A Cross-Layer Framework for Designing and Optimizing Deeply-Scaled FinFET-Based SRAM Cells under Process Variations

Alireza Shafaei, Shuang Chen, Yanzhi Wang, and Massoud Pedram
Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089
{shafaeib, shuangc, yanzhiwa, pedram}@usc.edu

***Abstract*—A cross-layer framework (spanning device and circuit levels) is presented for designing robust and energy-efficient SRAM cells, made of deeply-scaled FinFET devices. In particular, 7nm FinFET devices are designed and simulated by using Synopsys TCAD tool suite, Sentaurus. Next, 6T and 8T SRAM cells, which are composed of these devices, are designed and optimized. To enhance the cell stability and reduce leakage energy consumption, the dual (i.e., front and back) gate control feature of FinFETs is exploited. This is, however, done without requiring any external signal to drive the back gates of the FinFET devices. Subsequently, the effect of process variations on the aforesaid SRAMs is investigated and steps are presented to protect the cells against these variations. More precisely, the SRAM cells are first designed to minimize the expected energy consumption (per clock cycle) subject to the non-destructive read and successful write requirements under worst-case process corner conditions. These SRAM cells, which are overly pessimistic, are then refined by selectively adjusting some transistor sizes, which in turn reduces the expected energy consumption while ensuring that the parametric yield of the cells remains above some pre-specified threshold. To do this efficiently, an analytical method for estimating the yield of SRAM cells under process variations is also presented and integrated in the refinement procedure. A dual-gate controlled 6T SRAM cell operating at 324mV (in the near-threshold supply regime) is finally presented as a high-yield and energy-efficient memory cell in the 7nm FinFET technology.**

I. Introduction

On-chip SRAMs occupy a large portion of the chip area [1] [2]. Moreover, SRAM cells have relatively low activity factors, and hence long idle periods. As a result, the leakage power of SRAMs not only dominates the power consumption of the memory circuit but also becomes a major component of the overall chip power consumption [3]. An effective solution to reduce the leakage power without significantly sacrificing performance is to scale down the supply voltage to an operating point (typically in the near-threshold region) where energy consumption is minimized [4]. Accordingly, low voltage SRAMs are crucial for energy-efficient designs.

On the other hand, SRAM cells are required to have a small layout footprint in order to improve the memory density. Consequently, SRAM cells with ratioed designs (e.g., 6T SRAM cell) and minimum-size transistors are preferred. However, such conditions make SRAMs very sensitive to device mismatches which are exacerbated by process variations. Therefore, SRAM cells, if not designed properly, may fail to function. Since process variations are inevitable especially in deeply-scaled (i.e., sub 10nm) technologies, robust operation of SRAMs should be ensured during the design time.

In order to improve the robustness (i.e., read/write margins) of SRAM cells, FinFET-based SRAMs have been proposed [5] [6]. FinFET devices are currently one of the most effective ways to reduce short channel effects. This is due to the improved (three-dimensional) gate control over the channel, and less control by the source and drain terminals [7]. Moreover, FinFETs exhibit higher immunity to random variations and soft errors, superior scalability, and are recognized as the technology-of-choice beyond the 10nm regime [8].

The focus of this paper is thus on the cross-layer (device- and circuit-level) design of energy-efficient FinFET-based SRAMs which can tolerate process variations in order to achieve high yield. To this end, FinFET devices are designed for a 7nm process using the Synopsys Technology Computer-Aided Design (TCAD) tool suite, Sentaurus [9], which can generate accurate results with device simulators based on physics-driven models. Moreover, SPICE-compatible Verilog-A models for 7nm FinFET devices are extracted from the device simulations for performing fast circuit-level simulations.

The stability of an SRAM cell may be improved by the appropriate usage of *dual-gate control* of FinFET devices [5], and the leakage power consumption can be significantly reduced. More accurately, the front gate and the back gate of a FinFET device can be separately controlled with one gate controlling the on/off state of the device and the other gate adjusting the threshold voltage. We apply this dual-gate control[1] to conventional 6T and 8T [10] SRAM cells using 7nm FinFET devices without requirements of external signals, and compare their SNM, layout area, and leakage power.

Although FinFETs have reduced variability compared with the bulk CMOS counterpart, they still suffer from *line edge roughness* as well as *gate oxide thickness* variations. In order to protect FinFET-based SRAMs against these process variations and increase energy efficiency, a flow for designing high yield and energy-efficient SRAM cells is proposed. This design flow is a joint optimization of the supply voltage and device parameters of SRAM cells. For a given supply voltage, an optimal SRAM cell configuration is initially derived for the worst-case corner of process variation. This solution, despite offering a very high yield, is overly pessimistic, and hence it is refined later by applying transistor-level size adjustments to further reduce the expected energy consumption while satisfying yield requirements. Accordingly, an analytical method for measuring the yield of SRAM cells under process variations is also proposed. Using this framework, we present a dual-gate controlled 6T SRAM cell operating at 324mV as the high yield and energy-efficient memory cell in our 7nm FinFET process.

The rest of the paper is organized as follows. Section II introduces our 7nm FinFET devices. Section III presents the dual-gate control schemes applied to 6T and 8T SRAM cells. The proposed process variation tolerant design flow is discussed in Section IV, followed by simulation results in Section V. Finally, Section VI concludes the paper.

[1] We intentionally avoid using the term “independent gate control” as it implies the usage of a separate signal for controlling the back gate.

