Speed Binning Aware Design Methodology to Improve Profit under Parameter Variations*

Animesh Datta, Swarup Bhunia[1], Jung Hwan Choi, Saibal Mukhopadhyay, and Kaushik Roy
School of ECE, Purdue University, IN 47907, USA, <*adatta, choi56, sm, kaushik*> *@ecn.purdue.edu*
Dept EECS, Case Western Reserve University, Cleveland, OH 44106, USA, *Swarup.Bhunia@case.edu*

Abstract—*Designing high-performance systems with high yield under parameter variations has raised serious design challenges in nanometer technologies. In this paper, we propose a profit-aware yield model, based on which we present a statistical design methodology to improve profit of a design considering frequency binning and product price profile. A low-complexity sensitivity-based gate sizing algorithm is developed to improve the profitability of design over an initial yield-optimized design. We also propose an algorithm to determine optimal bin boundaries for maximizing profit with frequency binning. Finally, we present an integrated design methodology for simultaneous sizing and bin placement to enhance profit under an area constraint. Experiments on a set of ISCAS85 benchmarks show up to 26% (36%) improvement in profit for fixed bin (for simultaneous sizing and bin placement) with three frequency bins considering both leakage and delay bounds compared to a design optimized for 90% yield at iso-area.*

1. INTRODUCTION

Aggressive technology scaling has led to large uncertainties in device and interconnects characteristics for nano-scaled circuits [1]. Increasing variations (both inter-die and intra-die) in device parameters (channel length, gate width, oxide thickness, device threshold voltage etc.) produce large spread in the speed and leakage power consumption of integrated circuits (ICs) [1, 3].

Fig. 1 shows the distribution of operating frequency and leakage current over a large number of high-end micro-processor chips [3]. From the figure, it can be observed that a mature silicon technology like 130nm suffers from about 30% variation in maximum allowable frequency of operation and about 5*X* variation in leakage power. For newer technologies, the variations are reported to be much higher with about 20*X* variation in leakage power for 90nm technology [3]. Consequently, parametric yield of a circuit (probability to meet the desired performance or power specification) is expected to suffer considerably, unless an overly pessimistic worst-case design approach is followed. Therefore, design of high-performance circuits maintaining or enhancing yield under parameter variations has emerged as a major challenge in nano-scaled technologies [3].

In recent years, statistical analysis of timing and power has been extensively explored [2, 5, 8]. Several parametric yield models have been proposed to consider impact of different sources of variations on circuit delay and power [5, 6]. At the same time, multiple efforts have been made to develop statistical design methodology that either ensures or enhances parametric yield (e.g. with respect to delay or power) under specific design constraint (e.g. on area or power) [6, 10, 13].

Profitability of a design is conventionally equated with yield [1, 3]. However, large spread in the frequency distribution due to increasing uncertainties has led to the concept of speed-binning to improve the design profit [4]. Presently, speed-binning is widely used during manufacturing test to qualitatively sort the working (i.e. free from manufacturing defects) ICs based on their highest permissible frequency of operation. During the speed-binning process, functional or structural tests are run at multiple frequencies and parts are binned according to the highest speed test they pass. Working ICs are then priced based on their respective frequency bins [4].

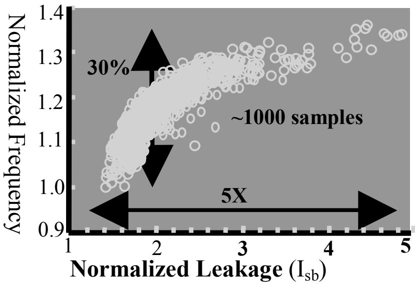


Figure 1: Leakage and frequency variations in 130nm technology (*source: Intel*)

Since high-frequency ICs correspond to higher price points, maintaining yield at a target circuit delay (i.e. frequency) under statistical delay distribution does not ensure high profit. Here, profit refers to the cumulative sum of price-weighted yield of the ICs across all frequency bins. Hence, there is a need to develop design methodology that can either maintain or improve the design profit (instead of yield at a target delay) under statistical delay variations. In [10] authors have used a linear profit function to capture the performance levels of different ICs. However, they did not consider the impact of design choices or delay distribution change on the profit function.

To the best of our knowledge, there is no design technique addressing optimization of design profit (instead of parametric yield at a fixed target delay), considering frequency bins and product price profile. In particular, this paper makes the following contributions:

- A weighted yield model to represent profit that considers price of ICs running at different frequencies and satisfying specific power dissipation requirement.
- A statistical design methodology to improve profitability of a design using a sensitivity-based low-complexity gate sizing under both delay and power bounds.
- An algorithm to determine optimal bin boundaries for a design to maximize profit for a given price profile and design specification.
- An integrated design flow for simultaneous gate sizing and optimal bin boundary placement to improve design profit for a given price profile with a constraint on area.

2. BACKGROUND AND MOTIVATION

In this section, we describe a parametric yield model of a design based on a target delay and leakage power constraint.

