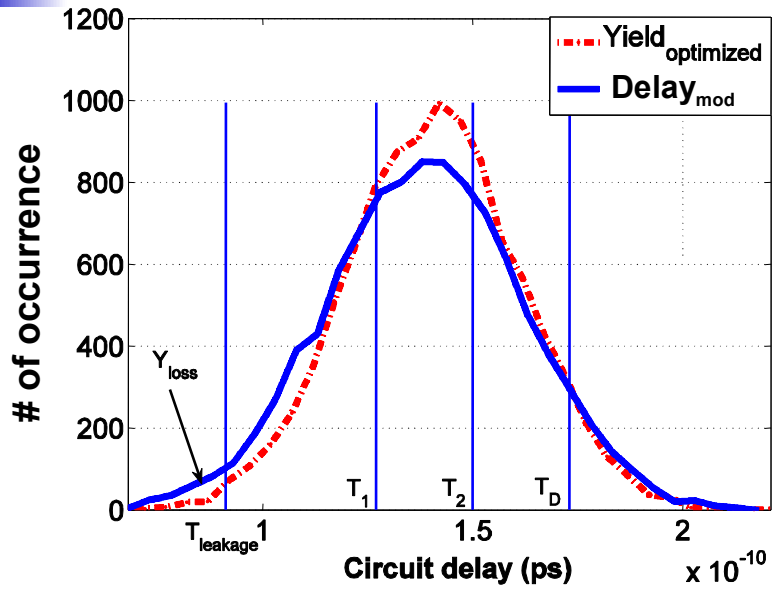
Speed-Binning

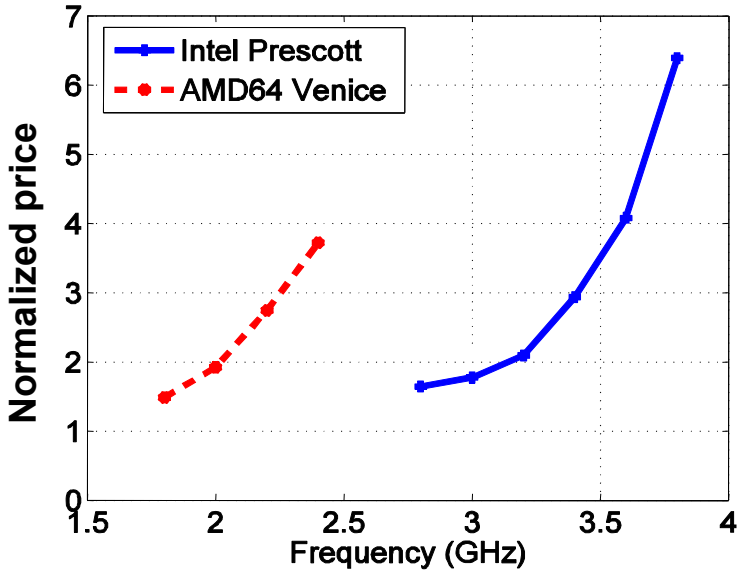
- **What is speed-binning?**
  - Test and classify each functionally correct chip to different bins
- **Why speed-binning?**
  - Improves design profit
  - Salvages low performing chips



How design methodology can be improved to take advantage of speed-binning?

Motivation





(Source: *www.newegg.com*)

- Yield optimized design => Higher $Y_{effective}$
- Design with higher spread => $Y_{effective}$ ↓, commercial value ↑

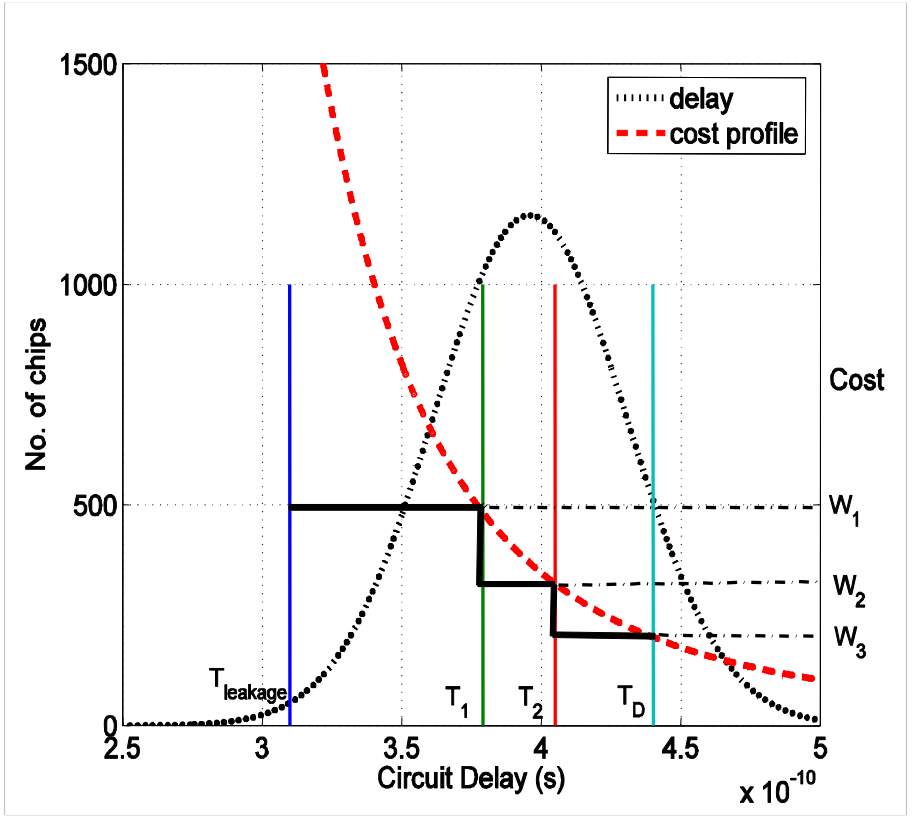
Variability can increase profit under non-linear price-profile

