# FinCACTI: Architectural Analysis and Modeling of Caches with Deeply-scaled FinFET Devices

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

***Abstract*—This paper presents FinCACTI, a cache modeling tool based on CACTI which also supports deeply-scaled FinFET devices as well as more robust SRAM cells. In particular, FinFET devices optimized using advanced device simulators for 7nm process serve as the case study of the paper. Based on this 7nm FinFET process, characteristics of 6T and 8T SRAMs are calculated, and comparison results show that under the same stability requirements the 8T cell has smaller area and leakage power. SRAM and technological parameters of the 7nm FinFET are then incorporated into FinCACTI. According to architecture-level simulations, the 8T SRAM is suggested as the choice of memory cell for 7nm FinFET. Moreover, a 4MB cache in 7nm FinFET compared with 22nm (32nm) CMOS under same access latencies achieves 5× (9×) and 11× (24×) reduction in read energy and area, respectively.**

## I. Introduction

The aggressive down-scaling of transistors to the sub-22nm regime exacerbates short channel effects as well as device mismatches [1]. Under such circumstances, conventional 6T SRAM cells suffer from poor read and write stabilities, which may result in functional failure during memory operations [2]. However, these stability issues can be improved by proper modifications at two abstraction levels as described next.

On one hand, at the device level, short channel effects reduce the ON current of planar CMOS transistors. This means that the SRAM cell cannot provide the desired read/write current levels in order to meet the stability requirements. Quasi-planar FinFET devices [3], however, offer higher ON current under the same channel width compared with CMOS counterparts due to the improved gate control over the channel (and less control by source and drain terminals). Moreover, FinFET devices show superior scalability, higher immunity to random variations and soft errors, and are perceived to be the technology-of-choice beyond the 10nm regime [1]. Consequently, FinFET-based SRAMs have attracted attention as a promising solution to a more robust and energy-efficient SRAM cell design [4] [5].

On the other hand, at the circuit level, read and write operations in conventional 6T SRAM cells share the same path. Functionality of the SRAM cell is thus achieved through the proper sizing of transistors. However, this ratioed design is vulnerable especially in technology nodes below 22nm where process variations and device mismatches become a severe issue. Accordingly, various alternative SRAM cell structures have been proposed [2] [6]. An example is the 8T SRAM cell [2] which enhances the cell stability by dedicating separate paths to each read and write operation.

As a result, future memory systems using deeply-scaled technology nodes necessitate FinFET support and more sophisticated SRAM cell structures. Based on these deeply-scaled devices, various characteristics of SRAM cells, e.g., *static noise margin* (SNM), layout area, leakage power, need to be analyzed in order to find a desirable SRAM cell that simultaneously achieves high stability and low leakage power. Furthermore, evaluating such memory systems at the architecture-level requires modifications to the existing memory models and analysis tools. CACTI [7] is the widely used cache modeling tool for estimating area, delay, and power consumption of on-chip caches. However, the current version of CACTI only supports planar CMOS for technology nodes from 90nm to 32nm. Moreover, technological parameters are extracted from the respective ITRS reports, which are essentially predicted values and thus may affect the accuracy of results. Additionally, CACTI only supports the conventional 6T SRAM.

In order to cope with technology scaling challenges, this paper presents FinCACTI, which enhances CACTI by adding accurate specifications for deeply-scaled FinFET devices, FinFET area/capacitance models, and architectural support for 8T SRAM cell. More precisely, technological parameters are calculated using advanced device simulators (Synopsys TCAD tool suite [8]) on 7nm FinFET process. SPICE-compatible Verilog-A models for 7nm FinFET transistors are extracted from the device simulations for performing fast circuit-level simulations. Using this 7nm FinFET technology, conventional 6T and 8T SRAM cells are compared in terms of SNM, layout area, and leakage power consumption, and comparison results reveal that under the same stability (SNM) requirements the 8T design achieves smaller area as well as lower leakage power.

SRAM characteristics, technological parameters, FinFET analytical models, and other modifications for supporting the 8T SRAM cell are then incorporated into FinCACTI in order to perform architecture-level simulations. Accordingly, areas, access latencies and energies and leakage power consumptions of a 4MB cache under 32nm and 22nm planar CMOS, as well as 7nm FinFET using 6T and 8T cells are computed. Based on architecture-level simulations, the 8T SRAM is suggested as the choice of memory cell for 7nm FinFET. Moreover, the specified cache in 7nm FinFET compared with 22nm (32nm) CMOS under same access latencies achieves 5× (9×) and 11× (24×) reduction in read energy and area, respectively.