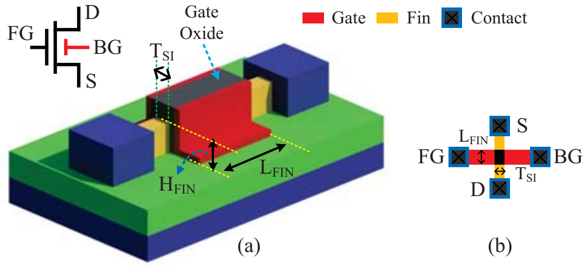

Fig. 1. (a) The structure and (b) layout of a dual-gate single-fin FinFET device. FG, BG, S, and D denote front gate, back gate, source and drain terminals, respectively.

TABLE I. GEOMETRIES OF 7NM FINFET DEVICES.

Parameter	Value (nm)	Comment
L_{FIN}	$2\lambda = 7$	Fin or gate length, also denoted by L
T_{SI}	5.5	Fin width, also known as Silicon thickness
H_{FIN}	14	Fin height
P_{FIN}	$2\lambda + T_{SI} = 12.5$	Fin pitch using the spacer lithography
t_{ox}	1.3	Oxide thickness

II. 7NM FINFET DEVICES

The structure and layout of a dual-gate FinFET device are shown in Figures 1(a) and 1(b), respectively. The main component is the *fin* which provides the channel for conducting current when the device is turned on. This vertical fin is surrounded by two gate terminals, the front gate and the back gate, where the front gate is used to turn on/off the device and the back gate can adjust the threshold voltage. Specifically, by connecting the back gate of an N-type (P-type) FinFET to a low (high) voltage such as Gnd (V_{dd}), the threshold voltage will increase when the front gate is turned on.

Major process-related FinFET geometries for 7nm technology are reported in Table I. Due to the lack of industrial data for such deeply-scaled FinFETs, our devices are designed and optimized using the Synopsys TCAD Sentaurus [9]. The supply voltage is 0.45V (0.3V) for super-threshold (near-threshold) operation, while the threshold values of our FinFET devices are between 0.2V and 0.25V.

III. BASELINE SRAM CELLS

The usage of deeply-scaled devices especially under low voltage operations makes SRAMs more vulnerable to process variations. Accordingly, we not only replace planar CMOS devices with FinFETs which have lower variability, but also take advantage of the dual-gate control to further improve the stability of SRAM cells. In order to avoid the cost of generating and routing extra signals, we will only use internal signals to the SRAM cell for back gate connections as described next.

A. Dual-Gate Control for Improving the Cell Stability

Read stability may be enhanced by strengthening the pull-down transistor by increasing the number of fins, and/or by weakening the access transistor during read operation using the dual-gate control. For the latter case, the back gate of the access transistor is connected to the corresponding storage node (i.e., $M5$ to Q, and $M6$ to $\overline{Q}$) [5]. Hence, during read operation the access transistor connected to bit '0', which is where the actual read happens, becomes weaker, thereby enhancing the read stability. For improving the write-ability, a weaker pull-up transistor compared with the access transistor is needed. For this purpose, we connect the back gates of P-type pull-up transistors to V_{dd}. Consequently, the threshold voltage of pull-up transistors will increase which is beneficial in terms of write margin improvement, and leakage power reduction.

The aforesaid schemes are applied to the conventional 6T cell. The 8T cell, however, relaxes the read stability requirement, which means there is no need to weaken the access transistor during read operation. In other words, back gate connections of access transistors are removed in the 8T cell, but still we adopt dual-gate control to weaken pull-ups. Moreover, to further reduce the leakage power of the 8T cell, the back gate of $M7$ is connected to the read word-line (RWL). Hence, during the standby mode when $RWL = 0$, the OFF current of $M7$ is (exponentially) decreased. Schematics of 6T and 8T SRAM cells after applying the discussed dual-gate controls are illustrated in Fig. 2.

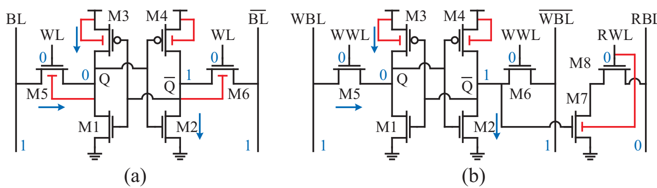

Fig. 2. Circuits of (a) standard 6T, and (b) 8T [10] SRAM cells. Red parts show back gate connections when dual-gate control is employed, whereas blue parts highlight subthreshold leakage paths in an idle SRAM cell (storing 0).

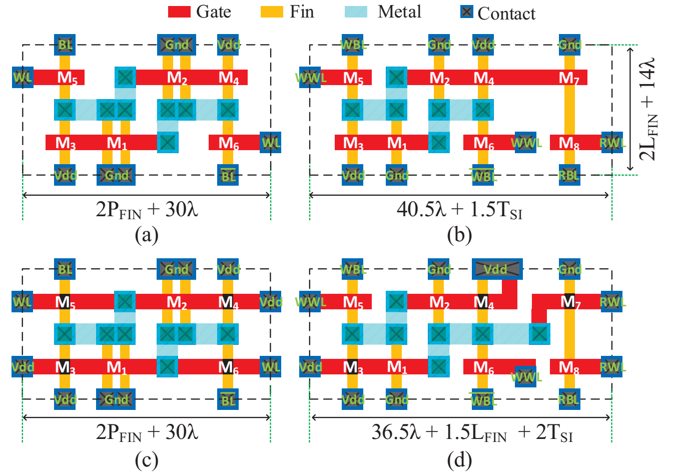

Fig. 3. Layouts of (a) 6T2-SG, (b) 8T-SG, (c) 6T2-DG, and (d) 8T-DG SRAM cells.

