Optoelectronics for refrigeration and analog circuits for combinatorial optimization

Patrick Xiao

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2019-74
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2019/EECS-2019-74.html

May 17, 2019

Optoelectronics for refrigeration and analog circuits for combinatorial optimization

by

Tianyao Patrick Xiao

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering – Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Eli Yablonovitch, Chair
Professor Ming C. Wu
Professor Christopher Dames

Spring 2019

Optoelectronics for refrigeration and analog circuits for combinatorial optimization

Abstract

Optoelectronics for refrigeration and analog circuits for combinatorial optimization

by

Tianyao Patrick Xiao

Doctor of Philosophy in Engineering – Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Eli Yablonovitch, Chair

The increasing efficiency and adoption rate of light-emitting diodes (LEDs) for lighting is forecasted to lead to large energy savings in that sector. Intrinsic self-refrigeration, which surprisingly accompanies the emission of light from an LED, might allow the same technology to be harnessed for new applications. In the first half of this dissertation, I will discuss whether refrigeration and cooling, which also account for a large share of world energy consumption, comprise a realistic future energy application for optoelectronics. To date, electroluminescent cooling has eluded direct observation at practical power densities, since a pre-requisite for net cooling is the near-complete elimination of LED internal losses that dissipate heat. I will propose a near-ideal-efficiency LED structure, realizable with existing optoelectronic material quality and device processing capabilities, to predict the technological limits of electroluminescent refrigeration. I will show that LED-based cooling can realistically be more efficient than solid-state alternatives like thermoelectric cooling, particularly for cryogenic applications and at moderate power densities.

In the second half of this dissertation, I will discuss a new hardware accelerator for difficult combinatorial optimization problems. These problems are abundant in modern society, in the fields of operations research, artificial intelligence, chip design, financial optimization, medicine, and many others. They are difficult in that no algorithm has yet been found that solves them efficiently, but the increasing computational power of digital computers has allowed these problems to be routinely tackled. As conventional computers reach their scaling limits, however, alternative hardware architectures are being explored to accelerate computationally intensive tasks like machine learning and combinatorial optimization. Here, I will discuss the design of a new optimization machine, implemented using a network of coupled analog electrical oscillators. The machine uses an unconventional search mechanism to discover the global minimum of the Ising problem, a difficult problem that maps quickly onto other hard optimization problems. I will present the simulated performance of this machine for moderately sized problems with all-to-all connectivity and discuss its potential to scale to larger problems.

To my loving parents and sister.

Contents

List of Figures

Acknowledgments

I am deeply grateful to have worked with so many talented and supportive people over the course of my graduate studies.

I must begin by thanking my thesis advisor, Professor Eli Yablonovitch, whose role in my development as a scientist is hard to overstate. Eli always seems to know where the important problems are, and as my mentor and my advocate, he has given me the confidence to seek them out and tackle them, even if it means venturing into an unfamiliar field (as I did at different points in my PhD). His guidance has taught me the importance of consolidating my work into ideas that are easy to communicate and easy to build from. I am incredibly fortunate to have worked in his group.

I would like to thank the other professors that I have had the opportunity to collaborate with over the past several years. Professor Paul Braun, at the University of Illinois at Urbana-Champaign, was instrumental in turning my designs for solar spectral splitting from an abstract promise into physical reality.[1] I appreciate my discussions with Professor Shanhui Fan, at Stanford University, on the thermodynamics of electroluminescent cooling. I would also like to thank the members of my dissertation and qualifying exam committees – Professor Ming Wu, Professor Chris Dames, and Professor Connie Chang-Hasnain – for their questions and feedback that have led me to evaluate my work more critically. I am also grateful to Professor Jaijeet Roychowdhury for his suggestions about my thesis.

My colleagues and friends in the Yablonovitch group, past and present, have contributed so much to my experience at Berkeley on both a professional and a personal level. I owe my gratitude to Samarth Bhargava, whose mentorship during my time as an undergraduate researcher helped me to transition quickly into a productive member of the group. I took after his research and presentation style in my early days, and still do in many ways. I have had many great discussions with Andy Michaels who, despite the relatively small overlap in our respective projects, has always been a reliable evaluator of my ideas and a source of many new insights. My collaboration with Sri Krishna Vadlamani has helped to advance our understanding of the Ising machine farther than either of us would have done alone. Gregg Scranton was a close collaborator in photovoltaics and inverse design, a great friend, a companion to heavy metal concerts, and a partner through some of the rougher times in graduate school. I am indebted to Vidya Ganapati, Owen Miller, and Sapan Agarwal for their foundational work that have helped get me started on my research projects. I would also like to acknowledge the support of Zunaid Omair, Sean Hooten, Luis Pazos, Chris Keraly, and Kevin Messer that came in the form of their company, their appetite for physics discussions, and their help in securing computational resources. I am also very grateful for two undergraduate interns – Ryan Brandt and John Trinh – who made valuable contributions to my electroluminescent cooling and spectral splitting projects, respectively.

I would like to thank everyone in the optoelectronics office, fellow members of Photobears, Shirley Salanio, Therese George, Josephine Ho, and the students and staff in the Center for

[1] This work did not ultimately find a suitable place in this thesis. For the details, see Ref. 1.

Energy Efficient Electronics Science (E^3S), who have helped make grad school more relaxing, enjoyable, and welcoming. I am grateful to Parthi Santhanam at Stanford for many insightful theoretical discussions about thermophotonics and about the experimental efforts, past and present, to make LED cooling real. I am also fortunate to have collaborated with Kaifeng Chen at Stanford on radiative cooling and Osman Safa Cifci at UIUC on the experimental characterization of spectral splitting optics. I would also like to acknowledge a number of others with whom I have enjoyed useful discussions on a number of topics, particularly: John Lloyd at Caltech, Winston Chern, Xin Zhao, and Redwan Sajjad at MIT, Tianshi Wang at Berkeley, and Professor Daniel Lidar at USC,

My research and my PhD education could not have been possible without the generous financial support from the National Science Foundation Graduate Research Fellowship program, the Office of Naval Research, the Department of Energy "Light-Material Interactions in Energy Conversion" Energy Frontier Research Center, and the Dow Chemical Company.

The great friends that I have made in my nine years at Berkeley, as well as my long-time friends from high school, have always been very supportive of my endeavors. Thanks for always being there to help balance the scales when the PhD got tough.

Finally, I am deeply grateful to my parents and sister for the love, support, and advice that they have shared over all these years. They have been my greatest influences, invested in my success at every step, and I am happy to make them proud.

Chapter 1

Introduction

1.1 Solid-state cooling: a future frontier for optoelectronics?

The increasing efficiency and adoption rate of the light-emitting diode (LED) as a light source is forecasted to lead to large energy savings in that sector. In 2015, lighting accounted for 15% of the total electricity consumption in the United States. The U.S. Department of Energy projects that by 2035, if the current level of investment in solid-state lighting continues, LED technology would be responsible for a 55% reduction in the total energy consumed by light sources (75% under more aggressive targets) [2]. Meanwhile, the solar cell, which is the reciprocal device to the LED, provides a global installed electricity generation capacity of over 400 GW as of 2017, a 50-fold increase from 2007 [3]. Rapid growth in these major energy applications has been fueled by the increasing efficiency of optoelectronic devices. LED efficiency has increased by several orders of magnitude since its invention, and blue LEDs today can operate with greater than 50% wall-plug efficiency [4]. By adapting some of the principles that are used in LED design, solar cells are also on a path to their theoretical limits of efficiency [5].

Surprisingly, the process of light emission in an LED is accompanied by an intrinsic self-refrigeration effect. This feature has long been recognized as a latent capability of LEDs [6], but one that has never been exploited or even observed at practical power densities. At all but the smallest power densities, cooling is a relatively weak effect that is rather easily beset by small internal losses. However, as LED technology approaches its fundamental efficiency limits, the prospect of solid-state electroluminescent cooling becomes more realistic. Cooling is an increasingly pervasive need in modern developed society: in 2015, space cooling accounted for an estimated 17% of residential electricity consumption in the United States, and refrigeration accounted for another 7% [7]. Much of this demand is presently met by technologies that make use of fluorinated gases, a major contributor to atmospheric pollution and climate change. If LED performance can be made viable enough to address these needs, clean cooling would represent a major future energy application for optoelectronics.

This is, of course, an extremely ambitious goal. In the shorter term, we can investigate whether the electroluminescent refrigeration mechanism is competitive with the adjacent solid-state methods of cooling. Solid-state refrigerators, though usually less efficient than their vapor-compression counterparts, possess the advantage of having no moving parts and can thus be highly portable. The most common of these technologies is thermoelectric cooling, which is in widespread use for the cooling of electronics and laser diodes. Solid-state laser cooling, which is more closely related to LED cooling, is currently an active area of research [8]. In the early part of this thesis, we will address whether LED cooling has the potential to become a viable alternative to these technologies. We will propose a highly optimized LED structure, realizable with present levels of optoelectronic material quality and device processing capability, and use this is a platform to assess the practical limits of this cooling mechanism.

1.2 Optimization machines: the renaissance of analog computing?

Analog computers, which process continuous-valued rather than discrete-valued information, long predate digital computers. Its electronic incarnation, introduced in the early part of the 20^{th} century, was used to perform mathematical functions and solve differential equations. They have also been shown to be efficient in solving continuous optimization problems such as linear and quadratic programming [9]. For general-purpose computing, however, the analog computer was superseded by the digital computer, which could offer higher precision and were more robust to practical issues like variability and drift in the electronic components.

There is a wide class of problems, however, that are discrete in nature but very difficult to solve efficiently using digital computers. In these combinatorial optimization problems, the objective is to find the best solution out of a finite set that minimizes or maximizes a prescribed figure-of-merit. Problems of this type arise routinely in modern society, where we must choose how best to allocate precious resources – such as time, space, and energy – to achieve a specific goal. For many of these problems, it is almost universally believed that no efficient algorithm can exist that can be implemented on a digital computer: as the size of the problem increases linearly, the resources that must be consumed to solve the problem grows exponentially. While the rapid pace of progress in digital hardware and algorithms has allowed large-scale optimization problems of this type to be tackled, the end of Moore's law might soon make scaling to larger, more difficult problems prohibitively resource-intensive.

There have therefore been many attempts to solve these problems using alternative hardware platforms that exploit the computational ability of physical systems, which are not constrained to process information in a sequential manner. Efforts in quantum computing have been driven in large part by its promise to accelerate solutions to difficult mathematical problems. In the related field of artificial intelligence – which is similarly computationally intensive and heavily reliant on optimization – the recent proliferation of applications has

spurred the exploration of new, specialized hardware architectures, both digital and analog [10].

It is in the domain of difficult optimization problems that analog computers, presently on a long hiatus from mainstream use, may find new life again. Generally speaking, these types of analog computers are classical machines that accept as input the parameters of the optimization problem, which are translated into a physical configuration of the hardware. The machine then finds a solution by relaxing via its internal dynamics to a physically favorable state, which can be measured to yield the answer to the problem. If we restrict both the input and the output of the machine to be digital, but exploit the internal analog processing of information, we can potentially accelerate the solutions to difficult optimization problems while retaining some robustness to component precision and drift. A digital output is already intrinsic to combinatorial problems, while a digital input will be needed as an additional constraint. A number of such analog machines have already been proposed and demonstrated, though it is not yet clear what approach is the most efficient or the most practical. Nonetheless, it is evident that analog computers are experiencing a resurgence, with potentially significant implications for the future of computation.

1.3 Thesis outline

This thesis is divided into two parts, addressing the two largely distinct areas of research introduced above.

The first part of this thesis (Chapters 2–4) explores the design and application of ultrahigh-efficiency optoelectronics. In Chapter 2, we introduce the external luminescence quantum efficiency as a figure-of-merit that is shared by all optoelectronics. Whether they convert electrons to photons or photons to electrons, the theoretical limit of device performance goes hand in hand with unity luminescence efficiency. In Chapter 3, we investigate how closely a practical device can approach this ideal. Our strategy is, informally speaking, a proof by construction: we strive to design a light-emitting diode with the highest luminescence efficiency that is practically achievable. To account for and minimize all of the imperfections that arise in real devices, we must carefully consider the operation of the device in the electronic, optical, and electrical domains. Therefore, this chapter is dedicated in equal measure to self-consistent device modeling and to a co-design approach that simultaneously minimizes losses of every type. Chapter 4 explores an emergent effect that is found only in devices with near-ideal luminescence efficiency: self-refrigeration. To exploit this mechanism as a technology for cooling, we draw upon both efficient light-emitting diodes and efficient photovoltaics. We use the device design in Chapter 3 as a platform to investigate the practical limits of electroluminescent cooling, and evaluate its technological viability in comparison to other solid-state methods of cooling. The analysis in Chapters 3 and 4 has been covered in part in a published manuscript, Ref. 11.

In the second part of this thesis (Chapters 5 and 6), we will consider analog computation as a potentially efficient means to solve problems that are intractably difficult for digital

computers. Chapter 5 serves as a high-level introduction to the subject, where we discuss the inherent difficulty of combinatorial optimization problems and survey the general approaches, both algorithmic and physically-based, that are used to solve them. We will visit the important issue of precision and the limitations that this poses on any analog hardware for solving digital optimization problems. In Chapter 6, we design a new type of analog optimization machine that is built from a network of electrical oscillators that are dissipatively coupled. The machine solves a specific problem known as the Ising problem, but can readily be adapted for other difficult optimization problems. We demonstrate our system's performance as an optimizer using circuit simulations and present a theoretical framework to describe its dynamical operation.

Part I

Optoelectronics at the edge of perfection

Chapter 2

Luminescence efficiency: a universal figure-of-merit for optoelectronics

The reciprocity between light emission and light absorption under thermal equilibrium was recognized more than 150 years ago by Gustav Kirchoff [12], long predating the birth of optoelectronics. If Kirchoff's passive light emitter were replaced by an excited electronic system, such as a light-emitting diode (LED) or a solar cell, the system no longer obeys the laws of equilibrium thermodynamics. Nonetheless, a more practical statement of reciprocity continues to hold that can be expressed as follows: all optoelectronics – absorbers and emitters alike – can be designed by maximizing the same figure-of-merit. For instance, an efficient solar cell should be designed in just the same way as an efficient LED, and vice versa [5], [13]. That universal figure-of-merit is the external luminescence efficiency, η_{ext}.

Whenever a steady-state population of excited charge carriers exists in a semiconductor, the carriers are in quasi-equilibrium with a population of luminescent photons. The photons are generated by radiative electron-hole recombination, and have the potential to escape the semiconductor into the surrounding environment. But neither radiative recombination nor the extraction of the emitted photons is guaranteed to occur, due to the presence of competing processes: the *non*-radiative removal of carriers, and the *non*-emissive removal of photons. We define the external luminescence efficiency as the probability of conversion from an excited electron-hole pair into a photon that escapes the device:

$$\eta_{\text{ext}} \equiv \frac{\#\text{ external photons} \cdot \text{s}^{-1}}{\#\text{ excited electron-hole pairs} \cdot \text{s}^{-1}} \tag{2.1}$$

We have expressed the above efficiency as a ratio of rates (or fluxes), with the denominator being the rate at which the device is pumped with excited carriers. The pump can be electrical, as in an LED, or it can be optical, as in a solar cell. We refer to the resulting radiation as electroluminescence and photoluminescence, respectively. The same definition for η_{ext} can be applied to both cases, though we must note that this quantity goes by different names for different technologies. In LEDs, it is usually called the external quantum efficiency (EQE) [14], while in photovoltaics it is often called the external radiative efficiency (ERE)

or photoluminescence quantum efficiency (PLQE) [15]. For clarity, we will always refer to it as the external luminescence efficiency.

In the following two sections, we will show how the performance of both the LED and the solar cell are connected directly to the external luminescence efficiency. We will then discuss how an extremely low-loss reflector – a necessity for high η_{ext} – greatly improves the efficiency of thermophotovoltaic power generation. This last result should be considered a practically significant byproduct of designing for high η_{ext}, rather than a direct consequence of high luminescence efficiency.

2.1 The luminescence efficiency of LEDs

Since its introduction in the middle part of the last century, the light-emitting diode (LED) has seen dramatic increases in energy efficiency. Commercial LEDs now convert electrical power to visible light with about 50% efficiency, an improvement by several orders of magnitude over the first red LEDs – see Fig. 2.1. From 1960 to 2000, the luminous efficiency of LEDs has doubled every four years [16]! This is a result of decades of progress in both optoelectronic material quality, which allows for the efficient generation of luminescence, and in light extraction strategies, which increase the probability that the internal luminescence escapes out of the device [14].

The external luminescence efficiency is a self-evident figure-of-merit for LED design, whose objective is to maximally convert the injected electrons and holes into outgoing photons. The wall-plug efficiency of the LED, defined as the ratio of optical output power to electrical input power, is related directly to the external luminescence efficiency η_{ext} by:

$$\eta_{\text{WPE}} = \frac{\langle E \rangle}{qV} \times \eta_{\text{ext}} \tag{2.2}$$

where $\langle E \rangle$ is the average energy of an emitted photon, V is the applied voltage bias across the device terminals, and q is the electronic charge. This expression can be intuitively

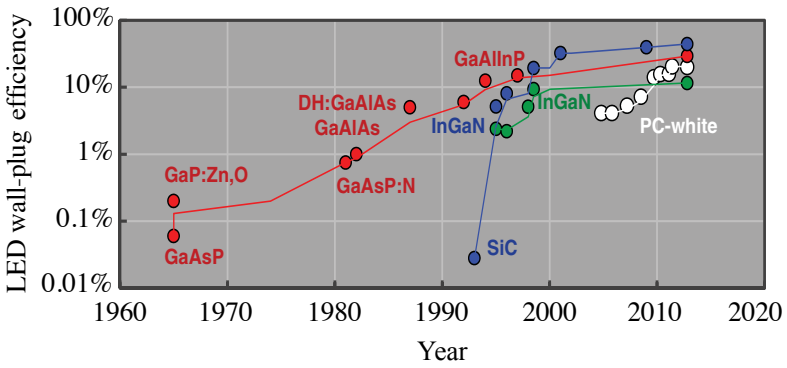

Figure 2.1: The trend of increasing efficiency over time in visible LEDs (taken from Ref. 4).

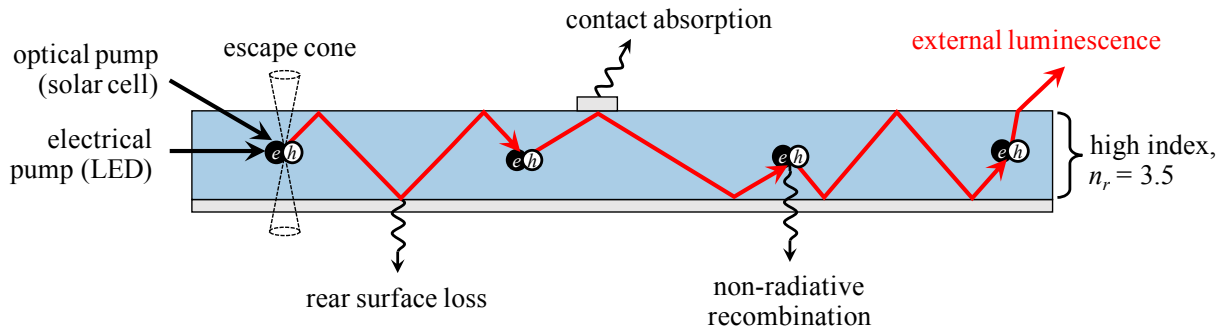

Figure 2.2: Light extraction from a high-index material is inherently difficult due to the small probability of emission into the escape cone. Extraction of the internal luminescence must compete with various loss mechanisms, both optical and electronic. High external luminescence efficiency η_{ext} demands very high internal luminescence efficiency η_{int} and very low optical losses.

understood on a per-carrier basis: the first term is the ratio of the outgoing photon energy to the work consumed to supply the device with a single electron, and the second term is the quantum efficiency of conversion from charge carriers into photons.

It is common to decompose the external luminescence efficiency as [14]:

$$\eta_{\text{ext}} = \eta_{\text{int}} \times C_{\text{ext}} \tag{2.3}$$

where η_{int} is the internal luminescence efficiency, the probability that an injected electron-hole pair recombines radiatively in the active region, and C_{ext} is the light extraction efficiency, the probability that an internally generated photon is ultimately converted into an externally emitted photon. Generally speaking, the internal luminescence efficiency depends primarily on material quality and is optimized separately from light extraction, which depends additionally on the optical design of the LED chip. We must recognize, however, that the two efficiencies above are not independent. As we will see, strong emitters (high η_{int}) benefit to a larger degree from optical design than do poor emitters (low η_{int}).

Let us use a ray optics picture to consider why light extraction is a problem of great inherent difficulty in LEDs. Fig. 2.2 depicts a device that emits light out of its top surface. Of the photons that are generated within the high-index semiconductor, only a small fraction lie in the cone of escape into the surrounding air or vacuum. This fraction is $1/4n_r^2$, which for GaAs is just 2%! The remainder is trapped inside by total internal reflection, and are ultimately absorbed by the device. The absorption can re-generate an electron-hole pair in the active region, or the absorption can be parasitic, generating heat. In the former case, it is possible through radiative recombination to recover the internal photon, which now has a renewed opportunity to enter the escape cone. Nonetheless, because of the narrow cone of escape, many such emission events are needed before the photon finally leaves. Thus, the external luminescence efficiency η_{ext} is heavily penalized by even a small probability of non-radiative recombination upon re-absorption, or by a slight amount of parasitic optical loss in the volume or at the surfaces of the device.

For each internally generated photon, let us assign a probability to each fate mentioned above: the probability of escape without being absorbed P_{esc}, the probability of band-to-band absorption in the active region P_{abs}, and the probability of parasitic absorption P_{par} by the rear surface, the electrical contacts, free carriers, or another mechanism. These must add to unity, $P_{\text{esc}} + P_{\text{abs}} + P_{\text{par}} = 1$. A good optical design for an LED is one that maximizes P_{esc} and minimizes P_{par}. Using these quantities, along with the probability of radiative recombination η_{int}, we can derive an expression for the external luminescence efficiency by following the fate of a single injected electron:

$$\eta_{\text{ext}} = \eta_{\text{int}} P_{\text{esc}} + (\eta_{\text{int}} P_{\text{abs}})\, \eta_{\text{int}} P_{\text{esc}} + (\eta_{\text{int}} P_{\text{abs}})^2\, \eta_{\text{int}} P_{\text{esc}} + \ldots \tag{2.4}$$

where the first term is the probability that external photon emission occurs on the first escape attempt, the second term is the probability that external emission occurs on the second escape attempt (following successful absorption and re-emission), and so on.

Summing over all of the escape attempts, we obtain:

$$\eta_{\text{ext}} = \frac{\eta_{\text{int}} P_{\text{esc}}}{1 - \eta_{\text{int}} P_{\text{abs}}} = \eta_{\text{int}} \times \frac{P_{\text{esc}}}{1 - \eta_{\text{int}} (1 - P_{\text{esc}} - P_{\text{par}})} \tag{2.5}$$

This expression is more commonly encountered in the photovoltaics literature [17], [18] but applies equally well to LEDs. Comparing this to Equation (2.3), we find that the second term above corresponds to the light extraction efficiency. The distinction between the escape probability P_{esc} and the light extraction efficiency C_{ext} is significant. The former is the likelihood that an internal photon escapes *without* first being absorbed. The latter is the overall escape probability summed over all subsequent re-absorption/radiative recombination events. In most LEDs, P_{esc} is often a small probability due to the high active-region refractive index; however, if the device has high internal luminescence efficiency and low optical parasitics, the light extraction efficiency can approach 100%. The fact that C_{ext} can greatly exceed P_{esc} is often called photon recycling. In poor emitters ($\eta_{\text{int}} \approx 0$), photon recycling is largely absent, and the two quantities are nearly equal: $C_{\text{ext}} \approx P_{\text{esc}}$. For such materials, there is little benefit in designing for low optical losses.

In Chapter 3, we will consider various design strategies to maximize the external luminescence efficiency of an LED by optimizing for high η_{int}, low P_{par}, and high P_{esc}. We also note that several alternative expressions exist in the optoelectronics literature for η_{ext}. The equivalence of these expressions to Equation (2.3) is shown in Appendix A.

2.2 The luminescence efficiency of solar cells

Solar cells, for much of their history, have been designed primarily to be excellent light absorbers. While this is clearly a necessary function, we now know that by following this strategy alone, a solar cell can never reach the theoretical limit of efficiency found by Shockley and Queisser [23]. The record power conversion efficiency for single-junction solar cells under

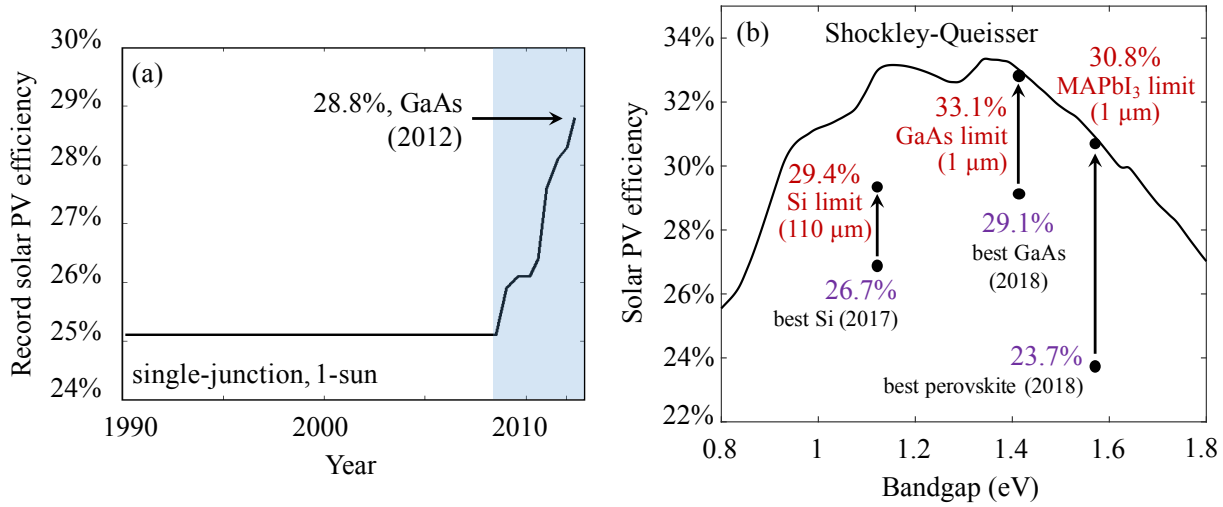

Figure 2.3: (a) The record efficiency of single-junction solar cells under 1-sun illumination, 1990-2012. In 2018, the record improved from 28.8% to 29.1%. (b) The present record solar cells [19] vs. their intrinsic efficiency limits when all but fundamental Auger recombination losses are eliminated from the device. The Si limit [20] assumes a thickness of 110 μm. The GaAs limit [5] and the $MAPbI_3$ limit [21] assume a thickness of 1 μm. The black curve is the Shockley-Queisser limit at each bandgap, assuming an ideal step-function absorber, calculated using Ref. 22.

one-sun illumination is shown in Fig. 2.3(a); for a long period (1990-2007), the record had been maintained at 25.1% [24], well short of the Shockley-Queisser limit of ∼33.5%. More recently, it has become widely recognized that to approach this limit, a solar cell must also be an efficient LED [5]. In other words, good light *extraction* must become a tenet of solar cell design, along with good absorption. While counter-intuitive, this principle has spurred the recent increases in solar cell efficiency to the present single-junction record of 29.1%, using GaAs devices [19], and has also led to record-efficiency multi-junction cells [25].

While the need for light extraction is self-evident in an LED, it is less obvious in a solar cell, where external emission appears to be a mechanism for energy loss that should be suppressed. However, the emission of light back toward the Sun is a fundamentally necessary pathway of carrier decay, while thermodynamics permits the elimination of all other channels of carrier removal. The necessity of external photon emission is most clearly seen in thermal equilibrium, where Kirchoff's law requires any medium to emit precisely as much light as it absorbs from its surroundings:

$$\Phi_{\text{ext},0} = \int_0^\infty a(E)\, b(E,T)\, dE \tag{2.6}$$

where $\Phi_{\text{ext},0}$ is the external luminescent emission rate in equilibrium, E is the photon energy, $a(E)$ is the band-to-band electronic absorptivity and $b(E,T)$ is the spectral radiance of blackbody radiation at the cell's temperature, T. When driven out of equilibrium by solar

illumination, the carrier density increases and the quasi-Fermi levels for electrons and holes separate. Since radiative recombination is a bimolecular process, the emission rate Φ_{ext} must increase in proportion to the product of carrier densities:

$$\Phi_{\text{ext}} = \Phi_{\text{ext},0} \cdot \frac{np}{n_0 p_0} = \Phi_{\text{ext},0} \cdot \exp\left(\frac{qV}{kT}\right) \tag{2.7}$$

where n and p are the electron and hole densities (respectively) under illumination, n_0 and p_0 are the corresponding densities in equilibrium, and k is the Boltzmann constant. The second equality follows from the mass action law in semiconductors, which relates the carrier density to the splitting of the quasi-Fermi levels qV under the Boltzmann approximation. A more rigorous detailed-balance argument, accounting for the Fermi-Dirac distribution of the carriers, yields nearly the same expression for Φ_{ext} – see Equation (3.17).

A rate of external light emission smaller than that given by Equation (2.7) is thermodynamically forbidden. Therefore, the maximum steady-state carrier concentration – and hence a maximum voltage – is attained by suppressing every pathway for carrier removal other than the thermodynamically required emission process. Suppose that solar illumination generates carriers at an areal rate given by Φ_{pump}, and that at open-circuit a fraction η_{ext} of these excited carriers are removed via external light emission. By balancing the rates of carrier generation and removal, we can obtain the open-circuit voltage [5], [13], [26]:

$$qV_{\text{oc}} \;=\; kT \ln\left(\frac{\Phi_{\text{pump}}}{\Phi_{\text{ext},0}}\right) - kT|\ln \eta_{\text{ext}}| \;=\; qV_{\text{oc,ideal}} - kT|\ln \eta_{\text{ext}}| \tag{2.8}$$

Inefficiencies in the creation or extraction of luminescent photons directly penalizes the voltage, as these would introduce additional pathways for loss and lead to a more rapid depletion of carriers. Improvements in η_{ext} allow the device to approach its ideal, maximum voltage, and $\eta_{\text{ext}} = 1$ is indeed a necessary condition to arrive at the Shockley-Queisser limit. In approaching that limit, a large burden falls upon the optical design of the solar cell, just as in LEDs [5]. The most likely optical loss in an unoptimized LED or solar cell is absorption by a thick growth substrate of comparatively low optoelectronic quality: if access to this substrate is permitted, a large fraction of the generated photons are lost through the rear surface, never to see the Sun. It is therefore extremely important to remove the photovoltaic film from its substrate by epitaxial lift-off [27] and to subsequently couple the film to a very low-loss reflector. Improvements in rear reflectivity translate to improvements in η_{ext} and therefore the voltage [28]: this has been precisely the technological driver for the increase in solar cell efficiency depicted in Fig. 2.3(a). A more quantitative treatment of how the rear reflectivity improves the external luminescence efficiency will be given in Chapter 3.

Designing for ideal light extraction allows solar cells to approach the Shockley-Queisser limit, but in practice they are limited by intrinsic material losses. Fig. 2.3(b) shows the gap between the present efficiency record and the intrinsic efficiency limit of the material – where all but Auger recombination losses have been removed – for Si, GaAs, and perovskite solar cells. For materials with a promising internal luminescence efficiency, such as GaAs [5], ideal

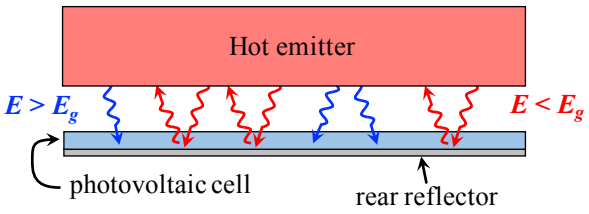

Figure 2.4: In thermophotovoltaics, a photovoltaic cell harvests energy from a locally heated emitter of thermal radiation. For high efficiency, the cell itself can be used as a spectrally selective filter that absorbs the above-bandgap thermal photons (blue) and uses a rear reflector to return the unabsorbed below-bandgap photons (red) back to the emitter.

light extraction brings the efficiency very close to the Shockley-Queisser limit. In Ref. 21, we find that this is also true for lead iodide perovskites ($MAPbI_3$). However, for poor light emitters like Si, where Auger recombination is inherently strong, the limiting efficiency is comparatively far from the Shockley-Queisser limit even with ideal light extraction.

2.3 Thermophotovoltaics

In addition to improved light extraction, which benefits both LEDs and solar cells, the development of a near-ideal reflector holds other important technological implications. Fig. 2.4 depicts an optoelectronic scheme for heat-to-electricity conversion that uses a photovoltaic cell to harvest thermal radiation from a hot object in its vicinity. This method of electricity generation is called thermophotovoltaics (TPV), to be contrasted with solar photovoltaics, where the radiation arrives from a distant source. As in solar PV, only those photons with energy above the semiconductor bandgap are absorbed and converted. But unlike solar PV, the unabsorbed sub-bandgap photons need not be wasted. If the photovoltaic cell were backed by a highly reflective mirror, those photons can be returned back to the hot emitter, where they are re-absorbed. By re-heating the emitter, the reflected below-bandgap photons reduce the amount of heat that must be supplied externally to maintain its high temperature. In this section, we will briefly review the theoretical and practical limits of this scheme for electricity generation. For more details on the following analysis, we refer the reader to our published preprint in Ref. 29: here, we will summarize the main results.

Efficient TPV conversion relies upon spectral selectivity: maximize the conversion of photons above the bandgap, and minimize the loss of photons below the bandgap. The reflective approach described above fulfills the requirement. Selectivity can alternatively be engineered by modifying the emissivity of the light source using a spectral filter (such as a photonic crystal) that inhibits sub-bandgap emission [30], or by exploiting a material whose emission band lies above the photovoltaic bandgap [31]. But the simpler approach in Fig. 2.4 has some advantages. It uses the inherently sharp band-edge absorption profile of the photovoltaic cell itself to provide spectral selectivity, allowing the use of an ordinary,

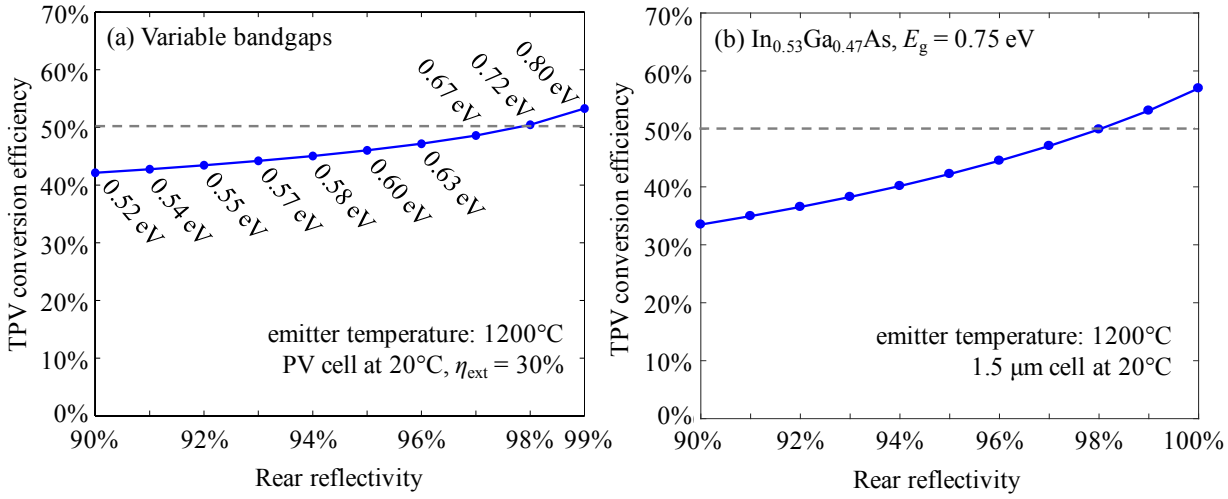

Figure 2.5: The limiting efficiency of TPV conversion as a function of a uniform rear reflectivity with an emitter at 1200°C. (a) An optimal bandgap is chosen at each reflectivity value to maximize the TPV efficiency, assuming a fixed value of $\eta_{ext} = 30\%$ for all bandgaps. (b) Rather than a variable bandgap and a fixed η_{ext}, the material parameters for $In_{0.53}Ga_{0.47}As$ are used to calculate the TPV efficiency. The efficiency of a Carnot engine at these temperatures is 80%.

unstructured thermal emitter that can readily withstand high temperatures.

The essential component in this scheme is a mirror with high reflectivity over a broad band, which returns nearly every sub-bandgap photon back to its source. Prior work making use of this scheme has reached 29% TPV conversion efficiency using Si cells at emitter temperatures above 2000°C [32], and 23.6% with InGaAs cells at a more moderate temperature of 1039°C [33]. With the ability to process high-quality thin films of III-V materials via epitaxial lift-off, which is now in widespread manufacturing use for solar cells and LEDs, much higher cell reflectivity is accessible than ever before. High reflectivity, which has been pivotal for the record-efficiency solar cells and for advancements in LED efficiency, now serves as as an impetus for new efficiency records in thermophotovoltaics.

Fig. 2.5 predicts the limiting efficiency of TPV conversion (thermal radiation to electricity) as a function of the average sub-bandgap reflectivity when illuminated by a 1200°C blackbody, whose intensity is equivalent to 267-sun concentration. We assume an idealized system in which the emitter and the cell have very large area, and in which the cell captures all of the emitted radiation. The emitter and cell are separated by a vacuum gap spanning many thermal wavelengths, permitting heat exchange only via far-field thermal radiation.

In Fig. 2.5(a), we evaluate a material-agnostic limit by leaving the bandgap as a free parameter and assuming a step-function absorptivity. We further assume that a luminescence efficiency of $\eta_{ext} = 30\%$ is attainable at all bandgaps: in reality, some materials are more ideal while others are less. As the reflectivity increases, the below-bandgap mirror losses are reduced in magnitude relative to the above-bandgap thermalization losses, favoring a

larger bandgap to optimize the TPV efficiency. At a reflectivity of 98%, which has likely been attained in the record GaAs solar cells, the optimal bandgap lies around 0.72 eV for a 1200°C emitter. This is very close to the bandgap of $In_{0.53}Ga_{0.47}As$ (0.75 eV), which can be grown lattice-matched to InP and, like GaAs, is known to be an efficient light emitter.

In Fig. 2.5(b), we plot the predicted TPV efficiency using an $In_{0.53}Ga_{0.47}As$ cell. We assume a uniform rear mirror reflectivity across all photon energies, and use the reported absorption coefficient of $In_{0.53}Ga_{0.47}As$ and the non-radiative recombination coefficients measured in Ref. 34 to evaluate the luminescence efficiency. For a rear reflectivity of 98% and an optimal cell thickness of 1.5 μm, the TPV conversion efficiency reaches $\eta_{\mathrm{TPV}} = 50\%$, which becomes competitive with that of various mechanical heat engines. The external luminescence efficiency at this operating point is $\eta_{\mathrm{ext}} = 32\%$. Relative to a less reflective cell, some gain in TPV performance results from the more efficient reflection of luminescent photons, which enhances light extraction and raises the cell's voltage. This benefit is secondary to the main effect, which is the increased recycling of unabsorbed photons back to the source, enabled by high reflectivity in the sub-bandgap region.

Experiments in a vacuum chamber using a graphite emitter at 1206°C and a 10 mm^2 $In_{0.53}Ga_{0.47}As$ cell have produced a TPV efficiency of 29.1% [35]. The cell's active layer was a thin film of $In_{0.53}Ga_{0.47}As$ (2.5 μm), which resides atop a gold reflector to yield a measured sub-bandgap reflectivity of 94.6%. Significant improvements in TPV efficiency are anticipated with a superior cell reflectivity.

Chapter 3

Design of ultra-efficient GaAs light-emitting diodes

Underlying decades of progress in LEDs and the increasing commercial success of solid-state lighting is the science of luminescence generation and extraction from a semiconductor. In recent years, the wall-plug efficiency of the best LEDs has advanced to within a factor of two from 100% (see Fig. 2.1): it is natural, then, to ask where the practical and theoretical efficiency limits of this technology lie. In this chapter, we will set out to find this practical limit by a design approach: we propose a device structure that, accounting for all realistic losses, comes as close as practically possible to the limit of unity external luminescence efficiency. At the end of this chapter, we will compare this practical limit to the theoretical limit of wall-plug efficiency, which surprisingly can exceed 100%. As we will explore in greater detail in the next chapter, above-unity wall-plug efficiency leads to the possibility of self-refrigeration, a property that emerges only in near-ideal LEDs.

The practical limit of external luminescence efficiency is an application-dependent limit, as the application casts an overarching influence on device design and dictates, to a large degree, the optimal choice of technology. Our design efforts in this chapter are motivated not by the goal of efficient solid-state lighting but by the more novel and as yet unrealized application of LED-based electroluminescent cooling. In Chapter 4, we will find that this type of cooling attains its greatest efficiency advantage over existing methods at more modest power densities. Thus, we will tailor our device design to maximize the luminescence efficiency at correspondingly moderate applied biases: qV between 80% to 90% of the bandgap energy. Many of the methods and principles described below can be applied to design high-efficiency LEDs for other applications at similar power densities.

Heat transfer need not require the emission of visible light; in fact, as we shall see in Chapter 4, electroluminescent cooling generally prefers a material whose bandgap energy is smaller than that of visible photons. In this chapter, we will therefore consider GaAs – rather than III-nitride systems – as the active light-emitting material. Aside from a more optimal bandgap, we choose GaAs primarily because it is well-known as an efficient, direct-bandgap light-emitter as well as a mature optoelectronic technology. Extremely high external lumi-

nescence efficiencies have been observed in high-quality GaAs samples in photoluminescence experiments [36], [37]. GaAs devices are also the most efficient single-junction solar cells [19], as well as efficient near-infrared light emitters: InAlGaAs LEDs at 850 nm (1.46 eV) have achieved an external luminescence efficiency of $\eta_{\text{ext}} = 68\%$ and a wall-plug efficiency of $\eta_{\text{WPE}} = 72\%$ [38].

Nonetheless, to our knowledge, the most efficient LED reported in the literature is a GaN-based device, grown on a bulk GaN substrate, emitting at 415 nm (2.99 eV) [39]. This device achieved a peak external luminescence efficiency of $\eta_{\text{ext}} = 80\%$ and a wall-plug efficiency of $\eta_{\text{WPE}} = 84\%$ at both 25°C and 85°C. In this chapter, we propose a design for a GaAs LED whose predicted external luminescence efficiency well exceeds 80%, allowing for above-unity wall-plug efficiency and hence net cooling of the device.

In anticipation of the use of the LED for room-temperature refrigeration applications, most of the results shown in this chapter will assume an operating temperature of 263K, well below ambient. The performance at 300K, which is generally slightly worse, will also be given for the major results of this chapter – the external luminescence efficiency and the wall-plug efficiency. The analysis and the results in this chapter have been included partially in Ref. 11.

3.1 Ultra-efficient GaAs LED structure

Fig. 3.1 shows an LED device structure that has been optimized for both high external luminescence efficiency and high wall-plug efficiency. This device represents – in our view – close to the most efficient LED that can be realized using existing optoelectronic materials. We qualify this statement to apply under *moderate* bias; the LED is operated at least several kT below its bandgap energy. The moderate bias regime, as we will show in Chapter 4, is the most relevant and practical regime of operation as a refrigerator. For the remainder of this chapter, we will discuss our approach to modeling and designing this device, culminating with an evaluation of its main efficiency metrics: η_{ext} and η_{WPE}. In this section, we provide a brief overview of the device features, possible fabrication strategies, and our modeling methodology, before investigating in more detail the specific design issues.

The device uses a thin film of GaAs as the active light-emitting material, which is cladded above and below by lattice-matched GaInP layers to form a double heterostructure. This heterostructure is chosen primarily for the very high quality of the GaAs/GaInP interface, for which extremely low surface recombination velocities have been measured [40], as well as the good carrier confinement provided by the large band offsets. Prior experiments on uncontacted samples have yielded very high photoluminescence quantum efficiencies of 96% at 300K [36] and 99.5% at 100K [37] using this double heterostructure. These results, which are indicative of near-ideal internal luminescence efficiency, suggest that GaAs LEDs should be capable of ultra-high external (electro)luminescence efficiency, far beyond what has been achieved in the best experimental demonstrations.

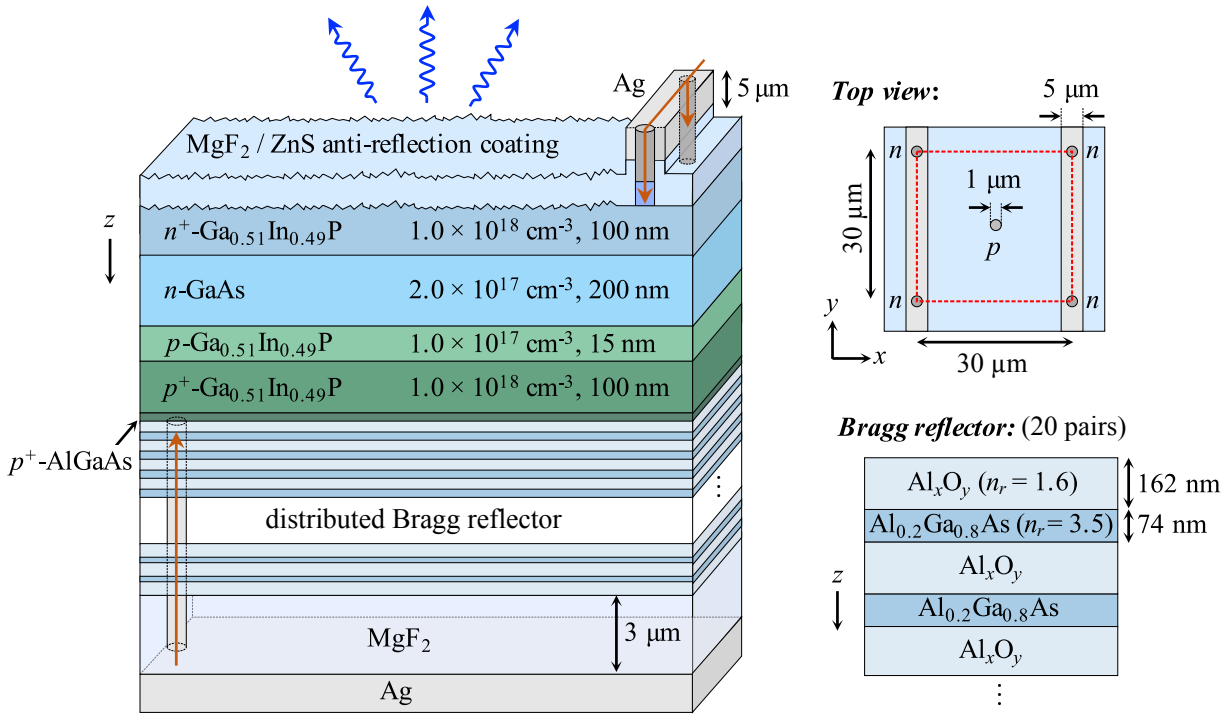

Figure 3.1: Proposed $Ga_{0.51}In_{0.49}P$/GaAs lateral-injection LED designed for efficient external uminescence (not to scale). The full device cross-section is shown on the left, with a close-up of the distributed Bragg reflector shown on the bottom right. The top view shows the layout of the point contacts to the LED (n contacts on front side, p contacts on rear side), where the red dashed square encloses a unit cell that can be repeated over a large area.

Various strategies have been implemented in the proposed device structure to enhance the efficiency of light extraction C_{ext}, which we will discuss in greater depth in Section 3.3. The device contains a textured surface, which allows all of the internal photons – many of which would otherwise be trapped inside by total internal reflection – to couple to external plane waves. On the rear side, the heterostructure is bounded by a hybrid mirror composed of an AlGaAs/oxide distributed Bragg reflector, a low-index MgF_2 total internal reflection layer, and finally a metallic backplane. Together, these layers provide exceptionally high reflectivity to the internal luminescence, promoting their extraction through the front surface.

In order to minimize the loss of luminescent photons at the electrical contacts, we adopt a lateral current injection scheme [16]. Current is supplied to the device from metallic gridlines on the front side and the metallic backplane on the rear side, entering the semiconductor through a two-dimensional array of point contacts. The carriers then spread out laterally in the GaInP cladding layers before recombining in the GaAs active region. At low to moderate current densities, the carriers are distributed uniformly in the lateral dimensions across the full device area, allowing most of the recombination to occur far away from the electrodes.

Thus, a point contact array with small surface coverage minimizes the electrodes' impact on light extraction. For the optimized geometry shown in the top view (top right) in Fig. 3.1, the contacts occupy less than 0.2% of the LED surface area. Though optimal for light extraction, the lateral injection design comes at the cost of large resistive losses at high current densities, since the carriers must travel long lateral distances in the semiconductor layers, which have relatively large sheet resistance. At large biases, the resistive losses also cause the recombination current to crowd near the contacts, greatly exacerbating the problem of contact absorption. Thus, this design is suited primarily for low to moderate power densities. Issues related to current spreading will be discussed in Section 3.4.

Though the structure requires superb material quality and sophisticated device processing to be realized, it is nonetheless designed within the constraints of present optoelectronic technological capabilities. The predicted high performance, presented at the end of this chapter, relies on the longest Shockley-Read-Hall (SRH) lifetime that has been experimentally observed in GaAs, which is 21 μs. This is representative of a material with very low defect density, grown by metal-organic chemical vapor deposition (MOCVD) [37]. The full semiconductor material stack in Fig. 3.1, including the distributed Bragg reflector, can be grown lattice-matched on a GaAs substrate. The Bragg reflector can first be grown as an $Al_xGa_{1-x}As$ stack with alternating Al-rich and Ga-rich layers.

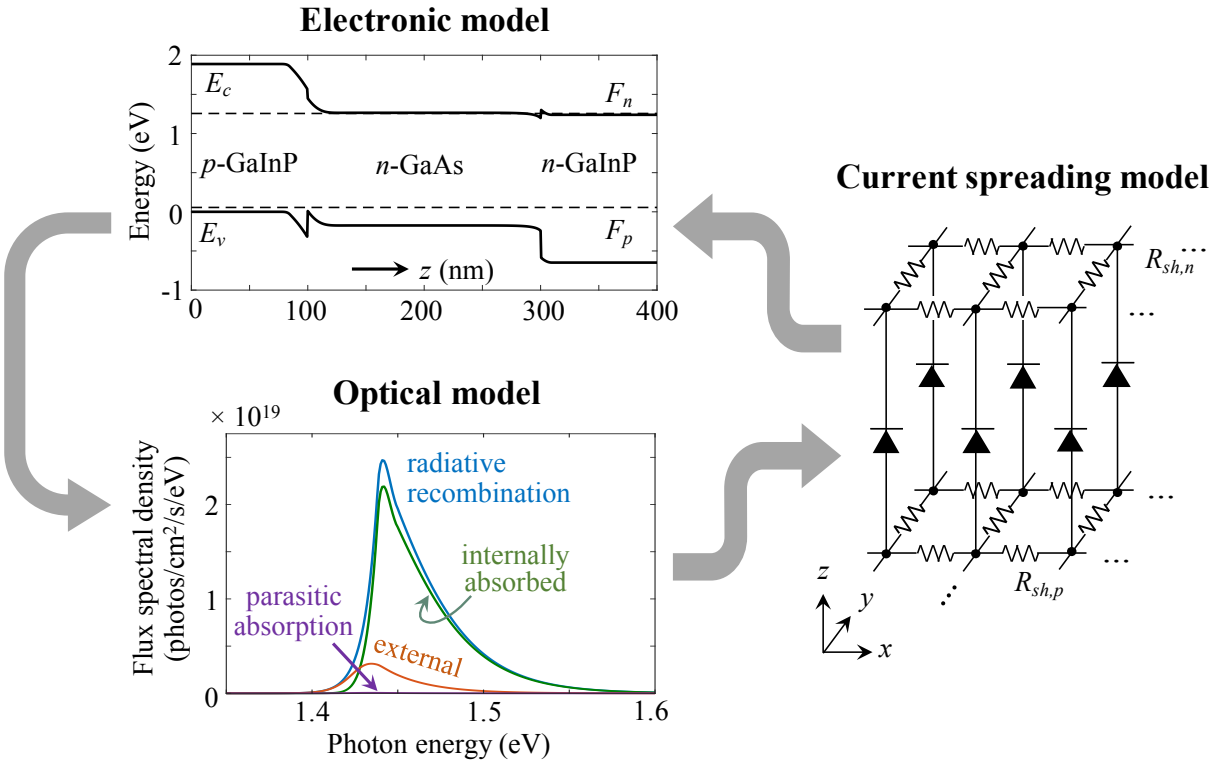

Figure 3.2: Self-consistent modeling methodology for the LED structure in Fig. 3.1.

The epitaxial LED structure can then be separated from the substrate using a selective chemical etch [27], then mounted onto a reflective host; in this case, a Ag surface coated with MgF_2. Further processing would then be needed on the front and rear sides of the device to texture the epitaxial film, deposit an anti-reflection coating, and define the Ohmic contacts, gridlines, and metal vias. Surface texturing at the scale of the emission wavelength can be accomplished in various ways, such as by etching a surface that is randomly coated with nanostructures [41]. The Al-rich AlGaAs layers in the Bragg stack can be converted into its native oxide Al_xO_y through selective wet oxidation [42] to create the oxide Bragg reflector shown in Fig. 3.1 (bottom right). This is a process often exploited in vertical-cavity surface-emitting lasers for optical communications. To maintain an uninterrupted light-emitting surface over a large area (as opposed to a mesa structure), these layers can be made accessible for lateral oxidation through the rear contact openings of the device, prior to metal deposition [43].

