Ultra-Dynamic Voltage Scalable (U-DVS) SRAM Design Considerations

by

Mahmut E. Sinangil

B.A.Sc. in Electrical & Electronics Engineering
Bogazici University, Istanbul 2006

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2008

Author .....
Department of Electrical Engineering and Computer Science
May 20, 2008

Certified by.....
Anantha P. Chandrakasan
Joseph F. and Nancy P. Keithley Professor of Electrical Engineering
Thesis Supervisor

Accepted by .....
Terry P. Orlando
Chairman, Department Committee on Graduate Students

Ultra-Dynamic Voltage Scalable (U-DVS) SRAM Design Considerations

by

Mahmut E. Sinangil

Submitted to the Department of Electrical Engineering and Computer Science
on May 20, 2008, in partial fulfillment of the
requirements for the degree of
Master of Science in Electrical Engineering and Computer Science

Abstract

With the continuous scaling down of transistor feature sizes, the semiconductor industry faces new challenges. One of these challenges is the incessant increase of power consumption in integrated circuits. This problem has motivated the industry and academia to pay significant attention to low-power circuit design for the past two decades. Operating digital circuits at lower voltage levels was shown to increase energy efficiency and lower power consumption. Being an integral part of the digital systems, Static Random Access Memories (SRAMs), dominate the power consumption and area of modern integrated circuits. Consequently, designing low-power high density SRAMs operational at low voltage levels is an important research problem.

This thesis focuses on and makes several contributions to low-power SRAM design. The trade-offs and potential overheads associated with designing SRAMs for a very large voltage range are analyzed. An 8T SRAM cell is designed and optimized for both sub-threshold and above-threshold operation. Hardware reconfigurability is proposed as a solution to power and area overheads due to peripheral assist circuitry which are necessary for low voltage operation. A 64kbit SRAM has been designed in 65nm CMOS process and the fabricated chip has been tested, demonstrating operation at power supply levels from 0.25V to 1.2V. This is the largest operating voltage range reported in 65nm semiconductor technology node. Additionally, another low voltage SRAM has been designed for the on-chip caches of a low-power H.264 video decoder. Power and performance models of the memories have been developed along with a configurable interface circuit. This custom memory implemented with the low-power architecture of the decoder provides nearly 10X power savings.

Thesis Supervisor: Anantha P. Chandrakasan
Title: Joseph F. and Nancy P. Keithley Professor of Electrical Engineering

Acknowledgments

Writing this thesis made me think about my first days at MIT. From those days to now, I have learnt an incredible amount, some about research and the rest about myself and life. Now, when I look back, I can confidently say that I made the right decision to come to the U.S.

I am so grateful to my advisor Professor Anantha Chandrakasan for including me in his group and for being a great advisor. I have learnt a tremendous amount from him not only about circuits but also about other things. Thank you Anantha, for helping me, guiding me and motivating me all the time. I am looking forward for my Ph.D studies and exciting projects under your supervision.

I am also grateful to my family, who were always there for me. Mom and dad, thank you for being such great parents. Although we are thousands of miles away from each other, hearing your voice over the phone is enough to keep my heart warm. I would also like to thank my best friend, my soulmate and my fiancee, Yildiz, for her continuous support.

It was great to know everyone in the group and I really find myself lucky to share the same lab with such special people. I am fortunate to be mentored by Naveen, a person with incredible technical expertise and impressive work discipline. Thank you Naveen, for all your help and support. Joyce, the one who knows everything. I can't remember how many times I came to your cube to ask you stupid questions and you helped me everytime politely. Masood: Thank you very much for being such a great friend. I am looking forward to work with you in the same lab again. Patrick: It was great to be classmates with you. Daniel and Vivienne: It was fun to work with you on the same project and chat during our spare time. Denis, Yogesh, Payam, Nathan, Manish, Courtney, Hye Won, Marcus: It was great to know all of you and I am looking forward to continue working with you. I would like to thank Margaret for helping me out with various things and I would also like to thank Texas Instruments and people on TI side (especially Alice Wang) for helping me with chip fabrication.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Gordon Moore made his famous observation about the annual doubling of the number of transistors on a die in 1965 [5]. Since then, the semiconductor industry accomplished to stay in this exponential trend of scaling. However, scaling of transistor feature size caused new problems, one of which being the ever-increasing power consumption of integrated circuits. The work in [6] shows that the CPU power has been increasing by 2.7X every two years. This continuous trend can be attributed to faster clock frequencies, more functionality embedded into chips and larger die areas. Furthermore, sub-threshold and gate leakage have also shown a continuously increasing trend because of lower threshold voltages (V_t) and thinner gate oxides with nearly every new process technology. Consequently, the total power of microprocessor chips is approaching the 'power-wall' which is set by the maximum power delivery and cost-effective cooling solutions of today's technology [6].

The improvement in battery capacity is not enough to support the ever-increasing power demand of the integrated circuits. This simply translates to shorter life-time for portable electronics, biomedical implants and any energy-starved applications. Energy harvesting is also a new and important technique which could enable self-powered circuits. However, the amount of energy harvested from the environment is highly limited requiring circuits to be very energy efficient. For the reasons stated

above, low-power circuit design has been a very active and important research area for over a decade.

Power consumption of digital circuits can be divided into two components:

- Active Power (P_{ACTIVE}) is the component related to active current drawn by the devices during circuit operation. It is given by

$$P_{ACTIVE} = C_{EFF} V_{DD}^2 f,$$

  where C_{EFF} is the effective switching capacitance, V_{DD} is the supply voltage and f is the frequency of operation.

- Leakage Power ($P_{LEAKAGE}$) is the component related to the off-current of the devices. It is given by

$$P_{LEAKAGE} = I_{LEAKAGE} V_{DD},$$

  where $I_{LEAKAGE}$ is the total off-current of the circuit and V_{DD} is the supply voltage level.

Lowering V_{DD} results in quadratic savings for P_{ACTIVE} and linear savings for $P_{LEAKAGE}$ from the above equations. In reality, however, $I_{LEAKAGE}$ also reduces as V_{DD} scales down and hence, $P_{LEAKAGE}$ exhibits a higher order dependance on V_{DD} . This is due to a second order effect on transistor operation known as drain-induced-barrier-lowering (DIBL). $P_{LEAKAGE}$ can be a significant portion of power consumption if the activity factor of the circuit is low.

SRAMs are consuming a large fraction of the total chip area because of their appealing features such as low activity factor and very high transistor density. Figure 1-1 shows the memory area as a percentage of total chip area over the past years [1]. The continuous increase in on-chip cache area is the main motivation for designing SRAMs with low power consumption.

As discussed above, lowering the supply voltage of digital circuits results in significant power savings. The minimum functional supply voltage of an integrated circuit is limited by on-chip SRAMs for the following two reasons:

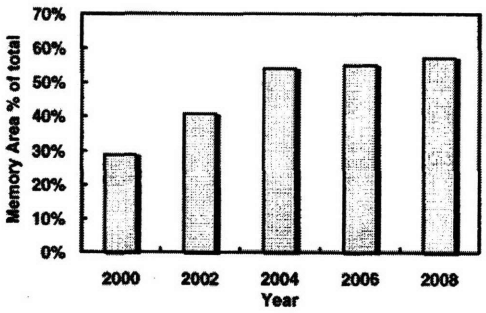

Figure 1-1: Predicted memory area as a percentage of total microprocessor area [1]

- With decreasing V_{DD} , SRAM delay increases at a higher rate than CMOS logic delay
- At low voltage levels, SRAM cell stability quickly degrades causing functional failures [7].

Thus, designing SRAMs operational down to low voltages is an important prerequisite to build a low-power system.

Conventional SRAMs fail to operate at low voltage levels due to cell stability problems. At the 65nm technology node, a conventional 6T design can operate down to 750mV. If reducing the supply voltage and further power savings are necessary for an energy-constrained application, new circuit topologies and techniques must be explored.

The array efficiency of a memory is given by the ratio of the cell array area to the overall memory area. As mentioned above, SRAM area is the dominant portion of modern integrated circuits so array efficiency of the memories has a direct effect on the overall area. The techniques developed to enable low voltage SRAM operation should also take array efficiency into account to be practical. For example, if SRAM accounts for 50% of the overall area in a design and a new low power technique increases SRAM area by 5%, this causes a 2.5% area increase for the overall design. In large scale manufacturing, this impacts the overall yield and cost of the system significantly.

1.2 Ultra-Dynamic Voltage Scalable Systems

Ultra-Dynamic Voltage Scalability (U-DVS) is an approach to reduce power consumption when the performance requirements of the system vary. In U-DVS systems, the supply voltage, and consequently the frequency, is adjusted to the system performance constraints to reduce power consumption. This approach was first proposed in [8] which uses a critical path replica in a feedback loop to tune the supply voltage to the lowest value while still satisfying a certain performance constraint.

Figure 1-2 shows a U-DVS scenario for a wireless sensor node that acquires data from its environment, processes this data and then stores it in memory. The sensor node transmits this data to a base station once the memory is full or once a request is received. Sensing and data processing tasks generally require very low performance levels and thus the sensor node can be operated at a low supply voltage during this phase. However, transmission of the acquired data must be done at a much higher speed and hence at a much higher voltage to maintain transmission efficiency. Adjusting the supply voltage for a given performance constraint provides power-efficient operation. U-DVS systems, however, face the challenges of optimizing circuits for a large voltage range.

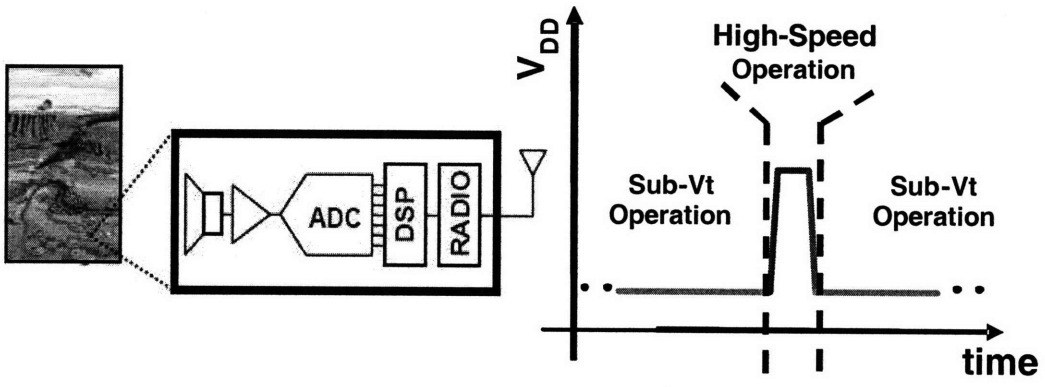

Figure 1-2: The plot shows the supply voltage for a U-DVS sensor node during its operation. Short cycles requiring high speed operation are enabled by increasing the supply voltage whereas low performance operation is done at a much lower supply level.

U-DVS operation is very important for memory circuits. Most SRAM designs

lower their array voltage during sleep mode to reduce leakage power. However, this does not necessarily provide the maximum possible power savings. Conventional SRAMs need to elevate the array voltage back to its nominal value when returning to the active mode. Entering and exiting sleep mode requires switching of large capacitances which are costly in terms of energy. Moreover, going into and coming out of the sleep mode requires some buffer time for settling. Due to above reasons, a memory can go into the sleep mode only in certain cases:

- Leakage energy savings should be larger than the energy consumption associated with entering and exiting the sleep mode.
- The end of the sleep mode should be known in advance so the memory has enough time to go back into the active mode.

These conditions might not be satisfied during all cases and consequently, a memory can enter the sleep mode only during long cycles with no accesses.

The U-DVS SRAM proposed in this work can adjust its voltage level depending on the performance requirement of the system. The low end of the voltage range for the U-DVS SRAM is very close to the data retention voltage level where the maximum power savings can be attained. In contrast to a sleep mode, U-DVS memory also allows accesses while working at very low voltages. Hence, this approach is superior to the conventional designs which employ power gating techniques during sleep modes.

1.3 Previous Work on Low-Power Digital Design

Operating digital circuits in the weak inversion region was first analyzed in [9]. The work in [10] showed that the minimum energy point for digital circuits occurs in sub-threshold region. This was revisited in [11], which plots constant energy contours for a ring oscillator as V_{DD} and V_t are varied. The minimum energy point was shown to lie in sub-V_t ($V_{DD} < V_t$) and varies with the activity factor of the circuit. Since energy and power are directly related to each other, low-V_{DD} and sub-V_t operation have been an important research direction for low-power design.

A sub-V_t FFT processor operational down to 180mV was presented in [12]. This design uses a mux-based memory which is the first example of sub-V_t memory reported. In 2006, [13] demonstrated a sub-V_t SRAM operational down to 400mV in 65nm CMOS. This design uses a 10T SRAM cell to have high density in the array and uses peripheral assist circuits to enable sub-Vt functionality. A year later, [14] proposed an 8T SRAM operational down to 350mV in 65nm. This design uses a smaller cell and also tries to address low voltage functionality issues with peripheral assist circuits. This work also uses sense-amplifier redundancy to improve the yield of the design which can be limited by the sense-amplifier failures at low voltage levels. In the same year, a 10T SRAM with a virtual ground replica sensing scheme operational down to 200mV [3] and a 6T SRAM in 130nm CMOS reporting functionality down to 200mV [15] were also reported.

The design in [16] works over a large range (0.41-1.2V) but only in the above-V_t ($V_{DD} > V_t$) regime. This work uses an 8T SRAM cell and multi-bit ECC scheme to achieve this range. By restricting the voltage range to the above-V_t regime, this design can use static topologies and well-known SRAM trade-offs. The design in [17] demonstrates a similar design working down to 0.42V in a 90nm CMOS. Similarly, this work only targets above-V_t operation and hence discusses trade-offs associated with this regime.

Above designs achieve low energy per access values due to operation at low voltage levels. Some of the previous work demonstrated sub-V_t SRAMs by using novel circuit techniques. However these designs only targeted sub-V_t or above-V_t operation, not both of them together. For U-DVS applications, memories should be designed to be functional down to sub-V_t levels and they should also be optimized for the above-V_t operation. This work focuses on this research problem and proposes new circuit techniques to enable U-DVS SRAMs. These proposed techniques are implemented on test chips which are tested to be fully functional.

1.4 Device Operation in Sub-threshold and Process Variation

1.4.1 Device Operation in Sub-V_t Regime

Sub-threshold current is given by Equation 1.1 [18]

$$I_{Dsub-threshold} = I_o e^{\left(\frac{V_{GS}-V_T+\eta V_{DS}}{nV_{th}}\right)} \left(1 - e^{\left(\frac{-V_{DS}}{V_{th}}\right)}\right), \tag{1.1}$$

where I_o is the drain current when $V_{GS} = V_T$ and is given in Equation 1.2 [18][19].

$$I_o = \mu_o C_{ox} \frac{W}{L} (n-1) V_{th}^2 \tag{1.2}$$

In Equation 1.1, I_D varies exponentially with V_{GS}, gate-to-source voltage and V_T, the device threshold voltage. V_{th} denotes the thermal voltage and $n = (1 + C_d/C_{ox})$ is the sub-threshold slope factor. η represents the drain-induced barrier lowering (DIBL) coefficient. The roll-off current at small V_{DS} is modeled by the term in the rightmost parentheses in Equation 1.1. Figure 1-3 shows the normalized drain current versus V_{GS} and demonstrates the exponential dependance of current to gate drive in sub-threshold explicitly.

