LOW ENERGY COMPUTER ARCHITECTURE DESIGNS

By

Mervat Mohamed Adel Mahmoud

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
DOCTOR OF PHILOSOPHY
in
Electronics and Communication Engineering

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2019

LOW ENERGY COMPUTER ARCHITECTURE DESIGNS

By

Mervat Mohamed Adel Mahmoud

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
DOCTOR OF PHILOSOPHY
in
Electronics and Communication Engineering

Under the Supervision of

Prof. Dr. Hossam A. H. Fahmy

Electronics and Communications Department
Faculty of Engineering, Cairo University,
Cairo, Egypt

Dr. Dalia A. El-Dib

Electrical and Computer Engineering Department
Faculty of Engineering, Dalhousie University, Halifax, NS, Canada

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2019

Engineer's Name: Mervat Mohamed Adel Mahmoud
Date of Birth: 12/8/1982
Nationality: Egyptian
E-mail: mervat-m@eri.sci.eg
Phone: 01001489626
Address: El-Harram, Giza

Registration Date: 1/10/2012
Awarding Date: 2019
Degree: Doctor of Philosophy
Department: Electronics and Communication Engineering

Supervisors:

Prof. Dr. Hossam Ali Hassan Fahmy
Dr. Dalia Abdel-Wahed Fouad El-Dib

Examiners:

Prof. Dr. Hossam Ali Hassan Fahmy (Thesis main advisor)
Dr. Dalia Abdel-Wahed Fouad El-Dib (Advisor)
Electrical and Computer Engineering Department, Faculty of Engineering, Dalhousie University, Halifax, NS, Canada
Prof. Dr. Amr Galal El-Din Ahmed Wassal (Internal examiner)
Dr. Magdy El-Moursy Ali (External examiner)
Associate Professor, Microelectronics Department, Electronics Research Institute. And Staff Engineer and Engineering Manager, Mentor, A Siemens Business

Title of Thesis:

Low Energy Computer Architecture Designs

Key Words:

low energy; pipeline processing; parallel architectures; lossless compression; main memory

Summary:

The continuous increase in chip integration and the associated energy consumption concerns made low power/energy design one of the main challenges facing VLSI systems. A low energy clock-gated pipelined dual base binary/decimal fixed-point multiplier is suggested extending a previously proposed non-pipelined design. A thorough study conducted on both the pipelined and non-pipelined designs versus other architectures in literature proves tremendous reductions in power, energy and area consumption. In addition, a new low energy lossless compression/decompression approach is suggested for main memory data. The proposed design lowers energy consumption due to its simplicity and low latency.

Disclaimer

I hereby declare that this thesis is my own original work and that no part of it has been submitted for a degree qualification at any other university or institute.

I further declare that I have appropriately acknowledged all sources used and have cited them in the references section.

Name: Date:

Signature:

Acknowledgments

بسم الله الرحمن الرحيم

"سُبْحَانَكَ لَا عِلْمَ لَنَا إِلَّا مَا عَلَّمْتَنَا إِنَّكَ أَنتَ الْعَلِيمُ الْحَكِيمُ"

صدق الله العظيم

Foremost, all thanks are due to Almighty, the merciful God. God blessed me and gave me the strength to finish this study.

Dr. Hossam Fahmy and Dr. Dalia El-Dib, my thesis supervisors, have guided my work insightfully, and supported me during the thesis journey. Dr. Hossam (مُعَلِمي) and I would like to write it in Arabic as he represents all the real meanings of this word and the verse of Ahmed Shawky's poem:

قُمْ للمُعلّمِ وَفِّهِ التبجيلا ... كادَ المُعلّمُ أن يكونَ رسولا

أعلِمتَ أشرفَ أو أجلَّ من الذي ... يَبني ويُنشئُ أنفُساً وعقولا

Dr. Dalia has enriched my research with her experience and has always supported, guided and encouraged me to achieve higher in my entire career path - not only in my research.