Fig. 3.2 summarizes our methodology for modeling the LED structure in Fig. 3.1. The model can be divided into three parts, whose results we ensure to be self-consistent. First, the electronic band profile of the GaAs/GaInP double heterostructure is found for a quasi-Fermi level separation qV in the active layer – initially assumed to be laterally uniform – which yields the carrier densities and recombination rates in the device. For the photons that are generated by radiative recombination, the optical model calculates the relative rates of external light emission, internal band-to-band re-absorption, and parasitic absorption, yielding an initial estimate for the external luminescence efficiency η_{ext}. Next, we model the lateral current flow and voltage drops in the device by using a distributed circuit model (see Fig. 3.2, right), where the current drawn by each distributed diode element is found from the results of both the electronic and optical models. The current spreading model allows us to calculate the Ohmic dissipation losses, as well as the lateral variation in the quasi-Fermi level separation $qV(x, y)$ that appears due to spreading resistance effects, which gives rise to current crowding at larger biases. If these lateral variations are present, we re-visit the electronic and optical models to re-calculate the carrier densities and recombination rates at every position in the device volume (x, y, z), as well as the optical fluxes at every lateral position (x, y) across the device area. These values are integrated to yield an updated value for the external luminescence efficiency that accounts for resistive effects.

In the following sections, we will provide a comprehensive discussion of the individual models and the design strategies that balance the losses that arise in the different domains of LED operation. Our overall goal is to maximize the wall-plug efficiency, given by:

$$\eta_{\text{WPE}} = \frac{\langle E \rangle}{qV} \times \eta_{\text{elec}} \times \left(\eta_{\text{int}} \times C_{\text{ext}}\right) \tag{3.1}$$

where the term in parentheses is the external luminescence efficiency η_{ext}, and we have introduced the electrical efficiency η_{elec}, which we define as the fraction of the injected electrical power that is not lost to resistive losses. In the above, qV refers to the quasi-Fermi level separation in the active region, not the voltage across the device terminals.

3.2 Electronic design

Following their injection into the active region, electrons and holes are confined by the potential barriers at the hetero-interfaces and eventually recombine. The internal luminescence efficiency is the fraction of the total recombination rate in the device that is radiative:

$$\eta_{\text{int}} = \frac{J_{\text{rad}}}{J_{\text{rad}} + J_{\text{srh}} + J_{\text{Auger}} + J_{\text{leak}}} \tag{3.2}$$

where the total supplied forward current per area of the light-emitting surface is resolved into: radiative recombination J_{rad}, Shockley-Read-Hall recombination J_{srh}, Auger recombination J_{Auger}, and carrier leakage out of the active region J_{leak} by diffusion and drift. In this section, we will first discuss the framework used to calculate these rates, then present the design considerations that led to the heterostructure layer stack in Fig. 3.1. The individual recombination rates will be given as a function of the quasi-Fermi level separation qV; the effect of lateral variations in qV (relevant at large biases) will be considered in Section 3.4.

In order to determine the recombination rates in the device, we must first find the band profile of the p^+-GaInP/p-GaInP/n-GaAs/n^+-GaInP material stack. This is done by solving Gauss's law along the growth (vertical) direction, z:

$$\frac{d\mathcal{E}}{dz} = \frac{\rho(z)}{\epsilon(z)} \tag{3.3}$$

where $\mathcal{E}$ is the electric field, ρ is the net charge density, and ϵ is the permittivity. We use a depletion approximation to make an initial guess for $\rho(z)$, and use Gauss's law to obtain the electric field $\mathcal{E}(z)$ and electric potential $\phi(z)$. The potential profile, along with the GaAs/GaInP band offsets [44], are used to determine the band edge profiles $E_c(z)$ and $E_v(z)$ [45]. Over the active region, we assume a uniform quasi-Fermi level splitting with z: $qV = F_n - F_p$. The carrier densities at each position z are then found using

$$\begin{aligned} n &= 2\left(\frac{m_e^* kT}{2\pi\hbar^2}\right)^{3/2} F_{1/2}\left(\frac{F_n - E_c}{kT}\right) \\ p &= 2\left(\frac{m_h^* kT}{2\pi\hbar^2}\right)^{3/2} F_{1/2}\left(\frac{E_v - F_p}{kT}\right) \end{aligned} \tag{3.4}$$

where m_e^* and m_h^* are the conduction and valence band effective masses, $\hbar$ is the reduced Planck's constant, and $F_{1/2}$ is the Fermi-Dirac integral of order 1/2. We use the calculated free carrier densities $n(z)$ and $p(z)$ to update the charge density profile $\rho(z)$, then iteratively solve for the band profile that satisfies charge neutrality in the presence of the free carriers.

The net rate of radiative recombination can be found from the van Roosbroeck-Shockley relation, which is a detailed-balance relationship between the rates of internal photon emission and absorption [46]:

$$J_{\text{rad}}(V) = qd\int_0^\infty \frac{8\pi n_r^2 E^2}{c^2 h^3}\left(\frac{\alpha(E,V)}{e^{(E-qV)/kT}-1} - \frac{\alpha(E,0)}{e^{E/kT}-1}\right)dE \tag{3.5}$$

where d is the active layer thickness, n_r is the active layer refractive index, E is the photon energy, h is Planck's constant, and α is the active layer band-to-band absorption coefficient. The absorption coefficient is given by,

$$\alpha(E,V) = \alpha_0(E) \times (f_\mathrm{v} - f_\mathrm{c}) \tag{3.6}$$

where $\alpha_0(E)$ is the absorption coefficient of intrinsic bulk GaAs in equilibrium, obtained from the experimental results in Ref. 47 at room temperature. The second term accounts for the occupancies of the conduction and valence band electronic states involved in the radiative transition, which depend on the bias qV and the doping density N_D. At a moderate bias $qV < E_g - 3kT$, the absorption coefficient is approximately independent of V.[1]

Often, the radiative recombination current is written in terms of the radiative recombination coefficient B:

$$J_\mathrm{rad}(V) \approx qdB\,(np - n_0 p_0) \tag{3.7}$$

where n_0 and p_0 are the carrier densities in equilibrium. This is an approximate expression that assumes a constant active-region np product, which is slightly violated wherever the carrier densities become degenerate, and a bias-independent value of B, which does not strictly hold due to band-filling effects. Nonetheless, under moderate bias with an n-type doping density of $N_D = 2 \times 10^{17}$ cm^{-3}, we obtain $B = 8 \times 10^{-10}$ cm^3/s for the GaAs layer at 300K, which is consistent with previous reports [48].

We next consider SRH recombination arising from defects in the semiconductor that trap mobile carriers and thereby act as non-radiative recombination centers. The full SRH recombination rate is:

$$J_\mathrm{srh}(V) = J_\mathrm{srh,b}(V) + J_\mathrm{srh,int}(V) + J_\mathrm{srh,per}(V) \tag{3.8}$$

where $J_\mathrm{srh,b}$ encompasses the defects within the bulk volume of the GaAs and GaInP layers, $J_\mathrm{srh,int}$ accounts for traps at the GaAs/GaInP interfaces, and $J_\mathrm{srh,per}$ accounts for traps at the exposed GaAs surfaces along the device perimeter.

The bulk SRH current is found by integrating over the depth z of the material stack, choosing $z = 0$ to mark the interface between p-GaInP and n-GaAs:

$$J_\mathrm{srh,b}(V) = q \int_{-(d_p'+d_p)}^{d+d_n} \frac{1}{\tau_\mathrm{srh,b}(z)} \frac{n(z)p(z) - n_0(z)p_0(z)}{n(z) + p(z) + n_0(z) + p_0(z)}\, dz \tag{3.9}$$

where d_n is the n^+-GaInP cladding thickness, d_p is the p^+-GaInP thickness, and d_p' is the lightly doped p-GaInP thickness. We allow for different bulk SRH recombination lifetimes $\tau_\mathrm{srh,b}$ in GaAs and in GaInP, but due to the large bandgap of GaInP (1.89 eV), the cladding layers are virtually devoid of minority carriers and the SRH recombination in those regions is practically negligible.

[1] We will therefore omit the explicit dependence of α on V in the following, though this is always included in our calculations.

At the GaAs/GaInP interfaces, the SRH recombination rate is given by:

$$J_{\text{srh,int}}(V) = qS\left(\frac{np - n_0p_0}{n + p + n_0 + p_0}\right)\Bigg|_{z=0^+} + qS\left(\frac{np - n_0p_0}{n + p + n_0 + p_0}\right)\Bigg|_{z=d^-} \tag{3.10}$$

where S is the surface recombination velocity at the GaAs/GaInP interface. The first term corresponds to the p-GaInP/n-GaAs interface, and the second term corresponds to the n-GaAs/n-GaInP interface.

The values of $\tau_{\text{srh,b}}$ and S for the GaAs/GaInP heterostructure have been extracted from photoluminescence measurements [37], [40], [49]. The mono-molecular SRH lifetime obtained from these measurements can be resolved into a bulk component and an interface component which scales with thickness:

$$\frac{1}{\tau_{\text{srh}}} = \frac{1}{\tau_{\text{srh,b}}} + \frac{2S}{d} \tag{3.11}$$

From these experiments, a remarkably low value of $S = 1.5$ cm/s has been measured for this interface [40], [49]. Ref. 37 measured a SRH lifetime of $\tau_{\text{srh}} = 21\,\mu$s for a 700 nm thick, MOCVD-grown GaAs film bounded by GaInP, which was separated from the substrate by epitaxial lift-off. This would imply a bulk SRH lifetime of at least 21 μs, depending on the value of S for the sample. For our calculations, we use a value of $\tau_{\text{srh,b}} = 21\,\mu$s to represent GaAs films of the highest quality, even though longer bulk lifetimes have likely been achieved. These values were measured at room temperature.

The perimeter recombination current can be modeled as:

$$J_{\text{srh,per}}(V) = \frac{4J'_{0p}}{L_{\text{total}}}\left(e^{qV/2kT} - 1\right) \tag{3.12}$$

where J'_{0p} is an empirical parameter with units of [A/cm], and the full LED has an area $L_{\text{total}} \times L_{\text{total}}$. Since the bare GaAs surface is highly defective, we desire a large area to minimize perimeter recombination. In Ref. 50, the authors passivated the exposed edges with a coating of As_2S_3, which reduced the surface recombination velocity by 100× and yielded a value of $J'_{0p} = 20$ fA/cm.[2] We use this value for our device, which has the same GaAs active thickness. For a 5 cm × 5 cm LED, which we estimate to be the maximum practical device area, the current density pre-factor in Equation (3.12) is 16 fA/cm^2.

The rate of Auger recombination is given by:

$$J_{\text{Auger}}(V) = q\int_0^d \Big(C_n n(z) + C_p p(z)\Big)\Big(n(z)p(z) - n_0(z)p_0(z)\Big)dz \tag{3.13}$$

where C_n and C_p are the Auger coefficients for the two-electron and two-hole processes, respectively. An overall Auger coefficient of $C = 7 \times 10^{-30}$ cm^6/s has been measured in

[2]This is claimed to be an upper bound, limited by experimental sensitivity: the real value is likely smaller.

intrinsic GaAs at room temperature, and we divide this equally between C_n and C_p [51]. Auger recombination in the wide-bandgap layers is negligible.

Finally, injected carriers may leak occur over the confinement barriers at the edges of the active region, and eventually recombine elsewhere. We treat all such recombination processes as lossy. The total leakage current is calculated by [16]:

$$J_{\text{leak}}(V) = qn_b(V)\left(\frac{D_n^P}{d_p} + \mu_n^P \mathcal{E}_p\right) + qp_b(V)\left(\frac{D_p^N}{d_n} + \mu_p^N \mathcal{E}_n\right) \tag{3.14}$$

where the two terms correspond to electron and hole leakage currents, respectively, each of which has a diffusion and a drift component. The leakage increases exponentially with bias through n_b and p_b, which are the densities of electrons (holes) at $z = 0^+$ ($z = d^-$) with sufficient energy to surmount the potential barrier into the p-GaInP (n-GaInP) cladding layers, respectively. μ_n^P is the electron mobility in p-GaInP, μ_p^N is the hole mobility in n-GaInP, and the corresponding carrier diffusion coefficients D_n^P and D_p^N can be found using the Einstein relation. The electric fields $\mathcal{E}_n$ and $\mathcal{E}_p$ which drive carrier drift into the n- and p-GaInP layers, respectively, are relevant only under high-level injection. We ultimately find that near room temperature, J_{leak} remains a very minor loss due to the large confinement potentials of the double heterostructure.

To set the appropriate doping density for the GaAs active region, we consider the relative rates of radiative and non-radiative recombination processes under low-level injection. With too large a doping level, Auger processes dominate even at moderate bias. At the extreme of low doping, the SRH recombination dominates, since it is roughly proportional to the minority carrier density. Therefore, an optimal doping density maximizes the internal luminescence efficiency. For this device, the optimum is $N_D = 2 \times 10^{17}$ cm^{-3}. We choose n-type rather than p-type doping of the active region in order to reduce the free-carrier absorption of luminescent photons, which is stronger for holes. This choice does not strongly influence the internal luminescence efficiency.

The active GaAs thickness of $d = 200$ nm is chosen both to minimize the non-radiative recombination rate, which prefers a thinner device, and to maximize the luminescent emissivity of the LED, which prefers a thicker device as explained in the next section. The GaAs double heterostructure, which does not exhibit the polarization effects found in GaN-based materials, is efficient enough on its own without the need for additional carrier confinement using multiple quantum wells. This approach also avoids the issue of non-uniform carrier distribution in the wells.

The depletion region of the p-n junction – where the carrier densities can differ dramatically from their values in the quasi-neutral regions – can have an outsized effect on the luminescence efficiency of the device. Depletion-region losses are particularly prominent at low and high biases, but in view of our very high targets for luminescence efficiency, these effects can limit the performance at moderate bias as well. As illustration, we first consider the case where the lightly doped p-GaInP layer in our material stack has been removed. The band profile and the spatially resolved recombination rates for this case are shown in

Fig. 3.3(a). At a moderate bias of $qV = 1.2$ eV, the radiative recombination rate is several orders of magnitude larger than the non-radiative recombination rates in the quasi-neutral region of the GaAs active layer. However, inside the depletion region, both the SRH and Auger recombination rates have large peaks, reducing the internal luminescence efficiency. The former is a well-known effect in any diode [52], and dominates the device current at low bias. This depletion-region (or space-charge) SRH current grows as $\exp(qV/2kT)$, and is eventually overtaken by the radiative and other recombination currents with increasing bias. In Fig. 3.3, we have used a bulk SRH lifetime of 21 μs in GaAs and assumed a much shorter lifetime of 100 ns in GaInP. In spite of the 200× shorter lifetime, SRH recombination in the wide-bandgap cladding is still clearly negligible.

The large peak in the Auger recombination rate in Fig. 3.3(a) is due to the accumulation of holes near the n-GaAs/p^+-GaInP interface, caused by strong band-bending on the GaAs side of the depletion region. We can suppress this hole-accumulation region by inserting a

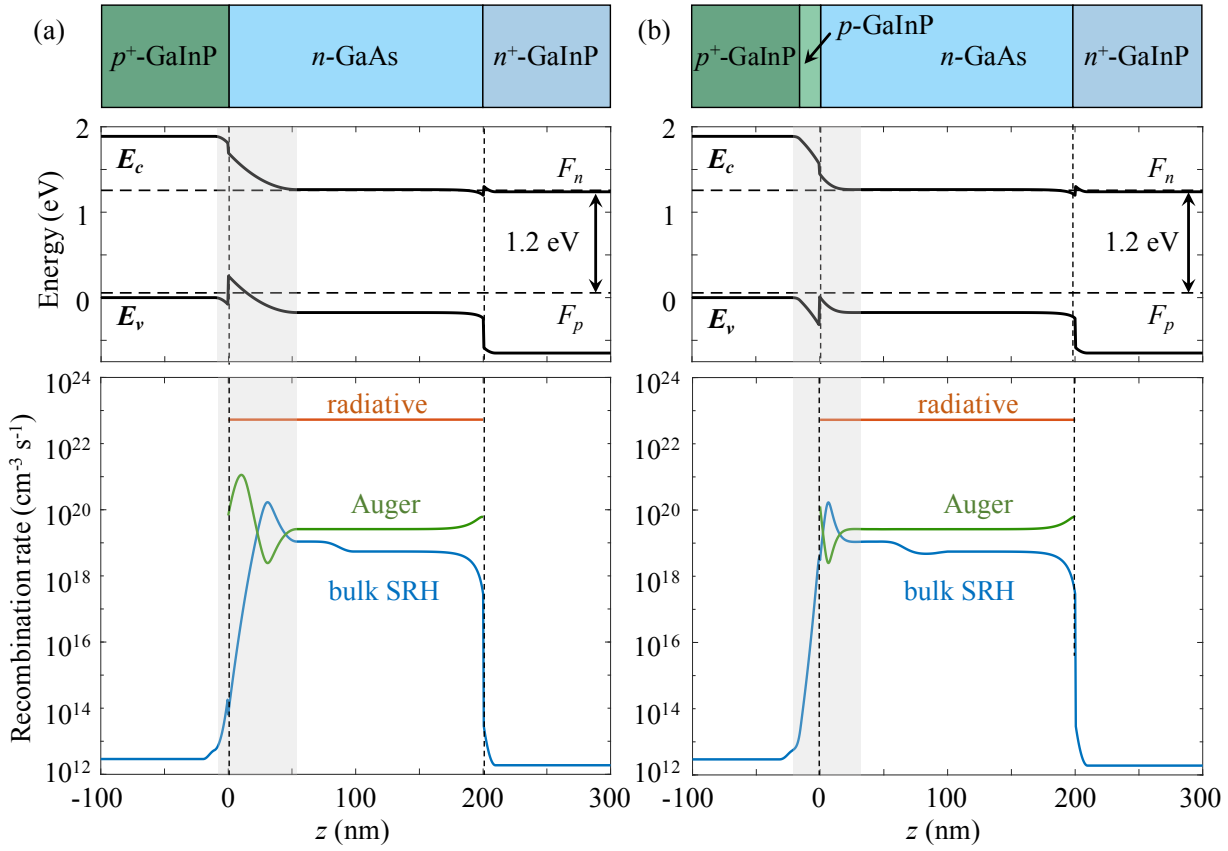

Figure 3.3: The spatial profile of the band edges and recombination rates in the p-GaInP/n-GaAs/n-GaInP double heterostructure under a forward bias of $qV = 1.2$ eV: (a) without and (b) with the thin, lightly doped p-GaInP Auger suppression layer. The depletion region of the p-n junction is denoted in gray. The interface recombination rates are not shown.

lightly p-doped GaInP layer, which acts as a thin buffer between n-GaAs and the heavily doped p^+-GaInP cladding. This layer moves the depletion region further into the wide-bandgap material and reduces the band-bending in GaAs, removing the Auger peak at moderate bias, as shown in Fig. 3.3(b).

Fig. 3.4(a) and (b) show the internal luminescence efficiency η_{int} of the LED structure in Fig. 3.1 as a function of the quasi-Fermi level splitting voltage at 263K and 300K, respectively. The non-radiative loss components due to SRH recombination, Auger recombination, and carrier leakage are resolved. The depletion-region and perimeter SRH recombination currents dominate at low voltage, while Auger recombination dominates at large bias. At 300K, a broad region of high efficiency lies between these two regimes, with a peak value of $\eta_{\text{int}} = 99.85\%$ at a moderate bias of 1.22V, which exceeds the very high value of 99.7% that has been measured in the AlGaAs/GaAs double heterostructure [48]. In Fig. 3.4(a), we evaluate the performance at a lower temperature of 263K, relevant for near-ambient refrigeration applications. The values of a large number of parameters in the model are adjusted for this lower temperature: the details are described more fully in Section 4.6. At this temperature, the device achieves a peak efficiency of $\eta_{\text{int}} = 99.91\%$ at a bias of 1.276V. The superior performance at lower temperature is a consequence of the suppression of both non-radiative recombination and hot carrier leakage, in combination with an enhancement in radiative recombination (at a given carrier density). We will discuss this point more fully in Section 4.6, which explores LED operation at cryogenic temperatures.

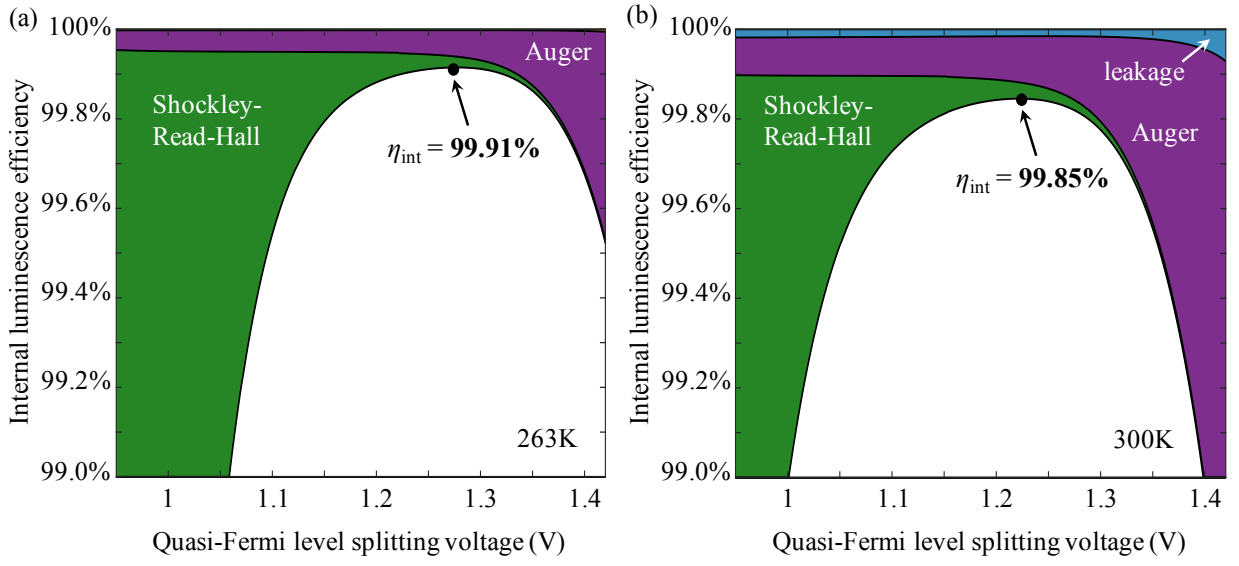

Figure 3.4: The internal luminescence efficiency η_{int} of the LED in Fig. 3.1 at (a) 263K and (b) 300K. The loss components responsible for deviations from $\eta_{\text{int}} = 100\%$ are shown, and the peak values of efficiency are labeled.

3.3 Optical design

The most important task in LED design is to ensure the rapid extraction of light out of the semiconductor. The longer a luminescent photon lingers inside the LED, the greater its chance to be lost by parasitic absorption or by non-radiative recombination if the photon is recaptured as an electron-hole pair. In this way, poor light extraction amplifies the effects of all internal carrier and photon losses. In simple planar device geometries, where the photons are confined as waveguide modes by total internal reflection, the external luminescence efficiency η_{ext} is very strongly sensitive to these losses by default. Though we cannot modify the size of the escape cone from GaAs ($n_r = 3.5$) into free space, it is possible to design the device geometry in ways that enhance the extraction of luminescence, thereby reducing the sensitivity of η_{ext} to the internal losses.

Radiative recombination pumps light into an internal reservoir of luminescent photons. The photon reservoir can also return its energy to the reservoir of charge carriers by band-to-band electronic absorption. In steady state, the *net* rate of photon addition to the reservoir must be balanced by the rate of irretrievable photon removal from the device by external emission and internal dissipation. The light extraction efficiency is the ratio of the external emission rate Φ_{ext} to the total rate of photon removal from the device:

$$C_{\text{ext}} = \frac{\Phi_{\text{ext}}}{\Phi_{\text{ext}} + \Phi_{\text{rear}} + \Phi_{\text{nr}} + \Phi_{\text{fc}} + \Phi_{\Omega}} \tag{3.15}$$

We have included here the photon losses that occur through rear surface absorption Φ_{rear}, non-radiative recombination following absorption in the active layer Φ_{nr}, free-carrier absorption Φ_{fc}, and absorption by the Ohmic contacts Φ_{Ω}. We note that C_{ext} is not to be confused with the escape probability P_{esc}, whose relationship to C_{ext} is explained in Section 2.1. As in the previous section, we will find the expression for each term above as a function of the quasi-Fermi level separation qV: effects of lateral variations in qV, such as current crowding, will be explored in the next section.

Randomly texturing the surface of the semiconductor, as featured in the structure in Fig. 3.1, is a well-known technique for enhancing light extraction in LEDs [41].[3] A textured surface randomizes the angle of a photon that scatters from it, giving the photon a renewed opportunity to escape with every pass through the film. The luminescence is thus extracted more rapidly and with fewer opportunities for loss in comparison to a planar device, where trapped internal photons must rely instead on electronic absorption and re-emission to enter the escape cone (see Fig. 2.2). To model the effect of the texture, we make the ergodic assumption: on being scattered many times by the texture, the internal photons experience complete randomization of their trajectories and collectively, they can be treated as an

[3]Other possible methods to enhance light extraction include the incorporation of a diffraction grating (e.g. by substrate patterning) and geometrical shaping the LED chip into a truncated inverted pyramid or hemispherical dome [14]. An ideal random texture provides the maximum possible angular randomization while remaining compatible with a thin film geometry.

isotropic photon gas [53]. Though the scattering properties of roughened surfaces can be complex [54], it has been found experimentally that by choosing the length scale of the texture to be comparable to the emission wavelength, optical absorption properties have been observed that are the signature of ergodicity [55]. This allows the internal photon dynamics to be studied using a simple statistical ray optics approach, where the individual photon outflow rates in Equation (3.15) can be found by integrating the spectral brightness b_{int} of the photon gas (which has units of photons per time, area, solid angle, and energy).

The external emission rate given by the isotropic internal photon gas model is:

$$\Phi_{\text{ext}}(V) = \int_0^\infty dE \int_0^{\theta_c} T(E,\theta_{\text{i}})\, b_{\text{int}}(E,V)\, 2\pi \sin\theta_{\text{i}} \cos\theta_{\text{i}}\, d\theta_{\text{i}} \tag{3.16}$$

where θ_{i} is the photon angle inside the GaAs film measured from the z-axis, $T(E,\theta_{\text{i}})$ is the front transmission coefficient averaged over polarization, and $\theta_{\text{c}} = \sin^{-1}(1/n_r) = 16.6^\circ$ is the critical angle of the emitting surface, which we will hereafter call the front surface. We assume that all quantities are independent of the photon's in-plane angle.

Alternatively, the rate of external emission from the device can always be found by applying the principle of detailed balance to the LED surface that faces the external environment:

$$\Phi_{\text{ext}}(V) = \int_0^\infty dE \int_0^{\pi/2} \frac{2E^2}{c^2h^3} \frac{a(E,\theta)}{e^{(E-qV)/kT}-1}\, 2\pi \sin\theta\, \cos\theta\, d\theta \tag{3.17}$$

where θ is emission angle into vacuum and $a(E,\theta)$ is the band-to-band electronic absorptivity of the device, which is equal to the luminescent emissivity of the device by Kirchoff's law. Unlike thermal radiation, the emitted luminescent radiation possesses a finite chemical potential μ that is equal to qV, the quasi-Fermi level separation of electrons and holes in the device active region [56]. This equality establishes a condition of quasi-equilibrium between the excited charge carriers and the luminescent photons that they generate via recombination. Note that in the case where $(E - qV) \gg kT$, relevant for moderate-bias operation, the equation reduces to the form given by Equation (2.7).

The two expressions for Φ_{ext} above must yield the same result. By enforcing this consistency, we find an expression for the internal spectral brightness:

$$b_{\text{int}}(E,V) = \frac{1}{\overline{T}(E)} \frac{2n_r^2E^2}{c^2h^3} \frac{a(E)}{e^{(E-qV)/kT}-1} \tag{3.18}$$

where $\overline{T}$ is the front transmission coefficient averaged over the escape cone. Because b_{int} is isotropic, the luminescent emissivity $a(E,V)$ must also be angle-independent for the randomly textured device. Under the ergodic assumption, the expression for the emissivity is found to be [53]:

$$a(E) = \frac{\overline{T}(E)\alpha(E)d}{\alpha(E)d + \overline{T}(E)/4n_r^2 + \mathcal{L}(E)} \tag{3.19}$$

where $\mathscr{L}$ contains the effect of the parasitic losses: $\mathscr{L}(E) = a_{\rm fc} + a_\Omega/4 + (1-\overline{R})/4$. These individual loss components will be discussed below. Equation (3.19) is strictly valid under weak internal absorption ($\alpha d \ll 1$), weak loss ($\mathscr{L} \ll 1$), and a narrow cone of escape ($\overline{T}/4n_r^2 \ll 1$), all of which must hold to maintain ergodicity. In the optimized LED of Fig. 3.1, we will find that the latter two conditions apply for all luminescent photon energies. For the strongly absorbed photons, which violate the first condition, the expression is still valid to a good approximation [5]. Relative to a planar device, surface texturing greatly enhances the emissivity; the factor could be as large as $4n_r^2$ for the weakly absorbed photons near the bandgap energy [53]. Equivalently, texturing allows the LED to achieve a high emissivity with a smaller active thickness, which helps mitigate the volumetric device losses such as non-radiative recombination and free-carrier absorption.

To further facilitate extraction, we add an optimized two-layer anti-reflection coating (145 nm MgF_2 with $n_r = 1.375$, 80 nm ZnS with $n_r = 2.3$) to the front surface, which has been used in III-V solar cells [57]. Using the transfer matrix method, we calculate a transmission coefficient of 79.9% for this material stack when averaged over the energy, angle, and polarization of luminescence and corrected for the obstruction of the front surface by metal gridlines. By comparison, an uncoated semiconductor surface has a transmission coefficient of 56.5%.

We now consider the parasitic optical losses, which must be kept to a minimum for ultra-high external luminescence efficiency, despite the reduced sensitivity to these losses provided by surface texturing. Most important of these losses is absorption by the rear surface, since the internal photons must still still experience many rear reflections on average before escaping. The rate of rear photon loss is given by

$$\Phi_{\rm rear}(V) = \int_0^\infty dE \int_0^{\pi/2} \Big[1 - R(E,\theta_{\rm i})\Big]\, b_{\rm int}(E,V)\, 2\pi \sin\theta_{\rm i} \cos\theta_{\rm i}\, d\theta_{\rm i} \tag{3.20}$$

where $R(E,\theta_{\rm i})$ is the polarization-averaged rear reflectivity. Fig. 3.5(a) shows the reflectivity of the composite rear reflector illustrated in Fig. 3.1, calculated using the transfer matrix method and averaged over polarization. Fig. 3.5(b) shows the angle-dependent reflectivity after averaging over the energy distribution $b_{\rm int}(E,V)$ of the luminescent photons. The low-index MgF_2 ($n_r = 1.375$) layer provides total internal reflection above the rear critical angle with GaAs of $\theta_{\rm cr} = 23°$, and is thick enough to fully suppress the penetration of evanescent waves. Meanwhile, the high index contrast of the $Al_{0.2}Ga_{0.8}As/Al_xO_y$ Bragg reflector[4] opens up a photonic bandgap that spans the nearly the full luminescence spectrum for angles below 23°[58]. A silver backplane reflects the remaining photons that penetrate through both the Bragg reflector and the MgF_2 layer. The angle-, energy- and polarization-averaged reflectivity of the mirror is 99.990%. This reflectivity can be maintained in the presence of parasitic absorption within the stack of 0.1 cm^{-1} or less, which can be expected for a Bragg stack that is left fully undoped. To achieve high reflectivity, we forego the conduction of current through the rear reflector.

[4]The Bragg reflector has twenty $AlGaAs/Al_xO_y$ layer pairs. The thickness of each layer is close to a quarter of the peak emission wavelength. The implementation of this structure was discussed in Section 3.1.

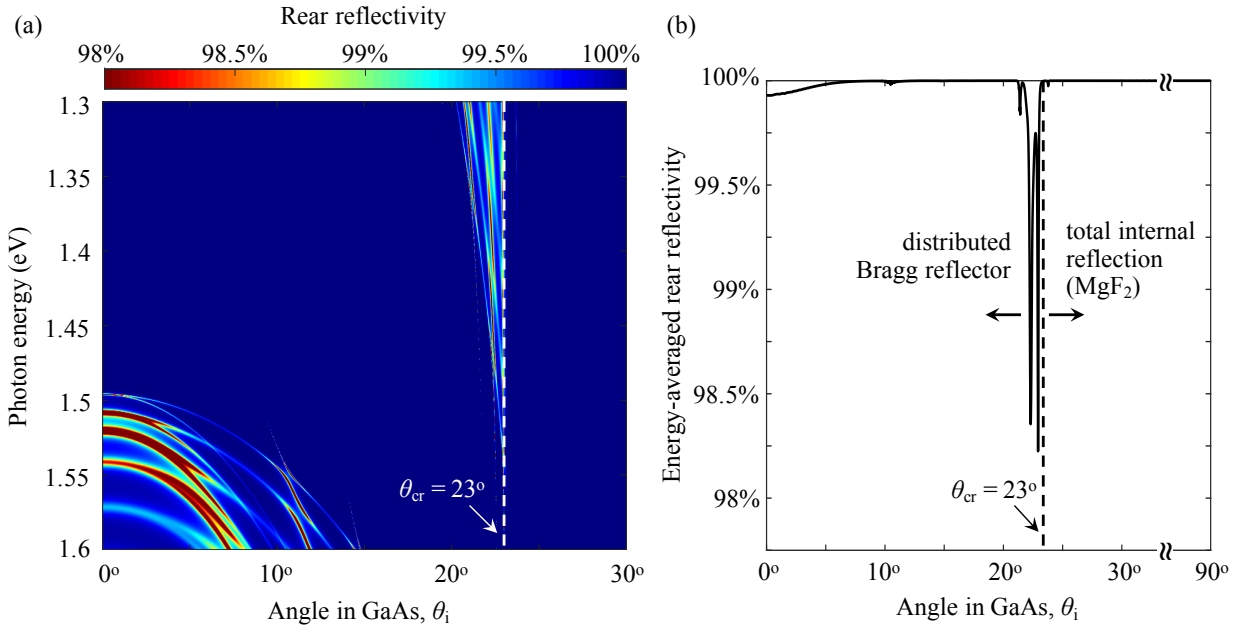

Figure 3.5: The angle-dependent reflectivity of the composite rear reflector of the LED in Fig. 3.1: (a) energy-resolved and (b) averaged over the spectrum of the internal photons. All quantities are calculated by the transfer matrix method and averaged over polarization. The low-index MgF_2 layer ensures total reflection for $\theta_i \geq \theta_{cr} = 23°$, the rear critical angle. For $0 \leq \theta_i \leq 21°$, the Bragg reflector (20 periods) enables total reflection over nearly the full luminescence bandwidth (see Fig. 3.6). The energy-, polarization- and angle-averaged reflectivity is 99.990%.

When a photon is absorbed within the active region volume, generating an electron-hole pair, it may fail to rejoin the internal photon gas if the carriers recombine non-radiatively. The effective rate of photon loss due to non-radiative recombination is given by:

$$\Phi_{nr}(V) = (1 - \eta_{int}) \int_0^\infty dE \int_0^\pi \alpha d\, b_{int}(E, V)\, 2\pi \sin\theta_i\, d\theta_i \tag{3.21}$$

This loss channel can be suppressed by improving η_{int} and by introducing surface texturing, which reduces the necessary thickness of the semiconductor to produce a given emitted flux. Through this term, the light extraction efficiency is closely coupled to the internal luminescence efficiency, as explained in Section 2.1.

Free-carrier absorption in the semiconductor volume is an intrinsic loss mechanism, with a rate given by:

$$\Phi_{fc}(V) = \int_0^\infty dE \int_0^\pi a_{fc}\, b_{int}(E, V)\, 2\pi \sin\theta_i\, d\theta_i \tag{3.22}$$

where the free-carrier absorptivity is

$$a_{fc} = \int_{-(d'_p + d_p + d_{p\Omega})}^{d+d_n} \Big[\sigma_n(z)n(z) + \sigma_p(z)p(z)\Big] dz \tag{3.23}$$

where σ_n and σ_p are the absorption cross-sections of free electrons and free holes, respectively. In GaAs, cross-sections of $\sigma_n = 3 \times 10^{-18}$ cm^2 and $\sigma_p = 7 \times 10^{-18}$ cm^2 have been measured for luminescent photons at room temperature [59]. Owing to lack of published data, we use the same values for the GaInP cladding layers, as well as a p^+-AlGaAs contact layer with thickness $d_{p\Omega} = 10$ nm and doped to $10^{18}\,\mathrm{cm}^{-3}$. Though all of the layers are assumed to have the same cross-sections at 300K, these values are adjusted differently with temperature using the known temperature dependence of their respective carrier mobilities, as will be explained in Section 4.6. Since the holes absorb the internal photons more strongly, the p^+-GaInP layer must be made no thicker than is necessary to provide adequate current spreading at moderate bias. This also motivates the use of an n-type GaAs active layer. The conflicting goals of low free-carrier absorption and low spreading resistance lead to one of the main trade-offs in the design of this LED, as we will discuss in the next section.

Finally, the rate of parasitic absorption by the LED contacts and front metallization is given by:

$$\Phi_\Omega(V) = \int_0^\infty dE \int_0^{\pi/2} a_\Omega \, b_{\mathrm{int}}(E, V) \, 2\pi \sin\theta_\mathrm{i} \cos\theta_\mathrm{i} \, d\theta_\mathrm{i} \tag{3.24}$$

We model the contact absorptivity as $a_\Omega = f_\Omega\,(1 - R_\Omega) + f_g\,(1 - R_g)$, where f_Ω and R_Ω are the surface coverage and reflectivity of the Ohmic contacts on both sides, and f_g and R_g are the corresponding quantities for the front-side metal gridlines. In the structure shown in Fig. 3.1, the metal penetrates the anti-reflection coating into the n-type semiconductor only at the positions of the Ohmic contacts, which form an array of small points. On the rear side, point contacts to the p-type semiconductor are connected to the Ag backplane by metal vias that punctuate the rear reflector. For the contact geometry shown in Fig. 3.1, the surface coverage of these point contacts is only $f_\Omega = 0.17\%$. This scheme, similar to the point contact silicon solar cell [60], greatly mitigates the otherwise detrimental effect of the electrodes on light extraction. The reflectivity of a typical Ohmic contact to GaAs can be as high as $R_\Omega = 90\%$, based on prior calculations [61]. The gridlines have a larger coverage of $f_g = 16.7\%$, but since the metal lies above the anti-reflection coating, total internal reflection by the front dielectric layers allows for a high internal reflectivity. Using the transfer matrix method, we obtain $R_g = 99.85\%$ (averaged over energy, angle, and polarization).

Like free-carrier absorption, contact absorption has a strong trade-off with spreading resistance. Placing the point contacts further apart can reduce their effect on light extraction, but this also means that the carriers must travel a long lateral distance across the resistive semiconductor layers: the Ohmic losses, therefore, increase sharply. If this effect is severe, the injected current – and therefore the emitted luminescence – will tend to crowd near the contacts, leading to both high resistive losses and very poor light extraction efficiency. To avoid these effects, light extraction must be carefully co-optimized with lateral current spreading, which will be discussed in the next section.

Fig. 3.6 shows the spectrum of the inflow and outflow rates to and from the internal photon gas for the optimized device in Fig. 3.1, evaluated under a bias of 1.2V. While the majority of the internal luminescence is re-absorbed, most of these photons rejoin the internal

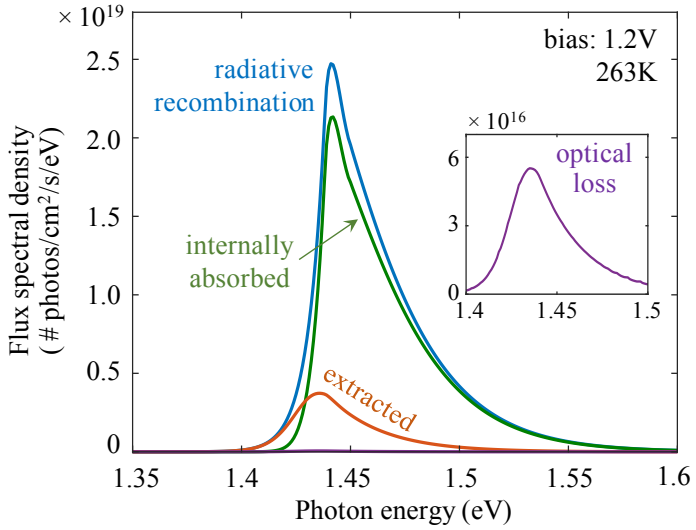

Figure 3.6: The spectrum of the steady-state rates of pumping and removal from the internal photon gas at a quasi-Fermi level separation of 1.2 eV. Most of the internally emitted light (blue) is electronically absorbed by the GaAs active region (green), but most of these carriers are recovered as photons. Optimal design of the LED allows the rate of photon extraction (red) to greatly exceed the rate of parasitic absorption (inset) by the rear surface, free carriers, and the electrical contacts.

photon gas by radiative recombination, owing to the high internal luminescence efficiency of the device. The external emission rate is considerably larger than the parasitic loss rate, leading to a high extraction efficiency – the relative strengths of the parasitic mechanisms will be shown in Section 3.5. The emission spectrum maintains the same shape at different biases until the quasi-Fermi level separation approaches the bandgap energy; at this point, the emission peak shifts to higher energies due to the band-filling effect.

3.4 Current spreading design

Efficient current spreading is a vital consideration for attaining high LED efficiency. The spatial profile of the lateral currents in our LED structure is shown in Fig. 3.7. The forward current is injected through the p-type terminals and collected at the n-type terminals. Though this lateral injection scheme is suited for maximal light extraction, it does necessitate the transport of charge over a considerable distance across thin semiconductor sheets. Compared to other designs for current spreading, the resistive losses become significant at a relatively low power density. Since the holes have lower mobility, these losses are incurred primarily in the p-type layers and the diode current tends to crowd near the p-type contacts. In this section, we will study the two main consequences of these losses on device performance: the direct effect of Ohmic losses on the electrical efficiency η_{elec}, and the effect of current crowding on external luminescence efficiency η_{ext}.

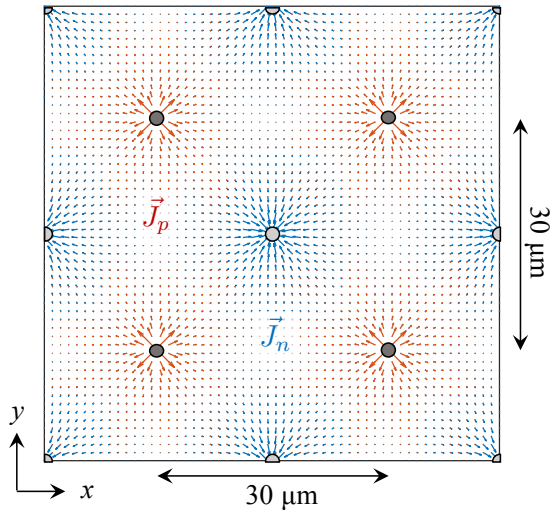

Figure 3.7: Calculated distribution of the lateral currents in the LED shown in Fig. 3.1. The currents $\vec{J}_n$ and $\vec{J}_p$ flow in the n-type and p-type current spreading layers, respectively, though they are overlaid here on the same plot. The n-type contacts are shown in light gray and the p-type contacts are in dark gray. A device area encompassing four unit cells is shown.

So far, we have assumed that the quasi-Fermi level separation qV in the active region is uniform across the device area. At larger biases, where the effects of spreading resistance cannot be ignored, this is an approximation: in reality, the quasi-Fermi level separation is a function of lateral position $V(x, y)$. This areal distribution is found from the coupled current continuity equations for the surface currents $\vec{J}_n$ and $\vec{J}_p$. We refer the reader to Appendix B for the details of this model. In the following analysis, it will be convenient to define an *internal voltage* $\overline{V}$:

$$J\left(\overline{V}\right)\overline{V} \equiv \frac{1}{L_c^2}\iint J(x,y)\,V(x,y)\,dx\,dy \tag{3.25}$$

where L_c is the separation between neighboring contacts of the same type, and the integral is over a unit cell of the device (red dashed square in Fig. 3.1). The current density for a given voltage, accounting for all internal carrier and photon losses, is given by:

$$J = q\Phi_{\text{ext}} + J_{\text{srh}} + J_{\text{Auger}} + J_{\text{leak}} + q\left(\Phi_{\text{rear}} + \Phi_{\text{fc}} + \Phi_{\Omega}\right) \tag{3.26}$$

where every term is now a function of the local quasi-Fermi level separation. We can interpret the internal voltage $\overline{V}$ as the laterally-averaged value of the quasi-Fermi level separation, weighted by the distribution of current over the device area. At low biases where spreading resistance effects are negligible, $V(x, y) \approx \overline{V}$.

For a given internal voltage, the electrical efficiency introduced in Equation (3.1) is:

$$\eta_{\text{elec}} = \frac{J(\overline{V})\overline{V}}{J(\overline{V})\overline{V} + Q_{\Omega}} \tag{3.27}$$

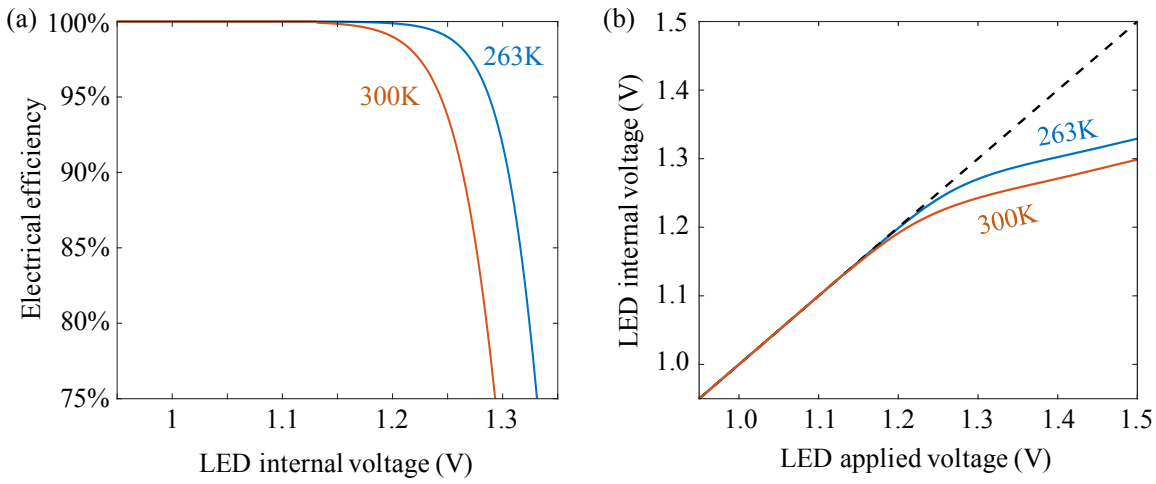

Figure 3.8: (a) The electrical efficiency as a function of the LED internal voltage $\overline{V}$. (b) The applied voltage V_{A} required to attain a given internal voltage. The dashed line represents $\overline{V} = V_{\mathrm{A}}$.

where Q_Ω is the power density that is lost to Ohmic heating of the structure. These losses arise primarily from the n-type and p-type spreading resistances, but we also include the series resistances at the device terminals: the Ohmic contact resistance (assuming a contact resistivity of $\rho_c = 10^{-6}\ \Omega\ \mathrm{cm}^2$ [62]), the front metallization resistance, and the rear metal via resistance. The full expression for Q_Ω is derived in Appendix B, and relies on the solution for the surface currents $\vec{J}_n$ and $\vec{J}_p$. In the following, we discuss the results of this model and their implications on design.

The electrical efficiency of the device is shown in Fig. 3.8(a) as a function of the internal voltage at 263K and 300K. Fig. 3.8(b) shows the voltage that must be applied across the device terminals to obtain the corresponding internal voltage. For $\overline{V} \gtrsim 1.25\mathrm{V}$ (for 263K), the electrical efficiency drops sharply due to Ohmic losses, and the internal voltage no longer directly tracks the applied voltage due to resistive voltage drops. This is a clear manifestation of the trade-off incurred in spreading resistance by optimizing for light extraction at moderate bias (1.15V to 1.3V at 263K). The device parameters most relevant for this trade-off are the contact spacing L_{c} and the p^+-GaInP thickness d_p. We have chosen to use $L_{\mathrm{c}} = 30\ \mu\mathrm{m}$ and $d_p = 100$ nm in order to minimize parasitic absorption by the contacts and by free holes, respectively, while still maintaining a high electrical efficiency of $\eta_{\mathrm{elec}} = 99.0\%$ at $\overline{V} = 1.25\mathrm{V}$. A smaller contact spacing or a thicker p-cladding layer would allow for a higher electrical efficiency at larger biases, but would lead to a smaller external luminescence efficiency at moderate bias. Ultimately, we choose to sacrifice the efficiency at high power density for superior performance at moderate bias: this is the most practical regime of operation for electroluminescent cooling, as we discuss in Section 4.5.

In addition to their effect on electrical efficiency, resistive losses are also harmful to the external luminescence efficiency at higher biases. Fig. 3.9 shows the lateral voltage and

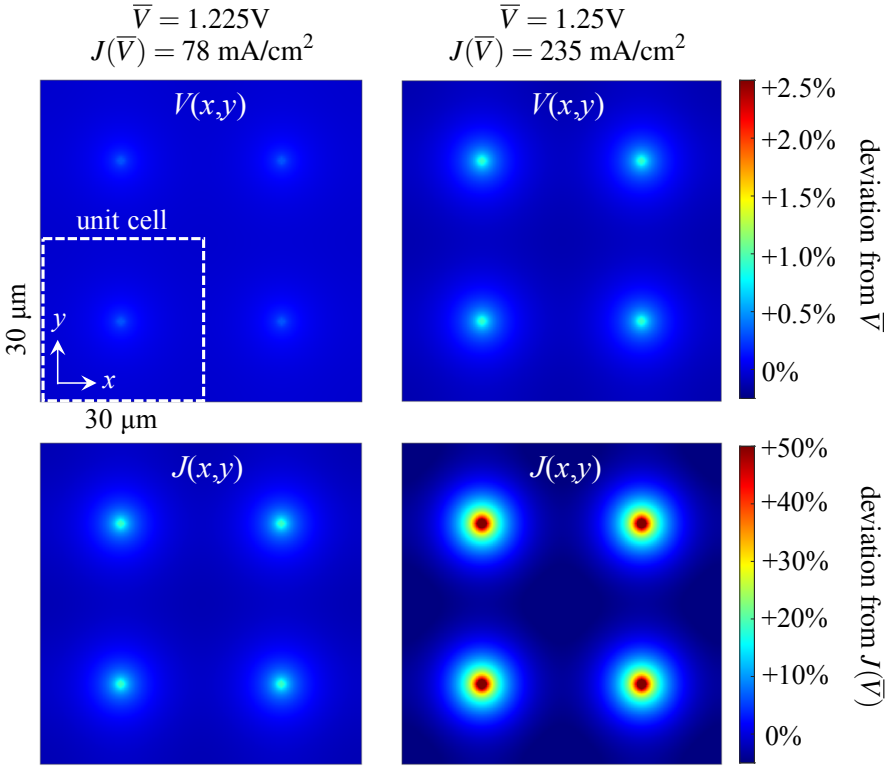

Figure 3.9: The lateral distribution of the quasi-Fermi level separation voltage $V(x, y)$ and the diode current density $J(x, y)$ when the device is biased to an internal voltage of $\overline{V} = 1.225$V and $\overline{V} = 1.25$V. Each plot is evaluated at 263K and shows four unit cells of the device.

current distributions $V(x, y)$ and $J(x, y)$ at two values of the internal voltage. At the smaller bias of $\overline{V} = 1.225$V, current crowding near the p-contacts starts to become visible, but the current density in these confined regions is only 20% greater than the average current density in the device: the effect is too small to influence η_{ext}. At the larger bias of $\overline{V} = 1.25$V, current crowding has a more significant effect, and absorption by the p-contacts becomes more pronounced, though not yet detrimental. Above this bias level, contact absorption rapidly comes to dominate the optical losses, as we will see in the next section.

To describe these effects, we must modify our models for the various recombination and optical fluxes in our device, which have so far assumed a uniform quasi-Fermi level separation. The modified fluxes can then be inserted into Equation (3.2) and Equation (3.15) for η_{int} and C_{ext}, respectively, to calculate the external luminescence efficiency η_{ext} for a given internal voltage $\overline{V}$. The expressions for the recombination currents introduced in Section 3.2 depend on lateral position only through the quasi-Fermi level separation. Therefore, they can be corrected in a straightforward manner as follows:

$$J_{\text{x}}(\overline{V}) = \frac{1}{L_{\text{c}}^2} \iint J_{\text{x}}\Big(V(x, y)\Big)\, dx\, dy \tag{3.28}$$

where the integral is taken over a unit cell and J_{x} refers to any of J_{rad}, J_{Auger}, $J_{\mathrm{srh,b}}$, $J_{\mathrm{srh,int}}$, and J_{leak}. The expression for the perimeter SRH current $J_{\mathrm{srh,per}}$ need not be modified, since the effect is isolated to the device edges. As long as the edges are closer in proximity to the n-type contacts than to the p-type contacts, we can calculate $J_{\mathrm{srh,per}}(\overline{V})$ by Equation (3.12) as before, taking the internal voltage to be the quasi-Fermi level separation at the edges.

