Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Εκτίμηση της Πιθανότητας Αποτυχίας FinFET SRAM κυττάρων υπό Στατική και Χρονικά Εξαρτώμενη Μεταβλητότητα

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ελένη Μ. Μαραγκουδάκη

Επιβλέπων : Δημήτριος Σούντρης
Αναπληρωτής Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2017

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Estimating the Failure Probability of FinFET-based SRAM Cells under Time-Zero and Time-Dependent Variability

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ελένη Μ. Μαραγκουδάκη

Επιβλέπων : Δημήτριος Σούντρης
Αναπληρωτής Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 16η Μαρτίου 2017.

Κιαμάλ Πεκμεστζή Καθηγητής Ε.Μ.Π.	Δημήτριος Σούντρης Αναπληρωτής Καθηγητής Ε.Μ.Π.	Ιωάννης Ξανθάκης Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2017

Ελένη Μ. Μαραγκουδάκη

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Στατική και χρονικά εξαρτώμενη μεταβλητότητα

Η όλο και αυξανόμενη ζήτηση για διεύρυνση των λειτουργιών των ολοκληρωμένων κυκλωμάτων έχει οδηγήσει στη ραγδαία συρρίκνωση των διαστάσεων των τρανζίστορ. Ο νόμος του Moore, ο οποίος υπέδειξε ότι ο αριθμός των τρανζίστορ ανά ολοκληρωμένο κύκλωμα θα διπλασιάζεται κάθε περίπου δύο χρόνια, έχει αποδειχθεί σωστός μέχρι σήμερα. Όμως, με αυτήν τη σμίκρυνση των διαστάσεων, το θέμα της αξιοπιστίας των ηλεκτρονικών συστημάτων έχει αναδειχθεί και έχει τραβήξει την προσοχή της επιστημονικής κοινότητας.

Η αξιοπιστία ορίζεται ως η πιθανότητα ενός συστήματος να λειτουργεί σωστά καθ' όλη τη διάρκεια 'ζωής' του [39]. Η αξιοπιστία επηρεάζεται από εσωτερικές αιτίες που υπήρχαν στο προϊόν εξαρχής αλλά και από εξωτερικούς παράγοντες, όπως η θερμοκρασία, η υγρασία, ο χρόνος. Όλα αυτά έχουν ως αποτέλεσμα τη μεταβλητότητα των παραμέτρων και συγκεκριμένα την αύξηση της απόλυτης τιμής της τάσης κατωφλίου, η οποία επιφέρει την αστάθεια του συστήματος. Από τη μία τα σταστικά φαινόμενα, όπως Random Dopant Fluctuations (RDF), προκαλούν την αρχική διασπορά των παραμέτρων [24] και από την άλλη η τάση κατωφλίου επηρεάζεται από χρονικά εξαρτώμενα φαινόμενα.

Η στατική μεταβλητότητα κατά τη διάρκεια της κατασκευαστικής διαδικασίας συνίσταται στο ότι η τάση κατωφλίου των τρανζίστορ δεν είναι σταθερή, αλλά διανέμεται και η Gaussian κατανομή είναι μία επαρκής προσέγγιση. Η μέση τιμή της μετατόπισης της τάσεως είναι μηδέν και η τυπική απόκλιση είναι ίση με:

$$\sigma_{V_{th,0}} = \frac{A_{VT}}{\sqrt{2WL}} \tag{0.1}$$

όπου A_{VT} είναι η παράμετρος του Pelgrom και W, L είναι το πλάτος και το μήκος της συσκευής αντίστοιχα. Στην περίπτωση των FinFET, το μήκος της πύλης αντιστοιχεί στο $LFIN$ που είναι το μήκος του fin και το πλάτος της πύλης είναι $W = 2HFIN + WFIN$ όπου $WFIN$ και $HFIN$ είναι το πλάτος και το ύψος του fin αντίστοιχα [25].

Ένας σημαντικός μηχανισμός που επιφέρει την χρονικά εξαρτώμενη υποβάθμιση ενός συστήματος είναι το φαινόμενο Bias Temperature Instability (BTI) [40]. Τα μοντέλα που το περιγράφουν είναι δύο, το reaction-diffusion (RD) και το ατομιστικό. Η προσέγγιση του RD μοντέλου επικεντρώνεται στα pFETs και στο σπάσιμο των δεσμών Si-H στην διεπαφή Si/oxide όταν βρίσκεται υπό τάση [38]. Τότε οι οπές αντικαθιστούν τα άτομα υδρογόνου ενώ αυτά διαχέονται στο gate-stack. Έτσι δημιουργούνται charge traps που οδηγούν στην μεταβολή της τάσης κατωφλίου. Επιπλέον, το μοντέλο αυτό προβλέπει τη φάση ανάκαμψης όταν τελειώνει ο χρόνος εφαρμογής τάσης στην πύλη, κατά την οποία οι Si-H δεσμοί αποκαθίστανται.

Το RD μοντέλο αποδείχθηκε ότι δεν μπορεί να συλλάβει το BTI φαινόμενο αρκετά αποδοτικά των σημερινών συρρικνωμένων συσκευών και γι' αυτό αναπτύχθηκε το πιο ακριβές ατομιστικό μοντέλο το οποίο δίνει έμφαση στα defects και την στοχαστική τους φύση, τα οποία προκαλούνται από το μηχανισμό charge traping [37],[40]. Σύμφωνα με αυτό το μοντέλο, μετά τη δημιουργία ενός defect, αυτό μπορεί να βρίσκεται σε δύο καταστάσεις, φορτισμένο ή μη ανάλογα με την εφαρμογή τάσης στην πύλη που δέχεται το τρανζίστορ. Μόνο τα defects που έχουν φορτίο μεταβάλουν την τάση κατωφλίου.

Σύμφωνα μ αυτήν τη θεωρία και τα πειραματικά δεδομένα [25], η μεταβολή της τάσεως κατωφλίου

ακολουθεί κανονική κατανομή με μέση τιμή:

$$\langle \Delta V_{th}(t) \rangle \cong A t^a E_{OX}^{\gamma} \tag{0.2}$$

όπου t ο χρόνος λειτουργίας, A ένας συντελεστής fitting, E_{OX} το ηλεκτρικό πεδίο κάθετα στο οξείδιο της πύλης και γ, a είναι εκθέτες επιτάχυνσης για το ηλεκτρικό πεδίο κάθετα στο οξείδιο της πύλης. Το E_{OX} υπολογίζεται ως $(V_G - V_{th})/T_{INV}$, όπου V_G είναι η τάση στην πύλη και T_{INV} το πάχος του inversion layer της πύλης. Όσον αφορά την τυπική απόκλιση, υπάρχει μία συσχέτιση μεταξύ της στατικής και της χρονικά εξαρτώμενης μεταβλητότητας [12] όπως περιγράφεται στην Εξίσωση 0.3. Έχει αποδειχθεί ότι υψηλότερη στατική μεταβλητότητα οδηγεί και στην αύξηση της χρονικά εξαρτώμενης μεταβλητότητας.

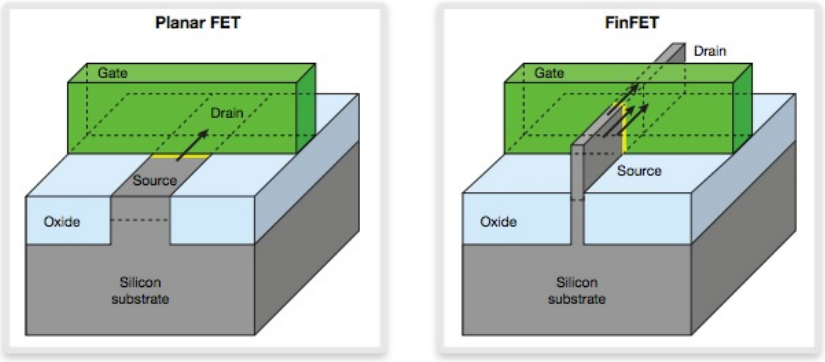
$$\sigma_{\Delta V_{th}}(t) = \sigma_{V_{th,0}} \sqrt{\frac{\langle \Delta V_{th}(t) \rangle}{100mV}} \tag{0.3}$$

Τέλος, η συνολική τυπική απόκλιση ισούται με:

$$\sigma_{V_{th,tot}}(t) = \sigma_{\Delta V_{th}}(t) + \sigma_{V_{th,0}} \qquad \sigma_{V_{th,tot}}(t) = \left(\sqrt{\frac{\langle \Delta V_{th}(t) \rangle}{100mV}} + 1 \right) \sigma_{V_{th,0}} \tag{0.4}$$

Σύγκριση FinFETs με planar FETs

Σήμερα, η περαιτέρω ελάττωση των διαστάσεων και η συνέχιση του νόμου του Moore αποτελεί μία πρόκληση. Η διαρκής σμίκρυνση των planar FETs μειώνει την αποδοτικότητά τους και οδηγεί στην αύξηση των ρευμάτων διαρροής και των short-channel effects (SCE) [18]. Γι' αυτόν τον λόγο, αναπτύχθηκαν αρχιτεκτονικές πολλαπλών πυλών. Η πιο διαδεδομένη είναι η τεχνολογία FinFET, που χρησιμοποιείται από 28nm και κάτω και στην οποία το κανάλι αγωγός τυλίγεται από ένα λεπτό "fin" πυριτίου. Το ρεύμα ρέει στην πάνω και στις πλάγιες επιφάνειες του fin όπως φαίνεται στην Εικόνα 0.1 [30].

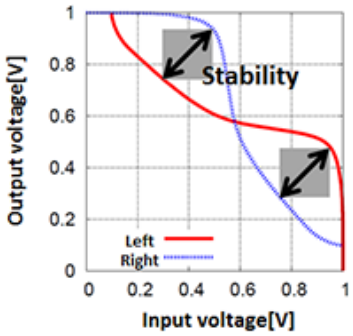


Εικόνα 0.1: Απλουστευμένη σύγκριση μεταξύ των planar FET και FinFET [30]

Η δομή του FinFET εκτός του ότι προσφέρει τη δυνατότητα μείωσης των διαστάσεων, λόγω του περιορισμού των ρευμάτων διαρροής και του SCE, έχει κι άλλα πλεονεκτήματα σε σχέση με τη δομή του planar FET. Μπορεί να λειτουργεί σε χαμηλότερη τάση τροφοδοσίας λόγω της μικρής τάσης κατωφλίου του, με αποτέλεσμα τη μείωση της κατανάλωσης ενέργειας. Το γεγονός αυτό και το ότι παρέχει μεγαλύτερη ταχύτητα καθιστά την τεχνολογία αυτή την πιο υποσχόμενη.

Η εισαγωγή των FinFET στα κύτταρα μνήμης Static Random Access Memory (SRAM) αυξάνουν τη δυσκολία της εκτίμησης της αξιοπιστίας τους. Κατά τη διάρκεια της λειτουργίας hold που το κύτταρο πρέπει να κρατάει το δεδομένο, η πτώση της τάσης τροφοδοσίας μπορεί να οδηγήσει στην καταστροφή του δεδομένου [23]. Για το λόγο αυτό, επιλέξαμε το κριτήριο για την αποτυχία του κυττάρου να βασίζεται στο Static Noise Margin (SNM) για τη λειτουργία hold, το οποίο εκφράζει το μέγιστο θόρυβο

που μπορεί να αντέξει το κύτταρο ενώ διατηρείται η σωστή λειτουργία του. Το SNM υπολογίζεται γραφικά μέσω του butterfly curve. Ορίζεται ως το μήκος της διαγωνίου του μικρότερο από τα δύο τετράγωνα που μπορούν να ενσωματωθούν μέσα στους δύο λοβούς, όπως φαίνεται στην Εικόνα 0.2 [32].



Εικόνα 0.2: Butterfly curve [32].

Εκτίμηση της αξιοπιστίας

Για τον υπολογισμό της αξιοπιστίας ενός συστήματος έχουν αναπτυχθεί διάφορες μετρικές, οι πιο διαδεδομένες ορίζονται παρακάτω.

Definition 0.0.1. Yield ορίζεται ως ο αριθμός των λειτουργικών κυκλωμάτων προς το συνολικό αριθμό των κυκλωμάτων που εξετάσθηκαν [41].

$$\text{Yield} = \frac{\text{Functional circuits}}{\text{Total examined circuits}}$$

Definition 0.0.2. Defects Per Million (DPM) είναι ο αριθμός των defects προς το συνολικό αριθμό των κυκλωμάτων που εξετάσθηκαν και αποτελεί άλλον έναν τρόπο μέτρησης της απόδοσης. Σε σχετικές μελέτες αξιοπιστίας, defect θεωρείται ένα δυσλειτουργικό κύκλωμα.

$$\text{DPM} = \frac{\text{Number of defects}}{\text{Total examined circuits}} \cdot 10^6$$

Definition 0.0.3. Mean Time To Failure (MTTF) αντιπροσωπεύει τη μέση τιμή του χρόνου λειτουργίας μέχρι την εμφάνιση κάποιας αστοχίας [43].

$$\text{MTTF} = \frac{\text{Operating hours}}{\text{Number of Failures}}$$

Definition 0.0.4. Failure rate ισούται με το πλήθος των αστοχιών του συστήματος κατά τη διάρκεια ζωής του, κανονικοποιημένο σε 1 δισεκατομύριο ώρες συσκευής και συνδέεται με το MTTF όπως περιγράφεται παρακάτω [43].

$$\text{Failure Rate} = \frac{10^9}{\text{MTTF}}$$

Στην εργασία αυτή επικεντρωνόμαστε στην πιθανότητα αποτυχίας P_{FAIL} για να εκτιμήσουμε την αξιοπιστία ενός κυττάρου SRAM, η οποία ορίζεται ως η πιθανότητα εμφάνισης αστοχίας σε ένα δεδομένο χρονικό διάστημα. Επειδή η πιθανότητα της μόνιμης αποτυχίας ενός κυττάρου μνήμης είναι πολύ μικρή, η ακριβής εκτίμησή της είναι πολύ περίπλοκη. Παρόλ' αυτά, λόγω του ότι η μετρική αυτή μπορεί να συμπεριλάβει όλα τα φαινόμενα που παράγουν τη μεταβλητότητα της τάσεως κατωφλίου, διάφορες μέθοδοι έχουν αναπτυχθεί για τον υπολογισμό της όπως οι Monte Carlo, Quasi Monte Carlo, Importance Sampling και Most Probable Failure Point.

Η γνωστή μέθοδος Monte Carlo έχει χρησιμοποιηθεί πολλές φορές για τον υπολογισμό του P_{FAIL} κυττάρων SRAM. Αποτελείται από τα παρακάτω στάδια [16]: πρώτα γίνεται η επιλογή N δειγμάτων, $x_1, x_2, \ldots, x_N$, στον εξεταζόμενο χώρο σύμφωνα με μία κατανομή, συνήθως την κανονική [9], έπειτα, η συνάρτηση I υπολογίζεται, η οποία υποδεικνύει αν ένα δείγμα αποτυγχάνει σύμφωνα με το κριτήριο που τέθηκε και τελικά το P_{FAIL} προσεγγίζεται ως:

$$P_{FAIL} = \frac{1}{N}\sum_{i=1}^{N} I(x_i) \tag{0.5}$$

Το μειονέκτημα αυτής της μεθόδου είναι το απαγορευτικό πλήθος των υπολογισμών που χρειάζονται για τον υπολογισμό μιας τόσο μικρής πιθανότητας.

Σε αντίθεση με τη συνηθισμένη Monte Carlo μέθοδο, η τεχνική Quasi-Monte Carlo χρησιμοποιεί low-discrepancy ακολουθίες, όπως οι ακολουθίες Halton, Sobol ή Faure, οι οποίες οδηγούν στην αύξηση της ταχύτητας σύγκλισης. Παρόλο που οι προσεγγίσεις του P_{FAIL} βασισμένες σε αυτήν τη μεθοδολογία είναι ικανοποιητικές για παλαιότερες τεχνολογίες, όταν πρόκειται για τεχνολογίες των σημερινών διαστάσεων στερούνται επαρκούς ακρίβειας [10].

Η μέθοδος Importance Sampling στοχεύει στη μείωση των προσομοιώσεων της τεχνικής Monte Carlo, επιλέγοντας τα δείγματα με βάση μία διαφορετική κατανομή, η οποία οδηγεί στην αύξηση των δειγμάτων που θεωρείται ότι απέτυχαν [4]. Το P_{FAIL} υπολογίζεται όπως φαίνεται παρακάτω:

$$P_{FAIL} = \frac{1}{N}\sum_{i=1}^{N} I(x_i)\frac{h(x_i)}{g(x_i)}$$

Επειδή τα περισσότερα από τα δείγματα της Monte Carlo βρίσκονται κοντά στη μέση τιμή, μία συνηθισμένη τακτική είναι η μετακίνηση της κατανομής προς την περιοχή των δειγμάτων που οδηγούν σε αποτυχία. Μία άλλη προσέγγιση είναι η παραγωγή των δειγμάτων με βάση μία μίξη κατανομών [2]. Η βασική ιδέα όλων αυτών των μεθόδων είναι η δειγματοληψία περισσότερων περιπτώσεων που προκαλούν αστοχία. Η επιλογή της εναλλακτικής κατανομής είναι κρίσιμη καθώς μπορεί να οδηγήσει σε ανακριβή υπολογισμό της πιθανότητας, η οποία συνήθως καταλήγει να είναι πιο απαισιόδοξη απ' ότι στην πραγματικότητα.

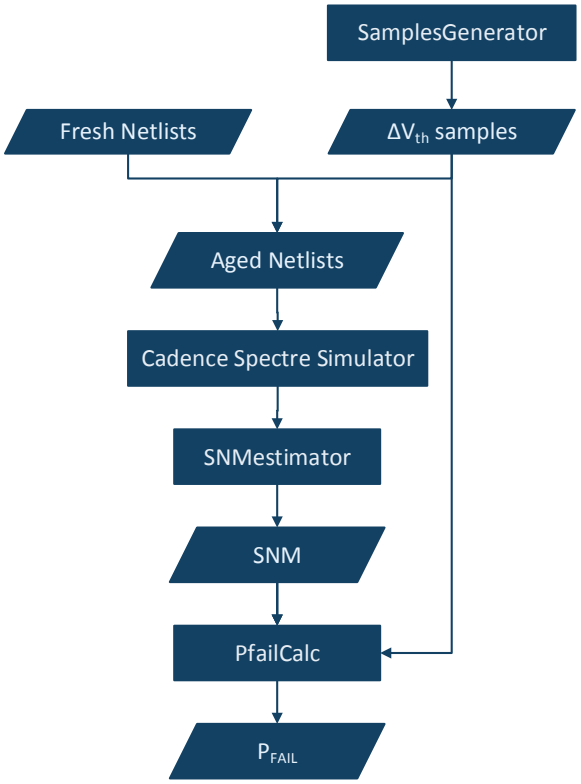
Η μέθοδος MPFP [26],[23],[29],[4],[11] αποτελεί μία πιο ακριβή τεχνική προσέγγισης του P_{FAIL}. Η ιδέα στην οποία στηρίζεται, είναι ο εντοπισμός του σημείου με τη μεγαλύτερη πιθανότητα εμφάνισης που οδηγεί στην αποτυχία του συστήματος.

Το κύριο πλεονέκτημα αυτής της μεθόδου είναι ότι δεν απαιτείται η προσέγγιση της κατανομής του κριτηρίου αποτυχίας. Σε άλλες τεχνικές, όπως οι Monte Carlo και Quasi Monte Carlo, θεωρείται ότι το SNM ακολουθεί κανονική κατανομή. Αυτή η υπόθεση αυξάνει την ανακρίβεια των αποτελεσμάτων τους [26].

