Thermally-induced soft errors in nanoscale CMOS circuits

H. Li, J. Mundy, W. Patterson, D. Kazazis, A. Zaslavsky and R.I. Bahar

Division of Engineering, Brown University, Providence, RI 02912

Abstract—Electrical noise will play an increasingly critical role in future nanoscale CMOS circuit operation characterized by lower supply voltages V_{DD} and smaller device sizes. Both of these downscaling approaches reduce the margin of immunity to thermal noise, alpha particle strikes, and threshold voltage variations. This paper investigates the noise probability distributions for both equilibrium and non-equilibrium logic states of advanced CMOS flip-flops operated at ultra-low V_{DD}. The theoretical distribution of transition times from one stable operation point to the other stable operation point is also derived, which is a useful representation of the soft error rate. It is shown that such nanoscale flip-flop designs are extremely sensitive to threshold variations, reducing average failure time to a few days. Monte Carlo simulations are provided to validate the theoretical model and its predictions.

Index Terms—noise, flip-flop, soft errors, birth-death queue

I. Introduction

A. Motivation

While noise margins have long been a reliability concern for digital circuits (especially for those designed using dynamic logic) the nature of the noise has been dominated by such factors as supply voltage fluctuations and capacitive cross-coupling. Although these sources of noise will certainly still exist in future circuits, as CMOS devices continue to scale down in both physical dimension and operating voltage V_{DD}, other noise sources – such as thermal noise, electromagnetic coupling, and hot electron effects coupled with a much wider process-related spread of gate threshold voltages – will further reduce the operational noise margins. This, in turn, will expose computation to higher soft-error rates.

The inevitability of reduced noise margins is generally accepted by device and circuit designers, but without an accurate means of modeling various sources of noise, the exact extent to which this may affect reliability is unknown. While the supply voltage V_{DD} scaling is considerably slower than transistor size scaling in order to maintain sufficient current drive, V_{DD} must still be reduced in order to limit overall power dissipation. For example, by 2015, CMOS device dimensions are targeted to scale down to gate lengths $L_G \sim 10$ nm, whereas the V_{DD} of low-power CMOS is only predicted to fall to 0.5 V by 2016 then stay at that level for the next several years [1]. Although extrapolations to an even lower $V_{DD} \sim 0.3$ V have also been published [2], the trend is still to limit supply voltage and threshold voltage scaling relative to L_G scaling as a means of maintaining current drive and reducing static leakage current, respectively. At these relatively high V_{DD} values, thermal noise is not expected to pose a problem even in ultimate CMOS.

However, if power reduction takes precedence over performance for a specific application, reducing V_{DD} further, into the subthreshold regime, may prove very useful. Indeed, a number of researchers have published works focused on ultra-low voltage digital circuits, mainly as a means of minimizing overall energy consumption (e.g., [12][13][14]). In this ultra-low V_{DD} scenario, noise margins will be necessarily small and soft error rate analysis due to all sources of noise becomes crucial, as designers attempt to balance the conflicting constraints of performance, power dissipation, and reliability. While tools for estimating power and performance are relatively mature, this is not true for reliability; at this point engineers do not have an analytical means of measuring reliability for a given V_{DD}, and hence tend to rely on empirical guesswork.

This paper addresses the need for tools that more accurately analyze soft error rates, including predicting unacceptable soft error rates for given device dimensions, logic design style, and V_{DD}. As a result, this analysis tool may allow for more efficiently optimized designs. For instance, it may allow a designer to choose different V_{DD} supplies for various parts of a chip in order to better balance performance, power, and reliability constraints.

While sources of noise in ultimate CMOS are many, this paper focuses on developing an analytical model specifically for thermal noise. Thermal noise is relatively unexplored in the digital CMOS context, because in state-of-the-art CMOS devices the number of electrons flowing in MOS transistors is sufficiently large to make random current fluctuations negligible. However, as V_{DD} and L_G are scaled down, current fluctuations corresponding to a few electrons will become more significant, thereby increasing the likelihood of soft errors. As a result, ultimate CMOS devices operated at ultra-low V_{DD} may suffer from unacceptable error rates. Similarly, fewer electron-hole pairs generated by alpha particle strikes may now lead to soft errors than would have disturbed earlier circuits.

Noise has been traditionally modeled using a Gaussian model (e.g., [3] added Gaussian noise of zero mean and 60mV RMS standard deviation to the output of each gate). While this model is based on noise predictions for ultimate CMOS [4], it is not based on any detailed analytical modeling of noise. In particular, if it were possible to analyze the charge on a storage node of a memory cell in terms of the probability of thermal fluctuations in the number of stored electrons, circuit designers would know whether random current fluctuations could flip the stored logic value. Preliminary analysis in developing an analytical noise model was done in [5][6], where the authors

978-1-4244-1791-9/08/$25.00 ©

used queuing theory to analyze thermal noise in a single transistor or inverter. However, these results did not provide enough information to understand the effects of thermal noise on storage cells or to quantitatively predict the mean time to storage errors. In this paper, the work of [5] is expanded to flip-flops and it is shown that thermal noise alone can be responsible for bit flips in a single flip-flop on the order of once every few days for aggressively scaled CMOS at ultra-low V_{DD}. This result is quite significant, since it puts reliability of digital circuits into question unless new strategies for reliable computing are employed.

