Study of the impact of variations of fabrication process on digital circuits

Tarun Chawla

Thèse

Présentée pour obtenir le grade de Docteur du
Télécom ParisTech

Spécialité: **Électronique et Communications**

Tarun CHAWLA

Titre: **Etude de l'impact des variations du procédé de fabrication sur les circuits numériques**

Soutenue le 30 Septembre 2010 devant le jury composé de:

Prof.	Lirida	NAVINER	Président de Jury
Dr.	Marc	BELLEVILLE	Rapporteurs
Dr.	Nadine	AZEMARD	Rapporteurs
Prof.	Amara	AMARA	Directeur de thèse
Prof.	Andrei	VLADIMIRESCU	Co-directeur de thèse
M.	Sebastien	MARCHAL	Tuteur industriel

Abstract

Designing digital circuits for sub-100nm bulk CMOS technology faces many challenges in terms of Process, Voltage, and Temperature variations. The focus has been on inter-die variations that form the bulk of process variations. Much work has been done to study their effects and to make circuits more robust by improvements in technology or design. In this work, we have focused on two particular kinds of variations- **Inter-die NMOS to PMOS mismatch** and **Intra-die local random mismatch**. Neither had a noticeable effect in industrial designs and has become a cause of worry only recently. The source of these variations lies in the basic process and is random in nature. Thus, their effect cannot be ameliorated without overhauling the complete process. The work in academia has mostly focused on process changes or architectural improvements. Our work is geared towards design improvements at gate and path level.

We looked at the basic phenomena behind these variations and using simulations observed how they affect the different parameters in a digital design. The focus was on synchronous systems, i.e. clock distribution system that is highly impacted by these variations. We proposed some design methods and optimization strategies to make the circuits more robust. Most of these methods are exploitable within existing design flows that minimizes the cost and allows for quick adoption in the industry. We included the effect of voltage and temperature changes on these two variations to put together a comprehensive understanding. We also proposed methods to verify the basis of our work by comparing against silicon test results. The results of this work have helped to shape the policy of how to handle local mismatch in industrial designs.

Acknowledgement

I would like extend my sincere gratitude to my thesis advisors Dr. Amara AMARA and Dr. Andrei VLADIMIRESCU, for their continuous guidance during this research. I am also greatly indebted to Sebastien MARCHAL, my industrial advisor, whose guidance and support made this thesis possible.

I wish to thank all my colleagues who helped to solve my queries and problems. I am thankful to all my friends in France who made my stay here a very pleasant one. I am especially grateful to my colleague and friend Nirmal PREGASSAME who has translated many a things for me as well as helped to improve my French.

Any endeavor in my life is incomplete without mentioning my family, especially my mother, who has taken great pains to help me become what I am today.

Finally, I would like to thank STMicroelectronics, Crolles that provided me an opportunity to pursue my dream and enabled me to work along and learn from some of the best people in the field.

Table of Contents

Résumé (en Français)

L’industrie microélectronique travaille actuellement sur la technologie 45 nm. Cette technologie est caractérisée par une taille de gravure plus petite que la résolution théorique de l'équipement lithographique. Il est de ce fait prévisible que les marges de variations absolues sur les paramètres caractéristiques du transistor ne vont pas s'améliorer de façon significative par rapport à technologies précédentes. Par conséquent, le transistor va subir une variation, par rapport à sa taille, plus importante que dans les technologies précédentes. La tendance des nœuds technologiques à venir n'est pas n’iront pas en s’améliorant. Les méthodes traditionnelles de mise en œuvre de la conception de circuits numériques utilisés dans l'industrie sont directement impliquées par ces variations. Pour des plus grands circuits, cela entraine une consommation d'énergie plus élevé ou alors une baisse de performance qui n'est pas souhaitable pour le marché semiconducteur. Il est donc impératif de trouver des techniques innovantes de conception de circuits intégrés pour réduire l'effet de ces variations.

Un exemple de un circuit synchronisé est monté dans la Figure I. La synchronisation de cet circuit dépend sur plusieurs paramètres comme délais d’insertion de horloge, délais de donne, setup time, hold time, skew entré deux chemin d’horloge, etc. Les paramètres sont montrés dans la Figure II. La synchronisation pour les circuits digitaux est affectée par de nombreux types de variations, comme le procédé de fabrication, la tension d'alimentation, la température, le vieillissement, ou l’exactitude des outils CAD, etc. Toutefois, la partie principale vient essentiellement de la varaition de ce que nous appelons PVT (procédé de fabrication, tension d'alimentation et la température). Celle-ci tente de paramétrer les effets des fluctuations de procédé de fabrication ainsi que celles provenant de sources externes comme la température ambiante ou la tension. Les variations PVT marquent la différence entre les circuits conçus et ceux qui sont fabriqués : Cette différence peuvent au meilleur cas, réduire l'efficacité d'un produit ou même au pire cas, le rendre complètement inutilisable. En général, les outils d'analyse temporelle calculent l'impact de ces différents types de variations et permettent d’établir

les cas idéaux et les cas pessimistes. En jugeant ces cas, on peut vérifier si le dessin se situe dans des limites acceptables.

Dans le cadre de cette étude, nous avons examiné l'effet des variations présentées ci-dessus sur des circuits numériques, notamment pour les réseaux d'horloge, de manière à minimiser les marges d'erreur et de réduire les configurations sensibles. Parmi les différents blocks d'un modèle synchrone, les réseaux d'horloge sont plus sensibles aux variations de mismatch en raison de leur nature différentielle. La présence de ce mismatch (local et global) dans un réseau d'horloge peut affecter tous les registres et donc limiter les performances réalisables et aussi la complexité de la conception. Dans ce travail, nous avons concentré nos efforts sur les réseaux d'horloge afin de caractériser l'effet des mismatchs pour la technologie CMOS 45nm en envisageant les différents scénarios possibles, comme le changement de la tension, les conditions de corners différents, l'impact sur la longueur de la période et le retard, le compromis entre délai, la taille des cellules et la consommation d'énergie, etc.

Nous avons travaillé principalement sur des variations aléatoires. La philosophie de la conception régulière ont grandement réduit l'impact des variations systématiques et peu d'erreur est possible dans le niveau de conception. Dans les variations aléatoires, nous avons décidé de travailler sur deux types de variations particulières, les variations aléatoires intra-die et les variations aléatoires inter-die déséquilibrée. Ces deux variations sont très importantes en fonction des différences de paramètres relatifs aux périodes d'horloge ou à la longueur de l'arbre d'horloge ou du skew.

Les variations aléatoires locales ou Intra-Die/Within-Die n'ont cessé d'augmenter en se mettant à l'échelle des dimensions du transistor. Jusqu'à présent, ses effets dans la conception pouvaient être négligés en toute sécurité en raison de l'impact global causé en moyenne par les petits effets des variations aléatoires. Toutefois, pour des dessins plus grands et pour des fréquences plus élevées, ces effets se font de plus en plus importants et son impact peut être vérifié.

Comme son nom l'indique, le mismatch crée une différence de propriétés électriques des transistors voisins, grâce à laquelle deux chemins similaires sur une même puce peuvent présenter un retard et des paramètres de puissance différents. Il peut provoquer des skew entre les deux chemins d'horloge qui peut limiter la fréquence et la complexité de la conception. Plus le skew est grand, plus les marges pou une période de l'horloge s'élargissent, et plus la période d'horloge devient importante. Pour une fréquence d'horloge, un skew plus important peut entrainer une limitation à à la profondeur du chemin d'horloge, réduisant alors la taille de la puce ou la complexité de conception. Pour chaque nœud de technologie, la taille relative de la puce et la fréquence d'horloge qui y est associée, sont en augmentation et leur mismatch peut affecter leur croissance.

Un grand nombre de travaux universitaires a été publié sur les origines et le comportement de mismatch, mais l'industrie a vu qu'un effet marginal jusqu'à tout récemment. Or, avec les dimensions du transistor atteignant quelques dizaines des nanomètres, l'effet devient beaucoup plus visible aujourd'hui. La plupart des circuits numériques, sauf les microprocesseurs, ne bénéficient pas de binning qui réduit le risque d'échec de synchronisation. En tant que tel, les dessins et modèles sont validés pour les cas pessimistes des processus ou de limites de coupe. La probabilité d'avoir les cas pessimistes de processus est inférieure à 1%, suite à la distribution gaussienne. En outre, le processus de fabrication est affiné et centré pour chaque produit. De ce fait, arriver à la conception des cas limites pessimistes est très rare. Toutefois, la présence du mismatch peut dégrader le rendement du processus, si elle n'est pas prise en compte dans le temps de conception. Une marge normale pour gérer ce mismatch sans tenir compte de ses caractéristiques peut entrainer plus de temps de conception.

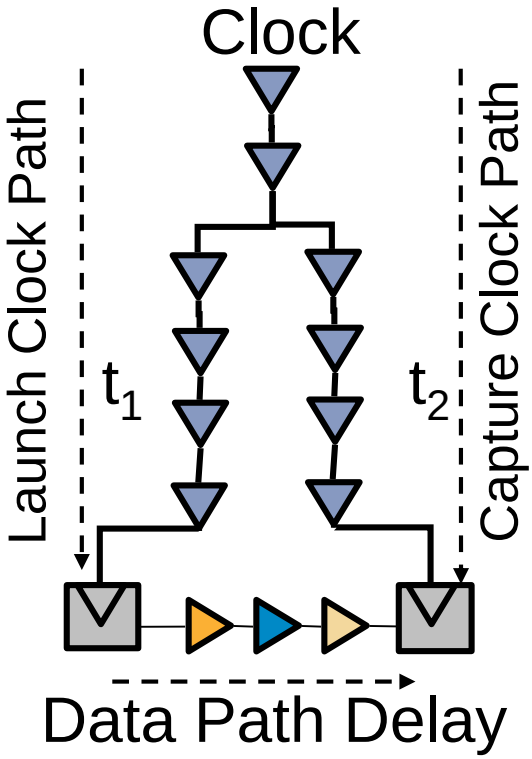

Fig. I: Typiques système synchrone avec les chemins d'horloge et de donne

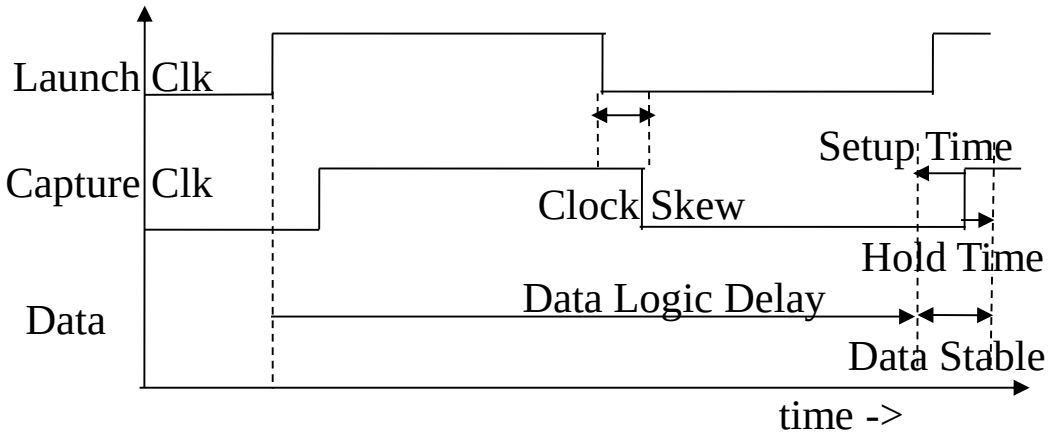

Fig. II: Une chronogramme qui montre da relation entré horloge, donne, setup, hold, et skew

Objectifs

- Identification du processus de variations et de leurs mécanismes

Dans un premier temps, il est nécessaire de comprendre les sources de variations et leurs mécanismes. Il est possible de séparer les sources de variation en deux catégories :

  - Les variations systématiques : variations par rapport aux performances simulées de l'élément dont la source est systématique sur tous les décès en cours de fabrication.
  - Des variations aléatoires : les variations dues aux fluctuations statistiques des performances de l'équipement de fabrication introduisant des variations de performances entre les différentes filières ou du centre d'une plaquette.

Un effort existe déjà pour simuler certaines variations systématiques provenant de la lithographie. Un effort existe aussi dans le domaine de l'analyse statique « timing statistique» qui permet de simuler les performances d'un circuit en tenant compte des variations aléatoires. Pendant cette phase, il s'agit de lister des sources de variations sur des transistors et des interconnexions, et de leurs mécanismes théoriques.

- Évaluation de l'impact des variations sur les performances d'un circuit numérique

Il est nécessaire d'être en mesure d'estimer ou de quantifier les conséquences des variations sur les performances des circuits numériques. Les métriques analysées sont les performances en vitesse, puissance et courant fuites.. La valeur absolue de la variation de la performance n'est pas nécessairement important. L'objectif de ces évaluations est d'être en mesure de quantifier la performance relative d'un circuit par rapport à l'autre pour choisir le meilleur. C'est plus simple que de simuler complètement l'effet d'une variation sur la cellule.

- Évaluation de la méthode et les techniques de conception pour limiter l'impact des variations de processus

Il est nécessaire d'évaluer diverses approches pour obtenir une amélioration quantifiable des performances d'un circuit en utilisant toutes les techniques appropriées pour réduire l'effet des variations sur les performances du circuit.

Les variations de procédé

Les variations global et de l'environnement ou dans les variations à court PVT comprennent le dé-to-die (D2D) les variations de processus, N-au-P mismatch de puce à l'autre, les variations de la température ambiante et des changements dans la tension d'alimentation. D2D variations ont été suffisamment bien expliquée dans de nombreuses publications. Les corners lente (SS) et rapide (FF) défini la limite des variations D2D sur le retardement d'insertion. Toutefois, ces corners traditionnels ne sont pas suffisants en cas de largeur d'période qui est composée de deux bords qui passe par différents transistors. Si la monte est plus rapide que la chute de pointe, l'impact sur la largeur de période est considérable, même si l'impact sur le délai d'insertion est moindre que pour le corner SS. Il ya deux possibilités à envisager N à P globale mismatch : marge supplémentaire (résultats sur les délais d'insertion ou moins réduit la fréquence d'horloge), de corners (dans les résultats des efforts accrus, le temps et l'argent). Il ya une grande corrélation entre les transistors N et P d'une cellule en raison de mesures masque commun. Cependant, l'étape de dopage est différente pour chacun et crée le n-à-p mismatch globale. Comme le dopage a un fort impact sur la tension de seuil et la mobilité, même de faibles variations peuvent entrainer des différences importantes entre les transistors N et P. L'impact du mismatch n-à-p globale est plus importante pour des paramètres comme la largeur d'période. Son impact sur le courant des transistors de type N et type P est montré dans la Figure III ou la mismatch globale est représenté par « Unbalanced Corners »

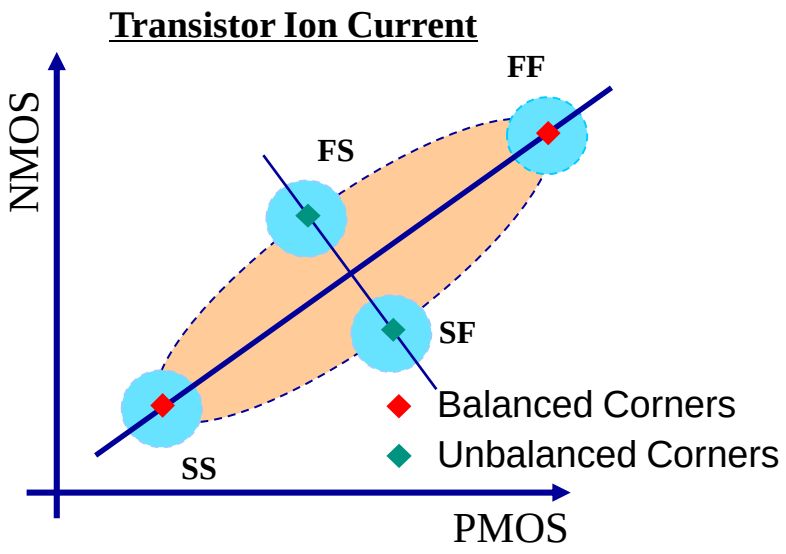

Fig. III: Courant transistor dans la mode saturation pour NMOS et PMOS

Les changements de tension d'alimentation peut être intentionnelle, comme dans le cas de la dynamique de tension et de fréquence mise à l'échelle (DVFS), ou non, comme les variations de régulateur de tension qui peut atteindre jusqu'à 12% autour de la tension d'alimentation nominale selon les spécifications de l'ITRS. Cependant, mise à l'échelle de tension intentionnelle peut être beaucoup plus grande en fonction de l'application et le mode d'alimentation. Les variations de température ambiante pour la plupart des applications industrielles varient de -40 ° C à 125 ° C.

Mismatch ou des variations intrinsèques n'ont pas de corrélation entre les dispositifs et proviennent principalement de la limite naturelle à l'élargissement. Il peut causer des différences dans les caractéristiques électriques de deux dispositifs identiques autrement même géométrie, l'aménagement, et le voisinage. Variations mismatch provient de l'incertitude inhérente liée à des atomes et des résultats dans les variations statistiques dans la structure d'un transistor et d'un cadre. Il existe trois principales sources de déséquilibre- Random Dopant Fluctuations, Line Edge Roughness, et Oxide Thickness Variations, montré dans la Figure IV. L'effet des variations locales et les variations globales peuvent voir dans la Figure V.

- **Random Dopant Fluctuations (RDF)**

RDF est le plus gros contributeur à l'ampleur du mismatch entre 45nm et 65nm transistors. Avec quelques centaines de dopants intérieur de l'appareil, les variations statistiques dans leur nombre et les résultats de localisation dans un potentiel non homogène dans le canal permettant début tournez-le dans les parties et affecter la barrière de fuite induite par abaissement de tension. L'incertitude sur la source et le drain bords des répercussions sur leur résistance et la capacité et consécutivement le transistor actuel. L'impact est principalement dans la région sous le seuil et augmente la variation de la tension de seuil ainsi que provoque un déplacement net de la valeur moyenne du courant de drain à la courbe de tension de grille vers l'axe négatif.

- **Line Edge Roughness (LER)**

LER provient de la rugosité inhérente des portes bords oxyde à l'échelle atomique. Elle influe sur la longueur de grille effective le long de la largeur du canal, ce qui affecte tensions de seuil local à l'intérieur d'un transistor. LER découle de la statistique des variations dans le nombre de photons incidents lors de l'exposition litho, le taux d'absorption, la réactivité chimique, et de résine photosensible composition moléculaire et joue un rôle dominant dans la détermination de la marge du champ électrique et l'accouchement charge l'interface. L'impact de la LER est plus prononcé pour les appareils à proximité de poinçonnement. L'ampleur des variations LER est mineur par rapport à RDF en technologie 65 nm, mais est censé devenir comparables dans les ganglions plus tard.

- **Oxide Thickness Variations (OTV)**

OTV se réfère à la variation moléculaire dans la porte de l'oxyde de surface et d'affecter l'épaisseur porte sur toute la surface. La porte-oxyde épaisseur physique est de l'ordre de l'espacement atomique 5-10 et peut varier de 1-2 espacements atomiques. L'impact de l'OTV est négligeable pour les nœuds en cours, mais sera important lorsque la longueur de grille périphérique devient comparable à la longueur de corrélation des fluctuations. OTV affecte de manière significative l'oxyde tunnel en cours et les causes de variation de la mobilité et le potentiel du canal.

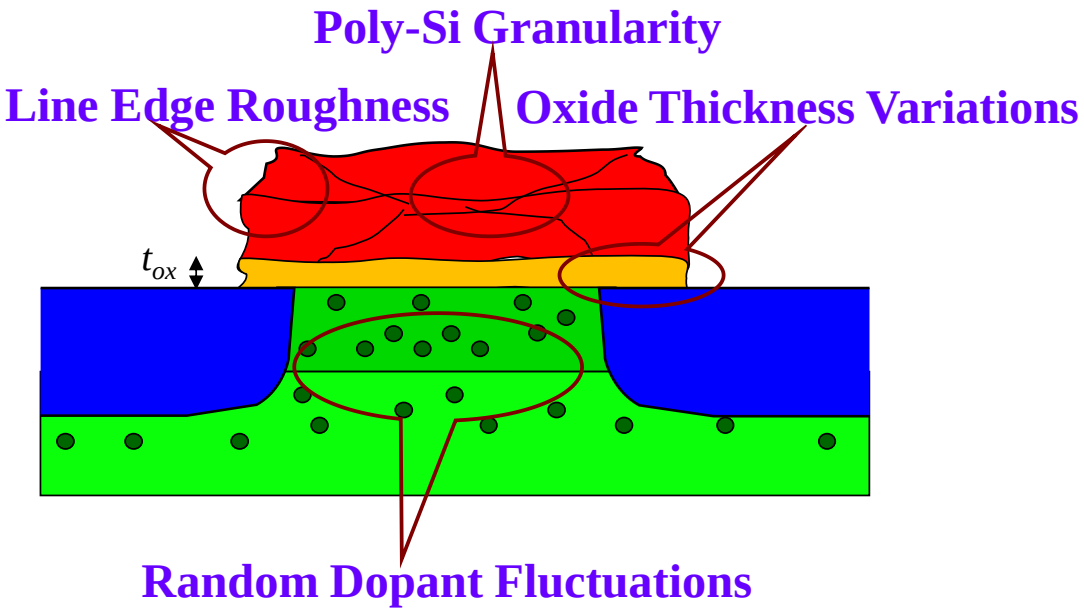

Fig. IV: Composants de mismatch locale

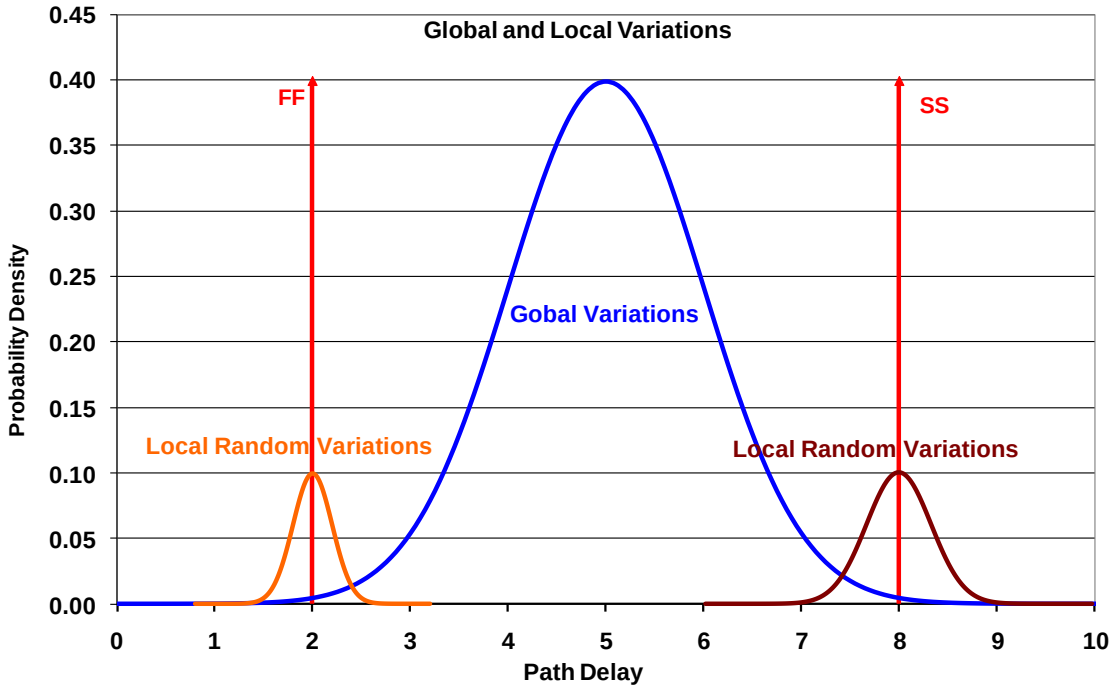

Fig. V: Un histogramme de délais qui montre l'effet des variations globale et locale avec les corners traditionnelles

Étude des variations aléatoires

Pour étudier l'impact des variations aléatoires, nous avons utilisé deux cas, l'un des mismatch locales, et l'autre pour mismatch globale.

- **Cas 1 : Mismatch Locales**

Les expériences sont basées sur des simulations utilisant des modèles spice industrielles qui incluent de silicium caractérisé mismatch. Les modèles utilisés sont de première génération montée en puissance des modèles de production et de processus en tant que telle pourrait montrer une plus grande ampleur de la variation par rapport au processus aujourd'hui. Toutefois, les tendances générales devraient être les mêmes. Nous avons utilisé les mêmes modèles de maintenir la cohérence sur toute la durée du projet comme cela se pratique dans des projets de conception.

Mismatch a été caractérisée par des simulations de Monte Carlo dans un simulateur spice industrielle avec 1000 échantillons de chaque série. Il existe deux approches pour caractériser mismatch. Première approche est une pleine Monte Carlo (MC), y compris les variations globales et mismatch, où l'effet de mismatch est extrait en différenciant les délais entre les deux voies similaires, l'un à le mismatch activé, et l'autre sans. En raison de même signal et de l'impact des variations globales égales, la différence donne directement l'effet de mismatch. Deuxième approche consiste à simuler mismatch que sur un corner de synchronisation dans un chemin avec un avantage de simulation en temps plus rapide et moins de ressources. Pour caractériser mismatch on soustrait la valeur nominale d'une quantité de sa valeur mesurée dans une course de MC. Les statistiques de distribution résultant nous donnent la valeur moyenne et l'écart type de l'impact de mismatch. Un modèle statistique complète avec globale et variations mismatch peut donner une valeur moindre en raison de l'effet de mismatch de réduire le plus rapide des échantillons, alors que les statistiques sur les corners mismatch donner des valeurs plus élevées en raison de limiter les cas de tensions de seuil.

Nous avons utilisé pour les bibliothèques de cellule standard CMOS 45nm processus de concentration sur les bibliothèques d'horloge. La moyenne, μ, et l'écart type, σ, de le mismatch des variations nous donnent les limites statistiques, μ ± 3σ, de la distribution. La pratique du design industriel utilise la variation en pourcentage par rapport au délai d'insertion. Utilisant les numéros de pourcentage, nous pouvons analyser l'impact de le mismatch long d'un chemin, qui est plus compréhensible pour un designer. Les valeurs x axe ont été normalisées avec l'insertion délai plus important (60 étapes) prises comme une seule et axe des y valeurs calculées pour l'insertion de délais normalisés pour préserver la forme de graphique.

Nous avons mesuré l'impact de mismatch sur l'insertion de délais, le skew et la largeur d'période, en faisant varier la tension d'alimentation, slew, la force d'entrainement, les types de cellules, corner traditionnelles et déséquilibré (SF, FS), et la profondeur de chemin d'accès (jusqu'à 60) pour trois cellules d'horloge pour différent taille des cellules (BF1 = 1x, 3x = BF2, BF3 = 6x) (Figure VI). Ces paramètres et ces mesures nous donnent une idée du compromis entre la puissance, de retardement, et la région, trois importants facteurs les plus à la conception. Nous avons utilisé une résistance au ratio capacité tirés du processus de gravure en 45 nm industrielle pour le routage des interconnexions pour modéliser l'impact de la dégradation et tua une référence comparable de la profondeur de chemin d'accès à la conception de taille.

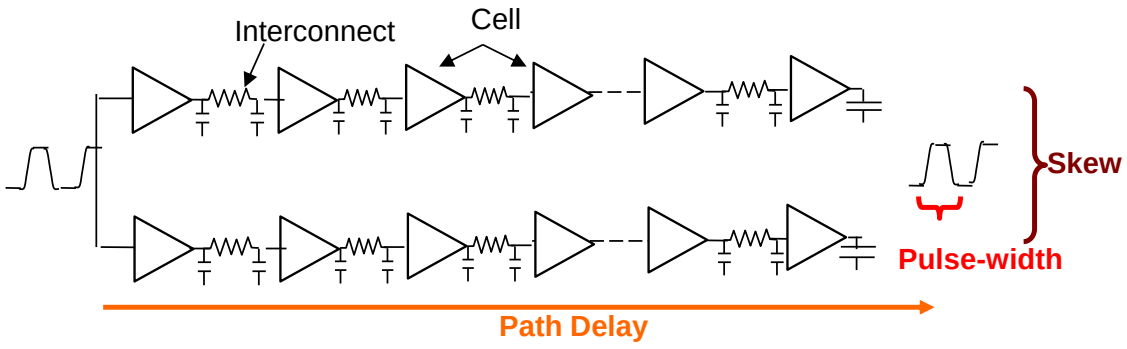

Fig. VI: Setup expérimental pour extraire la valeur de mismatch locale dans un chemin d'horloge sur le délai d'insertion, skew et largeur d'période

Cas 2 : Mismatch Global

Dans ce travail, nous avons caractérisé l'impact des variations sur un PVT CMOS 45 nm à faible puissance de cellules de bibliothèque horloge. La bibliothèque est spécialement conçue pour les arbres d'horloge et constitue un choix évident pour vérifier l'impact des variations PVT. Arbres des horloges ont des longueurs de parcours grandes, réparties sur l'ensemble de puce en passant par différents domaines de puissance qui les rend très sensibles à ces variations. La plupart des études sur l'impact des variations PVT se concentrer sur une ou l'autre de skew ou de retardement d'insertion. Toutefois, nous avons restreint l'analyse à largeur d'période dans le but de formuler des consignes d'optimisation. La fermeture de synchronisation dans la présence de ces variations est assurée par les corners et les marges ou sous forme de déclassement et les numéros de facteurs d'incertitude. Dans ce travail, nous utilisons les marges terme pour représenter tous les types de marges de manière à neutraliser les variations PVT.

La bibliothèque horloge utilisée dans cette expérience est un pouvoir faible bibliothèque 45nm avec une large gamme de tension d'alimentation qui lui permet de cibler plusieurs types d'applications. La demande varie de haute performance relativement à faible consommation énergétique très. Une bibliothèque d'horloge se compose de divers types de cellules requis pour conduire l'arbre d'horloge, les cellules combinatoires nécessaires à la génération d'horloge, la division et pulse shaping, horloge cellules ouverture de porte, flip-flops, etc. Ces cellules sont très optimisées et équilibrée pour atteindre l'équivalent du temps de montée et la chute du temps et des retardements respectifs. Considérant que le même est vrai pour d'autres cellules, il existe des concurrents objectifs d'optimisation en cause pour eux, comme le temps d'installation et temps de maintien, ce qui peut entrainé en moins que parfait caractéristiques de largeur d'période. À la tension nominale pour la bibliothèque, une cellule entièrement équilibré aura un impact minimum sur la largeur d'période pour les cas le pire corner. Toutefois, globale-à-p n mismatch peut entrainer la dégradation en largeur d'période plus élevée qui on peux voir dans la Figure VII.

Le travail est basé sur des simulations utilisant des modèles industriels spice avec des corners caractérisé à partir des résultats de silicium pour mesurer l'impact des changements dans le processus, de tension et de température. Le processus expérimental a été automatisé pour permettre des analyses multiples et de réduire la probabilité d'erreur. Simulations Spice fournir degré élevé de précision nécessaire pour mesurer l'impact des variations sur le retardement au niveau de la porte.

L'installation se compose d'un banc d'cellule dans un chemin d'horloge reliée à d'autres avec les interconnexions. Le signal d'entrée est une forme réaliste. Calcul de la différence de temps de propagation pour chaque étape entre l'entrée et la sortie de la cellule d'essai nous donne l'impact sur la largeur d'période. Les simulations ont été effectuées sur toutes les cellules dans une bibliothèque de cellule d'horloge en 45 nm. Nous avons également mesuré l'impact pour tous les lecteurs d'une cellule. La force d'entrainement est une meilleure mesure que la taille des cellules où il peut être directement perçu par le concepteur. Nous avons gardé le temps de transition standard à 55ps au pire corner, 0.90V, et -40 ° C. L'interconnexion de calcul et taux de résistance a été maintenu même que dans le 45nm industriels pour simuler la propagation des interconnexions réaliste.

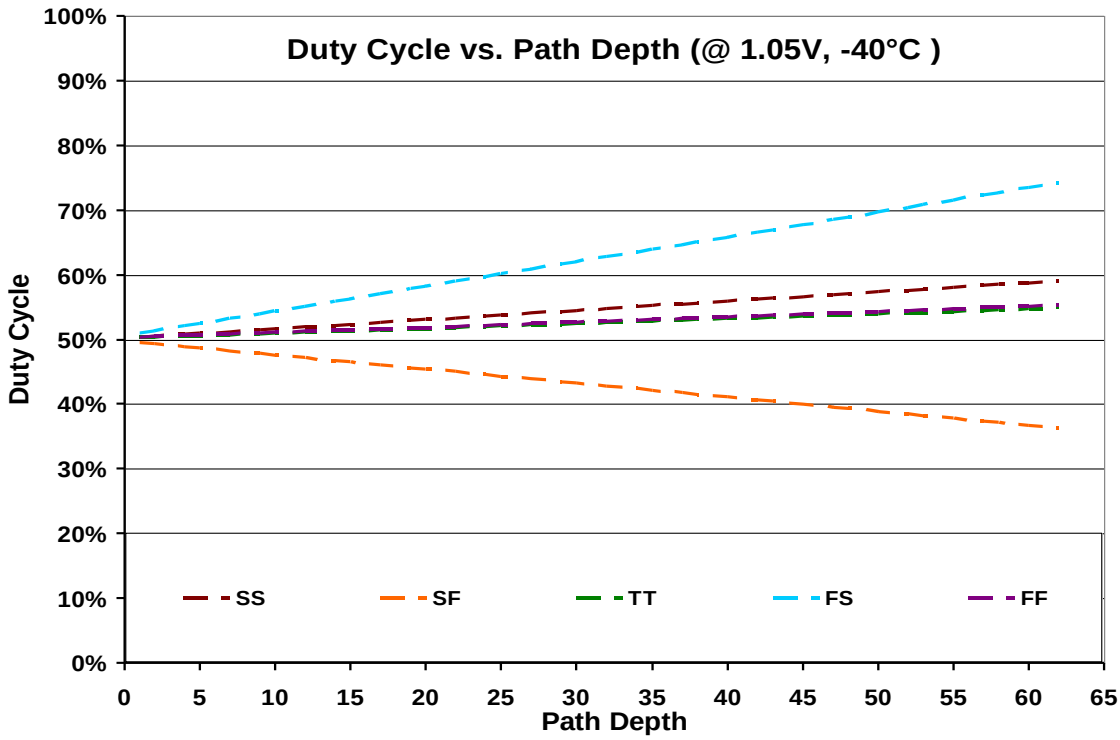

Fig. VII: L'effet de mismatch globale sur la largeur de période ou duty-cycle

Résultats et analyse

L'impact du mismatch sur le retardement, la valeur en pourcentage (ou de l'asymétrie) décroit exponentiellement avec la profondeur chemin, mais ne supprime pas complètement. Être une variation non corrélées, σ / μ était censé devenir négligeable pour de longs parcours (chemin de la profondeur de 60 pour nous). Toutefois, compte tenu des longueurs de parcours en cours de conception (moins de quelques ns), l'effet n'est pas négligeable. La valeur absolue de mismatch augmente le long d'un chemin, en ajoutant avec une moyenne quadratique (rms) la fonction à chaque étape. Il y a une décroissance exponentielle de mismatch en pourcentage (Figure VIII).

Il existe deux approches pour caractériser mismatch locales aléatoire en utilisant des simulations de Monte Carlo. Le premier est Monte Carlo avec des variations globales et locales, où l'effet de mismatch est extrait par différenciation de délais entre les deux

chemins, l'un avec mismatch activé et l'autre sans. L'impact des variations globales est annulé comme c'est la même pour les deux chemins. La deuxième approche consiste à simuler mismatch sur un corner. Toute variation de délais entrés deux exécutions est le résultat de variations locales. La valeur de mismatch peux obtenir en soustrayant la valeur nominale par la valeur mesurée pour chaque essai. L'avantage est que on garde les fonctionnalités corners et regarde juste les variations locales. Figure VIII montre les résulta pour les deux type. Le mismatch sur les corners encapsule l'on dans le cas réaliste, et donc toute les résulta dérivé pour un mismatch sur corner est valide pour le cas réaliste.

Il existe une relation non linéaire entre le retardement et des variations de mismatch qui provoque une valeur moyenne non-zéro pour le mismatch. L'effet est plus prononcé pour les cellules petites. L'effet est des marges inégales négatives et positives. En utilisant seulement variations (σ) pour les marges de variation peut entrainer l'échec dans le timing tout en utilisant la plus grande valeur pour les deux peut entrainer en sacrifier les performances réalisables. Au plus tardé, la différence est plus marquée pour les variations positives en raison de valeur non nulle en moyenne. Mismatch étant fonction de la tension de seuil (V_{th}) et la tension d'alimentation (VDD), V_{th} faible (LL) transistors ont un impact mismatch réduite que sur transistors standard (LS).

Mismatch est considérée comme critique pour setup et hold (dépendent sur skew), où même pour le non-pire cas des processus, une grande valeur de mismatch peut entrainer un échec de synchronisation. Le facteur le plus critique est la largeur d'période, où est la différence entre les deux bords en passant par les cellules mêmes, mais différents transistors. En outre, l'impact sur le bord en passant par petits transistors dans une chaine non-inversés est pire et fait donc un type d'période plus importante que l'autre (par exemple, haute de plus que de basse).

Corners débalancée (ou SF / FS) sont mauvais pour la largeur d'période en raison de grande différence dans les NMOS et PMOS courant qui affecte l'ascension et la chute différemment. Présence de mismatch sur les corners peut aggraver cette situation. En

outre, les pires conditions pour l'période peut changer avec la taille de la cellule et de retarder retardement. Donc, le corner SF peut être le pire des cas à un certain délai et SS sur autre. Une chaine mixte de cellules peut nécessiter des calculs complexes pour prédire l'effet.

Slew est un facteur important dans la construction de l'arbre d'horloge et affecte le retardement et mismatch absolu. L'augmentation de mismatch est corrélée à retardement de chemin. La même chose n'est pas vraie pour la largeur d'période que peut avoir un effet plus important en raison du retardement important d'insertion. Grand cellules peuvent être utilisées pour réduire les déséquilibres, mais plus interconnexions qui leur sont associés peuvent augmenter la dégradation. Ils sont plus adaptés aux grands fanout. Une solution de compromis pourrait consister à utiliser à moyen et à faible lecteur cellules à un stade proche de la racine qui composent les voies communes pour la plupart des registres reliés logique et les cellules de conduite élevée pour les stades à proximité des nœuds de feuilles qui composent le parcours hors du commun.

Tension d'alimentation a un impact important sur mismatch (<1V) (Figure IX). Basse tension sont principalement utilisés pour le mode basse puissance lorsque le système n'est pas nécessaire pour fonctionner à des fréquences élevées et le montant même de le mismatch pourrait être absorbée dans l'architecture du système. L'impact de la température sur asymétrie est bien inférieur à la tension et fait une différence que pour la basse tension et les températures élevées.

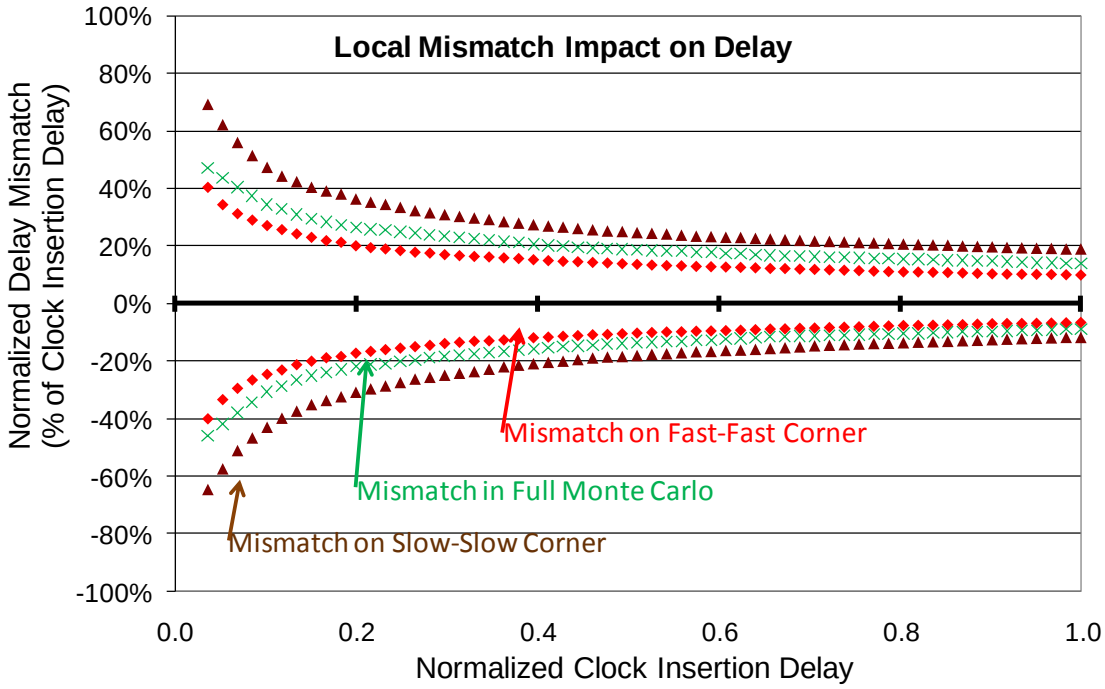

Fig. VIII: L'effet de mismatch dans un chemin d'horloge pour trois cas. 1) Dans un corner slow-slow, 2) Dans un corner fast-fast, et 3) Avec les variations globale et locales

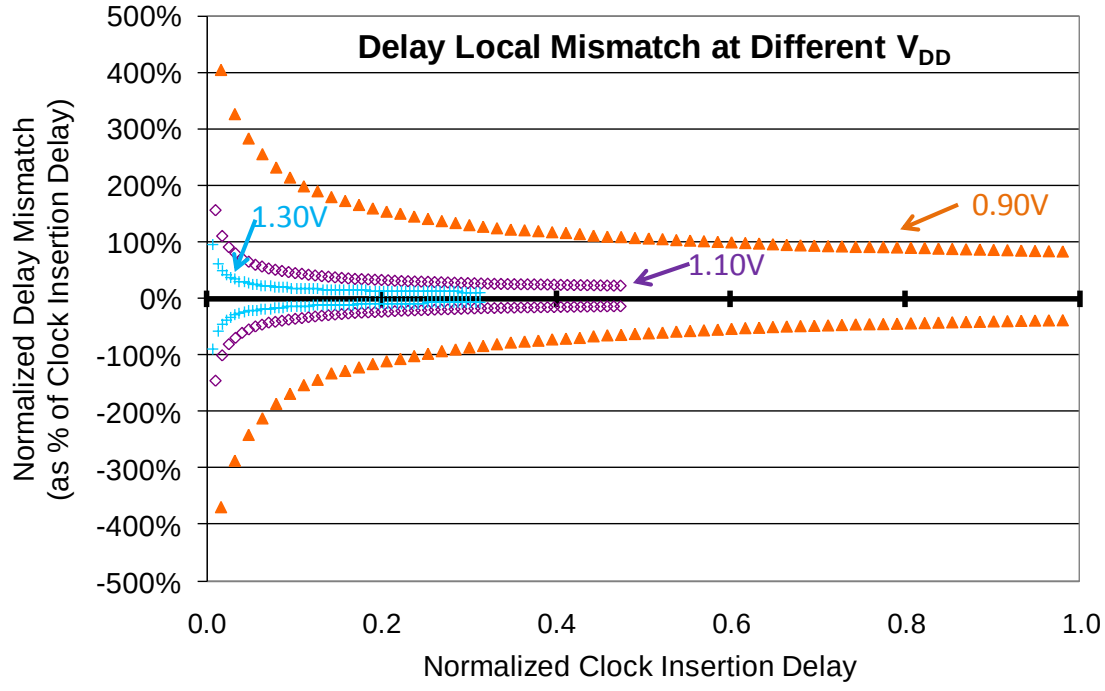

Fig. IX: L'effet de tension sur l'ampleur de mismatch sur délais d'insertion

L’effet de mismatch globale sur la largeur de période peut voir dans le Figure X sur un petit buffer pour plusieurs tensions. Il y a un grande effet pour les tensions < 1.0V. Lors de l'examen mismatch globale, nous avons ciblé trois domaines d'application différents, de haute performance à haute V_{DD} (HP), de faible puissance à faible V_{DD} (LP) et à différent modes de travail (HPLP). Dans les applications HP, par exemple les processeurs de télévision numérique, la puce nécessite des fréquences d'horloge élevées, mais les niveaux de tension élevés limitent l'impact des variations PVT. Dans les applications de LP, par exemple processeur mobile, le but est de réduire la consommation d'énergie et travaille donc à basse tension à fréquence d'horloge inférieure. Cependant, l'ampleur des variations PVT est beaucoup plus élevée à basse tension. Dans les applications HPLP, par exemple processeur du netbook, la performance change avec les besoins. Ces puces ont pour maintenir la fréquence d'horloge élevée ainsi que la fonctionnalité de basse tension.

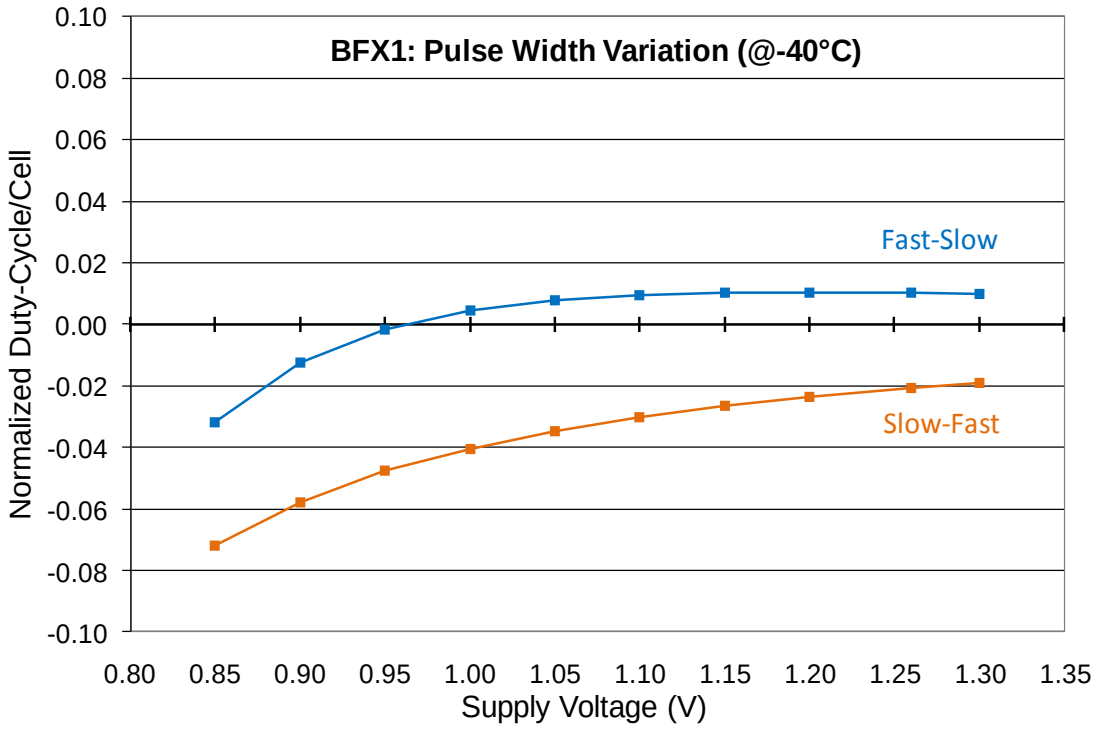

Fig. X: L’effet de mismatch globale sur la largeur d'période pour différent ampleur de tensions pour une petite cellule (buffer)

Une cellule de taille faible optimisée pour les applications HP est la cellule nominale. Elle a un impact marginal de la tension et la température sur la largeur d'période au-dessus de 1V à laquelle les demandes de HP travaillent habituellement. Les corners SF et FS représentent les limites de largeur d'période à la tension> 1V justifiant l'importance du mismatch n-à-p globale. Cependant, à basse tension (<1V), le variation de impulsion augmente fortement (Figure XI).

L'ampleur des variations de largeur d'période est importante pour virage lent à basse tension suggérant un plus mauvais comportement PMOS et peut être expliqué par le trou de la diffusion de petites cours. L'ampleur réelle varie selon le type cellulaire et lecteur. Un autre facteur pointant vers la diffusion du courant est l'inversion de température observée dans cette région. L'impact se situe surtout dans le premier et le plus petit stade d'une cellule à deux étapes. Comme la cellule-unité est augmenté, de même que la taille de la première étape et les cellules passe moins de temps en faible inversion. Après certaine taille / consommation de courant, il ya un impact marginal sur ΔPulse Largeur sur l'accroissement de la taille plus. La largeur ΔPulse reste à peu près la même tension et la température. Si une telle sorte de cellule qui est bon pour l'utilisation, il augmente la consommation d'énergie.

À basse tension, la réduction des forces actuelles de la fuite des transistors de rester en faible inversion de plus longue durée. Dans cette région, le courant de drain a une relation exponentielle avec prise de tension de seuil de l'impact du mismatch n-à-p globale plus importante. En outre, la température relation actuelle en faible inversion est opposée à celui de forte inversion. En fuite forte inversion actuelle est composée de la dérive actuelle tout en faible inversion, il est composé de diffusion du courant. Une augmentation de la température en forte inversion va augmenter l'agitation thermique des électrons qui empêche la dérive actuelle. Au contraire, une augmentation de la température en faible inversion augmente la distance moyenne parcourue par un porteur de charge, augmentant ainsi le courant de diffusion par gradient de concentration. Plus le

transistor reste en faible inversion plus sensibles que d'avoir une température un comportement inverse global.

Les changements dans la pente d'entrée ont un impact important sur la largeur d'période pour la mismatch n-à-p globale. Pour une transition rapide, il ya une différence négligeable. Toutefois, comme le temps de transition est augmenté, le montant de temps consacré à l'augmentation faible inversion et à la suite de la relation exponentielle avec la tension de seuil, la variation de largeur d'période augmente de façon spectaculaire.

Pour les applications HPLP, plus ΔPulse à basse tension peut être maintenue mais elle doit encore être dans les limites. Sur l'augmentation de la taille de la n-transistor dans la première étape de la cellule par 10%, ΔPulse devient ainsi moins sensible aux variations de tension. Elle déplace vers le haut pour donner une marge équivalente à faible période et d'périodes à haute. Une telle optimisation peut garantir la fonctionnalité de puces à toutes les tensions. L'impact sur le délai global de transition le temps est négligeable en raison de la deuxième phase dominante. L'augmentation de taille augmente la capacité de grille comme on le voit par la porte précédente, mais est négligeable par rapport à l'horloge capacité d'interconnexion (Figure XI).

Pour les applications LP, les variations de largeur d'périodes doivent être contrôlées pour la plupart des basses tensions. Sur l'augmentation de la taille des n-transistors dans la première étape par 20%, ΔPulse est devenu presque linéaire avec la tension d'alimentation. L'optimisation est pire pour la haute tension, mais être un cellule LP est acceptable (Figure XI).

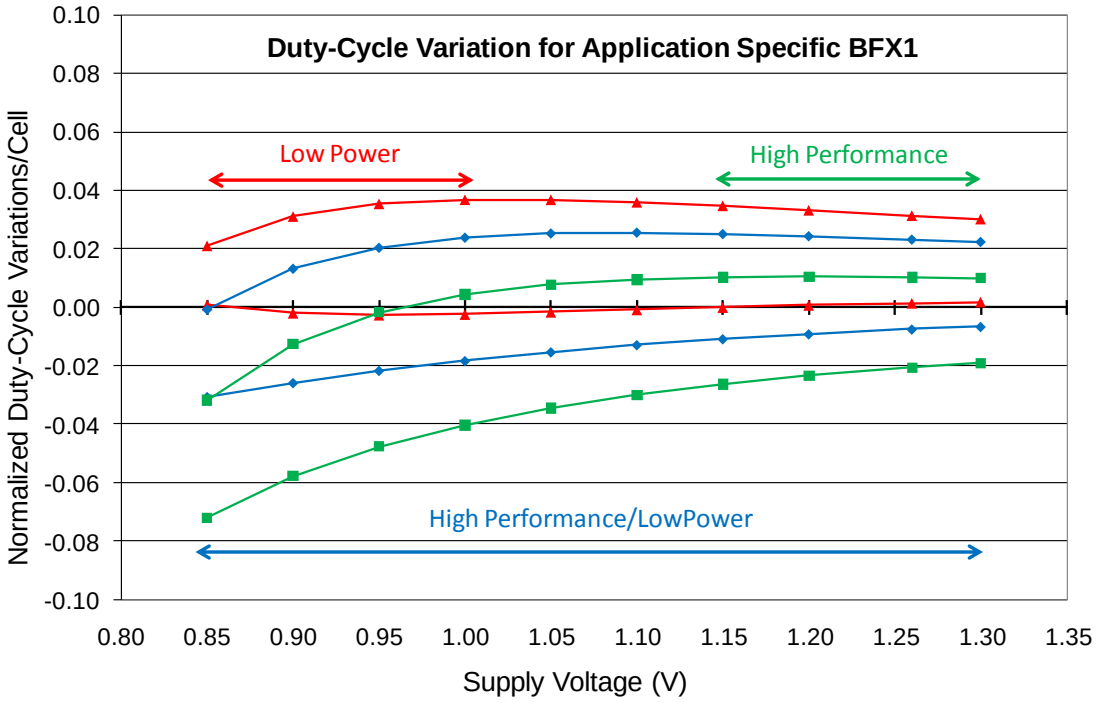

Fig. XI: Le variation de duty-cycle pour différent type des cellules optimisé pour trois applications: 1) High Performance (HP), 2) Low Power (LP), et 3) High Performance & Low Power (HPLP)

STA en présence de mismatch

Nous avons démontré une technique de STA y compris effet de mismatch qui peut servir de pont entre le STA traditionnel et SSTA. Elle est concentre vers la conception arbre d'horloge que c'est la quantité la plus touchée dans la conception numérique. Nous avons caractérisé l'impact du mismatch au niveau de la cellule et l'a utilisé pour prédire l'impact mismatch sur les chemins pour les réseaux d'horloge numérique. Nous avons été en mesure de prédire l'impact de retard dans la marge d'erreur de 10%. L'objectif est d'essayer de prédire le $\mu \pm 3\sigma$ statistiques (changement de moyenne, standard deviation) autour des cas limites.

Les deux équations ci-dessous représentent l'impact maximum et minimum de mismatch sur un chemin. Nous avons caractérisé le retardment de cellules pour en extraire μ (moyenne), σ (standard deviation) et M (valeur nominal).

$$t_{\max,\ mismatch\ \ (corner\ )} = (\mu_1 + \mu_2 ... + \mu_n) + (M_1 + M_2 ... + M_n) + 3 * \sqrt{\sigma_1^2 + \sigma_2^2 ... + \sigma_n^2}$$

$$t_{\min,\ mismatch\ \ (corner\ )} = (\mu_1 + \mu_2 ... + \mu_n) + (M_1 + M_2 ... + M_n) - 3 * \sqrt{\sigma_1^2 + \sigma_2^2 ... + \sigma_n^2}$$

Nous avons comparé les valeurs calculées avec les valeurs extraites de SPICE et a trouvé un bon match. Le procédé peut être appliqué dans des outils commerciaux pour STA avec minimal des frais.

Le nombre de points requis pour être qualifiée peut être réduit en utilisant des équations analytiques pour prédire mismatch des différents slew, tension et taille. L'impact du mismatch dépend de la valeur de ces paramètres et permet ainsi de prévoir le changement de la valeur de mismatch.

Conclusion

Cette thèse est centrée essentiellement sur l'estimation et la réduction globale et locale de l'effet mismatch aléatoire sur le timing dans les conceptions ASIC. L'aspect de différenciation, c'est que nous nous sommes limités à l'utilisation de techniques de conception pour réduire les délais. L'objectif est de réaliser des circuits plus robustes en gardant à l'esprit les compromis impliqués et de permettre ainsi une comparaison directe des couts et des avantages. Nous avons pris une approche multidimensionnelle pour réduire les marges de variation sur puce nécessaire dans l'approche corner. Nous avons analysé les principaux éléments touchés par le mismatch (local et global) et avons conclu que grâce à une conception robuste et les marges de variation sur puce, nous pouvons contrôler son impact dans des limites gérables pour les nœuds de courant. Les solutions exotiques comme l'utilisation de structures de transistors ou un autre procédé technologique peut être utilisée dans les ganglions de pointe lorsque l'amplitude des variations est trop élevée pour être maitrisée par les seules méthodes de conception.

Une variation de mismatch à la méthode actuelle d'analyse statique de temps a été proposée pour calculer les marges chemin spécifique adapté pour les corners individuels. La méthode de caractérisation des cellules nécessitant un minimum de temps a été proposé, tout en maintenant la précision. Équations analytiques pour accélérer le

processus de caractérisation ont été élaborées avec la marge d'erreur introduite par eux. Les simulations Spice a confirmé l'exactitude de la méthodologie proposée. Il peut être mis en œuvre dans les outils de CAO actuels avec un léger surcout.

Les stratégies d'optimisation spécifiques ciblant les retards ou la puissance pour les chemins d'horloge ont été proposées en utilisant une combinaison de paramètres, dont la tension de seuil, la longueur de grille, tension d'alimentation, et la force d'entrainement. Les avantages et inconvénients de chacun ont été répertoriés et peuvent aider à choisir la meilleure stratégie pour une application donnée en présence d'asymétrie. Un ensemble de règles de conception avec des gains subjective de limiter l'impact sur les chemins mismatch d'horloge ont été données qui aideront à créer un design plus robuste.

Une stratégie d'optimisation des applications spécifiques dans des ASIC a été proposée pour limiter l'impact du mismatch globale. Un sous-ensemble de cellules d'horloge dans la même bibliothèque optimisée avec les applications spécifiques à l'esprit peut limiter les variations de la largeur des périodes. La méthodologie proposée exige la caractérisation d'un petit sous-ensemble de cellules et de modifier quelques règles pour inclure un paramètre d'application cible qui vous aideront à choisir le sous-ensemble spécifique. L'approche se situe entre la conception full custom et la conception de cellule standard en utilisant le meilleur des deux. Le gain est plus dans la région de basse tension, où les variations d'période sont les plus élevés.

ASIC dessins en utilisant la méthode de cellule standard utilise généralement des arbres cellule d'horloge en raison de leur capacité de régénération du signal. Nous avons examiné les limitations favorisés et d'un arbre d'horloge inverseur en présence de le mismatch globale et locale. En considérant que les gains sont limités au niveau de la haute tension et/ou basse tension de conception, cela peut bénéficier de manière significative à réduire les déséquilibres variations impact. La réduction du nombre de transistors permettra de renforcer les économies d'énergie qui sont importantes dans cette région.

Nous avons également proposé une approche pour mesurer la précision du modèle avec de simples mesures de retardement RO. L'approche permet de vérifier en utilisant les circuits de test simple qui peuvent être et sont incorporés dans des plaquettes et meurt. Il permet la mesure rapide de mismatch local ou globale et confirme l'exactitude du modèle.

