Técnicas probabilísticas para análise de *yield* em nível elétrico usando propagação de erros e derivadas numéricas

Lucas Brusamarello [1]
Roberto da Silva [1]
Gilson I. Wirth [2]
Ricardo A. L. Reis [1]

Resumo: Em tecnologias nanométricas, variações nos parâmetros CMOS são um desafio para o projeto de circuitos com *yield*[a] apropriado. Neste trabalho nós propomos uma metodologia eficiente e precisa para a modelagem estatística de circuitos. Propagação de erros e técnicas numéricas são aplicadas para a modelagem em nível elétrico de variações aleatórias e sistemáticas durante o processo de fabricação. O modelo considera covariâncias entre os parâmetros e correlação espacial, e tem como saída os estimadores estatísticos que podem ser usados em ferramentas de mais alto nível, tais como ferramentas de análise estatística de atraso (SSTA). Além disso, desenvolvemos uma metodologia para a análise quantitativa da contribuição de cada parâmetro para a variância da resposta do circuito.
Como estudos de caso, modelamos o *yield* de uma memória SRAM e uma porta NOR dinâmica de pré-carga. No primeiro, consideramos o impacto do comprimento do canal e da tensão de limiar no tempo de acesso da célula de memória SRAM. Nós desenvolvemos um modelo probabilístico para o atraso de uma NOR dinâmica com *keeper*[b] estático, considerando variações na largura do canal e na tensão de limiar. Comparamos os resultados calculados pela metodologia proposta com dados estatístico obtidos a partir de simulações Monte Carlo. Reportamos ganho de desempenho de 70×, com um erro menor que 1%.
Palavras-chave: análise probabilística, estimação de *yield*, métodos de Monte Carlo, variabilidade no processo, VLSI, projeto visando *yield*.

[a] percentagem de circuitos funcionais em uma pastilha
[b] transistor utilizado em circuitos de lógica dinâmica para atenuar os efeitos da corrente de fuga

[1] Instituto de Informática, UFRGS
{lucas,rdasilva,reis@inf.ufrgs.br}
[2] Universidade Estadual do Rio Grande do Sul - UERGS
{gilson-wirth@uergs.edu.br}

Abstract: In nanometer scale CMOS parameter variations are a challenge for the design of high *yield* integrated circuits. In this work we propose an accurate and computer efficient methodology for statistical modeling of circuit blocks. Numerical error propagation techniques are applied to model random and systematic process variations at electrical level. The model handles co-variances between parameters and spatial correlation, and gives as output the statistical parameters that can be applied at higher level analysis tools, as for instance statistical timing (SSTA) analysis tools. Moreover, we develop a methodology to compute the quantitative contribution of each circuit random parameter to the circuit response variance.
As case studies, we model *yield* loss of a SRAM memory and a pre-charge dynamic-NOR. In the first case, we consider the impact of channel width and voltage threshold to the access time of a SRAM cell. Also, we develop a probabilistic model for time delay of a dynamic NOR with static keeper, considering channel length and voltage threshold variations. We compare results obtained using the proposed model with statistical results obtained by Monte Carlo simulation. A speedup of $50\times$ is achieved, with error less than 1%.
Keywords: Design for *yield*, Monte Carlo methods, Probabilistic analysis, Process variability, Very-large-scale integration, *yield* estimation

1 Introdução

Performance e confiabilidade de circuitos fabricados utilizando tecnologia sub - micrônica são cada vez mais afetados pelas variações no processo de fabricação [17]. Essas variações de natureza estatística devem ser levadas em consideração nas fases de projeto, e ferramentas de *Computer Aided Design* (CAD) devem ser capazes de prever o percentual de circuitos funcionais em uma pastilha ou lote. Assim, circuitos fabricados nessas tecnologias devem ser projetados a fim de atingir um determinado *yield* de produção.

A variabilidade dos parâmetros elétricos pode ser decomposta em parâmetros que apresentam correlação espacial (SC) e parâmetros que não apresentam correlação espacial (NSC)[19] [12]. Variabilidade nos parâmetros NSC pode ser originária de inúmeras fontes, tais como a discretude da matéria e energia (átomos de dopante, fótons, etc). Um exemplo de parâmetro NSC é a variabilidade na tensão de limiar dos transistores (Vt) causada por *Random Dopant Fluctuations* (RDF)[8]. Seja σ_{vt0} o desvio padrão da tensão de limiar de transistores projetados com dimensões mínimas, então a dependência funcional de σ_{vt} de um dado transistor em função das suas dimensões é dado por [18]:

$$\sigma_{vt} = \sigma_{vt0}\sqrt{\frac{L_{min} \times W_{min}}{L \times W}} \tag{1}$$

onde W e L são respectivamente o comprimento e a largura do canal do transistor, enquanto

W_{min} e L_{min} representam as dimensões mínimas para essas medidas.

Os parâmetros espacialmente correlacionados apresentam um componente *inter-die* e um componente *intra-die*. Variações *inter-die* podem acontecer devido a assimetria nos equipamentos (como assimetria na distribuição do gás dentro de uma câmara e gradientes de temperatura em um forno) ou imperfeições na operação de equipamentos e no fluxo de processo. Essas assimetrias afetam a média de um parâmetro entre pastilhas, *wafer* ou lote. Variações *intra-die* são o desvio de um parâmetro de seu valor nominal, as quais podem ser causadas por padrões de leiaute. Parâmetros como espessura do óxido (Tox), largura (L) e comprimento (W) do canal do transistor podem apresentar correlação espacial.

A fim de calcular média e variância do atraso do circuito utilizando análise de atraso estatística (SSTA), as portas lógicas precisam ser caracterizadas em nível elétrico. Em [7] é empregada caracterização de células lógicas utilizando propagação de erro e derivadas numéricas. Contudo, sua modelagem não considera correlação espacial em nível elétrico (embora esta seja considerada pelo algoritmo de SSTA). Além disso, os demais trabalhos da área não consideram a análise quantitativa da contribuição de cada parâmetro à variância, e apenas aproximações de primeira ordem são utilizadas para as derivadas numéricas.

Análise estatística de características elétricas de circuitos analógicos e digitais é comumente baseada no Método Monte Carlo (MC) [4] utilizando uma grande amostra de simulações em nível elétrico. Simulações Monte Carlo atualmente é o padrão empregado pela indústria para análise de variações em nível elétrico, e é suportado por simuladores elétricos [16].*Yield* em memórias SRAM utilizando simulações MC foi estudado em [1], [3] e [2].

Propagação de Erros (EP) é uma forma viável de computar a resposta estatística do circuito sem a necessidade de um grande número de simulações exigido por técnicas de amostragem. Através da propagação de erros é possível calcular o erro de uma medida tendo como entrada (1) as derivadas parciais da função de interesse em relação às variáveis dependentes e (2) as variâncias das variáveis dependentes (as quais são dadas pela *foundry*).

