IOSR Journal of VLSI and Signal Processing (IOSR-JVSP)
Volume 4, Issue 2, Ver. IV (Mar-Apr. 2014), PP 38-44
e-ISSN: 2319 – 4200, p-ISSN No. : 2319 – 4197
www.iosrjournals.org

SRAM based 28 nm Technology for consumer electronics applications

J. Ravibabu, P.Sairaghava Reddy, Perna Kishor Krishna,
M.TECH, Department of ECE, Sreenidhi Institute of Science and Technology
M.TECH, Department of ECE, Sreenidhi Institute of Science and Technology
M.TECH, Department of ECE, Sreenidhi Institute of Science and Technology

Abstract: *Processors for next generation mobile devices will need to operate across a wide supply voltage range in order to support both high performance and high power efficiency modes of operation. However, the effects of local transistor threshold variation, already a significant issue in today's advanced Process technologies, and further exacerbated at low voltages, complicate the task of designing reliable, manufacturable systems for ultra-low voltage operation. In this paper, we describe a 4-issue VLIW DSP system-on-chip (SoC), which operates at voltages from 1.0 V down to 0.6 V. The SoC was implemented in 28 nm CMOS, using a cell library and SRAMs optimized for both high-speed and low-voltage operating points. A new statistical static timing analysis (SSTA) methodology was also used on this design, in order to more accurately model the effects of local variation $V_T$and achieve a reliable design with minimal pessimism.*

I. Introduction

Dynamic voltage and frequency scaling is an important technique for meeting the conflicting requirements for mobile devices of increasingly higher peak computational performance, longer battery life, and smaller physical size. The energy efficiency of digital circuits is maximized at very low supply voltages, near or below the transistor threshold voltage. However, operating at such voltages exacerbates the effects of both global and local variation, which are already significant issues in today's advanced process technologies such as 28 nm. Local variations, also known as intra-die variations, are the class of variations in which transistor parameters are random variables that vary independently from transistor to transistor. Local variations result from random dopant fluctuation (RDF), line edge roughness (LER) and poly grain boundaries. At low voltage, however, variations in threshold voltage resulting from RDF are the dominant effect.

SRAM and logic gate designs can be optimized to counteract this variation and ensure reliable operation at very low voltage, but generally this is at the expense of significant performance loss at high voltage. Our goal in this work is to achieve a maximally efficient low-voltage operating point, with no or minimal loss of performance at high voltage.

In this paper, we describe the design of an ultra-low-power, voltage scalable, VLIW DSP system-on-chip (SoC) implemented in 28 nm CMOS, including cell library design techniques for low voltage and high performance, the design of a high-performance/low-voltage, high density 6T SRAM, and the use of a new stochastic timing analysis and closure flow to address the effects of local variation.

A. System Architecture

As a demonstration vehicle for the design techniques de-scribed in the remainder of this paper, we implemented a DSP system-on-chip (SoC), suitable for a cellphone application processor. The DSP core itself is a 32 bit, 4-issue VLIW processor based on the Texas Instruments C64x core. Fig. 1 shows a block diagram of the processor, as well as the remainder of the SoC, which consists of L1 and L2 caches (32 kB and 128 kB, respectively), a DMA engine, and an assortment of interfaces to off-chip components. The external interfaces were selected in order to provide the functionality required to implement asimple portable music player and include I^2S, SPI, UART and multimedia card (MMC).

Our performance goals were to scale from 400MHz (to support various multimedia applications) at 1.0 V down to 20 MHz (the minimal frequency required to run an MP3 decoder algorithm) at 0.5 V. The SoC overall utilizes 600 k NAND2-equiv-alent gates, and 43 SRAM macro blocks totaling 1.6 Mb.

B. 28 nm Technology

The 28 nm technology used for this work includes a custom Poly/SiON gate stack and minimum gate length transistors of L=32nm.This technology supports a complete suite of components for low power SoC design including multi $-V_T$ core transistors,1.8 V IO and analog transistors,0.12 μm^2 6T SRAM.