1.4.2 Process Variation

As with all manufacturing processes, semiconductor fabrication is subject to many sources of variation which can be divided into two major groups: global variation and local variation.

Global variation, as its name implies, affects all transistors on a die. This causes the device characteristics to vary from one die to another.

Local variation affects each transistor differently. The main components of local variation are random dopant fluctuation (RDF) and line edge roughness (LER). RDF, an important parameter that causes V_t variation, is shown to have a Gaussian distribution with its standard deviation being proportional to the square root of the

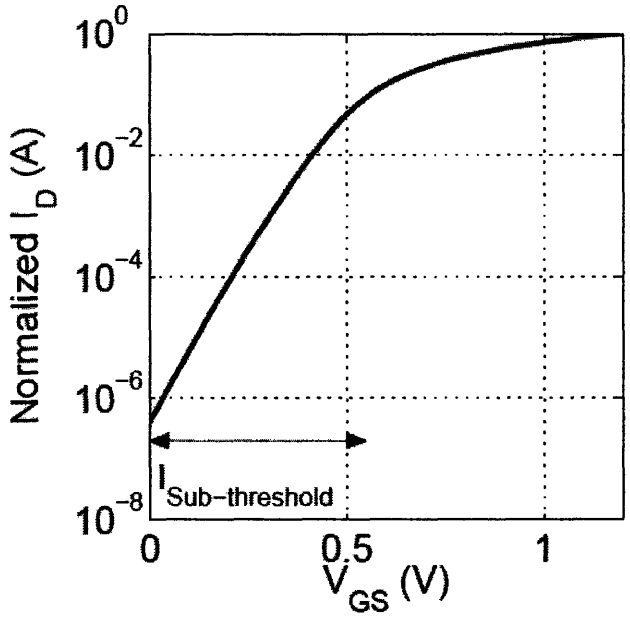

Figure 1-3: Normalized drain current versus V_{GS}. V_{DS} is at 1.2V for all points.

device channel area [20]. Thus, V_t of devices are governed by a Gaussian distribution. In sub-V_t regime, V_t has an exponential effect on I_D and hence, two identical devices can have orders of magnitude difference between their drive strengths.

1.5 Thesis Contributions

This thesis focuses on U-DVS SRAM design, an important research problem for energy-starved applications. The proposed innovations are implemented in two separate test chips and test results show that both designs are fully functional on silicon.

The first part of this thesis presents an SRAM that is designed for both sub-V_t and above-V_t operation. The design is operational from 250mV which is in deep sub-V_t region to 1.2V which is the nominal-V_{DD} for the process. An 8T bitcell is designed to construct a high density array. Hardware reconfigurability is proposed for assist circuitry to prevent power and area overheads. Three different reconfigurable write-assist schemes are implemented to enable functionality over the large voltage range. Multiplexed-sense-amplifiers minimize the sensing delay in the sensing net-

work. Lastly, the bitcell and peripheral circuits are designed for optimal operation over the voltage range.

The second part of this thesis focuses on the cache design considerations for a low-power H.264 video decoder chip. A full custom U-DVS SRAM is designed to specifically minimize the SRAM power consumption. Performance and power models are developed for the memories. A configurable interface circuit is proposed to create critical timing signals for the memories. The preliminary testing results show that the chip is fully functional down to 0.7V with significant power savings compared to previous work.

Chapter 2

U-DVS SRAM Design in 65nm CMOS

2.1 6T SRAM Cell Operation

Figure 2-1 shows the conventional 6T SRAM cell. This cell became nearly an industry standard over many years. The structure consists of two cross-coupled inverters (MN1, MN2, MP1 and MP2) and two NMOS access transistors (MN3 and MN4). Because of the positive feedback between the cross-coupled inverters, storage nodes N1 and N2 can hold a data indefinitely in a stable state if the access transistors are turned-off. Access transistors are used for write and read operations and turned-on only during accesses. NMOS device in the inverters are generally referred as "driver" and PMOS devices are referred as "load" in the literature

2.1.1 Notion of Static-Noise Margin (SNM) for SRAM cells

Static-Noise Margin (SNM) is a very widely used metric in SRAM design to characterize the stability of a cell. An analytical model and a simulation method to calculate the SNM of a memory cell are given in [21]. SNM quantifies the amount of voltage noise required at the storage nodes that causes the bitcell to lose its data. This noise can be present due to device mismatches or due to dynamic disturbances

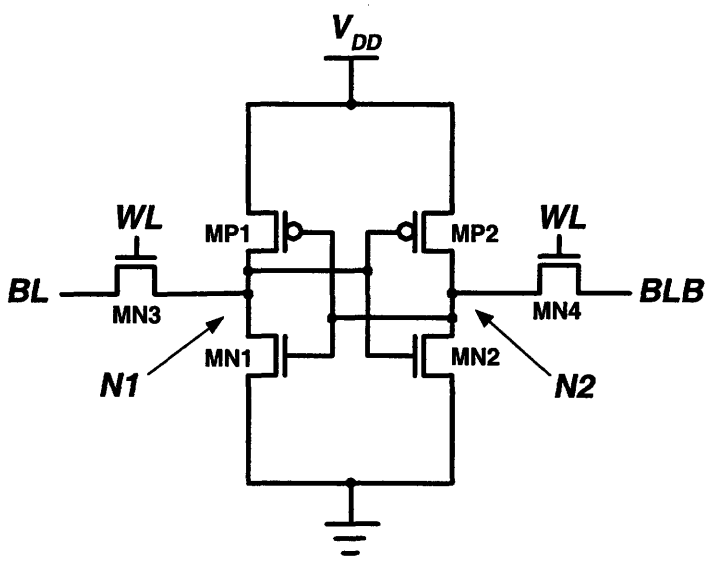

Figure 2-1: Conventional 6T SRAM cell

to the memory cell during accesses.

A graphical representation of SNM is shown in Figure 2-2. If SNM is calculated during "Hold mode" (i.e. when WL voltage is low), it is referred as the Hold SNM (Figure 2-2). Two curves on this plot are the Voltage Transfer Characteristic (VTC) of the first inverter and the inverse VTC of the second inverter. These two curves intersect at three points (two stable points and a metastable point). Two stable points represent the two values that can be stored in the memory cell (a logic '1' or '0'). Because of local and global variation, the relative strength of transistors inside the inverters can alter significantly. This changes projects to VTC shiftings and cause SNM degradation. In the graphical representation, SNM is defined as the length of the side of the largest square that can fit inside the lobes of the curve. In addition to the variation, accessing a memory cell also disturbs cell stability which can be analyzed by SNM calculations. Read and Write SNMs will be discussed in the subsequent sections.

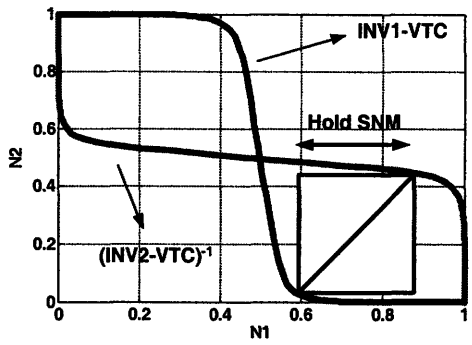

Figure 2-2: Graphical representation of SNM in 'Hold' mode. The side length of the largest square that can fit inside the lobes of butterfly curve gives SNM value.

2.1.2 Write Operation

During a write operation, BL and BLB terminals of the bitcell are driven to the data that is going to be written. Then, WL is pulled high, exposing the storage nodes to BL and BLB. Since NMOS access transistors can drive a '0' more easily, write operation begins by writing the '0' first. After writing '0', internal feedback between the inverters forces other storage node to '1'. An example write SNM figure along with the transient waveforms of the internal signals in a memory cell during a write access are shown in Figure 2-3. Write SNM in Figure 2-3-a can be counter-intuitive at first. During a write operation, memory cell is forced to store a data value which results in only one stable point in the butterfly-curve. Keeping the same graphical SNM definition, for write SNM, a negative value shows a successful write operation. The waveforms in Figure 2-3-b shows that '0' is written to N1 first, triggering the internal feedback and pulling N2 high.

A new write margin definition and a new simulation method were proposed in [22] in 2007. This definition sweeps the WL voltage until the both storage nodes have the same voltage and defines write margin as the difference of WL voltage at this trip voltage and the maximum WL voltage.

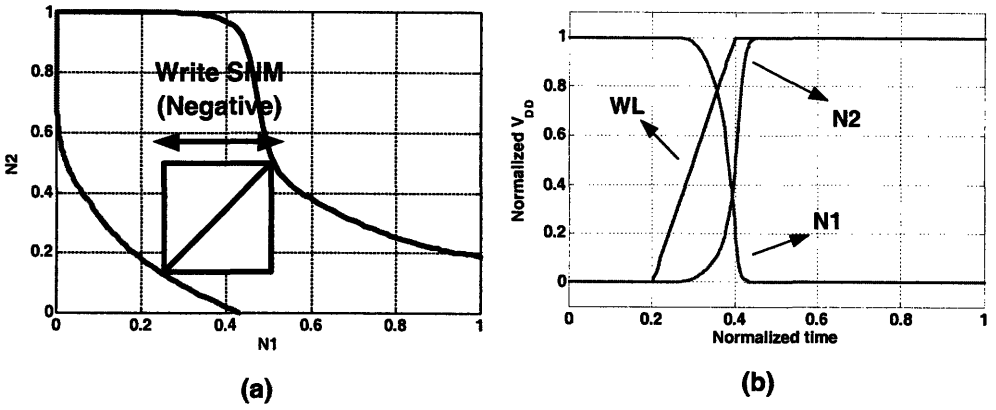

Figure 2-3: Graphical representation of SNM in 'Write' mode (a) and typical waveforms during a write operation (b). '0' is written through the NMOS access transistors first.

2.1.3 Read Operation

During a read operation, first, BL and BLB terminals of the bitcell are pre-charged to an initial voltage (generally to V_{DD}). After this pre-charge phase, both BL and BLB are kept floating and the WL signal is pulled high. The memory cell, depending on the data stored in its storage nodes, drive one of the BLs low whereas other BL stays high. Then, a sense-amplifier senses this change between the BLs and outputs the data.

Figure 2-4 plots an example Read SNM figure and the transient waveforms of the internal signals in a memory cell during a read access. Read SNM is a degraded version of Hold SNM. This is due to the extra noise injected into the internal nodes during a read access. As explained above, a read access is performed by pre-charging BL and BLB terminals and asserting WL signal. At the beginning of a read cycle, access transistor which is connected to the storage node holding a '0' tries to pull this node up. This causes a slight increase at this storage node voltage resulting in degradation of SNM.

6T cell, because of its compact layout, has been the dominant choice for SRAMs. However, scaling of process technologies caused a continuous degradation of SNM due to increased effect of local variation. Additionally, V_{DD} scaling causes further

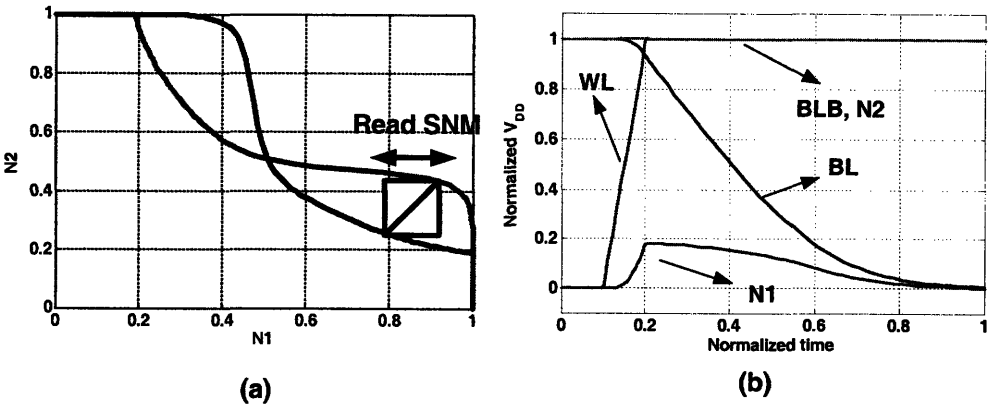

Figure 2-4: Graphical representation of SNM in 'Read' mode (a) and typical waveforms during a read operation (b). Read access causes a disturbance on one of the storage nodes.

degradation of cell stability. These two effects add up and cause functional failures for the 6T cell at low voltages. Analysis of these failures will be discussed next.

2.2 6T SRAM Cell at Low Supply Voltages

Power supply scaling makes devices more prone to the effect of variation. As V_{DD} approaches sub-V_t levels, V_t begins to have an exponential effect on I_D causing orders of magnitude fluctuations in drive strengths.

6T cell works properly under three main conditions:

- The cross-coupled inverters have a positive Hold SNM, i.e. the cross-coupled inverters can store both a '0' and a '1' statically.

- The access transistor can overpower the load transistor and break the internal feedback between the cross-coupled structure during a write operation.

- The disturbance on the storage nodes during a read access is low such that the cell is not accidentally flipped.

2.2.1 Data Retention Problems

Data retention can be maintained relatively easier. For an inverter, only one of the transistors (either NMOS or PMOS) is ON and the other device is OFF provided that the inverter is in a stable state. Even at 200mV of V_{DD} , if the devices are sized to have the same strength, one of them will have nearly 2 orders of magnitude larger current than the other under nominal conditions. This large difference is hard to be altered with variation and hence Hold SNM is relatively easier to satisfy. Upsizing memory cell transistors improves Hold SNM but is not desirable due to density considerations.

Finding the minimum data retention voltage (DRV) is an important problem since this voltage will determine the ultimate minimum of the operating voltage range. Moreover, DRV can be used to determine the idle-mode voltage to minimize the power consumption. A canary-replica feedback circuit is proposed in [23] to dynamically monitor DRV on-chip and adjust the supply voltage accordingly. The work in [24] proposes efficient statistical simulation methods to determine DRV during design stage.

2.2.2 Read Access Problems

At the beginning of a read access, voltage of the storage node holding a '0' is pulled-up through the access transistor. If this disturbance on the storage node is large enough to flip the cross-coupled inverters, a read access may cause the bitcell to lose its data (Figure 2-5-b).

The amount of voltage increase on the storage node depends on the relative strengths of the access and driver transistors. If these transistors are assumed to be resistors, the internal node voltage can be calculated from the resistive divider between two resistors.

In a 6T SRAM, driver transistor is sized to be stronger than the access transistor to minimize the extent of disturbance. However, as V_{DD} scales down, the effect of variation easily overpowers the effect of sizing. Figure 2-5-a shows 4σ Read SNM vs. V_{DD} in 65nm process for a minimum sized cell, i.e. all transistors are sized minimum.

This plot shows that 6T cell can work down to 0.6V for this process.

