Buffer Insertion Under Process Variations for Delay Minimization

ABSTRACT

This paper considers the buffer insertion problem under process variations. With continued technology scaling, it is necessary to model the physical parameters to be random variables. One approach to the buffer insertion problem under variations is to use the mean values of these parameters and solve the problem using traditional buffer insertion techniques for delay minimization. Another approach is to find a buffer insertion solution using a new method that can handle the probability distributions. Thus, the performance can be optimized with some yield constraint. In this paper, we present both analytical and experimental results to show that the two approaches give almost identical solutions. In other words, the more expensive statistical methods are not needed for the buffer insertion for delay minimization problem.

1. INTRODUCTION

With continued technology scaling, process variations have become a major factor that affects circuit performance and leads to excessive yield loss. Variations come from various sources. Geometric process variations are most significant [4]. Effective channel length L_{eff} can vary more than 50% because of the intra-die variations below 130nm technology node. The intra-die variations of interconnect geometric parameters also increase beyond 25%. Supply voltage and temperature variations are also becoming more prominent in nanometer design [5]. These variations can lead to 30% or more die-to-die performance difference.

Buffer insertion is a widely used interconnection optimization method. In nanometer designs, process variations become more and more significant. It is necessary to model the buffer and wire delay as random variables. Then there are two ways to find the optimal buffer insertion solution:

- One is to stick to the existing methods. Based on traditional deterministic algorithm for buffer insertion [6, 13], the mean values of delay distributions are used to guide the buffer insertion algorithm. This method totally ignores the variability of delays. But it can guarantee to minimize the mean delay value.
- The other way is to use new algorithms that can handle the probability distributions [1, 7]. For example, if the delay distributions are in Gaussian, it is reasonable to minimize the $\mu + 3\sigma$ value for the delay distribution. So optimal performance and maximum yield are both considered.

However, the statistical methods have much longer runtime. For nanometer designs, the number of buffers increases from thousands to millions. The new method will lead to more expensive design cost. If these two methods will achieve almost same solution, traditional buffer insertion algorithm is surely preferred.

Dynamic programming approaches have been proposed to find the optimal solution for buffer insertion [13]. When variations become a concern, statistical buffer insertion method was proposed [7]. However, in buffer insertion for delay minimization problem, coarse wire segmenting for dynamic programming algorithms will lead to a near optimal solution [3]. It indicates that the optimal solution is insensitive to the buffer position variation. We did further investigation on this problem considering process variations and get similar result: the buffer insertion for delay minimization is also insensitive to the process variations.

In this paper, we will theoretically analyze the buffer insertion under process variations. Results are derived to compare the cases with or without considering process variation. We also use both traditional and new statistical method to find the buffer insertion solutions. Both theoretical analysis and experimental results shows the buffer insertion problem is "immune" to the process variations. The deterministic buffer insertion method will give a near optimal solution for the buffer insertion considering process variations. Thus the more efficient traditional method is preferred. Process variations will not change buffer insertion solution in nanometer designs. It doesn't need to introduce new statistical buffer insertion method considering variations.

The rest of paper is organized as follows. In Section 2, process variation model for buffer insertion is stated. Theoretical analysis is applied based on this model in Section 3. In Section 4, we will compare experimental results from traditional and statistical buffer insertion algorithms. And we will conclude this paper in Section 5.

2. PROCESS VARIATIONS MODEL

The process variations, such as physical geometry variations, are usually modeled as probability distributions. Then, circuits' characteristics can be expressed as functions of these variations. For example, the delay d_k for circuit element k can be described as:

$$d_k = d_{inter-die}(p_i) + d_{intra-die}(x_k, y_k, p_i) \qquad (1)$$

where $d_{inter-die}$ is the delay distribution considering the inter-die variations. It can be well handled by traditional corner-based design methodology. $d_{intra-die}$ is the delay variation from within-die uncertainties. It could be a function of the position of element k, and p_i, which are parameters related to variations.