Before we proceed to modify the expressions for the optical fluxes in Section 3.3, we must check our core assumptions: is the isotropic photon gas still a valid model when a large lateral variation exists in the internal photon emission rate? For the photons to remain isotropic, they must continue to meet the requirements of ergodicity: on a single pass through the device, there must be a low probability of electronic absorption ($\alpha d \ll 1$), extraction ($\overline{T}/4n_r^2 \ll 1$), and parasitic loss ($\mathscr{L} \ll 1$). A higher emission rate near the contacts enhances the probability that any given photon in the device will be lost by contact absorption. The loss probability $\mathscr{L}$ remains much smaller than unity at moderate biases, but this is no longer the case at large biases ($\overline{V} \gtrsim 1.3\mathrm{V}$). In this regime, isotropy of the photon gas must be regarded as an approximation, since the contacts will more strongly absorb photons with angles close to the surface normal. Since we are primarily interested in the moderate-bias performance, we proceed with the isotropic photon gas model. The individual fluxes are modified as:

$$\Phi_{\mathrm{x}}(\overline{V}) = \frac{1}{L_{\mathrm{c}}^2} \iint \Phi_{\mathrm{x}}\Big(V(x,y),x,y\Big)\, dx\, dy \tag{3.29}$$

where Φ_{x} refers to Φ_{ext}, Φ_{rear}, Φ_{nr}, Φ_{fc}, or Φ_{Ω}. For the first four of these fluxes, the insertion of $V(x,y)$ into the internal spectral brightness $b_{\mathrm{int}}(E,V)$ provides an accurate correction to the model. For contact absorption, we must also modify the contact absorptivity as:

$$a_{\Omega}(x,y) = F_{\Omega}(x,y)\,(1-R_{\Omega}) + F_g(x,y)\,(1-R_g) \tag{3.30}$$

where $F_{\Omega}(x,y)$ and $F_g(x,y)$ are now binary functions – 0 or 1 – that indicate whether the semiconductor at the position (x,y) coincides with an Ohmic contact or the front metallization, respectively. At moderate bias, contact absorption is one of the primary losses in the device. At large biases, when current crowding is severe, it is the dominant source of loss along with Ohmic dissipation.[5]

[5]There is a possibility to reduce contact absorption, and to partially break the trade-off between contact absorption and spreading resistance, through epitaxial regrowth. An n-GaInP layer can be grown on the p-GaInP cladding, then etched away everywhere except for small patches close to the p-type contacts, defined by photolithography. Epitaxial growth of the remaining layers then proceeds. The buried n-GaInP current blocking layers introduce an additional potential barrier for carrier injection into the active region, suppressing light emission in the vicinity of the contacts [16]. Though this scheme would come at a cost to process complexity, similar buried structures have been realized as current apertures for vertical-cavity surface-emitting lasers in the same material system [63].

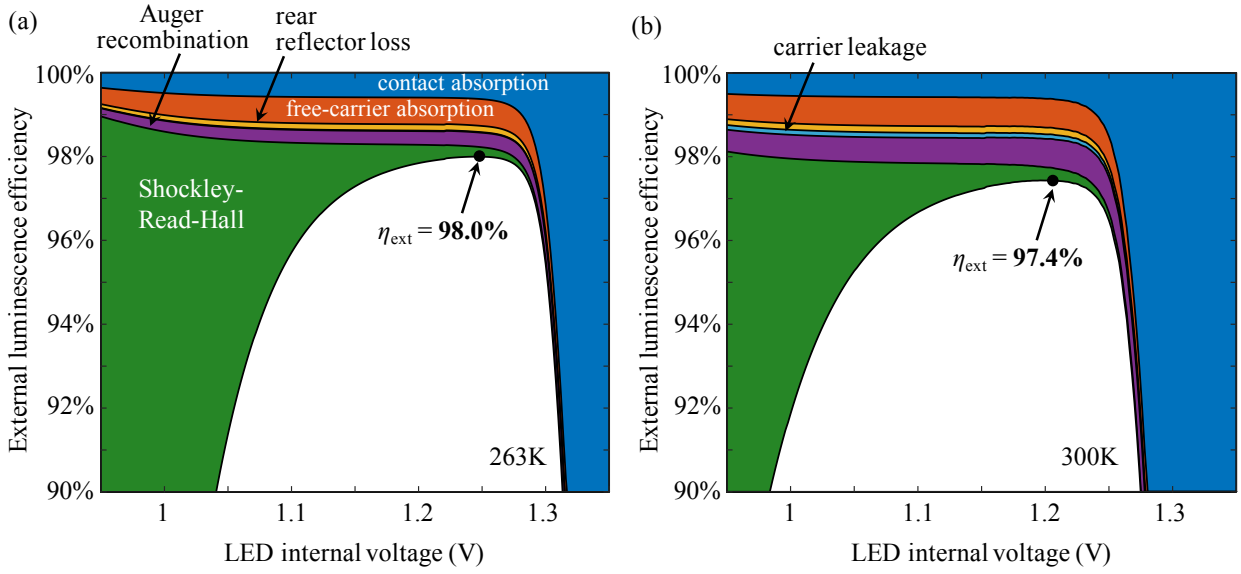

Figure 3.10: The external luminescence efficiency η_{ext} of the LED in Fig. 3.1 at (a) 263K and (b) 300K. The loss components responsible for deviations from $\eta_{\text{ext}} = 100\%$ are shown and the peak efficiencies are labeled.

3.5 Optimized LED performance

Fig. 3.10(a) and (b) show the external luminescence efficiency of the optimized LED in Fig. 3.1 at operating temperatures of 263K and 300K, respectively, as a function of the internal voltage. The various loss mechanisms contributing to non-ideal luminescence efficiency are resolved. As we saw in Section 3.2, Shockley-Read-Hall recombination dominates the losses at low voltage. In the large-bias limit, the efficiency drops sharply. This is overwhelmingly due to the loss of internal photons by contact absorption, as resistive losses act to concentrate the emission of light very close to the absorbing contacts. Peak luminescence efficiency therefore occurs at moderate bias, where the various bias-dependent loss mechanisms are similar in magnitude. At 263K, the peak efficiency is $\eta_{\text{ext}} = 98.0\%$, which occurs at $\overline{V} = 1.246\text{V}$. At 300K, the peak is $\eta_{\text{ext}} = 97.4\%$ at $\overline{V} = 1.204\text{V}$. Both of these values greatly exceed the highest LED external luminescence efficiency reported to date, which is 80% [39]! This suggests that large improvements in the state-of-the-art is within practical reach, for in designing our LED we have sought to keep the structure realistic, though not necessarily simple to implement.

Notably, we find in Fig. 3.10 that there is a broad range of biases, spanning more than 150 mV, where a high efficiency of $\eta_{\text{ext}} \geq 97\%$ can be maintained. This means that the LED remains extremely efficient over emission intensities spanning several orders of magnitude, although these intensities are smaller than those required in – for instance – many solid-state lighting applications. We will explore the practical applications of this ultra-efficient moderate-bias regime of LED operation in the next chapter.

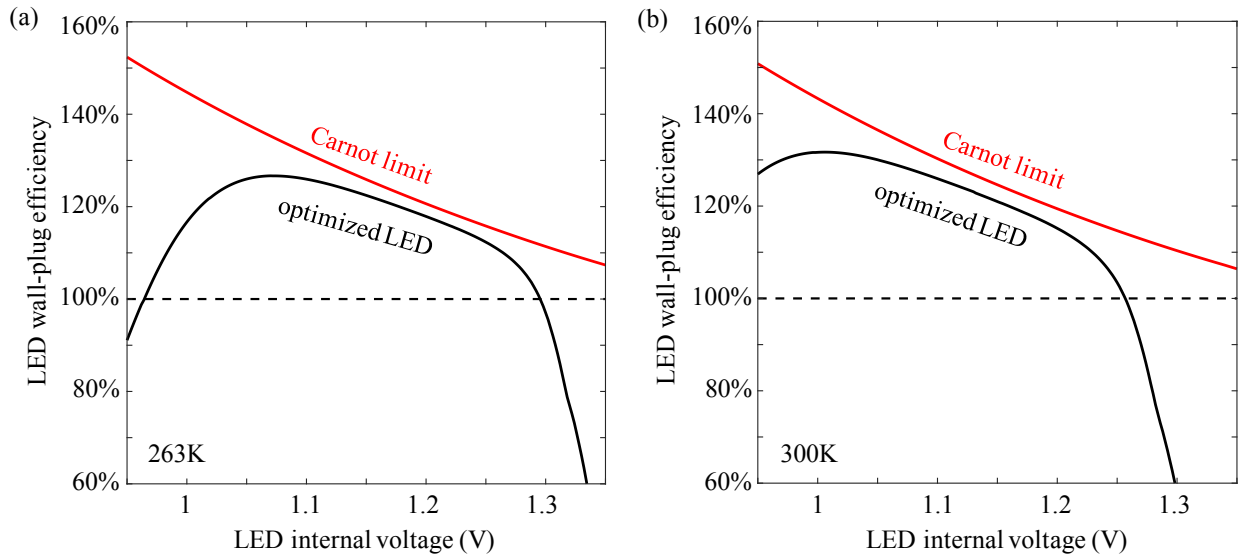

Figure 3.11: The wall-plug efficiency η_{WPE} of the LED in Fig. 3.1 at (a) 263K and (b) 300K. The dashed line corresponds to 100% wall-plug efficiency, while the bias-dependent Carnot limit is shown in red.

Fig. 3.11(a) and (b) show the wall-plug efficiency of the optimized LED at 263K and 300K, respectively, as a function of the internal voltage. As with the external luminescence efficiency, the wall-plug efficiency of this device far exceeds the best reported results in literature, which peak at 84% at 25° [39]. In particular, we note that there is a large bias range (about 300 mV) where the wall-plug efficiency actually exceeds 100%! The peak value is $\eta_{\mathrm{WPE}} = 127\%$ at 263K and 131% at 300K. By definition, this means that the optical power that is emitted by the LED is greater than the electrical power that is supplied to drive its operation. The source of the "missing" energy is the sea of phonons: thermally induced lattice vibrations in the semiconductor. Therefore, when operated at above-unity wall-plug efficiency, the LED refrigerates itself by its usual process of electroluminescent light emission. The origin of this cooling effect, and its practical uses, will be the subject of the next chapter.

The red curves in Fig. 3.11 represent the Carnot limit of LED wall-plug efficiency, which must indeed lie above unity due to the possibility of pumping power from the phonons to the photons. The value of the Carnot limit is obtained from Equation (2.2) by setting all of the constituent efficiencies (η_{int}, C_{ext}, η_{elec}) to unity, leaving only the factor $\langle E\rangle/q\overline{V}$, which exceeds unity for all biases below the photon energy. In Chapter 4, we will show more rigorously that this is the exact expression for the bias-dependent Carnot limit if we view the LED as a machine that pumps heat from a cold reservoir (lattice phonons) to a hot reservoir (luminescent photons).

Chapter 4

Electroluminescent cooling

Refrigeration is intrinsic to the operation of light-emitting diodes. This rather unexpected feature has long been recognized [6], [64]–[66], though to date it has eluded direct experimental observation. In the earliest work which hypothesized this effect, dating to 1953, the authors observed that the average photon energy $\langle E \rangle$ emitted from a green SiC light-emitting diode could be larger than the voltage qV used to bias the device [6]. As we saw in Chapter 3, this fact allows for the possibility of above-unity wall-plug efficiency: optical output power in excess of the supplied electrical power. In that original work, it was predicted that the energy shortfall is made up by lattice vibrations in the semiconductor, whose energy (and entropy) is pumped into the emitted photons; later experiments have confirmed this explanation, though without an explicit demonstration of net cooling [67]. We will begin this chapter by taking a careful look at this self-refrigeration effect from both a thermodynamic and a microscopic point of view.

If the refrigeration effect is indeed intrinsic rather than exotic, why is the experimental evidence lacking for a temperature reduction from ambient as a direct consequence of LED light emission? The problem is insufficient external luminescence efficiency η_{ext}: to attain net cooling, the losses in the device must be near zero, and a sufficiently efficient LED has yet to materialize. Furthermore, the sensitivity of cooling performance on luminescence efficiency is bandgap-dependent. We will show that GaAs is nearly ideally suited for electroluminescent cooling, as its bandgap (1.42 eV at 300K) is large enough to stave off significant Auger recombination, yet small enough to be penalized less heavily for luminescence losses in comparison to wider-bandgap materials, such as GaN. We will discuss the optimal technology choice for cooling in Section 4.2.

To probe the technological limits of electroluminescent cooling, we will make use of the ultra-efficient GaAs LED structure developed in Chapter 3, which we have already shown to be capable of above-unity wall-plug efficiency. For any refrigerator, the two primary figures-of-merit are the power density of cooling Q_{c} (in W/cm^2 of heat transferred) and the coefficient of performance (COP), which quantifies the energy efficiency of cooling:

$$\text{COP} = \frac{Q_{\text{c}}}{W} \tag{4.1}$$

where W is the work supplied per device area to drive the transfer of heat out of the LED.

With presently accessible levels of material quality and optoelectronic device processing, LED-based cooling is not only feasible but efficient enough to be competitive with thermoelectric cooling for moderate-power applications. To show this, we will adopt the thermophotonic cooling configuration, which enhances the COP by coupling the cold LED with its reciprocal device, a hot photovoltaic cell. The theoretical framework for this refrigerator, which uses luminescent photons as the working fluid, will be discussed in Section 4.3. A conceptual comparison with adjacent solid-state cooling technologies – thermoelectric cooling and solid-state laser cooling – will be provided in Section 4.4. We will present the performance of the GaAs cooling system in Section 4.5 for applications near room temperature and in Section 4.6 for cryogenic cooling applications. In cryo-cooling, the case for LEDs is more compelling, as electroluminescent cooling becomes more efficient (as a fraction of Carnot efficiency) while facing no competition from thermoelectrics, which cannot support net cooling at low temperature. In Section 4.7, we will discuss methods to potentially increase the power density of electroluminescent cooling to levels more comparable to what can be achieved with thermoelectrics. The analysis and results in Sections 4.1 to 4.6 have been included partially in Ref. 11, while the near-field cooling results in Section 4.7 are discussed in more detail in Ref. 68.

4.1 The self-refrigeration effect in light-emitting diodes

We have invoked the phenomenon of electroluminescent cooling as a way to account for the "missing energy" in LED operation and thereby satisfy the first law of thermodynamics. But what of the second law of thermodynamics? Cooling implies that the LED removes entropy from its lattice phonons, and this entropy cannot be annihilated; it must be transferred to another heat reservoir, and for net cooling of the LED to take place, this final reservoir must lie outside of the physical device.

Consider an LED that is subjected to electrical pumping, which populates the device active region with excited electrons and holes. As in equilibrium, the charge carriers continuously exchange energy with the semiconductor lattice via electron-phonon scattering. But the device, being excited, is not in equilibrium: instead, the carriers and the phonons are said to be in *quasi-equilibrium*. This condition implies that the energy distributions for the two ensembles – Fermi-Dirac for the carriers and Bose-Einstein for the phonons – can be described by the same temperature T. At the same time, the excited carriers recombine to produce luminescent photons, and are continuously re-generated by the absorption of said photons. This leads again to a condition of quasi-equilibrium, this time between the carriers and the luminescent photons. Unlike thermal radiation, luminescent radiation possesses a nonzero chemical potential μ [56]. In formal terms, quasi-equilibrium is the condition of equality between the luminescence chemical potential and the electron-hole quasi-Fermi level

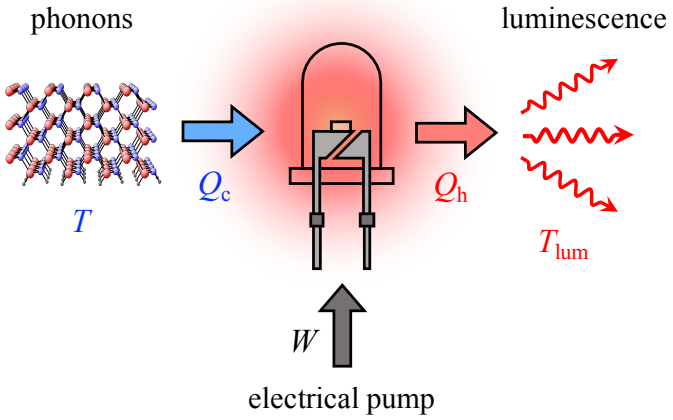

Figure 4.1: The LED can be considered as a thermodynamic refrigerator that uses electrical work to transfer heat from a low-temperature reservoir, the lattice phonons, to a high-temperature reservoir, the luminescent photons.

separation:

$$\mu = qV \tag{4.2}$$

The chemical potential, by definition, gives the free energy per photon in the luminescent radiation. It can therefore be written as the difference between the total energy E of the photon and an entropic energy component:

$$\mu = E - sT \tag{4.3}$$

Combining the above two quations, we can solve for the entropy s per photon:

$$s(E) = \frac{E - qV}{T} \tag{4.4}$$

Each luminescent photon therefore carries with it not only optical energy but also entropy, and thus heat. This entropy originates in the lattice phonons, and ultimately manifests in the spatial and temporal incoherence of LED emission. At the photon energies where the Bernard-Duraffourg transparency condition ($E = qV$) is met, the emission from the semiconductor is coherent and cooling is no longer possible.

If LED light emission transfers heat from the lattice heat reservoir (at temperature T) to the photon heat reservoir, it must be possible to assign a temperature to the luminescence and to describe the LED as a heat pump (or refrigerator) operating between the two reservoirs. This view is illustrated in Fig. 4.1. The temperature of the incoherent luminescent radiation can be found from an expression that is strongly reminiscent of the definition of thermodynamic temperature:

$$T_{\text{lum}} = \frac{dU/dt}{dS/dt} \tag{4.5}$$

In the literature, this has also been called a flux temperature [65]. The total power flux dU/dt and the total entropy flux dS/dt carried away from the device by the external luminescence can be found by integrating the energy E and the entropy $s(E)$ over the full bandwidth and angular spread of the emission:

$$Q_{\mathrm{h}} = \frac{dU}{dt} = \int_0^\infty dE \int_0^{\pi/2} E \cdot \frac{2E^2}{c^2h^3} \frac{a(E,\theta)}{e^{(E-qV)/kT}-1} \, 2\pi \sin\theta \, \cos\theta \, d\theta \tag{4.6a}$$

$$\frac{dS}{dt} = \int_0^\infty dE \int_0^{\pi/2} s(E) \cdot \frac{2E^2}{c^2h^3} \frac{a(E,\theta)}{e^{(E-qV)/kT}-1} \, 2\pi \sin\theta \, \cos\theta \, d\theta \tag{4.6b}$$

Using the definition of the average photon energy $\langle E \rangle$, we can evaluate the temperature of the luminescence as:

$$T_{\mathrm{lum}} = \frac{T}{1 - qV/\langle E \rangle} \tag{4.7}$$

As we expect, the larger the bias applied to the LED, the brighter its light emission and the greater its luminescence temperature. An LED at 300K and forward biased at 80% of its bandgap emits luminescence with a temperature of 1500K. When biased at 95% of the bandgap, which is typical for lighting applications, the luminescence temperature is 6000K.[1]

The heat pump picture of the LED in Fig. 4.1 allows us to derive the Carnot limit of wall-plug efficiency, which we first introduced at the end of Chapter 3. In the absence of all irreversible losses in the process of electroluminescent light emission, the wall-plug efficiency of the LED is the efficiency of a Carnot heat pump between the two reservoir temperatures T and T_{lum}:

$$\eta_{\mathrm{WPE,Carnot}} = \left(\frac{Q_{\mathrm{h}}}{W} \right)_{\mathrm{Carnot}} = \frac{T_{\mathrm{lum}}}{T_{\mathrm{lum}} - T} = \frac{\langle E \rangle}{qV} \tag{4.8}$$

This limiting value for η_{WPE} is exactly the same as what we obtain from Equation (3.1) by setting all of the device efficiencies equal to unity! This should not be surprising, since setting these efficiencies to unity is equivalent to eliminating all of the irreversible losses inside the LED. The refrigeration coefficient of performance of the LED can be easily found from Equation (4.1) by using the relation $Q_{\mathrm{c}} + W = Q_{\mathrm{h}}$, with a Carnot limit likewise given by the temperatures T and T_{lum}.

Let us return to the microscopic discussion of the LED internal dynamics to determine precisely how and where lattice cooling takes place within the device. Consider Fig. 4.2, which depicts the band diagram of the device in Fig. 3.1 at a small bias of 1.12V. In this device, the injected holes encounter a potential barrier as they diffuse from the p^+-GaInP

[1] Under *reverse* bias ($qV < 0$), the luminescence is cooler than the lattice, and the LED can refrigerate its surroundings by the net absorption of thermal radiation rather than by the emission of luminescence. We will not explore reverse-biased cooling in this work, as its available power density is very limited, but the effect has been discussed theoretically [66] and demonstrated experimentally [69].

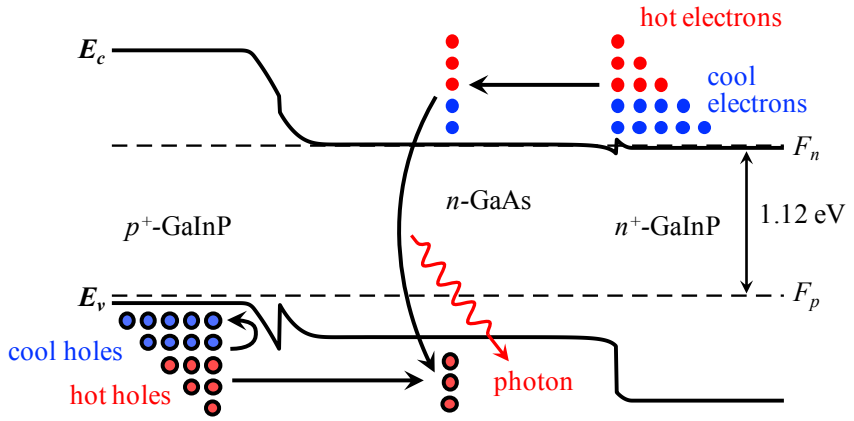

Figure 4.2: The band diagram of the LED in Fig. 3.1 is shown at 263K and a bias of $qV = 1.12$ eV. Cooling mainly takes place near the p-GaInP/n-GaAs interface, where hot holes are selectively transmitted over the potential barriers. There is consequently a net absorption of lattice heat by the carriers in these regions by electron-phonon scattering. The hot carriers subsequently recombine and transfer their heat to a luminescent photon.

cladding layer into the GaAs active region. The potential barrier ensures that only the hot holes comprising the high-energy tail of the Fermi-Dirac distribution can be transmitted into the active region; the remaining holes on the p^+-GaInP side do not have sufficient energy. As the region near the potential barrier becomes selectively depleted of hot holes, restoration of the quasi-equilibrium carrier distribution requires that hot holes be selectively generated in that same region. This latter process involves a net absorption of lattice heat by the carriers via electron-phonon scattering, leading to refrigeration that is localized near the potential barrier, as suggested in other works [67], [70], [71].[2] Some cooling occurs also in the injection of electrons from n-GaInP, but this is limited by the small size of the potential barrier at that junction. There is additionally a net exchange of heat between the carriers and the lattice at the device terminals, owing to the presence of semiconductor-metal potential barriers (not shown).

The above heat transfer process from phonons to carriers is known as the Peltier effect, which is the same phenomenon that is responsible for lattice cooling in thermoelectric devices [70]. In LEDs, this step is followed by the radiative recombination of the energetic carriers, which transfers heat from the carriers to the luminescent photons. These two processes are quite distinct in timescale. Peltier cooling, associated with carrier relaxation to quasi-equilibrium via electron-phonon scattering, occurs on a sub-picosecond timescale. Radiative

[2]Once the holes surpass the first potential barrier, they might fall into the triangular potential well at the edge of the n-GaAs layer, giving up energy and thus causing localized heating. To escape the potential well and ultimately recombine with electrons in the middle of the active region, the holes must overcome a second potential barrier (in n-GaAs), again leading to localized cooling. The net result from these processes is a cooling effect, as the holes in the active region are significantly higher in energy on average than the holes in the p-GaInP cladding.

recombination, on the other hand, typically occurs within a few nanoseconds. When the luminescent photons subsequently escape, they remove heat from the device. If extracted into a weakly scattering medium, such as air or vacuum, they can transport the heat ballistically from one structure to another. This contrasts fundamentally from thermoelectric devices, where the heat transport is everywhere diffusive between the cold and hot reservoirs. We will return to this important distinction between the two technologies in Section 4.4.

From conservation of energy as well as from Equation (4.4), we know that the average amount of heat extracted by an externally emitted photon is equal to $\langle E \rangle - qV$. When the applied bias is significantly smaller than the semiconductor bandgap, each photon has extracted many kT's of thermal energy from the lattice! Statistical mechanics tells us that such a process should be very unlikely, and indeed that is the case. However, this should hardly be considered an exotic effect. Recall that the carriers which participate in Peltier cooling reside in the high-energy tails of the electron and hole Fermi-Dirac distributions. The smaller the applied bias, the larger the internal potential barriers, and we must move towards ever higher energies in these distributions to find the carriers that participate in the cooling effect. The population of these carriers decreases exponentially at low values of qV according to the Fermi-Dirac distribution, and this small population of carriers is also involved in current conduction and light emission from the diode. Therefore, the low probability of the cooling process simply manifests in the exponential voltage dependence of the LED current and emitted photon flux. Framed in this way, it is hardly surprising!

In view of the fact that no demonstration of net cooling yet exists, is there any empirical evidence to suggest that the aforementioned heat pumping effects actually occur? Over the decades, alternative explanations have been offered for the missing energy in LED light emission, such as interfacial Auger processes [72] and two-electron impact ionization [73]. It was not until 2015 that the presence of Peltier cooling effects in LEDs was experimentally confirmed at practical bias levels [67]. In this work, which used a GaN-based LED, the authors demonstrated that while the external luminescence efficiency η_{ext} decreases at elevated temperatures, the wall-plug efficiency can remain high if the device moves to a lower applied bias qV. In effect, the larger amount of heat extracted from the lattice phonons compensates for the additional heat generation caused by a lower luminescence efficiency. In our GaAs device, this is the same reason that the wall-plug efficiency in Fig. 3.11 peaks at a lower voltage than the external luminescence efficiency in Fig. 3.10.

4.2 The luminescence efficiency requirement for refrigeration

In spite of the fact that electroluminescent cooling is an intrinsic effect in LEDs, it has never been directly observed. The reason is that to date, the external luminescence efficiency of even the most efficient devices is too small to allow for net cooling. In Fig. 4.3, we compare the cooling that results from electroluminescence to the heating caused by internal

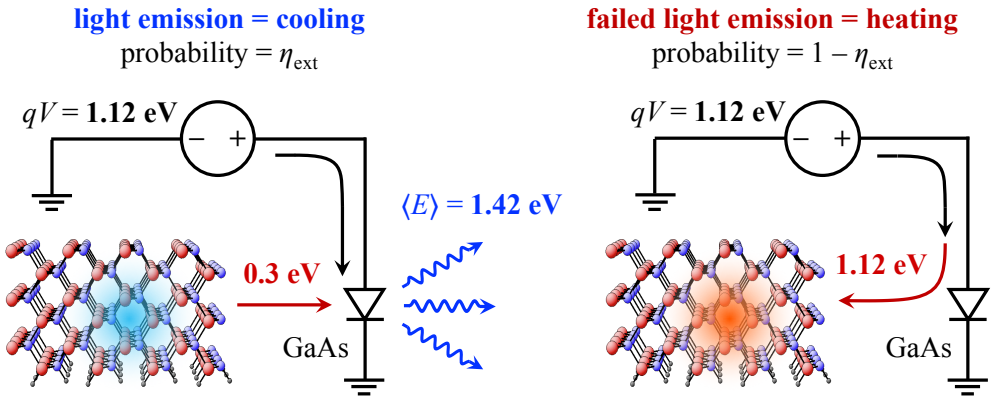

Figure 4.3: Each light emission event in an LED is accompanied by microscopic cooling of the lattice (left), while each failed conversion from an electron to an outgoing photon leads to microscopic heating of the lattice (right). The minimum luminescence efficiency required to produce net cooling is found by balancing the rates of heat transfer associated with these two outcomes.

losses. Every external luminescence event, occurring with a probability equal to η_{ext}, liberates $\langle E \rangle - qV$ of lattice heat from the device, where $\langle E \rangle$ can be approximated by the bandgap. On the other hand, every time an injected carrier fails to leave the LED as a photon as a result of non-radiative recombination or parasitic optical absorption, the full energy qV of the electron-hole pair is dissipated in the lattice: this occurs with a probability equal to $(1 - \eta_{ext})$. Comparing the cooling and heating rates associated with these processes, we can easily show that net cooling is possible only when:

$$\eta_{ext} > \frac{qV}{\langle E \rangle} \tag{4.9}$$

If we further account for Joule heating and other losses, the threshold efficiency becomes larger than the value predicted by Equation (4.9). Note that we could have reached the same conclusion from Equation (3.1) using the condition $\eta_{WPE} > 100\%$.

How high is this minimum efficiency? The equation above suggests that if the applied bias is considerably smaller than the photon energy, net cooling is possible even with a small value of η_{ext}. In fact, the requisite external luminescence efficiency has been achieved in experiment for ultra-low values of the applied bias ($qV \ll kT$) in GaInAsSb/GaSb LEDs [74] and mid-infrared InAs and InAsSb LEDs [75]. While these devices reached a wall-plug efficiency above unity, they operated at a power density that was too small to allow the observation of a net refrigeration effect.

To admit operation as a practical solid-state refrigerator, high external luminescence efficiency must be accessible at a large enough voltage bias qV to produce a practically useful light (or heat) output. In GaAs, a cooling heat flux of 1 mW/cm^2 – which is just large enough to be potentially useful – corresponds to a bias of about 300 mV below the bandgap at 300K. This leads to a minimum required efficiency of about $\eta_{ext} = 80\%$ according to Equation

(4.9), which exceeds the highest efficiencies reported in GaAs [38]. In other optoelectronic materials, the same value of heat flux will require – to a first-order approximation – the same deviation of the applied bias qV from the bandgap E_g.[3] Thus, for a heat flux of 1 mW/cm^2, a wider-bandgap material would require a larger minimum efficiency. In InGaN ($E_g = 2.99$ eV), for example, the minimum efficiency is greater than $\eta_{\text{ext}} = 90\%$, which is well in excess of the highest reported luminescence efficiency of $\eta_{\text{ext}} = 80\%$ in this material system [39]. Therefore, smaller bandgaps are preferred for electroluminescent cooling, as they demand a less stringent requirement on the luminescence efficiency to reach the same cooling performance.

In the small-bandgap limit, however, recombination in semiconductors tends to be dominated by Auger recombination, which rapidly degrades the internal luminescence efficiency and makes cooling infeasible. This problem is well-known in infrared photodetectors, where the leakage current has been observed to rise exponentially with decreasing bandgap and increasing temperature [76]. Thus, to stave off Auger recombination, the semiconductor bandgap for electroluminescent cooling should not be much smaller than about 1.0 eV. GaAs, with its moderately large bandgap and hence a small Auger recombination coefficient, is ideally suited for this application.

As we have seen in Chapter 3, GaAs is a mature optoelectronic technology that is capable of extremely efficient electroluminescence. Considering that luminescence efficiencies above 80% – better than the best LEDs to date – are needed merely to reach net zero cooling at practical fluxes, we should expect that electroluminescent cooling with a competitive coefficient of performance (i.e. a significant fraction of the Carnot limit) would demand η_{ext} very close to 100%. Will the ultra-high external luminescence efficiency that we found for our GaAs device in Chapter 3 – up to $\eta_{\text{ext}} = 98.0\%$ at 263K – be sufficient to make this a viable solid-state cooling technology? We will answer this question in the following sections. In the next section, we will first introduce a useful cooling scheme that can substantially increase the coefficient of performance for the same LED luminescence efficiency.

4.3 Thermophotonic cooling

The refrigeration effect that accompanies LED light emission can be leveraged to cool a load that is in thermal contact with the device active region. We can further bolster the efficiency of refrigeration by making use of the thermophotonic configuration, which couples the cold LED (at temperature T_{c}) to a hot photovoltaic cell (at temperature $T_{\text{h}} > T_{\text{c}}$), as shown in Fig. 4.4 [77], [78]. The photovoltaic (PV) cell, in absorbing the luminescence from the LED, captures the heat that was carried by the photons; this heat is deposited in the lattice and

[3] This can be seen as follows. The heat flux carried by the luminescence is $Q_{\text{c}} = (\langle E \rangle - qV)\,\Phi_{\text{ext}}$. If we integrate Equation (3.17) for the emitted photon flux assuming a step-function absorptivity, we will find an approximate bandgap dependence given by: $Q_{\text{c}} \sim (E_g - qV) E_g^2 \exp\big(-(E_g - qV)/kT\big)$. Therefore, for the same value of $E_g - qV$, Q_{c} is a function of bandgap only through the term E_g^2, whose variation is within one order of magnitude for bandgaps between 1 eV and 3 eV.

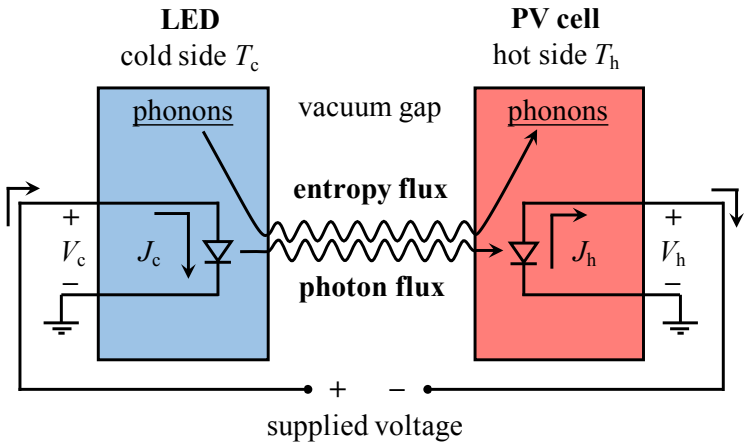

Figure 4.4: Schematic of the electroluminescent refrigerator combining an LED with a photovoltaic (PV) cell. Photons traveling between the two devices carry both energy and entropy; the directions of net photon and entropy fluxes are shown. The electrical power generated in the PV cell is returned to the LED to reduce the electrical work supplied externally. The single LED and PV cell represent in reality a series-parallel connection of multiple devices that matches the total current of the LED ensemble to that of the PV cells.

can subsequently be rejected to a heat sink. Additionally, the PV cell converts part of the absorbed optical energy back into electricity, which can be returned to the LED to partially offset its power consumption. The heat transfer from cold to hot can thus be driven by a much smaller supply of external power in comparison to an LED alone. The electrical power recovery enabled by the PV cell leads to a significantly larger cooling coefficient of performance (COP).

Since the system employs light as the working fluid, the cold and hot sides can be separated by a vacuum spacer, which eliminates the leakage of heat from hot to cold by direct thermal conduction or convection. This constitutes the chief advantage of this optoelectronic refrigerator over the thermoelectric cooler, where the thermoelectric material is itself the main conduit for heat leakage from hot to cold. Nonetheless, some parasitic heat leakage paths remain, as shown in Fig. 4.5. Thermal radiation transfers heat across the vacuum gap, and the electrical feedback connection that allows for power recovery is also an unwanted pathway for heat conduction. Fortunately, these leakage paths can be made negligibly small without compromising the power density or energy efficiency of the electroluminescent cooling effect. In Appendix C, we discuss the system-level design strategies for the suppression of heat leakage.

For most of this chapter, excepting Section 4.7, we will assume that the vacuum gap is much larger than a thermal wavelength ($\sim$10 μm at room temperature) so that the two devices reside in their respective far fields for both luminescent and thermal radiation. When this gap is reduced to sub-wavelength scales, the rate of radiative heat transfer increases

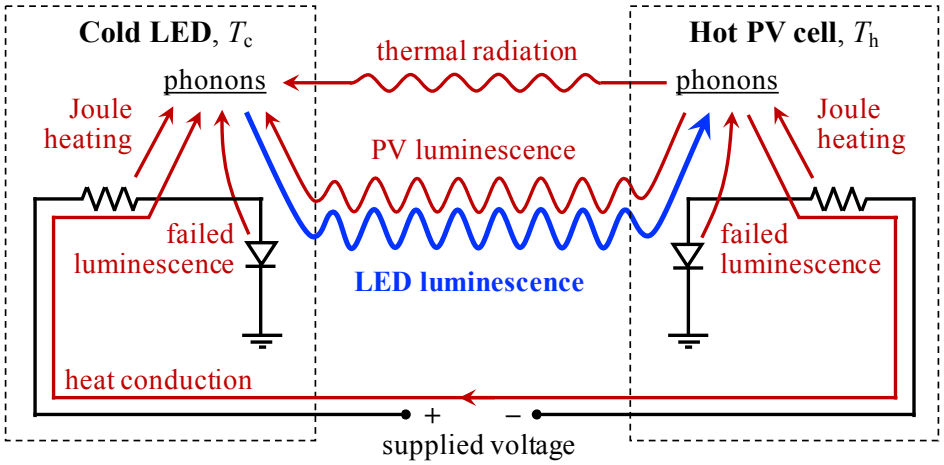

Figure 4.5: Heat flow paths in the thermophotonic system. Various mechanisms compete with the electroluminescent heat flux (blue), which is responsible for cooling. "Failed luminescence" refers to energy loss by non-radiative recombination and by the parasitic absorption of luminescence.

dramatically [79]. This effect can be exploited to enhance luminescent heat transfer, but this is preceded by a strong enhancement in the thermal radiative leakage at longer wavelengths. In the near field, this leakage manifests as the parasitic coupling of optical phonon-polaritons across the vacuum gap [68]. We will discuss the near-field cooling scheme in Section 4.7.

The paths for heat flow within the thermophotonic system, accounting for all practical non-idealities, are illustrated in Fig. 4.5. The net heat flow per area away from the cooled LED can be expressed as:

$$Q_c = \langle E \rangle_c \Phi_c - J_c V_c - Q_{\Omega c} - \langle E \rangle_h \Phi_h - Q_{leak} \tag{4.10}$$

where Φ is the external luminescent photon flux, J is the current density, V is the quasi-Fermi level splitting voltage in the device active region (i.e. the internal voltage in Chapter 3), Q_Ω is the Ohmic power dissipation, and Q_{leak} is the parasitic non-luminescent heat leakage. The subscripts "c" and "h" denote the cold (LED) and hot (PV) sides of the system, respectively. The first three terms together comprise the difference between the luminescent power output and the electrical power input to the LED, while the fourth and fifth terms correspond to the luminescent and non-luminescent heat fluxes from hot to cold, respectively. All power densities, photon fluxes, and current densities are expressed per area of the surface that faces the vacuum gap, which we will call the "front surface" of the device. We assume that the LED and PV have dimensions much larger than the vacuum gap, so that we can consider the geometry a closed system for the photons: any photon emitted by one device is either absorbed in the other device or is reflected and absorbed in the same device.

The net electrical work per area consumed to drive the electroluminescent heat transfer is given by

$$W = J_c V_c + Q_{\Omega c} - (J_h V_h - Q_{\Omega h}) \tag{4.11}$$

where V_h is the PV internal voltage. The coefficient of performance is then given by Equation (4.1). Since the thermophotonic refrigerator operates between the two temperatures T_c and T_h, the COP is bounded by the corresponding Carnot limit:

$$\mathrm{COP}_\mathrm{Carnot} = \frac{T_\mathrm{c}}{T_\mathrm{h} - T_\mathrm{c}} \tag{4.12}$$

The efficiency advantage of thermophotonic cooling over cooling using an LED alone can be viewed in one of two equivalent ways. The first and more apparent viewpoint is that the PV cell recovers energy to the LED, substantially reducing the amount of work needed to drive the same electroluminescent heat flux and thus increasing the absolute COP. The second viewpoint is to compare the Carnot COP above to the Carnot COP of the LED alone, which is given by $T_\mathrm{c}/\left(T_\mathrm{lum} - T_\mathrm{c}\right)$. Recall from Section 4.1 that the LED luminescence temperature T_lum is generally very large even when the LED is only moderately biased: when biased to a mere 80% of the bandgap at 300K, we have $T_\mathrm{lum} = 1500\mathrm{K}$. Therefore, the temperature difference $(T_\mathrm{lum} - T_\mathrm{c})$ also tends to be large, leading to a small Carnot COP. The thermophotonic system replaces this very hot heat reservoir with a much cooler reservoir – the PV cell – whose temperature is independent of the amount of heat transferred. The temperature difference $(T_\mathrm{h} - T_\mathrm{c})$ is much smaller than $(T_\mathrm{lum} - T_\mathrm{c})$ for all practical applications, and thus the Carnot limit itself is significantly higher.

The electrical current densities in the two devices are given by,

$$J_\mathrm{c} = q\left(\frac{\Phi_\mathrm{c}}{\eta_\mathrm{ext,c}} - \eta_\mathrm{abs,c}\Phi_\mathrm{h}\right) \tag{4.13a}$$

$$J_\mathrm{h} = -q\left(\frac{\Phi_\mathrm{h}}{\eta_\mathrm{ext,h}} - \eta_\mathrm{abs,h}\Phi_\mathrm{c}\right) \tag{4.13b}$$

where we define the PV current polarity such that $J_\mathrm{h} > 0$ if the cell operates in the power-producing quadrant, and a positive voltage denotes a forward bias in both devices. The first term in each equation is the forward current, which is converted to external luminescence with a quantum efficiency η_ext. The second term is the reverse current generated by the photovoltaic conversion of incident luminescence from the opposing device, with a quantum efficiency η_abs. This latter efficiency is analogous to the conventional definition of the external quantum efficiency of a solar cell, integrated over the illumination spectrum.

In order to generate electrical power, the PV cell operates in forward bias and must therefore emit its own luminescence, which carries heat from hot to cold. For a given LED voltage, which sets the cooling flux, the PV voltage is a free variable that we can adjust to maximize the COP: V_h should be large to increase the power returned to the LED, but small enough to ensure adequate suppression of the hot-to-cold luminescence flux: $\Phi_\mathrm{h} \ll \Phi_\mathrm{c}$. Conveniently, we can achieve the latter simply by choosing the same active material for the cold emitter and hot absorber. In III-V semiconductors, the bandgap narrows with temperature: for example, a GaAs LED at $T_\mathrm{c} = 263\mathrm{K}$ has a bandgap of 1.44 eV, while a GaAs PV cell at $T_\mathrm{h} = 313\mathrm{K}$ has a bandgap of 1.42 eV. This small bandgap offset is sufficient

to ensure that the PV cell is a near-complete absorber of the LED luminescence, while the LED is a weak absorber of the PV luminescence. Much of the PV luminescence is instead reflected by the rear mirror of the LED and returned to the PV cell, where it is re-absorbed. This asymmetry strongly enhances the cold-to-hot luminescent flux Φ_c relative to the hot-to-cold luminescence flux Φ_h. As a consequence, a larger value of the PV voltage V_h can be used that still keeps the reverse luminescence flux small, allowing for a higher COP.

As we reviewed in Chapter 2, an efficient LED is also an efficient PV cell. Therefore, we will adopt the device structure in Fig. 3.1 for both sides of the thermophotonic system with only slight modifications, to be described in Section 4.5. These modifications will turn out to reduce the external luminescence efficiency of the PV cell, but we note that the cooling performance is far less sensitive to $\eta_{ext,h}$ (PV) than to $\eta_{ext,c}$ (LED). The PV voltage V_h increases logarithmically with $\eta_{ext,h}$ as given by Equation (2.8), but since PV electricity generation is contingent upon successful light emission from the cold side, designing for high LED luminescence efficiency is of far greater importance. The presence of the PV cell does not change the requirement on $\eta_{ext,c}$ given by Equation (4.9).

Within the closed system of Fig. 4.4, the emissivity of each device must be modified to account for incomplete optical absorption in the opposing device. Since both devices are textured, the overall emissivity can be found by summing over a sequence of incoherent reflections between the LED and PV cell:

$$a_c(E) = a_{0,c}(E) \times \frac{a^\dagger_{0,h}(E)}{1 - \left(1 - a^\dagger_{0,c}(E)\right)\left(1 - a^\dagger_{0,h}(E)\right)} \tag{4.14}$$

where a_c is the corrected LED emissivity, used to calculate Φ_c via Equation (3.17), and $a_{0,c}$ is the unmodified emissivity given by Equation (3.19). The terms $a^\dagger_{0,c}$ and $a^\dagger_{0,h}$ represent the total absorptivity – including both electronic and parasitic absorption – of the LED and PV cell, respectively. This quantity is found by adding the loss term $\overline{T}\mathcal{L}$ to the numerator of Equation (3.19). The emissivity correction has a very small effect on the LED external luminescence efficiency values shown in Section 3.5. To calculate the PV absorptivity a_h, we can use the above equation with the subscripts "c" and "h" interchanged.

The quantum efficiency of reverse current generation in the PV cell is the fraction of the absorbed photons that produce electrons and holes, given by:

$$\eta_{abs,h} = \frac{1}{\Phi_c}\int_0^\infty \left(\frac{\alpha(E)d}{\alpha(E)d + \mathcal{L}(E)}\right)_h \frac{d\Phi_c}{dE}\,dE \tag{4.15}$$

where we have assumed that all photo-generated carriers are collected at the electrodes, and $d\Phi_c/dE$ denotes the spectral density of the LED luminescence, which is given by the energy integrand of Equation (3.17) after integrating over the angle θ. If the LED luminescence lies almost fully above the PV bandgap, as is the case with a 50K temperature difference, $\eta_{abs,h}$ is well over 99%. We can find $\eta_{abs,c}$ by interchanging the subscripts "c" and "h".

We will assume in our analysis that each device sees the other only through a single planar front surface, with the rear surface taken up by a highly reflective mirror. We can conceivably

use both the front and rear surface of the LED for light emission, with arrangements of PV cells on both sides of the LED to collect the emitted luminescence. This scheme increases the cooling power density by a factor of two and, more importantly, increases the LED external luminescence efficiency by doubling the light extraction rate relative to the internal optical loss rates. However, in doing so we would lose the ability to conduct heat from the load to the LED active region through the metallic backplane of the device; the heat would instead have to be conducted laterally over long distances through an optically transparent layer. While transparent, thermally conductive coatings exist, a lateral heat spreading scheme would impose further restrictions on the available cooling capacity if large temperature drops within the system are to be avoided. Therefore, we will not consider double-sided emission schemes in this work, though we do not dismiss this strategy as a potentially feasible (though more practically challenging) path to higher COP.

Having established a theoretical framework for thermophotonic cooling, we can now address the all-important question that was raised in the previous section: how sensitive is the cooling performance to the luminescence efficiency of the LED? For a first-order sensitivity analysis, we make the following assumptions: only luminescent losses are present (i.e. $Q_{\Omega c} = Q_{\Omega h} = Q_{leak} = 0$), the absorption quantum efficiency η_{abs} is unity, and the absorptivity of each device is approximated as a step function at the bandgap energy. The result of this calculation is shown in Fig. 4.6(a) for three bandgap values: $E_g = 0.93$ eV, 1.42 eV, and 2.99 eV.[4] The curves are calculated for temperatures of $T_c = 263$K and $T_h = 313$K and a fixed cooling heat flux of $Q_c = 1.0$ mW/cm^2. Along each curve, the biases V_c and V_h are chosen to achieve this power density and to maximize the COP, respectively.

We observe in Fig. 4.6(a) that net cooling is only possible above a minimum value of the external luminescence efficiency that is bandgap-dependent.[5] This agrees with our conclusion in Section 4.2 regarding technology selection: materials with a wider bandgap have a higher threshold efficiency for net cooling. Secondly, we note that above the threshold efficiency, the largest gains in COP are made at values of η_{ext} that are very close to unity, and this is especially true for wider-bandgap materials. This highlights a unique and very important property of electroluminescent cooling: the cooling performance benefits strongly from asymptotic approaches to optoelectronic perfection, represented by the limit $\eta_{ext} = 1$. Thus, efficient electroluminescent cooling is truly an emergent property of devices that operate very close to their theoretical limits. Fig. 4.6(b) shows the same sensitivity for different values of the cooling heat flux Q_c with a fixed bandgap of 1.42 eV. Consistent with Equation (4.9), a higher cooling flux requires a larger voltage bias and therefore a higher minimum value for the external luminescence efficiency.

In Section 4.5, we will more rigorously apply the theoretical framework introduced in this section to the LED device structure that we studied in Chapter 3. For now, we make one final observation about the first-order analysis in Fig. 4.6: in the limit of unity external

[4]Near 300K, $E_g = 0.93$ eV corresponds to the quaternary alloy $In_{0.66}Ga_{0.34}As_{0.65}P_{0.35}$, lattice-matched to InP. $E_g = 1.42$ eV corresponds to GaAs. $E_g = 2.99$ eV corresponds to InGaN at the same emission wavelength as the LED with the highest reported luminescence efficiency [39].

[5]The curves do not touch COP = 0 because we have imposed a nonzero value of the cooling flux.

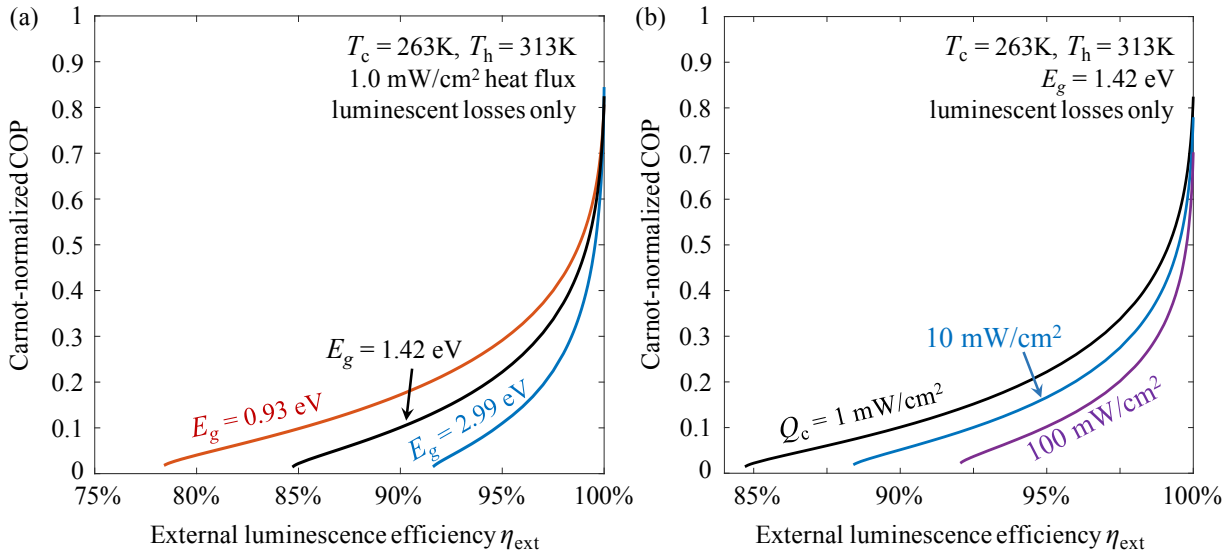

Figure 4.6: The cooling coefficient of performance (COP) is strongly sensitive to the external luminescence efficiency of the LED. (a) shows the sensitivity for three bandgap values, evaluated at a fixed cooling flux of $Q_c = 1.0$ mW/cm^2. (b) shows the sensitivity for three values of the heat flux using a bandgap of $E_g = 1.42$ eV. The COP is normalized to $\text{COP}_{\text{Carnot}} = 5.6$ for the temperatures $T_c = 263$K and $T_h = 313$K, and we have assumed $\eta_{\text{ext,c}} = \eta_{\text{ext,h}}$. We have set the heat leakage and Ohmic losses to zero.

luminescence efficiency, zero Ohmic losses, and zero heat leakage, the COP still fails to reach the Carnot limit. This is in contrast to the case of the LED alone, where we showed that $\eta_{\text{ext}} = 1$ and $\eta_{\text{elec}} = 1$ are sufficient conditions to attain Carnot efficiency. The remaining irreversibility is introduced by the presence of the photovoltaic cell. Upon absorption in the PV cell, the LED luminescence is converted to an electron-hole gas which quickly thermalizes to a distribution which has a different temperature and chemical potential (quasi-Fermi level splitting). Such a process always generates entropy via hot carrier thermalization [80]. In Appendix D, we will outline the theoretical requirements that must be met by the thermophotonic system in order to reach the Carnot limit. Actually attaining the full Carnot COP is not a practically important objective, since this is possible only in the quasi-static limit of zero heat flux.

We note that it is possible to operate the thermophotonic system as a heat engine that produces electricity. In this scheme, a room-temperature PV cell generates electricity from both the thermal radiation and the luminescent radiation that is emitted by an LED that is heated to a high temperature [77]. However, to realize the benefits of this heat engine over thermophotovoltaics – which requires no work input to the radiation source – very high LED luminescence efficiency would be needed, which is not practical at very high temperatures.[6]

[6] An optically pumped high-temperature emitter, rather than an LED, is another possibility [81].

4.4 Comparison with adjacent cooling technologies

In this section, we will compare the electroluminescent refrigerator – enhanced by the thermophotonic cooling scheme – with two adjacent solid-state refrigeration technologies: thermoelectric coolers and solid-state laser cooling. The relevant metrics of comparison will be the two primary cooling figures-of-merit: the cooling heat flux Q_c and the coefficient of performance. Our discussion in this section will be primarily conceptual: in the next section, we will provide a quantitative comparison of electroluminescent cooling performance with that of thermoelectric devices.

The most widely used solid-state technology for refrigeration is the thermoelectric cooler, which uses charge carriers as its working fluid as illustrated in Fig. 4.7. The carriers are transported from one heat reservoir to the other across a thermoelectric material, which is typically an alloyed semiconductor like Bi_2Te_3. Like electroluminescent cooling, thermoelectric cooling exploits the Peltier effect. Net cooling occurs at one of the semiconductor-metal junctions of the device, where the carriers must diffuse over a potential barrier, leading to a net absorption of heat from the lattice as discussed in Section 4.1. The charge carriers then carry the entropy through the thermoelectric material, and ultimately deposit the heat at the opposite contact of the device by means of thermal relaxation to the band edges. A typical thermoelectric cooler uses both n- and p-doped semiconductors in an arrangement that conducts electricity in series but heat in parallel.

