A Cross-Layer Design Framework and Comparative Analysis of SRAM Cells and Cache Memories using 7nm FinFET Devices

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

I. Introduction

FinFET devices are currently viewed as the technology-of-choice beyond the 10nm regime [1]. This is mainly due to the improved gate control over the channel which makes FinFETs more immune to short channel effects. On the other hand, SRAM caches, because of occupying a large portion of the chip area, and high sensitivity to device mismatches, are considered as the major bottleneck of the V_{dd} scaling [2]. Hence, FinFET-based SRAMs have emerged as a solution to a more robust and energy efficient memory design [3]. This paper thus adopts a cross-layer design framework (Fig. 1) in order to study the effect of different deeply-scaled (7nm) FinFET devices on memory designs: (1) at device-level, different FinFET devices for 7nm process are designed using TCAD tools [4], (2) at circuit-level, Verilog-A models are extracted from the device simulator for performing fast SPICE-based simulations, (3) and finally at architecture-level, the overall characteristics of an on-chip cache is assessed using a modified version of CACTI tool with FinFET support.

II. Cross-layer design framework

A. Device-level Design

Since no industrial data for deeply-scaled FinFET devices exist, 7nm FinFET devices are modeled (Fig. 2) and simulated using Sentaurus TCAD tools [4]. Gate underlap is introduced to mitigate the *direct source-to-drain tunneling* (DSDT) current [5]. We develop seven different designs of 7nm FinFET devices with different parameters such as gate length L_{FIN}, oxide thickness t_{ox}, fin width W, and underlap length ul. Table 1 reports the design parameters of the baseline (standard) FinFET device, whereas Table 2 shows the design parameters of other devices with only one parameter changed per device. Based on device simulations, we also extract SPICE-compatible Verilog-A models for fast circuit-level simulations, e.g. deriving ON/OFF currents of FinFET devices, *static noise margin* (SNM), as well as other parameters for integration into architecture-level simulators. According to Fig. 3(a), the highest ON current is achieved by the *high_w* device which has a larger fin width (which means a larger effective channel width) compared with the baseline device. On the other hand, as a result of the V_{th} roll-off effect, the lowest OFF current (Fig. 3(b)) and the highest ON/OFF current ratio (Fig. 3(c)) are obtained by using the *high_l* device (with a longer gate length).

B. Circuit-level Design

FinFET devices are next incorporated into 6T and 8T [6] SRAMs in order to find a robust and functional cell under this 7nm FinFET process. Since the P-type fin is (1.6x) weaker than the N-type counterpart, we only need to increase the number of fins of pull-down transistors for the 6T cell to ensure proper operation. Therefore, 6T-*n* is used to refer to a 6T cell whose pull-down transistors have *n* fins each, where $n>1$ since 6T-1 cell does not work properly in our 7nm FinFET process (because of weak pull-downs). On the other hand, 8T cell, by dedicating separate paths to read and write operations, does not need stronger pull-down transistors, and hence all transistors can be single-fin. Area (for memory density) and SNM (for robustness) of SRAMs are calculated based on cell layouts (Fig. 4, 5) and butterfly curves (Fig. 6), respectively. In general, the SNM is higher if the corresponding FinFET device has higher ON/OFF current ratio. The highest SNM is achieved by 8T cell using *high_l* devices (Fig. 7(b)) at the cost of 21% larger area compared with the smallest working 6T cell (Fig. 7(b)). The reason is higher SNM of 8T cell compared with 6T cell, and the highest ON/OFF current ratio in *high_l* devices.

C. Architecture-level Design

In order to evaluate SRAM cells at the architecture-level, deeply-scaled FinFET devices along with FinFET models are integrated into CACTI [7], which is a widely-used cache modeling tool. For this purpose, a 4MB cache (Table 3) is assumed. Cache area and access energies do not change significantly when using different FinFET devices, and are thus omitted. Access latency is mainly determined by the ON current of the underlying device, and hence, the shortest access latency is achieved by using *high_w* devices (Fig. 8(a)). On the other hand, the OFF current of the SRAM cell is the major component of the cache leakage power. Accordingly, the *high_l* device achieves lowest cache leakage power (Fig. 8(b)). Meanwhile, due to the usage of all single-fin transistors, 8T cell experiences less power consumption compared with working 6T cells. In summary, 8T cell using *high_l* devices has the lowest leakage power, with 18% latency penalty compared with the fastest 6T cell.

