An Extension of Probabilistic Simulation for Reliability Analysis of CMOS VLSI Circuits*

Farid N. Najm†, Ibrahim N. Hajj‡, and Ping Yang†

†Semiconductor Process & Design Center
Texas Instruments Inc.
Dallas, Texas 75265

‡Coordinated Science Laboratory
University of Illinois at Urbana-Champaign
Urbana, Illinois 61801

Abstract

The *probabilistic simulation* approach [1] is extended to include the computation of the *variance waveform* of the power/ground current, in addition to its *expected waveform*. To provide the motivation for doing this, we focus on the problem of estimating the median time-to-failure (MTF) due to *electromigration* in the power and ground busses of CMOS circuits. New theoretical results are presented that quantify the relationship between the MTF and the statistics of the stochastic current. This leads to a more accurate estimate of the MTF that requires both the expected and variance waveforms. A novel technique is then presented to compute the variance waveform for CMOS circuits, which has been incorporated into the probabilistic simulator CREST [1]. We show results of this implementation, demonstrating efficiency and accuracy on a number of circuits. We also use these results to study the importance of the variance waveform by estimating its contribution to the MTF relative to that of the expected waveform.

*This work was supported by Texas Instruments Inc. and the US Air Force Rome Air Development Center.

Submitted to IEEE Transactions on Computer-Aided Design, 1990.

1. Introduction

Reliability, already a major concern in integrated circuit design, can only become more important in the future. As higher levels of integration are used, metal line width and line separation will continue to decrease, thereby increasing a chip's susceptibility to failures resulting from line shorts or opens.

While the results to be presented can be used to study a variety of reliability problems, we will illustrate their utility by focusing on the problem of *electromigration* [2, 3]. Electromigration (EM) is a major reliability problem caused by the transport of atoms in a metal line due to the electron flow. Under persistent current stress, EM can cause deformations of the metal lines which may result in shorts or open circuits. The failure rate due to EM depends on the current density in the metal lines and is usually expressed as a median time-to-failure (MTF). There is a need for CAD tools that can predict the susceptibility of a given design to EM failures.

In [1] we presented a novel technique for MTF estimation based on a *stochastic current waveform* model. The implementation of this technique in the program CREST (CuRrent ESTimator) has proven to be very effective both in terms of accuracy and speed. In the interest of clarity, we will review some of the basic concepts behind this approach. The reader is referred to [1, 4] for a more detailed description.

We focus on the power and ground busses and derive currents for them to be used for MTF estimation. The argument presented in [1] is that the desired current waveform in any branch of the bus is one that combines the effects of all possible waveforms at the circuit primary inputs. By considering the set of logical waveforms allowed at the circuit inputs as a *probability space* [5], the current in any branch of the bus becomes a *stochastic process*. CREST derives the *expected* (or *mean) waveform* (not a time-average) of this process, which we call an *expected current waveform*, $E[\mathbf{i}(t)]$. This is a waveform whose value at any given time is the weighted average of all possible current values at that time. CREST uses *statistical* information about the inputs to directly derive $E[\mathbf{i}(t)]$. The resulting methodology is what we call a *probabilistic simulation* of the circuit. In general, it can be slightly more time consuming than a single timing simulation run, but it needs to be applied only once, resulting in significant speedup.

While the feasibility of performing probabilistic simulation was established in [1], the justification for using $E[\mathbf{i}(t)]$ to estimate the MTF was based mainly on qualitative arguments. In section 2 of this paper we quantify these arguments and present new theoretical results that specify the relationship between the MTF and the statistics of the stochastic current. This leads to an efficient and more accurate technique for deriving the MTF which requires the *variance waveform*, $V[\mathbf{i}(t)]$, of the stochastic current in addition to its expected waveform. $V[\mathbf{i}(t)]$ is a waveform whose value at any given time is the variance of the current values at that time, $V[\mathbf{i}(t)] = E\left[(\mathbf{i}(t) - E[\mathbf{i}(t)])^2\right]$. It is an indication of the spread of the real current waveforms around their expected waveform.

In section 3 we present a novel technique for deriving the variances of the individual gate currents in CMOS circuits. This has been implemented in the probabilistic simulator, CREST [1]. Section 4 and the appendix contain a discussion of how the bus current variance waveforms may be obtained from those of the individual gates. In section 5, we present some results of our implementation, and use them to study the importance of the variance waveform. This is done by estimating its contribution to the MTF relative to that of the expected waveform. The appendix also presents several approximations that can be used when handling large chips to simplify the variance computations, and thus make it possible to handle VLSI circuits.

Several simplifying assumptions and/or approximations will be made in the following sections to make the problem computationally tractable. Whenever possible, we will attempt to justify these assumptions. However, for lack of space, this will not always be possible, and the reader will be referred to appropriate references. Nevertheless, we will offer a verification of the overall approach on a *global* scale, by comparing the end result of the simulation (variance waveform) from CREST with that derived using SPICE. The extended CREST program, with the variance computation built in, maintains its excellent performance compared to traditional approaches : we demonstrate a speedup (over SPICE) of over 11500X on a 648-transistor CMOS parallel multiplier circuit. Preliminary results of this research have appeared in [6] and [7].

2. Stochastic Current Waveforms and the MTF

Consider a metal line of uniform width and thickness carrying a *constant* current. Due to electromigration (EM), the line will fail after a period of time. The time required for 50% of a large population of such lines to fail is called the median-time-to-failure (MTF), also denoted by t_{50}. Due to the nature of the distribution (log-normal) of EM failure times, it turns out that the MTF is also approximately equal to the mean-time-to-failure. The two names are used interchangeably in the EM literature. The relationship between the MTF, t_{50}, and the current density j in the line has been extensively studied, and shown to be a complex nonlinear function [8], as shown in Fig. 1. We will consider the MTF to be $t_{50} \propto 1/f(j)$ where j is in A/cm^2, and f is the dimensionless nonlinear function shown in Fig. 2.