The magnitude of Peltier cooling in thermoelectrics is governed by the Seebeck coefficient S, which can fundamentally be interpreted as the amount of entropy transported per charged particle [82]. This entropy resides in the carriers' internal degrees of freedom and is closely related to their kinetic energy, which is given by the difference between their average energy and their chemical potential; the latter is set by the position of the Fermi level

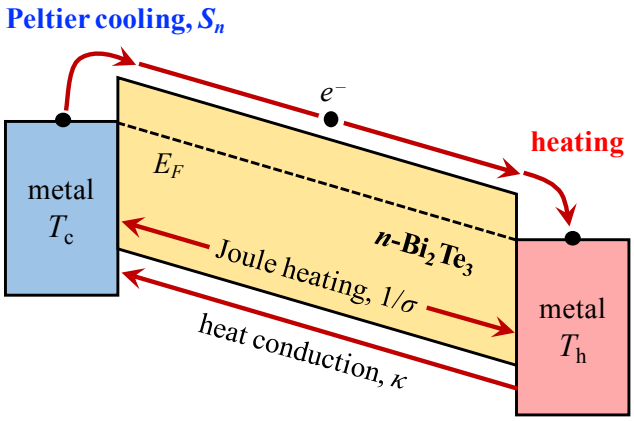

Figure 4.7: An n-type thermoelectric cooler. An electric current transports heat from cold to hot via the Peltier effect, but this cooling mechanism is intrinsically linked to the undesired Joule heating and heat conduction in the same thermoelectric material.

E_F, which can be controlled by doping.[7] The larger the separation between the Fermi level and the band edge, the greater the Seebeck coefficient. However, increasing this separation reduces the carrier concentration, making the material more resistive and thus increasing Joule heating. Additionally, since the thermoelectric forms a solid bridge between the two reservoirs, it conducts heat from hot to cold via both phonons and charge carriers. The thermoelectric cooling performance is a function of a dimensionless material figure-of-merit that encapsulates the relative strengths of these effects:

$$ZT = \frac{S^2\sigma}{\kappa} \times \frac{1}{2}\left(T_\mathrm{c} + T_\mathrm{h}\right) \tag{4.16}$$

where σ is the electrical conductivity and κ is the thermal conductivity of the thermoelectric material. For high ZT, the ultimate material is a phonon glass/electron crystal (i.e. low κ, high σ) that also supports a large Seebeck coefficient. Realizing these properties is a fundamental materials challenge [84], and many approaches have been explored [85]. Currently, the best commercial thermoelectric devices can achieve $ZT = 1$ at room temperature [86].

On the other hand, in thermophotonic electroluminescent cooling, the relative strengths of the desired cooling mechanism and the undesired heating mechanisms are largely decoupled. Fundamentally, this is because the transport of the working fluid is ballistic in the case of thermophotonics and diffusive in the case of thermoelectrics. Diffusive electronic transport forces the charge carriers to be in quasi-equilibrium with the phonons everywhere in the device. Consequently, heat is dissipated by electron-phonon scattering at every position between the two reservoirs, ultimately leading to intrinsic material tradeoffs in the quantity $S^2\sigma$. In the ballistic transport of photons, the absence of scattering processes in the vacuum gap implies that such a trade-off can be completely avoided. Furthermore, because no solid medium is needed to transport luminescent radiation over any distance, the problem of thermal conduction can be avoided as well. Thermal radiation across the gap remains a leakage mechanism, but because this leakage occurs at much longer photon wavelengths, it can be separately controlled without adversely affecting the desired luminescent mode of heat transfer (see Appendix C). This is to be contrasted with the intrinsic material trade-offs involved in the design of thermoelectrics.

The absence of fundamental trade-offs between the desired and parasitic heat fluxes in thermophotonics allows, in principle, for electroluminescent cooling to achieve higher values of the COP than thermoelectric coolers, particularly at low heat fluxes. A further advantage lies in the regime of low temperatures, where fundamental considerations cause the figure-of-merit ZT to rapidly decline [83]. At the same time, the condition of net cooling for thermoelectrics requires higher values of ZT to refrigerate across a large temperature difference, e.g. between cryogenic temperature and room temperature. Meanwhile, as we will show in Section 4.6, LEDs suffer no such performance penalties down to well below 100K; in fact, the device actually becomes more efficient. Thus, LEDs are well suited for low-temperature applications where thermoelectrics fail.

[7]Here, the average energy is taken with respect to a transport distribution, rather than the simple Fermi-Dirac distribution which holds in the absence of a temperature gradient. See Ref. 83, Section 2.3.

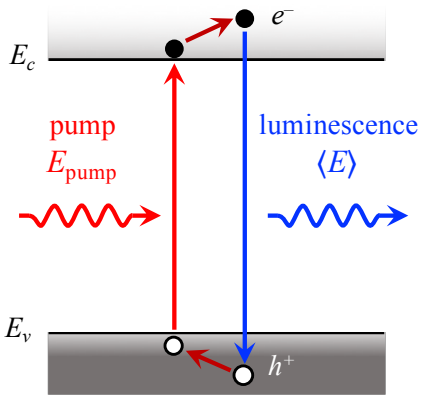

Figure 4.8: In laser cooling of solids, a semiconductor (or a rare-earth doped glass) is pumped by a laser whose photon energy is tuned to excite carriers from the valence band edge to the conduction band edge. The carriers thermalize to their quasi-equilibrium energy distributions and recombine radiatively to produce luminescence, whose energy exceeds the pump photon energy.

We next compare electroluminescent cooling to solid-state laser cooling, which is an alternative optoelectronic means of refrigeration that relies upon photoluminescence rather than electroluminescence. In laser cooling, an optically active material is excited by laser light and emits fluorescence whose energy is greater than that of the absorbed photons. The pump photon energy E_{pump} is chosen to excite carriers from the top of the valence band to the low-lying states in the conduction band as shown in Fig. 4.8. Lattice cooling occurs as the quasi-equilibrium carrier distributions are restored by electron-phonon scattering. The excited carriers must then recombine radiatively and the luminescence must be emitted externally in order for heat to be pumped out of the device. Therefore, the internal luminescence efficiency η_{int} and the light extraction efficiency C_{ext} are as important for laser cooling as they are for electroluminescent cooling. In laser cooling, the design is simplified as there is no need for electrical contacts. This can lead to a higher external luminescence efficiency and potentially a higher power density due to the lack of current spreading issues, both of which are advantages over electroluminescent cooling.

The amount of heat extracted per photon in laser cooling is given directly by the difference between the luminescence photon energy and the pump photon energy. In semiconductors, this blueshift is typically close to kT [8]. This arises from the fact that the bandgap energy E_g serves as a sharp lower limit on the pump photon energy: for lower E_{pump}, band-to-band electronic absorption falls off very rapidly and the optical pump is largely consumed by parasitic absorption. Meanwhile, the average luminescent photon energy is $\langle E \rangle \approx E_g + kT$. Thus, each emitted photon extracts about kT of thermal energy from the semiconductor. If the material fails to convert a pump photon into an external luminescent photon, the full energy E_g of the pump photon is wasted. We can perform a power balancing analysis, as we have done in Section 4.2 for LED cooling, to arrive at a minimum value of η_{ext} to achieve

net cooling via photoluminescence [8]:

$$\eta_{\text{ext}} > 1 - \frac{kT}{\langle E \rangle} \tag{4.17}$$

where we have assumed that the pump photon is electronically absorbed with unity probability. This is a very demanding efficiency requirement, especially in wider-gap materials where Auger recombination does not dominate. Nonetheless, net cooling by laser excitation has been achieved, and the observation of laser cooling in a semiconductor is an indicator of near-ideal optoelectronic material quality [87], [88].

While laser cooling extracts roughly kT of thermal energy per photon regardless of power density, the amount of heat extracted per photon in electroluminescent cooling is bias-dependent and is given by $\langle E \rangle - qV$. In the low to moderate bias regime, each photon carries away several (or many) kT's of thermal energy, though this comes at the cost of a lower luminescent heat flux. Nonetheless, for applications that call for a moderate power density, electroluminescent cooling can operate at much higher values of COP that are inaccessible by laser cooling.

4.5 Electroluminescent cooling at room temperature

In the following two sections, we evaluate the performance of the optimized GaAs LED structure discussed in Chapter 3 as an electroluminescent refrigerator, making use of the thermophotonic configuration in Section 4.3. As stated previously, we adopt the same device geometry for both the LED and the PV cell. On the PV side, we only make the following small changes to the device structure in Fig. 3.1: a thicker GaAs active layer ($d = 400$ nm), a thicker p^+-GaInP cladding layer ($d_p = 400$ nm), the absence of a thin lightly doped p-GaInP layer, and re-adjusted layer thicknesses in the distributed Bragg reflector to account for the small redshift in PV luminescence relative to the LED luminescence.

In this section, we operate the thermophotonic system between the temperatures of $T_{\text{c}} = 263\text{K}$ ($-10°\text{C}$) and $T_{\text{h}} = 313\text{K}$ ($40°\text{C}$), emulating a steady-state cooling application near room temperature. Its performance under larger and smaller temperature differences will be considered at the end of this section. In the results to follow, we assume a constant heat leakage of $Q_{\text{leak}} = 100\,\mu\text{W/cm}^2$, which is small enough to be irrelevant for practical cooling applications. In Appendix C, we discuss how the system can be engineered to reach this level of heat leakage.

Our methodology to characterize the cooling performance is as follows. We sweep the value of the LED internal voltage V_{c}, which in turn sweeps the cooling heat flux Q_{c} over several orders of magnitude. For each value of V_{c}, we find the optimum value of the PV bias V_{h} that maximizes the COP.[8] As shown in Fig. D.1 (solid black curve), the optimal PV bias

[8]We optimize the PV voltage without accounting for resistive effects for computational expedience. We find that in practice this is very close to the true optimum, since the choice of V_{h} has little effect on the resistive effects in either device.

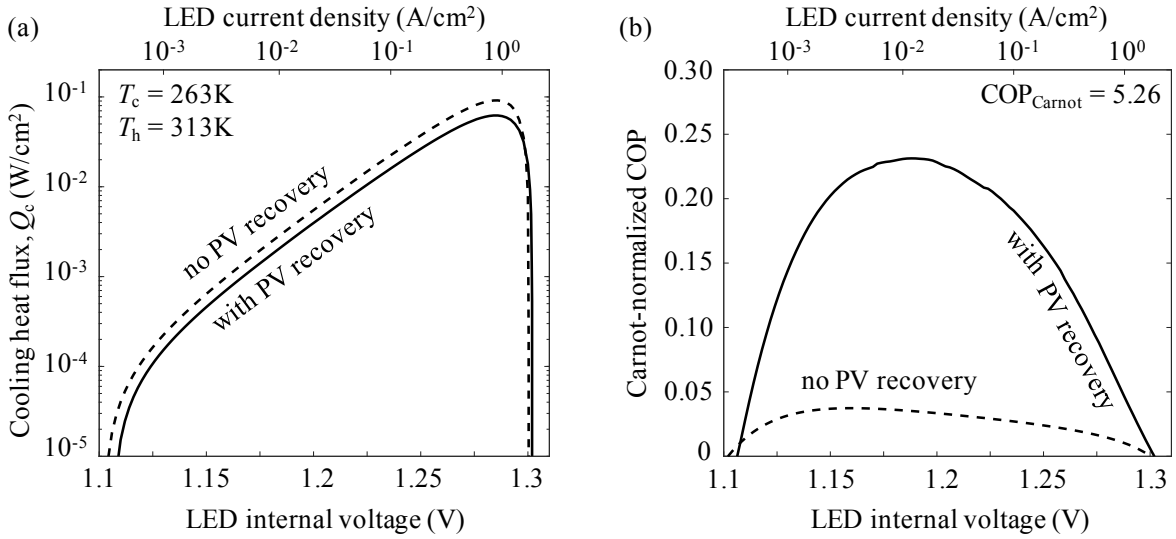

Figure 4.9: Voltage-dependent (a) cooling heat flux and (b) Carnot-normalized COP of the GaAs thermophotonic cooling system, operated between $T_c = 263$K and $T_h = 313$K. The performance is shown both with (solid) and without (dashed) energy recovery from the PV cell. The cooling heat flux is affected marginally by the PV cell, but the COP increases considerably.

tends to be 100 – 200 mV below the LED voltage. With the biases known, we calculate the effects of spreading resistance in the two devices, each of which carries both a forward current and a reverse current. The former dominates in the LED while the latter dominates in the PV cell: thus, the Ohmic losses in both devices grow with the intensity of LED luminescence. Finally, we correct for lateral voltage variations in the two devices, as explained in Section 3.4, to arrive at the final values of Q_c and the COP for this bias point.

Fig. 4.9(a) and (b) show the cooling heat flux Q_c and the Carnot-normalized COP, respectively, of the electroluminescent refrigerator as a function of the LED internal voltage. We show the performance with and without the energy recovery action of the PV cell; in the latter case, we let $V_h = 0$. The net heat flux from the LED rises above the level of the parasitic heat leakage ($100\,\mu W/cm^2$) at a bias of roughly 1.1V, then increases exponentially with LED voltage. At the upper extreme, near 1.3V, the net heat flux peaks and subsequently falls rapidly below zero; beyond this point, electroluminescent cooling cannot compensate for the heat generated by Ohmic dissipation and contact absorption. This system supports a maximum cooling power density of $Q_{c,max} = 62$ mW/cm^2. The cooling capacity diminishes slightly in the presence of the forward-biased PV cell, whose luminescence returns heat to the cold side. However, by virtue of the electrical power recovered from hot to cold, the COP improves considerably at all biases, as shown in Fig. 4.9(b). The COP peaks at 23.1% of Carnot when $V_c = 1.19$V (which yields $Q_c = 2.42$ mW/cm^2). Because more heat is pumped per photon under low bias, the COP peaks at a lower voltage than the external luminescence efficiency, which reaches its maximum near 1.25V (see Section 3.5).

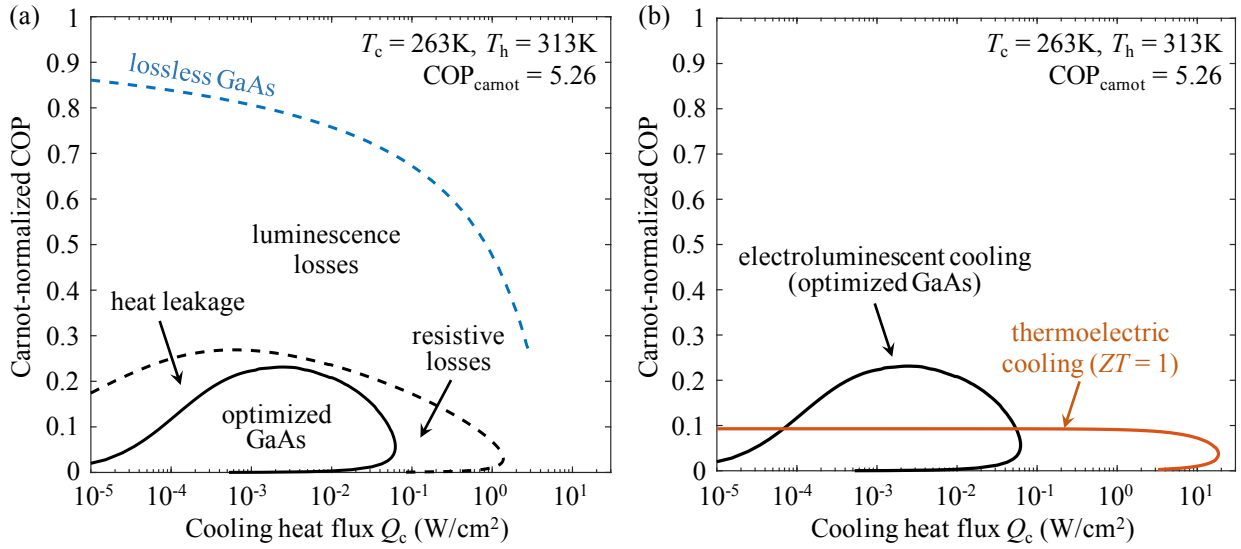

Figure 4.10: The principal cooling figures-of-merit are plotted against one another for the electroluminescent refrigerator, operated near room temperature. (a) The performance of the optimized GaAs system with all losses included (solid black), with zero heat leakage and electrical resistance (dashed black), and additionally with unity external luminescence efficiency (dashed blue). In (b), we compare the performance of the electroluminescent refrigerator with the highest-efficiency commercial thermoelectric coolers with $ZT = 1$. The solid black curves in (a) and (b) are identical.

In Fig. 4.10(a), we plot the two cooling figures-of-merit against one another, which more clearly reveals the trade-off between power density and energy efficiency in this refrigerator.[9] The black curve corresponds to the optimized GaAs system with the characteristics shown in Fig. 4.9. More efficient operation can be accessed at the expense of the heat flux, and vice versa, by adjusting the applied bias. We have also included two additional curves, representing more idealized performance: the dashed black curve is the system's performance after removing heat leakage and assuming that all device resistances are zero, and the dashed blue curve further assumes that both devices are lossless emitters with unity external luminescence efficiency. By comparing the two black curves, we can identify the regimes where heat leakage and resistive losses are dominant. We see that while these effects set practical upper and lower limits on the cooling flux, they have a relatively small effect on the peak COP. The large gulf in the peak COP of the optimized device and the lossless device (blue) can be largely attributed to luminescence losses; non-radiative recombination and parasitic absorption. The LED already attains a very high efficiency of $\eta_{ext} = 98\%$, but the remaining 2% accounts for the system's large deviation from ideal! Put in another way, the electroluminescent cooling application reaps large benefits in performance from asymptotic approaches to material and device perfection, as we have already seen in Fig. 4.6.

[9]Fig. 4.10 shows parametric curves, where the LED bias is the (hidden) independent parameter. The LED bias increases first from left to right then clockwise along the parametric curve.

In Fig. 4.10(b), we compare the performance of the optimized GaAs electroluminescent refrigerator (black) with a thermoelectric cooler operating between the same temperatures (red). We assume a material figure-of-merit of $ZT = 1$, representative of the best commercially available thermoelectric modules near room temperature [86]. The maximum possible COP of the thermoelectric cooler can be expressed as a simple function of ZT [83]:

$$\frac{\text{COP}}{\text{COP}_{\text{Carnot}}} = \frac{\sqrt{1+ZT} - T_{\text{h}}/T_{\text{c}}}{\sqrt{1+ZT}+1} \tag{4.18}$$

which evaluates to 9.3% of the Carnot limit for the selected temperatures. To represent the full range of cooling properties accessible with a $ZT = 1$ material, every point on the red curve in Fig. 4.10(b) corresponds to a different device thickness and bias current: see Appendix E for the details. The maximum normalized COP of 9.3% can be accessed for cooling fluxes up to 1 W/cm^2, above which the performance is degraded by electrical and thermal contact parasitics. We predict the maximum realistic thermoelectric cooling capacity to be between 10 and 100 W/cm^2 for this temperature difference, which is significantly larger than that of the electroluminescent refrigerator even in the lossless case. However, as suggested in Section 4.4, electroluminescent refrigeration possesses an efficiency advantage at low and moderate power densities. In Fig. 4.10(b), we find that the electroluminescent COP exceeds that of thermoelectrics for $Q_{\text{c}} < 50$ mW/cm^2. For power densities in the 1 – 10 mW/cm^2 range, the electroluminescent refrigerator with optimized GaAs devices operates with more than twice the energy efficiency of thermoelectric coolers.

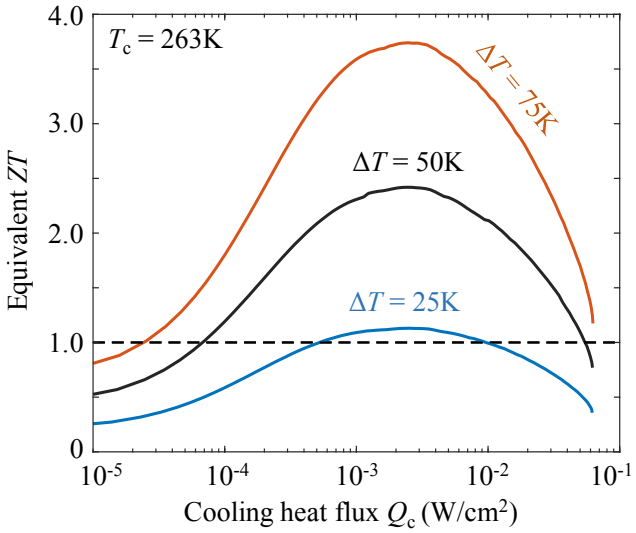

Figure 4.11: The equivalent ZT of the electroluminescent refrigerator (evaluated from the COP) as a function of heat flux for several temperature differences ΔT. The cold-side temperature is fixed at 263K. For larger ΔT and moderate heat fluxes, the equivalent ZT can greatly exceed unity, which represents the performance of the best commercial thermoelectrics.

To further explore the comparison, we can find the equivalent thermoelectric figure-of-merit ZT_{eff} of electroluminescent cooling by solving for ZT in Equation (4.18). The result is shown in Fig. 4.11 for several values of the temperature difference ΔT, assuming a fixed cold-side temperature of 263K. For the previously considered case of $\Delta T = 50$K, the electroluminescent refrigerator operates with a peak value of $ZT_{\text{eff}} = 2.4$, which greatly exceeds that of the best commercial thermoelectrics at room temperature. For larger ΔT, the equivalent ZT of electroluminescent cooling becomes even greater. This efficiency advantage diminishes as the hot and cold temperatures move closer together, eventually vanishing for $\Delta T < 20$K. This is because as ΔT approaches zero, the thermal conduction leakage in thermoelectric coolers decreases proportionally, improving the thermoelectric COP. Meanwhile, none of the primary losses in the electroluminescent refrigerator are driven by a temperature difference. The condition $\Delta T \to 0$ does not lead to any direct performance benefits, allowing thermoelectrics to gain the efficiency advantage.

What can we conclude about the practical limits of electroluminescence as a cooling mechanism? In Fig. 4.10, the optimized GaAs system operates at 20.8% of the Carnot limit while pulling 10 mW/cm^2 of heat from the LED active region. With efficient heat spreading, this high COP might be supported at a heat flux of up to 100 mW/cm^2 from a larger thermal load. While these fluxes are considerably lower than those provided by thermoelectrics for such applications as CPU and laser diode cooling, they are comparable to the power densities at which domestic refrigerators and air conditioning units operate. While evaporator units can attain a higher COP than our optoelectronic system, they tend to be bulky and use refrigerants that are eventually released as atmospheric pollutants. A portable, reliable, and efficient solid-state cooling technology could find niche applications at room temperature. For example, medicines often need to be kept below ambient temperature to maintain their structure (and function), and a compact electroluminescent refrigerator might offer a viable option to keep them cool while in transit.

4.6 Electroluminescent cooling at cryogenic temperatures

At cryogenic temperatures, where laser cooling has achieved success [8], [88], electroluminescent cooling likewise becomes more efficient. Semiconductor diodes are more ideal light emitters at reduced temperatures, as thermally activated non-radiative processes are suppressed and as radiative recombination is accelerated for the same carrier density. In this section, we evaluate the performance of electroluminescent cooling down to a temperature of 50K, which remains in the regime where carrier freeze-out effects are negligible. Operation at far lower temperatures than 50K should be possible, since GaAs has shallow donor states [89] and because the n-GaAs active region is nearly degenerately doped, so that almost no energy is needed for donor ionization. Nonetheless, we show that peak cooling efficiency is achieved at a cryogenic temperature above 50K.

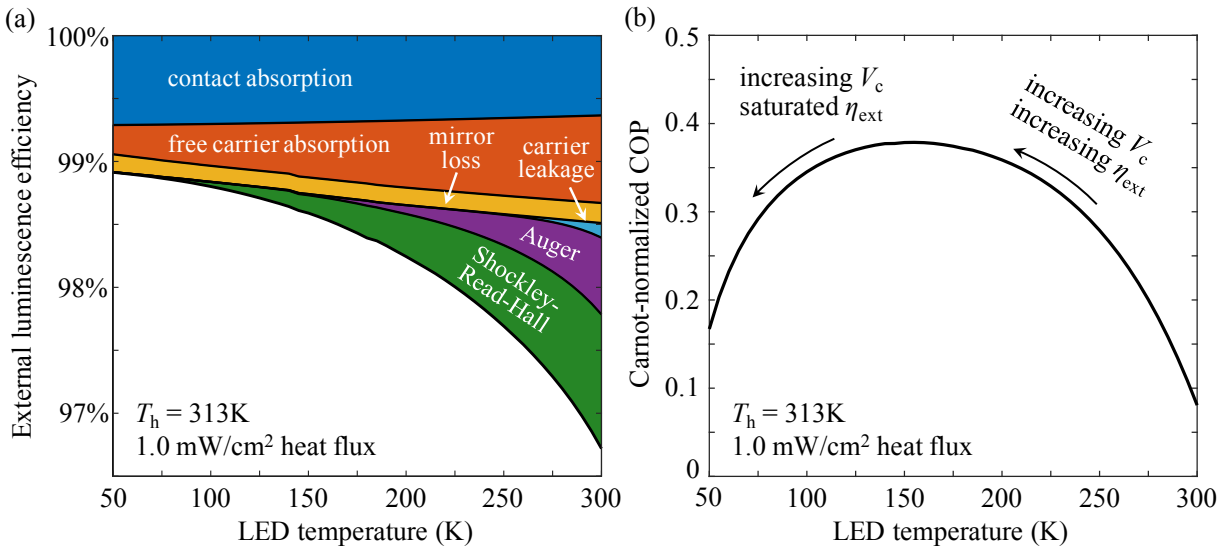

Figure 4.12: (a) The external luminescence efficiency of the GaAs LED (Fig. 3.1) increases at low temperatures, largely due to the suppression of non-radiative recombination. The LED applied bias must increase at low temperatures to provide a fixed heat flux of 1 mW/cm^2 in the thermophotonic system with $T_h = 313$K. (b) The cooling COP, evaluated at the same heat flux, increases below room temperature and reaches a peak near 150K. Below 150K, the increase in voltage – which reduces the cooling per photon – outpaces the increase in η_{ext}, and the COP decreases. 100 μW/cm^2 of heat leakage is assumed.

Following experimental measurements, we model the bulk Shockley-Read-Hall rate $1/\tau_{srh,b}$ and the surface recombination velocity S in GaAs to have a temperature dependence of the form $f(T) = f_0 \exp(-E_A/kT)$ with $E_A = 18$ meV [40]. The perimeter SRH recombination parameter J'_{0p} is modeled to follow the same activation energy, but since it is an integrated quantity over thickness, it additionally follows the strong temperature dependence of the intrinsic carrier density: $n_i \sim T^{3/2} \exp[-E_g(T)/2kT]$ [52]. The Auger process in GaAs is also thermally activated, and the Auger coefficients C_n and C_p decrease at low temperature with a larger activation energy of $E_A = 58$ meV [40], [90].

To model the absorption coefficient $\alpha_0(E)$ at low temperatures, we fit the data in Ref. 47 to a piece-wise model above and below the bandgap as in Ref. 5 and account for the temperature dependence of the GaAs bandgap and the Urbach energy [91]. Inserting these dependences into Equation (3.5), we find that the radiative recombination coefficient B increases at low temperatures approximately as $T^{-1.78}$. The electron and hole mobilities increase at low temperatures with a peak near 100K [92]. We take the absorption cross-sections of free electrons and holes to vary with the reciprocal of the corresponding mobilities, and thus decrease for temperatures below 300K [93].

Fig. 4.12(a) shows the improvement in the external luminescence efficiency of the optimized GaAs LED (Fig. 3.1) with decreasing temperature, resulting from the favorable trends

described above. We adjust the LED bias to yield the same heat flux of $Q_c = 1.0$ mW/cm^2 across the full range of LED temperatures, with the PV temperature fixed at $T_h = 313$K. These bias points do not necessarily coincide with the peak external luminescence efficiency of the device. Non-radiative recombination is strongly suppressed with decreasing temperature – carrier leakage most rapidly, then Auger recombination, and finally Shockley-Read-Hall recombination. As a result, the internal luminescence efficiency asymptotically approaches unity at low temperatures, exceeding $\eta_{int} = 99.999\%$ at 50K. The remaining luminescence losses are mainly optical and have weak temperature dependence, such as rear reflector absorption and contact absorption. These mechanisms cause the external luminescence efficiency to saturate nearly to a constant at very low temperatures. At 50K, the LED attains an external luminescence efficiency of $\eta_{ext} = 98.9\%$.

The cooling COP of the system at the same LED bias points, with optimally chosen PV biases, is shown in Fig. 4.12(b). As a result of increasing external luminescence efficiency, the COP improves as the system is cooled below room temperature. For the fixed value of heat flux that we have chosen (1 mW/cm^2), the COP peaks near 150K, then decreases. This can be understood as follows. Since the LED emits less thermal radiation at lower temperatures, the applied bias must increase in order to maintain the same luminescent heat flux:[10] from 300K to 50K, this voltage increase is about 370 mV. Meanwhile, over the same temperature range, the bandgap of GaAs increases by only 90 meV. Thus, electroluminescent cooling at low temperatures requires moving the LED bias closer to the bandgap to deliver the same total heat flux. As a consequence, the luminescence pumps less heat out of the device per photon, which has the effect of decreasing the COP. In lowering the temperature from 300K to 150K, this harmful effect is more than compensated by the increasing external luminescence efficiency. Below 150K, the external luminescence efficiency has nearly saturated while the required bias continues to rise: therefore, the COP begins to decline.

Fig. 4.13 shows the system's cooling characteristics when operated between the temperatures $T_c = 113$K and $T_h = 313$K, again assuming a leakage of $100\,\mu$W/cm^2. For this application, we use a lower doping density of $N_D = 9.0 \times 10^{16}$ cm^{-3} to ensure that the active region stays at the edge of degeneracy; a deeply degenerate doping density reduces the radiative recombination rate via the band-filling effect. The cryogenic cooling curve is similar in character to the room-temperature cooling curve in Fig. 4.10. The peak COP is 35.2% of the Carnot limit and occurs at a heat flux of 1.0 mW/cm^2; at a larger heat flux of 10 mW/cm^2, the COP is 25.3% of the Carnot limit. The LED external luminescence efficiency at both operating points is close to $\eta_{ext} = 98.8\%$. Owing to the more ideal luminescence properties at cryogenic temperatures, the optimized device operates closer to the lossless cooling curve (dashed blue) than in the room temperature case.

As we mentioned in Section 4.4, thermoelectric cooling is sustainable only above about 200K. For the cold and hot temperatures considered above, Equation (4.18) gives a requirement of $ZT > 6.7$ in order to produce net cooling. This is not only a very demanding material requirement but an impractical one, as thermoelectric effects are considerably weaker

[10] This can be seen directly from Equation (2.7).

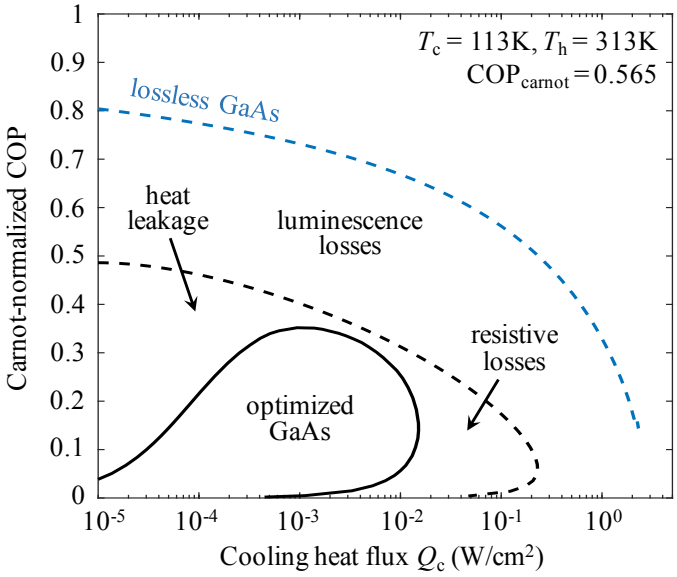

Figure 4.13: The cooling characteristics of the GaAs electroluminescent refrigerator operating between cryogenic and room temperature. The performance of the optimized GaAs system is shown with all losses included (solid black), with zero heat leakage and zero electrical resistance (dashed black), and additionally with unity external luminescence efficiency (dashed blue).

at low temperatures relative to the competing non-idealities [83]. Therefore, electroluminescent refrigerators can potentially displace thermoelectrics for the low-temperature cooling of solid-state components, such as infrared photodetectors. To efficiently detect photons whose energy is comparable or smaller than kT, photodetectors must be cooled for the same reason that LEDs operate more efficiently when cold: the suppression of non-radiative processes, particularly Auger generation [76]. Electroluminescent refrigeration systems might also be used for scientific experiments at cryogenic temperatures, down to 77K and possibly down to 4K, though we did not investigate the ultra-cold regime in this work. LED cooling competes in these applications with laser cooling, and may potentially be simpler to implement once the technology is optimized.

Finally, we note that because non-radiative recombination processes are naturally quenched at low temperatures, the requirement on material quality is less stringent in cryogenic cooling compared to room-temperature cooling. This opens the electroluminescent refrigeration application to a wider class of optoelectronic technologies with comparable or slightly inferior material quality (or technological maturity) to GaAs. Lead halide perovskites, for example, have been demonstrated as excellent light emitters in low-temperature laser cooling experiments [88], and rapid progress has been made in LEDs based on this material system [94].

4.7 Approaches to enhance cooling power density

If we observe the curves corresponding to ideal device operation in Fig. 4.10(a) and Fig. 4.13, we notice that the maximum heat flux that is achievable with electroluminescent cooling appears to be limited to the range $1-10$ W/cm^2. This maximum corresponds to operation just below the Bernard-Duraffourg condition, where the luminescence photon flux is large but the heat pumped by each individual photon is close to zero. What pathways are available to increase this upper bound on heat flux, potentially enabling new cooling applications?

One approach is to distribute the luminescent heat transfer over a wider band of photon energies. In principle, this can substantially increase the power density, since the geometry we have considered uses only a narrow slice of the spectrum (several kT's) surrounding the bandgap of GaAs. We can conceivably stack several LEDs of different bandgaps, whose luminescence is captured by a similar stack of PV cells on the hot side; each side is arranged in a manner similar to a tandem solar cell, but with independent electrical connections to each device. Ideally, this boosts the maximum power density by the number of different bandgaps used. This wavelength multiplexing scheme is destined to fail, however, because to maintain a high COP, every LED in the stack (and especially the wider-bandgap devices) needs to be exceptionally efficient. Only a small number of optoelectronic technologies can ever reach the efficiencies needed for cooling, and maintaining these high efficiencies while integrating them in the same system would be enormously challenging. Meanwhile, the benefit to the power density would be no more than a factor of three or four: it is impractical.

Another factor that sets an upper limit on the heat flux in our system is the need to emit photons from a high-index semiconductor into the low-index vacuum gap. As discussed in Section 3.3, this leads to the inherent difficulty of light extraction. It also fundamentally limits the photon flux that can leave the LED, as these photons must first couple to the radiative modes in vacuum before being transferred to another high-index medium, the PV cell. The vacuum gap therefore acts as an inconvenient obstacle for luminescence transfer, though it is necessary, of course, to prevent shorting the hot and cold sides thermally. If the gap distance is reduced below the luminescent wavelength, photons can tunnel directly across the vacuum gap while direct phonon conduction can still be cut off to a large degree. By allowing both the radiative and the evanescent modes of luminescence to couple directly between two high-index media, the photon flux and thus the heat flux can be enhanced by multiple orders of magnitude [79]. This is the concept of near-field radiative heat transfer; applied to our thermophotonic system, we refer to it as near-field electroluminescent cooling. In the following, we briefly summarize the main features of this scheme – for a more complete analysis, we refer the reader to our published manuscript, Ref. 68.

In addition to a significant enhancement in power density, near-field cooling enables other potential benefits. Since total internal reflection at the front surface ceases to be an issue, light extraction is greatly accelerated without the need for surface texturing or other light out-coupling techniques. Consequently, the luminescent photons linger for a far shorter time inside the device, leading to a reduction in the optical losses and an increase in the external

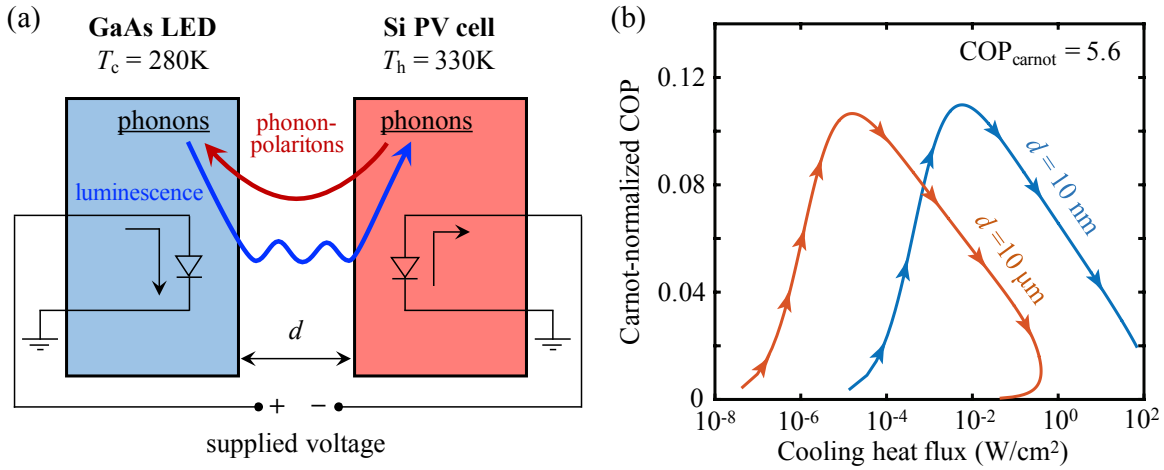

Figure 4.14: The near-field thermophotonic refrigeration scheme. (a) When the vacuum separation d between the LED and PV cell is sub-wavelength, luminescent photons as well as phonon-polaritons can couple directly between the two devices, greatly enhancing the transfer of heat by both mechanisms. To mitigate heat leakage, we choose two different semiconductors, GaAs and Si, with non-overlapping optical phonon resonances. (b) The theoretical limit of cooling performance for this system with a separation of 10μm (far field) and 10 nm (near field), reproduced from Ref. 68. Idealized devices are assumed where Shockley-Read-Hall and Auger recombination are the only losses. We assume the absence of an SiO_2 coating on the PV cell.

luminescence efficiency. Furthermore, since free-carrier absorption and contact absorption now play a more minor role, this allows for a more aggressive electrical design (such as a shorter contact separation) to reduce the Ohmic losses. This would potentially allow the device to actually utilize the larger electroluminescent heat fluxes that are now accessible.

The near-field approach, however, is not without drawbacks. While a very small vacuum gap enhances the direct coupling of luminescent photons, it also enhances the coupling of optical phonons from one device to another as phonon-polaritons. Without proper mitigation, this heat leakage channel potentially dominates the heat transfer. An approach to suppress this leakage is to select two different semiconductors for the LED and PV cell with non-overlapping optical phonon resonances. In particular, we can choose a GaAs LED and a Si PV cell as in Fig. 4.14(a). While GaAs supports optical phonons near 33 meV, the non-polar Si crystal possesses very weak optical phonon transitions. In practice, however, a layer of SiO_2 inevitably coats the surface of Si devices, whether intentionally for passivation or spontaneously upon exposure to oxygen. The surface phonon-polaritons of amorphous SiO_2 (at 57 meV and 133 meV) can couple to the optical phonons in GaAs, contributing a parasitic heat flux that depends on the thickness of the oxide layer.

The rate of near-field heat transfer between the two devices can be calculated by treating the electromagnetic radiation as originating from an ensemble of fluctuating dipoles that fill the volume of the two devices. This approach is known as fluctuational electrodynamics

[95], and the specific application of this theory to our system is detailed in Ref. 68. Unlike the calculations in Sections 4.5 and 4.6, we do not evaluate the performance of a realistic optimized LED device structure. We instead perform a more idealized calculation that assumes devices with no optical losses or Ohmic losses, but does include a model for Shockley-Read-Hall and Auger recombination. For optimal luminescence efficiency, we choose a 500 nm thick p-type GaAs layer and a 50 μm thick undoped Si layer.

Fig. 4.14(b) shows the calculated room-temperature cooling performance of this idealized system with a 10 μm vacuum separation, corresponding to the luminescent far field, and a 10 nm vacuum separation, which is in the near field. The curves are similar in shape to the far-field characteristics, but the heat leakage occurs by the transfer of phonon-polaritons. As the hot-to-cold separation decreases, both the luminescent heat flux and the phonon-polariton leakage are enhanced by two orders of magnitude, with a net cooling flux that reaches a theoretical maximum of about 100 W/cm^2. The COP, on the other hand, is lower for this system than for the GaAs-GaAs far-field refrigerator. This is caused primarily by the use of a Si PV cell, which has a smaller bandgap and therefore recovers a significantly lower voltage and power to the LED. Fig. 4.14(b) does not account for the presence of an SiO_2 coating on the PV cell. With a 20 nm SiO_2 coating, the peak COP drops from 11.0% to 4.3% of the Carnot limit.

The near-field refrigeration scheme can potentially raise the maximum cooling capacity of electroluminescent refrigeration to the power densities provided by thermoelectrics, though it remains to be seen whether this holds true once we include the full effects of resistive losses and, to a lesser degree, optical losses. In any case, it does not seem to provide an efficiency advantage over thermoelectrics. From the standpoint of cooling efficiency, the near-field refrigerator (assuming idealized devices) performs worse than the far-field refrigerator (with practical devices and accounting for all realistic losses).

Is there another cooling scheme that allows for the direct coupling of photons between high-index media without heavily compromising the COP? Ideally, this would require minimizing both the parasitic heat leakage and the difference in bandgaps. One possibility is to bridge the LED and a PV cell of the same material with another high-index material that is transparent to the luminescence. Of course, the conductive heat leakage across the high-index bridge needs to be suppressed. This can be done by elongating the bridge into a very long light pipe, though this not likely to be practical. A more realistic approach might be to introduce periodic sub-wavelength vacuum gaps into the high-index bridge, which are transparent to the luminescent photons but will effectively block the conduction of phonons across the bridge from hot to cold.[11]

[11] This idea is credited to Parthi Santhanam at Stanford.

Part II

Analog circuits for combinatorial optimization

Chapter 5

Analog machines for combinatorial optimization

There is no shortage of difficult optimization problems in modern society. Many of these are motivated by our desire to allocate precious resources such as time, space, energy, and money in the best way possible: finding the shortest route between points on a map, optimally planning a large number of aircraft routes and crew schedules, efficiently scheduling computational tasks in a data center, packing objects of variable size into finite-sized containers [96], and making the most profitable investments into a portfolio of financial assets [97]. These problems arise in integrated circuit design, where components should be optimally placed and routed to minimize delays, die area, and power dissipation [98]. They arise also in biology and medicine in the prediction of protein folding [99] and in the optimization of gene sequencing [100] and drug discovery [101] techniques. The success of many machine learning methods hinge on our ability to solve optimization problems of both the continuous and the combinatorial type [102], [103]. There are myriad examples of other problems with great practical as well as mathematical interest.

Two important features of combinatorial optimization problems are their universality and their difficulty [96]. It turns out that many of these problems, despite emerging in different fields, are mathematically similar. An algorithm that is efficient for solving a specific problem is likely to be useful for a wider class. Unfortunately, many of these problems are also difficult, sometimes intractably hard; we have not discovered, and we are unlikely to discover, any algorithm for these problems that is guaranteed to find the most optimal solution within a reasonable time.

It appears to be a miracle then that such hard problems are routinely tackled, often with great success. This is a testament to the power of our algorithms, which can find very good (if not the best) optima in an efficient manner, as well as the enormous capability of our computational resources. However, the imminent end of Moore's law and Dennard scaling means that our ability to perform computationally intensive tasks, like machine learning and combinatorial optimization, will likewise face significant scaling challenges. To address this scaling issue on a more fundamental level, we can view as a limitation the sequential

nature of our algorithms; they are designed to run, instruction by instruction, within the von-Neumann architecture of today's digital computers. As an alternative, there are already a number of analog hardware solutions, spanning a diverse set of architectures, that exploit the computational ability of physical processes to solve difficult optimization problems. These include simulated annealing accelerators using CMOS technology [104], coupled oscillator networks [105]–[108], memristor-based Boltzmann machines and neural networks [109], [110], networks of physical devices that implement invertible logic gates [111], [112], and quantum annealers or adiabatic computers [113], [114].

In Section 5.1, we give an overview of the general approaches that are used to solve combinatorial optimization problems, which encompass the methods implemented by some of the hardware systems listed above. These include simulated annealing and adiabatic optimization, both of which are very generally applicable. In Section 5.2, we present another general principle for optimization that can potentially be leveraged for global minimization, provided that a physical system can be realized whose underlying dynamics fully embodies the principle. This mechanism, proposed in Ref. 115 and Ref. 107 but which we will refer to as the "first-to-threshold" search in this work, does not yet truly have an exact physical implementation. In Chapter 6, we will design an analog optimization machine with the goal of putting this search method to practice.

We must not forget that decades ago, analog computers faded away in importance primarily because they lacked the numerical precision that digital computers could offer. To address these issues in the present class of analog machines, we must provide built-in robustness to the practical issues of component variability and drift. This is done by restricting both the input and the output of the machine to be digital, while exploiting the analog processing of information internally. Even then, we find that the achievable degree of analog precision limits the size of the optimization problems that can be reliably solved. We will discuss these analog precision requirements in Section 5.4.

5.1 General approaches to combinatorial optimization problems

The most famous example of a combinatorial optimization problem is the traveling salesman problem: given a list of N cities and their spatial coordinates, what is the shortest route that visits each city and returns to the starting city? One of the first relatively large-scale instances of this problem to be solved is shown in Fig. 5.1 along with its solution [116]. To the best of our knowledge, the traveling salesman problem is intractable. As more cities are added, the time it takes to find the shortest route increases exponentially (or faster). Many of the other optimization problems described in this chapter's opening are also apparently intractable.

More formally, intractability implies that the problem cannot be solved by a polynomial-time algorithm: that is, there is no algorithm which is guaranteed to find the globally

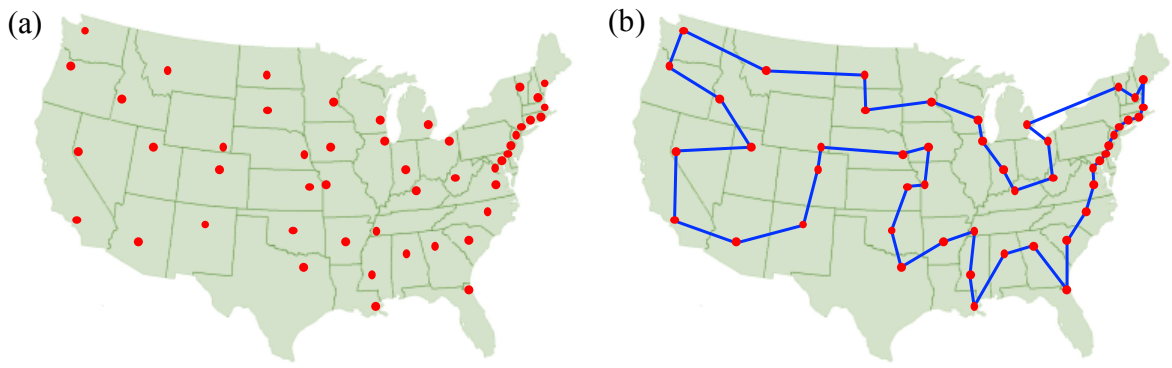

Figure 5.1: An instance of the traveling salesman problem: what is the shortest tour through the 49 U.S. cities shown in (a)? The solution is shown in (b). The list of cities and a method for determining the shortest route can be found in Ref. 116.

optimal solution to the problem using only resources (usually time, but potentially also space or energy) that scale as a polynomial function of the problem size.[1] Many of the optimization problems which are apparently intractable, in that the best known algorithms for solving them require an exponential runtime rather than polynomial, can be shown to belong to the NP-hard complexity class. A key property of NP-hard problems is reducibility: the transformation from one NP-hard optimization problem into any other requires only a polynomial-time algorithm [117].[2] Despite the large body of empirical evidence, the claim that NP-hard problems cannot be solved by a polynomial-time algorithm is unproven, and this remains an important open question in computer science, though the near-universal opinion is that they cannot be.

Nonetheless, these intractable problems can be tackled in ways that yield near-optimal solutions, though not necessarily the global optimum, even in the worst case. We will briefly review a few general strategies in this section. A popular and efficient method is to use deterministic approximation algorithms, which are polynomial-time algorithms that are proven to deliver a solution that is within a constant factor of the global optimum. The existence of such an approximation algorithm, and the value of the constant factor, is problem-dependent. For example, for the traveling salesman problem, an approximation algorithm exists if the distances between cities obey the triangle inequality; this applies to most practical applications such as the instance in Fig. 5.1, where a direct route between two cities is always shorter than an indirect route. In the more general case, with no such constraints, there is no polynomial-time approximation algorithm for the traveling salesman problem [118]. For the Max-Cut problem, which we will consider later in this chapter, a well-known approximation algorithm guarantees solutions with a value that is at least 87.856% of

[1]The problem size is the number of independent inputs used to specify the problem, and the guarantee of a global optimum must extend to the worst-case instance of the problem.

[2]To be more precise, polynomial-time reducibility has been shown from the complexity class NP to the class NP-complete. The decision problem versions of NP-hard optimization problems are NP-complete.

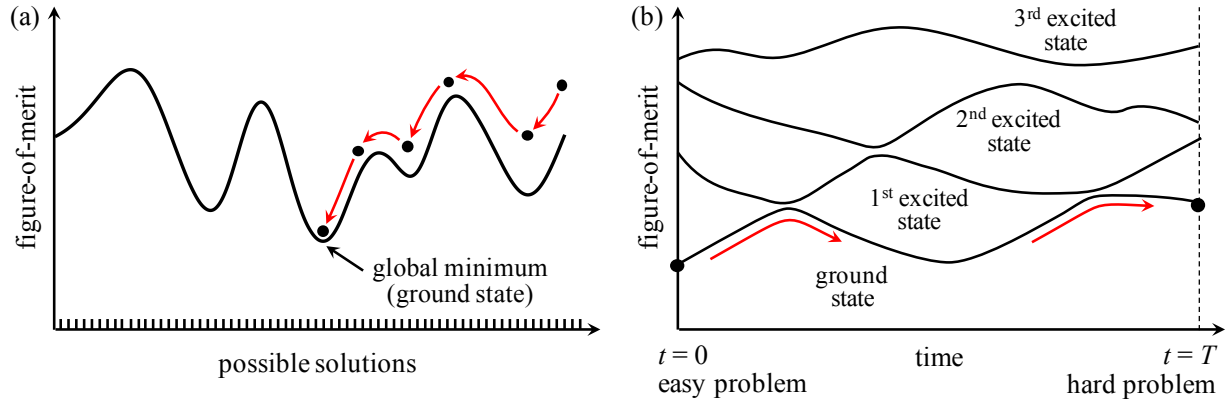

Figure 5.2: Two existing approaches to global minimization. (a) Simulated annealing combines a greedy local search with stochastic steps in the wrong direction to move from one minimum to another, potentially converging on the global minimum. The possible solutions are discrete, though a continuous figure-of-merit curve is shown for simplicity. (b) In adiabatic optimization, a physical system is first configured to represent a simple optimization problem with a known ground state. The system is then gradually transformed into a representation of the difficult optimization problem; the energies of all the states will change, but if evolved slowly enough, the system remains in its ground state.

the global optimum [119]. Since approximation algorithms are efficient and reliable if they exist, a competitive alternative method should be able to outperform them in a significant number of cases, or with a significant probability.

A very general class of methods, applicable to many large-scale optimization problems, relies on a probabilistic search of the solution space. Fig. 5.2(a) shows the typical figure-of-merit landscape of a difficult non-convex optimization problem. The challenge is to efficiently search the very large solution space for a good minimum, ideally arriving at the global minimum, without being trapped in one of the many poor local minima. In simulated annealing, a local (or greedy) search is used to move toward a local minimum, but the algorithm may also move the figure-of-merit in the wrong direction, allowing the system to escape from poor local minima and thus explore the solution space more broadly. The decision to accept such a step is probabilistic, and this probability is controlled by a global parameter that is analogous to physical temperature [120]. If the temperature is reduced over many iterations to zero, the system might converge to the global minimum by the end of the annealing process; the more iterations, the greater the likelihood of finding the global minimum, though this is never guaranteed. There are many other related heuristics for searching the solution space, such as tabu search and genetic algorithms, which are also widely used.

Simulated annealing is an optimization heuristic that can be readily implemented on a digital computer. There are, however, potential benefits to implementing this class of methods in analog hardware. The process of searching for local minima can be accelerated

by exploiting physical processes in an analog system, in which different parts of the machine can process information in parallel while still communicating continuously. When combined with a source of noise as in Refs. 105 – 106, this can lead to a more efficient implementation of probabilistic global minimization methods. Additionally, physical noise that is inherent in analog devices can be harnessed to provide true random number generation, which is a core ingredient for these probabilistic approaches [105], [110].

Another general method for global optimization is adiabatic computing or annealing. In this approach, a simple (e.g. convex) optimization problem is first encoded in the energy landscape of a physical system, which quickly relaxes to the global minimum. The simple problem is then continuously transformed into the difficult optimization problem, in the process modifying the energy levels of the physical states as shown in Fig. 5.2(b). If done slowly enough, the physical system remains in its ground state through the full evolution, and the final state can be measured to yield the global minimum of the problem. This is made possible by the existence of avoided crossings between the energy levels of the possible states in a physical system – in quantum mechanics, this leads to the adiabatic theorem for a time-dependent Hamiltonian. As it evolves, the system moves continuously across these avoided crossings as long as the adiabaticity condition is met: if the system changes too rapidly, or if energy is supplied from an external source (such as thermal energy), it can undergo a Landau-Zener transition from the ground state to an excited state [121].

The adiabatic method has been applied successfully to small-scale optimization problems both in numerical simulations [113] and in the D-Wave quantum computer based on superconducting flux qubits [114], though the mechanism of the latter has been the subject of some debate [122], [123]. In any case, it is hypothesized that as the problem size increases, the energy gap between the ground and excited state will in general decrease exponentially for difficult problems, leading to an exponential runtime; this has been proven for some specific problem instances [124]. While quantum computers can provide an exponential speedup over classical computers in specific problems like prime factorization [125], it does not seem likely that a quantum algorithm exists for solving NP-complete or NP-hard problems in polynomial time [126].