Η καθιερωμένη μεθοδολογία MPFP

Η παράμετρος που κατά κύριο λόγο προκαλεί αστοχίες του συστήματος είναι η τάση κατωφλίου [23] και επειδή οι αποτυχίες θεωρούνται σπάνιες, σκοπός της μεθόδου αυτής είναι η αναγνώριση του συνδυασμού των τάσεων κατωφλίου των συσκευών που οδηγεί στην αποτυχία και έχει τη μέγιστη πιθανότητα εμφάνισης. Θεωρώντας $\mathbf{x} = [x_1, x_2, ..., x_N]$ το συνδυασμό των ΔV_{th} των N εμπλεκόμενων τρανζίστορ που οδηγεί στην αποτυχία, το P_{FAIL} υπολογίζεται όπως φαίνεται στην Εξίσωση 0.6. Σύμφωνα με αυτήν, η πιθανότητα αποτυχίας για κάθε συσκευή αντιστοιχεί στο διάστημα όπου η μεταβολή της τάσεως κατωφλίου είναι μεγαλύτερη ή ίση με αυτή του σημείου MPFP. Η διαδικασία που ακολουθήσαμε για τον υπολογισμό της πιθανότητας φαίνεται στην Εικόνα 0.3

$$P_{FAIL} = \max\left\{\prod_{i=1}^{N} P(|\Delta V_{th,i}| \geq x_i)\right\} \tag{0.6}$$



Εικόνα 0.3: Η διαδικασία υπολογισμού του P_{FAIL} χρησιμοποιώντας τη μέθοδο MPFP.

Για να υπολογίσουμε την πιθανότητα $P(|\Delta V_{th}| \geq x_i)$ που αντιστοιχεί σε κάθε x_i, η κατανομή του ΔV_{th} απαιτείται. Όπως αναφέραμε, και η στατική και η χρονικά εξαρτώμενη μεταβλητότητα μοντελοποιούνται με την κανονική κατανομή. Η μέση τιμή, μ, και η τυπική απόκλιση, σ, υπολογίζονται από τις Εξισώσεις 0.2 and 0.4 αντίστοιχα. Επομένως, και τα pFETs και τα nFETs ακολουθούν κανονικές κατανομές: $\Delta V_{th,p} \sim \mathcal{N}(\mu_p, \sigma_p^2)$, $\Delta V_{th,n} \sim \mathcal{N}(\mu_n, \sigma_n^2)$.

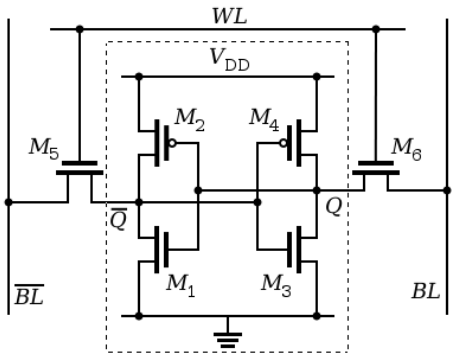
Για να βρούμε το MPFP σημείο, είναι απαραίτητη η επιλογή του κριτηρίου αποτυχίας. Για να απλοποιήσουμε το πρόβλημα χρησιμοποιούμε το SNM για τη λειτουργία hold και επικεντρωνόμαστε στα τέσσερα τρανζίστορ των cross-coupled inverters, όπως τονίζεται στην Εικόνα 0.4. Ο θόρυβος thermal noise αποτελεί την επικρατέστερη αιτία θορύβου για τα κύτταρα μνήμης SRAM και αυξάνεται με τη μείωση των διαστάσεων των τρανζίστορ. Συγκεκριμένα, για τις διαστάσεις που μελετάμε η διακύμανση του thermal noise έχει μετρηθεί εώς και 25mV [7]. Λαμβάνοντας υπόψιν αυτήν την τιμή, επιλέξαμε το SNM specification (threshold), Y, έτσι ώστε να διασφαλίζει την ανοχή του κυττάρου σε thermal noise. Επομένως, θεωρούμε ότι ένα δείγμα αποτυγχάνει όταν:

$$SNM(\mathbf{x}) < Y \tag{0.7}$$

Τα εργαλεία που χρησιμοποιήσαμε για την υλοποίηση της μεθόδου MPFP αναλύονται παρακάτω.

- Το πρώτο είναι το `SamplesGenerator` και παράγει τα ΔV_{th} δείγματα, $\mathbf{x}$, των εμπλεκόμενων τρανζίστορ. Για να μειώσουμε τον απαιτούμενο χρόνο των προσομοιώσεων, περιορίσαμε τα ΔV_{th} μεταξύ -0.45V και 0.45V. Το εργαλείο δέχεται το βήμα από το χρήστη, το οποίο επιλέξαμε ίσο με 0.05V. Μια πιο λεπτομερής ανάλυση θα μπορούσε να επιτευχθεί με μικρότερο βήμα, όμως, θα αυξανόταν πολύ ο απαιτούμενος χρόνος.

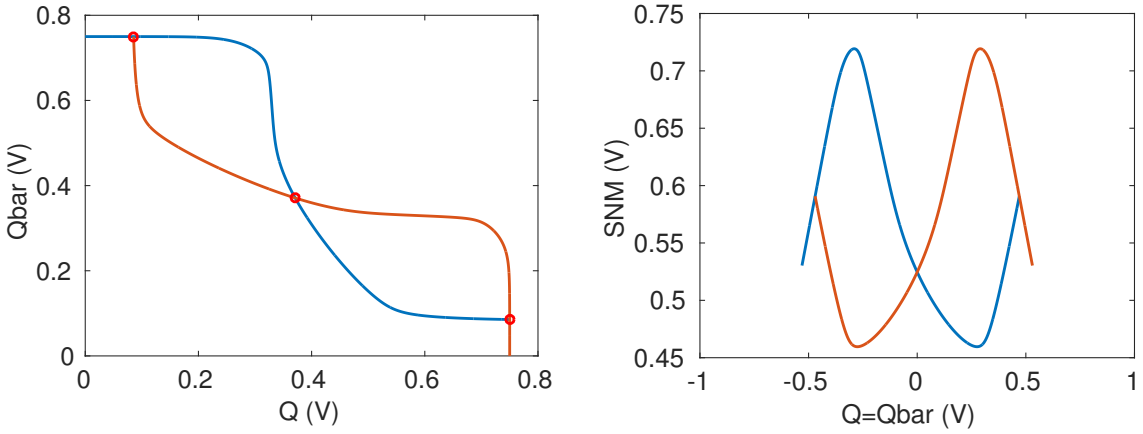
- Το `Cadence Spectre Simulator` [36] επεξεργάζεται το aged SRAM cell netlist, το οποίο παράχθηκε από το εργαλείο `SamplesGenerator`. Για το netlist χρησιμοποιήσαμε High Power



Εικόνα 0.4: SRAM cell circuit. In this thesis we focus on the four transistors as marked above.

(HP), PTM-MG modelfiles [35]. Το εργαλείο αυτό δημιουργεί το butterfly curve της κάθε μιας περίπτωσης, από το οποίο υπολογίζεται το SNM εκτελώντας δύο DC αναλύσεις. Σε κάθε μία, η τιμή του ενός από τους δύο internal storage κόμβους, Q ή $\bar{Q}$, αλλάζει και εξετάζεται η άλλη.

- Το επόμενο εργαλείο ονομάζεται `SNMestimator` και χρησιμοποιείται για την εκτίμηση του SNM για τη λειτουργία hold. Για να υπολογίσουμε την τιμή του SNM εκτελούμε μία πιο αποδοτική μέθοδο [6]. Σύμφωνα με αυτήν, περιστρέφουμε τις καμπύλες κατά 45° και αφαιρούμε την τιμή του ενός κόμβου από την τιμή του άλλου κι έτσι παράγεται μία καινούργια καμπύλη. Η μέγιστη και η ελάχιστη τιμή αντιπροσωπεύουν τις διαγώνιους που εφαρμόζονται στους λοβούς του butterfly curve [10]. Ένα αρχικό butterfly curve και ένα που έχει περιστραφεί απεικονίζονται παρακάτω.



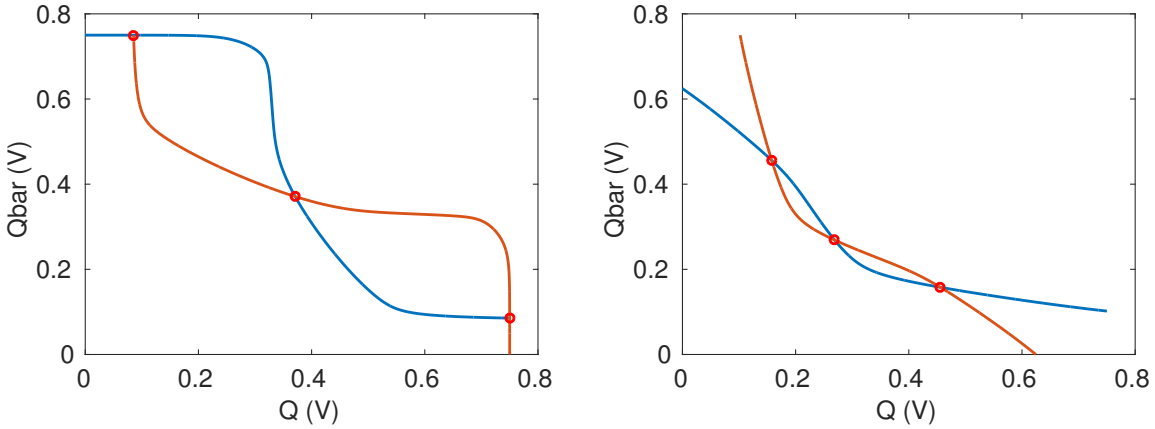
(i) Αρχικό butterfly curve

(ii) Μετά από περιστροφή 45°

Εικόνα 0.5: Η διαδικασία υπολογισμού του SNM

- Το τελευταίο εργαλείο είναι το `PfailCalc`, το οποίο αντιστοιχίζει τα ΔV_{th} δείγματα με τα SNM, αποφασίζει αν ένα δείγμα περνάει ή όχι το κριτήριο αποτυχίας και υπολογίζει το P_{FAIL}. Στην Εικόνα 0.6 φαίνεται η διαφορά του butterfly curve ενός δείγματος που αποτυγχάνει με ένα που επιτυγχάνει. Επειδή το P_{FAIL} ισούται με τη μέγιστη πιθανότητα, αρχικά το εργαλείο υπολογίζει την πιθανότητα εμφάνισης κάθε δείγματος που δεν περνάει το κριτήριο χρησιμοποιώντας την κανονική κατανομή και υπολογίζει τη μέση τιμή και την τυπική απόκλιση των nFET και pFET όπως περιγράφεται στις Εξισώσεις 0.2 και 0.4. Τέλος το P_{FAIL} εκτιμάται όπως φαίνεται στην Εξίσωση 0.6.

Είναι εμφανές ότι η πιθανότητα αποτυχίας επηρεάζεται πολύ από το κριτήριο αποτυχίας. Για να αντέχει το κύτταρο τον υψηλό θόρυβο θα πρέπει η τιμή του SNM spec να αυξηθεί, γεγονός που θα



(i) Περίπτωση που περνάει το κριτήριο αποτυχίας. **(ii)** Περίπτωση που δεν περνάει το κριτήριο αποτυχίας.

Εικόνα 0.6: Τα butterfly curves δύο διαφορετικών περιπτώσεων.

προκαλούσε την αύξηση του P_{FAIL}. Έτσι συμπεραίνουμε ότι στην περίπτωση μας το P_{FAIL} εξαρτάται άμεσα από την επιλογή της τιμής του Y.

Μελετήσαμε την πιθανότητα αποτυχίας του κυττάρου έπειτα από κατά προσέγγιση τρία χρόνια λειτουργίας. Τα αποτελέσματα της μέσης τιμής και της τυπικής απόκλισης των nFET και pFET μαζί με την παράμετρο του Pelgrom και κάποιες άλλες παραμέτρους φαίνονται στον Πίνακα 0.1. Η τιμή του P_{FAIL} καθώς και το σημείο με την μέγιστη πιθανότητα εμφάνισης απεικονίζονται στον Πίνακα 0.2.

	nFET	pFET
$A_{VT}[mV.\mu V]$	1.0	1.0
$\sigma_{V_{th,0}}[mV]$	26.46	26.46
$V_{DD}[V]$	0.75	0.75
$\langle \Delta V_{th} \rangle [mV]$	130.52	45.55
$\sigma_{V_{th,tot}}[mV]$	40.18	31.93

Table 0.1: Η μέση τιμή και η τυπική απόκλιση των nFET και pFET.

Devices	1	2	3	4
$\Delta V_{th}[V]$	0.4	0	0	0
P_{FAIL}	1×10^{-11}			

Table 0.2: Το σημείο MPFP και τοP_{FAIL}

Στον Πίνακα 0.1, είναι εμφανείς οι διαφορές της μεταβλητότητας των V_{th} των n-type και p-type FinFET. Ενώ και τα δύο παρουσιάζουν ίση στατική μεταβλητότητα, η απόκλισή τους σε χρονικά εξαρτώμενη είναι αισθητή. Συγκεκριμένα, τα nFETs φαίνεται να είναι πιο επιρρεπή στον υποβιβασμό που προκαλείται από το BTI. Επιπλέον, το τρανζίστορ M1, όπως φαίνεται στην Εικόνα 0.4, παίζει καθοριστικό ρόλο στην πιθανότητα αποτυχίας (Πίνακας 0.2). Το γεγονός αυτό είναι σε συμφωνία με τα αποτελέσματα προηγούμενης έρευνας [26], η οποία υπέδειξε ότι η πιθανότητα αποτυχίας της μίας κατεύθυνσης είναι συνήθως κυρίαρχη.

Μία ανασύνθεση της μεθόδου MPFP χρησιμοποιώντας την κατανομή χ^2

Παρόλο που η προηγούμενη τεχνική εντοπίζει σωστά το σημείο MPFP, δεν απομονώνει το σύνολο **F** των δειγμάτουν που οδηγούν σε αποτυχία με ακρίβεια. Το P_{FAIL} που υπολογίστηκε στο σημείο MPFP $\mathbf{x} = [x_1, x_2, ..., x_N]$ σύμφωνα με την προηγούμενη μεθοδολογία, αντιστοιχεί στην περιοχή: $|\Delta V_{th,i}| \geq x_i$, η οποία περιλαμβάνει μόνο ένα μέρος των περιπτώσεων που οδηγούν στην αποτυχία, επομένως, το αποτέλεσμα είναι πιο αισιόδοξο. Για αυτόν το λόγο, ακολουθούμε μία διαφορετική προσέγγιση της MPFP μεθόδου βασιζόμενοι σε προηγούμενη δουλειά [33] όπου προτείνεται ένας πιο ακριβής υπολογισμός του συνόλου **F**.

Το πρώτο βήμα αυτής της υλοποίησης είναι η εξέταση της κυρτότητας του χώρου, εντοπίζοντας ένα συνδυασμό ΔV_{th} των τρανζίστορ που οδηγεί σε ακρότατο του χώρου SNM. Ο έλεγχος της κυρτότητας του χώρου είναι καθοριστικό βήμα, καθώς η προσέγγιση του χώρου **F** βασίζεται σε αυτήν την υπόθεση. Παρόλο που η μαθηματική απόδειξη είναι αδύνατη, αφού τα αποτελέσματα του SNM προέκυψαν από προσομοιώσεις και όχι από μαθηματικό τύπο, η ύπαρξη ενός τοπικού ακρότατου είναι μία ένδειξη ότι ο χώρος είναι όντως κυρτός. Δεδομένου ότι η τιμή του SNM αναμένεται να μειώνεται με την μετατόπιση των V_{th}, επιλέξαμε να επικεντρωθούμε στην αναγνώριση ενός τοπικού μέγιστου ώστε να ελαττωθεί η πολυπλοκότητα του υπολογισμού του P_{FAIL}.

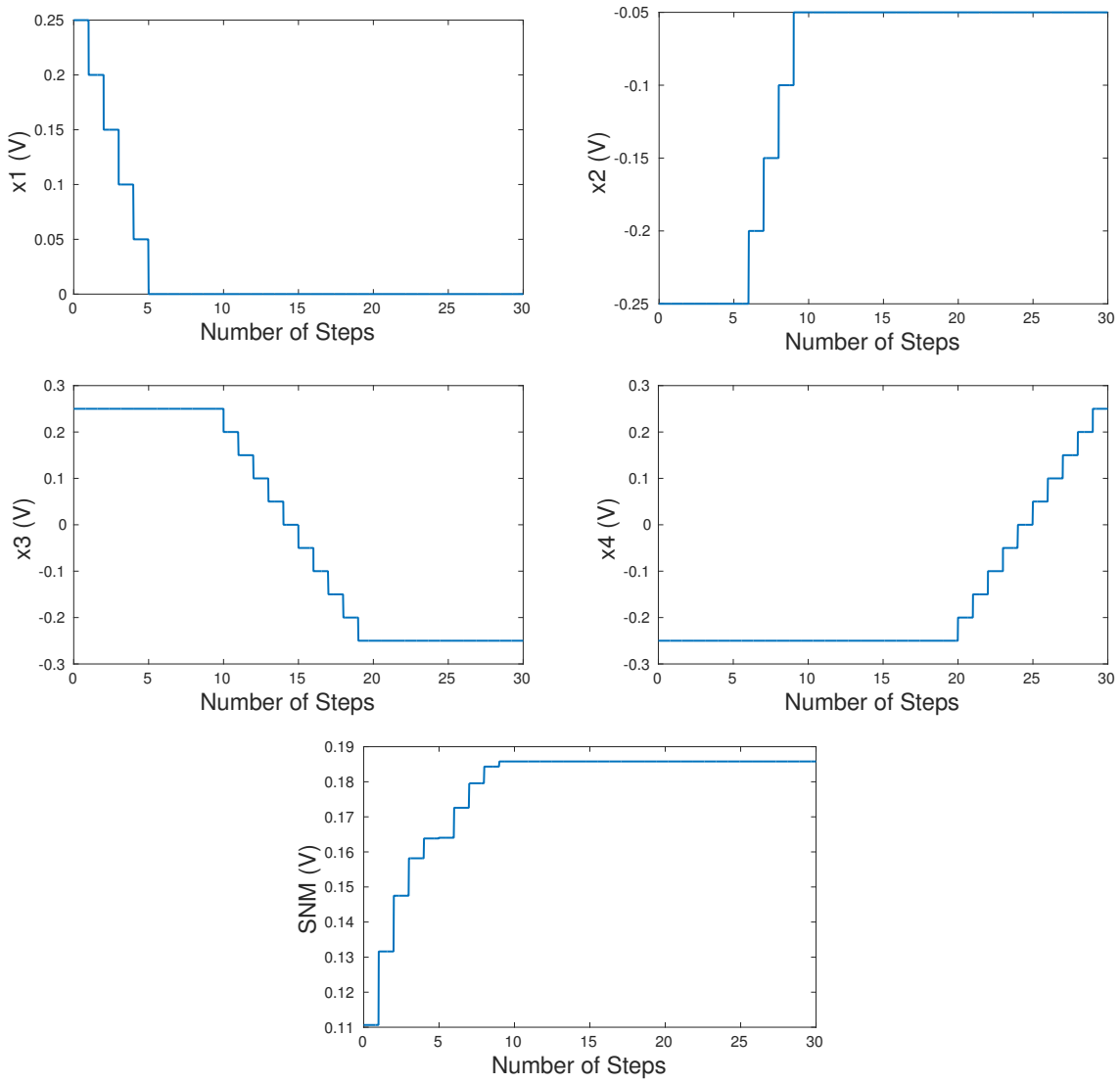
Έπειτα, συνεχίζουμε με τον εντοπισμό του ΔV_{th} συνδυασμού, $\mathbf{x}_Y$, με την μικρότερη απόσταση από το μέγιστο σημείο, r_Y, το οποίο οδηγεί στην αποτυχία του κυττάρου, δηλαδή $SNM(\mathbf{x}) < Y$. Για να το πετύχουμε αυτό, θέτουμε το μέγιστο ως σημείο εκκίνησης και κινούμαστε κάθε φορά προς την κατεύθυνση με τη μεγαλύτερη μείωση της τιμής του SNM μέχρι να φτάσουμε σε ένα ελάχιστο. Αφού αναγνωρίσαμε το πιο κοντινό σημείο στο μέγιστο που οδηγεί σε αποτυχία, είναι εμφανές ότι όλα τα σημεία που θεωρούνται αποτυχημένα βρίσκονται έξω από το hypersphere με κέντρο το μέγιστο σημείο και ακτίνα ίση με την απόσταση r_Y. Συνεπώς, αυτή η μέθοδος εκτιμά με περισσότερη ακρίβεια το χώρο **F** και προσφέρει ένα πιο ρεαλιστικό αποτέλεσμα.