If a metal line carries a *varying* current, of density $j(t)$, then the MTF is $t_{50} \propto 1/J_{\text{eff}}$, where J_{eff} depends both on f and on the waveform $j(t)$. It has been suggested [9] that, if the waveform is periodic, with period T, and if every period consists of a train of pulses $p = 1, \ldots, m$ of heights j_p and duration t_p, then $J_{\text{eff}} = \sum_{p=1}^{m} \frac{t_p}{T} \tilde{f}(j_p)$, where $\tilde{f}(j)$ is chosen to be one of the three dotted line approximations to $f(j)$ in Fig. 2, depending on the value of j_p, as follows. If $j \leq 10^5 A/cm^2$ then $\tilde{f}(j) = j$, if $10^5 A/cm^2 \leq j \leq 10^6 A/cm^2$ then $\tilde{f}(j) \propto j^{3/2}$, and if $j \geq 10^6 A/cm^2$ then $\tilde{f}(j) \propto j^2$. A more accurate expression, however, can be written in terms of f itself, as follows :

$$J_{\text{eff}} = \sum_{p=1}^{m} \frac{t_p}{T} f(j_p). \tag{2.1}$$

For a general periodic waveform, we take the summation to the limit and write :

$$J_{\text{eff}} = \frac{1}{T} \int_0^T f(j) dt. \tag{2.2}$$

If the current waveform is not periodic, then better estimates of J_{eff} are obtained by using larger values of T so that more features of the waveform are included. Therefore one can write :

$$J_{\text{eff}} = \lim_{T \to \infty} \frac{1}{T} \int_0^T f(j) dt. \tag{2.3}$$

Now suppose that the current waveform is *stochastic*, i.e., it is a stochastic process $\mathbf{j}(t)$, that represents a family of deterministic (real) current waveforms $j_k(t)$, with associated probabilities P_k, $k = 1, \ldots, N$, over the (finite) interval $[0, t_0]$. Based on this information, we can build a (non-stochastic) current waveform $j(t)$, over $[0, T]$ as $T \to \infty$, that is indicative of the current during typical operation as follows. Consider a random sequence of the waveforms $j_k(t)$, each being shifted in time, spanning an interval of length t_0, and occurring with its assigned probability P_k, as shown in Fig. 3. Let $n_k(T)$ be the (integer) number of occurrences of the waveform $j_k(t)$ in $[0, T]$, and let $n_T = \lfloor T/t_0 \rfloor$. If J_k, $k = 1, \ldots, N$ are defined as follows :

$$J_k \stackrel{\triangle}{=} \frac{1}{t_0} \int_0^{t_0} f(j_k) dt, \tag{2.4}$$

then :

$$J_{\text{eff}} = \lim_{T \to \infty} \frac{1}{T} \int_0^T f(j) dt = \lim_{n_T \to \infty} \sum_{k=1}^{N} J_k \frac{n_k(T)}{n_T} = \sum_{k=1}^{N} J_k \lim_{n_T \to \infty} \Big[\frac{n_k(T)}{n_T} \Big]. \tag{2.5}$$

By the law of large numbers [5], $\lim_{n_T \to \infty} [n_k(T)/n_T] = P_k$, which leads to :

$$J_{\text{eff}} = \sum_{k=1}^{N} \Big[\frac{1}{t_0} \int_0^{t_0} f(j_k) dt \Big] P_k = \frac{1}{t_0} \int_0^{t_0} \Big[\sum_{k=1}^{N} f(j_k) P_k \Big] dt \tag{2.6}$$

and, finally :

$$J_{\text{eff}} = \frac{1}{t_0} \int_0^{t_0} E\big[f(\mathbf{j}) \big] dt \tag{2.7}$$

where $E[\ ]$ denotes the expected value operator. Therefore, *the MTF due to a stochastic current depends only on the expected waveform of a nonlinear function of the current density.*

Since f is nonlinear, $E\big[f(\mathbf{j})\big]$ is not easy to evaluate. At low current density values, where f is linear (Fig. 2), $E\big[f(\mathbf{j})\big] = f(E[\mathbf{j}])$. In this case, the expected current waveform $E[\mathbf{j}]$ derived in [1] is indeed the correct waveform for MTF estimation. In general, f is nonlinear, and a generalized approach will be developed below.

At any time t, the process $\mathbf{j}(t)$ is a random variable $\mathbf{j}$ with mean $\eta_j \stackrel{\triangle}{=} E[\mathbf{j}]$, and variance $\sigma_j^2 \stackrel{\triangle}{=} E\big[(\mathbf{j}-\eta_j)^2\big]$. In general, the pth moment of $\mathbf{j}$ is $\mu_{j,p} \stackrel{\triangle}{=} E\big[(\mathbf{j}-\eta_j)^p\big]$. To estimate $E\big[f(\mathbf{j})\big]$, a Taylor series expansion of f gives [5] :

$$E\big[f(\mathbf{j})\big] \approx f(\eta_j) + f''(\eta_j)\frac{\sigma_j^2}{2} + \cdots + f^{(p)}(\eta_j)\frac{\mu_{j,p}}{p!}. \tag{2.8}$$

When f is linear, this reduces to :

$$E\big[f(\mathbf{j})\big] = f(E[\mathbf{j}]), \tag{2.9}$$

as observed above. Therefore, using the expected current waveform for MTF estimation [1] amounts to making a first-order approximation in (2.8). Naturally, higher order approximations would lead to better results. In particular, if f is approximated by a quadratic in the neighborhood of η_j, then :

$$E\big[f(\mathbf{j})\big] \approx f(\eta_j) + f''(\eta_j)\frac{\sigma_j^2}{2}. \tag{2.10}$$

This approximation becomes exact if $f(j)$ is represented by the straight lines corresponding to j^1 and j^2 in Fig. 2. It is more accurate than (2.9) since it covers a wider range of currents. As a result, *equations (2.10) and (2.7) offer a new, more accurate technique for computing the MTF.* In order to make use of this technique, we need to derive the variance of the current waveform in addition to its expected value. As pointed out in the introduction, the estimation of the expected current waveform has already been described in our previous work [1]; the following sections discuss the derivation of the variance waveform.