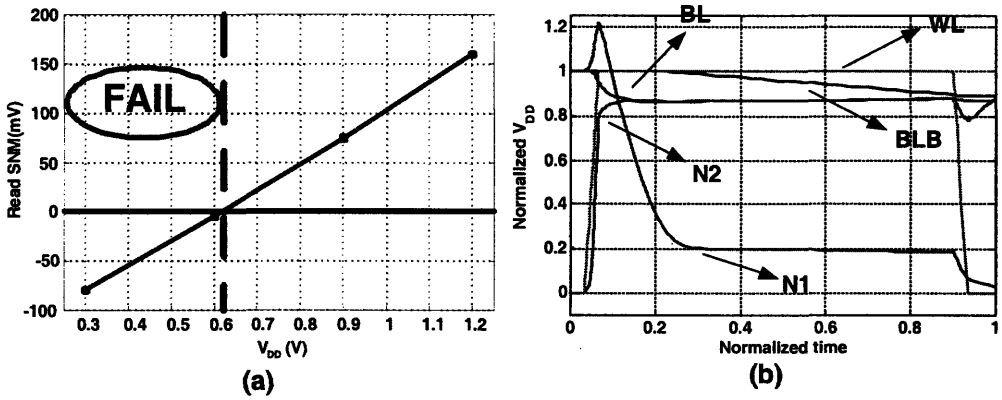

Figure 2-5: 4σ Read SNM vs. V_{DD} at 65nm node for a minimum sized cell (a) and waveforms showing a read failure (b). Disturbance on N2 is large enough to trip the cross-coupled inverters and the bitcell flips during the access.

Read margin, although analyzed statically in literature, is a dynamic phenomenon. When WL is asserted, BL creating the disturbance begins to discharge through the cell, causing the amount of disturbance to get smaller. If the BL capacitance is large, discharging will occur very slowly making the problem nearly a static phenomenon. However, in the case of a small BL capacitance, dynamic read margin should be considered for more accurate results.

2.2.3 Write Access Problems

Figure 2-6-b shows a write operation exerted on the 6T cell. For a successful write operation, first, previously stored '1' has to be written over. This is done by sizing the access transistors stronger than the load transistors. However with the effect of variation and scaling of V_{DD} , write failures for 6T cell begins to occur. Figure 2-6-a shows the 4σ Write SNM vs. V_{DD} in 65nm process for a minimum sized cell. Write stability is lost around 1V.

Read and Write problems associated with the 6T cell at low $V_{DD}s$ motivate for new topologies. The next sections will talk about previous work on low-V_{DD} SRAM design.

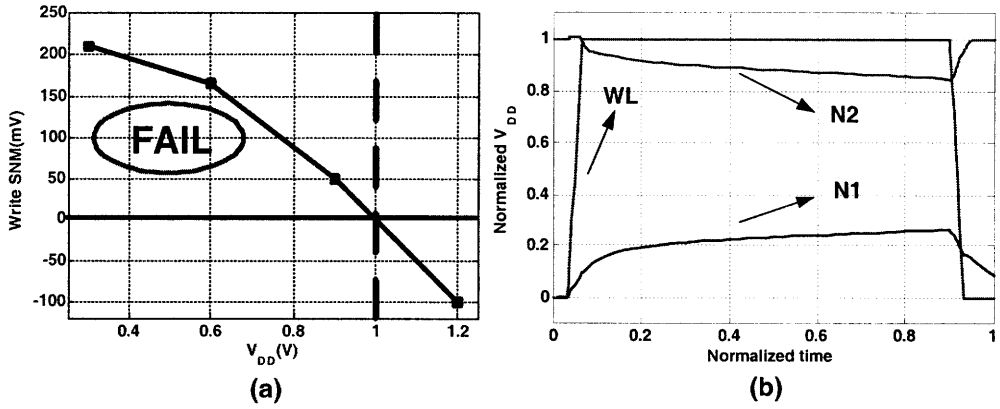

Figure 2-6: 4σ Write SNM vs. V_{DD} at 65nm node for a minimum sized cell (a) and waveforms showing a write failure (b). Access transistor cannot overpower the PMOS transistor preventing N2 voltage to go down.

2.3 Previous Work

The problems explained in the previous section motivate for new topologies to enable SRAM operation at low supply voltages. This section is devoted to summarize the previous work on low-power SRAM design.

2.3.1 Other Bitcell Topologies from Previous Work

The work in [7] proposes a 7T bitcell where the extra NMOS transistor is used to break the feedback between the cross-coupled inverters during a read operation. This makes the cell "read margin free". In order to have a compact area, cell layout is designed to be L-shaped and the extra space between the cells is used for the sense-amplifier. This ingenious layout idea brings the area overhead to an acceptable range of 11%.

An 8T SRAM cell is proposed in [25] and has been used in many low-power SRAM designs (Figure 2-7). 8T cell uses two extra NMOS devices to replicate the pull-down path of the 6T cell. This extra part is called "read-buffer" since it is used for read operations.

A read operation is exerted through pre-charging RDBL and assertion of RDWL. Since the storage nodes are decoupled from RDBL, accesses do not cause a disturbance

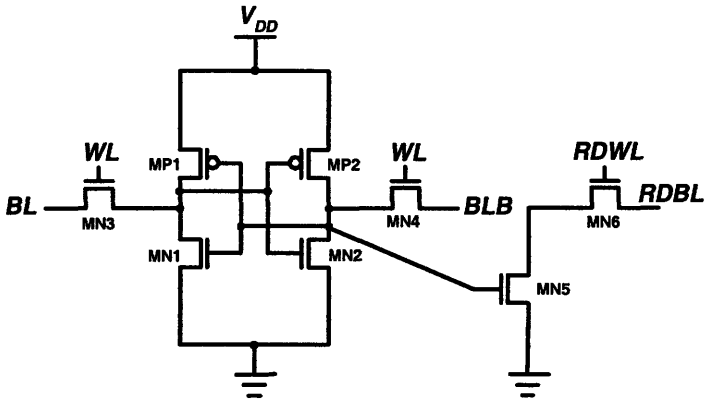

Figure 2-7: 8T SRAM cell. Two extra NMOS transistors form the "read-buffer" and decouple read and write ports of the cell.

on bitcell stability. So, Read SNM for this structure is the same as its Hold SNM, provided that the gate leakage through MN5 is negligible. A write operation is done through the 6T portion of the cell in the conventional way.

8T bitcell design has approximately 1.3X larger area than a 6T cell in 65nm [25] and in 90nm [17]. However making a fair comparison is not easy since the 8T cell can operate at much lower voltages than the 6T cell. The work in [17] predicts that, with the exacerbated effect of variation at 32nm node and beyond, 8T cell will have a smaller area than the 6T cell.

Other previous work proposes 10T SRAM cells operational in sub-V_t region ([26] and [3]). These designs use four extra transistors to address the sub-V_t functionality problems which will be discussed next.

2.3.2 Write Assist Circuits in Previous Work

One of the most widely used methods to assist a write operation is altering the voltage of one or both of the supply rails during an access. Lowering the supply voltage or elevating the ground voltage weaken the positive feedback between the cross-coupled inverters by reducing their drive strength. This results in improvement of write margin.

Changing the supply rail voltage during write operation can be done in a row-wise

manner or in a column-wise manner. The work in [27] uses a column-wise scheme where the accessed columns' supply voltage is floated during write accesses. Since WLs for the un-accessed rows are not asserted, lowering of the column supply by floating does not affect data retention for the un-accessed cells. This work also uses a replica circuit to determine the WL pulse width. Using two different supply voltages ($V_{DD,High}$ and $V_{DD,Low}$) for the SRAM array is proposed in [2] as shown in Figure 2-8. During a write operation, supply nodes for the accessed column is connected to $V_{DD,Low}$ whereas the un-accessed columns are connected to $V_{DD,High}$. This scheme improves the write margin of the accessed cells.

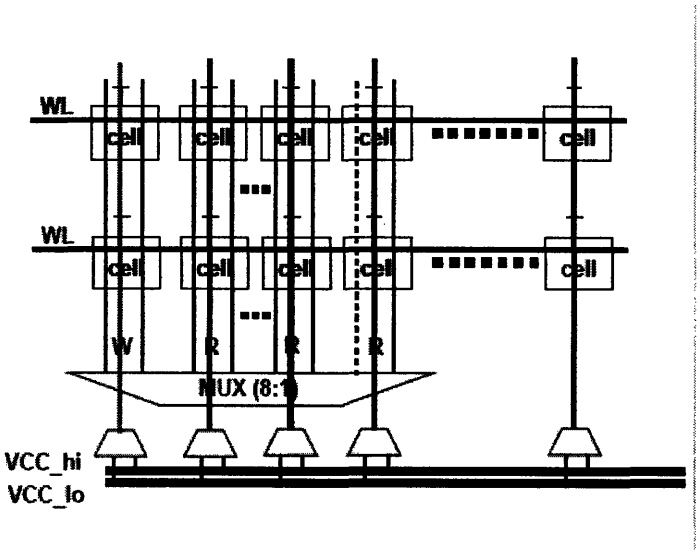

Figure 2-8: Multiple-V_{DD} scheme proposed in [2]. $V_{DD,Low}$ is selected to be 100mV lower than $V_{DD,High}$ to improve write margin.

Earlier examples of row-by-row power supply scheme are [28] [29] and [30].The designs in [26] and [14] also uses row-by-row V_{DD} , with [26] floating the virtual supply node and [14] actively pulling it down during a write operation.

The disadvantage of row-by-row approach is that it is harder to implement column interleaving together integrated into this scheme. Scaling of process technologies, results in smaller storage node capacitances causing increased soft error rates. This makes interleaving and error correction coding (ECC) schemes necessary for SRAMs. A multi-bit ECC scheme is proposed in [16] with limited area overhead which is

compatible with the row-by-row scheme.

Another effective method to ease a write operation is boosting the WL node for the accessed cells. This will increase the drive strengths of the access transistors and help them win the fight over the load transistors. However, this requires a second supply voltage which is very undesirable because of the increased complexity of power distribution network. [17] is an example of this approach. The WL voltage is kept at full V_{DD} (1.2V) in this work regardless of the array voltage.

2.3.3 Read Assist Circuits from Previous Work

Above-V_t designs does not require a read assist circuit since I_{ON}/I_{OFF} ratio is very large at these voltages. However, because of column interleaving, un-accessed cells on the same row are under "read-stress" during a write operation which can cause bitcells to flip and lose their data. This was addressed with architectural changes in previous work. An example is [2] which increases the supply voltage for the un-accessed columns during a write access. The design in [31] uses a sleep transistor at the footer node of an array to decrease leakage in sleep modes. This is shown to improve read margin by slightly increasing storage node voltages and degrading the drive strengths of access transistors.

For sub-V_t SRAMs, bitcell design requires consideration of other functional problems. In sub-threshold region, degraded I_{ON}/I_{OFF} ratio can result in erroneous readings and hence loss of functionality. During a read operation, aggregated leakage current through the bitcells can be comparable to the drive current of the cell. This causes discharging of both BLs at nearly the same rate making a correct sensing extremely difficult or impossible.

In order to address this issue, [26] uses stacking of transistors in the read-buffer. Stack effect decreases the leakage by orders of magnitude so it is very effective in leakage reduction. The work in [3], on the other hand, uses a data-independent leakage path as shown in Figure 2-9.

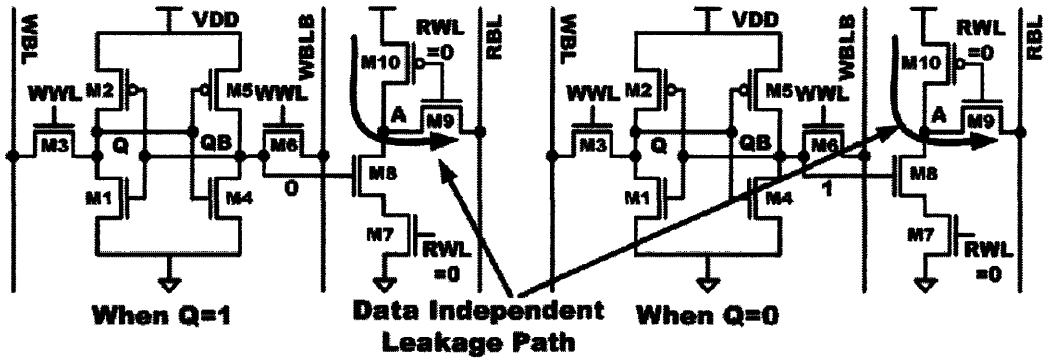

Figure 2-9: SRAM bitcell proposed in [3]. Un-accessed memory cells inject data-independent leakage to the RDBLs.

2.4 U-DVS SRAM Design

2.4.1 Proposed Bitcell Design

The bitcell used in the design is shown in Figure 2-10. This sub-V_t cell topology was first proposed in [14]. For U-DVS SRAM design, this topology is used by optimizing it for the requirement of large operating voltage range.

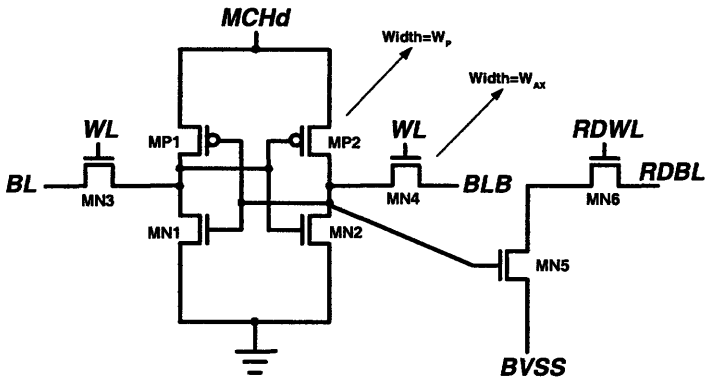

Figure 2-10: 8T SRAM cell used in the U-DVS SRAM.

BVSS is a virtual ground node for the read-buffer. This node is kept at V_{DD} if memory cell is not accessed. This makes the voltage drop across read-buffer devices zero and hence leakage through un-accessed cells becomes negligible. MCHd is the virtual supply node for the cross-coupled inverters and its voltage can be altered during an access to ease a write operation.

Pulling the MCHd node down during a write access is an effective way to improve the write margin. However, when this node is pulled down, a short-circuit current path emerges as shown in Figure 2-11. In the figure, the values shown with an arrow are the previous values stored in the storage nodes and BL and BLB are driven to the new values through BL drivers. At the same time, MCHd is also pulled-down through a driver. At the beginning of the write cycle, '1' is discharged through MN3 causing MP2 to turn-on. This creates a short-circuit path as shown in Figure 2-11 with dashed lines.

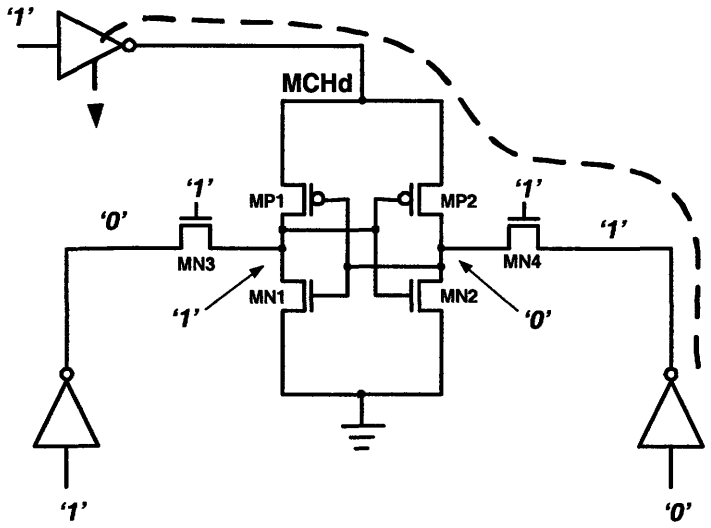

Figure 2-11: A short circuit path emerges during a write operation when MCHd node is pulled down.