Profit Aware Weighted-Yield

- **Price-weighted yield (Y_P)**

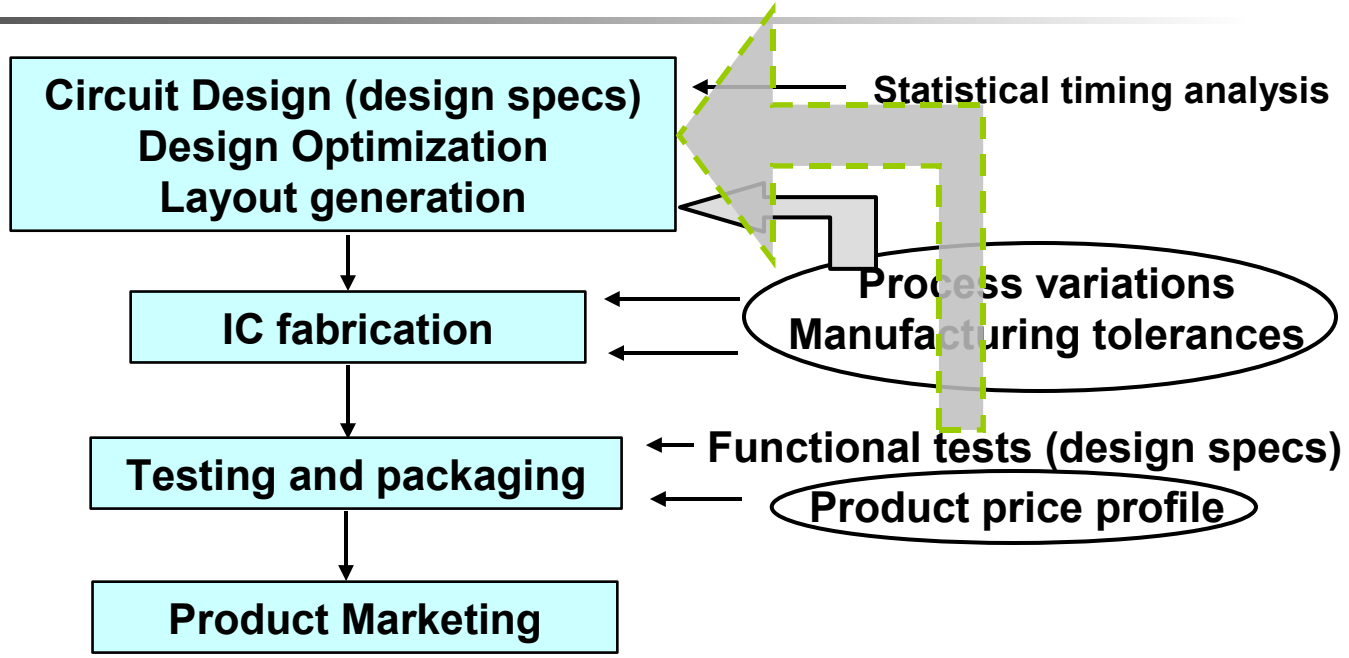
$$Y_P = \sum_{i=1}^{N} C(T_i) Y_{bin_i} \qquad Y_P = \sum_{i=1}^{N} w_i Y_i$$

- **Different price-profiles**
  - **Linear**
  - **Quadratic**
  - **Exponential**



