A Probabilistic Analysis of Pipelined Global Interconnect Under Process Variations

Navneeth Kankani Vineet Agarwal Janet Wang
Electrical and Computer Engineering Department
The University of Arizona, Tucson, AZ
{kankani, vagarwal, wml}@ece.arizona.edu

***Abstract*— The main thesis of this paper is to perform a reliability based performance analysis for a shared latch inserted global interconnect under uncertainty. We first put forward a novel delay metric named DMA for estimation of interconnect delay probability density function considering process variations. Without considerable loss in accuracy, DMA can achieve high computational efficiency even in a large space of random variables. We then propose a comprehensive probabilistic methodology for sampling transfers, on a shared latch inserted global interconnect, that highly improves the reliability of the interconnect. Improvements up to 125% are observed in the reliability when compared to deterministic sampling approach. It is also shown that dual phase clocking scheme for pipelined global interconnect is able to meet more stringent timing constraints due to its lower latency.**

I. Introduction

In multi-gigahertz system-on-chip designs, global interconnect wire proves to be a major bottleneck to the continual increase in clock frequency as it requires several clock periods of time to propagate a signal from source to sink. Extensive amount of work has been done in recent past to alleviate the global interconnect from the clock frequency constraints. And, it is an accepted fact that pipelining of global interconnects offers a promising solution to this problem [1], [2].

Although pipelining increases the throughput of interconnect, it comes with practical difficulties like the cycle level behavior changes in the RTL level which requires a lot of manual rework in design. Scheffer in [3] listed the challenges faced by the designer while pipelining an interconnect. In [4] Cong et al addressed the problem of automatic interconnect pipelining at RTL level by proposing a RDR-pipe(Regular Distributed Register) approach to efficiently support the multi-cycle on chip communication with interconnect pipelining.

For instance, consider a scheme where a single pipelined interconnect is shared between 3 computational units that need to transfer the data to another computational block. In a real-time system operated at a clock period of $285ps$, let the arrival time set of these 3 transfers to be (170, 350, 690)ps; each having a corresponding deadline. Now, if these transfers are scheduled to be sent on the interconnect such that at most one transfer is issued per clock cycle, then they can share the same interconnect, provided their respective deadlines are met.

However, it should be noted that the uncertainties in the arrival times and interconnect delay due to manufacturing variations causes considerable bit error rate (BER) on the global interconnect. For the above example, we observe that deterministic sampling of transfers at the μ or $(\mu + 3\sigma)$ values, can give an error rate as high as 59%, which is highly undesirable. However, there may exist other set of sampling times which can yield much higher reliability (as high as 99%). Hence, there is a need of a formal methodology that can find such sampling sets to address the problem of increasing error rates on shared global interconnects.

To this purpose, we propose a comprehensive probabilistic methodology for sampling transfers, to increase the reliability of shared latch inserted global interconnect under process variations. The proposed technique meets the timing constraints of the global interconnect with a much higher reliability compared to the deterministic sampling approach. Note that the overall reliability of each transfer depends upon its sampling reliability and transmission reliability. The latter depends upon the BER encountered by transfer on the pipelined interconnect. In order to compute the BER , we perform the statistical timing analysis on the interconnect. While doing so, we support a simple dual phase clocking scheme for pipelined global interconnect, as it proves to be robust in noisy environment. We also propose a novel delay metric based on ANOVA (DMA) for estimating the probability density function of the interconnect delay. It is an efficient and accurate metric when compared to other variational delay metrics and provides the designer with an explicit polynomial approximate expansion for the delay response.

To begin, we explain the new delay metric for interconnect and providing the results to prove its efficiency compared to other delay metrics in Section II. A probabilistic methodology for sampling transfers, to reliably transmit the data on a shared global interconnect is presented in Section III. The results of our technique are listed in Section IV and Section V concludes our paper.

II. Delay Metric based on ANOVA (DMA)

DMA serves as an efficient metric for finding the pdf of the global interconnect delay considering process variations. Given the uncertainty in parameters and degree p of the required model, the prototype DMA returns the p^{th} degree polynomial for the delay of the global interconnect along with its mean and variance.

