Measurement and Analysis of Variability in 45 nm Strained-Si CMOS Technology

Liang-Teck Pang, *Member, IEEE*, Kun Qian, *Student Member, IEEE*, Costas J. Spanos, *Fellow, IEEE*, and Borivoje Nikolić, *Senior Member, IEEE*

***Abstract*—A test-chip in a low-power 45 nm technology, featuring uniaxial strained-Si, has been built to study variability in CMOS circuits. Systematic layout-induced variation, die-to-die (D2D), wafer-to-wafer (W2W) and within-die (WID) variability has been measured over multiple wafers, analyzed and attributed to likely causes in the manufacturing process. Delay is characterized using an array of ring oscillators and transistor leakage current is measured with an on-chip ADC. The key results link systematic layout-dependent and die-to-die variability as being caused by gate patterning and material strain. In comparison to a previous 90 nm experiment, gate proximity now contributes less to frequency variability, causing a 2% change in overall performance, while strain has increased its contribution to about 5% of the overall performance.**

***Index Terms*—45 nm, CMOS, layout, leakage, ring oscillators, strain, variability.**

I. Introduction

SCALING of CMOS into deep submicron has increased the impact of process variability on circuits, to the point where it has emerged as a major technological barrier to further scaling [1]. Shrinking of critical dimensions has been possible due to advances in lithography and gate patterning. Resolution enhancement techniques such as optical proximity correction (OPC), phase shift masks, double patterning and immersion lithography have enabled scaling to the present 45 nm node and beyond. In addition, performance enhancing techniques such as the use of strained-Si have been introduced. These techniques have, however, contributed to an already complex manufacturing process and compounded the sources of process variability. To mitigate various systematic layout-dependent effects on devices, restricted design rules and more complex OPC have been introduced.

In this work, test structures have been designed to characterize variability in an early 45 nm process [2] and results were compared with a previous 90 nm process [3], [4]. Systematic, random, WID and D2D components of variability were studied by measuring the individual frequencies of an array of ring oscillators (ROs) and the sub-threshold leakage currents of an array of PMOS and NMOS transistors. Layout and spatial dependencies of variability were investigated. The nature of systematic variations in the presence of OPC and restricted layout design rules were evaluated.

Section II gives an overview of the characteristics and sources of variations in the manufacturing process and describes how variability impacts circuit performance. Section III describes special features of the low-power 45 nm process which influence transistor performance. The circuits in the test chip are described in Section IV, and the measurement and analysis results are presented in Section V. These results will be compared to a previous study, conducted in a 90 nm technology. Finally, Section VI concludes with a summary of the most significant results.

II. Variability in CMOS Technology

CMOS process parameter variability is generally classified into three categories: known systematic, known random and unknown [5]. Systematic variations are deterministic shifts in space and time of process parameters, whereas random variations change the performance of any individual instance in the design in an unpredictable manner. Systematic variations are, in general, spatially correlated. In practice, although many of the systematic variations have a deterministic source, they are not known at the design time, or are too complex to model, and are thus treated as random. As a result, many of the sources of variability are not modeled in the design kits and have to be treated as random in the design process. The resulting 'random' variation component, depending on the way systematic variability is modeled, will often appear to have a varying degree of spatial correlation. This is consistent with the findings of [24] and [26].

Tolerances in the manufacturing process are generally classified as within-die (WID), die-to-die (D2D), wafer-to-wafer (W2W) and lot-to-lot (L2L) [6]. Variations reflect both the spatial as well as the temporal characteristics of the process and cause different dies and wafers to have different properties. The performance of the manufacturing equipment, expressed through the dose, speed, vibration, focus, or temperature, varies within one die and from die to die. Those parameters that vary rapidly over distances smaller than the dimension of a die result in WID variations whereas variations that change gradually over the wafer will cause D2D variations. Similarly, even more parameters vary from wafer to wafer (W2W variations) and between different manufacturing runs (L2L variations). In a

Manuscript received December 16, 2008; revised March 01, 2009. Current version published July 22, 2009. This work was supported by the National Science Foundation Infrastructure Grant 0403427, wafer fabrication donation of STMicroelectronics, and the support of the Center for Circuit & System Solutions (C2S2) Focus Center, one of five research centers funded under the Focus Center Research Program, a Semiconductor Research Corporation program.

L.-T. Pang was with Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720-1770 USA, and is now with the IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 USA (e-mail: ltpang@us.ibm.com).

K. Qian, C. J. Spanos, and B. Nikolić are with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720-1770 USA (e-mail: bora@eecs.berkeley.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSSC.2009.2022217

typical design methodology, D2D variations can be due to intra-field variations or field to field variations over the same wafer and also encompass other inter-die variations such as W2W and L2L variations. Designs are made to satisfy the worst case corners which consist of the total WID and D2D variations.

A. Sources of Variability

Many sources of systematic variability can be attributed to the different steps of the manufacturing process. The photolithography and etching processes contribute significantly to variations in nominal lengths and widths due to the complexity required to fabricate lines that are much narrower than the wavelength of light used to print them [7]. Notable contributors in this area include temperature non-uniformities in the critical post-exposure bake (PEB) and etch steps. Variation in film thicknesses (e.g., oxide thickness, gate stacks, wire and dielectric layer height) is due to the deposition and growth process, as well as the chemical-mechanical planarization (CMP) step. Additional electrical properties of CMOS devices are affected by variations in the dosage of implants, as well as the temperature of annealing steps. In recent technologies, overlay error, mask error, shift in wafer scan speed, rapid thermal anneal and the dependence of stress on layout have become notable sources of systematic variations.

Random device parameter fluctuations stem mainly from line-edge roughness (LER) [8], Si/SiO_2 and polysilicon (poly-Si) interface roughness [9] and random dopant fluctuations (RDF) [10].

B. Impact of Variability

Variability affects IC yield. Yield is defined as the probability that a chip is both functional and meets the parametric constraints, such as timing and power. A circuit with more design margin will have a higher yield. The challenge is in finding the smallest margin necessary for the required yield so that performance is not overly constrained. In order to model the statistics of the circuit performance accurately, both the amount of parameter variation and any spatial characteristics of the parameter variation between different gates have to be known.

Characterizing the amount of variation involves making measurements of many devices and obtaining the statistical information, often expressed through standard deviation. This has traditionally been done in order to obtain "design corner" information. Presently, foundries measure a number of test-structures on the wafers and fit the I-V data into a model such as the BSIM SPICE model. Variability is captured in the statistics of the model parameters. This information is then used to generate the process corners whereby certain parameters are varied by a number of standard deviations from their nominal values.

In a typical VLSI design process, satisfying design corners is deemed necessary and assumed sufficient in order to validate a design. This approach typically regards all variations as D2D, with all devices on a chip having identical process parameters. Deep submicron scaling has compounded the impact of variability and increased the amount of design margin to cope with worst-case scenarios. Characterizing variability in a more detailed way would allow designers to reduce systematic variations and, with the help of statistical timing tools [11], use the right amount of margins to obtain an optimal design that maximizes performance, power and yield.

TABLE I
SUMMARY OF THE 45 NM PROCESS [12]–[14]

PROCESS FEATURE	45NM PROCESS	EFFECT
Si substrate	<100>-oriented channel	Higher PMOS mobility
Shallow trench isolation (STI)	Sub-atmospheric deposited oxide	Lower STI stress
Contact etch stop layer (CESL)	Nitride layer create high tensile strain	Higher NMOS mobility
Immersion	NA > 1	Improved resolution
Backend dielectric	Low k ~2.5	Low RC delay

C. Variability Characterization Goals

Test structures for characterizing variability, such as those devised in this paper, have the goals of classifying unknown sources of variation into systematic and random, performing model to hardware correlation and tracking of process performance in time. To achieve these goals, test structures have to be implemented early in the technology development cycle, and thus should have a short design cycle. Ring oscillators (ROs) are often used for this purpose, as they can be easily designed and have a fast and simple, digital readout through a frequency counter. ROs are also very compact test structures, resulting in good spatial resolution of measurements. An array of RO can produce large amounts of data in a short time over a small area. However, isolation of individual device or process parameters through variations in RO frequencies is not trivial.

