GaAs OPTOELECTRONIC INTEGRATED CIRCUITS FOR OPTICAL NEURAL NETWORK APPLICATIONS

Thesis by

Steven H. Lin

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1992

(Submitted Sept. 24, 1991)

Acknowledgments

I would like to thank first my advisor, Prof. Demetri Psaltis, for his guidance and support on this project. It has been always inspiring and challenging working under him. I am deeply grateful for the tremendous patience he had with me during the early phase of this work and the countless amount of time, energy and capital he has spent on me. Without any of these elements, my thesis would not have been possible. To show my sincere appreciation to him for being so supportive and understanding, I am dedicating the monolithically integrated device developed in this work, Photon-Sensing and Amplifying Light-emitting Thresholding Integrated Switch (PSALTIS), to him. It has indeed been my privilege and pleasure to be a member of his group.

I am also greatly indebted to Prof. Rutledge and his students, Dr. Yong Guo and Dr. Bobby Weikle, for more than generously letting me use and assisting me in operating their laboratory facilities extensively. Many thanks also go to Prof. Nicolet, Prof. Tai and Prof. Yariv for providing some of the device processing needs.

JPL has been my main place of conducting device research initially and has been supporting my project throughout the entire course of my work. My thesis would not have been started and eventually finished without the support and the generosity from the management and all members of the Sensor Technology Department of JPL. In this regard, I would like to thank first Dr. Jae. H. Kim, who shared many of his insightful ideas on many device and processing issues. The accomplishment of this thesis is a direct result of his sharp criticisms and invaluable advices. His moral support and friendship are also very much appreciated. I would also like to acknowledge the support from the management, Dr. Barbara Wilson, Dr. Robert Lang, and Dr. Joseph Katz, who inspired my initial interests in under-

taking this difficult optoelectronic integrated circuit project and provided a lot of initial research guidance and assistance in my exploration of the wonderful GaAs technology. The device presented in this thesis would not have been possible without the material generously provided and grown by Dr. Akbar Nouhi, Dr. Siamak Forouhar, Dr. Anders Larsson, Mr. Jeff Cody, and Mr. John Liu. Many technical discussions with these persons and Mr. Mohammad Mazed, who also provided much assistance in the device processing, are deeply appreciated. Technical supports and assistances from Ms. Susan Martin, Mr. Carlos Matus, Mr. Charles Radics, Ms. Judy Podosek and especially Mr. Charles Manning are thankfully acknowledged. I would also like to thank JPL for donating so many semiconductor processing equipments to Caltech. Finally, my support at Caltech by a JPL SDIO/ISTC scholarship is gratefully received.

I owe many thanks to the past and present members of the Optical Information Processing group at Caltech. One can not make progress without sharp criticisms from his fellow workers. In this regard, I would like to thank first, Mr. Francis Ho, who not only criticized a lot of my devices, but also contributed to a large portion of the MESFET-based optoelectronic neurons in both the design and process development. His development on the mask making process simplified a lot of tedious work. I am especially grateful to him for this. Many thanks also go to Ms. Annette Grot for sharing many technical discussions and assisting me in maintaining the processing laboratory and upgrading some of the old processes. I am also grateful for the assistance of Mr. Jia-Fu Luo, who performed most of the measurements on optical FET's and characterized our contact alloying machine. Mr. Alan Yamamura has more than generously donated much of his time, helping me with countless computer problems and sharing his refreshing and intuitive approaches in solving problems, especially on the circuit aspect of the MESFET-based neurons. The assistance from Mr. Sidney Li, who made putting this thesis together possible

by patiently showing me the "Fig" and "Grtool" software utilities, is deeply appreciated. I would also like to acknowledge the numerous help I received from Dr. Robert Snapp, Dr. Nabeel Riza, Dr. David Brady, Dr. Jeffrey Yu, Dr. Fai Mok, Dr. Scott Hodson, Dr. Mark Neifeld, Dr. Cheol-Hoon Park, Dr. Claire Gu, Dr. John Hong, Dr. Ken-Yuh Hsu, Mr. Charlie Stirk, Mr. Robert Denkewalter, Mr. David Marx, Ms. Chaunyi Ji, Mr. Donald Lie, Mr. Subrata Rakshit, Mr. Yong Qiao, and Mr. Kevin Curtis. Administrative support from Ms. Su McKinley, Ms. Linda Dozsa and Ms. Helen Carrier is gratefully acknowledged.

I would also like to express appreciation to Mr. Chung Ho of Vitesse Semiconductor Corp, also my former classmate at MIT and roommate in Silicon Valley, for many valuable suggestions and discussions on MESFET's and MESFET circuits. The double heterojunction bipolar transistors data presented in Ch. 3 have been taken mostly by Ms. Angela Lee. Her help is deeply appreciated. The donation of a MBE-grown wafer by the Hewlett-Packard Laboratory and the assistance of Dr. S. Y. Wang of the Hewlett-Packard Laboratory in some processes for the study of double-heterojunction bipolar transistors is gratefully acknowledged.

During the course of this work, I have received constant encouragement from a special friend, Anita, whose consideration and understanding are deeply appreciated. Finally, I would like to thank my family, especially my mother, Enid, for their support and love. Without them, my graduate study at Caltech would not have been so smooth and most of all this thesis would not have been possible.

Abstract

Optoelectronic integrated circuits (OEIC's) have emerged as a viable method in the implementation of optical neurons required for a neural network. This is due to the increased capability in both the material and the device engineering in GaAs technology, which has proliferated incredibly fast during the last decade. In this thesis, two different approaches to monolithically integrate various electronic and optical devices are explored for the implementation of optical neurons. The first approach utilizes the technology from double heterojunction bipolar transistor for its potentially high current gain and its structural compatibility with optical devices. In achieving the current gain required for optical neurons, modeling of the base leakage current, effect of surface passivation and diffusion characteristics is performed for Zn-diffused bipolar transistors. The second approach employs metal semiconductor field-effect transistors as the driver for the optical devices. It is found that, by properly designing the circuit, high optical gain, low electrical power dissipation and low optical switching energy thresholding devices can be accomplished in this approach with large input-output isolation. Such performance is required if large arrays of optoelectronic neurons are to be inserted into a neural network to perform tasks that make neural computation a unique approach in solving a certain class of problems. In this thesis, an optical gain of 80 is demonstrated along with an electrical power dissipation of 1.6 mW and an optical switching energy of 10 pJ. These results generate high promises and optimism for the realization of a physical neural computer in the near future.

Contents

Chapter 1

Introduction

The implementation of a neural computing system, whose structure and function is motivated by natural intelligence, will provide a unique way to solve problems that are typically too difficult for the conventional electronic computers to tackle. Interest in this type of computer has emerged largely because it is hoped that by building a computer that shares some of the characteristics of the biological systems, we will be able to address problems such as image recognition which animals do exceedingly well but current computers do not. There has been considerable progress on the theoretical research in neural network to justify the optimism of future applications. This has resulted in a focused attention on the hardware realization of neural architectures. The computational power of neural computers arises from matching the computer architecture and the physical properties of the devices used in the implementation to the requirements of the problem. In other words, a neural computer is highly specialized and it is therefore very difficult to derive its full advantages on a general purpose computer. This provides a strong impetus for advancing the technologies for the physical realization of neural computers in parallel and interactively with the development of theoretical neural network models.

A neural network consists of two basic components : a large collection of neurons and a dense network of interconnections among all the neurons. Neurons are usually modeled as thresholding elements. Information is stored in the weights of the connections largely through error driven learning algorithm. If, during a learning phase, the response of the network is correct, then the connections remain

unaffected. Otherwise they are modified to eventually produce a desired response. There are two contenders for the physical implementation of a neural network : electronics and optics. While the thresholding function of a neuron is relatively easy to implement in electronics, the massively dense interconnection network among the neurons is becoming the bottleneck in the realization of an electronic neural network. Furthermore, these interconnections are not dynamically modifiable because the interconnections are defined by metal wires in the integrated circuits, which form the basic building block for the neurons. Heat dissipation and interconnection delay are also serious limiting factors as the network gets denser and larger.

Optics, on the other hand, is well suited for a system in which a network of massively interconnected elements are required. This is achieved by arranging arrays of neurons in a planar geometry and using the third dimension to globally interconnect the neural planes with light. A variety of optical architectures for the realization of optical neural computers have been proposed [1-8] and most fit this basic architectural design. Figure 1.1 illustrates the schematic diagram of such a system. The feature of the optical implementation that gives it an advantage when compared to the electronic counterpart is the fact that it is constructed in three dimensions. This allows the active devices at the neural planes to be populated by processing elements only, since the interconnections are external to the planes of neurons. The third dimension is used to store the information that is required to specify the connections among the neurons. It is important to keep in mind that in a densely interconnected network the weights represent a large database. This large database can be easily implemented in the form of holographic interconnections [9]. A second advantageous feature of the optical implementation is the relative ease with which learning can be accomplished by dynamic holograms recorded in photorefractive crystals [10,11]. This has allowed the holograms to be programmed in real-time and the specific interconnections to be modified as the network is in its

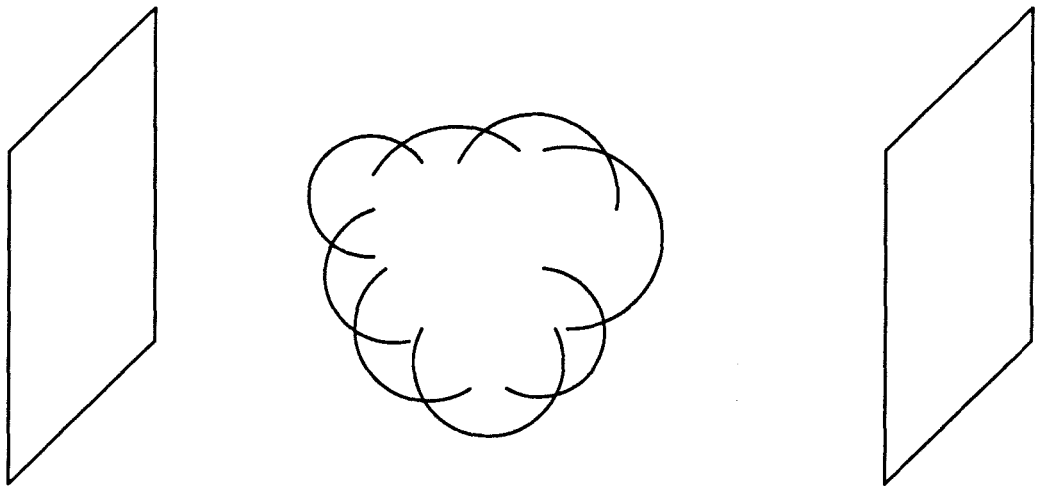


Fig. 1.1 Architecture of an optically implemented neural network.

learning phase.

The implementation of an optical network also requires physical devices that simulate the function of actual biological neurons. There are several possible candidates for the realization of optical neurons. The first is an optically addressed spatial light modulator such as the liquid crystal light valve [12]. Though large in density, liquid crystal light valves are not flexible in their use. The lack of variable threshold control and resolution also contribute to their functional inadequacy. In addition, the temporal response is in the milli-second range, which is slow for some applications. Other devices like ferroelectric liquid crystal spatial light modulators on silicon [13], electrooptic ceramics, such as PLZT, on silicon [14], heteroepitaxy of III-V material on silicon [15], and epitaxial lift-off hybridization of fabricated III-V devices on silicon [16] have also been tried. These devices are taking the advantage of the relatively mature silicon VLSI integrated circuits technology to provide the necessary functionality. However, to have an optical output, either light emitters, such as GaAs lasers or LED's, or spatial light modulators are either built somewhere else on different material and then transported to silicon circuits or are grown on the silicon substrate directly. The former method, which involves hybridization of two incompatible devices, requires complicated processes and procedures in properly connecting the optical devices to the silicon circuits. The latter method involves growing typically GaAs on silicon substrate. Because of the build-in 6% lattice mismatch in the lattice spacing between GaAs and Si, monolithically integrated devices based on GaAs on Si are subject to strain, which, if improperly controlled, will result in defects in the material and cause the degradation in the device performance.

The third candidate for the realization of the optical neurons is monolithically integrated optoelectronic circuits [17]. It can provide a better solution much faster and much easier. The optoelectronic approach is to construct a two dimensional

array of elements with each element in the array comprised of monolithically integrated detectors, electronic amplifiers and light sources. Each element simulates an individual neuron. The entire device can be built using well established fabrication techniques in GaAs and large two dimensional arrays can be constructed. The light sensitivity is excellent when compared to either liquid crystal spatial light modulators or hybridized devices. Potentially very high optical gains can also be achieved in optoelectronics to allow for the large fanout, which is required for a massively interconnected network. Moreover, there is flexibility in slowing the device down to any desirable speed in order to accommodate the large monolithic arrays to operate with reasonable electrical power dissipation. There is also flexibility in setting the function of each neuron, for example, threshold level and sharpness, electronically by designing the device and circuit appropriately. As a result, monolithically integrated optoelectronic circuit offers overall superior performance and greater flexibility when compared to the other candidates for optical neurons. Therefore, a neural network system, in which a planar array of GaAs optoelectronic integrated circuits with holographic optical elements located on top of the array to provide the interconnections among the neurons, can be envisioned. This is conceptually illustrated in Fig. 1.2.

This thesis is an investigation of various optical and electronic devices which, when monolithically integrated, give the best performance for the optoelectronic neurons. This involves detailed study on various discrete devices as well as integrated devices. Because of the nature of monolithic integration, the best discrete devices may not necessarily yield the best optoelectronic neuron when they are monolithically integrated. Consequently, trade-off analysis is required to determine the relative importance of each device in the integration. Nevertheless, high-performance discrete devices are generally desirable, regardless of the level of integration involved. Some examples of such performance are high-efficiency

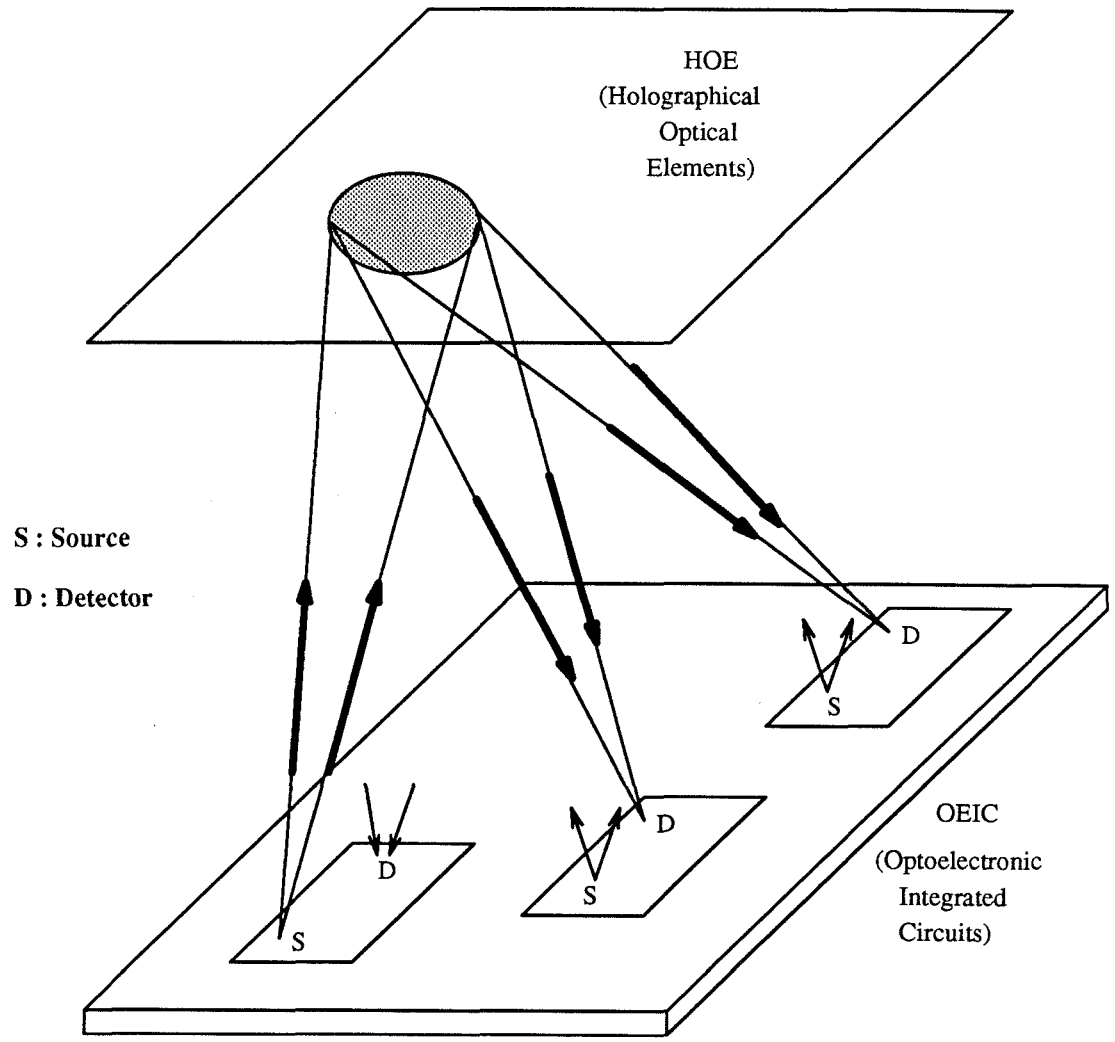


Fig. 1.2 Implementation of an optical neural network utilizing the well established GaAs optoelectronic integrated circuits as optical neurons and holographic optical elements as interconnection media.

LED's, laser diodes and photodiodes, high-current-gain bipolar transistors, and high-transconductance field-effect transistors. Therefore, this thesis starts by analyzing two discrete devices in detail. Chapter 2 gives a detailed analysis on the external quantum efficiency of LED's. The efficiency is investigated in terms of how efficiently photons are generated internally by the injected carriers, what factors limit the generated photons to escape from the device, and how easily the escaped photons can be collected. Simulations of external quantum efficiency as a function of various device parameters are illustrated. These are mostly background information that are needed in order to optimize the integration of the LED's and the transistors. Based on these results, a new way of confining the injected carriers into the LED, which results in an improvement in the external quantum efficiency, is also experimentally demonstrated at the end of Ch. 2. Chapter 3 analyzes the factors that limit the performance of double-heterojunction bipolar transistors (DHBT), specifically for Zn-diffused transistors. Mechanisms of the leakage currents, which reduce the current gain of the transistors, are identified and modeled. Proposed methods are experimentally verified and are shown to be very effective in increasing the current gain of the transistors. A systematic study on the condition of Zn-diffusion, which is required to make the contact to the base of the transistor, is also carried out. The results indicate a strong dependence of the current gain of the transistor on the condition at which the Zn-diffusion is performed.

Chapter 4 describes the integration scheme of two double-heterojunction bipolar transistors and a LED as a possible structure of an optoelectronic neuron. A parasitic problem is found in this integrated structure and is discussed and analyzed. The reason why double-heterojunction bipolar transistors are not suitable for neural network is also identified. This leads to metal semiconductor field-effect transistor (MESFET)-based optoelectronic neurons, which are presented in Ch. 5. In this chapter, two MESFET's, a LED and a detector are monolithically integrated

to perform the thresholding function needed by the neurons. Various detectors are compared and experimentally tested to determine the suitability of each type of detector in a neural network. Some of the background information is included in order to fully explore the trade-off among various schemes of integration. By properly designing the circuit and controlling the process, low-power and high-gain optoelectronic neurons are demonstrated. The advantages of the MESFET-based optoelectronic neurons over the bipolar transistor-based optoelectronic neurons are also presented. Chapter 6 compares the performance of the optoelectronic neuron, whether bipolar transistor-based or MESFET-based, that uses a LED as the light source and that uses a laser diode as the light source. Electrical power density dissipated by the LED-based neuron and the laser-based neuron is analyzed in cases where there is no optical feedback on the individual neuron and there is optical feedback through the holograms on the individual neuron. This power dissipation is then used to predict the maximum number of neurons that can be packed into an array without excessive heat dissipation for both types of neurons so that the trade-off between the speed and the array size can be understood. Finally, a proposal for improvement in the performance of the optoelectronic neurons, the complication of the fabrication steps involved, and the controllability of the uniformity across the neurons in an array is presented in Ch. 7, followed by some concluding remarks.

Chapter 2

Light-Emitting Diodes

2.1 Introduction

Light-emitting diodes (LED's) are semiconductor p-n junctions that under proper forward-biased conditions can emit external spontaneous radiation at a wavelength that depends on the material of the semiconductor. Their I-V characteristics are the same as those of the conventional p-n diodes. However, the energy released from the recombination of n-type and p-type carriers is transformed into light, as opposed to heat. This transformation is only possible in materials with direct bandgap in which conservation of momentum is maintained during the recombination process. Therefore, most of the LED's are fabricated in group III-V materials, such as GaAs and InP based materials. Being more mature in its material quality and technological development, GaAs has dominated the research and development efforts not only in LED's, but also in all spectrum of photonic devices, such as lasers and detectors. The advantage of fabricating LED's in GaAs-based material is, in addition to the previously mentioned direct bandgap property of GaAs, the ability to provide carrier confinement for both electrons and holes. These carriers are injected from the cathode and anode respectively. In order to promote efficient recombination, these carriers have to be confined in the same spatial region such that the wave functions of these two carriers have the maximum overlap. This feature of carrier confinement can be achieved easily in GaAs-based material. The active layer, within which the carrier recombination takes place, is sandwiched by two higher bandgap material such as AlGaAs, thus forming a double-heterojunction

(DH) structure. The upper AlGaAs layer is p-type doped and the lower AlGaAs layer is n-type doped. This forms the basic structure of all light-emitting devices. Depending on the doping concentration and thickness in each layer, the photon generation efficiency can be maximized. Since LED's are isotropic emitters, half of the photons generated are lost in the substrate. For the other half of the photons which are emitted vertically to the top surface, they are mostly reflected back into the LED's due not only to the small critical angle in GaAs which has an optical index of 3.6, but also the metals which inject the carriers into the LED's. Thus, LED's suffer from the disadvantage of not being able to extract the generated photons efficiently. Of course, there is the issue of collecting those photons which eventually make their ways outside the device. Because of the large optical index in GaAs, the divergence of the beam is large. Not all of these photons are usable and collectable. As a result, LED's are inherently inefficient devices. However, their structural simplicity and the lack of current threshold requirement make them practical and more efficient than lasers in certain applications, as we will demonstrate in subsequent chapters.

In this chapter, we will analyze the issues mentioned above in detail. Specifically, we will examine the factors that photon generation efficiency depends on so that this efficiency can be optimized. Next, we will investigate the difficulties involved in extracting the photons and ways to circumvent these difficulties. Discussion on ways to improve the collection of the photons once they have escaped the LED's are followed. The speed of the LED's is examined and methods of increasing the speed is investigated. This is followed by the computer simulations on the LED quantum efficiency as a function of various LED parameters. The results from the simulations provide the design rules for the optimum performance of the LED's as well as that of the integrated optoelectronic neurons. Finally, the section on LED will end in an experimental demonstration of the LED's and verification

of the improvement in the LED efficiency by using one of the methods suggested in the analysis.

2.2 Photon Generation Efficiency

In order to understand the process of photon generation, it is important to model the physics of carrier transport inside the LED's. Fig. 2.1(a) illustrates the cross section of a AlGaAs/GaAs/AlGaAs double heterojunction LED with active layer doped lightly to p-type. Light is extracted from the top, or the p-type side. As mentioned before, electrons and holes are injected into the active GaAs layer, where these carriers will survive, on the average, a lifetime τ before recombination takes place. Since there are radiative and nonradiative recombinations, the lifetime, τ, is given by

$$\frac{1}{\tau} = \frac{1}{\tau_r} + \frac{1}{\tau_{nr}}, \tag{2.1}$$

where τ_r and τ_{nr} denote radiative and nonradiative recombinations, respectively. The radiative recombination is promoted by the confinement of carriers provided by the barriers in the higher bandgap material, AlGaAs, which sandwich the GaAs active layer. This is illustrated in Fig. 2.2(b). Once the electrons and holes are trapped inside the lower bandgap region, they have nowhere else to go but to recombine with each other. It should be noted that the origin of the nonradiative recombination may be from the recombinations at the heterointerface and defect sites. These nonradiative recombinations reduce the overall lifetime of the carriers through Eq. (2.1) and decreases the quantum efficiency of the LED's as well.

From basic semiconductor physics in one dimension [18], the electron and hole

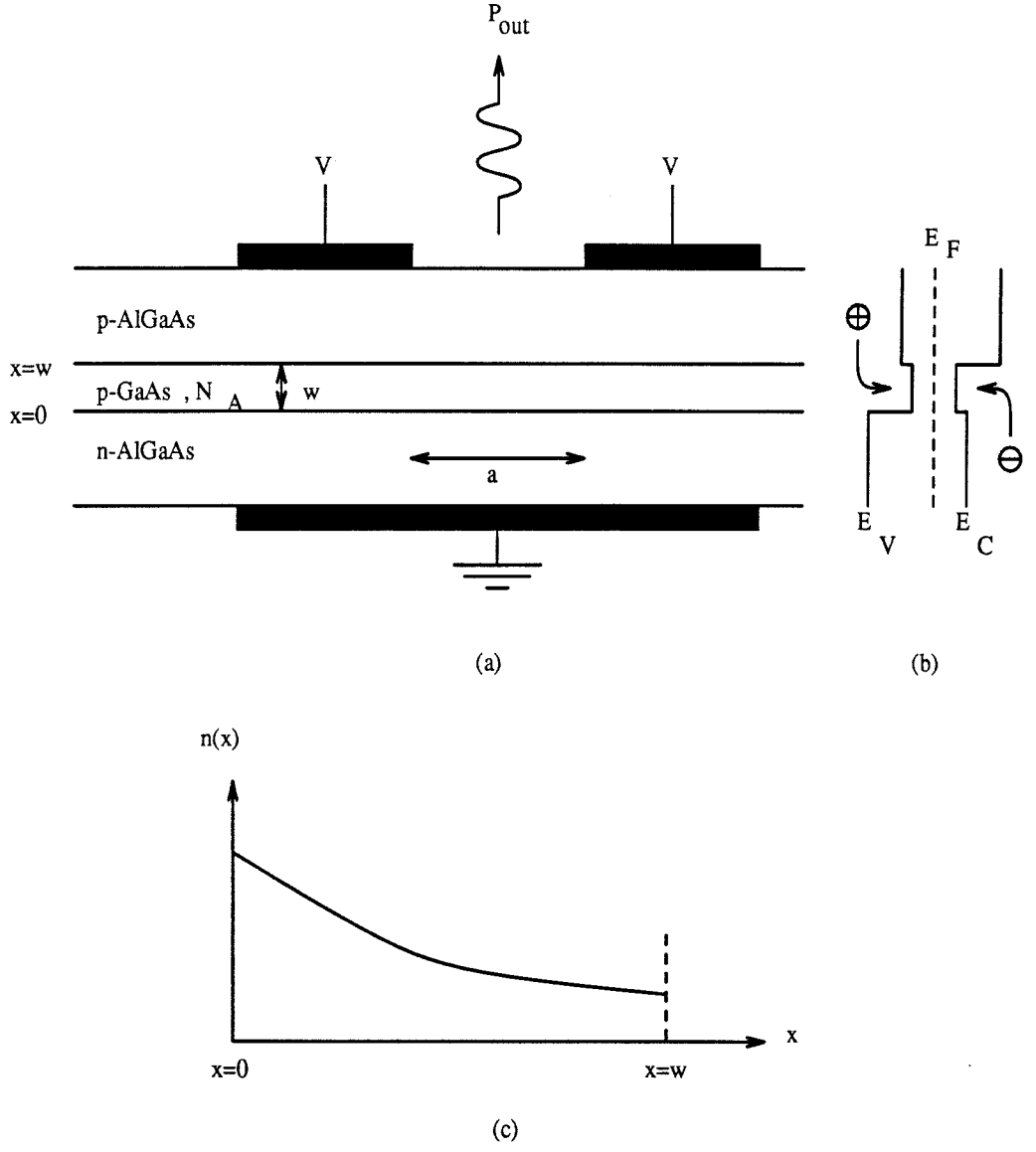


Fig. 2.1 **(a)** The cross section of the AlGaAs/GaAs/AlGaAs double heterojunction LED with lightly doped p-type active GaAs layer. **(b)** The band diagram of the double heterojunction LED, illustrating the concept of carrier confinement for both electrons and holes inside the active GaAs layer. **(c)** The minority carrier (electron) concentration inside the active layer.

conduction currents, J_n and J_p, are given by

$$J_n = q\mu_n nE + qD_n \frac{dn}{dx}, \tag{2.2}$$

$$J_p = q\mu_p pE + qD_p \frac{dp}{dx}, \tag{2.3},$$

where $q, \mu_n, \mu_p, n, p, D_n, D_p$, and E are electron charge, electron mobility, hole mobility, electron concentration, hole concentration, electron diffusion coefficient, hole diffusion coefficient, and electric field respectively. The first components in Eq. (2.2) and (2.3) are the drift currents due to the electric field. The second components in the same equations are the diffusion currents due to the gradient in carrier concentration. For low level injection, the conduction current of the minority carrier is dominated by the diffusion current as the minority carrier concentration is low enough for the drift current to be negligible. Thus, it is usually easier to work with minority carrier. Since the active layer is p-type, the minority carriers inside the active layer are electrons. This means that Eq. (2.2) would be the appropriate equation to work with and it can be approximated by

$$J_n \approx qD_n \frac{dn}{dx}. \tag{2.4}$$

There are also continuity equations, which describe the relations between the current and time variations in the carrier concentrations. They are

$$\frac{dn}{dt} = G_n - R_n + \frac{1}{q}\frac{dJ_n}{dx} \tag{2.5}$$

$$\frac{dp}{dt} = G_p - R_p + \frac{1}{q}\frac{dJ_p}{dx}, \tag{2.6}$$

where G_n, G_p, R_n, and R_p are the generation and recombination rate for the electrons and holes respectively. At steady state, Eq. (2.5) can be set to zero. Also, there is no generation of carriers in the LED's. Thus, Eq. (2.5) can be re-written as

$$R_n + \frac{1}{q}\frac{dJ_n}{dx} = \frac{n}{\tau_n} + \frac{1}{q}\frac{dJ_n}{dx} = 0. \tag{2.7}$$

In the above, the assumption that the recombination rate is equal to the electron concentration divided by the electron lifetime, which is given by Eq. (2.1), is used. Substituting Eq. (2.4) into Eq. (2.7) yields the steady state second order differential equation for the electrons and is shown as below.

$$\frac{d^2n}{dx^2} - \frac{n}{L_n^2} = 0, \tag{2.8}$$

where L_n is the electron diffusion length and is equal to

$$L_n = \sqrt{D_n\tau_n}. \tag{2.9}$$

The boundary conditions for the LED's are shown below [19]

$$-\frac{dn}{dx}(x=0) = \frac{J}{qD_n} - \frac{sn(x=0)}{D_n} \tag{2.10}$$

$$-\frac{dn}{dx}(x=w) = \frac{sn(x=w)}{D_n}, \tag{2.11}$$

where J, and s are the current density injected into the LED and the interfacial recombination velocity at the GaAs/AlGaAs heterointerface. The first boundary

condition states that the current density right after the edge of the heterojunction ($x = 0^+$) is equal to the total current density injected into the LED minus the recombination current which takes place at the same heterojunction. Similarly, the second boundary condition states that the current density right before the edge of the other heterojunction ($x = w^-$) is equal to the current density needed for the interfacial recombination at that heterojunction. Thus, the effect of the interfacial recombination is to reduce the magnitude of the originally injected current. This reduces the efficiency of the device. The parameter that describes the degree of recombination at the heterointerface, s, can be related approximately to the strain due to lattice mismatch at the heterointerface [20]. This relationship is described by

$$s \approx 2 \times 10^7 \cdot \frac{\Delta a}{a}, \tag{2.12}$$

where Δa and a are the mismatch in the lattice constant and the lattice constant respectively. For GaAs/AlGaAs system, an interfacial recombination velocity of 2000 cm/sec can be readily achieved [19]. Solving the second order differential equation for electrons in Eq. (2.8) subject to the two boundary conditions stated in Eq. (2.10) and (2.11) shows that [19]

$$n(x) = \frac{JL_n}{qD_n} \frac{cosh(\frac{w-x}{L_n}) + s\frac{L_n}{D_n} sinh(\frac{w-x}{L_n})}{((\frac{sL_n}{D_n})^2 + 1) sinh(\frac{w}{L_n}) + 2\frac{sL_n}{D_n} cosh(\frac{w}{L_n})}. \tag{2.13}$$

The electron concentration in the active region is schematically illustrated in Fig. 2.1(c). The electrons that get injected into the active region will remain free for an average of their lifetime before they get recombined. The generation rate of photon density can be easily derived by dividing the electron concentration by the

lifetime or $n(x)/\tau_r$. Each photon has an energy of hc/λ, where h, c and λ are Plank's constant, the speed of light in vacuum and the wavelength respectively. However, a portion of the generated photons will be re-absorbed when traveling through the active layer. Thus, the generation rate of photon density will be reduced by a factor of $exp(-\alpha(w-x))$ after traversing through the active layer, where α is the absorption coefficient for the photon in the active layer. The total optical power generated from the LED can then be obtained by integrating the product of these two factors over the thickness of the active layer and then multiplying it by the active area. This is shown below.

$$\begin{aligned} P_{out} &= A \cdot \frac{hc}{\lambda} \cdot \int_0^w \frac{n(x)e^{-\alpha(w-x)}}{\tau_r} dx, \\ &= \frac{\tau}{\tau_r} \cdot \eta_s \cdot \frac{hc}{q\lambda} \cdot I, \end{aligned} \tag{2.14}$$

where

$$\eta_s = \frac{\frac{1+S}{1-\alpha L_n}\left[1 - e^{-w\left(\frac{1-\alpha L_n}{L_n}\right)}\right] e^{\frac{w}{L_n}} - \frac{1-S}{1+\alpha L_n}\left[1 - e^{w\left(\frac{1+\alpha L_n}{L_n}\right)}\right] e^{-\frac{w}{L_n}}}{2\left[(S^2+1)sinh\left(\frac{w}{L_n}\right) + (2S)cosh\left(\frac{w}{L_n}\right)\right] \cdot e^{\alpha w}} \tag{2.15}$$

$$S = s \cdot \frac{L_n}{D_n}, \tag{2.16}$$

and I is the injected current. It should be noted Eq. (2.14)-(2.16) assume that the active layer of the LED generates and re-absorbs light at a single wavelength. This is not true as LED's are known to emit a broad spectrum and the re-absorption will not

be over a single wavelength either. Thus, techniquely speaking, an integration over the entire wavelength spectrum is needed in Eq. (2.14). However, this would require the knowledge of the emission spectrum, which is usually gaussian in wavelength [21] as well as the absorption profile. For simplicity of analysis, we assume this effect is negligible and treat the emission and the absorption of LED's to be single-wavelength.

In general, the mechanisms by which the photon generation efficiency of the LED's can be reduced can be categorized into 3 items. They are interfacial recombination current, self-reabsorption of photons in the active layer, and the non-radiative recombination process, such as Auger process [19]. As stated earlier, the interfacial recombination is a direct result of the defects at the heterointerface caused by the lattice mismatch between the GaAs and AlGaAs. This recombination obviously contributes to the non-radiative recombination process. Thus, the presence of interfacial recombination modifies the overall lifetime for the monority carriers, τ, from that stated in Eq. (2.1) to the following :

$$\frac{1}{\tau} = \frac{1}{\tau_r} + \frac{1}{\tau_{nr}} + \frac{2s}{w}. \tag{2.17}$$

This reduction of minority carrier lifetime is undesirable for the LED's because it directly translates into the reduction of the photon generation efficiency for the LED's, as evidence in the first factor in Eq. (2.14). Therefore it is extremely important to minimize the value of s, or the interfacial recombination velocity at the GaAs/AlGaAs heterojunction.

Self-reabsorption can also be a significant factor in reducing the efficiency. It cannot be avoided because photons generated in the active region can be immediately reabsorbed by the same material. Thus, while it is important to design the active layer of the LED not to be too thick so as to reduce self-reabsorption, it is

critical to realize that too thin an active layer will not generate enough photons. Therefore, an optimized thickness for the active layer should be designed. Reduction of other non-radiative recombination processes, which is predominantly Auger process that goes on in the bulk of the active layer, is also a necessary consideration in designing the LED's. Auger process is a process in which the energy released by the recombination of the electron-hole pair kicks another free electron in the conduction band to a higher energy state. This process can be reduced by reducing the doping concentration in the active layer, which makes the concentration of free electrons smaller, thus reducing the chance for Auger recombination to occur. By reducing all non-radiative recombination processes, including the interfacial recombination and the Auger recombination, the first term in Eq. (2.14), τ/τ_r, which is usually less than 1, can be made closer to 1. We use assumption, $\tau/\tau_r = 1$, in the simulation which will be shown in Sec. 2.6. Thus any non-radiative recombination process that is present in the LED's will reduce the efficiency by the factor τ/τ_r, or $w\tau_{nr}/[w(\tau_r + \tau_{nr}) + 2s\tau_r\tau_{nr}]$ if τ in Eq. (2.17) is substituted into τ/τ_r.

There is another issue that needs to be addressed about the photon generation efficiency of the LED's. That is the radiative recombination lifetime of the LED, τ_r. The value of τ_r can be affected by the doping level in the active layer, the thickness of the active layer as well as the injection current level and the probability of radiative recombination in the active layer. It can be derived based on the theory of bimolecular recombination [22], which states that the spontaneous band-to-band recombination rate R_{sp} inside the active layer under the conditions not requiring momentum conservation is given by

$$R_{sp} = Bnp, \tag{2.18}$$

where $B(incm^3/sec)$ is the radiative recombination probability, which is a char-

acteristic parameter of the band structure depending on the material, and n and p are the concentration of total electrons and holes, respectively. A typical value for B in GaAs is $10^{-10} cm^3/sec$ [23,24]. Under external excitation or injection, excess electrons and holes, Δn and Δp, are created, making the overall spontaneous recombination rate equal to

$$R_{sp} = B(n_o + \Delta n)(p_o + \Delta p), \tag{2.19}$$

where n_o and p_o are the intrinsic electron and hole concentrations at thermal equilibrium, and Δn is equal to Δp to maintain charge neutrality. Using $\Delta n = \Delta p$, Eq. (2.19) becomes

$$R_{sp} = Bn_o p_o + B\Delta n(p_o + n_o + \Delta n). \tag{2.20}$$

The first term in Eq. (2.20) represents the thermal equilibrium recombination rate, which is usually small compared to the other terms. In fact, if we assume the active layer is p-doped and its doping concentration is equal to N_A, meaning $p_o \gg n_o$, this thermal equilibrium recombination rate can be neglected as long as $\Delta n \gg n_o$, which is valid in most injection diodes whose doping concentration in the active layer is moderately high (at least $10^{14}/cm^3$). Making these approximations, Eq. (2.20) can be simplified to

$$R_{sp} = B\Delta n(N_A + \Delta n). \tag{2.21}$$

The radiative recombination lifetime, τ_r, is defined as

$$\tau_r = \frac{\Delta n}{R_{sp}}. \tag{2.22}$$

Substituting Eq. (2.21) into Eq. (2.22), we obtain

$$\tau_r = \frac{1}{B(N_A + \Delta n)}. \tag{2.23}$$

In case of low-level injection, $\Delta n \ll N_A$, and the radiative recombination lifetime is inversely proportional to the active layer doping concentration, or

$$\tau_r \approx \frac{1}{BN_A}. \tag{2.24}$$

However, at high injection levels, Δn becomes much bigger than the background doping concentration, N_A. Thus the radiative recombination lifetime becomes dominated by the excess carriers that get injected into the active layer, or

$$\tau_r \approx \frac{1}{B\Delta n}. \tag{2.25}$$

Thus, in general, the radiative recombination lifetime is determined by the greater of the background doping concentration, which is equal to the concentration of free carriers if 100% carrier ionization is achieved, and the free carrier concentration due to external injection. This statement can be easily understood if we just remember that whatever contributes to the total concentration of free carriers in the active layer, the radiative recombination lifetime is inversely proportional to that total concentration through a factor, which is the probability of radiative recombination.

For optimizing the LED efficiency, Δn is not a parameter that can be physically measured or easily inferred from the LED's. Thus, we need to further analyze the factors that influence the value of Δn so that the radiative recombination lifetime can be properly determined. If we start out with the minority carrier current in the active layer, which is dominated by the diffusion current, we have

$$\begin{aligned} J_n &= qD_n \frac{dn}{dx} \\ &\approx qD_n \frac{\Delta n}{L_n}, \end{aligned} \tag{2.26}$$

where L_n denotes the diffusion length for the electrons. Multiplying the numerator and denominator in Eq. (2.26) by τ_r, making use of the fact that $D_n\tau_r = L_n^2$, and assuming the thickness of the active layer is much less than the diffusion length (usually valid for lightly doped active layer), we obtain

$$\begin{aligned} J_n &= qD_n \frac{\Delta n \tau_r}{L_n \tau_r} \\ &= q\Delta n \frac{D_n \tau_r}{L_n \tau_r} \\ &= q\Delta n \frac{L_n^2}{L_n \tau_r} \\ &= q\Delta n \frac{L_n}{\tau_r} \\ &\approx q\Delta n \frac{w}{\tau_r}. \end{aligned} \tag{2.27}$$

Since the electrons and holes are confined in the low bandgap GaAs active layer, the injection of holes from the active layer into the lower n-type AlGaAs high bandgap cathode is negligible. Thus, the total current can be approximated by the electron

current that injects from the AlGaAs cathode into the GaAs active layer. As a result, Eq. (2.27) can be replaced by

$$J \approx J_n = q\Delta n \frac{w}{\tau_r}, \tag{2.28}$$

where J denotes the total current density. Eq. (2.28) can be solved to obtain Δn, which is

$$\Delta n = \frac{J\tau_r}{qw}. \tag{2.29}$$

Eq. (2.29) can then be substituted into Eq. (2.25) to obtain the radiative recombination lifetime in the case of high level injections. This is shown below.

$$\tau_r = \sqrt{\frac{qw}{BJ}}. \tag{2.30}$$

The only varying parameter in Eq. (2.30) is J, which indirectly affects the value of Δn in an inverse square root fashion. Thus, the higher the J is, the higher Δn becomes, which means the shorter the radiative recombination lifetime will be.

We have just discussed the extreme cases of $N_A \gg \Delta n$ and $\Delta n \gg N_A$. In the middle, the radiative recombination lifetime deserves more careful examination. Substituting Δn in Eq. (2.29) into Eq. (2.23) shows

$$\tau_r = \frac{1}{B(N_A + \Delta n)} = \frac{1}{B(N_A + \frac{J\tau_r}{qw})}. \tag{2.31}$$

Eq. (2.31) is a quadratic equation in τ_r. Thus, τ_r can be easily solved for.

$$\tau_r = \frac{qw\left(-N_A + \sqrt{N_A^2 + \frac{4J}{qwB}}\right)}{2J}. \tag{2.32}$$

It is important to check the appropriate limits of Eq. (2.32). To take the limit of $N_A \ll \Delta n$, we pull out the factor N_A^2 inside the square root and use the binomial expansion to obtain

$$\begin{aligned}
\lim_{N_A \ll \Delta n} \tau_r &= \lim_{N_A \ll \Delta n} \frac{qw(-N_A + N_A\sqrt{1 + \frac{4J}{qwBN_A^2}})}{2J} \\
&= \frac{qw\left(-N_A + N_A(1 + \frac{2J}{qwBN_A^2})\right)}{2J} \\
&= \frac{1}{BN_A}.
\end{aligned} \tag{2.33}$$

This agrees with Eq. (2.24). For the other limit, $\Delta n \ll N_A$, it can be shown that by setting N_A to zero,

$$\begin{aligned}
\lim_{\Delta n \ll N_A} \tau_r &\approx \frac{qw\sqrt{\frac{4J}{qwB}}}{2J} \\
&= \sqrt{\frac{qw}{BJ}},
\end{aligned} \tag{2.34}$$

which again agrees with Eq. (2.30). Thus, as we can see, the value of τ_r depends on all four parameters, J, w, N_A, and B, in a very complicated manner. However, for typical LED's, the concentration of the injected carriers in on the order of 10^{17} to 10^{18} cm^{-3}. Thus, it is not hard to satisfy the condition, $\Delta n \gg N_A$. Thus, the

radiative recombination lifetime will be predominantly determined by the amount of current injected into the active layer.

2.3 Photon Extraction Efficiency

In the previous section, we discuss the factors that affect the efficiency of the photon generation in the active layer. In this section, we will address the issue of extracting these photons that have been generated. For the particular type of the LED configuration that is of interest to us, light is emitted vertically upward into the air. There are five important considerations that will prevent these photons from being extracted from the LED. They are

1. Limitation of critical angle
2. Back reflection of photons into the LED
3. Limitation of emitting window dimensions
4. Waveguiding phenomenon
5. Emission of photons into the substrate

We will address each issue individually by analyzing the problem associated with each issue and proposing possible solutions to solve these problems.

2.3.1 Limitation of Critical Angle

From classical optics, it is well known that light impinging on another material having a smaller refractive index will undergo total internal reflection if the angle between the incident light and the surface normal is greater than the critical angle. This critical angle is given by

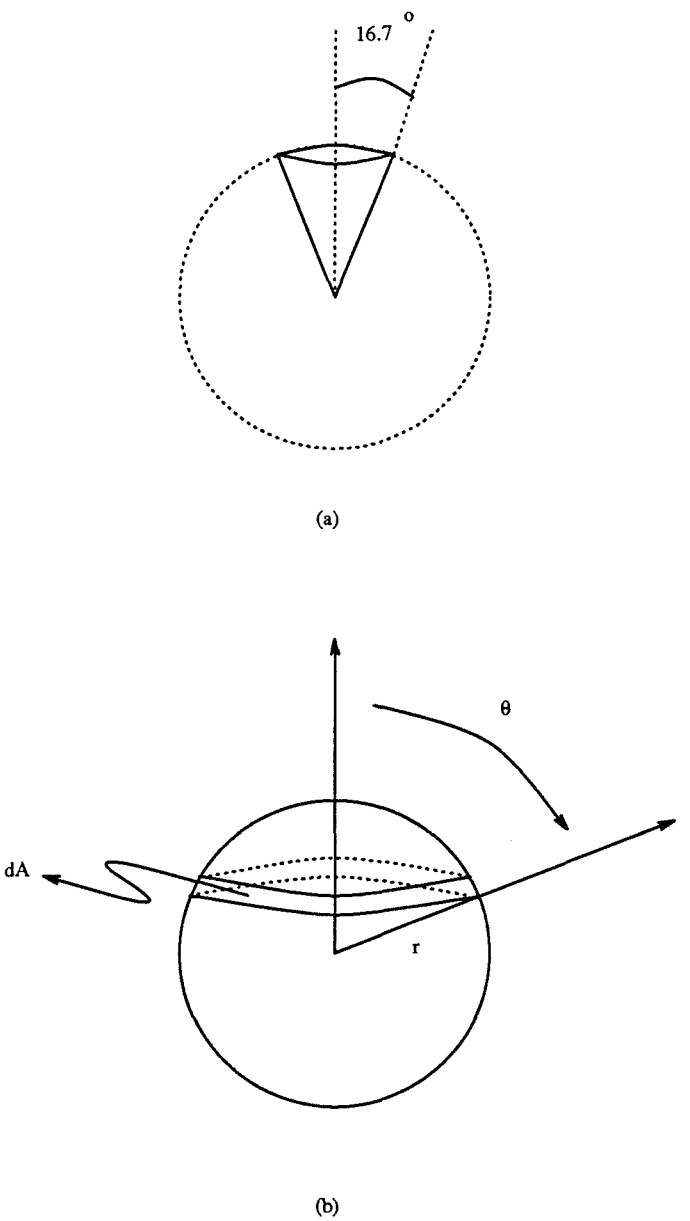


Fig. 2.2 **(a)** The schematic diagram illustrating the critical angle, θ_c, in the GaAs/air interface is only 16.7^o. **(b)** Schematic diagram showing the incremental area spanned by a circular strip on a sphere of radius, r, and located at an azimuthal angle of θ.

$$\theta_c = sin^{-1}(\frac{n_c}{n_s}), \tag{2.35}$$

where the index of refraction for the incident and transmitted materials are n_s and n_c respectively. The subscripts s and c denote substrate and cover. In GaAs LED's, this is a very severe problem because the refractive index of GaAs, which is about 3.6 at wavelength of 0.85 μm, is so high that the critical angle for transmitting light into air is only 16.7^o. This is schematically illustrated in Fig. 2.2(a), which shows that, for an isotropic emitter, only those photons that fall within the cone spanned by the critical angle can escape. Thus, it is not hard to imagine that the inefficiency of LED's is most attributable to the fact that most of the photons generated can not be transmitted outward into the air. Rather, they either undergo total internal reflection at the top surface or get absorbed in the substrate. The portion of the photons that will eventually be emitted through this small window can be calculated through the ratio of the solid angle spanned by this critical angle to the solid angle of a whole sphere, which is 4π. To calculate the solid angle spanned by a certain angle, θ, let's refer to Fig. 2.2(b). From Fig. 2.2(b), it is clear that the incremental area, dA, represented by the strip which wraps around the sphere at an azimuthal angle, θ, and of radius, r, can be given by

$$dA = 2\pi(rsin\theta) \cdot rd\theta. \tag{2.36}$$

The total spherical surface area spanned by a cone of angle θ can be obtained by integrating dA in Eq. (2.36) from 0 to an arbitrary θ. The result is

$$A = \int_0^\theta 2\pi r^2 sin\theta d\theta$$

$$= 2\pi r^2(1 - cos\theta). \tag{2.37}$$

Therefore the solid angle spanned by a cone angle of θ is

$$S(\theta) = 2\pi(1 - cos\theta). \tag{2.38}$$

For $\theta = 16.7^o$, the ratio of solid angle spanned by this angle to the solid angle of a whole sphere is thus

$$\frac{S(16.7^o)}{S(360^o)} = \frac{2\pi(1 - cos\theta)}{4\pi}\bigg|_{\theta=16.7^o} = 0.021. \tag{2.39}$$

This means that only 2 out of 100 photons will escape from the LED. Moreover, these 2 photons will not be transmitted with 100 % transmission. A certain percentage of these photons will be reflected because of the mismatch in the index of refraction of air and GaAs. This problem can be circumvented by properly coating the surface with a dielectric material, as we will see next. However, the problem of critical angle remains.

2.3.2 Back Reflection of Photons Into the LED

Reflection of light at any interface between two different materials is inevitable unless a third material with matching index of refraction and proper thickness is inserted between the two materials. For GaAs LED's, besides the problem of photons being trapped inside the LED's, back reflection of those photons which travel at angles less than the critical angle with respect to the surface normal is a problem that can not be overlooked. Because of the huge difference in the index

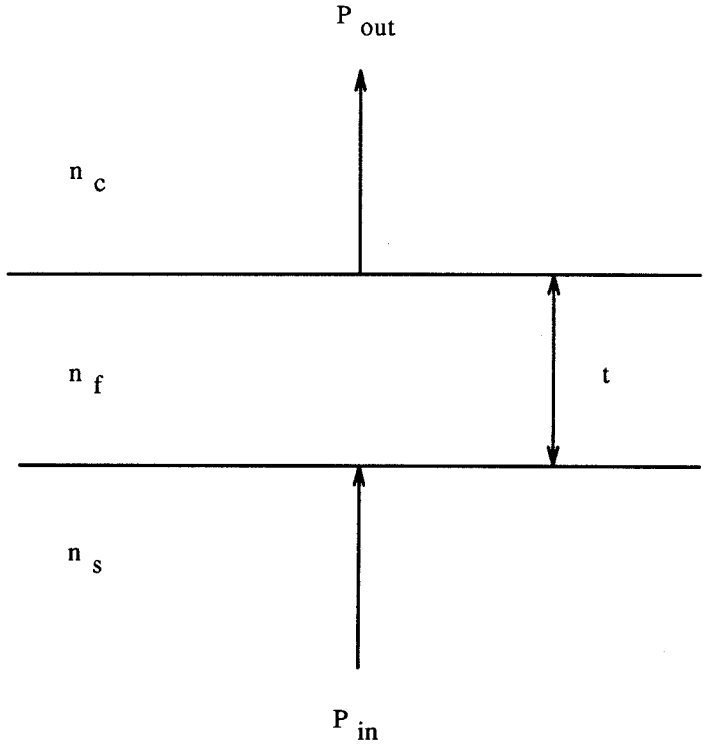


Fig. 2.3 Configuration of an anti-reflection coating system, in which a dielectric material of index of refraction of n_f and thickness of d is sandwiched between two materials having index of refraction of n_c and n_s.

of refraction between the GaAs LED's and air, a reflection of 32 % occurs for photons that are normally incident on the interface. The reflection increases as the angle of incidence increases. To alleviate this problem, we consider the problem of anti-reflection coating. Fig. 2.3 shows a dielectric film of thickness, d, sandwiched between two semi-infinite dielectric materials with refractive indices of n_c and n_s. Let's assume that n_s is greater than n_c and the refractive index of the dielectric film in the middle is n_f. It follows from the treatment in Hetch & Zajac [25] that the reflection for photons of normal incidence is

$$R = \frac{n_f^2(n_c - n_s)^2 cos^2 k_o h + (n_c n_s - n_f^2) sin^2 k_o h}{n_f^2(n_c + n_s)^2 cos^2 k_o h + (n_c n_s + n_f^2) sin^2 k_o h}, \tag{2.40}$$

where

$$k_o = \frac{2\pi}{\lambda} \qquad and \qquad h = n_f d. \tag{2.41}$$

One quick check of Eq. (2.40) is to let d equal to zero. If d is 0, then h is 0, too. This will immediately eliminate the $sin^2 k_o h$ factor and make the $cos^2 k_o h$ factor equal to 1. After cancelling out n_f^2, Eq. (2.40) is left with

$$R = \left(\frac{n_c - n_s}{n_c + n_s}\right)^2, \tag{2.42}$$

which is the familiar Fresnel coefficient. To reduce reflection down to zero, we need to make each factor in the numerator separately equal to zero. This requires making

$$n_f = \sqrt{n_o n_s} \qquad and \qquad t = \frac{\lambda}{4 n_f}. \tag{2.43}$$

In other words, the anti-reflection matching material has to have an index of refraction equal to the geometric mean of the refractive indices of the two host materials and its thickness has to be quarter-wavelength. By doing this, the transmission of light can be increased up to 100 %. For GaAs/air interface, this requires a dielectric material that has index of refraction of $\sqrt{(1)(3.6)}$ or 1.89. This index of refraction can be found in materials like Si_3N_4 or SiO_2, which are also popular materials for semiconductor processing. However, this does not mean that the problem of back reflection of light in the LED is solved because the process by which these films are deposited onto the surface of the LED's can very sensitively affect the performance of the LED's. Thus, one has to be careful when anti-reflection-coating the LED's with these dielectric films. Nevertheless, more photons can be extracted by using this technique. It is also interesting to note from Eq. (2.40) that as long as these anti-reflection coating films are odd multiples of a quarter wavelength, the transmission can be maximized.

For photons that are incident onto the interface at a slight angle but are still within the critical angle, the transmission characteristics due to this anti-reflection coating film are expected to change. This change is because of the slightly different optical path taken by these photons, which may destroy the constructive interference of the photons. As a result, the maximum transmission may be reduced. Furthermore, the thickness of the anti-reflection film at which the maximum transmission occurs may be different. This is clearly illustrated in Fig. 2.4, in which the transmission is plotted against the anti-reflection coating thickness for photons of incident angles from 0^o to the critical angle. From these plots, the following observations are noted.

1. The maximum transmission decreases as the incident angle for the photons increases from 0^o toward the critical angle. Furthermore, the transmission optimized for the normal incidence, which is usually the case of practical interest,

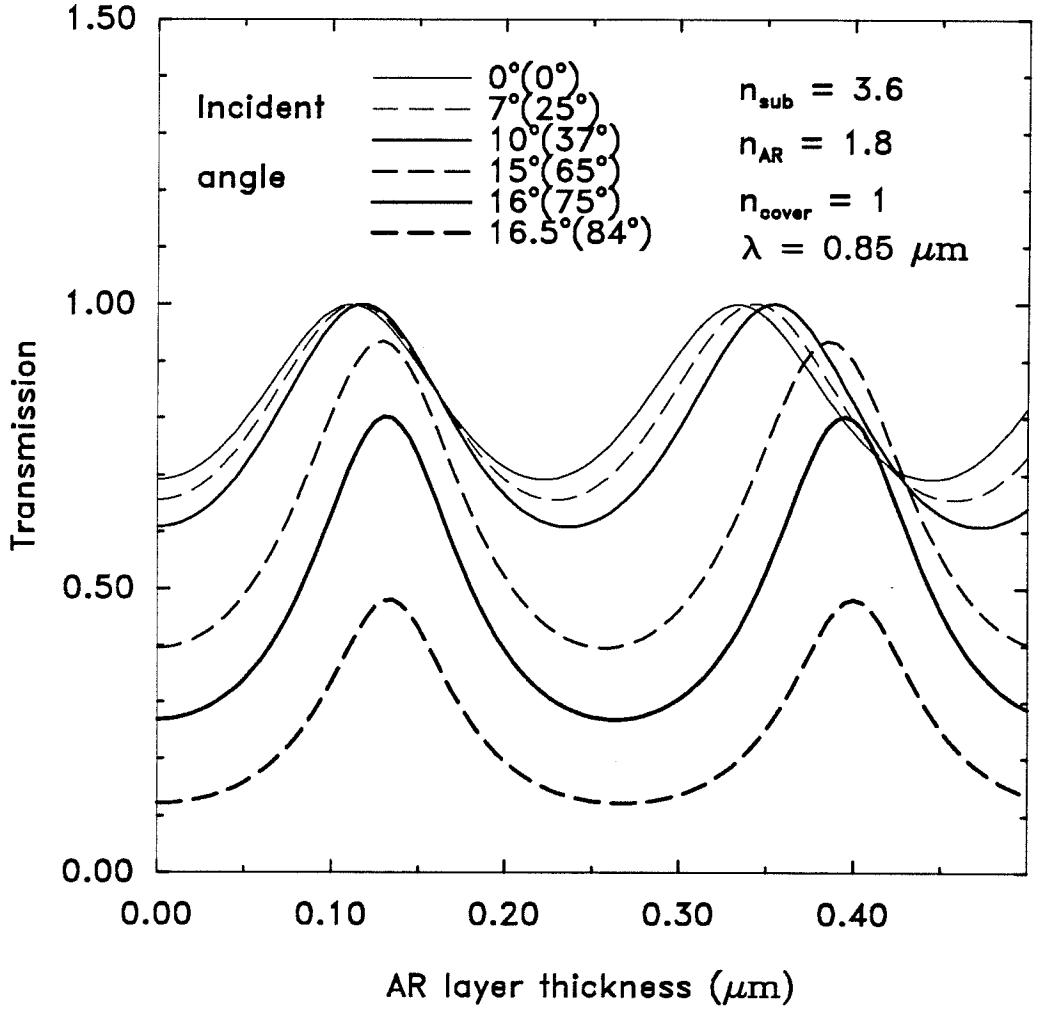


Fig. 2.4 Transmission of photons from GaAs into air through an anti-reflection coating layer vs. the thickness of the anti-reflection coating for various angles of incidence. The incident angles of 0^o, 7^o, 10^o, 15^o, 16^o, and 16.5^o correspond to the transmitted angles of 0^o, 25^o, 37^o, 65^o, 75^o, and 84^o respectively, which are in parentheses.

decreases at a faster rate than it does for the transmission optimized for a slightly off-normal incident angle as the angle of incidence increases.

2. The thickness of the anti-reflection film for maximum transmission shifts to a larger value as the angle of incidence increases. This is because the new optical path taken by these photons is shorter than the old optical path. Thus, in order to get constructive interference for the transmitted photons at the right propagation direction, the thickness of the anti-reflection film has to be increased to accommodate the deficiency in the path length.
3. The adjacent maxima in the transmission curve become farther apart as the angle of incidence increases. This can be explained by the same reason as in 2.
4. Despite the shift in the peak of the transmission toward thicker anti-reflection film as the angle of incidence is increased, the transmission coefficients at the maximums remain relative unchanged for angle of incidence of up to approximatedly 10^o. Since most of the usable light from the LED is concentrated within this cone angle, the tolerance in the thickness of the anti-reflection film for maximum transmission has become larger. From Fig. 2.4, for example, we can coat the dielectric film anywhere between 0.11 μm to 0.14 μm to obtain almost 100 % transmission for photons that impinge on the interface with angle of incidence of up to 10^o.

If LED's are indeed isotropic emitters, then the angle of incidence for the incoming photons should be uniformly distributed between 0^o and the critical angle. After being transmitted into air, the photons will be almost uniformly emitting in all angles between 0^0 and 90^0 due to the refraction at the interface. This is illustrated in Fig. 2.5, which shows the transmission coefficient as a function of the transmitted angle for different anti-reflection coating thickness. It is clear that from these plots the transmission coefficient is relatively constant for photons that emit at an angle of up to 60^o with respect to the surface normal, even though the thickness of the

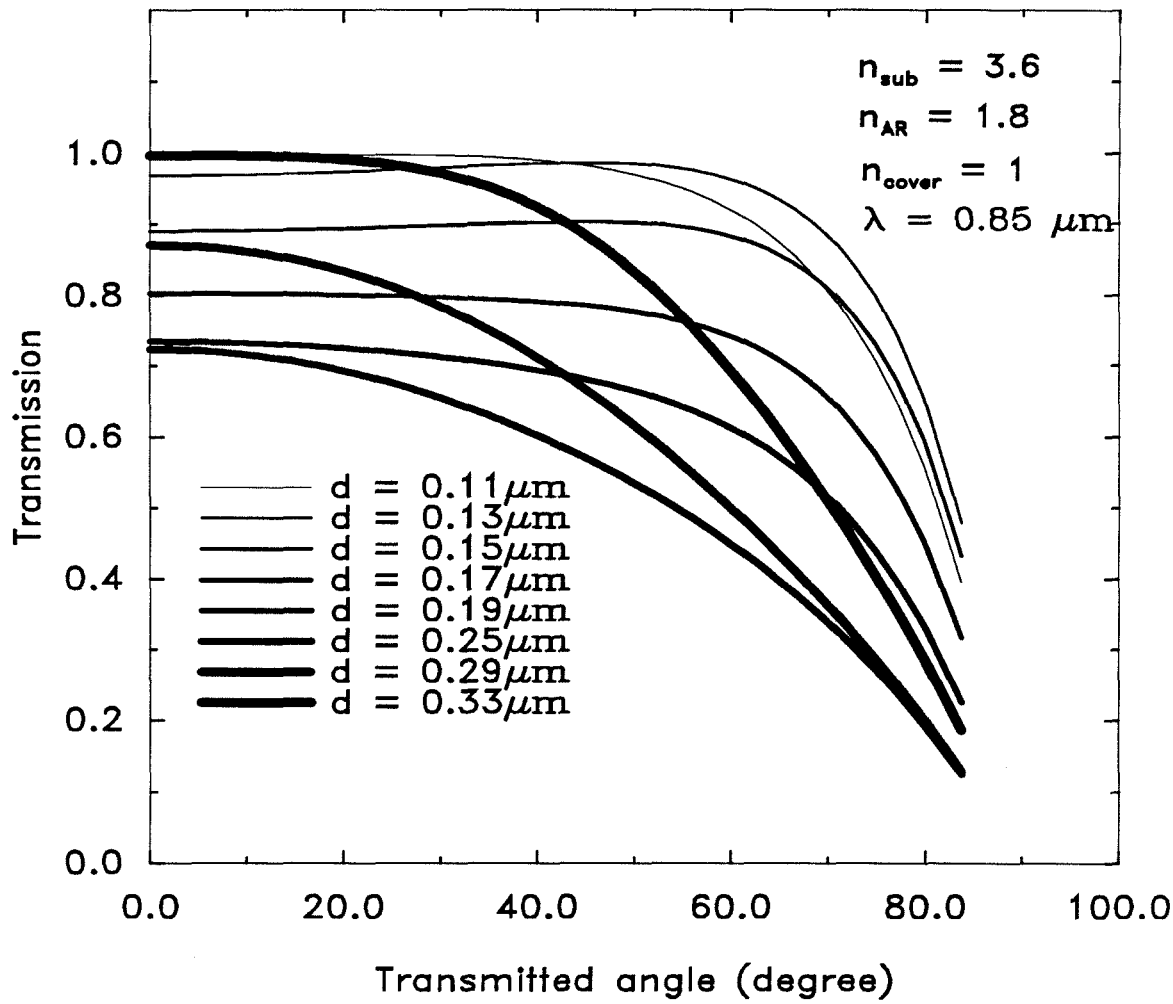


Fig. 2.5 Transmission coefficient vs. transmitted angle for various anti-reflection coating thickness.

anti-reflection film is wrong. This is useful to know as there is a practical limit on how much of the emitted photons can be collected to interface with the devices that the LED's are designed to communicate with. The results from the simulations shown in Sec. 2.6 reflect only the photons that are within a cone angle of 60^o.

2.3.3 Limitation of Emitting Window Dimensions

The analysis presented thus far is based on one dimensional analysis. If we take into the consideration of the second dimension, interesting effects can occur. One example of the significance of the second dimension is the size of the LED emitting window. Because of the low efficiency of the LED, LED's often have a rather large emitting window, which is on the order of 50 to 100 μm. As we shall find, large-window LED's may not be efficient.

Figure 2.6 shows two possible cases. The first case, shown in Fig. 2.6(a), depicts a LED with very small emitting window, or the dimension of the window is much less than the minority carrier diffusion length. Because most of the surface is covered by metals, generated photons are mostly totally reflected by the metals. As a result, the output power of the LED is expected to be very low. The second case, shown in Fig. 2.6(b), illustrates the opposite case, in which the dimension of the window is much greater than the minority carrier diffusion length. In this case, most of the current injected from the anode at the top flows almost vertically down to the cathode. Consequently, most of the photons are generated underneath the electrodes, thus are again internally reflected by the metals. In order to increase the photon density that can be transmitted into the air, the photons have to be generated underneath the emitting window area such that they would not be subject to any total reflection. However, to achieve generation of photons underneath the emitting window, injecting current must flow through the same region. As we can

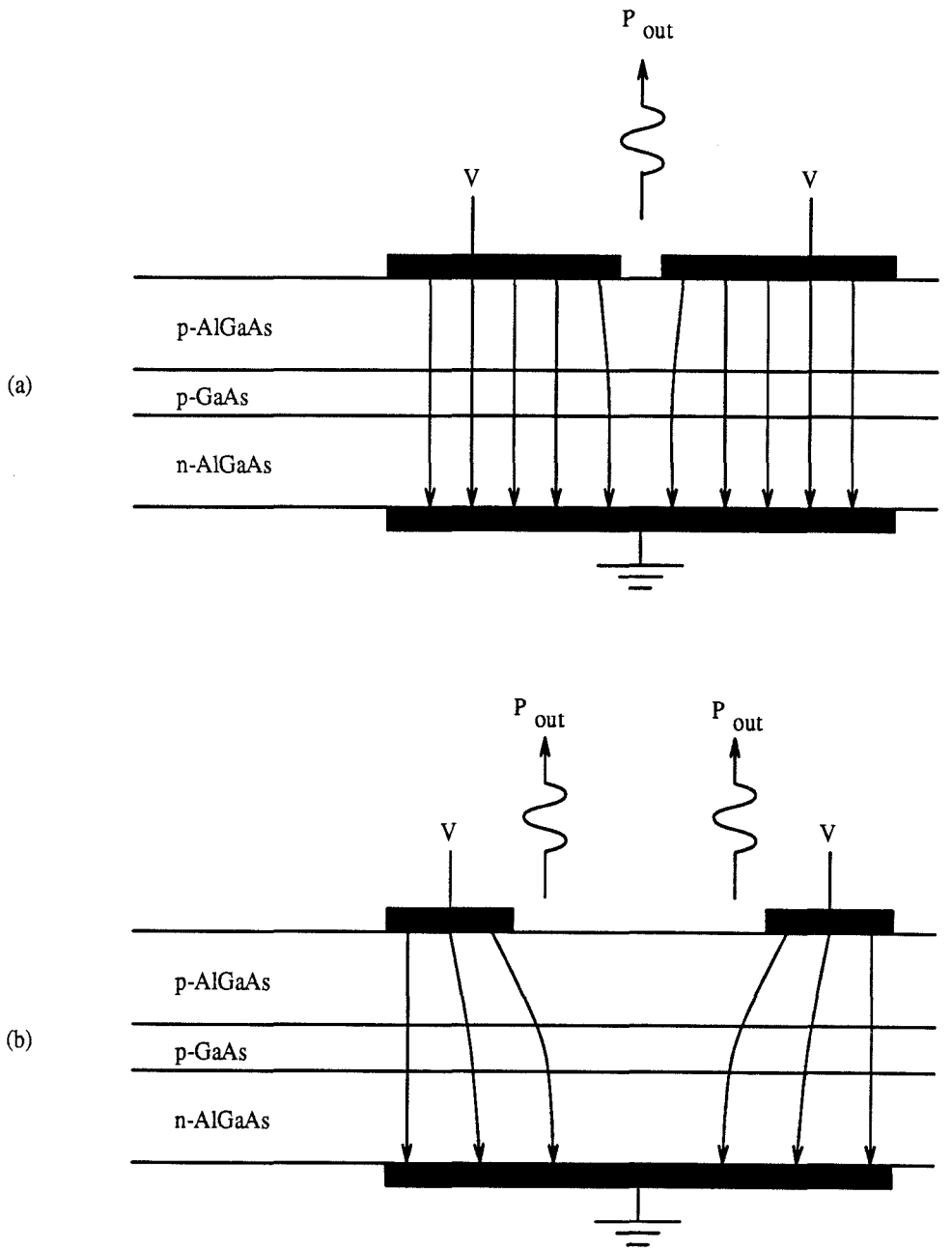


Fig. 2.6 Schematic illustrations showing the problems of the LED whose emitting window is **(a)** too small ($a \ll L_n$) or **(b)** too large ($a \gg L_n$).

see in Fig. 2.6(b), only limited current can be bent away from its vertical path. In fact, this lateral diffusion of current is on the order of the diffusion length. Thus, carrier recombination takes place only at the periphery of the emitting window. This is why the efficiency of this type of LED is also low.

Fortunately, there are 4 possible ways to circumvent this problem. These 4 ways are schematically illustrated in Fig. 2.7 (a)-(d). The first method is to deposit a transparent electrode on the top surface of the LED followed by the normal evaporation of the electrode. This is depicted in Fig. 2.7(a). The goal of this method is to hopefully inject some current from the middle of the emitting window and at the same time allow the photons generated by the injection current to be freely transmitted into the air through the transparent electrode. A candidate for the transparent electrode would be indium tin oxide (ITO). This method has been tried with limited success because of the high resistance of the ITO electrode layer. As a result, most of the injection current still concentrates underneath the conventional electrode.

The second method involving heavily doping the p-AlGaAs upper confinement layer so as to draw the injection current more toward the middle of the emitting window. This is shown in Fig. 2.7(b). Again, the success of this method is limited because of the limitation on the maximum doping levels that can be achieved in p-AlGaAs layers grown by either molecular beam epitaxy (MBE) or metalorganic chemical vapor deposition (MOCVD). This is due to the incorporation of Al, which reduces the maximum impurity doping concentration achievable in the material.

The third and fourth methods are the most promising. Fig. 2.7(c) shows the third method, in which a deep isolation implantation is performed down to the active layer of the LED, followed by appropriate thermal annealing. Because the isolation implantation changes the semiconducting material into insulating material, injection current is now forced to flow through the center of the LED, or right underneath the

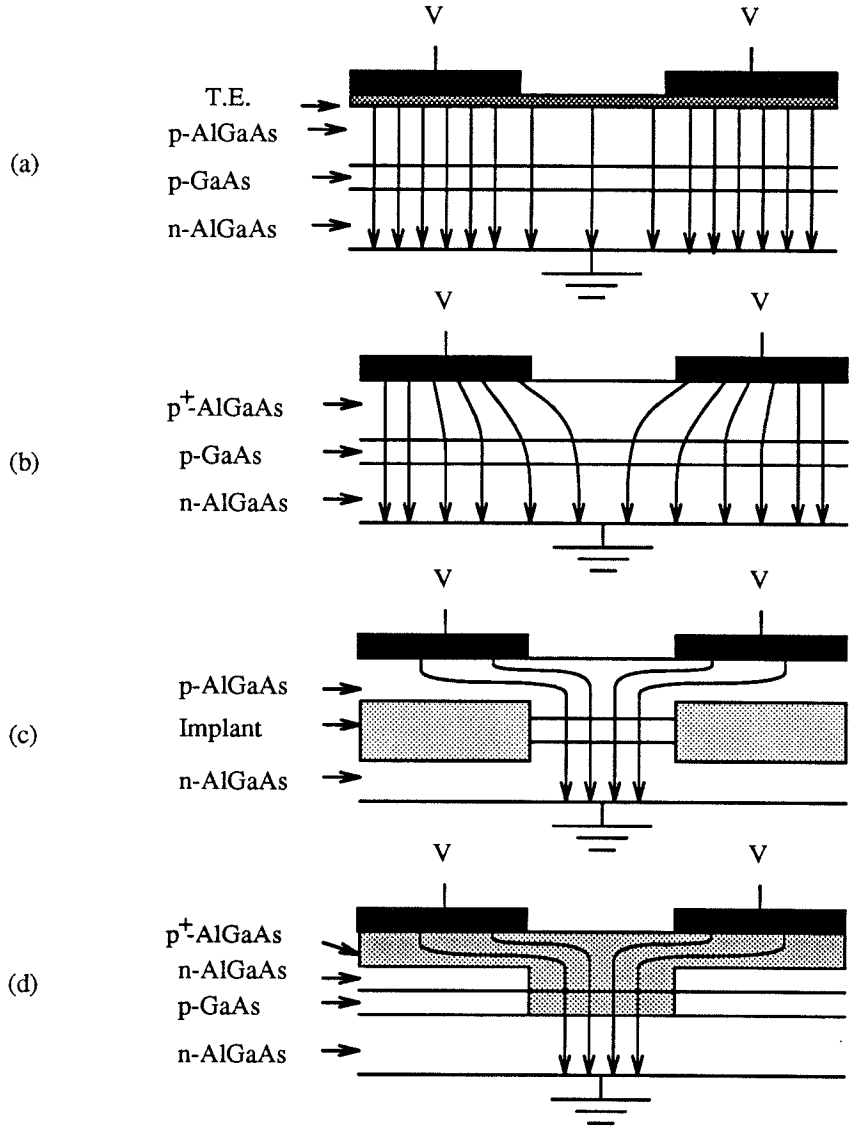


Fig. 2.7 4 possible ways to promote the injection current of the LED to flow through the region underneath the emitting window. **(a)** Insertion of a transparent electrode between the conventional electrode and the LED surface. **(b)** Heavy doping of the top p-AlGaAs top confinement layer. **(c)** Buried isolation implantation down to the active layer. **(d)** Double Zn-diffusion process.

emitting window. This has the advantage of injecting current into the LED from the side and extracting photons out efficiently from the middle. This method has been successfully adopted to fabricate vertical-cavity gain-guided surface emitting laser diodes [26].

The last method, which is shown in Fig. 2.7(d), shares a somewhat similar concept as the third method, however it starts with a different material structure and employs a different fabrication procedure. Basically, instead of fabricating the LED in a conventional double-heterojunction (DH) P-i-N structure, the LED can be fabricated in a similar double-heterojunction structure except the doping composition is N-p-N. This is accomplished by first diffusing Zn into the top n-AlGaAs confinement layer to convert it to p-AlGaAs, thereby forming the P-i-N doping composition again [27]. However, the difference is that the region into which Zn is diffused has finite dimensions. By doing the first Zn-diffusion down to the lightly doped p-GaAs active layer, followed by a second Zn-diffusion that has a larger area than the first diffusion and a diffusion front that stops inside the n-AlGaAs top confinement layer, the current is injected from the perimeter of the LED but it is forced to flow down to the cathode through the middle of the LED. It is important to realize that the existence of the reversed biased N-p junction in the P-N-p-N thyristor heterostructure underneath the injecting anode electrode is causing the current to direct itself toward the middle of the LED. To successfully form this P-N-p-N blocking thyristor, the diffusion front of the second Zn-diffusion has to be controlled to be within the n-AlGaAs top confinement layer. It is also interesting to note that by starting out with a N-p-N structure, double-heterojunction bipolar transistors or phototransistors can be easily monolithically integrated with LED's fabricated by the double Zn-diffusion technique.

2.3.4 Waveguiding Phenomenon

The waveguiding nature of the double heterostructure is another mechanism that accounts for loss in the LED's. Because the index of refraction of the AlGaAs confinement layer is lower than that of the GaAs active layer, light generated tends to be guided in the plane of the active layer. In fact, edge-emitting LED's are based on this principle. However, for surface-emitting LED's, this is not a desirable property. In actual testing, light can be observed from the perimeter of the LED, which is where the waveguide ends. In-plane superluminescence can sometimes develop in the active layer such that the properties of such superluminescent LED's resemble those of the edge-emitting lasers, except the stimulated emission is suppressed. Therefore, waveguiding should be reduced as much as possible to minimize the loss inside the LED.