Previous work on soft error analysis has mostly been confined to two frameworks. The first approach is the analysis of soft errors through simulation of the effect of charge bursts on circuit models, for example the work of V. Degalahal et al [7]. The second approach is to propagate discrete probabilities through logic networks using transition matrices or other propagation methods, for example the work by S. Krishnaswamy et al [8]. The work presented in this paper is a genuinely novel framework where time-varying logic signal noise probability distributions are represented analytically. This analytic framework allows analysis of error probability over vast intervals of time not achievable through simulation and with the ability to capture dynamic circuit behavior not achievable through discrete probability propagation.

B. Proposed Approach

Investigation of thermal noise phenomena in CMOS digital circuits has received little attention compared to analog circuit noise theory. The classical approach to noise analysis is based on computing the independent mean square current or voltage fluctuation in each device within the circuit bandwidth [9]. The underlying assumptions are that the noise statistics are stationary and the noise amplitude is sufficiently low that the circuit behaves linearly. Under those assumptions, the noise energies are added at the output of the circuit to find a gross mean square noise voltage. That limited information does not predict the time it would take to disturb a flip-flop state. More, a flip-flop that has sufficient noise fluctuation to change state does not satisfy either classical assumption, because the noise current distributions change instantaneously as the storage node potentials change and so do the noise gains.

A key background reference for the work described here is the paper by Sarpeshkar et al. [10] on Poisson-noise model in subthreshold MOS transistors. They show that in both the linear and saturation regimes the drain current noise fluctuation is equivalent to the difference between a forward and a reverse drain current each of which is characterized by pure shot noise. Thus, shot noise provides a unifying model for the main sources of random thermal effects. The model introduced in [10] for the distribution of shot noise is the Poisson process with independent distributions for the arrival and departure of carriers at a drain or source node.

This model has been extended by the application of queuing theory to represent the storage of charge in the gate or other parasitic capacitances [5]. The capacitance of a node is modeled as a queue where the CMOS drain current fluctuations are considered to be random arrival and departure events. In equilibrium, the mean queue population (capacitor charge) is constant, with equal arrival and departure rates. A queue with both arrival and departure events is called a birth-death queue and is modeled as a Markov chain [11].

The queuing analysis in [5] showed that CMOS logic circuits operating at subthreshold supply voltage are prone to errors. Theoretical results for the variance in circuit voltages due to thermal noise were compared to Monte Carlo simulations with good agreement. However, in [5] the translation of the voltage variance into actual digital error rates was qualitative.

In this paper, the theory is extended to a formal analysis of the time to a logic error due to thermal fluctuations, and the performance of a 65nm transistor is shown as an example that the thermal noise can be a major threat to the future nanoscale circuits. In current transistor technology, with $V_{DD} > 1$ V, thermally induced errors are not an issue, with the typical time to an error exceeding the age of the universe. However, as stated earlier, operating circuits at ultra-low (sub-threshold) V_{DD} can have clear energy advantages, especially for applications that do not require high performance, but at the cost of reduced noise margins. In our analysis, we will show that in projected ultimate CMOS devices at ultra-low V_{DD}, (i.e., $V_{DD} = 0.2$V), even minor variations in threshold voltage can reduce the mean time to a logic error to seconds.

Before proceeding, it is worth emphasizing that it is impossible to compute these error rates using conventional circuit simulation in SPICE [21]. Classical noise analysis in SPICE is subject to the limitations discussed above. Transient simulation requires external driving functions for the "noise" and it is difficult to couple those to the instantaneous circuit conditions in any realistic way. Moreover failure times are expected to be of the order of seconds from electron transition times on the order of picoseconds. Even if it were straightforward to set up the simulation, it is likely that the run time to expose exponentially rare logic errors caused by thermal fluctuations would be too long to be practical. Thus, the theory described here provides a significant new capability for accurate error prediction.

II. The Circuit Model

A. The flip-flop circuit

The theoretical approach is illustrated using a simple bistable flip-flop circuit, consisting of two cross-coupled inverters, as shown in Figure 1. The goal of this paper is to provide a theoretical framework for computing the mean and variance of the time for a flip-flop to change state due to thermal noise fluctuations. These results provide insight into CMOS soft errors in general circuit design.

Two inverters are connected with positive feedback, so the flip-flop resides in either one of two stable operation points: {V_{in} = logic "0", V_{out} = logic "1"} or vice versa. The state of the flip-flop is fully characterized by the amount of charge on the two load capacitors C_L. Every noise-induced variation causes the operation point to stray from the stable operating point, but large fluctuations are exponentially rare. If a fluctuation is large enough, the flip-flop input and output voltages will pass a critical threshold and the state will change spontaneously.

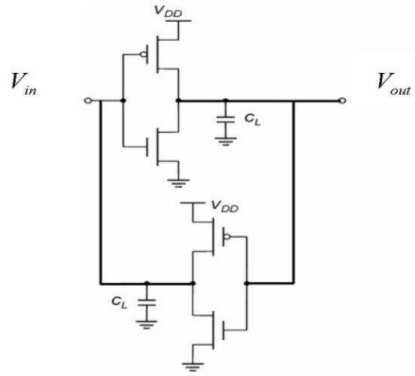


Figure 1 The flip flop circuit.

