

April 2014

Heat Dissipation Bounds for Nanocomputing: Methodology and Applications

Ilke Ercan
University of Massachusetts - Amherst

Follow this and additional works at: https://scholarworks.umass.edu/dissertations_2



Part of the [Electrical and Computer Engineering Commons](#)

Recommended Citation

Ercan, Ilke, "Heat Dissipation Bounds for Nanocomputing: Methodology and Applications" (2014).
Doctoral Dissertations. 6.
<https://doi.org/10.7275/2sx1-wc86> https://scholarworks.umass.edu/dissertations_2/6

This Open Access Dissertation is brought to you for free and open access by the Dissertations and Theses at ScholarWorks@UMass Amherst. It has been accepted for inclusion in Doctoral Dissertations by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.

HEAT DISSIPATION BOUNDS FOR NANOCOMPUTING: METHODOLOGY AND APPLICATIONS

A Dissertation Presented

by

İLKE ERCAN

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

February 2014

Electrical and Computer Engineering

© Copyright by İlke Ercan 2014

All Rights Reserved

HEAT DISSIPATION BOUNDS FOR NANOCOMPUTING: METHODOLOGY AND APPLICATIONS

A Dissertation Presented

by

İLKE ERCAN

Approved as to style and content by:

Neal G. Anderson, Chair

Eric Polizzi, Member

Qiangfei Xia, Member

Jonathan Machta, Member

Christopher V. Hollot, Department Chair
Electrical and Computer Engineering

ACKNOWLEDGMENTS

I owe in large measure the stimulation of my thoughts to my adviser and mentor Professor Neal G. Anderson. His unique vision allowed me to find a common ground between engineering, physics and philosophy of science, and pursue this dissertation. I am forever indebted to him for his generous support, encouragement and guidance.

I am thankful to Professor Csaba Andras Moritz and his students Prithish Narayanan, and Pavan Panchapakeshan for helpful discussions regarding the NASIC paradigm and Mostafizur Rahman for his assistance with the HSPICE simulation studies. I am grateful to Donald Blair for thought provoking discussions and his invaluable technical support. And, I thank my dissertation committee, Professor Eric Polizzi, Qiangfei Xia, and Jonathan Machta, for their useful comments and criticism of my work.

I am deeply grateful to my parents, Jülide and Nusret, for their unconditional love and support; to my sister and best friend, İris, for the joy and meaning she adds to my life; to my loyal friend Saime for providing me with a compass every time I need one; to Ozzie, Ann, Zeytin, Ronan and my beloved friends in the Valley for helping me redefine what family is and sustain my sanity, determination and resilience through rough times; to Nancy and Carlton for making 36 Triangle a home for me; to Lynette for her patience and guidance on my personal growth; to Chris for showing me that friendship can reach across any distance; and to Ceren for reminding me what is essential in life and that “one sees [it] clearly only with the heart.” Thank you; this dissertation could not have been completed without your loving support.

This work came to life through funding provided in part by the National Science Foundation under grant no CCF-0916156. I am grateful for generous resources provided by the Department of Electrical and Computer Engineering at UMass, Amherst.

ABSTRACT

HEAT DISSIPATION BOUNDS FOR NANOCOMPUTING: METHODOLOGY AND APPLICATIONS

FEBRUARY 2014

İLKE ERCAN

B.Sc., MIDDLE EAST TECHNICAL UNIVERSITY

M.Sc., UNIVERSITY OF MASSACHUSETTS AMHERST

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Neal G. Anderson

Heat dissipation is a critical challenge facing the realization of emerging nanocomputing technologies. There are different components of this dissipation, and a part of it comes from the unavoidable cost of implementing logically irreversible operations. This stems from the fact that *information is physical* and manipulating it irreversibly requires energy. The unavoidable dissipative cost of losing information irreversibly fixes the fundamental limit on the minimum energy cost for computational strategies that utilize ubiquitous irreversible information processing.

A relation between the amount of irreversible information loss in a circuit and the associated energy dissipation was formulated by Landauer's Principle in a technology-independent form. In a computing circuit, in addition to the information-theoretic dissipation, other physical processes that take place in association with irreversible information loss may also have an unavoidable thermodynamic cost that originates

from the structure and operation of the circuit. In conventional CMOS circuits such unavoidable costs constitute only a minute fraction of the total power budget, however, in nanocircuits, it may be of critical significance due to the high density and operation speeds required. The lower bounds on energy, when obtained by considering the irreversible information cost as well as unavoidable costs associated with the operation of the underlying computing paradigm, may provide insight into the fundamental limitations of emerging technologies. This motivates us to study the problem of determining heat dissipation of computation in a way that reveals fundamental lower bounds on the energy cost for circuits realized in new computing paradigms.

In this work, we propose a physical-information-theoretic methodology that enables us to obtain such bounds for the minimum energy requirements of computation for concrete circuits realized within specific paradigms, and illustrate its application via prominent nanocomputing proposals. We begin by introducing the unavoidable heat dissipation problem and emphasize the significance of limitations it imposes on emerging technologies. We present the methodology developed to obtain the lower bounds on the unavoidable dissipation cost of computation for nanoelectronic circuits. We demonstrate our methodology via its application to various non-transistor-based (e.g. QCA) and transistor-based (e.g. NASIC) nanocomputing circuits. We also employ two CMOS circuits, in order to provide further insight into the application of our methodology by using this well-known conventional paradigm. We expand our methodology to modularize the dissipation analysis for QCA and NASIC paradigms, and discuss prospects for automation. We also revisit key concepts in thermodynamics of computation by focusing on the criticisms raised against the validity of Landauer’s Principle. We address these arguments and discuss their implications for our methodology. We conclude by elaborating possible directions towards which this work can be expanded.

TABLE OF CONTENTS

	Page
ACKNOWLEDGMENTS	iv
ABSTRACT	v
LIST OF TABLES	x
LIST OF FIGURES	xii
CHAPTER	
INTRODUCTION	1
1. TECHNICAL BACKGROUND	8
1.1 Reversibility, Irreversibility, and Dissipation in Computation	8
1.2 First and Second Laws of Thermodynamics	9
1.3 Information is Physical	12
1.4 Landauer's Principle	13
1.5 Thermodynamic Relations for Quantum Systems	15
1.6 Information Erasure in Quantum Systems	18
1.7 Landauer's Principle Specialized for L -machines	24
2. HEAT DISSIPATION FOR NANOCOMPUTING:	
METHODOLOGY	27
2.1 Introduction	27
2.2 Abstraction	28
2.2.1 Physical Decomposition	29
2.2.2 Process Abstraction	32
2.3 Analysis	33
2.3.1 Operational Decomposition	33
2.3.2 Cost Analysis	35

3. HEAT DISSIPATION BOUNDS FOR NANOCOMPUTING:	
APPLICATION	39
3.1 Non-transistor-based Applications	40
3.1.1 Quantum Cellular Automata (QCA) Half Adder with Landauer Clocking	41
3.1.1.1 Abstraction	43
3.1.1.2 Analysis	47
3.1.2 QCA Half Adder with Bennett Clocking	54
3.2 Transistor-Based Applications	57
3.2.1 Nano-Application Specific Integrated Circuits (NASICs)	58
3.2.1.1 Abstraction	61
3.2.1.2 Analysis	66
3.2.1.3 NASIC Half Adder	72
3.2.1.4 NASIC Full Adder	76
3.2.1.5 Comparison with Physical Circuit Simulations	80
3.2.2 A Dynamically clocked np-CMOS Half Adder	82
3.2.2.1 Abstraction	87
3.2.2.2 Analysis	93
3.2.3 An Application to Static CMOS Circuits	96
3.3 Discussion	100
4. TOWARDS AUTOMATION: MODULAR DISSIPATION	
ANALYSIS	104
4.1 Modular Dissipation Analysis of a QCA Half Adder	104
4.1.1 Design Rules	105
4.1.2 Decomposition Rules	107
4.1.3 Dissipation Analysis	109
4.1.4 Prospects for Automation	112
4.2 Modular Dissipation Analysis of NASICs	114
4.3 Discussion	115

5. FOUNDATIONAL EPILOGUE: REVISITING LANDAUER'S PRINCIPLE AND THERMODYNAMICS OF COMPUTATION	117
5.1 A Historical Review Towards Landauer's Principle	117
5.2 Szilard Engine and Landauer's Exorcism of Maxwell's Demon	122
5.3 On the Validity of Landauer's Principle	124
5.4 On the Validity of Our Methodology	127
6. CONCLUSION	130
 APPENDIX: GRANULARITY OF HEAT DISSIPATION ANALYSIS FOR THE QCA HALF ADDER	 135
 BIBLIOGRAPHY	 141

LIST OF TABLES

Table	Page
3.1 State transformations for the QCA 1-bit half adder operated under Landauer clocking.	46
3.2 State transformations for the NAND-NAND NASIC 1-bit adder.	65
3.3 The truth table of the NAND-NAND NASIC 1-bit half adder.	73
3.4 Dissipation bound for the NASIC 1-bit half adder: Particle supply and information loss components	74
3.5 The truth table of the NAND-NAND NASIC 1-bit full adder.	76
3.6 Dissipation bound for the NASIC 1-bit full adder: Particle supply and information loss components	78
3.7 The truth table of the np-CMOS 1-bit half adder.	87
3.8 State transformations for the np-CMOS 1-bit half adder	91
3.9 Dissipation bound for the np-CMOS 1-bit half adder: Particle supply and information loss components.	95
3.10 The truth table of the CMOS 2-bit sorter.	98
5.1 Chronological list of significant events in the evolution of thermodynamics of computation.	121
A.1 Transition probabilities for a 1-bit half adder.	135
A.2 Transition probabilities for the second data zone of the QCA 1-bit half adder operated under Landauer clocking.	137
A.3 Transition probabilities for the AND gate with inputs AM and output N_1	137

A.4	Transition probabilities for the OR gate with inputs N_1N_2 and output S	138
A.5	Transition probabilities for the AND gate with inputs BM and output C	138

LIST OF FIGURES

Figure		Page
2.1	Physical abstraction of an information processing artifact and its surroundings in a globally closed and isolated universe.	30
3.1	Polarizations and logic states for four-dot and six-dot quantum cells.	40
3.2	Layout, clocking, and gate-level representation of a Landauer-clocked QCA half adder circuit with no line crossings.	42
3.3	Physical abstraction of the QCA circuit and its surroundings.	44
3.4	Fundamental lower bound on the cumulative dissipative cost for one computational cycle of the Landauer-clocking of QCA 1-bit half adder circuit.	53
3.5	Layout, clocking, and gate-level representation for Bennett clocking of the QCA 1-bit half adder.	55
3.6	The cross section of a crossed-nanowire Field Effect Transistor (xnwFET) [11].....	58
3.7	Layout, clocking, and logic diagram of the single-FET-type NAND-NAND NASIC 1-bit half adder circuit.....	60
3.8	Physical abstraction of the NASIC 1-bit adder situated in its surrounding environment.	62
3.9	Fundamental lower bound on the cumulative dissipation cost for one computational cycle of the NASIC 1-bit half adder.	75
3.10	Layout, clocking, and logic diagram of the single-FET-type NAND-NAND NASIC 1-bit full adder circuit.	77
3.11	Fundamental lower bound on the cumulative dissipation cost for one computational cycle of the NASIC 1-bit full adder.	79

3.12	Input-averaged cumulative energy consumption for one computational cycle of the NASIC 1-bit full adder obtained from HSPICE simulations.	81
3.13	Cross section of an nMOS (left) and a pMOS (right) transistor [39].	83
3.14	Layout, logic and timing diagram of the dynamically clocked np-CMOS half adder.	86
3.15	Physical abstraction of the CMOS 1-bit half adder situated in its surrounding environment.	88
3.16	Gershenfeld’s combinational 2-bit sorter [17].	98
4.1	Three example sections of QCA wires with “90-degree” cell orientation.	106
4.2	A complex inverter with two-cell input and output legs and specified clock zones.	106
4.3	QCA majority gates, with and without one fixed input, and associated clocking.	106
4.4	A dissipation zone, including placement of boundaries, for circuits designed according to the design rules of presented here.	108
4.5	Dissipation zones identified by application of the circuit decomposition rules to the QCA half adder of this work.	111
4.6	Schematic representation of the general (top) and modular (bottom) dissipation analysis procedures discussed in this work.	113
5.1	Schematics of the Szilard’s engine [65].	123
A.1	The lower bound value for the energy dissipation obtained by using circuit, data-zone, data-subzone and cell levels of analyses for QCA 1-bit half adder operated under Landauer clocking.	140

INTRODUCTION

Advancements in nanotechnology – emerging technologies with functional features at the one-billionth of a meter scale – promise to overcome limitations of current electronic circuits. Complementary Metal-Oxide Semiconductor (CMOS) transistors have been the fundamental building block of modern computer processors since 1980s and the size of these transistors has been shrinking steadily ever since,¹ enabling us to progressively develop smaller circuits operating at increasingly faster speeds. However, there are numerous limiting factors to this trend, which threaten the improved performance of this technology in the near future. The challenges encountered as the chips scale down to nanometer size have significant real-world implications,² and need to be addressed with urgency in all levels.

Numerous promising proposals are being put forth to overcome the limitations of current electronic technologies. The realization of these emerging nanoelectronic technology proposals faces a broad range of challenges cutting across device, circuit, architecture, and fabrication levels. Energy dissipation is perhaps the most critical one among these obstacles. The realization of post-CMOS nanoelectronic circuits³ depends on tackling such challenges. Advances in technology may allow us to overcome limitations at all levels for a proposed successor computational strategy and

¹The trend in the shrinking device size translates to an increase in the number of transistors on integrated circuits. Approximately every two years since 1960s, this number has doubled as predicted by the Moore’s law [1]. In the early years of this millennium, however, the scaling has reached a limit where making “a transistor smaller no longer meant it would be faster or less power hungry.” [2]

²In 2012, the revenue of the semiconductor industry in the US was about \$300 billion [2, 3].

³Integrated electronic circuit technologies that can potentially replace current Complementary Metal Oxide Semiconductor (CMOS) circuits.

perhaps one day we can realize complex nanocomputing circuits with atomic-scale perfection.

Even if we achieve such perfection, however, the operation of these circuits will unavoidably dissipate some minimum amount of energy that is determined solely by the logical irreversibility of the computations they perform. This stems from the fact that *information is physical* – it is encoded in the physical states of a system – and manipulating it irreversibly requires an unavoidable energy dissipation. This unavoidable dissipative cost of implementing logically irreversible operations fixes the minimum energy cost required for computation. Fundamental physical limits on energy dissipation associated with information processing has not been the focus of much attention until recently due to the minute contribution they make to the total power budget of conventional circuits. In order to determine the viability of investments best suited for the progress of computing technologies, however, it is critical to develop techniques that address these fundamental considerations in nanocircuits.

Rolf Landauer recognized the unavoidable cost of irreversible information processing, and formulated a relation between the amount of information loss and the associated energy dissipation in Landauer’s Principle [4] over fifty years ago. In the most general sense, Landauer’s Principle (hereafter LP) states that any irreversible manipulation of information in a circuit increases the thermodynamic entropy of the circuit and/or its environment, which necessarily dissipates heat. In post-CMOS nanocomputing circuits, the device densities and operating speeds will be extreme, as they are designed to supersede microelectronic circuits. Therefore, this unavoidable dissipation cost associated with irreversible computation may become significant and impose practical limitations on the ultimate capabilities and resource requirements of these circuits. The total energy dissipated per irreversible binary decision, continued at the historical rate observed for silicon technology, will reach Landauer’s lower bound on the physical cost of logical irreversibility alone ($\sim k_B T$) around 2020

[5]. Therefore, it is of paramount importance to approach emerging nanoelectronic devices from a fundamental energy requirement point of view to properly assess the viability of paradigms that have the potential to replace current CMOS technologies.

In its most general form, LP purely reflects the cost of irreversible information processing that is lower bounded by nature regardless of the computing technology used to execute it. In other words, LP is independent⁴ of any device or circuit parameters that play a functional role in information processing according to the underlying features of a computing paradigm. However, there may be physical operations in a circuit that cannot be disassociated from information erasure: physical operations (such as charge transport) that necessarily take place in order to process information. Therefore, for a given technology base, the fundamental lower bound on heat dissipation may contain additional unavoidable costs that are associated with thermodynamic processes required to execute computation within that paradigm. We define this cost as the heat dissipated by the “working substance” (such as information bearing electrons) of the computing circuit that carries out the physical operations required to process information. Landauer provided a connection between information processing and thermodynamics, however, this relation needs to be expanded further to capture the unavoidable cost of the working substance that is associated with the circuit structure and operation for a given paradigm. To put in another way, in addition to the lower bound on the irreversible information loss, the fundamental lower bound on the idealized circuit operation to execute that information processing according to a particular strategy needs to be considered as well. This juxtaposition of two interrelated components of unavoidable heat dissipation in computation allows us to obtain fundamental lower bounds on energy at the circuit level.

⁴Sagawa refers to this nature of LP as “model-independent” [6].

In order to obtain these fundamental lower bounds at the circuit level, our goal is to calculate heat dissipation in computation by identifying the cost of irreversible information loss, as well as the cost associated with the idealized circuit operation that is required to execute that information. This allows us to analyze computational paradigms from a fundamental energy requirement point of view and obtain lower bounds on energy at the circuit level. In this dissertation, we present a methodology that captures the fundamental lower bounds for a given nanocomputing technology. In a sense, the spirit of our approach can be explained with an analogy from thermodynamics: Carnot's bounds for heat engines. The bounds that Carnot obtained for a thermodynamic steam engine considers idealized, frictionless piston operations, and provides a bound for the physically achievable regime in the efficiency of a given engine. Similarly, the fundamental lower bounds on heat dissipation of information processing in a circuit considers computation with no parasitics, and is purely based on the nature of the irreversible information processing under an idealized circuit operation. Here, the challenge is to establish a solid connection between the energetic cost of information loss and unavoidable cost of associated physical operations by bringing the two, seemingly disassociated costs, together on the same ground. Furthermore, the dominating factors in the fundamental cost of information processing will certainly vary for different computing strategies. As such, the bounds must be formulated separately for a given paradigm and specified for each circuit in that paradigm.

The fundamental lower bounds of information processing that are tailored for a specific nanocircuit requires bringing physical laws to bear directly on the underlying computational strategy that defines the paradigm: treating information-bearing degrees of freedom and non-information-bearing degrees of freedom on an equal footing. In this work, we respond to this need this by developing a general physical-information-theoretic methodology for determination of fundamental lower bounds

on the dissipative costs of computation that can be applied directly to concrete, non-trivial nanocomputing paradigms. We, then, use this methodology to obtain lower bound on the energy cost of computation in non-transistor- and transistor-based post-CMOS technologies.

This dissertation is organized as follows. In Chapter 1, we introduce the unavoidable heat dissipation problem and discuss the limitations it imposes on emerging technologies. This requires a discussion of concepts such as reversibility, irreversibility and the laws of thermodynamics from a computational perspective. We explain how encoding of information in physical states of a system described and why its loss has a physical cost as captured by LP. Following Landauer’s inaugural work, various researchers have taken the bounds that establish a connection between energetic and entropic cost of information processing further to capture the cost of physical information processing in quantum mechanical systems [7, 8, 9]. We present these new bounds based on the “referential approach” at the end of this chapter. The studies presented in this chapter serve as the departure point for this study.

In Chapter 2, we develop a general methodology to determine fundamental lower bounds for the dissipative costs of computation. We begin by introducing the theoretical constructs, as well as the mathematical representations and dynamic assignments employed to develop our methodology. We present the most general form of the lower bounds independent of a given nanocomputing paradigm. We elaborate on the basis of our methodology and describe the steps that are required to obtain fundamental lower bounds for a broad class of computing technologies.

In Chapter 3, we apply our methodology to non-transistor- and transistor-based post-CMOS nanocomputing technologies to obtain lower bounds on the energy cost of computation. We first illustrate our methodology on a non-transistor-based circuit that can exchange energy – but not particles – with their surroundings. We consider a Quantum Cellular Automata (QCA) half adder operated under both Lan-

dauer (irreversible) and Bennett (reversible) clocking to illustrate the importance of the circuit operation on the fundamental bounds [10]. Then, we focus our attention on transistor-based nanocomputing circuits that exchange both energy and particles with their surroundings. Our methodology allows us to address this more challenging problem and obtain bounds for the NASIC (Nanoscale Application Specific Integrated Circuit) paradigm proposed by Moritz and co-workers ([11] and references therein). We consider the half [12] and full [13] adders implemented in the NASIC fabric. Illustrating our methodology via applications on transistor-based circuits allows for comparison of unavoidable costs associated with particle supply to those arising from logical irreversibility in such paradigms. In addition to such emerging paradigms, we also illustrate our methodology via an np-CMOS half adder circuit, and provide a pedagogic example for the application of our approach to the well-known and conventional CMOS paradigm. Furthermore, together with the dynamically clocked circuits, we discuss our methodology’s application to static circuits and comment on certain limitations involved.

In Chapter 4, we describe a modular approach that could allow automation for the determination of fundamental lower bounds on heat dissipation for the post-CMOS technologies. We illustrate modularized dissipation analysis for both QCA and NASIC paradigms. For the Landauer-clocked QCA half adder circuit designed according to a specified set of example design rules, we verify that the modular dissipation analysis yields identical results to that of our general approach [14]. Similarly, we generalize the fundamental lower bounds for the NASIC paradigm with multiple computational tiles. We elaborate on the dramatic analytical simplification that this modularization presents and discuss prospects for automation of our methodology. We suggest that automated implementation of our methodology could enable determination of fundamental lower bounds on dissipation for large and complex circuits such as full processors.

In Chapter 5, we provide an overview of the key concepts in thermodynamics of computation and comment on studies concerning the validity of LP. Our methodology has a distinct departure point than that of Landauer's; we employ the referential approach to obtain our lower bounds on dissipation. However, the final bounds we obtain qualitatively uphold LP and are of the same quantitative form for some cases. This resonance begs the question whether our methodology is also vulnerable to arguments leveled against the validity of LP. In this chapter, we provide an chronological list of fundamental concepts in thermodynamics of computation, and outline the arguments concerning the validity of LP. We comment on the extent to which these arguments have implications for our methodology.

In the final chapter of this dissertation, we summarize and conclude our work by elaborating on the significance of our contribution and hint at potential future directions.

CHAPTER 1

TECHNICAL BACKGROUND

In this chapter, we introduce fundamental concepts such as reversibility, irreversibility and laws of thermodynamics from a computational perspective. We explain how information is encoded in physical states of a computational system and why irreversibly discarding it has a physical cost. We also present the studies that take Landauer’s initial work further for quantum mechanical systems processing physical information. The concepts and formulations presented in this chapter serve as a departure point for this dissertation.

1.1 Reversibility, Irreversibility, and Dissipation in Computation

Heat dissipation in thermodynamics and computing is described in terms of reversibility and irreversibility. In thermodynamics, if a process is irreversible then it is defined to be dissipative. A thermodynamic system is studied within a universe in which it is situated together with its surroundings, and the thermodynamic relations governing the interactions and processes they undergo are defined within this universe. A reversible process in thermodynamic terms is an operation, which can be reversed after its transition from initial equilibrium state to final equilibrium state with no entropic or energetic cost to the system or its surroundings. An irreversible process, on the other hand, is a process, which the change in the thermodynamic state of a system and all of its surroundings cannot be precisely restored to their initial state without a resulting entropic and/or energetic cost to the universe.

From a computational point of view, a process is logically reversible if the inputs can be fully recovered from the outputs after computation. That is to say that the initial input states map one-to-one onto the final output states. Significantly, logical states in a computing system are represented by physical states of the system, i.e. the initial and final logic states correspond to the initial and final physical states of the system. The concept of logical reversibility was introduced by Rolf Landauer in 1961 [4] in connection with information-discarding process in computers. Information erasure and resetting are always logically irreversible. If the inputs to an information processing system cannot be obtained from its outputs then that computation is defined to be logically irreversible. As a consequence, throughout a computation, dissipation occurs when a part of system's energy that is capable of doing work is converted into heat during the operation.

Heat is a special form of energy measured by temperature and can be regarded as transformation of useful energy to useless energy. The perception of heat as energy, which is statistically distributed among the particles of a system, was provided by Rudolf Clausius. This interpretation of heat allowed a more refined approach to the implementation of principle of conservation of energy in both macroscopic and microscopic systems as it enables the inclusion of heat exchange of the system with its surroundings to the total energy exchange, in addition to the work performed by or on the system. The connection between the computational irreversibility and heat dissipation, therefore, allows us to use thermodynamic inequalities to obtain the fundamental energy requirements of a system to do computation.

1.2 First and Second Laws of Thermodynamics

The first and second laws of thermodynamics have an essential role in the discussion of relations governing the physical changes that a system undergoes as a consequence of any given computing process. In thermodynamics, an internal en-

ergy, U , is assigned to each macroscopic system. For isolated systems which do not exchange work or heat with their surroundings this internal energy, U , is equal to the total energy, E , of the system that we are familiar from classical mechanics and electrodynamics. However, if the system is able to exchange work or heat with its surroundings, we then take the principle of conservation of energy into account [15]. Based on the perception of heat as a form of energy introduced above, we can then state that the differential of the internal energy of a system, dU , is equal to the sum of the reversible work done on the system, dW , and the heat irreversibly exchanged with the environment, dQ ,

$$dU = dW + dQ. \quad (1.1)$$

This is the first law of thermodynamics. Here dQ is associated with the change in the entropy of the system and defined as $dQ = TdS$.¹ Integrating the above equation we obtain a fraction of the total internal energy that can reversibly be recovered to do work

$$F = U - TS. \quad (1.2)$$

Here, F represents the free energy of the system.² Erasing a bit consumes this free energy as well as changing the logical entropy dS of the system and hence dissipates heat [17].

In an information processing system, as a result of this heat dissipation, the number of logical states that is available to the system changes, and the associated thermodynamic cost can be explained by the second law of thermodynamics. The second law of thermodynamics states that the entropy of a system always increases or at most remains constant, but it never decreases;

¹This is the mathematical form of entropy expression that is initially provided by Clausius.

²Sklar, for example, uses the terms “high quality” and “low quality” for the energy and suggests that: “Energy must be available in a “high quality” state to do work. Performing the work degrades the energy into “low quality.” [16]

$$dS \geq 0, \tag{1.3}$$

therefore, it is also referred to as “the law of increase of entropy.” This law was discovered by Clausius in 1865 and its statistical account was given by Ludwig Boltzmann in the 1870s [18]. The formulation of the second law provided by Clausius is much more general and refers to any spontaneous process occurring in isolated systems and defines entropy as a state function that in an isolated system never decreases. In its essence, the second law is probabilistic as it indicates that the system will go from a relatively low probability state to a high probability state.³

Boltzman’s atomistic formulation of the second law, however, does not contain the word or concept of probability explicitly [20]. Although it is implied because the formulation is connecting the entropy with the total number of states, and there is a statistical probability distribution associated with particles’ occupation of the available states based on their type.

In the context of information erasure the second law has a significant role because erasure changes the number of logical states available to the system, which has an inevitably real thermodynamic cost. Gershenfeld [17] explains this as follows

“Entropy increases in spontaneous irreversible processes to the maximum value for the configuration of the system. If at any time the state of the bit is not close to being in its most probable state for the system, there will be extra dissipation as the entropy of the system increases to reach the local minimum of the free energy.”

This allows us to see the connection between information and thermodynamics more clearly. Below, we revisit the first and second laws of thermodynamics in a

³The probabilistic nature of the second law was poetically restated by Maxwell in a letter to John William Strutt as follows: “The second law has the same degree of truth as the statement that, if you throw a tumblerful of water into the sea, you cannot get the same tumblerful of water out again.” [19]

quantum mechanical setting during the discussion of Partovi’s work on “Quantum Thermodynamics” [7]. We now move our attention to the physicality of information and how these thermodynamic laws are used to explain the processes governing computing circuits and the information they process.

1.3 Information is Physical

Non-semantic information as processed by computing circuits is often thought of in abstract terms. However, every bit of non-semantic information, regardless of its semantic content, is encoded in the physical states of a computing circuit in order to be processed, manipulated or transmitted [21]. It’s often difficult to comprehend information independent of meaning and content in the conventional sense, and unfortunately the relation between semantic and non-semantic information is not as clear as one would expect. In the most general sense, semantic and non-semantic information are both associated with lack of information about the state of a certain system. In the context of semantic information, this lack of information is correlated with what a particular signal means as assigned by human convention. In the non-semantic sense, however, information is associated with physical correlation between two systems, and refers to the amount of information that one system has about the statistical likelihood of realization of a given signal chosen among many others in another system.⁴ It is important to underline this difference, however, further discussion on the distinction between different types of information is beyond the scope of our study.

In this work, we focus on non-semantic information that is carried in physical states of the system about the states of another system. This is defined as the

⁴In the literature, one can find slight differences in the distinction drawn between semantic and non-semantic information. For instance, Piccinini [22] refers to the concept of information, as we employ in this work, as “natural (semantic) information.”

mutual information between an information processing artifact, such as a circuit of interest, and an input referent that holds a physical instantiation of the input data to be processed by the artifact. Introducing a referent allows us to obtain the amount of information about the input referent that is irreversibly lost during circuit operation given a period of time. We discuss the details of such information processing systems and their quantum mechanical representation in the following chapter. We also introduce mathematical representations of appropriate measures as progress within our discussion. Here we simply underline that the statement “*information is physical*” implies that processing and manipulation of information in a system can be represented by the physical laws that govern the information processing system and its interaction with its environment. We employ this fact to obtain the fundamental lower bounds that can be tailored for specific nanocomputing devices.

1.4 Landauer’s Principle

Rolf Landauer used the physicality of information to provide a connection between the abstract logical computation and associated physical processes realized in a computing system. In his inaugural work, Landauer [4] proposes that

“computing machines inevitably involve devices which perform logical functions that do not have a single-valued inverse. This logical irreversibility is associated with physical irreversibility, and requires a minimal heat generation, per machine cycle, typically of the order of $k_B T$ for each irreversible function,”

where k_B is the Boltzmann constant. This is the initial and most general form of LP. The entropic and energetic bounds associated with LP are written as

$$\Delta S \geq k_B \ln(2) \Delta \mathcal{I}_{er} \quad (1.4)$$

$$\Delta E \geq k_B T \ln(2) \Delta \mathcal{I}_{er} \quad (1.5)$$

respectively where T is temperature and $\Delta\mathcal{I}_{er}$ is the amount of bits of classical information erased from a physical system.⁵

Therefore, the physics of information suggests that in addition to information always being encoded in a physical system, irreversible manipulation of it always results in a generation of at least $k_B T \ln(2)$ of heat per bit in the environment [21]. This statement has been refined over the years and a version of it as provided by Charles Bennett⁶ [24] states that

“any logically irreversible manipulation of information, such as the erasure of a bit or the merging of two computation paths, must be accompanied by a corresponding entropy increase in non-information bearing degrees of freedom of the information processing apparatus or its environment.”

This definition of LP is clearer than its original version as it introduces the entropy change in the non-information bearing degrees of freedom. We employ this refined definition of LP in the rest of this document, which helps us make our initial claim of treating information-bearing degrees of freedom and non-information-bearing degrees of freedom on an equal footing more clearly.

In a practical sense it can be stated that the manipulation of information in a circuit increases the thermodynamic entropy of the circuit and/or its environment, and hence dissipates heat. This principle provides a lower bound on the energetic cost of irreversible information loss, however, it does not provide insight into the unavoidable costs associated with implementing the physical information processing; i.e. the bound is independent of the structure and operation of the computational

⁵The definition of $\Delta\mathcal{I}_{er}$ is not provided explicitly, however, Landauer’s calculations show that the term represents the self entropy encoded in an information processing system.

⁶Bennett proved that logically irreversible operations can be embedded in and implemented by thermodynamically reversible operations [23]. It is important to note that, in this dissertation, the logically irreversible operations we study correspond to irreversible loss of information and are physically irreversible.

strategy that processes information. In addition, conventional devices have orders of magnitude higher thermodynamic degrees of freedom than the information-bearing ones. Emerging nanoelectronic devices present an opportunity, as well as an obligation, to study and understand these two levels of description together because of the fundamental limitations on the energy cost of computation that lies ahead.

1.5 Thermodynamic Relations for Quantum Systems

LP applies to the erasure of physical information from classical systems. Emerging nanoelectronic devices exhibit various quantum effects, which require quantum dynamic treatment to accurately capture the unavoidable dissipative cost of information processing. In order to address the fundamental energy requirement of quantum system, we need generalized thermodynamic relations to accommodate the treatment of these systems. Below, we present the quantum thermodynamic relations derived by Partovi [7].

In his paper Partovi focuses on the issue of determining the origin of irreversibility. First, he shows that “for microscopic systems, irreversibility and the second law follow from the dynamics of the system and do not have an independent status.” This allows him to limit his consideration to microscopic states, which also places emphasis on the fact that thermodynamic behavior is already present at quantum scale. He starts by introducing the system state based on the preparation state and measurement process. The state of the preparation is represented by a density matrix, $\hat{\rho}$, generally representing a mixed state. In Partovi’s approach, the uncertainty for this mixed states is interpreted as von Neumann entropy as

$$S(\hat{\rho}) = -Tr[\hat{\rho} \ln(\hat{\rho})]. \quad (1.6)$$

Partovi refers to this quantity as the Boltzmann-Gibbs-Shannon (BGS) entropy. The BGS entropy is a constant of the motion for a closed system, which evolves unitarily

according to a Hamiltonian. In quantum systems, the source of the irreversibility is the change in this entropy and is caused by an interaction with external systems. In general, for quantum systems, a thermodynamic interaction yields a different initial and final state. The equilibrium state of a closed system is explained by Gibbs states, which can be represented as

$$\hat{\rho} = \frac{\exp(-\beta\hat{H})}{Z} \quad (1.7)$$

where β and Z are real constants due to self-adjointness of $\hat{\rho}$. In his derivation of the thermodynamic inequalities Partovi omits Z where it is inessential, however, for the sake of consistency of notation we explicitly write it in this dissertation.