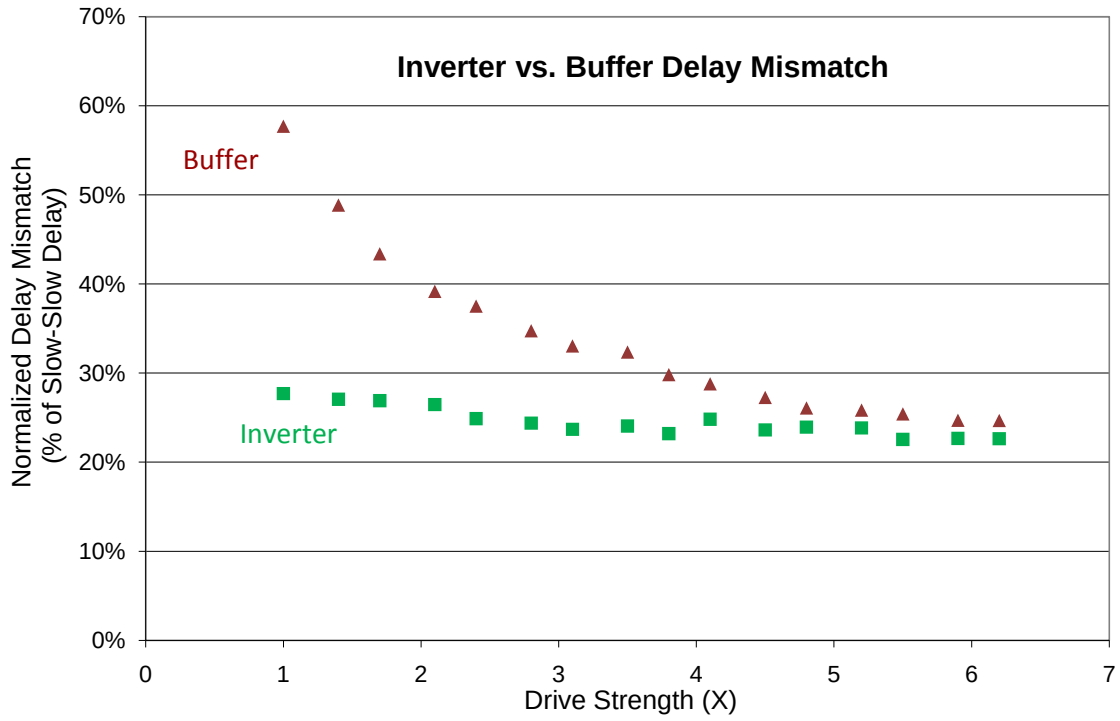

Fig. XII: L'effet de mismatch locale sur l'inverseur et le buffer pour différent tailles de transistor

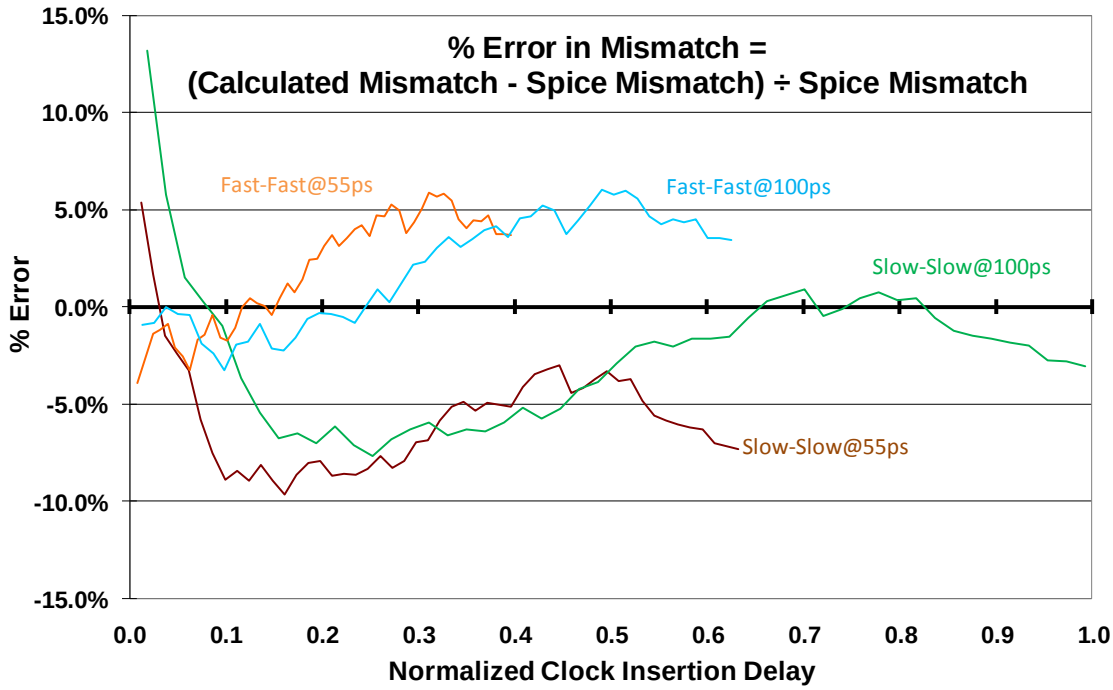

Fig. XIII: Le erreur pourcentage entré le ampleur de mismatch locale prévu par notre system et l'on extraire par spice sur le même chemin

List of Symbols

A	Gate area
a_S	Subthreshold swing coefficient
a_{Vth}	DIBL coefficients
b	Branching effort for a stage in a path
B	Path branching effort
C_D	Depletion layer capacitance
C_L	Output load capacitance
C_{ox}	Gate oxide capacitance = ε_{Ox}/t_{Ox}
E_C	Fitting factor for saturation field
E_{sat}	Critical electric field where carrier velocity saturates
E_{sw}	Switching activity factor
F	Path effort
f_{clk}	Clock frequency
g	Stage logical effort calculated based on topology
G	Path logical effort
h	Stage electrical effort calculated using input and output loads
H	Path electrical effort
I_{ds}	Drain to source current
$I_{ds,leak}$ / I_{leak}	Leakage current at VGS=0
$I_{ds,sat}$	Saturated drain current
I_{subth}	Subthreshold current
I_{subth}	Subthreshold current
k	Boltzman's constant
K	Loading
k_0, k_1, k_2	coefficients of loading K
L	Physical gate length
l_D	Debye length (~1.2nm)
L_{eff}	Effective gate length
n	Subthreshold parameter
Ñ	Optimal number of stages
N_A	Doped silicon carrier density
n_i	Intrinsic carrier concentration at 300 °K = $1.45 * 10^{-10}$ cm^{-3}
P	Gate perimeter
p	Stage parasitic delay
P	Path parasitic delay
$P_{Dynamic}$	Dynamic or Switching power
q	Electron charge
Q_{Dep}	Depletion charge under gate
Q_{fc}	Fixed charge due to imperfections in silicon-oxide interface and doping
S	Subthreshold swing
s_C - s_L	Skew difference for capture and launch flops
S_{subth}	Subthreshold swing

T	Temperature in Kelvin
T	Clock period
$t_{cq,L}$	Clock to Q time for launch register
T_d	Nominal gate delay
$t_{data,max}$	Maximum data delay between launch and capture flops
$t_{data,min}$	Minimum delay between launch and capture flops
$t_{hold,C}$	Hold time for capture flop
t_{Ox}	Gate oxide thickness
$t_{setup,C}$	Setup time for capture flop
V_{bs}	Body bias voltage
V_{DD}	Supply voltage
V_{DS}	Drain source voltage
V_{fb}	Flat band voltage
V_{GS}	Gate source voltage
V_S	Source voltage
V_t	Thermal voltage
V_{th} or V_{th0}	Threshold voltage
W	Physical gate width
w	Variational parameter ≈ depletion width
W_{eff}	Effective gate width
W_{sd}	Surface potential and source and drain ends of a channel
X_j	Junction depth
Γ	Body bias coefficient
ε_{Ox}	Gate oxide permittivity
ε_{Si}	Silicon permittivity = 1.06 * 10-12 Farads/cm
η	DIBL coefficient
λ	Body effect coefficient
λ	Fitting factor for DIBL and channel length modulation
λ_b	Models V_{th} roll-off
λ_d	Models DIBL effect and depend on L, t_{ox}, W_{sd}, X_j
σ/T_d	Gate delay variability
σ_P	Perimeter variance
σ_{Vth}	Threshold voltage variance
Φ_B	Bulk Potential
Φ_F	Surface potential
Φ_{ms}	Work function difference between gate and silicon substrate (= Φ_{gate} - Φ_{Si})
μ_{eff}	Effective mobility

1 Introduction to Variations in Digital Design

Semiconductor chips have been at the forefront of technological revolution and helped bring about a social change in the last few decades. They have enabled humans to tackle issues like climate modeling, weapons design, DNA sequencing, drug development, etc; while at the same time, they have aided in improving the quality of life for public with devices like digital high definition television, GPS, intelligent refrigerators, smart phones, multimedia entertainment systems, game consoles, etc. Semiconductor chips have become ubiquitous in the modern world.

This phenomenon has been the driving force behind increasing chip density and performance that has since come to known as the Moore's law. The size of basic building block in semiconductor chips, i.e. a transistor, has come down to 10's of nanometers, i.e. three orders or magnitude less than the diameter of a single strand of human hair. Fabricating a device of that size is a momentous challenge in itself [124], [32]. With human ingenuity, we have been able to find solutions to manufacture succeeding devices [102]. However, the solutions are not always perfect introducing new challenges in controlling the accuracy and precision of produced devices. Thus, the manufacturing process introduces some error between the desired and actual device, the relative importance of which has been increasing as devices become smaller and more difficult to fabricate [24].

With each technology node, more phenomena start to have a noticeable effect on the transistor characteristics. These manufacturing fluctuations constitute process variations and include any phenomenon that can create a deviation in physical properties of fabricated devices. In addition to manufacturing variations, transistor performance is also affected by voltage and temperature fluctuations during chip operation. Process, Voltage, and Temperature (PVT) variations constitute among the big challenges in path of transistor scaling [99], [124], [32]. Digital ASIC (Application Specific Integrated Circuits) companies have to guarantee the performance and yield of their products and thus designs do not benefit from binning used in microprocessors that reduce the risk of timing failure. As such, is necessary to have timing closure through timing analysis tools validating the designs at worst-case and best-case process along with any additional corners demanded by the customer. These corners represent the limits of process variations and thus characterize the maximum timing variations. They are combined with limiting operational temperature and voltage along with other parameters like jitter and on-chip variation margins to create the timing cut-off limits. Any die beyond these limits are discarded reducing yield and affecting cost per die.

This method was sufficient to guarantee timing and yield until recently. However, the number and magnitude of variations in chips are increasing with each technology that necessitates additional timing corners increasing the design flow effort and time. Moreover, either the total guard-band applied during static timing is increased, or the risk of affecting yield is increased. The aim of this project is to look at how these variations impact digital circuits and how we can reduce the overall impact. We limited the scope to gate/path level for two reasons- one, these parameters are understandable to designers and two, they have a direct relation to overall circuit performance. The approaches to make circuits more robust are limited to design to have minimum impact on design flow, thus enabling fast and easy implementation.

1.1 Process variations

Process variations can be defined as a difference in intended and actual physical makeup of a semiconductor device caused by fluctuations in fabrication process including equipment, material, and processing. It induces deviation of electrical properties from the targeted value creating a change in overall behavior of the circuit. Too large a deviation can result into functional failure reducing product yield [99] whereas smaller deviation can affect product efficiency and influence end user experience. Process variations are classified using two criteria- nature and predictability.

1.1.1 Nature

There are two categories of variations- Inter-die (Die-to-Die or D2D) and Intra-die (Within-die or WID).

1.1.1.1 Inter-die

Variations whose effects are different from one die to another but are constant for all devices inside the die are called Inter-die variations. They encapsulate lot-to-lot, wafer-to-wafer, and die-to-die variations in manufacturing process. As the impact is consistent inside a design, they can be clubbed together in a single entity to estimate the impact of these variations. Principally, they influence the spread of electrical properties and affect the product yield [65]. All devices inside the die share the same transistor I_{ON} current. However, the magnitude may differ from one device type to another, i.e. between NMOS and PMOS. Any die affected by Inter-die variations can lie on any single point in the elliptical area in Figure 1-1, which is a general representation of transistor I_{ON} current. The difference in NMOS and PMOS device determines the spread of ellipse's belly or the balanced and unbalanced nature of Inter-die variations. Most of the fabrication steps for both these devices are same correlating the variations caused by these steps. However, the doping step is inherently different to create different types of devices and thus the variations introduced in that step are different.

1.1.1.1.1 Balanced

If the variations affect NMOS and PMOS device in equal manner, it is known as balanced variations. Traditionally they constitute the limiting cases (worst and best) for delay. In designer's terminology, the worst-case delay or corner is known as SS (slow NMOS & slow PMOS) and the best-case delay or corner is known as FF (fast NMOS & fast PMOS), as shown in Figure 1-1. In any device, the rise and fall edges are affected in same manner.

1.1.1.1.2 Unbalanced

If the variations affect NMOS and PMOS device in different manner, it is known as unbalanced variations. They constitute limiting cases (worst and best) for pulse-width or duty-cycle as the rise and fall edges are affected in opposite manner. In designer's terminology, the two limiting cases or corners are known as SF (slow NMOS & fast PMOS) and FS (fast NMOS & slow PMOS), as shown in Figure 1-1.

1.1.1.2 Intra-die

Variations whose effects are different from one device to another inside a die are known as Intra-die variations. They encapsulate Within-die variations in manufacturing process as well as atomistic variations caused by limitations to material and process. The impact of these variations is not consistent inside a design, thus requiring a separate classification to estimate the impact of these variations. Principally, they affect the mean of variation distribution [65]. Any device affected by Intra-die variations can have a different transistor I_{ON} current and can lie at any point in circular area around a given point determined by Inter-die variations as shown in Figure 1-1. The amount of variations determines the diameter of the circular area and can vary from one point to another in the ellipse. The circle and ellipse in Figure 1-1 are general representation and the real shape might differ from them.

Intra-die variations can create functionality failure even in non-worst case corners as they affect the transistors differently. Differential parameters like skew and pulse-width are prone to these variations. Until recently, its effect in design could be safely neglected owing to the small overall impact caused by averaging effect in random variations as well as the small magnitude of variations [106]. However, rising intra-die variation magnitude and larger frequencies have made it increasingly important to consider its impact [112].

Intra-die variations create a difference in electrical properties of neighboring transistors, due to which two similar paths on the same die can exhibit different delay and power metrics. It can increase skew between two clock paths that can limit the design frequency or complexity. For a given design, larger skew means more margins on the clock pulse, which means larger clock period. For a given clock frequency, larger skew can put a limit on the clock path depth thus limiting the die size or design complexity. With increasing consumer demand, die-size and associated clock frequency are increasing and intra-die variations can create bottlenecks in their growth.

1.1.2 Predictability

There are two categories of variations-Systematic and Random variations.

1.1.2.1 Systematic variations

They can be defined as methodical variations in fabrication process due to equipment non-uniformity through space and time. The impact of these fluctuations on performance can be predicted using simulated models. Impact of systematic variations can be generally predicted using simulation models and design/equipment information. Their effect can then be minimized by modifying the fabrication process or mask data. Some of the systematic variations might have a small impact on overall performance that cannot justify the cost required to mitigate the impact. As such, they are put together with random variations.

1.1.2.2 Random variations

They can be defined as statistical fluctuations in manufacturing equipment introducing performance variations between different dies or between different elements in the same die. Random variations by name are unpredictable. However, being statistical in nature, their distribution, or minimum and maximum variations can be extracted. Using statistical minimum and maximum variations for different random variations, worst-case corners are created. If the fluctuations were small, worst-case scenarios for all parameters could be considered with little effect on chip performance. However, the magnitude of variations is quite large and considering the absolute worst case may not allow designs to meet targets. Thus, statistical worst-case variations are used to create the parameter worst-case scenarios, using which most of the chips will be within design specifications and a rare few will be outside those limits. Such an approach enables to gain maximum possible yield (thus reduce cost per die) without losing a great deal on performance. It should be kept in mind that larger the limiting case corners or more the yield, bigger the performance variation.

1.2 Voltage variations

In an ideal design, supply voltage is constant irrespective of the location of device in die or the moment in time, except the intentional cases where supply voltage may differ due to power-off, different voltage domains, different working modes, etc. However, in the real world there are always some amount of variations present for supply voltage in space and time due to glitches, line resistivity, supply fluctuations, etc. Even intentional changes in supply voltage can have unintentional effects on circuit performance. ITRS specifications allow for 12% fluctuations in nominal supply voltage [51], i.e. designers must allocate for that much supply voltage variations. However, intentional voltage changes can span a large scale from very low power standby mode to high performance active mode. As supply voltage is inherently linked to circuit performance, any changes have a direct impact on the same. Moreover, it also affects the impact of variations and thus constitutes an important factor to be included while studying variations.

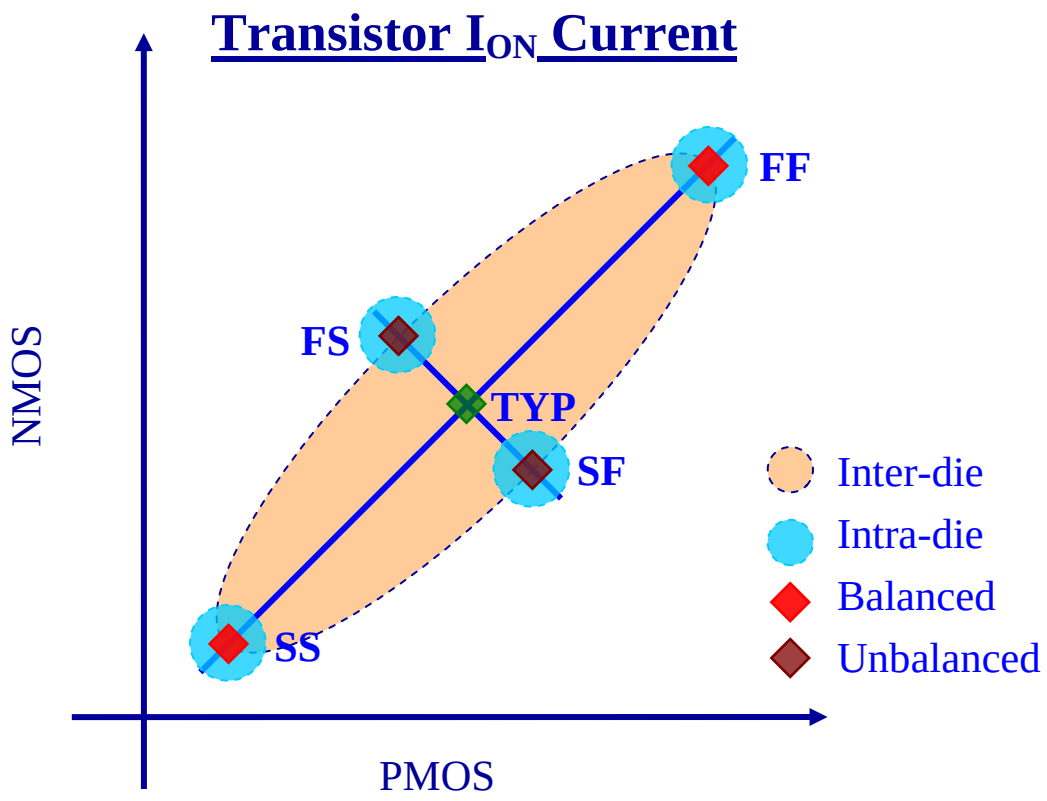

Figure 1-1: NMOS to PMOS transistor ION current

1.3 Temperature variations

Temperature variations include the ambient temperature variations lying between -40°C to 125°C for most industrial applications and the in-die temperature variations caused by activity affecting transistor mobility and issues like hot electron effect. Temperature has an effect on carrier mobility and through that on transistor current. Industrial circuits are rated to work within a given range to discount the location of use factor. Typically, it has not been a big factor on circuit performance but still has a non-negligible effect. Sub-100nm transistors can also see temperature inversion effect in low voltage region where a device may have worse performance at lower temperatures. As temperature has a higher impact in weak inversion region, it is necessary to consider its impact on other variations.

In the context of this project, we have focused on unbalanced inter-die variations, intra-die variations, changes in nominal supply voltage, and ambient temperature changes or block level changes. We have not considered the effect of in-die voltage and temperature variations like glitches, IR-drop, activity based variations, etc.

1.4 PVT variations in digital circuits

An ideal circuit will have the exact same specifications within the required limits for all the chips. However, variations introduce fluctuations in its specifications, i.e., chip frequency, power consumption, leakage power, reliable lifetime, etc. Keeping the impact of variations in mind, circuits are designed to sustain a certain level of fluctuations in its parameters without affecting the overall functionality and be within the required specifications.

Environmental i.e. voltage (V) and temperature (T), variations can induce fluctuations in transistor performance both at die-to-die and within-die level. These parameters are influenced by factors like power grid design, circuit placement, vector set, coupling capacitance, etc. With each technology node, impact of environmental variations on product performance is becoming a significant fraction. The impact of environmental variations is dependent on process corner. Traditionally, timing closure in ASIC design is verified at different combinations of voltage, temperature, and corner.

Limiting case corners may be good for inter-die variations. However, the same cannot be used for intra-die variations as the number of configurations is quite high and anyone of them can be the cause of specification failure. Testing such a large number of configurations is impractical. As such, to include the effects of intra-die variations, path specific margins are used inside the design. The role of margins is similar to corners, i.e. predict the maximum possible deviation but as a function of path configuration. While calculating the timing characteristics, these margins are included to obtain maximum and minimum delay for each path that in turn helps to calculate minimum and maximum design frequency and other specifications.

1.4.1 Variations in digital clock networks

Among the various components of a synchronous design, clock networks are most sensitive to intra-die variations owing to their differential nature, namely two parameters: pulse-width and skew [63]. Pulse-width can be defined as the arrival time difference of two opposite and consecutive edges passing through same path, as shown in Figure 1-2. However, the edges pass through different transistors, being opposite in nature, and thus perceive different amount of variations. This difference adds up along the path affecting pulse-width at the arriving flop. Flip-flops are made up of two stages, each working at either the low pulse or the high pulse. Thus, any reduction in size of pulse will affect the working of corresponding stage. This puts minimum pulse-width constraints for flip-flops and any violation result into wrongful latching of data. Variations in pulse-width will require guard bands or margins around pulse that may necessitate reducing clock path length (in turn reduce chip size or complexity or functionality) or increase the pulse period (in turn reduce the chip frequency).

Skew can be defined as the arrival time difference of same edge passing through two paths arriving at two different flops connected through a data path, as shown in Figure 1-3. As the edges pass through different cells, they perceive different amount of variations, which adds up along the path. In a perfectly synchronous system, the

edges should arrive at both flops together. Variations in skew will require guard bands or margins around arrival times that may necessitate reducing data logic between two flops (in turn increasing number of pipelined stages or chip size or reducing throughput) or increase the pulse period (in turn reduce the chip frequency). The relationship between minimum and maximum skew with other parameters like data, register parameters and frequency can be outlined with equations (1-1) and (1-2) [48], [35]. It is shown in Figure 1-4 where t_{C2Q} is source register Clock to Q delay, S is setup time of destination register, H is hold time of destination register, d_{max} and d_{min} represent the maximum and minimum time taken by data between two registers respectively.

$$(t_2 - t_1)_{min} \geq d_{max} + t_{C2Q} + S - T \quad (1\text{-}1)$$

$$(t_2 - t_1)_{max} \leq d_{min} + t_{C2Q} - H \quad (1\text{-}2)$$

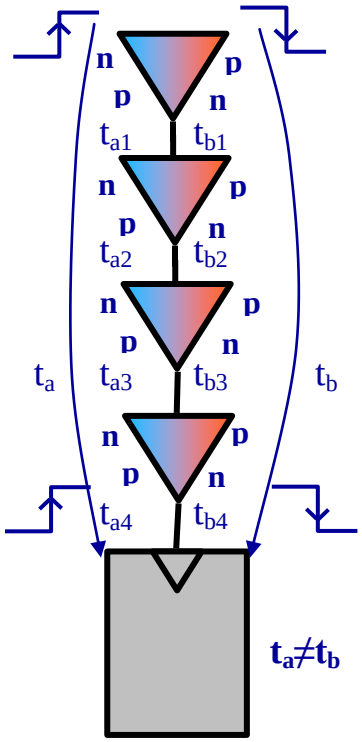

Figure 1-2: Intra-die variations in pulse-width along a path

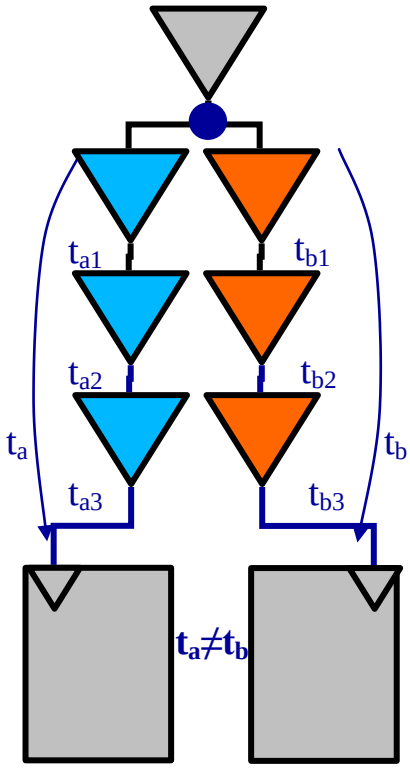

Figure 1-3: Intra-die variations in skew along two clock paths connecting a data path

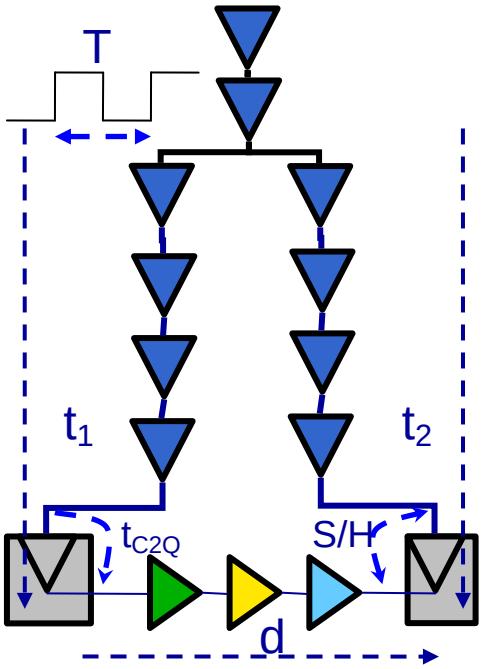

Figure 1-4: Impact of skew on setup/hold configurations

1.4.2 Variations vs. defects

Traditional way of determining a faulty chip is through tests like IDDQ that rely on measuring electrical parameters like supply current in the stable state. The measuring value should lie within pre-defined limits for the chip to be non-defective. Historically, these limits were quite large and outside the realm of functional chips. However, increasing magnitude of variations can force large values for these parameters even for a functional chip causing an overlap between the two sources. That makes it a tough task to determine the source, if it is because of parametric variations or because of a defect [97], [95]. Overlap between defect and design margin causes classes of defects that will behave as design margin and some defects might become un-testable due to increasing design margin. New and more complex tests have to be used to determine faulty chips increasing the overall cost as well as adding to the qualification time.

1.4.3 Analog behavior of digital networks

Electronic circuits are fundamentally analog in nature, i.e. it consists of continuous signals [60]. However, constraining the signals to a binary form of '1' and '0' has helped to accelerate the development and application of electronic systems. The analog behavior was still present but was factored out of the information transfer and was considered only as an unavoidable consequence designing the circuits to minimize this behavior. However, with advent of sub-100nm technology, the analog behavior has been steadily creeping back in the circuits in different forms. Exact signal shape during transition now constitutes a non-negligible factor in gate delay, glitches can take the signal level above '0' or below '1', rise and fall behavior differ from each other creating static noise margin like curves for simple gates, atomistic differences affect the shape of transistor characteristics causing difference in overall delay, etc. It is becoming important not only to study the behavior at digital '1' and '0' but also during the transition period. Circuits may not be completely off below threshold voltage and drain current may never reach the saturated value in the presence of high speed switching. These phenomena indicate that we have to look beyond the traditional digital parameters into analog domain to understand how the circuit is being affected in the presence of different variations.

1.5 Objectives

This thesis was divided into three main parts.

1.5.1 Identification of process variations and their mechanisms

The first stage involves studying and understanding the sources of variations and their mechanisms. Variation sources are generally divided into systematic and random. However, it is necessary to classify variations according to their impact on performance through space and time as well as their predictability and effect on

other elements. Such a classification can allow for easier identification and implementation of solutions. Rather than limiting the impact of a particular source, we can work on limiting the impact on circuits.

1.5.2 Estimation of variation impact on performance of digital circuits

The second stage involves estimation or quantification of variation consequences on performance of digital circuits. The metric are performance in terms maximum and minimal speed, dissipated power and leakage current. The absolute value of variation on a metric is not necessarily important. The goal is to be able to quantify the relative performance of a circuit compared to another to choose better configurations.

1.5.3 Evaluation of design methods and techniques to limit variation impact

The third stage involves identification and evaluation of various techniques to obtain a qualitative or quantitative improvement in circuit performance. The aim is to find design solutions that can help to reduce, limit, or predict the impact of variations without requiring a change in technology. Some of the proposed methods include regular design, variation aware cells, etc.
To limit the scope of work, we focused on digital clock networks in ASIC design using standard cells. Such an approach allows us to have a comparison with real industrial systems. It also allows us to consider the usability and ease of implementation of any proposed design optimization from industrial perspective.

2 State of the Art in ASIC Design

The following chapter will give an overview of different kinds of variations present in a digital design. It will also cover various aspects of digital design like power consumption, yield, clock network, timing analysis, etc. The goal is to give an outline of digital design and different perspectives that we can encounter during the course of this work.

2.1 Variation taxonomy

Variations can be classified based on two broad criteria: temporal and spatial.

2.1.1 Temporal

Temporal variations are defined as changes in design characteristics over a period of time. The time duration here is a relative factor and can be anything between clock cycle time to lifetime depending on the type of effect under consideration. Variations that tend to have an effect in a time-frame equivalent to clock-period are called short-term variations. Variations whose effect is seen over a period much longer than clock-period are called long-term variations. There are numerous clocks present in a design and the reference clock-cycle is determined by design component and variation under consideration. Time dependent variability is a strong function of capacitive loading, PMOS and NMOS device widths, activity factor, chip environment (VDD, T) and interconnect aspect ratio [106]. Temporal variations affect product reliability and thus determine the market perception of its quality.

2.1.1.1 Short term

Short-term variations occur during product operation from one cycle to another. Some of the main causes are change in activity affecting local temperature and glitches in local supply & signals, and presence of highly charged ions near critical paths. It can cause occasional errors and affect design performance over a small period. Repeated occurrence can affect overall design performance and user perception. The best defense against short-term variations is a robust design approach using redundancy, error correcting circuits and shielding for critical paths. However, the cost of using such methods is high due to larger silicon surface area. As such, their use may be limited to big or error critical designs. Error detection and re-execution methods like RAZOR can also be used [25]. It is difficult to do a real time activity based design simulation with 100% coverage due to the vast number of possible cases. Increasingly complex circuits and multiple power domains present on the same die affect the number of corners during timing check.

2.1.1.2 Long term

Long-term variations typically occur over a period much longer than the clock period. They are differentiated into two categories, reversible and irreversible, based

on the permanency of effect. The two categories are interrelated and can affect each other.

2.1.1.2.1 Reversible

Reversible variations are present over the operating period of a product that is much larger than the clock period but lesser than product lifetime. It consists of changes in supply voltage and ambient temperature that affect design performance. The effect is present only as long as the change in the concerned parameter is present. However, persistent high magnitude of temperature or voltage can expedite irreversible effects. Effect of reversible variations is included in design corners and appropriate worst-case corners are selected based on design targets.

2.1.1.2.2 Irreversible

Irreversible variations consists of effects occurring over product lifetime such as electromigration, hot spot degradation, hot-electron effect, negative bias temperature instability (NBTI) and general wear & tear [113]. These variations affect design performance permanently. Robust design techniques like maximum limit on line aspect ratio and worst-case simulation can find and remove bottlenecks. Ageing libraries are used to ensure product functionality over the contractual lifetime. The famous corollary to Murphy's Law that states, "A product fails soon after its warranty has expired" can be a consequence of these variations.

2.1.2 Spatial

Spatial variability is defined as difference between two devices based on their separating distance. The devices here can represent anything from individual transistors in a chip to whole lots and the separating distance is not necessarily in its literal sense. Spatial variability interacts with temporal variability to affect design performance. It can be classified based on two criteria: range of variation and degree of correlation between devices.

2.1.2.1 Range of Variation

Range of variation determines the separation between two devices. Typically, the effect can be grouped into two categories, Inter-Die and Intra-Die.

2.1.2.1.1 Inter-die

Inter-Die variation is defined as fluctuations of properties that are constant for one die but varies from one die to another. Broadly classified it consists of variations at die-to-die, wafer-to-wafer, and lot-to-lot level. Traditionally it is the dominant variation type affecting design objectives. The impact of Inter-Die variation is same for all transistors on a die. As such, worst-case process parameters can calculate the

limiting impact for a given design to meet the desired objectives. ASIC products are typically required to pass the contractual targets and have limited utility of achieving better performance thus promoting use of worst-case corners.
Inter-Die variations mainly consist of fluctuations in critical dimension, average doping, oxide thickness, sheet resistance, contact & via, etc. The main source of these variations is lack of manufacturing control caused by technological limitations. Prominent sources of inter-die variations include Rapid Thermal Anneal causing temperature gradients across the wafer, photoresist development and etching [106]. Inter-Die variations determine electrical characteristics variability of a die around a mean value [79]. Using combinations of worst and best process parameters for different quantities, corner cases are created that are then used to validate design specifications. Increasing number of variations have increased the number of design corners that need to be validated, which in turn affects the time and resources required to qualify a design driving up the cost per die. Intelligent corner selection along with derating & margins is becoming popular to find a compromise between yield and performance.

2.1.2.1.2 Intra-die

Intra-Die or within-die variations are defined as fluctuations of physical properties that affect the electrical properties of different transistors on the same die differently. It is also known as local mismatch. Intra-die variations are increasingly becoming an important factor in semiconductor chips. Using worst-case conditions for all transistors in a chip for local mismatch will be highly pessimistic as well as unrealistic. Moreover, differential parameters in a die may be more critical by a combination of worst and best case transistors. Thus, corner approach is generally rejected for such variations. Instead, bounding margins are used to limit the impact of within-die variations.
Principal sources of within-die variations are local dopant mismatch, line edge roughness, oxide thickness variations, polycrystalline granular structure, layout and neighborhood based effects, lens aberrations, etc [106]. Some of these effects are introduced by discrepancies in manufacturing equipment and others are caused by natural limits to current manufacturing process and materials. Local strain and RTA are also creating new sources of intra-die variations [111]. Currently, the proposed approach to include effect of within-die variations is through probabilistic models using statistical numbers to calculate the margins. The downside of such an approach is it can be expensively time consuming. Using worst-case margins can result in overly pessimistic and wasteful products. Margins built up through impact modeling of within-die variations and design-rules can be more realistic. Techniques can also be found to make the designs more robust against such fluctuations. These two approaches constitute the driving theme of this work.

2.1.2.2 Degree of Correlation

Degree of separation between two devices also determines the degree of correlation between them, which in turn decides the impact of variation in one device on the

other one. There are two types of variations under this category: systematic and random.

2.1.2.2.1 Systematic

Systematic variations have a correlation factor (>0) between two devices for a given source of variation. Generally, this factor changes with distance as well as neighborhood. In most cases, the variations are only partially correlated due difference in neighborhood induced stressed. Imperfections in manufacturing process and equipment are primarily responsible for systematic variations. Degradation of manufacturing equipment with time can cause systematic variations between different lots. Difference in temperature at different positions in annealing equipment can cause systematic variations among wafers [106]. Less than perfectly flat surface in Chemical Mechanical Polishing (CMP) equipment or regular fluctuations in stepper can cause systematic variations between different dies on a wafer [33]. Lithography defocus, lens imperfections, layout topology, and stepper induced illumination & imaging non-uniformity due to lens aberrations [30] can cause systematic variations between different transistors or blocks on the same die.
These variations can be divided into Inter-Die and Intra-Die. Inter-Die systematic variations are principally caused by manufacturing fluctuations. The correlation factor in Intra-Die systematic variations is highly dependent on physical separation between the devices and reduces with increasing distance. Intra-Die systematic variability can be differentiated into two types– spatial variability dependent on the location on the die and proximity based variability dependent on neighboring structures [79], [82]. Spatial systematic variations occur mostly due to equipment imperfections whereas proximity based systematic variations are mostly caused by lack of fidelity in reproducing mask patterns [107]. The correlation length can be in range of 1-3mm with high correlation below 1mm [43]. Due to threshold voltage correlation, it is the dominant factor for variations but have almost negligible effect on average value [42].
Systematic variability can change the critical path order based on location of block on the die. Intra-die spatial variability affects mean of die performance [79]. Impact of systematic variability can be reduced using variation modeling as well as through restrictive design rules and Resolution Enhancement Techniques (RET) in manufacturing. Modeling systematic variations can be an expensive task, both in terms of effort and in terms of resources, because of which some systematic variations can be bundled together with random variations. Regular design has reduced the impact of systematic variability largely [70].

2.1.2.2.2 Random

Random variations have no correlation between any two devices irrespective of the separating distance between them. As in systematic, they can be Inter-Die or Intra-Die. Inter-Die random variations affect all the transistors on a die in the same manner but differ from one die to another. Most Inter-Die variations fall in this category, like variations of critical dimensions due to lithography, average doping

variations, Inter Layer Dielectric (ILD) thickness variation due to deposition and planarization process, etc. These variations arise due to lack of accurate process. Limiting case corners can be used to take into account the impact of random Inter-die variations.

In contrast, Intra-die random variations, also known as local random mismatch, arise due to natural limits to materials and current manufacturing technology. It creates a difference in electrical characteristics of devices with identical geometry, layout, and neighborhood within the interaction distance of known sources of variation [107]. On such miniscule scale, factors like roughness of length and width, atomic changes in oxide thickness, locations of dopant atoms inside a transistor and difference in granular structure of poly gate affect the electrical properties [58]. Interface charge non-uniformities [67], Interface roughness, Random Dopant Fluctuations (RDF), Line Edge Roughness (LER) [43], Oxide Thickness variation (OTV) [106], and polycrystalline granular structure of the polysilicon [11] are some of the major constituents of intra-die random variations [75]. As the effect of such variations is not consistent over all transistors on a die, corner approach is not a viable solution. It can be overly pessimistic and computationally expensive to verify all combinations. As of now, On Chip Variation (OCV) margins are used to quantify the impact of local random mismatch [9]. Statistical Static Timing Analysis (SSTA) can be used to evaluate and bound the impact of local random mismatch [52]. Normally, any variation that cannot be modeled or is too difficult to model is treated as a random variation.

Improvements and changes in design practices can and are ameliorating systematic variations [70]. However, random variations require big changes in manufacturing techniques like using 13nm lithography, or new structures like Silicon-On-Insulator (SOI) [106] or modification in fabrication process like using carbon with halo and MDD implants [103]. Although it may be the only solution in long-term, it requires a paradigm shift and is costly to implement in short-term.

2.2 Manufacturing steps causing variations

Variations arise from fluctuations in various steps of fabrication process, major ones being deposition, etching, sub-wavelength lithography, and CMP [86]. This section lists most of these steps and their impact.

2.2.1 Photolithography

Photolithography refers to the process of using light to transfer a pattern on a photomask to a silicon wafer coated with light sensitive material (photoresist). Further chemical treatment carves the shape onto the silicon. Lithography is one of the main factors behind dimensional variations. Currently the wavelength of light being used for lithography (193nm) is much higher than the physical dimensions being drawn on the silicon (40-65 nm). Even with RETs, there is bound to be issues related to diffraction and accuracy. On top of that, stress, layout, and proximity related effects also play a part in lithography [15], making a bad situation worse. Lithography is present in all stages of semiconductor manufacturing, but gate length

and metal width are the main parameters being affected. Variations because of lithography are mostly inter-die or intra-die systematic but small random portion is also present.

2.2.2 Etching

Etching goes in-step with photolithography where it is used to remove layers from the die (wafer) to create desired shapes. Depending on the type of etching and etched material, variation magnitude and degree of correlation also changes. Etching contributes to gate and metal thickness variations as well as skewed length/width dimensions creating trapezoidal shapes [106]. Different types of etching include wet or chemical etching and plasma etching. They have different degree of compromise between selectivity and being anisotropic. Etching can over engrave or damage the surface requiring re-crystallization. Etching together with photolithography constitutes the source of most variations. Process change also changes the etching impact as in copper and aluminum interconnects. Copper is an oxide etch process while aluminum process is metal etch.

2.2.3 Doping

Doping is required to create the required amount of charge carriers in different regions. Doping is done either through thermal diffusion or through Ion-implantation depending on the step. Thermal diffusion is more isotropic and less damaging to the structure but less accurate whereas Ion-implantation is more accurate but damages the surface. Some small statistical variations still exists in doping that can create variations between different devices. Random variations in dopant locations inside the transistor can also create variations. In addition, any trapped ions in dielectric can change the permittivity affecting electrical parameters.

2.2.4 Deposition

Each new layer or intermediate layers required for the construction of the transistor are deposited by one process or other. It can be a general deposition like chemical vapor deposition (CVD) or highly accurate process like sputtering. However, as in doping both have their drawbacks. The energy levels in sputtering used for gate material deposition vary a lot thus introducing variations. The process is mostly inter-die but random components at the nano-scale level are becoming important.

2.2.5 Chemical Mechanical Polishing (CMP)

CMP is the process to flatten and smooth out the deposited layers. It is necessary for symmetric and proper functionality of device. It also reduces the layers to desired thickness. However, the process has its limitations at nano-scale and cannot remove the inherent roughness present in a surface that introduces random variations. Damascene process mostly used for CMP can manage local uniformity but cannot guarantee global uniformity [54]. Shallow Trench Isolation and Inter Layer

Dielectric cause differences in CMP also [67]. In addition, systematic variation is introduced in the wafer as well as in the die due to density differences. The main impact of CMP is on interconnect parasitics. However, recent publications have shown that CMP impact is minimal on the parasitics specially capacitance [28].

2.2.6 Annealing, Oxidation, Resist development

Annealing is performed to crystallize any damage surface as well as for the diffusion of dopant atoms. The process improves device properties. However, Rapid Thermal Anneal (RTA) generates proximity effects creating systematic variations [67].
Oxidation is used to create the silicon dioxide layers on the wafers. The thermal process used can introduce wafer-to-wafer or die-to-die variations because of temperature gradient.
Photoresist is deposited onto the wafer using spin coating process. Small spatial variations are possible due to equipment fluctuations.

2.3 Design Parameters at Different Levels of Abstraction

The meaning of variations changes with perspective. For e.g. an engineer working in fabrication considers gate length, one in technology considers drain current, one in library considers gate delay, one in design flow considers path delay and one in product group considers die frequency. Whereas each of them may be talking about the same variation, either its source or its effect, the parameter is different. In this section, major types of variation have been listed based on different perspectives or level of abstraction.

2.3.1 Manufacturing level

Manufacturing level parameters are physical quantities describing different structures in digital design like transistor or interconnect. It is the most basic level in semiconductor manufacturing where we can see the impact of manufacturing fluctuations. The following section will detail most of the physical or derived parameters and their main sources of variations.

2.3.1.1 Poly gate length and width

Gate length or width variation arises from lithography and etching fluctuations. Gate width variation has little impact in general due to large nominal value. However, in cases where small width transistors are used, narrow width effects may not be negligible. The variations are absolute with respect to the minimum drawn length i.e. depends on the lithography technology and not on drawn dimensions. Variations are typically inter-die in nature arising from lithography fluctuations. However, at nano-scale dimensions, roughness and proximity effects increase intra-die variations [79]. Systematic gate length variations have a large impact on the global clock

skew. Resolution Enhancement Techniques (RET) and regular design can reduce impact of proximity effects [15]. Random gate dimension variations due to atomic roughness can increase during upcoming nodes. Variations in gate length affect drive current, threshold voltage, Drain Induced Barrier Lowering effect and load capacitance to previous gate. PMOS devices are more sensitive to channel length variations due to higher Short Channel Effects that causes steeper V_{th} roll-off [73].

2.3.1.2 Gate thickness and composition

Transistor gate stack is one of the most complex structures in semiconductor fabrication. It consists of layers of multiple materials and compositions to achieve the desired characteristics. Gate oxide thickness for sub 100nm technologies is equivalent to few atomic layers. Multiple steps involving deposition, lithography and etching are required to attain the required structure. Gate thickness is a highly controlled parameter due to the high sensitivity of threshold voltage to even small changes in gate thickness. Atomic layer deposition process is used to control the gate stack accurately. Compared to the variations in critical dimensions, those in physical gate thickness are small. Atomistic fluctuations in gate thickness can become important in sub-30nm nodes.

2.3.1.3 Doping and implants

Doping is required to create wells, channel, gate, drain and source junctions, etc. Depending on degree of accuracy required, the process can be different. Channel doping has the most impact on transistor characteristics and any variation in average doping level can create large variations in the current and threshold voltage. Average doping variation is mostly inter-die phenomenon. However, nano-scale structures make doping a discreet phenomenon creating variations due to difference in atomic locations.

2.3.1.4 Mobility

Mobility is a derived physical parameter and defines the ability of charge carrier to move in the presence of an electric field. It is a strong function of material impurities & temperature and directly affects saturation velocity. Doping and transistor area differences will affect mobility. It is not a primary parameter for intra-die variations, but it does get affected by discreet dopant profiles.

2.3.1.5 Gate oxide capacitance

Gate oxide capacitance is a derived physical parameter and determines the capacitance seen by previous gate thus affecting signal transition rates. It is affected by fluctuations in gate area, oxide thickness, and material permittivity. Intra-die effects like line and gate edge roughness affects the capacitance.

2.3.1.6 Metal Interconnects

Interconnect cross-section consists of two parameters having different variation sources: metal width and thickness (or height). Metal width variations arise from etching and lithography variations whereas metal thickness variations arise from CMP and Inter Layer Dielectric (ILD) fluctuations. Line width has a non-liner and large impact on sheet resistance due to increased electron scattering on grain boundaries and interfaces in copper [86]. It has a large intra-die systematic part because of proximity effects and through etching creating difference in desired and obtained shapes. X-Y plane variations due to lithography and etch along with Z-plane variations due to tall and narrow shape create significant RC variability [86].
Metal pitch is the distance between two metal lines. It affects other parameters like metal width and coupling capacitance. Coupled with metal length, it is one of the principal parameters in proximity effects. Printed pitch is affected by length and width of the lines around the area [89].

2.3.1.7 Inter-Layer Dielectric

ILD thickness and permittivity variations affect metal thickness, current and coupling capacitance in metal lines. ILD thickness variations are caused by etching and CMP including erosion and dishing effects [28]. Permittivity variations are caused by non-uniform deposition process and impurities. Variations in dielectric permittivity create zones around metal where the current density changes affecting long-term electromigration and related effects. Strong coupling in such regions can also affect signal integrity.

2.3.1.8 Via width and liner

The thickness of liner used to protect via is in few nanometers and susceptible to variations because of deposition process imperfections. This thickness affects resistivity and via current. The magnitude of variations in liner thickness is comparatively large and is difficult to model.

2.3.1.9 Contact width

Contacts are by far one of the biggest challenges to reduce gate area. Contact size is very large as compared to transistor to keep a low resistance. Reducing dimensions put more pressure on contact dimensions leading to small margins. Contacts are prone to CMP effects.

2.3.1.10 Stress variation

Stress in transistors is both desirable and undesirable based on the parameter it is influencing [107]. It is deliberately introduced to improve the performance of transistors in microprocessors but it hampers function in large dimensional ratio

metal lines, gate stack, and source & drain junctions etc. Proximity stress is caused due to overlayers, PMOS epitaxy, and STI [67]. It can cause variations in temperature and electromigration behavior of affected regions as well as alter transistor characteristics.

2.3.1.11 Well proximity effect (WPE)

Devices near Well mask edge are affected by ion scattering from photoresist. These ions get implanted in silicon surface affecting threshold voltage, mobility, and body effect. This phenomenon, called Well Proximity Effect, is layout dependent, and can create inter-die as well as intra-die effects.

2.3.1.12 Random Dopant Fluctuations (RDF)

RDF is an atomistic phenomenon affecting transistors thus causing intra-die variations. It is the fluctuation in number and location of dopant atoms in each transistor in a die [46]. In sub-100nm technologies, the number of dopant atoms lies in the range of few hundred [42], [3]. Thus, the doping structure is discreet rather than continuous and affects transistor characteristics respectively. Discreet doping is a natural limitation to transistor scaling and manufacturing improvements will have nominal effect on improving the situation. It causes each transistor to behave like a combination of small transistors. RDF comprises 60-65% of V_{th} variations below 65nm [67], [39]. It can lead to an increase in V_{th} variations and shift the average I_D-V_G curve towards negative gate voltage axis [46]. There is a secondary effect of degradation and fluctuations in subthreshold slope. The impact of RDF is predominantly in subthreshold region and can cause inhomogeneous potential in channel allowing for early turn-on in parts and impact DIBL due to doping profile fluctuations [12]. Dopant atoms adjacent to active region of device influence the fluctuations most [2]. Latch based design has been proposed to reduce impact of local mismatch [29]. Magnitude of threshold voltage fluctuations due to RDF can be given by equation (2-1) (symbols explained in list of symbols). Most of the parameters given in the equation are constant for a technology. Only the effective transistor area is exploitable by designers to reduce the impact of RDF. However, bigger transistors mean more silicon increasing cost per die as well as larger dynamic power consumption.

$$\sigma_{V_{Th}} = \left(\frac{\sqrt[4]{4q^3 \varepsilon_{Si} \phi_B}}{2} \right) \cdot \left(\frac{t_{ox}}{\varepsilon_{ox}} \right) \cdot \left(\frac{N_A^{0.4}}{\sqrt{W_{eff} L_{eff}}} \right) \quad (2\text{-}1)$$

2.3.1.13 Line Edge Roughness (LER)

LER refers to the fact that in nano-scale domain, two transistors having same area and dimensions can still have different perimeter. It can be defined as atomistic roughness in the gate edges differing two similar looking transistors. LER arises

from statistical variation in incident photon count during lithography exposure, absorption rate, chemical reactivity, and photoresist molecular composition [106], [121]. Fringe electric field and charge confinement near the interface play dominant roles in determining the impact of LER [83] and is more important for devices reaching Punchthrough. LER causes fluctuations in threshold voltage and transistor capacitance given by equation (2-2) and equation (2-3) [83]. It can also increase the leakage current in shorter gate length devices [76]. RDF typically affects I_{DS} whereas LER affects V_{th} [43]. Simulations have predicted the LER impact to overtake RDF in nodes smaller than 32nm [126], [12], [56], [90], [114]. As seen in the equation, threshold and capacitance variations can be decreased by increasing the gate dimensions directly affecting the variability as well as by reducing perimeter standard deviation.

$$\sigma_{V_{th}} = \left[\frac{Q_{dep}}{C_{ox}^{2}} t_{ox} \frac{\varepsilon_{ox}}{l_D} \right] \sigma_P \tag{2-2}$$

$$\sigma_{C_{tot}} = \left\{ \left[\frac{\left(A \frac{\varepsilon_{Si}}{w} + w \frac{P}{4} \frac{\varepsilon_{Si}}{l_D} \right)}{A \frac{\varepsilon_{ox}}{t_{ox}}} + 1 \right]^{-2} w \frac{\varepsilon_{Si}}{l_D} + t_{ox} \frac{\varepsilon_{ox}}{l_D} \right\} \sigma_P \tag{2-3}$$

2.3.1.14 Oxide thickness variations (OTV)

Oxide thickness variations refer to molecular variations in gate surface. This atomistic roughness causes differences in two identical looking transistors affecting their characteristics. OTV is caused by statistical variations in deposition and planarization process [4]. The impact of OTV is limited until 32nm node but is expected to become significant below 22nm where it can cause variations of 1-2 atomic spacings in a gate that is just a few atoms thick [5]. OTV mainly affects oxide-tunneling current and causes mobility degradation at elevated transverse fields. Thickness variations can cause variation in potential across mosfet channel scattering carriers and decreasing mobility at high lateral electric fields [106].

2.3.1.15 Polysilicon Granularity

Polysilicon material used to create the transistor gates have a poly-crystalline granular structure that supports enhanced diffusion along the grain boundaries [11], [45]. This creates non-uniformity in poly-Si gate doping and localized penetration of dopants through gate oxide into channel creating potential barriers [98]. It induces threshold voltage variations as well as an increase in average threshold voltage [41]. The phenomenon behind magnitude of variations due to polysilicon granularity

between NMOS and PMOS also differs [1]. It is expected to be noticeable below 30nm and depending on gate configuration, might have a larger effect than LER [11]. Amorphous silicon gate [41] and uniformly structured metal gate [11] have been proposed as a way to negate the impact of polysilicon granularity.

2.3.2 Transistor level

Transistor level parameters are electrical parameters useful for technology design groups that are required to meet these specifications in manufacturing. Initially, transistor level parameters are decided for a given application flavor that are feasible with the current manufacturing technology. Anticipatory models are created based on the parameters that allow designers to start using the technology. Meanwhile, detailed manufacturing recipe is created to achieve the desired parameters. The recipe is tweaked regularly during process ramp-up to achieve higher yield and lower costs. Transistor models are also updated to be in line with process and become stable once the technology has achieved maturity. Transistor level parameters are kept roughly constant during ramp-up to enable designers up the chain to work without worrying about affect of process change on their design.

2.3.2.1 Saturated Drain Current

Saturated drain current, represented by (transistor ON current or I_{ON} or $I_{ds,sat}$ shown in Figure 2-1, is a measure of how much load a transistor can drive for a given slew. It also affects the transistor delay, i.e. the delay between input and output signal switching, to an extent. Larger the I_{ON} higher the load capacity (or faster the switching speed for a given load). Theoretically, supply current should decrease by the same factor as the gate length with technology scaling according to the constant field scaling [128]. However, in nanometer era, I_{ON} is mostly kept constant or it may increase with scaling, as supply and threshold voltage have remained almost constant to limit leakage power consumption while maintaining the drive capability. Saturated drain current can be approximately expressed using unified model mathematical equations (2-4) to (2-8). These equations provide a simplified view of the parameters affecting drain current and may not represent the magnitude in advanced technologies.

$$I_{ds,sat} = \mu_{eff} C_{ox} \left(\frac{W_{eff}}{L_{eff}} \right) \cdot \left[\left(V_{gs} - V_{th} \right) - \left(\frac{1}{2} \right) A_b V_{ds} \right] \cdot V_{ds} \tag{2-4}$$

$$V_{th} = V_{fb} + \phi_{S0} + \gamma \sqrt{\phi_{S0} - V_{bs}} \tag{2-5}$$

$$A_b = 1 + \frac{\gamma}{\left(2\sqrt{\varphi_{S0} - V_{bs}} \right)} \tag{2-6}$$

$$\gamma = \left(2 q \varepsilon_{Si} N_{ch} \right)^{1/2} / C_{ox} \tag{2-7}$$

$$\phi_{S0} = 2\phi_F = 2(kT/q) \ln\left(N_{ch} / n_i \right) \tag{2-8}$$

Complex doping profiles and nanoscale phenomena present in nanometer scale transistors make these equations at best just approximations. These equations lack the ability to model atomistic issues, quantum effects, etc. They are useful to determine the relationship between different parameters.

Drain current variability is principally dependent on threshold voltage, mobility, transistor dimensions, and supply voltage. It has a 2^{nd} order dependence on factors such as saturation velocity, leakage current, temperature, parasitics, etc. I_{ON} has a direct impact on switching speed and thus its variations are highly correlated to frequency variations [43]. I_{ON} variations can cause timing failures, excessive heating or high leakage, all of which affect the yield.

Before reaching the saturated value, drain current passes through the subthreshold and linear region. In subthreshold region, it is exponentially dependent on applied gate voltage and linearly proportional in linear region. In subthreshold region, the DIBL effect is very small due to small operating voltages and thus the threshold voltage has small dependence on gate length variation. Local dopant fluctuations are dominant in this region [29].

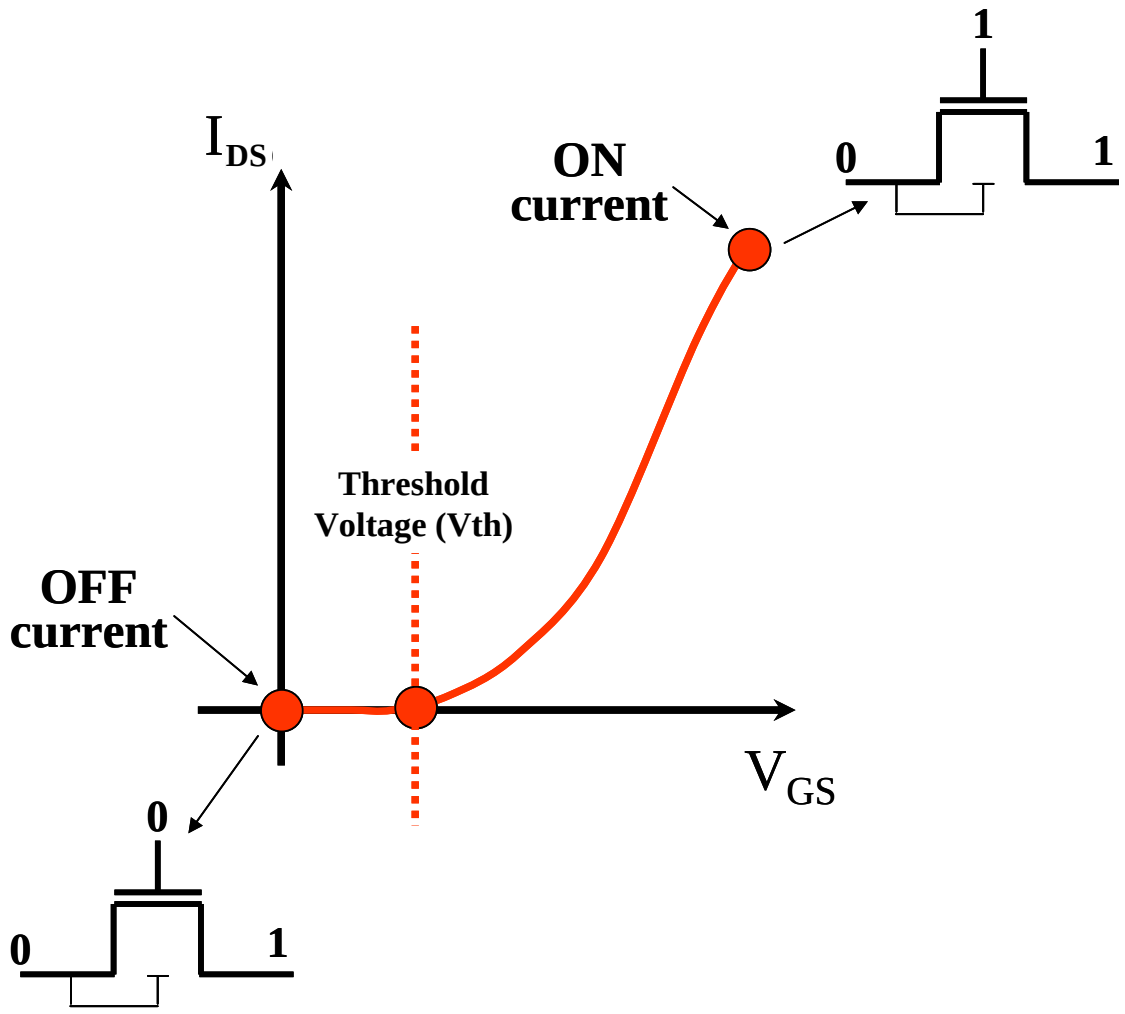

Figure 2-1: Drain current vs. gate voltage

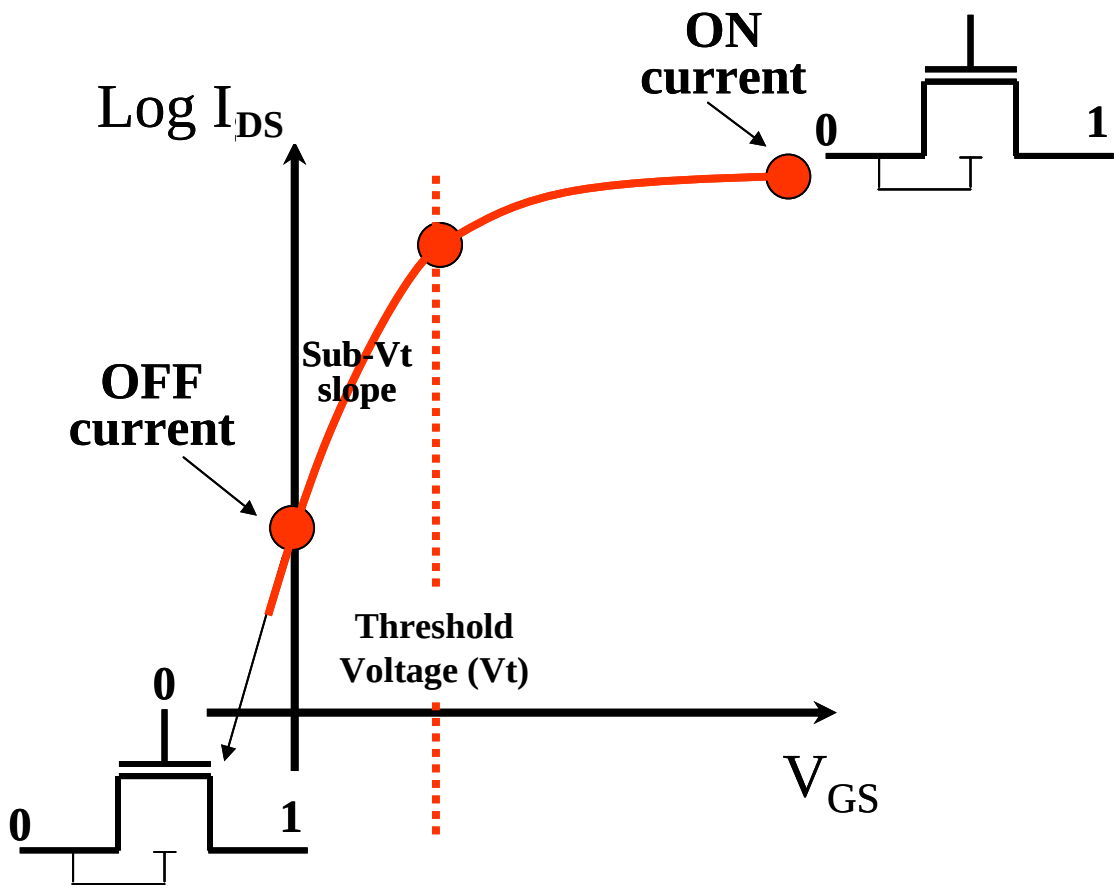

Figure 2-2: Log of drain current vs. gate voltage

2.3.2.2 Threshold voltage

Threshold voltage, as shown in Figure 2-2, is defined as the amount of gate voltage at which the transistor switches ON by the onset of inversion layer. At threshold voltage, instantaneous current passing through source to drain becomes larger than a predefined limit and with further increase in gate voltage increases exponentially. Threshold voltage is a dynamic quantity depending on physical, electrical, and environmental factors. It can be divided into two factors, a quantity defined as V_{th0} that is supposed to be constant for a given technology, flavor, & device and ΔV_{th} that is dependent on operational parameters. The relationship between V_{th} and different parameters can be specified through equations (2-9) to (2-13).

$$V_{th} = V_{th0} + \Delta V_{th} \tag{2-9}$$

$$V_{th0} = 2\phi_B + \frac{\lambda_b \sqrt{2\varepsilon_{Si} qN_A (2\phi_B + V_{SB})}}{C_{Ox}} + V_{fb} - \lambda_d V_{DS} \tag{2-10}$$

$$\phi_B = \frac{kT}{q} \ln\left(\frac{N_A}{n_i}\right) \tag{2-11}$$

$$V_{fb} = \phi_{ms} - \frac{Q_{fc}}{C_{Ox}} \tag{2-12}$$

$$\lambda_b = 1 - \left(\sqrt{1 + \frac{2W}{X_j}} - 1 \right) \frac{X_j}{L} \tag{2-13}$$

Transistor gates are made up of multiple layers of various materials and thickness to obtain the desired characteristics. These equations use representative values to obtain an approximate behavior while the exact characteristics can only be derived using physical simulators with TCAD. The given equations do help us to separate different sources of variations affecting threshold voltage. Variations in V_{th0} are dependent on the physical parameters like oxide thickness, doping concentration, doping profile and dielectric. Whereas variations in ΔV_{th} consists of Drain Induced Barrier Lowering, short channel effects, reverse short channel effects, narrow width effect, back bias dependent threshold shift, hot carrier effect, and mobility degradation impact on threshold voltage due to temperature and voltage variations.
V_{th0} variations by construction consist mostly of inter-die variations. However, nano-scale features induce systematic and random intra-die variations. A highly non–linear correlation exists between V_{th} and channel length variability [127]. Impact of physical or environmental variations on different phenomena is modeled mathematically introducing accuracy errors.
Newer models like PSP use physical modeling rather than mathematical modeling. PSP has eliminated threshold voltage and uses surface potential instead, allowing a higher level of accuracy and reduced discontinuities.

