A Reconfigurable 8T Ultra-Dynamic Voltage Scalable (U-DVS) SRAM in 65 nm CMOS

Mahmut E. Sinangil, *Student Member, IEEE*, Naveen Verma, *Student Member, IEEE*, and Anantha P. Chandrakasan, *Fellow, IEEE*

***Abstract*—In modern ICs, the trend of integrating more on-chip memories on a die has led SRAMs to account for a large fraction of total area and energy of a chip. Therefore, designing memories with dynamic voltage scaling (DVS) capability is important since significant active as well as leakage power savings can be achieved by voltage scaling. However, optimizing circuit operation over a large voltage range is not trivial due to conflicting trade-offs of low-voltage (moderate and weak inversion) and high-voltage (strong inversion) transistor characteristics. Specifically, low-voltage operation requires various assist circuits for functionality which might severely impact high-voltage performance. Reconfigurable assist circuits provide the necessary adaptability for circuits to adjust themselves to the requirements of the voltage range that they are operating in. This paper presents a 64 kb reconfigurable SRAM fabricated in 65 nm low-power CMOS process operating from 250 mV to 1.2 V. This wide supply range was enabled by a combination of circuits optimized for both subthreshold and above-threshold regimes and by employing hardware reconfigurability. Three different write-assist schemes can be selectively enabled to provide write functionality down to very low voltage levels while preventing excessive power overhead. Two different sense-amplifiers are implemented to minimize sensing delay over a large voltage range. A prototype test chip is tested to be operational at 20 kHz with 250 mV supply and 200 MHz with 1.2 V supply. Over this range leakage power scales by more than 50 X and a minimum energy point is achieved at 0.4 V with less than 0.1 pJ/bit/access.**

***Index Terms*—Cache memories, circuit reconfigurability, dynamic voltage scaling, low-power SRAM design.**

I. Introduction

ENERGY-EFFICIENT and low-power circuit design has been an important research area in industry and academia for many years. Portable electronics, wireless sensor networks and medical implants are just a few examples of a large variety of applications that require high energy efficiency. Energy scavenging, an important area beginning to attract increasing attention over the past years, can make self-powered electronics possible. However, for most scenarios, energy harvested from the ambient is in the orders of micro-watts, necessitating the circuits to be very efficient in terms of energy consumption [1]. Hence, aggressive voltage scaling and custom designs are required in order to allocate resources effectively for the overall design.

Ultra-Dynamic Voltage Scaling (U-DVS) is an approach to reduce energy consumption by adjusting the system supply voltage over a large range depending on the performance requirement [2]. U-DVS is suitable for systems with time-varying throughput constraint. In other words, in a digital system, if the throughput constraint is cycling between different operating modes, adjusting the supply voltage for the requirements of each mode can provide significant energy savings. Fig. 1 shows a scenario for a wireless sensor node with two different operating modes: *Silent Mode* and *Transmission Mode*. Sensor node stays mostly in *silent mode* throughout its life time during which it senses and processes a certain signal (e.g., acoustic) at a very low rate. However, during short time intervals (e.g., when a speech signal is recognized) , real-time data acquisition and transmission to a host is provided by switching to *transmission mode*. For most applications, during *silent mode*, data processing can be done at low speeds allowing aggressive voltage scaling down to system minimum energy point [3]. However, to maintain the throughput requirement of real-time data processing and transmission, high-speed operation at a raised V_{DD} is required.

In modern ICs, more area is allocated to on-chip caches with every new processor generation due to the appealing features of SRAMs such as low activity factor and very high transistor density [4]. Consequently, on-chip memories often dominate total energy consumption of a chip. For the U-DVS scenario mentioned above, it is important to have the on-chip memories capable of operating on a large voltage range. Previous work in the literature mostly focused on very low-voltage operation. For example work in [5]–[8] and [9] demonstrated subthreshold SRAMs. To address the challenges of ultra-low voltages, these designs use different topologies and certain peripheral assist circuits. For example, [5], [7] and [9] use 10T SRAM cells and [8] proposes an 8T bit-cell. Although these designs achieve very low energy consumption, they cannot operate efficiently at higher voltages since they are designed by targeting subthreshold functionality. The design demonstrated in [10] uses an 8T bit-cell in a novel array architecture and can operate over a large voltage range but stays only in the strong inversion region. However, more aggressive leakage power and active energy savings are required in highly energy-constrained applications and therefore extending the supply voltage range into the subthreshold region is crucial.

An SRAM designed for operation in both subthreshold and above-threshold regions is presented in this paper [11]. Reconfigurable circuit assists are used to address ultra-low-voltage

Manuscript received March 06, 2009; revised September 02, 2009. Current version published October 23, 2009. This paper was approved by Associate Editor Peter Gillingham. This work was supported by DARPA. Chip fabrication was provided by Texas Instruments.

The authors are with the Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: sinangil@mit.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSSC.2009.2032493

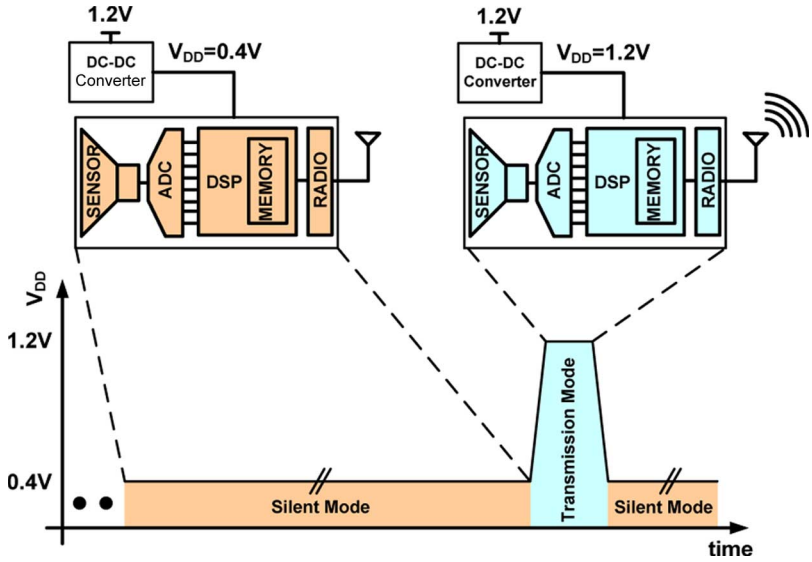


Fig. 1. A wireless sensor node with two operating modes: low-speed *Silent Mode* and high-speed *Transmission Mode*.

challenges and to minimize their adverse effect on high-voltage operation. Analysis of energy overhead due to supply voltage scaling is also included for this memory design.

This paper first discusses U-DVS design challenges mainly about SRAMs and then describes the circuits to address these issues. A suitable bit-cell design is presented, along with related read and write peripheral assists that are reconfigurable for the target $V_{\rm DD}$. Finally, analysis of energy consumption due to supply voltage scaling is discussed followed by the test chip measurement results.

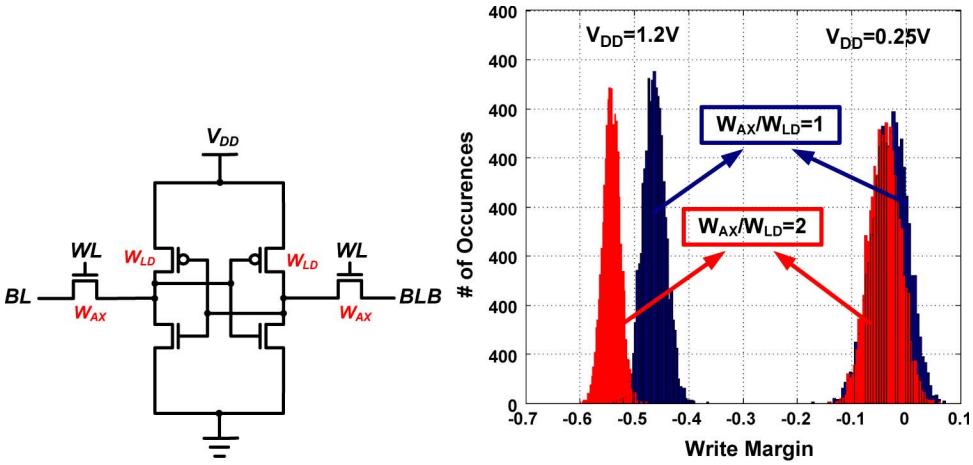
II. U-DVS Design Challenges for SRAMs

Designing U-DVS systems involves challenges at both system and circuit level. At the system level; adjusting the supply voltage involves charging and discharging the large $V_{\rm DD}$ capacitance. This raises two important considerations: i) energy overhead associated with changing the supply voltage level should at least be compensated by the energy savings in the low-voltage mode and ii) system operation should stall until the voltage transients dampens to an acceptable range. The latter problem can be solved by various techniques employing a feedback loop. However, energy overhead is a fundamental issue and cannot be avoided. Therefore, understanding the energy required to change the supply voltage level is necessary. Other system level challenges include a reconfigurable DC-DC converter design necessary to power the U-DVS system and necessity to have level converting circuits to interface with other components.