In contrast, voltage-current sweeps can be used to characterize device performance, resistances or capacitances in the process. These measurements are slower and have limited spatial coverage and density, as they require digital-to-analog conversion of the sweeping parameters and on- or off-chip analog-to-digital conversion of the target voltage or current values.

III. FEATURES OF THE LOW-POWER 45 NM STRAINED-SI PROCESS

Table I summarizes several techniques used in the low-power 45 nm process [12]–[14]. In the traditional $\langle 110 \rangle$-channel orientation, both NMOS and PMOS transistors are affected by strain. Compressive parallel strain increases PMOS mobility and decreases NMOS mobility whereas tensile parallel strain achieves the opposite effect; decreases PMOS mobility and increases NMOS mobility. In this 45 nm process, transistor channels are oriented in the $\langle 100 \rangle$ direction, which increases PMOS transistor mobility and makes it insensitive to strain [15]. NMOS mobility is still affected by strain as described above.

Shallow trench isolation (STI) is used to electrically isolate adjacent transistors. Traditional methods use SiO_2 in the STI trenches, which create compressive strain on the channel substrate that varies with distance from the edge of the STI/diffusion interface to the channel region [16]. In this 45 nm process, the use of sub-atmospheric chemical vapor deposition oxide

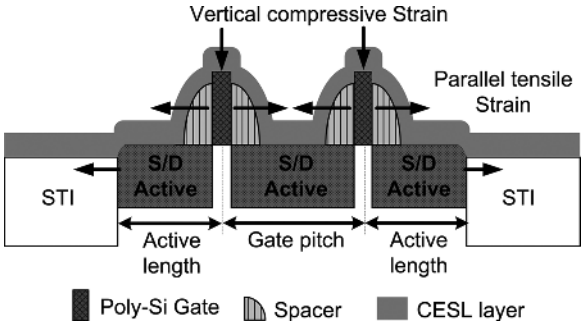

Fig. 1. Illustration of the strain caused by STI and the CESL layer. STI creates a weak parallel tensile strain that increases the mobility of the NMOS transistors while CESL strain creates a strong vertical compressive strain and strong parallel tensile strain, both of which increase NMOS mobility. STI strain depends on the length from the channel to the STI edge (active length) whereas CESL strain depends on the gate pitch, spacer width and the length of the active region.

(SACVD) for trench isolation further reduces stress effects [13]. Instead of a strong compressive strain, these trenches now exert a weak tensile strain on the transistors.

Nitride films are deposited on Si to act as the contact etch stop layer (CESL). In this 45 nm process, strong uniaxial tensile strain is created by the nitride layer in order to increase NMOS mobility. The amount of strain increases with the thickness of the nitride layer that can fill between the gates and the amount of contact of the nitride film with the source/drain region [17]. The latter depends on poly-Si pitches and spacer sizes. A smaller contact area between the CESL and the source/drain region results in less strain on the transistors. The amount of strain starts to drop quickly when the sidewalls of the CESL start touching each other. Fig. 1 illustrates the strain caused by the STI and CESL in this process. The impact of layout on strain can be different if the strain is created by other processes. Strain induced by the STI and the CESL nitride film are layout dependent and are investigated in the test chip.

Resolution in the 45 nm process is enhanced with immersion lithography and low-k dielectric is used for the copper interconnects. Our prior work in a dry lithography 90 nm process revealed significant impact of poly-Si gate pitch on transistor performance due to the lithography process [4]. In this 45 nm process, gate proximity effects are also investigated with a more detailed set of test-structures.

IV. Test-Chip

A 45 nm test-chip has been designed that contains an array of structures with different layout styles in order to characterize variability and evaluate the impact of layout-induced variations.

Seventeen layout styles were created to study the effects of layout. These are shown in Figs. 2 and 3 and described in Table II. Layouts P1, P2, P3 and P4 vary the spacing of the poly-Si nearest to the transistor's gate. SP1, SP2, SP3 vary the distance of the poly-Si that is the second nearest neighbor to the gate. The diffusion width in the layouts P1-4 differs from those of SP1-3. These layout styles investigate the effect of gate proximity on transistor performance. In the 90 nm test-chip, it was shown that line width variation caused by the poly-Si pitch

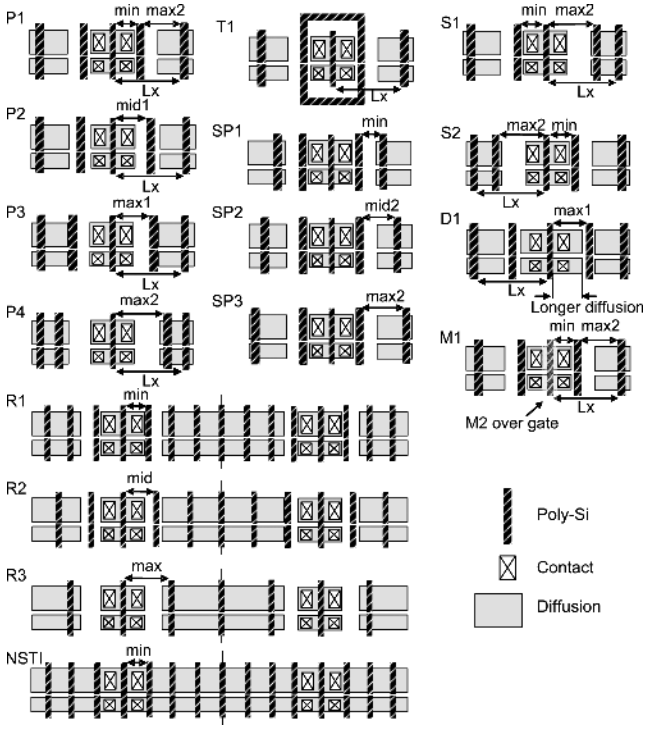

Fig. 2. Layout configurations.

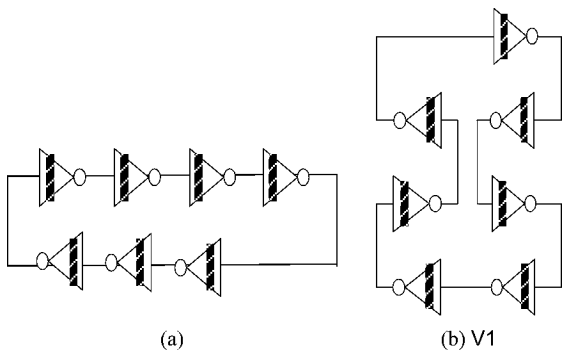

Fig. 3. (a) Horizontally versus (b) vertically arranged inverters. The dashed bar in the middle of an inverter represents the vertical poly-Si orientation.

TABLE II
Characteristics of the Layout Configurations

Layouts	Targeted Effect
P1, P2, P3, P4	Primary proximity
SP1, SP2, SP3	Secondary proximity
S1, S2	Symmetry
D1	Larger S/D area
M1	Metal coverage over gate
T1	Poly-Si at extremity of gate
R1, R2, R3	Regular pitch, different density
NSTI	No STI

density has a significant impact due to lithography and etch effects.

S1 and S2 are laterally inverted images of each other. It has been shown that odd optical aberrations (such as coma) can

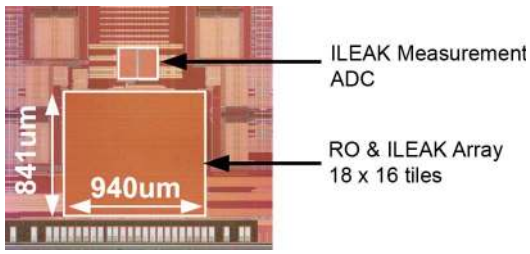

Fig. 4. 45 nm test-chip die photo.