For two interacting systems \mathcal{A} and \mathcal{B} which are initially in the Gibbs states, the density operators are $\hat{\rho}^{\mathcal{A}} = \exp(-\beta\hat{H}^{\mathcal{A}})$ and $\hat{\rho}^{\mathcal{B}} = \exp(-\beta\hat{H}^{\mathcal{B}})$. They undergo a joint unitary evolution, and their interaction can be explained accordingly. After this evolution, the total entropy which is the sum of BGS entropies of each system, increases. For example, consider the initial state $\hat{\Gamma} = \hat{\rho}^{\mathcal{A}} \otimes \hat{\rho}^{\mathcal{B}}$, the final state $\hat{\Gamma}'$, and the respective entropies of initial and final states, $S[\hat{\Gamma}] = S$ and $S[\hat{\Gamma}'] = S'$. The entropies of the partial systems $S^{\mathcal{A}}$ and $S^{\mathcal{B}}$ is defined in terms of density matrices $\hat{\rho}^{\mathcal{A}} = Tr_{\mathcal{B}}\hat{\Gamma}$ and $\hat{\rho}^{\mathcal{B}} = Tr_{\mathcal{A}}\hat{\Gamma}$ where $Tr_{\mathcal{A}}$ and $Tr_{\mathcal{B}}$ represent partial trace operation on the Hilbert space of system \mathcal{A} and \mathcal{B} , respectively. This allows us to write $S^{\mathcal{AB}} = S[\hat{\rho}^{\mathcal{AB}}]$ and $S^{(\mathcal{AB})'} = S[\hat{\rho}^{(\mathcal{AB})'}]$ for initial and final state, respectively. The change in the entropy is then defined as

$$S^{\mathcal{A}'} + S^{\mathcal{B}'} - S^{\mathcal{A}} - S^{\mathcal{B}} = -Tr \left[\hat{\Gamma}' \left(\hat{\rho}^{\mathcal{A}'} \hat{\rho}^{\mathcal{B}'} - \ln \hat{\Gamma}' \right) \right].$$

Due to the convexity property of the entropy function the right hand side of this equation is non-negative, which allows us to rewrite it as

$$\Delta S^{\mathcal{A}} + \Delta S^{\mathcal{B}} \geq 0. \quad (1.8)$$

The important consequence of this general result allows us to obtain an inequality for the interaction of two system that can be formulated as follows. Let's say the system a , can initially be in any state, interacts with another system b that is initially in a Gibbs state characterized by the parameter β where the density matrix is $\hat{\rho}^{\mathcal{B}} = \frac{\exp(-\beta\hat{H}^{\mathcal{B}})}{Z}$. The mean value of energy of the system b is given by $U^{\mathcal{B}} = \text{Tr}[\hat{\rho}^{\mathcal{B}}\hat{H}^{\mathcal{B}}]$, which refers to the internal energy we previously introduced in the above chapter. We can then write the difference between the entropy and the mean energy value as

$$\Delta S^{\mathcal{B}} - \Delta\beta U^{\mathcal{B}} = -\text{Tr} \left[\hat{\rho}^{\mathcal{B}'} \ln \hat{\rho}^{\mathcal{B}'} - \hat{\rho}^{\mathcal{B}} \ln \hat{\rho}^{\mathcal{B}} - \beta \left(\hat{\rho}^{\mathcal{B}'} - \hat{\rho}^{\mathcal{B}} \right) \hat{H}^{\mathcal{B}} \right]. \quad (1.9)$$

The right hand side of this equation can be rewritten by using the relation $\ln(\hat{\rho}^{\mathcal{B}}) + \beta\hat{H}^{\mathcal{B}} = -\ln(Z)$ as

$$\Delta(S^{\mathcal{B}} - \beta U^{\mathcal{B}}) = -\text{Tr} \left[\hat{\rho}^{\mathcal{B}'} \left(\ln \hat{\rho}^{\mathcal{B}'} - \ln \hat{\rho}^{\mathcal{B}} \right) \right]. \quad (1.10)$$

Partovi shows that, due to convexity property, right hand side of this equation is non-positive, i.e. $\Delta S^{\mathcal{B}} + \beta\Delta U^{\mathcal{B}} \leq 0$ where the term is equal to zero only when $\hat{\rho}^{\mathcal{B}'} = \hat{\rho}^{\mathcal{B}}$. This inequality, along with Eq. (1.8) and conservation of energy, $\Delta U^{\mathcal{A}} + \Delta U^{\mathcal{B}}$, gives us

$$\Delta(S^{\mathcal{A}} + \beta U^{\mathcal{A}}) \geq 0. \quad (1.11)$$

If both systems were to be in Gibbs states with corresponding parameters $\beta^{\mathcal{A}}$ and $\beta^{\mathcal{B}}$ for the system \mathcal{A} and \mathcal{B} , respectively, then Eq. (1.11) implies that $\Delta(S^{\mathcal{A}} + \beta U^{\mathcal{A}}) \geq 0$ as well as $\Delta(S^{\mathcal{B}} + \beta U^{\mathcal{B}}) \geq 0$. Given $\Delta S^{\mathcal{A}} + \Delta S^{\mathcal{B}} \geq 0$, it follows that $(\beta^{\mathcal{A}} - \beta^{\mathcal{B}}) \Delta U^{\mathcal{A}} \geq 0$. This means that when two Gibbs states interact, the state with the higher value of β^{-1} loses energy to the other. For equal values of β no flow of energy and no change of state occurs. This establishes the zeroth law of thermodynamics and the existence of temperature.

The Eq. (1.11) can also be used to establish the second law of thermodynamics. Consider a cyclic transformation of any system \mathcal{A} that undergoes through a series of interactions with an array of systems \mathcal{B}_n which are initially in equilibrium at temperatures β_n^{-1} . Eq. (1.11) applies to each of these interactions and implies $\Delta S_n^{\mathcal{A}} \geq \beta_n \Delta U_n^{\mathcal{A}}$. For a cyclic transformation we have $\sum_n \Delta S_n^{\mathcal{A}} = 0$, which then gives us

$$\sum_n \beta_n \Delta U_n^{\mathcal{A}} \leq 0. \quad (1.12)$$

This inequality represents the Clausius principle. In the absence of changes in the Hamiltonian of the system, which leads to exchange of work, ΔU is the *heat*.⁷

In his paper, Partovi establishes the two laws of thermodynamics that are statistical in nature by using relations provided by quantum dynamics. In addition, he also shows how and why systems that are not in equilibrium can approach equilibrium, however, that proof is beyond the scope of this work therefore we limit our discussion on Partovi’s paper with the quantum thermodynamic inequalities he defines.

1.6 Information Erasure in Quantum Systems

The fundamental bounds associated with LP, as presented in (1.5), regard the erasure of information from a physical system as a state transformation that reduces uncertainty in the system state associated with a self-referential information measure defined in terms of the statistical state of the “erasable” system alone. This definition has been restated and reformulated to address the information erasure in fully quantum mechanical framework within *referential approach* by Anderson [8]. The referential approach introduces a referent, defined as a physical system that unambiguously holds the input data throughout computation. The referent allows us to regard information erasure as loss of correlation between the state of an erasable

⁷The term ΔU is the expected value of internal energy and hereafter represented by $\Delta \langle E \rangle$.

quantum system and the initial input; i.e. it distinguishes the erasure of information from local changes in the structure of the erasable system's state.

The closed composite quantum system considered in this framework consists of a bipartite information bearing system composed of a referent and a system, \mathcal{RS} , and an environment \mathcal{E} . The initial state of the global system is then defined as $\hat{\rho} = \hat{\rho}^{\mathcal{RS}} \otimes \hat{\rho}^{\mathcal{E}}$, where $\hat{\rho}^{\mathcal{RS}}$ and $\hat{\rho}^{\mathcal{E}}$ density operators describing the initial states of the information bearing subsystem and environment, respectively. The states of the “referent system” \mathcal{R} and the “erasable” system \mathcal{S} are initially correlated. This means that there is information about the state of \mathcal{R} initially in the state of \mathcal{S} (and vice versa), but \mathcal{RS} is isolated from \mathcal{E} . Information processing and erasure is regarded by a unitary time evolution of \mathcal{RSE} which is governed by the Hamiltonian of the global, isolated system. Since the global system is closed, dynamical evolution of \mathcal{RSE} necessarily maps initial states $\hat{\rho}$ to final states $\hat{\rho}'$ via $\hat{\rho}' = \hat{U}\hat{\rho}\hat{U}^\dagger$, where \hat{U} is the unitary time-development operator governing Schrödinger evolution of the global system. During this process \mathcal{S} is coupled to \mathcal{E} , where as by definition, \mathcal{R} remains isolated through out computation. The correlation between \mathcal{S} and \mathcal{R} , as a result, is reduced by the interaction between \mathcal{S} and \mathcal{E} . This results in erasure of part of the information about the state of \mathcal{R} that is initially encoded in the state of \mathcal{S} . Based on the formalization of these notions the quantities needed for consideration of physical costs.

The information about the state of \mathcal{R} that is in the state of \mathcal{S} is the general measure of information erasure. This is quantified by the quantum mutual information, which is the correlation entropy, and given as

$$S(\hat{\rho}^{\mathcal{R}}; \hat{\rho}^{\mathcal{S}}) = S(\hat{\rho}^{\mathcal{R}}) + S(\hat{\rho}^{\mathcal{S}}) - S(\hat{\rho}^{\mathcal{RS}}) \quad (1.13)$$

Here $S(\hat{\sigma}) = -Tr[\hat{\sigma} \log_2 \hat{\sigma}]$ is the von Neumann entropy of density operator $\hat{\sigma}$ in “information theoretic” units, and $\hat{\rho}^{\mathcal{RS}} = Tr_{\mathcal{E}}[\hat{\rho}]$ and $\hat{\rho}^{\mathcal{E}} = Tr_{\mathcal{RS}}[\hat{\rho}]$ are the reduced

density operators of the information bearing subsystem and environment, respectively. Then, the amount of information lost in an erasure operation is

$$\Delta\mathcal{I}_{er} = S(\hat{\rho}^{\mathcal{R}}; \hat{\rho}^{\mathcal{S}}) - S(\hat{\rho}^{\mathcal{R}'}; \hat{\rho}^{\mathcal{S}'}). \quad (1.14)$$

This quantity is connected to the loss of classical information about the state of \mathcal{R} , which holds classical information, that is accessible via quantum measurement on the quantum state of \mathcal{S} . This can be codified by representing the i^{th} symbol in the source alphabet, generated with a priori probability p_i , by a pure state $\hat{\rho}_i^{\mathcal{R}} = |r_i\rangle\langle r_i|$ of \mathcal{R} and encoding this symbol in a generally mixed quantum state $\hat{\rho}_i^{\mathcal{S}}$ of \mathcal{S} . The $|r_i\rangle$ corresponding to the various source symbols are assumed to be mutually orthogonal, i.e. $\langle r_i | r_{i'} \rangle = \delta_{ii'} \forall i, i'$, to ensure their distinguishability, but no assumptions are made about the orthogonality or distinguishability of the encoding states $\hat{\rho}_i^{\mathcal{S}}$. This allows us to write the initial state of \mathcal{RS} as

$$\hat{\rho}^{\mathcal{RS}} = \sum_i p_i (|r_i\rangle\langle r_i| \otimes \hat{\rho}_i^{\mathcal{S}}). \quad (1.15)$$

The total entropy of the initial state is defined as

$$\tilde{S}_{tot}(\hat{\rho}) = k_B \ln(2) [S(\hat{\rho}^{\mathcal{RS}}) + S(\hat{\rho}^{\mathcal{E}})]. \quad (1.16)$$

Similarly, we can define the corresponding quantities for the final states. This then gives us a total entropy change of

$$\begin{aligned} \Delta\tilde{S}_{tot} = k_B \ln(2) \{ [S(\hat{\rho}^{\mathcal{S}'}) + S(\hat{\rho}^{\mathcal{E}'}) - S(\hat{\rho}^{\mathcal{SE}})] \} &+ k_B \ln(2) [S(\hat{\rho}^{\mathcal{R}'}) - S(\hat{\rho}^{\mathcal{R}})] \\ &+ k_B \ln(2) \Delta\mathcal{I}_{er} \end{aligned} \quad (1.17)$$

where, $S(\hat{\rho}^{\mathcal{S}\mathcal{E}'}) = S(\hat{\rho}^{\mathcal{S}\mathcal{E}})$ and $S(\hat{\rho}^{\mathcal{R}'}) = S(\hat{\rho}^{\mathcal{R}})$ for the erasure operation, since the systems evolve unitarily. Also by using the subadditivity of the von Neumann entropy

$$S(\hat{\rho}^{\mathcal{S}'}) + S(\hat{\rho}^{\mathcal{E}'}) \geq S(\hat{\rho}^{\mathcal{S}\mathcal{E}'}), \quad (1.18)$$

the lower bound on the total entropy change is then written as

$$\Delta\tilde{S}_{tot} \geq k_B \ln(2) \Delta\mathcal{I}_{er}. \quad (1.19)$$

This equation is obtained by the referential approach, however, the end result (1.19) is identical to the entropy relation provided by LP in (1.4).

The bound (1.19) can also be written in terms of the Holevo information.⁸ The quantum mutual information associated with a bipartite state of the form (1.15) is

$$S(\hat{\rho}^{\mathcal{R}}; \hat{\rho}^{\mathcal{S}}) = S(\hat{\rho}^{\mathcal{S}}) - \sum_i p_i S(\hat{\rho}_i^{\mathcal{S}}) = \chi(\epsilon) \quad (1.20)$$

where $\chi(\epsilon)$ is the Holevo information associated with the ensemble $\epsilon = \{p_i, \hat{\rho}_i^{\mathcal{S}}\}$ of the initial states of \mathcal{S} . The bound on the entropy cost of erasure can be defined based on the above definition as

$$\Delta\tilde{S}_{tot} \geq k_B \ln(2) [\chi(\epsilon) - \chi(\epsilon')] \quad (1.21)$$

for classical erasure operation. If the system were to be reset then the reset cost becomes

$$\Delta\tilde{S}_{tot} \geq k_B \ln(2) \chi(\epsilon). \quad (1.22)$$

Similarly, we can define the energy relations for the system. The system is coupled to a thermal bath and as a result of information erasure in the system some amount

⁸Holevo information is significant as it represents the maximum value of the mutual information and hence its reduction fixes the lower bound on the erasure entropy [8].

of energy flows to this environment. Assuming that the environment is initially in a Gibbs state at temperature T , the canonical density operator is then written as

$$\hat{\rho}^{\mathcal{E}} = Z^{-1} \exp\{-\hat{\mathcal{H}}/k_B T\} \quad (1.23)$$

where $Z = \text{Tr} [\exp\{-\hat{\mathcal{H}}^{\mathcal{E}}/k_B T\}]$ is the partition function. A lower bound on the energy increase of the environment due to information erasure can then be obtained by considering the quantity $\Delta\langle E^{\mathcal{E}} \rangle - T\Delta\tilde{S}^{\mathcal{E}}$, where $\Delta\langle E^{\mathcal{E}} \rangle = \langle E^{\mathcal{E}'} \rangle - \langle E^{\mathcal{E}} \rangle = \text{Tr} [\hat{\rho}^{\mathcal{E}'} \hat{\mathcal{H}}^{\mathcal{E}}] - \text{Tr} [\hat{\rho}^{\mathcal{E}} \hat{\mathcal{H}}^{\mathcal{E}}]$. This may read as a change in the free energy but cannot be interpreted as such since the final state of the environment, $\hat{\rho}^{\mathcal{E}'}$, is not assumed to be a thermal state. By using the unitary evolution of $\mathcal{RS}\mathcal{E}$ and the concavity of the quantum relative entropy, it can be shown that this quantity is nonnegative. This allows us to write the lower bound the energy as

$$\Delta\langle E^{\mathcal{E}} \rangle \geq T\Delta\tilde{S}^{\mathcal{E}}. \quad (1.24)$$

From (1.19) we have

$$\Delta\tilde{S}^{\mathcal{E}} \geq k_B \ln(2) \{\Delta\mathcal{I}_{er} - \Delta S^{\mathcal{RS}}\} \quad (1.25)$$

which then gives us

$$\Delta\langle E^{\mathcal{E}} \rangle \geq k_B T \ln(2) \{\Delta\mathcal{I}_{er} - \Delta S^{\mathcal{RS}}\}. \quad (1.26)$$

Unlike the inequality (1.19), the above relation (1.26) is different from that of the bound obtained by using LP. In addition to the information term that is also present in LP, the bound (1.26) has an extra entropy term that distinguishes the result obtained using the referential approach.

Note that $\Delta S^{\mathcal{R}} = 0$, and $\Delta\mathcal{I}_{er} - \Delta S^{\mathcal{RS}} = S(\hat{\rho}^{\mathcal{S}}) - S(\hat{\rho}^{\mathcal{S}'})$. This allows us to write the lower bound on energy cost of erasure as

$$\Delta\langle E^{\mathcal{E}} \rangle \geq k_B T \ln(2) \Delta S(\hat{\rho}^{\mathcal{S}}). \quad (1.27)$$

The entropy change in the system in resetting operation is $\Delta S(\hat{\rho}^{\mathcal{S}}) = S(\hat{\rho}_{reset}^{\mathcal{S}'}) - S(\hat{\rho}^{\mathcal{S}})$. We can then specialize the lower bound on energy for resetting as

$$\Delta \langle E^{\mathcal{E}} \rangle \geq k_B T \ln(2) \left(S(\hat{\rho}^{\mathcal{S}}) - S(\hat{\rho}_{reset}^{\mathcal{S}'}) \right). \quad (1.28)$$

Irrespective of the amount of information erased this energy is required to flow to the environment when the resetting to the standard state lowers the entropy of \mathcal{S} . If the states of the system are distinguishable that they have support on mutually orthogonal subspaces in quantum description we can further specialize the above bounds using $\Delta \mathcal{I}_{er} = H$ and $\Delta S^{\mathcal{S}} = S(\hat{\rho}_{reset}^{\mathcal{S}'}) - H - \sum_i p_i S(\hat{\rho}_i^{\mathcal{S}})$, where H is the Shannon entropy [25], or classical “encoding entropy,” $H = -\sum_i p_i \log_2 p_i$, associated with the ensemble $\epsilon = \{p_i, \hat{\rho}_i^{\mathcal{S}}\}$, of encoding states, and obtain the lower bounds on the entropy and energy cost as

$$\Delta \tilde{S}^{\mathcal{E}} \geq k_B \ln(2) H \quad (1.29)$$

$$\Delta \langle E^{\mathcal{E}} \rangle \geq k_B T \ln(2) \{ H - \langle \Delta S_i^{\mathcal{S}'} \rangle \}. \quad (1.30)$$

Here $\langle \Delta S_i^{\mathcal{S}'} \rangle = S(\hat{\rho}_{reset}^{\mathcal{S}'}) - \sum_i p_i S(\hat{\rho}_i^{\mathcal{S}})$ is the average entropy change in \mathcal{S} upon resetting averaged over all initial states.

In addition to “erasure-by-reset,” the erasure of information can be defined in association with “partial-erasure” scenarios. When information is erased by reset, the system is set back to its standard state with no trace of information left in the system after erasure operation. Partial-erasure, however, indicates an incomplete erasure operation after which a trace of information is left in the system, which indicates that the system is not completely reset to its standard state. This may be a result of a noisy communication or computation, for instance if the fidelity of the signal carriers is compromised by noise then the system would still carry some

amount of information after erasure. In the literature, “erasure-by-reset” has been widely studied for canonical systems. The approach presented by Anderson covers a more general class of state transformations, which allows us to go beyond the erasure of information encoded in distinguishable states, and treat partial-erasure scenarios in which the information is encoded in general quantum states.

1.7 Landauer’s Principle Specialized for L -machines

The above form of LP enables us to address the information erasure in quantum systems irrespective of the logic operation and computational strategy. In [9], Anderson further expanded this bound by specifying the logic transformations in a computing machinery, and specialized LP to account for fundamental energy requirements of specific logic operations in terms of entropy change in the information processing system as well as the amount of information erased.

Anderson’s paper is based on the notion of L -machine, that is introduced by Ladyman *et. al.* [26] as a part of their study on the connection between logical and thermodynamic irreversibility. In a later paper [27] Ladyman defined L -machine as a “hybrid-physical-logical entity that combines a physical device, a specification of which physical states of that device correspond to various logic states, and an evolution of that device which corresponds to the logical transformation \mathcal{L} .” The form of L -machine that is studied in Ladyman’s papers is classical and highly idealized. Anderson expanded this notion by generalizing it to classical logical transformations in quantum mechanical systems that can be noisy and hold imperfect representation of information.

The formal definition of quantum mechanical L -machine that Anderson develops allows us to obtain physical cost of “single-shot” implementation of logical transformation \mathcal{L} . In actual quantum mechanical systems that process physical information through classical logical transformations the computation is achieved by sequences of

cycles. Anderson's study on the quantum mechanical L -machines also sets the stage for such realistic computing scenarios.

Based on the generalized L -machine formulation, the fundamental lower bound on unavoidable heat dissipation in a quantum mechanical system as a result of a classical irreversible logic operation is obtained as follows. Consider a device \mathcal{D} that interacts with an environment \mathcal{E} . Assuming that the joint evolution of the composite system \mathcal{DE} is unitary allows us to use Partovi's result presented above, i.e. "any decrease in the von Neumann entropy $S^{\mathcal{D}}$ of the device results in an increase in the expected energy of the environment that is lower bounded as"

$$\Delta \langle E^{\mathcal{E}} \rangle \geq k_B \ln(2) \Delta S^{\mathcal{D}}, \quad (1.31)$$

where $\langle E^{\mathcal{E}} \rangle = \text{Tr} [\hat{\rho}^{\mathcal{E}} \hat{H}^{\mathcal{E}}] = \sum_{i=1}^M p_i \text{Tr} [\hat{\rho}_i^{\mathcal{E}} \hat{H}^{\mathcal{E}}] = \sum_{i=1}^M p_i \langle E_i^{\mathcal{E}} \rangle = \langle \langle E_i^{\mathcal{E}} \rangle \rangle$, with $\hat{H}^{\mathcal{E}}$ is the environment Hamiltonian. The entropy reduction in the device is defined as

$$-\Delta S^{\mathcal{D}} = S(\hat{\rho}^{\mathcal{D}}) - S(\hat{\rho}^{\mathcal{D}'}) . \quad (1.32)$$

Initial entropy of the device is expanded as

$$S^{\mathcal{D}} = H(X) + \sum_{i=1}^M p_i S(\hat{\mathcal{D}}_i^{\text{in}}), \quad (1.33)$$

where M is the number of device states and $\hat{\mathcal{D}}_i^{\text{in}}$ represents the initial device states which are mutually orthogonal. By using the the definition of Holevo information

$$\mathcal{X}(\epsilon_X^{\mathcal{D}'}) = S\left(\sum_{i=1}^M \Lambda(\hat{\mathcal{D}}_i^{\text{in}})\right) - \sum_{i=1}^M p_i S(\Lambda(\hat{\mathcal{D}}_i^{\text{in}})), \quad (1.34)$$

for the ensemble $\epsilon_X^{\mathcal{D}'} = \{p_i, \Lambda(\hat{\mathcal{D}}_i^{\text{in}})\}$. The information loss for a general quantum L -machine in terms of Holevo information is then

$$-\Delta\mathcal{I} = H(X) - \mathcal{X}(\epsilon_X^{\mathcal{D}}). \quad (1.35)$$

Based on these definition the change in entropy can be rewritten as

$$-\Delta S^{\mathcal{D}} = -\Delta\mathcal{I} - \langle \Delta S_i^{\mathcal{D}} \rangle, \quad (1.36)$$

where $-\Delta\mathcal{I}$ is the amount of information loss. The average device entropy reduction is

$$-\Delta S_i^{\mathcal{D}} = \sum_{i=1}^M p_i \left(S(\Lambda(\hat{\mathcal{D}}_i^{in})) - S(\mathcal{D}_i^{in}) \right). \quad (1.37)$$

Substituting these equalities into Partovi's inequality, Anderson obtains the lower bound on the expected valued of energy in terms of the information loss and the change in device entropy,

$$\Delta \langle E^{\mathcal{E}} \rangle \geq k_B \ln(2) (-\Delta\mathcal{I} - \langle \Delta S_i^{\mathcal{D}} \rangle). \quad (1.38)$$

This indicates that the average expected energy value of the environment necessarily increases as a result of information loss in a quantum L -machine as long as the entropy of the representative states, on average, increases more than the information loss after its evolution from its initial state to final state. If the device entropy does not increase as a result of information erasure then the above equation reduces to a simpler yet commonly encountered version of LP

$$\Delta \langle E^{\mathcal{E}} \rangle \geq k_B \ln(2) \Delta\mathcal{I}. \quad (1.39)$$

As stated earlier, the above inequality (1.38) can be used to obtained the fundamental lower bound on the physical cost of “single-shot” implementation of logical transformation L in a quantum mechanical system. We derive the fundamental lower bound on unavoidable heat dissipation from Eq. (1.38) and build our methodology based on the studies presented in this chapter.

CHAPTER 2

HEAT DISSIPATION FOR NANOCOMPUTING: METHODOLOGY

Our goal is to develop a general methodology for determination of lower bounds on the dissipative cost of computation that can be applied directly to concrete nanocomputing paradigms. In order to accomplish this, we need to bring physical laws directly to bear on the underlying computational strategy that defines the paradigm. In the previous chapter, we outlined laws and relations that govern the physical interactions required for computation. In this chapter, we introduce mathematical representations and dynamic assignments that allow us to tailor the fundamental bounds for a given nanocomputing paradigm. The theoretical constructs we define in this chapter are presented in their most general form, independent of a given nanocomputing technology. The methodology proposed here [12] is applied to various paradigms in Chapter 3 and 4.

2.1 Introduction

Information processing in a circuit involves various interactions, in both physical-information theoretic level and thermodynamic level, depending on the computational paradigm. These interactions are represented by information-bearing degrees of freedom and non-information-bearing degrees of freedom, respectively. We obtain fundamental lower bounds on heat dissipation for a given paradigm by treating these degrees of freedom on an equal footing. Below, we explain this detail by discussing the steps required for determination of fundamental bounds and introduce the math-

ematical representations and dynamic assignments that allow us to obtain the bounds for a given paradigm.

There are two main steps to obtain the fundamental lower bounds on the dissipative cost of computation of a given nanocomputing scenario using our methodology. First, an idealized physical abstraction of the circuit and its operation is constructed. Second, using a spacetime decomposition of computation in the circuit abstraction, physical-information-theoretic analyses of the global system evolution are performed for the time intervals relevant to each computational step. The abstract description of the circuit and its operation is constructed so that it captures nothing more and nothing less than the essential functional features of the underlying computational strategy, implemented precisely as envisioned in the computational paradigm. This is to say that the abstraction describes *paradigmatic operation* of the circuit. The analyses yield lower bounds on the energy that is unavoidably dissipated into the circuit’s local environment on each computational step, under paradigmatic operation, including both the dissipation required to execute logically irreversible operations and other unavoidable paradigm-dependent “overhead” costs (e.g. particle supply required to maintain the computational working substance in transistor-based paradigms). Bounds obtained for each computational step are finally summed over all steps in the computational cycle to obtain a fundamental lower bound on the (input-averaged) energy cost of each computation performed by the circuit. In the next chapter, we further articulate and illustrate applications to various nanocomputing circuits. First, we begin by introducing the lower bounds on the fundamental heat dissipation in the most general, paradigm-independent form below.

2.2 Abstraction

The abstraction is composed of two main stages. First, we create a physical abstraction of the circuit and its surroundings in a globally closed and isolated universe.

Second, in the process abstraction, we present assignments to identify local physical operations. The assignments are decomposed into two groups for control and restoration operations. The circuit’s interaction with the other information-bearing subsystems and the bath is described by the control operations. The restoration operations represent the coupling between the remote environment and the bath. The control and restoration operations provide the circuit evolution required to implement computation. We now discuss the abstraction procedure in detail.

2.2.1 Physical Decomposition

The physical abstraction of the circuit and its surroundings is depicted schematically in Fig. 2.1. The upper half of the figure represents the computationally relevant domain. This domain includes an information processing artifact \mathcal{A} which is the computing circuit of interest and computationally supporting subsystems such as external registers and adjacent circuit stages, as well as an input referent \mathcal{R} that holds a physical instantiation of the input data that will be processed by the artifact. The lower half represents the environmental domain consisting of a heat bath \mathcal{B} , which is the part of the environment that is in direct thermal contact with the artifact and nominally at temperature T . The greater environment $\bar{\mathcal{B}}$ includes heat reservoirs that “rethermalize” the bath and anything else that is required to ensure that the universe is globally closed. Fig. 2.1 depicts the location of each subsystem and its interaction with another subsystems.

Below, we elaborate on the definition of each domain and all the elements contained in these domains.

- *Computationally Relevant Domain* - The part of the universe that is directly relevant to consideration of computation by a physical artifact:
 - *Information Processing Artifact \mathcal{A}* - The nanocomputing circuit of interest, which, under paradigmatic operation, can be used to evaluate a specified

- *Supporting Computational Subsystems $\bar{\mathcal{A}}$* - Computationally relevant subsystems external to \mathcal{A} that exchange information about x_i or $L(x_i)$ with \mathcal{A} (e.g. input/output registers, previous and subsequent circuit stages).
- *Environmental Domain* - The environmental domain includes everything else, which is subdivided as follows:
 - *Heat Bath \mathcal{B}* - A part of the environment, nominally at temperature T , that interacts directly with \mathcal{A} , can exchange heat with \mathcal{A} , and can serve as a “sink” for information lost from \mathcal{A} .
 - *Environment $\bar{\mathcal{B}}$* - The greater environment, including remote heat reservoirs that “rethermalize” the bath (i.e. drive the bath toward a thermal equilibrium state at temperature T after it has been driven away from equilibrium), remote particle reservoirs that supply the electrical work required to “recharge” the artifact (i.e. drive particle densities of local particle reservoirs - such as supply wires - back to their nominal values after they have exchanged particles with the nanocircuit), and anything else that must be included to ensure that the global system $\mathcal{RA}\bar{\mathcal{A}}\mathcal{B}\mathcal{B}$ is isolated (i.e. that it cannot exchange heat, work, or matter with anything external to $\mathcal{RA}\bar{\mathcal{A}}\mathcal{B}\mathcal{B}$).

As we mentioned above, these subsystems are situated in a globally closed and isolated universe. Constructing a globally closed universe as presented above enables us to assume it evolves unitarily via Schrödinger’s equation. The global unitarity, together with identification of the subsystems that are coupled in each step, allows determination of the fundamental lower bounds that we are after in this work.

We now move on to the second stage of the abstraction and introduce the assignments that we use for process abstraction.

2.2.2 Process Abstraction

During computation, the subsystems are driven away from equilibrium but then rethermalized as a part of the process abstraction. We identify a set of local physical operations $\phi_t \in \{\phi_t\}$, each of which is decomposed into a control process and a restoration process. Control processes are the local operations that act during specified time intervals to change the states of representational elements in the artifact either unconditionally or conditioned on the states of other representational elements. Typically, they involve interaction between the artifact, other information-bearing subsystems, and the bath. Restoration processes are the local operations that couple the remote environment $\bar{\mathcal{B}}$ to the bath \mathcal{B} and local particle reservoirs in \mathcal{A} . These operations rethermalize the bath and recharge the reservoirs after they have been driven from their nominal states by control operations. Together, the control and restoration phases make up the sequence of global system evolutions required for implementation of computation in the circuit. The complete definition for these substeps is provided below.

- *Control Operations* - Local operations that act during specified time intervals to change the states of representational elements in the artifact either unconditionally or conditioned on the states of other representational elements. Typically involves interaction between the artifact, other information-bearing subsystems, and the bath. Denote as $\phi_t \in \{\phi_t\}$, where $\{\phi_t\}$ is the set of control operations employed by the artifact.
- *Restoration Processes* - Local operations that couple the remote environment $\bar{\mathcal{B}}$ to the bath \mathcal{B} and local particle reservoirs in $\bar{\mathcal{A}}$ that rethermalize the bath and recharge the reservoirs after they have been driven from their nominal states by control operations.

This completes the first step of our methodology. We can now discuss the analysis that we pursue based on the above assignments presented under the abstraction step.

2.3 Analysis

The second step in our approach involves spacetime decomposition of the circuit function (operational decomposition) and physical-information-theoretic analyses of local dissipation into the bath throughout the computational cycle (cost analysis). Any local information about the initial state of \mathcal{A} that is irreversibly lost during a computational step induces dissipation in \mathcal{B} before being swept completely into $\bar{\mathcal{B}}$ during the restoration process. Note that, loss of initial-state information from part of \mathcal{A} is locally irreversible if it is erased in the absence of interaction with other parts of \mathcal{A} or $\bar{\mathcal{A}}$ that hold or receive copies of the initial state during the clock step. This locally irreversible information loss affects the state of \mathcal{B} during an operation's control process, which is precisely the point at which the dissipation costs are “cached out” in our approach. The details of this procedure are outlined below.

2.3.1 Operational Decomposition

In this stage of the analysis, we perform a spacetime decomposition of the circuit. We first define clock zones, subzones, steps and cycles, which allows us to define computation steps and cycles. The definition of each concept is listed below.

- *Clocking* - We present the relevant concepts and definitions we employed for clocking below.
 - *Clock Zones and Subzones* - A clock zone is a set of representational elements that are simultaneously subjected to the same control operation in any given time step. Each clock zone may consist of physically disjoint subsets of representational elements - called clock subzones - that do not

interact with one another directly as they change state. We denote the u^{th} clock zone as $C(u)$ and the l^{th} clock subzone of $C(u)$ as $C_l(u)$.

- *Clock Step* -It represents a time step during which a specified set of control operations are applied to various clock zones. We denote the assignment of control operation ϕ_t to clock zone $C(u)$ as $(C(u); \phi_t)$, the v^{th} clock step ϕ_v is specifically defined as an assignment

$$\varphi_v : \{(C(u); \phi_t)\}_v$$

of control operations to all clock zones. The restoration processes that rethermalize the bath and recharge the artifact's local particle reservoirs after the associated control operation drives these subsystems from their nominal states is also included in a clock step.

- *Clock Cycle* - A clock cycle is one period of the periodic sequence $\Phi = \phi_1\phi_2\phi_3\dots$ of clock steps applied to the artifact to enable its operation.
- *Computation* - Above definitions we present for clocking allows us to define the computational steps and cycle. We list the relevant concepts and definitions below.

- *Computational Step*- The computational step c_k , defined for the η -th input $x_i^{(\eta)}$ in an input sequence $\dots x_i^{(\eta-1)} x_i^{(\eta)} x_i^{(\eta+1)} \dots$, is the k -th of K clock steps required for evaluation of $L(x_i^{(\eta)})$.
- *Computational Cycle*- The η -th computational cycle is the sequence

$$\Gamma^{(\eta)} = c_1 \dots c_k \dots c_K$$

of the K clock steps required to fully implement L for the η -th input $x_i^{(\eta)}$, including the phase that loads $x_i^{(\eta)}$ into the artifact, the phases that evalu-

ate $L(x_i^{(\eta)})$, and the phases that transfer $L(x_i^{(\eta)})$ to the outside world and erase all information about $x_i^{(\eta)}$ from the artifact. Denote c_1 as the LOAD phase and c_K as the phase in which all correlation between the computational state of the artifact and the i -th referent is lost. $\Gamma^{(\eta)}$ may include clock steps from multiple clock cycles, and, in artifacts that pipeline input data, $\Gamma^{(\eta)}$ may *exclude* clock steps that implement operations belonging only to other computational cycles (e.g. $\Gamma^{(\eta-1)}$ and $\Gamma^{(\eta+1)}$), i.e. clock steps that do not affect representational elements, whose states depend on the η -th input. Thus, the η -th computational cycle includes only clock phases that contribute directly to evaluation of $L(x_i^{(\eta)})$ in the information processing artifact.

We can now move on to the calculation of total dissipative cost associated with one computational cycle based on information dynamics, which involves data zones and subzones presented above.

2.3.2 Cost Analysis

We calculate the total dissipative cost associated with one computational cycle based on information dynamics. We present the concepts and definitions related to this dynamics analysis below.

- *Information Dynamics* - Here we define data zones and sub zones that allows us to track information flow and identify irreversible loss of information and hence calculate the associated unavoidable cost.
 - *Data Zones and Subzones* - For the η -th computational cycle, the k -th data zone is the set of representational elements that, at the completion of the k -th computational step c_k , hold information about the input data $x_i^{(\eta)}$. A data zone may contain clock subzones belonging to multiple clock

zones, and need not include all subzones belonging to any given clock zone. It may consist of physically disjoint subsets of representational elements - called *data subzones* - which do not interact with one another directly during some or all of the computational steps. We denote the data zone associated with computational step c_k as $D(c_k)$, and the w -th data subzone of $D(c_k)$ as $D_w(c_k)$. Note that, regardless of the circuit implementation, there is one data zone defined at the end of computational step of the computational cycle from c_1 to c_{K-1} . By definition, there are no data zones at the completion of step c_K . As a computation progress through steps in the computational cycle, the data zone changes its size and topology while propagating from input to output. Data subzones generally split and merge throughout a computation, generally changing in number from step to step.

- *Information Loss* - The total information loss is the sum of contributions from information lost to the bath in each computational step, with “information loss” for the k -th step defined as the amount of information *about the state of each subzone in the $k - 1$ -th data zone* that is not in the state of the corresponding subzone in the k -th data zone, summed over all data subzones. This implies assignment of individual “subzone referents,” with appropriate enumeration of states and corresponding probabilities, to each data subzone.

This conjecture follows from the assumption that data subzones that do not interact with one another essentially “act alone” for the purposes of dissipation calculations, as the interactions that erase information about the prior state are necessarily local. This highlights the importance of correctly identifying and classifying the relevant physical interactions that occur throughout the computational cycle, as identification of the data

subzones - and thus the appropriate level of analysis for obtaining dissipation bounds - depends on the nature of these interactions.

- *Dissipation Bounds* - In any given computational step, any information lost from \mathcal{A} that is not completely transferred to $\bar{\mathcal{A}}$ results in local energy dissipation into the bath. Information is “lost” from data subzone $D_w(c_{k-1})$ during computational step c_k if, at the conclusion of c_k , the initial states of erased clock subzones of $D_w(c_{k-1})$ cannot uniquely inferred from the final states of clock subzones $C_l(u) \in D(c_k)$ that interacted directly with $D_w(c_{k-1})$ during c_k . The total dissipative cost of one computational step is taken to be the sum of contributions from all informationally lossy data-subzone-to-data-subzone transitions, and the cost of processing one input is then the sum of contributions from each computational step. This is to say that the total energy cost is additive at the subzone level, i.e. the energy costs associated with these individual data-subzone information losses can be cashed out individually and summed over a full cycle to get:

$$\langle E \rangle_{TOT} = \sum_{k=1}^K \Delta \langle E \rangle_k = \sum_{k=1}^K \left[\sum_{w \in \{w\}_{k-1}} \Delta \langle E \rangle_{k-1}^{(w)} \right]. \quad (2.1)$$

where $\langle E \rangle_k$ and $\Delta \langle E \rangle_{TOT}$ represent the average energy transferred to the bath in the k -th computational step and over the full cycle, respectively, increase in the energy of the bath over one computational cycle.