At the circuit level, U-DVS systems face the challenge of optimizing functional blocks for a wide operating voltage range. Since the minimum energy point is shown to lie in subthreshold region for most digital systems [3], extending the voltage range into deep subthreshold provides the targeted energy savings while ensuring robust operation. However, designing circuits for subthreshold functionality often requires different topologies and certain assist circuits because of i) the exacerbated effect of transistor mismatches, ii) degraded $I_{\rm on}/I_{\rm off}$ and iii) smaller noise margins at low voltages. Additionally, the solutions specific to low-voltage operation can often be suboptimal for high voltages.

The main challenge for low-voltage operation is that relative sizing of transistors is a weak knob due to the exponential dependence of drive current on threshold voltage in subthreshold region. The basic and most important building block of a traditional SRAM, 6T bit-cell [Fig. 2(a)], is a ratioed structure and its correct operation depends on relative strength of its transistors. Fig. 2(b) shows the effect of access transistor drive strength on Write Margin (WM) [12] distribution for a 6T SRAM cell at different supply voltages. WM is a metric to numerically express the write-ability of a memory cell. A negative value represents successfully over-writing the bit-cell. When the ratio of access transistor width to load transistor width W_{AX}/W_{LD} increases, access transistor can drive the internal nodes stronger, resulting in an improvement in WM. In strong inversion region, the effect is very prominent. However, at low voltages, the effect of sizing is negated by transistor mismatches. Hence, a solution relying on only transistor sizing can be suitable for high-voltage operation but is insufficient for low-voltage functionality.

Exponential dependence of drain current (I_D) to gate drive in subthreshold region can favor nontraditional sizing over traditional sizing and a U-DVS design should consider different sizing approaches. I_D of a MOS transistor is affected by its gate length primarily through the $width/length$ (W/L) ratio. Additionally, some secondary effects determining device threshold voltage (V_t) also depend on L. Reverse-short-channel-effect (RSCE), for instance, causes a reduction in V_t with longer L [13]. This effect results from non-uniform doping of the channel area to alleviate the drain-induced barrier lowering (DIBL) effect. The placement of halo doping atoms close to the source/drain areas result in an increase in V_t if the channel length is very short. At high voltages, effect of V_t reduction on I_D is neutralized by the degradation of W/L ratio. However in subthreshold region where I_D changes exponentially with V_t, effect of RSCE is strong. Additionally, longer L increases the immunity of the



(a) (b)

Fig. 2. (a) 6T SRAM cell and (b) the effect of access transistor sizing on Write Margin (WM) distribution for a 6T SRAM cell at different supply voltages. $W_{AX}/W_{LD} = 2$ provides significant improvement at 1.2 V, whereas at 0.25 V, the effect of sizing is very small.

transistor against mismatches since the standard deviation of V_t varies inversely with the square root of the channel area [14]. Consequently, at low voltages, with longer L i) the standard deviation of V_t across transistors is lower because of larger channel area and ii) the mean of transistor V_t is smaller due to RSCE. Hence contrary to the traditional sizing, longer devices provide better performance over a large voltage range.

Another challenge related to U-DVS SRAM design is that the peripheral assist circuits can impose difficulties at the architectural level. For example a dual-$V_{\rm DD}$ SRAM would face the problem of allocating a fixed number of metal tracks to two different supply voltages (array-$V_{\rm DD}$ and periphery-$V_{\rm DD}$) instead of only one as in the case of a single-$V_{\rm DD}$ SRAM. This not only increases the complexity of the power grid in a complicated system but also adversely affects the supply voltage noise due to IR drop. Another example is the virtual-$V_{\rm DD}$ $(VV_{\rm DD})$ scheme for the memory cell array proposed in [5]. Row-wise $VV_{\rm DD}$ is not shared between adjacent rows in contrast to the traditional design where $V_{\rm DD}$ is shared across the array. Hence, $VV_{\rm DD}$ metal cannot be routed at the edge of the memory cell. To avoid excessive area overhead, a solution is proposed where each row is folded into two to route $VV_{\rm DD}$ in between the folded segments (Fig. 3). A similar approach is used in our design for row-wise signal routing in layout. The problem and proposed solution will be discussed in more detail in the subsequent sections.

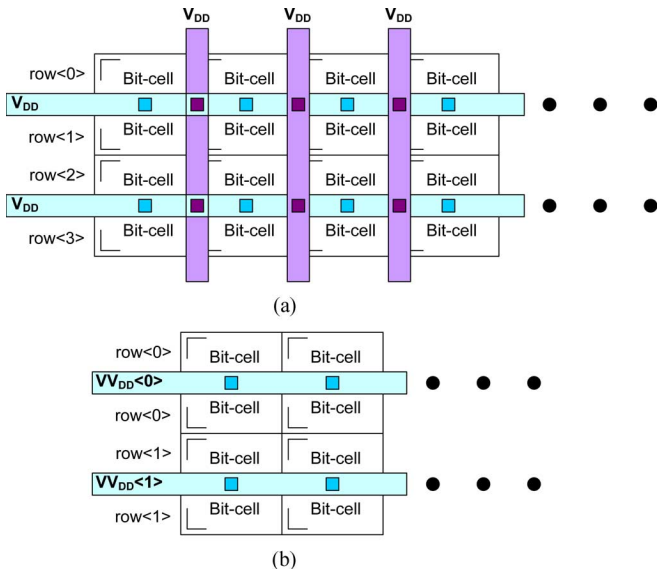


Fig. 3. (a) Traditional array architecture for shared $V_{\rm DD}$ across rows. (b) Folded-row architecture for row-wise $VV_{\rm DD}$ proposed in [5].

III. U-DVS SRAM Design

A. 8T Bit-Cell Design

The bit-cell used in this design is shown in Fig. 4(a) [8]. This topology, originally proposed for a subthreshold SRAM, is optimized for functionality and performance over a large voltage range in this design. Two nMOS devices (MN5 and MN6) constitute the read-buffer. A write operation is performed through WL, BL and BLB ports whereas single-ended read operation is exercised through RDWL and RDBL ports. RDBL is pre-charged at the end of each read cycle and kept pre-charged during a write cycle. In this bit-cell, read and write ports are decoupled in contrast to the traditional 6T cell so i) read-SNM (RSNM) problem is eliminated, ii) 6T part (MN1-4 and MP1-2) can be sized for better write-ability without trading-off RSNM and iii) 2T read-buffer can be sized for larger read-current (I_{READ}) independently. Fig. 4(b) shows the orientation of bit-cell transistors in layout implementation. In layout, rows are folded which enables sharing of MCHd, BVSS, WL and RDWL metal routing at the shared edge of the bit-cell. Five metal layers are used in bit-cell layout where WL is routed in second, BL/BLB and RDBL are routed in third and RDWL is routed in fifth metal layer.

