ΕΘΝΙΚΟ ΜΕΤΣΟΒΕΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ
ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Analysis of Performance Variation in 16nm FinFET FPGA Devices

Διπλωματική Εργασία
του
Έντρι Τάκα

Επιβλέπων: Δημήτριος Σούντρης
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2019

ΕΘΝΙΚΟ ΜΕΤΣΟΒΕΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Analysis of Performance Variation in 16nm FinFET FPGA Devices

Διπλωματική Εργασία

του

Έντρι Τάκα

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

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 19η Ιουλίου 2019.

(Υπογραφή)	(Υπογραφή)	(Υπογραφή)
....	....	....
Δημήτριος Σούντρης	Κιαμάλ Πεκμεστζή	Ευάγγελος Χριστοφόρου
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2019

....
Έντρι Τάκα

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

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

Περίληψη

Η εκθετική μείωση των διαστάσεων στην τεχνολογία των ημιαγωγών έχει οδηγήσει σε σημαντικές βελτιώσεις στην ισχύ, την απόδοση και το κόστος των ολοκληρωμένων κυκλωμάτων. Ωστόσο, αυτή η μείωση έχει οδηγήσει σε μη επιθυμητές διακυμάνσεις υλικού, λόγω της δυσκολίας ελέγχου της κατασκευαστικής διαδικασίας. Για το λόγο αυτό, η μεταβλητότητα υλικού καθίσταται ένα από τα μεγαλύτερα ζητήματα σε καινούριες τεχνολογίες προκαλώντας κυρίως διακυμάνσεις στις ηλεκτρικές ιδιότητες των κυκλωμάτων και έχοντας μεγάλη επίδραση στην αξιοπιστία και στην απόδοση των ολοκληρωμένων κυκλωμάτων. Όμως, διακυμάνσεις δεν προκύπτουν μόνο κατά τη διαδικασία κατασκευής αλλά και από μεταβλητότητες στην τάση τροφοδοσίας και την θερμοκρασία, όπως επίσης και από φαινόμενα γήρανσης που προκύπτουν από τη χρησιμοποίηση των ολοκληρωμένων κυκλωμάτων. Επιπρόσθετα, οι παραπάνω διακυμάνσεις αναμένεται να χειροτερεύσουν στις μελλοντικές τεχνολογίες.

Συνεπώς, η μελέτη της μεταβλητότητας των ολοκληρωμένων καθίσταται πολύ σημαντική. Ενώ όλες οι υπολογιστικές πλατφόρμες επηρεάζονται από τις μεταβλητότητες, οι Προγραμματιζόμενες στο Πεδίο Διατάξεις Πύλης (FPGAs) είναι ιδιαίτερης σημασίας λόγω της δυνατότητας επαναπρογραμματισμού τους στο επίπεδο της ψηφιακής σχεδίασης. Η ικανότητά τους αυτή, καθιστά τον προγραμματισμό κάθε συστατικού στοιχείου των FPGA σε πολύ χαμηλό επίπεδο. Εκμεταλλεύομενοι αυτήν την ιδιότητα, μπορούμε να αξιολογήσουμε τη μεταβλητότητα της απόδοσης με τη τοποθέτηση αισθητήρων σχεδιασμένων από το χρήστη, σε όλη την επιφάνεια του FPGA.

Σε αύτη τη μελέτη, εστιάζουμε στην ανάλυση της μεταβλητότητας της απόδοσης σε 16 nm FinFET FPGAs. Κατασκευάζουμε μια αξιολόγηση βασισμένη σε αισθητήρες ταλαντωτή δακτυλίου, οι οποίοι έχουν σχεδιαστεί με διαφορετικά χαρακτηριστικά πόρων του FPGA. Έχοντας ως σκοπό την απόκτηση ακριβών στοιχείων αλλά και την κατανόηση σε βάθος των διακυμάνσεων, διαχωρίζουμε τα συστηματικά και τα στοχαστικά μέρη και παράλληλα μοντελοποιούμε με μαθηματικό τρόπο τις μεταβλητότητες. Επιπροσθέτως, αξιολογούμε τις διακυμάνσεις υπό διάφορες περιβαλλοντικές συνθήκες, δηλαδή τάση τροφοδοσίας και θερμοκρασία, για να καταλάβουμε σε βάθος και να επεξηγήσουμε την επίδρασή τους στις διακυμάνσεις και την απόδοση των κυκλωμάτων.

Τα πειραματικά αποτελέσματα σε τέσσερα Zynq XCZU7EV FPGAs έδειξαν έως 7.3% ενδοψηφιδικές (intra-die) διακυμάνσεις, αυξάνοντας σε 9.9% για συγκριμένες συνθήκες λειτουργίας. Η μελέτη μας έδειξε ότι τα λογικά συστατικά στοιχεία και τα στοιχεία διασύνδεσης που απαρτίζουν τα FGPAs, παρουσιάζουν διαφορετικές διακυμάνσεις, ελαφρώς μη συσχετιζόμενες, κάτι που τονίζει τη σημασία τους στον προσανατολισμό υλοποίησης πιο περίπλοκων μεθόδων/εργαλείων άμβλυνσης των διακυμάνσεων.

Λέξεις Κλειδιά

FPGA, Μεταβλητότητα Υλικού, Ταλαντωτής Δακτυλίου, Αξιοπιστία, Μεταβλητότητα Απόδοσης, SoC FPGA, Θερμοκρασία, Τάση Τροφοδοσίας, Γήρανση

Abstract

The exponential scale down of the semiconductor technology has led to compelling improvements in power, performance and cost. This rapid scale down, however, exacerbated the unintended process fluctuations due to the difficulty in controlling the manufacturing process. Therefore, process variability has become a challenging issue in modern technologies, resulting in deviations of the electrical characteristics of circuits, impacting, mainly, the reliability and performance of chips. Although, variability does not solely occur from manufacturing, but also from fluctuations in supply voltage and temperature, as well as natural wear out phenomena resulting from utilization of chips, called aging effects. In addition, the aforementioned deviations are expected to become even more substantial in the future technology nodes.

Consequently, the study of chip variability becomes substantial. While all computing platforms divulge variability issues, Field Programmable Gate Arrays (FPGAs) are of particular interest due to their reconfigurable nature. This ability enables the programming of each resource at very low level by performing the so-called built-in-self-tests (BISTs). Exploiting this attribute, enables us to assess the actual performance variation by deploying custom sensors across the FPGA fabric.

In this work, we focus on the study of performance variation in 16nm FinFET FPGAs. We formulate a comprehensive assessment methodology based on the well-established ring oscillator sensors, which are designed utilizing diverse resource and delay characteristics. To obtain precise results and to comprehend the nature of the variability, we decouple variability to systematic and stochastic accompanied by adequate mathematical modeling of variations. Additionally, we assess the variability under different environmental conditions, i.e., supply voltage and temperature, to grasp and explain their effect on variability and circuit performance.

The experimental results on four Zynq XCZU7EV show up to 7.3% intra-die variation, increasing to 9.9% for certain operating conditions. Our approach demonstrates that logic and FPGA routing interconnect resources (including metal wires as well as switching transistors) present different variability, slightly uncorrelated, which highlights the necessity on the direction towards implementing more sophisticated mitigation methods/tools.

Keywords

FPGA, Process Variability, Ring Oscillator, Reliability, Performance Variation, SoC FPGA, Temperature, Supply Voltage, Aging

Acknowledgments

First and foremost, I would like to express my gratitude to Professor Dimitrios Soudris for the trust and the opportunity he offered me, to work on such a distinct and remarkable subject. His valuable advice and experience, both in the academic research field and in developing self-motivation skills, was indispensable for the successful attainment of this thesis.

Moreover, I would like to acknowledge the persisting supervision of doctoral researcher Konstantinos Maragos. Without his shrewd suggestions this thesis would not be possible. I would also like to express my appreciation to George Lentaris for his advice on mathematical issues and to Ioannis Stratakos for his software contribution.

Finally, I would like to thank my family and friends for their relentless love and support throughout my studies and life.

Εκτεταμένη Ελληνική Περίληψη

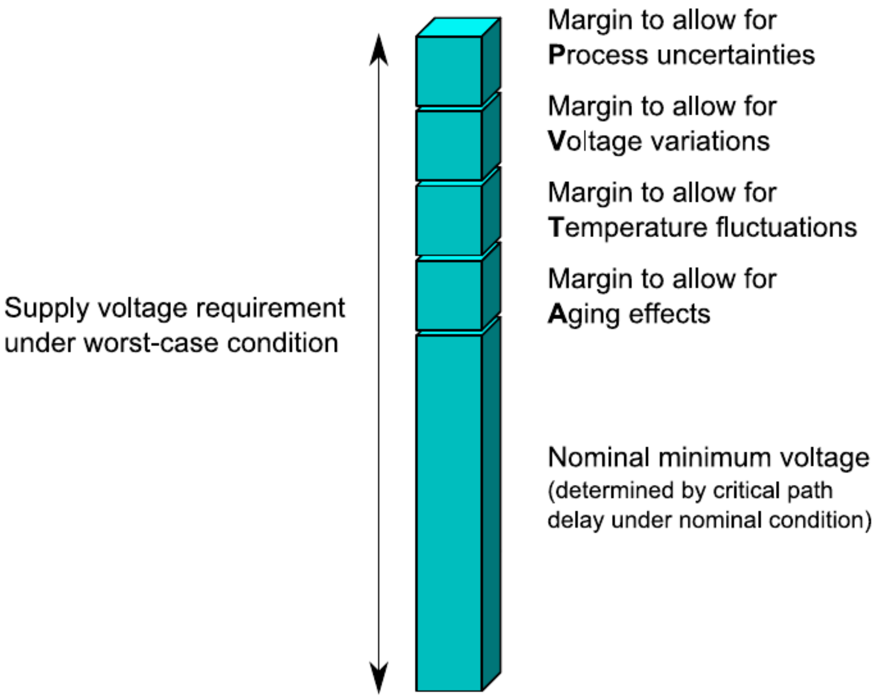
Μεταβλητότητες στα Ολοκληρωμένα Κυκλώματα

Ο νόμος του Moore υποδεικνύει ότι η πυκνότητα των τρανζίστορ ανά ολοκληρωμένο κύκλωμα διπλασιάζεται κάθε 18 μήνες [1]. Αυτό οφείλεται κυρίως στη μείωση των διαστάσεων των τρανζίστορ παρά στην κατασκευή μεγαλύτερων ολοκληρωμένων κυκλωμάτων [2]. Ωστόσο, καθώς οι διαστάσεις μικραίνουν, η αξιόπιστη ολοκλήρωση καθίσταται ως ένα από τα σημαντικότερα προβλήματα. Επομένως, οι μεταβλητότητες που προκύπτουν από την διαδικασία κατασκευής, δηλαδή οι μεταβλητότητες υλικού, αυξάνονται. Οι μεταβλητότητες υλικού προκύπτουν είτε κατά τα πολυάριθμα στάδια στη διαδικασία κατασκευής λόγω ανακριβειών είτε από διακυμάνσεις σε ατομικό επίπεδο των υλικών σε νανομετρική κλίμακα [3]. Αυτό οδηγεί σε μία κύρια κατηγοριοποίηση των μεταβλητοτήτων υλικού σε συστηματικές και στοχαστικές [4–6]. Οι συστηματικές πηγές διακυμάνσεων είναι ντετερμινιστικές και χωρικά εξαρτώμενες, προκαλώντας μία περιοχή του ολοκληρωμένου κυκλώματος να έχει παρόμοιες ηλεκτρικές ιδιότητες. Αντίθετα, οι στοχαστικές διακυμάνσεις είναι μη σχετιζόμενες χωρικά, μη προβλέψιμες πηγάζοντας από ατομικές διαφορές των υλικών.

Συνήθεις πηγές διακυμάνσεων που προκύπτουν από την κατασκευαστική διαδικασία, είναι η φωτολιθογραφία και η χάραξη [7, 8], η τοποθέτηση φωτοευαίσθητου υλικού [5], η εναπόθεση στρώματος φιλμ [8] και η χημική-μηχανική λείανση [9]. Οι διακυμάνσεις που προκύπτουν από ατομικές διαφορές οφείλονται κυρίως σε τυχαίες διακυμάνσεις των ατόμων νόθευσης [10], τραχύτητα γραμμών των άκρων [11] και μεταβλητότητες στο οξείδιο πύλης [12]. Αυτές οι διακυμάνσεις προκαλούν μεταβολές στα μεγέθη των τρανζίστορ, για παράδειγμα στο μήκος του καναλιού, στη συγκέντρωση των ατόμων νόθευσης, κτλ., οι οποίες τελικά μεταφράζονται σε μεταβλητότητες στην καθυστέρηση και στις διαρροές των κυκλωμάτων, αλλά και στην παραγωγή. Έκτος από τις διακυμάνσεις υλικού, οι περιβαλλοντικές διακυμάνσεις είναι ένα εξίσου σημαντικό πρόβλημα στη σχεδίαση ολοκληρωμένων κυκλωμάτων. Περιβαλλοντικές θεωρούνται οι διακυμάνσεις που οφείλονται στην τάση τροφοδοσίας και την θερμοκρασία σε όλη την έκταση του ολοκληρωμένου. Αυτοί οι τύποι διακυμάνσεων έχουν εξάρτηση από το χρόνο αλλά και από το χώρο στο ολοκήρωμένο κύκλωμα. Οι διακυμάνσεις στην τάση έχουν μικρότερες σταθερές χρόνου συγκριτικά με αυτές που οφείλονται στη θερμοκρασία [5, 13], αλλά και μικρότερες χωρικές κατανομές [2], επηρεάζοντας περισσότερο αρνητικά την απόδοση των κυκλωμάτων. Λόγω της εμφάνισης της νανομετρικής τεχνολογίας, η μη ιδεατή μείωση της τάσης τροφοδοσίας και της τάσης κατωφλίου λόγω περιορισμών στα ρεύματα διαρροών [14, 15] προκαλεί αύξηση στα ηλεκτρικά πεδία, γεγονός που επιταχύνει τη γήρανση των κυκλωμάτων. Οι σημαντικότεροι λόγοι γήρανσης των τρανζίστορ είναι η Αστάθεια Θερμοκρασίας Αρνητικής Πόλωσης [16], οι 'Καυτοί

Φορείς' [17] και η Εξαρτώμενη από το Χρόνο Διάσπαση του Διηλεκτρικού [18], ενώ για τα μέταλλα αστοχίες λόγω γήρανσης προκαλούνται κυρίως από το φαινόμενο της Ηλεκτρομετανάστευσης [19].

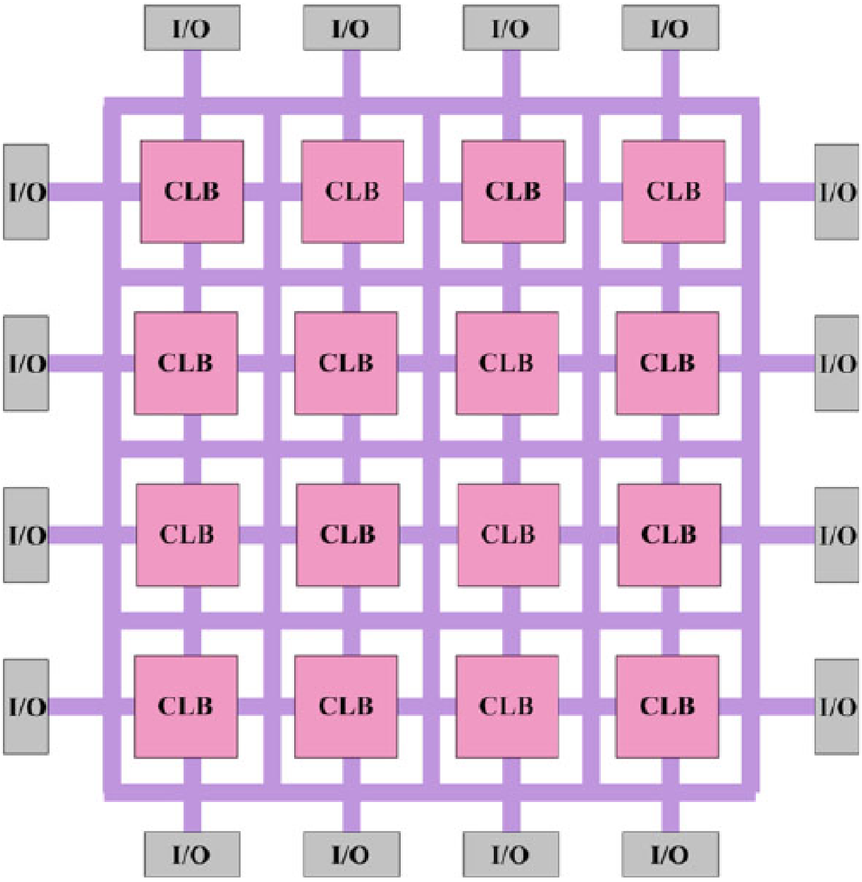
Για την αντιμετώπιση των παραπάνω φαινομένων η παραδοσιακή επίλυση τους βασίζεται στη συμπερίληψη της χειρότερης δυνατής περίπτωσης για τις διακυμάνσεις. Αυτό σημαίνει ότι επαρκή περιθώρια πρέπει να χρησιμοποιηθούν, κυρίως στην τάση τροφοδοσίας και στη συχνότητα λειτουργίας. Το Σχήμα 1 παρουσιάζει τη μεθοδολογία χειρότερης δυνατής περίπτωσης προσθέτοντας περιθώρια τάσης για όλα τους τύπους διακυμάνσεων. Επίσης, με τη μείωση των διαστάσεων τα περιθώρια αυτά αυξάνονται με αποτέλεσμα να οδηγούν σε μη αποδοτικούς σχεδιασμούς. Για το λόγο αυτό, σχεδιάσεις που αμβλύνουν τις διακυμάνσεις γίνονται πολύ σημαντικές. Μια τυπική κατηγορία είναι οι προσαρμοστικές τεχνικές [20]. Με τη μέτρηση διαφόρων παραμέτρων, όπως η τάση τροφοδοσίας και η συχνότητα λειτουργίας, σε ένα κλειστό κύκλωμα ανάδρασης, οι τεχνικές αυτές μπορούν να οδηγήσουν στη μείωση των πεσιμιστικών περιθωρίων.



Σχήμα 1: Περιθώρια τάσης για τη συμπερίληψη της χειρότερης δυνατής περίπτωσης για τις διακυμάνσεις [13].

Κίνητρο και Ερευνητικοί Στόχοι της Διπλωματικής

Ενώ όλα τα ολοκληρωμένα κυκλώματα επηρεάζονται από τις μεταβλητότητες υλικού, οι Προγραμματιζόμενες στο Πεδίο Διατάξεις Πύλης (FPGAs) παρουσιάζουν ιδιαίτερο ενδιαφέρον εξαιτίας της δυνατότητας επαναπρογραμματισμού τους στο επίπεδο της ψηφιακής σχεδίασης. Συγκεκριμένα, η ικανότητα προγραμματισμού των πόρων τους σε πολύ χαμηλό επίπεδο παρέχει τη δυνατότητα σχεδιασμού αισθητήρων από το χρήστη. Επιπρόσθετα, η ομογενής αρχιτεκτονική των πόρων που απαρτίζουν τα FPGAs σε όλη την έκτασή τους, τα καθιστά ικανά για τη μέτρηση των διακυμάνσεων υλικού με τη χρήση ειδικών αισθητήρων που σχεδιάζονται από το χρήστη και τοποθετούνται σε όλη την έκταση του ολοκληρωμένου [21–24].



Σχήμα 2: Αρχιτεκτονική των FPGAs [25].

Το Σχήμα 2 παρουσιάζει την αρχιτεκτονική των FPGAs. Τα θεμελιώδη συστατικά τους είναι τα επαναπρογραμματιζόμενα λογικά μπλοκ (CLBs), οι προγραμματιζόμενοι πόροι διασύνδεσης και τα μπλοκ εισόδου/εξόδου (I/O blocks). Τα CLBs υλοποιούν τις λογικές συναρτήσεις που καθορίζονται από το χρήστη, ενώ οι πόροι διασύνδεσης χρησιμοποιούνται για να συνδέουν τις λογικές συναρτήσεις. Τα I/O blocks υλοποιούν τη σύνδεση του FPGA με τον έξω κόσμο.

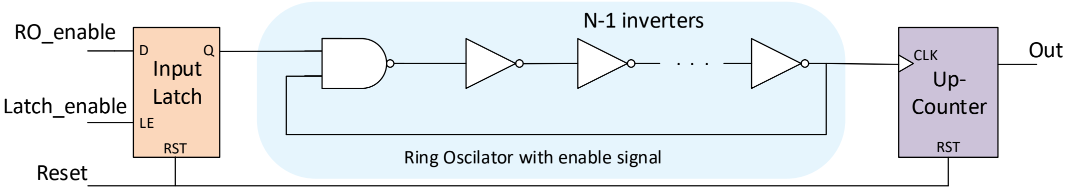
Η συγκεκριμένη διπλωματική εξετάζει τη μεταβλητότητα της απόδοσης σε 16nm FinFET FPGAs (πρώτη στη βιβλιογραφία). Οι κύριοι ερευνητικοί στόχοι είναι:

- Η αξιολόγηση της μεταβλητότητας απόδοσης σε πόρους λογικής και διασύνδεσης με το σχεδιασμό αισθητήρων σε πολύ χαμηλό επίπεδο.

- Ο διαχωρισμός της μεταβλητότητας στο συστηματικό και στοχαστικό της κομμάτι, με σκοπό την ανάλυση της επίδρασης του καθενός στην απόδοση των κυκλωμάτων.

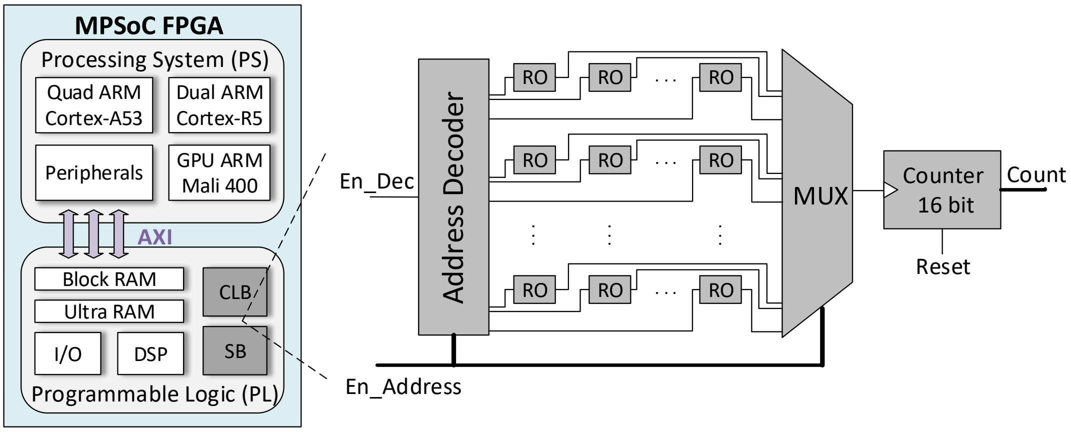
- Η αξιολόγηση της μεταβλητότητας υπό διάφορες περιβαλλοντικές συνθήκες τάσης τροφοδοσίας και θερμοκρασίας.

Μεθοδολογία Σχεδίασης Αισθητήριων Κυκλωμάτων



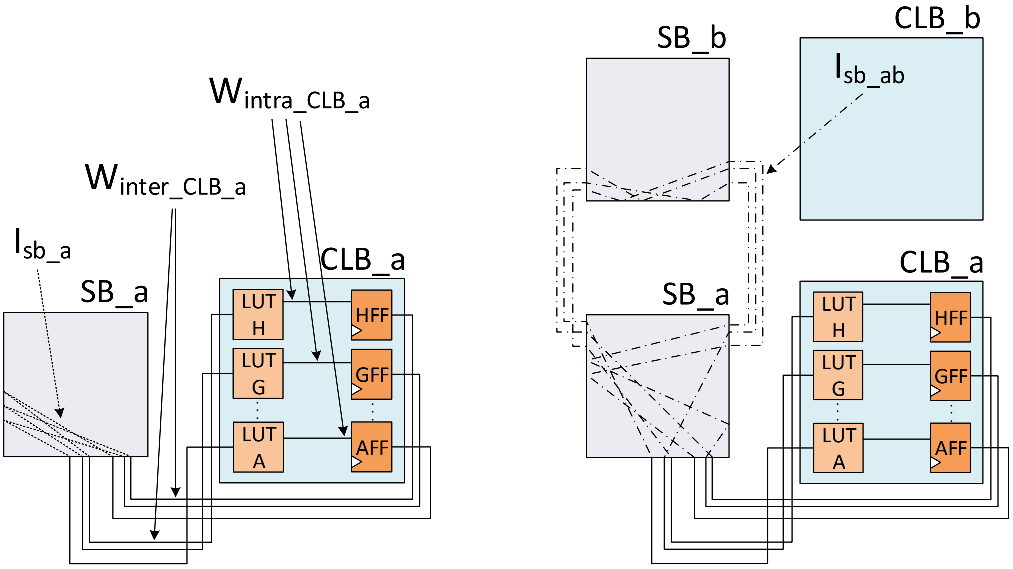
Σχήμα 3: Προτεινόμενη σχεδίαση ταλαντωτή δακτυλίου.

Σε αυτήν την παράγραφο επεξηγούμε τη μεθοδολογία που χρησιμοποιούμε για την ανάλυση της μεταβλητότητας υλικού. Το θεμελιώδες αισθητήριο κύκλωμα που χρησιμοποιήθηκε είναι ο ταλαντωτής δακτυλίου (ring oscillator), όπως έχει προταθεί σε αντίστοιχες εργασίες στη βιβλιογραφία [22, 26, 27]. Ένας ταλαντωτής δακτυλίου κατασκευάζεται από N στάδια πυλών αντιστροφέων, όπου όταν ο αριθμός N είναι περιττός το κύκλωμα ταλαντώνει και στην έξοδο του παράγει μια τετραγωνική κυματομορφή. Ο ταλαντωτής δακτυλίου που σχεδιάσαμε παρουσιάζεται στο Σχήμα 3, όπου μια πύλη NAND έχει χρησιμοποιηθεί για ενεργοποίηση/απενεργοποίηση της ταλάντωσης. Η έξοδος του ταλαντωτή τοποθετείται σε έναν άνω-μετρητή, ο οποίος μετράει τις θετικές ακμές του τετραγωνικού σήματος. Εάν ενεργοποιήσουμε τον ταλαντωτή για συγκεκριμένο χρονικό διάστημα T μπορούμε να μετρήσουμε την καθυστέρησή του.



Σχήμα 4: Μπλοκ διάγραμμα της προτεινόμενης αρχιτεκτονικής.

Με τη σχεδίαση ολόιδιων ταλαντωτών δακτυλίου, ως προς την άποψη των πόρων του FPGA που καταλαμβάνουν, μπορούμε να μετρήσουμε την καθυστέρηση σε διάφορα σημεία στην έκταση του FPGA, οπότε τελικά να μετρήσουμε την μεταβλητότητα απόδοσης. Το προτεινόμενο δίκτυο ταλαντωτών παρουσιάζεται στο Σχήμα 4. Αρχικά πρέπει να αναφέρουμε ότι η μελέτη πραγματοποιήθηκε σε συσκευές MPSoC FPGA οι οποίες παρέχουν τη δυνατότητα ολοκλήρωσης ετερογενών στοιχείων επεξεργασίας στο ίδιο ολοκληρωμένο κύκλωμα, όπως φαίνεται στο αναφερθέν σχήμα. Σε αυτήν την εργασία έχουν μελετηθεί μόνο οι πόροι των CLBs και της διασύνδεσης. Το δίκτυο των ταλαντωτών αποτελείται από ολόιδιους από άποψη πόρων ταλαντωτές, οι οποίοι ενεργοποιούνται σειριακά με τη χρήση ενός αποκωδικοποιητή για την αποφυγή φαινομένων πτώσης τάσεως. Στη συνέχεια η έξοδος ενός εν ενεργεία ταλαντωτή οδηγείται με τη χρήση ενός πολυπλέκτη σε έναν άνω-μετρητή. Η χρονική περίοδος ενεργοποίησης κάθε ταλαντωτή T έχει επιλεχτεί στα 50 μs για την αποφυγή φαινομένων αυτό-θέρμανσης [28] και τη μείωση του συνολικού λάθους στη διαδικασία μέτρησης. Η περίοδος T υπολογίστηκε με τη χρήση timer της ARM CPU.



(α′) “Nst_1sb”

(β′) “Nst_2sb”

Σχήμα 5: Αρχιτεκτονικές των ταλαντωτών δακτυλίου με ολόιδιους πόρους CLB και διαφορετικούς πόρους διασύνδεσης για την ίδια τιμή του N.

Η σχεδίαση των ταλαντωτών έχει πραγματοποιηθεί με διάφορα χαρακτηριστικά πόρων και καθυστερήσεων. Οι θεμελιώδεις αρχιτεκτονικές της σχεδίασής μας παρουσιάζονται στο Σχήμα 5, όπου κάθε στάδιο αντιστροφέα ακολουθείτε από ένα pass-through Flip Flop με σκοπό την αύξηση της καθυστέρησης που οφείλεται σε πόρους λογικής. Η πρώτη θεμελιώδης αρχιτεκτονική (“Nst_1sb”) έχει σχεδιαστεί ώστε να μειώνει την καθυστέρηση που οφείλεται σε διασύνδεση, ενώ η δεύτερη (“Nst_2sb”) χρησιμοποιεί πόρους διασύνδεσης από τα δύο κοντινότερα Switch Boxes (SBs). Η ανάλυση της καθυστέρησης έχει πραγματοποιηθεί με τη βοήθεια του εργαλείου Xilinx Vivado Static Timing Analysis (STA). Ένα σημαντικό χαρακτηριστικό της σχεδίασής μας είναι ότι οι πόροι των CLBs και των δύο αναφερθέντων αρχιτεκτονικών έχουν παραμείνει σταθεροί για την ίδια τιμή του N. Αναπτύσσοντας παραπάνω την προηγούμενη πρόταση, το Σχήμα 5 δείχνει ότι τα μεταλλικά καλώδια $W_{intra\_CLB\_a}$, $W_{inter\_CLB\_a}$ είναι ίδια και στις δύο περιπτώσεις και το ίδιο ισχύει για τους πόρους λογικής, που απεικονίζονται με πορτοκαλί χρώμα. Με τη σχεδίαση αυτή επωφελούμαστε την απομόνωση των πόρων διασύνδεσης καθώς μπορούμε να αφαιρέσουμε τις καθυστερήσεις των δύο παραπάνω ταλαντωτών. Τελικά, μπορούμε να κατασκευάσουμε νέα αισθητήρια κυκλώματα τα οποία τα ονομάζουμε “Nst_inter”. Ο Πίνακας 1 παρουσιάζει τους αισθητήρες που χρησιμοποιούμε στην ανάλυσή μας, όπου η τιμή των σταδίων (N) έχει επιλεχτεί να είναι 5 και 7. Όπως είναι εμφανές από τον πίνακα, οι αρχιτεκτονικές με 1

SB έχουν μεγαλύτερο ποσοστό καθυστέρησης οφειλόμενο σε λογική, π.χ. ο "7st_1sb" έχει 65.4% καθυστέρηση σε λογική και 34.6% σε διασύνδεση, ενώ οι αισθητήρες με 2 SBs έχουν μεγαλύτερο ποσοστό καθυστέρησης σε διασύνδεση, π.χ., ο "7st_2sb" έχει 36.3% καθυστέρηση σε λογική και 63.7% σε διασύνδεση. Οι "Nst_inter" έχουν καθυστέρηση που οφείλεται μόνο σε διασύνδεση και για την ακρίβεια έξω από το CLB (inter-CLB).

Πίνακας 1: Καθυστέρηση με βάση το STA των προτεινόμενων αισθητήρων.

sensor	**delay of logic resources**			**delay of interconnects**			**total (ps)**
conf.	LUTs	DFFs	Total	intra-CLB	inter-CLB	Total	
7st_1sb	707 ps	463 ps	65.4%	295 ps	325 ps	34,6%	1790
7st_2sb	707 ps	463 ps	36.3%	295 ps	1762 ps	63,7%	3227
7st_inter	-	-	-	-	1437 ps	100%	1437
5st_1sb	582 ps	309 ps	67%	196 ps	244 ps	33%	1331
5st_2sb	582 ps	309 ps	37,2%	196 ps	1305 ps	62,8%	2392
5st_inter	-	-	-	-	1061 ps	100%	1061

Διαχωρισμός της Μεταβλητότητας Υλικού

Διαχωρίζουμε την συνολική μεταβλητότητα στο συστηματικό και στο στοχαστικό της κομμάτι με σκοπό να μελετήσουμε την επίδρασή τους ξεχωριστά. Με την παρουσία διακυμάνσεων, η καθυστέρηση μίας κυκλωματικής τοπολογίας μπορεί να εκφραστεί ως μία τυχαία μεταβλητή [29]:

$$T_d = T_d^{\mu} + T_d^{S} + T_d^{R} \tag{1}$$

όπου T_d^{μ} είναι η μέση τιμή, T_d^{S} είναι το συστηματικό κομμάτι και T_d^{R} το στοχαστικό/τυχαίο κομμάτι. Η καθυστέρηση T_d^{μ} είναι μια σταθερή τιμή, ενώ η T_d^{S} είναι χωρικά συσχετιζόμενη αλλάζοντας σταδιακά από περιοχή σε περιοχή του ολοκληρωμένου κυκλώματος και η T_d^{R} δεν έχει χωρική συσχέτιση.