The inter-die variations can be formulated as deterministic values by traditional method, thus avoiding the statistical approaches. However, intra-die variations are now becoming more and more significant. Existing deterministic methods cannot model to these variations. Equation (1) can be further written as:

$$d_k = d_{k0}(p_{i0}) + \Delta d_{intra-die}(x_k, y_k, p_i) \tag{2}$$

where d_{k0} is the nominal value of delay. $\Delta d_{intra-die}$ are random variables with zero means. In most cases, $d_{intra-die}$ are not linear combinations of p_i. So even if the process variations can be modeled as Gaussian distribution, the intra die delays are not Gaussian. However, we can use first order approximation so the delays can still be Gaussian. Assume all parameters p_i are Gaussian variables with mean p_{i0}, and they are mutually independent. We have:

$$d_k = d_{k0}(p_i) + \sum_i \frac{\partial d_{k0}}{\partial p_i}\Delta p_i \tag{3}$$

where Δp_i are Gaussian distributions of parameters. And d_k is now a distribution in Gaussian.

If the delay distributions are Gaussian, it is easy to add two delays. The result is still Gaussian. Assume d_i, d_j are delay in Gaussian distributions with mean μ_i and μ_j and variation σ_i^2, σ_j^2 respectively. The correlation coefficient between d_i and d_j is $corr(d_i, d_j)$, and covariance $cov(d_i, d_j) = corr(d_i, d_j)\sigma_i\sigma_j$. $d_{sum} = sum(d_i, d_j)$ is still a Gaussian variable with:

$$\begin{aligned} \mu_{sum} &= \mu_i + \mu_j \\ \sigma_{sum}^2 &= \sigma_i^2 + \sigma_j^2 + 2cov(d_i, d_j) \end{aligned} \tag{4}$$

If d_i and d_j are independent variables, $cov(d_i, d_j) = 0$.

3. ANALYSIS ON BUFFER INSERTION UNDER VARIATIONS

Buffer insertion without consider variations could have closed form solution as proposed in [3]. Consider a delay model for buffer insertion as shown in Figure 1, assume that interconnect wire has resistance r_w and capacitance c_w per unit length. The buffer is a pair of cascaded inverters, with the minimum size inverter as input and a s size inverter to drive the load. The buffer has input capacitance C_L and output resistance R_d. We also assume the intrinsic delay for the buffer is τ_B. Elmore delay model is used here to analyze the delay of wire segment between two buffers. The length of the segment is l. And the delay t_d is from the first buffer input to the next buffer input, i.e, the delay of first buffer and the wire

$$t_d = \frac{1}{2}r_w c_w l^2 + (r_w C_L + R_d c_w)l + R_d C_L + \tau_B \tag{5}$$

Dividing it by l will give the delay value per unit length.

$$y = \frac{t_d}{l} = \frac{1}{2}r_w c_w l + (r_w C_L + R_d c_w) + \frac{R_d C_L + \tau_B}{l} \tag{6}$$

The optimal buffer insertion solution can be achieved by finding a length l s.t. t_d/l is minimized. This analytical result is useful to guide the buffer insertion for a long interconnect. We take the derivative of t_d/l and solve the equation below.

$$\frac{\partial y}{\partial l} = \frac{1}{2}r_w c_w - \frac{R_d C_L + \tau_B}{l} = 0 \tag{7}$$

The optimal length for buffer insertion is found to be:

$$l_{opt} = \sqrt{\frac{2(R_d C_L + \tau_B)}{r_w c_w}} \tag{8}$$

The $R_d C_L$ and τ_B in l_{opt} are determined by the device parameters. Using a linear transistor delay model, the intrinsic delay τ_B can be approximated as $R_{eff}C_{eff}$, where R_{eff} is the effective resistance of first inverter and C_{eff} is the gate capacitance of second stage. Compare to R_d and C_L, $C_{eff} = sC_L$ and $R_{eff} = sR_d$. A typical value for s is 5. This will lead to $\tau_B \gg R_d C_L$.