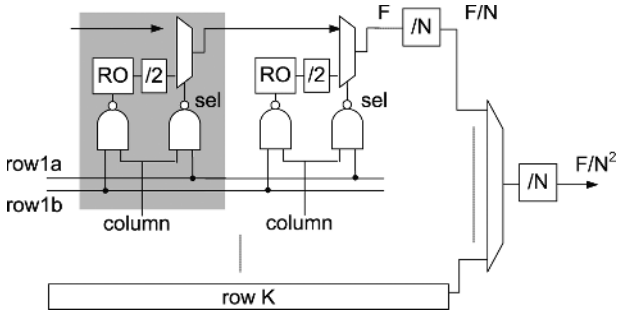

Fig. 5. RO frequency measurement.

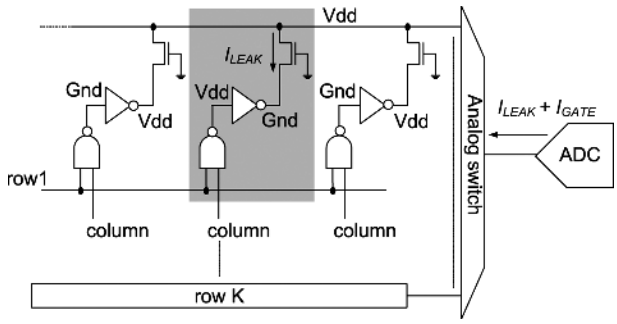

Fig. 6. NMOS sub-threshold leakage current measurement.

cause these structures to print differently, giving rise to different transistor performance [18]. D1 has a longer source/drain (S/D) diffusion area than P3, which has been observed to cause different strain in a transistor [16]. M1 has metal-2 coverage over its gate which has been shown to cause different annealing temperatures [19] during the process of rapid thermal annealing. The metal coverage could cause the gate to experience a lower annealing temperature, resulting in a higher V_T shift due to incomplete passivation of interface states. Metal-2 coverage has been used in this experiment, instead of metal-1 to obtain the smallest gate pitch possible. T1 has neighboring poly at the ends of its gate. R1, R2 and R3 have regular poly pitches that vary from minimum to maximum. The benefits of regular poly-Si pitch have been observed in lithography and recommended for improved manufacturability [20]. These structures permit characterization of the impact of regular pitches and pitch spacing on variability.

NSTI layout is the same as R1 except that there is no STI and isolation is achieved by turning off adjacent transistors [21]. STI causes strain in the substrate, which changes transistor mobility. It also creates unevenness on the surface of the substrate, resulting in systematic changes in transistor properties. Finally, in layout V1 the inverters of a RO are placed in the vertical direction instead of the horizontal direction as illustrated in Fig. 3. In the 90 nm test-chip, variations were dependent on the horizontal and vertical directions due to the step and scan photo-lithography which generates differences in the slit and scan directions [22]. Even though 90° gate rotations are not allowed by the design rules in this 45 nm process, certain properties of the two orthogonal directions can be investigated with this structure.

The die photo of the 45 nm test-chip is shown in Fig. 4. The array contains 18×16 tiles, each tile contains 17 ROs and 17 NMOS and PMOS transistors with $V_{GS} = 0$ in each of the 17 layouts. The measurement circuits in this chip have been adapted from [3] and [4]. In the RO array, row and column bits from a scan-chain enable the RO of interest and select the multiplexer to output its RO frequency which is multiplexed out to a row divider and further divided down before being output to a pad. A local divide-by-2 circuit within each RO allows for the use of small number of stages (13) by reducing the frequency of the signal that is multiplexed out, as shown in Fig. 5.

Both PMOS (I_{LEAKP}) and NMOS (I_{LEAKN}) leakage currents have been measured. Fig. 6 shows how the leakage current of a NMOS device in an array is measured. The gates of the NMOS devices are connected to ground. Row and column bits select the device to be measured by supplying either Vdd or Gnd to the source of the transistor using a large inverter. The selected NMOS will have Gnd applied to its source and the other NMOS transistors will have Vdd applied to their source. This will enable a sub-threshold current to flow from the drain terminal to the source of the selected NMOS. An analog multiplexer, constructed with thick oxide transistors, is used to select the row to be measured in order to reduce the magnitude of parasitic currents. Parasitic currents from the drain of the NMOS that are not selected can be significant as there can be ~50 devices in a row. These will be measured and subtracted from the final measurement. PMOS leakage currents are measured in a similar manner.

On-chip current measurement of ~ 1–10 nA with a precision better than 0.1 nA requires the use of an ADC. Precision analog blocks are hard to design and would generally not be ready in time for an early-phase technology characterization. However, in this process, the thick-oxide, 1.8 V transistors are similar to the thick oxide 2.5 V transistors used in the previous-generation, 90 nm process. Therefore, we were able to port the single-slope ADC circuit and current measurement procedures described in [3], [4] for the use in this chip. Two leakage current measurement circuits were designed, for NMOS and for PMOS, respectively. These two differ in the sizing of the ADC amplifier which is tuned to maximize DC gain at its respective range of input voltages. I_{LEAKP} measurement using a single slope ADC is illustrated in Fig. 7(a). Switches P1, P2 and P1b select the currents to be integrated. During integration, the output $(Vout)$ of the op-amp will ramp down. As it passes the threshold voltages of on-chip comparators, start and stop signals are generated. By timing the interval between these two signals, the currents are measured. Fig. 7(b) shows the folded cascode amplifier gain versus frequency plot for the PMOS leakage measurement,

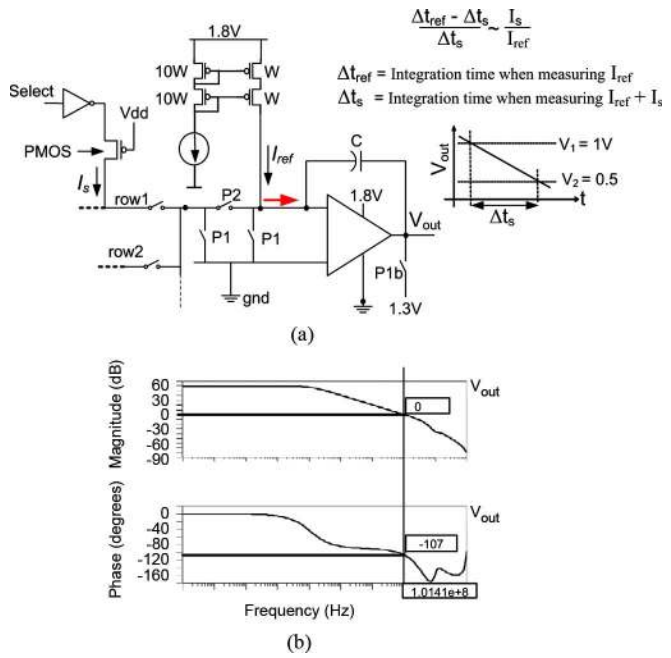

Fig. 7. (a) Single slope ADC for PMOS leakage current measurement. (b) Bode plots showing DC gain of 60 dB for the amplifier used in (a).

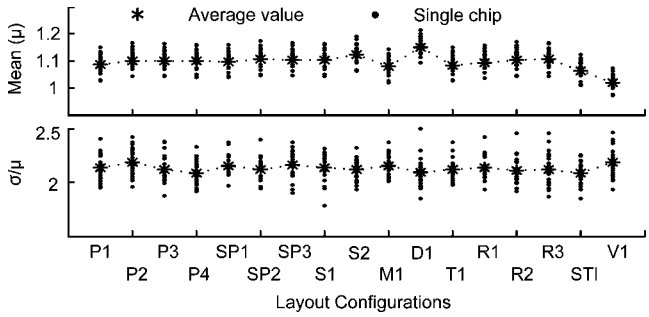

Fig. 8. 45 nm WID statistics of RO frequency for 22 dies in the first batch of chips. The frequency is normalized to the SS corner frequency.