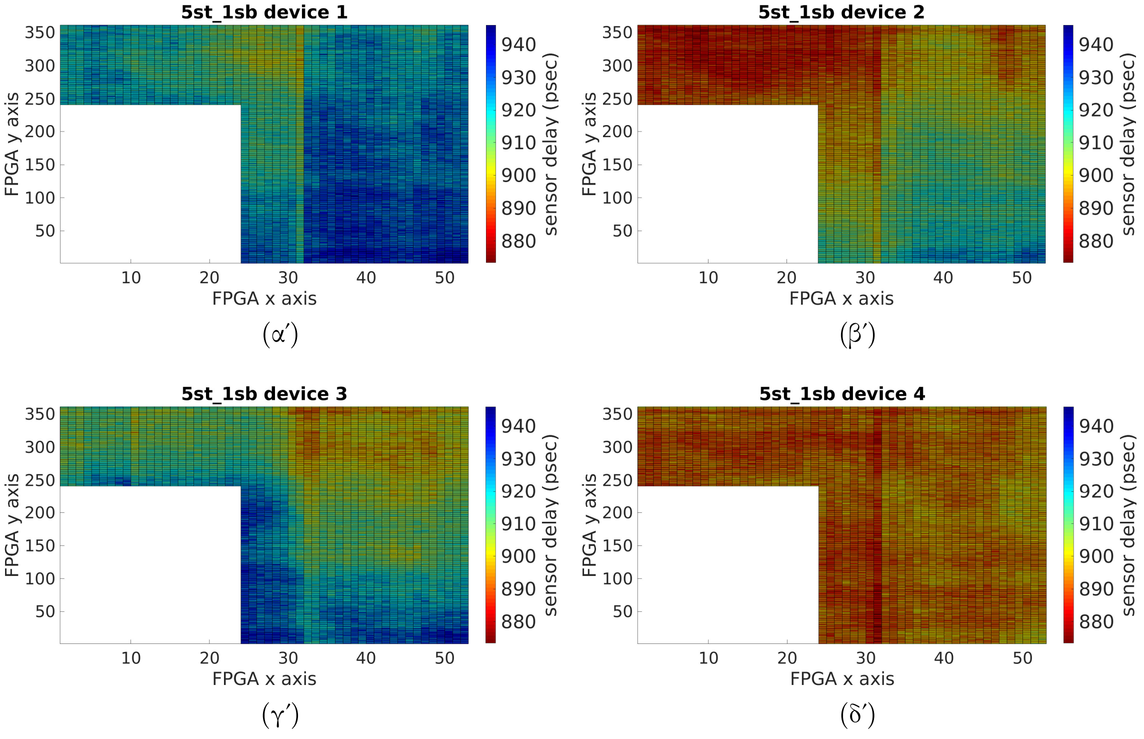
Για τους σκοπούς της ανάλυσής μας χρησιμοποιούμε το grid μοντέλο [30]. Σύμφωνα με αυτό, το FPGA μοντελοποιείται ως ένα X-Y πλέγμα, όπου κάθε σημείο του αναπαριστά έναν αισθητήρα. Το grid μοντέλο υποθέτει ότι η συσχέτιση μεταξύ των συστηματικών διακυμάνσεων όλων των τρανζίστορ και όλων των μεταλλικών καλωδίων είναι τέλεια μέσα σε κάθε σημείο του πλέγματος [30]. Για το λόγο αυτό θεωρούμε ότι όλοι οι πόροι που απαρτίζουν τους αισθητήρες έχουν τέλεια συσχετιζόμενες χωρικές διακυμάνσεις, αφού είναι τοποθετημένοι σε πολύ μικρή απόσταση μεταξύ τους (η υπόθεση μας έχει επαληθευτεί στην πράξη). Με μαθηματικούς όρους, χρησιμοποιούμε τον συντελεστή συσχέτισης (ρ), όπου ο παραπάνω ισχυρισμός μεταφράζεται ότι οι τυχαίες μεταβλητές που εκφράζουν τις συστηματικές διακυμάνσεις της καθυστέρησης μέσα στα όριο ενός σημείου του πλέγματος, έχουν συντελεστή συσχέτισης ακριβώς ίσο με 1. Επιπρόσθετα, το συστηματικό κομμάτι της καθυστέρησης T_d^{S}, λόγω του grid μοντέλου εκφράζεται ως μία συνάρτηση του (x, y), ενώ το στοχαστικό αφού δεν έχει χωρική εξάρτηση, εκφράζεται ως τυχαία μεταβλητή η οποία ακολουθεί την κανονική κατανομή [21, 23].

Ανάλυση Μεταβλητότητας και Αξιολόγηση

Σε αυτήν την παράγραφο παρουσιάζεται η ανάλυση των αποτελεσμάτων με βάση τη μεθοδολογία που περιγράφεται στην προηγούμενη παράγραφο. Η αξιολόγηση της μεταβλητότητας μελετήθηκε σε τέσσερα υποθετικά ίδια Zynq XCZU7EV FPGAs.

Αρχικά αξιολογούμε τη ολική μεταβλητότητα (χωρίς το διαχωρισμό της στο συστηματικό και στοχαστικό της μέρος) σε συνθήκες λειτουργίας: τάση τροφοδοσίας $V_{ccint} = 0.85V$ και θερμοκρασία ολοκληρωμένου $T_j = 30^{\circ}C$. Η διαφορά μεταξύ της μέσης τιμής της μετρούμενης καθυστέρησης και των αποτελεσμάτων του STA (βλ. Πίνακα 1) και για τα τέσσερα FPGAs υπολογίστηκε στο εύρος: 29.9% - 37.6%. Επιπρόσθετα, μεγαλύτερες διαφορές παρατηρήθηκαν στους αισθητήρες διασύνδεσης, που συνεπάγεται ότι το STA είναι πιο πεσιμιστικό για τις καθυστερήσεις των διασυνδέσεων. Στη συνέχεια, χρησιμοποιώντας τη μετρική $(max - min)/min$, όπου max, min είναι η μέγιστη και η ελάχιστη τιμή κάθε υλοποιημένου αισθητήρα, οι ενδοψηφιδικές (intra-die) διακυμάνσεις καθενός FPGA υπολογίστηκαν στο εύρος: 2.62% - 7.3%. Ομοίως, χρησιμοποιώντας την ίδια μετρική, αλλά αυτή τη φορά μεταξύ και των τεσσάρων FPGA υπολογίζουμε τις διαψηφιδικές (inter-die) διακυμάνσεις στο εύρος: 6.44% - 8%.

Στο Σχήμα 6 παρουσιάζονται οι χάρτες ολικής μεταβλητότητας του αισθητήρα “5st_1sb” μεταξύ των τεσσάρων FPGAs. Παρατηρούμε ότι η μορφή των χαρτών είναι διαφορετική για κάθε FPGA, καθώς και το γεγονός ότι σε κάθε χάρτη υπάρχουν συστηματικές περιοχές όπου η απόδοση (καθυστέρηση) είναι παρόμοια.



(α′)

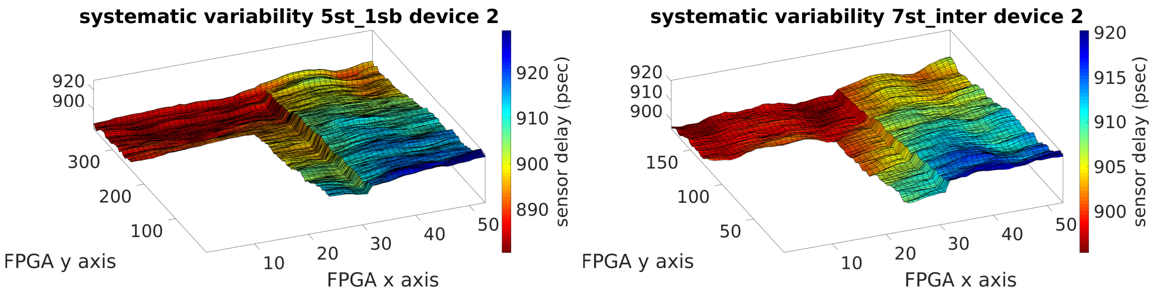
(β′)

(γ′)

(δ′)

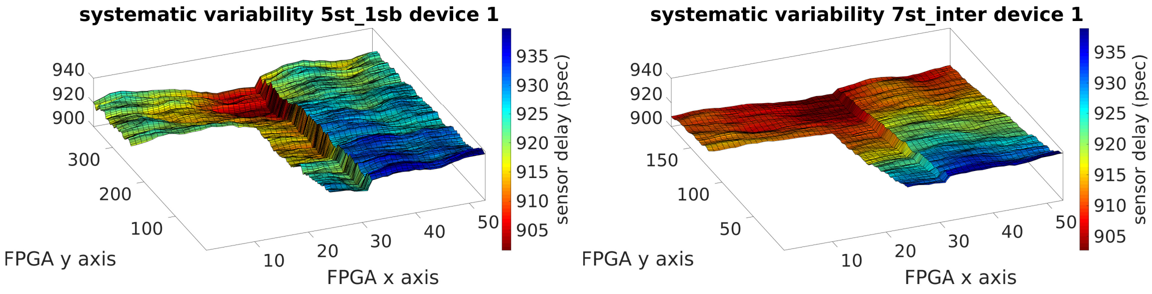
Σχήμα 6: Χάρτες ολικής μεταβλητότητας του αισθητήρα “5st_1sb” μεταξύ των τεσσάρων FPGAs (κοινή κλίμακα).

Το συστηματικό κομμάτι της μεταβλητότητας είναι αρκετά μεγαλύτερο από το στοχαστικό σε όλες τις περιπτώσεις των μετρήσεών μας. Στο Σχήμα 7 παρουσιάζονται οι συστηματικοί χάρτες μεταβλητότητας των αισθητήρων “5st_1sb”, “7st_inter” για ένα



(α′) (β′)

Σχήμα 7: Χάρτες συστηματικής μεταβλητότητας των αισθητήρων “5st_1sb”, “7st_inter” για ένα FPGA, το device 2.



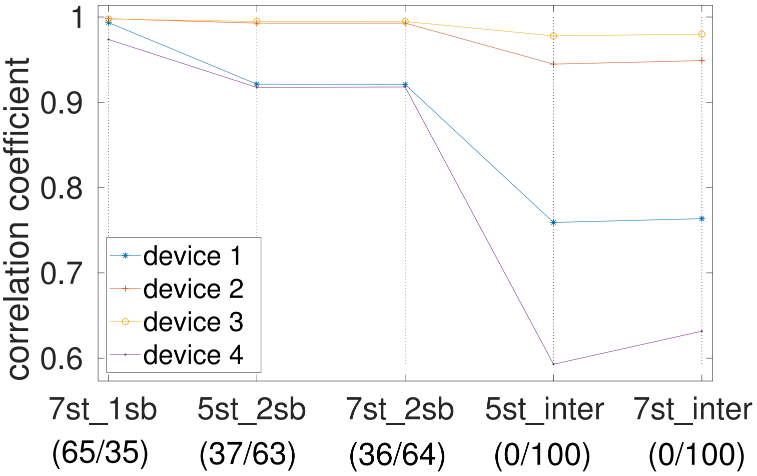
(α′) (β′)

Σχήμα 8: Χάρτες συστηματικής μεταβλητότητας των αισθητήρων “5st_1sb”, “7st_inter” για ένα FPGA, το device 1.

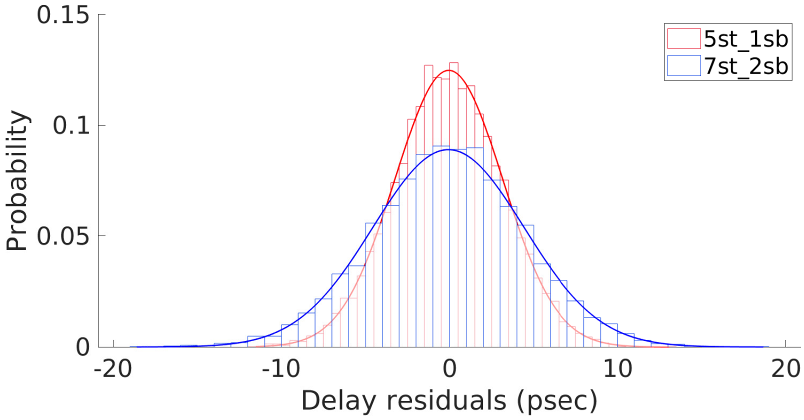
FPGA, ονομαζόμενο ως device 2. Η πρώτη παρατήρηση που μπορούμε να εξάγουμε είναι ότι ενώ οι αισθητήρες έχουν σχεδιαστεί με διαφορετικά χαρακτηριστικά, οι χάρτες τους είναι παρόμοιοι, οπότε το device 2 παρουσιάζει παρόμοιες μεταβλητότητες για τους πόρους λογικής και διασύνδεσης. Αντίθετα, στο Σχήμα 8, όπου οι αντίστοιχοι χάρτες παρουσιάζονται για το device 1, παρατηρούμε ότι οι γρήγορες/αργές περιοχές δεν αντιστοιχούν στα ίδια σημεία για τους δύο χάρτες, οπότε μπορούμε να εξάγουμε το συμπέρασμα ότι το συγκεκριμένο device δεν παρουσιάζει παρόμοιες μεταβλητότητες για τους πόρους λογικής και διασύνδεσης. Από αυτές τις δύο περιπτώσεις γίνεται αντιληπτή η σημασία της μελέτης των διακυμάνσεων κάθε FPGA ανεξάρτητα.

Η ανάλυση των συστηματικών διακυμάνσεων εξετάζεται με τη βοήθεια του συντελεστή συσχέτισης. Το Σχήμα 9 παρουσιάζει το συντελεστή συσχέτισης (Pearson correlation coefficient) έχοντας σαν αναφορά τον αισθητήρα “5st_1sb”. Παρατηρούμε ότι καθώς το ποσοστό (λογική/διασύνδεση) μειώνεται η συσχέτιση μεταξύ των χαρτών μειώνεται φτάνοντας στην τιμή 0.59 για τους αισθητήρες διασύνδεσης.

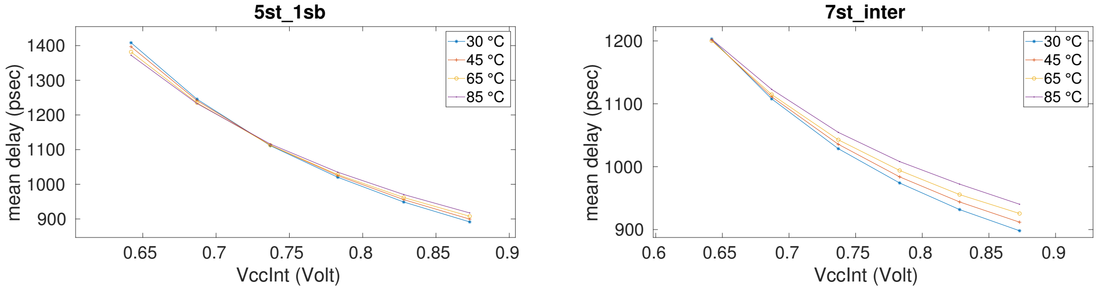
Το στοχαστικό κομμάτι της μεταβλητότητας δεν παρουσιάζει χωρική συσχέτιση και παρατηρήθηκε ότι ακολουθεί σε όλες τις περιπτώσεις κανονική κατανομή. Ο τελευταίος ισχυρισμός παρουσιάζεται στο Σχήμα 10, το οποίο απεικονίζει το ιστόγραμμα δύο αισθητήρων, όπου φαίνεται η ακρίβεια της κανονικής κατανομής. Επιπρόσθετα, πέραν του γεγονότος ότι οι στοχαστικές διακυμάνσεις έχουν συγκεκριμένη κατανομή, η επίδρασή τους μειώνεται καθώς αυξάνεται το μονοπάτι καθυστέρησης, λόγω της μετρίασής του, καθώς διαπερνά πολλαπλά στοιχεία λογικής και διασύνδεσης [23]. Τέλος, παρατηρήσαμε σε όλες τις περιπτώσεις ότι το στοχαστικό κομμάτι της μεταβλητότητας είναι μικρότερο για τους πόρους διασύνδεσης συγκριτικά με τους πόρους λογικής.



Σχήμα 9: Συντελεστής συσχέτισης μεταξύ του αισθητήρα "5st_1sb" (67/33) και των υπόλοιπων αισθητήρων (λογική/διασύνδεση).

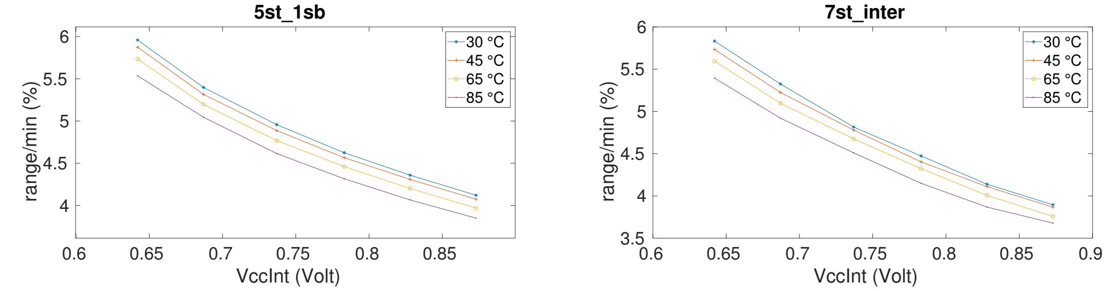


Σχήμα 10: Πιθανοκρατική κατανομή του στοχαστικού μέρους της μεταβλητότητας των αισθητήρων "5st_1sb", "7st_2sb" (device 1).



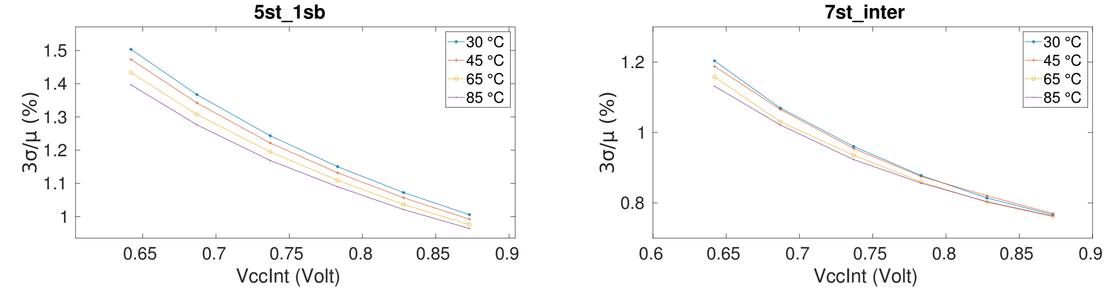
(α′) (β′)

Σχήμα 11: Μέση τιμή καθυστέρησης ως συνάρτηση της τάσης τροφοδοσίας για τέσσερις διαφορετικές θερμοκρασίες των αισθητήρων “5st_1sb”, “7st_inter” (device 1).



(α′) (β′)

Σχήμα 12: Συστηματική μεταβλητότητα ($(max - min)/min)$) ως συνάρτηση της τάσης και της θερμοκρασίας (device 1).



(α′) (β′)

Σχήμα 13: Στοχαστική μεταβλητότητα ($3\sigma/\mu$) ως συνάρτηση της τάσης και της θερμοκρασίας (device 1).

Η ανάλυσή μας συνεχίζεται με την αξιολόγηση της μεταβλητότητας υπό διάφορες συνθήκες θερμοκρασίας και τάσης. Για την ακρίβεια, στα πειράματά μας η τάση τροφοδοσίας έχει εύρος 0.640 – 0.875 V και η θερμοκρασία ολοκληρωμένου καθορίστηκε σε τέσσερις διακριτές τιμές: 30, 45, 65, 85 °C. Το Σχήμα 11 απεικονίζει τη μέση τιμή της καθυστέρησης ως συνάρτηση της τάσης τροφοδοσίας για τέσσερις τιμές της θερμοκρασίας. Παρατηρούμε ότι η καθυστέρηση αυξάνεται με την μείωση της τάσης, ενώ για τη θερμοκρασία παρατηρούμε το φαινόμενο της αντιστροφής θερμοκρασίας (temperature inversion), όπου η καθυστέρηση μειώνεται με την αύξηση της θερμοκρασίας. Παρατηρούμε ότι το σημείο αντιστροφής της θερμοκρασίας συμβαίνει προσεγγιστικά στα 0.72 V για τον αισθητήρα "5st_1sb", ενώ για τον "7st_inter" σε τάση χαμηλότερη των 0.65 V. Συγκριτικά με τους αισθητήρες λογικής, οι αισθητήρες διασύνδεσης παρουσιάζουν χαμηλότερη μείωση λόγω μείωσης της τάσης τροφοδοσίας και υψηλότερη μείωση λόγω αύξησης της θερμοκρασίας. Το τελευταίο εξηγείται λόγω του γεγονότος ότι η αντίσταση των μετάλλων αυξάνεται σχεδόν γραμμικά με τη θερμοκρασία, οπότε την ίδια συμπεριφορά έχει και η καθυστέρηση [2]. Λαμβάνοντας υπόψιν όλους τους αισθητήρες και τα FPGAs η μείωση λόγω της τάσης παρατηρήθηκε έως 33.9% ($7st\_inter$) - 57.9% ($7st\_1sb$), ενώ αντίστοιχα λόγω θερμοκρασίας έως 2.9% ($5st\_1sb$) - 4.8% ($5st\_inter$).

Τέλος, εξετάζουμε πως διαφοροποιείται η μεταβλητότητα υπό τις διάφορες συνθήκες τάσης και θερμοκρασίας. Τα Σχήματα 12, 13 παρουσιάζουν τις συστηματικές και στοχαστικές διακυμάνσεις συναρτήσει της τάσης και της θερμοκρασίας για τις αναφερθέντες τιμές αυτών. Και στις δύο περιπτώσεις παρατηρούμε ότι η μεταβλητότητα μειώνεται με την αύξηση της τάσης τροφοδοσίας. Επίσης, παρατηρούμε ότι η μεταβλητότητα μειώνεται με την αύξηση της θερμοκρασίας. Λαμβάνοτας υπόψιν όλες τις καταστάσεις λειτουργίας και τα FPGAs, η συστηματική μεταβλητότητα αυξάνεται έως 5.9% ($7st\_inter$) - 7.3% ($5st\_1sb$) και η στοχαστική έως 1.41% ($7st\_inter$) - 1.53% ($5st\_1sb$). Οι ολικές ενδοψηφιδικές διακυμάνσεις αντίστοιχα αυξήθηκαν έως 7.4-9.9% και οι ολικές διαψηφιδικές έως 9.5-12%.

Συμπεράσματα

Σε αυτήν την εργασία μελετήσαμε την μεταβλητότητα απόδοσης στα εμπορικά 16nm FinFET FPGAs (πρώτη στη βιβλιογραφία). Σχεδιάσαμε και τοποθετήσαμε αισθητήρες με διάφορα χαρακτηριστικά με σκοπό την αξιολόγηση των πόρων λογικής και διασύνδεσης των FPGAs. Η μεθοδολογία μας βασίζεται στο γνωστό κύκλωμα του ταλαντωτή δακτυλίου. Ωστόσο, παρουσιάζουμε μια νέα μέθοδο για να απομονώσουμε την καθυστέρηση των πόρων διασύνδεσης, χωρίς την ανάγκη να τοποθετήσουμε νέους αισθητήρες, αλλά με την προσεχτική σχεδίαση των ταλαντωτών δακτυλίου για να μπορέσουμε εν τέλει να αφαιρέσουμε τις καθυστερήσεις αυτών.

Στη συνέχεια, διαχωρίσαμε την μεταβλητότητα στο συστηματικό και στοχαστικό κομμάτι της και αξιολογήσαμε τα μαθηματικά μοντέλα που παρουσιάζει η βιβλιογραφία. Τα αποτελέσματα μας έδειξαν ότι το μεγαλύτερο κομμάτι της συνολικής μεταβλητότητας οφείλεται σε συστηματικές διακυμάνσεις. Επιπρόσθετα, εξήγαμε το συμπέρασμα της αυτόνομης μελέτης της μεταβλητότητας για κάθε FPGA καθώς και το γεγονός ότι οι πόροι της λογικής και της διασύνδεσης, γενικά, παρουσιάζουν διαφορετικές συστηματικές μεταβλητότητες. Αντίθετα, το στοχαστικό κομμάτι έχει γνωστή κατανομή (κανονική) και η επίδρασή του μειώνεται καθώς το κρίσιμο μονοπάτι αυξάνεται.

Επιπλέον, μελετήσαμε την μεταβλητότητα υπό διάφορες συνθήκες τάσης τροφοδοσίας και θερμοκρασίας. Τα αποτελέσματά μας έδειξαν έως 9.9% ενδοψηφιδικές διακυμάνσεις και 12% διαψηφιδικές υπό συγκεκριμένες καταστάσεις λειτουργίας. Αναμφισβήτητα, τα αποτελέσματα αυτά τονίζουν τη σημασία της πολύπλευρης ανάλυσης της μεταβλητότητας στα FPGAs και προάγουν ιδέες για την υλοποίηση πιο ανεπτυγμένων/σύνθετων μεθόδων για την άμβλυνση της μεταβλητότητας.

Μελλοντική Εργασία

Η πολύπλευρη ανάλυση των μεταβλητοτήτων στα FPGAs με τη μορφή που παρουσιάστηκε σε αυτήν την εργασία μπορεί να οδηγήσει σε πιθανές ερευνητικές κατευθύνσεις, κάποιες από τις οποίες αναλύονται συντόμως στη συνέχεια.

Αρχικά, μια σημαντική μελλοντική επέκταση είναι ο χαρακτηρισμός των μεταβλητοτήτων στις μονάδες ψηφιακής επεξεργασίας σήματος (DSPs) που υπάρχουν στα FPGAs, με τη χρήση της ίδιας μεθοδολογίας του ταλαντωτή δακτυλίου. Τυπικά, τα DSP μπλοκ βρίσκονται χωρικά μεταξύ των πόρων λογικής και διασύνδεσης, οπότε μπορεί να υποτεθεί ότι η μεταβλητότητα θα είναι παρόμοια με τους ήδη μελετημένους πόρους. Ωστόσο, αυτό μπορεί να μην είναι αληθές, καθώς διαφορετικά αποτελέσματα μπορεί να οφείλονται σε διαφορετικές μεθόδους κατασκευής αυτών κατά την διαδικασία παραγωγής. Για το λόγο αυτό, η μελέτη των διακυμάνσεων σε αυτά τα στοιχεία, καθίσταται σημαντική.

Στη συνέχεια, τα αποτελέσματα και η ανάλυση της παρούσας εργασίας μπορεί να οδηγήσουν σε υλοποίηση ή/και βελτίωση των εργαλείων CAD με σκοπό την εύρεση των γρήγορων/αργών περιοχών του FPGA και εν συνεχεία στην χρήση αυτής της πληροφορίας για σκοπούς αύξησης της απόδοσης της εφαρμογής. Η ανάλυσή μας δίνει ιδέες για την αξιολόγηση της μεταβλητότητας με τη χρήση πολλών χαρτών, αλλά ταυτόχρονα και περιορισμένων, όπως προκύπτει από την μαθηματική ανάλυση και μοντελοποίηση.

Τέλος, ένας πιθανός ερευνητικός προορισμός είναι η χρήση αισθητήρων για παρακολούθηση του FPGA στο χρόνο που εκτελείται η εφαρμογή. Αυτή η τεχνική μπορεί να οδηγήσει στην υλοποίηση μιας ανθεκτικής υποδομής, η οποία είναι ικανή να αξιολογεί τις μεταβλητότητες σε πραγματικό χρόνο, π.χ. λόγω αυξημένου υπολογιστικού φόρτου, και στη συνέχεια να μπορεί να λαμβάνει αυτόνομα αποφάσεις ώστε να διαβεβαιώνεται η σωστή λειτουργία των λογικών συναρτήσεων. Αναμφίβολα, τέτοιου είδους υποδομές μπορούν να μειώσουν σημαντικά τα πεσιμιστικά περιθώρια που έχουν συμπεριληφθεί λόγω των μεταβλητοτήτων.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Variability in Integrated Circuits

Moore's law indicates that transistor density per integrated circuit (IC) is doubled every 18 months [1]. This is mainly verified due to the shrinking of transistor's key dimensions, rather than manufacturing larger dies [2]. However, as dimensions are decreasing, reliable gigascale integration becomes one of the most significant challenges. Consequently, variability induced by manufacturing process, aka process variations, is increasing. Process variations result either from variations during the various processing steps due to inaccuracies of the equipment or from the intrinsic atomistic nature of materials in nanometer scale [3]. This leads to a main classification of process variations: *systematic* and *stochastic* [4–6]. Systematic sources are deterministic (induced from fabrication) and are, in general, spatially correlated, resulting in high likelihood for neighboring devices to present similar electrical properties. In contrast, stochastic refers to uncorrelated, unpredictable variations, originating from atomic scale differences.

Typical sources of variations resulting from manufacturing process, are photolithography and etching [7,8], photoresist development [5], rapid thermal annealing [42], film deposition and growth process [8] and chemical-mechanical polishing (CMP) [9]. Intrinsic variations derive mainly from random dopant fluctuations (RDF) [10], line-edge roughness [11] and oxide thickness variations [12]. These variations cause deviations in transistor parameters such as nominal lengths and widths, doping concentrations, oxide thickness etc., translating to variations in path delay, leakage power and yield.

Besides process variations, environmental variations have always been an issue for IC design. Environmental variations refer to voltage and temperature fluctuations across the die. These sources are spatially dependent across the die as well as time dependent. Voltage variations (also called power noise) have smaller time constants than temperature [5,13] and smaller spatial distribution across the chip [2], affecting more negatively circuits' performance.

Except from variations provoked by scaling in nanometer regime, the non ideal scaling of supply voltages and threshold voltages due to subthreshold leakage current constraints [14, 15] results in increased electric fields, accelerating wear-out failures. Prominent aging mechanisms for transistors are Bias Temperature Instability (BTI) [16], Hot Carrier Injection (HCI) [17] and Time Dependent Dielectric Breakdown (TDDB) [18], while metal interconnects failures are mainly caused by electromigration [19]. The aforementioned phenomena induce reliability concerns as they shorten the circuit's life, thus becoming pronounced for recent technology nodes [15].

The sources of variations are typically being reported in the literature with the abbreviation PVT, from Process, Voltage and Temperature. However, as mentioned, at deep sub-micron nodes the impact of aging in circuit performance can not be eliminated. Hence, a new abbreviation is introduced, i.e., PVTA [13].

The traditional methodology to cope with variability is the worst-case scenario of PVTA variations. To achieve that, sufficient worst case guard-band, usually in terms of operating frequency and supply voltage is utilized. Figure 1.1 depicts the worst case approach budgeting an adequate voltage margin to include all PVTA fluctuations. Nevertheless, as variability increases with device shrinking, safety margins should also increase, leading to inefficient designs. Therefore, variation aware designs are becoming more substantial. A typical category refers to adaptive control techniques [20]. By measuring operating conditions and adapting various parameters, e.g., supply voltage, operating frequency, in a closed loop framework, adaptive techniques are utilized to reduce pessimistic margins.

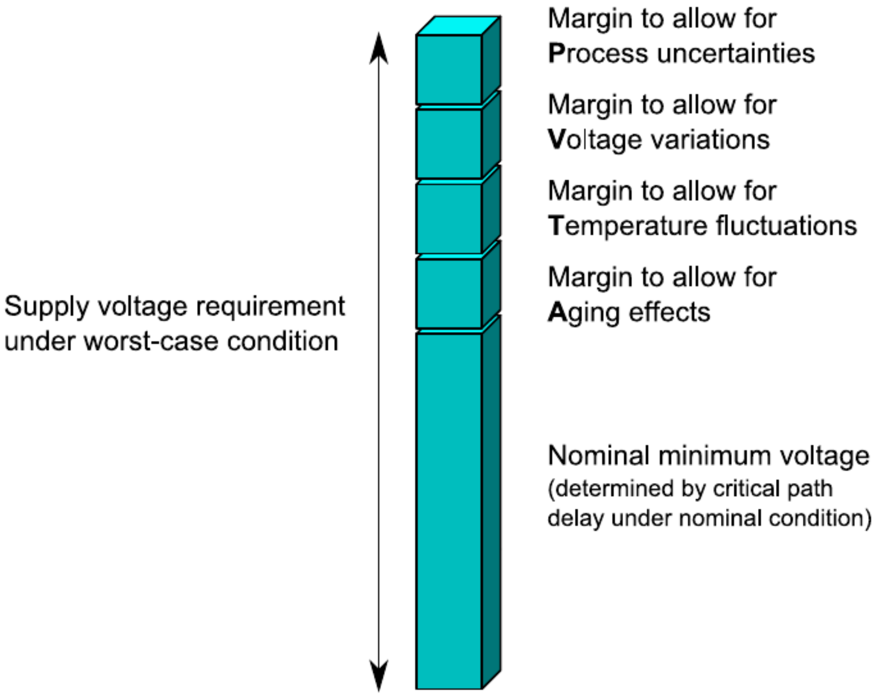

Margin to allow for
Process uncertainties
Margin to allow for
Voltage variations
Margin to allow for
Temperature fluctuations
Margin to allow for
Aging effects
Supply voltage requirement
under worst-case condition
Nominal minimum voltage
(determined by critical path
delay under nominal condition)

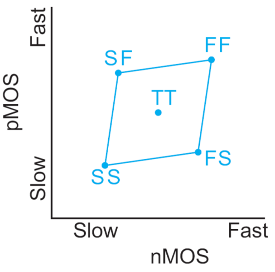
Figure 1.1: Voltage margin for worst case guard-banding scenario [13].

Resulting from the above analysis, the importance of understanding, modelling and quantifying variability in deep sub-micron technology nodes to avoid pessimistic designs is clarified. However, since worst case scenario significantly degrades performance and power dissipation, new techniques (adaptive) are proposed by academia to exploit variability margins.

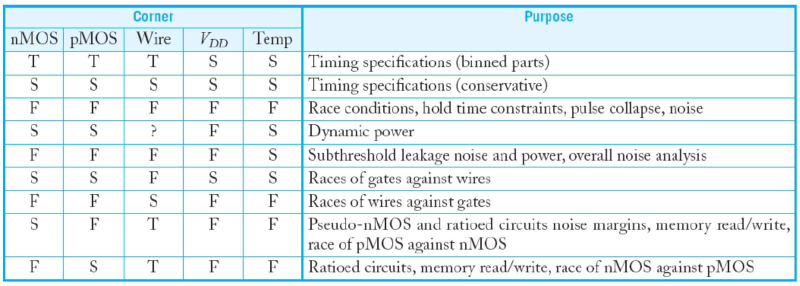
1.2 Variability Modelling and Simulation

To cope with process and environmental variations, specific models have been introduced. Typically, the impact on electrical properties of transistors, e.g., performance (speed) is marked as: typical (T), fast (F) and slow (S). In CMOS technology two types of transistors exist: nMOS and pMOS. Hence, the combination of performance levels for the different types of transistors leads to design/process corners [2]. Figure 1.2a illustrates the five possible corners: typical-typical (TT), fast-fast (FF), slow-slow (SS), fast-slow (FS), and slow-fast (SF). In the last two corners (FS, SF) devices (transistors) are not affected uniformly, causing asymmetric rising and falling edge of path delays. The opposite applies for the rest three corners (TT, FF, SS), where typically timing violations does not occur.

Corners refer also to variations in metal interconnects as well as to environmental parameters: supply voltage (V_{DD}) and temperature. Figure 1.2b shows some interesting design corners. Circuits are simulated in different corners to confirm different performance and correct operation. The aforementioned figure shows the purpose of simulation in each corner. More information can be found at [2]. In any case, timing constraints should be accomplished at the worst-case condition, i.e., corner SSSSS.