B. Comparison Results

Table II compares five SRAM cells in terms of SNM, layout area, and leakage power consumption. Process variation is not accounted for in this table. For the 6T SRAM cell, all transistors are single-fin except for pull downs which may have one (6T1) or two (6T2) fins each. All transistors in the 8T cell are single-fin. Furthermore, SG denotes an SRAM cell with all FinFETs in the single-gate mode (i.e., front and back gates are connected together), whereas DG represents a cell with dual-gate control added. Additionally, because of weak pull-down transistors, the 6T1-SG cell does not work properly in our 7nm technology.

Layouts of 6T2-SG, 6T2-DG, 8T-SG and 8T-DG SRAM cells are illustrated in Fig. 3. Layout of 6T1-DG is similar to that of 6T2-DG except that $M1$ and $M2$ are drawn with one fin, reducing the width to 30λ. After applying dual-gate control, the area of 6T SRAM cell does not change, whereas the area difference in the 8T cell mainly depends on the gate length. 6T1-DG SRAM has the smallest layout area, and using

TABLE II. COMPARISON OF SRAM CELLS IN 7NM FINFET. SNM AND LEAKAGE POWER ARE CALCULATED FOR SUPER-THRESHOLD REGIME.

SRAM Cell	Area (nm^2)	SNM (mV)	Leakage (nW)
6T2-SG	8,190	55	53.63
6T1-DG	6,615	82	39.59
6T2-DG	8,190	79.5	38.03
8T-SG	9,450	144.5	41.24
8T-DG	9,403	109.1	38.45

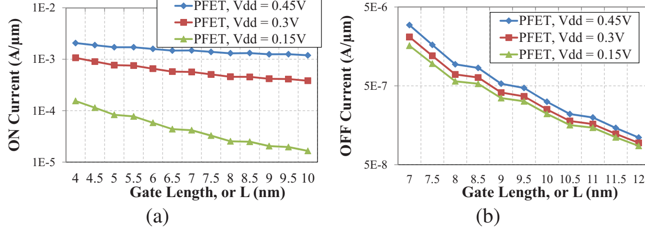

Fig. 4. The effect of process variations on 7nm P-type FinFETs: (a) ON and (b) OFF currents as a function of L. N-type FinFETs also have the same trend. Vertical axes in both figures are in logarithmic (base 10) scale.

our 7nm FinFET devices the areas of 6T2-DG, and 8T-DG SRAM cells are 24%, and 42% larger than the area of 6T1-DG cell, respectively.

The *static noise margin* (SNM) quantifies the amount of voltage noise required at the internal nodes of a bitcell to flip the SRAM cell's contents. As can be seen in Table II, dual-gate control can effectively improve the SNM of 6T SRAM cell. For 8T SRAM cell, SNM after applying dual-gate control will decrease. However, the degraded SNM value is still high, and in addition, leakage power is reduced.

IV. ROBUST SRAM CELL DESIGN UNDER PROCESS VARIATIONS

The main target of this paper is the design optimization of SRAM cells with minimum-size transistors based on FinFET devices with extremely small geometries (7nm) under low voltage operations (e.g. 0.3V for near-threshold computing). Hence, in order to reduce the process variation-induced failures (and achieve a high yield), a cross-layer framework for optimizing SRAM cells while tolerating various sources of process variation is needed.

A. Process Variations in FinFET Technology

FinFET devices are more immune to process variations compared with planar CMOS counterparts. The main reason is the undoped channel of FinFET devices which eliminates the *random dopant fluctuation*. However, FinFETs still suffer from *line edge roughness* (LER), which causes variations of the (effective) channel length L, as well as *oxide thickness variations*, which cause variations of the oxide thickness t_{ox}. Both types of variations will affect the threshold voltage and the subthreshold slope of FinFET devices. The effect of process variations becomes much more significant for deeply-scaled FinFET technologies. For example, reference work [11] and [12] predict that the standard deviation of L and the effect of LER variations are 0.8nm and more than 80mV variations in threshold voltage V_{th}, respectively, in deeply-scaled FinFETs.

In order to investigate the effect of process variations on deeply-scaled FinFET technology, we measured ON and OFF currents of our 7nm devices for various values of gate length (L) and gate oxide thickness (t_{ox}) using TCAD Sentaurus. Results for super-threshold ($V_{dd} = 0.45V$), near-threshold ($V_{dd} = 0.3V$), and sub-threshold ($V_{dd} = 0.15V$) regions are shown in Fig. 4 with a single fin and different L values. Two observations can be seen: (i) The effect of process variations is more significant in subthreshold regime compared with near-threshold and super-threshold regimes, because in the former case the ON current is (approximately) exponentially dependent on the threshold voltage and/or subthreshold slope, which is affected by LER; (ii) the effect of process variations is more significant on OFF currents compared with ON currents (similar observation also for bulk CMOS devices [13].) Simulation results on gate oxide thickness variations show that it has a similar, but much less significant, effect than LER, and thus is omitted due to space limitation.