In our formulation, we first approximate the delay response of a single RC model of an interconnect as a function of uncertainty in geometric parameters such as wire width(w), spacing(s), thickness(t) and inter-layer dielectric thickness(h). Our aim is to find a computationally efficient approximation $\hat{y}$ of actual output ($y = f(w, s, t, h)$), with a very small error margin.

$$\hat{y} = \hat{f}(w, s, t, h) \quad (1)$$

0-7803-9451-8/06/$20.00

where $\hat{f}$ is a finite order polynomial function that approximates the behavior of the model and w, s, t, h are random variables that can be expressed in terms of zero mean and unit variance vector $\overline{\xi} = (\xi_1, \xi_2, \xi_3, \xi_4)$ such that

$$\begin{array}{lcllcl} \text{w} & = & \mu_w + \sigma_w\xi_1 & \text{s} & = & \mu_s + \sigma_s\xi_2 \\ \text{t} & = & \mu_t + \sigma_t\xi_3 & \text{h} & = & \mu_h + \sigma_h\xi_4 \end{array}$$

In this work, we assume the pdf of geometric variations to be gaussian. Thus,without the loss of generality, this delay metric can be used for any random distribution, and hence we can approximate the output using a set of hermite polynomials [5]. Thus equation 1 can be written as

$$\hat{y} = \sum_{i=0}^{N} \alpha_i g_i\left(\overline{\xi}\right) \tag{2}$$

where $g(\overline{\xi})$ is a set of hermite polynomials, $\overline{\alpha}$ represents a coefficient vector, N is the order of approximation. To compute $\overline{\alpha}$ we use probabilistic collocation method (PCM) proposed in [6].

The polynomial constructed in (2) is expressed in terms of the input parameters and their interactions, not all of them may be significant in the approximation of the response. In fact, considering a large space of random parameters there may be certain parameters that have negligible effect on the response. Therefore we advocate that by detecting and eliminating such parameters from our design, we can reduce the computational complexity involved in evaluating the response without significant loss of accuracy.

We apply a technique named ANOVA (Analysis of Variance)[7] on this RC model to quantify the importance of variables on the variability of the response. As its name suggests, ANOVA analyzes the variances to test for significant differences between means by partitioning the total variability into component parts. The proportion of variance due to each input (or its correlation) towards the total variance can be used as a statistical significance parameter (F) of that particular input.

A. Overview of ANOVA

The statistical significance parameter (F) can be computed by applying the underlying notion of ANOVA. We explain the notation and implementation details by considering a simple model with single uncertain input ξ and response $y = f(\xi)$. As the input is random, let us assume that we consider 'm' different values for the input variable at which we will observe the response. Then the observed response 'y_{ij}' represents the j^{th} observation taken under i^{th} instance of ξ. The observation can be described by the linear statistical model as

$$\begin{array}{l} y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad i = 1, 2, \ldots, m \\ \qquad\qquad\qquad\quad\ \ j = 1, 2, \ldots, n \end{array} \tag{3}$$

μ is the overall mean, τ_i is called i^{th} instance effect and is unique to that instance, and ε_{ij} is the error component. The basic idea in ANOVA is the comparison of the variance in the response due to intra-instance and inter-instance variability. The null hypothesis defined as inter-instance variability (σ_t^2) is zero, implies that the response means are same for different instances.

$$H_0 : \sigma_t^2 = 0 \tag{4}$$

$$H_1 : \sigma_t^2 > 0 \tag{5}$$

where σ_t is variance of τ_i. $\sigma_t > 0$ implies that the variability exists between 'm' instances.

In order to calculate variance, we define between-instance sum of squares(SSF) and within-instance sum of squares (SSE) such that:

$$\text{SSF} = n\sum_{i=1}^{m}(\overline{y}_{ei} - \overline{y}_t)^2 \tag{6}$$

$$\text{SSE} = \sum_{i=1}^{m}\sum_{j=1}^{n}(y_{ij} - \overline{y}_{ei})^2 \tag{7}$$

$$\overline{y}_{ei} = \left(\sum_{j=1}^{n} y_{ij}\right)/n \tag{8}$$

$$\overline{y}_t = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} y_{ij}\right)/N \tag{9}$$

