Analysis and Optimization of SRAM Robustness for Double Patterning Lithography

Vivek Joshi, Kanak Agarwal, David Blaauw, Dennis Sylvester

Dept. of EECS, University of Michigan, Ann Arbor, MI. email: {vivekj, blaauw, dennis}@eecs.umich.edu, *IBM Research Lab, Austin, TX. email: kba@us.ibm.com

***Abstract* — Double patterning lithography (DPL) is widely considered the only lithography solution for 32nm and several subsequent technology nodes. DPL decomposes and prints the critical layout shapes in two exposures, leading to mismatch between adjacent devices due to systematic offsets between the two exposures. This results in adjacent devices with different mean critical dimension (CD), and uncorrelated CD variation. Such a mismatch can increase functional failures in SRAM cells and degrade yield. This paper analyzes the impact of DPL on functional failures in SRAM bitcells, and proposes a DPL-aware SRAM sizing scheme to effectively mitigate yield losses. Experimental results based on 45nm industrial models and test chip measurements show that DPL can significantly impact SRAM cell robustness. Using the proposed DPL-aware sizing scheme, the SRAM cell failure probability can be reduced by up to 3.6X. Also, for iso-robustness, cells optimized by the proposed approach have 7.9% lower dynamic energy as compared to non-DPL aware sizing optimization.**

1. INTRODUCTION

In nanometer CMOS optical lithography is being pushed to new extremes. The smallest printable feature size is defined by the Raleigh criterion to be $k_1\lambda/NA$ [1], where k_1 is the process difficulty factor, λ is the wavelength of the light source, and NA is the numerical aperture determined by lens size. Today's most aggressive single exposure production processes with off axis illumination have a k_1 factor of 0.36 - 0.4 for logic, and 0.29-0.30 for memory [2], which are quite close to the theoretical limit of 0.25. Currently, 193nm is the shortest wavelength in use for semiconductor production and is expected to continue its dominance for several technology nodes in the future. Using immersion lithography at 193nm (NA = 1.2), k_1 is required to be <0.2 to print 32nm pattern, which is lower than its theoretical limit. As a result, traditional lithography using 193nm wavelength light cannot print sub-32nm patterns. With significant technical hurdles delaying implementation of new lithography techniques, such as extreme ultraviolet (EUV), double patterning is the only viable solution to adhere to Moore's Law, despite the increased cost due to lower throughput and higher process complexity [3].

Double Patterning Lithography (DPL) [4, 5] partitions a critical-layer layout into two mask layouts and exposures, such that each individual exposure step takes place at a robust 0.35-0.4 k_1 factor, which is much more favorable for manufacturing compared to single exposure, ultra-low k_1 lithography. However, DPL incurs added complexity: more processing steps, throughput overhead, and tight overlay control between the two exposures. Several DPL schemes have been proposed in the past [6, 7], however the two most popular techniques are standard pitch-split DPL and the sacrificial self-aligned spacer with trim [8]. In pitch-split DPL either lines or spaces between lines are printed in two sequential processes. Thus, DPL is characterized by the existence of dual populations for critical dimension (CD), with uncorrelated variance and distinct means. So, for a fixed polysilicon gate pitch, devices on alternate poly tracks are correlated while devices on adjacent tracks are not. While such variation in gate length presents challenges to timing analysis and optimization of logic [9], it will have a much stronger negative impact on SRAM robustness where a mismatch between devices (e.g., access and pull-down devices) can cause significant yield loss.

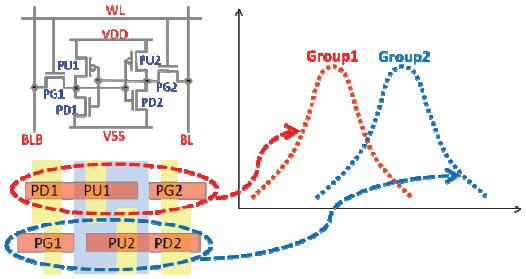

Figure 1: SRAM layout and DPL based variation.

Figure 1 shows the schematic and conventional layout of a typical six transistor SRAM cell. The access transistor and pull up/pull down (PU/PD) transistors for a given side lie on different poly tracks (e.g., PG1 lies on a different track than PU1/PD1) and will be printed with different exposures under DPL.[1] As a result the access and PU/PD transistors on a given side of the symmetric circuit structure will have uncorrelated gate length distributions, with such mismatch severely impacting the SRAM cell robustness by increasing functional failures. For example, if the access transistor becomes stronger than the PD transistor, the SRAM cell will be more prone to read failures. [10] presents modeling of SRAM failures, and statistical optimization to minimize yield loss, for single exposure lithography. In [11], the authors analyze the impact of lithographic variation on electrical yield of 32nm SRAM, for single patterning and cut-mask double patterning [12], where one exposure is used to print the polysilicon tracks and the other is used for cut-mask to print line ends. However, with technology scaling, the polysilicon pitch will go below the resolution limit of single exposure, and double patterning will be required to print the adjacent polysilicon tracks. To the best of our knowledge, this is the first work to analyze SRAM robustness under pitch-splitting double patterning, and propose a DPL-aware sizing scheme to mitigate yield loss due to DPL.