3. Derivation of the Gate Variance Waveforms

We will briefly review the probabilistic simulation approach [1], which follows an event-driven simulation strategy. *Probability waveforms*, which represent a large number of logic waveforms, are applied at the primary inputs and propagated through the circuit as a sequence of *probabilistic events.* A probabilistic event embodies a number of possible logical transitions. Whenever a gate is simulated, the events at its inputs are used to derive an event at its output and an *expected current pulse* $E[\mathbf{i}(t)]$, to be added to the global expected

current waveform. This pulse is modeled as a *triangular pulse* that starts with a *peak* value $E[\mathbf{I}]$ at the time of transition and decays linearly to zero after a time interval called the *time span*. The variance waveform can be derived with little modification to the overall simulation strategy. Whenever an expected pulse is derived for a gate, a *variance pulse* will be derived as well.

Figure 4 shows a generic CMOS gate. The p-block or p-part (n-block or n-part) of a gate refers to the p (n)-channel transistor mesh between its output node and the power supply (ground). A gate will be assumed to have *independent* inputs. While this may be true at the primary circuit inputs, it is not true in general. However, the general case is handled using the concept of a supergate [1], with the independent-inputs-gate-solver used as a subroutine.

As in [1], we only consider the *charging* component of the power supply or ground currents. The output node capacitance is split into two lumped capacitors C_p to V_{dd} and C_n to V_{ss}. Similarly, each internal gate node n_i has two capacitors C_{in} and C_{ip}. Capacitance values are derived from the circuit description and the transistor model parameters. On a low-to-high transition, the currents flowing through C_n and C_p at the output node are i_{p1} and i_{p2}, respectively, as shown in the figure. The corresponding i_{n1} and i_{n2} for a high-to-low transition are also shown. The currents i_{p2} and i_{n2} are discharging currents that redistribute locally, and we are interested in $i = i_{p1} + i_{n1}$, which is the current associated with the output node. The total gate current i_{tot} will be larger than i since it also contains the currents needed to charge/discharge the C_{in} & C_{ip} capacitors at the internal nodes. However, the output current plays a central role in the derivation.

The variance waveforms for the gate total and output currents will be modeled by triangular pulses $V[\mathbf{i_{tot}}(t)]$ and $V[\mathbf{i}(t)]$, respectively, with peak values of $V[\mathbf{I_{tot}}]$ and $V[\mathbf{I}]$. If an event occurs at the gate input at time t, then we denote by t^- and t^+ the instances of time immediately before and after the event, respectively. Focusing for now on the output current pulse, its variance waveform starts with a peak of $V[\mathbf{I}] = V[\mathbf{i}(t^+)]$ at time t and decays linearly to zero at time $t + \tau$. Since $V[\mathbf{I}] = E[\mathbf{I}^2] - E[\mathbf{I}]^2$ [5], and since CREST already derives the expected pulse peak $E[\mathbf{I}]$, we will concentrate here on the derivation of $E[\mathbf{I}^2]$.

Let $i_p = i_{p1} + i_{p2}$ and $i_n = i_{n1} + i_{n2}$. It is easy to verify that $i_{p1} = i_p \times C_n/(C_p + C_n)$, and $i_{n1} = i_n \times C_p/(C_p + C_n)$. Therefore :

$$E[\mathbf{i}^2(t)] = E[\mathbf{i_p}^2(t)]\left(\frac{C_n}{C_p + C_n}\right)^2 + E[\mathbf{i_n}^2(t)]\left(\frac{C_p}{C_p + C_n}\right)^2 \tag{3.1}$$

The term containing $E[\mathbf{i_p}(t)\mathbf{i_n}(t)]$ is omitted (it is zero) since at least one of the charging currents is zero at any given time. In particular, the value at the peak is :

$$E[\mathbf{I}^2] = E[\mathbf{I_p}^2] \times \left(\frac{C_n}{C_p + C_n}\right)^2 + E[\mathbf{I_n}^2] \times \left(\frac{C_p}{C_p + C_n}\right)^2 \tag{3.2}$$

The values of $E[\mathbf{I_p}^2]$ and $E[\mathbf{I_n}^2]$ are derived as follows. For $E[\mathbf{I_p}^2]$, consider the p-part of the gate, and let every transistor T_k be represented by a switch of on-conductance $g_{on,k}$ [1]. Based on this switch-network model of the p-block, let $\mathbf{G_p}(t)$ be the random conductance between the output node and V_{dd}. $\mathbf{G_p}$ is a function of the individual transistor random conductances $\mathbf{g_k}$, where $\mathbf{g_k}$ is 0 if the transistor is off and $g_{on,k}$ if it is on. If an event occurs at a gate input at time t, then the value of $\mathbf{G_p}(t^+)$ and the previous state of the output node, $\mathbf{V_o}(t^-)$, will determine $\mathbf{I_p}$. Formally, we have $E[\mathbf{I_p}^2] = E[(V_{dd} - \mathbf{V_o}(t^-))^2 \times \mathbf{G_p}^2(t^+)]$, which becomes :

$$E[\mathbf{I_p}^2] = V_{dd}^2 \times E[\mathbf{G_p}^2(t^+) \mid \mathbf{G_p}(t^-) = 0] \times P(\mathbf{G_p}(t^-) = 0) \tag{3.3}$$

where $P(A)$ is the probability of the event A, and $E[A \mid B]$ denotes the *conditional expected value* [5] of A given B. The formula is correct because if $G_p(t^-) = 0$ ($\neq 0$) then $V_o(t^-) = 0$ (V_{dd}). Similarly for the n-part of the gate, we get :

$$E[\mathbf{I_n}^2] = V_{dd}^2 \times E[\mathbf{G_n}^2(t^+) \mid \mathbf{G_n}(t^-) = 0] \times P(\mathbf{G_n}(t^-) = 0) \tag{3.4}$$

To derive the conditional expectations, consider a graph representation of the p-block (or n-block), where every edge in the graph is labeled with $E[\mathbf{g_k}^2(t^+) \mid \mathbf{G_p}(t^-) = 0]$, $E[\mathbf{g_k}(t^+) \mid \mathbf{G_p}(t^-) = 0]$, and the gate node probabilities of its corresponding transistors. The details of how these quantities can be derived for every transistor can be found in [4]. Then perform a *graph reduction* operation [1, 4], which, simply stated, involves a number of series/parallel combinations and node eliminations that reduce the graph to a single edge, whose labels are the required statistics $E[\mathbf{G_p}^2(t^+) \mid \mathbf{G_p}(t^-) = 0]$ and

$E[\mathbf{G_p}(t^+) \mid \mathbf{G_p}(t^-) = 0]$. Similar work can be done for the n-block. As a result, we have the peak value, $V[\mathbf{I}] = E[\mathbf{I^2}] - E[\mathbf{I}]^2$, of the output current variance pulse.