5.2 The first-to-threshold method of optimization

In this section, we present a method of global optimization that is distinct from the general approaches presented in the previous section, and which relies explicitly on the dynamics of a physical system. We will call this the first-to-threshold method of optimization. We note that this theoretical scheme has been proposed previously in the literature, where it has been called the minimum gain principle [107], [115].

The first-to-threshold method is illustrated in Fig. 5.3. This search mechanism relies on a physical system that satisfies two requirements. First, as in other hardware-based approaches, the possible solutions to the combinatorial optimization problem must be encoded as distinct physical states of the system. Each of these states can be characterized by a

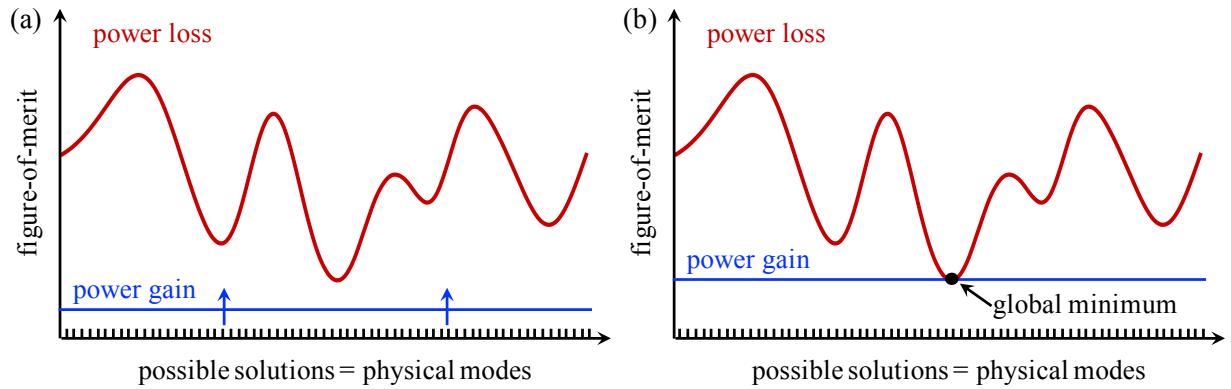

Figure 5.3: In the first-to-threshold method of optimization, the possible solutions to the problem are encoded as discrete physical states of the system, whose rate of power loss is proportional to the optimization figure-of-merit. An external source of power, ramped gradually from zero, allows all possible states to experience gain. (a) If the rate of power gain does not exceed the rate of power loss for any of these states, none of them can be sustained and noise dominates. (b) The first stable state to emerge is the one with the smallest loss; given the mapping from the loss to the figure-of-merit, this physical state encodes the global minimum of the problem.

rate of power loss or dissipation, with some states being lossier than others. The second critical requirement for the system, then, is that these loss rates must be proportional to the corresponding optimization figure-of-merit. More optimal solutions are less lossy, and the state with the least loss represents the global minimum, or the ground state of the problem. In general, establishing the mapping between the loss rates of these discrete physical states and the optimization figure-of-merit is a nontrivial design task.

If a machine that meets these requirements can be constructed, we can perform a search by gradually pumping external power into the system. Initially, with zero supplied power, none of the discrete states can be sustained since they are lossy; if excited momentarily by noise, they decay quickly back to zero. As the external pump increases, it imparts a rate of power gain to all of the possible states. At low pump power, all of the discrete states continue to experience a net power loss as shown in Fig. 5.3(a). Eventually, as the power gain increases, one of the states experiences a net loss of zero and can therefore be sustained, as shown in Fig. 5.3(b). The ramping of the pump power can be terminated when the first state to reach the threshold of zero net loss emerges from a sea of noise. This state is necessarily the one with the least loss, and by measuring the state of the system at the end of the ramping process, the global minimum of the problem can be discovered. The ramp must be slow enough that net gain is not supplied to the second-least lossy mode (or first excited state) before the system can settle to the global minimum. The first-to-threshold mechanism can be visualized as a "search from below," in contrast to methods like simulated annealing, which search for the global minimum from "above" – that is, starting from and passing through a number of less optimal solutions.

This optimization principle, though not a new idea, remains to date an idealization that has not been embodied by any physical system. The main difficulty lies in realizing an accurate mapping from the system's power loss rate to the optimization figure-of-merit. The first-to-threshold search was proposed in the context of competing oscillatory modes in optical cavities, such as those in lasers [115] and optical parametric oscillators [107], where an approximate mapping to the Ising Hamiltonian (described in the following section) was shown to hold under certain conditions. This mapping has been used as the basis for the coherent Ising machine, which uses a large number of time-multiplexed optical parametric oscillators that propagate through a long optical fiber [108], [127], [128]. While good performance on the Ising and Max-Cut problems have been experimentally demonstrated, the problem mapping has been recognized by the authors to be accurate only under restricted conditions, and not generally during the course of the system's evolution [129], [130].

In Chapter 6, we attempt to realize the first-to-threshold optimization principle using a network of coupled electrical oscillators, similar in structure to Ref. 105. We find, as in the coherent Ising machine, that this mechanism is challenging to implement; though in this we are ultimately unsuccessful, our system possesses dynamical properties that have shown to be useful in solving the Ising problem.

5.3 The Ising problem

The Ising problem is a combinatorial optimization problem that traces its origin to solid-state physics. It is of particular significance because its graphical nature allows it to be embedded somewhat naturally in a physical architecture, as exemplified by the popularity of Ising machines as hardware accelerators for combinatorial optimization. The Ising problem, in its most general formulation and in various restricted cases, has been proven to be an NP-hard optimization problem [131]. It has simple polynomial-time reductions to the canonical NP-hard optimization problems; for instance, the traveling salesman problem with N cities can be mapped to an Ising problem with $(N-1)^2$ spins [132]. Therefore, the implementation of a machine that can efficiently solve the Ising problem can be readily adapted to efficiently find the solution to any difficult combinatorial optimization problem.

The Ising problem can be stated as follows: consider a mutually interacting ensemble of N spins, $\{\sigma_1, \sigma_2, \sigma_3, ...\sigma_N\}$, as illustrated in Fig. 5.4(a). We will consider the classical problem in which each spin is binary-valued: $\sigma_i = \pm 1$, which we will call up and down. The spins are coupled via pairwise interactions J_{ij}, which can be collected into a coupling matrix J. A positive-valued interaction acts to align the two spins in parallel, while a negative-valued interaction acts to align them in opposite directions. An interaction that is left unsatisfied by a given spin configuration is considered frustrated.

Given the coupling matrix J, the problem is to find the configuration of spins $\{\sigma_i\}$ that minimizes the Ising Hamiltonian (or Ising energy) H, expressed as:

$$H = -\sum_{\langle i,j \rangle} J_{ij}\sigma_i\sigma_j - \sum_i h_i\sigma_i \tag{5.1}$$

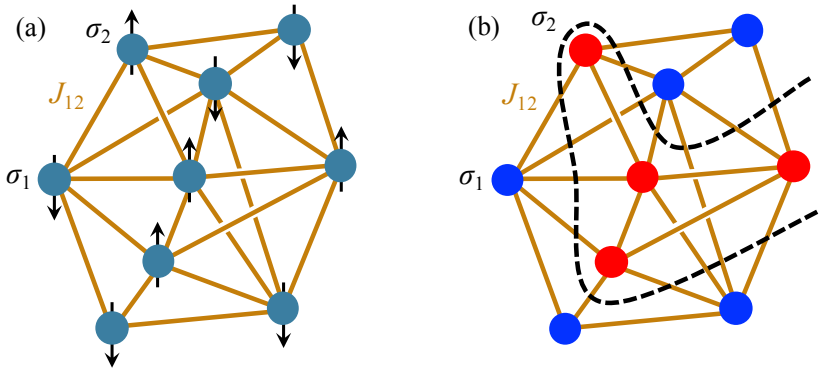

Figure 5.4: (a) An Ising graph is an ensemble of N binary spins $\{\sigma_i = \pm 1\}$ coupled by pairwise interactions, whose weights J_{ij} can be positive (ferromagnetic) or negative (anti-ferromagnetic). (b) The Ising problem is equivalent to the Max-Cut problem, which is to partition the vertices of a graph in such a way that maximizes the weighted sum of the edges between the two sets of vertices.

where the sum $\langle i, j \rangle$ is over all pairs of spins σ_i and σ_j, and h_i is the bias on each spin due to an external field. We will assume an undirected Ising graph with the property $J_{ij} = J_{ji}$. Without loss of generality, the effect of the bias can be absorbed into the coupling matrix J by adding one spin to the network. Thus, we will consider Ising Hamiltonians of the form:

$$H = -\sum_{\langle i,j \rangle} J_{ij}\sigma_i\sigma_j \tag{5.2}$$

This energy function has been used extensively to model the properties of magnetic materials. When the coupling weights J_{ij} are randomly distributed, the effective material that is modeled by the graph is called a spin glass. Except in the simple case of a two-dimensional lattice, the spin glass problem is NP-hard; there is no general algorithm to find the global minimum of H that is intrinsically faster than an exhaustive search over all 2^N possible spin configurations [131]. We note that for an undirected Ising graph, H remains unchanged if every spin is flipped; the number of unique solutions is thus 2^{N-1}.

It is worth noting that the Ising problem has a very close mathematical relationship to the Max-Cut problem, which is depicted in Fig. 5.4(b). Max-Cut is a widely studied graph problem with a well-known approximation algorithm based on semi-definite programming [119]. The figure-of-merit to be maximized in the Max-Cut problem is given by [127]:

$$S_c = -\frac{1}{2}\sum_{\langle i,j \rangle} J_{ij} - \frac{1}{2}H \tag{5.3}$$

Therefore, minimizing the Ising Hamiltonian H maximizes the cut size S_c. Max-Cut problems are commonly used to benchmark Ising machines, as we will discuss in the next chapter. The polynomial-time algorithm in Ref. 119 guarantees a value of S_c that is at least 87.856% of the global maximum. An Ising machine that cannot reliably yield a superior solution to this lower bound is of little practical use.

5.4 Analog precision requirements

In this section, we will investigate the all-important issue of precision, which was a primary factor in the demise of the analog computers of the 20$^{\text{th}}$ century. For any analog machine that is to be used for combinatorial optimization, the sensitivity to imprecision can be deduced from the corresponding sensitivity of an analog Ising machine, owing to the reducibility among NP-hard problems.

The method by which the Ising problem is programmed onto a physical system is, of course, hardware-dependent. In the machine that we propose in Chapter 6, the problem is embedded in the individual component values of a network of resistors that together implement the coupling matrix J. For any analog machine, errors in the programmed component values arising from variability or drift will appear directly as errors in the programmed weight values. For simplicity, we will consider how errors in the weight values δJ_{ij}, rather than the physical component values, propagate to the Hamiltonian that is minimized by the machine. This approach is general, tied neither to any specific scheme for implementing the weight values in hardware, nor to the mechanism by which the machine performs the optimization. In the presence of weight errors, the system seeks to minimize an erroneous Hamiltonian H' that may deviate significantly from the desired problem Hamiltonian H:

$$H' = H + \delta H = -\sum_{\langle i,j \rangle} \left(J_{ij} + \delta J_{ij}\right) \sigma_i \sigma_j \tag{5.4}$$

If the weight errors δJ_{ij} are uncorrelated and randomly distributed, the magnitude of the Hamiltonian error can be shown via the central limit theorem to be $|\delta H| = \delta J \sqrt{N_{\text{w}}}$, where N_{w} is the number of weights in the problem and δJ is the RMS size of the weight errors.

An Ising machine can tolerate small errors in the problem implementation if there is a lower bound on the energy gap H_{gap} between the global minimum and the next lowest value of the desired problem Hamiltonian. In this scenario, the problem Hamiltonian and the perturbed Hamiltonian can be minimized by the same ground state spin configuration, so long as the error δH is smaller than the minimum possible energy gap. This condition is illustrated in Fig. 5.5(a). Once the problem becomes large enough or the weight precision is poor enough such that $|\delta H| > \frac{1}{2} H_{\text{gap}}$, it becomes probable for the machine to mistake the first excited state for the ground state, leading to an incorrect solution even with an otherwise ideal machine for global optimization.

If the problem specification is fully analog – i.e. the values in the coupling matrix are continuous-valued – we can expect that the energy gap H_{gap} has no lower bound and that the machine has a very low tolerance for implementation error even in small problems. We must therefore provide built-in robustness to imprecision by constraining the weight values in the problem to be discrete, with a minimum difference of J_{step} between any two weights. It is then evident from Equation (5.2) that the energy gap has a lower bound equal to J_{step}. We therefore obtain a relationship between the problem size, the implementation error, and

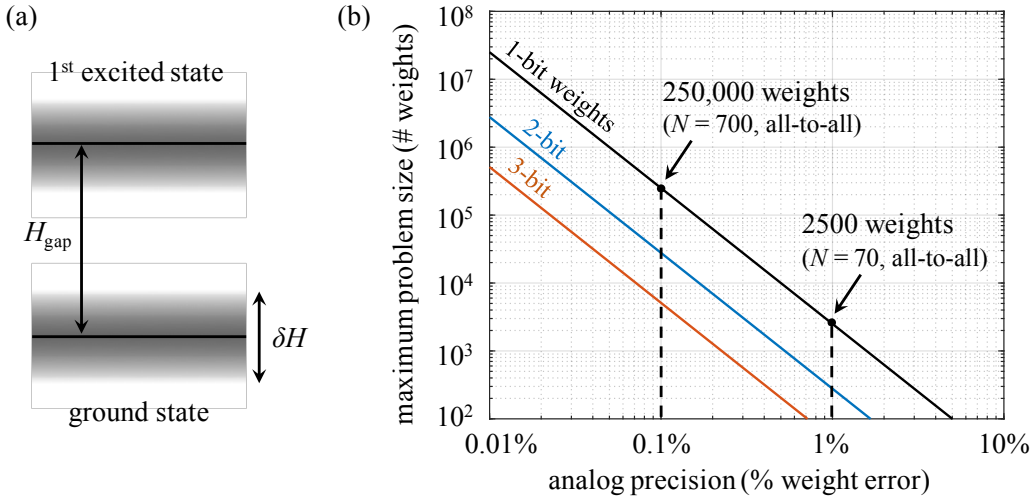

Figure 5.5: Errors in the implementation of coupling weights translate into errors in the Hamiltonian δH, shown in (a). When δH is large enough to allow the first excited state of the (error-free) Hamiltonian to be lower in energy than the ground state, the machine minimizes the wrong Hamiltonian. (b) Maximum problem size that can be embedded in our machine, in terms of the number of nonzero weights in the coupling matrix J, versus the analog precision in the weights.

the size of the Hamiltonian error:

$$\frac{|\delta H|}{H_{\text{gap}}} = \frac{\delta J}{J_{\text{step}}}\sqrt{N_{\text{w}}} \tag{5.5}$$

Setting $|\delta H| > \frac{1}{2}H_{\text{gap}}$ as the condition for failure, we can solve the above relationship for the maximum allowed number of weights that can be reliably implemented:

$$N_{\text{w,max}} = \frac{1}{4\left(2^{N_{\text{b}}}-1\right)^2\left(\delta J'\right)^2} \tag{5.6}$$

where $\delta J'$ is the weight error normalized to the total weight range $(J_{\max} - J_{\min})$, and N_{b} is the number of bits in the weight's digital representation. The former quantity, in percentage units, can be used as a measure of the implementation precision. Fig. 5.5(b) shows the maximum size of a problem, in terms of the number of weights, that can be solved with a given implementation precision. With a realistic analog component precision of 1%, a 2500-weight problem (70 fully-connected spins) with binary weights can be reliably implemented. For a more aggressive target of 0.1% precision, this increases to 250,000 weights (700 fully-connected spins) – a much more difficult problem. The maximum problem size decreases significantly as additional weight values are added, decreasing the energy gap H_{gap}. These results are hardware-agnostic and apply to any physical scheme for implementing the Ising problem. Similar conclusions have been reached, for example, in quantum computing. Ref. 133 has proposed the use of multiple redundant, ferromagnetically coupled Ising graphs to provide robustness to random errors in implementation.

Chapter 6

Dissipative coupled oscillator circuits as analog Ising machines

In this chapter, we present the design of an Ising machine, which solves the Ising optimization problem described in Section 5.3. When programmed to implement a given instance of the Ising problem, the machine stabilizes to a physical configuration that represents an optimum – ideally the ground state – of the Ising spin glass.

Our construction is a network of coupled phase-bistable electrical oscillators. We draw inspiration for this design from the oscillator-based Ising machine proposed in Ref. 105, which was shown to achieve good performance on the Ising problem and whose results can be generalized to any network of nonlinear oscillators that are self-sustaining and injection-locked. The authors also simulated and physically realized an Ising machine using electronic LC oscillators [105], [106]. Our electrical Ising machine uses a similar network topology and also exploits phase bistability. As a significant operational difference from this prior work, our machine does not assume that the oscillators are self-sustaining. Our system's search mechanism is therefore governed not only by the phase dynamics but also the amplitude dynamics of the oscillators. This stems from our aspiration to realize the first-to-threshold optimization principle described in Section 5.2, which relies heavily on the competition between gain and dissipative processes to increase the oscillators' amplitudes from the noise floor. In this sense, our system has similarities to the coherent Ising machine, implemented using optical parametric oscillators [107], [108], [127], [128]. In Section 6.3, we will discuss the relationship between our system and these closely related Ising machines in more detail. Though we find that our machine does not fully embody the desired first-to-threshold principle in its dynamics, it nonetheless possesses an unusual way of exploring the solution space that is often successful in leading the system to an optimal solution.

In Sections 6.1 and 6.2, we describe the architecture of our Ising machine and show how it can be programmed to represent any specific instance of the Ising problem. In Section 6.4, we demonstrate in simulation the performance of our analog oscillator circuit on Ising problems of various sizes. We present the theory of our machine's dynamics in Section 6.5, elucidating many of the operational features that are observed in the simulation results

and clarify precisely how the digital optimization problem is embedded in the machine's continuous-time dynamics. Finally, in Section 6.6, we will discuss the trade-offs that are encountered in the design of the machine as well as the practical considerations which – along with the issue of analog precision discussed in Section 5.4 – ultimately limit how far the machine can be scaled to accommodate large optimization problems.

6.1 Nonlinear reactive circuits as digital Ising spins

As a first step in our Ising machine implementation, we consider the representation of isolated binary spins using bistable electrical elements. We can draw upon a number of nonlinear phenomena in physics to realize bistability, one of which is parametric oscillation. When some parameter of a physical system is modulated by the influence of an external pump, a normally linear harmonic oscillator can be induced to become bistable in phase.[1] This effect, known as degenerate parametric oscillation, can occur in nonlinear oscillators of any type and has also been used to implement binary Ising spins in the optical [107] and electromechanical [134] domains. In the following, we will briefly review how phase bistability arises in a simple LC oscillator in the presence of a small nonlinearity in the circuit. These types of parametric oscillators were first proposed in the late 19$^{\text{th}}$ century [135] and were further developed in the 1950s [136] with the formulation of the Manley-Rowe relations [137]. Parametrically pumped LC oscillators, owing to their bistability and their relative ease of implementation, have also been exploited as computational elements for Boolean logic circuits that act on the phases of the oscillators [138]–[140].

Fig. 6.1 shows an LC oscillator containing a nonlinear capacitor. When the circuit is driven by a pump that modulates the nonlinear capacitor, a stable oscillation can develop, seeded by a small source of energy (such as thermal noise) inside the circuit. If we retain only the first-order nonlinear term in the capacitor, its charge-voltage relationship can be expressed as:

$$Q = (C_0 + a_1 V)\, V \tag{6.1}$$

where C_0 is the linear capacitance and a_1 is the strength of the first-order nonlinearity. The nonlinearity can be implemented using any voltage-dependent capacitor, such as a reverse-biased p-n junction. Suppose that the voltage on the capacitor can be decomposed into a component at the oscillator's fundamental frequency $\omega = \sqrt{LC_0}$ and a voltage that is supplied by an external pump circuit (not shown in Fig. 6.1) at the frequency ω_p:

$$V(t) = V_s(t) + V_p(t) = A(t)\cos(\omega t + \phi) + B(t)\sin(\omega t + \phi) + V_{p0}\cos(\omega_p t) \tag{6.2}$$

where $A(t)$ and $B(t)$ are the time-varying amplitudes of the oscillations at ω, ϕ is a fixed phase (described below), and V_{p0} is the pump amplitude. For now, we will assume that the

[1] In AC circuits, for example, the modulation is typically in the capacitance or the inductance. In optical systems and in mechanical systems, the modulated parameters are often the electric susceptibility and the mechanical stiffness, respectively, of the material that supports the oscillation.

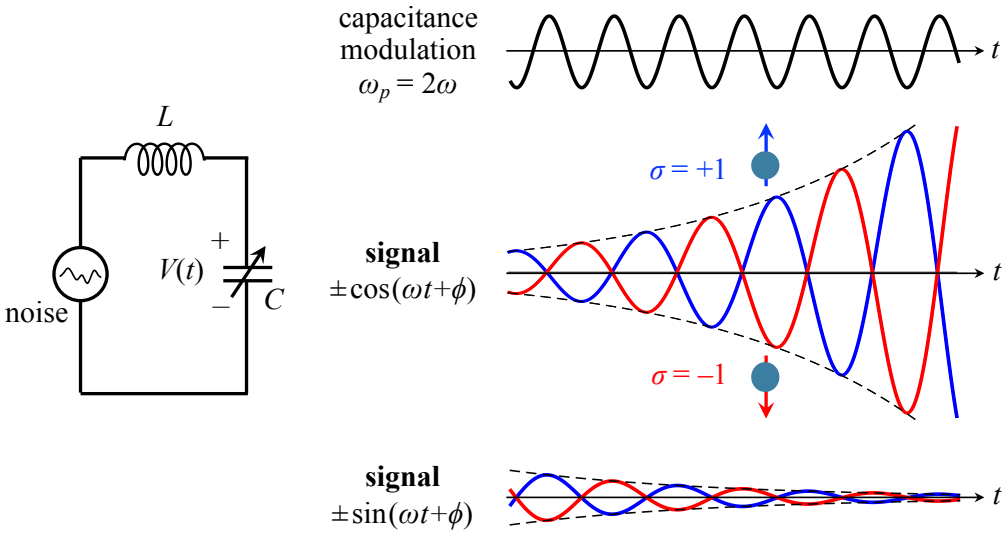

Figure 6.1: Degenerate parametric oscillation in an LC circuit with a nonlinear capacitance (left). When a pump signal at twice the oscillator's fundamental frequency is applied to the capacitor, the oscillation experiences phase-sensitive amplification: the cosine component is amplified, while the sine component is de-amplified. The result is a phase-bistable oscillation with a difference of π radians between the two phases, which can encode a binary-valued Ising spin.

pump amplitude is fixed and remains much larger than the signal amplitude: $V_{p0} \gg A, B$. In this regime, it follows from Equation (6.1) the the circuit's response at ω can be equivalently modeled using a sinusoidally modulated capacitance:

$$C(t) = C_0 + a_1 V_{p0} \cos(\omega_p t) = C_0 + \Delta C \cos(\omega_p t) \tag{6.3}$$

where ΔC is the modulation depth. Increasing (or decreasing) the capacitance involves work that must be done on (or by) the capacitor's internal electric field. Therefore, modulation of the capacitance inherently implies that an exchange of energy occurs between the pump and the oscillator.

Consider the case where the capacitance is modulated at twice the fundamental frequency of the circuit: $\omega_p = 2\omega$, shown in Fig. 6.1. Let us qualitatively examine how this modulation influences an oscillation at the frequency ω. If the capacitance falls, the voltage receives a boost, due to the basic relationship $V = Q/C$. The maximum boost is given to a signal whose amplitude peaks when the decrease in capacitance is fastest, while a signal that differs in phase by a quarter-cycle is close to zero at that moment and thus receives the minimum boost. Half a pump cycle (or a quarter of a signal cycle) later, the capacitance rises and the voltage drops. The same signal that had previously received the maximum boost is now close to zero, and thus suffers a negligible loss in amplitude; meanwhile, the signal that received the minimum boost now experiences the maximum drop. Over many cycles, one phase of the oscillation is amplified by the capacitance modulation, while the phase that

is a quarter-cycle ($\pm\pi/2$ radians) away is attenuated, as shown in Fig. 6.1. If we choose the value of ϕ in Equation (6.2) such that the waveform $\cos(\omega t + \phi)$ receives the maximum amplification from the pump, the amplitude A increases while the amplitude B decays to zero. This phase-sensitive effect is known as degenerate parametric amplification [136].

Amplification of the signal at ω continues until the power in the signal exceeds the power supplied by the external pump. When this occurs, the signal returns part of its power to the pump and decreases in amplitude until it is again weaker than the pump. This back-and-forth exchange of energy between the pump and the signal eventually stabilizes the signal amplitude. We note that this saturation effect is not captured by Equation (6.3), since the signal amplitude in this regime is no longer negligible in comparison to the pump.

If a given phase ϕ of the oscillation experiences amplification, the phase $\phi + \pi$ is also amplified, since the two phases react in an identical manner to the second-harmonic pump. Since these oscillations are opposite in phase, they cannot co-exist: the LC oscillator must choose one of the two phases (blue and red in Fig. 6.1) and the AC voltage $V(t)$ on the capacitor is phase-bistable. The two possible phases of the voltage oscillation can therefore encode a logical bit [138] or, equivalently, the two values of the Ising spin, $\sigma = \pm 1$. Which phase does it choose? In the isolated oscillator shown in Fig. 6.1, the two phases are equally favored and one is randomly selected by noise. When the oscillator is connected to an external circuit, it can be steered toward one phase over the other.

6.2 The dissipative coupled network of oscillators

Since the spin orientation of any one oscillator is determined by its connectivity to the external circuit, the coupling network among the oscillators is the physical embodiment of the Ising Hamiltonian. In order to implement the first-to-threshold search method proposed in Section 5.2, it is essential that the LC oscillators are coupled together by lossy elements. To this end, we use the resistive network shown in Fig. 6.2.

Fig. 6.2(a) and (b) show the implementation of a positive ($J_{ij} > 0$) and negative ($J_{ij} < 0$) interaction weight, respectively. These circuits, which use straight- and cross-linking resistors, are similar to that used in the oscillator-based Ising machine in Ref. 105, and are also similar to the buffer and NOT gates used in Goto's scheme for oscillator phase-based Boolean logic [138]. The dissipative nature of the connections allows the spin-spin coupling to be implemented in a manner that is compatible with the first-to-threshold mechanism that we propose to implement. In the straight-linking connection in Fig. 6.2(a), less power is dissipated in the resistors if the AC oscillations $V_i(t)$ and $V_j(t)$ are in phase (aligned spins), while the maximum amount of power is dissipated if the oscillators are π radians out of phase (opposite spins). In the cross-linking connection in Fig. 6.2(b), the opposite is true. The degree to which these connections are dissipative matters significantly for the machine's operation, as we will discuss in more detail in Section 6.6. Two oscillators with no coupling, as in Fig. 6.2(c), can implement a zero weight.

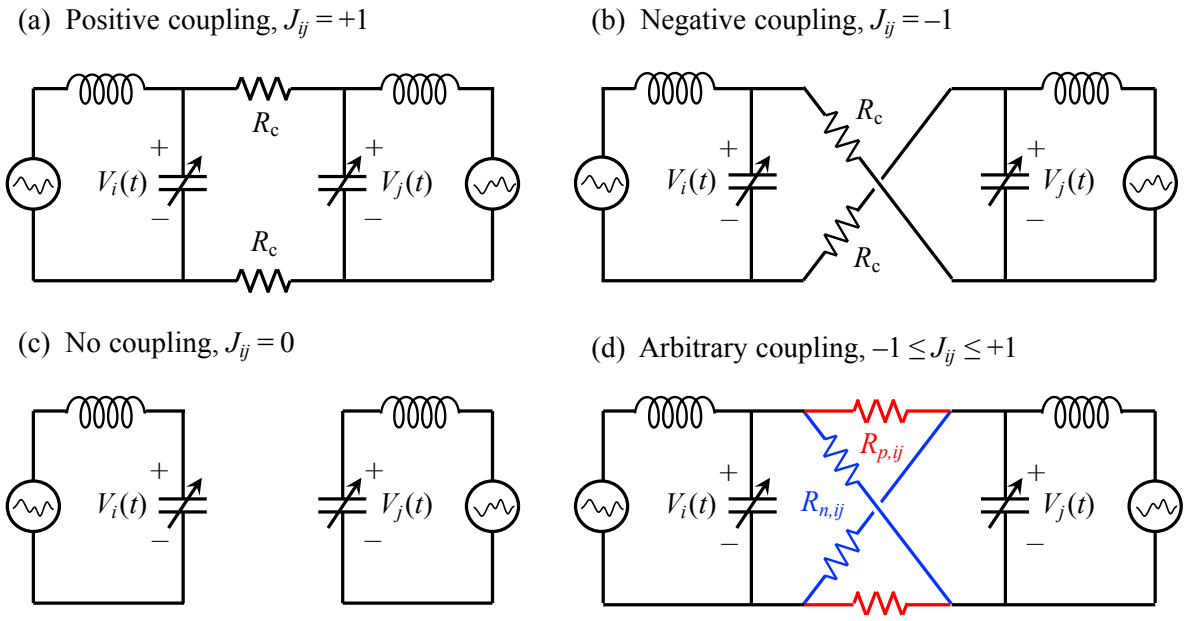

Figure 6.2: Scheme for resistively coupling two phase-bistable LC oscillators i and j. The straight-linking resistors in (a) implement a positive interaction $J_{ij} = +1$, while the cross-linking resistors in (b) implement a negative interaction of the same strength, $J_{ij} = -1$. Two unconnected oscillators, as in (c), have no direct interaction, $J_{ij} = 0$. An arbitrary real-valued weight $-1 \le J_{ij} \le +1$ can be implemented using four resistors with two different resistance values as shown in (d).

The two circuits in Fig. 6.2(a) and Fig. 6.2(b), using the same value for the coupling resistance R_c, implement two values of the interaction weight with the same magnitude: we can assign these to $J_{ij} = +1$ and $J_{ij} = -1$, respectively. An Ising problem that can be represented using only these two weight values (and $J_{ij} = 0$) requires at most two resistors per connection and a fixed resistance R_c. To program different problems of this type, the resistive network coupling each pair of oscillators can be switched among the circuits in Fig. 6.2(a)-(c). This scheme is sufficient, for example, to embed the unweighted Max-Cut problem, which has been shown to be NP-hard [96].

If we allow for multiple discrete (or continuously varying) values of the coupling resistance, we can represent multiple (or continuous) values of the coupling weights. A network that implements an arbitrary positive or negative weight, normalized to the range $-1 \le J_{ij} \le +1$, is shown in Fig. 6.2(d) and comprises two straight-linking resistors $R_{p,ij}$ and two cross-linking resistors $R_{n,ij}$. The values of these resistors are chosen as follows:

$$R_{p,ij} = 2R_c \left(1 + J_{ij}\right)^{-1} \tag{6.4a}$$

$$R_{n,ij} = 2R_c \left(1 - J_{ij}\right)^{-1} \tag{6.4b}$$

These equations imply that if R_c is a fixed parameter for the network, the conductance $1/R_{p,ij} + 1/R_{n,ij} = 1/R_c$ is the same for all pairs of oscillators. For the extreme values of

$J_{ij} = \pm 1$, one pair of resistors becomes open and the network reduces to Fig. 6.2(a) or (b). By choosing four equal resistors, we can couple the oscillators in an unbiased manner to implement $J_{ij} = 0$: this differs from the situation in Fig. 6.2(c), where the connection between the two oscillators has been severed. We note that the method of choosing four well-balanced resistors is more prone to variability than the implementation in Fig. 6.2(c), which introduces no weight error. Therefore, representing a zero weight using the latter scheme may be more advantageous in terms of scalability and energy efficiency, though the circuit's dynamics becomes harder to predict, as we will see in Section 6.5.

To see how these resistance values can map to any value of J_{ij}, we will make the assumption that every oscillator is self-sustaining with the same amplitude, and that all share a common pump so that they differ in phase only by 0 or π radians. This means that the voltage on every oscillator has the form $V_i(t) = \sigma_i V_0 \cos(\omega t + \phi)$, where V_0 is the uniform amplitude. In general, these conditions are *not* satisfied. However, if we proceed with these assumptions, the total power dissipation in the network is given by:

$$\begin{aligned} P_d &= \frac{1}{2}\sum_{\langle i,j\rangle}\left(\frac{V_0^2}{R_c} - \left(\frac{1}{R_{p,ij}} - \frac{1}{R_{n,ij}}\right)V_0^2\sigma_i\sigma_j\right) = P_{d0} - \frac{V_0^2}{2R_c}\sum_{\langle i,j\rangle} J_{ij}\sigma_i\sigma_j \\ &= P_{d0} + \frac{V_0^2}{2R_c}H \end{aligned} \tag{6.5}$$

where the power dissipation is averaged over one oscillation cycle, and P_{d0} is a constant. Note that under the restrictive conditions that we have imposed, the power dissipation maps directly onto the Ising Hamiltonian H.

We note, however, that the assumption of a uniform amplitude distribution is a vast over-simplification. There is no guarantee that the oscillators will ever settle to such a state, particularly if there are large voltage drops across the coupling resistors. This issue is identical to the problem of amplitude heterogeneity that has been discussed in the context of the coherent Ising machine [129]. Nonetheless, even with this fact we will select the resistance values according to Equations (6.4). While we cannot reply upon the mapping in Equation (6.5), we will find that this choice of resistors leads to a more subtle kind of mapping to the Ising Hamiltonian that holds in general, and not only in the saturated steady-state. A detailed dynamical analysis of the Ising machine will be presented in Section 6.5, where we will explore in a more precise manner how the Ising Hamiltonian is embedded in our circuit.

To perform a first-to-threshold search as described in Section 5.2, we initialize the system at rest; there are no self-sustaining LC oscillations, and only noise is present. We introduce gain through a second-harmonic pump that modulates the nonlinear capacitor, whose effect at the oscillators' fundamental frequency can be represented as a phase-sensitive negative resistance that simultaneously leads to amplitude gain and bistability in phase. Threshold is reached when the negative resistance supplied by the pump matches the positive resistance that is present in the coupling between the oscillators – in other words, when the power supplied parametrically meets the power lost by dissipation. At that point, phase-bistable

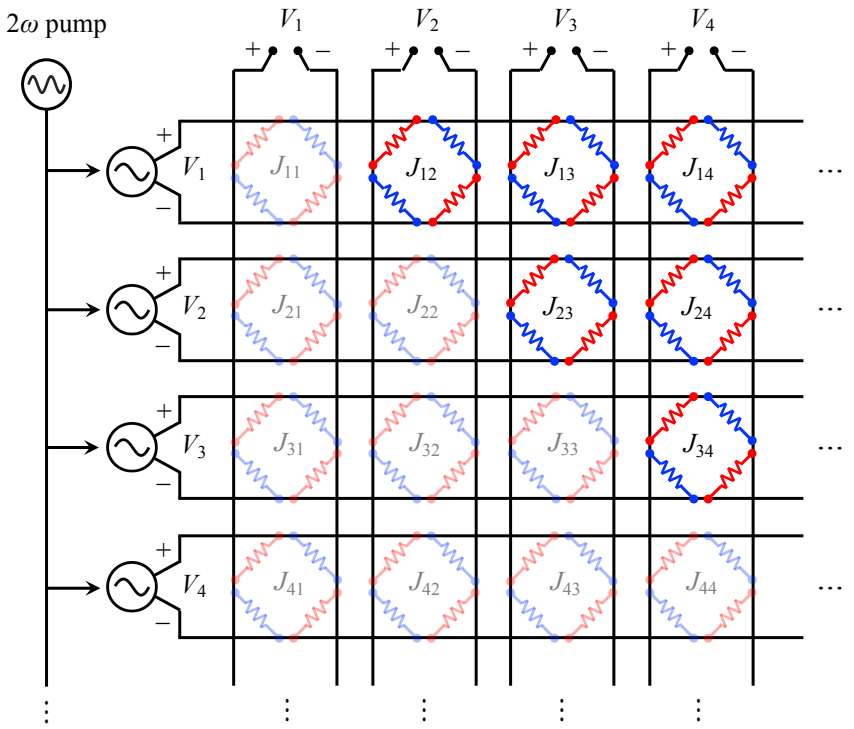

Figure 6.3: A crossbar array implementation of the analog electrical Ising machine, which can be used to solve fully-connected Ising problems. The oscillator voltages are copied onto the rows and columns of the array, while the resistors at the array crosspoints implement the interaction weights (red for positive, blue for negative as in Fig. 6.2(d)). Any unweighted Ising problem can be embedded in the array using the solid-colored coupling resistors. Weighted Ising graphs can also be implemented using both the solid-colored and faded resistors.

oscillations can develop from noise; the first mode to oscillate ideally represents the global minimum of the Ising problem.

Using the general scheme in Fig. 6.2(d), a new Ising problem can be programmed onto the network by using variable resistors, such as flash resistors or memristors. Since an Ising problem can be fully specified by a coupling matrix J, it is natural to represent the graph using a crossbar array architecture, similar to that shown in Fig. 6.3, which provides dense connectivity. This type of architecture is also used for dense data storage [141] and for vector-matrix multiplication, particularly in accelerating neural network computations [142]. The oscillator output voltages can be applied to the rows and columns of the crossbar array, while the coupling weights J_{ij} are placed at the array crosspoints and connect the oscillator terminals in the manner shown in Fig. 6.2(d). To implement a fully-connected Ising problem, it is necessary only to use the upper-triangular portion of the crossbar array (solid-colored resistors in Fig. 6.3). The remainder of the array can be used to embed a directed graph ($J_{ij} \neq J_{ji}$), which might be a more efficient representation for certain optimization problems.

If memristors or flash resistors are used to implement the coupling resistors, the interaction weights can be re-programmed by applying current or voltage pulses to the appropriate oscillator terminals, as typically done in resistive crossbars for data write and neural network training operations. In Section 6.6, we discuss some of the practical challenges related to technical implementation as we scale the Ising machine to large problem sizes.

6.3 Related Ising machines

There are a number of other Ising machine proposals that make use of coupled oscillator networks. In this section, we compare our proposed design to two closely related machines: the oscillator-based Ising machine in Refs. 105 and 106 and the coherent Ising machine, based on optical parametric oscillators and first proposed in Ref. 107.

As explained in Ref. 105, sub-harmonic injection locking is observed across a broad range of oscillator technologies, any of which can be chosen to implement an Ising machine. In particular, an implementation using LC oscillators was proposed and demonstrated in Refs. 105 and 106. In these works, the individual oscillators are self-sustaining and their phase dynamics in the presence of sub-harmonic injection locking was analyzed. It was shown that as injection locking takes hold and the oscillator phases are binarized, the phases collectively stabilize onto configurations that represent local minima of the Ising Hamiltonian H. The time to reach this minimum is found to be independent of problem size. This local search was combined with a source of noise to perform a probabilistic search, with the oscillators always remaining in self-sustaining mode, to obtain good performance on the Ising problem.

While the structure of our coupling network is similar to the LC oscillator implementation in Refs. 105-106, the physical dynamics of our machine is distinct. Since we propose to implement the first-to-threshold mechanism of optimization, we initialize all of the oscillators at rest and allow the competition between gain and dissipative processes to perform a search over the solution space. This means that the oscillators are not self-sustaining from the start, and both their amplitudes and phases evolve as the machine converges upon a solution. Unlike the oscillator-based Ising machine in Ref. 105, the amplitude dynamics is very important, as we show in Section 6.5. We do find, as in Ref. 105, that the system settles to a solution within a timescale that is independent of problem size. Furthermore, if implemented exactly, the first-to-threshold method offers a means toward a targeted search for the global minimum of the optimization problem; for our machine, however, we find that this has not yet been realized.

As we have previously noted, the coherent Ising machine [107], [108], [127], [128] was also proposed to operate by the first-to-threshold computational principle, using degenerate optical parametric oscillators that propagate around a very long fiber-ring cavity. As the name suggests, these oscillators encode the two spin states through the same nonlinear mechanism as our electrical oscillator, which is fully classical. An important difference, however, is that in the coherent Ising machine, the oscillators are coupled using a time-

multiplexed approach which ensures that the spins only communicate with one another at discrete time steps [108]. In the large-scale implementations of the coherent Ising machine, the oscillator states are discretely sampled and coupling is achieved digitally using a field-programmable gate array (FPGA) [127], [128] – outside of discrete moments of interaction, the spins operate independently. This is to be contrasted with our analog Ising machine, in which every oscillator communicates its state to every other oscillator in the network in continuous time. This inherent parallelism of analog systems, mentioned in other works [106], [112], may have significant operational advantages. We note that a benefit of implementing Ising spins using time-multiplexed oscillators, as opposed to a dedicated oscillator circuit for each spin, is the automatic guarantee of near-identical oscillators generated by the same hardware – in the case of the coherent Ising machine, a single nonlinear $LiNbO_3$ crystal. However, we will see in Section 6.6 that our system does possess some operational robustness to variability in the oscillator components.

More recent theoretical work on the coherent Ising machine concept has focused on the continuous-time dynamical analysis of the underlying computational principle [129], [130], [143]. We find that the equations of motion described in these works have some resemblance to the dynamical equations for our Ising machine, which we will derive in Section 6.5. However, to our knowledge this type of dynamics has not been directly embodied in any present physical implementation. Our electrical oscillator system, if indeed it is described by similar dynamics, is a working example of this specific computational principle, though to date it has not yet been demonstrated outside of a circuit simulation.

There are a number of other Ising machine proposals that are less closely related in their mechanism to our system. Machines that accelerate simulated annealing have been demonstrated on several different platforms, including SRAM cells connected by logic gates [104] and neural networks on memristor crossbars, which exploit the inherent noise of memristive devices [110]. Networks of devices that individually act as invertible logic elements have also been mapped onto combinatorial optimization problems like the Ising problem: implementations include stochastic nanomagnets [111] and deterministic memristors [112]. A direct performance comparison between our system and these various existing approaches will be possible as we scale up our benchmarking efforts, described in the next section.

6.4 Simulated performance

In this section, we will numerically demonstrate the operation of our analog Ising machine and discuss the intrinsic scaling properties of our search method. We will shed light on many of the features that we observe in these numerical results in the next section, where we more closely examine the machine's principle of operation from a theoretical approach.

Our numerical results are obtained from a transient simulation of the Ising machine circuit in LTspice. Each nonlinear oscillator is modeled by the circuit in Fig. 6.4 and contains an inductance of $L = 1.0\,\mathrm{nH}/2\pi$ and a linear capacitance of $C_0 = 1.0\,\mathrm{nF}/2\pi$, resulting in an oscillation frequency of $\omega = 2\pi \times 1.0$ GHz. We also allow for oscillator internal loss through

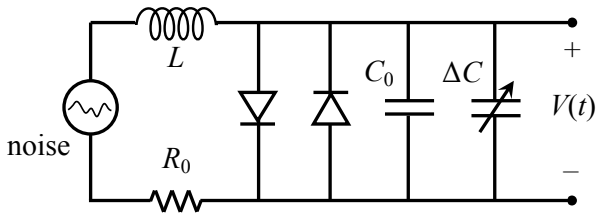

Figure 6.4: The nonlinear LC oscillator model used in the SPICE simulation. The two ideal diodes clamp the oscillation amplitude to their turn-on voltage, modeling amplitude saturation under parametric pumping. The output voltage appears across the capacitor.

the resistor R_0. The LC oscillators are coupled together using the resistive coupling scheme shown in Fig. 6.2(d), with the resistances chosen according to Equations (6.4). We use a coupling resistance of $R_c = 200\,\Omega$: the implications of this choice on the machine's operation will be discussed later in this section. The simulation ignores any phase delays associated with the distribution of the pump or the individual oscillator signals across the network.

To accelerate the simulation of a large-scale network of oscillators, we model the degenerate parametric amplification using a sinusoidal capacitance at 2.0 GHz – which implements the second term in Equation (6.3) – without explicitly including a model for the external pump circuit. We allow the amplitude ΔC to change with time to model a gain ramp. We will refer to the linear modulation depth $\Delta C/C_0$ as the "gain." We model the saturation of the oscillator amplitudes by adding two ideal (step function) diodes in parallel with the nonlinear capacitor. The diodes conduct no current when the oscillation amplitude is small, but once the output exceeds the diode turn-on voltage, all of the current is conducted through the diodes, clamping the peak-to-peak voltage amplitude.

An exact implementation of the first-to-threshold scheme would require ramping the gain from zero to a value that is not known a priori. We will see in the next section, however, that for our machine it is not necessarily advantageous to ramp the gain from zero. By choosing a value of the gain that is higher than the threshold value, we are still likely to excite the mode with the least loss while also enabling a broader search of the solution space. This does, however, mean that our system does not exactly implement the first-to-threshold mechanism described in Section 5.2; we will revisit this important point in the next section. For the circuit simulations shown in this section, we adopt both a linear gain ramp as well as a constant gain, with similar performance results. We choose the values of gain to be close to the threshold predicted by Equation (6.17), which we will derive in the next section. In general, for the same oscillation frequency and coupling resistance, the gain must increase linearly with the number of spins.

We begin by presenting the simulated performance of our machine on the simple, fully-connected Ising problem shown in Fig. 6.5(a) with eight spins. We choose the weights to be random and binary ($J_{ij} = \pm 1$). The full coupling matrix and the LTspice netlist for this example is provided in Appendix G. For this problem, the gain $\Delta C/C_0$ is ramped from a value of 2.5% to 3.5% over a duration of 1.0 μs.

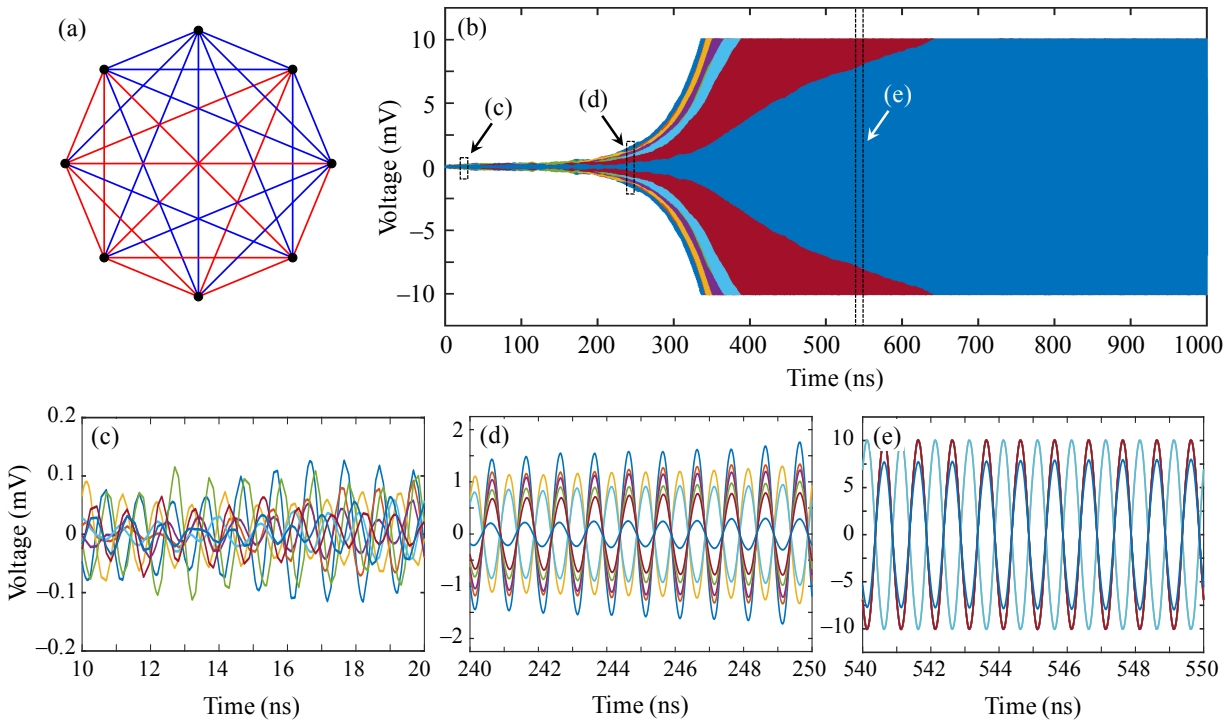

Figure 6.5: Oscillator output voltage waveforms generated by an LTspice transient simulation of an Ising machine circuit that solves the 8-spin problem shown in (a) with random binary-valued weights: red edges represent $J_{ij} = +1$, and blue edges represent $J_{ij} = -1$. Each color represents one of the eight LC oscillators. The waveforms are shown for (b) the full simulation and at various stages of the search in (c) – (e).

Fig. 6.5(b) shows the oscillator output voltage waveforms returned by the LTspice simulation, each oscillator represented by a different color. Since the circuit is initialized at rest with noise (with an RMS amplitude of 20 μV) as the initial condition, the oscillator waveforms at early times are noisy and mutually incoherent; this can be seen in the close-up view in Fig. 6.5(c). As the oscillators experience parametric gain, their amplitudes rise gradually above the noise floor. At the same time, because the amplification is phase-sensitive, the oscillators organize into two groups whose phases are separated by π radians. This is seen at $t = 240$ ns – see Fig. 6.5(d) – when the system has become bistable in phase while the amplitudes continue to increase. Beyond this point, the oscillator amplitudes saturate one by one to the limit of 10 mV that is set by the diode turn-on voltage. By $t = 540$ ns, shown in Fig. 6.5(e), the system is close to fully amplitude-stable as well as phase-bistable.

To obtain more useful insight into the circuit's operation, we extract the time-dependent oscillation magnitudes and phases for each of the eight oscillators. For both of these quantities, we use a moving window of 10 ns width to sample the oscillator voltage waveforms in 0.5 ns time steps. We obtain the magnitude by finding the maximum absolute voltage within the window. To find the phase, we first generate a reference sinusoidal waveform at 1.0 GHz

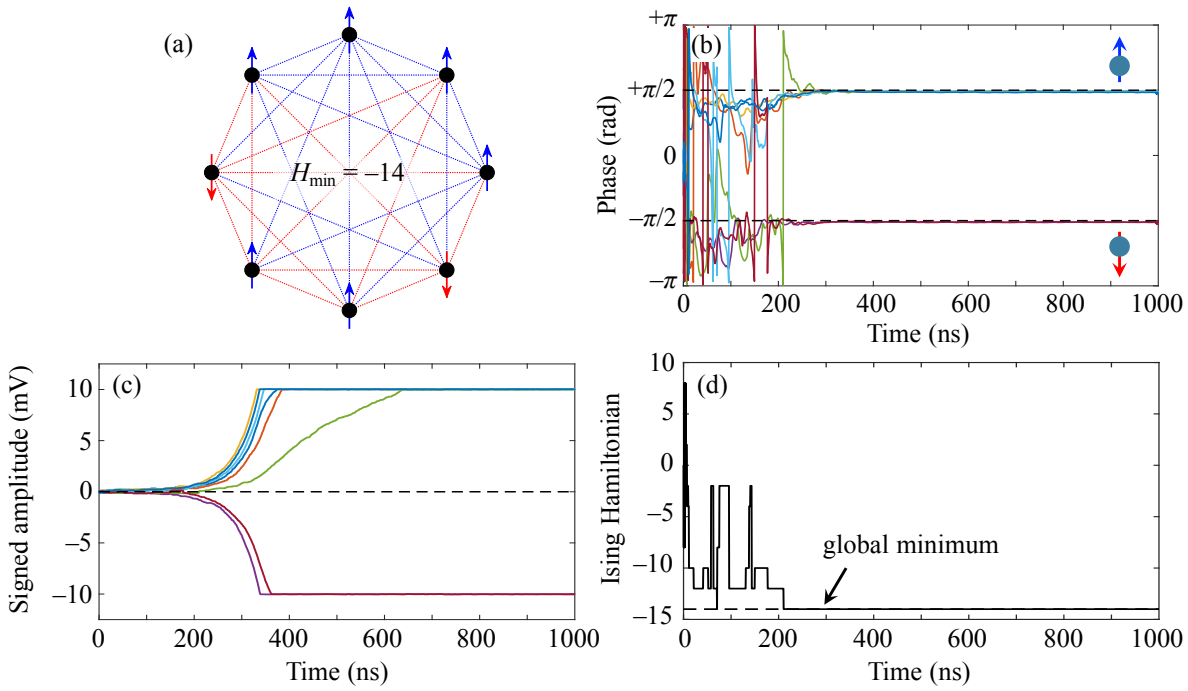

Figure 6.6: The (a) signed amplitudes and (b) phases of the eight oscillators, extracted from the voltage waveforms in Fig. 6.5(b). The oscillator phases fully bifurcate by ∼200 ns. We can calculate an Ising energy H from the equivalent spin configuration represented by the phase-bistable oscillators, shown in (c). The network settles within 220 ns to the global minimum of the 8-spin Ising problem, shown in (d).

whose phase is matched to that of oscillator 1 at the end of the simulation. At each time step, we evaluate the phase of the windowed oscillator waveform's Fourier component at 1.0 GHz. This phase is compared to that of the reference sinusoid over the same window. If the difference is zero, the waveform has a phase equal to that of oscillator 1 at the end of the simulation, which we arbitrarily assign to the value $+\pi/2$ (spin up); if the difference is π, the waveform is out of phase with oscillator 1 and has an absolute phase of $-\pi/2$ (spin down). These are plotted in Fig. 6.6(b). Early in the simulation, when the oscillation magnitudes are close to the noise floor, the phases fluctuate randomly and are essentially meaningless. By $t = 220$ ns, when the oscillation signal to noise ratio is large, all of the oscillators have locked onto one of the two bistable phases.