My parents have supported me with their love and prayers. I would never succeed in life without their early support. Owing them my success, I will never be able to thank them enough. To them, I would like to express my sincere thanks.

My colleagues, Safaa Ahmed Elawamy, Dina Ellithy, and Rasha Mahmoud have provided me with cooperation and help with the design simulations and CAD tools. Also, I would like to thank my sister Marwa. And finally, I would like to thank my dear friends Asmaa Adawy, Rasha Shoitan, Abeer Farouk, Nahla El-azab, Ebtsam Arafa, Heba Draz, Ratshih, Nahla El-Sayed, and Hoda Hosny, who endured this long journey with me, always offering unconditional support and love.

Dedication

I would like to dedicate my thesis to my beloved mother.

Table of Contents

ACKNOWLEDGMENTS ..... **I**

DEDICATION ..... **II**

TABLE OF CONTENTS ..... **III**

LIST OF TABLES ..... **V**

LIST OF FIGURES ..... **VI**

ABSTRACT ..... **VII**

CHAPTER 1 INTRODUCTION **1**

1.1. Power vs Energy 1

1.2. Power/Energy Measurement in Digital Design 2

1.3. Low Power/Energy Digital Design 4

1.4. Approximate Computing for Saving Energy 5

1.5. Thesis Overview 7

CHAPTER 2 : LOW ENERGY PIPELINED DUAL BASE (DECIMAL/BINARY), DBM, MULTIPLIER **9**

2.1. Introduction 9

2.2. Available Combined Binary/Decimal Multipliers 10

2.3. Low Energy Pipelined Dual Base (decimal/binary) Multiplier, DBM, Design 13

2.3.1. Multiplicand Multiples Generation Stage 13

2.3.2. Partial Products Selection Stage 16

2.3.3. Partial Products Accumulation Stage 18

2.3.3.1. Binary Column Tree *19*

2.3.3.2. Split Binary Addition *21*

2.3.3.3. Split Decimal Addition *21*

2.4. Summary 22

CHAPTER 3 : DBM MULTIPLIER COMPARISON AND RESULTS **23**

3.1. Introduction 23

3.2. FPGA Simulation Results 23

3.3. ASIC Implementation Simulation Results For The non-pipelined Designs 25

3.4. Binary/Decimal Pipeline Stages Selection Using NanGate-45 nm Technology 26

3.5. Pipelined DBM Design Simulation Results For Different Technologies 28

3.6. Power Distribution Analysis 30

3.7. Summary 32
CHAPTER 4 : LOW ENERGY DESIGN FOR MAIN MEMORY (LITERATURE IN DATA COMPRESSION) 33
4.1. Introduction 33
4.1.1. Memory Hierarchy 33
4.1.2. Lossless Compression 35
4.2. Literature Review on Lossless Compression Algorithms 36
4.2.1. Definitions 36
4.2.2. Basic Coding Techniques 38
4.2.2.1. Unary Coding 38
4.2.2.2. Binary Coding 38
4.2.2.3. Codes with Selector Part 38
4.2.2.4. Run Length Encoding (RLE) [1967] 39
4.2.3. Statistical Coding Techniques 40
4.2.3.1. Shannon-Fano Coding [1948] 40
4.2.3.2. Huffman Coding [1952] 40
4.2.3.3. Arithmetic Coding 42
4.2.4. Dictionary Based Coding Techniques 43
4.2.4.1. LZ77 and LZ78 [Lempel and Ziv, 1977 and 1978] 43
4.2.5. Differential Coding 44
4.2.6. Conclusion 44
CHAPTER 5 : SUGGESTED APPROACH FOR LOW ENERGY ASIC DESIGN FOR MAIN MEMORY DATA COMPRESSION/DECOMPRESSION 47
5.1. Recent Literature in Memory Compression 47
5.2. Suggested Methodology for Low Energy Main Memory Compression 50
5.3. Comparison and Results 53
5.4. Preparing design for fabrication 56
5.5. Summary 62
CONCLUSIONS AND FUTURE WORK 63
REFERENCES 65
APPENDIX A: RTL SYNTHESIS FLOW COMMANDS USING SYNOPSYS DESIGN COMPILER (DC) 71

