Overhead Reduction with Optimal Margining Using A Reliability Aware Design Paradigm

S. Mishra, P. Weckx, O. Zografos, J. Y. Lin, A. Spessot and F. Catthoor
IMEC vzw, Leuven 3001, Belgium
Phone: +32-477253988; email: subrat.mishra@imec.be

Abstract—Design margins are necessary to ensure reliable operation of integrated circuits over extreme ranges of environmental variations (Voltage, Temperature) and manufacturing Process variations. On top of these PVT variations, aging related parametric drift (e.g. due to BTI, HCI) also limits performance by requiring additional timing margin. In principle, corner based design methodology can be adopted. However, this approach is sub-optimal, because it applies margins which may be either too optimistic or pessimistic since it tends to ignore the correlation effects which exist inherently due to the circuit topology and the workload effects. In this paper, we propose a workload-dependent reliability aware optimization flow under the influence of NBTI aging by utilizing an optimal margining scheme. The proposed flow takes into account the relevant correlations in a design by modelling the degradation accurately and thus enables achieving the desired Power-Performance-Area (PPA) goals without severe reliability penalty.

I. INTRODUCTION

Power-Performance-Area (PPA) gains in advanced technology nodes are adversely impacted by introduction of more and more corners. These corners typically correspond to worst case scenarios which may occur during the operation of the chip, manifested by extreme operating ranges of Voltage, Temperature. Secondly, manufacturing Process introduces added sources of variability and consequently more corners. For timing closure during chip sign-off, designers have to ensure that the required constraints are met under all possible (critical) PVT corners. One way to deal with the problem is to provide timing margins or guard-bands (GB) during design phase so as to accommodate these variations and ensure safe operation [1]–[3]. Apart from these PVT variations, aging related timing drift due to degradation mechanisms like Bias Temperature Instability (BTI) impose further constraints by requiring additional timing GB [4]. The conventional approach of guardbanding assumes a fixed delay derate factor typically extracted from measurements on simple circuits like ring oscillators [5]. Consequently, all the timing arcs are slowed down by a constant factor during design optimization. Fig. 1(a) shows the schematic representation of the clock period (T) breakdown into typical GB components. Unless modelled properly, such “corner-based” aging margining may either yield chips with no reliability guarantee or chips without expected scaling benefits, as in designs D1 and D2 of Fig. 1(b) respectively. The above-mentioned design methodology tends to ignore the correlation effects which exist inherently due to the circuit topology and the workload patterns. For example in Fig. 2, two back to back inverters are shown to exhibit correlation among them in terms of the workload patterns seen by the PMOS transistors. A 20% activity (uniform) at INV1 input results in 80% activity at the input of INV2 and consequently, higher NBTI degradation for INV2. Therefore, it is important to propagate the “real” workload of the design into individual cell instances to model the degradation accurately.

As an important aspect, workload dependence of aging behaviour in digital designs has been extensively reported in several studies [6]–[8]. However, the major (common) drawback of these reports is the way in which workload has been abstracted by using a very simplified term called “Signal probability” or “Duty cycle” or “Switching activity” as a workaround for accurate abstraction of complex workloads. These terms essentially try to capture the activity at a node in a design by reflecting the probability of a “low” signal in the duration of a given workload or the fraction of the
time duration for which the transistor was effectively under NBTI stress. Even though this “effective” stress consideration is sufficient under uniform activity scenarios, as in Fig. 2, it turns out that non-uniform activity, typically encountered in realistic circuits under application of true workloads, may potentially cause major discrepancy [9], [10].

To this end, a novel complexity reduced simulation approach was introduced in [11]. The proposed Compact Digital Waveform (CDW) based signal representation methodology in [11] grouped consecutive signal regions into segments that feature similar signal characteristics i.e. frequency (f), duty factor (DF) and occupy a duration Δt. Since those segments were based on the numerical values of f and DF, this methodology suffers from limited compressibility and consequently scalability issues for large designs. For example, if two cycles with different DFs (say, 0.3 and 0.55) are separated beyond the set accuracy limit (say, 0.1), those 2 cycles will have to be simulated in cycle-accurate manner. The CDW based approach of workload abstraction was extended for reliability estimation of various CPU blocks for real applications [12]. However, the BTI evaluation framework in [12] was built on transistor level static timing analysis (STA) which is rather time-consuming. Hence the flow is unsuitable for large industrial designs which typically use cell level STA.