Η κυρτότητα του χώρου SNM

Για να εντοπίσουμε ένα μέγιστο υλοποιούμε τον coordinate ascent Αλγόριθμο 1 [33]. Σύμφωνα με αυτόν, για κάθε τρανζίστορ, επιλέγουμε ένα θετικό ή αρνητικό βήμα, αναλόγως προς ποια κατεύθυνση αυξάνεται το SNM και το προσθέτουμε στην τιμή του ΔV_{th} που επεξεργαζόμαστε κάθε φορά, μέχρι το σημείο που το SNM είναι μικρότερο από το προηγούμενο. Η διαδικασία αυτή επαναλαμβάνεται μέχρις ότου να μην υπάρχει προσκείμενο σημείο με μεγαλύτερο SNM, δηλαδή μέχρι να εντοπίσουμε ένα μέγιστο. Σε περίπτωση που επιλέγαμε να αναγνωρίσουμε ένα τοπικό ελάχιστο θα εκτελούσουμε τον coordinate descent αλγόριθμο και κάθε φορά θα μετακινούμασταν προς την κατεύθυνση με το μικρότερο SNM. Επιλέξαμε βήμα ίσο με 0.05V και περιορίσαμε τα ΔV_{th} μεταξύ -0.25V και 0.25V. Οι τιμές του SNM υπολογίστηκαν προηγουμένως με τα εργαλεία `SamplesGenerator`, `Cadence Spectre Simulator` και `SNMestimator`. Τα αποτελέσματα του Αλγόριθμου 1 παραθέτονται στην Εικόνα 4.1 βήμα προς βήμα.

Algorithm 1 Coordinate Ascent

```
while maximum not found do
    for i=1 to N do
        step=find ascenting direction of SNM for x_i
        repeat
            if currSNM(x_1, ..., x_i + step, ..., x_N) > preSNM then
                update x_i with x_i + step
            end if
        until currSNM < preSNM
    end for
end while
```



Εικόνα 0.7: Τα αποτελέσματα της υλοποίησης του coordinate ascent Αλγόριθμου 1

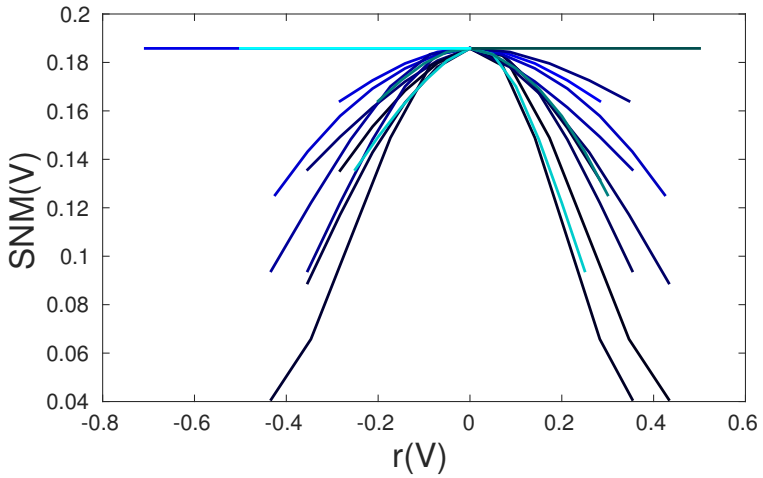
Για την ακρίβεια της μεθόδου, είναι σημαντικός ο εντοπισμός όλων των μέγιστων σημείων. Έπειτα από σύγκριση όλων των σημείων, καταλήξαμε ότι το σημείο που προηγουμένως αναγνωρίσαμε:

$$(\Delta V_{th,1}, \Delta V_{th,2}, \Delta V_{th,3}, \Delta V_{th,4}) = (0, -0.05, -0.25, 0.25)$$

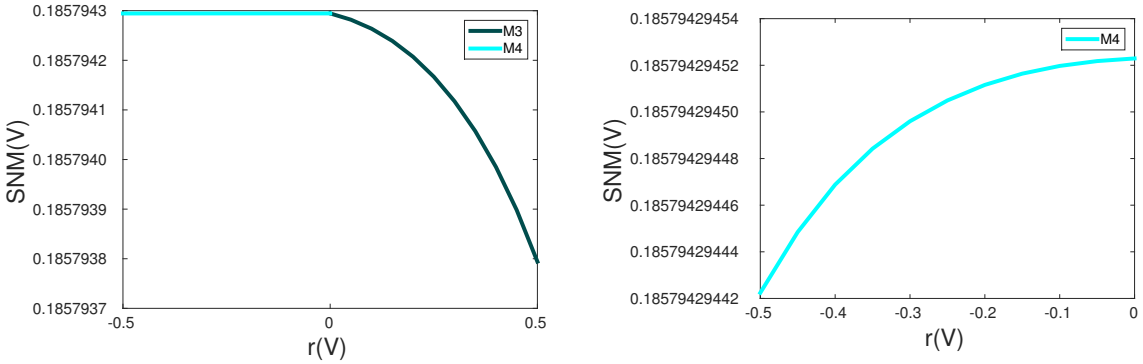
είναι το ολικό μέγιστο SNM στο χώρο που εξετάζουμε. Για την έρευνα του χώρου αναπτύξαμε ένα εργαλείο που σύγκρινε την τιμή του SNM του σημείου αυτού με τα SNM όλων των άλλων σημείων του χώρου και έδειξε ότι το μέγιστο που εντοπίστηκε είναι το ολικό μέγιστο. Παρόλο που υποθέσαμε ότι ο χώρος είναι συμμετρικός και ότι το συμμετρικό σημείο $(0, 0.05, 0.25, -0.25)$ αποτελεί κι αυτό μέγιστο, αποδείχθηκε ότι η σκέψη αυτή είναι λάθος καθώς υπάρχουν προσκείμενα σε αυτό σημείο με μεγαλύτερη τιμή SNM.

Στην Εικόνα 0.8 φαίνεται η αναπαράσταση του χώρου SNM κοντά στο μέγιστο. Οι καμπύλες απεικονίζουν το SNM καθώς μετακινούμαστε μακριά από το μέγιστο σε διαφορετικές κατευθύνσεις. Η παράμετρος r είναι η απόσταση των σημείων από το μέγιστο. Παρατηρούμε ότι δεν έχουν όλες οι καμπύλες την ίδια τάση. Στην πραγματικότητα, καθώς οι τάσεις κατωφλίου της συσκευής M3 και ειδικότερα της M4 απομακρύνονται από το μέγιστο, η μετατόπιση του SNM είναι ασήμαντη σε σχέση με τη ραγδαία μεταβολή που προκαλούν τα τρανζίστορ M1 και M2. Συνεπώς, όταν οι καμπύλες των $\Delta V_{th,3}$ και $\Delta V_{th,4}$ κατευθύνσεων απεικονίζονται στην ίδια κλίμακα με αυτές των δύο πρώτων

συσκευών, μοιάζουν οριζόντιες. Για το λόγο αυτό, η Εικόνα 0.9 είναι απαραίτητη για μία πιο λεπτομερή απεικόνιση του SNM. Έτσι, όλες οι καμπύλες επαληθεύουν ότι το σημείο που εντοπίσαμε είναι όντως το μέγιστο και ότι ο χώρος είναι κυρτός.



Εικόνα 0.8: Αναπαράσταση του χώρου του SNM γύρω από το μέγιστο.



Εικόνα 0.9: Αναπαράσταση του SNM στην κατεύθυνση των $\Delta V_{th,3}$ και $\Delta V_{th,4}$ παραμέτρων.

Εκτίμηση της πιθανότητας αποτυχίας

Αφού εντοπίσαμε το μοναδικό μέγιστο του SNM στο χώρο, προχωράμε στην εύρεση του σημείου $\mathbf{x}_Y$ με τη μικρότερη απόσταση από το μέγιστο, r_Y, το οποίο οδηγεί στην αποτυχία του κυττάρου. Αρχικοποιούμε τα ΔV_{th} στις τιμές που αντιστοιχούν στο μέγιστο και κάθε φορά κινούμαστε προς την κατεύθυνση με τη μεγαλύτερη πτώση της τιμής του SNM μέχρι να βρούμε ένα τοπικό ελάχιστο, σύμφωνα με τον Αλγόριθμο 2. Η Εικόνα 0.10 αποτυπώνει την εφαρμογή του αλγόριθμου. Οφείλει να σημειωθεί ότι αυτός ο αλγόριθμος ακολουθεί διαφορετική λογική από τον προηγούμενο. Αντί να ερευνά μία συντεταγμένη σε κάθε επανάληψη, τις εξετάζει όλες και έπειτα προχωρά στο σημείο με το χαμηλότερο SNM.

Έπειτα από τον εντοπισμό της κατεύθυνσης με την πιο ραγδαία πτώση του SNM, χρησιμοποιούμε την (χ^2) κατανομή για να προσεγγίσουμε την τιμή του P_{FAIL} λόγω του ότι αυτή η κατανομή απλοποιεί τους υπολογισμούς. Σύμφωνα με αυτήν, υπολογίζουμε την πιθανότητα της τυχαίας μεταβλητής z^2, όπως φαίνεται στην Εξίσωση 0.8 [34], όπου N ο αριθμός των τρανζίστορ και σ_i η τυπική απόκλιση του καθενός, όπως υπολογίστηκαν προηγουμένως. Η παράμετρος non-centrality αντιστοιχεί στο λ και

Algorithm 2 Greatest Descent

Initialize **x** to the value that leads to maximum
while not reach minimum **do**
  for i=1 to N **do**
    calculate $SNM_i(\dots, x_i - step, \dots)$
    calculate $SNM_i(\dots, x_i + step, \dots)$
    calculate $SNM_i(\dots, x_i, \dots)$
  end for
  for i=1 to N **do**
    find the minimum SNM_i
    update x_i with the value that leads to minimum
  end for
end while

μ είναι το μέγιστο σημείο, το οποίο εμπεριέχει και τη μέση τιμή του κάθε τρανζίστορ.

$$z^2 = \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} \qquad \lambda = \sum_{i=1}^{N} \frac{\mu_i^2}{\sigma_i^2} \tag{0.8}$$

Σε περίπτωση που εντοπίζαμε ελάχιστο αντί για μέγιστο, θα υλοποιούσαμε τον αλγόριθμο greatest ascent και αναλόγως, θα κινούμασταν προς την κατεύθυνση με την μεγαλύτερη αύξηση του SNM μέχρι να φτάσουμε στο σημείο x'_Y, με απόσταση από το ελάχιστο r'_Y το οποίο δεν οδηγεί στην αποτυχία του κυττάρου. Θα χρησιμοποιούσαμε ξανά την χ^2 κατανομή για την εκτίμηση της πιθανότητας, αλλά αυτή τη φορά το σύνολο **F** θα βρισκόταν μέσα στο hypersphere με κέντρο το ελάχιστο και ακτίνα ίση με r'_Y.

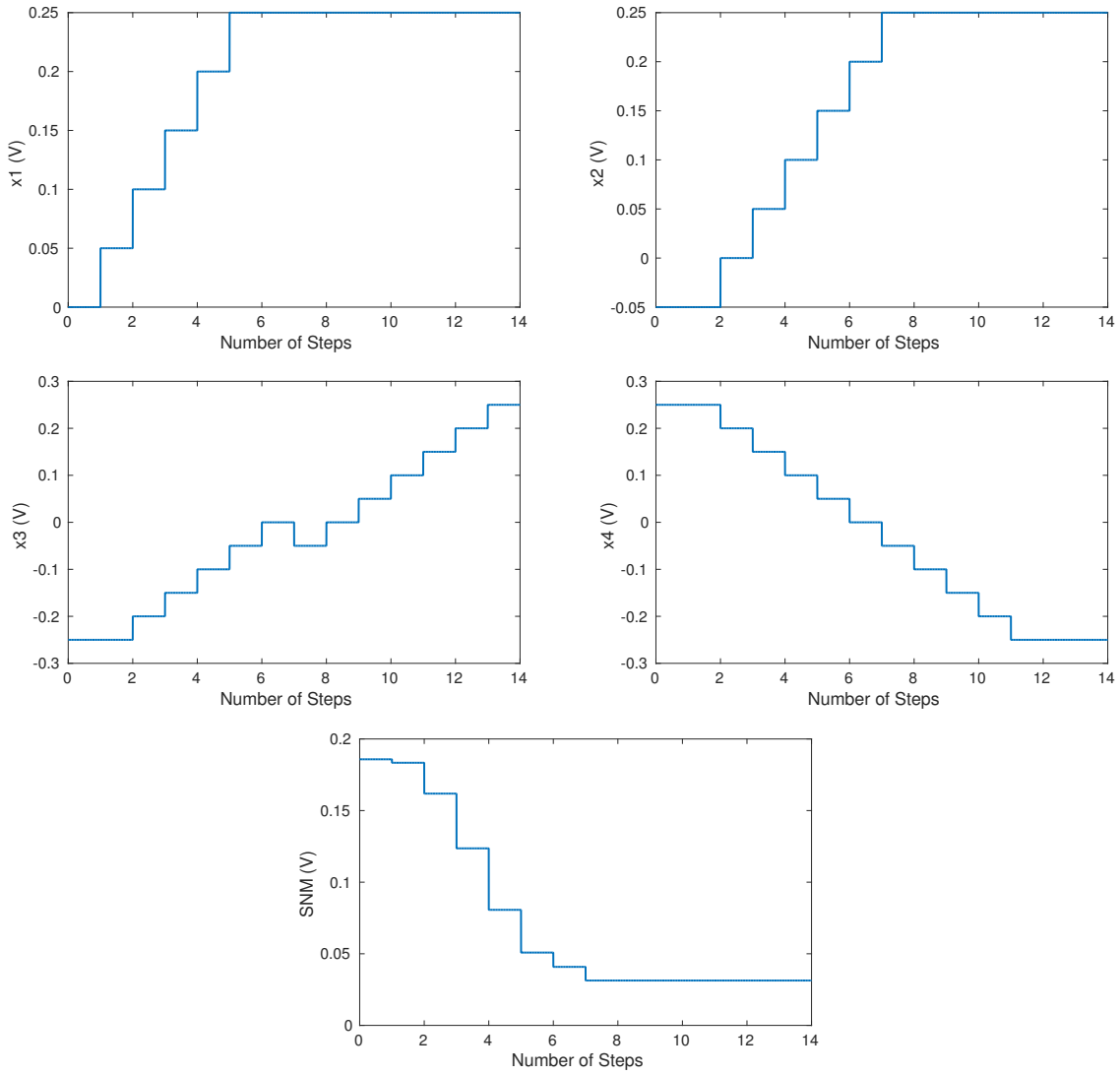
Για να προσεγγίσουμε την κατανομή του P_{FAIL} ακολουθούμε την παρακάτω διαδικασία. Για κάθε σημείο $\mathbf{x}_Y$, με απόσταση από το μέγιστο r_Y, που ο Αλγόριθμος 2 υπέδειξε, υπολογίζουμε το P_{FAIL} θεωρώντας ως Y το αντίστοιχο SNM και χρησιμοποιώντας την κατανομή χ^2. Το P_{FAIL} για κάθε spec Y ικανοποιεί την Εξίσωση 0.9, όπου CDF και PDF η αθροιστική συνάρτηση κατανομής και η συνάρτηση πυκνότητας πιθανότητας αντίστοιχα. Τα αποτελέσματα του P_{FAIL} προς το SNM spec φαίνονται στην Εικόνα 0.11. Επιλέξαμε να μην προσεγγίσουμε την κατανομή του P_{FAIL}, καθώς απαιτούνται περισσότερα σημεία για να υπολογιστεί με ακρίβεια.

$$P_{FAIL} = 1 - CDF(r < r_Y) = 1 - \int_{-\infty}^{r_Y} PDF_r dr \tag{0.9}$$

Είναι εμφανές ότι η μείωση του Y οδηγεί σε μικρότερο P_{FAIL} και αντίστροφα. Κάτι τέτοιο είναι αναμενόμενο, αφού η πιθανότητα αποτυχίας εξαρτάται άμεσα από το κριτήριο και το SNM spec καθορίζει το θόρυβο που μπορεί να ανεχτεί το κύτταρο. Ο αποφασιστικός ρόλος του SNM spec φαίνεται έντονα στις χαμηλότερες τιμές του Y όπου η πτώση του P_{FAIL} είναι ραγδαία.

Αν συγκρίνουμε τις πιθανότητες που κατέληξαν οι δύο τεχνικές για το ίδιο Y, καταλήγουμε στο συμπέρασμα ότι η τυπική MPFP μέθοδος οδηγεί σε ένα πολύ πιο αισιόδοξο αποτέλεσμα. Συγκεκριμένα, το P_{FAIL} της πρώτης μεθοδολογίας εκτιμήθηκε 9.9×10^{-12} ενώ της δεύτερης 1.3×10^{-10}. Η διαφορά αυτή οφείλεται στην ανεπάρκεια της τυπικής MPFP τεχνικής να προσεγγίσει το χώρο **F**. Αντιθέτως, η νέα προσέγγιση του MPFP παρέχει μία πιο ακριβή εκτίμηση του **F**. Επειδή κινούμαστε από το μέγιστο προς την κατεύθυνση με τη μεγαλύτερη πτώση του SNM, σε κάθε βήμα μπορούμε να διαχωρίσουμε τα περισσότερα δείγματα **x** με $SNM(\mathbf{x}) < Y$.