Here $N = mn$ is the total number of observations. $\overline{y}_{ei}$ represents the average of the observations taken under i^{th} instance whereas $\overline{y}_t$ represents the average of observations taken under all instances. The significance of SSF is that it explains the variability in response due to difference in mean of instances whereas SSE is referred to as error variance. The variance or the mean square for between and within instance is defined as:

$$\text{MSF} = \text{SSF}/(m-1) \tag{10}$$

$$\text{MSE} = \text{SSE}/(N-m) \tag{11}$$

To test the null hypothesis H_0, a F-ratio is defined as F = MSF/MSE. If the null hypothesis is true, then both MSF and MSE estimate the same quantity and thus F-ratio must be 1. Assuming the observations are normally distributed, it can be shown that SSF/σ^2 and SSE/σ^2 are independently distributed chi-square random variables [8]. Thus, if null-hypothesis is true then the F-ratio must also be chi-square distributed with ($m-1$,$N-m$) number of samples. We can find this F-ratio (F_0) using a look up table [7] for a given significance level (α). The probability P of obtaining the computed F-ratio (MSF/MSE) greater than F_0 is used as a significance parameter such that the null hypothesis is rejected if P is lower than α (set here as 0.05). Another important quantity that is used in determining the proportion of variability in response explained by the model is defined as R^2 = SSF/SST.

B. Reduction using ANOVA

We apply ANOVA technique on the polynomial generated by PCM in equation (2) to find the insignificant terms in the model. The analytical expression for the delay of single RC segment, generated by PCM is given as:

$$\begin{array}{rl} delay = & 19.65 - 2.28\xi_1 - 0.9\xi_2 - 1.82\xi_3 - 0.32\xi_4 \\ & +0.28(\xi_1^2 - 1) + 0.1(\xi_2^2 - 1) + 0.12(\xi_3^2 - 1) \\ & +0.05(\xi_4^2 - 1) + 0.17(\xi_1\xi_2) + 0.03(\xi_1\xi_4) \\ & +0.2(\xi_2\xi_3) - 0.17(\xi_2\xi_4) + 0.17(\xi_3\xi_4)\ ps \end{array} \tag{12}$$

TABLE I
COMPARISON OF MONTE CARLO(MC),PCM AND DMA RESULTS

Number of segments	Mean delay (ps)			Delay Variation (ps)			Number of Spice Runs			Error%	
	MC	PCM	DMA	MC	PCM	DMA	MC	PCM	DMA	μ	σ
2	40.06	39.92	40.11	3.59	3.45	3.63	10000	45	38	0.12	1.11
4	163.67	163.94	164.04	9.24	9.37	9.42	15000	153	86	0.22	1.94
8	496.36	496.66	494.62	26.59	26.36	27.16	20000	561	261	0.35	2.15
16	1616.29	1617.15	1625.82	43.45	42.81	44.86	25000	2145	912	0.59	3.26

The mean and standard deviation of the delay are $19.62ps$ and $3.15ps$ respectively. At first, we apply a primary level of screening to determine the individual effect of each ξ_i on the delay. Figure 1 shows the individual significance of w, s, t, h in the delay of a RC segment which is used in primary level of screening. We then compute the delay gradient of individual effects of ξ_i. The set of ξ_i, for which the delay gradient is below a certain threshold, is used in secondary level of screening. For example, it is noted that h has negligible effect on delay compared to other parameters. And thus, h will be passed on to secondary level of screening.

We use ANOVA on the Equation (12) in the secondary level of screening to remove insignificant terms. Removing the insignificant terms(Ω), we generate a reduced analytical equation such that its R^2 value is at least 98.5%. In this case ANOVA gives us that the terms $\{\xi_4, \xi_2^2, \xi_4^2, \xi_1\xi_2, \xi_1\xi_3, \xi_1\xi_4, \xi_2\xi_4, \xi_3\xi_4\}$ are insignificant and the ANOVA table for the corresponding reduced model is given in Table II. The reduced analytical equation is of the form

$$\begin{aligned} delay = \quad & 19.65 - 2.28\xi_1 - 0.9\xi_2 - 1.82\xi_3 + 0.28(\xi_1^2 - 1) \\ & +0.12(\xi_3^2 - 1) + 0.2(\xi_2\xi_3)\ ps \end{aligned} \tag{13}$$

which has a mean and standard deviation of $19.64ps$ and $3.13ps$ respectively, there by giving a mere error of 0.02% in mean value and 1.2% in standard deviation. The definitions of the terms in Table II can be found in Section II-A.