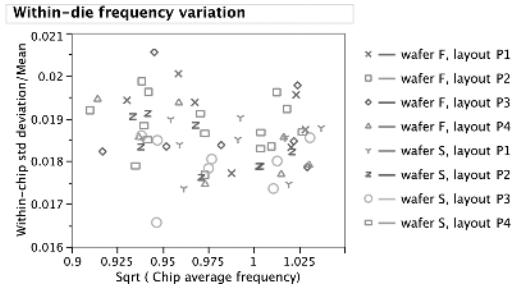

Fig. 9. Within-chip (Standard Deviation/Mean) versus Sqrt (Mean freq) for the dies in the second batch.

verifying that a gain of 60 dB has been achieved without compromising stability. The large integration capacitors are implemented using metal fringe structures.

V. Measurement Results

Two batches of dies originating from four wafers were evaluated. The first batch consists of dies from two wafers with no information on their position on the wafers. These dies were used for analysis of the impact of layout and general view of D2D and WID variations in this process. The second batch of dies comes from two wafers with the die position on the wafer known, allowing for in-depth study of across-wafer variability. These two wafers were selected to have a nominal 4 nm difference in effective gatelength (Leff), representing different process corners. The slower of the two wafers will be referred to as "Wafer S", and the faster, will be referred to as "Wafer F". For extraction of the systematic across-wafer variation component, a set of chips was carefully sampled to cover the largest area across the wafer, while assigning adequate samples for all distances to the center of the wafer. To speed up the measurement, only $8 \sim 16$ devices per layout are measured on each of these chips.

The measurements reveal several important trends. Systematic layout-induced variations, in particular those related to poly-Si density, are significantly reduced in this early 45 nm process compared to the early 90 nm process. Variations due to strain and other factors are now dominant. Finally, random WID variation has increased proportionally to transistor area reduction, while systematic D2D variation has decreased.

A. W2W, D2D, WID Variation Components

Fig. 8 plots the WID mean and σ/μ of the RO frequency for each of the 22 dies in the first batch of chips as a function of the layout configuration. The mean frequency for each layout has been normalized to the respective SS corner in order to compare differences in layouts that have not been captured by the layout extraction. WID variation is approximately 2.2% which is more than twice that in 90 nm ($\sigma/\mu \sim 1.1\%$) for the RO of the same length. This is consistent with the transistor area reduction by a factor of 4 between the two processes.

Fig. 9 plots the RO frequency standard deviation normalized to mean frequency of each chip against the square root of mean frequency. Direct gate CD measurement shows about 10% variation in $\mathrm{L_{EFF}}$ across the wafer leading to varying device size $\mathrm{W} * \mathrm{L}$. According to Pelgrom's model [27], σ_{VT} will show inverse proportionality to square root of device size $\mathrm{W} * \mathrm{L}$. However if we use $(\sigma/\mu)_{\mathrm{freq}}$ to approximate σ_{VT} (as $\mathrm{V_T}$ cannot be measured directly) and use μ_{freq} to replace 1/L, no significant correlation between $(\sigma/\mu)_{\mathrm{freq}}$ and $\mathrm{sqrt}(\mu_{\mathrm{freq}})$ is observed, indicating that the local randomness cannot be entirely from random dopant fluctuation (RDF). Other effects as line edge roughness (LER) and gate oxide interface roughness must have also played a role.

Systematic variations in the mask are investigated by normalizing the data of each chip by the mean and standard deviation of that chip to zero mean and unity standard deviation, followed by averaging the normalized data from the first batch of 22 chips to remove random variations. The normalization process retains the relative frequency variation of each RO with respect to its position within the chip. No significant systematic variation was found, partially due to the fact that random WID variation has increased [2]. Finally, there is no significant spatial structure across the die in the measurements.

D2D and WID $3\sigma/\mu$ of RO frequency from the second batch of dies are listed in Table III. The mean RO frequencies of the

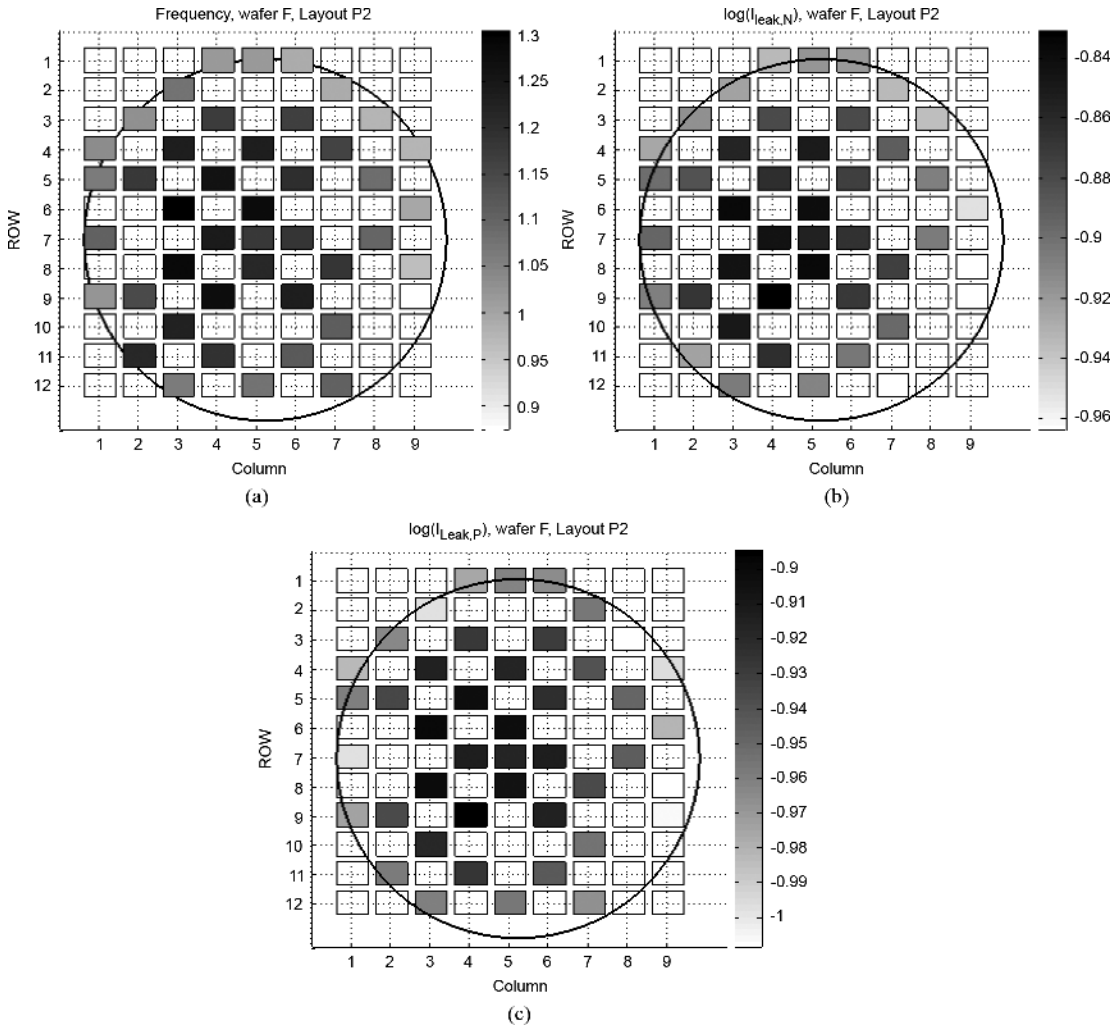

Fig. 10. Wafer maps of (a) RO frequency, (b) NMOS leakage current, (c) PMOS leakage current in the second batch of dies. Non-shaded dies are not measured.