Πρέπει να επισημανθεί ότι σε μεγαλύτερους χώρους είναι πολύ πιθανό να υπάρχουν περισσότερα μέγιστα ή ελάχιστα. Αυτά μπορούν να εντοπιστούν υλοποιώντας τον Αλγόριθμο 1 (ή τον Αλγόριθμο coordinate descent αν πρόκειται για ελάχιστα) με διαφορετικά σημεία εκκίνησης κάθε φορά. Στην περίπτωση αυτή, η διαδικασία που ακολουθεί μετά πρέπει να επαναληφθεί για όλα ώστε να έχουμε μία ακριβή εκτίμηση της κατανομής του P_{FAIL}. Αλλιώς, στην περίπτωση πολλαπλών μεγίστων, το



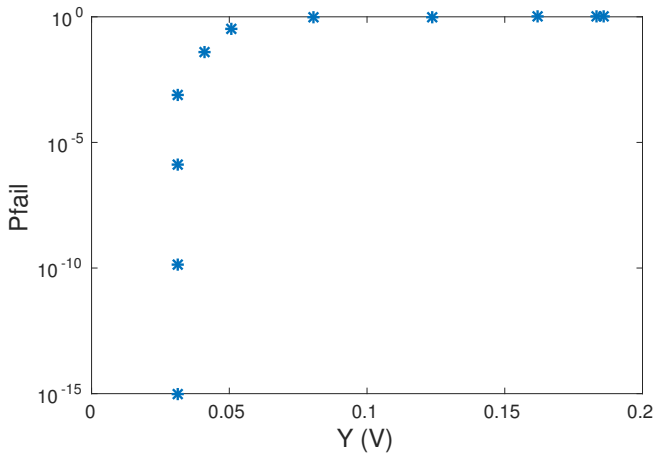
Εικόνα 0.10: Αποτελέσματα της εφαρμογής του Αλγόριθμου 2

σύνολο **F** θα περιέχει τα υπόλοιπα μέγιστα και συνεπώς, πολλές περιπτώσεις που δεν προκαλούν την αποτυχία του κυττάρου, γεγονός που θα οδηγήσει σε ένα πολύ απαισιόδοξο αποτέλεσμα. Αντίστοιχα, στην περίπτωση πολλαπλών ελαχίστων, θα καταλήξουμε με ένα πολύ αισιόδοξο P_{FAIL}.

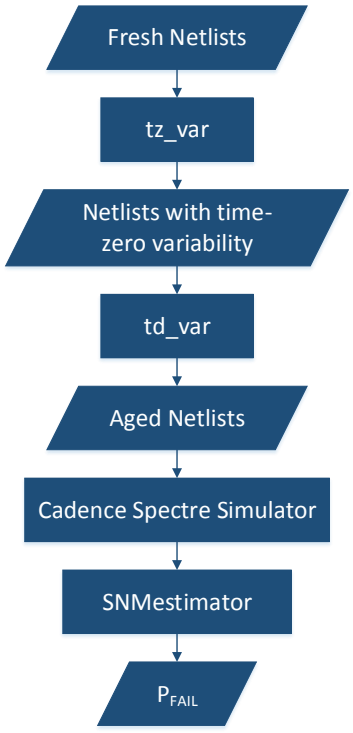
Σύγκριση των αποτελεσμάτων της μεθόδου MPFP και της Monte Carlo

Η τεχνική Monte Carlo έχει χρησιμοποιηθεί ευρέως για της εκτίμηση της αξιοπιστίας των CMOS κυκλωμάτων [16],[17],[9]. Παρόλο που παρέχει ακριβή αποτελέσματα για πολαιότερες τεχνολογίες, όταν πρόκειται για σύγχρονες, πολύ μικρών διαστάσεων τεχνολογίες, αποτυγχάνει να προσεγγίσει με επαρκή ακρίβεια πολύ χαμηλές τιμές του P_{FAIL} . Σε αυτήν την Ενότητα, ελέγχουμε την ακρίβεια αυτής της μεθοδολογίας για μία υψηλότερη τιμή της πιθανότητας αποτυχίας για ένα SRAM κύτταρο βασισμένο σε FinFET. Χρησιμοποιούμε το σκελετό που ακολουθεί (Εικόνα 0.12) για να υπολογίσουμε το P_{FAIL} και να το συγκρίνουμε με το αποτέλεσμα της μεθόδου MPFP.

Αρχικά, παράγουμε τα aged netlists χρησιμοποιώντας τα εργαλεία `tz_var` και `td_var`. Πρώτα, προσθέτουμε στατική μεταβλητότητα σε ένα πλήθος fresh SRAM netlists με το εργαλείο `td_var` και έπειτα, εισάγουμε χρονικά εξαρτώμενη μεταβλητότητα με το εργαλείο `td_var` στα netlists που προέκυψαν. Στη συνέχεια, παράγουμε τα butterfly curves κάθε περίπτωσης χρησιμοποιώντας το `Ca-`



Εικόνα 0.11: Η πιθανότητα αποτυχίας του SRAM κυττάρου για διάφορες τιμές του SNM spec.

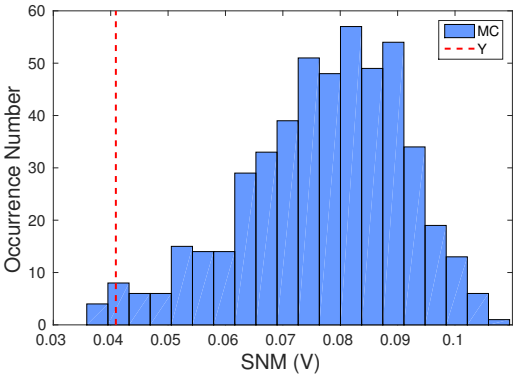


Εικόνα 0.12: Ο σκελετός προσομοίωσης της τεχνικής Monte Carlo.

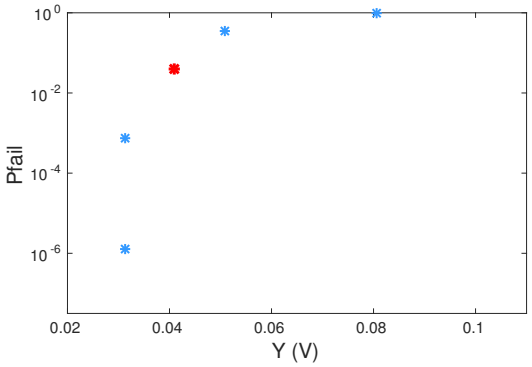
`dence Spectre Simulator` και εκτιμάμε την τιμή του SNM με το εργαλείο `SNMestimator`. Η πιθανότητα αποτυχίας ισούται με το πλήθος των περιπτώσεων που οδηγούν σε αποτυχία (σύμφωνα με την Εξίσωση 0.7) προς το συνολικό πλήθος περιπτώσεων (Equation 0.5).

Εφαρμόσαμε τη Monte Carlo μεθοδολογία και πάλι έπειτα από τρία περίπου χρόνια λειτουργίας και επικεντρωθήκαμε στα τέσσερα τρανζίστορ των cross-coupled αντιστροφέων. Το ιστόγραμμα του SNM παρουσιάζεται στην Εικόνα 0.13i μαζί με το spec Y σε σύγκριση με το αντίστοιχο P_{FAIL} της μεθόδου MPFP (Εικόνα 0.13ii). Επιλέξαμε μία υψηλή τιμή Y για την σύγκριση αυτή, έτσι ώστε να ελέγξουμε την ακρίβεια της τεχνικής Monte Carlo για μεγάλες τιμές P_{FAIL}. Το P_{FAIL} εκτιμήθηκε σχετικά παρόμοιο για τις δύο μεθοδολογίες που εξετάζουμε.

Για μικρότερες πιθανότητες η τεχνική Monte Carlo δε θεωρείται αποδοτική, ειδικά για σύγχρονες τεχνολογίες. Πρώτον, το κόστος υπολογισμού αυξάνεται δραματικά καθώς μειώνεται η πιθανότητα. Για παράδειγμα, για την εκτίμηση πιθανότητας της τάξεως 10^{-5}, είναι απαραίτητες τουλάχιστον 10^6



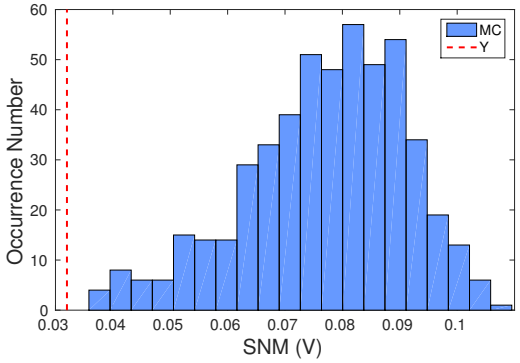
(i) Το ιστόγραμμα του SNM μαζί με το Y.



(ii) Το P_{FAIL} της MPFP μεθόδου για το ίδιο Y.

Εικόνα 0.13: Σύγκριση μεταξύ των αποτελεσμάτων των Monte Carlo και MPFP για μία υψηλή τιμή του spec Y.

προσομοιώσεις, γεγονός που καθιστά τη μέθοδο αυτή μη αποδοτική. Αυτό το γεγονός εισάγει μεγάλη ανακρίβεια για πολύ μικρές τιμές του P_{FAIL}, το οποίο επιβεβαιώνεται και από τους υπολογισμούς μας, αφού ο αριθμός των αποτυχημένων δειγμάτων είναι μηδέν όταν επιλέγουμε ένα πιο ρεαλιστικό spec, όπως φαίνεται στην Εικόνα 0.14. Για να μειώσουν το χρόνο υπολογισμού, οι μηχανικοί επεξεργάζονται ένα μικρότερο πλήθος δειγμάτων και έπειτα βρίσκουν την κατανομή που τα προσεγγίζει καλύτερα (συνήθως την κανονική) [27]. Όμως, το συνολικό SNM προεκτείνεται περισσότερο προς τις χαμηλές τιμές [44] τις οποίες η κατανομή δεν τις συλλαμβάνει.



Εικόνα 0.14: Το ιστόγραμμα του SNM μαζί με μία ρεαλιστική τιμή Y.

Συμπεράσματα και προτάσεις για μελλοντική έρευνα

Η εκτίμηση της αξιοπιστίας των σύγχρονων συρρικνωμένων συσκευών αποτελεί μία σημαντική πρόκληση. Για να παραχθεί ένα ακριβές αποτέλεσμα θα πρέπει να ληφθούν υπόψιν και τα στατικά και τα χρονικά εξαρτώμενα φαινόμενα. Το κυριότερο φαινόμενο που δημιουργεί χρονικά εξαρτώμενη μεταβλητότητα είναι το BTI, το οποίο προκαλεί τη διασπορά των παραμέτρων των τρανζίστορ και μπορεί να προκαλέσει αποτυχίες στο σύστημα. Ως αποτέλεσμα της όλο και αυξανόμενης εμφάνισης αποτυχιών συστήματος, έχουν αναπτυχθεί πολλές μέθοδοι που υπολογίζουν την πιθανότητα αποτυχίας, όπως οι Monte Carlo, Quasi Monte Carlo, Importance Sampling. Όμως, αυτές οι μεθοδολογίες δεν προσφέρουν αρκετή ακρίβεια όταν χρησιμοποιούνται σε μοντέρνες τεχνολογίες.

Σκοπός αυτής της διπλωματική εργασίας ήταν η έρευνα μιας αποδοτικής τεχνικής για την εκτίμηση της πιθανότητας αποτυχίας με ικανοποιητική ακρίβεια, ενός κυττάρου SRAM βασισμένο σε FinFET, λαμβάνοντας υπόψιν τη μετρική SNM για τη λειτουργία hold. Αρχικά αναπτύξαμε ένα σύνολο εργαλείων

και υλοποιήσαμε την τυπική MPFP μεθοδολογία. Σύμφωνα με αυτή, σκοπός είναι η εύρεση του σημείου στο χώρο μετατόπισης των V_{th}, που οδηγεί σε αποτυχία και έχει τη μεγαλύτερη πιθανότητα εμφάνισης. Η τεχνική αυτή, παρόλο που εντοπίζει σωστά το σημείο αυτό, δεν απομονώνει επαρκώς το σύνολο **F** των σημείων που προκαλούν αποτυχία.

Στη συνέχεια, μελετήσαμε μία ανασύνθεση της προηγούμενης τεχνικής, η οποία βασίζεται στην κυρτότητα του χώρου του SNM, με σκοπό την καλύτερη εκτίμηση του συνόλου **F**. Πρώτα, εστιάσαμε στην αναγνώριση ενός μεγίστου του χώρου και εξετάσαμε την κυρτότητα γύρω από αυτό. Αφού βεβαιωθήκαμε ότι αποτελεί το μοναδικό μέγιστο του χώρου, σκοπός ήταν η εύρεση του σημείου με μικρότερη απόσταση από αυτό που οδηγεί στην αποτυχία γι' αυτό και αναπτύξαμε ένα εργαλείο που ξεκινά από το μέγιστο και κινείται κάθε φορά προς το σημείο με το μικρότερο SNM. Με αυτό τον τρόπο απομονώσαμε τα σημεία που οδηγούν σε αποτυχία έξω από το hypersphere με κέντρο το μέγιστο και ακτίνα ίση με την εν λόγω απόσταση. Η σύγκριση των αποτελεσμάτων έδειξε ότι αυτή η μεθοδολογία καταλήγει σε ένα πιο ρεαλιστικό αποτέλεσμα. Τέλος, ελέγξαμε την ακρίβεια της τεχνικής Monte Carlo για μεγάλες τιμές του P_{FAIL}, συγκρίνοντάς τα αποτελέσματά της με την δεύτερη MPFP μέθοδο.

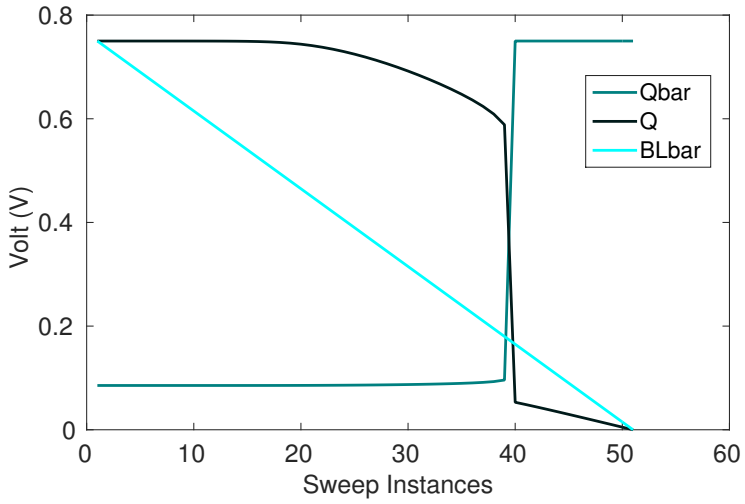
Μερικές βελτιώσεις και επεκτάσεις που μπορούν να γίνουν στην καινοτόμα προσέγγισση της MPFP τεχνικής, η οποία οδήγησε στο πιο ακριβές αποτέλεσμα, είναι οι παρακάτω:

Παραλληλοποίηση της εκτίμησης του SNM

Λόγω του μεγάλου υπολογιστικού κόστους της δημιουργίας των butterfly curves και της εκτίμησης του SNM με τα εργαλεία `Cadence Spectre Simulator`, `SNMestimator`, είναι ουσιαστικής σημασίας η παραλληλοποίηση τους. Η απουσία εξάρτησης των σημείων ΔV_{th} μας επιτρέπει την ταυτόχρονη επεξεργασία των netlist, καθώς επίσης και τον παράλληλο υπολογισμό του SNM των διάφορων butterfly curves. Με αυτόν τον τρόπο μπορούμε να αυξήσουμε την ακρίβεια της μεθόδου μας λαμβάνοντας υπόψιν και τις έξι συσκευές του SRAM κυττάρου και μειώνοντας το βήμα μετατόπισης των V_{th}.

Η ένταξη της μετρικής Write Trip Point

Εκτός από την αντοχή στο θόρυβο και την ευστάθεια του κυττάρου κατά την λειτουργία hold, εξίσου σημαντική είναι και η εγγύηση της σωστής λειτουργίας write με τη μικρότερη κατανάλωση ενέργειας. Συνεπώς, το κύτταρο πρέπει να έχει μία λογική τιμή Write Trip Point (WTP), η οποία ορίζει τη μέγιστη τάση που απαιτείται για να αλλάξει το περιεχόμενο του κυττάρου [45] (Εικόνα 0.15). Επομένως, η ενσωμάτωση του WTP στο κριτήριο αποτυχίας οδηγεί στην αύξηση της ακρίβειας.



Εικόνα 0.15: Παράδειγμα του Write Trip Point

Εκτίμηση του P_{FAIL} βασικών λογικών πυλών

Μία ενδιαφέρουσα επέκταση θα ήταν η εκτίμηση του P_{FAIL} άλλων δομικών στοιχείων όπως λογικών πυλών. Αν και αυτό έχει ήδη προταθεί [33], ο υπολογισμός της πιθανότητας αποτυχίας βασικών λογικών πυλών και η διάδοση αυτών των εκτιμήσεων σε logic paths δεν έχει ακόμα μελετηθεί. Στην περίπτωση αυτή, η καθυστέρηση μπορεί να χρησιμοποιηθεί ως κριτήριο και η κυρτότητα του χώρου της καθυστέρησης κάθε πύλης θα πρέπει να επαληθευθεί. Έτσι μπορούμε να υπολογίσουμε την πιθανότητα αποτυχίας όχι μόνο μνημών, αλλά και επεξεργαστών.

Λέξεις κλειδιά

Πιθανότητα αποτυχίας, SRAM, Bias Temperature Instability, Most Probable Failure Point, FinFET, στατική μεταβλητότητα, χρονικά εξαρτώμενη μεταβλητότητα, Static Noise Margin, αξιοπιστία.

Abstract

Due to the aggressive downscaling of devices, the reliability issue has come to surface. The everlasting demand for the shrinking of dimensions has led to the introduction of the multi-gate, three dimensional FinFET device. The complexity of the new device compared to planar CMOS renders the estimation of the failure probability (P_{FAIL}) of integrated circuits (IC) even more difficult. The main threat of a system's reliability is the variability provoked by time-zero phenomena during the manufacturing process and time-dependent effects, with Bias Temperature Instability (BTI) being the dominant one. The model that accurately explicates BTI is the atomistic one which efficiently captures the stochastic nature of this degradation mechanism.

Techniques that were widely-used for the evaluation of the P_{FAIL} although they were effective for older technologies, when it comes to modern downscaled devices either require a colossal number of simulations or lead to inaccurate results. Therefore, in this work we focus on the Most Probable Failure Point (MPFP) methodology and we explore the accuracy limits of the standard approach against a state-of-the-art one. We examine the stability of a 6T Static Random Access Memory (SRAM) cell since it is a component highly vulnerable to degradation and we use the Static Noise Margin (SNM) for the hold operation as metric. We compare the results of the two concepts and we verify our claim that the MPFP methodology is much more realistic, using the Monte Carlo technique.

Key words

Failure Probability, SRAM, Bias Temperature Instability, Most Probable Failure Point, FinFET, time-zero variability, time-dependent variability, Static Noise Margin, reliability.

Acknowledgements

First of all, I would like to express my sincere thanks to Professor Dimitrios Soudris, who offered me the opportunity to carry out this thesis under his supervision. His valuable and constructive advices along with his research experience inspired me and strengthened my motivation to work on this subject.

This thesis would not be completed without the guidance and continuous support of Michail Noltsis. I am really grateful for the time he devoted to share his knowledge with me. His cooperation and his willingness to help me overcome each difficulty, encouraged me and made it possible to accomplish this work.

Furthermore, I would like to offer my gratitude to Dimitrios Rodopoulos, who provided his expertise. His insightful suggestions greatly assisted the research.

Last but not least, I would like to thank my family for encouraging me throughout my studies in NTUA and helping me to fulfill my goals, as well as my friends for supporting me.