B. The CMOS transistor model

In order to ground the analysis in a practical set of device parameters, a specific advanced bulk 65 nm CMOS technology from Intel [15] is used to model the effects of ultra-low supply voltage on the sub-threshold noise characteristics of a CMOS memory circuit operated at $V_{DD} = 0.2$ V. We note that for this device size thermal noise is not expected to dominate, but as CMOS is scaled towards the ultimate limit thermal noise will become more significant. The DC transistor characteristics of both NMOS and PMOS transistors with 35 nm gate lengths are shown in Fig. 2(a). In the subthreshold regime, the drain current I_D can be described by the usual diffusion-dominated expression, including the drain-induced barrier lowering (DIBL) shift:

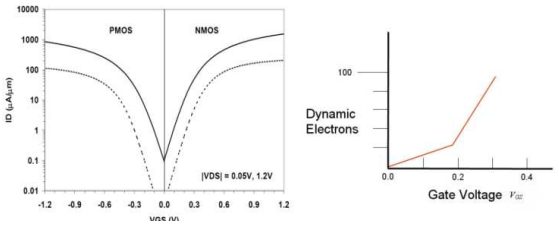
$$I_D = I_0 e^{\frac{q(V_{GS}-\Delta V_T)}{nkT}}\left(1-e^{\frac{-qV_{DS}}{kT}}\right), \qquad (1)$$

where $\Delta V_T < 0$ is the V_{DS}-dependent decrease in threshold voltage due to the DIBL. For an NMOS device width of 180 nm (assuming 5:1 width to gate length sizing), I_D is on the order of one electron per picosecond at $V_{GS} = 0.2$ V. The slope factor n is extracted by matching this equation to the reported device characteristics [15], and a good fit to the subthreshold characteristics $V_{GS} < V_T = 0.28$ V in Figure 2(a) is obtained with $I_0 = 0.006$ μA/μm and $n = 1.6$ (we have chosen kT to correspond to 100 °C). The effective capacitance of a gate load can be represented by the incremental change in the total gate charge with gate voltage V_{GS}, corresponding to the sum of changes in the substrate depletion charge beneath that gate and in the inversion layer charge. The model for this variation expressed as a number of electrons as a function of V_{GS} is shown in Figure 2(b), where there are about 28 additional electron-charges in the NMOS device when $V_{GS} = 0.2$ V. Although the weaker subthreshold current drive of PMOS in fabricated devices has often led to greater PMOS to NMOS gate width ratios [16], here we assume the standard 2:1 ratio, resulting in the storage of 28×2=56 electrons in the PMOS device. Ignoring parasitic capacitance, the capacitive load of an inverter input corresponds to a population of 56+28 = 84 electrons in logic state "1", when the inverter is operating at $V_{DD} = 0.2$ V.

III. The Noise Model

A. The Poisson distribution

The Poisson distribution is designed to portray the random arrival of elements, such as customers arriving at a service desk queue. The basic assumption is that the probability of an event in an interval dt is independent of time and is given by λdt where λ is a rate constant. A CMOS transistor has two independent Poisson processes that represent charge and discharge drain currents. For example, in an NMOS transistor these forward and reverse currents have rate constants in electrons per picosecond $\mu = \frac{I_0 \cdot 10^{-12}}{q} e^{\frac{q(V_{GS}-\Delta V_T)}{nkT}}$ and $\lambda = \frac{-I_0 \cdot 10^{-12}}{q} e^{\frac{q(V_{GS}-\Delta V_T)}{nkT}} e^{\frac{-qV_{DS}}{kT}}$ that correspond to the two terms in equation (1). In equilibrium, the sum of the Poisson rates at each gate node is zero. When the gate and drain voltages are expressed in terms of the number of charges, k_i, on a given capacitor C_i, then the system can be represented as shown in Figure 3.



a) b)

Figure 2 The CMOS transistor model: a) drain current I_D as a function of V_{GS} at constant V_{DS} [15]; b) the number of electrons under the gate of a single NMOS transistor in weak inversion vs. V_{GS} at low V_{DS}.

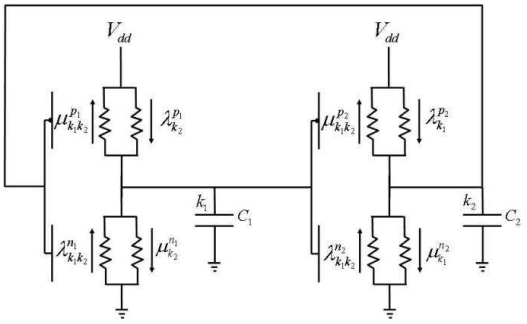


Figure 3 The Poisson noise model for the flip-flop

First, consider the current components in one inverter. The rates depend on the number of charges on both capacitors defined as k_i, $i = 1, 2$. The voltage on a load capacitor is $V_i = qk_i/C_i$. Define m_i as the number of charges on the capacitor when the voltage $V_C = V_{DD}$, thus $m_i = V_{DD}C_i/q$. For the PMOS transistor in stage 2,

$$\lambda_{k_1}^{p_2} = N_0 e^{\frac{q(V_{DD}-V_{C_1})}{nkT}} = N_0 e^{\frac{U_0}{n}(\frac{m_1-k_1}{m_1})} \quad (2)$$

$$\mu_{k_1k_2}^{p_2} = N_0 e^{\frac{q(V_{DD}-V_{C_1})}{nkT}} e^{\frac{-q(V_{DD}-V_{C_2})}{kT}} = N_0 e^{\frac{U_0}{n}(\frac{m_1-k_1}{m_1})} e^{-U_0(\frac{m_2-k_2}{m_2})} \quad (3)$$

