Yield Modeling and Analysis of Bundled Data and Ring-Oscillator Based Designs

Yang Zhang, Ji Li, Huimei Cheng, Haipeng Zha, Jeffrey Draper, and Peter A. Beerel

Computer Engineering Technical Report Number CENG-2018-01

Ming Hsieh Department of Electrical Engineering – Systems
University of Southern California
Los Angeles, California 90089-2562

April 2018

Yield Modeling and Analysis of Bundled Data and Ring-Oscillator Based Designs*

Yang Zhang, Ji Li, Huimei Cheng, Haipeng Zha, Jeffrey Draper, and Peter A. Beerel*
* University of Southern California - Los Angeles, United States
{zhan808, jli724, huimeich, hzha}@usc.edu, draper@isi.edu, pabeerel@usc.edu

***Abstract*—Both ring-oscillator based clocks and bundled-data designs mitigate the ill effects of process, voltage, and temperature (PVT) variations. They both rely on delay lines which, when made post-silicon tunable, offer the opportunity to add test margin into the design in which the delay line in shipped products is set slower than that which is successfully tested. By adopting the uniform and per-chip test margin methods to asynchronous designs, this paper mathematically analyzes the resulting yield and shipped product quality loss and compares them to traditional synchronous design, quantifying the potential benefits that arise from the correlation in delay among paths in the delay line and combinational logic.**

I. Introduction

PVT variations introduce statistical fluctuations in physical properties of the MOS devices which result in degrading the parametric yield and logic characteristics of the logic gates [1]. One effective approach to combat the PVT variations is using bundled data (BD) design [2], where the programmable delay line tracks the delay of the critical path [3], [4]. Although BD designs have been studied in several test schemes (see e.g., [5]), there is a serious lack of analysis and optimization of associated manufacturing test metrics for BD designs.

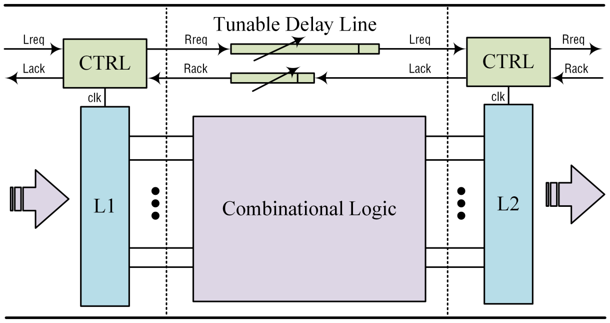

Fig. 1. Bundled data design with forward and backward delay lines

In Figure 1, two programmable delay lines are employed. One is placed on the forward latency path which accounts for the maximum delay of combinational logic to ensure the setup time constraint is met; the other is placed on the backward latency path and is used to control the non-overlap period of the local clocks, thereby mitigating hold violations.

In this paper, our focus is on delay faults [6], [7]. We expect the programmable delay lines to be analyzed during chip characterization, tested at a particular delay setting, and shipped at possibly a different, longer-delay setting. *Test margin* is the difference between the test frequency and the chip's shipped frequency, which is designed to mitigate 1) incomplete test coverage in which the critical path under test may be different from the actual critical path; and 2) the temperature and voltage during actual operation may be different from the ones under test.

As shown in Figure 2, there are four types of chip [8]:

- Good chips whose test paths pass test and chip performance meets the customer specification.
- Bad chips whose test paths fail to pass test and chip performance does not meet the customer specification.
- Yield loss chips whose test paths fail to pass the test but whose chip performance meets the customer specification.
- *Shipped product quality loss* $(SPQL)$ chips whose test paths pass the test but whose chip performance does not meet the customer specification.

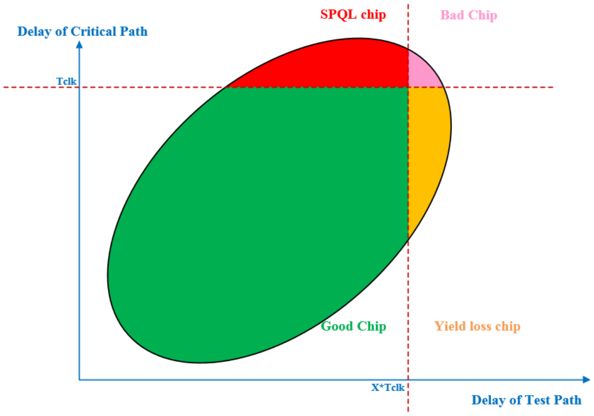

Fig. 2. Four types of chips (modified from [8])

Yield and SPQL can be calculated as

$$Yield = \frac{Good\ chips + SPQL\ chips}{All\ chips} \tag{1}$$

$$SPQL = \frac{SPQL\ chips}{SPQL\ chips + Good\ chips} \tag{2}$$

This paper analyzes and compares the yield and SPQL of asynchronous BD designs and ring oscillator (RO)-based designs to traditional synchronous designs. The mathematical

*This research has been supported in part by NSF Grant #1619415.
*Peter A. Beerel is also Chief Scientist at Reduced Energy Microsystems.

TABLE I
ANALYSIS DIMENSIONS

Test Margin	uniform	**per-chip**
Performance limit	average case	**worst case**
Time	right after fab	**years after fab**
Violations	setup	**hold**
variables	as measured	**with variations by 10%**

delay model in [9] is adopted and the yield advantage of the correlated designs is quantified, given the correlation coefficient between the combinational logic and delay line for test. Based on this model, we propose methods to determine the optimal setting needed to maximize yield while meeting a required SPQL. Monte Carlo simulation is run on a sample circuit to complement and support the mathematical model. It is observed that the BD/RO design has up to a 55% yield advantage over synchronous design given same test margin, and up to a 50% yield advantage given the same required SPQL. Speed binning the synchronous designs improves their yield but the asynchronous design can still shows significant advantages when the delays of the delay line and combinational logic are highly correlated.

Besides the analysis at the time of being shipped, an aging-aware Monte Carlo simulation flow is presented to accurately account for the Negative Biased Temperature Instability-induced timing difference in the analysis of the delays and correlation coefficients between T and L over the circuit's lifetime. Our analysis shows that the ratio of the critical path under test over delay line remains the same over the lifetime of the circuit. In other words, the correlation coefficient between them remains constant. This indicates that when a performance degradation over the lifetime of the circuit is allowed, there is no need of tuning delay lines to combat the aging effect for BD/RO designs. In contrast, synchronous circuits may require additional circuitry to track the performance degradation [10] in order to tune the clock and/or voltage to ensure aged circuits remain functional.

This paper extends initial work presented in [11]. For example, in addition to analyzing *uniform test margin* [11], this paper also considers *per-chip test margin*, meaning the optimal test margin may vary from chip to chip based on post-silicon per-chip measurements. Moreover, in addition to considering average-case performance constraints [11], this paper also considers worst-case performance requirements. A summary of the various dimensions we consider is illustrated in Table I. The first column was considered in [11] whereas this paper also considers the second column.

The rest of the paper is organized as follows. Section II conceptually defines and discusses the parameters for BD/RO designs that affect yield and SPQL. Section III introduces the basic mathematical model used for modeling correlations among the parameters defined in Section II. Next, Section IV analyzes yield given average performance and shipped product quality loss constraints and Section V focuses on where worst-case performance constraints are given. Section VI then describes our sample circuit, Monte Carlo simulation setup, and the correlations obtained. Section VII describes a method to merge analysis of aging into the Monte Carlo simulation. Next, Section VIII addresses how variations affect yield and graphically illustrates the analyses in Sections IV and V, quantifying the benefits of BD/RO designs for the specific correlations obtained in Section VI. Finally, Section IX discusses future work and concludes the paper.

II. KEY PARAMETERS

To mathematically analyze bundled data and ring-oscillator based designs, a model for the basic parameters is needed. All parameters are introduced conceptually in this section and a more formal model that captures their variation is described in the next section.

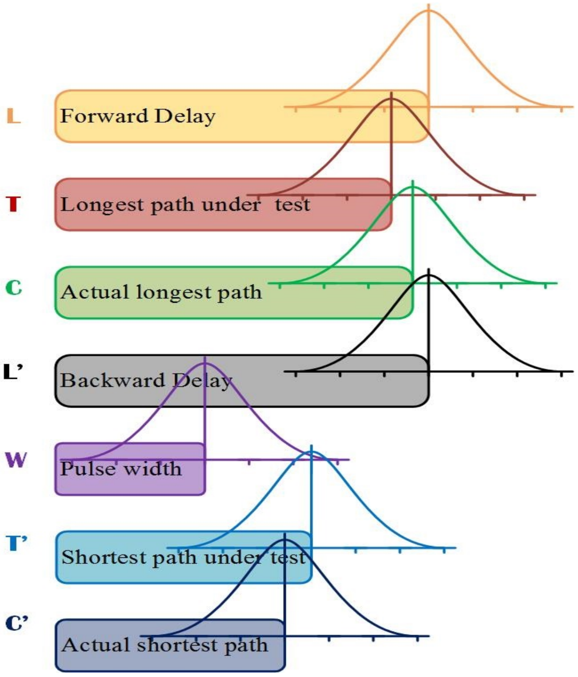

Fig. 3. Normally distributed Test Parameters

A. Critical paths

The critical paths represent longest paths during setup time analysis. The path with the maximum delay in the circuit, known as the *actual longest path*, determines the clock period for synchronous circuits and the minimum forward delay line length for BD circuits.

During test, selected paths on chip are tested to achieve a balance between fault coverage and test time. The longest path under test (T) is the slowest logic path exercised among all test vectors. In some cases, the actual longest path (C) is triggered by the applied test vectors. In other cases, however, due to increasing test data volume, process variations, and test times [12], the actual longest path may not be exercised by the applied test vectors. In such cases, C differs from T. The forward delay line L should be sufficiently long to ensure the shipped chip works with actual longest path C. In this paper, T, C and L are modeled using Gaussian distributions, as shown in Figure 3.