In conclusion, the analytical strategies developed in this chapter rely on the assumption that the circuit abstractions we constructed appropriately captures all computationally functional features of the underlying computational strategy. The abstractions and decompositions used in this methodology are essential for proper localization and quantification of the resulting dissipation costs. This requires constant scrutiny and reevaluation as our approach is applied in various contexts. The

dissipation bounds reflect truly the fundamental costs associated with the circuit operation based on this abstraction. We acknowledge that certain circuit features that are deemed “non-functional” in an initial analysis, such as a parasitic capacitance not intended as a memory element, may play a functional role, such as energy-reducing storage of charge for later reuse. Therefore, a redefinition of the circuit abstraction and the notion of “paradigmatic operation” need to be developed to address such features to obtain tighter bounds.

We apply the foundations developed in this chapter to obtain the fundamental energy requirements of specific nanoelectronic technology proposals. In the next chapter, we illustrate applications of this methodology to various nanocomputing strategies.

CHAPTER 3

HEAT DISSIPATION BOUNDS FOR NANOCOMPUTING: APPLICATION

In this chapter, we illustrate application of the approach presented in Chapter 2 to specific nanocomputing paradigms. We consider both non-transistor- and transistor-based paradigms, and use prominent examples from each nanocomputing technology to illustrate application of our methodology. First, we focus on non-transistor-based circuits, which can exchange energy with their surroundings but not particles. As an illustrative example we employ a Quantum Cellular Automata (QCA) half adder [10]. We also consider nanocomputing circuits that exchange both energy and particles with their surroundings, which allows us to calculate the fundamental bounds on energy cost in transistor-based technologies. We apply our methodology to an early version of the NASIC (Nanoscale Application Specific Integrated Circuit) paradigm proposed by Moritz and co-workers ([11] and references therein) by using 1-bit half and full adder implementations of the fabric. In addition to these post-CMOS technologies, we also calculate the fundamental lower bounds for conventional and well-known CMOS circuits for pedagogic purposes; we illustrate our methodology via its application to an np-CMOS half adder. The circuit examples we study throughout this chapter are dynamically clocked, in the last section of the chapter we also discuss our methodology's application to static circuits and comment on certain limitations involved.

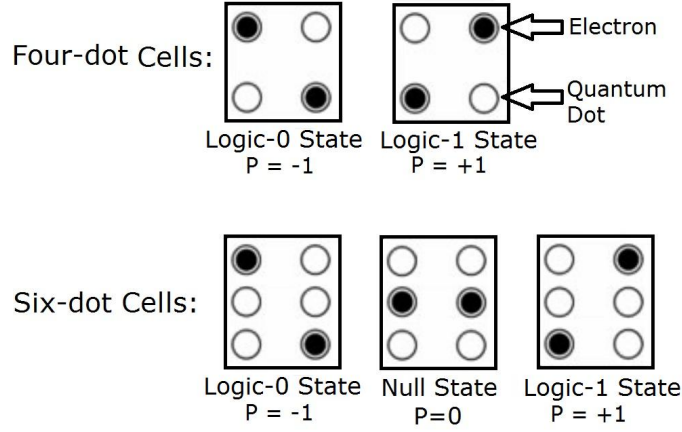


Figure 3.1. Polarizations and logic states for four-dot and six-dot quantum cells.

3.1 Non-transistor-based Applications

The non-transistor-based technology we employ to illustrate our methodology is the QCA paradigm. QCA is an emerging technology that is proposed as an alternative to transistor-based computing technologies. QCAs are composed of arrays of bistable cells that support binary computing. The technology relies on “arrays of quantum device cells in a locally-interconnected architecture” [28] that are used to implement logical devices [29]. Each cell contains quantum dots that are either unoccupied or occupied by electrons. The logic states are represented by using different configuration of electron occupancy. The four-dot cell can represent logic-0 and logic-1 states by using the polarization of the charge in opposite cells. The null state in the four-dot cell is a mixture of 0 and 1 states. The six-dot quantum cells can represent a null state, logic-0 and logic-1 states orthogonally as shown in Fig. 3.1

QCA circuits can be composed of four- or six-dot cells. In our study, we assume that the input states are orthogonal, therefore require the circuit to made up of six-dot cells.¹ operation we consider does not depend on the cell type, therefore the QCA

¹In the four-dot cells, the null state is a mixture of 0 and 1 states, therefore is not orthogonal.

half adder we present can be constructed by using either four- or six-dot quantum cells. The stability of cell polarization and coupling to neighboring cells is the essence of computation in QCA. The polarization of charges in a cell changes according to the polarization of charges in the neighboring cells due to electrostatic interactions; i.e. information flow does not involve charge transport. Information in a QCA circuit propagates by means of the electrostatic interaction between the neighboring cells. The direction and flow of interaction between the QCA cells can be controlled by using various clocking schemes. In this study we obtain the fundamental lower bounds on energy dissipation in QCA half adder circuit that is operated under two different clocking schemes; Landauer (irreversible) and Bennett (reversible) clocking. Below we discuss the operation of QCA circuit Landauer and Bennett clocking, and demonstrate how difference in clocking scheme effects the fundamental lower bounds on the heat dissipation in this circuit.

3.1.1 Quantum Cellular Automata (QCA) Half Adder with Landauer Clocking

The cell layout and logic diagram of the QCA half adder circuit used in this study is illustrated in Fig. 3.2 along with its associated timing diagram for the Landauer clocking scheme. This layout is based on the circuit used in our earlier study (see Ref. [10])² as depicted in Fig. 1 of Ref. [10]. The timing diagram depicts the Landauer clocking of the circuit using the four-phase adiabatic clocking scheme, which is color coded to indicate the corresponding clock domains on the cell layout and functional features of the logic diagram. The cells in each clock region *switch* to data states determined by the states of cells at the edges of neighboring regions during the leading clock edge, *hold* their states when the clock is high, *release* during the falling edge, and *remain* relaxed in a null state when the clock is low. The clocking of the various

²Note that the circuit employed in Ref. [10] was based on the layout presented in Ref. [30]

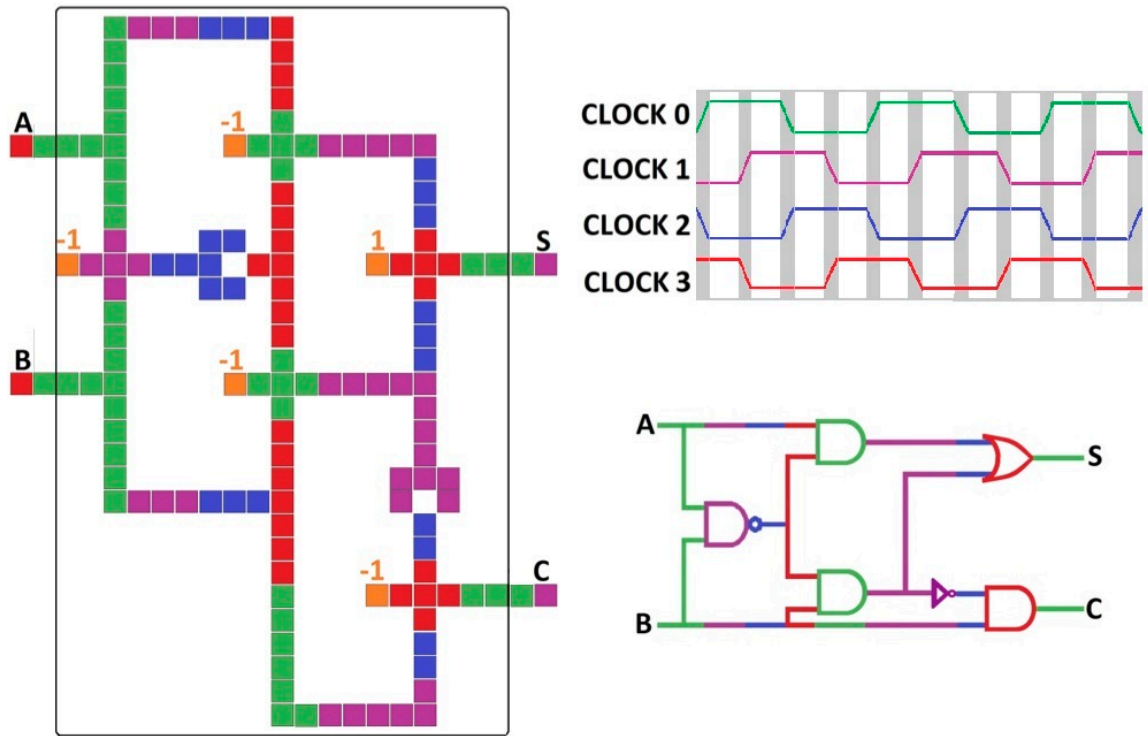


Figure 3.2. Layout, clocking, and gate-level representation of a Landauer-clocked QCA half adder circuit with no line crossings.

regions is synchronized as shown in Fig. 3.2. This allows the data to move from left to right through the adder. The four clock waveforms and corresponding clock zones for Landauer clocking are indicated in green, magenta, blue, and red, respectively. The gray shaded clocking regions in the timing diagram correspond to the eleven steps of one computational cycle. Note that eleven computational steps are required for one computational cycle, from loading of input from the external source register AB through evaluation and transfer output to the external destination register CS . The pipelining is supported in this circuit under Landauer clocking, which means that while input data for any given computational cycle is loaded from register AB into the green cells on the left side of the circuit (on the leading edge of $CLOCK$ 0), intermediate results for the previous computational cycle are being clocked into the green cells on the right side of the circuit and are still two clock steps away from the output. Let us now present these structure and process details by using our methodology.

3.1.1.1 Abstraction

The first step is construction of physical abstractions of the circuit and its surroundings (physical decomposition) and the physical processes that enable paradigmatic operation (process abstraction) as presented in Chapter 2. The intention is to capture the essential computational strategy of the QCA adder within a physical description that is compatible with physical-information-theoretic analysis, from which the dissipation bound can be obtained.

Physical Decomposition – The universe that our system of interest is situated in is depicted in Fig. 3.3. In the circuit abstraction corresponding to paradigmatic operation of the QCA adder, the artifact \mathcal{A} is the QCA adder circuit and the representational elements are the individual QCA cells in this circuit. Each cell is regarded as an elementary physical system that can support two distinguishable antipodal “data”

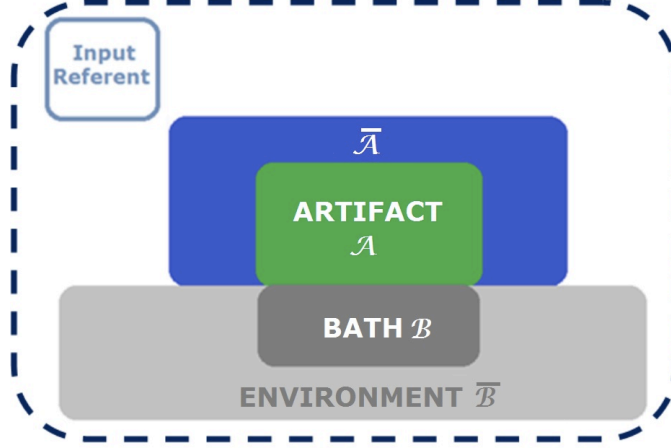


Figure 3.3. Physical abstraction of the QCA circuit and its surroundings.

states and one null state. Key subsystems external to \mathcal{A} include the input/output register cells AB and SC (which belong to $\bar{\mathcal{A}}$) and the underlying substrate in thermal contact with cells in the adder circuit (which belongs to \mathcal{B}), and other environmental subsystems (belonging to $\bar{\mathcal{B}}$) that interact with \mathcal{B} (but not \mathcal{A}) and work always to drive \mathcal{B} toward equilibrium at temperature T . The set of computational states is the set of all combinations of cell states that are accessed when all possible combinations of addend and augend bits are clocked through a full computational cycle, defined so it extends from the step that loads inputs into the circuit from the external register AB through the step that erases all information about the initial state of AB (i.e. about \mathcal{R}) from the circuit after the resulting sum and carry values have been evaluated and stored externally in SC .

As we mentioned in the previous chapter, constructing a globally closed system enables us to assume the system evolves unitarily via Schrödinger's equation. The global unitarity allows us to obtain fundamental lower bounds that we are after in this work.

Process Abstraction – During computation the subsystems are drawn away from equilibrium but then rethermalized as a part of the process abstraction. The four ele-

mentary operations relevant to operation of the QCA adder correspond to the four phases of the adiabatic clock cycle – switch (ϕ_1), Hold (ϕ_2), Release (ϕ_3), and Relax (ϕ_4) – each of which acts on specified parts of the circuit in any given clock step. The restoration processes defined for the ϕ_t unconditionally rethermalize the bath if it has been driven from equilibrium during the control operation. These operations provide the circuit evolution as required to implement computation.

We can elaborate further on the state transformations associated with the control and restoration process for all elementary operations performed throughout one computational cycle of the QCA adder as follows. Initially, all cells in the adder circuit are in NULL states, the bath \mathcal{B} is in a thermal state, and the η -th input referent is in the state

$$\hat{\rho}^{\mathcal{R}_\eta} = \sum_{i=1}^4 p_i |x_i^{\mathcal{R}_\eta}\rangle \langle x_i^{\mathcal{R}_\eta}| \quad (3.1)$$

where the $|x_i^{\mathcal{R}_\eta}\rangle$ are orthogonal pure states encoding the inputs $\{x_i\} = \{A_i, B_i\}$ and p_i is the i^{th} input probability. This referent is initially correlated with the external input cells A and B .

On each computational step, the globally closed composite system evolves according to Shrödinger equation. The initial state of each computational step c_k is shown together with the structure of the control operation relevant to that state transformation in Table 3.1. Recall here that \mathcal{A}_k denotes the subsystem of \mathcal{A} that participates in the k -th computational step c_k . Each computational step concludes with a restoration process. For computational steps that need not be dissipative, \hat{U}^{rest} is simply identity operation. However, for the dissipative steps c_6 , c_9 and c_{11} the restoration operator is $\hat{U}^{rest} = \hat{U}^{\mathcal{B}\bar{\mathcal{B}}} \otimes \hat{I}^{\mathcal{R}_\eta \mathcal{A}_k \bar{\mathcal{A}}_k}$.

The sequence of state transformations specified in Table 3.1. is relevant to a single shot computation, so $\mathcal{R}_{\eta+1}$ is not loaded into the circuit as \mathcal{R}_η is being processed as would be the case in pipelined operation of the circuit.

Computational Steps	Initial State	State Transformation	Control Operation
c_1	$\hat{\rho}_0 = \left(\sum_{i=1}^4 p_i \hat{\rho}_i^{\mathcal{R}_\eta} \otimes \hat{\rho}_i^{\bar{\mathcal{A}}_k} \right) \otimes \hat{\rho}^{\mathcal{A}_k} \otimes \hat{\rho}^{\mathcal{B}} \otimes \hat{\rho}^{\bar{\mathcal{B}}}$	$\hat{\rho}_1 = \hat{U}^{rest} \hat{U}_1 \hat{\rho}_0 \hat{U}_1^\dagger \hat{U}^{rest\dagger}$	$\hat{U}_1 = \hat{U}^{\mathcal{R}_\eta \mathcal{A}_k \bar{\mathcal{A}}_k} \otimes \hat{I}^{\mathcal{B} \bar{\mathcal{B}}}$
c_2	$\hat{\rho}_1 = \left(\sum_{i=1}^4 p_i \hat{\rho}_i^{\mathcal{R}_\eta} \otimes \hat{\rho}_{0,i}^{\mathcal{A}_k} \right) \otimes \hat{\rho}^{\bar{\mathcal{A}}_k} \otimes \hat{\rho}^{\mathcal{B}} \otimes \hat{\rho}^{\bar{\mathcal{B}}}$	$\hat{\rho}_2 = \hat{U}^{rest} \hat{U}_2 \hat{\rho}_1 \hat{U}_2^\dagger \hat{U}^{rest\dagger}$	$\hat{U}_2 = \hat{U}^{\mathcal{R}_\eta \mathcal{A}_k} \otimes \hat{I}^{\mathcal{B} \bar{\mathcal{A}}_k \bar{\mathcal{B}}}$
c_3	$\hat{\rho}_2 = \left(\sum_{i=1}^4 p_i \hat{\rho}_i^{\mathcal{R}_\eta} \otimes \hat{\rho}_{1,i}^{\mathcal{A}_k} \right) \otimes \hat{\rho}^{\bar{\mathcal{A}}_k} \otimes \hat{\rho}^{\mathcal{B}} \otimes \hat{\rho}^{\bar{\mathcal{B}}}$	$\hat{\rho}_3 = \hat{U}^{rest} \hat{U}_3 \hat{\rho}_2 \hat{U}_3^\dagger \hat{U}^{rest\dagger}$	$\hat{U}_3 = \hat{U}^{\mathcal{R}_\eta \mathcal{A}_k} \otimes \hat{I}^{\mathcal{B} \bar{\mathcal{A}}_k \bar{\mathcal{B}}}$
c_4	$\hat{\rho}_3 = \left(\sum_{i=1}^4 p_i \hat{\rho}_i^{\mathcal{R}_\eta} \otimes \hat{\rho}_{2,i}^{\mathcal{A}_k} \right) \otimes \hat{\rho}^{\bar{\mathcal{A}}_k} \otimes \hat{\rho}^{\mathcal{B}} \otimes \hat{\rho}^{\bar{\mathcal{B}}}$	$\hat{\rho}_4 = \hat{U}_{diss}^{rest} \hat{U}_4 \hat{\rho}_3 \hat{U}_4^\dagger \hat{U}_{diss}^{rest\dagger}$	$\hat{U}_4 = \hat{U}^{\mathcal{R}_\eta \mathcal{A}_k} \otimes \hat{I}^{\mathcal{B} \bar{\mathcal{A}}_k \bar{\mathcal{B}}}$
c_5	$\hat{\rho}_4 = \left(\sum_{i=1}^4 p_i \hat{\rho}_i^{\mathcal{R}_\eta} \otimes \hat{\rho}_{3,i}^{\mathcal{A}_k} \right) \otimes \hat{\rho}^{\bar{\mathcal{A}}_k} \otimes \hat{\rho}^{\mathcal{B}} \otimes \hat{\rho}^{\bar{\mathcal{B}}}$	$\hat{\rho}_5 = \hat{U}^{rest} \hat{U}_5 \hat{\rho}_4 \hat{U}_5^\dagger \hat{U}^{rest\dagger}$	$\hat{U}_5 = \hat{U}^{\mathcal{R}_\eta \mathcal{A}_k} \otimes \hat{I}^{\mathcal{B} \bar{\mathcal{A}}_k \bar{\mathcal{B}}}$
c_6	$\hat{\rho}_5 = \left(\sum_{i=1}^4 p_i \hat{\rho}_i^{\mathcal{R}_\eta} \otimes \hat{\rho}_{4,i}^{\mathcal{A}_k} \right) \otimes \hat{\rho}^{\bar{\mathcal{A}}_k} \otimes \hat{\rho}^{\mathcal{B}} \otimes \hat{\rho}^{\bar{\mathcal{B}}}$	$\hat{\rho}_6 = \hat{U}_{diss}^{rest} \hat{U}_6 \hat{\rho}_5 \hat{U}_6^\dagger \hat{U}_{diss}^{rest\dagger}$	$\hat{U}_6 = \hat{U}^{\mathcal{R}_\eta \mathcal{A}_k \mathcal{B}} \otimes \hat{I}^{\bar{\mathcal{A}}_k \bar{\mathcal{B}}}$
c_7	$\hat{\rho}_6 = \left(\sum_{i=1}^4 p_i \hat{\rho}_i^{\mathcal{R}_\eta} \otimes \hat{\rho}_{5,i}^{\bar{\mathcal{B}}} \otimes \hat{\rho}_{5,i}^{\mathcal{A}_k} \right) \otimes \hat{\rho}^{\mathcal{B}} \otimes \hat{\rho}^{\bar{\mathcal{A}}_k}$	$\hat{\rho}_7 = \hat{U}^{rest} \hat{U}_7 \hat{\rho}_6 \hat{U}_7^\dagger \hat{U}^{rest\dagger}$	$\hat{U}_7 = \hat{U}^{\mathcal{R}_\eta \mathcal{A}_k} \otimes \hat{I}^{\mathcal{B} \bar{\mathcal{A}}_k \bar{\mathcal{B}}}$
c_8	$\hat{\rho}_7 = \left(\sum_{i=1}^4 p_i \hat{\rho}_i^{\mathcal{R}_\eta} \otimes \hat{\rho}_{6,i}^{\bar{\mathcal{B}}} \otimes \hat{\rho}_{6,i}^{\mathcal{A}_k} \right) \otimes \hat{\rho}^{\mathcal{B}} \otimes \hat{\rho}^{\bar{\mathcal{A}}_k}$	$\hat{\rho}_8 = \hat{U}_{diss}^{rest} \hat{U}_8 \hat{\rho}_7 \hat{U}_8^\dagger \hat{U}_{diss}^{rest\dagger}$	$\hat{U}_8 = \hat{U}^{\mathcal{R}_\eta \mathcal{A}_k} \otimes \hat{I}^{\mathcal{B} \bar{\mathcal{A}}_k \bar{\mathcal{B}}}$
c_9	$\hat{\rho}_8 = \left(\sum_{i=1}^4 p_i \hat{\rho}_i^{\mathcal{R}_\eta} \otimes \hat{\rho}_{7,i}^{\bar{\mathcal{B}}} \otimes \hat{\rho}_{7,i}^{\mathcal{A}_k} \right) \otimes \hat{\rho}^{\mathcal{B}} \otimes \hat{\rho}^{\bar{\mathcal{A}}_k}$	$\hat{\rho}_9 = \hat{U}^{rest} \hat{U}_9 \hat{\rho}_8 \hat{U}_9^\dagger \hat{U}^{rest\dagger}$	$\hat{U}_9 = \hat{U}^{\mathcal{R}_\eta \mathcal{A}_k \mathcal{B}} \otimes \hat{I}^{\bar{\mathcal{A}}_k \bar{\mathcal{B}}}$
c_{10}	$\hat{\rho}_9 = \left(\sum_{i=1}^4 p_i \hat{\rho}_i^{\mathcal{R}_\eta} \otimes \hat{\rho}_{8,i}^{\bar{\mathcal{B}}} \otimes \hat{\rho}_{8,i}^{\mathcal{A}_k} \right) \otimes \hat{\rho}^{\mathcal{B}} \otimes \hat{\rho}^{\bar{\mathcal{A}}_k}$	$\hat{\rho}_{10} = \hat{U}_{diss}^{rest} \hat{U}_{10} \hat{\rho}_9 \hat{U}_{10}^\dagger \hat{U}_{diss}^{rest\dagger}$	$\hat{U}_{10} = \hat{U}^{\mathcal{R}_\eta \mathcal{A}_k} \otimes \hat{I}^{\mathcal{B} \bar{\mathcal{A}}_k \bar{\mathcal{B}}}$
c_{11}	$\hat{\rho}_{10} = \left(\sum_{i=1}^4 p_i \hat{\rho}_i^{\mathcal{R}_\eta} \otimes \hat{\rho}_{9,i}^{\bar{\mathcal{B}}} \otimes \hat{\rho}_{9,i}^{\mathcal{A}_k} \right) \otimes \hat{\rho}^{\mathcal{B}} \otimes \hat{\rho}^{\bar{\mathcal{A}}_k}$	$\hat{\rho}_{11} = \hat{U}_{diss}^{rest} \hat{U}_{11} \hat{\rho}_{10} \hat{U}_{11}^\dagger \hat{U}_{diss}^{rest\dagger}$	$\hat{U}_{11} = \hat{U}^{\mathcal{A}_k \bar{\mathcal{A}}_k \mathcal{B}} \otimes \hat{I}^{\mathcal{R}_\eta \bar{\mathcal{B}}}$

Table 3.1. State transformations for the QCA 1-bit half adder operated under Landauer clocking.

The computationally active region of the half adder \mathcal{A}_k is correlated with the supporting computational system $\bar{\mathcal{A}}_k$ only after the inputs are loaded in the first computational step and before the outputs are erased in the last step.

The referent is correlated with \mathcal{A}_k alone from the end of c_1 until the end of c_4 , after which the loss of information to the bath correlates \mathcal{R}_η to \mathcal{B} and, after restoration steps, to $\bar{\mathcal{B}}$. At the end of the final step c_8 , the circuit has lost all information about the input referent and \mathcal{R}_η and \mathcal{A} are no longer correlated.

3.1.1.2 Analysis

We start our analysis by performing a spacetime decomposition of the circuit (operational decomposition), which is then followed by a physical- information-theoretic analysis of local dissipation into the bath throughout the computational cycle to capture and lower bound the dissipative costs.

Operational Decomposition – The clock cycle is a periodic sequence $\Phi = \varphi_1\varphi_2\varphi_3\varphi_4$ of four clock steps φ_v , each of which is an assignment of operations ϕ_t to the four independently controlled “clock zones” $C(u)$ depicted in green, magenta, blue, and red in Fig. 3.2. Denoting these clock zones as $C(1)$, $C(2)$, $C(3)$ and $C(4)$, respectively, the assignment corresponding to the adder clocking described in above is

$$\varphi_1 : \{(C(1); \phi_1), (C(2); \phi_4), (C(3); \phi_3), (C(4); \phi_2)\}$$

$$\varphi_2 : \{(C(1); \phi_2), (C(2); \phi_1), (C(3); \phi_4), (C(4); \phi_3)\}$$

$$\varphi_3 : \{(C(1); \phi_3), (C(2); \phi_2), (C(3); \phi_1), (C(4); \phi_4)\}$$

$$\varphi_4 : \{(C(1); \phi_4), (C(2); \phi_3), (C(3); \phi_2), (C(4); \phi_1)\}$$

The computational cycle Γ associated with a single input, which requires two full clock cycles, is then

$$\Gamma = c_1 c_2 c_3 c_4 c_5 c_6 c_7 c_8 c_9 c_{10} c_{11} = \varphi_1^{(1)} \varphi_2^{(1)} \varphi_3^{(1)} \varphi_4^{(1)} \varphi_1^{(2)} \varphi_2^{(2)} \varphi_3^{(2)} \varphi_4^{(2)} \varphi_1^{(3)} \varphi_2^{(3)} \varphi_3^{(3)}$$

where c_k denotes the k -th step in the computational cycle. Here c_1 is the initial LOAD step and c_{11} is the final erase step.

Each step of the computational cycle is necessarily a unitary evolution of the global system state, as must be the case for any isolated system evolving according to the time-dependent Schrödinger equation, i.e. according to physical law, governed by Hamiltonians that selectively couple various parts of \mathcal{A} and external subsystems to one another. At each computational step, the system starts out with \mathcal{A} in a computational state and \mathcal{B} in its nominal (thermal) state. The control processes evolves the system unitarily such that \mathcal{A} transitions to the next computational state specified by paradigmatic operation, with the state of \mathcal{B} adjusting as necessary to accommodate this computational state change within the global constraints imposed by physical law. The restoration process then returns \mathcal{B} to its nominal state through interaction with $\bar{\mathcal{B}}$, leaving \mathcal{A} in its new computational state.

To track information flow through the circuit, and to isolate the sources of irreversible information loss within the computational cycle, we refer back to section 2.3.2 and use the definition *data zone* $D(c_k)$ for each computational step c_k as follows: The k -th data zone is the union of all clock zones that, at the conclusion of computational step c_k , hold *some* information about the input associated with the cycle. There are ten data zones, $D(c_1) \dots D(c_{10})$ (since no information about the relevant input remains in the circuit at the conclusion of computational step c_{11}), each of which is the union of multiple disjoint *data subzones* $D_w(c_k)$. For the adder circuit of the present work, the data zones and subzones are

$$D(c_1) = D_1(c_1) \cup D_2(c_1)$$

$$D(c_2) = D(c_2)$$

$$D(c_3) = D_1(c_3) \cup D_2(c_3) \cup D_3(c_3)$$

$$D(c_4) = D_1(c_4) \cup D_2(c_4) \cup D_3(c_4)$$

$$D(c_5) = D(c_5)$$

$$D(c_6) = D_1(c_6) \cup D_2(c_6) \cup D_3(c_6)$$

$$D(c_7) = D_1(c_7) \cup D_2(c_7) \cup D_3(c_7) \cup D_4(c_7)$$

$$D(c_8) = D_1(c_8) \cup D_2(c_8)$$

$$D(c_9) = D_1(c_9) \cup D_2(c_9)$$

$$D(c_{10}) = D_1(c_{10}) \cup D_2(c_{10})$$

where each data subzone corresponds to clock subzones³ as

$$D_1(c_1) = C_1(1)$$

$$D_2(c_1) = C_2(1)$$

$$D_1(c_2) = C_1(1) \cup C_1(2) \cup C_2(2) \cup C_2(1) \cup C_3(2)$$

$$D_1(c_3) = C_1(2) \cup C_1(3)$$

$$D_2(c_3) = C_2(2) \cup C_2(3)$$

$$D_3(c_3) = C_3(2) \cup C_3(3)$$

$$D_1(c_4) = C_1(3) \cup C_1(4)$$

$$D_2(c_4) = C_2(3) \cup C_2(4)$$

³Numbered from top to bottom and then from left to right.

$$D_3(c_4) = C_3(3) \cup C_3(4)$$

$$D_1(c_5) = C_1(4) \cup C_5(1) \cup C_2(4) \cup C_2(5) \cup C_3(4) \cup C_3(5)$$

$$D_1(c_6) = C_1(5) \cup C_1(6)$$

$$D_2(c_6) = C_2(6) \cup C_2(5)$$

$$D_3(c_6) = C_3(5) \cup C_4(6)$$

$$D_1(c_7) = C_1(6) \cup C_1(7)$$

$$D_2(c_7) = C_2(7) \cup C_2(6) \cup C_3(7)$$

$$D_3(c_7) = C_3(6) \cup C_4(7)$$

$$D_1(c_8) = C_1(7) \cup C_1(8) \cup C_2(7)$$

$$D_2(c_8) = C_3(7) \cup C_2(8) \cup C_4(7)$$

$$D_1(c_9) = C_1(8) \cup C_1(9)$$

$$D_2(c_9) = C_2(8) \cup C_2(9)$$

$$D_1(c_{10}) = C_1(10)$$

$$D_2(c_{10}) = C_2(10).$$

The data zones exhibit considerable splitting and merging throughout the computational cycle.

Cost Analysis – The dissipative cost per computational cycle is the sum of contributions from each step

$$\Delta\langle E \rangle_{TOT} = \sum_{k=1}^{11} \Delta E(c_k) = \sum_{k=1}^{11} \Delta\langle E^{\mathcal{B}} \rangle_k \quad (3.2)$$

where $\Delta\langle E^{\mathcal{B}} \rangle_k$ is the change in the expected energy of the bath \mathcal{B} during the control phase of the k -th computational step c_k . A lower bound on this quantity is obtained

by summing single-step lower bounds on $\Delta\langle E^{\mathcal{B}}\rangle_k$ obtained from physical-information-theoretic analyses. These single-step bounds, and thus the full-cycle bound, derive exclusively from global unitary dynamics of the coupled subsystems, entropic inequalities, fundamental thermodynamic considerations assuming paradigmatic operation of the abstracted circuit as described in Chapter 2 (see Eq. (2.1)).

Nonzero lower bounds on $\Delta\langle E^{\mathcal{B}}\rangle_k$ are obtained for computational steps in which data-bearing blocks of cells relax to null states *in the absence of* interaction with other blocks of cells that hold a *complete* record of the data initially encoded in the relaxing cells. To formalize this, note that the clock subzones that are nullified during computational step c_k are those common to data zones $D(c_{k-2})$ and $D(c_{k-1})$. Assuming that there is one such clock subzone for each data subzone $D_w(c_{k-2})$, as is the case for the QCA adder considered here, these clock subzones can be identified, denoted, and indexed as $\mathcal{C}_w^{(k)} = D_w(c_{k-2}) \cap D(c_{k-1})$. Noting also that the union of “successor” subzones with which $\mathcal{C}_w^{(k)}$ interacts as it is nullified during step c_k is $\mathcal{D}_w^{(k)} = \bigcup_{w'} \mathcal{D}_{w'w}^{(k)}$, where $\mathcal{D}_{w'w}^{(k)} \in \{\mathcal{D}_{w'w}^{(k)}\} = \{D_{w'}(c_{k-1}) | D_{w'}(c_{k-1}) \cap \mathcal{C}_w^{(k)} \neq \emptyset\}$. Specifically, some information is irreversibly lost during computational step c_k if, at the conclusion of c_k , the initial states of clock subzones $\mathcal{C}_w^{(k)}$ cannot uniquely inferred from the final states of the successor subzones $\mathcal{D}_w^{(k)}$ with which they interacted during c_k . This allows us to treat each $\mathcal{C}_w^{(k)} \cup \mathcal{D}_w^{(k)}$ as an independent L -machine (denoted $\mathcal{L}_w^{(k)} \equiv \mathcal{C}_w^{(k)} \cup \mathcal{D}_w^{(k)}$). We assume that the $\mathcal{C}_w^{(k)}$ interact locally with the bath, since by definition they do not interact with one another, the dissipative cost of information loss from subzones in a given computational step is additive and the net dissipative cost of erasure for all subzones erased during the k -th computational step can be lower bounded as [9]

$$\Delta\langle E^{\mathcal{B}}\rangle_k \geq \sum_{\mathcal{L}_w^{(k)}} -k_B T \ln(2) \Delta \mathcal{I}^{\mathcal{R}_\eta \mathcal{L}_w^{(k)}} \quad (3.3)$$

where k_B is the Boltzmann constant, $-\Delta\mathcal{I}^{\mathcal{R}_\eta\mathcal{L}_w^{(k)}}$ is the amount of information about the initial state of clock subzone $\mathcal{C}_w^{(k)}$ that is locally and irreversibly lost from $\mathcal{C}_w^{(k)} \cup \mathcal{D}_w^{(k)}$ during the k -th computational step. In specializing the bound presented above in Eq. (1.38) for quantum mechanical L -machines to this scenario, we have taken each $\mathcal{C}_w^{(k)} \cup \mathcal{D}_w^{(k)}$ to be an ideal classical L -machine and regarded the two data states and the null state of each subzone to be pure (and thus zero entropy).

The single-step dissipation bound (3.3) is nonzero only for computational steps c_4 , c_7 and c_{10} , arising from information loss in the clock subzones $\{\mathcal{C}_w^{(4)}\} = \{C_2(2)\}$, $\{\mathcal{C}_w^{(7)}\} = \{C_3(1), C_4(1)\}$, and $\{\mathcal{C}_w^{(10)}\} = \{C_4(2), C_5(2)\}$. Evaluation of the corresponding bounds for equiprobable adder inputs yields.⁴

$$\Delta\langle E^{\mathcal{B}} \rangle_4 \geq 1.19k_B T \ln(2) \quad (3.4)$$

$$\Delta\langle E^{\mathcal{B}} \rangle_7 \geq 1.38k_B T \ln(2) \quad (3.5)$$

$$\Delta\langle E^{\mathcal{B}} \rangle_{10} \geq 1.19k_B T \ln(2). \quad (3.6)$$

These results are summarized in Fig. 3.4, which shows the lower bound on the *cumulative* dissipative cost of processing one input, averaged over the four possible two-bit inputs (assumed equiprobable), for the eleven steps of the computational cycle under Landauer clocking.

Adding the single-step costs via Eq. (3.3), we obtain the fundamental lower bound on the local dissipative cost of computation for one computational cycle of the Landauer-clocked QCA half adder:

$$\Delta\langle E \rangle_{TOT} \geq (3.76)k_B T \ln(2). \quad (3.7)$$

⁴Note that the bound $\Delta\langle E^{\mathcal{B}} \rangle_{10}$ accounts for the fact that the subzones $\mathcal{C}_w^{(10)} = \{C_4(2), C_5(2)\}$ are nullified in interaction with the output registers S and C , respectively, which hold partial copies of the information erased from these subzones.