(a)



Corner					Purpose
nMOS	pMOS	Wire	V_{DD}	Temp	
T	T	T	S	S	Timing specifications (binned parts)
S	S	S	S	S	Timing specifications (conservative)
F	F	F	F	F	Race conditions, hold time constraints, pulse collapse, noise
S	S	?	F	S	Dynamic power
F	F	F	F	S	Subthreshold leakage noise and power, overall noise analysis
S	S	F	S	S	Races of gates against wires
F	F	S	F	F	Races of wires against gates
S	F	T	F	F	Pseudo-nMOS and ratioed circuits noise margins, memory read/write, race of pMOS against nMOS
F	S	T	F	F	Ratioed circuits, memory read/write, race of nMOS against pMOS

(b)

Figure 1.2: Transistors process corners (a). Corner checks (b) [2].

Traditionally, Static Timing Analysis (STA) is employed in all process corners to meet timing constraints [43]. STA tools are deterministic and the circuit's delay is computed for each specific corner in order to find the [44]. Consequently, all parameters of variations, e.g., threshold voltage, channel length, temperature, are considered to be fixed across the die for each corner simulation. The drawback of STA is that variations across the die (called within-die variations) are not taken into consideration. This was not an issue in the past, as global variations (refer to variations from die-to-die) were much larger than within-die variations [2]. However, with technology down-scaling into nanometer regime, within-die variations have been considerable.

The inability of STA to model within-die variations results in over- or underestimate of circuit path delays [44], and thus a new model becomes vital. Therefore, Statistical STA (SSTA) has been proposed, where the sources of fluctuations are treated statistically. Most research works on SSTA focus only on statistical models of process variations and uncertainties due to environmental and aging causes, which are typically modeled using worst-case margins [44]. Physical parameters (e.g., channel length, doping concentration, oxide thickness) are modeled as random variables. Consequently, path delays are considered as a sum of independent random variables [2,44]. The objective of SSTA is to compute the probability density function (PDF) of the random variable representing circuit path delay [45]. Afterwards, results can be employed to determine yield as well as design timing constraints.

1.3 Focus of Thesis & Research Goals

All semiconductor chips are affected by process variations. However, Field programmable Gate Arrays (FPGAs) are of particular interest due to their reconfigurable nature. In fact, due to the ability to program every single resource at a very low level, it enables performing built-in-self-tests (BISTs). Additionally, another appealing characteristic is the homogeneous architecture of an FPGA consisting of identical resources, placed uniformly in the entire fabric. Taking advantage of the above characteristics, the actual performance variation can be measured via the deployment of custom sensors across the fabric [21–24].

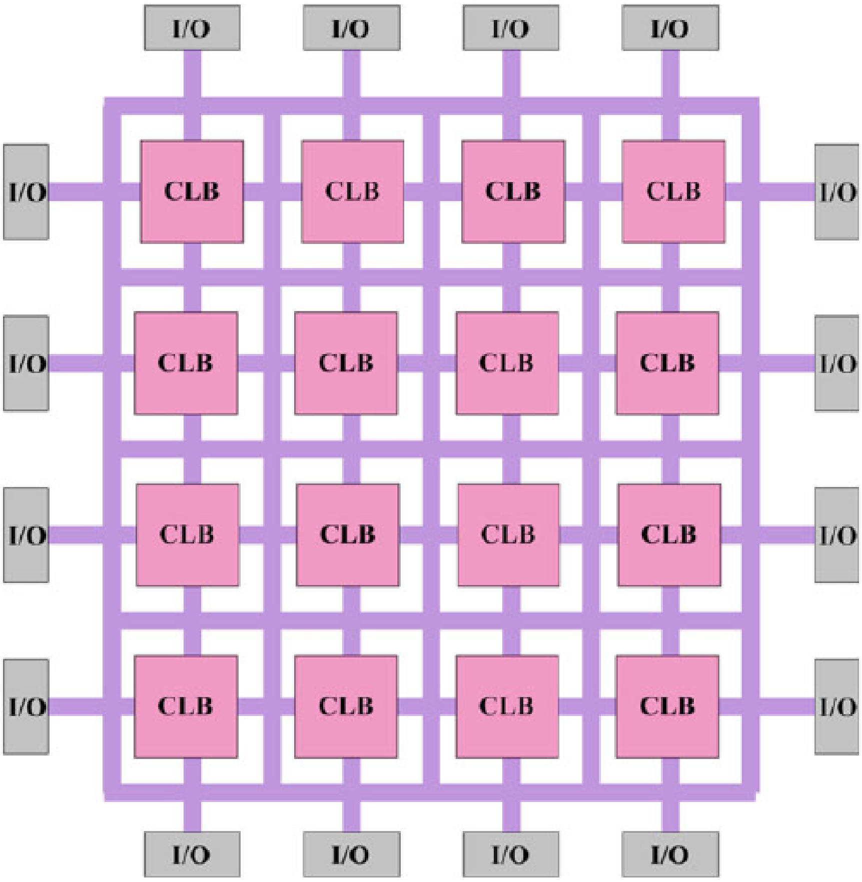


Figure 1.3: FPGA architecture overview [25].

In order to clear up the above statements, an overview of FPGA architecture is illustrated in Figure 1.3. Its basic components are configurable logic blocks (CLBs), programmable routing resources and I/O blocks. A two dimensional grid is arranged with CLBs being interconnected with routing resources. The reconfigurability is achieved by CLBs, which implement user-defined logic functions. Programmable routing resources are utilized to connect the implemented logic functions and finally, input/output blocks (I/O) are used to make off-chip connections.

To cope with variability and provide acceptable solutions, the industrial electronic design automation (EDA) tools consider the extreme case process corners for FPGA designs. Additionally, as mentioned in the previous section within-die variability has increased acutely into nanometer regime, thus these conservative STA approaches lead to significant performance loss [27, 46]. Considering that, the importance of characterizing, quantifying and finally exploiting the actual performance capabilities of the individual chips becomes appealing towards potential performance improvements. This potential can be reinforced by the fact that the total measured variability is spatial correlated [44], due to to systematic variations, which can presumably be exploited by the reconfigurable nature of the FPGA.

This thesis studies the performance variation in commercial state-of-the-art 16nm FinFET FPGAs (literature's first). The main research goals of this work are:

- The assessment of performance variation in logic and interconnect resources via variability maps. To achieve that, we design custom sensors at very low level and map them across the FPGA fabric.
- The decoupling of variability into systematic and stochastic, to analyze and quantify their impact on circuit performance.
- The assessment of variability under different environmental conditions, i.e., voltage and temperature, which have major impacts on circuit's performance.

In this work, we focus on a thorough analysis of the variability. However, the target vision is performance enhancement by exploiting the existing variability. There exists a number of works in the literature demonstrating the performance improvement by exploiting the process variability in FPGAs, either in-the-field via frequency/voltage scaling methods [27, 47] or by adapting the computer-aided design (CAD) tools to the specifics of the underlying chips [23, 26, 48, 49]. Our work aims for highlighting the importance of multifaceted evaluation of variability and give insights for future implementations of more accurate methods/tool for its exploitation.

We clarify that the target of this study is any potential contribution toward improving CAD tools, e.g., in guiding the place & route process, even on a per-board (FPGA) basis (assuming feedback from the device itself), and not just another mere evaluation of process variability in chip manufacturing. Instead, we are interested in analyzing the performance of the constituent parts of critical paths, i.e., their routing and logic parts, as well as their behavior with respect to process, voltage and temperature, referring to real-world designs' critical paths.

1.4 Dissertation Outline

The remainder of the dissertation is structured as follows:

Chapter 2 presents a brief review of process variability sources, their classification and the effect that they have in the final measured values of interest, i.e., path delay, power consumption. Furthermore, makes an introduction to the FPGA architecture and finally refers to the related work on variability evaluation of FPGAs.

Chapter 3 presents the proposed methodology for the assessment of the variability and the mathematical models utilized for the analysis and decoupling of variability into systematic and stochastic. In addition, methods for the aforementioned decoupling are exhibited.

Chapter 4 provides the experimental results and the variability maps, occurred from the deployed custom sensors. A thorough analysis of the results is presented, including mathematical tools, i.e., the Pearson correlation coefficient. Finally, the variability is assessed under voltage and temperature alterations, and the explanations for the presented results are clarified.

Chapter 5 draws the conclusions and the highlights of this thesis and in addition addresses presumable future research directions.

Chapter 2

Background

2.1 Variability Classification

Variability originates from fluctuations in process, voltage, temperature and aging (PVTA). The variations can be categorized in various ways. An applicable way can be the division in environmental, temporal and spatial variations [5, 31]. Environmental variations arise typically from alterations in temperature, voltage and even cosmic radiation [31]. Temporal refers to aging and transistor wear-out [5], being reversible (e.g. self-heating), as well as irreversible (e.g electromigration). Spatial variations depend on the distances between transistors and metal interconnects (wires). Hence, different locations on a chip have different electrical properties, affecting die's performance and leakage. Typically, spatial variations appear from deviations in manufacturing process, e.g. channel length, threshold voltage, wire width. However, spatial variations can arise from environmental sources, e.g. on-die hot spots, activity factor [5].

Resulting from the previous paragraph, variability categorization is not commonplace and can be frustrating. Thus, in order to clarify the variability classification and the impact on electrical properties both for active (transistors) and passive (wires) components of chips, an extensive variability analysis is presented in the following sections.

2.2 Process Variations

Process variations result either from variations in fabrication parameters or from the transistors' intrinsic atomistic nature [3]. These variations can be categorized as follows [2]:

- Lot-to-Lot (L2L)
- Wafer-to-Wafer (W2W)
- Die-to-Die (D2D)
- Within-Die (WID)

Figure 2.1 illustrates the above classification. However, as circuit designers are interested in the final characteristics of dies, L2L and W2W fluctuations are lumped with D2D and are called "global" or "inter-die" variations, being both systematic, as well as stochastic in nature. Modelling of global variations is traditionally accomplished by design/process corners. In the same way, WID variations, which are called "local" or "intra-die", also consist of systematic and stochastic parts. Intra-die variations have become significant in nanometer regime, and can not longer be ignored [2]. This type of variations are treated statistically (SSTA).

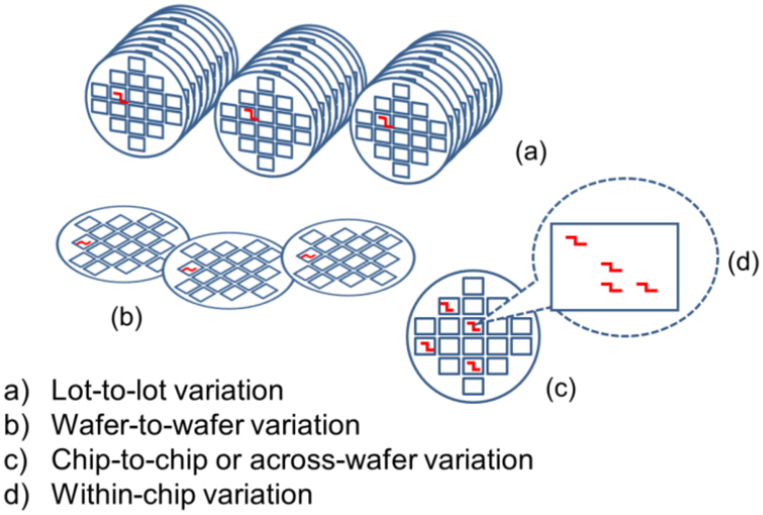

(a)
(b)
(c)
(d)
a) Lot-to-lot variation
b) Wafer-to-wafer variation
c) Chip-to-chip or across-wafer variation
d) Within-chip variation

Figure 2.1: Classification of process variations [31].

As mentioned, both inter-die and intra-die variations are divided into two classes:

- **Systematic variations**: are deterministic variations, spatially dependent, i.e., on the spatial position on the die and on the wafer, as well as layout dependent [31]. Typical sources are lithographic process, etching and Chemical Mechanical Polishing (CMP).

- **Stochastic variations**: are unpredictable and random in nature resulting either from the atomic layer differences, e.g. random dopant fluctuations (RDF), Line Edge Roughness (LER), or from random fluctuations in fabrication process.

2.2.1 Common Sources of Process Variations

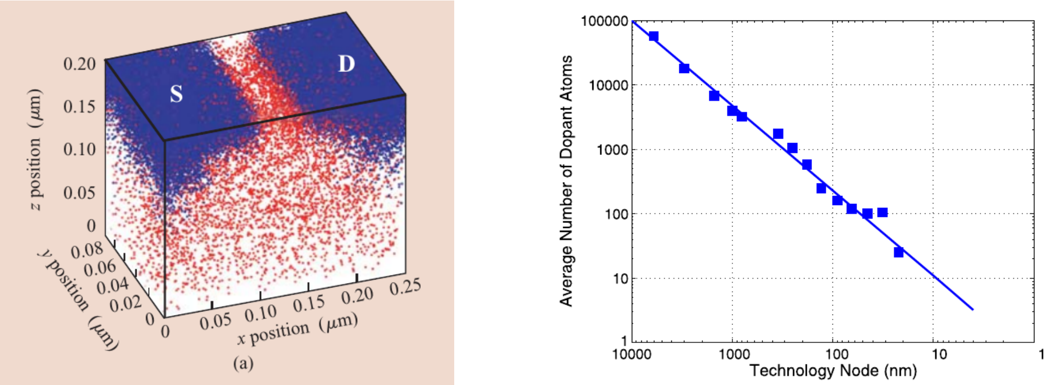
Process variability can be divided into intrinsic, which expresses the atomic level differences (stochastic variations) and extrinsic, occurring form fabrication. In the following subsections both are being reviewed.

2.2.1.1 Intrinsic Transistor Variability

The main sources of intrinsic transistor variations have typically been: random dopant fluctuations (RDF), line-edge roughness (LER) and gate oxide thickness variations [5].

2.2.1.1.1 Random Dopant Fluctuations

Ion implantation and annealing process determine the doping procedure in a channel [5]. However, the position and the number of these atoms are random in nature (Figure 2.2a), resulting in a random distribution of threshold voltage (V_{th}). Also, it causes capacitance and resistance variations in the source/drain region [5]. In older technologies, the number of dopant atoms per channel region were in the order of thousands (Figure 2.2b), and hence the impact of RDF was negligible. Instead, in recent deca-nanometer nodes the number has been reduced to the range of tens, denoting that RDF is the most prominent source of stochastic variations in modern technologies [31, 50].



(a) (b)

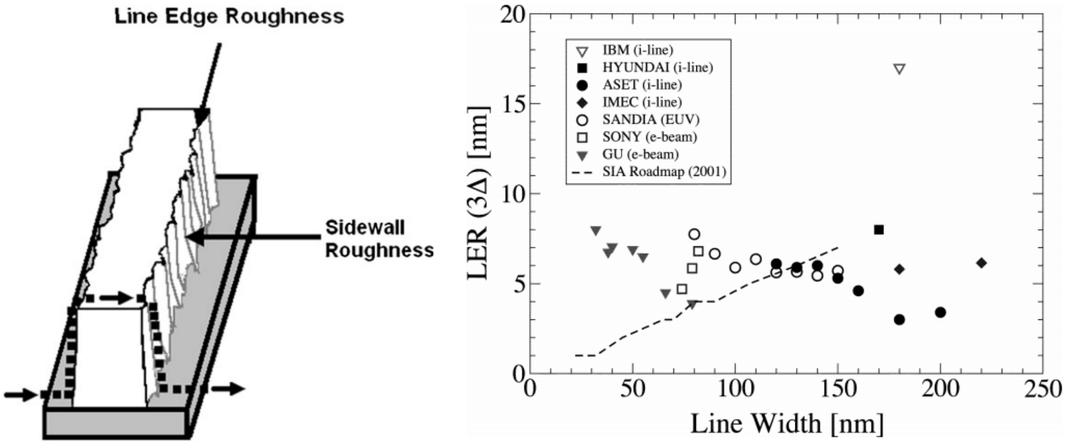
Figure 2.2: Randomly placed dopants in 50-nm MOSFET technology [5] (a). Number of dopant atoms per channel region over technology nodes [32] (b).

2.2.1.1.2 Line-Edge Roughness

The uncertainty in width of patterned lines is increased with technology downscaling. The deviation of the line edge from a straight line, is known as line edge roughness [33](Figure 2.3a). In sub 50 nm technology, LER has become a considerable source of variations [34]. It arises from variation in the incident photon count during lithography exposure, the absorption rate, chemical reactivity, and molecular composition of the photoresist [5]. LER leads to a non-uniform channel length, affecting transistor current and V_{th}. In Figure 2.3b the actual data from different lithography processes are illustrated. As shown, LER uncertainty does not scale according to SIA Roadmap, and is typically considered as the second most significant intrinsic variability issue following RDF.

2.2.1.1.3 Gate Oxide Thickness Variations

The mean gate oxide thickness can be controlled with high accuracy; the uncertainty is in the order of a fraction of one atomic layer [2]. In [5] it is referred that the uncertainty induced by oxide variability, leads to an approximately 10% increase in standard deviation of V_{th}. Thus, in contrary with the aforementioned sources,



(a) (b)

Figure 2.3: LER effect of a patterned line feature [33] (a). Actual data from lithography processes reported by different labs and SIA Roadmap (2001) [34] (b).

this variation is secondary. However, the impact on oxide tunneling leakage current is prominent, since it varies exponentially with gate thickness [5].

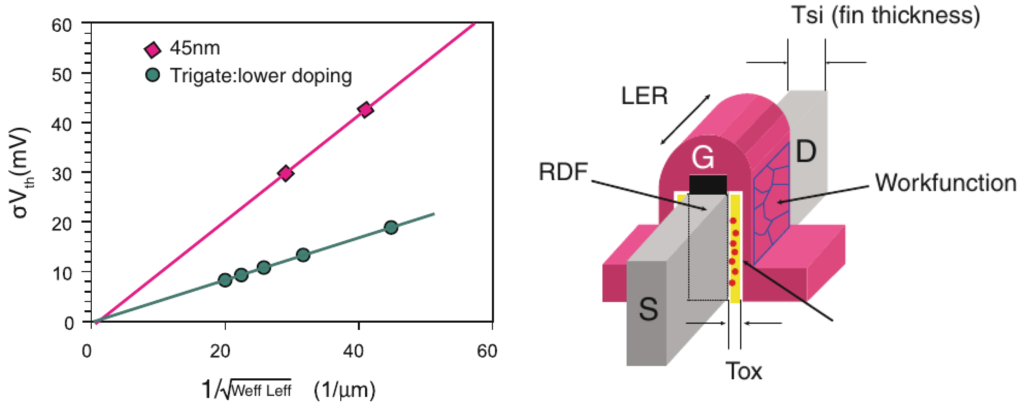
2.2.1.1.4 Emerging Technologies and Variations: FinFET

The fundamental transistor architecture dominated in digital design is the MOSFET transistor. However, with the aggressive scale-down in deep sub-micron technology have led to short-channel effects (SCEs). To mitigate these phenomena, new architectures have been proposed. FinFET transistors introduce a fundamental change in CMOS technology, moving from traditional planar transistors (MOSFET) to 3D structures. FinFET technology can significantly improve SCE, and thus transistor's performance [11, 35]. Its main advantage is the stronger coupling to the channel offering better control with lower channel doping [35]. This contributes to reduced effects on variations arising from RDF, and consequently to reduced uncertainty of V_{th}. Figure 2.4a depicts the comparison of V_{th} variation due to RDF from a 45 nm technology, between planar and FinFET (also referred as trigate). The latter has lightly doped channel, compared to the planar.

In contrast, FinFET architecture induces new sources of intrinsic variations, in comparison with the planar. LER does not affect only the gate length, but also the fin thickness (Figure 2.4b). In addition, the metal gate, which has been principal for deep sub-micron technology with the introduction of high-K gate dielectrics [51], introduces another major source of variation: workfunction fluctuation (WKF) [50]. It has been shown [50], that WKF is a major source of V_{th} variation both for n-type and p-type FinFETs. It must be clarified that WKF does not only refer to FinFET, but also to every MOSFET with metal gate, typically beyond 45 nm technology [2].

2.2.1.2 Extrinsic Variability

Extrinsic variability occurs due to shifts in the manufacturing process. It does not have association with atomic differences but with fabrication's dynamic and technologies. These type of sources are present in multiple fabrication processes,



(a) (b)

Figure 2.4: Comparison of V_{th} variation due to RDF between planar and FinFET (trigate) (45 nm) (a). FinFET variations including RDF, gate LER, fin LER (fin thickness), oxide thickness and workfunction variations (b) [35].

e.g., lot, wafer processes steps, but also occur from the layout design [5, 31]. This manufacturing variability leads typically to systematic variations [5]. For instance, Figure 2.5 shows the frequency distribution utilizing ring oscillators (ROs), as a function of their position in the wafer. The frequency distribution can be analyzed into two unambiguous components: a systematic spatial radial component and a smaller random. The main sources of extrinsic variations are analyzed below.

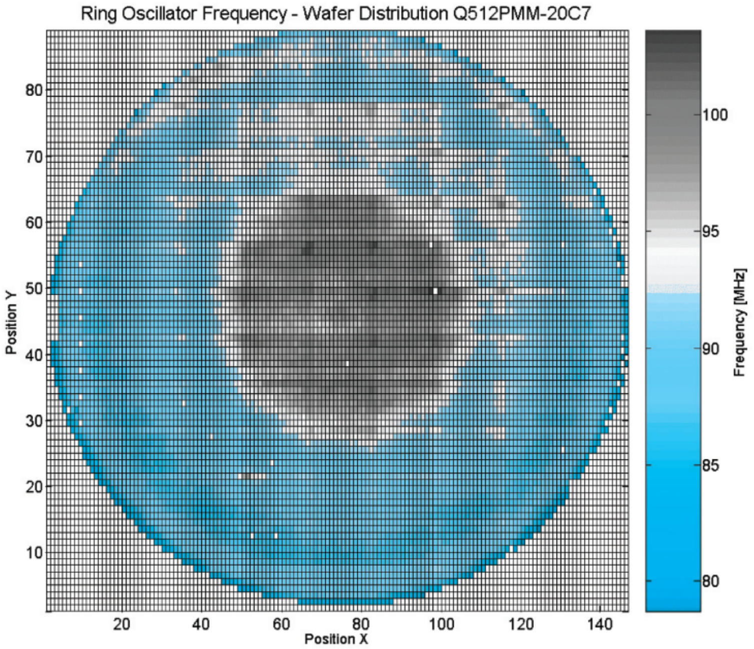
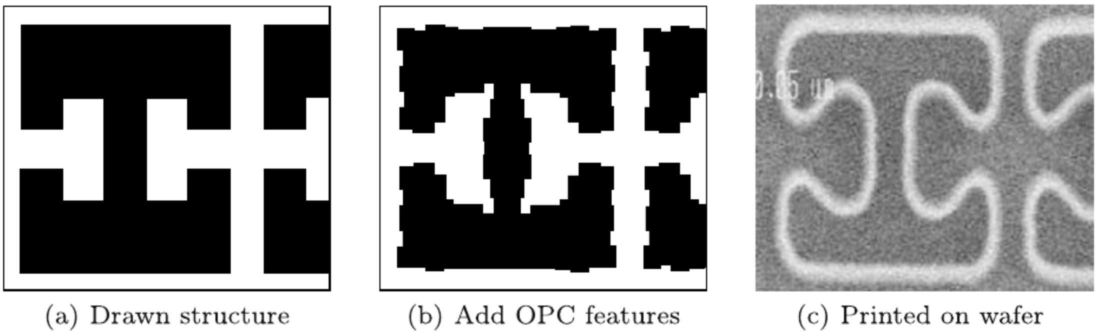


Figure 2.5: Ring oscillator frequency distribution for a CMOS 90 nm wafer [2].

2.2.1.2.1 Lithography variations

The wavelength of light at lithography process has remained at 193 nm wavelength since 130 nm process node [13]. However, when the wavelength is greater or equal to the minimum printed feature size, i.e., critical dimension (CD), then CD is distorted [2]. To avoid these sources of variations, resolution enhancement techniques (RET) have been developed. In particular, a prominent technique is optical proximity correction (OPC). OPC introduces small alterations to the mask patterns (Figure 2.6), to reduce the unintended rounding on edges [2]. Furthermore, the opposite movement between the mask reticle and the wafer can cause tiny vibrations leading to non-uniformities in the depth of focus (DOF) and the light-exposure dose [31]. This results in non-uniformity of CD, leading to delay and leakage variations.


(a) Drawn structure

(b) Add OPC features

(c) Printed on wafer

Figure 2.6: Optical proximity correction (OPC) is used to alter the patterns of masks for distortions compensation [13].

In addition, the post exposure bake (PEB) is another essential source of variation in the lithographic process [31]. The rapid change of temperature in PEB step of the wafer activates unwanted chemical reactions and the diffusion of the chemicals within the photoresist. This results in an unequal temperature which can cause significant CD variations.

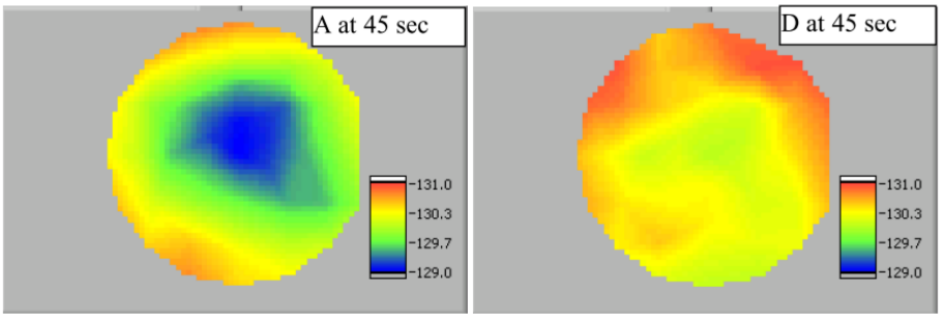


Figure 2.7: Temperature non-uniformity near the end of PEB step for two wafers (A and D) [36].

2.2.1.2.2 Well proximity effect

The well proximity effect is a layout dependent effect. It is a phenomenon caused by the lateral scattering of implantation ions during the ion implantation step for wells [2, 31]. A number of ions collide at the edge of photoresist, on top of the shallow trench isolation (STI), and disperse at the well edges, as depicted in Figure 2.8. This results in a greater concentration of dopant atoms at the edge of the well, which is translated to higher threshold voltages in that area. Well proximity effect is essential in deep sub-micron nodes, due to the small number of dopant atoms (Figure 2.2b).

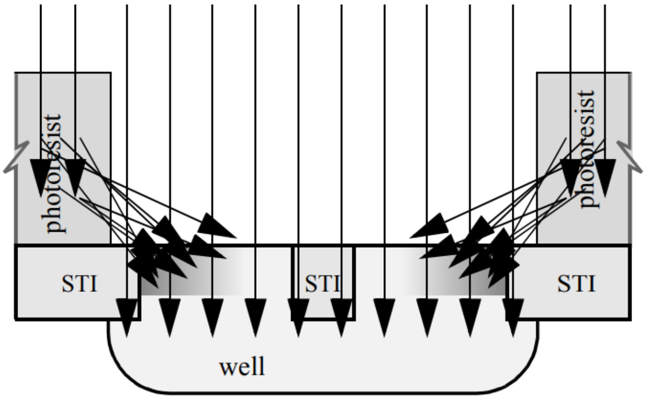


Figure 2.8: Well proximity effect increases, due to scattering, the doping concentration near the edge of the well [37].

2.2.1.2.3 Chemical Mechanical Polishing

Chemical mechanical polishing (CMP) is used to flatten the topography on the wafer, making feasible the integration of seven or more layers of metal interconnects [2]. Traditionally, aluminum metal is patterned and inter-layer dielectric (ILD) is polished. Nevertheless, in deep sub-micron technologies aluminum has been replaced by copper and a new technology of metal interconnection, named damascene process, has been utilized. In this process the oxide is patterned and etched, and metal is deposited followed by metal CMP [2, 9]. When ILD is polished, variability occurs in dielectric. On the contrary, in damascene process, variation occurs in copper wire. This results in two variations effects: dishing and erosion (Figure Typically, wide lines experience significant metal dishing, whereas fine pitch lines experience oxide erosion [9]. Both of these effects are layout dependent and result in metal thickness loss. Results have shown that CMP variation can increase bus delay more than 30% [9].

2.2.1.2.4 Other Sources

Other sources of systematic spatial variations due to fabrication are photoresist development and etching [5], strained silicon effects (used for carrier mobility enhancement), oxide thickness non uniformity [31], etc. Some of these sources, e.g.,

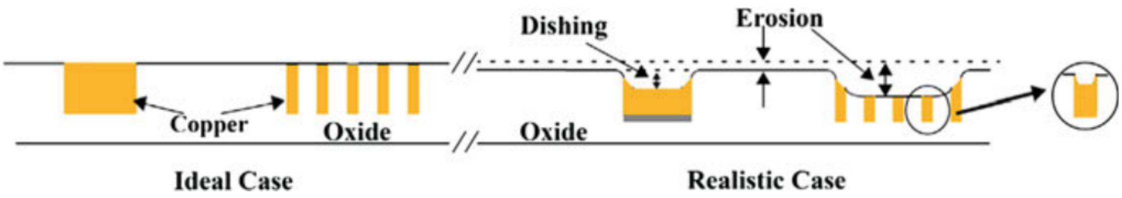


Figure 2.9: Ideal scenario in contrast to realistic, where metal thickness is decreased due to CMP [35].

strained silicon effects, affect only transistors, while others, like etching variations, have a negative impact on metal interconnects as well.

2.3 Voltage Variations

Supply voltage variations are caused mainly by supply regulator's tolerances, IR drops and di/dt noise [2]. Voltage supply regulator's offsets from nominal voltage can lead to fluctuations, which are caused either from inaccuracies of the regulator, or from the voltage reference circuit [13]. IR drops are caused mainly by the parasitic resistance of metal interconnects inside the chip, but also there is a small contribution outside the chip [2]. Additionally, IR drops can be caused by switching activity, when multiple transistors of the chip operate simultaneously. IR drop obeys the Ohm's law [13]. Finally, di/dt noise is caused by the parasitic inductance, which results from metal interconnects inside the chip and interconnects that connect chip with package.

Voltage fluctuations are both spatial and temporal in nature. Figure 2.10a illustrates the simulated spatial distribution of the supply voltage within an ASIC design. Typically, voltage variations have very short time constants in the range of nano- to microseconds [13].

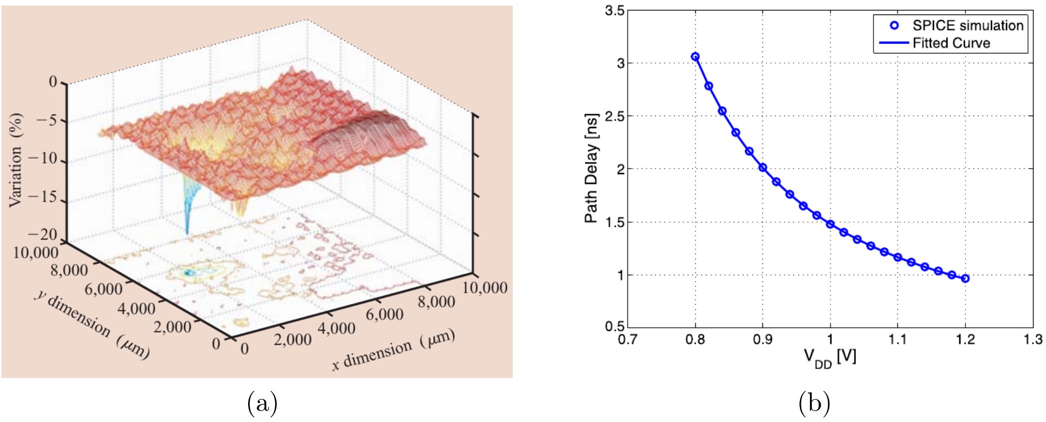


Figure 2.10: Simulation of percentage of supply voltage variation within an ASIC design [5] (a). Simulated path delay versus supply voltage and fitted curve by eq. 2.1 of the critical path of a multiplier in 65 nm [13] (b).

The impact of voltage in circuit delay can be derived from the alpha-power law model of the CMOS logic gate delay [52]:

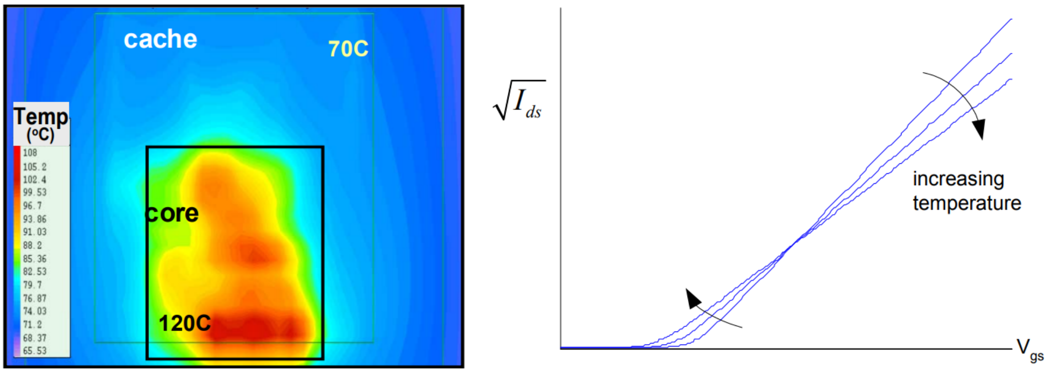
$$t_d = \frac{V_{DD}}{K \cdot (V_{DD} - V_{th})^a} \tag{2.1}$$

where V_{DD} is the supply voltage, V_{th} is the threshold voltage, a is a fitting parameter and K is a process dependent parameter.

For a critical path delay the same equation can be approximately utilized for diverse supply voltages. Figure 2.10b shows the simulation of the most critical path of a multiplier circuit, where the solid line exhibits the fit by the equation 2.1. This picture reveals how accurately the aforementioned equation is utilized for fitting data of critical paths. Notice that, in this case the parameters of the equation (V_{th}, a, K) are obtained by fitting and does not have the physical meaning mentioned above.

2.4 Temperature Variations

Junction temperature is the summation of ambient temperature and the increase in temperature from power dissipation [2]. Power dissipation leads to differentiated spatial temperature distribution, called hot spots, where commonly high transition activity occurs. On the contrary, ambient temperature leads to global shifts in junction temperature. Temperature fluctuations are spatial and time dependent like voltage fluctuations. However, spatial temperature variations are more gradually distributed contrary to spatial voltage variations [2] and their time constants are in the range of milliseconds to seconds [13]. Figure 2.11a illustrates the temperature variation for a microprocessor. Thermal hot spots inside the core have a maximum value of 120 °C, while inside the caches the difference is approximately 50 °C lower.



(a) (b)

Figure 2.11: Thermal image of a microprocessor [38] (a). I-V characteristics of a transistor for diverse temperatures [2] (b).