2.3.5 Emission of Photons Into the Substrate

Lastly, we will address the issue of those photons that propagate down to the substrate and subsequently get absorbed. Because LED's are isotropic emitters, half of the photons generated are absorbed in the substrate because the substrate is GaAs. There is, however, a method by which a mirror can be placed between the substrate and the high-bandgap AlGaAs lower confinement layer so that these photons can be reflected back up. This is schematically illustrated in Fig. 2.8(a). Since metals cannot be grown in the same chamber with the GaAs and AlGaAs layers, this mirror has to be a dielectric mirror with composition made up of N-pair of alternating layers of low-index and high-index materials, such as AlGaAs and GaAs, which are in-situ grown on the same substrate. Let's designate the refractive index of the low-index and high-index materials to be n_l and n_h respectively, and the thickness of each layer to be a quarter-wavelength. There are two ways to arrange these alternating layers. The first is to place the low-index material below

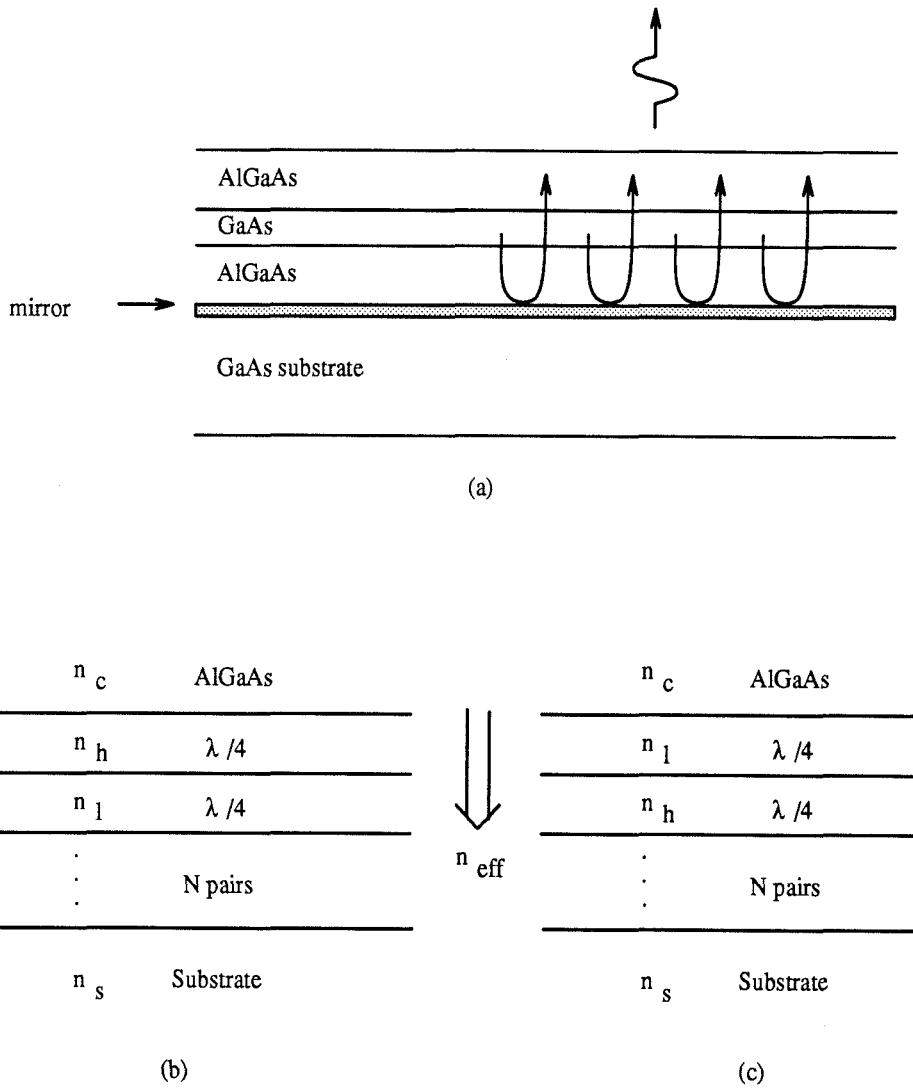


Fig. 2.8 **(a)** Schematic diagram showing that by placing a dielectric mirror between the substrate and the LED structure, photons that are originally absorbed in the substrate will be reflected back up. **(b)** Structure of the dielectric mirror that has the low-index material below the high-index material. **(c)** Structure of the dielectric mirror that has the high-index material below the low-index material.

the high-index material. The second way is the other way around. These two cases are shown in Fig. 2.8(b) and 2.8(c) respectively. Let us restrict our attention to the case that the low-index material is below the high-index material first. Let us further define a parameter, n_{eff}, that represents the effective index of all layers looked down from the interface between the lower AlGaAs layer and all the layers beneath it. From fundamental optics, the reflection coefficient for light traversing from the AlGaAs layer (n_c material) toward the substrate is given by

$$R = (\frac{n_c - n_{eff}}{n_c + n_{eff}})^2. \tag{2.44}$$

From the analysis of multiple interference in Hecht and Zajac [25], the effective index, n_{eff}, can be further expressed as

$$n_{eff} = n_s(\frac{n_h}{n_l})^{2N}. \tag{2.45}$$

Since n_h is larger than n_l, as more and more pairs of these alternating layers are added, the reflection coefficient becomes closer to 1 because n_{eff} approaches ∞ as N goes to ∞. Surprisingly, exactly the same results can be derived for the other case, in which the high-index material is below the low-index material. In this case, Eq. (2.44) still remains valid. However, the expression for n_{eff} needs to be modified into

$$n_{eff} = n_s(\frac{n_l}{n_h})^{2N}. \tag{2.46}$$

Taking the limit of n_{eff} in Eq. (2.46) as N goes to ∞, we see that n_{eff} approaches zero, which makes the reflection coefficient, R, in Eq. (2.44) equal to 1 again. Thus

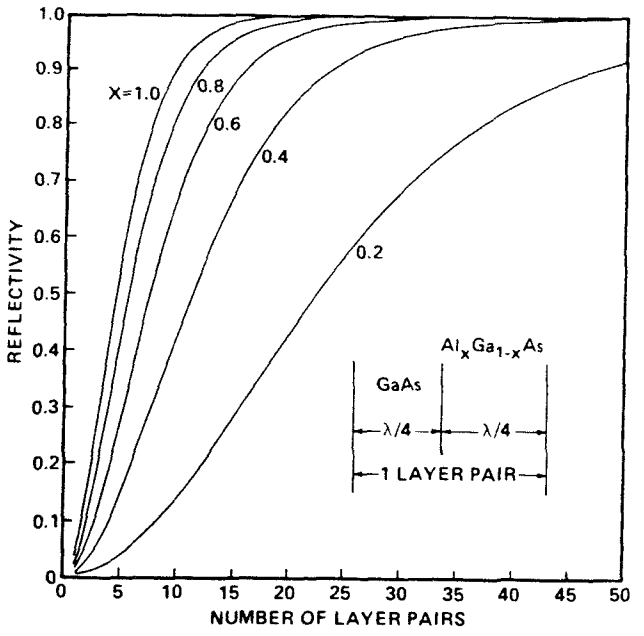
the conclusion is that regardless of the order that these quarter-wavelength alternating layers are grown, the reflectivity of the dielectric mirror improves as the number of pairs increases. There is however a subtle difference. If we look at Eq. (2.44), we notice that the reflectivity can be equal to zero when n_c is equal to n_{eff}. When N is equal to zero, which corresponds to the case of no alternating layers, n_{eff} is equal to n_s. By adding pairs of these alternating layers, n_{eff} either becomes larger or smaller depending on the order these alternating layers are arranged. If the low-index material is below the high-index material, then n_{eff} becomes larger through the multiplication of the factor, $(n_h/n_l)^2$, which is evidence in Eq. (2.45). Since n_s (GaAs) is greater than n_c (AlGaAs) to start out with, it is impossible to make n_{eff} equal to n_c. Thus, the reflectivity is a monotonically increasing function of the number of pairs added. On the other hand, if the high-index material is below the low-index material, the value of n_{eff} is reduced by the factor $(n_l/n_h)^2$, as seen by Eq. (2.47). This reduction in n_{eff} has the effect of decreasing the overall reflectivity. As n_{eff} continues to decrease, for some value of N, n_c will eventually equal to n_{eff}, which makes the reflectivity equal to zero. If N continues to increase, the reflectivity increases again and goes to 1 as n goes to infinity. Thus, in this case, the overall reflectivity decreases first and eventually diminishes for some N, and then monotonically increases to 1 as N goes to ∞. Because of this subtle difference, it is not hard to conceive that to achieve the same reflectivity, it would require slightly fewer number of pairs of these alternating layers for the case shown in Fig. 2.8(b) than it would in the other case. (This is analogous to the situation in which the reflection for the TM-wave is always less than that for the TE-wave because there exists a Brewster angle for the TM-wave, for which the reflection is zero.) It is also important to realize that the larger the difference between the index of refraction for the two materials is, the higher the reflectivity is for the same number of pairs added, or the fewer the number of pairs would be required

to achieve the same reflectivity. For the GaAs/AlGaAs system, this means that to achieve a given reflectivity with minimum number of layers, AlAs and GaAs should be used as the dielectric mirror materials. However, there are practical considerations, such as light absorption in GaAs and electrical resistance associated with the high-barrier superlattice AlAs/GaAs system. Nevertheless, the concepts discussed above are valid and are illustrated by plots shown in Fig. 2.9(a) [28], which shows the reflectivity as a function of the number of pairs needed for various aluminum mole fraction in the ternary AlGaAs compound. Aluminum mole fraction equal to 1 means that the compound is AlAs.

The spectral response of the dielectric mirror is also an important design consideration since LED's are known to have a broad emitting spectrum. This can be seen in Fig. 2.9(b), which shows the reflectivity of a 20-pair GaAs/$Al_{0.6}Ga_{0.4}As$ dielectric mirror as a function of wavelength [28]. It is interesting to note from Fig. 2.9(b) that the high reflectivity region spans a finite spectral width. In fact, as more layers are added, not only does the absolute reflectivity at the designed wavelength increases, but the spectral width over which the high reflectivity is obtained also increases [28] . However, this increase in the spectral width is limited by the locations of the same nulls. Thus, by adding more layers, the spectral response changes from more sin(x)/x-like to more rectangle-like and the locations of the nulls will remain unchanged.

The same concepts can be applied to making anti-reflection coating, which is useful for extracting more photons out the LED's. As mentioned previously, this can be accomplished by placing the low-index material underneath the high-index material and choosing an appropriate number of pairs such that $n_c = n_s(n_l/n_h)^{2N}$. It is however not always possible to find an integer N that satisfies this equation. If this is the case, n_l and n_h have to be changed so that for some integer N, the equation is satisfied. The last factor to consider when designing the mirror is that

(a)



(b)

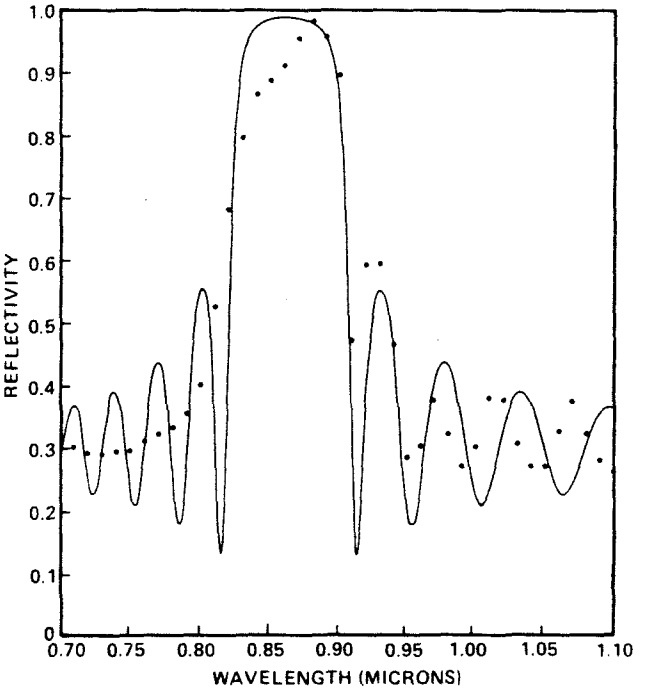


Fig. 2.9 **(a)** Reflectivity as a function of N, which is the number of pairs of alternating layers, $GaAs/Al_xGa_{1-x}As$, with each layer being $\lambda/4$ thick for various aluminum mole fraction, x. **(b)** Experimental and theoretical reflectivity of a 20-pair $GaAs/Al_{0.6}Ga_{0.4}As$ dielectric mirror [28].

the effective index, n_{eff}, seen by the active layer at its boundary with the lower AlGaAs confinement layer will be different from that seen at the interface between the lower AlGaAs confinement layer and the dielectric mirror unless this AlGaAs layer has a thickness of integral multiple of $\lambda/2$. Thus, besides the active layer, whose thickness should be chosen to optimize the LED output power, and the two AlGaAs confinement layers, which are integral multiple of $\lambda/2$ thick, the rest of the layers should be $\lambda/4$ thick in order to maximize the photon extraction efficiency. An example of the LED epitaxial layers that are designed to have a dielectric mirror at the bottom and a dielectric anti-reflection coating on the top using GaAs/AlGaAs $\lambda/4$ superlattice is proposed and is shown in Fig. 2.10.

2.4 Photon Collection Efficiency

Because the index of refraction of GaAs is much higher than that of air, light that eventually penetrates into the air is radiated almost in all directions. Furthermore the portion of the light that is traveling at very shallow angles with respect to the surface of the LED's is not likely to be usable, the efficiency of the LED's is further limited. However, this dilemma can be efficiently eliminated to certain extent by the incorporation of an integrated device that would focus the photons radiated photons from the LED. One such device is an integrated lens fabricated on top of the LED. Depending on the severity of beam divergence, lenses of different numerical aperture can be monolithically fabricated in GaAs-based or InP-based materials. Ostermayer et al. [29] have developed a photoelectrochemical method for defining and etching integral lenses on InP LED's. An LED wafer is immersed in an electrolyte and biased at a potential at which the etch rate is directly proportional to light intensity. The image of a photomask is projected onto the surface of the wafer to produce a spatial variation of light intensity to etch out the desired shape. Figure

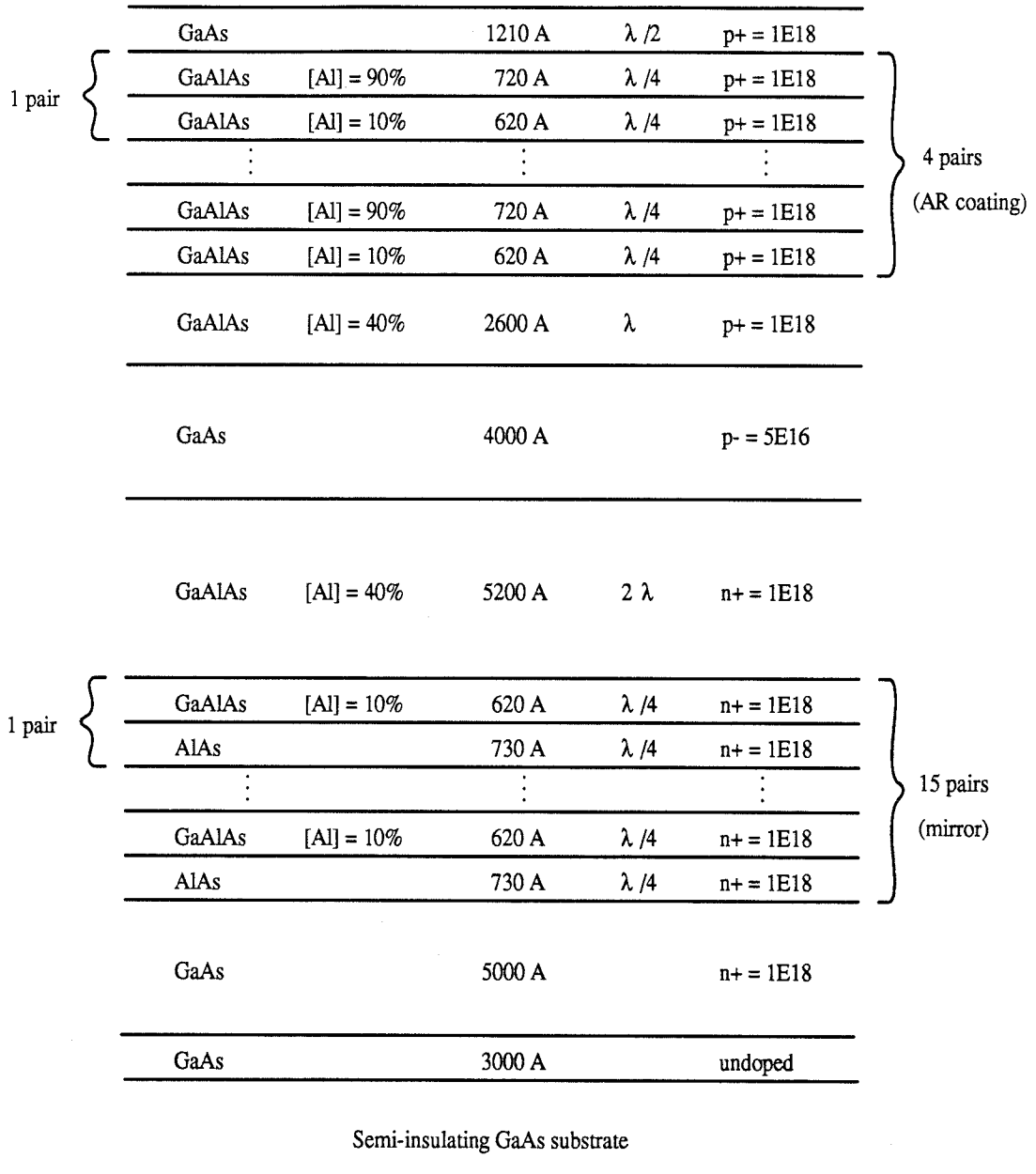


Fig. 2.10 Structure of the LED using GaAs/AlGaAs superlattice to form a reflecting mirror at the bottom and an anti-reflection coating above the LED in order to maximize the photon extraction efficiency.

(a)

(b)

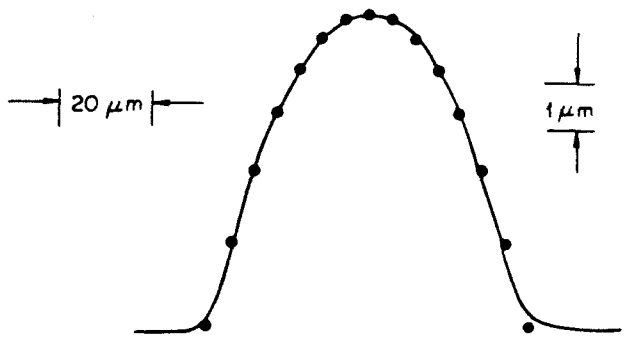
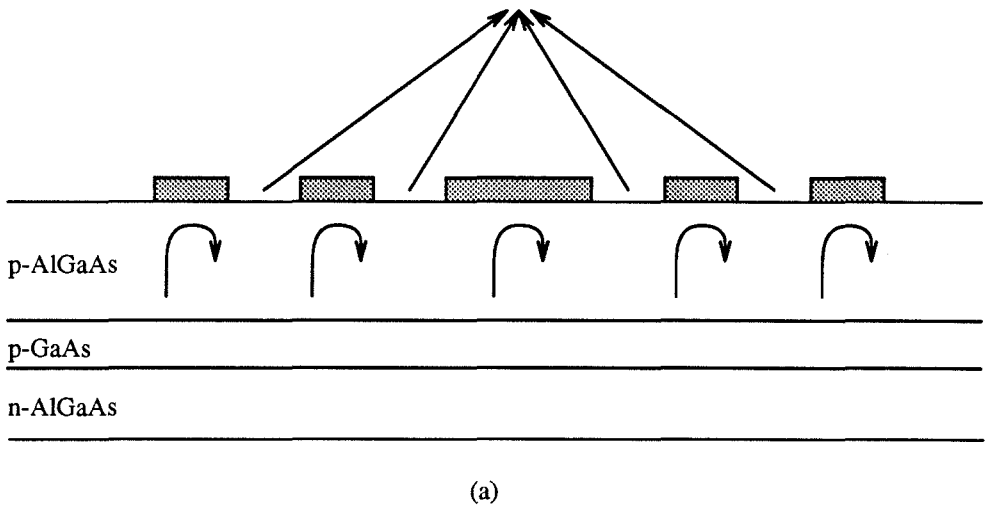


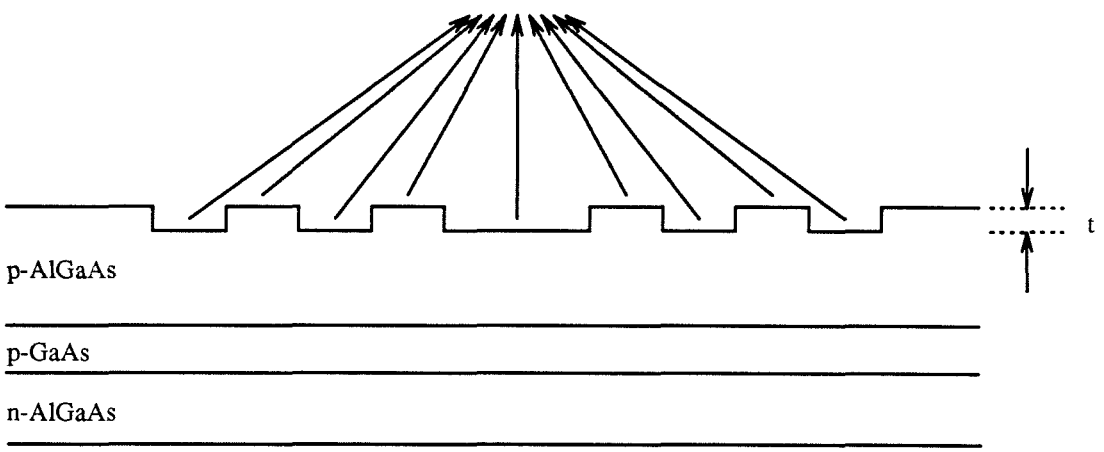
Fig. 11 **(a)** Scanning electron micrograph of the integrated lenses fabricated in a InGaAsP/InP double heterostructure by photoelectrochemical etching technique [29]. **(b)** The surface profile of the integrated lens fabricated by the same technique.

2.11(a) shows the scanning electron micrograph of the lenses on the LED's defined by this method. Figure 2.10(b) shows the surface profile of a lens. As we can see from these figures, very smooth surfaces are obtained on the lenses. As a result, the efficiency of the LED has been improved by 60 %. Heinen [30] has also developed a chemical etching technique to define the lenses that are integrated with the LED. An improvement factor of 300 % has been demonstrated by using this technique. Wada et al. [31] demonstrated similar improvements by using ion beam etching on InGaAsP/InP LED's. While these results seem encouraging, it is important to realize that these techniques were developed for InP-based LED's, which are an important element in fiber-optic communication systems. Though similar in principle, methods of fabricating monolithically integrated lenses on GaAs-based LED's involve finding appropriate etching solutions and gases. Therefore, this should not constitute a problem. However, of more concern is the issue of microlithographic control in the fabrication of GaAs-based LED's. The techniques just mentioned for fabricating lenses on InP-based LED's are appropriate for LED's of large emitting diameters. For the GaAs LED's that are of interest of us, their emitting areas typically range from 10 to 30 μm in diameter. Thus, etching a profile within the diameter of the LED's has to be precisely controlled in order to obtain smooth profile as well as the correct numerical aperture.

Another way to improve the collection of emitted photons is to build Fresnel zone plates on the emitting surface of the LED's. Shown in Fig. 2.12 (a) and (b) are two ways of building Fresnel zone plates in GaAs. The first way, depicted in Fig. 2.12(a), employs metalizations spaced by a certain gap from each other to define each individual zone. The gap between adjacent metals and the width of the metal are so designed that light emitted from adjacent zones interfere destructively and light emitted from every other zone interfere constructively. In order to block off the destructively interfering components, metals are evaporated to reflect off these



(a)



(b)

Fig. 2.12 **(a)** Configuration of a Fresnel zone plate fabricated by evaporation of metals spaced by a certain distance from each other. **(b)** Configuration of a Fresnel zone plate fabricated by etching into the material by a depth, t. The collection efficiency of this Fresnel zone plate is twice that in (a) since a π phase shift is introduced between adjacent zones.

components of light. This results in allowing only those components of light that are constructively interfering to radiate outward and achieving a focusing effect. This effect is clearly illustrated in the figure. The drawback of this method is the limited efficiency in collecting the photons because practically half of the light is blocked off by the metals that define the Fresnel zones. Thus, a more clever way to implement a Fresnel zone plate is to introduce a π phase shift between the adjacent zones of the Fresnel zone plate. This added π phase shift allows the originally destructively interfering components, which are π out of phase with respect to the adjacent zone, to be in phase again at the focusing point. Thus, a factor of 2 in the improvement of photon collection efficiency can be expected. This is illustrated clearly in Fig. 2.12(b), in which a simple etching process is employed to introduce the π phase shift between adjacent zones of the zone plate. The depth of the etch can be easily calculated. If we denote the depth of etch to be t, then the following equation can be formulated to produce the π phase shift.

$$\frac{2\pi}{\lambda}(n_{GaAs} - n_{air})t = \pi. \tag{2.47}$$

Using $n_{GaAs} = 3.6$, $n_{air} = 1$ and $\lambda = 0.87\mu$m, we obtain $t = 0.17\mu$m, which means the recess is only 0.17 μm deep. However, there is a concern over the resolution of the photolithography because it is difficult to define many zones on a tiny LED window region. Nevertheless, it is a very practical way of implementing Fresnel zone plate in GaAs devices.

2.5 LED Speed

The speed of the LED is governed by two factors. The first is the RC time

constant associated with the double-heterostructure P-i-N diode. The second factor is the minority carrier lifetime in the active layer. When the step current is applied to the LED, because of the capacitance and the series resistance associated with the turn-on process of the LED, the junction voltage rises to its final value with a delay of RC. Thereafter, minority carriers are injected into the active layer. However, carrier recombination does not take place immediately. Instead, these carriers remain on an average of τ before they are recombined. Thus, there is a total delay of (RC + τ) before photons are generated. The dominant contribution to the capacitance is from the capacitance due to the depletion region, which narrows as the junction voltage increases, and the capacitance associated with the carrier charge-up by the injecting minority carriers. For large-area LED's, this can be a severely limiting factor in increasing the speed of the device. However, for small-area LED's, the minority carrier recombination lifetime is the dominating delay. Its expression was given earlier in Eq. (2.17), which depends on radiative recombination lifetime, nonradiative recombination lifetime and interfacial recombination. The presence of nonradiative recombination and interfacial recombination act to reduce the overall lifetime, thus the risetime, of the LED's. However, this is achieved at the expense of lower optical output power from the LED's because these recombination processes do not contribute to the photon generation process. In the absence of the nonradiative and interfacial recombinations, the overall recombination lifetime is reduced to the radiative recombination lifetime, τ_r, which is given by Eq. (2.32). In this case, the doping concentration in the active layer as well as the injection current density play important roles in determining the overall speed of the LED. For a LED that is modulated by a small signal, the 3-dB bandwidth is

$$f_{3dB} = \frac{1}{2\pi\tau}. \tag{2.48}$$

Computer simulations on the speed of the LED's will be shown in the next section to highlight the significance of active layer doping concentration and the injection current density on the speed of the LED.

2.6 Computer Simulations

It is useful to be able to observe the behavior of the LED power generation efficiency as a function of various material and circuit parameters and the dependence of the LED speed on the same parameters. Eq. (2.14)-(2.16) and (2.32) are used to predict the photon generation efficiency of the LED. The same Eq. (2.32) and (2.48) are used to describe the speed of the LED. The photon extraction efficiency is taken into the consideration of the overall LED external quantum efficiency by incorporating Eq. (2.38)-(2.40). In these simulations, nonradiative recombination is assumed to be negligible. Thus, the factor τ/τ_r in Eq. (2.14) is assumed to be unity. The parameters used in the simulation have the following values :

$\tau_{nr} = \infty$

n_{GaAs} at 0.87μm $= 3.6$

N_A, $= 10^{16}$ cm^{-3}

W $= 1.5$ μm,

J $= 1000$ A/cm^2,

$\lambda = 0.87$ μm

s $= 1000$ cm/sec [19]

α at 0.87 μm $= 10^4$ /cm

$D_n = 100$ sec^2/cm.

Unless otherwise stated, these values are used in all simulations. For plots in which the external quantum efficiency of the LED with anti-reflection coating is shown, these anti-reflection coating layers are assumed to have a refractive index of 1.8 and

a thickness of exactly quarter wavelength. The refractive index of 1.8 is close to the actual refractive index of silicon nitride deposited by a thermal chemical vapor deposition system. Finally, Auger recombination is not taken into consideration in the plot where the external quantum efficiency of the LED is plotted against the active layer doping concentration.

Figure 2.13 shows the radiative recombination lifetime as a function of the active layer doping concentration for injection current density from 0 to 10^4 A/cm^2. When injection current density is zero, there are minority carriers in the active layer. Thus, its radiative recombination lifetime is expected to depend on the doping concentration. In Eq. (2.33), we see that the relationship between the radiative recombination lifetime and the active layer doping concentration is inversely proportional. This is evidenced in this figure in which a slope of -1 on a log-log plot is observed for the J = 0 curve. However, as the injection current density increases, these injected minority carriers start to affect the recombination process. Because there are more total minority carriers present in the active layer now, probability of recombination increases. Thus, the radiative recombination lifetime decreases. Nevertheless, for a heavily doped active layer, the overall lifetime is still dominated by the doping concentration unless the injection current density has contributed to a larger percentage of the total minority carriers in the active layer. This behavior is clearly manifested in these plots.

Figure 2.14 shows the radiative recombination lifetime as a function of the injection current density for various active layer doping concentration varied from 10^{16}/cm^3 to 10^{19}/cm^3. This is similar to the previous plot except the parameters of variation are interchanged. When the active layer doping concentration is low, such as 10^{16}/cm^3, the radiative recombination lifetime varies with injection current density in an inverse square root fashion, as seen in Eq. (2.34). This dependence is observed by the -1/2 slope for this curve on a log-log plot. However, as the active

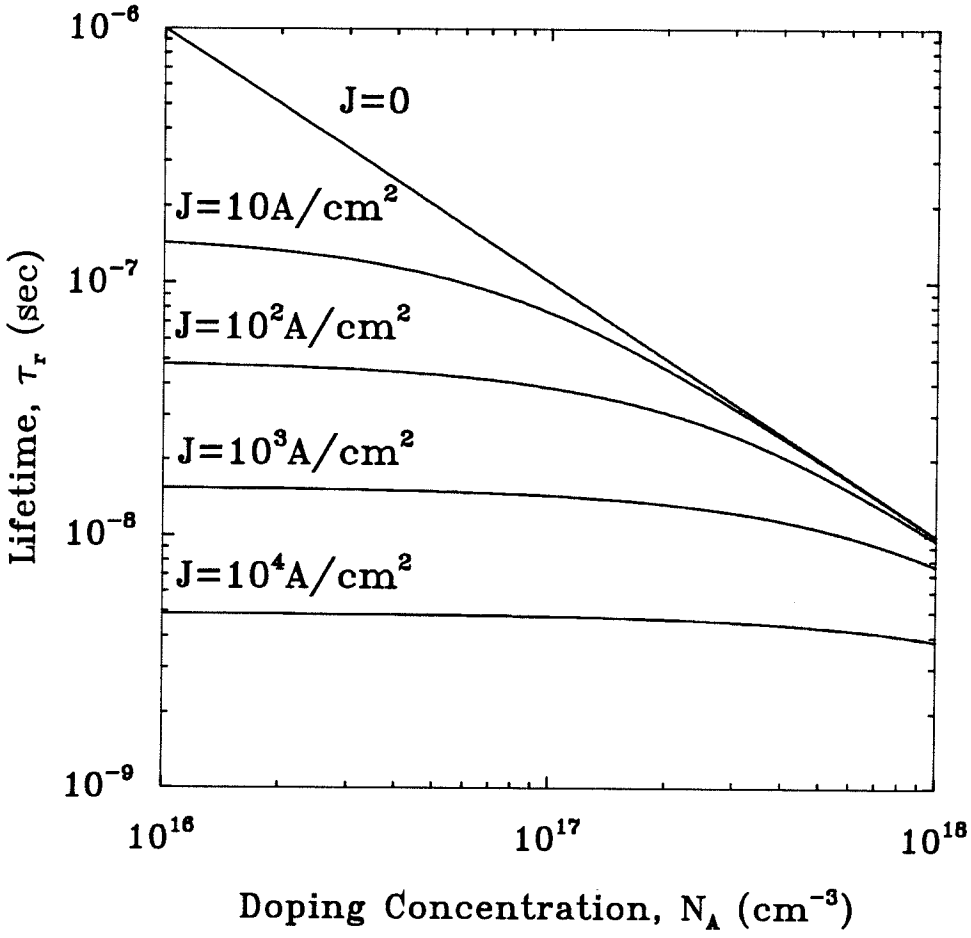


Fig. 2.13 Radiative recombination lifetime as a function of the active layer doping concentration for various injection current densities.

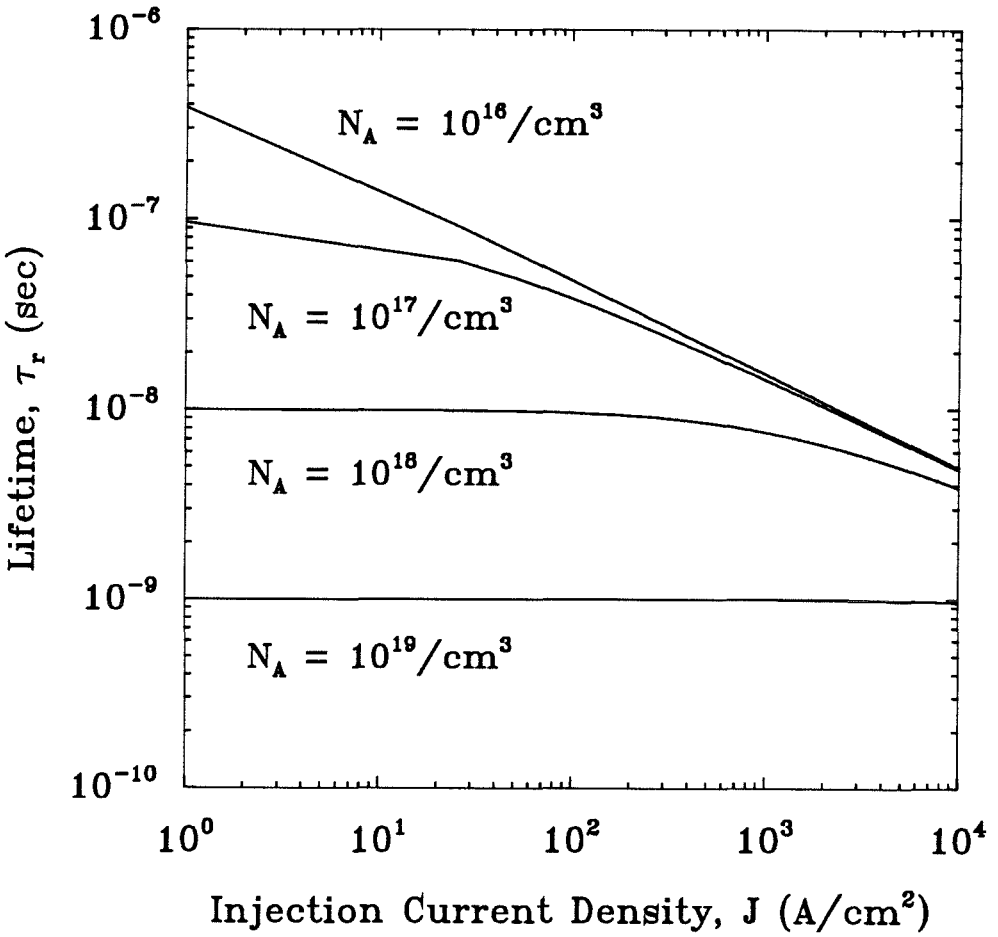


Fig. 2.14 Radiative recombination lifetime as a function of the injection current density for various active layer doping concentrations.

layer doping concentration increases, the radiative recombination lifetime decreases more dramatically at low injection current density than at high injection current density. When the doping concentration increases to $10^{19}/\text{cm}^3$, the dependence of radiative recombination lifetime on injection current density has diminished. Again, these results are consistent with the physical intuitions.

Figure 2.15 illustrates a somewhat less obvious plot. It plots the radiative recombination lifetime as a function of the active layer thickness for various injection current densities. It is not at all clear initially what role the active layer thickness plays in determining the radiative recombination lifetime. A closer examination reveals that as the active layer thickness is decreased, the injected carriers have to build up their concentration inside the active layer in order to maintain the same total charge. Maintaining the total charge inside the active layer is necessary because, according to Eq. (2.28), the current flowing through LED is equal to the total charge inside the active layer divided by the recombination lifetime. Furthermore, according to Eq. (2.25), as the concentration of these excess injected carriers increases, the overall recombination lifetime decreases. Thus, the increase in the recombination lifetime due to the increase in the active layer thickness as shown in Fig. 2.15 is consistent with this argument. This is why the active layer thickness is a very important parameter in the design of the LED. Not only will it determine the speed of the LED, but also it will significantly affect its quantum efficiency as we shall see in the next few plots. One minor detail on the process of the increase in the carrier concentration should be pointed out. As the carrier concentration is being increased due to the decrease in the active layer thickness, the slope of the carrier concentration profile is maintained constant. This is again because of the constant current flowing through the LED, which fixes the slope of the minority carrier concentration profile. Thus, as the active layer thickness gets smaller, we can envision the whole process to be an increase in the minority carrier concentra-

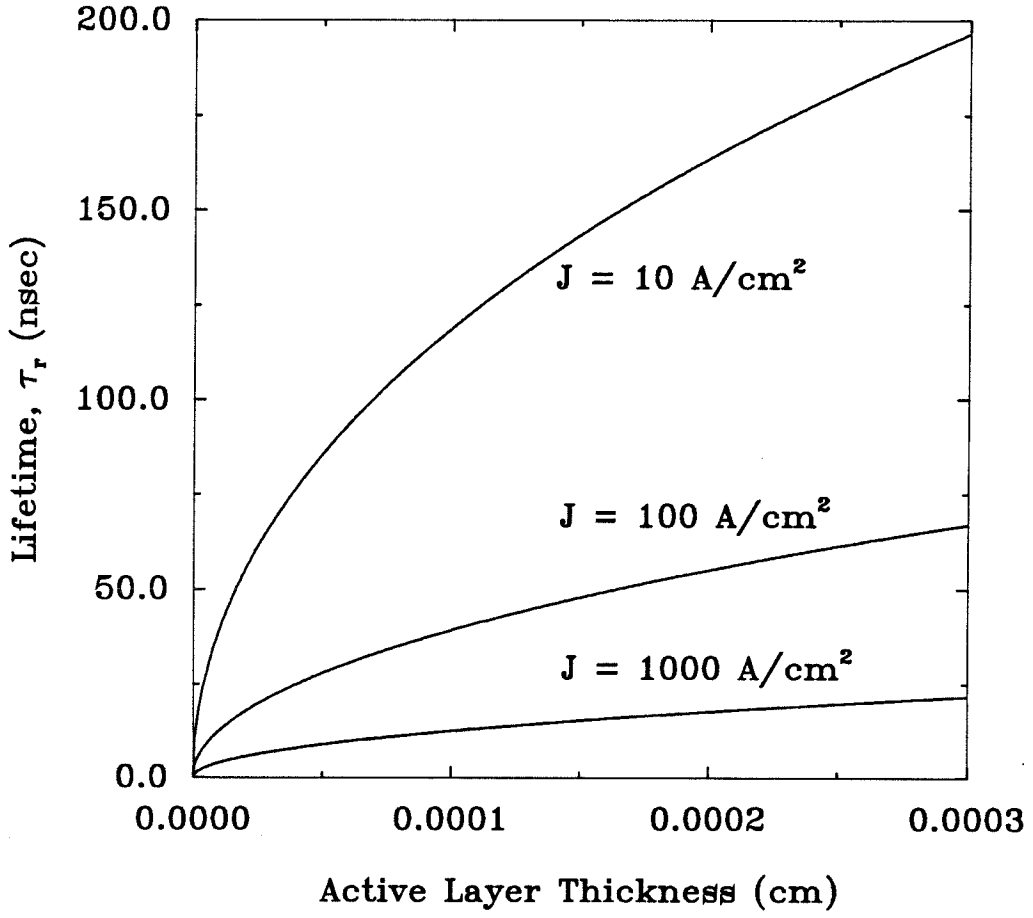


Fig. 2.15 Radiative recombination lifetime as a function of the active layer thickness for various injection current densities.

tion inside the active layer with a constant slope in the concentration profile and the same number of charges inside the active layer.

We have seen from the previous three figures that, in order to shorten the radiative recombination lifetime so as to decrease the risetime of the LED, we can increase the doping concentration of the active layer, increase the injection current density or decrease the thickness of the active layer. The same principles can be applied to increasing the modulation bandwidth of the LED's. Figure 2.16 illustrates the 3-dB bandwidth of the LED as a function of the active layer doping concentration for various injection current densities. As expected, the bandwidth increases as a function of the doping concentration for regions in which the injection current density is low. However, if the injection current density becomes a significant factor in contributing to the total minority carrier concentration in the active layer, the bandwidth becomes limited by the injection current density until such a condition is no longer valid again. This is again clearly seen in the curve with an injection current density of 10^4 A/cm^2.

Similar characteristics are shown in Fig. 2.17, in which the 3-dB bandwidth of the LED as a function of injection current density for various the active layer doping concentrations is plotted. In this case, the dependence of the bandwidth on the injection current density seems to be more pronounced at low active layer doping levels, such as 10^{16} /cm^{-3}. With increasing doping concentration, the bandwidth's dependence on the injection current density diminishes as the doping concentration in the active layer starts to play a dominating effect in the speed of the LED.

Next, we show the factors that affect the LED external quantum efficiency. First, the external quantum efficiency of the LED as a function of the active layer thickness for various interfacial recombination velocities and zero self-absorption coefficient is plotted Fig. 2.18. Let's look at the curve with zero interfacial recombination velocity first. When the interfacial recombination velocity and the

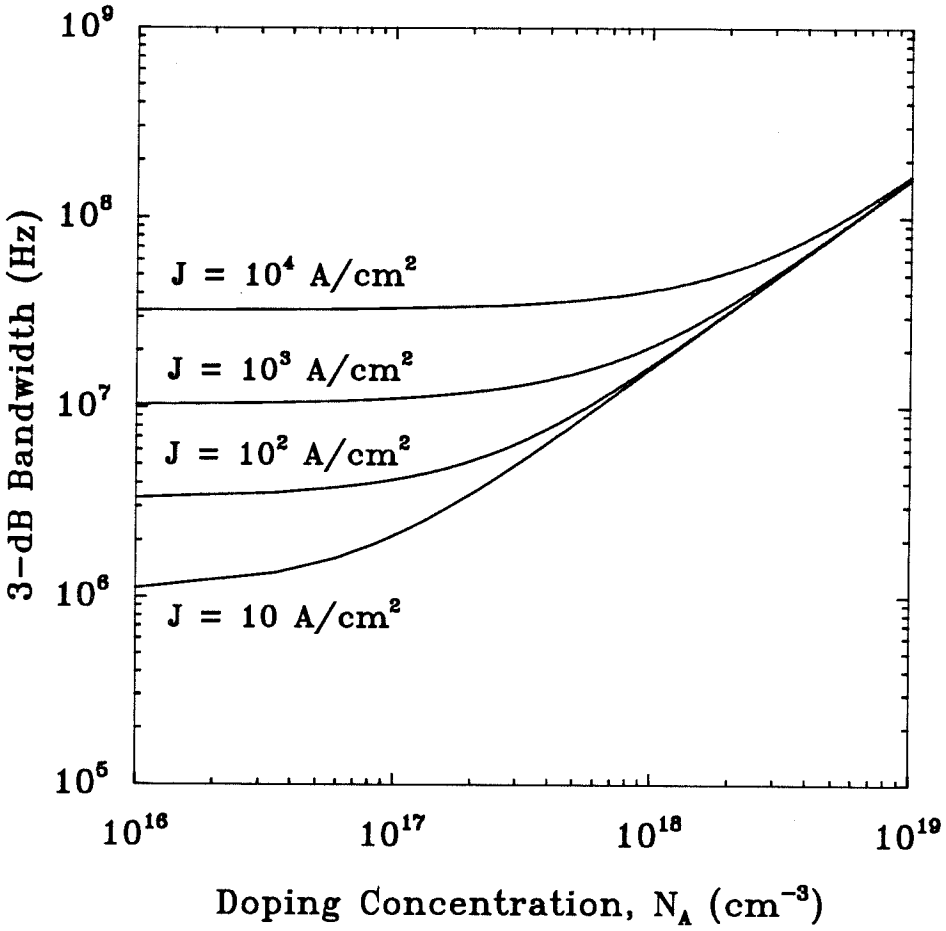


Fig. 2.16 3-dB bandwidth of the LED as a function of the active layer doping concentration for various injection current densities.

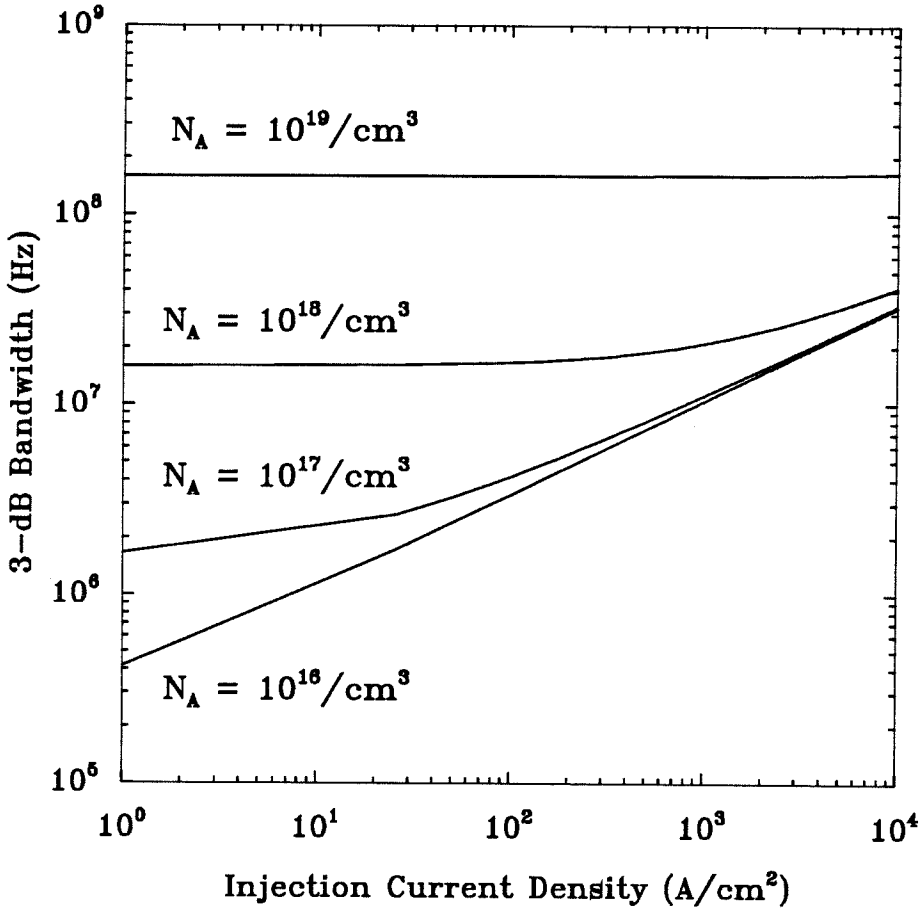


Fig. 2.17 3-dB bandwidth of the LED as a function of injection current density for various the active layer doping concentrations.

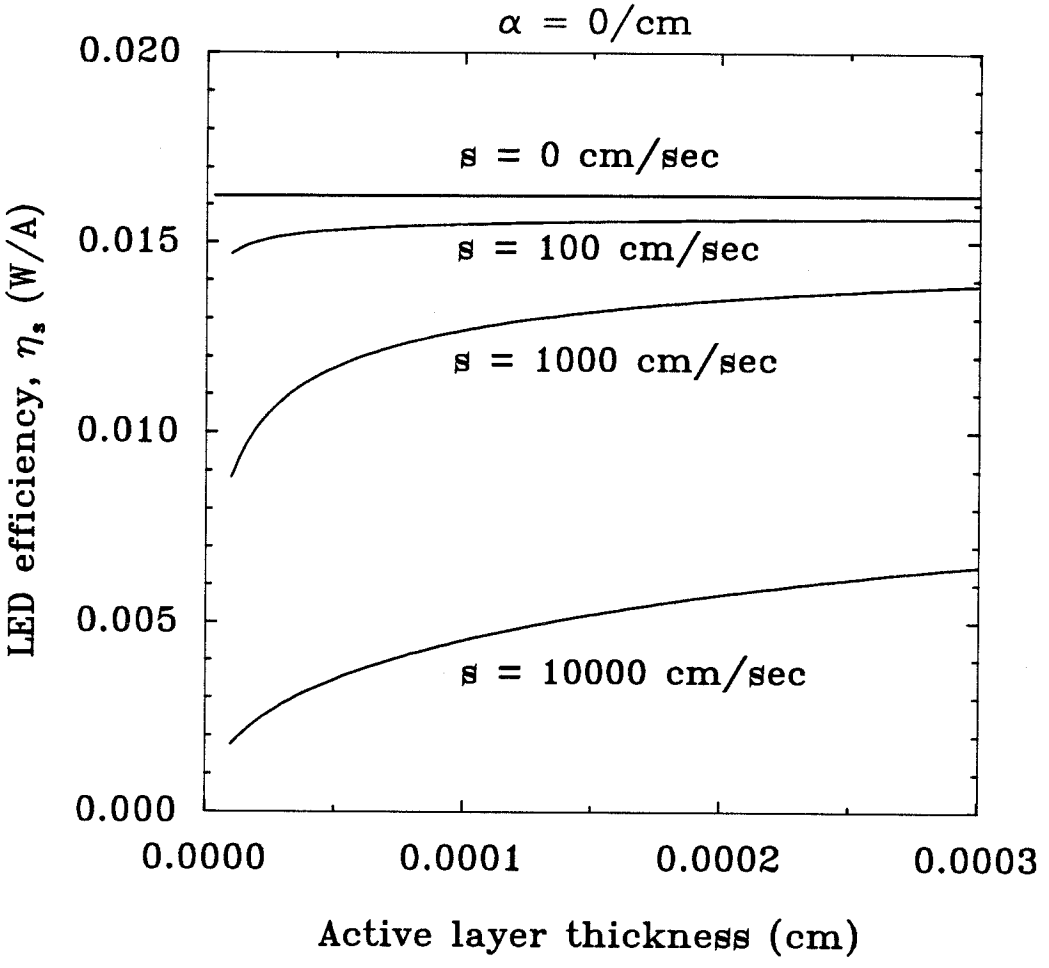


Fig. 2.18 LED external quantum efficiency as a function of the active layer thickness for various interfacial recombination velocities and zero self-absorption coefficient.

self-absorption coefficient are both zero, then all the current injected goes into generating photons which escape from the active layer without any self-absorption at all. This is why the external quantum efficiency of the LED does not depend on the active layer thickness at all. However, as a finite interfacial recombination is introduced, the effect of decreased LED external quantum efficiency for thin active layers is immediately apparent. This is because as the active layer gets thinner, the effect of interfacial recombination becomes very important for the whole LED. In other words, the interfacial recombination current becomes a larger portion of the total current. Thus, the remaining current to generate photons is correspondingly reduced. This is why interfacial recombination has a detrimental effect in the LED with thin active layers. In fact, as this interfacial recombination is increased, the effect of reduced LED external quantum efficiency for thin active layers becomes more severe. It not only pulls down the LED external quantum efficiency for thin active layers, but also drastically reduces overall the LED external quantum efficiency across the whole range of the active layer thickness, even for very thick active layers. Thus, interfacial recombination should be minimized in order to maximize the power efficiency of the LED.

Figure 2.19 shows similar plots as in Fig. 2.18 except the parameters of variation are reversed. Specifically, the LED external quantum efficiency as a function of the active layer thickness for various self-absorption coefficients and zero interfacial recombination velocity is shown. Again, let's start with the curve which has zero self-absorption coefficient. When both the self-absorption coefficient and the interfacial recombination velocity are zero, the external quantum efficiency is expected to remain constant as the active layer thickness is varied according to the argument presented in the previous paragraph. However, as a finite self-absorption is introduced, the external quantum efficiency of the LED decreases as the active layer becomes thicker. This is because the photons generated have to travel a longer

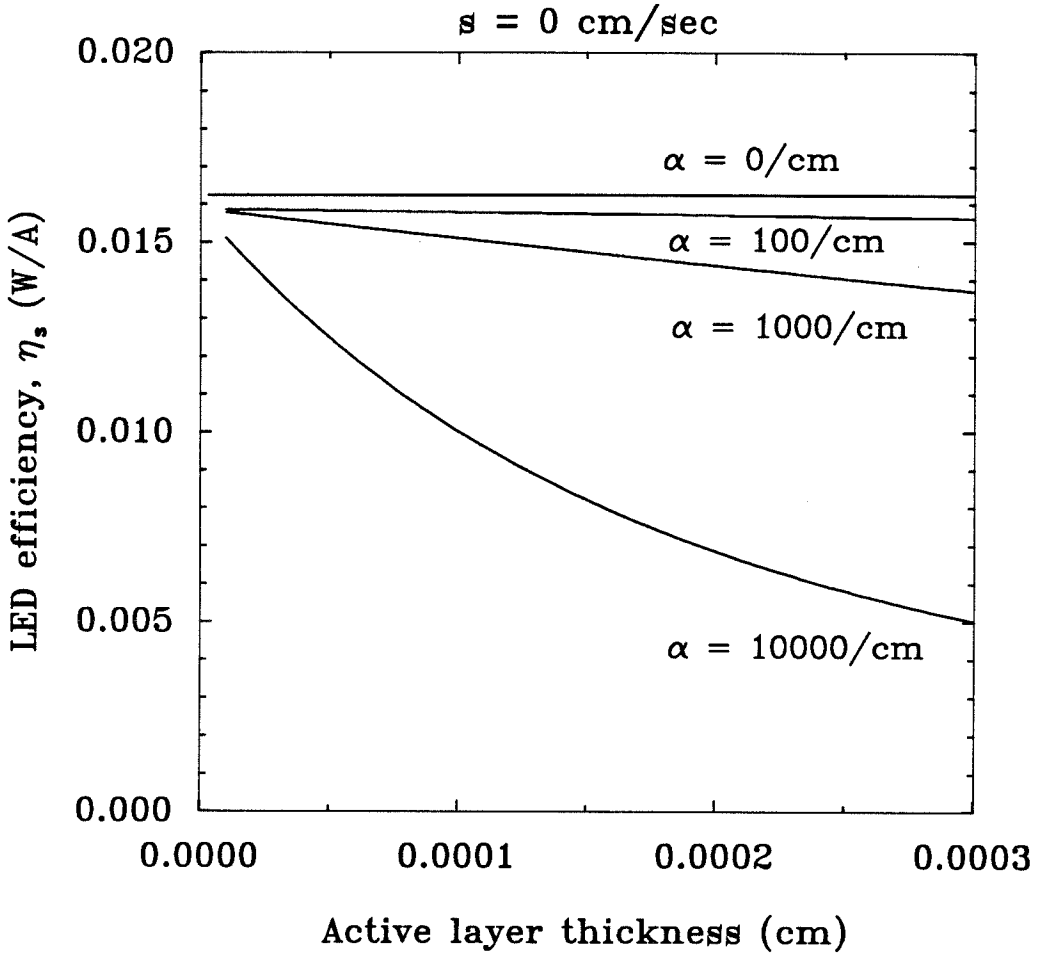


Fig. 2.19 LED external quantum efficiency as a function of the active layer thickness for various self-absorption coefficients and zero interfacial recombination velocity.

distance before totally escaping from the active region of the LED. Because photons are traversing in the same material that generates them, these photons are subjected to some absorption. Thus, the longer they have to travel, the more they will be absorbed. This is why thick active layers are on the other hand very undesirable for the power efficiency of the LED's. As the self-absorption coefficient is increased, the effect of reduced LED external quantum efficiency, especially for larger active layer thickness, becomes more pronounced and eventually severely limits the usefulness of the LED.

When we combine the effect of both nonzero self-absorption and interfacial recombination, an optimized active layer thickness which maximums the LED external quantum efficiency can be expected. This is shown in Fig. 2.20. Because interfacial recombination plays a dominating role for the LED with a thin active layer and likewise self-absorption plays a dominating role for the LED with a thick active layer, an active layer that is too thin or too thick is undesirable for the LED. Consequently, appropriate thickness of the active layer should be carefully designed. For the parameters used in this simulation, an active layer thickness of approximately 0.3 μm should be used. The external quantum efficiency of the LED with proper anti-reflection coating is also shown to contrast the improvement that can be obtained by anti-reflection coating.

Next, we consider the effect of active layer doping concentration on the external quantum efficiency of the LED. If we ignore first the existence of Auger recombination, then according to Eq. (2.32), an increase in the active layer doping concentration will result in a decrease in the minority carrier lifetime. This has the effect of also decreasing the minority carrier diffusion length as the diffusion length is given by the square root of the product of the diffusion coefficient and the minority carrier lifetime. Since diffusion length is a characteristics measure of how deep the minority carriers are injected into the active layer, a smaller diffusion

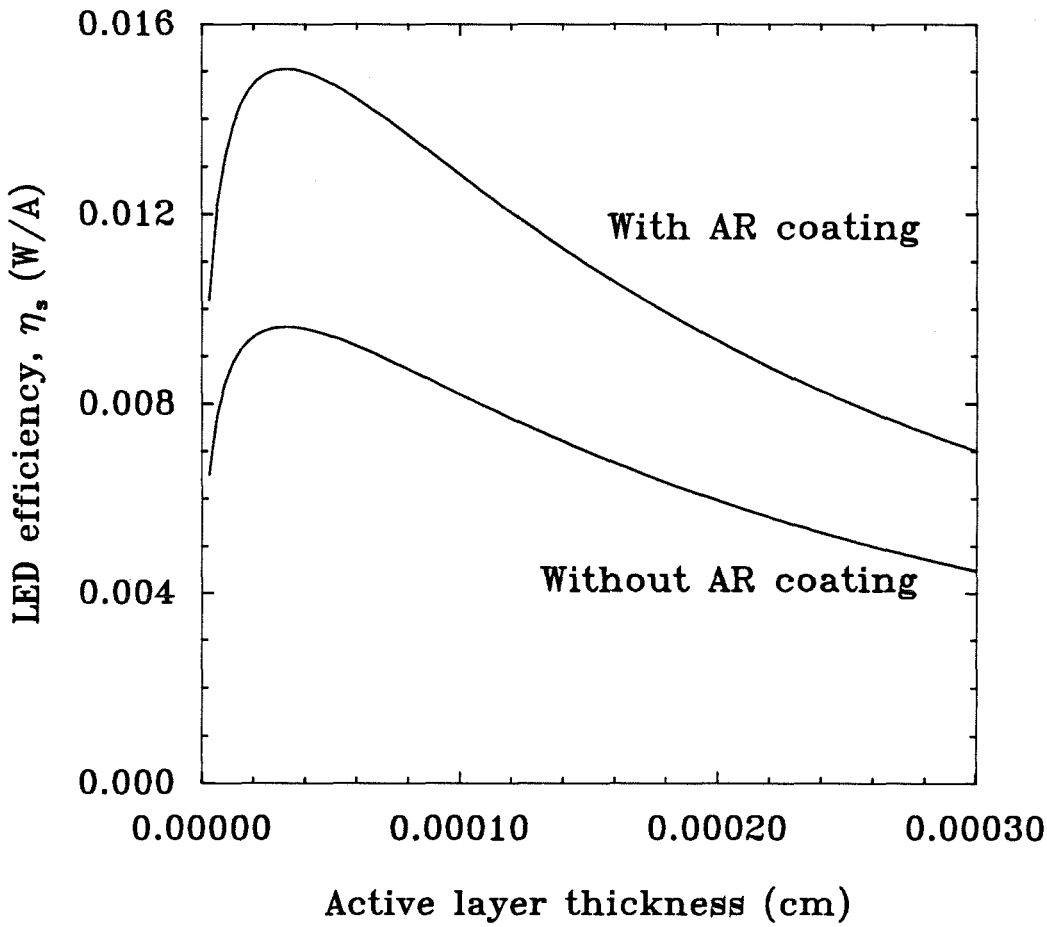


Fig. 2.20 LED external quantum efficiency as a function of the active layer thickness with nonzero self-absorption coefficient and interfacial recombination velocity.

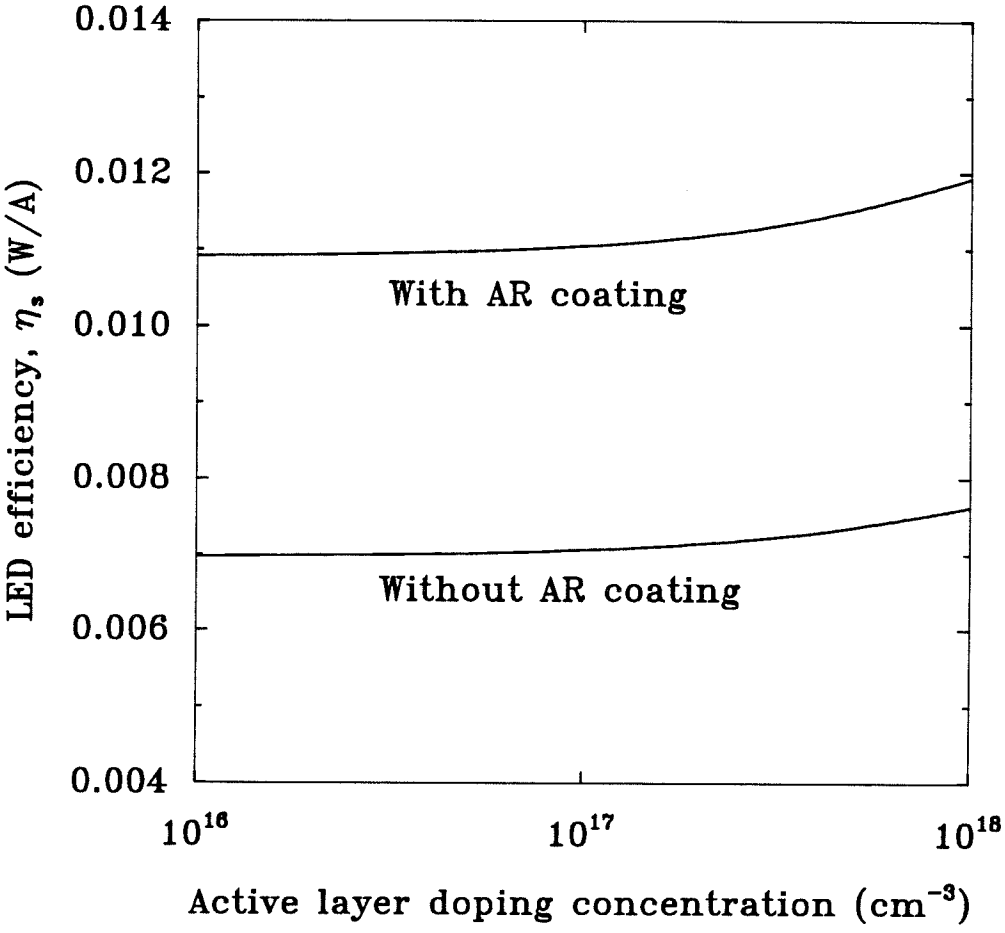


Fig. 2.21 LED external quantum efficiency as a function of the active layer doping concentration without considering the effect of Auger recombination. Because Auger recombination dominates at high doping levels, the LED efficiency should decrease as the doping concentration increases.

length means that minority carriers are closer to the injection junction. In terms of Fig. 2.1(a), this means that the minority carrier distribution has changed from that shown in Fig. 2.1(c) to a distribution that has a higher concentration of carrier near x = 0 and a lower concentration near x = w. This has two consequences. The first is that the significance of the interfacial recombination at the x = w heterojunction is less because the minority carriers are now farther away from this heterojunction. Therefore, the external quantum efficiency should increase. The second consequence however creates an opposite effect. Because the minority carriers are now farther away from the x = w heterojunction, generated photons are also farther away from the x = w heterojunction. This means that these photons have to travel a longer distance before they can escape from the active region of the LED. Thus, self-absorption of photons is also higher. As a result, it is not clear as to whether the external quantum efficiency will increase or decrease as the active layer doping concentration increases. For the parameters used in this simulation, there is a slight increase in the LED external quantum efficiency, as shown in Fig. 21. What this implies is that the effect of interfacial recombination is bigger than that of self-absorption for the particular values of parameters used. If the values of the interfacial recombination velocity and self-absorption are changed, then the external quantum efficiency of the LED may not necessarily show the same trend any more.

If we now consider the Auger effect, then a very different picture will emerge. As we discuss before, Auger recombination is a nonradiative recombination process. Its presence will decrease the overall external quantum efficiency of the LED. Increasing the active layer doping concentration however acts to promote Auger recombination. Thus, increasing the doping concentration will actually decrease the LED external quantum efficiency. In fact, heavy doping concentration is used to increase the speed of the LED at the expense of its efficiency. This is because Auger recombination

reduces the overall minority carrier recombination lifetime as shown in Eq. (2.17). As the overall minority carrier lifetime is reduced, the risetime of the LED shortens and the bandwidth of the LED increases at the same time. Therefore, this presents a tradeoff between the speed and power of the LED. If speed is not an important consideration, then the active layer doping concentration should be minimized in order to minimize the Auger recombination and maximize the output power of the LED. If speed is an important consideration in the application of the LED, then the active layer doping concentration should be as high as possible.

Lastly, we consider the effect of injection current density on the external quantum efficiency of the LED. Injection current density affects the LED external quantum efficiency in the same manner that the active layer doping concentration affects the LED external quantum efficiency. Increasing injection current density decreases the minority carrier lifetime through the same Eq. (2.32). As the minority carrier lifetime is decreased, the minority carrier diffusion length is also decreased. Thus, using the argument presented previously for the doping concentration, we come to the same conclusion as before. That is, a shorter minority carrier diffusion length helps to increase the LED external quantum efficiency due to the carriers being farther from the x = w heterojunction and thus being affected less by the interfacial recombination. On the other hand, a shorter minority carrier diffusion length hurts the LED external quantum efficiency due to the carrier having to travel a longer distance before escaping from the active region. The end result using the current LED parameters shows that increasing injection current density increases the LED external quantum efficiency. This effect is shown in Fig. 2.22. The increase in this case is more dramatic than that shown in the previous figure, in which the active layer doping concentration is varied. This is because minority carrier lifetime for the current LED parameters is more sensitive to the change in the injection current density than to the change in the doping concentration, which is only 10^{16} /cm^3 in

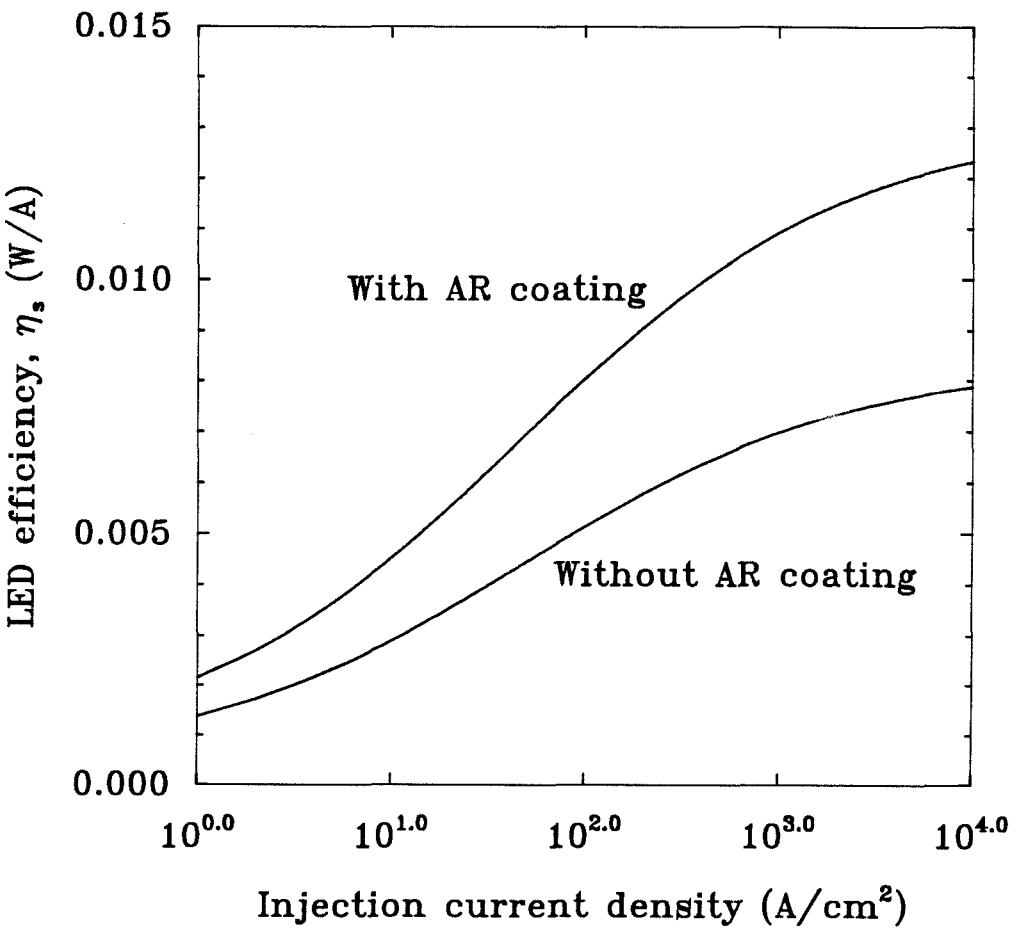


Fig. 2.22 LED external quantum efficiency as a function of the injection current density.

this simulation.

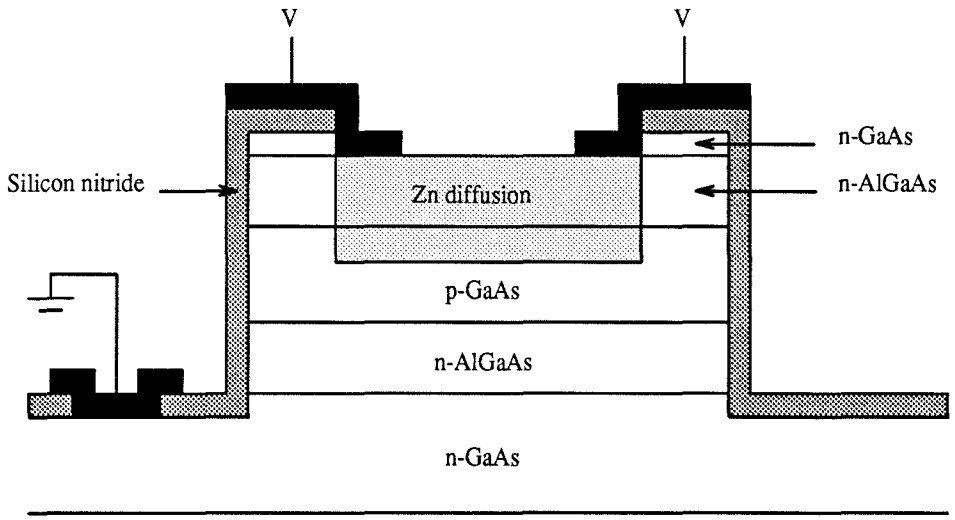
Overall, it is worthwhile to note that the LED external quantum efficiency is on the order of 1 to 2 percent W/A. This is a very low number due primarily to the fact that most of the photons generated are trapped inside the LED because of the refractive index of GaAs being much higher than that of the air. Even with proper anti-reflection coating, the best that efficiency can be obtained is no more than 0.02 W/A. Therefore, LED may not be the solution to all problems. However, the lack of threshold current may make LED a much preferred light emitter as compared to laser diode in certain applications.

2.7 Experimental Results

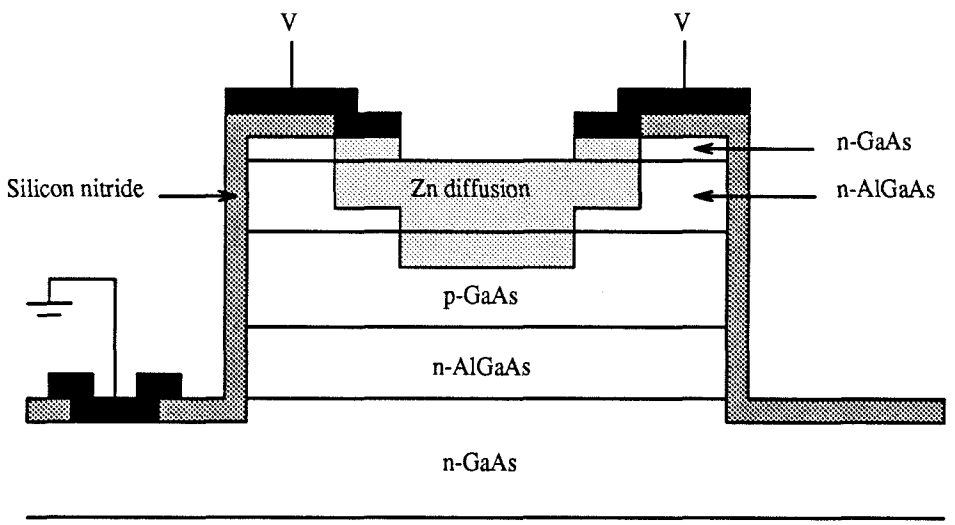
Based on the previous discussion on the lateral current confinement in the LED's, we present experimental results in this section on the external quantum efficiency of the LED's with and without double Zn-diffusion. As discussed previously, the double Zn-diffusion helps to confine the current flow through the middle region of the device, in which the generated photons can easily be transmitted into the air. The structure of the LED with double Zn-diffusion and that of the conventional LED are shown in Fig. 23. The fabrication process for these LED's is relatively simple. A blank deposition of Si_3N_4 on a N-p-N double heterostructure was performed by using a chemical vapor deposition system, in which silane and ammonia gases were mixed in a 610^o environment enclosed by a belljar. The window areas of the LED's were then defined by etching away the Si_3N_4 in a CF_4 plasma. This was followed by loading the LED samples with $ZnAs_2$ in an ampoule, which was sealed at a pressure of 5×10^{-8} torr. Putting the sealed ampoule into a 640^o furnace for 10 minutes allowed the Zn to diffuse down to the active layer of the LED's so that the structure shown in Fig. 2.23(a) was obtained. A second diffusion of shallower

depth was performed on one of the samples by using the same procedure except the diffusion window was larger. Its final diffusion profile is shown in Fig. 2.23(b). Finally, proper metalizations were applied to the cathode and the anode of the LED's for terminal contacts.

These two types of LED's were electrically probed and the emitted optical power was measured by a detector positioned at approximately 1 cm above the LED. A current of 10 mA was applied to the LED. Since there was a gap of 1 cm between the LED and the detector, the measured power would not reflect the real absolute optical power emitted from the LED. However, the improvement in external quantum efficiency of the LED could be inferred by comparing the relative powers measured by the detector from these types of LED's. Table 2.1 summarizes the results for 10 pairs of LED. As we can see, an average of 55.5 % in the improvement of the external quantum efficiency was observed in the LED's with double Zn diffusion. Thus, it was concluded that by properly designing the path of the current flow, the LED external quantum efficiency could be greatly improved.



(a)



(b)

Fig. 2.23 **(a)** The cross section of a LED with single Zn diffusion. **(b)** The cross section of a LED with double Zn Diffusions.

Pair number	Single diffusion	η_{LED}	Double diffusion	η_{LED}	Improvement
1	71μW	0.0071 W/A	105μW	0.0105 W/A	48%
2	68μW	0.0068 W/A	108μW	0.0108 W/A	59%
3	75μW	0.0075 W/A	115μW	0.0115 W/A	53%
4	60μW	0.006 W/A	95μW	0.0095 W/A	58%
5	66μW	0.0066 W/A	103μW	0.0103 W/A	56%
6	74μW	0.0074 W/A	119μW	0.0119 W/A	60%
7	69μW	0.0069 W/A	109μW	0.0109 W/A	58%
8	63μW	0.0063 W/A	99μW	0.099 W/A	57%
9	68μW	0.0068 W/A	105μW	0.0105 W/A	54%
10	73μW	0.0073 W/A	111μW	0.0111 W/A	52%
Average					55.5%

Table 2.1 Comparison showing the optical power emitted from two different LED's at 10 mA, one with single Zn diffusion and the other with a double Zn diffusion. There is a gap of 1 cm between the emitting window of the LED and the detector. Thus, the measured optical power does not reflect the total optical power emitted from the LED.

Chapter 3

Double-Heterojunction Bipolar Transistors

3.1 Introduction

For the implementation of optoelectronic neurons which incorporate heterojunction bipolar transistors (HBT's), it is extremely important that the current gain of the transistors be as high as possible, as we shall see in the next chapter. With the advent of molecular beam epitaxy (MBE) and metalorganic chemical vapor deposition (MOCVD), many promised advantages [32] of heterojunction bipolar transistors, including enhanced current gain [33,34], β, and high speed [35-37], have been gradually realized. While these improvements are impressive, there remain some problems that cause these HBT's not to perform as well as the theory predicts. One of the disappointments is the current gain, β, of the HBT's. According to the theory of HBT, the emitter-base heterojunction provides a hole barrier so that the emitter injection efficiency is no longer affected by this unwanted hole injection. This results in the current gain of a HBT being determined only by its base transport factor, which depends on the base layer thickness and the minority carrier diffusion length in the base. As long as the base layer thickness is kept very thin, a current gain of at least 1000 should be readily obtainable. However, typical published results on current gains of HBT's have revealed values that are on the order of 10's or 100's [38-42]. Among the possible reasons for the degradation of the current gains are poor quality in the epitaxial layers and their respective metallurgical junctions with the neighboring material, leakage currents through carrier recombinations in the depletion region, in the bulk region and on the surface, and

out-diffusion of base dopants, which ultimately destroys the integrity of emitter-base heterojunction. While all of these are devastating causes for the low values of β, it is important to identify the effects of leakage currents and hopefully to clarify some confusing issues relating to the carrier transport across the base-emitter junction in a HBT.