BVSS is a virtual ground node for the read-buffer and is kept at $V_{\rm DD}$ if a memory cell is not accessed. This makes the voltage drop across unaccessed read-buffers zero and hence leakage on read-bit-lines $(RDBL)$ is highly reduced. MCHd is the virtual supply node for the cross-coupled inverters and its voltage can be brought down during a write access to weaken pMOS load

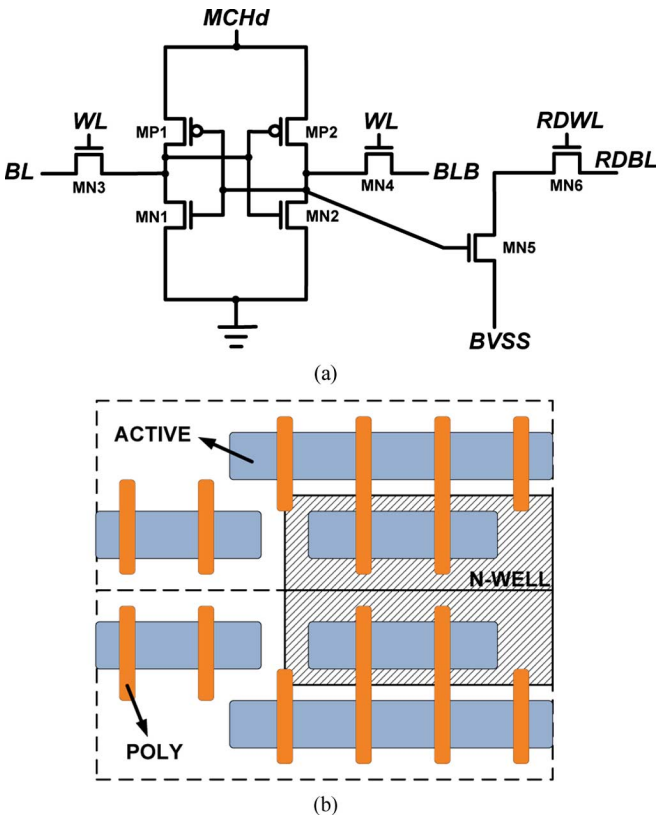


Fig. 4. (a) 8T SRAM cell used in this design. (b) Orientation of transistors in layout for two bit-cells.

devices (MP1 and MP2) and ease write-ability problem at low voltages. Since all the bit-cells on a row are written and read at the same time, MCHd is shared across one row of memory cells.

Performance of an SRAM heavily depends on the time it takes for the cell I_{READ} to discharge the bit-line by an amount that is larger than the sense-amplifier offset. Therefore, I_{READ} of the worst case cell in the array determines the maximum performance that a memory can run at. Conventionally, read-buffer devices are designed as wide transistors with minimum L. However, for this U-DVS design, second order effects that begin to be more prominent at low voltage levels are also considered carefully and RSCE is exploited to improve the drive-strength of read-buffer devices at low voltages by increasing their gate lengths.

Fig. 5 shows the effect of L on 4σ read-buffer I_{READ}. At high voltages, longer L results in a slight degradation in I_{READ} whereas at low voltages, because of the RSCE and reduced variation, longer L provides considerable improvement in 4σ I_{READ}. For this design, L for read-buffer devices are chosen to be approximately twice the minimum size to achieve 7 X I_{READ} improvement at 250 mV which causes less than 10% degradation at 1.2 V.

B. Reconfigurable Write Assist Scheme

Although the bit-cell is sized for improved write-ability, at low voltages, effect of sizing is negated by transistor mismatches. Fig. 6 shows the WM distribution of the bit-cell used in this design at three different supply voltage levels. As the supply voltage scales down, two trends are observed for WM

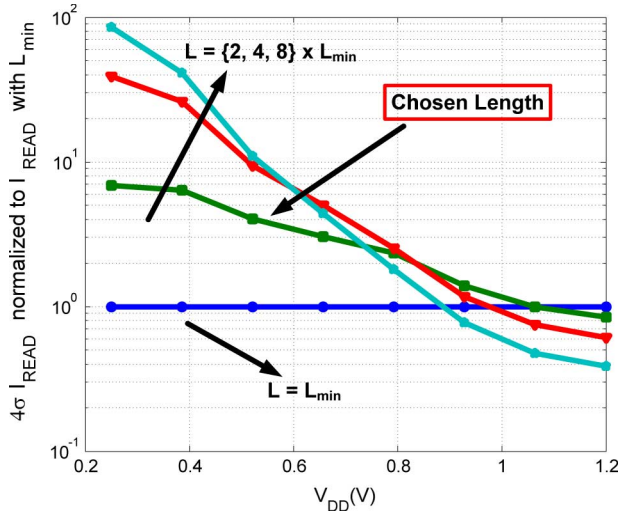


Fig. 5. Effect of read-buffer L on 4σ I_{READ}. At high voltages, longer L slightly degrades performance whereas at low voltages, a prominent improvement is observed with longer L.

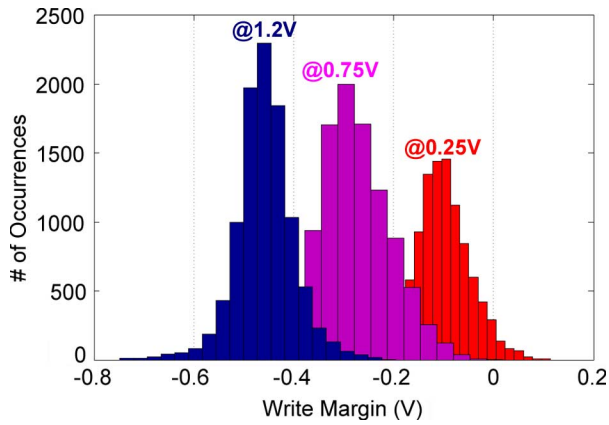


Fig. 6. WM distribution at three different supply voltages (1.2 V, 0.75 V, and 0.25 V) suggests that a reconfigurable write-assist scheme should be used.

distributions: i) mean of the distributions shifts to the right and ii) distributions become wider. Both trends cause the tail of the distribution to get closer to the zero point, and at around 0.7 V, write failures begin to emerge. At 0.25 V, a significant portion of the cells cannot be written using the conventional technique because of the severe degradation of the WM.

To provide robust operation at low voltages, a write-assist scheme is necessary. However, it is very important to consider the overheads associated with the assist circuits in terms of area, power and performance on the entire voltage range. Subthreshold designs are motivated by lower power and energy efficiency whereas above-threshold designs generally target better performance and smallest area. For our design, assist circuitry is designed with circuit reconfigurability which consumes a small area and causes negligible power overhead but provides the necessary robustness and adaptability over the entire voltage range.

Lowering memory cell $V_{\rm DD}$ during a write operation improves write-ability of the cell [8]. However, during an access, this creates a short-circuit current path from the BL drivers to the MCHd driver as shown in Fig. 7(a). At low voltages where

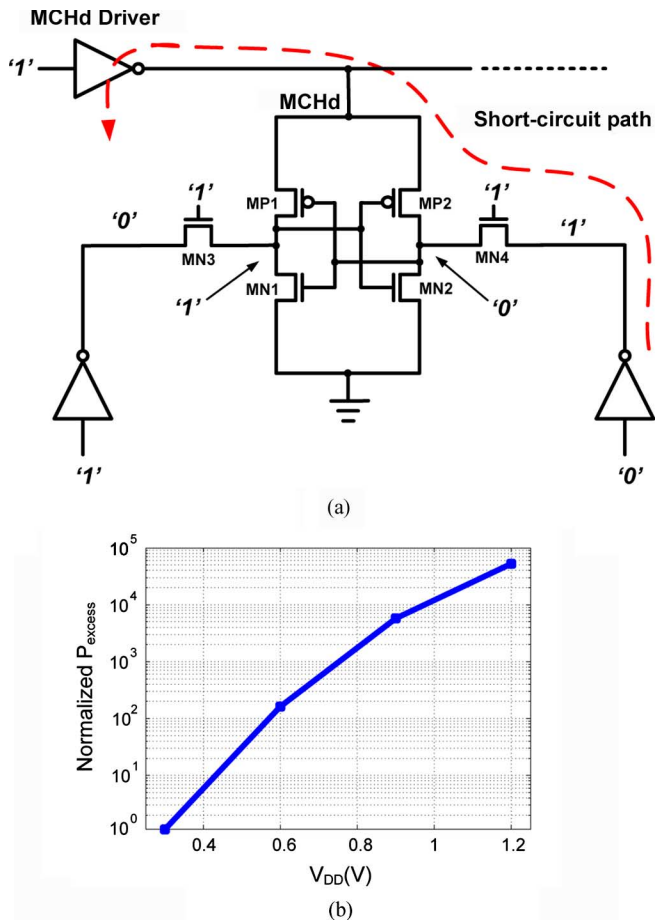


Fig. 7. (a) Short-circuit current path. (b) Power consumption resulting from this path.