List of Tables

Table 2.1 Decimal Multiplicand multiples selection.....17
Table 3.1 Worst path delay, area and power consumption of the proposed design and the previously published designs.....24
Table 3.2 Detailed power dissipations distribution (in Watts) for the proposed pipelined design and the previously published designs.....24
Table 3.3 Area, total power dissipation, and PDP for non-pipelined designs at 100MHz. 25
Table 3.4 Minimum path delay for non-pipelined designs at maximum operating frequency.....26
Table 3.5 Delay breakdown of each stage in the non-pipelined DBM design for NanGate-45 nm technology at 100MHz.....27
Table 3.6 Comparison among different pipelining schemes for NanGate 45nm technology at 100MHz.....28
Table 3.7 Area, power, and PDP for pipelined DBM design at 100MHz for different technologies.....29
Table 3.8 Detailed power dissipation for pipelined DBM design for NanGate 45nm at 100 MHz.....29
Table 3.9 Power distribution of the pipelined DBM at 100 MHz for different technologies.....31
Table 4.1 Memory technologies' at 2012.....33
Table 4.2 Three simple codes and their expected length.....37
Table 4.3 minimal binary, Elias, Golomb, and Rice codes (i.e. the blanks in the codewords do not appear in the coded bit stream).....39
Table 4.4 Huffman coding example.....41
Table 4.5 Example of arithmetic coding for message of 10 symbols (a) statistical model of the data, (b) coding of the input message.....43
Table 4.6 Comparison between different basic coding techniques.....45
Table 5.1 Similarity between memory lines (reference: 1st line of each 4K page).....55
Table 5.2 Compression performance for different designs.....55
Table 5.3 Compression ratios for different benchmark applications for the suggested approach.....55
Table 5.4 Compression ratios for the different designs.....56

List of Figures

Figure 1.1 Energy consumption versus power consumption [7] 2
Figure 1.2 Switching power 3
Figure 1.3 Short circuit power 3
Figure 1.4 Example of how glitches occur 4
Figure 1.5 Leakage Power 4
Figure 1.6 Approximate computing in architecture level of abstraction [29] 7
Figure 2.1 non-pipelined Dual Base (decimal/binary) Multiplier, non-pipelined DBM. 12
Figure 2.2 Pipelining Schemes 14
Figure 2.3 Proposed pipelined Combined Binary/Decimal Multiplier. 15
Figure 2.4 Decimal multiples generation 16
Figure 2.5 Partial products selection 17
Figure 2.6 Binary column tree scheme. 19
Figure 2.7 “Column 15” binary CSA tree 20
Figure 2.8 The four binary bit-vectors after rearranging. 21
Figure 2.9 Scheme of the final Binary CPA. 21
Figure 2.10 Scheme of the final Decimal CPA. 22
Figure 2.11 The three decimal, BCD, bit-vectors after rearranging. 22
Figure 3.1 Total power breakdown in terms of the structure for 45nm technology at 100 MHz 30
Figure 4.1 Levels in memory hierarchy 35
Figure 5.1 Benini et al. basic compression block diagram [91] 48
Figure 5.2 General block diagram for “base-delta-immediate compression” Compressor Unit (CU) 49
Figure 5.3 “Base-delta-immediate compression” block diagram 49
Figure 5.4 Block diagram of 4KB memory page 51
Figure 5.5 Compression block diagram for a 32B memory line 51
Figure 5.6 Details of the 8-bits subtractor, the basic unit of the compression design 52
Figure 5.7 Decompression block diagram for 32B memory line 53
Figure 5.8 Compression design block diagram after editing for fabrication 58
Figure 5.9 FPGA post translate simulation results 59
Figure 5.10 ASIC post synthesis simulation results 60
Figure 5.11 ASIC post synthesis simulation results for four memory lines 61

Abstract

The continuous increase in chip integration and the associated power consumption concerns with moving towards portability made low energy design one of the main challenges facing VLSI systems. As low energy design has low power with high operating frequency. Power/Energy management has various strategies at all design process levels. That includes energy optimization at technology, circuit, logic, architecture and system levels of abstraction. In this thesis, we present two low energy designs. A low energy design at circuit level of abstraction is proposed for combined binary/decimal multipliers. And a low energy design at architecture level of abstraction is proposed for main memory data compression.