Another practical issue of obtaining a high-gain HBT is the reliability and the ease with which the electrical contacts can be made to the base of the HBT. Current fabrication technology allows three different approaches. The first approach, also the most commonly used approach, is by wet chemical etching to define two mesas and making the base contact to the second mesa, where the base layer is exposed [43-45]. The second approach is by ion implantation, in which a certain dosage of ion is implanted to reach the base of the HBT [46,47]. The last approach is by diffusion [48-51]. The wet chemical etching technique is popular because of its ease and time-saving feature. However, if the current gain of the HBT is to be increased further, the base layer thickness has to be reduced. This presents a serious problem of etching down to the base layer and properly exposing it for electrical contact. This problem can be circumvented by use of selective etchants which will only etch either GaAs or AlGaAs, but not both [52-54]. Such etchants exist. But, their etch rates and selectivities critically depend on the pH values of the etchants and the temperature. Moreover, wet chemical etching destroys the planarity of the device, which makes subsequent processing more difficult. Therefore, a more reliable and repeatable method would be ion implantation or diffusion. These two processes involve either physically damaging the lattice structure of the host material, and thereby creating excess defect centers or subjecting the material to high temperatures for certain duration. Both of these processes have the effect of lowering the minority carrier lifetime, which, in turn, degrades the current gain of the HBT [47,55]. However, if the mechanisms of degradation can be understood

and controlled, these two methods will present themselves as the preferred process in making the base contact to the HBT. Thus, Zn-diffusion is chosen in this investigation to study this degradation effect. The results of which will be explained and quantified.

This investigation begins with a background description, in which detailed description on the sources and paths of the leakage currents in a HBT and on the issue of why diffusion is preferred over wet chemical etching are given. This is followed by the device fabrication and testing procedures, and a discussion and an interpretation on the results obtained. A model is proposed to explain the data.

3.2 Physical Modeling

A high-gain transistor requires maximization of two transistor parameters : emitter injection efficiency and base transport factor. Emitter injection efficiency is defined as the ratio of the injected current from the emitter into the base to the total emitter current, and the base transport factor is defined as the ratio of the current collected by the collector to the injected emitter current into the base. Since the total emitter current is composed of the injected currents from the emitter into the base and likewise from the base into the emitter, the emitter injection efficiency is a number between 0 and 1. Similarly, because part of the injected emitter current is lost through the recombination with the base current, the base transport factor also ranges between 0 and 1. For a high-gain transistor, its emitter injection efficiency and base transport factor are very close to 1. The product of the two numbers represents the percentage of the total emitter current that makes it to the collector, or the more well-known common base current gain, α:

$$\alpha = \gamma_e \cdot \gamma_b \tag{3.1}$$

$$\gamma_e = emitter\ \ injection\ \ efficiency$$

$$= \frac{1}{1 + \frac{N_b D_e W_b}{N_e D_b L_e}} \tag{3.2}$$

$$\gamma_b = base\ \ transport\ \ factor$$

$$= 1 - \frac{1}{2}\left(\frac{W_b}{L_b}\right)^2, \tag{3.3}$$

where γ_e, γ_b, N_b, N_e, D_b, D_e, W_b, L_b, and L_e represent the following parameters :

γ_e : *emitter injection efficiency*,

γ_b : *base transport factor*,

N_b : *doping concentration in the base*,

N_e : *doping concentration in the emitter*,

D_b : *minority carrier diffusion coefficient in the base*,

D_e : *minority carrier diffusion coefficient in the emitter*,

W_b : *base layer thickness*,

L_b : *minority carrier diffusion length in the base*,

and L_e : *minority carrier diffusion length in the emitter.*

Eq. (3.2) and (3.3) are valid only for $W_b << L_b$, which usually holds for high-gain transistors because the base of these transistors is very thin. The current gain of a transistor, β, is defined by :

$$\beta = \frac{I_c}{I_b}$$

$$= \frac{\alpha}{1-\alpha}. \tag{3.4}$$

For a N-p-N AlGaAs/GaAs/AlGaAs double heterojunction bipolar transistor (DHBT), the discontinuity in the valence band, ΔE_v, helps suppress the hole injection from the base into the emitter. This results in an improvement in the emitter injection efficiency, γ_e, as the injected current from the emitter into the base represents a larger portion of the total emitter current. Taking into consideration this effect, Eq. (3.2) is modified by :

$$\gamma_e = \frac{1}{1 + \frac{N_b D_e W_b}{N_e D_b L_e} \cdot exp(\frac{-\Delta E_v}{kT})}. \tag{3.5}$$

Quantitatively, for a 30% Al mole fraction system, the bandgap difference between GaAs and AlGaAs is approximately 0.38 eV [56]. Assuming 15% of this bandgap difference appears across the valence band [57], and the following numbers : $N_b = 2 \times 10^{17}$ cm^{-3}, $N_e = 4 \times 10^{17}$ cm^{-3}, $D_b = 100$ cm^2/sec, $D_e = 10$ cm^2/sec, $W_b = 0.15\mu$m, $L_e = 3\mu$m, the resulting γ_e is 0.9997256, which, for practical purpose, is usually approximated by 1. Therefore, for DHBT's, the current gain, β, is dominated by the base transport factor. Unfortunately, the expression for the base transport factor as given by Eq. (3.3) represents an idealized situation in which the recombination of the injected electron current from the emitter with the hole current from the base takes place only in the quasi-neutral region of the base. In actuality, the situation is complicated by the addition of leakage currents, which comprise the majority of the base current, This has the effect of increasing the base current and at the same time reducing the collector current. Since β is I_c/I_b, the detriment to β is two-fold. This can be understood qualitatively by referring to Fig. 3.1. I_1 is the recombination of the injected electron current with the hole current in

the quasi-neutral region of the base (or intrinsic base region). I_2 is the recombination of the same two currents in the extrinsic base region due to the lateral diffusion of electrons. I_3 is the recombination current in the emitter-base depletion region. I_4 is the recombination current in the bulk emitter region. I_5 is the recombination current that takes place on the exposed surface and I_6 is the base-collector reverse leakage current. In term of these currents, the current gain, β can be written as :

$$\beta = \frac{I_c}{I_b} = \frac{I_e - I_1 - I_2 - I_3 - I_4 - I_5 + I_6}{I_1 + I_2 + I_3 + I_4 + I_5 + I_6}. \tag{3.6}$$

This is in contrast to the β that would have been without the leakage currents :

$$(\beta)_{ideal} = \frac{I_c}{I_b} = \frac{I_e - I_1}{I_1}. \tag{3.7}$$

Clearly, the β in Eq. (3.6) will be less than that in Eq. (3.7). In fact, it will be less by at least an order of magnitude because these leakage currents dominate the actions of the transistor, especially at low bias conditions [58]. Thus, it is crucial to understand the role of each leakage current.

I_1 is the recombination current in the quasi-neutral region of the base. This is the base current that is mentioned in most semiconductor textbooks. Its magnitude is determined in part by the minority (or electrons in the case of N-p-N transistors) carrier lifetime, which, in turn, is a characteristic parameter of the material. Electrons and holes can recombine either directly or indirectly through traps inside the bandgap. Regardless which recombination process dominates, the overall minority carrier lifetime under low-level injection conditions is fixed once the epitaxial layers have been grown. For GaAs, direct recombination is the dominant process. Thus,

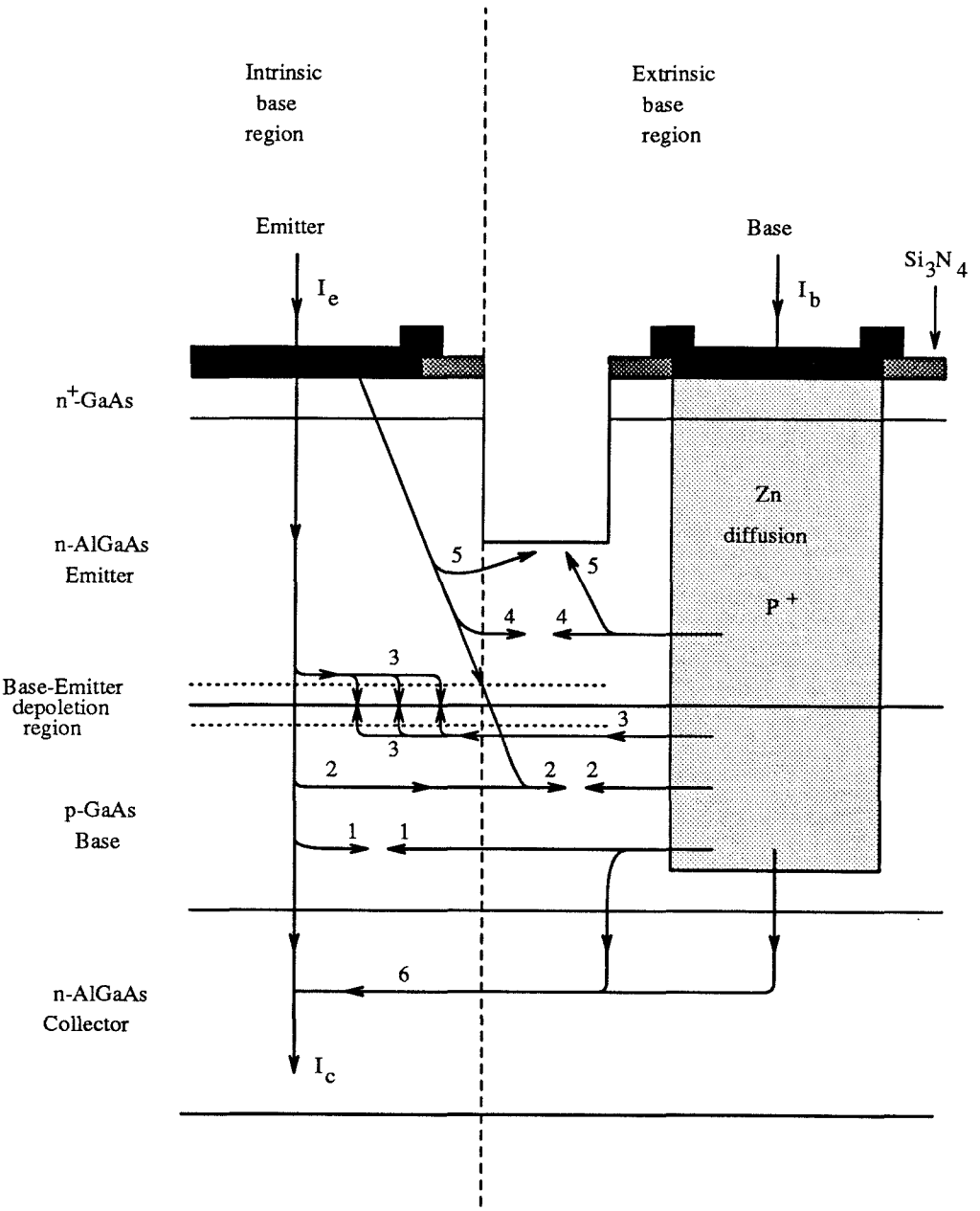


Fig. 3.1 Paths of possible leakage currents across the base-emitter junction in a DHBT.

if one is to increase the minority carrier lifetime in hope for an improvement in β, one should reduce the base doping concentration because there will be fewer free carriers for recombination [59].

I_2 is similar to I_1 in nature except it takes place in the extrinsic region of the base. It is comprised of emitted electrons from the emitter, which laterally spread into the neighboring regions. This means that before the holes can flow to the intrinsic region of the base to make the base-emitter junction more forward biased, they get annihilated by the electrons that laterally diffuse from their main path. This effect has been systematically studied by Nakajima et al. [60]. Their results indicate that the electron lateral diffusion becomes worse as the ratio of the emitter perimeter to emitter area increases. Thus, to reduce this leakage current, a small ratio of emitter perimeter to area is required. For a given area, a circle has the smallest perimeter. Thus, a circular emitter would be most effective in eliminating I_2. Moreover, the bigger the circle, the more I_2 is reduced relative to the other components of the base current, thus the bigger the β. This is in agreement with the finding of Hiraoka et al. [55]. Intuitively, this means that as the emitter gets bigger, the part of the current that diffuses laterally represents a smaller portion of the total emitter current, which, in turn, implies that I_2 has become a less significant leakage current. However, the actual size of the circle is limited by the base-emitter capacitance. Another effective way of reducing I_2 is reducing the thickness of the base layer. This will reduce the extent of lateral diffusion by the injected electrons. Hiraoka et al. [55] also suggest that a graded-bandgap base will reduce I_2 because the build-in electric field due to the graded bandgap will swiftly sweep the injected electrons across the base, and thus giving these electrons a less chance for lateral diffusion.

I_3 represents the recombination current in the base-emitter depletion region. In this depletion region, carriers recombine through deep-level traps, which might

be created by imperfection of heterojunction interface or interfacial stress. It is shown that carriers recombining in this fashion do not obey the ideal diode law. In fact, they obey [61]

$$I = I_o exp(\frac{qV}{2kT}). \tag{3.8}$$

Thus, by measuring the I-V characteristics of the base-emitter junction, one can easily determine the approximate magnitude of this leakage current relative to the other part of the currents that obey the ideal diode law.

This is a major source of leakage current. Its magnitude for homojunction transistors is given as follows [62] :

$$I_3 = \frac{qn_i x_d}{2\tau_0} \cdot exp(\frac{qV_{be}}{2kT}), \tag{3.9}$$

where n_i, x_d, and τ_0 are the intrinsic carrier concentration, the depletion layer thickness of the base-emitter junction, and the minority carrier lifetime in the depletion region, respectively. However, for a heterojunction bipolar transistor, it is modified by :

$$I_3 = \frac{q}{2}\left(\frac{n_{ie}x_e}{\tau_{0e}} + \frac{n_{ib}x_b}{\tau_{0b}}\right) \cdot exp(\frac{qV_{be}}{2kT}), \tag{3.10}$$

where n_{ie}, n_{ib}, x_e, x_b, τ_{0e}, and τ_{0b} denote the following parameters :

n_{ie} : *intrinsic carrier concentration in the emitter,*

n_{ib} : *intrinsic carrier concentration in the base,*

x_e : *depletion layer thickness in the emitter,*

x_b : *depletion layer thickness in the base,*

τ_{0e} : *minority carrier lifetime in the emitter side of the depletion region,*

and τ_{0b} : *minority carrier lifetime in the base side of the depletion region.*

Substituting Eq. (3.10) into Eq. (3.6) and neglecting all other leakage currents, $I_2, I_4, I_5, and\ I_6$, yields :

$$\frac{1}{\beta} = \frac{1}{2\left(\frac{L_b}{W_b}\right)^2 - 1} + \frac{N_b W_b \left(\frac{n_{ie} x_e}{\tau_{0e}} + \frac{n_{ib} x_b}{\tau_{0b}}\right)}{2D_b (n_{ib})^2} \cdot exp(-\frac{qV_{be}}{2kT}). \qquad (3.11)$$

Assuming $L_b = 6\mu$m, $n_{ib} = 1.8 \times 10^6$ cm^{-3}, $n_{ie} = 10^3$ cm^{-3}, $V_{be} = 0.9$ V, $x_b = 0.02\mu$m, $x_e = 0.01\mu$m, $\tau_{0e} = \tau_{0b} = 5 \times 10^{-8}$ sec, and the previously assumed values for the other parameters, one finds that the β of such a transistor equals to 110. Comparing this with the β of 3200 obtained by neglecting I_3, we can see that the effect I_3 is dramatic. Again, there is very little that can be done to eliminate I_3 once the emitter-base junction is grown. Thus, one should be very careful in growing this junction so that interfacial defects and stress are minimized.

I_4 is the leakage current that flows laterally through the AlGaAs P-N homojunction. It has been argued [63] that electrons in the emitter will be restricted from injecting into the AlGaAs portion of the Zn diffusion region because this is high bandgap P-N homojunction. Thus, electrons injection will preferentially take

place at the vertical base-emitter heterojunction, but not at the lateral base-emitter homojunction. However, if one looks at this leakage current from the standpoint of hole injection from the base into the emitter, the picture might be slightly different. There are two possible paths which holes from the base might take to inject into the emitter. The first path is the lateral injection across the AlGaAs P-N homojunction. The second path is for holes to traverse down to the extrinsic region of the base, then traverse horizontally into the intrinsic region of the base, followed by injection into the emitter layer above it through the GaAs-AlGaAs p-N heterojunction. These two paths are depicted in Fig. 3.2(a) and the associated energy band diagrams for these two paths are also shown in Fig. 3.2(b).

For path 1, the energy barrier that holes need to overcome in order to inject into the emitter is E_1. For path 2, holes need to first overcome a small barrier, E_2, in order to reach the extrinsic region of the base layer. Once holes are in the extrinsic part of the base layer, most of them lose energy through scattering and eventually relax down to the energy level of the valence band in the GaAs. In order to inject into the AlGaAs emitter through the heterojunction, holes need to overcome a barrier height of E_3. Thus, a total energy $E_2 + E_3$ is needed before holes can inject into the emitter through path 2. The relative magnitudes of E_1 and E_3 are comparable and are shown as follow :

$$\begin{aligned} E_1 &= (E_g)_{AlGaAs} - \big((E_c)_{AlGaAs} - E_f\big) - \big(E_f - (E_v)_{AlGaAs}\big) \\ &= (E_g)_{AlGaAs} - kT \cdot ln\frac{(N_c)_{AlGaAs}}{(N_d)_{emitter}} - kT \cdot ln\frac{(N_v)_{AlGaAs}}{(N_a)_{Zn-diff}}, \end{aligned} \tag{3.12}$$

$$\begin{aligned} E_3 &= (E_g)_{AlGaAs} - \big((E_c)_{AlGaAs} - E_f\big) - \big(E_f - (E_v)_{GaAs}\big) \\ &= (E_g)_{AlGaAs} - kT \cdot ln\frac{(N_c)_{AlGaAs}}{(N_d)_{emitter}} - kT \cdot ln\frac{(N_v)_{GaAs}}{(N_a)_{base}}, \end{aligned} \tag{3.13}$$

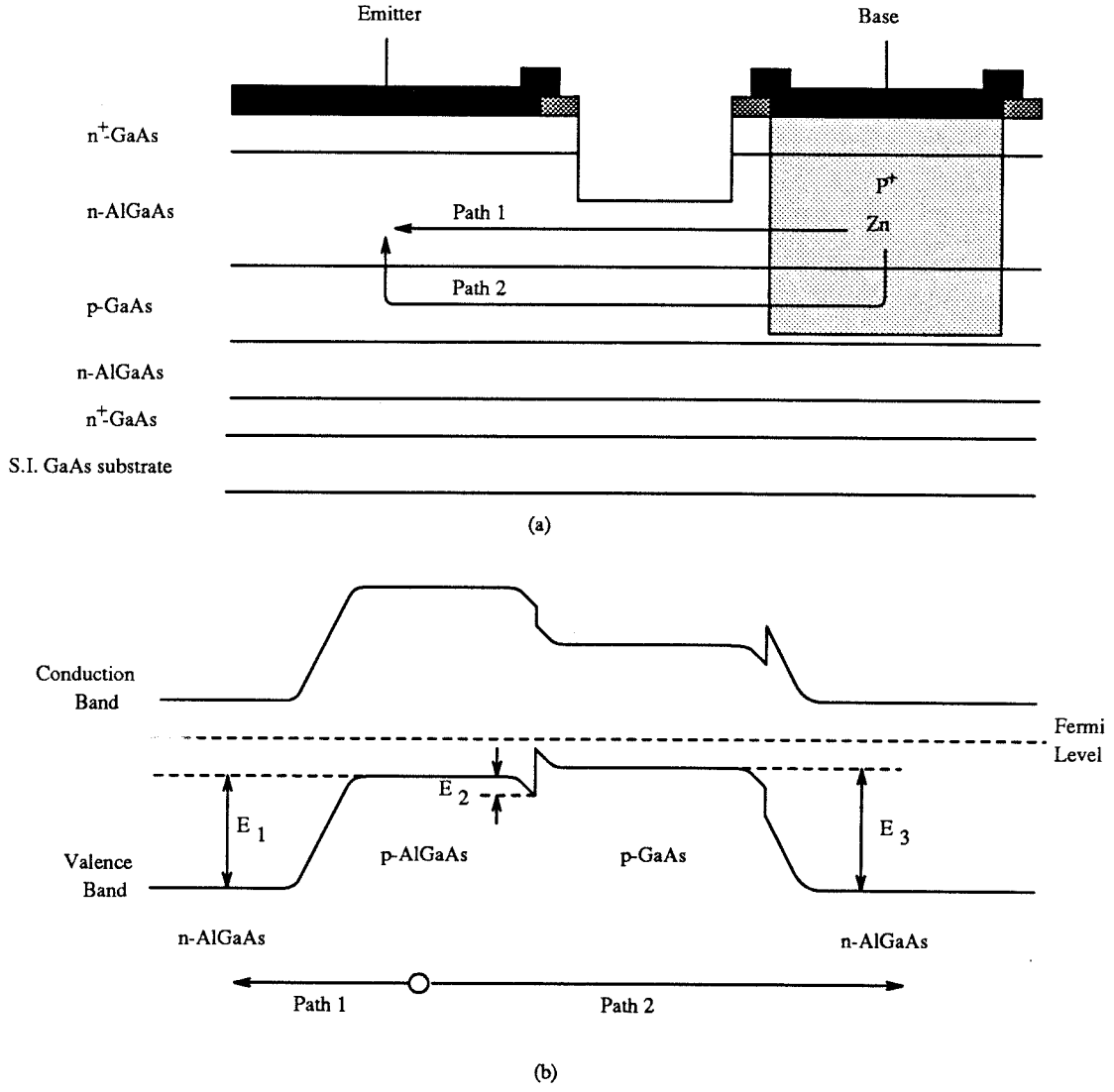


Fig. 3.2 **(a)** Two possible paths for holes to inject into the AlGaAs emitter region. One is to inject laterally through the AlGaAs p-n homojunction and the other is for holes to traverse down to the extrinsic region of the base and then traverse horizontally into the intrinsic region of the base and finally inject into the AlGaAs emitter through the heterojunction. **(b)** Energy band diagrams for the two different paths.

where N_c, N_v, N_d, and N_a are the effective density of states for the conduction band and valence band, and doping concentrations for the donors and the acceptors respectively. E_g, E_c, E_v, and E_f stand for bandgap, conduction band, valence band, and the fermi level respectively. A close examination reveals that E_1 and E_3 differ only by the last term in Eq. (3.12) and (3.13). As a matter of fact, the last terms in Eq. (3.12) and (3.13) are approximately equal for most practical doping situations. As a result, holes traversing through path 2 need an additional energy of E_2 than traversing through path 1. This translates into a relative magnitude of 1 to $\exp(-E_2/kT)$ for holes injection into the emitter through path 1 and path 2. Therefore, depending on the value of E_2, the leakage current due to I_4 may not be negligible.

I_5 is the recombination current that occurs on the exposed surface. This is another important leakage current. Because of the exposed surface, surface states are formed, which, in turn, attract electrons and holes to the surface and promote recombination. Because the concentration of these surface states is enormous (usually around 10^{12} cm^{-2}), the fermi level is pinned to the energy level of these states. Since these states typically occupy levels close to midband, carriers recombining in this fashion also show "2kT" dependence in their I-V characteristics, as given by Eq. (3.8). Tiwari et al. [58] have shown two dimensional potential diagrams for both electrons and holes of a double-mesa transistor having an exposed base. In their finding, the maximum recombination rate occurs on the exposed part of the base due the formation of a potential well for both electrons and holes there. Along the exposed surface, they further find that the majority of the carriers that recombine on the surface recombine around the corner near the emitter. Passivation of the exposed surface is an effective way in getting rid of these surface states. Si_3N_4 and SiO_2 are candidates for passivation materials. Si_3N_4 is particularly suited for DHBT's with a diffused base because Si_3N_4 can also be used as the diffusion mask

for Zn. However, because of the large thermal mismatch between Si_3N_4 and GaAs [64], which leads to large interfacial stress and creates more defects if not processed carefully, transistors passivated with Si_3N_4 do not show as good an improvement in β as expected [65]. It has been demonstrated that there is a much more promising and reliable way of passivating the exposed surface [66]. The idea is to leave a very thin layer of AlGaAs emitter layer on top of the GaAs base layer so that this thin layer of AlGaAs is totally depleted due to the surface states on the exposed surface. This has the effect of not only blocking the lateral leakage, I_4, but also completely suppressing the recombination current on the the surface, I_5. By using this technique, HBT's with β of 12500 have been demonstrated [66]. Comparison of the effectiveness of Si_3N_4 and depleted AlGaAs as passivation materials has been studied [65]. The result indicates that transistors passivated by thin layers of depleted AlGaAs show β's that are 10 times higher than transistors passivated by Si_3N_4. Thus, the concept of using a thin layer of depleted AlGaAs as a passivation material has been adopted for double-mesa type of transistors, which usually have large areas of exposed base. For transistors with diffused base or implanted base, this concept has not been incorporated because the general belief is that the base is protected by the AlGaAs layer above it. Also, the potential barrier across the P-N homojunction in the AlGaAs layer will prevent any bulk leakage current. Thus, the typical process that prevents leakage current flowing between the emitter and the base in transistors with diffused or implanted base is simply a wet chemical etch to remove the top n^+ cap layer, which might otherwise shunt the emitter-base junction. As will be shown in the current investigation, an etch of just the n^+ cap layer is not enough. One needs to etch all the way down to leave only a very thin layer of depleted AlGaAs on top of the GaAs base in order to completely eliminate these leakage currents, just like the case for double-mesa type of transistors.

I_6 is the base-collector reverse leakage current. This leakage current is again

fixed once the base and collector have been grown. Its I-V characteristics follow that of a ideal diode. In practical transistors, the magnitude of I_6 is relatively small compared to the other leakage currents, and thus it is usually neglected.

Another issue that needs to be addressed is the technology of making a DHBT. As mentioned previously, there are three different technologies in making the base contact to the transistors. One is by wet chemical etching, and the other two are by ion implantation and diffusion. For the benefit of understanding the advantages and the disadvantages of each method, a brief discussion on the actual process and its effect on transistor performance is given. Ion implantation will not be discussed here because its effect on the transistors is similar to that of diffusion.

Wet chemical etching down to the base provides an easy and simple way to make contacts to the base. This method typically results in a transistor geometry that has two mesas, with the top mesa providing the emitter contact and the second mesa providing the base contact. Although its popularity is supported by the simplicity, it has its own drawbacks. Firstly, a double-mesa type of transistor has different junction areas for the emitter-base and the collector-base junctions. This leads to different electrical injection characteristics at the two junctions. It has been shown that these differences are responsible for the emitter-collector offset voltages observed when operating the transistors in the common-emitter mode [67]. This phenomenon gets worse for high-frequency transistors because they usually have very small base-emitter junctions in order to minimize the base-emitter junction capacitance. The consequence of having an offset collector-emitter voltage is that extra power consumption by the transistor is required to achieve the same performance. In addition, the useful operating range of the transistors is reduced by the same offset voltage. As a result, the performance of the DHBT is severely limited. The second drawback is the destruction of the planarity of the transistors. Non-planar geometry causes non-uniform photoresist coating upon spinning. Thus,

exposure and development of this photoresist will not be uniform. Consequently, any subsequent processing steps, whether it is etching or deposition, will be affected by this non-uniformity and sometimes turn out to be detrimental for the device. This is a difficult problem to remedy. Thus, the best way is to avoid non-planar design altogether. The third drawback is the issue of passivation required on the large exposed second mesa, which is the base. This was already discussed previously. As it turns out, all transistors having any exposed surface suffer from this effect. The last drawback, and the most important one, is the difficulty involved in stopping the etch once the etch front reaches inside the base. In carrying out this difficult task accurately, one needs to know first the thickness of each epitaxial layer accurately, which can be obtained from the examination of a SEM (scanning electron microscope) picture on the cross section of the sample. Then, by using a very well controlled etching system, one can obtain the correct etching depth and expose the base properly. However, as the thickness of the base layer gets thinner, proper exposure of the base becomes a challenge in the fabrication process due to instability of the etchants used and the resolution and the accuracy of the etch depth measurement system. Thus, one has to resort to the use of selective etching in properly exposing such a thin layer of base. Selective etchants will etch only either GaAs or AlGaAs, but not both. This assures the complete removal of the emitter AlGaAs layer and yet leaves the GaAs base layer totally intact. A lot of chemical solutions have been proposed as selective etchants [52-54]. Typically, an oxidizing agent is used to oxidize the material first, followed by dissolving the oxidation products. The oxidizing strength depends on mixture ratio of the oxidizing agents to the reducing agents. The dissolution rate of the oxidation products depend on the pH value of the solution, which, is controlled by a buffer solution [68]. Both processes of oxidation and dissolution are a sensitive function of temperature. Moreover, the selectivity of an etchant critically depends on the relative mixture

ratio of the oxidizing agents to the reducing agent and the pH value of the buffer solution. For instance, the same etchant intended to etch AlGaAs will etch GaAs if its pH value in the buffer solution is slightly off. Similarly, for certain selective etchants, their selectivities will reverse totally if the mixture ratio of the oxidizing agents to the reducing agents is off. Thus, care has to be exercised in preparing these selective etchants.

While wet chemical etching presents itself as a simple, and yet less controllable method in making contact to the base, diffusion is a more involved and controllable alternative. Moreover, the process of using diffusion to make contacts to the base of the transistor can be simultaneously carried out while performing the diffusion process to convert the N-p-N structure for the transistor to the P-p-N structure for the LED. This saves an extra processing step. This issue is discussed in Ch. 4. In making a diffusion to contact the base (Zn is typically used as the diffusant), one starts by depositing a layer of dielectric material to be used as the diffusion mask. Among the candidates for this dielectric material are SiO_2, PSG (phosphosilicate glass) and Si_3N_4. Since Zn-diffusion is at an elevated temperature, the diffusion mask used has to be able to inhibit the decomposition of GaAs, which melts at high temperatures. Of these three masks, SiO_2 has the worst properties because SiO_2 permits the rapid diffusion of Ga through it at high temperatures. Thus, decomposition of GaAs is not prevented. SiO_2 is also very transparent to Zn so that masked diffusion is very hard to accomplish. Finally, SiO_2 is very poorly matched to GaAs in their thermal expansion coefficients. Thus, a thick film often cracks during the diffusion process. Thermal expansion problems can be greatly reduced by the use of PSG [69]. Moreover, its effectiveness in blocking Zn is good. However, its blocking characteristics for Ga is in doubt. On the other hand, Si_3N_4 offers excellent blocking characteristics for both Zn and Ga. In fact, it is almost impermeable to both of them. However, it has a bad thermal mismatch with GaAs [70]. Thus,

a thin film of Si_3N_4 has to be used to prevent cracking during diffusion. A side effect associated with masked diffusion is the enhanced lateral diffusion observed in the samples that show large interfacial stress with the mask above. This interfacial stress is due the bad thermal mismatch between the mask and GaAs. Thus, this is another reason why a thin layer of Si_3N_4 ($\approx 1000\mathring{A}$) should be used.

After depositing the diffusion mask and the subsequent patterning, the sample is loaded into an ampoule containing the diffusants, usually $ZnAs_2$, and then sealed under vacuum for diffusion. During the high temperature diffusion process, the minority carrier lifetime is affected. As a result, the current gain of the transistors made in this fashion usually will be affected as well. Part of this study is to find out how different diffusion conditions affect the current gain of the transistors. The other part of this study deals with the effects leakage currents have on the current gains of the DHBT's.

3.3 Device Fabrication

The double heterostructure was epitaxially grown on (100) Cr-doped semi-insulating GaAs substrates ($\rho \geq 5\text{x}10^7$ ohm·cm) by a metalorganic chemical vapor deposition system (SPIRE-450) using a vertical barrel reactor. The GaAs and AlGaAs layers were grown by trimethyl gallium (TMG), trimethyl aluminum (TMA), and 10% AsH_3 in 90% H_2. Zinc and silicon were used for p- and n-type dopants, respectively. The substrate temperature during the growth was about 730^o C. The double heterostructure consists of : 0.5 μm of Si-doped (10^{18} cm^{-3}) n-GaAs subcollector/buffer, 1.2 μm of Si-doped ($1.6\text{x}10^{17}$ cm^{-3}) n-$Al_{0.3}Ga_{0.7}As$ collector, $100\mathring{A}$ of undoped GaAs spacer layer, 0.15 μm of Zn-doped ($2\text{x}10^{17}$ cm^{-3}) p-GaAs base, 100 $\mathring{A}$ of GaAs undoped spacer layer, 1.0 μm of Si-doped ($4.2\text{x}10^{17}$ cm^{-3}) n-$Al_{0.3}Ga_{0.7}As$ emitter, and 0.23 μm of Si-doped ($1.4\text{x}10^{18}$ cm^{-3}) n-GaAs cap layer.

The doping profile and thickness of each DHBT layer were plotted by a Polaron PN-4200 electrochemical profiler as shown in Fig. 3.3. It should be noted that the doping concentration of each layer is extremely uniform.

Four samples were cleaved from the same part of the wafer, followed by the standard post-growth wafer cleaning procedures. Each transistor was first photolithographically defined by etching the epilayers down to the n^+ GaAs collector layer with nonselective etchants, $H_3PO_4 + H_2O_2 + CH_3COOH$ (1 : 1 : 3). A 10-minute deposition of Si_3N_4 was performed using SiH_4 and NH_3 in room atmosphere at 700° C. Zn-diffusion areas were defined by photolithography in allowing this part of the Si_3N_4 film to be etched away in a CF_4 plasma. Measurements on the film thickness indicated that Si_3N_4 was about 700 $\AA$ to 800 $\AA$ thick, which agreed with the blue color of the film. Prior to being loaded into the ampoules, the 4 samples were slightly etched in $NH_4OH + H_2O_2 + H_2O$ (20 : 7 : 973) solutions to remove any native oxide that might have been present. Zn diffusion at 640°C for 25, 40, 88, and 218 minutes were then performed on these 4 samples in sealed ampoules using $ZnAs_2$ as the source in order to provide an overpressure of As to prevent As from decomposing or evaporating from the GaAs epitaxial layers. After diffusing for the designated time, the ampoules were quenched immediately by water so that As vapor condensed quickly onto the wall inside the ampoule. This was a good indication that Zn-diffusion had taken place. Ampoules were then cut open and samples were cleaned again. Examination under the microscope revealed slight, but very uniform brown to yellowish colors in the diffusion areas. This was another excellent indication that Zn had uniformly diffused. More Si_3N_4 was subsequently etched away for the emitter and collector contacts in a CF_4 plasma. Base contacts were then defined by photolithography, followed by soaking the transistor samples in chlorobenzene for 10 minutes prior to development. A slight etch of the exposed GaAs n^+ cap layer was performed to remove any oxide before the samples were

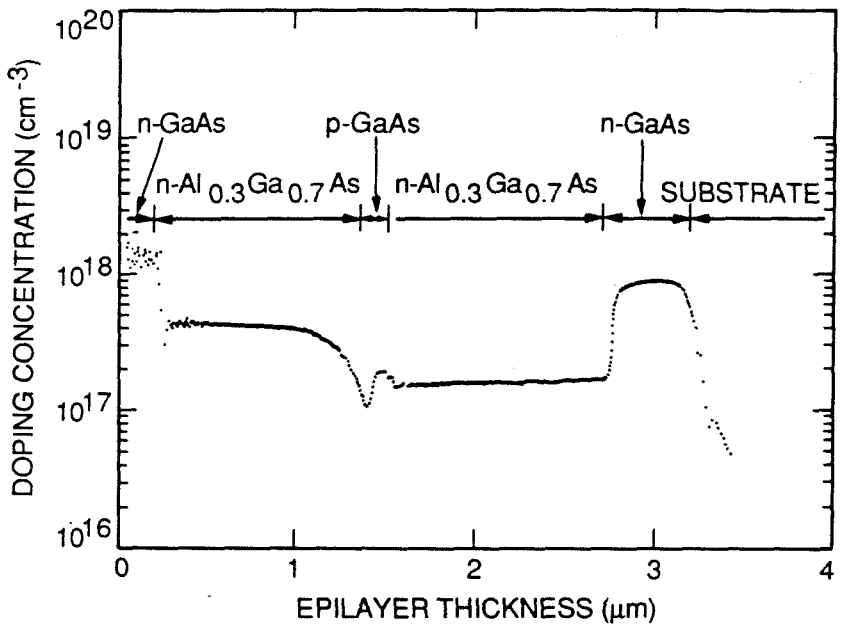


Fig. 3.3 Doping profile and thickness of each DHBT epitaxial layer as plotted by a Polaron PN-4200 electrochemical profiler.

mounted for evaporation. A 300 Å layer of Cr, followed by a 1700 Å layer of Au overlayer were evaporated. Lift-off was used to remove the unwanted metal. When defining the geometry for the base contacts, care was taken to make sure that the Zn-diffused areas were completely covered by the metals so that the base region was not exposed to the air. Similar procedures were repeated for the emitter and collector contacts except for two things. First, an initial 300-Å layer of AuGe was substituted for Cr. Second, when defining the emitter contact areas, a small exposed region of the n^+ cap layer was purposely left uncovered by the emitter metalization so that an isolation etch could be performed in this region to remove any leakage currents that would have been flowing in the n^+ cap layer. While allowing this exposed region to be uncovered by the metalization, the emitter contact area was at the same time made as big as possible so the effect of electron lateral diffusion discussed earlier was minimized. Following the metalization, transistors were tested for their common-emitter characteristics. A rather large turn-on voltage was routinely observed. Thus, contacts were alloyed for 4 minutes in a N_2 ambient and the transistors were tested again. The turn-on voltage disappeared and showed no offset voltage right after alloying. The cross section of a completely fabricated transistor is shown in Fig. 3.4.

After the transistors were alloyed, their common-emitter I-V characteristics were tested and the collector current was recorded at a particular base current. Then, the transistor sample was immersed in an etching solution to etch away part of the exposed n^+ GaAs cap layer. Immediately after the etch, the transistor was put back on the probe station for testing. New common-emitter I-V characteristics were recorded on the same transistor and the new collector current at the same base current on the same transistor was also recorded. After the test, the same process of etching more of the exposed surface followed immediately. The testing was repeated until the transistors' I-V characteristics or current gains did not improve anymore.

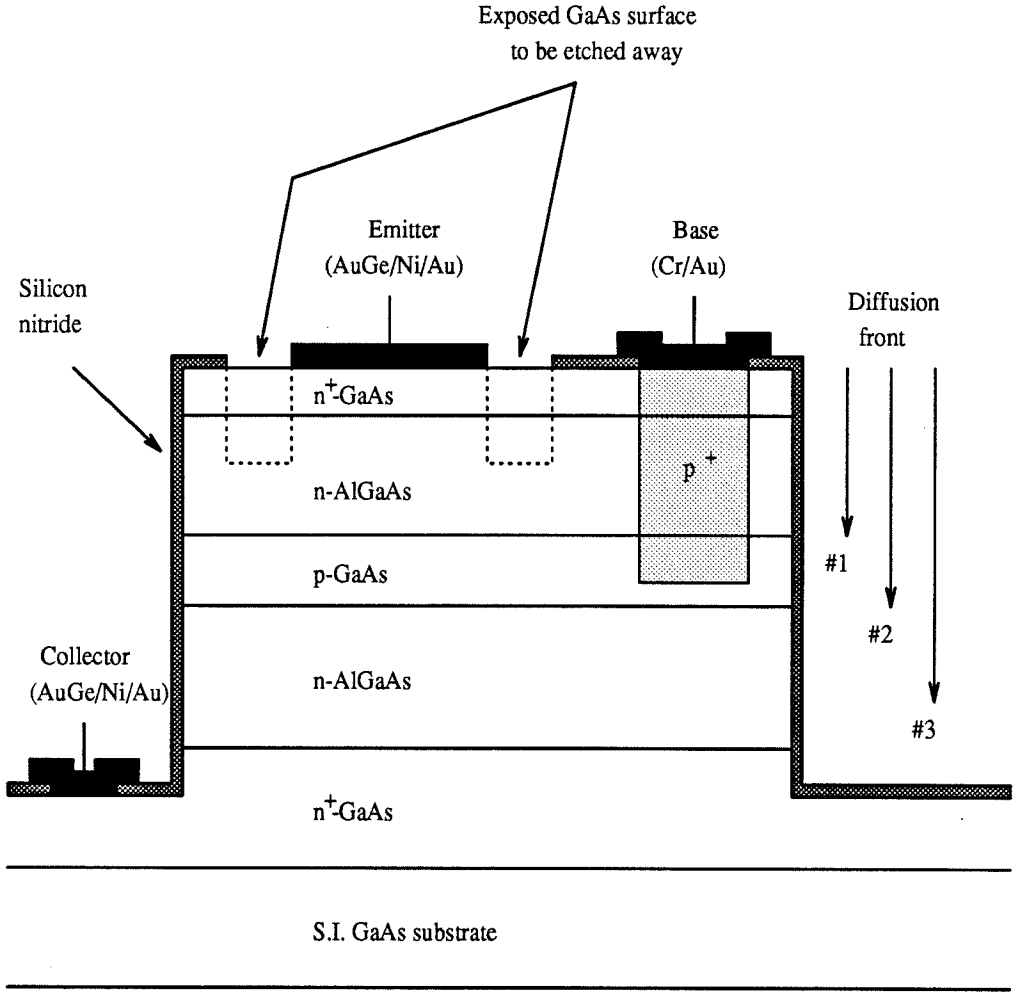


Fig. 3.4 Cross section of a completely fabricated DHBT before isolation etch.

Each time the etch was performed, the etch depth was measured by a surface profiling system. Thus, the correlation of transistor performance and this etch depth could be obtained.

The same process was applied to each one of the 4 transistor samples, except for sample # 3, in which the exposed area of the transistors was continuously etched until the base and the emitter of the transistors were totally disconnected. Data on these measurements are presented in the next section.

3.4 Experimental Results

DHBT's from sample # 1, # 2, and # 3 showed very low current gain before any exposed surface was etched. Typical β's were less than 10, and in some cases they were less than unity. However, as the etch isolation depth increased, β started to increase, but very slowly. As the etch front approached the emitter-base heterointerface, β increased dramatically to a maximum value, and then saturated. At this point, the etch front was only approximately 1000 $\AA$ away from the base-emitter interface. Etching beyond this point did not improve β anymore. In fact, for sample # 3, as the etch front reached inside the base layer, thus exposing the base to air, β started to decrease. These data can be seen in Fig. 3.5 and 3.6. Fig. 3.5 shows the current gain, β, for sample # 2 as a function of the isolation etch depth. The heterointerface between the cap GaAs layer and the AlGaAs emitter layer , and that between the AlGaAs emitter layer and the GaAs base layer are also drawn in for better visualization. The biases were $V_{ce} = 4V$ and $I_b = 0.9mA$. Fig. 3.6 shows the same plot for sample # 3 at $V_{ce} = 4V$ and $I_b = 0.2mA$. It should be noted that $\beta's$ shown on these plots are the DC current gains, not the small signal current gains. There was no corresponding plot for sample # 1 because the ohmic contacts were not stable. Most data measured on the current gains from sample

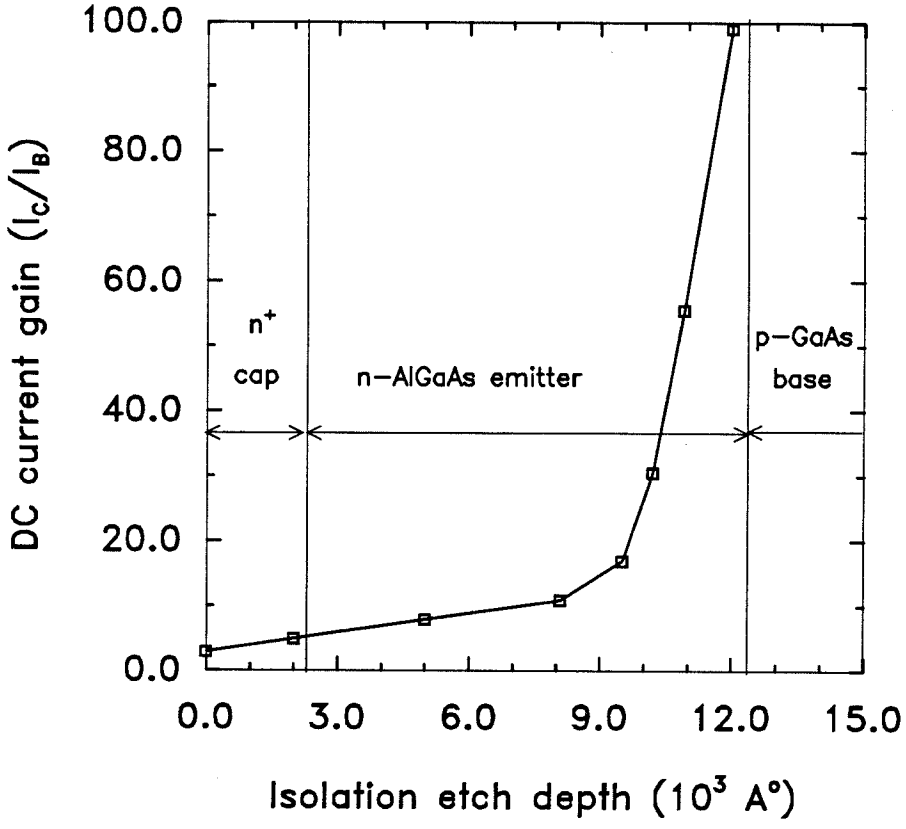


Fig. 3.5 DC current gain as a function of the isolation etch depth for transistors from sample # 2, which has a diffusion time of 40 minutes and a diffusion front that just reaches the base-collector junction.

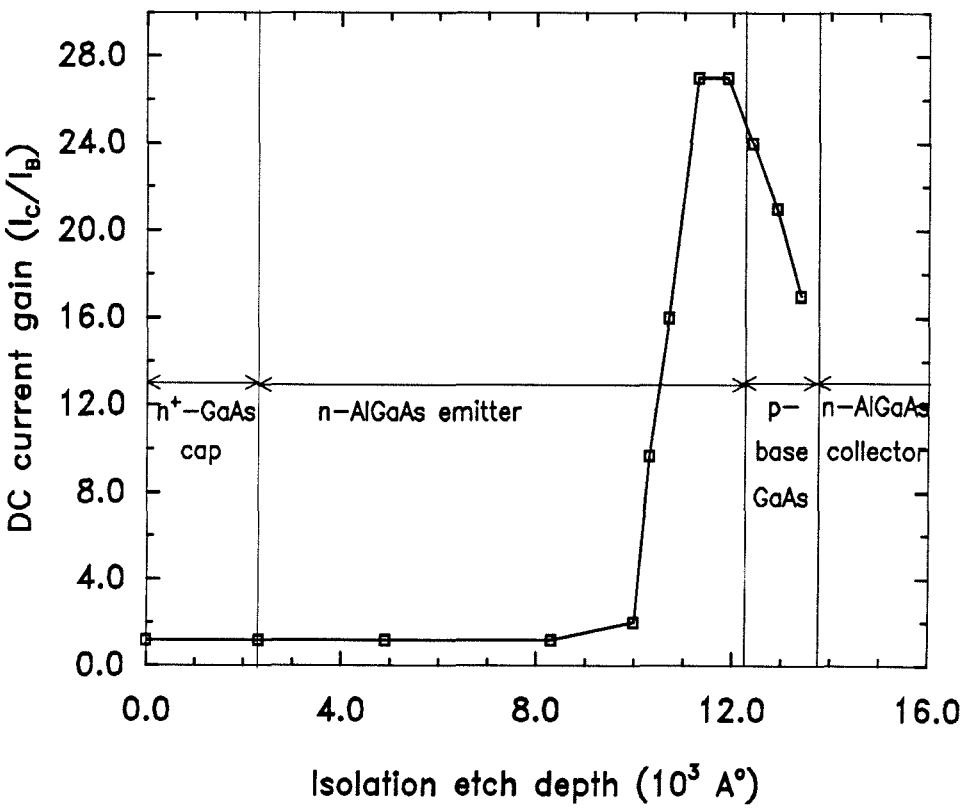


Fig. 3.6 Similar plot with Fig. 3.5 except it is for transistors from sample # 3, which has a diffusion time of 88 minutes and a diffusion front that just reaches inside the collector.

1 were inconsistent . Thus, a plot like Fig. 3.5 or 3.6 could not be produced. However, β vs. I_c plots were produced without any difficulty for samples # 1, # 2, and # 3. They are shown in Fig. 3.7, 3.8, and 3.9 respectively. In Fig. 3.7, β was plotted against I_c for two different isolation etch depths. Curve 1 had an isolation etch depth of 10000 $\AA$, which corresponded to 2300 $\AA$ of AlGaAs emitter left, and curve 2 had an etch depth of 11600 $\AA$, which corresponded to 700 $\AA$ of AlGaAs emitter left. For each curve, β increased with increasing I_c until a maximum was reached. The ideality factors evaluated from the relationship, $\beta \sim I_c^{1-\frac{1}{n}}$, were 2.46 and 1.61 respectively. Because of the previously mentioned problem of unstable contacts in sample # 1, only two curves were obtained. However, the general trend on the transistors of this sample was that β remained relatively unchanged until the etch front reached the promixity of the base-emitter junction. Similar curves for DHBT's from sample # 2 and # 3 were also shown in Fig. 3.8 and 3.9.

Several things should be noted on the behavior of these DHBT's. Firstly, the $\beta's$ on all transistors improved with increasing I_C until reaching a maximum, then degraded slightly afterward. Secondly, the β's improved with deeper isolation etch between the emitter and the base region, especially at low collector current levels. Thirdly, the ideality factor started out close to one before the isolation etch. However, as the isolation etch became deeper, the ideality factor degraded to close to two and then eventually recovered to close to one again. Lastly, the maximum β achieved in each sample degraded with increasing diffusion times. Quantatitively, the maximum β achieved in sample # 1 , # 2 and # 3 were 200, 120, and 60 and the diffusion times were 25, 40 and 88 minutes, respectively. These four observations will be discussed in more detail in the next section.

Transistors from sample # 4 did not produce any meaningful data. The β's measured were on the order of unity and did not improve with either increasing I_C or deeper isolation etch between the emitter and the base. This was probably

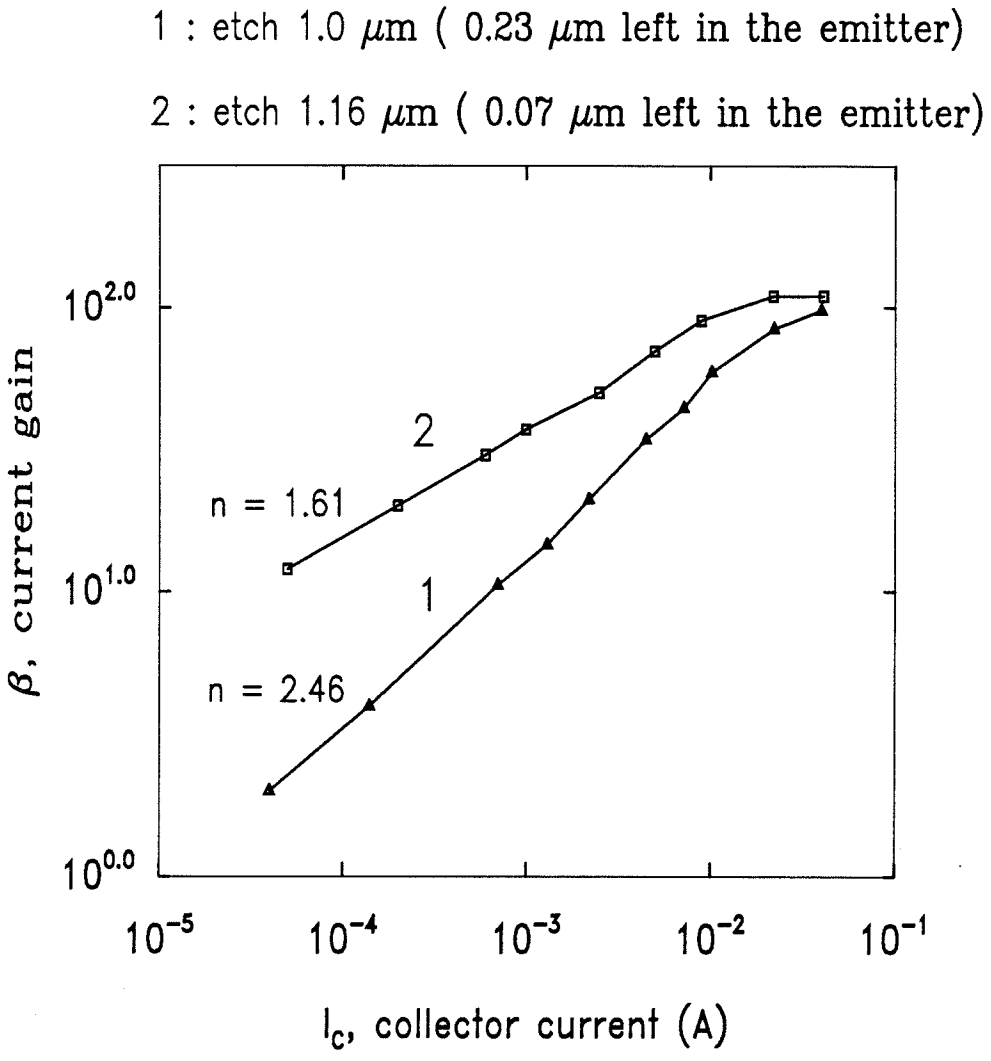


Fig. 3.7 β vs. I_c plots with different isolation etch depths for transistors from sample # 1. Because of the unstable contacts, limited data points are presented

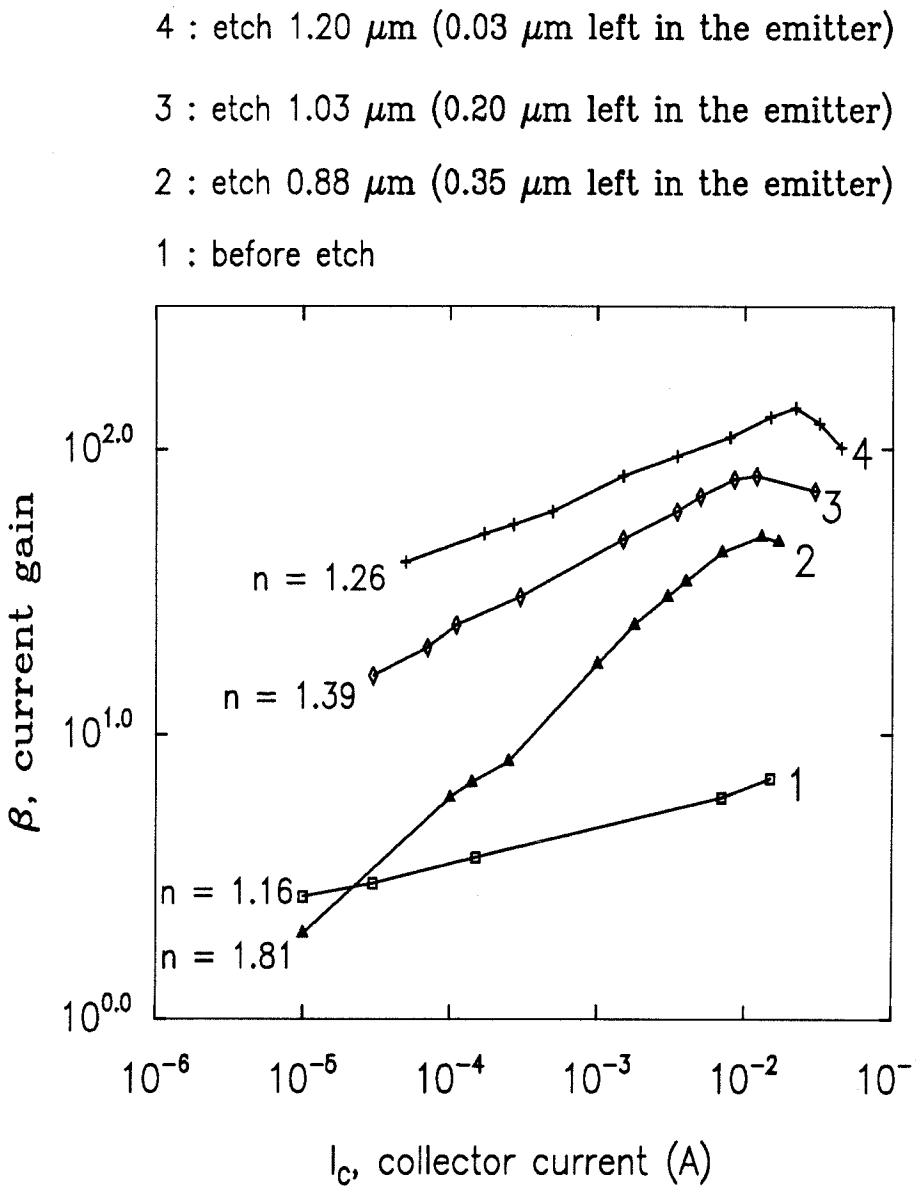


Fig. 3.8 β vs. I_c plots with different isolation etch depths for transistors from sample # 2.

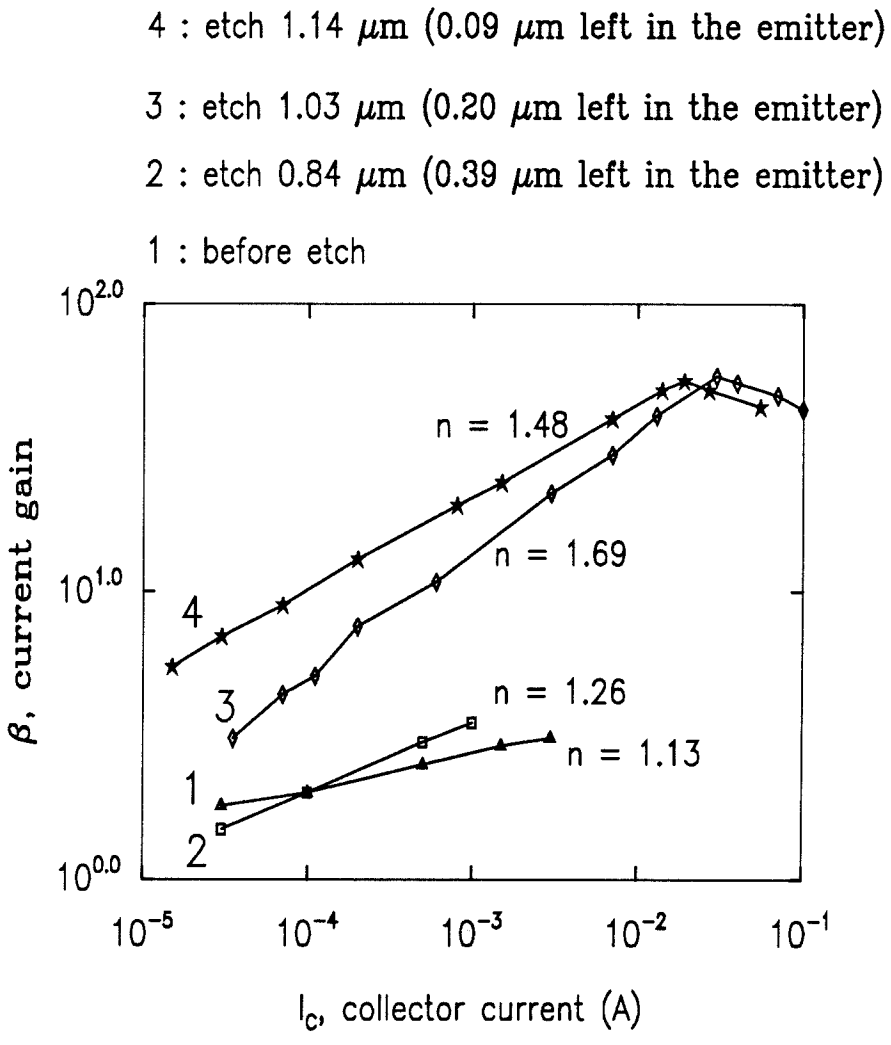


Fig. 3.9 β vs. I_c plots with different isolation etch depths for transistors from sample # 3.

due to the long diffusion time of 218 minutes, which might have introduced other complicating effects. For this reason, data obtained from sample # 4 were not compared with the other three samples in this report.

3.5 Discussion

3.5.1 Dependence of β on I_c

In ideal transistors, β should be a constant and not depend on I_C. The non-ideality comes in due to the fact that not all base currents inject into the emitter or recombine with injected electron current from the emitter with the same electrical characteristics. In fact, there are six components in the base current, as discussed in Sec. 3.2. Also as mentioned in the same section, any carrier that recombines through deep-level traps and defects will show I-V characteristics with an ideality factor equal to 2. Thus, in categorizing the six components in the base current, I_1, I_2, I_4, and I_6 obey the ideal diode law with an ideality of one. I_3 and I_5, which are currents that account for the carriers recombining in the depletion regions, obey Eq. (3.8), which has an ideality factor of two. Thus, in general, I_b can be written as [71]

$$\begin{aligned} I_b &= I_{b0} \cdot exp(\frac{qV_{be}}{kT}) + I'_{b0} \cdot exp(\frac{qV_{be}}{2kT}) \\ &= (I_{b0})_{eff} \cdot exp(\frac{qV_{be}}{nkT}), \end{aligned} \tag{3.14}$$

where I_{b0} and I'_{b0} are the reverse saturation currents associated with carriers recombining in the quasi-neutral regions of the transistor and those recombining in

the depletion region. $(I_{b0})_{eff}$ and n represent the effective reverse saturation current and the overall ideality factor for the junction. It should be noted that n is V_{be}-dependent because the recombination current in the depletion regions becomes less significant as V_{be} increases. Since $I_c \sim I_e$,

$$\begin{aligned}
\beta = \frac{I_c}{I_b} &\sim \frac{I_e}{I_b} \\
&= \frac{I_{e0} \cdot exp(\frac{qV_{be}}{kT})}{(I_{b0})_{eff} \cdot exp(\frac{qV_{be}}{nkT})} \\
&= \frac{I_{e0}}{(I_{b0})_{eff}} \cdot exp(\frac{qV_{be}}{kT}(1 - \frac{1}{n})) \\
&\sim I_e^{(1-\frac{1}{n})} \\
&\sim I_c^{(1-\frac{1}{n})}.
\end{aligned} \tag{3.15}$$

Thus, according to Eq. (3.15), β should increase as I_c increases. This is in agreement with the data obtained as shown in Fig. 3.7, 3.8, and 3.9. The ideality factor obtained through this fashion indicates how ideal the base-emitter junction is in a transistor. A number close to one implies most carriers are recombining in the quasi-neutral regions of the base or emitter. In other words, I_1, I_2, I_4, and I_6 dominate over I_3 and I_5. A ideality factor close to two means carriers are recombining in the depletion regions, or I_3 and I_5 dominate over I_1, I_2, I_4, and I_6. This ideality factor can be a very useful indication of how carriers are transported across the base-emitter junction of the transistors.

As I_c increases further, β eventually declines due to other effects introduced by high-level injection, such as the Kirk Effect. This phenomenon will not be investigated in this work.

3.5.2 Dependence of β on Isolation Etch Depth

Typical HBT's using diffusion or implantation to make contacts to the base involve etching only the highly doped n^+ GaAs cap layer away after the transistors have been fabricated in order to eliminate the leakage currents that might have flown from the base contact through this n^+ layer to the emitter contact directly [39,49-50]. In doing so, I_4 and I_5 are neglected. This assumption is valid if I_4 and I_5 are insignificant compared to the rest of the base current components. However, as shown in Fig. 3.5 and 3.6, a contradiction is strongly evident. From these plots, β increases monotonically to a maximum value as the isolation etch front reaches to a point where the remaining AlGaAs emitter layer is totally depleted. This depletion layer, usually on the order of 1000 $\AA$ thick, is the sum of two depletion mechanisms : one being the depletion region from the base-emitter p-n junction and the other being due to the surface states on the exposed surface. As the isolation etch front reaches this depth, I_4 and I_5 are totally eliminated. As a result, a maximum β is obtained. This remaining layer of depleted AlGaAs emitter acts as a passivation film, which effectively suppresses the recombination taking place on the surface. This is consistent with the observation of Lin et al. [66], who employed the same technique in demonstrating high-gain in HBT's with a double-mesa geometry. The effect of using a thin, but depleted AlGaAs layer as passivation material is further proved in Fig. 3.6, which shows as the isolation etch gets deeper, but before exposing the base layer, the maximum value of β obtained remains the same. However, as the etch exposes the base layer to the air, β degrades until the etch reaches the collector layer, in which case, the transistor is destroyed. The results obtained from this experiment clearly indicate not only the existence of I_4 and I_5, but also their significance. To quantitatively understand the behavior on the improvement of β

as a function of the isolation etch depth, let's denote the isolation etch depth and the depth at which the remaining AlGaAs emitter starts to be totally depleted as measured from the top of the n^+ GaAs cap layer by X_{etch} and X_{E-dep}, respectively. Then,

$$I_4 = (I_4)_{max} \cdot (1 - \frac{X_{etch}}{X_{E-dep}}), \qquad 0 \leq X_{etch} \leq X_{E-dep}, \tag{3.16}$$

where $(I_4)_{max}$ is the amount of emitter bulk current with no isolation etch. Eq. (3.16), which implies that I_4 decreases linearly with X_{etch}, is valid as long as the current density across the AlGaAs base-emitter lateral junction is not affected by the isolation etch. As the carriers are flowing laterally across this junction, some of the carriers will be attracted toward the exposed surface, and recombine on the surface. This is because due to the large number of surface states, there exist potential wells for both the electrons and the holes on the surface [58]. As a result, a portion of I_4 flows toward the exposed surface, which gives rise to I_5. This proportionality can be decribed to the first order by :

$$I_5 = kI_4. \tag{3.17}$$

Subsituting Eq. (3.16) and (3.17) into Eq. (3.6) yields :

$$\beta = \frac{I_c}{I_b} = \frac{I_e - I_1 - I_2 - I_3 + I_6 - (1+k) \cdot (I_4)_{max} \cdot (1 - \frac{X_{etch}}{X_{E-dep}})}{I_1 + I_2 + I_3 + I_6 + (1+k) \cdot (I_4)_{max} \cdot (1 - \frac{X_{etch}}{X_{E-dep}})}. \tag{3.18}$$

The maximum and minimum β's shown in Fig. 3.5 thus correspond to setting $X_{etch} = X_{E-dep}$ and $X_{etch} = 0$ in Eq. (3.18). Ignoring I_6, the results are :

$$\beta_{max} = 100 = \frac{I_E - I_1 - I_2 - I_3}{I_1 + I_2 + I_3} \tag{3.19}$$

$$\beta_{min} = 3 = \frac{I_E - I_1 - I_2 - I_3 - I_4 - I_5}{I_1 + I_2 + I_3 + I_4 + I_5}. \tag{3.20}$$

Solving Eq. (3.19) and (3.20) simultaneously shows :

$$(I_E)_{sample2} = 101(I_1 + I_2 + I_3)_{sample2} \tag{3.21}$$

and

$$(I_4 + I_5)_{sample2} = \frac{97}{4}(I_1 + I_2 + I_3)_{sample2}. \tag{3.22}$$

Performing the same operations on sample 3 reveals :

$$(I_E)_{sample3} = 29(I_1 + I_2 + I_3)_{sample3} \tag{3.23}$$

and

$$(I_4 + I_5)_{sample3} = 13.5(I_1 + I_2 + I_3)_{sample3}. \tag{3.24}$$

Eq. (3.22) and (3.24) demonstrate how important I_4 and I_5 are. They simply state that the recombination currents taking place in the emitter bulk region and on the exposed surface are at least an order of magnitude bigger than the sum of I_1, I_2, and I_3. Thus, eliminating I_4 and I_5 would increase β by at least an order of magnitude.

Another effect that isolation etch has on β is that the extent of electron lateral diffusion is reduced as the isolation etch gets deeper. In terms of Fig. 3.1, this means that the part of I_2 that spreads laterally in the emitter before injecting into the base is reduced because the isolation etch etches away the region into which the electrons might otherwise diffuse. As a result, electrons are better confined in the horizontal direction and the base transport factor is improved.

A subtle difference in β-improvement between sample # 2 and # 3 should be pointed out. In sample # 2, β improves gradually as the isolation etch increases and eventually increases at a much faster rate as the isolation etch front approaches the base-emitter junction. This increasing behavior can be approximately described by Eq. (3.18). This is, however, not so for sample # 3. As shown in Fig. 3.6, β remains relatively constant until the isolation etch depth approaches a certain depth. Then, the β increases drastically to the maximum value. The distinct behavior in the improvement of β for sample # 3 can be qualitatively understood in terms of the shortening of minority carrier lifetime due to a longer diffusion time. This will be explained Sec. 3.5.4.

3.5.3 Dependence of the Ideality Factor on Isolation Etch Depth

As shown in Fig. 3.8 for sample # 2, ideality factor for the base-emitter junction seems to change as the isolation etch between the base and emitter changes. This is an indication that the isolation etch affects the transport mechanisms of electrons and holes across the base emitter junction. Before the etch, the junction exhibits an ideality factor of 1.16 and the values of the β's are less than 10. This suggests that the majority of the base current flows throught the $p^+ - n^+$ junction in GaAs cap layer to the emitter. This will account for the low β observed because only a little portion of the base current contributes to the transistor gain action. Furthermore,

a low ideality factor of 1.16 indicates that the holes and electrons are being transported across the p-n junction by diffusion, which obeys the ideal diode law. Thus, it is important to eliminate this current path. An interesting phenomenon occurs when this $p^+ - n^+$ junction inside the GaAs cap layer is etched away. As shown in curve 2 of Fig. 3.8, the ideality factor degrades to 1.81 as the isolation etch depth increases to 0.88 μm. At this point, the current that originally flows in the cap layer is forced to flow in the AlGaAs emitter layer, which is partially exposed to air due to the isolation etch. As carriers are flowing throught the emitter bulk region, part of them are pulled toward the exposed surface and recombine there. This part of the current contributes to an ideality factor of 2. Thus, an ideality factor of 1.86 implies majority of the base current is flowing toward the exposed surface. This current, again, contributes no transistor gain action. Thus the β's remain low at low biases. As the bias increases, this current becomes less significant due to the $exp(qV/2kT)$ factor as compared to the gain-contributing part of the base current, which varies as $exp(qV/kT)$. As a result, β increases as the bias increases. Further etch on the isolation trench improves the ideality factor from 1.39 at an etch depth of 1.03 μm to 1.26 at an etch depth of 1.2 μm. This can be explained by the fact that as the partially exposed AlGaAs emitter gets thinner, the bulk current flowing through the emitter gets less, which, in turn, decreases the amount of the current flowing to the exposed surface. Consequently, the ideality factor improves and β increases dramatically especially at low biases because the gain-contributing part of the base current has become the dominating component of the total base current. Further etch on the isolation trench does not improve the ideality factor any more due to the existence of I_3, which is not affected by the isolation etch, takes place inside the base-emitter junction depletion region and has an ideality factor of 2. Thus, the transport mechanisms of electrons and holes across the base-emitter junction can be pictured clearly with the aid of the ideality factor.

General agreements are also obtained for sample # 1 and # 3 as shown in Fig. 3.7 and 3.9. However, in Fig. 3.7, an ideality factor of 2.46 has been observed for an isolation etch depth of 1 μm. The phenomenon of having an ideality factor greater than 2 has been reported by Ghannam et al. [72]. This peculiarity can not be explained by the conventional means. In Fig. 3.9, the trend in the improvement of the ideality factor is in agreement with that in Fig. 3.8, supporting the validity of the explanation given above.

3.5.4 Dependence of Maximum β on Diffusion Time

There is a clear degradation in the maximum value of β obtained as the diffusion time increases. The maximum β's are 200, 120 and 54 for sample # 1, # 2 and # 3, which have diffusion times of 25, 40 and 88 minutes, respectively. This degradation may be due to two factors, one being the out-diffusion of Zn into the emitter layer, and the other being the degradation of minority carrier lifetime in the base. These two possibilities will be investigated separately.

The problem of base dopants out-diffusion during the growth process or subsequent high temperature processing is a notorious phenomenon [73,74]. Since Zn has the highest diffusivity than any other base dopant, it is of prime importance to be able to confine Zn in the base so that the integrity of the heterojunction is maintained. Thus, undoped GaAs spacer layers have been inserted betweeen the base-emitter and base-collector junctions to prevent the out-diffused Zn from reaching the emitter region [34]. While a thin spacer layer does not effectively block the out-diffusion of Zn, a thick one, which presumably is totally depleted under the normal transistor biasing condition, introduces more space charge recombination current, which is characterized by an ideality factor of 2. If Zn does out-diffuse into the emitter layer even with the insertion of spacer layers of proper thickness, the

emitter injection efficiency of the HBT is governed by that of a homojunction instead of the original heterojunction. Quantitatively, the emitter injection efficiency, γ_e, is given by Eq. (3.2) instead of (3.5). At the same time, the base transport factor is also decreased due to the widening of the base, as seen in Eq. (3.3). Thus, the longer the diffusion is, the more out-diffusion of Zn there is and the more β decreases. This analysis assumes that the base doping concentration is higher than that in the emitter so the out-diffused Zn will be able to convert a thin layer of n-type AlGaAs emitter layer near the base-emitter junction to p-type AlGaAs layer, which forms part of the base. As one can see from the material parameters of the sample used in this experiment, this is not possible because the base doping concentration of 2×10^{17} cm^{-3} is less than the emitter doping concentration of 4.2×10^{17} cm^{-3}. Thus, the degradation of β is probably due to the decrease of minority carrier lifetime as a result of heat treatment in the diffusion process.

Assuming the integrity of heterojunction is still maintained after the Zn-diffusion process, the β of the HBT continues to be dominated by the base tranport factor. A possible explanation of the inverse relationship between the β and the diffusion time is the following. Upon substituting Eq. (3.3) into Eq. (3.4) with the assumption of $\gamma_e = 1$, one obtains :

$$\beta = 2\left(\frac{L_b}{W_b}\right)^2 = 2\frac{D_b\tau_b}{W_b^2}, \tag{3.25}$$

where D_b and τ_b are minority carrier diffusion coefficient and lifetime, respectively, in the quasi-neutral region of the base. Thus, β is linearly proportional to τ_b. τ_b is further composed of two factors ;

$$\frac{1}{\tau_b} = \frac{1}{(\tau_b)_{intrinsic}} + \frac{1}{(\tau_b)_{defect}}, \tag{3.26}$$

$(\tau_b)_{intrinsic}$ and $(\tau_b)_{defect}$ are the original electron lifetime in the base and the electron lifetime due to newly created defects from the heat treatment in the Zn-diffusion process, respectively. $(\tau_b)_{intrinsic}$ is related to the base doping concentration, N_a, by [59] :

$$(\tau_b)_{intrinsic} = \frac{1}{b_{intrinsic} \cdot N_a}. \tag{3.27}$$

Similarly, $(\tau_b)_{defect}$ is related to the defect concentration, N_{defect}, by [75]

$$(\tau_b)_{defect} = \frac{1}{\sigma_n v_{th} N_{defect}}, \tag{3.28}$$

where $B_{intrinsic}$ (sec$\cdot$ cm^{-3}) is the probability of recombination, σ_n is the electron capture cross section, N_{defect} is the defect concentration, and v_{th}, the thermal velocity for electrons, is equal to $\sqrt{\frac{3kT}{m^*}}$, where m^* is the effective mass for the electrons. The concentration of defects as a function temperature and time can be further expressed as [76,77]:

$$N_{defect} = N \cdot exp(-\frac{w}{kT}) \cdot (1 - exp(-\frac{t}{\tau_g})), \tag{3.29}$$

where

$$\frac{1}{\tau_g} = A \cdot exp(-\frac{Q_r}{kT}) \tag{3.30}$$

and w and Q_r are energy needed to generate one defect and activation energy for recrystallization, respectively, N is the volume density of atoms, and A is some constant. In other words, defect population depends on temperature in an $exp(-w/kT)$

fashion and on time in an exponential fashion with the time contant given by Eq. (3.30).

Figure 3.10 shows the plot of log β versus log t for sample # 1, # 2, and # 3, with t being the diffusion time. A straight line with slope equal to -1 is obtained. This implies that the maximum β varies inversely proportional to the amount of time the transistors are exposed to high temperature ambient. This result can be explained by the following arguments and assumptions.

Assuming τ_g is a very big number, Eq. (3.29) can be approximated by :

$$N_{defect} \approx N \cdot exp(-\frac{w}{kT}) \cdot \frac{t}{\tau_g}, \tag{3.31}$$

This assumption is consistent with the fact that the degradation of lifetime due to the generation of defects in a high temperature environment is seldom recovered after the temperature is lowered to original low temperature [78,79]. This is because the τ_g for subjecting a piece of semiconductor to high temperature is much lower than that for subjecting the same sample to low temperature througth the T dependence in the exponent of Eq. (3.30). For the condition of Zn diffusion in the present study, the time constant for defect annihilation is approximately the cube of the time constant for defect generation because the room temperature is roughly 1/3 the temperature used in Zn-diffusion (300 K versus 900 K). Thus, the assumption of τ_g being a big enough number so that Eq. (3.31) is valid is consistent with the irrecoverability of lifetime in a sample after going through a high temperature treatment. The result of this assumption predicts a linear dependence of defect concentration on time, or an inversely proportional relationship between $(\tau_b)_{defect}$ and time. In the present case, the assumption is valid as long as $\tau_g >> 88$ minutes.