In contrast to setup time analysis, hold time analysis considers shortest paths as critical. During test, the shortest path under test (T') is the fastest logic path exercised among all test vectors. The actual shortest path (C') may not be triggered by the applied test vectors. The backward delay line L' should be tuned to ensure the actual shortest path C' is longer than the hold time requirement. T', C and L' are modeled using Gaussian distributions as well.

After test, the longest test path of a passing chip must, by definition, be shorter than the delay line or clock period. Similarly the shortest test path of it is supposed to be long enough to satisfy the hold time constraint. In contrast, the actual longest path and shortest path have a small chance of violating setup or hold time constraint. The chance that setup or hold time of a passing chip is not met, is also known as $SPQL$.

B. Ratio of the delay line for test to delay line (X and X')

During test of a BD or RO circuit, the forward delay line is tuned to have a smaller delay that is used for shipped chips. This introduces a test delay ratio as defined below:

$$X = \frac{Delay\ Line\ for\ Test}{Actual\ Delay\ Line} \tag{3}$$

Ideally X, the ratio of delay during test to shipped delay, is constant. However, because of process variation, X itself varies from chip to chip. The variance of X depends on the correlation coefficient between the delay line for test (XL) and the forward latency delay line (L). If the correlation coefficient equals 1, X is a constant, and thus has a variance of 0. If it is close to 0, X can be a variable with larger variance and thus the analysis based on a constant X can be incorrect. Fortunately, our experimental results show that, if the delay line is designed carefully, XL and L are indeed highly correlated.

$$L = XL + Test\ margin \tag{4}$$

The difference between the actual delay line and delay line for test is the test margin for BD and RO designs, as shown in Equation 4.

However, when we analyze the hold time test delay ratio, the delay line during test is longer than the delay line on working mode. We use X' instead to represent this ratio, that is larger than 1 naturally.

$$L' + Test\ margin = X'L' \tag{5}$$

Notice that only BD design has the ability of tuning hold time delay line and use X'.

C. Yield

To compare the yields of SYNC, BD, and RO designs, the SYNC design is assumed to have a nominal clock period of T_{clk} and the nominal test clock period is XT_{clk} . However, we also model performance binning of the synchronous designs which allows chips to be sold at different target frequencies to increase yield [13]. In particular, speed binning enables us to ship synchronous chips with a frequency range $T_{clk}(1-\beta)$ to $T_{clk}(1+\beta)$, where $T_{clk}(1+\beta)$ is the slowest shipped clock period. The yield of the SYNC design is thus

$$Yield_{SYNC} \ = \ P(T + s < XT_{clk}(1+\beta), T' > h). \tag{6}$$

where s and h represents setup and hold time of SYNC design.

Similarly, BD/RO designs are assumed to have a delay line delay of L and L' where the nominal delay during test is XL and $X'L'$, where $X < 1$ and $X' > 1$. However, the definition of the yield of a BD/RO design depends on the system requirements.

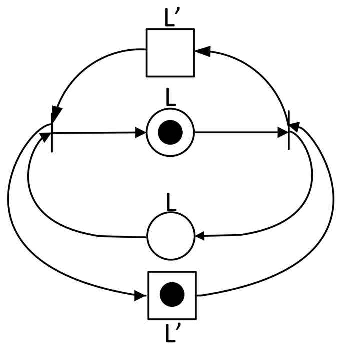

Fig. 4. Full-buffer channel net model of the performance

In this paper, the performance of BD designs is modeled using the Full Buffer Channel Net (FBCN) model [4] of a typical master-slave latch bundled-data configuration [14], illustrated in Figure 4. In this marked graph model, the forward latency represents the datapath delay from the master to slave latches as well as the datapath delay from the slave to master latches. It is captured by the delay line in the forward path L and labelled on the round places in the marked graph. The backward latency is the delay determined by the handshaking overhead in BD designs and is not present in RO designs. It is captured by the delay line in the backward path L' and labelled on the square places in the marked graph. The performance of the circuit is determined by the longest cycle in this graph [4] and thus equals $max(2L, L + L', 2L')$.

In particular, if it is acceptable to ship chips whose performance varies with PVT variations but on average has the same delay as synchronous designs, then L and L' can be assumed to be normally distributed whose means equal $\frac{T_{clk}(1+\beta)}{/}2$ and the BD/RO yield is the probability of having the longest path under test (T) smaller than the delay line for test (XL) and the shortest path under test (T') bigger than the overlapping period ($W - X'L'$), as shown in Figure 5.

$$Yield_{B-AVE} = P(T + s < XL, T' - h > W - X'L')^1 \tag{7}$$

This definition may be best suited for many-core or multi-chip designs for non real-time applications.

Note that, in this case, the larger the delay of the programmable delay line during test, the higher the chance that T

[1] It is assumed that time borrowing is not allowed

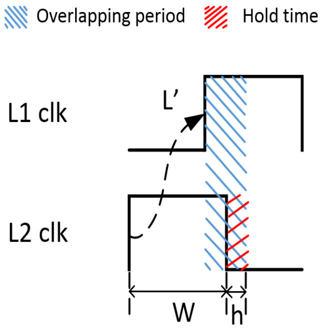

Fig. 5. Illustration of the hold time constraint

will be smaller than XL. In other words, the yield of a BD/RO design is a monotonically increasing function of X. Given a certain X, the yield is determined by the correlation between T and XL, $\rho_{T,XL}$. If $\rho_{T,XL}$ equals 1, the delay line (XL) tracks the critical path (T) for every chip and the chance of having a chip that does not pass the test is 0. If $\rho_{T,XL}$ equals 0, there is a good chance of having larger T and smaller XL, i.e., a test failure.

If, on the other hand, a worst-case performance constraint is also given, the yield of a BD/RO circuit can be expressed as

$$\begin{aligned} Yield_{B-WC} = P(&T + s < XL, T' - h > W - X'L', \\ &L + L' < T_{clk}(1+\beta), 2L < T_{clk}(1+\beta)). \quad (8) \end{aligned}$$

Note here we omit the constraint that $2L' < T_{clk}(1+\beta)$ because, in practice, the nominal delay of L' is much smaller than that of L and thus this constraint is typically redundant. To appreciate the difference between worst-case and average-case constraints, consider the case where β is set to zero. If the mean of L and L' are naively set to $T_{clk}/2$, as is optimal when considering average-case performance, the worst-case yield would be close to 50% and we would lose approximately half of the manufactured chips due to setup violations. Thus a more sophisticated approach to optimize L for this case is needed and our specific proposed approach is discussed in Section V.

Interestingly, the worst-case yield definition can be further classified into two sub-categories. $Yield_{B-WCP}$ is the yield considering performance violations caused only by process variations and $Yield_{B-WCPVT}$ considers performance violations also caused by (temporary) changes in operating voltage and temperature. In some applications, such as mobile and IoT, we may allow performance to change with changes in voltage and temperature and for such applications $Yield_{B-WCP}$ may be suitable. In other applications with strict real-time constraints, however, $Yield_{B-WCP}$ may be a better measure.

As discussed by Cortadella et al. [15], the delays of paths that are physically close to each other are highly correlated. Given that the delay line (T) and the combinational paths (XL) that it is supposed to match are often physically close, their delays are often highly correlated, i.e., $\rho_{T,XL}$ is close to 1. Consequently, as we will show below, given an average performance constraint, BD/RO designs have a higher $Yield_{B-AVE}$ than SYNC designs for the same test margin. More precisely, Cortadella et al. [15] suggest that the clock margin required need only be used to compensate the *local* process variation (i.e., mis-match) between the delay line under test and critical path in combinational logic. Similarly, we show that for BD/RO designs only local variations motivate a larger test margin and affect chip yield. Conversely, to achieve the same yield as SYNC design, we show that BD/RO designs can have a smaller test delay ratio X. On the other hand, because the delay line, which dictates the performance of BD/RO designs, is affected by voltage and temperature similarly to that of synchronous combinational logic, we show the yield advantage of BD/RO designs disappears when strict worst-case performance constraints are given.

D. Shipped product quality loss

Shipped product quality loss ($SPQL$) determined the quality of shipped chips. Thus manufacture generally puts a limit on it, in order to achieve an acceptable failure rate of shipped products.

The $SPQL$ of SYNC design is defined as

$$\begin{aligned} &SPQL_{SYNC} = \\ &\quad P\big((C + s > Tclk(1+\beta) \ or \ C' > h\ ) \mid pass\ test\big) \quad (9) \end{aligned}$$

where the condition pass test is as used in Equation 6.

Similarly, we define the $SPQL$ of BD/RO design as

$$\begin{aligned} &SPQL_{B-AVE} = \\ &\quad P\big((C + s > L \ or \ C' - h < W - L'\ ) \mid pass\ test\big) \quad (10) \end{aligned}$$

where the condition BD/RO passes the test is the same as used in Equation 7. Finally, to define $SPQL_{B-WC}$, we simply apply the stricter performance constraint for passing the test, as expressed in Equation 10.

E. Aging effects

Aging effects lead to the increase of delays from their values when shipped, resulting in a gradual performance degradation. Aging does not change the definition of yield, SPQL etc., but does change the distribution of the yield-determining parameters, including T, C and L.

Our simulations show that aging affects T, C, and L similarly. Thus, for BD/RO designs in applications that allows chips to slow down as they age, we can determine yield using Monte Carlo simulation results that do not include aging, as addressed in Section VIII-C. For BD/RO designs in applications that require a chip to meet a fixed performance constraint throughout its lifetime, we need to apply variations after aging on transistor width, length and threshold voltage, run Monte Carlo Simulation, and use the resulting aged distributions. For SYNC designs, the clock or power supply must be conservatively set based on the aged distribution or adjusted as the chip ages using both distributions.