We use measurement (from a 45nm test chip) and simulation to analyze the impact of DPL-based variation on SRAM robustness, as compared to traditional single exposure lithography. We show that the DPL impact on cell robustness is substantial, and there is a need for DPL variation aware SRAM robustness analysis and optimization framework. We propose a DPL-aware SRAM sizing technique that iteratively sizes the SRAM cell to achieve desired robustness while changing the read and write energies by a very small amount (<5%). The proposed technique is very effective in mitigating the negative impact of DPL on SRAM robustness, and can improve V_{th} σ failure numbers by up to 9.8%, which translates to a 3.6X reduction in SRAM cell failure probability, for an area

[1] Additional techniques such as cut-mask [12] might be required to define line ends, but cut-mask does not change the printed polysilicon line widths (gate lengths). Hence, the use of cut-mask or similar techniques will not affect the analysis presented in this work.

overhead of only 1.6% (sizing optimization is performed on an industrial 45nm cell optimized for single exposure). For iso-robustness, cells optimized using the proposed technique have 7.9% lower dynamic when compared to non-DPL aware sizing optimization of SRAM cell.

The rest of the paper is organized as follows. Section 2 discusses the background and analysis of DPL impact on SRAM cell robustness, while Section 3 introduces the proposed DPL aware SRAM sizing technique. Experimental results are discussed in Section 4, and Section 5 concludes the paper.

2. BACKGROUND AND ANALYSIS

As mentioned in the previous section, device mismatch due to DPL can result in increased failure probability of the SRAM cell. This mismatch depends on the mean (μ) and standard deviation (σ) of the two line width distributions, for the two exposures used to print adjacent polysilicon tracks. Based on hardware results, [13] reported 3σ/μ numbers as high as ~16.5% for DPL line width distributions in 32nm technology. Parametric failures in SRAM cell are principally due to:

1. *Destructive Read/Read Failure* – flipping of the stored data in the cell while reading. Flipping occurs when bump in the read voltage is higher than the trip point of the other inverter (e.g., when the bump in the output of inverter PU2-PD2 (V_{read}) > V_{trip} for inverter PU1-PD1, while reading out a 0).
2. *Write failure* – failure to write to a cell within the time when wordline (WL) is high.
3. *Access time failure* – an increase in the access time of the cell violating the delay requirements.
4. *Hold failure* – destruction of the cell content in standby mode due to the application of lower supply voltage (in order to suppress leakage in standby mode).

In this section, we analyze the impact of DPL on SRAM variability and robustness through measurement and simulation.

2.1 Test Chip Measurement based Analysis

Figure 2 shows the stick diagram of polysilicon layer depicting how rows of SRAM cells are laid out. Each row of SRAM cell is the mirror image of adjacent rows. As a result, if transistors on polysilicon track A are stronger (have smaller channel lengths) than those on track B in Row 1, transistors on track B will be stronger in Row 2. Hence, for a given operation (read 1, write 1, etc.), even and odd rows are expected to show opposite behavior, which could be quantified in terms of failure count. However, two subsets of even/odd rows are supposed to show similar behavior. 75 test chips, implemented in a 45nm CMOS technology that uses DPL, were measured for write 1 failure count at lowered V_{DD}. Figure 3 shows the failure count distribution for even and odd rows, as well as two subsets of even rows (subset 1 comprising of rows 4, 8, 12, etc. and subset 2 comprising of rows 2, 6, 10, etc.). As expected, the distributions for even and odd rows are very different, while those for the two subsets are much more similar. The difference in mean number of failures for even and odd rows is ~14.5%, while the difference in mean between the two subsets of even rows is ~3%. Student's t-test [14] was performed on the two sets of data, and the probability of the even and odd rows assuming the null hypothesis was found to be 0.025, and for the two subsets of even rows this probability was found to be 0.62. If two sets of data points follow the null hypothesis, it means that they belong to the same kind of data, and any difference in mean is due to chance/random variation. Null hypothesis can be rejected in case of even and odd rows (probability < 0.05), while it is followed by two subsets of even rows. So, Student's t-test results conclusively prove

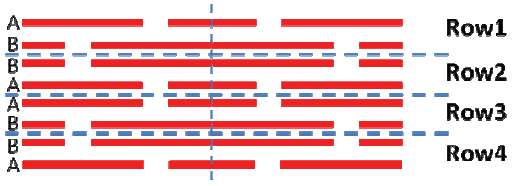

Figure 2: Stick diagram (polysilicon only) showing how rows of SRAM cells are laid out.

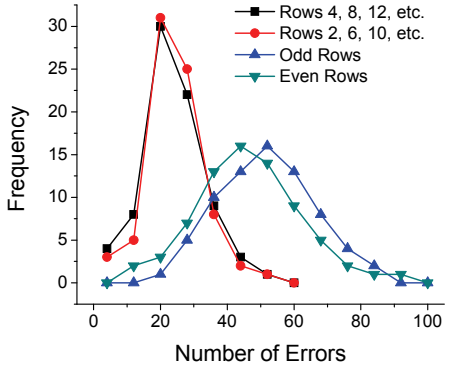

Figure 3: Write 1 failure count distribution for even, odd rows, and two subsets of even rows.

that under DPL based mismatch can have a significant impact on SRAM failure count. If we assume that the two line width distribution curves for the two exposures have the same mean (μ) and standard deviation (σ), then based on simulation and measured difference in error count 3σ/μ is roughly estimated to be ~12.8%.