In the sub-V_t region, this short-circuit current is smaller than the leakage current of the whole array and hence it is negligible. At higher voltages, however, short circuit current can cause a significant amount of power consumption. The bitcell should be optimized for writability so that this peripheral assist can be activated at a lower V_{DD} resulting in less short-circuit power consumption. The U-DVS SRAM chip does not employ column interleaving so this 8T bitcell does not suffer from degraded read margin at low supply levels. Hence, load devices can be made weaker and access transistors can be designed to be stronger.

The effect of sizing on Hold SNM and Write SNM is shown in Figure 2-12. Write

margin improves significantly with increasing access transistor width and decreasing load transistor width. Hold SNM, however, is minimally affected for these four different sizings. Figure 2-12 suggests that write margin can be improved without degrading Hold SNM for the 8T design. This sizing approach is used for the design.

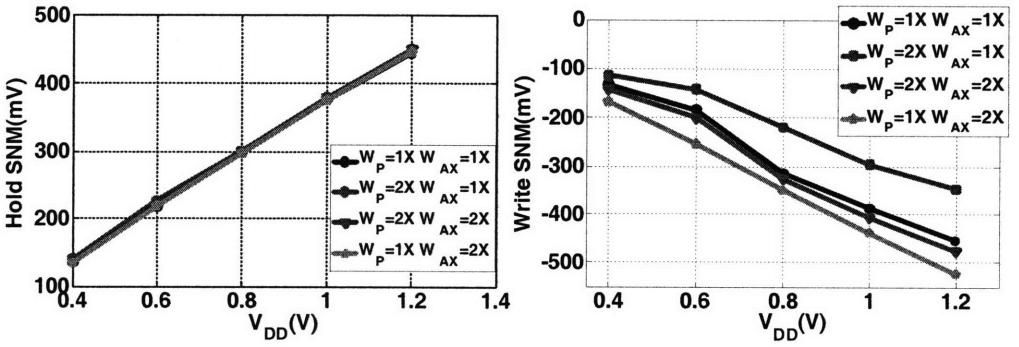

Figure 2-12: Effect of access and load transistor sizing on Hold and Write Margin. Making load PMOS weaker and making access NMOS stronger improves write margin considerably without degrading HSNM.

Sizing of the read-buffer devices should be done by considering the large voltage range. Conventionally, read-buffer devices are sized with minimum length. However, for the U-DVS design, second order effects that begin to emerge at low-voltage levels need to be considered carefully.

Reverse short channel effect (RSCE) is a secondary effect that causes a reduction in V_t with longer channel length [32]. 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 causes an increase in V_t if the channel length is very short.

Figure 2-13 shows the effect of longer channel lengths on the SRAM performance over the voltage range. At high voltages, longer channel length results in degraded performance whereas at low voltages, due to the RSCE, longer channel length provides improved performance. For this design, gate lengths for read-buffer devices are chosen to be approximately twice the minimum size to achieve a good low voltage performance without affecting the high voltage performance.

Although the bitcell is optimized for low voltage functionality and high voltage

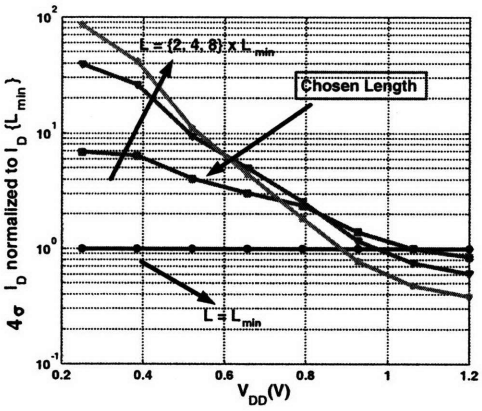

Figure 2-13: 4σ I_D of a read-buffer normalized to I_D of minimum length read-buffer. Longer channel lengths results in improved performance at low voltages but degraded performance at high voltages.

performance, the effect of variation in sub-V_t region is so severe that peripheral assists are also required. The next sections will talk about the peripheral read and write assist circuits for this work.

2.4.2 Proposed Write Assist Design

U-DVS SRAM design requires write assists for low voltage functionality. However, as mentioned in the previous section, peripheral assists generally introduce overheads in terms of power consumption and area. This design considers possible overheads and tries to address them with new circuit techniques.

Previous work is intended to be operational in either sub-V_t region or the above-V_t region. Sub-V_t designs try to enable functionality at very low voltage levels. Since energy efficiency is the primary target for these designs, introducing area overhead to some extent is acceptable. For above-V_t designs, however, the primary target is area efficiency and performance and hence the overheads are kept at very low levels. Lastly, for the U-DVS SRAM covering both sub-V_t and above-V_t regions, assist circuitry is definitely required. But these circuits should be designed for minimal overhead especially at high V_{DD} levels.

For the peripheral assist design of the U-DVS SRAM, "reconfigurable" circuits are

proposed to enable low voltage functionality with minimal overhead at high voltages. Figure 2-14 shows write margin or WSNM distributions of the memory cell used in the design at V_{DD} = 1.2V, 0.75V and 0.25V.

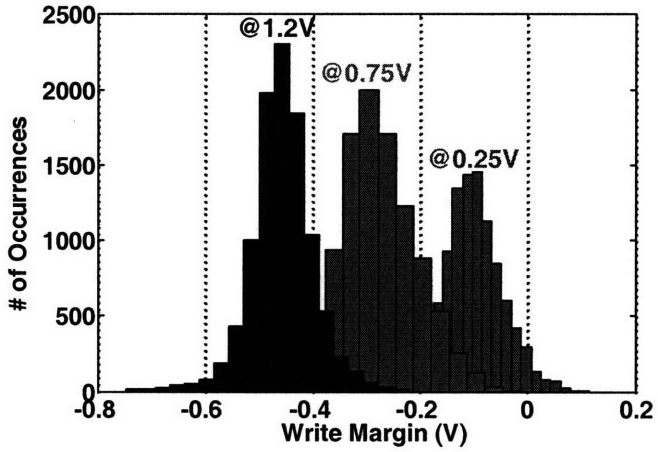

Figure 2-14: Write margin distribution at three different supply voltages (1.2V, 0.75V and 0.25V) suggests that a reconfigurable write assist scheme should be used.

The distribution at 1.2V shows that at this voltage, virtually all memory cells have enough write margin and do not require write assist. As V_{DD} scales down, the mean of the distributions move to the right and the σ of the distribution increases. This is due to the increased effect of variation at low voltage levels. At 0.75V the tail of the distribution becomes positive, indicating write failures. Since only a small portion of the cells are failing at this voltage, a slight improvement in write margin can recover these cells. At 0.25V, a significant portion of the memory cells suffer write failure. To operate correctly at this voltage, these cells require even more improvement in write margin.

Having three different scenarios over the voltage range motivates the design of reconfigurable write assist for U-DVS SRAM. Figure 2-15-a shows the MCHd driver design and Figure 2-15-b shows three different write assist schemes implemented.

At high voltage levels, memory cells have enough write margin to operate correctly so the MCHd node which is the virtual supply node for the memory cells on a row, is kept at V_{DD} .

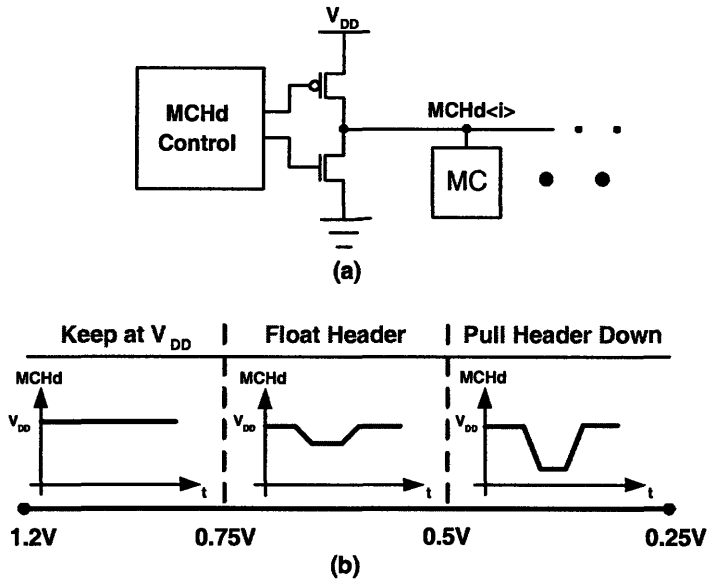

Figure 2-15: MCHd driver (a) and three different write assist schemes for U-DVS SRAM (b).

At lower voltage levels, memory cells at the tail of the distribution show write failures. A slight improvement of write margin which can be achieved by floating the MCHd node, enables these cells to be written successfully. When the MCHd node is floated, the charge on this node is shared among the bitcells on the same row. A load device trying to fight its counterpart access device will pinch from this fixed amount of charge, causing the MCHd voltage to lower and helping access transistors to overpower more easily.

At even lower supplies including the sub-V_t region, the write margin needs to be further improved. At these voltage levels, the MCHd driver pulls MCHd node down, causing it to drop to a very low voltage. Access transistors can easily flip the cells under these conditions, enabling write functionality even in sub-V_t region. MCHd voltage is actively pulled-up in all schemes before WL goes low. This enables positive feedback inside memory cells to charge/discharge internal nodes all the way to the rails before the end of a write operation.

Figure 2-16 shows the short-circuit power consumption due to driving MCHd voltage down at all voltages. Using this scheme instead of the proposed reconfigurable

scheme results in extra power consumption in the mW range at full-V_{DD} level. For systems requiring low power SRAMs, this overhead is not acceptable.

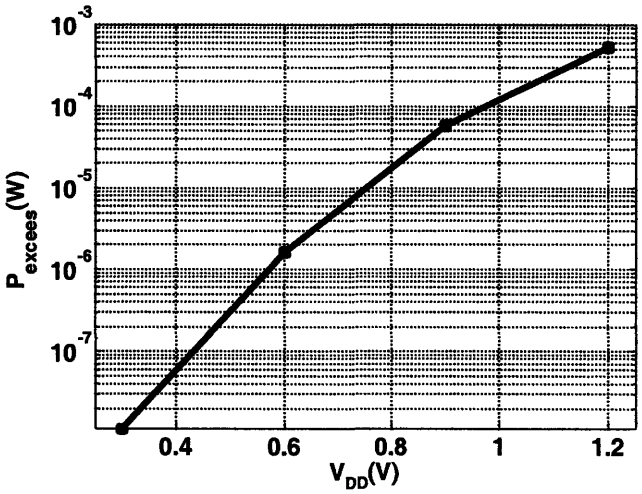

Figure 2-16: Power overhead due to the short-circuit power during pulling MCHd node low over the V_{DD} range. At high voltages, this overhead becomes significant requiring a reconfigurable write assist scheme.

2.4.3 Proposed Read Assist Design

Degraded I_{ON}/I_{OFF} ratio in the sub-threshold region and its effect on SRAM functionality is addressed by pulling up the BVSS node for un-accessed rows. By making the voltage drop across read-buffer devices close to 0V, leakage from RDBL is highly suppressed. During a read access, however, the BVSS node should be rapidly pulled down and remain low in order to not introduce performance overhead. This requires that the pull down device should be strong enough to sink the current from all accessed columns discharging their corresponding RDBLs.

The design in [14] uses a charge-pump circuit to boost the gate drive of the pull-down NMOS connected to BVSS node. By doubling the gate voltage, the drive strength of this transistor can be increased by ~500X since current has an exponential dependence on gate drive in the sub-threshold region. This enables rapid discharging of the BVSS node and keeping it low during the read access.

A similar scheme is used for this work. However, the charge-pump circuit cannot be activated beyond V_{DD} = 0.6V due to reliability concerns. Applying a voltage higher than the nominal V_{DD} = 1.2V for the process might cause oxide breakdown and functional failure for the SRAM. Moreover, even if the charge-pump is activated and oxide break-down is not be an issue, the current would not increase by 500X in above-V_t region by doubling the gate voltage.

Figure 2-17 shows the BVSS driver and the effect of the pull-down device width in this driver on the performance of the memory. At V_{DD} = 0.6V, there is a sudden drop in the performance of the memory due to the de-activation of charge-pump circuit. This causes an off-region for the voltage range of the memory where it does not make sense to operate. The width of this NMOS pull-down device is increased to make the off-region smaller (Figure 2-17-b). This work uses 10X larger width for the driver NMOS to make the off-region small enough that a continuous performance vs. V_{DD} response can be acquired from the memory.

2.4.4 Sensing Network Design for U-DVS SRAM

The sensing network is an important part in SRAM design. Since a sense-amplifier lies in the critical path of a read-cycle, its delay directly adds up to the total read delay. Figure 2-18 shows a read-cycle in a conventional SRAM design. First, large BL capacitances are pre-charged through PMOS devices during the "pre-charge" phase. After this, depending on the value stored in the bitcell, BLs are discharged or stay high during the "discharging" phase. Lastly, the sense-amplifier is activated to output the data.

In SRAMs, sensing can be single-ended or differential depending on the array architecture and bitcell design. For a 6T cell, BL and BLB are both pre-charged. During the discharging period, a differential voltage develops between BLs and a sense-amplifier amplifies this differential voltage. However, for an 8T cell, single-ended sensing is necessary since only RDBL port is used for read-accesses.

Another important concept is the amount of discharging necessary on the BLs to do a successful sensing. Small-signal and large-signal sensing schemes are described

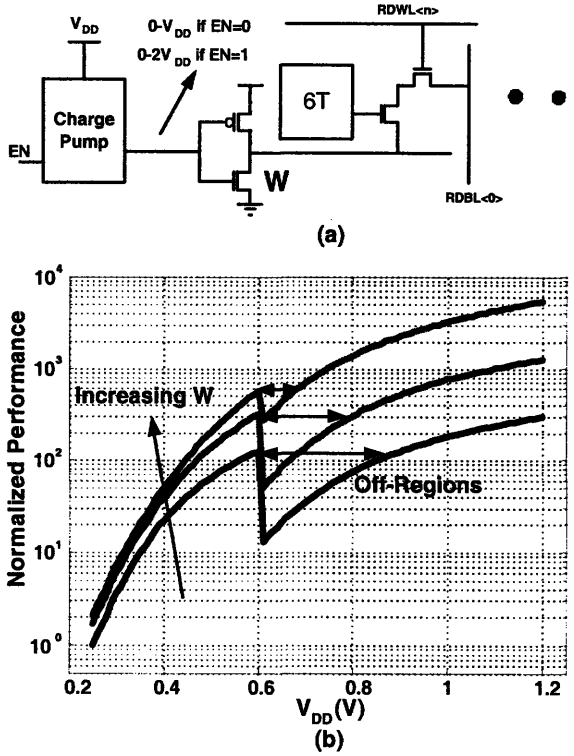

Figure 2-17: BVSS driver (a) and performance vs. V_{DD} plot for the memory with increasing width of pull-down device in the driver (b). Increasing the width by ~10X is necessary to get a nearly continuous plot by minimizing the off-region.

in [33]. In small-signal sensing scheme, the number of bitcells on the BL is chosen to be large resulting in a large BL capacitance. This is why only a small differential voltage on the BLs is allowed to keep the discharging period shorter. For small-signal sensing, a rigorous sense-amplifier structure should be used to sense this small voltage difference. Moreover, the sense-amplifier's offset voltage should be small enough to detect small changes on BL voltages. With the scaling of technology, the amount of transistor mismatch continuously increase causing larger offset voltages for sense-amplifiers. For large-signal sensing, BL capacitance is relatively small so BLs are allowed to discharge all the way to ground. The advantage of this scheme is that a simple logic gate can be used to sense the BL voltage difference.