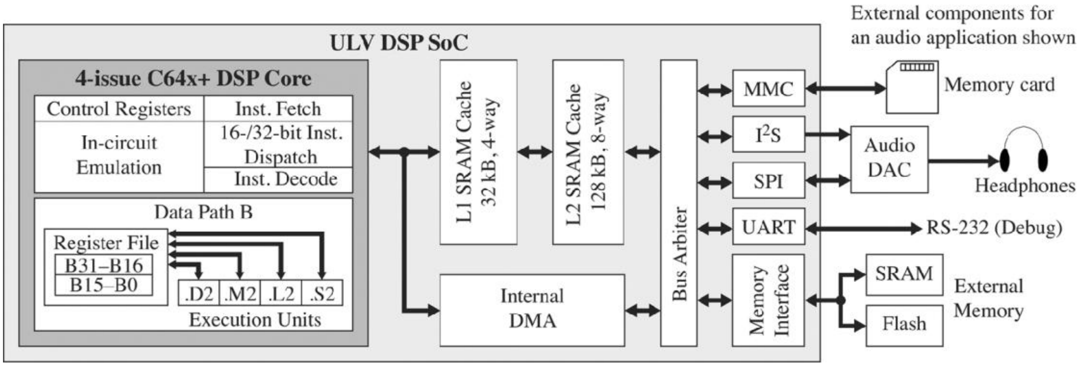


Fig. 1. SoC block diagram.

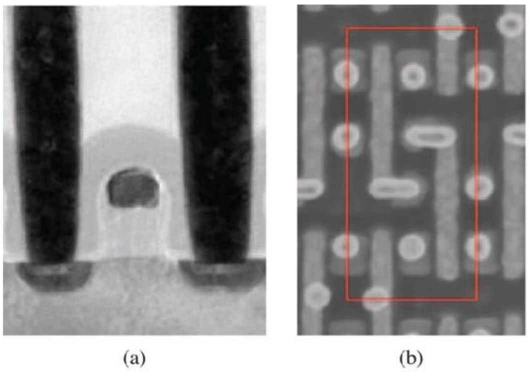
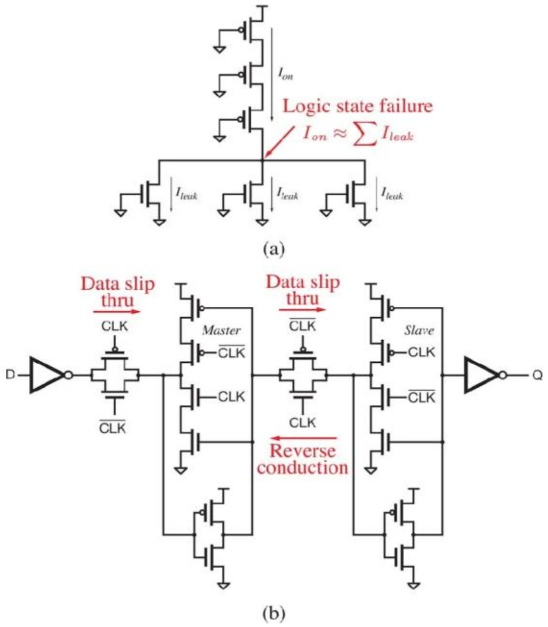


Fig.2. (a) Poly/SiON n-FET with Lg=32nm,

(b) 28 nm 0.12 μm2 6T SRAM bit cell.



n-FET transistor, and an overhead view of the bit cell. The performance gain of this technology over the 45 nm technology node is 1.3 for the SVT logic library and with the LVT logic library. 193 nm immersion lithography is used with dual-patterning at the gate level. The metallization includes up to eight dual-damascene ULK copper levels with an additional thick top aluminum level for power and signal routing.

For the logic n-FET performance attainment, emphasis was placed on optimizing the inversion oxide thickness (TINV), short-channel effect (SCE) control through halo implant and co-implants, and channel mobility enhancement through stress memorization technique (SMT) and tensile contact-etch-stop-layer (t-CESL). Mobility enhancement for the p-FET was achieved with epitaxial grown SiGe source/drains on non-rotated substrates and compressive contact-etch-stop-layer (c-CESL).

Further performance gain through TINV optimization and source-drain resistance (RSD) reduction for the p-FET device were enabled through co-optimizing source-drain (S/D) implants and S/D laser rapid thermal anneal (RTA). while SCE was controlled through co-implant

II. Standard Cell Design

To maintain sufficient reliability and performance at ultra-low voltage (ULV), a custom digital cell library was developed, with significant attention given to the effects of local variation. Performance of logic circuits operating in or near sub-threshold is highly sensitive to any variation in , and some circuits can cease to function at the extremes of V_T variations.