2.2 Simulation based Analysis

In order to analyze the effect of DPL on SRAM parametric failures, we must begin with the analysis of the failure triggering mechanisms under DPL based mismatch. First, we study the effect of DPL based mismatch on V_{read} and write time of a 45nm SRAM, and compare the effect to that of a single exposure based system. V_{read} is defined as bump in the output voltage of inverter PU2-PD2, while reading out a 0 from SRAM. Higher the value of V_{read}, more prone is the cell to read failure under random V_{th} variation. Similarly, a cell with higher write time will be more likely to experience write failure under random V_{th} variation. For cell level DPL based analysis, we assume that gate length distribution of PD1, PU1, and PG2 (mean μ1, standard deviation σ1), is uncorrelated with the gate length distribution of PD2, PU2, and PG1 (μ2, σ2), which lie on separate polysilicon track. Our analysis focuses on a 45nm industrial SRAM cell which is optimized for single exposure based patterning. Based on V_{th} corner analysis, the nominal cell experiences read failure at a V_{th} σ value of $4.23\sigma_{VT0}$, and write failure σ value is $6.36\sigma_{VT0}$, where σ_{VT0} is the standard deviation of intra-die V_{th} variation specified for the technology. These numbers establish that, in general, write operation is much more robust for the industrial SRAM being analyzed, which provides the designer an opportunity to make the read operation more robust at the cost of degrading write robustness by a small amount. A similar opportunity will exist in case the read operation is more robust than write. We exploit this property later on in our DPL-aware sizing optimization.

Figure 4 shows the V_{read} distribution for the simple case of equal means (μ1 = μ2) and standard deviations (σ1 = σ2), with 3σ1/ μ1 = 3σ2/ μ2 = 10%, for the two line width distribution curves. Also shown in Figure 2, is the distribution of V_{read} for single exposure

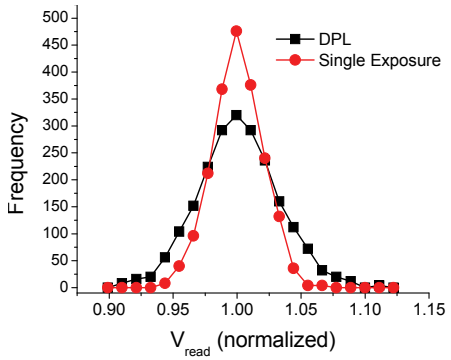

Figure 4: V_{read} distribution for DPL and single exposure system.

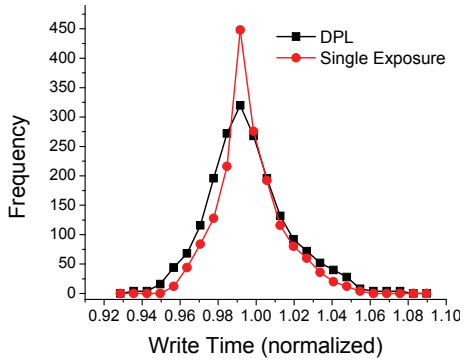

Figure 5: Write time distribution for DPL and single exposure system.

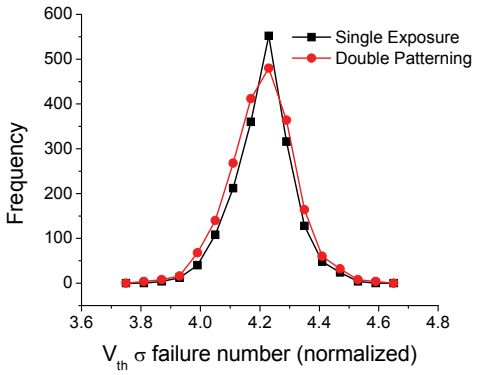

Figure 6: Read V_{th} σ failure number distributions for DPL and single exposure lithography.

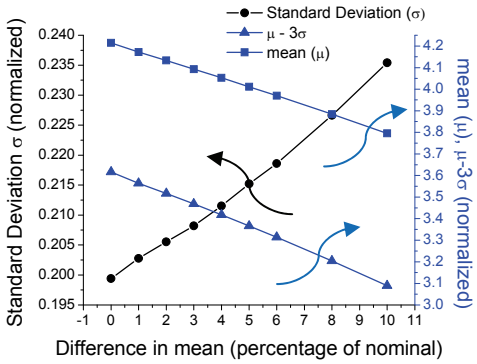

Figure 7: Read V_{th} standard deviation, mean and μ-3σ as a function of difference in means of the line width distribution curves for DPL.

case assuming 3σ/ μ is 10%. All the V_{read} values are normalized to V_{read} for the nominal cell (V_{read0}). The mean and variance of V_{read} distribution for DPL is found to be ~1 and 0.035 (normalized to nominal V_{read}), while the mean and variance for single exposure are ~1 and 0.018, respectively. Standard deviation in the case of DPL based technology is almost twice the standard deviation for single exposure system. DPL μ + 3σ (1.11 V_{read0}) is higher than the single exposure case (1.06 V_{read0}). Hence, it is important to consider both gate length and V_{th} variation for accurate variability/robustness analysis in DPL systems. Figure 5 shows similar plots for write time analysis. DPL based distribution has a mean value of ~1 and a standard deviation of 0.024, while the single exposure mean and standard deviation are ~1 and 0.014, respectively (normalized to nominal write time). Again, the standard deviation of DPL case is higher than the single exposure, suggesting that there is a need for DPL variation aware analysis. DPL almost doubles the standard deviation observed in the case of single exposure system, when the means of the two gate length distribution curves are identical. In case there is a difference in means, impact of DPL increases even further due to increase in mismatch between transistors on adjacent polysilicon tracks.