III. Conclusion

Seven FinFET devices optimized for 7nm technology along with three SRAM cells were evaluated and compared. The *high_l* device has the lowest OFF current and the highest ON/OFF current ratio. Moreover, 8T SRAM cell achieves the highest SNM which guarantees its robust operation. Hence, 8T SRAM cell using *high_l* devices is suggested as the choice of memory cell for the discussed 7nm FinFET process.

Acknowledgment

This work is funded by the DARPA PERFECT program.

References

[1] E. Nowak et al., *IEEE Circuits and Devices Magazine*, 20(1), 2004. [2] Baravelli et al., *Solid-State Electronics*, 54(9), 2010. [3] Guo et al., *ISLPED*, 2005. [4] http://www.synopsys.com/tools/tcad. [5] A. Goud et al., *DRC*, 2013. [6] L. Chang et al., *Symposium on VLSI Technology*, 2005. [7] http://www.hpl.hp.com/research/cacti/.

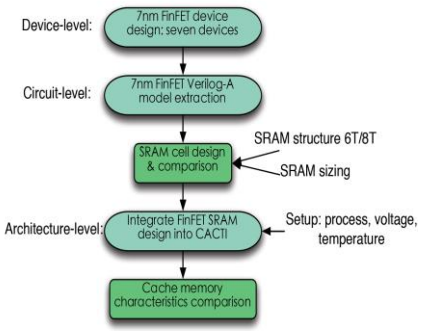
Device-level:
7nm FinFET device design: seven devices
Circuit-level:
7nm FinFET Verilog-A model extraction
SRAM structure 6T/8T
SRAM cell design & comparison
SRAM sizing
Architecture-level:
Integrate FinFET SRAM design into CACTI
Setup: process, voltage, temperature
Cache memory characteristics comparison

Fig. 1. Cross-layer design framework.

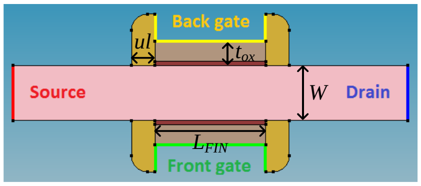

Fig. 2. 2-D model for 7nm FinFET in TCAD device simulator [4].

Table 1. Design parameters of the baseline 7nm FinFET device.

Parameter name	Value
Gate length (L_{FIN})	7nm
Fin width (W, or T_{SI})	3.5nm
Fin height	14nm
Gate oxide material	SiO_2+HfO_2
Gate oxide thickness (t_{ox})	1.3nm
Gate underlap (ul)	1.5nm
Source/Drain doping	$1\times10^{20}cm^{-3}$
Gate work function (NFET)	4.4eV
Gate work function (PFET)	4.9eV

Table 2. Design parameters of other 7nm FinFET devices. For each device, only one parameter is changed.

Device	Parameter	Value
low_w	*W*	3.2nm
high_w	*W*	3.8nm
low_tox	t_{ox}	1.1nm
high_tox	t_{ox}	1.5nm
high_ul	*ul*	2.25nm
high_l	L_{FIN}	8nm

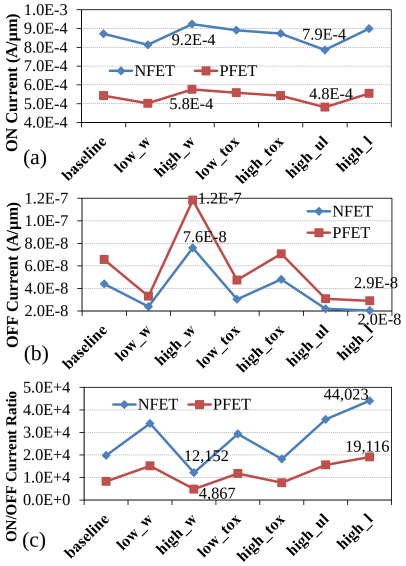

Fig. 3. (a) ON currents, (b) OFF currents, and (c) ON/OFF current ratios of N- and P-type FinFET devices. (†)

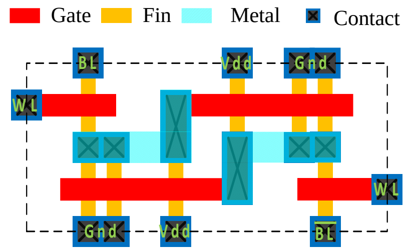

Fig 4. Layout of 6T SRAM cell. Pull-down transistors have 2 fins each. Other transistors have one fin.