Bin-Aware Statistical Circuit Design

Use binning price profile information during design optimization



Bin-aware circuit design optimization under a price profile

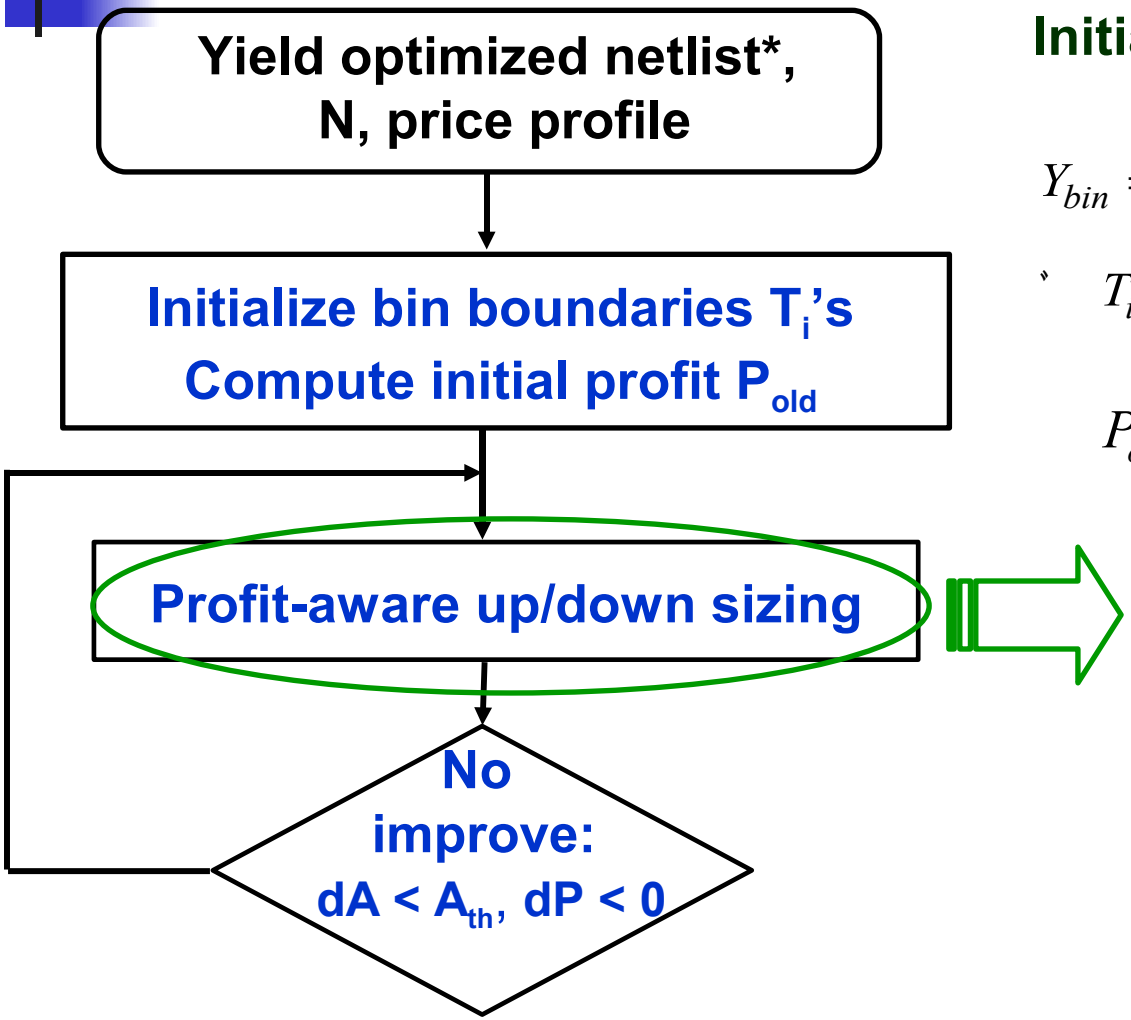
Maximize $Y_P = \sum_{i=1}^{N} C(1/T_i) Y_i = \text{Profit-estimate}$

Subject to: $A = \sum_{i=1}^{n} x_i < A_{init} + \Delta A$

N: Bins; *n*: gates

x: size factor of gate

Profit-Aware Design



Initial bin boundaries (T_i)