2.3.2.3 Leakage current

The quality of a transistor is defined by I_{ON} as well as the I_{ON}/I_{OFF} ratio where I_{OFF} is defined as the transistor leakage current or the amount of current passing through the transistor when it is supposed to be in switched-off mode. Leakage currents are a byproduct of CMOS technology imperfections. In theory, the ‘off’ state or zero in digital systems is considered to have zero current. However, in practice, a small amount of current still pass through the transistor reducing efficiency, consuming power, and creating reliability issues. With each new technology, more types of leakage currents are becoming significant. Various types of mechanisms [6], [37] aiding leakage are listed below.

2.3.2.3.1 Reverse bias p-n junction

When a p-n junction is in reverse bias, the majority charge carriers are pulled away from the junction causing the width of the depletion zone to increase. The large voltage barrier causes a resistance to flow of majority charge carriers thus limiting the current across p-n junction. However, the minority charge carriers have a favorable condition to flow across the junction and the related current constitutes the reverse bias leakage.

2.3.2.3.2 Minority carrier diffusion

A transistor can have a drift current that happens because of applied electric field and a diffusion current that happens because of difference in doping levels. Majority charge carriers move from high concentration regions towards low concentration regions where they are become minority carriers. Diffusion current does not need an electric field. Minority carriers see a potential slide when moving across p-n junction facilitating their diffusion.

2.3.2.3.3 Band to band tunneling

Band to band tunneling current happens in deeply depleted region induced between oxide and heavily doped junctions by an accumulating gate field and a lateral field (the drain-bulk bias).

2.3.2.3.4 Weak inversion or subthreshold leakage

In weak inversion mode, the channel surface potential is almost constant across the channel and the current flow is determined by diffusion of minority carriers due to a lateral concentration gradient. In short-channel devices, this current is strongly influenced by the channel length due to the drain-induced-barrier-lowering effect.
Subthreshold leakage is the dominant cause of worry for transistor variability. The relationship between subthreshold leakage and other parameters can be given through equations (2-14) to (2-17).

$$I_{subth} = \mu_{eff} C_{ox} \frac{W}{L} V_t^2 \times \exp\left(\frac{V_{GS} - V_{th}}{nV_t}\right)\left(1 - \exp\left(-\frac{V_{DS}}{V_t}\right)\right) \quad (2\text{-}14)$$

$$V_t = kT/q \quad (2\text{-}15)$$

$$I_{ds,leak} = \mu_{eff} C_{ox} \frac{W}{L}(n-1)V_t^2 e^{-\frac{V_{th}}{V_t}} \quad (2\text{-}16)$$

$$P_{leak} = V_{DD} I_{leak} \quad (2\text{-}17)$$

As can be seen, leakage current has a direct relation to supply voltage, temperature, mobility, oxide capacitance, and transistor dimensions. It has an exponential relationship to threshold voltage and thus is highly sensitive to V_{th} variations. A typical chip can have millions of transistors, which when added up consume a significant amount of leakage power. Leakage current variation is a lognormal distribution [107]. Intra-die leakage can create local errors. Architectural techniques can reduce leakage power but increases overall cost.

2.3.2.3.5 Gate induced drain leakage (GIDL)

GIDL is caused by high electric field in drain junction of MOS transistors under the gate/drain overlap region. In a NMOS with low gate voltage and high drain voltage, significant band bending in the drain allows electron-hole pair generation through

avalanche multiplication and band-to-band tunneling. A deep depletion condition is created since the holes are rapidly swept out to the substrate. At the same time, electrons are collected by the drain, resulting in GIDL current [36].

2.3.2.3.6 Punchthrough

Punch through is an extreme case of channel length modulation where the depletion layers around the drain and source regions merge into a single depletion region. The field underneath the gate then becomes strongly dependent on the drain-source voltage, as is the drain current. Punch through causes a rapidly increasing current with increasing drain-source voltage.

2.3.2.3.7 Oxide leakage

The gate oxide, which serves as insulator between the gate and channel, should be made as thin as possible to increase the channel conductivity and performance when the transistor is on and to reduce subthreshold leakage when the transistor is off. However, with current gate oxides (thickness around 1.2nm or ~5 atoms thick), the quantum mechanical phenomenon of electron tunneling occurs between the gate and channel.

2.3.2.3.8 Hot carrier injection

Hot carrier injection is a phenomenon when a charge gains sufficient kinetic energy to overcome a potential barrier between different areas of the device and migrates from one area to another. The kinetic energy is directly related to the matter temperature.

2.3.2.3.9 Drain induced barrier lowering (DIBL)

When a high drain voltage is applied to a short-channel device, it lowers the barrier for electrons between the source and the channel, resulting in further decrease of the threshold voltage. The source then injects carriers into the channel surface independent of gate voltage increasing leakage.

2.3.2.4 Subthreshold Slope

Subthreshold slope is defined as the amount of gate voltage required to increase the subthreshold leakage current by an order of magnitude. It can be represented as the slope of log I_D vs. V_{GS} curve in subthreshold region and given by equation (2-18). Theoretical values for S lie between 60-100mV/decade. Inter-die and random intra-die variations affects S that in turn creates variations in I_{subth}.

$$S_{subth} = \left(1 + \frac{C_D}{C_{OX}}\right) V_t \ln 10 \qquad (2\text{-}18)$$

2.3.2.5 Saturation velocity

At high electric fields easily achievable in nano scale devices, scattering effects cause the carrier velocity to saturate resulting in mobility degradation. It defines the limit after which a further increase in applied gate voltage will have a reduced impact on drain current. As electric field depends on effective channel length, any variations in channel voltage or electrical channel length will affect the point of reaching velocity saturation. Saturation velocity itself is a function of temperature and is thus affected by design activity. Overall, it is a secondary parameter when considering transistor variability.

2.3.3 Logic gate level

Logic-gate level parameters are useful for library designers who have to minimize cell area, power consumption, and delay while maximizing drive capacity in order to achieve the best possible tradeoff. Standard cells libraries are created for given technology flavors. Using transistor models and field solvers, gate level parasitics are extracted and included in spice models to improve timing accuracy. The libraries are then characterized for timing, power, and transition times at different corners to create timing libraries. These timing libraries form the building blocks of Static Timing Analysis. Timing characteristics of gates are kept almost constant during process ramp-up to have minimum effect on design timing. However, it may be required to update the library to be in line with the process.

2.3.3.1 Logic gate delay

Logic gate delay or switching time is the amount of time required for a change on input to reflect on output. It is dependent on load and input transition time for a given cell at a given corner. It is a strong function of drain current and thus reflects any variations in the current. Smaller the logic gate delay, faster is the design. However, a faster logic gate can result into higher leakage or higher power. Thus, it is a compromise between speed and power consumption linked with heat dissipation costs.

Inverter is the most basic logic gate and determines the technology speed. Delay of other logic gates can be represented in terms of inverter delay. The fanin effect is more important for NPOS than PMOS and affects the amount of variations in complex logic gates [66]. Inverter delay and its variations can be expressed using the equations (2-19) to (2-23).

$$T_d = \frac{CV_{DD}}{I} = \frac{K.V_{DD}.\left\{1+\ln\left[1+\exp\left(\frac{V_{DD}-V_{th}}{E_{sat}L}\right)\right]\right\}}{\left\{\ln\left[1+\exp\left(\frac{V_{DD}-V_{th}}{2S}\right)\right]\right\}^2} \quad (2\text{-}19)$$

$$\frac{\sigma_T}{T_d} = \sqrt{\left(\frac{\partial \ln T_d}{\partial L}\right)^2 \sigma_L{}^2 + \left(\frac{\partial \ln T_d}{\partial V_{th}}\right)^2 \sigma_{V_{th}}{}^2} \quad (2\text{-}20)$$

$$V_{th} = V_{th0} - V_{DD}.\exp\left(-a_{V_{th}}.L\right) \quad (2\text{-}21)$$

$$S = S_0.[1 + \exp\left(-a_S.L\right)] \quad (2\text{-}22)$$

$$K = \left(k_0 + k_1.L.C_{load} + k_2.L^{a_k}\right)/W \quad (2\text{-}23)$$

Logic gate delay is the principal parameter used for design timing analysis. Path delays in a design can be calculated by adding successive logic gate delays [120] computed from their timing libraries. Any variations in this quantity will directly affect design performance. For most part, gate delay variability can be divided into two mutually orthogonal parts – channel length variability and threshold voltage variability, as shown in equation (2-20). It takes into account the variation of two parameters as well as the degree of relationship between the logic gate delay and the respective parameters. Threshold voltage will include the effects of temporal variability and the intra-die variability while logic gate length will include the inter-die variability effects. Supply voltage and load will have a direct impact on logic gate delay while effects like saturation velocity and leakage are second order. Largest source for delay variability is effective channel length variability while it is most sensitive to threshold voltage variability, especially in subthreshold region [127].

2.3.3.2 Slew rate

Input slew rate is a primary factor to calculate effective cell delay. Output slew rate depends on the driving capacity of the cell and load capacitance primarily, and on input slew rate depending on number of stages. For a single stage cell, input slew rate will influence the propagation delay as well as the output slew rate. For multistage cells, load capacitance plays a major role in determining the output slew rate of each stage. Higher the drive capacity (or lower the output load), larger the logic gate output slew rate. Output slew rate is characterized in a similar manner as to logic gate delay for timing library construction. Slew rate also affects the dynamic power consumption through switching time as shown in equation (2-24).
Supply voltage, channel length and width, threshold voltage, and parasitics have a direct impact on slew rate variability. Rise time or fall time for a given cell are different and are affected by intra-cell mismatch variations [42].

2.3.3.3 Dynamic Power

Dynamic power, or switching power, is the amount of power required to charge up a given load. In principal, the load here represents the output load of the logic gate. However, it also includes some parasitics and wire loads. Dynamic power affects the maximum instantaneous power in the design and thus determines the required heat dissipation capacity. Frequency has a direct correlation with dynamic power, which in turn is impacted by logic gate delays. As such, the fastest logic gate delay

causes the largest power consumption. Any variation increasing the logic-gate power consumption can take the design in excess of its maximum rated heat dissipation capacity. Whereas there is little overall impact of intra-die random variations on dynamic power, systematic variations can affect local areas. Inter-die variations are bounded by corners. Dynamic Power can be specified using equation (2-24) where C is the capacitance, V is supply voltage, and f is switching frequency.

$$P_{Dynamic} = CV^2 f \quad (2\text{-}24)$$

2.3.3.4 Leakage Power

Leakage power is the amount of power dissipated when logic gates are supposed to be in stable state, i.e. no transition. Theoretically, when the logic gate is not in transition in CMOS logic, current is supposed to be zero. However, there is a small amount of current still flowing through the transistors called I_{leak}. Leakage current has been increasing with technology and is not a negligible factor these days. It directly affects battery efficiency of product as well as can cause reliability issues if not controlled properly. Leakage power can be specified by equation (2-25).

$$P_{Leakage} = V_{DD} * I_{Leak} \quad (2\text{-}25)$$

2.3.3.5 Logic gate parasitics

Logic gate parasitics are mainly of two types: capacitance and resistance. Parasitic capacitances are generated due to interaction of different conducting and non-conducting regions present in a cell as well as from interaction with neighboring cells and structures. Parasitic resistances are generated from the metal and poly lines connecting different transistors. Parasitics affect the switching time due to required charging/discharging of different capacitances as well as increased resistance to current. Parasitics have evolved with technology and have become an important factor in timing characteristics today. Complex solvers and extractors are used to extract the various types of parasitics present in a cell. These parasitics allow to back-annotate the spice netlist to attain better timing accuracy. As these parasitics are created by interaction of various physical structures in and around the cell, any variations in them affect the parasitics also.

2.3.3.5.1 Intrinsic and Extrinsic capacitance

For any cell or logic gate, the capacitance inside the cell that does not affect the load of the fanin logic gates or the input transition of the fanout gates is called intrinsic capacitance. The capacitance seen by the fanin gates, the capacitance of the first input transistors and that seen by the fanout transistors, the capacitance of the last output transistors is called the extrinsic capacitance. Both intrinsic and extrinsic capacitances have different roles in delay calculation. The intrinsic capacitance is more important to calculate the delay for the given cell and the extrinsic capacitance to calculate the delay for the neighboring cells as it forms part of loading (input or output) to that cell. Logic-gate capacitance variations depend on its dimensions and

dielectric. Inter-die variations and intra-die systematic variations are the main phenomena occurring in its variations. Variations in photolithography are the chief cause of capacitance variations.

2.3.3.5.2 Logic gate parasitic capacitance (e.g. miller)

Until recent technology nodes, the input to output capacitance of an inverter (C_{gd}) could be safely neglected as only one transistor at a time conducted through the load capacitance. However, with increasing threshold voltage to supply voltage ratio and large variations in the threshold voltage, the NMOS and PMOS transistors have overlap among their different operating regions. The impact is different for rise and fall transitions because of different sizes. As such, the impact of C_{gd} becomes significant and it cannot be neglected anymore. Furthermore, it is a highly non-liner capacitance and any variations in threshold voltage cause large variations in its impact. This input to output capacitance is called the miller capacitance and is made up of two parts, static overlap capacitance [40] and dynamic channel capacitance. Miller capacitance depends on a whole lot of factors like channel length, threshold voltage, supply voltage, operating conditions, etc.

2.3.4 Path level

Path level parameters are useful for design engineers who have to minimize skew, manage interconnect routing, meet target delays, reduce critical paths, etc. Timing analysis is done for different paths in a design using standard cell libraries as well as timing models at given conditions. Margins and derating [7] are used for different kinds of variations. Design speed is limited by critical path delay and thus reducing that delay is one of the prime tasks at this stage. Other constraint violations like setup & hold times are also checked using timing analysis and corrected. Input libraries and models are provided by technology and library groups. Input paths are provided by designers through design data at various stages of implementation (netlist, pre-layout, post-layout). Path level parameters enable designers to construct the whole design within allowed limits.

2.3.4.1 Path delay

Path delay (or path propagation delay) is the amount of time taken for a change in input to reflect on output of a path. It is one of the principal output parameters in timing analysis through which it is determined if the design meets given timing constraints. It also helps to determine the maximum operable speed. It can be calculated either using logic gate delay, input slew rate and load capacitance (timing analysis) or measured using spice simulations. Time to market constraints makes it necessary to use timing analysis for timing closure. Because of its dependency on load capacitance, parasitics, input slew rate and logic gate delay, any variations affecting these parameters will in turn affect path delay also. Path delay distribution depends on output transition and delay of previous logic gates [42].

Timing analysis algorithms are highly complex and mostly proprietary. One of the general algorithms that give an insight into path delay calculations is based on the principal of Logical Effort (Sutherland, Sproull, and Harris) [53]. It is useful for custom designers to create a path with minimal achievable delay. For a path with same type of logic gates (but not the same size), the minimal path delay is calculated using equations (2-26) to (2-34).

$$D = \tilde{N} F^{1/\tilde{N}} + P \quad (2\text{-}26)$$

$$d = g.h + p \quad (2\text{-}27)$$

$$h = C_{out} / C_{in} \quad (2\text{-}28)$$

$$F = G.B.H = \prod f_i \quad (2\text{-}29)$$

$$H = C_{out-path} / C_{in-path} \quad (2\text{-}30)$$

$$G = \prod g_i \quad (2\text{-}31)$$

$$B = \prod b_i \quad (2\text{-}32)$$

$$P = \sum p_i \quad (2\text{-}33)$$

$$\tilde{N} = \log_4 F \quad (2\text{-}34)$$

For a single stage logic gate, logical effort ‘g’ determines the capacity of the logic gate to produce output current based on its topology, electrical effort ‘h’ determines the load-driving capability, and parasitic delay ‘p’ determines the internal delay of the logic gate because of its intrinsic capacitance. ‘d’ is the single gate delay, ‘b’ is the gate branch effort, ‘f’ is the stage effort. ‘G’ represents the path logical effort, ‘H’ represents the path electrical effort, ‘F’ represents the path effort, ‘D’ represents path delay, ‘P’ represents path parasitic delay, ‘N’ is the number of stages, and ‘B’ is the branching effort. ‘C_{out}’ and ‘C_{in}’ are the output and input capacitance of the path.

For a path made of these logic gates, the total delay is distributed among the path effort and path parasitic delay. The number of stages is determined using path effort with size of logic gates being the variable parameter. Even for a simple path, calculating the delay requires a complex set of equations. When multiple types of cells, different fanouts, branching, feedback loops, etc. are included, the task becomes exponentially complex. In general using Logical Effort, an optimum delay along a path can be achieved that is the basic requirement for any design.

2.3.4.2 Setup & Hold time

Setup time for a path is defined as the minimum amount of time that the data signal should arrive before the clock signal for correct latching. If the period is so small that the next clock signal after launch signal arrives at capture flop before data has reached, then there is a setup violation. Setup violations can be removed by increasing clock period.

Hold time is defined as the minimum amount of time for which the data signal should be stable after the clock signal arrives for correct latching. If the skew between two synchronous flops is so large that the same clock signal that launched

the data arrives at the capture flop after the data from that clock has arrived then there is a hold violation. For a system, hold violations are a bigger issue as they require inserting more delay in data path by adding extra buffers thus changing the variation.

2.3.4.3 Skew

Clock skew is the difference in arrival times of clock signal at two data connected clocked components. Generally, designers target a zero value for skew as it increases system complexity. However, it cannot be avoided totally and some unintentional value might be present due to unbalanced configurations, intra-die variations, device delay scaling with environmental conditions etc [109]. Intentional skew can also be present in particular paths to gain from time borrowing and reduce clock period. Unbalanced loads in clock increase global variability effect [110]. Due to systematic and random mismatch, effect of environmental variations is different that in turn impacts skew.

2.3.4.4 Wire (interconnect) delay

In sub-100nm, interconnect delay has gained as much importance as logic gate delay. In a design, it constitutes 30-50% of the total delay with wire length going in hundreds of kilometers. Whereas local interconnect delay is still less compared to logic delay, global delay is comparatively much higher because of long lines.
Wire delay consists of resistance and capacitance delay, contact and via delay. Sensitivity of interconnects to variations is less than transistors [43]. Dense interconnect performance variance depends on a lot of factors including CD bias, metal thickness, sheet resistance, Low K permittivity between, above and below metal lines, low K thickness above, below and in between metal lines and via resistance [57]. Any variations in metal cross-section, dielectric permittivity, and coupling capacitance will induce variations in wire delay. For high performance interconnects, inductance also plays a role and is affected by the neighboring wires as well as by the wire structure and length. Principally the inter-die variations and temporal variations are important for wire delays but intra-die systematic variations are also becoming important [43] because of large wire lengths necessitating use of statistical models. Random variations are comparatively small. Main factors present in wire delay are given below.

2.3.4.4.1 Wire resistance

Wire cross-section and metal resistivity are the main factors to calculate the resistance. The cross-section is affected by etching/deposition/planarization process and the resistivity by temperature and material imperfections created by ions and induced strain.

2.3.4.4.2 Wire capacitance

Dielectric constant and thickness, and metal width are the main factors affecting wire capacitance. Etching/deposition/planarization process is responsible for the variations in dielectric thickness and damaged dielectric and temperature affect the dielectric constant.

2.3.4.4.3 Contact and via resistance

Via resistance is affected by the metal cross-section and the thin film width at the base of via. Contact resistance is mainly affected by its cross-section and the thickness. Etching/deposition/planarization processes are again responsible for any variations seen here.

2.3.4.4.4 Parasitic capacitance (e.g. coupling, fringe)

One of the main factors in calculating wire delay is impact of parasitics. Calculating delay for an isolated wire is an easy task but when you consider impact of neighboring wires, it becomes complex. Parameters like temperature, activity, distance between two wires and the dielectric between two wires impact coupling capacitance. With decreasing scales, impact of fringe capacitance becomes important too. Environmental and temporal effects with lithography are the major constituents. Parasitic extractors normally do not consider the trapezoidal conductor cross section present in realistic designs created by etching, and are thus constitutes a large factor in parasitic inaccuracy by increasing total capacitance [28].
There are many ways to reduce coupling capacitance [23], some of which are listed below.

2.3.4.4.4.1 Shielding

It makes horizontal wire capacitance independent of adjacent wires switching and provides noise immunity and signal integrity. However, it is area expensive and its effectiveness is reduced by dummy fills

2.3.4.4.4.2 Wire spacing

It reduces delay and energy by minimizing coupling capacitance. It is also area expensive

2.3.4.4.4.3 Swizzling

It reduces worst-case delay by realizing all possible adjacencies within a swizzle group. It requires extra routing and vias increasing the total delay and is not effective in controlling variations.

2.3.4.4.4.4 Skew signals on alternate lines

It avoids same time switching between alternate lines. Effective for long wires and relaxed clock frequency where ΔT overhead is small compared to delay reduction.

2.3.4.4.4.5 Repeater staggering

It offset inverters on adjacent lines. Coupling capacitance at end of a segment is driven by more resistance than at beginning, and contributes more to total wire delay. Optimal repeater insertion point is 70% of the segment. Issues with unidirectional buses, layout constraints, larger delay variations.

2.3.4.4.5 Inductance

Most parasitic calculations do not include inductance as a basic component. However, with smaller scales, relatively large dimensions and high speed, inductance can play a role also. All factors affecting line resistance and capacitance affect inductance also. Only major issue is how to include it in the parasitic calculations effectively without affecting the computing speed and accuracy. Inductance increases clock skew, max timing and noise in bus signals. Affect of inductance due to process variations ranges from 6% to 13% [123].

2.3.4.4.6 Conclusions on Interconnects

There is a high amount of correlation found between parameters for a given interconnect line e.g. line resistance and thickness [34]. Such correlations can cause an over-estimation of the impact of variations on interconnect delay if not taken into account. Some of these correlations arise from systematic effects. Metal thickness is a function of density and width (high dishing for wider lines and erosion for higher metal pattern density [28]). Thus, any variations in width will also affect thickness during the process that in turn causes lithography defocus. In addition, metal resistivity has a dependence on line width due to surface scattering. On the other hand, random variation alters the geometry or material properties of interconnect causing variations in electrical resistance, capacitance, signal delay and 1/f noise. One important affect seen in recent works is averaging effect of variations if metal interconnects are broken into several layers because of larger number of independent parameters [34], [57].
Pattern density has a substantial impact on the interconnect characteristics [91]. Density information can also be grouped with metal layers [92]. Standard cell routing is mostly done in first and second metal layers and has high density due to small cell area with a very narrow PDF due to very compact cell designs. Random logic routing uses layers on level three and four mostly, which have lower density with a narrow PDF also as it uses automatic routing tools and is restricted to small areas in blocks connecting different logic gates. Global inter connects and power distribution use the rest of metal layers and have a higher density than logic routing

but wider PDF also due to the fact that it has to cover almost the whole chip and connect all the cells.
Metal density affects resistance and capacitance in opposite ways. As density increases, resistance increases and capacitance decreases. However, on the overall resistance is more sensitive to density. Wire delay can also be differentiated based on the interconnect type. Short wire delay decreases with increasing density as it depends more on resistance while long wire delay increases as it depends more on capacitance. The sensitivity of longer wires is also higher [91]. Dummy metal fills used to achieve density uniformity improves uniformity but increases coupling capacitance to a high degree also [28]. Dense fills cause a higher variation in capacitance and depends on fill patterns, minimum inter-fill spacing, and minimum conductor to fill spacing.

2.3.5 Circuit level

Circuit level parameters are useful for product design, validation, and test engineers who need to verify design functionality and evaluate if performance targets are being met. Design timing analysis along with power and reliability analysis forms the core of testing and design validation. Fluctuations in fabrication will induce variations in circuit parameters. However, the goal is to qualify if the product lies within tolerance limits.

2.3.5.1 Design specifications

2.3.5.1.1 Clock Frequency

Design frequency typically represents the overall functional clock frequency visible to the external connections. A design can have more than one clock but most of them are internal to the design. A product engineer might have to validate all internal clocks and individual block functionality, while a test engineer will be concerned with overall design frequency. Variations in gate delay, supply voltage, temperature, parasitic delay, etc will affect the clock frequency. Any variations in clock frequency have to be within pre-defined tolerance limits so that it would not affect design functionality with any external connection. Internal clock frequency depends on variation margins, clock skew, setup and hold time constraints, insertion delay, etc.

2.3.5.1.2 Power consumption

Die power consumption determines the maximum power that needs to be dissipated thus determining the heat sink capacity. Heat sinks are costly and add a lot to the cost per die. Moreover, there are threshold limits to the amount of power a particular heat sink is rated for after which another one with higher rating will have to be used. If thermal power is not dissipated properly, it can affect product reliability. Larger dynamic power at transistor/gate level can adversely affect the design power

consumption. Each design is itself made up of multiple blocks having different power consumption levels. Designers verify the thermal effect in corresponding area so that it remains within the acceptable levels.

2.3.5.1.3 Leakage power

Many products like mobile phones remain in standby state for most of the time and have to be in active state for a relatively small duration of time. Leakage power is a major consideration for them as it determines battery lifetime. Even in off or non-transitioning state, cmos gates dissipate a small amount of power. Adding up over millions of gates, it becomes a significant quantity and drains power from battery. Variations in leakage current at transistor/gate level will affect the chip leakage power and thus affects the battery life.

2.3.5.2 Internal Parameters

2.3.5.2.1 Signal Integrity

With large amount of decoupling capacitance, signal integrity is a major concern for any design. Some principal components of signal integrity are given below.

2.3.5.2.1.1 Crosstalk

Coupling between neighboring wires generating glitches in signals of one wire due to transitions in neighboring signals is called crosstalk. The two wires act as a parallel plate capacitor. Larger the coupling capacitance, larger is the crosstalk. Length of wires running parallel and their width determines amount of crosstalk. A switch in neighboring lines can flip the value latched in a flip-flop or two neighboring wires switching in opposite directions can effectively double the coupling capacitance delaying the signal [93]. With signals going in gigabit rates, protecting signal integrity is of utmost importance. Parameter variations increase logic gate vulnerability to crosstalk by decreasing its ability to recover from charge collection due to particle strikes and by increasing its ability to propagate transient pulse un-attenuated [77]. Metal thickness, width and intra-layer dielectric variations, as well as damaged dielectric regions on side of metal lines are important factors to determine crosstalk contribution [34].

2.3.5.2.1.2 Noise (Substrate, Thermal, Flicker, Shot)

Substrate noise is generated because of coupling between different regions through substrate. For mixed signal designs with both analog and digital areas, it is an important issue. The large amount of noise in more robust digital signals can affect the sensitive analog signals via substrate if they are not properly shielded.

Thermal noise is generated because of thermal agitation of charge carriers in the material. With very high charge densities present in current nodes, it affects the performance of the designs.
Flicker noise is generated because of large amount of (direct) current present in the channel.
Shot noise has become significant lately because of small number of dopant atoms and thus electrons gives rise to significant statistical fluctuations in measurement.
Fluctuations in doping, dimensions, activity, and parasitics affect coupling, current density, etc. thus affecting the magnitude of noise.

2.3.5.2.1.3 Static Noise Margin

Static Noise Margin (SNM) defines stability of the cell (SRAM) in presence of noise. It is useful to determine how much noise can a cell withstand without inversing its output. Mostly memories use this parameter to characterize their limits. A better replacement to SNM is Noise rejection Curve (NRC) that represents combination of magnitude and duration at input to drive the logic gate to point of instability [77].

2.4 Dynamic variations

Dynamic variations consist of environmental and operating factors and constitute a large factor of variations present in semiconductor designs. Whereas they typically affect characteristics at block level, they have started to influence the impact of intra-die variations on transistor characteristics also. Three main factors constituting dynamic variations are

2.4.1 Supply voltage

Supply voltage consists of all the power/voltage domains in a design and their respective power supply. Fluctuation in external supply, glitches, spikes, magnitude degradation, etc constitute supply variations. Supply voltage has a direct impact on delay, power, leakage, transition time, hot spot, etc. It can also affect individual transistor characteristics due to intra-die variations. Current designs have multiple supplies, threshold voltage transistors, and power domains that make managing supply voltage variations a tough task. Reducing the effect of supply variations require extensive shielding of global nets, low resistivity lines, multiple entry voltage lines, coupling capacitance, etc. Low supply voltage increases the impact of variations on delay even more through DIBL and V_{th} fluctuations [127].

2.4.2 Temperature

Temperature variations can occur from changes in ambient temperature, dense interconnects, high activity, voltage variations, excessive leakage, poor design etc. Whereas short-term effects can influence threshold voltage, carrier mobility,

saturation velocity [23], and drain current, long-term effects due to continuous high temperature in a region increases electromigration and other related effects [106]. Both delay and leakage variability increases with temperature [125].

2.4.3 Activity

Activity defines the workload on a given block in a die at any time. It results into frequent transitions in signals across a broad range of transistors. High activity periods can result into increased temperature due to large dynamic power consumption. It is a usage dependent factor. Variations in activity across the design can cause temperature hot spots affecting subthreshold leakage [100]. Schemes like Dynamic Voltage and Frequency Scaling has been proposed to alter the chip functioning based on activity requirement.

2.5 Power

2.5.1 Power mechanisms

Power budgeting has become an important issue along with performance in designs today. Most applications are limited by one or other kind of power mechanisms. High-performance microprocessors are limited by cost of heat dissipation system caused by dynamic power consumption. Mobile applications are limited by recharge time restricted by leakage during standby mode. Architecture and implementation choices in technology, logic and circuit design dictates tradeoff between power and performance. Furthermore, leakage variations are much higher than delay variations because of exponential dependency and can vary by orders of magnitude. Various mechanisms through which power is dissipated are listed below.

2.5.1.1 Dynamic power

Traditionally dynamic power or switching power per transistor was scaled with each technology limiting the amount of energy consumption. However, in recent generations, supply voltage scaling has slowed to keep sufficient drive affecting scaling of dynamic power per transistor. Combined with increasing transistor density, dynamic power dissipation (or thermal power) has reached levels where it can melt the system if not properly controlled. Various factors that affect dynamic power can be given by equation (2-35). Principally V_{DD} and E_{sw} fluctuations are responsible for variations in dynamic power. C_L and f_{clk} indirectly depends on process variations. Dynamic power can also be specified using drain current affected by process variations [13] as shown in equation (2-36). Dynamic power is mainly reduced through architectural improvements.

$$P_{Dynamic} = \frac{1}{2} C_L V_{DD}^2 f_{clk} E_{sw} \quad (2\text{-}35)$$

$$I_{DSat} = \frac{W}{L} \frac{\mu_{eff} C_{ox} E_C L}{2} \frac{(V_{GS} - V_{th})^2}{(V_{GS} - V_{th}) + E_C L} (1 + \lambda V_{DS}) \quad (2\text{-}36)$$

2.5.1.2 Short circuit power

Theoretically, there should not be a short-circuit power. However, difference in NMOS and PMOS characteristics can create short circuit conditions. In addition, local mismatch effect on transistors can exacerbate the situation. The effect is relatively small compared to switching and leakage power.

2.5.1.3 Leakage power

Leakage is an ever-increasing issue in scaling limiting the full potential of scaled devices. Any further reduction in threshold voltage will result into large amount of leakage causing slow but continuous drain on the battery power. It can also result into logic error if the charge leaks away. Principal leakage mechanisms are subthreshold leakage and oxide leakage. The exponential dependence of leakage currents also results into high sensitivity to parameter variations.

2.5.2 Power management

Controlling power is one of the biggest issues along with variations. Yield is based not just on passing the timing checks but also the chip power consumption. Many different techniques are used at various level of design to reduce the amount of power consumed. Most of these techniques, weather at transistor or gate or architectural level have an effect on variations also. The combination of techniques to use depends on the application, necessity, and effect on yield.

2.5.2.1 Clock gating

The fastest way to reduce the power consumption in a design is to reduce the clock switching activity. One way to do is to shut down the clocks to block when they are not active for many cycles. To do the same, an enable signal is included in the flip-flops using which the output transitions can be stopped based on control signals. Clock gating does not add penalty on delay but an extra signal needs to be propagated plus some control circuitry. The benefits are high for blocks that are working only a part of the time. Gate level clock gating has some delay penalty as it works for individual gates based on pervious and next output. Gate level gating is more useful for power critical designs. Block-level clock gating does not have a large impact on variations but systematic effects because of extra input in flip-flop and routing does come into picture. Gate-level clock gating will have a noticeable impact on variations of cell as the output depends on at least three inputs now two of which, clock and enable, will be very close together. Two inputs switching close together changes the variations of the output signal.

2.5.2.2 Adaptive Body Bias (ABB)

ABB is used to compensate the leakage/frequency spread post-production [59], [55]. ABB is a mix of two different techniques, Forward body bias (FBB) and Reverse body bias (RBB) used ensemble to attain the best results. Both of them require a bias applied to the body of the transistor with respect to the gate. Forward bias lowers the threshold voltage and makes the transistor faster while reverse bias increases the same and reduces the leakage. These techniques were used in static mode separately in earlier nodes but the tradeoff is not very beneficial in current nodes. FBB has a large leakage current in off state and RBB slows the switching in on state. FBB reduces the delay variations due to reduction in threshold voltage but RBB increases the same. Using ABB, the advantages of both are combined along with dynamic usage to adapt to the environment. ABB can be used at either chip level or block level based on tradeoff benefits for the application. Chip-level ABB is used to apply a bias voltage for the whole chip automatically determined by measurements on one or more blocks, typically the most critical block [59]. The bias voltage only considers D2D variations but can track dynamic variations and adjust the bias accordingly. Other factors including leakage can also be taken as input but it increases the control complexity. Block-level ABB is more flexible and includes within die variations. Each block has its own bias generator separate from rest of the blocks obtained by tracking delay/leakage for that block. A central controller can track the overall chip frequency to direct the individual blocks to meet overall performance targets. Using block-level ABB systematic WID effects are reduced to an extent. Body biasing has shown promising results for ultra-low voltage subthreshold design [62].
DIBL has an exponential dependence on depletion width and thus depends on body bias through it. RBB aggravates SCE and thus increases leakage variability due to deteriorating effect on DIBL [73] but the increase in leakage is small compared to the reduction in total leakage. FBB reduces V_{th} roll-off and decreases sensitivity of V_{th} to L variations.

2.5.2.3 Dynamic voltage scaling (DVS)

Scaling supply voltage has the maximum benefits in terms of power. In ABB, we scale the threshold voltage while in DVS we scale the supply voltage [17]. As in ABB, voltage can also scale adaptively to just meet the performance required. Ring oscillators (RO) can be put on the chip at various points to measure the current performance. DVS can be again of two types: chip-level with lower cost and benefits and block-level with higher cost and benefits [101]. Supply voltage for individual blocks can be controlled using isolation cells to keep the effect of scaled signals inside the block only. It is possible to include variation as another parameter in the algorithm [14].
An extreme usage of DVS is in Razor [25]. Using this approach, an extra time-borrowed flip-flop is inserted with critical flops. Using the two, the error rate is monitored and used to control the supply voltage. The approach is better suited to designs with high pipelining where on the cost of few clock-cycles the correct state

can be restored. As the scheme is data dependent, it has higher accuracy, better savings and able to reduce variation impact largely. However, it comes with high cost for using extra buffers, control logic and routing. It is able to account for global and local variations as well as environmental variations. The issue here is to find a compromise between shadow-latch coverage ratio and its costs. Mismatch errors can convert a non-critical path into a critical one that is again missed by Razor approach.

2.5.2.4 Logic gate sizing

Larger logic gates mean larger drive, lesser delay, and higher power. Using minimum size logic gates can save power but will have higher mismatch variations also due to inverse square root dependence on logic gate area. The compromise is to find an optimal logic gate size and load capacity while still meeting timing requirements. Large mismatch between drive capacity and actual load present also increases the amount of variations. A better approach is to break down large fanout gates into two stages and to add dummy gates to low fanout ones. Consistent drive strength and load will have better matching.

2.5.2.5 Channel length

Channel length is among the largest contributor to threshold variations and leakage. An increase in effective channel length results into lesser power consumption and lesser delay variations on cost of some delay [97], [10], [29], [125], and [67]. It has been found a 15% increase in nominal channel length provides an optimal tradeoff point at 1.0V [127]. Any cost on delay can be recovered using larger logic gates or other techniques. Optimal length calculations are more application dependent and will need some trial runs before a basic knowledge database can be collected for tradeoff issues.

2.5.2.6 High-V_{th} transistors

Threshold voltage level works on opposite sides for delay variability and leakage. Larger V_{th} means higher delay variations but lesser leakage [125]. One possibility is to replace all low-V_{th} logic gates in non-critical paths with high-V_{th} ones so that the variation impact is still limited. In a well-designed ASIC system, most paths are near the critical edge and as such, the benefits of this system are marginal only.

2.5.2.7 Dynamic Voltage and Frequency Scaling (DVFS)

DVFS is an advanced version of DVS [17] where both clock frequency and voltage scaling are employed together on the block level to obtain maximum gains. Algorithms running on a small core decide the best frequency/voltage group based on the computational intensity of the task [61]. Variation related error detection schemes could be used along with to provide inputs to the controller. As the costs

are on higher side with larger benefits also, this approach is more suitable for bigger designs like microprocessors.

2.5.2.8 Power gating

Power gating uses various power/voltage domains present in a design to turn them off when not in use. A small cost of routing and isolation cells plus extra routing is present but the power savings are huge. A small wake-up time is necessary, delaying the tasks for that period. It is effective only in systems where independent blocks are present that are not working for many clock cycles. Multi-core microprocessors and larger mobile systems are some of the examples. The important thing here is to make sure that the blocks are completely isolated. There is no direct impact on variations.

2.5.2.9 Pulse-width Modulation

PWM as the name suggests controls the pulse-width to data/power transfer. This principal is being considered as a replacement to DVFS in next generation technology because of its simplicity and lesser cost to benefit ratio. The approach here is to run the CPU as fast as possible for a task and then go into sleep mode. A master PWM control signal monitors the input task and output to determine when to turn on the clock/high pulse and when to turn off the clock/low pulse. During low pulse, back bias is applied to increase V_{th} and thus reduce leakage. The same principal is also under consideration for next generation network controller. As we are just applying back-bias/stopping the clocks, it is relatively fast to remove the bias/start the clock to go into maximum performance mode. In addition, as the voltage is fixed at one level, the variations dependent on voltage level are relatively fixed also. With mismatch having dependence on voltage also, varying DVFS will change the amount of variations with each shift and impossible to analyze all possible situations arising from it.

2.6 Integrated Circuit Design

2.6.1 Modeling

2.6.1.1 Device model

Physical models (like TCAD) are the most accurate and form the basic transistor model where the actual transistor structure including various regions, doping, etc can be recreated [15]. The model is simulated for a single transistor (or more) behavior with different parameters using charge carrier models. It is also used to extract compact models. This model is used for exploring advanced transistor designs and atomic level issues like quantum effects, RDF, LER, OTV, etc. It is also used to verify device characteristics before actually fabricating the device. Once the technology and model are calibrated, it is easier to verify the effect of any minor

modifications. Physical models require huge amount of resources & time and are mostly used for advanced level research. Simulating variation in physical models require generating new transistor for each fluctuation. As such, it is quite difficult to generate statistically significant number of samples. Mathematical techniques have been used to simplify the process.

2.6.1.2 Compact model

Compact or empirical models use transistor characteristics to fit different parameters to obtain a fitting model for the transistor. It can be either mathematical fitting like in BSIM model or with more physical basis like in PSP. These models are used for logic gate or circuit level simulation to determine unit behavior under given conditions. Compact models can determine the impact of parameter variations on device as well as block level characteristics. However, they lack the functionality to implement atomic level fluctuations accurately. Indirect ways to include local mismatch are under study but their accuracy is questionable [16]. Compact models are used for creating timing characterization of logic gates that helps in high-level timing analysis.
Compact models categorize variations in three types and deal with them accordingly [107]. Predictable variations are modeled by adding additional layout parameters to include effects like stress, orientation and WPE, correlated variations are handled by random variables whose distribution is dependent on layout & spatial parameters and uncorrelated variations is handled by independent random variables whose distribution do not depend on layout and spatial parameters.

2.6.1.3 Numerical models

Numerical models have been created to fill the gap between compact and tabular models and are typically based on charge-current equations. A similar approach is using analytical models where equations are used to derive the circuit behavior [127]. Analytical models can also link process parameters directly to performance metrics. Monte Carlo or Response Surface Modeling combined with Principal Component Analysis [19] or Domain Decomposition Analysis can be used to decouple variation sources. These models can boast of sufficiently high accuracy with reasonable runtimes. Circuit level analysis can be performed using numerical models for timing, power, noise, etc.

2.6.1.4 Tabular models

Tabular models are timing models used for circuit level simulations and timing analysis. Logic gates are characterized to create timing and power tables based on corner, supply voltage, temperature, load, slew, etc. The approach is mostly used in digital ASIC timing. For a path, appropriate values are selected to obtain logic gate delay & interconnect delay according to load and slew conditions. The delays are then added in succession to compute overall timing and power metrics. Tabular models have been very successful in doing corner analysis to determine worst and

best case scenarios to calculate the performance limits of a design. Recent updates to tabular models have added statistical simulations using probabilistic data.

2.6.1.5 Behavioral models

Behavioral models are used at block level to verify functionality of a whole block. These models are used to verify logic, check block timing rules, global clock frequency, expected power consumption, intellectual property (IP) blocks, etc. Behavioral models are also used to synthesize the gate level netlist.

2.6.2 Timing analysis

Meeting timing requirements is a tough task for increasingly complex and bigger circuits with millions of transistors. The design should be functional as well as meet the performance targets within given constraints and be reliable enough for its contracted lifetime. Timing analysis is used to check and meet all these requirements. There are different constituents of timing analysis.

2.6.2.1 Device timing and full chip timing

Variations affect device timing directly but in most cases, we are more concerned about full-chip timing. If the full chip timing is within limits even with high device timing variations, then the design is okay. The issue is how to combine various device variation distributions to obtain the path/block/chip level distributions. Actual microprocessor blocks were simulated in [74] to obtain this given relationship. It was observed that for majority of paths in any block, delay is just below one clock cycle i.e. close to zero setup margins. Random variations smooth out large number of paths near zero setup time to distribute in proximity, causing non-linear variability. Systematic variations impact is higher compared to random variations as cells will be spatially correlated in the critical path and thus if any cell is in worst case, rest of cells in the path are more likely to be in worst case also. Block level timing margins are decided by combining path timing margins and random variation effects, with negative margins being most sensitive. The impact of variability also increases with block size, as in larger blocks the probability of outliers increases with increasing number of critical paths making it more likely to create slower paths. Globally asynchronous and locally synchronous architecture is somewhat resilient to many variations and thus more robust.

2.6.2.2 Corners

Corners are used to define and check the limiting case performance. It is typically consisted of limiting case inter-die process, voltage, and temperature parameters. They are defined so that most of the dies lie within the limiting cases and as such are directly linked to functional yield. In other words, it is the best and worst case dies possible in production. As the design is validated for limiting cases, all parameter values that lie in between will pass the test. The number of different types

of varying parameters like transistor, interconnects, environmental, etc have forced to increase the number of corners to be validated. That in turn has increased the time to market.
The biggest drawback of corner cases is they do not cover intra-die variations, as number of different configurations will be unattainably high. Traditionally corners are supposed to be pessimistic as they take worst-case values for everything that is statistically improbable. However, they provide a fast, comparatively cheap, and assured way of design validation to semiconductor industry and as such is the most common method of validating a design.

2.6.2.3 Margins

Margins are used to tackle any variations not already covered in corners. Typically, it consists of variations that are delay or design dependent like clock jitter, mismatch, etc. Margins are required to assure functionality. However, application of margins impact design performance and thus needs to be minimized. They are added on top of corners to create the limiting cases for a design.

2.6.2.4 Static Timing Analysis (STA)

STA uses tabular models to do timing analysis for a circuit under given condition. It does not involve any dynamic signals but simply estimate the delay for a given operational point or corner, thus "Static". STA helps to find which paths in a design may not meet the requirements or are the bottlenecks [7]. The advantage of STA is its linear runtime with design size. In addition, as it propagates signals through all paths, no test vectors are required removing test patterns from the picture. STA also offers an incremental operation useful for optimization where we do not have to redo a full analysis for a small change in design. However, STA based optimization results into a large number of paths just below critical limit increasing chances of failure in presence of large variations [9].

2.6.2.5 Statistical Static Timing Analysis (SSTA)

SSTA, as the name suggests, is the statistical use of STA. The main principal is still the same, i.e. tabular models [129]. However, it includes parameter sensitivity to variations in each cell that it propagates to calculate correlations among path delays. SSTA does not improve timing characteristics of a path but it helps to identify the weak paths that can then be made robust. SSTA aims to reduce the pessimism in design thus helping to boost efficiency and performance. However, it may be necessary to use SSTA only in combination with At-Speed tests to sort out any outliers.
SSTA has two flavors – path based and block based. Path based approach is more accurate but computationally expensive whereas block based is faster at small cost of accuracy. Most commercial tools prefer block based approach that is more suitable for large designs. In path-based approach, logic gate and wire delays along a path are added statistically to obtain the delay. The algorithm is simple but for this

approach, paths have to be pre-defined that is not feasible for a large set of paths. Correlation between logic gates due to reconvergence or spatial relations can be taken into account with this approach [9]. In block-based approach, arrival time and required time for each node are generated starting from clocked elements using mathematical max/min operation [22]. It covers the whole design but the algorithm is more difficult and requires many inputs including correlation between nodes for accurate calculation. Some of the required data may be related with process and thus dynamic in nature making it necessary for frequent updates. The variation-delay relation may not be linear which will require non-normal distributions to be used in SSTA [9]. In addition, the min/max operation in block-based approach is not very accurate.

SSTA needs statistical libraries whose characterization requires foundry specific information. Moreover, most fabrication changes improve reliability and yield rather than performance. As such, there is a marginal impact of process change on delay mismatch. However, in case a SSTA tool uses process information for delay calculation, any process change will require re-characterization of libraries. Any parameter variations that do not add to the cell's overall statistical distribution need to be removed from characterization effort else, they will increase runtime for no change in accuracy.

SSTA was supposed to alter the design approach but the industry has not been able to realize any major gains of fine-grained statistical analysis techniques over intelligent corner selection [84]. As such, it prefers a gradual implementation of SSTA [21].

2.7 Interconnects

Routing of interconnects is one of the toughest challenges of any design in deep submicron region [15]. Even a basic design starts with thousands of logic gates and connecting all of them while making sure of all the design rules, timing requirements, and density issues is a tough task. Interconnect delay constitutes a large fraction of the total delay with wire lengths going in kilometers on a single chip. Thus, any variations in interconnect have a non-negligible impact. Interconnects can be classified into different categories based on range and type of carrying signal.

2.7.1 Range

Interconnect wires on a chip can be roughly divided into three categories by range using length of the segment. With range, we know the average distance a signal has to travel and the type of variations it is sensitive to.

2.7.1.1 Global level

Global level interconnects are those traversing the whole chip like clock signals, power lines, and control inputs. These signals have to travel long distance, the order

of chip size. These lines have large capacitances and thus large crosstalk affect also. Clock lines have frequent transitions that can cause glitches in other signals. Power lines have to be least resistive to provide equal voltage level all across the chip. Control signals are highly sensitive and affect working of big blocks, and thus needs to be well protected from any crosstalk effect. Variations in global lines can increase the insertion delay by significant amount as it accumulates over the length. Systematic variations can result in unequal voltage levels across the chip. Delay in global lines is wire length dominated and wires are capacitance dominated.

2.7.1.2 Block level

Block level interconnects lies in between global and local interconnects. They are responsible for power supply within the block, clock distribution, primary input signals required over the block, some control signals like clock gating & enable, and bias distribution. These lines stay within the block or connect two blocks but do not extend over a large range. Device and wire delay are comparable while wire resistance and capacitance are also similar. These lines are less sensitive to chip level systematic effects but highly sensitive to proximity effects. Mismatch variations are also present to an extent. Clock signals suffer from unequal variations in these lines due to different rise and fall device variations causing pulse degradation.

2.7.1.3 Local level

Local level interconnects are lines present between various logic gates in a block. Their primary function is of data transfer. These are very short lines, a fraction of block size but the number of such wires is huge. These wires are prone to timing failures and hotspot effects. Delay in local line is device dominated and wires are resistance dominated. Mismatch is a bigger factor in local wires as they are very short and no averaging effect takes place.

2.7.2 Type of signal

Interconnects can again be differentiated by the type of signal they are carrying. Each signal has its own property and sensitive to different kind of effects or variations [15]. Routing of each follows different rules.

2.7.2.1 Clock connections

Clock lines are by far among the most distributed set. They are present at all levels right from the clock generator to the flip-flop. Other than the primary clock, there are many secondary clocks present on the chip running at different frequencies. Primary clock defines the chip frequency and the interaction between big blocks. Secondary clocks normally work inside the blocks or for connected blocks working together. With many power control schemes, these secondary clocks are operated separately. The primary clock is normally always on while the secondary ones can

be turned off. Clocks at global level are more prone to insertion delay variation whereas those at lesser levels are prone to pulse degradation. Clocks are normally high capacitance lines and as such have to be well decoupled so that they do not affect analog and RF blocks present on the chip through substrate noise.

2.7.2.2 Power connections

Power lines differ from other types of signals as in they do not have any transitions in general. They are purely metal lines with transistors present only in decoupling capacitors. However, for generating different voltage levels and for making isolated power and voltage islands on a chip transistors are used. The important thing in power connections is to maintain the same level all over the distribution network. For that reason, they are always distributed through highest metal layers on global level providing them with least resistivity. Most of the designs use grids to distribute the power. For power connections inside the blocks, they also serve a purpose of isolating the critical lines by shielding. The major issue for these lines is systematic effects.

2.7.2.3 Control connections

Control signals are trickiest of the lot. They are present mostly at global or block level and are responsible for controlling functions of the various blocks. Control lines generally have large fanouts and thus high capacitance. These are low activity lines but insertion delay and skew are very important. Any glitch in these lines can create significant issues relating to functionality of the whole block. They have equivalent device and metal delay but device variations tend to be a little larger. Random mismatch is not a big issue for these lines due to averaging but systematic & proximity effects as well as global and dynamic variations are present.

2.7.2.4 Data connections

Data connections are mostly present at block levels with a few input lines at global level. These lines are prone to almost all kind of variations including proximity and systematic effects, random, environmental, and global variations. Random mismatch tends to average out for long data paths but when combined with clock signals at flip-flop, it can be a cause for worry. Data lines have small metal connections and mostly device delay. Thus, the variations are highest in these lines. Long data lines can be prone to crosstalk noise.

2.8 Yield and Design for Manufacturability

2.8.1 Yield

Yield is defined as the proportion of good dies out of the total number of dies. Product yield can be classified into three parts- random (defect), systematic (design

dependent) and parametric (process related). Defect means random events that cause silicon failure like impurity particles causing line rupture. Design errors mean issues in circuit design that prevent it from functioning properly under given conditions like a wrongly latched flop. Process related yield loss depends on fluctuations in manufacturing that cause extensive fluctuations in circuit performance creating dies that lie outside the acceptable limits. Parametric yield can be improved through improved design techniques and through post-production tuning [59]. Most electrical design margin translates into improved systematic yield. Most layout margins translate into improved random yield. Margin and performance are competing requirements. Margin decides robustness and performance decides market requirements.

Process yield can be demonstrated in Figure 2-3 that shows a typical variation histogram of clock frequency. Maximum number of samples lies at the nominal frequency of the product and as we go away from the mean value, the percentage decreases. For a general digital product the cutoff point lies at +3σ value (except memories where it can reach as high as 7σ -9σ). In presence of power restrictions then anything below -3σ is also discarded thus the usable products constituting 99.73% of total. The following depiction is useful for design perspective to create the worst-case scenarios. However, it may differ from production perspective where the variation is considered for physical parameters and defect levels. Using product specific centering and process optimization, process yield can be increased. For a finished product, the rejection levels are counted in ppm scale (10^{-6}).

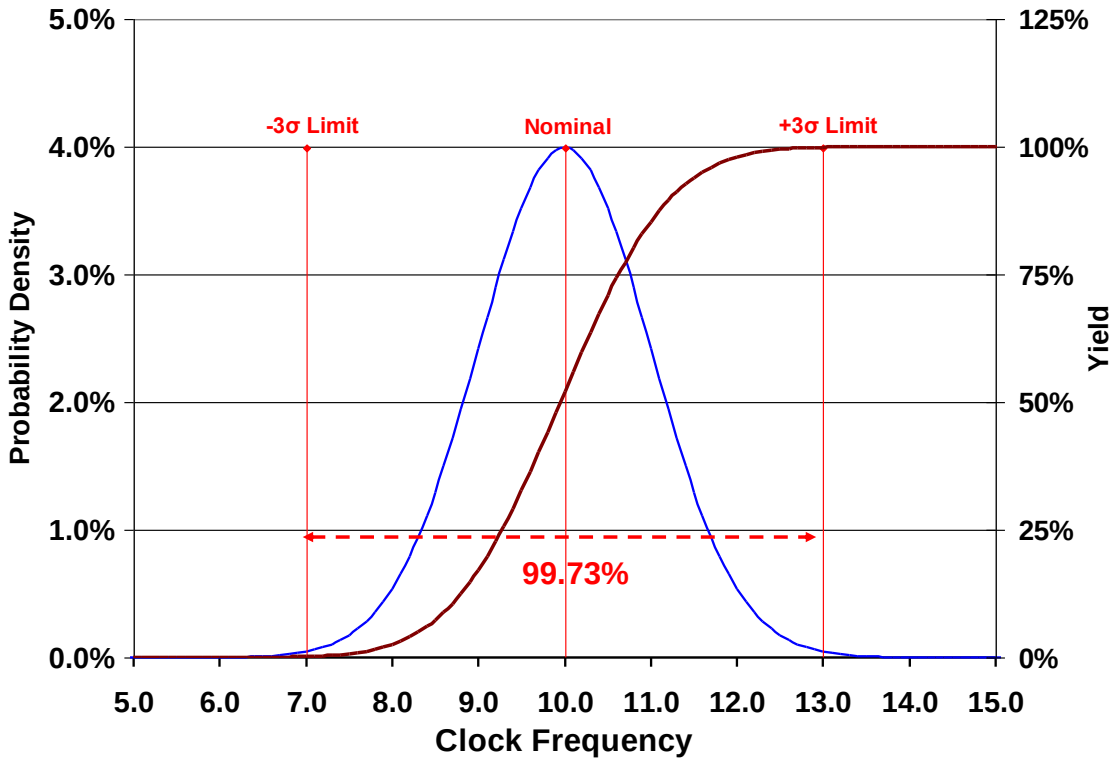

Figure 2-3: Yield histogram

2.8.2 Design for manufacturability

In nanometer scale, design is not independent of the technology. Already many mask data preparation steps have to be done after the design has been passed to production to make it usable. That converts to more time to market. Even after following all the design rules given by the fabrication plant, there may still be structures prone to parametric failures or defects. In all it can result into lower yield and higher cost. To increase the manufacturability of the designs, DFM approach is being used now. DFM promotes a set of guidelines that reduces the probability of failure post-production using existing knowledge accrued over time from the technology.

2.8.2.1 Rules vs. guidelines

Design rules are given by fabs to designers so that their design is within the limits of production. Traditionally these rules have always been pessimistic. Collected over the time from one generation to another, it is a huge set of rules including all the legacy rules [15]. Many of these rules may not be true anymore but to verify each and every rule requires huge amount of time. For current nodes, the set of rules have actually started to hurt the yield in an indirect manner. To follow all the rules while still being able to match the performance targets required, designers are using many techniques some of which are prone to parametric variations. The chip may pass the functionality tests but give errors later on during application use. Design rules limit the amount of performance that can be extracted from a chip making it necessary for designers to compromise on other aspects like power.
In contrast to design rules, design guidelines tell designers what they should do in their design to increase its manufacturability [18]. A designer may selectively decide to neglect few that he knows wont impact his design much. It is possible to extract higher performance levels thus making it easier to do a tradeoff with other parameters. These guidelines give the designer a probability of failure for his choices. He may be willing to forgo some chip yield so that his parametric yield may increase.

2.8.2.2 Manufacturability rank

Manufacturability (or DFM) rank gives the designer an objective parameter to know the manufacturability of their design. The DFM rank is calculated for all standard cells in the library based on their layout and structure. Using these ranks, tools compute the probability of failure for a given set of cells and paths. It is also used to optimize the design to decrease the probability of failure. DFM rank is based on the design guidelines and is not a hard and fast rule, so designers based on their experience and knowledge can bypass some of them. The guidelines are much more flexible and can be grouped based on their properties.

2.9 Reliability

Reliability in our case defines the ability of semiconductor devices to function error free. Different mechanisms affecting reliability and dependent on wafer processing are listed below. These mechanisms are directly linked to device level parameters and thus are directly impacted by process variations. Change in transistor parameters will change the limits for each mechanism.

2.9.1 Negative Bias Temperature Instability (NBTI)

NBTI affects PMOS device and is caused due to negative bias stress applied on the logic gate. It arises due to generation of interface and positive trapped charge while device is in operation [106]. It causes an increase in the threshold voltage and thus a decrease in drain current and transconductance. With time, the degradation increases and is facilitated by high temperatures. It depends on persistent high activity/temperature regions on a chip that can show failures over time.

2.9.2 Electromigration

High current density in aluminum interconnects causes gradual movement of ions due to current flow stripping away the material and resulting into increased resistivity of that area [106]. Aluminum lines have a polycrystalline configuration with many grain boundaries that aid diffusion of metal atoms [113]. The effect is aided by high temperatures and stress. High activity regions and large aspect ratio lines are prone to this effect. The effect is visible only when large-scale migration leads to a rupture in line.

2.9.3 Hot Carrier

Increasing electric fields inside transistors due to reduced scaling of supply voltage causes hot carrier degradation near drain areas. Large electric field can cause some carriers to gain sufficient energy to overcome electric potential barrier existing between the Si substrate and gate oxide film [113] and are called hot carriers. These hot carriers are injected into the gate oxide film and some of them can be trapped. Trapped carries form a space charge, and over a period cause a change or degradation of MOSFET characteristics such as threshold voltage (V_{th}) and transconductance (g_m). Un-trapped become gate current causing substrate current [113].
In contrast to other reliability effects, hot carrier degradation tends to increase as temperature decreases, especially in presence of stress. At low temperature, thermal vibrations of silicon lattice are reduced, in turn reducing the probability of collisions. This increases the mean free path and allows larger energy absorption increasing the number of hot carriers [113].

2.9.4 Time dependent dielectric breakdown

Over time, dielectrics tend to degrade. An electric field applied to an oxide film causes the injection of holes into the oxide film to occur on the anode side, and it consequently causes traps to be made in the oxide film. As the number of traps increases, an electric current via the traps is observed as a Stress Induced Leakage Current due to hopping or tunneling. If the number of traps continues to increase and the traps connect between the gate electrode and the Si substrate, the connection carries a high current that causes the gate oxide film to break down [113].

2.9.5 Stress Migration

Stress migration is the phenomenon in which metal atoms migrate in the presence of thermal stress alone, with no electric current applied. It is caused by stress that occurs from a difference of the thermal expansion coefficients between interlayer dielectric and metal wiring. ILD causes tensile stress on the wiring, resulting in movement of metal atoms, formation of voids, and eventually a disconnection. The lower the temperature, the greater the stress; the higher the temperature, the easier it is for the metal atoms to move.

2.10 Different approaches to counter variations

Variations can only be handled by using a combination of techniques at various levels of design and fabrication. Locally systematic can be handled through extraction techniques, spatially systematic can be handled using spatial proximity techniques, random treated statistically across the die and the unknown will require margins [94]. Some of these techniques are elaborated below.

2.10.1 Manufacturing and Test

It involves techniques that are used either in fabrication or at verification stage.

2.10.1.1 Immersion lithography

Immersion or Wet lithography replaces air by liquid in the gap between the lens and the wafer surface to increase the resolution limit. Immersion lithography for 65nm and 45nm nodes can achieve lower critical dimension variations.

2.10.1.2 DFM guidelines

By relaxing design rules and using more DFM type guidelines for manufacturing, the actual yield can be improved [30]. Design rules can be too restrictive and result in wastage of capacity. DFM guidelines require a good information exchange between foundry and CAD tools and make a design more foundry dependent.

2.10.1.3 Regular poly

Regular poly ensures equal density and coupling capacitance over the die. This in turn reduces systematic and lithography effects. However, there can be an increase in coupling capacitance.

2.10.1.4 Resolution Enhancement Techniques (RET)

RETs are a collection of techniques to increase the lithographic resolution of printed lines by modifying mask data [30].

2.10.1.4.1 Phase-shift mask (PSM)

PSM are photomasks with alternating thinner and thicker regions to produce interference between the passing light to generate a higher resolution at desired points. The constructive and destructive interference is caused because of scattering from these alternating regions [47].

2.10.1.4.2 Optical Proximity Corrections (OPC)

OPC is used to get final printed shapes similar to the desired shape. There is a big difference between the line shape present in design and the line shape printed on the wafer. OPC tries to obtain the final printed line as close as possible to that present in design by working on the intermediate mask data [88].

2.10.1.4.3 Sub resolution assist features

They are features placed near the edges of patterns on the reticle or are embedded into the pattern to increase or decrease scattering of light in specific regions. These features alter the slope of the aerial images of isolated and semi-isolated lines to match those of densely nested lines. They help to maintain adequate depth of focus across pitch to help reduce aberration effects and CD dispersion [68].

2.10.1.4.4 Double Patterning

It involves decomposing the design across multiple masks to allow printing of tighter pitches, i.e. split a task into more number of masks than in use today. It consists of multiple techniques including double exposure, self aligned spacer, and double exposure & double etch [49].

2.10.1.5 Hi-K and metal gate

Using a higher permittivity dielectric than SiO_2 and metal gates to match the corresponding work function, the electric thickness of the gate can be reduced for

larger gate thickness. This technique helps to achieve very low electrical thickness without having gate breakdown at atomic thickness.