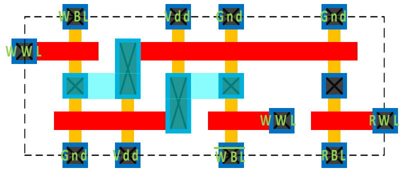

Fig. 5. Layout of 8T SRAM cell with single-fin devices. Separate read path increases the SNM.

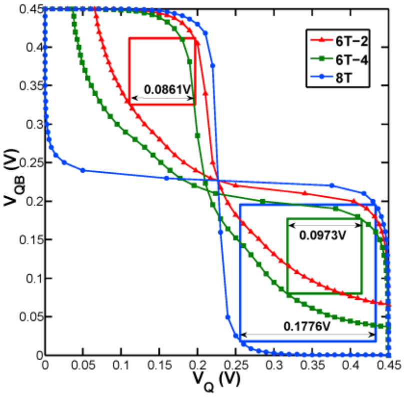

Fig. 6. Butterfly curves of SRAM cells during read access using the baseline 7nm FinFET device. The butterfly curve is derived by combining the *voltage transfer curves* (VTCs) of the two inverters with one VTC inverted. *Static noise margin* (SNM) values are also shown. SNM of 8T cell is 1.8x higher than that of the best 6T. (*)

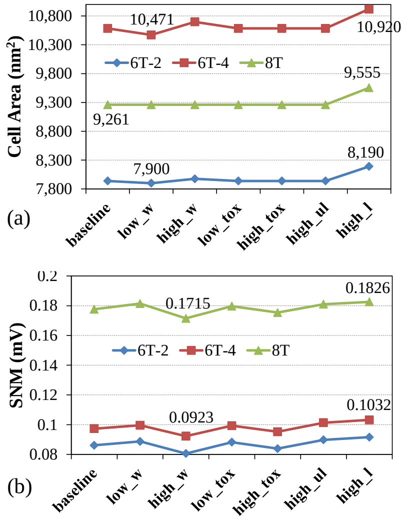

Fig. 7. (a) Layout areas, and (b) SNM values of SRAM cells using different 7nm FinFET devices. (†) (*)

Table 3. Cache configuration.

Parameter	Value
Cache size	4MB
Cache level	L3
Block size	64B
Associativity	8
Number of read/write ports	1
Cache model	UCA
Number of banks	4
Output/input bus width	512
Temperature	300K

UCA: Uniform Cache Access

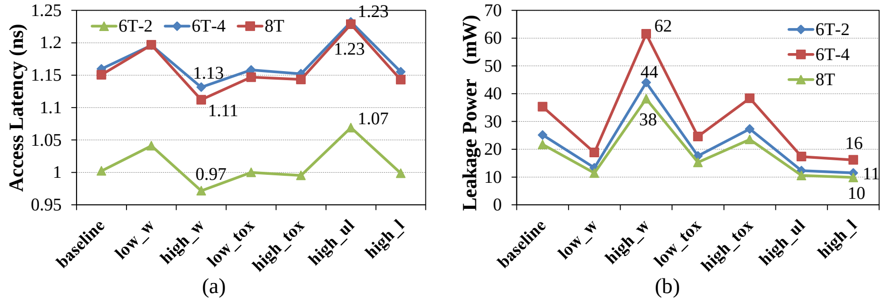

Fig. 8. (a) Access latency, and (b) leakage power of the cache for various combinations of SRAM cells and FinFET devices. Higher ON current leads to shorter access latency (a), whereas higher OFF current causes larger leakage power dissipation (b). (†) (*)

(†) Numbers on each plot show maximum and minimum values.

(*) 6T-*n* denotes a 6T SRAM cell whose pull-down transistors have *n* fins each.