TABLE III
D2D AND WID RO FREQUENCY VARIATION OF THE SECOND BATCH OF CHIPS FOR LAYOUT P2

Process	Wafer S	Wafer F
Number of chips	32	44
D2D ($3\sigma/\mu$)	18%	28%
WID ($3\sigma/\mu$)	5.4%	6%

two wafers differ by $\sim$12%. The significantly larger D2D variation in wafer F is due to a larger systematic component of across-wafer variation in the faster wafer.

Device parameter and circuit performance vary with their position on the chip and wafer. Due to the multi-level nature of the manufacturing variations, it is natural to use a hierarchical structure to model process variability. This model is verified by observations from the 90 nm test chip measurement data [24]. This methodology has been applied to the second batch of chips from two wafers. The frequency and leakage current map across the wafer is shown in Fig. 10.

All three types of across-wafer variations can be approximated by a dome-shaped deterministic function, whose shapes strongly correlate. This can be explained by a systematic gate length variation across the wafer, and the corresponding threshold voltage roll-off. A plausible cause is the PEB step, where the wafer may be subject to non-uniform heating, at least during the heating transient step [23]. During the plasma etching process, a typically observed temperature non-uniformity pattern may also cause over etch (and therefore faster devices) near the center of the wafer [25].

As shown in Fig. 11, the slower wafer (S) and the faster wafer (F) share a similar across-wafer function in the frequency measurement results, while a significant difference in magnitude and curvature. This supports our assumption that gate length is one of the underlying mechanisms of across-wafer variation, because ring oscillator frequency becomes more sensitive to effective gate length change as transistors get shorter. However gate length alone is not enough to explain for this big difference. Strain effect and doping variation may also contribute to the across-wafer variation as shown in the measurements. It should be noted, however, that this variation is reduced in a mature process.

B. Layout Effects

Measurement results show systematic variations for different layouts that are not captured by the layout extraction tool. Accurate analysis of the causes of variability is difficult as the

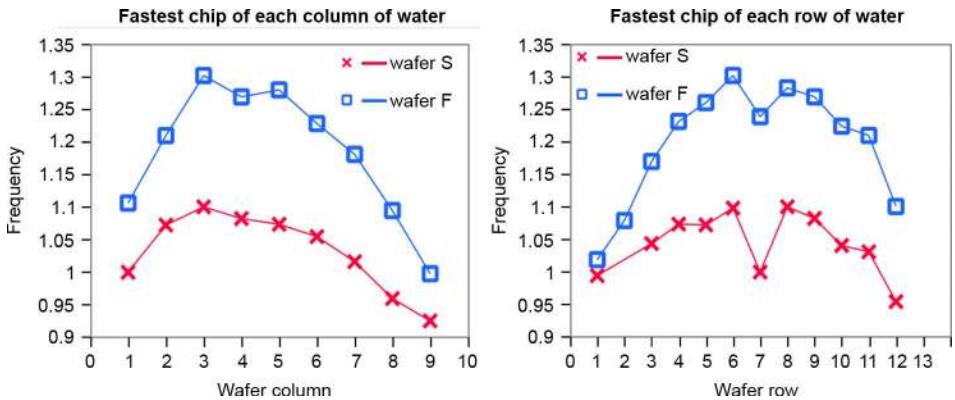

Fig. 11. Normalized across-wafer frequency variation in the second batch of dies: fastest chips for layout P2. Measurement is normalized to the same constant value for both wafers.

variations are relatively small. Nevertheless, we shall propose plausible explanations for these observations.

In the following analysis using the first batch of dies, RO frequencies are normalized to the SS corners in order to remove the differences in parasitics that are captured by the layout extraction. Parasitic R and C have been extracted from the layout using the Mentor Calibre extraction tool for both devices and interconnects, and simulations were performed with device corner parameters. BSIM stress parameters (SA, SB) were extracted but do not have an impact in the model that was provided by the foundry. Leakage currents are not aligned with the corners as they are independent of parasitics. Distributions of normalized frequency and normalized leakage in the log domain are plotted in Fig. 12 through Fig. 16. Each shaded histogram represents the distribution for a die and the overall distribution is plotted as a continuous curve. Measured results of 22 dies from two wafers in the first batch of dies are plotted. Only the histograms of the fastest and slowest die are shown. Vertical lines labeled SS and TT represent simulation results from the extracted layout for SS and TT corners.

Measured results show that the impact of layout on performance is small. After compensating for the parasitics, only 2% shift in frequency ($\Delta F = 2\%$) exists due to proximity of poly-Si. The most significant effects were deemed to be stress related due to a larger S/D area ($\Delta F = 5\%$) and the removal of STI ($\Delta F = 3\%$).

1) Proximity Effects: Fig. 12 shows the distributions for four layouts with different poly-Si gate pitches. Maximum systematic shift in frequency is $\sim 2\%$. Leakage currents also experience small systematic shifts. This effect could be due to small gate-length (L) variation due to the corrective influence of OPC, and layout dependent variation of the strain caused by the CESL. PMOS transistors have a sharper V_T roll-off and hence PMOS leakage currents are more sensitive to L variation. An isolated gate will generally experience more strain due to CESL. In the plot of PMOS leakage, poly density induced L variation is observed. The P2 layout on the second row likely has a shorter L than the others resulting in increased PMOS leakage. The effect of L variation on NMOS leakage is weaker and is completely compensated by the CESL stress. This can be due to a smaller change of L and a smaller dependence of V_T on L variation as it is possible that the NMOS device lies on a flatter part of the

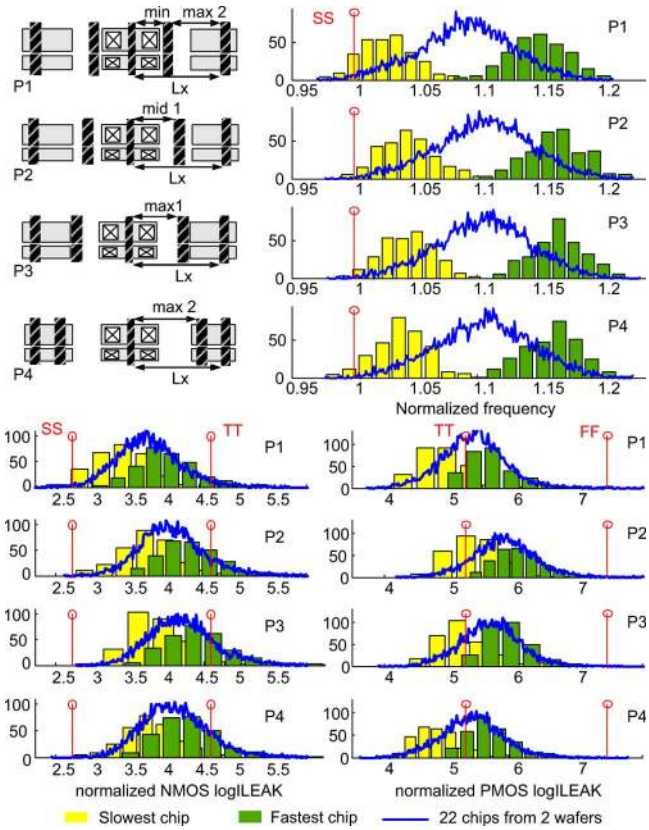

Fig. 12. Effect of nearest neighbor poly-Si pitch on RO frequency and transistor leakage currents. Plots of RO frequency distribution, NMOS $\log(I_{I,EAK})$ distribution and PMOS $\log(I_{I,EAK})$distribution for 22 dies.

V_T roll-off curve. As the poly-Si pitch increases, more tensile stress is applied, increasing the mobility of NMOS transistors and raising the amount of NMOS leakage thereby offsetting the effect of increased gate length. PMOS leakage is not affected since it is insensitive to stress in a $\langle 100 \rangle$-oriented channel. At the same time, RO frequency also increases and this offsets the effect of a longer L.