To quantify the functionality of each combinational cell at a given voltage, the static noise margin (SNM) is determined using the procedure described in [3] and illustrated in Fig. 4. For a maximum device stacking height of three, the NAND3 and NOR3 gates represent the worst-case examples for evaluating the input threshold voltages.

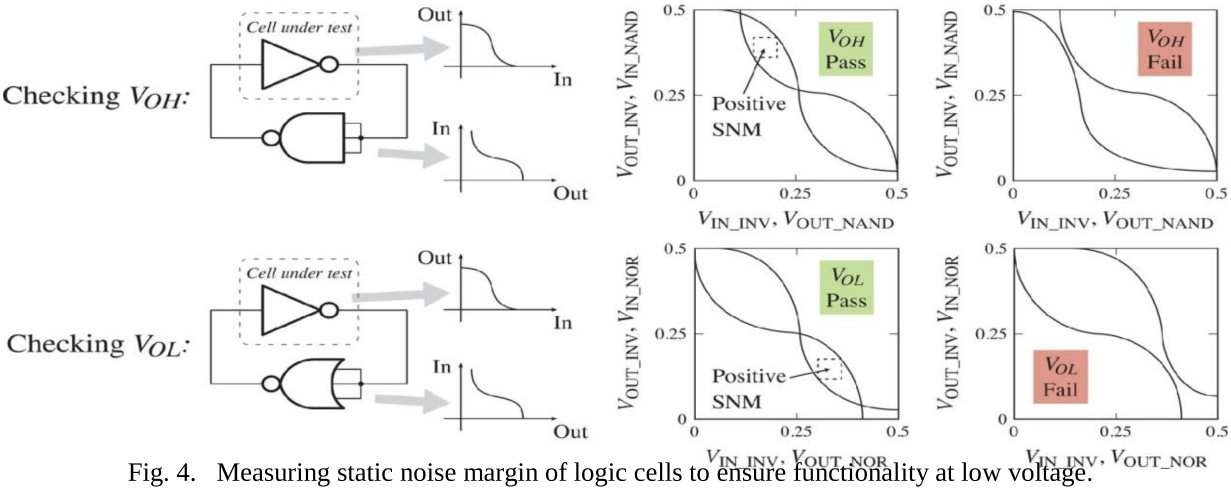


Fig. 4. Measuring static noise margin of logic cells to ensure functionality at low voltage.

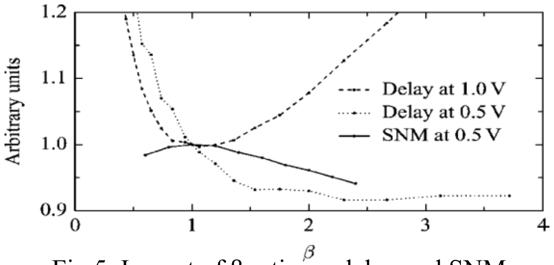


Fig.5. Impact of β ratio on delay and SNM

equivalent to constructing an infinite logic chain formed by the two cells connected alternatively. If the back-to-back cells are not functionally stable (SNM is negative), then the output of infinite chain of the two cells will eventually fail to switch. We used this SNM analysis technique to analyze our logic library and ensure low-voltage functionality. Failing cells were either modified or not used. To alleviate flipflop failures from data slip through or reverse conduction, an inverter is inserted between the master latch and pass-gate in order to delay the turn-on of the master stages and avoid reverse current flow. The static stability of the flipflops was verified using a SNM analysis similar to the one used for the combinational cells.

It illustrates the effect of β, the ratio of p-FET to n-FETwidth, on the propagation delay and SNM of a typical logic cell.

VDD=1.0V (β=1) and VDD=0.5V (β=2.4). The increase in the optimal ratio is driven by the increased difference in drive current between PMOS and NMOS transistors at ultra-low supply voltages. Because one of our design goals was to maintain near-maximal high-voltage performance, we designed our cell library using β=1.0,minimizing delay for V_{DD}=1.0V. This had the additional advantage of maximizing the SNM at low voltage.