Next we look at the distribution of V_{th} σ failure numbers for DPL and single exposure systems, for read operation. Figure 6 shows the read V_{th} failure distributions for both the cases. These distributions are generated by performing V_{th} corner based failure analysis at each gate length sample, to find the smallest V_{th} σ number at which the cell experiences functional failure. As expected based on V_{read} analysis, double patterning leads to much worse V_{th} failure numbers (or the V_{th} failure σ distribution has higher variance). Mean of the read V_{th} failure curve for DPL is $4.20\sigma_{VT0}$, with a standard deviation of 0.2 σ_{VT0}. Single exposure read mean is $4.23\sigma_{VT0}$, and standard deviation is $0.1\sigma_{VT0}$, and the standard deviation is half of that in the case of DPL. For the μ-3σ point in the distribution, the probability of failure increases by ~3.3X due to DPL, as compared to single exposure lithography.

We look at another way to analyze the read stability instead of the computationally intensive analysis involving V_{th} corner analysis on every length sample. We pick the worst few (1%) length sample points for read operation, by using the V_{read} distribution curve, and the knowledge that high values of V_{read} make the sample point more vulnerable to V_{th} variation based failure. On these selected points, we run V_{th} corner analysis and take an average of the failure numbers**.** The average roughly corresponds to μ - 2.8σ for the case of read failure. This shows that the corner cases of the V_{read} distribution are the ones which generate lowest (worst) V_{th} σ failure numbers. In other words, V_{th} corner analysis on the worst cases of DPL based V_{read} distribution captures most of the worst cases (lowest V_{th} σ failure values) of the complete analysis involving finding V_{th} failure numbers at each length sample. So, if the aim of an analysis is to capture the worst case, then length-based analysis and the V_{th} corner analysis can be decoupled while still capturing most of the bad cases (more prone to functional failure).

Figure 7 shows how the mean (μ), standard deviation (σ), and μ-3σ points of the read V_{th} σ failure distribution vary if a difference in mean is introduced between the two curves ($\mu1 \neq \mu2$), for read operation. All the values are plotted against the difference in means of the two DPL length distributions (expressed as a percentage number of the nominal value). As the difference in mean of the two length distributions increases, mean of the V_{th} σ failure distribution decreases, and variance increases which means that the cell becomes less robust. For difference in mean of 4%, μ-3σ value of the V_{th} failure distribution goes down to as low as $3.41\sigma_{VT0}$, which ~13% smaller than the value for single patterning, and the probability of failure for the μ-3σ point increases by ~7X. Hence, the impact of DPL on SRAM cell robustness greatly increases with

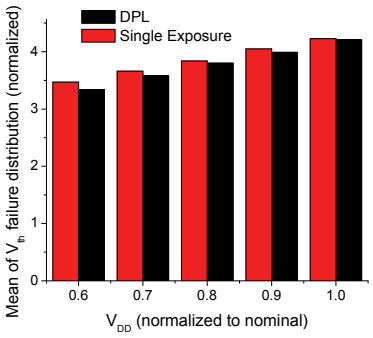

Figure 8: Mean of V_{th} failure distribution as a function of V_{DD} scaling for DPL and single exposure techniques.

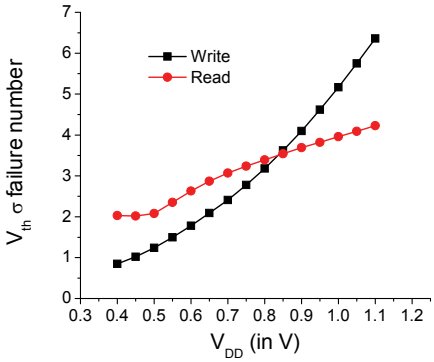

Figure 9: Read and write V_{th} σ failure numbers as a function of V_{DD}.

the increase in the difference between the mean of two gate length distribution curve. This is expected since increasing the difference in mean of the two length curves increases the mismatch between SRAM devices, thereby making them more prone to failure.

An interesting analysis examines the mean and variance of the V_{read} distribution if PG1, PU1, and PD1 were on the same poly track, assuming that the layout could be changed in such a manner. Now the access transistor and PU/PD transistors will have identical lengths and there would be no mismatch there due to DPL. As a result, we would expect the V_{read} distribution to be much closer to the single exposure case. For the simple case of equal means ($\mu 1 = \mu 2$), and variances, with $3\sigma 1/\ \mu 1 = 3\sigma 2/\ \mu 2 = 10\%$, V_{read} distribution has a mean of ~1 and standard deviation of 0.02 (normalized to nominal V_{read}). These values are very close to single exposure case as expected ($\mu = 1$, $\sigma = 0.018$). However, the actual V_{th} σ failure numbers are higher than the single exposure case. This is because of the mismatch between the two inverters due to different distributions. For example, in case of a reading a zero, although V_{read} is not affected much (as PG1, PU1, and PD1 lie on the same litho track), the trip voltage (V_{trip}) of the other inverter (PU2-PD2) depends on the other uncorrelated length distribution (for PU2, PD2, and PG2). This mismatch between V_{read} and V_{trip} can cause samples with high V_{read} and low V_{trip}, which are very vulnerable to random V_{th} variation based failure. Even despite this mismatch, the failure numbers are much better than DPL based results for the original layout, and this could potentially be a useful design optimization to mitigate yield losses due to DPL. However, such a change in layout leads to very high area penalty (~2×), making this an unattractive design change.