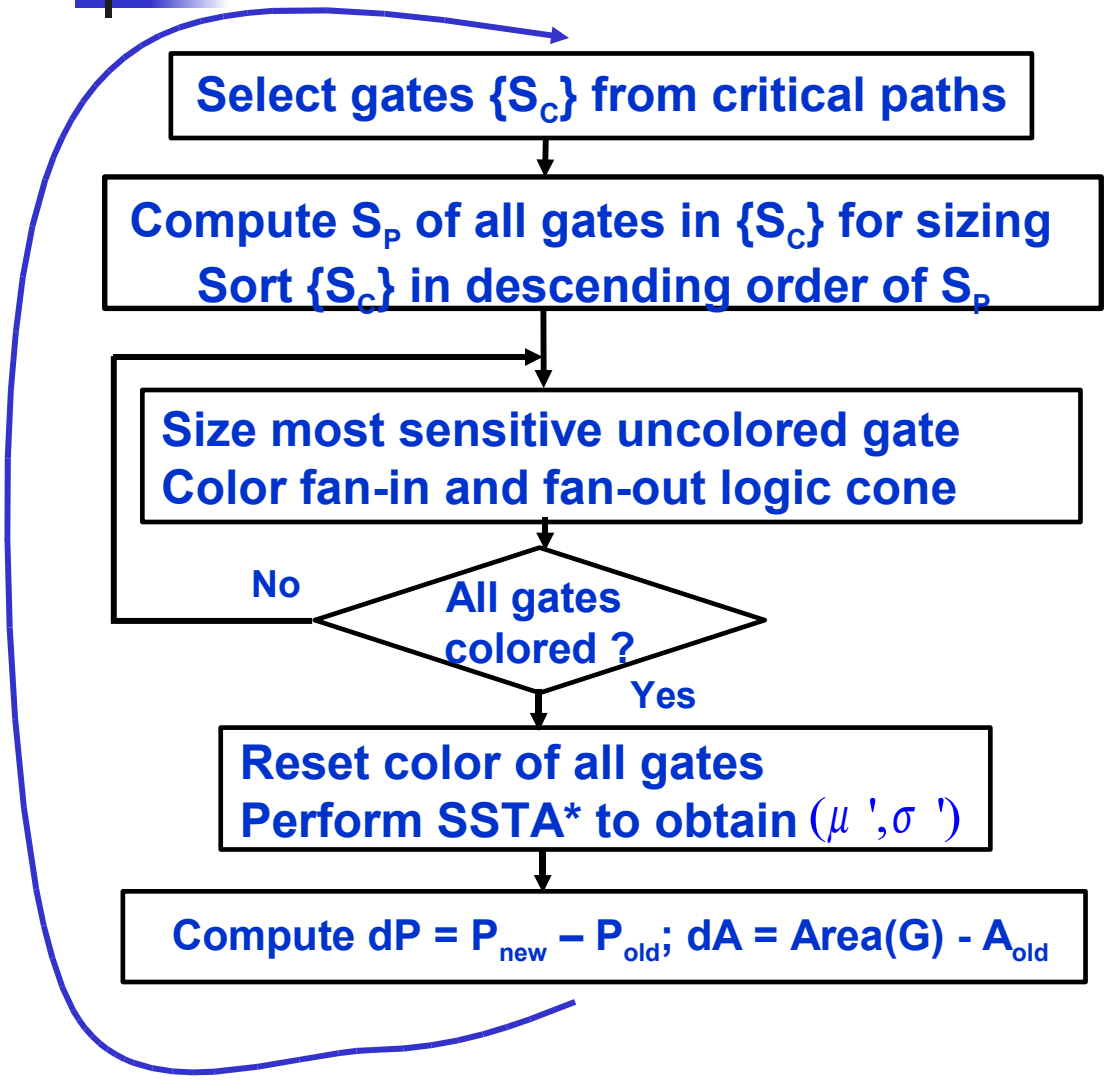
$$Y_{bin} = \frac{Y_T - Y_{leakage}}{N}$$