In this paper, we assume Gaussian variation on L and t_{ox} of a single fin with standard deviations of $\sigma_L = 0.8nm$ and $\sigma_t = 5\%$, respectively. Then the ON and OFF currents approximately satisfy a log-normal distribution in the sub- and near-threshold regimes. Please note that in the proposed optimization framework, we use an analytical method to estimate the yield of SRAM cells under process variations, instead of performing Monte Carlo simulation or importance sampling on the SRAM cells. This is because the latter method necessitates generating Verilog-A models for different devices (NFET and PNET, with and without dual-gate control) at different (and fine-grained) L and t_{ox} combinations using TCAD Sentaurus, the computation complexity of which is prohibitive.

B. Optimization Framework

We intend to find the supply voltage level along with transistor sizings and device-level parameters for 6T and 8T SRAM cells (with dual-gate control) in order to minimize the (expected) cell energy consumption while satisfying a certain yield constraint under process variations. Activity factors of SRAM cells are relatively low, e.g. 2% as reported in [14], which means SRAM cells are idle for most of the time. Accordingly, the leakage power becomes the dominant component of the power consumption of SRAMs [3]. Hence, without loss of generality, our objective is to minimize the expectation of the leakage energy consumption (in each clock cycle) of an SRAM cell such that read stability and write-ability requirements under process variations are met.

Optimization variables of the optimal design problem include the supply voltage (V_{dd}), the gate length of the pull-up transistor (L_{PU}), as well as the number of fins of the access (N_{AC}) and pull-down (N_{PD}) transistors. For the pull-up transistor a single-fin device (i.e., $N_{PU} = 1$) is assumed, because a weaker pull-up is desired to enhance SRAM yield and reduce leakage. On the other hand, N_{AC} and N_{PD} may be greater than one and should be judiciously optimized because we need to satisfy the SRAM yield constraints under process variations. Moreover, in order to simplify the fabrication process, standard-length devices are used for access and pull-down transistors. On the other hand, we use a larger length L_{PU} for the pull-up transistor since it can (i) further weaken the pull-up transistor, thereby enhancing the yield and reducing leakage power consumption, and (ii) mitigate the effect of process variation on the pull-up transistor [15]. A combination of L_{PU}, N_{AC}, and N_{PD} creates an SRAM cell configuration.

The motivation of the joint optimization of V_{dd} and SRAM cell configuration is as follows. When V_{dd} decreases to the near/sub-threshold regime, we have the following three effects: (i) both leakage power and dynamic energy consumptions reduce [4]; (ii) the circuit delay increases [4]; and (iii) the access and pull-down transistors need to be heavily sized up in order to satisfy the yield constraint since the process variation

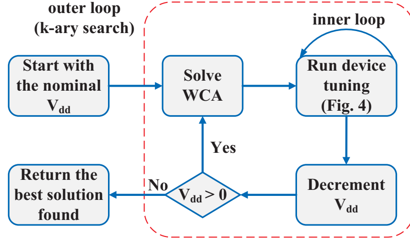

Fig. 5. The proposed flow for designing high yield and energy-efficient SRAMs.

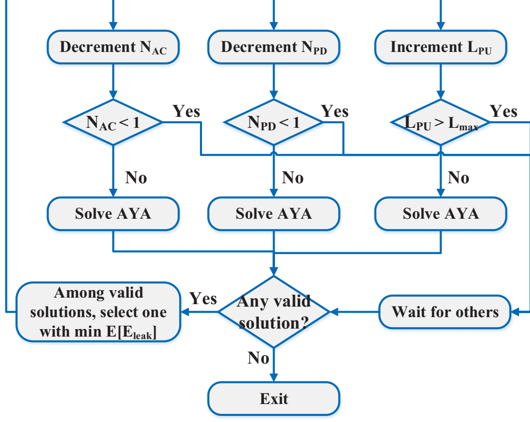

Fig. 6. Design flow of the device tuning step.

is more phenomenal when V_{dd} is low. The first effect will reduce energy consumption whereas the second and third will have the opposite effect. Therefore, it is critical to find the best-suited V_{dd} level and corresponding SRAM cell configuration through a joint optimization framework.

The proposed flow for optimizing process variation tolerant SRAM cells is shown in Fig. 5. The optimization flow is composed of two nested loops where the outer loop iterates over V_{dd} values in a k-ary search fashion. For each V_{dd}, the optimal SRAM cell configuration for the worst-case corner of process variation is initially calculated. We will refer to this problem as the *worst-case analysis* (WCA). The solution of the WCA problem satisfies yield constraints, but results in an overly pessimistic SRAM cell design. Hence, this solution is refined later using the inner loop which contains the *device tuning* step described next.

Values of L_{PU}, N_{AC}, and N_{PD} are refined in a fine-grained manner as shown in Fig. 6 in order to further reduce the expected leakage energy while satisfying a yield constraint. The yield constraint is estimated from (i) the yield requirement of the whole SRAM array and (ii) the number of redundant rows/columns and error correction mechanism in the SRAM array. Then, the yield of the newly derived SRAM cell configuration is analytically measured under process variations. This step is called *analytical yield analysis* (AYA) which enables us to perform fast yield analysis as opposed to time-consuming sampling-based methods including Monte Carlo sampling or importance sampling [4] [16]. If we can find a solution with lower expected energy consumption and satisfying the yield constraint, we will accept it as the new solution. However, when device parameters cannot be tuned anymore (because of hitting the limits), or a solution satisfying the yield constraint cannot be found, we exit from the inner loop. The solution that minimizes the objective function is finally returned as the optimal high-yield and energy-efficient SRAM cell configuration.