where $U_0=qV_{DD}/kT$ is the ratio of logic energy to thermal energy, and $N_0=I_0/q$ is a constant. For the NMOS transistor in stage 2,

$$\lambda_{k_1k_2}^{n_2} = N_0 e^{\frac{U_0}{n}(\frac{k_1}{m_1})} e^{-U_0(\frac{k_2}{m_2})} \quad (4)$$

$$\mu_{k_1}^{n_2} = N_0 e^{\frac{U_0}{n}(\frac{k_1}{m_1})}. \quad (5)$$

The rate expressions are similar for stage 1. The difference of charge and discharge rates as a function of the number of carriers, k, at one node of a symmetric flip-flop is shown in Figure 4. Note that the flip-flop exhibits a balance in the charge and discharge rates at k = 1 and 83 electrons (corresponding to stable logic states for our chosen V_{DD} = 0.2 V using the transistors of Fig. 2), as well as at k = 42. The latter balance point is in the middle of the transition band and is the usual metastable point.

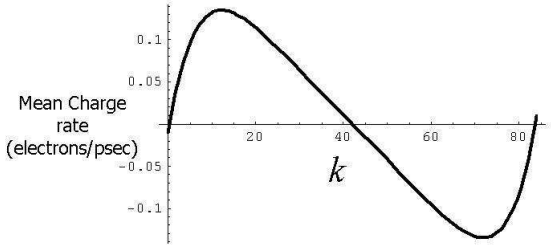


Figure 4 The existence of three equilibrium states of charge/discharge balance.

B. The one-dimensional (1D) birth-death queue model

The birth-death process is a special case of a continuous-time Markov process where the states represent the current size of a population and where the transitions are limited to births and deaths [17]. When a birth occurs, the process goes from state i to i+1, and when a death occurs, the process goes from state i to state i-1. The process is specified by the birth and death rates.

Figure 5 shows the one-hop concept of a birth-death process, that is, a process that can only jump to adjacent states in one step, but after a long time, the process can reach a remote state through a series of strictly local state moves.

C. The concept of first passage time

Given any two states of a system, the first passage time between them, T_{ij}, is the time from when the system is known to be in the first state, i, until it enters the second state, j, for the first time. Passage time is a random variable with some mean and variance for each possible state pair. In the case of a flip-flop, if the first state is one of the stable equilibrium states and the second state is the other stable state, then the mean of the first passage time, $\overline{T_{0N}}$, is the mean time to a bit flip error. A flip-flop or any other system that makes state transitions through simple Poisson processes is sufficiently well behaved that we can define a probability density function, $p_{ij}(t)$, for it such that if the circuit is in state i at t = 0 then the probability that it enters state j between t and t + dt is $p_{ij}(t)dt$.

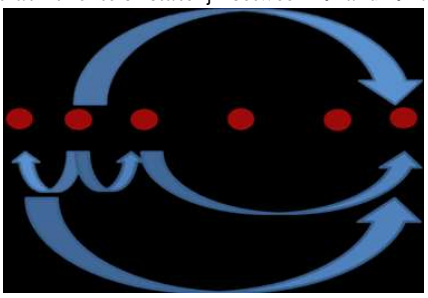


Figure 5 The state transition diagram

If $p_{ij}(t)$ were known then the mean passage time is its first moment and the variance or standard deviation can be computed from the second moment. The purpose of this paper is to estimate this probability density function.

D. Properties of Poisson processes applied to a birth-death queue

For simplicity, consider a 1D birth-death queue like that shown in Figure 6. Transitions between states are by Poisson distributed processes representing the charge or discharge of a nodal capacitance. In Figure 6, the state i is associated with a charging rate λ_i and with a discharging rate μ_i. If the system is known to be in state i at t = 0, then the probability that it is still in state i some time later is $p_i(t)=\exp(-(\lambda_i+\mu_i)t)$. The probability density for transitions out of this state into either i+1 or i-1 is $p_i(t)=(\lambda_i+\mu_i)\exp(-(\lambda_i+\mu_i)t)$. The sojourn or residence time for the state is the mean duration of the state given by the first moment of this transition probability density

$$\tau_i = \int_0^\infty t\, p_0(t)dt = (\lambda_i+\mu_i)\int_0^\infty t e^{-(\lambda_i+\mu_i)t}dt = \frac{1}{(\lambda_i+\mu_i)}$$

Finally, let us define a rate connecting two states as

$$r_{ik} = \begin{cases} \lambda_i & \text{if } k = i+1 \\ \mu_i & \text{if } k = i-1 \\ 0 & \text{if } |k-i| > 1 \end{cases}$$

With this definition the probability density for transitions from state i to state k is $r_{ik}p_i(t)/(\lambda_i+\mu_i)$. The factor $r_{ik}/(\lambda_i+\mu_i)$ can be regarded as the probability that when the system leaves state i, it does so to state k. The extension of these expressions to two-dimensional (2D) queues only requires the use of four rather than two rate constants for each state since four rather than two states are accessible in single jumps.

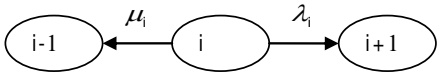


Figure 6 Three adjacent states in a one-dimensional birth-death queue.