$$T_i = \Phi^{-1}((Y_{leakage} + i * Y_{bin}), \mu, \sigma)$$

$$P_{old} = \sum_{i=1}^{N} Y_i C(1/T_i)$$

* S. Choi et al. DAC 2004

Profit-Aware Up/down Sizing



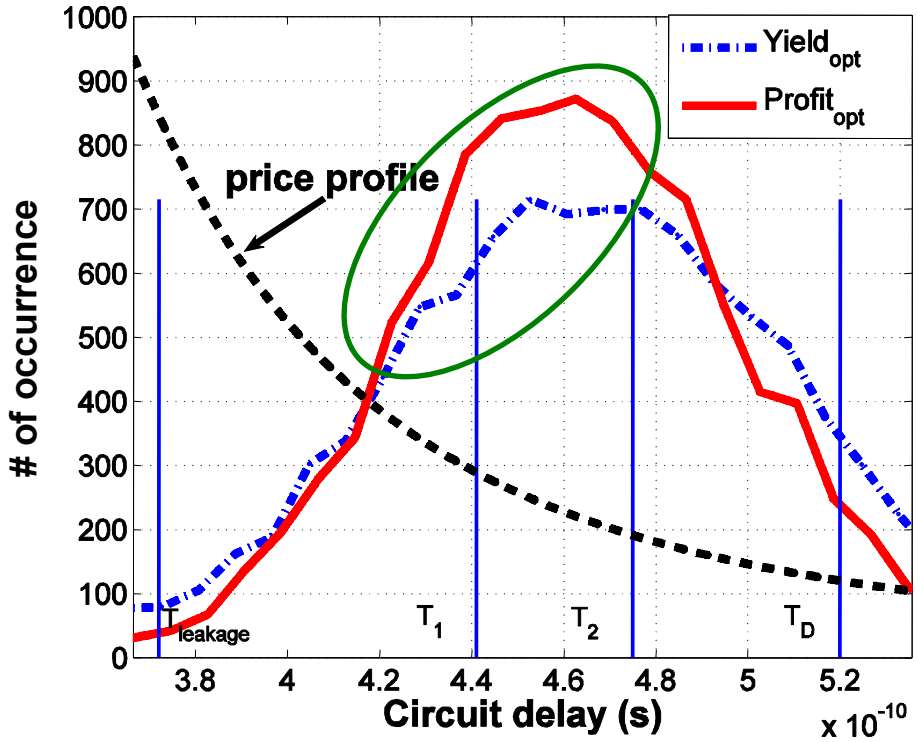
Profit sensitivity S_P = dP / dx

dx = change in gate size factor

$$P_{old} = P(T_0, T_1, \ldots, T_N, \mu, \sigma)$$
$$P_{new} = P(T_0, T_1, \ldots, T_N, \mu', \sigma')$$

* K. Kang et al., Date 2005

Delay Distribution Change

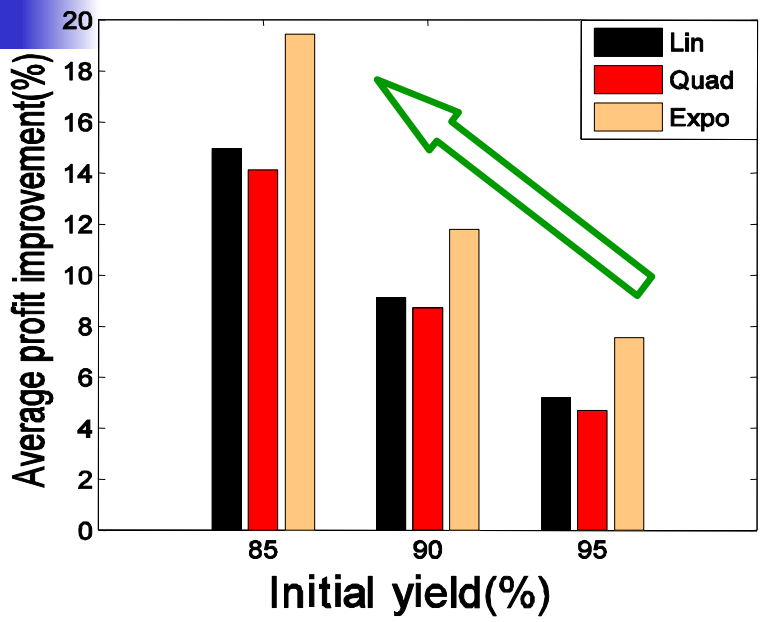


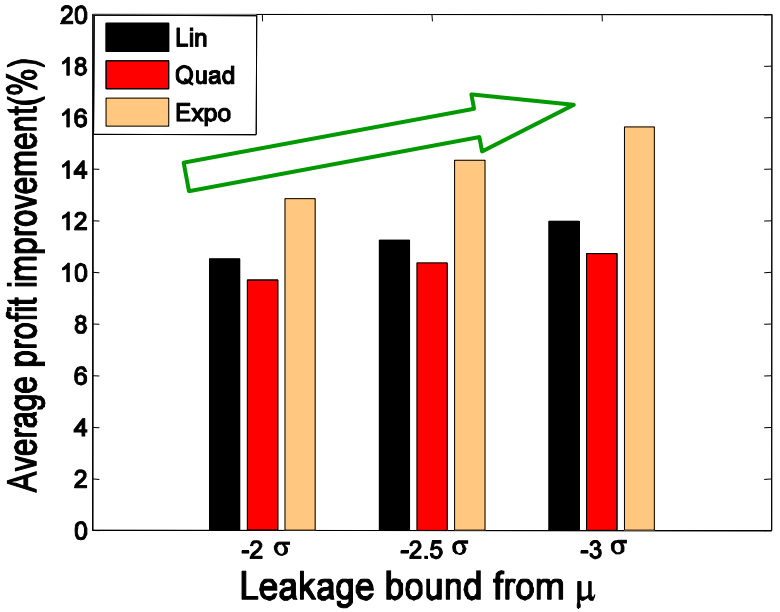
Exponential bin pricing

- Y_{init} = 90%
- $T_{leakage}$ = $\mu - 2.5\sigma$
- c2670 circuit

- Circuit delay distribution change by bin-aware sizing
- Profit optimized design has 9.1% higher Y_P