Previous work proposes different sense-amplifier structures for SRAMs. First of all, the area of the sense-amplifier should be small so that it will not degrade the

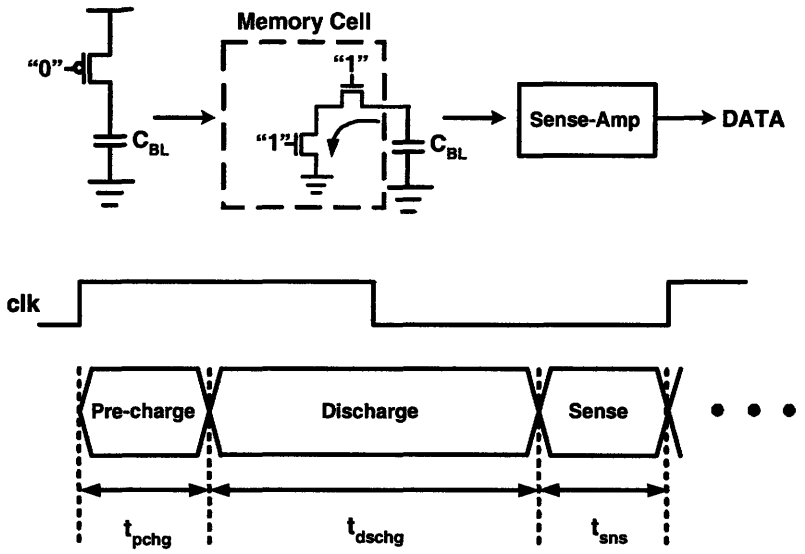

Figure 2-18: Critical path of an SRAM during read operation. Sense amplifier is in this critical path so its delay directly affects total read delay.

array efficiency too severely. The structure shown in Figure 2-19 is proposed in [34] and has been widely used in SRAMs. This design consists of a differential pair with cross-coupled inverters as a load. Four PMOS transistors are used to pre-charge internal nodes to V_{DD} before enabling the sense-amplifier. This prevents any memory effect from affecting the output. A footer NMOS transistor is disabled to reset the state of the structure when the sense-amplifier is not active. If single-ended sensing is required, one of the ports of this structure can be connected to a global reference voltage.

For the U-DVS SRAM, the sensing network should be designed for a large voltage range. At low-voltage levels, the offset and the delay of the sense-amplifier become worse and these problems should be analyzed carefully. Figure 2-20 shows the signals during a read operation for two scenarios. In the first scenario (Figure 2-20-a), the drive strength of the memory cell is much larger than the aggregated leakage of un-accessed memory cells. This causes a differential voltage to develop very quickly. For a single-ended sensing scheme, the reference voltage can be set close to V_{DD} level. In the second scenario (Figure 2-20-b) which is the case in the sub-V_t region, total leakage is comparable to the drive strength of the memory cell. Sensing margin requires the

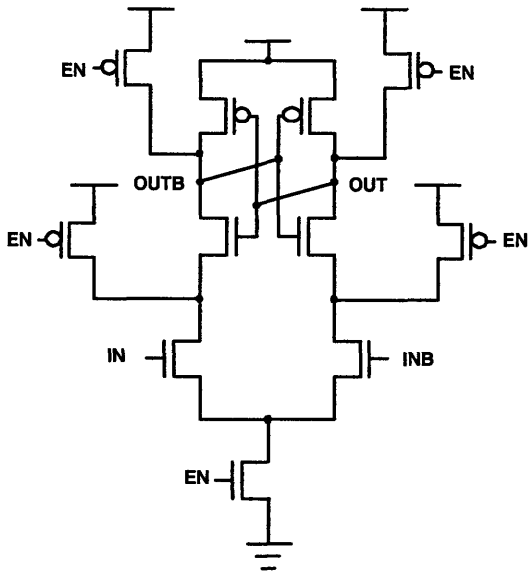

Figure 2-19: Schematic view of a widely-used latch type sense-amplifier. The structure consists of a differential pair which is loaded by a cross-coupled inverter and pre-charge transistors.

reference voltage to be set to a voltage closer to ground level. In the first scenario, the common-mode of the BL voltage and reference voltage is close to V_{DD} whereas in the second scenario, it is closer to 0V.

[14] uses the latch-type sense-amplifier shown in Figure 2-19 with PMOS input devices. For only sub-V_t operation, this is understandable because of the following two reasons:

- In sub-threshold, the drive strength of a PMOS device is higher for this specific technology and
- At very low voltage levels, the common-mode of the input and reference signals is closer to ground level which makes a PMOS input device better with respect to NMOS input devices.

However, at higher voltage levels both of these reasons do change in favor of NMOS devices. This motivates for an approach that will give low sensing delay over a very large voltage range.

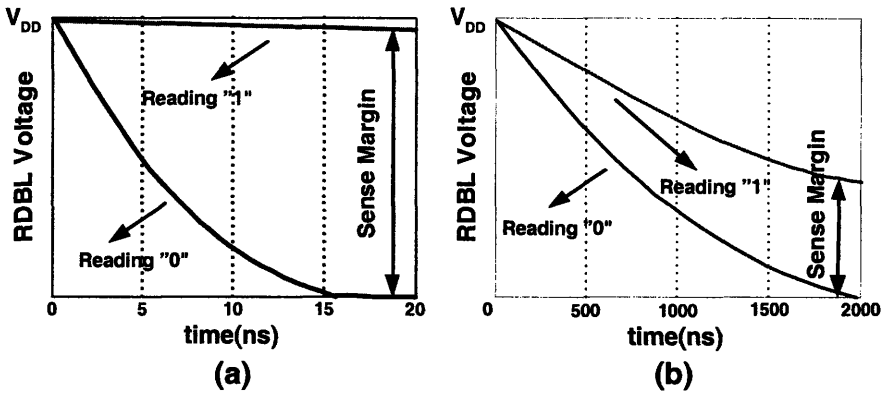

Figure 2-20: Signals during a read operation for two different scenarios. (a) shows the case where drive strength of the memory cell is much larger than the aggregated leakage and (b) shows the case where they are comparable.

Figure 2-21-b shows the sensing network design proposed in this work. Two sense-amplifiers, one with NMOS input devices and the other with PMOS input devices, are implemented with a simple selection logic. At low-voltage levels, the PMOS input sense-amplifier is activated whereas at higher voltage levels, the NMOS input sense-amplifier is activated. The valid output of one of the sense-amplifiers' is multiplexed to the output. There is an area overhead associated with using two sense-amplifiers but it is less than 10% and this is necessary to get an acceptable performance from memory at both the low and high ends of the voltage range.

Figure 2-21-a shows the delay of the NMOS and PMOS input sense-amplifiers for different reference voltage levels. As the reference voltage decreases, the PMOS input sense-amplifier's delay decreases and becomes less than the NMOS input sense-amplifier's delay. At low voltage levels, the reference voltage is close to ground, so using the PMOS sense amplifier provides better performance.

2.5 Test Chip Architecture

A 64kbit 65nm CMOS test chip is designed by using 8T SRAM cell. The array architecture is shown in Figure 2-22. The memory consists of eight 8kbit blocks, where each block contains 64 rows and 128 columns of bitcells along with row and

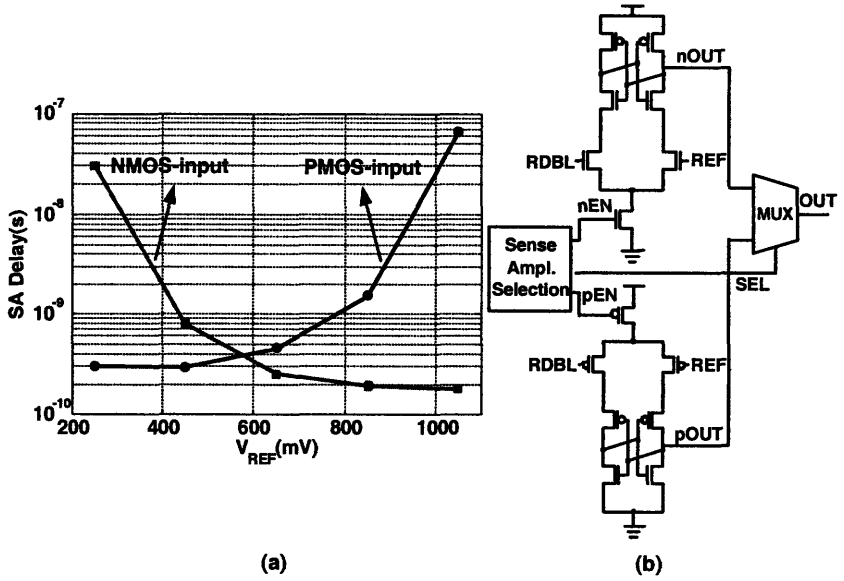

Figure 2-21: Schematic view of the sensing network used in this work (a) and delay of the NMOS input and PMOS input sense-amplifiers at different reference voltage levels (b). One of the two sense-amplifiers is activated depending on the voltage range and the valid output is multiplexed to the output.

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. Depending on the access type, local RDWL or WWL is pulled high through the WL driver circuitry during an access. MCHd Driver, BVSS Driver and WL Driver constitute the row circuitry. The column circuitry, on the other hand, contains the sensing network and BL/BLB drivers. PMOS devices are used to pre-charge RDBLs and they are controlled by pchgB signal.

The power routing for the MCHd Driver and WL Driver in row circuitry and power routing for the column circuitry are done separately. This gives the flexibility to boost WL driver with respect to the memory cell array voltage, and also ease write/read problems. The sense-amplifier can also be operated at a higher voltage level to decrease its delay. A 128 bit word is written and read back at the same time. Column interleaving is not implemented in this design. Soft errors can be addressed with multi-bit ECC schemes. Sixty-four memory cells share BLs to have a good area efficiency without increasing the BL capacitance too much and hence degrading speed

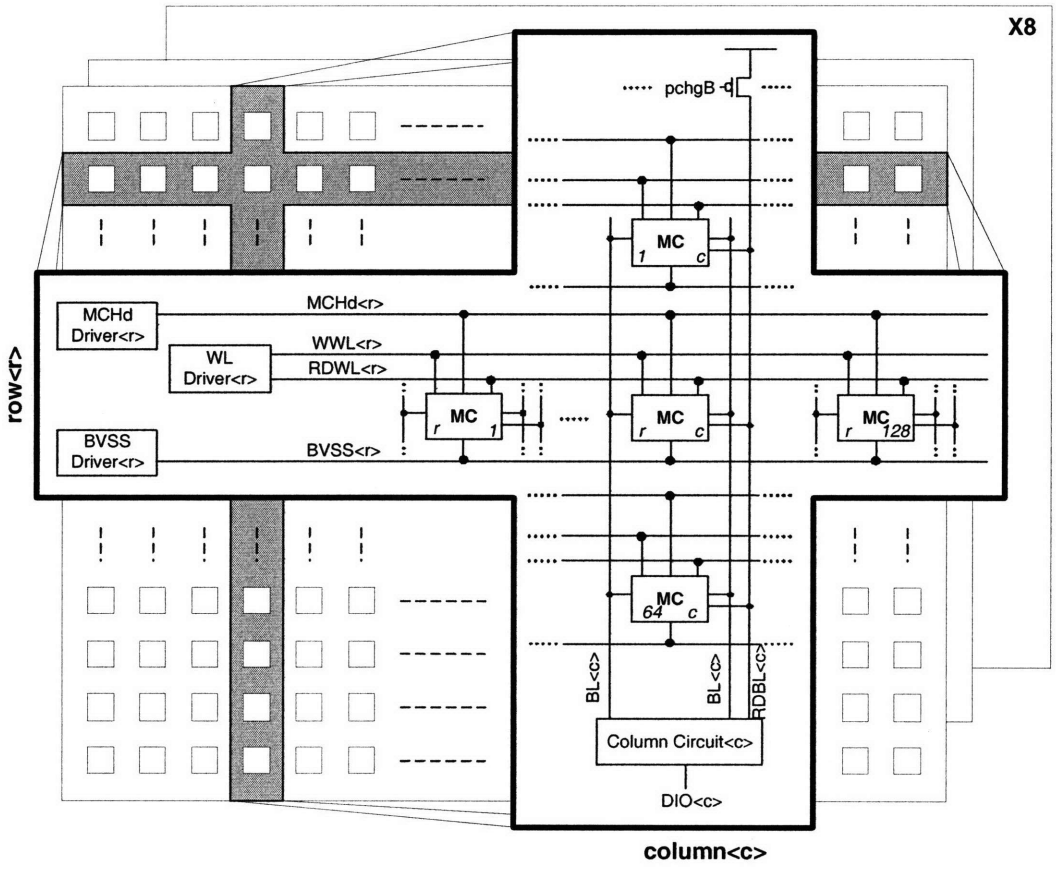

Figure 2-22: Architecture diagram of 64kbit memory is shown. The array consists of eight 8kbit blocks, and each sub-block is composed of 64 rows and 128 columns.

and functionality.

Figure 2-23 shows the architecture of the test chip. Eight sub-blocks are connected to each other through the DIO bus. This DIO bus is also connected to the Write-Registers and Read-Drivers as shown in Figure 2-23. The Write-Registers block is a 128-bit shift-register. In order to ease testing, data1sel signal forces the registers to store a checker-board pattern. When data1sel is '1' registers store '1010...10', whereas when data1sel is '0', registers hold '0101...01'. The Read-Drivers block contain a multiplexing stage and output drivers. Output words of 16-bit length are routed to the pads so three levels of multiplexing and three select bits (rdBrst<0:2>) are used.

Decoders used in the design are constructed by using static CMOS gates to ensure functionality down to very low voltage levels. A 9 bit address word is decoded, and

this decoding is pipelined in order not to increase the critical path. The first three bits of the address selects the active block and causes signals to switch only in the selected block. The rest of the address bits are used to decode the active row. A global WL is asserted causing local WLs to be pulled-high by the row circuitries.

Lastly, Figure 2-24 shows the chip micrograph of the test chip fabricated in a 65nm low-power CMOS process. The die size is 1.4mm by 1mm.

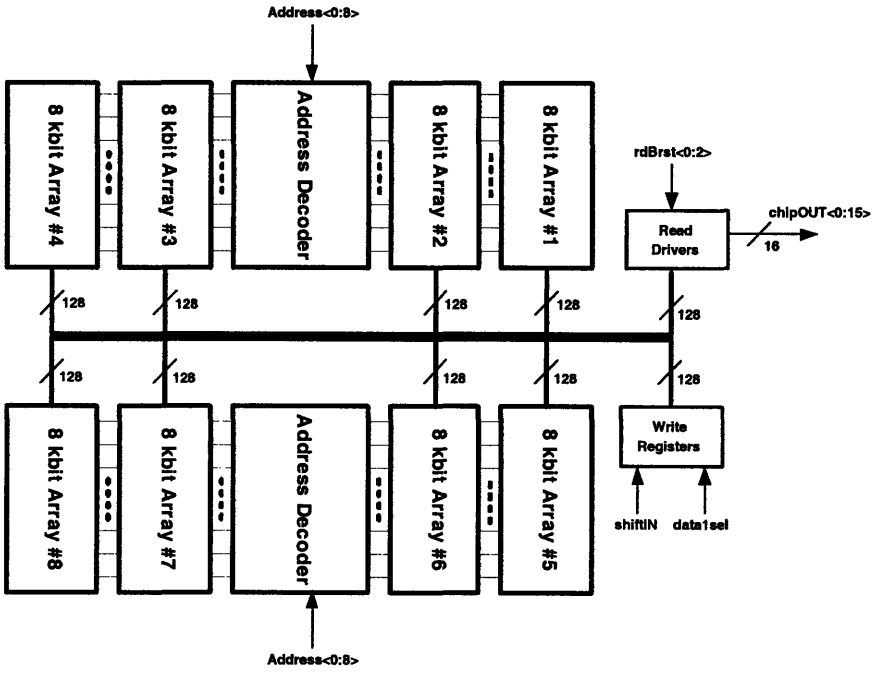

Figure 2-23: Architecture diagram of the test chip is shown. DIO bus, Read-Drivers, Write-Registers and Address Decoder blocks are shown.