Typically an increase in temperature leads to a decrease in circuit speed due to reduced carrier mobility (degrades non linearly) and to an increase in the metal interconnect resistance (almost linear). However, for low supply voltage (typically from 0.7 to 1.1) transistors can operate in temperature inversion: speed is increased with temperature elevation. This is explained by the fact that V_{th} degrades almost linearly with temperature, and for low supply voltages the V_{th} degradation dominates the carrier mobility degradation. Figure 2.11b depicts this phenomenon by plotting the drain current versus the gate-source voltage, where the thermal inversion spot is denoted.

2.5 Aging

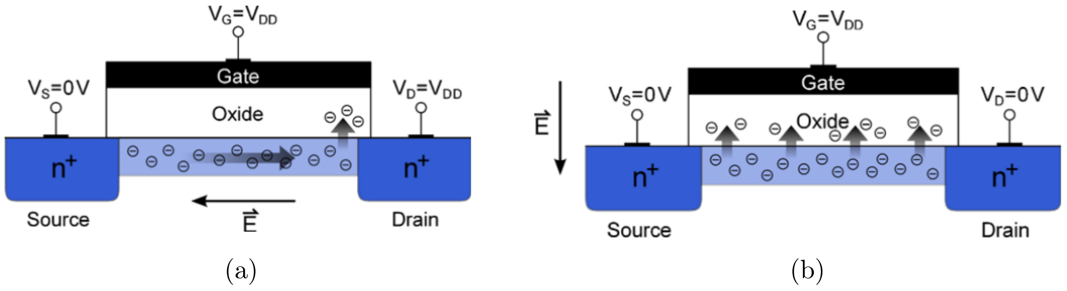
Electric fields are increasing due to the non-ideal scaling of supply voltages and threshold voltages [14, 15]. In addition, the usage of new materials has increased wear-out phenomena. Designers address these aging problems by adding sufficient safety margins, so that circuits can operate in the long term (typically more than 10 years [2]). Aging effects have time constants in the order of days, weeks or even years.

2.5.1 Gate Oxide Wear-Out

The prevailing mechanisms that cause gate oxide wear-out are analyzed briefly in the following Subsections.

2.5.1.1 Hot Carrier Injection (HCI)

Due to transistor switching, carriers are accelerated and obtain high energy ("hot carriers") due to the high lateral electric field. Hence, some of them can be injected into the gate dielectric [2]. The hot carriers that are trapped inside the dielectric cause shifts in the threshold voltage, reducing the transistor's speed. Figure 2.12a illustrates the HCI effect for a n-MOS transistor.



(a) (b)

Figure 2.12: HCI effect for a n-MOS transistor (a). BTI effect for a n-MOS transistor (b) [13].

2.5.1.2 Bias Temperature Instability (BTI)

BTI occurs when high vertical electric fields are applied to gate oxide of a transistor. In this case, bonds are developed at the Si/SiO_2 interface, called "traps", where charge is trapped [2]. Contrary to HCI, BTI does not occur at switching time, besides it arises when transistor is "ON" for a long period of time. For n-MOS transistors, this phenomenon is referred as Positive Bias temperature Instability (PBTI) and is more essential at higher temperatures. Accordingly, for p-MOS is called NBTI (N refers to Negative). Figure 2.12b depicts the PBTI phenomenon. Both PBTI and NBTI cause shift in the threshold voltage, leading to transistor performance degradation. This effect has become the most essential wear-out mechanism for nanometer technologies [2].

2.5.1.3 Time-Dependent Dielectric Breakdown (TDDB)

When a vertical electric field is applied to the gate oxide, the gate leakage current is increased. This phenomenon is called time-dependent dielectric breakdown (TDDB) and may lead to a gate short circuit, destroying the transistor [2]. The physical mechanisms of this effect are not completely comprehensible. More information about TDDB can be found at [53].

2.5.2 Metal Interconnect Wear-Out

Due to the high current densities in metal interconnects some atoms can migrate, causing vacuums inside the metal wire. This effect is illustrated in Figure 2.13, where electromigration has occurred at the connection (via) between the two metal layers. It is mostly significant for unidirectional currents (direct currents (DC)) and has exponential dependence on temperature [2]. On the other hand, for bidirectional (AC) metal interconnects, self-heating is the most essential effect. Current through wire dissipates power and because the dielectric is a thermal insulator, temperature can increase substantially, leading to slower metal interconnects [28]. As mentioned, since electromigration is very sensitive to temperature, self-heating can cause electromigration problems in AC wires [2].

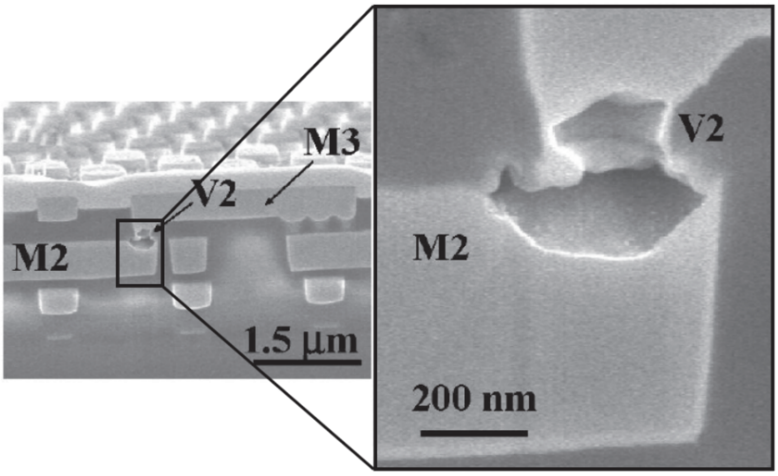


Figure 2.13: Electromigration vacuum in via between metal layers M2-M3 [19].

2.6 FPGA Architecture

Field Programmable Gate Arrays (FPGAs) are integrated silicon transistors that can be electrically programmed to develop almost any digital circuit or system. The FPGA configuration is specified using a hardware description language (HDL), analogous to that used for fixed-function Application-Specific Integrated circuits

(ASICs). Compared to an ASIC, an FPGA requires significant cost in area, delay and power consumption [39]. However, the reconfigurable nature of an FPGA provides a tremendous reduction in Non Recurring Engineering (NRE) cost and time to market. In addition to the previous advantages, the increased performance over microprocessors renders the FPGAs to be the appealing computing platforms for server or cloud applications, hardware acceleration (especially machine and deep learning), telecommunications, signal processing, Internet of Things (IoT) and ASIC Prototyping.

2.6.1 Overview

FPGAs consist of an array of programmable logic blocks, which typically are diverse in types. Figure 2.14 depicts an FGPA structure, where programmable blocks are general logic blocks, as well as memory and multiplier blocks [39]. These programmable logic blocks implement logic functions, which are connected by programmable routing resources. Thus, all logic blocks are surrounded by programmable routing fabric, as illustrated with grey color in the aforementioned Figure. Finally, the 2-D array periphery of an FPGA is arranged by programmable input/output blocks (I/O), that allow the connection with the outside world.

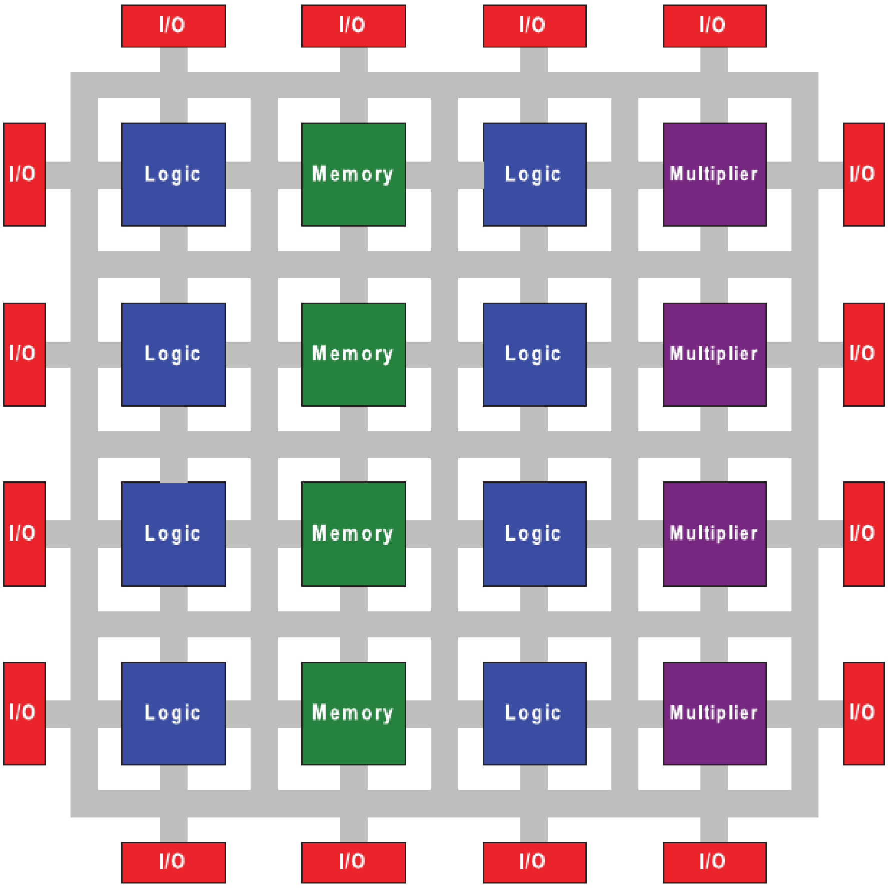
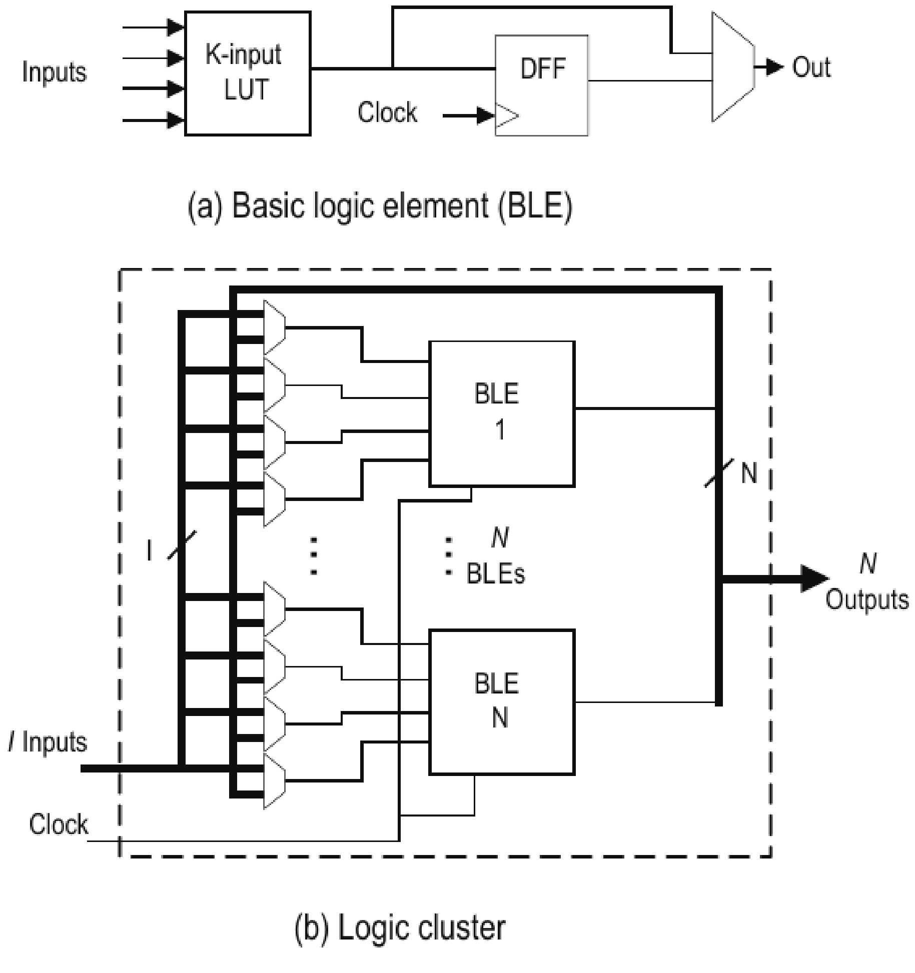


Figure 2.14: FPGA basic structure [39].

The "programmable/reconfigurable" term indicates the ability of the programming of logic functions after silicon fabrication. This reconfigurable characteristic of FPGAs is achieved by programmable switches, implemented with various technologies. The most known programming technologies for programmable switches are: static memory, flash and anti-fuse [25]. Static memory technology uses static memory (SRAM) cells for programmable switches and are fabricated with standard CMOS technology. Flash technology uses flash memory cells, while anti-fuse can not be reprogrammed. Both of them are manufactured in a different technology compared to the conventional CMOS technology. Modern integrated circuits typically use SRAM-based programming technology, primarily for the reason of being manufactured by the standard CMOS process. More information and comparison between programming technologies can be found at [39].

2.6.2 Logic Block Architecture



(a) Basic logic element (BLE)

(b) Logic cluster

Figure 2.15: FPGA BLE and logic cluster [39].

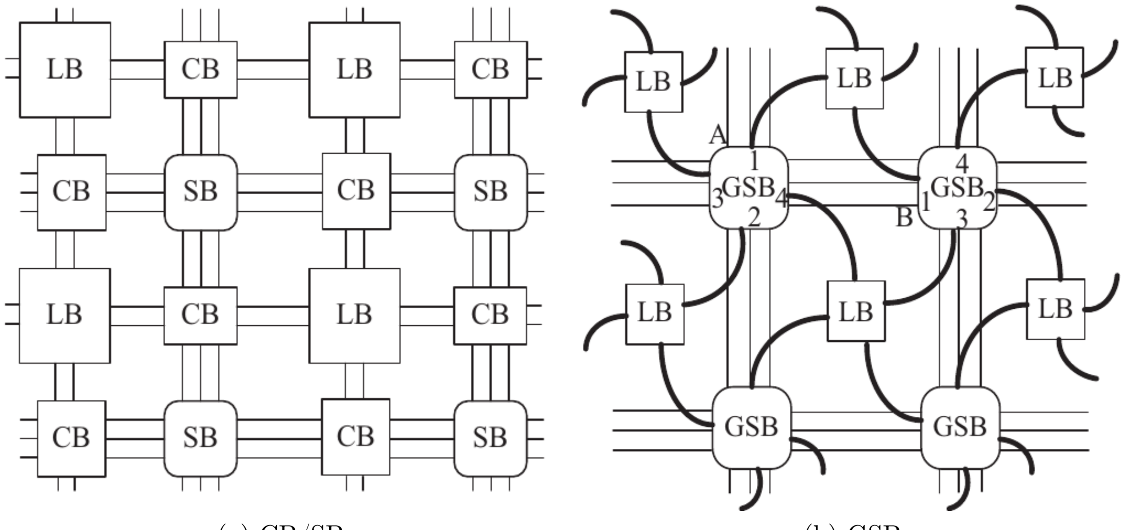
Logic blocks, also referred as configurable logic blocks (CLBs), are basic components of FPGAs, since they implement the logic functions and are used for storage purposes. Commercial vendors, like Xilinx and Altera, use Look-Up Table (LUT) based CLBs [25]. A CLB can consist of a single basic logic element (BLE), or a cluster of locally interconnected BLEs (Figure 2.15). The basic architecture of a BLE consist of a LUT and a Flip-Flop. The output of a BLE is selected by a multiplexer, in order to implement combinational or sequential logic functions. A k-input LUT contains 2^k configuration bits to implement any logic function of k inputs.

Modern commercial FPGAs utilize the logic cluster approach for CLBs in order to exploit the gains that arise in the total critical path delay [39], and can contain a large number of BLEs, typically in the order of ten. In addition, many commercial FPGAs contain heterogeneous mixture of logic blocks with specific functionalities [25]. These specific blocks, also referred as hard blocks, include memory, adders, carry logic, multipliers, DSP blocks etc. Hard blocks are utilized to implement specific logic, arithmetic and memory functions, evading the waste in logic and routing resources.

2.6.3 Routing Architecture

The programmable routing network of an FPGA provides the required connections between logic and I/O blocks to implement the user's intended logic functions. The routing interconnects comprise of wire segments and programmable switches to accomplish the required connection. The switches are implemented by utilizing the programmable technology referred in the Subsection 2.6.1.

Routing interconnects must be very flexible, since FPGAs claim to be computing platforms able to implement almost any digital circuit. Many designs require local connections between logic blocks, hence short, fast, routing wires are necessary. However, when more distant connections are required, e.g., connecting logic with I/O blocks, the routing interconnect architecture should provide longer wires. It is clear that the accommodation of a wide variety of circuits establishes the necessity of flexible routing interconnects, as well as the efficiency in terms of area, speed performance and power consumption.



(a) CB/SB

(b) GSB

Figure 2.16: FPGA routing architectures [40].

The design of FPGA routing interconnects is critical, as more than 70% of the chip area is occupied by routing resources [40]. Furthermore, about 80% of the critical path delay arises from inter-CLB routing delay [54]. The most commonly used architecture both in academia and industry, is illustrated in Figure 2.16a. This architecture is referred as island-style routing architecture [39], and logic blocks (LBs) are connected utilizing connection blocks (CBs) and switch blocks (SBs). The routing architecture comprises of horizontal and vertical wires, which are interconnected through SBs. An input or output of a LB can only connect to the routing network through CBs. Besides the traditional island-style architecture, new approaches have been proposed [55]. In particular, the combination of SB and CB into a new general switch box (GSB) [40] has been studied. Figure 2.16b illustrates this new approach, where LBs can connect to each other through GSBs, achieving reduction in critical path delay compared to the island-style architecture [40].

2.7 Evaluation of Process Variability in FPGAs

Performance variation in commercial FPGAs has been studied by several works in the past. The most established method relies on ring oscillator (RO) sensors. In [21] and [23], the authors employed ROs to analyze the stochastic and systematic intra-die process variability in 90nm Cyclone II and 65nm Virtex-5 FPGAs, respectively. Furthermore, in [22,27,56,57] the authors used ROs to measure the intra-die variation in 90-28nm Spartan-3E, Virtex-4 ,Virtex-5 and Zynq FPGAs. In [22], the authors used 112 ROs and they measured 2.3% (σ/μ) intra-die delay variation in two 65nm Virtex-5 FPGAs. In [56], 6400 ROs were employed in the fabric of an 90nm Virtex-4 FPGA and the intra-die delay variation was measured 2% (σ/μ). Similar in [57], by using 2688 RO sensors in two 90nm Spartan-3E FPGAs, the intra-die variation was measured up to 14.1% ($3\sigma/\mu$) and the inter-die 7.6% (average value over the ROs).

Another method for variability evaluation is based on shadow registers. In [58], they evaluated the delay variation of 336 logic paths on a 65nm Virtex-5 FPGA, by placing additional shadow registers alongside the main paths' registers. To estimate the minimum delay of the respective paths, they were finely increasing the clock frequency until an error is detected in the comparison between the data of main and shadow registers. According to their experiments on a 65nm Virtex-5 FPGA, the maximum correlated variation was measured at 6.88%. Similar results arrived, i.e., 6.82% when used RO sensors as well. In [59], using the method of negative-skewed shadow registers, they evaluated the delay of three different logic paths of a floating point adder circuit, which was placed in five different locations on two 130nm Virtex-II FPGAs. They measured up to 25.7% intra-die ($(max - min)/avg$) and up to 16.6% inter-die variation.

In [24], an alternative technique is proposed for the evaluation of delay variability in FPGAs. The key idea is based on the placement of a combinatorial circuit under test (CUT) between a launch and a sampling register. A clock generator drives the clock of the registers and a stimuli generator provides inputs to the CUT. While stepping up the frequency, a custom circuit monitors the outputs of the CUT and the sampling register detects the occurrence of timing errors. Consequently, the maximum error-free frequency is derived. Utilizing this technique, they measured the delay variation of LUTs, carry-chain units and embedded multipliers in Cyclone II and Cyclone III FPGAs. Similarly in [60], by using the same method they measured the intra-die delay variation of 1024 logic CUTs in 12 65nm Virtex-5 FPGAs.

Contrary to the aforementioned, the differentiating parts of our work are:

- We study the performance variation in 16nm FinFET FPGAs under various voltage and temperature operating conditions.

- We evaluate process variability in a multifaceted fashion considering diverse types of RO and interconnect sensors.
- We analyze systematic and stochastic variability for both logic and interconnect resources.
- We perform correlation analysis on the variability results derived by the different sensors demonstrating the inconsistency in variation between the logic and interconnect resources.

Chapter 3

Sensing Infrastructure & Methodology

In this chapter, we describe our methodology used for the analysis of variability. We assess the performance variation of the configurable logic blocks (CLBs) and the routing interconnects, which are the most prevalent resources in the FPGA fabric. The proposed methodology is based on the generation of multiple variability maps for the characterization of process variations of the underlying FPGA and the performance variation under various operating (voltage, temperature) conditions. The variability maps are extracted by measuring multiple small, compact sensors deployed across the FPGA fabric.

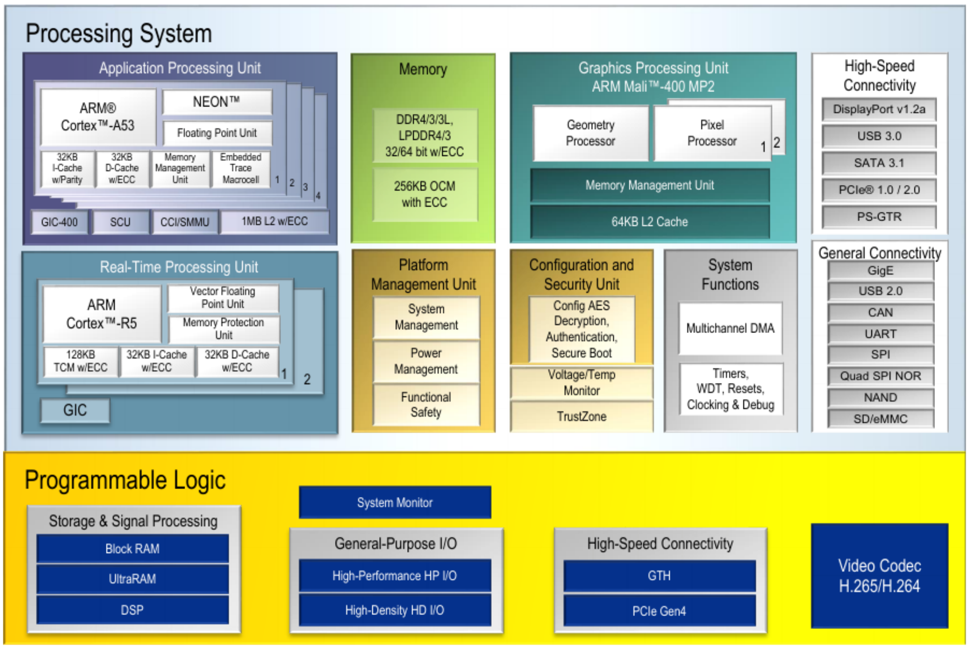


Figure 3.1: Zynq UltraScale+ MPSoC EV: Block diagram [41].

For our analysis, we employ the 16nm Zynq XCZU7EV devices, which are member of the Zynq UltraScale+ MPSoC EV family. As illustrated in Figure 3.1, these MPSoC (Multi-processor system-on-chip) devices, consist of the Processing System (PS) and the Programmable Logic (PL) part. The PS is equipped with two ARM CPUs, a quad core applications processor (Cortex-A53) and a dual core real-time processor (Cortex-R5) along with an embedded GPU (ARM Mali-400 MP2) and a variety of units, like DMA, voltage/temperature monitoring, timers etc. The PL comprise the traditional resources of the FPGA fabric, i.e., CLBs, Interconnect resources, DSPs, Block RAM etc. The communication between the PS and PL is established by the AXI protocol, which is based on the ARM Advanced Microcontroller Bus Architecture (AMBA). In our case, we utilize one of the ARM Cortex-A53 CPU core to control the operation of the sensors, collect their data and forward them to an external Host PC for further analysis.

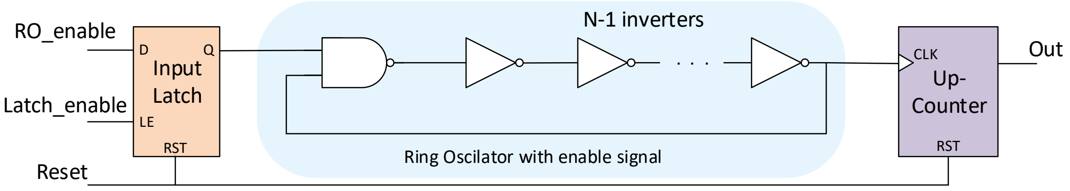


Figure 3.2: Proposed ring oscillator design.

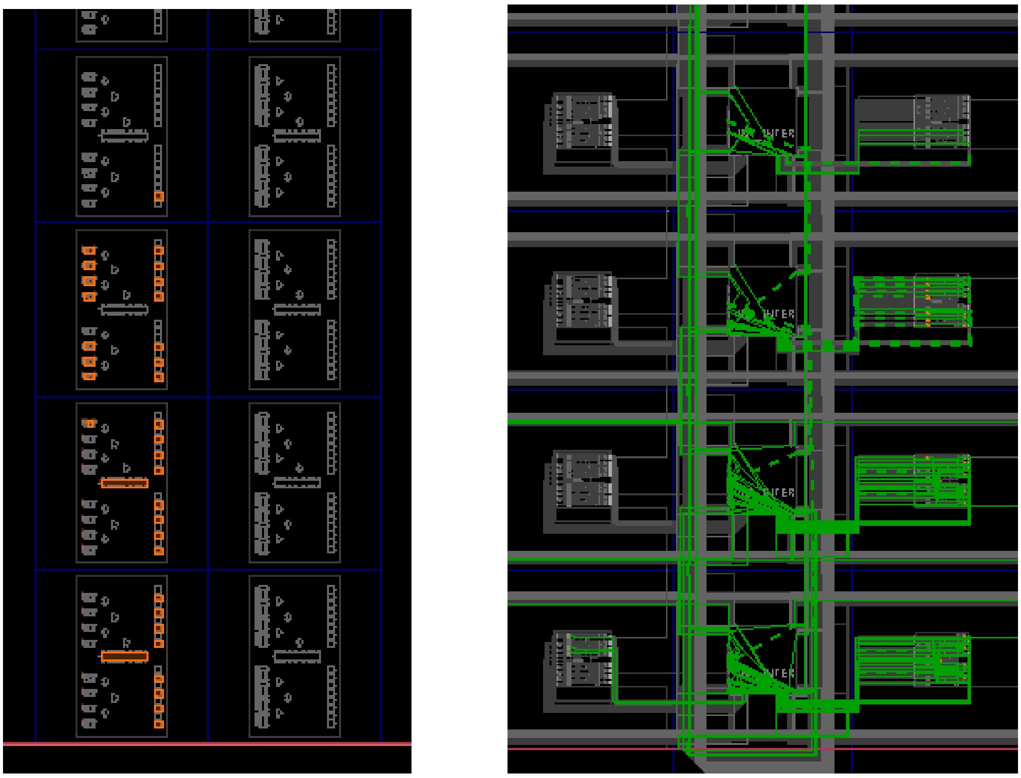
3.1 Custom Sensor Design & Network

The fundamental sensing circuit is based on the well established ring oscillator (RO) approach, as proposed by other similar works in the literature [22, 26, 27]. A traditional RO is an asynchronous loop of N-stage inverter gates, where N is an odd number, such that the loop becomes unstable and a square wave signal is generated in the output. Our sensing RO infrastructure is depicted in Figure 3.2, where $N-1$ inverters are followed by a NAND gate. The role of the NAND gate is to activate the RO operation (oscillation) by a given signal (the "*RO_ enable*" signal, which results from an input latch). The square wave output of the RO is fed to an up-counter in order to operate as a clock signal. In this way, the rising edges of the square signal can be measured. Upon initialization of the RO, it's activation is maintained for a predefined time period T. By representing the counter output as C_{ro}, we can approximately calculate the RO loop delay, T_d, as:

$$T_d = \frac{T}{2 \cdot C_{ro}} \tag{3.1}$$

The factor "2" arises from the fact that only rising edges are measured by the counter. The actual RO delay depends on the electrical properties of the region where the RO is mapped on. By placing multiple identical ROs across the FPGA fabric, we can evaluate the speed of the corresponding regions and hence, calculate the intra-die variability.

It is essential that all the employed ROs must be constructed with the exact same CLB resources and routing connections to obtain precise results regarding variability. To ensure this identity, we build an RO soft-macro block, which is replicated accompanied by particular physical constraints that specify its placement and routing on the FPGA fabric. First, to prevent the optimization of the inverters chain, we use synthesis constraints in the VHDL code (for Xilinx tools the attributes *dont_ touch*, *keep* and the constraint *flatten_ hierarchy*). Second, the underlying RO should be placed and routed at specific logic and routing resources in the FPGA fabric, respectively. Figure 3.3 illustrates the floorplan view of an 7-stage ring oscillator, accompanied by a 16-bit counter, as proposed above (Figure 3.2). However, a subtly different design to the proposed has been developed: each inverter is followed by a pass-through D Flip Flop (DFF), operating as an open-latch, to increase the portion of the delay arising from logic resources. Figure 3.3a depicts the constrained logic resources (input Latch, LUT, pass-through DFFs), which are placed in the two upper CLBs, while the two lower are utilized for the counter mapping. All resources are mapped to predefined, by the user, locations, using the constraint commands (for Xilinx tools the constraints *LOC*, *BEL*). Additionally, Figure 3.3b shows the routing resources (with green color). The constrained routing resources are noted with dotted lines, and are indispensable in order to achieve the intended identical routing resources for the RO. In the same way like the logic resources, constraint commands are used for fixing LUT pins and routing paths (for Xilinx tools the constraints *LOC_PIN*, *FIXED_ROUTE*).



(a) Constrained logic resources (orange color)

(b) Constrained routing resources (green color)

Figure 3.3: Floorplan view of a 7-stage ring oscillator accompanied with a 16-bit counter (Figure 3.2).

The implemented constrained RO macro-block, is deployed in numerous copies and placed across the FPGA fabric. The deployment is automated by a parametric VHDL code and a custom Python script, which generates the constraint file of the RO network. The Python script receives as input the constraints for an individual RO and the coordinates of the locations where ROs are going to be placed across the FPGA. As a result, our sensor macro-block and the sensor network are fully parametric in terms of RO stages, number of sensors and mapping locations.

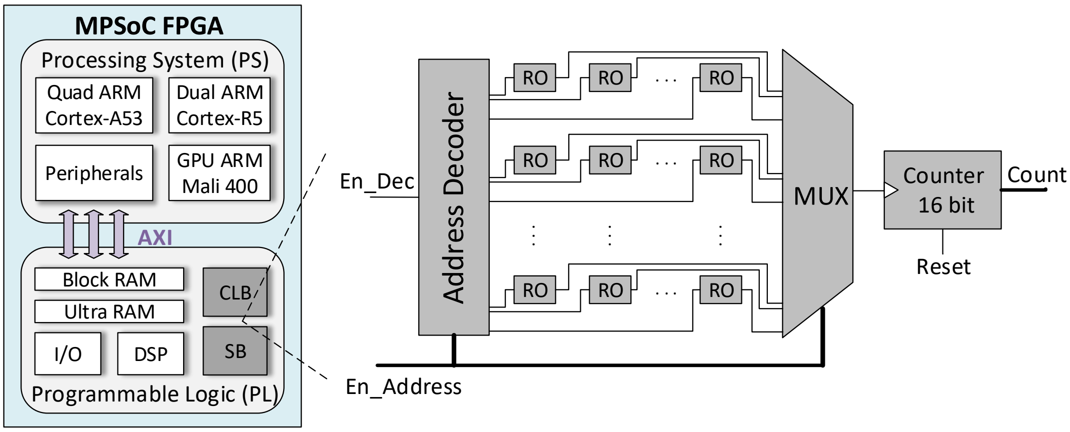


Figure 3.4: Block diagram of the proposed architecture.

The proposed sensor network is illustrated in the Figure 3.4. As already mentioned, in this work the CLB and routing resources of the Programmable Logic have been studied. The abbreviation SB in the Figure stands for Switch Box, a term that is used for the tested MPSoC FPGA devices, which utilizes the GSB architecture that has been discussed in the sub-section 2.6.3. Our custom architecture employs a shared up-counter which is multiplexed with all the sensors of the network to measure the delay of each individual RO sensor. Contrary to the scenario of a private counter per RO [61], the shared use of a single counter results in reduced resource overhead, and enables the employment of higher number of ROs and consequently, a more fine-grain evaluation of variability. The operation of ROs is performed sequentially to avoid potential voltage drops that could affect the evaluation of the results. The selection of a specific RO sensor for operation is specified by the address decoder unit, which is controlled by the ARM CPU via the *En_ Address* signal. Notice that the same signal is also applied as a selection signal in the multiplexer. The communication between the ARM CPU and the sensor network is obtained via the AXI-Lite interface. The operation period T, where each individual RO remains active is calculated by the private timer of the ARM CPU and is selected to be 50 μs as proposed by [21] to avoid self-heating phenomena [28] and mitigate the error in the measurement process. Overall, when including the non-ideal timer operation, the quantization issue due to the non-aligned operation of timer and sensors [62] and the micro-fluctuations in voltage and temperature, the measurement error in our procedure is calculated to be less than 0.2%. To alleviate this error, we determine the RO delay as the average value of 10 consecutive 50 μs runs.

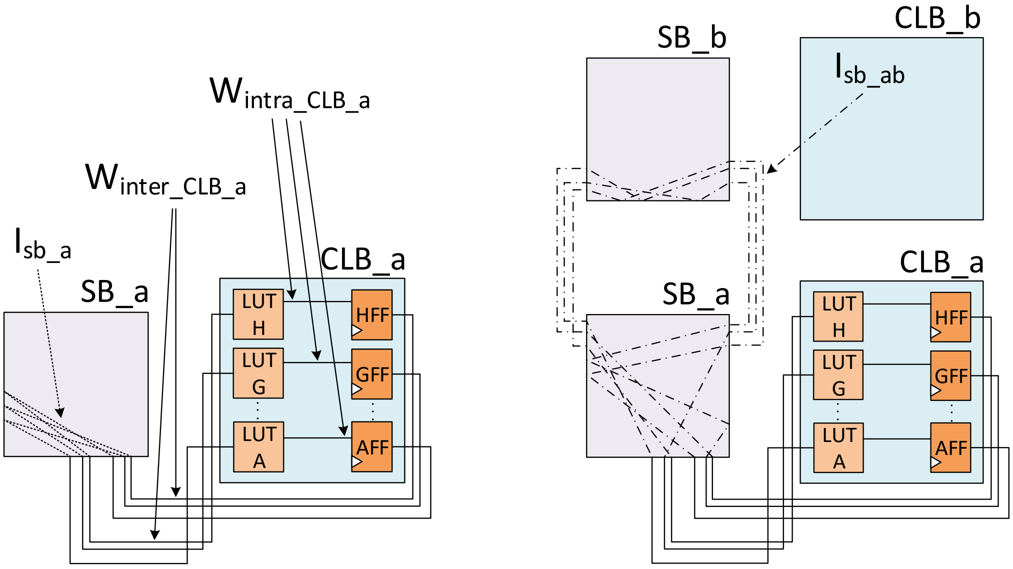
3.2 Assessment Approach

We provide an extensive variability assessment methodology which includes:

- Various sensor configurations.
- Decoupling of variability into systematic and stochastic.
- Diverse voltage and temperature conditions.