2.1 Modeling Yield with respect to a Target Delay

Under parameter variations, circuit delay is the maximum of all path delays in the circuit [1]. The overall circuit delay, T_{ckt} thus follows a distribution and can be modeled as a random variable with mean μ and standard deviation (STD) σ (i.e. $T_{ckt} \sim N(\mu, \sigma)$) [1]. Hence, for a given target delay, the circuit will have a certain

* The work is sponsored in part by Marco Gigascale Systems Research Center (GSRC) and Semiconductor Research Corp.

0-7803-9451-8/06/$20.00 ©2006 IEEE.

probability to meet it depending on its delay distribution parameters. Conventionally, yield of a design is defined as its probability to meet the target delay (T_D) [6]:

$$Y = P_D = \Pr\{T_{ckt}(\mu,\sigma) \le T_D\} \qquad (1)$$

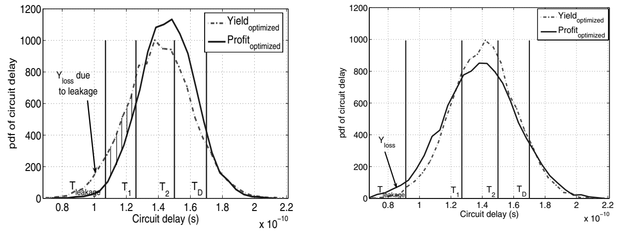
However, in nano-scaled circuits, power consumption of the system also varies from chip to chip along with performance due to variations in transistor threshold voltage*,* channel length, gate width and gate oxide thickness, resulting in parametric yield loss. The yield loss occurs due to the fact that devices with lower V_t (and/or lower channel length) suffer from an exponential increase in sub-threshold leakage. Therefore, there exists a strong correlation among the maximum operating frequency and leakage power of a system. In [5], authors show that when both power and performance constraints are considered, maximum yield loss occurs in the highest frequency bin, while negligible yield loss occurs in other frequency bins. Hence, we can effectively use a minimum delay $T_{leakage}$ (or maximum operating frequency) value as a bound on the leakage power dissipation of the design. Mathematically, when $T_{leakage}$ bound is used together with T_D bound in (1), effective yield of a design can be given by:

$$Y_{effective} = \Pr\{T_D \ge T_{ckt}(\mu,\sigma) \ge T_{leakage}\} \qquad (2)$$

2.2 Motivation

Fig. 2 shows two possible circuit delay distributions (Gaussian) with the corresponding three frequency bin boundaries for a small test circuit. In this figure, Yield$_{\text{optimized}}$ distribution corresponds to a design obtained by optimizing for certain yield constraint at a target delay, T_D. The other distribution (Profit$_{\text{optimized}}$) corresponds to a design where the distribution is changed (by properly sizing the netlist) to improve profit with respect to an exponential price profile (i.e. price point of ICs have exponential dependence on their operating frequency). Fig 2(a) shows that although both distributions meet the yield constraint at T_D, the process of improving profit deterministically changes the distribution (μ increases, while σ decreases) in a way that increases the profit with respect to the given price-profile. In this case, yield loss suffered by profit-optimized design in the highest frequency bin, is easily amortized by significant gain in yields at the other two lower frequency bins. In Fig. 2(b), interestingly, yield degrades due to increase in σ though profit increases due to adequate increase in the number of highest frequency ICs (that more than offsets yield loss at the two lower frequency bins).

Thus, instead of modeling and optimizing yield at a target delay, it is important to consider modeling of profit and to explore design space to optimize profit for high performance systems, which employ frequency binning.



(a) (b)

Figure 2: Circuit delay distributions for yield- and profit-optimized design with a) iso-yield at the target delay, and b) profit-optimized design of smaller yield

3. PROFIT-AWARE YIELD MODEL

We propose a new profit-aware yield model that represents product profitability considering both frequency binning and product price profile.

3.1 Yield Model Considering Frequency binning and Price Profile

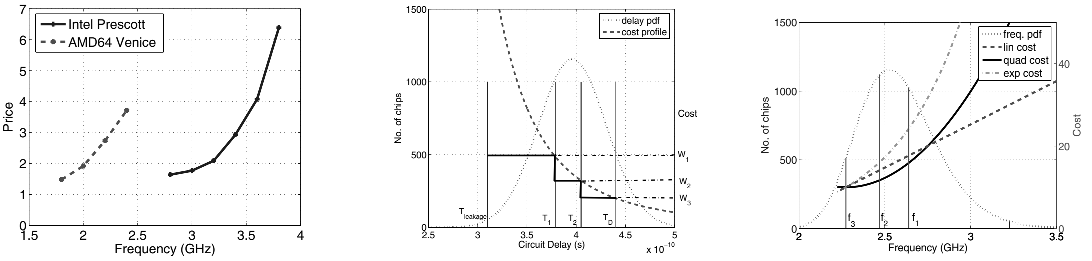
Fig. 3(a) shows a normalized price vs. frequency specification of two recent high-end processors [14]. We observe that for both products, price of the highest frequency part is about three times higher than that of the lowest frequency parts. Hence, to capture the quality of different chips in the design objective function, we model design profit as a price-weighted cumulative sum of yields at different frequency bins. Thus, considering N frequency bins, the profit-aware design yield can be expressed as:

$$\begin{aligned} Y_P &= \sum_{i=1}^{N} C(T_i) Y_{bin_i}; \text{ where } \mathrm{T_N} = \mathrm{T_D} = \text{Target design delay} \\ Y_P &= \sum_{i=1}^{N} w_i Y_i; \text{ where } \mathrm{w_i} = C(T_i) \end{aligned} \qquad (3)$$

where, weighing parameter $w_i = C\ (T_i)$ is the price of a chip in the i^{th} frequency bin. It can be noted that the weighted yield Y_P directly represents the profitability of the design. Fig. 3(b) shows delay distribution vs. exponential product price profile with three frequency bins (i.e. N = 3) for an ISCAS85 benchmark circuit (c499) realized in 70nm BPTM technology [9]. The delay distribution is computed considering both systematic and random variations in threshold voltage. Since all the ICs in a particular frequency bin are sold at the same price, product price profile becomes a stair-case function of delay (or frequency) irrespective of the nature of the price function (Fig. 3(b)). As different circuits have different delay distribution parameters (μ, σ), for a given circuit, we choose price weights (w_i) in such a way that the ratio of the prices at the highest and lowest frequencies is constant for all circuits. Mathematically this can be represented as:

$$\begin{aligned} &\frac{w_{\max}}{w_{min}} = \text{Constant } = R_{\text{price\_profile}} \\ &\text{where } w_{min} = C(f_{\min}),\ f_{\min} = 1/T_D;\ \ w_{\max} = C(f_{\max}), f_{\max} = 1/T_{leakage} \end{aligned} \qquad (4)$$

Four delay specifications ($T_{leakage}$, T_1, T_2, T_D) are used to consider three frequency bins (Fig. 3(b)). Yield of a frequency bin is defined as the fraction of the chips that lies within a specified



(a) (b) (c)

Figure 3: (a) Price and frequency comparisons of two recent high-end processors; (b) Exponential price profile vs. delay distribution for benchmark c499 and (c) Frequency distribution vs. different cost profile

delay (frequency) range. Assuming a Gaussian circuit delay distribution with mean (μ) and standard deviation (σ), yields of different bins can be expressed as:

$$Y_1 = Y(T_1) - Y_{leakage} = \Phi\left(\frac{T_1-\mu}{\sigma}\right) - \Phi\left(\frac{T_{leakage}-\mu}{\sigma}\right)$$
$$Y_2 = \Phi\left(\frac{T_2-\mu}{\sigma}\right) - \Phi\left(\frac{T_1-\mu}{\sigma}\right);\; Y_3 = \Phi\left(\frac{T_D-\mu}{\sigma}\right) - \Phi\left(\frac{T_2-\mu}{\sigma}\right) \quad (5)$$

where, Φ is the Cumulative Distribution Function of circuit delay. In (3) and (4), price function '*C*' can represent any price profile depending on the specific product. In our experiments, we have considered a variety of price profiles with linear, quadratic, exponential dependence on operating frequency. Any other price profiles and even discrete bin prices can also be modeled in the similar manner to compute the profit. Typical example of an exponential price profile vs. delay distribution for an ISCAS85 benchmark c499 is shown in Fig. 3(b). Fig. 3(c) shows different price profiles vs. operating frequency for this circuit.

The profit-aware yield model (3) can help us to design high-performance circuits under variations such that design profit is maximized instead of the yield at a specific target delay. Using (3), profit optimization problem with respect to a given price profile and the number of frequency bins N can be formulated as:

$$\boxed{\begin{array}{l} \text{Maximize } Y_R = \sum_{i=1}^{N} C(1/T_i) Y_i, \text{ where } Y_i \text{ and } T_i \text{ are defined in (3)} \\ \text{Subject to: } A = \sum_{i=1}^{n} x_i = \text{const, where } A = \text{total area of circuit;} \\ \qquad x_i = \text{size for the i}^{\text{th}} \text{ gate; } n = \text{total number of gates} \end{array} \quad (6)}$$

3.2 Statistical Delay Model

To compute the delay distribution of a circuit based on the information of parameter variations, we have employed the statistical static timing analysis (SSTA) algorithm proposed in [8], where delay distribution of a circuit is calculated using Levelized Covariance Propagation (LCP). It was shown in [8] that, using this technique, the effect of both inter-die and intra-die variations can be taken into account. The simulation results on several ISCAS85 benchmarks show average error of 0.21% and 1.07% compared to the Monte-Carlo analysis for mean and standard deviation of delay, respectively. Gate delays are modeled with 70nm BPTM [9] parameters using analytical expression as in Sakurai et al. [11]. For simplicity, we ignore interconnect delays and assume a constant capacitive load for each net. However, our algorithm can be easily extended to incorporate interconnect delays using conventional π-type RC model as used in [7].

4. PROFIT-AWARE DESIGN OPTIMIZATION

Using our profit-aware yield model presented in section 3.1, we propose a statistical design flow for profit optimization under an area constraint.