III. Mathematical model

A canonical delay model [9] for gate delays, slacks, and slews can be expressed as

$$a_0 + \sum_{i=1}^{n} a_i \Delta Y_i + a_{n+1} R_a \tag{11}$$

where a_0 represents the mean value μ, ΔY_i models global process variations, and ΔR_a models other variations. ΔY_i and ΔR_a are assumed to be zero-mean, unit-variance Gaussians. Coefficients a_1 to a_{n+1} are sensitivities to the corresponding variations. The critical path under test (T), the actual critical path (C), and the delay of the delay line (L) are modeled using form 11 with different parameters. We assume to some degree that T, C, and L are correlated. We thus introduce correlation coefficients, $\rho_{T,C}$, $\rho_{T,L}$, and $\rho_{C,L}$ to quantify their correlations.

$$\rho_{T,C} = \frac{cov[T,C]}{\sigma_T \sigma_C} \tag{12}$$

where

$$cov[T,C] = \sum_{i=1}^{n} a_{T,i} a_{C,i} \tag{13}$$

$a_{T,i}$ and $a_{C,i}$ are sensitivities to globally correlated variations of distributions T and C respectively. Additionally, $\rho_{T,L}$ and $\rho_{C,L}$ are similarly computed.

Form 11 is a linear combination of Gaussian distributions. Its variance σ^2 is

$$\sum_{i=1}^{n} a_i^2 + a_{n+1}^2 \tag{14}$$

The probability density function of a Gaussian distribution can be expressed as

$$f_X(x) = \frac{1}{\sqrt{2\sigma_x^2 \pi}} e^{-\frac{(x-\mu_x)^2}{2\sigma_x^2}} \tag{15}$$

where μ_x and σ_x denote mean and standard deviation of distribution respectively. The joint probability density function of k-variate Gaussian distribution can be expressed as

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{\mathbf{1}}{\sqrt{\mathbf{2\pi^k|\Sigma|}}} \mathbf{exp}(-\frac{\mathbf{1}}{\mathbf{2}}(\mathbf{x} - \mu_{\mathbf{X}})^{\mathbf{T}} \mathbf{\Sigma^{-1}} (\mathbf{x} - \mu_{\mathbf{X}})) \tag{16}$$

where

$$\mathbf{X} = (\begin{smallmatrix} X_1 \\ \vdots \\ X_k \end{smallmatrix}),\ \mu_{\mathbf{X}} = (\begin{smallmatrix} \mu_{X_1} \\ \vdots \\ \mu_{X_k} \end{smallmatrix}) \tag{17}$$

and

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{X_1}^2 & . & \rho_{X_1,X_k}\sigma_{X_1}\sigma_{X_k} \\ . & . & . \\ \rho_{X_1,X_k}\sigma_{X_1}\sigma_{X_k} & . & \sigma_{X_k}^2 \end{pmatrix}$$

Thus, the joint probability density distribution of T and C is represented as $f_{T,C}(t,c)$. We define $f_{T,L}(t,l)$ and $f_{T,C,L}(t,c,l)$ in a similar fashion.

IV. Optimal test margin given an average performance constraint

Yield and its relation to test margin has been conceptually introduced in Section II for both SYNC and BD/RO designs. In this section, we analyze the optimal test margin that maximizes yield subject to a given average performance constraint and SPQL. It analyzes both uniform and per-chip test margins.

A. SYNC design with speed binning

The optimal uniform test margin X for SYNC design can be obtained by setting the SPQL to its maximum and solving Equation 9 for X [11]. To implement per-chip test margins, we propose to measure performance-sensitive ring oscillators during wafer test. This information can help improve yield by enabling the application of a per-chip test margin computed for each individual chip. For the purposes of this paper, we assume the chip performance is estimated with delay of a ring oscillator and denoted as L, the same notation we use for the forward delay line in BD designs. Test margin on the forward latency path only affects the setup time related metrics, thus the problem of finding the optimal test margin given a required SPQL can be re-expressed as follows.

$$\max_{X(L)} P(T + s < Tclk(1+\beta)) \tag{18}$$

subject to

$$P(C + s > Tclk(1+\beta) | T + s < Tclk(1+\beta)) \leq q \tag{19}$$

where q is an upper limit on SPQL given by the user and determines the quality of shipped products. It is proved in [8] that the optimal yield for SYNC is achieved when the SPQL reaches its maximum. Thus, the inequality in Constraint 19 can be replaced with an equality without loosing optimality. By analytically solving this optimization problem, we can write X in the form of

$$X = \gamma L + \eta \tag{20}$$

where

$$\gamma = \frac{\sigma_{CT}\sigma_{TL} - \sigma_{CL}\sigma_T^2}{(1+\beta)T_{clk}(\sigma_{CT}\sigma_L^2 - \sigma_{CL}\sigma_{TL})} \tag{21}$$

and η can be obtained by substituting X in Inequality 19 as $\gamma L + \eta$ and use its equation form. The detailed proof of this derivation is given in Appendix A.

B. BD/RO design

As suggested by the FBCN model of BD/RO designs in Section II-C, their master-slave latch-based nature implies that both the forward and backward delay lines can be configured to have their mean delay equal half of the synchronous clock period to meet the average performance constraint. Similar to the SYNC design, the test ratio X of a BD/RO design can be configured to optimize yield. In contrast, however, BD/RO designs also can configure the hold time test delay ratio X' which makes the analysis more complicated.

1) Monotonicity of SPQL: The problem of optimizing the yield of BD/RO designs subject to an SPQL requirement is somewhat simplified by the monotonicity of SPQL versus X and X', as first shown in [11] and formalized as follows:

Theorem I: The $SPQL_{B-AVE}$ of a BD/RO design is a monotonically increasing function of X if the correlation coefficient between T and C satisfies $\rho_{T,C} > 0$, and a monotonically decreasing function of X' if the correlation coefficient between T' and C' satisfies $\rho_{T',C'} > 0$.

As described in [11], the correlation constraint is easily satisfied. Consequently, decreasing X and increasing X' causes an increase BD/RO yield. It also causes an increase in SPQL. Thus, similar to SYNC design, we can achieve the maximum yield subject to an SPQL constraint when the SPQL hits its maximum limit. This result can guide designers and CAD tools. It also helps us find a unique analytical solution when X or X' is the only unknown variable in the equation for SPQL.

2) Uniform X: Due to the monotonicity of SPQL, the optimal test margin X for BD/RO designs is obtained when its SPQL is set to its maximum limit [11]. In particular, by setting Equation 10 to q, we are able to obtain a unique optimal value for X.

3) Uniform X and X': To achieve the optimal joint values of X and X', we sweep them and identify pairs whose SPQL equals its limit q. By plugging the satisfying pairs of X and X' into Equation 7, we are able to obtain a set of yields, and record the pair that leads to the maximum yield.

4) Per-chip X: As with SYNC design, optimizing the BD/RO test margin X on a per-chip basis requires an easily obtainable measure of chip performance. Fortunately, the delay of the delay line is a naturally good candidate to estimate chip performance. In this subsection, we assume the delay line can be configured into a ring oscillator during test and tune the test margin parameter X based on the measured ring oscillator delay (L).

The problem of finding the optimal test margin given a required $SPQL$ can be expressed as follows.

$$\max_{X(L)} P(T + s < XL) \tag{22}$$

subject to

$$P(C + s > L | T + s < XL) \leq q \tag{23}$$

Due to the monotonicity proof in Section IV-B1, the yield of BD/RO is optimal when SPQL of BD/RO reaches its maximum value. Thus, similar to the above analysis, the less-than-or-equal-to sign in Inequality 23 can be safely replaced by equality. By solving the optimization problem, we can determine the optimal setting of X as a function of the delay line L as follows:

$$X = \frac{\gamma}{L} + \eta \tag{24}$$

where

$$\eta = \frac{\sigma_T^2 \sigma_L^2 - \sigma_{TL}^2 + \sigma_{CT}\sigma_{TL} - \sigma_{CL}\sigma_T^2}{\sigma_{CT}\sigma_L^2 - \sigma_{CL}\sigma_{TL}}. \tag{25}$$

We can then express γ as a function of q by substituting the expression for X in terms of γ into Inequality 23. The derivation details are given in Appendix B.

It is interesting to note that the per-chip X Equations 20 and 24 for SYNC and BD/RO have opposite dependencies on L. In particular, because the SYNC clock period is fixed at $Tclk$, a SYNC chip has a lower chance of passing its test as L increases. In contrast, with a larger L, the BD/RO constraints on T and C are relaxed, making its test somewhat easier to pass.

5) Per-chip X and X': Simultaneously finding the optimal per-chip configuration of X and X' is more complex because the both X and X' are modeled as functions with η and γ parameters. Defining a finite grid-search over this space is difficult because there are no clear bounds on the parameters η and γ. One alternative heuristic is to sweep X' over a predefined range and for each point obtain the set of equations for per-chip X and apply the analysis above in Section IV-B4. From the results, we can find the optimal combination of uniform X' and per-chip X, as a function of L. Then, we can replace the uniform X' by a per-chip X' using a similar process. Because this two-step optimization procedure does not explore the entire design space, the result may not be the optimal per-chip solution. However, the result is guaranteed to be better than the yield using uniform test margins.

V. Optimal test margin given worst-case performance constraints

Performance constraints vary from application to application. Ensuring an average performance constraint may be acceptable in cases in multi-core systems in which individual cores can have varying performance or where voltage scaling can compensate for varying performance. However, in other applications, a manufacturer may be required to meet certain worst-case performance constraints. With this motivation, this section focuses on the following problem: given a required SPQL and worst-case performance constraint, configure the setup and hold delay lines as well as their uniform/per-chip test margins to maximize yield. Due to the fact that average and worst-case performance for SYNC designs are the same, this section focuses on BD/RO design. Unfortunately, optimizing the BD/RO delay lines for the worst-case performance constraint is more complicated than for the average-case performance constraint because the yield may no longer be a monotonic function of the delay lines L and/or L'. Instead, we need to consider variations and carefully balance setup and hold time violations with the worst-case performance constraint to find the optimal setting of L and L' and their associated test delay ratios X and X'.