Propagação de erro para análise de *yield* em memória SRAM foi explorado em [9] e [10]. Contudo, estes trabalhos apresentam solução somente para a modelagem de parâmetros NSC e variações na tensão de limiar (Vt) são consideradas. Além disso, a metodologia apresentada nesses trabalhos somente pode ser aplicada a memória SRAM, pois as derivadas são analíticas e assim específicas para esse circuito. Nosso objetivo é estender trabalhos passados sobre propagação de erros para suportar variáveis SC. Além disso, o *framework* desenvolvido é genérico devido ao uso de derivadas numéricas (calculadas através de simulação elétricas) ao invés do uso de derivadas analíticas.

Este trabalho apresenta uma metodologia genérica para caracterização de blocos em nível elétrico, capaz de considerar variáveis SC e NSC e correlação entre parâmetros elétricos. O método mantêm a generalidade de técnicas Monte Carlo, ainda largamente empre-

gadas em simuladores elétricos comerciais[16], com uma diminuição drástica no tempo de processamento. Além disso, nós implementamos um método capaz de apontar os parâmetros que apresentam maior contribuição à variabilidade do circuito.

O *framework* aqui desenvolvido pode ser aplicado a uma diversa gama de blocos combinacionais e seqüenciais. As fórmulas para o cálculo da variâncias são independentes da topologia do circuito (célula SRAM, multiplexador, portas complexas, ...), a função de interesse (atraso, corrente de fuga, potência, ...) e os parâmetros elétricos considerados como variáveis aleatórias (Vt, W, L, Tox, ...). Contudo, uma vez que o número de simulações elétricas requeridas é linear com o número de variáveis aleatórias, o método só apresenta ganho de desempenho em relação a Monte Carlo para circuitos com pequeno número de transistores.

Como estudo de caso para a metodologia desenvolvida, nós a aplicamos a dois estudos de caso: (1) análise e otimização de *yield* de uma memória SRAM e (2) análise de *yield* de lógica dinâmica.

Este trabalho está organizado da seguinte forma. A seção 2 mostra a modelagem estatística de circuitos e introduz conceitos básicos de teoria de erros e simulação Monte Carlo. Na seção 3 são apresentadas as fórmulas para o cálculo das derivadas. As seções 4 e 5 mostram dois estudos de caso: análise de *yield* de uma memória SRAM análise de *yield* de uma porta NOR dinâmica de pré-carga. Finalmente, expomos nossas conclusões sobre este trabalho na seção 6.

2 Modelo

Considere um circuito ω, composto por n transistores representados como componentes do vetor $\overrightarrow{\tau} = (\tau_1, \ldots, \tau_n)$, interconectados de acordo com uma topologia Γ. Por definição, a resposta do circuito é dada pela função $F(\overrightarrow{\alpha}_1, \ldots, \overrightarrow{\alpha}_n, \overrightarrow{\beta}_1, \ldots, \overrightarrow{\beta}_n, \omega)$ onde os vetores $\overrightarrow{\alpha}_i = (\alpha_i^{(1)}, ..., \alpha_i^{(p)})$ e $\overrightarrow{\beta}_i = (\beta_i^{(1)}, ..., \beta_i^{(q)})$ representam respectivamente os parâmetros NSC e SC do transistor i, onde p é o número de parâmetros independentes e q é o número de parâmetros espacialmente-correlacionados. Por exemplo, os casos $\overrightarrow{\alpha}_3 = (V_t)$ e $\overrightarrow{\beta}_3 = (T_{ox}, L, W)$ representam parâmetros do transistor τ_3, incluindo espessura do óxido, tensão de limiar, e as dimensões do transistor.

Com a presença de variabilidade no processo de fabricação, características elétricas e dimensões do circuito são variáveis aleatórias, e conseqüentemente a resposta do circuito (função de interesse) é uma variável aleatória. Considere, sem perda de generalidade, que os parâmetros (por exemplo T_{ox}, Vt, L, W) são variáveis aleatórias com distribuição Normal, cada uma com média (μ) e variância (σ^2). Por exemplo, $\alpha_i^{k_1} = N\left(\mu(\alpha_i^{k_1}), \sigma^2(\alpha^{k_1})\right)$ e $\beta_i^{k_2} = N\left(\mu(\beta_i^{k_2}), \sigma^2(\beta^{k_2})\right)$, onde $i = 1, ..., n$, $k_1 = 1, ..., p$ e $k_2 = 1, ..., q$.

A resposta estocástica S é uma função que depende de $N = n*(p+q)$ variáveis aleatórias (incluindo parâmetros SC e NSC), dada por

$$S = F(\overrightarrow{\alpha}_1, \ldots, \overrightarrow{\alpha}_n, \overrightarrow{\beta}_1, \ldots, \overrightarrow{\beta}_n, \omega)$$

2.1 Parâmetros SC

Correlação espacial faz com que todos os transistores tenham seus parâmetros sincronizados. Por exemplo, se a dimensão W_1 do transistor τ_1 mudar uma quantidade δW, a dimensão W_2 do transistor τ_2 tem um acréscimo da mesma quantidade δW, embora suas médias sejam diferentes. As variáveis correlacionadas espacialmente podem ser escritas como

$$\beta_i^j = \mu(\beta_i^j) + \xi_j \cdot \sigma(\beta^j)$$

onde $\xi_j = N(0,1)$ é uma variável normal padrão independente do transistor $0 \leq i \leq n$. Ou seja, para o mesmo circuito, a variável j apresenta uma mesma variação $\xi_j \cdot \sigma(\beta^j)$, independente do transistor. Em outras palavras, as variáveis $\beta_1^j, ..., \beta_n^j$ são as mesmas variáveis aleatórias, exceto das suas médias. Buscando a contribuição destas vaiáveis para a estimativa do erro, é importante definir a variável geral $\beta^j = \mu(\beta^j) + \xi_j \cdot \sigma(\beta^j)$, onde $\mu(\beta^j)$ é uma constante que independe do transistor. A partir disso, podemos concluir que para cada transistor $\beta_i^j(k) = \mu(\beta_i^j) + \xi_j \cdot \sigma(\beta^j) = \mu(\beta_i^j) + \beta^j - \mu(\beta^j)$. Isso leva à oportuna simplificação $F(\overrightarrow{\alpha}_1, \ldots, \overrightarrow{\alpha}_n, \overrightarrow{\beta}_1, \ldots, \overrightarrow{\beta}_n, \omega) = F(\overrightarrow{\alpha}_1, \ldots, \overrightarrow{\alpha}_n, \beta^1, \ldots, \beta^q, \omega)$, e utilizando a regra da cadeia, a computação das derivadas parciais é dada por

$$\frac{\partial F}{\partial \beta^j} = \sum_{i=1}^{n} \frac{\partial F}{\partial \beta_i^j} \frac{\partial \beta_i^j}{\partial \beta^j} = \sum_{i=1}^{n} \frac{\partial F}{\partial \beta_i^j} \tag{2}$$

uma vez que $\partial \beta_i^j / \partial \beta^j = 1$, para todo $i \in \{1, ..., n\}$.