4.1 Yield Optimization using Gate Sizing

Gate sizing is conventionally used in different circuit synthesis tools for area/power optimizations while meeting the desired timing constraint, or for minimizing the maximum delay under constraint on area/power [6, 7, 10, 13]. Mathematically, gate sizing problem for achieving mean delay A_0 with minimum active area can be formulated as [7]:

$$\boxed{\begin{array}{l} \text{Minimize } \sum_{i=1}^{n} x_i \quad \text{where } L_i \le x_i \le U_i, \text{ for } i = 1, \ldots, n \\ \text{Subject to: } \sum_{i \in p} \text{mean}(D_i) \le A_0, \;\; \forall p \in \text{set of Paths,} \\ \text{Yield } Y \ge Y_T; \text{ where } Y = \Phi(\frac{T_D-\mu}{\sigma}) \end{array} \quad (7)}$$

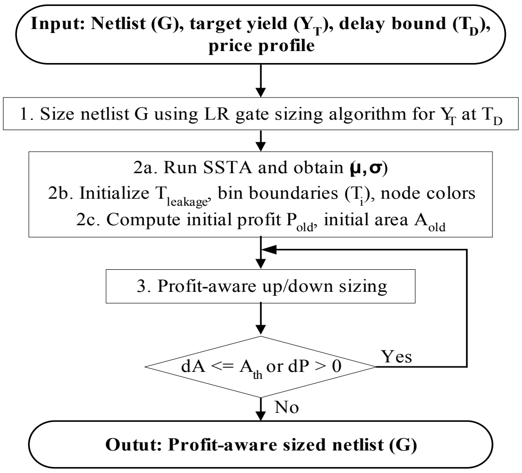

Input: Netlist (G), target yield (Y$_T$), delay bound (T$_D$), price profile
1. Size netlist G using LR gate sizing algorithm for Y$_T$ at T$_D$
2a. Run SSTA and obtain (μ,σ)
2b. Initialize T$_{leakage}$, bin boundaries (T$_i$), node colors
2c. Compute initial profit P$_{old}$, initial area A$_{old}$
3. Profit-aware up/down sizing
dA <= A$_{th}$ or dP > 0
Yes
No
Outut: Profit-aware sized netlist (G)

Figure 4: Profit-aware statistical gate sizing algorithm

where, U_i, L_i, are the bounds of maximum and minimum gate size, respectively, the value of A_0 depends upon the target yield (Y_T), target delay T$_D$, and (μ, σ) of the delay distribution. In [7], a solution for convex gate-level sizing problem is proposed to minimize maximum delay under an area constraint. Starting from the minimum-sized netlist, we iteratively use the Lagrangian Relaxation (LR) based sizing algorithm [7] in conjunction with the statistical timing analysis to achieve the yield target with minimum area. In the m^{th} iteration mean target delay is set to μ_m and at the next iteration we update the target delay by small steps to μ_{m+1} as:

$$\Phi(\frac{T_D-\mu_{m+1}}{\sigma_{m+1}}) \ge Y_T \;\Rightarrow \mu_{m+1} \le T_D - k\sigma_m \Phi^{-1}(Y_T);\text{ where } \sigma_{m+1} = k\sigma_m \quad (8)$$

where, σ_{m+1} is the STD of the circuit delay after sizing in the next iteration, σ_m is STD of circuit delay with current sizes and $k < 1$ is a constant. The solution in [7] provides a globally optimal solution for the problem of area minimization under a static delay constraint. We have observed that with up/down sizing of a logic gate, the mean and standard deviation of the circuit delay shift in the same direction. Based on this observation, we obtain a yield optimized initial design with the minimum area.

4.2 Profit-Aware Gate Sizing

In this section, we propose a sensitivity-based profit-aware gate sizing methodology for the optimization problem proposed in (6). Fig. 4 shows principal steps of the proposed profit-aware sizing methodology. Step 1 has been detailed in section 4.1. Once the design is optimized for a target yield with minimum area, we perform SSTA to determine the delay distribution parameters (step 2a). We define leakage bound based on delay distribution parameters (μ, σ) of the design as:

$$T_{leakage} = f(\mu,\sigma) = \mu - (l*\sigma);\;\text{ where, } l = \text{constant} \quad (9)$$

Since the delay parameters (μ, σ) vary with sizing, the leakage bound also changes with sizing. We define the fixed bin boundaries based on these delay parameters so that Y$_{\text{effective}}$ (effective yield between the highest and lowest permissible frequencies) is equally distributed among *N* frequency bins as:

$$Y_T = \Phi\left(\frac{T_D-\mu}{\sigma}\right) \Rightarrow T_D = \Phi^{-1}(Y_T,\mu,\sigma);\;\; Y_{bin} = \frac{Y_T - Y_{leakage}}{N}$$
$$\Rightarrow T_i = \Phi^{-1}((Y_{leakage} + i*Y_{bin}),\mu,\sigma),\;\; \forall \text{i} = 1, \ldots, \text{N} \quad (10)$$

Then we compute initial design profit P_{old} (using (3)), initial design area A_{old} (computed as active area). In step 3, we compute sensitivity of logic gates with respect to profit (*dP/dx*) for up and down sizing. Next, starting from the most sensitive gate, we perform up/down sizing of logic gates to improve the overall profit of the circuits. This process is performed iteratively until there is no improvement in profit ($dP \leq 0$) or the area constraint cannot be satisfied ($dA > A_{th}$) for further improvement in profit.