As combined binary/decimal arithmetic is optimal in supporting binary and decimal high speed and low power applications. A low energy clock-gated pipelined dual base binary/decimal fixed-point multiplier is suggested extending a previously proposed non-pipelined design. A thorough study conducted on both the pipelined and non-pipelined designs versus other architectures in literature proves tremendous reductions in energy consumption. The pipeline stages are chosen to achieve energy reductions with acceptable latency. In addition, clock gating the pipelined multiplier design is introduced to provide a total of 43% energy reduction for the pipelined design if compared to the lowest energy design in the literature.

In addition, a new low energy lossless compression/decompression approach is suggested for the data of main memory. The proposed approach depends on the delta coding and the observation that, for many applications, the lines of the main memory pages are mostly similar. The target is to achieve a simple low energy compression design for exact storage of memory data. The proposed design lowers energy consumption by up to 66% when compared to previous designs. This is due to its simplicity and low latency. Furthermore, the frequency of operation is increased from 300 MHz to 800 MHz. The new design also allows the main memory to store up to 30% more data according to PARSEC and PERFECT benchmarks applications data.

Chapter 1 Introduction

In predicting the future of integrated electronics, Gordon Moore predicted in 1965 that the number of components per chip will double every year in the period till 1975 [1] reaching 65,000 components on a single quarter-inch semiconductor. In 1975, Moore reduced the rate to a doubling every two years due to integrating more microprocessors which are in general less dense in electronic circuits [2]. In 1995, Moore's stated that his projection is not going to stop soon [3]. In fact, Moore's rule was considered one of the driving forces of electronics industry. It challenged technologists to deliver annual breakthrough in manufacturing Integrated Circuits (ICs) to comply with Moore's law. In 2014, a die was able to hold over seven billion transistors. Moore's law worked perfectly and was continuously fulfilled and has caused many of the most important changes in the electronics manufacturing technology.

Since 1970s, The most dominant electronics manufacturing technologies used were bipolar and nMOS transistors [4]. Nevertheless, these consume non-negligible power even in static (non-switching) state. Consequently, by 1980s, the power consumption of bipolar designs and its cooling solution costs were considered too high to be sustainable. This caused an expected switch to a slower, but lower-power Complementary Metal Oxide Semiconductor (CMOS) technology. At that time, CMOS transistors consumed lower power largely because static (leakage) power was negligible if compared to dynamic (switching) power. Along with fulfilling Moore's law, the aim is always to increase processing power of electronic circuits. This is achieved by scaling down the technology, increasing the number of components per chip, and increasing the frequency of operation. In the late 2004 with scaling down the CMOS fabrication technology to 45-nm and downwards, we encountered a high increase in leakage power to the extent that it is comparable to dynamic power and can even dominate the overall power dissipation. Also, with integrating more and more components, the power increases dramatically, and causes a challenge regarding excessive thermal dissipation. Thus, another paradigm shift in computing electronics was inevitable. The shift to multi-core computing was in the aim to increase performance while keeping the hardware simple, retain acceptable power consumption and transfer complexity to higher levels of the system design abstraction, including software level. [4]. Approximate computing (AC) is one of the energy efficient computing paradigms that use the inaccuracy tolerance inherited in applications for significant performance improvements [5][6]. It leads to another tradeoff, energy and performance versus computing quality. Where, slightly losing computing quality can improve energy and/or density.