3.2.1 Variety of sensor configurations

Owning the parametric implementation of our RO sensor, we utilize various configurations with different resource and delay characteristics. This serves a twofold purpose. First, we need to investigate how the derived variability results are affected by the footprint of the sensor. Second, we need to analyze the impact of variability on logic and interconnect resources. To do so, we utilize RO configurations with different fraction of logic and interconnect delays. The employed RO sensors have been designed with respect to the referred, in the previous section, architecture: each inverter stage is followed by a pass-through Flip Flop. The delay attributed to logic and interconnect resources is specified via the custom mapping of the sensor on the FPGA fabric by using the floorplan utility of the Xilinx Vivado tool. We clarify that the term "interconnects" in our work refers to intra-, inter-CLB wires and switch boxes (SB)[1]. Since the Vivado tool does not distinguish between inter-CLB wires and SBs, we put their delays together under the same category of inter-CLB interconnects.



(a) "Nst_1sb"

(b) "Nst_2sb"

Figure 3.5: RO architectures with identical CLB resources and different routing resources for the same value of N.

The principal architecture of our RO sensors is illustrated in Figure 3.5. To a great extent, the designed sensors occupy exactly one CLB with respect to the RO loop. An important remark is that the UltraScale CLB consists of eight BLEs (LUT and its corresponding Flip Flop), labelled from "A" to "H", from bottom to top. For instance, in the aforementioned Figure the bottom Look-Up table is labelled as "LUT A", while the respective Flip Flop as "AFF". More information about the CLB UltraScale architecture can be found in [63]. In our approach, we have designed two

[1] The term "interconnects" can be frustrating as it may refer to metal interconnects (wires) or to routing interconnects, including the switch boxes, which comprise of switching transistors besides metal wires. The policy used in this thesis is to refer to wires severely as "metal interconnects".

diverse RO architectures. The first one, referred as “Nst_1sb”, is designed with the restrict of minimizing as much as possible the delay of routing interconnect resources. This has been achieved with the assistance of the Xilinx Vivado tool, evaluating the minimum achievable delay as reported by the tool, while simultaneously placing each individual inverter stage to a distinct LUT and its corresponding Flip Flop (Figure 3.5a). The second RO architecture is referred as “Nst_2sb” and utilizes routing from two SBs: the directly connected to the occupied CLB and the exactly upper as shown in Figure 3.5b. We note that, all SB-SB routing is based on short wire segments (direct connection between the SB tiles) [64]. An important attribute of the proposed sensors is that the CLB resources for a constant value of the number of stages N, are exactly the same for a given location. Elaborating further on that, Figure 3.5 depicts such a case, with $W_{intra\_CLB\_a}$ and $W_{inter\_CLB\_a}$ being identical in both sensors. Taking advantage of our ROs feature, by subtracting the measured delays of the two different sensors, referring on the exact same location, we isolate and calculate the delay of the remaining inter-CLB interconnects, i.e., $I_{sb\_ab} - I_{sb\_a}$. The I_{sb_ab} and I_{sb_a} have been carefully selected to avoid any overlap between their routing. Thus, we create extra sensors, named “Nst_inter”, which enable us to measure and characterize the interconnects individually [65].

Summarizing, “Nst_1sb” ROs have been designed with as small as possible proportion of routing resources, which as a result, leads to the augmentation of the delay of logic resources. “Nst_2sb” RO employs exactly the same logic resources as “Nst_1sb” with the same value of N, while its routing resources have been designed to utilize the resources of two SBs. Finally, the subtraction of the common part (logic resources) of the two aforementioned RO sensors gives us the ability to isolate the interconnects and assess them individually.

Table 3.1: STA delay of various sensor configurations.

sensor conf.	**delay of logic resources**			**delay of interconnects**			**total (ps)**
	LUTs	DFFs	Total	intra-CLB	inter-CLB	Total	
7st_1sb	707 ps	463 ps	65.4%	295 ps	325 ps	34,6%	1790
7st_2sb	707 ps	463 ps	36.3%	295 ps	1762 ps	63,7%	3227
7st_inter	-	-	-	-	1437 ps	100%	1437
5st_1sb	582 ps	309 ps	67%	196 ps	244 ps	33%	1331
5st_2sb	582 ps	309 ps	37,2%	196 ps	1305 ps	62,8%	2392
5st_inter	-	-	-	-	1061 ps	100%	1061

In Table 3.1, we provide details regarding sensor configurations. We employ ROs of N=5 and N=7 stages, as explained above. For each sensor configuration, we distinguish the delay attributed to logic and interconnect resources as reported by static timing analysis (STA) tool. Notice that RO sensor configurations with the same number of stages, e.g., “7st_1sb”, “7st_2sb”, have the same delay of logic resources, as proposed above. Furthermore, in the configurations that consist of a single SB, the logic delay dominates the total delay, i.e., 65,4% and 67% for 7 and 5 stages, respectively, as expected. The opposite applies in the case of two SBs, i.e., interconnects dominate the total delay with 63.7% and 62,8%, respectively. Finally, the “7st_inter” and “5st_inter” sensors, apparently have only interconnect delay, as their resources take part outside the CLB (inter-CLB).

3.2.2 Decoupling of Variability

We decouple the total measured variability into systematic and stochastic in order to study the impact of each individual type separately. In presence of variability, the delay of a path can be expressed as a random variable following the first order canonical form, as arises from SSTA research [29]:

$$T_d = T_d^{\mu} + T_d^{S} + T_d^{R} \tag{3.2}$$

T_d^{μ} represents the mean or nominal value, T_d^{S} represents the systematic part and T_d^{R} the random/stochastic part. The T_d^{μ} is a constant value, while the T_d^{S} is spatially correlated, changing gradually from one location to other and T_d^{R} has no spatial correlation. Spatial correlated variations arise from manufacturing process, such as systematic lithography variations and chemical-mechanical polishing (CMP), while stochastic variations result from intrinsic, atomic scale fluctuations, such as random dopant fluctuations (RDF) and line-edge roughness (LER) (Section 2.2). Therefore, the delay of all sensors can be expressed by equation 3.2.

The above modelling for variations is used in SSTA, where typically process variations are treated statistically, while environmental, i.e., temperature and voltage, and aging fluctuations are modeled using safety margins [44]. An important remark of our work is that we minimize environmental and aging variations, due to the measurement techniques we utilize:

- We enable an individual RO each time, thus avoiding any potential voltage drop, so voltage fluctuations are minimized (Section 3.1).
- We left the measurement system to reach its thermal equilibrium and each RO is activated for a small period of time, avoiding self-heating phenomena (Section 3.1). Therefore, temperature variations are minimized as well.
- From the above statement, the small period of time that ROs are enabled does not cause aging phenomena, due to the fact that they have, at least, time constants in the order of days (Section 2.5).

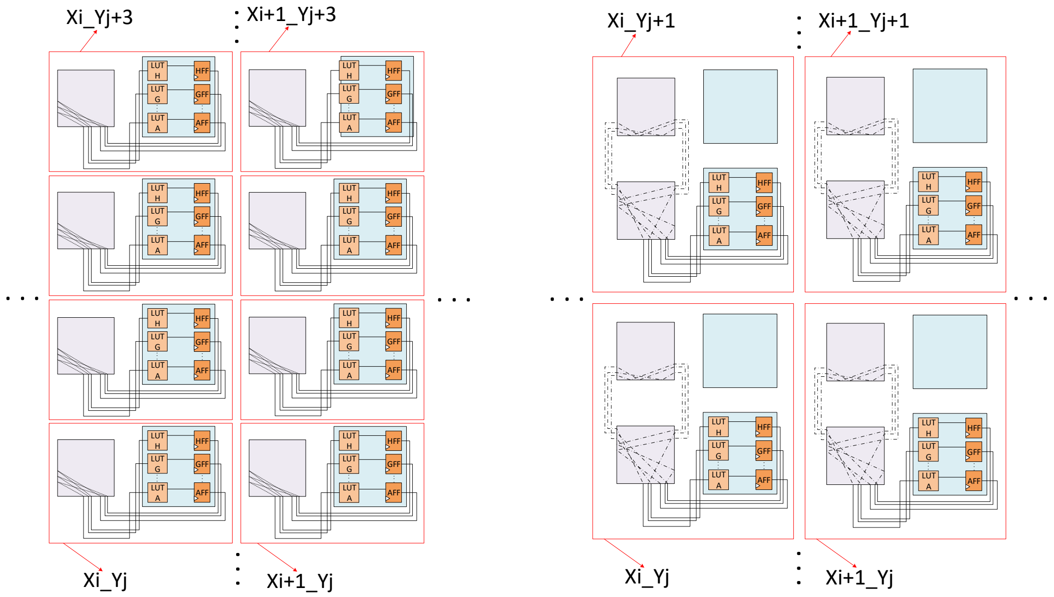
As a matter of fact, we can adequately assess process variations with the proposed technique and equation 3.2. In SSTA approach device parameters such as the gate length, doping concentration, gate oxide thickness and wire width and thickness, are treated as random variables due to process variation. These random variables represent both systematic and stochastic variations. A more detailed expression of the canonical form (equation 3.2) is [29, 44]:

$$T_d = \mu_d + \sum_{n=1}^{n} d_i z_i + d_{n+1} R \tag{3.3}$$

μ_d is the mean or nominal delay, z_i represents the n *independent* random variables used to express the spatially correlated parameter variations, both for transistors and wires, R represents the residual independent variation, and coefficients d_i represent the sensitivity of the delay to each of the random variables.

Since equations 3.2, 3.3 are equivalent, apparently:

$$T_d^{\mu} = \mu_d \qquad T_d^S = \sum_{n=1}^{n} d_i z_i \qquad T_d^R = d_{n+1} R \tag{3.4}$$



(a) "Nst_1sb"

(b) "Nst_2sb"

Figure 3.6: The FPGA fabric is modeled as a grid, each point (red orthogonal border) representing an RO sensor.

For the purpose of our analysis, we utilize the grid model [30]. According to that, the FPGA fabric is modeled as a X-Y grid, where each point on the grid represents an sensor. Figure 3.6 represents the grids for our two RO sensor design architectures: "Nst_1sb" and "Nst_2sb". Each grid for the RO sensor "Nst_1sb" contains exactly one CLB and its corresponding SB, while for "Nst_2sb" the occupied area is twofold, as shown in Figure 3.6b. According to the grid model, we assume perfect correlations among all transistors and among all wires in the same grid [30]. Therefore, all sensor's resources have perfectly correlated spatial variation, as they are closely located[2]. The perfect correlation among parameters, physically means that the systematic variations of two transistors (or wires respectively) inside the grid are identical, thus they exhibit proportional speeds, and their actual values depends on their dimensions (e.g., gate length and width for transistors). Mathematically, this can be expressed with the correlation coefficient (ρ) inside the boundary of a grid, of the random variables expressing the spatial correlated variations for a determined physical parameter, e.g., doping concentration, being exactly 1. Furthermore, since the systematic part of the delay T_d^S, is spatially correlated due to the grid model (FPGA fabric is modeled as a X-Y grid), can be expressed as a function of (x, y), while T_d^R has no spatial correlation and can be expressed as a random variable following a normal distribution $(0, \sigma^2)$ [21, 23].

[2] This assumption was also verified in practice for both of the two RO architecture design approaches.

For instance, consider two path delays T_a and T_b, for paths "a" and "b", respectively which are located inside the boundary of a grid. Then, according to equation 3.3 the path delays can be expressed as:

$$T_a = \mu_a + \sum_{n=1}^{n} a_i z_i + a_{n+1} R \tag{3.5}$$

$$T_b = \mu_b + \sum_{n=1}^{n} b_i z_i + b_{n+1} R \tag{3.6}$$

The sum of the aforementioned delay distributions, $T_c = T_a + T_b$, can also be expressed in canonical form and its coefficients can be easily computed [44]:

$$\mu_c = \mu_a + \mu_b \tag{3.7}$$

$$c_i = a_i + b_i, \quad \forall i : 1 \leq i \leq n \tag{3.8}$$

$$c_{n+1} = \sqrt{a_{n+1}^2 + b_{n+1}^2} \tag{3.9}$$

Two important remarks should be noted here. First, the coefficients representing the spatial correlated variations (equation 3.8) are added linearly, due to their perfectly correlated spatial variations inside the boundaries of the grid. In particular, assuming $a_k z_k$ and $b_k z_k$ to be the two random variables expressing the doping concentration variations impact on the path delays, accordingly, then $\rho(a_k z_k, b_k z_k) = 1$. Note that, $a_k z_k$ and $b_k z_k$ are random variables, and consequently the addition or subtraction of them cannot be expressed as two actual values, but on the contrary, should be considered as sample spaces. However, assuming the correlation coefficient to be 1, we can obtain the linearity on the coefficients due the addition of these random variables[3]. Second, notice that since the last term represents independent stochastic variations, the standard deviation is computed by the square root summation of the individual independent contributions[4].

As mentioned, each path delay can be expressed with the first order canonical form (equation 3.2). Hence, when specifically considering the interconnects, we can assert that for each interconnect path delay, applies:

$$T_{inter} = T_{inter}^{\mu} + T_{inter}^{S} + T_{inter}^{R} \tag{3.10}$$

In our case, the methodology requires the subtraction of RO delays, i.e., the delay of "Nst_2sb" minus the delay of "Nst_1sb", to derive the "Nst_inter" sensor for sufficient isolation of routing interconnect resources. That results also in a random variable and is expressed as:

$$\begin{aligned} T_{Nst\_inter} &= T_{Nst\_inter}^{\mu} + T_{Nst\_inter}^{S} + T_{Nst\_inter}^{R} \\ = \left(T_{Nst\_sb\_ab}^{\mu} - T_{Nst\_sb\_a}^{\mu}\right) &+ \left(T_{Nst\_sb\_ab}^{S} - T_{Nst\_sb\_a}^{S}\right) + \left(T_{Nst\_sb\_ab}^{R} - T_{Nst\_sb\_a}^{R}\right) \end{aligned} \tag{3.11}$$

[3] The proof can be found in the Appendix's equation A.5.

[4] The proof can be found in the Appendix's equation A.6.

Notice that, the above expression entirely comprises of delay interconnects terms (see Figure 3.5). Each term expresses the mean delay $T^{\mu}_{Nst\_inter}$, the systematic spatial correlated delay $T^{S}_{Nst\_inter}$ and the residual stochastic delay $T^{R}_{Nst\_inter}$. These variables/paths have overlapped parts (identical CLB resources, Figure 3.5), but their subtraction leads to the random variable T_{Nst_inter}, derived by two independent parts ($T_{Nst\_sb\_ab}$ and $T_{Nst\_sb\_a}$ with no physical overlap). The subtraction inside $T^{S}_{Nst\_inter}$ expresses accurately the systematic part of T_{Nst_inter} as the spatial correlation of $T^{S}_{Nst\_sb\_ab}$ and $T^{S}_{Nst\_sb\_a}$ is assumed to be 1, as mentioned above due to their closely located routing (grid model). That is to say, directly subtracting RO delays is sufficient for calculating $T^{S}_{Nst\_inter}$. However, when considering the stochastic parts, since $T^{R}_{Nst\_inter}$ is the difference of two *statistically independent* variables (no overlap in their resources), the subtraction of individual RO delays would typically follow the normal distribution $(0, \sigma^2_{Nst\_sb\_ab} + \sigma^2_{Nst\_sb\_a})$[5] and would not be correct for our analysis. Instead, we need to derive $\sigma^2_{Nst\_sb\_ab} - \sigma^2_{Nst\_sb\_a}$, due to the subtraction that we attempt for our purposes. In order to achieve that, we first calculate the variances of "Nst_2sb" and "Nst_1sb" independently for each RO set, and we subtract them afterwards [65].

3.2.2.1 Variability Decoupling Methods

Additionally to the above analysis for variability modeling, we consider two distinct methods for the decoupling of variability, as proposed in the literature: the regression method [21] and the down-sampled moving average estimator (DSMA) [23]. In both methods, systematic variability is modelled as a function of (x, y), according to the grid model, while T^{μ}_{d} is constant and T^{R}_{d} is a random variable following the normal distribution $(0, \sigma^2)$. Hence, equation 3.2 is further elaborated as:

$$T_d(x, y) = T^{\mu}_{d} + T^{S}_{d}(x, y) + T^{R}_{d} \tag{3.12}$$

Furthermore, the residuals in both methods are utilized to estimate the stochastic variability. In this work, we test both methods, evaluate the results with what is expected from the literature and choose the most accurate one for our analysis. The two methods are analyzed briefly in the following.

3.2.2.1.1 Regression method

In regression method, systematic variation is modelled by a quadratic polynomial function of x and y, as being proposed in several works [21, 31, 66]. According to that, the delay of each individual sensor in the coordinate system (x, y) can be described as:

$$T_d(x, y) = \left(c_{00} + c_{10} \cdot x + c_{01} \cdot y + c_{20} \cdot x^2 + c_{11} \cdot xy + c_{02} \cdot y^2\right) + T^{R}_{d} \tag{3.13}$$

where the coefficients $c_{ij}, i, j = 0, 1, 2$ are computed by a least-square curve fitting algorithm in MATLAB. The residuals after the computation of the quadratic function are utilized to derive the stochastic variability T^{R}_{d}, which is not expected to have dependence on the (x, y) coordinates, because it expresses the stochastic variations.

[5] See Appendix's equation A.7.

3.2.2.1.2 DSMA method

In contrast, DSMA applies a moving average window across the die to calculate the values for each (x, y) coordinate. The DSMA value computes the systematic part of the delay and is expressed as the average delay of all oscillators inside the window in each location [23]:

$$DSMA(x, y) = \frac{\sum_{i=x-z}^{x+z} \sum_{j=y-z}^{y+z} T_d(x, y)}{(2z + 1)^2} \tag{3.14}$$

where z is the size of the square moving window. With this computation, the random part is attenuated by a factor of $(2z+1)$. The choice of the moving average is crucial for an essential estimation: a window that is too small would not remove sufficiently the stochastic variation components, while a window that is too large may remove some of the systematic components as well. In our case, a 5x5 window size was found to be the optimal according to some previous applications of the DSMA [23, 66] and our own experimentation.

Having computed the DSMA value in each coordinate, we obtain the ability to compute the stochastic variation since equation 3.12 can be rewritten as:

$$T_d(x, y) = DSMA(x, y) + T_d^R \tag{3.15}$$

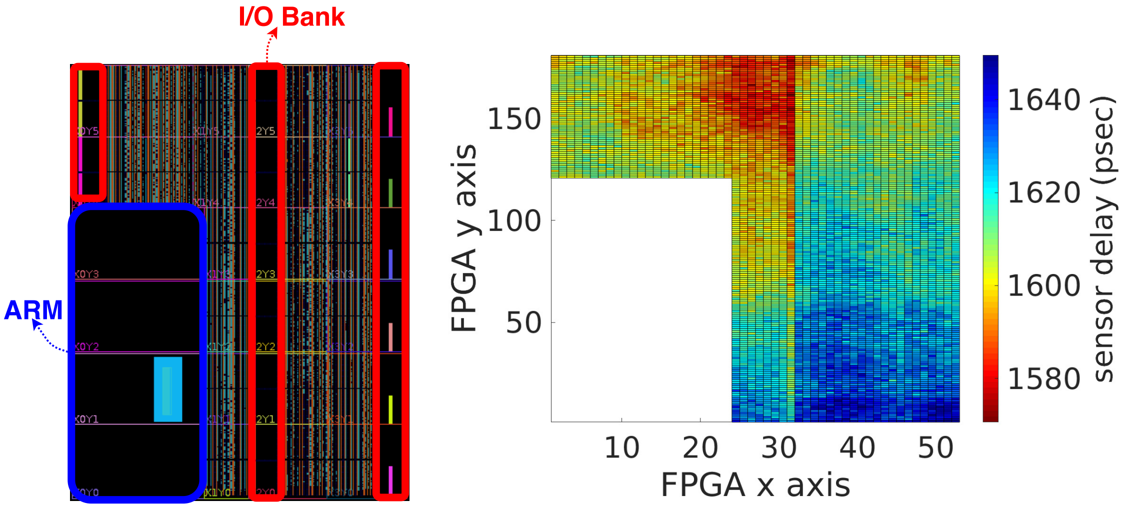
3.2.3 Diverse operating conditions

We assess performance variation under different voltage and temperature operating conditions. The assessment regards all sensor configurations and decoupling of variability. To perform voltage scaling we utilize the built-in I2C controller of the Zynq Ultrascale+ device (PS-subsystem), as well as the power management units (PMU) IRPS5401M from Infineon that supply the core and the auxiliary voltages of the Zynq US+ device on the ZCU104 development board. Specifically, through the I2C controller and the use of a custom-made software implementing the PMBus protocol, we have access to the PMU, which is responsible for supplying the FPGA fabric voltage. This enables us to alter the supply voltage, as well as to perform power consumption measurements on the specific voltage rail. To modify temperature, we employ a thermal chamber of uniform thermal distribution.

Chapter 4

Variability Analysis & Evaluation

In this chapter we present the variability analysis with respect to the methodology being described in the previous Chapter. We perform the assessment of variability in four supposedly identical Zynq XCZU7EV FPGAs. For each FPGA, we generate multiple variability maps by utilizing the manifold sensors of Table 3.1. For "Nst_1sb" RO configurations, which are implemented with one SB, we deploy 13200 sensors, while for the counterparts "Nst_2sb" with two SBs we deploy 6600 sensors due to their larger footprint on the fabric (Figure 3.6). All sensors are uniformly placed over the grid to cover the die sufficiently. An example of variability map with the corresponding floorplan view for the arbitrary selected device 1 is illustrated in Figure 4.1. The red color denotes the faster regions of the device (smaller RO delay), while the blue color denotes slower regions. We observe a smooth distribution of the performance across the die due to systematic variation, with a noticeable discontinuity in the middle column of our variability map (Figure 4.1b). This is explained by looking at the floorplan of the FPGA (Figure 4.1a): the corresponding physical area is utilized by I/O Banks.



(a) Floorplan

(b) Variability map

Figure 4.1: FPGA floorplan and variability map of sensor "5st_2sb".

4.1 Total Variability Analysis

In this section we assess the total variability (without decoupling into systematic and stochastic) in nominal environmental conditions: supply voltage $V_{ccint} = 0.85V$ and junction temperature $T_j = 30°C$. Ambient temperature was held constant as much as possible and the system was left until it reached thermal equilibrium. This is very important, as we compare various configurations and we subtract to estimate the performance of interconnects, according to the analysis in the Subsection 3.2.1. To achieve that, supply voltage and junction temperature has been sampled by the Xilinx integrated system monitor (SYSMON) with precision +/-1% and +/-4°C for supply voltage and temperature respectively.

Table 4.1: Total measured performance variation results for nominal conditions.

(a) Intra-die variability for each device

Sensor	**device 1 (ps)**					**device 2 (ps)**				
	σ_{total}	μ	σ_{total}/μ	μ vs STA	$range/min$	σ_{total}	μ	σ_{total}/μ	μ vs STA	$range/min$
7st_1sb	11.1	1255.3	0.89%	29.9%	5.51%	15.8	1227.2	1.29%	31.4%	6.55%
7st_2sb	19	2172.5	0.87%	32.7%	5.05%	21.7	2133	1.02%	33.9%	4.95%
7st_inter	8.9	917.2	0.97%	36.2%	4.82%	6.4	905.8	0.70%	37%	3.85%
5st_1sb	8.2	922.4	0.89%	30.7%	5.8%	11.8	901.4	1.31%	32.3%	7.3%
5st_2sb	14.1	1611.7	0.87%	32.6%	5.02%	16.1	1581.4	1.02%	33.9%	4.96%
5st_inter	6.8	689.4	0.99%	35%	5.19%	4.8	680	0.71%	35.9%	4.15%

Sensor	**device 3 (ps)**					**device 4 (ps)**				
	σ_{total}	μ	σ_{total}/μ	μ vs STA	$range/min$	σ_{total}	μ	σ_{total}/μ	μ vs STA	$range/min$
7st_1sb	14	1243.3	1.13%	30.5%	6.15%	5.7	1216.3	0.47%	32.1%	3.95%
7st_2sb	19.7	2144.7	0.92%	33.5%	4.77%	7.4	2112.7	0.35%	34.5%	2.92%
7st_inter	6.2	901.3	0.68%	37.3%	4.04%	3	896.3	0.33%	37.6%	2.62%
5st_1sb	10.4	913.3	1.14%	31.4%	6.29%	4.4	893.1	0.50%	32.9%	4.09%
5st_2sb	14.7	1589.9	0.92%	33.5%	4.82%	5.6	1566.2	0.36%	34.5%	2.94%
5st_inter	4.7	676.6	0.69%	36.2%	3.98%	2.4	673.1	0.36%	36.6%	2.76%

(b) Inter-die variability among the 4 devices

Sensor	**7st_1sb**	**7st_2sb**	**7st_inter**	**5st_1sb**	**5st_2sb**	**5st_inter**
$range/min_{among\_devs}$	8%	7.01%	6.44%	8.31%	6.93%	6.83%

Table 4.1 provides detailed total variability results for our four devices. The metrics that have been reported for quantifying the intra-die variability are (Table 4.1a): the mean sensor delay μ, the standard deviation σ_{total}, the ratio σ_{total}/μ, the difference between the STA estimation (Table 3.1) versus the actual mean sensor delay (vs STA), as well as the estimation of variability expressed by the $range/min$ metric, where $range = max - min$ refers to the maximum and minimum sensor delays. Furthermore, in Table 4.1b the inter-die variability among devices is presented as the ratio of $range/min = (max - min)/min$, where max, min attribute to the extreme values among the four devices. Note that, this Table includes a negligible error in the σ_{total}/μ of interconnects stochastic parts of variation, due to our methodology of subtracting delays (Subsection 3.2.2). However, it is insignificant for the total variability, because it occurs only in stochastic parts and it is measured to be 0.03% in the σ_{total}/μ.

The first important observation is the great difference between the STA and the actual measured delay of the sensors, which tends to rise as the portion of the delay attributed to interconnects, increases, reaching up to 37.6% for “7st_inter”. Essentially, this indicates that the STA tool introduces more pessimism in interconnects rather than logic.

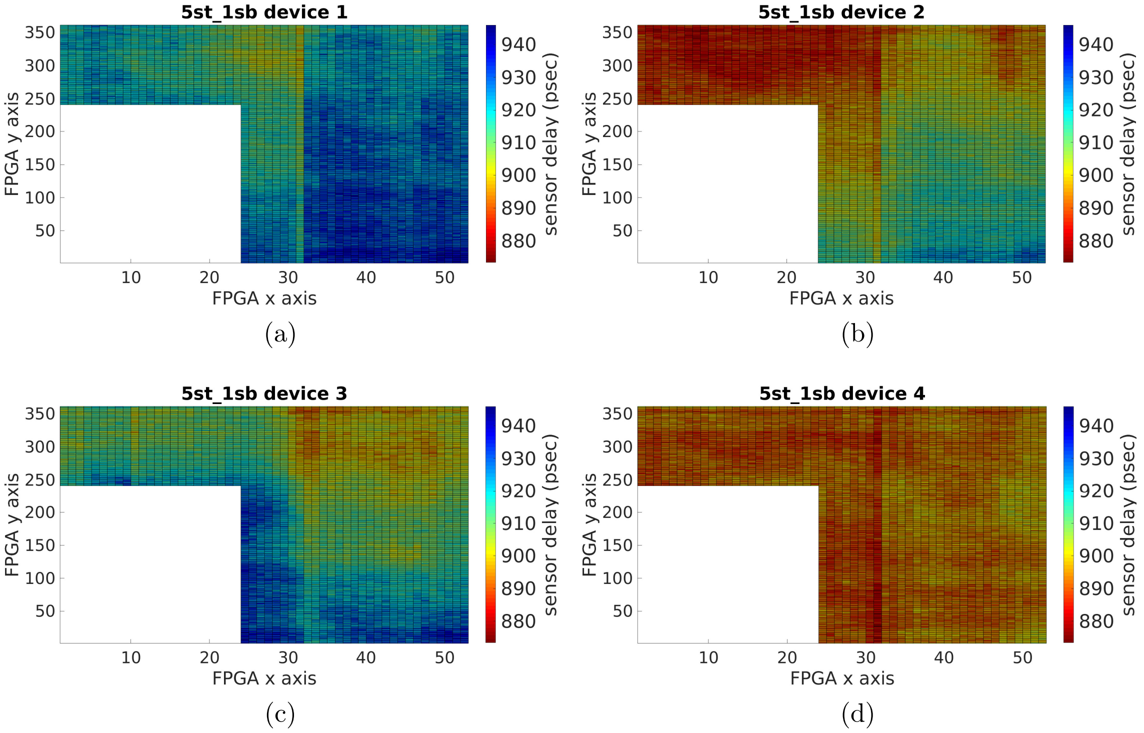
Second, concerning the mean delay (μ), device 4 is the fastest and device 1 is slowest among the others, for all sensor configurations. However, notice that for RO sensors with the highest portion of logic[1], i.e., “7st_1sb” and “5st_1sb”, device 2 is faster than device 3, while for interconnect sensors, i.e., “7st_inter” and “5st_inter”, device 3 is faster than device 2. This result indicates the importance of assessing manifold configurations for precise variability results, as interconnects and transistors are affected in a different manner from variability (see Chapter 2.7). The RO configurations “7st_2sb” and “5st_2sb” seem to follow the same trend as those with 1 SB (device 2 faster than device 3), but the relative difference among devices 2 and 3 is smaller. This is reasonable due to the fact that for these RO sensors the portion of interconnects is augmented contrary to the ones with 1 SB, thus interconnects affect them more.

Finally, regarding variability, we provide two metrics to measure it. The ratio σ_{total}/μ is a statistic metric which reveals how much does the delay disseminate over the mean delay, while the $range/min$ is a quantitative metric about how much the delay varies, in terms of extreme values. Among our devices, device 4 seems to be less affected by variations. For all devices except device 1, RO sensors with 1 SB point out greater variability than interconnect sensors. The same applies for RO sensors with 2 SBs, but the difference against interconnects sensors is smaller. For device 1, notice that when comparing the sensors “5st_1sb” and “7st_inter”, that have approximately the same mean delay (implying fair comparison), i.e., 922.4 ps and 917.2 ps respectively. When considering σ_{total}/μ, “7st_inter” has greater variability (0.97% $>$ 0.89%), while when considering the $range/min$ metric, “5st_1sb” has greater value (5.8% $>$ 4.82%). We should clarify that each metric is evaluated for different purposes, and when comparing sensor measurements, σ_{total}/μ provides more precise results about the comparison as it is a statistic metric, while the $range/min$ is an estimator about the value of variability. Consequently, it is obvious that for device 1 the variability of transistors is approximately as great as the variability in the measured interconnects[2]. The highest intra-die variation is measured in the smaller RO configuration ($5st\_1sb$) for device 2, reaching up to $range/min = 7.3\%$ and $\sigma_{total}/\mu = 1.31\%$. For inter-die variability the same applies as well, since “5st_1sb” has the greatest value of variability and interconnects sensors reveal smaller variations, compared to RO sensors, among all sensors. This is reasonable due to the fact that three out four of our devices have grater variability in transistor than wires, as explained above.

For a more comprehensible perspective of the variability, we provide the variability maps of the sensors with different architectures. We do not present the other variability maps because it is expected that sensors with same features, e.g., “5st_1sb” and “7st_1sb”, provide similar variability; this statement will be analyzed in the following sections, where a meticulous variability decomposition will be studied. Figures 4.2, 4.3 and 4.4 depict the variability maps of sensors “5st_1sb”, “7st_inter” and “5st_2sb” respectively. The first observation being made, is that

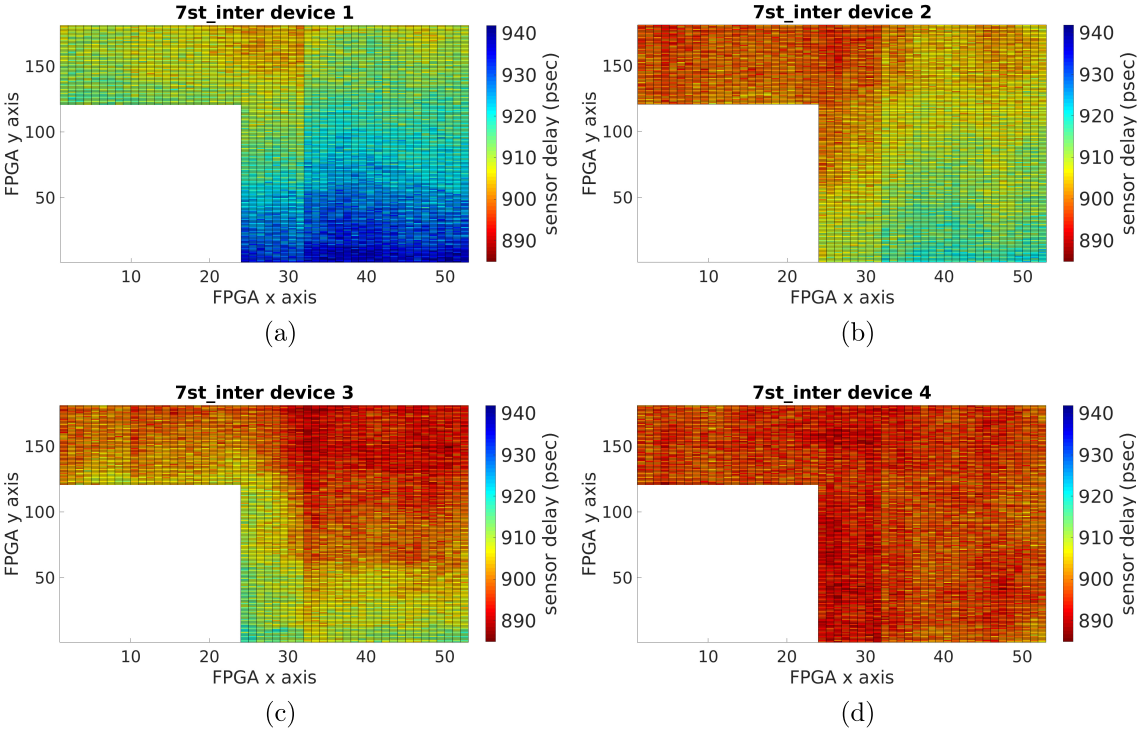
[1] For analytic delay portions of sensors see Table 3.1.

