DESIGNING LOW POWER SRAM SYSTEM USING ENERGY COMPRESSION

Approved by:

Prof. Saibal Mukhopadhyay, Committee Chair, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Prof. Sudhakar Yalamanchili
School of Electrical and Computer Engineering
Georgia Institute of Technology

Prof. Moinuddin K. Qureshi
School of Electrical and Computer Engineering
Georgia Institute of Technology

Date Approved: April 1st 2013
I dedicate this thesis to my mother Mythili Jayaprakash, father Jayaprakash Nair and brother Pradeesh Nair, because of whom, I am what I am today. They motivated and encouraged me at every step. I owe immense gratitude to my friends Darshan, Amar, Melroy, Santosh, Shobhit, Amit, Ankit, Ankita, Anagha, Allhad, Amey, Aditi and Bhakti who always stood by me at all times during my master’s degree, motivating me.
The power consumption in commercial processors and application specific integrated circuits increases with decreasing technology nodes. Power saving techniques have become a first class design point for current and future VLSI systems. These systems employ large on-chip SRAM memories. Reducing memory leakage power while maintaining data integrity is a key criterion for modern day systems. Unfortunately, state of the art techniques like power-gating can only be applied to logic as these would destroy the contents of the memory if applied to a SRAM system. Fortunately, previous works have noted large temporal and spatial locality for data patterns in commercial processors as well as application specific ICs that work on images, audio and video data. This thesis presents a novel column based Energy Compression technique that saves SRAM power by selectively turning off cells based on a data pattern. This technique is applied to study the power savings in application specific integrated circuit SRAM memories and can also be applied for commercial processors. The thesis also evaluates the effects of processing images before storage and data cluster patterns for optimizing power savings.
ACKNOWLEDGEMENTS

I would like to thank the committee for providing me valuable insights in my meeting with them. I also thank my advisor, Prof. Saibal Mukhopadhyay, for giving me constructive directions during my thesis. I owe my gratitude to members of the GREEN Lab namely Denny Lie, Wen Yueh and Amit Trivedi for helping me gain insights into images and compression techniques and Muneeb Zia who I worked with during the development phase of this thesis. I am grateful to Amit K. for reviewing my thesis. I would like to thank everyone who gave me valuable suggestions during the course of my thesis.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEDICATION</td>
<td>iv</td>
</tr>
<tr>
<td>PREFACE</td>
<td>v</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>ix</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>x</td>
</tr>
<tr>
<td>I INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>II BACKGROUND AND MOTIVATION</td>
<td>4</td>
</tr>
<tr>
<td>2.1 Memory System using SRAM: Organization</td>
<td>4</td>
</tr>
<tr>
<td>2.2 Leakage Power in SRAM</td>
<td>5</td>
</tr>
<tr>
<td>2.3 Image Storage for Application Specific Integrated Circuits</td>
<td>6</td>
</tr>
<tr>
<td>III SYSTEM DESIGN: CONSIDERATIONS AND IMPLEMENTATION</td>
<td>9</td>
</tr>
<tr>
<td>3.1 Design Considerations and Working</td>
<td>12</td>
</tr>
<tr>
<td>3.2 7T SRAM Cell Design</td>
<td>13</td>
</tr>
<tr>
<td>3.3 Compress Group</td>
<td>14</td>
</tr>
<tr>
<td>3.4 Zero-Switch Cell</td>
<td>14</td>
</tr>
<tr>
<td>3.5 Dual Write Pulse Generator</td>
<td>15</td>
</tr>
<tr>
<td>3.6 Modification in Row Decoder</td>
<td>15</td>
</tr>
<tr>
<td>3.7 Modification in the Write Circuitry</td>
<td>16</td>
</tr>
<tr>
<td>IV RESULTS AND ANALYSIS</td>
<td>17</td>
</tr>
<tr>
<td>4.1 Data Read</td>
<td>17</td>
</tr>
<tr>
<td>4.2 Data Write</td>
<td>18</td>
</tr>
<tr>
<td>4.3 Compress Group: Turning OFF</td>
<td>19</td>
</tr>
<tr>
<td>4.4 Compress Group: Functionality Analysis</td>
<td>19</td>
</tr>
<tr>
<td>4.5 Power Analysis</td>
<td>20</td>
</tr>
<tr>
<td>4.5.1 Power Consumption of SRAM cells</td>
<td>20</td>
</tr>
</tbody>
</table>
4.5.2 Power Consumption of Compress-Group .......................... 21
4.5.3 En-Com System: Power Analysis ................................. 22
4.6 Dynamic Power Consumption ....................................... 23
4.7 Enhancement to Improve Power Savings ............................ 24
4.8 Layout of the 7T SRAM Cell ....................................... 25