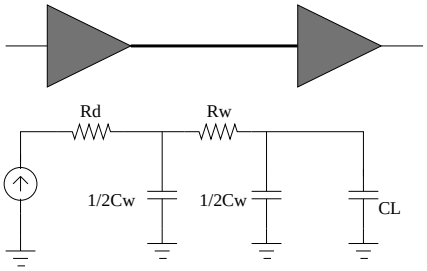


Figure 1: Interconnect model for buffer insertion

Considering the process variations, t_d is no longer a deterministic delay value. Under the variations, the capacitance c_w and resistance r_w are now random variable. Because of the large effective gate length variation, devices show more uncertainty than interconnection. So we also include the buffer delay variations in our calculation. Assume c_w, r_w and τ_B becomes Gaussian distribution under process variations:

$$\begin{aligned} c_w &= c_{w0} + \Delta c_w \\ r_w &= r_{w0} + \Delta r_w \\ \tau_B &= \tau_{B0} + \Delta \tau_B \end{aligned} \tag{9}$$

where c_{w0}, r_{w0} and τ_{B0} are the nominal values of wire capacitance, resistance and intrinsic buffer delay, respectively. Δc_w, Δr_w and $\Delta \tau_B$ are random variables in Gaussian with zero means. Now consider the process variations on c_w and r_w. Use t_{wd} to notate the delay distribution only considering the interconnect variations. First order approximation is used to calculate t_{wd}.

$$t_{wd} = t_{wd0} + (\frac{\partial t_d}{\partial c_w})_0 \Delta c_w + (\frac{\partial t_d}{\partial r_w})_0 \Delta r_w \tag{10}$$

where t_{wd0} is the mean value of delay. We can get t_{wd0} from Equation (5) by plug-in c_{w0}, r_{w0} and τ_{B0}. The derivations are also from Equation (5):

$$\begin{aligned} (\frac{\partial t_d}{\partial c_w})_0 &= \frac{1}{2}r_{w0}l^2 + R_d l \\ (\frac{\partial t_d}{\partial r_w})_0 &= \frac{1}{2}c_{w0}l^2 + C_L l \end{aligned} \tag{11}$$

The delay for the wire segment now is a random variable in Gaussian. So the delay per unit length t_{wd}/l is also in

Gaussian. It has mean value t_{wd0}/l. The variance of t_{wd}/l is:

$$\text{var}(\frac{t_{wd}}{l}) = \frac{\text{var}(t_{wd})}{l^2} \tag{12}$$

Thus, from Equations (10)-(12), we can calculate the variance of t_{wd}/l

$$\text{var}(\frac{t_{wd}}{l}) = (\frac{1}{2}r_{w0}l + R_d)^2\text{var}(c_w) + (\frac{1}{2}c_{w0}l + C_L)^2\text{var}(r_w) \tag{13}$$

We define σ_W as the standard deviation of t_{wd}/l. So:

$$\sigma_W = \sqrt{(\frac{1}{2}r_{w0}l + R_d)^2\text{var}(c_w) + (\frac{1}{2}c_{w0}l + C_L)^2\text{var}(r_w)} \tag{14}$$

The variance of c_w and r_w are mainly from the geometric process variations. Figure 2 shows the cross section of a wire. W is the wire width. S is the spacing between two adjacent wires. T is thickness of the wire. H is the spacing from wire to substrate. The capacitance c_w is sensitive to the variations of these parameters, especially W and S. r_w is sensitive to the variation of W and T. Moreover, the conductor resistivity ρ can have 18% variation in 100nm technology.

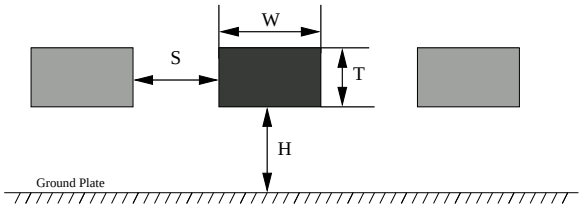


Figure 2: Physical model for interconnect

We can express the distribution c_w and r_w in terms of these process parameters. Closed-form equations $C_w(P_i)$ and $R_w(P_i')$ for calculated capacitance and resistance from those physical parameters are used.