[2] Remember that only short wires are studied in this thesis, as described in the methodology.



(a) (b) (c) (d)

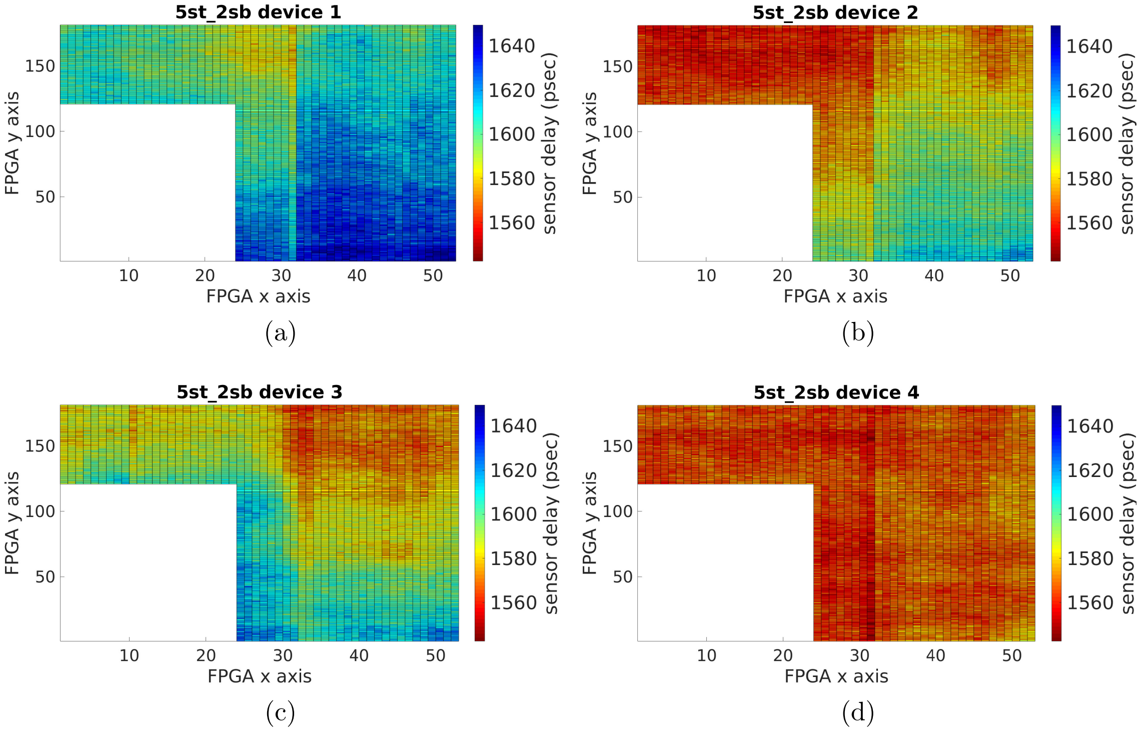
Figure 4.2: Total variability maps of sensor "5st_1sb" among our four devices (common scale).



(a) (b) (c) (d)

Figure 4.3: Total variability maps of sensor "7st_inter" among our four devices (common scale).

each device has systematic areas that could be characterized either as fast or as slow. This is very important because it implies that certain areas have similar variability, something encouraging for variation aware tools that can be implemented to take advantage of this attribute. Moreover, an important mention is that having as vision the exploiting of variability for application performance improvement, each device should be characterized individually, because the morphology of the variability is disparate for each of them, as can be pointed out from the presented maps.



(a)

(b)

(c)

(d)

Figure 4.4: Total variability maps of sensor "5st_2sb" among our four devices (common scale).

Note that we present the variability maps for sensors "5st_1sb" and "7st_inter" (Figures 4.2 and 4.3); we do that because they have approximately the same mean delay value as already mentioned (Table 4.1). However, notice that at a first sight, without mathematically identification, fast and slow areas do not map identically. In particular, in Figures 4.2a and 4.3a for device 1, there is a noticeable difference. This observation indicates the significance of obtaining multiple variability maps for device characterization. More results about these observations will be provided in the following sections, where variability is decomposed and each device is studied individually.

To have a connection with the remarks stated solely from the Table 4.1, it is irrefutable that device 4 has the lower variability among the others, and it seems, in all presented maps, that variability is equivalent and uniform across the device. However, this statement is insufficient because all maps presented are in the same scale for each sensor. Since device 4 has the smallest variations, it seems like the effect of variability is negligent. This is another implication about the importance of studying the variability individually for each device in order to obtain precise results. For the other devices, we can not obtain any unambiguous connection between the Table 4.1 and the variability maps. Thus, the importance of studying them both is compelling: variability maps are utilized to point out the morphology of the variation and its area extent, while on the contrary, a quantification can be presented without having the knowledge of the fast and slow locations, allowing an untimely quantified estimation of variations.

4.2 Analysis of Systematic and Stochastic Variability

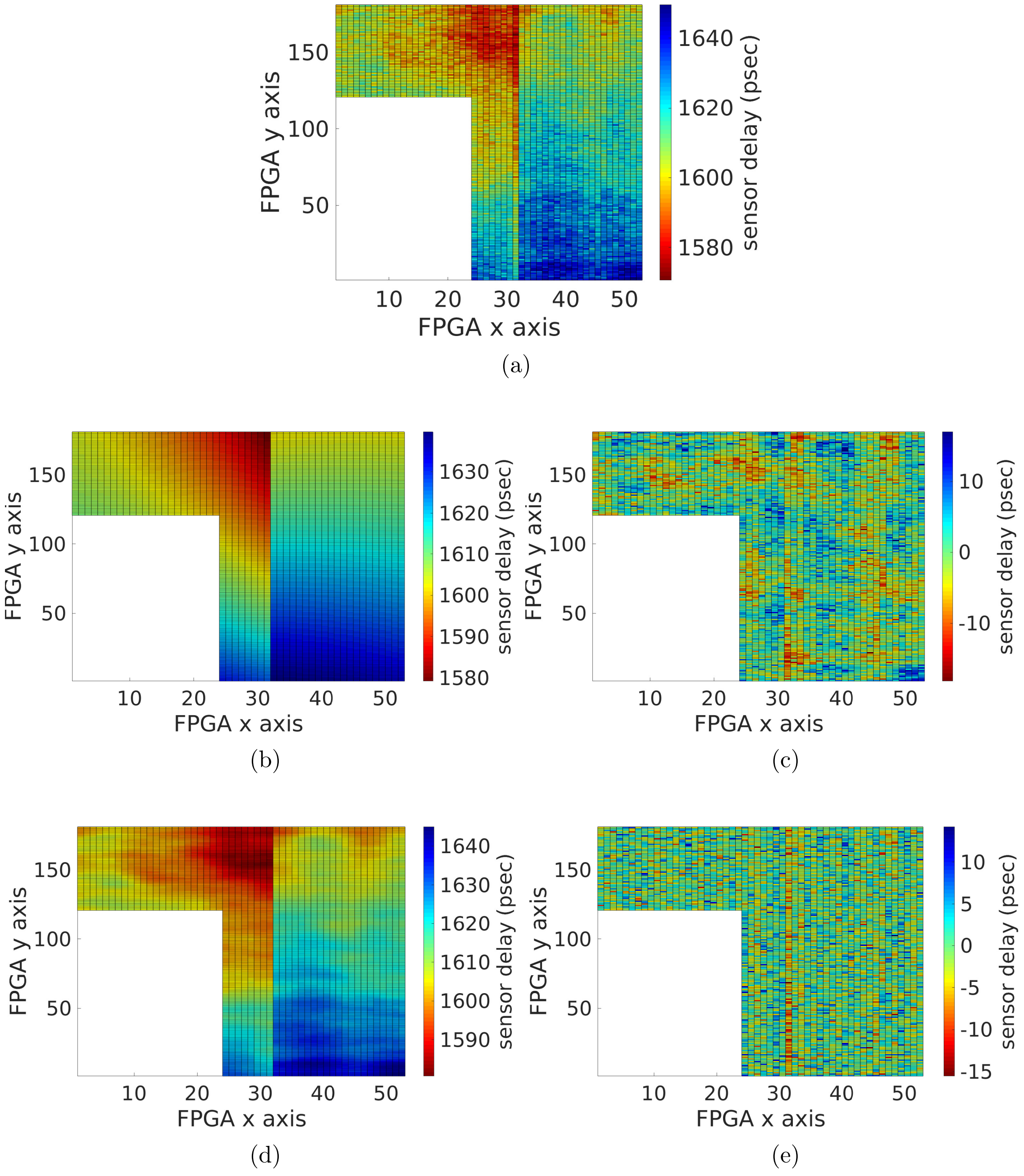
A step further in our analysis is to decouple the variability into systematic and stochastic using both the regression and the DSMA method. To address the discontinuity in variability maps (Figure 4.1), we empirically insert null columns until the systematic impact on the stochastic maps is minimized. In Figure 4.5, we show the extracted systematic (Figure 4.5b, 4.5d) and stochastic (Figure 4.5c, 4.5e) variability maps for the arbitrarily selected sensor "5st_2sb" of device 1. By comparing the initial map (Figure 4.5a) with the systematic resulted from both methods, DSMA extracts more precisely the systematic components and simultaneously highlights the random nature of the stochastic variations. On the contrary, the stochastic map computed by the regression method reveals systematic aberrations in various spatial regions. Unambiguously DSMA can perform more accurate decoupling of variations and therefore it is preferred. Thus, in the remainder of the dissertation, we continue our analysis based on the DSMA method.

Table 4.2: Systematic variation results for nominal conditions.

Sensor	**device 1 (ps)**				**device 2 (ps)**			
	σ_{sys}	$\sigma_{sys}/\sigma_{total}$	σ_{sys}/μ	$range/min$	σ_{sys}	$\sigma_{sys}/\sigma_{total}$	σ_{sys}/μ	$range/min$
7st_1sb	10.3	92.8%	0.82%	4.28%	15.4	97.5%	1.26%	5.47%
7st_2sb	18.3	96.3%	0.84%	4.05%	21.3	98.2%	1.00%	4.12%
7st_inter	8.6	96.6%	0.94%	4.01%	5.9	92.2%	0.65%	2.78%
5st_1sb	7.5	91.5%	0.81%	4.25%	11.4	96.6%	1.26%	5.54%
5st_2sb	13.4	95.0%	0.83%	4.07%	15.7	97.5%	0.99%	4.14%
5st_inter	6.5	95.6%	0.95%	4.12%	4.4	91.7%	0.64%	2.83%
Sensor	**device 3 (ps)**				**device 4 (ps)**			
	σ_{sys}	$\sigma_{sys}/\sigma_{total}$	σ_{sys}/μ	$range/min$	σ_{sys}	$\sigma_{sys}/\sigma_{total}$	σ_{sys}/μ	$range/min$
7st_1sb	13.3	95.0%	1.07%	5.25%	4.2	73.7%	0.34%	2.22%
7st_2sb	18.9	95.9%	0.88%	4.06%	5.65	76.4%	0.27%	1.66%
7st_inter	5.7	91.9%	0.63%	3.02%	1.9	63.3%	0.21%	1.34%
5st_1sb	9.8	94.2%	1.07%	5.15%	3.1	70.5%	0.34%	2.16%
5st_2sb	14	95.2%	0.88%	3.98%	4.1	73.2%	0.26%	1.7%
5st_inter	4.2	89.4%	0.63%	3.00%	1.4	58.3%	0.21%	1.41%

4.2.1 Systematic Variability

As already mentioned, systematic variability is a significant portion of the total variability. The systematic parts are very essential because they reveal the implication of a potential variation aware utility, targeting application performance improvement. In this perspective, the systematic parts across an FPGA can be potential,



(a)

(b)

(c)

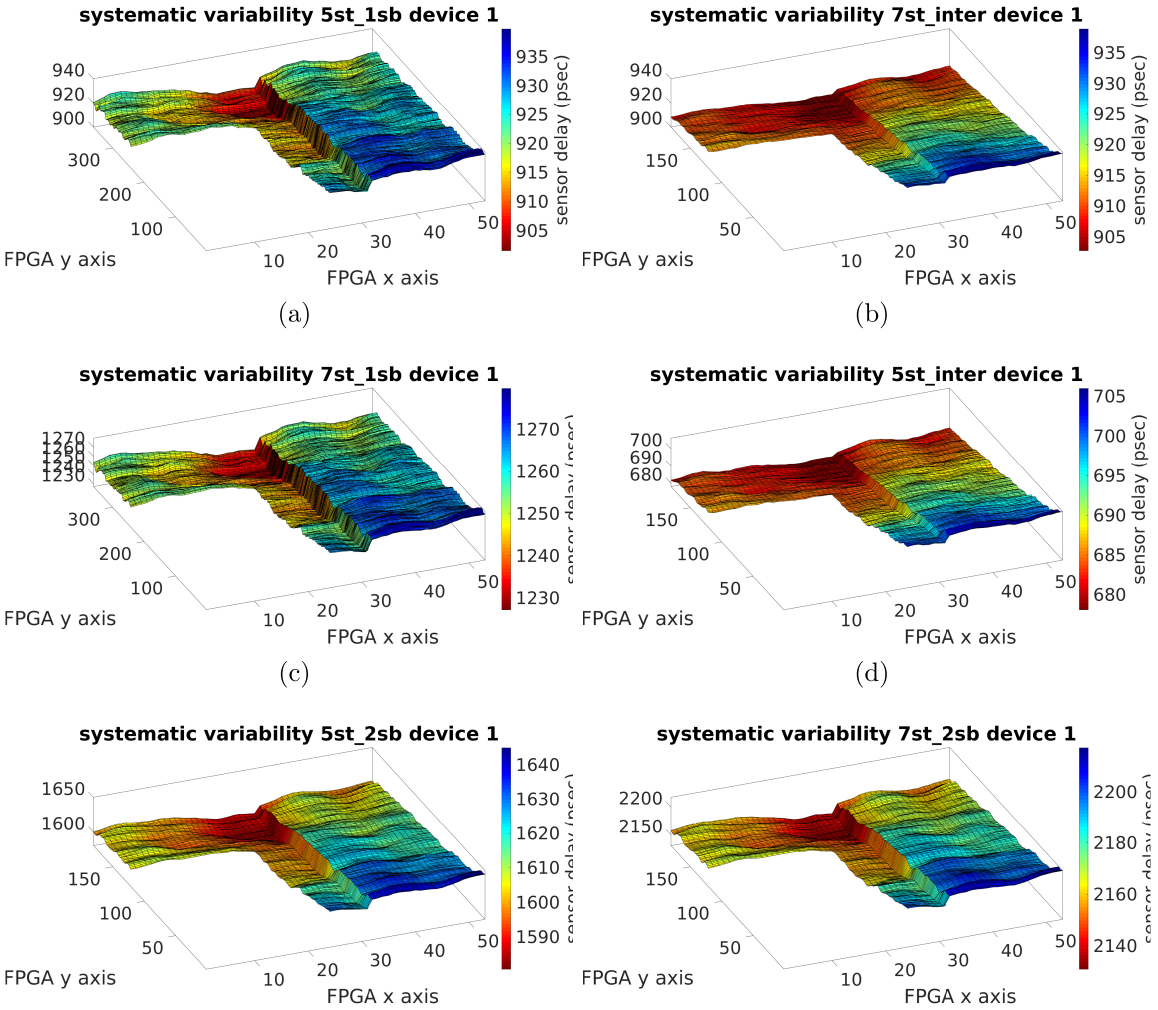
(d)

(e)

Figure 4.5: Systematic and stochastic variability maps of sensor “5st_2sb” using regression (b),(c) and DSMA (d),(e) methods.

for instance, in guiding the placement and routing tool towards the fastest area. Consequently, calculating and acutely understanding the systematic parts across the die is substantial for the design of variation aware methods/tools.

In our work, each sensor is designed to measure different parameters, as we make the assumption that variability maps are not identical among configurations. This has already been shown timidly in the previous section (see Figures 4.2a, 4.3a). In this Subsection, however, we analyze the systematic variability to obtain and measure the differences among the sensors which are utilized to sample the devices.



(a) (b)

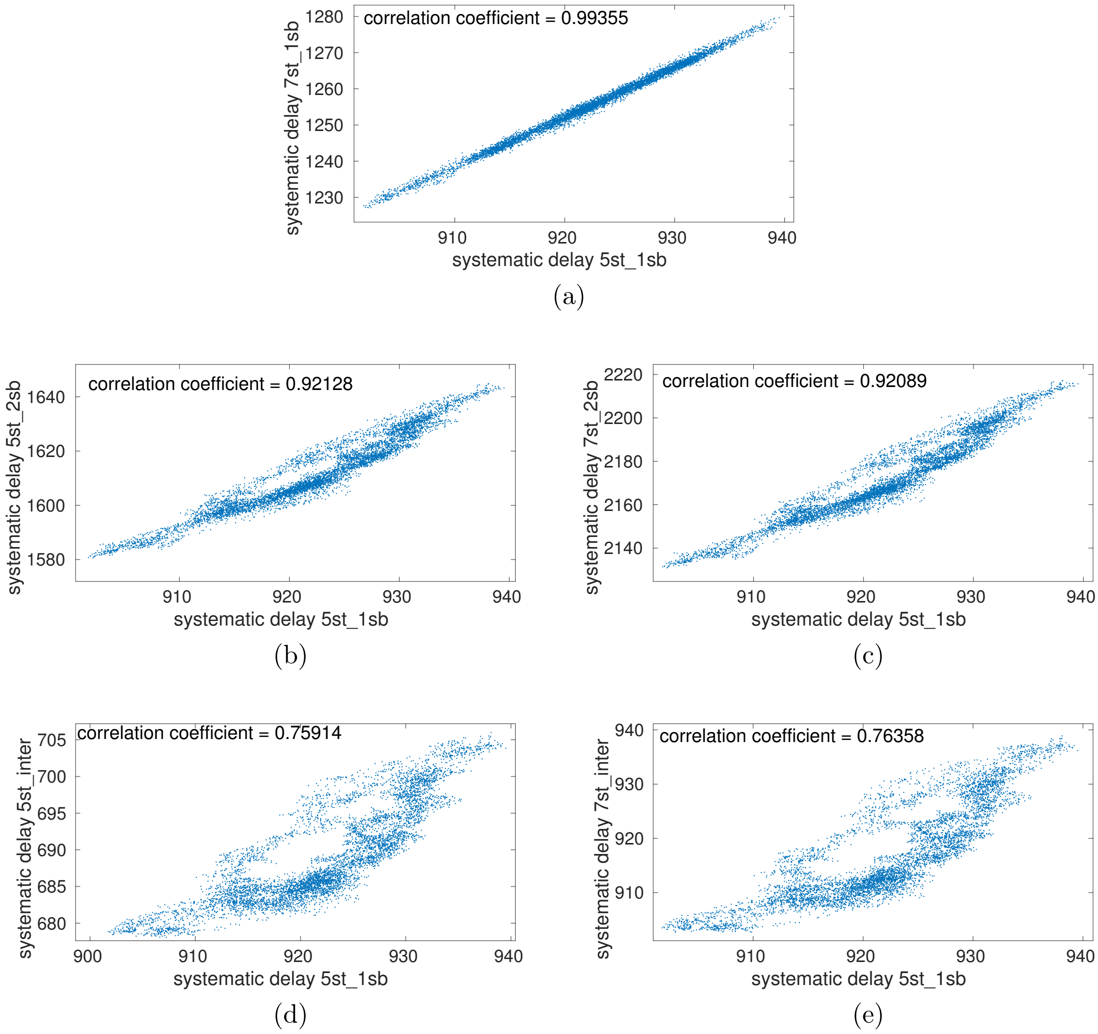
(c) (d)

(e) (f)

Figure 4.6: Systematic variability maps of all sensors for device 1.

Table 4.2 provides results about the systematic variability, afterward the decomposition is computed. We utilize similar metrics as in the total variability (Table 4.1). Furthermore, we provide the portion of the standard deviation of the systematic variability to the total, i.e., $\sigma_{sys}/\sigma_{total}$. Notice that for all devices except device 4, this portion is approximately or greater than 90%, which indicates that systematic variability is the dominating part of variability. For device 4 this portion is smaller; this occurs due to the fact that the total variability of this particular die is not as significant as the other devices (see Table 4.1), thus stochastic variability will have a greater relatively portion. This fact will be better comprehend in the following

Subsection, where stochastic variability will be scrupulously examined. Moreover, as the mean value (μ) is constant, the value σ_{sys}/μ is the same portion of the total related variability (σ_{total}/μ) as the aforementioned portion ($\sigma_{sys}/\sigma_{total}$) and hence it is not presented in the Table. Finally, the metric $range/min$ is smaller in all cases for systematic variability compared to the total, as obviously expected.



(a)

(b)

(c)

(d)

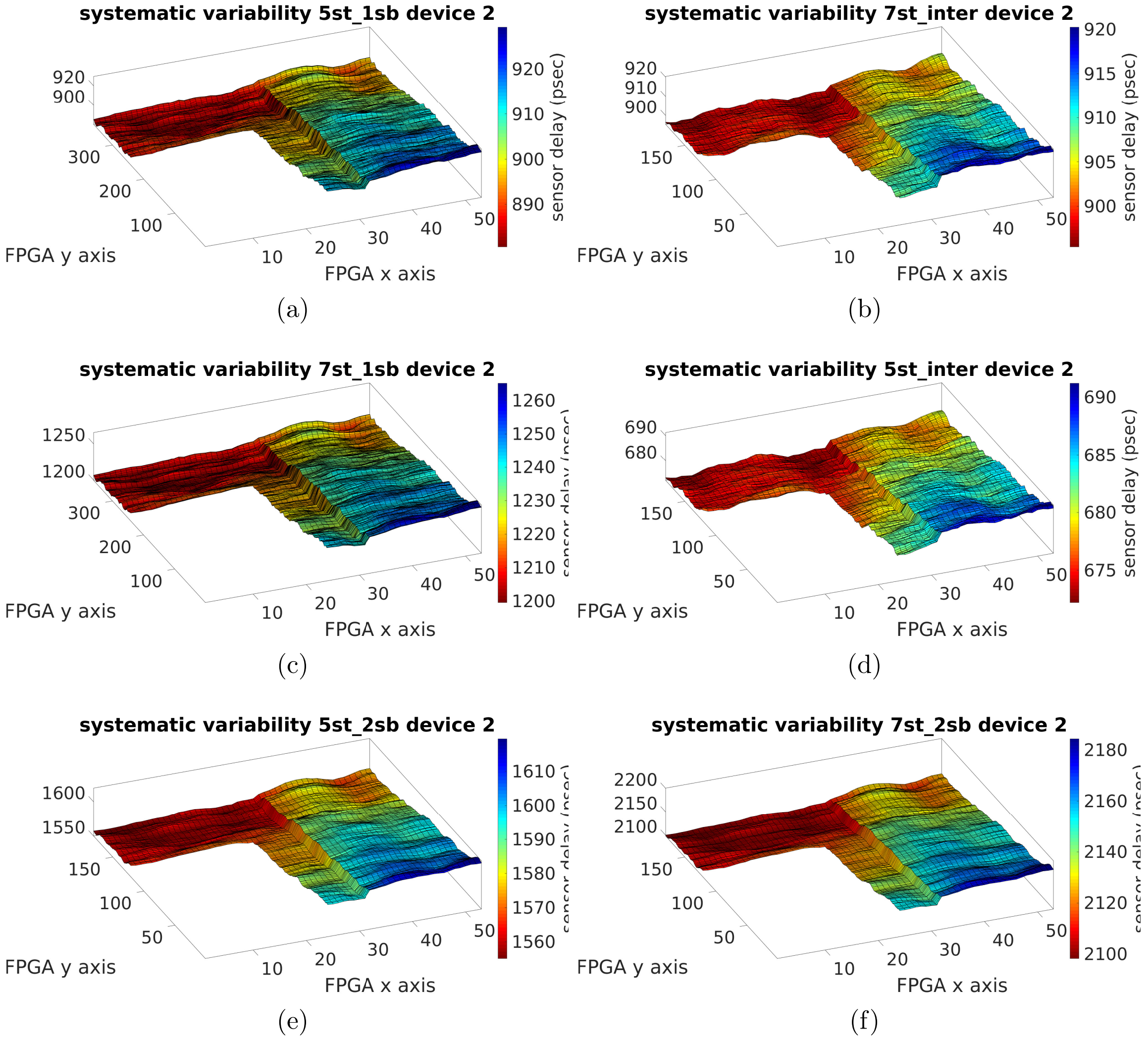
(e)

Figure 4.7: Correlation graphs with reference sensor the “5st_1sb” for device 1.

Additionally to the previous, we present the systematic variability maps for all the sensors of device 1 (Figure 4.6). The first significant observation is that the sensors that have been designed with the same architecture principals point out optically very similar systematic variability maps: “5st_1sb” and “7st_1sb” (Figures 4.6a and 4.6c), “5st_2sb” and “7st_2sb” (Figures 4.6e and 4.6f) and “5st_inter” and “7st_inter” (Figures 4.6d and 4.6b), even though their ranges (maximum and minimum delays) are different. To bolster this observation, we use the Pearson correlation coefficient among sensors[3]. Figure 4.7a depicts the correlation graph between “5st_1sb” and “7st_1sb”, where the linear relationship among them is irrefutable, and thus the experimental correlation coefficient is acutely close to 1 (0.99355 in

[3] The discontinuity in our maps was removed manually in order to avoid the biasing and obtain more accurate/fair correlation coefficients.

fact). These results verify the assumption of the grid model stated in Subsection 3.2.2 of the previous Chapter. The same applies for the other mentioned pair of sensors, and in Figure 4.7 we present the correlation computation with respect to the sensor “5st_1sb” among the others. A simple way to claim that the correlation is very close to 1 for the other pair of sensors is to observe that between Figures 4.7b and 4.7c the correlations are very close to another, i.e., 0.92128 and 0.92089, respectively. This means that, because the reference sensor is common (“5st_1sb”), the two aforementioned maps should be almost identical, i.e., have correlation coefficient very close to 1. The same exists for the interconnect sensors “5st_inter” and “7st_inter” (Figures 4.7d and 4.7e) having correlation coefficients 0.75914 and 0.76358, respectively. Deductively, the way we managed to implement the grid model is sound for all sensors.



(a)

(b)

(c)

(d)

(e)

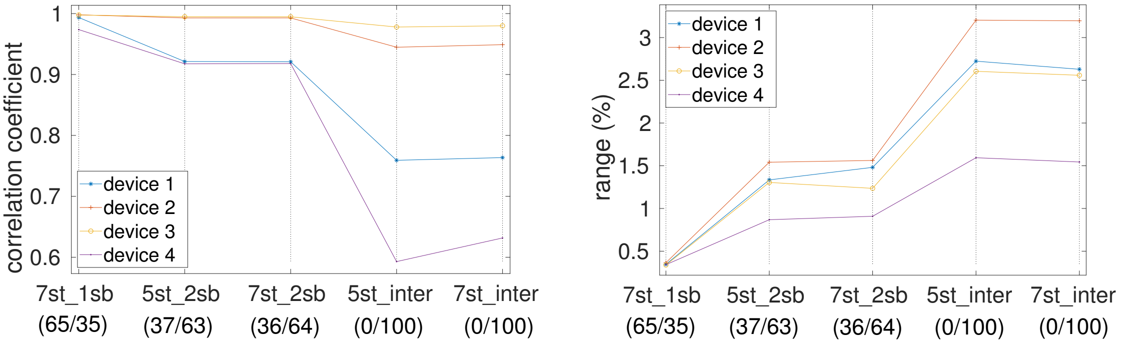
(f)

Figure 4.8: Systematic variability maps of all sensors for device 2.

Most importantly, by optically observing the maps of Figure 4.6, we notice that sensors “5st_1sb” and “7st_inter” having approximately the same mean delay (μ in Table 4.1), and hence connoting fair comparison, do not reveal correlated areas in terms of speed (delay). This is also presented in Figure 4.7e, where the points in the correlation graph (representing systematic delay of the underlying sensors) are abstained from developing an almost linear line, but on the contrary they are spread in a systematic manner, revealing a correlation of approximately 0.76. Furthermore, remember that on the one hand, the first sensor is designed with a major portion of logic delay, while the other with solely interconnect resources. Consequently, there is an irrefutable evidence that the characteristics of the implemented sensor are very important for the morphology of the map, while our results denote that interconnects (mainly wires) follow a different process than logic components (mainly transistors). Overall, these results demonstrate the importance of utilizing multiple sensors implemented with diverse characteristics, in order to accurately analyze the variability and to predict the performance variation in FPGAs. This information could be very useful for the prediction of the delay of realistic designs' paths with different routing and logic resource characteristics, having the prior knowledge of their mapping on the FPGA fabric.

Considering the systematic variability maps and their correlation graphs, we can undoubtedly claim that when considering two sensors, the more similar is the portion of their logic delay to the interconnect delay, the more their variability maps reveal similarities, which means that correlation coefficient is more close to 1. In particular, from Figure 4.7 we can observe that as sensors differ even more from the reference sensor "5st_1sb", they present lower correlation coefficient: sensors "5st_2sb" and "7st_2sb" have greater correlation coefficient and their graphs indicate a more linear relationship contrary to "5st_inter" and "7st_inter". Since this statement is only confirmed for device 1, we have to present the same results for the other devices as well, in order to be able to claim more general assertions about the characteristics of the sensors and the relationship to their systematic variability maps.

Figure 4.8 presents the systematic maps for device 2. In this case we can obtain optically that the underlying device presents more correlated variability maps than device 1 (e.g., Figures 4.8a and 4.8b). In addition, note that the sensors being designed with the same architecture provide similar (optically identical) maps, exactly as occurred in device 1; which is an expected observation as mentioned in the previous paragraph. From the two presented devices, we can educe the significance of the individual study of the variability in each device independently and the identification of the fast and the slow areas of the die with respect to sensors' architecture.

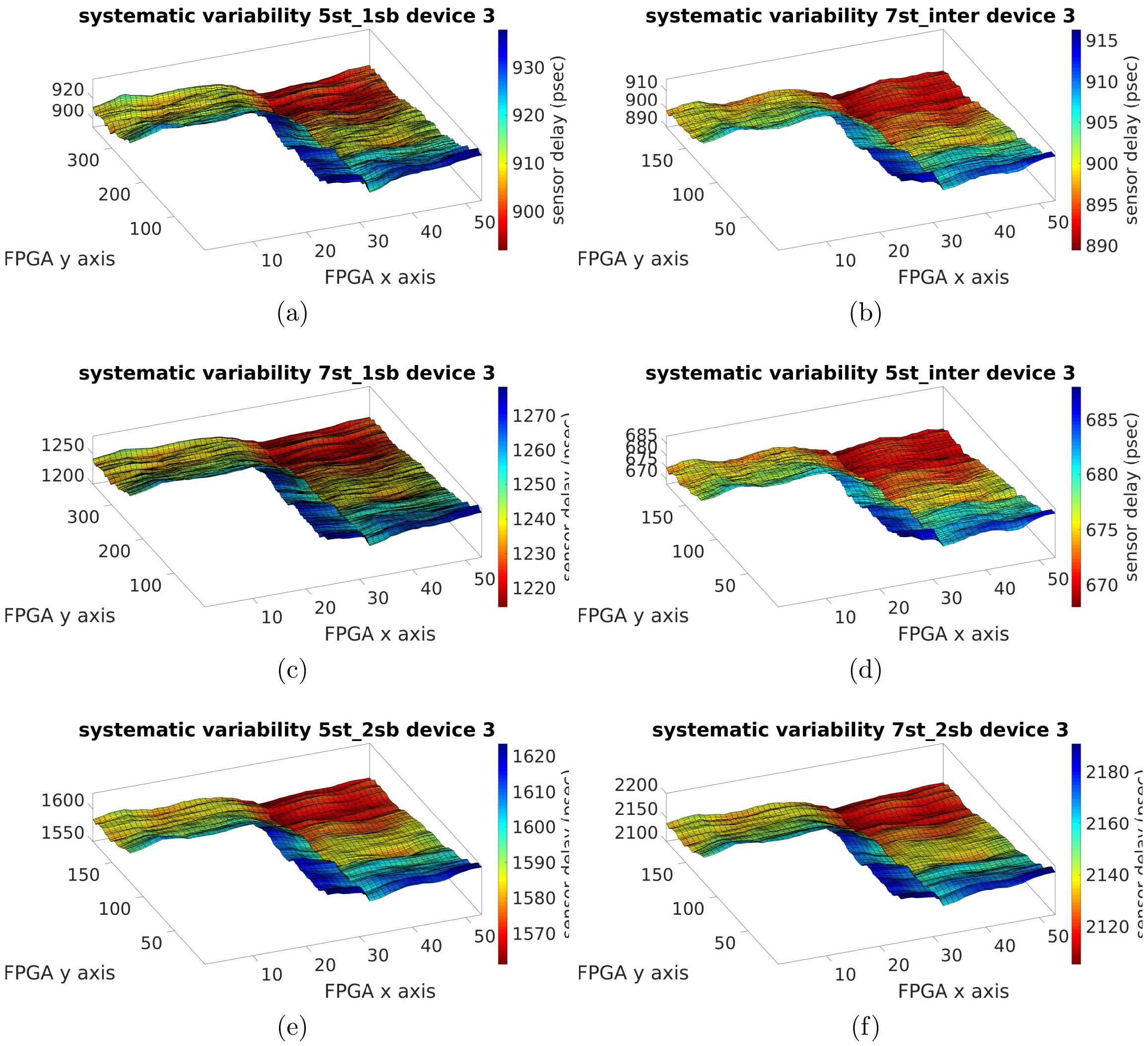


(a) Correlation

(b) Performance estimation error

Figure 4.9: Correlation results and performance estimation error between $5st\_1sb$ (67/33) and rest sensors (logic/interconnect).