2.2 Propagação de Erro e Monte Carlo

Dada a natureza estatística do processo de fabricação, a resposta do circuito é Gaussiana e uma estimativa do erro na variável S é dada pela fórmula da propagação de erro [13], que está de acordo com a equação 2 e provê

$$
\begin{aligned}
\sigma_S^2 &= \sum_{i=1}^{n}\sum_{j=1}^{p}\left(\left.\frac{\partial F}{\partial \alpha_i^j}\right|_{\alpha_i^j=\mu(\alpha_i^j)}\right)^2 \sigma^2(\alpha^j) + \sum_{j=1}^{q}\left(\sum_{i=1}^{n}\left.\frac{\partial F}{\partial \beta_i^j}\right|_{\beta_i^j=\mu(\beta_i^j)}\right)^2 \sigma^2(\beta^j) \\
&+ 2\sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{k=j}^{p}\left(\left.\frac{\partial F}{\partial \alpha_i^j}\right|_{\alpha_i^j=\mu(\alpha_i^j)}\left.\frac{\partial F}{\partial \alpha_i^k}\right|_{\alpha_i^k=\mu(\alpha_i^k)}\right)\sigma(\alpha^j,\alpha^k) \\
&+ 2\sum_{i=1}^{n}\sum_{j=1}^{q}\sum_{k=j}^{q}\left(\left.\frac{\partial F}{\partial \beta_i^j}\right|_{\beta_i^j=\mu(\beta_i^j)}\left.\frac{\partial F}{\partial \beta_i^k}\right|_{\beta_i^k=\mu(\beta_i^k)}\right)\sigma(\beta^j,\beta^k) \\
&+ 2\sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{k=1}^{q}\left(\left.\frac{\partial F}{\partial \alpha_i^j}\right|_{\alpha_i^j=\mu(\alpha_i^j)}\left.\frac{\partial F}{\partial \beta_i^k}\right|_{\beta_i^k=\mu(\beta_i^k)}\right)\sigma(\alpha^j,\beta^k) \qquad (3)
\end{aligned}
$$

O leitor deve perceber que a correlação entre os parâmetros elétricos não exige acréscimo no número de simulações. Como prova, basta verificar que as derivadas que multiplicam os coeficientes de correlação precisam ser calculados mesmo na ausência destes.

O desvio padrão obtido de uma amostra de n_{sample} medidas experimentais de S – denotadas por S_1, S_2, ..., $S_{nsample}$ – calculado pela expressão

$$\delta_S = \sqrt{\frac{1}{(n_{sample}-1)}\sum_{i=0}^{n_{sample}}(S_i - \langle S_i\rangle)^2}$$

é numericamente igual a σ_S se o tamanho da amostra n_{sample} for suficientemente grande, ou seja

$$\lim_{n_{sample}\to\infty}\delta_S = \sigma_S$$

Simulação Monte Carlo [16] é comumente empregada para computar a função de densidade de probabilidades (PDF) de alguma resposta do circuito (atraso, potência, corrente de fuga, ...). Mas para isso é necessário um grande número de simulações elétricas, pois o erro em simulações Monte Carlo é $O(1/\sqrt{n_{sample}})$. Propagação de erro tem como entrada (1) as derivadas parciais da função de resposta do circuito em relação às variáveis aleatórias, (2) o desvio padrão dos parâmetros e (3) os coeficientes de correlação entre os parâmetros. Os desvios padrão e os coeficientes de correlação são dados pela *foundry*, e estimativas das derivadas $\left.\frac{\partial F}{\partial k_i}\right|_{k_i=\overline{k_i}}$ podem ser calculadas numericamente com o simulador elétrico.

2.3 Contribuição dos parâmetros à variabilidade do circuito

Através de propagação de erro é possível obter valores quantitativos para a contribuição de cada variável aleatória à variabilidade do circuito. Revendo a equação 3, observa-se que a contribuição individual de uma variável que não apresenta correlação espacial é dada por

$$\left(\frac{\partial F}{\partial \alpha^k}\right)^2 \sigma^2_{\alpha^k} \tag{4}$$

Para uma variável espacialmente sincronizada β_i^k, a contribuição é dada por

$$\left(\frac{\partial F/\partial \beta_i^k}{\sum_{j=1}^{m} \partial F/\partial \beta_j^k}\right)\left(\frac{\partial F}{\partial \beta^k}^2 \sigma^2_{\beta^k}\right) \tag{5}$$

para m variáveis sincronizadas.

3 Estimativas numéricas das Derivadas

Neste trabalho nós utilizamos aproximações numéricas das derivadas a fim de apresentar uma metodologia genérica, independente da topologia do circuito. Nós exploramos aproximações lineares usando 1 e 2 pontos próximos dos valores nominais, a fim de obter a sensibilidade da resposta do circuito em relação às variáveis aleatórias de interesse.

Considerando uma função de n variáveis $f = f(x_1, x_2, ..., x_n)$, tem-se que

$$\sigma_f^2 = (\partial f/\partial x_1)^2_{x_1=\overline{x}_1} \sigma^2_{x_1} + ... + (\partial f/\partial x_n)^2_{x_n=\overline{x}_n} \sigma^2_{x_n},$$

e assim

$$\left.\frac{\partial f}{\partial x_i}(x_1, ..., x_n)\right|_{x_i=\overline{x}_i} = \frac{f(x_1, .., \overline{x}_i + \varepsilon, ...x_n) - f(x_1, .., \overline{x}_i, ...x_n)}{\varepsilon} + O(\varepsilon) \tag{6}$$

para todo $i = 1, \ldots, n$.

Neste caso precisamos de 2 simulações elétricas para calcular cada derivada, uma para $f(x_1, .., \overline{x}_i + \varepsilon, ...x_n)$ e outra para $f(x_1, .., \overline{x}_i, ...x_n)$. Como $f(x_1, .., \overline{x}_i, ...x_n)$ pode ser calculada apenas uma vez, então o cálculo de todas as derivadas exige $n+1$ simulações.

A fim de obter uma aproximação mais precisa, manipulações algébricas de expansões em Série de Taylor resultam em uma fórmula com precisão $O(\varepsilon^2)$:

$$\left.\frac{\partial f}{\partial x_i}(x_1,...,x_n)\right|_{x_i=\overline{x}_i} = \frac{f(x_1,..,\overline{x}_i+\varepsilon,...x_n)-f(x_1,..,\overline{x}_i-\varepsilon,...x_n)}{2\varepsilon}+O(\varepsilon^2) \qquad (7)$$

e neste caso $2n$ simulações são necessárias para o cálculo de todas as derivadas.