More details are discussed below.

WCA formulation. We use 6T SRAM as an example. The purpose of the WCA problem is to provide an initial solution, and to guide the optimization process in the right direction. This problem is formulated as follows.
Find the L_{PU}, N_{AC}, and N_{PD} values.
Minimize the expectation of leakage energy consumption:

$$\mathbb{E}[E_{leak}] = \mathbb{E}[V_{dd} \cdot I_{leak} \cdot t_{clk}] = V_{dd} \cdot \mathbb{E}[I_{leak} \cdot t_{clk}], \quad (1)$$

where t_{clk} represents the clock cycle of the SRAM array, and the leakage current of the SRAM cell, according to subthreshold leakage paths shown in Fig. 2, is estimated as

$$I_{leak} = I_{OFF,PU} + N_{AC} \cdot I_{OFF,AC} + N_{PD} \cdot I_{OFF,PD}, \quad (2)$$

where I_{OFF} denotes the OFF current of the corresponding single-fin transistor.

We use the following observations to simplify the estimation of $\mathbb{E}[I_{leak} \cdot t_{clk}]$:

1) t_{clk} is dominated by peripheral devices [17], and hence is not sensitive to process variations of SRAM cells.
2) $I_{OFF,PU} << N_{AC} \cdot I_{OFF,AC}$ or $N_{PD} \cdot I_{OFF,PD}$ because the PU transistor has (i) a single fin and (ii) longer length L_{PU}.
3) The PU transistor is more immune to process variations since it has a longer length. Hence the expected value of $I_{OFF,PU}$ is closer to the nominal value.

So we have the following approximation:

$$\mathbb{E}[I_{leak} \cdot t_{clk}] \approx \mathbb{E}[I_{leak}] \cdot t_{clk}, \quad (3)$$

where

$$\begin{aligned} \mathbb{E}[I_{leak}] \approx & I_{OFF,PU}(V_{dd}, L_{PU}) + \mathbb{E}[N_{AC} \cdot I_{OFF,AC}] \\ & + \mathbb{E}[N_{PD} \cdot I_{OFF,PD}] \\ \approx & I_{OFF,PU}(V_{dd}, L_{PU}) + N_{AC} \cdot \beta_{AC} \cdot I_{OFF,AC}(V_{dd}) \\ & + N_{PD} \cdot \beta_{PD} \cdot I_{OFF,PD}(V_{dd}) \quad (4) \end{aligned}$$

where $I_{OFF,PU}(V_{dd}, L_{PU})$, $I_{OFF,AC}(V_{dd})$, and $I_{OFF,PD}(V_{dd})$ are all nominal values without considering process variation, and β_{AC} and β_{PD} are coefficients accounting for the effect of process variations **on a single fin with standard length**, and defined as the ratio of the expected value of the OFF current to its nominal value. In this way, we separate the effect of process variations from the optimization of the L_{PU}, N_{AC}, and N_{PD} values, thereby significantly simplifying the optimization procedure.

The WCA problem is subject to the following constraints:

$$\begin{aligned} & N_{PD} \cdot I_{ON,PD}(V_{dd}, L + 6\sigma_L, t_{ox} + 6\sigma_t) > \\ & \quad \alpha_r \cdot N_{AC} \cdot I_{ON,AC}(V_{dd}, L - 6\sigma_L, t_{ox} - 6\sigma_t), \quad (5) \\ & N_{AC} \cdot I_{ON,AC}(V_{dd}, L + 6\sigma_L, t_{ox} + 6\sigma_t) > \\ & \quad \alpha_w \cdot I_{ON,PU}(V_{dd}, L_{PU} - 6\sigma_L, t_{ox} - 6\sigma_t), \quad (6) \\ & N_{PD}, N_{AC} \in \{1, 2, \cdots, N_{max}\}, \quad (7) \\ & L_{PU} \in \{L_{min}, L_{min} + s_L, \cdots, L_{max}\}, \quad (8) \end{aligned}$$

where I_{ON} denotes the ON current of the corresponding transistor with a single fin in the 6T SRAM cell. Moreover, α_r (α_w) represents the strength ratio of PD (AC) to AC (PU) transistors such that the read stability (write-ability) constraint is met (they also account for leakage currents that weaken the corresponding stability constraint); N_{max} is the maximum allowable number of fins in a FinFET device; L_{min} (L_{max}) is the minimum (maximum) allowable gate length in a P-type FinFET device; s_L denotes the step value for incrementing L_{PU}. Constraints (5) and (6) are designed for the worst-case corner of process variations. Please recall that the ON current of a FinFET device will decrease by increasing L or t_{ox}.

Therefore, for the worst-case of the read stability, we assume that the ON current of the PD transistor is at the lowest corner whereas the ON current of the AC transistor is at the highest corner. The constraint on write-ability is obtained similarly.

To accelerate the optimization process, we store the I_{ON}, I_{OFF}, and t_{clk} values in lookup tables (LUT). In fact, due to the complicated gate control mechanism over the channel, the current of a deeply-scaled FinFET device is typically expressed with LUTs instead of analytical models [18]. For this purpose, advanced device simulators such as TCAD tools [9] are used. However, these device simulators are too expensive in terms of runtime for simulating SRAM cells. As a result, we generate ON and OFF currents of single-fin NFET and PFET for a limited combination of V_{dd}, L and t_{ox} values, and store them into corresponding LUTs. By using multivariate interpolation methods (which are used for interpolating functions with more than one input variables), we can perform rapid circuit-level simulations as well as explore a wider region of the design space in order to enhance the performance and characteristics of the final solution.