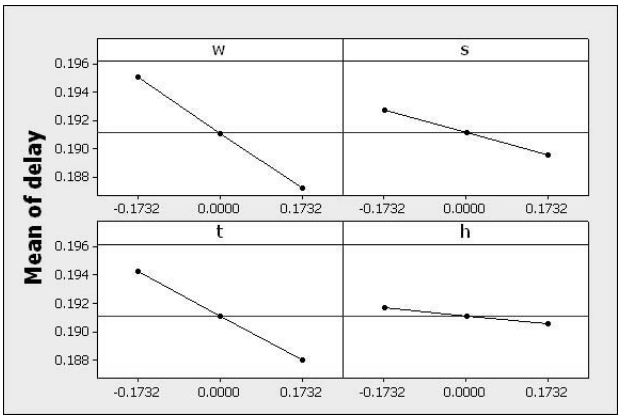

Fig. 1. Significance of geometric parameters on delay

C. DMA Implementation

To begin our DMA analysis, we divide the global interconnect into smaller identical modules (Figure 2) where each smaller module has n random variables. To find the delay equation for interconnect, we use PCM technique in which, the

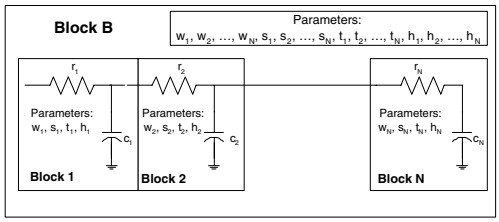

Fig. 2. RC tree Model for Global Interconnect

TABLE II
ANALYSIS OF VARIANCE FOR DELAY

Source	Υ	Sum of Squares	Mean Squares	F value	p
Model	7	454×10^{-6}	76×10^{-6}	8059.65	0.0
Residual Error	8	12×10^{-9}	26×10^{-10}	5.28	
Total	15			R^2 = 99.73%	

Υ →Degree of Freedom

collocation points are computed for these n random variables, for a given degree p.

An important step in DMA is analysis of this smaller blocks using ANOVA to determine the parameters that have insignificant effect on the variability of the response. This is performed by the technique outlined in Section II-B. As all the segments of the macro-model (Block B) are identical we can extrapolate the information about insignificant parameters of the other smaller blocks based on the ANOVA results of the first block. A set of all insignificant parameters $\Gamma = \{\Omega_1 \cup \Omega_2 \ldots \cup \Omega_n\}$ is then used in evaluation of response of larger block. It should be noted that DMA preserves significant correlations among the design parameters. Since we know from [9] that the computational complexity is directly proportional to the number of significant terms in the model, the information in Γ can be used to decrease the complexity. Hence, we perform the model runs on only those input sampling points that correspond to the significant parameters in Block B. In this way, by hierarchically removing the insignificant terms and thus reducing the model runs, the computational complexity of DMA is decreased. A comparison of delay values of Monte Carlo, PCM and DMA is done in Table I.

III. RELIABILITY AWARE GLOBAL INTERCONNECT SHARING

The latch inserted global interconnect can be shared between two computational units in order to achieve higher levels of resource utilization. As shown in Figure III, the three transfers from different computational units in Block A can be transferred to Block B using a shared pipelined interconnect.

Each block has a sampler (multiplexer) that sends the transfers on the interconnect according to a priori sampling times such that at most one transfer is issued per clock cycle on the interconnect.

A feasible schedule would be the one in which each transfer starts after its arrival time and completes before the specified deadline. While constructing such a schedule of transfers on the interconnect, it is also necessary to keep the BER below a minimum allowable value, to ensure correct signal transmission in the presence of uncertainty in the arrival times and global interconnect delay. Hence, in this section we formalize a probabilistic methodology to reliably transmit the data on the interconnect with the consideration of process variations. Before discussing our approach, the basic notation

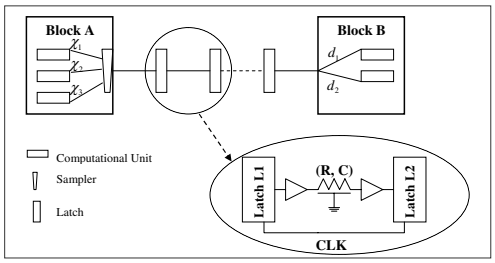

Fig. 3. Global Interconnect Sharing Stage

and terminology is presented.