Clock tree cells required separate consideration because equalizing rising and falling delays is more important than simply minimizing the delay. the effect of on the duty cycle of a typical clock tree. Once again, different ratios are favored for high-voltage (β=1.5) and low-voltage (β=2.0) operation, however, represents a reasonable compromise favouring high-voltage operation.

When targeting low-voltage operation, there is the temptation to only use high drive strength or large area cells so as to minimize the impact of local transistor variation. The impact of local variations at 0.6 V on the corner delay of a cell for various drive strengths.

As the transistor area increases, the impact of local variations decreases rapidly. However, the area overhead and leakage power increase resulting from upsizing low drive strength cells are significant. We chose not to eliminate low drive strength cells entirely, but rather to ensure they were only used where their increased variation sensitivity would not be problematic. First, we restricted our clock cells to drive strength or higher. This was to ensure a robust clock tree and minimize clock skew and duty cycle degradation. Secondly, we used a derate methodology during high-voltage synthesis and optimization. By derating the performance of low-drive cells we ensured that they are not used in critical paths. In this way we improve the low-voltage performance while still allowing minimum area cells when not critical. The area impact of these drive-strength increases was less than 5%.

The derating factors were determined as follows. Our design goals required on average no more than a 20 increase in propagation time when scaling from 1.0 V to 0.5 V. The minimum drive strength inverter capable of meeting this scaling require-ment, including both global and local variation. Therefore, derating factors were applied to all cells containing transistors smaller than those used in a 4 drive strength inverter. The derating factor for a given cell was calculated as the ratio of the local variation delay impact of the cell to the local variation delay impact of the 4 drive strength inverter, at the worst case global corner.

III. Low-Voltage Sram Design

The 6T bit cell has been the workhorse in modern SRAMs for many years. Because of its simple, efficient design and advantageous layout implementation, its area has been continuously scaled down for every process node providing an area-efficient memory solution. However, the 6T cell is a ratioed topology and transistor variation severely affects its operation. Moreover, the use of near-minimum size devices inside the bit cell and exacerbated effects of variation at low voltage levels cause this topology to fail as the supply voltage scales down.

SRAM designs using 8T bit cells have been proposed as low voltage solutions.

However, 8T designs occupy a larger macro area compared to 6T designs mainly because of up to 40% larger bit cell area. Although an 8T cell has decoupled read and write ports, using it in a column-interleaved architecture causes a half-select cell stability problem. Special techniques can be used to avoid this but they cause additional area overhead. Moreover, an 8T design has a single-ended read port that requires a single-ended sensing scheme and special precautions to prevent bit line (BL) leakage from affecting the output. However in 6T designs, differential BLs allow simple, small-area solutions to this problem (such as the cross-coupled PMOS devices used in this design).

In this work, a standard high-density (0.12 m) 6T bit cell is made to work at low voltage (down to 0.6 V) using area-efficient peripheral assist circuits. The principle mechanisms used to enable low voltage operation of this SRAM are 1) short local BLs, which minimize read disturbances on the bit cell, 2) word-line (WL) voltage boosting to ensure write-ability at low voltage levels and 3) large-signal sensing of the local BLs, to maximize area efficiency.

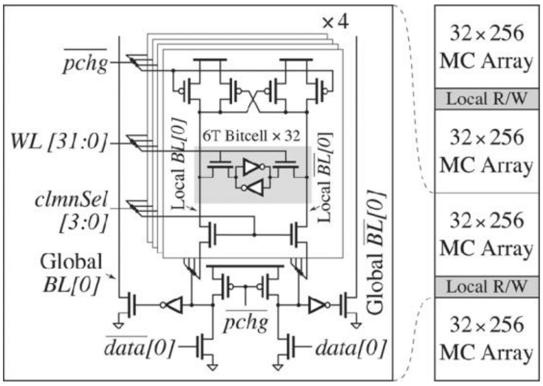


Fig. 6. Array architecture used in this design. Each local R/W circuit stripe is shared across two sub-arrays of 32 rows and 256 columns.