2.10.1.6 SOI and double gate

Silicon-on-Insulator does away with dopants completely in the channel, reducing mismatch effect by a large extent. Further, it provides capability to do two or more gates on a single transistor, enabling the designer to improve on currents or to increase the functionality [106], [102].

2.10.1.7 Post production corner realignment

Using test structures, dies lying at extreme can be shifted towards the mean through post processing. It involves varying voltage and frequency to realign them within the limits. Extra logic and control circuitry needs to be added but it will improve functional yield [85].

2.10.1.8 At-Speed test

AT-Speed test is a necessary addition to full SSTA. SSTA can make most dies more robust but at the cost leaving out outliers that need to be filtered out using At-Speed test [20].

2.10.1.9 Asymmetric source/drain extension CMOS transistors

Atomistic variations cause different variations in S/D regions affecting overlap capacitance and effective source resistance [106]. There is almost one sigma difference between the source and drain regions in worst case due to dopant fluctuations [46]. The symmetrical S/D structure intended in the transistor is lost due to atomistic variations. Instead of trying to achieve a symmetrical structure, an asymmetrical transistor with different source and drain properties to mitigate the variation impacts and to achieve the best performance can be a much better choice.

2.10.1.10 Retrograde doping profile

Retrograde doping profiles give smaller variation as they keep the dopants away from the channel thus reducing the impact on threshold variation [106].

2.10.2 Modeling and Characterization

It involves techniques that improve the matching between designs and fabrication.

2.10.2.1 Lithography simulation

Lithography simulation allows simulating the impact of lithography on a design that can be used to find out critical areas and improve manufacturability [30]. Lithography simulation is heavily technology dependent and needs a stable technology for efficient use. Using lithography simulation, costly redesigns can be prevented.

2.10.2.2 CMP simulation

CMP simulation of a design provides data about systematic effects on a die/wafer due to CMP that can be used to analyze its impact [121].

2.10.2.3 PSP – physical modeling

PSP is the new model of choice for 45nm and beyond in the semiconductor industry. It is based on physical parameters and is more accurate than BSIM model for deep submicron effects [38].

2.10.2.4 Compact modeling of fully random phenomena

Intra-die random effects by definition are tough to model but their effects can be modeled using compact model that can provide better and more accurate timing analysis [12].

2.10.2.5 Numerical models

Numerical models provide a bridge between full spice simulation and table based timing analysis. The results are accurate and fast, necessary for industrial usage. Numerical models use charge-current equations to model the transistor or cell [127].

2.10.3 Library

It involves techniques that make standard cells more robust.

2.10.3.1 Anisotropic layout

Use of stepper in manufacturing causes more variations in one direction of the wafer. Anisotropic layout involves cells with all their poly lines in a single direction. It can then minimize the impact of systematic variations by aligning the cells with low variation direction on the wafer.

2.10.3.2 $L > L_{min}$ cells

Increasing the transistor channel length from nominal for a given technology gives a lot of advantage in variations and leakage [40]. Studies have shown that about 15% increase in L_{eff} from the nominal value is optimal at VDD 1.0V for 45nm technology [127]. It reduces the first and second order effects [125]. With increasing length, threshold voltage roll-off reduces the leakage. Transistor width can be reduced to preserve capacitance. For channel lengths higher than nominal in a transistor with halo doping, reverse threshold voltage roll-off is observed. However, for high DIBL doubling channel length is less effective than stack forcing. A larger gate length can be used along with FBB to get smaller V_{th} roll-off and DIBL [125].

2.10.3.3 Variation robust cell layout

Self compensated cell layout using iso and dense lines in the same cell is able to compensate for systematic variability effects to an extent [30]. Variability of these two types of lines works opposite to each other. If the iso lines go faster, the dense lines will go slower reducing the cell delay spread. Cell layout can be altered by using reduced variability transistor structures at cost of area and power consumption [26].

2.10.3.4 Variation aware cells

Cells can be made more robust to variations with static and dynamic capabilities [105]. In a flip-flop, variations are much higher for transitions of data signal close to the clock-edge [42]. Similar analogy can be applied to two or more input logic gates. The cells could be layout to offset the signals with respect to each other dynamically based on output variations. In addition, correlation of series connected transistors in a path is important for distribution. Similarly, it can also scale the threshold voltage accordingly to variations. Dynamic cell error-correction capability can be used like in RAZOR but for cell level [25].

2.10.3.5 Synchronized Level shifters

Multiple voltage/power domains require use of level shifters that have higher sensitivity of delay to process and voltage variations [13]. Using flip-flop in level shifter to have a timing boundary associated with conversion will limit the impact of variations on the signals inside one domain only.

2.10.3.6 Forced stacking

Stack forcing converts a single transistor into two and halves width of each. Forcing stacks in the cell layout has a dual advantage for leakage and variations [106]. A small delay penalty makes stack forcing useful for non-critical paths only. Leakage current through series connected transistors with at least one device off is an order

of magnitude less than through a single device [40]. The stack effect factor is given by ratio of single device leakage to stack leakage as shown through equations (2-37) & (2-38) and increases with increasing DIBL factor and supply voltage. DIBL effect is stronger in UDSM technology and thus stack effect is more effective.

$$\frac{I_{Single}}{I_{Stacked}} = 10^{U} \tag{2-37}$$

$$U = \frac{\eta(1+\eta)V_{DD}}{S(2\eta+1)} \tag{2-38}$$

V_{th} variations have two important factors, V_{th} roll off (SCE) and DIBL that depend on channel length as well as VDD. η is a linearized DIBL coefficient inversely proportional to channel length. With forced stacking, effective channel length decreases thus η decreases and so U decreases. DIBL coefficient has lesser sensitivity to channel length variation in stack case [10]. Stack effect decreases V_{DS} for lower transistor decreasing the DIBL effect and thus leakage variability with channel length [73]. The variation delay is higher for output gated by the lower transistor in a ground stack [106]. This fact can be used in logic gates with more than one input to keep the fastest signal in the lowest transistor.

2.10.4 Design

It involves techniques that can be used during design, routing, layout, etc.

2.10.4.1 Anisotropic placement of cells and interconnects

There is a strong correlation between cells in the scan direction [71]. Using similar analogy to anisotropic cell layout, the placement of cells can be constrained to align in a single direction. Furthermore, clock and data paths can be differentiated to profit from anisotropic correlation.

2.10.4.2 Regular layout

Dummy cells introduced to produce a regular layout might create antennas by collecting charge during manufacturing in the metal traces [89]. Dummy fill effect on capacitance depends on the size of block and distance between interconnects and fill. Poly fill for uniform density have a small effect on variations but there is a large negative shift of the mean frequency [71]. A denser fill can reduce the random variations but again with a large average penalty.

2.10.4.3 Alternating repeater insertion

Constant effective coupling capacitance for input transitions can be achieved by combining inverting and non-inverting repeaters (buffers and inverters) in a clock tree. Such a configuration also helps in reducing variation effects [23]. With

alternating repeaters, a worst-case delay on first half causes a best-case delay on the second half. It is also less sensitive to placement variations.

2.10.4.4 Path delay compensation

By using iso and dense cells alternatively, the delay spread along a path can be restricted [30]. The two types of cells occurring alternatively will compensate the systematic variation in each other.

2.10.4.5 Redundant/adaptive architecture

Reconfigurable architecture using redundant and adaptive blocks like in ElastIC can be used [31]. Such blocks can be post-processed to reduce variation impact as well as provide more functionality for less effort in redesigning [99].

2.10.4.6 Dynamic schemes for variation reduction

In schemes like RAZOR [25], supply voltage or bias can be dynamically scaled for individual blocks based on variation or output. The scheme ensures a high yield without extensive redesign at the cost of leakage and power consumption. However, increasing V_{dd} can result into reliability and lifetime concerns as degradation in gate oxide integrity and electromigration [66]. A mix of variation tolerant logic synthesis and dynamic scaling has been proposed to improve low power functionality and improved yield [104].

3 Comprehensive Overview of Clock Networks in Digital Synchronous System

ASIC designs are generally synchronous in nature. Although asynchronous designs are more tolerant to process variations [81], they pose big challenges in implementation and analysis and require a big shift in design methodology. Clock network constitutes the central nervous system of a synchronous design. The following chapter will give an overview of a typical clock system detailing the components, parameters, constraints, distribution methods, issues, etc. We also studied a typical clock network in a product CPU core to understand the clock network better and to able to judge the cost to benefit ratio of any design optimization. The aim is to look from a designer's perspective.

3.1 Synchronous system

3.1.1 Clock path

Clock path consists of the path between the clock root and the clock pin of the flip-flop as shown in Figure 3-1. The path principally contains balanced buffers and some gating elements. Limited type of cells in the path ensures better correlation among different clock paths. In a synchronous system, a clock path can be differentiated between launch path and capture path. The two are not absolute but relative to the pipelined stage. A capture path in one stage is a launch path in the next. The aim of clock paths is to ensure synchronous signaling with certain margins so that the pipelined stages can pass the data from one to next seamlessly. Mostly, the leaf node of a clock path is a flip-flop.

3.1.2 Data path

The time delay between output pin of launch flop to the input pin of capture flop is known as data path delay showed in Figure 3-1. Data path consists of various types of elements including buffers, logic gates, gating elements, multiplexer, etc. It also contains elements to correct any hold time violation. This large number of different elements present in the path reduces correlation between logic gates. However, smaller paths have larger correlation. With designs going for larger pipelines requiring lesser path depth, mismatch averaging affect reduces. Uncorrelated variations in logic path delays result in reduction of relative path delay variation through averaging, as such longer paths are beneficial for variation reduction [13] but the number of stages in critical path have been decreasing with time to increase clock frequency [9]. For correlated variations, sigma/average delay does not change. The delay distributions of driving and driven logic gate are correlated through the intermediate node's transition slope [42].

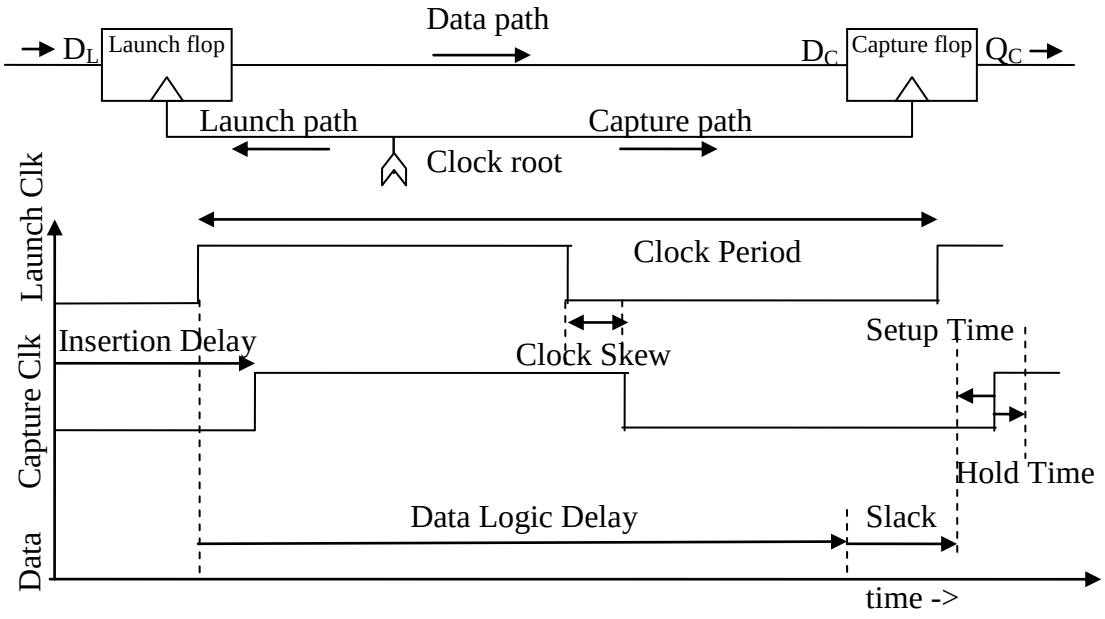

Figure 3-1: A typical synchronous system

3.2 Clock parameters

Local synchronous system mainly consists of three paths: launch, data, and capture. Each path has different structures and affected by variations in a different manner. Some of the concepts involved in such a system are described below. Equations (1-1) and (1-2) have to be true for all cases for correct timing in a synchronous system. Each of the concepts is demonstrated in the Figure 3-1.

3.2.1 Insertion delay

The delay time from clock root to the leaf node is known as insertion delay. It is important when considering bigger blocks and functionality at large. Insertion delay has a direct relation to the performance of the system. Larger insertion delay means the signal has to pass through larger number of logic gates thus increasing the signal variations as well as clock jitter. Some of the properties of local synchronous system are a function of the insertion delay also.

3.2.2 Clock period

Clock period defines the maximum amount of data path delay. Smaller the period smaller the amount of logic that can be put in between two synchronous elements. Clock period has to satisfy the relation given in equation (1-1). As the rising and falling edges do not pass through the same transistors, the amount of variation is different for two causing a variation in clock pulse. N/P transistor mismatch plays a big role in this variation as well as random variation. Proximity effects are limited as N and P transistors are close together.

3.2.3 Clock skew

Clock skew is said to be present when in a synchronous system, the clock signal arrives at different moments on two different clocked components. The amount of skew is the difference in their arrival times. Normally, designers strive to achieve zero skew as its presence increases system complexity and timing margins. Achieving zero skew is very difficult and most of the time designers include skew margin in arrival time. Skew can be intentional or unintentional. Unintentional skew arises from unbalanced configurations as well as variations present in the system that differentiates two paths [109]. Unbalanced loads in clock increase global variability effect [110]. Due to systematic and random mismatch effect of environmental variations changes, which in turn changes the amount of skew present in a system. Large number of buffers in clock tree makes mismatch very important.

3.2.4 Setup and Hold time

Setup time for a path is defined as the minimum amount of time that the data signal should arrive before the clock signal for correct latching. If the period is so small that the next clock signal after the launch signal arrives at the capture flop before the data has reached, then there is a setup violation. Hold time is defined as the minimum amount of time for which the data signal should be stable after the clock signal arrives for correct latching. If the skew between two synchronous flops is so large that the same clock signal that launched the data arrives at the capture flop after the data from that clock has arrived then there is a hold violation. For a system, hold violations are a bigger issue as they require inserting more delay in the data path by adding extra buffers thus changing the variation. Setup violations can be removed by increasing the clock period.

3.2.5 Slack

Slack is defined as the difference between required time and the actual arrival time at the capture flop. In other words, it is the amount of margin for path delay for a given data path. For the design to work perfectly, slack should be greater than zero. In practice, for well-optimized ASIC designs slack is a very small positive value in the worst case. As the slack variations are context dependent, a single margin for all paths is overly pessimistic design that still has a probability or error. It also has to be within the limits defined by the equations given above to respect the setup and hold time constraints. Thus, any variation in slack will affect the other parameters and vice versa.

3.2.6 Jitter

In digital design, clock jitter is the variance of clock period from cycle to cycle. It can be caused by either environmental variations or variations in clock generation. It can result in an uncertainty in clock-signal arrival time requiring additional margins.

3.3 Clock distribution

Clock distribution requires getting the same clock signal everywhere in the design at about the same time. A lot of different structures and concepts are used in clock networks some of which are detailed below [35].

3.3.1 H-Tree

A fully balanced H-Tree clock structure is considered the most robust clock distribution network against skew and variation effects. Each branch in H-Tree has four sub branches and extends in all directions equally starting from the centre. Most ASIC designs use some form or other of H-Tree for top-level clock network. For local distribution systems, it depends on the application. Though the H-Tree is balanced by levels, buffers insertion in the branches is done automatically and may not be equivalent for all branches. Differences in buffer placement as well as proximity effects on buffers will result into different amount of variations at the end of branches. For global H-Tree structures, systematic effects are also a cause for worry, as branches on one side will see it differently on the other side. With buffers going 40-50 stages, random mismatch may average out.
Maximum impact of variability is the in second and third stages of a 5-stage H-tree [8]. The main factors responsible for variations in clock network are lithography, RDF, and power supply – temperature variations in which V-T variations are time, location, and context dependent. Recent results have showed a 30% clock skew variation at leaf nodes in 45nm designs in which transistor variations are dominant [8]. There is a significant increase in interconnect variation contribution due to large interconnects, poor feature control, increasing wire resistance and variation. The maximum impact of variations on clock skew is at the border of global and local clock distribution network i.e. the 2-3 levels of a 5-level tree. Thus, an optimum place for variation reduction techniques is at the beginning of local clock network. The variations in wires are high but the contribution is less.

3.3.2 Tree

A normal tree structure has two sub branches for each branch and extends downwards from the top. For local clock distribution the structure is more similar to tree but the number of sub-branches vary between the branches with unbalancing from the buffer insertion present also. It can also be used as a feeder tree to a mesh structure. Tree structure at local level is more prone to proximity effects and less for global systematic effects as it is limited to a block only. Random mismatch do not

average out due to lesser number of buffers and intersection of clock and data paths at flip-flop. Local tree structures are highly unbalanced wherein the variations are different also. The implementation of tree is easier compared to H-Tree.

3.3.3 Mesh

To reduce skews for highly synchronous designs, mesh is used to provide the clock signals. A mesh is generally fed by a tree. Due to interlinks between the nodes, any skew is suppressed. It is also less prone to variation effects. The overhead of implementing a mesh is high due to high surface area. A lot of care must be taken for implementing feeder tree as large differences between different branches can cause short circuits. The power overhead of a mesh is high. New techniques involving local mesh or link insertion between leaf nodes of a tree to make it more skew and variation robust are being implemented [87]. These structures have lesser overheads and larger benefits.

3.3.4 Balanced and Unbalanced network

When two branches of same clock network at same level have zero skew at a given configuration and the paths for the two nodes are exactly equal in terms of cells/interconnects, such that they vary in the same way for global variations, they are called balanced clock paths. In general if two leaf nodes have almost zero skew and vary in the same direction with global variation, they are taken as balanced. However, differences in type and number of cells can introduce different variations in the two resulting in a non-zero skew seen at the leaf nodes making them unbalanced. For fixing hold issue at a flip-flop, extra buffers are inserted which can make two balanced leaf nodes unbalanced. Even if the two nodes have zero skew at one configuration, it can be non-zero at another. Random mismatch and proximity variations are the major issue here. Systematic global effects are small as two nodes that have to be in synchronization are generally close together.

3.4 Clock network components

Clock network consists or various elements other than buffers and interconnects required for clock distribution across the chip.

3.4.1 PLL and DLL

Phase-locked Loop (PLL) is used to synchronize clock signals using phase detection to control the clock frequency and phase with respect to a reference signal. PLLs are used to generate stable clock signals at global level and to synchronize them over big chips. Delay-locked Loop (DLL) is used to generate phase shifted clocks or for clock recovery. A DLL does not detect the phase but affects it directly. Both PLL and DLL are part of a clock system in a chip. As PLL uses a reference, any difference between intended and actual reference signal because of variations will

result into an unsynchronized clock. Similarly, DLL uses delay elements whose delay can vary due to variations and affect the generated clock. Random mismatch will play a big role in DLLs as output of each delay element in chain is considered. PLL and DLL are big elements and affected by systematic variations as well as proximity effects. Environmental variations can result into differences between the actual clock generated and the desired clock.

3.4.2 Primary and Secondary clocks

Product data sheet specify a single clock frequency for any design. However, there are multiple clocks inside any design and they may not be working at the same frequency. Once the base clock is generated, it is divided into multiple clocks through a hierarchical structure. Each clock is altered to its required frequency, wave shape, amplitude, etc suitable for its designated function. Primary clock is the basic chip frequency by which it communicates with the outside world. Normally a chip has one primary frequency. Other than that, there are many other clocks present inside the chip for different purposes known as secondary clocks. These clocks are not meant to communicate outside and just help with the chip functionality.

3.4.3 Clock domains

In a design, it is possible to run different blocks at different frequencies fixed relative to each other. It is necessary to generate the clocks from the basic clock and using different elements to obtain the desired frequency and phase. These clocks are not available over the whole chip but restricted to small blocks. These blocks are called clock domains. Two clock domains may communicate with each other in which case it is necessary to synchronize both of them. Systematic variations can be an issue in such a case affecting blocks differently over the chip area.

3.5 Pipeline vs. Logic depth

Increasing design frequency requires decreasing the logic depth and increasing number of pipeline stages to achieve more number of operations per second. On the other hand, increasing parallelism i.e. decreasing pipeline stages and increasing logic depth gives higher throughput rate where more number of operations per cycle can be completed. The current trend is towards higher pipelining that is more energy efficient but increases the impact of random mismatch. However, parallel designs are easier to include power saving features for selective operations. A typical microprocessor can have about 10-15 stages in critical paths. For 16 stages, WID critical path delay sigma is comparable to the NMOS/PMOS I_{on} sigma [59]. Increasing parallelism and/or functionality requires an increase in the total number of critical paths whereas increasing frequency through deeper pipelining requires increase in number of critical paths as well as decrease in logic depth [100].

3.6 FMAX vs. Number of critical paths

FMAX is defined as the maximum operational frequency of a design. The total number of critical paths (N_{cp}) present in a design has a direct correlation with FMAX. The impact of an increment in N_{cp} on FMAX is more important when N_{cp} is smaller, i.e. it follows the law of diminishing returns [65], [64]. An increase in N_{cp} causes a reduction in the magnitude of variations, but also reduces the average value of FMAX due to within die variations [80]. Systematic within-die variations do not average out over the path length as random variations and are thus a bigger challenge for FMAX variations.

The relationship between average and standard deviation of critical path delay for a path made up of NAND gates to that for a single NAND gate can be given by equations (3-1) and (3-2) for systematic and random variations. T_{cp} and σ_{Tcp} are the nominal delay and standard deviation of delay for critical path whereas T_{Nand} and σ_{Nand} represent the same for a NAND gate.

$$\frac{\sigma_{T_{cp}}}{T_{cp}} = \frac{\sigma_{T_{Nand}}}{T_{Nand}} (systematic \quad) \tag{3-1}$$

$$\frac{\sigma_{T_{cp}}}{T_{cp}} = \frac{\sigma_{T_{Nand}}}{T_{Nand} \cdot \sqrt{N_{cp}}} (random \quad) \tag{3-2}$$

Thus systematic within die variations are one of the largest performance degradation factor among parametric variations. Deviations in critical path delay in turn directly impact the FMAX. Within die variations largely determine the average value of the FMAX and die-to-die variations its variance. However, within die variations skew the shape of the distribution to a non-normal shape [65]. Recent nodes have seen significant increase in random variations that in turn have increased the variance of FMAX.

3.7 Synchronous system in a microprocessor core

Microprocessor core constitutes a part of many ASIC products and is a big block in itself. It thus forms a good candidate to study a typical synchronous system including clock network and data logic. We studied a CPU core (more than 200K cells) implemented in 45nm to observe the distribution of cells, nets, path lengths, clock network, metal layers, parasitics, number of stages, number of levels, fanout, etc. The purpose is to look at the statistical behavior of different quantities that can help to determine the efficacy of any approach to reduce variations. Any optimization having a large individual effect but a small target footprint may be less effective than that with smaller individual effect but a larger target footprint. The goal is to optimize from a product point of view. Using the data extracted from the microprocessor core, we plotted a number of graphs to see the relationship between different quantities. These graphs (Figure 3-2 to Figure 3-8) are analyzed below considering different types of cells and interconnects.

3.7.1 Distribution of cells

Cells can be divided into few broad categories between clock and logic. Each has its own particular type of cells that dominate the category.

3.7.1.1 Clock buffers

Most of the delay in a clock tree comes from clock buffers. To minimize variation difference among diverse branches of a clock network, only a limited number of clock buffers are allowed. A general rule is to use larger buffers in critical paths to reduce insertion delay and use smaller buffers elsewhere to limit power usage. Larger buffers are also used at leaf nodes to drive a large number of flops. In the given core, clock buffers consisted of less than 0.5% of total number of cells out of which 50% were that of the smallest drive. Dynamic power consumption is directly related to the buffer size and a high transition rate can make it a major part of the chip power consumption. As such, smaller buffers are generally used in non-critical paths. The goal in Clock Tree Synthesis (CTS) is also to reduce the overall number of buffers and thus save power. Large drive buffers are used at penultimate nodes to drive many flops also reducing skew between neighboring flops and consisted of about 20% of all buffers.

3.7.1.2 Clock gate

Clock gates are an integral part of clock networks and stop the clock signal from propagating to any non-functional block. Typically, clock gates are present at multiple levels of tree starting from root, clock divider, block level, function specific, memories, etc down to the group of flops. Clock gates act as the leaf node of clock tree connecting large buffers driving multiple clock gates and itself driving multiple flops. The fact is also corroborated by small drive clock gates constituting more than 70% of all cases and a generally decreasing proportion with increasing drive. Larger clock gates are typically used to drive memories, large fanout, higher up in clock distribution and critical paths.

3.7.1.3 Shift registers

The third major constituent of a synchronous system are shift registers including SR flip-flops, D flip-flops, latches and their variants. Their purpose is to control and organize the flow of data synchronized to the clock signal. Like clock gates, small drive cells constitute the majority with well over 85% of total registers with SR-type claiming bulk of the proportion and D-type lagging far behind. Lathes constitute a miniscule part, as they are level sensitive and not edge sensitive. Flip-flops do not have to drive large loads in general.

3.7.1.4 Memories

Another major category of cells in a design is memories. Whereas they are not many in number, each memory cell consists of a large number of transistors occupying a large area. Memories also have big input capacitance for signals and clock. Being usage specific they do not have much trends and are largely design specific.

3.7.1.5 Logic gates

By and large, the biggest category of cells is logic cells and consists of all cells excluding clock cells, shift registers, and memories. Most of the logic cells involve simple functions like inverter, NAND, NOR, XOR, etc but also include complex cells like multiplexer, adders, programmable cells, tri-state buffers, etc. Figure 3-2 shows the distribution of cells in the given design. As we can see, a few cells make up the bulk and then there is a long logarithmically decreasing trail. The top 10 out of some 700 different cells make up a little more than 1/4th of the 160K instances present in the design. Typically logic cells do not drive large loads and are thus of small drive. The most common cell is also the most simple i.e. an inverter. In general, the top 10 cells are simple logic gates like inverter, buffer, NAND with few exceptions consisting of multiplexer required for input selections.
Cell type and drive are a big factor in determining the impact of global and local variations. Moreover, the scaling of delay with voltage and temperature is not same for all cells. Individual layout is susceptible to systematic effects.

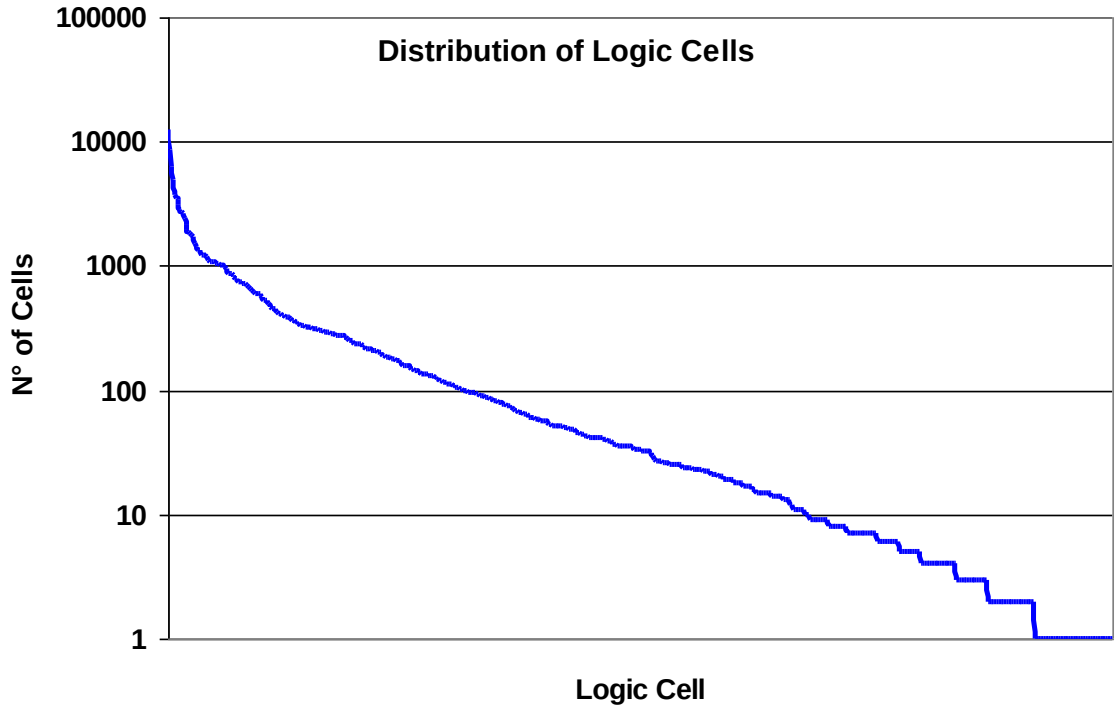

Figure 3-2: Distribution of logic cells in the given microprocessor core

3.7.2 Distribution of nets

In sub-100nm, technology nodes, ratio of interconnect delay to total delay has become very important hovering around 0.5, i.e. half the path delay is coming from interconnects. Moreover, connecting cells and providing them with power and clock connections are increasingly getting tougher. As such, the bottlenecks are not cell area anymore, but interconnect area or routing. Thus, it is important to understand how interconnects affect insertion delay. Typically the principal parameters that define interconnects are capacitance and resistance. In addition, wire length and fanout characterize nets also. For leaf nodes, number of levels and insertion delay are important criteria. The contribution of a net to total delay can be approximately represented by its resistance and capacitance product called RC product.

3.7.2.1 Net distribution

Most of the nets are very small connecting cells right next to each other like in case of leaf nodes, skew balancing, large load cells etc. As in cell distribution, the number of nets decreases logarithmically with increasing RC product as seen in Figure 3-3. The second largest group of nets is mostly a mix of medium to large fanout nets and long length nets. Being such a large collection of nets, it is rather difficult to find many trends at this level. Leaf nodes tend to have very large fanout and a large RC product. Few outliers are also there consisting of very long nets or very large fanout nets.

3.7.2.2 RC product

Wire capacitance is highly correlated to wire length as can be seen in Figure 3-4 whereas degree of correlation between wire resistance and wire length is comparatively less, as seen in Figure 3-5, probably due to very low resistance of higher metal layers used in power supply nets. Wire capacitance and resistance are prone to parasitic and systematic effects. They are also affected by reliability issues.

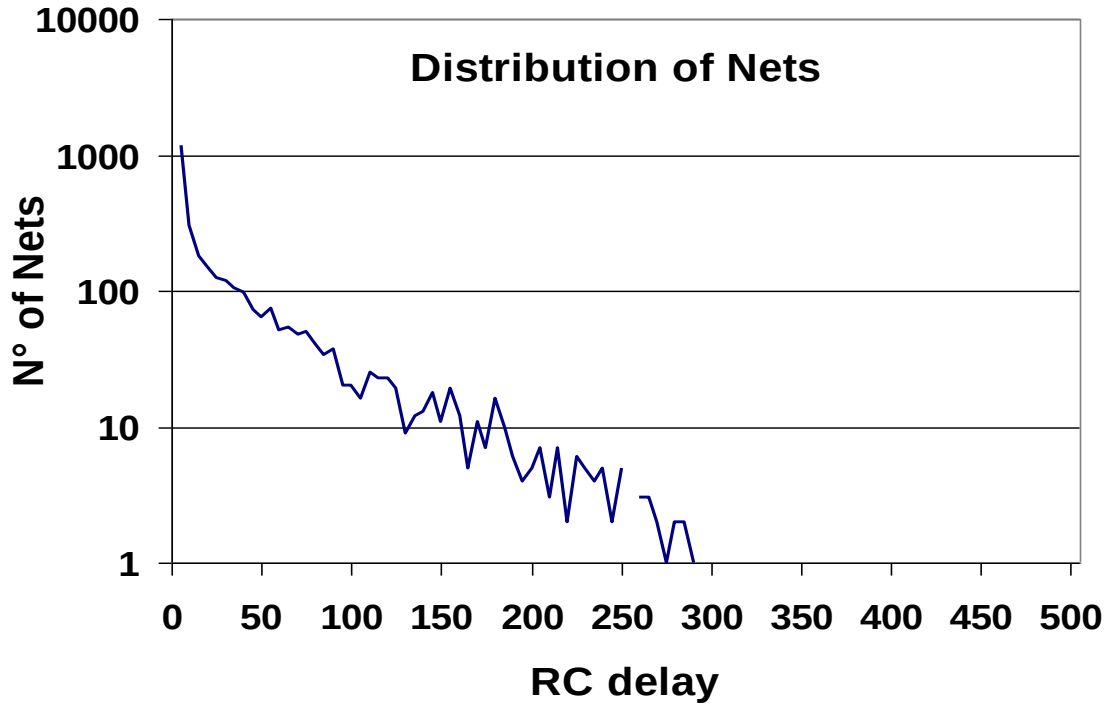

Figure 3-3: Distribution of nets in the given microprocessor core

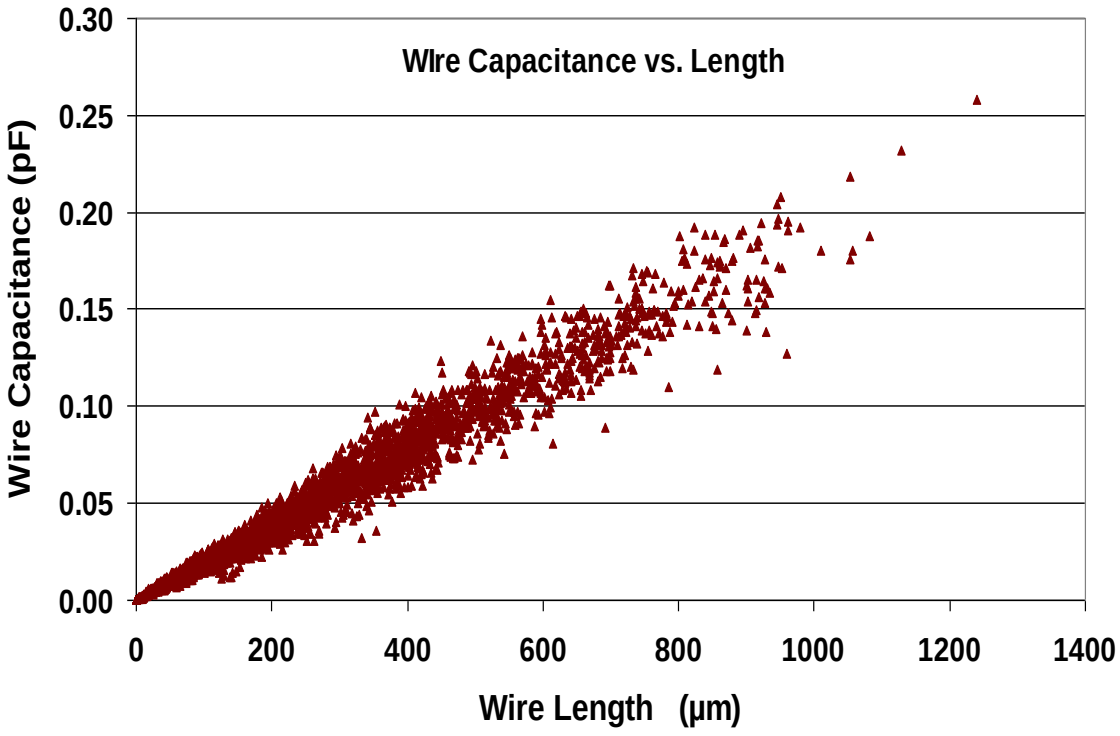

Figure 3-4: Correlation between net or wire capacitance and wire length in the given microprocessor core

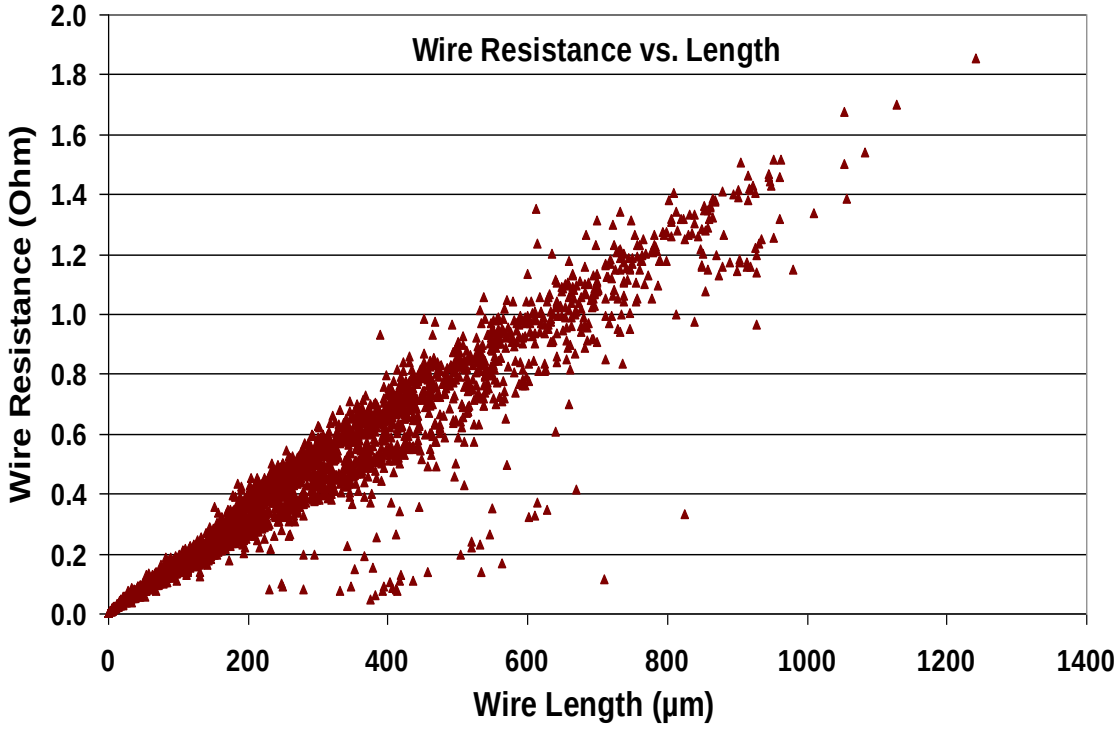

Figure 3-5: Correlation between net or wire resistance and wire length in the given microprocessor core

3.7.2.3 Fanout

Figure 3-6 demonstrates the fanout distribution for nets. Most of the nets have low fanout (below 10). These are the nets typically used in clock distribution from one level to next as well as in between logic gates in a data path. Nets driven by small cells dominate the distribution (fanout below 5) consisting of almost 50% of the total number of nets. Data path generally uses small cells and thus a low fanout whereas clock networks aim to minimize insertion delay and limit maximum slew because of which there is a limit on maximum fanout. A second group of nets visible in the graph is around fanout 35. This group consists mostly of leaf nodes in a clock network driving many cells to minimize the skew. Very high fanout nets (>50) generally consists of asynchronous networks like set/reset as well as test pins. Other than these three major classes, the fanouts are usage dependent.

The point to note here is that critical paths do not have large a fanout. Critical path delay has to be minimized and as such, the fanout is kept to a minimum, even at leaf nodes. Local variations can change drive current and input capacitance that can affect delay. Minimizing the number of cells in the path, also keeps local variations in check. Large fanout increases the slew that in turn is a factor in local mismatch.

3.7.2.4 Level

Level defines the depth of a path or the number of buffers present in that path. Traditionally, buffers were the dominant factor in path delay and defined the number of levels in a path. Although it may appear that insertion delay has a direct relationship with number of levels, it is not the case. Contribution of net delay to total delay skews the relationship between insertion delay and number of levels. Most of the leaf nodes are approximately around the same level and thus their insertion delay will vary between a given limit. Clock Tress Synthesis aims for a delay-balanced tree. There can be paths with larger number of levels but lie in the same delay range because of reduced net contribution. However, there will be few outliers having delay outside the limits. These are paths that either constitute the critical path or are non-delay sensitive. The critical paths with highest delay will probably have larger number of levels to maintain the signal.
Figure 3-7 shows the histogram of leaf node levels on a logarithmic scale. As seen in the figure, most of the nodes lie between levels 16 to 20 and very few nodes exist after level 25. A similar trend can be seen in Figure 3-8 showing the histogram of leaf-node insertion delay where most of the leaf nodes lie between 1.5 and 1.75 delay units. The result of clock tree synthesis can be seen on comparing Figure 3-7 and Figure 3-8. The spread of histogram for insertion delay is smaller than that of levels. Even the number of outliers are smaller for insertion delay. As the number of levels varies for leaf nodes, the impact of local variations as well as global n-to-p mismatch will differ also causing delay unbalancing between different branches.

3.7.2.5 Metal Layers

A net is made up of combination of multiple metal layers with via connections between two adjacent metal layers. Each layer has its own characteristic width increasing from bottom layer up. The lowest metal layer is generally reserved for intra-cell connections and is rarely used for inter-cell connections. The layer is characterized by very small width and thus very high resistance making it ineffective for longer connections. The next two layers are responsible for most of small connections between geometrically close cells and very few instances of longer lengths. The next two layers form the backbone of long connections, taking signal across long lengths without a lot of degradation. Clock distribution happens mostly in these layers. The last two layers are used mostly for power distribution and thus not preferred for signals especially in case of last layer that is rarely used due to its high capacitance. Sensitivity of each layer differs for different variations and affects timing differently [94]. Lower layers are more susceptible to reliability issues whereas higher layers are more susceptible to systematic effects. The parasitic coupling is also higher for upper layers due to their large capacitance. Delay difference between lower and higher metal layers is quite large making the layer composition of a net quite important in determining the percentage of net delay in total delay.

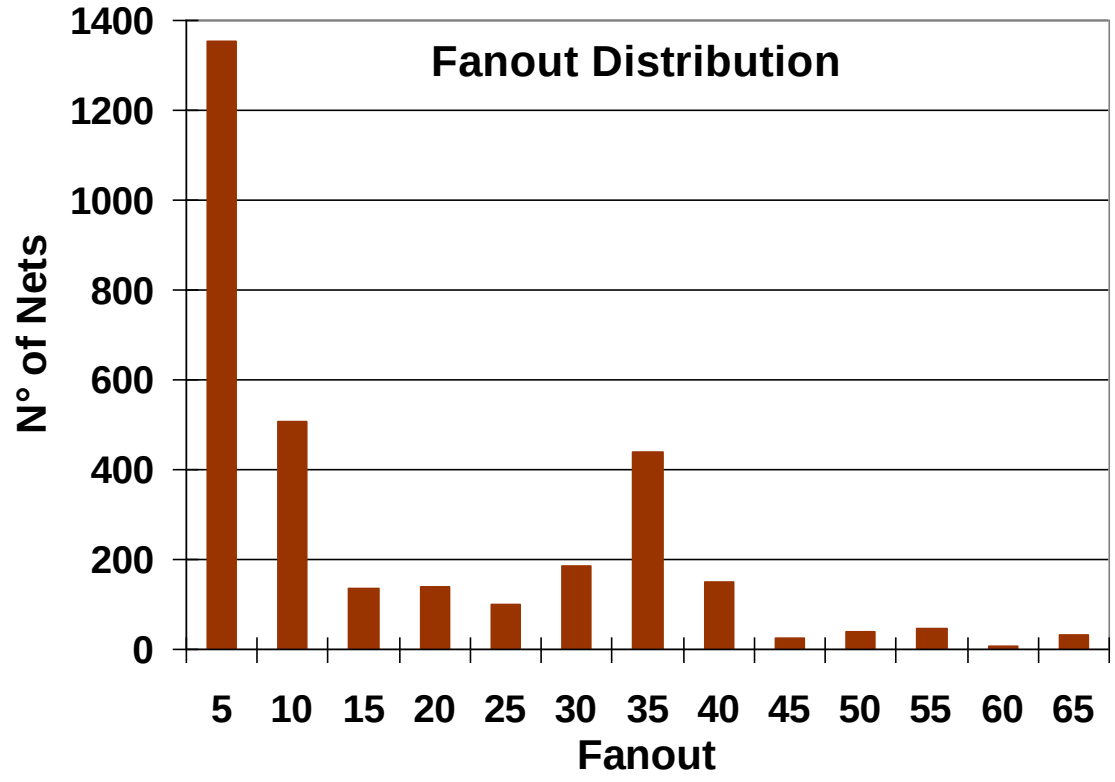

Figure 3-6: Distribution of net fanout in the given microprocessor core

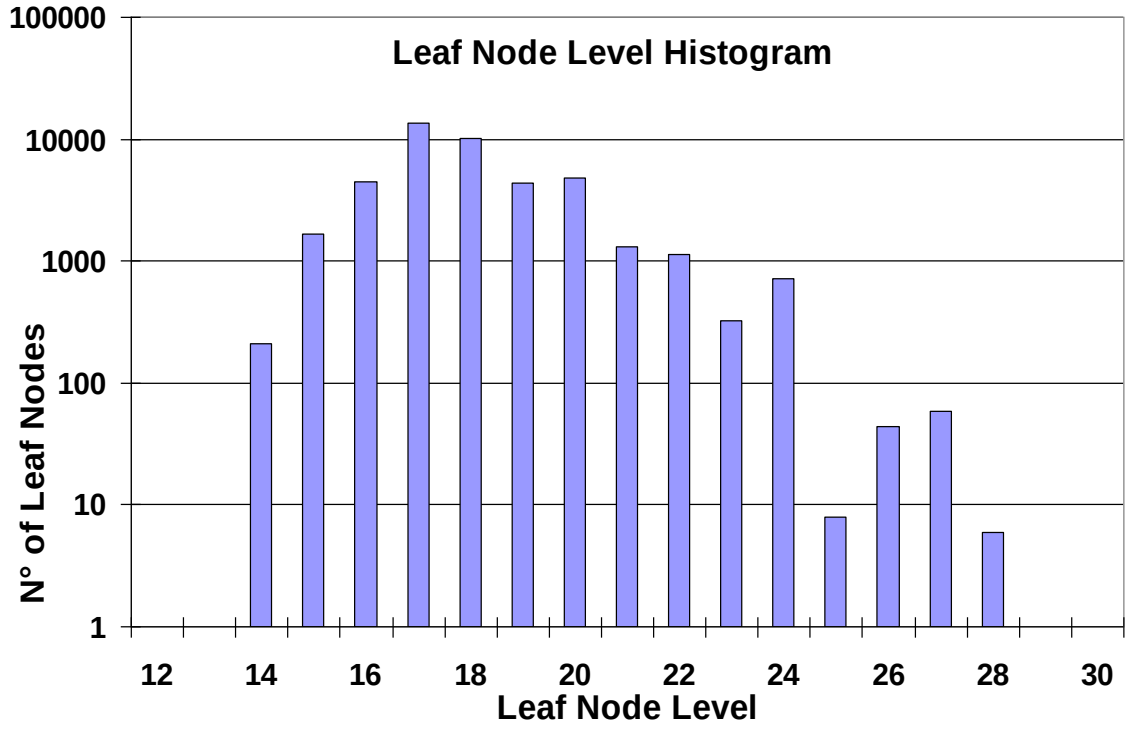

Figure 3-7: Histogram of leaf-node level in the given microprocessor core

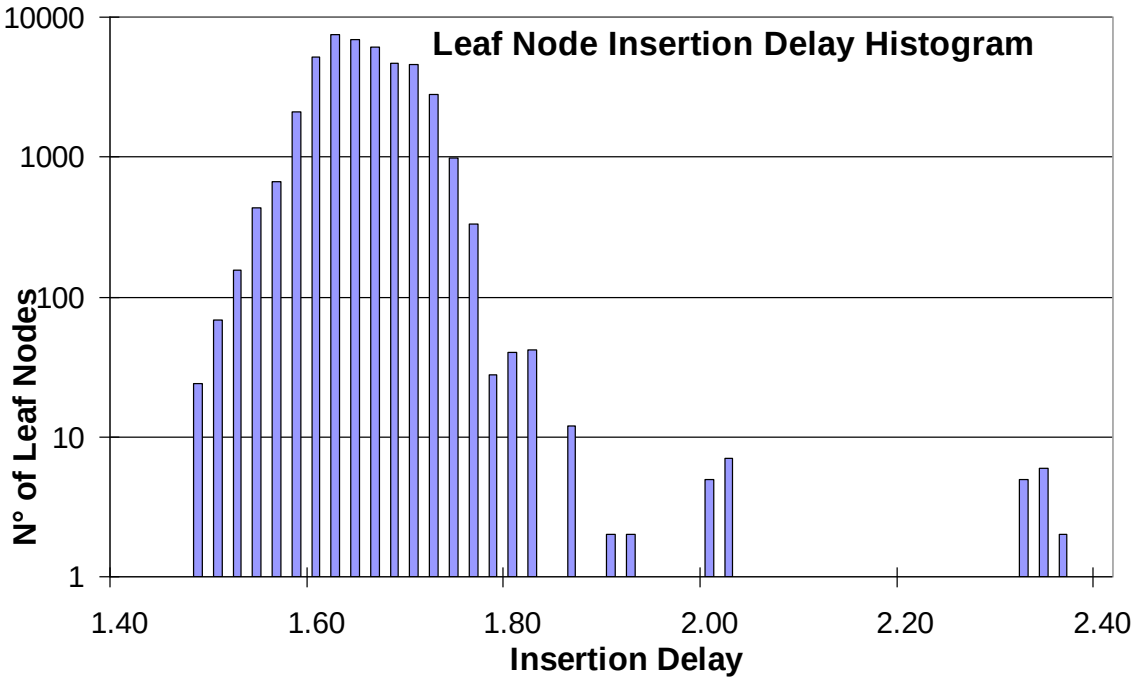

Figure 3-8: Histogram of leaf-node insertion delay in the given microprocessor core

3.8 Multi-voltage systems

An example of a general multi-V_{DD} system is shown in Figure 3-9 that uses a low power block inside a high power block. Supply voltage of two can vary separately. Connecting two different supplies require level shifter in between. The two blocks share a common clock as the low power block forms a sub-block of the high power one. Increasing integration can make such a system a common occurrence.

The double V_{DD} system in Figure 3-9 consists of a low power block (LP) and a high power block (HP) made of two types of transistors TS1 and TS2 respectively. The system is divided into two parts – input and output, to ease analysis. Clocks are balanced to have equal insertion delay (=1.2ns) in LP (1.1V) & HP (1.0V) with LP varying from 0.9V to 1.3V and HP varying from 0.9V to 1.15V. Hold time in data logic is fixed to have minimum slack in worst case that comes out to be SS,-40°C, HP=1.15V, LP=0.9V for input and SS,-40°C, HP=0.9V, LP=1.3V for output system. These conditions correspond to fastest launch and slowest capture. Slack is defined as the difference in arrival times of clock and data signal at destination register. Higher the slack, smaller is the probability of timing failure due to variations in arrival times. However, it also results into an under-optimized and thus a slower system. Multiple V_{DD} makes hold fixing a complex task as it is not the smallest voltage now that creates the hold condition. Moreover, the voltage levels are different depending on towards which direction you are going. Different temperature sensitivity of two transistors can affect timing.

The two systems support different minimum frequencies at different conditions- Input system 765MHz at SS,125°C, HP=0.9V, LP=1.3V and output system

430MHz at SS,-40°C, HP=1.15V, LP=0.9V. Minimum working frequency is what defines the bottleneck of the system that turns out to be the output system in this case. It may be possible to increase the output frequency by skewing the output clock but will affect hold fix conditions. As seen, optimizing such a system is a complex task that requires multiple variables. Minimum frequency will vary from one launch-capture pair to next based on specific path delays and composition.

Slack dependencies on LP and HP voltages at worst (SS) and best (FF) corner for input system are shown in Figure 3-10 and Figure 3-11 respectively, whereas those for output system are shown in Figure 3-12 and Figure 3-13 respectively. Larger the difference in available slack between minimum and maximum temperature for different LP/HP pairs, larger the temperature sensitivity. A system can have a nominal operating point anywhere in the 2D LP/HP region. The temperature sensitivity of slack can be seen at and around that point. For input system, the temperature sensitivity of slack is higher at minimum slack conditions, i.e. slower LP block and faster HP block. For output system, the temperature sensitivity of slack is higher for maximum slack conditions, i.e. slower LP block and faster HP block. Same LP/HP conditions for different (input or output) block can cause similar temperature sensitivity of slack. However, the slack is least in input block and maximum in output block. The worst working temperature for slack is also different for two blocks, 125°C for input whereas -40°C for output. Thus, the temperature sensitivity is worse for output block at minimum working frequency further affecting the manageable system frequency. Although, presence of maximum slack conditions in output block will mitigate the impact to some extent.

The given system considers fluctuations in voltage separately. The two transistors, LP & HP, can also have different mask layers that will reduce the correlation for process variations i.e. we can have a faster HP transistor and a slower LP transistor. Add on top temperature difference due to block activity and number of corners will get non-viable soon. On Chip Variations (OCV) provide a solution for reducing the number of corners. Instead of doing STA at all corner combinations, only specific combinations are simulated and extra margin is added on the paths to cover for fluctuations. The two blocks will have a certain level of correlation in PVT and thus doing corner with completely separate conditions for both can be pessimistic.

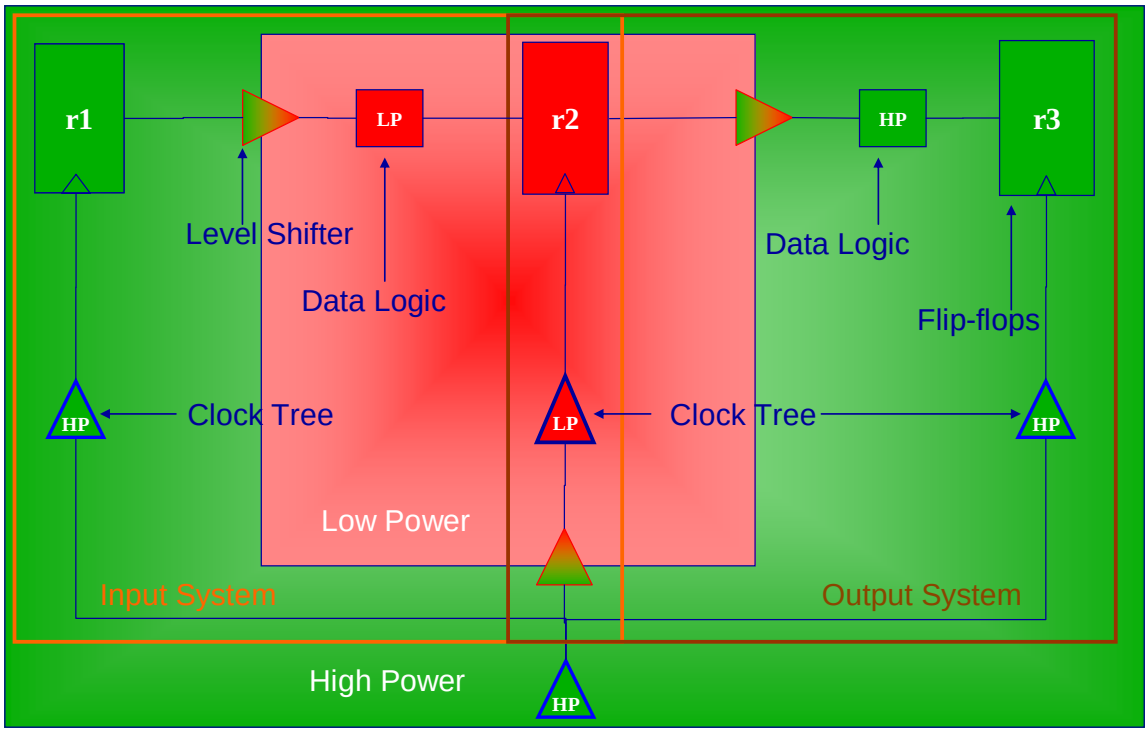

Figure 3-9: Multi-V_{DD} system

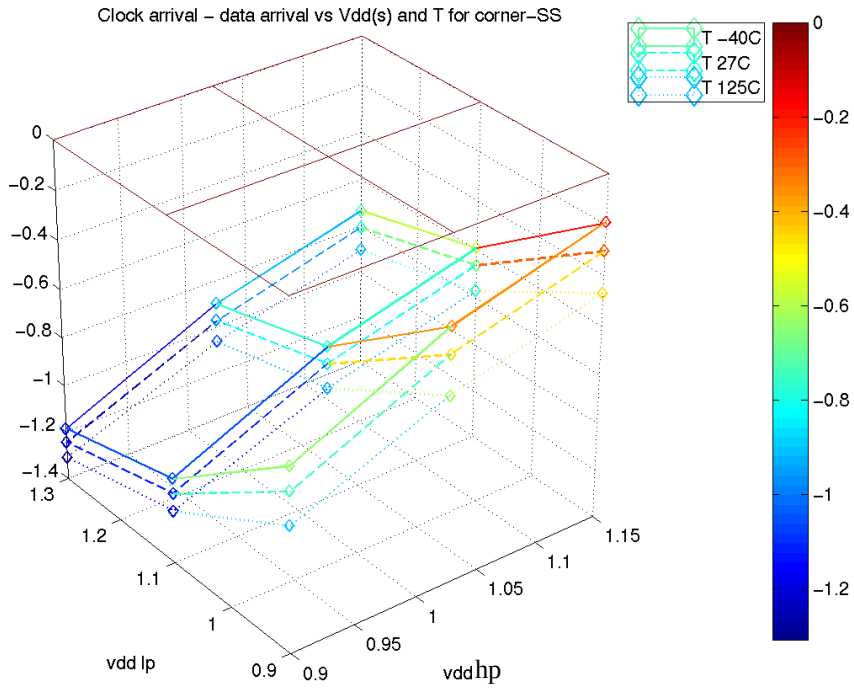

Figure 3-10: Input system: Slack at SS corner for different V_{DD} (LP, HP) & T

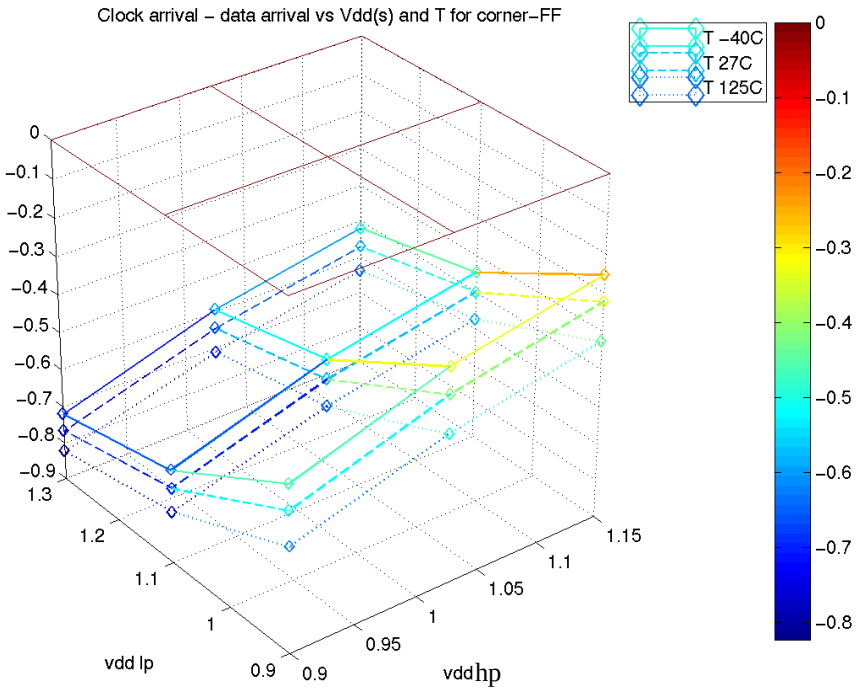

Figure 3-11: Input system: Slack at FF corner for different V_{DD} (LP, HP) & T

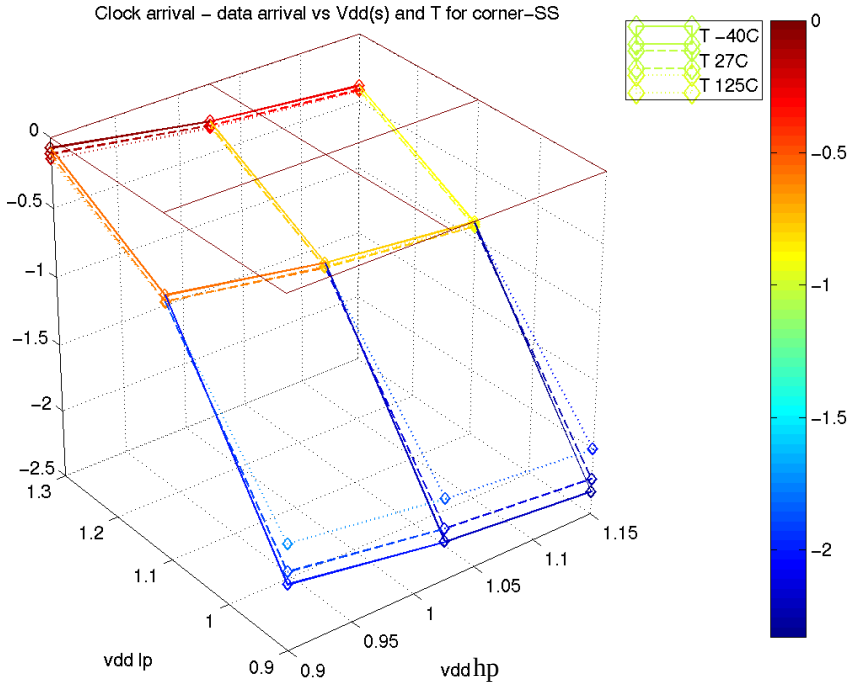

Figure 3-12: Output system: Slack at SS corner for different V_{DD} (LP, HP) & T

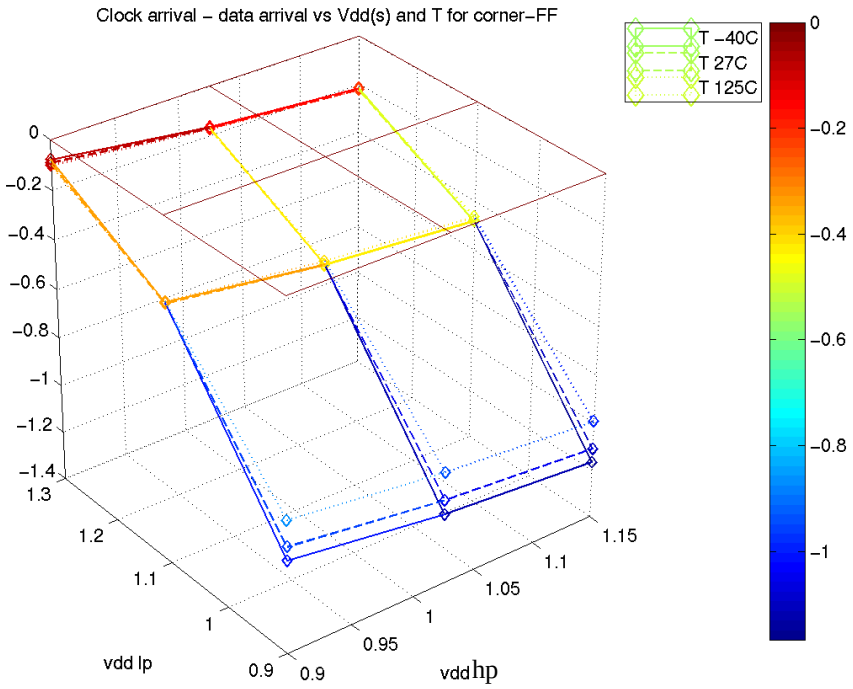

Figure 3-13: Output system: Slack at FF corner for different V_{DD} (LP, HP) & T

3.9 Unbalanced clock configuration

An ideal design would have perfectly balanced paths i.e. equal delay and scaling factor. However, placement and clock tree synthesis will induce certain degree of unbalancing into the system. Moreover, designers have to make compromises to attain desired objectives and further unbalance the paths. Figure 3-14 shows the two cases where unbalanced path uses a different type of cell in logic path. The arrival delay at capture flop in both paths from root node at nominal condition (1.10V, 125°C) is equal for both balanced and unbalanced configurations.