Finally, we look at the effect of DPL on voltage scaling. With the SRAM cell fixed at the nominal sizing, V_{DD} is scaled to analyze the impact of DPL on V_{th} σ failure number. Figure 8 shows how the mean of read V_{th} σ failure distribution varies with V_{DD} for DPL and single exposure systems. V_{DD} values are normalized to the nominal V_{DD}. Supposing we require the mean of the V_{th} failure distribution to be greater than 3.5 σ_{VT0}, we find that the supply voltage can be scaled down to 0.62 of the nominal VDD under single exposure. However, degraded robustness under DPL leads to less favorable voltage scalability, requiring a supply voltage of at least 0.68 of the nominal VDD, leading to approximately 20% energy penalty. Figure 9 shows how the V_{th} σ failure number changes for read and write operations as V_{DD} is lowered, at nominal value of gate length. Write operation is more stable at nominal V_{DD} (higher value of V_{th} σ failure number), however as V_{DD} is lowered it becomes less robust than read operation. At nominal V_{DD}, DPL-aware sizing optimization can make read operation more robust at the cost of degrading write robustness by a small amount. However, at lower values of V_{DD}, write operation needs to be optimized and made more robust, and this could be done at the cost of read robustness.

3. DPL-AWARE SRAM SIZING

Based on the intuition developed through analyzing the impact of DPL on SRAM cell robustness, we now propose a DPL-aware SRAM sizing scheme to mitigate the negative impact of DPL on SRAM robustness. Key points to remember from the analysis section are:

- Typically read and write robustness numbers are very different for SRAM, providing the designer an opportunity to trade the robustness of one operation off for the other. For the SRAM cell under consideration write is more robust than read at nominal V_{DD}, which is the common case in modern SRAMs (read is more stable at lower V_{DD}).
- The length-based analysis and the V_{th} corner analysis can be decoupled, and the DPL sizing optimization can focus on optimizing the worst cases (say $\mu+3\sigma$) of the V_{read} and write time distributions.
- Given a range in which the means and variances of the two gate length distributions could lie, there is a worst case combination that creates maximum mismatch (highest values of mean and standard deviation for line width distribution curves). Any sizing optimization should be directed at this worst case, and the other intermediate cases are expected to improve by using the resulting sizes. This fact is verified in the experimental results.

The DPL-aware SRAM sizing optimization problem can be viewed as that of shifting the V_{th} σ failure number distribution (Eg. Figure 6) to the right for the less robust operation, while meeting the constraints on read and write times (to avoid access failures), and read/write energies. Shifting the distribution to the right would increase the value of V_{th} σ failure number, and hence decrease the probability of failure, for any given point on the distribution (higher value of failure σ means lower probability of failure). We can choose a representative point on the distribution (of the form μ-aσ), and try to shift it to the right (or increase its V_{th} σ failure number) to achieve this goal. This is based on the assumption that variance of the V_{th} σ failure number distribution would not change drastically during the sizing optimization, and so increasing the V_{th} σ failure number for one point is the same as shifting the entire curve to the right. This assumption is validated by the experimental results discussed in the next section, where the variance of the curve is almost the same before and after the optimization. For our experiments, we choose this representative point as the mean (μ) of the failure number distribution curve. Hence, the problem can now be seen as maximizing the mean (μ) of V_{th} σ failure distribution for the less robust operation, while meeting the constraints on read and write times (to avoid access failures), and read/write energies. Hold failures were demonstrated to have much lower occurrence, and they can be further controlled by appropriately choosing standby

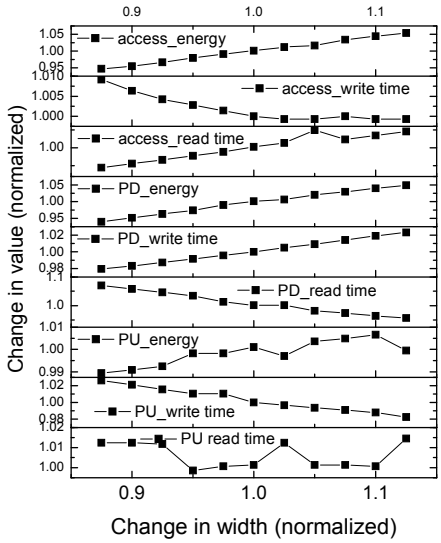

Figure 10: Variation in read, write times, and average energy with change in w_{PU}, w_{PG}, or w_{PU} for nominal gate lengths.

mode supply voltage [10], so they were not included in the sizing optimization as constraints. Only the widths of the SRAM devices were used as optimization variables. The problem can then be stated as:

$$\begin{aligned} &Maximize \quad f(w_{PU}, w_{PD}, w_{PG}) = \min(\mu_{Vth,read}, \mu_{Vth,write}) \\ &subject\ to: \\ &\quad E_{read} + E_{write} < E_0 \\ &\quad T_{read} < T_{r,0} \\ &\quad T_{write} < T_{w,0} \end{aligned} \tag{1}$$

where w_{PU}, w_{PD}, w_{PG} are the widths of pull up, pull down, and access device, respectively, $\mu_{Vth,read}$, $\mu_{Vth,write}$ are the mean of V_{th} σ failure number distribution for read and write, respectively, E_{read}, E_{write} are the read and write energies, while T_{read} and T_{write} are read and write times, respectively. The function f chooses the less robust of the two operations (read/write) to optimize as the one with lower value of mean of V_{th} σ failure number distribution, while E_0, $T_{r,0}$, and $T_{w,0}$ are the constraints on total energy, read time, and write time, respectively. In order to solve this problem, we can use the intuition developed during analysis, and decouple the length based analysis and V_{th} corner analysis. As a result, we can try to minimize the worst case values of V_{read} distribution if write is more robust than read, or write time distribution if read is more robust than write, while considering only DPL based gate length (line width) variation. This can be seen as shifting the V_{read} (Figure 4) or write time (Figure 5) distribution to the left, thereby making the cell more robust. Eg. if we shift the V_{read} curve to the left, we decrease the V_{read} value for any given point on the curve. Lower the value of V_{read}, less prone is the cell to read failure under random V_{th} variation. Thus shifting the V_{read} curve to the left would increase the robustness which results in higher value of V_{th} σ failure numbers. Again, we can choose a representative point on the V_{read}/write time curve and minimize it (shift it to the left). This minimization can be seen as minimizing the value of a point say P=μ+aσ on the V_{read}/write time curve (for our analysis a=3).