From Eq. (3.27), $(\tau_b)_{intrinsic}$ is on the order of 10^{-7} sec because $B_{intrinsic}$ is roughly 10^{-10} sec$\cdot$ cm^{-3} for GaAs and the base doping concentraion is approxi-

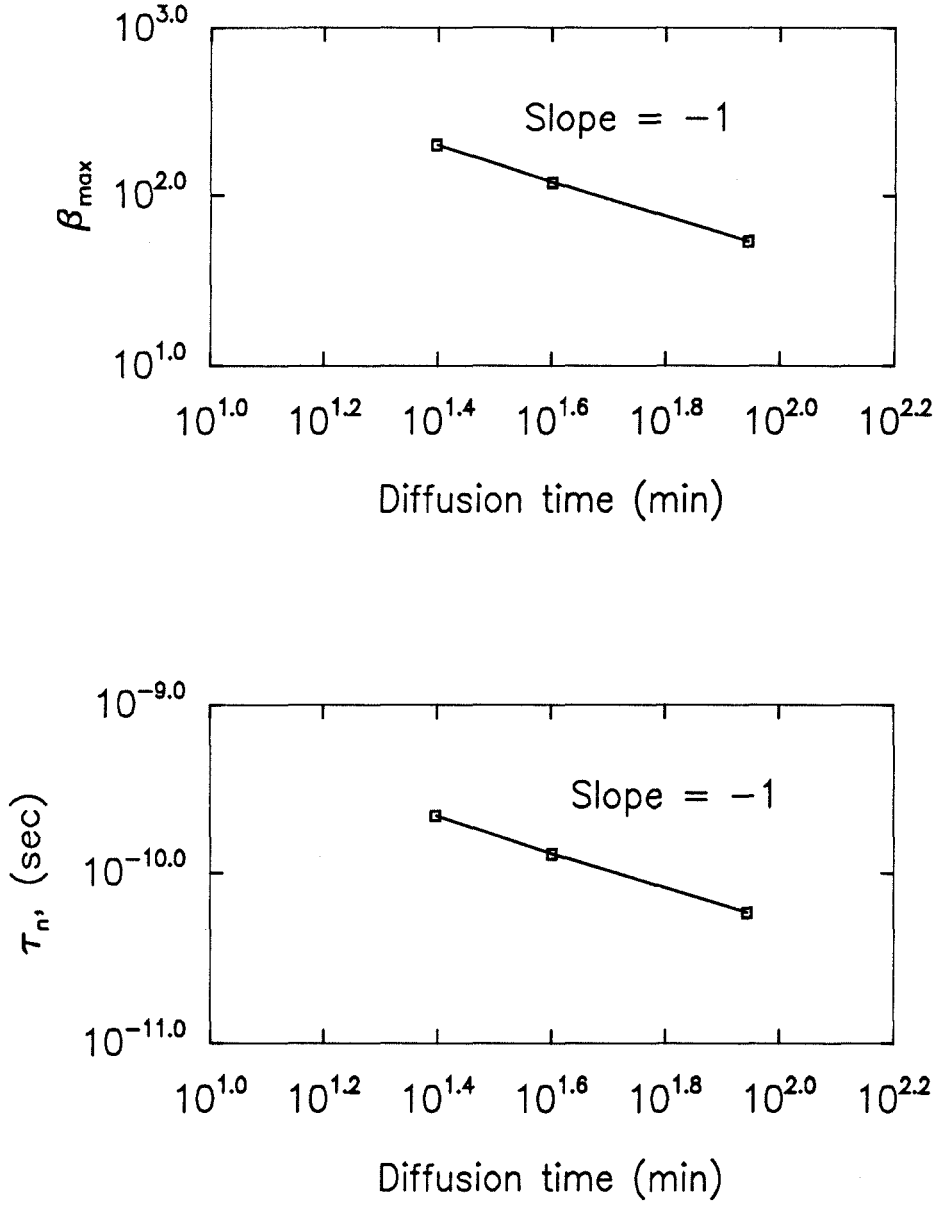


Fig. 3.10 Log to log plot for maximum β obtained in samples # 1, # 2, and # 3 as function of the diffusion time. The slope of -1 indicates that $\beta_{max} \sim 1/t$

mately 10^{17} cm^{-3}. Thus, as long as $(\tau_b)_{defect} << (\tau_b)_{intrinsic}$ holds, Eq. (3.26) can be re-written as :

$$\tau_b \approx (\tau_b)_{defect}. \tag{3.32}$$

Substituting Eq. (3.32), (3.31), and (3.28) into Eq. (3.25) yields :

$$\beta \sim \frac{1}{t}, \tag{3.33}$$

which agrees with Fig. 3.10.

The validity of Eq. (3.33) depends on the assumption that $(\tau_b)_{defect}$ is much smaller than 10^{-7} seconds. This puts a requirement on defect concentration generated by the high temperature treatment in the Zn diffusion process. For ion implanted samples, this requirement is easily met as the minority carrier lifetime is on the order of 100 ps [55] after the implemention. Since ion implantation involves physically damaging the atoms and subsequently annealing at high temperatures, it is not unreasonable to assume the minority carrier lifetime in the diffused HBT is degraded by the same mechanism and perhaps to a comparable extent. In addition, the experimental results presented in Fig. 6.10 strongly suggests this assumption.

The shortening of minority carrier lifetime due to heat treatment can also be applied to explain the difference in the improvement of β as a function of isolation etch depth in samples # 2 and # 3, as previously mentioned in Sec. 3.5.2. The β of the transistors from sample # 2 improves gradually as the isolation etch increases and eventually increases at a much faster rate as the isolation etch approaches the base-emitter junction. However, the β of transistors from sample # 3 remains relatively constant until the isolation etch depth almost reaches the base-emitter

junction. Then β increases dramatically to the maximum value. The sluggishness in the improvement of β for transistors from sample # 3 can be attributed to the degradation of minority carrier lifetime in the emitter by the heat treatment. This can be possibly explained by considering the magnitude of leakage current, I_4, in both samples.

$$\begin{aligned} I_4 &= Aq(D_e\frac{dp}{dx} + D_b\frac{dn}{dx}) \\ &= Aq(D_e\frac{p_{n0}}{L_e} + D_b\frac{n_{p0}}{L_b})(exp(\frac{qV_{be}}{kT}) - 1) \\ &= Aqn_i^2(\frac{D_e}{L_eN_e} + \frac{D_b}{L_bN_b})(exp(\frac{qV_{be}}{kT}) - 1) \\ &= Aqn_i^2(\sqrt{\frac{D_e}{\tau_e}}\frac{1}{N_e} + \sqrt{\frac{D_b}{\tau_b}}\frac{1}{N_b})(exp(\frac{qV_{be}}{kT}) - 1), \end{aligned} \tag{3.34}$$

where τ_e and τ_b are the minority carrier lifetime in the emitter and the base, and p_{n0} and n_{p0} are the intrinsic carrier concentration of holes in the emitter and of electrons in the base, respectively. For sample # 2 and # 3, $N_b >> N_e$ because of the Zn-diffusion. Thus, Eq. (3.34) can be approximated by :

$$I_4 \approx Aqn_i^2\sqrt{\frac{D_e}{\tau_e}}\frac{1}{N_e}(exp(\frac{qV_{be}}{kT}) - 1). \tag{3.35}$$

Using the argument presented earlier, $(\tau_e)_{sample2} > (\tau_e)_{sample3}$ because sample # 2 has a shorter diffusion time than sample # 3. As a result, $(I_4)_{sample3} > (I_4)_{sample2}$. This means that at exactly the same conditions, the β of transistors from sample # 3 will be lower than that of transistors from sample # 2. This is why the β of the transistors from sample # 3 does not improve gradually as the isolation etch depth gets deeper, unlike the transistors from sample # 2. As the isolation etch front approaches the base-emitter junction, I_4 in transistors from sample # 3, though

still larger than that in transistors from sample # 2, is eventually eliminated by the isolation etch. At which point, the β increases dramatically to the maximum β. This is in agreement with the observation as shown in Fig. 3.6. This effect is further amplified by the fact that I_5 is proportional to I_4. Thus, the sluggish improvement in the β for sample # 3 is the result of having a larger I_4 and I_5 than the same counterparts for sample # 2. To summarize this phenomenon, the long diffusion time, overall, has two undesired effects on the performance of the HBT : degradation of the maximum achievable β, and inresponsiveness in the improvement of β as a function of the isolation etch depth.

3.6 DC Switching Characteristics

In order to understand the switching dynamics of a DHBT-based optoelectronic neuron, it is necessary to understand the switching behavior of an individual transistor. In particular, the turn-on and turn-off delays in the output current, I_c, in response to a step current in the input, I_b, are of great interest. Shown in Fig. 3.11 is the schematic circuit diagram of a bipolar transistor and the relative timing of the base and collector currents. In the circuit, a resistor of resistance, R, has been included to account for all the series, parasitic and contact resistances seen by I_c. If we apply a step base current, I_b, the collector current generally rises exponentially with a time constant of t_r. However, if the base current is suddenly decreased to zero, the collector current will not decrease immediately. Instead, it will remain unchanged for a certain delay time of t_d and then decrease exponentially to zero with a time constant of t_f. In the analysis that follows, we will examine the factors that affect these timing parameters, t_r, t_d, and t_f in detail as the ultimate speed of the neuron which incorporates the integration of bipolar transistors will be invariably limited by these parameters.

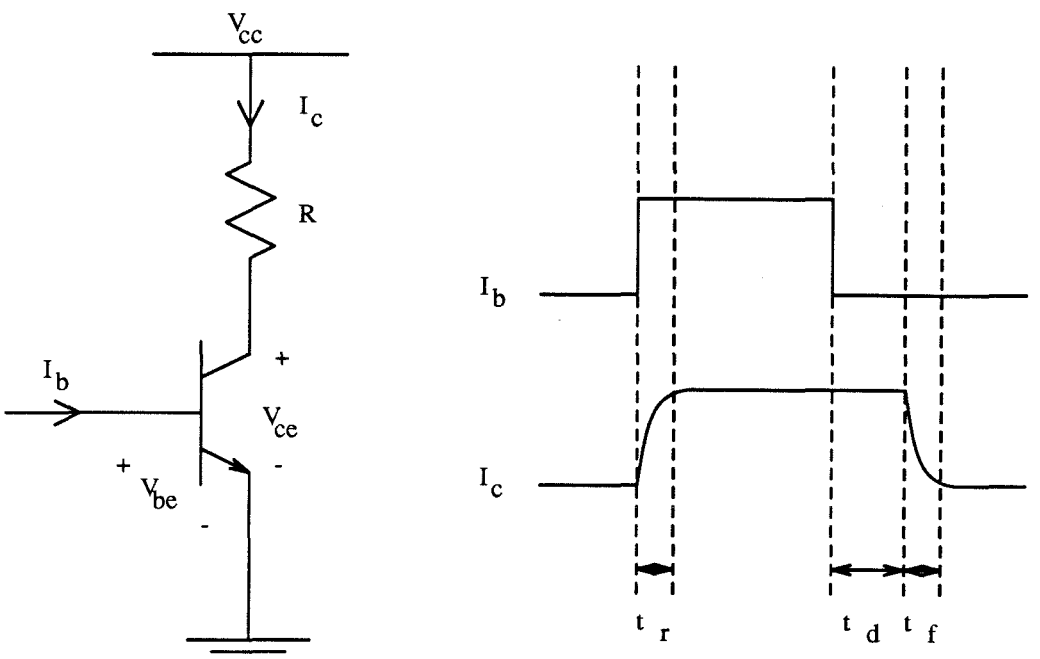


Fig. 3.11 The schematic circuit diagram of a bipolar transistor with a series resistance, R, which accounts for all series, parasitic and contact resistances and the timing diagram of I_b and I_c, showing the delays in I_c in responding to a step I_b.

The operation of a bipolar transistor falls into 4 different regions. They are cut-off, normal active, onset of saturation, and saturation regions. The minority carrier distributions in the emitter, base and the collector regions of a N-p-N transistor are graphically illustrated in Fig. 3.12(a)-(d). In the cut-off region, the base-emitter and the base-collector junctions are reverse biased, causing the minority carriers in the base, which are electrons, to be totally depleted. There is no current flowing through the transistor except for the leakage current that flows across the reverse biased pn junctions. Thus, the external current is very small. This case is illustrated in Fig. 3.12(a). When the base-emitter junction becomes forward biased and the base-collector junction still remains reverse biased, minority carriers in the base start to build up as a result of the injection of electrons from the emitter. Some of these injected carriers recombine with the holes from the base, while the rest of the carriers diffuse across the base before they are swiftly swept across the base-collector depletion region and are collected by the collector. This situation is shown in Fig. 3.12(b). Since the minority carriers remain on an average of their lifetime, τ_b, in the base before they recombine with the holes in the base, the recombination current is equal to the total charge in the base divided by the carrier lifetime. If we further assume that the hole injection from the base into the emitter is negligible, the base current is then equal to this recombination current. Thus, the following relationship can approximately established:

$$I_b = \frac{Q_{base}}{\tau_b}. \tag{3.36}$$

As the base current continues to increase, the minority carriers in the base build up in proportion. However, the slope of the minority carrier concentration in the base is proportional to the magnitude of the collector current, which is equal to the product of the base current and the current gain, β. Eventually, the minority

carriers in the base build up to a level that removes all the reverse bias across the base-collector junction. This is the onset of saturation, which is shown in Fig. 3.12(c). At this point, the slope of the minority carrier concentration in the base is determined by the maximum current that can flow through the transistor, which is given by

$$I_{c,sat} = \frac{V_{cc} - V_{ce,sat}}{R}. \tag{3.37}$$

Thus, the collector saturation current is predominantly controlled by the value of the series resistance. Beyond this point, the charge in the base continues to increase as the base current increases. However, the slope of the base charge profile maintains constant because the collector current has already reached its maximum value. As the transistor is driven far into the saturation, the common emitter saturation voltage of the transistor becomes close to zero. This simplifies Eq. (3.37) to

$$I_{c,sat} \approx \frac{V_{cc}}{R}. \tag{3.38}$$

Moreover, the base-collector junction becomes forward biased. Thus, electrons are injected from both the emitter and the collector. This can be thought of as the superposition of two transistors. One is in the forward active mode and the other one is in the reverse active mode because the collector starts to act like the emitter and vice versa. This case is graphically illustrated in Fig. 3.12(d).

By applying a square-wave current with a magnitude greater than $I_{c,sat}/\beta$ in the base, the transistor can be switched between the cut-off and the saturation regions because any further increase in I_b beyond this value will not result in any

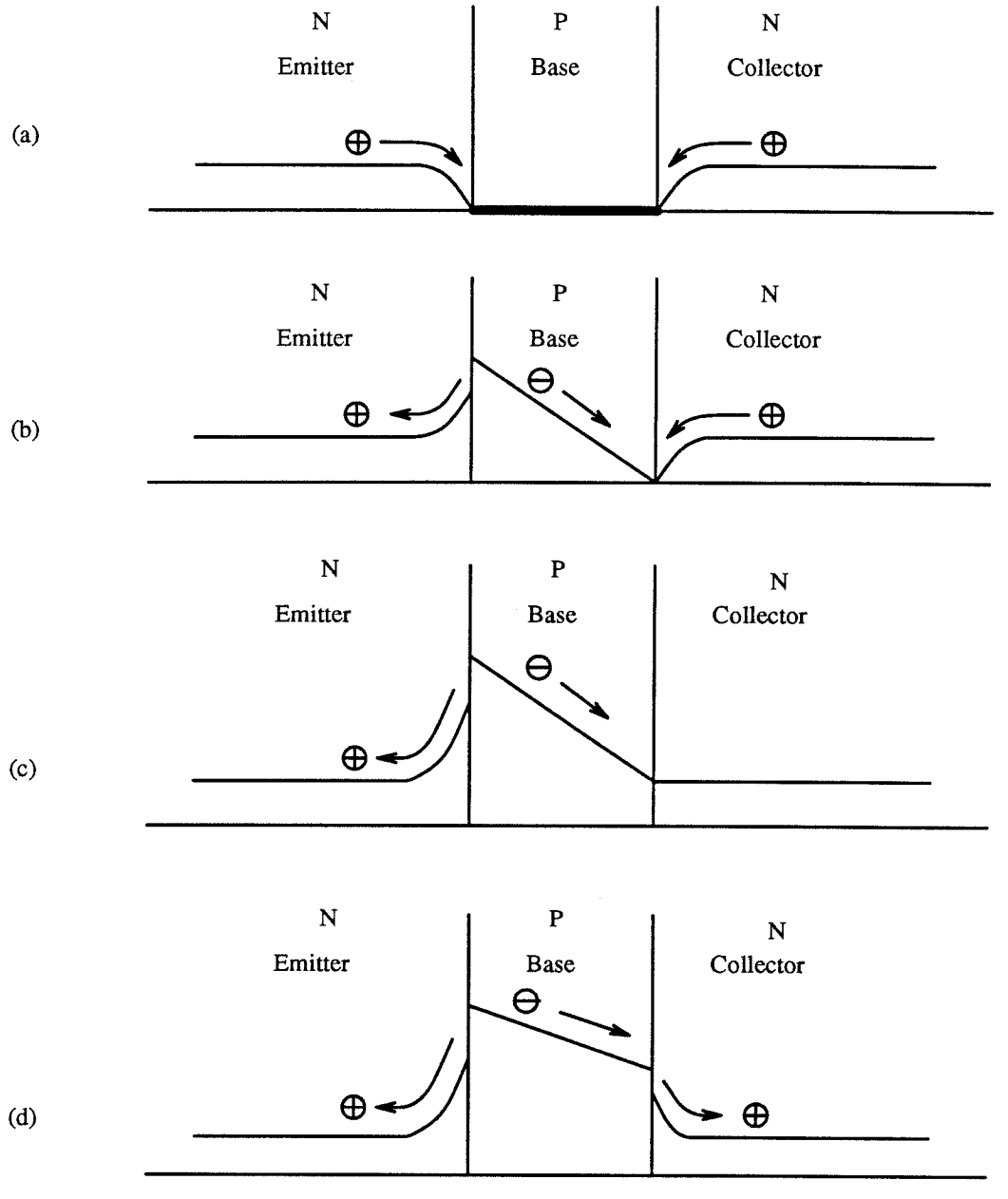


Fig. 3.12 The minority carrier concentration profile in the emitter, base and collector of a transistor in **(a)** cut-off region. **(b)** forward active region. **(c)** onset of saturation. **(d)** saturation region.

increase in I_c as I_c is clamped by V_{cc}/R. This is shown in Fig. 3.13, in which the common emitter I-V characteristics of a transistor for various I_b and its load curve are illustrated. To switch the transistor from cut-off to saturation, electrons are injected into the base until the saturation depicted in Fig. 3.12(d) is reached. It can therefore be seen that the risetime in the collector current depends on how quickly the electrons in the base establish themselves to be a certain profile governed by the magnitude of the base current as well as the series resistance. If we denote the total charge stored in the base at any time t to be $Q_b(t)$, the base and collector currents at any time t are respectively given by

$$I_b(t) = \frac{Q_b(t)}{\tau_b} \tag{3.39}$$

$$I_c(t) = \frac{Q_b(t)}{\tau_{tr}}, \tag{3.40}$$

where τ_{tr} is the base transit time for the electrons. Again, the assumption that base current is dominated by the recombination current in the base is employed in deriving Eq. (3.39). That is, the hole injection from the base into the emitter contributes negligibly to the total base current. This assumption is valid especially in heterojunction bipolar transistor because the valence band discontinuity effectively eliminates hole injection into the base. Equation (3.39) describes the relationship between the total charge stored in the base and the base current at steady-state. However, during the transient state, an extra term accounting for the time transient has to be added because the steady-state Q_b can not instantaneously follow any variation in the base current. Thus, Eq. (3.39) becomes

$$I_b(t) = \frac{Q_b(t)}{\tau_b} + \frac{dQ_b(t)}{dt}. \tag{3.41}$$

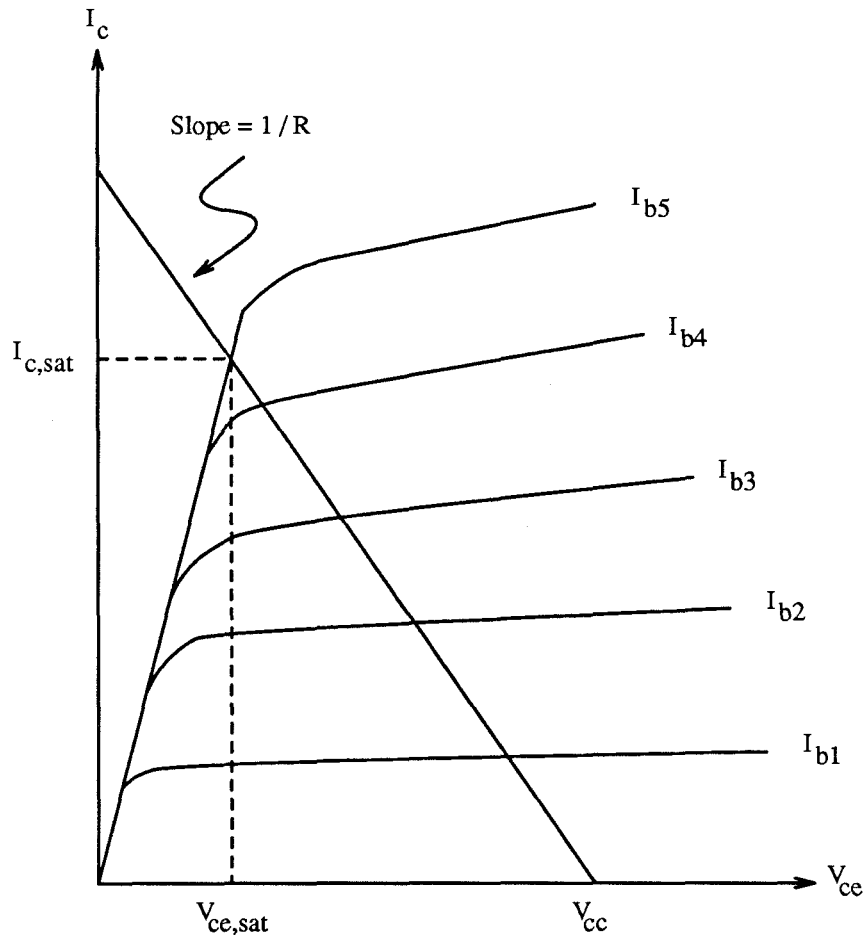


Fig. 3.13 I-V characteristics of a common-emitter bipolar transistor loaded with a resistor of resistance R. The maximum current that the collector current can attain is $I_{c,sat}$, which is determined by the loading resistance and $V_{ce,sat}$. Further increase in I_b will not increase I_c as I_c is clamped by $I_{c,sat}$.

Solving the differential equation in Eq. (3.41) yields

$$Q_b(t) = I_b\tau_b(1 - e^{-t/\tau_b}). \tag{3.42}$$

Substituting Eq. (3.42) into Eq. (3.40) shows

$$\begin{aligned} I_c(t) &= \frac{Q_b(t)}{\tau_{tr}} \\ &= \frac{\tau_b}{\tau_{tr}} \cdot I_b \cdot (1 - e^{-t/\tau_b}) \\ &= \beta \cdot I_b \cdot (1 - e^{-t/\tau_b}). \end{aligned} \tag{3.43}$$

It is interesting to note that β can be expressed as τ_b/τ_{tr}. In fact, β can be re-written in a more familiar fashion as shown below.

$$\beta = \frac{\tau_b}{\tau_{tr}} = \frac{\tau_b}{\frac{W_b^2}{2D_n}} = \frac{2D_n\tau_b}{W_b^2} = 2(\frac{L_n}{W_b})^2, \tag{3.44}$$

where L_n, D_n and W_b are the electron diffusion length, electron diffusion coefficient and base width respectively. Equation (3.43) implies that the value of I_c reaches its steady state with a characteristic time equal to the lifetime of the electrons in the base. This characteristic time is independent of the magnitude of the I_b applied to the base of the transistor. Thus, regardless of how large the input step is, the collector current rises with a time constant equal to τ_b. However, in reality, the maximum current that can be sustained by the collector is given in Eq. (3.38), which is fixed by the external circuit parameters. Therefore, for base current exceeding $I_{c,sat}/\beta$, the collector can only reach its saturation current level and Eq.

(3.43) is no longer valid. However, if we continue to increase the base current, the time over which the collector current reaches the collector saturation current will be decreased owing to a faster rate of rise in the collector current. This is illustrated in Fig. 3.14(a). In this figure, two base currents of different magnitudes, I_{b1} and I_{b2}, where $I_{b2} > I_{b1}$, are applied to the transistor. If we now ignore the fact the maximum collector current is clamped at $I_{c,sat}$, both collector currents will increase to their steady-state collector currents I_{c1} and I_{c2} respectively with the same time constant, τ_b. However, because, in reality, collector current can only reach $I_{c,sat}$, as soon as both collector currents reach this value, they saturate. As a result, I_{c2} has a shorter overall risetime than I_{c1} as shown in the figure. This risetime can be calculated by equating Eq. (3.43) to 90% of $I_{c,sat}$ and then solving for t. The result is the following :

$$t_r = \begin{cases} \tau_b ln \dfrac{1}{1 - \dfrac{0.9 I_c}{\beta I_b}} & I_b < \dfrac{I_{c,sat}}{\beta} \\ \\ \tau_b ln \dfrac{1}{1 - \dfrac{0.9 I_{c,sat}}{\beta I_b}} & I_b > \dfrac{I_{c,sat}}{\beta} \end{cases} . \tag{3.45}$$

Equation (3.45) assumes that the risetime is defined as the time needed to reach 90 % of the final value. Also, for the case $I_b < I_{c,sat}/\beta$, I_c is equal to βI_b. Therefore, t_r can be re-expressed by $\tau_b ln 10$ by cancelling these two factors. The functional dependence of risetime, t_r, on the base current, I_b, is shown in Fig. 3.14(b). As can be clearly seen, the risetime remains constant until the the transistor is driven into saturation, which causes the risetime to decrease as the base input current increases. This presents a tradeoff between the power and speed. This consideration will be important in the design of an optoelectronic neuron.

After the turn-on transient has settled down in a transistor that has been driven

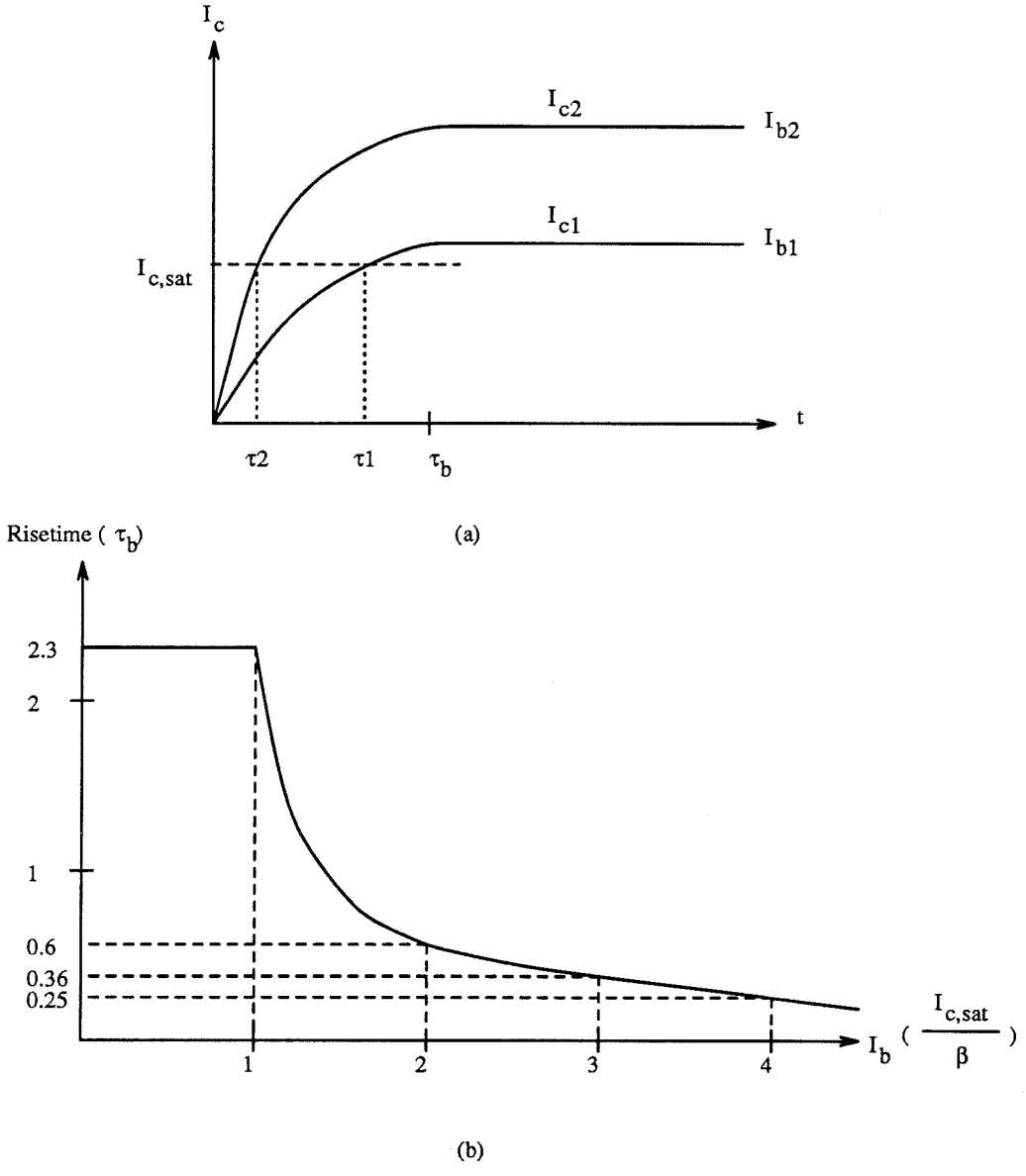


Fig. 3.14 **(a)** Turn-on characteristics of collector current for two different base input currents. Because the transistor is driven into saturation, the large base input current results in a smaller risetime in the collector current. **(b)** Risetime as a function of the magnitude of the base current. In the forward active mode, the risetime is constant. However, if the transistor is driven into saturation, the risetime decreases with increasing base current.

into saturation, the total charge stored in the base from the injection of electrons from the emitter is determined by the magnitude of the base current and the lifetime of the electrons. Specifically, it can be found by taking the limit of t going to infinity in Eq. (3.42). The result is

$$Q_b = \lim_{t\to\infty} i_b\tau_b(1 - e^{-t/\tau_b}) = I_b\tau_b. \tag{3.46}$$

As the base current is suddenly decreased to zero, I_c remains unchanged for a period of time, as denoted by t_d shown in Fig. 3.11, and then decreases to zero exponentially with a time constant of t_f. This can be explained as follows. As the base current is removed, the stored charge in the base decays exponentially with a time constant equal to the electron recombination lifetime in the base. This decrease in the stored charge takes place with the slope of the electron concentration profile in the base remained constant. This is due to the fact that I_c is still clamped at $I_{c,sat}$ in the saturation region of the transistor. I_c remains constant until the stored charge in the base has decreased to the onset of the saturation. Beyond which, the transistor returns to its forward active mode and I_c decreases exponentially to zero. To calculate the time delay in the collector current in response to a decrease in the base current, we have to calculate the time it takes to decrease the steady-state stored charge in the base shown in Eq. (3.46) to the stored charge at the onset of the saturation, which, according to Eq. (3.40), is equal to $I_{c,sat}\tau_{tr}$. Thus, the following equation can be set up :

$$I_{c,sat}\tau_{tr} = I_b\tau_b e^{t_d/\tau_b}. \tag{3.47}$$

Solving for t_d in Eq. (3.47), we obtain

$$t_d = \tau_b ln \frac{i_b \tau_b}{I_{c,sat} \tau_{tr}}. \tag{3.48}$$

For a transistor operating between the cut-off and the forward active modes, there is no delay time, t_d, as the discharge of the charge built up in the base results in an immediately decrease in the collector current. Thus, while a larger base input decreases the risetime of the collector current, it also increases the delay time in shutting off the collector current after the base current is turned off. After the transistor has reached the forward active mode in the process of discharging its base stored charge, the collector current falls off with the same time constant as it goes up. Thus, the rate at which the collector current falls off can be described by

$$I_c = I_{c,sat} e^{-t/\tau_b}. \tag{3.49}$$

Again, if we define the falltime, t_f, to be the time it takes to fall to 10% of its steady-state value, then

$$t_f = \tau_b ln \frac{I_{c,sat}}{0.1 I_{c,sat}} = 2.3 \tau_b. \tag{3.50}$$

In conclusion, the turn-on and turn-off transients are determined by the rate at which the base stored charge is built up or removed. By switching the transistor between cut-off and saturation, the risetime can be reduced by applying a base current of larger magnitude. However, this is gained at the expense of a larger turn-off time because the total charge stored in the base can not be easily and quickly removed. If the transistor is designed to switch between the cut-off and the forward active mode, then the risetime and falltime are independent of the

magnitude of the base current and are equal to each other. In which case, the risetime and falltime are solely determined by the electron recombination lifetime in the base.

3.7. Conclusion

A systematic study of the effects of leakage currents and diffusion on β has been qualitatively and experimentally presented in this chapter. It is found that an isolation deeper than the n^+ cap layer is necessary to maximize the β of the transistors. The optimum depth is when a thin and depleted layer of AlGaAs emitter is left on top of the base and functions as a passivation film. The improvement in β through this method is especially impressive at low-bias conditions, where the surface recombination current dominates. The maximum β achievable is also found to vary inversely proportional with the diffusion time. The degradation process has been examined and suggested to be due to the shortening of minority carrier lifetime in the base. In arriving at this suggestion, two assumptions have been made. One is the assumption of long time constant in generating defects due to the elevated temperatures. The other one is the carrier lifetime due to the newly generated defects along is much shorter than that without these defects. Thus, the overall recombination process is dominated by the newly generated defects, which, in turn, are created by the high-temperature process. To avoid these problems, future transistors should have an isolation deep enough between the emitter and base so that transistor performance will not be dominated by the surface action. In addition, the future emitter layer thickness should be decreased so that the total time required for Zn to diffuse down to the base layer is less. As a result, less degradation in the minority carrier lifetime and thus higher β can be expected.

Chapter 4

Double-Heterojunction Bipolar Transistor-Based Neurons

4.1 Introduction

The input-output characteristics of an optoelectronic neuron is approximated by a thresholding function, whose output remains zero until the input exceeds a predetermined threshold. Beyond this threshold, the output level saturates. This input-output functionality can be easily implemented in integrated circuits fabricated by standard GaAs processing technology. Since the neuron has to have an optical input and an optical output, additional elements, such as the light detector and the light emitter, have to be added to the integrated circuits in order to obtain a complete optical neuron. Bipolar transistors are suitable for this type of integration because, in addition to the inherent gain provided by the transistors, they can function as phototransistors, which detect light with high efficiency. Thus, the need for transistor amplifiers and detectors can be simultaneously satisfied by bipolar transistors. The remaining issue is the integration of these transistors with a light emitter. For reasons of reducing overall electrical power dissipation in an array of optoelectronic neurons, which is discussed in detail in Ch. 6, light-emitting diodes (LED's) are chosen as the light emitter. Monolithic integration of bipolar transistors and LED's present a problem in the material compatibility. The structure of a LED usually consists of an active material, such as GaAs, sandwiched between two higher bandgap cladding materials, such as AlGaAs, in a double-heterostructure fashion. The doping requirement for the LED is heavy p^+ doping in the top AlGaAs cladding layer, intrinsic doping in the GaAs active layer and heavy n^+ doping

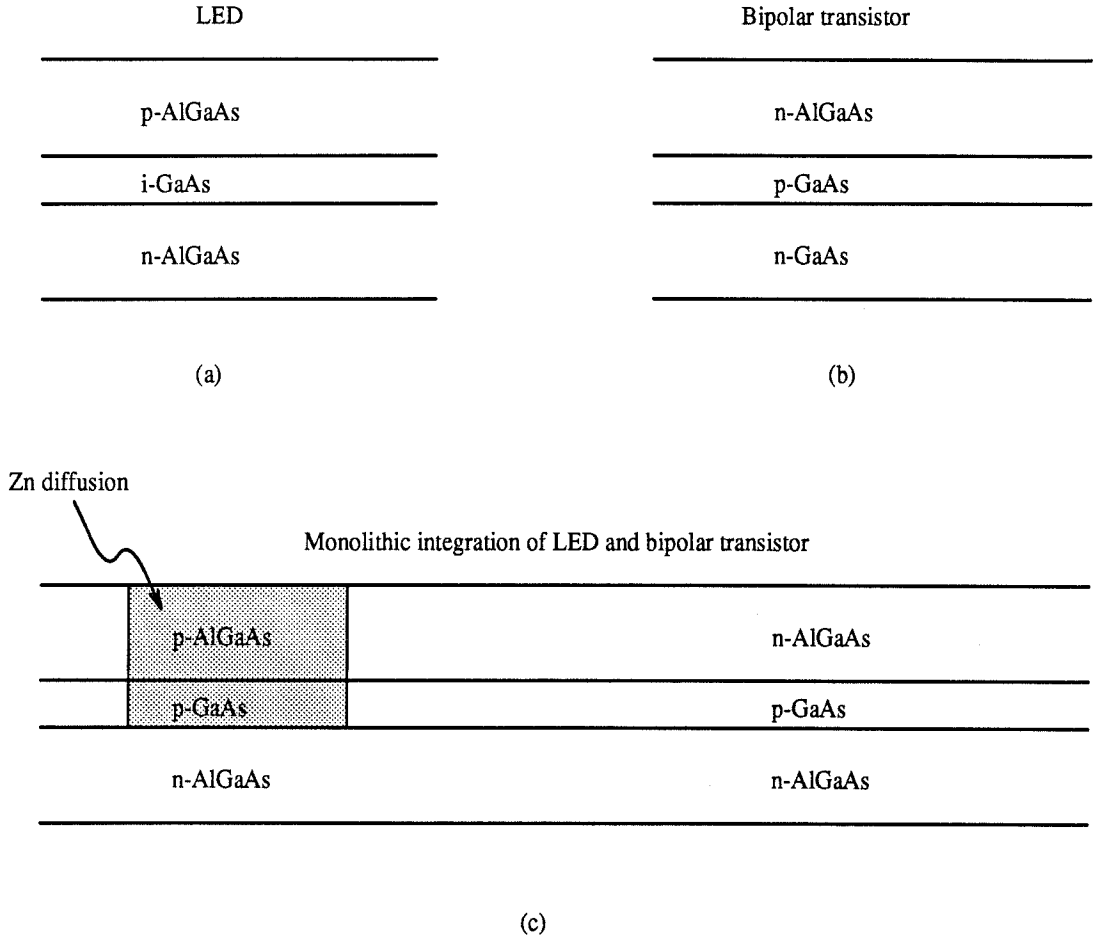


Fig. 4.1 Structure of typical epitaxial layers for **(a)** LED and **(b)** heterojunction bipolar transistor. By converting the collector to a higher bandgap material in the transistor and n-type upper cladding layer to p-type in the LED, both the LED and the bipolar transistor can be fabricated in the same epitaxial layer as shown in **(c)**.

in the bottom AlGaAs cladding layer, forming a P-i-N diode. This is illustrated in Fig. 4.1(a). For the bipolar transistor, however, the structure is somewhat different. The emitter is usually a high bandgap material, such as AlGaAs, and the base and the collector are low bandgap material, such as GaAs. The doping composition in a bipolar transistor is n-type, p-type and n-type in emitter, base, and collector respectively in order to utilize the high electron mobility in the base. This is illustrated in Fig. 4.1(b). As seen in the figure, the material and doping requirements for the LED and the bipolar transistor are somewhat different from each other. For the top layer, which is AlGaAs for both device, LED is doped p-type and the transistor is doped n-typed. For the second layer, which is GaAs in both devices, the LED is intrinsic and the transistor is p-type. For the last layer, both devices are doped n-type. However, the LED requires a high bandgap material whereas the transistor usually has a small bandgap material. In order to successfully integrate these two devices in a planar fashion, some compromise from each device has to be made. For example, the collector of the bipolar transistor does not have to be GaAs. It can be AlGaAs, which will make the transistor a double-heterojunction bipolar transistor (DHBT) and matches the lower cladding AlGaAs in the LED. In addition, the active layer of the LED can be doped p-type as long as it is sandwiched between two large bandgap materials. For the top layer, a compromise can not be made easily because each device requires a totally opposite doping composition from each other. Thus, a conversion from n- to p-type or vice versa has to be performed in order to obtain both devices simultaneously. While n-type diffusion is harder to perform and is not characterized as well, p-type diffusion can be performed in a very controlled manner on the LED to convert the originally n-type AlGaAs to the p-AlGaAs, which is required for the LED. This is graphically depicted in Fig. 4.1(c). In fact, Katz. et al. has experimentally demonstrated the feasibility of fabricating both the bipolar transistor and the LED from the same epitaxial material by performing Be

implantation to convert the n-type AlGaAs to p-AlGaAs [80]. This structure will form the basic material structure to build the DHBT-based optoelectronic neurons.

4.2 Design Considerations

The integration of LED's with heterojunction bipolar transistors presents a unique approach to realizing optoelectronic neurons needed for the neural network implementation. Shown in Fig. 4.2(a) is the schematic circuit diagram of the optoelectronic neuron that incorporates both devices. The bipolar transistor not only functions as the amplifier, but also as the photodetector. The thresholding is provided by applying a reverse biased current, I_{bb}, on the base of the transistor such that the transistor will not be turned on until the photogenerated current has exceeded the reverse biased current. After which, the transistor amplifies the signal received to produce an output current that drives the LED. This process continues until the transistor saturates, which causes the neuron to saturate as well. In order for this circuit to work in a neural network properly, several issues have to be addressed. Firstly, this optoelectronic neuron has to be able to provide sufficient optical gain in order for the signal to propagate to the next neuron without dying down. This implies that high current gain from the bipolar transistor is a requirement. If we assume η_H, η_D, η_L, and β are the efficiencies of the hologram that specifies the interconnections, the LED, the detector, and the current gain of the bipolar transistor respectively, then it is necessary to mandate the following relationship in order to close the loop without any attenuation :

$$\eta_H \cdot \eta_D \cdot \eta_L \cdot \beta \geq 1. \tag{4.1}$$

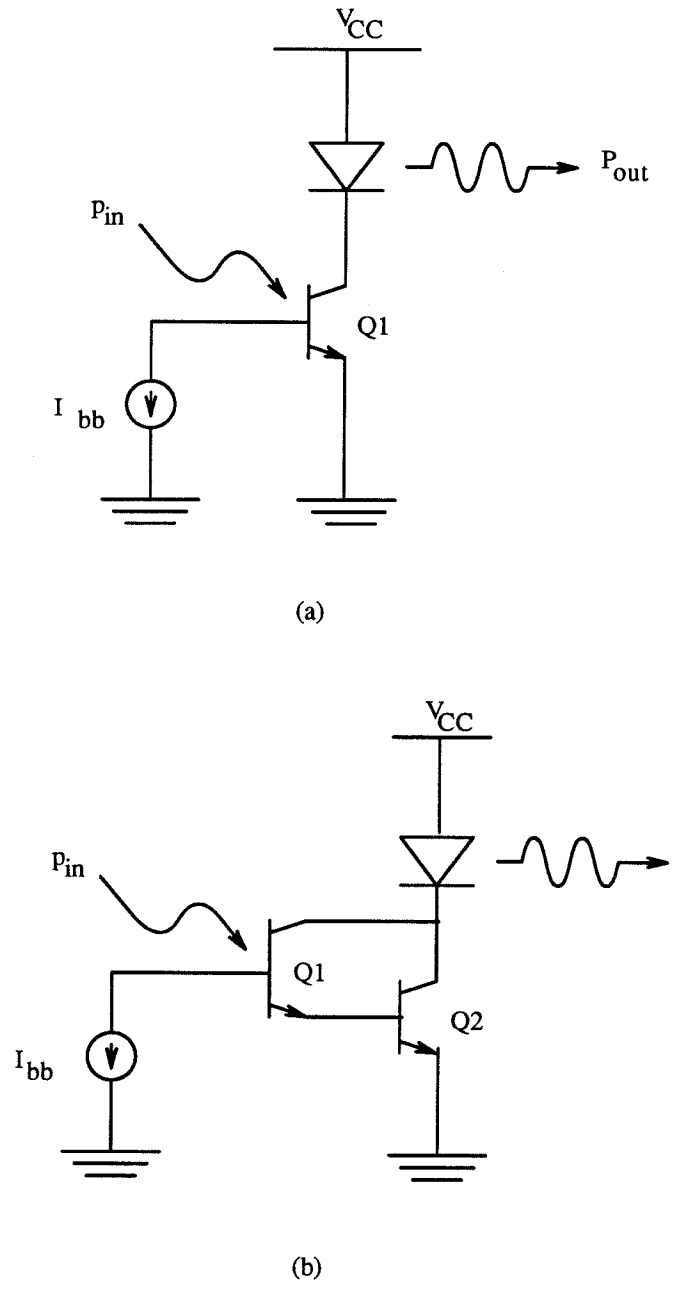


Fig. 4.2 **(a)** Schematic circuit diagram of an optoelectronic neuron incorporating only one bipolar transistor. **(b)** Schematic circuit diagram of an optoelectronic neuron incorporating two bipolar transistors to provide the gain needed to satisfy the loop gain requirement.

For $\eta_H = 0.1$, $\eta_D = 0.3A/W$, and $\eta_L = 0.01W/A$, β has to be at least 3333. Though a current gain of greater than 5000 has been reported in GaAs heterojunction bipolar transistors [66,81], it may be difficult to fabricate transistors that satisfy this gain reliably and consistently. Thus, two heterojunction bipolar transistors connected in a Darlington pair configuration has been proposed to meet the current gain requirement for the neuron and at the same time provide a reliable and practical way of fabricating the transistors. This is shown in Fig. 4.2(b). By using a Darlington pair, a combined current gain of 3333 can be obtained more easily as the product of the current gains from each transistor, $\beta_1 \cdot \beta_2$, need only be greater than 3333.

The second issue of concern is the ability of these bipolar transistors to provide gains at low driving power. It is well-known that the current gain of the transistor is dependent upon the collector current. In fact, the higher the current gain required, the higher the collector current needed, which, in turn, increases the power dissipation of the transistors. Approximately, the relationship can be expressed as

$$\beta \sim I_c^{1-\frac{1}{n}}, \tag{4.2}$$

where n is the ideality factor for the base-emitter junction, which ranges from 1 for the ideal junction to 2 for the non-ideal junction. The case of n = 2 corresponds to the situation in which the base current is dominated by recombination taking place through deep level traps in the space charge region. If the base-emitter junction is ideal, it can be seen from Eq. (4.2) that β is independent of the collector current. However, if the junction is not ideal, β's dependence on I_c can be as dramatic as square root. Thus, to achieve the high gain needed by the neuron, it is not surprising if the level of collector current needed is higher. To circumvent this problem, we need to decrease the ideality factor to as close to one as possible. In other words, the current component that contributes to the ideality factor of two

should be minimized. This can be achieved by designing the base-emitter junction such that the depletion region is narrow enough to disallow recombination within this region. Therefore, high β's can be attained at low collector currents.

The third issue is on the performance compromise of the LED and the transistor as a result of sharing the same epitaxial layers required for monolithic integration. This compromise arises from the fact that the p-GaAs is shared by all three devices. For the photodetector, this p-GaAs is the absorption layer, which needs to be thick to allow for the complete absorption of incoming photons. However, for the transistor, this layer is the base, which should be as thin as possible to maximize the current gain. Similarly, for the LED, it can not be too thick or too thin due to self-reabsorption and interfacial recombination as discussed in Ch. 2. Thus, the thickness of the p-GaAs layer needs to be carefully chosen so that the overall performance of the neuron, not each individual device, is maximized.

To quantify the parameters of the transistors before monolithically integrating them with the LED on the same substrate, individual transistors were first characterized to ensure the β's measured was sufficient for the Darlington transistor pair to provide the current gain needed for the neurons.

4.3 Discrete Double-Heterojunction Bipolar Transistors

GaAlAs/GaAs/GaAlAs double heterojunction bipolar transistors (DHBT's) are very attractive for high-gain applications and optoelectronic monolithic integration because of their structural compatibility with laser diodes [80,82] and LED's. Very high current gain ($\beta \sim 10^4$) has been demonstrated in single heterojunction bipolar transistors (SHBT's) grown by liquid phase epitaxy [81,49]. However, most of the SHBT's grown by molecular beam epitaxy (MBE) [38] and MOCVD [83] show much lower gains. The current gain is even lower in DHBT's [66,84]. In spite of

some encouraging results [33,63] and recent progress in the crystal growth by MBE and MOCVD, the reproducibility of high-gain heterojunction bipolar transistors, especially DHBT's, is not sufficiently good mainly because the heavy base doping concentration incorporated in these transistors has resulted in an out-diffusion of these base dopants during crystal growth or subsequent high-temperature processing. Consequently, the heterojunction is displaced and its integrity is degraded. Secondary ion mass spectroscopy (SIMS) [74] has shown that even with a spacer layer of 100Å sandwiched between the GaAs and the GaAlAs layer, Be has been found to diffuse into the AlGaAs layer at a growth temperature of 720° C, whereas no diffusion is observed in the AlGaAs layer at a growth temperature of 650° C. Since the diffusion coefficient of Be is lower than that of Zn, an even lower growth temperature is preferable to prevent the out-diffusion of Zn. However, a higher growth temperature is necessary to suppress nonradiative recombination centers in achieving a high crystal quality [81]. Therefore, in addition to the insertion of undoped GaAs spacer layers at the emitter-base and the collector-base junctions, the base doping level must be reduced in order to eliminate the problem of Zn out-diffusion if nonradiative recombination centers are to be suppressed at the same time. This assures that the p-n junction is properly placed at the heterointerface and prevents intermixing of GaAlAs-GaAs. Furthermore, the reduction in the base doping will increase the maximum current gain achievable in DHBT's [32]. Thus, the advantage is two-fold.

The double heterojunction structure was epitaxially grown on (100) Cr-doped semi-insulating GaAs substrates ($\rho \geq 5\times10^7$ ohm·cm) by a metalorganic chemical vapor deposition system (SPIRE-450) using a vertical barrel reactor. The GaAs and AlGaAs layers were grown by trimethyl gallium (TMG), trimethyl aluminum (TMA), and 10% AsH_3 in 90% H_2. Zinc and silicon were used for p- and n-type dopants, respectively. The substrate temperature during the growth was about

730° C. DHBT's structure consists of : 0.5 μm of Si-doped (10^{18} cm^{-3}) n-GaAs subcollector/buffer, 1.2 μm of Si-doped (1.6×10^{17} cm^{-3}) n-$Al_{0.3}Ga_{0.7}As$ collector, 100$\AA$ of undoped GaAs spacer layer, 0.15 μm of Zn-doped (2×10^{17} cm^{-3}) p-GaAs base, 100 $\AA$ of GaAs undoped spacer layer, 1.1 μm of Si-doped (4.2×10^{17} cm^{-3}) n-$Al_{0.3}Ga_{0.7}As$ emitter, and 0.23 μm of Si-doped (1.5×10^{18} cm^{-3}) n-GaAs cap layer. The doping profile and thickness of each DHBT layer were plotted by a Polaron PN-4200 electrochemical profiler as shown in Fig. 4.3. It should be noted that the doping concentration of each layer is extremely uniform.

The structure of the DHBT is schematically shown in Fig. 4.4. The devices with an emitter area of 2.4x10^{-4} cm^2 were fabricated by standard lift-off and wet chemical etching processes. The base was properly exposed by first etching the GaAs cap layer in H_3PO_4 + H_2O_2 + CH_3COOH, followed by etching the $Al_{0.3}Ga_{0.7}As$ emitter in NH_4OH + H_2O_2 + H_2O. H_3PO_4 + H_2O_2 + CH_3COOH was used to etch epilayers down to the subcollector layer for collector contacts. AuGe/Au and AuZn/Au were evaporated for emitter/collector and base contacts, respectively and alloyed separately.

Typical common-emitter current characteristics are shown in Fig. 4.5 at several different current levels. Current gains of 40, 100, 300, and 500, which was the highest current gain reported for MOCVD-grown DHBT's without base or junction grading then, were obtained at collector currents of 0.2, 10, 70, and 120 mA, respectively as shown in Fig. 4.5(a), (b), (c), and (d). The collector current density at which the current gain of 500 was obtained was 500 A/cm^2 based on the base-emitter junction area of 2.4x10^{-4} cm^2. Figure 5(d) also shows the inverted-mode DHBT characteristics with a current gain of 10 at an emitter current of 2 mA. The forward I-V characteristics showed an offset voltage of 0.3 V. However, no offset voltage was observed in the inverted mode. Reverse breakdown voltages of 8 and 10 volts were observed for the emitter-base and the collector-base junctions, respectively. The

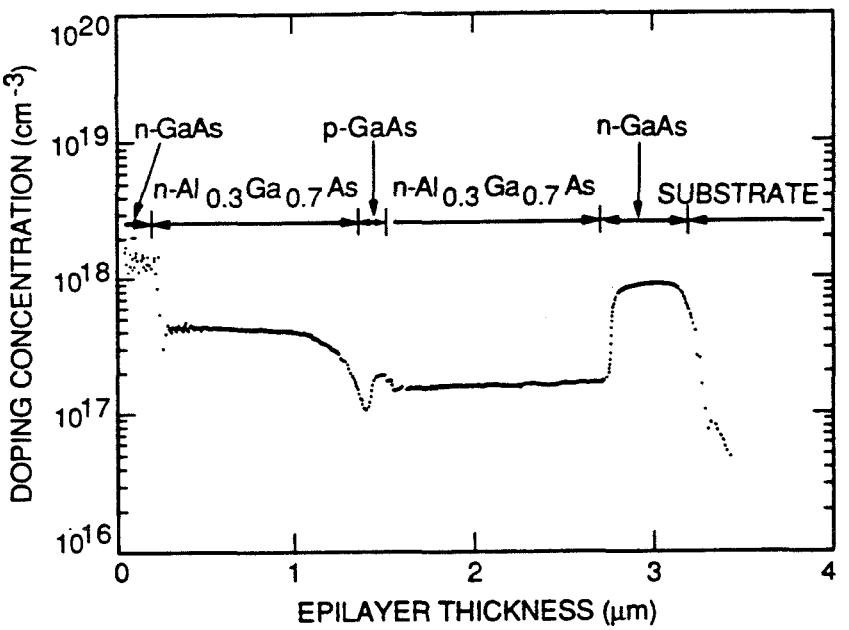


Fig. 4.3 The doping profile and the thickness of each epilayer measured by a Polaron electrochemical profiler.

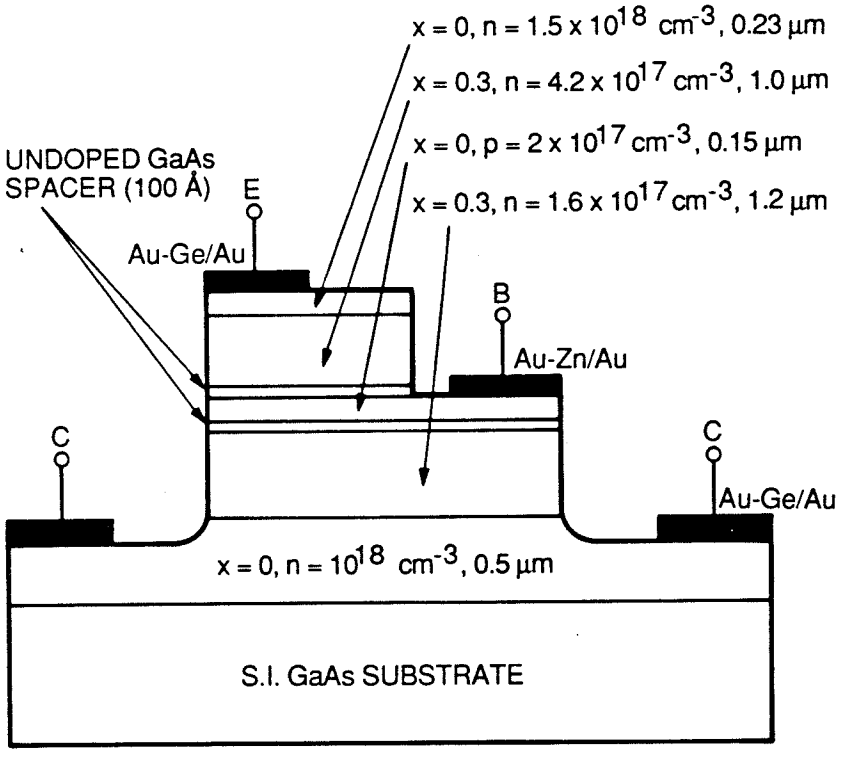


Fig. 4.4 The schematic cross sectional view of the $Al_xGa_{1-x}As/GaAs/Al_xGa_{1-x}As$ double heterojunction bipolar transistor.

collector-emitter breakdown voltage (BV_{CEO}) in the common-emitter configuration was 5 V.

Figure 4.6 shows the logarithmic plot of the measured common-emitter current gain as a function of the collector current. The ideality factor evaluated from the relation of $\beta \sim I_c^{1-\frac{1}{n}}$ was approximately 1.4. This value indicates that the recombination current in the emitter-base junction depletion region is not negligible at small collector currents. It should be noted however that the current gain increases continuously with increasing collector current. This means that no serious base push-out or emitter crowding effect exists. In this device operating region, a maximum current gain of 500 was obtained. We observed a higher current gain of 750 at a higher collector current. However, DHBT's operated at this current level were not thermally stable. These high current gains are evidence of effective blockage of Zn out-diffusion as a result of the reduced base doping concentration and the insertion of the undoped GaAs spacer layers. However, these inserted undoped spacer layers may be responsible for not having a smaller base-emitter junction ideality factor of 1.4 because of the recombination current taking place within these depleted regions.

By hooking up two of these transistors in a Darlington fashion, a combined current gain of 4000 has been measured. This is shown in Fig. 7. Because of the electrical contact problem, a large offset voltage in V_{CE} was observed. Nevertheless, it showed the feasibility of achieving the current gain required for the integrated optoelectronic neuron by using a Darlington transistor pair.

4.4 Monolithically Integrated Optoelectronic Neurons

With the current gain of 500 demonstrated in the discrete double-heterojunction bipolar transistors, monolithically integrated optoelectronic neurons consisting of two double-heterojunction bipolar transistors with a LED on a common

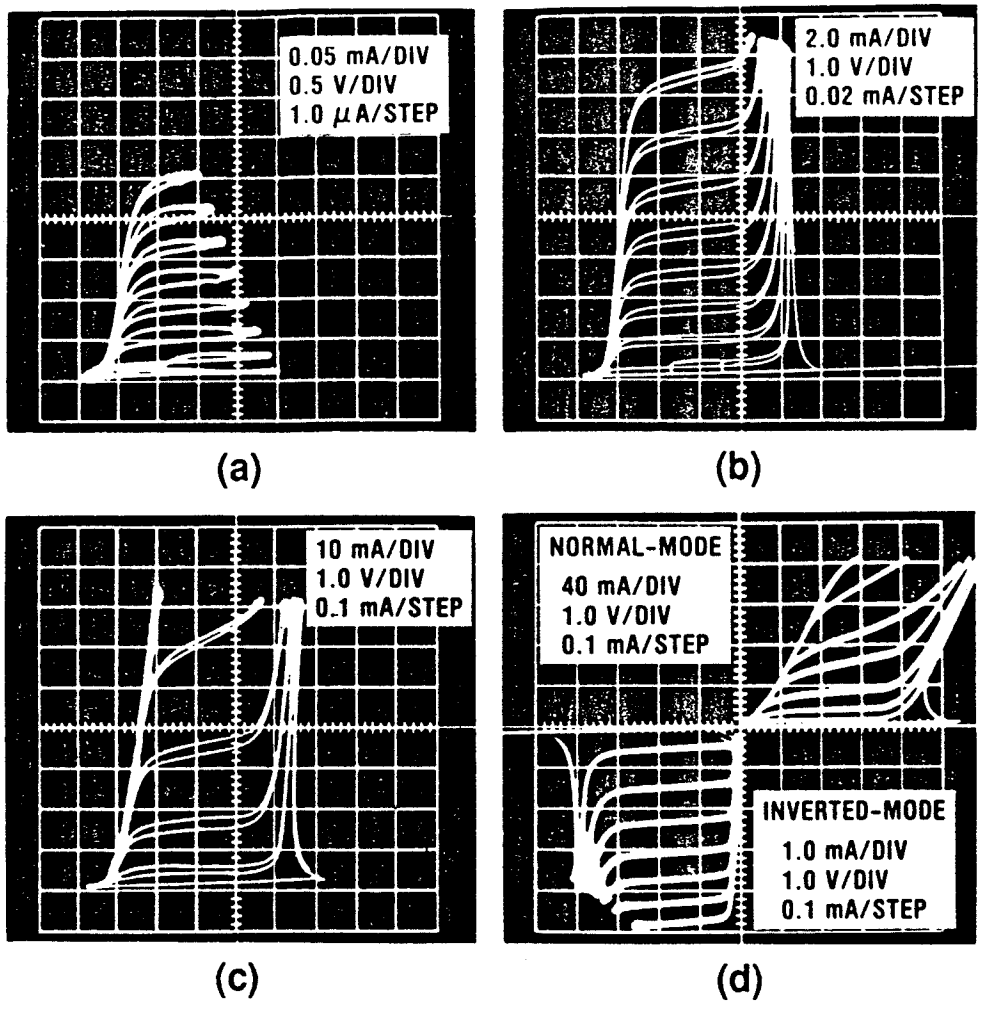


Fig. 4.5 Typical common-emitter I-V characteristics of the DHBT at: (a) low current levels; (b) normal current levels; (c) high current levels; (d) normal and inverted modes.

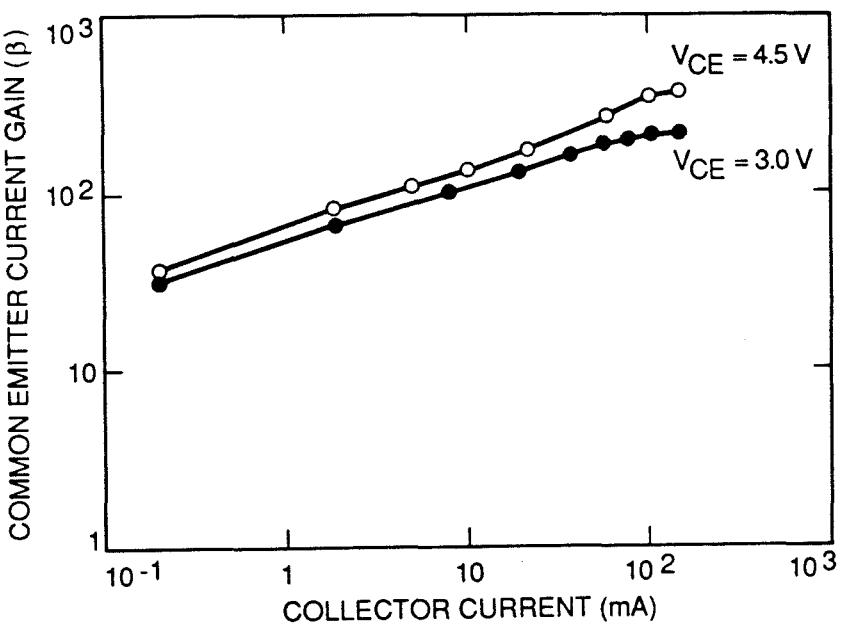


Fig. 4.6 The common-emitter current gain as a function of the collector current at $V_{CE} = 3$ and 4.5 V.

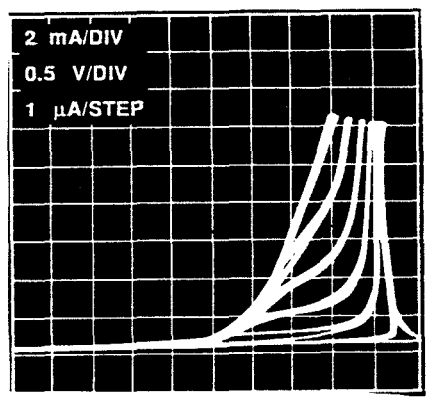


Fig. 4.7 The common-emitter I-V characteristics of two discrete DHBTs' connected a Darlington pair fashion. A combined current gain of 4000 has been measured. An offset voltage of 2.5 V in V_{CE} was observed because of the electrical contact problem of the probes.

GaAs substrate as shown in the circuit in Fig. 4.2(b) are next fabricated. Since the design and material parameters for the discrete transistors are approximately the same as those of the integrated Darlington transistor pair, the integrated Darlington transistor pair is expected to exhibit the same high current gain as that observed in discrete transistors. However, a minor difference exists between the discrete and the integrated transistor. In the discrete transistor, base contacts are defined by etching down to expose the base, followed by evaporation of proper metals. For the integrated transistor, Zn-diffusion is used to facilitate the making of the contact to the base of the transistor. This avoids the need for etching down to expose the base, which is a very sensitive and delicate process. This is because the base layer is so thin that it is very easy to over-etch it. This Zn-diffused bipolar transistor has been experimentally described and analyzed earlier in Ch. 3. While Zn-diffusion is performed to make contact to the base of the transistor, it also serves as a necessary step in converting the n-AlGaAs upper cladding layer to p-AlGaAs, thus forming a P-i-N diode for the LED. In fact, the reason why Zn-diffusion is chosen to make contact to the base of the transistor is because the n-AlGaAs upper cladding layer needs to be Zn-diffused in converting to p-type AlGaAs anyway. Thus, while this Zn-diffusion is necessary for formation of the LED, it also facilitates making the contact to the base of the transistor so that one extra step in the processing of this integrated optoelectronic neuron can be eliminated.

Figure 4.8 shows the cross sectional view of the optoelectronic neuron. The structure of the epitaxial layers is the same as that for the discrete transistors and is described earlier in Sec. 4.3. Figure 4.9(a) through 4.9(c) illustrate the step-by-step process in fabricating the optoelectronic neurons. Following the standard post-growth wafer cleaning procedure, each neuron was first photolithographically defined by etching the epilayers into the semi-insulating substrate with a nonselective etchant, $H_3PO_4 + H_2O_2 + CH_3COOH$. Each individual device in a neuron was

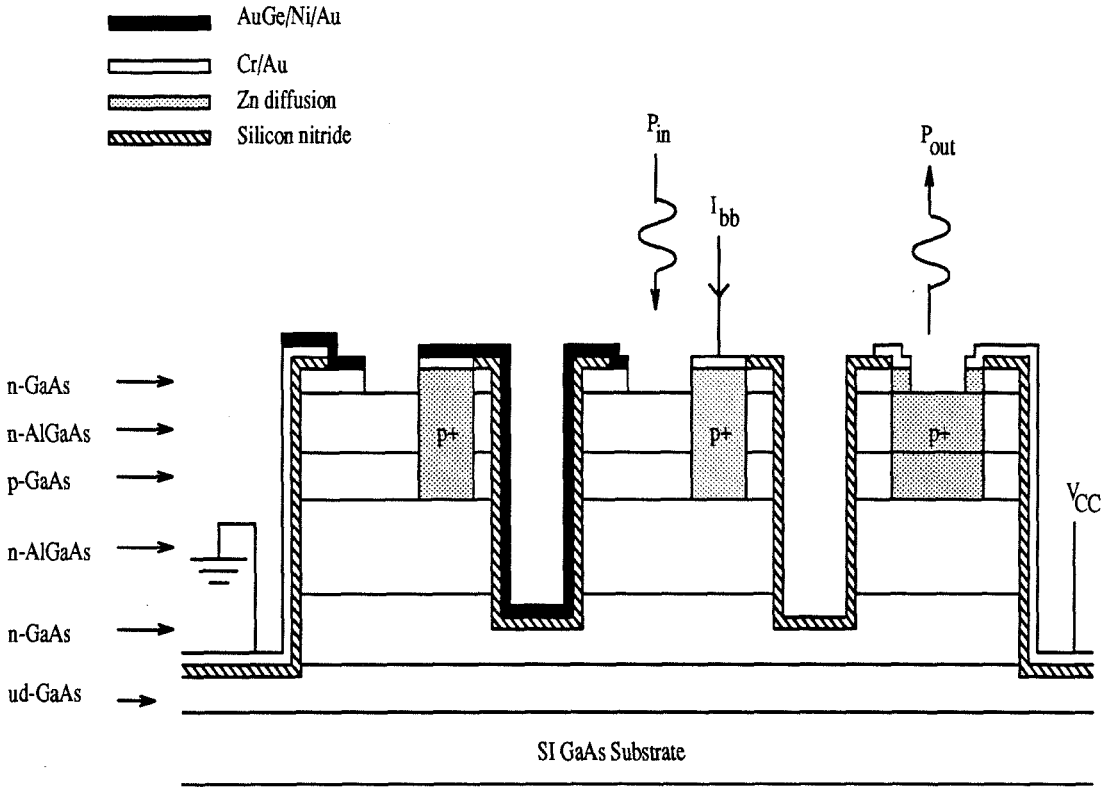


Fig. 4.8 Cross sectional view of the monolithically integrated optoelectronic neuron that is consisted of two Zn-diffused double-heterojunction bipolar transistors, which form a Darlington transistor pair, and a LED.

subsequently defined by etching the epilayers into the AlGaAs collector layer using the same etchant so that the DHPT, DHBT and LED were mutually connected by the n-type GaAs sub-collector layer only. Zn-diffusion at 650°C for 45 minutes was then performed in a sealed ampoule using $ZnAs_2$ as the source in order to provide external base contacts for the Darlington transistor pair as well as to convert the n-type cap and emitter layers to p-type for the LED. The mask used for diffusion was Si_3N_4 grown at 680°C by a thermal chemical vapor deposition system. The diffusion regions were defined by etching the Si_3N_4 mask in a CF_4 plasma. Following Zn-diffusion, the photosensitive area of the DHPT was opened by removing Si_3N_4. Cr/Au was evaporated and lifted off for both the p-type ohmic contacts and the interconnection lines. AuGe/Au was evaporated and lifted off for the n-type ohmic contacts and alloyed subsequently at 380° for 1 minute. Photographs of a fabricated neuron is shown in Fig. 10. Each array has dimensions of 5×5 mm^2 and each neuron has dimensions of 250×250 μm^2. The light-emitting area of the LED is 8×8 μm^2 and the light-detecting area of the DHPT is 50×130 μm^2. The emitter areas for the DHBT and the DHPT are 3.5×10^{-5} cm^2 and 1.6×10^{-4} cm^2 respectively.

A 10×10 array of the optoelectronic neurons was also fabricated by using the same process procedure and its picture is shown in Fig. 4.11. It has dimensions of 5mm × 5mm with 40 bond pads surrounding the array. These 100 neurons were grouped in a certain fashion so that some of the neurons were not electrically connected. This was purposely designed so as to avoid the "host" image which would have been created when the output of one neuron was diffracted by the grating intended for its neighboring neuron.

When the integrated optoelectronic neuron was tested, semiconductor controlled rectifier (SCR) characteristics were observed as shown in Fig. 4.12. A forward breakdown voltage of 75 V, a forward holding voltage of 25 V and a reverse

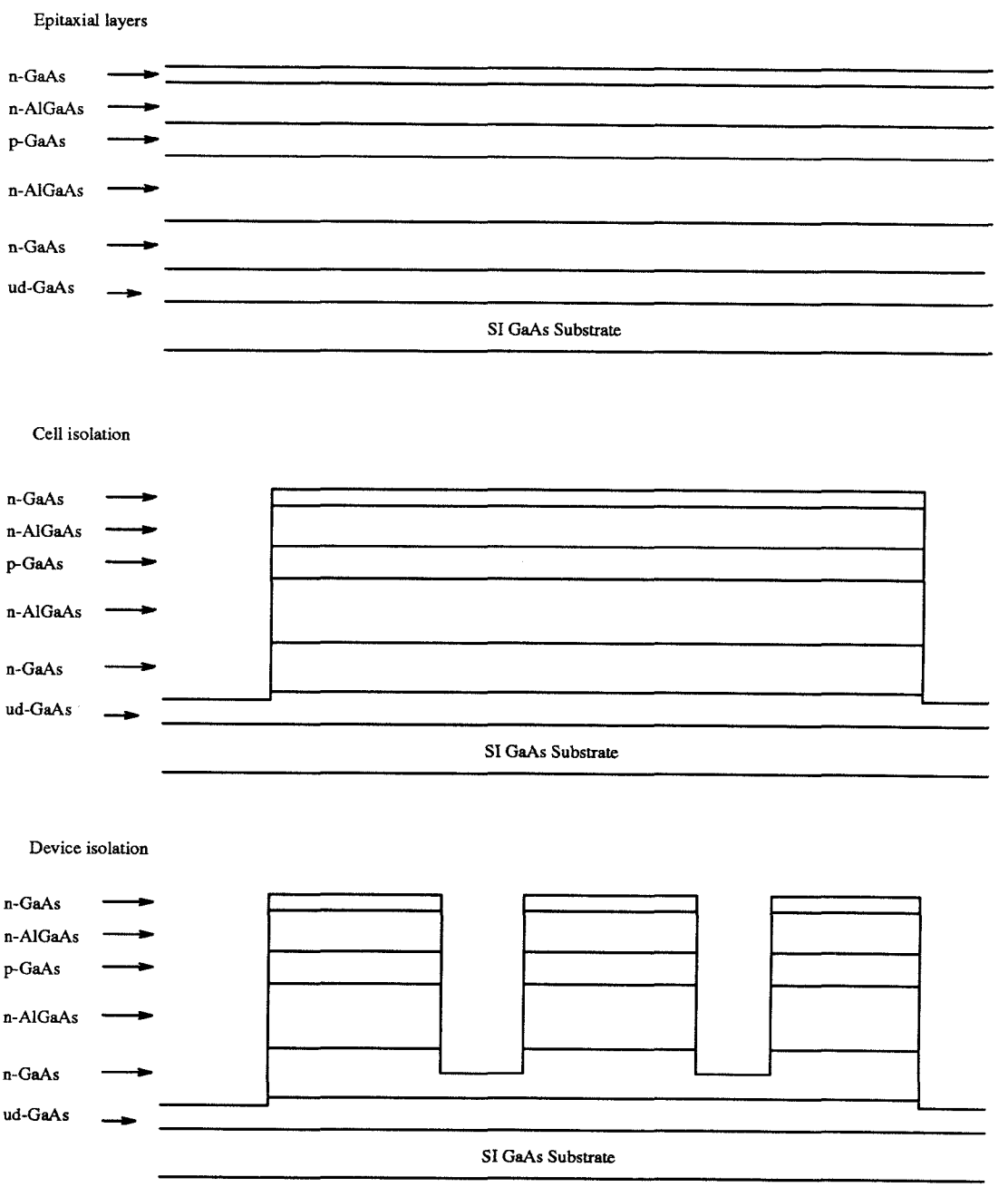


Fig. 4.9(a) Fabrication process of the optoelectronic neuron, illustrating the cell isolation and the device isolation.

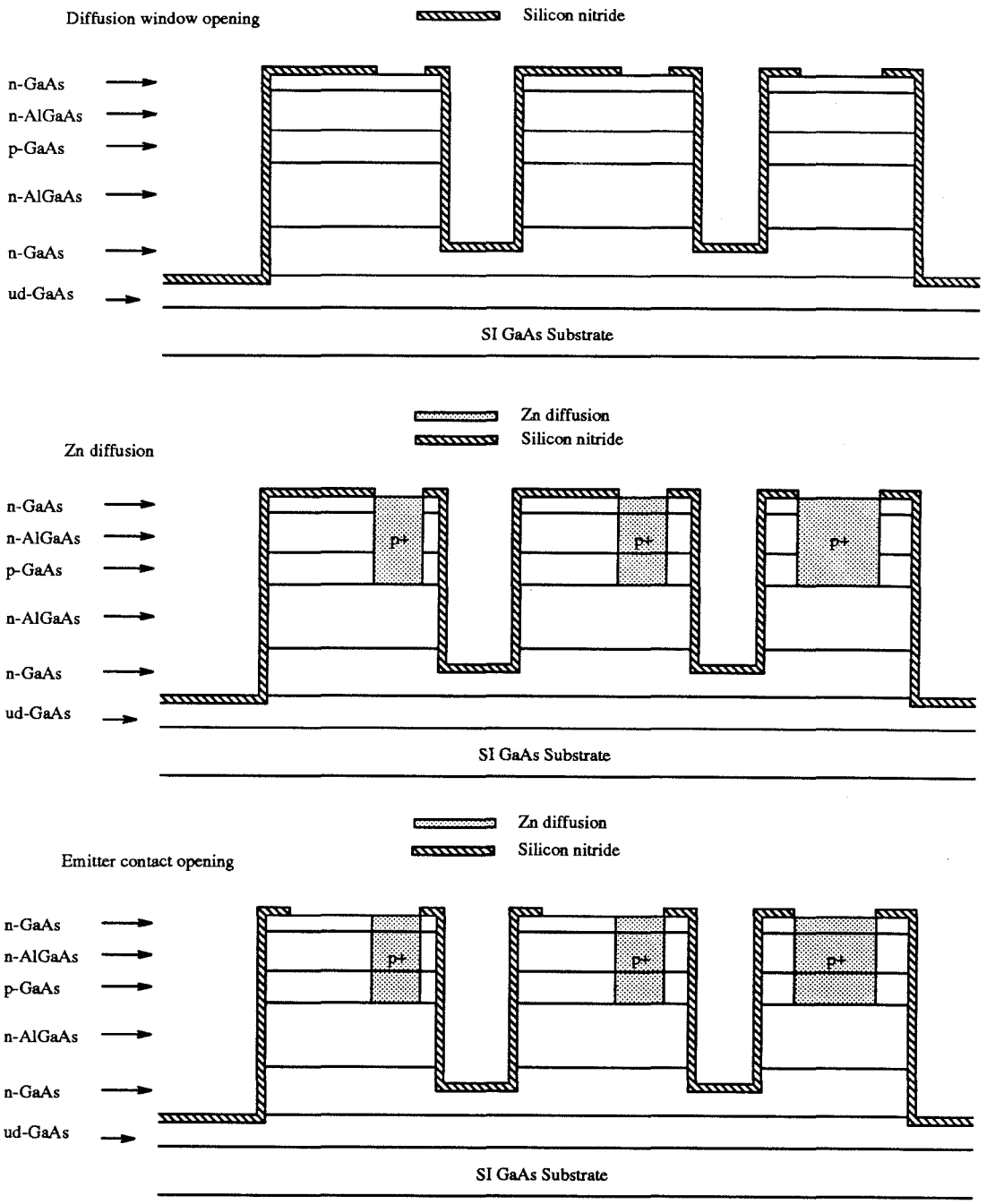


Fig. 4.9(b) Fabrication process of the optoelectronic neuron, illustrating the process of Zn-diffusion and emitter contact opening.