Figure 3-15 and Figure 3-16 shows the impact of die-to-die process variations in the presence of varying voltage and temperature on balanced and unbalanced configurations respectively. The y-axis in both graphs shows the normalized 1-σ variation of difference of arrival delays at capture flop. Different applications may require different amount of σ variation and as such, statistical timing models use a 1-σ variation, which we are showing in the given figures. It is a qualitative number representing hold time variation. As seen in Figure 3-15, the delay difference increases with voltage but remains steady with temperature for balanced configuration. The delay difference is higher for unbalanced configuration for same voltage as the two paths do not vary in the same way (Figure 3-16). Moreover, unbalanced configuration is affected by temperature variations also increasing the delay difference as the two cell types do not scale in the same way (Figure 3-16). This hold value needs to be fixed for all cases requiring extra cells.

In sub-50nm technology, temperature inversion happens at low voltages. An unbalanced circuit working at very low voltage will have a higher hold. Temperature inversion is dependent on the drain current and thus the transistor size. Thus, not every cell will see a temperature inversion at low voltage complicating the task.

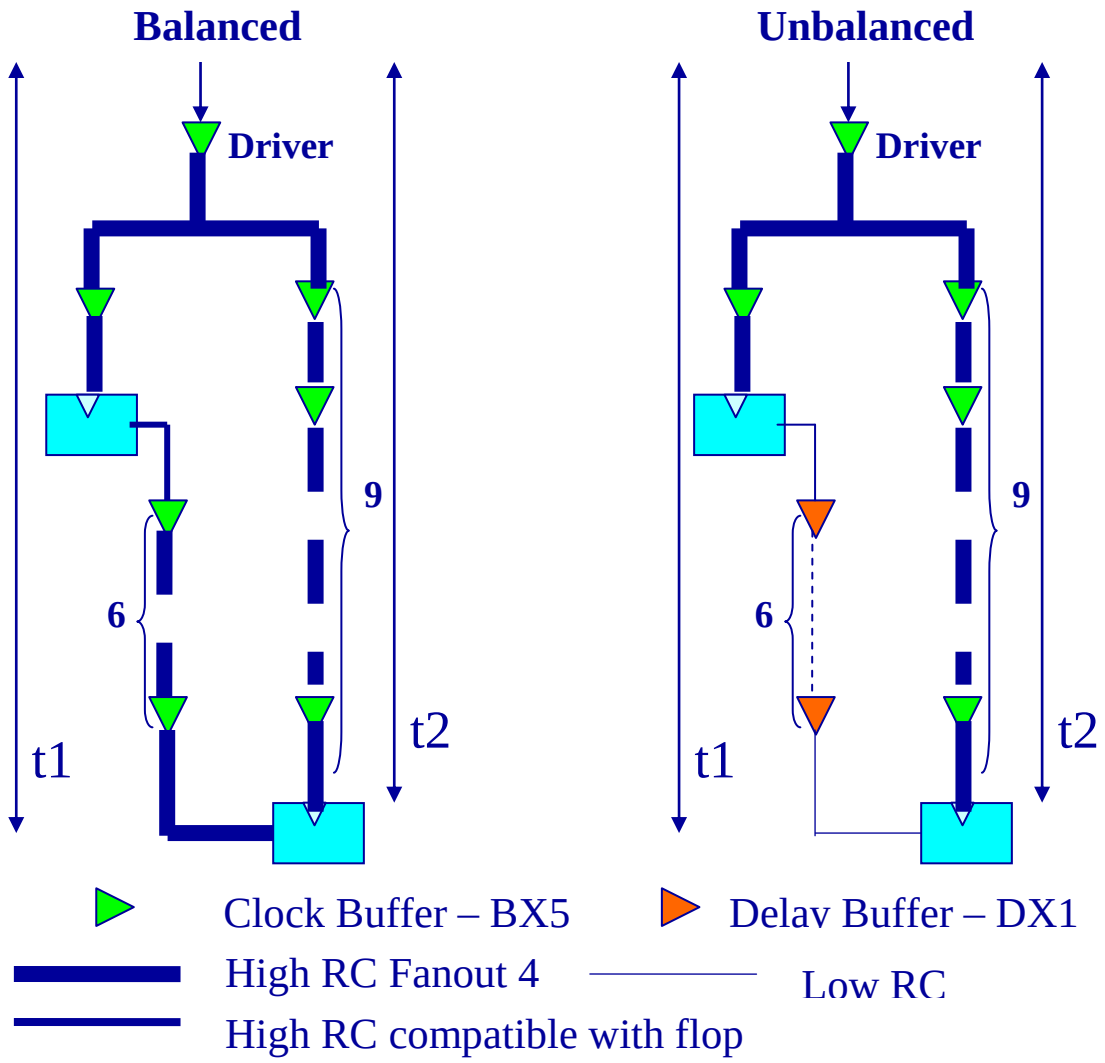

Figure 3-14: Balanced and Unbalanced clock skew configurations

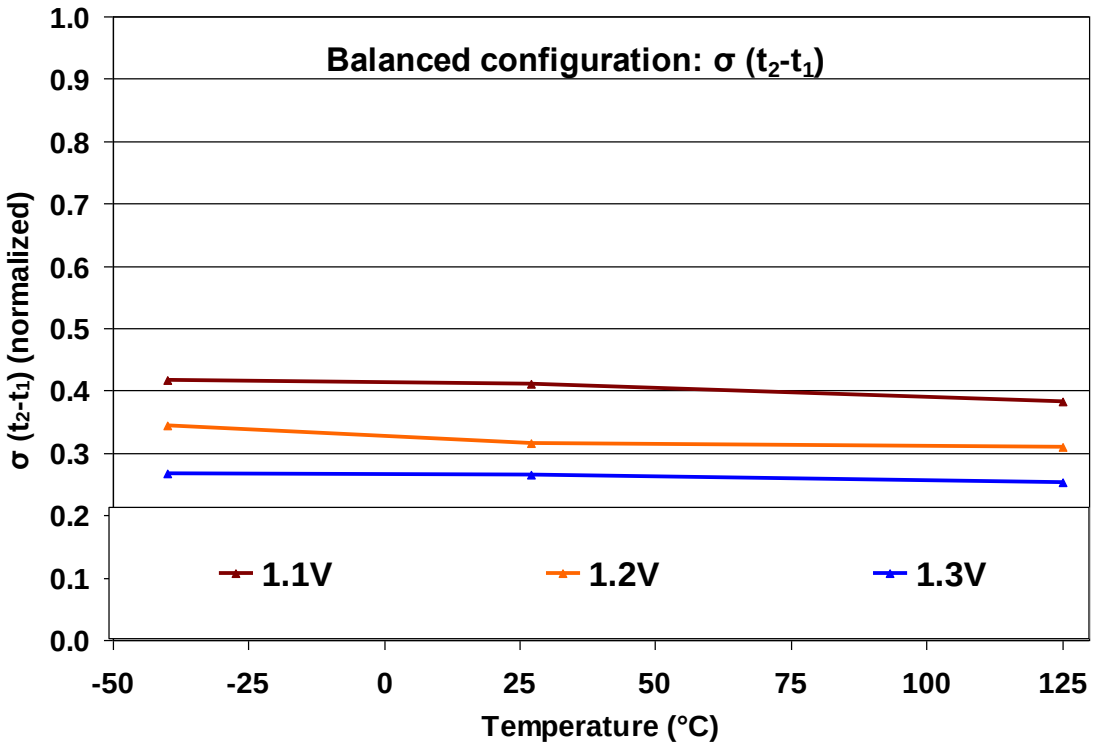

Figure 3-15: Global process variation impact on a balanced configuration for different V_{DD} & T

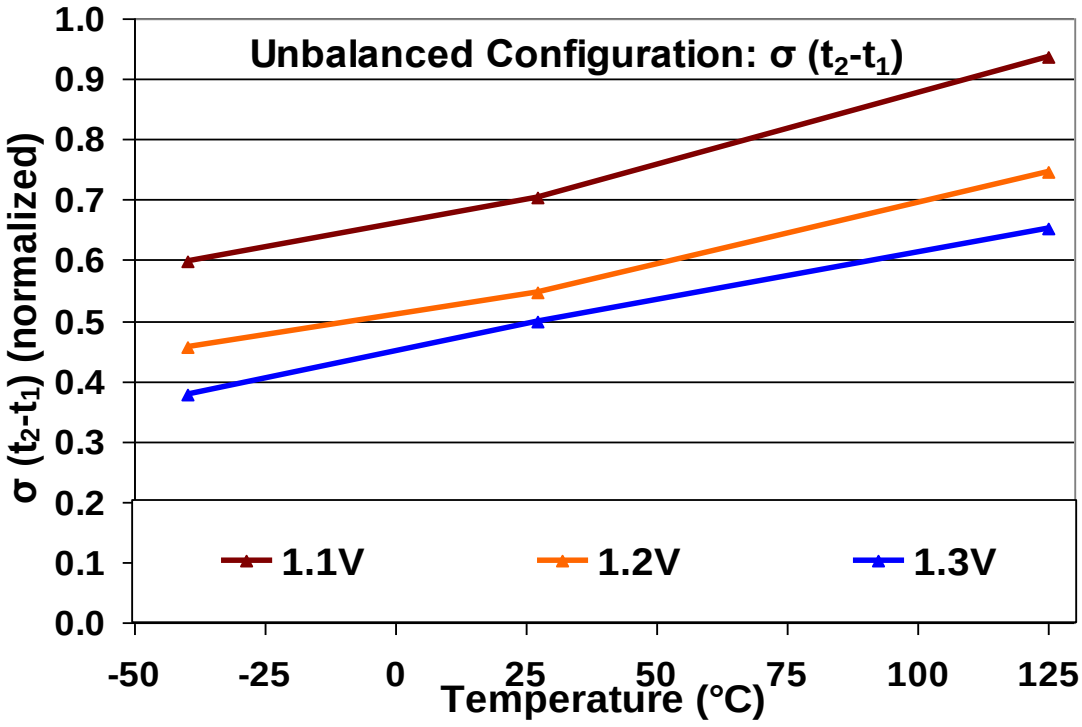

Figure 3-16: Global process variation impact on an unbalanced configuration for different V_{DD} & T

4 Experimental Framework used in the Research

Variation impact is an elusive quantity to measure accurately in real designs. Although its impact can be seen on physical quantities like delay, these parameters are affected by other phenomena like process centering, etc. Further extracting the effect of a single type of variation like local mismatch is very difficult. Practically it may not be an impossible task but time and cost limitations make it complicated to work on real designs. Simulations offer an alternative to silicon. However, their result is only as good as the models.
This work is mostly based on spice simulations using industrial production models. The choice of spice as the basic test bench is based on cost in terms of time and effort to accuracy compromise. The best options for accuracy may lie in TCAD models but they will be expensively time consuming for path level analysis that we are interested in. STA tools offer the other end of the approach with fast results but a reduced accuracy. The amount of control over parameters possible in spice allows us to separate out individual effects thus allowing us to verify efficacy of optimization strategies. Moreover, spice models allow physical simulation providing picoseconds level accuracy.
The following chapter will detail the models, simulation framework, parameters, etc that were required during the experiments. It will give an understanding at the issues that need to be considered during this project.

4.1 Spice model

The spice models used are production level industrial models, i.e. they have a stable set of parameters required to specify the technology. Although the models are first generation ramp-up process production models and might show a larger magnitude of variation compared to the process today, the general trends are expected to be same. We used the same models to maintain consistency over the lifetime of the project as is the practice in design projects. Transistor models are PSP based encapsulated within a shell that also consists of equations to model various phenomena giving it a high level of accuracy. All the PSP parameters are referenced by a second set of parameters that are specific to the industrial model. Equations related to the different phenomenon like Well Proximity Effect, aging, systematic variations, temperature effect, etc that are specific to their technology and built into these parameters and applied on top of standard PSP model. Model parameters are extracted by characterizing silicon test chips to enable a realistic behavior. Variation models including mismatch are built into the spice models itself by extracting more than 60 different variation parameters for NMOS & PMOS devices over a wide range of test circuits. We used the same models as provided to keep them realistic. Spice simulations provide a high level (ps level) of accuracy needed to measure the impact of variations on delay at logic gate level.

4.1.1 Global NMOS-to-PMOS mismatch model

Traditionally, global variations could be bounded by slow (SS) and fast (FF) corners. It was sufficient to obtain the worst and best case delay using the two corners. However, increasing global variations have given rise to global NMOS-to-

PMOS (or N-to-P) mismatch or inter-die unbalanced variations caused by doping fluctuations between NMOS and PMOS devices. They cause the elliptical shape (instead of a straight line) of NMOS-to-PMOS I_{ON} curve in Figure 1-1. Their principal effect is on pulse-width as it is made up of opposite edges traveling through different transistors. If the two edges travel at different rates, then the pulse-width or duty cycle changes along the path limiting the potential path depth.

Traditional corners are constructed by taking the limiting parameters for same case (best or worst) for both device types. However, unbalanced corners are constructed by taking limiting parameters for opposite cases for both devices i.e. best-NMOS/worst-PMOS or worst-NMOS/best-PMOS. The point to note here is that only parameters affected by doping differ between two devices and the rest, like critical dimension, remain the same. The model is constructed by extracting the best-worst pairs in silicon test chips. Final implementation and usage is same as traditional corners.

Corner models are necessary to reduce the computation time and complexity of timing analysis for a design. They are typically but not necessarily closer to 3-σ values for parameters variations. Corner variations are derived through statistical and analytical analysis of test chips. However, corners are not suitable for all applications like yield analysis. Statistical models are created for such conditions using similar process as corner models but with 1-σ variation value and distribution characteristic for each parameter. It is necessary to use 1-σ values as design application can require different value of sigma. Digital designs use 3-σ but analog and mixed uses 4 or 5-σ and memories even higher. Thus, using a 1-σ value gives a common model for all.

4.1.2 Local random mismatch model

Local random mismatch or intra-die random variations are caused by statistical differences between different transistors on same die. Some of the principal causes are RDF, LER, OTV, and polysilicon granularity. Currently, surface potential (replacing threshold voltage) and mobility are most affected, in line with RDF being the major phenomenon. Local random mismatch model is created by extracting its impact over different transistors on same die. Unlike global variations, local variations are by default statistical in nature. The model is created using 1-σ value and distribution characteristic for both parameters extracted from test results. The two parameters are interdependent but with varying degree of correlation. Thus, a third parameter varying randomly with a normal distribution is created to assign the degree of correlation between two.

4.2 Standard cells

We used cmos standard cell libraries created in 45nm technology used in production. Most of our work is concentrated on clock libraries but includes logic cell libraries. A clock library consists of various types of buffers required to drive the clock tree, combinational cells required for clock generation, division and pulse shaping, clock gating cells, flip-flops, etc. These buffers are optimized for driving a

clock tree and balanced to achieve equivalent rise and fall time and respective delays. These are low power libraries that function at a wide range of supply voltages allowing them to target multiple applications. The applications can vary from relatively high performance to very low power consumption. Cell models are created by post layout extraction of spice parameters and parasitics. Regular design strategy has been used in these libraries to minimize systematic effects. Using standard cells aligns us with real design issues.

4.3 Monte Carlo simulations

Monte Carlo simulations using statistical variation models for spice parameters provide a good way to derive the impact of process variations on circuits. Random sampling assures a realistic mix of samples. We used a sample size of 1000 for our runs to obtain results with 99% confidence (3σ) [50] that is the standard practice in industry. Verification simulations were conducted on a clock path to measure delay with 100, 1000 and 1M samples. The error ratio with 100 samples as compared to 1M samples was found to be more than 10% whereas the same with 1000 samples was less than 0.5% in case of global variations only and local variations only. The error percentage was approximately 0.7% for 1000 samples as compared to 1M samples in case of global and local variations combined. Thus, a sample size of 1000 provided us with a good compromise between simulation time and accuracy. The runtime for 1M samples was more than 10 days on a server farm using 100 machines.

4.3.1 Variation calculation

From extracted dataset of a quantity (e.g. delay), we calculate the nominal (zero variations), mean (μ) and standard deviation (σ) of the distribution. $\mu \pm 3\sigma$ gives us the statistical limits of the distribution (99.63% coverage). If only mismatch is being calculated, then nominal value is zero. Industrial design practice uses percentage variation with respect to insertion delay to characterize variations. X-axis values have been normalized with the largest insertion delay for the chain taken as 1 and y-axis values calculated for normalized insertion delay to preserve the shape of graph. The normalization procedure can be seen in Figure 4-1 where maximum x-axis value was taken as 1 and the y-axis multiplied by same factor. The percentage values in y-axis are only representative and do not have any absolute significance. However, the relationship between different curves in the graph is maintained.

4.3.2 Local random mismatch characterization

There are two approaches to characterize local random mismatch using Monte Carlo simulations. First is a full Monte Carlo (MC) simulation with global and local variations, where mismatch effect is extracted by differentiating delays between two equal paths in the same run, one with mismatch activated and the other without. Equal impact of global variations cancels out in the difference leaving only

mismatch. Second approach is to simulate only mismatch on a timing corner. Any variation of delay from one run to next is a result of local variations. The advantage is faster simulation time and lesser resources due to reduced circuit size and lesser number of varying parameters. Subtracting the nominal from measured value for each run gives mismatch.

A full statistical model with global and mismatch variations can give a smaller value of standard deviation due to averaging effect caused by reduced mismatch on faster samples. This effect can be seen in Figure 4-1 where mismatch on corners (MM@SS, MM@FF) bounds the upper and lower limits and full Monte Carlo mismatch (MM@MC) lies in between.

We measured the impact of mismatch on insertion delay, skew, and pulse-width, while varying supply voltage, slew rate, drive strength, cell types, traditional and cross corners (SF, FS), and path depth (up to 60). These parameters and measurements give us an idea of the compromise between power, delay, and area that determine the optimum PPA (performance-power-area) point in a design.

4.4 Computational systems

Monte Carlo simulations are computationally heavy. To minimize the simulation time, we distributed runs on a server farm consisting of many machines. To verify the integrity of results, we compared a simulation on server farm and single machine and found no difference in result. The speedup was directly correlated to number of machines used and allowed us to do very large simulations.

4.5 Wave model

In sub-100nm technologies, the interconnect resistance is of the same order as the gate output resistance and wire capacitance dominates the gate capacitance [69]. As such, the cell output waveforms cannot appear as saturated ramps like sine wave that have a curved waveform only when the output is close to saturation. A realistic waveform is closer to a two pole saturated exponential as shown in Figure 4-2 that has curved waveform both at the beginning and close to saturation. For a unit saturated exponential input given by equation (4-1), the two pole saturated exponential waveform is given by equation (4-2) [108]. The analytical model is not feasible to use in spice. The solution was to have an equivalent RC circuit with a modified input wave controlling a Voltage Controlled Voltage Source. The RC values and input wave were calculated by applying Newton-Raphson method on these equations to obtain a converging solution. $Slew_{Input}$ for a waveform is defined as the delay from 20% to 80% of maximum amplitude.

$$v_{input}(t) = V_{dd}\left(1 - e^{(-t/Slew_{input})}\right) \tag{4-1}$$

$$v_{output}(t) = V_{dd}\left\{1 - \frac{\left[Slew_{input} * e^{(-t/Slew_{input})} - RC * e^{(-t/RC)}\right]}{Slew_{input} - RC}\right\} \tag{4-2}$$

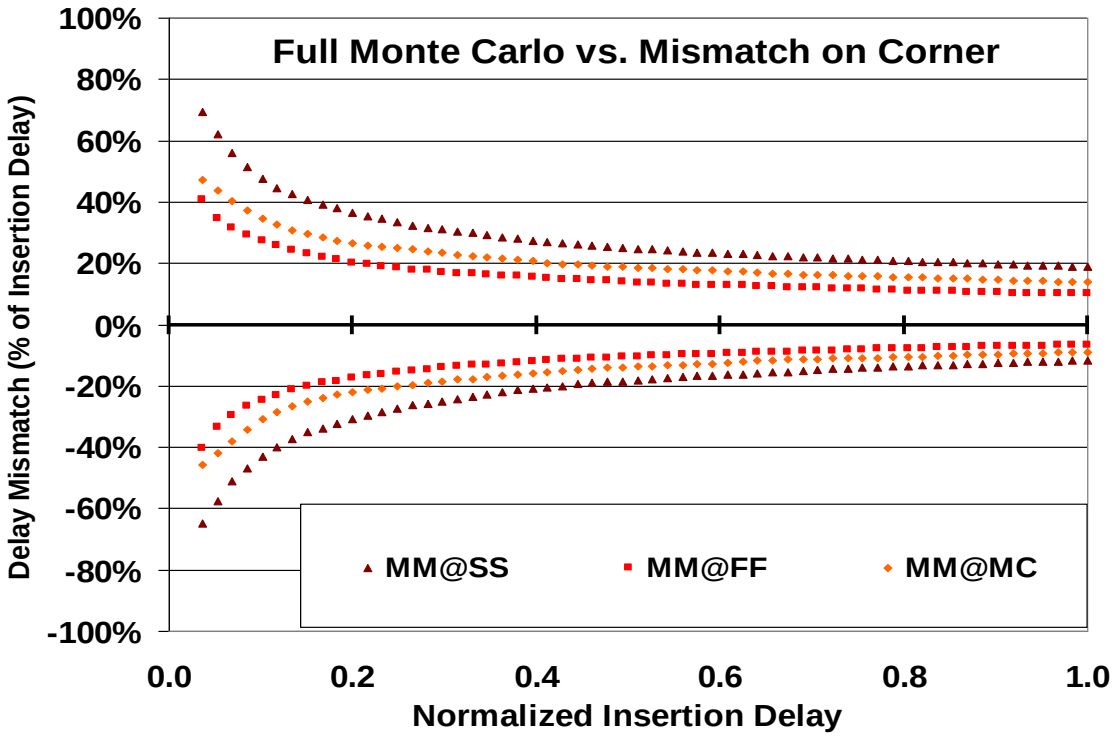

Figure 4-1: Full Monte Carlo Mismatch vs. Mismatch on Corners

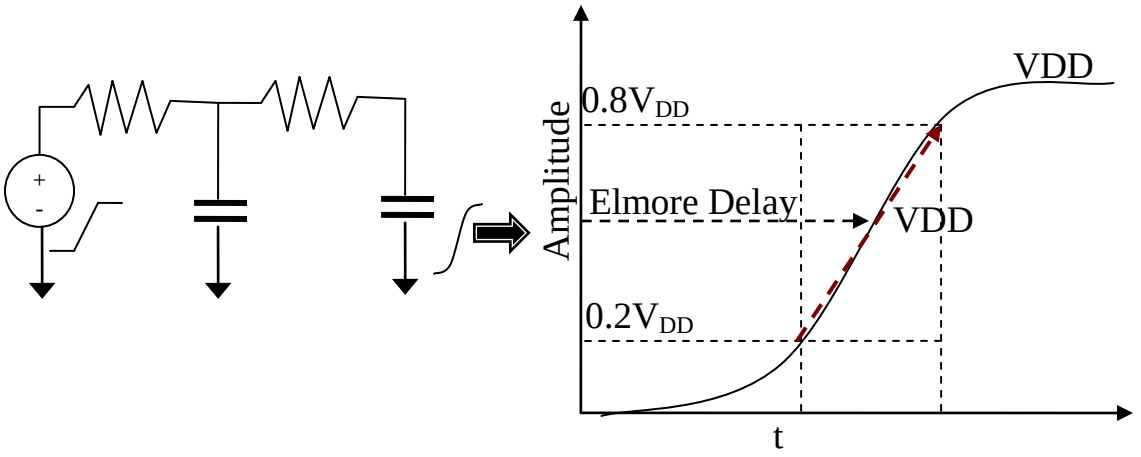

Figure 4-2: Distributed RC network output-saturated exponential waveform

4.6 Slew degradation in RC network

In sub-100nm, wire resistance & capacitance have a large impact on signal during transmission. Slew degrades as the signal propagates in the wire and is worse for highly resistive nets. Clock tree uses large buffers to drive big interconnects having high resistance. To have a realistic circuit, it is important to consider the impact of slew degradation in the RC interconnects. Clock trees are designed with a maximum slew limitation for a clock signal arriving at the input node of any buffer. Thus, we

designed our circuits with this slew limitation. The challenge is to ensure the arrival slew that depends on fanout load on previous buffer, drive strength of previous buffer, interconnect capacitance and resistance and input capacitance of load cell for a given supply voltage and temperature. To emulate the slew degradation, we wrote a program to reverse calculate the output slew of previous buffer and RC values from arrival slew, drive strength of first buffer and input capacitance of load buffer. Using library characterization tables (delay as a function of input slew and output capacitance) of these cells, we used a convergence algorithm to find the appropriate RC values. We used a single π-type structure for each interconnect to limit the complexity. As we do not consider interconnect variations and only take the worst-case RC values, it gives acceptable results. We used a resistance to capacitance ratio extracted from industrial 45nm process for routing interconnects.

Figure 4-3 shows typical interconnect model between two connecting cells. Input signal is applied at Point A, thus defining the input slew. In clock networks, the important consideration is arrival slew and thus we are considering Point C as the output port (that will function as input port to next cell) and calculate the output slew at that point. Point B forms the intermediate port. Figure 4-4 shows a typical timing table to calculate output slew for a given standard cell for given PVT conditions. The two input axes are formed by input slew (at Point A) and output capacitance (at Point B). The aim is to obtain a given output slew (at point C) for a given input slew (Point A). If the network had been purely capacitive, slew at point A & B would have been same. However, the resistance between these two points cause slew degradation. Thus, we need to reverse calculate an output slew at Point B from a given Output slew at Point C and Input Slew at Point A. The solution to this system can be arrived at through a converging algorithm described in Figure 4-5. The error in output slew at Point C was less than 10% on comparing with spice simulations for the calculated RC values.

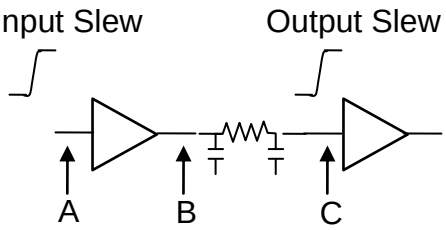

Figure 4-3: Typical cell and interconnect model for slew analysis

Output Capa (pF) Input Slew (pS)	1	10	50	100	200
1	5	30	100	125	250
10	12	45	125	150	280
50	20	55	140	165	305
100	40	65	150	175	325
200	50	80	160	180	340

Figure 4-4: Typical standard cell timing table to calculate output slew (pS) for given PVT condition

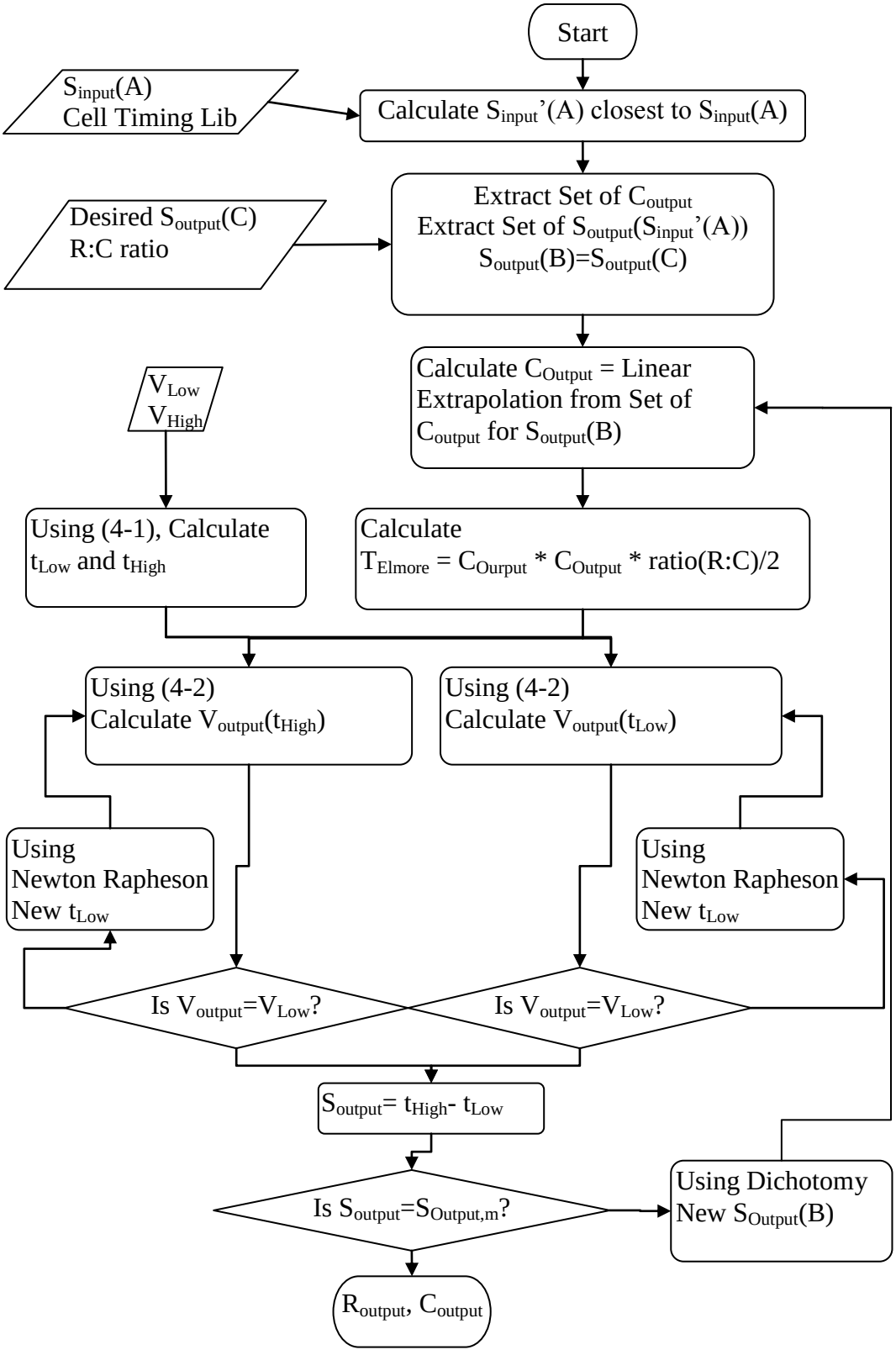

Figure 4-5: Flowchart to calculate required interconnect RC for a given input and output slew considering the slew degradation in interconnects

4.7 Automation scripts

Writing spice files by hand can be a very time consuming and error prone method. As such, we wrote our own compiler that takes an input format describing the overall structure/schematic of the circuit in minimum words and then transforms it into a spice netlist. It can reduce multiple reoccurring cells to a single line; create proper fanout connections without any need for extra lines. It automatically inserts interconnects, references proper libraries, connect pins to proper targets, etc. It can reduce a big spice file to just a few lines. As our test structures deal with a lot of regular formations, it helped to reduce and streamline the effort. The compiler can also take a library and generate test circuits to measure mismatch on single cells with minimum human intervention. The architecture allowed us to easily expand our targets and concentrate on testing optimization strategies rather than spending time on creating spice files. A simple example for the input model in the form of a tcl list is given below. The structure expands into a path branching out into two separate paths using respective interconnect loads. The branches have a depth of 60 cells with fanout of 2 and 1 respectively. It also assigns the starting conditions, input waveform, and parameters to extract. The list can be easily modified for different cells and structures.

```
list   \
      {cell CELL_A   structure single   load 0.05   resistance 50  from A  to Z
supply_value {vdd *}   supply_node {vgnd vgnd vvdd vvdd}   net drv    instance
driver   stimuli_height  vdd   stimuli {{0.5 R} {3.0 F} {6.0 R}}  stimuli_style {
dsm 0.055 0.7 1000.0 }    iccond { driver:A 0 } } \
 { branch branch_1 } \
      {cell CELL_B  structure tree   depth 60   fanout 2   load 0.01   resistance 5
from A   to Z   supply_value {vdd *}    supply_node {vgnd vgnd vvdd vvdd} net
net1   instance cell1   extract { delay { {rise driver:Z rise Z 1 1 20.0 0.0} {fall
driver:Z fall Z 1 1 20.0 0.0} } } } \
{ branch branch_1 } \
{ branch branch_2 } \
       {cell CELL_C structure tree depth 60 fanout 1 load 0.02 resistance 20 from
A   to Z   supply_value {vdd *}   supply_node {vgnd vgnd vvdd vvdd}   net net2
instance cell2    extract { delay { {rise driver:Z rise Z 1 1 20.0 0.0} {fall driver:Z
fall Z 1 1 20.0 0.0} } } } \
{ branch branch_2 }
```

Handling a large amount of results is very difficult. For that reason, we created scripts to convert the datasets into statistical numbers easily importable into graphing tools. These scripts parsed the data and extracted useful information predetermined by us. The extracted data was analyzed for statistical information. Furthermore, we created macros to create general model of each graph that helped us to obtain similar and comparable graphs having same properties allowing for easier identification and analysis.

4.8 Metrology

Waveforms today do transit linearly and have a more exponential part in beginning and end as shown in Figure 4-2. Thus, the transition time of the waveform is calculated from 20% to 80% of maximum magnitude for rise time and 80% to 20% for fall time. The number is a representative figure and gives an idea about how fast the signal inverses.

Transistor threshold voltage and intrinsic delay has been falling. As such, even before an input wave may reach its maximum amplitude, the output wave start to invert. Thus, the delay calculation use 40% & 60% as threshold levels for a rise & fall edge respectively. For an inverter with rising input and falling output, delay is calculated from when signal reaches 40% at rising input to 60% at falling output ($\mathbf{d_{r->f}}$ in Figure 4-6). For an inverter with falling input and rising output, delay is calculated from when signal reaches 60% at falling input to 40% at rising output ($\mathbf{d_{f->r}}$ in Figure 4-6).

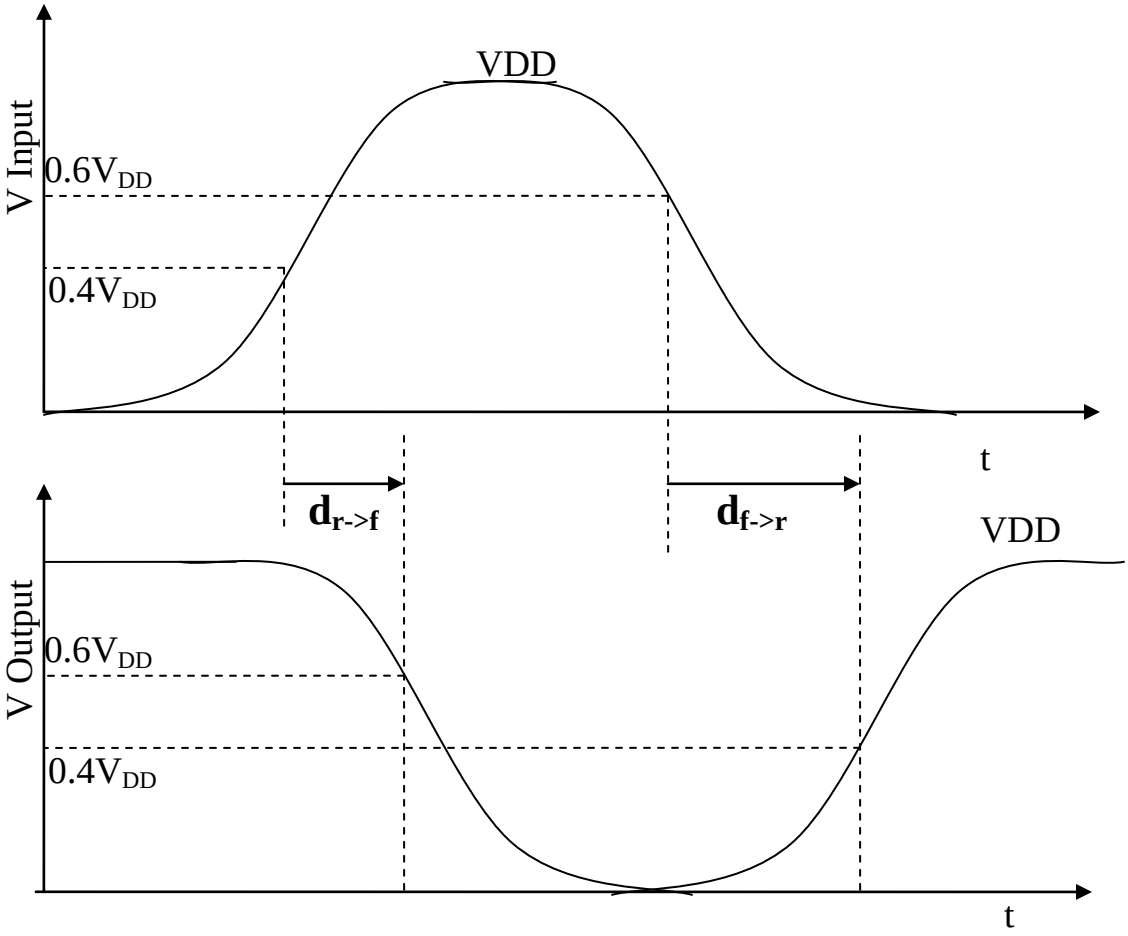

Figure 4-6: Inverter rise & fall delay

4.9 Setup for die-to-die NMOS-to-PMOS mismatch

The purpose is to study the impact of global mismatch on clock cells to understand what configurations are more susceptible to it and thus find optimization solutions that can be implemented without sacrificing on clock speed or design size. Being a die-to-die phenomenon, any improvement in a cell will have a proportional improvement in the chain.

We created a standard setup used for all test cells to maintain consistency in results. The setup consists of a test cell in a clock path connected to other buffers through interconnects as shown in Figure 4-7 to create realistic signal conditions. The arrival signal is affected by previous logic gates and as such changes with PVT conditions. The same drive and load buffer were used for all cells. Interconnect load was taken to get 55ps rise time at input node of load cell at worst (SS) corner, 1.05V, and -40°C. The given PVT condition specifies a common worst-case delay scenario in many ASIC chips for 45nm. Slew degradation due to RC load was considered in calculations to achieve the arrival slew. The interconnect load to resistance ratio was extracted from routing interconnects in the 45nm design detailed earlier and thus provides a comparable reference of path depth to design size. The work is based on spice simulations using production models.

The input signal is a realistic waveform with equal rise and fall delay. Calculating the difference in propagation delay for each edge between the input and output of the test cell gives us the impact on pulse-width (or duty cycle). The simulations were conducted on the cells in a 45nm standard cell clock library that includes buffers, clock gates, combinational cells, flip-flops, etc and their different drive strengths. Drive strength is a better measure than cell size as it is understood by a designer and gives us an idea about the given technology flavor.

We measured the impact for different configurations by varying various parameters including process, temperature, voltage, and slew for different test cells. Presence of two buffers before the test cell and a load cell ensures a realistic waveform. Slew value specifies the arrival slew at the input node of test and load cell at 1.05V and -40°C. As the temperature and voltage changes, the slew will change also. Clock networks are designed for maximum arrival slews and thus we took the same approach to measure global mismatch impact.

Process: SS (slow-slow), FF (fast-fast), TT (Typical), SF (slow-fast) & FS (fast-slow)

Temperature: -40°C, 25°C & 125°C

Voltage: 0.85V to 1.30V in steps of 50mV

Worst-case slew at arrival node=20ps, 55ps, 100ps, 150ps.

Test cells:

BF (X1 to X6): Clock buffer drive strength from 1 to 6 (normalized)

INV (X1 to X6): Clock inverter drive strength from 1 to 6 (normalized)

DBF (2Y1 to 6Y3): Delay buffer with 2/4/6 inverters for drive 1 to 3 where Y represents the drive strength.

CG (X1 to X6): Clock gate drive strength from 1 to 6 (normalized)

AND, NAND, OR, NOR, XOR, Flip-flop, MUX: Varying number of inputs for different cells with various drive strengths ranging from X1 to X6.

We measured impact on high pulse-width (or duty cycle or duration of '1') at each combination of inputs. The resulting change in pulse-width (output pulse – input pulse) has been normalized with worst-case delay of a standard buffer. Positive value signifies an increase in pulse-width and negative value signifies a decrease in pulse-width. The net impact on clock period is zero and thus, impact on low pulse is opposite to high pulse.

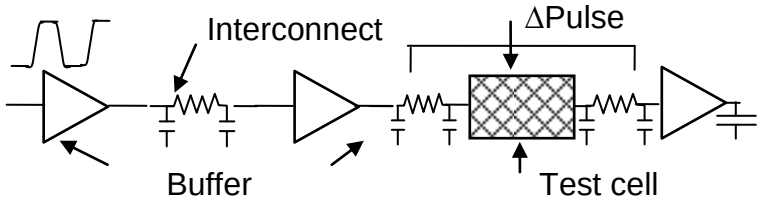

Figure 4-7: Setup to measure impact of die-to-die n-to-p mismatch on pulse-width

4.10 Setup for within-die local random mismatch

The objective here is to study the impact of local mismatch on clock paths to determine the sensitive configurations and the basis of their susceptibility. It will allow us to realize the optimization solutions that can be implemented without sacrificing on critical parameters. Path delay mismatch can be reduced either by improving path configurations or by improving cell delay mismatch. For that reason, we looked at both cell level mismatch and path level mismatch to find a combination of optimization methods implementable in design.

4.10.1 Cell level analysis

We created a standard setup used for all test cells to maintain consistency in results. The setup consists of a test cell in a clock path connected to other buffers through interconnects as shown in Figure 4-8 to create realistic signal conditions. The arrival signal is affected by previous gates and as such changes with PVT conditions as well as is affected by local mismatch on previous logic gate. A perfect input slew underestimates the impact of local mismatch. The same drive and load buffer were used for all cells. Interconnect load was taken to get a specific rise time (55ps by default) at input node of test and load cell at worst (SS) corner, 1.05V, and -40°C. The given PVT condition specifies a typical worst-case delay scenario. Slew degradation due to RC load was considered in calculations to achieve the arrival slew. Equivalent slew for same cell type for different drive strengths produces equivalent delay. The interconnect load to resistance ratio was extracted from routing interconnects in the 45nm design detailed earlier and thus provides a comparable reference of path depth to design size. The work is based on spice simulations using production models.
Delay between input and output node of the test cell is calculated at each sample point of a Monte Carlo simulation with mismatch activated at a given corner. Delay difference with nominal delay calculated at same corner but without mismatch gives

us the impact of local mismatch for that test cell. Using statistical calculation on the data set, we extracted the nominal delay M, mean shift μ, and standard deviation σ of resulting distribution. Minimum and maximum statistical delay for a test cell at a given corer in presence of mismatch is specified by M+μ-3σ and M+μ+3σ respectively.

The simulations were conducted on the cells in a 45nm standard cell clock library and logic cell library that includes buffers, combinational cells, flip-flops, etc and their different drive strengths. We measured the impact for different configurations by varying various parameters including process, temperature, voltage, and slew for different test cells. Presence of two buffers before the test cell and a load cell ensures a realistic waveform. Slew value specifies the arrival slew at the input node of test and load cell at 1.05V and -40°C. As the temperature and voltage changes, the slew will change also.

Technology: 65nm (only where mentioned), 45nm (default)

Process: SS (slow-slow), FF (fast-fast), TT (Typical), SF (slow-fast) & FS (fast-slow)

Temperature: -40°C, 25°C & 125°C

Voltage: 0.85V to 1.30V in steps of 50mV

Worst-case slew at arrival node=20ps, 55ps, 100ps, 150ps.

Threshold voltage: LVT (low), SVT (standard), HVT (high)

Gate length: Standard, Large L

Test cells:

BF (X1 to X6): Clock buffer drive strength from 1 to 6 (normalized)

INV (X1 to X6): Clock inverter

DBF (2Y1 to 6Y3): Delay buffer with 2/4/6 inverters for drive 1 to 3

AND, NAND, OR, NOR, XOR, AND-OR, etc: Varying number of inputs for different cells with various drive strengths ranging from X1 to X6.

Miscellaneous cells

We measured impact on rise and fall delay at each combination of inputs. The statistical extracted result has been normalized with worst-case delay of a standard buffer. Positive value signifies an increase in delay and negative value signifies a decrease in delay.

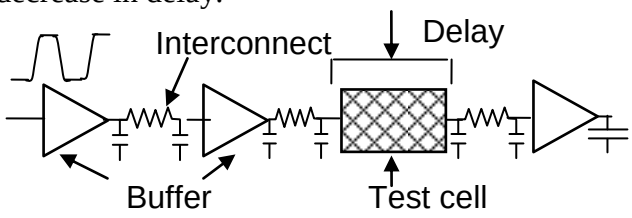

Figure 4-8: Setup to measure local random mismatch on rise and fall delays and transition time

4.10.2 Path level analysis

Path level analysis requires extracting the impact of local mismatch on delay, skew and pulse-width for different path depths under various conditions. The general setup for three cases is shown in Figure 4-9. In case of path delay or pulse-width, we require a single path whereas in cases of skew, we need both the paths. A typical

path consists of cells (same or different type) connected through RC interconnects. Path depth (or number of cells in the path) is about 60. Interconnect load was determined to provide a specific rise slew (55ps by default) at worst-case corner (SS, 1.05V, -40°C). PVT conditions for load determination for a specific slew can change in cases when required. The path replicates clock paths in a design and the traversing signal is affected by local mismatch at each logic gate. Slew degradation due to RC load was considered in calculations to achieve the required slew. Interconnect load to resistance ratio was extracted from routing interconnects in the 45nm design detailed earlier and thus provides a comparable reference of path depth to design size. The work is based on spice simulations using production models.

There are two approaches to characterize mismatch. First approach is a full Monte Carlo (MC), including global and local variations, where the mismatch effect is extracted by differentiating the delays between two similar paths, one with mismatch activated, and the other without. Because of the same signal and equal impact of global variations, the difference directly imparts the effect of mismatch. Second approach is to simulate mismatch only on a timing corner in a path with an advantage of faster simulation time and lesser resources. To characterize mismatch we subtract the nominal value of a quantity from its measured value in each MC sample. The resulting statistical distribution gives us the average value and standard deviation of mismatch impact.

Monte Carlo statistical simulation with global and local variations together can have a smaller distribution of local variations (σ) due to the reduced mismatch effect on faster samples whereas statistical mismatch on corners have a larger mismatch distribution (σ) due to worst corner delay. The same can be seen in Figure 4-1 where mismatch on corners (MM@SS, MM@FF) bounds the upper and lower limits and full Monte Carlo mismatch (MM@MC) lies in between. The numbers might vary with two approaches but the overall trends remain the same. Mismatch on SS corner is more pessimistic than full Monte Carlo mismatch whereas mismatch on FF corner is more optimistic. The value to be used will depend on the required assured yield that in turn depends on application. Memory requires high yield and thus mismatch on corner may be a better choice. A NAND chip can work on lesser yields and full Monte Carlo mismatch can be a better choice.

Technology: 65nm (only where mentioned), 45nm (default)

Process: SS (slow-slow), FF (fast-fast), TT (Typical), SF (slow-fast) & FS (fast-slow)

Temperature: -40°C, 25°C & 125°C

Voltage: 0.85V to 1.30V in steps of 50mV

Worst-case slew =20ps, 55ps, 100ps, 150ps.

Threshold voltage: LVT (low), SVT (standard), HVT (high)

Test cells:

BF (X1 to X6): Clock buffer drive strength from 1 to 6 (normalized)

INV (X1 to X6): Clock inverter

We measured impact on rise/fall delay, skew and pulse-width at each combination of inputs and plotted against insertion delay of the path. X-axis values have been normalized with the largest insertion delay (60 stages) taken as 1 and y-axis values calculated for normalized insertion delay to preserve the shape of graph. Each

distribution can be expressed in terms of nominal delay M, mean shift μ, and standard deviation σ with smallest and largest value being M+μ-3σ and M+μ+3σ respectively. We plotted the graphs using only μ±3σ values, i.e. only mismatch impact.

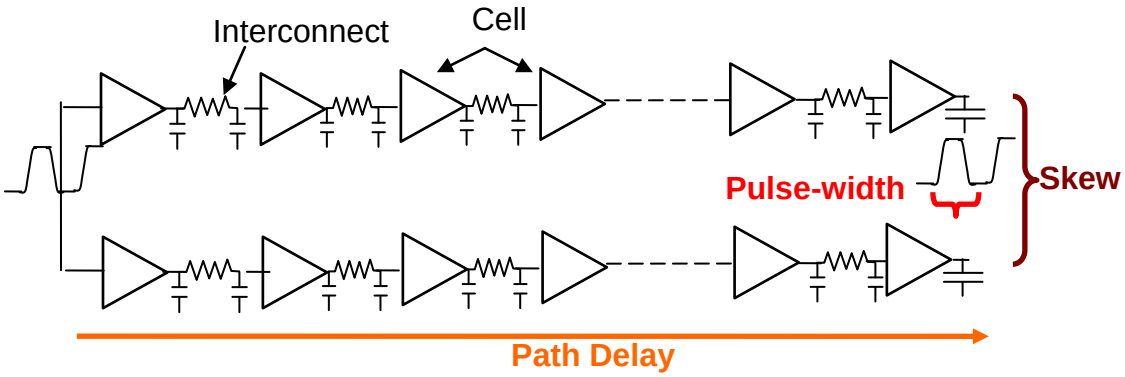

Figure 4-9: Setup to measure local random mismatch on path delay, skew and pulse-width

5 Impact of and Design Solutions for Die-to-Die NMOS-to-PMOS Mismatch

Traditionally, global variations have been the biggest factor in process variations. As the effect of these variations is same all over the die, balanced corners like SS (slow-NMOS, slow-PMOS) and FF (fast-NMOS, fast-PMOS) are sufficient to constitute the statistical limiting cases for delay. Global NMOS-to-PMOS mismatch [94] (or die-to-die NMOS-to-PMOS mismatch or global unbalanced variations) have always been present in CMOS fabrication process. Pulse-width scaling has always lagged behind transistor intrinsic delay scaling in ASIC designs due to stable clock latencies caused by increase in wire resistance, larger logic content, and increasing margins. However, its increasing relative importance has become a critical factor now [94].

Global mismatch primarily affects pulse-width or the clock duty cycle. As the path length increases, effects of global mismatch increases. Consumer demand has been growing for more functionality in a product that requires higher complexity requiring bigger chips. A synchronous design requires same clock to drive most of the logic and thus the length of clock paths increases along with the chip size. Thus global mismatch have become an important factor to be taken into account.

The following chapter studies the impact of global mismatch on digital circuits by looking at the impact on single cells. Global mismatch being die-to-die variation, we can extrapolate the results to a chain of cells. We looked at susceptible configurations and formulated different approaches to make design more robust without sacrificing performance. Optimization methods targeting specific parameter or design use are given considering pulse-width.

5.1 Origin

CMOS device fabrication requires multiple steps. Most of the steps are common between NMOS and PMOS devices that create an excellent correlation for the parameters affected. For example, photolithography step is common for both devices and thus create a good correlation in gate length variations. Such variations are responsible for balanced variations with extreme corners being SS & FF as shown in Figure 1-1. However, to create different majority charge-carrier regions, different doping steps are used that allows for statistical variability of average doping levels in NMOS and PMOS devices. The doping level for all transistors of same type on a die is same but is uncorrelated to doping level of the other device type and varies from die to die. Such variations can create unbalancing in device characteristics and are responsible for deviations from straight line in NMOS to PMOS transistor I_{ON}-V_G curve in Figure 1-1. The extreme cases can create corners like fast-NMOS & slow-PMOS (FS) or slow-NMOS & fast-PMOS (SF). The impact of doping variations is predominantly on threshold voltage and mobility.

5.2 Effect on design

Global mismatch impacts insertion delay along a path but the variation is less than traditional limiting cases (SS, FF). Reason being, a signal passing through a path

will see at least one transistor type whose delay is better than the worst case or worse than the best case. Thus, the value addition of doing maximum delay analysis for global mismatch is minimal. Our experiments have shown that intrinsic delay of cells under SF & FS corners is much closer to TT corner than SS or FF corners. The same is not true when considering pulse-width or duty cycle that can be defined as the difference in insertion delays of two opposite and consecutive edges passing through the same path calculated from a single point of time. Ideally, the rise and fall delays should be exactly equal. However, the two edges pass through same cells but opposite transistors, as shown in Figure 5-1, and thus the impact on their end delay is different. The two delays vary independently of each other to some extent. Being a global effect, all instances of a cell in clock path will have the same pulse behavior (decrease or increase), thus aggravating the arrival pulse-width. Figure 5-2 shows the variation of duty cycle along a path caused by global mismatch (SF & FS). It may be noted that even the balanced corners like SS, FF & TT have some effect on duty cycle typically because the rise and fall edge see different amount of drive currents creating a difference in rise time and fall time that affects their respective delays. Although duty-cycle and pulse-width have different values, their change represents the same quantity.

Externally, flip-flops may be activated by either rise edge or fall edge but internally both the edges are required to manage the data shift. Any change in duty cycle will decrease the available time-period for one of the two stages that in turn can affect setup or hold time. Thus, each flip-flop has a Minimum Pulse Width (MPW) constraint, i.e. the minimum amount of time required in each stage for successful data shift. Global mismatch reduces either the high pulse or the low pulse and for long paths can violate the MPW constraint. To keep the violations in check may require smaller clock paths affecting design size or a lower clock frequency affecting product performance.

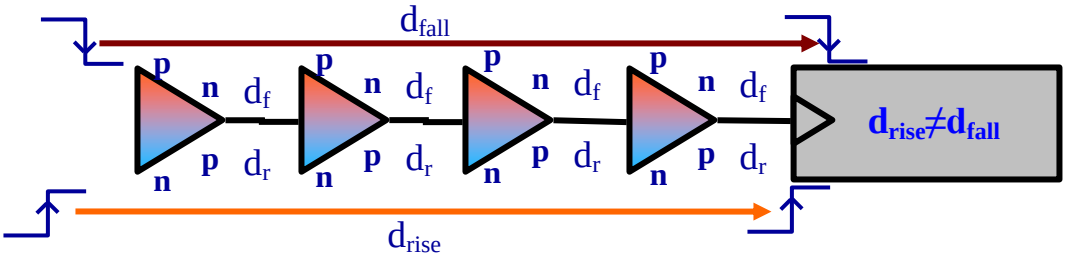

Figure 5-1: Rise and fall delay in presence of die-to-die n-to-p mismatch in a clock path

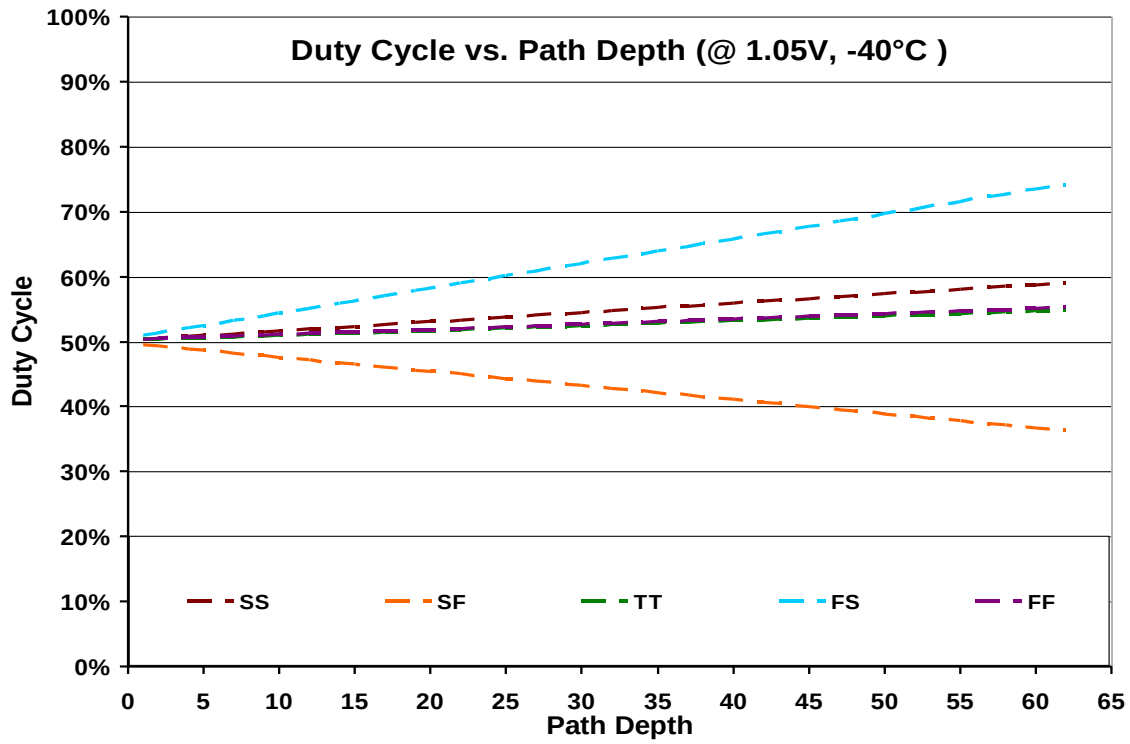

Figure 5-2: Effect of die-to-die n-to-p mismatch on clock duty cycle in a chain

5.3 Clock cells vs. logic cells

Clock and logic cells are made up of same transistors but they differ in their purpose and thus their construction. Clock cells are designed to maintain the pulse-width i.e. equal rise and fall delay and are thus called balanced cells. They use an inverter P/N ratio between 2:1 and 3:1 to balance the rise and fall delays taking into account their respective mobility [53]. Logic cells on the other hand are designed to minimize the average cell delay by using a smaller P/N ratio to reduce the input load significantly while only somewhat slowing the rising output. Thus, the average delay of a logic gate decreases, though the rise and fall times become unbalanced. The best inverter P/N ratio for logic cells to minimize average delay is between 1.4:1 and 1.7:1 [53]. For a balanced corner, clock cells will typically maintain the pulse-width while logic cells will not.

Clock cells are typically balanced for worst-case condition through PMOS to NMOS sizing as well as output to input stage sizing. Majority instances of clock cells are there to drive the clock and clock buffer constitute the most important. Clock cells are sized to have least input to output delay while in balanced condition. The input stage is small to reduce the input capacitance thus allowing for higher fanout and output stage are bigger to allow for larger drive current. According to principal of Logical Effort [53], the output to input stage ratio between 2.7 and 3 provides the least parasitic delay and thus the best sizing ratio.

Transistor drain current scales non-linearly to some extent with device size. Moreover, the scaling of NMOS and PMOS are not exactly same. These two trends

can affect the behavior of two different drive buffers. Buffer drive represents the output drive current that determines size of output stage and consequently input stage transistors.

5.4 Analysis & Inferences

5.4.1 Clock buffer

Figure 5-3 demonstrates the change in pulse-width for a cell (BFX1) at different supply voltages and for four major corners-SS, FF, SF, & FS. As seen in the given figure, unbalanced corners make up the worst-case for pulse-width in most cases as expected. From best-case to worst-case, there is a four-five times degradation in pulse-width. A design working at 1GHz frequency at maximum voltage will have to work below $1/5^{th}$ frequency i.e. 200MHz at minimum voltage to work properly. Whereas low voltage operation generally has low load, an upper limit to frequency can put constraints on design applications. For example, a mobile microprocessor may easily have clock frequency lower than 200MHz in standby mode but a laptop microprocessor will require more than that in low-power mode. Moreover, the impact may increase in next node that will be working at even higher frequencies putting a greater strain on pulse-width.

For a two-stage buffer (Figure 5-4), the 1^{st} stage is smaller than the 2^{nd} stage to keep low input capacitance and higher output drain current. Typically, balanced corners should maintain same rise & fall delay and thus maintain the pulse-width even though the intrinsic delay is higher. Unbalanced corners should slow down one edge compared to other but in a two stage buffer, each edge pass through both NMOS and PMOS transistors and thus should maintain the pulse-width. However, the difference in size of 1^{st} and 2^{nd} stage becomes a factor where the effect of slow corner is worse for smaller transistors in 1^{st} stage. Thus, the 1^{st} stage is responsible for most of the pulse change in a cell. The 2^{nd} stage should reduce the impact to some extent. However, the transistors are larger in 2^{nd} stage and thus have a smaller impact. The difference of mobility between NMOS and PMOS also differentiates the required device size to have similar rise and fall delay for slow NMOS and PMOS.