Now if we try to size the SRAM cell iteratively where we can change one width value by one step (say 1% of the nominal size) at a time; at each step of the iteration we will have the choice to change either w_{PU}, w_{PG}, or w_{PD}. But changing each of them by one step has a different effect on the read/write times, energies, and V_{read}. Figure 10 shows the variation of read and write times, and average energy ((E_{read}+E_{write})/2) with change in w_{PU}, w_{PG}, or w_{PU}, for nominal value of gate length. For each sub-plot, width of one of access (w_{PG}), pull up(w_{PU}), or pull down(w_{PD}) transistors is varied, while keeping the other two values fixed at nominal, and one of the values out of read time, write time, or average energy is plotted as a function of the width being varied. All the values are normalized to their value for the nominal cell. The key conclusion from the figure is that, the decision on which transistor to size at a given step in the optimization iteration, depends on the actual values of w_{PU}, w_{PG}, and w_{PD}. For example in order to increase read robustness, we can choose increase either of w_{PU} or w_{PD}, or reduce the size of the access transistor w_{PG} by one step, but the best choice depends upon the actual values of w_{PU}, w_{PG}, and w_{PD} at that point. In order to choose the best width value to change, we define a sensitivity metric G, based on the decrease in the value of point P (ΔP) and change in the value of a constraint function C (ΔC), where point P is the representative point chosen on the V_{read}/write time curve.

$$\begin{aligned} &G = \frac{\Delta C}{\Delta P}, \qquad P = \mu + a\sigma \quad (a = 3) \\ &C = w_1 \frac{E_{read}}{E_{r,nom}} + w_2 \frac{E_{write}}{E_{w,nom}} + w_3 \frac{T_{read}}{T_{r,nom}} + w_4 \frac{T_{write}}{T_{w,nom}} \end{aligned} \tag{2}$$

where $E_{r,nom}$, $E_{w,nom}$, $T_{r,nom}$, and $T_{w,nom}$ are the nominal values of read energy, write energy, read time and write time, respectively. $w_{1,2,3,4}$ are positive numbers less than 1, such that

$$w_1+w_2+w_3+w_4 = 1 \tag{3}$$

At each step we calculate G for a single step change in each of w_{PU}, w_{PG}, and w_{PD}, and accept the change that yields the minimum value of G. Using different weights, we can define the relative importance of constraints. To calculate P, we need the mean and variance of V_{read} or write time distribution curve, given the mean and variances of the two length distribution curve (μ_1, σ_1, μ_2, σ_2), and a set of device widths. We use a Taylor series expansion to calculate the mean and variance of a function $y = f(l_1, l_2)$, where l_1 and l_2 are the two independent random variables representing the two length distributions.

$$\begin{aligned} &\mu_y = f(\mu_1, \mu_2) + \frac{1}{2}\sum_{i=1}^{2} \frac{\partial^2 f(l_1,l_2)}{\partial l_i^2}\Big|_{\mu_i} \sigma_i^2 \\ &\sigma_y^2 = \sum_{i=1}^{2} \left(\frac{\partial f(l_1,l_2)}{\partial l_i} \right)^2 \sigma_i^2 \end{aligned} \tag{4}$$

To verify the accuracy of this expression in our analysis setup, we calculated the value of mean and variance for the V_{read} distribution curve of Figure 2 for the nominal cell using (4). The mean was calculated to be 1, and the variance was 0.036 (normalized to nominal value of V_{read}), which is very close to the simulated values for the curve (mean = 1.0, variance = 0.035). The mean and variance values from the Taylor Series expansion are merely used to guide the iterative optimization in the right direction (through the sensitivity metric), and so the fact that these values are slightly inaccurate (based on approximation) does not affect the final result of the sizing optimization significantly. Hence, using Taylor series expansion is a reasonable approximation to make. A flowchart for the proposed SRAM sizing optimization algorithm is shown in Figure 11. The next section discusses the experimental results using the proposed algorithm.

4. EXPERIMENTAL RESULTS

We use our proposed technique to optimize an industrial 45nm SRAM bitcell optimized for single patterning lithography, considering DPL based variation. For the purpose of analysis, we assume that the two gate length distribution curves (for the adjacent polysilicon tracks) have the same standard deviation, with 3σ1/ μ1 = 3σ2/ μ2 = 10%, and their means can differ by up to 5% (|μ1-μ2| <=5%). As discussed in Section 3, we run our DPL-aware sizing optimization algorithm for worst case combination of mean

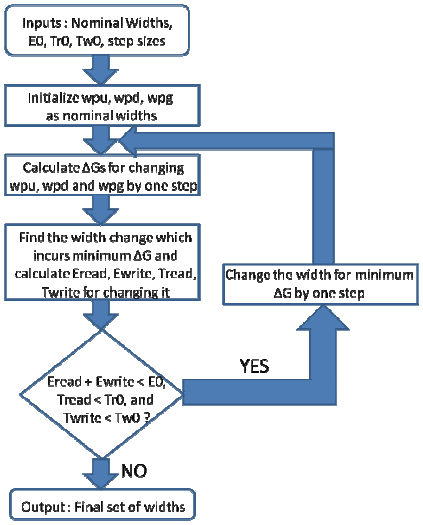

Figure 11: Proposed SRAM sizing optimization algorithm.