Eleni M. Maragkoudaki,

Athens, March 16, 2017

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

The reliability of hardware components has been a major concern for digital designers. The continuous downscaling of device dimensions, although allows designers to increase the functionality per unit area, has negative impact on system reliability due to the parameters variation [3] and has led to a shift to probabilistic design. The fluctuation of the parameters is initially provoked by time-zero phenomena during the imperfect front-end manufacturing process [13], while degradation phenomena like Hot-Carrier Injection (HCI) [14], Time-Dependent Dielectric Breakdown (TDDB) [15] and Bias Temperature Instability (BTI) [40] effectuate the time-dependent variability.

A component most vulnerable to degradation is the Static Random Access Memory (SRAM) [1]. The shrinking of device dimensions along with the decrease of supply voltage further escalate the difficulty to address the stability of the SRAM cell, which "determines the soft-error rate and the sensitivity of the memory to process tolerances and operating conditions" [6]. Thus, the evaluation of the cell stability has been a crucial issue and the SRAM cell is one of the most studied cases.

The endless demand for larger integration of transistors on a chip has led to the introduction of multi-gate devices, with FinFET being the most promising one [31]. The general idea of this technology is that the gate is wrapped around the channel instead of being above the channel as in the standard planar FET. This new device architecture allows the further reduction of dimensions. FinFET has been introduced since 1999 [19], Intel has already used 22nm FinFET devices in the Avoton processors [20] and Samsung and GlobalFoundries have announced that they will use FinFETs at their 14nm process node [21],[22].

In this thesis we evaluate the reliability of a FinFET-based SRAM cell under both time-zero and time-dependent variability. We study the BTI-induced degradation of the cell, since it is the dominant aging phenomenon that generates fluctuations in the power and the delay of electronic circuits. Specifically, we focus on developing two different tools that calculate the failure probability (P_{FAIL}) of the cell. The first one is implemented based on the standard Most Probable Failure Point (MPFP) methodology, as described in [29],[4], while the other tool follows a state-of-the-art approach of the MPFP which can provide more realistic results, as mentioned in [33]. The metric under study is the Static Noise Margin (SNM) for the hold operation, which describes the stability of the cell and defines failure in maintaining the correct bit value.

In Chapter 2, we first analyze the time-zero and time-dependent variability and we focus on the BTI phenomenon and the two main models that describe it, the reaction-diffusion (RD) model and the atomistic one. The insufficiency of the first model to accurately capture the degradation of the modern

downscaled devices is also highlighted. In addition, the structure of the FinFET device is described along with its advantages in comparison to the standard planar FET. In the final part, the metrics that have been developed to estimate the reliability of a system are summarized along with the P_{FAIL} and some of the widely used techniques that aim to evaluate it.

Moving on to the next Chapter, we elaborate on the simulation framework of the typical MPFP concept that we implemented and we present the tools that were developed. The method of computing the SNM is pointed out and also the failure criterion, which is based on the SNM specification (threshold). Eventually, the result of the P_{FAIL} is demonstrated after approximately three years of operation.

In Chapter 4, the inability of the previous technique to isolate all the failed cases in the variation space is explained, which affects the calculation of the P_{FAIL} leading to more optimistic results. Thus, we follow a new approach to implement our tool. In addition, the novelty of our work lies in the study of the concavity of the SNM space, since it is essential for the precision of this methodology. We also describe in detail the algorithms that were used and we present the results of the concavity and the P_{FAIL}. Finally, we compare this methodology with the Monte Carlo for high P_{FAIL} values to confirm that the Monte Carlo offers accurate enough results these values of probability.

In the last Chapter, the conclusions on the accuracy, the computational expense that derived from the implementation of the two techniques and the comparison of their results are underlined. Moreover, some thoughts for future work are proposed that could improve accuracy by better describing the stability of the SRAM cell and restrict the computational time.

CHAPTER 2

Prior Art

2.1 Introduction

The aggressive shrinking of device dimensions, has deteriorated variability and degradation phenomena in modern hardware components. A component heavily affected by such phenomena is the SRAM cell[10]. Thus, in the current section we elaborate on both time-zero and time-dependent variability phenomena that affect system reliability and can lead to failure in the operation. We primarily focus on studying Bias Temperature Instability (BTI), since it is the dominant cause of time-dependent variability. Therefore, the two main, state-of-the-art models that explicate BTI, namely the reaction-diffusion and the atomistic model are discussed.

Furthermore, in the need for following Moore's law, engineering industry has focused on new device architectures which render the evaluation of the cell's reliability even more critical. Among these architectures the most promising one is the FinFET. Hence, the structure of FinFET is studied, along with the necessary factors that have led the industry to focus on this technology. In addition, the differences between FinFET devices and planar ones are also highlighted. To describe cell's correct operation and resilience to degradation effects, we will use the metric of the Static Noise Margin for the hold operation (SNM).

Finally, various metrics describing reliability are overviewed. In this thesis we focus on the failure probability (P_{FAIL}). Some of the existing P_{FAIL} estimation techniques are demonstrated: Monte Carlo (MC), Quasi Monte Carlo (QMC) and Importance Sampling, while their incapability to accurately compute the P_{FAIL} of the modern aggressively downscaled devices is indicated. Towards this direction, we exploit the Most Probable Failure Point (MPFP) methodology for our Pfail estimations.

2.2 Time-Zero and Time Dependent-Variability

Nowadays, the increasing demand for electronic devices to perform numerous tasks, has led to an aggressive downscale of transistor's dimensions. Moore's law, which indicated that the number of transistors in an integrated circuit doubles approximately every two years, has been proved accurate so far. However, with this scaling of devices, the issue of reliability has come to surface and has attracted the attention of the electronic engineering community.

Reliability is defined as the possibility that a system will perform correct service along its operating period [39]. The reliability of an electronic device is affected by both internal causes that existed in the

product from the beginning and by external, such as temperature, humidity and workload stress. Modern downscaled devices are even more affected by all these determinants, causing the generation of system failures more and more often.

The major factor that provokes instability and system failures is the variability of the device's parameters and especially the absolute increase of threshold voltage (V_{th}). On the one hand, time-zero effects like Random Dopant Fluctuations (RDF) add to the initial spread of the transistor's parameters [24] and on the other hand, threshold voltage is affected by the aging of the transistor (time-dependent variability).

The effect of the time-zero variability during the front-end manufacturing process is that the transistor's threshold voltage is not fixed, but is rather distributed, with the Gaussian distribution being the most sufficient approximation [10]. The mean value is equal to the intended threshold voltage and the standard deviation is derived from Pelgrom's mismatch parameter A_{VT} and is equal to:

$$\sigma_{V_{th,0}} = \frac{A_{VT}}{\sqrt{2WL}} \tag{2.1}$$

where W and L are the effective device width and length respectively. In the case of FinFET the effective gate length corresponds to the $LFIN$ which is the fin length and the width is $W = 2HFIN + WFIN$ where $WFIN$ and $HFIN$ are the fin width and height respectively [25].

A key degradation mechanism that can induce time-dependent variability is Bias Temperature Instability (BTI) [40]. The main models that attempt to describe it are two, the reaction-diffusion (RD) model and the atomistic one. The RD approach focuses on pFETs and the breaking of the Si-H bonds at the Si/oxide interface when it is under stress. [38] Minority carriers (holes) are replacing the H atoms, while the last are diffusing to the gate-stack. Thus, charge traps are generated which lead to a shift of the V_{th}. In addition, the conventional RD theory predicts a recovery phase at the end of the stress, during which some of the Si-H bonds are annealed.

Since the RD model proved unsuccessful to capture BTI efficiently in aggressively downscaled technology nodes, the reliability research community proposed the more accurate defect-centric atomistic model and focus on the stochastic nature of these defects which are provoked by the charge trapping mechanism [37],[40]. According to it, once a defect is created, it can change between two different states, charged and discharged, in other words occupied and unoccupied, depending on the gate stress. Only the occupied defects cause shift to the V_{th}, which contradicts the RD's concept of trapped charges.

According to the atomistic theory and experimental data [25] the ΔV_{th} can be modeled through a normal distribution with mean value:

$$\langle \Delta V_{th}(t) \rangle \cong A t^a E_{OX}^{\gamma} \tag{2.2}$$

where t is the operating time, A a fitting coefficient, E_{OX} the electric field across gate oxide and γ, a are acceleration exponents for the electric field across gate oxide. E_{OX} is calculated as $(V_G - V_{th})/T_{INV}$, where V_G is the gate stress voltage and T_{INV} the thickness of gate inversion layer. Regarding the standard deviation, there is a correlation between time-zero and time-dependent variability [12] which is described by Equation 2.3. It is has been proven that higher level of time-zero variability leads to the

increment of time-dependent variability.

$$\sigma_{\Delta V_{th}}(t) = \sigma_{V_{th,0}} \sqrt{\frac{\langle \Delta V_{th}(t) \rangle}{100mV}} \tag{2.3}$$

Finally, the total standard deviation is equal to:

$$\sigma_{V_{th,tot}}(t) = \sigma_{\Delta V_{th}}(t) + \sigma_{V_{th,0}} \qquad \sigma_{V_{th,tot}}(t) = \left(\sqrt{\frac{\langle \Delta V_{th}(t) \rangle}{100mV}} + 1 \right) \sigma_{V_{th,0}} \tag{2.4}$$

2.3 Comparing FinFETs to planar FETs

Today, the semiconductor industry is confronted with serious challenges in order to continue the downscaling of the devices and thus follow the pace of Moore's law. "The most important among these challenges is the diminishing gate control over the channel, which manifests itself in the form of increased short-channel effects (SCE) and leakage currents"[18]. Hence, the continued scaling of planar FETs reduces their efficiency and therefore, multiple gate device architectures are explored. One of these is the FinFET device which is used for dimensions lower than 28nm.

The FinFET technology promises to ensure that the current progress with increased levels of integration can be maintained. In the FinFET device, the conducting channel is wrapped by a thin silicon "fin" and current flows along the top and side surfaces of the fin, as shown in Figure 2.1 [30]. This form of gate structure provides improved electrical control over the channel conduction and it provokes the reduction of the SCE and leakage current.

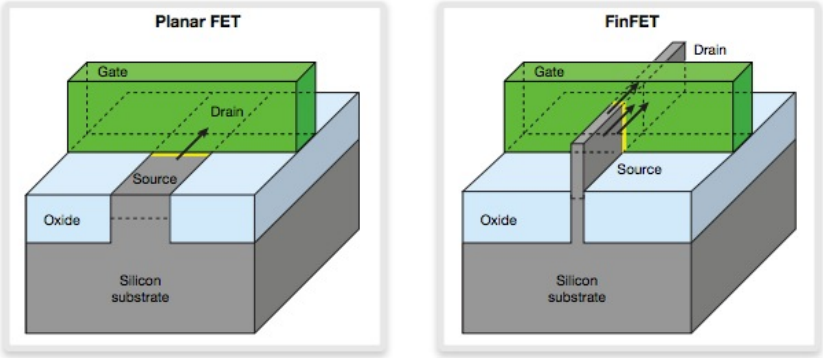


Figure 2.1: Simplified comparison between planar FET and FinFET [30]

Besides the fact that FinFET structure offers better scalability, due to diminution of SCE and leakage current, it has more advantages in comparison with the planar gate structure devices. A FinFET can operate from a lower supply voltage since it has a lower threshold voltage. This drop in supply voltage can improve dynamic power consumption significantly and along with the fact that provides faster switching speed, it is not surprising that it is considered the most promising technology.

The introduction of downscaled FinFET devices in SRAM cells renders the evaluation of its reliability even more difficult. The operations of an SRAM cell that can lead to a potential failure are three: the read, write and hold operations. The hold operation is the stand-by mode during which the cell is supposed to hold the data. However, when the supply voltage is reduced, the stored data may be destroyed [23]. Thus, in order to estimate the stability of the cell for the hold operation, a permanent

failure criterion based on the cell's Static Noise Margin (SNM) is used, which is the maximum voltage that the cell can accept while still maintaining its correct operation [6].

The SNM can be approximated graphically directly through the butterfly curve. It is defined as the diagonal's length of the smallest of the two maximum squares that can be embedded inside the two lobes [8], as shown in Figure 2.2 [32]. However, we use a more efficient method to calculate it, as described in Section 3.3.

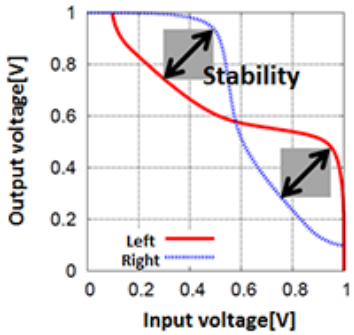


Figure 2.2: Butterfly curve [32].

2.4 Estimating Reliability

Reliability analysis has gained significant attention by the research community and BTI-induced degradation has become one of the most important threats in modern VLSI design. Thus, various metrics have been used in order to describe reliability, such as yield, defects per million, failure rate and mean time to failure.

Definition 2.4.1. Yield is defined as the ratio of the number of functional circuits to that of the total examined circuits [41].

$$\text{Yield} = \frac{\text{Functional circuits}}{\text{Total examined circuits}}$$

Definition 2.4.2. Defects Per Million (DPM) is the number of defects to the total number of examined circuits. It is another measure to quantify yield. In relevant analysis, a defect is considered a failed circuit.

$$\text{DPM} = \frac{\text{Number of defects}}{\text{Total examined circuits}} \cdot 10^6$$

Definition 2.4.3. Mean Time To Failure (MTTF) represents the mean value of operating time before a failure occurrence [43].

$$\text{MTTF} = \frac{\text{Operating hours}}{\text{Number of Failures}}$$

Definition 2.4.4. Failure rate corresponds to the number of failures during lifetime operation, normalized to one billion device hours and its connection with MTTF is shown in the equation below [43].

$$\text{Failure Rate} = \frac{10^9}{\text{MTTF}}$$

In our work we focus on the aggregate metric Failure Probability (P_{FAIL}) for the efficient reliability analysis. P_{FAIL} is defined as the probability of a failure occurrence in a particular time period.

Considering that the probability of a cell's permanent failure is too low, an accurate appraisal is hardly impossible and requires a massive computational complexity. Nevertheless, due to the fact that P_{FAIL} incorporates all the phenomena that generate V_{th} variability, various methods have been developed that aim to calculate it, such as Monte Carlo, Quasi Monte Carlo, Importance Sampling and the Most Probable Failure Point which is the technique we will use.

2.4.1 Monte Carlo and Quasi-Monte Carlo

The typical Monte Carlo method has been widely used for the P_{FAIL} calculation of SRAM cells. It consists of the following stages [16]: first the random selection of the N samples, $x_1, x_2, \ldots, x_N$, in the space under examination according to a distribution, usually the Gaussian [9], secondly the function I is calculated which indicates whether the samples pass or not the failure criterion and finally the P_{FAIL} is estimated as:

$$P_{FAIL} = \frac{1}{N} \sum_{i=1}^{N} I(x_i) \tag{2.5}$$

The disadvantage of this technique is the prohibitive computation time needed to capture low failure probability estimations.

In contrast to the regular Monte Carlo method which is based on sequences of pseudorandom numbers, Quasi-Monte Carlo uses low-discrepancy sequences, such as the Halton sequence, the Sobol sequence, or the Faure sequence, which lead to a faster rate of convergence. Although the estimations of P_{FAIL} based on this methodology are efficient for older technologies, when it comes to modern downscaled devices they lack of sufficient accuracy [10].

2.4.2 Importance Sampling

While Monte Carlo requires a large number of simulations in order to be efficient, "Importance Sampling is a method of reducing the variance of the Monte-Carlo estimator based on the idea of sampling from an alternative distribution with more failure samples, instead of the original one" [4]. The failure probability according to this technique is given below:

$$P_{FAIL} = \frac{1}{N} \sum_{i=1}^{N} I(x_i) \frac{h(x_i)}{g(x_i)}$$

Since most of Monte Carlo's samples are around the mean, a common practice is to shift the initial distribution into the failure region. Another approach is the generation of variables based on a mixture of distributions [2]. The idea of all methods is to sample more extreme cases. The selection of the alternative distribution is critical since it can lead to inaccurate estimation of the failure probability, which usually ends up to be more pessimistic.

2.4.3 Most Probable Failure Point

P_{FAIL} is very low and an extremely large amount of computations is essential to produce an accurate result using old techniques. Therefore, a more accurate method has been proposed, namely the Most Probable Failure Point [26],[23],[29],[4],[11]. The concept of MPFP is to locate in the variation

space the point with the higher probability of occurrence for which the cell fails according to a failure criterion.

The principal advantage of this method is that there are not any assumptions about the distribution of the failure criterion. In other techniques, such as the Quasi-Monte Carlo [10], it is presumed that the failure criterion follows a known distribution, more frequently the Gaussian. This approximation escalates the inaccuracy of the results due to the deviation of the actual distribution and the approximated one [26].

2.5 Conclusions

The reliability analysis of SRAM cells has become a critical issue with the aggressively downscaling of devices. The reaction-diffusion model that first described the BTI degradation mechanism and worked efficiently in older technologies, fails to capture the deterioration observed in modern devices. Therefore, the atomistic theory was developed that contains both time-zero and time-dependent variability and incorporates the stochastic nature of BTI. It focuses on the defects and considers them as switching oxide traps, that can be charged or uncharged, depending on the stress time among others. Only the charged traps add to the spread of the transistor's parameters which is the dominant reason of system faiures.

In the need for following Moore's Law, engineering industry has focused on the FinFET device. The continued scaling of the standard planar FET has negative consequences and increases short-channel effects and leakage currents. Hence, new, multiple gate architectures are explored and one of these is the FinFET which not only allows the further shrinking of transistor's dimensions but also offers improvement in many other aspects like power consumption and speed.

However, in view of the above, the evaluation of reliability has become even more complex. Among many metrics describing the reliability of a system, in our work we focus on the failure probability, since it is a more aggregate measure. The old techniques that compute P_{FAIL} such as Monte Carlo, Quasi-Monte Carlo, Importance sampling, either fail to make an accurate estimation, or require a tremendous amount of calculations. Thus, the motivation behind this thesis is to develop a computationally feasible tool that accurately computes the P_{FAIL} using the MPFP methodology for netlists based on FinFET devices.

CHAPTER 3

The standard MPFP methodology

3.1 Introduction

This Chapter describes the typical MPFP methodology which we followed in order to estimate the failure probability of an SRAM cell that is based on downscaled FinFET devices. The failure criterion that we chose in order to decide which samples lead to the failure of the cell is discussed, as well as the method of locating the MPFP.

Furthermore, the tools which were developed in order to compute the P_{FAIL} are overviewed. First the way we sampled the variation space of the transistor's parameter V_{th} is described and then the tools, that decide whether a sample leads to failure or not and calculate the P_{FAIL} according to the MPFP technique, are explained.

In the final part of this Chapter the result of the implementation of this technique is presented.

3.2 Premises

In the previous Chapter, the insufficiencies of the widely-used methods of estimating the failure probability of modern downscaled nodes, such as Monte Carlo, Quasi Monte Carlo and Importance Sampling were mentioned. Therefore, in this thesis we leverage the MPFP concept to evaluate the impact of both time-zero and time-dependent variability on P_{FAIL}. The methodology we follow to approximate P_{FAIL} is demonstrated in Figure 3.1.