Uma aproximação com $O(\varepsilon^4)$ pode ser obtida nas aproximações numéricas, realizando um cálculo de acordo com

$$\begin{aligned}\left.\frac{\partial f}{\partial x_i}(x_1,...,x_n)\right|_{x_i=\overline{x}_i} = & \frac{f(x_1,..,\overline{x}_i+2\varepsilon,...x_n)-f(x_1,..,\overline{x}_i-2\varepsilon,...x_n)}{4\varepsilon} \\ & -\frac{4}{3}\frac{f(x_1,..,\overline{x}_i+\varepsilon,...x_n)-f(x_1,..,\overline{x}_i-\varepsilon,...x_n)}{2\varepsilon}+O(\varepsilon^4)\end{aligned} \qquad (8)$$

contudo nesse caso são necessárias $4n$ simulações elétricas.

Em problemas de caracterização estatística em nível elétrico tem-se um número pequeno de transistores, e portanto essas fórmulas significam uma drástica diminuição no número de simulações, comparado a processos de amostragem. Fórmulas com 2 e 4 pontos ao redor do valor nominal apresentam uma precisão maior, ao custo de aumento no tempo de processamento.

4 Estudo de Caso: Memória SRAM

Esta seção apresenta a aplicação da metodologia previamente descrita para análise e otimização de *yield* de uma memória SRAM. Utilizamos propagação de erros e derivadas numéricas (1, 2 e 4 pontos ao redor da média) para caracterização do tempo de acesso de uma célula SRAM [5]. A figura 1 apresenta o esquemático de uma célula SRAM e a nomenclatura adotada neste trabalho.

Nós analisamos a influência da variabilidade no comprimento do canal e na tensão de limiar dos transistores na variabilidade do tempo de acesso da célula de memória. Utilizamos o software HSPICE e tecnologia BPTM 70nm [6] para as simulações elétricas. Considere os seguintes valores nominais e desvios padrões:

Parâmetro	valor nominal	3σ
Vt(PMOS)	-0.22 V	40 mV
Vt(NMOS)	0.2 V	40 mV
W	100 nm	33 nm

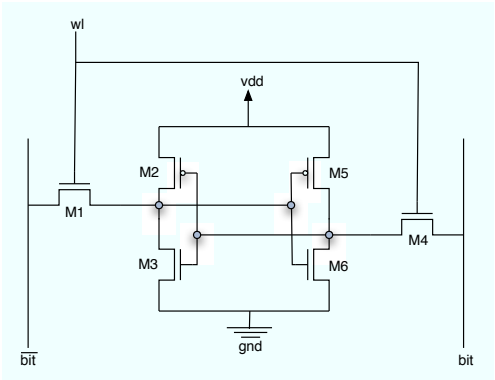

Figura 1. Célula SRAM

Esses valores estão de acordo com [14] e [11]. Para este estudo de caso, supomos a inexistência de coeficientes de correlação entre parâmetros, mas consideramos correlação espacial no comprimento dos transistores.

Tempo de acesso é o *tempo necessário para ler o valor armazenado na célula de memória*. Uma *falha de tempo de acesso* acontece se o tempo de acesso da célula é maior que o maior tempo de acesso permitido pelo projeto [8] [9]. O tempo de acesso pode ser escrito como uma função das variáveis aleatórias, assim temos

$$T_{AC} = T_{AC}(W_{M1}, \ldots, W_{M6}, Vt_{M1}, \ldots, Vt_{M6},)$$

Reescrevendo a equação 3 para o problema do tempo de acesso da célula SRAM considerando correlação espacial em W e assumindo Vt como variáveis aleatórias independentes (sem correlação espacial), temos que a variância no tempo de acesso é dada por

$$\sigma^2_{T_{AC}} = \left(\sum_{i=1}^{6} \frac{\partial T_{AC}}{\partial W_{Mi}}\bigg|_{W_{Mi}=\overline{W_{Mi}}} \right)^2 \sigma^2_W + \sum_{i=1}^{6} \left(\frac{\partial T_{AC}}{\partial Vt_{Mi}}\bigg|^2_{Vt_{Mi}=\overline{Vt}_{Mi}} \right) \sigma^2_{Vt} \tag{9}$$

Considere T_{MAX} uma constante de projeto relacionada à freqüência do circuito, então a probabilidade p da célula SRAM não apresentar falha de tempo de acesso é dada por

$$p = P(T_{AC} \leq T_{MAX}) = \int_{-\infty}^{T_{MAX}} \frac{1}{\sigma_{T_{AC}}\sqrt{2\pi}} e^{-\frac{(x-\mu_{T_{AC}})^2}{2\sigma^2_{T_{AC}}}} dx \tag{10}$$

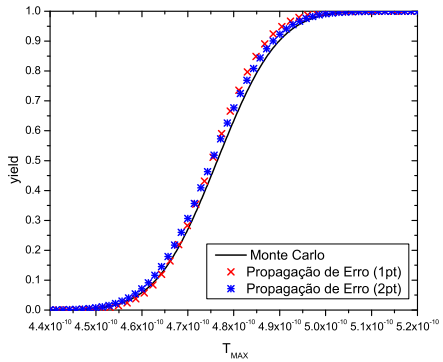

Figura 2. *Yield* de uma célula SRAM em função de T_{MAX}, estimado a partir de MC (10^3 simulações) e propagação de erro usando 1 ponto para derivada (13 simulações) e 2 pontos para derivada (25 simulações).

4.1 *Yield* da célula SRAM

Usando estimativas numéricas para as derivadas e desvio padrão dado pela *foundry*, a fórmula da propagação de erros pode ser empregada a fim de calcular a variância do tempo de acesso. A partir da variância (calculada por EP) e o valor médio (calculado por uma simulação utilizando os valores nominais dos parâmetros), a PDF do tempo de acesso pode ser estimada. A figura 2 mostra o *yield* de uma célula SRAM considerando falhas no tempo de acesso em função de T_{MAX}. As curvas foram calculadas utilizando os valores obtidos através de propagação de erros (1 e 2 pontos ao redor dos valores nominais) e Monte Carlo (10^3 simulações).

4.2 *Yield* da memória SRAM

Memórias SRAM apresentam uma arquitetura regular na qual a maior parte do *chip* é composta por células SRAM dispostas como numa grade. Considere uma memória com N_{COL} colunas, N_{ROW} linhas e N_R colunas redundantes, conforme mostra a figura 3. Se uma ou mais células de memória falham em uma coluna, aquela coluna deve ser substituída por uma coluna redundante – isso pode ser feito durante a fase de teste do circuito, ajustando um conjunto de fusíveis. Se mais do que N_R colunas falham, então o circuito é defeituoso e deve ser descartado, diminuindo o *yield* do projeto.