In this paper, we try to address the accuracy-runtime trade-off associated with considering real workloads during chip design. We propose a workload dependent reliability aware design optimization flow by utilizing an optimal margining scheme under the influence of NBTI aging. The proposed flow takes into account the relevant correlations in a design by modeling the degradation accurately under specific workload conditions and thus enables achieving the desired PPA goals without severe reliability penalty, as in design D3 of Fig. 1(b).

II. DISCUSSION

A. The proposed workload dependent aging aware design flow

Aging related degradation mechanisms like NBTI are strongly a function of workload/activity pattern at the gate input of the transistor due to the observed recovery effects [4], [13]. In [10], an Adaptive Workload Splitting (AWS) algorithm and a long-term extrapolation methodology were proposed for fast calculation of NBTI degradation due to an arbitrary non-uniform activity pattern with excellent accuracy. It was also demonstrated in [10] at block level that, timing path violations could be minimized using these techniques. We utilize these schemes in this paper, in order to evaluate the PPA gains on a full chip, implemented using industry-standard EDA tool flow.

Fig. 3 depicts the proposed flow highlighting the workload analysis block which is coupled with the traditional design flow. BTI reliability analysis is done by using foundry calibrated Capture Emission Time (CET) maps based on the modeling framework of [14]. The mission profiles or workload traces describe the specific workloads or application data to be used for assessing reliability under various functional loads. These traces are provided to the gate level simulator (GLS) in the form of test vectors. Thus workload at any cell input can be extracted from the GLS following which appropriate transistor level workload propagation is done. Workload dependent aging analysis is set to be carried out during the physical design phase to evaluate timing degradation of each instance due to the workload its individual transistor sees. As a consequence, the parasitic loading and signal transition time are taken into account while translating the device $V_T$ shift into instance specific standard cell timing drift.

In our proposed flow, the physical design phase starts with applying a pessimistic/worst case (WC) delay derate to all the cells in the design during placement stage. With the WC derate, the design constraints (timing in this case) are now tighter and the optimization engine has to over-design, e.g. by use of more high-drive cells in the critical paths to converge, thus claiming more area (power). However, during post-route timing optimization, the accurate instance-based timing derate factors obtained as above are provided as newer constraints. Since these derate values are less pessimistic (better than WC), with the relaxed constraints, the optimization engine can recover some of the area (power) and hence the initial redundancy is partly undone. The reclaim of area (power) is achieved by downsizing some of the high-drive cells, logic remapping, moving instances etc. These benefits are also associated with faster convergence or shorter turn-around-time (TAT). It is to be noted that, the flow is demonstrated here by a timing closed design with engineering change order (ECO) flow in place which offers doing incremental changes in the later stages of the design cycle. Additionally, contrary to [6], this flow alleviates the need for performing an instance based aged library pre-characterization step to do timing checks during STA, since timing derate factors are applied on-the-fly.

B. Simulation setup

In Fig. 4, the schematic of the test design block after implementation is shown with the design parameters and simulation setup used. The design under test (DUT) is a Discrete Cosine Transform (DCT) block inside a jpeg encoder [15]. The circuit accepts a 24-bit raw image file in RGB format to perform image compression. Multiple corners with different supply voltage requirements have been considered.
to do timing signoff at a foundry 28nm technology. Analysis views are constituted so that setup checks are performed on the SS corner with less-than-nominal supply voltage, whereas hold checks are done on FF corner with higher-than-nominal supply voltage. Apart from FEOL corners, BEOL parasitic RC delay was assumed to originate only from typical corner across all analysis views. To account for on-chip-variations (OCV), flat/constant timing derate is applied as a standard approach at 28nm node. Six metal layers (M1-M6) have been used for routing the design.