Since the parametric failures are primary caused by the variation of the threshold voltages, in the case of the SRAM cell [23], the objective of this methodology is to locate the combination of the involved V_{th} that leads to failure and has the maximum probability of occurrence. Considering as $\mathbf{x} = [x_1, x_2, ..., x_N]$ the combination of ΔV_{th} of the N involved transistors that leads to failure, the P_{FAIL} is appraised as stated in Equation 3.1. According to it, the product of the probability of each device is having a V_{th} shift equal or larger than the one of the MPFP.

$$P_{FAIL} = \max \left\{ \prod_{i=1}^{N} P(|\Delta V_{th,i}| \geq x_i) \right\} \tag{3.1}$$

To calculate the probability that corresponds to each x_i, namely $P(|\Delta V_{th}| \geq x_i)$, the distribution of ΔV_{th} is required. According to prior art, both time-zero and time-dependent variability can be modeled through the Gaussian distribution while the mean value, μ, and the total standard deviation,

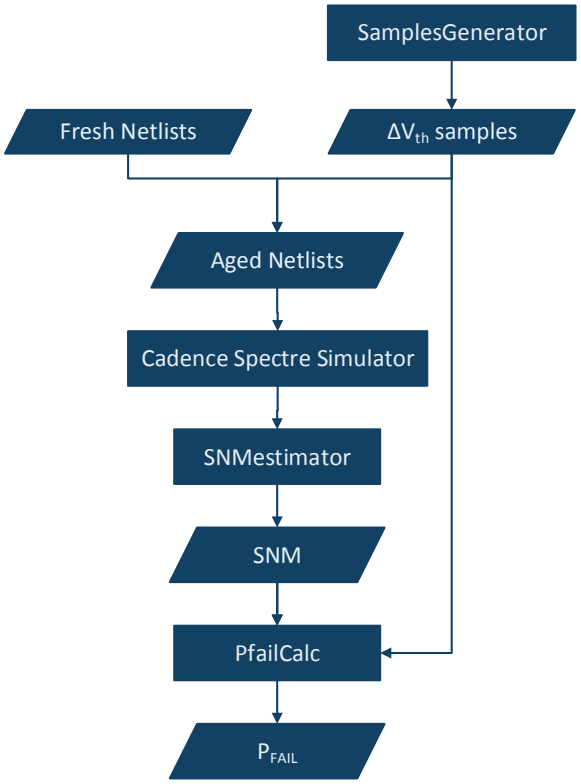


Figure 3.1: The procedure of P_{FAIL} estimation using the MPFP concept.

σ, are formulated in the Equations 2.2 and 2.4 respectively. Therefore, both pFETs and nFETs follow Gaussian distributions: $\Delta V_{th,p} \sim \mathcal{N}(\mu_p, \sigma_p^2)$, $\Delta V_{th,n} \sim \mathcal{N}(\mu_n, \sigma_n^2)$.

In order to identify the MPFP, it is necessary to choose a failure criterion, which can be either static [28],[6] or dynamic [26]. For the sake of simplicity we use the SNM for the hold operation and we emphasize the four transistors of the cross-coupled inverters, as highlighted in Figure 3.2. Thermal noise consists the prevailing cause of noise for SRAM cells and it rises with the shrinking of transistor's size. In particular, for the dimensions we study, the variance of thermal noise should be at least 25mV [7]. On account of this, we chose a SNM specification (threshold), Y, that ensures the tolerance of the cell to thermal noise. Thus, we consider a failed sample when:

$$SNM(\mathbf{x}) < Y \tag{3.2}$$

3.3 Tool description

In this section the tools we used for processing the MPFP concept (Figure 3.1) are further explained.

- The first one is `SamplesGenerator` and it generates the ΔV_{th} samples, $\mathbf{x}$, of the involved transistors, which are the exploration space for the `Cadence Spectre Simulator`. In order to reduce the computational time we limited the variation space of ΔV_{th} between -0.45 and 0.45 Volts. The tool allows a user-defined step which in our case was set to 0.05 Volts. A more fine-grained analysis can be achieved that will, however, cost in terms of simulation time.

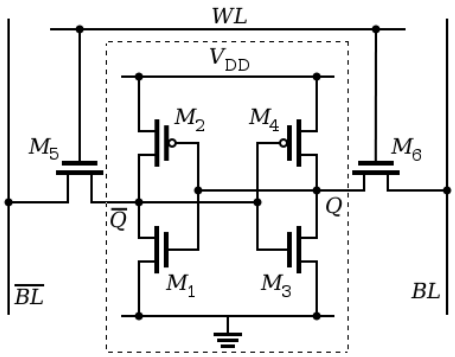
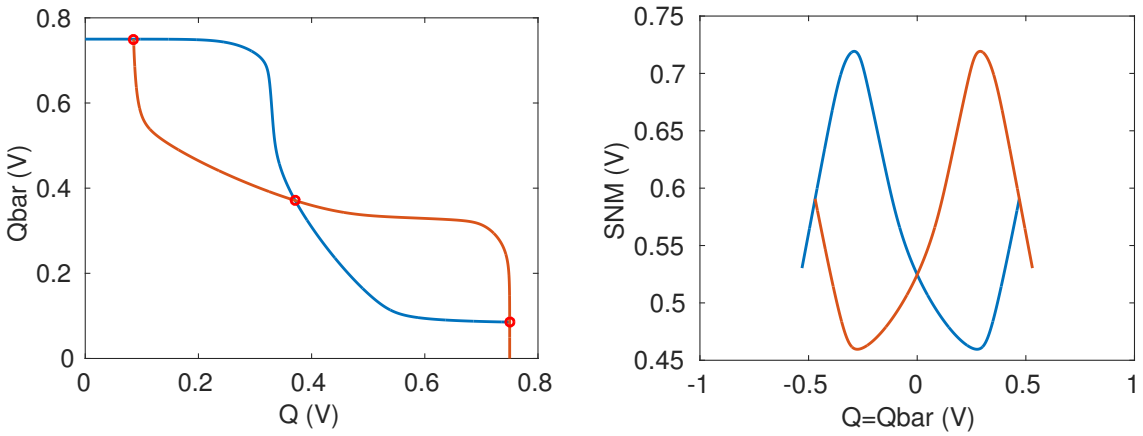


Figure 3.2: SRAM cell circuit. In this thesis we focus on the four transistors as marked above.

- The `Cadence Spectre Simulator` [36] employs the aged SRAM cell netlist that was previously produced by the `SamplesGenerator`. For the netlist we used High Power(HP), PTM-MG modelfiles [35]. This tool creates the butterfly curve for each instance, from which the SNM derives, by performing two DC analyses. In each one, it sweeps the value of one of the internal storage nodes, Q or $\bar{Q}$, and probes the other.

- The next tool, namely `SNMestimator`, is used for the calculation of the SNM for the hold operation. Apart from the methodology described in Section 2.3 to estimate SNM we chose a more efficient approach [6]. According to it, after we rotate the curve by $45°$ and we subtract the values of one node from the other, a new SNM graph is produced. "The maximum and minimum values represent the diagonals of the maximum squares that can be fitted on the butterfly curve" [10]. An original butterfly curve and a rotated one are displayed in Figure 3.3.



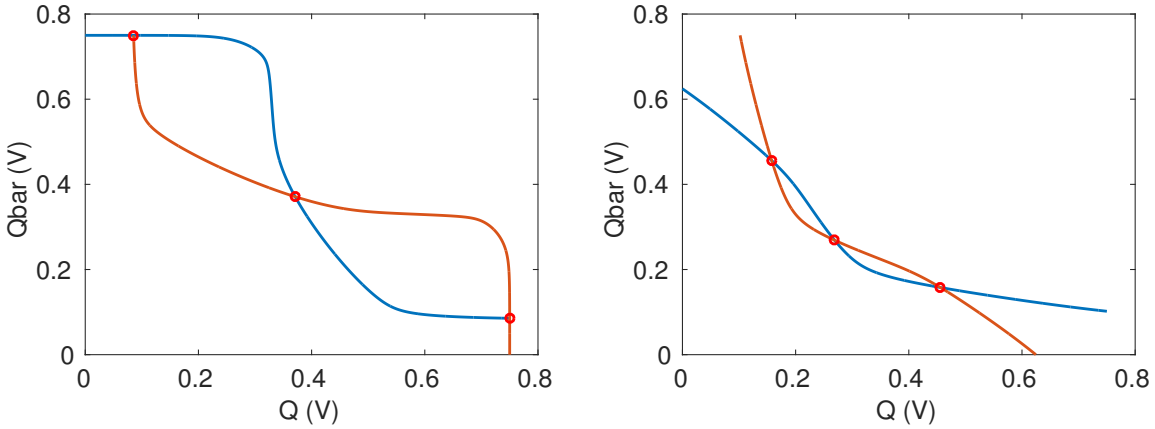
(i) Original butterfly curve

(ii) Rotated by 45°

Figure 3.3: The process of calculating the SNM

- The final tool is `PfailCalc`. It matches the ΔV_{th} samples with the respective SNMs, decides whether a sample is considered failed according to the failure criterion and estimates the P_{FAIL}. The difference of the failed cases with the passed can be seen in Figure 3.4, in which the butterfly curves of both cases are displayed. Since P_{FAIL} is equal to the maximum probability, first our tool computes the probability of occurrence of each failed sample using the Gaussian distribution

and calculates the mean values and standard deviation values of both nFETs and pFETs as stated in Equations 2.2 and 2.4. Then P_{FAIL} is derived through the Equation 3.1.



(i) Passed case **(ii)** Failed case

Figure 3.4: Butterfly curves of passed and failed cases.

3.4 Results

It is self-evident that the failure probability is highly affected by the failure criterion. A cell more tolerant to high amounts of thermal noise means that the value of the SNM spec should increase, which would provoke the rise of the P_{FAIL}. Hence, in our case study the P_{FAIL} is directly depending on the selection of Y.

We studied the failure probability of the cell after approximately three years of operation. The results of the mean values and the standard deviations of both nFET and pFET along with the Pelgrom's mismatch parameter are shown in Table 3.1 and the estimation of P_{FAIL} as well as the failure point with the maximum probability of occurrence are indicated in Table 3.2.

	nFET	pFET
$A_{VT}[mV.\mu V]$	1.0	1.0
$\sigma_{V_{th,0}}[mV]$	26.46	26.46
$V_{DD}[V]$	0.75	0.75
$\langle \Delta V_{th} \rangle [mV]$	130.52	45.55
$\sigma_{V_{th,tot}}[mV]$	40.18	31.93

Table 3.1: Results of the mean values and standard deviations of both nFET and pFET.

In Table 3.1, the differences in the variability of the V_{th} of the n-type and p-type FinFETs can be seen. Although both of them perform equal time-zero variability, the scale of the BTI-induced degradation is different. In fact, the nFETs seem to have a higher acceleration to aging. In addition, the M1 device, as shown in Figure 3.2, has an essential role in the failure probability (Table 3.2). This is in agreement with previous work [26], which indicates that the failure probability of only one device is usually dominant.

Devices	1	2	3	4
$\Delta V_{th}[V]$	0.4	0	0	0
P_{FAIL}	1×10^{-11}			

Table 3.2: The failed sample with the maximum probability of occurrence and the P_{FAIL}

3.5 Conclusions

In this Chapter, we discussed the simulation framework of the typical MPFP methodology. Since the major factor that provokes the degradation of the cell is the V_{th} variability, the concept of this technique is to locate the point in the variation space which causes the failure of the cell and has the maximum probability of occurrence. For this study, we took under consideration the time-zero variability and the BTI effect according to the atomistic approach.

Afterwards, we presented the tools that we developed to implement the MPFP approach. First, we used the `SamplesGenerator` tool to sample the variation space of the ΔV_{th} of the devices and create the aged netlist. Next, we used the `Cadence Spectre Simulator` to engender the butterfly curves by performing two DC analysis for each instance. To calculate the SNM we developed the `SNMestimator` which instead of estimating the SNM directly through the butterfly curve, uses a more efficient approach, as described in Section 3.3. In the final part, the `PfailCalc` tool identifies the MPFP and evaluates the P_{FAIL} according to the failure criterion, as described in Equation 3.2.

Subsequently, we presented the results of the MPFP implementation after approximately three years of operation time. We also demonstrated the mean values and the standard deviations of the Gaussian distributions that the ΔV_{th} of the nFETs and pFETs follow. From these results, it became clear that the impact of time-dependent variability is higher on nFETs. Finally, the results of the MPFP and the P_{FAIL} manifested that the first device, namely M1, has a greater influence on the P_{FAIL}.

CHAPTER 4

A MPFP methodology utilizing the χ^2 distribution

4.1 Introduction

In this Chapter, in order to research the accuracy limits of the previous widely-used technique, we implement a state-of-the-art approach of the MPFP concept [33]. We describe the methodology that we followed for the implementation of the second tool along with the differences with the first one and we explicate the fact that it achieves a better approximation of the set of the failed cases, **F**, and consequently a more accurate estimation of the P_{FAIL} in comparison with the standard MPFP.

In addition, we explain the importance of the concavity of the SNM space in this particular technique. On the assumption that the space is not concave, our methodology is rendered of questionable accuracy, since the identification of the **F** space is based on that. To this end, we locate a local maximum and we extensively examine its nearby space.

Moreover, we present the method of calculating the failure probability by locating the closest failed case to the maximum point that was earlier identified. Then, we demonstrate the distribution of the P_{FAIL} and the cumulative distribution function by applying different SNM specifications and we compare the results of the this technique with the typical MPFP for the same SNM spec. Finally, we use the Monte Carlo methodology to estimate a high value of P_{FAIL} and we compare it with the respective P_{FAIL} of the MPFP to test whether the Monte Carlo can provide accurate results for high probability values.

4.2 Differences between the two approaches

Although, the previous technique (Chapter 3) correctly estimates the most probable failure point, it does not isolate the **F** set accurately. In fact, the P_{FAIL}, which was evaluated at the point $\mathbf{x} = [x_1, x_2, ..., x_N]$ according to the previous methodology, corresponds to the area: $|\Delta V_{th,i}| \geq x_i$ and includes only a part of the failed cases. In view of this, the MPFP technique leads to an optimistic result. Hence, to estimate P_{FAIL} we follow a second approach based on [33], where the MPFP concept is reformulated and a more accurate approximation of the **F** space is proposed.

The first step of this approach is to examine the concavity of the search space, after locating the combination of V_{th} shifts that leads to an extremum of the SNM space. The examination of the concavity of the space is a fundamental step regarding this methodology, since the approximation of the set **F** is based on this assumption. Although, a strict mathematical proof is undoubtedly unfeasible without the

formulation of the SNM, the existence of a local extremum is an evidence that the SNM is indeed concave. In this work we chose to identify a local maximum, considering that the SNM space has a lesser number of local maxima compared to minima and therefore, the complexity of the P_{FAIL} calculation is alleviated. This assumption is based on the fact that the V_{th} shifts of the cell's devices normally are expected to drop the SNM value rather than to increase it.

Subsequently, we continue by locating the ΔV_{th} combination, $\mathbf{x}_Y$, with the lowest distance from the maximum, r_Y, that is considered failed, $SNM(\mathbf{x}) < Y$. To achieve it we set the maximum as starting point and we move towards the direction with the greatest descent of SNM until we reach a local minimum. Since we detect the closest failed point to the maximum, it is clear that all the failed samples are located outside the hypersphere that is centered at the maximum and its radius is equal to r_Y. As a result, this methodology estimates more accurately the **F** space, containing now more failed cases compared to the previous approach. Hence, this leads to a more realistic result regarding the P_{FAIL}.

4.3 Concavity of the SNM space

Locating an extremum in the space under study and having an idea on the concavity of the space is a crucial issue in order to isolate the **F** space accurately. To achieve locating a maximum, we implement the coordinate ascent Algorithm 3 [33]. According to it, for each transistor, we select a positive or negative step, depending on which direction SNM has a greater value, and subsequently we add it to the ΔV_{th} that we are currently deal with, up to the point SNM's value is lower than the previous. This whole process is repeated until there are no points with higher SNM value than the current in any direction of all the involved transistors, in other words, until we locate the maximum in our space. In case of a local minimum identification, we would execute the coordinate descent algorithm and would move each time to the descending direction instead of the ascending. We chose the step value equal to 0.05V and we narrowed ΔV_{th} among -0.25V and 0.25V. The values of SNM were previously calculated using the `SamplesGenerator`, `Cadence Spectre Simulator` and `SNMestimator` tools as described in Section 3.3. The results of the Algorithm 3 are presented in Figure 4.1 step by step.

Algorithm 3 Coordinate Ascent

```
while maximum not found do
    for i=1 to N do
        step=find ascenting direction of SNM for x_i
        repeat
            if currSNM(x_1, ..., x_i + step, ..., x_N) > preSNM then
                update x_i with x_i + step
            end if
        until currSNM < preSNM
    end for
end while
```

For the accuracy of this methodology, it is important to detect all the maximum SNM points. After comparing all the points, we came to the conclusion that the local maximum that was previously identified for:

$$(\Delta V_{th,1}, \Delta V_{th,2}, \Delta V_{th,3}, \Delta V_{th,4}) = (0, -0.05, -0.25, 0.25)$$

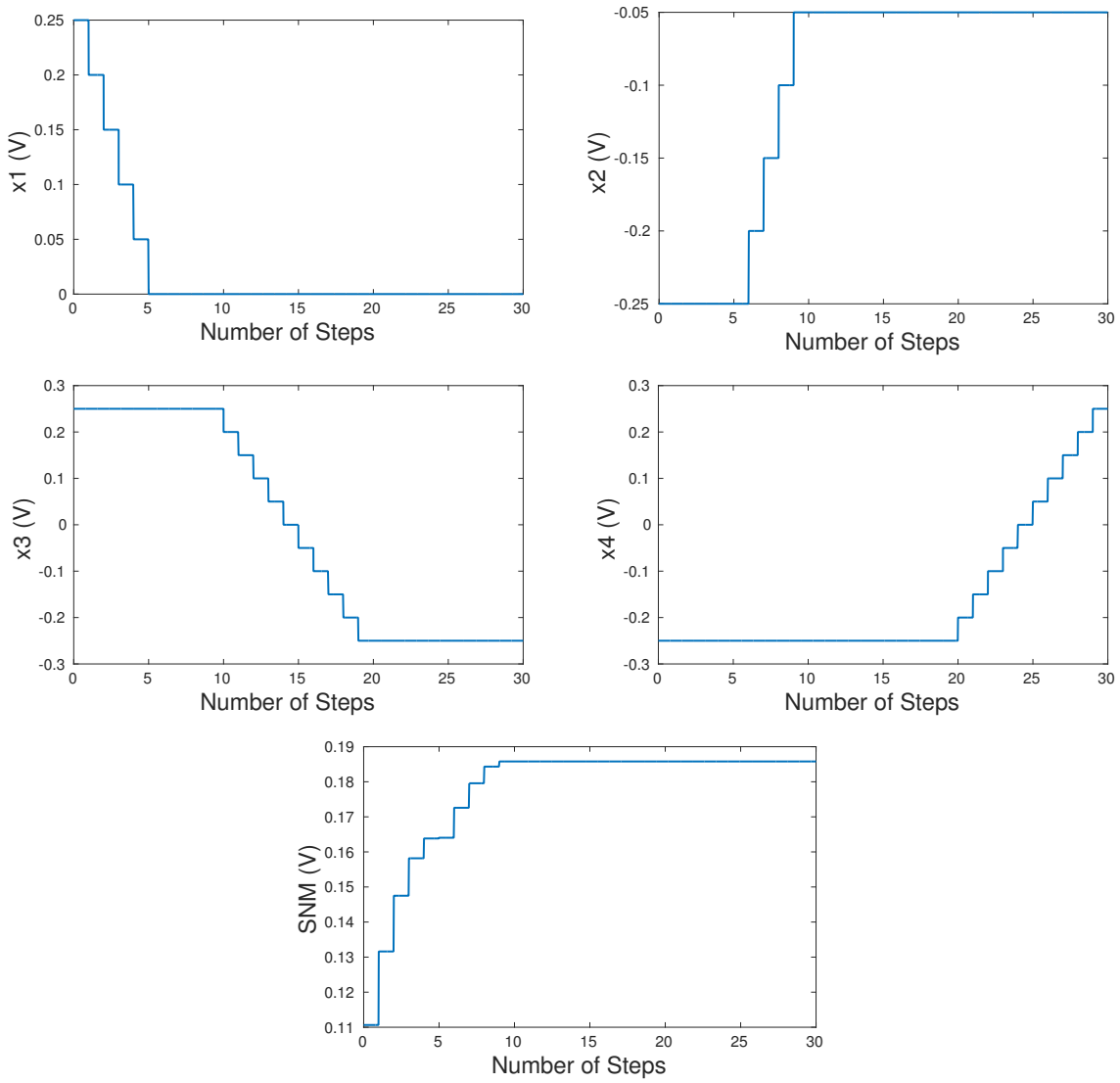


Figure 4.1: Results of the coordinate ascent algorithm implementation

is the global maximum of SNM in this particular space we examine. The exhaustive search was performed by developing a tool that compared the aforementioned point with all the other sampled points of the space in terms of SNM value and proved that the detected maximum is the global maximum. Although we assumed that the space is symmetric and therefore we expected its symmetric point $(0, 0.05, 0.25, -0.25)$ to be a local maximum, we proved wrong since there are adjoining samples with higher SNM value.