- $\overline{\chi}$: Arrival Times$\rightarrow N(\overline{\mu}_A, \overline{\sigma}_A)$
- $\overline{\mu}_A$: Mean
- $\overline{\sigma}_A$: Standard Deviation
- $\overline{\lambda}$: Deadline Time Vector
- clk : Clock period of Global Interconnect
- n : Number of Pipelined stages
- $\overline{\zeta}$: Sampling Times
- $\overline{D}$: Delivery Times
- σ_p : Standard Deviation of Sampler
- $\overline{\omega}$: Bit Error Rate
- $\overline{\psi}_s$: Sampling Confidence
- $\overline{\psi}_b$: Transmission Confidence = $(1 - \overline{\omega})$
- $\overline{\psi}_t$: Total Confidence Level = $\overline{\psi}_s * \overline{\psi}_b$
- Ψ_f : Root Mean Square $(\overline{\psi}_t)$

We assume that each arrival time vector is associated with a given deadline time vector where deadline time is defined as the latest time by which the data must be received by the other end of the interconnect. The sampling time vector is the time when the data is sent over interconnect by the sampler. The sum of sampling time and the interconnect latency is termed as delivery time.

Definition 1: (Sampling Confidence = ψ_s): The Sampling Confidence level ψ_s of x is defined as

$$\psi = 100 * \int_{-\infty}^{\pi_\psi} f(x)dx \tag{14}$$

where f(x) is the pdf of x and π_ψ $(or\ \pi_\psi(x))$ is ψ^{th} percentile of x, meaning that a designer can be $\psi\%$ assured that the random parameter x will be less than π_ψ $(or\ \pi_\psi(x))$.
Transmission confidence $(\overline{\psi}_b)$ gives us the reliability of transmitting data over pipelined interconnect considering the effect

Algorithm 1 probabilistic Scheduling $(clk, \overline{\chi}, \overline{\lambda}, \omega_{min}, \sigma_P)$

```
  τ̄ ← MOD(μ̄_A, clk) − 0.5 * clk      γ̄ ← N(τ̄, σ̄_A)
  p ← LUT(clk, σ_p)                   α ← QUOTIENT(μ̄_A, clk)
{ᾱ_new, δ̄} ← COMPUTESLACK(ω_min, α)
if ∃i such that δ_i < 0 then
  β̄ ← FINDSAMPLINGCYCLE(α_new, δ̄)
  if ∃(i, j, i ≠ j) such that β_i = β_j then
    for i = 1 to N do
      if α_i = β_i then
        if π_99(γ_i) ≤ p then
          k ← 3
        else if π_50(γ_i) ≤ p then
          k ← SOLVE(τ_i + kσ_i = p)
        else
          k ← SOLVEFORMAXIMUM(ψ_{t_i}(k))
        end if
        ζ_i ← μ_i + kσ_i
      else
        ζ_i ← ((β_i − 0.5) * clk + p)
      end if
    end for
    ω̄ ← LUT(clk, ζ̄)
    Ψ_f ← COMPUTECONFIDENCE(ζ̄, ω̄)
  else
    print 'Abort: Deadline Constraint Violation'
  end if
else
  print 'Error: No feasible schedule possible'
end if
return (Ψ_f, ζ̄)
```

of process variations. It is computed using BER encountered while transferring data over interconnect. We calculate the root mean square confidence (Ψ_f) and choose it as an optimization criteria because it maximizes both sampling confidence and transmission confidence (which in-turn means BER (ω) is minimal).