C. Workload analysis for transistor $V_T$ shift

Workload traces obtained from GLS on post-route design database are fed to the AWS algorithm [10] for estimation of degradation for each transistor in the design. Based on the workload averaging effect, the AWS algorithm essentially splits the waveform into segments which have high toggle rate and segments with lower toggle rate. Then the CET map based compact model (calibrated such that degradation results $\sim 30\text{mV} V_T$ shift @10 years under DC stress at 1.05V and 125C) is used to simulate high toggling segments by using the average workload within the segment and the low toggling segments in cycle-accurate manner. In this way, by adaptively splitting the stress waveform based on the toggling behaviour, fewer simulations are needed by the aging-based compact model to reach the end-point degradation. For more details about this methodology, see [10].

In order to deal with complex workload patterns for a large multi-million-instance design, we propose a hierarchical approach as presented in Fig. 5. The workload can be composed of several applications ordered in a predefined sequence.
Each application can be broken down into smaller packets of workload or so-called “scenarios”. The concept of system scenarios were introduced in the late 1990s in the context of designing cost reduced systems by exploiting run-time workload information. An extensive overview of the state-of-the-art in this domain is available in [16]. In general, the system scenarios may be defined in an N-dimensional objective space where different metrics, assigned to separate objective axes are established, like energy, maximal power, footprint, throughput latency and so on. The meaningful/useful segment mappings (ordered sequences of segments on the execution platform including processor allocation and assignment) may be characterized in terms of their metrics and those who are sufficiently close to each other in terms of the distance (e.g. Manhattan distance) in the N-dimensional objective space will be clustered in the same scenario. Based on profiling of realistic applications, the expected order of scenario instances may then be established to arrive at the final scenario sequence. In the context of our framework, an important metric to be included during the scenario clustering may be the amount of degradation due to aging on the platform for the associated segment mapping. These scenarios are usually ~ millisecond long making them suitable for doing a scenario characterization which is performed by simulating degradation due to the workload of the individual scenario. This scenario characterization needs to be done as a function of initial condition or the amount of degradation to start with. Since the degradation is in general, non-linear in nature w.r.t. time, the same scenario will have temporally different impact, as seen in the schematic of Fig. 6. So the simulated value of degradation is stored in look up tables (LUT) for all the unique scenarios as a function of initial conditions. By utilizing these LUTs and prior information of the occurrence of the scenarios (by workload profiling), it is possible to evaluate the end point degradation using simple analytical methods. This 3-step approach is summarized in Fig. 6.

D. Obtaining the cell-based derate factor

For a fixed transistor $V_T$ shift, the relative rise delay degradation for an inverter cell is dependent on input signal slew and output load capacitance.

Cell timing degradation is a function of transistor $V_T$ shift, the cell input slew rate and cell output load capacitance as shown for an inverter cell in Fig. 7. Moreover, sensitivity of the cell timing shift also depends on the PVT corner being used, e.g. FF/SS, supply voltage and temperature have an impact as well. Therefore, the workload analysis is done on the routed design which allows accurate analysis with parasitic data to capture these signal slew and wire loading effects. Fig. 8 shows an INV cell with these four parameters which have to go into SPICE simulation environment to evaluate the actual timing degradation. Fig. 9 shows the distribution of timing delay degradation for AOI21D1 cell instances in the design under max delay corner (for setup checks) and min delay corner (for hold checks). The max delay corner, on an average shows higher timing degradation than the min delay corner even if they have the same amount of transistor degradation.

The timing derate factor is calculated as:

$$\text{DerateFactor} = \left(1 + \frac{\Delta D}{D_0}\right)$$

where $\Delta D$ is the absolute shift in delay due to degradation, $D_0$ is the initial delay parameter of a standard cell. For example, a standard cell with 5% delay degradation will have a derate factor of 1.05 ($= 1 + 0.05$).

E. Overview of the Binning methods

The combination of degradation and the input design data to be fed into SPICE-based simulator can be humongous
considering a large number of standard cell instances in a design. In order to limit the number of fine-grained SPICE simulations while retaining reasonable accuracy, two different approaches can be followed:

1) Hierarchical Clustering: This method is suitable when the objects or data points in the input variable span a wide range without having any well-defined distribution. Data points are grouped in different clusters if they are sufficiently dissimilar from each other. The measure of dissimilarity is set by a metric called a distance criteria, for example, Euclidean or Manhattan distance. SPICE simulations are only performed around the cluster centroids as shown in Fig. 10(a) for a 2-dimensional parameter space. The hierarchy of clusters is represented in a tree or a dendrogram. The set cutoff threshold can be used as an accuracy parameter to decide the number of clusters.