E. The Laplace transform representation

The Laplace transform is often used to enable an algebraic representation of the integral-differential equations that arise in time varying circuit analysis [18]. In this paper, Laplace analysis will provide a closed form approximation to the flip-flop passage-time probability density function. The following analysis is similar to that used in the analysis of dislocations in metals [19].

Suppose that $p_{kq}(t)$, the probability density for the transition from state k to q, is known for all states k and q. Consider the special case that when the queue leaves state i it enters state k at some time t_k. The probability density for the time of reaching state q via state k, is given by,

$$p_{\substack{iq \\ i\to k\to q}}(t) = \frac{r_{ik}}{\lambda_i+\mu_i}\int_0^t p_i(t_k)\,p_{kq}(t-t_k)\,dt_k .$$

This integral is a convolution and its upper limit may be extended to infinity because $p_{ik}(t)$ vanishes for $t < 0$. The Laplace transform converts such a convolution into a simple product, so it follows that

$$L\left(p_{\substack{iq \\ i\to k\to q}}(t)\right) = \frac{r_{ik}}{\lambda_i+\mu_i}L(p_i(t))L(p_{kq}(t)).$$

The probability density function of the sojourn time is $p_i(t)=(\lambda_i+\mu_i)\exp(-(\lambda_i+\mu_i)t)$, so the Laplace transformation of the sojourn time in the state i is:

$$L(p_i(t)) = \int_0^\infty p_i(t)e^{-st}dt = \frac{(\lambda_i+\mu_i)}{s+\lambda_i+\mu_i}.$$

To reach state q, the system must first go through some particular first state, so the probability density to reach q must be the sum of all possible densities through the different possible first steps (See Figure 5). Thus the Laplace transform of $p_{iq}(t)$ is given by,

$$L(p_{iq}(t)) = \sum_{k\neq q}\frac{r_{ik}}{\lambda_i+\mu_i}\frac{(\lambda_i+\mu_i)}{s+\lambda_i+\mu_i}L(p_{kq}(t)) + \frac{r_{iq}}{\lambda_i+\mu_i}\frac{(\lambda_i+\mu_i)}{s+\lambda_i+\mu_i}$$

The common factor $(\lambda_i+\mu_i)$ can be eliminated and the final expression is,

$$L(p_{iq}(t)) = \sum_{k\neq q}\frac{r_{ik}}{s+\lambda_i+\mu_i}L(p_{kq}(t)) + \frac{r_{iq}}{s+\lambda_i+\mu_i}$$

All the above analysis is for the 1D birth-death queue, but can be applied to the 2D birth-death queue in the same way, albeit with four rates – while a 1D queue can model a single inverter [5], a 2D queue will be needed for a flip-flop.

F. A simple example

To illustrate the computation of passage time for a 2D birth-death queue, consider the simple symmetric two-electron flip-flop queue in Figure 7. It represents a circuit with two nodal capacitances that only hold one electron each at full V_{DD}. The following system of equations arises in analyzing the passage time from state 0 to state 3.

$$L(p_{03}(t)) = \frac{\lambda_0}{s+\lambda_0+\mu_0}L(p_{13}(t)) + \frac{\lambda_0}{s+\lambda_0+\mu_0}L(p_{23}(t))$$

$$L(p_{13}(t)) = \frac{\mu_1}{s+\lambda_1+\mu_1}L(p_{03}(t)) + \frac{\lambda_1}{s+\lambda_1+\mu_1}$$

$$L(p_{23}(t)) = \frac{\mu_2}{s+\lambda_2+\mu_2}L(p_{03}(t)) + \frac{\mu_2}{s+\lambda_2+\mu_2}$$

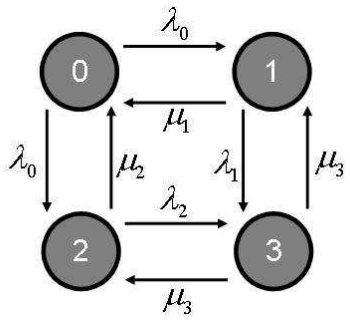


Figure 7 A four-state 2D birth-death queue.

Rearranging in matrix form,

$$\begin{bmatrix} s+\lambda_0+\mu_0 & -\lambda_0 & -\lambda_0 \\ -\mu_1 & s+\lambda_1+\mu_1 & 0 \\ -\mu_2 & 0 & s+\lambda_2+\mu_2 \end{bmatrix}\begin{bmatrix} L(p_{03}(t)) \\ L(p_{13}(t)) \\ L(p_{23}(t)) \end{bmatrix} = \begin{bmatrix} 0 \\ \lambda_1 \\ \lambda_2 \end{bmatrix}$$

and $L(p_{03}(t)) = L_{03}(s)$ can be found using linear algebra. We used equations (2) to (4) with V_{DD} = 0.2 V to get the rates as $\lambda_0=\mu_0=0.1622$ and $\lambda_1=\mu_1=\lambda_2=\mu_2=0.1039$. The plot of $L_{03}(s)$ as a function of log(1/s) is shown in Fig. 8. The probability density function $p_{03}(t)$, which is obtained from the inverse Laplace transform, is shown in Fig. 9.

The mean value and the variance are easy to derive from the probability density function and are 15.7873 and 199.7387, respectively. Note that the distribution is very much skewed, indicating a large variance in transition times. The Monte Carlo simulation is also shown in Figure 9 with the density function overlaid, showing good agreement for the two methods of analysis. For 1000 samples, the mean of the Monte Carlo simulation is 15.8769 and the variance is 208.2768.