A. Uniform X

The optimal yield given both SPQL and worst-case performance constraints depends on L, L', X and X'. If the hold time requirement is easily met, e.g. the shortest paths are sufficiently long to satisfy the hold constraint, however, no hold time test margin is needed. As a first step, this subsection makes this assumption and therefore focuses on setting X and L.

Given this assumption, we set SPQL to its maximum limit q and optimize test delay ratio X and L. Based on $SPQL_{B-WC}$, by assuming $X' = 1$ and $\mu_{L'}$ is constant, we can obtain the optimal test delay ratio X as a function of L and q. By sweeping L, we obtain its corresponding test delay ratio X. More specifically, all combinations of test delay ratio X and L are plugged into Equation 8 to achieve multiple possible yields given a certain q and the X, L pair that leads to the maximum yield is recorded. Note also that we can also run this procedure multiple times do determine how the optimal yield varies as a function of q.

B. Uniform X and X'

In Section V-A, we assumed $X' = 1$ and kept $\mu_{L'}$ constant, sweeping L to achieve the optimal X and yield. In this subsection, we wish to optimally set X' and L' as well as L and X. To do this, we propose to simultaneously sweep all but one of X, X', L and L'. For example, for each sample point of X, L and L', X' can be calculated from $SPQL_{B-WC}$. Each four tuple can then be plugged into Equation 8 to obtain multiple possible yields given a specified q. The maximal yield can be picked from these results, concluding the optimization procedure.

C. Per-chip X

In Section IV-B4, we found that given an average-case performance constraint, we could express the optimal per-chip X as a function of two parameters γ and η and L by manually solving the optimization problem expressed in Equation 22. An important observation is η, expressed in Equation 24, is independent of the lower and upper limits on L. Thus, the worst-case performance limit on the forward delay line, which bounds the upper limit on L, does not effect the value of η. Consequently, the optimal γ can be derived by substituting X in $SPQL_{B-WC}$ by $\frac{\gamma}{L} + \eta$, where X' is assumed to be 1.

D. Per-chip X and X'

Similar to the average-case situation described in Section IV-B5, the per-chip optimization problem is complicated because defining a finite grid search over all possible models of X and X' is difficult to construct. To simplify the optimization, we first run the analysis described in Section V-B by assuming X and X' are uniformly set. We then assume X' is set to the optimal uniform value and obtain the optimal per-chip X as a function of L using the analysis in Section V-C. Lastly, we can find the optimal per-chip X' by fixing X to this optimal value. This heuristic approach does not guarantee an optimal solution, but does lead to a better yield compared to using uniform test margins.

VI. Monte Carlo simulation and measuring correlations

The yield of both BD/RO and SYNC circuits depend on the correlation coefficients between the test parameters T, C, XL, and L. This section discusses how we use Monte Carlo simulations on an example combinational circuit and programmable delay line to estimate these values for a particular process. In particular, all circuits were designed in the IBM 65 nm CMOS technology and were sized to achieve equal rising and falling propagation delays.

Our example combinational circuit, illustrated in Figure 1, is a 16-bit carry select adder (CSA). CSAs are a simple circuit that have multiple potentially critical paths and thus represents the case in which it may not be practical to test all possible paths. In particular, the structure of the carry select adder, shown in Figure 6, has 17 inputs and 17 outputs. By assuming that the delay of a MUX is comparable to the delay of a 1-bit full adder, the critical path is from the lowest significant bit of one of the grouped ripple carry adders (RCAs) to the most significant bit of the primary outputs.

Our example programmable delay line is the MUX-based delay line shown on Figure 7. It is analyzed to quantify the correlation between XL and L. We assume the I1 is selected as the valid input of the MUX during test and I2 for shipped chips. Thus the delay line for test (XL) uses 38 inverters and the delay line (L) uses 40 inverters. 40 is picked to obtain a slightly longer delay line than the critical path of the CSA. Based on different requirements of the $SPQL$, we can pick any even number smaller than 40. And 38 is one of possible value which results in a reasonable yield and $SPQL$.

The different types of performance constraints discussed in Section II warrant different setups to the Monte Carlo simulation process. The first MC variation setup is where we randomly vary process, voltage, and temperature. Voltage is varied between 0.9V to 1.1V and the temperature from -55°C to 120°C. For each MC run, the delays of all potentially critical paths are recorded. The largest delay among these paths is the actual critical path delay, one sample point of C. The maximum of path delay from Cin to $Cout$ and from $A[2]$ to $Cout$, a subset of all potentially critical paths, is one sample point of T. We simulated 9,000 sample points with randomly set PVT variations. The i^{th} sample point provides $a_{T,i}$ and $a_{C,i}$ in Equation 13. By plugging these sample values into the equation, $cov[T, C]$ is estimated. Then we use Equation 12 to calculate $\rho_{T,C}$, where σ_T and σ_C can be estimated from all sample points. With all above parameters, the joint distribution of T and C is obtained using Equation 16. A similar procedure is used to obtain the joint distributions of T and C with L.

The resulting correlation coefficients and joint distributions are used to compute the yield of SYNC design $Yield_{SYNC}$ and the average and strict worst-case yields for ASYNC designs, $Yield_{B-AVE}$ and $Yield_{B-WCPVT}$. However, when computing the yield $Yield_{B-WCP}$ we must alter the MC setup to not include variations in temperature and voltage, fixing them to their nominal values. This is because 1) we allow fluctuations in performance caused by changes in voltage and temperature and 2) changes in voltage and temperature do not cause the BD/RO circuit to malfunction. The latter fact is because the correlation coefficients of all variables under variations in voltage and temperature are 1, meaning their

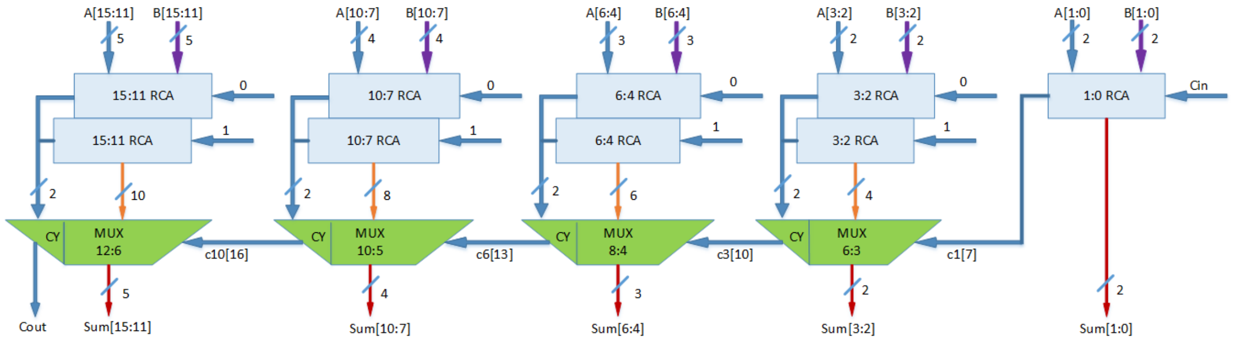

Fig. 6. 16-bit carry select adder

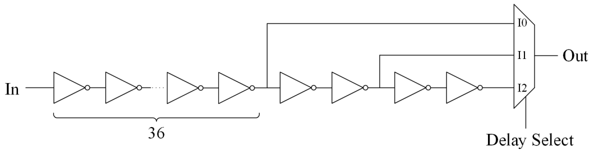

Fig. 7. Programmable delay line designed using a MUX

TABLE II
Analysis of the Correlation Matrix due to PVT Variations

	PVT variation			1.0V, 27°C varying P		
	T	C	L	T	C	L
T	1	0.99	0.98	1	0.97	0.86
C	0.99	1	0.98	0.97	1	0.87
L	0.98	0.98	1	0.86	0.87	1
	T'	C'	L'	T'	C'	L'
T'	1	0.99	0.83	1	0.98	0.68
C'	0.99	1	0.84	0.98	1	0.69
L'	0.83	0.84	1	0.68	0.69	1

variations affect the delay of the delay line and combinational logic equally.

Table II shows the final correlation matrix of T, C and L, as well as T', C' and L'. Compared to only considering process variation, PVT variation leads to higher correlation coefficients. This is because the impact of local mismatch is reduced when global systematic variations are introduced. Depending on the actual PVT variation in real circuits, the correlation coefficients may change. However, the remainder of this paper shows results based on these obtained parameters.

Finally, it is important to recall that the mathematical model presented in Section III assumes that the test delay ratio X is constant in BD/RO designs. Fortunately, our MC simulations justify this assumption. The Monte Carlo simulation shows that XL and L are highly correlated with $\rho_{XL,L} = 0.999$. This is largely because in our example delay line the tested delay line XL is actually part of the shipped delay line L. The high correlation between XL and L is illustrated Figure 8 which shows the linear nature of the ratio of the delay XL over L. The slope $X = \frac{XL}{L}$ has a mean of 0.97 and variance is 2.4×10^{-5}, suggesting that X is close to a constant.

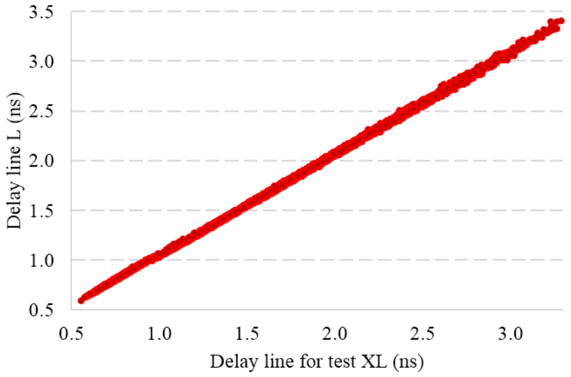

Fig. 8. Delay of the delay line L vs. the delay of the delay line for test XL

VII. Aging Analysis

With the aggressive downscaling of CMOS technology, Negative Biased Temperature Instability (NBTI) becomes one of the most critical aging effects threatening the reliability of nanoscale CMOS circuits [16]–[19]. NBTI is caused by the stress on PMOS transistors ($V_{gs} = V_{dd}$) and leads to an increase in both the threshold voltage (V_{th}) of the PMOS transistor and the delay of the associated gate. Due to the NBTI effect, many circuit paths that are not critical in the design stage may turn critical over time, causing timing violations during the operation [17]. The NBTI-induced timing difference will significantly affect the accuracy of the proposed yield and shipped product quality loss analysis, and therefore, it is imperative to consider the NBTI effect in the proposed evaluation framework.