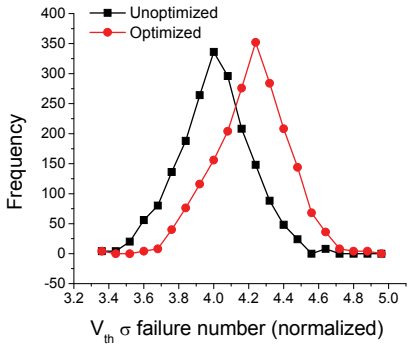

Figure 12: V_{th} σ failure distribution for read operation before and after the proposed optimization.

and standard deviation that creates maximum mismatch (3σ1/ μ1 = 3σ2/ μ2 = 10%, and difference in mean is maximum = 5%), and then examine at the V_{th} σ failure improvement at intermediate values to examine the effectiveness of the algorithm in mitigating yield losses at intermediate values.

The read operation is less robust for the analyzed SRAM. Figure 12 shows the V_{th} σ failure number distribution for read operation before and after the application of the proposed approach. Sizing optimization shifts the curve to the right, thereby making the SRAM cell more robust (higher value of V_{th} σ failure number means lower probability of failure). The mean of the V_{th} failure number distribution for optimized SRAM is 4.22 σ_{VT0} (very close to the mean in the case of single exposure for 3σ/μ = 10%), while the standard deviation observed is ~0.21σ_{VT0}. The absolute value of variance remains almost the same before and after the optimization (standard deviation = ~0.21σ_{VT0} for unoptimized case), just the curve is shifted to the right through optimization. This validates the assumption made during the sizing optimization that variance of the V_{th} σ failure number distribution would not change drastically during the sizing optimization. The μ-3σ of V_{th} distribution for application of proposed sizing is 3.57σ_{VT0}, which is a 6.2% improvement over the μ-3σ of the unoptimized DPL curve. This corresponds to a 2.17X reduction in failure probability of the μ-3σ point in the distribution. These values are for a maximum allowed change of 5% in E_{read}+E_{write}, compared to the nominal value (E_0 = 1.05($E_{r,nom}$ + $E_{w,nom}$)). Figure 13 shows the variation of the percentage improvement in the μ-3σ of V_{th} σ failure number

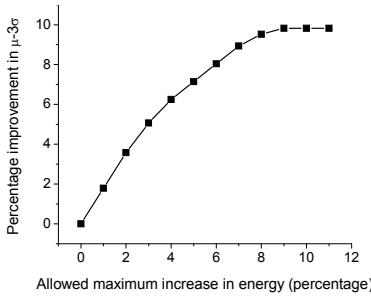

Figure 13: Variation of percentage improvement in the μ-3σ point of the V_{th} σ failure distribution with maximum allowed change in energy for the optimization algorithm.

distribution achieved by the proposed optimization over unoptimized SRAM, as a function of maximum allowed normalized value of E_{read}+E_{write} (E_0/($E_{r,nom}$ + $E_{w,nom}$)). The improvement number goes up with increase in maximum allowed change in dynamic energy, and saturates for an allowed change of ~9%. Maximum improvement in μ-3σ point is ~9.8%, which corresponds to 3.6X reduction in cell failure probability, for an SRAM cell area overhead of only 1.6%. Beyond this value, increasing the allowable energy penalty gives no further improvement because any further sizing violates the write time constraint (T_{write} < $T_{w,0}$). For this maximum improvement sizing point, μ-3σ for the write V_{th} σ failure number distribution decreases to ~5.19σ_{VT0} from its value of ~5.43σ_{VT0} for the unoptimized case. Even after the sizing optimization, write operation remains the more robust operation, but read robustness improves significantly.

Table 1 summarizes the percentage improvement in mean and μ-3σ points for V_{th} failure distribution curve of the optimized case over non optimized SRAM, for intermediate variation numbers (3σ1/ μ1 <=10%, 3σ2/ μ2 <= 10%, |μ1-μ2| <= 5%), and maximum allowed change in dynamic energy of 5% (E_0 = 1.05($E_{r,nom}$ + $E_{w,nom}$)). The improvement numbers obtained decrease as the mismatch is decreased (low variances and difference in mean). This is because there is less room for optimization, as the mean and μ-3σ values are closer to the nominal case. However, the proposed technique does ensure that we get almost all of the possible improvement given the constraints. Hence, SRAM cell optimized for worst case variation provides good improvement in SRAM robustness at intermediate points.

Next, we compare our approach to an approach where SRAM cell is over-optimized under single exposure based variation, in order to achieve better robustness under DPL based variation. For this purpose we use an algorithm similar to our proposed algorithm, but with a single length distribution curve instead of two (as in DPL). We find that in order to achieve similar robustness as the DPL aware sizing scheme, the constraints on energy and access times have to be relaxed. Such a technique results in higher energy and slower access times as compared to DPL-aware sizing optimization for the same value of improvement over the unoptimized case. For iso-robustness, such a technique results in 7.9% higher energy (E_{read}+E_{write}), and 4.6% larger access times as compared to the proposed technique.