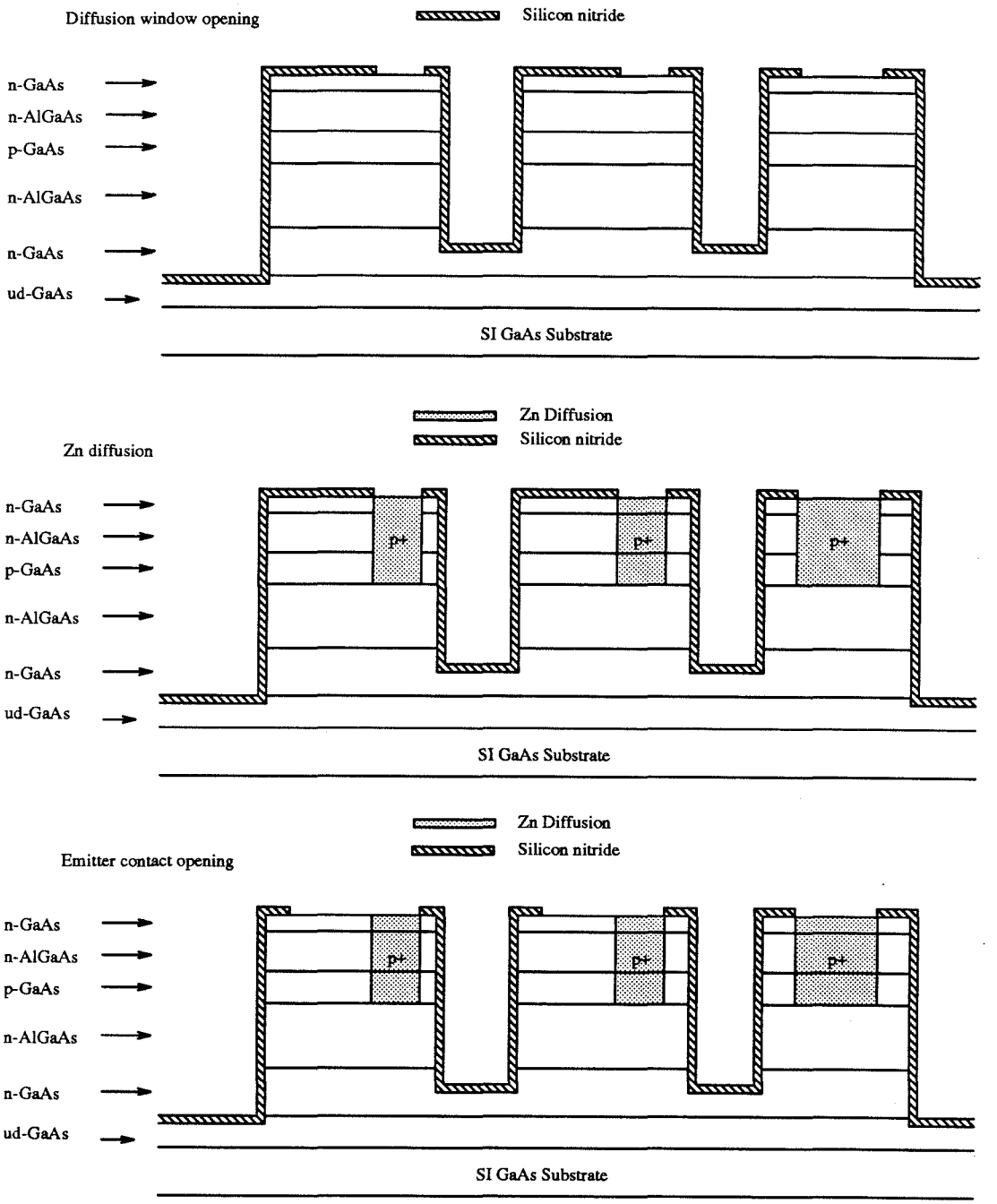


Fig. 4.9(c) Fabrication process of the optoelectronic neuron, illustrating the process of p-type and n-type metalizations.

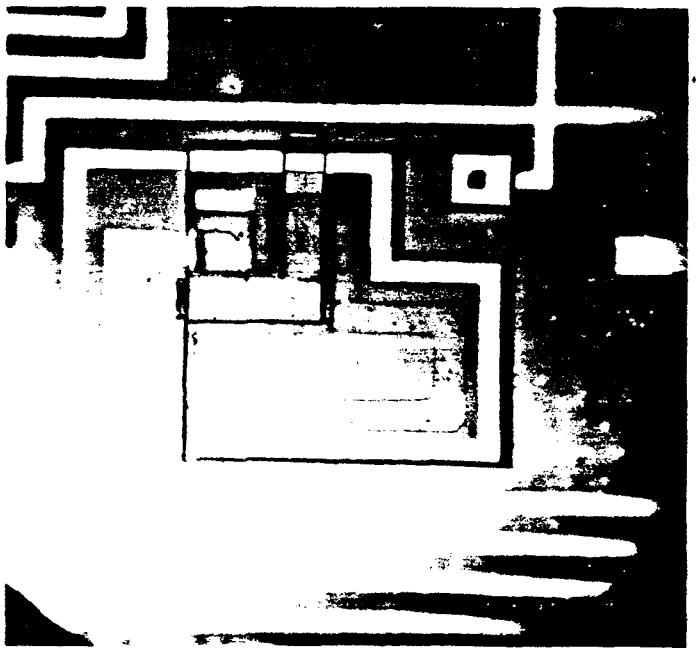


Fig. 4.10 Photograph of a fabricated optoelectronic neuron, consisted of two Zn-diffused double-heterojunction bipolar transistors and a LED.

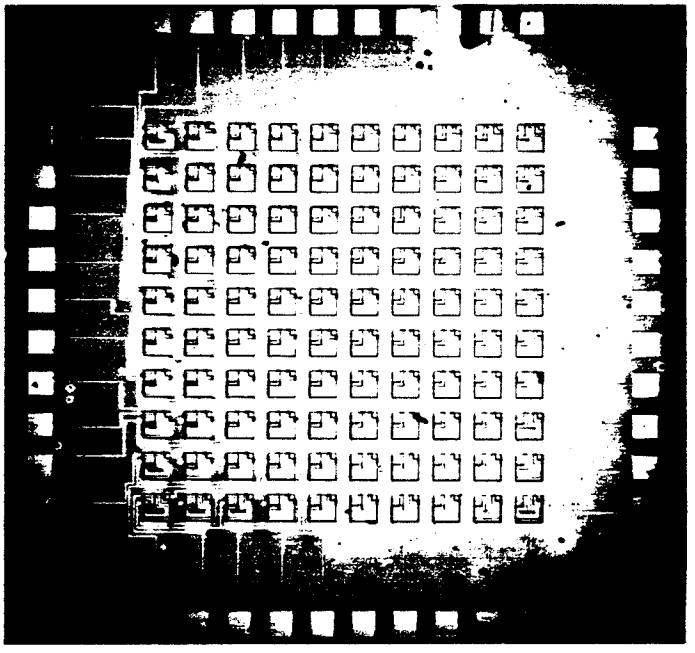


Fig. 4.11 Photograph of a fabricated 10×10 array of optoelectronic neurons with each element in the array consisted of the integrated circuits shown in Fig. 4.10.

breakdown voltage of 60 V were measured. By either increasing the base current or the external illumination on the neuron, the forward breakdown decreased. The LED was observed to emit light in the forward breakdown mode, implying the carriers were recombing in the low bandgap GaAs layer.

Careful inspection of the integrated LED with Darlington transistor pair revealed that the SCR was present in the device due to the parasitic p-n-p transistor coupled to the n-p-n DHBT. The anode of the SCR was the Zn-diffused area in the original LED region and the cathode was the original emitter (ground). This parasitic p-n-p transistor existed because the LED and the Darlington transistor pair shared the same collector. The effective base width of this parasitic p-n-p transistor was at least the separation between the LED and the bipolar transistor, which was 20 μm.

In order to better understand the latch-up of the parasitic p-n-p transistor, it is necessary to first understand the requirement for latching. The basic structure of a SCR consists of four alternating p-n-p-n epitaxial layers [85] as shown in Fig. 4.13(a). It can be modeled as a n-p-n transistor coupled to a p-n-p transistor. This is illustrated in Fig. 4.13(b), in which the collector current from the p-n-p transistor (transistor 1) becomes the base current to the n-p-n transistor (transistor 2) and the base current of the p-n-p transistor partially becomes the collector current of the n-p-n transistor. If we consider the leakage current across the base-collector junction and label them to be I_{co1} and I_{co2} for transistor 1 and 2 respectively, we have from transistor 1 :

$$I_{b1} = (1 - \alpha_1)I_A - I_{co1}, \tag{4.3}$$

where α_1 is the common-base current gain of the p-n-p transistor. This current is also the collector current of the n-p-n transistor, which can be expressed by

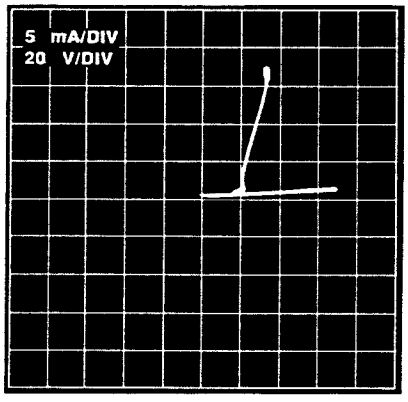


Fig. 4.12 I-V characteristics of the integrated optoelectronic neurons, exhibiting the behavior of a semiconductor controlled rectifier (SCR).

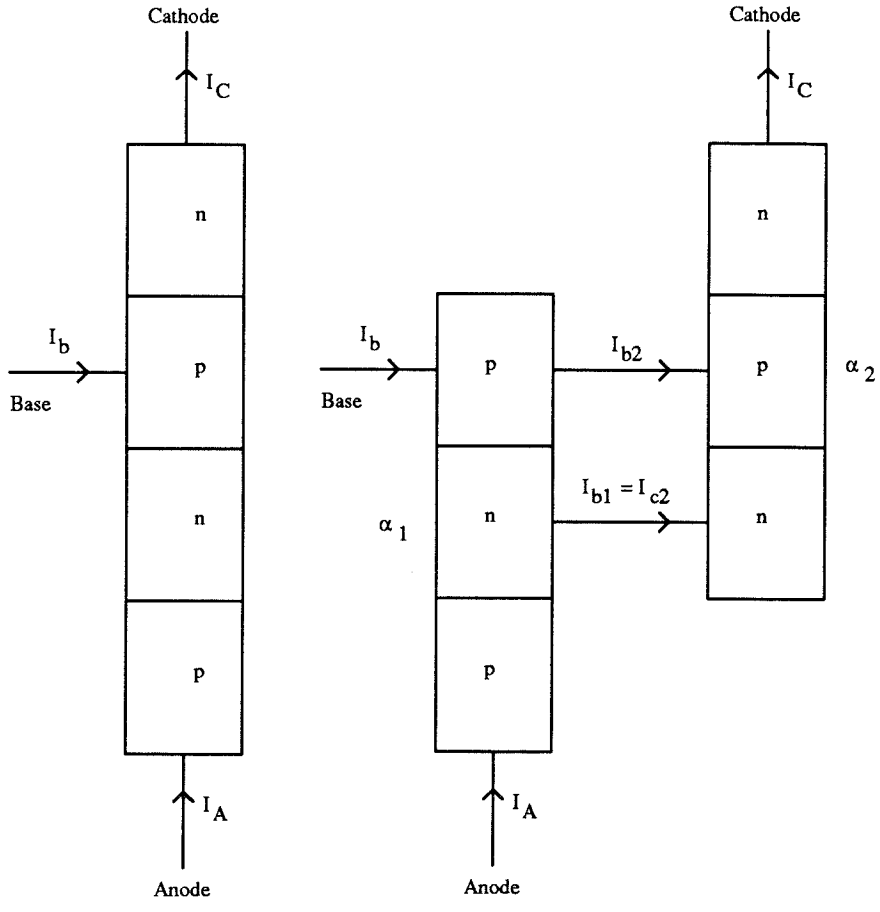


Fig. 4.13 **(a)** The structure of a semiconductor controlled rectifier, consisting of alternating p-n-p-n layers. **(b)** The device model of an SCR, illustrating that a SCR can be modeled as being composed of a p-n-p and n-p-n transistor connected in a fashion as shown in the figure.

$$I_{c2} = \alpha_2 I_C + I_{co2}. \tag{4.4}$$

Since $I_C = I_b + I_A$, Eq. (4.4) can be re-written as

$$I_{c2} = \alpha_2(I_b + I_A) + I_{co2}. \tag{4.5}$$

Equating Eq. (4.5) and (4.3) and solving for I_A, we obtain

$$I_A = \frac{\alpha_2 I_b + I_{co1} + I_{co2}}{1 - \alpha_1 - \alpha_2}. \tag{4.6}$$

Therefore, the current through a SCR is relatively small until the condition $\alpha_1 + \alpha_2 = 1$ is reached. At this point, the current increases to infinity which, in turn, causes the device to switch to a forward conduction mode. Before this switching occurs, α_1 and α_2 increase monotonically with I_A. However, increasing α_1 and α_2 also increases I_A through Eq. (4.6). As a result, a positive feedback takes place and quickly causes the SCR to switch into the forward conduction mode. It should be noted that all current components in the numerator of Eq. (4.6) are small. Thus, I_A is small before breakdown occurs. It is also interesting to note that increasing I_b or illuminating light on the the SCR will add positively to the feedback process and causes the SCR to switch at a lower voltage. This is because applying I_b and illumination of light increase the overall current flowing through the SCR and as a result, α_1 and α_2 increase.

Having understood the switching condition of a SCR is that the sum of the common-base current gains from the two n-p-n and p-n-p transistors, which makes up the structure of a SCR, is equal to one, it is not difficult to see that in our

current optoelectronic neuron, the sum of α_{npn} and α_{pnp} must be equal to 1, where α_{npn} and α_{pnp} are the common base gains of the designed npn transistor and pnp parasitic transistor respectively. Since the effective base width of the parasitic p-n-p transistor was 20 μm, the existence of this transistor would have normally been negligible, compared to the designed transistor, which had a base width of only 0.15 μm. The fact that the latching occurred suggested that this parasitic p-n-p transistor contributed to a non-negligible α_{pnp} and subsequently caused the switching. This switching occurred despite the fact that this p-n-p transistor had a wide-bandgap base and small-bandgap emitter and collector, which were counter-productive in maximizing the current gain of the transistor. This problem could be eliminated by degrading the gain of the parasitic p-n-p transistor further through further separating the LED from the transistors. This method is not feasible because not only would it not guarantee the complete removal of the parasitic p-n-p transistor, it also would consume a larger chip area. Thus, a more reliable and efficient method would be to electrically isolate the LED from the Darlington transistor pair and then employ metalization to connect them as required. This method was pursued in our later version of the optical neurons. The process involved was to perform an additional isolation etch between the LED and the transistors down into the semi-insulting GaAs substrate, followed by evaporation of n-type metalization to appropriately connect the LED and the transistors up. The device cross sectional view after the remedial process was applied is shown in Fig. 4.14.

By etching into the semi-insulating substrate and employing metalization to connect the LED to the transistors, the optoelectronic neuron showed the correct I-V characteristics as shown in Fig. 4.15, which shows the common-emitter I-V characteristics for the Darlington transistor pair monolithically integrated to the LED. Even though the combined current gain was measured to be 2 at best, it exhibited the proper transistor characteristics offset by a voltage caused by the

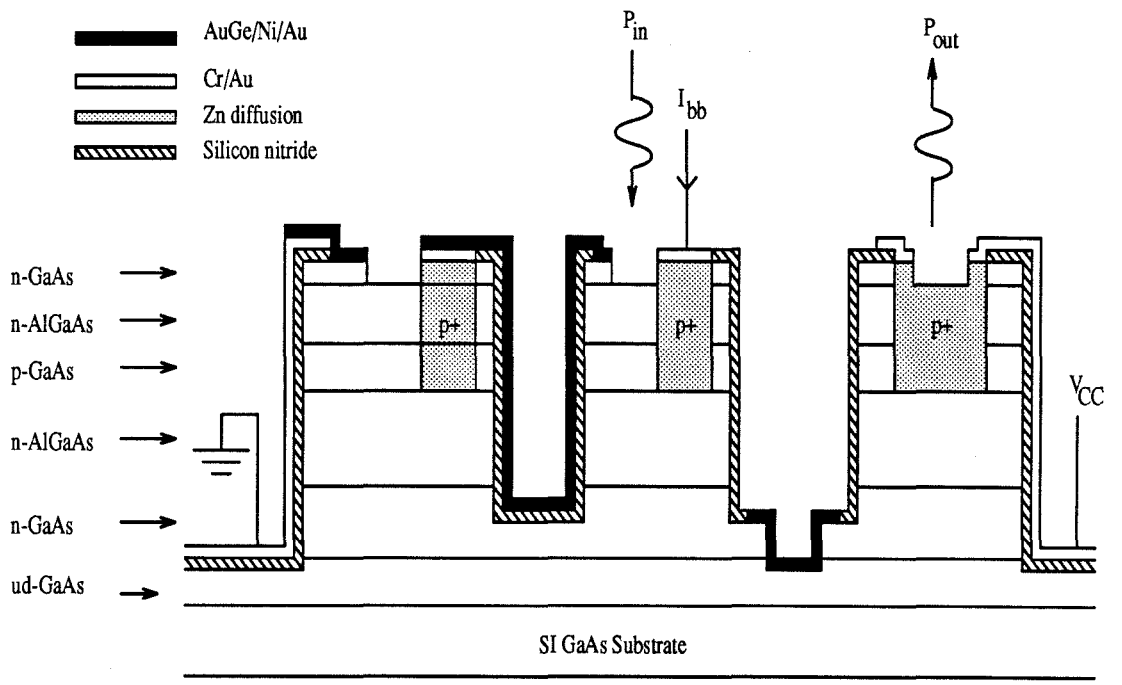


Fig. 4.14 The cross sectional view of the optoelectronic neuron after the the LED has been isolated from the transistors by an etch into the substrate and subsequently connected to the transistors by metalization.

turn-on voltage of the LED. Thus, the origin of thyristor latching was successfully verified to be due to the parasitic coupling of the p-n-p transistor. This has an important consequence on the design of the integration. Namely, bipolar devices, such as bipolar transistors, LED's, and lasers, should be totally isolated from one another when integrating these devices on a common substrate. The integrated laser and bipolar transistor on an n^+ GaAs substrate as demonstrated by Katz et al. [80] might eventually be limited by the thyristor latching.

The low current gain measured from the Darlington transistor pair suggested that base leakage current dominated the current transport in the base, which led to inefficient electron injection from the emitter. As a result, most of the base current was recombining either through the surface or inside the depletion region. To prevent carriers from recombining in these regions, the method of etching down to the depleted AlGaAs in area between the base and the emitter of the transistor has been introduced and analyzed in Ch. 3. Figure 4.16 shows the region of the Darlington transistor pair to be etched down. Since the region to be etched was the only region that was exposed to air, no extra mask was needed as the etching was done in a self-aligned manner. By applying this technique to the current monolithically integrated Darlington transistor pair, a plot of the combined current gain vs. the etch depth can be obtained. This is shown in Fig. 4.17.

It is worthwhile to note that, from Fig. 4.17, the dramatic improvement obtained in the current gain as the isolation etch depth increased was similar to the improvements obtained in individual discrete transistor as discussed in Ch. 3 and suggested the same mechanism by which the base current was transported in the transistor. Before the etch, the current gain was only measured to about 10. This was an indication that the majority of the base current was flowing through the n^+ GaAs cap layer. As the etch front penetrated into the AlGaAs emitter layer, the current gain improved steadily until the remaining AlGaAs layer was totally

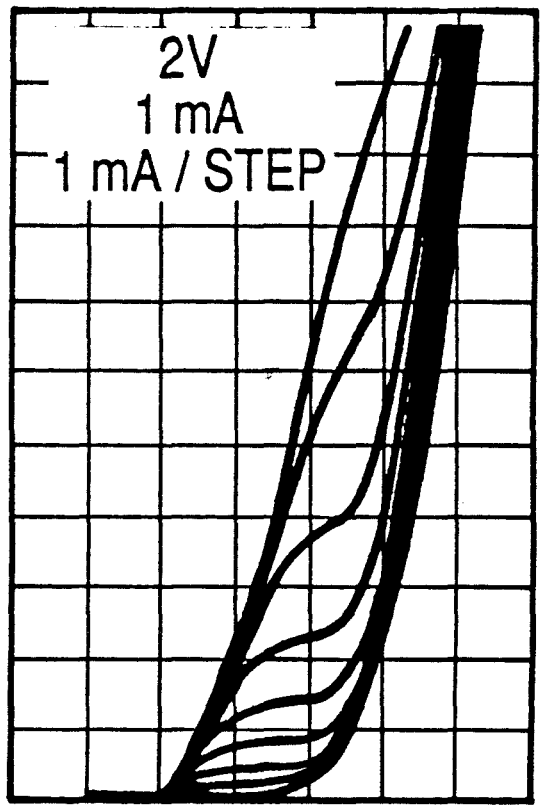


Fig. 4.15 I-V characteristics of the monolithically integrated Darlington transistors and LED after the LED has been electrically isolated from the transistors. The offset voltage in the V_{CE} of 2 V was due to the combination of the turn-on voltage of the LED and the intrinsic offset voltage in the transistors.

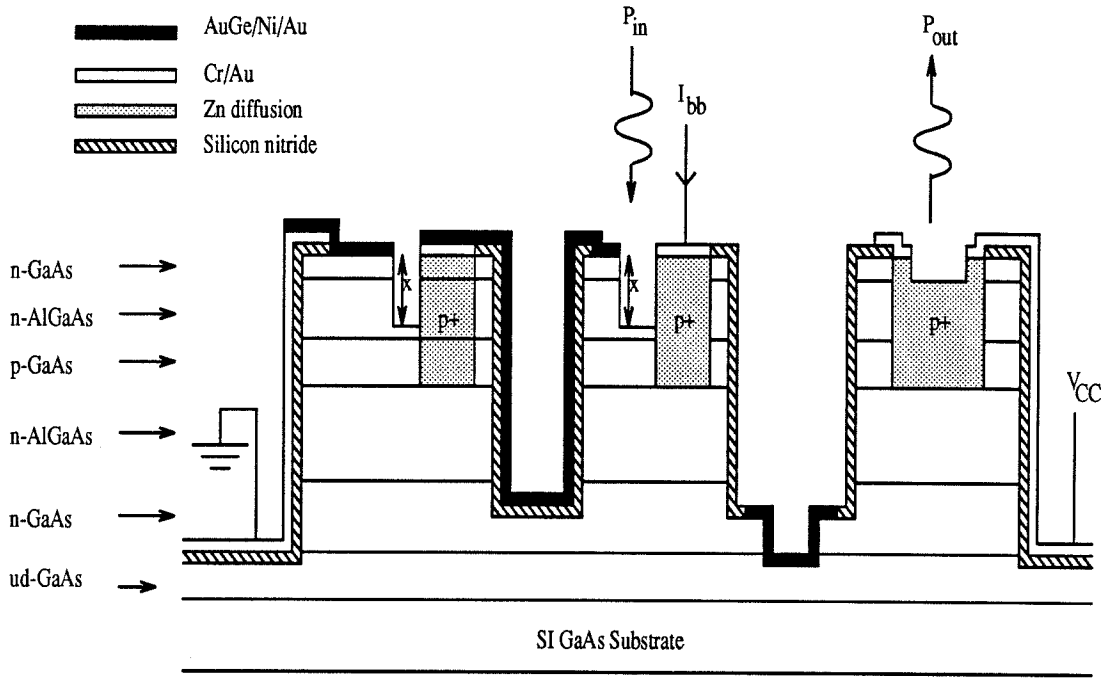


Fig. 4.16 Cross sectional view of the Darlington transistor pair monolithically integrated with the LED. In order to maximize the current gain of the transistors, the bulk emitter region between the emitter and the base contacts were etched away to leave only a thin and depleted AlGaAs as a passivation layer.

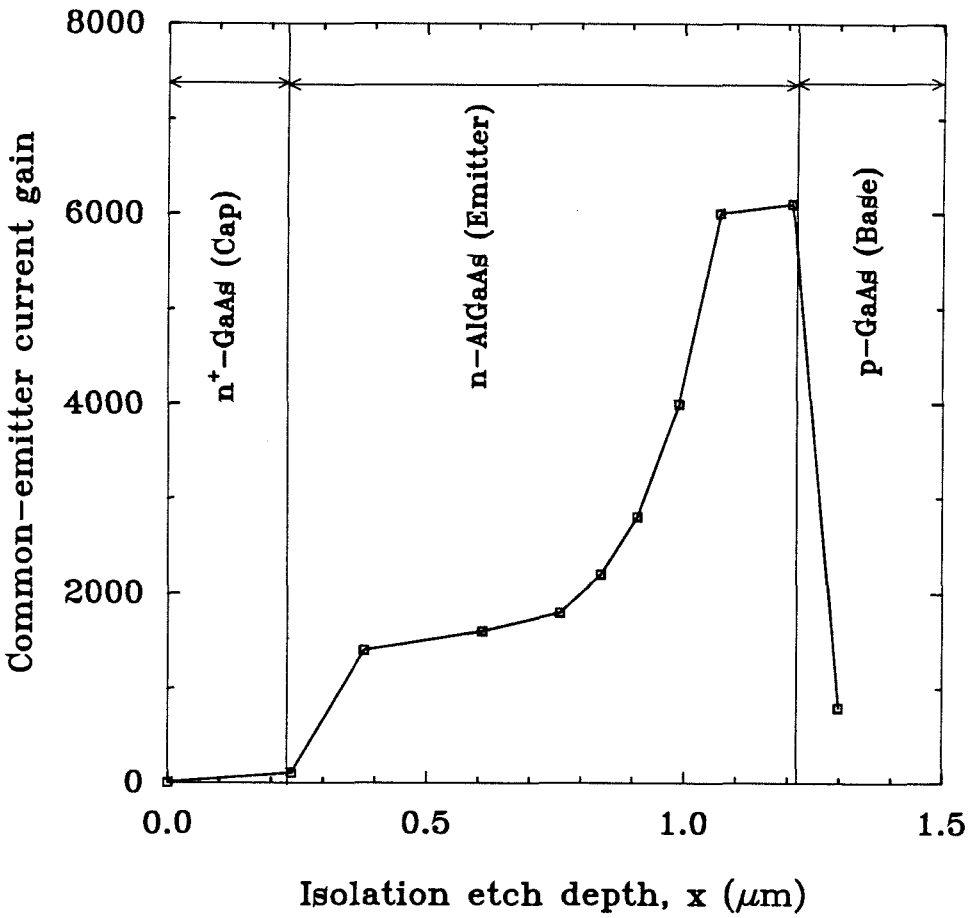


Fig. 4.17 Common-emitter current gain as a function of the isolation etch depth. x corresponds to the etch depth depicted in Fig. 4.16. Before the etch, the current gain was only 10, indicating the majority of the base current was flowing through the n^+ GaAs cap layer. As the etch depth increased, the current improved steadily until the remaining AlGaAs emitter was totally depleted. At which point, the current gain saturated. Thereafter, the current degraded dramatically due to the exposure of the base to the air, which promoted the surface recombination current.

depleted, which, in turn, served as a surface passivation layer, and effectively eliminated all the recombination current that flowed on the surface. This was clearly illustrated by the flat plateau that indicated the saturation of the current gain at this situation. The width of the plateau approximately corresponded to the maximum thickness of the AlGaAs emitter layer that would be used to passivate the surface. As long as there was a depleted AlGaAs layer covering the base layer, the current remained constant, suggesting the independence of the current gain by the surface effect. However, as the isolation etch reached into the base layer, the extrinsic base region was exposed to air, thereby enhancing the recombination of electrons and holes on the exposed surface. This resulted in a detrimental reduction in the current gain of the transistor as the current gain plummeted to about 10% of the maximum value. By using this technique, a maximum current gain of 6000 was obtained in the Darlington transistor. This would more than satisfy the loop gain requirement imposed by the network. However, the current level at which this gain of 6000 was measured was at 20 mA. With a 5-volt power supply, the electrical power dissipation was 100 mW. Without a special cooling design, the heat generated would seriously limit the density of the neurons as eventually the generated heat would cause the device to fail. Thus, unless the same current gain could be obtained at however a much lower current level, the integration density for the neurons that were based on bipolar transistors would be severely hampered.

Chapter 5

Metal Semiconductor Field-Effect Transistor-Based Neurons

5.1 Introduction

In the previous chapter, we described the integrated DHBT-based optoelectronic neurons. Because of the high electrical power dissipation by the neuron, it is concluded that a large array of these neurons would present a very severe heat dissipation problem, which eventually would lead to the failure of the chip. This limitation originates from the fact that the current gain of a bipolar transistor depends on how hard the transistor is driven. The larger the current is, the higher the current gain will be. This has the undesirable consequence of obtaining the gain required at the expense of power dissipation. In addition, this circuit does not have any input-output isolation. The input signal is directly amplified to obtain the output signal. Thus, any gain required will have to be directly provided by the Darlington transistor pair. This puts a very stringent requirement on the Darlington transistor pair. Namely, it has to provide a sufficient current gain and yet dissipate little power. This however contradicts the gain-power tradeoff rule stated earlier. Thus, an alternative design that decouples the relationship between the gain and the power as well as the relationship between the input and the output has to be developed.

One such possible alternative is to use a voltage-controlled transistor, such as a metal-semiconductor field-effect transistor (MESFET), to drive the LED. This MESFET is in turn controlled by an input switching circuit, composed of a detector, which accepts the input light, and a loading transistor. Thus, as the detector

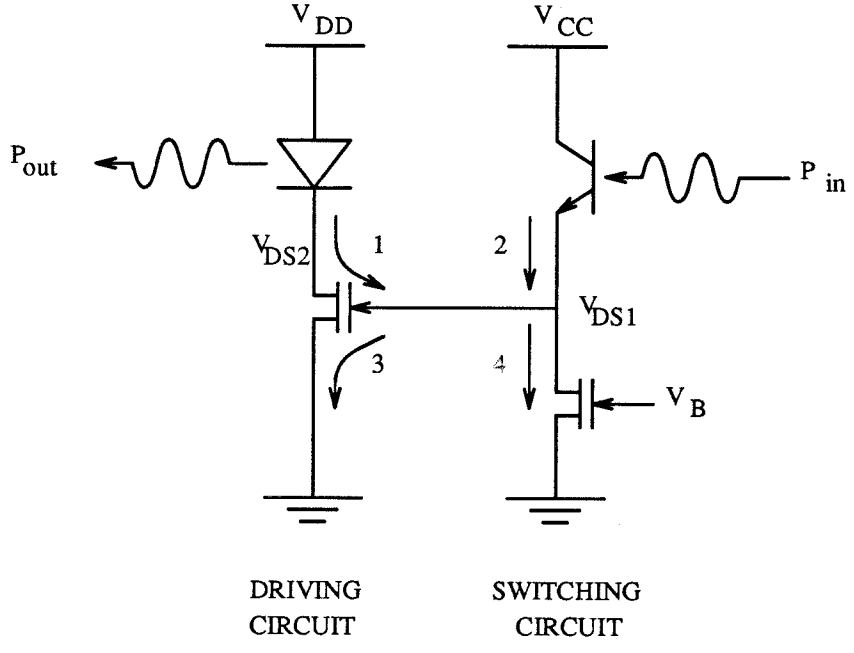


Fig. 5.1 Schematic circuit diagram of the optoelectronic neuron that incorporates two MESFET's, a phototransistor and a LED.

detects a sufficient input light, it pulls up the loading transistor. As a result, the LED-driving MESFET is turned on and it drives the LED. The advantage of this circuit design is that the LED is indirectly controlled by the input light, as opposed to the direct amplification of the input to produce the output in the DHBT-based neuron. Figure 5.1 depicts one such possible circuit, where a phototransistor is used as the detector and another MESFET is used as the loading transistor. As seen in Fig. 5.1, this circuit can be divided into the output driving circuit and the input switching circuit. The gate of the output driving MESFET is controlled by the voltage between the phototransistor and the loading MESFET. This voltage fluctuates between ground and V_{CC}, depending upon the the photocurrent generated. As the input optical power increases, the photocurrent increases. At a certain point, the generated photocurrent has surmounted the current drawn by the loading MESFET. At this point, this voltage changes from ground to V_{CC}. This turns on the output driving MESFET. A direct consequence of this circuit design is that the isolation of the output and the input. This would enhance the sensitivity of the circuit as the circuit can be designed to switch by a very weak input light. Another consequence of this switching action is that the optical gain is now determined by the relative output impedance of the phototransistor and the loading MESFET. As will be shown later, if these two transistors have infinite output impedances, the neuron can be turned on instantaneously. Other advantages of this circuit include the relatively mature technology in fabricating the high-gain MESFET's and a much lower electrical power dissipation required to turn on the neuron. These will be discussed in more detail later.

5.2 Analysis of MESFET-Based Neurons

The MESFET-based optoelectronic neuron consists of an output driving circuit

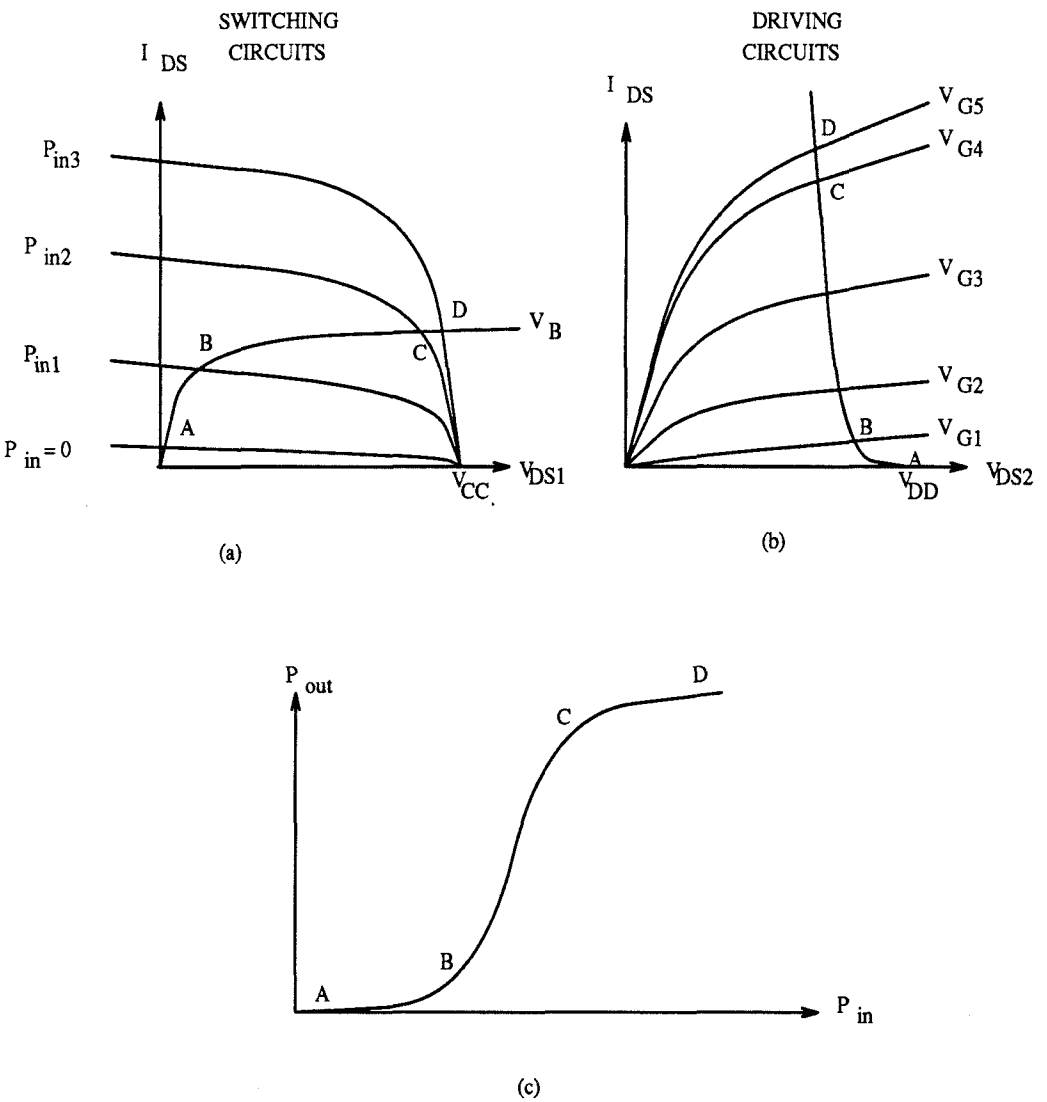


Fig. 5.2 **(a)** I-V characteristics of the input switching circuit in the MESFET-based neurons. The intersection point determines the operating point of the neuron. **(b)** I-V characteristics of the output driving circuit, which determines the output power level emitted by the LED. **(c)** Overall input-output characteristics of the neuron.

that is relatively simple in circuit design and easy to understand. The LED is driven by a MESFET, which, in turn, is controlled by an input switching circuit. To fully appreciate this circuit, one has to understand how the input switching circuit works and how it affects the operation of the output driving circuit. If we restrict ourselves first to analyze the input switching circuit, we see that the voltage in the middle of the two devices, V_{DS1}, can be designed to swing between ground and V_{CC}, depending on the relative currents drawn by each device, namely the phototransistors and the loading MESFET. This is best understood by analyzing the I-V characteristics of both transistors plotted on the same graph. Shown in Fig. 5.2(a) is one such plot. The I-V curve for the loading MESFET is plotted in a conventional way with the vertical axis being the drain-source current and the horizontal axis being the drain-source voltage, which is labeled as V_{DS1}. In order to plot the I-V curve for the phototransistor on the same plot, we use the fact that the voltage across the emitter and the collector, V_{CE}, is given by $V_{CC} - V_{DS1}$. Therefore, the I-V curve for the phototransistor is first flipped with respect to the vertical axis to get the $-V_{DS1}$ and then linearly translated to the right by V_{CC}. The resultant plot is shown Fig. 5.2(a). The voltage at the middle node, V_{DS1}, which is also the gate voltage of the output MESFET, is determined by the intersection point of the two transistor curves. For input light power equal to zero, the value of V_{DS1} is almost equal to zero, as indicated by point A in the figure. However, as the input light power gradually increases from zero to P_{in3}, the voltage, V_{DS1}, changes from point A through point B and point C and to point D. Thus, V_{DS1}, swings from almost ground to almost V_{CC}. To see how this swing in V_{DS1} affects the output circuit, a similar I-V curve of the output driving MESFET with the LED plotted backward is shown in Fig. 5.2(b). The swing in V_{DS1} corresponds to a swing in the gate voltage, which is designated by V_{G1} through V_{G5}. Assuming the output driving MESFET is an enhancement-mode transistor, the initial V_{DS1}, or the gate voltage,

of zero volt (point A) will not turn on the transistor. Thus the transistor is still in cut-off as indicated by point A in the output I-V curve. Since there is no current between the source and the drain, the LED is off. As V_{DS1} swings from ground to the voltage corresponding to point D, the gate voltage changes from ground to V_{G5}, which puts the MESFET in a strong forward conduction mode. The current flowing between the source and the drain is used to drive the LED, which emits light with an intensity that is linearly proportional to the current which passes through the LED. If we take the output power emitted by the LED as the output and retain the input power to the phototransistor as the input, we obtain the input-output characteristics of the MESFET-based optoelectronic neuron. This is shown in Fig. 2(c). Points A, B, C, and D are also labeled to indicate the various states that the neuron is in. It should be noted that the output power from the LED does not increase too much as the input power increases from 0 to P_{in1} (from point A to point B). This is because the increase in the input power does not generate enough photocurrent in the phototransistor to cause a significant change in V_{DS1}. However, from point B to point C, a dramatic increase in the LED output power is observed. This is due to the large change in V_{DS1} which is, in turn, caused by the phototransistor current overtaking the current drawn by the loading MESFET. This dramatic increase in the LED output power simulates the thresholding characteristics in the neurons with the level of threshold controlled by the biasing voltage, V_B, which is the gate voltage of the loading MESFET in the input switching circuit. From point C to point D, there is only a small change in V_{DS1}. Thus, the change in the LED output power is small. This provides a saturation effect, which is desirable for simulating the thresholding operation of the neurons. Therefore, the thresholding and saturation behaviors of the neurons can be easily controlled and simulated by using these four devices.

The level of the threshold can be adjusted by applying a different voltage

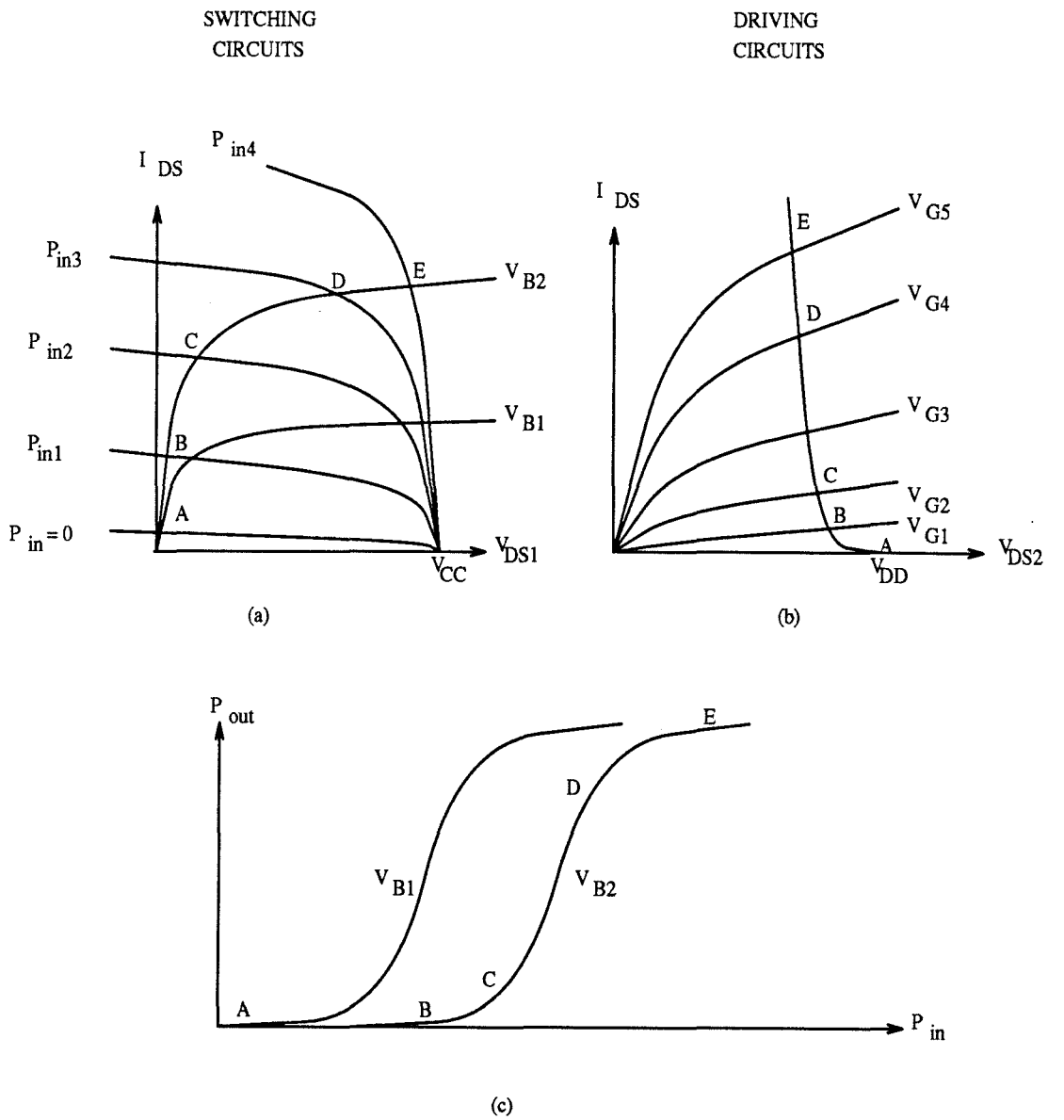


Fig. 5.3 **(a)** I-V characteristics of the input switching circuit in the MESFET-based neurons with a different biasing voltage applied to the gate of the loading MESFET. **(b)** I-V characteristics of the output driving circuit, which determines the output power level emitted by the LED. **(c)** Overall input-output characteristics of the neuron.

to the gate of the loading MESFET in the input switching circuit. The various operating states labeled by A through E are illustrated in Fig. 5.3 (a)-(c). From these plots, it is clear that this circuit does not only provide the desired input-output characteristics for the neurons, but it also enables the threshold level of the neuron to be electronically tuned through this bias terminal. This feature will be needed for the dynamics in a massively interconnected network of neurons. Generally speaking, the higher the biasing voltage, V_B, the higher the level of threshold in the neuron since a higher V_B induces a higher source-drain current through the MESFET, which, in turn, requires a higher input power in order to establish the onset of the threshold. From the physical standpoint, this loading MESFET provides a reference current against which the generated photocurrent from the phototransistor is compared. If the photocurrent generated by the phototransistor is not sufficient to meet the reference current, the voltage at V_{DS1} is pinned to ground. However, if the photocurrent is greater than the reference current, the voltage at V_{DS1} is pulled up to V_{CC}, which causes the switching to occur. By varying the magnitude of this reference current through V_B, one can obtain a set of thresholding input-output curves with different threshold levels.

The optical gain in the MESFET-based neuron is determined by the ability of the input circuit to switch for a given input intensity. Qualitatively, as long as the photo-generated current is larger than the reference current drawn by the loading MESFET, the switching occurs. However, there is a region in which the output rises gradually from zero to maximum. The slope of the rise defines the differentially optical gain. In this region, the output power level depends critically on the gate voltage of the output driving MESFET. If the gate voltage rises sluggishly, the output of the LED is expected to rise sluggishly as well. On the other hand, if this gate voltage rises instantaneously, the output of the LED rises instantaneously. Thus, the differential optical gain is determined by the sensitivity with which the

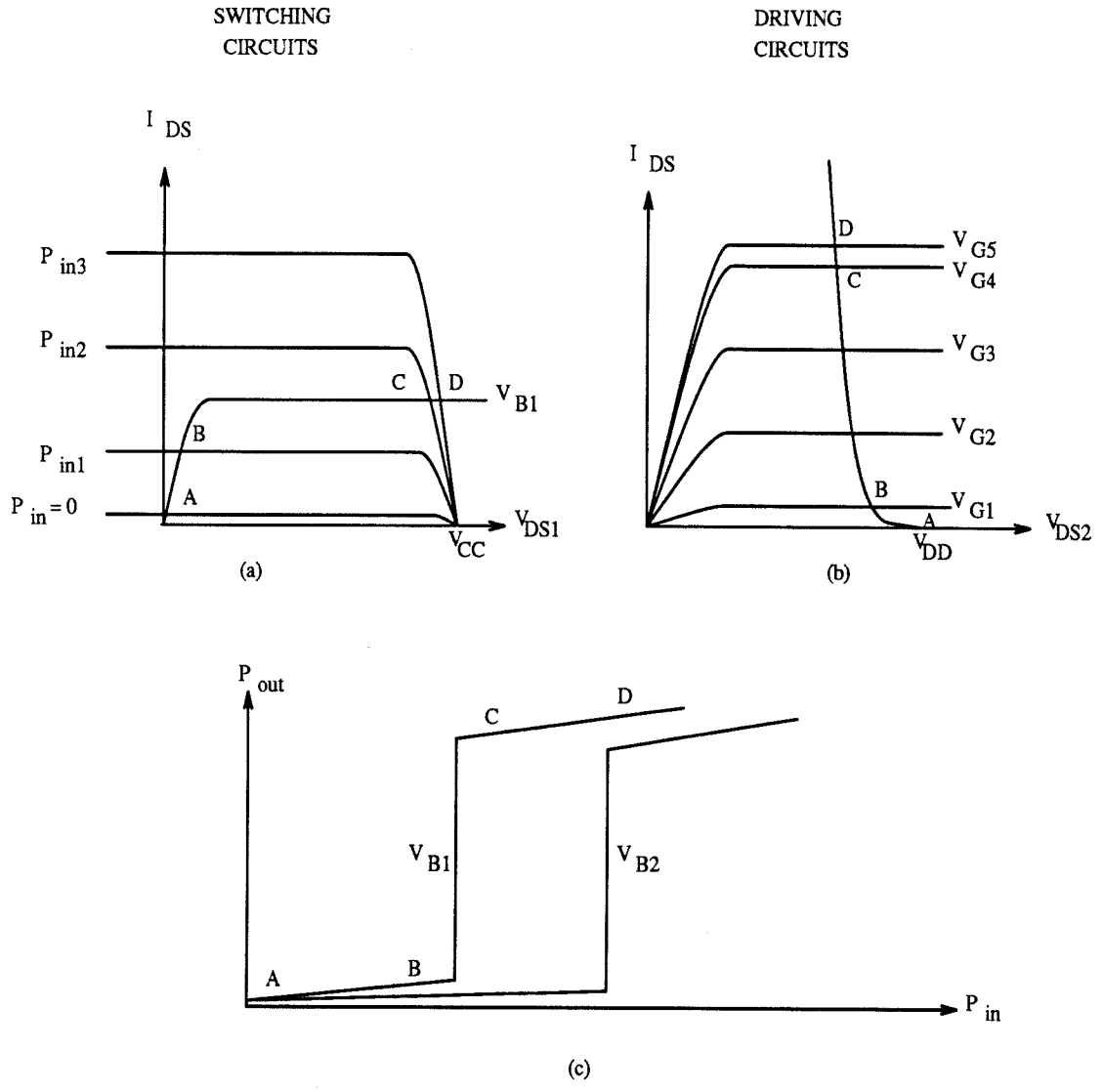


Fig. 5.4 **(a)** I-V characteristics of the input switching circuit in the MESFET-based neurons with an infinite output impedance in both the phototransistor and the loading MESFET. **(b)** I-V characteristics of the output driving circuit again with an infinite output impedance in the driving MESFET. **(c)** Overall input-output characteristics of the neuron showing an instantaneous switching as a result of having an infinite output impedance in the transistors.

voltage, V_{DS1}, can be raised for a given input light. This is further determined by the relative flatness in the I-V curves of both the phototransistor and the loading MESFET. If the output saturation currents of both transistors are not constant as the voltage varies, the switching will be a soft one. This is because the rise in current in one transistor accompanied by the same rise in current in the other transistor has to be accomplished by a change in the transistor voltage. Otherwise, current continuity will not be satisfied. This is the case that corresponds to finite output impedances in the transistors. If, however, the output saturation currents of both transistors remain constant, the switching characteristics can be expected to be an abrupt one because the photo-generated current and the reference are now independent of the voltage across the transistors and a comparison of the relative magnitude of the currents will uniquely determine the state of the switch. If the photocurrent is slightly less the than the reference current, the circuit will not switch. If, however, the photocurrent is just slightly larger than the reference, the circuit will switch. This instantaneous switching characteristic, caused by the infinite output impedances in the transistors, translates into an infinite differential optical gain. Therefore, it is extremely desirable to make these transistors with very high output impedances so that high-gain neurons can be obtained. The switching characteristics of the circuits for the infinite impedance case are illustrated in Fig. 5.4 (a)-(c) again with points A through D again to show the various states the circuit is in.

The nature of the output impedance can be quite complicated. It can be due to improper design in the material that causes the non-saturating current. For example, a low base doping concentration in the phototransistor will cause a severe sloping in the output current. However, increasing the base doping concentration unfortunately decreases the current gain, β, of the transistor. For MESFET's, the non-saturating output current is usually due to the source-drain current that spills

into the substrate [86] and causes a bias-dependent source-drain current. The origin of the non-saturating output current can be also caused by the leakage current in the transistors, particularly the reverse leakage current across the gate and the drain. As the voltage, V_{DS1}, is being raised from ground to V_{CC}, the gate-drain Schottky diode experiences a stronger reverse bias as the gate voltage is kept constant. This introduces a larger leakage current, which flows from the drain to the gate. From the point of view of V_{DS1}, this leakage current is no different from the reference current drawn by the same MESFET because both of these currents flow out of the node at V_{DS1}. As a result, this leakage current is mixed into the reference current, which is usually bias-dependent to start with. Therefore, the total current becomes even more bias-dependent and consequently the output impedance decreases.

Leakage currents complicate the analysis of the switching behavior significantly if the input switching circuit is connected to the output driving circuit. This is because the isolation between the input and the output circuits is not complete. Having a Schottky diode at the gate, MESFET's inevitably draw leakage current across the gate, either from the source or from the drain, depending on the bias of the transistor. The switching characteristics presented above is an overly simplified picture of the real device. In reality, there are 4 basic current components that determine the switching characteristics of the neuron (instead of just two as previously mentioned). Referring to Fig. 5.5(a), I_2 is the photo-generated current from the phototransistor, and I_4 is the reference current drawn by the loading MESFET. In addition, there is a current, I_1, that represents the leakage current across the gate and the drain in the output driving MESFET and an I_3 that represents the other leakage current component in the MESFET, which is the gate-source leakage current. At any time, the sum of I_1 and I_2 has to equal the sum of I_3 and I_4 in order to satisfy the current continuity equation. Since these four current components depend on V_{DS1}, V_{DS1} will adjust itself such that the current continuity equation

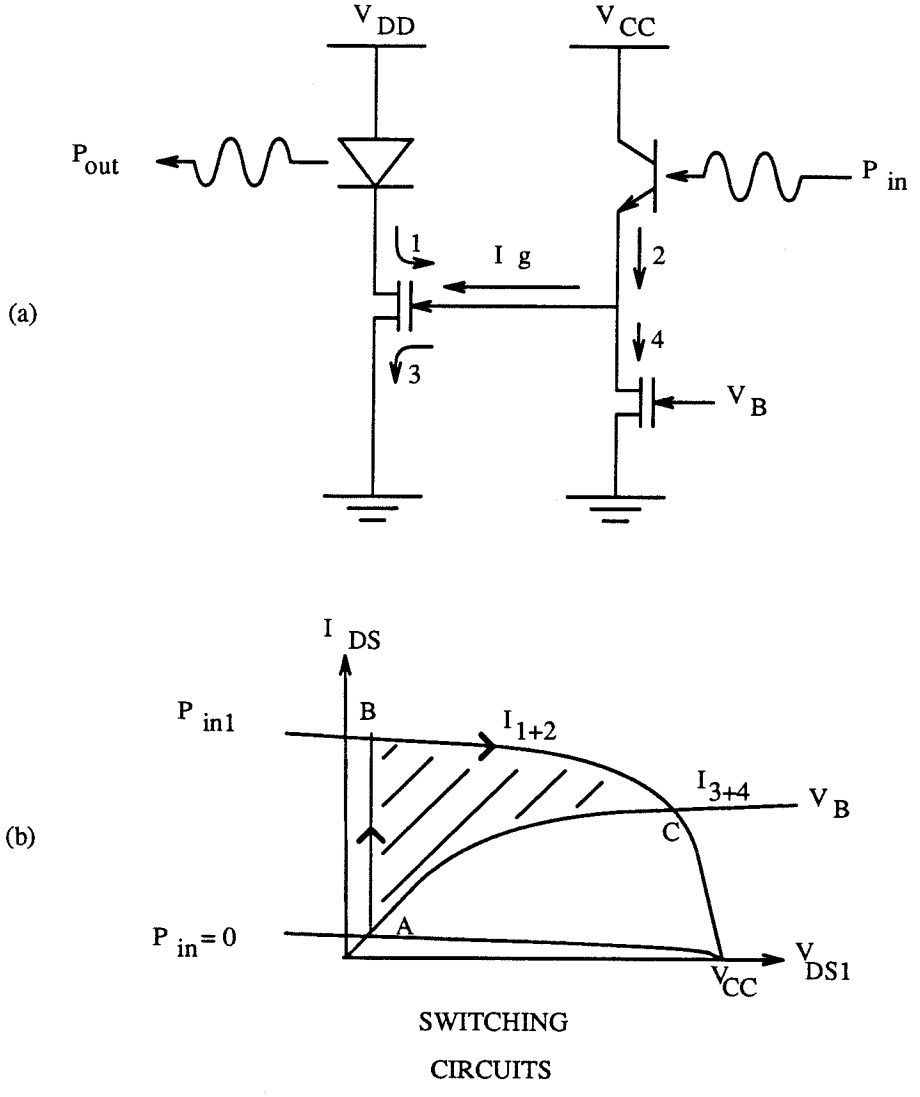


Fig. 5.5 **(a)** Schematic circuit diagram of the MESFET-based neuron showing the four current components that affect the gate voltage of the output driving MESFET, V_{DS1}. **(b)** Evolution of current dynamics as the neuron is being turned on. The neuron starts from point A in the off-state and ends at point C in the on-state through the state at point B. The shaded region represents the excess current available to charge the gate of the output driving MESFET.

is satisfied. Any perturbation in any one of the current components will cause the re-adjustment in V_{DS1}. As the input power is increased, I_2 increases proportionally. Thus, V_{DS1} reacts to this imbalance by increasing its value, which, in turn, decreases I_2 and I_1 and increases I_3 and I_4 at the same time so that the current continuity is satisfied again. As far as V_{DS1} is concerned, there is no difference between I_1 and I_2 because these two current components both flow into the node, providing the excess carriers needed by the other two current components. Nor is there any difference between I_3 and I_4 from the standpoint of V_{DS1} because these two current components both flow out of the same node, removing carriers that are injected by I_1 and I_2. Thus, at the end, we can still treat this switching circuit as being consisted of two current components; one flowing into the gate of the output driving MESFET and the other one flowing out of the gate. When the input light illuminates on the phototransistor, there will be an excess current that flows in the gate. This excess current is used to charge up the capacitance associated with increasing the gate voltage to its proper value. However, as the gate voltage increases, the magnitude of the excess current decreases owing to a smaller current that flows into the gate and a larger current that flows out of the gate. Eventually, as the final gate voltage is established, the current flowing into the gate is again equal to that flowing out of the gate. This process is illustrated in Fig. 5.5(b). Initially, the neuron is in the off-state, which is indicated by point A. As the input power jumps from zero to P_{in1}, the current that flows into the gate, $I_1 + I_2$, all of sudden increases to a value dictated by the amount of the input power, labeled as point B. This increase can not be accommodated immediately by the current that flows out the gate, $I_3 + I_4$. Therefore, the gate voltage has to increase in trying to balance the two current components. However, the gate voltage can not be raised immediately because these is a capacitance associated with charging up the gate. As a result, this excess current goes to charge up the capacitance of the gate in bringing up the

gate voltage until the the two current components balance each other. The time over which this switching takes place depends on the relative magnitude of the two current components. From this plot, it can be easily inferred that the stronger the input power is, the faster it will be for the neuron to reach the steady-state because there is more excess current available to charge up the gate. In fact, the switching time can be found by solving for τ_{charge} in the following equation.

$$CV_{CC} \approx \int_0^{\tau_{charge}} (I_1 + I_2 - I_3 - I_4)dt, \tag{5.1}$$

where C is the total gate capacitance that needs to be charged up. From this equation, we see that in order to decrease the switching time, one needs to decrease the capacitance and V_{CC} and increase the input power.

With this circuit, it is sometimes possible to have a situation in which the neuron is already on without any input power. This is due to the fact that I_1 is so large that it overcomes the combined currents of I_3 and I_4. As a result, the gate is fully charged up to almost V_{CC} and the LED is emitting. This situation is especially likely to occur when the output driving MESFET is very wide and the input loading MESFET is very narrow. The narrow-width MESFET is needed to increase the sensitivity of the input circuit. Thus, there is an optimal width in the loading MESFET that will prevent this phenomenon from happening and yet provide sufficient sensitivity. When this problem is present, it can be cured by increasing the biasing voltage, V_B, applied to the gate of the loading MESFET. This will increase the reference current, which provides a sink for I_1 to bring V_{DS1} down to the ground in order to shut the neuron off. In an opposite situation where the neuron can not be turned on by the input power, a bias optical beam can be applied to the phototransistor to generate more photocurrent, I_2. The magnitude of this optical beam can be just sufficient to bias the neuron to a point that the

original input power will be able to turn on the neuron. This situation is illustrated in Fig. 5.6. Thus, by using either the electrical bias to increase the reference current so that the neuron can not be turned on without any input power or the optical bias to decrease the amount of input power needed to turn on the neuron, the MESFET-based neuron can be properly tuned for maximum sensitivity and fault-tolerance.

5.3 Characterization of Discrete Devices

5.3.1 Metal Semiconductor Field-Effect Transistors

Metal semiconductor field-effect transistors are three-terminal devices in which one of the terminals, the gate, is used to control the current flow between the other two terminals, the source and the drain. The operational principle of the MESFET's is very similar to that of the junction field-effect transistors (JFET's) [87] except in MESFET's, Schottky diodes, as opposed to p-n diodes, are used to control the width of the depletion region, beneath which the current flows. In addition, because of the nature of the Schottky diodes, leakage currents through the gate tend to be higher in MESFET's as compared to those in JFET's. Nevertheless, the fabrication of MESFET's is much simpler because the formation of the control terminal, gate, is by metalization rather than by diffusion as in the case for JFET's.

A typical MESFET has one of the structures shown in Fig. 5.7. The first structure, shown in Fig. 5.7(a), is the simplest. It basically involves metalizing the source, the drain, and the gate appropriately on a properly doped material. The drawback is the relatively low breakdown voltage between the gate and the drain. Another disadvantage of this structure is the difficulty in placing the gate down accurately between the source and the drain. Since the spacing between the

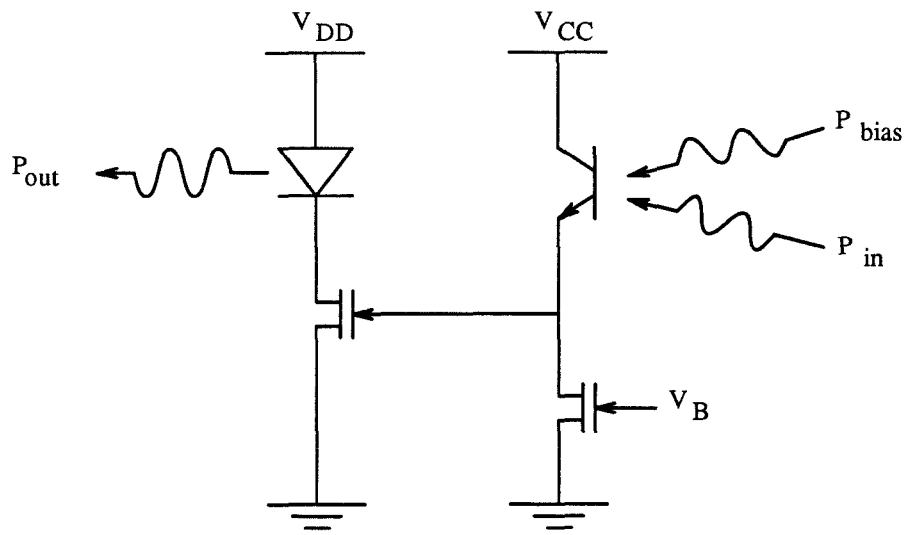


Fig. 5.6 Sensitivity of the MESFET-based neuron can be increased by applying an external optical beam to bias the neuron to just right before the threshold.

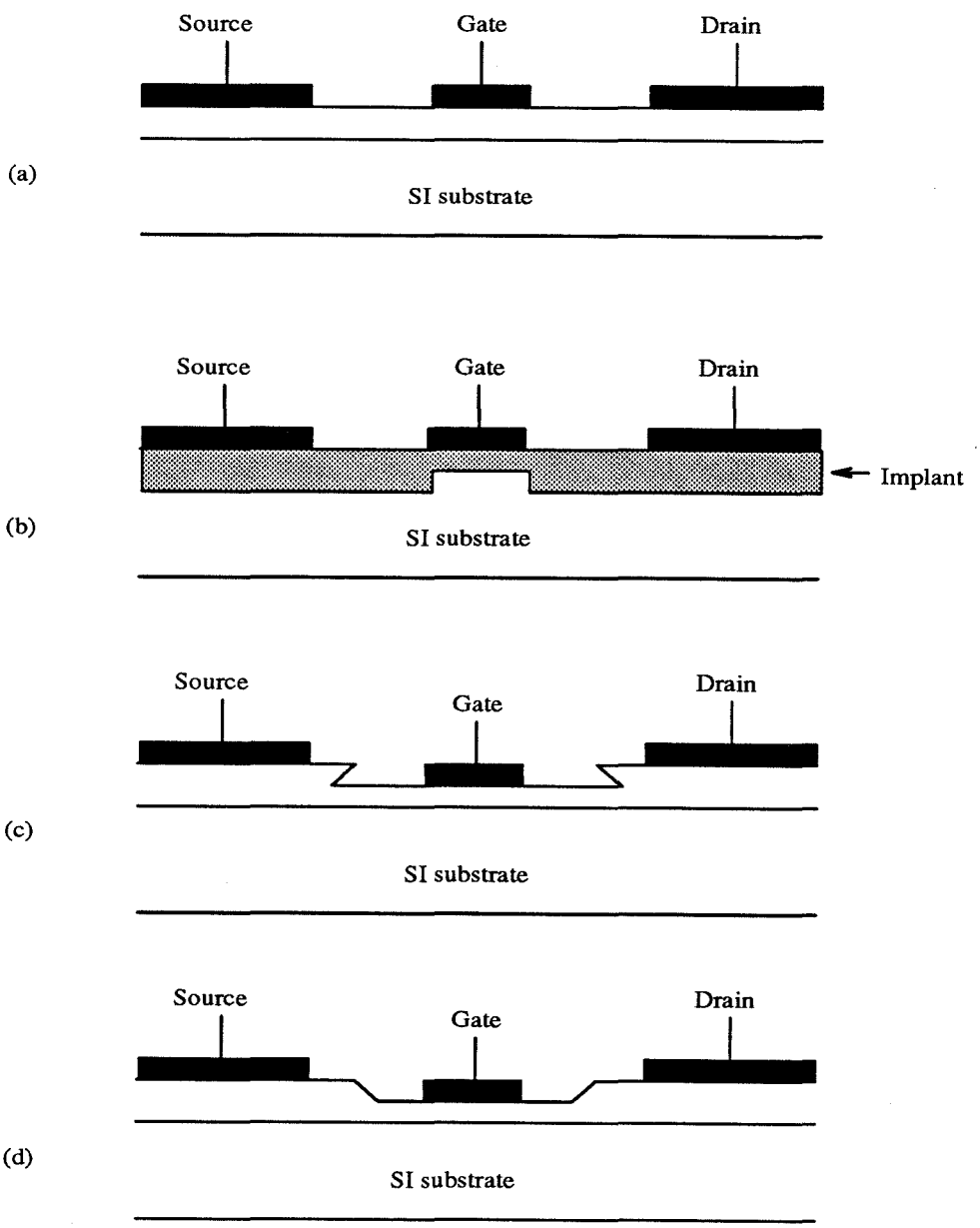


Fig. 5.7 Typical structure of a MESFET :**(a)** Conventional structure with planar geometry. **(b)** Self-aligned implanted structure. **(c)** & **(d)** Recessed gate structure.

source and the drain is typically less than 10μm and the gate length is already few μm long, a tight control on photolithography is very crucial. As a result, a self-aligned structure, such as the one shown in Fig. 5.7(b), has been developed [88,89]. It involves implanting an appropriate dosage of n-dopants into the material first to define the thickness and the doping level of the MESFET conduction channel layer. Then a special refractory gate material, typically made of Ti/Pt/Au, is evaporated, followed by a deeper implantation with a stronger dosage to define the ohmic regions for the drain and the source. This step is accomplished in a self-aligned manner because the gate metals are used as the implantation mask. Once the highly conductive ohmic regions for the source and the drain are defined, the actual source and drain metalizations can be defined without much precise control as long as they fall within the implanted region. Though the process is more tolerant, it does suffer from low breakdown voltage between the gate and the drain as the highly conductive ohmic drain region is very close to the gate. This is the direct result of using self-aligned implantation. However, if one designs the circuit properly so that the MESFET's will never be driven close to that breakdown voltage, this structure might prove to be very useful and yield very consistent device performance. In fact, this is the structure employed by the commercial MESFET company, such as Vitesse Semiconductor Corp [90]. Another way of making the MESFET, which will have a higher breakdown voltage, is to recess the gate slightly into the MESFET conduction channel layer by etching, such as the one shown in Fig. 5.7(c) and 5.7(d). Because of the property of GaAs, the side of the recess will make an obtuse or an acute angle with the surface depending on the orientation of the GaAs [91]. The effect of the recessed gate is not only to increase the breakdown voltage, but also to increase the transconductance of the MESFET through the reduction of the parasitic source-gate resistance [92]. However, between the structure in Fig. 5.7(c) and 5.7(d), the one in Fig. 5.7(c) tends to be less reliable as the sharp corners

resulted from etching generate high electric fields locally around the corners. Thus, in this work, the MESFET structure in Fig. 5.7(d) is used.

For a recessed-gate MESFET, it is extremely crucial that the etch depth be controlled as precisely as possible because the remaining channel will directly determine the pinch-off condition of the MESFET. Thus the operational mode of the MESFET, for example, either enhancement-mode or depletion-mode, will be affected by the amount of the recessed etch. For a recess that is shallow, the channel is not totally depleted. Therefore, a negative voltage is needed at the gate to pinch the channel off. This is the depletion-mode operation. On the other hand, if the recess is excessive such that the channel is already totally depleted, then a positive voltage is needed at the gate to induce a current flow between the source and drain. This is the depletion-mode operation of the MESFET.

Once the configuration of the MESFET is determined, there remains several detailed issues that need to be addressed for the optimization of the MESFET performance. Firstly, the parasitic resistance between the source and the gate contributes to a reduction in the effective voltage between the source and the gate. This can be explained by referring to Fig. 5.8. Because of the finite separation between the gate and the source, there is a parasitic resistance, R_{gs}, which accounts not only for this separation, but also the contact resistance and the distributed bulk resistance contributed by the source metalization. This resistance causes a voltage drop along the channel even before the channel current gets to the edge of the gate metalization. As a result, the effective voltage between the gate and the source is less. Therefore, the overall transconductance drops. The magnitude of the drop can be determined by the following expression.

$$g_m = \frac{I_{ds}}{V_{gs}}$$

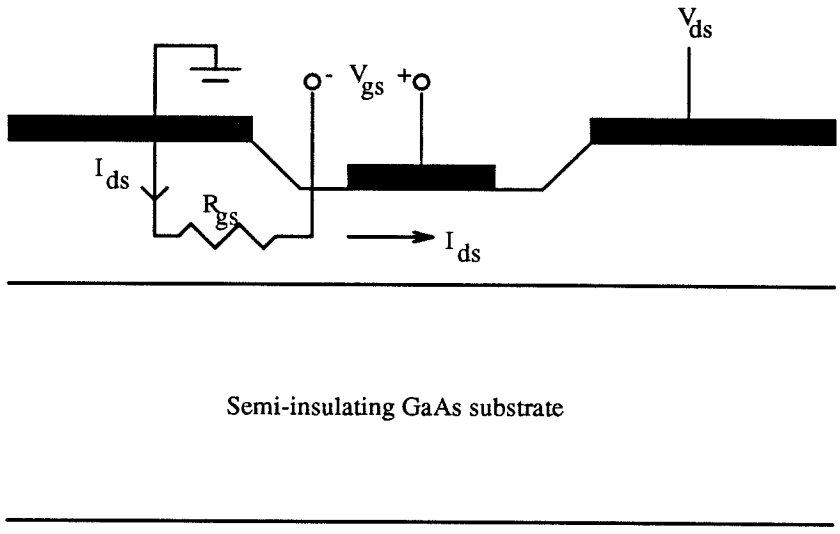


Fig. 5.8 Model of a MESFET including the parasitic R_{gs}, source-gate resistance.

$$= \frac{I_{ds}}{V'_{gs} + I_{ds}R_{gs}}$$
$$= \frac{g'_m}{1 + g'_m R_{gs}}, \tag{5.2}$$

where V'_{gs} and g'_m represent the true source-gate voltage and the intrinsic transconductance of the MESFET respectively. Therefore, the larger the parasitic resistance, R_{gs}, is, the more reduction there is in the transconductance of the transistor. This is very undesirable for the transistor. Thus, one should minimize this parasitic resistance. One way is to abridge the gap between the gate and the source metalizations. In the extreme case where the gap is zero, a self-aligned structure is obtained in which the edge of the gate metalization is aligned to the edge of the source metalization. This is illustrated in Fig. 5.9(a). This requires using the evaporated source metals as the etching mask in recessing the gate down to the appropriate depth. The area of the gate metalization has to also overlap slightly over the source metalization. This is because only by overlapping the two metalizations can a truly self-aligned structure be obtained. The overlapping portion of the gate metalization becomes part of the source contact with remaining non-overlapping gate physically defining the size of the gate metalization.

If we simply increase the length of the gate to overlap the source in obtaining the self-aligned structure, two problems arise. One is the degradation of the transconductance due to the increased gate length. The other one is the small breakdown voltage between the gate and the drain because of the small separation between the two terminals. In fact, the breakdown in a MESFET is usually dominated by the breakdown between the gate and the drain. If one measures the breakdown voltage of a MESFET, one would find it almost equal to the reverse breakdown voltage in the gate-drain Schottky diode. This is experimentally verified and shown in Fig. 5.10(a) and (b), in which the breakdown voltage of approximately

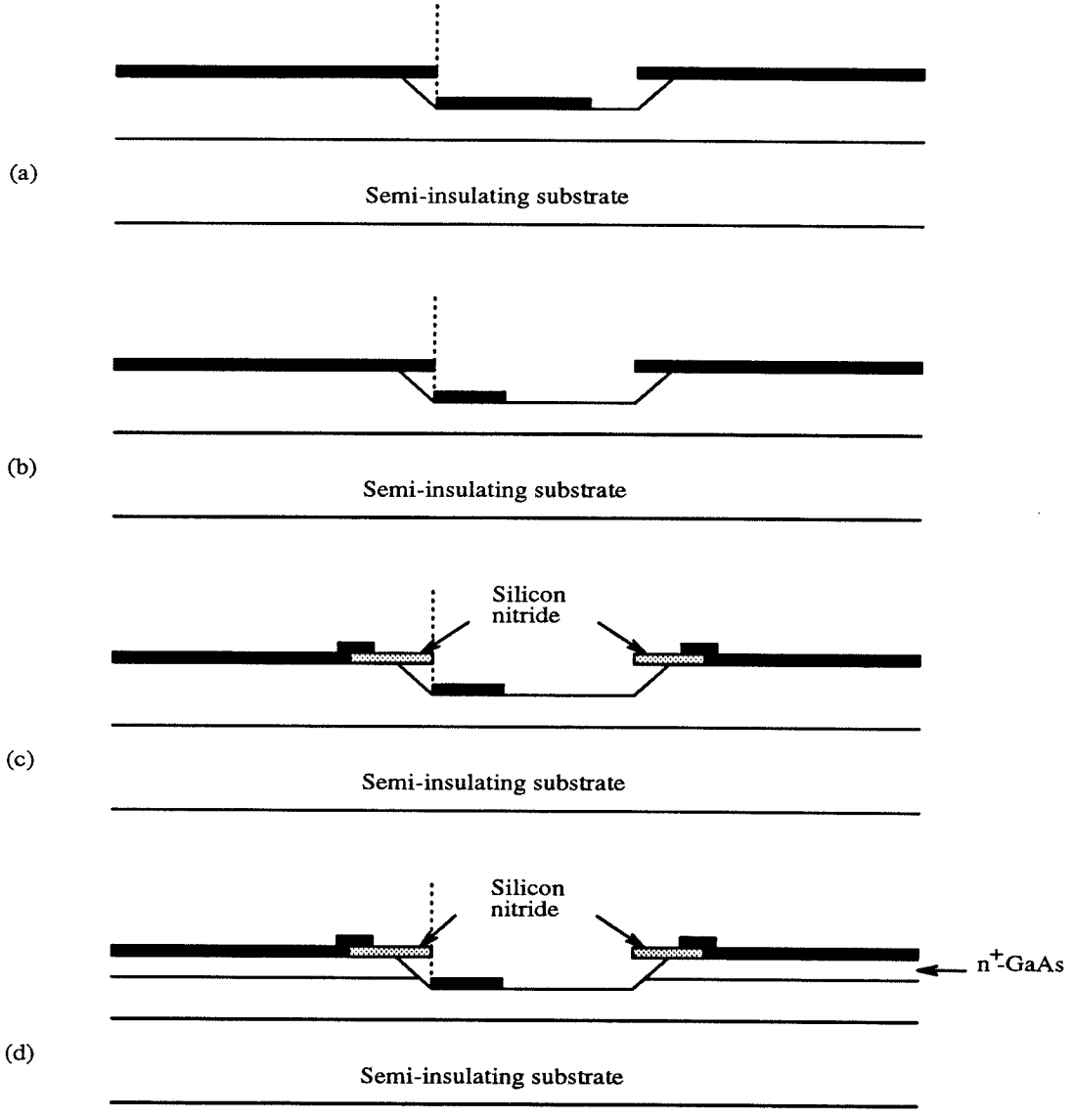
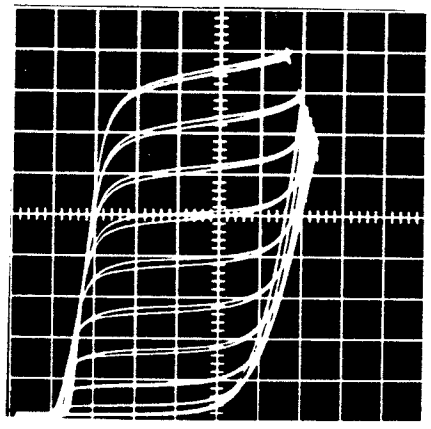


Fig. 5.9 The MESFET structure in which **(a)** the gate is aligned to the source to decrease the R_{gs}. **(b)** the separation between the gate and the drain is maximized for increased breakdown voltage. **(c)** a Si_3N_4 film is inserted to prevent the possible shorting between the gate and the source while still maintaining the self-aligned structure. **(d)** a n^+ GaAs layer is inserted to again decrease the R_{gs} as well as to facilitate ohmic contacts.

(a)



(b)

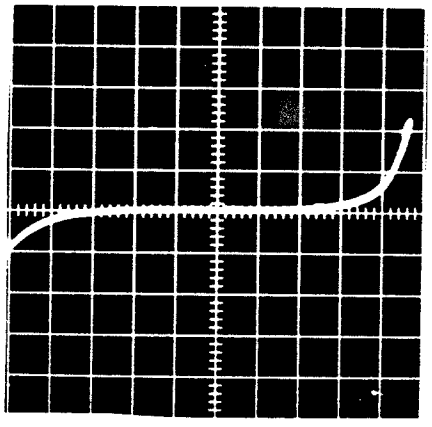


Fig. 5.10 **(a)** Common-source I-V characteristics of a MESFET showning a breakdown voltage of ≈ 4 V. The initial turn-on voltage of 1 V is due the LED which is in series with the MESFET. Scales are 500 μA/div vertically and 1 V/div horizontally. **(b)** The reverse breakdown characteristics of the gate-drain Schottky diode (first quadrant) and the gate-source Schottky diode (third quadrant). Scales are 10 μA/div vertically and 1 V/div horizontally.

4 Volts in a MESFET shown in Fig. 5.10(a) matches well the breakdown voltage of its gate-drain Schottky diode shown in Fig. 5.10(b). This is indicative of the strong correlation between the two breakdown phenomena. To eliminate this problem, one has to place the gate farther away from the drain. Therefore, reducing the size of the gate not only achieves a higher breakdown voltage in a MESFET, but also increases the transconductance. This improved structure is seen in Fig. 5.9(b).