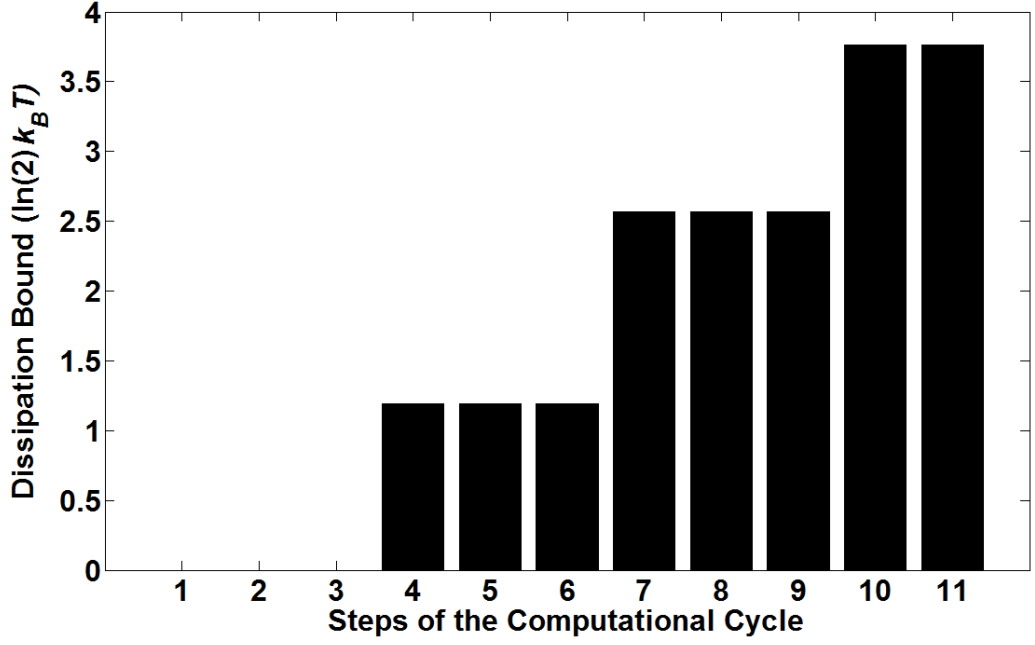


Figure 3.4. Fundamental lower bound on the cumulative dissipative cost for one computational cycle of the Landauer-clocking of QCA 1-bit half adder circuit.

This full-cycle bound is about $7.5\times$ the $0.5k_B T \ln(2)$ lower bound for any half adder that erases its input, with the excess resulting from local irreversibility associated with the specific circuit structure and clocking scheme. We emphasize that the localization and quantification of information loss that enable determination of this dissipation bound requires the abstractions and decompositions used in this analysis: A half adder with equiprobable inputs necessarily loses 0.5 bits of information taken as a whole, whereas the 135 cells in the QCA adder lose a total of 114 bits when taken individually. The intermediate value of 3.76 bits follows from our detailed analysis of the structure and operation of this particular circuit. A detailed discussion on the granularity of heat dissipation is presented in Appendix A.

To give the bound (4.3) a quantitative meaning in a computational setting, we assume a clock (cycle) frequency of 50 GHz (55 ps computational cycle), a cell footprint of 10 nm^2 (circuit area of $5.7 \times 10^{-11} \text{ cm}^2$), and operation at $T = 300K$. For this

case, the bound (4.3) implies an areal power dissipation of no less than $3.45\text{W}/\text{cm}^2$ per computational cycle.⁵ Since data is pipelined through the circuit, and portions of other computational cycles are active during the same time period, the lower bound on the *total* energy dissipated during the eleven clock steps taken to implement one computational cycle is $(8.9)k_B T \ln(2)$,⁶ and the lower bound on the corresponding areal power dissipation for the above parameters and conditions is $8.14\text{ W}/\text{cm}^2$.

3.1.2 QCA Half Adder with Bennett Clocking

We now sketch application of our approach to operation of the QCA adder using Bennett clocking, which proposed by Lent and co-workers [31] as a way to implement reversible computation in QCA circuits and improve power efficiency. In principle, Bennett clocking allows for dissipation-free single-shot computations by retaining the input data in the circuit throughout the computation. Dissipation-free *sequences* of computations are also possible, but only if special efforts are made to “unload” the input data at the end of each cycle [32]. Dissipative costs do, however, arise in more practical Bennett clocking scenarios, including large circuits with independent Bennett-clocking of individual stages to enable pipelining, where input data is erased from each stage at the end of its Bennett-clocked cycle. A scheme for pipelining data through multiple Bennett-clocked stages of a QCA circuit has recently been proposed by Ottavi and co-workers [33], who investigated trade-offs between stage depth, computational throughput, and the dissipative cost of erasing intermediate results of each stage. Here, we consider a relatively simpler Bennett clocking scheme.

⁵We calculate this by using $P = \frac{\Delta E}{A \cdot t}$, where ΔE is the amount of energy dissipated, A is the circuit area, and t is the amount of time it takes for one full computational cycle ($2\frac{3}{4}$ clock cycles). At $T = 300\text{K}$ the constant $k_B T \ln(2)$ is 2.869×10^{-21} Joules, therefore $P = \frac{3.76 \cdot 2.869 \times 10^{-21}}{5.7 \times 10^{-11} \cdot 55 \times 10^{-12}}$.

⁶The three dissipative steps associated with the processing of the η^{th} input overlaps with the last two and first two of the dissipative steps associated with the $(\eta + 1)^{th}$ and $(\eta - 1)^{th}$ input, respectively. Therefore, dissipative costs outlined in Eq. (3.4), (3.5), and (3.6) appear two-, three-, and two-times in this calculation, respectively.

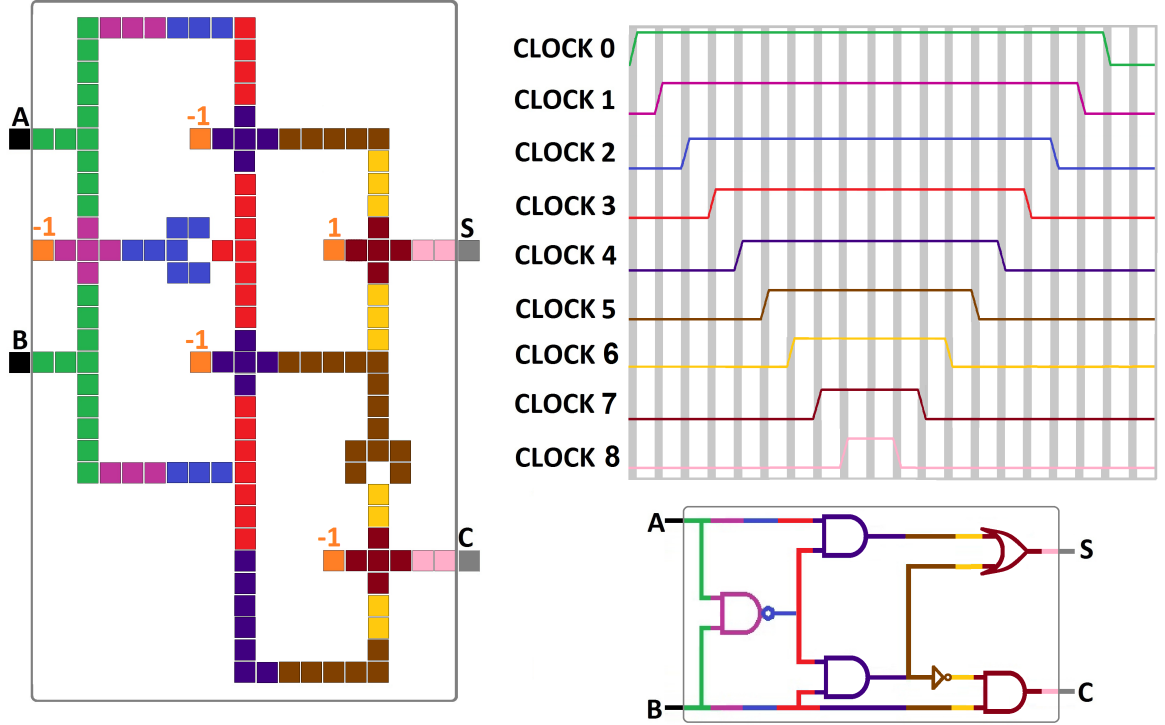


Figure 3.5. Layout, clocking, and gate-level representation for Bennett clocking of the QCA 1-bit half adder.

The scheme for Bennett clocking of the QCA adder considered here enables pipelining and is depicted in Fig. 3.5. The nine clock waveforms and corresponding zones for Bennett clocking are indicated in green, magenta, blue, red, navy, and brown, yellow, maroon, pink, respectively. The external input and output registers are shown in black and gray since they are not simultaneously operated with clock zones listed above. The shaded clocking regions correspond to the twenty steps of one computational cycle. There are nine clock zones and nineteen clock steps ϕ_v (grey bars), each corresponding to a unique computational step c_k of the nineteen-step computational cycle. Input data is loaded from the external register AB in step c_1 , the result is computed during steps $c_2 - c_7$. During these steps, the data remains intact behind the forward propagating clock wave, the result is read out to external register CS in step c_9 , the computation is reversibly “undone” during steps $c_8 - c_{19}$.

This leaves the original input data for the cycle in the leftmost (green) clock zone, with the remainder of the adder in its null state, which is then irreversibly erased in step c_{19} in the *absence* of a copy assuming that register AB holds data for the next cycle at this point. This would also be the case in a simple pipelining scheme. This is the only step in the computational cycle where information is irreversibly lost and dissipation costs are necessarily incurred. For instance, recent work on molecular QCA suggests that dissipation from switching will dominate dissipation costs, far exceeding parasitic clocking losses [34].

For equiprobable inputs, the dissipative cost of each computational cycle is that of erasing the two bits of input information, which is lower bounded simply as

$$\Delta\langle E\rangle_{TOT} \geq (2)k_B T \ln(2). \quad (3.8)$$

Here the coefficient two comes from the two bits of input information that is unavoidably discarded at the end of a full computational cycle. We can explain this in terms of the general bound (2.1); only the final computational step contributes to the total bound –when the circuit loses all the information at once and completely the end of the computational cycle. The final fundamental bound for the QCA half adder operated under Bennett clocking does not have any connection to dissipation from the data zones or subzones we identified for the same circuit operated under Landauer clocking. The fundamental lower bound on the dissipative cost per computational cycle for Bennett clocking is smaller than the corresponding bound for Landauer clocking by a factor of 0.53 ($= 2/3.76$). This is despite the fact that the Bennett cycle requires twenty computational steps compared to eleven for Landauer clocking. For a fixed computational throughput, with the Bennett clocked circuit operated at a rate 4.75 ($= 19/4$) times faster than the Landauer clocked circuit to compensate for the increased latency inherent in Bennett clocking, the fundamental lower bound

on areal power dissipation for Bennett clocking is less than the corresponding bound for Landauer clocking.

This comparative study, which is specific to one particular circuit controlled via Landauer clocking scheme and Bennett clocking scheme, demonstrates application of our approach to concrete scenarios in the QCA nanocomputing paradigm. The result of this study provides fundamental support for the conclusion that Bennett clocking can provide power efficiency (operations per Watt [35]) advantages in multi-stage, pipelined circuits for a given computational throughput. The resulting power dissipation bounds with a Bennett clocking rate increases as necessary (4.75x the Landauer rate) to achieve the same computational throughput in the two schemes. For the particular pipelining granularities considered, we found the lower bound on power dissipation for Landauer clocking to exceed that for Bennett clocking.

3.2 Transistor-Based Applications

In this section, we expand our methodology to address the fundamental bounds on energy costs for circuits that exchange not only energy but also particles with their surroundings. This enables us to apply our approach to transistor-based complex circuit structures. We employ 2D grid single nanowire type NASIC (Nanoscale Application Specific Integrated Circuit) proposed by Moritz and co-workers ([11] and references therein) for illustrative purposes. We use both half and full adder implementations of the fabric. We also apply our methodology to a dynamically clocked CMOS circuit to illustrate our methodology’s application to a conventional paradigm. Lastly, we discuss the limitations involved in applying our methodology to static CMOS circuits.

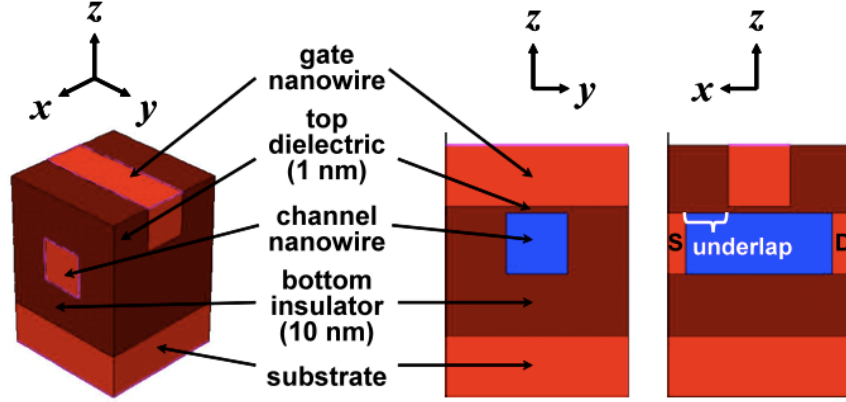


Figure 3.6. The cross section of a crossed-nanowire Field Effect Transistor (xnwFET) [11].

3.2.1 Nano-Application Specific Integrated Circuits (NASICs)

NASIC is a promising CMOS-replacement technology that has been extensively researched by multiple groups for various applications. The architecture relies on a new hybrid CMOS-nanoscale circuit style which is shown to have higher performance, as well as less manufacturing cost by using single type Field Effect Transistor in nanoscale portions [36]. The fabric is built from 2-D semiconductor nanowire grids with crossed-nanowire FET (xnwFET) at certain crosspoints. The channel of an xnwFET is aligned along one nanowire while the perpendicular nanowire above it acts as a gate [11], [37]. The cross section of an xnwFET⁷ is shown in Fig. 3.6. In xnwFET devices, the gate, source, drain, and substrate regions are all doped n-type (high in electron concentration), and the channel is doped p-type. This device is an inversion mode device similar to conventional MOSFETs: applying a positive voltage at the gate terminal attracts electrons into the p-doped channel leading to n-type FET behavior.⁸ NASICs use a dynamic circuit style with control signals driven

⁷The Fig. 3.6 displays a rectangular nanowire, however, depending on the performance target as well as the manufacturing and synthesis process used cylindrical nanowires too are possible [11].

⁸The “inversion mode” corresponds to inversion layer-like behavior similar to that of conventional CMOS. Inversion layer in CMOS is a channel through which electron can pass. Further details, as

from external reliable CMOS circuitry. This emerging technology is one of the most well-developed post-CMOS circuits and its analysis is tractable within our approach.

The physical circuit structure of the NASIC half adder along with the associated timing and logic diagram used in this study is depicted in Fig. 3.7. The nanowire grid is represented by thin blue lines and the peripheral microwires carrying V_{DD} , V_{SS} and dynamic control signals (PRE1, EVA1, PRE2, EVA2) are depicted as thick yellow, green and blue lines, respectively. Channels of the crossed-nanowire FETs, shown at certain cross points, are oriented horizontally in Stage 1 (left NAND plane) and vertically in Stage 2 (right NAND plane). Inputs are received from vertical nanowires in Stage 1, which act as gates for horizontal crossed-nanowire FETs implementing the first stage of the dynamic circuit. The horizontal nanowires act as the outputs of Stage 1 and as gates of the Stage 2 transistors, whose channels are aligned in the vertical direction. Independent latching of the gate inputs and outputs in separate transistor stages allows for pipelined operation of the adder. Note here that the control signals coordinate the flow of data through NASIC tiles. The horizontal and vertical signals are different which supporting cascading. Fig. 3.7 (middle) shows NASIC control scheme considered in this study. A logic diagram of the half adder implemented by this circuit is also shown in Fig. 3.7 (bottom). The transistors in Stage 1 register the input values for the first level of NAND gates, whereas the Stage 2 transistors register the outputs of these NAND gates (hereafter the “minterm complements”), which are, of course, the inputs to the second level of NAND gates.

We trace the processing of a single adder input through one full computational cycle.⁹ Only one of the control signals is active at any time in this scheme, affecting

well as the difference between the operation of n-type and p-type devices will be explained under the discussion of CMOS paradigm in Sec. 3.2.2. Here, we focus on the single type FET used in the NASIC paradigm, and provide information on its structure and operation.

⁹A variety of clocking schemes are possible for NASIC circuits: the one we consider here (*cf.* Fig. 3.7) is of Ref. [11] which is designed for single wire type circuits in the original paper.

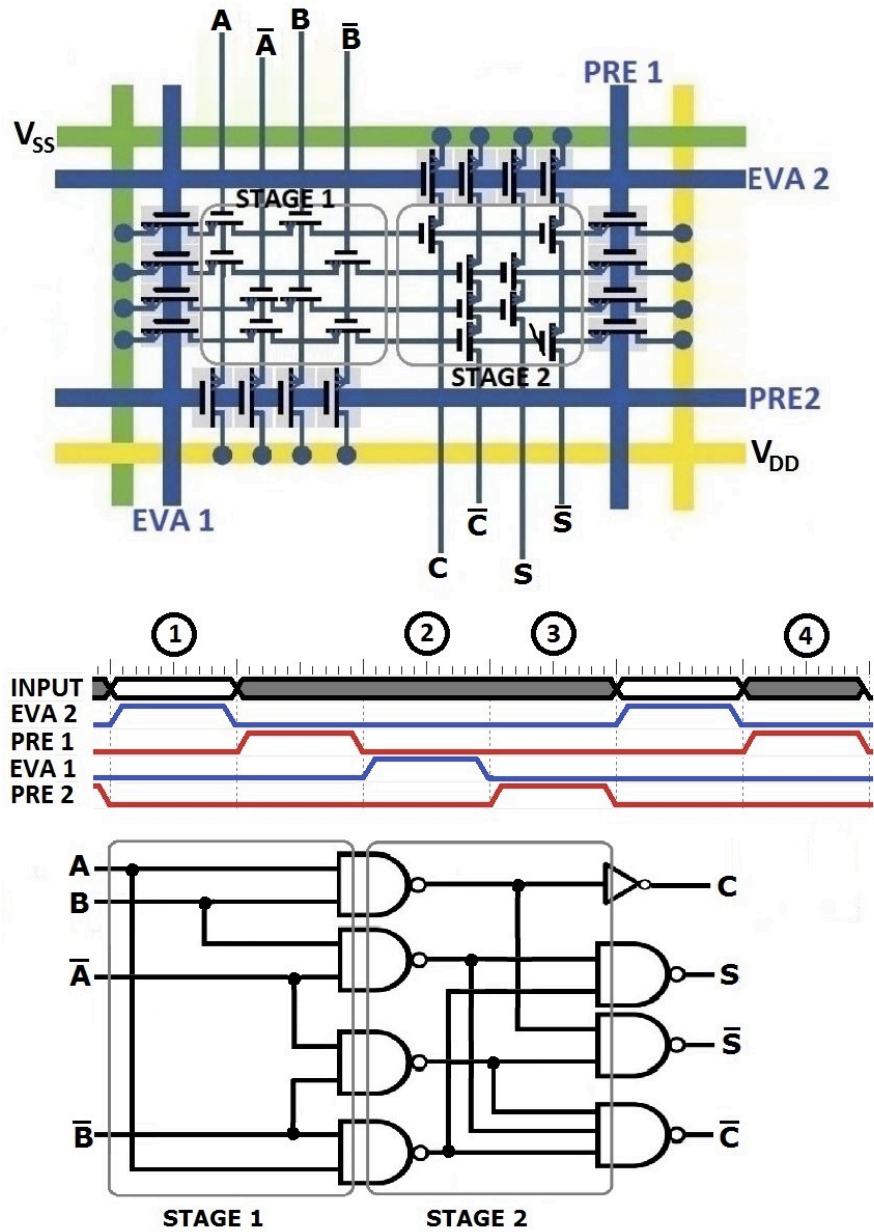


Figure 3.7. Layout, clocking, and logic diagram of the single-FET-type NAND-NAND NASIC 1-bit half adder circuit.

either the Stage 1 transistors *or* the Stage 2 transistors but not both. Ideally, PRE2 and PRE1 *unconditionally* precharge the transistors in Stages 1 and 2, respectively. This erases any information held in the transistors of these circuit stages. During the evaluation steps, EVA2 *conditionally* discharges the transistors in Stage 1, conditioned on the adder input values, and EVA1 discharges the transistors in Stage 2 conditioned on the states of the Stage 1 transistors. From a functional point of view, transistor states are the repositories for bit values in NASIC circuits.

We now construct the physical decomposition and process abstraction of the circuit scheme explained for NASIC half adder and introduce our cost analysis. We present the final fundamental lower bounds for both circuits in sections 3.2.1.3 and 3.2.1.4, respectively.

3.2.1.1 Abstraction

We begin by constructing the abstraction of the circuit and other surrounding subsystems to capture the essential functional features of the underlying computational strategy of NASIC adders. We do this within a physical description that is compatible with physical-information-theoretic analysis. This allows us to obtain the dissipation bounds we are after.

Physical Decomposition – The abstraction used to describe the NASIC adders and its environment is depicted in Fig. 3.8. The artifact \mathcal{A} is the NASIC tile, or, more specifically, the system of electrons in the nanowire grid and at the surface of the underlying substrate (collectively \mathcal{C}) together with the source (V_{SS}) and drain (V_{DD}) microwires (\mathcal{S} and \mathcal{D}). The source and drain are nominally regarded as idealized Fermi gases at temperature T with associated chemical potentials μ_{SS} and μ_{DD} , respectively, with $\Delta\mu = \mu_{SS} - \mu_{DD} = qV_{DD}$, where q is the electric charge. The bath \mathcal{B} is (the phonon gas in) the underlying substrate in direct thermal contact with the NASIC tile, and is nominally in a thermal state at temperature T . The

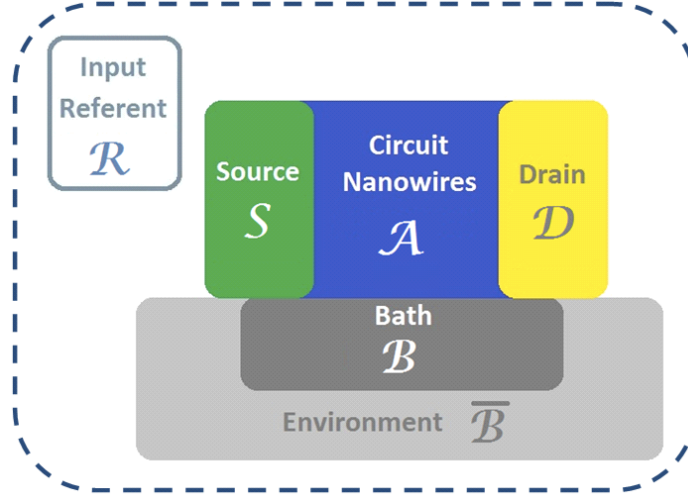


Figure 3.8. Physical abstraction of the NASIC 1-bit adder situated in its surrounding environment.

greater environment $\bar{\mathcal{B}}$ includes subsystems that drive \mathcal{B} toward thermal equilibrium and supply the energy and particles required to maintain the nominal populations of \mathcal{S} and \mathcal{D} and a chemical potential difference $\Delta\mu = qV_{DD}$ when these subsystems are driven from their nominal states during computation. The nanowire grid exchanges particles with the source and drain microwires \mathcal{S} and \mathcal{D} and heat with the bath \mathcal{B} as it processes input data held in register (referent) \mathcal{R} . The greater environment $\bar{\mathcal{B}}$ provides the energy, particles, and heat removal that enable circuit operation, and everything else required to thermodynamically isolate the global system.

The interaction between the subsystems can be summarized as follows. The precharge and evaluate operations selectively open the nanowire grid to particle exchange with the drain \mathcal{D} and source \mathcal{S} , respectively, with \mathcal{S} open to substrate electrons at all times. The subsystems \mathcal{C} , \mathcal{S} , and \mathcal{D} can also exchange heat with \mathcal{B} during each computational step. Heat exchange between $\bar{\mathcal{B}}$ and the subsystems \mathcal{B} , \mathcal{S} and \mathcal{D} , and particle exchange between $\bar{\mathcal{B}}$ and \mathcal{S} and \mathcal{D} , are assumed to restore \mathcal{B} , \mathcal{S} and \mathcal{D} to their nominal states at the conclusion of each computational step.

Process Abstraction – We associate physical operations ϕ with each clock step, decomposing each operation into a control process and a restoration process. Both processes are necessarily unitary evolutions of the global system state, as must be the case for any isolated system evolving according to the time-dependent Schrödinger equation (i.e. according to the physical law) each involving selected subsystem interactions. At each step, the system starts out with \mathcal{C} in a computational state and \mathcal{S} , \mathcal{D} , and \mathcal{B} in their nominal states. The control process evolves the system unitarily such that \mathcal{C} transitions to the next computational state specified by paradigmatic operation, exchanging precisely the minimum number of particles required for the operation (if any) with \mathcal{S} and \mathcal{D} and adjusting the state of \mathcal{B} as necessary to accommodate the computational state change while obeying physical law. The restoration process returns \mathcal{S} , \mathcal{D} , and \mathcal{B} to their nominal states through interaction with $\bar{\mathcal{B}}$ and leaves \mathcal{C} in its new computational state.

Similar to the QCA half adder, we outline the sequence of state transformations that comprise a single computational cycle of the NASIC. The initial state of the NASIC half adder circuit and the surrounding systems is defined as

$$\hat{\rho}^{initial} = [\hat{\rho}^{\mathcal{C}_k} \otimes \hat{\rho}^{\bar{\mathcal{C}}_k} \otimes \hat{\rho}^{\mathcal{S}} \otimes \hat{\rho}^{\mathcal{B}} \otimes \hat{\rho}^{\mathcal{D}} \otimes \hat{\rho}^{\bar{\mathcal{B}}}],$$

where subsystem density operators are separable. The reduced density operators are $\hat{\rho}^{\mathcal{R}\mathcal{A}} = Tr_{\mathcal{S}\mathcal{B}\mathcal{D}}[\hat{\rho}]$, $\hat{\rho}^{\mathcal{S}} = Tr_{\mathcal{R}\mathcal{A}\mathcal{B}\mathcal{D}}[\hat{\rho}]$, $\hat{\rho}^{\mathcal{B}} = Tr_{\mathcal{R}\mathcal{A}\mathcal{S}\mathcal{D}}[\hat{\rho}]$ and $\hat{\rho}^{\mathcal{D}} = Tr_{\mathcal{R}\mathcal{A}\mathcal{S}\mathcal{B}}[\hat{\rho}]$. Initially, all nanowire FETs in the circuit are precharged (set to their logic 1 states), the bath \mathcal{B} is in a thermal state, and the referent state for the η -th input is represented as defined in Eq. (3.1).

On each step, the global system evolves unitarily via the time-dependent Schrödinger equation. Table 3.2. outlines the initial states and the structure of the state transformation associated with each computational step, with \mathcal{C}_k indicating the clock zone

that changes state during computational step c_k . The state of \mathcal{C}_k is determined by the pattern of discharged transistors in the computationally relevant circuit stage.

The control operation of every computational step is followed by a restoration phase. For evaluation steps, the restoration operator is $\hat{U}_{EVA}^{rest} = \hat{U}^{\mathcal{B}\bar{\mathcal{B}}} \otimes \hat{I}^{\mathcal{R}_\eta \mathcal{C}_k \mathcal{S} \mathcal{D} \bar{\mathcal{C}}_k}$, which rethermalizes the bath and \mathcal{S} . The restoration operation for precharge steps, which is of the form $\hat{U}_{PRE}^{rest} = \hat{U}^{\mathcal{B} \mathcal{S} \mathcal{D} \bar{\mathcal{B}}} \otimes \hat{I}^{\mathcal{R}_\eta \mathcal{C}_k \bar{\mathcal{C}}_k}$, rethermalizes \mathcal{B} and recharges and rethermalizes \mathcal{S} and \mathcal{D} .

As was the case for the QCA adder, we consider single shot computation to obtain a lower bound on the amount of dissipation associated with processing a single input. The computationally relevant stage \mathcal{C}_k of the half adder is correlated with the supporting computational system $\bar{\mathcal{C}}_k$ only when the inputs are loaded in the first computational step and when the outputs are obtained in the last step. Starting from c_2 the greater environment $\bar{\mathcal{B}}$ is correlated with \mathcal{R}_η due to the restoration of \mathcal{B} after evaluation. At the end of c_4 , the final step of computation, the circuit loses all the information about the η -th input and \mathcal{C} is no longer correlated with \mathcal{R}_η .

Three classes of clock operations defined for the NASIC adder are precharge operations ϕ_P , evaluate operations ϕ_E , and hold ϕ_H . Each of these operations act on a specified part of the circuit in any given clock step. During precharge ϕ_P , part of the nanowire grid is selectively opened to \mathcal{D} and electrons flow from \mathcal{S} into \mathcal{C} and from \mathcal{C} into \mathcal{D} . A subsequent “reinvestment” of energy is required by $\bar{\mathcal{B}}$ during the restoration to recharge \mathcal{S} and \mathcal{D} . During evaluation ϕ_E , part of the nanowire grid is selectively opened to \mathcal{S} , which is in contact with substrate electrons at all times, allowing discharging via electron exchange *within* the artifact (*through* \mathcal{S}) at no cost to the greater environment. During hold operations ϕ_H , part of the nanowire grid remains isolated from \mathcal{S} and \mathcal{D} by keeping past transistors OFF.

Computational Steps	Initial State	State Transformation	Control Operation
c_1	$\hat{\rho}_0 = \left(\sum_{i=1}^{M=4} p_i \hat{\rho}_i^{\mathcal{R}_\eta} \otimes \hat{\rho}_i^{\bar{\mathcal{C}}_k} \right) \otimes \hat{\rho}^{\mathcal{C}_k} \otimes \hat{\rho}^{\mathcal{S}} \otimes \hat{\rho}^{\mathcal{D}} \otimes \hat{\rho}^{\mathcal{B}} \otimes \hat{\rho}^{\bar{\mathcal{B}}}$	$\hat{\rho}_1 = \hat{U}_{EVA}^{rest} \hat{U}_1 \hat{\rho}_0 \hat{U}_1^\dagger \hat{U}_{EVA}^{rest\dagger}$	$\hat{U}_1 = \hat{U}^{\mathcal{R}_\eta \mathcal{C}_k \bar{\mathcal{C}}_k \mathcal{B}} \otimes \hat{I}^{SD\bar{\mathcal{B}}}$
c_2	$\hat{\rho}_1 = \left(\sum_{i=1}^4 p_i \hat{\rho}_i^{\mathcal{R}_\eta} \otimes \hat{\rho}_{1,i}^{\mathcal{C}_k} \otimes \hat{\rho}_{1,i}^{\bar{\mathcal{B}}_k} \right) \otimes \hat{\rho}^{\mathcal{S}} \otimes \hat{\rho}^{\mathcal{D}} \otimes \hat{\rho}^{\mathcal{B}} \otimes \hat{\rho}^{\bar{\mathcal{C}}_k}$	$\hat{\rho}_2 = \hat{U}_{EVA}^{rest} \hat{U}_1 \hat{\rho}_1 \hat{U}_1^\dagger \hat{U}_{EVA}^{rest\dagger}$	$\hat{U}_2 = \hat{U}^{\mathcal{R}_\eta \mathcal{C}_k \mathcal{B}} \otimes \hat{I}^{SD\bar{\mathcal{C}}_k \bar{\mathcal{B}}}$
c_3	$\hat{\rho}_2 = \left(\sum_{i=1}^4 p_i \hat{\rho}_i^{\mathcal{R}_\eta} \otimes \hat{\rho}_{2,i}^{\mathcal{C}_k} \otimes \hat{\rho}_{2,i}^{\bar{\mathcal{B}}_k} \right) \otimes \hat{\rho}^{\mathcal{S}} \otimes \hat{\rho}^{\mathcal{D}} \otimes \hat{\rho}^{\mathcal{B}} \otimes \hat{\rho}^{\bar{\mathcal{C}}_k}$	$\hat{\rho}_3 = \hat{U}_{PRE}^{rest} \hat{U}_3 \hat{\rho}_2 \hat{U}_3^\dagger \hat{U}_{PRE}^{rest\dagger}$	$\hat{U}_3 = \hat{U}^{\mathcal{R}_\eta \mathcal{C}_k SD\mathcal{B}} \otimes \hat{I}^{\bar{\mathcal{C}}_k \bar{\mathcal{B}}}$
c_4	$\hat{\rho}_3 = \left(\sum_{i=1}^4 p_i \hat{\rho}_i^{\mathcal{R}_\eta} \otimes \hat{\rho}_{3,i}^{\mathcal{C}_k} \otimes \hat{\rho}_{3,i}^{\bar{\mathcal{B}}_k} \right) \otimes \hat{\rho}^{\mathcal{S}} \otimes \hat{\rho}^{\mathcal{D}} \otimes \hat{\rho}^{\mathcal{B}} \otimes \hat{\rho}^{\bar{\mathcal{C}}_k}$	$\hat{\rho}_4 = \hat{U}_{PRE}^{rest} \hat{U}_4 \hat{\rho}_3 \hat{U}_4^\dagger \hat{U}_{PRE}^{rest\dagger}$	$\hat{U}_4 = \hat{U}^{\mathcal{C}_k SD\mathcal{B} \bar{\mathcal{C}}_k} \otimes \hat{I}^{\mathcal{R}_\eta \bar{\mathcal{B}}}$

Table 3.2. State transformations for the NAND-NAND NASIC 1-bit adder.

The crucial feature of this decomposition and abstraction, which ultimately enables determination of fundamental dissipation bounds within our approach, is this: Any local information about the initial state of \mathcal{C} that is irreversibly lost during a clock step induces dissipation in \mathcal{B} before being swept completely into $\bar{\mathcal{B}}$ during the restoration process.¹⁰ This locally irreversible information loss affects the state of \mathcal{B} during the control process, which is precisely the point at which the dissipation costs are cashed out in our approach. All dissipation costs resulting from logical irreversibility *and* particle supply can be captured by “monitoring” energy flow into the bath during control operations. It is important to note here that the cost we capture by this approach is subject to several key assumptions regarding the relative time scales of the various physical processes involved. These assumptions are concerned with the time it takes for the information in the contacts to get dissipated into the bath as well as the interactions between the other subsystems. Specifically, we assume that the information in the contacts get dissipated into the bath before the next computational operation and rethermalization of the subsystems involved.

3.2.1.2 Analysis

In order to properly capture and lower bound the dissipative costs, we perform a spacetime decomposition of the circuit (operational decomposition) and a physical-information-theoretic analysis of local dissipation into the bath throughout the computational cycle.

Operational Decomposition – The clock cycle $\Phi = \varphi_1\varphi_2\varphi_3\varphi_4$ is a periodic sequence of four clock phases φ_v , each of which is an assignment of operations ϕ to the two independently controlled clock zones identified as Stage 1 and Stage 2 of Fig. 3.7. The duration of the full computation requires six clock steps, however, only four among

¹⁰Loss of initial-state information from part of \mathcal{C} is locally irreversible if it is erased in the absence of interaction with other parts of \mathcal{C} or $\bar{\mathcal{C}}$ that hold or receive copies of the initial state during the clock step.

the six steps is associated with the computation cycle of a given input. Only the steps labeled as ①, ②, ③ and ④ involve manipulation of data related to a given input within the tile.

Denoting these clock zones as $C(1)$ and $C(2)$, respectively, the assignment corresponding to the adder clocking described above is

$$\varphi_1 : \{(C(1); \phi_E), (C(2); \phi_H)\}$$

$$\varphi_2 : \{(C(1); \phi_H), (C(2); \phi_P)\}$$

$$\varphi_3 : \{(C(1); \phi_H), (C(2); \phi_E)\}$$

$$\varphi_4 : \{(C(1); \phi_P), (C(2); \phi_H)\}.$$

The computational cycle $\Gamma^{(\eta)}$ for the η -th input is then the sequence

$$\Gamma^{(\eta)} = c_1 c_2 c_3 c_4 = \varphi_1^{(1)} \varphi_3^{(1)} \varphi_4^{(1)} \varphi_2^{(2)} \quad (3.9)$$

of four computational steps c_k associated with this cycle, which are drawn from the first six steps of two consecutive clock cycles

$$\Phi^{(1)} \Phi^{(2)} = \varphi_1^{(1)} \varphi_2^{(1)} \varphi_3^{(1)} \varphi_4^{(1)} \varphi_1^{(2)} \varphi_2^{(2)} \varphi_3^{(2)} \varphi_4^{(2)} \quad (3.10)$$

as explained above. Only dissipative costs incurred during the four computational steps contribute to the total dissipative cost of processing the η -th input in the NASIC adder tile.