2.6 Measurement Results

Measurements of the test chip shows that it functions from 0.25V to 1.2V. No bit errors are encountered over this range. Below 250mV, some of the sense-amplifiers begin to fail. By keeping sensing network's supply voltage at 250mV, the array voltage can be further brought down.

For testing the chip, two different test setups have been created. The first setup is done to do functional testing at low performance levels. The PCB board and

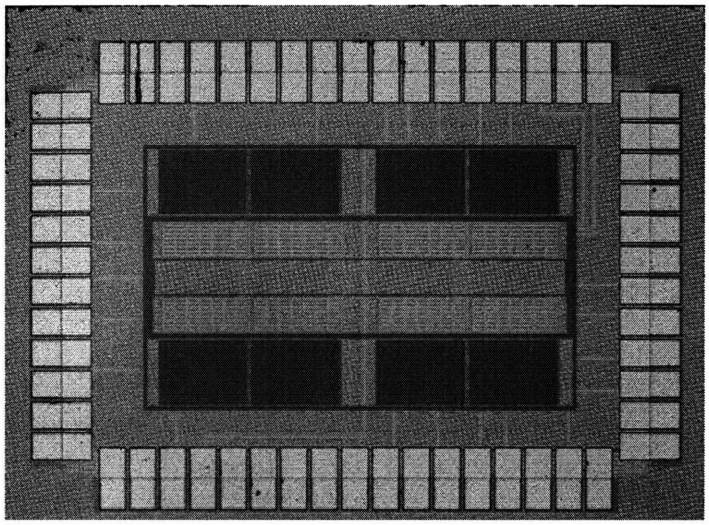

Figure 2-24: Chip micrograph for the 64kbit U-DVS SRAM fabricated in 65nm. Die area is 1.4mm x 1mm.

components were chosen for low-performance measurement. For example, the socket on this board was chosen to be a low speed socket and all inputs to the board were given by the pattern generator. This first board proved functionality and also utilized for low voltage testing from 250mV to 600mV.

The second PCB is designed with SMA connectors and 50 ohms terminations on the board. The socket is also chosen to be compatible with high speed operation. Only inputs starting and ending a write and read operation is input as high speed signals whereas the address inputs were given at ~10MHz. Even though the address inputs switch at a low frequency, two rising edges of clock separated by a few nanoseconds is used to capture the high speed response of the memory. Since the address decoder is pipelined, the delay of decoding is not in the critical path so access period measurement is a valid way to test the high frequency operation.

Figure 2-25 shows the frequency and leakage power vs. V_{DD} plot for the U-DVS SRAM. The memory functions from 20kHz to 200MHz over the 250mV to 1.2V voltage range. This very large frequency range makes this SRAM compatible with numerous low-power applications. The leakage power scales down by ~50X as shown in Figure 2-25. 50X scaling of leakage power is not only due to V_{DD} scaling but also due to DIBL effect. Since leakage power is the main source of power consumption in

SRAMs, this result is very important for low-power circuit design.

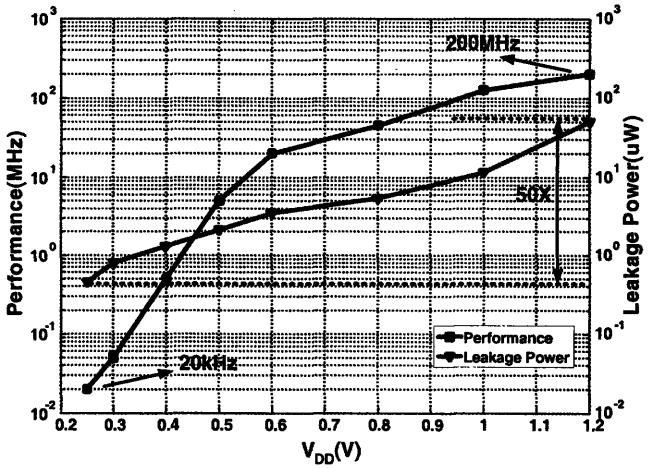

Figure 2-25: Performance and leakage power vs. V_{DD} plot for the U-DVS SRAM. Leakage power scales down by > 50X over the voltage range.

Figure 2-26 plots the active and leakage components of energy/access along with the sum of these two components.

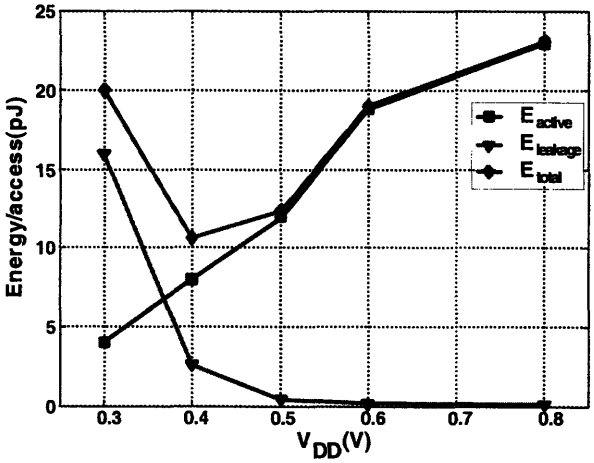

Figure 2-26: Active and leakage components of the energy/access for the test chip. Minimum energy point occurs at ~ 400mV.

The active energy scales down quadratically as expected. The 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. The total energy makes a minimum around 400mV which is also known as the "minimum energy point".

Chapter 3

Cache Design for Low-Power H.264 Video Decoder

This chapter will focus on design considerations for an application specific dynamic-voltage scalable memory.

3.1 Motivation for low-power H.264 video decoder

As discussed in the Introduction section, power consumption of digital systems is an important metric for various reasons. Applications that run from a battery or that are powered-up through energy harvesting techniques need to be very energy efficient to stay within the energy budget of the system. For example, a medical implant that runs from a battery should be replaced through surgery once the battery is depleted. This situation not only causes severe inconveniences, but also can be life threatening. Another example is a handheld device such as a mobile phone. Consumer demand for more functionality from a cell phone caused companies to add more features to these devices. However, customers also do want to have longer battery life for obvious reasons. Designing low-power and energy efficient digital circuits is the only way to fulfill both of these demands.

The work in [35] talks about the "insatiable appetite" for computing power in efficiency-constrained applications and uses mobile phones as an application example.

The projected data shows that by the year of 2011, mobile phones will have 100Mbit/s data bandwidth [36], high speed multimedia video processing, high-definition video-coding, 3D graphics and etc. All these features require beyond 25 giga-operations per second [37]. However battery capacity is predicted to increase by only 44% during the same time frame [38]. Obviously, this large gap cannot be filled by micro-architectural innovations for the overall design. This work suggests that "application-domain-specific platforms (with processing elements optimized for targeted work-load)" are the main candidate to enable all the computing needs with high energy efficiency.

The work in [39] shows a 3.5G baseband-and-multimedia applications processor fabricated in 45nm process. The processor employs many different cores for power efficiency. Multiple power domains and custom designed circuits result in a power scaling of 37% and a performance scaling of 155% at the same time, compared to the previous technology node. This work, for example, uses seperate power routing for the array and memory peripheral circuits in order to be able to get the maximum leakage power savings during sleep mode.

H.264 standard has been getting increasing popularity because of its higher video quality for the same allocated bandwidth. Many companies try to support this video format for higher consumer satisfaction in mobile applications. This new standard has 1.5X more compression capability but 4X more complexity compared to the previous MPEG2 standard. Higher compression is very important for mobile applications since it relaxes the bandwidth requirement. 4X more complexity, of course, should be addressed with architectural innovations to improve energy efficiency. Also custom memories that can operate as low as core logic voltage should be designed specifically for the application to

- lower the dominating portion (memory) of active power consumption and
- simplify the power network of this large system.

The work in [4] shows the power-breakdown of a H.264 video decoder chip in Figure 3-1. On-chip memories account for 22% of the total power consumption of the module. Low-power architectures generally try to use more on-chip caches to reduce

external memory bandwidth which increases the significance of memory power. This makes the custom SRAM design very crucial for low-power H.264 decoder design.

H.264 decoder should also be able to support multiple resolutions which brings a time-variable throughput constraint on the system. This requirement can be satisfied by dynamically adjusting the voltage of the core depending on the throughput constraint. For the highest resolution setting, logic might need to operate at full-V_{DD} levels whereas for the smallest resolution, V_{DD} /2 can be adequate. On-chip caches for the decoder should also be dynamic-voltage-scalable to be able to be compatible with the logic. This requires the memory to be designed for a large voltage range, too. A DVS design similar to the one discussed in Chapter 2 can be used for the caches of this decoder.

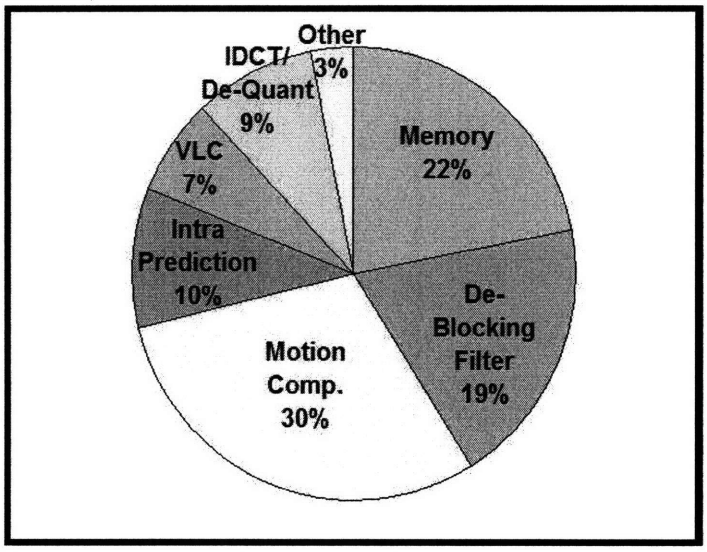

Figure 3-1: Power breakdown of a H.264 video decoder [4]. Memory module accounts for 22% of the total power consumption.

3.2 Design considerations for the DVS caches

Figure 3-2 shows the architecture of the video decoder. The algorithm is implemented in multiple pipeline stages. The majority of the energy is consumed during the motion compensation(MC block), deblocking (DB block) and frame buffer accesses. Frame

buffers are implemented off-chip in order to have a smaller die area.

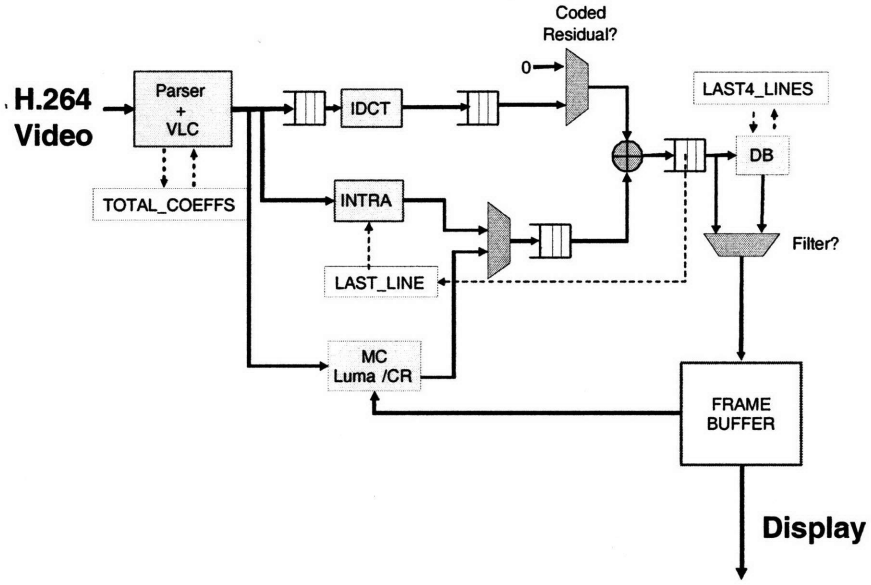

Figure 3-2: Architecture of the H.264 video decoder. Frame buffers are implemented as off-chip components to reduce the die size.

In order to reduce energy/operation in the decoder, supply voltage can be lowered and the performance loss due to voltage scaling can be compensated through parallelism. Because of the opposing trends of active energy and leakage energy, total energy makes a minimum close to sub-V_t region. The latency, on the other hand, is directly related to the drive strength of the transistors and hence increases almost linearly in above-V_t regime and then worsens exponentially in sub-V_t region. By operating at a lower V_{DD} , significant energy savings can be achieved and the increase in latency can be compensated by employing parallelism. For example, deblocking filter architecture can be re-designed to have four filters in parallel. On top of this, luma and chroma components of the frame can be filtered in parallel. These architectural innovations result in a reduction of cycles from 192cycles/MB to 42cycles/MB which is more than 4X.

Since H.264 decoder has to feed the frame buffer with a pre-defined frames/sec rate, there exists a throughput constraint for the system. By leveraging parallelism, in 720p resolution mode, performance required from the system should be around

15MHZ. This speed can be satisfied at or above $\sim 0.6V$. Since 0.6V is in above-V_t regime, SRAMs should be designed for above-V_t operation only, where trade-offs in SRAM design changes significantly compared to the U-DVS design trade-offs as discussed in Chapter 2.

3.3 On-chip Cache Design for H.264 Decoder

The bitcell design for this design should be different than the design presented in Chapter 2. For the U-DVS SRAM, BVSS node inside the cell was required to offset the voltage droop on the BLs through leakage. Since this aggregated leakage can be comparable to cell current in sub-V_t regime, this was necessary. However for a design targeting a minimum operating voltage of 0.6V, aggregated leakage is far from being comparable to the cell current. Figure 3-3 shows the drive current of a 5σ cell divided by the total leakage of 63 memory cells. Even in this worst case scenario, cell current is orders of magnitude larger than the aggregated leakage current above V_{DD} =0.6V.