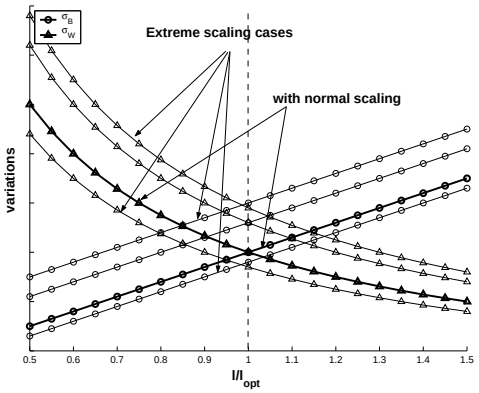
$$\begin{aligned} c_w &= c_{w0} + \sum_i (\frac{\partial C_w}{\partial P_i})_0 \Delta P_i \\ r_w &= r_{w0} + \sum_i (\frac{\partial R_w}{\partial P_i'})_0 \Delta P_i' \end{aligned} \tag{15}$$

where P_i are the parameters related to c_w, including W, H, T and S. P_i' are W, T and ρ, parameters related to r_w. var(P_i) is the measured variations of one parameter. In current or future process, these process variations may bring 20-30% changes for capacitance or resistance, which mean the 3σ for c_w and r_w could be 30% of their nominal value. So it's reasonable to approximate the variance for c_w and r_w as below:

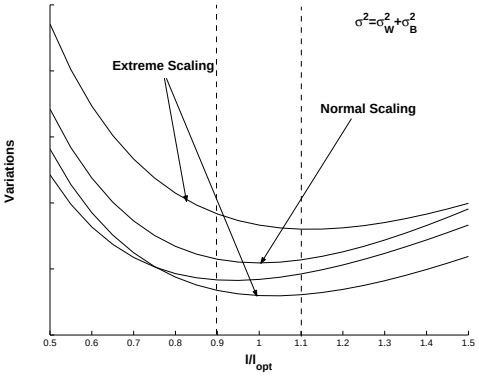
$$\begin{aligned} \text{var}(c_w) &\approx \alpha^2 c_{w0}^2 \\ \text{var}(r_w) &\approx \beta^2 r_{w0}^2 \end{aligned} \tag{16}$$

We assume the 3σ value of all P_i and P_i' to be 20%-30% of nominal value [2]. Then α and β are typically $0.1 - 0.15$ and relatively close to each other. Let $\gamma = \max(\alpha, \beta)$. Plug Equation (16) into Equation (14).

$$\sigma_W = \sqrt{\alpha^2(\frac{1}{2}r_{w0}c_{w0}l + R_d c_{w0})^2 + \beta^2(\frac{1}{2}c_{w0}r_{w0}l + r_{w0}C_L)^2} \tag{17}$$



(a)σ_W v.s. σ_B



(b)$\sigma^2 = \sigma_W^2 + \sigma_B^2$

Figure 3: The variation of t_d/l around the l_{opt}

Since C_L is relatively very small, we can approximate this equation as follows:

$$\begin{aligned} \sigma_W &\approx \sqrt{(\frac{\alpha}{2}r_{w0}c_{w0}l + \alpha R_d c_{w0} + \frac{\beta}{2}r_{w0}c_{w0}l + \beta r_{w0}C_L)^2} \\ &\approx \gamma(r_{w0}c_{w0}l + R_d c_{w0} + r_{w0}C_L) \end{aligned} \tag{18}$$

Consider Equation (5) and (18). If we want to minimize the $\mu + 3\sigma$ value of t_{wd}/l, we can take derivative of $t_{wd}/l + 3\sigma_W$ over l and we can get:

$$l'_{opt} = \sqrt{\frac{R_d C_L + \tau_B}{(0.5 + 3\gamma)r_{w0}c_{w0}}} \tag{19}$$