V SUMMARY ................................................................. 27
LIST OF TABLES

1  Compress-Group Area Overhead ........................................ 10
2  Parameters of the Energy Compressed System ...................... 12
3  Power Consumption Comparison ....................................... 22
LIST OF FIGURES

1 A 6T SRAM Cell .............................................. 4
2 SRAM Memory System ........................................ 5
3 The size of memory occupied by the raw-image ............. 7
4 The proportion of SRAM cells that store a value of ‘0’ to total number of SRAM cells ........................................... 7
5 The size of the group is varied and the proportion of compress-groups is compared to total groups. The proportion of compress group is found to exponentially decrease with increasing group size ......................... 8
6 Row Based Compression increases the per cell height by four ‘Layer 3’ metal lanes per cell. Column Based Compression increases the per cell width by two Layer 2 metal layers only ............................................ 9
7 ‘En-Com’ implementation (Schematic). The top and bottom pins in the decoder and the SRAM array represent the lines to the Zero-Switch Cells ........................................ 11
8 7T SRAM cell (Logical View) .................................. 13
9 An 8 Cell Compress Group ...................................... 14
10 En-Com requires one NOT gate and two AND gates to select between the top or bottom Zero-Switch Cell in case of a write .............................................. 15
11 The write circuit is modified to support dual write in case of data value ‘1’. The pulse generator enables the dual write ......................... 16
12 Waveforms that show that data being read from a row in En-Com .......... 17
13 Waveforms that show the read operation of a row in En-Com ............ 18
14 The waveforms showing the turning OFF of the compress-group. At 3ns, the 7T SRAM cell is tristated due to power gating. The 7T cell stores 0 strongly. ................................................................. 19
15 Steady state operation is shown. The system is run for 30µS and the compress group is switched OFF and ON. The data value in the 7T SRAM cell is retained .............................................................. 20
16 Comparison of the Power Consumption with varying voltage for a) Zero-Switch Cell b) 7T SRAM Cell c) 6T SRAM (Baseline) .......... 21
17 Comparison of the Power Consumption when the Compress Group is switched ON and OFF. The power consumption of the compress-group is reduced by roughly 6.5 times, this includes the power consumed by the Zero-Switched cell. .............................................................. 22
<table>
<thead>
<tr>
<th>Page</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>Normalized power saved due to <em>En-Com</em> when compared with baseline</td>
</tr>
<tr>
<td>19</td>
<td>Dynamic power comparison of <em>En-Com</em> with baseline system. The write</td>
</tr>
<tr>
<td></td>
<td>power in <em>En-Com</em> is two times higher than the baseline in the worst</td>
</tr>
<tr>
<td></td>
<td>case. The read power of <em>En-Com</em> remains almost same</td>
</tr>
<tr>
<td>20</td>
<td><em>En-Com</em> discussed in this paper until now uses Run-length information</td>
</tr>
<tr>
<td></td>
<td>to form compress-group. Another way to form this group is to use</td>
</tr>
<tr>
<td></td>
<td><em>Clustered En-Com</em></td>
</tr>
<tr>
<td>21</td>
<td>Normalized power saving due to <em>Clustered En-Com</em> when compared with</td>
</tr>
<tr>
<td></td>
<td>baseline</td>
</tr>
<tr>
<td>22</td>
<td>The baseline split wordline 6T layout</td>
</tr>
<tr>
<td>23</td>
<td>The 7T layout occupies 15% more area when compared to the 6T layout</td>
</tr>
</tbody>
</table>
CHAPTER I

INTRODUCTION

SRAM (Static Random Access Memories) is used in processing image frame buffers\([1, 2, 3]\). Videos have multiple frames that need to be stored. SRAM provide a low latency solution for comparing master video frames with subsequent frames. SRAM have low access time and can service a request for data elements quicker than DRAM (Dynamic Random Access Memories). They are fabricated on-chip, thus allowing them to use the same technology process in their manufacture as that of the core and other on-chip logic. SRAM is a volatile memory element, but does not require refreshing like that of DRAM. As there is high amount of temporal locality between subsequent video frames, usually the differential between the master frame and adjacent frames is stored on disk. The master frame is stored in a SRAM system that predominantly only reads this frame\([1, 2]\).SRAM memory system for image and video processing in application specific integrated circuits (ASICs) consume upto 81% of the power as standby/leakage power\([4]\).