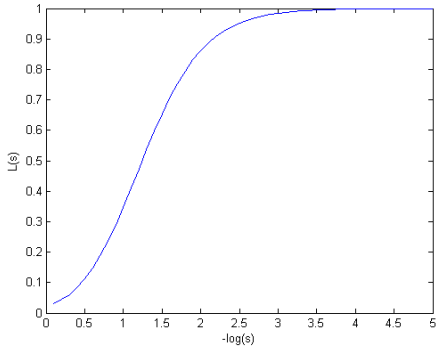


Figure 8 The plot of $L_{03}(s)$ vs. log(1/s).

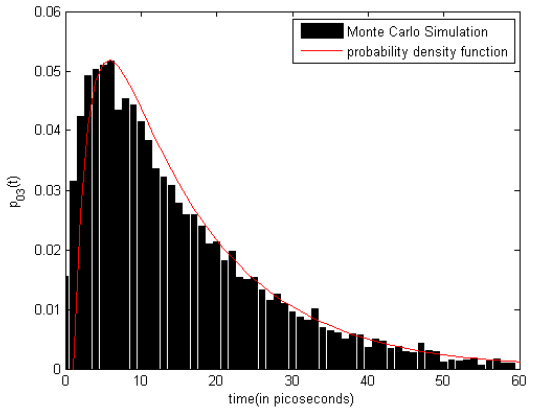


Figure 9 Monte Carlo Simulation overlaid with the passage time probability density function $p_{03}(t)$.

G. Approximation using a 1D queue

The linear system of equations arising from the 2D queue analysis becomes very ill-conditioned when the size of the queue grows past about ten electrons. A 14×14 flip-flop queue, where we artificially assume that there are only 14 electrons on the load capacitor of each inverter, is the largest that we can solve in two dimensions numerically with good accuracy. However, a lower bound on the passage time for any size system can be found by a 1D queue model that represents a diagonal path through the 2D queue as shown in Figure 10. This approximation ignores less probable paths through the 2D queue deviating from the 1D diagonal shown.

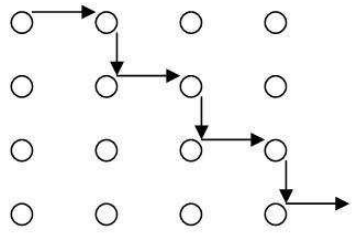


Figure 10 A typical 1D path in the 2D queue.

The 1D path approximation can be solved symbolically and takes the form,

$$L_{ij}(s) = \frac{1}{1 + a_1 s + a_2 s^2 + \ldots + a_m s^m}$$

The terms a_i are functions of the charge and discharge rates for states along the 1D path. Only a few terms in the denominator need to be evaluated to give a good approximation to $L_{ij}(s)$ since higher order terms vanish rapidly with decreasing s (increasing time). One can also show that the first moment of the inverse function, which is the mean passage time, is given exactly by $\overline{T_{ij}} = a_1$ and its variance is $\mathrm{Var}[T_{ij}] = a_1^2 - 2a_2$.

Figure 11 demonstrates the performance of the 1D approximation by comparing its results to the 2D numerical solution using the 14x14 case. The results of the 1D approximation are $\overline{T_{ij}} = 2.18 \cdot 10^7$ picoseconds, $a_2 = 3.7149 \times 10^9$, and $\mathrm{Var}[T_{ij}] = 4.75 \times 10^{14}$ picosecond^2. From the lateral separation of the two Laplace transform curves, the mean passage time from the full 2-D solution is only 30 % larger than the 1D approximation, a result that strongly suggests the symbolic 1D queue model is a close approximation.

IV. Soft Error Rates

A. The first-passage time of the symmetric flip-flop

Returning to the flip flop circuit model using transistor characteristics from Figure 2 the prediction for $L_{ij}(s)$ is shown in Figure 12. The mean first passage time is $a_1 = 1.53 \times 10^{32}$ picoseconds. Thus, in the balanced flip-flop soft errors are extremely rare, even when the supply voltage $V_{DD} = 0.2$ V and the full charge on a loaded capacitive output node consists of only 84 electrons. However, this stability arises from assuming perfectly matched current drive of the NMOS and PMOS transistors in the CMOS inverter, which is impossible to guarantee in any realistic technology.

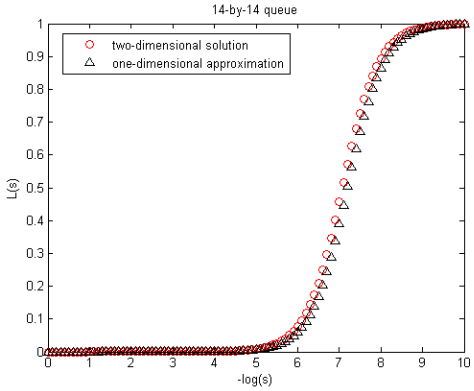


Figure 11 Comparison of the 1D queue approximation to the 2D queue result obtained by numerical solution for a symmetric flip-flop with 14 stored charges per node.