The pulse-width change for BFX1 in Figure 5-3 shows the SS corner to be the worst-case at 0.85V, which is not in line with our assumptions. However, it can be explained because of the small sized 2^{nd} stage. The small PMOS transistor is highly affected by slow corner as compared to the NMOS transistor. The behavior at low voltage changes as the drive strength increases. As seen in Figure 5-5, the higher drive strength buffer BFX2 has a much lower pulse-width impact at low voltage due to larger 1^{st} stage that is less susceptible to slow transistors. Transistors at low voltage have a low V_{DD}/V_{th} ratio keeping them in weak inversion for much longer. Moreover, smaller transistors have a much lower drain current due to low V_{DD} greatly increasing the time required to charge or discharge a load. All these factors point to a threshold limit for transistor size for a given technology below which the impact of slow corner is very high.

5.4.1.1 Increase in drive strength

As we increase the drive strength, above a certain limit there is a marginal improvement in pulse-width variations. The same can be seen in Figure 5-5 where BFX2 and BFX6 have almost same pulse behavior except for very low voltage where we see an inversion between SF and FS behavior. The increased drive strength also increases the power consumption that is a critical parameter at very low voltage design.

5.4.1.2 Temperature dependency

Figure 5-6 shows the impact of temperature on pulse-width in BFX1. There is marginal impact on FF corner for the whole range of voltage but SS corner is highly impacted at low voltage, as much as 33% difference between -40°C and 125°C. SF corner is more impacted at low voltage whereas FS corner is more impact (less than SF though) at higher voltage. Temperature inversion happens at low voltage although the point of inversion is a factor of specific corner. At low voltages, reduced current drive forces the transistors to remain in weak inversion for longer duration where the current-temperature relationship is opposite as compared to strong inversion [78]. In strong inversion drain current is made up of drift current while in weak inversion it is made up of diffusion current. An increase in temperature in strong inversion will increase the thermal agitation of electrons that hinders the drift current. On the contrary, an increase in temperature in weak inversion increases the average distance traveled by a charge carrier, thus increasing the diffusion current through concentration gradient. Longer a transistor stays in weak inversion more susceptible it is to have an overall inverse temperature behavior.

5.4.1.3 Impact of slew

Slew or transition time consists of another major factor in determining pulse-width as they directly affect the rise or fall delay. The difference in NMOS and PMOS current allows for different slews and thus the delay. Larger slew rates increase the time spent in weak inversion that combined with slow transistor can increase pulse-width variation. The effect of slew on pulse-width for BFX1 at different voltages and at SF & FS corner can be seen in Figure 5-7. We have plotted three different slews- 55ps, 100ps, & 150ps. To keep in mind, the given slew is at the input nodes of the driver and driven cell at 1.05V & -40°C. As the voltage or temperature changes, the slew may change accordingly. There was little change in pulse variation below 55ps.

As can be seen in Figure 5-7, increasing slew increases the pulse-width variation for unbalanced corners. The effect is pronounced for SF corner and is increased even more at low voltage. The fact corroborates with our earlier observations that smaller NMOS in 1st stage is highly sensitive at slow corner. For BFX1, 3-times increase in slew caused 5-times degradation in pulse-width [118]. Designs aimed mostly at low power operation may use higher slews at nominal supply. However, when working

in ultra-low power mode, the impact on pulse-width will be much higher that can constrain design usage.

Although the results clearly support using higher drive buffers, small drive buffers are still used in designs for all non-critical paths. Their smaller size gives a big advantage in terms of area. These non-critical paths can have large variations and under extreme conditions can become critical. As such, we kept the smaller buffers in our trials.

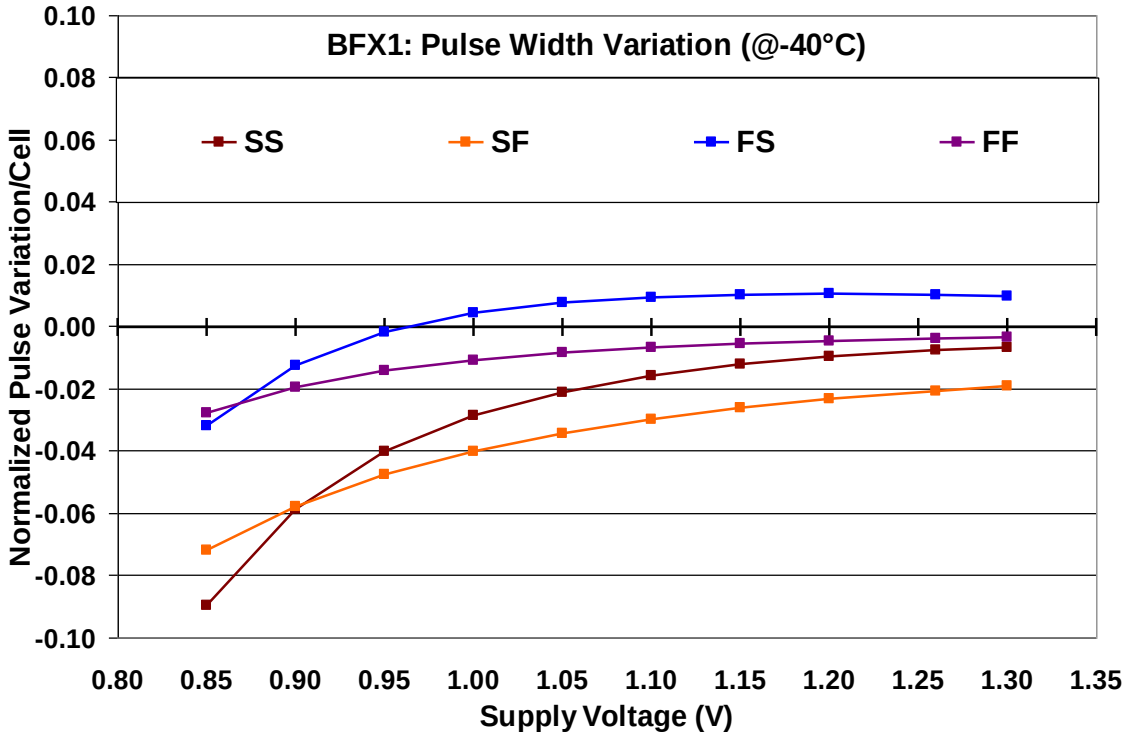

Figure 5-3: Buffer pulse-width variation for die-to-die n-to-p mismatch

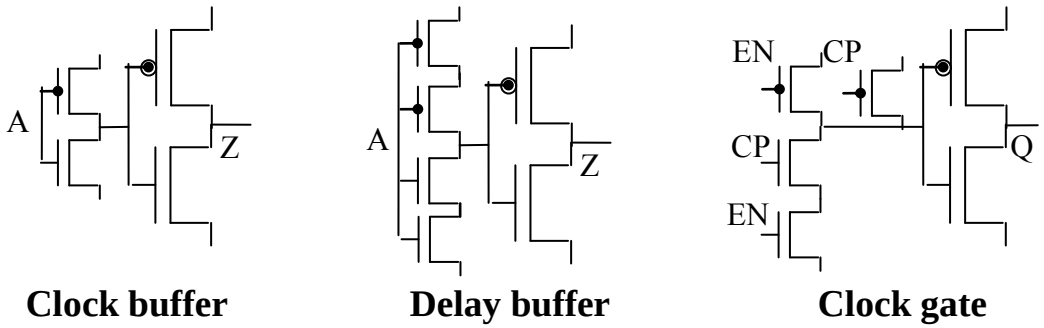

Figure 5-4: Schematics of clock buffer, delay buffer and clock gate

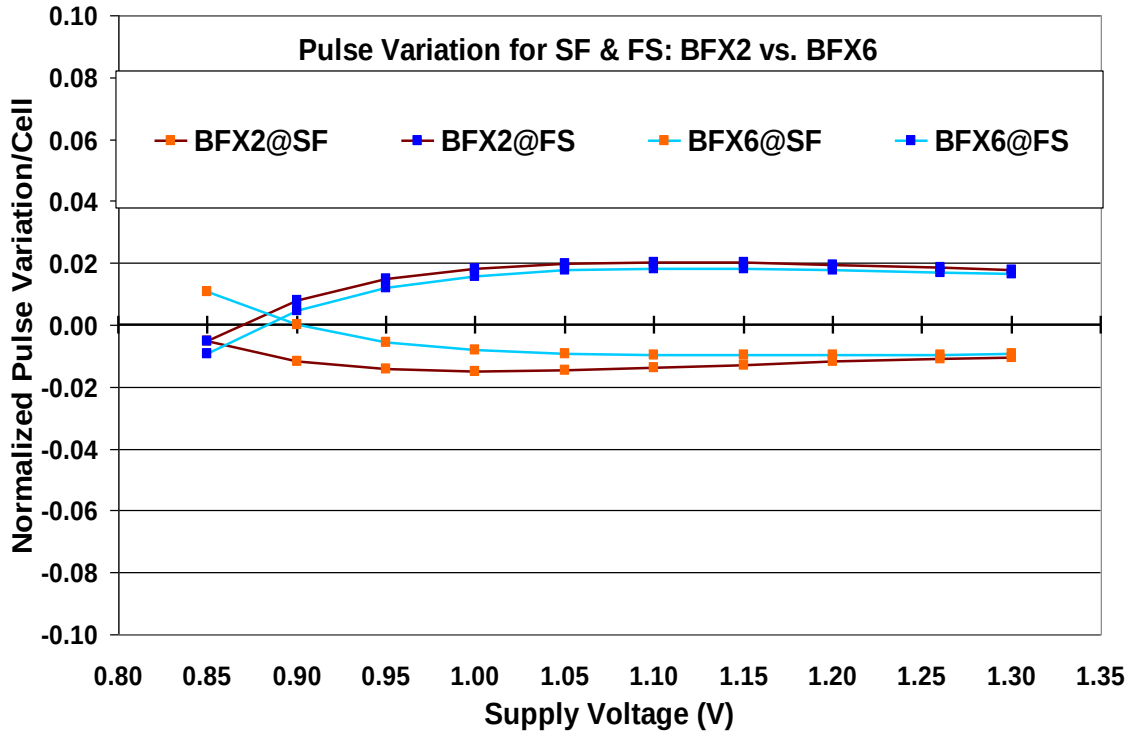

Figure 5-5: Impact of buffer drive on pulse-width variation for die-to-die n-to-p mismatch

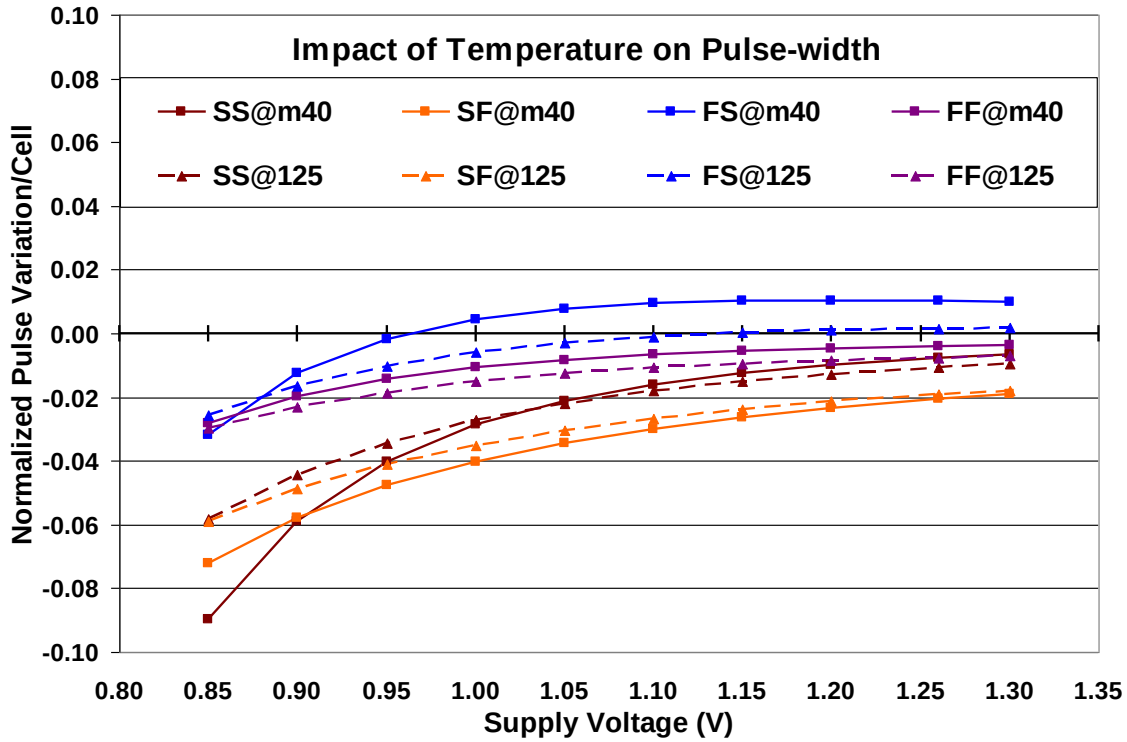

Figure 5-6: BFX1 Pulse-width variation due to die-to-die variations for different temperatures

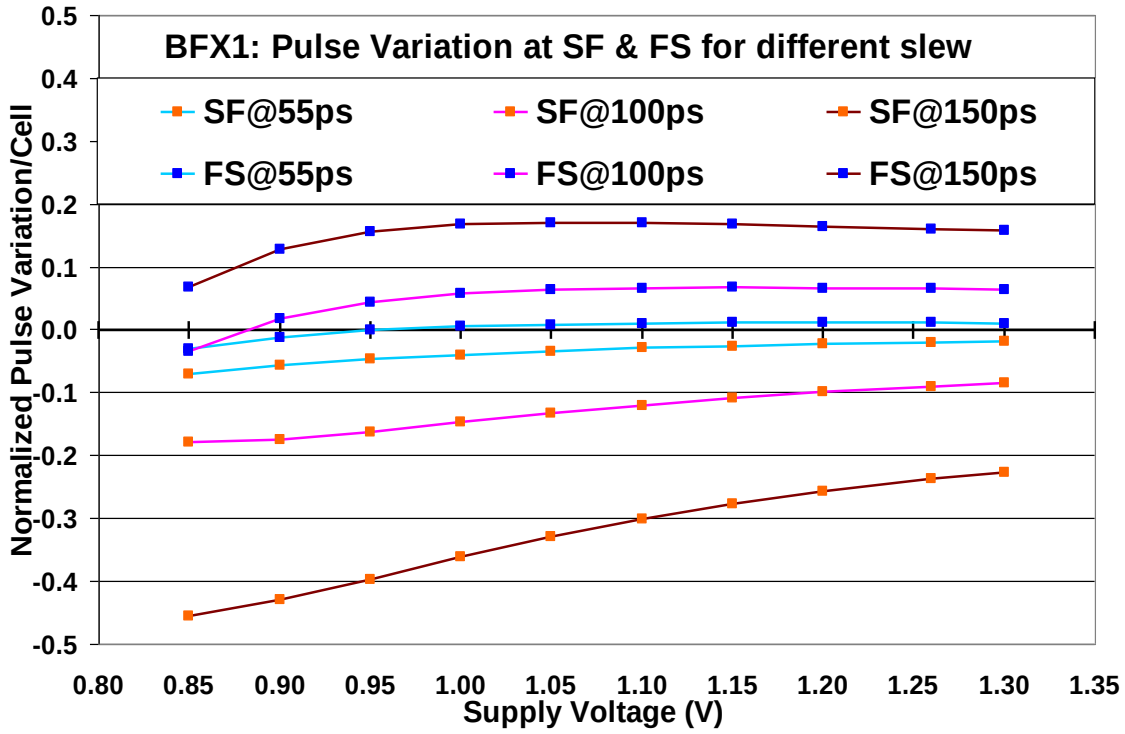

Figure 5-7: Pulse-width variation due to die-to-die n-to-p mismatch for different slews

5.4.2 Clock inverter

Technically a clock buffer is simply two inverters connected in series though the sizes of two inverters are different. A clock inverter has similar function as to buffer i.e. to drive the clock in addition to inverting the signal when required. Traditionally, clock inverters are not the preferred choice of clock driver as they have higher input capacitance and poor slew regeneration capacity in spite of their low intrinsic delay. Clock inverter output is highly correlated to its input and thus a difference in slew for rise and fall edge will manifest itself in the output. Miller capacitance effect [40] can increase the local variations of slew.

Clock inverter has a big advantage over clock buffer. Its pulse response is very symmetric as can be seen in Figure 5-8 and opposite in consecutive inverters. Two consecutive and same inverters in a chain negate any pulse-width variations and maintaining the duty cycle throughout. An odd number of inverters will have a pulse-width variation equal to a single inverter whereas an even number of inverters will have almost zero pulse-width variation. A clock inverter can be particularly useful in low voltage designs to limit pulse variation as well as temperature sensitive applications.

Inverters have an advantage of cell area over other cells. In addition, the efficiency of area utilization is highest in an inverter, i.e. poly covered area by total cell area. It is also a highly symmetric cell and has straight poly lines mostly. Such a regular structure helps to reduce systematic effects and increases cell matching.

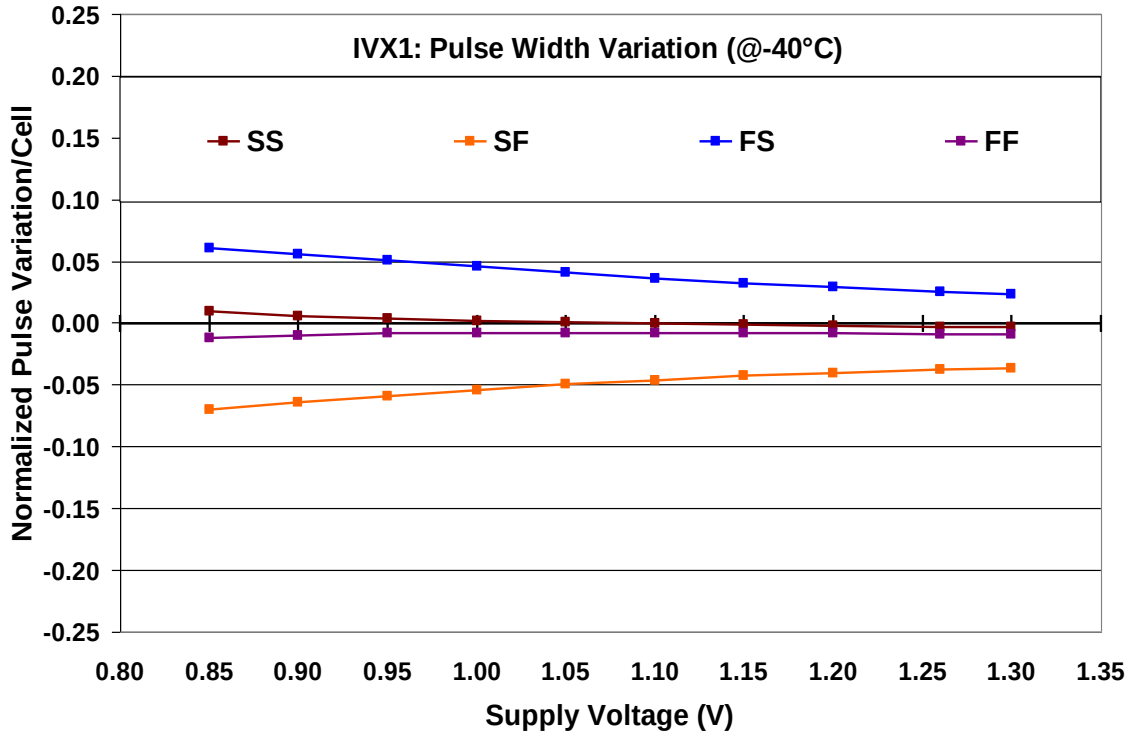

Figure 5-8: Inverter pulse-width variation due to die-to-die n-to-p mismatch

5.4.3 Clock gate

Designs these days utilize clock gates to reduce the non-essential circuit switching. A clock gate is typically an AND gate with one input being the clock signal and second being an enable signal that controls the output switching, as shown in Figure 5-4. The enable signal is usually '1' except when specifically made asked by control logic to shut down the clock signal. As shown in earlier chapter, clock gates are present mostly before the flip-flops and at higher levels of distribution hierarchy.

Clock gates have a dual functionality to minimize the average delay and maintain balanced rise and fall transitions at output node. The 1st stage of a clock gate is sized differently for 'high skew' or rising edge critical [53] to reduce intrinsic delay of enable signal. Such a configuration becomes more critical for pulse though and the same can be seen in Figure 5-9. The magnitude of pulse-width change is much higher as compared to a buffer for same drive strength. Lack of sufficient drive current also makes it more vulnerable to temperature variations at low voltage. Pulse variation at low voltage for clock gate is almost 2.5 times that of buffer. Although very few clock gates are present in a path, they can have a large effect on pulse-width. Unlike buffer, clock gate behavior does not change with drive strength at low voltage due to its unique functionality and even the magnitude change from CG1X to CG6X is only 25%. Clock gates are quite large generally due to their extra enable signal logic.

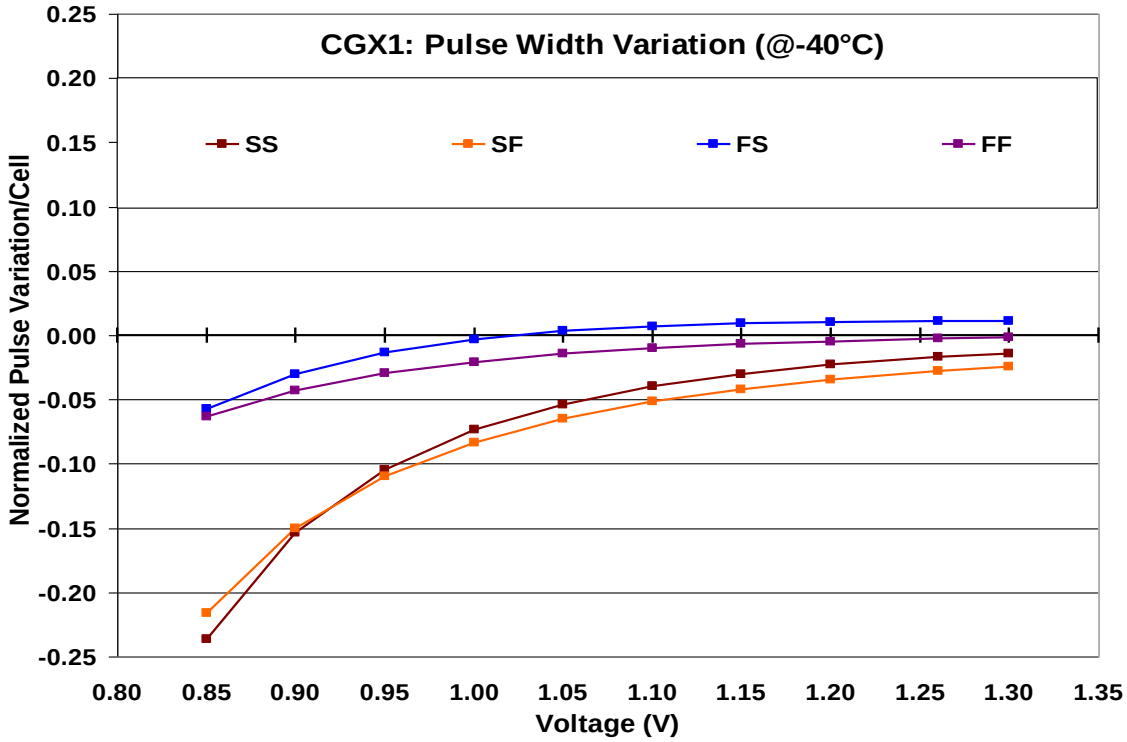

Figure 5-9: Clock gate pulse-width variation due to die-to-die n-to-p mismatch

5.4.4 Stacked logic gates

Stacked logic gates represent NAND, NOR, AND, OR type cells that have at least two series connected NMOS or PMOS. These logic gates are used for pulse shaping & frequency division. They are generally sized to minimize average delay in input and balance output. These cells are not used to drive clock and have few instances in a clock tree. As seen in Figure 5-10, the magnitude of pulse-width change is equivalent to that of a buffer. Therefore, unless the pulse-width change is very critical, there is little to be gained by optimizing stacked logic gates. However, stacks can allow for easier manipulation of pulse-width by changing PMOS and NMOS sizes in 1st stage. AND/OR gates have a preference over NAND/NOR as they do not invert the signal as well as have a higher output drain current.

5.4.5 Delay buffer

Delay buffer is a unique cell among all the logic gates. Generally, the objective is to minimize the intrinsic delay of the logic gate but delay buffers are used to increase the intrinsic delay. Delay buffers have normal inverters in the output stage and split inverters in all the previous stages as shown in Figure 5-4. The split inverters have their PMOS and NMOS divided into two transistors connected in series to increase the intrinsic delay. Delay can be increased by adding more number of split inverters. They are used to fix hold violations and can be used to reduce skew of a highly

unbalanced clock. Their unique characteristics justify their large pulse impact as shown in Figure 5-11. If not for their large insertion delay, they can be used to manage pulse-width. The difference between a 2, 4 & 6 inverter delay buffer is relatively small making a 6-inverter cell more attracting for same amount of delay.

5.5 Design impact of global mismatch

Let us take an example of how global mismatch can affect a clock path working at 200MHz at 1.30V & -40°C. The path is 15 levels deep with 2 clock gates, 2 AND gates, and 6 high drive buffers and 6 low drive buffers. The distribution is similar to what we saw in a real design. 200MHz frequency translates to 2500ps of pulse-width at 50% duty-cycle. Adding the pulse-width change for different cells, the arrival pulse-width can vary between 2260ps and 2780ps, i.e. between 45.2% to 55.6% duty-cycle. Such a path requires at least 5.6% margins on pulse-width at arrival flop. The path at 0.85V & -40°C will have a pulse-width variation of -1300ps that will translate to 43MHz maximum frequency at 5.6% margins. The calculations are just a representative of the effect. It should be kept in mind that other effects can come into picture like increased data path delay at low V_{DD}.

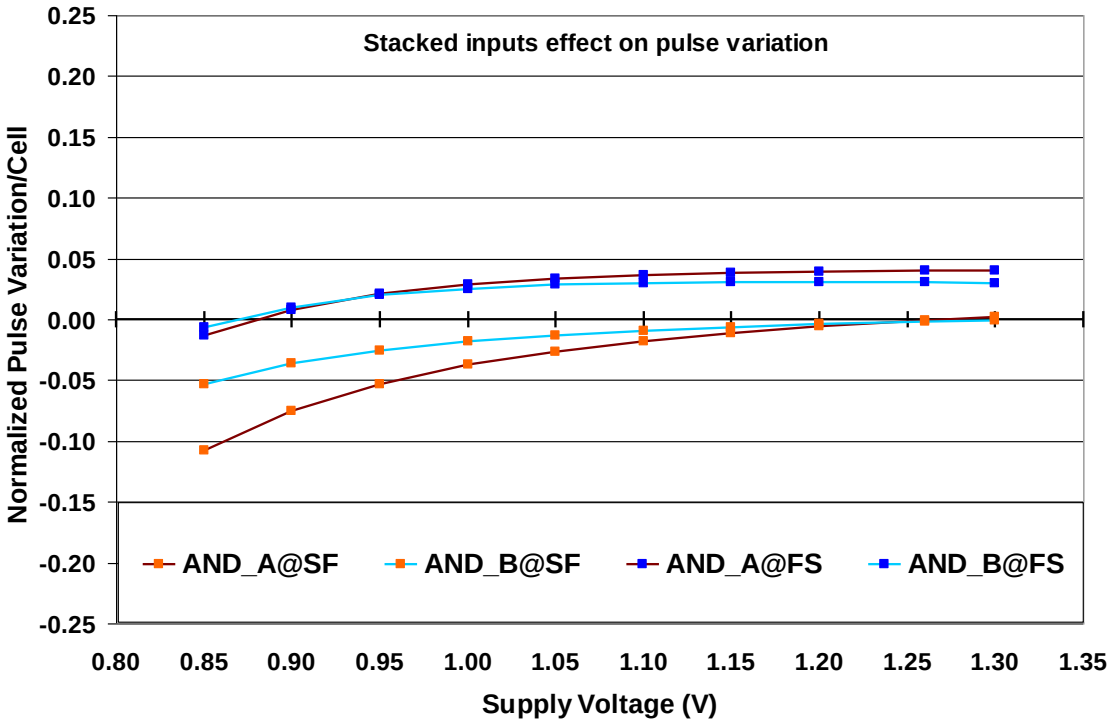

Figure 5-10: Difference between stacked inputs on pulse-width due to die-to-die n-to-p mismatch

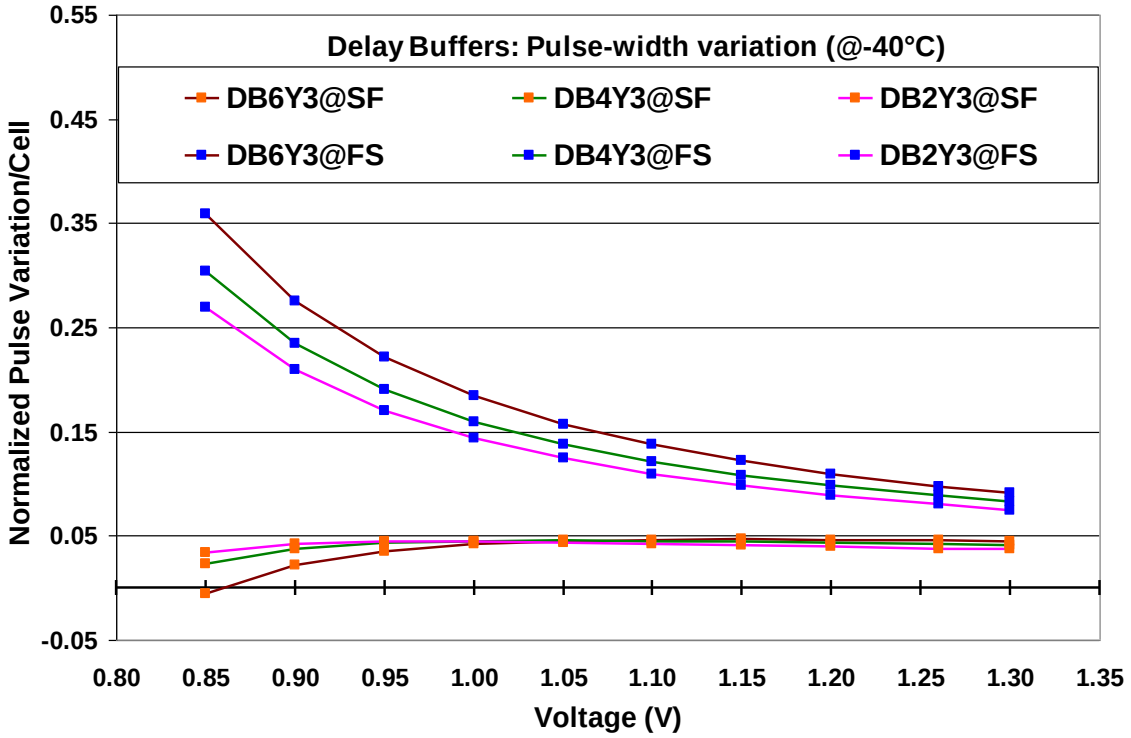

Figure 5-11: Delay buffers pulse-width variations due to die-to-die n-to-p mismatch

5.6 Optimization solutions

Global mismatch is a cause of concern for pulse-width in specific cases. Its impact is a big factor of cell schematic and transistor sizes. However, we can optimize the design for a reduced impact of global mismatch using different approaches. The optimization strategies are not stand-alone and can change the PPAY (Performance-Power-Area-Yield) point in the design space. According to the critical parameter in a design, the priority and method of optimizations will change. A power critical design cannot use high drive cells and will have to sacrifice on performance. Similarly, a parametric yield critical design may have to use larger cells and sacrifice on area.

5.6.1 Application specific unbalanced cells

One approach to reduce global mismatch effect is to re-look the design strategy. Currently, ASIC designs use 2-3 different libraries like high performance or low power. High performance library targets applications like laptop microprocessors. Low power library is calibrated towards a wide range of applications like automotive controllers, multimedia chips, smart phone microprocessors, etc. The aim is to reduce power consumption and provide performance only when needed. There can be sub flavors of a library based on different threshold voltages depending on their purpose but their application domain remains the same. A low

power library is created to work on a wide range of supply voltages and do not differentiate with specific application types.
We have proposed to create a subset of cells in a low power clock library based on their specific application types [119]. The subset of cells are those highly affected by global mismatch like clock gate or have a large number of instances in the path like clock buffers. We used a subset to minimize characterization effort and limit required design rules. As global mismatch has a major effect on pulse-width only, we limited the subset to clock library cells.
We differentiated the cells into three different application domains based on performance/power requirements- high performance with high V_{DD} (HP), low power with low V_{DD} (LP) and variable power and performance working at the whole range of V_{DD} (HPLP). In HP applications, e.g. digital TV processors, the chip requires high clock frequencies. In LP applications, e.g. low-end mobile phone processor, the aim is to reduce the power consumption and requires reasonable clock frequency. In HPLP applications, e.g. net book processor, the required performance level changes with time.
The impact of a low drive buffer optimized for each of the application category on pulse-width is shown in Figure 5-12. The optimization principal is very simple. By unbalancing the first stage of a cell with respect to rise and fall edge, we can change the pulse-width characteristics for specific supply voltage regions. The effect on average output delay and slew is minimal.
A buffer optimized for HP applications form the base line. It has consistent characteristics in the operational region i.e. between 1.05V to 1.30V. Pulse-width limits lie from 3% to 4.5% (the numbers are normalized and thus dimensionless) in this region. The normalization is done with respect to a worst-case (SS, 0.90V, -40°C) standard buffer delay. Thus, the percentage variation represents the change of pulse-width with respect to a given delay and useful only for comparative purposes. However, in the lower V_{DD} region, the pulse-width is highly impacted and required limits reach almost 7% (Figure 5-12).
A low drive buffer is optimized for HPLP applications by increasing the size of the NMOS transistor in 1st stage by 10% as compared to the HP cell. The design has to work at a large range of supply voltages with varying performance levels. The requirements are less restrictive at high voltages but more at low voltages as compared to HP applications. The different PMOS to NMOS ratio in 1st stage skews the intrinsic delay of rise & fall edge to compensate the effect of global mismatch. The average delay is slightly worse than the HP buffer but pulse-width limits lie from 3% to 4% over the whole range of supply voltages (Figure 5-12).
A low drive buffer is optimized for LP applications by increasing the size of the NMOS transistor in 1st stage by 20% as compared to the HP cell. The design typically works at low V_{DD} from 0.85V to 1.05V. Further skewing of rise & fall delay compensates low drain current. Pulse-width variations lie from 2% to 4% in this region but is generally less compared to HP or HPLP (Figure 5-12) at the expense of average delay.

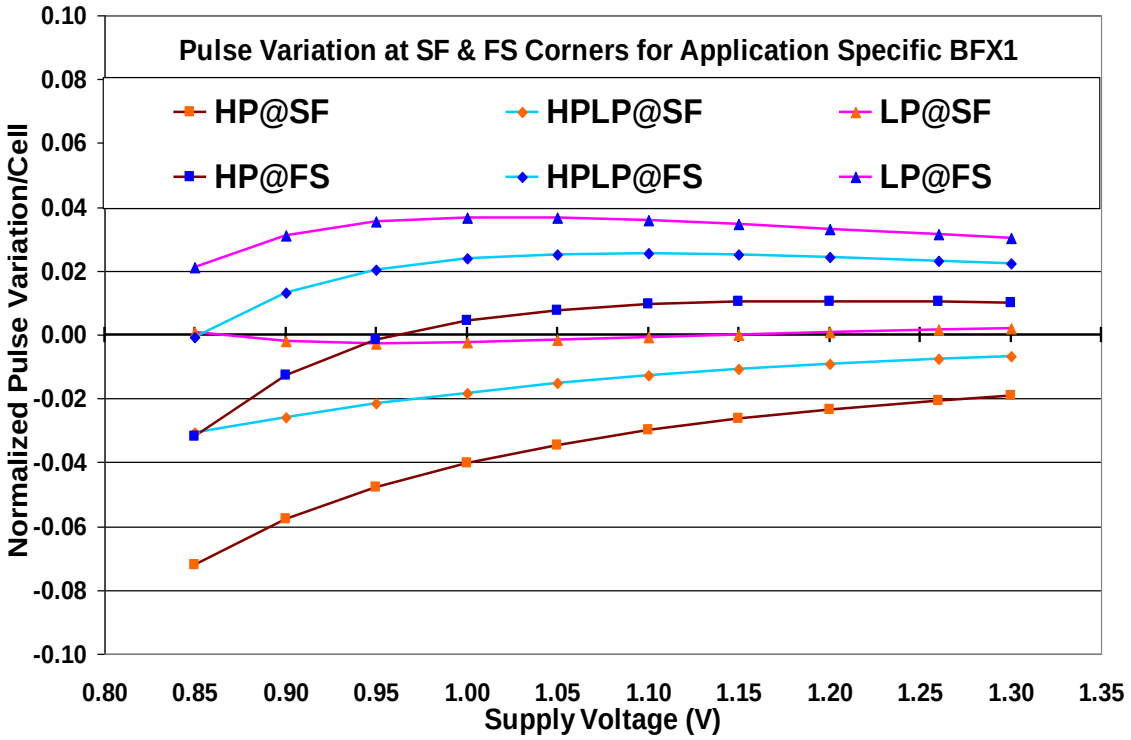

Figure 5-12: Application specific cells pulse-width variation due to die-to-die n-to-p mismatch

5.6.2 Design optimization in presence of global mismatch

Another approach to optimize for global mismatch on pulse-width is to reduce its impact in a standard design. We have made two board categories for such an approach- pulse management and design rule limitations.

5.6.2.1 Pulse-width management

Pulse-width management comprises of reducing the bottleneck in a design. It can be done using skewed duty-cycle, custom cells to change pulse-width, or by using inverters for low power region. The aim is either to reduce the pulse-width variation in the whole tree or to manage it at the end without touching the rest of the tree. Each technique has its own advantages and disadvantages.

5.6.2.1.1 Skewed pulse signal

Skewed pulse i.e. a duty-cycle of less or more than 50% can be used to mitigate the impact of SF or FS corner. The skewing is in direction opposite to the one favored by the critical corner. For example, if the SF corner reduces the pulse-width, a pulse of duty-cycle greater than 50% can be used. However, skewing can be done only until the other corner does not get critical. As skewed pulse will affect the whole design, it can be used effectively only in cases where a large number of leaf nodes

are in critical condition for a particular corner. Although it is easy to implement in design, it requires a lot of verification analysis.

5.6.2.1.2 Clock inverters

Clock inverters are excellent pulse-width managers at low V_{DD} designs. A low power design working at low V_{DD}, can use a clock tree with inverter as its driver. As shown in previous section, inverters have negligible pulse-width impact at low V_{DD} but negligible gain over buffers at high V_{DD}. Dynamic power consumption of an inverter clock tree as compared to a buffer clock tree is a complex issue. There are more cells but lesser number of inverters in the path. Inverter clock tree uses larger load cells everywhere as compared to buffer clock tree where load cells have smaller capacitance. Factor in the input capacitance difference between the two along with reduced drive capacity of inverters into dynamic power calculation. Inverter clock tree will require doing it from scratch and with more number of constraints.

5.6.2.1.3 Custom cells

Customized cells to change pulse shape can be used to manage the pulse-width in a specific corner critical path. The cells are customized by altering the PMOS-to-NMOS ratio of both stages in a buffer to favor a particular pulse change characteristic. For example, if the path is SF critical, then we put a buffer with larger 1^{st} stage NMOS and 2^{nd} stage PMOS and smaller 1^{st} stage PMOS & 2^{nd} stage NMOS as compared to a standard buffer. If the path is FS critical, then we put a buffer with smaller 1^{st} stage NMOS and 2^{nd} stage PMOS and larger 1^{st} stage PMOS & 2^{nd} stage NMOS as compared to a standard buffer. The amount of change will determine the affect on pulse-width. Thus, multiple cells can be created that support different magnitudes of pulse change. These cells are not meant to drive the tree and can be used just before leaf node to manage a MPW violation. Inserting extra cells will increase insertion delay but will not affect the rest of the tree. Thus, it is useful for post layout corrections for non-delay critical paths. Insertion delay can be minimized by splitting last levels and exchanging the standard buffers with custom cells.

5.6.2.1.4 Increase NMOS & PMOS of 1st stage

Output-to-input stage size ratio can be decreased in smaller cells by increasing input stage size to reduce the impact on pulse-width. It will result into increased input capacitance but will have a minimal impact on area. These cells are useful only for low power designs as they will increase the overall delay and are not suitable for high performance design. The PMOS-to-NMOS ratio of each stage will remain the same.

5.6.2.1.5 Clock stack cells with balanced input stage

Cells like AND/OR do not drive clock tree and thus try to reduce the intrinsic delay by minimizing input stage delay and using a balanced output delay. Such a configuration tends to skew the pulse-width in presence of global mismatch. Using a balanced input stage will improve the pulse characteristics but will also increase the intrinsic delay and input capacitance. Other option is to change the output stage p-to-n ratio to negate the pulse skew induced by input stage.

5.6.2.2 Design rule limitations

CAD tools use design rules for synthesis using various input parameters. They help to find an optimum solution in the design space by constraining the choices and guiding the outcome. Using design rule limitations, we can avoid situations where global mismatch can be limiting factor.

5.6.2.2.1 Avoid low drive buffer for low V_{DD} designs

Minimum drive may be constrained for designs working mostly at low V_{DD}. Higher drive strength can increase dynamic power consumption but will limit global mismatch impact. Power consumption can be limited by constraining the maximum drive strength forcing the tools to split larger delay lines into smaller ones. However, it can be compensated by increasing the load on minimum drive buffers.

5.6.2.2.2 Upper limit on slew at low V_{DD}

Low V_{DD} operation has a very large impact of global mismatch in case the slew exceeds 100ps. Slew constrains are put for nominal V_{DD} operation and slew changes are left onto voltage scaling. It may be necessary to put constraints on the maximum slew allowed for low V_{DD} also. It can affect the maximum slew at nominal or high V_{DD} or may require altering clock tree to meet the low V_{DD} slew constraints.

5.6.2.2.3 Smaller delay input in stacked logic gates

In stacked logic gates the input closer to output , for example input ‘CP’ in clock gate in Figure 5-4, provides a more balanced rise and fall delay. Two or more input cells with one clock input should always use the smaller delay input.

5.7 Approach: Silicon vs. Simulations

Global mismatch is a type of random variation that makes it difficult to have a one-to-one matching between spice and silicon. Various phenomena can induce difference between silicon and simulation results or constrain the testability of paths. Parasitic difference between model and silicon, systematic effects, lack of knowledge about exact point in global variation space, silicon to model error,

process centering, test equipment error, non-testable paths, limited testable pads, maximum test frequency, etc do not allow a one-to-one matching between silicon and simulation. Corner spice models define the 3σ variation limits. Ideally, silicon results should lie within these boundaries. If results from a set of different dies create a shape encapsulated by the boundaries created from spice corner simulations, it proves the validity of the variation model. In case there is a big difference between spice and silicon results, it can point to a fault in variation model.

The parameter most easily accessible in test results is delay. Thus, any test strategy should be based on the same. Generally, test structures are constrained to ring oscillators (RO) with one point accessible from test pad, as shown in Figure 5-13. Global mismatch affects pulse-width but ROs are based on delay. Thus, we need to extract pulse-width from RO delays. There are two measurable parameters in RO: $t_{1\text{-}to\text{-}1}$ delay (or rise-to-rise delay at one point in RO or inverse of oscillation frequency) and $t_{1\text{-}to\text{-}0}$ delay (or rise-to-fall delay at one point in RO or single oscillation delay). $t_{1\text{-}to\text{-}1}$ is simply sum of $t_{1\text{-}to\text{-}0}$ and $t_{0\text{-}to\text{-}1}$ in a RO. Therefore, by having any two of these delays, we can calculate the third delay ($t_{1\text{-}to\text{-}1} = t_{1\text{-}to\text{-}0} + t_{0\text{-}to\text{-}1}$). In an ideal RO, $t_{1\text{-}to\text{-}0}$ and $t_{0\text{-}to\text{-}1}$ should be equal but global mismatch will create difference between the two. This difference between $t_{1\text{-}to\text{-}1}$ & $t_{1\text{-}to\text{-}1}$ is equal to pulse-width change in an equivalent clock path. Thus, $\Delta PW = t_{1\text{-}to\text{-}1} - 2*t_{1\text{-}to\text{-}0}$.

5.7.1 Silicon test

To measure global mismatch effect, we need a long chain (about 1000 cells) RO made up of a single type of cell (preferably buffer) except the inverting cell. It should be kept in mind that although the number of inverting sets in RO is always odd, the number of cells, that may include buffers, may be even or odd along with an inverting cell. In practice, the inverting cell in a RO has 2 inputs, like NAND gate, to allow the initialization of RO. The test cell should be of large drive to minimize local mismatch effect. Long chain will further average out local mismatch effect. Multiple instances of the test RO need to be placed on each die. Their average delay value should negate systematic effects. Samples from statistically significant number of dies will be needed (few hundreds) to obtain a wide range of process corners. Samples may be needed from same wafer, different wafer in same lot and from different lots to cover the whole range of variations. From each sample, we need to extract the 1-to-1 delay and 1-to-0 delay. Calculate Δ Pulse Width from two values.

5.7.2 Simulation

Spice simulation netlist needs to match the silicon RO. We use the netlist of the RO used in silicon with transistor level extracted parasitics, necessary to minimize difference between two. Test point in silicon and simulation needs to be same as 1-to-0 delay varies with point to point in a RO. Using same V_{DD} and T conditions as in silicon test, we can measure 1-to-1 delay and 1-to-0 delay from the test point for 5 major corners (SS, FF, TT, SF, FS). To achieve better results, custom corners can

be created with 1σ and 2σ variations. Calculate Δ Pulse Width from two delays at each simulation point.
We will do a Monte Carlo simulation with global variations only on the test netlist and extract the required delays.

5.7.3 Matching silicon to simulation

Plot the points obtained from simulation on a ΔPulse-Width vs. 1-to-1 delay graph and join the points to create the encapsulation. Now plot all the points obtained from silicon on this graph as in Figure 5-14. If most of the points lie within the encapsulation and follow a Gaussian kind of distribution with very few points lying near the boundaries, it proves the validity of the model. Small deviations in X-Y or angular direction are possible due to process shift and systematic effects. The figure represents the variation effects but is not an accurate picture of the variations. Real shape may vary from the one shown in figure.
Further, a correlation analysis between silicon results and spice Monte Carlo results for 1-to-1 delay and 1-to-0 delay can provide us an idea if the spice Monte Carlo simulations are in accordance to silicon test measures.

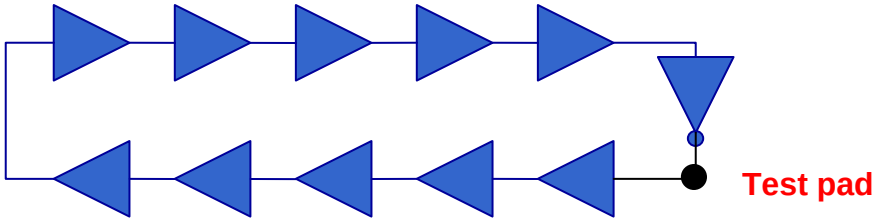

Figure 5-13: Ring oscillator test circuit

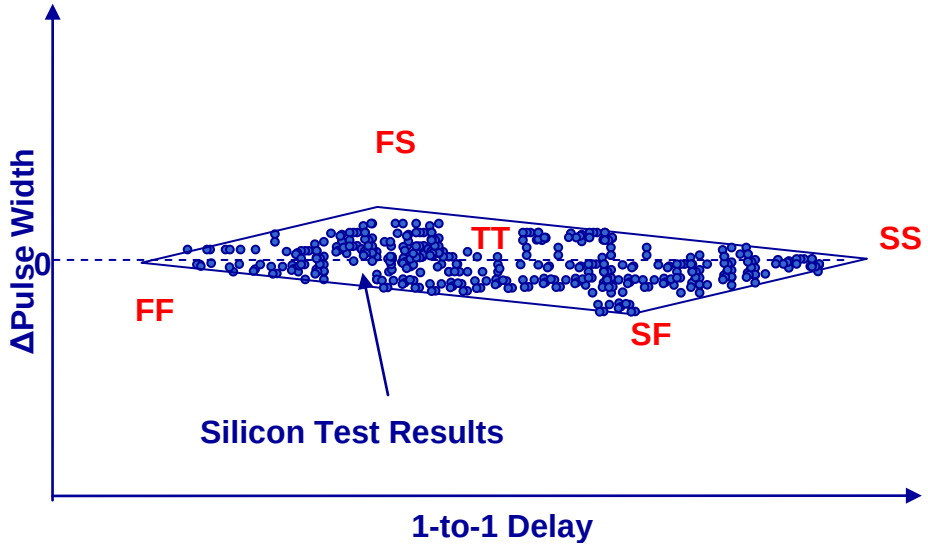

Figure 5-14: Silicon vs. Spice verification method

6 Impact of and Design Solutions for Within-Die Local Random Mismatch

Local random variations (or local mismatch or intra-die random variations or within-die random variations) is an emergent challenge in microelectronics industry [44], especially in the ASIC designs. Being a random variation, it was supposed to average out along a path until recently. However, its increasing magnitude [106], complex impact on transistor characteristics [12] and increasing clock frequency have made it a non-negligible factor when considering design variations. 3σ local random delay variations had already reached 5.5% in 90nm [94]. Decreasing systematic local variations due to regular design has left random local variations as a major contributor to intra-die variations [70], [112].

Local mismatch creates a difference in electrical properties of neighboring transistors, due to which two similar paths on the same die can exhibit different delay and power metrics. The resulting skew between two clock paths creates limitations on amount of data logic between consecutive flops. As the impact of local mismatch is different for NMOS and PMOS, it affects pulse-width creating limitations on design frequency and die size.

Local mismatch is a fully random within-die phenomenon and thus cannot be modeled as systematic variations or included in corners like global variations. Standard approach to handle mismatch is using On Chip Variation (OCV) margins in corner conditions. Local mismatch margins behave differently than clock jitter or global mismatch that can be applied as percentage values of insertion delay. It is also more of an absolute value than a percentage value. OCV margins reduce achievable performance. Any over-budgeting will increase the design effort as well as force to make compromises in other parameters.

Although, SSTA can handle statistical variations, there is marginal benefit of doing it over corner analysis for current designs [84]. Optimization using SSTA favors microprocessor designs that benefit from binning. Improvements in process spread or mean frequency can be done during ramp-up reducing turnaround time. However, timing optimization has little advantage in ASIC designs that have hard performance cutoff limits. ASIC design optimizations are focused towards improving additional parameters like power after achieving timing objectives.

To achieve timing closure in ASIC designs in presence of local random variations without vastly increasing the time to market requires a good understanding of how local mismatch affects various configurations in a design. In this chapter, we looked at how local mismatch affects design performance, power consumption, area, etc in a clock tree. Clock tree makes a logical choice for studying local mismatch as it can provide maximum benefits through improvement in clock frequency and data path logic. We looked at the origins of local mismatch at physical level and how it affects transistor characteristics in turn affecting cell and path level parameters. We looked at the impact of local mismatch under different configurations to find robust choices that can reduce timing failure probability. We also proposed an approach to predict local mismatch in a path that is consistent with current STA methods. Using analytical equations, we predicted local mismatch for cells under different conditions reducing the characterization effort. In the end, we proposed optimization methods considering the whole design and looking at PPAY (Performance-Power-Area-Yield) metrics [116], [117].

6.1 Origin

Any production process has inherent statistical variability to some extent due to non-ideal nature of the process and its components. The variability can be related to equipment or material imperfections caused by natural or technological limitations. There are four principal causes of local mismatch identified as of yet- Random Dopant Fluctuations (RDF), Line Edge Roughness (LER), Polysilicon Granularity, Oxide Thickness Variations (OTV). More phenomena may be present that affects transistor characteristics within-die but are tough to identify due to their nanoscale nature.

RDF is defined by inherent fluctuations in dopant locations and statistical variation in number of dopant atoms inside a transistor. The difference in voltage potential profile between transistors affects the drain current creating variations as well as an average shift. LER is defined by the atomistic roughness in gate edges affecting the effective gate length and width. The net effect is on threshold voltage and oxide capacitance. Polysilicon granularity is caused by the granular nature of polysilicon that affects doping profile and can create potential barriers in channel region affecting drain current. OTV is caused by atomistic roughness of gate surface that changes the effective oxide thickness over the surface of the gate. It can affect tunneling current as well as surface potential and mobility. Until 35nm node, RDF is the dominant cause of local mismatch with LER constituting little less than half of RDF variations. OTV has a negligible impact above 22nm and Polysilicon granularity will be equivalent to that of RDF at 35nm.

These variations create difference between any two transistors on a die and affect their intrinsic delay. Their impact is limited to few parameters that can be measured but cannot be differentiated based on its source. The physical variations follow Gaussian distribution [27] though the impact on delay may be closer to lognormal distribution [115] due to exponential relationship between delay and threshold voltage.

6.2 Effect on design

Local random variations affect electrical characteristics of a transistor. This affect is propagated up the hierarchy affecting cell delay, path delay and design frequency. There is some averaging effect as we move up the hierarchy. Moreover, global variations still dominate the variation space. If the impact of local mismatch on delay does not cross the limiting cases for global variations, then it can be neglected. However, in case of pulse-width or skew, the delay component of global variations is negated leaving local variations only.

The impact of local mismatch variations is a function of global process point and attains maximum value at SS corner. The larger standard deviation arises due to smaller number of dopant atoms at SS corner resulting into reduced statistical averaging. Least drive current and V_{DD}/V_{th} ratio also plays a significant role in increasing local variations at SS corner.

6.2.1 Effect at cell level

Local mismatch affects transistor I_{DS}-V_{GS} characteristics differently than global variations. One of the principal components of local mismatch, RDF, can cause variations of threshold voltage as well as an average shift of the resulting distribution as shown in Figure 6-1 [46]. It also affects the drain current value in linear region for a given gate voltage. The net impact is on transistor drive current and threshold voltage, thus affecting transistor switching delay and transition time. LER has a similar impact on transistor characteristics with a net shift and variations in gate perimeter affecting threshold voltage and gate capacitance. Poly silicon granularity increases the threshold voltage and causes variations depending on grain size. A 35nm mosfet can have a 1-σ threshold voltage fluctuation of 49mV for all within-die random sources combined [11].

Variations in transistor characteristics affect cell electrical parameters like delay, leakage, dynamic power, transition time, etc. Presence of multiple transistors with uncorrelated threshold voltage variations reduces cell delay variations as compared to a single transistor delay variation. 3σ delay variation of a cell as compared to its nominal delay can reach 15% for a 70nm device [42] and 35% for a 35nm device [16]. However, delay variations of one gate are correlated to some extent to delay variations of previous gate through transition at the intermediate stage. Non-linear relationship between cell delay and threshold voltage affects the delay distribution, causing a mean shift towards slower values, i.e. there are more number of cells with slower than average delay. Thus, random intra-die delay variations of a cell are expressed using nominal M, mean shift μ, and standard deviation σ. The statistical limiting cases will be: M+μ+3σ and M+μ-3σ.

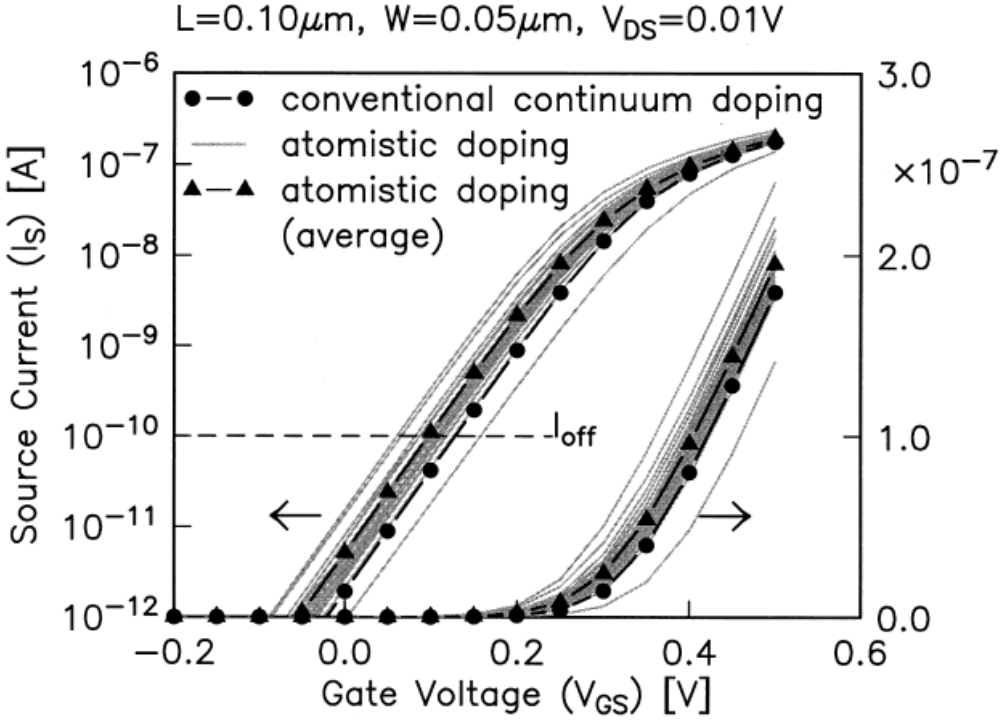

Figure 6-1: RDF impact on mosfet transistors with different atomistic doping distributions [46]

6.2.2 Effect at path level

Random delay component for each cell in a path combines to affect the final path delay. As the random component cannot be determined individually, delay distribution for each cell is combined to get the path delay distribution as shown in Figure 6-2 [117]. The delay distribution is like Gaussian but the probability of achieving limiting cases is much lower than global variations as it will require every single transistor in path to be at the limiting case. Delay variations caused by local random mismatch combine with that caused by global variations to achieve the final path delay distribution. Impact of local mismatch on a path that does not lie near the corners can be easily neglected for delay, as the combined worst-case will still be less than the corner case. The impact of local mismatch on corner delay will have to be taken into account, as the combined worst-case delay is higher than corner delay, (Figure 6-3). However, the probability of 3σ local mismatch on corner is extremely low and it may be more useful to use only 2σ local mismatch at corners.

The same is not true for skew and pulse-width that are calculated as a difference of two delays- different path delays for skew and different edge delays for pulse-width as shown in Figure 1-3 and Figure 1-2 respectively. For similar paths, the impact of global variations will be equivalent for both delays and will negate largely leaving only the impact of local mismatch. Being random in nature, the local mismatch component of two paths is mostly uncorrelated. The difference between two will vary as a RMS addition of two delay variations. Although, worst-case skew or pulse-width can occur at any point in global variations, the magnitude is largest at slow corner only. Thus, doing a mismatch on corner analysis can provide sufficient information about timing characteristics of any path. Skew variations will affect the required setup and hold time (Figure 1-4) for timing closure as shown in equations (1-1) and (1-2). The percentage value of local mismatch with respect to insertion delay may reduce with path depth, but the absolute value increases that affect skew and pulse-width. Larger percentage value may be acceptable for mismatch over a small path as compared to a smaller value on a large path.

For a given design, larger skew or pulse-width variations means higher margins on clock pulse increasing clock period or limit data size logic. With each technology node, mismatch variations are increasing and reduce the gains of scaling. Skew and pulse-width variations will require extra OCV margins to avoid chip failure. These extra margins eat up into the size of data logic and design frequency. Although a chip may be functionally working, it can have large static or dynamic power consumption affecting reliability. Within-die variations can reduce the chip mean frequency by almost 15% in 50nm technology [65].