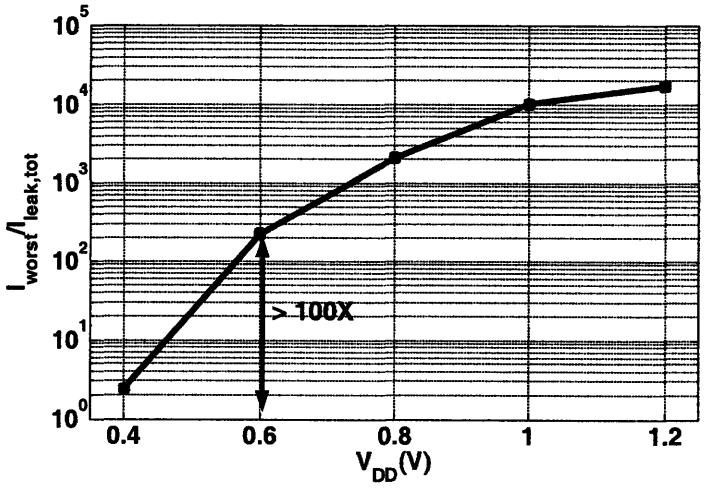

Figure 3-3: 5σ current divided by total leakage through 63 memory cells over V_{DD} range. At 0.6V and above, the aggregated leakage is orders of magnitude less than the worst case drive current.

Consequently, BVSS node can be statically connected to ground inside the cell for this design. This change brings two major advantages:

- Memory performance increases slightly.
- Bitcell design is simplified by using less metal routing

Read performance of the memory highly depends on the discharging time of the RDBL. Since BVSS node is not a static ground and it is pulled down through an NMOS device inside BVSS driver circuit, finite resistance of this transistor causes BVSS voltage to slightly increase during accesses. This causes a degradation in read performance due to a weaker drive current through read-buffer devices. Statically connecting BVSS node to ground ensures that the footer of the read-buffer is always at ground level regardless of the current through read-buffer devices. Figure 3-4 shows the normalized performance vs. V_{DD} plot with and without the virtual ground node (BVSS) inside the cell.

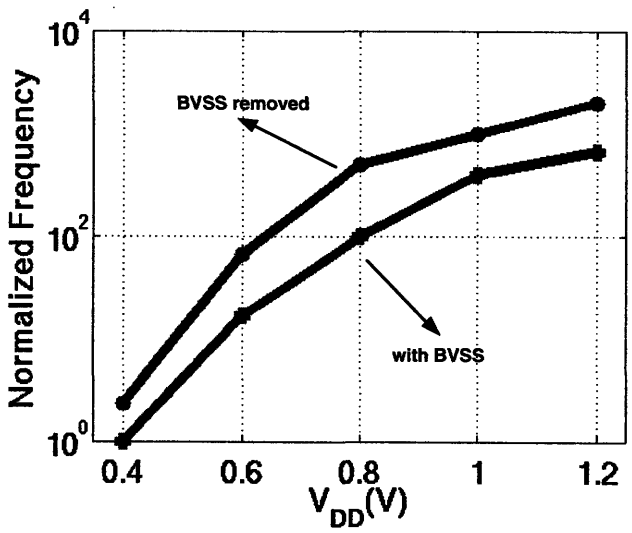

Figure 3-4: Normalized performance vs. V_{DD} plot with and without BVSS node. The resistance of the pull-down driver for BVSS node causes some degradation in performance.

The row-wise BVSS signal should be routed on a separate metal layer and removing this node results in using less metal layers inside the cell. A simpler design means less coupling and higher reliability and hence it is preferred over a more complex design most of the time. Especially for SRAM cells where layout has to be very small, it is preferred to use less metal routing.

The periphery for the H.264 decoder caches also do change due to a larger minimum voltage requirement. First of all, BVSS driver is taken out of the design as explained in the previous paragraphs. Secondly, the sensing network is simplified for above-V_t operation. The motivation for using two sense-amplifiers and activating one of them depending on the voltage level is not valid for an above-V_t SRAM. A single sense amplifier can be used to simplify the design and increase area efficiency. For this design, an NMOS input latch type sense amplifier (Figure 2-19) is used.

The U-DVS chip presented in Chapter 2 is designed as a standalone SRAM chip so some of the critical timing signals are given off-chip. This provides flexibility during testing. However, for a large system, having all signals coming off-chip is not practical. An interface between the SRAM and the rest of the logic should be designed to ensure coherence between blocks.

The interface for the caches is designed with reconfigurability in order to provide some flexibility without bringing excessive area overhead. Since the interface is intended to work over a large voltage range and since the SRAM delay scales much faster than the logic delay, variable delay lines are used to create critical timing signals. Figure 3-5 shows an example circuit which uses an inverter chain and multiplexer connected to different positions on the delay line. By applying different select signals to the multiplexers, different delays can be generated easily.

Some of the peripheral circuits are also designed to have self-timing feature. For example, the sensing network is designed such that whenever the sense-amplifier is activated, its final value is stored into a latch and then sense-amplifier is put back to reset mode automatically. BLs are also pre-charged back to V_{DD} and get ready for another read-operation. All these sequence is self-timed and triggered by the sensing signal which is created by the interface circuit.

3.4 Energy Models of On-Chip Caches

In a large digital system, power consumption of each block can be quite different from each other. In order to make some crucial decisions during design time, energy models

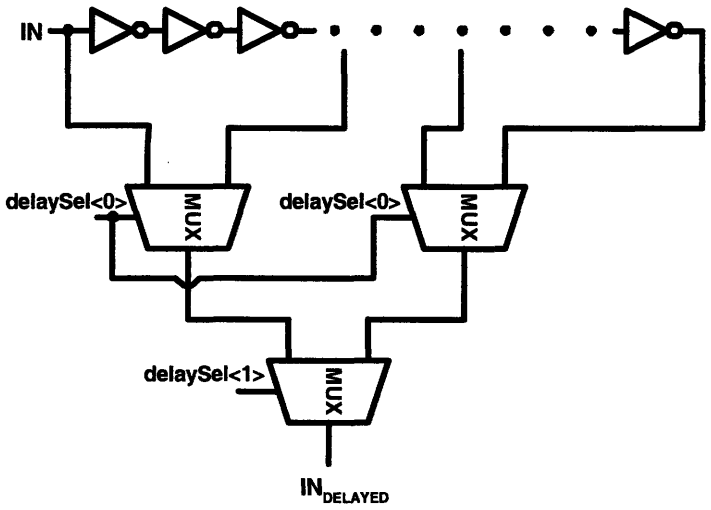

Figure 3-5: Variable delay line implementation used in the design. Inputs of the multiplexers are used to generate different delays. This provides flexibility for the interface circuit which needs to be operational over a large voltage range.

are used. Energy models of memories are created for the H.264 decoder project to provide guidance during the design stage.

For the energy modelling of caches on H.264 decoder chip, three parameters are used:

- Total capacity of the design
- Number of memory cells on a BL
- Number of memory cells sharing a WL

The first parameter determines the number of blocks in the memory and has a direct effect on the total energy consumption of the memory due to leakage. During an access, only one block is activated so the active energy consumption does not depend on the number of blocks in a design. However, since all memory blocks are leaking during idle or active mode, leakage energy is directly proportional to the size of the memory.

Number of memory cells on a BL determines the BL capacitance, which has a direct effect on the performance. Also, during a read/write access, the amount of BL capacitance charged/discharged determines the active energy consumption.

Number of memory cells sharing a WL has relatively less effect on the energy consumption. A larger WL capacitance will only bring an extra delay for the WL assertion. Since total read delay is not dominated by WL assertion time, its effect is marginal.

In order to estimate the energy consumption of the memory, the following assumptions are done. On the average, number of read and write accesses are the same and energy consumption of each access is dominated by charging/discharging of the BLs.

During a read access, RDBL capacitances are pre-charged. On the average, half of these RDBLs are discharged because half of the cells are assumed to store a '0' and the other half is assumed to store a '1'. During read-access period some of the cells storing a '0' can discharge their RDBLs to ground and some of them cannot. Effective switching RDBL capacitance is found by multiplying the number of RDBL by a factor, which is denoted by p in the below equation.

$$C_{RDBL,eff} = 0.5(No.ofColumns)C_{RDBL}p$$

The value of p can be calculated by examining the distribution of the cell read currents. For simplicity, this factor is taken to be close to unity for this model.

During a write access, on the other hand, BL and BLB capacitances are driven to '0's and '1's. Assuming that every cycle half of the BLs will be driven to the opposite of its previous value, effective number of BLs charging up can again be found by multiplying the total number by 0.5.

$$C_{BL,eff} = 0.5(No.ofColumns)C_{BL}$$

For both read and write accesses, WL switching is not included since it is generally negligible compared to BL switching.

For leakage power calculations, the following method is used. First, the leakage of each building block is simulated with their idle mode conditions. Then these leakage power numbers are scaled depending on the number of rows, number of columns and total capacity parameters.

Figure 3-6 shows the energy/access vs. V_{DD} calculated with the generated model with different number of bitcells on a BL. As expected, an increase in the number of cells on a BL results in an increase in the total BL capacitance and hence an increase in energy/access.

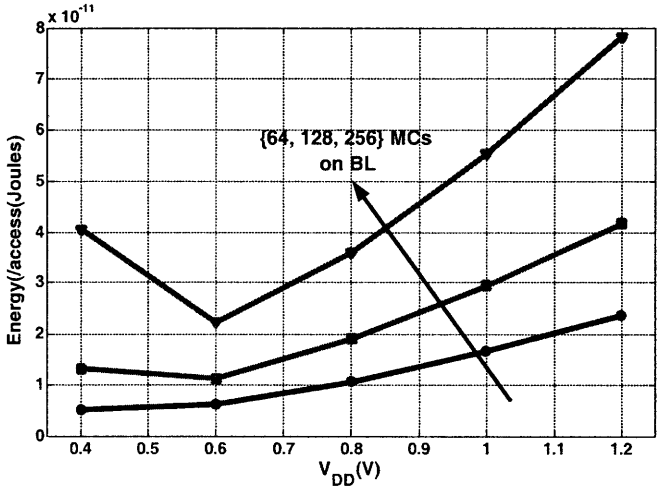

Figure 3-6: Total energy/access vs. V_{DD} plot calculated by the model with different number of bitcells on a BL. Minimum energy point nearly stays the same since both leakage and active components of energy increase.

3.5 Performance Models of On-Chip Caches

In order to estimate the performance of the memory, the following assumption is done. Total read delay is nearly twice the minimum time it requires BLs to be discharged for a successful sensing.

$$t_{read} = 2t_{discharge}$$

This assumption is generally valid in above-V_t regime. However in sub-V_t region, the discharging phase can have more significant effect on the total delay.

For $t_{discharge}$, worst case cell should be considered since it will be limiting the performance of the overall design. To find $t_{discharge}$, first, an average offset voltage(V_{off}) for the sense-amplifier is calculated from simulations. Assuming the offset voltage of

the sense-amplifier is V_{off}, reference voltage should be set to V_{DD} -V_{off}. In order to read a '1' correctly, RDBL should be discharged by at least V_{off} below the reference level which is V_{DD} -$2V_{off}$. This means that, the worst case cell should discharge the RDBL capacitance by at least $2V_{off}$ for a successful read operation. The time it takes for the BL capacitance to be discharged through the worst-case read-buffer by $2V_{off}$ is defined as the half of the total read latency.

This estimate is a pessimistic way of looking at the situation because having a worst-case cell on the same column with a sense-amplifier having V_{off} offset is less probable. However, in order to get an estimate of the performance, this approach is good enough.

Figure 3-7 shows the performance vs. V_{DD} calculated with the generated model with different number of bitcells on a BL. As expected, an increase in the number of cells on a BL results in an increase in the total BL capacitance and hence a degradation in performance.

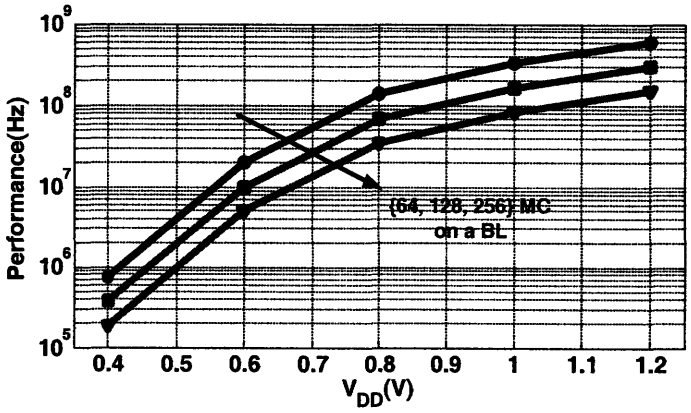

Figure 3-7: Performance vs. V_{DD} plot calculated by the model for different number of bitcells on a BL.

3.6 Test chip architecture

The test chip for the low-power H.264 video decoder is taped-out in November 2007 so the testing results are not ready to be included to this thesis. This chapter will

talk about the chip implementation without giving silicon measurement results.

Table 3.1: On-chip memories used in low-power H.264 video decoder chip

Cache	Cache	Purpose
1	104kbits	last 4 line of pixels for DB filtering
2	21kbits	last 1 line of pixels for intra prediction
3	1.3kbits	last line of intra prediction modes for each 4x4
4	9.4kbits	last line of motion vectors for each 4x4
5	1.6kbits	last line of total-coefficients for each 4x4

Table 3.1 shows the on-chip memories used in the design along with their capacities and their purpose in the design. As can be seen from this table, many memories with different aspect ratios are used in this design. In order to be able to generate/modify these layouts quickly, a memory compiler code is written.

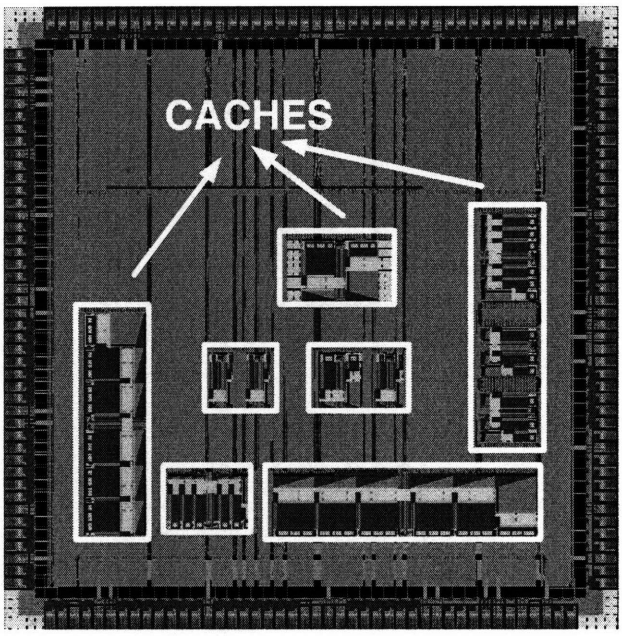

Figure 3-8: Layout of the low-power H.264 video decoder chip. On-chip caches are highlighted on the image.

The memory compiler is coded in Skill. The code takes number of blocks, number of bitcells on a BL and number of bitcells sharing a WL as inputs and generate the schematic and layout view of the memories. The x- and y-dimensions of the basic

building blocks are also parametrized inside the code. For example, the bitcell height and width are both parameters and can be changed very easily. The layout of the bitcell and peripheral circuits are custom designed and these layouts are compatible with the compiler code. Power routing of the cell array, row and column circuits are also created by the compiler depending on the input parameters. Coding a simple compiler gives the designer two advantages:

- First, it allows the designer to generate different memory layouts with a few clicks.
- Secondly and more importantly, by changing the spacing parameters, the same code can be used even if the bitcell and/or peripheral circuit layout is changed.

Figure 3-8 shows the layout of the H.264 video decoder chip. The design is pad limited and have 160 pads. On-chip caches are annotated on the layout view. The memories account for 52% of the total area whereas the standard cell logic only accounts for 14%.