The power dissipation of the chip is dictated by the amount of power consumed by unit area. The total power in the system is comprised of two components, active power and idle power. Active power is consumed when the processor chip is busy. This power is used while processing data. Since the usage patterns of commercial electronic appliances imply that a device need not be busy all the time, there are gaps of idle periods. Transistors consume idle power (leakage power) during all times. At sub-nanometer nodes, the leakage power of transistors increases and their threshold voltage \(V_{th}\) is found to reduce. A lower \(V_{th}\) results in more leaky transistors as the pressure differential of Drain-Source voltages induces a small sub-threshold current even when the device is not active. This results in increased idle power, thus increasing the overall system power even when the processor or
application specific IC is not active.

SRAM consumes significant area of the on-chip real estate. For video processing, this number has increased as the number of processing engines has increased. Since each processing engine can operate on independent portion of memory, an increase in number of processor engines require an increase in SRAM size for better throughput[2]. The size of SRAM impact system performance and most processing engines are allocated significant on chip area for SRAM system. Leakage power is determined by the number of SRAM cells. As the size of SRAM system in commercial appliances increases leakage power for SRAM is found to dominate the total power consumption.

Commercial processors and ASICs reduce operating power using techniques like power gating and frequency scaling for cores or processing engines and other on-chip logic circuits during idle modes [5]. Power gating drives the supply nodes (VDD/GND) into high impedance by using power gating transistors. This results in reduced power into the logic circuitry but also results in tri-stating these logic circuitry. Memory elements store data and validity of data cannot be compromised. Most large memory elements like L3 (Level Three) caches have additional error correcting circuits to ensure data fidelity. Since for memories, one cannot afford to have a loss in data reliability as it results in incorrect execution or crashing of applications that are run by the cores or incorrect processing of video frames and images by the processing engines. Thus we cannot use techniques like power gating for SRAM system, as tri-stating these elements result in data loss.

Studies show that processors and image,audio or video processing engines operate on data that is biased towards a particular data value zero or one [6, 7, 8]. This results in SRAM memories having larger number of data cells that these biased values. Thus architecting memories that have biased properties for storing biased values has been suggested in some previous literature.

Some techniques that are used to mitigate this problem include lowering the supply voltage across the SRAM cells. As leakage exponentially reduces with decrease in supply
voltage, this results in reduced idle power. Data retention voltage (DRV) of an SRAM cell defines the minimum voltage that must be applied across the SRAM cell before it will flip. This voltage places a limit on supply voltage reduction. Due to process variations, across the chip, huge variations in DRV are observed[9]. This thesis analyzes traditional 6T SRAM designs for image storage, specifically for long idle periods of time and also suggests a novel implementation of 7T SRAM cell. Oriented for low nanometer nodes (sub 22nm), the proposal targets system power and suggests an alternate design called ‘Energy Compressed SRAM system’ that exhibits 15% lower power consumption. These proposals are targeted to be designed for large area and low activity memories consuming large amount of idle power. The thesis will consider ASIC’s which operate on images.

**Thesis Goal: To suggest a novel SRAM system targeted for image/video processing applications for saving leakage power**

This thesis is organized as follows, Chapter 2 describes the Background and Motivation, Chapter 3 introduces *Energy Compressed* System design for low power operations. Chapter 4 evaluates system performance. Chapter 5 give a summary of the implementation and suggests areas to be investigated.
CHAPTER II

BACKGROUND AND MOTIVATION

A SRAM cell employs a cross coupled inverter structure to store data. Figure 1 shows the cross coupled inverter structure and its equivalent 6T (6 transistor) circuit. This structure is prevalent in processor caches and ASIC memory due to its simple design [10].

![Figure 1: A 6T SRAM Cell](image)

2.1 Memory System using SRAM: Organization

A memory system with SRAM consists of banks and each bank consists of multiple subarrays. Subarrays are internally organized into rows and columns. Columns consist of SRAM cells connected to a pair or bitlines. The columns are grouped and multiplexed to a sense amplifier. Rows consist of SRAM cells that share a common word Line. Figure 2 shows the full system that using SRAM. To read data, the pair of bit lines need to be precharged. To access an SRAM cell, the address is decoded upto the subarray level, asserting the appropriate word line and selecting one column from the group. The column address helps select one pair of column bit lines to the sense amplifier. Depending on the
value of the data stored, on activating the cell, one of these bitlines will start to discharge. Data is detected using the sense amplifier which amplifies the voltage differential. The output of the sense amplifier is the stored bit value (having voltage levels VDD [logic 1] or Zero [logic 0]).