In this paper, we use an NBTI aging model for a 65nm process of a commercial foundry, where NBTI is identified as the most critical aging effect for this process. In this model, the NBTI-induced threshold voltage shift ΔV_{th} of a PMOS transistor is calculated as

$$\Delta V_{th} = f_{NBTI}(V_{dd}, t_{on}, D_{load}, Lg, T) \qquad (26)$$

where V_{dd}, t_{on}, L_g, D_{load}, and T represent the supply voltage, total “on” state time, gate length, load, and temperature, respectively. The aging model is similar to other accessible NBTI aging models in the literature [16], [18], [19].

Next, we propose an aging-ware Monte Carlo simulation flow with the NBTI model. We assume the circuit operates under a constant supply voltage V_{dd} throughout its lifetime. For each PMOS transistor, the load D_{load} and gate length L_g, which are determined in the design stage, are extracted from the netlist. The “on” state time t_{on} is calculated by multiplying the total circuit operation time t_{op} by the probability of “on” state p_{on} (i.e., $V_{gs} = 0$) of the PMOS, i.e., $t_{on} = t_{op} \cdot p_{on}$. According to [20], the probability of logic “on” state can be calculated using two approaches: (i) the correlation coefficient method (CCM) approach proposed in [21], or (ii) simulations over a large set of typical vectors (possibly obtained by running a set of benchmark programs). In this paper, the first approach is adopted.

One important observation is that the temperature parameter T appears in both Equation 26 and the PVT variation analysis. In the proposed aging-aware analysis, for each PVT corner, the NBTI-induced ΔV_{th} of all the PMOS transistors in the circuit of interest is re-calculated based on Equation 26 with the T in that corner. Furthermore, for each user-specified circuit operation time, the Monte Carlo simulation (mentioned in Section VI) is executed once with the updated ΔV_{th} drift applied to each PMOS transistor. Algorithm 1 provides the pseudo code of the flow.

Algorithm 1: Pseudo code for the aging-aware Monte Carlo simulation flow

```
Load netlist of interest and technology library;
foreach PMOS in the netlist do
    Extract $D_{load}$, $L_g$;
    Calculate $p_{on}$ using the correlation coefficient method [21];
end
foreach corner of the technology do
    foreach user-specified circuit operation time $t_{op}$ do
        foreach PMOS do
            Calculate $t_{on} = t_{op} \times p_{on}$;
            Update $\Delta V_{th}$ via Equation 26;
            Update width and length based on process
              variation and Equation 26;
        end
    end
    Run simulations according to Section VI;
end
```

VIII. Results and discussion

This section first presents the yield analysis of SYNC and BD/RO design given an SPQL and performance constraints and then explores the impact of aging.

A. Test delay ratio and yield analysis given an average performance constraint

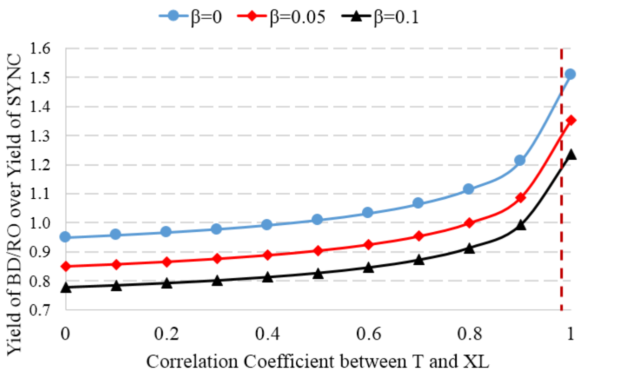

Fig. 9. Ratio of BD/RO yield to SYNC yield vs. $\rho_{T,XL}$ given an average-case performance constraint

Figure 9 plots the ratio of optimal yields of BD/RO (Equation 7) over SYNC (6) designs as a function of the correlation coefficient between T and XL with no SPQL constraint. Notice that as the correlation becomes closer to perfect, the yield advantage of BD/RO designs increases. For example, the curve labeled as $\beta = 0$ shows that the ratio is larger than 1 when the correlation coefficient is larger than 0.51. The red dashed line towards the right side of the plot indicates the actual $\rho_{T,XL}$ measured from our sample circuit under the measured PVT variations summarized in Table II. With a 5% slowest speed bin, i.e., $\beta = 0.05$, the yield of SYNC after binning increases, but the yield of BD/RO is still larger than SYNC if the correlation coefficient is larger than 0.8. As we increase β, the threshold correlation coefficient for which point the yields are equal increases. For example, $\beta = 0.1$ leads to a larger threshold value of 0.89. The result shows the importance of high correlation coefficient between combinational logic and delay line, which leads to the yield advantage of BD/RO over SYNC.

To appreciate the impact of the SPQL constraint on yield, Figure 10 plots the $SPQL$ vs. uniform test delay ratio X graphically using our measured statistical results. In particular, the mean, covariance, and correlation matrix of T, C and L is computed from our Monte Carlo simulation data and the joint distribution of T, C and L is mathematically derived. By integrating the joint distribution, we can plot the SPQL of BD/RO versus a uniform test delay ratio X. Notice that, as predicted by Theorem I, it shows that $SPQL$ is a monotonically increasing function of X.

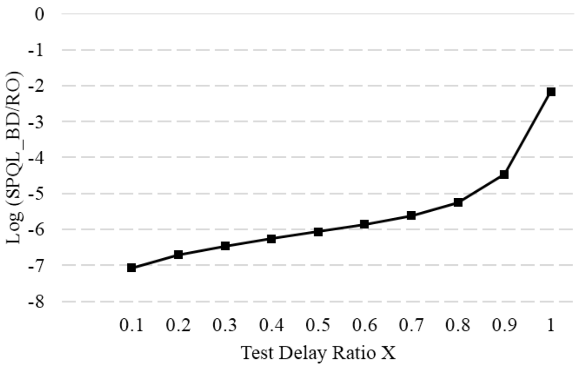

Fig. 10. log(SPQL) vs. X

Also we know that the optimal yield for BD/RO and SYNC designs is achieved when the $SPQL$ is set to its maximum value which thereby determines X. Thus, we can now compare yield at different values of SPQL. In particular to obtain the yield vs. $SPQL$ curve for BD/RO designs, For each desired $SPQL$, we first determine the corresponding test margin X from Figure 10. Based on this X, the BD/RO yield, $P(T < XL)$, is calculated using the joint distribution of the critical path under test and the delay line delay. The yield vs. $SPQL$ curve for SYNC is obtained similarly. Figure 11 plots the resulting yield versus required SPQL for both SYNC with binning and BD/RO designs. When the required SPQL is larger than 0.001, the yield of a BD/RO design is 50% higher than the comparable SYNC design without binning. A larger β allows more slow chips to pass the test. As shown in Figure 11, $\beta = 1$ boosts the SYNC yield higher, but it is still not as good as the equivalent BD/RO circuit.

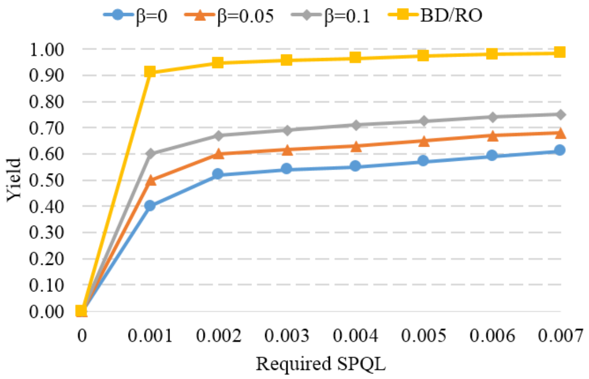

Fig. 11. Yield vs. required SPQL given an average-case performance constraint

TABLE III
γ AND η VERSUS REQUIRED SPQL

q	0.001	0.002	0.003	0.004	0.005	0.006	0.007
γ	-41.8	-28.3	-18.7	-10.6	-3.25	3.83	10.9
η	0.992	0.992	0.992	0.992	0.992	0.992	0.992

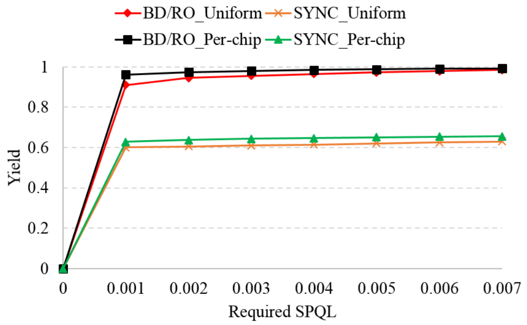

Fig. 12. Yield vs. SPQL for both per-chip and uniform test margins given an average performance constraint

To extend this analysis to per-chip test margins, Table III shows the optimal values of γ and η for the optimal per-chip test margin X based on the analysis in Section IV-B. Using these results, Figure 12 plots both the per-chip and uniform test margins for SYNC and BD/RO circuits and illustrates the increase in yield that per-chip test margins provides. In particular, for SYNC designs, as described in [8], per-chip test margins increases yield by about 10%. For BD/RO designs per-chip test margins increases yield by about 5%.

More generally, BD/RO yields with average performance constraints are significantly larger than their SYNC counterparts. They are approximately 40% larger when using uniform test margins and 37% larger when using per-chip margins. This yield advantage stems from two factors. First the combinational logic and delay line in BD/RO designs are highly correlated. The higher correlation in BD/RO designs leads to smaller X for the same desired yield. Second, the smaller X indicates smaller SPQL, as initially described in [11]. Conversely, given the same SPQL, BD/RO designs have larger X and thus increased yield.