Denotando p como a probabilidade da célula SRAM funcionar corretamente (computado através de EP ou MC) diante da variabilidade no processo, $P_{COL} = (p)^{N_{ROW}}$ é a probabi-

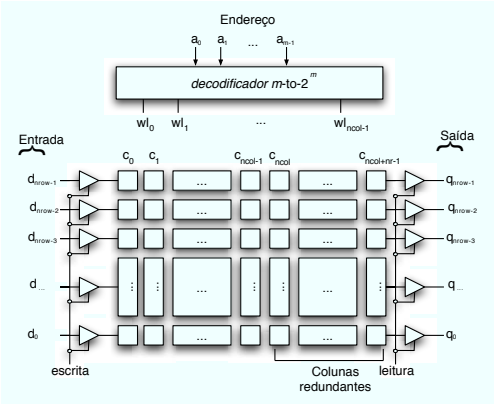

Figura 3. Esquemático de uma memória SRAM.

lidade de nenhuma célula falhar em uma coluna. O *yield* do *chip* é dado pela probabilidade de funcionarem pelo menos N_{COL} de um total de $N_{COL}+N_R$ colunas, que é dado por um somatório de distribuições binomiais [8]:

$$P_{MEM} = \sum_{i=N_{COL}}^{N_{COL}+N_R} \binom{N_{COL}+N_R}{i} (P_{COL})^i (1-P_{COL})^{N_{COL}+N_R-i} \tag{11}$$

A figura 4 apresenta o *yield* de uma memória SRAM com $N_{COL} = 512$, $N_{ROW} = 32$ e $N_R = 24$, em função de T_{MAX}. O erro absoluto de EP utilizando 1 ponto para as derivadas é 4×10^{-12}, enquanto o erro para 2 pontos é 1.5×10^{-12} (ambos em relação a MC utilizando 10^3 simulações). Propagação de erros exige 13 e 25 simulações considerando, respectivamente, 1 e 2 pontos para as aproximações das derivadas. Em nossas simulações, MC teve tempo de processamento de 3400s, enquanto EP usando 2 pontos teve tempo menor que 90s. As simulações foram realizadas em uma Sun Fire V240 (UltraSPARC IIIi 1 GHz) com dois processadores.

A partir das equações 5 e 4, a contribuição individual da variável NSC Vt_{Mi} para a variância no tempo de acesso é dada por $\frac{\partial T_{AC}}{\partial Vt_{Mi}}^2 \sigma^2_{VtMi}$ e para a variável SC W_{Mi} a contribuição é dada por $(\frac{\partial T_{AC}/\partial W_{Mi}}{\sum_{j=1}^{m} \partial T_{AC}/\partial W_{Mj}})(\frac{\partial T_{AC}}{\partial W}^2 \sigma^2_W)$ para m variáveis sincronizadas. A figura 5 mostra a contribuição de cada parâmetro para a variância do tempo de acesso da célula SRAM. A contribuição do comprimento do canal é ordens de magnitude mais importante do que a variabilidade na tensão de limiar. Além disso, os transistores M1,M4 e M3,M6 (é importante ressaltar que a célula é simétrica e portanto os transistores devem ser analisados aos pares) apresentam contribuições que são ordens de magnitude maiores do que o par M2,M5.

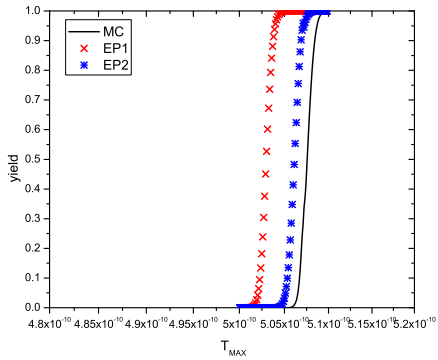

Figura 4. *Yield* de uma célula SRAM em função de T_{MAX}

Assim, a partir dos dados obtidos pela análise da contribuição de cada parâmetro, induz-se que um redimensionamento dos transistores M1,M4 ou M3,M6 leva ao aumento do *yield* da memória. A figura 6 mostra o *yield* em função do comprimento dos transistores M1 e M4. Conforme esperado, o *yield* é aumentado drasticamente através do aumento das dimensões desses transistores.

5 Estudo de caso: Lógica Dinâmica com pré-carga

A topologia de uma porta NOR dinâmica com pré-carga e *keeper* estático [15] é conforme representado na figura 7. Este circuito é composto por uma saída dinâmica, um inversor estático CMOS e um transistor *keeper*. Quanto mais robusto o transistor M_k, maior deve ser o atraso da porta e maior a tolerância desta a ruído.

Podemos escrever o atraso de uma transição $00...00 \rightarrow 10...00$ no circuito (transição $0 \rightarrow 1$ em uma das entradas da porta) em função dos parâmetros elétricos dos $(n+4)$ transistores: $\{Mi\}_{i=0}^{n-1}$, Mc, Mk, Mp e Mn. As tensões de limiar são representadas por Vt_{Mc}, Vt_{Mk}, Vt_{Mp}, Vt_{Mn} e $\{Vt_{Mi}\}_{i=0}^{n-1}$. Assumimos que quanto a largura do canal, a variabilidade possui duas componentes: SC e NSC. Assim, o atraso de uma transição t_i (atraso de uma transição na entrada I_i para i arbitrário) pode ser escrito como

$$t_i = t_i\left(L^{sc}, L_{Mc}^{nsc}, L_{Mp}^{nsc}, L_{Mn}^{nsc}, L_{Mk}^{nsc}, \{L_i^{nsc}\}_{i=0}^{n-1}, Vt_{Mc}, Vt_{Mp}, Vt_{Mn}, Vt_{Mk}, \{Vt_{Mi}\}_{i=0}^{n-1}\right). \quad (12)$$

E, a partir da equação 3 e assumindo simetria nas entradas (todos os transistores têm

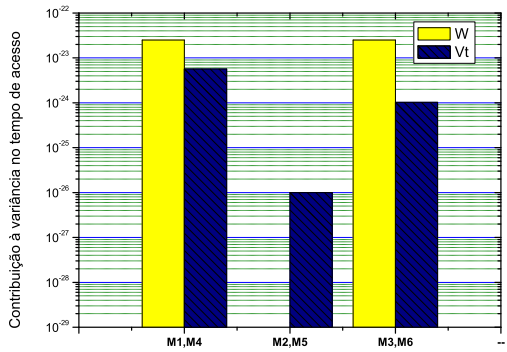

Figura 5. Contribuição de cada parâmetro na variância do tempo de acesso.

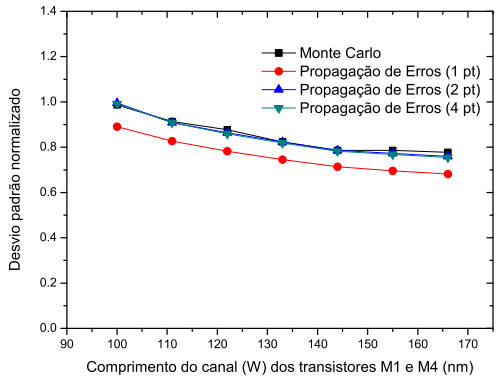

Figura 6. Impacto do comprimento dos transistores M1 e M4 no desvio padrão do tempo de acesso da célula SRAM, obtido com MC e EP.