In order to quantify these results for all the examined devices, we present how the correlation varies for all devices, taking as reference sensor the "5st_1sb" (Figure 4.9a). We observe that the correlation weakens as the ratio of logic/interconnects[4] sensor delay decreases reaching down to approximately 0.59 for interconnect only sensors (device 4, sensor "5st_inter"). However, note that devices 1 and 4 point out dissimilar (uncorrelated), in terms of speed (delay), areas, while the others do not (correlation coefficient is above 0.9 in all cases, indicating strong correlation among sensors). Moreover, in Figure 4.9b we illustrate the maximum difference (error) in relative performance estimation between the various systematic variability maps. To compute the relative performance estimation, we perform the subtraction of the minimum delay of each corresponding map and then dividing by the same value each coordinate point (x, y). Afterward, the new relative maps are subtracted to compute the maximum relative difference among them. By keeping the "5st_1sb" sensor as reference, we calculate that the average error of the entire map ranges in 0.01-1.22% and extends to 0.34-3.2% considering particular (x, y) points on the



(a)

(b)

(c)

(d)

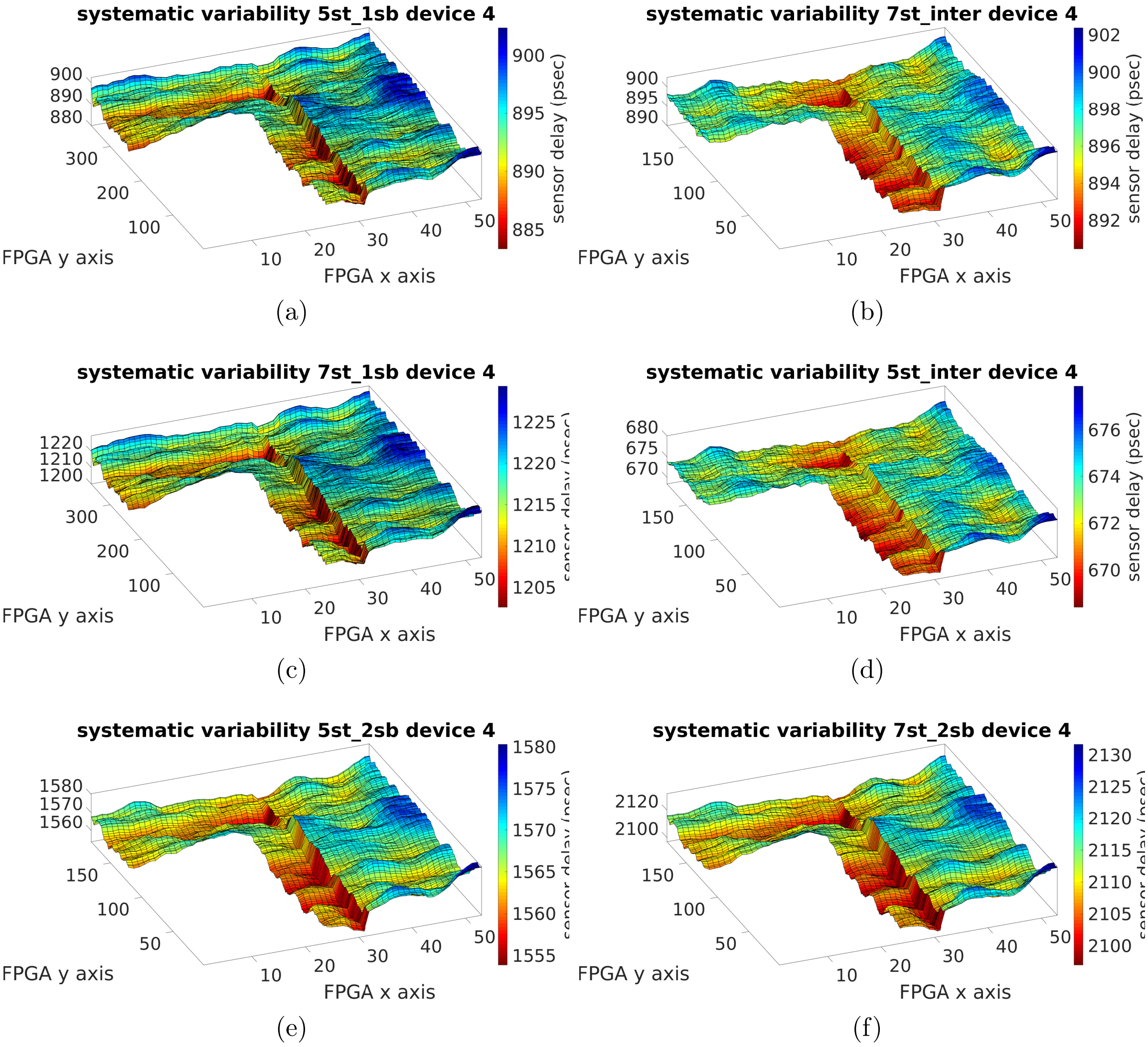
(e)

(f)

Figure 4.10: Systematic variability maps of all sensors for device 3.

[4] The (logic/interconnect) indicates the approximate portion (in percentage), of the delay in logic and interconnect resources, as provided by STA tool (Table 3.1).

maps. Notice that, device 2 reveals the greatest difference while device 4 the lowest. This is reasonable, as device 2 is the most varied device in our RO sensors and simultaneously it provides small variance in interconnect sensors compared to the other devices (σ/μ in Table 4.2). Alternatively, device 4 reveals the lowest variability of all sensors contrary to the other devices, and it has the lowest performance error. One important remark is that the maximum difference is not associated with the correlation coefficient, due to the fact that they provide different evaluation metrics: the correlation coefficient indicates how maps vary in terms of speed, ignoring the ranges, while maximum difference quantifies the difference in terms of performance of the corresponding sensors in each coordinate.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 4.11: Systematic variability maps of all sensors for device 4.

For completeness reasons, we depict the systematic maps for devices 3 and 4 in Figures 4.10 and 4.11 respectively. Obviously, similar assertions to those that have been made for devices 1 and 2, can be claimed as well.

Overall, systematic variability is the major portion of the total variability and its individual study for each device, both qualitatively and quantitatively, becomes substantial for comprehending and sequentially evaluating the performance variation of different paths, which could presumably refer to real-world designs' critical paths.

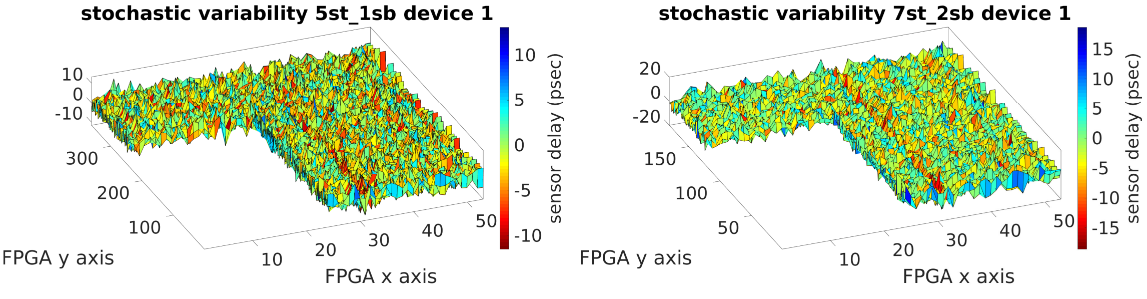
4.2.2 Stochastic Variability

Stochastic variability does not exhibit any spatial correlation and it is assumed that it follows a Gaussian distribution (Subsection 3.2.2). To verify the first statement, we provide the stochastic variability maps of sensors "5st_1sb" and "7st_2sb" of device 1 (Figure 4.12). Obviously, both maps do not exhibit any spatial correlation indicating the random nature of stochastic variability. In the following, to assert the statement of the Gaussian distribution, we plot the probability histogram of the residuals versus the theoretical probability distribution with μ and σ, the mean and standard deviation of the computed residuals respectively. Figure 4.13 illustrates the histogram probability distribution of the aforementioned sensors and the solid line depicts the theoretical probability distribution. Indubitably, the calculated stochastic variability is normally distributed (Gaussian) with high accuracy. In addition, we compute the correlation coefficient among the two depicted stochastic maps. Figure 4.14 shows this case, where obviously there is no correlation between the two maps (correlation coefficient is approximately 0). That said, this is an expected result since stochastic variability is from its nature uncorrelated and random for two distinct delay paths, and hence the two maps should not have any correlation to each other. We should mention that, while only sensors of device 1 are presented, we verified these statements in all measurements that have been provided for all of our devices, and all cases asserted the same statements as above. Hence, we do not present other cases in order to avoid repetitiveness.

Table 4.3: Stochastic variation results for nominal conditions.

Sensor	device 1 (ps)				device 2 (ps)			
	σ_{stoch}	$3\sigma_{stoch}/\mu$	$3\sigma_{stoch}/\mu \cdot \sqrt{N}$	$\sqrt{\sigma_{sys}^2 + \sigma_{stoch}^2}$	σ_{stoch}	$3\sigma_{stoch}/\mu$	$3\sigma_{stoch}/\mu \cdot \sqrt{N}$	$\sqrt{\sigma_{sys}^2 + \sigma_{stoch}^2}$
7st_1sb	3.8	0.90%	0.78%	11	3.5	0.86%	0.76%	15.8
7st_2sb	4.5	0.62%	0.55%	18.8	4.2	0.60%	0.52%	21.7
7st_inter	2.4	0.79%	0.70%	8.9	2.3	0.77%	0.68%	6.4
5st_1sb	3.2	1.04%	0.77%	8.1	2.9	0.98%	0.73%	11.7
5st_2sb	3.7	0.70%	0.52%	14	3.5	0.66%	0.49%	16.1
5st_inter	1.9	0.84%	0.63%	6.8	1.8	0.81%	0.61%	4.8
Sensor	**device 3 (ps)**				**device 4 (ps)**			
	σ_{stoch}	$3\sigma_{stoch}/\mu$	$3\sigma_{stoch}/\mu \cdot \sqrt{N}$	$\sqrt{\sigma_{sys}^2 + \sigma_{stoch}^2}$	σ_{stoch}	$3\sigma_{stoch}/\mu$	$3\sigma_{stoch}/\mu \cdot \sqrt{N}$	$\sqrt{\sigma_{sys}^2 + \sigma_{stoch}^2}$
7st_1sb	3.9	0.93%	0.82%	13.8	3.6	0.88%	0.78%	5.5
7st_2sb	4.7	0.65%	0.58%	19.5	4.3	0.60%	0.53%	7.1
7st_inter	2.7	0.88%	0.78%	6.3	2.3	0.77%	0.68%	3
5st_1sb	3.2	1.06%	0.79%	10.3	3	1.01%	0.75%	4.3
5st_2sb	3.9	0.73%	0.55%	14.5	3.6	0.68%	0.51%	5.4
5st_inter	2.2	0.96%	0.71%	4.8	1.9	0.85%	0.64%	2.4

Table 4.3 provides results that quantify the stochastic variations. Since stochastic variations are normally distributed the most useful quantification metric is $3\sigma_{stoch}/\mu$ because three times the standard deviation bilateral from the mean value in a Gaussian distribution, contains approximately all the possible values (with 99.7% probability). Important notes can be extracted from this Table. In particular, the standard deviations for "5st_1sb" and "7st_1sb" of device 1, which have the same architecture, are 3.2 ps and 3.8 ps respectively, and thus we can draw the conclusion that as the stages of an RO increase, stochastic standard deviation increases as well. This is reasonable, because as discussed in Subsection 3.2.2, the standard deviation



(a)

(b)

Figure 4.12: Stochastic variability maps of sensors "5st_1b" and "7st_2sb" of device 1.

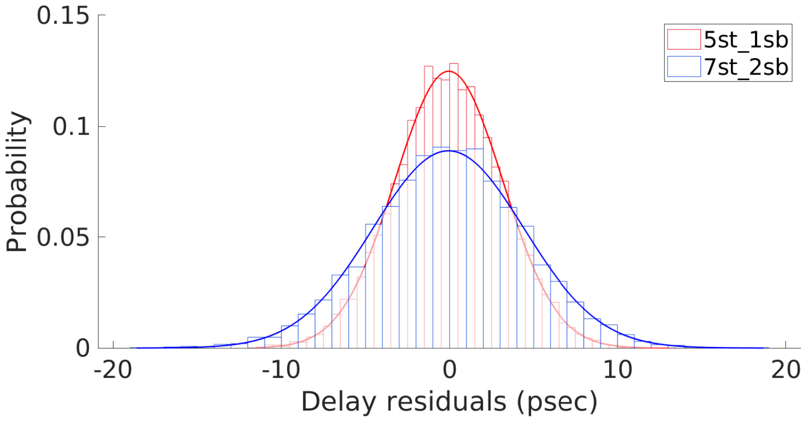


Figure 4.13: Probability distribution of stochastic variation (device 1).

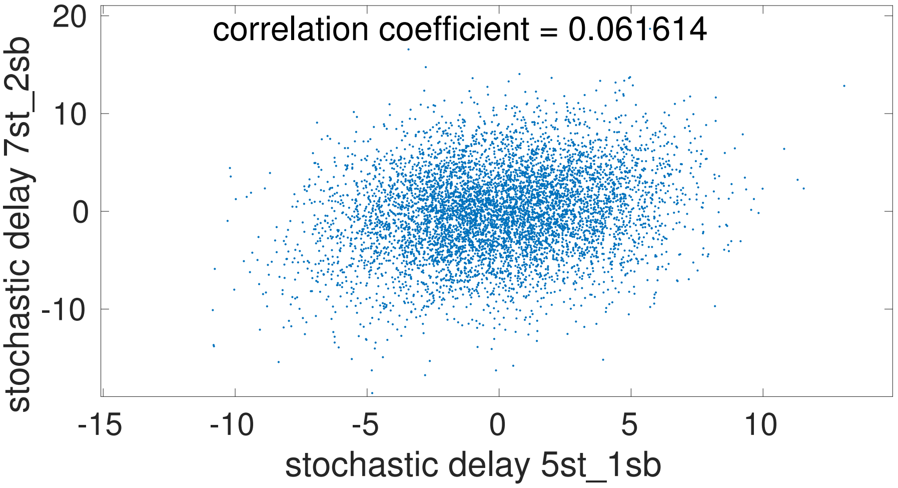


Figure 4.14: Correlation graph of the two stochastic maps (Figure 4.12).

is increased by the square root of the summation of squared standard deviations. Consequently, as the footprint increases we anticipate higher standard deviation. However, the $3\sigma_{stoch}/\mu$ is 1.04% and 0.90% respectively, which means that the relative standard deviation, as percentage of the mean, is reduced. The justification for this observation is that stochastic variations become smoother as we increase the resources that compose the sensor and hence the path delay becomes higher. The following formula proves the aforementioned statement [13,44,62]:

$$\frac{\sigma_{stoch}}{\mu} \propto \frac{1}{\sqrt{N}} \tag{4.1}$$

where in our case N is the number of RO stages. From the above formula we deduce that, as the delay of a path increases the magnitude of the uncorrelated stochastic variation is attenuated due to averaging over multiple gate and interconnect delays [23]. This has been delineated in Table 4.3, where we have multiplied the relative to the mean standard deviation with the square root of N, e.g., for the above-mentioned sensors the values are very close to each other: 0.77% and 0.78% respectively. By observing the values of the Table, we can elicit that our results verifies the equation 4.1 in all cases, having a negligent experimental error. Finally, by observing the results of the Table, we can educe that interconnect sensors induce lower stochastic variations than RO sensors, e.g., for sensors "5st_1sb" and "7st_inter" of device 3, which have approximately the same mean delay, $3\sigma_{stoch}/\mu$ is 1.06% and 0.78% respectively. This may not surprise, as interconnect sensors induce delay primarily from metal wires, while the aforementioned RO sensor primarily from transistors, and with respect to the analysis of variability in Chapter 2.7, transistors provide greater intrinsic variations, while most of variations in metal wires result from spatial systematic effects [9].

Unlike systematic, stochastic variability tends to be much more similar. Resulting form the Table 4.3, we can elicit that in all cases the magnitude of stochastic variability is similar in all devices, e.g., for sensor "5st_2sb" the relative standard deviation $3\sigma_{stoch}/\mu$ is 0.70%, 0.66%, 0.73% and 0.68% for devices 1 to 4 respectively. The same applies for all sensors correspondingly. By comparing the systematic with stochastic, the latter indicates that it has a much more predictable and well-known distribution, and its impact is abated as the critical path delay increases. This is essential and leads to the conclusion that systematic is the type of variability that delineates and describes the location of fast and slow areas on the FPGA.

Regarding the two variability types, a sound question that comes into existence is the relationship among their standard deviations. Considering the first order canonical form of a critical path delay (equation 3.2) we can derive that the two independent random variables circumscribing systematic and stochastic variability, are summed. Hence, in accordance with equation A.6 the variances should be summed. Consequently, for standard deviations the formula has to be:

$$\sigma_{total} = \sqrt{\sigma_{sys}^2 + \sigma_{stoch}^2} \tag{4.2}$$

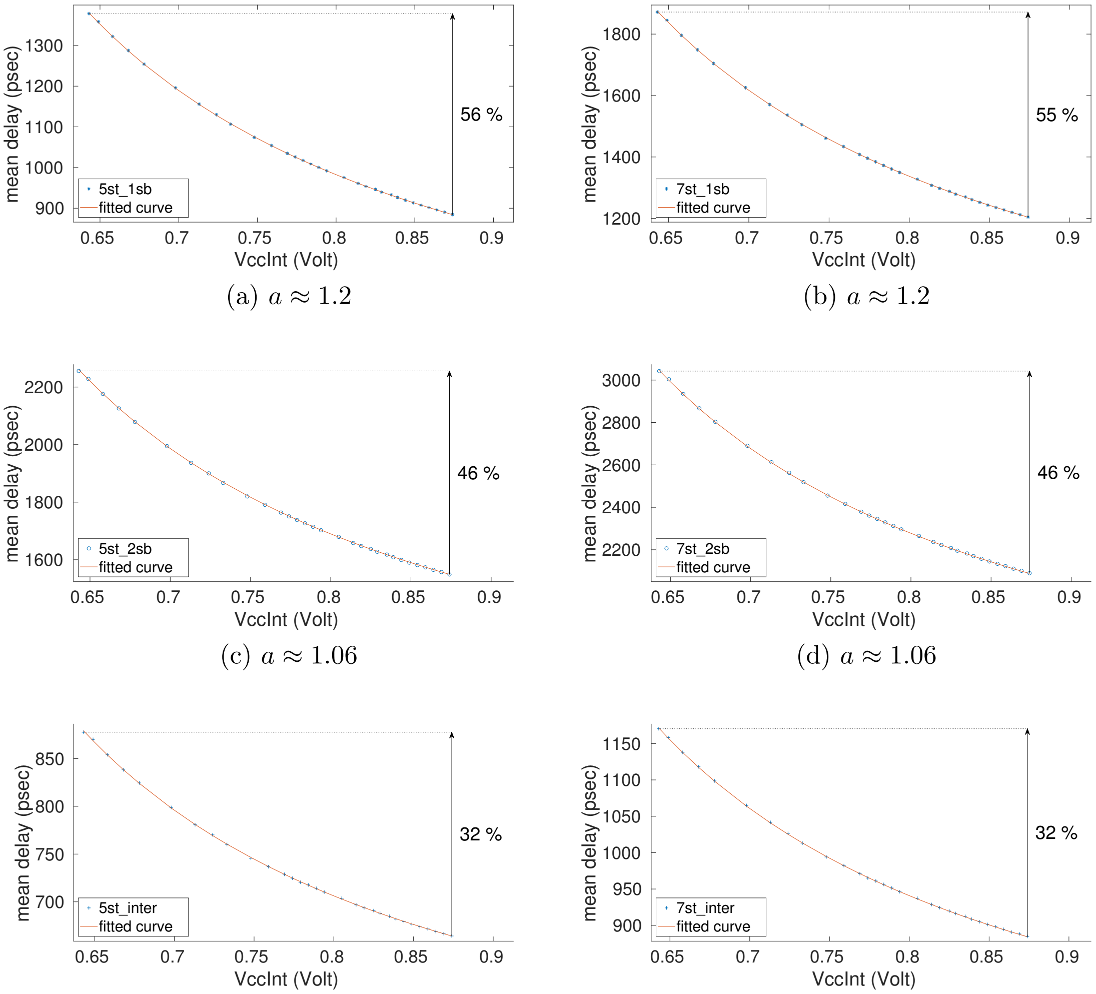
That is examined in Table 4.3, where this square root is calculated. By comparing the total standard deviation from Table 4.1 we assert that equation 4.2, is valid in all cases, considering the negligible experimental error, i.e., 4%.

In Conclusion, the stochastic variability of our FPGA devices is normally distributed in all cases. Our results verify that as the footprint of the sensor increases and the total delay becomes higher, the σ^2_{stoch} increases (Figure 4.13), but stochastic variation as ratio of mean value ($3\sigma_{rand}/\mu$) attenuates due to averaging over multiple gate and interconnect delays. Furthermore, interconnects sensors, which are dominated by wire resources, provide lower stochastic variations than RO sensors which are dominated by transistors.

4.3 Variability Under Voltage and Temperature Alteration

Our analysis is continued by assessing the performance variation under voltage and temperature alterations. First of all, we ascertain the validity of the equation 2.1 (alpha-power law model) by plotting the mean sensor delay versus voltage and by fitting the above equation, using the custom equation fitting model in MATLAB. To achieve that, we provide fine-grain, in terms of the supply voltage, measurements for one device, especially the one labeled as device 3, with the following operating conditions: V_{ccint} in the range of 0.640-0.875 V and T_j constant at 30 °C. Figure 4.15 illustrates the relationship of the measured delay values, where manifestly, we observe the accuracy of the fitting models in all cases. Note that, different values of the fitting parameter a occur for each architecture type sensor: sensors with one SB have approximately $a \approx 1.2$, while those occupying two SBs have $a \approx 1.6$ and the interconnect sensors have $a \approx 0.9$. Unambiguously, as the ratio of the logic to the interconnect delay increases, the parameter a increases as well, and thus the path delay is affected by from supply voltage. This is quantified by the increase (in percent) of the mean delays for the extreme voltage values: 56% for sensors with one SB, 45% with two SBs and 32% for interconnect sensors. However, since the parameter a in the equation 2.1 denotes the mobility degradation of the transistor due to the augmented lateral electric field [2], in our case where we use the fitting equation for the entire path delay (not for each individual transistor) we should clarify that it is not a value utilized to express the mobility saturation. Rather, due to the precise fitting, we exploit this parameter to acquire a mathematical model of the mean delay as a function of the supply voltage. Hence, it is reasonable the value of the parameter a to be lower than 1 (in our case for interconnect sensors), while for transistors the lowest possible value is 1, occurring for a transistor in extreme velocity saturation. That said, this value of a for interconnects is an implication for what is expected: the dominated portion of the delay is induced by metal wires rather than switch transistors. In addition, we noticed that in all sensors cases the delay decreases almost linearly ($a \approx 1$) for higher V_{ccint} values in the range 0.81-0.875V, while for the lower voltages the value of the fitting parameter is as discussed above (Figure 4.15).

Afterward, we exhibit the performance in a more coarse-grain aspect than the previous analysis, with respect to the supply voltage in the range of 0.640-0.875 V and the junction temperature at four constant values: 30, 45, 65 and 85 °C. In particular, Figure 4.16 illustrates the mean delay of the proposed sensors (device 1) versus voltage in coarse-grain (six values of voltages are utilized in this analysis) for the aforementioned temperature values. As expected below, a certain V_{ccint} value, the *temperature inversion* phenomenon occurs (Subsection 2.4); delay decreases with elevated T_j. The temperature inversion point as well as the value of fitting parameter



(a) $a \approx 1.2$

(b) $a \approx 1.2$

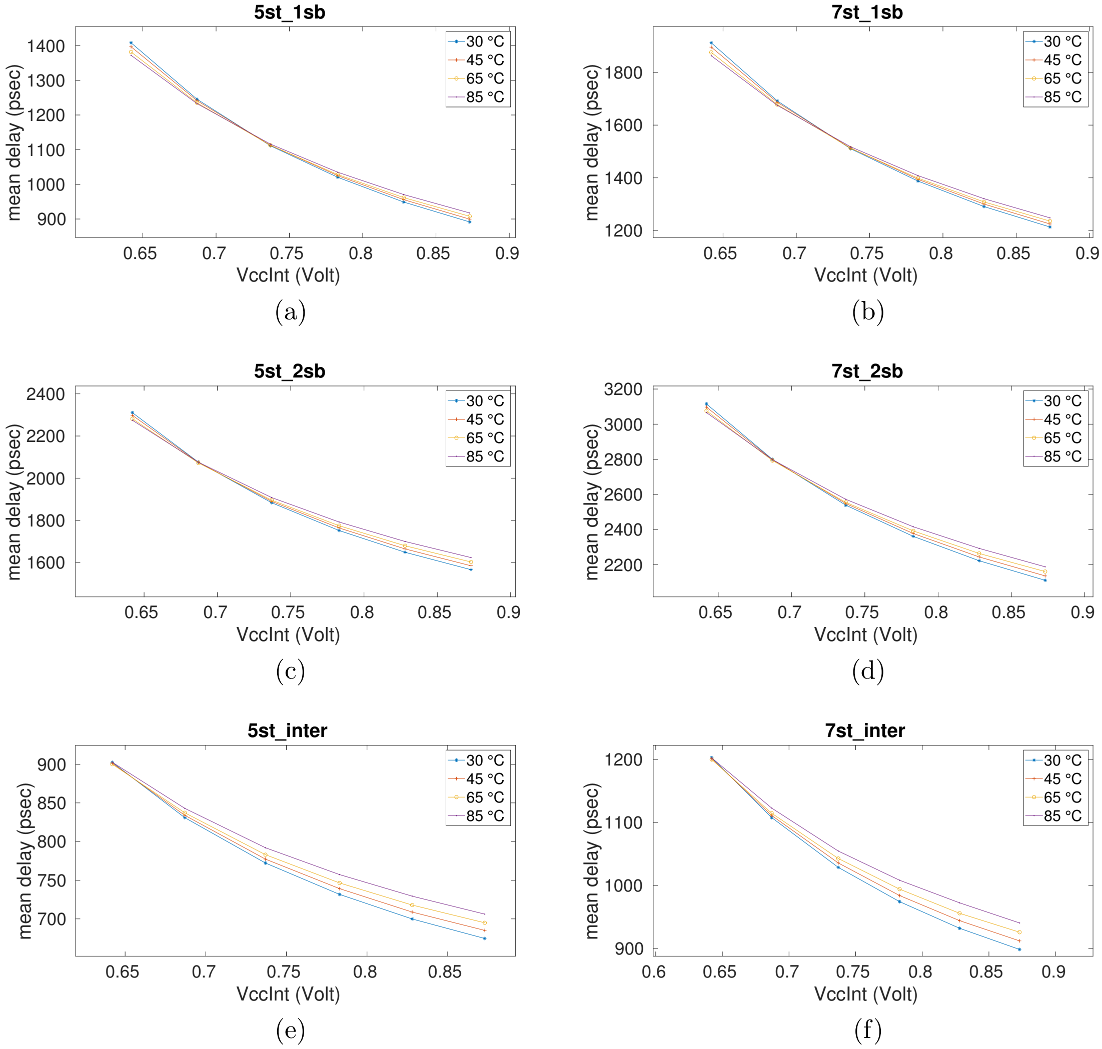
(c) $a \approx 1.06$

(d) $a \approx 1.06$

(e) $a \approx 0.9$

(f) $a \approx 0.9$

Figure 4.15: Mean path delays of the proposed sensors versus supply voltage and the fitted curves utilizing the equation 2.1 (device 3, $T_j = 30$ °).



(a)

(b)

(c)

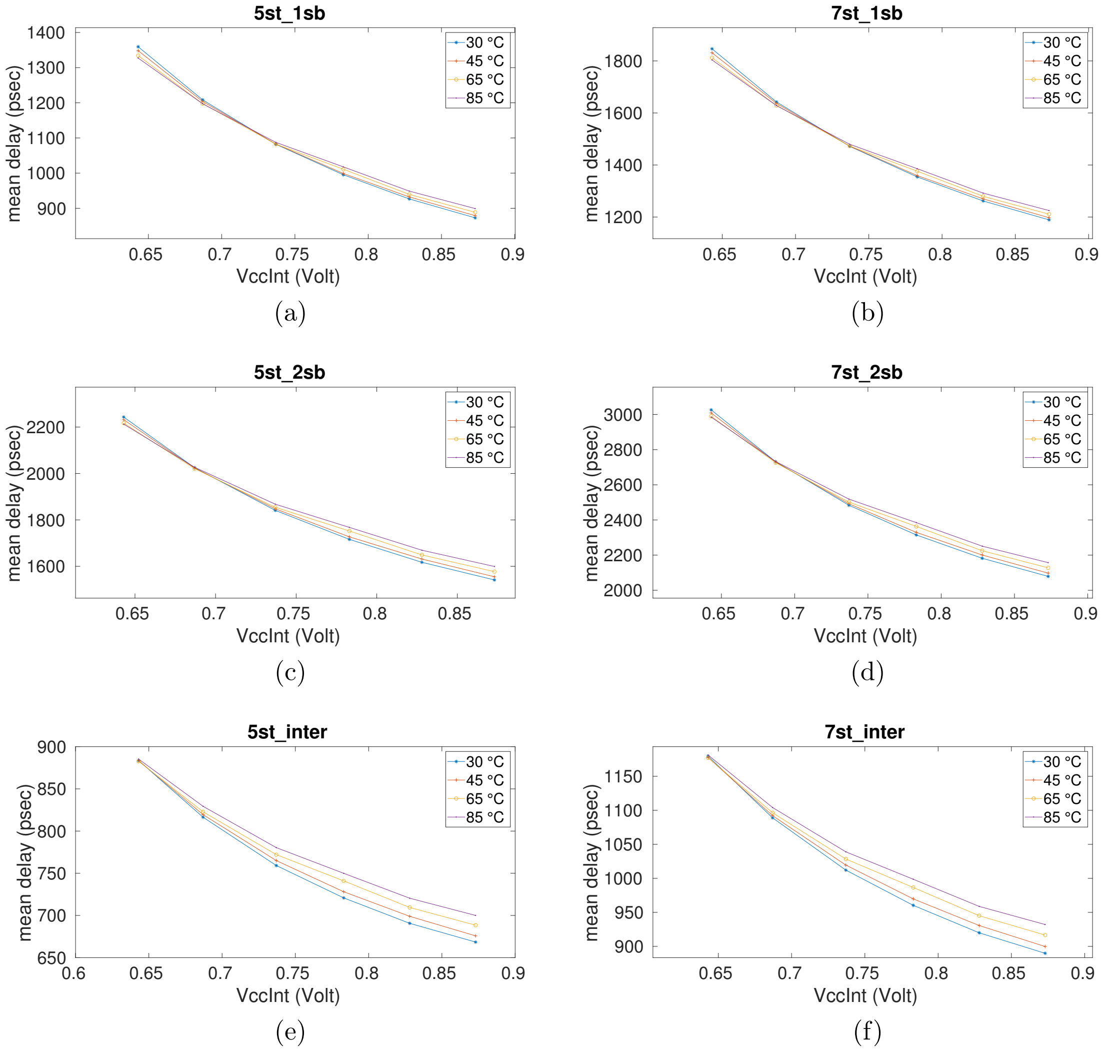
(d)

(e)

(f)

Figure 4.16: Mean path delays of the proposed sensors versus supply voltage for four distinct junction temperature values (device 1).

a vary depending on the sensor configuration. In sensors with higher portion of logic delay (greater amount of transistors), we measure higher performance degradation and temperature inversion manifests in higher V_{ccint} values: approximately, the temperature inversion spot is at 0.72V for “5st_1sb” and “7st_1sb”, 0.69V for “5st_2sb” and “7st_2sb”, and bellow 0.65 for “5st_inter” and “7st_inter”. In contrast to the RO sensors, the interconnect sensors show lower performance degradation with voltage decrease and higher degradation with temperature increase; resistance of wires increases almost linearly, and such does the path delay [2].



(a)

(b)

(c)

(d)

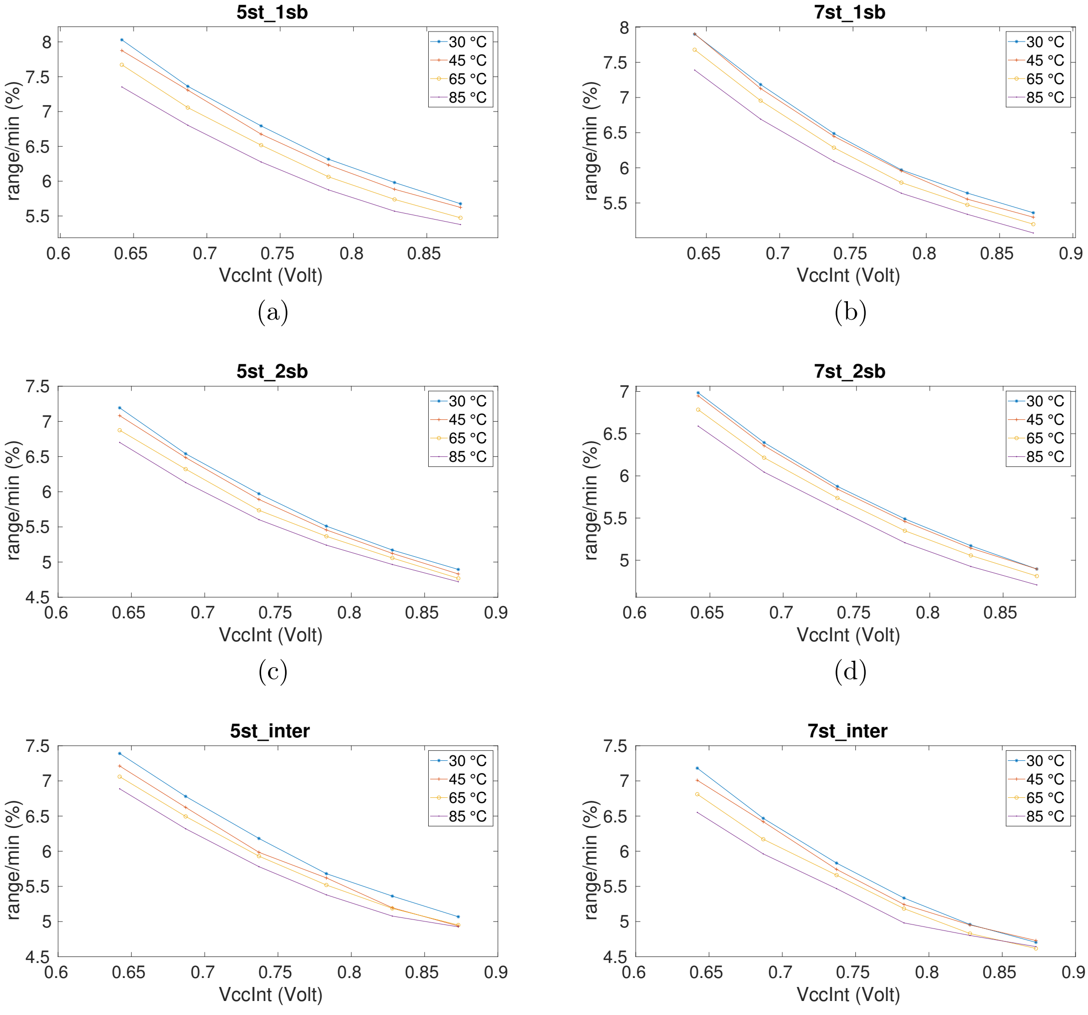
(e)

(f)

Figure 4.17: Mean path delays of the proposed sensors versus supply voltage for four distinct junction temperature values (device 2).