A. Problem Statement

Given a set of n arrival times $\overline{\chi} = \{\chi_1, \chi_2, \ldots, \chi_n\}$, their corresponding deadlines $\overline{\lambda} = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, and a global interconnect clock period clk, the sampling time set $\overline{\zeta} = \{\zeta_1, \zeta_2, \ldots, \zeta_n\}$ is found such that the Ψ_f is maximized. The problem can thus be formulated as formulated as:

$$\begin{array}{lrcll} & & \max & \Psi_f & \\ \texttt{subject to:} & \omega_i & \leq & \omega_{min} & \forall i \in n \\ & \mathrm{d}_i & \leq & \lambda_i & \forall i \in n \\ & \psi_{s_i} & \geq & 50\% & \forall i \in n \end{array}$$

Our goal is to maximize the total confidence Ψ_f so that the timing constraints are met with some guaranteed probability in the presence of process variations. In order to maximize Ψ_f, we have to optimize sampling confidence ψ_s.

It is evident from the third constraint ($\psi_{s_i} \geq 50\%$) that we want the sampling confidence of each arrival time to be higher than 50%. For simplicity, we hereafter refer to each arrival time as an event to be scheduled on the interconnect. Note

that the overall reliability of each transfer depends upon its sampling reliability and transmission reliability. The sampling reliability is evaluated using the confidence level at which the sampler selects the arrival times. The latter depends upon the BER encountered by transfer on the pipelined interconnect. In order to compute the BER associated with each event, we need to perform the statistical timing analysis on the interconnect which is found using the following section.

B. Bit Error Rate Computation

The statistical timing analysis is performed on the latch inserted global interconnect for computation of BER as formulated in [10]. The notations which are used in section are also kept same. For a given latch of stage i, the opaque region (R_{opaque}) is defined as the time at which clock to the latch is low. And the region of high-clock where data can be sampled correctly by latch is termed as transparent Region (R_{tran}). All the other region of clock apart from R_{opaq} and R_{tran} is termed as faulty region (R_{faulty}). Based on the region where the propagation delay of previous stage lies, the propagation delay of stage i is written as:

$$p_i = \begin{cases} \tau_{wire} + \tau_{data} - T_{clk} & p_i \in R_{opaque} \\ p_{i-1} + \tau_{wire} + \tau_{prop} - T_{clk} & p_i \in R_{tran} \end{cases} \quad (15)$$

Using, the pdf of propagation delays the total probability of correct transmission and thereby the BER on the pipelined interconnect can be found using the following equation:

$$BER = 1 - \underbrace{q_1 q_2 \dots q_N}_{N} = 1 - \prod_{i=1}^{N} q_i$$

where q_i is the probability that stage i of the pipeline will transmit the data correctly.The details for the notation and analysis can be found in [10].

With the use of single phase clocking in latches, the flexibility in timing provided by latch based methodology is not fully utilized. A dual-phase clock system can be used instead of single phase clocking. The advantages of using such a scheme are reduced latency, clock skew tolerance and higher performance. Nevertheless, these advantages come along with area overhead for generating two different clocks.

C. Estimation of Interconnect Reliability

Once the BER is computed, we build a Look-up table (LUT) for a given clock frequency of the pipelined global interconnect and the delay variation of the sampler (σ_s). The inputs to our algorithm are the pdf of the events, their deadlines, and the maximum clock frequency at which the pipelined interconnect can be operated for a given BER (ω_{min}). At this clock frequency, and for a particular sampler delay variation (σ_s), we find the value p using LUT (figure 4, where p is the mean arrival time beyond which BER starts increasing and approaches one. Thus, it is important to select an optimum value of p for obtaining minimum BER.

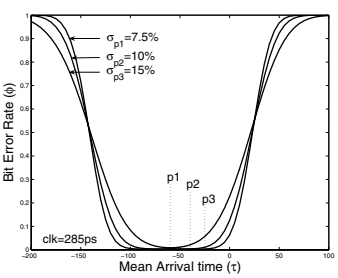

Fig. 4. Graph showing maximum allowable arrival time

The throughput of pipelined interconnect is one event per clock cycle for a single phase clocking scheme. Hence, we must find a sampling clock number for each event such that there is at least one clock period difference between any two events in the set to guarantee no overlap in pipelined stages. We define $\alpha = \texttt{quotient}(\overline{\mu}_A, clk)$ as the earliest sampling clock number for an event. If $\omega_i > \omega_{min}$ for this α_i, we assign $\alpha_{new,i} = \alpha_i + 1$. Based on the deadline ($\overline{\lambda}$) and number of pipelined stages, we compute slack $\overline{\delta}$ which is defined as:

Definition 2: (Slack = $\overline{\delta}$): It is the difference between the deadline time and the delivery time,that is,the absolute time at which the event reaches the other end of interconnect.

$$\overline{\delta} = \left(\overline{\lambda} - \overline{\alpha}_{new} - n * clk\right) \quad (16)$$

The foremost requirement for any event to be eligible for scheduling is that the $\overline{\delta}$ must be non-negative because it otherwise violates the deadline constraint. We then compute the actual sampling clock number β using available slack for each event, such that there is no overlap in clock number among any two events. If there is an overlap of actual sampling clock numbers, we conclude that there cannot be a feasible schedule that meets the constraint of minimum BER (ω_{min}). Now, for this sampling clock number (β) we want to find an optimum sampling time (ψ_s) for each event in order to maximize Ψ_f. It is assumed that the sampler that is used can select these ψ_s from the continuous arrival time pdf.

If α_i remains unmodified ($\alpha_i = \beta_i$), then we calculate the sampling time $\overline{\zeta} = \overline{\mu}_A + k\sigma_A$, where k is found on the basis of 50^{th} and 99^{th} percentile values of pdf of event set and p. The value of k is computed for the three cases as shown in the algorithm. For instance, when 50th% $\geq p$, we solve for k using LUT for finding transmission confidence (ψ_b) and sampling confidence simultaneously, until ψ_t is maximized. And in case α is modified (provided there is slack available), $\overline{\zeta}$ is found using sampling clock number (β) and mean arrival time (p). Once the sampling time ($\overline{\zeta}$) is computed, we check for the BER (ω) at these sampling times. The total root mean square confidence is finally computed using sampling times and BER.

IV. Results

We perform our analysis on global interconnect based on the $0.18\mu m$ technology parameters given by Berkeley PTM model [11]. There are 8 pipelined stages and the length of wire in each stage is taken to be 1.4μm. The mean and variance of delays are computed using DMA and are listed in Table IV.

Next, we use Section III-B to find the maximum operable clock clk so that $\omega_i < \omega_{min}$. We proceed in our analysis using this clock period and assuming that $\omega_{min} = 2\%$. For comparison of deterministic sampling and our approach, we consider three cases of 4 randomly generated events that needs to be transferred between two blocks. The 3 cases correspond to $\overline{\delta} > 0$, $\overline{\delta} = 0$ and $\overline{\delta} < 0$. Without the loss

TABLE III
COMPARISON OF DETERMINISTIC AND PROBABILISTIC SCHEDULING

	μ_A	λ	δ		ζ		Ψ_f						Gain %			
							μ Sampling		$\mu+3\sigma$ Sampling		probabilistic Sampling		μ Sampling		$\mu+3\sigma$ Sampling	
	ps	ps	SP	DP	SP	DP	SP	DP	SP	DP	SP	DP	SP	DP	SP	DP
Case 1																
E1	70	2850	1	1	385	650										
E2	395	3700	3	1	1240	1110	44.64	43.13	69.99	70.62	99.47	99.79	122.77	131.37	42.09	41.29
E3	690	3450	1	2	955	1570										
E4	1020	4100	2	2	1810	2030										
Case 2																
E1	70	2600	0	0	90	90										
E2	395	2900	0	0	405	650	49.47	40.65	0	81.54	73.52	99.74	48.62	145.36	-	22.30
E3	690	3150	0	1	695	1110										
Case 3																
E1	70	2600	0	0	-	90										
E2	395	3450	2	2	-	1570	×	43.13	×	0	×	89.38	-	107.23	-	-
E3	690	2850	-1	0	-	710										
E4	1020	3100	-2	0	-	1165										

SP → Single Phase clocking DP → Dual Phase clocking × → Deadline Constraint Violation