The rest of the paper is organized as follows. Section 2 reviews basic concepts of FinFET devices and prior work. Characteristics of 6T/8T SRAM cells under our deeply-scaled FinFET devices are presented in Section 3. Section 4 introduces the FinCACTI. Architecture-level simulation results are

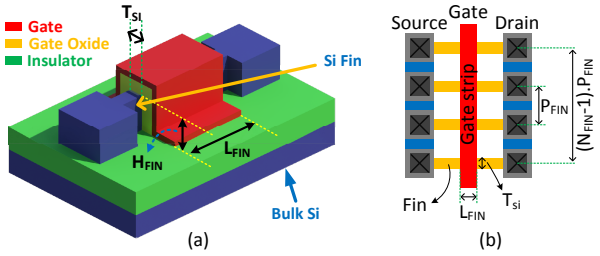


Fig. 1. FinFET device: (a) structure, (b) layout.

presented in Section 5. Finally, Section 6 concludes the paper.

## II. Preliminaries

### A. FinFET Devices

FinFET device is a quasi-planar double-gate transistor [3]. This structure allows FinFET devices to enhance the energy efficiency, ON/OFF current ratio, and soft-error immunity compared with bulk CMOS counterparts. As a result, FinFET technology is currently viewed as the substitute for the bulk CMOS for technology nodes below 32nm [1], [9]. The structure of a three-terminal FinFET device is shown in Figure 1(a). The main component is the *fin* which provides the channel for conducting current when the device is switched on. This vertical fin is surrounded by the gate, and hence a more efficient control over the channel is achieved which in turn helps to suppress short channel effects.

The key geometric parameters of a FinFET device are related to the fin which include the height ($H_{FIN}$), width ($T_{SI}$), and length ($L_{FIN}$) of the fin (cf. Figure 1(a)). The effective channel width of a single fin, $W_{min}$, is thus (approximately) equal to

$$W_{min} \approx 2 \times H_{FIN}. \quad (1)$$

Increasing the width (strength) of a FinFET device is achieved by connecting more fins in parallel. More precisely, the number of fins, $N_{FIN}$, in order to obtain a FinFET with channel width of $W$ is calculated as

$$N_{FIN} = \lceil \frac{W}{W_{min}} \rceil. \quad (2)$$

This is known as the *width quantization property* of FinFET devices (i.e., the FinFET width can only take discrete values), which may result in an over-sized transistor if the required width is not a multiple of $W_{min}$.

Layout of a three-terminal FinFET with four fins is shown in Figure 1(b) [10]. A single strip is used for the gate terminal. Moreover, source (and also drain) terminals of multiple fins are connected together through a metal wire to make a wider FinFET device. A critical process-related geometry in Figure 1(b) is the *fin pitch*, $P_{FIN}$, which is defined as the minimum center-to-center distance of two adjacent parallel fins. The value of $P_{FIN}$ is determined by the underlying FinFET technology. More precisely, there are two types of FinFET technologies: (1) *Lithography-defined* technology where lithographic constraints limit the fin pitch spacing, and (2) *spacer-defined* technology which relaxes the constraints on $P_{FIN}$, and obtains $2\times$ reduction in the value of $P_{FIN}$ at the cost of a more elaborate and costly lithographic process [11].

Major process-related FinFET geometries for 5nm [12] and 7nm technologies are reported in Table I. The table also includes process design rules which are similar for FinFET and CMOS technologies (their difference is in the fin fabrication, which does not influence design rules [10]). In the rest of the paper, we adopt our 7nm FinFET as the case study in order to assess the characteristics of SRAM cells as well as cache structures. Due to the lack of industrial data for such deeply-scaled FinFETs, our device specifications are designed and optimized using the Synopsys TCAD tools [8].

TABLE I. FinFET-specific geometries and design rules

| Parameter | Value in 7nm FinFET (nm) | Value in 5nm FinFET (nm) [12] | Comment |
|---|---|---|---|
| $L_{FIN}$ | 7 | 4.9 | Fin length |
| $T_{SI}$ | 3.5 | 2.725 | Silicon thickness, or fin width |
| $H_{FIN}$ | 14 | 10.9 | Fin height |
| $P_{FIN}$ | $2\lambda + T_{SI}$ = 10.5 | $2\lambda + T_{SI}$ = 7.5 | Fin pitch using spacer lithography |
| $t_{ox}$ | 1.55 | 1.09 | Oxide thickness |
| $W_C$ | $3\lambda$ = 10.5 | $3\lambda$ = 7.5 | Minimum contact size |
| $W_{M2M}$ | $3\lambda$ = 10.5 | $3\lambda$ = 7.5 | Minimum space between metal wires |
| $W_{G2C}$ | $2\lambda$ = 7 | $2\lambda$ = 5 | Minimum space of gate to contact |

### B. Prior Work

Integrating FinFET devices into CACTI has been done previously by a tool called CACTI-FinFET [13], which also considers the effect of process variations on FinFET-based caches. This tool relies on look-up tables in order to generate a design library of device- and gate-level parameters for FinFETs. However, in order to ease the support of new technologies, we adopt analytical models for calculating gate-level information from technology-dependent device-level parameters.