Fig. 5 shows pseudo-code for our sensitivity-based up/down sizing routine. We apply sizing step (satisfying the bounds on maximum and minimum gate size) to these nodes and perform SSTA to re-compute the delay distribution. A node with higher sensitivity is sized before the node with lower sensitivity. The runtime complexity of the routine depends on the number of SSTA calls in sensitivity analysis during an up/down sizing iteration. We have employed three optimization techniques to improve the runtime of the proposed gate sizing method.

1. Considering the fact that gates lying in the critical path are most sensitive in terms of delay variation and hence, change in profit, we select a set of gates $\{S_C\}$ on the critical paths. After the SSTA run, we choose a fixed number of critical paths (10,000 in our case) and compute their delay failure probabilities (with respect to T_D) from the path delay parameters (mean and STD). Next, we compute the worst-case delay failure probability for each gate based on the paths crossing it. Gates with high delay failure probability are chosen in an iteration to improve profit (line 1, Fig. 5).
2. We determine profit sensitivity (S_P) of all gates in $\{S_C\}$ by changing one gate size at a time, with a small step (*dx*) and computing the corresponding change in profit (i.e. $S_P = dP/dx$). The sensitivity analysis is performed for both sizing directions (i.e. up and down). If a logic gate has unacceptable profit sensitivities (i.e. profit drops with upsizing or degrades too much with down sizing) in an iteration, we remove it from $\{S_C\}$ in the subsequent calls of up/down sizing (line 3, Fig. 5).
3. Multiple up/down sizing steps are performed after each sensitivity analysis by selecting successive gates not lying in the fan-in and fan-out logic cone of the previously sized gates. In each iteration, we color the fan-in and fan-out cone of a sized logic gate in the graph (*G*) (line 8, Fig. 5). The colored nodes are not considered for sizing in that iteration. When no suitable uncolored gate exists in $\{S_C\}$, the iteration terminates. We then mark all gates uncolored and perform a SSTA to update the increment of profit (*dP*) and area (*dA*) values of the given circuit (line 10-15, Fig. 5).

```
up-downSizing ( )
Input: Netlist (G)
Output: Sized netlist (G) after one sizing iteration
Direction: down = 0;  up = 1
1.  Select gates {$S_C$} from critical set of paths with high failure probability;
2.  Compute profit sensitivity of all gates {$S_C$} for up/down sizing;
3.  Remove the gates with unacceptable profit sensitivity form {$S_C$};
4.  Sort gates of set {$S_C$} in the descending order of profit sensitivity;
5.  while (not all gates are colored)
6.      Choose the most sensitive uncolored gate;
7.      Size the gate by dx in proper direction;
8.      Color its fan-in and fan-out logic cone;
9.  end  //while
10. Reset color of all gates;
11. Perform SSTA (G) to obtain $\mu', \sigma'$;
12. $P_{new} = P(T_0, T_1, \ldots, T_N, \mu', \sigma')$;
13. $dP = P_{new} - P_{old}$;
14. $P_{old} = P_{new}$;
15. $dA = Area(G) - A_{old}$;
```

Figure 5: Pseudo code for up/down sizing

Table I: Profit-aware design results compared to 90% yield-optimized design (N = 3)

Circuit	Target Delay T_D (ps)	Profit improvement (%)			Runtime (sec)
		R_{Lin}	R_{Quad}	R_{Exp}	
c432	520	9.53	9.20	12.80	1.12
c499	440	7.75	8.58	9.66	15.5
c880	400	16.28	15.16	20.89	11.88
c1908	520	7.02	6.50	9.07	51.69
c2670	425	8.11	7.15	10.32	9.91
c3540	640	18.78	18.04	26.18	56.71
c6288	1725	12.98	12.04	17.89	55.74
c74181	200	7.27	6.49	9.28	0.24
c74L85	150	13.83	13.29	19.06	0.04
c74283	170	2.77	2.57	3.53	0.08
Avg.		**10.43**	**9.92**	**14.15**	**18.45**

These three techniques reduce the number of gates used to compute profit sensitivity in an iteration with negligible degradation (< 1%) in the over-all profit improvement compared to a case where sensitivity analysis is performed for all gates. It is important to note that, we perform downsizing of least profit sensitive (by assigning higher weight to the downsizing sensitivity) gates to recover from the area overhead incurred during upsizing phase and to satisfy area overhead constraint A_{th}. Note that the proposed profit-aware sizing methodology (up/down sizing as described in Fig. 5) can be used to improve the profit independent of initial sizing of the design. For example, profit optimization can be performed starting from a minimum-sized design instead of a yield optimized design (Section 4.1).