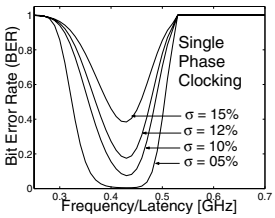

Fig. 5. Effects of σ_{clk} on BER for single phase scheme

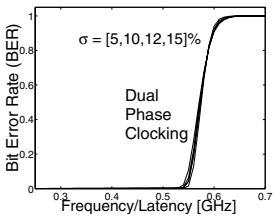

Fig. 6. Effects of σ_{clk} on BER for dual phase scheme

of generality, the events are selected such that they cover all possible regions (opaque, transparent, faulty) of sampling within a clock period. The results are tabulated in Table III. We consider two types of deterministic samplings which are 50% and 99% sampling. 50% and 99% samplings corresponds to sampling at μ and $\mu+3\sigma$ values respectively. For deterministic and probabilistic sampling, we show the results when either single phase or dual phase clocking is used. It should be noted in Case 3 that for the same set of arrival times and deadlines, the single phase clocking succumbs because the deadline constraint does not meet. However, its dual phase clocking counterpart gives a feasible schedule with a much higher confidence than mean sampling value. This is possible because the latency of shared global interconnect in dual phase clocking scheme is lower than that of single phase clocking. Thus that more stringent timing constraints can be met with dual phase clocking yielding a low error rate. Furthermore a comparison of dual and single phase clocking, in Figure 5 and 6, for various values of σ_{clk} proves that dual phase is more robust in noisy environments.

TABLE IV
DELAY VALUES USED FOR 0.18μM TECHNOLOGY

	DMA Delay	
	$\mu(ps)$	$\sigma(ps)$
τ_{wire}	109.12	9.13
τ_{data}	119.60	13.19
τ_{prop}	115.42	13.62
τ_{setup}	119	13
T_{clk}	μ_c	22

V. CONCLUSIONS

The major contribution of this work is two-fold. First, we provided an efficient and accurate delay metric DMA for estimating the delay pdf of an interconnect under process variations. This delay metric uses ANOVA technique to reduce its computational complexity. Then, a probabilistic methodology for sampling events to reliably transmit the data on the shared latch inserted global interconnect was presented. Using our approach, the error rates are dramatically reduced and significant improvements are observed in total confidence compared to 50^{th} or 99^{th} percentile deterministic sampling approach.

REFERENCES

[1] R. McInerney, K. Leeper, T. Hill, H. Chan, B. Basaran, and L. McQuiddy, "Methodology for repeater insertion management in the rtl layout, floorplan and fullchip timing databases of the itanium microprocessor," in *Proc. of 2000 International Symposium on Physical Design*, 2000.

[2] "Semiconductor industry association," *International Technology Roadmap for Semiconductors*, 2004.

[3] L.Scheffer, "Methodologies and tools for pipelined on-chip interconnect," in *Proc. of International Conference on Computer Design*, 2002, pp. 152–157.

[4] J. Cong, Y. Fan, X. Yang, and Z. Zhang, "Architecture and synthesis for multi-cycle communication," in *ISPD '03: Proceedings of the 2003 international symposium on Physical design*, 2003.

[5] J. M. Wang, P. Ghanta, and S. Vrudhula, "Stochastic analysis of interconnect performance in the presence of process variations," in *ICCAD*, 2004, p. 880.

[6] M. Webster, M. A. Tatang, and G. J. McRae, "Application of the probabilistic collocation method for an uncertainty analysis of a simple ocean modeltesting multivariate uniformity and its applications," *MIT Joint Program on the Science and Policy of Global Change*, 1996.

[7] D. C. Montgomery, *Design and Analysis of Experiments*. John Wiley and Sons, 1997.

[8] Cochran, G. William, and G. M.Cox, *Experimental Designs*, 2nd ed. John Wiley Sons, Inc., 1957.

[9] Y. S. Kumar, J. Li, C. Talarico, and J. Wang, "A probabilistic collocation method based statistical gate delay model considering process variations and multiple input switching." in *DATE*, 2005.

[10] L. Zhang, Y. Hu, and C. C. Chen, "Statistical timing analysis in sequential circuit for on-chip global interconnect pipelining," in *Proc. Design Automation Conf.*, 2004, pp. 904–907.

[11] "Berkeley ptm-interconnect." [Online]. Available: www-device.eecs.berkeley.edu/~ptm/interconnect.html