Moreover, while CACTI-FinFET uses compact models to extract technological parameters for a 22nm process, we use Synopsys TCAD to obtain accurate device-level parameters for a deeply-scaled (7nm) FinFET technology. Additionally, we perform Verilog-A-based SPICE simulations as well as layout-aware calculations to precisely characterize SRAM cells in our 7nm FinFET process. We believe the strength of CACTI-FinFET is its process variation models and the ability to analyze FinFET-based caches under such models. Our emphasis, however, is on accurate device-level parameters and analytical gate-level models for deeply-scaled FinFET devices which are explained in more detail in the following sections.

## III. SRAM Cell Characteristics Using Deeply-Scaled FinFET Devices

The conventional 6T SRAM cell, as shown in Figure 2(a), is composed of two cross-coupled inverters ($M1$–$M4$) which form the storage element. Moreover, in order to read from and write into the memory cell, the SRAM cell is equipped with two access transistors ($M5$ and $M6$). Since access transistors are shared between read and write operations, careful sizing of transistors is critical for attaining high read and write stabilities. More precisely, during read operation, access transistor should be weaker than the pull-down transistor such that the access transistor cannot flip the stored data. On the other hand, for a successful write operation, access transistor has to be able to change the stored data, and hence access transistor should be stronger than the pull-up transistor.

This ratioed design, however, is vulnerable especially in technology nodes below 22nm where process variations become a severe issue. Accordingly, various alternative SRAM cell structures have been proposed. Among them, the 8T cell design (cf. Figure 2(b)) [2] utilizes an additional read bit-line in order to separate read and write paths, and by decoupling the storage node from the read bit-line, stability is improved.

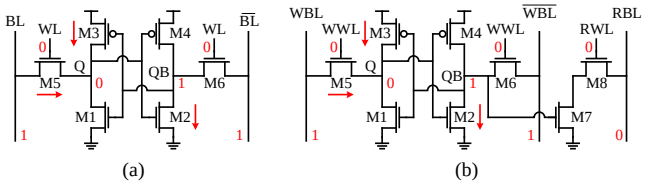


Fig. 2. Circuit schematics of (a) conventional 6T, and (b) 8T [2] SRAM cells. $BL$ and $WL$ denote bit-line and word-line, respectively. The 8T design contains separate $BL$ and $WL$s for read and write operations. Arrows highlight subthreshold leakage paths in an idle SRAM cell (storing 0).

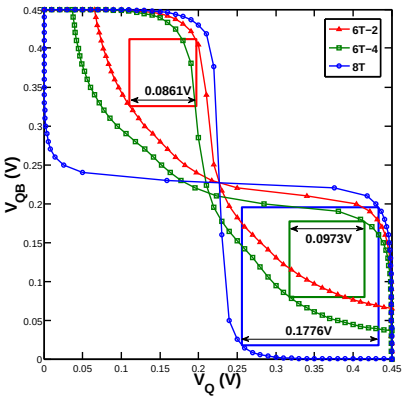


Fig. 3. Butterfly curves for 6T–2 (△), 6T–4 (□), and 8T(○) SRAM cells in 7nm FinFET. SNM values for each cell is included in the figure.

In this section, we investigate (and compare) the characteristics of 6T and 8T SRAM cells in order to find a robust SRAM cell for our deeply-scaled FinFET devices. For the 6T cell, we consider four variations where in all cases a single-fin device is used for the access and pull-up transistors ($M3$–$M6$), and thus their difference is in the number of fins of pull-down transistors ($M1$ and $M2$) which varies from one to four. We will use 6T–$n$ ($1 \leq n \leq 4$) in the paper to refer to the 6T SRAM cell whose pull-down transistors have $n$ fins each. All transistors in the 8T design are assumed to be single-fin.

**Butterfly curves.** 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. Figure 3 provides the common graphical representation of SNM, i.e., the butterfly plot, for 6T–2, 6T–4, and 8T SRAM cells during read access. The butterfly plot is derived through combining the *voltage transfer curves* (VTCs) of the two inverters with one VTC inverted, while taking into account the effect of access transistors $M5$ and $M6$. The SNM is found graphically as the length of the side of a square fitted between the VTCs and having the longest diagonal. As shown in the figure, 6T–4 SRAM cell slightly increases the SNM by 13% compared with the 6T–2 SRAM cell. On the other hand, the 8T SRAM cell significantly improves the SNM from 0.0861V to 0.1776V compared with the 6T–2 cell. SNM values of all SRAM cells expecpt for 6T–1 cell which does not work properly in our 7nm technology (because of the weak $M1$ and $M2$) are reported in Table II.