To track information flow through the circuit, and to isolate the sources of irreversible information loss within the computational cycle, we define a *data zone* $D(c_k)$ for each computational step c_k . The k -th data zone is defined as the union of all clock zones that, at the conclusion of computational step c_k , hold *some* information about the η -th input. Each data zone $D(c_k)$ is, in general, the union of multiple clock zones. For the computational cycle of the NASIC adder, we have

$$D(c_1) = C(1)$$

$$D(c_2) = D_1(c_2) \cup D_2(c_2) = C(1) \cup C(2)$$

$$D(c_3) = C(2)$$

$$D(c_4) = \emptyset.$$

Cost Analysis – The total dissipative cost associated with one computational cycle is

$$\Delta \langle E \rangle_{TOT} = \sum_{k=1}^4 \Delta E(c_k) = \sum_{k=1}^4 \Delta \langle E^{\mathcal{B}} \rangle_k \quad (3.11)$$

where $\Delta \langle E^{\mathcal{B}} \rangle_k$ is the change in the expected energy of the bath \mathcal{B} during the k -th computational step c_k . At most one clock zone in the data zone $D(c_{k-1})$ changes its state during step c_k , and the dissipative cost can be lower bounded as [9]

$$\Delta \langle E^{\mathcal{B}} \rangle_k \geq -k_B T \ln(2) \left(\Delta \mathcal{I}^{\mathcal{R} \mathcal{C}_k \mathcal{SD}} + \langle \Delta S_i^{\mathcal{C}_k \mathcal{SD}} \rangle \right) \quad (3.12)$$

where \mathcal{C}_k is the clock zone that changes state during the k -th computational step ($\mathcal{C}_1 = \mathcal{C}_3 = C(1)$, $\mathcal{C}_2 = \mathcal{C}_4 = C(2)$), $\Delta \mathcal{I}^{\mathcal{R} \mathcal{C}_k \mathcal{SD}}$ is the information about \mathcal{R} that is lost from $\mathcal{C}_k \mathcal{SD}$ during the k -th step and $\langle \Delta S_i^{\mathcal{C}_k \mathcal{SD}} \rangle$ is input-averaged entropy change of $\mathcal{C}_k \mathcal{SD}$ for the k -th step. Because, for all steps, the initial and final states of \mathcal{C}_k and the number ΔN_k of particles transferred (from \mathcal{S} to \mathcal{C}_k and \mathcal{C}_k to \mathcal{S}) are precisely specified for all steps under paradigmatic operation, and because \mathcal{S} and \mathcal{D} are in thermal states at the beginning and end of the control process for each step, the bound (3.12) simplifies to

$$\Delta \langle E^{\mathcal{B}} \rangle_k \geq -k_B T \ln(2) \left(\Delta \mathcal{I}^{\mathcal{R} \mathcal{C}_k} + \Delta \langle S^{\mathcal{S}} \rangle_k + \Delta \langle S^{\mathcal{D}} \rangle_k \right). \quad (3.13)$$

We use this general bound in the steps where the artifact is charged and information is irreversibly erased (i.e.; the precharge steps) to analyze the cost. In the evaluation steps, we can simply use the energy conservation relation

$$\Delta \langle E^B \rangle_k = -\Delta \langle E_i^{C_k} \rangle \quad (3.14)$$

since no information is erased, the initial and final states of \mathcal{S} are the same, and \mathcal{D} is isolated during this process.

Since clock zone $C(1)$ changes state only during computational steps c_1 and c_3 ($\mathcal{C}_1 = \mathcal{C}_3 = C(1)$), and since clock zone $C(2)$ changes state only during computational steps c_2 and c_4 ($\mathcal{C}_2 = \mathcal{C}_4 = C(2)$), we can isolate contributions from each clock zone (or stage) to the full dissipation bound:

$$\Delta \langle E^B \rangle_1 + \Delta \langle E^B \rangle_3 \geq -\Delta \langle E_i^{C_1} \rangle - k_B T \ln(2) \left(\Delta \mathcal{I}^{\mathcal{R}_\eta C_3} + \Delta \langle S_i^{C_3} \rangle + \Delta \langle S_i^{\mathcal{S}} \rangle_3 + \Delta \langle S_i^{\mathcal{D}} \rangle_3 \right) \quad (3.15)$$

$$\Delta \langle E^B \rangle_2 + \Delta \langle E^B \rangle_4 \geq -\Delta \langle E_i^{C_2} \rangle - k_B T \ln(2) \left(\Delta \mathcal{I}^{\mathcal{R}_\eta C_4} + \Delta \langle S_i^{C_4} \rangle + \Delta \langle S_i^{\mathcal{S}} \rangle_4 + \Delta \langle S_i^{\mathcal{D}} \rangle_4 \right) \quad (3.16)$$

These bounds can be simplified greatly through two minimal assumptions that recognize the nature of the NASIC circuit operation, as we now illustrate in detail for Stage 1. First, since selected transistors in Stage 1 (i.e. subsystem \mathcal{C}_1) simply discharge (through \mathcal{S}) in the evaluation step c_1 , and since this process involves only \mathcal{C}_1 , \mathcal{B} , and \mathcal{S} (with the initial and final states of \mathcal{S} unchanged), electrostatic energy capacitively stored in \mathcal{C}_1 is necessarily dissipated into the bath in Step c_1 . A minimal assumption on the energy loss from \mathcal{C}_1 is thus that it is nonnegative:

$$-\Delta \langle E_i^{C_1} \rangle \geq 0. \quad (3.17)$$

Second, we assume that the interactions between \mathcal{C}_1 and \mathcal{B} are such that, for initial charge configurations of \mathcal{C}_1 that encode *every* input x_i , precharging of \mathcal{C}_1 can never

decrease the entropy of \mathcal{B} . This recognizes that a nonnegative amount of energy is necessarily transferred *to* the surroundings during capacitive charging, and requires that this energy is transferred in the form of heat. We thus require that, for all i

$$\Delta S_i^{\mathcal{B}} \geq 0 \quad (3.18)$$

in computational step c_3 . Noting that (i) Step 3 is a unitary evolution of $\mathcal{C}_3\mathcal{SDB}$, (ii) von Neumann entropy is preserved under unitary evolution, (iii) the initial and final states of $\mathcal{C}_3\mathcal{SDB}$ for every i , and (iv) the entropy is additive for composite systems in separable states, we also have

$$\Delta S_i^{\mathcal{C}_3} + \Delta S_i^{\mathcal{S}} + \Delta S_i^{\mathcal{D}} + \Delta S_i^{\mathcal{B}} = 0 \quad (3.19)$$

Thus

$$\Delta S_i^{\mathcal{B}} = -\Delta S_i^{\mathcal{C}_3} - \Delta S_i^{\mathcal{S}} - \Delta S_i^{\mathcal{D}} \geq 0 \quad (3.20)$$

for step c_3 , and consequently

$$-\langle \Delta S_i^{\mathcal{C}_3} \rangle_3 - \langle \Delta S_i^{\mathcal{S}} \rangle_3 - \langle \Delta S_i^{\mathcal{D}} \rangle_3 \geq 0 \quad (3.21)$$

for the input averages.

Applying (3.12) and (3.14) to (3.3), and analyzing Stage 2 along precisely the same lines, we have

$$\Delta \langle E^{\mathcal{B}} \rangle_1 + \Delta \langle E^{\mathcal{B}} \rangle_3 \geq -k_B T \ln(2) \Delta \mathcal{I}^{\mathcal{R}_\eta \mathcal{C}_3} \quad (3.22)$$

$$\Delta \langle E^{\mathcal{B}} \rangle_2 + \Delta \langle E^{\mathcal{B}} \rangle_4 \geq -k_B T \ln(2) \Delta \mathcal{I}^{\mathcal{R}_\eta \mathcal{C}_4}. \quad (3.23)$$

Combining these results, we finally have the full-cycle bound with the total energy dissipated

$$\Delta \langle E^{\mathcal{B}} \rangle_{TOT} = \sum_{k=1}^4 \Delta \langle E^{\mathcal{B}} \rangle_k \geq -k_B T \ln(2) \left(\Delta \mathcal{I}^{\mathcal{R}_\eta \mathcal{C}_3} + \Delta \mathcal{I}^{\mathcal{R}_\eta \mathcal{C}_4} \right). \quad (3.24)$$

This reflects only the costs of irreversible information erasure, can be tightened so it includes additional costs related to particle supply by the source-drain subsystem \mathcal{SD} . To achieve this, we model \mathcal{SD} as a capacitor - initially charged to a voltage V_{DD} – that partially discharges into the NASIC circuit during the precharge phases. For a discharge of magnitude $q\Delta N$, i.e. a transfer of ΔN electrons to \mathcal{D} from the circuit and ΔN electrons to the circuit from \mathcal{S} , the change in the amount of energy stored in \mathcal{SD} is given by

$$\Delta \langle E^{\mathcal{SD}} \rangle = -qV_{DD}\Delta N \quad (3.25)$$

provided that ΔN is small compared to the total electron populations of \mathcal{S} and \mathcal{D} when it is charged to V_{DD} . The electron transfer ΔN associated with any given precharge step depends on the number of FETs that are charged in the step and the amount of charging required to change the state of each FET.

Now, part of the energy $\Delta \langle E^{\mathcal{SD}} \rangle$ transferred out of \mathcal{SD} during precharge steps is stored in the artifact and part of this energy is irreversibly dissipated into the bath. The fraction f of this energy that is stored in the artifact during precharge is ultimately dissipated into the bath during subsequent evaluation steps, so its dissipative contribution can be cashed out during the evaluation stages; the energy lost from the first circuit stage (during EVA2) is thus $\Delta \langle E_i^{\mathcal{C}_1} \rangle = -fqV_{DD}\Delta N_1$ and the energy lost from the second circuit stage (during EVA1) is $\Delta \langle E_i^{\mathcal{C}_2} \rangle = -fqV_{DD}\Delta N_2$. Note here that the total number of electron transfer for a NASIC circuit operated under this clocking scheme is $\Delta N_{TOT} = \Delta N_1 + \Delta N_2$. Using these expressions in place of the weaker conditions

$$\Delta \langle E^{\mathcal{B}} \rangle_{TOT} \geq -k_B T \ln(2) \Delta \mathcal{I}^{\mathcal{R}^{TOT}} + f q V_{DD} \Delta N_{TOT}. \quad (3.26)$$

The general term, ΔN corresponds to the number of transistor times the electron transfer required to switch each FET, Δn . Since Δn cannot be less than unity, $\Delta n \geq 1$, the lower bound on $\Delta \langle E^{\mathcal{B}} \rangle_{TOT}$ can be obtained in terms of Δn . Note that substitution of $\Delta n = 1$ into the above bound simply ensures that the resulting expression is a fundamental lower bound on the dissipation; it does not amount to a claim that $\Delta n = 1$ is technologically achievable in this context. Below, we obtain the numerical form of this bound by considering the irreversible information erasure and associated particle cost in both half and full adder NASIC circuits.

3.2.1.3 NASIC Half Adder

We now present the fundamental lower bounds on heat dissipation for the NASIC half adder circuit introduced above. We refer back to Fig. 3.7 for the physical circuit structure, timing and logic diagrams. We consider pipelined operation of the adder as it processes a sequence of inputs $\dots X^{(\eta+1)} X^{(\eta)} X^{(\eta+1)} \dots$ (where $X \in \{X_i\} = \{AB\} = \{00, 01 \dots 11\}$). Let the η -th computational cycle begin with EVA2, when the η -th input $X^{(\eta)}$ is loaded into the Stage 1 transistors. Next, in PRE1, the minterm complements left over from the $(\eta - 1)^{th}$ input are erased from Stage 2 without affecting Stage 1. Next, in EVA1, the minterm complements for $X^{(\eta)}$ are evaluated and latched into Stage 2. Next, in PRE2, Stage 1 is precharged and $X^{(\eta)}$ is erased from this stage. Next, in EVA2, the complemented minterms held in Stage 2 are preserved as the sum and carry outputs for input $X^{(\eta)}$ are evaluated and latched into the (implied) “downstream” circuit stage while input $X^{(\eta+1)}$ is loaded into Stage 1 from the “upstream” tile. Finally, PRE1 precharges Stage 2 and erases the minterm complements for input $X^{(\eta)}$. This completes the computational cycle for the η^{th} input in the adder tile.

A	B	\bar{A}	\bar{B}	D_0	D_1	D_2	D_3	C	S	\bar{C}	\bar{S}
0	0	1	1	0	1	1	1	0	0	1	1
0	1	1	0	1	0	1	1	0	1	1	0
1	0	0	1	1	1	0	1	0	1	1	0
1	1	0	0	1	1	1	0	1	0	0	1

Table 3.3. The truth table of the NAND-NAND NASIC 1-bit half adder.

Although the duration of the full computation requires six clock steps, we associate only four of these steps NASICse labeled as ①, ②, ③ and ④ on the timing diagram in Fig. 3.7 – with the computational cycle for the η -th input, since only these steps involve manipulation of data related to the η -th input within the tile. (The remaining clock steps – the first PRE1 and the second EVA2 – belong to computational cycles for inputs $X^{(\eta-1)}$ and $X^{(\eta+1)}$, respectively.)

The truth table associated with the NASIC half adder is presented in Table. 3.3. Computational steps c_1 and c_2 are both evaluation steps, during which charge redistributes within \mathcal{C} without altering the particle numbers in \mathcal{S} or \mathcal{D} and information about the input $X^{(\eta)}$ (i.e. about \mathcal{R}_η) is transferred within \mathcal{C} without irreversible information loss from $C(1)$ or $C(2)$. Thus, $\Delta N_1 = \Delta N_2 = 0$ and $\Delta \mathcal{I}^{\mathcal{R}_\eta \mathcal{C}_1} = \Delta \mathcal{I}^{\mathcal{R}_\eta \mathcal{C}_2} = 0$, so $\Delta \langle E^{\mathcal{B}} \rangle_1 = \Delta \langle E^{\mathcal{B}} \rangle_2 = 0$. The fundamental lower bound on the dissipative cost of the first two computational steps thus vanishes. Computational steps c_3 and c_4 are precharge steps. During c_3 , half of the twenty four transistors in $C(1)$ are precharged (at one electron per transistor) and all of the information about input $X^{(\eta)}$ is irreversibly lost from $C(1)$, yielding $\Delta N_3 = 4$ electrons and $\Delta \mathcal{I}^{\mathcal{R}_\eta \mathcal{C}_3} = -2$ bits (assuming equiprobable inputs). During c_4 , two of the sixteen transistors in $C(2)$ are precharged and all of the information about input $X^{(\eta)}$ is irreversibly lost from $C(2)$, yielding $\Delta N_4 = 2$ electrons and $\Delta \mathcal{I}^{\mathcal{R}_\eta \mathcal{C}_4} = -2$ bits.

We add the single-step costs as shown in the bound (3.11). Contributions to this bound from information loss and particle supply at each computational step are

Computational Step	Particle Supply	Information Loss
c_1 : EVA 2	0	0
c_2 : EVA 1	0	0
c_3 : PRE 2	$4qV_{DD}$	$2k_B T \ln(2)$
c_4 : PRE 1	$2qV_{DD}$	$2k_B T \ln(2)$
Cycle Total	$6qV_{DD}$	$4k_B T \ln(2)$

Table 3.4. Dissipation bound for the NASIC 1-bit half adder: Particle supply and information loss components

tabulated in Table 3.4., and the cumulative cost is shown for one computational cycle in Fig. 3.9 assuming $T = 300K$, $V_{DD} = 0.8$ V, and equiprobable inputs.¹¹

The dissipative cost of processing a single input through a full computational cycle of the NASIC half adder, controlled via the clocking scheme described above, is lower bounded as

$$\Delta \langle E^{\mathcal{B}} \rangle_{TOT} \geq 4k_B T \ln(2) + f6qV_{DD}, \quad (3.27)$$

where $\Delta \langle E^{\mathcal{B}} \rangle_{TOT}$ is the total amount of heat dissipated into the circuit's surroundings during one computational cycle. Note here that we assume Δn , the number of electrons required to switch each FET, to be 1 in order to obtain the lowest bound. This bound is obtained from fundamental physical considerations for the NASIC half adder working under paradigmatic operation with no parasitic losses and the most optimistic possible assumption for the amount of charge required to switch each transistor's state. Therefore the cost reflected in the bound reflect the unavoidable physical cost of computation inherent in the underlying computational strategy employed by the NASIC adder. Significantly, the cost of local information loss associated with irreversible information processing (4 bits) and the supply of (particles dropping

¹¹For the values considered here, the total dissipation from the information loss is $4k_B T \ln(2) = 0.072eV$ and the total dissipation from the working substance is $6qV_{DD} = 4.8eV$. Also note that, the coefficient in front of the particle supply cost is taken to be $f = 1$ while calculating the total dissipation for this graph.

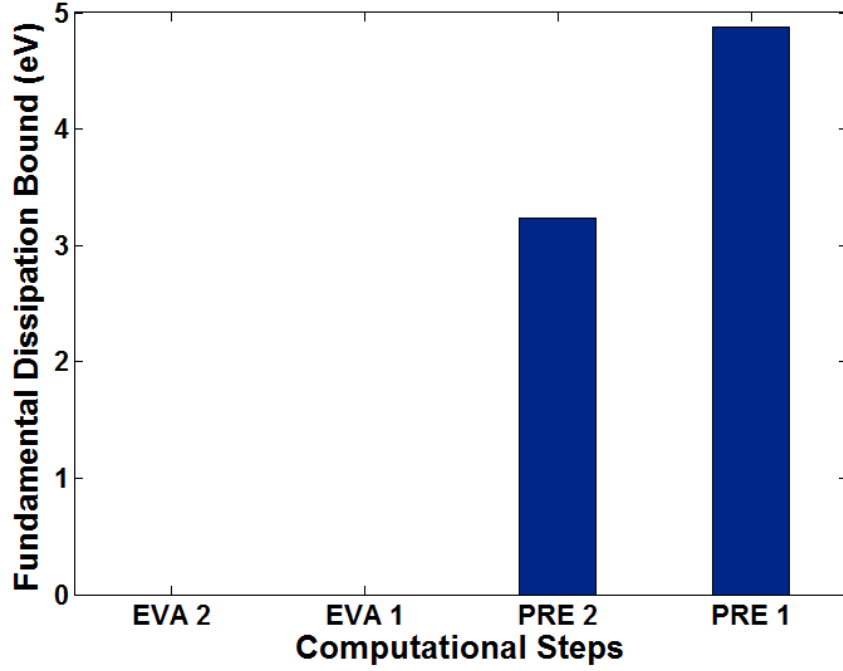


Figure 3.9. Fundamental lower bound on the cumulative dissipation cost for one computational cycle of the NASIC 1-bit half adder.

through potential difference of V_{DD}) maintaining the computational working substance (6 electrons) are distinctly reflected in the two separate components of the fundamental bound.

Note that the bound is independent of assumptions regarding material, device, or circuit dimensions or parameters. Even in this idealized scenario – where the particle supply requirements are at their absolute minimum – particle-supply accounts for 99% of the lower bound on the dissipative cost, far exceeding that required for implementation of the irreversible computation. The dominance of the particle-supply cost in the bound can be reduced but not eliminated by reducing V_{DD} , since the directional circuit/microwire electron injection and extraction processes that enable proper circuit operation and reliable computation require that $V_{DD} \gg k_B T$.

In the next section, we present a similar inequality for NASIC full adder which operates based on same principles in a more complicated structure. There, we also

A	B	C^{in}	\overline{A}	\overline{B}	$\overline{C^{in}}$	S	C^{out}	$\overline{S_i}$	$\overline{C^{out}}$
0	0	0	1	1	1	0	0	1	1
0	1	0	1	0	1	1	0	0	1
1	0	0	0	1	1	1	0	0	1
0	0	1	1	1	0	1	0	0	1
0	1	1	1	0	0	0	1	1	0
1	0	1	0	1	0	0	1	1	0
1	1	0	0	0	1	0	1	1	0
1	1	1	0	0	0	1	1	0	0

Table 3.5. The truth table of the NAND-NAND NASIC 1-bit full adder.

discuss the terms on the right hand side of the inequality and compare them to provide a better understanding of the fundamental lower bound.

3.2.1.4 NASIC Full Adder

We consider a single-FET-type NAND-NAND NASIC full adder shown in Fig. 3.10 (top), which is adapted from Ref's [11] and [36]. Based on the circuit details and the application of our methodology to the NASIC half adder presented in the previous section, we now extend our analysis to a full adder. The physical circuit structure, associated timing and logic diagrams are presented in Fig. 3.10. The truth table associated with the full adder is also presented in Table 3.5.

Computational steps c_1 and c_2 are both evaluation steps, during which charge redistributes within \mathcal{C} without altering the particle numbers in \mathcal{S} or \mathcal{D} and information about the input¹² $X^{(\eta)}$ (i.e. about \mathcal{R}) is transferred within \mathcal{C} without irreversible information loss from $C(1)$ or $C(2)$. Thus, $\Delta N_1 = \Delta N_2 = 0$ and $\Delta \mathcal{I}^{\mathcal{R}C_1} = \Delta \mathcal{I}^{\mathcal{R}C_2} = 0$, so $\Delta \langle E^{\mathcal{B}} \rangle_1 = \Delta \langle E^{\mathcal{B}} \rangle_2 = 0$. The fundamental lower bound on the dissipative cost of the first two computational steps thus vanishes. Computational steps c_3 and c_4 are precharge steps. During c_3 , half of the twenty four transistors in $C(1)$ are

¹²Note here that, for the full adder circuit the inputs correspond to $X \in \{X_i\} = \{ABC^{in}\} = \{000, 001, \dots, 111\}$ as listed in Table 3.5.

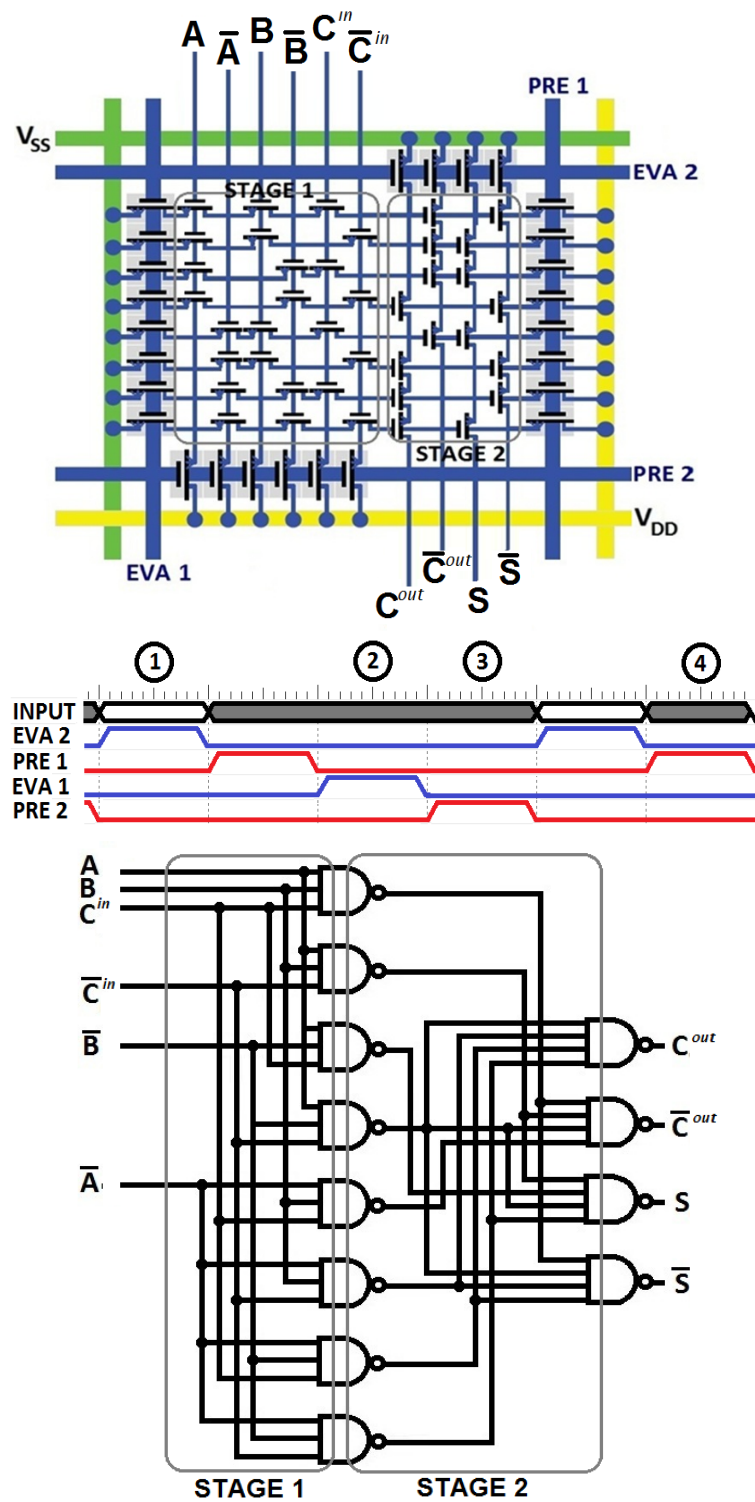


Figure 3.10. Layout, clocking, and logic diagram of the single-FET-type NAND-NAND NASIC 1-bit full adder circuit.

precharged (at one electron per transistor) and all of the information about input $X^{(\eta)}$ is irreversibly lost from $C(1)$, yielding $\Delta N_3 = 12$ electrons and $\Delta \mathcal{I}^{\mathcal{R}C_3} = -3$ bits (assuming equiprobable inputs). During c_4 , two of the sixteen transistors in $C(2)$ are precharged and all of the information about input $X^{(\eta)}$ is irreversibly lost from $C(2)$, yielding $\Delta N_4 = 2$ electrons and $\Delta \mathcal{I}^{\mathcal{R}C_4} = -3$ bits.

Similarly, we add the single-step costs to obtain the total bound on the dissipative cost. Contributions to this bound from information loss and particle supply at each computational step are tabulated in Table 3.6., and the cumulative cost is shown for one computational cycle in Fig. 3.11 assuming $T = 300K$, $V_{DD} = 0.8$ V, and equiprobable inputs.

Computational Step	Particle Supply	Information Loss
c_1 : EVA 2	0	0
c_2 : EVA 1	0	0
c_3 : PRE 2	$12qV_{DD}$	$3k_B T \ln(2)$
c_4 : PRE 1	$2qV_{DD}$	$3k_B T \ln(2)$
Cycle Total	$14qV_{DD}$	$6k_B T \ln(2)$

Table 3.6. Dissipation bound for the NASIC 1-bit full adder: Particle supply and information loss components

The dissipative cost of processing a single input through a full computational cycle of the NASIC full adder, controlled via the clocking scheme described above, is lower bounded as

$$\Delta \langle E \rangle_{TOT} \geq 6k_B T \ln(2) + f14qV_{DD}. \quad (3.28)$$

Similar to the fundamental bound for the half adder, this inequality too has the form of Eq. (3.26). The first term represents the cost of local information loss associated with irreversible information processing (6 bits) and the second term represent the cost of the supply of maintaining the computational working substance (14 electrons),

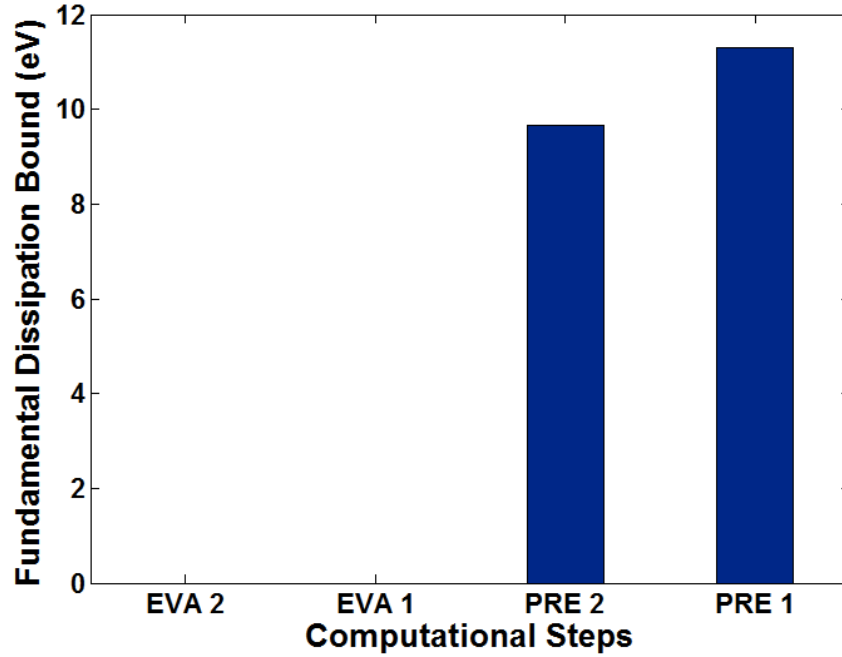


Figure 3.11. Fundamental lower bound on the cumulative dissipation cost for one computational cycle of the NASIC 1-bit full adder.

which are, as mentioned before, distinctly different components of the fundamental bound.

We substitute the constant values considered for the half adder case¹³ and get $6k_B T \ln(2) = 0.108$ eV and $14qV_{DD} = 11.2$ eV – see Fig. 3.11.¹⁴ For quantitative perspective, note that the bound $\Delta\langle E \rangle_{TOT} \geq 11.3$ eV on the net cost per computational cycle implies a theoretical minimum power density of $P_{TOT} \geq 6.8 \text{ W/cm}^2$ for an average (nanowire and control) transistor density of 10^{11} cm^{-2} and clock speed of 10 GHz.

In order to acquire further perspective, we can compare this result with dissipation in a full adder circuit level and NAND gate level. A full adder with equiprobable

¹³For $T = 300\text{K}$ and $V_{DD} = 0.8$ V. We also assume that one electron $\Delta n = 1$, is required to switch each transistor's state

¹⁴The coefficient in front of the particle supply cost is taken to be $f = 1$ while calculating the total dissipation for this graph.

inputs necessarily loses 1.19 bits of information from input to output, while the net local information loss associated with the 11 NAND gates in the logic circuit of Fig. 3.7 (with all gate input probabilities reflecting equiprobable adder inputs) is 34.3 bits. The 6 bit loss identified here for the NASIC implementation, and the simple additivity of the associated logical irreversibility cost ($6k_B T$) and particle supply cost ($14qV_{DD}$) observed in the bound, *result* from our detailed physical analysis of the circuit and its operation.

Finally, we can also generalize the bound (3.26) for any number of inputs and outputs. Any combinational function with M input bits and N output bits can be implemented using a NASIC logic block with general structure employed in the half adder, with $2M$ input nanowires (for M logical inputs and their complements) $2N$ output wires (for $2N$ logical outputs and their complements) and $2^M(M+N)$ crossed-nanowire FETs appropriately placed at crosspoints. Application of our methodology to such a logic block yields the dissipation bound

$$\Delta E_{TOT} \geq (2M) k_B T \ln(2) + f q V_{DD} [2^{M-1} M + N] \quad (3.29)$$

with the assumption of uniformly distributed input. This simple bound holds for arbitrary M and N , and for any M -input, N output Boolean function.

3.2.1.5 Comparison with Physical Circuit Simulations

The theoretical *lower bounds* on dissipative costs that we present here and *calculated values* for energy consumption based on explicit device models are of a fundamentally different nature. However, comparing trends in the fundamental bounds with those observed in analogous results from physical circuit simulations may provide us further insight into the overhead costs incurred in real implementations. In this subsection we present the results we have obtained on the cumulative energy delivered by the power source, throughout one computational cycle, from HSPICE simulations

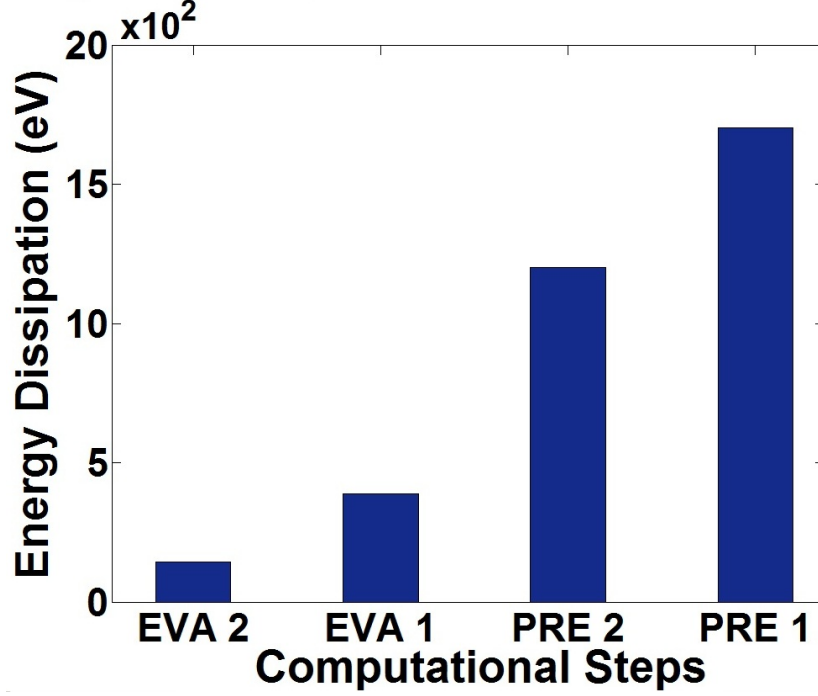


Figure 3.12. Input-averaged cumulative energy consumption for one computational cycle of the NASIC 1-bit full adder obtained from HSPICE simulations.

of a nanowire full adder with layout, clocking, and operation as described in Fig. 3.7 and in above discussion. This simulation provides us a basis for comparison.

The simulated adder uses the circuit style of Fig. 3.7, but uses a grid of paired Si nanowires (pitch 22 nm). Each nanowire pair consists of two 5-nm-diameter nanowires with an 11 nm pitch. Ref.'s [11] and [37] provide more information on NASIC HSPICE simulations, and additional details of the device model used can be found in Ref. [?]. Circuit operation at $T = 300K$, a supply voltage of $V_{DD} = 0.8$ V, and clock frequency of 10 GHz were assumed for the simulations. It is important to note here that the computational cycles associated with successive inputs overlap one another as described above, therefore, we have taken special measures, such as use of an independent V_{DD} source to charge the circuit output nodes, into consideration in order to isolate the cost of processing a single input throughout a single computational cycle.

Results of HSPICE simulations for the cumulative, input-averaged energy dissipation at each step of the computational cycle are shown in Fig. 3.12. Note that the values shown in Fig. 3.12 are actually *double* averages, since, in the simulated circuit, the dissipative cost associated with processing of one input in one computational cycle may depend on the binary values of the input being processed in that cycle *and* the input processed in the previous cycle. The input sequences used in the simulations were constructed so these dependences are captured in the simulation results, but effects of capacitive memory in the circuit that extend beyond one previous input are not reflected. The cumulative dissipative cost shows qualitative similarities to the fundamental bound of Fig. 3.11, with the largest steps in the cumulative cost occurring in the third and fourth computational steps (where Stage 2 and Stage 1 are precharged, respectively). Conspicuous differences include the nonzero calculated energy costs for the first two phases, which can only result from parasitic leakage during evaluation, and the orders-of-magnitude discrepancy between the calculated energy dissipation and the fundamental bound in the last two phases (expected from charge transfers far exceeding one electron per transistor in the simulated circuit). In any case, the qualitative similarities that do exist in the results of Fig. 3.11 and Fig. 3.12 suggest that the physical abstraction used to obtain the fundamental bound captures essential features of the NASIC circuit operation, while the differences highlight what is missed.