Compare l'_{opt} to l_{opt}:

$$\frac{l'_{opt}}{l_{opt}} = \sqrt{\frac{1}{1 + 6\gamma}} < 1 \tag{20}$$

This will lead to $l'_{opt} < l_{opt}$, which means more aggressive buffer insertion are needed when variations are considered. However, we have totally ignored the buffer delay variation in discussion above. In nanometer designs, devices have larger parameter variations and hence larger intrinsic delay variation than interconnects. Moreover, τ_B cannot be ignored to the wire delay t_w when the wire length is close to l_{opt}. Similarly, we assume the variance of τ_B is $\text{var}(\tau_B) = \theta^2\tau_B^2$. So considering the relation of τ_B and l_{opt}

in Equation (8) the standard deviation of τ_B/l is:

$$\sigma_B = \frac{\theta\tau_B}{l} \approx \frac{\theta(\tau_B + R_d C_L)}{l} = \frac{\theta r_{w0} c_{w0} l_{opt}^2}{l} \quad (21)$$

where θ is about $0.1 - 0.15$ for designs using 70nm-130nm technology [8]. For future technology θ could be even larger.

Now the variation of t_d/l could come from both interconnect variation and the device variation:

$$\text{var}(\frac{t_d}{l}) = \sigma^2 = \sigma_B^2 + \sigma_W^2 \quad (22)$$

Compare the σ_B to σ_W when $l = l_{opt}$. Depending on the scaling strategy, the ratio σ_W/σ_B may change. Normally, σ_W/σ_B will remain close to 1 for different technologies [9].

Now we compare the changes of σ_B and σ_W at $l = l_{opt}$.

$$\frac{\partial\sigma_W}{\partial l} = \gamma r_{w0} c_{w0} \quad (23)$$

$$\frac{\partial\sigma_B}{\partial l} = -\theta\frac{l_{opt}^2}{l^2} r_{w0} c_{w0} \quad (24)$$

σ_W is a linear function of l. The slope is $\gamma r_{w0} c_{w0}$. σ_B is a function proportional to $1/l^2$, as plotted in Figure 3

When l decreases from l_{opt}, the interconnect variations make the σ_W decrease linearly. However, shorter distance between two buffers makes the buffer delay variations more critical. σ_B will increase when l decreases. At $l = l_{opt}$, they increase or decrease at the same rate because γ and θ are close to each other. So the var(td/l) will almost remain unchanged if the buffer delay variations are considered. When l decreases more from l_{opt}, the increase of σ_B are dominant. We can see the variance of delay per unit length increases when $l \ll l_{opt}$.

Consider l increase from l_{opt}, σ_W increase linearly. σ_B will decrease with $1/l^2$. So the increase of σ_W is dominant. We can see the overall variance σ^2 increase when $l \gg l_{opt}$ in Figure 3.

In Figure 3, we also plot some extreme cases [9] for σ_W and σ_B. Consider the extreme case that either scaling only the devices or only the interconnections, the ratio σ_W/σ_B will vary from 2-0.67. So we plot these σ_W, σ_B and σ^2. The σ^2 in Figure 3(b) still have a relatively flat curve between $0.9l_{opt}$ to $1.1l_{opt}$.

From the previous discussion, we find the variance for the unit length delay will remain the same value when l is very close to l_{opt} because the decrease (or increase) of σ_B is similar to the increase (or decrease) of σ_W. However, when l becomes much larger or smaller than l_{opt}, σ^2 always increases from its value at l_{opt}.

In Figure 4, we plot the mean value and standard deviation of t_d with the parameters at 100nm technology (same as those in the Section 5). Elmore delay model is used and close-form equations in [10, 11] are used to calculate the mean and variance of delay. Other technologies also have the similar curves as in Figure 4.

From Figures 3 and 4, variance has a flat curve around the l_{opt} point. And it will increase when l becomes much larger or much smaller than l_{opt}. This indicates that the optimal solution of buffer insertion to minimize the mean delay value will also have minimum delay variability.