1.1. Power vs Energy

Both terms energy and power are ex-changeably used although energy is different from power. For example, a specific task needs a specified amount of energy E to complete over time T. Its power consumption P is the rate at which energy is consumed (E/T). The time needed to complete the task can be increased by reducing the frequency of operation for example. Whereas, the same amount of energy is still needed

to complete the task. Thus, the power consumption is reduced; however the energy consumption (area under the graph) is still the same as shown in Figure 1.1.

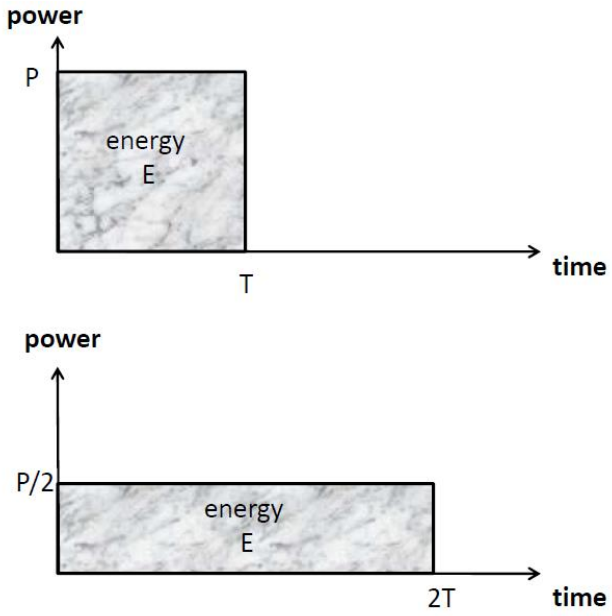

Figure 1.1 Energy versus power [7]

While energy measures the total quantity of work done, it doesn't say how fast the work done. As a loaded semi-trailer can be moved across the country with a lawnmower engine if time is not essential. If other things being equal, the tiny engine would do the same amount of energy and burn the same amount of fuel as the truck's big one. But the bigger engine has more power, so it can get the job done faster.

The question now, in computer architecture design, which is more important power or energy?. Desktop computers or wired digital devices are permanently connected to a power supply. Power supply feeds the design components with power ($P = V.I$), whereas these design's components consume this power. That makes energy efficiency here a bonus compared to a functional necessity. However, laptop computers or portable devices have limited battery life time ($E = P.T$), before get rid of the battery or recharge it.

1.2. Power/Energy Measurement in Digital Design

An IC's energy consumption is defined as its power consumption by the operating frequency ($E = P/f$) while power consumption is mainly composed of static power and dynamic power.

$$P_{total} = P_{dynamic} + P_{static} \quad (1.1)$$

Dynamic power consumption is frequency-dependent and results from one of the following three sources: Switching power, short circuit power and glitching power [11]. The dominant part of the dynamic power is the switching power which is consumed

during the charging and discharging of capacitive nodes, Figure 1.2. It can be represented with the following equation;

$$P_{dynamic} = \alpha . C_l . V_{dd}^2 . f \quad (1.2)$$

Where α is the switching activity of the circuit; C_l is the effective capacitance of the circuit; V_{dd} is the supply voltage; and f is the operating frequency.

Short circuit power occurs during the momentary current flow that occurs when two complementary transistors conduct during a logic transition, which arises from long rise or fall times of input signals, Figure 1.3. Moreover, glitching power occasionally arises due to the finite delay of the logic gates that cause spurious transitions at different nodes in the circuit, Figure 1.4.