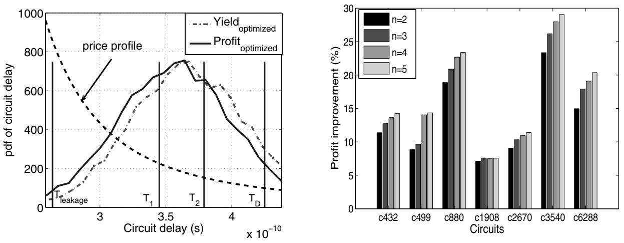
4.3 Experimental Results

We have applied the proposed profit-aware statistical design on several ISCAS85 benchmarks for different number of frequency bins. We have considered design profit improvement for three different price profiles with different price ratios as defined in (4) (e.g. $R_{lin} = 3$, $R_{quad} = 5$, $R_{expo} = 10$). The final product profit (Y_P) is then computed using (3). The area overhead threshold (A_{th}) is taken to be a very small value (0.3%) to observe the scope of profit optimization at iso-area over a yield-optimized design. In Table I columns 3 to 5 present profit improvement for different price profiles as a percentage of profit for a 90% yield-optimized design with $T_{leakage} = \mu - 2.5*\sigma$.

Using the proposed method, we obtain up to 26% profit improvement (for c3540). On average, we observe profit improvements of 10.4% for the linear, 9.9% for the quadratic, and 14.2% for the exponential price profile at iso-area (Table I). Smaller slope of quadratic price function than the linear one (Fig 3(c)) for lower frequency bins is responsible for smaller profit improvement under quadratic price profile over linear one, considering equal yield bin boundaries. However, profit improvement for a particular price profile varies widely across the benchmarks and it largely depends on the design specifications as well as on the circuit topology (Table I).

Profit improvement at the iso-area comes from the proper distribution of the right amount of yield in the right frequency bin (depending on the price profile). Average runtimes of a sizing iteration considering linear price profile are reported in column 6. Similar runtimes are observed for the other price profiles since they are dominated by the number of calls to the SSTA routine. The number of sizing iterations required by the proposed sizing scheme varies from 3 to 21.

Fig. 6(a) plots the delay distributions for c2670 circuit after initial yield optimized design and profit-aware sizing along with its exponential price profile. Note that considerable profit



(a) (b)

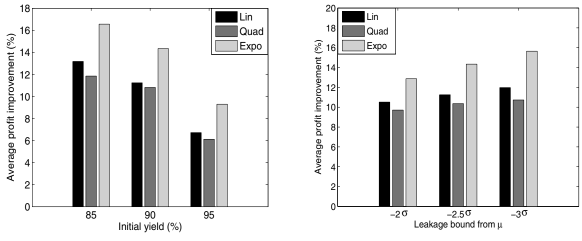
Figure 6: (a) Delay distributions change for c2670 with an exponential price profile. (b) Profit improvement of different number of bins for fixed bin boundaries

improvement (10.3%) is achieved for increased high frequency bin yield (due to reduction in mean). Figure 6(b) shows profit improvements of the various benchmarks as the number of frequency bins is varied under an exponential price profile. Note that we have not considered the smaller benchmarks (i.e. c74 series) in this plot because they have small delay spread (Table I). The trend of increasing profit improvement with number of bins can be attributed to the fact that with fine-grained frequency binning, the high frequency bin prices (and the average bin price) increase considerably under a given price profile.

We have obtained three sets of average profit improvement results for three different target yields (Y_T), while keeping other design specifications ($T_{leakage}$, T_D, N) unchanged (Fig. 7(a)). We observe that profit improvement decreases with the increase in Y_T (Fig. 7(a)). This is due to the fact that with low initial yield, design profit is usually low for the specific price profile, thus has good scope of improvement. Fig. 7(b) shows the profit improvement trend with different leakage bounds. We observe that when we change the leakage bound ($T_{leakage}$) from $\mu - 2*\sigma$ to $\mu - 3*\sigma$ for $N = 3$, average profit improvement increases for all the cost functions. This happens because scope of change in high frequency bin yields and thus scope of profit improvement for a given price profile increase with the smaller leakage bounds.

4.4 Complexity of Profit-Aware Sizing Algorithm

In our proposed sensitivity-based gate sizing approach, we run SSTA multiple times during sensitivity computation of different gates. The complexity of the SSTA is $O(m)$ [7], where m is the maximum number of gates in any level in the levelized netlist of the given circuit. It is shown that even for a very large circuit this number grows very slowly [7]. The complexity of a LR sizing run is $O(n^a)$, where n is the number of gates in the circuit and $a \approx 1.7$ [8]. Moreover, since after the 1st iteration of up/down sizing, the subsequent sizing steps start with a reduced set of gates, the effective runtime of the subsequent up/down sizing iteration reduces. Hence, the over-all complexity of the design flow is dominated by the runtime complexity of the LR-based sizing (section 4.1) and SSTA routine. Let us assume that the LR-sizing routine is called r times and the number of iterations for profit improvement is s. Hence, the total runtime complexity can be given by $O(r*n^a + r*m + s*m*l)$ for a design with average number of sensitive gates in set $\{S_C\}$ equal to l.