Finally, we analyze the improvement in voltage scalability of the SRAM using the proposed technique. We fix the desired mean of the V_{th} σ failure number distribution to be greater than 3.5σ_{VT0}, and calculate the V_{DD} to which the SRAM can be scaled given the two length distribution curves before the mean goes below the desired value, with and without the application of proposed sizing optimization. Figure 14 plots the ratio of minimum V_{DD} allowed in

Table 1: Robustness improvement numbers for intermediate values of mean and standard deviation for DPL length distributions

DPL Length Distribution			Improvement in V_{th} failure numbers	
$3\sigma_1/\mu_1$	$3\sigma_2/\mu_2$	Mean difference	Mean	$\mu - 3\sigma$
10%	10%	5%	5.2%	6.2%
10%	10%	3%	3.1%	4.3%
10%	10%	1%	1.0%	2.4%
10%	10%	0%	0.1%	1.5%
9%	10%	3%	3.2%	4.8%
10%	8%	2%	2.1%	3.5%

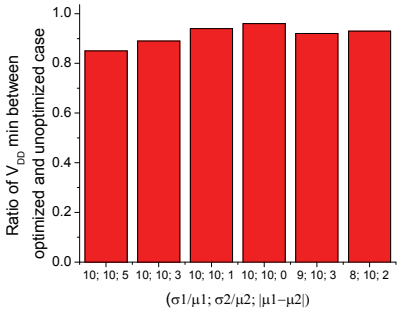

Figure 14: Ratio of minimum V_{DD} allowed in optimized and unoptimized case for a variety of mean and standard deviation combinations for the gate length distributions.

the optimized and unoptimized case for a variety of mean and standard deviation combinations of the two gate length distribution curves. On an average, the proposed technique improves V_{DD} scalability by 9.5% (18.1% reduction in energy). Thus, the proposed DPL-aware sizing optimization approach is shown to effectively mitigate the yield loss at very small area and energy penalty.

5. CONCLUSION

Double patterning lithography results in adjacent devices with different mean critical dimension (CD), and uncorrelated CD variation. Such a variation can increase functional failures in SRAM cells, which are very sensitive to mismatches, and degrade yield. In this paper, we analyze the impact of DPL on parametric functional failures in SRAM cells and propose a DPL-aware SRAM sizing scheme to effectively mitigate the yield losses for a very small energy and area overhead. Experimental results based on 45nm models show that DPL can significantly impact the SRAM cell robustness and hence there is a need for DPL aware analysis and optimization of SRAM cells. The proposed technique is very effective in mitigating the negative impact of DPL on SRAM robustness, and can improve Vth failure numbers by up to 9.8%, which translates to a 3.6X reduction in SRAM cell failure probability, for a very small area penalty of 1.6%.

References

[1] A. K. K. Wong, Resolution Enhancement Techniques in Optical Lithography, SPIE PRESS, 2001, p. 28.

[2] Sematech High Index Workshop, Oct. 2006, Kyoto.

[3] G. Capetti et al.," Sub k1 = 0.25 Lithography with Double Patterning Technique for 45nm Technology Node Flash Memory Devices at 193nm," Proc. SPIE Optical Microlithography, vol. 6520, pp. 65202K-1 - 65202K-12.

[4] J. Finders, et. al, "Double Patterning Lithography: The Bridge Between Low k1 ArF and EUV," Microlithography World, Feb. 2008.

[5] M. Drapeau, et. al, "Double Patterning Design Split Implementation and Validation for the 32nm Node," Proc. SPIE Design for Manufacturability through Design-Process Integration, Vol. 6521, 2007.

[6] W.-Y. Jung, et. al, "Patterning with amorphous carbon spacer for expanding the resolution limit of current lithography tool," Proc. SPIE 6156, 6156J1, 2006.

[7] C.-M. Lim, et. al, "Positive and negative tone double patterning lithography for 50-nm flash memory," Proc. SPIE 6154, 615410, 2006.

[8] M. Maenhoudt, et.al, "Double Patterning Scheme for Sub-0.25 k1 Single Damascene Structures at NA=0.75, λ=193nm," Proc. SPIE Conference on Optical Microlithography, 2005, pp. 1508-1518.

[9] K. Jeong et. al, " Timing analysis and optimization implications of bimodal CD distribution in double patterning lithography," Proc. ASPDAC 2009, pp. 486-491.

[10] S. Mukhopadhayay et. al, "Modeling of Failure Probability and Statistical Design of SRAM Array for Yield Enhancement in Nanoscaled CMOS," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, Volume 24, Issue 12, pp. 1859-1880, December 2005.

[11] S. Verhaegen et. al, "Litho variations and their impact on the electrical yield of a 32nm node 6T SRAM cell," Proc. SPIE, vol. 6925, 69250R, 2008.

[12] C. Sarma et. al, "Double exposure double etch for dense SRAM: a designer's dream," Proc. SPIE, vol. 6924, pp. 692429-692429-9, 2008.

[13] M. Dusa et al., "Pitch Doubling Through Dual-Patterning Lithography: Challenges in Integration and Litho Budgets," Proc. SPIE Conference on Optical Microlithography, pp. 65200G-1 - 65200G-10, 2007.

[14] R. R. Wilcox, "Introduction to robust estimation and hypothesis testing," Elsevier academic press, 2005, p. 155.