AYA formulation. The analytical yield analysis (AYA) problem is formulated as follows. *Given* a supply voltage V_{dd}, and an SRAM configuration N_{AC}, N_{PD}, and L_{PU}, we estimate the yield of the SRAM cell under process variations (with standard deviations given in Section IV-A). An SRAM cell is *functional* if (i) it satisfies the read stability and write-ability requirements, and (ii) the SNM in the hold state is higher than a pre-defined value. Other requirements are flexible to be added to this analytical framework.

We use the read stability requirement as an example, whereas the write-ability requirement can be estimated similarly. We first derive the distribution of the ON current of the PD transistor with N_{PD} fins, denoted by $f(I_{PD,total})$, under process variations, through the following steps: (i) derive the ON current distribution of a single fin, based on the above-mentioned $I_{ON,PD}$ LUTs and process variation parameters given in Section IV-A, and (ii) derive $f(I_{PD,total})$ using the principle of probability distribution of random variables summations [19], which is the convolution of probability distributions of individual random variables. Similarly we derive the ON current distribution of the AC transistor with N_{AC} fins, denoted by $f(I_{AC,total})$, under process variations.

Then we estimate the probability that the read stability requirement is satisfied using the following equation:

$$\iint f(I_{PD,total}) f(I_{AC,total}) \cdot \mathbf{I}[I_{PD,total} > \alpha_r \cdot I_{AC,total}] \cdot \mathrm{d}I_{PD,total} \mathrm{d}I_{AC,total}, \quad (9)$$

where $\mathbf{I}[x]$ is the *indicator function*, which equals to one if the boolean variable x is true.

For the hold SNM estimation, we adopt an analytical method to derive an effective estimation, with a brief procedure discussed as follows: First we define the *imbalance factor* of each of the two cross-coupled inverters in the SRAM cell as the ON current ratio of the PD transistor and PU transistor. Next we derive the distribution of the imbalance factor based on the ON current distributions. We (approximately) derive the *voltage transfer characteristics* (VTC) of the two cross-coupled inverters based on the imbalance factor. For each pair of imbalance factors (of the two cross-coupled inverters), we estimate the SNM graphically by plotting the butterfly plot and calculating the length of the square fitted between the VTCs and having the longest diagonal. We estimate the probability that the SNM is higher than the pre-defined value. In this way we can perform effective SNM estimation using only the ON current LUTs. Details are omitted due to space limitation.

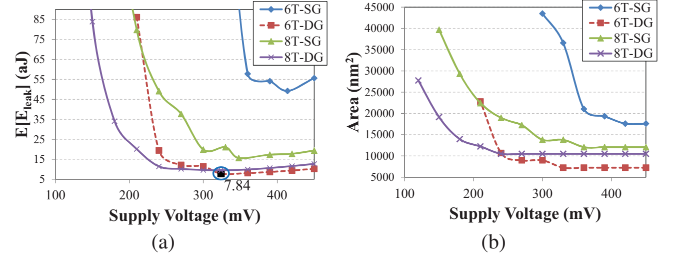

Fig. 7. (a) Optimal expected leakage energy consumption of each SRAM cell at different V_{dd} values. Best result (dual-gate controlled 6T SRAM cell operating at 324mV) is highlighted in the figure. (b) Layout area of the corresponding SRAM cell in (a).

TABLE IV. CACHE CONFIGURATION.

Parameter	Value	Parameter	Value	Parameter	Value
Cache size	4MB	Associativity	8	Number of banks	4
Block size	64B	Bus width	512	Read/write ports	1

By assuming that the read stability, write-ability, and SNM are independent of each other, we can estimate the yield of the SRAM with given configuration at given supply voltage V_{dd}.

V. SIMULATION RESULTS

We implemented the proposed design flow in C++. For all simulations, we used $N_{max} = 20$, $L_{min} = 4$nm, $L_{max} = 11.5$nm, $s_L = 0.5$nm, and we derived the values $\alpha_r = \alpha_w = 1.25$ and $\beta_{AC} = \beta_{PD} = 3$. For architecture-level simulations, we adopted an L3 cache memory with configurations given in Table IV. V_{dd} is varied from 450mV to 102mV by steps of 6mV. Moreover, SRAM cell configurations are shown as a triplet (N_{PD}, N_{AC}, L_{PU}), and the `6T-SG` SRAM operating at $V_{dd} = 450$mV is considered as the baseline cell in this section.

Optimal (i.e., with minimal $\mathbb{E}[E_{leak}]$) SRAM cell configurations for different combinations of V_{dd} and SRAM cell structures are reported in Table III. As can be seen, the overall optimal result is achieved by the (1,1,10) 6T SRAM cell equipped with the proposed dual-gate control scheme (`6T-DG`), operating at 324mV V_{dd} level, which is in the near-threshold regime. In fact, the table compares the effectiveness of the 8T SRAM design as a circuit-level solution, and the dual-gate control as a unique feature of FinFET devices, in enhancing the stability of 7nm FinFET SRAM cells under process variations. 8T design, because of higher SNM (due to relaxing the read stability constraint), is able to find a valid (i.e., high yield) solution even under very low operating voltages, which can be seen in Fig. 7(a).