The time span τ is found by first solving for the area under the $V[\mathbf{i}(t)]$ pulse. Notice that, if $i(t)$ is a triangular pulse of height I and area q, then :

$$\int_0^\infty i^2(t)dt = \frac{2}{3}Iq \tag{3.5}$$

In this case, q is equal to the charge delivered to (or from) the output node capacitors. From this it follows that :

$$\int_0^\infty E[\mathbf{i}^2(t)]dt = \frac{2}{3}E[\mathbf{Iq}], \text{ and } \int_0^\infty E[\mathbf{i}(t)]^2dt = \frac{2}{3}E[\mathbf{I}]E[\mathbf{q}]. \tag{3.6}$$

The second equation follows since $E[\mathbf{i}(t)]$ is a triangular pulse of height $E[\mathbf{I}]$ and area $E[\mathbf{q}]$. Therefore, the variance pulse has an area :

$$\frac{V[\mathbf{I}] \times \tau}{2} = \int_0^\infty V[\mathbf{i}(t)]dt = \frac{2}{3}\Big(E[\mathbf{Iq}] - E[\mathbf{I}]E[\mathbf{q}]\Big). \tag{3.7}$$

The value of $E[\mathbf{Iq}]$ can be written as :

$$E[\mathbf{Iq}] = E[(\mathbf{I_{p1}} + \mathbf{I_{n1}}) \times (\mathbf{q_{p1}} + \mathbf{q_{n1}})]. \tag{3.8}$$

where $\mathbf{I_{p1}}$ ($\mathbf{I_{n1}}$) is the peak of $\mathbf{i_{p1}}(t)$ ($\mathbf{i_{n1}}(t)$), and $\mathbf{q_{p1}}$ ($\mathbf{q_{n1}}$) is the charge delivered by $\mathbf{i_{p1}}(t)$ ($\mathbf{i_{n1}}(t)$). Since $\mathbf{q_{p1}}$ ($\mathbf{q_{n1}}$) is equal to $V_{dd}C_n$ ($V_{dd}C_p$) if $\mathbf{i_{p1}}(t)$ ($\mathbf{i_{n1}}(t)$) is non-zero, and is otherwise zero, then :

$$E[\mathbf{Iq}] = \frac{V_{dd}C_n^2}{C_p + C_n}E[\mathbf{I_p}] + \frac{V_{dd}C_p^2}{C_p + C_n}E[\mathbf{I_n}]. \tag{3.9}$$

The time span of the gate output current variance pulse is, therefore :

$$\tau = \frac{4}{3}\left(\frac{E[\mathbf{Iq}] - E[\mathbf{I}]E[\mathbf{q}]}{V[\mathbf{I}]}\right). \tag{3.10}$$

An expression similar to (3.7) can be written for the gate total current, $i_{tot}(t)$, as follows :

$$\int_0^\infty V[\mathbf{i_{tot}}(t)]dt = \frac{2}{3}\Big(E[\mathbf{I_{tot}q_{tot}}] - E[\mathbf{I_{tot}}]E[\mathbf{q_{tot}}]\Big). \tag{3.11}$$

Unfortunately, $E[\mathbf{I_{tot}q_{tot}}]$ does not have as simple an expression as was found for $E[\mathbf{Iq}]$. We have chosen to use a conservative estimate based on the following assumption : whenever a node in the p-block (n-block) is charged to V_{dd} (V_{ss}), then every other node in the p-block (n-block) is also charged to V_{dd} (V_{ss}). This assumption is true for simple gates, and may over-estimate the current-charge product in more complex cases. Based on this assumption, one can show [4] that :

$$E[\mathbf{I_{tot}q_{tot}}] \approx \frac{E[\mathbf{q_{p,tot}}]}{E[\mathbf{q_p}]}\frac{Q_p C_n}{C_p + C_n}E[\mathbf{I_p}] + \frac{E[\mathbf{q_{n,tot}}]}{E[\mathbf{q_n}]}\frac{Q_n C_p}{C_p + C_n}E[\mathbf{I_n}], \tag{3.12}$$

where $E[\mathbf{q_p}]$ and $E[\mathbf{q_n}]$ are available as equations (3.12) & (3.13) in [1], $E[\mathbf{q_{p,tot}}]$ and $E[\mathbf{q_{n,tot}}]$ are, respectively, the first and second summations in equation (3.10) in [1], and

$$Q_p = \sum_{i \in P \text{ block}} V_{dd} C_{in}, \; Q_n = \sum_{i \in N \text{ block}} V_{dd} C_{ip}. \tag{3.13}$$

As was assumed for the expected current pulse [1], we let the time span of the gate total current variance pulse be equal to that derived for the gate output current. Therefore :

$$V[\mathbf{I_{tot}}] = \left(\frac{E[\mathbf{I_{tot}q_{tot}}] - E[\mathbf{I_{tot}}]E[\mathbf{q_{tot}}]}{E[\mathbf{Iq}] - E[\mathbf{I}]E[\mathbf{q}]} \right) \times V[\mathbf{I}]. \tag{3.14}$$

4. Estimating the Variance Current Waveforms in the Bus

In the two previous sections, we presented the motivation for computing the variance waveform, and a procedure for computing the variance pulse for a CMOS gate. Ultimately, the variance waveform (of the current density) in every branch of the power/ground bus is required. In this section we present a technique for deriving the bus variance waveforms from those of the individual gates. Since the current density $\mathbf{j}(t)$ in any branch of the power or ground bus is directly proportional to the current $\mathbf{i}(t)$ in that branch, we will discuss the derivation of $\sigma_i^2(t)$ (i.e., $V[\mathbf{i}(t)]$) rather than $\sigma_j^2(t)$ (i.e., $V[\mathbf{j}(t)]$). Furthermore, we will only discuss the power bus since the ground bus analysis is similar.