![Diagram of SRAM Memory System](image)

Figure 2: SRAM Memory System

### 2.2 Leakage Power in SRAM

Power consumption of a large memory system with SRAM is dominated by leakage power at nanometer nodes. In ASICs used in cameras and in commercial processors (specifically last level caches), memories store data for long time and have low activity. Their power consumption is due to idle power which is dissipated due to leakage currents. At nanometer nodes, subthreshold, gate and reversed-biased junction leakage currents exist. As the technology node get reduced, the threshold voltage ($V_{th}$) of transistors reduces. This increases sub-threshold leakage at nanometer nodes and it dominates the total leakage[11].

There are two components of subthreshold leakage in SRAM, cell leakage and bit line leakage. Figure 1 describes the direction of cell leakage and bit line leakage. Bit line leakage can occur in any direction and it depends on the data stored within the cell. Cell leakage
occurs from VDD towards GND [12]. Cell leakage is caused due to the supply voltage differential across these SRAM cells. Bit line leakage is due to voltage differential between the storage nodes and the bit lines and is found to be much lesser than cell leakage[13]. Techniques such as body biasing of access transistors ensure that bit line leakage can be reduced [14]. Low power SRAM schemes focus on reducing the supply voltage or increasing the $V_{th}$ of individual transistors in the SRAM cell.

Initial SRAM low power techniques proposed employing dual $V_t$ for reducing the leakage power and some innovations in decoding[15]. Most of the modern SRAM cells employ these techniques already and still have high leakage currents. This is because in nanometer nodes, most of these techniques for controlling leakage become ineffective. Leakage power in 6T SRAM cells can be reduced by using reducing gate voltage and body biasing the access transistors. Dual biasing and PMOS transistors can also be used to reduce leakage[14].

SRAM can be power gated towards a nominal supply voltage at sub-array granularity to reduce the idle power consumption [16]. This proposal leverages on data retention characteristics of these SRAM cells, such that the supply voltage is maintained above the data retention voltage (DRV) of a majority of these SRAM cells. Data retention voltage is the voltage above which data integrity for a SRAM cell is assured with a high probability. However the DRV value tends to vary within a large cluster of cells. As technology node decrease DRV increases causing bit flips in SRAM cells leading to retention failures. Thus current last level caches employ ECC to provide some protection in case of errors [17, 18]. At a system level, many manufacturers use DRV characteristics to employ fault tolerant SRAM systems [19].

2.3 Image Storage for Application Specific Integrated Circuits

For the purpose of initial design, image benchmarks (using unprocessed raw images) are analyzed. These benchmarks are publically available and is used for research on image and
video processing[20, 21, 22, 23]. The size of the SRAM system that will be required to store a single image is shown in Figure 3. On an average, the image sizes in the benchmark tend to be nearly 1MB for raw bitmaps.

![Image Benchmarks](image1)

**Figure 3:** The size of memory occupied by the raw-image

Figure 4 shows the total number of cells that store logic ‘0’ is almost 50% on an average. Since images have high amount of correlation between between adjacent pixels, adjacent data sets can be grouped and grouped elements can be analyzed. In this thesis, data is inverted and stored, so that this work is compatible other systems such as commercial processors. This thesis considers data sets containing zeros in the same position between adjacent pixels. Figure 5 shows groups having 2, 4, 8 and 16 data bytes and shows the proportion of groups that will have only zeros scanning bit positions one by one for all bytes in the group, called a *compress-group*. For instance, if a group is constructed by using two adjacent pixels, where each pixel is a byte long and has zeros exactly at MSB

![Proportion of Zeros](image2)

**Figure 4:** The proportion of SRAM cells that store a value of ‘0’ to total number of SRAM cells
and LSB positions. Other positions all have ones. The proportion of compress-groups to total groups for the image will be 0.25, as 2 (LSB and MSB only) in 8 bits are a zero on an average. It is interesting to see that this number will remain the same even if the group is constructed by using 4, 8 or 16 adjacent pixels, if all pixel elements have MSB and LSB positions have a zero.

Figure 5: The size of the group is varied and the proportion of compress-groups is compared to total groups. The proportion of compress group is found to exponentially decrease with increasing group size.

However, Figure 5 shows that as the size of group increases the proportion of compress-groups decreases. Since the benefits saturate around a size of 8, in this thesis, compress group with a size of 8 is analyzed.

**Thesis Statement:** Use data patterns to compress the energy in SRAM (*En-Com System*) by selectively turning OFF groups of cells SRAM to save leakage power.