2) Statistical Design of Experiments (DOE) and Response Surface Modelling (RSM): This approach is suitable when the input variables can be modelled using multicomponent and multivariate Normal distribution. By careful design of experiments (DOEs), the idea is to identify a set of points \( N_{\text{doe}} \) which are representative of the n-dimensional parameter space. The timing characterization (SPICE simulation) needs to be performed only on those \( N_{\text{doe}} \) points. Once the timing characterization is complete, response surface modelling (RSM) is done to get a maximum likelihood fit around the DOEs. The fitted RSM would be used to model the entire parameter space. The RSM model equations can be a very simple linear model to linear model with interaction terms or more complex quadratic model. This is shown below for a 3 variable RSM fit.

\[
y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \tag{2}
\]

\[
y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 \tag{3}
\]

\[
y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{33} x_3^2 \tag{4}
\]

where, \( y \) is the response variable (timing for example), \( x \)'s are input variables (\( V_T \) shift and load slew parameters) and the \( \beta \)'s are the model coefficients. It can be seen that, as the number of variables (number of transistors in a cell) increases

### Table I: Efficacy of hierarchical clustering

| Cell name | Number of PMOS transistors (n) | Number of sim. points with hierarchical clustering | Number of sim. points with uniform binning (\( 7^n \) - with 7 bins) |
|-----------|--------------------------------|-----------------------------------------------|-------------------------------------------------------------|
| NR2D0     | 2                              | 12                                           | 49                                                          |
| OR2D1     | 3                              | 8                                            | 343                                                         |
| AO22D0    | 5                              | 35                                           | 16807                                                       |
| XOR2OPTND1| 6                              | 44                                           | 117649                                                      |

Fig. 9: Timing degradation distribution for AOI21D1 cell under (a) max delay corner (setup) and (b) min delay corner (hold). Even if the cell has same \( V_T \) shifts for transistors, cell timing impact is different in different delay corners.

Fig. 10: (a) Data clusters (5 in no.) with two variables along with their centroid points (b) Dendrogram with Euclidean distance on the X-axis. The number of clusters (7 in this case) can be defined by the set cutoff threshold.

Fig. 11: Brussel DOE approach to find corner points for cell timing characterization.
the number of fitting coefficients also increases substantially making it difficult to converge in stipulated time.

Out of several DOE selection methods, the method of Brussel DOE (BDOE) [17] is particularly useful, because some of the correlation among input variables can be nullified with change of basis using principal component analysis (PCA). Hence the number of fitting parameters in RSM can be minimized. The method of PCA is based on 2 principles (1) Eigenvectors of covariance matrix represent direction of largest variance of data (2) Eigenvalues of covariance matrix represent the magnitude of the variance. The method of BDOE involves building an n-dimensional probability density function (PDF) and selection of \((2n+1) N_{doe}\) points. Fig. 11 represents the possible \(N_{doe}\) points \((5 = 2.2 + 1)\) in a 2-dimensional parameter space. Out of 5 points, 4 are corner points and one centre point. For more details about the methodology, please refer to [17].

\[ F. \textbf{Applying the aging derate} \]

Timing derate factor is applied appropriately to different path groups (reg2reg, in2reg, reg2out) under suitable operating conditions to do setup and hold checks. For instance in Fig. 12, during setup check launch clock path and datapath logic are treated with late delay constraints whereas receive clock path is treated with early delay constraint. The converse is true for hold check.

Once all the instance based derate factors are obtained, they are propagated to the individual instances by using in-house automation scripts prior to the design optimization. Fig. 13 shows the aging derate map for the chips with corner based (typically, a constant/flat 10% timing shift @10 years) derate propagation and an instance based derate propagation. A significant number of instances having a large timing degradation with the corner-based approach, now have to endure a smaller timing degradation factor with the instance-based approach. The effect is that the optimization engine has less effort in meeting timing constraints, e.g. upsizing the cells of the critical paths, logic remapping, moving instances etc. This results in reclaiming some of the area (and associated power benefits, see Fig. 15), faster convergence and shorter TAT.