We assume that gates are tied to the bus at certain points, called *contacts*. Several gates may be tied to the same contact. The current in a branch of the bus, $\mathbf{i}(t)$, is a function of the

currents being drawn off the contacts, $\mathbf{i_j}(t), j = 1, \ldots, n$. Each of these is, in-turn, simply the sum of the individual gate currents tied to that contact :

$$\mathbf{i_j}(t) = \mathbf{i_{j1}}(t) + \cdots + \mathbf{i_{jk}}(t). \tag{4.1}$$

In the framework of our probabilistic simulation technique, the process of deriving the variance waveforms consists of three steps :

-1- Using the probabilistic events at the inputs to each logic gate, derive its variance pulse.

-2- Combine the pulses at each contact point to derive the variances of the contact currents.

-3- Using the bus topology, and the variances of the contact currents, derive the variances of the bus branch currents.

Step 1 has been described in section 3; the other two steps will be described below.

A critical issue in computing the variance is the correlation between the different current waveforms. Since such correlation is too expensive to derive for VLSI circuits, we will occasionally be making conservative approximations to simplify the problem. Our experience with the probabilistic simulation approach [1] suggests that neglecting the correlation between different current waveforms gives good results in most cases, especially for large circuits. In general, there will be cases where this becomes a poor assumption, as in clocked circuits, for example, where different parts of the circuit may switch in unison in response to a central clock. However, since the variance is of secondary importance in (2.10), and since keeping track of the correlation is too expensive for large VLSI circuits, we have opted to make this accuracy-efficiency trade-off.

Based on this, we assume that the gate currents tied to the same contact are *uncorrelated*. This immediately provides a simple solution for step 2, using (4.1), as follows :

$$\sigma^2_{i_j}(t) = \sigma^2_{i_{j1}}(t) + \cdots + \sigma^2_{i_{jk}}(t). \tag{4.2}$$

Thus, the variance pulses from the individual gates are simply added to provide the contact current variance waveform (step 2). The implementation and results presented in the next section are based on (4.2).

Step 3 is considerably more complex, and is left to the appendix, where a number of approximations are also proposed to make possible an efficient implementation. The present

implementation of CREST provides the user with the expected and variance waveforms for the contact currents (steps 2 and 3). The proper place for the implementation of step 3 is in the SPIDER [10] program, which takes the currents generated by CREST and uses the bus layout information to estimate the MTF.

5. Implementation and Results

The variance calculation technique outlined above has been implemented in CREST. We present below the results of CREST runs on a variety of circuits, showing both waveform comparisons and timing performance.

To assess the accuracy of the results, it is important to make a fair comparison with a variance waveform derived using a valid simulation tool. To do so, we have generated the variance waveform for a variety of examples by running SPICE for every set of input voltage signals allowed by the probability vectors (see [1], section 2.1), deriving the expected current waveform $E[\mathbf{i}(t)]$ by doing a time-point averaging of the results, and then using that to find the variance as the time-point average of $(i - E[\mathbf{i}(t)])^2$. Since electromigration models for non-dc waveforms are still controversial, it makes little sense to shoot for perfect accuracy in the current waveforms. Furthermore, it is important to make a trade-off between accuracy and efficiency to make possible the solution of large circuits. Hence, our objective has been to derive, in a very *short time*, a waveform that matches reasonably well the peak and general shape of the SPICE waveform. Since the number of required SPICE simulation runs grows exponentially with the number of circuit inputs, the comparisons to be presented will necessarily be limited to medium sized circuits. There is no reason to suspect, however, that the accuracy observed on these circuits will deteriorate on larger ones.

Waveform comparisons are shown in Figs. 5 and 6 for a single nand gate and a complex gate, respectively. The comparisons for two larger circuits are shown in Fig. 7 (for an exclusive-or circuit) and Fig. 8 (for a 54-MOSFET 2-bit ripple adder circuit).

For our next example, we consider a 648-MOSFET 4-bit parallel multiplier. This circuit is too big to make the 2^{16} required SPICE simulations. We will, therefore, show the results of two different CREST runs on this circuit. In Fig. 9 we compare a full accuracy CREST run and a heuristic [1] CREST run in which all internal nodes of the multiplier

were assumed independent (uncorrelated). The excellent agreement demonstrates that the correlation between different current pulses in large circuits may indeed be neglected.

These results can be used to assess the significance of the contribution of the variance waveform to the MTF, i.e., to see whether the gain in accuracy is worth the effort. Recall that the expected and variance waveforms combine to provide a J_{eff} effective current density value for MTF estimation, according to (2.10) and (2.7). We have measured the variance contribution as the increase in J_{eff} due to the variance waveform, divided by J_{eff} using only the expected waveform, as follows :

$$\frac{\Delta J_{\text{eff}}}{J_{\text{eff}}} \triangleq \frac{J_{\text{eff}}(\text{using } E[\mathbf{j}] \ \& \ V[\mathbf{j}]) - J_{\text{eff}}(\text{using only } E[\mathbf{j}])}{J_{\text{eff}}(\text{using only } E[\mathbf{j}])} = \frac{\int_0^{t_0} \frac{f''(E[\mathbf{j}])}{2} V[\mathbf{j}] dt}{\int_0^{t_0} f(E[\mathbf{j}]) dt} \tag{5.1}$$

This expression depends on current *density* and not simply on current. Consequently, the expected and variance waveforms have to be augmented with metal line *width* information in order to evaluate (5.1). Table 1 shows the results for all the test cases presented above, for different values of line width, with a line thickness of $0.3\mu m$ throughout. Note that (5.1) has been tabulated as a percentage.