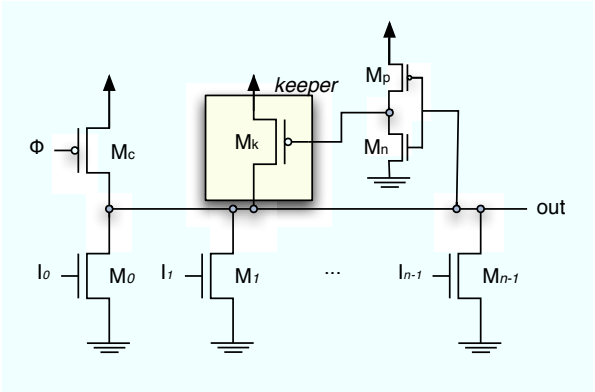

Figura 7. Nor dinâmica de pré-carga com keeper estático

as mesmas dimensões), temos que a variância no atraso dessa transição é dado por

$$\begin{aligned}
\sigma_t^2 &= \left(\frac{\partial t}{\partial L_{Mc}^{sc}} + \frac{\partial t}{\partial L_{M0}^{sc}} + (n-1)\frac{\partial t}{\partial L_{M1}^{sc}} + \frac{\partial t}{\partial L_{Mn}^{sc}} + \frac{\partial t}{\partial L_{Mp}^{sc}} + \frac{\partial t}{\partial L_{Mk}^{sc}}\right)^2 \sigma_{L_{sc}}^2 \\
&+ \left(\frac{\partial t}{\partial L_{Mc}^{nsc}}\right)^2 \sigma_{L_{Mc}^{nsc}}^2 + \left(\frac{\partial t}{\partial Vt_{Mc}}\right)^2 \sigma_{Vt}^2 + \left(\frac{\partial t}{\partial L_{M0}^{nsc}}\right)^2 \sigma_{L^{nsc}}^2 + \left(\frac{\partial t}{\partial Vt_{M0}}\right)^2 \sigma_{Vt}^2 \\
&+ (n-1)\left(\frac{\partial t}{\partial L_{M1}^{nsc}}\right)^2 \sigma_{L^{nsc}}^2 + (n-1)\left(\frac{\partial t}{\partial Vt_{M1}}\right)^2 \sigma_{Vt}^2 + \left(\frac{\partial t}{\partial L_{Mk}^{nsc}}\right)^2 \sigma_{L^{nsc}}^2 + \left(\frac{\partial t}{\partial Vt_{Mk}}\right)^2 \sigma_{Vt}^2 \\
&+ \left(\frac{\partial t}{\partial L_{Mn}^{nsc}}\right)^2 \sigma_{L^{nsc}}^2 + \left(\frac{\partial t}{\partial Vt_{Mn}}\right)^2 \sigma_{Vt}^2 + \left(\frac{\partial t}{\partial L_{Mp}^{nsc}}\right)^2 \sigma_{L^{nsc}}^2 + \left(\frac{\partial t}{\partial Vt_{Mp}}\right)^2 \sigma_{Vt}^2 \qquad (13)
\end{aligned}$$

A fim de modelar estatisticamente o atraso de portas lógicas, a metodologia deve ser capaz de caracterizar a probabilidade do atraso da porta T ser menor que um dado τ_{max}. Além disso, é importante a caracterização da própria distribuição do atraso da porta. Podemos escrever que a probabilidade do atraso de uma transição em I_i ser igual ou menor a τ_{max} é dada por $f_i(t_i \leq \tau_{max}) = \int_{-\infty}^{\tau_{max}} p_{\bar{t}_{Mi},\sigma_{Mi}}(t_i)dt_i$, onde $p_{\bar{t}_i,\sigma_i}(t)$ é uma PDF Normal com média $\bar{t}_i$ e desvio padrão σ_i. Esses valores podem ser obtidos por simulações MC ou por EP. Supondo que todas as entradas t_i são variáveis aleatórias independentes, temos que a probabilidade do atraso da porta ser menor que τ_{max} é

$$P(T < \tau_{max}) = f_0(t_0 \leq \tau_{max})\, f_1(t_1 \leq \tau_{max}) \ldots f_{n-1}(t_{n-1} \leq \tau_{max}).$$

Estamos interessados na probabilidade do atraso da porta pertencer ao intervalo $[\tau_{max} - d\tau_{max}, \tau_{max}]$, que para $h \lll 1$ é dado por

$$\begin{aligned} F(\tau_{max}) &\sim \text{Prob}\ (t_0 \leq \tau_{max}, t_1 \leq \tau_{max}, ..., t_{n-1} \leq \tau_{max} \mid \\ &\quad \text{ao menos um } t_i \in [\tau_{max}, \tau_{max} - h]) \\ &= \text{Prob}\ (\max\{t_i\} \in [\tau_{max}, \tau_{max} - h]) \\ &= \text{Prob}\ (\{t_0 \leq \tau_{max}, t_1 \leq \tau_{max}, ..., t_{n-1} \leq \tau_{max}\} / \\ &\quad \{t_0 \leq \tau_{max} - h, t_1 \leq \tau_{max} - h, ..., t_{n-1} \leq \tau_{max} - h\}) \\ &= \text{Prob}\ (\{t_0 \leq \tau_{max}, t_1 \leq \tau_{max}, ..., t_{n-1} \leq \tau_{max}\}) - \\ &\quad \text{Prob}\ (\{t_0 \leq \tau_{max} - h, t_1 \leq \tau_{max} - h, ..., t_{n-1} \leq \tau_{max} - h\}) \\ &= P(T < \tau_{max}) - P(T < \tau_{max} - h). \end{aligned}$$

Assim, a probabilidade do atraso da porta dinâmica ser menor que a constante τ_{max}, considerando simetria nas entradas, é dado por

$$P(T < \tau_{max}) = f(T \leq \tau_{max})^n = \left(\int_{-\infty}^{\tau_{max}} p_{\bar{t},\sigma}(\tau) d\tau \right)^n \tag{14}$$

A partir dos resultados previamente obtidos, conclui-se que a PDF do atraso da porta dinâmica, o qual é dado pela transição t_i mais lenta da porta, considerando $h \lll 1$, é dada por

$$F(T) = \left(\int_{-\infty}^{\tau_{max}} p_{\bar{t},\sigma}(T) dT \right)^n - \left(\int_{-\infty}^{\tau_{max}-h} p_{\bar{t},\sigma}(T) dT \right)^n \tag{15}$$

5.1 Resultados

Considere a NOR dinâmica com pré-carga conforme mostrado na figura 7. Seja W_{Mi} o comprimento dos transistores de *pull-down*, $W_{M_{clk}}$ o comprimento do transistor de *clock*, $W_{M_{clk}}$, W_{Mp}, W_{Mn} e W_{Mk} são o comprimento dos transistores M_p, M_n e M_k, respectivamente. Então, seja $W_{Mi} = W_{M_n} = 1\mu m$ e $W_{M_{clk}} = W_{M_p} = 2.5\mu m$.