**Layout Area.** Based on FinFET-specific geometries and design rules reported in Table I, layouts of 6T–2 and 8T SRAM cells are shown in Figures 4(a) and 4(b), respectively. Other variations of the 6T cell have a layout similar to that of shown in Figure 4(a) except that $M1$ and $M2$ transistors in the 6T–$n$ cell are drawn with $n$ parallel fins. In all cases, the height of

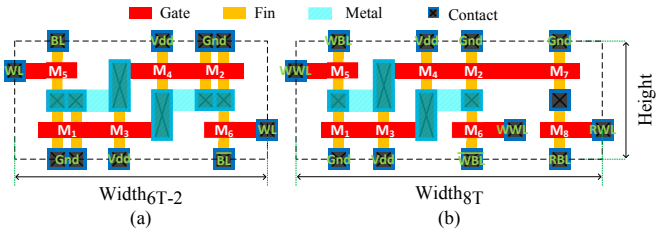


Fig. 4. Layouts of (a) 6T–2, and (b) 8T SRAM cells. Name of transistors and signals are shown on gate and contact locations, respectively.

TABLE II. COMPARISON OF SRAM CELLS IN 7NM FINFET

| SRAM Cell | SNM (V) | Area ($nm^2$) | Aspect Ratio | Leakage Power (nW) |
|---|---|---|---|---|
| 6T–1 | – | 6,615 | 0.60 | 0.67 |
| 6T–2 | 0.0861 | 7,938 | 0.50 | 1.58 |
| 6T–3 | 0.0925 | 9,261 | 0.43 | 3.20 |
| 6T–4 | 0.0973 | 10,584 | 0.38 | 1.92 |
| 8T | 0.1776 | 9,261 | 0.43 | 1.32 |

the SRAM cell is obtained by

$$Height = 2L_{FIN} + 4W_{G2C} + 2W_C = 2L_{FIN} + 14\lambda. \quad (3)$$

On the other hand, the width of the 6T–$n$ cell is given by

$$\begin{aligned} Width_{6T-n} &= 2(n-1)P_{FIN} + 5W_{M2M} + 5W_C \\ &= 2(n-1)P_{FIN} + 30\lambda, \end{aligned} \quad (4)$$

whereas the width of the 8T cell is calculated by

$$Width_{8T} = 7W_{M2M} + 7W_C = 42\lambda. \quad (5)$$

According to aforementioned equations, the area and the aspect ratio (defined as $Height/Width$) of 6T-$n$ ($1 \leq n \leq 4$), and 8T cells for 7nm FinFET process technology are given in Table II. Although the 6T–1 SRAM cell does not work with our deeply-scaled FinFET devices, the area of the 6T–1 cell is reported for comparison purposes. Accordingly, the areas of 6T–2, 8T, 6T–3, and 6T–4 SRAM cells are 20%, 40%, 40%, and 60% larger than the area of 6T–1 cell, respectively. Moreover, area of the 8T cell is 17% larger than the smallest working 6T (i.e., 6T–2), but 14% smaller than the 6T cell with the highest SNM (i.e., 6T–4).

**Leakage Power.** During the idle mode of an SRAM cell, $BL$ and $BLB$ (or $WBL$ and $WBLB$) are precharged to $Vdd$, $RBL$ is predischarged to 0, and all word-lines are deactivated. Therefore, for a 6T SRAM that stores bit '0', $M2$, $M3$, and $M5$ as shown in Figure 2(a) have subthreshold leakage paths from $Vdd$ to 0. Similarly, $M1$, $M4$, and $M6$ establish subthreshold leakage paths in a 6T SRAM storing bit '1'. As a result, due to symmetry of 6T SRAM cell, leakage power in both cases is identical. For 8T cell, since $RBL$ is predischarged to 0, both ends of $M7$ and $M8$ are 0, which in turn reduces the subthreshold current on that path.

In order to validate above discussions, we calculated leakage power of SRAM cells using Verilog-A models extracted from our 7nm FinFET specification. Results are reported in the last column of Table II. For the 8T SRAM, average leakage power of the cell when storing '0' and '1' is calculated. Among the functional SRAMs under our 7nm FinFET devices, 8T cell has the smallest leakage power.

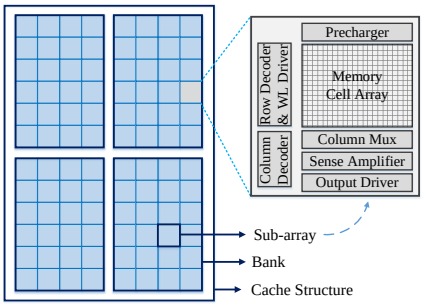


Fig. 5. Cache structure consisting of multiple banks, where each bank contains several sub-arrays. Main components of a sub-array are also shown on the right-hand side of the figure.

## IV. FinCACTI

CACTI [7] is a widely used delay, power, and area modeling tool for cache and memory systems. In CACTI, large cache memories are divided into multiple banks that can be accessed simultaneously in order to improve the memory access bandwidth. Additionally, to reduce the delay and power consumption of word-lines and bit-lines, each bank is partitioned into several sub-arrays, where only one sub-array in a bank can be accessed at any time. A combination of the number of banks, the number of sub-arrays in each bank row and column, along with the aspect ratio of sub-arrays forms the cache structure. Using such cache structures and a description of the process technology node, CACTI can estimate cache access and cycle times, leakage and dynamic power consumptions, and the area of the cache. Alternatively, we can configure CACTI in order to derive the most desirable cache structure that minimizes a user-defined cost function of delay, power and area.