Experimental results



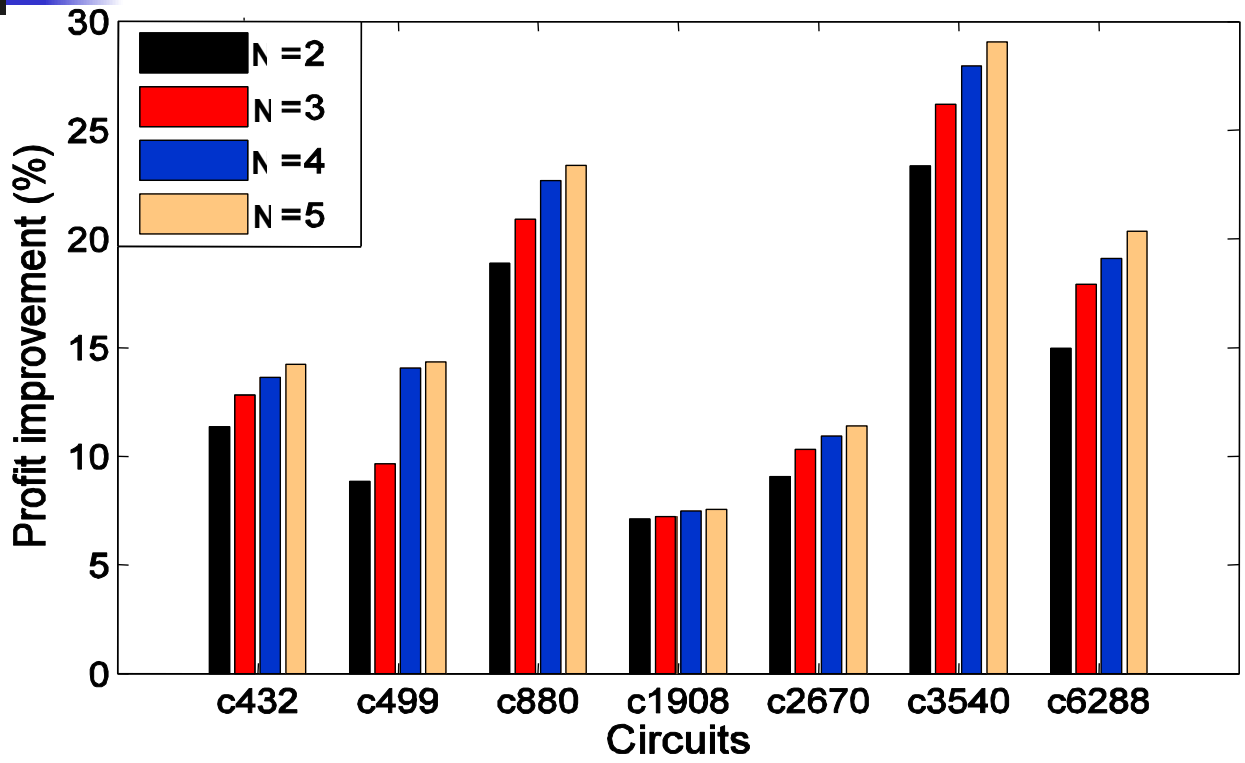


Average profit improvement over ISCAS85 benchmarks

- 8% P_{Improv} even at 95% yield
- P_{Improv} increases with $Y_{initial}$ ↓
- P_{Improv}: a weak function of $T_{leakage}$

- 70nm BPTM
- On average A_{th} ~ 0.3% of $A_{initial}$
- $N_{bin}(3)$, $Y_{init}(90\%)$, $T_{leakage}(\mu - 2.5\sigma)$

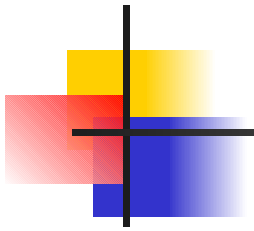
Profit Improvement for Different N



N = number of bins

Exponential price profile

- **Profit improvement is circuit topology dependent**
- **With N_{Bin} ↑ fine grained bin-aware optimization is possible**



How the choice of bin-boundaries affects the profit improvement ?