For completeness purposes we present the mean delay versus voltage and temperature for device 2 as well (Figure 4.17). We observe that while the actual sensors’ delays of the two devices are different, the temperature inversion spots occur at the same value of the supply voltage. Considering all devices and sensors, the performance degradation due to voltage ranges up to 33.9% ($7st\_inter$) - 57.9% ($7st\_1sb$), while the degradation due to temperature up to 2.9% ($5st\_1sb$) - 4.8% ($5st\_inter$).

The next step of our assessment is to figure out the variation alteration due to temperature and voltage. For this reason, Figure 4.18 illustrates the total variability for the reported voltage-temperature conditions. We observe that variability increases with the decrease of V_{ccint}. This was expected because, according to equation 2.1, as V_{ccint} scales down, the delay of slower transistors (higher V_{th}) increases relatively higher than faster transistors (lower V_{th}) thus, leading to higher variability. On the other hand, variability decreases with the elevation of T_j: the V_{th} decreases almost linearly to T_j increase [2], i.e., the V_{ccint}–V_{th} increases more for slow transistors and decreases their delay more (see equation 2.1) than the delay of fast transistors [65]. Considering all devices and operating conditions, intra-die and inter-die total variability is increased up to 7.4% ($7st\_inter$) - 9.9% ($5st\_1sb$) and 9.5% ($7st\_inter$) - 12% ($5st\_1sb$).



(a)

(b)

(c)

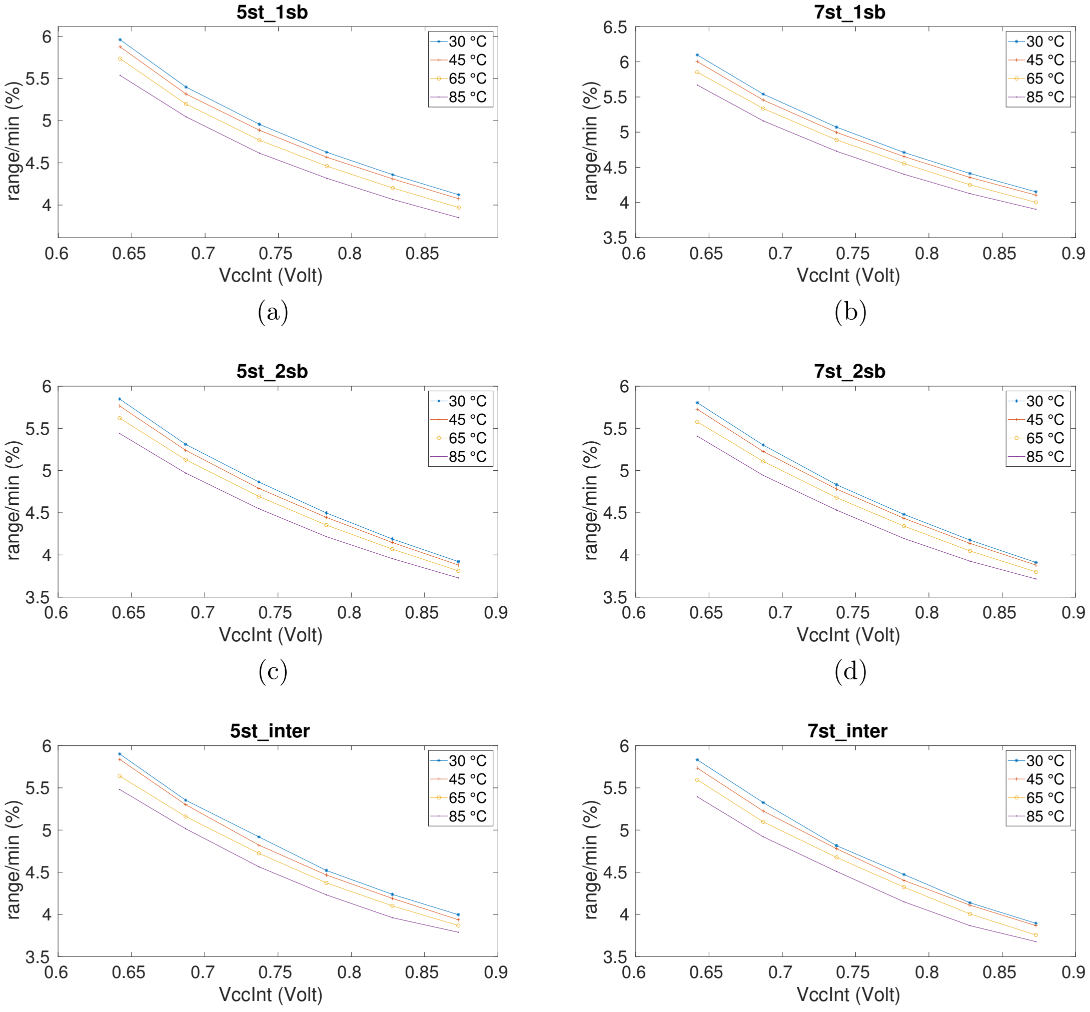
(d)

(e)

(f)

Figure 4.18: Total variability ($range/min$) as a function of voltage and temperature (device 1).

On account of decoupling the variability we present the systematic and the stochastic variability as a function of voltage and temperature distinctly in Figures 4.19 and 4.20, respectively. Apparently, the same observations exist as well for the decoupled variability. However, we notice that decoupling offers the ability to assess the variability in a more lucid way, because in this way the systematic is isolated and the stochastic is assessed by exploiting the standard deviation, due to

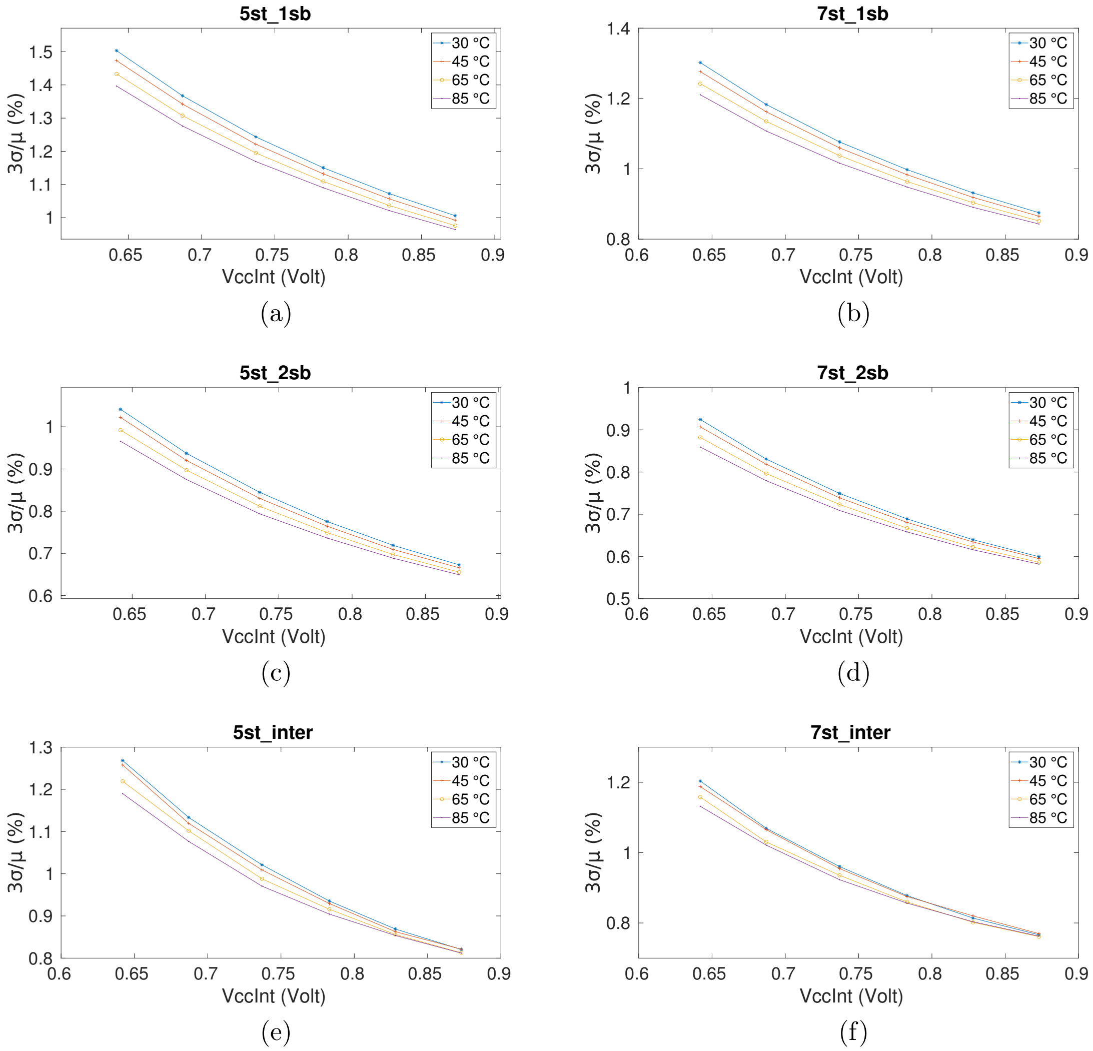


(a) (b)

(c) (d)

(e) (f)

Figure 4.19: Systematic variability ($range/min$) as a function of voltage and temperature (device 1).



(a)

(b)

(c)

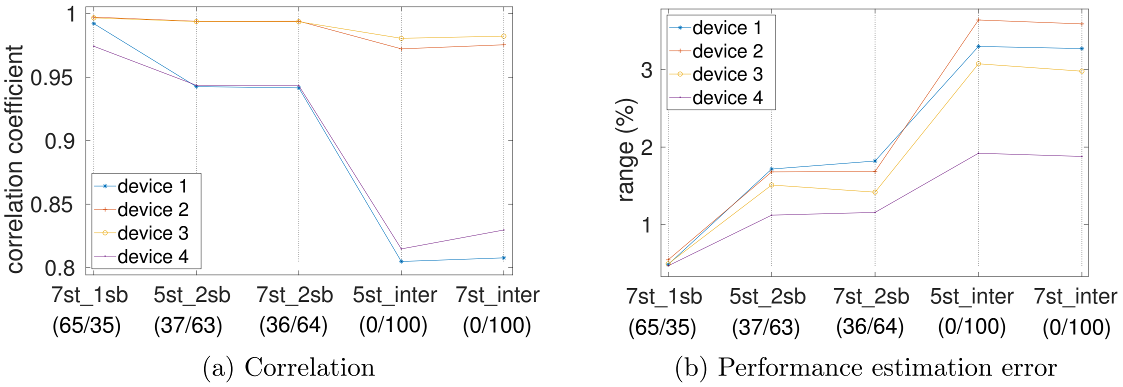
(d)

(e)

(f)

Figure 4.20: Stochastic variability ($3\sigma/\mu$) as a function of voltage and temperature (device 1).

the fact that is normally distributed as verified in the previous section. Thus, the evaluation metric ($range/min$) in this case is not affected by stochastic components, giving the ability to draw precise diagrams and calculations for both systematic and stochastic parts. Considering all devices and operating conditions, the systematic variability is increased up to 5.9% ($7st\_inter$) - 7.3% ($5st\_1sb$) and the stochastic variability is increased up to 1.41% ($7st\_inter$) - 1.53% ($5st\_1sb$).



(a) Correlation

(b) Performance estimation error

Figure 4.21: Correlation results and performance estimation error between “5st_1sb” (67/33) and other sensors for 0.640V, 30℃(logic/interconnect).

Figure 4.21a represents the Pearson coefficients and performance estimation errors (maximum) between the “5st_1sb” reference and the rest sensors for $V_{ccint} = 0.64V$. In contrast to the corresponding plots in Figure 4.9 for nominal conditions ($V_{ccint} = 0.85V$), the correlation between RO sensors remains almost the same, however, the correlation with interconnect sensors has been greatly increased to 0.82 (from 0.59). This is explained by the fact that, with voltage under-scaling, the change in the interconnect delay is attributed mainly to the transistors residing in SBs, hence, the variability maps tend to follow the behavior of transistors. Nevertheless, note that even though the correlation is improved, the error in performance estimation (Fig. 4.21b) is increased to 3.6% (from 3.2%).

Chapter 5

Conclusion

5.1 Concluding Remarks

In this work, we studied the performance variation in 16nm FinFET commercial FPGAs. In order to attain a thorough examination of the variability, we employed multiple types of sensors, which had been designed to assess the logic and interconnect resources of the FPGA fabric. Our methodology relies on the well-established ring oscillator approach. However, we provide a new technique to assess precisely the variability in routing interconnect resources by completely isolating them, without the necessity of deploying other sensors. To do so, we have meticulously designed the ring oscillators in a manner that allows the subtraction of their delays afterward, to obtain the intended results for the interconnects.

In addition, we decoupled the variability into its systematic and stochastic parts and we utilized and assessed the mathematical modeling furnished by the literature. Our results showed that systematic variability is the major portion of the total variability. Furthermore, an extracted implication is the necessity of the individual study of each device due to the fact that variability affects integrated circuits in a dissimilar manner. On the other hand, the stochastic variability has a more predictable and well-know distribution, as well as its impact is abated as the critical path delay increases. Consequently, the characteristics of the two aforementioned types of the variability lead to the conclusion that systematic delineates and describes the locations on the FPGA, where the fast and slow areas appear. Overall, our study indicates that the comprehensive analysis and modeling of the variability could presumably lead to potential performance improvements.

Afterward, we evaluated the impact of diverse voltage and temperature conditions. We analyzed and explained mathematically the way that mean delay is affected by the environmental conditions (voltage and temperature). Furthermore, besides the exhibition of the way that variability is affected by the environmental conditions, we also provide briefly comprehensive explanations of the reasons that lead to variability alterations.

Our experimental results showed up to 9.9% intra-die and 12% inter-die performance variation under certain operating conditions. Moreover, we deduced that logic and interconnect resources present different variation, with low correlation, and a maximum error of 3.6% in performance estimation. Our results accentuate the importance of a multifaceted assessment of variability in FPGAs and provide insights for the implementation of more sophisticated mitigation methods.

5.2 Future Work

The multifaceted study of variability in FPGA devices that is presented in this work can lead to potential research directions, which are examined briefly in the following.

First of all, an important future extension work is the characterization of variability of the DSP blocks across the FPGA fabric, by exploiting the proposed methodology of the ring oscillator approach. Typically the DSP blocks are located between the logic and routing interconnect resources in the FPGA fabric, and an assumption being made is that they are affected in similar way like the examined resources close to them. However, this assumption is contentious because different results can be expected due to dissimilar masks and layers used for the development of the individual components comprising each chip. In order to be able to determine the effect of variability in DSP blocks their variability examination becomes significant.

Second, this work could contribute to the implementation or improvement of CAD tools that indicate the fast/slow areas of the FPGA and utilize this information to guide the placement and routing process for presumable application performance improvement. Our analysis give insights to assess the variability by considering multiple variability maps, whose number can be limited as provided by our mathematical modeling and explanation. Since the results of this work reveal the importance of the mathematical modeling of variability, both of systematic and stochastic, and its individual analysis assuming feedback from the device itself, any potential design of a CAD infrastructure should consider these aspects.

Finally, a presumable research orientation could be the on-line monitoring of FPGAs by exploiting various sensors to measure alterations in performance on a real-time basis. This technique can contribute to the implementation of a robust closed-loop framework that is able to measure the variability at real-time capturing all potential changes, e.g., in workload, and take indispensable decisions to assure the correct operation of the logic functions. Indubitably, these type of implementations lead to healthier FPGAs because they exploit variations by sensing them at real-time and hence they minimize the safety margins as much as possible.

Appendix A

Variance of the summation of random variables

Considering n random variables, $X_1, X_2, \cdots, X_n$, with mean values $m_1, m_2, \cdots, m_n$, respectively, the variance of the random variable representing their summation can be computed as:

$$
\begin{aligned}
&Var(X_1 + X_2 + \cdots + X_n) = E\Big[\Big(X_1 + X_2 + \cdots + X_n - E(X_1 + X_2 + \cdots + X_n)\Big)^2\Big] \\
&= E\Big[\Big(X_1 + X_2 + \cdots + X_n - E(X_1) - E(X_2) - \cdots - E(X_n)\Big)^2\Big] \\
&= E\Big[\Big((X_1 - m_1) + (X_2 - m_2) + \cdots + (X_n - m_n)\Big)^2\Big] \\
&= E\Big[(X_1 - m_1)^2 + (X_2 - m_2)^2 + \cdots + (X_n - m_n)^2 + 2\sum_{i<j}(X_i - m_i)(X_j - m_j)\Big] \\
&= E[(X_1 - m_1)^2] + E[(X_2 - m_2)^2] + \cdots + E[(X_n - m_n)^2] + 2\sum_{i<j} E[(X_i - m_i)(X_j - m_j)] \\
&= Var(X_1) + Var(X_2) + \cdots + Var(X_n) + 2\sum_{i<j} Cov(X_i, X_j), \quad i, j = 1, 2, \cdots, n
\end{aligned}
\tag{A.1}
$$

where the covariance of the variables X_i, X_j, i.e., $Cov(X_i, X_j)$, is used to measure their joint variability. A useful parameter in probability theory is the correlation coefficient, which expresses the linear interdependence of two random variables and is given by the following formula:

$$\rho(X_i, X_j) = \frac{Cov(X_i, X_j)}{\sqrt{Var(X_i)}\sqrt{Var(X_j)}} \tag{A.2}$$

Correlation coefficient is a dimensionless number between -1 and +1.

Special cases for two random variables

For the special case of two random variables, equation A.1 becomes:

$$Var(X_1 + X_2) = Var(X_1) + Var(X_2) + 2 \cdot Cov(X_1, X_2) \tag{A.3}$$

The equation can be restated in a more essential way, considering the equation A.2:

$$Var(X_1 + X_2) = Var(X_1) + Var(X_2) + 2 \cdot \rho(X_1, Y_2) \cdot \sqrt{Var(X_1) \cdot Var(X_2)} \tag{A.4}$$

Perfectly correlated random variables

Two random variables are considered perfectly correlated when their correlation coefficient is exactly $+1$. Then, equation A.4 becomes:

$$\begin{aligned} Var(X_1 + X_2) &= Var(X_1) + Var(X_2) + 2 \cdot 1 \cdot \sqrt{Var(X_1) \cdot Var(X_2)} \\ &= \left(\sqrt{Var(X_1)} + \sqrt{Var(X_2)}\right)^2 \end{aligned} \tag{A.5}$$

Independent random variables

When two random variables are independent, then their correlation coefficient is 0. Hence, equation A.4 becomes:

$$Var(X_1 + X_2) = Var(X_1) + Var(X_2) \tag{A.6}$$

Furthermore, due to the attribute of the variance, when considering a constant number a, then, for a random variable X applies: $Var(aX) = a^2 Var(X)$, equation A.6 can be transformed for the subtraction as follows:

$$\begin{aligned} Var(X_1 - X_2) &= Var(X_1 + (-X_2)) \\ &= Var(X_1) + Var(-X_2) \\ &= Var(X_1) + (-1)^2 Var(X_2) \\ &= Var(X_1) + Var(X_2) \end{aligned} \tag{A.7}$$

Bibliography

[1] G. Moore, "Cramming more components onto integrated circuits," *Proceedings of the IEEE*, vol. 86, no. 1, pp. 82–85, 1998.

[2] N. Weste and D. Harris, *CMOS VLSI design: a circuits and systems perspective.* Pearson Education, Inc., 2011.

[3] ITRS, "International technology roadmap for semiconductors," http://www.itrs2.net/2011-itrs.html, 2011.

[4] K. J. Kuhn, M. D. Giles, D. Becher, P. Kolar, A. Kornfeld, R. Kotlyar, S. T. Ma, A. Maheshwari, and S. Mudanai, "Process technology variation," *IEEE Transactions on Electron Devices*, vol. 58, no. 8, pp. 2197–2208, 2011.

[5] K. Bernstein, D. J. Frank, A. E. Gattiker, W. Haensch, B. L. Ji, S. R. Nassif, E. J. Nowak, D. J. Pearson, and N. J. Rohrer, "High-performance CMOS variability in the 65-nm regime and beyond," *IBM Journal of Research and Development*, vol. 50, no. 4.5, pp. 433–449, 2006.

[6] S. K. Saha, "Modeling process variability in scaled CMOS technology," *IEEE Design & Test of Computers*, vol. 27, no. 2, pp. 8–16, 2010.

[7] A. Kahng and Y. Pati, "Subwavelength lithography and its potential impact on design and eda," in *Proceedings 1999 Design Automation Conference.* IEEE, 1999.

[8] L.-T. Pang, K. Qian, C. J. Spanos, and B. Nikolic, "Measurement and analysis of variability in 45 nm strained-Si CMOS technology," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 8, pp. 2233–2243, 2009.

[9] V. Mehrotra, S. L. Sam, D. Boning, A. Chandrakasan, R. Vallishayee, and S. Nassif, "A methodology for modeling the effects of systematic within-die interconnect and device variation on circuit performance," in *Design Automation Conference (DAC).* IEEE, 2000, pp. 172–175.

[10] D. Frank, Y. Taur, M. Ieong, and H.-S. Wong, "Monte carlo modeling of threshold variation due to dopant fluctuations," in *Symposium on VLSI Circuits. Digest of Papers.* IEEE, 1999, pp. 169–170.

[11] E. Baravelli, A. Dixit, R. Rooyackers, M. Jurczak, N. Speciale, and K. De Meyer, "Impact of line-edge roughness on FinFET matching performance," *IEEE Transactions on Electron Devices*, vol. 54, no. 9, pp. 2466–2474, 2007.

[12] K. Bowman, X. Tang, J. Eble, and J. Meindl, “Impact of extrinsic and intrinsic parameter variations on cmos system on a chip performance,” in *Twelfth Annual IEEE International ASIC/SOC Conference.* IEEE, 1999, pp. 267–271.

[13] M. Wirnshofe, *Variation-aware adaptive voltage scaling for digital CMOS circuits.* Springer, 2013.

[14] R. Kumar and V. Kursun, “Reversed temperature-dependent propagation delay characteristics in nanometer CMOS circuits,” *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 53, no. 10, pp. 1078–1082, 2006.

[15] J. Srinivasan, S. Adve, P. Bose, and J. Rivers, “Lifetime reliability: toward an architectural solution,” *IEEE Micro*, vol. 25, no. 3, pp. 70–80, 2005.

[16] T. Grasser, B. Kaczer, W. Goes, H. Reisinger, T. Aichinger, P. Hehenberger, P.-J. Wagner, F. Schanovsky, J. Franco, P. Roussel, and M. Nelhiebel, “Recent advances in understanding the bias temperature instability,” in *International Electron Devices Meeting.* IEEE, 2010, pp. 82–85.

[17] A. Bravaix, V. Huard, D. Goguenheim, and E. Vincent, “Hot-carrier to cold-carrier device lifetime modeling with temperature for low power 40nm si-bulk nmos and pmos fets,” in *International Electron Devices Meeting.* IEEE, 2011, pp. 622–625.

[18] F. Monsieur, E. Vincent, D. Roy, S. Bruyre, G. Pananakakis, and G. Ghibaudo, “Time to breakdown and voltage to breakdown modeling for ultra-thin oxides (tox<32/spl aring/),” in *IEEE International Integrated Reliability Workshop. Final Report.* IEEE, 2001, pp. 20–25.

[19] C. Christiansen, B. Li, J. Gill, R. Filippi, and M. Angyal, “Via-depletion electromigration in copper interconnects,” *IEEE Transactions on Device and Materials Reliability*, vol. 6, no. 2, pp. 163–168, 2006.

[20] A. Wang and S. Naffziger, *Adaptive Techniques for Dynamic Processor Optimization.* Springer US, 2008.

[21] P. Sedcole and P. Y. Cheung, “Within-die delay variability in 90nm FPGAs and beyond,” in *International Conference on Field Programmable Technology (FPT).* IEEE, 2006, pp. 97–104.

[22] K. M. Zick and J. P. Hayes, “On-line sensing for healthier FPGA systems,” in *International Symposium on Field Programmable Gate Arrays (FPGA).* ACM, 2010, pp. 239–248.

[23] T. Tuan, A. Lesea, C. Kingsley, and S. Trimberger, “Analysis of within-die process variation in 65nm FPGAs,” in *International Symposium on Quality Electronic Design (ISQED).* IEEE, 2011, pp. 1–5.

[24] J. S. Wong, P. Sedcole, and P. Y. Cheung, “Self-measurement of combinatorial circuit delays in FPGAs,” *ACM Transactions on Reconfigurable Technology and Systems*, vol. 2, no. 2, pp. 10:1–10:22, 2009.

[25] U. Farooq, Z. Marrakchi, and H. Mehrez, *Tree-based Heterogeneous FPGA Architectures.* Springer-Verlag New York, 2012.

[26] L. Cheng, J. Xiong, L. He, and M. Hutton, "FPGA performance optimization via chipwise placement considering process variations," in *International Conference on Field Programmable Logic and Applications (FPL).* IEEE, 2006, pp. 1–6.

[27] K. Maragos, G. Lentaris, and Soudris, "In-the-field mitigation of process variability for improved FPGA performance," *IEEE transactions on Computers*, pp. 1–15, 2019.

[28] T.-Y. Chiang, B. Shieh, and K. Saraswat, "Impact of Joule heating on scaling of deep sub-micron Cu/low-k interconnects," in *Symposium on VLSI Technology. Digest of Technical Papers.* IEEE, 2002, pp. 38–39.

[29] C. Visweswariah, K. Ravindran, K. Kalafala, S. Walker, S. Narayan, D. Beece, P. Piaget, I. Zamek, J. Fan, J. Beetner, N. Venkateswaran, and J. Hemmett, "First-order incremental block-based statistical timing analysis," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 25, no. 10, pp. 2170–2180, 2006.

[30] H. Chang and S. Sapatnekar, "Statistical timing analysis considering spatial correlations using a single pert-like traversal," in *International Conference on Computer Aided Design.* IEEE, 2003, pp. 621–625.

[31] K. Qian, "Variability modeling and statistical parameter extraction for CMOS devices," Ph.D. dissertation, EECS Department, University of California, Berkeley, Jun 2015. [Online]. Available: http://www2.eecs.berkeley.edu/Pubs/TechRpts/2015/EECS-2015-165.html

[32] K. Kuhn, C. Kenyon, A. Kornfeld, M. Liu, A. Maheshwari, W. kai Shih, S. Sivakumar, G. Taylor, P. VanDerVoorn, and K. Zawadzki, "Managing process variation in intel's 45nm cmos technology," *Intel Technology Journal*, vol. 12, no. 2, pp. 93–109, 2008.

[33] N. G. Orji, T. V. Vorburger, J. Fu, R. G. Dixson, C. V. Nguyen, and J. Raja, "Line edge roughness metrology using atomic force microscopes," *Measurement Science and Technology*, vol. 16, no. 11, pp. 2147–2154, 2005.

[34] A. Asenov, S. Kaya, and A. R. Brown, "Intrinsic parameter fluctuations in decananometer mosfets introduced by gate line edge roughness," *IEEE Transactions on Electron Devices*, vol. 50, no. 5, pp. 1254–1260, 2003.

[35] M. Abu-Rahma and M. Anis, *Nanometer Variation-Tolerant SRAM.* Springer-Verlag New York, 2013.

[36] D. A. Steele, A. Coniglio, C. Tang, B. Singh, S. Nip, and C. J. Spanos, "Characterizing post-exposure bake processing for transient- and steady-state conditions, in the context of critical dimension control," in *Proc. SPIE*, vol. 4689, 2002, pp. 517–530.

[37] P. Drennan, M. L. Kniffin, and D. R. Locascio, "Implications of proximity effects for analog design," in *IEEE Custom Integrated Circuits Conference.* IEEE, 2006.

[38] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De, "Parameter variations and impact on circuits and microarchitecture," in *Design Automation Conference.* IEEE, 2003, pp. 338–342.

[39] I. Kuon, R. Tessier, and J. Rose, *FPGA Architecture: Survey and Challenges.* Now Foundations and Trends, 2008.

[40] K. Ma, L. Wang, X. Zhou, S. X.-D. Tan, and J. Tong, "General switch box modeling and optimization for fpga routing architectures," in *International Conference on Field-Programmable Technology.* IEEE, 2010, pp. 320–323.

[41] Xilinx, "Ultrascale+ fpga product tables and product selection guide," https://www.xilinx.com/support/documentation/selection-guides/zynq-ultrascale-plus-product-selection-guide.pdf, 2018.

[42] I. Ahsan, N. Zamdmer, O. Glushchenkov, R. Logan, E. Nowak, H. Kimura, J. Zimmerman, G. Berg, J. Herman, E. Maciejewski, A. Chan, A. Azuma, S. Deshpande, B. Dirahoui, G. Freeman, A. Gabor, M. Gribelyuk, S. Huang, M. Kumar, K. Miyamoto, and D. Mocuta, "Rta-driven intra-die variations in stage delay, and parametric sensitivities for 65nm technology," in *Symposium on VLSI Technology, 2006. Digest of Technical Papers.* IEEE, 2006, pp. 170–171.

[43] S. Onaissi and F. N. Najm, "A linear-time approach for static timing analysis covering all process corners," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 27, no. 7, pp. 1291–1304, 2008.

[44] D. Blaauw, K. Chopra, A. Srivastava, and L. Scheffer, "Statistical timing analysis: From basic principles to state of the art," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 27, no. 4, pp. 589–607, 2008.

[45] A. Agarwal, K. Chopra, D. Blaauw, and V. Zolotov, "Circuit optimization using statistical static timing analysis," in *Proceedings of the 42nd annual Design Automation Conference.* ACM, 2005, pp. 321–324.

[46] J. L. Nunez-Yanez, M. Hosseinabady, and A. Beldachi, "Energy optimization in commercial FPGAs with voltage, frequency and logic scaling," *IEEE Transactions on Computers*, vol. 65, no. 5, pp. 1484–1493, 2016.

[47] K. Maragos, G. Lentaris, I. Stratakos, and D. Soudris, "A framework exploiting process variability to improve energy efficiency in FPGA applications," in *Great Lakes Symposium on VLSI (GLSVLSI).* ACM, 2018, pp. 87–92.

[48] Z. Guan, J. S. Wong, S. Chaudhuri, G. Constantinides, and P. Y. Cheung, "A two-stage variation-aware placement method for FPGAs exploiting variation maps classification," in *International Conference on Field Programmable Logic and Applications (FPL).* IEEE, 2012, pp. 519–522.

[49] Z. Guan, J. S. Wong, S. Chaudhuri, G. Constantinides, and P. Cheung, "Exploiting stochastic delay variability on FPGAs with adaptive partial rerouting," in *International Conference on Field-Programmable Technology (FPT)*. IEEE, 2013, pp. 254–261.

[50] Y. Li, C.-H. Hwang, and M.-H. Han, "Simulation of characteristic variation in 16 nm gate FinFET devices due to intrinsic parameter fluctuations," *Nanotechnology*, vol. 21, no. 9, pp. 1–7, 2010.

[51] R. Chau, S. Datta, M. Doczy, B. Doyle, J. Kavalieros, and M. Metz, "High-k/metal-gate stack and its mosfet characteristics," *IEEE Electron Device Letters*, vol. 25, no. 6, pp. 408–410, 2004.

[52] T. Sakurai and A. Newton, "Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas," *IEEE Journal of Solid-State Circuits*, vol. 25, no. 2, pp. 584–594, 1990.

[53] J. Hicks, D. Bergstrom, M. Hattendorf, J. Jopling, J. Maiz, S. Pae, C. Prasad, and J. Wiedemer, "45nm transistor reliability," *Intel Technology Journal*, vol. 12, no. 2, pp. 131–144, 2008.

[54] V. Betz, J. Rose, and A. Marquardt, *Architecture and CAD for Deep-Submicron FPGAs*. Boston, MA: Springer, 1999.

[55] C. Zhou, R. Cheung, and Y.-L. Wu, "What if merging connection and switch boxes - an experimental revisit on fpga architectures," in *International Conference on Communications, Circuits and Systems*. IEEE, 2004.

[56] Y. Pino, V. Jyothi, and M. French, "Intra-die process variation aware anomaly detection in FPGAs," in *International Test Conference (ITC)*. IEEE, 2014, pp. 1–6.

[57] H. Yu, Q. Xu, and P. H. Leong, "Fine-grained characterization of process variation in FPGAs," in *International Conference on Field-Programmable Technology (FPT)*. IEEE, 2010, pp. 138–145.

[58] P. Sedcole, J. S. Wong, and P. Y. Cheung, "Characterisation of FPGA clock variability," in *Computer Society Annual Symposium on VLSI (ISVLSI)*. IEEE, 2008, pp. 322–328.

[59] J. Li and J. Lach, "Negative-skewed shadow registers for at-speed delay variation characterization," in *International Conference on Computer Design (ICCD)*. IEEE, 2007, pp. 354–359.

[60] M. Majzoobi, E. Dyer, A. Elnably, and F. Koushanfar, "Rapid FPGA delay characterization using clock synthesis and sparse sampling," in *International Test Conference (ITC)*. IEEE, 2010, pp. 1–10.

[61] K. Maragos, G. Lentaris, D. Soudris, K. Siozios, and V. F. Pavlidis, "Application performance improvement by exploiting process variability on FPGA devices," in *Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2017, pp. 452–457.

[62] K. M. Zick and J. P. Hayes, "Low-cost sensing with ring oscillator arrays for healthier reconfigurable systems," *ACM Transactions on Reconfigurable Technology and Systems*, vol. 5, no. 1, pp. 1:1–1:26, 2012.

[63] Xilinx, "Ultrascale architecture configurable logic block user guide," https://www.xilinx.com/support/documentation/user_guides/ug574-ultrascale-clb.pdf, 2017.

[64] V. Betz, J. Rose, and A. Marquardt, *Architecture and CAD for deep-submicron FPGAs.* Springer Science & Business Media, 2012, vol. 497.

[65] K. Maragos, E. Taka, G. Lentaris, I. Stratakos, and D. Soudris, "Analysis of performance variation in 16nm FinFET FPGA devices," in *International Conference on Field Programmable Logic and Applications (FPL).* IEEE, 2019.

[66] B. Stine, D. Boning, and J. Chung, "Analysis and decomposition of spatial variation in integrated circuit processes and devices," *IEEE Transactions on Semiconductor Manufacturing*, vol. 10, no. 1, pp. 24–41, 1997.