The current version of CACTI only supports planar CMOS process for technology nodes between 90nm to 32nm using the standard 6T SRAM cell design. As a result, for technology nodes below 32nm, support for FinFET devices as well as more robust and sophisticated SRAM cell designs are needed. In the following subsections, we describe how deeply scaled FinFET models and the 8T SRAM cell are incorporated into CACTI. We refer to this new version of CACTI as FinCACTI (FinFET integration into CACTI).

### A. FinFET Support

**Technological Parameters.** CACTI uses the ITRS MASTAR tool[1] [14] to estimate technological parameters for 90nm to 32nm CMOS processes. These are predicted values based on the device scaling trends, and hence may influence the accuracy of results. On the other hand, we design our deeply-scaled (7nm) FinFET devices using Synopsys Technology Computer-Aided Design (TCAD) tool suite [8], which can generate accurate results with device simulators based on physics-driven models. Hence, we extract from TCAD simulations the major technological parameters including FinFET-specific geometries, supply and threshold voltages, ideal gate capacitance, and ON/OFF currents of N- and P-type fins (for temperatures from 300K to 400K). Moreover, we also extract SPICE-compatible Verilog-A models, which are much faster for circuit-level simulations compared with device simulators,

[1]MASTAR (*model for assessment of CMOS technologies and roadmaps*) is a computing tool released by the ITRS for calculating the electrical characteristics of advanced CMOS transistors [14].

TABLE III. Important technological parameters for 7nm FinFET devices. Currents are obtained for 300K.

| Parameter | Value | Comments |
|---|---|---|
| $Vdd$ (V) | 0.45 | Supply voltage |
| $Vth$ (V) | 0.235 | Threshold voltage |
| $I_{ON,NMOS}$ (A/$\mu$m) | 8.82e-04 | ON current of a N-type FinFET § |
| $I_{ON,PMOS}$ (A/$\mu$m) | 5.50e-04 | ON current of a P-type FinFET § |
| $I_{OFF,NMOS}$ (A/$\mu$m) | 7.62e-08 | OFF current of a N-type FinFET † |
| $I_{OFF,PMOS}$ (A/$\mu$m) | 1.16e-07 | OFF current of a P-type FinFET † |
| $L_{phy}$ (nm) | 7 | Physical gate length |
| $C_{g,ideal}$ (A/$\mu$m) | 1.59e-16 | Ideal gate capacitance = $\frac{\epsilon_{ox}}{t_{ox}} \cdot L_{phy}$ |
| PMOS to NMOS size ratio | 1.6 | |
| NAND2 stack effect factor | 0.4 | Stack effect of two N-type FinFETs |
| NAND3 stack effect factor | 0.2 | Stack effect of three N-type FinFETs |
| NOR2 stack effect factor | 0.4 | Stack effect of two P-type FinFETs |

§ ON current is defined as the drain current when $V_G = V_D = Vdd$, and $V_S = 0$ for NMOS, and $V_G = V_D = 0$, and $V_S = Vdd$ for PMOS transistors.
† OFF current denotes the leakage current when $V_G = 0$, and $V_{DS} = Vdd$ for NMOS, and $V_G = Vdd$, and $V_{DS} = -Vdd$ for PMOS transistors.

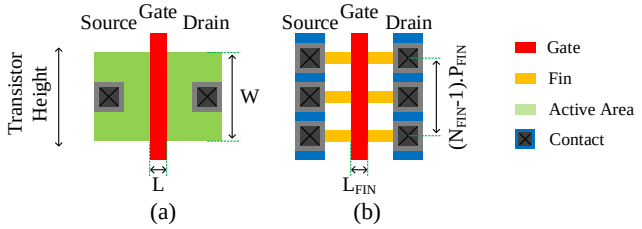


Fig. 6. Layouts of a transistor with channel width of $W$ in (a) planar CMOS and (b) FinFET process technologies.

in order to derive gate- and circuit-level parameters such as the PMOS to NMOS size ratio, and the stack effect factor (leakage current ratio of single to stacked device).

For other parameters that we could not derive using device simulators (e.g., effective mobility, fringing capacitance, and electrical gate length), we used double-gate (FinFET) device profiles from the MASTAR tool, which are prediction results on 7nm FinFET devices. Furthermore, technological parameters of wires in conservative and aggressive technology scaling projections for 7nm process technology are extrapolated from data for 180nm to 13nm technologies [15]. Important technological parameters are summarized in Table III. For FinFET-specific geometries please refer to Table I.