The *En-Com System* proposed should be used at low nanometer designs (below 22nm) as the leakage power becomes dominant at these nodes and DRV technique with SECDED does not help due to large DRV values and variance. In this thesis, I evaluate the design at 130nm IBM process.
CHAPTER III

SYSTEM DESIGN: CONSIDERATIONS AND IMPLEMENTATION

There is high amount of spatial and temporal locality in video/image data. The data in these structures can be compressed by grouping together adjacent pixels. At the system level, these pixels may be stored in adjacent columns in an SRAM. Every row in SRAM can represent a line of pixels. There can be one bit assigned for every 8 bits of image pixels. A frequent pattern like 00000000 can be compressed using this one bit to represent the data for the entire 8 bits. These 8 bits can then be turned off saving leakage power. These can be done in 2 ways; the frequent pattern can be identified on a row basis or column basis. Each of these techniques have trade-offs in layout and implementation.

Figure 6: Row Based Compression increases the per cell height by four ‘Layer 3’ metal lanes per cell. Column Based Compression increases the per cell width by two Layer 2 metal layers only.
The *En-Com* system employs a novel column based data based compression instead of row based techniques that have already been proposed. The row based techniques involve the use of large number of metal wires that require to be connected to each SRAM cell. This involves an area overhead in the design of the SRAM cells as shown in the Figure 6. The column based technique involves using parallel column based metal lines. The row based compression technique increases the cell height by four ‘Layer 3’ metal lanes. Compared to this, the column based technique will only increase the cell width by two ‘Layer 2’ metal lanes, saving layout area, an important design point in this thesis.

The *En-Com* system stores the compresses value on an additional cell, called the Zero-Switch Cell. The compression pattern for this system is 00000000. When the Zero-Switch Cell stores the compressed value (data value 0), it switches OFF the compress-group. The compress-group consists of 7T SRAM cells that hold their values (data value 0), even though they are switched OFF. These switched off cells can be read without turning them ON and does not require any modifications in the read circuitry. This allows *En-Com* to be used as a perfect framework for video/image processing applications that usually compare master frames/images with subsequent ones. This comparision involves reads from the SRAM. By selectively switching OFF, *En-Com* can save leakage power in these applications. On any other pattern, the Zero-Switch Cell should not switch off the compress-group. To implement this, the Zero-Switch Cells are initialized to ‘0’ at start. This ensures that almost all of the SRAM is switched OFF in the beginning. On a write of a ‘1’, it is written to the Zero-Switch Cell along with the original cell.

<table>
<thead>
<tr>
<th>Table 1: Compress-Group Area Overhead</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compress-Group Size</td>
</tr>
<tr>
<td>---------------------</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>16</td>
</tr>
<tr>
<td>32</td>
</tr>
</tbody>
</table>
En-Com uses one cell per compress-group, the area overhead will increase as the group size reduces. The additional overhead is because every compress-group will require an additional SRAM cell to store its information. Table 1 gives the area overhead as the group size is varied.

The En-Com SRAM System requires minor modifications in the decoders, SRAM column array and write circuitry. En-Com requires an architectural change that includes having a double write signal in case of a writing binary value ‘1’ into the array. A write of binary value ‘1’ requires one extra cycle. Since the decoder tree of a large SRAM memory system has a delay of many clock cycles, an additional cycle is a reasonable trade off for the energy savings. Figure 7 shows the schematic of En-Com system.

Figure 7: ‘En-Com’ implementation (Schematic). The top and bottom pins in the decoder and the SRAM array represent the lines to the Zero-Switch Cells

The SRAM based memory system is designed at the 130nm node using the IBM process. A 16x32 SRAM subarray consists of dynamic NAND based decoders for operating the array at 500MHz. Table 2 shows the specifications of the system.
### Table 2: Parameters of the Energy Compressed System

<table>
<thead>
<tr>
<th>Feature</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supply Voltage</td>
<td>1.2V</td>
</tr>
<tr>
<td>Row Decoder</td>
<td>4:16 working at 500MHz (Dynamic)</td>
</tr>
<tr>
<td>Column Decoder</td>
<td>2:4 working at 500 MHz (Dynamic)</td>
</tr>
<tr>
<td>Sense Amplifier</td>
<td>voltage based</td>
</tr>
<tr>
<td>Data Granularity</td>
<td>1 Byte access/cycle</td>
</tr>
</tbody>
</table>

#### 3.1 Design Considerations and Working

The *En-Com* System relies on energy compression using a frequently appearing data pattern. We choose a data pattern of **0000000** (eight zeros) to enable the compression. This number is selected based on Figure 5 and table 1. A group of cells storing this *pattern of eight zeros* form the compress-group for the design. Each cell in the *compress-group* is a 7 transistor SRAM (7T SRAM). The design also requires a pivot cell to turn off the compress-group and store its value called the *Zero-Switch Cell*. Even though images contain more compress groups having only ‘1’s, the *En-Com* system stores inverted values of data for images. This is to make the design consistent even for commercial processors, that have higher number of ‘0’s over ‘1’s in the caches. Commercial processors can ignore the inversion of data. Other major design considerations are:

- The *Zero-Switch Cell* will require a small driver for operating on the compress-group and a power-gate transistor

- The power gate transistor sizing determines the turn on/off time of 8 cells

- The compress group must hold the data-value (0 for each cell) and requires a 7T SRAM structure. This ensures that reads to the compress group will work without any modifications to the read circuitry at the SRAM.