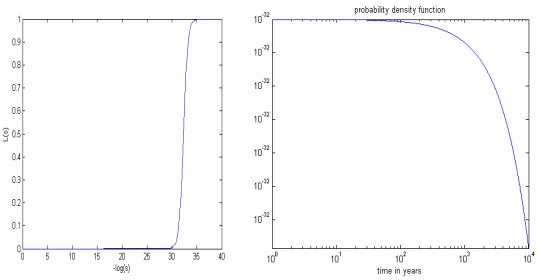


Figure 12 The Laplace transform of the probability density function of the first passage time from state (1,83) to state (83,1) is provided on the left, and the temporal-domain probability density function is on the right.

B. The effect of threshold variations

The rapid scaling of CMOS technology inevitably leads to increased process variations [1]. These variations manifest themselves as fluctuations in transistor parameters, such as L_G,

V_T and parasitic source/drain resistance. The effect of transistor threshold voltage shift δV_T is particularly critical, as it affects the current drive exponentially, with $I_0 \sim \exp[-\delta V_T / nkT]$ (recall that for the technology of Fig. 2, $V_T \sim 0.28$ V and $I_0 = 0.006$ μA/μm). If the thresholds of NMOS and PMOS devices are shifted in opposite directions in an inverter by the same magnitude δV_T, the asymmetrical charge and discharge rates shift the gate threshold and lower one of the input noise margins. When threshold shifts are of opposite sign in the two flip-flops, they produce a worst-case situation in which one state of the flip-flop is less stable and more subject to thermal upset. Fig. 14 shows that the mean time for the flip-flop to leave that state due to thermal noise decreases very rapidly with δV_T. For example, if $\delta V_T = 0.025$ V, a soft error can occur with high probability in several days. Note that 0.025V threshold shift represents a ~10% variation, which is quite reasonable in terms of projected ultimate CMOS technology.

C. The effect of voltage scaling

The supply voltage also has a dramatic effect on soft error rates. The 1D queue approximation model was used to predict the mean time to an error for the symmetric flip-flop case, i.e. no variation from nominal threshold voltage. The result is summarized in Table 1, and the effects of the threshold variations are shown in Figure 13.

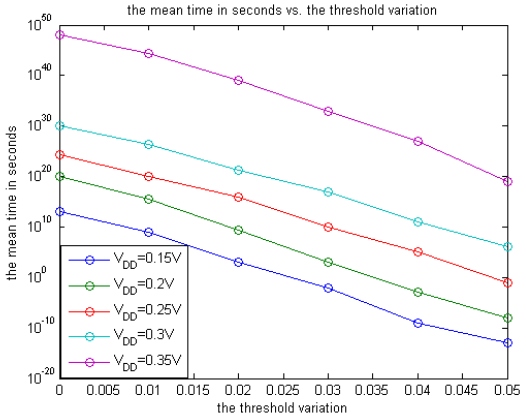


Figure 13 Mean passage time as a function of threshold variation δV_T.

V_{DD} (V)	Number of electrons in NMOS	Number of electrons in CMOS	Mean first passage time (in seconds)
0.15	20	60	10^{13}
0.2	28	84	10^{20}
0.25	40	120	10^{24}
0.3	60	180	10^{30}
0.35	80	240	10^{48}

Table 1 The effect of V_{DD} on soft error rate.

D. Forecast for ultimate CMOS

According to the ITRS roadmap [1], the 2015 workhorse ultra-thin body SOI transistor will have $L_G = 10$ nm and $V_T = 175$mV. The parameters of this future device predicted with the help of the MASTAR 4 tool [22], can be used in solving the Poisson equation for a CMOS load capacitor in order to obtain a rough estimate of the number of electrons in our system. For an operating supply voltage of $V_{DD} = 0.2$V, there are about 72 electrons on the CMOS load capacitor. Based on symmetrical inverters with these theoretical parameters, the prediction for the first passage time is on the order of 10^{16} seconds, a reduction by a factor of 10^4 over the 65 nm technology. Moreover, as in section IV.B, the effect of threshold variation δV_T will further reduce the mean time to failure. The ratio of passage times for the two technologies is approximately constant with δV_T.

E. The basic relation of stored charge to data retention

While it is difficult and somewhat problematic to predict the course of CMOS scaling, the current theory sets a clear limit on the amount of scaling, however achieved, that will still result in satisfactory flip-flops. Figure 14 shows the first passage time for a thermally induced change of state for a flip-flop operating at $V_{DD} = 0.2$ volts. Two curves are shown, one for the symmetric case and the other for leaving the less stable state of a flip-flop with $\delta V_T = .025$ volts. The Poisson rates were derived from the currents of the transistors from the technology of Fig. 2. Because the passage time scales inversely with rate, it is trivial to adjust the assumptions for different rates. In fact, as long as the devices have a current dependence representative of thermal diffusion or thermal emission over a barrier, this figure is applicable to the storage capacity needed for a given level of reliability regardless of whether the devices are CMOS transistors or are other more exotic devices.

Figure 14 shows that even a modest reduction in stored charge from 84 electrons to 55 increases the thermally induced errors to once a year in perfectly symmetric circuits. A 10 % worst-case shift of thresholds ($\delta V_T = 0.025$ V) with 55 stored electrons further reduces the mean time to an error to only 0.01 seconds. As circuits of this type are most often used for registers and cache memory, they are present in large numbers on even fairly modest processors. An array of 10^8 such circuits, only 1% of which are characterized by worst-case threshold shifts, would still have an error rate of ~10^8 per second!