(a) (b)

Figure 7: Average profit improvement for different (a) initial yield targets and (b) leakage bounds for (N = 3)

The complete profit-aware design flow takes 0.12 second for the smallest benchmark (c741L85, 33 gates) and 15 minutes for the largest benchmark (c6288, 2503 gates). All the simulations have been run on a Linux server with 3.06 *GHz* Pentium Xeon processor and 2*GB* RAM. Hence, this sensitivity-based profit-aware gate sizing can be efficiently applied to improve design profit of large scale industrial circuits containing large number of gates. Note that we can apply incremental timing analysis (realized by incremental timing refinement considering only the gates with modified size) [2] for a large number of SSTA runs to further improve the runtime.

5. SIMULTANEOUS GATE SIZING AND OPTIMAL BIN PLACEMENT

In this section we present key observations on how the choice of frequency bin boundaries affects design profitability under process parameter variations. Next, we propose a design method, which maximizes design profit by simultaneous gate sizing and optimal bin-boundary placement.

5.1 Optimal bin placement

In case the bin boundaries are not available or designers are allowed to change it, they can be chosen appropriately such that the profit metric is optimized for a given price profile. Assuming fixed number of frequency bins (say N), given a delay distribution $D \sim N(\mu, \sigma)$ and price profile C, the problem of finding optimal bin boundaries can be expressed as:

$$\begin{aligned} &\text{Maximize } Y = \sum_{i=1}^{N} C(T_i)Y(T_i), \\ &\text{Subject to: } T_{leakage} \le T_i \le T_D, \text{ for } i = 1 \text{ to } N \end{aligned} \tag{11}$$

In order to solve the problem of optimal bin boundary determination (11), we start with equal yield bin boundaries as defined in (10). We employ a greedy approach and at a time we search for one bin boundary that optimizes the design profit keeping other boundaries fixed. First, we search the optimal boundary for the highest allowable frequency bin. We repeatedly perform such optimization for all other bins in descending order of bin frequencies. At the end, we obtain modified bin boundaries

```
findOptBinBoundary ( )
Input: Number of bins (N), delay params: μ,σ
Output: Optimal bin boundaries ($T_0$, … , $T_N = T_D$)
Find equal yield bin boundaries ($T_0$, $T_1$ …, $T_N$) for N bins
$P_{old}$ = P ($T_0$, $T_1$ …, $T_N$, μ, σ)
while (no change in $T_i$ )
    for each $T_i$ (0 < i < N)
        $P_{new+}$ = P ( …, $T_i$+ dT,…, $T_N$, μ, σ);
        dP+ = $P_{new+}$ - $P_{old}$;
        $P_{new-}$ = P ( …, $T_i$- dT,…, $T_N$, μ, σ);
        dP- = $P_{new-}$ - $P_{old}$;
        if (dP+ > 0)
            $T_i$ = $T_i$+ dT; $P_{old}$ = $P_{new+}$;
        else if (dP- > 0)
            $T_i$ = $T_i$- dT; $P_{old}$ = $P_{new-}$;
        end //if
    end //for
end //while
```

Figure 8: Pseudo-code to find optimal bin boundaries

for all the bins that locally optimize the profit. The pseudo-code for optimal bin boundary determination is given in Fig. 8. Results obtained from this algorithm match very closely with optimal bin placement results obtained from an exhaustive search using an implementation in MATLAB.

It is important to note that this technique does not have any design overhead. It only requires an extra design step to be incorporated after the final up/down sizing routine (Fig. 4).

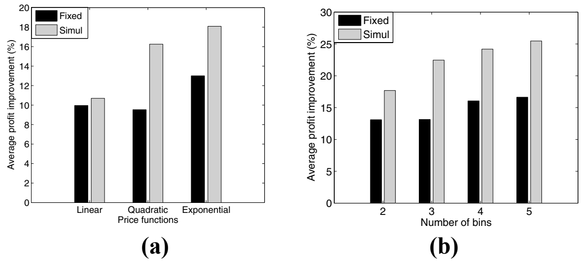
5.2 Simultaneous Sizing and Optimal Bin Placement

Finally, we integrate optimal bin placement procedure (section 5.1) and our profit-aware design flow (section 4.2) to develop an integrated design methodology that simultaneously perform gate sizing as well as optimal bin placement. Basic steps of the design flow are similar to that shown in Fig. 4, except that now we use optimal bin placement routine for profit computation during sensitivity analysis in each up/down sizing routine (step 2, Fig. 5). This means that after obtaining μ and σ corresponding to sizing of a logic gate we perform optimal bin placement to compute the profit sensitivity of the gate. This method when employed to different ISCAS85 benchmarks shows up to 36% profit improvement with three frequency bins (Table II), considering a leakage bound of μ - 2.5*σ for an 90% yield-optimized design at equal area. For all price profiles, additional improvements in profit with simultaneous sizing and bin placement over fixed bin boundaries are also shown in Table II under the columns labeled **Imp**. We observe that using integrated approach up to 10.1% (c3540) more improvement in profit over fixed bin boundaries can be achieved for exponential price profile. Note that profit improvements do not change much with optimal bin-placement for a linear price profile, since linear price profile has the smallest bin price ratio (i.e. R_{Lin}<R_{Quad}<R_{Expo}). However, the integrated approach shows significant average profit improvement compared to fixed bin boundary results for both quadratic (from 10% to 16.7%) and exponential price profiles (from 14.1% to 18.8%) for a set of ISCAS85 benchmarks (Table I, and II).