Figure 4.2 shows a representation of the SNM space near the maximum. The curves depict SNM as we are moving away from the maximum in multiple directions. The parameter r is the distance of ΔV_{th} from the maximum. We observe that not all the curves seem to follow the same trend. In fact, as the ΔV_{th} of the M3 device and especially of the M4 are moving away from the maximum, the SNM shift is minor in comparison with the rapid shift that cause the M1 and M2 transistors. Therefore, when the curves of the $\Delta V_{th,3}$ and $\Delta V_{th,4}$ directions are illustrated in the same scale with these of the first two devices, they seem to be horizontal. As a result, Figure 4.3 is necessary for a more detailed depiction of SNM. Now all the curves confirm that the point we identified is indeed a maximum since the SNM

value for all the specified directions is lessening while we are moving away from it and hence, that the space seems concave.

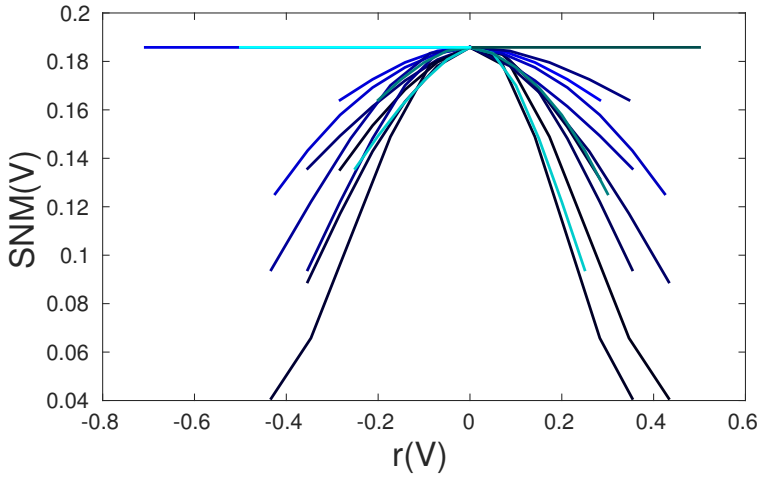


Figure 4.2: Represantation of the SNM space near maximum

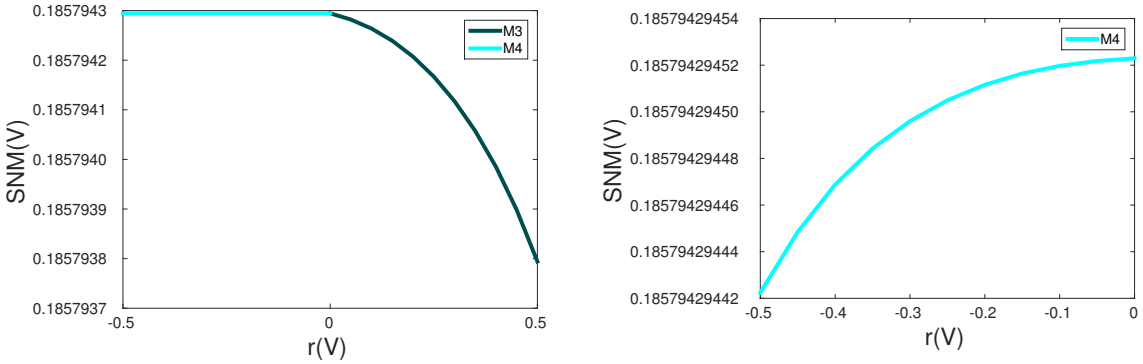


Figure 4.3: Representation of SNM in the direction of $\Delta V_{th,3}$ and $\Delta V_{th,4}$

4.4 Estimation of the Failure Probability

4.4.1 Moving away from the maximum

Since we have detected the only existing maximum SNM of the space under examination, in order to estimate the failure probability, we continue by locating the ΔV_{th} combination, $\mathbf{x}_Y$, with the minimum distance from the maximum, r_Y, that leads to failure. To achieve it we initialize the V_{th} shifts to the values of the maximum and each time we move towards the most descending direction of SNM until we reach a local minimum, according to Algorithm 4. Figure 4.4 illustrates the implementation of the Algorithm.

It has to be noted that this Algorithm follows a different concept than the previous (Algorithm 3). Instead of searching one coordinate at a time, it examines all the coordinates and then moves to the point with the lower SNM value in each step. This methodology, although requires more computations in each iteration, provides a better approximation of the most descending direction of SNM. In Section

4.3 we chose the coordinate ascent Algorithm, since the path that lead the maximum was not of interest.

Algorithm 4 Greatest Descent

```
Initialize x to the value that leads to maximum
while not reach minimum do
    for i=1 to N do
        calculate SNM_i(..., x_i − step, ...)
        calculate SNM_i(..., x_i + step, ...)
        calculate SNM_i(..., x_i, ...)
    end for
    for i=1 to N do
        find the minimum SNM_i
        update x_i with the value that leads to minimum
    end for
end while
```

After the identification of the most descending direction, we utilize the non central chi-square distribution (χ^2) to estimate the P_{FAIL} since it requires a low computational cost. It will provide the probability of the **F** that corresponds to the set in the ΔV_{th} variation space outside of the hypersphere which is centered at the maximum and has radius equal to r_Y [33]. Hence, according to it, we calculate the probability of the random variable z^2, as stated in Equation 4.1 [34], where N is the number of transistors and σ_i are the standard deviations of the transistors, that were earlier computed in Chapter 3. The non-centrality parameter is λ and μ is the maximum point which also includes the mean ΔV_{th} values of each transistor, that were also evaluated in Chapter 3.

$$z^2 = \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} \qquad \lambda = \sum_{i=1}^{N} \frac{\mu_i^2}{\sigma_i^2} \tag{4.1}$$

In case we have located a minimum, we would implement the greatest ascent algorithm instead of the greatest descent and accordingly, we would move towards the most ascending direction until we identify a point x'_Y, with distance from the minimum r'_Y which does not lead to failure. We would use again the χ^2 distribution for the estimation of the failure probability, but the set of the failed cases would be inside the hypersphere that is centered at the minimum point and its radius is equal to r'_Y.

4.4.2 Distribution of failure probability

To estimate the distribution of the P_{FAIL} we use the following methodology. For each point $\mathbf{x}_Y$, with distance from the maximum r_Y, the Algorithm 4 indicated, we calculate the failure probability regarding as Y the respective SNM and using the χ^2 distribution. The P_{FAIL} for each spec Y satisfies Equation 4.2, where CDF and PDF are the cumulative distribution and probability density functions respectively. The results of the P_{FAIL} vs the SNM spec are demonstrated in Figure 4.5. Due to the small number of points, an accurate approximation of the distribution of P_{FAIL} was not possible.

$$P_{FAIL} = 1 - CDF(r < r_Y) = 1 - \int_{-\infty}^{r_Y} PDF_r dr \tag{4.2}$$

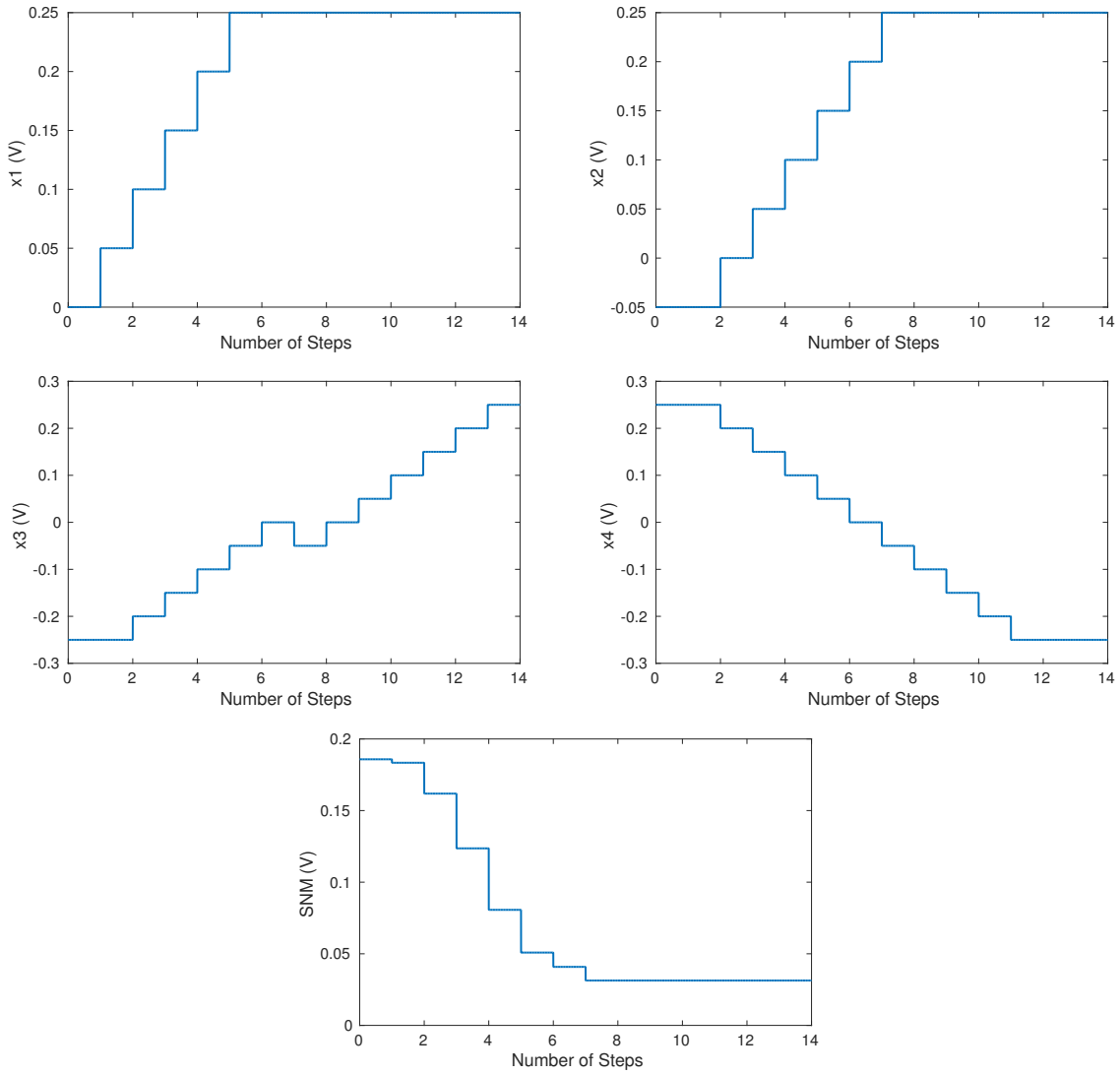


Figure 4.4: Results of the implementation of the Algorithm 4

We can see in Figure 4.5 that the decrement of the spec Y leads to a lower P_{FAIL} and vice versa. This is expected considering that the P_{FAIL} is directly depending on the failure criterion and the SNM spec determines the amount of noise that the cell can tolerate. The decisive role of the SNM spec is obvious for lower Y values, which cause a rapid drop to the P_{FAIL} value.

Comparing the failure probabilities of this technique with the previous one (Chapter 3) for the same SNM spec Y we come to the conclusion that the standard MPFP leads to a much more optimistic result. Specifically, the P_{FAIL} with the first methodology was estimated 9.9×10^{-12} while with the reformulated MPFP it is calculated 1.3×10^{-10}. This difference is a result of the insufficiency of the typical MPFP concept to approximate **F**. In fact, it isolates only the points with V_{th} shifts higher than these of the located MPFP. On the contrary, the new MPFP approach provides a more accurate evaluation of **F**. Regarding that we move away from the maximum towards the direction with the greatest descent, the χ^2 formulation can distinguish all samples $\mathbf{x}$ with $SNM(\mathbf{x}) < Y$ in each step.

It should be pointed out that in larger variation spaces there might be multiple local maxima or minima. These can be identified by implementing Algorithm 3 (or coordinate descent algorithm for

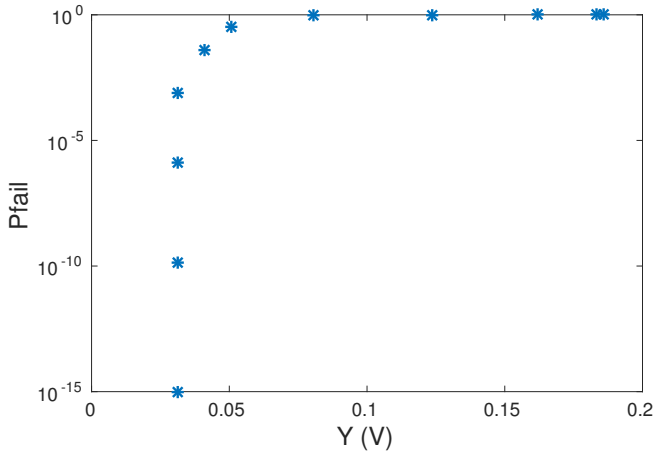


Figure 4.5: The failure probability of the SRAM cell for various values of SNM specification (threshold).

minima) with different starting points each time. In this instance, the process of Section 4.4.1 has to be repeated for all of them in order to have an accurate estimation of the P_{FAIL} distribution. Otherwise, in case of several maxima, the **F** space will include the rest of the maxima and thus, a lot of passed cases which will lead to a very pessimistic result. Accordingly, in case of minima, we will end up with a very optimistic P_{FAIL}.

4.5 Comparing the failure probability of the MPFP against the Monte Carlo

The Monte Carlo technique has been largely employed to evaluate the reliability of CMOS designs [16],[17],[9]. Although, it provides accurate results for older technologies, when it comes to modern heavily scaled devices, it fails to deliver an accurate approximation of very low failure probability. In this Subsection we test the accuracy of this methodology for a high value of P_{FAIL} for a FinFET-based SRAM cell. We use the following simulation framework (Figure 4.6) to calculate P_{FAIL} and compare it with the result of the reformulated MPFP method.

We first engender the aged netlists using the `tz_var` and `td_var` tools. We add time-zero variability to a population of fresh SRAM netlists with the `tz_var` and subsequently, we "inject" time-dependent variability with the `td_var` tool to the derived netlists. Afterwards, we produce the butterfly curves of each instance with the `Cadence Spectre Simulator` and we assess the SNM using the `SNMestimator` tool as described in Section 3.3. The P_{FAIL} is equal to the number of failed cases (according to Equation 3.2) to the total number of cases (Equation 2.5).

We implemented this methodology again after approximately three years of operation and we focused on the four devices of the cross-coupled inverters. The histogram of the SNM is presented in Figure 4.7i along with the spec Y in comparison with the respective P_{FAIL} of the MPFP (Figure 4.7ii). We chose a high Y value for this comparison in order to test the accuracy of Monte Carlo for high P_{FAIL} values. The estimated failure probability between the two methodologies for the specified Y is rather insignificant as seen from Figure 4.7.

For lower probabilities the Monte Carlo is not considered efficient, especially for modern technolo-

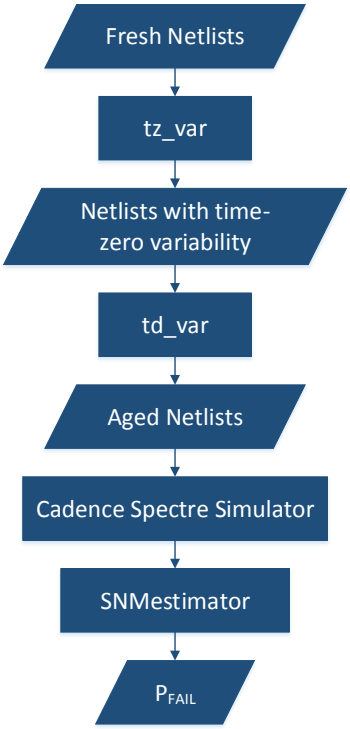
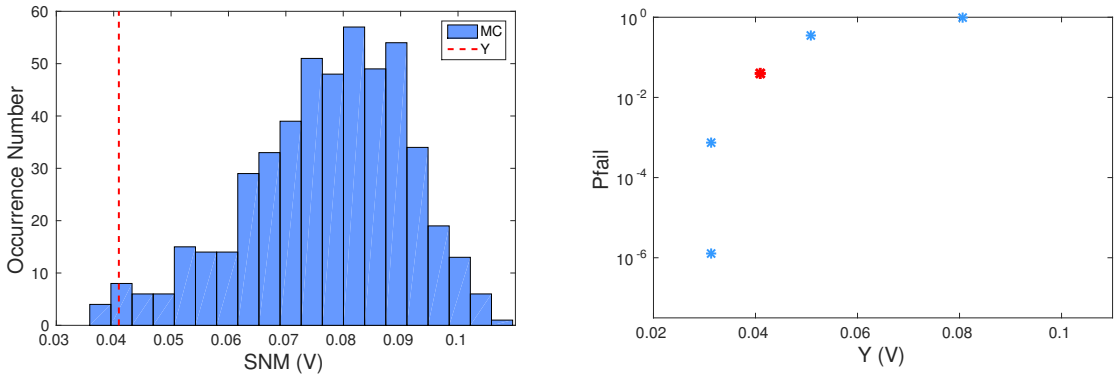


Figure 4.6: Simulation framework of the Monte Carlo technique.



(i) The histogram of SNM along with a high Y value. **(ii)** The P_{FAIL} of MPFP method for the respective Y.

Figure 4.7: A comparison between the results of Monte Carlo and MPFP for a high value of spec Y.

gies. First, the computational cost rises dramatically, while the probability decreases. For example, for the calculation of a probability equal to 10^{-5}, the number of simulations are at least 10^6, which makes the Monte Carlo inefficient. This fact introduces large inaccuracy for extremely low failure probabilities, which is confirmed from our computations, since the number of failed samples is zero when we choose a more realistic spec as shown in Figure 4.8. To alleviate prohibitive computation times, engineers work a specific population of Monte Carlo samples and fit the output estimations to a known distribution (usually the Gaussian) [27]. However, the SNM metric is observed to have a rather prolonged tail towards lower values [44] that the aforementioned fitting fails to capture.