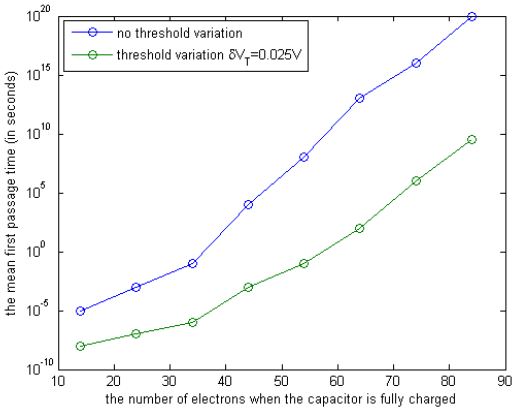


Figure 14 Mean passage time as a function of total stored charge per flip-flop node with and without a threshold variation $\delta V_T = .025$ volts.

V. CONCLUSIONS

This paper provides an analytical model of thermally-induced soft errors in the memory elements of ultimate nanoscale CMOS circuits. Such errors are difficult to estimate with conventional simulation. The analytical model shows that errors will become a serious problem at ultra-low $V_{DD} \sim 0.2$ V. The error rate is extremely sensitive to the total stored charge within the flip-flop memory element and to process-induced variations of threshold voltage. Thus the theory sets a limit to the size scaling possible in low power systems that must achieve a given level of reliable data retention.

[1] International Technology Roadmap for Semiconductors, http://public.itrs.net.

[2] For example, see H. Iwai, "The Future of CMOS Downscaling", in S. Luryi, J. M. Xu, and A. Zaslavsky, eds., Future Trends in Microelectronics: The Nano, the Giga, and the Ultra, New York: Wiley, 2004, p. 26.

[3] K. Nepal, R. I. Bahar, J. Mundy, W. Patterson, A. Zaslavsky, "Designing Logic Circuits for Probabilistic Computation in the Presence of Noise", Design Automation Conference, June 2005.

[4] V. M. Polyakov and F. Schwierz, "Excessive noise in nanoscaled double-gate MOSFETs: A Monte Carlo study", Semicond. Sci. Technol. 19, 145 (2004).

[5] Hua Li et al, "A model for soft errors in the subthreshold CMOS inverter", IEEE workshop on silicon errors in logic-system effects, 2006. http://selse2.selse.org/papers/li.pdf

[6] J. Wyatt, G. Coram, "Nonlinear device noise models:satisfying the thermodynamic requirements", IEEE transactions on electron devices, pp.184-193, 1999.

[7] V. Degalahal, R. Rajaram, N. Vijaykrishanan, Y. Xie , M. J Irwin, " The effect of threshold voltages on soft error rate", in the Proc. 5th International Symposium on Quality Electronic Design (ISQED), March 22-24, 2004 at San Jose, California. Page (s): 503-508

[8] S. Krishnaswamy, G. F. Viamontes, I. L. Markov and J. P. Hayes "Accurate Reliability Evaluation and Enhancement via Probabilistic Transfer Matrices" Proceedings of Design, Automation and Test in Europe Conference and Exhibition (DATE'05)

[9] A. van der Ziel, Noise: Sources, Characterization, Measurement, Prentice Hall, New York, 1970.

[10] R. Sarpeshkar, T. Delbruck, and C. Mead, "White Noise in MOS Transistors and Resistors", IEEE Circuits and Devices Magazine, Vol. 9, No. 6, pp. 23-29, November 1993.

[11] W. J. Anderson, Continuous Time Markov Chains: An Applications-Oriented Approach, Springer-Verlag, New York, 1991.

[12] Scott Hanson et al, "Ultra-low voltage minimum energy CMOS", IBM Journal of Research and Development, June 2006.

[13] Paul, B. C., Raychowdhury, A., and Roy, K., "Device optimization for ultra-low power digital sub-threshold operation", Proceedings of the 2004 international Symposium on Low Power Electronics and Design, August 2004, pp. 96 - 101.

[14] K. Granhaug, S. Aunet, "Improving yield and defect tolerance in multifunction subthreshold CMOS gates", Proceedings of the 21st IEEE International Symposium on Defect and Faulty-Tolerance in VLSI Systems, October 2006.

[15] P. Bai et al, "A 65nm logic technology featuring 35nm gate lengths, enhanced channel strain, 8 Cu interconnect layers, low-k ILD and 0.57 μm^2 SRAM cell", Tech. Digest IEDM 2004, pp. 657-660.

[16] B. Calhoun, A. Wang, and A. Chandrakasan, "Device sizing for minimum energy operation in subthreshold circuits", Proceedings of IEEE 2004 Custom Integrated Circuits Conference.

[17] Samuel Karlin, A First Course in Stochastic Processes, Academic Press, New York, 1975.

[18] D. V. Widder, "What is the Laplace Transform?", The American Mathematical Monthly, 1945, pp.419-425.

[19] C. S. Deo and D. J. Srolovitz, "First passage time Markov chain analysis of rare events for kinetic Monte Carlo: double kink nucleation during dislocation glide", Modeling and Simulation in Materials Science and Engineering, pp.581-596, 2002.

[20] W. H. Press et al, Numerical Recipes in C++, Cambridge University Press, 2nd Edition, 2002.

[21] L. Nagel, "SPICE2: a Computer Program to Simulate Semiconductor Circuits," Memo ERL-M520, Dept. Elect. and Computer Science, University of California at Berkeley, 1975.

[22] The Model for Assessment of CMOS Technologies and Roadmaps (MASTAR), http://www.itrs.net/models.html.