A equação 13 se refere à variância do atraso de uma transição $0 \rightarrow 1$ em uma das entradas da NOR dinâmica. Sem perda de generalidade (pois as entradas são simétricas)

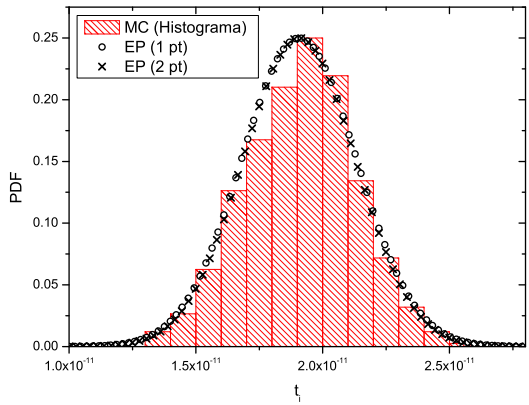

Figura 8. PDF obtida com EP comparada a histograma por MC para o atraso de transição em uma entrada em NOR dinâmica de 8 entradas com keeper ($W_k = 500nm$).

consideramos transição em I_0. As derivadas parciais de 6 transistores (sendo que cada um contém 3 parâmetros) devem ser calculadas, resultando na necessidade de 18 simulações elétricas. A figura 8 apresenta uma comparação entre propagação de erros usando 1 ou 2 pontos ao redor da média e MC com 10^3 iterações. O circuito considerado é uma NOR com 8 entradas, onde $W_{Mk} = 500nm$. Propagação de erro utilizando 1 ponto para derivadas numéricas requer 19 simulações Spice, enquanto a abordagem utilizando 2 pontos requer 36 simulações.

O gráfico 9 mostra a variância do atraso em uma NOR dinâmica com 8 entradas em função do comprimento do canal do transistor keeper. Propagação de erros utilizando 1 e 2 pontos ao redor da média é comparado com MC utilizando 10^3 iterações. EP usando 1 ponto ao redor da média apresenta erro de até 2% comparado a MC, enquanto a abordagem com 2 pontos para as derivadas apresenta erro sempre menor a 1%. O desvio padrão relativo é normalizado pelo desvio padrão relativo de uma NOR dinâmica sem *keeper*, a fim de possibilitar a análise do impacto do tamanho do keeper no projeto. A curva mostra que existe um ponto ótimo para a robustez do keeper, onde a variância é mínima. O caso onde $W_{Mk} = 400nm$ apresenta diminuição de 3% no desvio padrão quando comparado a um projeto sem keeper estático, enquanto um projeto usando $W_{Mk} = 1\mu m$ apresenta acréscimo de 6% na variabilidade do atraso.

A partir do desvio padrão obtido por EP e pela aproximação da média utilizando uma simulação com os valores nominais dos parâmetros (ou estes estimadores podem ser

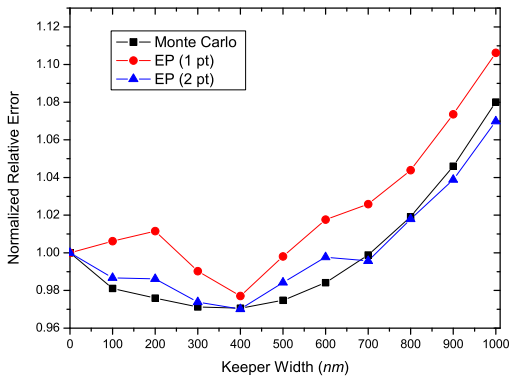

Figura 9. Erro relativo para NOR dinâmica com 8 entradas em função da robustez do keeper, normalizado pelo desvio padrão relativo de uma NOR dinâmica de 8 entradas sem keeper.

computados em um processo de amostragem) a probabilidade do desvio da porta T ser menor que τ_{max} pode ser calculado utilizando a expressão 14. A partir disso, a figura 10 mostra o *yield* de uma NOR dinâmica de 8 entradas com keeper estático ($W_{Mk} = 500nm$) em função da constante de projeto τ_{max}. A PDF do atraso da porta T é calculada com a expressão 15. A figura 11 expõe a PDF do atraso da NOR dinâmica com keeper estático. Ambas as fórmulas requerem o cálculo de σ_τ e $\overline{\tau}$, que são os estimadores estatísticos para o atraso em uma transição em uma entrada I_0, os quais podem ser calculados com EP ou MC. Enquanto a abordagem de EP usando 1 ponto para a derivada numérica requer 19 simulações, EP usando 2 pontos requer 37. Assim, nos nossos experimentos foi relatado um ganho de performance de até $50\times$.

Uma metodologia de análise da contribuição dos parâmetros para a variabilidade do circuito é importante para projetos visando otimização de *yield*. Nós aplicamos as fórmulas expostas na seção 2.3 para a NOR dinâmica com keeper estático. A figura 12 mostra a contribuição de cada parâmetro para a variância do atraso, considerando uma porta de 8 entradas e $W_k = 500nm$. A contribuição de L^{nsc} e L^{sc} do transistor da entrada que tem transição são ordens de magnitude mais significativos que outros parâmetros.

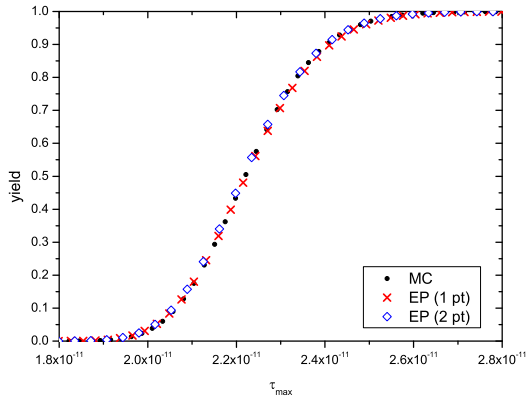

Figura 10. $p(t_0 < \tau_{max}, t_1 < \tau_{max} \ldots, t_{(7)} < \tau_{max})$ para uma NOR dinâmica de 8 entradas, calculada usando MC e EP com 1 ou 2 pontos para as derivadas.

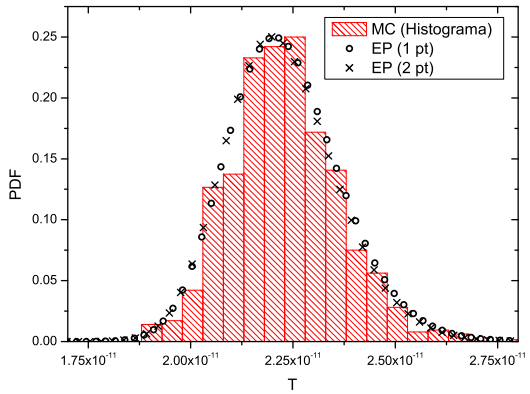

Figura 11. PDF de atraso de NOR dinâmica com 8 entradas e keeper estático ($W_k = 500nm$), calculando por MC e EP usando 1 ou 2 pontos para as derivadas.