Because of the self-aligned nature in defining the gate in a MESFET, it is sometimes inevitable that the gate metalization is shorted to the source metalization due to the close proximity these two metalization are with respect to each other. Therefore, it is necessary to insert a dielectric layer, such as Si_3N_4, which acts as a spacer in preventing the shorting between the gate and the source, and still maintains the self-aligned gate structure with respect to the nitride layer. This is shown in Fig. 5.9(c). The insertion of the Si_3N_4 layer however increases the physical spacing between the gate and the source metalizations, which, in turn, increases the parasitic gate-source resistance, R_{gs}, as mentioned before. Therefore, it is necessary to insert an n^+ GaAs layer beneath the drain and the source metalizations to reduce the actual distance between the source and the gate as well as to decrease the resistance for the drain and the source ohmic contacts. As a result, a MESFET structure shown in Fig. 5.9(d) is obtained. It is a self-aligned and passivated MESFET with a recessed asymmetric gate.

The composition of the material required for this self-aligned and passivated MESFET with a recessed asymmetric gate consists of an n^- GaAs layer beneath an n^+ GaAs layer on a semi-insulating GaAs substrate. The fabrication process of the MESFET is outlined in Fig. 5.11. Firstly, a blank deposition of Si_3N_4 was performed on the wafer followed by etching away the Si_3N_4 at the source and drain ohmic contact regions in a CF_4 plasma. AuGe/Ni/Au were evaporated onto the wafer and lifted off to define the source and the drain. The wafer was then subjected

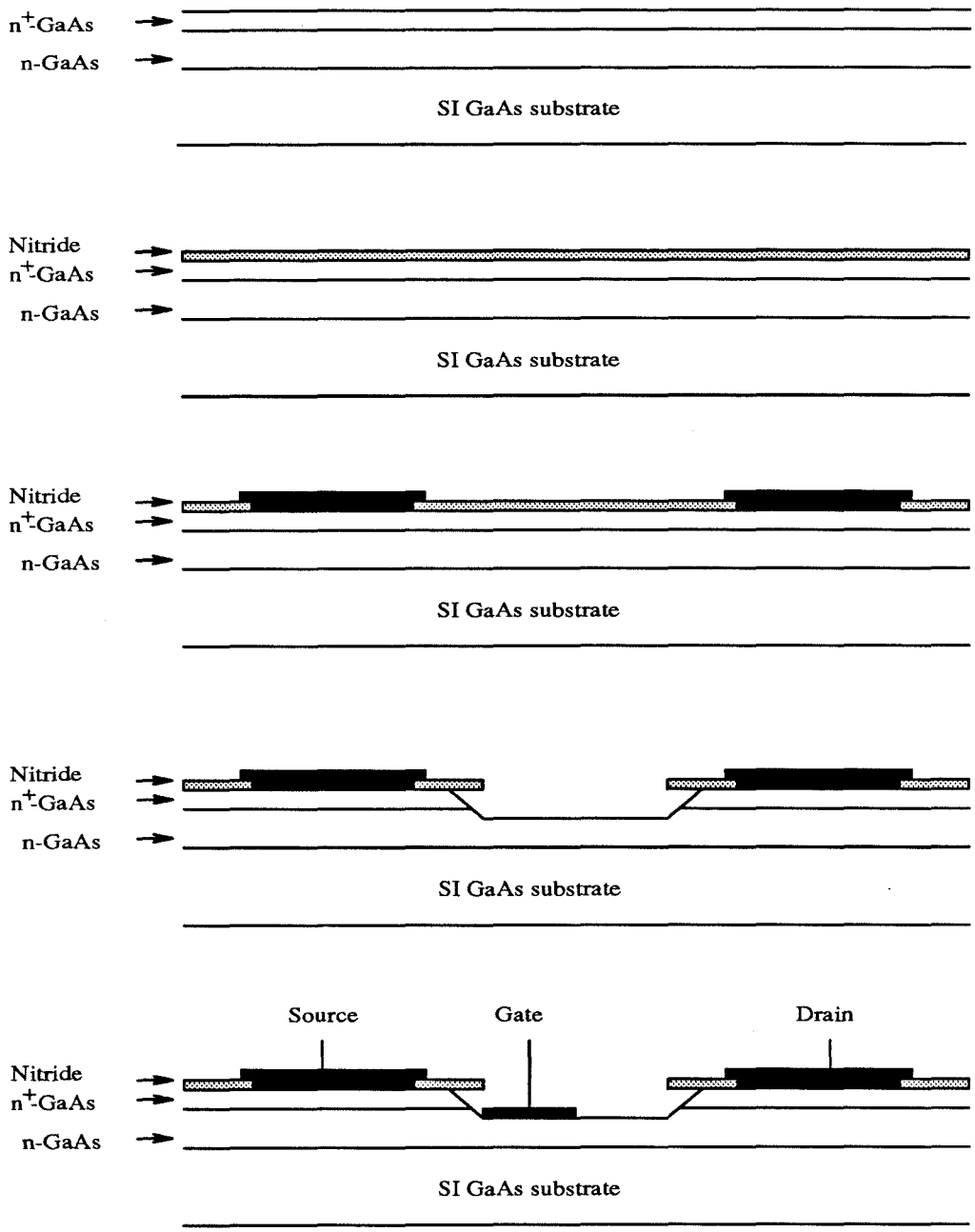


Fig. 5.11 Fabrication steps of a self-aligned and passivated MESFET with a recessed asymmetric gate.

to alloying in a N_2 ambient at 430^o C for 4 minutes. The gate recess region was photolithographically defined next and the exposed Si_3N_4 was again etched away in a CF_4 plasma. Once the Si_3N_4 in the gate region was removed, the wafer was immersed in a chemical etchant, consisting of NH_4OH, H_2O_2, and H_2O in a 20 : 7 : 973 ratio, to recess the gate region. While the gate was being recessed, the amount of current flowing between the source and the drain was monitored. Initially, the effect of the etch was probably not apparent from the current measured. However, as the recess became deeper, the source-drain current started to saturate due to the fact the remaining conduction layer began to be pinched off at the drain end. As the etch got even deeper, the saturation current became even smaller. This etching process was continued until the desired saturation current was obtained. Because the gate was to be subsequently evaporated onto this recessed region, there would be an additional depletion region developed underneath the gate due to the gate metalization. This additional depletion region reduced the height of the conduction channel layer, which, in turn, reduced the source-drain saturation current. Therefore, the recessed etching was stopped slightly before the desired source-drain saturation current was reached, so that after the gate metalization the appropriate source-drain saturation current was obtained. Finally, Ti/Pt/Au was evaporated to define the gate in a self-aligned manner as described earlier. The dimensions of the gate was 7 $\times$ 100 μm^2 with a gate to drain spacing of 11 μm.

The I-V characteristics of the MESFET was shown in Fig. 5.10(a) earlier. A transconductance of 30 mS/mm and a source-drain breakdown voltage of 4 V were measured. The initial offset in the V_{DS} was due to the turn-on voltage of the LED, which was in series with the MESFET. These results were consistent with the expectation except for the low breakdown voltage of 4 V. This was probably caused by the surface-induced breakdown instead of the true gate-drain Schottky diode breakdown because any dirt or particles in the vicinity of or underneath the

gate would cause the premature breakdown.

5.3.2 Phototransistors

Phototransistors are bipolar transistors with a floating base. As the light is incident onto the phototransistor window, it traverses through the transparent and high-bandgap emitter region and is absorbed in the small-bandgap base layer. The photocurrent generated acts as the base current and is amplified through the normal amplification process in a heterojunction bipolar transistor to produce the collector current. Since there is an initial efficiency involved in detecting the incoming photons, the overall optical gain of a phototransistor is $\eta_D\beta$, where η_D is the efficiency of generating electrons from the incoming photons and β is the common-emitter transistor current gain. In designing a phototransistor, there is an issue that should be noted. It is the issue of the base layer thickness. A thin base layer should be used to maximize the current gain. However, if the base is too thin, the incoming photons will not be fully absorbed inside the base. Therefore, there is an optimal thickness for the base. Fortunately, the reduction of the current gain of the transistor with a thick base can be compensated by using as more lightly doped base. Consequently, the thickness of the base in the transistor should be chosen first to accommodate the absorption of photons and then to optimize the current gain. There is a disadvantage in using a lightly doped base, however. A lightly doped base causes the base width to be modulated by the reverse biased base-collector junction. This modulation results in a reduction in the effective base width, which, in turn, causes the collector current to rise. A rise in the collector current causes the output impedance of the transistor to decrease because the output saturation current is now an increasing function of the emitter-collector voltage. This is the Early effect.

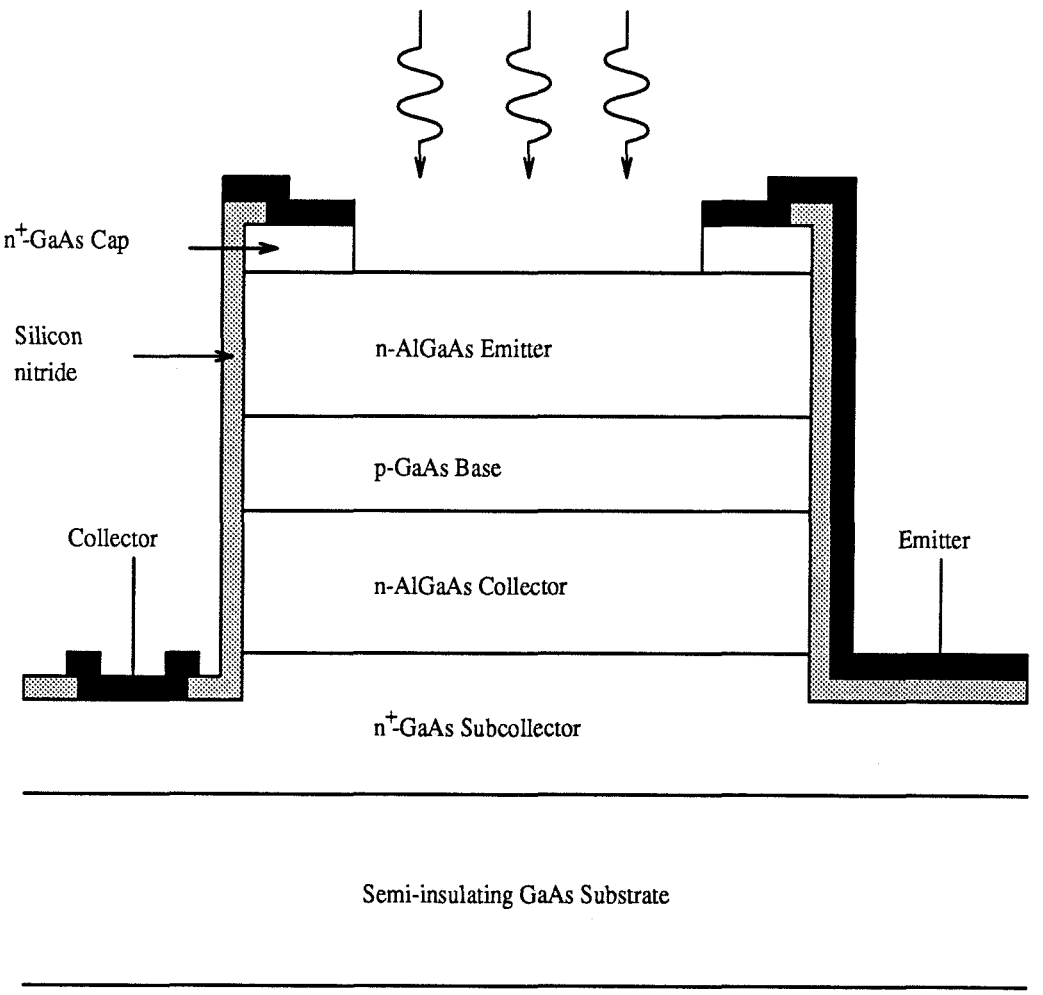


Fig. 5.12 The structure of a double-heterojunction phototransistor incorporating a p-doped GaAs layer as the base.

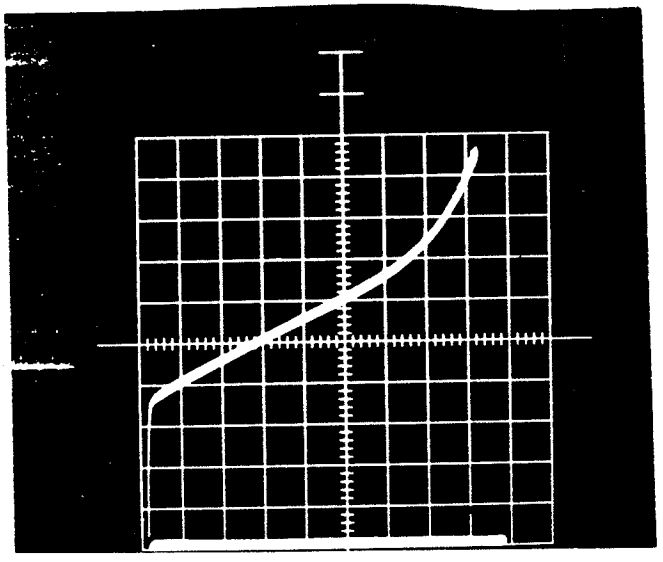


Fig. 5.13 The I-V characteristics of the phototransistor. The intensity of the incoming laser beam is 90 μ W. The scales for the vertical and horizontal axes are 20 μA/div and 2 V/div.

The structure of the phototransistor is shown in Fig. 5.12. It consisted of a lightly p-doped GaAs base layer sandwiched by two higher bandgap n-doped AlGaAs layers, namely the emitter and the collector. n^+ GaAs layers were used for emitter and collector contacts. The side wall of the phototransistor was passivated with a Si_3N_4 dielectric film over which the emitter and the collector metalizations ran. The window within which the incoming photons were incident was transparent to the photons by etching away the absorbing n^+ GaAs cap layer. The I-V characteristics of the phototransistor were obtained by monitoring the intensity of a GaAs laser onto the phototransistor and measuring the emitter-collector current simultaneously. This is shown in Fig. 5.13. Because of the low doping concentration used in this phototransistor, there was a severe Early effect, which caused the output impedance of the phototransistor to decrease substantially. From this measurement, an output impedance of 175 KΩ was obtained. Nevertheless, the I-V characteristics of the phototransistor shown in Fig. 5.13 was typical of all phototransistors fabricated. As the intensity of the laser beam increased, the current level increased as well. For this particular measurement, the input laser beam intensity was 90 μW and measured current was approximately 90 μA at a collector-emitter voltage of 4 V. This corresponded to an external efficiency of 1 A/W. Assuming an absorption efficiency of 0.3 A/W, we obtained a current gain, β, of only 3. This was a result of having a very thick base layer, which was 1.5 μm. The breakdown voltage of the phototransistor was 20 V, indicating the effectiveness of using a high bandgap and lightly doped AlGaAs collector layer.

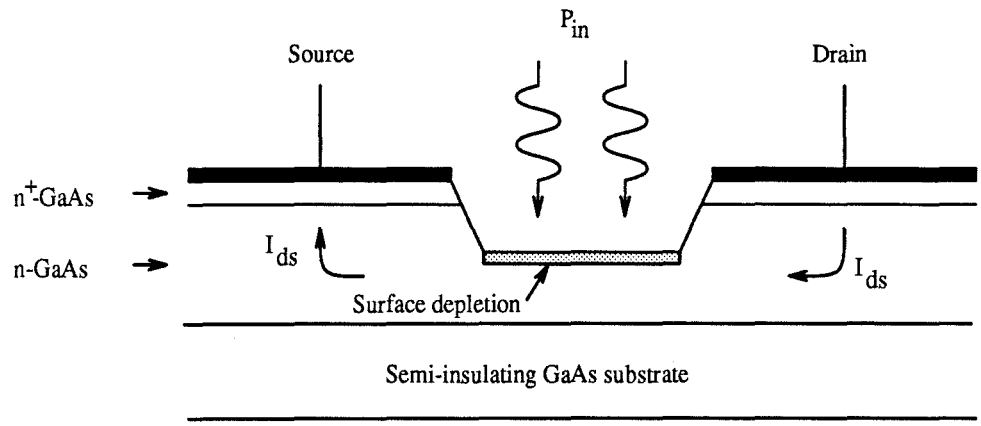
5.3.3 Optical Field-Effect Transistors

Another candidate for the detector is the optical FET. It is essentially identical structurally with the conventional MESFET except the optical FET does not

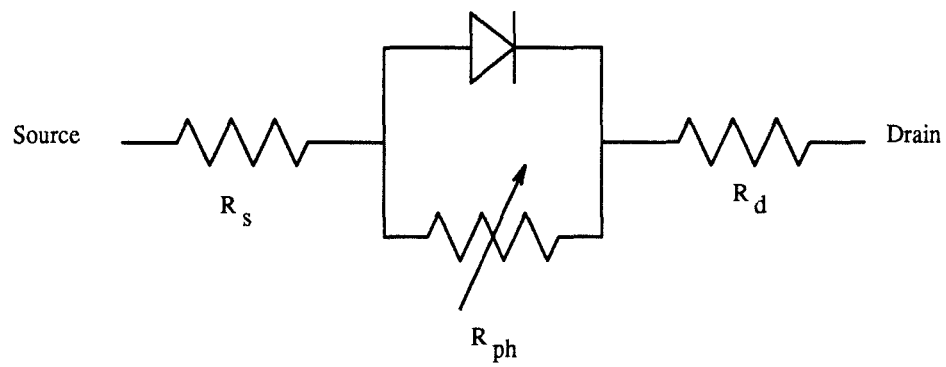
have a gate. Instead, it uses an optical input, which is incident on the gate area, to control the channel current underneath. There are two possible mechanisms in explaining the operation of an optical FET. One is based on MESFET-like mechanism [93]. As the electron-hole pairs are excited by the incoming laser beam, a portion of the electrons flow to the surface depletion region, in which the ionized donors are positively charged. This changes the surface potential with respect to the channel potential. As a result, the channel current is modulated by the transconductance of the underlying MESFET, which leads to an increase in the output source-drain current. However, Gummel et al. [94,95] have argued for a different mechanism, which is based on photoconductivity. As the electron-hole pairs are generated by the laser beam, external carriers are injected into the bulk of the photo-excitation region through the source and drain ohmic contacts in order to satisfy the steady-state recombination and generation requirement. In doing so, at rate at which these external carriers are injected into the photo-excitation region depends upon the carrier recombination lifetime and the device transit time. The longer the carrier recombination lifetime is, the more carriers are injected because the recombination rate is the ratio of the carriers injected over the carrier lifetime. Thus, longer carrier lifetime leads to a large carrier injection per unit time, which is current. However, on the other side, these injected carriers are swiftly removed by the electric field inside the device so that the there will be no carrier build-up over time. The faster the carriers are removed from the device, the larger the current is. Thus, the measured current due to the photo-excitation is expected to be inversely proportional to the carrier transit time across the device. In Sec. 5 of this chapter, a detailed analysis of the gain mechanism of a photoconductor is presented. In that analysis, it is found that the optical gain of a photoconductor is given by the ratio of carrier recombination lifetime over the carrier transit time. Thus, in maximizing the gain from the optical FET, the gap between the source and the drain should

be minimized and yet should still allow sufficient input light to be detected. Of these two mechanisms, the photoconductivity is the more likely explanation for the operation of the optical FET. This is based on the observation that some of the gain measured [94] is too high to be attributable to MESFET-like amplification.

Figure 5.14(a) shows the cross sectional structure of an optical FET. Since it is very similar to the conventional MESFET, the fabricational procedure of an optical FET is identical to that of a MESFET, except the gate metalization is not defined. Instead, the amount of the recessed depth is used to control the sensitivity and dark current desired. Generally speaking, the deeper the recess is, the smaller the dark current becomes and the smaller the optical gain is. To explain this dependency, a circuit model shown in Fig. 5.14(b) is proposed to model the optical FET assuming it is dominated by the photoconductivity mechanism. The current channel region is modeled as a reverse biased photodiode in parallel with a photoconductor. The existence of the photodiode is attributed to the fact that there is a surface depletion layer on the exposed surface of the recessed region. Carriers generated in this region are collected by the build-in electric field in the depletion region. Thus, the principle is the same as that of a conventional reverse biased p-n photodiode. Underneath this surface depletion layer, there lies a undepleted ohmic conduction channel made out of n-GaAs. Carriers absorbed in this region contribute to the photoconductivity action as described earlier. To complete the modeling, the source and the drain contact and bulk resistances are added in series. Since the photoconductor is a high-gain detector and the photodiode does not have any gain, the overall efficiency of the optical FET would be bounded by the efficiency of these two devices. If the photoconducting channel region is thick, a high-gain optical FET can be expected. As this region gets thinner by the recessed etch, the photoconductivity effect starts to decrease due to the smaller absorption region. As a result, the optical gain decreases gradually. This process continues until this photoconducting channel



(a)



(b)

Fig. 5.14 **(a)** The cross sectional view of an optical FET. The measured source-drain current is a combination of the p-n photodiode current and the photoconductor current. **(b)** The circuit model of the optical FET.

totally disappears, leaving only the surface depletion region. At this point, the optical FET does not exhibit any optical gain because the remaining photodiode does not have any gain. These characteristics are observed in actual experimental data, which are shown in Fig. 5.15 and 5.16.

Figure 5.15 shows the measured source-drain current as a function of the input laser beam power for four different dark currents, which correspond to four different recessed depths. The higher the dark current, the shallower the recess is. It is clear from this figure that the current increases monotonically as a function of the input power. However, there is a slight saturation in the current as the input power increases. Moreover, as the recessed depth gets larger, the measured current decreases. This is consistent with the argument presented earlier in the modeling of the optical FET. Figure 5.16 shows the efficiency of the optical FET in A/W for the same four sets of measurement with the same dark current. The efficiency of the optical FET increases initially as the input power increases. However, it decreases as the input power continues to increase. The mechanism of this unexpected behavior is not clear. But, this trend is consistent within each curve. However, the drop in efficiency of the optical FET as the recessed depth increases is in agreement with the model presented earlier.

5.4 Experimental Results of MESFET-Based Neurons

Having discussed the design consideration and analysis of the MESFET-based neuron as well as the individual devices in the neuron, we next describe the process of fabricating the optoelectronic neuron and the testing results. Figure 5.17 shows the device cross sectional view of the monolithically integrated optoelectronic neurons consisting of the device elements shown in Fig. 5.1. Basically, on the semi-insulating GaAs substrate, an undoped GaAs buffer layer was first grown. Upon which, an

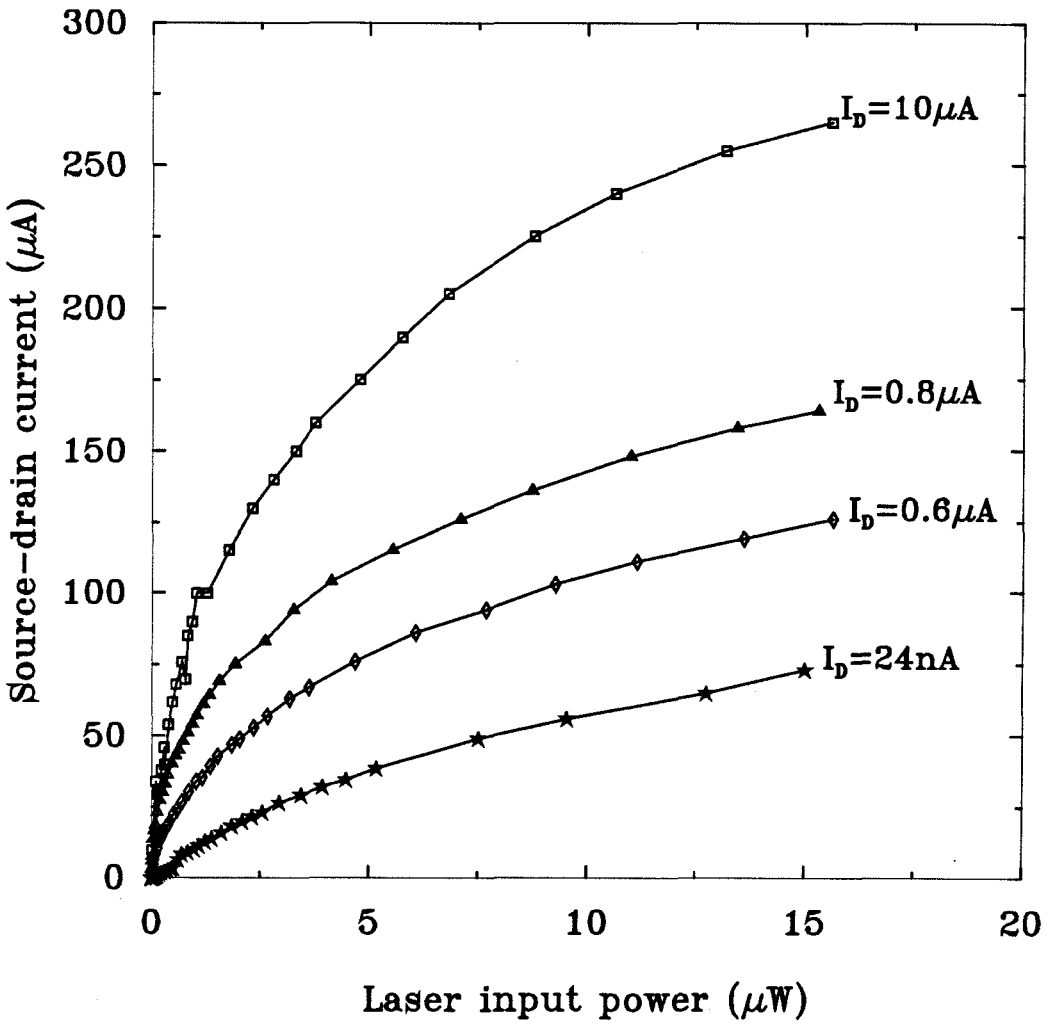


Fig. 5.15 Input-output characteristics of an optical FET. The output is the measured source-drain current and the input is the intensity of the laser beam. The four curves correspond to four different dark currents, which are due to the different recessed depth.

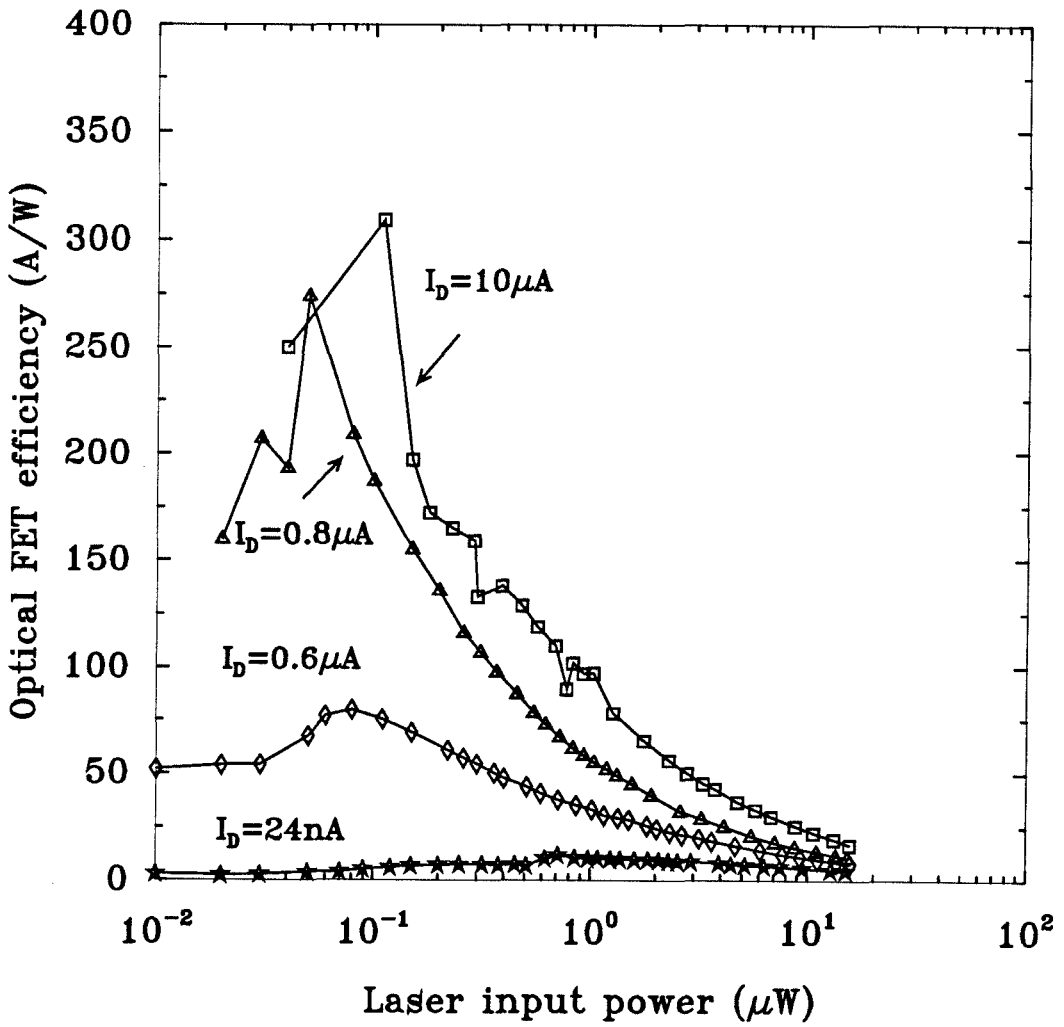


Fig. 5.16 Efficiency of the optical FET plotted in term of A/W as a function of the input power. The vertical axis is re-plotted from Fig. 5.15 by taking the slope of the curve. The horizontal axis remains the same.

n^+-GaAs acting as the source and the drain ohmic contact layer on top of an n^--GaAs current conduction channel layer were grown. These two layers form the structure of the MESFET. On top of the n^+-GaAs layer, a conventional double heterojunction bipolar transistor structure was grown. It consisted of an n^+-GaAs layer as the subcollector layer, an n-$Al_{0.35}Ga_{0.65}As$ layer as the collector layer, a p-GaAs layer as the base layer, an n-$Al_{0.35}Ga_{0.65}As$ layer as the emitter layer and an n^+-GaAs layer as the contact layer. The doping concentration and the thickness of each layer are listed in Fig. 5.18. The formation of the LED was completed by diffusing Zn twice over different areas to create the confinement for the current, as discussed in Ch. 2.

The fabrication of the MESFET-based optoelectronic neuron began by applying the standard degreasing and cleaning procedure to the surface of the GaAs epitaxial layers. A non-selective etch, consisting of a mixture of H_3PO_4 , H_2O_2 and CH_3COOH in the ratio of 1 : 1 : 3 was used to etch down to the n^+-GaAs layer to define the LED and the phototransistor. After this, the same etch was used to etch down the semi-insulating substrate to define the MESFET. A blank deposition of Si_3N_4 was then applied to the surface of the device by using a thermal chemical vapor deposition system heated to 610°. The gases used were silane diluted to 1% by nitrogen, ammonia and nitrogen. A thickness of approximately 1200 $\AA$ to 1500 $\AA$ of Si_3N_4, which exhibited a color of blue to light blue, was deposited. The next step was Zn-diffusion to convert the n-AlGaAs emitter layer to p-AlGaAs for the upper cladding layer for the LED as well as to provide the current confinement. This was achieved by selectively removing the Si_3N_4 over the the LED window area in a CF_4 plasma and performing a sealed ampoule Zn-diffusion process. The ampoule, in which the neuron device and the diffusion source, $ZnAs_2$ were placed, was pumped to a vacuum of 8×10^{-8} torr before it was sealed with a torch. The ampoule was then inserted into a furnace of 640° to promote the diffusion of Zn

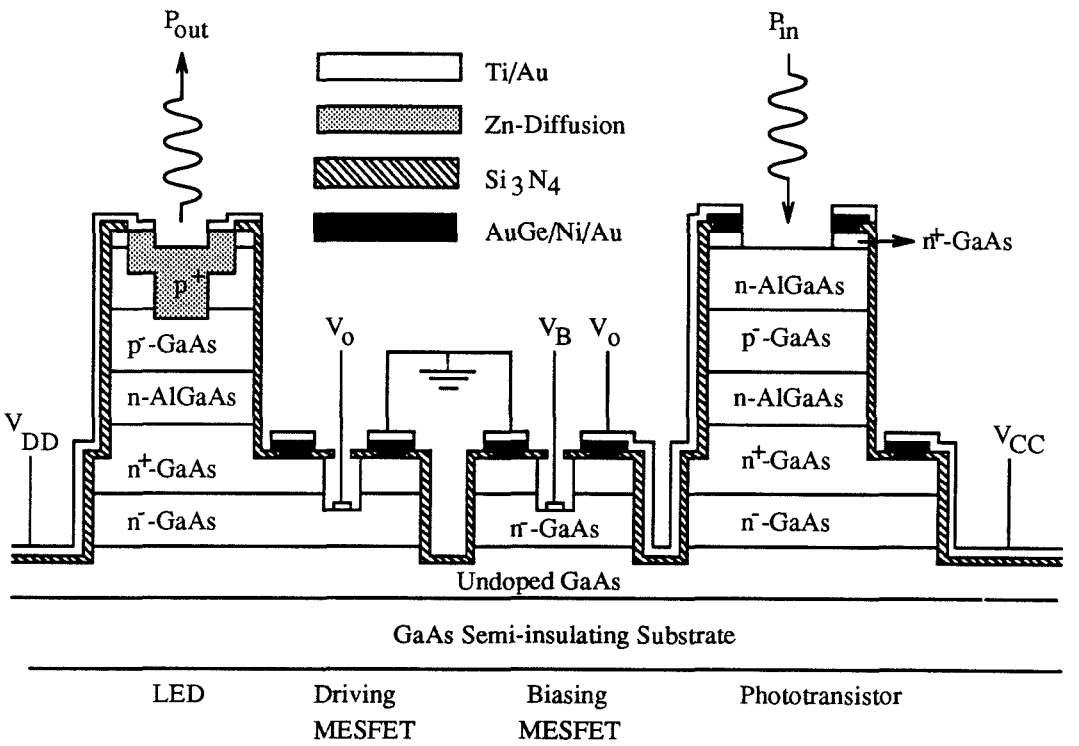
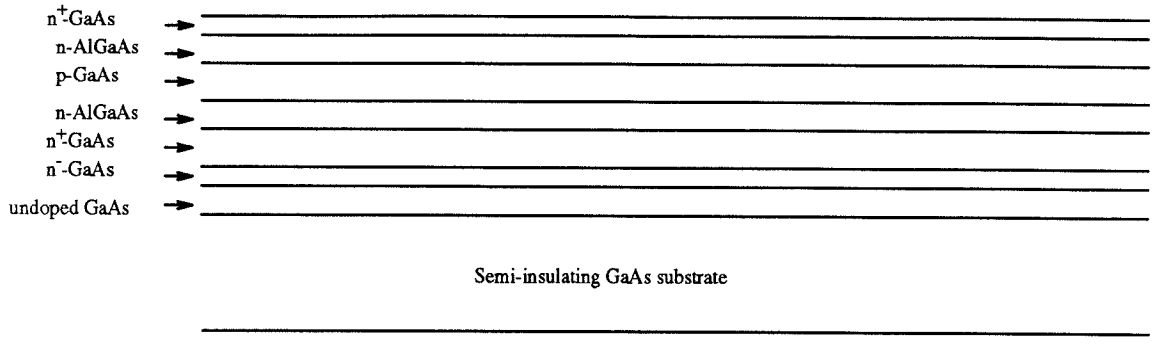


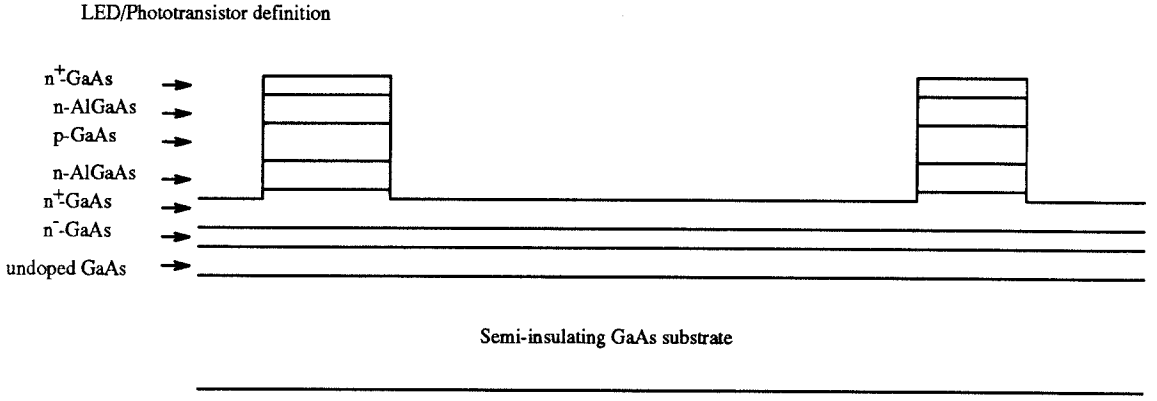
Fig. 5.17 The cross sectional view of the MESFET-based optoelectronic neuron monolithically integrating 2 MESFET's, a LED and a phototransistor.

GaAs	n = 5×10^{18} cm^{-3}	0.15 μm
$Al_{0.35}Ga_{0.65}As$	n = 2×10^{17} cm^{-3}	0.3 μm
GaAs	p = 1×10^{16} cm^{-3}	1.5 μm
$Al_{0.35}Ga_{0.65}As$	n = 5×10^{16} cm^{-3}	0.6 μm
GaAs	n = 5×10^{18} cm^{-3}	1 μm
GaAs	n = 2×10^{17} cm^{-3}	0.25 μm
GaAs	undoped	0.4 μm

Semi-insulating GaAs Substrate

Fig. 5.18 The epitaxial material composition of the MESFET-based optoelectronic neuron.





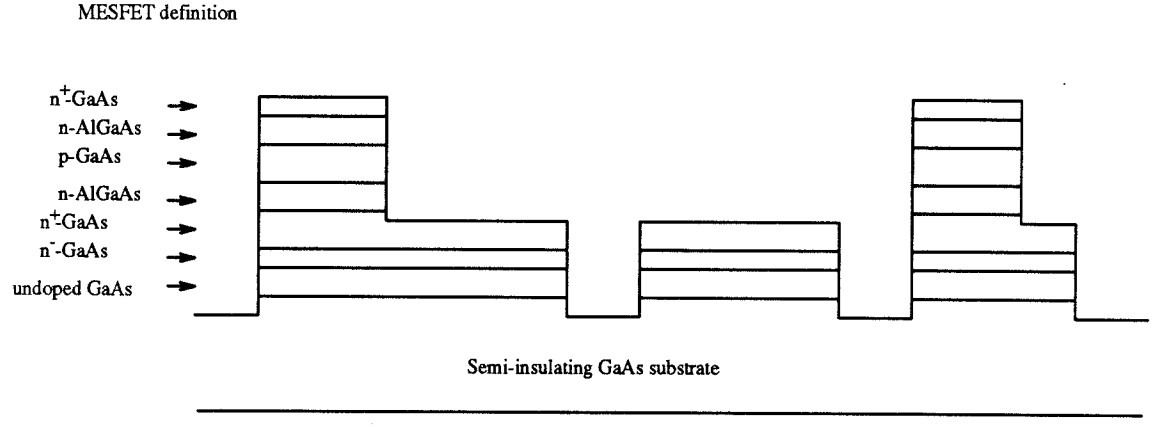


Fig. 5.19(a) The fabrication steps of the MESFET-based optoelectronic neuron.

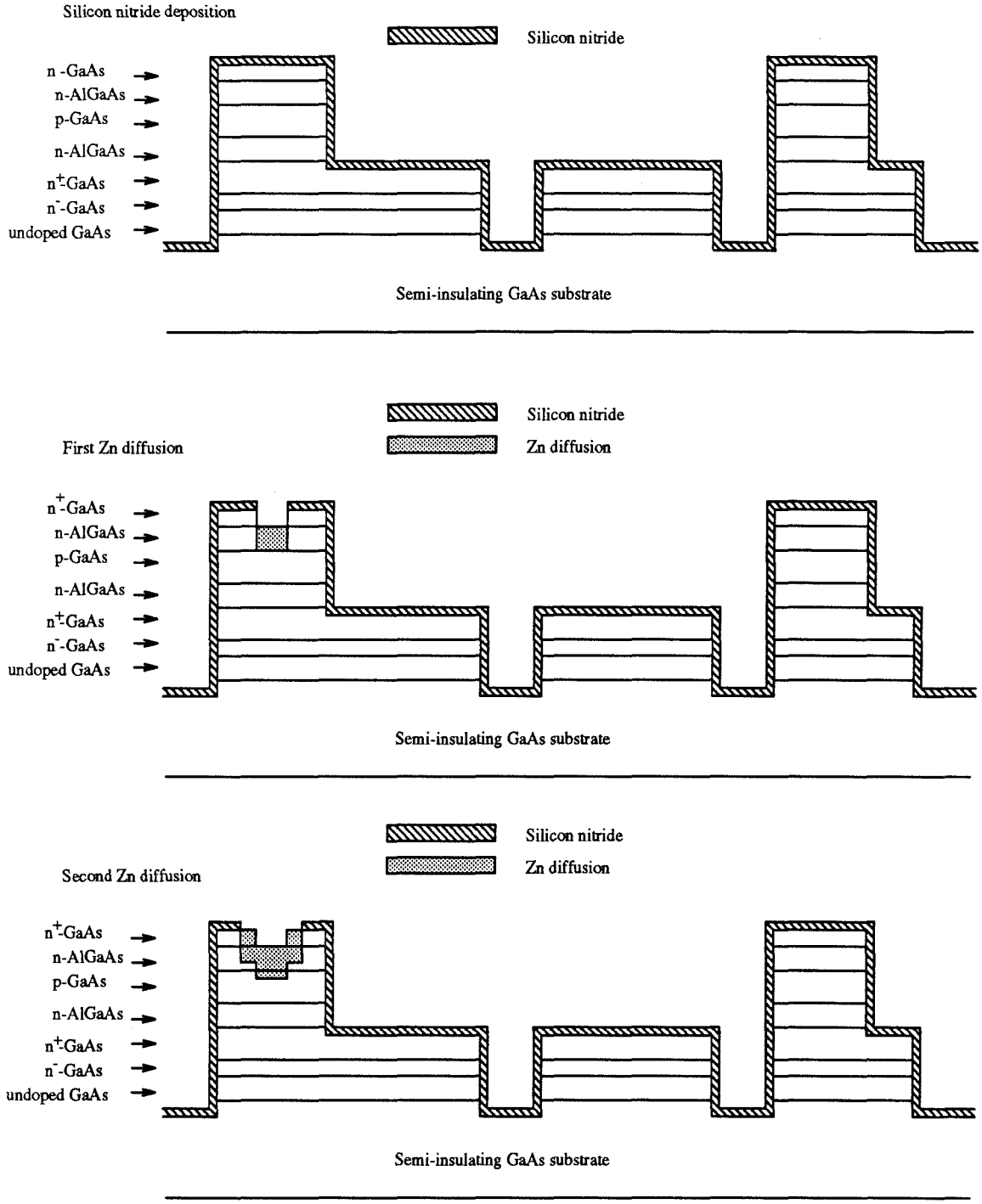


Fig. 5.19(b) The fabrication steps of the MESFET-based optoelectronic neuron.

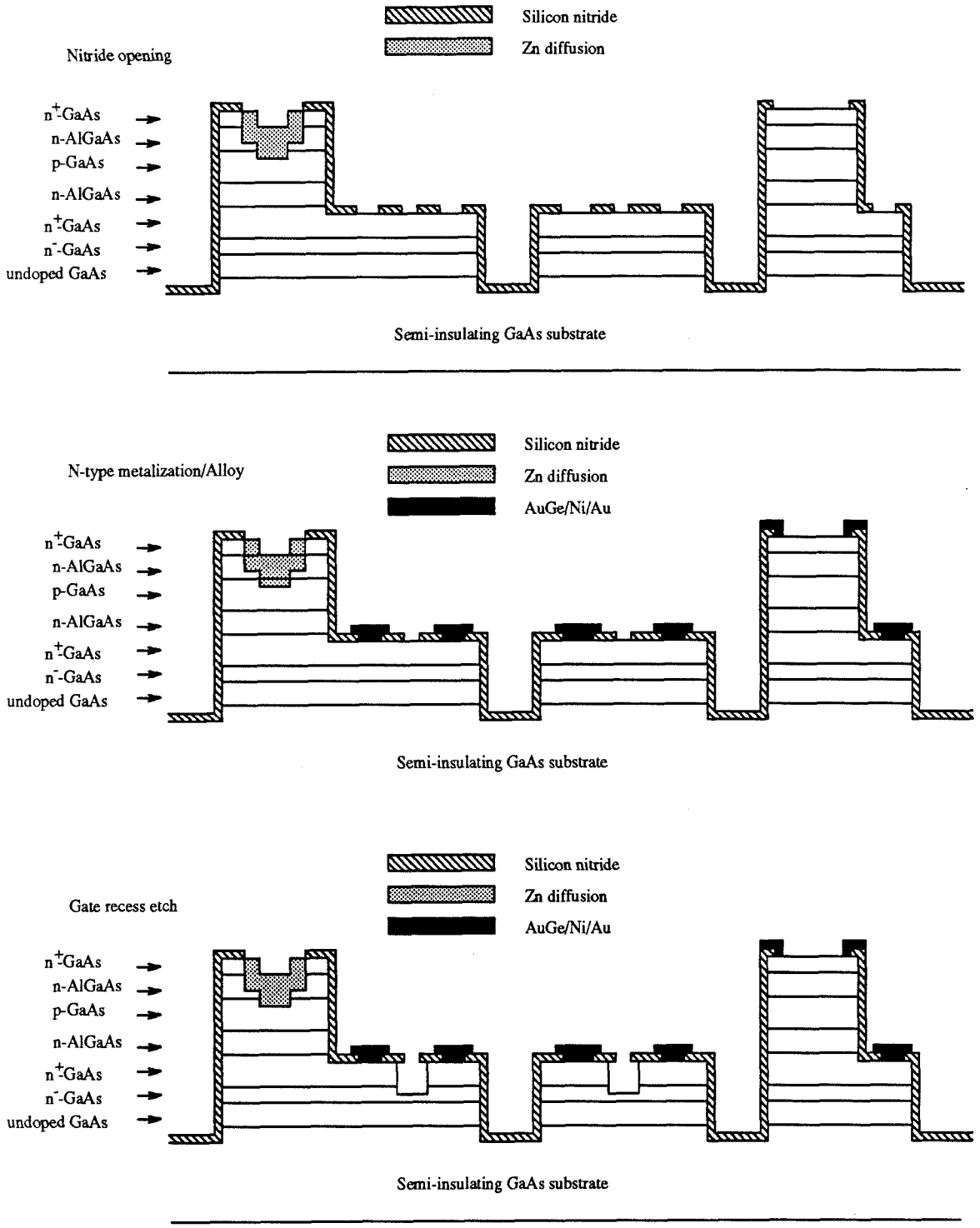


Fig. 5.19(c) The fabrication steps of the MESFET-based optoelectronic neuron.

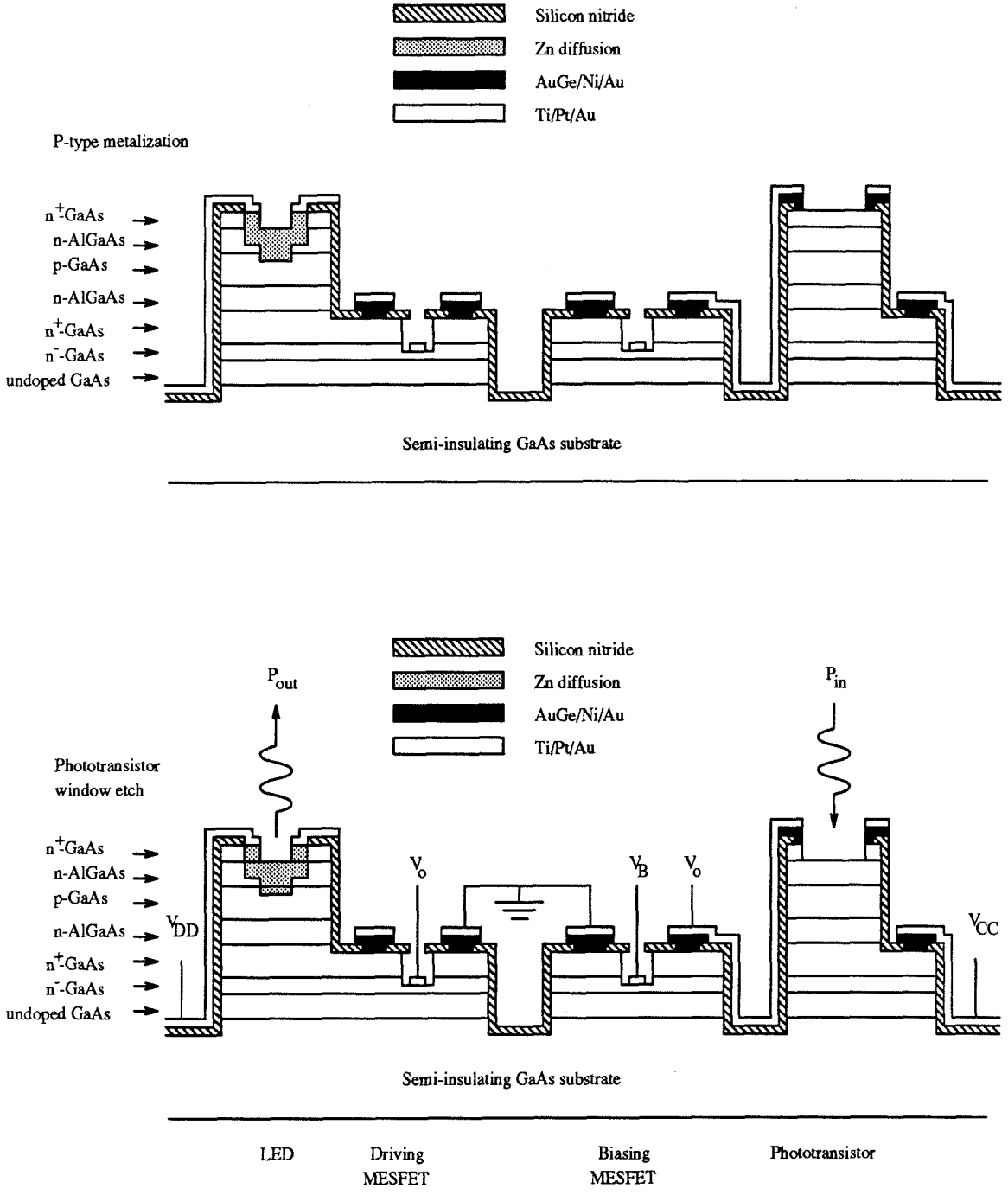


Fig. 5.19(d) The fabrication steps of the MESFET-based optoelectronic neuron.

into the exposed area of the LED for approximately 9 minutes. Afterwards, the ampoule was quickly quenched. Usually there was As condensation on the inner wall of the ampoule after quenching to indicate the proper diffusion of Zn into the LED. A second diffusion process was carried out by using exactly the same procedure except the diffusion time was approximately 5 minutes and the area of diffusion was slightly larger. After the Zn-diffusion step, selective area of Si_3N_4 was again removed in CF_4 plasma to facilitate the subsequent ohmic contacts for the source and the drain as well as the gate recessed area of the MESFET, and the contacts for the emitter and the collector of the phototransistor. The transistor contact terminals, including the source and the drain, and the emitter and the collector of the phototransistor, were metalized by evaporating AuGe/Ni/Au of 200Å, 100 Å, and 1500 Å, respectively, by using the lift-off technique and subsequently alloyed at 430 ^oC in an N_2 ambient for 4 minutes to drive in the Ge in forming the ohmic contacts. The gate recess etching process was then performed by monitoring the source-drain current in the MESFET. This etching used the Si_3N_4 as the mask in order to obtain a self-aligned recess. The etching was stopped when the desired source-drain saturation was obtained. The etchants used in recessing the gate was NH_4OH, H_2O_2 and H_2O in a ratio of 20, 7 and 973 respectively. The etch rate was approximately 30 Å/second. Next, the gate was defined by evaporating first 150 Å of Ti and then 1500 Å of Au by an electron beam evaporator. The excess metals were lifted off in acetone. The area of the gate was 6 $\times$ 100 μm^2 and was self-aligned asymmetrically to the edge of the source inside the recessed region. The last step was to remove the light-absorbing n^+-GaAs cap layer in the phototransistor by wet etching. The entire processing of the MESFET-based optoelectronic neuron utilized 9 masks and took about 2 weeks to complete. Figure 5.20 shows the photograph of a complete neuron. The entire area, including the contact pads, measured approximately 400 $\times$ 400 μm^2. However, the active device area was only about 150 $\times$ 250 μm^2.

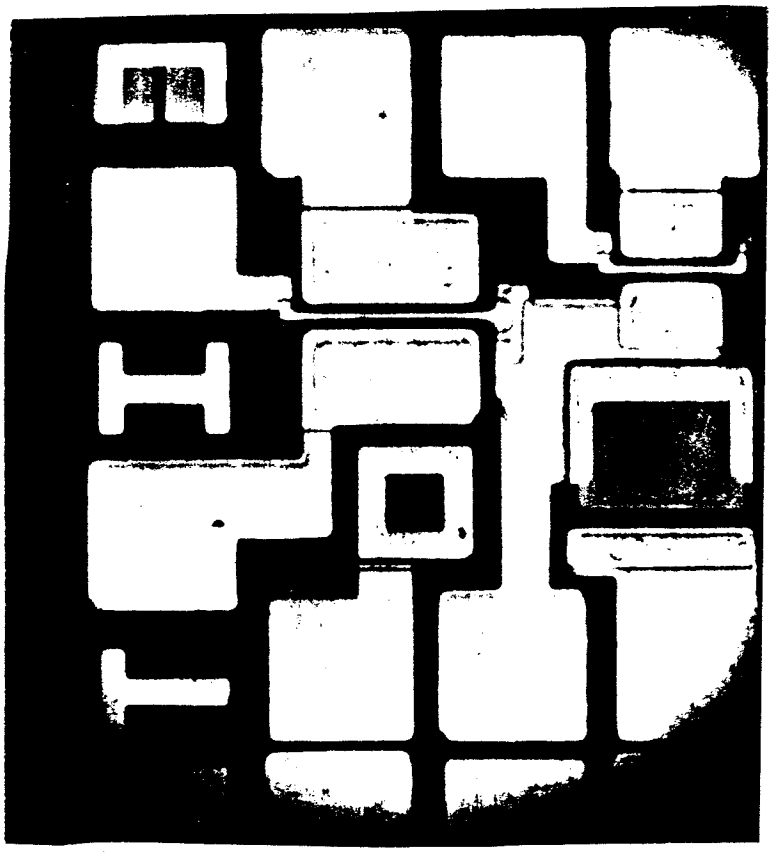


Fig. 5.20 Photomicrograph of a completely fabricated optoelectronic neuron. The input switching circuit is on the right side of the picture and the output driving circuit is on the left side of the picture. The bottom left square is the LED, which is monolithically connected to the drain of the driving MESFET. The gate of the same MESFET is controlled by the combination of the phototransistor located at the bottom right corner of the picture and the loading MESFET, which is located right above the phototransistor. The windows of the LED and the phototransistor are 40 $\times$ 40 μm^2 and 60 $\times$ 80 μm^2 respectively.

The neuron was tested by illuminating the phototransistor window area by a GaAs laser diode. This was achieved by splitting the output beam of the laser diode by a beam splitter into 2 equal-intensity beams. One of the beams was focused onto the phototransistor and the other beam was focused into the detector in order to monitor the power of the beam incident on the phototransistor. The input switching circuit was first tested. The voltage between the phototransistor and the loading MESFET, V_{DS1}, was monitored as the intensity of the input laser beam was varied. A power supply of 2.5 V was connected to the collector of the phototransistor, while the source of the MESFET was electrically grounded. By varying the gate voltage of the MESFET, V_{DS1} was measured at a fixed laser output power incident on the phototransistor. Then the laser output power was changed and the same measurement was performed again. The results are shown in Fig. 5.21, in which V_{DS1} is plotted against V_B for five different laser powers incident on the phototransistor. For a laser power of 10.8 μW incident on the phototransistor, the voltage, V_{DS1}, was pulled up to the power supply voltage less the phototransistor saturation voltage. As the gate voltage of the MESFET increased, V_{DS1} stayed relatively unchanged until the current drawn by the MESFET had exceeded the photocurrent provided by the phototransistor. At which point, V_{DS1} dropped and was pulled down to ground. As the laser power became smaller, the value of the gate voltage at which V_{DS1} dropped from 2.5 V to ground decreased. This was in consistency with the analysis shown earlier because as the photocurrent became smaller, the current needed by the loading MESFET to pull down V_{DS1} also became smaller. However, due to the leakage current from the drain to the gate of the MESFET, V_{DS1} could not be pulled up completely. In fact, as the gate bias decreased, this leakage current increased because the gate-drain had became more reverse biased. Since this leakage current was no different in nature compared to the source-drain current drawn by the MESFET from the standpoint of V_{DS1}, V_{DS1} was pulled

down as a result. This imperfection was clearly evidenced for laser input power of 3.2 μW and less. Thus, it was extremely important that the leakage current be minimized in the MESFET, especially the one across the gate-drain terminals.

The overall input-output characteristics of the optoelectronic neuron was obtained by monitoring the current through the LED as a function of the laser input power incident on the phototransistor at a fixed gate voltage on the loading MESFET. Figure 5.22 shows two of these plots. One of them was taken at a gate voltage of -3.0 V and the other one was taken at a gate voltage of -2.4 V. The measured LED current was converted into optical output power by assuming an external quantum efficiency of the LED to be 0.01 W/A. The reason the output power of the LED could not be directly measured was because the beam of the LED diverged too fast and it was difficult to collect all of it into the detector. If, however, we brought the detector very close to the neuron, the input laser beam could not easily illuminate the phototransistor. Thus, the current through the LED was monitored. The external efficiency of 0.01 W/A was typical for the LED of double Zn-diffusions, as experimentally verified at the end of Ch. 2. For the curve with a gate voltage of -3.0 V, the output remained zero until the input power reached approximately 3 μW. Beyond this point, the output power increased rapidly to 12 - 15 μW over an additional input of 2 μW. This represented a differential optical gain of 6. The threshold of the neuron was controlled by applying a different voltage to the gate of the loading MESFET, as clearly seen in the plot. Because of the leakage in both the loading MESFET and the output driving MESFET, a minimum 3 μW was necessary to turn on the neuron. By reducing the leakage currents through the gate in both MESFET's, this number is expected to drop substantially. During the on-state of the neuron, the LED current was measured to be 1.2 mA. This implied an electrical power dissipation of 2.4 mW by using a 2-volt power supply on the the output driving circuit. When the input laser beam was pulsed to a level just

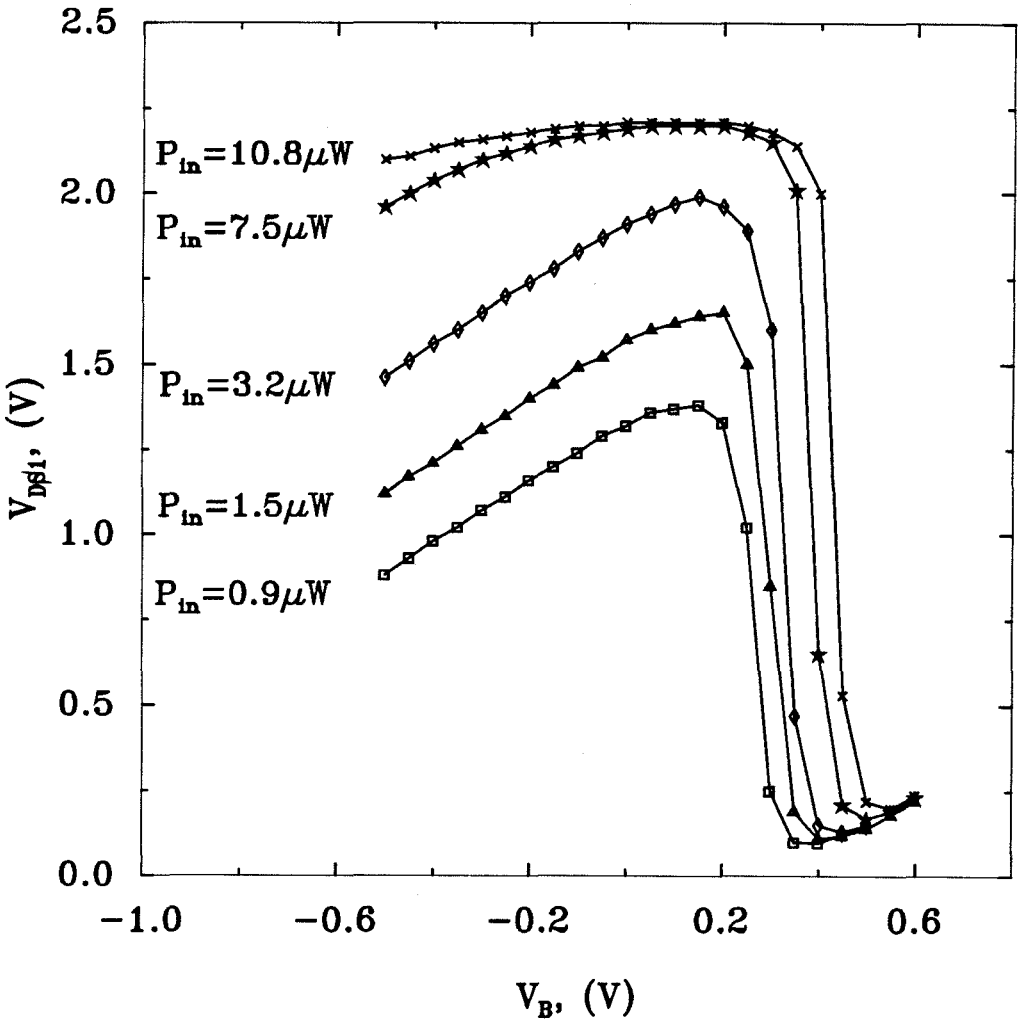


Fig. 5.21 V_{DS1} as a function of the gate voltage on the loading MESFET for various input laser power incident on the phototransistor.

enough to turn on the neuron, the output LED current showed a rise time of 5 μsec. This is shown in Fig. 5.23. This meant that the neuron could be turned on with an optical switching power of only (2 μW) $\times$ (5 μsec) = 10 pJ [97]. Not only did this MESFET-based optoelectronic neuron exhibit comparable switching energy as compared to the SEED devices [96], it also dissipated only 2.4 mW of electrical power. This was a factor of 40 less when compared to the DHBT-based neuron. Figure 5.24 shows the picture of the LED being lit up electrically by an external bias voltage.

The differential optical gain of 6 was limited by the finite output impedance of the phototransistor and the loading MESFET as well as the leakage currents in the MESFET's. The output impedance of the phototransistor could be increased by doping the base more heavily. However, this reduced the current gain of the phototransistor. As a result, the base thickness had to be reduced to compensate for the increased doping concentration in order to maintain the same current gain. Unfortunately, reducing the thickness of the base layer adversely affected the absorption efficiency of the phototransistor. Therefore, an optimized design had to be used. The output impedance of the loading MESFET could be increased by using a more insulating substrate as well as reducing the leakage current through the gate. It was interesting to note that reducing the leakage current has a lot of benefits in terms of improving the optical gain and the sensitivity of the neuron. Thus, the same MESFET-based optoelectronic neuron was fabricated again by carefully cleaning the surface before the gate metalization was defined. Furthermore, a different gate metalization composition was employed. This consisted of the same Ti/Au metals except an 100-Å layer of Pt was inserted between the Ti and the Au. The doping concentration of the MESFET conduction layer was also reduced to 5×10^{16} cm^{-3} for less leakage current across the gate. Figures 5.25 - 5.27 show the results of the neuron, which incorporated the above-mentioned simple changes [98].

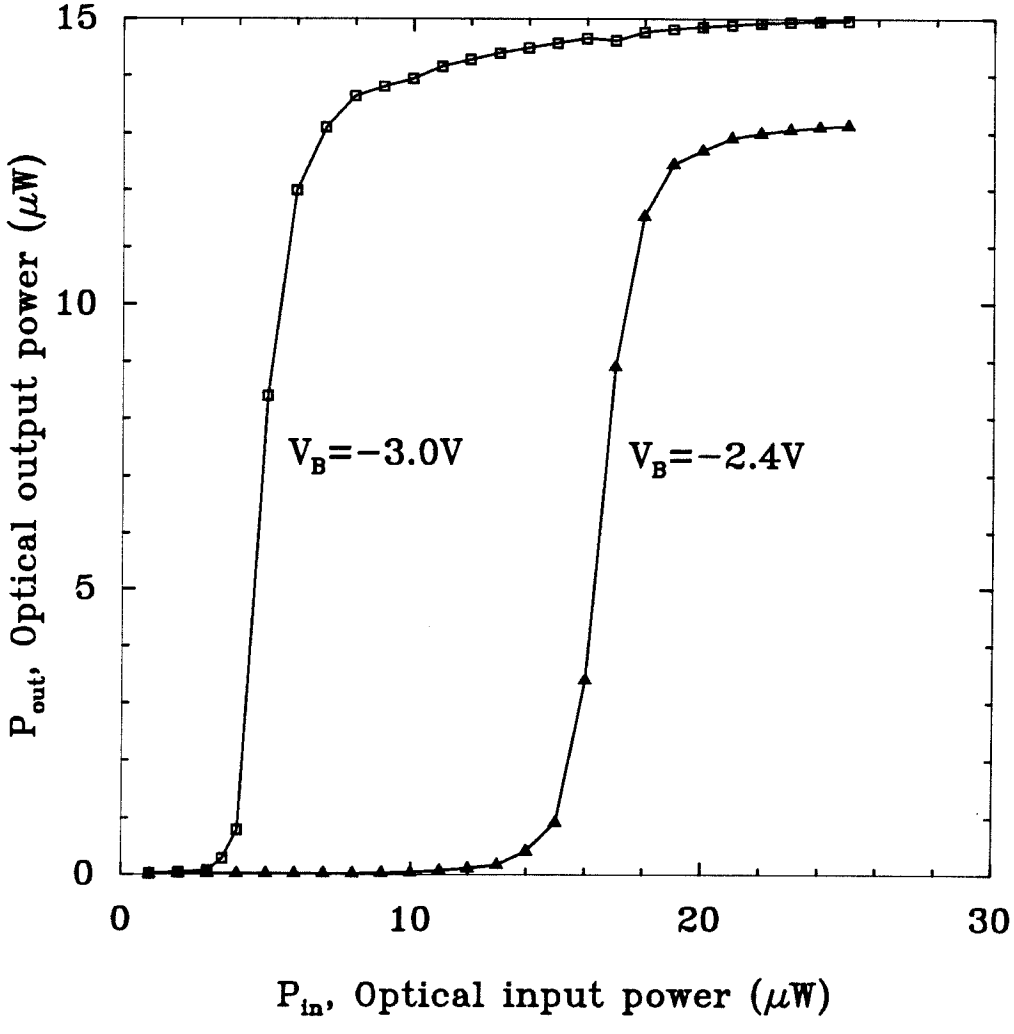


Fig. 5.22 Input-output characteristics of the MESFET-based optoelectronic neuron. V_B is the bias voltage on the gate of the input switching circuit.

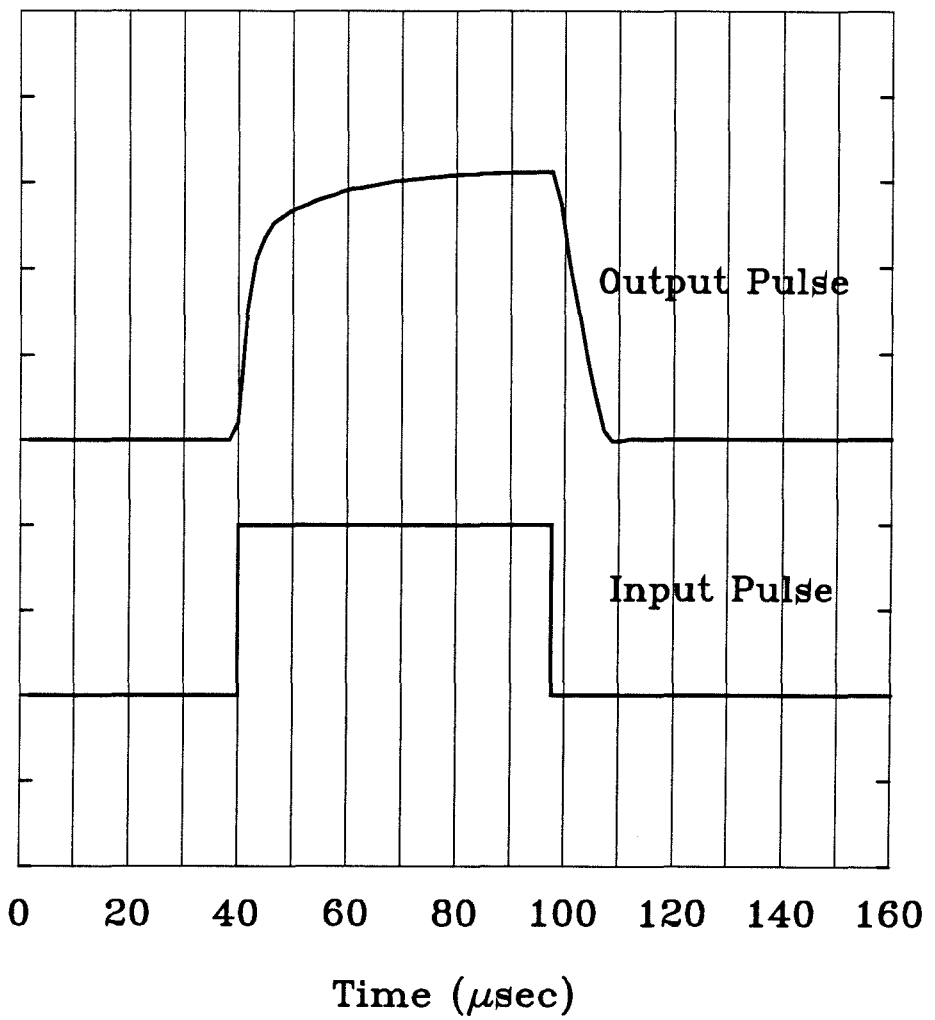


Fig. 5.23 Time measurement of the MESFET-based optoelectronic neuron in response to a step input in the laser power incident on the phototransistor. The rise time was measured to be 5 μsec.

Fig. 5.24 Photograph of the MESFET-based optoelectronic neuron, showing the LED being lit up electrically.

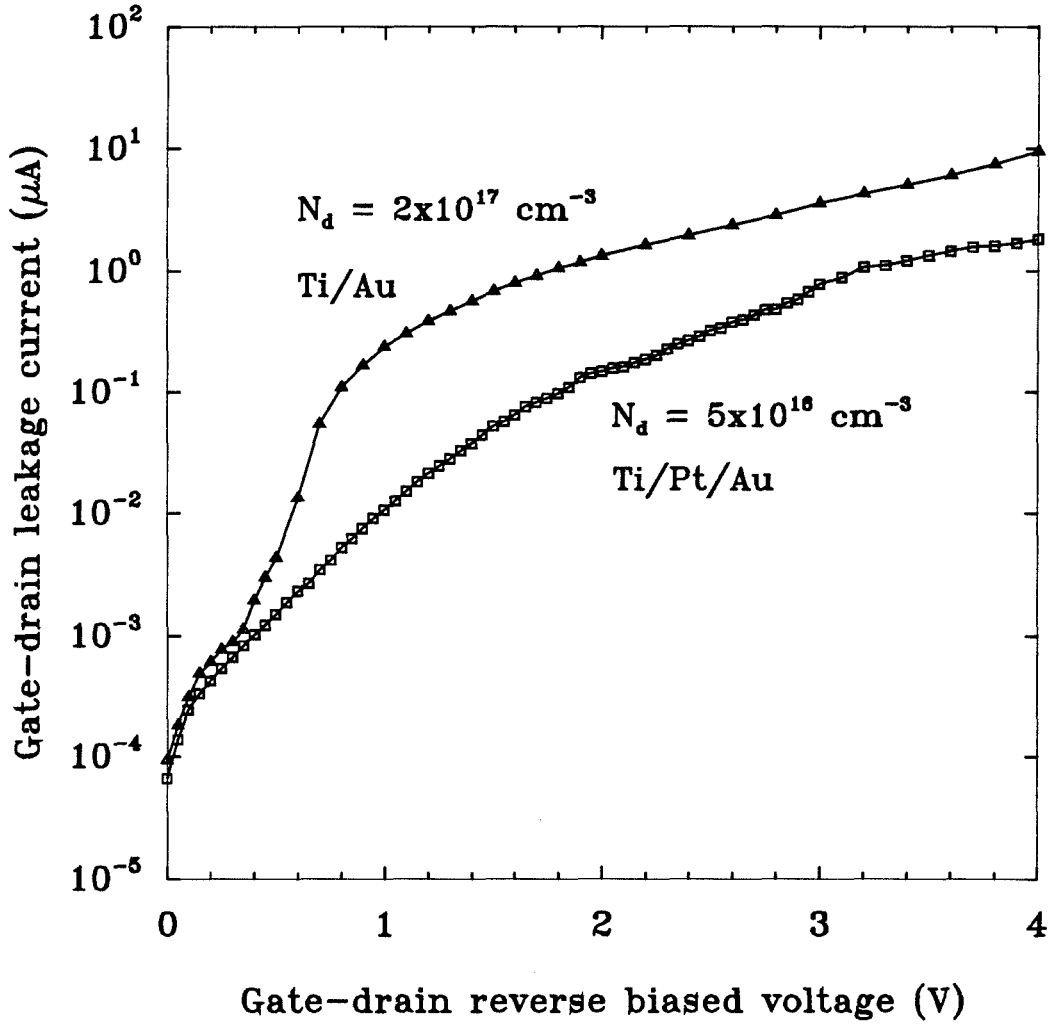


Fig. 5.25 Comparison of the gate-drain leakage currents as a function of the gate-drain reverse biased voltage in the loading MESFET for the old and the new neurons.

First of all, a comparison of the gate leakage current in the old and the new neurons was made. Figure 25 shows the measured gate-drain reverse leakage current as a function of the reverse biased voltage for the two Schottky diodes. At a typical operating voltage of 1 volt across the gate and the drain, the leakage current for the new MESFET was at least an order of magnitude lower. This reduction in the leakage current directly translated into an increase in the optical gain of the neuron because the phototransistor needed a smaller photocurrent to pull up the gate voltage of the output driving MESFET. Figure 5.26 shows the input-output relationship of the improved MESFET-based optoelectronic neuron. The testing conditions were the same as before except the gate of the loading MESFET was floating. This was intended to reduce the gate-drain leakage current further. It is evident from the plot that, by reducing the gate-drain leakage current, the minimum input power needed to turn on the neuron was reduced. In this case, an 1 μW input power was measured. Moreover, the differential optical gain also increased dramatically to 40 as an additional input power of 0.2 μW beyond the threshold caused a change of 8 μW at the output. This improvement was remarkable considering the only improvement made was to reduce the gate leakage current of the loading MESFET. Not only was the differential optical gain improved, but also the absolute optical gain had increased to 8. The current drawn by the LED during the on-state of the neuron was measured to be 0.8 mA. Therefore, the electrical power dissipation per neuron was 1.6 mW by using a 2-volt power supply. The speed of the neuron was also measured by applying an electrical pulse to the laser diode that illuminated the phototransistor. Figure 5.27 shows the measured response of the neuron. A rise time of 65 μsec was obtained in this neuron. This implied a total optical switching energy of (65 μsec $\times$ 0.2 μW) = 13 pJ. This optical switching energy was comparable to that of the previous neuron, which was 10 pJ. This is expected because the total charges needed to charge up the gate of the output

driving MESFET's, which had the same gate width and length, in both neurons were the same. Since the voltage swings at the same gate from the off-state to the on-state of the neuron were also the same, the switching energy, which was equal to QV, remained unchanged. Thus, overall, the neuron became more sensitive and provided more gain. However, this was achieved at the expense of a lower switching speed.

As the leakage current problem was improved, the limiting factor in the performance of the MESFET-based optoelectronic neurons shifted to the efficiency of the detector, which, in this case, was the double-heterojunction bipolar phototransistor. Because of the relatively thick base layer, the current gain, β, of the phototransistor was only about 3. This current gain dropped further as the input power level was reduced to the sub-μW regime. Since the overall goal of the MESFET-based optoelectronic neuron was to achieve a high-gain optical thresholding device at a low input power level, high-efficiency or even high-gain detectors at low input power level was vital to the success of the neuron. For this reason, optical FET's were developed. The operational principle of the optical FET was described in Sec. 5.3.3. In addition to the inherent high optical gain achievable in the optical FET, the structure of the optical FET was identical to that of a conventional MESFET. This meant that, for our MESFET-based neurons, optical FET's could be easily implemented into the existing material and process. This was a very important advantage of having the optical FET.

The fabricational steps of the MESFET-based neuron incorporating the optical FET as the detector were very similar to those of the conventional MESFET-based neuron. The difference was the definition of the optical FET rather than the phototransistor. Figure 5.28(a)-(d) show the sequential fabrication steps of the new neuron with Fig. 5.28(d) illustrating the entire device cross section of the neuron.