We can further use the value of the phase to obtain a *signed* amplitude for each oscillator; we simply multiply the oscillation magnitude by $+1$ if the phase is positive (closer to $+\pi/2$) and -1 if the phase is negative (closer to $-\pi/2$). These are plotted in Fig. 6.6(c). Of course, the signed amplitudes are only truly meaningful after the phase values have settled to $\pm\pi/2$. Nonetheless, this quantity provides a convenient, though approximate, way to visualize both the phase and amplitude dynamics of the system.

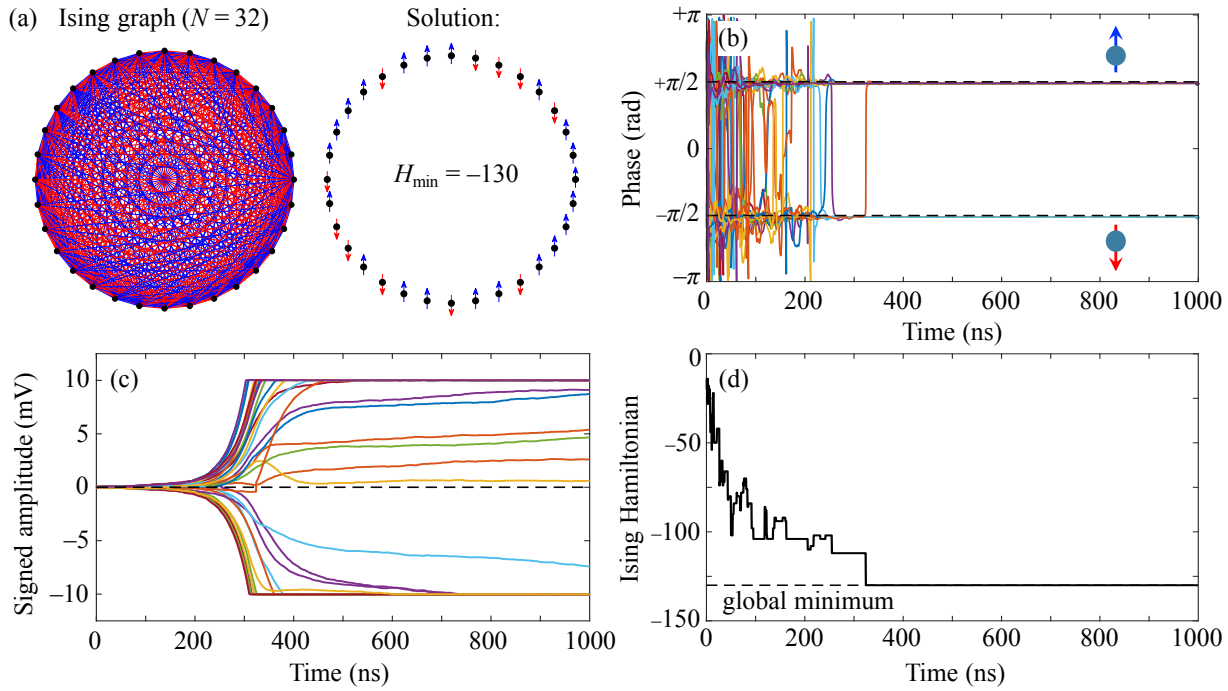

Figure 6.7: Circuit simulation results for the analog Ising machine with 32 coupled LC oscillators, which solves the fully-connected Ising problem depicted in (a). The phases and signed amplitudes of the 32 oscillators are shown in (b) and (c), respectively, each color representing a different oscillator. (d) shows the Ising energy H of the equivalent spin configuration. The network settles to the global minimum of the Ising problem within 330 ns.

We can assign a spin configuration to the collective state of the oscillators once phase bistability is established over the full network. Fig. 6.6(d) shows the evolution of the Ising Hamiltonian corresponding to these spin configurations. Like the signed amplitude, this quantity is meaningful only after every oscillator has locked to one of the two bistable phases. For this problem, the system finds and ultimately settles to a configuration with $H = -14$, shown in Fig. 6.6(a). To verify the machine's usefulness as an optimizer, we independently solved the problem by sequentially calculating the energy of all possible solutions – a trivial task in this case – and found the same solution. The circuit discovers the correct solution within 220 ns of the simulation's start, and maintains this phase distribution, remaining at the global minimum, as the gain is ramped further.

Having demonstrated the basic features of the LTspice simulation, we now test our Ising machine's performance on two larger problems. The LTspice netlist files and the coupling matrix J for both of these examples can be found at a Git repository: `https://github.com/ptxiao/analogIsing`. The resulting LTspice voltage waveforms are shown in Appendix G. Here, we will show only the evolution of the amplitude and phase distributions, extracted as described above from the voltage waveforms.

Fig. 6.7(a) shows a moderately sized Ising graph with 32 fully-connected spins and 496 random binary weights ($J_{ij} = \pm 1$). The binary weight constraint does not change the NP-hardness of the Ising problem, and as we showed in Section 5.4, some discretization of the coupling weights is desired in practice to ensure an accurate representation of the optimization problem. For this problem size, there are 2^{32} possible spin configurations, half of which are degenerate due to the symmetry of the coupling matrix: thus, there are $2^{31} \approx 2.15 \times 10^9$ candidate spin configurations. As before, we independently found the global minimum by an exhaustive search; this solution is shown in Fig. 6.7(a). This global minimum is non-degenerate.

For this larger circuit, we linearly increase the gain from 12% to 13% over a duration of 1.0 µs. Fig. 6.7(b) and (c) show the phases and the signed amplitudes of the 32 individual oscillators. Compared to the previous example, it is more conspicuous here that the oscillation amplitudes do not all grow at the same rate, nor do all of the oscillators saturate in amplitude by the end of the simulation. We generally find that the oscillators that are the smallest in amplitude belong to the spins that possess the highest degree of frustration (number of unsatisfied interactions) in the corresponding Ising spin configuration. Due to the unsatisfied connections, these oscillators dissipate more power than average in their resistive connections, thus requiring more parametric gain to reach saturation. We also notice in this case that even after all 32 oscillators have locked onto one of the two bistable phases, the oscillators can respond to changes in amplitude elsewhere in the network and switch to the opposite phase. While these appear as sudden jumps in phase in Fig. 6.7(b), they are seen in Fig. 6.7(c) as somewhat more gradual transitions in amplitude across zero. We argue in the next section that the most frustrated spins are the most susceptible to these late-time spin flips. As shown in Fig. 6.7(d), the circuit ultimately finds and settles to the global minimum of the problem within 320 ns of the simulation's start.[2]

We next evaluate our problem on an unweighted Max-Cut problem with 60 vertices and 50% edge density, which is part of the Biq-Mac library in Ref. 144.[3] For this moderately large problem, shown in Fig. 6.8(a), the global maximum is known and has a value of 529. Using Equation (5.3), we can convert this into an Ising problem with two possible weight values ($J_{ij} = -1$ and 0) and a global minimum of $H_{\min} = -179$. For this trial, we use a constant gain of $\Delta C/C_0 = 29.5\%$, estimated using Equation (6.6) as explained below.

Fig. 6.8(b) and (c) show the oscillator amplitudes and phases. One outstanding feature of these plots is that while almost all of the oscillators have saturated in amplitude by the end of the simulation, there are exactly two oscillators whose amplitudes remain close to zero. On closer inspection, we find that these two oscillators represent the two spins in the ground state configuration (labeled spin 33, green, and spin 58, red) that have exactly 50% of their interactions left frustrated – the highest possible proportion. Consequently, there

[2]For this problem, the exhaustive search for the global minimum, written in Python, took 30 minutes to complete on two Intel Xeon E5 processors, parallelized over 28 cores. Meanwhile, the full LTspice simulation in Fig. 6.7 was carried out an Intel Core i7-8700 processor, with no parallelization, within 5 minutes.

[3]The problem attempted here is identified as `g05_60.2` and is available at `http://biqmac.uni-klu.ac.at/biqmaclib.html`.

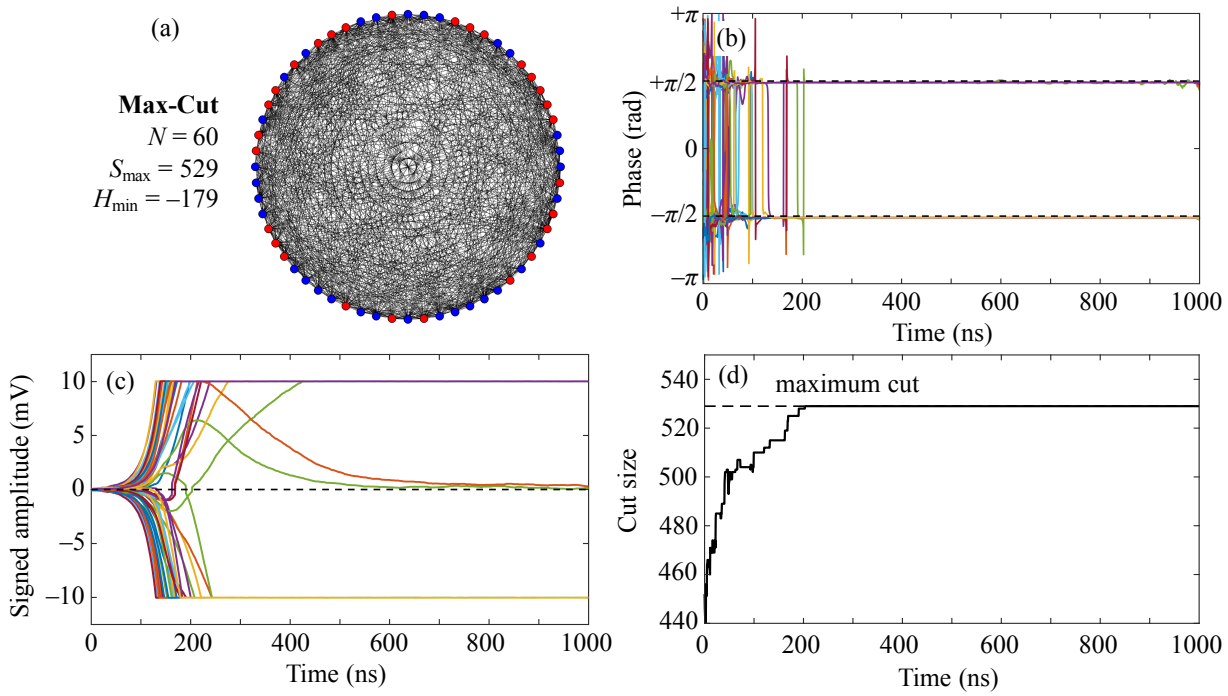

Figure 6.8: Circuit simulation results for the analog Ising machine with 60 coupled LC oscillators, which solves the Max-Cut problem depicted in (a), with an edge density of 50%. The problem comes from the Biq-Mac library, and has a known maximum cut size of 529 [144]. The phases and signed amplitudes of the 60 oscillators are shown in (b) and (c), respectively. (d) shows the evolution of the cut size. The network settles to the Max-Cut solution within 203 ns; this solution, which is one of several possible ways partition the vertices to obtain the maximum cut size, is shown in (a).

are exactly as many interactions that pull the oscillators towards a phase of $+\pi/2$ as there are towards $-\pi/2$. Since the problem is unweighted, this also means that flipping either of these spins does not change the value of the Ising Hamiltonian or the cut size. The machine discovers the maximum cut solution shortly after $t = 200$ ns. If either of the two spins with near-zero amplitude flips its phase, the resulting spin configuration would still yield the maximum cut.

We tested the performance of our Ising machine on a number of problems with varying sizes, up to 32 spins.[4] Fig. 6.9(a) shows the simulated success probability as a function of the number of spins N for Ising graphs with all-to-all connectivity and random binary weights. For each problem size on this plot, 100 different random Ising graphs were generated, their global minima were found by an exhaustive search, and the machine was given exactly one attempt to independently solve each problem. For these problems, we use a constant gain

[4]For problem sizes larger than 32 spins, we have performed too few trials to date for the extraction of a meaningful value of success probability. Expanding the trials at these moderate size scales is a goal for the near future.

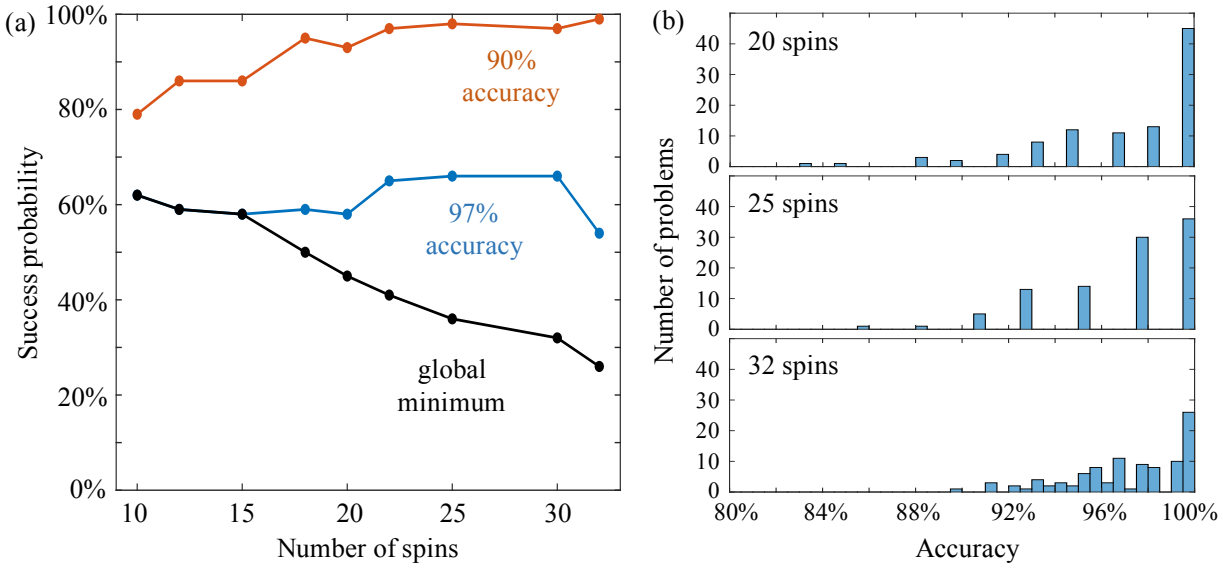

Figure 6.9: (a) The success probability of the analog Ising machine vs. problem size (up to $N = 32$) for fully-connected Ising graphs with random binary weights. Probabilities are shown for finding the global minimum (black) and solutions with at least 97% (blue) or 90% (red) of the energy of the global minimum. For each problem size, 100 different random problems were generated, and the machine was given one attempt to solve the problem. (b) Histogram of the solution accuracies (out of 100 problems) for $N = 20$, 25, and 32. The width of each bin is 0.5%.

that is fully agnostic to the coupling matrix, other than its size:

$$\frac{\Delta C}{C_0} = \frac{N-1}{\omega R_c C_0} \tag{6.6}$$

The above expression tends to over-estimate the value of the threshold gain for a given coupling matrix. Nonetheless, it is often advantageous to use a value of gain that exceeds this threshold, as this helps to avoid one of the failure modes of the optimization that we will discuss in Section 6.5.

The probability of finding the global minimum using our Ising machine is shown by the black curve in Fig. 6.9(a), and decreases for larger problems as expected. The problem sizes in our trials are as yet too small to allow the extraction of a meaningful scaling law for the success probability. We note nonetheless that our performance is comparable to the coherent Ising machine's success probability on dense random graphs of the same size [145].

The probabilities of finding an approximate solution, whose Ising energy is within 97% or 90% accuracy of the global minimum, are also shown in Fig. 6.9(a) by the blue and red curves, respectively. The histogram of solution accuracies are shown in Fig. 6.9(b).[5] The

[5]Accuracy is defined as: $1 - (H - H_{\min}) / (H_{\max} - H_{\min})$. The histograms are more discrete for smaller N simply due to the relatively small number of distinct energy levels.

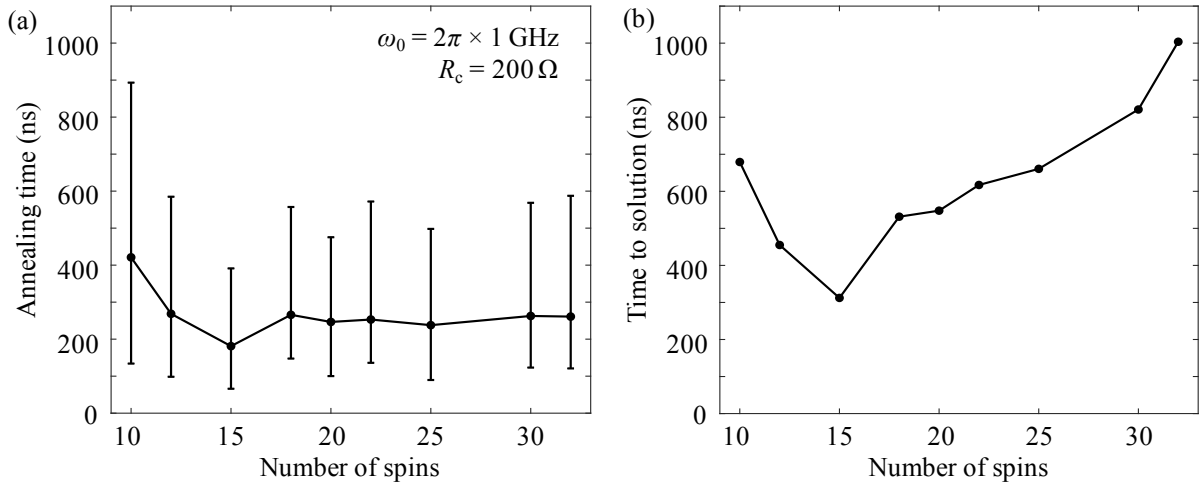

Figure 6.10: (a) The characteristic annealing time of the Ising machine – the average time taken to converge onto a solution after the onset of threshold – is shown vs. problem size The error bars encompass annealing times that are within ±25% of the mean. (b) The time-to-solution vs. problem size of the Ising machine, given by the characteristic annealing time divided by the success probability.

97% success probability stays relatively steady with problem size, while the 90% success probability actually increases for larger problems, reaching a near-guarantee for $N \geq 25$. These trends may be explained by the observation that for problems similar to the Ising spin glass, the number of high-quality local minima in the solution space increases with problem size, and these minima become more tightly clustered in energy around the global minimum [146]. For applications like machine learning, where there is relatively little to be gained by solving exactly for the global minimum, our machine can be used to rapidly sample a large number of good solutions.

How quickly does our machine search through the solution space? In Fig. 6.10(a), we plot the characteristic annealing time of the Ising machine as a function of problem size. We define the annealing time as the time taken to converge onto a solution to the Ising problem after the onset of threshold. Since we have used a constant, above-threshold gain in these trials, we simply measure the annealing time from the starting point of the simulation. The average annealing times in Fig. 6.10(a) were calculated only over the searches in Fig. 6.9 that successfully found the global minimum, though the annealing time in the unsuccessful attempts are similar.

The numerical results suggest that the characteristic annealing time of the machine is a constant that is independent of problem size. The annealing time is closely related to the oscillation bandwidth, which sets the timescale over which an oscillator can rotate its phase to lock into a spin-up or a spin-down state. Since each oscillator is an RLC circuit with a resistance that is dominated by the coupling network, this bandwidth is roughly approxi-

mated by $4R_cC_0 \approx 127$ ns for our choice of parameters. Consistent with the results of our simulations, this time-scale does not increase with the number of spins in the problem. We will justify this property more rigorously in the following section. Thus, as the problem size increases, the Ising machine samples a local minimum of the solution space in constant time. This describes the machine's intrinsic optimization search speed, ignoring any delays associated with communicating information across the oscillator network, which cannot realistically be neglected when the system is large. The characteristic time to find the global minimum, called the time to solution, is given by product of the annealing time and the expected number of trials before the global minimum is found. The time to solution is plotted in Fig. 6.10(b) and increases with problem size as the probability of success declines.

To validate our future results for problems much larger than 32 spins, for which an exhaustive search on our machines is impractical, there are multiple approaches. The first is to run our machine on benchmark instances of the Max-Cut problem, which is directly equivalent to the Ising problem, that are widely used by the optimization community. For moderately large problems (up to $N = 250$), instances have been published with known global minima [144]. For larger problems, the quality of our solutions can be compared to that of other algorithms and hardware systems on the benchmarking set described in Ref. 147: results on 2000-spin problems, whose ground states are not exactly known, have already been reported by other Ising machines [105], [128], [143], [148]. Yet another approach is to generate large, difficult instances of the Ising problem with a known planted solution; a procedure for this has been outlined in Ref. 149.

6.5 Dynamics of the analog Ising machine

In this section, we will formulate a more complete and rigorous theory for the operation of our analog Ising machine. The goal is to shed more light on the simulated performance presented in the previous section, as well as to address the important question of whether the machine truly implements the principle of first-to-threshold search that was described in Section 5.2. We saw previously that the power dissipation in the circuit maps to the Ising Hamiltonian when the system reaches a steady state with uniform, saturated oscillation amplitudes. But that is, ideally, the final state of the circuit; the search over the solution space takes place while the state of the circuit is still evolving, and here we will find a dynamical equation that describes this evolution. To preserve clarity, we omit much of the mathematical derivation from this section and focus instead on interpreting the equation. For a more detailed derivation, we refer the reader to Appendix F.

We begin by applying the basic equations of circuit analysis to our oscillator network. Consider the circuit in Fig. 6.11, which shows the resistive coupling network between two oscillators labeled k and j. The oscillator output voltage V_k is used as the indicator of the Ising spin. We use a four-resistor network, which can encode any value of the weight J_{jk} by selecting the resistance values $R_{p,jk}$ and $R_{n,jk}$ by means of Equations (6.4). Inside each LC oscillator is a noise voltage source $V_{n,k}$, a linear inductor L, and a nonlinear capacitor that

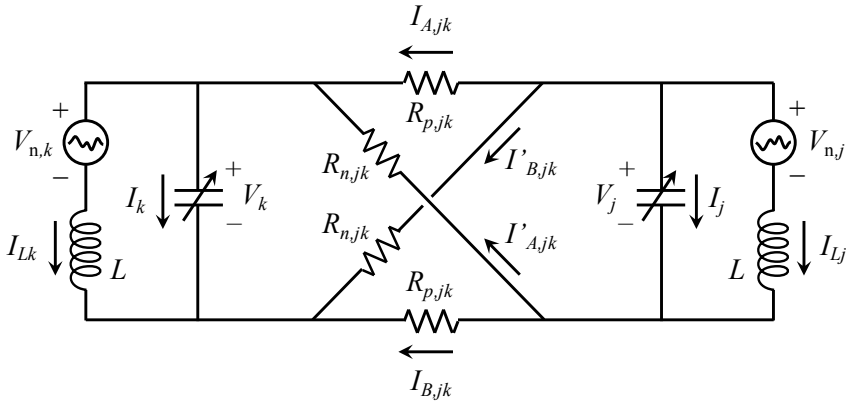

Figure 6.11: Circuit diagram for two coupled oscillators k and j inside the Ising machine, implementing a weight J_{jk}. The polarities of the coupling currents are selected as shown for the case $j > k$: for $j < k$, the currents are reversed.

draws a current $I_k(V_k)$. An internal series resistance in the oscillator can also be included in I_k, though in our analysis we assume that the oscillators are lossless. The coupling currents between two oscillators are chosen to flow from oscillator j to oscillator k, where $j > k$. We designate the negative terminal of oscillator 1 to be ground.

By applying Kirchoff's circuit laws to the circuit in Fig. 6.11, we obtain the following set of equations:

$$0 = V_k - V_{\mathrm{n},k} - L\dot{I}_{Lk} \tag{6.7a}$$

$$0 = V_k - V_j + (I_{A,jk} - I_{B,jk})\, R_{p,jk} \qquad (j > k) \tag{6.7b}$$

$$0 = V_k + V_j + \left(I'_{A,jk} - I'_{B,jk}\right) R_{n,jk} \qquad (j > k) \tag{6.7c}$$

$$0 = V_k + I'_{A,jk} R_{n,jk} - I_{B,jk} R_{p,jk} \qquad (j > k) \tag{6.7d}$$

$$0 = V_k - V_j + I_{A,jk} R_{p,jk} - I_{B,j1} R_{p,j1} + I_{B,k1} R_{p,k1} \qquad (j > k > 1) \tag{6.7e}$$

$$0 = \textstyle\sum_{j>k} \left(I_{A,jk} + I'_{A,jk}\right) - \sum_{j<k} \left(I_{A,jk} + I'_{B,jk}\right) - I_k - I_{Lk} \tag{6.7f}$$

$$0 = \textstyle\sum_{j>k} \left(I_{B,jk} + I'_{B,jk}\right) - \sum_{j<k} \left(I'_{A,jk} + I_{B,jk}\right) + I_k + I_{Lk} \tag{6.7g}$$

Equation (6.7a) and Equations (6.7b)-(6.7d) are found by applying Kirchoff's voltage law (KVL) inside each oscillator and to all possible current loops between the two oscillators, respectively. Equation (6.7e) is found by applying KVL to the current loop that passes from ground to the positive terminals of the two oscillators j and k. We obtain Equations (6.7f) and (6.7g) by applying Kirchoff's current law (KCL) to the positive and negative nodes of each oscillator, respectively. One of the $2N$ KCL equations is redundant by current conservation. In total, there are $2N^2$ independent equations, exactly equal to the number of unknowns in the circuit: $\{V_k,\, I_{Lk},\, I_{A,jk},\, I'_{A,jk},\, I_{B,jk},\, I'_{B,jk}\}$.

We can eliminate all of the extraneous unknowns from the above equations, leaving only a single set of dynamical equations for the oscillator voltages V_k. Here, we omit the algebraic steps, which can be found in Appendix F. The result is:

$$V_k - V_{\mathrm{n},k} = \frac{L}{2R_\mathrm{c}} \left[\sum_{j \neq k} J_{jk} \dot{V}_j - (N-1)\dot{V}_k \right] - L\dot{I}_k \tag{6.8}$$

This exact equation holds for any arbitrary coupling matrix J as long as the coupling resistances are chosen according to Equations (6.4). Notably, this means that a weight of zero must be implemented using four matched resistors, rather than by fully decoupling the oscillators as in Fig. 6.2(c).

We now introduce the effect of the parametric nonlinearity. As in our SPICE simulations, we will model the parametric pump purely as a sinusoidal capacitance modulation, given by Equation (6.3). The current through the time-varying capacitor is:

$$I_k = \frac{\partial}{\partial t}\left(CV_k\right) = \dot{C}V_k + C\dot{V}_k \tag{6.9}$$

Phase-sensitive parametric amplification allows us to map the binary Ising spin to the phase of the oscillation, as described in Section 6.1. It will therefore be convenient to separate the oscillatory part of the voltage V_k from its time-varying *real-valued* amplitude A_k:

$$V_k(t) = A_k(t)\cos\left(\omega t + \phi\right) \tag{6.10}$$

To simplify our analysis, we have made two significant assumptions in writing the above equation: (1) under the influence of the pump oscillator, all of the oscillators lock to the frequency $\omega = \frac{1}{2}\omega_p$, and (2) the phase dynamics is faster than the amplitude dynamics, so that the oscillators lock to one of the two bistable phases while the amplitudes continue to evolve. The first assumption holds for our simulation, where the oscillators have no variability in their frequencies, though in practice the spread in fundamental frequency would need to be smaller than the bandwidth of any individual oscillator. The second assumption is based on the fact that only the oscillation with the correct phase relative to the pump can have net gain, while the incorrect phase, separated by $\pm\pi/2$ radians, always has a nonzero net loss. Thus, the oscillators can be considered to have stabilized in phase once their amplitudes have risen well above the noise floor (e.g. by $\sim$10$\times$). This is validated by the voltage waveforms in Fig. 6.5(d), where we see that the network as a whole becomes phase-bistable while the amplitudes continue to evolve. This assumption allows us to absorb the difference between the two bistable phases into the sign of the real-valued amplitude A_k. The additional phase delay ϕ is a constant that is shared by all oscillators and determines the phase relationship between the ensemble of oscillators and the second-harmonic pump. To be consistent with assumption (2) above, it must be chosen so that the oscillators collectively experience parametric gain. Spin flips are still allowed under Equation (6.10) by changing the sign of A_k, though continuous changes in phase are not modeled.

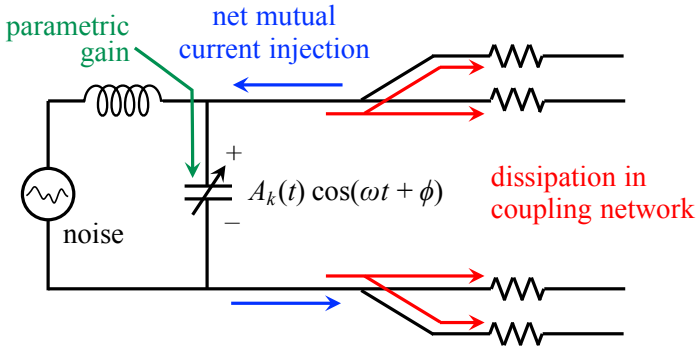

Figure 6.12: Circuit diagram showing the physical meaning of the individual terms in the Ising machine dynamical equation, Equation (6.11). The polarity of the blue arrows indicates net current injection from the other oscillators *into* oscillator k.

Inserting the expressions for the oscillator voltages V_k and the nonlinear currents I_k into Equation (6.8), we obtain a dynamical equation describing the evolution of the signed oscillation amplitudes (see Appendix F for the mathematical details):

$$\dot{A}_k = \frac{1}{4R_{\rm c}C_0}\sum_{j\neq k} J_{jk}A_j + \left(-\frac{N-1}{4R_{\rm c}C_0} - \frac{\omega\Delta C}{4C_0}\sin 2\phi\right)A_k \tag{6.11}$$

In arriving at this equation, we have ignored the noise term, $V_{{\rm n},k}$. While noise can seed the initial oscillation in the circuit, it quickly becomes very small in comparison to the signal once the oscillation has been sufficiently amplified – this is part of assumption (2) mentioned above. We have also made the approximation that ΔC is a weak modulation: $\Delta C/2C_0 \ll 1$. Finally, we have made the slowly-varying amplitude approximation: since the natural resonance oscillation of the voltage has been separated from the amplitude in Equation (6.10), A_k has a negligible second derivative $\ddot{A}_k$.

Let us now interpret this dynamical equation. The first term on the right-hand side represents the net injection of current into or out of oscillator k, resulting from its coupling to the other oscillators in the network, which pulls the voltage in a positive or negative direction. This is shown by the blue arrows in Fig. 6.12. The relative injection strengths to or from these neighbors are determined by the coupling weights J_{jk} of the embedded Ising problem, as well as the amplitudes A_j of the neighboring oscillators. This term can be written to show that it leads to a minimization of an *analog* Ising Hamiltonian $H_{\rm a}$:

$$\dot{A}_k = -\frac{1}{4R_{\rm c}C_0}\frac{\partial H_{\rm a}}{\partial A_k} + \left(-\frac{N-1}{4R_{\rm c}C_0} - \frac{\omega\Delta C}{4C_0}\sin 2\phi\right)A_k \tag{6.12}$$

where

$$H_{\rm a} = -\sum_{\langle j,k\rangle} J_{jk}A_jA_k \tag{6.13}$$

This analog Ising Hamiltonian is analogous to the digital Ising Hamiltonian H given by Equation (5.2), but with the binary spins replaced by real-valued spin amplitudes A_k. Only in the limit that the oscillators have saturated to the same steady-state amplitude is H_a directly proportional to H. What happens during the system's evolution? If we ignore the second term in Equation (6.12) for the moment, we find that the stationary points of the oscillation amplitude ($\dot{A}_k = 0$) correspond to minima of the analog Ising Hamiltonian ($\partial H_\mathrm{a}/\partial A_k = 0$). Whatever the initial condition on A_k, this term drives all of the oscillation amplitudes in the circuit to evolve in a manner that relaxes H_a to a minimum; this is the actual nature of the circuit's implementation of the Ising problem! Importantly, we also notice in Equation (6.12) that the characteristic timescale of the resistive interaction between oscillators is $4R_\mathrm{c}C_0$, a factor that is independent of the problem size if we ignore parasitic reactances in the connections. This is consistent with our empirical finding in Fig. 6.10(a) that the annealing time does not depend on the number of spins.

Turning our attention to the second term on the right-hand side and ignoring the first term, we find an equation of the form $\dot{A}_k = \gamma A_k$, whose solution either grows ($\gamma < 0$) or decays ($\gamma > 0$). The first term in the brackets accounts for power loss in the coupling resistors, shown by the red arrows in Fig. 6.12. If a capacitance modulation ΔC is present, the second term inside the brackets can either be negative (parametric de-amplification) or positive (parametric amplification, shown in green in Fig. 6.12), depending on the value of the phase delay ϕ. The oscillators collectively experience maximum gain if $\phi = 3\pi/4$ or $\phi = 7\pi/4$, and these are the only values that are compatible with our assumptions in Equation (6.10). Which of the two is selected is inconsequential: choosing one or the other simply flips the sign of every oscillation, which has no effect on the digital or the analog Ising Hamiltonian. Thus, we can simplify the dynamical equation to:

$$\dot{A}_k = -\frac{1}{4R_\mathrm{c}C_0}\frac{\partial H_\mathrm{a}}{\partial A_k} + \left(-\frac{N-1}{4R_\mathrm{c}C_0} + \frac{\omega\Delta C}{4C_0}\right)A_k \tag{6.14}$$

Note that while the mutual injection term above can act to increase the amplitude, $\dot{A}_k$ cannot be positive in the absence of parametric gain. If the gain $\Delta C/C_0$ is not large enough to overcome the network losses, the only steady-state solution to Equation (6.14) is that of zero amplitude: $A_k = 0$ for all k, and there is no oscillation above the noise floor.

Let us now predict how the distribution of the signed amplitudes, collected into a vector $\vec{A}$, evolves over time. Provided that none of the oscillators have reached saturation, we can express Equation (6.14) for the pre-saturation (or early-time) dynamics as:

$$\dot{\vec{A}} = \frac{1}{\tau_1}J\vec{A} + \frac{1}{\tau_2}\vec{A} \tag{6.15}$$

where $\tau_1 = 4R_\mathrm{c}C_0$ and $\tau_2 = \tau_1\left(\omega R_\mathrm{c}\Delta C - N + 1\right)^{-1}$ are constants with units of time. Since the coupling matrix J is real and symmetric, its eigenvalues are given by a diagonal matrix $\Lambda = QJQ^\mathrm{T}$, where Q is an orthogonal basis matrix ($Q^\mathrm{T} = Q^{-1}$). With this knowledge, we can manipulate Equation (6.15) as follows:

$$\begin{aligned} Q\dot{\vec{A}} &= \frac{1}{\tau_1} QJ\vec{A} + \frac{1}{\tau_2} Q\vec{A} = \frac{1}{\tau_1} QJQ^{-1}(Q\vec{A}) + \frac{1}{\tau_2} Q\vec{A} \\ \frac{\partial}{\partial t}(Q\vec{A}) &= \left(\frac{1}{\tau_1}\Lambda + \frac{1}{\tau_2}\right) Q\vec{A} \end{aligned} \tag{6.16}$$

where $Q\vec{A}$ is simply the amplitude vector in the orthogonal basis. Expressed in this way, the dynamical equation states that if the amplitude distribution $\vec{A}$ is proportional to an eigenvector of the matrix J, this distribution is preserved by the system's time evolution; we can therefore call these the modes of the system. For the i^{th} eigenvector of J, the corresponding amplitude distribution will experience gain if $\frac{1}{\tau_1}\Lambda_{ii} + \frac{1}{\tau_2} > 0$. The modes with the largest positive eigenvalues will grow the most rapidly, until saturation is reached. Similar findings have been made using an independent model of the coherent Ising machine [143]. To simplify our notation, suppose that the eigenvalues are arranged in descending order along the diagonal of Λ. The value of gain at which the m^{th} mode of the system reaches threshold is given by:

$$\left(\frac{\Delta C}{C_0}\right)_{\text{th},m} = \frac{(N-1) - \Lambda_{mm}}{\omega R_{\text{c}} C_0} \tag{6.17}$$

The gain threshold corresponding to the first mode, $m = 1$, is the absolute threshold for the system, below which no oscillation can occur. Other modes, corresponding to smaller eigenvalues of J, are excited at higher values of the gain threshold. In a sense, this enables a first-to-threshold search. However, it does not truly represent the first-to-threshold optimization mechanism proposed in Section 5.2. Since the modes with distinct gain thresholds are eigenvectors of J, there are only N competing modes, much fewer than the 2^N possible solutions to the Ising problem. Also, the spin configuration corresponding to the maximum eigenvalue is, of course, not generally the global minimizer of the Ising Hamiltonian. Therefore, there is not necessarily a benefit allowing only the first mode to oscillate. In the likely scenario that this first mode lies far away from the global minimum, it would be advantageous to use a higher gain, exciting a superposition of multiple oscillation modes. The linearity of the pre-saturation dynamics also implies that unless the global minimum lies close to these system modes, we must rely heavily on the nonlinearity provided by gain saturation in our search.

By modeling parametric gain purely as a modulated capacitance, we have so far not accounted for gain saturation in our equations. In our SPICE simulations, we model saturation by adding two ideal diodes with opposite polarities to our oscillator, shown in Fig. 6.4. Their effect is to force $\dot{A}_k = 0$ when the oscillation magnitude attempts to increase above the diode turn-on voltage A_{sat}. We can model this effect by modifying Equation (6.14) to:

$$\dot{A}_k = \tilde{\dot{A}}_k \times \left[1 - u\big(A_k^2 - A_{\text{sat}}^2\big)\, u\Big(A_k \tilde{\dot{A}}_k\Big)\right] \tag{6.18}$$

where $u(\cdot)$ is the Heaviside step function and $\tilde{\dot{A}}_k$ is the value of $\dot{A}_k$ given by Equation (6.14). In a physical implementation, the amplitude saturates more smoothly.

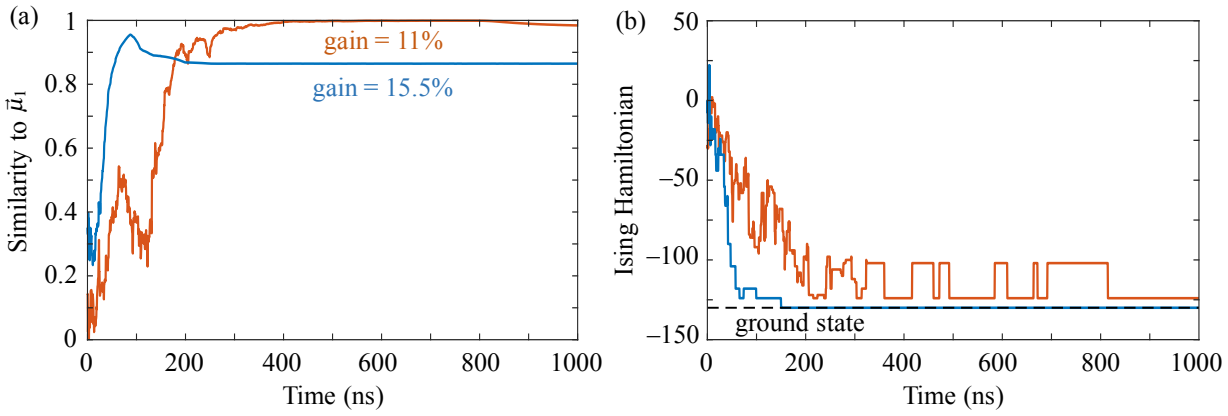

Figure 6.13: Evolution of the Ising machine which solves the 32-spin problem in Fig. 6.7(a), under two constant values of gain. (a) Evolution of the oscillator amplitude distribution's similarity to $\vec{\mu}_1$, the eigenvector of the coupling matrix J with the largest positive eigenvalue. This is evaluated using Equation (6.19). (b) The corresponding evolution of the Ising Hamiltonian.

As we saw in Section 6.4, it is possible for an oscillator to reverse its direction while its amplitude is small, momentarily moving back toward $A_k = 0$: see Fig. 6.7(b). This occurs in response to the nonlinear saturation of one or more of the other oscillators in the network. The oscillators most susceptible to these phase reversals are those which represent spins with a high degree of frustration. These oscillators teeter on the edge between giving and receiving current from the oscillators to which they are coupled, and thus are very sensitive to nonlinear disturbances elsewhere in the circuit. These late-time spin flips are not simple to predict, as they can no longer be described by a linear equation as in Equation (6.16). It is evident, however, that these highly nonlinear dynamics are responsible for the convergence to the global minimum in the 32-spin problem shown in the previous section. We observe in Fig. 6.7(b) that the first oscillators reach saturation near $t = 300$ ns. These saturation events drive a few of the lower-amplitude (highly frustrated) oscillators to flip their phases, ultimately yielding the global minimum, which is first found at around $t = 330$ ns.

To illustrate this more explicitly, where we operate the same circuit in Fig. 6.7 with two constant values of the gain. In Fig. 6.13(a), we plot the evolution of the quantity:

$$S = \left| \vec{A}^{\mathrm{T}} \vec{\mu}_1 \right| \; / \; \left\| \vec{A}^{\mathrm{T}} Q \right\| \tag{6.19}$$

which measures how closely the signed amplitude distribution $\vec{A}(t)$, extracted using the moving window method described in the previous section, matches $\vec{\mu}_1$, which represents the amplitude distribution of the first collective mode of oscillation. A gain of $\Delta C/C_0 = 11\%$ lies between the thresholds for $m = 1$ and $m = 2$: thus, only the first mode is allowed to oscillate. We find in Fig. 6.13(a) that in this case, as the oscillators increase in amplitude above the noise floor, the similarity of the amplitude distribution to $\vec{\mu}_1$ increases to 100%. For this problem, the spin configuration $\mathrm{sgn}(\vec{\mu}_1)$ has a Hamiltonian value of $H = -124$, while

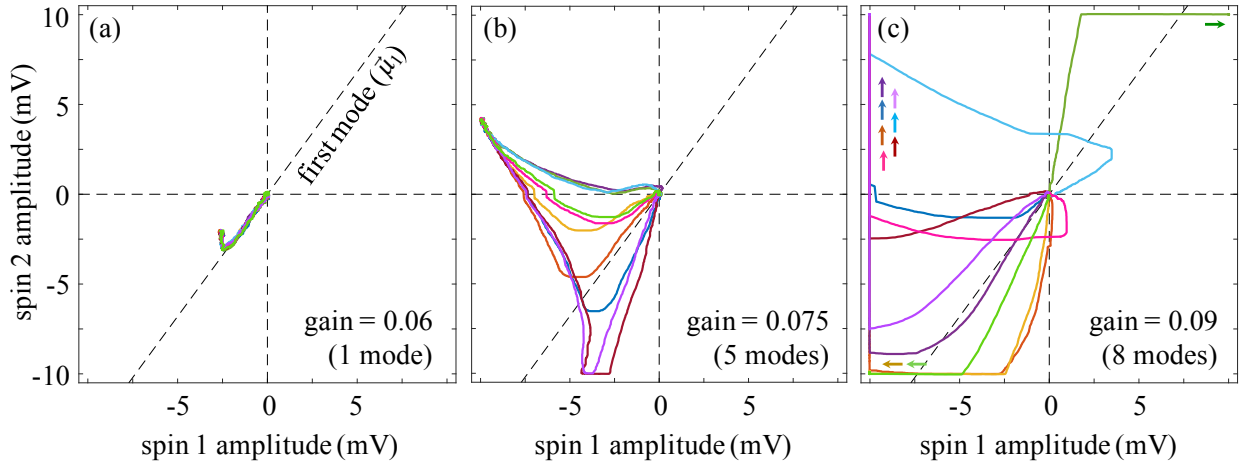

Figure 6.14: The Ising machine's evolution through a two-dimensional slice of phase space as it solves a fully-connected, binary Ising problem with 20 spins. For the selected spins, the second and fourth quadrants represent the correct solution. The simulation uses a fixed gain of (a) $\Delta C/C_0 = 0.06$, (b) 0.075, and (c) 0.09, corresponding to one, five, and eight oscillation modes above threshold, respectively. Each of the ten colors represents a trajectory with different random initial conditions provided by the noise sources. The machine becomes more sensitive to initial conditions at larger gain. The saturation voltage is 10 mV, and the RMS noise amplitude is 25 μV.

$H_{\text{min}} = -130$. Thus, the machine ultimately yields the wrong answer. The similarity to $\vec{\mu}_1$ decreases slightly at the end of the simulation, as some of the oscillators reach saturation.

On the other hand, a gain of $\Delta C/C_0 = 15.5\%$, given by Equation (6.6) from the previous section, allows all modes with a positive eigenvalue to oscillate. Here, the amplitudes increase more quickly due to the higher gain; the similarity to $\vec{\mu}_1$ also increases, since the first mode experiences the largest net gain according to Equation (6.16). This similarity reaches a peak of 95.5% when the first saturation event occurs. Afterwards, nonlinear dynamics leads the system to relax to the global minimum, whose similarity to $\vec{\mu}_1$ is less than 90%. We hypothesize that the search is successful in this scenario because additional modes of the system had been excited at the onset of amplitude saturation, effectively providing more degrees of freedom to the system's evolution. We can also attribute this success partially to the proximity of the first mode to the global minimum: the spin configuration $\text{sgn}(\vec{\mu}_1)$ is only one spin flip away from the ground state.

Our Ising machine is a dynamical system that evolves through the N-dimensional phase space defined by the real-valued oscillation amplitudes $\{A_1, A_2, ...A_N\}$. Below threshold, the system fluctuates around the origin of phase space. Above threshold, noise initially drives the system in a random direction away from the origin, which becomes unstable. Fig. 6.14 shows the system's evolution through a two-dimensional slice of phase space for a fully-connected, 20-spin problem. First consider Fig. 6.14(a), where the gain is chosen to be slightly larger than the threshold given in Equation (6.17) with $m = 1$, allowing only the first mode $\vec{\mu}_1$ to

oscillate. Each color represents a trajectory with different noisy initial conditions. Though the trajectories diverge as they leave the origin, they quickly collapse onto $\vec{\mu}_1$ since no other mode has gain. After one or more of the oscillators in the network reaches saturation (spins other than 1 or 2 in this case), the dynamics becomes nonlinear and the trajectories can deviate from $\vec{\mu}_1$. For this value of gain, since all of the trajectories initially collapsed onto $\vec{\mu}_1$, they all eventually stabilize to the same point in phase space. This point yields an incorrect relative orientation of the two spins.

Fig. 6.14(b) shows the dynamics in phase space with a higher gain, which allows the modes corresponding to the five largest eigenvalues of J to grow. In this case, noise injects a different initial amplitude to these multiple modes, leading to a greater diversity of initial trajectories in phase space. Nonetheless, $\vec{\mu}_1$ remains the most preferred direction as it has the largest gain coefficient. The sudden changes in trajectory correspond to the onset of nonlinear dynamics. Despite the diversity in initial trajectories, all paths lead to the same point in phase space with a large basin of attraction. In Fig. 6.14(c), which uses a still larger gain, the initial trajectories become more diverse and both spins are eventually driven to saturation. This time, the trajectories terminate at different points in phase space representing different solutions to the Ising problem: of the ten trajectories, seven have selected the correct orientations of the two spins. Some of the trajectories are very similar initially but eventually settle to different outcomes. Therefore, it appears that while physical noise is a source of stochasticity in the optimization process, the machine only becomes stochastic in outcome if the system's evolution begins with a large enough value of gain that multiple modes of the circuit can oscillate. A slow linear ramp from zero, which gives the network enough time to settle to the mode $\vec{\mu}_1$, might be largely deterministic in outcome.

We note that dynamical equations which are similar in form and in behavior to Equations (6.14) and (6.18) for our Ising machine have been proposed as an idealized continuous-time model for the coherent Ising machine – both by the original authors who conceived the machine [129], [130] as well as independently by another group [143]. However, as we stated in Section 6.3, the coherent Ising machine that has been experimentally demonstrated cannot be described by these dynamics since the oscillators in that system interact very sparsely in time. Our machine, which is fully analog, appears to be a closer physical embodiment of the dynamics discussed in these works. However, there is much about our system's nonlinear evolution after the onset of saturation that remains to be better understood and placed on a more rigorous mathematical footing; we leave this to the realm of future work.

We reiterate, as a final point, that our circuit does not faithfully reproduce the first-to-threshold optimization scheme proposed in Section 5.2. It remains to be seen what physical system, if any, can truly implement or more closely approximate this search method. In principle, if heterogeneity in the oscillation amplitudes can be kept small during the full course of the system's evolution, a direct mapping between power dissipation and the digital Ising Hamiltonian – given by Equation (6.5) – can be maintained while the hardware searches for an optimal solution. An approach was recently proposed, within the continuous-time theoretical framework for the coherent Ising machine, for the mitigation of amplitude heterogeneity. This approach involves continuously sampling all the oscillator amplitudes, then using this

information to adaptively control the gain supplied to each oscillator to maintain close to a uniform amplitude distribution. This method has been implemented computationally and has shown some success in improving the success probability [130].

6.6 Design trade-offs and considerations for scaling

Up to now, we have considered only the intrinsic performance of our Ising machine, ignoring many of the non-idealities that would arise in a physical hardware implementation. In this section, we will study the key issues in design that must be considered in our implementation of the analog Ising machine. We find that design trade-offs arise among a number of metrics: the accuracy of the problem implementation, optimization speed, robustness to oscillator variability, robustness to circuit parasitics, energy efficiency, reliability, and optimization accuracy. Some of these trade-offs are intrinsic to the machine's operation, while others are purely practical (i.e. can be eliminated in an idealized simulation of the machine). These non-idealities ultimately impose limits on the size of the optimization problem that can be realistically embedded onto and ultimately solved by the machine. Our discussion in this section will be qualitative, deferring further analytical or numerical exploration of these issues to the realm of future work.

We found in Section 5.4 that analog imprecision in the parameters of the optimization problem, arising in our case from variability in the coupling resistors, limits the size of the problem that can be accurately represented. The amount of variability in the coupling resistors depends on the magnitudes of the resistance values that are used: in our implementation, this is set by the single parameter R_c. Lower resistance is expected to yield lower variability, as current is conducted through a greater number of quantum channels. A lower value for resistance is also desirable for ensuring robustness to variability in the reactive components, L and C_0. We find empirically that variability that appears in the oscillators' natural frequencies, given by $\omega_0 = \sqrt{LC_0}$, translates to a spread in phase around the two bistable values. These phase errors are not destructive so long as the spread in the natural frequencies is smaller than the bandwidth $\Delta\omega$ of each oscillator, which is given roughly by $1/R_cC_0$. Thus, a lower resistance allows oscillators with a greater variability in their reactive components to join the ensemble of injection-locked oscillators, which is critical for the success of the optimization.

In terms of the intrinsic speed of the optimization, a lower resistance value is again advantageous. As we found in our analysis in Section 6.5, and confirmed in our numerical results in Fig. 6.10(a), the characteristic timescale for coupling between oscillators is set by the quantity $4R_cC_0$, which is independent of problem size. A lower resistance value means that a greater current is mutually injected among the coupled oscillators, allowing them to respond more quickly to changes in oscillator states in the network.

On the other hand, the higher currents introduced by low-resistance connections can be problematic. With a lower value for R_c or a larger problem size, larger resistive voltage drops are generated between oscillators. We observe in simulation that for the same gain,

networks with lower R_c settle to a steady-state configuration with fewer saturated oscillators – this can already be seen to some extent in Fig. 6.7(c) with $R_c = 200\,\Omega$. It is not yet fully understood how the success probability is influenced by the resistance. Larger voltage drops imply that the oscillators are less accurately sampling the amplitudes of their neighbors. Having fewer saturated oscillators also means that the Hamiltonian H_a that is minimized by the system is less likely to take on a digital character. For the system's dynamics, however, the voltage drops may potentially be beneficial. When the first oscillators reach saturation, a greater fraction of the remaining oscillators remain at low amplitude. As discussed in Section 6.5, this means that the subsequent evolution of these more frustrated oscillators becomes susceptible to nonlinear dynamical effects, which we have found to be important in the discovery of global minima. In loosely related technologies such as resistive crossbars for machine learning, these voltage drops are known to be harmful to performance, driving the push for higher-resistance implementations of connection weights [150].

A clear disadvantage of lower resistance is energy efficiency – a larger amount of power is dissipated in the coupling resistors and consequently, a higher gain must be supplied to overcome these losses and reach threshold. If the resistance is low enough, the threshold gain given by Equation (6.6) can no longer truly be considered a weak capacitance modulation ($<$10%) even for a moderate problem size. In this regime, the approximations we have made to arrive at our dynamical equation will no longer hold, though it is not clear that the dynamics in this case will lead to worse (or even substantially different) performance. The regime of strong modulation can be moved to larger problem sizes by increasing the inductance. On a more practical level, the high currents enabled by low resistance can potentially cause reliability issues such as electromigration in highly scaled technologies.

An effect that we have mentioned but neglected in our simulations is delay – time delays in communication between spins as well as phase delays in the second-harmonic pump as it is distributed to the various oscillators in the network, analogous to clock skew in digital systems. These delays are caused by reactive and resistive parasitics which scale with the system's interconnect lengths. As the circuit grows in area to accommodate larger problems, these delays will become more important and may set the limits to scaling. Between oscillators, the introduction of delay may severely limit the system's performance and reduce its speed, but is unlikely to render it useless as an optimizer; the coherent Ising machine, for example, has achieved good optimization results while allowing the physical spins to communicate very sparsely in time. The effect of delay on accuracy was also found to be small for the CMOS annealing machine [104]. On the other hand, for our system, significant asynchrony in the received pump signal can be detrimental to performance with large problem sizes. While a small phase delay in the pump would make little difference, a delay that is close to π radians would assuredly destroy bistability, as the same oscillator phase can experience parametric gain or parametric loss depending on its position in the network. A compromise to mitigate both types of delay is to use a lower oscillation frequency ω, slow enough to fit the longest delay in the network into one cycle; this would, however, reduce the optimization search speed and lead to a narrower locking range.