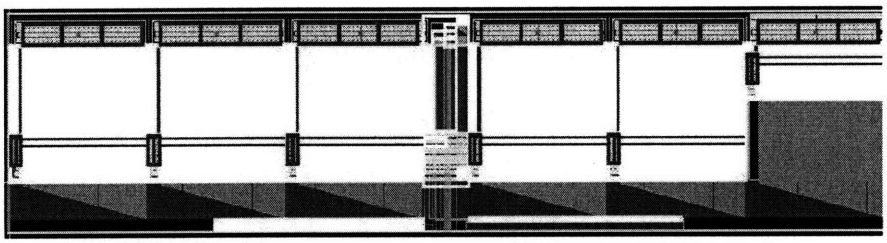

Figure 3-9: Layout of one of the memories used in the decoder.

Figure 3-9 shows the layout view of a memory block used in H.264 video decoder. The same architecture used for the U-DVS memories is implemented for this design, too. A DIO line connects all the sub-blocks and the address decoder selects one of the sub-blocks and the active row during an access.

The number of bitcells sharing a BL is chosen to be 64 and kept fixed for all memories since this number provides a good balance between area efficiency and performance. Different word lengths ranging from 8 to 160 are used for different memories. A single power supply is used for the memory array and all peripheral

circuits to simplify the power routing. An off-chip reference voltage is shared between all sense-amplifiers in the design.

Chapter 4

Conclusions

This thesis focuses on low-power SRAM design considerations in 65nm CMOS process. On-chip memory area as a percentage of total chip area is continuously increasing and causing chip power to be dominated by memory power consumption. Therefore, in order to build a low-power system, designing low-power memories is a necessity.

Theoretical analysis of sub-V_t operation for digital circuits in literature shows that sub-V_t circuits are valuable for energy efficiency. The performance degradation due to lower supply voltages can be addressed by implementing circuits with U-DVS capability. Adjusting the supply voltage of the digital circuits to meet system performance requirement improves energy efficiency of a system. Designing U-DVS SRAMs is an important research problem and crucial for various applications.

For complex digital systems, energy efficiency improvement that can be acquired with micro-architectural innovations is incremental. Instead, application-specific modules optimized for target workload should be designed. Although this approach requires longer design time, it also provides significantly higher energy savings.

This thesis presents two different SRAM designs: A U-DVS design operating across a very large voltage range and an application-specific memory optimized for the requirements of an H.264 video decoder.

U-DVS designs are very challenging since these systems try to operate in two different domains, sub-threshold and above-threshold, with very different trade-offs. This requires circuits to adapt dynamically to the voltage domain that they are

working in. This adaptability is provided by reconfigurable circuits proposed in this thesis. Simple changes to the conventional design can enable reconfigurability and can make low-overhead U-DVS systems possible. For the areas where reconfigurability cannot be employed (i.e. device sizing), the optimum point for the voltage range should be found through simulations. Secondary effects such as DIBL and RSCE can also be exploited to improve design metrics so these effects should be considered carefully.

Designing a memory as a part of a large system introduces more considerations. Developing a voltage-scalable interface is a challenging task but is also a requirement for a U-DVS SRAM. Further, performance and energy models are useful for the memory designer to fully understand the dynamics and trade-offs of the SRAMs.

The following sections summarize the contributions of each test-chip and discuss future work.

4.1 U-DVS SRAM Design in 65nm CMOS

In Chapter 2, an 8T U-DVS SRAM is presented. Fabricated in 65nm low-power process, 64kbit array operates from 0.25V-1.2V range continuously. This large voltage range, including sub-V_t and above-V_t regions, is enabled by hardware reconfigurability proposed in this work. Peripheral assist circuits necessary for sub-V_t functionality are reconfigured by control bits and the overheads due to these assist circuits at high voltage levels are highly suppressed.

Specifically, a novel multiplexed sense-amplifier approach is proposed which provides low sensing delay over the large voltage range. One of the three different write assist schemes are activated for different supply voltage levels to prevent power overhead. An 8T cell is optimized for the large voltage range, and utilization of RSCE for read-buffer devices is proposed to improve low-voltage performance.

The design achieves four orders of magnitude performance scaling (from 20kHz to 200MHz) over the voltage range. Leakage power scales by more than 50X from 1.2V to 0.25V and this result emphasizes the necessity of U-DVS memories for low-power

applications.

4.2 Cache Design for Low-Power H.264 Video Decoder

Chapter 3 presents cache design considerations for low-power H.264 video decoder. Fabricated in 65nm low-power process, preliminary test results show that the chip operates down to 0.7V enabling more than 4X energy savings for the core logic.

The H.264 video decoder application imposes a throughput constraint on the memories to support 30frames/sec display rate. In order to provide this data rate, memories are designed to be scalable from 1.2V down to 0.6V. This voltage range includes only above-V_t region and allows a simpler and application specific design. Performance and energy models of the memories are developed. A voltage scalable interface circuit is designed with programmable delays to create critical timing signals for the memories. A memory compiler code to automatically create schematic and layout views is written in Skill.

4.3 Future Work

Scaling of transistor feature sizes makes low-power memory design more challenging. The effect of variation is exacerbated at low voltage levels and alters device operation significantly. Simply increasing design margins helps make circuits more robust but also results in non-optimized designs. More intelligent circuit techniques should be employed to address these issues.

Sense-amplifier offset, an important variable determining the performance of memories, worsens with increased levels of variation at every process node. Offset compensation schemes generally require large areas and are not compatible with memories' array efficiency targets. An area efficient offset compensation scheme for SRAM sense-amplifiers is an important and open question.

8T memory cell is getting increased attention for its improved low voltage operation capability compared to the traditional 6T design. Current column interleaving schemes, however, prevent the 8T cell from being functional at low voltages due to the read margin problem. Architectural innovations with minimal area overhead are desired to enable low voltage operation for an interleaved memory array composed of 8T cells. This problem is also an open question at the present time and future research can propose possible solutions.

Bibliography

[1] S. Borkar, "Obeying Moore's Law beyond 0.18 micron [microprocessor design]," in *IEEE International ASIC/SOC Conference*, Sept. 2000, pp. 26–31.

[2] K. Zhang, U. Bhattacharya, Z. Chen, F. Hamzaoglu, D. Murray, N. Vallepalli, Y. Wang, B. Zheng, and M. Bohr, "A 3-GHz 70Mb SRAM in 65nm CMOS Technology with Integrated Column-Based Dynamic Power Supply," in *IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers*, Feb. 2005, pp. 474–475.

[3] 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 International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers*, Feb. 2007, pp. 330–331.

[4] T.-M. Liu, T.-A. Lin, S.-Z. Wang, and C.-Y. Lee, "A low-power dual-mode video decoder for mobile applications," *IEEE Communications Magazine*, vol. 44, no. 8, pp. 119–126, Aug. 2006.

[5] G. E. Moore, "Cramming More Components onto Integrated Circuits," *Electronics*, vol. 38, no. 8, pp. 114–117, Apr. 1965.

[6] R. K. Krishnamurthy, A. Alvandpour, S. Mathew, M. Anders, V. De, and S. Borkar, "High-performance, Low-power, and Leakage-tolerance Challenges for Sub-70nm Microprocessor Circuits," in *IEEE European Solid-State Circuits Conference (ESSCIRC) Digest of Technical Papers*, 2002, pp. 315–321.

[7] K. Takeda, Y. Hagihara, Y. Aimoto, M. Nomura, Y. Nakazawa, T. Ishii, and H. Kobatake, "A Read-Static-Noise-Margin-Free SRAM Cell for Low-Vdd and High-Speed Applications," in *IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers*, Feb. 2005, pp. 478–479.

[8] P. Macken, M. Degrauwe, M. V. Paemel, and H. Oguey, "A Voltage Reduction Technique for Digital Systems," in *IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers*, Feb. 1990, pp. 238–239.

[9] R. M. Swanson and J. D. Meindl, "Ion-Implanted Complementary MOS Transistors in Low-Voltage Circuits," *IEEE Journal of Solid-State Circuits*, vol. SC-7, no. 2, pp. 146–153, Apr. 1972.

[10] J. Burr and A. Peterson, "Ultra Low Power CMOS Technology," in *3rd NASA Symposium on VLSI Design*, 1991, pp. 4.2.1–4.2.13.

[11] A. Wang, A. Chandrakasan, and S. Kosonocky, "Optimal Supply and Threshold Scaling for Sub-threshold CMOS Circuits," in *IEEE Computer Society Annual Symposium on VLSI*, Apr. 2002, pp. 7–11.

[12] A. Wang and A. Chandrakasan, "A 180mV FFT Processor Using Sub-threshold Circuit Techniques," in *IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers*, 2004, pp. 292–293.

[13] B. Calhoun and A. Chandrakasan, "A 256-kbit Sub-threshold SRAM in 65nm CMOS," in *IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers*, Feb. 2006, pp. 628–629.

[14] N. Verma and A. Chandrakasan, "A 65nm 8t sub-vt sram employing sense-amplifier redundancy," in *IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers*, Feb. 2006, pp. 328–329.

[15] B. Zhai, D. Blaauw, D. Sylvester, and S. Hanson, "A Sub-200mV 6T SRAM in 0.13um CMOS," in *IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers*, Feb. 2007, pp. 332–333.

[16] 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 5.3ghz 8t-sram with operation down to 0.41v in 65nm cmos," in *Symposium on VLSI Circuits (VLSI) Digest of Technical Papers*, June 2007, pp. 252–253.

[17] Y. Morita, H. Fujiwara, H. Noguchi, Y. Iguchi, K. Nii, H. Kawaguchi, and M. Yoshimoto, "An Area-Conscious Low-Voltage-Oriented 8T-SRAM Design under DVS Environment," in *Symposium on VLSI Circuits (VLSI) Digest of Technical Papers*, June 2007, pp. 256–257.

[18] C. Enz, F. Krummenacher, and E. Vittoz, "An Analytical MOS Transistor Model Valid in All Regions of Operation and Dedicated to Low-Voltage and Low-Current Applications," *Journal on Analog Integrated Circuits and Signal Processing*, pp. 83–114, July 1995.

[19] K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand, "Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits," *Proceedings of the IEEE*, vol. 91, no. 2, pp. 305–327, Feb. 2003.

[20] M. J. M. Pelgrom, A. C. J. Duinmaijer, and A. P. G. Welbers, "Matching Properties of MOS Transistors," *IEEE Journal of Solid-State Circuits*, vol. 24, no. 5, pp. 1433–1439, Oct. 1989.

[21] E. Seevinck, F. List, and J. Lohstroh, "Static Noise Margin Analysis of MOS SRAM Cells," *IEEE Journal of Solid-State Circuits*, vol. SC-22, no. 5, pp. 748–754, Oct. 1987.

[22] K. Takeda, H. Ikeda, Y. Hagihara, M. Nomura, and H. Kobatake, "Redefinition of Write Margin for Next-Generation SRAM and Write-Margin Monitoring Circuit," in *IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers*, Feb. 2006.

[23] J. Wang and B. H. Calhoun, "Canary Replica Feedback for Near-DRV Standby VDD Scaling in 90nm SRAM," in *Custom Integrated Circuits Conference (CICC) Digest of Technical Papers*, Sept. 2007, pp. 29–32.

[24] A. Singhae and R. Rutenbar, "Statistical Blockade: a novel method for very fast Monte Carlo simulation of rare circuit events and its application," in *DATE*, 2007, pp. 1–6.

[25] L. Chang, D. M. Fried, J. Hergenrother, W. Sleight, R. H. Dennard, R. K. Montoye, L. Sekaric, S. J. McNab, N. W. Topol, C. D. Adams, K. W. Guarini, and W. Haensch, "Stable SRAM Cell Design for the 32nm Node and Beyond," in *Symposium on VLSI Circuits (VLSI) Digest of Technical Papers*, June 2005, pp. 128–129.

[26] B. Calhoun and A. Chandrakasan, "A 256-kbit Sub-threshold SRAM in 65nm CMOS," in *IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers*, Feb. 2006, pp. 628–629.

[27] M. Yamaoka, N. Maeda, Y. Shinozaki, Y. Shimazaki, K. Nii, S. Shimada, K. Yanagisawa, and T. Kawahara, "Low-Power Embedded SRAM Modules with Expanded Margins for Writing," in *IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers*, Feb. 2005, pp. 480–481.

[28] K. Itoh, A. Fridi, A. Bellaouar, and M. Elmasry, "A Deep Sub-V, Single Power-Supply SRAM Cell with Multi-V_T, Boosted Storage Node and Dynamic Load," in *Symposium on VLSI Circuits (VLSI) Digest of Technical Papers*, June 1996, pp. 132–133.

[29] K. Kanda, T. Miyazaki, M. K. Sik, H. Kawaguchi, and T. Sakurai, "Two Orders of Magnitude Leakage Power Reduction of Low Voltage SRAM's by Row-by-Row Dynamic V_{DD} Control (RRDV) Scheme," in *IEEE International ASIC/SOC Conference*, Sept. 2002, pp. 381–385.

[30] M. Yamaoka, K. Osada, and K. Ishibashi, "0.4-V Logic Library Friendly SRAM Array Using Rectangular-Diffusion Cell and Delta-Boosted-Array-Voltage Scheme," in *Symposium on VLSI Circuits (VLSI) Digest of Technical Papers*, 2002, pp. 170–173.

[31] C. Kim, J. Kim, I. Chang, and K. Roy, "PVT-Aware Leakage Reduction for On-Die Caches with Improved Read Stability," in *IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers*, Feb. 2005, pp. 482–483.

[32] B. Yu, E. Nowak, K. Noda, and C. Hu, "Reverse Short-Channel Effects and Channel-Engineering in Deep-Submicron MOSFETs: Modeling and Optimization," June 1996, pp. 162–163.

[33] K. Zhang, K. Hose, V. De, and B. Senyk, "The Scaling of Data Sensing Schemes for High Speed Cache Design in Sub-0.18μm," in *Symposium on VLSI Circuits (VLSI) Digest of Technical Papers*, 2000, pp. 226–227.

[34] 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 Journal of Solid-State Circuits*, vol. 28, no. 4, pp. 523–527, Apr. 1993.

[35] M. Muller, "Embedded Processing at the Heart of Life and Style," in *IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers*, June 2008, pp. 32–37.

[36] "3^{rd} Generation Partners Project (3GPP), Long Term Evaluation of the 3GPP radio technology ." [Online]. Available: http://www.3gpp.org/Highlights/LTE/LTE.htm, 2006

[37] Y. Lin, H. Lee, and M. W. al, "SODA: A Low-power Architecture For Software Radio," in *33rd International Symposium on Computer Architecture (ISCA'06)*, 2006, pp. 89–101.

[38] "International Technology Roadmap for Semiconductors (ITRS) Systemdrivers 2005." [Online]. Available: http://www.itrs.net/Links/2005ITRS/Home2005.htm

[39] G. Gammie, A. Wang, M. Chau, S. Gururajarao, R. Pitts, F. Jumel, S. Engel, P. Royannez, R. Lagerquist, H. Mair, J. Vaccani, G. Baldwin, K. Heragu, R. Mandal, M. Clinton, D. Arden, and U. Ko, "A 45 nm 3.5G Baseband-and-Multimedia Applications Processor using Adaptive Body-Bias and Ultra-Low-Power Techniques," in *IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers*, Feb. 2008, pp. 258–259.