However, larger layout area at super- and near-threshold regimes (cf. Fig. 7(b)) compared with 6T design is the main bottleneck. More precisely, the smaller cell area of the 6T SRAM reduces the WL and BL lengths, and hence their capacitances, which in turn reduces the access latency and

TABLE V. ARCHITECTURE-LEVEL SIMULATION RESULTS OF NOMINAL AND OPTIMAL V_{dd} VALUES FOR EACH SRAM CELL STRUCTURE. BEST RESULT OF EACH CACHE CHARACTERISTICS IS BOLD-FACED.

SRAM Cell	V_{dd} (mV)	SRAM Config	Access Latency (ns)	Access Energy (nJ)	Leakage Power (mW)	Area (mm^2)
`6T-SG`	450	(6,2,10)	0.902	0.155	54	1.603
`6T-SG`	414	(6,2,10)	0.948	0.132	45	1.603
`6T-DG`	450	(1,1,10)	**0.557**	0.073	19	**0.703**
`6T-DG`	324	(1,1,10)	0.793	**0.041**	**10**	**0.703**
`8T-SG`	450	(1,2,10)	0.757	0.122	28	1.172
`8T-SG`	348	(1,2,10)	1.040	0.067	15	1.111
`8T-DG`	450	(1,1,10)	0.691	0.093	20	0.980
`8T-DG`	324	(1,1,10)	0.959	0.052	11	0.980

TABLE III. OPTIMAL SRAM CELL CONFIGURATIONS SHOWN AS (N_{PD}, N_{AC}, L_{PU}) FOR DIFFERENT COMBINATIONS OF V_{dd} AND SRAM CELL STRUCTURES. OPTIMAL RESULT UNDER EACH CELL STRUCTURE IS BOLD-FACED, AND THE OVERALL OPTIMAL RESULT IS MARKED BY (*).

	6T-SG			6T-DG			8T-SG			8T-DG		
V_{dd} (mV)	SRAM Config	$\mathbb{E}[E_{leak}]$ (aJ)	Area (nm^2)	SRAM Config	$\mathbb{E}[E_{leak}]$ (aJ)	Area (nm^2)	SRAM Config	$\mathbb{E}[E_{leak}]$ (aJ)	Area (nm^2)	SRAM Config	$\mathbb{E}[E_{leak}]$ (aJ)	Area (nm^2)
450	(6,2,10)	55.65	17,595	(1,1,10)	10.25	7,245	(1,2,10)	19.39	12,075	(1,1,10)	12.70	10,505
414	**(6,2,10)**	**49.24**	**17,595**	(1,1,10)	9.24	7,245	(1,2,10)	17.44	12,075	(1,1,10)	11.39	10,505
348	(9,2,10)	64.97	22,770	(1,1,10)	7.97	7,245	**(1,2,10)**	**15.66**	**12,075**	(1,1,10)	9.71	10,505
324	(16,3,10)	155.04	36,570	**(1,1,10) (*)**	**7.84**	**7,245**	(1,3,10)	20.81	13,800	**(1,1,10)**	**9.49**	**10,505**
300	(20,3,10)	211.04	43,470	(2,1,10)	11.49	8,970	(1,3,10)	19.87	13,800	(1,1,10)	9.68	10,505
192	N/A	N/A	N/A	(19,3,10)	256.21	41,745	(1,10,10)	119.05	25,875	(1,3,10)	30.18	13,955
144	N/A	N/A	N/A	N/A	N/A	N/A	(1,20,10)	403.72	43,125	(1,7,10)	105.13	20,855
102	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	(1,17,10)	407.43	38,105

energy consumption of the whole SRAM array. Accordingly, the combination of 6T design, because of the smaller cell area, and the dual gate control, due to significantly improving the cell stability, emerges as the high yield and most energy efficient memory cell under our deeply-scaled 7nm FinFET process. The optimal (1,1,10) 6T-DG SRAM cell at 324mV even satisfies the WCA constraints which ensures its high immunity to process variations. We also observe that all solutions return $L_{PU} = 10$nm, which is due to the small impact of L_{PU} on the clock cycle, but its higher effect on the leakage power consumption.

Fig. 7(a) shows the optimal expected energy consumption for each SRAM cell structure as a function of V_{dd}. Here, sudden increases in the value of $\mathbb{E}[E_{leak}]$ can be seen. These are caused by the *width quantization property* of FinFET devices which restricts FinFET width to only take discrete values. As a result, when the current SRAM cell configuration cannot satisfy yield constraints in a (slightly) lower V_{dd} value, more fins are needed to increase the ON current of the corresponding (PD or AC) transistor, thereby resulting in a sudden increase in the value of $\mathbb{E}[E_{leak}]$. However, by moving to lower V_{dd}'s this effect is mitigated.