In addition to yield comparison, it is also useful to study how variations, in particular process mismatch, affects the yield given average-case constraint. In Section VI, we explained that we used Monte Carlo simulations varying process, voltage, and temperature to compute $Yield_{B-AVE}$. However, it is important to note that the global-variation-induced delay changes on the delay line and critical path are identical. In particular, additional Monte Carlo simulations showed that the pair-wise correlations between T, C, and L under global variations are all exactly one. Thus, (global) voltage and temperature variations have no effect on $Yield_{B-AVE}$. Similarly, global process variation does not affect $Yield_{B-AVE}$. The only variation that affects correlation coefficients and thus yield is (local) process mismatch, where severe mismatch leads to a low $Yield_{B-AVE}$.

The intuition behind this result is discussed in [15] in the context of margins for a ring-oscillator-based clock. Because global variation changes the delay of the ring oscillator/delay line and combinational logic in the same manner, it does not warrant increasing the clock margin. We show that for the same reason, global variations do not adversely affect the yield of BD/RO designs. Thus, as long as the mean of the delay line under test is longer than the critical path of the combinational logic, the resulting yield is close to 1.

In contrast, for SYNC designs the yield under global variations behaves similarly to under PVT variations and is significantly less than 1 when the test margin is not sufficiently large. This is because the period of the global clock is fixed and thus does not track the delay of the combinational logic. Binning the synchronous circuit reduces the impact (see e.g., [13]), but the fundamental differences remain.

To further explore how the correlation coefficients change yields, Figures 13 and 14 show how different correlation-related factors affect the yield. The factors we studied are classified into two categories: 1) the mean and standard deviation of the underlying delays, and 2) the correlation coefficients between delays. To simplify the analysis, we either change the mean and standard deviation, or correlation coefficients by multiplying them by a scaling factor. Figure 13 shows the $Yield_{SYNC}$ is mainly affected by mean and variance. Correla-

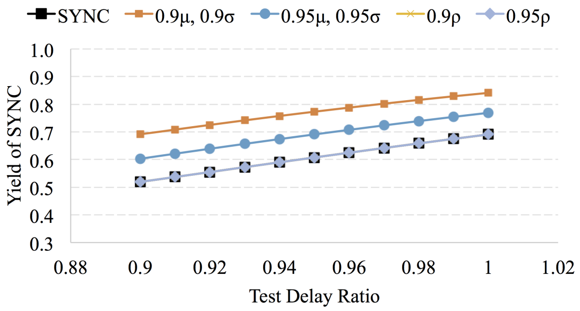

Fig. 13. Yield of SYNC versus Test Delay Ratio

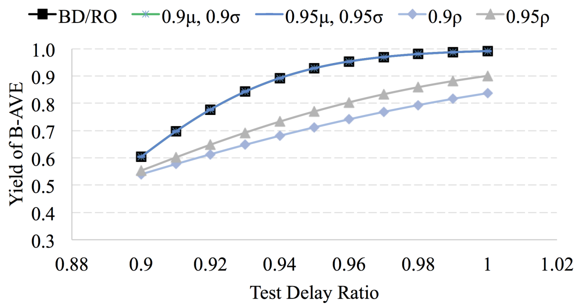

Fig. 14. Yield of B-AVE versus Test Delay Ratio

tion coefficients have no influence on it. In contrast, Figure 14 show that the correlation between delays and not their mean and variance affects the $Yield_{B-AVE}$. In particular, note that the two curves derived from the change in mean and standard deviation fall directly on top of the original BD/RO curve with no change of parameters. This result illustrates the fact that $Yield_{SYNC}$ is the probability that the critical path is shorter than a fixed number, where as $Yield_{B-AVE}$ compares the critical path to a delay line, whose delay will track that of the critical path under global variations.

We also explored the cases where the hold time constraint is as large as 10% or 20% of $Tclk$. In these scenarios, we need to either set a minimum constraint on shortest path or tune L' to resolve the hold time issue. To simplify the analysis, we assume that minimum constraint can improve the mean of T' by at most $\sigma_{T'}$.

TABLE IV
YIELD COMPARISON UNDER LARGE HOLD TIME AND AVERAGE-CASE PERFORMANCE CONSTRAINTS

	Tune L’			Add min delay constraint	
hold time (h)	0	$10\%Tclk$	$20\%Tclk$	$10\%Tclk$	$20\%Tclk$
$Yield_{BD-AVE}$	0.99	0.99	0.96	0.99	0.99
$Yield_{SYNC}$	0.61	0.42	0.01	0.61	0.05

Table IV shows that yield of bundled-data design is still close to 1 when hold time is 20% of $Tclk$ whereas the yield of comparable SYNC designs is more challenged.

B. Test delay ratio and yield analysis given worst-case performance constraints

Optimizing yield under worst-case performance constraints is more complex and, as described in Section V, generally requires a brute-force search through a subset of parameters.

As an example of an intermediate result of such a search, Figure 15 shows the normalized mean of the delay line versus test delay ratio, given two different SPQL requirements. They arise from the search sweep step, where X' is 1 and $\mu_{L'}$ is $Tclk(1+\beta)/2$. Notice that the mean of the delay line increases as X increases. The values above the plot are the corresponding yield of a BD/RO circuit under worst-case performance requirement set to $T_{clk} * 1.05$ with temporary performance changes due to fluctuations in temperature and voltage allowed. Notice as X increases the yield initially rises, reaches a maximum, and then begins fall. This makes finding the optimum yield straight forward.

Similar results can be obtained for strict worst-case performance constraints which do not allow temporary performance changes due to voltage and/or temperature fluctuations, as illustrated in Figure 16. Here, the yield varies in a similar manner but with smaller values than those in Figure 15. The difference between these two figures is summarized in Table V. In particular, BD/RO designs under worst-case process variation leads to 8% higher yield when SPQL equals 0.0005 and 12% higher yield when SPQL equals 0.005. However, a BD/RO design under worst-case process, voltage and temperature variation leads to 14% less yield when SPQL equals 0.0005 and 12% less yield when 0.005. We can further improve the yield for BD/RO design by applying per-chip analysis in Section V-C, which improve the yield by 5% to 8% and shown in the last row of Table V. In both cases, if we assume temporary performance changes caused by voltage and/or temperature fluctuations are allowed, we see significant yield advantages for BD/RO designs over SYNC designs. However, if a stricter criteria for performance is required, BD/RO designs lose their advantage over SYNC designs.

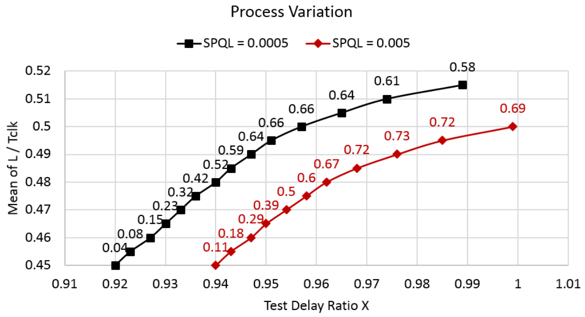

Fig. 15. Yield of BD/RO designs under worst-case performance constraints and process variation

Compared to average-case yield $Yield_{AVE}$ which is only affected by correlation coefficients, $Yield_{B-WC}$ is affected by both correlation coefficients and the mean and variance of the delays. This is illustrated in Figure 17, where lower correlation coefficients leads to smaller yield, and lower mean

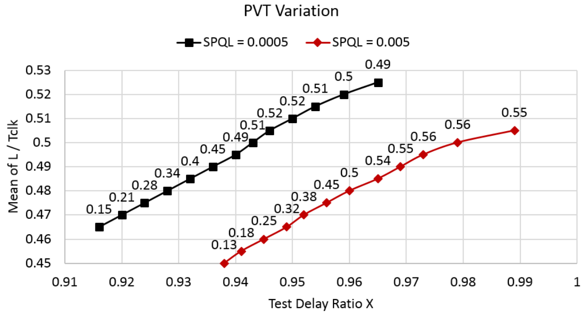

Fig. 16. Yield of BD/RO designs under worst-case performance constraints and PVT variations

TABLE V
YIELD OF BD/RO OVER SYNC GIVEN SPQL AND WORST-CASE PERFORMANCE CONSTRAINTS

	SPQL = 0.0005			SPQL = 0.005		
Yield	$SYNC_{PVT}$	B_{WCP}	B_{WCPVT}	$SYNC_{PVT}$	B_{WCP}	B_{WCPVT}
Uniform	0.61	0.66	0.52	0.65	0.73	0.56
Per-chip	0.65	0.71	0.56	0.69	0.77	0.59

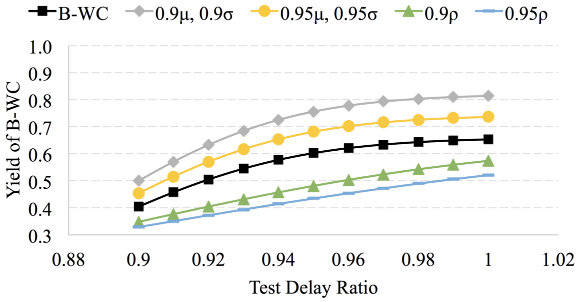

Fig. 17. Yield of B-WC versus Test Delay Ratio

and variance leads to higher yield. It indicates that if the mean of critical paths changes over years, or the technology node change affects correlation coefficients, $Yield_{B-WC}$ needs to be re-considered.