This neuron was tested at the same conditions as the previous one. Again,

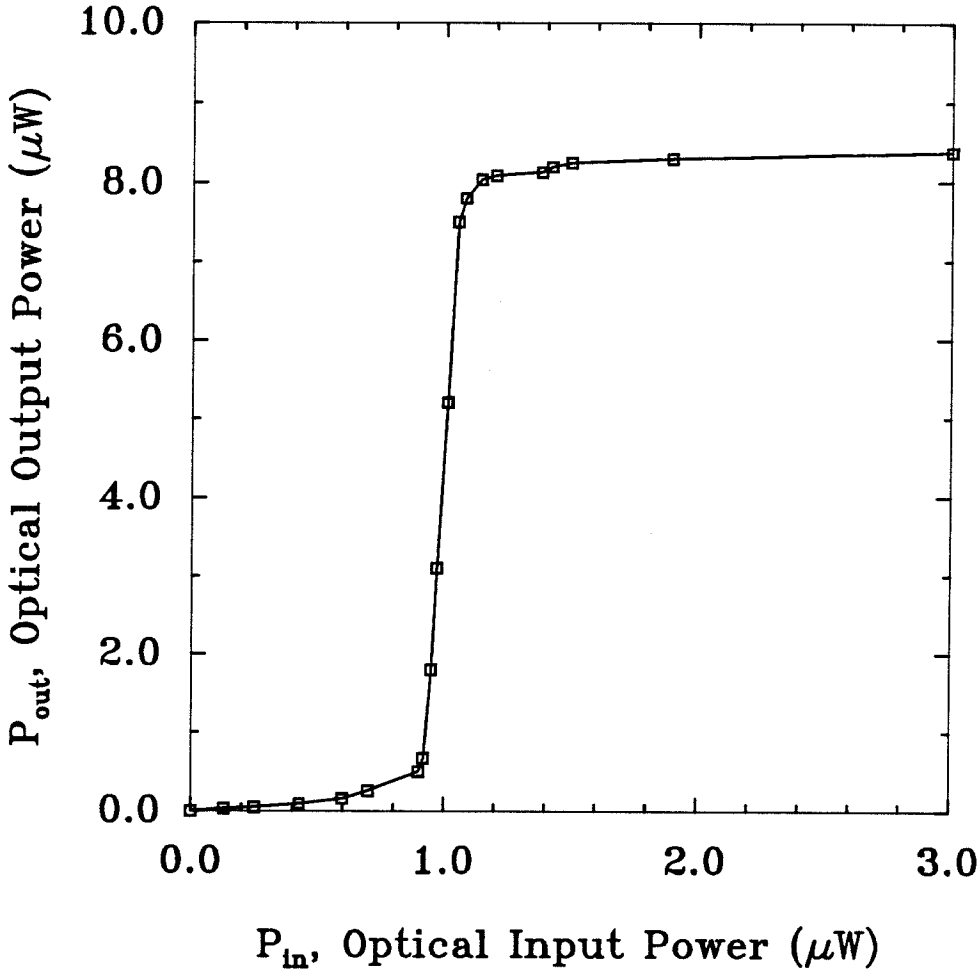


Fig. 5.26 Input-output characteristics of the improved MESFET-based optoelectronic neuron, showing a differential optical gain of 40 and an absolute optical gain of 8.

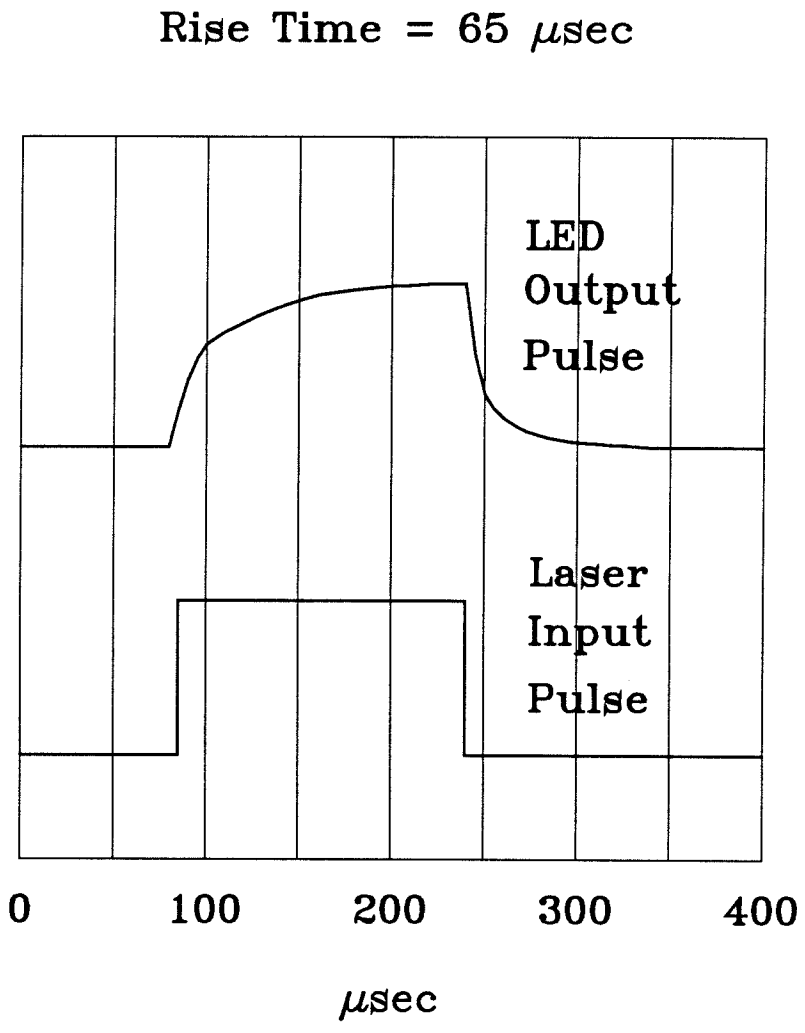


Fig. 5.27 Time response of the improved MESFET-based optoelectronic neuron. The rise time was measured to be 65 μsec.

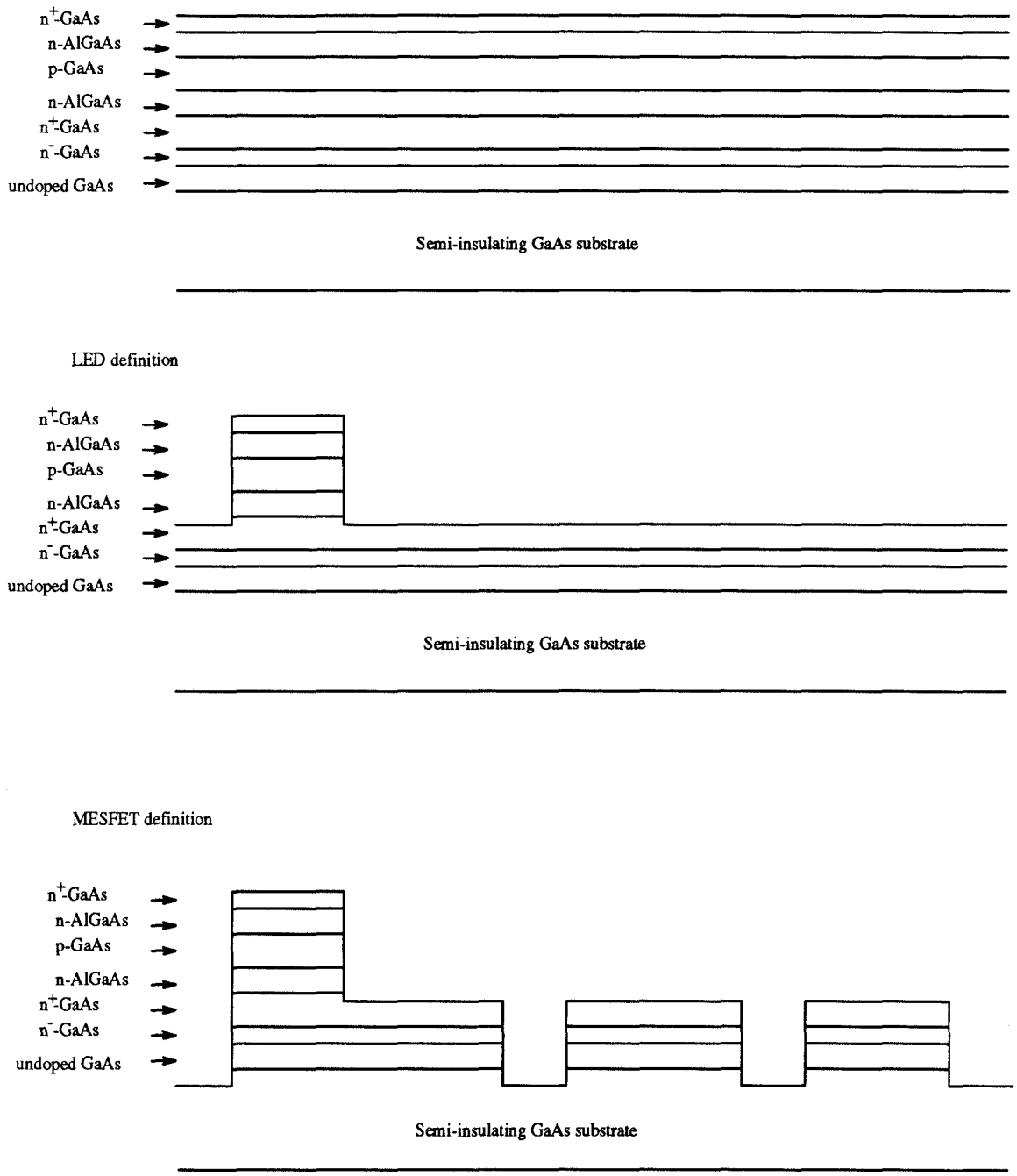


Fig. 5.28(a) Fabricational steps of the new MESFET-based optoelectronic neuron incorporating the optical FET as the detector.

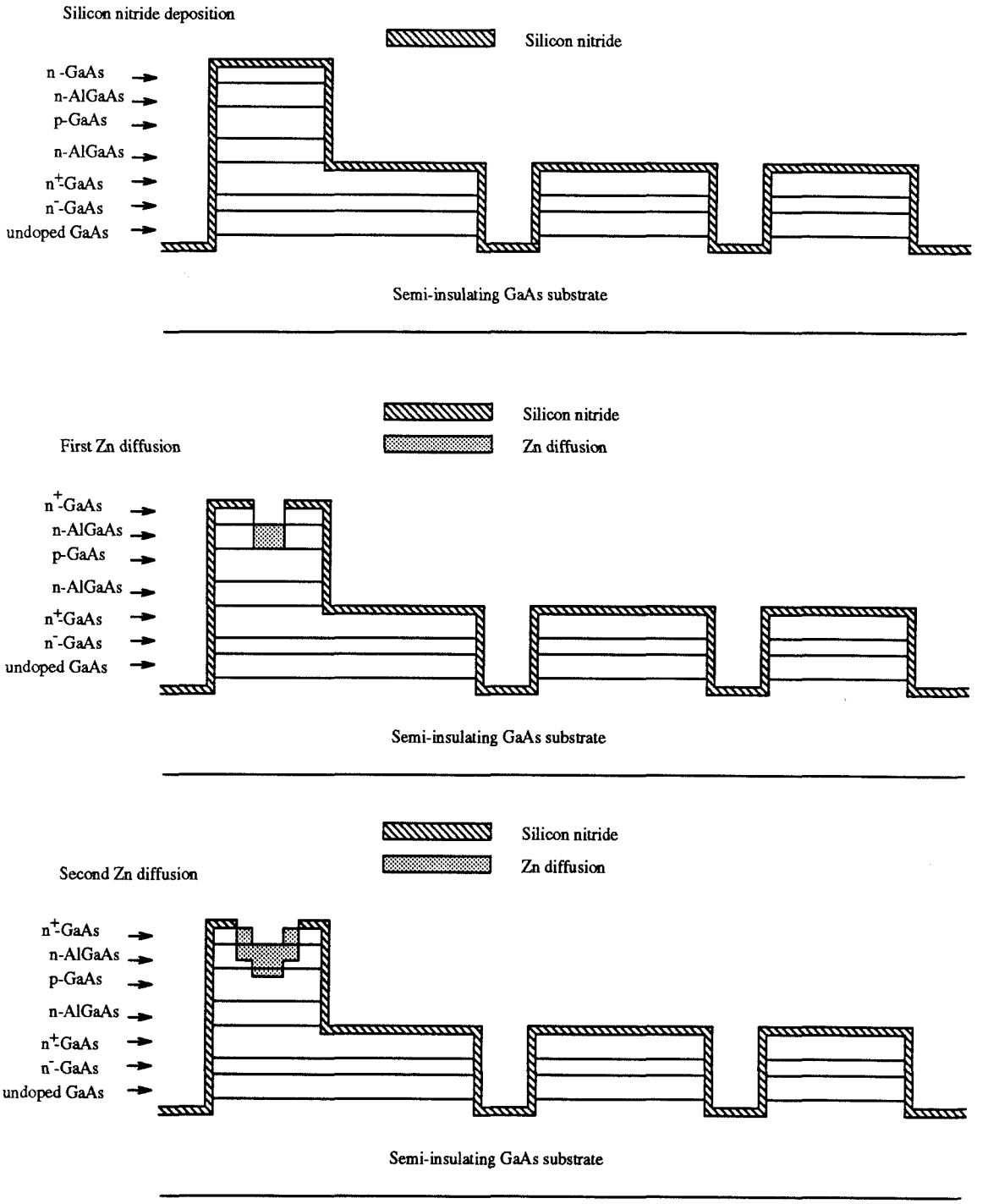


Fig. 5.28(b) Fabricational steps of the new MESFET-based optoelectronic neuron incorporating the optical FET as the detector.

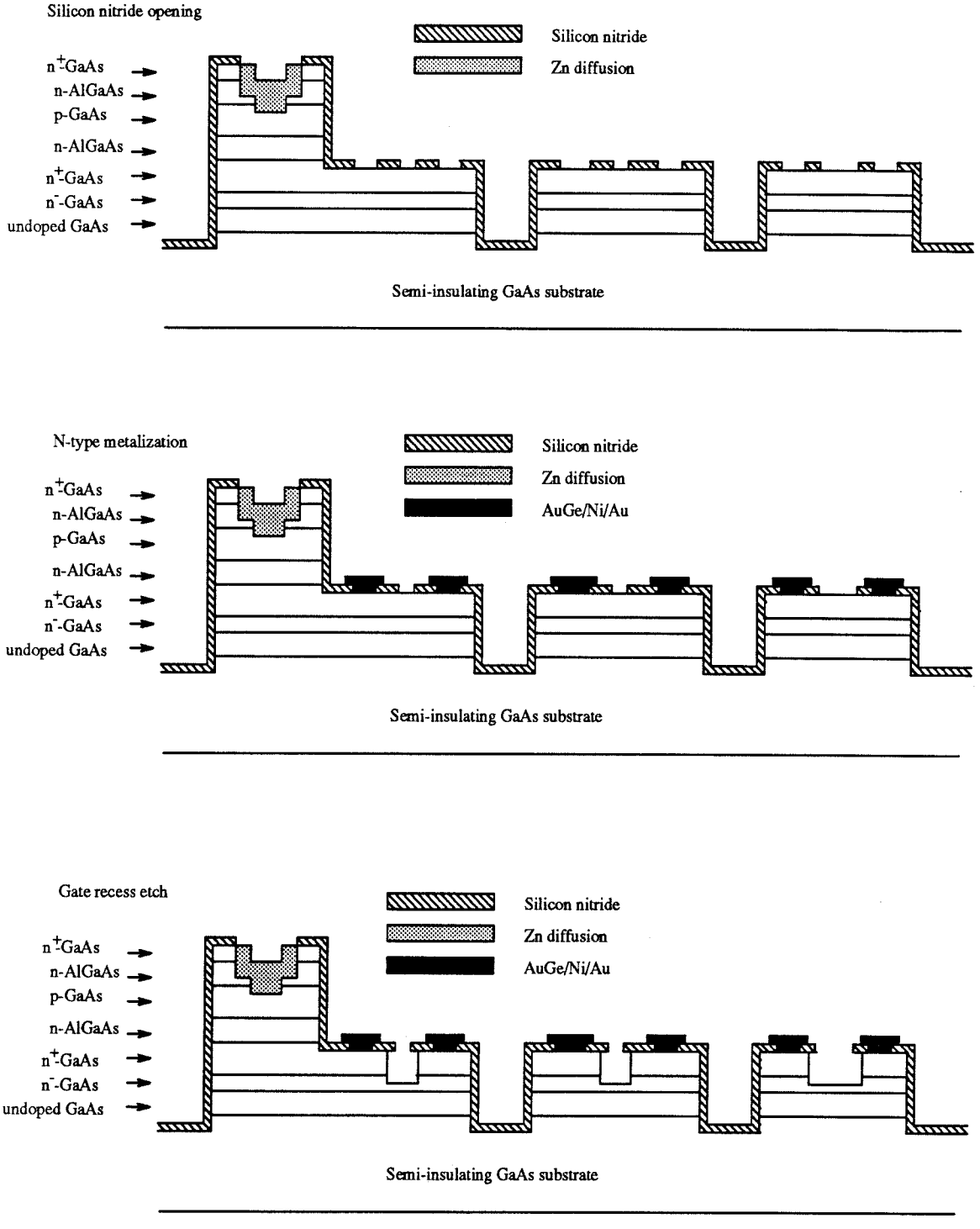


Fig. 5.28(c) Fabricational steps of the new MESFET-based optoelectronic neuron incorporating the optical FET as the detector.

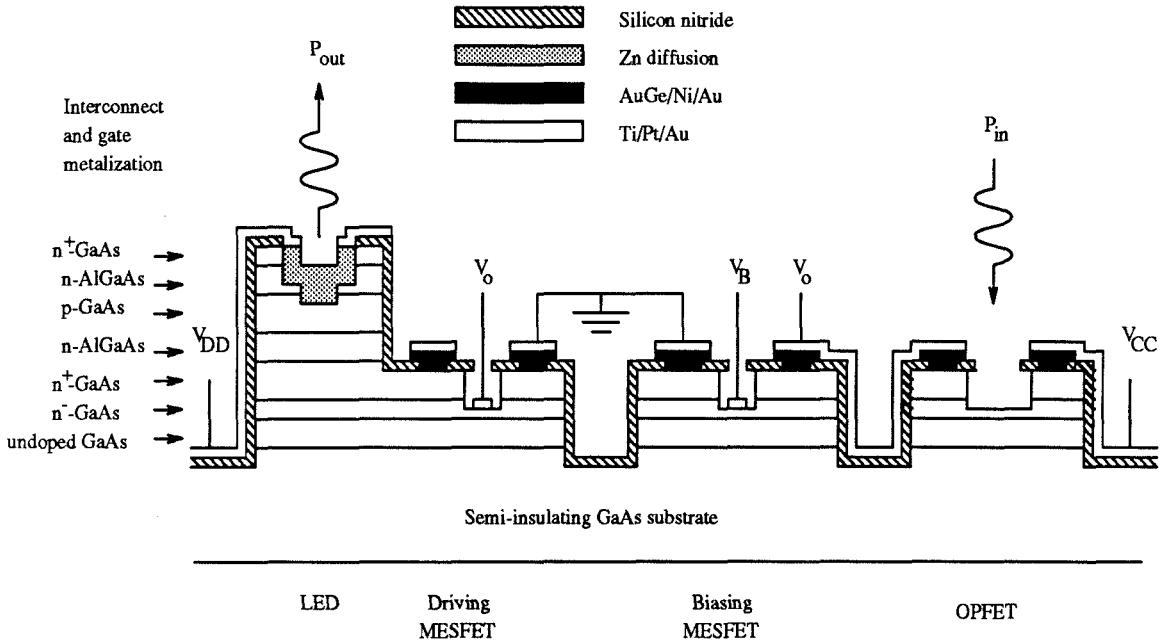


Fig. 5.28(d) Complete device cross sectional view of the new MESFET-based optoelectronic neuron incorporating the optical FET as the detector.

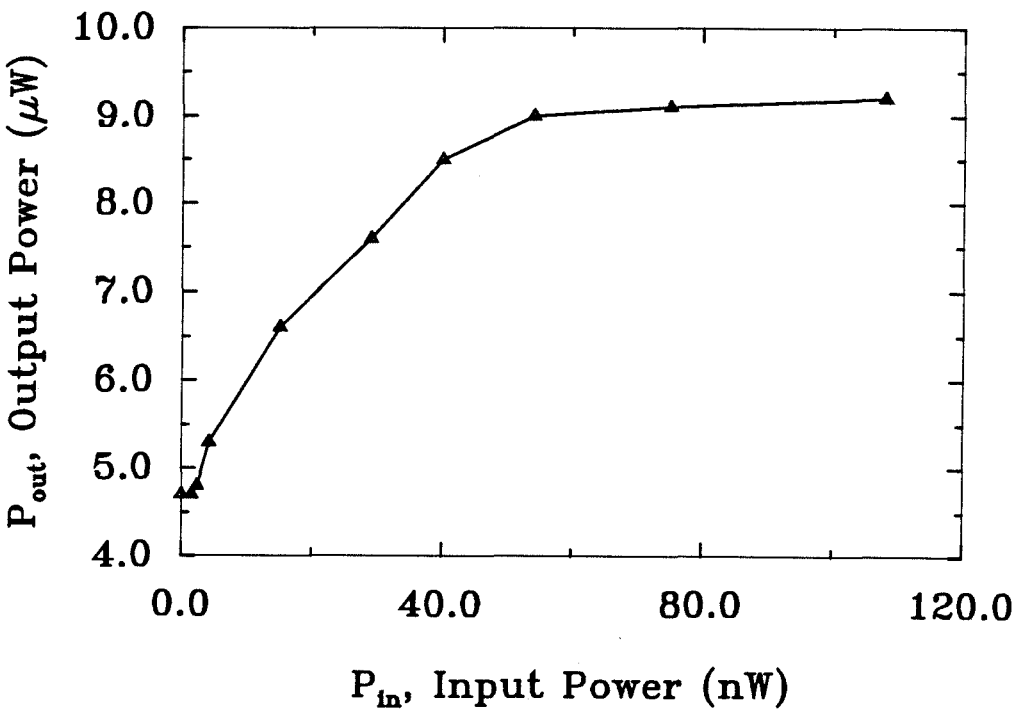


Fig. 5.29 Input-output characteristics of the MESFET-based optoelectronic neuron that incorporated the optical FET in replacing the phototransistor. A differential optical gain of 80 was measured in this neuron.

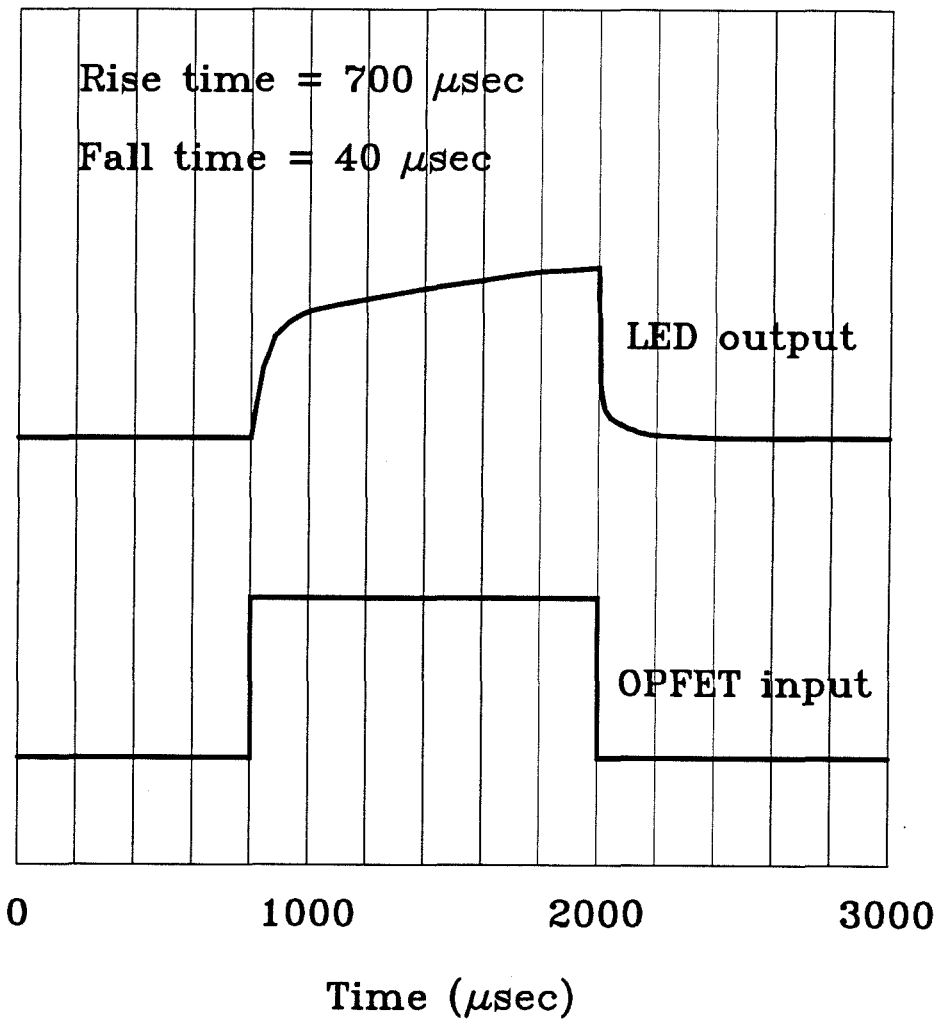


Fig. 5.30 Time response of the MESFET-based optoelectronic neuron incorporating the optical FET as the detector. The rise time was measured to be 700 μsec.

the gate of the loading MESFET in the input switching stage was left floating to minimize the gate leakage current. The optical input-output characteristics are shown in Fig. 5.29. Because of the insufficient recess in the gate of the output driving MESFET, this MESFET was not pinched off at zero gate bias. As a result, a current flowed between the source and the drain with zero input power onto the optical FET. This caused a non-zero LED output power at zero input power level. The remedy to this problem was to recess the gate of the LED-driving MESFET further until the current was close to zero at zero gate bias. This would shift the entire input-output curve shown in Fig. 5.29 down to the origin so that a normal neuron input-output characteristics could be obtained. Despite the gate recess problem, the differential optical gain measured was quite impressive. The output rose by 4.3 μW over an input swing of 54 nW. This corresponded to a differential optical gain of 80 [99]. It is also worth noting that the minimum input power needed to turn on the neuron had dropped significantly from the previous 1 μW down to about 5 nW. This could be attributed to the higher efficiency of the detector as well the overall reduction in the gate leakage current. Since this initial thresholding power was very small, the absolute optical gain was approximately the same as the differential optical gain, assuming the gate of the output MESFET was properly recessed. During the on-state of the neuron, the total current drawn by the LED was 0.9 mA, which implied an electrical power dissipation of 1.8 mW/neuron. Again, if the gate were properly recessed, this dissipation power would be reduced by 50%. The time response of the neuron was measured and is shown in Fig. 5.30. A rise time of 700 μsec was measured. When this was multiplied by the optical switching power of 54 nW, an optical switching energy of 38 pJ was obtained. Again, this was on the same order of magnitude as the previous optical switching energies. This indicated that the speed of the MESFET-based optoelectronic neurons was limited by the charging process of the gate capacitance and varied inversely proportional

with the input power level. Table 5.1 summarizes the results of the MESFET-based optoelectronic presented in this chapter.

5.5 Neuron Switching Characteristics

The input-output characteristics of the optoelectronic neuron critically depend on the performance of each of the devices in the integrated circuit. The ultimate performance of the neuron will be limited by the best performance that can be achieved in each of the elements. For the neuron that consists of a LED, a driving MESFET, a biasing MESFET and an optical FET as a detector, such as the one shown in Fig. 5.31(a), it is important that the input switching circuit, which is the combination of an optical FET and a MESFET in series, is able to provide the proper switching characteristics that are needed to drive the LED. In addition to the proper switching characteristics needed at the input, the output driver circuit has to meet several requirements in order for the whole integrated neuron to work properly. Firstly, the driving MESFET has to be in the enhancement mode (E-mode) so that there is no current flowing through the transistor during the off-state of the neuron. However, the same driving MESFET has to also provide enough amplification through it transconductance (g_m) so that a small change in its gate voltage will be sufficient to drive the LED, whose emitted optical power needs to be detectable by the detector after suffering the diffraction loss through the connection medium. Thus, getting the correct input-output characteristics of the optoelectronic neuron requires each of the device elements in the circuit performing at its designed level. Shown in Fig. 5.31(b) is the general input-output characteristics of the neuron. As can be seen, there are 3 distinct slopes in the transfer curve, $S1, S2$, and $S3$. The non-zero $S1$ and $S3$ are due to the non-zero saturation voltage from both the optical FET and the biasing MESFET. And the finite value of $S2$ is a result of the

	Phototransistor-based neuron	Phototransistor-based neuron with gate leakage reduction	Optical FET-based neuron
OUTPUT POWER	12 μW	8 μW	9 μW
DIFFERENTIAL OPTICAL GAIN	6	40	80
ABSOLUTE OPTICAL GAIN	2.5	8	150
RISE TIME	5 μsec	65 μsec	700 μsec
SWITCHING POWER	2 μW	0.2 μW	54 nW
MINIMUM THRESHOLD	3 μW	1 μW	5 nW
OPTICAL SWITCHING ENERGY	10 pJ	13 pJ	38 pJ
ELECTRICAL POWER DISSIPATION	2.4 mW	1.6 mW	1.8 mW

Table 5.1 Summary of the neuron characteristics for three versions of MESFET-based optoelectronic neurons discussed in this chapter.

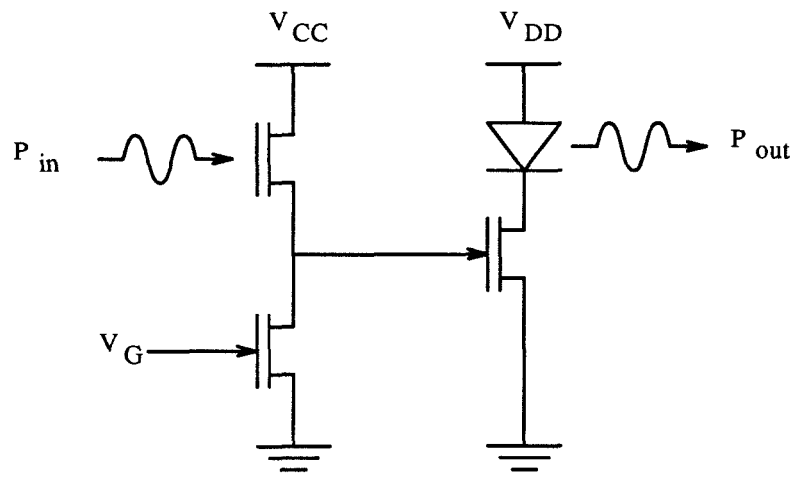
inability of the input switching circuit to switch instantaneously. The only effect that the output driving circuit has on the neuron input-output characteristic is the level of output power from the LED or point D in the figure.

Although the role of the output driver circuit has less of an impact on the neuron transfer characteristic curve compared to that of the input switching circuit, its functionality is simpler and almost linear in the gate voltage and LED output power relationship. Let's denote $\Delta V_g, g_m, \eta_{LED}$ to be the change in the gate voltage due to the input light incident on the optical FET in volts, transconductance of the driving MESFET in amp/volt, and the external quantum efficiency of the LED in watts/amp, respectively. The differential optical output power from the LED can then be expressed by

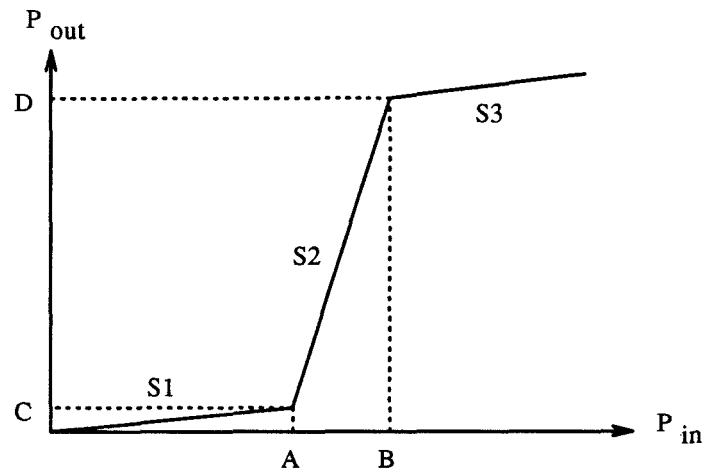
$$\Delta P_{out} = \eta_{LED} \cdot g_m \cdot \Delta V_g. \tag{5.3}$$

Thus, the bigger the change in the gate voltage is, the bigger the output power from the LED would be. This linear relationship is approximately valid as long as the leakage currents in the output driving MESFET are negligible compared to the current flowing through the LED.

The operation of the input switching circuit is not, however, so simple. The interaction between the detecting optical FET and its biasing MESFET is highly nonlinear. Since the gate voltage of the driving MESFET is the voltage at the node between these two input switching devices, it is important to understand to relationship between this voltage and the input power. This circuit is shown in Fig. 5.32(a). In principle, the voltage at this node, denoted by V_o, will be pulled up to V_{CC} if there is sufficient light incident on the detector, where V_{CC} is the power supply voltage for the two transistors. Otherwise, V_o should remain at 0 V, which forces the output driving MESFET to be in cutoff. The function of the gate



(a)



(b)

Fig. 5.31 **(a)** Integrated optoelectronic neuron consisting of a LED as the light source, a MESFET which drives the LED, and an optical FET in combination with another MESFET as the input switching circuit. **(b)** The input-output characteristics of the integrated optoelectronic neuron. The optical gain is determined by the slope, $S2$ and the threshold value, A, is electronically controlled by the voltage, V_G.

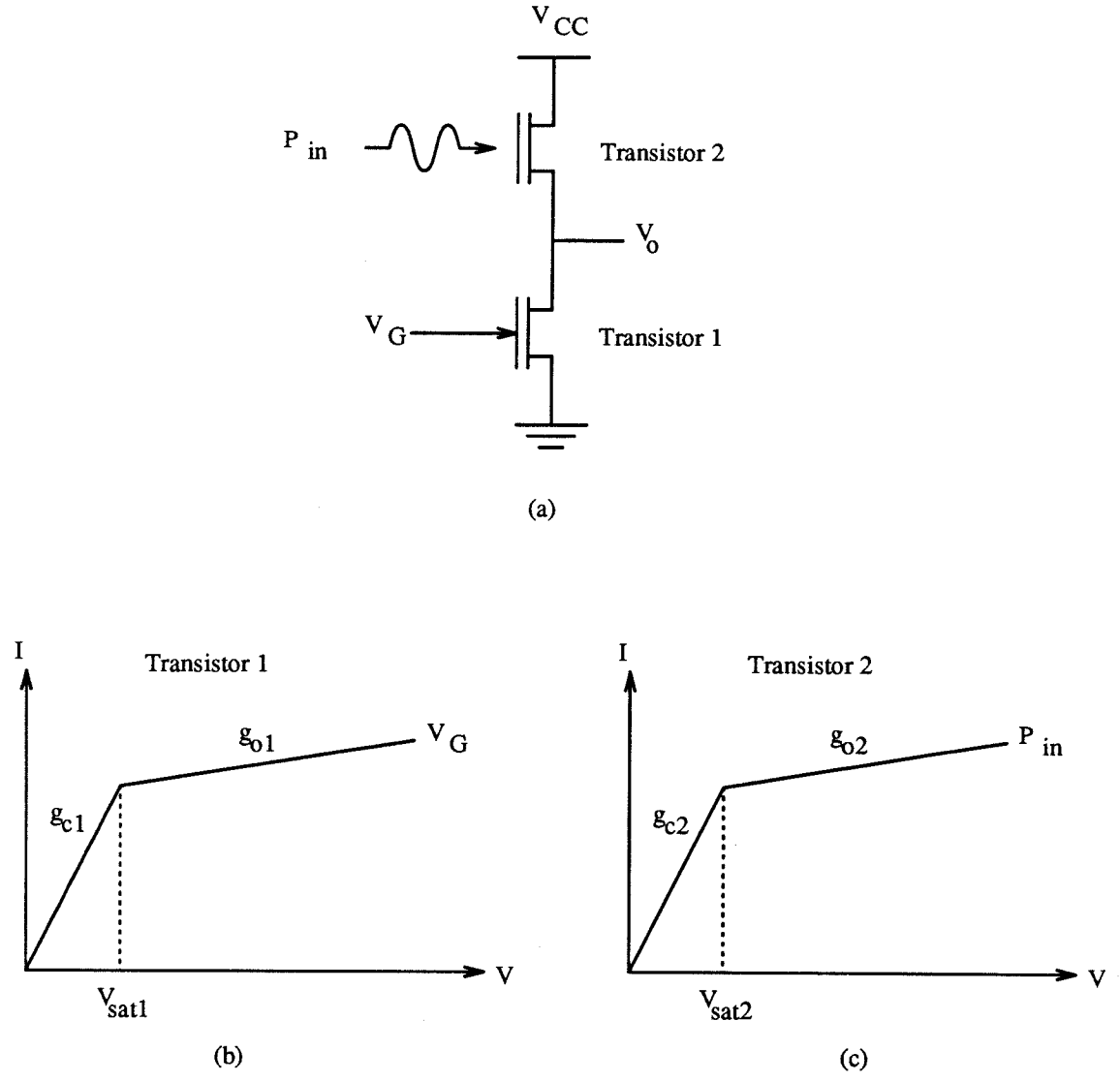


Fig. 5.32 **(a)** The input switching circuit of the neuron, which consists of an optical FET as the detector and a MESFET as an active load for the optical FET. The voltage, V_o, directly controls the the gate of the LED-driving MESFET. **(b)** The DC I-V characteristics of the MESFET at a given V_G. g_{c1} and g_{o1} are the conductances of the transistor before and after becoming saturated. **(c)** The DC I-V characteristics of the optical FET at a given P_{in}. g_{c2} and g_{o2} are the conductances of the transistor before and after becoming saturated.

voltage on the biasing MESFET, V_G, is to control the threshold for turning on the neuron. The higher V_G is , the higher P_{in} needs to be in order to turn on the neuron because the detector needs to produce more current to satisfy the current drawn by the biasing MESFET. Since the LED output power is linearly proportional to the change in the gate voltage, ΔV_g, of the driving MESFET, which is the same as the change in V_o, we will restrict our analysis to the change in V_o as a function of the input power, P_{in}. It should be noted that the maximum change in V_o would be V_{CC} since V_o is restricted between ground and V_{CC}. In order to understand and gain insight into the switching characteristics of this circuit, models of the optical FET and the MESFET have to be developed. Shown in Fig. 5.32(b) and 5.32(c) are the DC I-V curves for the MESFET and optical FET respectively. The channel and output conductances for both transistors are labeled as g_{c1}, g_{c2}, g_{o1}, and g_{o2}, where transistor 1 is MESFET, transistor 2 is optical FET, and the subscripts 'c' and 'o' mean channel and output, respectively. It is important to relate these conductances, g_{c1}, g_{c2}, g_{o1}, and g_{o2} to actual external input parameters, such as V_G and P_{in}. Thus, we will first derive the theoretical expressions for these conductances and quantify their functional dependence on V_G and P_{in}.

Figure 5.33(a) shows the device model of an enhancement-mode MESFET. Because of the positive V_G applied onto the gate of the transistor, the originally cutoff transistor has a finite conductance now as a result of the depletion front, which has retracted toward the metal-semiconductor interface on the top. The conductance between the source and the drain of the MESFET depends on the dimensions of this conduction channel layer. Assume the width, length and thickness of this channel layer are given by W, L, and t, respectively. Then, the channel conductance, g_{c1}, of the transistor is given by

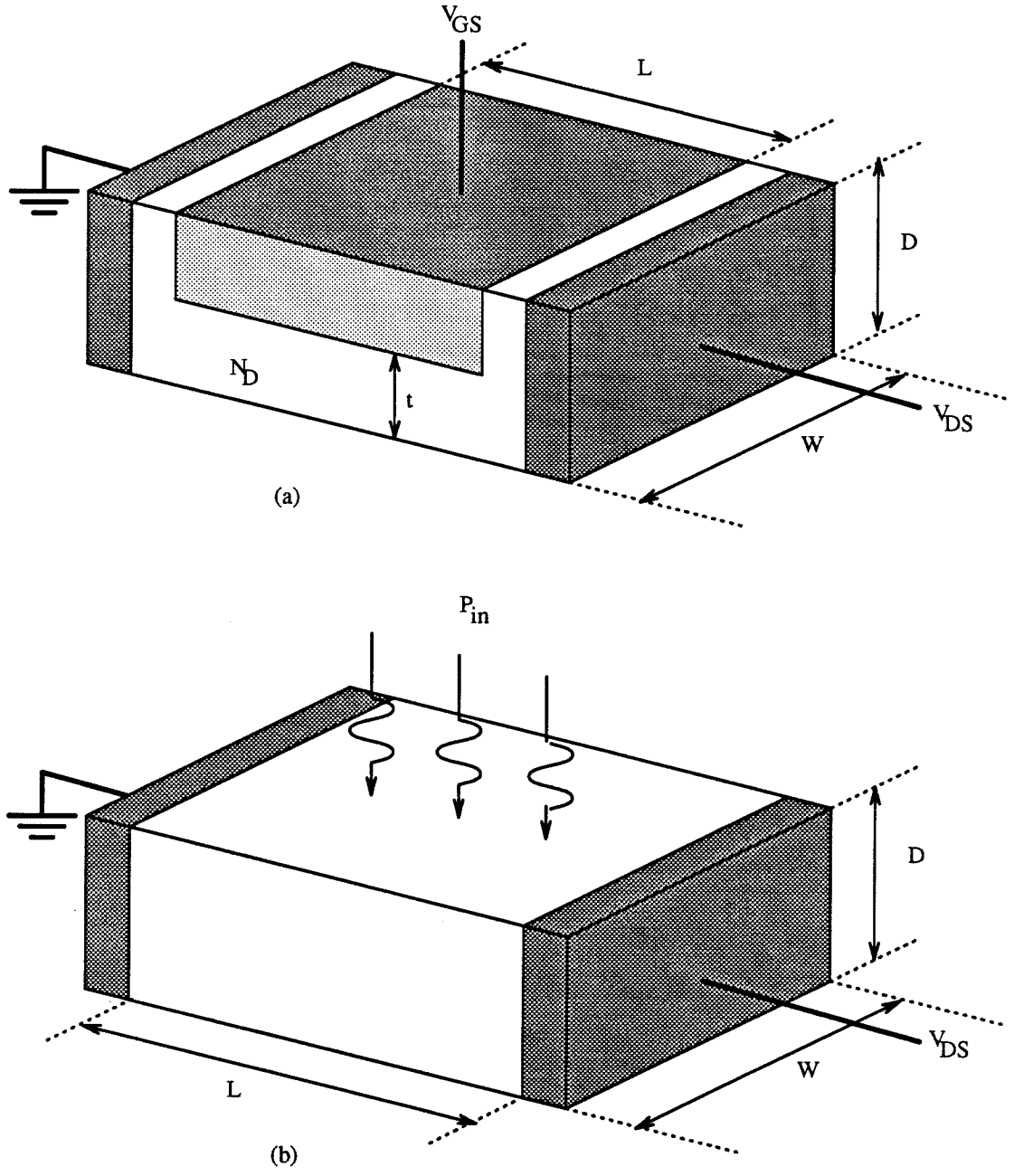


Fig. 5.33 **(a)** Model of an enhancement-mode MESFET at a positive V_{GS}. The conductance channel thickness, t, is a result of the depletion region removal caused by the positive V_{GS} action. **(b)** Model of an optical FET. Its operational principle is the same as that of a photoconductor.

$$\begin{aligned} g_{c1} &= \sigma_n \frac{Wt}{L} \\ &= q\mu_n N_D \frac{W}{L} t, \end{aligned} \tag{5.4}$$

where σ_n, q, μ_n, and N_D are the conductivity of the channel, electron charge, mobility of the electrons in the channel, and the n-type doping concentration in the channel, respectively. The thickness of the channel, t, is further dependent upon the voltage of the gate, V_G, and can be expressed by

$$t = \sqrt{\frac{2\epsilon V_G}{qN_D}}, \tag{5.5}$$

where ϵ is the permitivity of the semiconductor material. Substituting Eq. (5.5) into Eq. (5.4) immediately yields

$$\begin{aligned} g_{c1} &= \frac{W}{L} \cdot \mu_n \cdot \sqrt{2\epsilon q N_D} \cdot \sqrt{V_G} \\ &= p_1 \cdot \sqrt{V_G} \end{aligned} \tag{5.6}$$

with

$$p_1 = \frac{W}{L} \cdot \mu_n \cdot \sqrt{2\epsilon q N_D}. \tag{5.7}$$

Thus, we find that the channel conductance increases in proportion to $\sqrt{V_G}$. This conductance will continue to increase with increasing V_G until V_G eventually turns on the Schottky diode across the gate and the source. In which case, the majority

of the channel current comes from the gate-source Schottky diode forward current, rather than the ohmic current that flows from the drain to the source.

In a similar fashion, the channel conductance of the optical FET before becoming saturated, g_{c2}, can be derived. The operational principle of an optical FET is the same as that of a photoconductor. Assuming the incident light, P_{in}, is fully absorbed in the semiconductor, which has dimensions of D, W, and L, respectively for its thickness, width and length as shown in Fig. 5.33(b). Then, the carrier generation rate per unit volume, U_G, is

$$U_G = \frac{\eta_D \cdot \frac{P_{in}}{h\nu}}{WDL} = \frac{dn}{dt} = \frac{dp}{dt}, \tag{5.8}$$

where n and p are the electron and hole concentrations, and η_D, and $h\nu$ are the detector quantum efficiency and the photon energy respectively. These photo-generated electrons and holes will survive on average for a time equal to their lifetimes, τ_n and τ_p, before they recombine with each other. Thus, the carrier recombination rate can be given by

$$U_R = \frac{n}{\tau_n} = \frac{n}{\tau_p}. \tag{5.9}$$

At steady state, the carrier generation rate must be equal to the carrier recombination rate. Otherwise, there will be carrier build-up or depletion. By equating Eq. (5.8) and Eq. (5.9), we obtain an expression for the steady-state electron and hole concentration per unit volume.

$$\begin{aligned} n &= \frac{\eta_D \cdot \tau_n}{WDL} \cdot \frac{P_{in}}{h\nu} \\ p &= \frac{\eta_D \cdot \tau_p}{WDL} \cdot \frac{P_{in}}{h\nu} \end{aligned}. \tag{5.10}$$

Upon application of an external electric field, E_{ext}, a current will flow according to the ohmic law as if these photo-generated carriers were intrinsic carriers. The magnitude of the current is

$$\begin{aligned} I_{DS} &= J_{DS} \cdot A \\ &= (q\mu_n n E_{ext} + q\mu_n n E_{ext}) \cdot (WD) \\ &\approx (q\mu_n n E_{ext}) \cdot (WD), \end{aligned} \tag{5.11}$$

where J_{DS}, A, and μ_n are the source-drain current density, area of the cross section, and the electron mobility. The fact that $\mu_n \gg \mu_p$ is used to obtain the approximation. Upon substituting the expression for n in Eq. (5.10) into Eq. (5.11) and let $V_{DS} = E_{ext} \cdot L$, we find

$$I_{DS} = q \cdot \frac{\eta_D}{h\nu} \cdot \frac{\mu_n \tau_n}{L^2} \cdot P_{in} \cdot V_{DS}. \tag{5.12}$$

Therefore,

$$\begin{aligned} g_{c2} &= q \cdot \frac{\eta_D}{h\nu} \cdot \frac{\mu_n \tau_n}{L^2} \cdot P_{in} \\ &= p_2 \cdot P_{in} \end{aligned} \tag{5.13}$$

and

$$p_2 = q \cdot \frac{\eta_D}{h\nu} \cdot \frac{\mu_n \tau_n}{L^2}. \tag{5.14}$$

As we can see from Eq. (5.13), the value of g_{c2} depends linearly on P_{in}, unlike MESFET, for which a square root dependence on the external control parameter, such as V_G, is found. It should be noted further that Eq. (5.12) can be re-written as

$$\begin{aligned} I_{DS} &= \left(q \cdot \eta_D \cdot \frac{P_{in}}{h\nu}\right) \cdot \left(\frac{\tau_n}{L^2/\mu_n V_{DS}}\right) \\ &= I_{ph} \cdot \frac{\tau_n}{\tau_{tr}}, \end{aligned} \tag{5.15}$$

where I_{ph} represents the primary photocurrent and the factor $\tau_n/\ \tau_{tr}$ represents the photoconductive gain with τ_{tr} denoting the carrier transit time. Thus, in general, it is desirable to make the gap between the electrodes, L, as small as the amount of light that needs to be detected permits and at the same time keep the electron lifetime large in order to maximize the photoconductive gain.

To obtain the output conductances, g_{o1} and g_{o2}, for both transistors after they have become saturated, we need to consider the equivalent of the Early effect from bipolar transistors as it applies to field effect transistors. Figure 5.34(a) and 5.34(b), which show the I-V characteristics of the MESFET and the optical FET, illustrate the idea behind it. In Fig. 5.34(a), there is an Early voltage, V_{E1}, such that if the I-V curves in the saturation region of the transistor are extended to the left, they would all intersect at $-V_{E1}$. Because of the finite V_{E1}, the output conductance of the transistor is generally not zero. This causes an undesired rise in current as the voltage is raised in the saturation region of the transistor. The origin of this rise in current can be explained by several physical phenomena in FET's, such as leakage current through the substrate [86] and channel length modulation [87]. However, if we are only interested in the behavior of the output conductance in relation to the external control parameter, such as V_G, a simple knowledge of the empirical value of

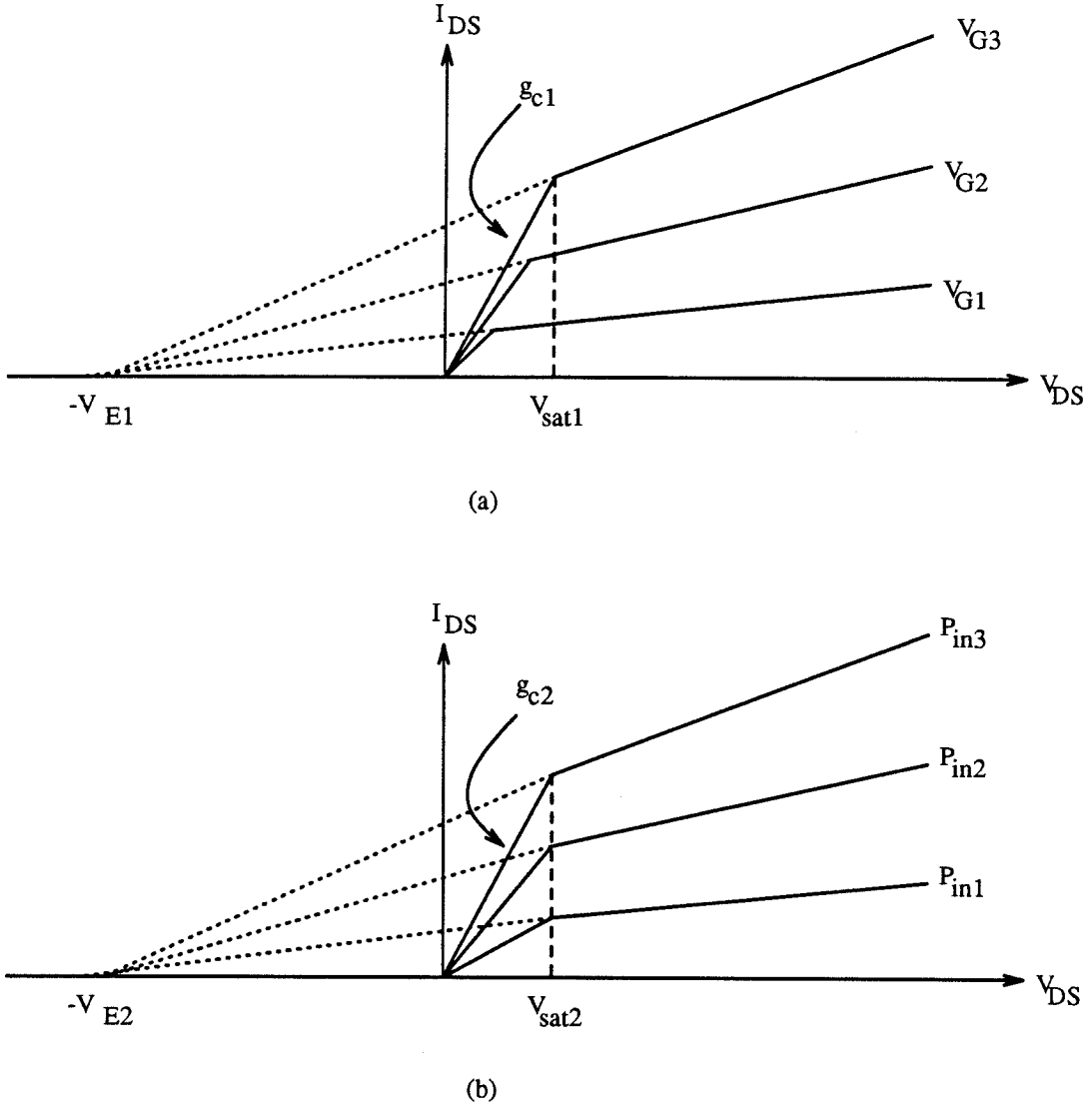


Fig. 5.34 **(a)** The DC I-V characteristics of the MESFET at various V_G values. The Early effect is indicated by the dotted lines, which intersect at $-V_{E1}$. **(b)** The DC I-V characteristics of the optical FET at various P_{in} values. The same Early effect is also indicated by the dotted lines, which intersect at $-V_{E2}$.

the Early voltage will suffice. Going back to Fig. 5.34(a), the output conductance of the MESFET can be easily calculated from the triangle formation due to the extension of the I-V curve to $-V_{E1}$. For the curve with $V_G = V_{G3}$, the output conductance is given by

$$g_{o1} = \frac{g_{c1} \cdot V_{sat1}}{V_{E1} + V_{sat1}}. \tag{5.16}$$

If we assume this enhancement-mode MESFET has a threshold voltage of 0 volts, which means that the depletion front is such that the source-drain current, I_{DS}, is just totally cutoff, any slight positive V_G would retract the depletion front and induce a non-zero I_{DS}. With this assumption, V_{sat1} is simply equal to V_G because any undepleted channel caused by a positive V_G can be depleted again by the same amount of voltage applied at the drain. This is why V_{sat1} is equal to V_G under the condition that the threshold voltage is zero. If we substitute V_{sat1} by V_G in Eq. (5.16) and replace g_{c1} by Eq. (5.6), we obtain

$$\begin{aligned} g_{o1} &= \frac{p_1 \cdot (V_G)^{3/2}}{V_{E1} + V_G} \\ &\approx q_1 \cdot (V_G)^{3/2}, \end{aligned} \tag{5.17}$$

where

$$q_1 = \frac{p_1}{V_{E1}}. \tag{5.18}$$

The approximation assumes V_G being negligible compared to V_{E1}, which is valid for $V_G < 1V$. Eq. (5.17) tells us that the output conductance of a MESFET depends on V_G to the 3/2 power.

A similar operation can be performed to obtain the dependence of g_{o2} on the input power, P_{in} for the optical FET. The only difference is that the saturation voltage for the optical FET, V_{sat2}, is a constant of the device because there is no gate to affect the extent of the depletion. Thus, V_{sat2} is completely determined by the source-drain voltage, V_{DS}. This distinction is clearly illustrated in Fig. 5.34(a) and 5.34(b). Following Eq. (5.16), the output conductance, g_{o2}, of the optical FET can be expressed by

$$g_{o2} = \frac{g_{c2} \cdot V_{sat2}}{V_{E2} + V_{sat2}}. \tag{5.19}$$

Substituting g_{c2} in Eq. (5.13) into Eq. (5.19) reveals

$$\begin{aligned} g_{o2} &\approx \frac{p_2 \cdot V_{sat2}}{V_{E2}} \cdot P_{in} \\ &= q_2 \cdot P_{in}, \end{aligned} \tag{5.20}$$

where

$$q_2 = \frac{p_2 \cdot V_{sat2}}{V_{E2}}. \tag{5.21}$$

Again, the assumption that $V_{sat2} << V_{E2}$ is used to obtain the linear dependence of the output conductance, g_{o2}, on the optical input power, P_{in}, of the optical FET.

Finally, a set of equations that relate the channel and output conductances to the external control parameters are obtained. For the analysis that follows, only these equations will be used and referred to. They are displayed again for summary.

$$\begin{aligned} g_{c1} &= p_1 \cdot (V_G)^{1/2} \\ g_{c2} &= p_2 \cdot P_{in} \\ g_{o1} &= q_1 \cdot (V_G)^{3/2} \\ g_{o2} &= q_2 \cdot P_{in} \end{aligned} \tag{5.22}$$

with p_1, p_2, q_1, and q_2 given by Eq. (5.7), (5.14), (5.18), and (5.21) respectively.

When the optical FET, powered by V_{CC}, is connected in series with the MESFET, the voltage, V_o, which is the voltage between the two transistors as shown in Fig. 5.32(a), is determined by the intersection point of the MESFET I-V curve with the I-V curve from the optical FET, which is plotted backward with V_{CC} being its new origin. This is illustrated in Fig. 5.35(a) through 5.35(c). Clearly, there are three distinct regions into which the intersection point falls. They are $0 \leq V_o \leq V_{sat1}, V_{sat1} \leq V_o \leq V_{CC} - V_{sat2}$, and $V_{CC} - V_{sat2} \leq V_o \leq V_{CC}$, which we should label them as region 1, 2, and 3 respectively.

For region 1, the intersection point can be found by equating the current from each transistor. Explicitly,

$$g_{c1} \cdot V_o = V_{sat2} \cdot g_{c2} + g_{o2} \cdot (V_{CC} - V_{sat2} - V_o). \tag{5.23}$$

Solving for V_o, we obtain

$$V_o = \frac{V_{sat2} \cdot g_{c2} + g_{o2} \cdot (V_{CC} - V_{sat2})}{g_{c1} + g_{o2}}. \tag{5.24}$$

Substituting the expressions for g_{c1}, g_{c2}, and g_{o2} as outlined in Eq. (5.22) into Eq. (5.24) shows

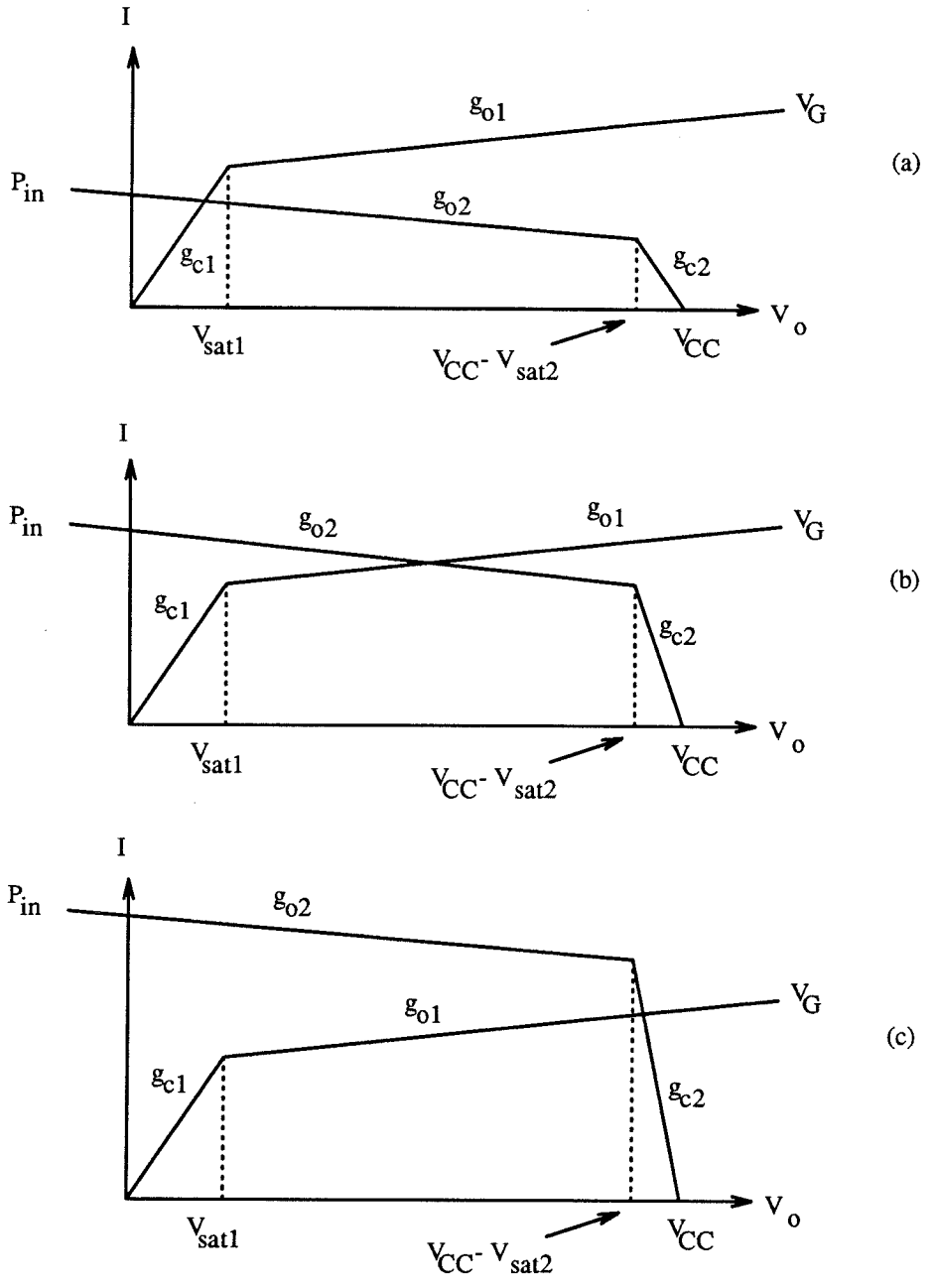


Fig. 5.35 Illustration of the DC operating point for the input switching circuit for the optoelectronic neuron. The I-V curve for the MESFET is plotted against that of the optical FET. The intersecting point determines the voltage, V_o, which is the gate voltage of the output MESFET and varies between 0 V and V_{CC}. The figures in **(a)**, **(b)** and **(c)** corresponds to three distinct regions in which the voltage, V_o, is determined.

$$V_o = \frac{p_2 \cdot V_{sat2} + q_2 \cdot (V_{CC} - V_{sat2})}{p_1 \cdot \sqrt{V_G} + q_2 \cdot P_{in}} \cdot P_{in}. \tag{5.25}$$

Eq. (5.25) is valid as long as $0 \leq P_{in} \leq (P_{in})_{1,2}$, where. $(P_{in})_{1,2}$ is the value which makes $V_o = V_{sat1}$. The values of $(P_{in})_{1,2}$ can be easily found by setting $V_o = V_{sat1}$ in Eq. (5.25) and solving for P_{in}, which is given below :

$$(P_{in})_{1,2} = \frac{p_1 \cdot (V_G)^{3/2}}{p_2 \cdot V_{sat2} + q_2 \cdot (V_{CC} - V_{sat2}) - q_2 \cdot V_G}. \tag{5.26}$$

Similar equations can be derived for region 2 and 3. For region 2,

$$g_{c1} \cdot V_{sat1} + g_{o1} \cdot (V_o - V_{sat1}) = g_{c2} \cdot V_{sat2} + g_{o2} \cdot (V_{CC} - V_{sat2} - V_o). \tag{5.27}$$

With the aid of Eq. (5.22), V_o can be written as

$$V_o = \frac{P_{in} \cdot V_{sat2} \cdot (p_2 - q_2) + q_2 \cdot P_{in} \cdot V_{CC} + q_1 \cdot (V_G)^{5/2} - p_1 \cdot (V_G)^{3/2}}{q_1 \cdot (V_G)^{3/2} + q_2 \cdot P_{in}}, \tag{5.28}$$

with the region of validity being $(P_{in})_{1,2} \leq P_{in} \leq (P_{in})_{2,3}$, where $(P_{in})_{1,2}$ is given in Eq. (5.26) and $(P_{in})_{2,3}$ is given below :

$$(P_{in})_{2,3} = \frac{p_1 \cdot (V_G)^{3/2} - q_1 \cdot (V_G)^{5/2} - V_{sat2} \cdot q_1 \cdot (V_G)^{3/2} + V_{CC} \cdot q_1 \cdot (V_G)^{3/2}}{V_{sat2} \cdot p_2}. \tag{5.29}$$

For region 3,

$$g_{c1} \cdot V_{sat1} + g_{o1} \cdot (V_o - V_{sat1}) = g_{c2} \cdot (V_{CC} - V_o). \tag{5.30}$$

Again, with the aid of Eq. (5.22), V_o is given by

$$V_o = \frac{P_{in} \cdot p_2 \cdot V_{CC} + q_1 \cdot (V_G)^{5/2} - p_1 \cdot (V_G)^{3/2}}{q_1 \cdot (V_G)^{3/2} + p_2 \cdot P_{in}}, \tag{5.31}$$

with the region of validity being $(P_{in})_{2,3} \leq P_{in} \leq \infty$. A simple check of Eq. (5.31) can be performed by taking the limit of P_{in} to infinity. This operation reveals

$$\lim_{p_{in} \to \infty} \frac{P_{in} \cdot p_2 \cdot V_{CC} + q_1 \cdot (V_G)^{5/2} - p_1 \cdot (V_G)^{3/2}}{q_1 \cdot (V_G)^{3/2} + p_2 \cdot P_{in}} = V_{CC}, \tag{5.32}$$

which agrees with the expectation. Eq. (5.25), (5.28), and (5.31) together describe the behavior of V_o due the variation of P_{in}. It is therefore important to have a comprehensive understanding of the effect each parameter has on the overall characteristics of the input-output relationship for the neuron. Shown in Fig. 5.36 - 5.41 are examples of the neuron input-output characteristics with one parameter varied at a time see the effect of each parameter on the overall neuron characteristics. The corresponding parameters varied in Fig. 5.36 - 5.41 are p_1, p_2, q_1, q_2, q_1 and q_2 at the same time, and V_G respectively. The actual numbers used in each figure are shown in the caption of the figure.

In Fig. 5.36, because both q_1 and q_2 are set to 0, the transition from the off to the on state is very abrupt. The optical input power, P_{in}, at which this transition occurs increases with increasing p_1 because as the channel conductance in the MESFET gets bigger, the optical FET needs to produce more photocurrent before turning on the neuron. It is also noted that the value of V_o right before the

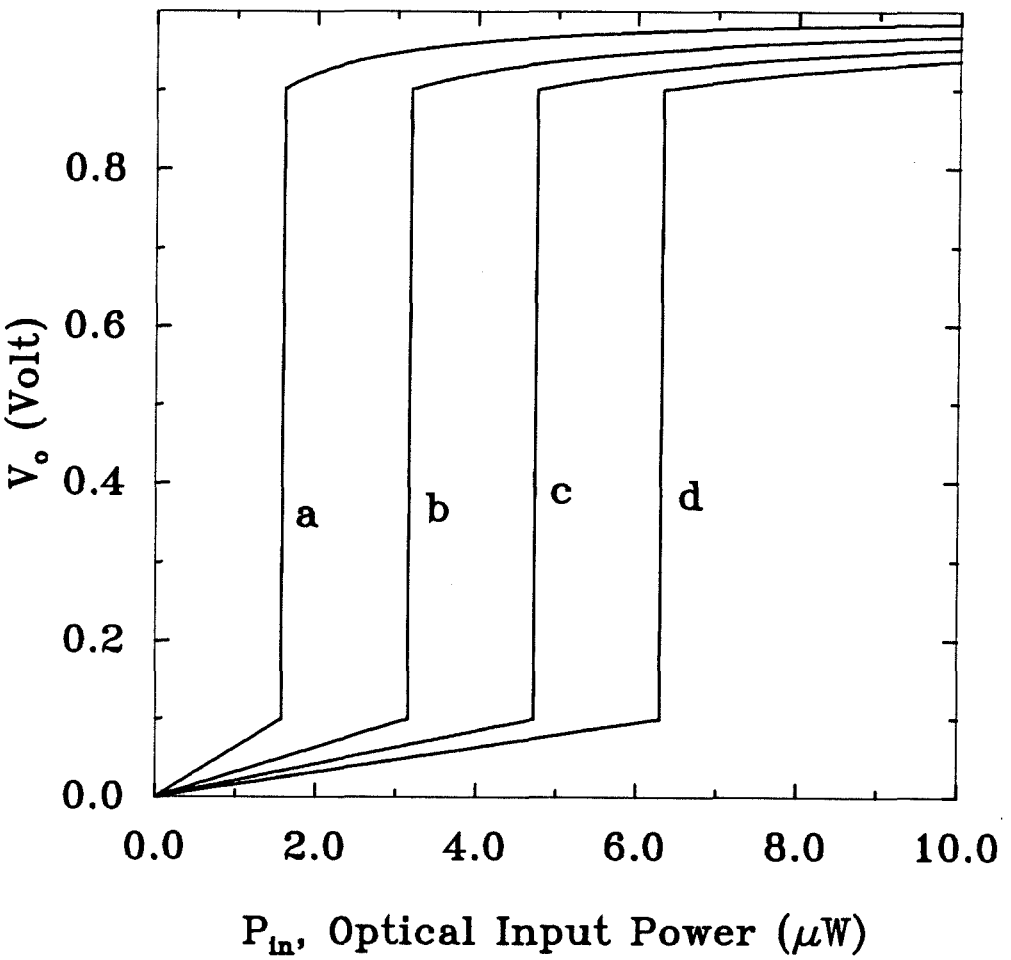


Fig. 5.36 Input-output characteristics of the neuron for various p_1 values and the following parameters : p_1 = (**a**) 500 (**b**) 1000 (**c**) 1500 (**d**) 2000, p_2 = 100, $q_1 = 0$, $q_2 = 0$, $V_{sat2} = 0.1\ V$, $V_G = 0.1\ V$, and $V_{CC} = 1\ V$

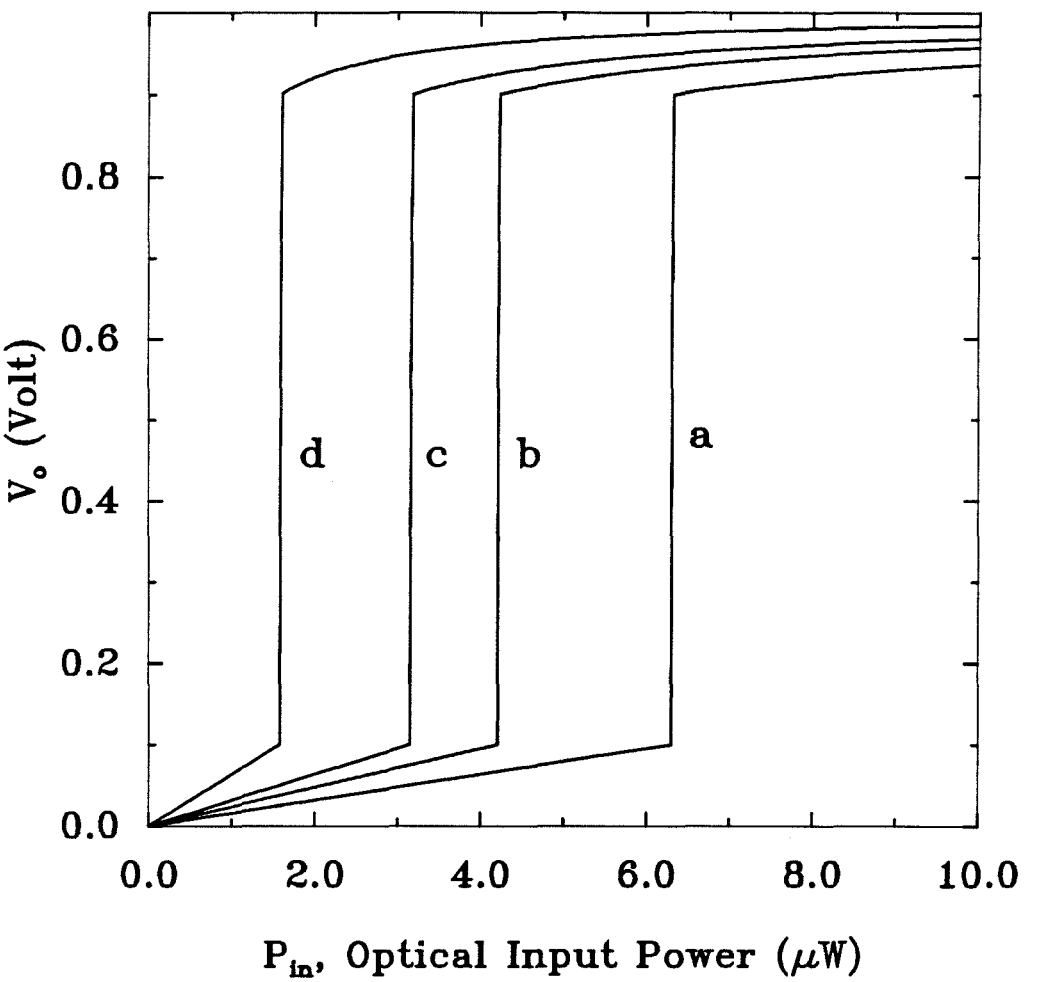


Fig. 5.37 Input-output characteristics of the neuron for various p_2 values and the following parameters : $p_1 = 1000$, $p_2 =$ (**a**) 50 (**b**) 75 (**c**) 100 (**d**) 200, $q_1 = 0$, $q_2 = 0$, $V_{sat2} = 0.1\ V$, $V_G = 0.1\ V$, and $V_{CC} = 1\ V$

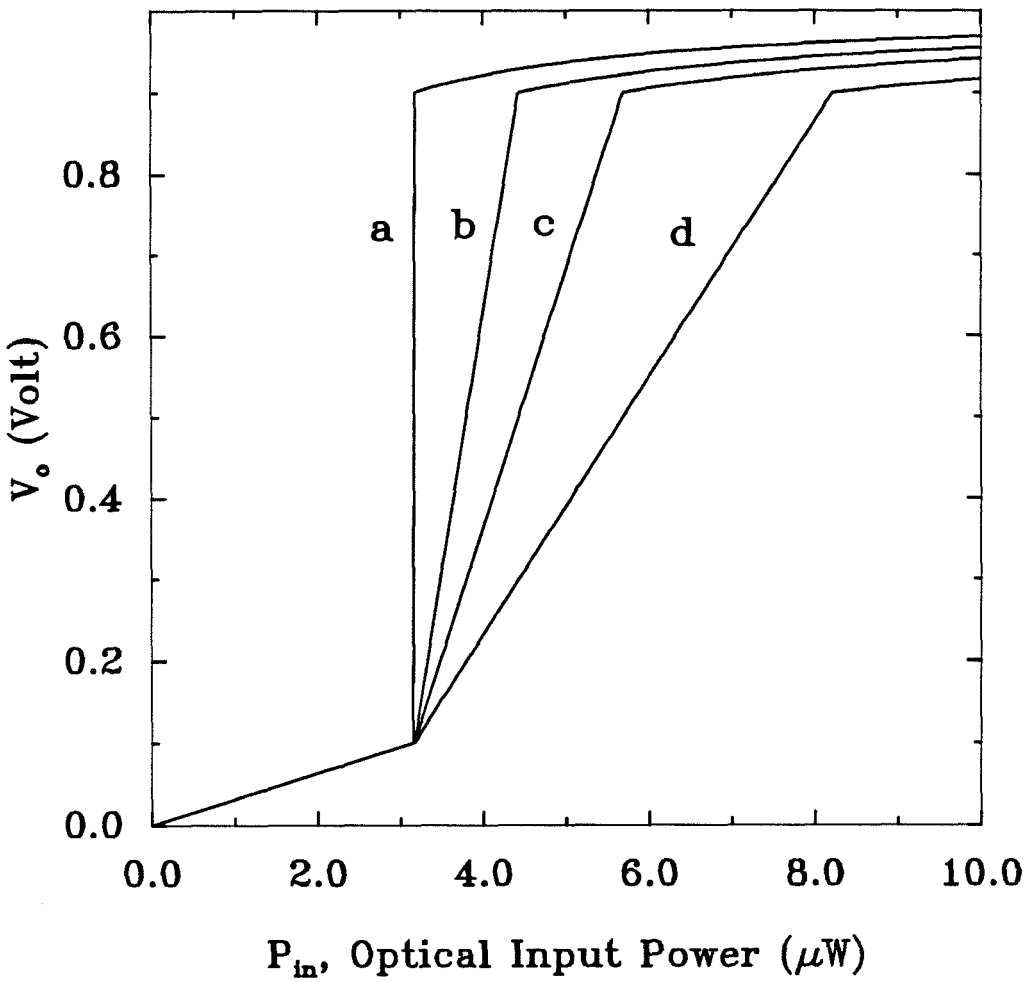


Fig. 5.38 Input-output characteristics of the neuron for various q_1 values and the following parameters : $p_1 = 1000$, $p_2 = 100$, $q_1 =$ (**a**) 0 (**b**) 500 (**c**) 1000 (**d**) 2000, $q_2 = 0$, $V_{sat2} = 0.1\ V$, $V_G = 0.1\ V$, and $V_{CC} = 1\ V$

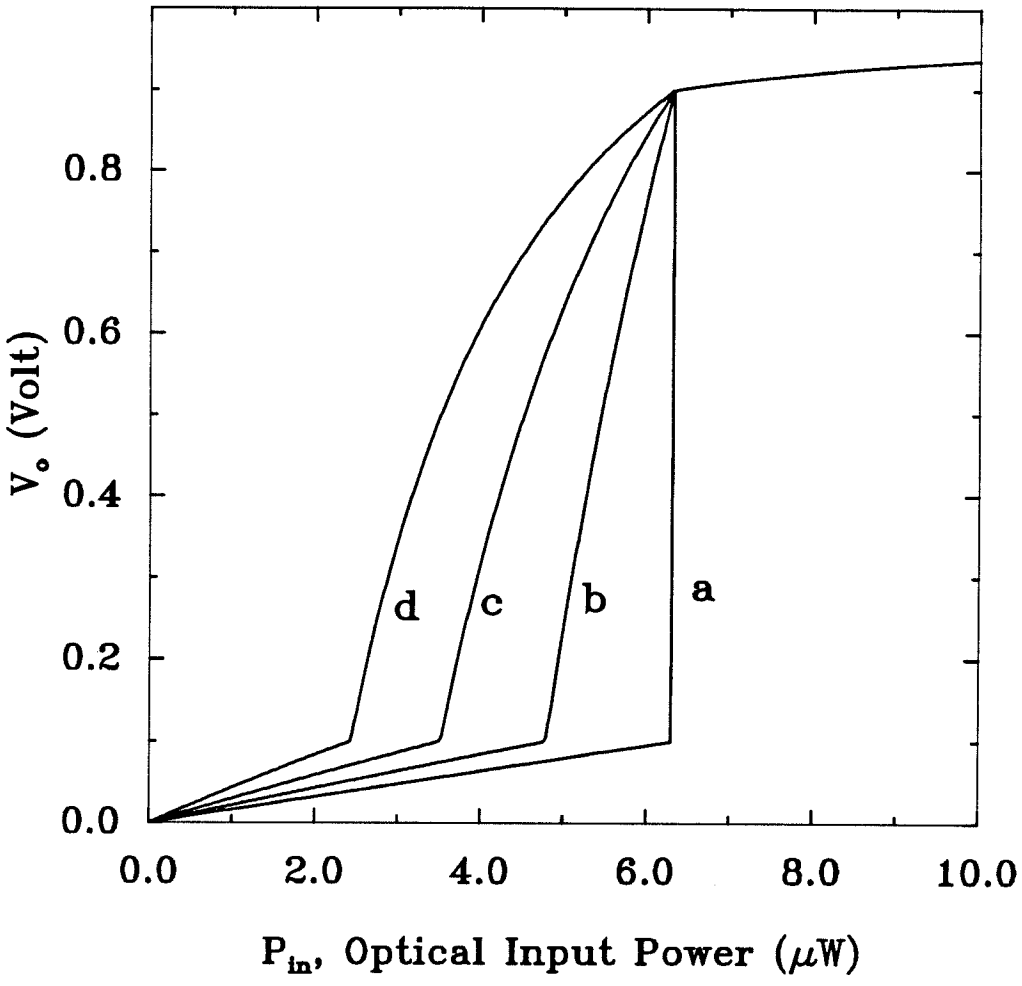


Fig. 5.39 Input-output characteristics of the neuron for various q_2 values and the following parameters : $p_1 = 1000$, $p_2 = 50$, $q_1 = 0$, $q_2 =$ (a) 0 (b) 2 (c) 5 (d) 10, $V_{sat2} = 0.1\ V$, $V_G = 0.1\ V$, and $V_{CC} = 1\ V$

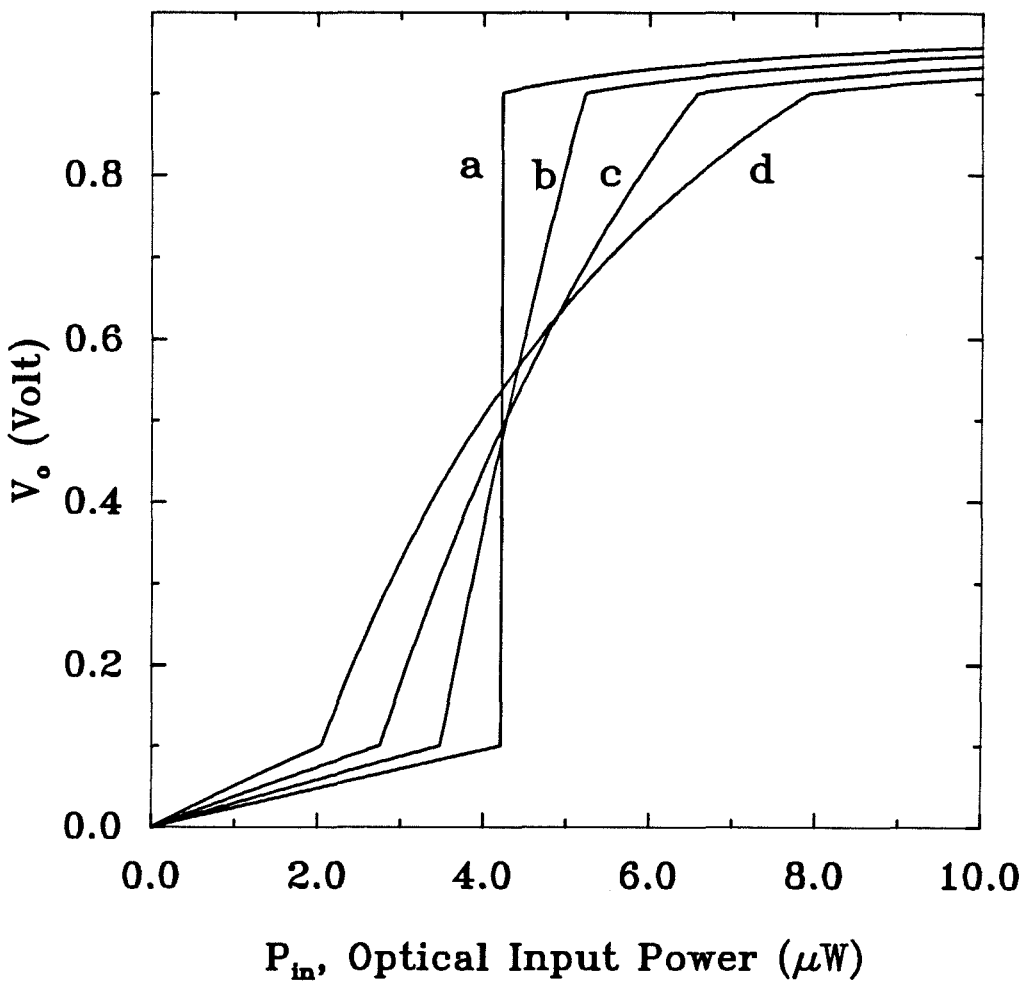


Fig. 5.40 Input-output characteristics of the neuron for values of q_1 and q_2 simultaneously varied and the following parameters : $p_1 = 1000$, $p_2 = 75$, $(q_1, q_2) =$ **(a)** $(0,0)$ **(b)** $(300,2)$ **(c)** $(700,5)$ **(d)** $(1100,10)$, $V_{sat2} = 0.1\ V$, $V_G = 0.1\ V$, and $V_{CC} = 1\ V$

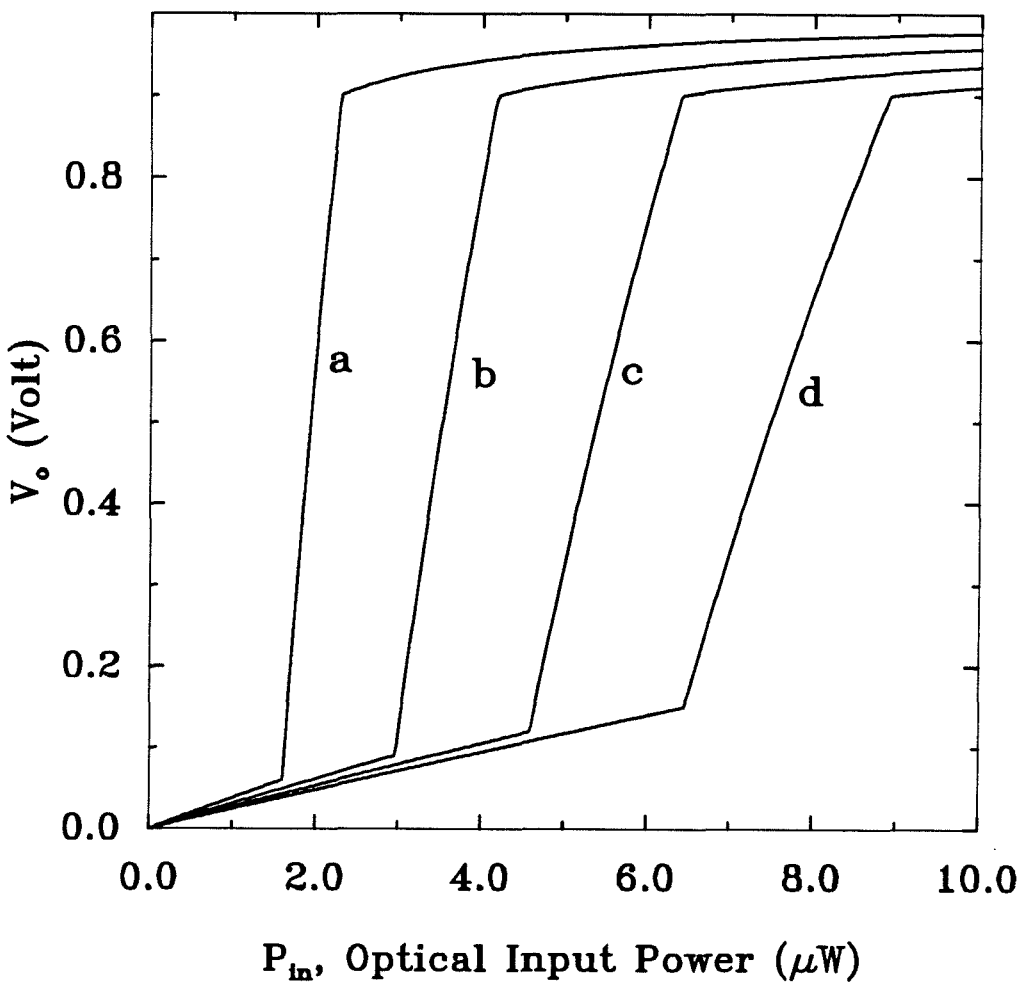


Fig. 5.41 Input-output characteristics of the neuron for various V_G values and the following parameters : $p_1 = 1000$, $p_2 = 75$, $q_1 = 200$, $q_2 = 2$, $V_{sat2} = 0.1\ V$, $V_G =$ (**a**) $0.06\ V$ (**b**) $0.09\ V$ (**c**) $0.12\ V$ (**d**) $0.15\ V$, and $V_{CC} = 1\ V$

transition takes place is always 0.1 V. This is due to the fact that V_{sat1}, which is equal to V_G, is set to 0.1 V. Consequently, the MESFET would not be in saturation until V_o is at least 0.1 V, a condition for neuron thresholding. Similarly, the value of V_o at the tip of transition is always 0.9 V because the value of V_{sat2} and V_{CC} are set to 0.1 V and 1 V respectively. This implies that the optical FET would not come out of the saturation until V_o reaches 0.9 V, a condition for neuron output saturation. A similar trend is observed when p_2 is varied. As p_2 is increased, the optical FET becomes more conductive in the linear region. This has the consequence of generating more photocurrent with the same input light power and surpassing the level of current drawn by the MESFET sooner. Thus, the value of P_{in} at which the neuron thresholding transition occurs decreases. This is clearly evident in Fig. 5.37.