3.2.2 A Dynamically clocked np-CMOS Half Adder

The circuits we studied so far are prominent examples of post-CMOS nanoelectronic technology proposals. We obtained the fundamental lower bounds on energetic cost of computation for these potential CMOS replacement technologies, and our results illustrate the factors affecting the bounds such as the clocking scheme and thermodynamic processes taking place as a result of information loss. Even though

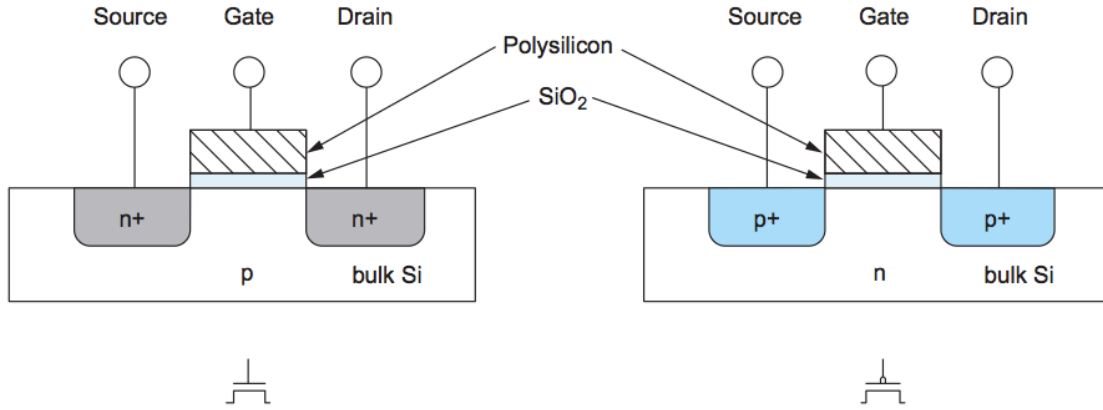


Figure 3.13. Cross section of an nMOS (left) and a pMOS (right) transistor [39].

these emerging proposal are well-developed and studied extensively, they are not as well-know and familiar as the conventional CMOS technology. In order to provide further insight into the application of our methodology, we now employ a CMOS circuit adapted from widely used textbooks [38], [39]. We calculate the fundamental lower bound for a dynamically clocked np-CMOS half adder as a pedagogic illustrative example. Below, we first provide a brief introduction on the CMOS circuits and NP domino logic operation, and then move on the np-CMOS half adder circuit example we employed to illustrate our methodology via a conventional paradigm.

Metal oxide semiconductor field effect transistors (MOSFETs) are composed of n-type or p-type semiconductors, i.e. semiconductors with high electron and hole concentration, respectively. Transistors are built on a silicon semiconductor substrate. Pure silicon has no free charge carriers and conducts electricity poorly. However, by adding dopants the conductivity of silicon can be increased. CMOS technology employs complementary and symmetrical pairs of negatively and positively doped silicon semiconductors in order to implement logic functions. This complementary n-type and p-type pairs are referred to as nMOS and pMOS, respectively. MOS structure is created by superimposing various layers of conducting and insulating materials. Fig. 3.13 shows the cross section of an nMOS and a pMOS transistor. The

insulating oxide layer between the bulk and the metal gate allows MOS structure to function as a capacitor. The gate voltage allows for the formation of a conductive channel below the oxide layer that allows charge transport from source to drain. The three terminals of a transistor structure is the base (bulk of the semiconductor where conducting channel forms), source (where the electrons come from) and drain (where the electrons go to).

The NASIC half and full adder circuits studied in Sec. 3.2.1.3 and 3.2.1.4, respectively, are all composed of xnwFET transistor structure that is operated similar to nMOS transistors. As we have seen in the NASIC examples, an nMOS transistor is ON when there is a positive charge on the transistor gate, i.e. the switch is closed to build a conducting path between the source and drain. The positive charge on the nMOS transistor is selectively neutralized during evaluation via the path from the source to the gate. During precharge, an nMOS gate is connected to the drain and if the gate is neutral, negative charge flows into the drain, leaving positive charge on the nMOS gate, setting it to logic 1-state. The charge flow mechanism in pMOS transistors is similar yet asymmetric (with reversed polarity) to that of nMOS transistors. A pMOS transistor is ON when the gate is neutral, which, provides a path from the source to the drain.¹⁵ The charge flow in nMOS and pMOS networks will be explained further as we introduce the circuit and its operation.

The nMOS and pMOS transistor gates provide a connection between the output and the source, V_{SS} , and between the output and the drain, V_{DD} , respectively. Transistors gates can be used to build networks that perform logic functions, and based on their nMOS and pMOS characteristic they are called a pull-down (PDN) or pull-up

¹⁵There are two modes in FET devices that determine whether the transistor is ON or OFF at zero gate-source voltage; depletion mode and enhancement mode. If the device is ON at zero gate voltage then it is referred to as a depletion-mode device, and if it is OFF at zero gate voltage then it is referred to as an enhancement-mode device. An enhancement-mode MOSFET can be turned on by changing the gate voltage towards V_{DD} . This change is positive for nMOS and negative for pMOS devices.

(PUN) network, respectively. These MOSFETs networks are built to perform logic operations. In this section, the circuit we employ is an example of a dynamically clocked NP domino logic composed of nMOS and pMOS that performs half adder operation. The simple half adder circuit is designed and operated in a way that it will allow for comparison with the bound obtained for NASIC half adder. This application to a conventional technology will provide further insight to the key features of our approach.

We designed a dynamically clocked half adder composed of two stages as shown in Fig. 3.14 (top). The first stage corresponds to two n-block transistor networks encircled in red on the left, and the second stage corresponds to two p-block transistor networks encircled in blue on the right of the NP domino logic circuit. Fig. 3.14 (bottom) displays the logic diagram of the circuit. In its simplest and most general form a half adder is composed of an XOR and an AND gate connected in parallel. In order to perform this operation, we employ nMOS transistors to build the XNOR and NAND gate in the first stage, each of which is followed by a dynamically clocked inverter composed of pMOS transistors in the second stage.

The XNOR gate in the first stage is obtained by using two NAND gates and an AND gate as shown in the logic diagram. The logical expression for XNOR is $\overline{(A \oplus B)}$. This expression can be rewritten by using logical tautologies¹⁶ as

$$\begin{aligned} \overline{(A \oplus B)} &\equiv \overline{(\overline{A \cdot B}) + (A \cdot \overline{B})} \\ &\equiv \overline{(\overline{A \cdot B})} \cdot \overline{(A \cdot \overline{B})}. \end{aligned} \tag{3.30}$$

The final form of the expression represents the two NAND and an AND gate we used to implement the XNOR operation in the half adder circuit as depicted in the logic

¹⁶Recall that, for variables (or logic statements) X and Y , the logical expression for the exclusive disjunction, XOR, is $X \oplus Y = (\overline{X} \cdot Y) + (X \cdot \overline{Y})$, and the De Morgan's Law states that $\overline{(X + Y)} \iff \overline{X} \cdot \overline{Y}$.

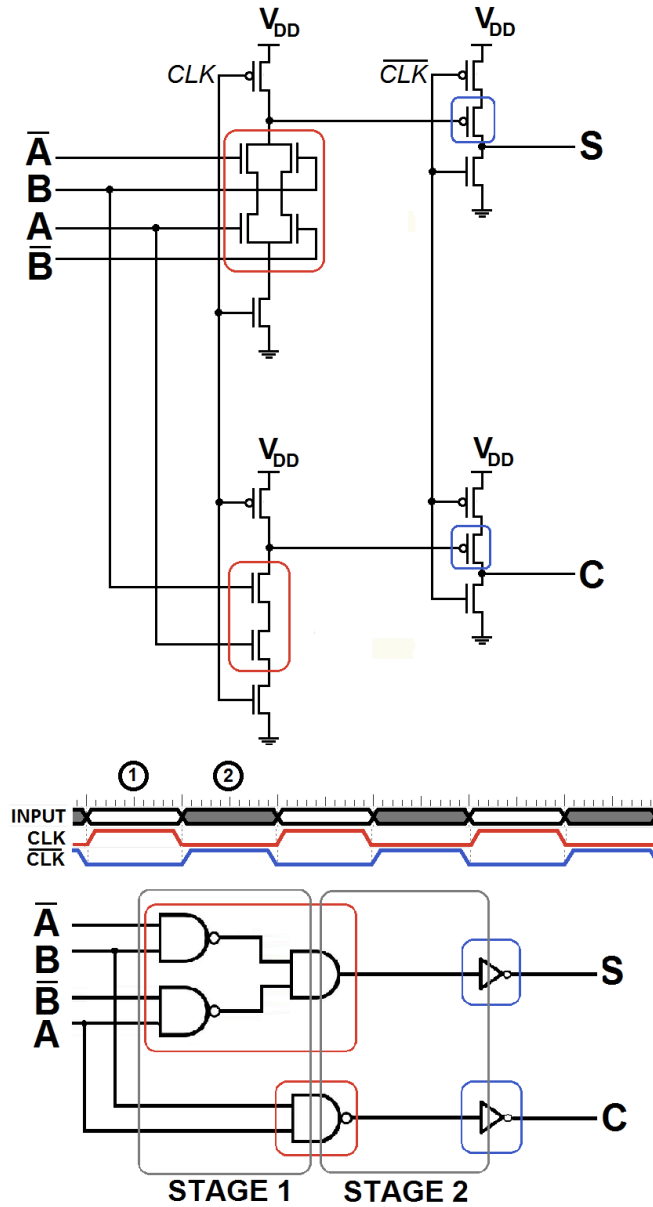


Figure 3.14. Layout, logic and timing diagram of the dynamically clocked np-CMOS half adder.

A	B	\bar{A}	\bar{B}	C	S	\bar{C}	\bar{S}
0	0	1	1	0	0	1	1
0	1	1	0	0	1	1	0
1	0	0	1	0	1	1	0
1	1	0	0	1	0	0	1

Table 3.7. The truth table of the np-CMOS 1-bit half adder.

diagram. The truth table associated with the np-CMOS half adder is presented in Table. 3.7.

The clocking scheme of the np-CMOS circuit is depicted in Fig. 3.14 (middle). The operation of this circuit is slightly different, unlike the NASIC half and full adder circuits, both stages of the np-CMOS circuit is activate at a given clock step, i.e. the stages do not go through a hold phase. The circuit stages are driven by complementing clocking signals CLK and \overline{CLK} . When the clock signal is low, $CLK = 0$, the first stage precharge high while the second stage predischarges low. When the clock signal is high, $CLK = 1$, both stages evaluate. It is important to emphasize here that the logic gates in the pMOS blocks precharges low and discharges high. In other words, the evaluation phase is when the bottom control transistor in PDNs and the top control transistor in PUNs is ON –the ① phase shown in Fig. 3.14 (middle). And, the precharge occurs when the top control transistor in PDNs and bottom control transistor in PUNs is ON –the ② phase shown in Fig. 3.14 (middle) – as can be inferred from the asymmetry in the PDNs and PUNs. We elaborate further on the circuit operation in the process abstraction below.

3.2.2.1 Abstraction

The abstraction of the CMOS circuit and its surrounding subsystems is similar to the abstraction of the NASIC paradigm. We situate the circuit in an isolated and closed universe, and the abstraction of this universe allows us to capture the essential functional features of the underlying computational strategy.

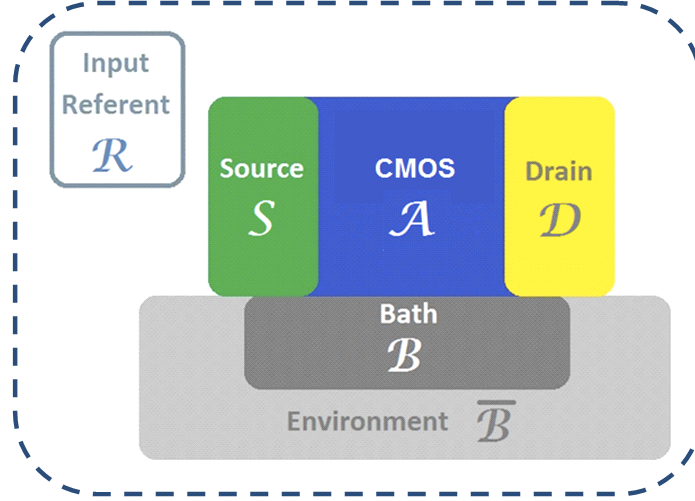


Figure 3.15. Physical abstraction of the CMOS 1-bit half adder situated in its surrounding environment.

Physical Decomposition – The abstraction used to describe the np-CMOS adder and its environment is similar to that of NASIC paradigm, and is depicted in Fig. 3.15. The artifact \mathcal{A} is the CMOS half adder, \mathcal{S} correspond to the ground, low voltage plate (particle source), and \mathcal{D} represents the plate at a higher, drain (V_{DD}), potential. Similar to the previous abstraction, we nominally regard the source and drain as idealized Fermi gases at temperature T with associated chemical potentials μ_{SS} and μ_{DD} , respectively, with $\Delta\mu = \mu_{SS} - \mu_{DD} = qV_{DD}$. The bath \mathcal{B} is the underlying substrate in direct thermal contact with the CMOS half adder, and is nominally in a thermal state at temperature T . It is important to note here that for the PDN the substrate corresponds to the p-type Silicon bulk and for the PUN it corresponds to the n-type bulk as shown in Fig. 3.13. The greater environment $\bar{\mathcal{B}}$ includes subsystems that drive \mathcal{B} toward thermal equilibrium and supply the energy and particles required to maintain the nominal populations of \mathcal{S} and \mathcal{D} and a chemical potential difference $\Delta\mu = qV_{DD}$ when these subsystems are driven from their nominal states during computation. The transistor in the CMOS half adder exchanges particles with the source and drain \mathcal{S} and \mathcal{D} and heat with the bath \mathcal{B} as it processes input data

held in the register (referent) \mathcal{R} . The greater environment $\bar{\mathcal{B}}$ provides the energy, particles, and heat removal that enable circuit operation, and everything else required to thermodynamically isolate the global system.

The interactions between subsystems is control by the clock signal, CLK . The state $CLK = 0$ signals precharge for the first stage, PDN blocks, and predischarge for the second stage, PUN blocks, and $CLK = 1$ loads the input as well as evaluating the output at the same step. The precharge and evaluate operations selectively open the PDN and PUN in the CMOS to particle exchange with the drain \mathcal{D} and source \mathcal{S} , respectively. The subsystems \mathcal{C} , \mathcal{S} , and \mathcal{D} can also exchange heat with \mathcal{B} during each computational step. Heat exchange between $\bar{\mathcal{B}}$ and the subsystems \mathcal{B} , \mathcal{S} and \mathcal{D} , and particle exchange between $\bar{\mathcal{B}}$ and \mathcal{S} and \mathcal{D} , are assumed to restore \mathcal{B} , \mathcal{S} and \mathcal{D} to their nominal states at the conclusion of each computational step. The details of this operation is discussed below.

Process Abstraction – The interactions between the subsystems occurs in a similar fashion to that of NASIC paradigm. The order of clock operations and circuit stage activation is different to some extent and can be outlined as follows. The PDNs and PUNs are controlled by complementing clock signals, CLK and \overline{CLK} . The PDN gates can directly drive PUN gates, and vice-versa. During the precharge phase, $CLK = 0$, the outputs of the PDN are set to 1, i.e. precharged, and the outputs of the PUN are pre-discharged. As we mentioned above, this step takes place by connecting the PDN outputs to V_{DD} and the PUN outputs to the source, $V_{SS} = 0$. The PDN precharge control transistors are located on top of the network, whereas for the PUNs the precharge control transistor is located at the bottom. At this time the PUN is turned off since the n-tree gate connects PMOS pull-up devices [38]. When the clock signal is high, $CLK = 1$, the PDN outputs make a conditional transition from 1 to 0, and during this evaluation some transistors in the PUN are turned on. In the mean time PDN inputs are precharged to 0.

In other words, the precharging of all the PDN blocks, and predischarging of all the PUN blocks occur in parallel when the clock signal is low, i.e. $CLK = 0$. When the clock signal is high, i.e. $CLK = 1$ *all* the blocks perform evaluation; PDN and PUN each type maintaining the monotonically rising or monotonically falling property, respectively. Of course there is a cascading effect to propagate the effect of inputs (at the first stage) to the output (at the last stage), but all of this happens during the high CLK signal phase. $CLK = 1$. The inputs must be stable during this phase.¹⁷

The sequence of state transformations that comprise a single computational cycle of the CMOS half adder can be described similar to that of the transformations of the NASIC paradigm. Initially, all MOSFETs in the PDNs are precharged, set to their logic 1-state, and the ones in PUNs are predischarged, the bath \mathcal{B} is in a thermal state, and the referent state for the η -th input is represented as defined in Eq. (3.1).

On each step, the global system evolves unitarily via the time-dependent Schrödinger equation. Table 3.8. outlines the initial states and the structure of the state transformation associated with each computational step, with \mathcal{C}_k indicating the clock zone that changes state during computational step c_k . The state of \mathcal{C}_k is determined by the pattern of discharged transistors in the computationally relevant circuit stage.

The control operation of every computational step is followed by a restoration phase. For evaluation steps, the restoration operator is $\hat{U}_{EVA}^{rest} = \hat{U}^{\mathcal{B}\bar{\mathcal{B}}} \otimes \hat{I}^{\mathcal{R}_\eta \mathcal{C}_k \mathcal{S} \mathcal{D}}$, which rethermalizes the bath and \mathcal{S} . The restoration operation for precharge steps, which is of the form $\hat{U}_{PRE}^{rest} = \hat{U}^{\mathcal{B} \mathcal{S} \mathcal{D} \bar{\mathcal{B}}} \otimes \hat{I}^{\mathcal{R}_\eta \mathcal{C}_k}$, rethermalizes \mathcal{B} and recharges and rethermalizes \mathcal{S} and \mathcal{D} .

¹⁷As one can see, the output of the final stage is obtained in the same evaluation cycle as the loading of the input to the initial stage. This is the case regardless of the number of the cascaded NP stages. Surely, there is a delay depending on the number of stages but the loading of the input to the initial stage and obtaining the output from the final stage happens at the same clock step, unlike the NASIC paradigm.

Computational Steps	Initial State	State Transformation	Control Operation
c_1	$\hat{\rho}_0 = \left(\sum_{i=1}^{M=4} p_i \hat{\rho}_i^{\mathcal{R}_\eta} \right) \otimes \hat{\rho}^{\mathcal{C}_k} \otimes \hat{\rho}^{\mathcal{S}} \otimes \hat{\rho}^{\mathcal{D}} \otimes \hat{\rho}^{\mathcal{B}} \otimes \hat{\rho}^{\bar{\mathcal{B}}}$	$\hat{\rho}_1 = \hat{U}_{EVA}^{rest} \hat{U}_1 \hat{\rho}_0 \hat{U}_1^\dagger \hat{U}_{EVA}^{rest\dagger}$	$\hat{U}_1 = \hat{U}^{\mathcal{R}_\eta \mathcal{C}_k \bar{\mathcal{C}}_k \mathcal{S} \mathcal{D} \mathcal{B}} \otimes \hat{I}^{\bar{\mathcal{B}}}$
c_2	$\hat{\rho}_1 = \left(\sum_{i=1}^4 p_i \hat{\rho}_i^{\mathcal{R}_\eta} \otimes \hat{\rho}_{1,i}^{\mathcal{C}_k} \otimes \hat{\rho}_{1,i}^{\mathcal{S}} \otimes \hat{\rho}_{1,i}^{\mathcal{D}} \otimes \hat{\rho}_{1,i}^{\bar{\mathcal{B}}} \right) \otimes \hat{\rho}^{\mathcal{B}}$	$\hat{\rho}_2 = \hat{U}_{PRE}^{rest} \hat{U}_1 \hat{\rho}_1 \hat{U}_1^\dagger \hat{U}_{PRE}^{rest\dagger}$	$\hat{U}_2 = \hat{U}^{\mathcal{C}_k \mathcal{S} \mathcal{D} \mathcal{B} \bar{\mathcal{C}}_k} \otimes \hat{I}^{\mathcal{R}_\eta \bar{\mathcal{B}}}$

Table 3.8. State transformations for the np-CMOS 1-bit half adder

At the end of second and final step of computation, the circuit loses all the information about the η -th input and \mathcal{C} is no longer correlated with \mathcal{R}_η .

The np-CMOS half adder circuit structure presented in Fig. 3.14 does not support multi tile operation.¹⁸ In this example, we study a single tile without defining a supporting computational system $\bar{\mathcal{C}}_k$. Also, unlike the QCA and NASIC examples we studied above, the number of inputs and outputs in this circuit is not symmetric. However, one can slightly modify the structure to accommodate computation with upstream and downstream tiles that will be activated simultaneously. Below, we obtain a lower bound on the amount of dissipation associated with processing a single input in this single shot computation. Based on the operation of the np-CMOS half adder, following the initial computational step the greater environment $\bar{\mathcal{B}}$ is correlated with \mathcal{R}_η due to the restoration of \mathcal{B} after evaluation. Two classes of clock operations for the np-CMOS half adder is defined in terms of the two states of the clock signal $CLK = 0$ and $CLK = 1$ as precharge, ϕ_P , and evaluate, ϕ_E , respectively. During precharge ϕ_P , PDN inputs and PUN outputs are connected to \mathcal{S} , and PDN outputs are connected to \mathcal{D} , i.e. electrons flow in to PDN through its inputs, out from PDN through its outputs (inputs to PUN) and in from PUN towards its output. During evaluation ϕ_E , PDN inputs are connected to \mathcal{D} , PDN outputs (PUN inputs) are connected to \mathcal{S} , and PUN outputs are connected to \mathcal{D} . After each of these steps a subsequent “reinvestment” of energy is required by $\bar{\mathcal{B}}$ during the restoration to recharge \mathcal{S} and \mathcal{D} with a cost to the greater environment.

Below, the dissipation costs resulting from both the logical irreversibility and particle supply cost are captured by following the guidelines provided in Sec. 2.3.

¹⁸The circuit operation involves simultaneous activation of all stages.

3.2.2.2 Analysis

Operational Decomposition – The clock cycle $\Phi = \varphi_1\varphi_2$ is a periodic sequence of two clock phases φ_v , each of which is an assignment of operations ϕ to the two simultaneously controlled clock zones identified as Stage 1 and Stage 2 of Fig. 3.14. The duration of the full computation requires two clock steps. The steps that involve the manipulation of data related to a given input within the tile are labeled as ① and ② on the figure, corresponding to c_1 and c_2 , respectively.

Denoting these clock zones as $C(1)$ and $C(2)$, respectively, the assignment corresponding to the adder clocking described above is

$$\varphi_1 : \{(C(1); \phi_E), (C(2); \phi_E)\}$$

$$\varphi_2 : \{(C(1); \phi_P), (C(2); \phi_P)\}.$$

The circuit stages are activated simultaneously. Here ϕ_E corresponds to high clock signal, $CLK = 1$, and ϕ_P corresponds to low clock signal, $\overline{CLK} = 1$, . The computational cycle $\Gamma^{(\eta)}$ for the η -th input is the straightforward sequence

$$\Gamma^{(\eta)} = c_1 c_2 = \varphi_1^{(1)} \varphi_2^{(1)} \quad (3.31)$$

of two computational steps c_k associated with this cycle.

The operation of the dynamically clocked CMOS circuit we study here does not involve copying of information and the information is lost in a single step. The information is loaded in the circuit and erased in each clock step, the complete circuit acts as a single data zone. Therefore, in this paradigm, we do not define data zones to track information flow through the circuit, and to isolate the sources of irreversible information loss within a computational cycle. However, for consistency, the data zones can be represented in terms of the circuit stages

$$D(c_1) = C(1) \cup C(2)$$

$$D(c_2) = \emptyset.$$

Cost Analysis – The total dissipative cost associated with one computational cycle for the np-CMOS half adder is similar to that of NASIC paradigm. The particle transfer and the amount of information erasure is different to some extent and can be explained as follows. Computational steps c_1 is the evaluation step, during which the information about the input $X^{(\eta)}$ (i.e. about \mathcal{R}_η) is transferred to \mathcal{C} without irreversible information loss from $C(1)$ or $C(2)$. Thus, $\Delta\mathcal{I}^{\mathcal{R}_\eta\mathcal{C}_1} = \Delta\mathcal{I}^{\mathcal{R}_\eta\mathcal{C}_2} = 0$. The outputs for the PDNs are obtained by an electron flow into the PDN (towards PUN) from the \mathcal{S} . In order to prevent double counting, we account for the particle cost when an electron completes its transport from the ground to V_{DD} – this particle cost will be accounted for during precharge. However during evaluation, the inputs to PDNs are loaded by three electrons flowing into \mathcal{D} from the PDN gates (two electron from the top PDN, and one from the bottom), we take $\Delta N_1 = 3$. The computational step c_2 is the precharge step. During c_2 , the two particles that entered the circuit in the previous step are transported to the drain, this is, of course, assuming that each FET requires one electron to switch,¹⁹ and all of the information about input $X^{(\eta)}$ is irreversibly lost from the circuit, simultaneously from $C(1)$ and $C(2)$, yielding $\Delta N_2 = 2$ electrons and $\Delta\mathcal{I}^{\mathcal{R}_\eta\mathcal{C}_3} = -2$ bits. Note here that we assume equiprobable inputs.

Contributions to this bound from information loss and particle supply at each computational step are tabulated in Table 3.9. The dissipative cost of processing a single input through a full computational cycle of the np-CMOS half adder is then obtained similar to the fundamental lower bounds of NASIC adders,

$$\Delta\langle E^{\mathcal{B}} \rangle_{TOT} \geq 2k_B T \ln(2) + f5qV_{DD}, \quad (3.32)$$

¹⁹In Sec. 3.2.1, we made the assumption that the number of electrons required to switch each FET is $\Delta n = 1$.

Computational Step	Particle Supply	Information Loss
c_1 : EVA	$3qV_{DD}$	0
c_2 : PRE	$2qV_{DD}$	$2k_B T \ln(2)$
Cycle Total	$2qV_{DD}$	$2k_B T \ln(2)$

Table 3.9. Dissipation bound for the np-CMOS 1-bit half adder: Particle supply and information loss components.

where $\Delta \langle E^{\mathcal{B}} \rangle_{TOT}$ is the total amount of heat dissipated into the circuit's surroundings during one computational cycle.

It is important to note here that the bound (3.32) assumes that the np-CMOS half adder circuit has both input and output complements present. The circuit structure presented in Fig. 3.14 (top) depicts two inputs with complements, and two outputs without complements. The second term in the bound (3.32), $f5qV_{DD}$, is calculated by considering output complements to make the analysis inline with the analysis of the NASIC half adder in Sec. 3.2.1.3. This symmetry can be achieved by slight modification in the circuit diagram in Fig. 3.14.

The lower bound (3.32) may be interpreted as, in the best case scenario, CMOS can be more energy efficient than NASIC paradigm given that both the cost of local information loss associated with irreversible information processing (2 bits) and the supply of (particles dropping through potential difference of V_{DD}) maintaining the computational working substance (5 electrons) are less than that of NASIC half adder. However, it is important to underline once again that this bound is independent of assumptions regarding material, device, or circuit dimensions or parameters. In practice, the manufacturing techniques proposed for the NASIC allow aggressive scaling that can mean higher performance, density, and power efficiency that can go far beyond the performance of CMOS technology. Analyzing CMOS circuits from the fundamental energy requirement point of view stands as an academic effort due to the the practical limitations imposed on the CMOS technology. Here, we studied the CMOS adder simply to provide further guidelines for the implementation of

our methodology via a conventional paradigm, by making no claims regarding the performance of CMOS in relation to the post-CMOS paradigms we analyzed earlier.

The analysis of the np-CMOS half adder does not only serve as an application to a well-known and familiar circuit structure but also provides us valuable insights towards the limitations of our methodology. The assumption we made regarding the need for the complemented inputs and outputs initially originated from the intention to make the logic operation similar the operation of the NASIC half and full adder examples we studied in Sec.s 3.2.1.3 and 3.2.1.4. However, further scrutiny of our assumption revealed that this symmetry is also the key for the accurate evaluation of the particle cost component in the bound. Our calculations show that, if the output complements are not a part of the circuit then the number of electrons flowing in and out of the circuit depends on the input processed at a given time. The presence of output complements creates such symmetry in the circuit that it makes the particle flow and associated cost to be independent of the probability of the η^{th} input. The advantage of the NASIC examples studied is that the number of inputs and outputs are identical and each stage has the same number of electrons flowing in and out of the circuit regardless of the input. It is important to note here that the non-transistor based paradigms are free of this requirement to include output complements. The breadth of our methodology presented here can be expanded further to accommodate transistor-based circuits without output complements, however, such a modification is beyond the scope of this work. In the next section, we study a static CMOS circuit example and elaborate further on the limitations of our methodology.

3.2.3 An Application to Static CMOS Circuits

Above we presented applications of our methodology to various post-CMOS nano-electronic proposals as well as a conventional CMOS circuit. The illustrative examples we studied so far are circuits that are dynamically clocked. Our analysis on the QCA

paradigm showed that the operation of the circuit significantly affects the fundamental lower bound. In non-transistor based paradigms, as long as the irreversible information loss is accurately localized, the static implementations too can easily be treated with our methodology. However, for transistor-based applications, identifying the particle supply cost in static implementations can be challenging. In this section, we discuss these challenges as well as other limitations of our methodology via another conventional CMOS circuit example.

We consider a combinational two-bit sorter proposed by Gershenfeld [17]. In 1996, Gershenfeld studied the information dissipation of a two-bit sorter circuit implemented in CMOS from a thermodynamic perspective. He explored the connection between the information and thermodynamic properties of a system that can implement it, emphasizing the importance of studying information bearing and non-information bearing thermodynamic degrees of freedom together to obtain the fundamental physical limits for a given technology base in emerging electronic circuits. He proposed a sketch for the unified theory of degrees of freedom and calculated the minimum energy cost of computation in a combinational and a sequential two-bit sorter.

The circuit, as shown in Fig. 3.16, is composed of an AND and OR gate connected in parallel which are composed of a combination of parallel MOSFETs. The AND and OR operations are obtained by a NAND and NOR logic gate each followed by an inverter as shown in the figure. The NAND and NOR gates consist of two nMOS and two pMOS transistor gates where as the inverters are made up of an nMOS and a pMOS gate. Overall the circuit is composed of six nMOS and six pMOS gates.

The truth table associated with the logic operation is presented in Table 3.10. From input to output mapping, the number of zeros and ones remain the same for each operation, however, the self entropy of output distribution is one-half less than

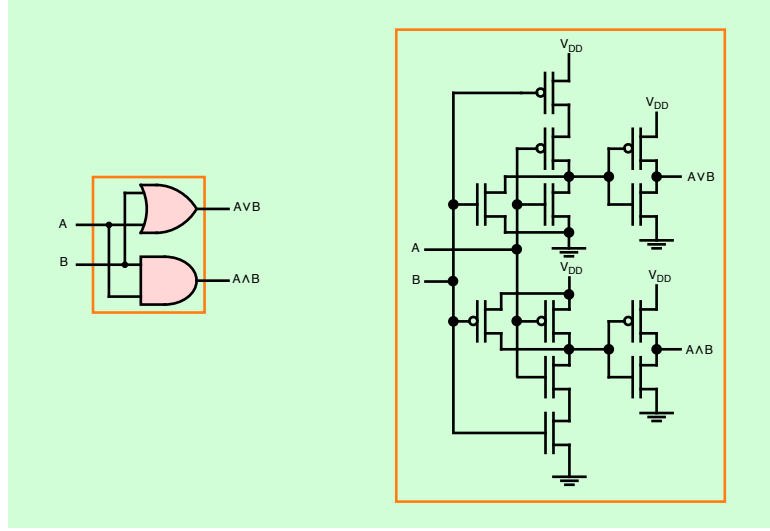


Figure 3.16. Gershenfeld's combinational 2-bit sorter [17].

IN		OUT	
0	0	0	0
0	1	1	0
1	0	1	0
1	1	1	1

Table 3.10. The truth table of the CMOS 2-bit sorter.

the self entropy of the input distribution, assuming that inputs are equiprobable.²⁰ The circuit dissipates energy every time an input is loaded. The output capacitors charge and discharge the inputs to the consecutive transistor; inputs and outputs exist simultaneously and independently as charge stored in the corresponding gates.

The static CMOS sorter bears certain structural similarities with the np-CMOS adder circuit studied in Sec. 3.2.2, however, the design and operation of the circuit renders this example radically different from the earlier transistor-based implementations we studied. First, this circuit is asynchronous, the computation of an input is

²⁰The self entropy of the input and output distributions is $H(X) = -\sum_{i=1}^3 p_i \log_2 p_i = 2$ bits and $H(Y) = -\sum_{j=1}^3 q_j \log_2 q_j = 1.5$ bits, respectively.

composed of one single step rather than multiple computational steps operated by the clocking scheme as we have seen in QCA, NASIC and np-CMOS adder circuits above. Second, each new input is loaded in the circuit without a prior reset, i.e. there is no intermediate erasure between computational cycles. The erasure of a given input is done via overwriting of the next input, which means that the energy associated with erasure of a given input is dissipated over the operation of two inputs. This makes the cost of computation for a given input dependent on the previous input. These characteristics make the fundamental analysis of the static CMOS circuit immune to the analysis by the approach presented in Chapter 2 in its current form.

One may be prompted to think that the physical abstraction of the circuit can be constructed in a similar fashion to that of NASIC and np-CMOS adders. However, the CMOS sorter is not divided into circuit stages, all the transistor gates in the circuit interact with the source and drain simultaneously. We can compare this to the operation of the NASIC and np-CMOS adders. In NASIC, at a given computational step, one circuit stage is active and the other stage is in hold state; the circuit interacts with either the source or drain at a given step through the active stage. In the np-CMOS half adder, at a given computational step, both stages of the circuits are active, however, they operate independently; i.e. one stage interacts with the source and the other stage interacts with the drain. As opposed to the NASIC and np-CMOS adders, the static CMOS adder circuit interacts with the source and drain simultaneously during the processing of a given input. The physical decomposition and associated subsystem interactions therefore cannot be constructed based on the guidelines provided in Sec. 2.2.1 and as demonstrated for the dynamically clocked

circuit examples above.²¹ And the asynchronous nature of the sorter renders the operational decomposition presented in Sec. 2.2.2 inapplicable to the static circuit.

Furthermore, we assume that inputs are equally probable for the static CMOS sorter as was the case for the QCA, NASIC and np-CMOS circuit examples. However, in the static circuit, the number of electric charges flowing in and out of the circuit depends on both the current, η^{th} , and previous, $(\eta - 1)^{th}$, input. Four possible combinations of AB for $(\eta - 1)^{th}$ input followed by the four possible combination of η^{th} input gives us sixteen scenarios for the number of charges flowing in and out of the circuit. Therefore, the processing of a given input involves one of sixteen possible scenarios of charge transport. In the dynamic transistor-based circuit examples we studied, the number of electrons flowing in and out of the circuit was the same regardless of the input. Due to this characteristic of the static circuit, the second term in the fundamental bound (3.12) should be evaluated to accommodate for the amount of information embedded in the probabilistic distribution of charge transport scenarios.

The evaluation of the fundamental cost of computation in static circuits is possible after certain modifications to our work.²² However, such expansions in the methodology are beyond the scope of this dissertation.

3.3 Discussion

In this chapter, we presented various illustrative applications of our methodology via three distinct computational paradigms. We studied a QCA half adder controlled using Landauer and Bennett clocking schemes that support pipelining, and compared

²¹The heterogeneous structure of the substrate and simultaneous interactions that involve charge exchange with multiple subsystems require the methodology to be expanded to accommodate the physical decomposition of a static circuit.

²²For instance, the guidelines presented in Ref. [40] can be use to calculate the dissipative cost of the overwriting information.

the resulting power dissipation bounds with the relative clock speeds adjusted to achieve the same computational throughput in the two schemes. For the particular pipelining granularities considered, we found the lower bound on power dissipation for Landauer clocking to exceed that of Bennett clocking by nearly a factor of two. Our study demonstrate the benefits that Bennett clocking can provide in terms of fundamental energy dissipation even in small circuits..