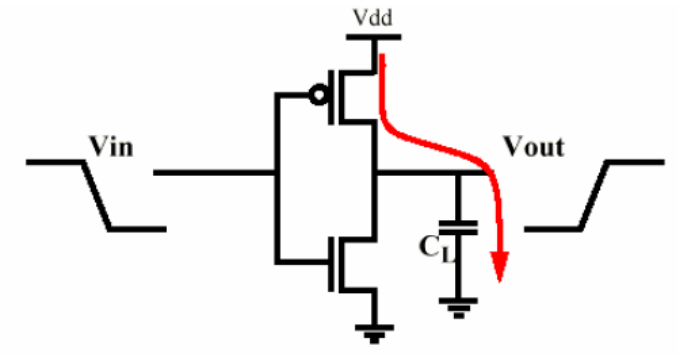

Figure 1.2 Switching power

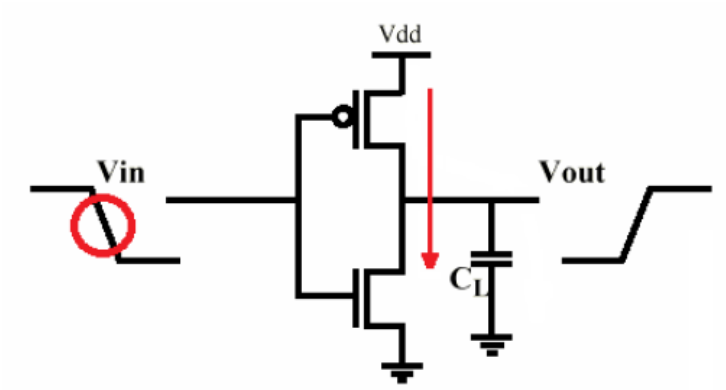

Figure 1.3 Short circuit power

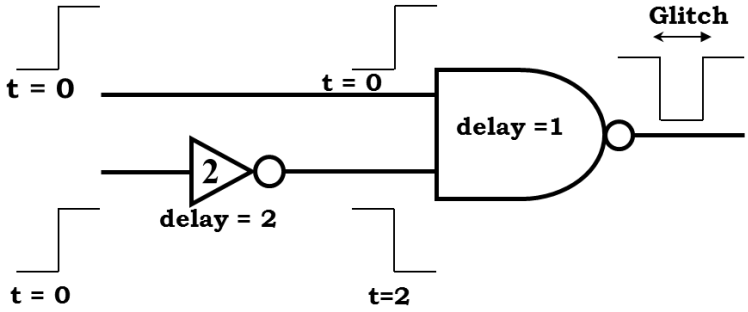

Figure 1.4 Example of how glitches occur

Static power typically comes from leakage current and dc current sources. Static power consumption has many components and has many paths, Figure 1.5. The most important contributor to static power in CMOS is the subthreshold leakage which is exponentially dependent on $(V_{gs} - V_T)$, where V_{gs} is the gate to source voltage and V_T is the threshold voltage. Another part of leakage is caused by reduced gate oxide thickness t_{ox} which increases gate oxide tunneling current. All parts of leakage current are increased excessively due to scaling down of technology which requires reducing V_T and t_{ox} to keep up with higher processing requirements.

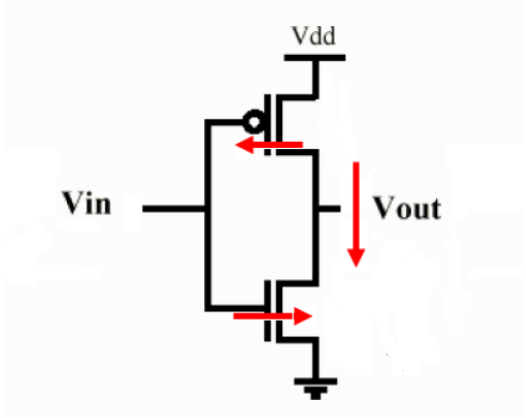

Figure 1.5 Leakage Power

1.3. Low Power/Energy Digital Design

Low power/energy design methodologies can be applied at all design levels of abstraction including system, algorithm, architecture, logic, circuit, device and technology levels. Low energy design can be defined as a design that has low power at high operating frequency. However, some of the low energy techniques reduce energy in exchange of reduced performance. Eventually, one has to reach a compromise between energy, power, performance, and cost to satisfy overall design requirements. Nevertheless, improvement at a higher level design abstraction will definitely affect all subsequent design abstraction levels to comply with the changes at that higher level. At