2) Effect of a Longer Source/Drain Diffusion: Fig. 13 studies the impact of a longer S/D diffusion. Layout D1 has its S/D diffusion length increased by approximately 75%. Differences in S/D capacitances are captured by the layout extraction. After

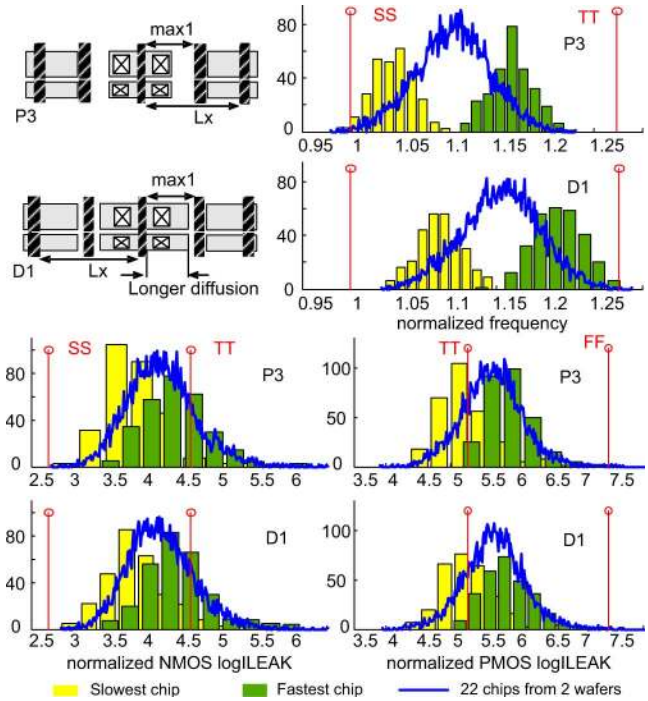

Fig. 13. Effect of a longer S/D diffusion on RO frequency and transistor leakage currents.

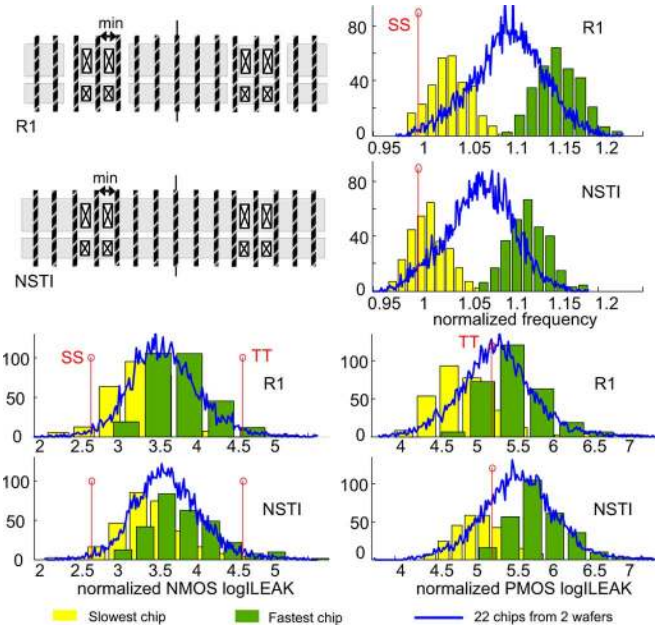

Fig. 14. Effect of STI on RO frequency and transistor leakage currents. NSTI uses gate isolation [21] instead of STI.

normalization, the layout with larger S/D area is about 5% faster, while the leakage currents remain approximately unchanged. This can be explained by the fact that a larger S/D area will allow the CESL to exert more tensile strain on the transistors, thereby increasing the mobility of the NMOS transistors and resulting in faster RO. Since mobility varies linearly with leakage current, its effect on leakage current in the log scale is small.

In addition to increased S/D diffusion length, the length of STI next to the transistor is reduced, resulting in reduced STI strain. Since STI in this process generates a weak tensile strain, layout D1 would experience less NMOS mobility enhancement due to STI strain compared to layout P3, partially offsetting the impact of increased tensile strain due to longer S/D diffusion. In the next Section V-B-III, measurement results show a relatively weak effect on mobility due to STI strain compared to the effect due to a longer S/D diffusion. Hence, we conclude that a longer S/D diffusion generates slightly more than the 5% impact that is observed in these measurements.

3) Effect of STI: Fig. 14 compares the layouts with and without STI. The layout without STI is slower by about 3% and has higher PMOS leakage current. Since PMOS leakage current is strongly dependent on L and not affected by strain, it is likely that the layout without STI (NSTI) has shorter L. As this process uses SACVD trench oxide that generates a low tensile strain, STI stress increases the mobility of NMOS transistors and causes the layout with STI to be faster than layout NSTI, thereby compensating the effect of longer L. Increased NMOS mobility also increases NMOS leakage current, offsetting the impact of a longer L. Variation in L could be due to the STI step that causes unevenness on the surface of the chip.

4) Symmetry: In the 90 nm test-chips, the difference between structures that are mirror image of each other is small and is likely due to the coma effect. Roughly 1–2% shift in the mean RO frequency was observed for single gate stages in the 45 nm process. The top layout in Fig. 15 has a higher RO frequency (∼1% higher) and higher PMOS leakage but lower NMOS leakage compared to the bottom layout. This could be due to a combination of the coma effect and asymmetry in the pocket doping process. However, since the difference is small, it is difficult to infer the exact cause.

5) Fixed Poly-Si Pitch: The use of fixed gate pitch has negligible impact on variability as shown in Fig. 16. RO frequency of fixed gate pitch layout (R1) has a σ/μ of 2.2% which is the same as for non-fixed gate pitch layout (P1). This shows that, as long as there is regularity, the use of fixed poly-Si pitches in a grid-like layout does not reduce variability significantly.

6) Other Effects: The remaining layout-dependent effects are metal coverage over gate, the impact of poly-Si near the ends of the gate and gate placement in the vertical direction. The impact of metal-2 coverage over gate (layout M1) and the impact of poly-Si near the ends of the gate (layout T1) on transistor performance after normalization are negligible. The impact of placement in the vertical direction (V1) on performance is significant (10% variation in frequency). However, this is likely due to the parasitics in the metal interconnects which have not been accurately extracted. In this case, comparison of the mean values of frequency is not valid.

VI. Conclusion

In 90 nm, the largest impact of layout on performance comes from gate poly-Si density, which causes a systematic shift in frequency of up to 10%. D2D variation is significant resulting in a $3\sigma/\mu$ of 15% over half a wafer. Finally, WID variation for identical structures is relatively small $(3\sigma/\mu \approx 3.5\%)$. The WID apparent spatial correlation of RO frequency is significant and shows a dependency on the direction of spacing and the

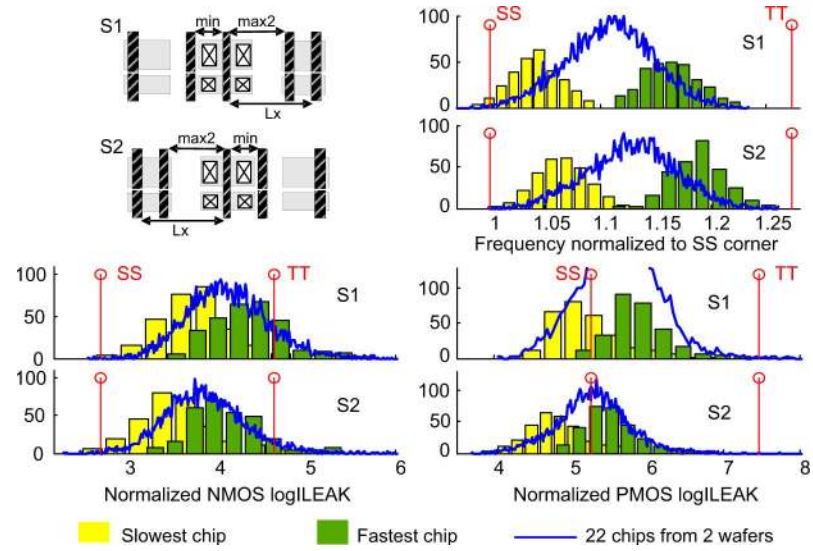

Fig. 15. Effect of symmetrical layouts which are mirror images of each other in the 45 nm test-chip.