Table 1. Variance contribution (with relevant line width). Size refers to the number of transistors. μ stands for micro-meter.

Circuit	Size	$\Delta J_{\text{eff}}/J_{\text{eff}}$ (width)	
Nand	4	110% (.067μ)	0.00% (1.0μ)
Complex	6	185% (.033μ)	0.00% (1.0μ)
Xor	16	236% (.067μ)	0.00% (1.0μ)
Adder	54	107% (0.17μ)	0.43% (1.0μ)
Multiplier	648	219% (0.25μ)	3.70% (1.0μ)
Multiplier*	648	200% (0.25μ)	3.40% (1.0μ)

*Heuristic CREST run, all others are full accuracy CREST.

For a given circuit, the variance contribution increases with a decrease in line width. The two width values used in the table for each circuit are meant to demonstrate that the variance contribution to the MTF can vary from insignificant to very important. For the types of circuits, and technology, that we have examined, it seems that one can use a practical lower limit of 1μ on line width, and do without a variance waveform. However, in cases where such limits cannot be guaranteed (as in manual layout with unlimited designer freedom) and/or where the circuit and metal technologies involve high enough current density, the variance waveform can be as important (or more) than the expected waveform.

In any case, and regardless of the significance of its contribution to the MTF, the variance waveform is important in its own right. Its relevance for studying other reliability problems will be briefly discussed in the next section.

Finally, we illustrate the speed performance of CREST with the variance estimation built in. Table 2 shows the speed comparisons between CREST and SPICE for all the examples presented above. The speedup is excellent, and becomes much better for larger circuits (1529X for the adder and 11595X for the multiplier). In fact, the speedup grows exponentially, because an exponential number of deterministic simulation runs are replaced by a single probabilistic simulation run. We point out the case of the multiplier circuit (the largest circuit in the table) with the heuristic CREST run (last row in Table 2). The excellent waveform comparison in Fig. 9, along with the speedup of 11595X and execution time of 1 minute in Table 2, establish the feasibility of solving large VLSI chips.

6. Summary and Conclusions

We have extended the probabilistic simulation approach [1] to include the computation of the variance of the power/ground current waveform, in addition to its expected waveform. To provide the motivation for this, we have focused on the problem of estimating the median time-to-failure (MTF) due to electromigration in the power and ground busses of CMOS circuits. This requires knowledge of the current density in these busses.

In previous work [1], we presented a novel technique for MTF estimation based on a *stochastic current waveform* model. We derived the expected waveform of this current model and gave a qualitative argument suggesting that it is the appropriate waveform to be used

Table 2. Execution time comparisons. Time is in CPU seconds on a VAX-11/780; size refers to the number of transistors.

Circuit	Size	SPICE	CREST	Speedup
Nand	4	41.75	0.90	46X
Complex	6	244.15	1.09	224X
Xor	16	456.00	3.13	146X
Adder	54	32620.42	21.33	1529X
Multiplier	648	697530.88†	1871.99	373X
Multiplier*	648	697530.88†	60.16	11595X

*Heuristic CREST run, all others are full accuracy CREST.
†Estimated (2^8 times the cost of a typical SPICE run).

for MTF estimation. In this paper, we have quantified that argument by presenting new theoretical results which show the exact relationship between the MTF and the statistics of the current. Equation (2.7) relates J_{eff}, required for estimating the MTF, to the mean waveform of a nonlinear function of the stochastic current. Coupled with (2.10), it provides an efficient and more accurate technique for computing the MTF, which requires both the expected and variance waveforms. A novel technique was then presented to compute the variance waveform for CMOS circuits, which has been incorporated into the probabilistic simulator CREST [1].

The results of several CREST runs were presented, demonstrating good waveform agreement with SPICE, as well as excellent speedups over traditional approaches : a speedup of over 11500X was demonstrated on a 648-transistor circuit.

Using these results, we have studied the significance of the variance waveform by evaluating its contribution to the MTF, $\Delta J_{\text{eff}}/J_{\text{eff}}$. We have found this to be highly dependent on whether a minimum metal line width can be guaranteed, in which case, this contribution may be neglected. Otherwise, and/or if a high current density technology is used, then the variance contribution can be as important (or more) than the expected waveform.

In any case, and regardless of the significance of its contribution to the MTF, the variance waveform is important in its own right as a measure of the spread of the real current waveforms. Such information may be useful for studying voltage drop (glitches) on the

power/ground lines. If stochastic techniques are to be used for this purpose, then it seems definite that the expected current by itself will not be sufficient.

Appendix

In this appendix, we are concerned with the problem of deriving the variance waveforms in the power/ground bus branches given those at the bus contacts. The bus can be modeled as a *linear*-time-invariant (LTI) system with inputs $\mathbf{x_j}$ and outputs $\mathbf{y_i}$. The inputs $\mathbf{x_j}(t)$, $j = 1, \ldots, n$ represent the contact currents, and carry the stochastic processes $\mathbf{i_j}(t)$ of known variance waveforms $\sigma^2_{i_j}(t)$. The outputs $\mathbf{y_i}(t)$, $i = 1, \ldots, m$ represent the bus branch currents at which the variance waveforms, $\sigma^2_{y_i}(t)$, are required. Let $h_{ij}(t)$ be the impulse response function relating $\mathbf{y_i}(t)$ to $\mathbf{x_j}(t)$:

$$\mathbf{y_i}(t) = \sum_{j=1}^{n} h_{ij}(t) * \mathbf{x_j}(t), \qquad i = 1, \ldots, m \tag{A.1}$$

where "$*$" denotes the *convolution* operation.

It is well known (see [5], page 209) that the variances of the system inputs are not enough to derive the variances of its outputs. The *auto-correlation* of each input, $R_{x_j x_j}(t_1, t_2) \stackrel{\triangle}{=} E\left[\mathbf{x_j}(t_1)\mathbf{x_j}(t_2)\right]$, is also required. Since the input processes are not *wide-sense stationary* [5], an exact analytical solution can be quite complex, even if the auto-correlations were known. Therefore, as is often necessary, we will make certain simplifying assumptions about the structure of $R_{x_j x_j}$.