Finally, we observe that while non-idealities like component precision and parasitic re-

actances pose practical limits to scaling for the physical Ising machine, these effects can be eliminated altogether in an idealized circuit simulation. Our SPICE simulation returns voltage waveforms that encode the machine's solution to the Ising problem and thus, the simulation is itself an optimization algorithm. While this method of solving the problem will inevitably be slower than a physical realization of the ideal system being simulated, it bypasses the practical difficulties of implementation and thus is potentially more scalable. The overhead of a circuit simulation can be largely removed by solving instead a smaller set of dynamical equations, such as Equations (6.14) and (6.18), that capture the essence of the system's operation. Since these equations are fully classical, they do not inherently require exponential resources to model on a conventional computer. The important question is whether the time complexity of the optimization is fundamentally altered by emulating the analog Ising machine, in which every spin communicates with every other spin continuously in time, on a digital computer, which is inherently sequential. For our system, this question remains not fully resolved. We note that for the coherent Ising machine, efforts to emulate the machine's dynamics on digital processors have led to similar optimization performance with a speed that is superior to the actual physical machine [143], [148].[6]

[6]The present implementation of the coherent Ising machine is itself not a purely analog machine; it is a hybrid processor with an analog optical component and a digital FPGA [127].

Chapter 7

Conclusion

In Part I of this thesis, we examined the emergent properties of optoelectronic devices as they approach their fundamental limits. The threshold at which electroluminescent cooling is observed, which is the condition of unity wall-plug efficiency, can be loosely treated as the "ultra-efficient" regime. The highest-performance light-emitting diodes that have been reported to date have not yet crossed this efficiency threshold, though we expect that this will change in the near future. Why the optimism? In Chapter 3, we showed that within reach of the present-day capabilities of the optoelectronics industry – in terms of material quality and device processing – is a GaAs device with 97.4% external luminescence efficiency at room temperature. Such a device, which is well into the ultra-efficient regime, will certainly not be simple to realize. However, we can expect that devices will approach this level of performance as continued progress is made in the coming years in optoelectronics.

Treating this device as a practical limit of LED technology, we foresee some promise in electroluminescent cooling as a viable mechanism of cooling. At room temperature, this type of cooling can surpass thermoelectric devices in efficiency when operated at moderate heat fluxes (in the 10 mW/cm^2 range), leading to niche applications. LED cooling seems better matched to low-temperature applications, such as the cryogenic cooling of infrared photodetectors, as LEDs improve in efficiency for the same reasons that photodetectors also improve, all while thermoelectric performance sharply declines. The practical limit of LED cooling, predicted in this thesis for present technology, should be treated as a moving goalpost as optoelectronics improves. An important feature of LED cooling is that the practical limit departs sharply from the theoretical limit with only a very small amount of internal loss. While this may seem like a discouraging fact, it does mean that once we enter the regime of ultra-efficient devices, every small improvement in the luminescence efficiency towards ideal will be met by a large gain in the accessible cooling efficiency.

In Part II, we presented an oscillator-based analog computer that is capable of finding good solutions to difficult combinatorial optimization problems. In designing this machine, we set out to implement the first-to-threshold mechanism of searching for the global minimum; a closer look at the system's dynamics, however, reveals that its principle of operation is rather different. Nonetheless, our analog Ising machine explores its phase space for the op-

timal solution in a unique way, and is often successful. A key property is that it can reliably settle to a good solution (though not necessarily the global minimum) within a timescale that is independent of the problem size, allowing the local minima of the problem to be sampled very rapidly.

Our results so far should be treated as somewhat preliminary, with many questions that remain to be explored in future work. While our dynamical model of the Ising machine elucidates many of its operational features, we have yet to fully understand its failure modes, especially with large problem sizes. Gaining more insight into these failure modes will require a more detailed investigation of the strongly nonlinear (or saturated) regime of the machine's dynamics. Furthermore, to understand how our system compares in performance to other analog approaches as well as to digital algorithms, we must scale up our benchmarking efforts to much larger problem sizes. It is also worthwhile to investigate what modifications must be made, or whether a new hardware architecture is needed, to more accurately implement the first-to-threshold mechanism. Finally, since the problems to be solved are computational, to what extent is a digital simulation of the machine an efficient and scalable manifestation of the underlying idea, in comparison to a physical hardware implementation?

Appendix A

Equivalent expressions for the external luminescence efficiency

In Section 2.1, we showed the equivalence between two commonly used expressions for the external luminescence efficiency,

$$\eta_{\text{ext}} = \eta_{\text{int}} \times C_{\text{ext}} \tag{A.1}$$

$$= \eta_{\text{int}} \times \frac{P_{\text{esc}}}{1 - \eta_{\text{int}} \left(1 - P_{\text{esc}} - P_{\text{par}}\right)} \tag{A.2}$$

Here, we will briefly establish the consistency of these equations with other expressions that are found in the literature for the same quantity. The following analysis applies to both light-emitting diodes and photovoltaic cells.

We will denote the total rates of radiative and non-radiative recombination per unit area of the semiconductor device by R_{rad} and R_{nr}, respectively. The internal luminescence efficiency is then given by $\eta_{\text{int}} = R_{\text{rad}} / \left(R_{\text{rad}} + R_{\text{nr}}\right)$. Additionally, we will denote the removal rates of internal photons by Φ_{ext}, Φ_{abs}, and Φ_{par} for external emission, internal band-to-band absorption, and parasitic absorption, respectively. In steady-state, we must balance the rates of photon addition and removal from the reservoir of luminescent photons inside the semiconductor:

$$R_{\text{rad}} = \Phi_{\text{ext}} + \Phi_{\text{abs}} + \Phi_{\text{par}} \tag{A.3}$$

An excited electron-hole pair in the device active region ultimately either leaves as an external photon (with a rate Φ_{ext}) or is lost through one of the undesired dissipative processes, whether electronic (with a rate R_{nr}) or optical (with a rate Φ_{par}). Therefore, it is possible to write the external luminescence efficiency as,

$$\eta_{\text{ext}} = \frac{\Phi_{\text{ext}}}{\Phi_{\text{ext}} + R_{\text{nr}} + \Phi_{\text{par}}} \tag{A.4}$$

This expression can be used directly to calculate η_{ext}, as in Ref. 5. To see that it is consistent with Equation (A.1), we can manipulate it into a product of two efficiencies. We express the non-radiative recombination rate as $R_{\text{nr}} = (\eta_{\text{int}}^{-1} - 1)R_{\text{rad}}$, then insert Equation (A.3) for R_{rad}. The result, which is used in Section 3.3, is:

$$\eta_{\text{ext}} = \eta_{\text{int}} \times \frac{\Phi_{\text{ext}}}{\Phi_{\text{ext}} + \Phi_{\text{par}} + \Phi_{\text{nr}}} \tag{A.5}$$

where $\Phi_{\text{nr}} = (1 - \eta_{\text{int}})\,\Phi_{\text{abs}}$ is the rate of non-radiative recombination of those carriers generated by the absorption of internal luminescent photons. The second term above is the light extraction efficiency C_{ext}, since Φ_{nr} can effectively be considered a mechanism of internal photon loss. Thus, Equations (A.4) and (A.5) are fully equivalent to Equations (A.1) and (A.2).

In some works, a simplifying assumption is made that recombination occurs through three mechanisms whose rates scale with a power of a single photo-generated carrier density n: Shockley-Read-Hall recombination An, radiative recombination Bn^2, and Auger recombination Cn^3, where A, B, and C are rate coefficients specific to each process. In this case, the internal luminescence efficiency becomes:

$$\eta_{\text{int}} = \frac{Bn^2}{An + Bn^2 + Cn^3} \tag{A.6}$$

If we insert this into Equation (A.2), we obtain another commonly used expression for η_{ext} [8], [37]:

$$\eta_{\text{ext}} = \frac{P_{\text{esc}} Bn^2}{An + (P_{\text{esc}} + P_{\text{par}})\,Bn^2 + Cn^3} \tag{A.7}$$

We note that treating internal recombination in this way fails to account for asymmetric electron and hole concentrations (e.g. due to doping) and depletion-region effects, which can dominate the recombination processes in LEDs and photovoltaic cells in certain bias regimes, as we show in Section 3.2.

Appendix B

LED spreading resistance calculation

In this appendix, we describe the model used to calculate the Ohmic power dissipation Q_Ω and the lateral current and voltage distributions in the LED structure shown in Fig. 3.1. We use this model to generate the results in Section 3.4. The model is also applied to calculate the Ohmic dissipation in the photovoltaic cell, which is used as the hot-side absorber in the thermophotonic cooling configuration (see Section 4.3). In the LED, the net current flows in the forward direction (p to n contact); in the PV device, the reverse current (n to p contact) dominates.

To model the effects of current spreading, we treat the n-type and p-type layers in the device as two parallel, planar sheets across which current can flow, as shown in Fig. B.1. Current travels from the positive contacts on the p-surface to the negative contacts on the n-surface, and passes vertically between the two surfaces through the p-n junction diode, distributed across the device area. At a specific lateral position (x, y), the diode has a quasi-Fermi level separation equal to $V(x, y)$ and draws a vertical current density equal to $J(x, y)$. Here, we focus our attention on a unit cell of the device area (red dashed square in Fig. 3.1), with the approximation that current is not shared between adjacent unit cells. The unit cell has a side length equal to the contact separation of $L_c = 30$ μm.

The thicknesses D_n and D_p of the two sheets are taken to be the total thickness of the n-type and p-type layers in the device epitaxial stack, respectively. We find the sheet resistances $R_{\square n}$ and $R_{\square p}$ of the two surfaces by adding the sheet resistances of the constituent layers in parallel. The resistivity of a layer j is found using $\rho_j = (q\mu_j N_j)^{-1}$, where μ_j and N_j are the mobility and the doping density within the layer. The 15 nm p-GaInP layer, which tends to be fully depleted, is not included in the calculation. Using Ref. 92 to determine the carrier mobilities in each layer, accounting for the dependences on doping and temperature, we obtain $R_{\square n} = 240\ \Omega/\square$ and $R_{\square p} = 12\ \mathrm{k}\Omega/\square$ for the LED at 263K.

We solve for the lateral current densities $\vec{J}_n(x, y)$ and $\vec{J}_p(x, y)$ by applying the current continuity equations. On the p-type surface of the LED, lateral current decreases with distance from the positive contact as charge is drawn out vertically through the diodes. The opposite is true for the n-type surface, which receives the current that is injected through

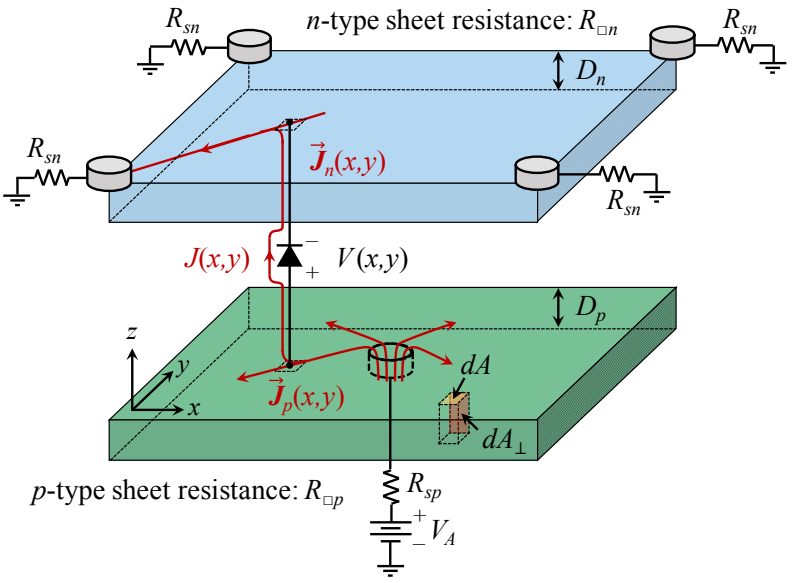

Figure B.1: The electrical model used to calculate the Ohmic dissipation rate and the lateral current and voltage distributions for the device structure shown in Fig. 3.1.

the diodes. Current continuity then yields for the two surfaces,

$$\nabla_{xy} \cdot \vec{J}_n(x,y) = D_n^{-1} J(x,y) \tag{B.1a}$$

$$\nabla_{xy} \cdot \vec{J}_p(x,y) = -D_p^{-1} J(x,y) \tag{B.1b}$$

where $(\nabla_{xy} \cdot)$ is the divergence operator in the x-y plane. We then use Ohm's law to relate the lateral current density to the potential distributions V_p and V_n on each surface:

$$\nabla_{xy} V_n(x,y) = -\rho_n \vec{J}_n(x,y) \tag{B.2a}$$

$$\nabla_{xy} V_p(x,y) = -\rho_p \vec{J}_p(x,y) \tag{B.2b}$$

where ρ_n and ρ_p are the effective electrical resistivities of the n-type and p-type surfaces, respectively. Combining the above four equations and using the relation $V_p - V_n = V$, we obtain a single differential equation for the diode voltage:

$$\nabla_{xy}^2 V(x,y) = (R_{\square n} + R_{\square p})\, J(x,y) \tag{B.3}$$

The current density in its general form is given by Equation (3.26), and has a complicated dependence on the voltage $V(x,y)$ at every position. In our electrical model, we use the expressions in Sections 3.2 and 3.3 to calculate the current density only for a single value of the quasi-Fermi level separation, equal to the desired internal voltage $\overline{V}$. Using the value of $J(\overline{V})$, we then extrapolate to nearby voltage values using a one-diode model:

$$J(x,y) = J(\overline{V}) \exp\left(\frac{qV(x,y) - q\overline{V}}{n_{\mathrm{d}}(\overline{V})kT}\right) - J_{\mathrm{R}} \tag{B.4}$$

where the diode ideality factor $n_{\mathrm{d}}(\overline{V})$ is estimated from the voltage dependence of the forward current at $V = \overline{V}$. We recalculate the values of J and n_{d} when different average voltages $\overline{V}$ over the device area are considered. The ideality factor is unity over the moderate bias regime, and gradually increases above unity as the device enters high-level injection. Equation (B.4) makes the approximation that the ideality factor of the device is uniform over its surface area. This is a valid approximation under most operating regimes of interest. This model starts to become inaccurate only at large voltages ($\overline{V} > 1.35$V) when both the resistive voltage drops across the device become large *and* the ideality factor changes by a non-negligible amount over the device area due to the onset of high-level injection.

The term J_{R} is the reverse current density in the device. If the LED is considered alone, as in Chapter 3, J_{R} is equal to the reverse saturation current (or dark current), which is negligible under forward bias. In the thermophotonic system considered in Chapter 4, J_{R} is the photocurrent that is generated by the absorption of luminescence from the opposite device, assuming that the absorption is laterally uniform. When the system is biased for optimal cooling efficiency, J_{R} is a small fraction of the forward current in the LED but dominates in the photovoltaic cell.

Using Equation (B.4), we convert Equation (B.3) into a second-order non-linear differential equation for the lateral voltage distribution $V(x,y)$. Additionally, we must ensure that our calculation is consistent with the internal voltage $\overline{V}$ that we have used in Equation (B.4). This is done by setting the total power consumption of the device – with a non-uniform voltage distribution – equal to that of the same device with a laterally uniform voltage of $\overline{V}$:

$$J\left(\overline{V}\right)\overline{V} = \frac{1}{L_{\mathrm{c}}^2}\iint J(x,y)\,V(x,y)\,dA \tag{B.5}$$

where the integral is taken over a unit cell. Note that we have used the same equation in Chapter 3 to define the internal voltage. This can then be used as a boundary condition on Equation (B.3). We solve the equation using the successive over-relaxation method, evaluating the voltages $V(x,y)$ on a uniform grid whose resolution is $dx = dy = 0.5\,\mu$m.

From the solution for $V(x,y)$, we can write a differential equation for the voltage of the n-surface by combining Equations (B.1) and (B.2):

$$\nabla_{xy}^2 V_n(x,y) = -R_{\square n} J(x,y) \tag{B.6}$$

With the right side now known, we again use the successive over-relaxation method to solve for $V_n(x,y)$. The p-surface voltage distribution is given by $V_p(x,y) = V_n(x,y) + V(x,y)$. The lateral current densities $\vec{J}_n$ and $\vec{J}_p$ are then readily found using Equation (B.2).

The Ohmic power dissipation due to the spreading resistance on each surface is found by integrating the contribution from each differential element of the surface currents. The total

dissipation also includes the losses from the series resistances R_{sn} and R_{sp}:

$$Q_\Omega = \frac{1}{L_\mathrm{c}^2} \iint \left((dI_n)^2 R_{\square n} + (dI_p)^2 R_{\square p} \right) + \frac{I^2 (R_{sn} + R_{sp})}{L_\mathrm{c}^2} \tag{B.7}$$

where I is the total current injected into the unit cell:

$$I = \iint J(x,y)\, dA \tag{B.8}$$

The differential lateral current is given by $dI_p = |\vec{J}_p| \times dA_\perp$, where $dA_\perp = dx \times D_p$ is the differential area normal to the direction of lateral current flow. Inserting this into Equation (B.7), we arrive at:

$$Q_\Omega = \frac{1}{L_\mathrm{c}^2} \iint \left(\left|\vec{J}_n(x,y)\right|^2 D_n^2 R_{\square n} + \left|\vec{J}_p(x,y)\right|^2 D_p^2 R_{\square p} \right) dA + \frac{I^2 (R_{sn} + R_{sp})}{L_\mathrm{c}^2} \tag{B.9}$$

For the device in Fig. 3.1, Ohmic dissipation is dominated by the spreading resistance of the p-type surface, due to the low hole mobility in the p-GaInP layer.

The n- and p-side series resistances are calculated by:

$$R_{sn} = \frac{\rho_\mathrm{c}}{\pi (d_\mathrm{c}/2)^2} + \frac{\rho_\mathrm{m} N_\mathrm{c}^2 L_\mathrm{c}}{3 L_\mathrm{g} h_\mathrm{g}} \tag{B.10a}$$

$$R_{sp} = \frac{\rho_\mathrm{c}}{\pi (d_\mathrm{c}/2)^2} + \frac{\rho_\mathrm{m} h_\mathrm{via}}{\pi (d_\mathrm{c}/2)^2} \tag{B.10b}$$

where the first term in each expression is the Ohmic contact resistance, assuming a contact resistivity of $\rho_\mathrm{c} = 1.0 \times 10^{-6}\ \Omega\,\mathrm{cm}^2$ for both the n-type and p-type contacts [62], and $d_\mathrm{c} = 1\,\mu\mathrm{m}$ is the contact diameter. The second term in R_{sn} is the resistance associated with the front metal grid lines, where $\rho_\mathrm{m} = 1.6 \times 10^{-8}\ \Omega\,\mathrm{m}$ is the resistivity of Ag, $L_\mathrm{g} = 5\,\mu\mathrm{m}$ is the grid line width, $h_\mathrm{g} = 5\,\mu\mathrm{m}$ is the grid line depth, and N_c is the total number of contacts connected to each grid line. Assuming a total dimension of $L_\mathrm{total} = 5$ cm for the LED array, we have $N_\mathrm{c} = 1667$. This resistance can be made lower by using a square mesh rather than a linear grid, but this would also double the surface coverage f_g of the grid, which increases parasitic absorption. The second term in R_{sp} is the resistance of the metal via connecting the p-GaInP layer to the rear Ag backplane, where h_via is the depth of the via trench, equal to the total thickness of the Bragg reflector and MgF_2 layer. We assume negligible spreading resistance in the Ag backplane.

Fig. B.2 shows the electrical efficiency – defined by Equation (3.27) – calculated using the model presented in this section. The dashed line was found by calculating Q_Ω using Equation (B.9). Note that in this solution, the electrical efficiency has an unphysical plateau near $\overline{V} = 1.32$V. This is caused by the fact that with increasing bias and current crowding, the gradient of the surface voltage near the contact becomes larger, and the voltage needs to be sampled more finely to accurately evaluate this gradient. Therefore, a progressively

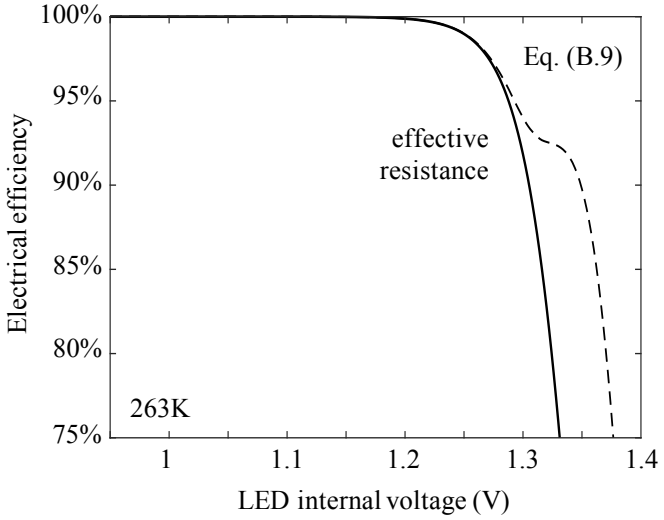

Figure B.2: The electrical efficiency as a function of the LED internal voltage $\overline{V}$. The dashed curve is calculated from Equation (B.9), whereas the solid curve is found using an effective resistance as in Equation (B.11).

finer mesh (smaller dx) is needed to calculate the surface currents $\vec{J}_n$ and $\vec{J}_p$ using Equation (B.2). If the mesh is not sufficiently fine, we underestimate the surface currents and hence Q_Ω. Note that this problem does not appear in the solution for $V(x, y)$, which can be determined without knowing the surface currents and thus does not require the calculation of a voltage gradient.

Rather than using an extremely fine mesh to accurately calculate the Ohmic dissipation at large biases, we use the fact that for lower voltages ($\overline{V} < 1.3$V), the solution for Q_Ω is accurate. These solutions can then be used to extract an effective resistance:

$$Q_\Omega(\overline{V}) = J^2(\overline{V})R_{\text{eff}} \tag{B.11}$$

which is valid for all biases. This solution is shown by the solid curve in Fig. B.2. For the same geometry, the effective resistance is a function of temperature.

Appendix C

Heat leakage in the thermophotonic system

As mentioned in Sections 4.3 and 4.4, one of the chief advantages of thermophotonic refrigeration over thermoelectric cooling is the fact that the luminescent heat flux from cold to hot can be decoupled from the undesired non-luminescent heat leakage from hot to cold. The primary mechanisms of heat leakage in this system are (1) thermal radiation and (2) heat conduction through the electrical feedback connection which passes electricity from the PV cell to the LED. In this appendix, we engineer the system so that the combined effect of these leakage mechanisms can be reduced to $Q_{\rm leak} < 100$ μW/cm^2, independent of the amount of luminescent heat transfer in the system. At this level, heat leakage has a very small influence on cooling performance for all practical power densities ($Q_{\rm c} \geq 1.0$ mW/cm^2).

C.1 Radiative heat leakage

The vacuum gap does nothing to inhibit the passive heat transfer by radiation from the hot side to the cold side, which occurs in opposition to the active heat transfer by luminescent radiation. Since we have assumed a far-field separation between the two devices, the net amount of heat transferred by thermal radiation is bounded by the blackbody limit: $\max\left(Q_{\rm leak,rad}\right) = \sigma\left(T_{\rm h}^4 - T_{\rm c}^4\right) = 27.3$ mW/cm^2 for temperatures of $T_{\rm c} = 263$K and $T_{\rm h} = 313$K, where σ is the Stefan-Boltzmann constant. In reality, the heat leakage will be smaller than this value, since the devices have less than unity emissivity at the far-infrared wavelengths of the thermal photons. The exact amount of heat transfer depends on the thicknesses and the dielectric functions of every layer in both devices.

Since the most practical regime of operation of the electroluminescent refrigerator lies at moderate fluxes of $1-10$ mW/cm^2, we must take steps to suppress the thermal radiative heat leakage. One strategy, shown in Fig. C.1, is to insert a metal mesh in the vacuum gap between the two devices, which can act as an optical high-pass filter [151]. The periodicity of the mesh can be chosen to maximally reflect the low-energy thermal photons and transmit

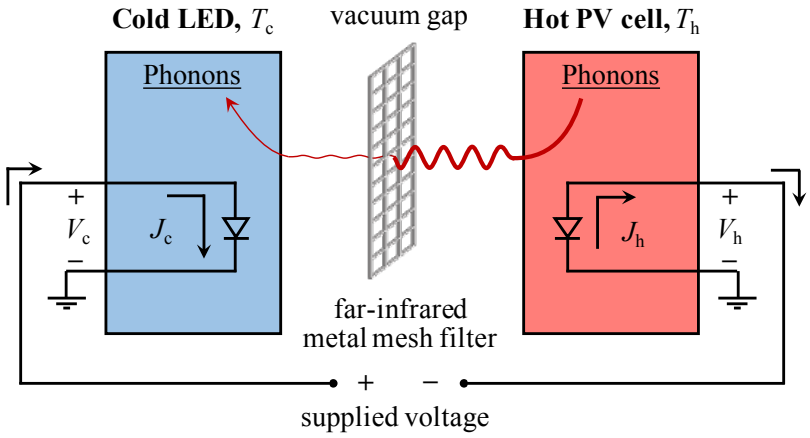

Figure C.1: Heat leakage by thermal radiation can be suppressed by inserting a far-infrared metal mesh filter in the vacuum spacer between the hot and cold sides, which reflects a large fraction of the thermal photons back to their source.

the high-energy luminescent photons. Assuming blackbody thermal radiators, the required far-infrared reflectivity is 99.8% in order to reduce the leakage below 50 μW/cm^2. For temperatures of ($T_c = 113$K, $T_h = 313$K) or ($T_c = 263$K, $T_h = 338$K), the required reflectivity is 99.9%. These requirements can be considerably relaxed after accounting for the below-unity emissivity of the two devices in the far-infrared.

C.2 Heat leakage via the electrical feedback connection

In the thermophotonic configuration, the electrical connection that supplies the generated photovoltaic power to the LED is a source of heat leakage by direct thermal conduction. This leakage can be suppressed by using a narrower or longer feedback wire, which has a lower thermal conductance. However, this would also have the effect of making the wire more electrically resistive, contributing to greater Ohmic heating of the LED. To break this trade-off, we can connect an ensemble of individual LEDs or PV cells in series on the cold and hot sides of the system, respectively. A large voltage builds up across the ensemble, but the total current through it is reduced, permitting the use of a narrow wire to attain both low heat leakage and low resistive losses.

The heating of the LED caused by the electrical feedback connection is the sum of the heat leakage through the wire and half of the Joule heating along the wire (the other half heats the PV side):

$$Q_{\text{leak,fb}} = \frac{1}{A}\left[\frac{\kappa_{\text{fb}} A_{\text{fb}}}{l_{\text{fb}}}\left(T_h - T_c\right) + \frac{1}{2} I_h^2 \frac{l_{\text{fb}}}{\sigma_{\text{fb}} A_{\text{fb}}}\right] \tag{C.1}$$

where A is the total area of the device served by a feedback connection, κ_{fb} and σ_{fb} are the thermal and electrical conductivities of the wire, l_{fb} and A_{fb} are the wire length and cross-sectional area, and I_{h} is the total PV current that passes through the connection.

Suppose that a single feedback connection carries the power generated in N series-connected PV cells, each with an area A_0. The total current is equal to the current generated in a single cell, $I_{\mathrm{h}} = J_{\mathrm{h}} A_0$, where J_{h} is the current density in the PV device, but the total area served by the connection is $A = NA_0$. We also notice that the two terms in Equation (C.1) depend in opposite directions on the wire geometric parameter ($A_{\mathrm{fb}}/l_{\mathrm{fb}}$). By choosing the value of this geometric parameter that minimizes $Q_{\mathrm{leak,fb}}$ for a given J_{h}, we can simplify the equation to:

$$Q_{\mathrm{leak,fb}} = \frac{J_{\mathrm{h}}}{N}\sqrt{\frac{2\kappa_{\mathrm{fb}}}{\sigma_{\mathrm{fb}}}\Big(T_{\mathrm{h}} - T_{\mathrm{c}}\Big)} = \frac{J_{\mathrm{h}}}{N}\sqrt{L\Big(T_{\mathrm{h}} + T_{\mathrm{c}}\Big)\Big(T_{\mathrm{h}} - T_{\mathrm{c}}\Big)} \tag{C.2}$$

where we have approximated the ratio of thermal to electrical conductivity in a metal using the Wiedemann-Franz law: $\kappa_{\mathrm{fb}}/\sigma_{\mathrm{fb}} = L \times \frac{1}{2}\left(T_{\mathrm{h}} + T_{\mathrm{c}}\right)$, where $L = 2.44 \times 10^{-8}\ \mathrm{W\,\Omega\,K^{-2}}$ is a constant, known as the Lorenz number.

Using the GaAs device in Fig. 3.1 and operating the thermophotonic system between $T_{\mathrm{c}} = 263\mathrm{K}$ and $T_{\mathrm{h}} = 313\mathrm{K}$, we find that a net cooling flux of 10 $\mathrm{mW/cm^2}$ corresponds to a PV current density of $J_{\mathrm{h}} \approx 60\ \mathrm{mA/cm^2}$. To reduce the parasitic heating to $Q_{\mathrm{leak,fb}} = 50$ $\mu\mathrm{W/cm^2}$, where it is negligible compared to the electroluminescent cooling flux, we need to connect $N = 32$ cells in series. When the same configuration is biased to provide a lower cooling flux, Equation (C.1) implies that the heat leakage will also be smaller than 50 $\mu\mathrm{W/cm^2}$. In practice, the LEDs and PV cells must be connected both in series and in parallel in order to properly match their total currents.

Appendix D

The Carnot limit in thermophotonics

In this section, we will show analytically how the cooling performance of the thermophotonic system depends upon the applied biases and the luminescence efficiencies of the two devices. We will then strip away all of the inefficiencies in the system's operation to determine what conditions must be met to obtain a coefficient of performance equal to the Carnot limit given in Equation (4.12). For this idealized derivation, we will assume that Ohmic losses and non-luminescent heat leakage are absent ($Q_{\Omega c} = Q_{\Omega h} = Q_{leak} = 0$); this will keep our expressions analytical, and in any case we are ultimately interested in the limit of zero losses. We will further assume that the quantum efficiency of reverse current generation is ideal: $\eta_{abs,c} = \eta_{abs,h} = 1$.

Given these simplifications, we can re-write Equations (4.10) and (4.11) for Q_c and W, respectively, by inserting the expressions for the current densities in Equation (4.13):

$$Q_c = \left(\langle E\rangle_c - \frac{qV_c}{\eta_{ext,c}}\right)\Phi_c - \left(\langle E\rangle_h - qV_c\right)\Phi_h \tag{D.1}$$

$$W = \left(\frac{qV_c}{\eta_{ext,c}} - qV_h\right)\Phi_c + \left(\frac{qV_h}{\eta_{ext,h}} - qV_c\right)\Phi_h \tag{D.2}$$

Combining these, we obtain an expression for the coefficient of performance:

$$\mathrm{COP} = \frac{Q_c}{W} = \frac{\left(\langle E\rangle_c - \frac{qV_c}{\eta_{ext,c}}\right) - \frac{\Phi_h}{\Phi_c}\left(\langle E\rangle_h - qV_c\right)}{\left(\frac{qV_c}{\eta_{ext,c}} - qV_h\right) + \frac{\Phi_h}{\Phi_c}\left(\frac{qV_h}{\eta_{ext,h}} - qV_c\right)} \tag{D.3}$$

This equation summarizes many of the operational characteristics of the thermophotonic refrigerator in the absence of Ohmic losses. In particular, since the reverse photon flux Φ_h reduces the COP, the optimal value of the photovoltaic voltage V_h generally results in the condition $\Phi_h \ll \Phi_c$. When this holds, the external luminescence efficiency $\eta_{ext,h}$ of the PV cell is of relatively little importance to the overall cooling efficiency, as noted in Section 4.3, while the external luminescence efficiency of the LED is always important.

We now proceed to eliminate the remaining losses in the system in hopes of attaining the Carnot limit. We begin by removing non-radiative recombination and parasitic luminescence absorption, leading to unity external luminescence efficiencies: $\eta_{\text{ext,c}} = \eta_{\text{ext,h}} = 1$. In this limit, the COP simplifies to:

$$\text{COP} = \frac{\Big(\langle E\rangle_{\text{c}} - qV_{\text{c}}\Big) - \dfrac{\Phi_{\text{h}}}{\Phi_{\text{c}}}\Big(\langle E\rangle_{\text{h}} - qV_{\text{c}}\Big)}{\Big(qV_{\text{c}} - qV_{\text{h}}\Big) + \dfrac{\Phi_{\text{h}}}{\Phi_{\text{c}}}\Big(qV_{\text{h}} - qV_{\text{c}}\Big)} \tag{D.4}$$

The Carnot limit is fundamentally the limit of zero entropy generation. When applied individually to the LED and the PV cell, this condition implies zero net current extraction out of the device [80]. The Carnot limit must therefore occur at the open-circuit condition of the PV cell: $J_{\text{h}} = 0$. Since we have assumed unity quantum efficiency, Equation (4.13) then directly yields the requirement of equal luminescence fluxes at open-circuit: $\Phi_{\text{c}} = \Phi_{\text{h}}$.

Additionally, a reversible refrigeration cycle must be quasi-static, allowing no heat transfer to take place between two reservoirs with a finite temperature difference:

$$Q_{\text{c}} = \Big(\langle E\rangle_{\text{c}} - qV_{\text{c}}\Big)\Phi_{\text{c}} - \Big(\langle E\rangle_{\text{h}} - qV_{\text{c}}\Big)\Phi_{\text{h}} = 0 \tag{D.5}$$

When the above equation is combined with the condition of equal luminescence fluxes, we find that the average photon energies must be equal: $\langle E\rangle_{\text{c}} = \langle E\rangle_{\text{h}} = \langle E\rangle$. Therefore, at the Carnot limit, the two devices must also emit the same luminescence intensity. These twin requirements – equal luminescence flux and equal luminescence intensity – can be expressed using Equation (3.17) as:

$$\int_0^\infty \frac{2\pi E^2}{c^2h^3}\frac{a_{\text{c}}(E)\,dE}{e^{(E-qV_{\text{c}})/kT_{\text{c}}} - 1} = \int_0^\infty \frac{2\pi E^2}{c^2h^3}\frac{a_{\text{h}}(E)\,dE}{e^{(E-qV_{\text{h}})/kT_{\text{h}}} - 1} \tag{D.6}$$

$$\int_0^\infty \frac{2\pi E^3}{c^2h^3}\frac{a_{\text{c}}(E)\,dE}{e^{(E-qV_{\text{c}})/kT_{\text{c}}} - 1} = \int_0^\infty \frac{2\pi E^3}{c^2h^3}\frac{a_{\text{h}}(E)\,dE}{e^{(E-qV_{\text{h}})/kT_{\text{h}}} - 1} \tag{D.7}$$

where we have assumed that the absorptivity of each device is angle-independent, so that the solid angle integral evaluates to π.

Let us recall Equation (4.14) for the emissivity within the closed thermophotonic system. If neither device has any parasitic absorption ($\mathscr{L} = 0$), then we have $a_{0,\text{c}} = a_{0,\text{c}}^\dagger$ and $a_{0,\text{h}} = a_{0,\text{h}}^\dagger$. This condition implies that the net emissivities of the LED and PV cell are exactly identical: $a_{\text{c}}(E) = a_{\text{h}}(E)$, even if the two devices have different structures. Under this condition, Equations (D.6) and (D.7) can be simultaneously satisfied only if the luminescence lies in an infinitely narrow spectral band centered at $E = \langle E\rangle$. This reduces both integrals to a simple equality at this single energy value, and both equations can be satisfied if:

$$\exp\left(\frac{\langle E\rangle - qV_{\text{c}}}{kT_{\text{c}}}\right) - 1 = \exp\left(\frac{\langle E\rangle - qV_{\text{h}}}{kT_{\text{h}}}\right) - 1 \tag{D.8}$$

For a given value of the LED voltage, this yields an ideal open-circuit voltage on the PV cell that is equal to:

$$V_{\mathrm{h,Carnot}} = V_{\mathrm{c}} - \left(\frac{\langle E \rangle}{q} - V_{\mathrm{c}} \right) \frac{T_{\mathrm{h}} - T_{\mathrm{c}}}{T_{\mathrm{c}}} \tag{D.9}$$

To validate that this PV voltage indeed corresponds to the Carnot limit, we set the COP in Equation (D.4) equal to the Carnot COP, using the known requirements of $\Phi_{\mathrm{c}} = \Phi_{\mathrm{h}}$ and $\langle E \rangle_{\mathrm{c}} = \langle E \rangle_{\mathrm{h}} = \langle E \rangle$:

$$\mathrm{COP}_{\mathrm{Carnot}} = \frac{T_{\mathrm{c}}}{T_{\mathrm{h}} - T_{\mathrm{c}}} = \frac{\langle E \rangle - qV_{\mathrm{c}}}{qV_{\mathrm{c}} - qV_{\mathrm{h,Carnot}}} \tag{D.10}$$

This equation is readily shown to be equivalent to Equation (D.9). $V_{\mathrm{h,Carnot}}$ represents the theoretical maximum value of the PV voltage that can be attained only at the Carnot limit. Note that even in this ideal limit the PV voltage is smaller than the LED voltage by an amount that is proportional to the temperature difference. Therefore, it is impossible to construct a thermophotonic system that pumps heat against a temperature difference with no additional voltage provided by an external power supply. Of course, this is to be fully expected as such a system would be a perpetual motion machine.

Fig. D.1 shows the optimal PV voltage V_{h} as a function of the LED voltage V_{c}. The red curve is the Carnot upper bound on the PV voltage, given by Equation (D.9), for the cold-

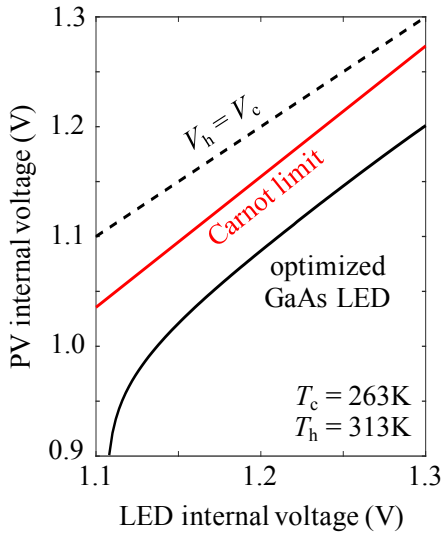

Figure D.1: The optimal forward bias V_{h} on the GaAs PV cell is shown as a function of the bias V_{c} on the GaAs LED. The red curve represents the Carnot limit corresponding to the temperatures $T_{\mathrm{c}} = 263\mathrm{K}$ and $T_{\mathrm{c}} = 313\mathrm{K}$, used in Section 4.5. The solid black curve shows the actual optimal values of V_{h} using non-ideal devices (shown in Fig. 3.1) and providing finite heat transfer.

and hot-side temperatures used in Section 4.5. The limit $V_h = V_c$ is only attainable when there is no temperature difference. The solid black curve shows the optimal values of the PV bias that yield the cooling performance shown in Section 4.5. This curve deviates from the Carnot limit, as an optimized but non-ideal GaAs device structure (shown in Fig. 3.1) has been used. Even if the device were ideal, it would still be necessary to operate below the Carnot PV voltage in order to obtain a nonzero cooling flux. We note that the PV voltage decreases sharply at low LED fluxes, as the PV device enters a regime dominated by Shockley-Read-Hall recombination.

We review the conditions that must be met to operate the thermophotonic refrigerator with Carnot efficiency:

1. Both devices must have zero non-radiative recombination, zero parasitic photon absorption, zero Ohmic losses, and ideal carrier extraction.
2. The parasitic non-luminescent heat leakage from hot to cold is zero.
3. Both devices must operate at the open-circuit condition, drawing zero net current.
4. Both devices must absorb and emit light over an infinitely narrow energy band.

These conditions are very strict indeed, and are clearly of only theoretical significance. The final requirement can be understood by the fact that entropy is generated in the conversion from one photon energy (or optical mode) to another by the processes of electronic absorption and luminescent emission. Classically, the uncertainty that is created in this process can only be fully eliminated by filtering the emission and absorption of photons to a single energy and a single spatial mode. However, in view of quantum optics, incoherent LED radiation that is perfectly filtered in this manner is a mixed quantum state that is distinct in nature from a single-mode coherent state that emerges from an ideal laser, which is a pure quantum state. The former still carries nonzero entropy, while the latter does not.

Appendix E

Thermoelectric cooling performance

In this appendix, we briefly discuss the calculation of the thermoelectric cooling properties shown in Fig. 4.10(b). As mentioned in Section 4.5, the curve represents the cooling performance that is accessible to a family of thermoelectric coolers, which have different device dimensions but the same material figure-of-merit $ZT = 1$. This calculation uses as its basis the equations for heat transfer across the thermoelectric module, accounting for contact resistances. The heat pumped out of the cold side and the heat rejected at the hot side are given by [152]:

$$AQ_{\mathrm{c}} \;=\; (S_p - S_n)\, IT'_{\mathrm{c}} - \tfrac{1}{2} I^2 R - \frac{A\kappa}{l}\,(T'_{\mathrm{h}} - T'_{\mathrm{c}}) \;=\; AK_{\mathrm{c}}\,(T_{\mathrm{c}} - T'_{\mathrm{c}}) \tag{E.1a}$$

$$AQ_{\mathrm{h}} \;=\; (S_p - S_n)\, IT'_{\mathrm{h}} + \tfrac{1}{2} I^2 R - \frac{A\kappa}{l}\,(T'_{\mathrm{h}} - T'_{\mathrm{c}}) \;=\; AK_{\mathrm{c}}\,(T'_{\mathrm{h}} - T_{\mathrm{h}}) \tag{E.1b}$$

where A is the cross-sectional area of a single unit of a thermoelectric module, which contains one n-type (Seebeck coefficient $S_n < 0$) and one p-type thermoelement ($S_p > 0$). We assume that the two thermoelements have the same cross-sectional area $A/2$ and the same thickness l. A total current I flows through the module. The supplied electrical work is the difference between AQ_{h} and AQ_{c}.

The first term in the middle expressions in Equations (E.1) corresponds to the cooling (heating) of the cold (hot) side of the module by the Peltier effect. The second term accounts for the Joule heating of the two devices arising from the electrical resistance R:

$$R = 2\,\frac{\rho l}{A/2} + 4\,\frac{\rho_{\mathrm{c}}}{A/2} \tag{E.2}$$

where ρ is the average bulk resistivity of the thermoelements and ρ_{c} is the average resistivity of the four semiconductor-metal contacts in the module.

To account for the effects of thermal contact resistance, we define the temperatures T'_{c} and T'_{h} to be the cold- and hot-side temperatures internal to the thermoelectric module, prior to contact with the external world. T_{c} and T_{h} are the externally accessible cold- and hot-side

temperatures at the module's thermal contacts. κ is the average bulk thermal conductivity of the thermoelements and K_c is the thermal conductance of the contacts.

For the various material parameters in Equations (E.1), we choose values that are representative of state-of-the-art commercial thermoelectric modules. Specifically, to yield $ZT = 1$ using Equation (4.16) at the temperatures used in Fig. 4.10(b), we assume a Seebeck coefficient of $S = 250\,\mu\text{V/K}$,[1] a bulk electrical resistivity of $\rho = 1.8 \times 10^{-3}\,\Omega\,\text{cm}$, and a bulk thermal conductivity of $\kappa = 1.0$ W/m/K, which are accessible with optimized thermoelectric materials such as Bi_2Te_3 [152]. We assume an electrical contact resistivity of $\rho_c = 10^{-6}$ Ω cm^2 and a thermal contact conductance of $K_c = 100$ kW/m^{2}K.

Equations (E.1) contain two useful free device parameters: the bias current I and the thermoelement length l. We generate the thermoelectric cooling characteristics by sweeping the thickness, and for each thickness we optimize the bias current to yield the maximum COP. Somewhat similarly to the selection of the bias point in thermophotonics, the thickness l sets the power density Q_c and the bias current optimizes the COP. The result is the thermoelectric cooling curve in Fig. 4.10(b). Each point on the curve represents a different thermoelement thickness.

[1] This is an average over the n- and p-type thermoelements: $S = \frac{1}{2}\left(|S_n| + S_p\right)$.

Appendix F

Derivation of the Ising machine dynamical equation

In this appendix, we outline a detailed derivation of the dynamical equation of the analog Ising machine, filling in the missing mathematical steps in Section 6.5. We begin by reproducing Equations (6.7), which are found by applying Kirchoff's circuit laws to the two coupled oscillators k and j shown in Fig. 6.11:

$$0 = V_k - V_{\mathrm{n},k} - L\dot{I}_{Lk} \tag{F.1a}$$

$$0 = V_k - V_j + (I_{A,jk} - I_{B,jk})\, R_{p,jk} \qquad (j > k) \tag{F.1b}$$

$$0 = V_k + V_j + \left(I'_{A,jk} - I'_{B,jk}\right) R_{n,jk} \qquad (j > k) \tag{F.1c}$$

$$0 = V_k + I'_{A,jk} R_{n,jk} - I_{B,jk} R_{p,jk} \qquad (j > k) \tag{F.1d}$$

$$0 = V_k - V_j + I_{A,jk} R_{p,jk} - I_{B,j1} R_{p,j1} + I_{B,k1} R_{p,k1} \qquad (j > k > 1) \tag{F.1e}$$

$$0 = \sum_{j>k} \left(I_{A,jk} + I'_{A,jk}\right) - \sum_{j<k} \left(I_{A,jk} + I'_{B,jk}\right) - I_k - I_{Lk} \tag{F.1f}$$

$$0 = \sum_{j>k} \left(I_{B,jk} + I'_{B,jk}\right) - \sum_{j<k} \left(I'_{A,jk} + I_{B,jk}\right) + I_k + I_{Lk} \tag{F.1g}$$

Our first step is to eliminate I_{Lk} by combining Equations (F.1a) and (F.1f). We can also solve for the cross-linking currents $\{I'_{A,jk}, I'_{B,jk}\}$ in Equations (F.1c) and (F.1d) and eliminate them from the other equations, being careful to reverse the indices k and j when treating the case $j < k$. The equations above reduce to:

$$\begin{aligned} 0 = V_k - V_{\mathrm{n},k} + L \sum_{j \neq k} \frac{\dot{V}_k}{R_{n,jk}} + L\dot{I}_k - L \sum_{j>k} \left(\dot{I}_{A,jk} + \frac{1 - J_{jk}}{1 + J_{jk}}\, \dot{I}_{B,jk} \right) \\ + L \sum_{j<k} \left(\dot{I}_{A,jk} + \frac{1 - J_{jk}}{1 + J_{jk}}\, \dot{I}_{B,jk} \right) \end{aligned} \tag{F.2a}$$

$$0 = \sum_{j\neq k} \frac{V_j - V_k}{R_{n,jk}} + \sum_{j>k} \left(I_{A,jk} + \frac{3 - J_{jk}}{1 + J_{jk}} I_{B,jk} \right) - \sum_{j<k} \left(I_{A,jk} + \frac{3 - J_{jk}}{1 + J_{jk}} I_{B,jk} \right) \tag{F.2b}$$

$$0 = V_k - V_j + (I_{A,jk} - I_{B,jk}) R_{p,jk} \qquad (j > k) \tag{F.2c}$$

$$0 = V_k - V_j + I_{A,jk} R_{p,jk} - I_{B,j1} R_{p,j1} + I_{B,k1} R_{p,k1} \qquad (j > k > 1) \tag{F.2d}$$

where we have used Equations (6.4) to convert the coupling resistances to the interaction weights. Next, we can solve Equations (F.2c) and (F.2d) for the straight-linking currents $\{I_{A,jk}, I_{B,jk}\}$ and eliminate these variables from the equations. We are then left with:

$$\begin{aligned} 0 = {} & V_k - V_{\mathrm{n},k} - L \sum_{j\neq k} \left(\frac{\dot{V}_j - \dot{V}_k}{R_{p,jk}} - \frac{\dot{V}_k}{R_{n,jk}} \right) + L\dot{I}_k \\ & - 2L \sum_{j>k} \left(\frac{\dot{I}_{B,j1}}{1 - J_{j1}} - \frac{\dot{I}_{B,k1}}{1 - J_{k1}} \right) + 2L \sum_{j<k} \left(\frac{\dot{I}_{B,j1}}{1 - J_{j1}} - \frac{\dot{I}_{B,k1}}{1 - J_{k1}} \right) \end{aligned} \tag{F.3a}$$

$$0 = \sum_{j\neq k} \frac{V_j - V_k}{R_{\mathrm{c}}} + 4 \sum_{j>k} \left(\frac{I_{B,j1}}{1 - J_{j1}} - \frac{I_{B,k1}}{1 - J_{k1}} \right) - 4 \sum_{j<k} \left(\frac{I_{B,j1}}{1 - J_{j1}} - \frac{I_{B,k1}}{1 - J_{k1}} \right) \tag{F.3b}$$

Finally, we can substitute Equation (F.3b) into (F.3a) to eliminate the currents $\{I_{B,j1},\, I_{B,k1}\}$. The result is a single set of equations for the unknown oscillator voltages V_k:

$$V_k - V_{\mathrm{n},k} = \frac{L}{2R_{\mathrm{c}}} \left(\sum_{j\neq k} J_{jk} \dot{V}_j - (N-1)\dot{V}_k \right) - L\dot{I}_k \tag{F.4}$$

which is identical to Equation (6.8).

We can express this set of equations more concisely in vector form:

$$\vec{V} - \vec{V}_{\mathrm{n}} = \frac{L}{2R_{\mathrm{c}}} J' \dot{\vec{V}} - L\dot{\vec{I}} \tag{F.5}$$

where $\vec{V}$, $\vec{V}_{\mathrm{n}}$, and $\vec{I}$ are vectors containing the oscillator output voltages, noise voltages, and capacitor currents, respectively. For brevity, we have introduced the matrix J', defined as:

$$J' = J - (N-1) I_N \tag{F.6}$$

where I_N is the $N \times N$ identity matrix. In Section 6.5, we introduced the parametric nonlinearity through Equations (6.9) and (6.10). These can also be written in vector form:

$$\vec{I} = \dot{C}\vec{V} + C\dot{\vec{V}} \tag{F.7}$$

$$\vec{V} = \vec{A} \cos(\omega t + \phi) \tag{F.8}$$

where $\vec{A}$ is the vector of the signed oscillator amplitudes and the time-varying capacitance $C(t)$ is given by Equation (6.3).

Inserting Equation (F.7) for the nonlinear current into Equation (F.5), we obtain:

$$\begin{aligned} 0 = -2R_{\mathrm{c}}\Big[1 - \omega_p^2 L\Delta C\cos(\omega_p t)\Big]\vec{V} \;+\; L\Big[J' + 4\omega_p R_{\mathrm{c}}\Delta C\sin(\omega_p t)\Big]\dot{\vec{V}} \\ -\; 2R_{\mathrm{c}}L\Big[C_0 + \Delta C\cos(\omega_p t)\Big]\ddot{\vec{V}} \end{aligned} \tag{F.9}$$

where we have assumed $\vec{V}_{\mathrm{n}} \ll \vec{V}$ as explained in Section 6.5. We then perform a sequence of algebraic manipulations, outlined as follows. First, we evaluate the time derivatives of $\vec{V}$ and insert them into Equation (F.9). This generates terms that are proportional to the products of sinusoids with frequencies ω and $\omega_p = 2\omega$, which can be expressed as sums of sinusoids with frequencies ω and 3ω. The 3ω oscillations can be discarded, since the spin information is encoded in the oscillations at ω. Next, we expand terms of the form $\cos(\omega t \pm \phi)$ and $\sin(\omega t \pm \phi)$ so that every term in the equation contains a pre-factor of $\cos(\omega t)$ or $\sin(\omega t)$. Since these pre-factors cannot simultaneously be zero, the corresponding terms must separately sum to zero:

$$\begin{aligned} 0 = \Big[\omega^2 R_{\mathrm{c}}\Delta C\vec{A} + J'\dot{\vec{A}} - 2R_{\mathrm{c}}\left(C_0 + \tfrac{1}{2}\Delta C\right)\ddot{\vec{A}}\Big]\cos\phi \\ +\Big[-\omega J'\vec{A} + 4\omega R_{\mathrm{c}}\left(C_0 - \tfrac{1}{2}\Delta C\right)\dot{\vec{A}}\Big]\sin\phi \end{aligned} \tag{F.10a}$$

$$\begin{aligned} 0 = \Big[-\omega J'\vec{A} \;+\; 4\omega R_{\mathrm{c}}\left(C_0 + \tfrac{1}{2}\Delta C\right)\dot{\vec{A}}\Big]\cos\phi \\ +\Big[\omega^2 R_{\mathrm{c}}\Delta C\vec{A} - J'\dot{\vec{A}} + 2R_{\mathrm{c}}\left(C_0 - \tfrac{1}{2}\Delta C\right)\ddot{\vec{A}}\Big]\sin\phi \end{aligned} \tag{F.10b}$$

where we have assumed that the oscillators lock to a frequency close to their natural resonance, $\omega = \sqrt{LC_0}$, to eliminate the inductance. The two equations can be re-combined, and the remaining equation can be manipulated into the following form:

$$\left(1 + \frac{\Delta C}{2C_0}\cos 2\phi\right)\dot{\vec{A}} = \left(\frac{1}{4R_{\mathrm{c}}C_0}J' - \frac{\omega\Delta C}{4C_0}\sin 2\phi\right)\vec{A} + \frac{\Delta C}{8\omega C_0}\sin 2\phi\,\ddot{\vec{A}} \tag{F.11}$$

Our equation so far is exact. We now make the approximation that ΔC is a weak modulation, such that $\Delta C/2C_0 \ll 1$ and the second term on the left side vanishes. We further assume that since the natural resonance oscillation of the voltage has been separated from the time-dependent amplitude in Equation (F.8), the second derivative $\ddot{A}$ is also very small. Thus, we have:

$$\dot{\vec{A}} = \left(\frac{1}{4R_{\mathrm{c}}C_0}J' - \frac{\omega\Delta C}{4C_0}\sin 2\phi\right)\vec{A} \tag{F.12}$$

Converting this equation from vector form back to index notation, and expanding J' using Equation (F.6), we obtain Equation (6.11). In Section 6.5, we refine this equation further and investigate its implications on the system's optimization performance.