III. \textbf{SIMULATION RESULTS}

\[ \text{A. IR drop profile} \]

Dynamic vector less IR drop simulations were also performed to compare three design approaches, (1) reference with no aging margin applied (2) corner design with 10% timing margin and (3) workload (WL) aware design. The heat map in Fig. 14 shows marginally cooler hotspots in WL aware design compared to a corner-based design, even though peak IR drop is \(\sim 20\text{mV}\) lower than corner case.

Dynamic vector less IR drop simulations were also performed to compare three design approaches, (1) reference with no aging margin applied (2) corner design with 10% timing margin and (3) workload (WL) aware design. The heat map in Fig. 14 shows marginally cooler hotspots in WL aware design compared to corner based design, even though the peak IR drop is significantly improved (\(\sim 20\text{mV}\)) for WL aware case. This can be attributed to the use of less high-drive strength cells/buffers in WL aware design.

\[ \text{B. PPA Analysis} \]

The relative PPA results are compared for benchmarking our design as shown in Fig. 15. Performance is defined by the target frequency, calculated from the slowest timing path. Area is defined as the standard cell area, which is obtained by multiplying the total core area (almost fixed for all 3 designs) with the final utilization density. Under iso-performance condition, WL aware design clearly shows better power and area.
Marie Skłodowska-Curie grant agreement No. 843546. Horizon 2020 research and innovation programme under the iso-performance condition using a foundry 28nm technology.

In terms of area and power (energy) was demonstrated under timing closure. Making more routing resources available during final stages of timing closure. This helps reducing routing congestion and density is reduced in WL aware design compared to the corner case.

Overhead reduction compared to a 10% corner based design. It results in as much as 42.8% area and 38.4% power overhead reduction compared to a 10% corner based design. It is also to be noted in Fig. 16 that, the final achieved utilization density is reduced in WL aware design compared to the corner based design. This helps reducing routing congestion and making more routing resources available during final stages of timing closure.

IV. CONCLUSION

In conclusion, we proposed a workload dependent aging aware system optimization flow from logic design till final implementation and signoff stage. It is able to handle applications with complex activity patterns with tractable runtime and reasonable accuracy. Using this flow, ~40% overhead reduction in terms of area and power (energy) was demonstrated under iso-performance condition using a foundry 28nm technology.

ACKNOWLEDGMENT

This project received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 843546.

REFERENCES

[1] M. S. Gupta, J. A. Rivers, P. Bose, G. Wei, and D. Brooks, “Tribeca: Design for pvt variations with local recovery and fine-grained adaptation,” in 2009 42nd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO), 2009, pp. 435–446.

[2] V. J. Reddi, S. Kane, W. Kim, S. Campanoni, M. D. Smith, G. Wei, and D. Brooks, “Voltage smoothing: Characterizing and mitigating voltage noise in production processors via software-guided thread scheduling,” in 2010 43rd Annual IEEE/ACM International Symposium on Microarchitecture, 2010, pp. 77–88.

[3] S. Jain, S. Khare, S. Yada, V. Ambili, P. Salihundam, S. Ramani, S. Muthukumar, M. Srinivasan, A. Kumar, S. K. Gb, R. Ramanarayanan, V. Erraguntla, J. Howard, S. Vangal, S. Dighe, G. Ruhl, P. Aseron, H. Wilson, N. Borkar, V. De, and S. Borkar, “A 280mV-to-1.2v wide-operating-range ia-32 processor in 32nm cmos,” in 2012 IEEE International Solid-State Circuits Conference, 2012, pp. 66–68.

[4] A. T. Krishnan, F. Cano, C. Chancellor, V. Reddy, Zhangfen Qi, P. Jain, J. Carulli, J. Masin, S. Zuhoski, S. Krishnan, and J. Ondruszek, “Product drift from nbti: Guardbanding, circuit and statistical effects,” in 2010 International Electron Devices Meeting, 2010, pp. 4.3.1–4.3.4.

[5] V. Reddy, J. Carulli, A. Krishnan, W. Bosch, and B. Burgess, “Impact of negative bias temperature instability on product parametric drift,” in 2004 International Conference on Test, 2004, pp. 148–155.