- At the system level, the implementation of write logic and the read logic for the SRAM Array requires a few changes. Reads require that the pulse generator is not
activated by the system and writes require it to be active in case of writing data value ‘1’, the pulse generator must be active for an additional cycle to enable dual write.

The 16 rows in the 16x32 array are divided into 2 sections (vertically) of 8 rows each. Every column in a section has a Zero-Switch Cell. All Zero-Switch Cells are reset to a ‘0’ at the beginning. Every write of a 1 to the SRAM section is also written to its Zero-Switch Cell. Only when all 8 cells in the column have ‘0’s i.e the compress-group has the data pattern 00000000, the Zero-Switch Cell switches off the compress-group.

3.2 7T SRAM Cell Design

The compress-group consists of cells with 7 transistor SRAM. This enhancement over traditional 6T SRAM is to store the compressed data value, instead of putting the SRAM in a high impedance state. The 7T SRAM cell holds logic zero ‘strongly’ when the compress-group gets activated. The activation of the highlighted transistor (Figure 8) ensures that the cell can be switched off without SRAM going into the high impedance state. Without this transistor if the cell is switched off, the sense amplifier may produce an erroneous output on a read. Due to positive feedback, there is unpredictability of the final state once the cell is turned on. By storing a strong ‘0’, 7T SRAM ensures that these problems do not occur. Figure 8 shows the 7T SRAM cell in its logical view.
### 3.3 Compress Group

A Zero-Switch Cell is connected on each column for a group of 8 cells called the *compress group*. Writing a ‘0’ to the Zero-Switch Cell power gates the compress group. All cells in the compress group will be held to data value ‘0’ even after power gating. The implementation of the Zero-Switch Cell and the compress group shown in Figure 9.

![Figure 9: An 8 Cell Compress Group](image)

### 3.4 Zero-Switch Cell

The 6T SRAM cell along with an inverter driver forms the Zero-Switch Cell. Its job is to store the information for the compress group. If the compress group contains data that is all ‘0’, the Zero-Switch Cell is activated. A ‘0’ in the Zero-Switch Cell implies a pattern of all ‘0’s in the compress group. For any other pattern, the Zero-Switch Cell stores a ‘1’.
3.5 Dual Write Pulse Generator

The pulse generator, enabled when dual write must take place, produces a clocked signal that selects the **Zero-Switch Cell** from the section called the Dual Write Pulse Generator. After a write into the original cell, if the data value is ‘1’, the Dual Write Pulse will also write into the **Zero-Switch Cell**.

3.6 Modification in Row Decoder

The row decoder is modified to accommodate additional 1 cell per group. This requires a select signals that are derived by taking the Most Significant Bit of the 4:16 row decoder. The cells are selected on the assertion of the Dual Write Pulse Generator. Figure 10 shows the implementation of the 2 select signals.

![Diagram of 4:16 decoder and select signals](image)

**Figure 10:** *En-Com* requires one NOT gate and two AND gates to select between the top or bottom **Zero-Switch Cell** in case of a write
3.7 **Modification in the Write Circuitry**

The write circuitry is modified to perform 2 writes in case of writing a ‘1’. Since writing is done by isolating the sense amplifier circuitry, a dual write will involve maintaining this isolation for 2 cycles. The logic to implement dual writing is shown in Figure 11. The 3 input AND gate for writing into the ‘BIT’ position ensures that dual write is activated only when Dual Write Pulse is generated.

![Figure 11](image)

Figure 11: The write circuit is modified to support dual write in case of data value ‘1’. The pulse generator enables the dual write
CHAPTER IV

RESULTS AND ANALYSIS

The En-Com system saves leakage power by selectively switching off SRAM cells while retaining data. This section discusses the power saving in En-Com system. The baseline SRAM system uses 6T SRAM cells with ECC protection.