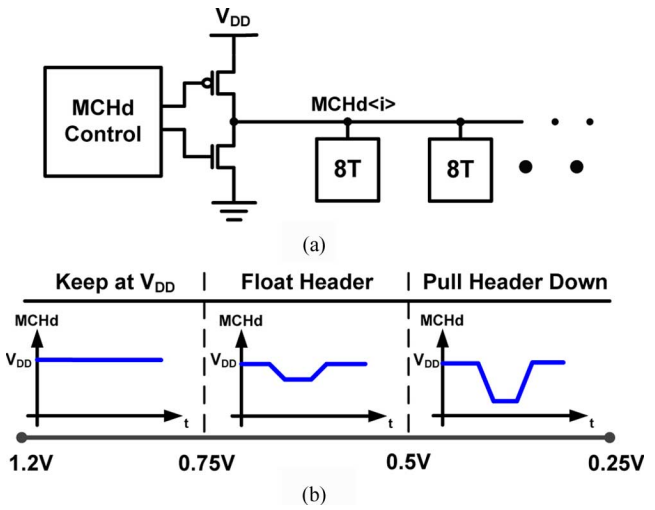


Fig. 8. (a) MCHd driver. (b) Three different write assist schemes used in the U-DVS SRAM.

drive currents are comparable to the leakage currents, this short-circuit current is negligible since leakage of the cell array dominates the power consumption. However, at high voltages, as shown in Fig. 7(b), this current can account for a significant amount of power consumption on the orders of milliWatts. Quantitatively, ratio of short-circuit power consumption to array power is 6% at 1.2 V whereas this ratio is 1% at 0.3 V. To address this problem, a reconfigurable write-assist scheme is implemented with a simple control circuit and a driver as shown in Fig. 8(a). MCHd is a row-wise virtual supply node and connected to all memory cells on the same row which will be accessed during a write. In layout, row of memory cells are folded and MCHd metal is routed between the folded segments [5]. The area overhead associated with this assist circuitry is less than 6%.

Depending on the operating voltage range, reconfigurable write-assist circuit activates one of the three different schemes shown in Fig. 8(a). At high voltage levels, memory cells have enough write margin to operate correctly so MCHd is kept at $V_{\rm DD}$. This avoids unnecessary switching in the control and row circuitry and avoids power consumption associated with them. As the supply voltage scales down, memory cells at the tail of the distribution begin to show write failures but a slight WM improvement can enable these cells to be overwritten successfully. This slight improvement can be provided by keeping MCHd node floating during a write-access. A fixed amount of charge on MCHd is shared among the bit-cells on the same row and every pMOS transistor trying to fight its counterpart access transistor causes MCHd voltage to droop. This droop helps access transistors to overpower PMOSs more easily. At very low supply voltages close to or inside the subthreshold region, WM needs to be further improved. Floating MCHd node can provide this improvement but at low voltage levels where drive currents are comparable to leakage currents, voltage droop on MCHd requires significantly larger amount of time. Hence, MCHd is actively pulled down bringing its voltage close to 0 V. Under this condition, access transistors can easily flip the cell and enable functionality even in subthreshold region. Drivers for MCHd nodes are sized to prevent a performance penalty during write acceses. MCHd voltage is actively pulled-up in all schemes before WL goes low. This ensures that feedback inside the bit-cell is established again and internal nodes are fully charged or discharged to the supply voltage levels before the end of the cycle.

Fig. 9 shows the simulation waveforms of the write assist scheme at 0.3 V and 50 kHz. A write operation is followed by a read operation. WLenable signal triggers the RDWL and WL assertion whereas WrAsstB signal triggers the selected write assist on the active row. After WL assertion, WrAsstB signal is pulled low which causes MCHd node to be actively pulled down. If a different assist scheme is selected, MCHd is kept floating or held at $V_{\rm DD}$. During a read operation, MCHd is always actively pulled high to provide correct operation. Finally, signals for write assist timing are supplied from an off-chip pattern generator.

C. BVSS Driver Design

Small gate drive at low voltages results in low on-current-to-off-current (I_{ON}/I_{OFF}) ratio for transistors. Moreover, in the presence of local transistor mismatches, worst case I_{ON} can be comparable to I_{OFF}. This imposes a bottleneck in terms of signal development on BLs and sensing for SRAMs since it becomes impossible to distinguish a valid discharging from

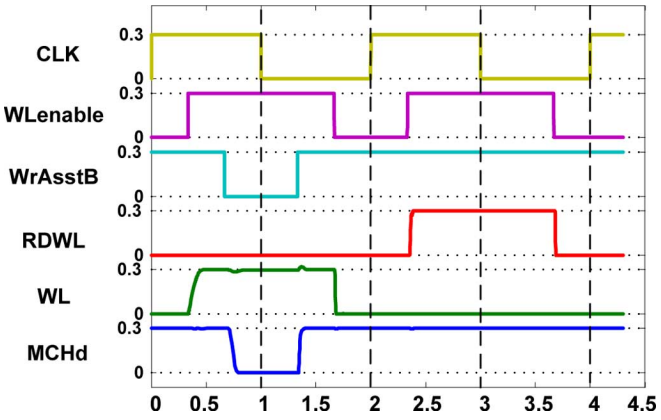


Fig. 9. Waveforms associated with the write assist scheme during a write followed by a read cycle.

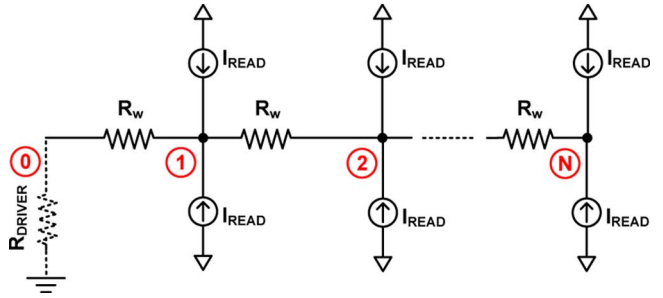


Fig. 10. First order analysis of the IR drop across BVSS node. R_w is the wiring resistance and read-buffers are modeled as ideal current sources.

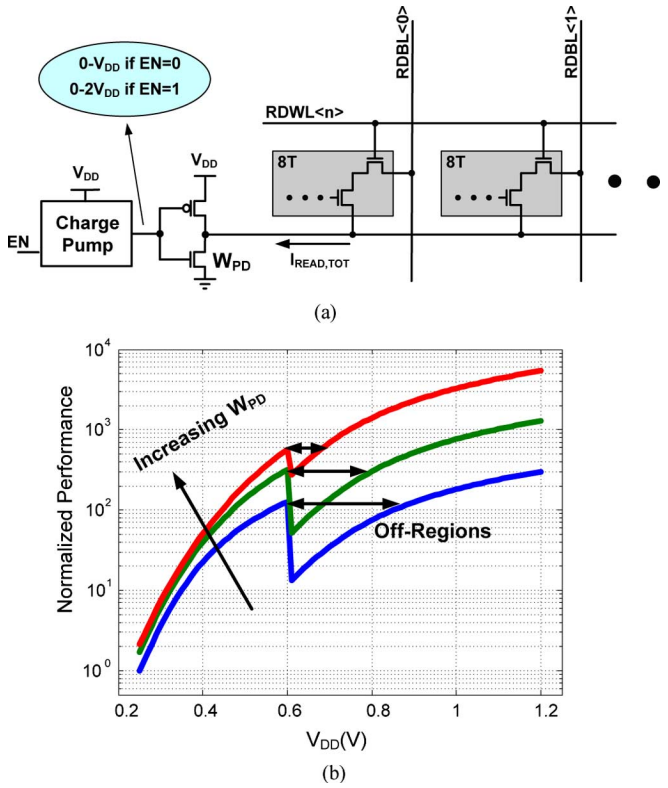


Fig. 11. BVSS driver (a) and performance of the U-DVS SRAM for different widths (W) of the pull-down device of this driver (b).

voltage droop due to leakage. This problem can be addressed in two different ways: i) by increasing I_{READ} through up-sizing or voltage-boosting or ii) by lowering the leakage on BLs. Designs in [5], [8] and [7] address this problem by using the latter option. Specifically, works in [5] and [7] add transistors inside each bit-cell to limit the leakage from BLs whereas [8] and our work implements a solution in the periphery.