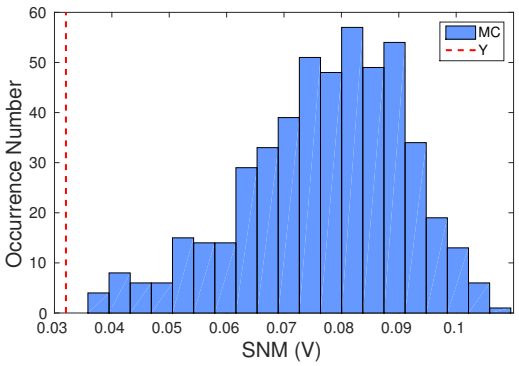


Figure 4.8: The histogram of SNM along with a realistic Y value.

4.6 Conclusions

Previously we evaluated the failure probability of a SRAM cell with the widely-used MPFP technique. In this Chapter we implemented a state-of-the-art approach of the MPFP concept which can provide a more accurate approximation of the set of the failed samples **F** and hence, of the P_{FAIL}. First, we described the process that we followed and the differences with the first approach in the isolation of the **F**.

For the accuracy of this methodology the concavity of the SNM space is of major importance. Consequently, we focused on locating a maximum point in terms of SNM value in the ΔV_{th} variation space using the coordinate ascent algorithm. Then we applied an exhaustive search to verify that this space has a unique maximum and we extensively examined the concavity of the space near it by moving to multiple directions from it. It was also noticed that we chose to identify a maximum instead of a minimum considering that the value of SNM is more likely to diminish rather than to increase as the ΔV_{th} becomes higher.

Furthermore, in order to isolate the set **F** we moved away from the maximum to the most descending direction of the SNM according to the greatest descent algorithm, until we reached a point that leads to the failure of the cell. The **F** was approximated as the set of the points outside of the hypersphere that is centered at the maximum and its radius is equal to the distance of the point that the greatest descent algorithm located from the maximum. Afterwards, we used the χ^2 distribution to calculate the P_{FAIL}.

Finally, to approximate the distribution of the P_{FAIL}, we repeated the above estimation for various SNM specifications (thresholds). Specifically, we calculated the P_{FAIL} for all the points that the greatest descent algorithm indicated regarding each time as Y the SNM of the respective point. Moreover, we verified that this technique does indeed a more accurate evaluation of the **F** space by comparing the estimated P_{FAIL} with this of the previous methodology for the same SNM spec. In addition, we assessed the P_{FAIL} with the Monte Carlo approach for a high value of spec Y to confirm that it coincides with the respective P_{FAIL} of the MPFP and thus, the Monte Carlo provides accurate results for high probability values.

CHAPTER 5

Conclusions

5.1 Overall

The estimation of reliability of today's aggressively downscaled devices has become a significant challenge for VLSI design. In order to engender an accurate result, both time-zero and time-dependent variability effects have to be taken under consideration. The BTI is the primary aging phenomenon which leads to the spread of transistor's parameters and can cause system failures due to the temperature and voltage stress. As a result of the augmenting frequency of system failures, many methodologies have been developed that calculate the failure probability as accurate as possible, while aiming to keep the computational expense to the minimum. However, when it comes to modern devices either they lack of precision or they require a tremendous amount of simulations.

The purpose of this thesis was to research for an efficient methodology that can accurately evaluate the failure probability of modern hardware components. We examined the variation of the V_{th} of a FinFET-based SRAM cell under time-zero and time-dependent variability. We chose to study a SRAM cell since memories are heavily susceptible to performance fluctuation and degradation. We approximated the P_{FAIL} of the cell with two different techniques, first we used the standard MPFP concept and then we compared the results with a state-of-the-art approach of the MPFP which enhanced the accuracy.

In this work we explained the reasons that induce the variation of a system's parameters. We elaborated on the effects of time-zero variability during the manufacturing process and the BTI effect which is the dominant source of time-dependent variability. The two main models that explicate it were discussed. First, the RD model and its insufficiency to capture the degradation mechanism of modern devices and then the atomistic theory which focuses on the stochastic nature of the defects that are created in the substrate and depending on the stress period, they can be occupied and cause a shift to the V_{th} or not. Our work is based on the second model, since it describes BTI more accurately.

Furthermore, we discussed the shift of the semiconductor industry to multi-gate device architectures and specifically the FinFET device in order to continue the shrinking of transistors dimensions. We compared it with the established planar-FET and we pointed out its advantages. We also, described the SNM metric for the hold operation, which is the criterion that we chose to define the stability of the cell. In addition, some of the existing techniques of estimating the failure probability are overviewed.

Subsequently, we presented the simulation framework of the standard MPFP methodology. Since system failures are considered a rare event, the objective of the MPFP is to locate the point that leads to

failure and has the highest probability of occurrence. To achieve it, first we generated the variation space of the ΔV_{th} and the aged netlist with the `SamplesGenerator` tool. To minimize the computational time we focused on the four transistors of the cross-coupled inverters. Afterwards, the `Cadence Spectre Simulator` processed the aged netlist and produced the butterfly curves of each instance which the `SNMestimator` tool used to engender the respective SNM values. The final tool is the `PfailCalc` which first calculated the mean and standard deviation values of nFETs and pFETs and then located the MPFP and estimated the P_{FAIL}. The results of the implementation of this methodology indicated that although both nFETs and pFETs perform the same time-zero variability, the nFETs are more vulnerable to the BTI phenomenon and moreover, the influence of the M1 device (Figure 3.2) on the P_{FAIL} is higher.

Moving to the next Chapter, we deliberated a reformulation of the standard MPFP technique, in order to search the accuracy limits of the previous methodology. First, we examined the concavity of the SNM space, since it is essential for the accuracy of the results. In this direction, we used the coordinate ascent algorithm to locate a maximum and we extensively studied the space near it, which allowed us to demonstrate a representation of the concavity of the SNM space. To isolate the set of the failed cases, **F**, we continued by identifying the closest point to the maximum that leads to failure with the greatest descent algorithm. We utilized the χ^2 distribution to approximate the **F** and calculate the P_{FAIL}. By repeating this process for several SNM spec (thresholds) we estimated the distribution of the P_{FAIL}. We also discussed the case of multiple extrema and the methodology we would use. In addition, the comparison between the results of the two techniques, indicated that the second state-of-the-art approach offers a better evaluation of the **F** set and consequently, a more accurate assessment of the P_{FAIL}. Finally, we tested the accuracy of the Monte Carlo technique for high P_{FAIL} values by comparing its result with the respective one of the reformulated MPFP method.

5.2 Future Work

In this thesis we implemented two different techniques, the typical MPFP and an alternative approach of this concept, which led to a much more realistic result than the first one. However, there are some improvements that can be taken into consideration to enhance the computational cost and the accuracy of the second technique. Hence, in the following Subsections some topics are discussed including the parallelization of the computations in order to reduce the required simulation time, the extension of our tool to incorporate not only the tolerance of the cell to noise during the hold operation but also during the write operation using the Write Trip Point (WTP) metric and finally, the implementation of the second technique for basic gate components.

5.2.1 Parallelization of the SNM estimation

The study of the concavity of the SNM in the ΔV_{th} variation space can be a time-consuming procedure and constitutes an obstacle to a more detailed analysis. The examination of the whole space requires a large amount of computations, especially if we include the spread of all six devices and not only of the four of the cross-coupled inverters. Particularly, the processes that introduce a large delay are the two DC analysis that the `Cadence Spectre Simulator` performs to engender the butterfly curves for each instance from which the SNM derives. Although, each one of these analysis does not

demand a large amount of time, the total time is very high due to the massive number of the ΔV_{th} points.

In order to reduce the execution time, the simulations of the `Cadence Spectre Simulator` can be parallelized. The absence of dependence between the points of the ΔV_{th} space allows the simultaneous process of the netlist for different aging levels. As a result, for the further shrinking of the consuming time, the evaluations of the SNM for each point can also be parallelized. Thus, depending on the number of cores, the total time for the evaluation of SNM can be diminish, which can lead to the decrease of the step of the V_{th} shifts and eventually to the increment of the accuracy. In addition, it can help to the extension of our tool for bigger circuits.

5.2.2 Incorporation of the Write Trip Point metric

Besides the stability of the SRAM cell during the hold operation, it is equally important the guarantee of the write ability while spending the minimum amount of energy. Therefore, the cell should have a reasonable WTP which defines the maximum voltage that is required to flip the content of the cell [45] (Figure 5.1). Higher values of the WTP lead to the rise of the needed power for the write operation and consequently the increase of the cell stability. Thus, a realistic WTP threshold should be chosen that can balance the write stability and the low power consumption.

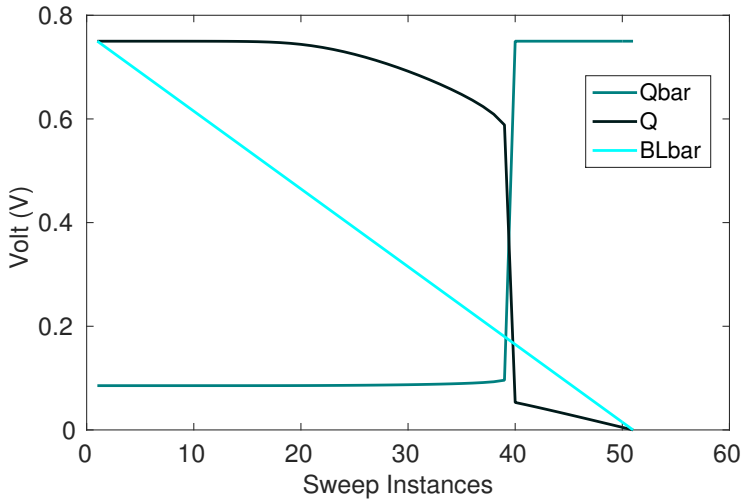


Figure 5.1: Example of Write Trip Point

To have an even more accurate result of the failure probability, the failure criterion should be determined by both SNM and WTP. In order a test case to be considered passed, SNM and WTP should be above the respective thresholds. The procedure for the estimation of the WTP is similar to this of the SNM. First the aged cell is simulated and then the WTP is calculated as the maximum bit-line voltage needed to flip the state of the cell, as shown in Figure 5.1. In previous work [10], the P_{FAIL} was evaluated based on both of these metrics, however, it was calculated with the Quasi-Monte Carlo technique that fails to deliver an accurate result for today's technologies.

5.2.3 Estimation of P_{FAIL} of basic gate components

An interesting work would be the assessment of the P_{FAIL} of other basic hardware components such as logic gates. Although such an approach was also proposed in [33], estimating failure probabilities of basic logic gates and propagating such estimations to standard logic paths has not yet been studied. In this case, the metric under study can be the delay of each gate and the concavity of the delay space has to be validated.

This being said, we can move from the evaluation of the failure probability of memories to processors. However, complex circuits will probably have multiple minima and hence, they should all be included in the estimation of the **F**. Otherwise, the accuracy of this implementation could significantly decrease and end up with a pessimistic result.

Bibliography

[1] L. Chang, et al., "Stable SRAM cell design for the 32nm node and beyond", *Digest of Technical Papers. 2005 Symposium on VLSI Technology, 2005.* pp. 128–129, June 2005.

[2] R. Kanj, et al., "Mixture Importance Sampling and its application to the analysis of SRAM designs in the presence of rare failure events", *Design Automation Conference, 2006 43rd ACM/IEEE*, July 2006.

[3] S. Borkar, et-al., "Parameter variations and impact on circuits and microarchitecture", *Design Automation Conference, 2003. Proceedings*, pp. 338–342, June 2003.

[4] M. Rana, R. Canal, "Ssfb: A highly-efficient and scalable simulation reduction technique for SRAM yield analysis", *Design, Automation and Test in Europe Conference and Exhibition (DATE), 2014*, March 2014.

[5] D. Hardy, et al., "The performance vulnerability of architectural and non-architectural arrays to permanent faults", *2012 45th Annual IEEE/ACM International Symposium on Microarchitecture*, pp. 48–59, Dec. 2012.

[6] E. Seevinck, et al., "Static-noise margin analysis of MOS SRAM cells", *IEEE Journal of Solid-State Circuits*, vol. 22, pp. 748–754, Oct. 1987.

[7] V. B. Suresh, S. Kundu, "On analyzing and mitigating SRAM BER due to random thermal noise", *VLSI (ISVLSI), 2013 IEEE Computer Society Annual Symposium on*, pp. 159–164, August 2013.

[8] N. Rahman, B.P. Singh, "Static-Noise-Margin Analysis of Conventional 6T SRAM Cell at 45nm Technology", *International Journal of Computer Applications*, vol. 66, pp. 19–23, March 2013.

[9] T.S. Doorn, et al., "Importance sampling Monte Carlo simulations for accurate estimation of SRAM yield", *Solid-State Circuits Conference, 2008. ESSCIRC 2008. 34th European*, Sept. 2008.

[10] M. Noltsis, et al., "Accuracy of Quasi-Monte Carlo technique in failure probability estimations", *IC Design and Technology (ICICDT)*, June 2016.

[11] M. Rana, R. Canal, "REEM: Failure/non-failure region estimation method for SRAM yield analysis", *2014 IEEE 32nd International Conference on Computer Design (ICCD)*, Oct. 2014.

[12] P. Weckx, et al., "Implications of BTI-Induced Time-Dependent Statistics on Yield Estimation of Digital Circuits", *IEEE Transactions on Electron Devices*, vol. 61, pp. 666–673, Jan. 2014.

[13] K.J. Kuhn, et al., "Process Technology Variation", *IEEE Transactions on Electron Devices*, vol. 58, pp. 2197–2208, April 2011.

[14] P. Heremans, et al., "Consistent model for the hot-carrier degradation in n-channel and p-channel MOSFETs", *IEEE Transactions on Electron Devices*, vol. 35, pp. 2194–2209, August 2002.

[15] J.W. McPherson, et al., "Time dependent dielectric breakdown physics – Models revisited", *Special Issue 23rd European Symposium on the Reliability of Electron Devices, Failure Physics and Analysis*, vol. 52, pp. 1753–1760, Sept.–Oct. 2012.

[16] Jian Wang, et al., “SRAM parametric failure analysis”, *Design Automation Conference, 2009. DAC ’09. 46th ACM/IEEE*, July 2009.

[17] A. Khosropour, et al., “Process variation tolerant sram cell design using additive model considering NBTI effect”, *2012 4th Asia Symposium on Quality Electronic Design (ASQED)*, July 2012.

[18] V. Subramanian et al., “Planar Bulk MOSFETS Versus FinFETs: An Analog/RF Perspective”, *IEEE Transactions on Electron Devices (IEEE T ELECTRON DEV)*, vol. 53, pp. 3071–3079, Dec. 2006.

[19] Xuejue Huang et-al., “Sub 50-nm FinFET: PMOS”, *International Electron Devices Meeting 1999. Technical Digest*, Dec. 1999.

[20] S. Wasson, “Inside Intel’s Atom C2000-series ’Avoton’ processors”, *The Tech Report*, Sept. 2013.

[21] “14nm FinFET Technology - Not All FinFETs are Created Equal.” `http://www.samsung.com/semiconductor/foundry/process-technology/14nm/`.

[22] “14nm FinFET Leading Edge Technologies”, `http://www.globalfoundries.com/technology-solutions/leading-edge-technology/14-lpe-lpp`.

[23] S. Ganapathy, et al., “INFORMER: An integrated framework for early-stage memory robustness analysis,” in DATE, March 2014, pp. 1–4.

[24] Toledano-Luque M. et al., ”Degradation of time dependent variability due to interface state generation,” VLSIT, 2013, pp. T190–T191.

[25] H. Kukner et al., ”Scaling of bti reliability in presence of time-zero variability,” IEEE IRPS, June 2014, pp. CA.5.1–CA.5.7.

[26] D. Khalil et al., “Sram dynamic stability estimation using MPFP and its applications,” Microelectronics J., vol. 40, no. 11, pp. 1523–1530, Nov. 2009.

[27] D. Khalil et al., “SRAM dynamic stability estimation using MPFP”, *2007 Internatonal Conference on Microelectronics*, Dec. 2007, pp. 167–170.

[28] A. Bhavnagarwala et al., “Fluctuation limits amp; scaling opportunities for CMOS SRAM cells,” *Electron Devices Meeting, 2005. IEDM Technical Digest. IEEE International*, Dec. 2005, pp. 659–662.

[29] D. Rodopoulos et al., “Sensitivity of SRAM Cell Most Probable SNM Failure Point to Time-Dependent Variability,” *SELSE-11 Silicon Errors in Logic - System Effects*, March 31–April 1, 2015 Austin, TX.

[30] “Designing with FinFETs.” `https://www.semiwiki.com/forum/content/1709-designing-finfets.html/`.

[31] D. Hisamoto, et al., “Finfet—a self-aligned double-gate mosfet scalable to 20 nm”, *IEEE Transactions on Electron Devices*, vol. 47, pp. 2320–2325, Dec. 2000.

[32] “Circuit-characteristics analysis system capable of reflecting lithography patterns.” `http://phys.org/news/2013-06-circuit-characteristics-analysis-capable-lithography-patterns.html`.

[33] D. Rodopoulos et al., “Approximating standard cell delay distributions by reformulating the Most Probable Failure Point”, *ERMAVSS 2016 Early Reliability Modeling for Aging and Variability in Silicon Systems*, Dresden–Germany, March 18 2016, pp. 13-16.

[34] N.L. Johnson, et al., “Continuous Univariate Distributions”, Volume 2, May 1995.

[35] “Predictive Technology Model (PTM)”, http://ptm.asu.edu/

[36] “Spectre Circuit Simulator”, https://www.cadence.com/content/cadence-www/global/en_US/home/tools/custom-ic-analog-rf-design/circuit-simulation/spectre-circuit-simulator.html#section6

[37] B. Kaczer et al., “Atomistic approach to variability of bias-temperature instability in circuit simulations”, *Reliability Physics Symposium (IRPS), 2011 IEEE International*, pp. XT.3.1 – XT.3.5, 2011.

[38] D. Rodopoulos, et al., “Atomistic pseudo-transient BTI simulation with inherent workload memory”, *IEEE, Transactions on Device and Materials Reliability*, pp. 704 – 714, June 2014.

[39] A. Avizienis, “Basic concepts and taxonomy of dependable and secure computing”, *IEEE, Transactions on Dependable and Secure Computing*, vol. 1, no. 1, pp. 11–33, 2004.

[40] T. Grasser, et al., “The paradigm shift in understanding the bias temperature instability: From reaction–diffusion to switching oxide traps”, *IEEE, Transactions on Electron Devices*, vol. 58, no. 11, pp. 3652–3666, 2011.

[41] F. N. Najm, N. Menezes, and I. A. Ferzli, “A yield model for integrated circuits and its application to statistical timing analysis”, *IEEE, Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 26, pp. 574 – 591, March 2007.

[42] S. V. Shinde, “Intel-22nm squelch yield analysis and optimization”, *International MultiConference of Engineers and Computer Scientists*, vol. II, March 2014.

[43] S. Mukherjee, *Architecture Design for Soft Errors*. 2008.

[44] X. Wang, et al., “Statistical variability and reliability and the impact on corresponding 6T-SRAM cell design for a 14-nm node SOI FinFET technology”, *IEEE Design & Test*, vol. 30, pp. 18–28, Dec. 2013.

[45] E. Grossar, et al., “Read Stability and Write-Ability Analysis of SRAM Cells for Nanometer Technologies”, *IEEE Journal of Solid-State Circuits*, vol. 41, pp. 2577–2588, Oct. 2006.