The application of our approach to transistor-based paradigms showed that due to external particle supply required for computation the resulting bounds account both for dissipative cost of logical irreversibility and additional particle-supply cost required to minimally maintain the computational working substance in these paradigms. Implementation of our approach in the NASIC resulted in an inequality with two terms. The first term in the bound, which corresponds to the logical irreversibility cost, is of the same order as that obtained here for the QCA adder. The second term, which has no analog in QCA, corresponds to the cost of charge flow required to maintain the computational working substance in the circuit. Comparing the two terms in the NASIC bound for $T = 300\text{K}$ and $V_{DD} = 0.8\text{ V}$, and taking $f = 0.5$, we find that irreversible information loss accounts for $4k_B T \ln(2) = 0.07\text{ eV}$ and $6qV_{DD} = 2.40\text{ eV}$ (for $f = 1$). Thus, even in this idealized scenario – where the particle supply requirements are at their absolute minimum²³ and half of the energy provided by the power supply encodes information – particle-supply accounts for $> 97\%$ of the lower bound on the dissipative cost. The dominance of the particle-supply cost in the bound can be reduced but not eliminated by downscaling V_{DD} . The directional electron injection and extraction processes that drive particle exchange between the circuit and microwires require that $V_{DD} \gg k_B T$, in order to enable proper circuit

²³Recall that, we take the number of electrons required to switch a FET to be $\Delta n = 1$.

operation and reliable computation the particle supply cost has to exceed the logical irreversibility cost.

This work shows the importance of systematically isolating and localizing irreversible information loss even for circuits as simple as those considered here. In detailed analysis of the structure and operation of the QCA adder, five blocks of cells were identified as the data subzones that independently contribute to the 3.76 bits of information irreversibly lost in each computational cycle. This is well above ($7.5\times$) the minimum of 0.5 bits of information that must be lost by *any* physical implementation of a half adder that erases its inputs, and is well below ($0.03\times$) the 114 bits of information loss that would be expected if the 135 cells in the circuit were erased individually. The intermediate bound obtained through systematic analysis of the QCA adder using our methodology thus differs substantially from dissipation bounds that would be obtained by “guessing” either extreme – the entire adder or individual cells – to be the appropriate levels of granularity for quantifying irreversible information loss. We discuss the granularity of the heat dissipation for the QCA half adder circuit in Appendix A in detail.

Although, the individual logic gates are identified as the appropriate level of granularity in the QCA adder studied here, this need not be the case in QCA in general or in other paradigms; the corresponding analysis for the NASIC adders tie irreversible information loss to circuit regions that do not even correspond to individual logic gates. It is important to note here that, this explicit dependence on circuit structure and operation (among other things) distinguishes our approach from that of Zhirnov and Cavin ([41] and references therein). The systematic isolation of independent sources of dissipation distinguishes our methodology from approaches that do consider explicit circuit structure and operation, but that “pre-assign” individual gates as the sources of dissipation and neglect the dependence of the dissipation on the statistics of the circuit and gate inputs (e.g. [42] and references therein).

We also note that the energy cost of irreversible information loss reflected in the dissipation bounds for the QCA and NASIC adders are of very similar magnitudes for the two very different adder implementations. The additional term appearing in the NASIC bound ($f q V_{DD} \Delta N_{TOT}$), which has no analog in QCA,²⁴ reflects the additional cost of irreversible particle transfer required to maintain the computational working substance in the circuit.

Aside from the post-CMOS technology proposals, we also studied two examples of CMOS paradigm. Our analysis of these conventional circuits served as a tool to provide further insight into the application of our approach. The final results we obtained for the np-CMOS half adder circuit are not intended for comparison with the post-CMOS examples from a fundamental energy requirement perspective. The CMOS circuits are best treated by other approaches due to the excess particle cost.

Lastly, we observed that the more difficult it is to design a circuit structure and its operation, easier it is to analyze the circuit from the fundamental energy requirement point of view due to the advantage it presents in localizing the irreversible information loss and associated physical operations taking place in the circuit.

²⁴The absence of paradigm-specific overhead costs in the QCA bound assumes that energy invested by the clock is reversibly recoverable in principle, thus does not represent a *fundamental* cost. Recent calculations suggest that, in molecular QCA, the *parasitic* losses associated with clocking can be much smaller than the unavoidable dissipative cost cell switching that enables computation [34].

CHAPTER 4

TOWARDS AUTOMATION: MODULAR DISSIPATION ANALYSIS

In Chapter 2, we introduced a general methodology that enables determination of fundamental heat dissipation bounds for concrete and non-trivial nanocomputing scenarios. In Chapter 3, we showed how irreversible information loss can be isolated in a given non-transistor or transistor based circuit, and how lower bounds on the resulting energy dissipation can be obtained. Our studies have shown that, even for the single-tile adder circuits, localizing irreversibility requires hard work, and for large and complex circuit the general methodology can be extremely laborious. In order to accommodate evaluation of fundamental lower bounds for large and dense circuit structures we propose to “modularize” our approach. Here, we present a modularized analysis to facilitate – and possibly automate – the determination of fundamental dissipation bounds for large, complex circuits designed according to specified rules.

4.1 Modular Dissipation Analysis of a QCA Half Adder

In Section 3.1 we applied our general approach to QCA half adder operated under Landauer and Bennett clocking schemes. Here, we show how this approach can be modularized by decomposing a circuit into smaller zones – such that dissipative contributions can be evaluated separately for each zone and summed – and how this can simplify dissipation analysis of QCA circuits (hereafter the “modular approach”). We stress the enabling feature of this decomposition; that it preserves the effects of field interactions across tile boundaries that influence the reversibility of information

loss. We provide a comparative comparison between the general analysis presented in Sec. 3.1 and modular analyses presented here. We briefly discuss prospect for automation of QCA dissipation analysis using the modular approach. Automated analysis could, for example, enable evaluation of dissipation bounds for full QCA-based processor architectures [43] via the approach of [44] by simplifying circuit-level dissipation analyses of the constituent circuit blocks.

4.1.1 Design Rules

We begin discussion of our modular approach by articulating an example set of QCA design rules, since circuit decomposition rules defined for a set of design rules can be applied to modular dissipation analysis of any circuit designed according to these rules. For each set of design rules, there is a modular dissipation analysis procedure; we will demonstrate a modular analysis specific to design rules presented here. Our example design rules, which are specific to Landauer-clocked combinational QCA circuits with no wire crossings, are as follows:

1. *Wires*: All wires are linear arrays of “90-degree” wires, i.e. with adjacent cells oriented as in Fig. 4.1¹, and with right-angle corners. Wire segments corresponding to individual clock zones are of length

$$2 \leq N \leq \exp[E_k/k_B T]$$

(in units of cells), where E_k is the kink energy and k_B is Boltzmann’s constant.

The minimum allowable wire pitch is three cells.

¹We do not consider “45-degree” wires since the design rules we propose are not intended for wire crossings in a plane.

2. *Inverters*: Inverters are of the “complex” form shown in Fig. 4.2, with identically clocked, two-cell input leg and an identically-clocked two-cell output leg as shown.
3. *Majority Gates*: Majority gates are of the standard three-input configuration in Fig. 4.3. The four input and output legs adjacent to the central cell – hereafter the “device cell” – are identically clocked, and the four identically clocked input and output legs are of equal length. The majority gates with cell polarizations fixed as -1 and +1 function as two-input AND or OR gates, respectively.



Figure 4.1. Three example sections of QCA wires with “90-degree” cell orientation.

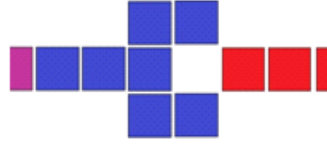


Figure 4.2. A complex inverter with two-cell input and output legs and specified clock zones.



Figure 4.3. QCA majority gates, with and without one fixed input, and associated clocking.

The lower and upper bounds on the number of identically clocked cells in the wire segments, which are based on considerations discussed in Ref.s [45] and [46] respectively, are selected to help ensure reliable information transfer. The minimum pitch is selected to minimize crosstalk from adjacent wires. The requirement of equal-length input and output legs in majority gates helps to ensure simultaneous arrival of new inputs at the device cell, and thus fair voting [45].

We emphasize this simple set of design rules is presented for the purposes of illustrating our modular dissipation analysis. Having said that, these simple rules allow for construction of QCA circuits that implement any desired Boolean function² and are generally consistent with common QCA design practice. It is easily verified that the adder design of Fig. 3.2 adheres to these simple design rules.

4.1.2 Decomposition Rules

Decomposition of a QCA circuit for modular analysis requires that the circuit first be segmented into zones according to a design-rule-specific set of decomposition rules. These rules stipulate how boundaries between the zones are to be placed. For the set of example design rules presented above, the decomposition rules are simply as follows:

1. Every cell in the circuit must belong to one and only one zone.
2. All zone boundaries must be placed *between* adjacent, identically clocked cells, perpendicular to the direction of information flow. The same applies to the boundary enclosing the full circuit.

²AND, OR, and NOT form a universal set of primitives, and three-input majority gates can implement two-input AND and OR functions if one of the inputs is appropriately biased as shown in Fig. 4.3.

3. Zone boundaries are to be placed between the two cells of the input legs and between the two cells of the output legs of inverters, with no boundaries in between.
4. Majority gates must be enclosed within a single zone as in Fig. 4.4, i.e. with zone boundaries placed so they enclose (a) the device cell and the identically clocked, equal-length input and output legs, (b) one cell adjacent to each input leg in the neighboring clock zone³, and (c) one cell adjacent to the output leg in the neighboring clock zone. We refer to this zone in Fig. 4.4 as the dissipation zone.

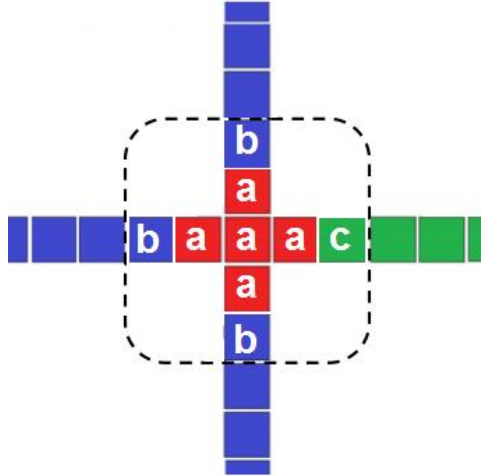


Figure 4.4. A dissipation zone, including placement of boundaries, for circuits designed according to the design rules of presented here.

It is important to note that, for circuits designed according to the rules presented here, all dissipation zones resulting from modular dissipation analyses performed according to the decomposition rules of this work have this form. The design rules preclude irreversibility in all circuit structures other than majority gates. Use of the

³Alternatively, this cell could be fixed if the corresponding input is to be biased.

decomposition rules presented here greatly simplifies dissipation analysis, as shown below.

4.1.3 Dissipation Analysis

The modular approach aims to simplify evaluation of the fundamental dissipation bounds obtained via the general approach by partitioning the circuit into smaller zones – once and for all at the beginning of the analysis – and applying the general approach piecemeal to determine the dissipative contributions from each zone. The partitioning process can, however, introduce an artifact that causes the modular approach to overestimate the dissipative contributions from the individual zones, and thus from the circuit as a whole when the individual contributions are summed. We now describe the origin of this artifact, and show that it is avoided in circuits that are designed according to the above design rules and partitioned according to the above decomposition rules.

Information propagation in Landauer-clocked QCA is not dissipative under paradigmatic operation. Information lost from a block of adjacent, identically clocked cells in a QCA wire during the “relax” phase of the clocking cycle are always erased in the presence of (and in interaction with) an identical copy that has already been transferred to – and is locked in – an adjacent block of cells that is immediately downstream. This is the reversible ERASE WITH COPY operation. If the block of cells being erased belongs to a particular circuit zone, but the downstream copy does not, then the erasure is *irreversible* – and thus dissipative – in a dissipation analysis that treats the zone including the erased cells as independent and isolated. Neglect of the cross-boundary interactions that renders the erasure reversible are lost, causing the simplified modular analysis to fail.

If the decomposition rules stated above are followed, however, any group of identically clocked cells is necessarily a wire segment that belongs to two circuit zones.

Information in the furthest downstream cell(s) of the upstream circuit zone is always also held in the furthest upstream cell(s) of the downstream circuit zone, since these two groups of adjacent “boundary cells” are identically clocked. Dissipation analysis of the upstream zone can thus neglect any apparent contributions from the clock phase where information is erased from the furthest downstream cells, since this information also belongs to the furthest upstream cells of the downstream circuit zone, and any dissipation that would result from irreversible erasure of this information in a subsequent clocking phase is captured in analysis of the downstream zone.

This simple constraint on circuit decomposition results in major simplification of the dissipation analysis. Dissipative contributions from each circuit zone can be calculated independently and added, and the effects of cross boundary interactions captured by the general analysis are properly reflected. Furthermore, the only circuit zones that are necessarily dissipative – those designated as “dissipation zones” – are those enclosing majority gates; there is no irreversible information loss in zones that correspond to wire segments and inverters. Dissipation analysis thus requires only that the dissipation zones be identified and their contributions calculated.

The analysis is simplified even further by the fact that, on each “use,” dissipation zones defined as above irreversibly lose information during one and only one clock transition: the clock transition in which the information-bearing state of the “core” of the dissipation zone – the device cell and surrounding identically clocked cells belonging to the input and output legs ((a) in Fig. 4.4) – is relaxed.

We proceed to modular dissipation analysis of the adder circuit of Fig. 3.2. One can immediately identify $N_{diss} = 5$ dissipation zones, which are delineated and labeled in Fig. 4.5. These dissipation zones can be analyzed separately, with each regarded as an independent information processing artifact. The amount of energy dissipated in the processing of each input by the circuit is

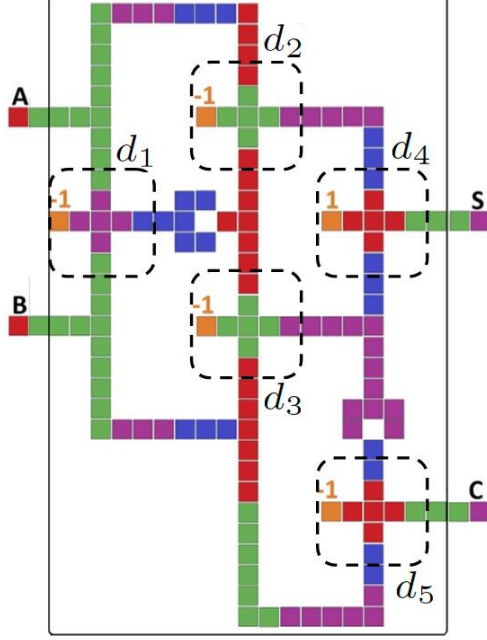


Figure 4.5. Dissipation zones identified by application of the circuit decomposition rules to the QCA half adder of this work.

$$\Delta\langle E\rangle_{TOT} = \sum_{n=1}^5 \Delta\langle E^{\mathcal{B}}\rangle_{d_n} \quad (4.1)$$

where $\Delta\langle E^{\mathcal{B}}\rangle_{d_n}$ is the amount of energy dissipated during the critical clock phase in dissipation zone d_n . $\Delta\langle E^{\mathcal{B}}\rangle_{d_n}$ is lower bounded as [9]

$$\Delta\langle E^{\mathcal{B}}\rangle_{d_n} \geq k_B T \ln(2) \Delta\mathcal{I}_{d_n} \quad (4.2)$$

where $\Delta\mathcal{I}_{d_n}$ is the amount of information irreversibly lost from zone d_n during the dissipative clock phase. Using the same assumptions of pure, orthogonal QCA data states that were made in general analysis, $\Delta\mathcal{I}_{d_n} = H_n(X|Y)$ where $H_n(X|Y)$ is the conditional Shannon entropy for the zone (gate) input and output random variables X and Y . Obtaining the probability mass functions (pmfs) for the various gate inputs that result from a uniform adder input pmf, and evaluating the five required conditional entropies, we have

$$\Delta\langle E^{\mathcal{B}}\rangle_{d_1} \geq 1.1887k_B T \ln(2)$$

$$\Delta\langle E^{\mathcal{B}}\rangle_{d_2} \geq 0.6887k_B T \ln(2)$$

$$\Delta\langle E^{\mathcal{B}}\rangle_{d_3} \geq 0.6887k_B T \ln(2)$$

$$\Delta\langle E^{\mathcal{B}}\rangle_{d_4} \geq 0.5k_B T \ln(2)$$

$$\Delta\langle E^{\mathcal{B}}\rangle_{d_5} \geq 0.6887k_B T \ln(2)$$

for the five dissipation zones. Summing these results, we obtain the bound

$$\Delta\langle E\rangle_{TOT} \geq (3.76)k_B T \ln(2) \tag{4.3}$$

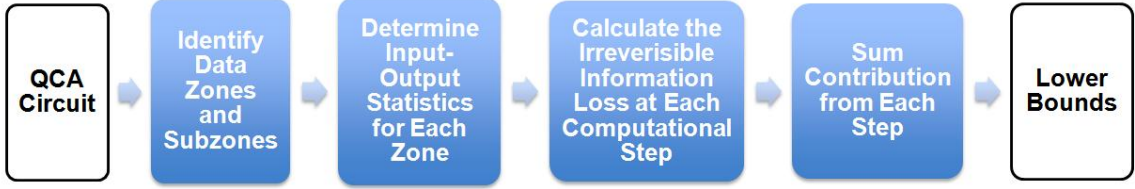
on the dissipative cost of processing one adder input, which is indeed identical to that obtained from the general approach.

Evaluation of this dissipation bound, which was shown in previous chapter to be somewhat involved even for this simple circuit in general approach (and is laborious in more complex circuits like the $> 10^5$ -cell QCA ALU studied in [35]), is straightforward and simple in the modular approach. The vast analytical simplification was enabled by the consistency of the circuit structure with stated design rules, and the identification and formulation of an appropriate set of circuit decomposition rules specific to these design rules. With decomposition rules in hand for our design rules, modular dissipation analyses could be performed in exactly the same manner for any Landauer-clocked QCA circuit constructed according to the same design rules.

4.1.4 Prospects for Automation

The modular dissipation analysis presented here is much better suited to automation than is the general approach. For easy comparison, the flow of the general and modular dissipation analysis procedures is shown schematically in Fig. 4.6. An

GENERAL APPROACH



MODULAR APPROACH

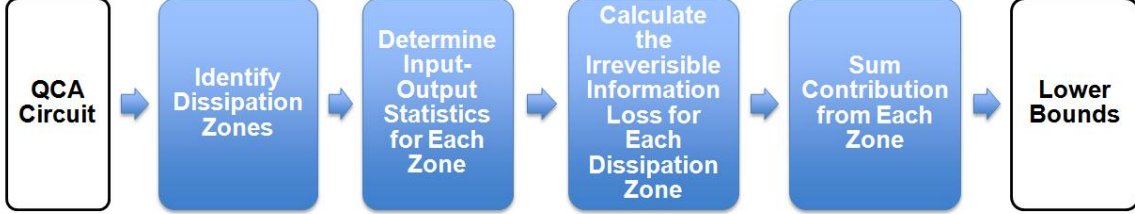


Figure 4.6. Schematic representation of the general (top) and modular (bottom) dissipation analysis procedures discussed in this work.

algorithm could certainly be devised that would enable automation of the general approach, but it would be complex and difficult to formulate and implement. The general approach requires that data zones and subzones be identified at each step, which requires that the flow of input information be tracked in space and time throughout the computational cycle. Irreversible information loss and associated energy dissipation depend upon changes in the amount of correlation between the states of these zones during each step.

It would be comparatively straightforward to formulate an algorithm for modular dissipation analysis of QCA circuits designed according to the design rules of this work and driven by a random input string with specified input pmf. The first step is simply to identify the device cells in the circuit, which could easily be performed by searching a simple matrix representation of the circuit layout. The second step is to determine the required joint pmfs and marginal (input and output) pmfs for the gates corresponding to the dissipation zones associated with each device cell. This

could be achieved, perhaps in a simulator embedding QCADesigner simulation of the circuit, by building appropriately weighted input-output histograms for all gates in simulations that step through all adder inputs. The third step is to evaluate the conditional entropies, and thus the corresponding lower bounds on the dissipative contributions, for each dissipation zone, which is easily done once the joint input-output pmfs have been determined. The fourth and final step, is simply to sum the zone contributions to obtain a dissipation bound for the full circuit. None of these steps pose formulation or implementation difficulties. We leave implementation of modular dissipation analysis in a QCA simulator for future work.

4.2 Modular Dissipation Analysis of NASICs

Above, we presented our study on enhancing the tractability of dissipation analysis for QCA circuits through modularization and possible automation. We can develop a similar strategy for the NASIC paradigm. We showed that, for circuit designs with sufficient regularity, our methodology can yield generalized, scalable bounds that are easily evaluated. In the QCA circuit obtaining this regularity requires defining a set of sample design rules. In the NASIC paradigm, however, we can obtain the modular analysis without the need for specified design rules since the physical circuit structure of the NASIC is inherently premodularized.

Our studies on the localization of information loss in the NASICs showed that each circuit stage in the circuit corresponds to a dissipation zone. The most general form of the fundamental heat dissipation bound specified for the NASIC paradigm is given in inequality (3.26). The first term in the bound is simply the sum of information loss from each circuit stage, i.e. dissipation zone. Similarly, the second term in (3.26) corresponds to the total particle loss from each dissipation zone. The nature of the NASIC is therefore premodularized. This allows us to automatically rewrite the general form of the fundamental dissipation bound in the modular form as

$$\Delta \langle E^{\mathcal{B}} \rangle_{TOT} \geq -k_B T \ln(2) \sum_{n=1}^2 \Delta \mathcal{I}_{d_n} + f q V_{DD} \sum_{n=1}^2 \Delta N_{d_n} \quad (4.4)$$

where $\Delta \mathcal{I}_{d_n}$ corresponds to the amount of information stored in and completely lost from a dissipation zone d_n , and ΔN_{d_n} is the number of particles transported from the source to drain through the given dissipation zone in each computational cycle.⁴

Localization of information loss in a circuit depends on the structure and operation of the circuit. In the NASIC, this characteristic gives rise to each circuit stage to function as a dissipation zone, which makes this circuit inherently modularized, and suitable for automation. As demonstrated by the bound (4.4), the modularization of the fundamental heat dissipation analysis for the NASIC paradigm is straightforward, and independent of any requirement for additional set of design rules on the circuit structure.

4.3 Discussion

In this chapter, we presented an approach to modularize our general methodology in order to accommodate easier determination of fundamental dissipation bounds for large, complex circuits designed according to specified rules. We compared the modularized approach with the general methodology presented in Chapter 2, and showed that the modularized analysis of QCA and NASIC paradigms gives us the same results as the ones we obtained in Chapter 3.

First, we introduced an example set of QCA design rules which allowed us to easily identify the dissipation zones that contribute to the total fundamental lower bound on the energy dissipation for the Landauer-clocked QCA circuits. We applied

⁴Here, the lower case n represents the index label corresponding to a dissipation zone, not to be confused with the Δn we defined to represent the number of electrons required to switch each FET in the transistor-based circuits.

the modular analysis to the QCA half adder circuit and show that modular analysis gives us the same results as obtained by our general approach.

Second, we modularized the fundamental heat dissipation bounds for the NASIC paradigm. Our studies showed that the structure of the circuit makes the NASIC pre-modularized and hence the analysis is straightforward and – unlike the QCA paradigm – does not require additional set of design rules.

Our calculations show that the results obtained by the general and modular approach are consistent. The modular approach can provide dramatic analytical simplification over the general approach to dissipation analysis, provided that circuits are designed according to specified design rules. We argue that the modular dissipation analysis is well suited for automation, which could enable determination of fundamental lower bounds on dissipation for large and complex circuits such as full processors.

CHAPTER 5

FOUNDATIONAL EPILOGUE: REVISITING LANDAUER’S PRINCIPLE AND THERMODYNAMICS OF COMPUTATION

Landauer’s inaugural work [4] on the relation between information and energy is an invaluable contribution to thermodynamics of computation. Despite the differences between the nature of LP and our approach, under certain circumstances our methodology yields results that resonate with LP to a degree. There are various scholars and researchers, who address flaws in the interpretation of Landauer’s argument and question its validity. The similarities between the results obtained by our approach and LP begs the question whether our methodology too is vulnerable to such criticisms.

In this chapter, we first provide an historical review of studies that play a significant role in the field of thermodynamics of computation. We focus our attention to emergence of LP and consecutive research that take Landauer’s work further, and address some foundational questions surrounding its validity. We study some key arguments and discuss their implication for our methodology. Certain concepts presented in this chapter do not have immediate consequences for the circuits we are interested in, however, we address them to provide insights into the interdisciplinary nature of thermodynamics of computation.

5.1 A Historical Review Towards Landauer’s Principle

In this section, we provide an overview of some key results that play a crucial role in emergence of thermodynamics of computation, and led to and motivated this

dissertation. The cast of important characters¹ and their contribution are chronologically listed in Table. 5.1. A set of researchers listed here play a key role in the inception and refinement of LP, we elaborate further on their studies in the following sections of this chapter.

Time	Event
1703	Mathematician and philosopher Gottfried Leibniz published “Explication de l’arithmétique binaire” and presented an explanation of the modern binary system. The paper appeared in “Memoires de l’Academie royale des sciences” in 1705 [47].
1833	Michael Faraday recorded the first semiconductor effect; he discovered that the electrical conduction increases with temperature in silver sulfide crystals, which is the opposite to that observed in copper and other metals [48].
1847	Mathematician George Boole published “The mathematical analysis of logic” where he introduced Boolean Logic. This played an essential role in the further development of the modern binary system by Claude Shannon in later years [49].
1850	Physicist Rudolf Clasius published “On the mechanical theory of heat” introducing the second law of thermodynamics. The validity of this law has become the subject of arguably more argument than any other theory in physics [50].

¹As much as we admire and appreciate the genius of individuals listed here, we also acknowledge that these achievements are collaborative work of numerous scientists and thinkers spanning over many decades, and listed dates and people are rather symbolic milestones.

- 1867 Physicist James Clerk Maxwell wrote a letter to Peter Guthrie Tait in which he talked about the Maxwell's Demon thought experiment for the first time. The idea is presented to public in Maxwell's Theory of Heat in 1871 [51].
- 1865 Physicist Rudolf Clausius coined the term *entropy* [52].
- 1872 Physicist Ludwig Boltzmann proposed H-theorem, and with it the formula $\Sigma p_i \log p_i$ for the entropy of a single gas particle [53].
- 1878 Physicist J. Willard Gibbs defined the Gibbs entropy: the probabilities in the entropy formula are now taken as probabilities of the state of the whole system [54].
- 1927 Mathematician and polymath John von Neumann defined the von Neumann entropy, extending the classical notion of entropy to the field of quantum mechanics [55].
- 1928 Electronics researcher Ralph Hartley introduced Hartley information as the logarithm of the number of possible messages, with information being communicated when the receiver can distinguish one sequence of symbols from any other (regardless of any associated meaning) [56].
- 1929 Physicist Leo Szilard analyzed Maxwell's Demon, showing how the Szilard engine can transform information into the extraction of useful work [57].
- 1936 Mathematician and computer scientist Alan Turing introduced Turing machines as a thought experiment representing a computing machine in "On computable numbers, with an application to the Entscheidungs problem" [58].

- 1937 Electronic engineer and mathematician Claude E. Shannon invented digital electronics as his Master's thesis entitled "A symbolic analysis of relay and switching circuits" [59].
- 1947 Physicists William Shockley, John Bardeen and Walter Brattain invented the first transistor at Bell Laboratories [48].
- 1948 Claude E. Shannon published "A mathematical theory of communication." This paper became one of the most influential works of the Information Theory field [25].
- 1950 The first of London Symposiums on Information Theory is organized where prominent scientist discussed the quantification of information. Later, physicist Donald MacKay published his "Information, Mechanism and Meaning" book where he made some of the studies from some of these symposiums available and explained the history of how scientists started to think of information separate from meaning [60].
- 1956 Physicist Leon Brillouin in his "Science and Information Theory" book expressed that the Clausius' entropy and Shannon's entropy are identical [61].
- 1958 The precursor ideas to the integrated circuits (ICs) were improved sufficiently and the revolutionary design of the ICs emerged.²
- 1961 Physicist Rolf Landauer published his "Irreversibility and heat generation in the computing process" paper in which he presented the LP [4].

²There is no consensus over who invented the ICs. Jack Kilby, Kurt Lehovec, Robert Noyce and Joerni Hoerni are all credited for their contribution. Kilby was awarded the Nobel Prize in Physics in 2000 "for his part in the invention of the IC" [62].

- 1965 Gordon Earle Moore published his infamous paper describing Moore’s Law. His conjecture successfully predicted the trend in the increase in number of transistors on ICs and guided the long term strategies in semiconductor technology until the early years of this century [1].
- 1982 Physicist Charles Bennett exorcised Maxwell’s Demon with “Thermodynamics of computation - a review” in which he clarified the basis of unavoidable dissipation in computation, and showed that it is not the measurement but erasure that cannot be done reversibly [63].
- 2020 Projected time for the end of scaling for the current CMOS technology. The total energy dissipated per irreversible binary decision, continued at the historical rate observed for silicon technology, will reach Landauer’s lower bound on the physical cost of logical irreversibility alone ($\sim k_B T$) [5].

Table 5.1: Chronological list of significant events in the evolution of thermodynamics of computation.

The chronological list of events provides us with perspective on the multidisciplinary nature of thermodynamics of computation. It also allows us to see how recently this field has emerged, and that the fundamental concepts and formulations constituting the very core of this field are still being refined and improved to apply to emerging technologies. Among various studies, we now briefly focus on the Szilard engine that serves as a tractable model to allow thermodynamic analysis of information and memory, and later as an application for LP.

5.2 Szilard Engine and Landauer's Exorcism of Maxwell's Demon

In Oxford English Dictionary [64], the definition of Maxwell's Demon is given as "an entity imagined by Maxwell as allowing only fast-moving molecules to pass through a hole in one direction and only slow-moving ones in the other direction, so that if the hole is in a partition dividing a gas-filled vessel, one side becomes warmer and the other cooler, in contradiction of the second law of thermodynamics." In 1867, James Clerk Maxwell wrote a letter to Peter Guthrie Tait in which he talked about the this hypothetical intelligent being and the associated thought experiment for the first time. The idea is presented to public in Maxwell's Theory of Heat in 1871. It was William Thompson who referred to this imaginary being as "Maxwell's intelligent demon" in 1874 [51]. The association between entropy and information was loose in those days given that Shannon's "A mathematical theory of communication" [25] paper had not yet come into existence. Therefore, Maxwell himself did not discuss the connection between the Second Law and information explicitly. Although, it was implied by his definition of demon's capability of using information to lower information entropy [20]. As a practical example, Maxwell's demon problem hints at the need for unified theory of thermodynamics and information.

In 1929, Leo Szilard analyzed the implications of Maxwell's demon to the Second Law. Szilard proposed an engine that explains the workings of the Maxwell's thought experiment [57]. This engine is depicted in Fig. 5.1. As shown in the figure, the position of the particle is not known initially. The demon measures the location of the particle and with that information brings the box into contact with the heat bath and allows the particle to do work on a piston with the thermal energy it gains from the bath. The piston is then removed, and the demon can start over. It then appears as if the demon is extracting heat from the bath and converting it into work,

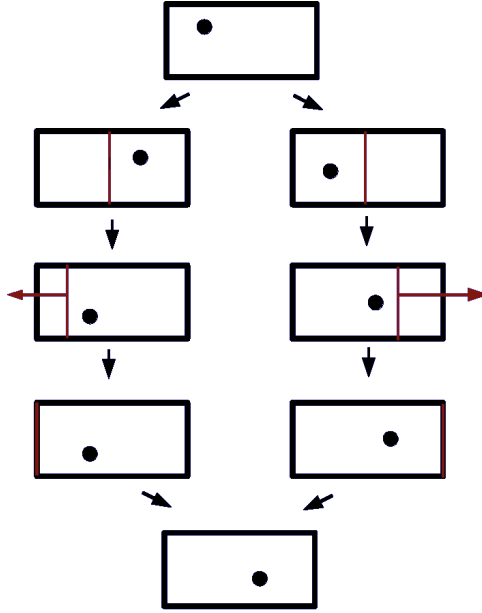


Figure 5.1. Schematics of the Szilard’s engine [65].

thus decreasing the entropy of the bath with no corresponding increase in entropy anywhere else, which threatens the Second Law.

The ingenuity of Szilard’s engine comes from its ability to turn the Maxwell’s demon problem into a “binary” decision process. This binary decision making process requires information acquisition, memory and subsequent information erasure which are necessary to couple the piston and thermal bath. It is important to note that Szilard did not identify the demon’s specific role as saving the second law but rather simply regarded *memory* as an important feature in a demon’s operation [51]. Szilard’s design has fundamentally influenced the way we think about entropy, and Maxwell’s demon led to the concept of a “bit” of information and to key concepts in information theory.

In 1982, Bennett [63] provided a clarification to the basis of unavoidable dissipation. He showed that the fundamentally dissipative step in Demon’s process is actually not the measurement. The measurement *can* be done reversibly. What is unavoidably dissipative is the logically irreversible erasure of Demon’s memory which is required

to make room for new measurements. This erasure cannot be done reversibly, and hence is unavoidably dissipative. It is based on this definition that LP could be used to exorcise Maxwell’s Demon and save the second law of thermodynamics.

Szilard’s unique contribution shed light on the relation between the laws of thermodynamics and theories regarding the physicality of information. However, this study is based on a simple information processing system composed of two chambers; it does not address the technologically relevant question regarding the effect of information erasure on the dissipative cost in actual electronic circuit examples that can be used in computation.

5.3 On the Validity of Landauer’s Principle

Landauer’s pioneering work has set the stage for numerous foundational studies on the relation between information and energy in the last half-century. The majority of research is theoretical in nature [24], [26], [66-71], however, recently the advancements in science with the emergence of nanotechnology allow for experimental work concerning the relation between information and thermodynamics [72-74].³ In this section, we outline a set of studies regarding LP and its validity. We objectively present the arguments made in each study to the best of our ability. In the next section, we will comment on their implications for our approach.

In the most general sense, the point raised by researchers who argue against the validity of LP is based on certain issues concerning the interpretation of thermodynamics. From an information theoretic point of view, when we consider storing a bit of information in a macroscopic thermodynamic system, thermodynamics cannot tell us which of the possible states is used to store that bit. In the microscopic sense one can interpret the storing of a bit as occupation of any of the available states or

³Please note that, the studies are grouped based on their theoretical and experimental nature, as opposed to their stance towards – for or against – the validity of LP.

even can assume that each of these states are occupied with a certain probability. Common point raised by both Shenker [66] and Norton [67] is the issue of properly treating the macroscopic representations of information-representing states. They argue against the use of macroscopic variables as the probability weight magnitude of thermodynamic physical states. The use of probability distributions over macroscopically distinct states, such as the thermodynamic entropy of a system, can only be accurately defined by considering the region of state space accessible to the given microscopic state. The physical representations of logic states assume that a physical system cannot jump from one logic state to another. This means that certain regions of state space representing different logic states are not accessible to others. Shenker refers to this as “interaccessibility” of states.

Furthermore, Norton [67] argues against the assumption that bits with random values correspond to an ensemble. He states that LP “depends on the incorrect assumption that memory devices holding random data occupy a greater volume in phase space and have greater entropy than when the devices have been reset to default data.” Shenker [66] also points out a similar problem and suggests that “the collection of cells carrying random data is being treated illicitly as a canonical ensemble.”

Another interesting point about LP is raised by Bacon [68]. Bacon suggests that LP “treats the two macrostates 0 and 1 in a very error-free manner” and that nothing is perfectly digital therefore it is inaccurate to treat these variables as digital values. To support his argument he gives the example of information stored in the hard drive which are not only composed of 1s and 0s but also include some small fluctuations for each bit. This fact is at odds with the statement that two bit states should not be accessible to each other. Bacon makes a valid argument by drawing attention to the idealization made about the true digital nature of information for computers.

Aside from these points, Maroney [69] too raised an argument concerning the interpretation of LP. The author drew attention to absence of a relationship between ther-

modynamic and logical irreversibility and provided further derivations to generalize LP. Later, he expanded the formulation to [70] “input and output states represented by macrostates with variable mean internal energies, entropies and temperatures and took in the account consequences for a larger group of logic operations than the reset operation” [65] and pointed to the restrictions on LP. The derivation provided by Turgut [71] can be used to address these restrictions. Turgut obtained results using the Gibbs microcanonical distribution approach, which provides a stronger constraint over LP without the need to consider probability distributions over the input logic states.