TABLE VI
YIELD COMPARISON UNDER LARGE HOLD TIME AND WORST-CASE PERFORMANCE CONSTRAINTS

	Tune L'			Add min delay constraint	
hold time (h)	0	$10\%Tclk$	$20\%Tclk$	$10\%Tclk$	$20\%Tclk$
$Yield_{B-WCP}$	0.66	0.54	0.47	0.54	0.50
$Yield_{B-WCPVT}$	0.52	0.42	0.25	0.43	0.32
$Yield_{SYNC}$	0.61	0.42	0.01	0.61	0.05

Similar to the yield comparison under significant hold time constraints in Section VIII-A, Table VI shows the yield comparison under both worst-case performance constraints as well as hold constraints. When hold time is 0, $Yield_{B-WCPVT}$ is smaller than $Yield_{SYNC}$ but $Yield_{B-WCP}$ is slightly bigger than $Yield_{SYNC}$. However, because of the ability to tune L', BD designs have a significant yield advantage when the hold time constraints increase to 20% of $Tclk$. Adding hold buffers improves the obtainable yields, but the yield advantage of BD/RO designs remains significant.

TABLE VII
ANALYSIS OF THE CRITICAL PATH UNDER TEST AND DELAY LINE OVER 9 YEARS

Year	0	3	6	9
μ_T/μ_L	0.903	0.904	0.903	0.903
σ_T/σ_L	0.931	0.931	0.931	0.931
$\rho_{T,L}$	0.899	0.899	0.899	0.899
Delay at year N over year 0	1.000	1.011	1.013	1.015

C. Aging analysis

Based on the NBTI model in Section VII, we run Monte Carlo simulation with global and local variations over a period of 9 years. Our goal is to determine how the mean and variance of the relative delays changes over the lifetime of the part. We explored whether these changes will impact the failure rate over time and how should we set the delay line in order to ensure functionality as the circuit ages.

Table VII shows the trend of delay of the critical path under test with a step size of 3 years. The delay of T and L increases 1% at the third year after being shipped. The delay then increases more slowly, becoming 1.5% larger at year 9 after being shipped. Both the mean and standard deviation of the delay ratio T over L remains the same. This means that aging can be viewed as a global variation that affects T and L quite similarly. Consequently, the correlation coefficient of T and L remains constant and aged asynchronous chips will likely remain functional as they age, although run a bit slower.

In comparison, to ensure synchronous chips remain functional over their life-time, the clock period or voltage must be conservatively set when shipped or altered over time. Otherwise, there is a significant chance that aged chips will fail.

In both SYNC and BD/RO design, however, if the performance constraint applies to the entire lifetime of the circuit, we should use the joint distribution of T, C and L from the Monte Carlo simulation that includes the aging variations. The analysis methods, however, are the same as in the non-aging case.

IX. CONCLUSION AND FUTURE WORK

Despite the plethora of research in bundled-data asynchronous designs and ring-oscillator-based synchronous circuits, their yield and $SPQL$ has not been mathematically explored in the literature. This paper proposes a mathematical model of their yield and compares them to that of comparable traditional synchronous designs under both average and worst-case performance constraints. The analysis is validated and quantified using the joint probability distributions obtained using Monte Carlo simulations of a carry select adder and MUX-based delay line in a 65nm technology.

The theory can guide designers to set the test margin in their designs to achieve a given SPQL as well as predict their resulting yield. The analysis can also guide design decisions by quantifying the benefits of co-locating delay lines and the associate combinational logic thereby increasing their correlation and using delay lines for which the test margins are programmable with high resolution.

More generally, this work describes a mathematical framework for analyzing test metrics of designs in which the delays of the clocking circuitry is correlated to the delays of the associated combinational logic. It thus forms the basis of several directions of future work. First, we can extend the theory to apply to other delay models, including log normal which may be more accurate in sub-threshold regions of operation [22], an increasingly important region of operation for asynchronous designs. Second, beyond these theoretical advances, our future work includes completing the physical design flow and post-silicon tuning procedures that target the programmable delay lines. An open source version of this flow is under development [23], [24].

References

[1] V. K. Sharma, M. Pattanaik, and B. Raj, "Pvt variations aware low leakage indep approach for nanoscale cmos circuits," *Microelectronics reliability*, vol. 54, no. 1, pp. 90–99, 2014.

[2] I. E. Sutherland, "Micropipelines," *Communications of the ACM*, vol. 32, no. 6, pp. 720–738, 1989.

[3] G. Russell, A. Yakovlev, A. Bystrov, D. Kinniment, and O. Maevsky, "On-chip structures for timing measurements and test," in *Proceedings of the 8th International Symposium on Asynchronus Circuits and Systems*. IEEE Computer Society, 2002, p. 190.

[4] P. A. Beerel, R. Ozdag, and M. Ferretti, *A Designer's Guide to Asynchronous VLSI*. Cambridge University Press, 2010.

[5] O. A. Petlin and S. B. Furber, "Scan testing of micropipelines," in *Proceedings of the 13th IEEE VLSI Test Symposium*, 1995, pp. 296–301.

[6] M. Abramovici, M. Breuer, and A. Friedman, *Digital Systems Testing and Testable Design*. John Wiley & Sons, 1990.

[7] N. K. Jha and S. Gupta, *Testing of Digital Systems*. Cambridge University Press, 2003.

[8] J. Xiong, V. Zolotov, C. Visweswariah, and P. A. Habitz, "Optimal test margin computation for at-speed structural test," *IEEE Trans. on CAD*, vol. 28, no. 9, pp. 1414–1423, 2009.

[9] C. Visweswariah, K. Ravindran, K. Kalafala, S. G. Walker, S. Narayan, D. K. Beece, J. Piaget, N. Venkateswaran, and J. G. Hemmett, "First-order incremental block-based statistical timing analysis," *IEEE Trans. on CAD*, vol. 25, no. 10, pp. 2170–2180, 2006.

[10] Y. You and J. Gu, "Exploiting accelerated aging effect for on-line configurability and hardware tracking," in *Design Automation Conference (ASP-DAC), 2017 22nd Asia and South Pacific*. IEEE, 2017, pp. 348–353.

[11] Y. Zhang, H. Zha, V. Sahir, H. Cheng, and P. A. Beerel, "Test margin and yield in bundled data and ring-oscillator based designs," in *23rd IEEE International Symposium on Asynchronous Circuits and Systems (ASYNC)*, May 2017.

[12] L. Li, K. Chakrabarty, S. Kajihara, and S. Swaminathan, "Efficient space/time compression to reduce test data volume and testing time for IP cores," in *18th IEEE International Conference on VLSI Design*, 2005, pp. 53–58.

[13] S. P. Mu, M. C. T. Chao, S. H. Chen, and Y. M. Wang, "Statistical framework and built-in self-speed-binning system for speed binning using on-chip ring oscillators," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 24, no. 5, pp. 1675–1687, May 2016.

[14] J. Cortadella, A. Kondratyev, L. Lavagno, and C. P. Sotiriou, "Desynchronization: Synthesis of asynchronous circuits from synchronous specifications," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 25, no. 10, pp. 1904–1921, 2006.

[15] J. Cortadella, M. Lupon, A. Moreno, A. Roca, and S. S. Sapatnekar, "Ring oscillator clocks and margins," in *ASYNC*, 2016.

[16] F. Kriebel, S. Rehman, M. Shafique, and J. Henkel, "ageOpt-RMT: compiler-driven variation-aware aging optimization for redundant multithreading," in *Proceedings of the 53rd Annual Design Automation Conference*. ACM, 2016, p. 46.

[17] J. B. Velamala, V. Ravi, and Y. Cao, "Failure diagnosis of asymmetric aging under NBTI," in *Computer-Aided Design (ICCAD), 2011 IEEE/ACM International Conference on*. IEEE, 2011, pp. 428–433.

[18] D. Gnad, M. Shafique, F. Kriebel, S. Rehman, D. Sun, and J. Henkel, "Hayat: Harnessing dark silicon and variability for aging deceleration and balancing," in *Design Automation Conference (DAC), 2015 52nd ACM/EDAC/IEEE*. IEEE, 2015, pp. 1–6.

[19] T.-B. Chan, W.-T. J. Chan, and A. B. Kahng, "Impact of adaptive voltage scaling on aging-aware signoff," in *Proceedings of the Conference on Design, Automation and Test in Europe*. EDA Consortium, 2013, pp. 1683–1688.

[20] J. Li and J. Draper, "Accelerating soft-error-rate (SER) estimation in the presence of single event transients," in *Proceedings of the 53rd Annual Design Automation Conference*. ACM, 2016, p. 55.

[21] S. Ercolani, M. Favalli, M. Damiani, P. Olivo, and B. Ricco, "Estimate of signal probability in combinational logic networks," in *European Test Conference, 1989., Proceedings of the 1st*. IEEE, 1989, pp. 132–138.

[22] P. Corsonello, F. Frustaci, M. Lanuzza, and S. Perri, "Over/undershooting effects in accurate buffer delay model for sub-threshold domain," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 61, no. 5, pp. 1456–1464, 2014.

[23] Y. Zhang, L. S. Heck, M. T. Moreira, D. Zar, M. Breuer, N. L. V. Calazans, and P. A. Beerel, "Challenges in building an open-source flow from rtl to bundled-data design," in *24th IEEE International Symposium on Asynchronous Circuits and Systems (ASYNC)*, May 2018.

[24] Y. Zhang, H. Zha, and H. Cheng, "Edge 1.0.2," https://github.com/nobodybutyou1/Edge, 2018.

APPENDIX A

The optimization problem for per-chip SYNC design given an SPQL limit q is described as follows

$$\max_{X(L)} P(T+s<(1+\beta)XT_{clk}) \tag{27}$$

subject to

$$P(C+s>(1+\beta)T_{clk} \mid T+s<(1+\beta)XT_{clk}) \leq q \tag{28}$$

where X is a function of L.