Stripes of local read/ write (R/W) circuitry are inserted inside the memory cell array. Each local R/W stripe is shared across two sub-arrays of 32 rows and 256 columns. These R/W circuits are coupled to the local bit-lines (BLs) of the sub-arrays. A detailed view of four columns of memory cell array and the implementation of the local R/W circuit is also during a read operation, signal development on the local BLs is sensed through a pair of inverters and pull-down devices connected to the global BLs. Similarly, for the global BLs, large signal sensing is used through cross-coupled static NAND gates. During a write operation, the local BLs are driven by one of the two NMOS devices in the R/W circuitry and the cross-coupled PMOS devices connected to the local BLs.

A. *Short Local Bit lines Reduce Read Disturbance*

Traditionally, read margin is characterized through static noise margin analysis which does not consider any dynamic effects. However, read disturbance of a bit cell involves dynamic transitions of the BLs and recent work focuses on the dynamic nature of read margin .To capture the effect of BL transitions on the bit cell stability, read margin must be modeled and analyzed through transient simulations.

To characterize read margin, two DC noise sources are placed between the cross-coupled inverters. Then their values are swept in consecutive transient simulations up to the point where a read operation alters the state of the bit cell. To reduce simulation time, a coarse-to-fine three-step search approach with 100, 10 and 1 mV step sizes is used. Results of the transient and static DC simulations are shown in Fig. 7. A smaller number of cells on a bit-line results in a faster bit-line transition and exposes the bit cell to read disturbance for a shorter period of time. This significantly improves the read margin and enables correct operation at a lower supply voltages. In this work, a limit of 32 cells/BL is necessary to ensure operation down to 0.6 V.

Although using smaller number of cells on BLs improves read margin, the resulting repetition of the local R/W circuit reduces area efficiency. To minimize area overhead, the local R/W circuit is designed carefully and has a height of only 2.4 m. Fig. 9 shows the area overhead of the local BL architecture compared to a conventional implementation with 512 cells/BL. Area overhead increases rapidly with smaller number of cellser BL. In this work, selection of 32 cells/BL provides significant improvement of read margin at the expense of 15% area overhead.

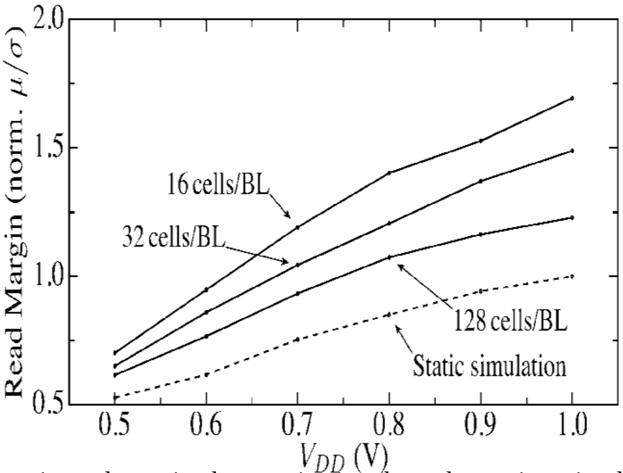


Fig. 7. Results of dynamic read margin characterization through transient simulations and the static simulation results. Smaller number of cell/BL provide significant reduction in read disturbance

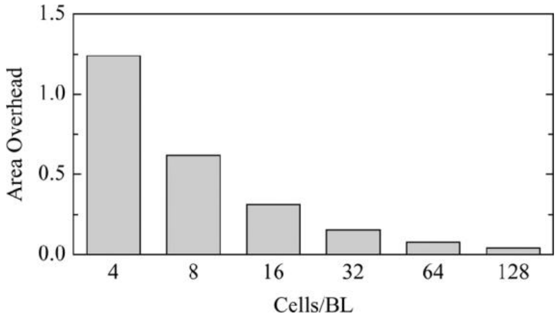


Fig. 8. Area overhead due to local BL architecture compared to a conventional architecture with 512 cells/BL. For 32 cells/BL, the area overhead is 15%.

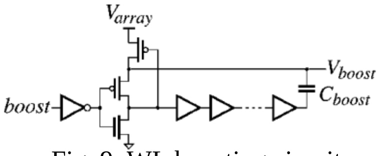


Fig. 9. WL boosting circuit.