4.1 Data Read

![Waveforms showing data read in En-Com](image)

Figure 12: Waveforms that show that data being read from a row in En-Com

The read operation is similar to the baseline SRAM system. Figure 12 depicts the reading operation. For simplicity only Row 0 is selected and 4 columns are selected repeatedly.
It is seen that the introduction of additional logic like the Dual Pulse Generator and Write Circuitry does not influence the read operation. Since data is read with fidelity, the 7T SRAM cell is found to be stable for a read operation. Even if the cells are switched off, all cells are read correctly.

4.2 Data Write

![Waveforms showing the read operation of a row in En-Com](image)

Figure 13: Waveforms that show the read operation of a row in En-Com

Data is written similar to the baseline in the *En-Com* system. Figure 13 depicts the writing operation. For simplicity only Row 0 is selected and 4 columns are selected repeatedly. Writing requires the Write Enable signal to be synchronized with Precharge. Asymmetry in the write logic due to pulse generator does not impact the write operation.
4.3 Compress Group: Turning OFF

Figure 14 shows the operation of the compress-group when it is turned OFF. The Zero-Switch Cell is made to store a 0 thereby power gating the compress-group. The additional transistor in the 7T structure stores 0 strongly. The tristated node, shows a sudden small increase in the voltage due to increased capacitance.

![Waveforms Diagram](image)

Figure 14: The waveforms showing the turning OFF of the compress-group. At 3ns, the 7T SRAM cell is tristated due to power gating. The 7T cell stores 0 strongly.

4.4 Compress Group: Functionality Analysis

The compress group should continue to retain the value of 00000000 even after the Zero-Switch Cell switches to 1. This can happen when a ‘1’ needs to be written into one of the cells in the compress group. This write of ‘1’ will involve writing to the Zero-Switch Cell 1st. On a write of a ‘1’, the compress group will switch on with the original pattern.
Figure 15 shows the voltage at node storing ‘1’ decreases due to leakage within the SRAM. On switching ON the compress group, the 7T cell will continue retaining the same data value.

4.5 Power Analysis

*En-Com* system draws benefits from its power benefits. Until now the thesis presented the functional analysis. To truly understand the potential of *En-Com*, a power analysis is done at the cell, group and system level.

4.5.1 Power Consumption of SRAM cells

Leakage power consumption of three categories of SRAM cells are analyzed. The first one is the baseline 6T SRAM cell. The maximum leakage power for this cell is found to be 743pW at 1.2V. The second one is the 7T SRAM cell, this cell will have leakage power that is slightly greater than the 6T SRAM cell due to the presence of an additional NFET. When
the cell is storing a ‘1’, two NFETs (7th Transistor and turned OFF NFET from the cross coupled inverter) start leaking towards ground path. To mitigate this, the other terminal of the NFET is connected to Zero-Switch Cell, this reduces the VDD pressure across this NFET and mitigates the subthreshold leakage. The maximum power consumption of the 7T SRAM is 759pW at 1.2V. The third cell is the Zero-Switch cell that consumes additional power due to the inverter. However, if designed to be able to store a ‘0’ easily and overall changing the device dimensions for all elements for inverter and the cell, this cell can have a lower power consumption. This cell consumes 850pW at 1.2V. The power consumption analysis of these three cells with varying voltage is shown in the Figure 16.

![Figure 16: Comparison of the Power Consumption with varying voltage for a) Zero-Switch Cell b) 7T SRAM Cell c) 6T SRAM (Baseline)](image)

4.5.2 Power Consumption of Compress-Group

The power consumption of the compress-group depends on the number of cells in the group and the pattern of data. The power consumption of the compress group for a case when the group is ON and OFF with varying operating voltages is shown in the Figure 17. During the ON state, the maximum power consumption is 6.5nW at 1.2V. This is reduced by 6.5
times when the *compress group* is turned OFF. The maximum power consumption in the OFF state is nearly 1nW at 1.2V. This saving in power is because the 7T SRAM cells are power gated.

### 4.5.3 En-Com System: Power Analysis

The En-Com power consumption is analyzed with a baseline system without any energy saving features. The average power consumption is shown in table 3.

<table>
<thead>
<tr>
<th>Implementation</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>7.2mW</td>
</tr>
<tr>
<td>En-Com</td>
<td>6.3mW</td>
</tr>
</tbody>
</table>

The power saving of system for image benchmarks is shown in the Figure 18. It can seen that on an average 11% power can be saved using *En-Com* over the baseline. Some images can save upto 48% of the power when stored in the *En-Com* system. These values
are found to be correlated to the number of \textit{compress-groups} that can be turned OFF.