We also performed architecture-level simulations. For this purpose, we modified the CACTI, which is a widely-adopted simulation tool for cache memories [17], to add support for FinFET devices and voltage scaling capability. More precisely, we incorporated our 7nm FinFET devices (including geometries and ON/OFF currents), and FinFET models for calculating gate and drain capacitances, gate area, etc, into CACTI. We picked the process variation tolerant SRAM cell configurations for nominal and optimal V_{dd} values of each SRAM cell structure. Results are given in Table V, which show the effectiveness of the optimal 6T-DG cell in reducing the energy and leakage power consumptions. More accurately, the optimal 6T-DG SRAM cell, compared with the 6T-SG (6T-DG) operating at the normal V_{dd} of 450mV, achieves 3.8× (1.8×) and 5.4× (1.9×) lower access energy and leakage power consumptions, respectively. However, because of operating at the near-threshold regime, the 6T-DG SRAM cell experiences 42% longer access latency compared with the 6T-DG SRAM cell operating at the super-threshold regime.

VI. CONCLUSION

We proposed a cross-layer (device- and circuit-level) framework for designing high yield and energy-efficient SRAM cells using deeply-scaled FinFET technology. Advanced device simulators from Synopsys TCAD tool suite were used to design 7nm FinFET devices, and to extract Verilog-A models for fast SPICE simulations. Dual-gate control was employed in conventional 6T and 8T SRAM cells in order to improve the cell stability and increase energy efficiency. Moreover, we proposed a design flow for minimizing the expected energy consumption of SRAM cells while yield constraints under process variation are satisfied. In our 7nm FinFET process, 6T SRAM cell equipped with the proposed dual-gate control, operating at 324mV, achieves the lowest expected leakage energy consumption under process variation.

ACKNOWLEDGMENT

This research is supported by grants from the PERFECT program of the Defense Advanced Research Projects Agency and the Software and Hardware Foundations of the National Science Foundation.

REFERENCES

[1] H. Pilo *et al.*, "A 64MB SRAM in 32nm High-k Metal-Gate SOI Technology with 0.7V Operation Enabled by Stability, Write-ability and Read-ability Enhancements," *IEEE Journal of Solid-State Circuits*, vol. 47, no. 1, pp. 97–106, Jan 2012.

[2] M. Manoochehri, M. Annavaram, and M. Dubois, "Extremely low cost error protection with correctable parity protected cache," *IEEE Transactions on Computers*, vol. 63, no. 10, pp. 2431–2444, Oct 2014.

[3] H. Pilo *et al.*, "A 64MB SRAM in 22nm SOI Technology featuring Fine-Granularity Power Gating and Low-Energy Power-Supply-Partition Techniques for 37% Leakage Reduction," in *International Solid-State Circuits Conference (ISSCC)*, Feb 2013.

[4] G. Chen *et al.*, "Yield-Driven Near-Threshold SRAM Design," in *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, Nov 2007.

[5] Z. Guo *et al.*, "FinFET-based SRAM Design," in *International Symposium on Low Power Electronics and Design (ISLPED)*, Aug 2005.

[6] F. Moradi *et al.*, "Asymmetrically Doped FinFETs for Low-Power Robust SRAMs," *IEEE Transactions on Electron Devices*, vol. 58, no. 12, pp. 4241–4249, Dec 2011.

[7] S. Tang *et al.*, "FinFET - A Quasi-Planar Double-Gate MOSFET," in *IEEE International Solid-State Circuits Conference (ISSCC)*, 2001, pp. 118–119.

[8] E. Nowak *et al.*, "Turning Silicon on its Edge," *IEEE Circuits and Devices Magazine*, vol. 20, no. 1, pp. 20–31, 2004.

[9] Synopsys Technology Computer-Aided Design (TCAD). [Online]. Available: http://www.synopsys.com/tools/tcad

[10] L. Chang *et al.*, "Stable SRAM Cell Design for the 32nm Node and Beyond," in *Symposium on VLSI Technology*, June 2005, pp. 128–129.

[11] X. Wang *et al.*, "Statistical Variability and Reliability in Nanoscale FinFETs," in *IEEE International Electron Devices Meeting (IEDM)*, Dec 2011.

[12] K. Patel, T.-J. K. Liu, and C. J. Spanos, "Gate Line Edge Roughness Model for Estimation of FinFET Performance Variability," *IEEE Transactions on Electron Devices*, vol. 56, no. 12, pp. 3055–3063, Dec 2009.

[13] H. Kaul *et al.*, "Near-Threshold Voltage (NTV) Design — Opportunities and Challenges," in *Design Automation Conference (DAC)*, June 2012.

[14] M. Seok, D. Sylvester, and D. Blaauw, "Optimal Technology Selection for Minimizing Energy and Variability in Low Voltage Applications," in *ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED)*, 2008.

[15] J. Kwong and A. Chandrakasan, "Variation-Driven Device Sizing for Minimum Energy Sub-threshold Circuits," in *International Symposium on Low Power Electronics and Design (ISLPED)*, Oct 2006.

[16] A. Makosiej *et al.*, "Stability and Yield-Oriented Ultra-Low-Power Embedded 6T SRAM Cell Design Optimization," in *Design, Automation Test in Europe Conference Exhibition (DATE)*, March 2012.

[17] N. Muralimanohar, R. Balasubramonian, and N. Jouppi, "Optimizing NUCA Organizations and Wiring Alternatives for Large Caches With CACTI 6.0," in *International Symposium on Microarchitecture (MICRO-40)*, Dec 2007.

[18] A. Goud *et al.*, "Atomistic Tight-Binding based Evaluation of Impact of Gate Underlap on Source to Drain Tunneling in 5nm Gate Length Si FinFETs," in *71st Annual Device Research Conference (DRC)*, June 2013.

[19] A. Papoulis, *Probability, Random Variables and Stochastic Processes*. McGraw-Hill, 2002.