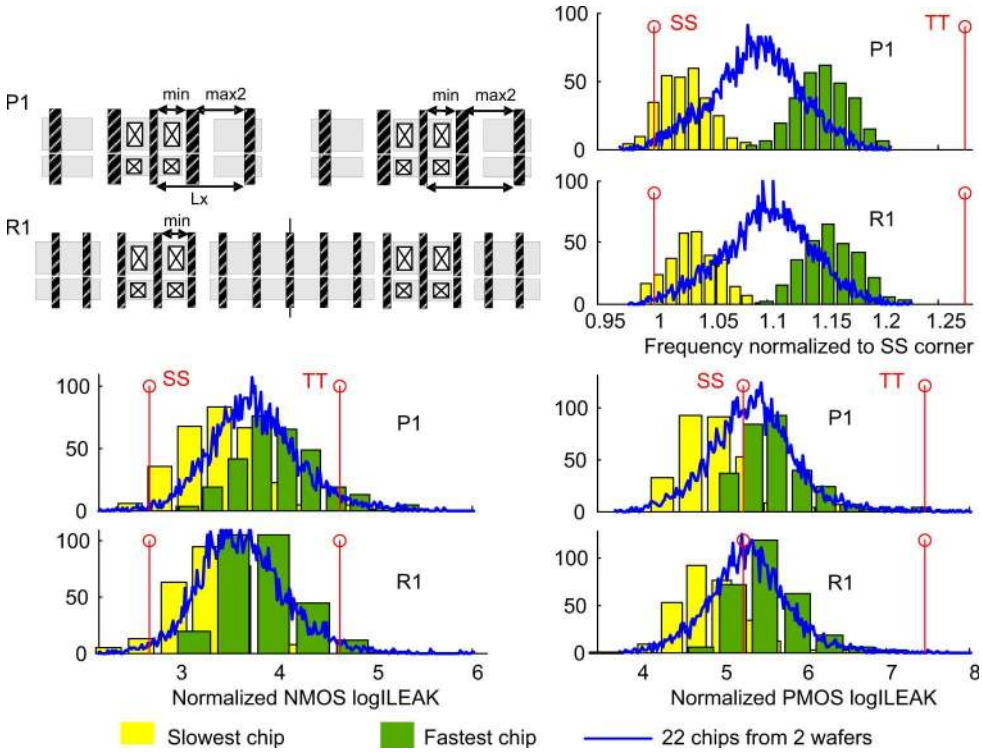

Fig. 16. Effect of fixed poly-Si gate pitch in the 45 nm test-chips.

orientation of the gates. This is because in the 90 nm data set we had a statistically significant systematic WID [24].

In 45 nm, systematic layout-induced variations, in particular those related to poly-Si density, are significantly reduced compared to the early 90 nm process. Variations due to strain and other factors are now dominant. Finally, random WID variation has increased proportionally to transistor area reduction, while systematic D2D variation has increased. Table IV compares the results of the two technologies.

Restricted design rules and likely better OPC in the 45 nm process reduce the layout induced performance variations by limiting variations in poly-Si density. However, other layout-induced variations due to CESL and STI stress have become more significant and are added on to the total variation.

TABLE IV
COMPARISON OF 45 NM AND 90 NM TECHNOLOGY RESULTS

Process	90nm	45nm
Proximity Effects	10%	2%
D2D ($3\sigma/\mu$)	15%	18% - 28%
WID ($3\sigma/\mu$)	3.5%	5.4% - 6.6%
WID Spatial Dependency (*apparent* spatial correlation)	Significant	Insignificant

Measurements show a trend towards less systematic but more random WID variations. From 90 nm to 45 nm, random WID variation has more than doubled whereas systematic layout-dependent variations have decreased. D2D variation is still important and depends strongly on the process corner. One way to reduce random WID variations is to increase channel area. As this increases area and power, only critical gates should be wider. Also, averaging more gates in a critical path will reduce the uncertainty over the path delay.

ACKNOWLEDGMENT

The authors wish to acknowledge the contributions of the students, faculty and sponsors of the Berkeley Wireless Research Center. The authors would also like to thank Ernesto Perea, Richard Ferrant, and the team of engineers in STMicroelectronics for their support; Professors Andrew Neureuther and Tsu-Jae King and students Andrew Carlson and Zheng Guo for helpful discussions and chip integration work; and students Jason Tsai, Kenneth Duong, and Emmanuel Adeagbo for their help with physical design and data collection.

REFERENCES

[1] K. A. Bowman, S. G. Duvall, and J. D. Meindl, "Impact of die-to-die and within-die parameter fluctuations on the maximum clock frequency distribution for gigascale integration," *IEEE J. Solid-State Circuits*, vol. 37, no. 2, pp. 183–190, Feb. 2002.

[2] L.-T. Pang and B. Nikolić, "Measurement and analysis of variability in 45 nm strained-Si CMOS technology," in *Proc. IEEE Custom Integrated Circuits Conf. (CICC'08)*, Sep. 2008, pp. 129–132.

[3] L.-T. Pang and B. Nikolić, "Impact of layout on 90 nm CMOS process parameter fluctuations," in *2006 Symp. VLSI Circuits Dig. Tech. Papers*, Honolulu, Hawaii, Jun. 2006, pp. 69–70.

[4] L.-T. Pang and B. Nikolić, "Measurements and analysis of process variability in 90 nm CMOS," *IEEE J. Solid-State Circuits*, vol. 44, no. 5, pp. 1655–1663, May 2009.

[5] S. Nassif, "Delay variability: Sources, impacts and trends," in *2000 IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, San Francisco, CA, Feb. 2000, pp. 368–369.

[6] J. W. Tschanz *et al.*, "Adaptive body bias for reducing impacts of die-to-die and within-die parameter variations on microprocessor frequency and leakage," *IEEE J. Solid-State Circuits*, vol. 37, no. 11, pp. 1396–1402, Nov. 2002.

[7] A. B. Kahng and Y. C. Pati, "Subwavelength lithography and its potential impact on design and eda," in *Proc. 36th Design Automation Conf. 1999*, New Orleans, LA, Jun. 1999, pp. 799–804.

[8] P. Oldiges, Q. Lin, K. Petrillo, M. Sanchez, M. Ieong, and M. Hargrove, "Modeling line edge roughness effects in sub 100 nanometer gate length devices," in *Proc. 2000 Int. Conf. Simulation of Semiconductor Processes and Devices*, Seattle, WA, Sep. 2000, pp. 131–134.

[9] A. Asenov, S. Kaya, and J. H. Davies, "Intrinsic threshold voltage fluctuations in decanano MOSFETs due to local oxide thickness variations," *IEEE Trans. Electron Devices*, vol. 49, pp. 112–119, 2002.

[10] D. J. Frank, Y. Taur, M. Ieong, and H.-S. P. Wong, "Monte Carlo modeling of threshold variation due to dopant fluctuations," in *1999 Symp. VLSI Circuits Dig. Tech. Papers*, Kyoto, Japan, Jun. 1999, pp. 171–172.

[11] M. Orshansky and A. Bandyopadhyay, "Fast statistical timing analysis handling arbitrary delay correlations," in *Proc. 41st Design Automation Conf. 2004*, San Diego, CA, Jun. 7–11, 2004, pp. 337–342.