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure18.png}
\caption{Normalized power saved due to \textit{En-Com} when compared with baseline.}
\end{figure}

\section{Dynamic Power Consumption}

There is increase in the dynamic power of the \textit{En-Com} system due to dual writes. Since dynamic power in a large SRAM system only contributes to a small percentage of the total power, the effect of dual writes on total power consumption is low. Figure 19 shows the variation of dynamic power with activity. \textit{En-Com} read power is slightly lower than the baseline system as most cells are switched off while reading and the internal node voltage swing of these cells does now does not consume power.

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure19.png}
\caption{Dynamic power comparison of \textit{En-Com} with baseline system. The write power in \textit{En-Com} is two times higher than the baseline in the worst case. The read power of \textit{En-Com} remains almost same}
\end{figure}
4.7 Enhancement to Improve Power Savings

The amount of power saved depends on the way data is organized. An organization that turns OFF more compress-groups will lead to greater power savings. The En-Com system that is discussed until now uses runlength information, taking pixels within the same row (x direction) while forming the compress-group. This type of En-Com, called Run-Length En-Com, does not fully utilize the spatial locality that is exhibited by images and is a generic method. For images, this thesis proposes an enhancement called the Clustered En-Com.

Run-Length and Clustered En-Com system implement the compress groups differently is shown in the Figure 20. Clustered En-Com achieves greater spatial locality by working on data pixels that are adjacent in x and y directions. While using Clustered En-Com, data must be read from the image sensor in a clustered format and stored in the SRAM. The Figure 21 shows the normalized power saving due to Clustered En-Com. Clustered En-Com achieves 18% power saving when compared with the baseline. A few benchmarks that had greater locality in the x direction when using Run Length En-Com is seen to be having reduced locality while using Clustered En-Com. Most image benchmarks benefit
from *Clustered En-Com* with a maximum saving nearly 41% and 15% on an average.

![Figure 21: Normalized power saving due to *Clustered En-Com* when compared with baseline.](image)

### 4.8 Layout of the 7T SRAM Cell

For lower nanometer nodes, a split wordline layout is preferred[24]. This type of layout allows the *poly* to be laid out straight and the cell can be made with higher density. This is a wide cell type of a layout which increases the distance between the BIT and the BIT-BAR metals. This reduces the coupling noise between BIT and BIT-BAR. The 7T SRAM cell is laid out using split wordline technique. Figure 22 shows the 6T SRAM layout using split word line.

![Figure 22: The baseline split wordline 6T layout](image)

To accommodate the additional line to turn OFF the 7T SRAM, we require a metal that runs parallel to the BIT line in the 7T layout. We also require another parallel metal that connects the source of the additional metal to the *Zero-Switch* Cell. Figure 23 shows the
7T SRAM layout using split word line. An additional 15% area overhead is incurred due to the extra metal lanes.

Figure 23: The 7T layout occupies 15% more area when compared to the 6T layout.
CHAPTER V

SUMMARY

SRAM systems have leakage which is dominated by cell leakage. When these systems are used to store data they tend to have high power dissipation to leakage currents. They are usually used to store images or video frames in ASICs or to store application data in commercial processors. ASICs have processing engines that operate on this data. As the complexity of these processing engines have increased, they operate on larger data to improve performance. To supply this large amount of data, the size of the SRAM have increased in these ASICs along with increasing its leakage power. As these SRAM systems store images/video frames for long periods of time, leakage power mitigation is an important design point.

Some leakage mitigation techniques such as body biasing and power gating are found to be ineffective for SRAM based systems. It is observed that data patterns in these images or data tend to follow high amount of spatial and temporal locality. By compressing data based on these patterns we can reduce leakage power in SRAM. This thesis suggests a compression based approach for saving power.

The thesis implements a pattern based Energy Compression System (En-Com) that switches OFF cells to save leakage power. Contrary to row based compression scheme, En-Com is implemented as a column based scheme. En-Com allows reads to the SRAM to take place without any modifications and does not require turning ON the switched OFF cells. Traditional 6T SRAM is modified to a 7T structure to allow such reads to happen. En-Com requires an additional Zero-Switch Cell that is activated on compression and switches of a group of cells.

En-Com only requires minor modifications in the read and write circuitry and shows
a performance improvement of up to 15% when applied to images. The area overhead in laying out the 7T SRAM cell is nearly 15%. By using modifications in En-Com, such as clustered or runlength, we can try to extract greater locality in the images thereby getting higher savings in power.

This thesis presents En-Com technique for images, however it should be noted that this technique will work well for any system that stores data for long idle periods and have much larger reads than writes. Applications such as video frame buffer processing and digital speech sampling also exhibit data properties of high locality.
REFERENCES