Optimal Bin Boundary Placement

Iteratively optimized T_i for maximizing Y_P form the highest frequency bin → Largest change in Y_P

$$P_{old} = P(T_0, T_1, \ldots, T_N, \mu, \sigma) = \sum_{i=1}^{N} Y_i C(1/T_i)$$

$$P_{new+} = P(\ldots, T_i + dT, \ldots, T_N, \mu, \sigma)$$

$$P_{new-} = P(\ldots, T_i - dT, \ldots, T_N, \mu, \sigma)$$

for each T_i $(0 < i < N)$

  $dP+ = P_{new+} - P_{old}$;

  $dP- = P_{new-} - P_{old}$;

  if ($dP+ > 0$)

    $T_i = T_i + dT$;

  else if (dP- > 0)

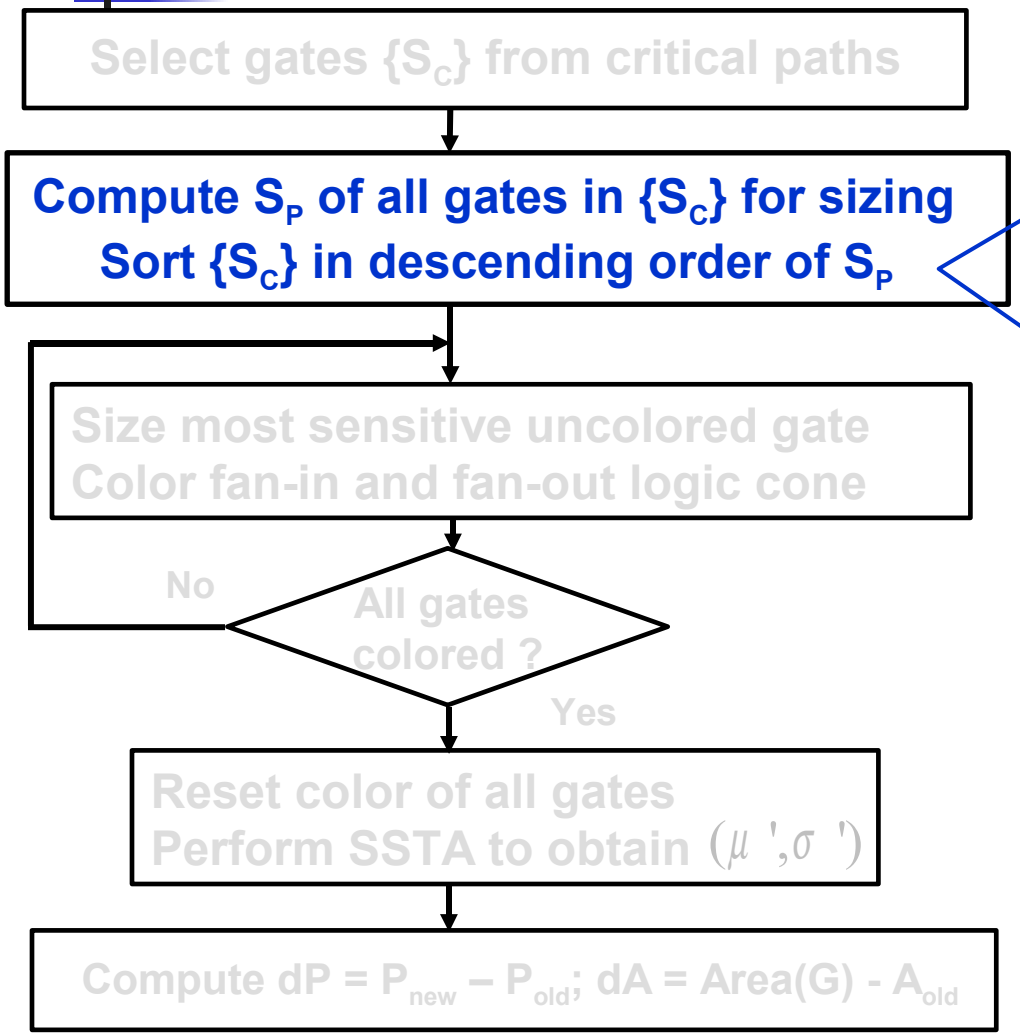
    $T_i = T_i - dT$;