4. EXPERIMENTAL RESULTS

To verify the results in last section, we carry out the buffer insertion experiment with/without considering the process

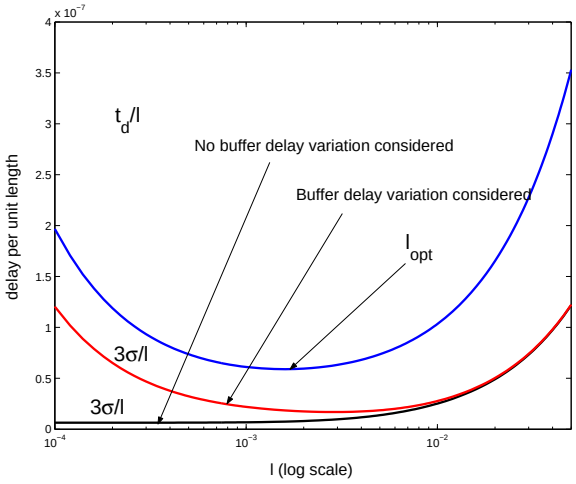


Figure 4: Mean value and standard deviation per unit length under 100nm technology parameters

variations. The buffer insertion problem for a pin-to-pin net can be formulated into a shortest path problem in directed acyclic graph (DAG) [6]. With process variations, it becomes a statistical shortest path problem and can be solved by the method proposed in [1].

Edge weight t_d is a function of physical parameters of the buffer and the wire. Sakurai's closed form equation [11] is used to calculate capacitance for the wire segment. Elmore's delay model is used to calculate the delay of wire segment. Drive resistance R_d and the load capacitance C_L are related to the buffer.

$$R_d = Vdd/I_{eff}$$

$$C_L = C_{ox} L_g W_g$$

where C_{ox} is gate capacitance of unit area, L_g and W_g are gate length and width, respectively. I_{eff} and intrinsic buffer delays are calculated by empirical equations in [10].

Those algorithms in [1, 6] can be extended to handle the buffer library. So we will consider the buffer insertion with different buffer types, which is different from the analysis in Section 3.

The parameters are chosen from [8]. To show the larger variation will give the same result, we set the larger parameters variations for some technologies as labeled 100nm and 70nm larger variations in Table 1. For these test data, we set the 3σ value to be 1/3 of the mean value. All nominal parameter values follow the data in ITRS01 [12].

We use the driver resistance $R_D = 1k\Omega$ and the load capacitance $C_L = 10fF$. Buffer library with 4X, 8X, 16X, 32X and 64X sizes is used to optimize the delay of 5-20mm long wires. 100 possible buffer locations are assumed to locate uniformly along the wire. We notice the buffer insertion solution will converge when the number of possible locations is larger than 50. Analysis in [3] shows the reason why the coarse grid also leads to an optimal solution.

We optimize the $\mu + 3\sigma$ delay with considering process variation first. An algorithm for statistical shortest path is used to find the optimal buffer insertion solution under process variations.

To compare to the statistical approach, we also run the traditional shortest path algorithm to minimize the mean delay to get the buffer insertion solution without considering process variation. Then, the μ and σ value for this buffer

	our method				worst case method				Comparison	
L	BUF#	$\mu(ps)$	$\sigma(ps)$	$\mu+3\sigma$(ps)	BUF#	$\mu(ps)$	$\sigma(ps)$	$\mu+3\sigma$(ps)	$\Delta\mu$	$\Delta(\mu+3\sigma)$
100nm technology										
5mm*	2	268.9	14.78	313.3	2	268.9	14.78	313.3	-	-
10mm*	3	527.5	24.12	599.9	3	527.5	24.12	599.9	-	-
20mm*	7	1040.0	36.6	1149.8	7	1040.0	36.6	1149.8	-	-
70nm technology										
5mm	2	248.1	13.9	289.9	2	248.0	14.1	290.2	+0.03%	-0.1%
10mm*	4	488.1	22.7	556.1	4	488.1	22.7	556.1	-	-
20mm*	9	962.4	33.7	1063.6	9	962.4	33.7	1063.6	-	-
45nm technology										
10mm*	10	536.7	18.7	592.9	10	536.7	18.7	592.9	-	-
20mm*	19	1065.2	27.9	1145.8	19	1065.2	27.9	1145.8	-	-
100nm technology with larger variations										
5mm*	2	268.9	17.0	319.9	2	268.9	17.0	319.9	-	-
10mm	3	528.4	27.4	610.6	3	527.5	28.4	612.6	+0.02	-0.003%
20mm*	7	1040.0	42.1	1166.2	7	1040.0	42.1	1166.2	-	-
70nm technology with larger variations										
5mm*	2	248.1	16.0	296.0	2	248.1	16.0	296.0	-	-
10mm*	4	488.1	25.5	564.5	4	488.1	25.5	564.5	-	-
20mm*	9	962.4	37.6	1075.1	9	962.4	37.6	1075.1	-	-