[6] E. Mintorno, V. Chandra, D. Pietromonaco, R. Aiiken, and R. W. Dutton, “Workload dependent nbti and pbt analysis for a sub-45nm commercial microprocessor,” in 2013 IEEE International Reliability Physics Symposium (IRPS), 2013, pp. 3A.1.1–3A.1.6.

[7] A. Sivadasan, S. Mhira, A. Notin, A. Benhassain, V. Huard, E. Maurin, F. Cacho, L. Anghel, and A. Bravai, “Architecture- and workload-dependent digital failure rate,” in 2017 IEEE International Reliability Physics Symposium (IRPS), 2017, pp. CR–8.1–CR–8.4.

[8] S. Morita, S. Bian, M. Shintani, M. Hiromoto, and T. Sato, “Efficient worst-case timing analysis of critical-path delay under workload-dependent aging degradation,” in 2018 23rd Asia and South Pacific Design Automation Conference (ASP-DAC), 2018, pp. 631–636.

[9] M. Chen, V. Reddy, J. Carulli, S. Krishnan, V. Rentala, V. Srinivasan, and Y. Cao, “A tdc-based test platform for dynamic circuit aging characterization,” in 2011 International Reliability Physics Symposium, 2011, pp. 2B.2.1–2B.2.5.

[10] S. Mishra, P. Weckx, J. Y. Lin, B. Kaczer, D. Linten, A. Spessot, and F. Catthoor, “Fast accurate methodology for aging incorporation in circuits using adaptive waveform splitting (aws),” in 2020 IEEE International Reliability Physics Symposium (IRPS), 2020, pp. 1–5.

[11] D. Rodopoulos, P. Weckx, M. Noltsis, F. Catthoor, and D. Soudris, “Atomistic pseudo-transient bti simulation with inherent workload memory,” IEEE Transactions on Device and Materials Reliability, vol. 14, no. 2, pp. 704–714, 2014.

[12] D. Stamoulis, S. Corbetta, D. Rodopoulos, P. Weckx, P. Debacker, B. H. Meyer, B. Kaczer, P. Raghavan, D. Soudris, F. Catthoor, and Z. Zilic, “Capturing true workload dependency of bti-induced degradation in cpu components,” in 2016 International Great Lakes Symposium on VLSI (GLSVLSI), 2016, pp. 373–376.

[13] K. T. Lee, W. Kang, E. A. Chung, G. Kim, H. Shin, H. Lee, H. Kim, M. Choe, N. I. Lee, A. Patel, J. Park, and J. Park, “Technology scaling on high-k amp; metal-gate finfet bti reliability,” in Proc. IEEE Int. Rel. Phys. Symp., Apr. 2013, pp. 2D.1.1–2D.1.4, doi:10.1109/IRPS.2013.6531956.

[14] T. Grasser, P. Wagner, H. Reisinger, T. Aichinger, G. Pobegen, M. Nelliebel, and B. Kaczer, “Analytic modeling of the bias temperature instability using capture/emission time maps,” in 2013 IEEE International Reliability Physics Symposium (IRPS), 2013, pp. CR–8.1–CR–8.4.

[15] V. Dutton, “Workload dependent nbti and pbti analysis for a sub-45nm commercial microprocessor,” in 2013 IEEE International Reliability Physics Symposium (IRPS), 2013, pp. 3A.1.1–3A.1.6.

[16] A. Sivadasan, S. Mhira, A. Notin, A. Benhassain, V. Huard, E. Maurin, F. Cacho, L. Anghel, and A. Bravai, “Architecture- and workload-dependent digital failure rate,” in 2017 IEEE International Reliability Physics Symposium (IRPS), 2017, pp. CR–8.1–CR–8.4.

[17] A. T. Krishnan, F. Cano, C. Chancellor, V. Reddy, Zhangfen Qi, P. Jain, J. Carulli, J. Masin, S. Zuhoski, S. Krishnan, and J. Ondruszek, “Product drift from nbti: Guardbanding, circuit and statistical effects,” in 2010 International Electron Devices Meeting, 2010, pp. 4.3.1–4.3.4.

Fig. 15: Under iso-performance condition, the area and energy/power overhead are reduced by ~40% in WL aware design when compared to the 10% corner design.

Fig. 16: The achieved utilization density for various design paradigms showing reduction in WL aware case compared to a corner case.