We assume that the correlation between $\mathbf{x_j}(t)$ and $\mathbf{x_j}(t+\tau)$ goes to zero as $\tau \to \infty$. In terms of the *auto-covariance*, $C_{x_j x_j}(t_1, t_2) \stackrel{\triangle}{=} R_{x_j x_j}(t_1, t_2) - \eta_{x_j}(t_1)\eta_{x_j}(t_2)$, where $\eta_{x_j}(t) \stackrel{\triangle}{=} E[\mathbf{x_j}(t)]$, this is formulated as :

$$C_{x_j x_j}(t_1, t_1) = \sigma^2_{x_j}(t_1), \text{ and } C_{x_j x_j}(t_1, t_2) = 0 \text{ for } |t_1 - t_2| \geq T, \tag{A.2}$$

where T is a (typically small) time interval.

Consider the *discrete time* system obtained by sampling, with period T, the continuous time system defined by (A.1). If $\mathbf{x_j}[k] \stackrel{\triangle}{=} \mathbf{x_j}(kT)$ are the discrete processes at the inputs, and $\mathbf{y_i}[k] \stackrel{\triangle}{=} \mathbf{y_i}(kT)$ are the discrete output processes, then :

$$\mathbf{y_i}[k] = \sum_{j=1}^{n} h^{(d)}_{ij}[k] * \mathbf{x_j}[k], \qquad i = 1, \ldots, m \tag{A.3}$$

where $h_{ij}^{(d)}[k]$ is the discrete impulse response function relating $\mathbf{y_i}[k]$ to $\mathbf{x_j}[k]$. As shown below, the discretized output variance waveforms can be derived irrespective of the shape of $C_{x_j x_j}(t_1, t_2)$ for $|t_1 - t_2| < T$. The continuous variance waveforms can then be obtained by interpolation. Strictly speaking, therefore, the sampling period T should be small : $1/T$ should be larger than the largest frequency component of the inputs. However, since fine waveform details are not of paramount importance in this work, we need only restrict T to be small enough so that waveform features in that small an interval are inconsequential.

To simplify the notation, define $\mathbf{y_{ij}}[k] \stackrel{\triangle}{=} h_{ij}^{(d)}[k] * \mathbf{x_j}[k]$. Furthermore, as pointed out above, we will neglect the correlation between the contact currents. Hence the $\mathbf{x_j}$ inputs are uncorrelated, and :

$$\sigma_{y_i}^2[k] = \sum_{j=1}^{n} \sigma_{y_{ij}}^2[k], \qquad i = 1, \ldots, m. \tag{A.4}$$

We have thus reduced the problem to analyzing a single-input single-output discrete LTI system :

$$\mathbf{y_{ij}}[k] = h_{ij}^{(d)}[k] * \mathbf{x_j}[k] = \sum_{\kappa=0}^{\infty} h_{ij}^{(d)}[\kappa] \mathbf{x_j}[k - \kappa]. \tag{A.5}$$

Let $\tilde{\mathbf{x}}_\mathbf{j}[k] \stackrel{\triangle}{=} \mathbf{x_j}[k] - \eta_{x_j}[k]$ and $\tilde{\mathbf{y}}_\mathbf{ij}[k] \stackrel{\triangle}{=} \mathbf{y_{ij}}[k] - \eta_{y_{ij}}[k]$. Then $\sigma_{y_{ij}}^2[k] = E[\tilde{\mathbf{y}}_\mathbf{ij}[k]^2]$ and $\tilde{\mathbf{y}}_\mathbf{ij}[k] = h_{ij}^{(d)}[k] * \tilde{\mathbf{x}}_\mathbf{j}[k]$, hence :

$$\sigma_{y_{ij}}^2[k] = E\left[\Big(\sum_{\kappa=0}^{\infty} h_{ij}^{(d)}[\kappa] \tilde{\mathbf{x}}_\mathbf{j}[k - \kappa]\Big)^2\right] = \sum_{\kappa_1=0}^{\infty} h_{ij}^{(d)}[\kappa_1] \sum_{\kappa_2=0}^{\infty} h_{ij}^{(d)}[\kappa_2] E\Big[\tilde{\mathbf{x}}_\mathbf{j}[k - \kappa_1] \tilde{\mathbf{x}}_\mathbf{j}[k - \kappa_2]\Big]. \tag{A.6}$$

Furthermore, it is easy to see that $E\big[\tilde{\mathbf{x}}_\mathbf{j}[k_1] \tilde{\mathbf{x}}_\mathbf{j}[k_2]\big] = C_{x_j x_j}(k_1, k_2)$, which, using (A.2), gives :

$$\sigma_{y_{ij}}^2[k] = \sum_{\kappa=0}^{\infty} \Big|h_{ij}^{(d)}[\kappa]\Big|^2 \sigma_{x_j}^2[k - \kappa] = \Big|h_{ij}^{(d)}[k]\Big|^2 * \sigma_{x_j}^2[k]. \tag{A.7}$$

And, finally, the variance waveforms for the system outputs are, using (A.4) :

$$\sigma_{y_i}^2[k] = \sum_{j=1}^{n} \Big|h_{ij}^{(d)}[k]\Big|^2 * \sigma_{x_j}^2[k], \qquad i = 1, \ldots, m. \tag{A.8}$$

In other words, *the variances of the system outputs (bus branch currents) can be obtained from the convolution of the variances of its inputs (contact currents) with the squares of its discrete impulse response functions.* This discrete convolution can be easily performed once the discrete impulse response functions are found. Of course the summation need not be taken to infinity, and may be conveniently truncated after $\left|h_{ij}^{(d)}[\kappa]\right|$ is less than some small value. To obtain the discrete impulse response functions, note that if a unit-step input current is applied at contact j, with all other contact currents held at zero, and if the resulting outputs $y_i(t)$ are monitored, then :

$$h_{ij}^{(d)}[k] = y_i(kT) - y_i\big((k-1)T\big) = \int_{(k-1)T}^{kT} h_{ij}(\tau)d\tau, \qquad i = 1, \ldots, m. \tag{A.9}$$

This suggests two methods for deriving $h_{ij}^{(d)}[k]$. The first uses a simulation program such as SPICE to simulate the bus with unit-step input currents applied at each contact (one at a time), while monitoring the bus branch currents. This gives the mn functions $h_{ij}^{(d)}[k]$ using (A.9). Another (approximate) method would be to make use of the second equality in (A.9) : if the continuous impulse response functions are approximated using some RC time-constant analysis of the bus, then the discrete impulse response functions can be obtained from them.