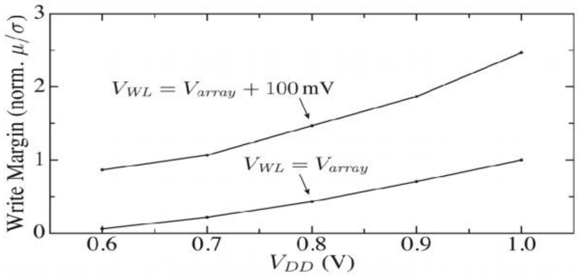


Fig. 10. Effect of WL boosting on the write margin

B. Voltage Boosting Increases Write-Ability

At low voltage levels, transistor variation causes write-ability problems. We use WL voltage boosting to overcome this issue. The voltage boosting circuit used to generate the WL voltage is shown in Fig. 9. The node is used to power up last level of buffers in WL driver circuits. During a write operation, in the first half of the cycle, the *boost* signal is kept low and the WL is first asserted then, after the negative edge of the clock, voltage boosting takes place. Triggered by the *boost* signal, the over drive of 100 mV is necessary for correct operation down to 0.6 V as shown in Fig. 10. The size of the capacitor is selected carefully with respect to the capacitive loading of the WL driver to ensure correct amount of voltage boosting.

Since only one row of a sub-array is active at any given time, the boosting circuit and capacitor can be shared across all 32 rows of a sub-array. To further reduce the area impact, the large capacitors are placed underneath the metal wiring of the address decoder. Similar shared boosting circuits are used in the column-select (CS) and data-line signal generation to achieve robust low voltage operation. For data boosting circuit, the load capacitance is smaller and can be correspondingly smaller. The area overhead due to this capacitor for the data signals is less than 4%.

Timing of the *boost* signal is crucial, as turning on the boost circuit too early will result in loss of charge stored on the capacitor and turning on the boost circuit too late can degrade the performance of the SRAM. For this design, we used the negative edge of the clock to trigger voltage boosting using inverter delay lines in the timing circuit. The pulse width of the boost signal is chosen through transient Monte-Carlo simulations on the bit cell to allow enough time for write operation to be performed. For voltage boosting circuit, the *boost* signal is sent from the timing circuit to trigger WL boosting operation. Similar *boost* signals are used for CS and data boost circuits. The charging/discharging time of the boosted signals can be overlapped with BL pre-charging time and does not require additional margin for functionality.

The power lines need to be routed such that the metal parasitic capacitances on it are minimized, as this will add to the load capacitance and affect the amount of voltage boosting. However, the lines also need to be dense enough to pre-vent IR drop when the current drawn from this node spikes.We chose to route a strap every 8rows as a balance. Fig. 11 shows a sketch of the layout of the boost circuit components.

C. Improving Read Access Time

The differential read path from the bit cell to global BLs is shown in Fig. 11. Since a large-signal development is necessary on the local BLs, NMOS switches are selected for column-multiplexing. Secondly, sensing inverters are designed to favor a low-to-high transition to speed up the signal propagation from local BLs to global BLs. This is done by designing the PMOS devices larger than the NMOS devices of the sensing inverters. Finally, differential global BLs are used to read data from the sub-arrays to maintain signal integrity even at low

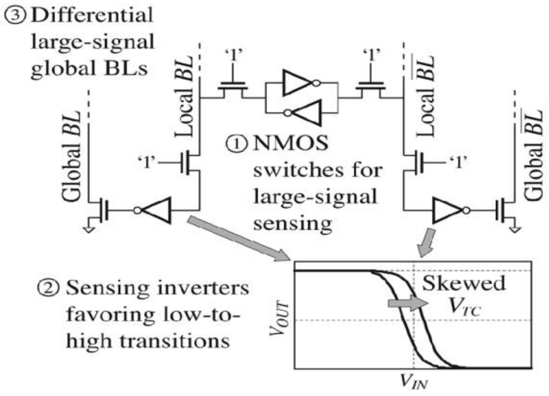

③ Differential large-signal global BLs
'1'
'1'
Local BL
Local $\overline{BL}$
'1'
'1'
Global BL
Global $\overline{BL}$
① NMOS switches for large-signal sensing
② Sensing inverters favoring low-to-high transitions
Skewed V_{TC}
V_{OUT}
V_{IN}

Fig. 11. Layout of the boost circuitry.