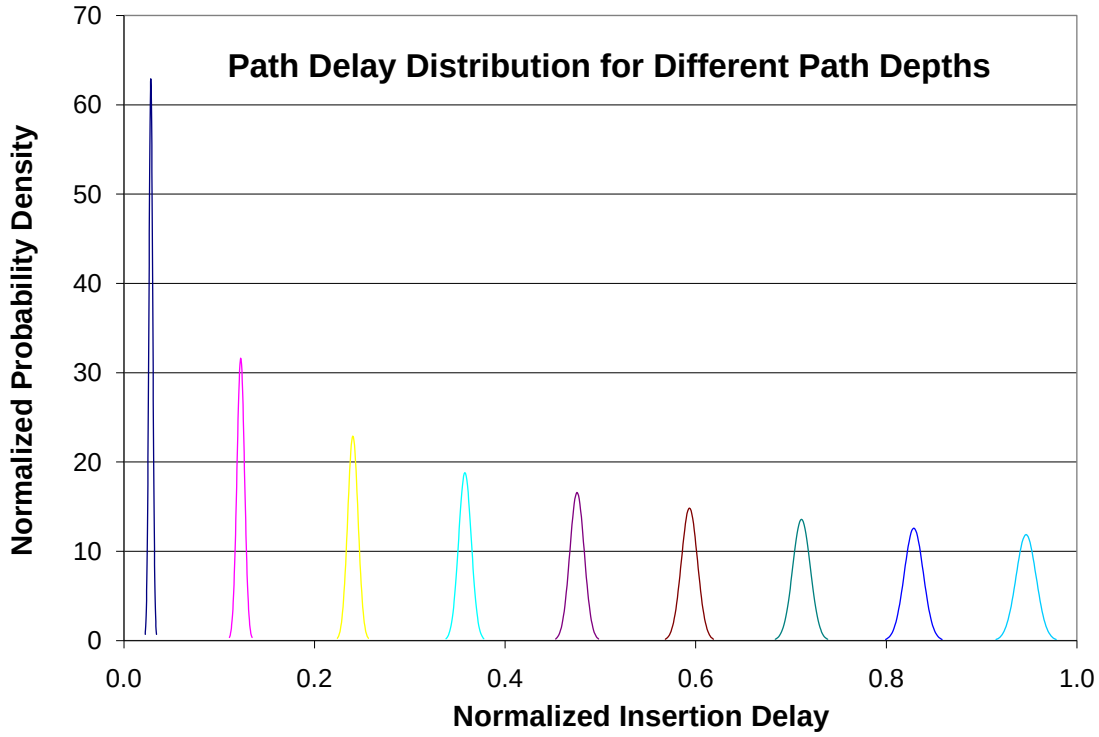

Figure 6-2: Distribution of path delay at different path depths due to local random variations

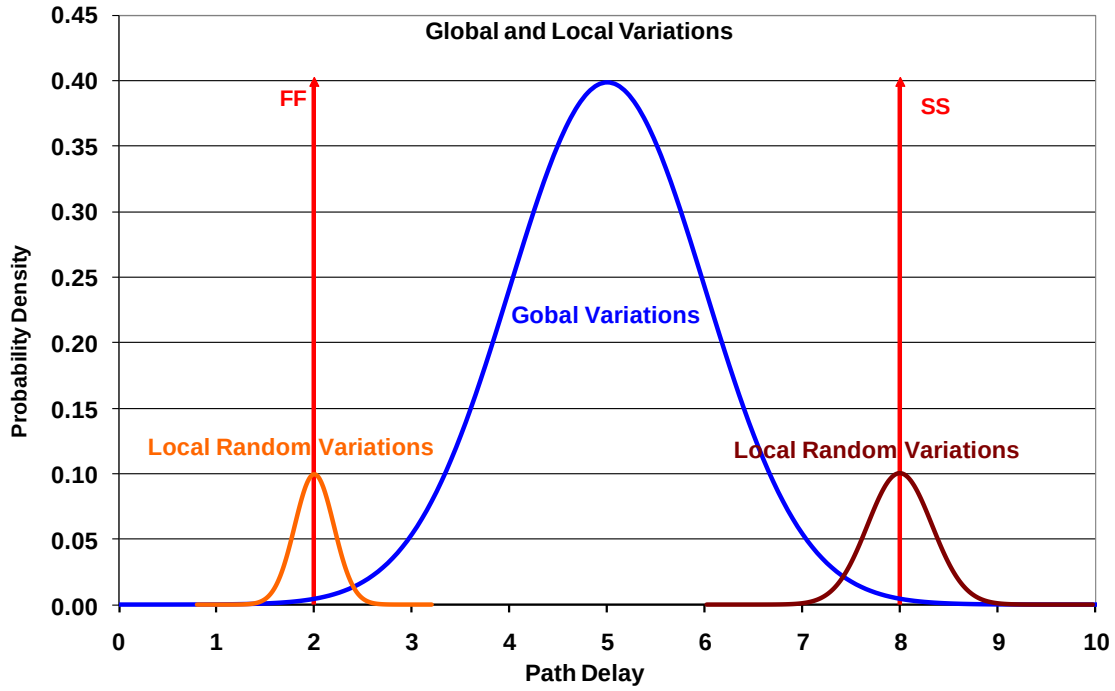

Figure 6-3: Impact of global and local variations on path delay

6.3 Cell level analysis

Mismatch simulation for different buffers and inverters revealed a decreasing σ with drive strength as shown in Figure 6-4. σ-values (as percentage of normalized SS delay) are plotted against drive strength for buffers and inverters at SS and FF corners. Figure 6-4 shows the σ-values for rise edge output while Figure 6-5 shows the σ-values for fall edge output. The difference in two edges arises due to different active transistors inside the cell. As different drives of same cell type are loaded for equal slew, their respective delays are equivalent. As seen in Figure 6-4, mismatch for low drive buffer at SS corner is quite high because of small cells both in input and output stage. As the drive strength increases, mismatch reduces but the rate of decline also reduces and has minimal change for cells above 4.5X. There is marginal difference between different drives for FF corner. That means the impact of mismatch for very large drive cells is unaffected by process changes as seen by equivalent values for SS & FF process. Similar trend can be seen for inverter but with lower magnitude as now the edge passes through a single transistor as compared to two in buffer. Input stage of a buffer is also smaller whereas the output stage should be equivalent to that in inverter. Thus, it can be approximated that inverter shows the maximum mismatch contribution of output stage of a buffer of same drive strength. It should be kept in mind that mismatch σ has a RMS addition.

Figure 6-5 shows mismatch for same cells and conditions but for fall edge with a lower value as compared to rise edge. The impact is noticeable for lower drive cells and minimal for large drive. The difference between low drive and high drive is also lesser means rise edge is more susceptible to mismatch than fall edge. A rise edge in buffer passes through NMOS in 1^{st} stage and PMOS in 2^{nd} stage either, or both of which can be responsible for the higher amount of mismatch. Difference in rise and fall edge mismatch will affect pulse-width and opposite edge skew. The impact is higher for buffers and although it is not negligible for inverters, the alternating edge behavior will negate the impact largely.

As drive strength increases, the size of input stage also increases reducing the amount of mismatch. It may be possible to increase the input stage size in multi-stage cells like buffers to reduce mismatch sensitivity. For larger buffers, there is little advantage of increasing size. For smaller cells, the reduction in σ is larger than reduction in delay when moving from SS to FF corner. The rise and fall delays for a cell are approximately same but the σ-values differ. The σ-values for a two-stage cell are slightly smaller than the RMS addition of σ-values of cells equivalent to its stages arising from smaller slew in between the stages.

Mismatch being a function of threshold voltage (V_{th}) and supply voltage (V_{DD}), low V_{th} (LVT) transistors have a reduced mismatch impact due to higher V_{DD}/V_{th} ratio than standard V_{th} (SVT) or high V_{th} (HVT) transistors [116], [117]. The proportionate change in σ from SVT to HVT is much larger as compared to that from LVT to SVT. Thus, it is more advantageous to move from HVT cells to SVT cells. Although the increase in HVT delay can reduce percentage mismatch, absolute delay mismatch is much higher. Large L cells increase the intrinsic delay but also reduce local mismatch almost to the level of LVT cells. Thus, it may be advantageous to use large-L cells instead of HVT cells for leakage reduction.

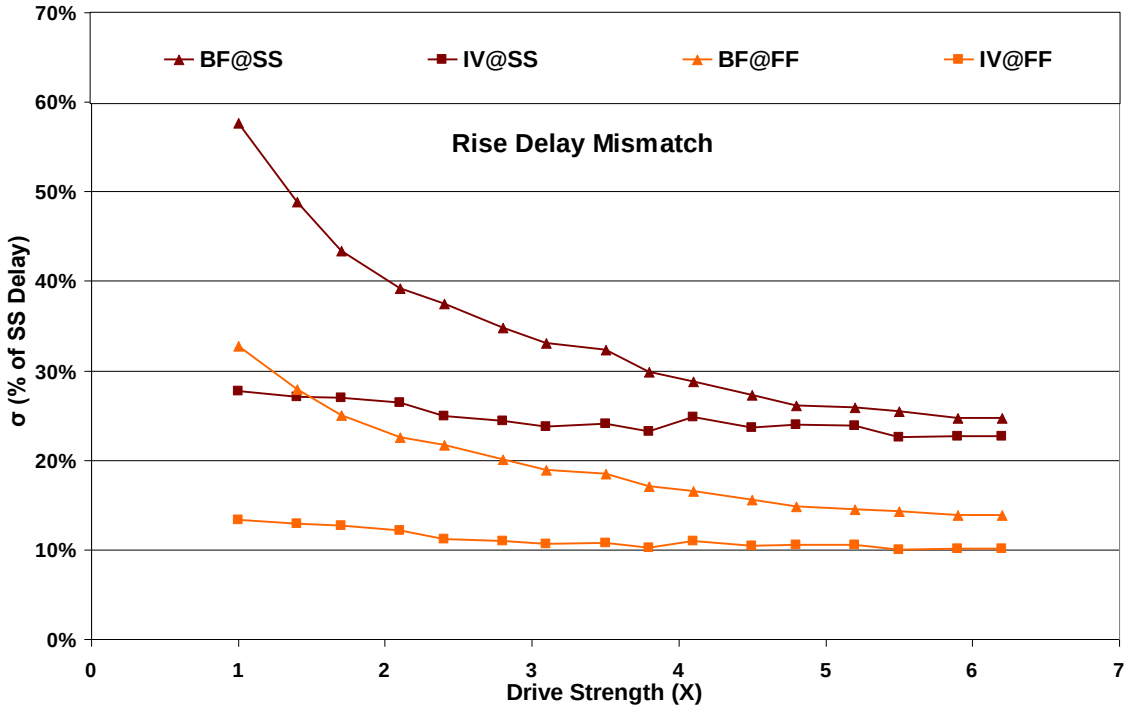

Figure 6-4: Rise delay mismatch σ for buffer and inverter for different drive strengths at SS and FF corner normalized against buffer SS rise delay

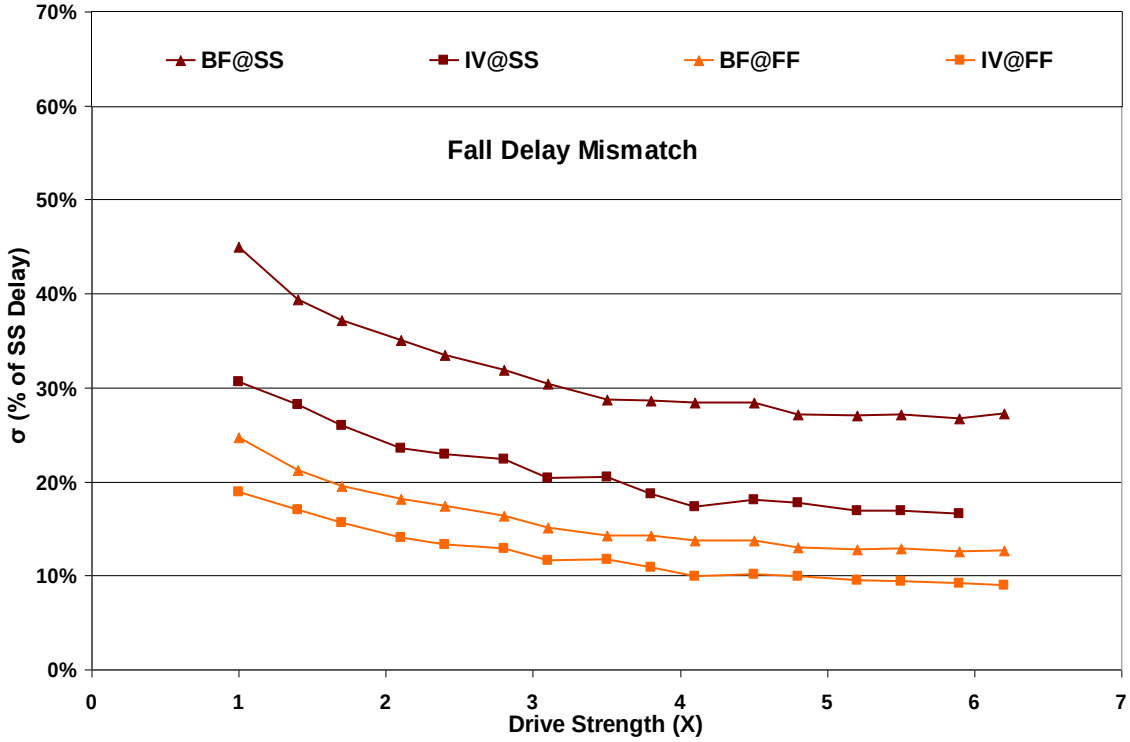

Figure 6-5: Fall delay mismatch for buffer and inverter for different drive strengths at SS and FF corner normalized against buffer SS rise delay

6.4 Path level analysis

Cell level analysis demonstrated amount of mismatch in individual cells thus illustrating robust and weak configurations. However, it is only a part of the story. What matters in the end run is amount of mismatch in a path. Although, path mismatch is made up of cell mismatch, there are other factors including interconnect delay, clock frequency, path usage, design application, etc that determines the amount of reasonable delay mismatch in a path. The following section shows local random mismatch in different clock paths under various configurations.

The impact of local random mismatch on path delay can be seen in Figure 4-1 where x-axis shows the normalized insertion delay and y-axis shows the corresponding percentage delay mismatch for two cases- local variations and global+local variations. As seen in figure, percentage value of mismatch decreases exponentially with path depth but does not average out completely. Being an uncorrelated random variation, σ/μ was supposed to become negligible for long paths (60-path depth for us). However, increasing mismatch magnitude and decreasing cell delay & clock frequency makes it a non-negligible factor in design considerations. The absolute value of mismatch increases continuously along a path, as shown in Figure 6-2, adding with a root mean square (rms) function with each stage. Linear incremental delay and rms incremental absolute mismatch combine together to give $1/\sqrt{n}$ decay for percentage mismatch. A path with depth 60 will still have a residual delay mismatch of 4.5% at the leaf node in 35nm [16].

Figure 4-1 also demonstrates the impact of a non-zero mean in cell delay mismatch caused by a non-linear relationship between delay and mismatch variations on path delay distribution. The mean value of path delay distribution is shifted, i.e. there are more cells with delay higher than corner delay than cells with delay smaller than corner delay. It also illustrates that the difference between statistical worst-case mismatch delay and corner delay is higher than that between statistical best-case mismatch delay and corner delay. The net effect is unequal positive and negative margins. Using only the standard deviation (σ) for variation margins can result into timing failure. Standard deviation in a path increases as a $\sqrt{n}$ function but mean shift increases as a linear function of depth. Thus, even a small cell mean-shift can result into a non-negligible mean-shift in a path. The effect is more pronounced for small drive buffers (BFX1) where the magnitude of mismatch is higher whereas it is negligible for large drive buffers (BFX6) with a much smaller deviation.

Figure 6-6 plots percentage delay mismatch against normalized insertion delay for HVT, SVT and LVT high drive buffer (BF5) paths in 65nm. We can also see the impact is much higher on HVT path as compared to others even considering the increased insertion delay. Thus, presence of HVT cells may have a very large impact on absolute mismatch. HVT buffers are typically used for leakage power reduction and are not present in critical paths. However, as we will show later, they can make a path skew or pulse-width critical. Replacing SVT buffers by HVT buffers will require mismatch timing analysis for all affected paths.

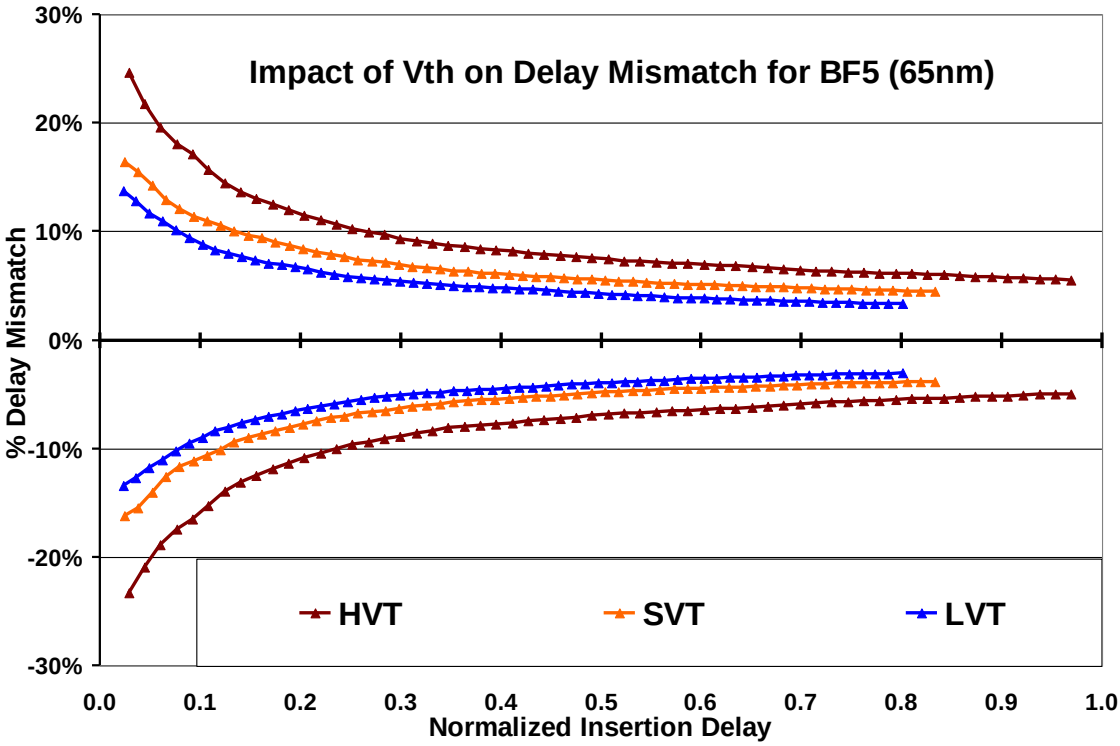

Figure 6-6: Delay local random variations for HVT, SVT & LVT buffer BF5 in 65nm

As we saw through Figure 1-3 and Figure 1-2, mismatch is a bigger concern for skew (thus setup and hold time - equations (1-1) and (1-2)) and pulse-width. Figure 6-7 shows impact of local random variations on delay and skew of same path at SS corner in 45nm [117]. The x-axis plots normalized insertion delay and y-axis plots corresponding delay and skew mismatch as a percentage of normalized insertion delay. As seen in figure, skew follows a similar trend as delay, reducing in $1/\sqrt{n}$ fashion. However, it differs markedly in fact that there is negligible mean shift, i.e. the upper and lower margins are almost equal. That is because the delay mean shift for two equal paths is same and cancels out in the difference t_2-t_1. If the paths are unbalanced, i.e. different type or number of logic gates, then the mean shift wont cancel out completely but will still be much less than delay. The magnitude of skew variations is larger than delay, and as we will see later comes out to be $\sqrt{2}$ times that of delay being difference of two random uncorrelated delays with same σ.

Unlike delay, skew cancels out corner delay contribution in paths and only the mismatch part is left. As equation (1-1) shows, increased magnitude of mismatch on path will increase the minimum possible value and will require an increase in clock period or in other words a reduction in clock frequency. On the other extreme, equation (1-2) shows an increased magnitude of mismatch can cause hold violation. Although, hold time violation due to mismatch is not very likely for most paths, it can cause timing failure for very small data paths or skewed paths. Fixing hold violations do not affect clock timing but are area limited and require very large delay cells that can have large mismatch magnitude that have to be accounted for and can even increase the required number of cells.

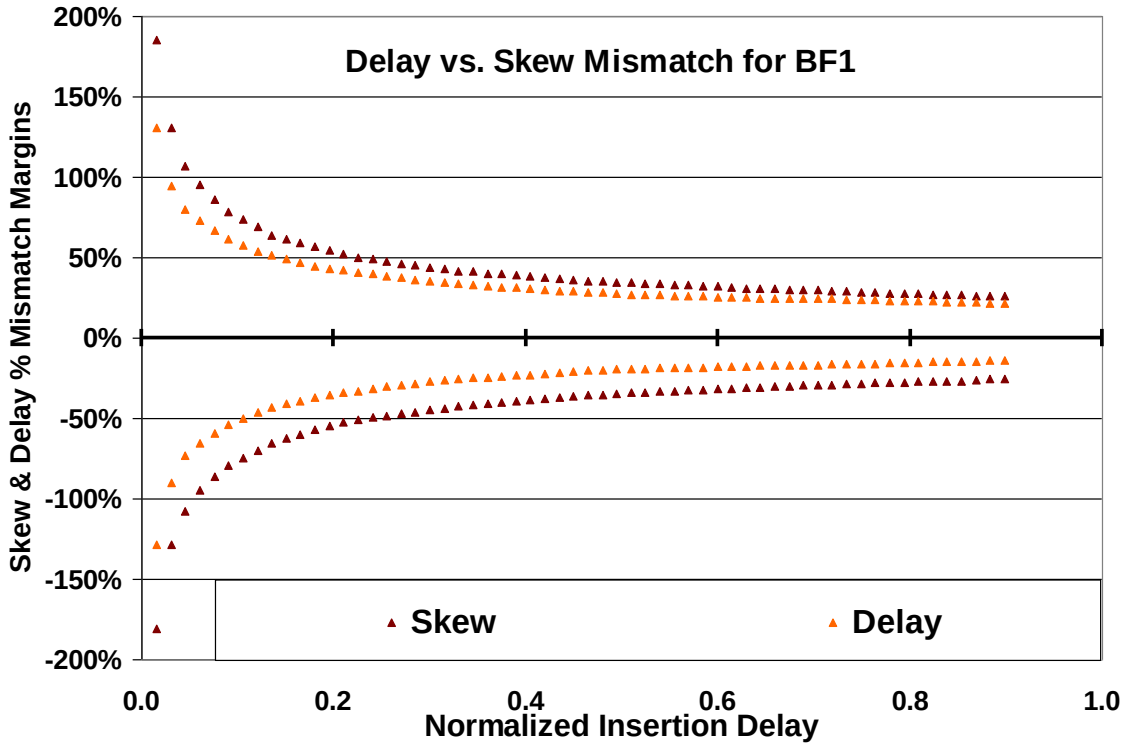

Figure 6-7: Delay vs. Skew local random variations for BF1 in 45nm

The second critical parameter affected by local mismatch is pulse-width where rise and fall delays passing through different transistors experience different amount of mismatch. The impact of local random variations on pulse-width for BFX1 is shown in Figure 6-8 that plots percentage change in pulse-width (of normalized SS delay) at different corners (SS, FF, SF & FS) against the normalized insertion delay [117]. It allows us to compare the overall impact of variations on pulse-width. The declining trend is similar to delay but the slope changes according to the corner. The most striking factor here is that limiting case conditions is mostly at the SS corner. In previous section, we saw the impact of global mismatch that makes SF and FS corners as limiting cases. However, the presence of mismatch alters the relation between different corners. The amount of mismatch is highest for SS corner and lowest for FF corner that dominates over global mismatch. There is a big shift in mean value of pulse-width distribution caused by combination of local mismatch and global corners. Local mismatch tends to increase the probability of pulse-width reduction. Whereas the SS, SF & FF corner also have similar tendency, the FS corner has opposite nature. Thus, local mismatch can aggravate the situation for SS, SF, and FF corner but improve in cases of FS corner.

Compare the impact of local mismatch on pulse-width in BFX1 in Figure 6-8 to BFX6 in Figure 6-9. Due to large size transistors, local mismatch variations are minimal in BFX6 and as such, global mismatch effect on pulse-width dominates causing a shift in its distribution. In this case, unbalanced corners (SF & FS) form the limiting cases. For a mixed cell path containing both low and high drive buffers, it can become a complex task to predict the limiting case for pulse-width and can

require extensive timing analysis. A path consisting of a single buffer type can help to reduce the number of corners that need to be verified for timing closure.

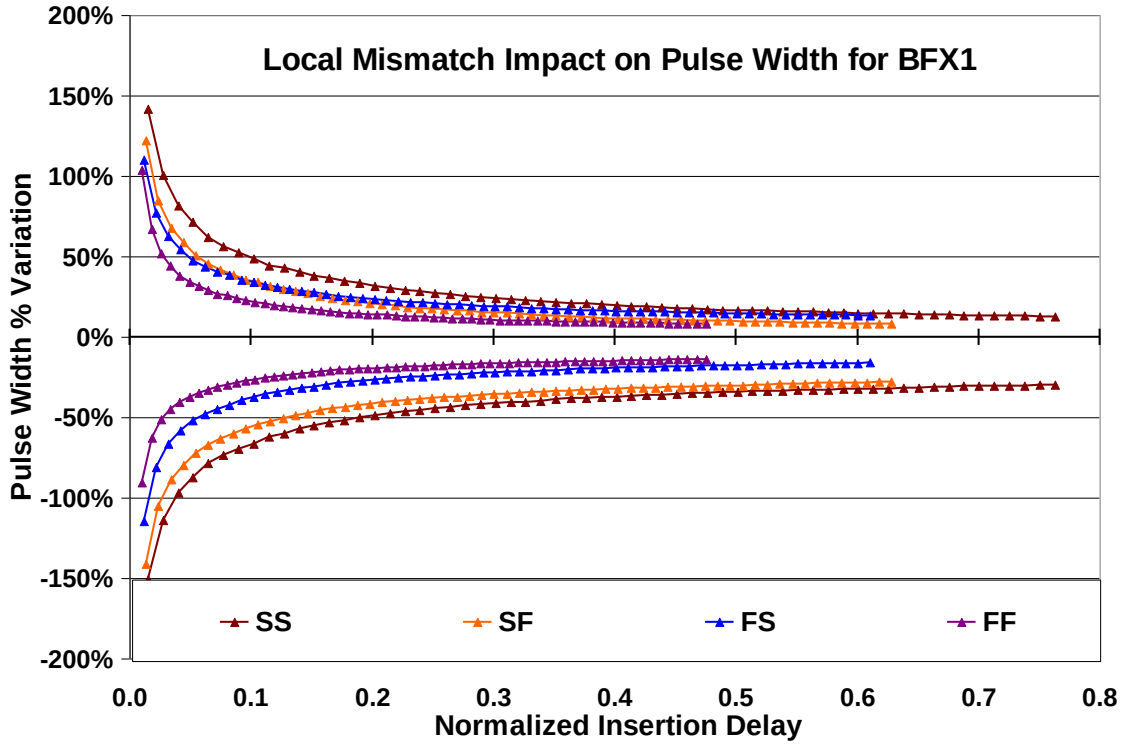

Figure 6-8: BFX1 pulse-width variations (% of SS delay) due to local mismatch on global corners

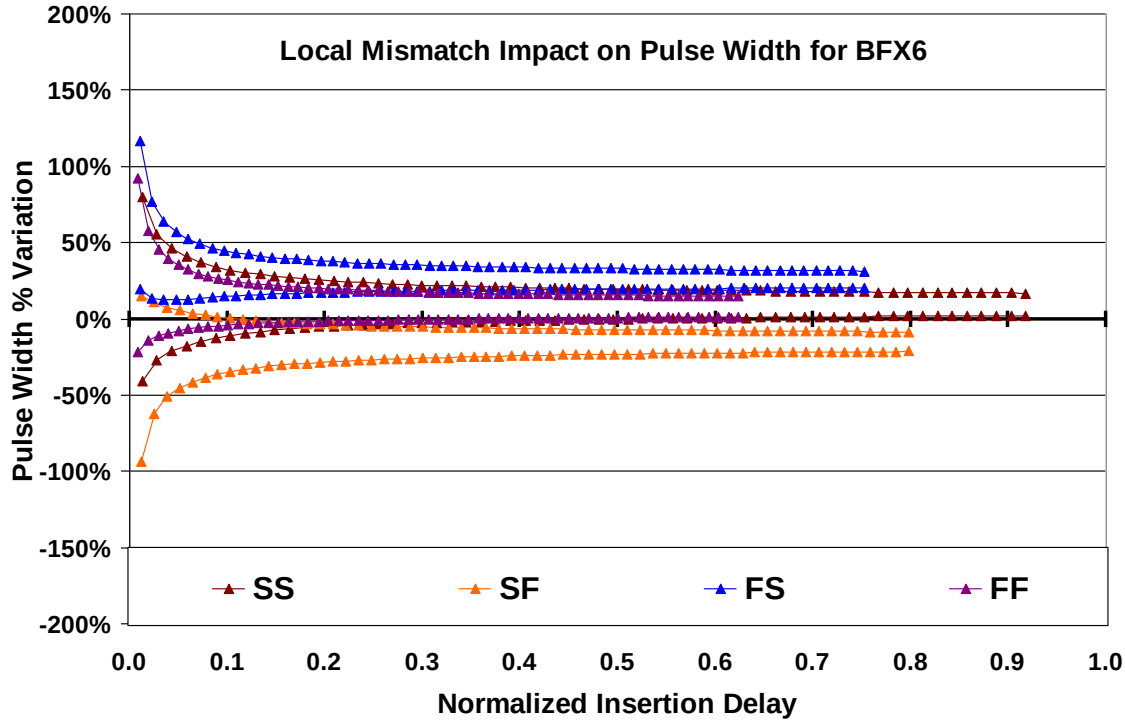

Figure 6-9: BFX6 pulse-width variations (% of SS delay) due to local mismatch on global corners

Other parameters that affect the magnitude of local mismatch in a path include slew rate [72], supply voltage, temperature, fanout, interconnect load, etc. Slew rate directly affects delay as well as absolute mismatch due to larger time spent in subthreshold region where transistor current is exponentially dependent on gate voltage. However, the impact of larger slew on % delay mismatch variations is minimal due to proportionate increase in delay and absolute mismatch for a given path depth. Fanout and interconnect load also affect the slew rate at cell input and output nodes. A larger fanout will increase the slew at output node of a driver thus increasing mismatch. A larger interconnect load will cause slew degradation and including the maximum allowed slew at input node of load cells will lower the slew at output node of a driver. Thus it may be better to have larger interconnect load than higher fanout on clock trees. Allowed signal degradation can define the resistance limit and thus interconnect load.

Figure 6-10 plots percentage delay mismatch against normalized insertion delay for three different slew rates-20ps, 55ps, 100ps. As we can see for a given path depth, the percentage mismatch is same for different insertion delays. Mean shift is also a function of slew and as we can see increases with slew. The shift tends to keep the upper limit constant for different slews. Higher slew continue to show similar mismatch trends. Although the impact on delay may be neglected, the same on skew or pulse-width cannot as they are affected by absolute mismatch. It may be possible to define the maximum slew rate on a clock tree as a function of driving buffer. Large buffers can afford larger slew without increasing mismatch significantly and smaller buffers can have a much smaller maximum allowed slew reducing the mismatch. The combination of larger and smaller slew can compensate each other for any change in delay and we can have a net reduction in delay mismatch.

Supply voltage impacts mismatch directly and through V_{DD}/V_{th} ratio. The impact is considerable for supply voltage $\leq$ 1.00V in 45nm low power process as seen in Figure 6-11. It plots percentage delay mismatch (as a % of normalized SS delay at 1.05V) against normalized SS delay at respective voltages. As we can see, there is a large increment in delay as well as mismatch for 1.00V and 0.90V. The mean shift is also much more pronounced for lower voltages. Supply voltage determines the formation of inversion layer as well as drain current of previous driver thus affecting rise/fall time. Larger the time spent in subthreshold region, larger the impact of mismatch. As seen in Figure 6-11, after a certain depth the percentage mismatch becomes stable. Timing analysis based on percentage OCV margins can benefit from the minimal path depth after which the percentage value is constant. For high voltages, mismatch achieves a stable percentage value within 15 stages. However, for lower voltages, it can go as high as 30 stages to achieve a stable value. Low voltages are mostly used for low power mode when the system is not required to function at high frequencies and even high amount of mismatch might be absorbed in the system architecture. However, percentage margins required at low voltage will be much higher.

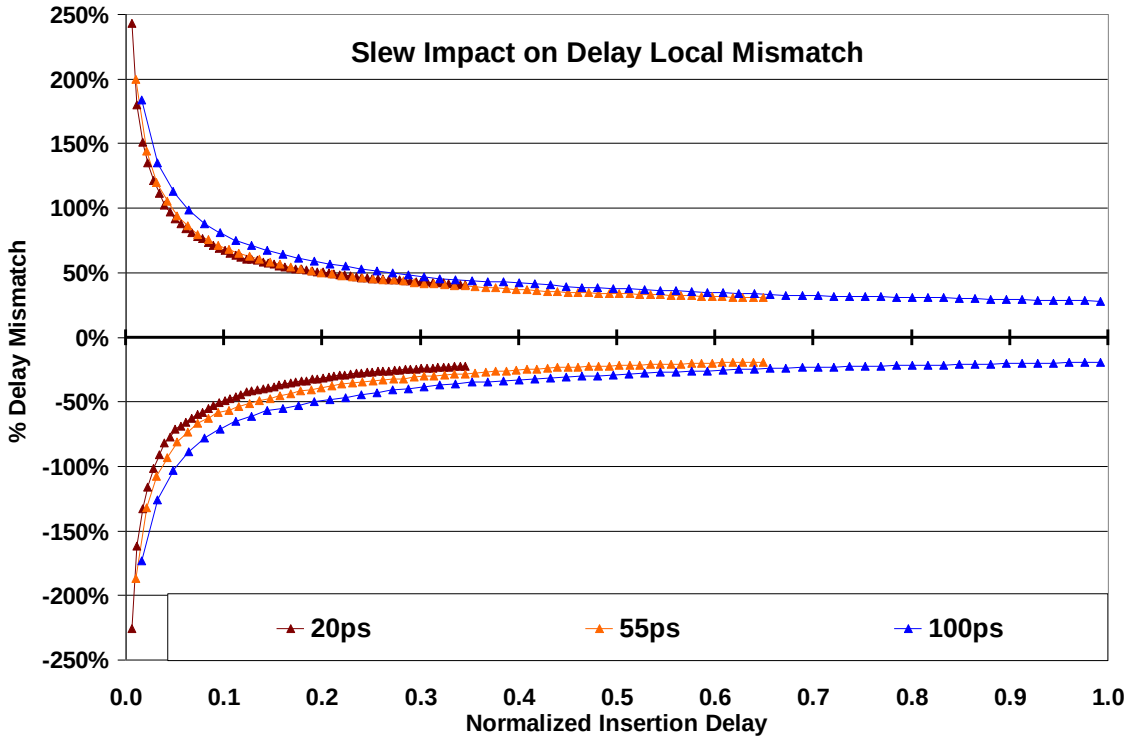

Figure 6-10: Effect of slew on delay due to local random variations in a BFX1 chain

Temperature has a negligible impact on mismatch above 1.00V as seen in Figure 6-12. The figure plots percentage delay mismatch (as percentage of normalized SS delay at 1.05V & -40°C) against normalized insertion delay at each temperature for two voltages. At 1.20V, there is very small difference between -40°C and 125°C. We also see that the minimum mismatch is for 125°C. The difference is slightly more important at 1.05V, but now the minimum mismatch is at -40°C, in line with temperature inversion in delay in 45nm. It points out the change in slew caused by change in drain current with temperature as the main cause of difference.

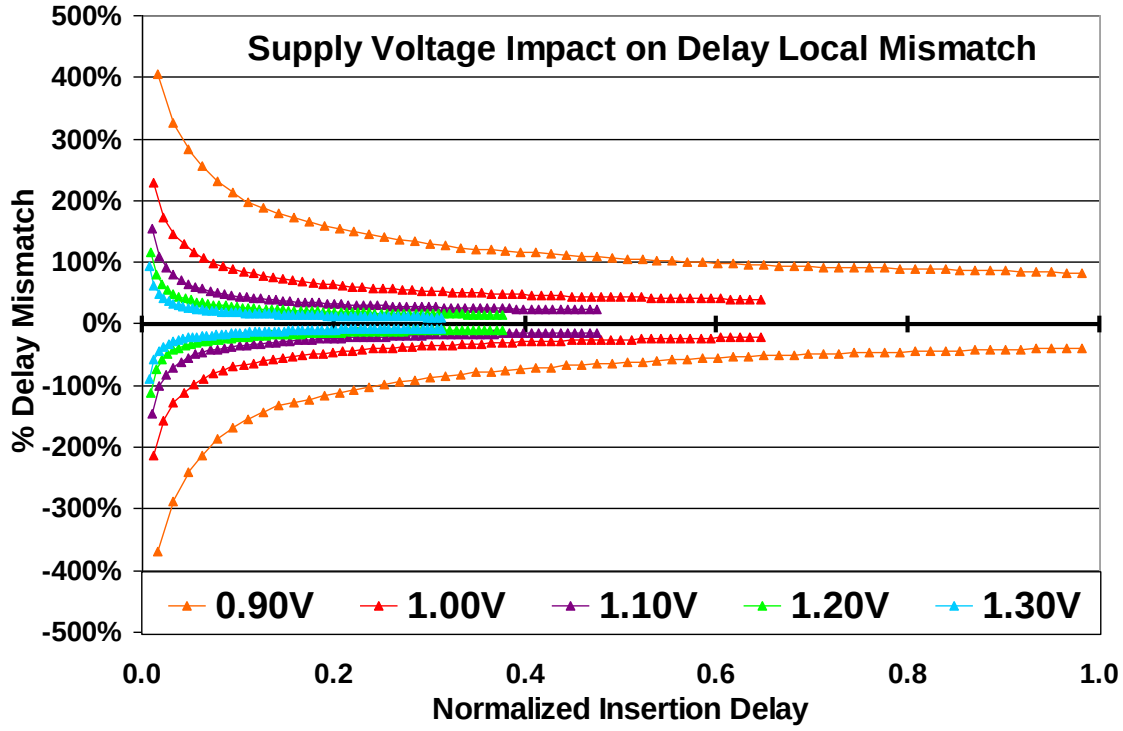

Figure 6-11: Effect of supply voltage on delay local random variations for BFX1

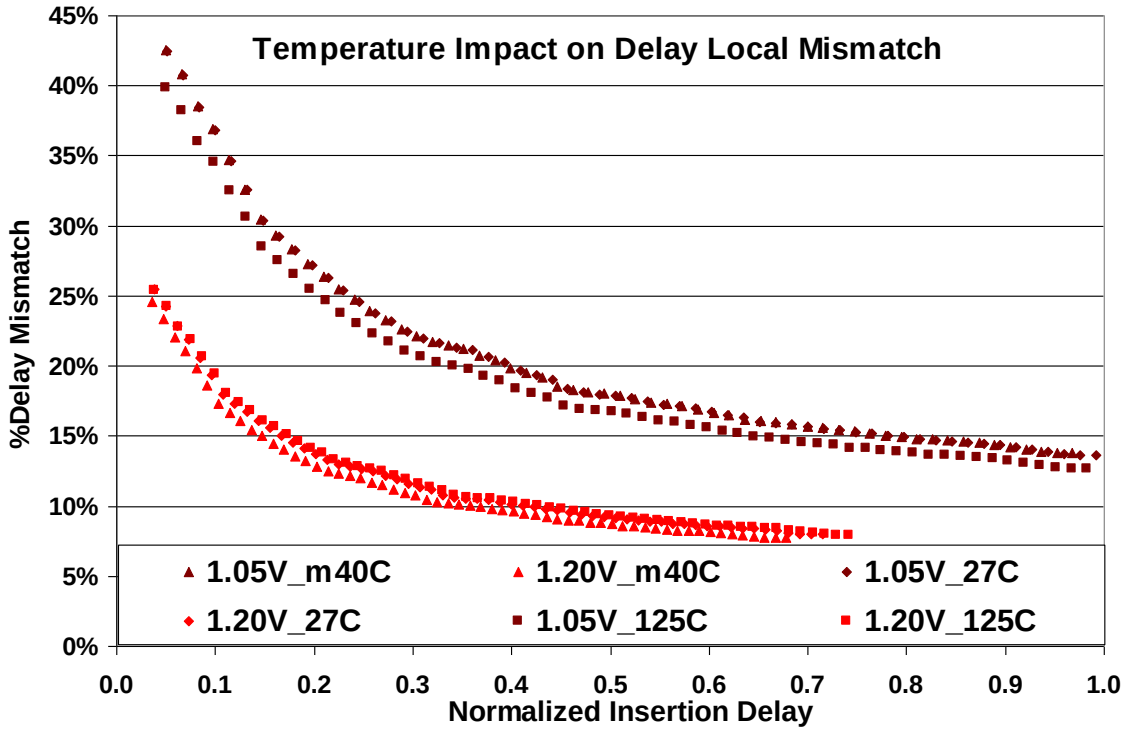

Figure 6-12: Effect of temperature on delay local random variations for BFX1

6.5 Local mismatch aware STA

Local random mismatch is a fully random within-die phenomenon that is path dependent and thus cannot be modeled or included in corners. Standard approach to handle mismatch would be through cell or OCV margins in corner conditions. However, these margins can be wasteful if impact of different conditions on local mismatch is not taken into account. ASIC designs today are working at GHz frequencies and even 10's of picoseconds are important. Using larger cell size to reduce mismatch conflicts with area and power constraints. SSTA as mentioned in earlier section is not yet ready for full-scale timing analysis.

We have demonstrated a local random mismatch aware Static Timing Analysis (STA) technique that can provide a bridge between traditional STA and SSTA. It is focused towards clock tree design as that is the most impacted quantity in digital design. Clock network design affects the worst-case clock frequency and thus determines the pass/fail conditions for ASIC chips. Moreover, clock trees are designed under restrictive conditions of cell type, cell size, and slew and thus make an easy test bed. We characterized the impact of mismatch at cell level and used it to predict the mismatch impact on paths for digital clock networks. We were able to predict the delay impact within 10% error margin (~few picoseconds in absolute terms).

Clock networks are responsible for synchronous working of a design and any unaccounted for variation in clock timing can cause a design failure. There are three important timing parameters associated with clock paths that are affected by within-die variations – insertion delay, clock skew, and pulse width. These parameters further determine setup time, hold time, clock frequency, and duty cycle at any given node.

6.5.1 Range based design vs. SSTA

The approach of handling on chip variations in STA is to use percentage margins. However, the contribution of mismatch to delay variations (in %) decreases with increasing path length. Moreover, there is a shift in mean value of mismatch impact on delay caused by non-linear relationship in subthreshold region that can increase pessimism. The amount of timing uncertainty added by mismatch for each timing node depends on path and varies from one to next. For that reason, it is difficult to apply a single rule for all paths. To measure mismatch impact and reduce design margins to minimum, Statistical Static Timing Analysis (SSTA) has been proposed. Instead of using deterministic delays as in STA, SSTA passes delay distributions through the circuit to obtain probability density function of circuit performance. Depending on the implementation, it can handle path correlations, non-Gaussian distributions, etc. However, SSTA has its own limitations. It can cause a larger design cost in terms of resources required for library characterization, runtime, and user training. SSTA libraries depend on detailed process recipe and thus it can be difficult to have a stable set of parameters during process ramp-up. Moreover, ASICs have a fixed performance point that can make it difficult to benefit from SSTA methodology.

Most fabrication changes improve product reliability and yield rather than performance and as such does not affect mismatch on delay. SSTA characterization based on process parameters is more susceptible to process changes. Statistical timing characterization typically requires foundry information and is a long computation intensive task. The proposed method is based on spice simulation of a small test circuit and thus required cells can be fully characterized within short duration without any foundry specific information. The method is robust against small process changes. It is applicable to skew and pulse width that typically are the limiting factors in presence of local random mismatch. The novelty of this approach is to compromise on error margins to enable fast implementation time and minimal overhead. The error margins are within acceptable limits.

Our proposed method builds on STA and is easily implementable in current design flow. Mismatch calculation along a path can be done with STA output and tabular models as inputs using scripts in current timing analysis tools. Computation time overhead will be minimal. Although mismatch variations can be Gaussian due to their random nature, their impact on delay is not symmetric. The impact on a slow corner will be higher than a fast corner due to a larger supply to threshold voltage ratio. The parameters used to describe a distribution are average value or nominal value (M) and standard deviation (σ). Along with that mismatch introduces a third parameter, mean shift (μ). Using these three parameters, we can describe the impact of mismatch around a corner, i.e. the statistical (99.7%) maximum ($\mu+3\sigma$) and the minimum ($\mu-3\sigma$) values possible around a given corner. Nominal value refers to the delay at the given corner without any mismatch component. Average shift represents the difference in the average value of the mismatch delay distribution and the nominal delay. Standard deviation determines the impact of mismatch variations.

6.5.2 Methodology

The goal of mismatch aware STA is to try to predict the statistical $\mu\pm3\sigma$ (average shift, standard deviation) limits around corner cases. The basic idea is to use mismatch variations of individual cells and calculate the impact on a chain made up of those cells. The standard deviation of a chain (σ_{chain}) made of "n" uncorrelated random variables is equal to the root mean square addition of the "n" individual standard deviations (σ_1, σ_2…) (6-1). We can calculate nominal delay (M_{chain}) (6-3) and average shift (μ_{chain}) (6-2) of a chain by linearly adding cell nominal delays including interconnect delay (M_1, M_2…) and average shift (μ_1, μ_2…) for each cell. Equations (6-4) and (6-5) represent the maximum and minimum impact of mismatch on a path for a given corner and slew. We characterized the cell parameters (μ and σ) for delay using the setup shown in Figure 4-8 doing a Monte Carlo simulation with a sample size of 1000. Although random variations are uncorrelated, their impact on two neighboring cells may have a small correlation factor due to slew variations induced by first cell. Thus, we decided to characterize the mismatch impact on third cell to include this slew effect and obtain realistic measures.

$$\sigma_{chain} = \sqrt{\sigma_1^2 + \sigma_2^2 + ... + \sigma_n^2} \qquad (6\text{-}1)$$

$$\mu_{chain} = \mu_1 + \mu_2 + ... + \mu_n \quad (6\text{-}2)$$

$$M_{chain} = M_1 + M_2 + ... + M_n \quad (6\text{-}3)$$

$$t_{\max,\ mismatch\ (corner)} = \mu_{chain} + M_{chain} + 3\sigma_{chain} \quad (6\text{-}4)$$

$$t_{\min,\ mismatch\ (corner)} = \mu_{chain} + M_{chain} - 3\sigma_{chain} \quad (6\text{-}5)$$

Whereas nominal delay for a cell at a given slew includes the interconnect delay in our case, we can separate it out without affecting the methodology. We used spice analysis to verify the numbers. Theoretically, in an implementation, we can use the input and output slew calculated by STA tool to estimate the mismatch impact for each cell. The nominal chain delay is calculated by the STA tool itself. Cells can be characterized for a given corner to obtain a table with input slew and output load as the two axes for parameters, μ and σ. Subsequently, using table lookup method, we can calculate the required margins for a path with individual cell values.

6.5.3 Analytical prediction of mismatch to reduce characterization effort

To build the tabular models, we need to characterize cells at different supply voltage, temperature, corner, input slew and output load. However, that increases the amount of time required to characterize a library. Optimization techniques including mathematical interpolation for different slews, sizes and supply voltages can be used to reduce library characterization time.

From earlier analysis, we found that absolute mismatch increases linearly with slew as shown in Figure 6-14. The figure plots normalized σ-value (smallest σ-value taken as 1) of delay mismatch for a cell at SS corner, 1.05V and -40°C. As shown, the relationship holds true for various cells. Using linear interpolation, we can calculate the σ-value of a cell at any slew using just two slews. Equation (6-6) shows how we can calculate the mismatch σ using slew characterizations at two points only. Using this optimization method, the number of characterization runs for a cell can be reduced by 60% (for a library using 5 slews) as compared to normal characterization methods.

$$\sigma_x = \sigma_1 + \left(\frac{\sigma_2 - \sigma_1}{s_2 - s_1}\right).(s_x - s_1) \quad (6\text{-}6)$$

The second optimization possible for mismatch characterization is for supply voltage. Typically, each cell has to be characterized at multiple voltage points for a given corner. From the dataset, we extracted an analytical relationship between σ and nominal delay for different supply voltages. Equation (6-7) shows this relationship where d_1 and d_2 are the nominal delay at given supply voltage. The given equation holds true for the whole range of supply voltage from 0.85V to 1.35V for different buffers within 5% error margin. An error margin higher than 5% is possible for very large buffers at 1.35V where mismatch has negligible value. As nominal delays are already characterized for cells at different supply voltages, we need to characterize just one point for mismatch σ.

$$\sigma_2 = \left(\frac{d_2}{d_1}\right)^{5/3} \cdot \sigma_1 \quad (6\text{-}7)$$

Third optimization method is applicable to drive strength. Using data available for one buffer, we tried to predict the amount of mismatch in delay for higher drive buffers. Figure 6-13 shows the schematic of a simple buffer with PMOS and NMOS transistor widths marked as WP and WN respectively. Each transistor itself is made up of multiple smaller transistors to achieve larger effective size. As clock networks are predominantly made of buffers, any reduction in number of buffers to characterize is beneficial. Equation (6-8) shows equivalent width calculation for NMOS or PMOS at each stage where W_1, $W_2 \ldots W_n$ represent width of individual fingers (or parallel transistors). The equivalent width calculation is based on fact that n parallel transistors of equal width will have a smaller probability distribution curve than a single transistor of same width. However, the distribution will be wider than a single transistor of width equal to the sum of widths of n transistors. RDF mismatch is inversely proportional to square root of transistor area and we have used the same fact for equivalent width calculation. Equation (6-9) shows calculation of k-factor for rise or fall edge for a given buffer. For rise edge, it uses the equivalent width of 1st stage NMOS and 2nd stage PMOS whereas for fall edge it uses 1st stage PMOS and 2nd stage NMOS. Equation (6-10) shows how using k-factors for two buffers, we can calculate σ-value of a buffer given σ-value of another buffer. The error difference remains less than 0.8ps in absolute terms for all buffers. Mismatch calculation based on transistor widths may allow optimizing a buffer for mismatch value without characterizing it first. It may be less useful for reducing characterization runs.

$$W_{eqN/eqP} = \sqrt{W_1^2 + W_2^2 + \ldots + W_n^2} \tag{6-8}$$

$$k = \sqrt{\frac{1}{W_{eqN}^2} + \frac{1}{W_{eqP}^2}} \tag{6-9}$$

$$\sigma_2 = \frac{k_2}{k_1}\sigma_1 \tag{6-10}$$

Skew is difference in delay of two paths. For two balanced paths, the cells should have same standard deviation of their respective delays. Average shift seen by both paths should be same and thus cancel out. Moreover, standard deviation of difference of two random independent variables (σ_{skew}) is equal to rms addition of individual standard deviations (σ_{path1}, σ_{path2}). Pulse width is similar to skew as in it is the difference in delay of two edges for a path. Pulse standard deviation (σ_{pulse}) is calculated similar to skew using edge delays (σ_{rise}, σ_{fall}). However, the average shift for two edges will be different because of different transistor sizes in their respective paths. As such, the net average shift for the pulse (μ_{pulse}) will be a difference in rise and fall average shifts (μ_{rise}, μ_{fall}). Equations (6-11) and (6-12) show how mismatch mean shift and standard deviation can be calculated for skew whereas equations (6-13) and (6-14) demonstrate the same for pulse-width.

$$\sigma_{skew} = \sqrt{\sigma_{path\,1}^2 + \sigma_{path\,2}^2} \tag{6-11}$$

$$\mu_{skew} = \mu_{path\,2} - \mu_{path\,1} \tag{6-12}$$

$$\sigma_{pulse} = \sqrt{\sigma_{rise}^2 + \sigma_{fall}^2} \tag{6-13}$$

$$\mu_{pulse} = \mu_{fall} - \mu_{rise} \quad (6\text{-}14)$$

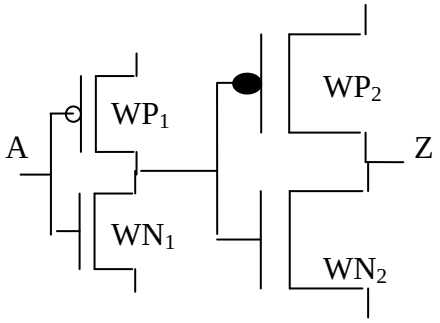

Figure 6-13: Schematic of a clock buffer with N & P transistor widths labeled

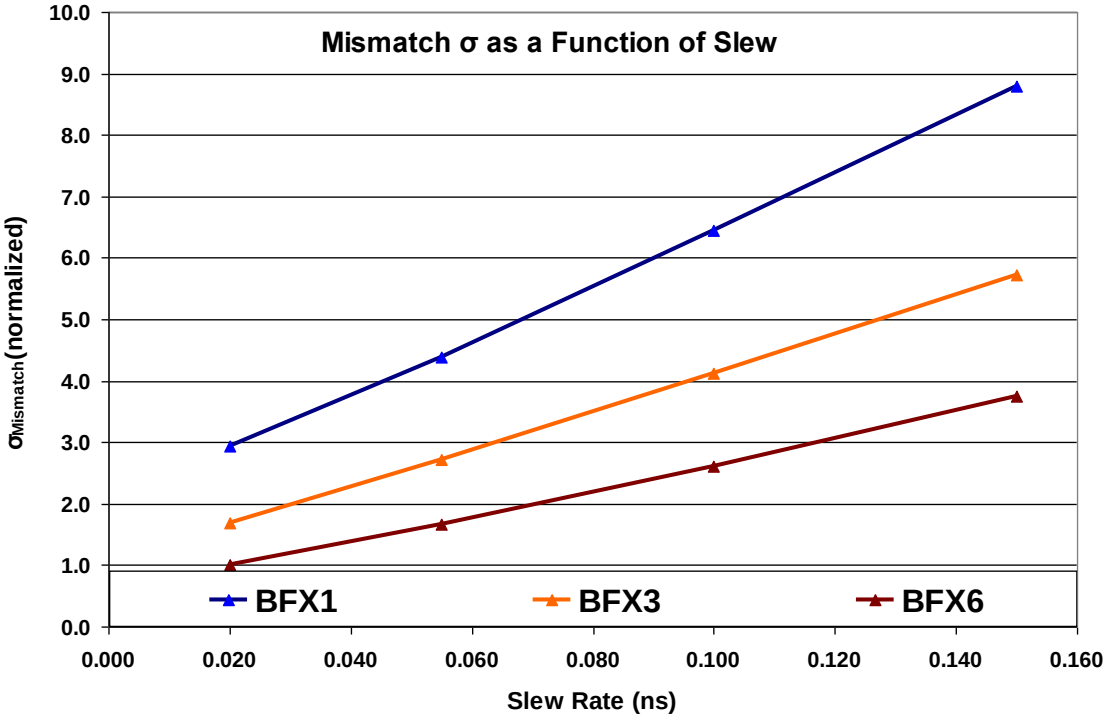

Figure 6-14: Mismatch σ (normalized with smallest value as 1) with slew at SS, 1.05V, -40°C.

6.5.4 Prediction vs. Monte Carlo method

We calculated and compared the impact of mismatch on insertion delay, skew, and pulse width at different corners and slew rates for different clock buffers. The values have been normalized with the largest x-axis value taken as 1 and y-axis values calculated for the new number to preserve the shape of the graph. We calculated the impact of mismatch on different configurations through our approach using cell characterized mismatch and then extracted the same in spice paths. The normalized $\mu \pm 3\sigma$ limits along a path have been compared and plotted in following graphs. Calculated mismatch has a good accuracy for most paths except for initial stage where the variation characteristics are smaller due to non-varying input pulse. There

is a small error (in absolute term) in average shift within Monte Carlo accuracy. The calculated mismatch on delay for a clock path composed of different buffers agrees well with the observed mismatch at slow corner and 55ps slew as shown in Figure 6-15.

CAD tools synthesize clock networks typically by maintaining a constant slew along a path. The load factor from one drive to another will change to maintain this slew. We calculated the impact of different slews on a path and compared with the observed value. The approach still maintains the required accuracy. STA differentiates between different corners through characterization files. To be consistent, we compared the mismatch impact between calculated values from our approach and observed values from simulation for a path on different corners with good accuracy. Figure 6-16 shows delay mismatch error percentage with respect to spice extracted mismatch for different configurations of SS/FF corner and 55ps/100ps slew rate. Except the 1st stage, error percentage stays below 10%. As we are comparing mismatch, it is less than few ps of error.

Figure 6-17 plots calculated and observed value of mismatch impact on skew and Figure 6-18 plots them for pulse-width. They show a good consistency between spice extracted and calculated values. Mean shift has lesser accuracy in calculation than standard deviation and affects pulse-width mostly in μ-3σ region.

We obtained a good level of accuracy through mismatch aware STA to calculate the impact of mismatch on delay, skew and pulse-width in a clock path under various configurations varying corner, supply voltage, temperature, slew rate, etc. Verification using Monte Carlo simulations validates the said approach. The given approach can be easily implemented in current design flows with small overheads.

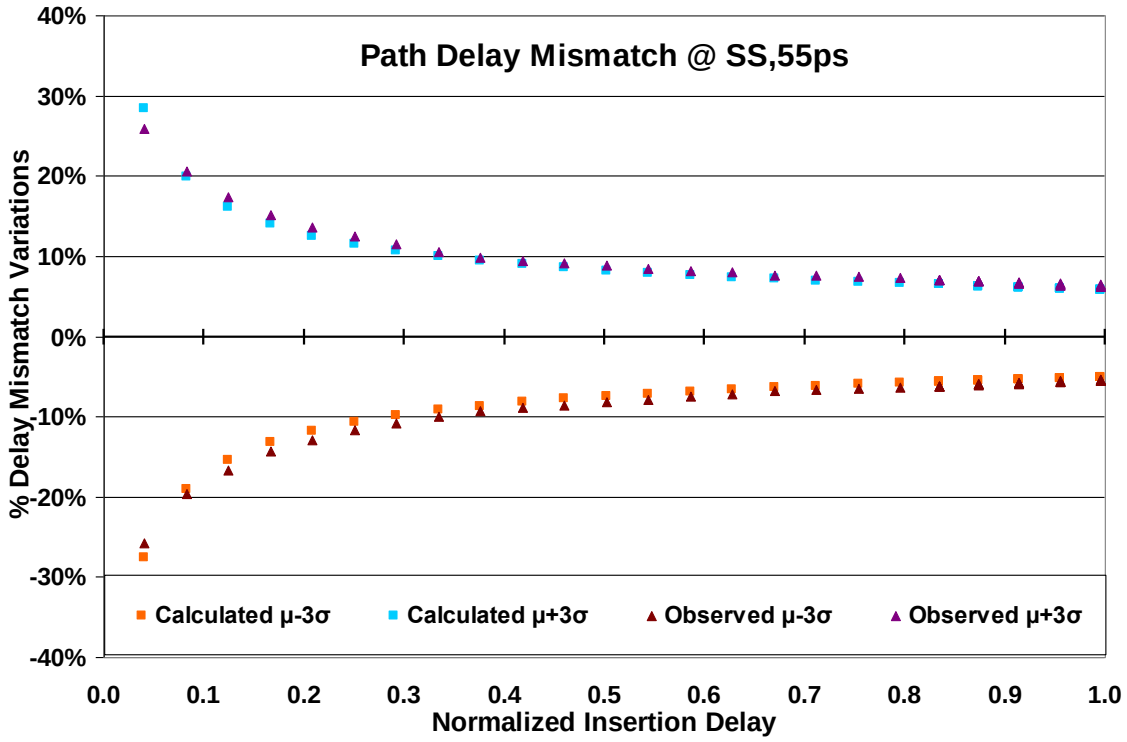

Figure 6-15: Calculated and spice extracted delay mismatch in a clock path at SS, 1.05V

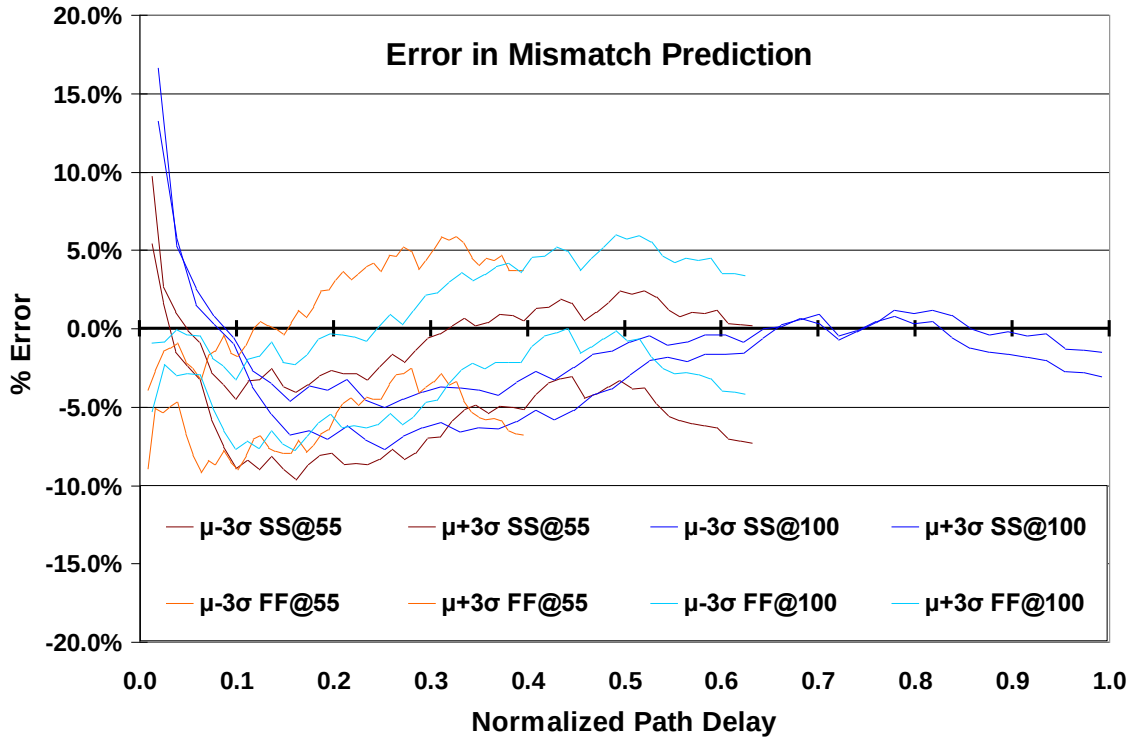

Figure 6-16: Error percentage for calculated mismatch for SS/FF corner and 55ps/100ps slew

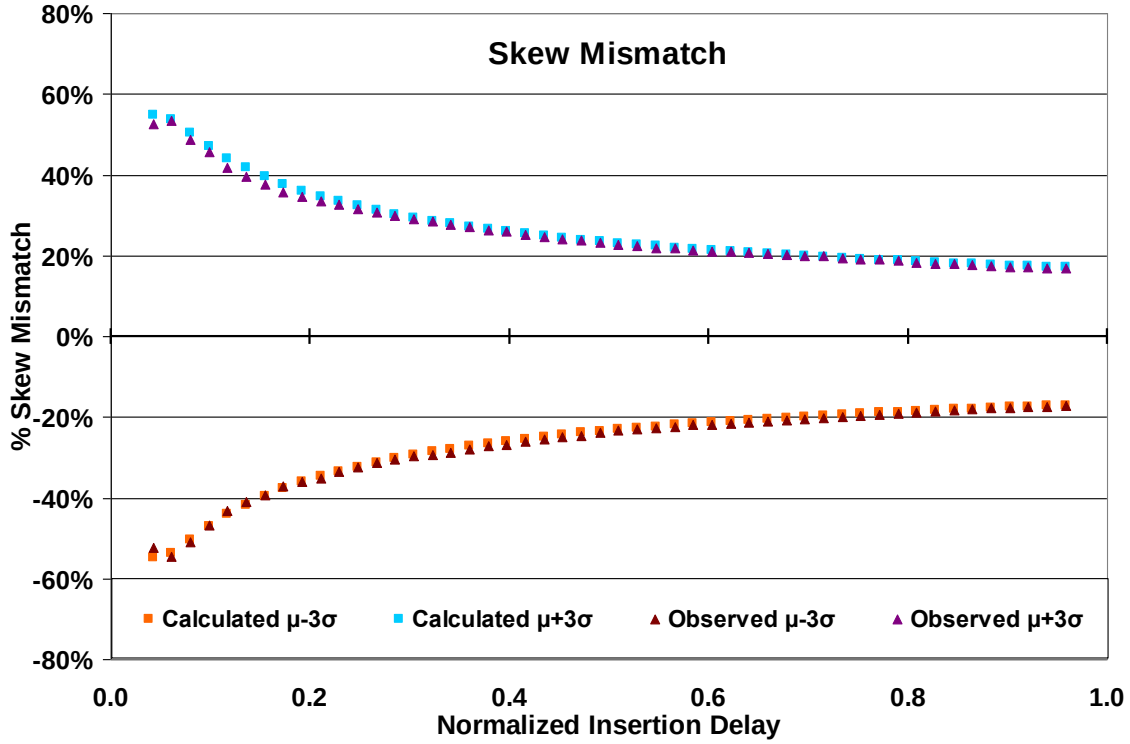

Figure 6-17: Calculated and spice extracted skew mismatch in a clock path at SS, 1.05V

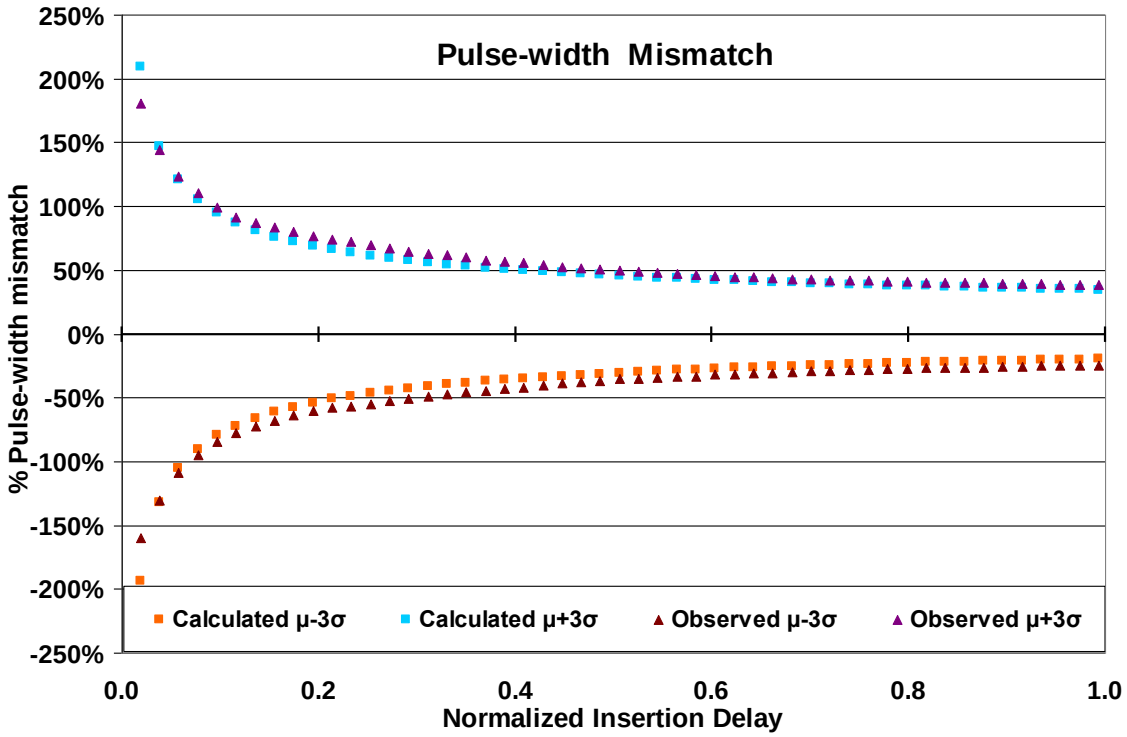

Figure 6-18: Calculated and spice extracted pulse-width mismatch in a clock path at SS, 1.05V

6.6 Hold fix analysis

Presence of hold violation in a clock system can cause timing failure. Thus, hold margins are calculated based on delay variations and accordingly extra cells are padded up to remove any violation. These hold fix cells are special as their purpose is to increase the delay in minimum possible area. Hold fixing is done after routing is over and as such, these cells do not have much room. They have minimum interconnect load as they are padded up one after next. Hold fix cell usage is minimized as their delay tends to vary a lot with variations and presence of extra cells can affect setup time margins.