[12] E. Josse *et al.*, "A cost-effective low power platform for the 45-nm technology node," in *2006 IEEE Int. Electron Devices Meeting Tech. Dig.*, San Francisco, CA, Dec. 2006, pp. 1–4.

[13] C. Cam *et al.*, "A low cost drive current enhancement technique using shallow trench isolation induced stress for 45-nm node," in *Symp. VLSI Technology Dig. Tech. Papers*, Honolulu, HI, Jun. 2006, pp. 82–83.

[14] B. L. Gratiet *et al.*, "Process control for 45 nm CMOS logic gate patterning," in *Metrology, Inspection, and Process Control for Microlithography XXII, Proc. SPIE*, J. A. Allgair and C. J. Raymond, Eds. Bellingham, WA: SPIE, 2008, vol. 6922.

[15] A. V.-Y. Thean *et al.*, "Uniaxial-biaxial stress hybridization for super-critical strained-Si directly on insulator (SC-SSOI) PMOS with different channel orientations," in *2005 IEEE Int. Electron Devices Meeting Tech. Dig.*, San Francisco, CA, Dec. 2005, pp. 509–512.

[16] R. A. Bianchi, G. Bouche, and O. R. dit Buisson, "Accurate modeling of trench isolation induced mechanical stress effects on MOSFET electrical performance," in *2002 IEEE Int. Electron Devices Meeting Tech. Dig.*, San Francisco, CA, Dec. 2002, pp. 117–120.

[17] G. Eneman, M. Jurczak, P. Verheyen, T. Hoffmann, A. D. Keersgieter, and K. D. Meyer, "Scalability of strained nitride capping layers for future CMOS generations," in *Proc. 35th European Solid State Device Research Conf., 2005*, Grenoble, France, Sep. 2005, pp. 449–452.

[18] T. A. Brunner, "Impact of lens aberrations on optical lithography," *IBM J. Research and Development*, vol. 41, pp. 57–67, Jan./Mar. 1997.

[19] H. Tuinhout, M. Pelgrom, R. P. de Vries, and M. Vertregt, "Effects of metal coverage on MOSFET matching," in *1996 IEEE Int. Electron Devices Meeting Tech. Dig.*, San Francisco, CA, 1996, pp. 735–738.

[20] V. Kheterpal *et al.*, "Design methodology for IC manufacturability based on regular logic-bricks," in *Proc. 42nd Design Automation Conf. 2005*, Anaheim, CA, Jun. 2005, pp. 353–358.

[21] I. Ohkura, T. Noguchi, K. Sakashita, H. Ishida, T. Ichiyama, and T. Enomoto, "Gate isolation—a novel basic cell configuration for CMOS gate arrays," in *Proc. IEEE Custom Integrated Circuits Conf. (CICC'82)*, May 1982, pp. 307–310.

[22] J. D. Plummer, M. D. Deal, and P. B. Griffin, *Silicon VLSI Technology*. Upper Saddle River, NJ: Prentice Hall, 2000.

[23] Q. Zhang *et al.*, "Across wafer critical dimension uniformity enhancement through lithography and etch process sequence: Concept, approach, modeling, and experiment," *IEEE Trans. Semiconduct. Manufact.*, vol. 20, pp. 488–505, 2007.

[24] K. Qian and C. J. Spanos, "A comprehensive model of process variability for statistical timing optimization," in *Design for Manufacturability Through Design-Process Integration II, Proc. SPIE*, V. K. Singh and M. L. Rieger, Eds. Bellingham, WA: SPIE, 2008, vol. 6925, pp. 1G-1–11.

[25] Kanno *et al.*, "Controlling gate-CD uniformity by means of a CD prediction model and wafer-temperature distribution control," *Thin Solid Films*, vol. 515, no. 12, pp. 4941–494, 2007.

[26] Q. Y. Tang and C. J. Spanos, "Layout optimization based on a generalized process variability model," in *Proc. SPIE*, San Jose, CA, 2008, vol. 6925.

[27] M. Pelgrom, A. Duinmaijer, and A. Welbers, "Matching properties of MOS transistors," *IEEE J. Solid-State Circuits*, vol. 24, no. 5, pp. 1433–1439, Oct. 1989.

Liang-Teck Pang (S'02–M'08) received the Dipl.Ing. from Ecole Centrale de Paris in France and the Master of Philosophy (M.Phil.) from Cambridge University, U.K., in 1997. Between 1998 and 2002, he worked in the DSO National Labs in Singapore on VLSI implementation of signal processing algorithms and high performance micro-architecture and circuit design. In 2002, he started his Ph.D. program in the department of Electrical Engineering and Computer Sciences at the University of California, Berkeley. In the summers of 2005-2007, he interned in IBM Austin Research Labs where he was an inventor on several US patent applications. He was presented with the IBM Invention Achievement Award in recognition of those contributions. Upon completion of his Ph.D. program in 2008, he joined the IBM T. J. Watson Research Center as a research staff member. His Ph.D. research involved the design of circuits to measure and characterize CMOS performance variability due to fluctuations in the manufacturing process. His emphasis is on the effects of layout on CMOS performance, measurement of the spatial correlation of logic gates, and analysis of variability data.

Kun Qian (S'06) received the B.S. degree from the Department of Microelectronics at Peking University, Beijing, China, in 2005. Between 2003 and 2005 he did his undergrad research with the Novel Devices Research Group of Institute of Microelectronics at PKU, which involves modeling of high-K dielectric based non-volatile memory devices. In 2005, He started his graduate study at the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, and is working towards the Ph.D. degree. His Ph.D. research is about characterization, analysis and modeling of integrated circuit manufacturing process and performance variability. His emphasis is on mechanisms of systematic and random spatial variation and layout dependent effects. He is also interested in general design for manufacturability techniques.

Costas J. Spanos (S'77–M'85–SM'95–F'00) received the Electrical Engineering Diploma with honors from the National Technical University of Athens, Greece and the M.S. and Ph.D. degrees in electrical and computer engineering from Carnegie Mellon University.

In 1988 he joined the faculty at the Department of Electrical Engineering and Computer Sciences of the University of California at Berkeley, where he is now a Professor and Associate Chair in the Department of Electrical Engineering and Computer Sciences, and the Chair of the Electrical Engineering Division at UC Berkeley. From 1994 to 2000 he was the Director of the Berkeley Microfabrication Laboratory, and from 2004 to 2008 has been the Associate Dean for Research in the College of Engineering. His present research interests include the development of flexible manufacturing systems, the application of statistical analysis in the design and fabrication of integrated circuits, and the development and deployment of novel sensors and computer-aided techniques in semiconductor manufacturing.

Borivoje Nikolić (S'93–M'99–SM'05) received the Dipl.Ing. and M.Sc. degrees in electrical engineering from the University of Belgrade, Serbia, in 1992 and 1994, respectively, and the Ph.D. degree from the University of California at Davis in 1999.

He lectured electronics courses at the University of Belgrade from 1992 to 1996. He spent two years with Silicon Systems, Inc., Texas Instruments Storage Products Group, San Jose, CA, working on disk-drive signal processing electronics. In 1999, he joined the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, where he is now a Professor. His research activities include digital and analog integrated circuit design and VLSI implementation of communications and signal processing algorithms. He is co-author of the book *Digital Integrated Circuits: A Design Perspective*, 2nd ed. (Prentice-Hall, 2003).

Dr. Nikolić received the NSF CAREER award in 2003, College of Engineering Best Doctoral Dissertation Prize and Anil K. Jain Prize for the Best Doctoral Dissertation in Electrical and Computer Engineering at University of California at Davis in 1999, as well as the City of Belgrade Award for the Best Diploma Thesis in 1992. For work with his students and colleagues he received the Best Paper Award at the ACM/IEEE International Symposium of Low-Power Electronics in 2005, and the 2004 Jack Kilby Award for the Outstanding Student Paper at the IEEE International Solid-State Circuits Conference.