Fig. 11(a) shows the scheme used to limit BL leakage problem. If a row of memory cells is not accessed, footer of their read-buffers (BVSS) are pulled-up through the BVSS driver. This makes the voltage drop across read-buffer devices close to 0 V and consequently reduces the leakage from RDBLs to unaccessed rows. In contrast, if a row is accessed, footer node should be pulled-down quickly and stay low while the BVSS driver sinks aggregated current from every read-buffer on its row. The design in [8] uses a charge-pump circuit to boost the gate-drive of the pull-down path in BVSS driver and acquire a gain of ~500X in strength while operating at 350 mV. In our design, however, a comparable gain cannot be achieved for higher voltages since doubling gate-drive translates into exponential current increase only in subthreshold region. Moreover, above $V_{\rm DD} = 0.6$ V, due to reliability concerns, charge-pump circuit cannot be activated since, in this case, the output of charge-pump will exceed the process nominal of 1.2 V.

Fig. 11(b) shows the performance versus $V_{\rm DD}$ plot for different BVSS driver nMOS widths (W_{PD}). Disabling the charge-pump circuit causes a suddent drop in the performance at $V_{\rm DD} = 0.6$ V. Increasing W_{PD} results in continuous performance increase but also results in larger leakage power. In this work, W_{PD} is up-sized by ~ 10 X to make the off-region small enough that a continuous performance versus $V_{\rm DD}$ response can be acquired from the design.

A related problem arises from the layout of the row-wise BVSS signals. Since the aggregated drive currents from all columns flow through the BVSS node as shown in Fig. 11(a), the IR drop on this metal line can be very high especially at $V_{\rm DD} = 1.2$ V. Fig. 10 shows a first order analysis of the effect of wiring resistance on the internal node voltages. To share the BVSS node across the rows effectively, a folded-row layout similar to the one explained in Section II is implemented. Each read-buffer is modeled as a current source for simplicity. Assuming the resistance of the BVSS driver is small, node voltage at 0 is approximately 0 V. Finally, the worst-case situation where all read-buffers are discharging RDBLs (all memory cells holding a logic 0) should be considered. Then the voltage at node N is

$$V_N = N(N+1) \times I_{READ} \times R_W.$$

For this design, a word-length of 128 $(N = 64)$ is used. Hence, the voltage at the source of the right-most read-buffer is $V_{64} = 4032 \times I_{READ} \times R_W$ and can cause a serious voltage drop. In reality, read-buffers are not ideal current sources so if the voltage increases at the BVSS node, it will cause I_{READ} to get smaller since the V_{GS} of read-buffer transistors are directly

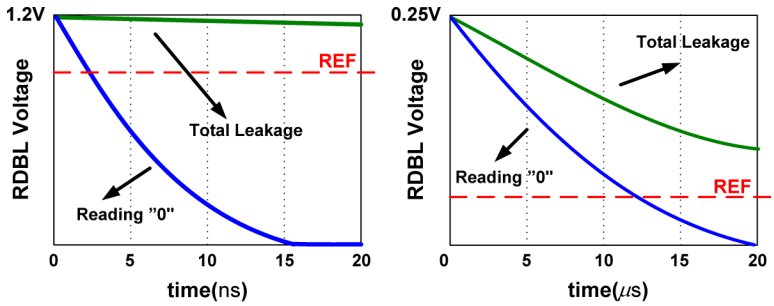


Fig. 12. RDBL voltage development during a read access (a) in high-voltage and high-performance mode and (b) in low-voltage low-performance mode.

affected by this. The IR drop on BVSS node can cause a severe degradation of I_{READ} and consequently a loss in SRAM performance. To address this problem, the resistance of row-wise BVSS is minimized in the layout. First, two metal layers are assigned to this signal with multiple vias connecting these layers. Second, layout of the bit-cell is done by considering this problem and allows wide BVSS metal routing.

D. Reconfigurable Sensing Network

The sensing network is an important part in SRAM design. Since a sense-amplifier lies in the critical path of a read-cycle, in some SRAM designs, sense-amplifier delay directly adds up to the total read delay and determines the maximum attainable performance.

In SRAMs, sensing can be single-ended or differential depending on the array architecture and bit-cell topology. For a 6T cell, conventionally, BL and BLB are both pre-charged at the beginning of a read-cycle and then during the evaluation period, a voltage differential develops between these two ports. Finally, sensing network amplifies this differential voltage and outputs the result. The work in [15], in contrast, uses an asymmetric 6T cell and employs a single-ended sensing scheme with fine-grained bit-line segmentation. Shorter bit-lines and improved I_{READ} through asymmetrically up-sizing bit-cell transistors are the key ideas to enhance memory performance and bit-cell stability with minimum area overhead. For 8T cells, read operation is inherently single-sided and an appropriate sensing scheme is required.

Another important issue concerns the amount of signal development on the BLs to do a successful sensing. Small-signal and large-signal sensing schemes are described in [16]. Small-signal sensing often requires a rigorous sense-amplifier structure to catch a small differential slowly developing on long BLs. However, in large-signal sensing, a simple logic gate can be used to resolve the voltage on shorter BLs. Hence, the choice of sensing scheme can be different depending on the array architecture and area optimization of each design.

In our design, three different voltage levels can be observed on the RDBLs so a sensing scheme capable of differentiating between them is necessary. First, a logic 0 causes pre-charged RDBLs to be discharged to 0 V through a read-buffer [Fig. 12(a)]. In contrast, a logic 1 inside the cell leaves RDBLs floating during access. For high-speed operation at high voltages, RDBLs stay close to the pre-charged value

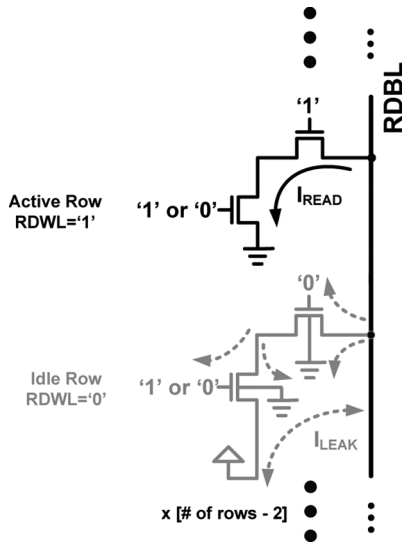


Fig. 13. Schematic showing the gate, junction and subthreshold leakage paths from RDBL to unaccessed rows.

of V_{DD} [Fig. 12(a)] since the access periods are short and the effect of leakage is very small. However, at ultra-low voltages, very long cycles cause the leakage to discharge RDBLs. Due to the BVSS scheme used in this design as shown in Fig. 13, RDBLs are not discharged to 0 V but their voltage droops to a medium level due to reverse leakage from BVSS [Fig. 12(b)]. Hence, for our U-DVS SRAM, the sensing network should be capable of resolving these different conditions explained above. Moreover, the sensing delay should also be considered carefully since, over a large voltage range, one topology might not provide an optimum solution.

For the U-DVS SRAM, a reconfigurable sensing network is designed to optimize sensing delay over the entire voltage range (Fig. 14). Two similar implementations of a widely used sense-amplifier [17] are used in our design along with simple selection logic. RDBL is fed to one of the inputs of both sense-amplifiers and a global off-chip reference signal (REF) is connected to the second input. Since the NMOS-input structure utilizes nMOS transistors for the differential pair, inputs with a higher common-mode voltage results in faster resolution of the outputs. In contrast, PMOS-input sense-amplifier is faster if the input common-mode is closer to ground. At high voltage levels, as shown in Fig. 12(a), charge stored on RDBL capacitance cannot

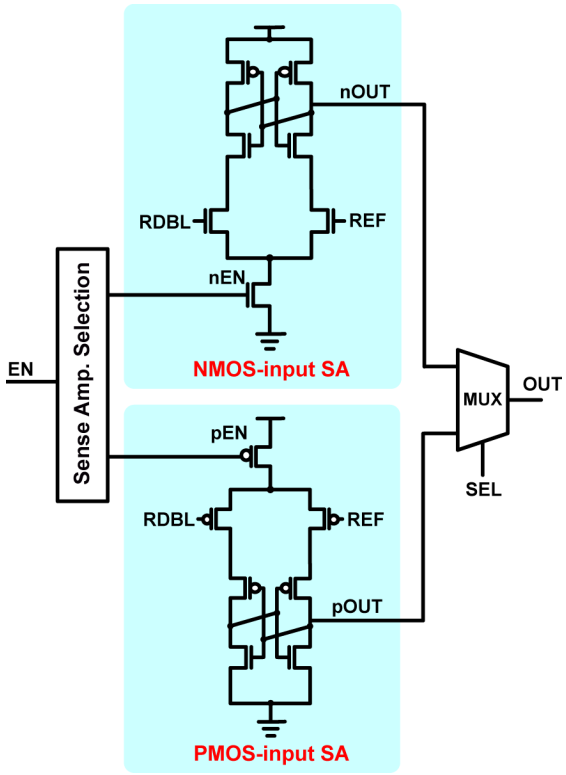


Fig. 14. Reconfigurable sensing network employing two sense-amplifiers (NMOS-input and PMOS-input) and a selection logic. Pre-charge and pre-discharge transistors necessary for sense-amplifier operation are not shown in the figure for simplicity.