Table 1: Experimental results

insertion scheme is simulated with the same parameters.

Table 1 shows the results of different methods. μ and σ are mean and standard deviation of delay under process variations, respectively. We use "*" in first column if results from two method are identical.

From Table 1, we can find that the traditional method could have relatively smaller mean delay for some cases. It may also get a lager variation. However, the differences are very minor. In most cases two methods give identical solutions. So the traditional methods without considering parameter variations are still working when variations are becoming more dominant. The buffer insertion solution from existing algorithms will NOT lead to excessive yield loss.

5. CONCLUSION

In this paper, we proved that the buffer insertion problem is "immune" to the process variations. Based on our variation model, we analytically compared the optimal buffer insertion solution with or without considering process variation. They share the same buffer insertion results. We also carry out the experiments on buffer insertion by traditional and statistical algorithms. The results are still similar. Thus, we claim that process variations will not change buffer insertion solution in nanometer designs. It doesn't need to introduce new buffer insertion method considering variations for minimizing the 2-pin net delay.

6. REFERENCES

[1] Removed for blind review.

[2] K. Agarwal, D. Sylvester, D. Blaauw, F. Liu, S. Nassif, and S. Vrudhula. Variational delay metrics for interconnect timing analysis. In *DAC 2004*, pages 381–384. ACM Press, 2004.

[3] C. Alpert and A. Devgan. Wire segmenting for improved buffer insertion. In *Proceedings of the 34th annual conference on Design automation*, pages 588–593. ACM Press, 1997.

[4] D. Boning and S. Nassif. *Design of High-Performance Microprocessor Circuits*. 2002.

[5] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De. Parameter variations and impact on circuits and microarchitecture. In *DAC 2003*, pages 338–342. ACM Press, 2003.

[6] Y. Gao and D. Wong. A graph based algorithm for optimal buffer insertion under accurate delay models. In *Proceedings of the conference on Design, automation and test in Europe*, pages 535–539. IEEE Press, 2001.

[7] V. Khandelwal, A. Davoodi, A. Nanavati, and A. Srivastava. A probabilistic approach to buffer insertion. In *Computer Aided Design, International Conference on*, pages 560– 567, 2003.

[8] S. Nassif. Delay variability: sources, impacts and trends. In *Solid-State Circuits Conference, Digest of Technical Papers*, pages 368–369, 2000.

[9] S. R. Nassif. Modeling and forecasting of manufacturing variations (embedded tutorial). In *Statistical Metrology, 2000 5th International Workshop on*, pages 2–10, 2000.

[10] M. Orshansky, L. Milor, P. Chen, K. Keutzer, and C. Hu. Impact of spatial intrachip gate length variability on the performance of high-speed digital circuits. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on,*, 21, May 2002.

[11] T. Sakurai. Closed-form expressions for interconnection delay, coupling, and crosstalk in vlsis. *Electron Devices, IEEE Transactions on*, 40(1):118–124, 1993.

[12] Semiconductor Industry Association. *International Technology Roadmap for Semiconductors*, 2001.

[13] L. van Ginneken. Buffer placement in distributed rc-tree network for minimal elmore delay. In *Inrl. Symp. Circuits and Systems*, pages 865–868, 1990.