By re-writing the function using integrals we get

$$\max_{X(L)} \int_{-\infty}^{+\infty} \int_{-\infty}^{(1+\beta)XT_{clk}-s} P_{T,L}(t,l)\ dt\ dl \tag{29}$$

subject to

$$\int_{-\infty}^{+\infty} \int_{(1+\beta)T_{clk}-s}^{+\infty} \int_{-\infty}^{(1+\beta)XT_{clk}-s} P_{T,C,L}(t,c,l)\ dt\ dc\ dl$$
$$-\ q \int_{-\infty}^{+\infty} \int_{-\infty}^{(1+\beta)XT_{clk}-s} P_{T,L}(t,l)\ dt\ dl\ = 0$$

By solving the problem using Lagrangian equation we get

$$\begin{aligned} &L(a,b,\lambda) \\ =&(1+\lambda q) \int_{-\infty}^{+\infty} \int_{-\infty}^{(1+\beta)XT_{clk}-s} P_{T,L}(t,l) dt\ dl\ - \\ &\lambda \int_{-\infty}^{+\infty} \int_{(1+\beta)T_{clk}-s}^{+\infty} \int_{-\infty}^{(1+\beta)XT_{clk}-s} P_{T,C,L}(t,c,l) dt\ dc\ dl \end{aligned}$$

We define $H(X,l,\lambda)$ as

$$\begin{aligned} &(1+\lambda q) \int_{-\infty}^{(1+\beta)XT_{clk}-s} P_{T,L}(t,l) dt\ - \\ &\lambda \int_{(1+\beta)T_{clk}-s}^{+\infty} \int_{-\infty}^{(1+\beta)XT_{clk}-s} P_{T,C,L}(t,c,l) dt\ dc \end{aligned} \tag{30}$$

To reach optimal yield, it requires X to satisfy

$$\frac{\partial H(X,l,\lambda)}{\partial X} = 0 \tag{31}$$

Through the optimality condition we obtain the following for X

$$\begin{aligned} &(1+\lambda q) P_{T,L}((1+\beta)XT_{clk}-s,l)\ - \\ &\lambda \int_{(1+\beta)T_{clk}-s}^{+\infty} P_{T,C,L}((1+\beta)XT_{clk}-s,c,l) dc\ = 0 \end{aligned} \tag{32}$$

The equations can be further simplified as

$$\int_{(1+\beta)T_{clk}-s}^{+\infty} P_{C|T,L}(c|(1+\beta)XT_{clk}-s,l) dc\ = q + \frac{1}{\lambda} \tag{33}$$

The mean of the conditional Gaussian distribution can be expressed as

$$\hat{\mu} = \Phi^{-1}(q+\frac{1}{\lambda})\hat{\sigma} + (1+\beta)T_{clk} - s \tag{34}$$

$$\mathbf{V} = \left(\begin{smallmatrix} C \\ T \\ L \end{smallmatrix}\right) = \left(\begin{smallmatrix} C \\ V_{TL} \end{smallmatrix}\right), \quad \mu_{\mathbf{V}} = \left(\begin{smallmatrix} \mu_C \\ \mu_T \\ \mu_L \end{smallmatrix}\right) = \left(\begin{smallmatrix} \mu_C \\ \mu_{TL} \end{smallmatrix}\right) \tag{35}$$

and

$$\begin{aligned} \mathbf{\Sigma_V} &= \begin{pmatrix} \sigma_C^2 & \sigma_{CT} & \sigma_{CL} \\ \sigma_{CT} & \sigma_T^2 & \sigma_{TL} \\ \sigma_{CL} & \sigma_{TL} & \sigma_L^2 \end{pmatrix} \\ &= \begin{pmatrix} \sigma_C^2 & \mathbf{\Sigma_{C,TL}} \\ \mathbf{\Sigma_{C,TL}^T} & \mathbf{\Sigma_{TL}} \end{pmatrix} \end{aligned}$$

$$\hat{\mu} = \mu_c + \Sigma_{C,TL}\Sigma_{TL}^{-1}(V_{TL} - \mu_{TL}) \tag{36}$$

By combining two Equations 34 and 36, we can write X as

$$X = \gamma L + \eta \tag{37}$$

where

$$\gamma = \frac{\sigma_{CT}\sigma_{TL} - \sigma_{CL}\sigma_T^2}{(1+\beta)T_{clk}(\sigma_{CT}\sigma_L^2 - \sigma_{CL}\sigma_{TL})} \tag{38}$$

η can can be obtained by substituting X in equation 28 by $\gamma r + \eta$.

$$\begin{aligned} \eta =& -\gamma\mu_L + \frac{1}{\sigma_{CT}\sigma_L^2 - \sigma_{CL}\sigma_{TL}} + \frac{\mu_T}{(1+\beta)T_{clk}} \\ &+ \frac{(\sigma_T^2\sigma_L^2 - \sigma_{TL}^2)[\Phi^{-1}(q+\frac{1}{\lambda})\hat{\sigma} - s - \mu_C]}{(\sigma_{CT}\sigma_L^2 - \sigma_{CL}\sigma_{TL})(1+\beta)T_{clk}} \end{aligned} \tag{39}$$

APPENDIX B

The problem of maximizing yield for per-chip BD/RO designs given an SPQL limit q is described as follows

$$\max_{X(l)} P(T+s<XL) \tag{40}$$

subject to

$$P(C>L \mid T+s<XL) \leq q \tag{41}$$

where X is a function of l, a per-chip measure of the delay line.

Recall that both yield and $SPQL$ are monotonically increasing functions of X [11]. Consequently, the yield is maximized when $P(C>L|T<XL)$ is set to q, the required SPQL. The above maximization problem can be re-written as follows.

$$\max_{X(l)} \int_{-\infty}^{+\infty} \int_{-\infty}^{Xl-s} f_{T,L}(t,l)\ dt\ dl \tag{42}$$

subject to

$$\begin{aligned}&\int_{-\infty}^{+\infty}\int_{l-s}^{+\infty}\int_{-\infty}^{Xl-s} f_{T,C,L}(t,c,l)\ dt\ dc\ dl\\ &-q\int_{-\infty}^{+\infty}\int_{-\infty}^{Xl-s} f_{T,L}(t,l)\ dt\ dl\ = 0\end{aligned}$$

The problem can be solved using the Lagrangian method as follows.

$$\begin{aligned}&L(a,b,\lambda)\\ =&(1+\lambda q)\int_{-\infty}^{+\infty}\int_{-\infty}^{Xl-s} f_{T,L}(t,l)dt\ dl\\ &-\lambda\int_{-\infty}^{+\infty}\int_{l-s}^{+\infty}\int_{-\infty}^{Xl-s} f_{T,C,L}(t,c,l)dt\ dc\ dl\end{aligned}$$

We define $H(X,l,\lambda)$ as

$$\begin{aligned}&(1+\lambda q)\int_{-\infty}^{Xl-s} f_{T,L}(t,l)dt-\\ &\qquad\lambda\int_{l-s}^{+\infty}\int_{-\infty}^{Xl-s} f_{T,C,L}(t,c,l)dt\ dc\end{aligned} \tag{43}$$

To obtain the optimal yield, X should satisfy

$$\frac{\partial H(X,l,\lambda)}{\partial X} = 0 \tag{44}$$

Through the optimality condition we obtain the following equation for X

$$(1+\lambda q)f_{T,L}(Xl-s,l) - \lambda\int_{l-s}^{+\infty} f_{T,C,L}(Xl-s,c,l)dc = 0 \tag{45}$$

This equation can be further simplified as follows.

$$\int_{l-s}^{+\infty} f_{C|T,L}(c|Xl-s,l)dc = q + \frac{1}{\lambda} \tag{46}$$

To solve this equation, we introduce the following definitions.

$$\mathbf{V} = \left(\begin{smallmatrix}C\\T\\L\end{smallmatrix}\right) = \left(\begin{smallmatrix}C\\V_{TL}\end{smallmatrix}\right),\quad \mu_{\mathbf{V}} = \left(\begin{smallmatrix}\mu_C\\\mu_T\\\mu_L\end{smallmatrix}\right) = \left(\begin{smallmatrix}\mu_C\\\mu_{TL}\end{smallmatrix}\right) \tag{47}$$

and

$$\begin{aligned}\mathbf{\Sigma_V} &= \begin{pmatrix}\sigma_C^2 & \sigma_{CT} & \sigma_{CL}\\ \sigma_{CT} & \sigma_T^2 & \sigma_{TL}\\ \sigma_{CL} & \sigma_{TL} & \sigma_L^2\end{pmatrix}\\ &= \begin{pmatrix}\sigma_C^2 & \mathbf{\Sigma_{C,TL}}\\ \mathbf{\Sigma_{C,TL}^T} & \mathbf{\Sigma_{TL}}\end{pmatrix}\end{aligned}$$

The mean of the conditional Gaussian distribution can be calculated from the above definitions.

$$\hat{\mu} = \mu_C + \mathbf{\Sigma_{C,TL}}\mathbf{\Sigma_{TL}^{-1}}(\mathbf{V_{TL}} - \mu_{\mathbf{TL}}) \tag{48}$$

Based on the conditional Gaussian distribution in Equation 46, the mean of the distribution can also be expressed as

$$\hat{\mu} = \Phi^{-1}(q + \frac{1}{\lambda})\hat{\sigma} + l - s \tag{49}$$

By combining Equations 48 and 49, we can write X as

$$X = \frac{\gamma}{l} + \eta \tag{50}$$

where

$$\eta = \frac{\sigma_T^2\sigma_L^2 - \sigma_{TL}^2 + \sigma_{CT}\sigma_{TL} - \sigma_{CL}\sigma_T^2}{\sigma_{CT}\sigma_L^2 - \sigma_{CL}\sigma_{TL}} \tag{51}$$

η is known and γ is a function of λ. To obtain the value of λ, we can directly substitute X in Equation 41 by $\frac{\gamma}{l} + \eta$.

$$\begin{aligned}\gamma =&\mu_T + \frac{(-\sigma_{CT}\sigma_{TL} + \sigma_{CL}\sigma_T^2)\mu_L}{\sigma_{CT}\sigma_L^2 - \sigma_{CT}\sigma_{TL}}\\ &+ \frac{(\sigma_T^2\sigma_L^2 - \sigma_{TL}^2)[\Phi^{-1}(q+\frac{1}{\lambda})\hat{\sigma} - s - \mu_L]}{\sigma_{CT}\sigma_L^2 - \sigma_{CT}\sigma_{TL}}\end{aligned} \tag{52}$$