Appendix G

The Ising machine SPICE model

In this appendix, we discuss our Ising machine SPICE model in more detail and provide the LTspice netlist for the 8-spin problem shown in Fig. 6.5 and 6.6. This compact problem allows us to illustrate the features of the SPICE model in a concise way. The coupling matrix for the 8-spin problem is given by:

$$J = \begin{pmatrix} 0 & +1 & +1 & -1 & -1 & -1 & +1 & +1 \\ 0 & 0 & +1 & +1 & +1 & +1 & +1 & +1 \\ 0 & 0 & 0 & -1 & -1 & +1 & -1 & +1 \\ 0 & 0 & 0 & 0 & -1 & -1 & +1 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & -1 & +1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & +1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{G.1}$$

The LTspice netlist for the Ising machine circuit that implements this matrix is provided in Listing G.1. The individual oscillators have a natural frequency of $\omega_0 = 2\pi \times 1.0$ GHz and are implemented by the circuit shown in Fig. 6.4. The first such oscillator in the machine is specified in lines 5 – 10 of the following netlist; afterwards, a new oscillator is specified every six lines until line 52. Each oscillator contains a linear capacitor $C_0 = 1.0\,\text{nF}/2\pi$, a linear inductor $L_0 = 1.0\,\text{nH}/2\pi$, a very small internal resistance $R_0 = 1.0\,\text{n}\Omega$, and two ideal diodes (whose model is defined on line 4) with opposite polarities and a forward turn-on voltage of 10 mV. Finally, in parallel with the linear capacitor is a sinusoidally varying capacitor that implements the second term in Equation (6.3) with $\omega_p = 2\pi \times 2.0$ GHz. We allow the gain $\Delta C/C_0$ to increase linearly with time from a starting value of `G0` to a final value of `G0 + dG` over the ramp time `Tramp`, which in this case is also the duration of the transient simulation.

The output voltage of the j^{th} oscillator appears between node $4j-3$ (high) and node $4j-2$ (low). The oscillator output nodes are connected by the resistive coupling scheme shown in Fig. 6.2(d); in this case, since the weight values are binary, these reduce to the two-resistor connections in Fig. 6.2(a) and Fig. 6.2(b). These connections are implemented

on lines 53 – 108 for the netlist below. The circuit uses a coupling resistance of $R_c = 200\,\Omega$, as specified on line 1.

Each oscillator additionally has two internal nodes: $4j-1$ (high) and $4j-4$ (low), across which the noise voltage source is connected. We designate the internal low node of oscillator 1 (node 0) as the ground reference. The noise sources for the N oscillators are specified on lines 109 – 116 in the netlist below. We use the `random` function in LTspice to generate a noise waveform. Since this is a deterministic function, we pass into its argument a random frequency (near 1 GHz) and a random phase, which are generated externally in MATLAB before every problem trial. In this example, the noise voltage source on each oscillator is a sum of five such random noise sources, each of which has an amplitude of $20\,\mu\text{V}/\sqrt{5}$. The circuit is therefore instantiated with 80 random parameters for this example, which introduce some degree of stochasticity.

Listing G.1: LTspice netlist for the Ising machine that solves the problem in Equation (G.1).

```
1  .param Rc 200
2  .params C0=1.5915e-10 L0=1.5915e-10 R0=1e-09
3  .params omegaPump=2*(2*pi)*1000000000 G0=0.025 dG
       ↪ =0.01 Tramp=1e-06 Vn=8.9443e-06
4  .model idealDiode D(Ron=0.0 Roff=1Gig Vfwd=0.01)
5  L001 N003 N001 {L0} Rser=0 Rpar=1e12
6  C001 N001 N002 {C0} Rser=0 Rpar=1e12 RLShunt=0
7  R001 0 N002 {R0}
8  Df001 N001 N002 idealDiode
9  Dr001 N002 N001 idealDiode
10 Cp001 N001 N002 Q=({G0}+{dG}*time/{Tramp})*{C0}*
       ↪ cos({omegaPump}*time)*x
11 L002 N007 N005 {L0} Rser=0 Rpar=1e12
12 C002 N005 N006 {C0} Rser=0 Rpar=1e12 RLShunt=0
13 R002 N004 N006 {R0}
14 Df002 N005 N006 idealDiode
15 Dr002 N006 N005 idealDiode
16 Cp002 N005 N006 Q=({G0}+{dG}*time/{Tramp})*{C0}*
       ↪ cos({omegaPump}*time)*x
17 L003 N011 N009 {L0} Rser=0 Rpar=1e12
18 C003 N009 N010 {C0} Rser=0 Rpar=1e12 RLShunt=0
19 R003 N008 N010 {R0}
20 Df003 N009 N010 idealDiode
21 Dr003 N010 N009 idealDiode
22 Cp003 N009 N010 Q=({G0}+{dG}*time/{Tramp})*{C0}*
       ↪ cos({omegaPump}*time)*x
23 L004 N015 N013 {L0} Rser=0 Rpar=1e12
24 C004 N013 N014 {C0} Rser=0 Rpar=1e12 RLShunt=0
25 R004 N012 N014 {R0}
26 Df004 N013 N014 idealDiode
27 Dr004 N014 N013 idealDiode
28 Cp004 N013 N014 Q=({G0}+{dG}*time/{Tramp})*{C0}*
       ↪ cos({omegaPump}*time)*x
29 L005 N019 N017 {L0} Rser=0 Rpar=1e12
30 C005 N017 N018 {C0} Rser=0 Rpar=1e12 RLShunt=0
31 R005 N016 N018 {R0}
32 Df005 N017 N018 idealDiode
33 Dr005 N018 N017 idealDiode
34 Cp005 N017 N018 Q=({G0}+{dG}*time/{Tramp})*{C0}*
       ↪ cos({omegaPump}*time)*x
35 L006 N023 N021 {L0} Rser=0 Rpar=1e12
36 C006 N021 N022 {C0} Rser=0 Rpar=1e12 RLShunt=0
37 R006 N020 N022 {R0}
38 Df006 N021 N022 idealDiode
39 Dr006 N022 N021 idealDiode
40 Cp006 N021 N022 Q=({G0}+{dG}*time/{Tramp})*{C0}*
       ↪ cos({omegaPump}*time)*x
41 L007 N027 N025 {L0} Rser=0 Rpar=1e12
42 C007 N025 N026 {C0} Rser=0 Rpar=1e12 RLShunt=0
43 R007 N024 N026 {R0}
44 Df007 N025 N026 idealDiode
45 Dr007 N026 N025 idealDiode
46 Cp007 N025 N026 Q=({G0}+{dG}*time/{Tramp})*{C0}*
       ↪ cos({omegaPump}*time)*x
47 L008 N031 N029 {L0} Rser=0 Rpar=1e12
48 C008 N029 N030 {C0} Rser=0 Rpar=1e12 RLShunt=0
49 R008 0 N030 {R0}
50 Df008 N029 N030 idealDiode
51 Dr008 N030 N029 idealDiode
52 Cp008 N029 N030 Q=({G0}+{dG}*time/{Tramp})*{C0}*
       ↪ cos({omegaPump}*time)*x
53 Ra001_002 N001 N005 {Rc}
54 Rb001_002 N002 N006 {Rc}
55 Ra001_003 N001 N009 {Rc}
56 Rb001_003 N002 N010 {Rc}
57 Ra001_004 N001 N014 {Rc}
58 Rb001_004 N002 N013 {Rc}
59 Ra001_005 N001 N018 {Rc}
60 Rb001_005 N002 N017 {Rc}
61 Ra001_006 N001 N022 {Rc}
62 Rb001_006 N002 N021 {Rc}
63 Ra001_007 N001 N025 {Rc}
64 Rb001_007 N002 N026 {Rc}
65 Ra001_008 N001 N029 {Rc}
66 Rb001_008 N002 N030 {Rc}
67 Ra002_003 N005 N009 {Rc}
68 Rb002_003 N006 N010 {Rc}
69 Ra002_004 N005 N013 {Rc}
70 Rb002_004 N006 N014 {Rc}
71 Ra002_005 N005 N017 {Rc}
72 Rb002_005 N006 N018 {Rc}
73 Ra002_006 N005 N021 {Rc}
```

```
Rb002_006 N006 N022 {Rc}
Ra002_007 N005 N025 {Rc}
Rb002_007 N006 N026 {Rc}
Ra002_008 N005 N029 {Rc}
Rb002_008 N006 N030 {Rc}
Ra003_004 N009 N014 {Rc}
Rb003_004 N010 N013 {Rc}
Ra003_005 N009 N018 {Rc}
Rb003_005 N010 N017 {Rc}
Ra003_006 N009 N021 {Rc}
Rb003_006 N010 N022 {Rc}
Ra003_007 N009 N026 {Rc}
Rb003_007 N010 N025 {Rc}
Ra003_008 N009 N029 {Rc}
Rb003_008 N010 N030 {Rc}
Ra004_005 N013 N018 {Rc}
Rb004_005 N014 N017 {Rc}
Ra004_006 N013 N022 {Rc}
Rb004_006 N014 N021 {Rc}
Ra004_007 N013 N025 {Rc}
Rb004_007 N014 N026 {Rc}
Ra004_008 N013 N030 {Rc}
Rb004_008 N014 N029 {Rc}
Ra005_006 N017 N022 {Rc}
Rb005_006 N018 N021 {Rc}
Ra005_007 N017 N026 {Rc}
Rb005_007 N018 N025 {Rc}
Ra005_008 N017 N029 {Rc}
Rb005_008 N018 N030 {Rc}
Ra006_007 N021 N026 {Rc}
Rb006_007 N022 N025 {Rc}
Ra006_008 N021 N029 {Rc}
Rb006_008 N022 N030 {Rc}
Ra007_008 N025 N030 {Rc}
Rb007_008 N026 N029 {Rc}
B001 N003 0 V=2*{Vn}*(random(7.046e+09*time
    ↪ +(0.8005))+random(6.318e+09*time+(2.148))+
    ↪ random(6.405e+09*time+(0.06335))+random
    ↪ (5.549e+09*time+(-2.1))+random(6.998e+09*
    ↪ time+(1.347))-0.5*5)
B002 N007 N004 V=2*{Vn}*(random(6.35e+09*time
    ↪ +(2.557))+random(6.462e+09*time+(-1.768))+
    ↪ random(6.087e+09*time+(2.331))+random
    ↪ (6.839e+09*time+(-1.811))+random(5.21e+09*
    ↪ time+(2.115))-0.5*5)
B003 N011 N008 V=2*{Vn}*(random(6.522e+09*time
    ↪ +(2.258))+random(7.193e+09*time+(0.1469))+
    ↪ random(6.155e+09*time+(-0.1422))+random
    ↪ (6.667e+09*time+(2.45))+random(5.789e+09*
    ↪ time+(-2.733))-0.5*5)
B004 N015 N012 V=2*{Vn}*(random(6.553e+09*time
    ↪ +(0.05941))+random(5.728e+09*time+(0.759))
    ↪ +random(7.036e+09*time+(1.468))+random
    ↪ (7.028e+09*time+(-1.696))+random(7.424e
    ↪ +09*time+(-3.004))-0.5*5)
B005 N019 N016 V=2*{Vn}*(random(6.143e+09*time
    ↪ +(-2.268))+random(6.175e+09*time+(1.693))+
    ↪ random(6.534e+09*time+(2.952))+random
    ↪ (7.144e+09*time+(-0.7113))+random(5.105e
    ↪ +09*time+(3.1))-0.5*5)
B006 N023 N020 V=2*{Vn}*(random(5.497e+09*time
    ↪ +(-1.091))+random(7.398e+09*time+(-2.28))+
    ↪ random(7.409e+09*time+(-0.7241))+random
    ↪ (6.165e+09*time+(0.3936))+random(7.064e
    ↪ +09*time+(0.8409))-0.5*5)
B007 N027 N024 V=2*{Vn}*(random(7.361e+09*time
    ↪ +(0.2614))+random(6.717e+09*time+(-1.162))
    ↪ +random(5.962e+09*time+(-2.141))+random
    ↪ (6.046e+09*time+(-2.183))+random(6.129e
    ↪ +09*time+(-2.281))-0.5*5)
B008 N031 0 V=2*{Vn}*(random(6.732e+09*time
    ↪ +(1.318))+random(6.195e+09*time+(-0.2208))
    ↪ +random(7.422e+09*time+(-2.43))+random
    ↪ (5.918e+09*time+(1.262))+random(5.879e+09*
    ↪ time+(-2.011))-0.5*5)
.tran 0.0001ns 1000ns
.save V(*)
.backanno
.end
```

The LTspice netlists for the 32-spin Ising problem (Fig. 6.7) and the 60-spin Max-Cut problem (Fig. 6.8) shown in Section 6.4 can be found as downloadable files at `https://github.com/ptxiao/analogIsing`. The full coupling matrix for the 32-spin problem is also available at the repository. The list of edge weights for the Max-Cut problem can found at `http://biqmac.uni-klu.ac.at/library/mac/rudy/g05_60.2`.

For further verifiability, Fig. G.1 and Fig. G.2 show the oscillator voltage waveforms returned by LTspice for the 32-oscillator and 60-oscillator Ising machines, respectively. The amplitude and phase profiles in Fig. 6.7 and Fig. 6.8 are extracted from these waveforms.

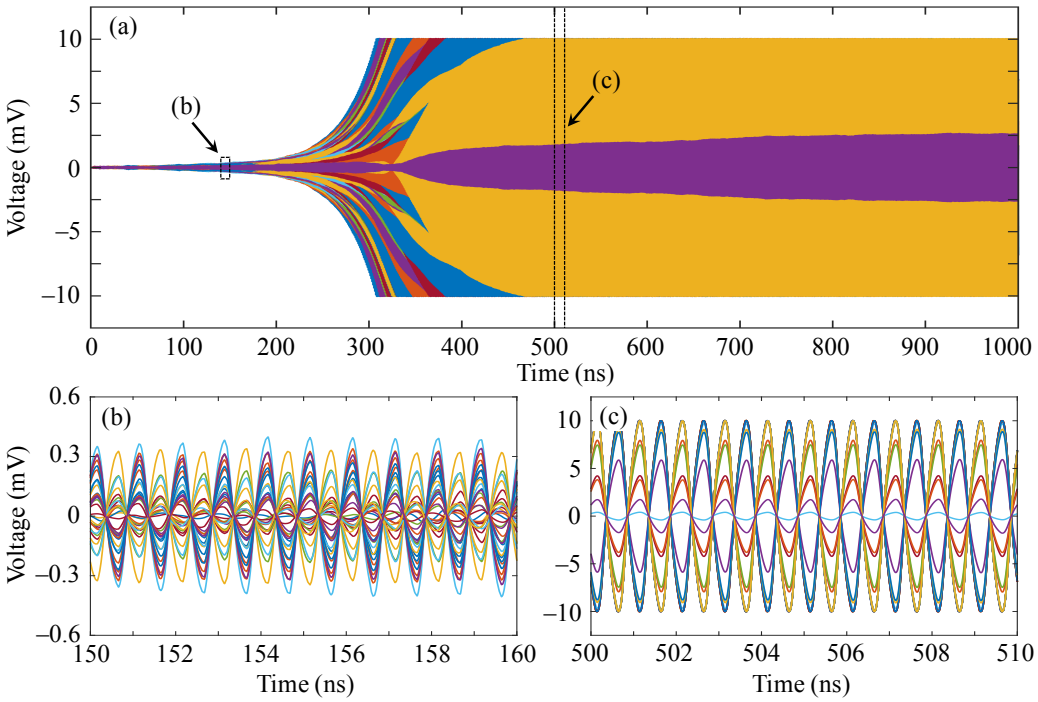

Figure G.1: Oscillator output voltage waveforms corresponding to the results in Fig. 6.7. The oscillator color labels are not identical to those in Fig. 6.7(b) and (c).

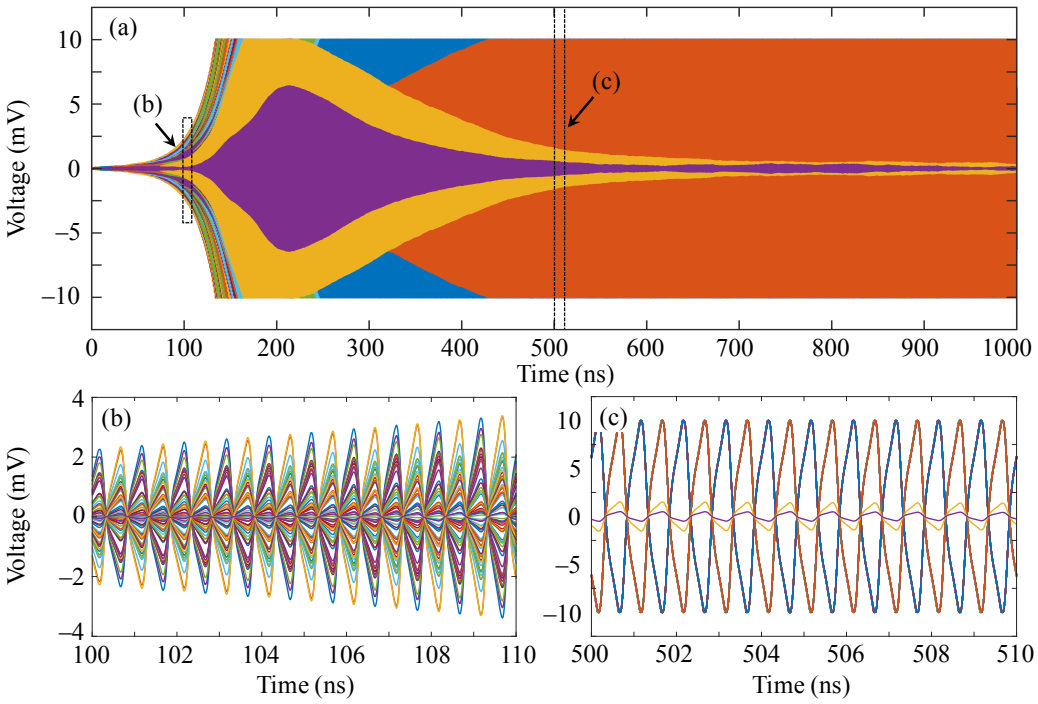

Figure G.2: Oscillator output voltage waveforms corresponding to the results in Fig. 6.8. The oscillator color labels are not identical to those in Fig. 6.8(b) and (c).

Bibliography

[1] T. P. Xiao, O. S. Cifci, S. Bhargava, H. Chen, T. Gissibl, W. Zhou, H. Giessen, K. C. Toussaint, E. Yablonovitch, and P. V. Braun, "Diffractive spectral-splitting optical element designed by adjoint-based electromagnetic optimization and fabricated by femtosecond 3D direct laser writing," *ACS Photonics*, vol. 3, no. 5, pp. 886–894, 2016.

[2] J. Penning, K. Stober, V. Taylor, and M. Yamada, "Energy savings forecast of solid-state lighting in general illumination applications," Office of Energy Efficiency & Renewable Energy, U.S. Department of Energy: Washington, D.C., Tech. Rep., 2016.

[3] G. Masson and I. Kaizuka, "Trends 2018 in photovoltaic applications," International Energy Agency, Photovoltaic Power Systems Programme, Tech. Rep., 2018.

[4] M. H. Crawford, J. J. Wierer, A. J. Fischer, G. T. Wang, D. D. Koleske, G. S. Subramania, M. E. Coltrin, R. F. Karlicek, and J. Y. Tsao, "Solid-state lighting: Toward smart and ultra-efficient materials, devices, lamps, and systems," in *Photonics*. Wiley-Blackwell, 2015, ch. 1, p. 2.

[5] O. D. Miller, E. Yablonovitch, and S. R. Kurtz, "Strong internal and external luminescence as solar cells approach the shockley-queisser limit," *IEEE Journal of Photovoltaics*, vol. 2, no. 3, pp. 303–311, 2012.

[6] K. Lehovec, C. A. Accardo, and E. Jamgochian, "Light emission produced by current injected into a green silicon-carbide crystal," *Phys. Rev.*, vol. 89, pp. 20–25, 1 1953.

[7] "Residential Energy Consumption Survey," U.S. Energy Information Administration, U.S. Department of Energy: Washington, D.C., Tech. Rep., 2015.

[8] M. Sheik-Bahae and R. I. Epstein, "Optical refrigeration," *Nature Photonics*, vol. 1, pp. 693–699, 12 2007.

[9] S. Vichik and F. Borrelli, "Solving linear and quadratic programs with an analog circuit," *Computers & Chemical Engineering*, vol. 70, pp. 160 –171, 2014.

[10] J. Dean, D. Patterson, and C. Young, "A new golden age in computer architecture: Empowering the machine-learning revolution," *IEEE Micro*, vol. 38, no. 2, pp. 21–29, 2018.

[11] T. P. Xiao, K. Chen, P. Santhanam, S. Fan, and E. Yablonovitch, "Electroluminescent refrigeration by ultra-efficient GaAs light-emitting diodes," *Journal of Applied Physics*, vol. 123, no. 17, p. 173 104, 2018.

[12] G. Kirchhoff, "Ueber das verhältniss zwischen dem emissionsvermögen und dem absorptionsvermögen der körper für wärme und licht," *Annalen der Physik*, vol. 185, no. 2, pp. 275–301, 1860.

[13] R. T. Ross, "Some thermodynamics of photochemical systems," *The Journal of Chemical Physics*, vol. 46, no. 12, pp. 4590–4593, 1967.

[14] M. R. Krames, O. B. Shchekin, R. Mueller-Mach, G. O. Mueller, L. Zhou, G. Harbers, and M. G. Craford, "Status and future of high-power light-emitting diodes for solid-state lighting," *Journal of Display Technology*, vol. 3, no. 2, pp. 160–175, 2007.

[15] M. A. Green, "Radiative efficiency of state-of-the-art photovoltaic cells," *Progress in Photovoltaics: Research and Applications*, vol. 20, no. 4, pp. 472–476, 2012.

[16] E. Schubert, *Light-Emitting Diodes*. Cambridge University Press, 2006.

[17] M. A. Steiner, J. F. Geisz, I. Garca, D. J. Friedman, A. Duda, and S. R. Kurtz, "Optical enhancement of the open-circuit voltage in high quality GaAs solar cells," *Journal of Applied Physics*, vol. 113, no. 12, p. 123 109, 2013.

[18] L. M. Pazos-Outón, M. Szumilo, R. Lamboll, J. M. Richter, M. Crespo-Quesada, M. Abdi-Jalebi, H. J. Beeson, M. Vrućinić, M. Alsari, H. J. Snaith, B. Ehrler, R. H. Friend, and F. Deschler, "Photon recycling in lead iodide perovskite solar cells," *Science*, vol. 351, no. 6280, pp. 1430–1433, 2016.

[19] M. A. Green, Y. Hishikawa, E. D. Dunlop, D. H. Levi, J. Hohl-Ebinger, M. Yoshita, and A. W. Ho-Baillie, "Solar cell efficiency tables (version 53)," *Progress in Photovoltaics: Research and Applications*, vol. 27, no. 1, pp. 3–12, 2019.

[20] A. Richter, M. Hermle, and S. Glunz, "Reassessment of the limiting efficiency for crystalline silicon solar cells," *IEEE Journal of Photovoltaics*, vol. 3, pp. 1184–1191, Jul. 2013.

[21] L. M. Pazos-Outón, T. P. Xiao, and E. Yablonovitch, "Fundamental efficiency limit of lead iodide perovskite solar cells," *The Journal of Physical Chemistry Letters*, vol. 9, no. 7, pp. 1703–1711, 2018.

[22] O. D. Miller. (2013). ShockleyQueisser – MATLAB File Exchange, [Online]. Available: `https://www.mathworks.com/matlabcentral/fileexchange/43879-shockleyqueisser` (visited on 01/30/2019).

[23] W. Shockley and H. J. Queisser, "Detailed balance limit of efficiency of *p-n* junction solar cells," *Journal of Applied Physics*, vol. 32, no. 3, pp. 510–519, 1961.

[24] M. A. Green, K. Emery, Y. Hishikawa, and W. Warta, "Solar cell efficiency tables (version 31)," *Progress in Photovoltaics: Research and Applications*, vol. 16, no. 1, pp. 61–67, 2007.

[25] M. A. Steiner, J. F. Geisz, J. S. Ward, I. Garca, D. J. Friedman, R. R. King, P. T. Chiu, R. M. France, A. Duda, W. J. Olavarria, M. Young, and S. R. Kurtz, "Optically enhanced photon recycling in mechanically stacked multijunction solar cells," *IEEE Journal of Photovoltaics*, vol. 6, no. 1, pp. 358–365, 2016.

[26] U. Rau, "Reciprocity relation between photovoltaic quantum efficiency and electroluminescent emission of solar cells," *Phys. Rev. B*, vol. 76, p. 085 303, 8 2007.

[27] E. Yablonovitch, T. Gmitter, J. P. Harbison, and R. Bhat, "Extreme selectivity in the lift-off of epitaxial GaAs films," *Applied Physics Letters*, vol. 51, no. 26, pp. 2222–2224, 1987.

[28] V. Ganapati, M. A. Steiner, and E. Yablonovitch, "The voltage boost enabled by luminescence extraction in solar cells," *IEEE Journal of Photovoltaics*, vol. 6, no. 4, pp. 801–809, 2016.

[29] V. Ganapati, T. P. Xiao, and E. Yablonovitch, "Ultra-efficient thermophotovoltaics exploiting spectral filtering by the photovoltaic band-edge," *ArXiv e-prints*, 2016. arXiv: `1611.03544 [physics.optics]`.

[30] I. Celanovic, N. Jovanovic, and J. Kassakian, "Two-dimensional tungsten photonic crystals as selective thermal emitters," *Applied Physics Letters*, vol. 92, no. 19, p. 193 101, 2008.

[31] E. S. Sakr, Z. Zhou, and P. Bermel, "High efficiency rare-earth emitter for thermophotovoltaic applications," *Applied Physics Letters*, vol. 105, no. 11, p. 111 107, 2014.

[32] R. M. Swanson, "Recent developments in thermophotovoltaic conversion," ser. International Electron Devices Meeting, vol. 26, 1980, pp. 186–189.

[33] B. Wernsman, R. R. Siergiej, S. D. Link, R. G. Mahorter, M. N. Palmisiano, R. J. Wehrer, R. W. Schultz, G. P. Schmuck, R. L. Messham, S. Murray, C. S. Murray, F. Newman, D. Taylor, D. M. DePoy, and T. Rahmlow, "Greater than 20% radiant heat conversion efficiency of a thermophotovoltaic radiator/module system using reflective spectral control," *IEEE Transactions on Electron Devices*, vol. 51, no. 3, pp. 512–515, 2004.

[34] R. K. Ahrenkiel, R. Ellingson, S. Johnston, and M. Wanlass, "Recombination lifetime of $In_{0.53}Ga_{0.47}As$ as a function of doping density," *Applied Physics Letters*, vol. 72, no. 26, pp. 3470–3472, 1998.

[35] Z. Omair, G. Scranton, L. M. Pazos-Outón, T. P. Xiao, M. A. Steiner, V. Ganapati, P. F. Peterson, J. Holzrichter, and E. Yablonovitch, "Ultra-efficient thermophotovoltaic power conversion by band-edge spectral filtering," *Proceedings of the National Academy of Sciences (Submitted)*, 2019.

[36] H. Gauck, T. H. Gfroerer, M. J. Renn, E. A. Cornell, and K. A. Bertness, "External radiative quantum efficiency of 96% from a GaAs/GaInP heterostructure," *Applied Physics A*, vol. 64, no. 2, pp. 143–147, 1997.

[37] D. A. Bender, J. G. Cederberg, C. Wang, and M. Sheik-Bahae, "Development of high quantum efficiency GaAs/GaInP double heterostructures for laser cooling," *Applied Physics Letters*, vol. 102, no. 25, p. 252 102, 2013.

[38] P. S. Markus Broell, A. Rudolph, W. Schmid, A. Vogl, and M. Behringer, "New developments on high-efficiency infrared and InGaAlP light-emitting diodes at OSRAM Opto Semiconductors," *Proc. SPIE*, vol. 9003, p. 90030L, 2014.

[39] C. A. Hurni, A. David, M. J. Cich, R. I. Aldaz, B. Ellis, K. Huang, A. Tyagi, R. A. DeLille, M. D. Craven, F. M. Steranka, and M. R. Krames, "Bulk GaN flip-chip violet light-emitting diodes with optimized efficiency for high-power operation," *Applied Physics Letters*, vol. 106, no. 3, p. 031 101, 2015.

[40] B. Imangholi, M. P. Hasselbeck, M. Sheik-Bahae, R. I. Epstein, and S. Kurtz, "Effects of epitaxial lift-off on interface recombination and laser cooling in GaInP/GaAs heterostructures," *Applied Physics Letters*, vol. 86, no. 8, p. 081 104, 2005.

[41] I. Schnitzer, E. Yablonovitch, C. Caneau, T. J. Gmitter, and A. Scherer, "30% external quantum efficiency from surface textured, thin-film light-emitting diodes," *Applied Physics Letters*, vol. 63, no. 16, pp. 2174–2176, 1993.

[42] M. H. MacDougal, H. Zhao, P. D. Dapkus, M. Ziari, and W. H. Steier, "Wide-bandwidth distributed bragg reflectors using oxide/GaAs multilayers," *Electronics Letters*, vol. 30, no. 14, pp. 1147–1149, 1994.

[43] C. L. Chua, R. L. Thornton, and D. W. Treat, "Planar laterally oxidized vertical-cavity lasers for low-threshold high-density top-surface-emitting arrays," *IEEE Photonics Technology Letters*, vol. 9, no. 8, pp. 1060–1062, 1997.

[44] M. A. Haase, M. J. Hafich, and G. Y. Robinson, "Internal photoemission and energy-band offsets in GaAs/GaInP *p-I-N* heterojunction photodiodes," *Applied Physics Letters*, vol. 58, no. 6, pp. 616–618, 1991.

[45] S. Chuang, *Physics of Photonic Devices*, ser. Wiley Series in Pure and Applied Optics. Wiley, 2012, ISBN: 9781118585658.

[46] W. van Roosbroeck and W. Shockley, "Photon-radiative recombination of electrons and holes in germanium," *Phys. Rev.*, vol. 94, pp. 1558–1560, 6 1954.

[47] M. D. Sturge, "Optical absorption of Gallium Arsenide between 0.6 and 2.75 eV," *Phys. Rev.*, vol. 127, pp. 768–773, 3 1962.

[48] I. Schnitzer, E. Yablonovitch, C. Caneau, and T. J. Gmitter, "Ultrahigh spontaneous emission quantum efficiency, 99.7% internally and 72% externally, from AlGaAs/GaAs/AlGaAs double heterostructures," *Applied Physics Letters*, vol. 62, no. 2, pp. 131–133, 1993.

[49] J. M. Olson, R. K. Ahrenkiel, D. J. Dunlavy, B. Keyes, and A. E. Kibbler, "Ultralow recombination velocity at $Ga_{0.5}In_{0.5}P$/GaAs heterointerfaces," *Applied Physics Letters*, vol. 55, no. 12, pp. 1208–1210, 1989.

[50] E. Yablonovitch, T. J. Gmitter, and B. G. Bagley, "As_2S_3/GaAs, a new amorphous/crystalline heterojunction for the III-V semiconductors," *Applied Physics Letters*, vol. 57, no. 21, pp. 2241–2243, 1990.

[51] U. Strauss, W. W. Rühle, and K. Köhler, "Auger recombination in intrinsic GaAs," *Applied Physics Letters*, vol. 62, no. 1, pp. 55–57, 1993.

[52] A. Grove, *Physics and Technology of Semiconductor Devices*, ser. Wiley International edition. Wiley, 1967.

[53] E. Yablonovitch, "Statistical ray optics," *J. Opt. Soc. Am.*, vol. 72, no. 7, pp. 899–907, 1982.

[54] A. David, "Surface-roughened light-emitting diodes: An accurate model," *Journal of Display Technology*, vol. 9, no. 5, pp. 301–316, 2013.

[55] H. W. Deckman, C. B. Roxlo, and E. Yablonovitch, "Maximum statistical increase of optical absorption in textured semiconductor films," *Opt. Lett.*, vol. 8, no. 9, pp. 491–493, 1983.

[56] P. Wurfel, "The chemical potential of radiation," *Journal of Physics C: Solid State Physics*, vol. 15, no. 18, p. 3967, 1982.

[57] S. Essig, S. Ward, M. A. Steiner, D. J. Friedman, J. F. Geisz, P. Stradins, and D. L. Young, "Progress towards a 30% efficient gainp/si tandem solar cell," *Energy Procedia*, vol. 77, no. Supplement C, pp. 464–469, 2015.

[58] J. N. Winn, Y. Fink, S. Fan, and J. D. Joannopoulos, "Omnidirectional reflection from a one-dimensional photonic crystal," *Opt. Lett.*, vol. 23, no. 20, pp. 1573–1575, 1998.

[59] H. Casey and M. Panish, *Heterostructure Lasers: Fundamental Principles*, ser. Quantum electronics. Academic Press, 1978, ISBN: 9780121631017.

[60] R. M. Swanson, S. K. Beckwith, R. A. Crane, W. D. Eades, Y. H. Kwark, R. A. Sinton, and S. E. Swirhun, "Point-contact silicon solar cells," *IEEE Transactions on Electron Devices*, vol. 31, no. 5, pp. 661–664, 1984.

[61] A. Olsson, J. Tiira, M. Partanen, T. Hakkarainen, E. Koivusalo, A. Tukiainen, M. Guina, and J. Oksanen, "Optical energy transfer and loss mechanisms in coupled intracavity light emitters," *IEEE Transactions on Electron Devices*, vol. 63, no. 9, pp. 3567–3573, 2016.

[62] A. G. Baca, F. Ren, J. C. Zolper, R. D. Briggs, and S. J. Pearton, "A survey of ohmic contacts to iii-v compound semiconductors," *Thin Solid Films*, vol. 308-309, no. Supplement C, pp. 599 –606, 1997.

[63] J. Wang, J. Kapraun, N. Cabello, P. Tingzon, K. Cook, J. Qi, E. Kolev, and C. J. Chang-Hasnain, "Single-mode buried InGaP aperture VCSEL emitting at 980 nm," in *Compound Semiconductor Week 2018*, We2A2:Lasers, 2018.

[64] J. Tauc, "The share of thermal energy taken from the surroundings in the electroluminescent energy radiated from a *p-n* junction," *Cechoslovackij fiziceskij zurnal*, vol. 7, no. 3, pp. 275–276, 1957.

[65] M. A. Weinstein, "Thermodynamic limitation on the conversion of heat into light," *J. Opt. Soc. Am.*, vol. 50, no. 6, pp. 597–602, 1960.

[66] P. Berdahl, "Radiant refrigeration by semiconductor diodes," *Journal of Applied Physics*, vol. 58, no. 3, pp. 1369–1374, 1985.

[67] J. Xue, Y. Zhao, S.-H. Oh, W. F. Herrington, J. S. Speck, S. P. DenBaars, S. Nakamura, and R. J. Ram, "Thermally enhanced blue light-emitting diode," *Applied Physics Letters*, vol. 107, no. 12, p. 121 109, 2015.

[68] K. Chen, T. P. Xiao, P. Santhanam, E. Yablonovitch, and S. Fan, "High-performance near-field electroluminescent refrigeration device consisting of a GaAs light emitting diode and a Si photovoltaic cell," *Journal of Applied Physics*, vol. 122, no. 14, p. 143 104, 2017.

[69] L. Zhu, A. Fiorino, D. Thompson, R. Mittapally, E. Meyhofer, and P. Reddy, "Near-field photonic cooling through control of the chemical potential of photons," *Nature*, vol. 566, no. 7743, pp. 239–244, 2019.

[70] K. P. Pipe, R. J. Ram, and A. Shakouri, "Bias-dependent peltier coefficient and internal cooling in bipolar devices," *Phys. Rev. B*, vol. 66, p. 125 316, 12 2002.

[71] J. Piprek and Z.-M. Li, "Electroluminescent cooling mechanism in InGaN/GaN light-emitting diodes," *Optical and Quantum Electronics*, vol. 48, no. 10, p. 472, 2016.

[72] M. I. Nathan, T. N. Morgan, G. Burns, and A. E. Michel, "High-energy emission in GaAs electroluminescent diodes," *Phys. Rev.*, vol. 146, pp. 570–574, 2 1966.

[73] J. I. Pankove, "Blue anti-Stokes electroluminescence in GaN," *Phys. Rev. Lett.*, vol. 34, pp. 809–812, 13 1975.

[74] P. Santhanam, D. J. Gray, and R. J. Ram, "Thermoelectrically pumped light-emitting diodes operating above unity efficiency," *Phys. Rev. Lett.*, vol. 108, p. 097 403, 9 2012.

[75] P. Santhanam, D. Huang, R. J. Ram, M. A. Remennyi, and B. A. Matveev, "Room temperature thermo-electric pumping in mid-infrared light-emitting diodes," *Applied Physics Letters*, vol. 103, no. 18, p. 183 513, 2013.

[76] W. Tennant, D. Lee, M. Zandian, E. Piquette, and M. Carmody, "Mbe hgcdte technology: A very general solution to ir detection, described by "rule 07", a very convenient heuristic," *Journal of Electronic Materials*, vol. 37, no. 9, pp. 1406–1410, 2008.

[77] N.-P. Harder and M. A. Green, "Thermophotonics," *Semiconductor Science and Technology*, vol. 18, no. 5, S270, 2003.

[78] J. Oksanen and J. Tulkki, "Thermophotonic heat pump - theoretical model and numerical simulations," *Journal of Applied Physics*, vol. 107, no. 9, 093106, 2010.

[79] D. Polder and M. Van Hove, "Theory of radiative heat transfer between closely spaced bodies," *Phys. Rev. B*, vol. 4, pp. 3303–3314, 10 1971.

[80] T. Markvart, "Thermodynamics of losses in photovoltaic conversion," *Applied Physics Letters*, vol. 91, no. 6, p. 064 102, 2007.

[81] A. Manor, N. Kruger, T. Sabapathy, and C. Rotschild, "Thermally enhanced photoluminescence for heat harvesting in photovoltaics," *Nature Communications*, vol. 7, no. 13167, 2016.

[82] H. B. Callen, "The application of Onsager's reciprocal relations to thermoelectric, thermomagnetic, and galvanomagnetic effects," *Phys. Rev.*, vol. 73, pp. 1349–1358, 11 1948.

[83] H. Goldsmid, *Thermoelectric refrigeration*, ser. International cryogenics monograph series. Plenum Press, 1964.

[84] T. J. Scheidemantel, C. Ambrosch-Draxl, T. Thonhauser, J. V. Badding, and J. O. Sofo, "Transport coefficients from first-principles calculations," *Phys. Rev. B*, vol. 68, p. 125 210, 12 2003.

[85] A. J. Minnich, M. S. Dresselhaus, Z. F. Ren, and G. Chen, "Bulk nanostructured thermoelectric materials: Current research and future prospects," *Energy Environ. Sci.*, vol. 2, pp. 466–479, 5 2009.

[86] D. Zhao and G. Tan, "A review of thermoelectric cooling: Materials, modeling and applications," *Applied Thermal Engineering*, vol. 66, no. 1-2, pp. 15 –24, 2014.

[87] J. Zhang, D. Li, R. Chen, and Q. Xiong, "Laser cooling of a semiconductor by 40 kelvin," *Nature*, vol. 493, no. 7433, pp. 504–508, 2013.

[88] S.-T. Ha, C. Shen, J. Zhang, and Q. Xiong, "Laser cooling of organic-inorganic lead halide perovskites," *Nature Photonics*, vol. 10, no. 2, pp. 115–121, 2016.

[89] A. Milnes, *Deep Impurities in Semiconductors*, ser. A Wiley-Interscience publication. Wiley, 1973, ISBN: 9780471606703.

[90] M. Takeshima, "Effect of Auger recombination on laser operation in $Ga_{1-x}Al_xAs$," *Journal of Applied Physics*, vol. 58, no. 10, pp. 3846–3850, 1985.

[91] M. Beaudoin, A. J. G. DeVries, S. R. Johnson, H. Laman, and T. Tiedje, "Optical absorption edge of semi-insulating GaAs and InP at high temperatures," *Applied Physics Letters*, vol. 70, no. 26, pp. 3540–3542, 1997.

[92] M. Sotoodeh, A. H. Khalid, and A. A. Rezazadeh, "Empirical low-field mobility model for III-V compounds applicable in device simulation codes," *Journal of Applied Physics*, vol. 87, no. 6, pp. 2890–2900, 2000.

[93] J. I. Pankove, *Optical Processes in Semiconductors*. Prentice Hall, 1971.

[94] B. Zhao, S. Bai, V. Kim, R. Lamboll, R. Shivanna, F. Auras, J. M. Richter, L. Yang, L. Dai, M. Alsari, X.-J. She, L. Liang, J. Zhang, S. Lilliu, P. Gao, H. J. Snaith, J. Wang, N. C. Greenham, R. H. Friend, and D. Di, "High-efficiency perovskite-polymer bulk heterostructure light-emitting diodes," *Nature Photonics*, vol. 12, pp. 783–789, 2018.

[95] S. Rytov, *Theory of Electric Fluctuations and Thermal Radiation*. Electronics Research Directorate, Air Force Cambridge Research Center, Air Research and Development Command, U.S. Air Force, 1959.

[96] R. M. Karp, "Reducibility among combinatorial problems," in *Complexity of Computer Computations*, R. E. Miller, J. W. Thatcher, and J. D. Bohlinger, Eds. Boston, MA: Springer US, 1972, pp. 85–103, ISBN: 978-1-4684-2001-2.

[97] H. Markowitz, "Portfolio selection," *The Journal of Finance*, vol. 7, no. 1, pp. 77–91, 1952.

[98] N. Sherwani, *Algorithms for VLSI Physical Design Automation*. Springer US, 2012, ISBN: 9781461523512.

[99] J. T. Ngo, J. Marks, and M. Karplus, "Computational complexity, protein structure prediction, and the Levinthal paradox," in *The Protein Folding Problem and Tertiary Structure Prediction*, K. M. Merz and S. M. Le Grand, Eds. Boston, MA: Birkhäuser Boston, 1994, pp. 433–506, ISBN: 978-1-4684-6831-1.

[100] J. D. Kececioglu and E. W. Myers, "Combinatorial algorithms for DNA sequence assembly," *Algorithmica*, vol. 13, no. 1, p. 7, 1995.

[101] D. B. Kitchen, H. Decornez, J. R. Furr, and J. Bajorath, "Docking and scoring in virtual screening for drug discovery: Methods and applications," *Nature Reviews Drug Discovery*, vol. 3, pp. 935–949, 2004.

[102] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, `http://www.deeplearningbook.org`.

[103] J. J. Hopfield and D. W. Tank, "'Neural' computation of decisions in optimization problems," *Biological Cybernetics*, vol. 52, no. 3, pp. 141–152, 1985.

[104] M. Yamaoka, C. Yoshimura, M. Hayashi, T. Okuyama, H. Aoki, and H. Mizuno, "A 20k-spin Ising chip to solve combinatorial optimization problems with CMOS annealing," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 1, pp. 303–309, 2016.

[105] T. Wang and J. Roychowdhury, "Oscillator-based Ising machine," *ArXiv e-prints*, 2017. arXiv: `1709.08102 [cs.ET]`.

[106] ——, "OIM: Oscillator-based Ising Machines for solving combinatorial optimisation problems," *ArXiv e-prints*, 2019. arXiv: `1903.07163 [cs.ET]`.

[107] Z. Wang, A. Marandi, K. Wen, R. L. Byer, and Y. Yamamoto, "Coherent Ising machine based on degenerate optical parametric oscillators," *Phys. Rev. A*, vol. 88, p. 063 853, 2013.

[108] A. Marandi, Z. Wang, K. Takata, R. L. Byer, and Y. Yamamoto, "Network of time-multiplexed optical parametric oscillators as a coherent ising machine," *Nature Photonics*, vol. 8, pp. 937–942, 2014.

[109] M. N. Bojnordi and E. Ipek, "Memristive Boltzmann machine: A hardware accelerator for combinatorial optimization and deep learning," in *2016 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2016, pp. 1–13.

[110] F. Cai, S. Kumar, T. V. Vaerenbergh, R. Liu, C. Li, S. Yu, Q. Xia, J. J. Yang, R. Beausoleil, W. Lu, and J. P. Strachan, "Harnessing intrinsic noise in memristor Hopfield neural networks for combinatorial optimization," *ArXiv e-prints*, 2019. arXiv: `1903.11194 [cs.ET]`.

[111] B. Sutton, K. Y. Camsari, B. Behin-Aein, and S. Datta, "Intrinsic optimization using stochastic nanomagnets," *Scientific Reports*, vol. 7, p. 44 370, 2017.

[112] M. Di Ventra and F. L. Traversa, "Perspective: Memcomputing: Leveraging memory and physics to compute efficiently," *Journal of Applied Physics*, vol. 123, no. 18, p. 180 901, 2018.

[113] E. Farhi, J. Goldstone, S. Gutmann, J. Lapan, A. Lundgren, and D. Preda, "A quantum adiabatic evolution algorithm applied to random instances of an NP-complete problem," *Science*, vol. 292, no. 5516, pp. 472–475, 2001.

[114] S. Boixo, T. F. Rønnow, S. V. Isakov, Z. Wang, D. Wecker, D. A. Lidar, J. M. Martinis, and M. Troyer, "Evidence for quantum annealing with more than one hundred qubits," *Nature Physics*, vol. 10, pp. 218–224, 2014.

[115] S. Utsunomiya, K. Takata, and Y. Yamamoto, "Mapping of Ising models onto injection-locked laser systems," *Opt. Express*, vol. 19, no. 19, pp. 18 091–18 108, 2011.

[116] G. Dantzig, R. Fulkerson, and S. Johnson, "Solution of a large-scale traveling-salesman problem," *Journal of the Operations Research Society of America*, vol. 2, no. 4, pp. 393–410, 1954.

[117] M. Garey, D. Johnson, and M. S. M. Collection, *Computers and Intractability: A Guide to the Theory of NP-completeness*. W. H. Freeman, 1979, ISBN: 9780716710448.

[118] V. Vazirani, *Approximation Algorithms*. Springer Berlin Heidelberg, 2002, ISBN: 9783540653677.

[119] M. X. Goemans and D. P. Williamson, "Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming," *J. ACM*, vol. 42, no. 6, pp. 1115–1145, Nov. 1995.

[120] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, pp. 671–680, 1983.

[121] V. Bapst, L. Foini, F. Krzakala, G. Semerjian, and F. Zamponi, "The quantum adiabatic algorithm applied to random optimization problems: The quantum spin glass perspective," *Physics Reports*, vol. 523, no. 3, pp. 127 –205, 2013.

[122] S. W. Shin, G. Smith, J. A. Smolin, and U. Vazirani, "How "quantum" is the D-Wave machine?" *ArXiv e-prints*, 2014. arXiv: `1401.7087 [quant-ph]`.

[123] T. F. Rønnow, Z. Wang, J. Job, S. Boixo, S. V. Isakov, D. Wecker, J. M. Martinis, D. A. Lidar, and M. Troyer, "Defining and detecting quantum speedup," *Science*, vol. 345, no. 6195, pp. 420–424, 2014.

[124] W. van Dam, M. Mosca, and U. Vazirani, "How powerful is adiabatic quantum computation?" In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, 2001, pp. 279–287.

[125] P. W. Shor, "Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer," *SIAM J. Comput.*, vol. 26, no. 5, pp. 1484–1509, Oct. 1997.

[126] C. Bennett, E. Bernstein, G. Brassard, and U. Vazirani, "Strengths and weaknesses of quantum computing," *SIAM Journal on Computing*, vol. 26, no. 5, pp. 1510–1523, 1997.

[127] P. L. McMahon, A. Marandi, Y. Haribara, R. Hamerly, C. Langrock, S. Tamate, T. Inagaki, H. Takesue, S. Utsunomiya, K. Aihara, R. L. Byer, M. M. Fejer, H. Mabuchi, and Y. Yamamoto, "A fully programmable 100-spin coherent Ising machine with all-to-all connections," *Science*, vol. 354, no. 6312, pp. 614–617, 2016.

[128] T. Inagaki, Y. Haribara, K. Igarashi, T. Sonobe, S. Tamate, T. Honjo, A. Marandi, P. L. McMahon, T. Umeki, K. Enbutsu, O. Tadanaga, H. Takenouchi, K. Aihara, K.-i. Kawarabayashi, K. Inoue, S. Utsunomiya, and H. Takesue, "A coherent Ising machine for 2000-node optimization problems," *Science*, vol. 354, no. 6312, pp. 603–606, 2016.

[129] T. Leleu, Y. Yamamoto, S. Utsunomiya, and K. Aihara, "Combinatorial optimization using dynamical phase transitions in driven-dissipative systems," *Phys. Rev. E*, vol. 95, p. 022 118, 2 2017.

[130] T. Leleu, Y. Yamamoto, P. L. McMahon, and K. Aihara, "Destabilization of local minima in analog spin systems by correction of amplitude heterogeneity," *Phys. Rev. Lett.*, vol. 122, p. 040 607, 4 2019.

[131] F Barahona, "On the computational complexity of Ising spin glass models," *Journal of Physics A: Mathematical and General*, vol. 15, no. 10, pp. 3241–3253, 1982.

[132] A. Lucas, "Ising formulations of many NP problems," *Frontiers in Physics*, vol. 2, p. 5, 2014.

[133] K. C. Young, R. Blume-Kohout, and D. A. Lidar, "Adiabatic quantum optimization with the wrong Hamiltonian," *Phys. Rev. A*, vol. 88, p. 062 314, 6 2013.

[134] I. Mahboob, H. Okamoto, and H. Yamaguchi, "An electromechanical Ising Hamiltonian," *Science Advances*, vol. 2, no. 6, 2016.

[135] S. Hong, *Wireless: From Marconi's Black-box to the Audion*, ser. Transformations (M.I.T. Press). MIT Press, 2001, ISBN: 9780262275637.

[136] H. Heffner and G. Wade, "Gain, band width, and noise characteristics of the variable-parameter amplifier," *Journal of Applied Physics*, vol. 29, no. 9, pp. 1321–1331, 1958.

[137] J. M. Manley and H. E. Rowe, "Some general properties of nonlinear elements – part i. general energy relations," *Proceedings of the IRE*, vol. 44, no. 7, pp. 904–913, 1956.

[138] E. Goto, "The parametron, a digital computing element which utilizes parametric oscillation," *Proceedings of the IRE*, vol. 47, no. 8, pp. 1304–1316, 1959.

[139] J. von Neumann, "Non-linear capacitance or inductance switching, amplifying, and memory organs," pat. US2815488A, 1954.

[140] T. Wang and J. Roychowdhury, "Phlogon: Phase-based logic using oscillatory nano-systems," in *Unconventional Computation and Natural Computation*, O. H. Ibarra, L. Kari, and S. Kopecki, Eds., Cham: Springer International Publishing, 2014, pp. 353–366, ISBN: 978-3-319-08123-6.

[141] "Crossbar ReRAM Technology," Crossbar, Inc, Tech. Rep. [Online]. Available: `https://www.crossbar-inc.com/assets/resources/white-papers/Crossbar-ReRAM-Technology.pdf`.

[142] T. M. Taha, R. Hasan, C. Yakopcic, and M. R. McLean, "Exploring the design space of specialized multicore neural processors," in *The 2013 International Joint Conference on Neural Networks (IJCNN)*, 2013, pp. 1–8.

[143] E. S. Tiunov, A. E. Ulanov, and A. I. Lvovsky, "Annealing by simulating the coherent Ising machine," *Opt. Express*, vol. 27, no. 7, pp. 10 288–10 295, 2019.

[144] A. Wiegele, "Biq mac library – a collection of Max-Cut and quadratic 0-1 programming instances of medium size," Alpen-Adria-Universität Klagenfurt, Tech. Rep., 2007.

[145] R. Hamerly, T. Inagaki, P. L. McMahon, D. Venturelli, A. Marandi, T. Onodera, E. Ng, C. Langrock, K. Inaba, T. Honjo, K. Enbutsu, T. Umeki, R. Kasahara, S. Utsunomiya, S. Kako, K. ichi Kawarabayashi, R. L. Byer, M. M. Fejer, H. Mabuchi, D. Englund, E. Rieffel, H. Takesue, and Y. Yamamoto, "Experimental investigation of performance differences between Coherent Ising Machines and a quantum annealer," *ArXiv e-prints*, 2018. arXiv: `1805.05217 [quant-ph]`.

[146] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun, "The loss surfaces of multilayer networks," in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, G. Lebanon and S. V. N. Vishwanathan, Eds., ser. Proceedings of Machine Learning Research, vol. 38, San Diego, California, USA, 2015, pp. 192–204.

[147] C. Helmberg and F. Rendl, "A spectral bundle method for semidefinite programming," *SIAM Journal on Optimization*, vol. 10, no. 3, pp. 673–696, 2000.

[148] A. D. King, W. Bernoudy, J. King, A. J. Berkley, and T. Lanting, "Emulating the coherent Ising machine with a mean-field algorithm," *ArXiv e-prints*, 2018. arXiv: 1806.08422 [quant-ph].

[149] I. Hen, J. Job, T. Albash, T. F. Rønnow, M. Troyer, and D. A. Lidar, "Probing for quantum speedup in spin-glass problems with planted solutions," *Phys. Rev. A*, vol. 92, p. 042 325, 4 2015.

[150] S. Agarwal, S. J. Plimpton, D. R. Hughart, A. H. Hsia, I. Richter, J. A. Cox, C. D. James, and M. J. Marinella, "Resistive memory device requirements for a neural algorithm accelerator," in *2016 International Joint Conference on Neural Networks (IJCNN)*, 2016, pp. 929–938.

[151] R. Ulrich, "Far-infrared properties of metallic mesh and its complementary structure," *Infrared Physics*, vol. 7, no. 1, pp. 37 –55, 1967.

[152] H. Lee, *Thermoelectrics: Design and Materials*. Wiley, 2016, ISBN: 9781118848920.