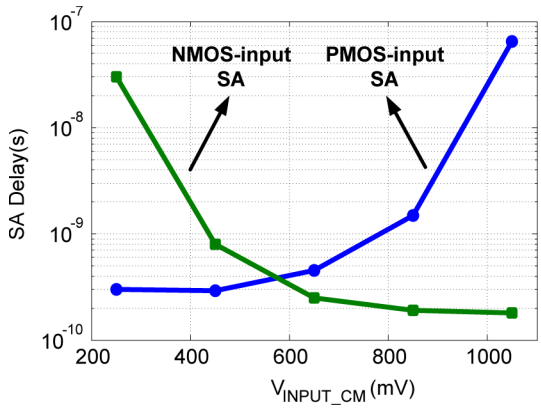


Fig. 15. Delay of the sense-amplifiers used in the design for different input common-mode levels.

be discharged through leakage mechanisms during the short access cycles so REF can be close to $V_{\rm DD}$. Hence, common-mode of the input signals are larger in this case and using NMOS-input structure provides smaller delays. In contrast, at low voltages, REF should be placed closer to 0 V to avoid erroneous readings in the case of long access cycles causing RDBL voltage to droop to a medium voltage [Fig. 12(b)]. So, for low voltages, common-mode of input signals are closer to 0 V and PMOS-input sense-amplifier provides a smaller sensing delay. Fig. 15 shows the delay of both sense-amplifiers for different common-mode voltages. Using a reconfigurable sensing network enables significant delay improvement over a large voltage range. Both versions of the sense-amplifiers are sized to consume the same area. The area overhead of the second sense-amplifier and the selection logic is less than 10%.

IV. Analysis of Energy Consumption Due to Power-Supply Scaling in a U-DVS Scenario

Dynamic voltage scaling involves charging and discharging of capacitances associated with the power supply node. Moreover, power grid in modern chips are designed to be very dense causing this capacitance and consequently the energy consumed during voltage scaling to be larger. In other words, voltage scaling provides energy savings at the expense of this energy consumption. The main benefit of power supply scaling is attributed to the quadratic savings in active energy given by the well-known formula $E = C \times V_{\rm DD}^2$. Moreover, leakage power is also reduced at low voltages not only because of smaller potential difference across devices but also due to second order effects determining device V_t such as drain-induced barrier lowering (DIBL). In this section, the energy consumption and savings associated with supply voltage scaling is analyzed for this work.

In this analysis, a scenario with system imposed time-varying performance requirements will be analyzed. For simplicity, only two discrete performance modes will be assumed: *high-performance* (200 MHz) and *low-power* (500 kHz). Next, two different power saving schemes will be considered: i) no voltage scaling and ii) ultra-dynamic voltage scaling (Fig. 16). First scheme provides operation at the voltage level satisfying the high-performance mode all the time and avoid the energy overhead of supply voltage scaling. Second scheme, in contrast, utilizes U-DVS operation to dynamically adjust the supply voltage to the time-varying performance requirements. To avoid differences in topology and architecture of designs, we will consider the energy efficiency of our U-DVS design under these different power saving schemes. For the U-DVS SRAM, performance requirements of 200 MHz and 500 kHz are satisfied at 1.2 V and 0.4 V, respectively. So the first scheme operates at 1.2 V all the time, whereas the second scheme switches between 1.2 V and 0.4 V depending on the operation mode.

To calculate the energy overhead associated with the supply voltage scaling, the amount of capacitance connected to the supply node needs to be determined. The contributors to this capacitance are: i) pMOS transistor gate-to-drain capacitance (C_{gd}), ii) NWELL diode capacitance (C_{NWELL}) assuming NWELL body terminals are shorted to $V_{\rm DD}$, iii) wiring capacitance (C_{WIRE}) and iv) decoupling capacitors (C_{DC}). For our design and most SRAMs, every bit-cell has two pMOS transistors and an NWELL strip. Hence, the capacity of the memory has a direct effect on C_{gd} and C_{NWELL} capacitances. Source/Drain areas for each pMOS transistor as well as NWELL area can be calculated from the bit-cell layout. Peripheral circuits (WL drivers, BL drivers, sensing network and address decoder) also contribute to capacitance of the supply node. For wiring capacitance, a parasitic extraction is necessary since it is not a straightforward analysis to calculate C_{WIRE} of the power grid. Finally, allocating around 10% of

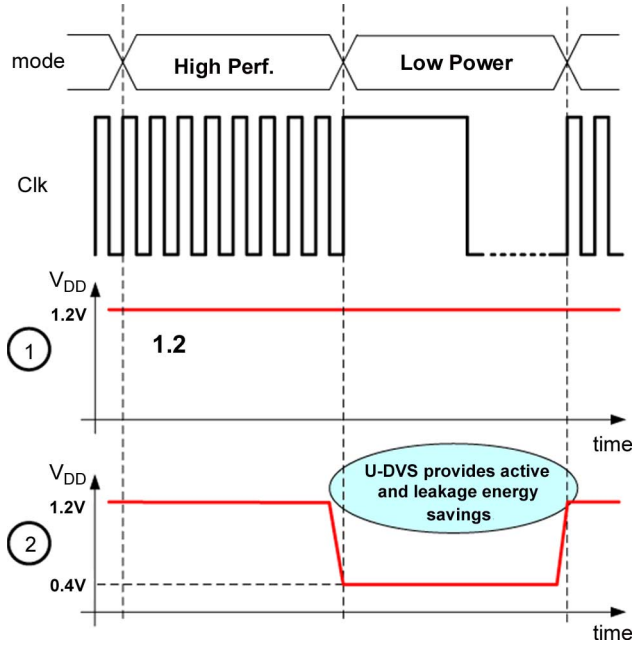


Fig. 16. Two different power saving schemes analyzed: (a) no voltage scaling and (b) ultra-dynamic voltage scaling.

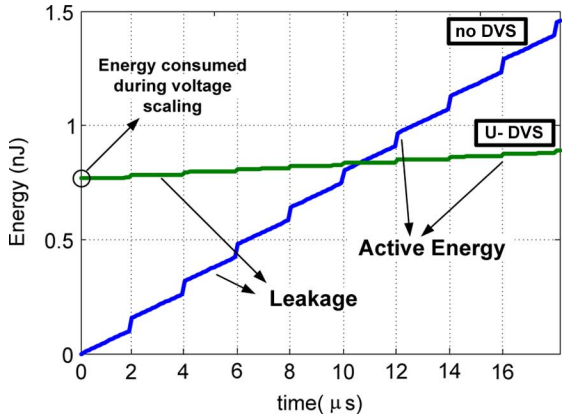


Fig. 17. Comparison of simulated energy consumption during low-power mode with no-DVS and U-DVS power saving schemes. Both schemes have the same energy consumption after 10 μs which is the break-even time for this design.

the active area to decoupling capacitors is a rule of thumb used in the industry [18], [19], which is also used in this analysis.