**Transistor Area.** Figures 6(a) and 6(b) show layouts of a transistor with channel width of $W$ in planar CMOS and FinFET process technologies, respectively. In planar CMOS, transistor height is equal to the channel width ($W$). However, the channel width in FinFET devices is along the $z$-axis which does not impact the device area, and hence is not equal to the transistor height. Basically, transistor height for a FinFET device is proportional to $(N_{FIN} - 1) \cdot P_{FIN}$, where $N_{FIN}$ denotes the number of fins obtained from Equation (2), and $P_{FIN}$ is a process related geometry that restricts the minimum space between two adjacent fins [10]. In our area model for FinFETs, we use $N_{FIN} \cdot P_{FIN}$ as the transistor height.

*Example*. Consider a transistor with channel width of $W$ = 56nm in 7nm process technology. The transistor height in CMOS will be 56nm, whereas in our 7nm FinFET technology the transistor height is calculated as $\lceil 56\text{nm}/(2 \times 14\text{nm}) \rceil \cdot 10.5\text{nm} = 21\text{nm}$.

Transistor width, on the other hand, is determined by contact-related design rules (i.e., $W_C$ and $W_{G2C}$) and the channel length. As mentioned above, design rules are identical in CMOS and FinFET technologies for wires and contacts.

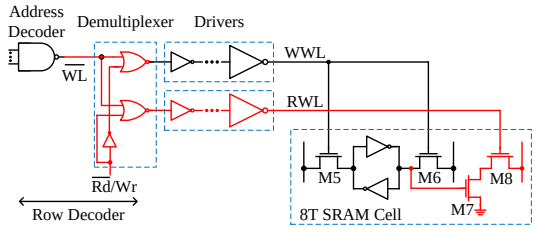


Fig. 7. Row decoder for an 8T SRAM cell. Highlighted parts with red color (demultiplexer, RWL drivers, $M7$ and $M8$ transistors) do not exist in conventional 6T cell designs. $\overline{Rd}/Wr$ is 0 for read, and 1 for write operation.

Thus, by assigning the proper value to the channel length based on the chosen process technology, transistor width calculations remain the same for both CMOS and FinFET technologies.

**Gate and Diffusion Capacitances.** Due to the width quantization effect of FinFETs, the effective channel width may become larger than the required width. For a FinFET device, the total effective channel width is calculated as $N_{FIN} \cdot W_{min}$. By plugging this equation into the MASTAR's gate capacitance equation, we can compute the gate capacitance ($C_G$) of a FinFET device as follows:

$$C_G(N_{FIN}) = (C_{g,ideal} + C_{ov} + C_{fr}) \cdot N_{FIN} \cdot W_{min}, \quad (6)$$

where $C_{g,ideal}$, $C_{ov}$, and $C_{fr}$ are ideal gate, overlap, and total fringing capacitances, respectively. Furthermore, in order to calculate the drain capacitance ($C_D$) of a FinFET transistor, we use bias-independent drain-side junction capacitance calculations from the BSIM-CMG [16]:

$$\begin{aligned} C_D(N_{FIN}) &= C_j \cdot A_D + C_{jsw} \cdot P_D \\ &\quad + C_{jswg} \cdot N_{FIN} \cdot W_{min}, \\ A_D &= (W_D \cdot T_{si}) \cdot N_{FIN}, \\ P_D &= 2(W_D + T_{si}) \cdot N_{FIN}. \end{aligned} \quad (7)$$

In the above equations, $C_j$ is the unit area drain junction capacitance, $C_{jsw}$ and $C_{jswg}$ are unit length sidewall and gate sidewall junction capacitances, respectively, $W_D$ is the total drain width, and $A_D$ and $P_D$ are the area and perimeter of the drain juction, respectively. As can be seen in Equations (6) and (7), $C_G$ and $C_D$ are functions of the number of fins.

### B. 8T SRAM Cell

CACTI only supports the conventional 6T SRAM cell design, which is specified by defining the width (size) of the access, pull-down, and pull-up transistors, along with the area and aspect ratio of the cell. For FinFET technology, we use the number of fins to specify the size of each transistor in the SRAM cell. While word-line ($WL$) and bit-line ($BL$) are shared between read and write operations in the 6T design, the 8T cell dictates separate $WL$ and $BL$ for read and write accesses. According to Figure 7, capacitance of read and write $WL$s (i.e., $RWL$ and $WWL$), and read and write $BL$s (i.e., $RBL$ and $WBL$) are modeled as follows for an 8T cell-based sub-array with $n$ rows and $m$ columns.

$$\begin{aligned} C_{RWL} &= m \cdot (C_G(N_{FIN,M_8}) + W_{Cell} \cdot C_W), \\ C_{WWL} &= m \cdot (2 \cdot C_G(N_{FIN,M_5}) + W_{Cell} \cdot C_W), \\ C_{RBL} &= n \cdot (C_D(N_{FIN,M_8})/2 + H_{Cell} \cdot C_W), \\ C_{WBL} &= n \cdot (C_D(N_{FIN,M_5})/2 + H_{Cell} \cdot C_W), \end{aligned}$$

TABLE IV. CACHE CONFIGURATION.