Figure 5-4 shows the schematic of a delay buffer. Their unique configuration tends to increase the impact of local random mismatch and have relatively high mismatch σ. A large μ value is beneficial as it increases the minimum delay. Thus, it is important to verify that mismatch does not create a hold situation in a hold fixed path. The probability is stronger in very small paths requiring large hold fix or paths where margin after hold fixing is small. We looked at the impact of mismatch on different cells that can be used for hold fix including delay buffers, small inverters, small buffers, NAND, NOR, etc. For delay buffers, we also looked at different delay length cells created by adding extra stages in the cell. The purpose is to find the most suitable cell in presence of mismatch to fix a hold of 100ps, 500ps, 1ns and 2ns. The setup used is similar to Figure 4-8, except the interconnect load that is kept negligible. We look at amount of mismatch delay variations for each cell at 1.05V &

-40°C for SS and FF corners. Inverting cells are combined to have a net non-inverting cell.
Results show that corner delay dominates over mismatch delay in case of hold fix. Minimum delay including mismatch for delay buffers is still more than that for other cells when optimizing for minimum area. For a single cell, mismatch impact is in few ps as compared to corner delay that is in many 10s of ps. For 100ps hold, delay buffer with minimum delay length comes out to be the best solution in terms of area. Cells with larger delay length increase area used. As the amount of hold to fixed increases, delay buffers with larger delay length gives the best hold-area compromise as they have better delay to area ratio.
Although very small inverters and buffers do not give the best solution for fixing hold, they are not far behind in case of smaller hold values. Typically, clock system has only small hold values to be fixed. Only in exceptional cases can it be in ns region. The area required to fix a given hold value using inverters and buffers is larger, but they provide a much better scaling behavior with supply voltage and temperature as shown in Figure 3-15 and Figure 3-16. To achieve larger delay to area ratio, HVT cells or Large L cells can be used that increase the net delay. In case of HVT cells, mismatch is higher than normal cells but it is still less than delay buffers. In case of Large L cells, mismatch is even lesser than normal cells for larger delay value proving to be a very good replacement for delay buffers. Large L cells may still not be able to compete with delay buffers in terms of area but they come a close 2nd and provide a robust solution to local as well as global variations effect. Cells may be created with higher gate length to increase the nominal delay but they may add extra mask steps.

6.7 Optimization solutions

Local mismatch has maximum effect at extreme corners. However, it is as much an issue in between as it is on corners. SS corner is typically equal to 3σ point in global delay distribution and has a probability of less than 0.3%. In physical terms, arriving at such a point has a much lower probability being limiting case variation of multiple individual factors. On top of that, foundries use product centering that can shift the whole distribution towards left. A 3σ point in local delay distribution also has a probability less than 0.3% on top of SS corner. Combining the two cases, the probability of having worst-case global variation and worst-case local variation is in ppm range. Thus, using worst-case mismatch margins in a design can be overly pessimistic. A smaller amount of mismatch margin equivalent to 3σ at typical or 2σ at corner can be used with relatively low occurrence probability and without sacrificing too much performance.
The second approach to mismatch optimization is through robust design practices pre-empting chances of large local mismatch effect. Such an approach requires identifying and eliminating susceptible structures based on usage and optimizing devices to minimize local mismatch effect. ASIC designs are highly dependent on intrinsic cell delay and any optimization that results in a net increase in intrinsic delay may not be practical. That is to say, a slower transistor with reduced mismatch is not necessarily better than a faster transistor with large mismatch. As stated

earlier, any optimization will affect the PPAY (Performance-Power-Area-Yield) point in the design space. Getting the right balance between them is the key to a feasible and viable design.
Optimization has to be done with a specific target in mind that in turn will define how much local mismatch is acceptable. We have listed multiple approaches to optimization that affects local mismatch although not necessarily for reduction.

6.7.1 Frequency optimization

Frequency optimization is the primary focus of an ASIC design followed by power. If a design meets the cut-off performance level under worst-case conditions including, then it is acceptable. How, getting a good enough design in the first run is becoming more and more difficult. A part of lost performance goes into margins. Reduced margins mean better performance. Frequency optimization is done to improve the overall performance level of a design. However, if speed is not the critical parameter in some part of design, then we can sacrifice on performance to optimize other parameters.

6.7.1.1 Different V_{th} cells

Threshold voltage has a direct impact on mismatch and delay. As seen earlier absolute mismatch $\sigma_{HVT} > \sigma_{SVT} > \sigma_{LVT}$ but intrinsic delay $M_{LVT} < M_{SVT} < M_{HVT}$. Clock paths are most susceptible to mismatch and best suitable for LVT cells. It will help to reduce insertion delay as well as mismatch in pulse-width and skew. Percentage mismatch may be higher depending on interconnect delay but the absolute mismatch is smaller and thus the net delay is smaller than SVT cells. Smaller threshold voltage will affect leakage power in a clock network but presence of clock gating cells can constrain that leakage. Larger drive current may also allow for larger load and thus longer interconnects reducing the number of required cells in a path. Longer the feasible interconnect length, longer the clock signal can travel without affecting its waveform. Longer interconnects are specific to clock tree distribution where buffer delay constitutes most of the insertion delay and minimizing this delay is important.
LVT cells can be used in setup delay critical paths. Instead of increasing the frequency, some of the data path cells can be replaced by LVT cells to reduce max delay through reduced mismatch as well as reduced intrinsic delay. Again, it affects the leakage power but as not many paths should have a setup violation, the impact is limited. Leakage power can be compensated by using HVT cells to replace data cells in hold violating paths instead of padding extra cells.
SVT cells are typically used everywhere except for specific purposes like clock where LVT or leakage reduction where HVT cells are used. In terms of mismatch, SVT cells are best suited for data paths where there is no concept of skew or pulse-width. Data paths are mostly concerned with worst-case delay including mismatch that can be quite high due to small cell sizes. SVT cells can also be used for slower clocks that have higher available margins. Any design has multiple clocks of which only the principal clock works at maximum frequency generally. If certain clock or

paths have high available margin, then LVT cells can be replaced by SVT cells in those cases.
HVT cells have the highest intrinsic delay and mismatch. They are avoided on all delay sensitive paths but frequently used for leakage reduction. HVT cells can be used in standby or sleep mode clocks that work at very low frequency and are just required to maintain the basic system functionality. When powered up, drive responsibility can be shifted to faster clocks. Designs also use built in test circuits that can use HVT cells, as they are delay insensitive. Similar logic can be applied to any asynchronous system in the design. Control logic like set/reset lie in this category. Use of HVT cells can help to reduce overall leakage power consumption. HVT cells also make good candidates as delay cells for hold fixing. They have smaller difference between rise and fall edge than delay buffers and thus provide almost equal hold fix value for both edges. Mean shift due to mismatch will only help to increase the minimum delay.

6.7.1.2 Drive strength

Fundamental principal behind local mismatch says larger size (or drive strength) means smaller mismatch. However, it also means larger dynamic power consumption as well as larger leakage power. Most designs today are aggressively optimized for power reduction and simply increasing cell drive is not a feasible option. Although the drive capacity of a higher drive buffer is larger, it does not guarantee a proportional increase in actual interconnect load due to increased resistance. Thus, the marginal utility of increasing buffer drive reduces. Even from mismatch point of view, there is negligible gain moving from BFX4 to BFX6. A clock network can be constrained to use only medium drive buffers (BFX3 and BFX4) to reduce overall mismatch. Removing both high drive and low drive buffers will balance load distribution and power consumption. Very large fanout loads can be divided between two buffers maximizing the marginal utility. Lesser number of buffer drives will also help to balance the tree and reduce skew variability. Equivalent interconnect load for all buffers will provide a regular structure and a good scaling behavior with process, supply voltage, and temperature. At low V_{DD}, pulse-width mismatch is high and thus low drive buffers may not be best suited in spite of power restrictions.

6.7.1.3 Adaptive body bias

Adaptive body bias (ABB) has been proposed as a technique to reduce leakage and control frequency variations in microprocessors [59]. In ASIC products, the need is more to pass the cutoff frequency. Thus, a simpler form of ABB with only single level of forward or reverse body bias applied at different clock levels can be used. The technique uses only two biasing levels (forward or reverse) easily available from existing power lines. Moreover, bias is level specific and not buffer specific simplifying the control circuitry. Clock levels here represent the different distribution levels of a clock tree [8]. Forward bias helps to improve the V_{DD}/V_{th} ratio and thus reduce mismatch. Reverse bias acts in the opposite direction and helps

to reduce leakage. The applied body bias can be controlled by DVFS (Dynamic Voltage and Frequency Scaling) system applying forward bias at high frequency and reserve bias at low frequency. Differentiated applied bias at different clock levels can allow for reduced skew variations without sacrificing too much on leakage power through larger forward bias voltage at leaf nodes. Pulse-width variations can be reduced by applying larger forward bias in levels with smaller buffers.

6.7.1.4 Supply voltage

Supply voltage is one of the biggest factor affecting mismatch variations. As we have seen earlier, for a given threshold voltage, higher supply voltage leads to smaller mismatch variations. Schemes like DVFS already control supply voltage. However, they do not take into account variations. Techniques have been proposed to control supply voltage and body bias using ring oscillator frequency as references in post-production [85]. Using separate supply for clock and data paths can allow for better variation control. We can apply higher voltage on clock and lower on data to control and balance frequency and power consumption. Voltage control can be activity based using schemes similar to DVFS. Fine grain control can help to eliminate worst cases only. The issue with separate clock and data supplies is current mesh architecture used for standard cell design. It puts all standard cells in between power & ground lines and then connects the cells according to required routing. The overhead of implementing two separate supplies everywhere can be high. The overhead can be limited by implementing separate clock supplies only for local clock distribution limiting the required area with two supply lines.
Dynamic block level supply voltage scaling based on error rate provided by software input or RO frequency can help to have a broader frequency control with necessitating separate supplies. Dynamic structure will help to reduce power consumption in low power regions.

6.7.2 Power optimization

Low power has become the buzzword in microelectronics industry today and designs are aggressively optimized to minimize dynamic as well as leakage power. Based on application being standby time critical or peak power critical, one type of power optimization may get priority over other. Like performance, power consumption may also have a cutoff value. However, unlike performance smaller the power consumption better it is as it increases marketing value of the product. Power reduction techniques like HVT cells or low supply voltage increase mismatch that in turn increases variation in performance. We have already seen the usable configurations of both and their impact. Other than that, there are other methods that used solely for power reduction.

6.7.2.1 Large L cells

Standard cells in 45nm technology typically have 45nm transistor gate length as per the technology. There are also large L cells that use a gate length higher than 45nm; typically 1.5-2 times the minimum gate length. Large L cells have higher intrinsic delay but smaller power consumption. As such, they are used to replace nominal cells in non-critical paths to reduce power consumption. Large L cells also have smaller global variations as their drawn gate length is larger than the critical length. Moreover, larger gate area and length results into smaller local random variations. Thus, these cells prove to be more robust.

Large L cells can be used in small data paths where there is a big margin available for path delay as compared to clock frequency. They can replace nominal cells in hold violating data paths and serve as delay cells to provide extra delay. Large L cells can replace nominal cells in non-critical paths like slow clocks and asynchronous paths like set/reset. They can replace HVT cells for power reduction. One major place large L cells can be used is in flip-flops and clock gate. It can help to reduce the variations in required setup and hold time but may increase the nominal clock to output delay. Thus, it is useful only in non-critical paths.

6.7.2.2 Stack forcing

Stack forcing is another method used for power reduction. The NMOS transistor of a cell is broken into two half-width transistors connected in series between ground and output. Figure 6-19 shows the method. Increased nominal delay and mismatch variations make it necessary to use it only in paths with big margins compared to frequency. It can be used in data paths where leakage power is a big concern. Stack forcing reduces global variations.

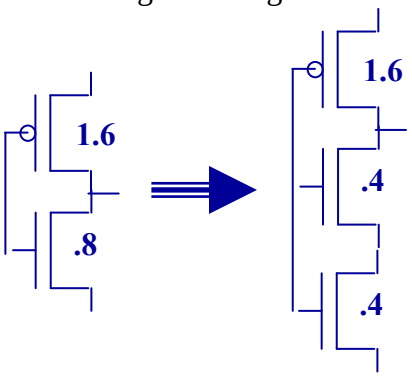

Figure 6-19: Stack forcing

6.7.3 Clock network optimization

Clock network optimization includes design rules/guidelines as well as clock buffer optimization. The purpose is to make clock tree more robust against mismatch variations. We have already seen that there is negligible change in mismatch above BFX4. In addition, the 1st stage is mainly responsible for mismatch variations in a

clock buffer, except in low drive buffer like BFX1 where the output stage also has a significant role. Restricting clock buffers to medium drive at BFX3 and BFX4 can produce an optimum clock tree from skew and pulse-width perspective.

6.7.3.1 Clock buffer

Clock buffers themselves can be improved for mismatch by increasing the input transistor sizes. Reducing the output to input stage ratio can increase intrinsic delay and thus the path insertion delay. It will also increase the input capacitance and thus reduce the interconnect load. For delay critical paths, it may not be viable but can be used in other paths. In designs that are limited by skew or pulse-width, it can be a good option. This optimization can be limited to low drive cells that are most sensitive to mismatch. These cells show a larger improvement ratio in low V_{DD} designs where mismatch can have a big effect.

6.7.3.2 Design guidelines

Design guidelines/rules constrain and steer clock network construction. They are meant to provide an optimum solution that may not be the best one but limits effort and resources. Instead of reducing mismatch of susceptible paths or limiting its effect through margins, we can make the paths more robust against mismatch. They vary in their scope and ease of utilization. A typical timing analysis tool applies OCV margins on delay on each path and accordingly calculates the available margin for pulse-width or skew. However, mismatch being a random phenomenon, there is some averaging effect in skew and pulse-width along a path. Thus, delay margins are pessimistic by definition. If we apply margins on skew and pulse-width directly, we can reduce margins and still be within the statistical limits. It should also be noted that for skew and pulse-width, absolute mismatch is more important as it determines the maximum data path length through absolute pulse-width and setup & hold margins through absolute skew. Percentage mismatch tends to be misleading as high mismatch percentage on a small path may be better than low percentage on a long path.

6.7.3.2.1 Slew rate

Clock paths typically have only a maximum transition constraint. Thus, it is possible to have two paths one having maximum slew and the other having very low slew. Just from mismatch perspective, the worst-case difference is less than that for two paths at maximum slew. However, different slew amounts to different delays increasing nominal skew. Even if the two paths were to have same delay at nominal conditions, PVT variations can rapidly increase the difference. Mismatch values at two paths unbalanced due to PVT variations can make the case worse. Applying a minimum slew constraint, we can ensure balanced clock paths that scale similarly for skew.

Maximum slew constraint in a clock path is independent of the driving cell. However, the impact of slew on mismatch is not the same on a low drive buffer and

a high drive buffer. It is possible to apply slew constraints as a function of the driving buffer size. Larger buffers can sustain high interconnect loads with buffer fanout to regenerate the signal. Smaller buffers will have a lesser maximum slew constraint limiting the amount of mismatch.

6.7.3.2.2 Fanout

Penultimate buffers generally have very large fanout to drive multiple clock-gate/flip-flops. Such an arrangement reduces the amount of skew among neighboring flip-flops. Typically, large fanout generates higher output slew rate at the driving buffer increasing mismatch effect. However, penultimate buffers are very high drive buffers and have minimal mismatch. Thus, the arrangement will have minimal effect on skew or pulse-width due to mismatch. Larger fanout for small buffers will have a large mismatch effect. It may be possible to increase the maximum slew constraint for penultimate stage reducing the number of paths and thus skew without having adverse effects on mismatch. It may be possible to replace some large buffers driving large fanout with inverters. As the paths are not highly resistive, signal will not degrade much.

6.7.3.3 Inverter clock tree

Clock tree are generally driven by clock buffers as they can regenerate a degraded waveform as well as decouple slew variations between input and output pins to some extent. The same is not true for inverters that also provide a larger input capacitance for same drive capacity. An inverter tree can make it obligatory to use combination of inverters to avoid pulse-width variations. Two same inverters in series will give a net pulse change of zero though two different inverters in series can give a net non-zero pulse change. Thus, a clock inverter tree needs to be fully balanced and regular to be useful.
Clock inverters are much less affected by local mismatch especially in low V_{DD} region where clock buffers encounter high mismatch. Due to their dual configuration, they also show a much smaller mismatch impact on pulse-width. It may be useful to implement inverter clock tree for low V_{DD} designs. Their reduced insertion delay can compensate for extra cells required to drive the tree. Clock inverters can also replace low drive cells and break the path into two segments maintaining the delay with an improved mismatch.

6.7.4 Data path optimization

Data path consists of logic cells connected between two flip-flops in a synchronous design responsible for all the logic functions. The objective in a data path is to have maximum logic in smallest path as compared to clock path where the objective is to minimize the number of cells for maximum path. Data path is predominantly intrinsic delay with a small interconnect delay required for signal transmission across the circuits. Data path cells are typically small and the principal parameter is delay. As explained earlier, mismatch affects delay but has a very small probability

of having a worst-case corner as well as worst-case mismatch. However, the magnitude of mismatch in data paths can be larger due to small cells and small parasitic delay. Data path lengths are limited by clock frequency. Data paths can be made more robust to avoid adding extra margins for mismatch.

6.7.4.1 Multi stage cells

Mismatch variation of a multi stage cell is less than equal to the rms addition of mismatch variation of its individual stages. For e.g., $\sigma^2_{AND} \leq \sigma^2_{NAND} + \sigma^2_{INVERTER}$. Mismatch variation can be less as the slew rate on the inside node of a multi-stage cell will be smaller than that seen on the connecting node between two cells separately. Thus, using long multi-stage cells computing various functions will provide a smaller overall mismatch as well as smaller path delay.

6.7.4.2 Complex vs. Simple cells

Logic functions require many complex calculations that can be performed using multiple simpler cells or lesser number of complex cells. Typically, the functionality requires at least four inputs. A 4-input functionality can be implemented as a single or dual stage 4-input cell or multi-stage 2-input cells. Overall percentage mismatch for both will be similar but a simple cell implementation will have higher intrinsic delay and thus higher absolute mismatch. Signal transmission is better in simple cells than in complex cells. Complex cells also have higher input capacitance.

6.8 Approach: Silicon vs. Simulations

Local mismatch is a random variation that makes it difficult to have a one-to-one matching between spice and silicon. Various phenomena can induce difference between silicon and simulation results or constrain the testability of paths. Parasitic difference between model and silicon, systematic effects, lack of knowledge about exact point in local and global variation space, silicon to model error, process centering, test equipment error, non-testable paths, limited testable pads, maximum test frequency, etc do not allow a one-to-one matching between silicon and simulation.

Monte Carlo spice simulations for mismatch on corners define delay and mismatch σ. The delay-σ pairs at corners create the encapsulation within which all silicon results should lie. A delay-σ pair is unique for each die on silicon. If results from a set of different dies create a shape encapsulated by the boundaries created from spice simulations, it proves the validity of the variation model. In case there is a big difference between spice and silicon results, it can point to a fault in variation model.

The parameter most easily accessible in test results is oscillation frequency (inverse of delay). Generally, test structures are constrained to ring oscillators (RO) with one point accessible from test pad, as shown in Figure 6-20. RO delay or rise-to-rise delay is a sum of rise-to-fall and a fall-to-rise delay through the same RO. Thus,

measured delay takes into account both rise and fall delay mismatch. A set of similar ROs on a die, lie on the same point in global variation space and should have a single nominal delay. However, mismatch variations will change the oscillation frequency for each RO. Each set delays will have a Gaussian like distribution defined by a mean value M and standard deviation σ.

6.8.1 Silicon test

To measure local mismatch effect, we need a small chain (< 100 cells) RO made up of a single type of cell (preferably buffer) except the inverting cell. The test cell should be of small drive to maximize local mismatch effect. Small chain will have a reduced averaging effect. Significant number of instances (~1000) of the test RO needs to be placed on each die to enable a good distribution and enable accurate statistical data. The ROs need to be placed geometrically closed together to suppress systematic effects. Samples from statistically significant number of dies will be needed (few hundred) to obtain a wide range of process corners. Samples may be needed from same wafer, different wafer in same lot and from different lots to cover the whole range of variations. From each sample, we need to extract the oscillation frequency. For each set of geometrically closed together ROs on a single die, we calculate the mean and standard deviation of the delay distribution. The resulting data can be put together in a tabular format as shown in Figure 6-21.

6.8.2 Simulation

Spice simulation netlist needs to match the silicon RO. We use the netlist of the RO used in silicon with transistor level extracted parasitics, necessary to minimize difference between the two. Using same V_{DD} and T conditions as in silicon test, we measure the oscillation frequency for 5 major corners (SS, FF, TT, SF, FS). To achieve better results, custom corners can be created with 1σ and 2σ variations. Monte Carlo simulations with local mismatch only are done at each corner with at least 1000 samples and rise-to-rise delay is extracted. From each set, we can obtain mean and standard deviation.

6.8.3 Matching silicon to simulation

Plot the points obtained from simulation on σ vs. mean delay graph and join the points to create the encapsulation. Now plot all the statistical data obtained from silicon results on this graph as in Figure 6-22. If most of the points lie within the encapsulation and follow a Gaussian kind of distribution with very few points lying near the boundaries, it proves the validity of the model. Small deviations in X-Y or angular direction are possible due to process shift and systematic effects. The figure represents mismatch effect but real shape may vary from the one shown in figure. Further verification can be done by extracting 1-to-0 delay from silicon results and creating a similar graph.

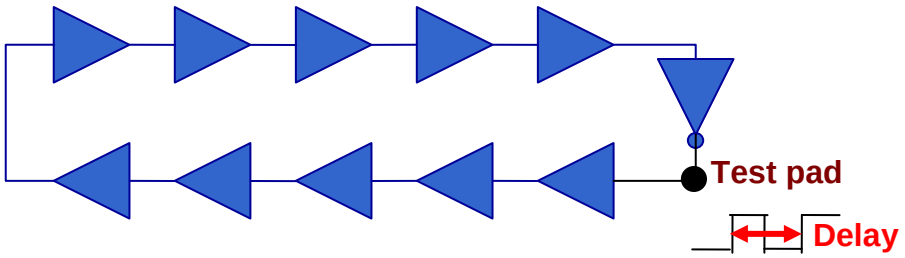

Figure 6-20: Ring oscillator test circuit

Die Number	RO number	Delay	Mismatch, Delay
1	1	0.41	σ_1, M_1
1	2	0.52	
1	3	0.43	
2	1	0.62	σ_2, M_2
2	2	0.53	
2	3	0.46	
3	1	0.57	σ_3, M_3
3	2	0.45	
3	3	0.43	

Figure 6-21: Example of silicon data set and its statistical output

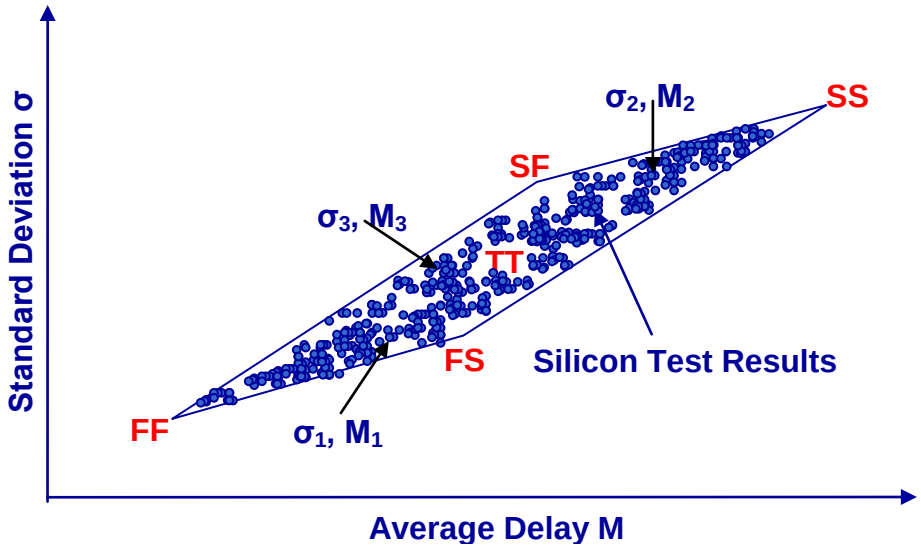

Figure 6-22: Silicon vs. Spice verification method

7 Conclusions and Future Work

7.1 Conclusions

This thesis is focused at estimating and reducing global and local random mismatch effect on timing in ASIC designs. The differentiating aspect is we limited ourselves to using design techniques to reduce turnaround time. The aim is to make circuits more robust keeping in mind the tradeoffs involved and thus enable a direct comparison of cost and benefits. We took a multipronged approach for minimizing on-chip variation margins required in timing corner approach. We analyzed the principal components affected by mismatch (local & global) and concluded that through robust design and on-chip variation margins, we can control its impact to within manageable limits for current nodes. Exotic solutions like using different transistor structures or process technology can be used in advanced nodes when the magnitude of variations will be too high to be contained through design methods only.

A mismatch variation aware static timing analysis methodology was proposed to calculate path specific margins tailored for individual corners. Characterization method for cells requiring minimum time was proposed while maintaining the accuracy. Analytical equations to speed up the characterization process were elaborated along with the error margin introduced by them. Spice simulations confirmed the accuracy of proposed methodology. It can be implemented in current CAD tools with small overhead.

Specific optimization strategies targeting delay or power for clock paths were proposed using a mix of parameters including threshold voltage, gate length, supply voltage, and drive strength. Advantages and disadvantages of each were listed that can help to choose the best strategy for a given application in the presence of mismatch. A set of design rules with subjective gains to limit mismatch impact on clock paths were given that will help to create a more robust design.

An application specific optimization strategy in ASICs was proposed to limit the impact of global mismatch. A subset of clock buffers in the same library optimized with specific applications in mind can limit pulses-width variations. The proposed methodology requires characterization of a small subset of cells and modifying few rules to include a target application parameter that will help to choose the specific subset. The approach lies in between full custom design and standard cell design using the best of both. The gain is most in low voltage region where pulse variations are highest.

ASIC designs using standard cell method typically use buffer clock tree because of their signal regeneration capability. We looked into the advantaged and limitations of an inverter clock tree in the presence of global and local mismatch. Whereas the gains are limited in high voltage region, low voltage designs can benefit significantly to reduce mismatch variations impact. Reduced number of transistors will enable increased power savings that are important in this region.

We also put forward an approach to measure the model accuracy with simple RO delay measurements. The approach allows verification using simple test circuits that can be and are embedded in wafers and dies. It allows for fast measurement of local or global mismatch and confirms model accuracy.

7.2 Future work

This work was limited to looking at global and local random mismatch at design level. There have been successful efforts to contain mismatch at transistor or architecture level. However, most of our work was focused towards techniques to reduce mismatch at cell or path level. Such an approach has benefits in terms of fast implementation time and minimum overheads, i.e. minimum cost to benefit ratio. However, it also limits the options and efficacy to reduce mismatch. It may be of interest to have a comparative study of different methods with cost and benefit as target parameters. Amount of available time and resources will determine which method is more useful for a given design. High cost-high benefits approach involving technology modifications can be better suited for future technology development whereas new architecture development can be used in advanced design development during technology ramp-up.

A comparative study for effectiveness of different approach mentioned in this thesis will require actual implementation of full design using those methods. Many different implementations have to be done to enable a clear distinction on advantages and tradeoffs. Some of the principal methods are inverter clock tree, clock tree using medium drive buffer only, use of large L cells, separate clock and data supply with dynamic voltage control, minimum slew constraint, application oriented standard cells, clock cells with non-equal rise and fall delay, etc.

Variation aware STA can be implemented in some CAD tool and included in the design flow. A comparative study of STA, variation aware STA, and SSTA can be done to demonstrate the cost to benefit ratio of any approach. Without demonstrating the costs associated with an approach and the advantages in comparison to other approaches, it will be difficult to justify or promote use of any method.

Other methods involving perfectly (almost) balanced clock trees using same buffers, equal load/fanout, regular buffer insertion, consistent slew, etc. have to be implemented to illustrate its advantages and disadvantages. Globally unbalanced but locally balanced clock tree similar to Globally Asynchronous Locally Synchronous design can give a good approach to reduce skew with lesser efforts. Mesh and differential clock distribution systems are good candidates to the traditional clock system for more variation control.

8 Bibliography

[1] A. Asenov, A. Cathignol, B. Cheng, K. P. McKenna, A. R. Brown, A. L. Shluger, D. Chanemougame, K. Rochereau and G. Ghibaudo, "Origin of the Asymmetry in the Statistical Variability of n- and p-channel Poly Si Gate Bulk MOSFETs," IEEE Electron Device Letters, Vol.29, No.8, 2008, pp. 913-915.

[2] A. Asenov, A.R. Brown, J.H. Davies, S. Kaya, G. Slavcheva, "Simulation of intrinsic parameter fluctuations in decananometer and nanometer-scale MOSFETs," Tran. on Elec. Dev., Vol. 50, Issue 9, Sept. 2003, pp. 1837 – 1852.

[3] A. Asenov, G. Slavcheva, A.R. Brown, J.H. Davies, S. Saini, "Increase in the random dopant induced threshold fluctuations and lowering in sub-100 nm MOSFETs due to quantum effects: a 3-D density-gradient simulation study," Tran. on Electron Devices, Vol. 48, Issue 4, April 2001, pp. 722 – 729.

[4] A. Asenov, S. Kaya, J.H. Davies, "Intrinsic threshold voltage fluctuations in decanano MOSFETs due to local oxide thickness variations," IEEE Tran. on Electron Devices, Vol. 49, Issue 1, Jan 2002, pp. 112 – 119.

[5] A. Asenov, S. Kaya, J.H. Davies, and S. Saini, "Oxide thickness variation induced threshold voltage fluctuations in decanano MOSFETs: a 3D density gradient simulation study," Superlattices and Microstructures, 2000, vol. 28, pp. 507-515.

[6] A. Chandrakasan, W.J. Bowhill, and F. Fox (Editor), Design of High Performance Microprocessor Circuits, 2001, IEEE Press

[7] A. Dasdan, S. Kolay, M. Yazgan, "Derating for static timing analysis: Theory and practice," ISQED'09, March 2009, pp. 719-727.

[8] A. Narasimhan, R. Sridhar, "Impact of Variability on Clock Skew in H-tree Clock Networks," ISQED '07, March 2007, pp. 458 – 466.

[9] A. Ripp, M. Buhler, J. Koehl, J. Bickford, J. Hibbeler, U. Schlichtmann, R. Sommer, M. Pronath, "DFM/DFY Design for Manufacturability and Yield - influence of process variations in digital, analog and mixed-signal circuit design" DATE'06, Vol. 1, March 2006, pp. 1 – 6.

[10] A.N. Chandorkar, C. Ragunandan, P. Agashe, D. Sharma, H. Iwai, "Impact of Process variations on Leakage Power in CMOS Circuits in Nano Era," ICSICT '06, Oct. 2006, pp. 1248 – 1251.

[11] A.R. Brown, G. Roy, A. Asenov, "Poly-Si-Gate-Related Variability in Decananometer MOSFETs With Conventional Architecture," Tran. on Electron Devices, Vol. 54, Issue 11, 2007, pp. 3056-3063.

[12] B. Cheng, S. Roy, G. Roy, F. Adamu-Lema, A. Asenov, "Impact of intrinsic parameter fluctuations in decanano MOSFETs on yield and functionality of SRAM cells," JSSE, Vol. 49, 2005, pp. 740-746.

[13] B. Nikolic, "Design in the Power-Limited Scaling Regime," Tran. on Elec. Dev., Jan. 2008, Vol. 55, Issue 1, pp. 71-83.

[14] B. Stefano, D. Bertozzi, L. Benini, E. Macii, "Process Variation Tolerant Pipeline Design Through a Placement-Aware Multiple Voltage Island Design Style," DATE'08, March 2008, pp. 967-972.

[15] B.H. Calhoun, Yu Cao, Xin Li, Ken Mai, L.T. Pileggi, R.A. Rutenbar, K.L. Shepard, "Digital Circuit Design Challenges and Opportunities in the Era of Nanoscale CMOS," Proc. of IEEE, vol. 96, no. 2, Feb. 2008, pp. 343-365.

[16] B.J. Cheng, S. Roy, G. Roy, A. Asenov, “Integrating ‘atomistic’, intrinsic parameter fluctuations into compact model circuit analysis,” ESSDERC’03, Sept. 2003, pp. 437 – 440.
[17] Bo Zhai, D. Blaauw, D. Sylvester, K. Flautner, "Theoretical and practical limits of dynamic voltage scaling," DAC'04, 2004, pp. 868 – 873.
[18] C. Chiang, and J. Kawa, Design for Manufacturability and Yield for Nano-Scale CMOS, 2007, Springer Publications
[19] C. Cho, D.D. Kim, J. Kim, J.-O. Plouchart, D. Lim, S. Cho, R. Trzcinski, “Decomposition and Analysis of Process Variability Using Constrained Principal Component Analysis,” IEEE Tran. on Semi. Manu., Vol. 21, Issue 1, Feb. 2008, pp. 55 – 62.
[20] C. Visweswariah “Death, taxes and failing chips,” DAC’2003, pp. 343 – 347.
[21] C. Visweswariah, “Fear, uncertainty and statistics,” ISPD’07, March 2007, pp 169
[22] C. Visweswariah, K. Ravindran, K. Kalafala, S.G. Walker, S. Narayan, D.K. Beece, J. Piaget, N. Venkateswaran, J.G. Hemmett, “First-Order Incremental Block-Based Statistical Timing Analysis,” IEEE Tran. on CAD of Integrated Circuits and Systems, Vol. 25, Issue 10, Oct. 2006, pp. 2170 – 2180.
[23] C.J. Akl, M.A. Bayoumi, “Reducing Delay Uncertainty of On-Chip Interconnects by Combining Inverting and Non-Inverting Repeaters Insertion,” ISQED'07, March 2007, pp. 219 – 224.
[24] D. Boning, and S. Nassif, "Models of Process Variations in Device and Interconnect," Design of High-Performance Microprocessor Circuits, A. Chandrakasam (ed.), 2000.
[25] D. Ernst, N.S. Kim, S. Das, S. Pant, R. Rao, T. Pham, C. Ziesler, D. Blaauw, T. Austin, K. Flautner, T. Mudge, “Razor: a low-power pipeline based on circuit-level timing speculation,” MICRO-36, 2003, pp. 7 – 18.
[26] D. Iparraguirre-Cardenas, J.L. Garcia-Gervacio, and V. Champac, “A design methodology for logic paths tolerant to local intra-die variations,” ISCAS’08,
[27] D. Reid, C. Millar, G. Roy, S. Roy, and A. Asenov, "Statistical enhancement of combined simulations of RDD and LER variability: What can simulation of a 10^5 sample teach us?" IEDM’09, Dec. 7-9, 2009.
[28] D. Sinha, J. Luo, S. Rajagopalan, S. Batterywala, N.V. Shenoy, “Impact of Modern Process Technologies on the Electrical Parameters of Interconnects,” 20th Intl. Conference on VLSI Design’07, Jan. 2007, pp. 875 – 880.
[29] D. Sylvester, “Analysis and mitigation of variability in subthreshold design,” Proc. of ISLPED'05, Aug. 2005, pp. 20 – 25
[30] D. Sylvester, “Design for manufacturability: challenges and opportunities,” ASICON’05, Vol. 1, Oct. 2005, pp. 1169 – 1171.
[31] D. Sylvester, D. Blaauw, and E. Karl, “ElastIC: An Adaptive Self- Healing Architecture for Unpredictable Silicon,” IEEE Design and Test of Computers, Vol. 23, No. 6, Dec. 2006, pp. 484-490.
[32] D.J. Frank, R.H. Dennard, E. Nowak, P.M. Solomon, Y. Taur, Hon-Sum Philip Wong, “Device scaling limits of Si MOSFETs and their application dependencies,” Proc. of the IEEE, Vol. 89, Issue 3, March 2001, pp. 259 – 288.

[33] E. Chang, et al, “Using a statistical metrology framework to identify systematic and random sources of die- and wafer-level ILD thickness variation in CMP process,” IEDM, Dec. 1995, pp. 499-502.
[34] E. Demircan, “Effects of Interconnect Process Variations on Signal Integrity,” Intl. Conference on SOC’06, Sept. 2006, pp. 281 – 284.
[35] E.G. Friedman, “Clock distribution networks in synchronous digital integrated circuits,” Proc. of IEEE, vol. 89, Issue 5, May 2001, pp. 665-692.
[36] F. Fallah and M. Pedram, "Standby and active leakage current control and minimization in CMOS VLSI circuits," IEICE Trans. on Elec., Special Section on Low-Power LSI and Low-Power IP, Vol. E88-C, Issue 4, Apr. 2005, pp. 509-519.
[37] Farzan Fallah, Massoud Pedram, “Standby and Active Leakage Current Control and Minimization in CMOS VLSI Circuits,” IEICE transactions on electronics, 2005, vol. 88, no4, pp. 509-519.
[38] G. Gildenblat, X. Li, W. Wu, H. Wang, A. Jha, A.R. Van Langevelde, G.D.J. Smit, A.J. Scholtena, and D.B.M. Klaassen, "PSP: An advanced surface-potential-based MOSFET model for circuit simulation," IEEE Trans. Electron Devices, vol. 53, pp. 1979–1993, Sep. 2006.
[39] G. Roy, A.R. Brown, F. Adamu-Lema, S. Roy, A. Asenov, “Simulation Study of Individual and Combined Sources of Intrinsic Parameter Fluctuations in Conventional Nano-MOSFETs,” Tran. on Elec. Dev., Vol. 53, Issue 12, Dec. 2006, pp. 3063 – 3070.
[40] G. Sery, S. Borkar, V. De, “Life is CMOS: why chase the life after?,” DAC’02, June 2002, pp. 78 – 83.
[41] H. Fukutome, Y. Momiyama, T. Kubo, E. Yoshida, H. Morioka, M. Tajima, T. Aoyama, "Suppression of Poly-Gate-Induced Fluctuations in Carrier Profiles of Sub-50nm MOSFETs," IEDM '06, 2006, pp.1-4.
[42] H. Mahmoodi, S. Mukhopadhyay, K. Roy, “Estimation of delay variations due to random-dopant fluctuations in nanoscale CMOS circuits,” JSSC, Vol. 40, Issue 9, Sept. 2005, pp. 1787 – 1796.
[43] H. Masuda, S. Okawa, M. Aoki, “Approach for physical design in sub-100 nm era,” ISCAS’05, May 2005, Vol. 6, pp. 5934 – 5937.
[44] H. Tuinhout, “Impact of parametric mismatch and fluctuations on performance and yield of deep-submicron CMOS technologies,” ESSDERC 2002, pp. 95-I01.
[45] H. Watanabe, “Statistics of Grain Boundaries in Polysilicon,” IEEE Tran. on Elec. Dev., vol. 54, Issue 1, Jan 2007, pp. 38-44.
[46] H.-S.P. Wong, Y. Taur, and D. J. Frank, “Discrete Random Dopant Distribution Effects in Nanometer-Scale MOSFETs,” J. of Microelectronics Reliability, Vol. 38, Issue 9, Sept. 1998, pp. 1447-1456.
[47] H.Y. Liu, L. Karklin, Y.T. Wang, Y.C. Pati, "The application of alternating phase shifting masks to 140nm gate patterning: Line width control improvements and design optimization," Proc. of SPIE Symp. on Photomask Technologies, Vol. 3236, 1998, pp. 328-337.
[48] http://en.wikipedia.org/wiki/Clock_skew
[49] http://en.wikipedia.org/wiki/Double_patterning
[50] http://www.eecs.harvard.edu/~ellard/Q-97/HTML/root/node38.html
[51] http://www.itrs.net/Links/2009ITRS/Home2009.htm

[52] I. Nitta, T. Shibuya, and K. Homma, "Statistical static timing analysis technology," FUJITSU Sci. Tech J., vol. 43, pp. 516--523, Oct 2007.
[53] I. Sutherland, B. Sproull, and D. Harris, Logical Effort – Designing fast CMOS Circuits, Morgan Kaufmann Series in Computer Architecture and Design
[54] J. Toney Pan, Ping Li, K. Wijekoon, S. Tsai, F. Redeker, "Copper CMP and Process Control," CMP-MIC'99, Feb 1999.
[55] J. Tschanz, K. Bowman, V. De, "Variation-tolerant circuits: circuit solutions and techniques" DAC'05, June 2005, pp. 762-763.
[56] J.A. Croona, G. Storms, S. Winkelmeierc, I. Pollentier, M. Ercken, S. Decoutere, W. Sansen, H.E. Maes, "Line Edge Roughness: Characterization, Modeling and Impact on Device Behavior," IEDM'02, pp. 307-310.
[57] J.-F. Huang, V.C.Y. Chang, S. Liu, K.Y.Y. Doong, K.J. Chang, "Modeling Sub-90nm On-Chip Variation Using Monte Carlo Method for DFM," ASP-DAC'07, Jan. 2007, pp. 221 – 225.
[58] J.H. Kim, W. Kim, Y.H. Kim, "Effect of Local Random Variation on Gate-Level Delay and Leakage Statistical Analysis," ASQED, July 2009, pp. 255-258.
[59] J.W. Tschanz, J.T. Kao, S.G. Narendra, R. Nair, D.A. Antoniadis, A.P. Chandrakasan, V. De, "Adaptive body bias for reducing impacts of die-to-die and within-die parameter variations on microprocessor frequency and leakage," JSSC, Vol. 37, Issue 11, Nov. 2002, pp.
[60] Juan J. Becerra and Eby G. Friedman, "Analog Design Issues in Digital VLSI Circuits and Systems," Analog Integrated Circuits and Signal Processing, vol. 14, No. 1-2, Sept. 1997, pp 5-8.
[61] K. Choi, R. Soma, M. Pedram, "Dynamic Voltage and Frequency Scaling based on Workload Decomposition," ISLPED'04, 2004, pp. 174-179.
[62] K. Roy, J.P. Kulkarni, Hwang Myeong-Eun, "Process-Tolerant Ultralow Voltage Digital Subthreshold Design," SiRF'08, Jan 2008, pp. 42-45.
[63] K. Takeuchi, et al, "Clock-Skew Test Module for Exploring Reliable Clock-Distribution Under Process and Global Voltage-Temperature Variations," IEEE Tran. on VLSI Systems, Vol. 16, Issue 11, p. 1559–1566, 2008.
[64] K.A. Bowman, J.D. Meindl, "Impact of within-die parameter fluctuations on future maximum clock frequency distributions," CICC'01, May 2001, pp. 229 – 232.
[65] K.A. Bowman, S.G. Duvall, J.D. Meindl, "Impact of die-to-die and within-die parameter fluctuations on the maximum clock frequency distribution for gigascale integration," JSSC, Vol. 37, Issue 2, Feb. 2002, pp. 183 – 190.
[66] K.A. Bowman, T. Xinghai, J.C. Eble, J.D. Menldl, "Impact of extrinsic and intrinsic parameter fluctuations on CMOS circuit performance," JSSC, Vol. 35, Issue 8, Aug. 2000, pp. 1186 – 1193.
[67] K.J. Kuhn, "Reducing Variation in Advanced Logic Technologies: Approaches to Process and Design for Manufacturability of Nanoscale CMOS," IEDM'07, Dec. 2007, pp. 471 – 474.
[68] L. Liebmann, G. Northrop, J. Culp, L. Sigal, A. Barish, and C. Fonseca, "Layout optimization at the pinnacle of optical lithography," Proc. SPIE, Vol. 5042, pp. 1–14, 2003.
[69] L. T. Pileggi, "Achieving Timing Closure for Giga-Scale IC Designs," in Proc. Intl. Symp. On Timing Issues, Mar. 1999, pp. 25-28.

[70] L.T. Pang, and B. Nikolic, “Measurement and analysis of variability in 45nm strained-Si CMOS technology,” *CICC’08*, 2008, pp. 129-132.
[71] L.T. Pang, B. Nikolic, “Impact of Layout on 90nm CMOS Process Parameter Fluctuations,” Sym. on VLSI Circuits’06, 2006, pp. 69 – 70.
[72] M. Abu-rahma and M. Anis, “A Statistical Design-Oriented Delay Variation Model Accounting for Within-Die Variations,” Tran. on CAD of Integrated Circuits and Systems, vol. 27, Nov. 2008, pp. 1983-1995.
[73] M. Anis, M.H. Aburahma, “Leakage current variability in nanometer technologies,” Proc. 5th Intl. Workshop on SoC for Real-Time App., July 2005, pp. 60 – 63.
[74] M. Annavaram, E. Grochowski, P. Reed, “Implications of Device Timing Variability on Full Chip Timing,” HPCA 2007, Feb. 2007, pp. 37 – 45.
[75] M. Hane, T. Ikezawa, T. Ezaki, “Atomistic 3D process/device simulation considering gate line-edge roughness and poly-Si random crystal orientation effects,” IEDM’03, Dec. 2003, pp. 9.5.1 - 9.5.4.
[76] M. Hane, T. Ikezawa, T. Ezaki, “Coupled atomistic 3D process/device simulation considering both line-edge roughness and random-discrete-dopant effects,” SISPAD’03, Sept. 2003, pp. 99 – 102.
[77] M. Mondal, K. Mohanram, Y. Massoud, “Parameter-Variation-Aware Analysis for Noise Robustness,” ISQED'07, March 2007, pp. 655 – 659.
[78] M. Nishida, H. Ohyabu, “Temperature Dependence of MOSFET Characteristics in Weak Inversion,” in IEEE Transactions on Electron Devices, Vol. 24, Issue 10, p. 1245-1248, 1977.
[79] M. Orshansky, L. Milor, C. Pinhong K. Keutzer, C. Hu, “Impact of spatial intrachip gate length variability on the performance of high-speed digital circuits,” Tran. on Comp. Aided Des. of Intg. Circ. & Sys., Vol. 21, Issue 5, May 2002, pp. 544 – 553.
[80] M.H. Abu-Rahma, M. Anis, “Variability in VLSI Circuits: Sources and Design Considerations,” ISCAS’07, May 2007, pp. 3215 – 3218.
[81] Mohsen Raji, B. Ghavami, Hossein Pedram, "Statistical Static Performance Analysis of Asynchronous Circuits Considering Process Variations", ISQED’09, 2009, pp. 291-296.
[82] N. Borivoje, “Measurements and analysis of process variability in 90nm CMOS,” ICSICT '06, Oct. 2006, pp. 505 – 508.
[83] N. Gunther, E. Hamadeh, D. Niemann, I. Pesic, M. Rahman, “Modeling intrinsic fluctuations in decananometer MOS devices due to gate line edge roughness (LER),” ISQED’05, March 2005, pp. 510 – 515.
[84] N. Menezes, “The good, the bad, and the statistical,” ISPD’07, March 2007, pp 168
[85] N. Moubdi, et al, “Product On-Chip Process Compensation for Low Power and Yield Enhancement,” Integrated Circuit and System Design, Power and Timing Modeling, Optimization and Simulation, Vol. 5953/2010, Springer Pub., 2010, pp. 247-255.
[86] N. Verghese, P. Hurat, “DFM reality in sub-nanometer IC design,” ASP-DAC'07, Jan. 2007, pp. 226 – 231.
[87] N.A. Kurd, J.S. Barkatullah, R.O. Dizon, T.D. Fletcher, and P.D. Madland, “A multi-gigahertz Clocking scheme for the pentium 4 microprocessor,” IEEE J. Solid-State Circuits, vol. 36, no. 11, Nov. 2001, pp. 1647–1653.

[88] P. Burggraaf, "Optical lithography to 2000 and beyond," J. Of Solid State Technology, Vol. 42, N°2, Feb. 1999, pp. 31-41.
[89] P. Gupta, A.B. Kahng, "Manufacturing-aware physical design," ICCAD'03, Nov. 2003, pp. 681 – 687.
[90] P. Oldiges, Q. Lint, K. Petrillot, M. Sanchez, M. Ieong, M. Hargrove, "Modeling Line Edge Roughness Effects in sub 100 Nanometer Gate Length Devices," SISPAD 2000, pp. 131-134.
[91] P. Zarkesh-Ha, S. Lakshminarayanan, K. Doniger, W. Loh, P. Wright, "Impact of interconnect pattern density information on a 90 nm technology ASIC design flow," ISQED'03, March 2003 pp. 405 – 409
[92] P. Zarkesh-Ha, S. Lakshminarayanan, K. Doniger, W. Loh, P. Wright, "Impact of interconnect pattern density information on a 90 nm technology ASIC design flow," ISQED'03, March 2003, pp. 405 – 409.
[93] P.R. Groeneveld, "Physical design challenges for billion transistor chips," Conf. on Computer Design: VLSI in Computers and Processors, Sept. 2002, pp. 78 – 83.
[94] P.S. Zuchowski, P.A. Habitz, J.D. Hayes, J.H. Oppold, "Process and environmental variation impacts on ASIC timing," ICCAD'04, Nov. 2004, pp. 336 – 342.
[95] R. Aitken, "Defect or Variation? Characterizing Standard Cell Behavior at 90nm and below," ISQED'07, March 2007, pp. 693-698.
[96] R. Kumar, V. Kursun, "Impact of temperature fluctuations on circuit characteristics in 180nm and 65nm CMOS technologies," ISCAS'06, May 2006, pp. 410-415.
[97] R.C. Aitken, "Defect or Variation? Characterizing Standard Cell Behavior at 90 nm and Below," Tran. on Semiconductor Manufacturing, Feb. 2008 Vol. 21, Issue 1, pp. 46-54.
[98] R.Difrenza, J.C Vildeuil, P. Llinares, G. Ghibaudo, "Impact of grain number fluctuations in the MOS transistor gate on matching performance," Intl. Conf. on Microelectronic Test Structures, March 2003, pp. 244-249.
[99] S. Bhunia, S. Mukhopadhyay, K. Roy, "Process Variations and Process-Tolerant Design," VLSID'07, Jan. 2010, pp. 699-704.
[100] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, V. De, "Parameter variations and impact on circuits and microarchitecture," DAC'03, June 2003, pp. 338 – 342.
[101] S. Das, S. Pant, D. Roberts, S. Lee, D. Blaauw, T. Austin, K. Flautner, T. Mudge, "A Self-Tuning DVS Processor Using Delay-Error Detection and Correction," JSSC, vol. 41, Issue 4, April 2006, pp. 792-804.
[102] S. Deleonibus (Editor), Electronic Device Architectures for the nano-CMOS Era-From Ultimate CMOS Scaling to Beyond CMOS Devices
[103] S. Ekbote, et al, "45nm low-power CMOS SoC technology with aggressive reduction of random variation for SRAM and analog transistors," Symp. on VLSI Tech., June 2008, pp. 160-161.
[104] S. Ghosh, S. Bhunia, K. Roy, "A New Paradigm for Low-power, Variation-Tolerant Circuit Synthesis Using Critical Path Isolation," ICCAD'06, Nov. 2006, pp. 619-624.
[105] S. Mukhopadhyay, Kim Keejong H. Mahmoodi, K. Roy, "Design of a Process Variation Tolerant Self-Repairing SRAM for Yield Enhancement in Nanoscaled CMOS," JSSC, Vol; 42, Issue 6, June 2007, pp. 1370-1382.

[106] S. Nassif, K. Bernstein, D.J. Frank, A. Gattiker, W. Haensch, B.L. Ji, E. Nowak, D. Pearson, N.J. Rohrer, "High Performance CMOS Variability in the 65nm Regime and Beyond," IBM Journal of Research and Development, Vol. 50, Issue 4/5, July 2006, Advanced silicon technology, pp. 433 – 449.

[107] S. Saxena, C. Hess, H. Karbasi, A. Rossoni, S. Tonello, P. McNamara, S. Lucherini, S. Minehane, C. Dolainsky, M. Quarantelli, "Variation in Transistor Performance and Leakage in Nanometer-Scale Technologies," Tran. on Elec. Dev., Vol. 55, Issue 1, Jan. 2008, pp. 131 – 144.

[108] S. Sayil, M. Rudrapati, "Accurate Prediction of Crosstalk for RC Interconnects," Turk J Elec Eng & Comp Sci, Vol.17, No.1, 2009, pp. 55-67.

[109] S. Sundareswaran, L. Nechanicka, R. Panda, S. Gavrilov, R. Solovyev, J.A. Abraham, "A Timing Methodology Considering Within-Die Clock Skew Variations," SOC Conference, sept. 2008, pp. 351 – 356

[110] S. Zanella, A. Nardi, A. Neviani, M. Quarantelli, S. Saxena, C. Guardiani, "Analysis of the impact of process variations on clock skew," Tran. on Semiconductor Manufacturing, Vol. 13, Issue 4, Nov 2000, pp. 401 – 407.

[111] S.K. Springer, et al, "Modeling of Variation in Submicrometer CMOS ULSI Technologies," Tran. on Elec. Dev., Vol. 53, Issue 6, Sept. 2006, pp. 2168-2178.

[112] S.R. Stg, J. Srivatsava, Narahari Tondamuthuru R, "Process Variability Analysis In DSM Through Statistical Simulations And Its Implications To Design Methodologies," ISQED'07, March 2008, pp. 325-329.

[113] Semiconductor Reliability Handbook, Renesas technology

[114] Seong-Dong Kim, H. Wada, J.C.S. Woo, "TCAD-based statistical analysis and modeling of gate line-edge roughness effect on nanoscale MOS transistor performance and scaling," IEEE Tran. on Semiconductor Manufacturing, Vol. 17, Issue 2, May 2004, pp. 192 – 200.

[115] T. Chawla, A. Amara, A. Vladimirescu, "Yield, Power and Performance Optimization for Low Power Clock Network under Parametric Variations in Nanometer Scale Design," MWSCAS'06, Vol. 2, Aug. 2006, pp. 231 – 235.

[116] T. Chawla, S. Marchal, A. Amara, A. Vladimirescu, "Impact of Intra-Die Random Variations on Clock Tree" NORCHIP'09, Nov. 2009, pp. 1 – 4.

[117] T. Chawla, S. Marchal, A. Amara, A. Vladimirescu, "Local Mismatch in 45nm Digital Clock Networks," ISIC'09, Dec. 2009, pp. 466 – 469.

[118] T. Chawla, S. Marchal, A. Amara, A. Vladimirescu, "Pulse Width Degradation in 45nm ASIC Design due to Global and Environmental Variations," ICM'09, Dec. 2009, pp. 302-305.

[119] T. Chawla, S. Marchal, A. Amara, A. Vladimirescu, "Pulse width variation tolerant clock tree using unbalanced cells for low power design," MWSCAS'09, Aug. 2009, pp. 443-446.

[120] T. Sakurai, A.R. Newton, "Delay analysis of series-connected MOSFET circuits," JSSC, Vol. 26, Issue 2, Feb. 1991, pp. 122-131.

[121] T. Yamaguchi, K. Yamazaki, M. Nagase, H. Namatsu, "Line-edge roughness: characterization and material origin," Jpn. J. Appl. Phys. Part 1, Vol. 42, Issue 6B, 2003, pp. 3755-3762.

[122] V. Mehrotra, S. Nassif, D. Boning, and J. Chung, ‘‘Modeling the Effects of Manufacturing Variation on High-Speed Microprocessor Interconnect Performance,’’ IEDM Tech. Digest, pp. 767–770 (1998).
[123] X. Qi, A, Gyure, Y. Luo S.C. Lo, M. Shahram, K. Singhal, “Simulation of interconnect inductive impact in the presence of process variations in 90 nm and beyond,” IEEE Elec. Dev. Letters. Vol. 27, Issue 8, Aug. 2006, pp. 696 – 698.
[124] Y. Taur, D.A. Buchanan, Wei Chen, D.J. Frank, K.E. Ismail, Shih-Hsien Lo, G.A. Sai-Halasz, R.G. Viswanathan, H.-J.C. Wann, S.J. Wind, S.J., Hon-Sum Wong, “CMOS scaling into the nanometer regime,” Proceedings of the IEEE, Volume 85, Issue: 4 April 1997 Page(s):486 – 504.
[125] Y.-F. Tsai, N. Vijaykrishnan, Y. Xie, M.J. Irwin, “Influence of leakage reduction techniques on delay/leakage uncertainty,” 18th Intl. Conf. on VLSI Design, Jan. 2005, pp. 374 – 379.
[126] Yang Fu-Liang, Hwang Jiunn-Ren, Li Yiming, “Electrical Characteristic Fluctuations in Sub-45nm CMOS Devices,” CICC’06, Sept. 2006, pp. 691 – 694.
[127] Yu Cao, L.T. Clark, “Mapping statistical process variations toward circuit performance variability: an analytical modeling approach,” DAC’05, June 2005, pp. 658 – 663.
[128] http://www.sp.phy.cam.ac.uk/~SiGe/Scaling.html
[129] http://www.chipestimate.com/techtalk.php?d=2008-12-30

9 Publications

[1] T. Chawla, S. Marchal, A. Amara, A. Vladimirescu, “Impact of Intra-Die Random Variations on Clock Tree” NORCHIP’09, Nov. 2009, pp. 1 – 4.
[2] T. Chawla, S. Marchal, A. Amara, A. Vladimirescu, “Local Mismatch in 45nm Digital Clock Networks,” ISIC’09, Dec. 2009, pp. 466 – 469.
[3] T. Chawla, S. Marchal, A. Amara, A. Vladimirescu, “Pulse Width Degradation in 45nm ASIC Design due to Global and Environmental Variations,” ICM’09, Dec. 2009, pp. 302-305.
[4] T. Chawla, S. Marchal, A. Amara, A. Vladimirescu, “Pulse width variation tolerant clock tree using unbalanced cells for low power design,” MWSCAS’09, Aug. 2009, pp. 443-446.

Thank You