Fig. 17 shows the comparison of energy consumptions of two SRAMs operating with and without ultra-dynamic voltage scaling. There is an energy overhead associated with supply voltage scaling which corresponds to the y-intercept of the second (U-DVS) scheme whereas first scheme starts from the origin. A memory access (happening every 2 μs) causes active energy consumption and appears as a jump in the figure. Between accesses, however, leakage determines the slope. 40X leakage power scaling occurs from 1.2 V to 0.4 V which accounts for the important difference between two schemes. Energy consumption in both schemes become equal after five accesses (or 10 μs) which is the break-even time for this SRAM. If the system stays in the low-power mode longer than 10 μs, it is more advantageous to apply U-DVS since active and leakage energy savings provide better energy efficiency beyond this point.

V. Test Chip Measurement Results

A 64 kbit 65 nm CMOS test chip is designed to demonstrate the ideas discussed above. The array architecture is shown in Fig. 18(a). The design consists of eight 8 kbit blocks, where each block contains 64 rows and 128 columns of bit-cells along with row and column peripheral circuits. A single data input/output (DIO) bus is connected to each block since only a read or write operation is done during an access. MCHd Driver, BVSS Driver and WL Driver constitute the row circuitry whereas column circuitry contains the sensing network and BL/BLB drivers. PMOS devices are used to pre-charge RDBLs and they are controlled by pchgB signal. Finally, Fig. 18(b) shows the chip micrograph of the test chip fabricated in 65 nm low-power CMOS process. The die size is 1.4 mm by 1 mm.

Fabricated test chips achieve read and write functionality from 1.2 V down to 0.25 V. This large voltage range is enabled with the proposed reconfigurable assist circuits. Fig. 19 shows the frequency and leakage power versus $V_{\rm DD}$ plot for the design. Memory operates from 20 kHz to 200 MHz over the voltage range. Very large frequency range makes this design compatible with numerous low-power applications. Leakage power scales down by ~50X as shown in Fig. 19 over the same voltage range.

Fig. 20 plots the active and leakage components of *energy/access* along with the sum of these two components. Active energy scales down quadratically as expected. Leakage energy increases with decreasing supply voltage. This is due to an increase in access period at low voltages which causes the leakage power to be integrated over a longer period of time. Since the drive strengths of the transistors are exponentially dependent on the gate drive in sub-V_t regime, the performance of the memory also degrades exponentially causing an exponential increase in the leakage energy. Total energy makes a minimum around 400 mV with less than 11 pJ/access. This value corresponds to less than 0.1 pJ/bit/access since 128 bit words are accessed at every cycle. Measurement results show a good match with the simulations.

Twenty-five chips were tested and all proved functional down to 0.3 V. Only two chips failed to operate down to 0.25 V due to sense-amplifier failures at this voltage level. Raising sense-amplifier supply voltages by 50 mV enabled these failing chips to operate correctly at 300 mV.

VI. Conclusion

Ultra-dynamic voltage scaling is an important technique to reduce energy consumption of circuits under time-varying performance constrained scenarios. SRAMs occupy an important portion of total area and energy in modern ICs and hence it is especially important to design voltage scalable memories. To enable highest possible energy savings and a wide performance range, it is crucial to design memory circuits that can do voltage scaling from deep subthreshold region to full-$V_{\rm DD}$ levels. However optimizing circuits for a large voltage range is challenging

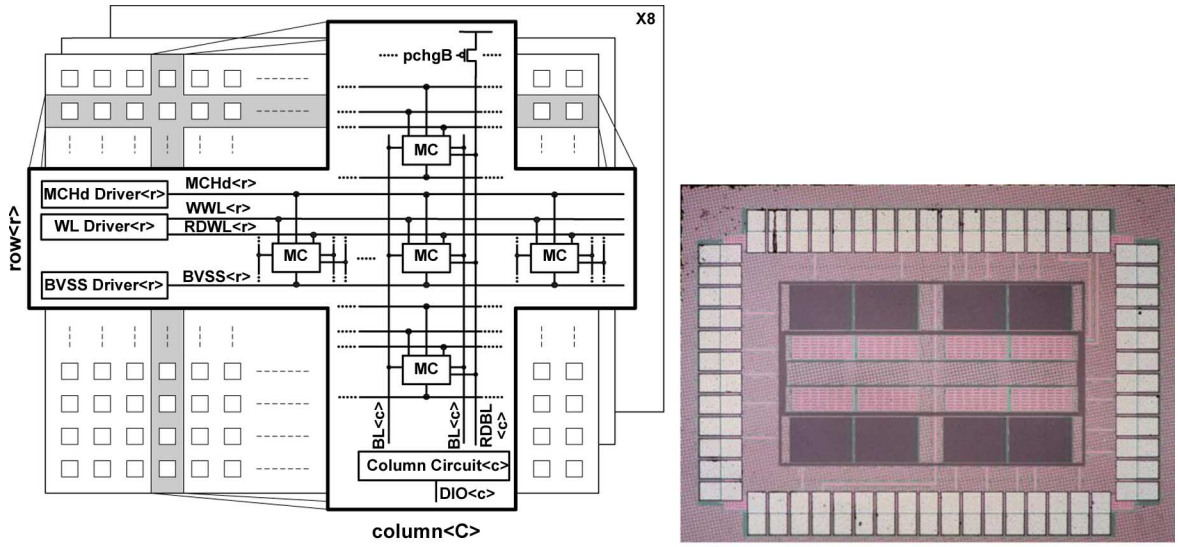


Fig. 18. (a) Architectural diagram and (b) die photo of the test chip fabricated in 65 nm low-power CMOS process. Each array consists of 8 kbit memory cells composed of 64 rows and 128 columns.

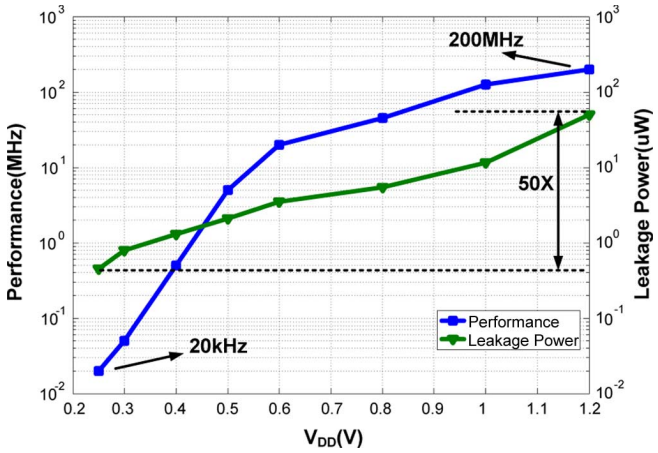


Fig. 19. Performance and leakage power versus $V_{\rm DD}$ plot for the U-DVS SRAM. Leakage power scales down by > 50 X over the voltage range.

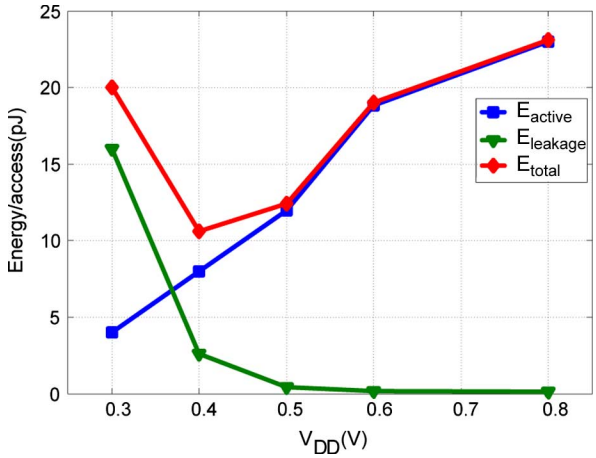


Fig. 20. Active and leakage components of the energy/access for the test chip. Minimum energy point occurs at $\sim$ 400 mV.

because of the conflicting nature of functionality problems at low voltages and performance concerns at high voltages. In this work, we address this problem by employing reconfigurable peripheral assist circuits and 8T bit-cells. 64 kbit SRAM module fabricated in 65 nm CMOS process achieves four orders of magnitude performance scaling from 200 MHz down to 20 kHz at 1.2 V and 0.25 V respectively. Minimum energy point occurs at 0.4 V with 11 pJ/access where 128-bit words are read from or written to the memory.

ACKNOWLEDGMENT

The authors would like to thank Joyce Kwong for valuable discussions.