| Parameter | Value | Parameter | Value |
|---|---|---|---|
| Cache size | 4MB | Device type | HP |
| Block size | 64B | Associativity | 8 |
| Read/write ports | 1 | Bus width | 512 |
| Cache model | UCA | Number of banks | 4 |
| Temperature | 330K | Objective | Energy-Delay Product |

where $W_{Cell}$ and $H_{Cell}$ denote the width and height of the SRAM cell, respectively, which are obtained from the cell area and cell aspect ratio as described in Section III, $C_W$ represents unit length wire capacitance, and $N_{FIN,M_i}$ is the number of fins in transistor $Mi$. Since BL contancts are shared between two adjacent cells, half of the drain capacitance is considered for each cell. Moreover, the row decoder of the 8T design should be modified such that one of the $RWL$ or $WWL$ is activated at any time. For this purpose, a 1-to-2 demultiplexer is added to each output of the address decoder. Accordingly, distinct drivers are needed for $RWL$ and $WWL$. The structure of this modified row decoder is shown in Figure 7, and is incorporated in FinCACTI for 8T FinFET cell support.

For planar CMOS process technologies, specifications of the 8T SRAM are obtained from [2]. Accordingly, we use $195F$ as the area, and 0.36 as the aspect ratio of the 8T CMOS-based SRAM cell, where $F$ denotes the feature size. The width of all transistors is $1F$, except for $M7$ which is $3.5F$.

## V. SIMULATION RESULTS

Architecture-level simulation results are presented in this section. For all simulations, a 4MB, 8-way L3 cache with configurations given in Table IV is assumed. Areas, access latencies, read energies and leakage power consumptions of this cache under 32nm and 22nm planar CMOS, as well as 7nm FinFET using 6T and 8T SRAM cells are computed and compared. Technological parameters of 22nm CMOS are extracted from McPAT [17]. For each process technology and SRAM cell pair, the cache structure with optimal Energy-Delay product is found. Moreover, results of 6T–1 cell under 7nm FinFET are reported for comparison purposes.

Cache area results are shown in Figure 8(a). Considering 8T cell-based caches, the area of 22nm CMOS is about 2× smaller than that of 32nm CMOS. On the other hand, shifting from 22nm CMOS toward 7nm FinFET reduces the area by a factor of 11. This 11× area reduction not only comes from technology down-scaling, but also from the smaller footprint of FinFET devices as discussed in Section IV-A. Moreover, the area overhead of 8T-based caches compared with 6T counterparts in both CMOS and FinFET are about 30%, which is mainly due to the area overhead at the cell-level. In other words, the area of the extra hardware that was added to the row decoder of the 8T design can be considered negligible.

We next compare access latencies and read energies which are shown in Figures 8(b) and 8(c), respectively. FinFET-based caches achieve considerably lower access energies compared with CMOS counterparts, due to smaller feature size, shorter access latencies, and because of excellent control over the channel which allows them to operate at lower supply voltages (e.g., 0.45V in our 7nm FinFET process). In particular, the 8T-based cache in 7nm FinFET compared with 22nm (32nm) CMOS has 5× (9×) lower read energy and 1.7× (2×) shorter access latency, respectively. Furthermore, larger area of the 8T cell increases $WL$ and $BL$ capacitances, which in turn results in higher latency and energy for accessing a memory cell.