Fig. 9(a) shows effectiveness of two proposed profit-aware sizing methodologies (i.e. fixed bin and simultaneous sizing and optimal bin placement) for different price functions with N = 3 and $T_{leakage}$= μ - $2.5*\sigma$. Fig. 9(b) shows a consistent increasing trend in average profit improvements for both the methods under an exponential price profile as the number of bins is increased. Note that the runtime of the simultaneous sizing and optimal bin placement algorithm is almost indistinguishable from the runtime of profit optimization with fixed bin boundaries (Table I). This is because optimal bin placement routine has negligible impact on overall runtime, which is dominated by the number of SSTA runs. It is worth noting that although our delay models follow Gaussian (normal) distribution, the proposed methods can be easily extended to a recently-proposed non-normal delay distribution [12], which is reported to have higher accuracy in representing delay variations. With non-normal distribution, the steps on making sizing decision using profit sensitivity do not change. However, computation of effective yield in each frequency bin needs to be modified based on the nature of the distribution. Determination of optimal bin boundaries using the greedy approach section 5.1) also remains valid for non-normal delay distributions.

Table II: Simultaneous profit-aware sizing and optimal bin placement compared to 90% yield-optimized design (N = 3)

Circuit	Profit improvement (%)					
	R_{Lin}	Imp.	R_{Quad}	Imp.	R_{Exp}	Imp.
c432	9.98	0.33	15.58	6.38	17.98	5.18
c499	8.01	0.26	14.58	6.00	13.13	3.47
c880	17.01	0.73	22.63	7.47	27.17	6.28
c1908	7.13	0.11	8.32	1.82	9.52	0.45
c2670	8.59	0.48	13.17	6.02	14.68	4.36
c3540	20.97	2.19	30.00	11.96	36.31	10.13
c6288	14.02	1.04	20.69	8.65	26.07	8.18
c74181	7.66	0.39	11.93	5.44	13.30	4.02
c74L85	14.85	1.02	22.48	9.19	27.60	8.54
c74283	2.93	0.16	6.59	4.02	6.06	2.53
Avg.	**11.11**	**0.67**	**16.70**	**6.70**	**18.78**	**5.31**



(a) (b)

Figure 9: Average profit improvements in different methods (a) for different price functions by (N=3), and (b) for different N (exponential price profile)

6. CONCLUSIONS

We have proposed a profit-aware yield model and a statistical design methodology to optimize design profit for a given price profile under an area constraint. The methodology can be applied to any price profile and any delay distribution models (normal/non-normal). We have demonstrated that optimal bin-boundary determination can be used to increase the design profit. Experimental results on ISCAS85 benchmarks show that the proposed profit-aware design flow that incorporates information on price profile and frequency binning during the design phase can be very effective to improve design profit under parameter variations.

REFERENCES

[1] K. A. Bowman et al., "Impact of Die-to-Die and Within-Die Parameter Fluctuations on the Maximum Clock Frequency Distribution for Gigascale Integration", *JSSC,* 2002, pp. 183-190.

[2] L.-C. Chen et al., "A New Framework for Static Timing Analysis, Incremental Timing Refinement, and Timing Simulation", *ATS,* 2000, pp. 102-107.

[3] S. Borkar et al., "Parameter Variations and Impact on Circuits and Micro-architecture", *DAC,* 2003, pp. 338-342.

[4] B. Cory et al., "Speed binning with path delay test in 150-nm technology", *IEEE Design and Test of Computers*, 2003, pp. 41-45.

[5] R. R. Rao et al., "Parametric Yield Estimation Considering Leakage variability", *DAC, 20*04, pp. 442-447.

[6] S. Choi et al., "Novel Sizing Algorithm for Yield Improvement under Process Variation in Nanometer Technology", *DAC, 20*04, pp. 454-459.

[7] C. P. Chen et al., "Fast and Exact Simultaneous Gate and Wire Sizing by Lagrangian Relaxation," *IEEE TCAD*, 1999, pp. 1014-1025.

[8] K. Kang et al., "Statistical Timing Analysis using Levelized Covariance Propagation", *DATE,* 2005, pp. 764-769.

[9] "Technology Models", *http://www-device.eecs.berkeley.edu/~ptm.*

[10]A. Agarwal et al., "Circuit Optimization using Statistical Timing Analysis", *DAC,* 2005, pp. 321-324.

[11]T. Sakurai et al., "Delay Analysis of Series-connected MOSFET Circuits", *IEEE JSSC*, vol. 26, no. 2, 1991, pp. 122-131.

[12]X. Li et al., "Asymptotic Probability Extraction for Non-Normal Distributions of Circuit Performance", *ICCAD,* 2004, pp. 2-9.

[13]X. Bai et al., "Uncertainty-Aware Circuit Optimization", *DAC,* 2002, pp. 58-63.

[14] "Processor Retail prices", *http://www.newegg.com.*