For very large chips, it may be prohibitively expensive to perform the required convolutions. One can simplify the calculations by making an additional assumption as follows. If the bus is known to be "fast", i.e., if $h_{ij}^{(d)}[k]$ dies down faster than changes in $\sigma_{x_j}^2[k]$, then (A.7) reduces to :

$$\sigma_{y_{ij}}^2[k] \approx \sigma_{x_j}^2[k] \sum_{\kappa=0}^{\infty} \left|h_{ij}^{(d)}[\kappa]\right|^2. \tag{A.10}$$

So the convolutions in (A.8) can be replaced by simple multiplications, and the constants $\sum_{\kappa=0}^{\infty} \left|h_{ij}^{(d)}[\kappa]\right|^2$ can be derived in a pre-processing step from the impulse response functions and stored in a single $m \times n$ constant matrix.

If the chip is too big to even derive $h_{ij}^{(d)}[k]$, then one further simplification can be made as follows. If $h_{ij}^{(d)}[k]$ dies down faster than changes in $\mathbf{x_j}[k]$ then (A.5) reduces to $\mathbf{y_{ij}}[k] = \mathbf{x_j}[k] \sum_{\kappa=0}^{\infty} h_{ij}^{(d)}[\kappa]$, and so :

$$\sigma^2_{y_{ij}}[k] \approx \sigma^2_{x_j}[k]\Big(\sum_{\kappa=0}^{\infty} h^{(d)}_{ij}[\kappa]\Big)^2. \tag{A.11}$$

The constants $\left(\sum_{\kappa=0}^{\infty} h^{(d)}_{ij}[\kappa]\right)^2$ can be very easily obtained as follows. Notice that $\sum_{\kappa=0}^{\infty} h^{(d)}_{ij}[\kappa]$ is the *steady state* current in branch i in response to a unit-step input current at contact j, with all other contact currents held at zero. If the bus is modeled as a resistive network, then the steady state node voltages in response to such inputs are the entries of the driving point impedance matrix. So if the node-admittance matrix is built by simple inspection of the bus and then inverted to produce the driving point impedance matrix, the steady state currents are immediately available.

References

[1] F. Najm, R. Burch, P. Yang, and I. Hajj, "Probabilistic simulation for reliability analysis of CMOS VLSI circuits," *IEEE Transactions on Computer-Aided Design*, vol. 9, no. 4, pp. 439–450, April 1990.

[2] F. M. d'Heurle, "Electromigration and failure in electronics : an introduction," *Proceedings of the IEEE*, vol. 59, no. 10, pp. 1409–1418, October 1971.

[3] J. R. Black, "Electromigration failure modes in aluminum metallization for semiconductor devices," *Proceedings of the IEEE*, vol. 57, no. 9, pp. 1587–1594, September 1969.

[4] F. Najm, "Probabilistic simulation for reliability analysis of VLSI circuits," Ph.D. thesis, Dept. of Electrical and Computer Engineering (also available as report UILU-ENG-89-2221, DAC-14), University of Illinois at Urbana-Champaign, June 1989.

[5] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 2nd Edition. New York, NY: McGraw-Hill Book Co., 1984.

[6] F. Najm, I. Hajj, and P. Yang, "Electromigration median time-to-failure based on a stochastic current waveform," *IEEE International Conference on Computer-Design*, Cambridge, MA, pp. 447–450, October 2–4, 1989.

[7] F. Najm, I. Hajj, and P. Yang, "Computation of bus current variance for reliability estimation of VLSI circuits," *IEEE International Conference on Computer-Aided Design*, Santa Clara, CA, pp. 202–205, November 6–9, 1989.

[8] J. D. Venables and R. G. Lye, "A statistical model for electromigration induced failure in thin film conductors," *Proc.* 10^{th} *Annual IEEE Reliability Physics Symposium*, Las Vegas, NV, pp. 159–164, April 5–7, 1972.

[9] J. W. McPherson and P. B. Ghate, "A methodology for the calculation of continuous DC electromigration equivalents from transient current waveforms," *The Electrochemical Society, Proc. Symp. on Electromigration of Metals*, New Orleans, LA, pp. 64–74, Oct. 7–12, 1984.

[10] J. E. Hall, D. E. Hocevar, P. Yang, and M. J. McGraw, "SPIDER - a CAD system for modeling VLSI metallization patterns," *IEEE Transactions on Computer-Aided Design*, vol. CAD-6, pp. 1023–1031, Nov. 1987.

Figure Captions

Figure 1: The dependence of MTF on current density, reproduced for convenience from [8]. The dashed lines show the results of the approximation $t_{50} \propto j^{-n}$ for $n = 1,\ 3/2,$ and 2.

Figure 2: A plot of $f(j)$, obtained from Fig. 1 by inverting and appropriately scaling the ordinate axis.

Figure 3: A (non-stochastic) current waveform, $j(t)$, built as a sequence of the waveforms $j_k(t)$, each occurring with its assigned probability P_k.

Figure 4: A generic CMOS gate structure.

Figure 5: CREST variance pulse result for a 2-input CMOS nand gate, compared to SPICE.

Figure 6: Variance results for a 3-input CMOS complex gate (inset).

Figure 7: Variance results for a 16-MOSFET exclusive-or (xor) CMOS circuit.

Figure 8: Variance results for a 54-MOSFET 2-bit ripple adder CMOS circuit.

Figure 9: Variance results for a 648-MOSFET 4-bit parallel multiplier CMOS circuit.