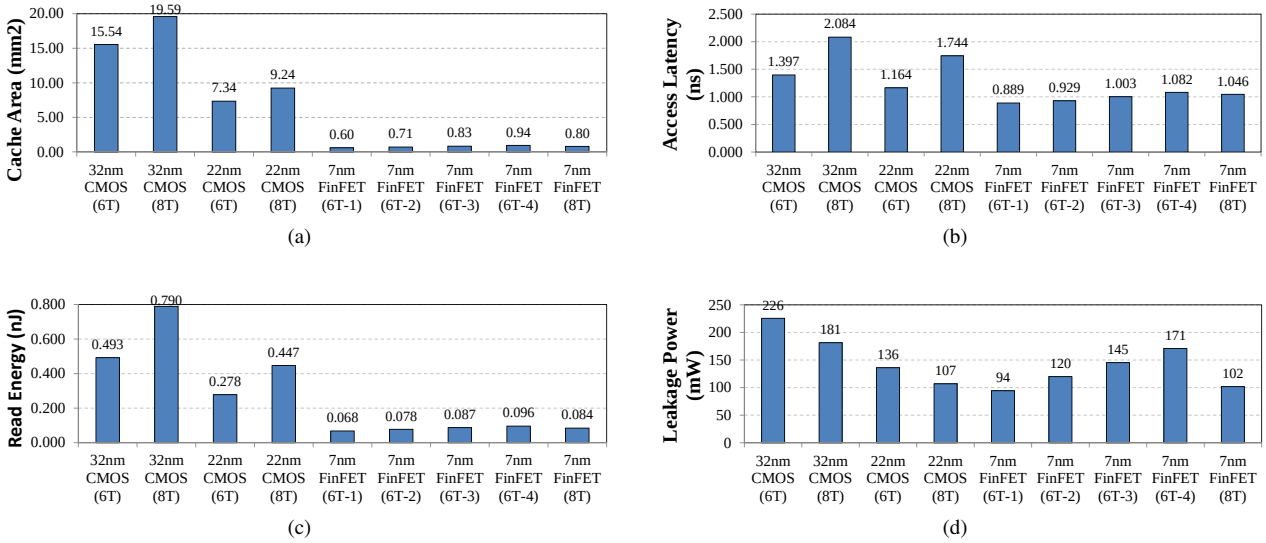


Fig. 8. (a) Area, (b) access latency, (c) read energy, and (d) leakage power of a cache with configuration parameters shown in Table IV. Results are reported for 32nm and 22nm planar CMOS, as well as 7nm FinFET using 6T and 8T SRAM cells.

Finally, we discuss the leakage power results which are shown in Figure 8(d). As can be seen, leakage power has not scaled well with technology nodes by shifting toward 7nm FinFET. This may be due to the usage of inaccurate leakage models in CACTI. Since SRAM cell array is the major source of leakage power in caches, we used results of Table II (which are based on SPICE simulations) as the leakage power of SRAM cell instead of CACTI's model. We observed that leakage power values for FinFET-based caches are on average four times smaller than what is shown in Figure 8(d). However, because we do not have such information for SRAM cells in 32nm and 22nm CMOS processes, we are not able to perform a fair comparison.

Nevertheless, for CMOS-based caches, 8T design has lower leakage power. The reason is that all transistors that contribute to the leakage power in the 8T design are minimum-sized. Additionally, among all functional SRAMs in our 7nm FinFET process, 8T-based cache has the smallest leakage power, and increasing the number of fins of the pull-down transistor in the 6T cell causes significant increase in the leakage power.

## VI. Conclusion

We introduced FinCACTI, a cache modeling tool built on top of the widely used CACTI tool. FinCACTI adopts analytical FinFET models, accurate technological parameters for deeply-scaled (7nm) FinFET devices, and architectural support for 8T SRAM cell in order to assess characteristics of on-chip caches for future deeply-scaled FinFET technologies. We showed that for 7nm FinFET process, 8T SRAM compared with 6T counterpart achieves higher read stability at cell-level, and higher energy efficiency at architecture-level. Furthermore, a 4MB, 8-way L3 cache in 7nm FinFET compared with 22nm (32nm) CMOS under same access latencies achieves 5× (9×) and 11× (24×) reduction in read energy and area, respectively.

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

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

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

[4] Z. Guo *et al.*, "FinFET-based SRAM Design," in *International Symposium on Low Power Electronics and Design (ISLPED)*, Aug 2005, pp. 2–7.

[5] 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.

[6] K. Takeda *et al.*, "A Read-static-noise-margin-free SRAM Cell for Low-Vdd and High-Speed Applications," *IEEE Journal of Solid-State Circuits*, vol. 41, no. 1, pp. 113–121, Jan 2006.

[7] CACTI: An Integrated Cache and Memory Access Time, Cycle Time, Area, Leakage, and Dynamic Power Model. [Online]. Available: http://www.hpl.hp.com/research/cacti/

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

[9] T. Sairam, W. Zhao, and Y. Cao, "Optimizing FinFET Technology for High-Speed and Low-Power Design," in *17th GLSVLSI*, pp. 73–77, 2007.

[10] M. Alioto, "Comparative Evaluation of Layout Density in 3T, 4T, and MT FinFET Standard Cells," *IEEE Trans. on VLSI Systems*, vol. 19, no. 5, pp. 751–762, 2011.

[11] Y.-K. Choi, T.-J. King, and C. Hu, "Nanoscale CMOS Spacer FinFET for the Terabit Era," *IEEE Electron Device Letters*, vol. 23, no. 1, pp. 25–27, 2002.

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

[13] C.-Y. Lee and N. Jha, "CACTI-FinFET: An Integrated Delay and Power Modeling Framework for FinFET-based Caches under Process Variations," in *48th Design Automation Conference (DAC)*, June 2011.

[14] The Model for Assessment of CMOS Technologies and Roadmaps (MASTAR). [Online]. Available: http://www.itrs.net/models.html

[15] R. Ho, "On-chip Wires: Scaling and Efficiency," Ph.D. Dissertation, Stanford University, 2003.

[16] V. Sriramkumar *et al.* BSIM-CMG 107.0.0: Multi-Gate MOSFET Compact Model (Technical Manual). [Online]. Available: http://www-device.eecs.berkeley.edu/bsim/?page=BSIMCMG

[17] S. Li *et al.*, "McPAT: An Integrated Power, Area, and Timing Modeling Framework for Multicore and Manycore Architectures," in *42nd International Symposium on Microarchitecture (MICRO-42)*, Dec 2009.