  end if

end for

Avg. 5% P_{Improv} for Expo profile without any change in design

Simultaneous Sizing and Bin Placement



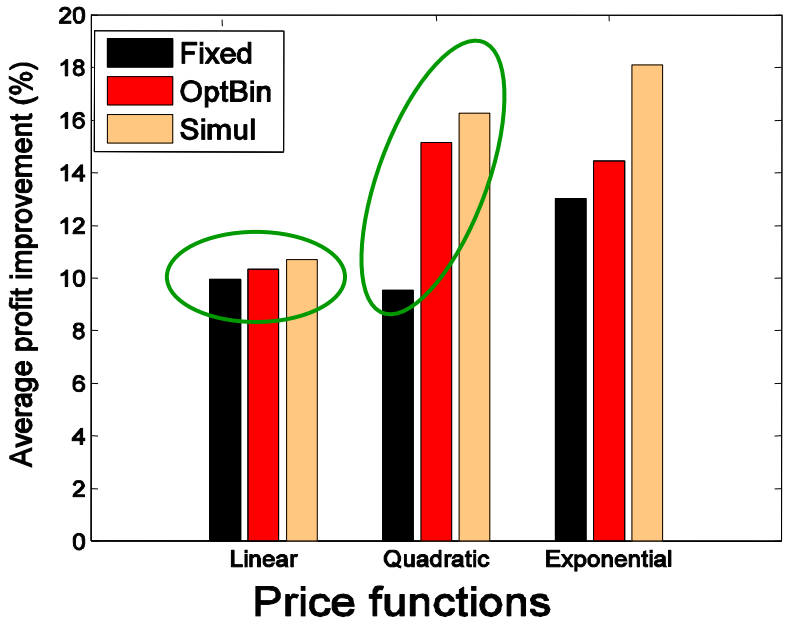
Profit sensitivity $S'_P = dP_{opt} / dx$;

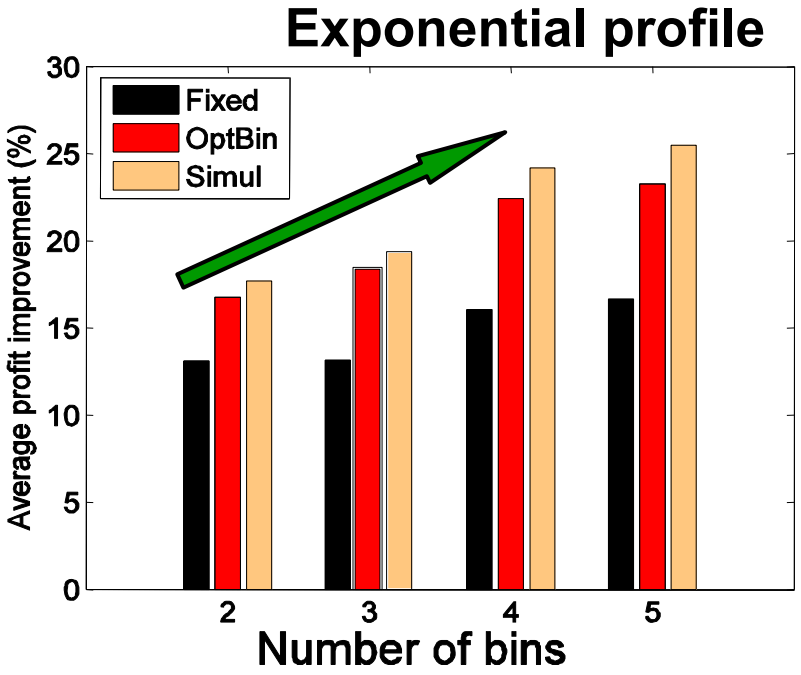
1. **dx: change in gate size factor**
2. **Opt. bin placement**
3. **Compute dP_{opt}**
4. **Compute S'_P**

- **A simple modification in the basic sizing algorithm**
- **Negligible effect in the runtime of simultaneous sizing**

Integrated Bin-Aware Design

- Y_{init} = 90%; $T_{leakage}(\mu - 2.5\sigma)$



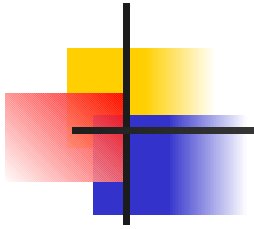


- Initial equal bin boundaries (T_i) are near optimal for Lin. profile
- Initial T_i's are not optimal for Quad. profile
- Consistent profit improvement with increasing N

Conclusions

- Proposed an price-weighted yield model
- Developed an efficient sizing methodology that consider speed-binning price profile in design phase
- Proposed bin-aware design scheme can optimize large scale circuits to improve weighted yield by 19% with small runtime

Proposed a circuit optimization scheme leverages benefit of the increased variations in the scaled technology



Thank you

Contact author: Animesh Datta (adatta@purdue.edu)

Bin-Aware Sizing Routine Runtime

- Runtime of the algorithm is circuit topology dependent

Proposed algorithm takes

- 0.12 secs for c74L85 (33 gates)
- 15 min. for c6288 (2503 gates) ISCAS85 circuits.

Avg. runtime over ISCAS85 benchmarks : 18.45 sec

M/C Specs: Linux server, 3.06 GHz Pentium Xeon, 2GB RAM

- Sensitivity computation → expensive in runtime
- Number of up/down sizing iterations : 4~8 (ISCAS85)