In addition to the foundational studies presented above, experimental studies concerning LP has also emerged in the recent years. Berut *et. al.* [72] reported a measurement of small amount of heat release, when an individual bit of data is erased, and demonstrated the relation between information and energy dissipation. Bokor *et. al.* [73] too obtained a similar result using magnetic memories. There are also studies where the measured dissipation for information erasure is less than $k_B T \ln(2)$ (Eg. Snider *et. al.* [74]). In order to accurately interpret the outcome of these studies, a clear and consistent definition of information is required. Overall, majority of studies concerning the relation between information and energy do not provide a definition for the terms associated with information erasure.⁴ The question “information about what?” becomes pivotal in interpreting the outcome of studies concerning LP. In Snider’s work, for instance, the erasure operation is done with a presence of a copy of the information, which means that information is not irreversibly lost; i.e. no unavoidable dissipation cost is associated with this operation. In the next section, we elaborate further on the distinction between our approach and LP and comment on the implications of studies listed here for our approach.

⁴As we mentioned in Sec. 1.4, Landauer himself did not provide a definition for ΔI_{er} in Eq. (1.4) and (1.5).

5.4 On the Validity of Our Methodology

The referential approach serves a departure point for our methodology. As we noted in Chapter 1, this approach and LP are fundamentally different in nature, however, the resulting bounds resonate under certain conditions. The definition of the concept of *information* plays a crucial role in the distinction between our approach and that of Landauer’s. In the referential approach, the information⁵ is defined in relation to an input referent, unlike LP where information corresponds to self entropy (Eq. (1.4) and (1.5)). This distinction is the key in interpreting the unavoidable energy cost associated with information erasure as demonstrated in the experimental studies. In Snider’s paper erasure of information is performed with a presence of the initial copy of the input, i.e. the correlation between the circuit and the input referent is not lost. Therefore, the calculated values for the information-theoretic dissipation are less than $k_B T \ln(2)$ [40]. This can be accurately captured by the referential approach since such erasures are indeed do not lead to information loss and hence do not have an associated unavoidable dissipation cost.

In order to further elaborate on the distinction between our methodology and LP, we refer to Anderson’s study based on the referential approach [75]. Anderson emphasizes that the results obtained by using the referential approach

“... follow from the specified forms of the global system states at specific times throughout the cycle, unitary Schrödinger dynamics, established entropic inequalities, energy conservation and the assumption that the bath, \mathcal{B} , is in a thermal state characterized by temperature T at the beginning of each process step.”

⁵Recall that in Sec. 1.6, we defined the referent as a physical system that unambiguously holds the input data throughout computation. The referent allows us to regard information erasure as loss of correlation between the state of an erasable quantum system and the initial input; i.e. it distinguishes the erasure of information from local changes in the structure of the erasable system’s state [8].

It is important to underline that, in our methodology, the state of the information bearing system, as well as the state of the surrounding subsystems and the interactions between them are classified and obtained by using quantum dynamics. This makes our approach immune to the majority of arguments criticizing the foundations of LP (eg. Shenker [66] and Norton [67]) based on a classical mechanics and classical thermodynamics.⁶ In order to thoroughly address the extent to which criticism raised along these lines apply to our approach, a common language and consistent definition of key terms need to be developed.

Furthermore, in order to address Bacon’s argument, we emphasize that our methodology is applied to information processing artifacts where the inputs are represented by orthogonal states⁷. In addition, the idealized circuit operation consider for the artifact excludes any affect of error or noise in the fundamental bounds. Due to the orthogonal and error-free nature of information states our methodology is not subject to the criticism raised in [68]. The methodology presented here can be expanded to address the fundamental energy requirements of noisy computation [9], however, such modification is beyond the scope of this work. We conclude that based on the functional features of the information processing artifact considered in our approach, the argument made by Bacon does not have an implication for our methodology in its current form.

In conclusion, a brief survey of studies concerning information processing and the associated energetic cost suggests that the referential approach provides us with a strong basis to determine the fundamental lower bounds on energy at the circuit level. The criticisms leveled against the validity of LP have no direct implications

⁶Hemmo and Shenker [76], in their book chapter on erasure, argue that LP cannot be obtained by using the principles of classical mechanics. The referential approach, however, logically follows from the principles of quantum dynamics.

⁷Eg., recall Sec. 3.1.1, where we employed six-dot quantum cells to represent the logic-0, logic-1 and null states orthogonally.

for our methodology. However, as we expand the breadth of our methodology to accommodate the treatment of a wider range of nanocomputing proposals, we will continue to pay attention to such counter arguments and persistently scrutinize the foundations of our approach.

CHAPTER 6

CONCLUSION

In this dissertation, we proposed a methodology for determining fundamental lower bounds on the dissipative energy cost of logically irreversible computation, which can be tailored to specific nanocomputing paradigms. The bounds we evaluated derive solely from the underlying computational strategy employed in a given circuit and the physical laws that govern its operation. They represent the fundamental physical limits for a given technology base which can be obtained by analyzing the connection between the logic operations and the physics of information processing carried out by the circuit.

In Chapter 1, we provided a technical background for concepts, theories, and laws that govern the physical interactions required for computation and are required for the determination of energy dissipated to erase information irreversibly. The physical cost of information loss is often studied independent of thermodynamic processes that take place as a result of the erasure. To this date, a number of studies have been pursued to shed light on the relation between the laws of thermodynamics and theories regarding the physicality of information. However, these studies are all based on simple information processing systems composed of pistons and chambers; they do not address the technologically relevant question regarding the effect of information erasure on the dissipative cost in actual electronic circuit examples that can be used in computation.

In Chapter 2, we laid out the foundations of our methodology for establishing dissipation bounds for arbitrary circuits. We introduced the theoretical constructs

of our approach in the most general form, independent of a given nanocomputing technology. We introduced mathematical representations and dynamic assignments that allow us to tailor the fundamental bounds for a nanocomputing paradigm. We accomplished this by bringing physical laws to bear directly on the underlying computational strategy that defines the paradigm.

In Chapter 3, we illustrated applications of our methodology to non-transistor- (QCA half adder) and transistor-based (NASIC half and full adders, np-CMOS half adder) circuits. The lower bounds we obtained are tailored for specific features of the underlying computational paradigm. Our results show that the fundamental energy requirements of a given technology depend not only on the physical structure but also on the details of the circuit operation. This kind of technology-dependent analysis can be of paramount importance in developing long-term design strategies for potential CMOS replacement technology proposals. We also analyzed an np-CMOS half adder from a fundamental energy requirement point of view, which stands out as an academic endeavor to illustrate an application of our methodology via a widely-known conventional circuit structure. In addition to these dynamically clocked circuits, we also discussed application to static circuits, and certain limitations of our methodology. We showed that the analysis of static circuits is tractable within our approach, however, the methodology needs to be expanded to accommodate this computing strategy, which lies beyond the scope of this work.

In Chapter 4, we presented a modularized approach to accommodate the treatment of large and complex circuits. We discussed the scalability of our methodology to multi-tile circuits by means of modular analysis, and compared our results with the calculations presented in Chapter 3, where we addressed the problem of determining fundamental heat dissipation bounds associated with irreversible logic transformations in a single tile computation. Our study of the modular analysis showed that details of the circuit design play a significant role in localizing the information loss

that allows us to obtain the fundamental bounds. We illustrated our modular approach via applications to QCA and NASIC circuits, and demonstrated that modular analysis of QCA requires us to specify design rules, whereas the NASIC structure is inherently premodularized. The approach presented here allows us to facilitate and possibly automate the determination of fundamental dissipation bounds for large, complex circuits designed according to specified rules.

In Chapter 5, we provided a brief and qualitative background on thermodynamics of computation, and a list of arguments directed against the validity of LP. This chapter provides an overview of the disputes concerning the study of irreversible information processing and associated energy cost, as well as the foundations of thermodynamics of computation. We discussed the extent to which the arguments against LP have implications for our methodology. We concluded that the referential approach provides a strong basis to study fundamental lower bounds on energy dissipation at the circuit level, and makes our methodology immune to the criticism leveled against the validity of LP.

We provided guidelines on how to obtain the lower bounds on the fundamental energetic cost of computation at the circuit level for post-CMOS nanocomputing technology proposals. This was done by juxtaposing the energetic cost of irreversible information processing and physical operations that take place as a result of it. The bounds resulting from our approach are truly fundamental and they depend only on the circuit structure, clocking scheme, and temperature of the circuit's environment. Therefore, implementation-specific quantities (such as the kink energy for QCA cells, or parasitics for the NASICs) do not appear in these energy bounds. Combining results from such analyses with assumptions on circuit scale and clock rate for specified circuits, lower bounds on areal heat dissipation can be obtained at any desired computational throughput for arbitrary circuits. These fundamental dissipation bounds can serve as tools for the assessment of proposed post-CMOS nanocomputing

technologies. They can be used to verify the fundamental compatibility of speed and density projections (e.g. those required to outperform ultimate CMOS) with assumptions regarding the capacity for heat removal, which will become crucial for emerging technologies.

The methodology we presented here can be expanded to increase the breadth of circuit structures and operations that can be treated by using our approach. In Chapter 3, we commented on the limitations of our methodology in obtaining fundamental lower bounds on energy for static transistor-based circuits and dynamically clocked transistor-based circuits that do not have output complements. The abstraction can be generalized to accommodate a wider range of subsystem interactions as present in static transistor-based circuits. In addition, the relation between the two terms in the general fundamental lower bound for the transistor-based circuits, the information-theoretic cost and particle supply cost, can be generalized. This would allow for the treatment of circuits in which the number of particles and the associated transport cost for each input depends on the probability of that input as is the case for circuits without output complements. The analysis of these circuits, as well as other paradigms, can be tractable within our methodology if the necessary modifications are made on the present formulation.

We emphasize that the viability of any nanocomputing technology proposal will hinge on a complex mix of various physical, technological, and economic factors. The fundamental dissipation bounds provide us one necessary but insufficient condition regarding the performance limits of a proposed technology; the physical possibility that specified performance targets can be met under best case assumptions regarding circuit fabrication and computational control. These lower bounds can be used as a “litmus test” in nanocomputing technology assessment, providing a check for consistency of performance projections with assumed resources and fundamental physical constraints. Analyzing the thermodynamic limits of a nanocomputing paradigm pro-

vides insights into how far it can be improved in principle and how much room there is at the bottom.¹

The emergence of nanoscale electronics brings about a new era in the evolution of computer processors where the Moore’s law remains arrears in predicting the trajectory of computation. Proposals for a candidate successor technology face a broad range of disparate concerns at practical and conceptual levels where overcoming these challenges will require the involvement of a diverse community of researchers. The post-CMOS nanocomputing paradigms are expected to operate near the energetic limits of what nature will allow, therefore the fundamental performance limits are likely to be of key importance in assessing the viability of a proposed successor technology. The evolution of the field will depend on breakthroughs in nanoelectronics and emerging trends in computing with implications for nanoscale computation. By pursuing this research we hope to have contributed to the performance assessment tools that are essential for determining the course of post-CMOS computing. As Patterson [78] suggests, the endeavors on the path to the future of computation is much like a sports game full of surprises and uncertainties, and “no matter how the ball bounces, it’s going to be fun to watch, at least for the fans. The next decade is going to be interesting.”

¹Feynman’s lecture [77], “There is plenty of room at the bottom,” inspired the conceptual foundations of nanotechnology, and is considered to be a seminal work in the field. As we approach the end of scaling, however, it is a good time to ask “how much room there is at the bottom?”

APPENDIX

GRANULARITY OF HEAT DISSIPATION ANALYSIS FOR THE QCA HALF ADDER

We assume that the circuit inputs are equally probable, i.e. for the input vector $x_i = \{00, 01, 10, 11\}$ the associated probabilities are $p_i = \{\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\}$. Below, based on this assumption, we map the transition probabilities for the complete circuit, as well as for each gate and cell.

Circuit Level

In the circuit level, the inputs map out to the outputs as shown in Table. A.1.

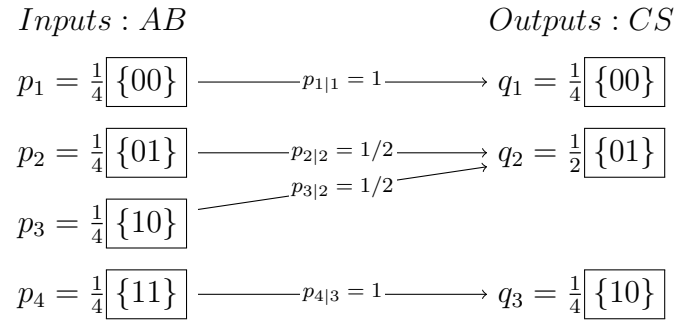


Table A.1. Transition probabilities for a 1-bit half adder.

Based on the probabilities outlined above, the Shannon self entropy of the input set and the output set are

$$H(X) = - \sum_{i=1}^4 p_i \log_2 p_i = -\frac{1}{4} \log_2 \left(\frac{1}{4} \right) \times 4 = 2bits. \quad (\text{A.1})$$

$$H(Y) = - \sum_{j=1}^3 q_j \log_2 q_j = -\frac{1}{4} \log_2 \left(\frac{1}{4} \right) \times 2 - \frac{1}{2} \log_2 \left(\frac{1}{2} \right) = 1.5bits. \quad (A.2)$$

And the information loss associated with the logic operation is

$$H(X|Y) = -q_j \sum_{j=1}^3 p_{i|j} \log_2 p_{i|j} = -\frac{1}{2} \times \frac{1}{2} \log_2 \left(\frac{1}{2} \right) \times 2 = 0.5bits. \quad (A.3)$$

Therefore, for the circuit level analysis, the dissipation associated with a computational cycle of a half adder is $\boxed{0.5 \times k_B T \ln(2)}$.

Data-Zone Level

Data-Zone level analysis assumes that certain zones of the circuit which get activated simultaneously at a given clock step form a data-zone and that the total dissipative cost of a computational cycle of the QCA half adder is equal to the sum of individual dissipations associated with these zones. Based on this division, first zone corresponds to the NAND gate, second zone corresponds to the two AND gates with inputs A , B , and M , and outputs N_1 and N_2 , the third zone corresponds to the AND and OR gates with inputs M , N_1 and \bar{N}_2 , and outputs S and C .

Accordingly, the dissipation at the first zone is the dissipation of an individual NAND gate, i.e. $H(X|Y) = 1.1887bits$. The dissipation in the second zone is calculated based on the probability distribution of the three input two output logic operation shown in Table. A.2 Hence, the information loss associated with the second data zone logic operation is $H(X|Y) = 0.5bits$. Finally the dissipation in the third zone is calculated similar to the second data zone and the information for the four input two output logic operation is obtained, i.e for the third data zone $H(X|Y) = 0.5bits$. In total, based on the data-zone level analysis, dissipation throughout a computational cycle of the QCA half adder is $\boxed{2.1887 \times k_B T \ln(2)}$.

Data-Subzone Level

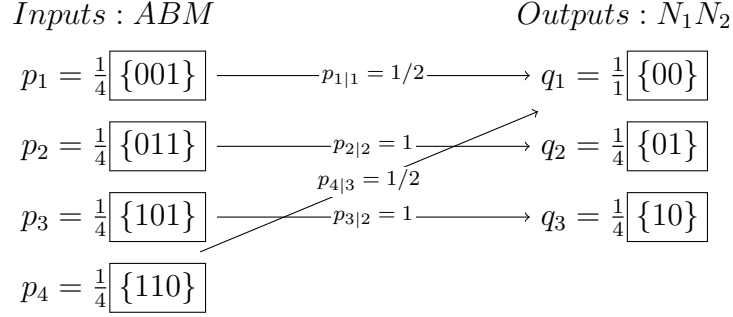


Table A.2. Transition probabilities for the second data zone of the QCA 1-bit half adder operated under Landauer clocking.

Data-subzone level analysis assumes that each gate within the circuit dissipates heat independent from one another, therefore the total dissipative cost of a computational cycle of the QCA half adder is the sum of dissipations associated with each gate. Above, the dissipation of a NAND gate is given as $H(X|Y) = 1.1887bits$. For the two AND gate with outputs N_1 and N_2 the probability distribution can be obtained as shown in Table A.3.

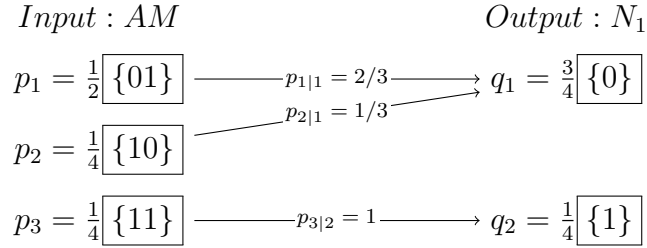


Table A.3. Transition probabilities for the AND gate with inputs AM and output N_1 .

Based on the above probability distribution the information loss associated with this operation is $H(X|Y) = 0.6887bits$. The dissipation from the second AND gate with input BM and output N_2 is also $H(X|Y) = 0.6887bits$. For the OR gate with output S the probability distribution is given in Table A.4.

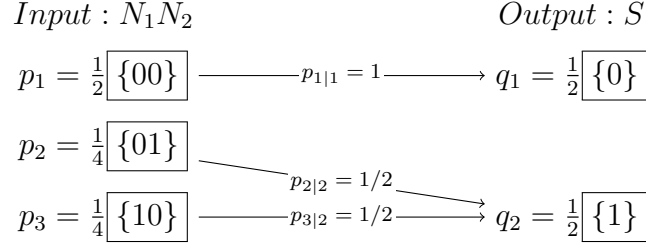


Table A.4. Transition probabilities for the OR gate with inputs N_1N_2 and output S .

Hence, the information loss in this operation, based on the above probability distribution, is $H(X|Y) = 0.5bits$. For the AND gate with output C the probability distribution is given in Table A.5.

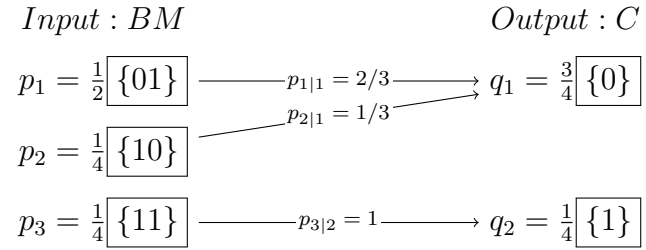


Table A.5. Transition probabilities for the AND gate with inputs BM and output C .

Therefore, the information loss is $H(X|Y) = 0.6887bits$. In total, based on the data-zone level analysis, dissipation throughout a computational cycle of the QCA half adder is $\boxed{3.7546 \times k_B T \ln(2)}$.

Cell Level

Cell level analysis corresponds the summation of the dissipative cost of erasing the memory of input from each cell, which corresponds to the sum of self entropy of each cell in the circuit. Each of the inputs A and B occur with equal probabilities, i.e. $x_i = \{0, 1\}$ with the associated probabilities being $p_i = \{\frac{1}{2}, \frac{1}{2}\}$, i.e. $H(X) = 1bit$.

Based on this assumption there are 23 cells that hold the *1bit* memory of A and 42 cells that hold the *1bit* memory of B. The self entropy of the 4 cells carrying the information regarding S is also *1bit*. As for the rest of the cells, we assume that the probability distribution of the cells at the gate regions change at the 'center' cell of that gate, i.e. the self entropy of the cells change after the center cell of the AND gate with the output M . The self entropy of these cells is calculated from $x_i = \{0, 1\}$ with the associated probabilities $p_i = \{\frac{1}{4}, \frac{3}{4}\}$ and obtained to be $H(X) = 0.8113\text{bits}$. The probability distribution is the same for cells carrying the information regarding N_1 , N_2 , \bar{N}_2 and C . Therefore, in total, There are 69 cells in the circuit carrying 1-bit memory, and 56 cells carrying 0.8113 bit, based on the cell level analysis, dissipation throughout a computational cycle of the QCA half adder is $\boxed{114.4328 \times k_B T \ln(2)}$.

Conclusion

The calculations presented above outlines various interpretations of the fundamental heat dissipation bounds for a QCA half adder circuit. The three orders of magnitude discrepancy between the circuit and cell level analysis underlines the significance of identifying the accurate level of dissipation analysis for a given circuit. The lower bound on unavoidable energy dissipation obtained by using all four levels of analyses for the QCA circuit is depicted in Fig. A.1 on the next page. The signature feature of our methodology is that it allows us to identify the irreversibility based on the structure and operation of the information processing artifact and assign the accurate level of analysis for the associated unavoidable energy dissipation.

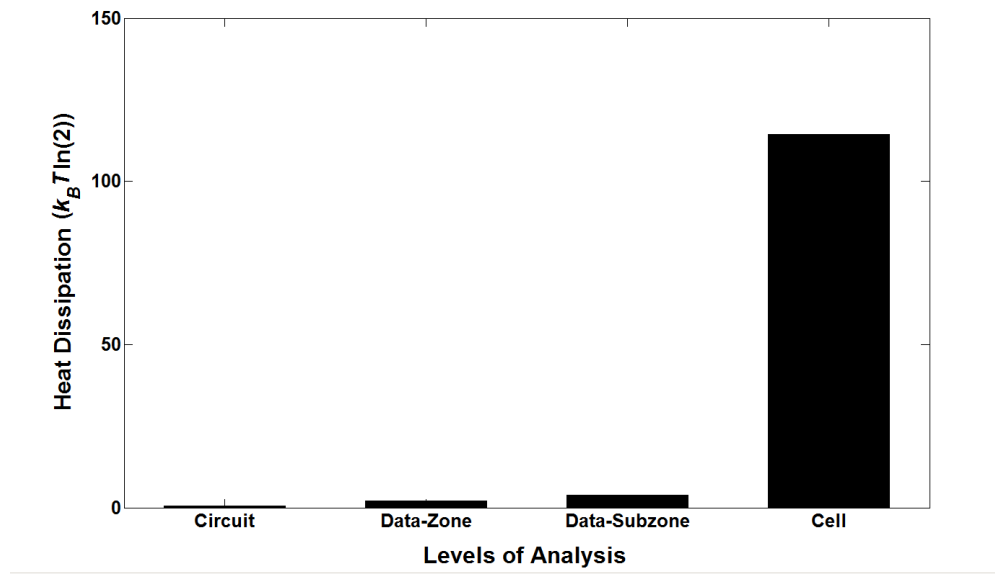


Figure A.1. The lower bound value for the energy dissipation obtained by using circuit, data-zone, data-subzone and cell levels of analyses for QCA 1-bit half adder operated under Landauer clocking.

BIBLIOGRAPHY

- [1] G. E. Moore, “Cramming more components onto integrated circuits,” *Electronics*, pp. 114-117, April 19, 1965.
- [2] R. Courtland, “The status of Moore’s Law: It’s complicated,” *IEEE Spectrum*, October, 2013.
- [3] R. Courtland, “The end of the shrink,” *IEEE Spectrum*, November, 2013.
- [4] R. Landauer, “Irreversibility and heat generation in the computing process,” *IBM J. Res. Dev.*, vol. 5, pp. 183-191, 1961.
- [5] R. W. Keyes, “Physical limits of silicon transistors and circuits,” *Rep. Prog. Phys.*, vol. 68, pp. 2701-2746, 2005.
- [6] T. Sagawa, *Thermodynamics of information processing in small systems*, Springer Thesis, 2013.
- [7] M. H. Partovi, “Quantum thermodynamics”, *Phys Letts A* vol. 137, no. 9, pp.440-444, 1989.
- [8] N. G. Anderson, “Information erasure in quantum systems,” *Phys. Lett. A*, vol. 372, pp. 5552-5555, 2008.
- [9] N. G. Anderson, “On the physical implementation of logical transformations: Generalized L -machines,” *Theoretical Computer Science*, vol. 411, pp. 4179-4199, 2010.
- [10] İ. Ercan and N.G. Anderson, “Heat dissipation in nanocomputing: Theory and QCA application,” *Proceedings of the 11th IEEE Conference on Nanotechnology (IEEE NANO, 2011)*, pp.1289-1294, 2011.
- [11] C. A. Moritz, P. Narayanan, and C. O. Chui, “Nanoscale application specific integrated circuits,” in *Nanoelectronic Circuit Design*, Niraj K. Jha and Deming Chen, Eds. New York, Springer, pp. 215-275, 2011.
- [12] İ. Ercan, N. G. Anderson, “Heat dissipation in nanocomputing: Lower bounds from physical information theory,” *IEEE Transactions on Nanotechnology*, Vol. 12, Issue 6, pp. 1047 - 1060, 2013 (doi: 10.1109/TNANO.2013.2276938).

- [13] İ. Ercan, M. Rahman, and N. G. Anderson, “Determining fundamental heat dissipation bounds for transistor-based nanocomputing paradigms,” *NANOARCH 11, Proceedings of the 2011 IEEE/ACM International Symposium on Nanoscale Architectures*, pp. 169-174, 2011.
- [14] İ. Ercan, and N. G. Anderson, “Modular dissipation analysis for QCA,” presented at *FCN’13: The 2013 Workshop on Field-Coupled Nanocomputing*, Tampa, FL, February 7-8 2013. In *Field-Coupled Nanocomputing*, N.G. Anderson and S. Bhanja. Eds. (Lecture Notes in Computer Science, Vol. 8280). Heidelberg: Springer (forthcoming).
- [15] W. Greiner, L. Neise and H. Stocker, *Thermodynamics and statistical mechanics*, Springer-Verlag New York Inc. 1995.
- [16] L. Sklar, *Physics and chance: Philosophical issues in the foundations of statistical mechanics*, Cambridge University Press, p.19, 1995.
- [17] N. Gershenfeld, “Signal entropy and thermodynamics of computation”, *IBM Systems Journal*, vol. 35, no.s 3&4, 1996.
- [18] L. D. Landau and E. M. Lifshitz, *Statistical physics*, 2nd Revised English Edition, Pergamon Press, 1978.
- [19] E. Garber, S. G. Brush and C. W. F. Everitt, *Maxwell on heat and statistical mechanics: On "avoiding all personal enquiries" of molecules*, Bethlehem: Lehigh University Press; London Associated University Presses, p. 205, 1995.
- [20] A. Ben-Naim, *A farewell to entropy: Statistical thermodynamics based on information*, World Scientific, 2008.
- [21] M. B. Plenio and V. Vitelli, “The physics of forgetting: Landauer’s erasure principle and information theory,” *Contemporary Physics*, vol. 42, no. 1, pp. 25-60, 2001
- [22] G. Piccinini and A. Scarantino, “Information processing, computation, and cognition,” *Journal of Biological Physics*, vol. 37.1, pp. 1-38, 2011.
- [23] C. H. Bennett, “Demons, engines and the second Law” *Scientific American*, vol. 257, 108-116, 1987.
- [24] C. H. Bennett, “Notes on Landauer’s principle, reversible computation, and Maxwell’s demon,” *Studies in History and Philosophy of Modern Physics*, vol. 34, pp. 501-510, 2003.
- [25] C. E. Shannon, “A mathematical theory of communication,” *Bell Syst. Tech. J.*, vol. 27, pp. 379-423, 623-656, 1948.
- [26] J. Ladyman, S. Presnell, A.J. Short, and B. Groisman, “The connection between logical and thermodynamic irreversibility,” *Studies in History and Philosophy of Modern Physics*, vol. 38, pp. 58-79. 2007.

- [27] J. Ladyman, "What does it mean to say that a physical system implements a computation?," *Theoretical Computer Science*, vol. 405, pp. 376-383, 2008.
- [28] C. S. Lent, P. D. Tougaw, W. Porod, and G. H. Bernstein, "Quantum Cellular Automata," *Nanotechnology*, vol. 4, pp. 49, 1993.
- [29] P. Tougaw, and C. S. Lent, "Logical devices implemented using quantum cellular automata," *J. Appl. Phys.*, vol. 75, pp. 1818-1825, 1994.
- [30] N. G. Anderson, F. Maalouli, and J. Mestancik, "Computational efficacy measures for nanocomputing channels," *Nano Communication Networks*, vol. 3, pp. 139-150, 2012.
- [31] C. S. Lent, M. Liu and Y. Lu, "Bennett clocking of quantum-dot cellular automata and the limits to binary logic scaling," *Nanotechnology*, vol. 17, 4240-4251, 2006.
- [32] N. G. Anderson, "Reversible computation via Bennett vlocking in QCA circuits: Input-output requirements", *Proceedings of the 2009 International Workshop on Quantum-Dot Cellular Automata (IWQCA)*, pp. 12-13, 2009.
- [33] M. Ottavi, S. Pontarelli, E. DeBenedictis, A. Salsano, P. Kogge and F. Lombardi, "High throughput and low power dissipation in QCA pipelines using Bennett clocking," *6th IEEE/ACM International Symposium on Nanoscale Architectures*, Anaheim, CA, June 17-18, 2010.
- [34] E. P. Blair, E. Yost, and C. S. Lent, "Power dissipation in clocking wires for clocked molecular quantum-cellular automata," *J. Comput. Electron.*, vol. 9, pp. 49-55, 2009.
- [35] K. Stearns and N. G. Anderson, "Throughput-dissipation tradeoff in partially reversible nanocomputing: A case study," *Proceedings of the 2013 IEEE/ACM International Symposium on Nanoscale Architectures*, 101-105, 2013.
- [36] P. Narayanan, M. Leuchtenburg, T. Wang, and C. A. Moritz, "CMOS control enabled single-type FET NASIC," *IEEE Computer Society Annual Symposium on VLSI*, April 2008.
- [37] P. Narayanan, C. A. Moritz, K. W. Park and C. O. Chui, "Validating cascading of crossbar circuits with an integrated device-circuit exploration," *NANOARCH '09: Proceedings of the 2009 IEEE/ACM International Symposium on Nanoscale Architectures*, pp. 37-42, 2009.
- [38] J. M. Rabaey, A. Chandrakasan and B. Nikolic, *Digital integrated circuits*, Prentice Hall 2003.
- [39] N. H. E. Weste, D. M. Harris, *CMOS VLSI design: A circuits and systems perspective*, 4th Edition, Addison-Wesley, 2010.

- [40] N. G. Anderson, “Overwriting information: Correlations, physical costs, and environment models,” *Phys. Lett. A*, vol. 376, pp. 1426-1433, 2012.
- [41] V. V. Zhirnov and R. K. Cavin III, “Scaling beyond CMOS: Turing-Heisenberg rapprochement,” *Solid-State Electronics*, vol. 54, pp. 810-817, 2010 (and references therein).
- [42] I. Hanninen and J. Takala, “Irreversibility induced density limits and logical reversibility in nanocircuits,” *Proceedings of the 2012 IEEE/ACM Symposium on Nanoscale Architectures*, pp. 50-54, 2012.
- [43] K. Walus, M. Mazur, G. Schulhof, G. A. Jullien, “Simple 4-bit processor based on Quantum-Dot Cellular Automata,” *Proceedings of the 16th IEEE International Conference on Application-Specific Systems, Architectures, and Processors* pp. 288–293 2005.
- [44] N. G. Anderson, İ. Ercan, N. Ganesh, “Toward nanoprocessor thermodynamics,” *Proceedings of 12th IEEE International Conference on Nanotechnology*, 2012. Extended version forthcoming in IEEE Transactions on Nanotechnology.
- [45] W. L. Liu, L. Lu, M. O’Neill, and E. E. Swartzlander Jr., “Design rules for Quantum-dot Cellular Automata,” *IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 2361–2364, 2011.
- [46] M. T. Niemier and P. M. Kogge, *Nano, Quantum and Molecular Computing: Implications to high level design and validation*, Sandeep K. Shukla, R. Iris Bahar Eds, Springer, pp. 267–293, 2004.
- [47] Retrieved on March 14, 2012, from Leibniz Translations: <http://www.leibniz-translations.com/binary.htm>
- [48] Retrieved on October 18, 2013, from Computer History Museum, <http://www.computerhistory.org/semiconductor/timeline/1833-first.html>
- [49] G. Boole, *The mathematical analysis of logic: Being an essay towards a calculus of deductive reasoning*, Cambridge: Macmillan, Barclay & Macmillan; 1847.
- [50] R. Clausius, *The mechanical theory of heat*, Macmillan and co., London, 1879.
- [51] H. Leff and A. F. Rex, *Maxwell’s Demon 2: Entropy, classical and quantum information, Computing*, Institute of Physics Publishing, Bristol and Philadelphia, 2003.
- [52] K. J. Laidler, *The Physical World of Chemistry*, Oxford University Press, 1995.
- [53] L. Boltzmann, “Further Studies on the Thermal Equilibrium of Gas Molecules,” *The Kinetic Theory of Gases*, vol.1, pp. 262-349, 2003.

- [54] J. W. Gibbs, "On the Equilibrium of Heterogeneous Substances," *Transactions of the Connecticut Academy of Arts and Sciences*, vol. 3, pp. 108-248 and 343-524, 1874 and 1878.
- [55] J. von Neumann, "Thermodynamik quantummechanischer Gesamtheiten," *Gött. Nach.* vol. 1, pp. 273-291, 1927.
- [56] R. V. L. Hartley, "Transmission of information," *Bell System Technical Journal*, vol. 7, pp 535-563, 1928.
- [57] L. Szilrd, "ber die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen," *Zeitschrift fr Physik*, vol. 53, pp. 840-856, 1929.
- [58] A. Turing, "On computable numbers, with an application to the Entscheidungs problem," *Proceedings of the London Mathematical Society*, ser. 2, vol. 42, 1937.
- [59] C. E. Shannon, "A symbolic analysis of relay and switching circuits," *Trans. AIEE*, vol. 57 no..12, pp. 713723, 1938.
- [60] D. M. MacKay, *Information, mechanism and meaning*, The M.I.T. Press, 1969.
- [61] L. Brillouin, *Science and information theory*, Academic Press, 1956.
- [62] J. S. Kilby, "Turning potentials into realities: The invention of the integrated circuit," *Int. J. Mod. Phys. B*, vol. 16, p. 699 2002.
- [63] C. H. Bennett, "Thermodynamics of computation - a review," *International Journal of Theoretical Physics*, vol. 21, no. 12, 1982.
- [64] Retrieved on August 14, 2011, from Oxford English Dictionary, <http://www.oed.com>
- [65] Retrieved on August 18, 2011, from Stanford Encyclopedia of Philosophy, <http://plato.stanford.edu/entries/information-entropy/>
- [66] O. R. Shenker, "Logic and entropy," [Preprint] 2000.
- [67] J. D. Norton, "Eaters of the lotus: Landauer's principle and the return of Maxwell's demon," *Studies in History and Philosophy of Science Part B* vol. 36 no. 2, pp.375-411, 2005.
- [68] Retrieved on September 19, 2011, from <http://dabacon.org/pontiff/?p=977>
- [69] O. J. E Maroney, "The (absence of a) relationship between thermodynamic and logical irreversibility," *Studies in the History and Philosophy of Modern Physics*, vol. 36, pp. 355-374, 2005.
- [70] O. J. E Maroney, "Generalising Landauer's principle," *Physical Review E*, vol. 79, pp. 031-105, 2009.

- [71] S. Turgut, “Relations between entropies produced in non-deterministic thermodynamic processes,” *Physical Review E*, vol. 79, pp. 041-102, 2009.
- [72] A. Berut, A. Arakelyan, A. Petrosyan, S. Ciliberto, R. Dillenschneider and E. Lutz, “Experimental verification of Landauers principle linking information and thermodynamics,” *Nature*, vol. 483, pp. 187-189, 2012.
- [73] B. Lambson, D. Carlton, and J. Bokor, “Exploring the thermodynamic limits of computation in integrated systems: Magnetic memory, nanomagnetic logic, and the Landauer limit,” *Phys. Rev. Lett.*, vol. 107, 2011.
- [74] A. O. Orlov, C. S. Lent, C. C. Thorpe, G. P. Boechler, and G. L. Snider, “Experimental test of Landauer’s principle at the Sub- $k_B T$ Level,” *Jpn. J. Appl. Phys.*, vol. 51 2012.
- [75] N. G. Anderson, “Conditioning, correlation and entropy generation in Maxwell’s Demon,” *Entropy*, vol. 15, pp. 4243-4246, 2013.
- [76] M. Hemmo and O. R. Shenker, *The road to Maxwell’s demon: Conceptual foundations of statistical mechanics*, Cambridge University Press, 2012.
- [77] Editorial, “Plenty of room revisited,” *Nature Nanotechnology*, vol. 4, p. 781 2009.
- [78] D. Patterson, “The trouble with multicore,” *IEEE Spectrum*, June, 2010.