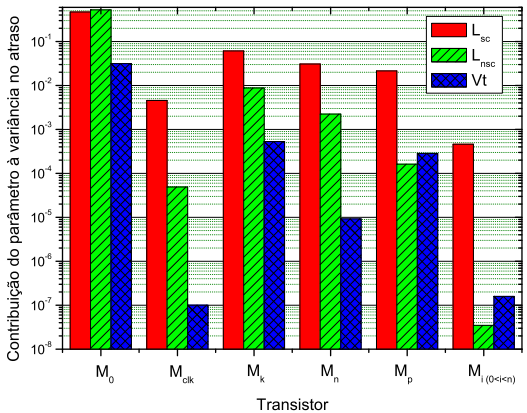

Figura 12. Contribuição de cada parâmetro para a variabilidade do atraso de uma NOR dinâmica com keeper estático

6 Conclusões

Este trabalho apresenta uma metodologia para análise estatística de blocos elétricos combinacionais e seqüenciais. A variância é calculada através de propagação de erro, e as derivadas são calculadas numericamente através dos valores obtidos por simulação elétrica. O emprego de derivadas numéricas confere generalidade a metodologia: as fórmulas para cálculo da variância e das derivadas são gerais.

A principal contribuição do trabalho é a elaboração de uma metodologia para cálculo de yield que apresenta qualidade semelhante a Monte Carlo, mas com ganho em desempenho. Nossos resultados reportam diferença de 1% entre o desvio padrão calculado usando a metodologia proposta comparado a Monte Carlo. Quanto a performance, reportamos ganho de desempenho de $50\times$ no caso da lógica dinâmica e $70\times$ no caso da memória SRAM.

O número de iterações utilizadas no Método Monte Carlo é função da ordem do erro esperado – e não do número de variáveis aleatórias – e portanto o número de iterações requeridos por este independe do tamanho do circuito. Como o número de simulações elétricas requeridas pela propagação de erros é função linear do número de variáveis aleatórias, esta apresenta ganhos de desempenho – comparativamente a MC – quando aplicada a caracterização de células (célula SRAM, multiplexador, portas complexas, ...). Contudo, propagação de erros e derivadas numéricas não apresenta vantagem de desempenho sobre MC se aplicada a

blocos com grande número de transistores.

Além disso, a metodologia permite quantificar a contribuição de cada parâmetro à variabilidade do circuito. A partir da análise da contribuição dos parâmetros à variância, é possível fazer uma etapa de re-projeto a fim de aumentar o *yield*.Nas simulações da memória SRAM, verificamos que o comprimento W do transistor é o parâmetro que mais contribui à variância. Na lógica dinâmica, verificamos que a largura (ambas as componentes SC e NSC) para o transistor da entrada que tem transição é o parâmetro que mais contribui para a variabilidade do atraso. Além disso, a partir de nossas simulações, verificamos a existência de um W ótimo para o keeper estático.

7 Referências

[1] A. Agarwal, B. C. Paul, H. Mahmoodi, A. Datta, and K. Roy. A process-tolerant cache architecture for improved yield in nanoscale technologies. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 13(1):27–38, 2005.

[2] A. Agarwal, B. C. Paul, and K. Roy. Process variation in nano-scale memories: failure analysis and process tolerant architecture. *Custom Integrated Circuits Conference, 2004. Proceedings of the IEEE 2004*, pages 353–356, 2004.

[3] A. Agarwal, B.C. Paul, S. Mukhopadhyay, and K. Roy. Process variation in embedded memories: failure analysis and variation aware architecture. *IEEE Journal of Solid-State Circuits*, 40(9):1804– 1814, September 2005.

[4] Jacques G. Amar. The monte carlo method in science and engineering. *Computing in Science and Engineering*, 8(2):9–19, 2006.

[5] Yang Byung-Do and Kim Lee-Sup. A low-power sram using hierarchical bit line and local sense amplifier. *IEEE Journal of Solid-State Circuits*, 40(6):1366– 1376, June 2005.

[6] Y. Cao, T. Sato, D. Sylvester, M. Orchansky, and C. Hu. New paradigm of predictive mosfet and interconnect modeling for early circuit design. In *Custom Integrated Circuit Conference*, pages 201–204, June 2000.

[7] Kunhyuk Kang, B. C. Paul, and K. Roy. Statistical timing analysis using levelized covariance propagation. *Design, Automation and Test in Europe, 2005. Proceedings*, pages 764–769 Vol. 2, 2005.

[8] H. Mahmoodi, S. Mukhopadhyay, and K. Roy. Estimation of delay variations due to random-dopant fluctuations in nanoscale cmos circuits. *Solid-State Circuits, IEEE Journal of*, 40(9):1787–1796, 2005.

[9] S. Mukhopadhyay, H. Mahmoodi, and K. Roy. Statistical design and optimization of sram cell for yield enhancement. In *ICCAD-2004. IEEE/ACM International Conference on Computer Aided Design, 2004.*, pages 10– 13, November 2004.

[10] S. Mukhopadhyay, H. Mahmoodi-Meimand, and K. Roy. Modeling and estimation of failure probability due to parameter variations in nano-scale srams for yield enhancement. In *2004 Symposium on VLSI Circuits Digest of Technical Papers*, pages 64–67, June 2004.

[11] S.R Nassif. Design for variability in dsm technologies [deep submicron technologies]. In *Quality Electronic Design, 2000. ISQED 2000. Proceedings. IEEE 2000 First International Symposium on*, pages 451–454, 2000.

[12] M. Orshansky, L. Milor, Pinhong Chen, K. Keutzer, and Chenming Hu. Impact of spatial intrachip gate length variability on the performance of high-speed digital circuits. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 21(5):544–553, 2002.

[13] L. G. Parrat. *Probability and Experimental Errors on Science*. John Wiley and Sons Inc, New York, NY, USA, 1961.

[14] Semiconductor Industry Association. *The International Technology Roadmap for Semiconductors*, 2005 Edition.

[15] Shang-Jyh Shieh, Jinn-Shyan Wang, and Yuan-Hsun Yeh. A contention-alleviated static keeper for high-performance domino logic circuits. *Electronics, Circuits and Systems, 2001. ICECS 2001. The 8th IEEE International Conference on*, 2:707–710 vol.2, 2001.

[16] Synopsys Inc. *HSPICE Simulation and Analysis User Guide*, 2005.

[17] Yuan Taur, D.A. Buchanan, Wei Chen, D.J. Frank, K.E. Ismail, Shih-Hsien Lo, G.A. Sai-Halasz, R.G. Viswanathan, H.-J.C. Wann, S.J. Wind, and Hon-Sum Wong. Cmos scaling into the nanometer regime. In *Proceedings of the IEEE*, volume 85, pages 486–504, Apr 1997.

[18] Yuan Taur and Tak H. Ning. *Fundamentals of modern VLSI devices*. Cambridge University Press, New York, NY, USA, 1998.

[19] P. S. Zuchowski, P. A. Habitz, J. D. Hayes, and J. H. Oppold. Process and environmental variation impacts on asic timing. In *Computer Aided Design, 2004. ICCAD-2004. IEEE/ACM International Conference on*, pages 336–342, 2004.