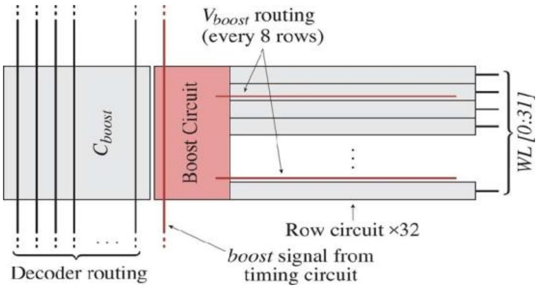

V_{boost} routing (every 8 rows)
C_{boost}
Boost Circuit
$WL\ [0:31]$
Row circuit ×32
boost signal from timing circuit
Decoder routing

Fig. 12. Differential read path from the bit cell to the global BLs.

voltage levels. The last level of sensing on the global BLs is done through a pair of cross-coupled NAND gates which are not shown in the figure.

Although the cell variation affects both hierarchical sensing and conventional small-signal schemes, in the small-signal scheme, signal propagation on the long BL (512 cells/BL) would dominate the delay, and is severely sensitive to cell variation. However, in the hierarchical sensing approach, the local BL driven by the bit-well is much shorter (32 cells/BL) and then the propagation continues through the local sense and NMOS pull-down device in the peripheral circuitry. These devices can be designed to be larger with minimal area overhead and this reduces their susceptibility to variation. Thus, hierarchical sensing approach provides better worst-case delay at low-voltage levels.

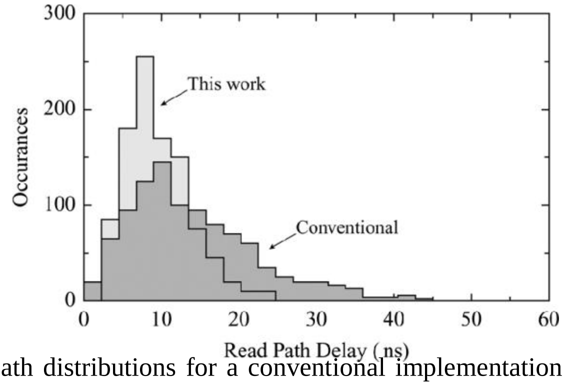


Fig. 13. Simulated read path distributions for a conventional implementation and the large-signal approach used in this design

The reference conventional implementation employs small-signal sensing with 512 cells/BL and a sense-amplifier input offset of 50 mV. At 0.6 V, distributions show that the improvements explained in this subsection result in a reduction in worst-case read access time.

The waveforms of critical signals during write and read operations. Inverter delay lines are used to create different edges from the clock. The write operation starts with the rising edge of the clock. The WL is asserted after a short delay.

IV. Conclusion

The ability to operate at ultra-low supply voltages for maximize energy efficiency will be an important advantage for next generation mobile devices. However, achieving this in a manufacturable design requires carefully addressing the issue of local variation. We have presented the cell library and SRAM design techniques for ensuring reliable low-voltage operation, as well as a new stochastic timing analysis methodology which more accurately probabilistic delays without generating prohibitive runtimes. These techniques have been validated by their use on a full-scale DSP SoC test chip, which operates down to 0.6 V.

References

[1] B. H. Calhoun and A. Chandrakasan, “Characterizing and modelingminimum energy operation for subthreshold circuits,” in *Proc.ISLPED*, 2004, pp. 90–95.

[2] B. Cheng, D. Dideban, N. Moezi, C. Miller, G. Roy, X. Wang, S. Roy,and A. Asenov, “Benchmarking statistical compact modeling strategies for capturing device intrinsic parameter fluctuations in BSIM4 and PSP,” *IEEE Design and Test of Computers*, vol. 27, no. 2, pp. 26–35, Mar. 2010.

[3] J. Kwong and A. Chandrakasan, “Variation-driven device sizing forminimum energy sub-threshold circuits,” in *Proc. ISLPED*, 2006, pp.8–13.

[4] H. Nho, P. Kolar, F. Hamzaoglu, Y. Wang, E. Karl, N. Yong-Gee, U.Bhattacharya, and K. Zhang, “A 32 nm high- metal gate SRAM with adaptive dynamic stability enhancement for low-voltage operation,” in*IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, 2010,pp. 346–347.