REFERENCES

[1] A. P. Chandrakasan, D. Daly, J. Kwong, and Y. K. Ramadass, "Next generation micro-power systems," in *Symp. VLSI Circuits Dig. Tech. Papers*, Jun. 2008, pp. 2–5.

[2] P. Macken, M. Degrauwe, M. V. Paemel, and H. Oguey, "A voltage reduction technique for digital systems," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 1990, pp. 238–239.

[3] A. Wang, A. Chandrakasan, and S. Kosonocky, "Optimal supply and threshold scaling for subthreshold CMOS circuits," in *Proc. IEEE Computer Society Annual Symp. VLSI*, Apr. 2002, pp. 7–11.

[4] S. Borkar, "Obeying Moore's law beyond 0.18 micron, microprocessor design," in *Proc. IEEE Int. ASIC/SOC Conf.*, Sep. 2000, pp. 26–31.

[5] B. Calhoun and A. Chandrakasan, "A 256-kbit subthreshold SRAM in 65 nm CMOS," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2006, pp. 628–629.

[6] B. Zhai, D. Blaauw, D. Sylvester, and S. Hanson, "A sub-200 mV 6T SRAM in 13 μm CMOS," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2007, pp. 332–333.

[7] T.-H. Kim, J. Liu, J. Keane, and C. H. Kim, "A high-density subthreshold SRAM with data-independent bitline leakage and virtual ground replica scheme," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2007, pp. 330–331.

[8] N. Verma and A. Chandrakasan, "A 65 nm 8T sub-Vt SRAM employing sense-amplifier redundancy," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2006, pp. 328–329.

[9] I. J. Chang, J. Kim, S. P. Park, and K. Roy, "A 32 kb 10T subthreshold SRAM array with bit-interleaving and differential read scheme in 90 nm CMOS," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2008, pp. 388–389.

[10] L. Chang, Y. Nakamura, R. K. Montoye, J. Sawada, A. K. Martin, K. Kinoshita, F. Gebara, K. Agarwal, D. Acharyya, W. Haensch, K. Hosokawa, and D. Jamsek, "A 5.3 GHz 8T-SRAM with operation down to 0.41 V in 65 nm CMOS," in *Symp. VLSI Circuits Dig. Tech. Papers*, Jun. 2007, pp. 252–253.

[11] M. E. Sinangil, N. Verma, and A. P. Chandrakasan, "A reconfigurable 65 nm SRAM achieving voltage scalability from 0.25–1.2 V and performance scalability from 20 kHz-200 MHz," in *Proc. European Solid-State Circuits Conf. (ESSCIRC)*, Sep. 2008, pp. 282–285.

[12] E. Seevinck, F. List, and J. Lohstroh, "Static noise margin analysis of MOS SRAM cells," *IEEE J. Solid-State Circuits*, vol. SC-22, pp. 748–754, Oct. 1987.

[13] B. Yu, E. Nowak, K. Noda, and C. Hu, "Reverse short-channel effects and channel-engineering in deep-submicron MOSFETs: Modeling and optimization," in *Symp. VLSI Technology Dig. Tech. Papers*, Jun. 1996, pp. 162–163.

[14] M. J. M. Pelgrom, A. C. J. Duinmaijer, and A. P. G. Welbers, "Matching properties of MOS transistors," *IEEE J. Solid-State Circuits*, vol. 24, pp. 1433–1439, Oct. 1989.

[15] A. Kawasumi *et al.*, "A single-power-supply 0.7 V 1 GHz 45 nm SRAM with an asymmetrical unit-β-ratio memory cell," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2008, pp. 382–383.

[16] K. Zhang, K. Hose, V. De, and B. Senyk, "The scaling of data sensing schemes for high speed cache design in sub-0.18 μm," in *Symp. VLSI Circuits Dig. Tech. Papers*, 2000, pp. 226–227.

[17] T. Kobayashi, K. Nogami, T. Shirotori, and Y. Fujimoto, "A current-controlled latch sense amplifier and a static power-saving input buffer for low-power architecture," *IEEE J. Solid-State Circuits*, vol. 28, no. 4, pp. 523–527, Apr. 1993.

[18] H. H. Chen and S. E. Schuster, "On-chip decoupling capacitor optimization for high-performance VLSI design," in *Int. Symp. VLSI Technology, Systems, and Applications, Proc. Tech. Papers*, Jun. 1995, pp. 99–103.

[19] M. D. Pant, P. Pant, and D. S. Wills, "On-chip decoupling capacitor optimization using architectural level prediction," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 10, no. 6, pp. 319–326, Jun. 2002.

Mahmut E. Sinangil (S'06) received the B.Sc. degree in electrical and electronics engineering from Bogazici University, Istanbul, Turkey, in 2006, and the S.M. degree in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, MA, in 2008. He is currently pursuing the Ph.D. degree at MIT, where his research interests include low-power digital circuit design in the areas of SRAMs and video coding.

Mr. Sinangil was the recipient of the Ernst A. Guillemin Thesis Award at MIT for his Master's thesis in 2008, and co-recipient of 2008 A-SSCC Outstanding Design Award and 2006 Bogazici University Faculty of Engineering Special Student Award.

Naveen Verma (S'04) received the B.A.Sc. degree in electrical and computer engineering from the University of British Columbia, Vancouver, Canada, in 2003, and the M.S. and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, in 2005 and 2009, respectively.

Since July 2009, he has been an Assistant Professor of electrical engineering at Princeton University, Princeton, NJ. His research focuses on ultra-low-power integrated circuits including low-voltage digital logic and SRAMs, low-noise analog instrumentation and data-conversion, and energy-efficient processing algorithms especially for biomedical applications.

Dr. Verma was a co-recipient of the 2008 ISSCC Jack Kilby Award for Outstanding Student Paper, and 2006 DAC/ISSCC Student Design Contest Award. During his doctoral research, he was an Intel Foundation Ph.D. Fellow and an NSERC Fellow. He has coauthored book chapters in *Embedded Memory for Nano-Scale VLSI* (Springer, 2009) and *Adaptive Techniques for Processor Optimization* (Springer, 2008).

Anantha P. Chandrakasan (M'95–SM'01–F'04) received the B.S., M.S., and Ph.D. degrees in electrical engineering and computer sciences from the University of California, Berkeley, in 1989, 1990, and 1994, respectively.

Since September 1994, he has been with the Massachusetts Institute of Technology, Cambridge, where he is currently the Joseph F. and Nancy P. Keithley Professor of Electrical Engineering. He is the Director of the MIT Microsystems Technology Laboratories. His research interests include low-power digital integrated circuit design, wireless microsensors, ultra-wideband radios, and emerging technologies. He is a coauthor of *Low Power Digital CMOS Design* (Kluwer Academic Publishers, 1995), *Digital Integrated Circuits* (Pearson Prentice-Hall, 2003, 2nd edition), and *Sub-threshold Design for Ultra-Low Power Systems* (Springer 2006). He is also a co-editor of *Low Power CMOS Design* (IEEE Press, 1998), *Design of High-Performance Microprocessor Circuits* (IEEE Press, 2000), and *Leakage in Nanometer CMOS Technologies* (Springer, 2005).

Prof. Chandrakasan was a co-recipient of several awards including the 1993 IEEE Communications Society's Best Tutorial Paper Award, the IEEE Electron Devices Society's 1997 Paul Rappaport Award for the Best Paper in an EDS publication during 1997, the 1999 DAC Design Contest Award, the 2004 DAC/ISSCC Student Design Contest Award, the 2007 ISSCC Beatrice Winner Award for Editorial Excellence and the 2007 ISSCC Jack Kilby Award for Outstanding Student Paper. He has served as a technical program co-chair for the 1997 International Symposium on Low Power Electronics and Design (ISLPED), VLSI Design'98, and the 1998 IEEE Workshop on Signal Processing Systems. He was the Signal Processing Sub-committee Chair for ISSCC 1999-2001, the Program Vice-Chair for ISSCC 2002, the Program Chair for ISSCC 2003, and the Technology Directions Sub-committee Chair for ISSCC 2004-2009. He was an Associate Editor for the IEEE JOURNAL OF SOLID-STATE CIRCUITS from 1998 to 2001. He served on SSCS AdCom from 2000 to 2007 and he was the meetings committee chair from 2004 to 2007. He is the Conference Chair for ISSCC 2010.