It would be interesting to see how the input-output characteristics change as the output conductance of the MESFET, which depends on q_1, and of the optical FET, which depends on q_2, are changed. Figure 5.38 and 5.39 show the results. In Fig. 5.38, the parameter, q_1, is varied. This means physically that the output conductance of the MESFET increases with increasing V_G. Because of the increase in the output conductance, the slope of the I-V curve for the MESFET in the saturation region becomes steeper. This has the consequence of making the optical FET more difficult to come out of the saturation and go into the linear region. Thus, the input power level at which the neuron output saturates is increased. This effect is clearly manifested in Fig. 5.38, in which the slope of the neuron thresholding transition loses its steepness as q_1 is increased. On the other hand, if we keep q_1 constant and vary q_2, the factor that affects the output conductance of the optical FET, a very similar but opposite effect is observed. Because the output conductance of the optical FET is increased, which makes the slope of the I-V curve in the optical FET steeper, it is easier to drive the MESFET into saturation, which,

in turn, pulls up the voltage, V_o, and thus turns on the neuron. This is why with increasing q_2, the threshold value in P_{in} decreases in a manner depicted in Fig. 5.39. If we combine the effect of output conductance due to both transistors, a set of curves shown in Fig. 5.40 would be obtained. Thus, from these results, it is clear that any output conductance in the transistor is undesirable. It especially hurts the differential optical gain in the neuron. However, it does gain in reducing the threshold value in optical input power if the output conductance of the optical FET is allowed to increase. As a result of the reduced differential optical gain, the thresholding characteristics become softer, which may be desirable for certain neurons, in which the issue of robustness is of some concern.

Finally, if we increase the value of V_G in the MESFET, an increase in the output saturation current across the source and drain is expected. This is because an increase in V_G will increase the channel conductance in the linear region, g_{c1}, as seen from Eq. (5.22) as well as the saturation voltage, V_{sat1}, which is equal to V_G. In fact, according to Eq. (5.22), g_{c1} varies with V_G to the 1/2 power. Thus the output saturation current in the MESFET, $(I_{DS})_{sat}$, grows with V_G to the 3/2 power because it is equal to the product of g_{c1} and V_{sat1}. Consequently, the optical FET needs to generate more photocurrent to meet the threshold criteria for the neuron and a shift in the threshold level to the right as V_G is increased is expected. Once the threshold criteria are met, the neuron thresholding characteristics, which are a measure of the differential optical gain, are determined by the output conductance of the MESFET and the optical FET. According to Eq. (5.22) again, increasing V_G results in an increase in the output conductance of the MESFET, g_{o1}, which, causes the softening of the thresholding curve and the decreasing in differential optical gain. This decrease in the differential optical gain can be observed in Fig. 5.41.

Overall, we learn that, from this model, the output conductance of both tran-

sistor should be kept low in order to maximize the optical gain. The value of V_G should also be as low as possible to allow rapid turn-on of the neuron and also to minimize the output conductance of the MESFET as we find that the factors that govern the thresholding characteristics are adversely affected by an increase in V_G to the 3/2 power. However, it needs to be large enough to overcome the dark current of the optical FET. Otherwise, the neuron will be on all the time. This model is applicable to any two transistors or transistor-like devices that are connected in series. If the two devices that need to be analyzed are different from the MESFET-optical FET combination, the only modification is in the form of Eq. (5.22), which is specific to MESFET and optical FET. So one needs to derive new expressions for g_{c1}, g_{c2}, g_{o1} and g_{o2} and the rest of the equations will remain unchanged. It should also be pointed out that this model will allow only a qualitative understanding of the neuron thresholding characteristics. A quantitative analysis and modeling would require a detailed knowledge on the physical phenomena that govern these devices at the microscopic level and very accurate derivation and estimate of the parameters used in the model.

Chapter 6

LED Vs. Laser

6.1 Introduction

On-chip light sources are important devices for the implementation of optoelectronic neurons. Two possible candidates exist. They are light-emitting diodes (LED's) and laser diodes (LD's). Both the LED and the laser diode array provide spatially incoherent illumination, which is actually desirable in many neural systems in order to avoid speckles. For most applications, both devices have sufficiently narrowband spectrum. However, LED's emit light in all directions with very low efficiencies, and lasers require threshold currents before significant light is emitted. Thus, there are relative advantages and disadvantages in LED's and lasers when applied to neural networks. In evaluating the merits of each device, electrical power dissipation and efficiency are the two most important parameters to consider. For a neural network, these two parameters will ultimately limit the density of neurons in a given area due to limited heat sinking capability of the semiconductor chip. There is also a geometrical limitation on the number of neurons that can be packed into a given area. Although, our current optoelectronic neurons fall under the latter limitation, with a better controlled photolithography, limitation by the electrical power dissipation is expected to dominate the density of the neuron array. Thus, the analysis presented in this chapter solely assumes the power dissipation limitation factor. Other limitations, such as photolithography, optical setup and application are not considered here. Based on this assumption, LED's are compared with laser diodes in terms of their efficiencies, which are then

analyzed as a function of the array size.

The analysis begins by comparing the two devices in general in a neural system framework. No specific circuits are assumed. However, as we attempt to analyze in more detail, the efficiency of the neurons incorporating the LED or the laser diode will depend on the specific circuit. Therefore, we will compare them by using the circuit presented in Ch. 4, which is the DHBT-based neuron, and in Ch. 5, which is the MESFET-based neuron. These two circuits are shown again in Fig. 6(a) and 6(b) respectively.

Generally speaking, the size of the optoelectronic neuron array should be as large as possible. One of the limitations is the heat dissipation in the chip as previously mentioned. As the density of the neuron array increases, the current drawn by the neurons increases proportionally. Thus, the electrical power dissipation, which turns into heat, increases as well. At some point, the heat removal rate is smaller than the heat generation rate. As a result, the optoelectronic integrated circuit starts to heat up and the performance begins to degrade. To understand what role this limitation has played in a neuron array, let us designate the maximum power dissipation allowed by the chip and the number of neurons in the array to be $P_{elect,max}$ and N, respectively. Then, the maximum current that can be drawn to drive either the LED or the laser diode will be

$$I = \frac{P_{elect,max}}{NV}, \tag{6.1}$$

where V is the power supply voltage for each neuron. Since Eq. (6.1) gives the maximum current available to drive the light source, the total optical power generated by the LED and the laser would be respectively

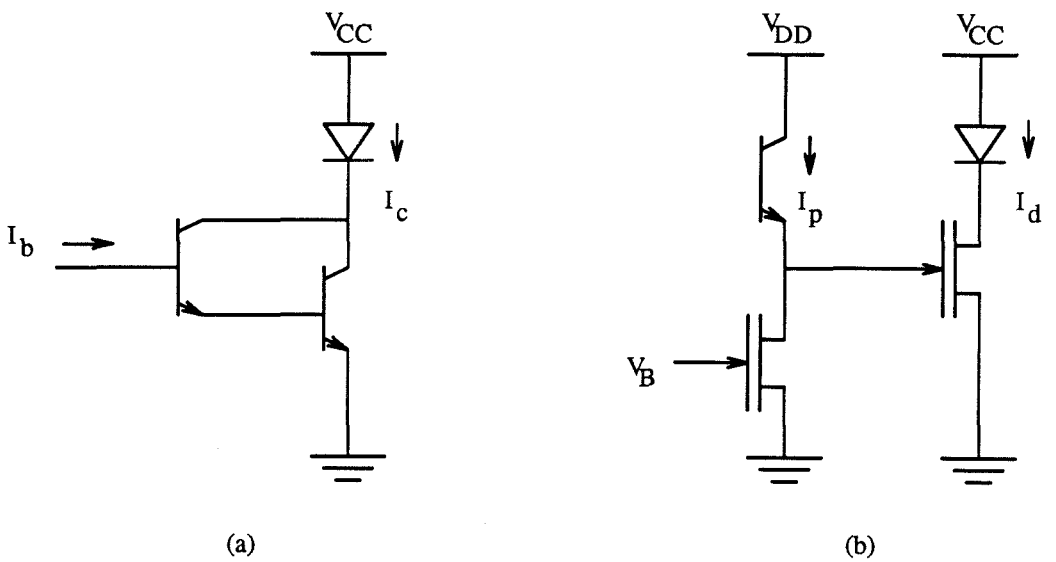


Fig. 6.1 **(a)** Schematic circuit of the optoelectronic neuron that is based on double heterojunction bipolar transistors (DHBT-based neurons). **(b)** Schematic circuit of the optoelectronic neuron that is based on metal-semiconductor field-effect transistors (MESFET-based neurons).

$$P_{opt,LED} = \eta_{LED} I N, \tag{6.2}$$

$$P_{opt,LD} = \eta_{LD}(I - I_{th})N, \tag{6.3}$$

where η_{LED}, η_{LD}, I_{th} are the external quantum efficiencies of the LED and the laser diode and the threshold current of the laser diode. Substituting Eq. (6.1) into (6.2) and (6.3), we obtain

$$P_{opt,LED} = \eta_{LED} \frac{P_{elect,max}}{V} \tag{6.4}$$

$$P_{opt,LD} = \eta_{LD}(\frac{P_{elect,max}}{V} - I_{th}N). \tag{6.5}$$

From Eq. (6.4) and (6.5), we see that the total maximum optical power emitted by the LED is independent of the number of neurons in the array. This is expected because LED is a linear device. Thus, as the total electrical power dissipation, which is proportional to the current flowing through the LED, is fixed, the total optical power emitted by the LED array is also fixed. However, this is not the case for the laser diode. Because the laser diode has a threshold current requirement, the driving current has to bias to this threshold current before the laser diode is turned on. As a result, the majority of the electrical power is used to satisfy this threshold requirement. Therefore, the total optical power emitted by a laser diode is expected to decrease as the number of neurons increases. These results are graphically illustrated in Fig. 6.2(a).

On the other hand, if the total optical power emitted from the neuron array is constrained, the total electrical power dissipated by either the LED or the laser diode is

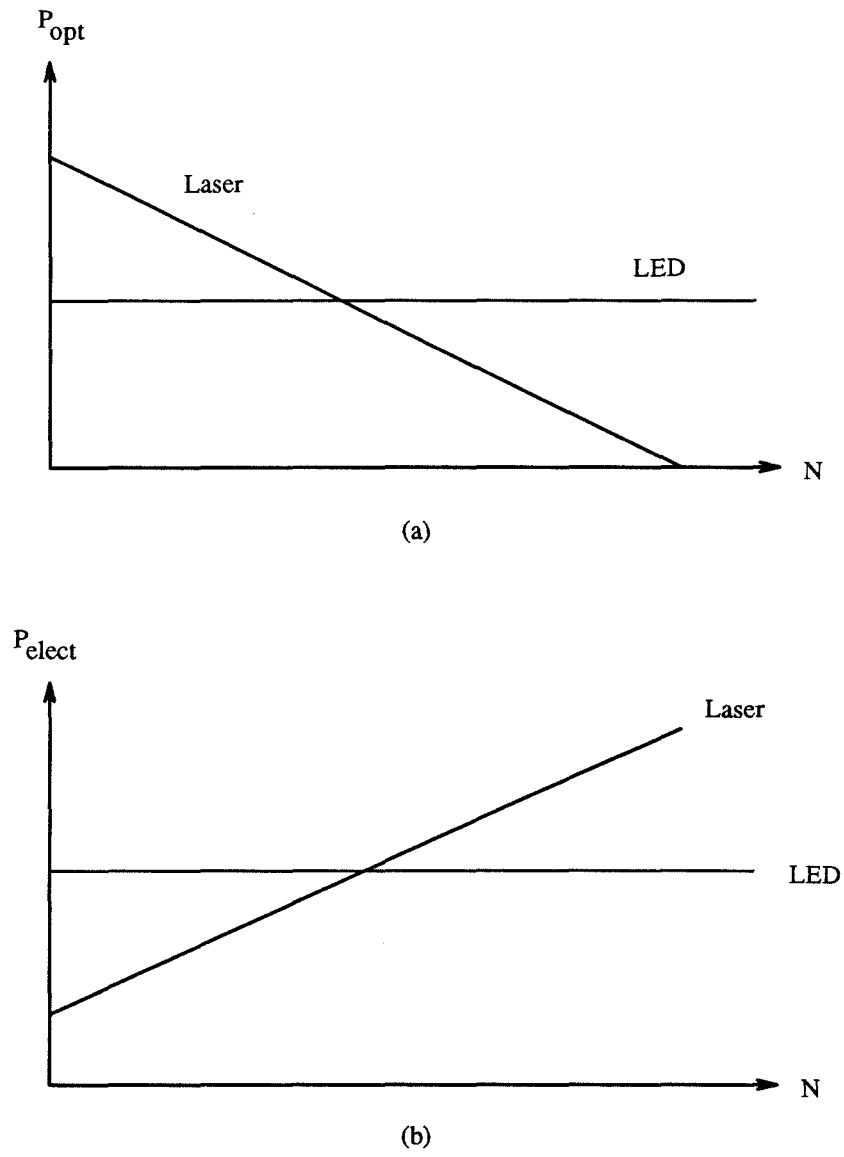


Fig. 6.2 **(a)** Comparison of the total optical power emitted by the LED and the laser diode as a function of the number of neurons in an array for a fixed electrical power dissipation in the chip. **(b)** Comparison of the total electrical power dissipation by the LED and the laser diode as a function of the number of neurons in an array for a fixed optical power emitted from the chip.

$$P_{elect,LED} = P_{elect,LD} = NVI. \tag{6.6}$$

Substituting I in Eq. (6.2) for LED and (6.3) for laser diode into Eq. (6.6), we obtain

$$P_{elect,LED} = V\frac{P_{opt}}{\eta_{LED}} \tag{6.7}$$

$$P_{elect,LD} = V(\frac{P_{opt}}{\eta_{LD}} + NI_{th}). \tag{6.8}$$

Again, we see that the electrical power dissipated by the LED is independent of the number of neurons in the array. However, the electrical power dissipated by the laser diode array increases with increasing the number of neurons in the array due to the fact that generating a fixed amount of optical power with fewer laser diodes requires less total threshold currents, thus less electrical power dissipation. These results are graphically shown in Fig. 6.2(b).

From these arguments, LED's are more power efficient if a large array of neurons are required. However, if we only need a few neurons, then laser diodes become the better choice. If we further impose other requirements, such as feedback and loop gain requirements, and specify the circuit elements in the neurons, more interesting tradeoffs may surface. These will be discussed in the subsequent sections.

6.2 DHBT-Based Neurons

In Ch. 4, we discussed optoelectronic neurons which incorporate two double heterojunction bipolar transistors (DHBT's) to drive the LED. In this circuit, the

bipolar transistors need to provide enough current gain to compensate for the inefficiency of the LED. This is usually achieved at the expense of a higher electrical power dissipation as the current gain of the transistor increases monotonically with increasing collector current until the high current injection condition, which causes the Kirt effect, becomes significant. In fact, this current dependent gain has been discussed in detail in Ch. 3 and the approximate relation between the current gain, β, and the collector current, I_c, is

$$\beta \sim I_c^{1-\frac{1}{n}}, \tag{6.9}$$

where n is the ideality factor of the base-emitter junction in the transistor and ranges between 1 and 2. For simplicity in analysis, we will assume n to be 2 in this section so that Eq. (6.9) can be rewritten as

$$\beta = \beta_0 \sqrt{I_c}. \tag{6.10}$$

Therefore, in order to get a higher current gain, a higher collector current is needed. However, a higher collector current results in higher electrical power dissipation, which will eventually limit the collector current available to drive the LED as the maximum electrical power dissipation on the chip is fixed. As a result, the maximum current gain is limited by the number of neurons in the array by

$$\beta = \beta_0 \sqrt{I_c} = \beta_0 \sqrt{\frac{P_{elect,max}}{NV}}, \tag{6.11}$$

where the relation $P_{elect,max} = NVI_c$ is used. Eq. (6.11) is valid for both the LED and the laser diode. The dependence of the maximum current gain available on the

number of neurons in the array is shown in Fig. 6.3, where we assume a $P_{elect,max}$ of 1 W/cm^2, a V of 2 volts, and a β_0 of $1000/\sqrt{mA}$. For a neuron array density of 1000/cm^2, the maximum β is only 700, which may not be enough to close the optical loop in the LED case or to overcome the threshold current in the laser diode case.

Another relevant parameter to monitor is the overall optical gain as defined by the ratio of the optical output power over the optical input power in the DHBT-based neuron. This overall optical gain is also expected to decrease as the array density increases. For LED, the optical gain can be found as follows,

$$P_{opt,out} = \eta_{LED} \cdot I_c = \eta_{LED} \cdot \beta \cdot I_b = \eta_{LED} \cdot \beta \cdot \eta_D \cdot P_{opt,in}, \tag{6.12}$$

where η_D is the phototransistor external quantum efficiency. Substituting β in Eq. (6.11) into Eq. (6.12) and moving $P_{opt,in}$ to the other side of the equation, we obtain

$$G_{opt} = \frac{P_{opt,out}}{P_{opt,in}} = \eta_{LED} \cdot \eta_D \cdot \beta_0 \sqrt{\frac{P_{elect,max}}{NV}}. \tag{6.13}$$

For the laser diode, the optical output power can be expressed as

$$\begin{aligned} P_{opt,out} &= (I_c - I_{th}) \cdot \eta_{LD} \\ &= (P_{opt,in} \cdot \eta_D \cdot \beta - I_{th}) \cdot \eta_{LD}. \end{aligned} \tag{6.14}$$

Moving $P_{opt,in}$ to the other side and substituting β in Eq. (6.11) into Eq. (6.14), we have

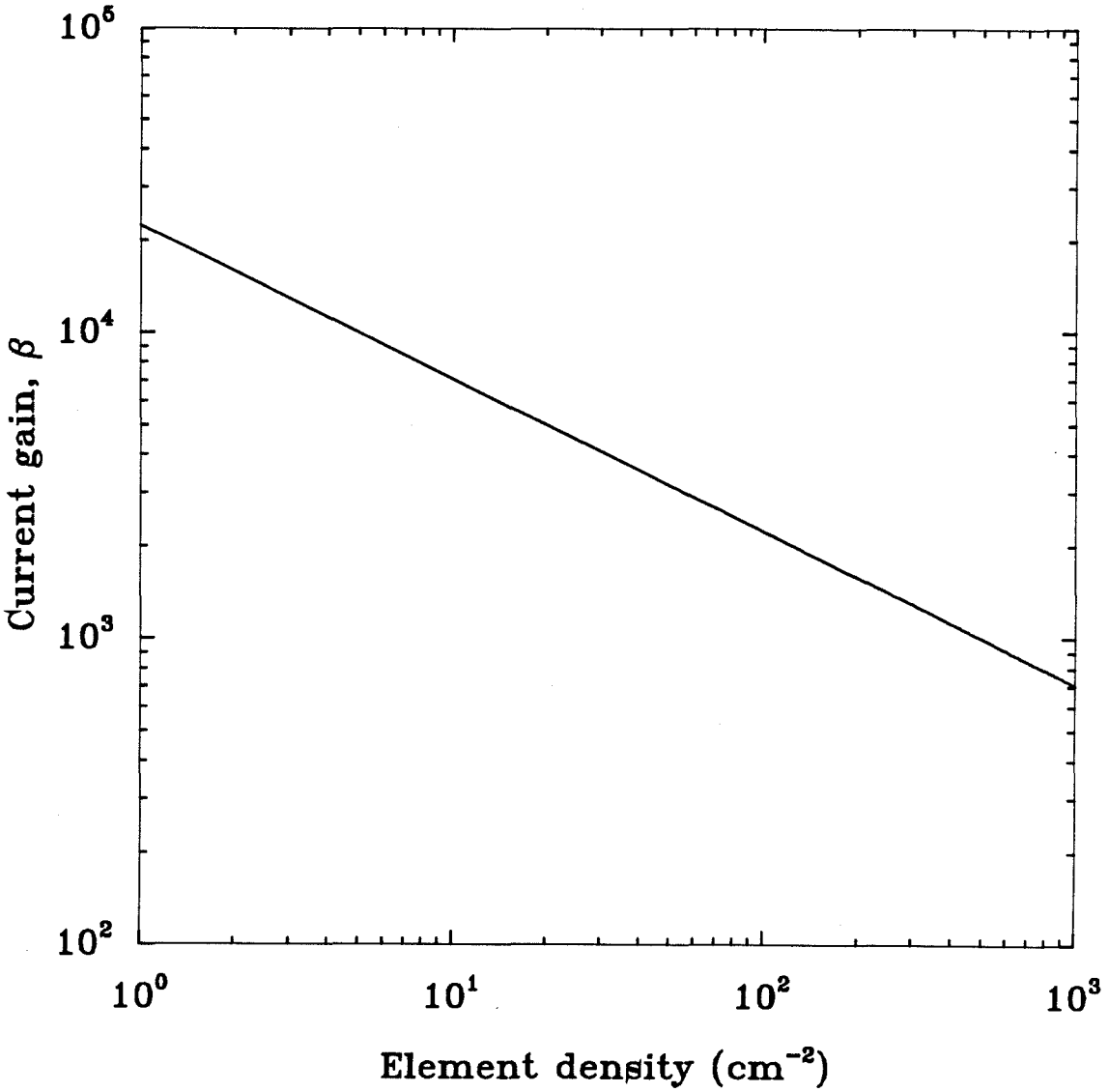


Fig. 6.3 Maximum current gain of the DHBT-based neuron as a function of the array density. This is a plot of Eq. (6.11) with the assumption of 1 W/cm^2 maximum electrical power dissipation in the chip and a driving voltage of 2 volts. Current gain coefficient, β_0, is set to $1000/\sqrt{mA}$.

$$G_{opt} = \frac{P_{opt,out}}{P_{opt,in}} = \eta_{LD} \cdot (\eta_D \cdot \beta_0 \sqrt{\frac{P_{elect,max}}{NV}} - \frac{I_{th}}{P_{opt,in}}). \tag{6.15}$$

From Eq. (6.13), we see that the overall optical gain in the LED case can possibly be less than one if the array density gets too large. Likewise in the case of laser diode, the overall optical gain, as shown in Eq. (6.15), can even become negative. This simply means that the collector current generated is less than the threshold current of the laser diode. A plot of these two cases is illustrated in Fig. 6.4. The parameters used in this plot are $P_{opt,in} = 20\mu W$, $\eta_{LED} = 0.01 W/A$, $\eta_{LD} = 0.1 W/A$, and $\eta_D = 0.3 A/W$ and also are used throughout the rest of this discussion unless otherwise specified. The external quantum efficiency of 0.1 W/A for the laser diode is typical for vertical-cavity surface-emitting laser diodes [100]. From this plot, it is clear that for a small array, the laser diode is the obvious choice because it is able to provide very high optical gain. However, as the array gets large, the LED becomes superior. This is consistent with the argument stated earlier in this chapter.

If we now try to close the optical loop by feeding the optical output back to its input through holographic connections with an efficiency η_H, the signal will be locked in the loop as long as the overall loop gain is at least one. This is accomplished by imposing the product of individual device efficiencies to be equal to one in the case of LED, or

$$\eta_H \cdot \eta_D \cdot \eta_{LED} \cdot \beta_0 \sqrt{I_c} = 1. \tag{6.16}$$

Solving for I_c, we have

$$I_c = (\frac{1}{\eta_H \cdot \eta_D \cdot \eta_{LED} \cdot \beta_0})^2. \tag{6.17}$$

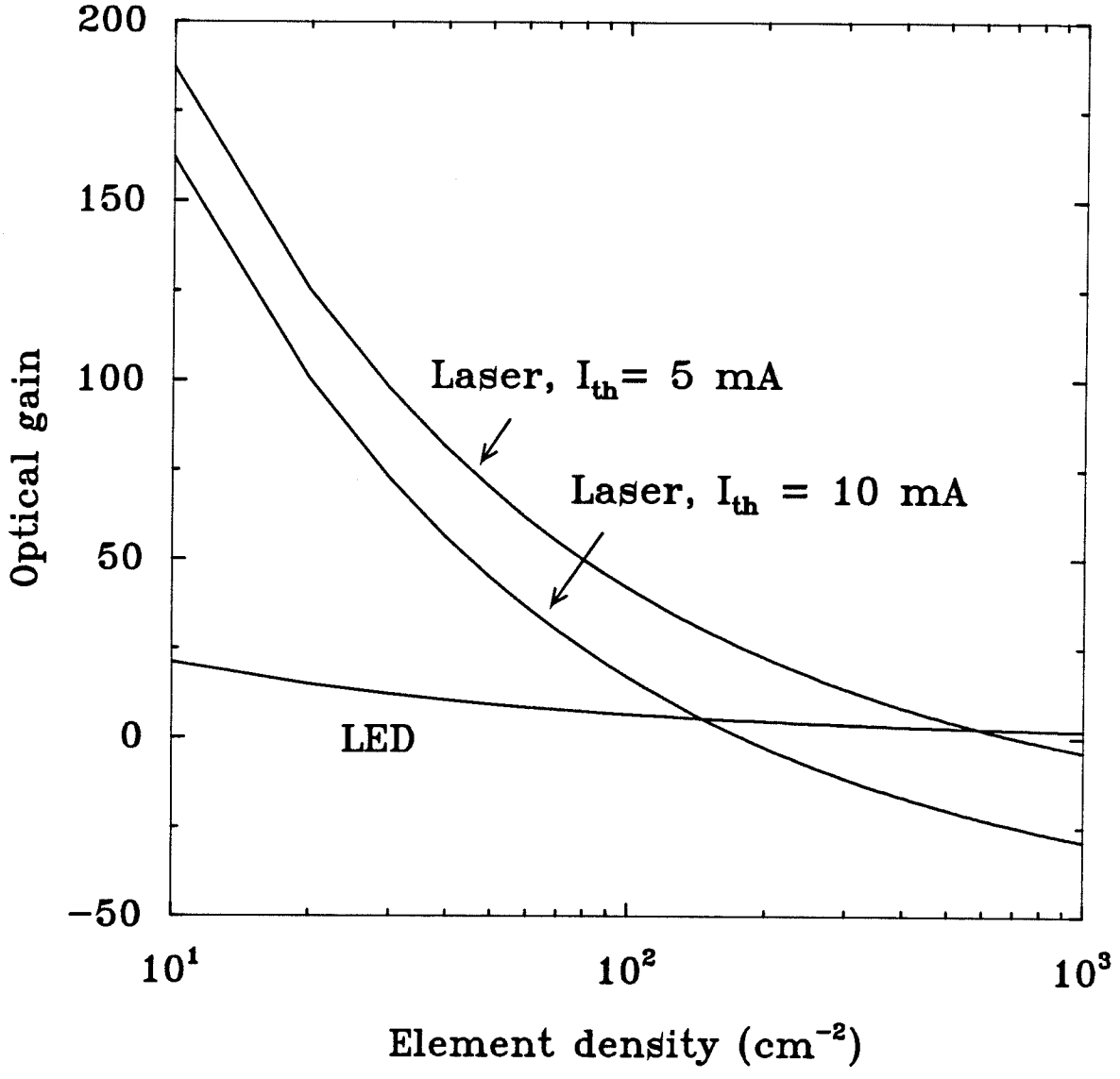


Fig. 6.4 Overall optical gain for the DHBT-based neuron as a function of the array density. The parameters used are stated in the text.

The I_c expressed in Eq. (6.17) is the current needed to close the optical loop with a loop gain of one. I_c also determines the maximum array density, because of the finite electrical power dissipation. However, if we can somehow make the transistor better by making it provide more current gain at the same current level, then the current level needed to close the loop may decrease substantially as the current varies inversely with the square of the current gain coefficient, β_0. Similar analysis can be carried out for the laser diode. In order to close the loop, the excess of the collector current over the threshold current, when scaled by the various device efficiencies, has to produce the same collector current. Thus,

$$(I_c - I_{th}) \cdot \eta_{LD} \cdot \eta_H \cdot \eta_D \cdot \beta_0 \sqrt{I_c} = I_c. \tag{6.18}$$

Solving for I_c in the quadratic equation in Eq. (6.18), we obtain

$$I_c = \frac{2I_{th} + \frac{1}{(\eta\beta_0)^2} + \sqrt{\frac{4I_{th}}{(\eta\beta_0)^2} + \frac{1}{(\eta\beta_0)^4}}}{2}, \tag{6.19}$$

where

$$\eta = \eta_{LD} \cdot \eta_H \cdot \eta_D. \tag{6.20}$$

To compare the relative magnitude of the collector current in the case of LED and laser diode with the optical loop closed, we plot the collector current as a function of the β coefficient, β_0, for both the LED and the laser diode in Fig. 6.5. Once the collector currents are determined, the density of the neuron array are also determined through the relation, $P_{elect,max} = NVI_c$, where I_c is the collector

current given in Eq. (6.17) and (6.19) respectively for the LED and the laser diode. Figure 6.6 shows the density of the array as a function of the β coefficient for the LED and the laser diode. It is clear from Fig. 6.6, that as the transistor is made better, the density for the LED-based neuron is larger than that for the laser-based neuron. This is because regardless how high the current gain of the transistor is, the laser-based neuron needs to draw at least the threshold current. Therefore, the density of the laser-based neuron array is limited. Whereas for the LED-based neuron, the increased β allows the neuron to draw less current. Thus, a higher density array is possible.

Usually, a loop gain of only one is not sufficient in a neural network. For a network in which large fanout is needed, a much higher loop gain is required. Thus, if we designate the loop gain to be g, then the corresponding collector current needed to achieve a loop gain of g in each case is

$$I_{c,LED} = \left(\frac{g}{\eta_H \cdot \eta_D \cdot \eta_{LED} \cdot \beta_0}\right)^2 \tag{6.21}$$

and

$$I_{c,LD} = \frac{2I_{th} + \frac{g}{(\eta\beta_0)^2} + \sqrt{\frac{4I_{th}g}{(\eta\beta_0)^2} + \frac{g}{(\eta\beta_0)^4}}}{2}. \tag{6.22}$$

From these two equations, the maximum array density while achieving a loop gain of g can be calculated. Figure 6.7 illustrates the collector current needed to close the loop as a function of the loop gain. Figure 6.8 shows the maximum array density as a function of the loop gain. In Fig. 6.7 and Fig. 6.8, a β_0 of $1000/\sqrt{mA}$ is assumed. Fig. 6.7 and 6.8 are similar to Fig. 6.5 and 6.6 except a unity gain is assumed in Fig. 6.5 and 6.6. As we can see from Fig. 6.8, as the loop gain increases,

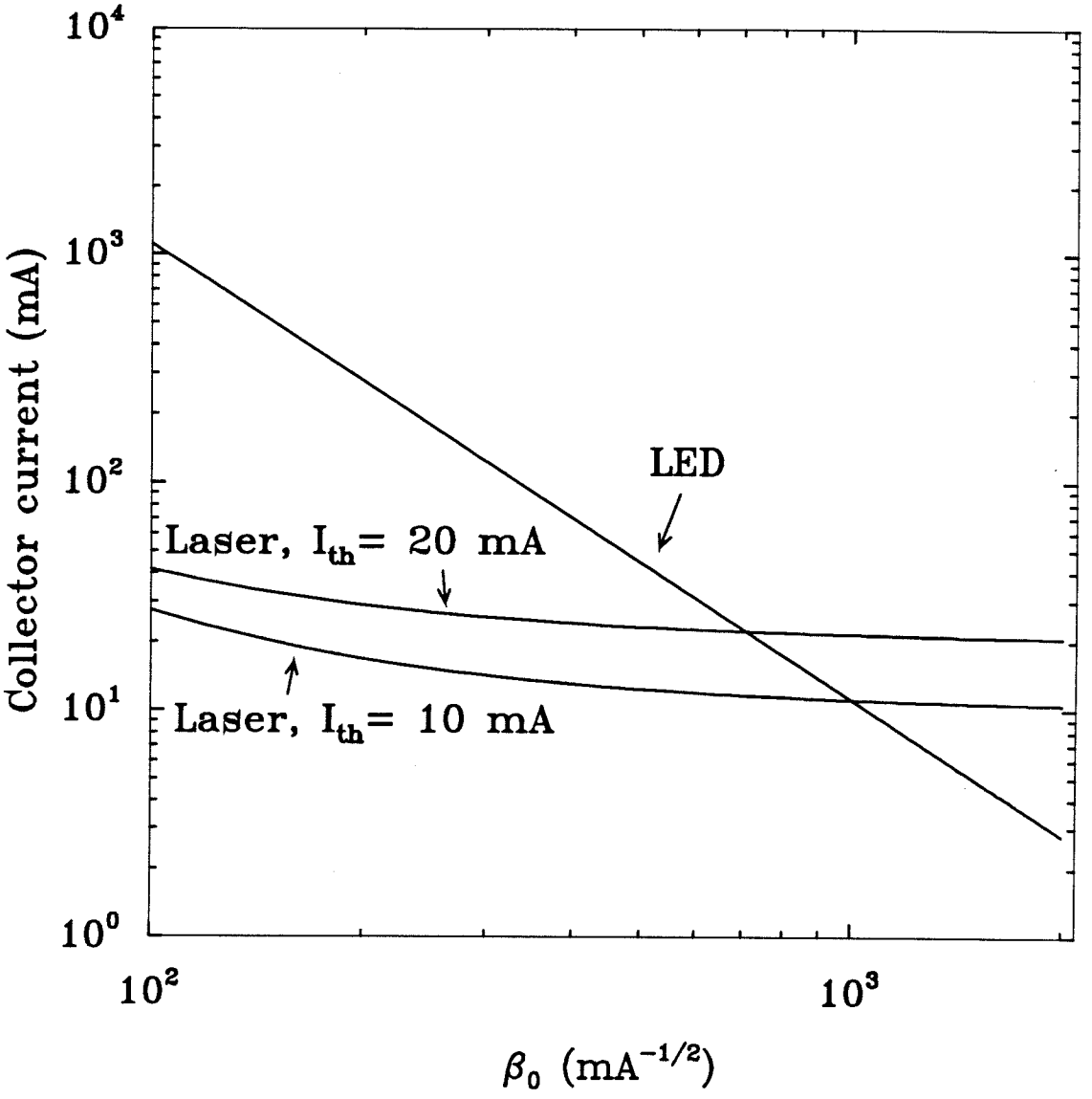


Fig. 6.5 Collector current needed to close the optical loop with a loop gain of one as a function of the transistor current gain coefficient, β_0. The holographic connection efficiency is assumed to be 0.1.

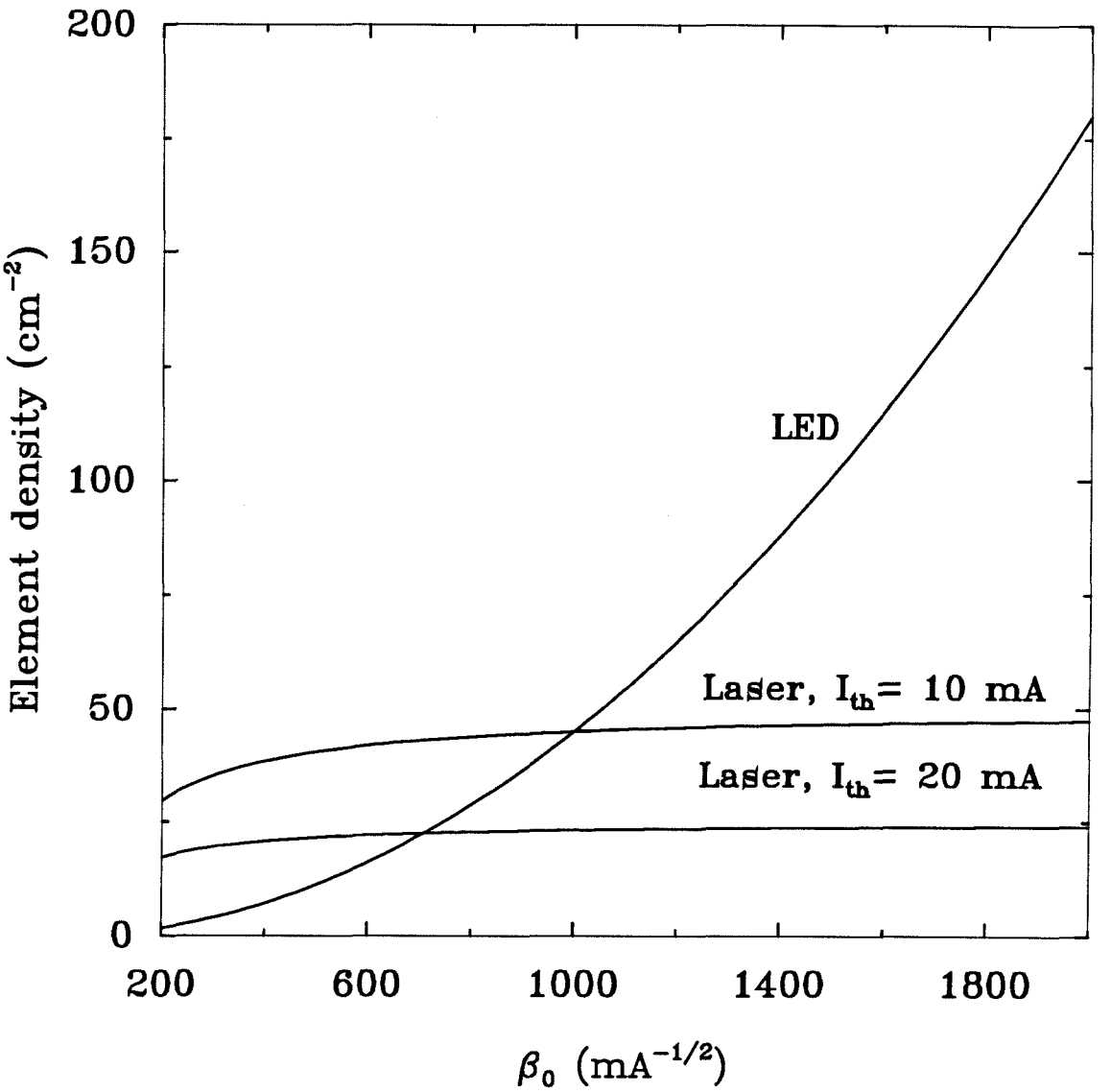


Fig. 6.6 Maximum array density allowed in order to close the optical loop with a loop gain of one as a function of the transistor current gain coefficient, β_0. The holographic connection efficiency is assumed to be 0.1.

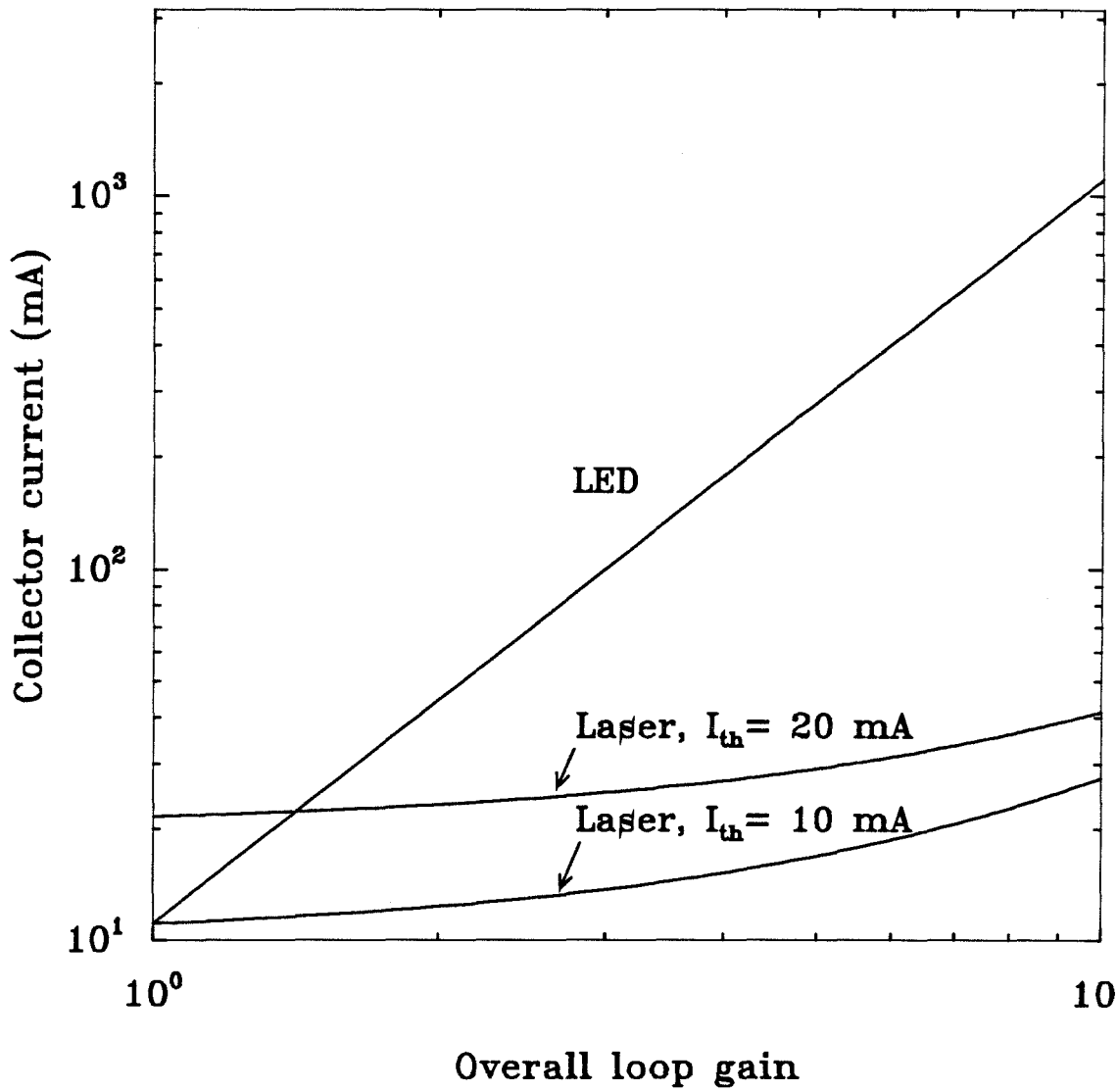


Fig. 6.7 Collector current needed to close the optical loop as a function of the loop gain.

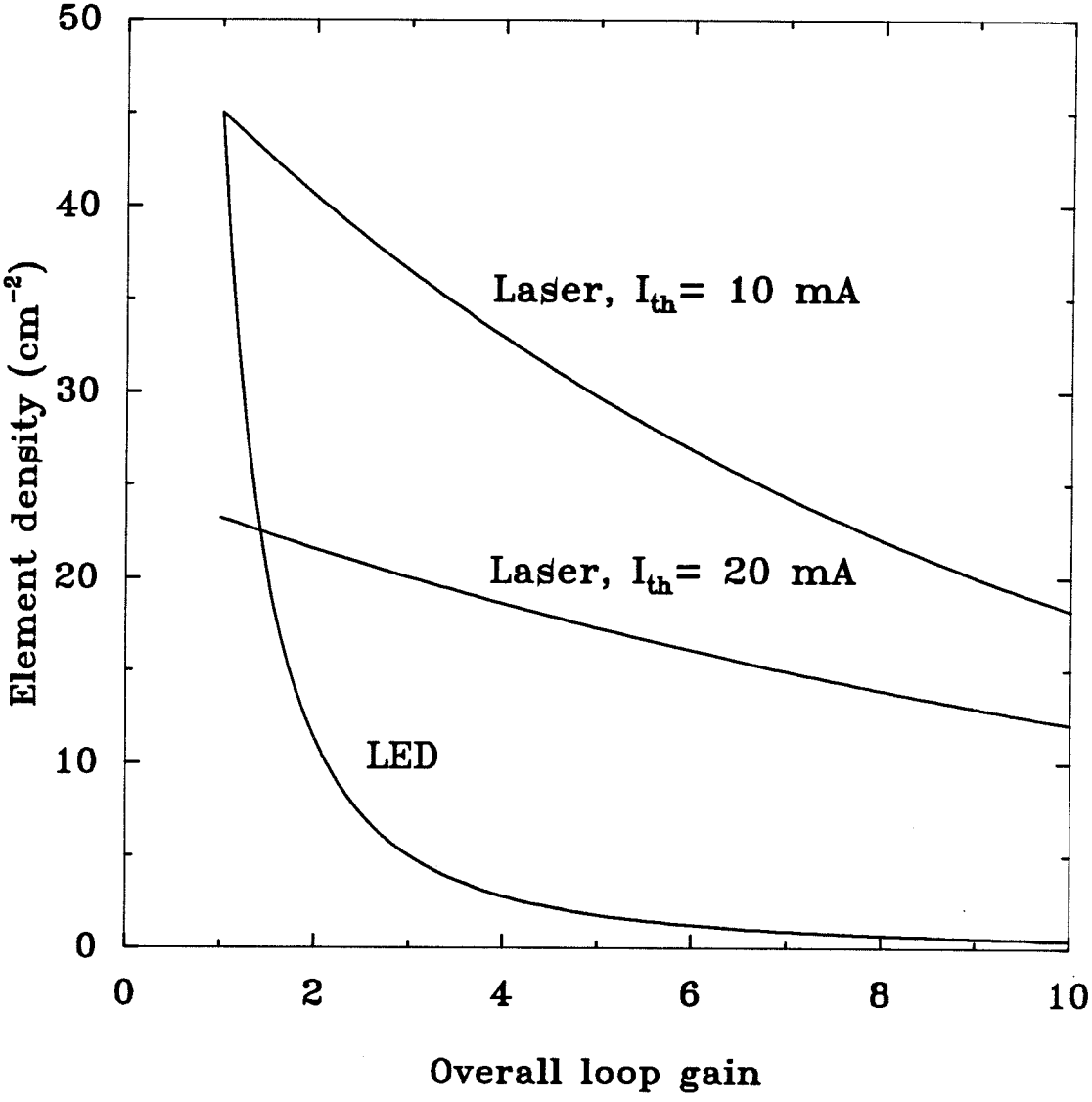


Fig. 6.8 Maximum array density under the condition that the optical loop is closed as a function of the loop gain.

the array density for the LED-based neuron decreases dramatically, owing to the fact that a very large current is needed to meet the loop gain requirement. Thus, the electrical power dissipation increases significantly. Even though laser diode seems to be the better choice as the loop gain increases, the density array for the laser-based neurons is also very small. This is limited by the dependence of the bipolar transistor's current gain on the driving current level. As a result, it will be very difficult to obtain a very large array of DHBT-based neurons that operate at a satisfactory electrical power dissipation level. Nevertheless, the analysis in this section shows clearly the tradeoff between the LED and laser diode.

6.3 MESFET-Based Neurons

The comparison of LED vs. laser diode for the MESFET-based optoelectronic neurons can be analyzed in a similar fashion as in the case for the DHBT-based neurons. Referring back to Fig. 6.1(b) in which the schematic circuit diagram of the MESFET-based neuron is shown, we see that the total electrical power dissipation of a MESFET-based neuron is given by

$$P_{elect} = V_{DD} I_p + V_{CC} I_d, \tag{6.23}$$

where I_p and I_d are currents flowing in the input switching circuit and output driving circuit respectively. Because I_d is usually much larger than I_p, Eq. (6.23) can be approximated by

$$P_{elect} \approx V_{CC} I_d. \tag{6.24}$$

Therefore, the maximum array density, assumed to be limited by the electrical power dissipation, is given by

$$\begin{aligned} N &= \frac{P_{elect,max}}{V_{CC} I_d} \\ &= \frac{P_{elect,max}}{V_{CC} V_{DD} g_m}, \end{aligned} \tag{6.25}$$

where the source-drain current flowing through the output driving MESFET, I_d, is substituted by the product of the transconductance of the MESFET, g_m, and the swing in the gate voltage, which is approximately V_{DD}. Therefore, to increase the array density, each of the parameters in the denominator of Eq. (6.25) should be decreased. However, decreasing g_m has an adverse effect on the speed of the neuron and should be considered as the last resort. The value of V_{CC} is limited by the forward turn-on voltage of either the LED or the laser diode and the source-drain voltage of the MESFET. Consequently, it is usually not less than 2 volts. As a result, the most logical choice for obtaining a denser array is to decrease the value of V_{DD}. Not only is the reduction in V_{DD} beneficial for the density of the array, but also the amount of the input light needed to switch the gate voltage of the output driving MESFET by V_{DD} is less as V_{DD} is reduced. This directly translates into an improvement in the optical gain of the neuron as long as the reduced input light power is not limited by the detector noise power. It should be noted that Eq. (6.25) is derived strictly from the circuit's point of view and is valid for both LED and laser diode as the light emitter.

If we now consider the optical feedback of the output back to the input detector, a slightly different picture develops between the LED and the laser diode. For LED, the array density is still given by Eq. (6.25). However, by considering the optical feedback, we can re-derive a different version of Eq. (6.25), which involves the speed

of the neuron. This is shown in the following.

$$\begin{aligned} P_{elect,max} &= NV_{CC}I_d \\ &= \frac{NV_{CC}P_{out}}{\eta_{LED}} \\ &= \frac{NV_{CC}P_{in}}{\eta_{LED}\cdot\eta_H} \\ &= \frac{NV_{CC}I_p}{\eta_{LED}\cdot\eta_H\cdot\eta_D} \\ &= \frac{NV_{CC}V_{DD}C_g}{\eta_{LED}\cdot\eta_H\cdot\eta_D\cdot\tau}, \end{aligned} \tag{6.26}$$

where the LED current, I_d, is sequentially converted into the photocurrent generated by the phototransistor, I_p, at the input through the various optical elements whose efficiencies are η_{LED}, η_H, and η_D. I_p is further substituted by the expression, C_gV_{DD}/τ, where τ is the time taken to charge up the gate to V_{DD}. This substitution can be justified by the fact that the total charge required to charge the gate voltage by V_{DD} is the product of this voltage and the gate capacitance, C_g. This product has to equal to the product of I_p and τ. Re-arranging Eq. (26), we obtain

$$N = \frac{P_{elect,max}\cdot\eta_{LED}\cdot\eta_H\cdot\eta_D}{V_{CC}V_{DD}C_g}\cdot\tau. \tag{6.27}$$

Eq. (6.27) has the following interpretation. From the consideration of optical feedback, the array density is linearly proportional to the response time of the neuron. The longer the response time can be, the weaker the input light power is allowed, which, in turn, implies that the LED does not have to be driven as hard. Therefore, the array density can be increased due to the reduction in the driving current. This represents a trade-off between the array density and the response time of the neuron. However, this trade-off ceases to exist if we combine the result obtained

from the optical consideration to that obtained from the circuit consideration. In other words, the implication that the response time of the neuron can be arbitrarily traded off to obtain a very high density array, as derived from Eq. (6.27), is not valid if we consider the fact that there is an upper bound on the array density as dictated by Eq. (6.25). The result presented in Eq. (6.25) is irrespective of how the outside optical interconnections are specified and designed. If we designate τ_o to be the switching time corresponding to this transition, we can solve for τ_o by substituting N from Eq. (6.25) into Eq. (6.27) and solve for τ. The result is shown in the following equation.

$$\tau_o = \frac{C_g}{g_m \cdot \eta_{LED} \cdot \eta_H \cdot \eta_D}. \tag{6.28}$$

It is interesting to note that C_g/g_m represents the intrinsic response time of the MESFET. The response time shown in Eq. (6.28) is equal to the intrinsic response time of the transistor degraded by the various inefficiency of the optical elements. What this means physically is that, when the input power is so large that gate charging time is negligible, the response time of the neuron is primarily dominated by the intrinsic response time of the MESFET. As the input power is allowed to decrease so as to increase the array density, the overall response time of the neuron starts to increase and will continue to increase only up to the value given by Eq. (6.28). Beyond which, any increase allowed in the response time of the neuron will not help increase the array density because of the circuit power dissipation constraint. Therefore, Eq. (6.28) represents an upper bound over which Eq. (6.27) is valid.

To get a feeling for the magnitudes of N and t, let's assume the following parameters : $P_{elect,max}$ = 1 W/cm^2, V_{CC} = 2 V, V_{DD} = 1 V, g_m = 2 mA/V, η_{LED} = 0.01 W/A, η_H = 0.1, η_D = 0.3 A/W, and C_g = 0.7 pF, which represents an 1000

Å depletion layer beneath a gate of $6\mu m \times 100\mu m$. From Eq. (6.25), the maximum array density that can be achieved is $250/cm^2$ and the corresponding τ_o is 1.15 μsec. From these numbers, we see that if we would decrease V_{DD} from the present 1 V to 0.1 V, the ratio of N to τ would increase by a factor of 10. This factor of 10 gain can be used to increase the array density or to decrease the neuron response time or both. However, it is probably best to increase the array density by a factor of 10, since the response time of the neuron may be decreased by decreasing the gate capacitance, as shown in Eq. (6.28). By doing this, a large array of neurons with small response time, that leads to small optical switching energy, can be obtained without too much compromise.

The analysis for laser diode is straight forward and follows the same procedure as shown above. It can be shown that, for laser diode,

$$P_{elect,max} = NV_{CC}\left(\frac{V_{DD}C_g}{\eta_{LD} \cdot \eta_H \cdot \eta_D \cdot \tau} + I_{th}\right). \tag{6.29}$$

Re-arranging Eq. (6.29), we obtain

$$N = \frac{P_{elect,max} \cdot \eta_{LD} \cdot \eta_H \cdot \eta_D \cdot \tau}{V_{CC}\left(V_{DD}C_g + \eta_{LD} \cdot \eta_H \cdot \eta_D \cdot I_{th} \cdot \tau\right)}. \tag{6.30}$$

The corresponding τ_o for the laser diode-based neuron is

$$\tau_o = \frac{C_g V_{DD}}{\eta_{LD} \cdot \eta_H \cdot \eta_D \cdot \left(V_{DD}g_m - I_{th}\right)}. \tag{6.31}$$

As a simple check, the expression in Eq. (6.31) reduces to Eq. (6.28) as I_{th} is set to zero. To compare the maximum array density that can be achieved as a function of the response time of the neuron for both the LED and the laser diode cases, Eq.

(6.25), (6.27) and (6.30) are illustrated in Fig. 6.9, which show both the circuit and the optical limitations of the neuron. In Fig. 6.9, the array density for the laser diode-based neuron is assumed to be limited by the threshold current. Thus, it is not constrained by Eq. (6.25).

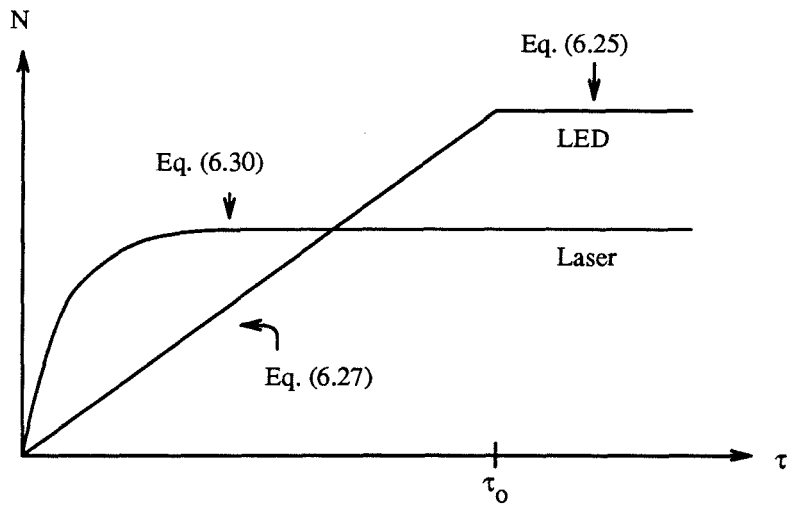


Fig. 6.9 The maximum array density achievable as a function of the response time for the LED-based neuron and the laser diode-based neuron.

Chapter 7

Conclusion

In summary, two different versions of monolithically integrated optoelectronic neurons are presented in this thesis. The first one utilized two double-heterojunction bipolar transistors, cascaded in a Darlington transistor pair configuration, to drive a LED, which provided the optical output. The input bipolar transistor also functioned as a phototransistor, on which the input light was focused. Due to the surface recombination current in the bipolar transistors and the emitter bulk current, the current gain of the Darlington transistor pair was not sufficiently high enough to provide useful optical gain for the overall optical neuron. It was found that, by etching away all but a thin layer of the high bandgap AlGaAs emitter layer, the remaining AlGaAs, which was completely depleted due to surface depletion, acted very effectively as a surface passivation material for the bipolar transistors. This resulted in a dramatic improvement in the overall current gain of the Darlington transistor pair, namely from 10 to 6000. Although the current gain of 6000 was sufficient to close the optical loop, the electrical power dissipation required to achieve this gain was 100 mW per neuron. This severely limited the density of the optoelectronic neurons in an array. Thus, an alternative integration approach was explored.

Being mature in its technology and process requirement, MESFET's were used as the driver device in the next generation optoelectronic neurons. In order to avoid direct coupling of output and input, an input switching circuit, composed of a phototransistor and a MESFET, was designed. The switched voltage was then used to drive the output MESFET, which, in turn, drove the LED. By isolating

the output from its input, an optical differential gain of 6 was measured in this MESFET-based neuron. In addition, only 2.4 mW electrical power dissipation was needed to drive this neuron. This represented a reduction by a factor of 40 when compared to the DHBT-based neuron. Further refinement in this MESFET-based neuron included the reduction in the gate leakage current and the substitution of the phototransistor by an optical FET. These resulted in dramatic improvements, such as an increase in the differential optical gain to 80, a decrease in the optical switching power to 54 nW, and a decrease in the electrical power dissipation to 1.6 mW per neuron, while maintaining a low optical switching energy of 10 pJ.

There is still a lot of room for improvement in the MESFET-based optoelectronic neurons. One easy modification is the replacement of the loading MESFET in the input switching circuit by another optical FET. This is illustrated in Fig. 7.1. In this circuit, there are two optical inputs, P_1 and P_2. P_1 functions as the conventional input with the threshold determined by the photocurrent produced by P_2. This produces an input-output characteristic for an excitatory neuron. The measured input-output characteristics for the neurons shown in Ch. 5 are all for excitatory neurons. However, the role of P_1 and P_2 can be reversed to obtain an input-output characteristic for an inhibitory neuron. The application of light at P_2 brings the gate voltage of the output driving MESFET down to ground. Thus, the LED is turned off. In this case, the threshold at which this turn-off occurs is determined by P_1. Not only does this modification represent an additional flexibility in the use of the neuron, it also will greatly simplify the fabrication process of the neuron array since it only requires two electrical wires connected to every neuron. This avoids the necessity of implementing a two-layer metalization scheme in which the complexity of the process increases dramatically and more complicated and expensive semiconductor processing equipments are required, such as low-temperature chemical vapor deposition systems. However, the price paid for the extra flexibility

and simplicity is the complication in the optical setup as two optical inputs are now required.

Another improvement that can increase the overall uniformity of all the neurons across the array is the employment of an etch stop layer, inserted between the MESFET conduction n^- GaAs layer and the n^+ GaAs layer above it. Currently, the variation in the gate threshold voltage stems from the variation in the gate recessed depth for all MESFET's. This variation introduces a non-uniform channel thickness. As a result, some MESFET's are in the enhancement mode, while some are in the depletion mode. The degree of enhancement or depletion also varies. This results in a wide range of pinch-off voltage for all MESFET's. This is a serious problem because this will cause significant deviations and errors in the performance of the neurons in the array. To remedy this problem, a very controlled etch down to the gate is absolutely necessary. Unfortunately, most of the wet chemical etching schemes are not consistent enough to yield the same etch rate or depth every time mainly due to the different conditions under which the chemicals are prepared and agitated before each use. Therefore, a selective etching scheme has to be developed. This is accomplished by inserting a high aluminum mole fraction material at the location where the etching is intended to stop. This is shown in Fig. 7.2(a). The inserted AlAs layer underneath the n^+ GaAs ohmic contact layer will stop the etching process, which only etches the GaAs and leaves AlAs unattacked. Next, a different selective etchant, which only etch AlGaAs and leaves GaAs unattacked, selectively removes only the AlAs layer. This leaves the n^- GaAs conduction layer exposed for gate contact, In this way, the pinch-off voltage is determined solely by the doping concentration and the thickness of this n^- GaAs layer. Thus, the factors of uncertainty due to processing are completely eliminated so that a very uniform neuron array can be achieved.

While the previously mentioned two suggestions promise significant improve-

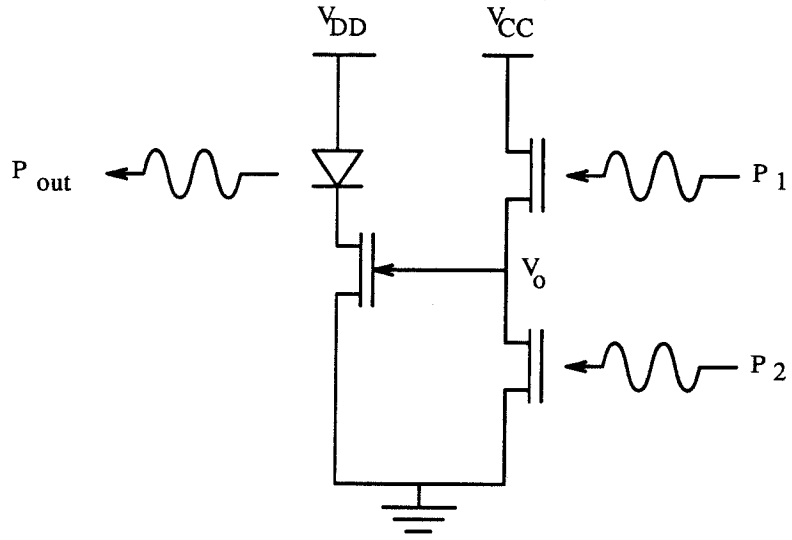


Fig. 7.1 The replacement of the loading MESFET in the input switching circuit by an optical FET makes the overall optoelectronic neuron much more flexible to use and simpler to fabricate.

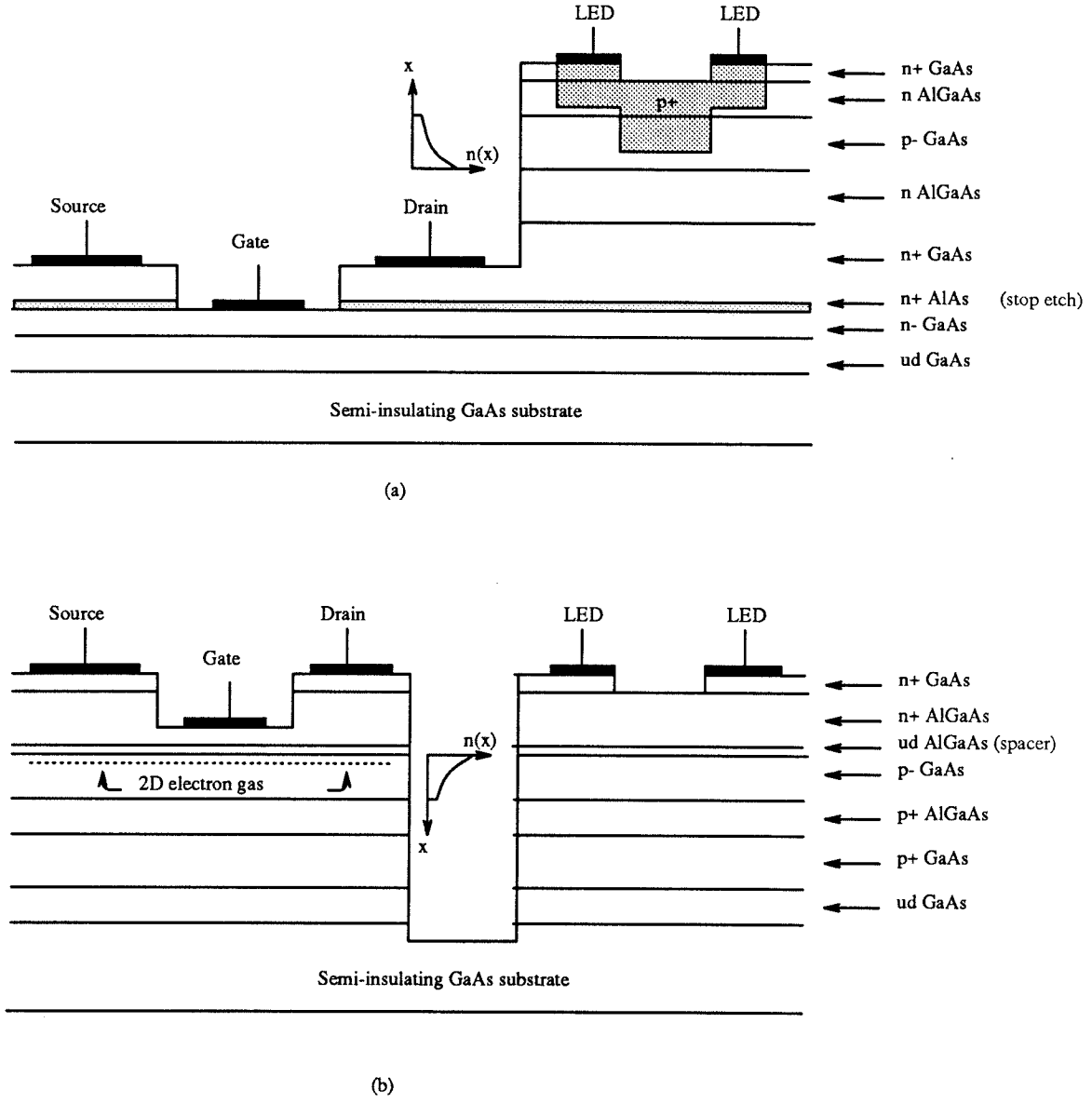


Fig. 7.2 **(a)** The integrated neuron structure employing a stop-etch scheme in which an AlAs layer above the n^- GaAs layer is inserted. **(b)** The proposed integrated neuron structure employing a combination of LED and HEMT to eliminate the non-planar surface topology and to increase the optical gain of the neuron.

ments in the performance of either the individual neuron or the overall neuron array, they still suffer from two drawbacks. Firstly, the overall integrated neuron has a very non-planar topology. This is very undesirable for process controllability. Because of the non-planarity due to various recessed etching, uniform coating of photoresist becomes extremely difficult. Non-uniform photoresist introduces non-uniform UV exposure during the photoresist exposure step. This leads to a non-uniform post-development photoresist profile. Such a problem is most often seen in a situation where some part of the photoresist is properly developed while the other part is either under-developed or over-developed. As a result, the original mask pattern, however sharp on the original mask plate, loses some accuracy when transferred onto the photoresist. This subsequently leads to uncontrolled processing of the device. The second drawback is the fact that, in the LED, the generated photon density is higher near the bottom GaAs-AlGaAs heterojunction because electrons are injected from the bottom n-type AlGaAs layer. In order to extract these photons out from the top, they have to traverse through the thick p^- active GaAs layer, which is absorptive. In order to solve this problem, electrons have to be re-directed to inject from the top down into the active layer. This requires placing the n-AlGaAs layer on the top and p-AlGaAs layer at the bottom of the GaAs active layer. Thus, the conventional P-i-N LED diode becomes an N-i-P in structure, which, coincidentally, also solves the dilemma of non-planar surface topology in the conventional MESFET-based neuron. The proposed structure is illustrated in Fig. 7.2(b). Because the i-GaAs active layer in the LED is actually p^-, there is a two dimensional quantum well at the hetero-interface between the active layer and the n-type AlGaAs upper cladding layer. Thus, a two dimensional electron gas is formed there. By placing two ohmic contacts on the top and recessing the region in between slightly, we obtain the structure of a high electron mobility transistor (HEMT) [101]. The two ohmic contacts are the source and drain contacts, and the

recessed region defines the gate of the HEMT. HEMT's have much higher transconductance compared to MESFET's. This is important for optical neurons because this means that the swing in the gate voltage does not need to be as large to obtain the same output current to drive the LED. A smaller swing in the gate voltage implies a smaller input optical power required to turn on the neuron. Therefore, the optical gain of the neuron based on HEMT's is expected to be higher than the MESFET counterpart. The optical detector required for the neuron can also be easily fabricated by removing the gate in the HEMT. Therefore, the ungated region is the photo-sensitive region and the incident light modulates the population of the electrons in the quantum well thus the current through in the HEMT. This is somewhat analogous to the MESFET-optical FET relationship. As can be seen from Fig. 7.2(b), a very planar structure is obtained in this HEMT-LED integration except for the etch down to the substrate in order to isolate each neuron and the very slight recessed etch for the gate of the HEMT, which makes threshold voltage in the HEMT much more controllable.

All in all, with the incremental improvements outlined above, the performance of the optoelectronic neuron is expected to achieve a level that makes possible a dense integration of these devices in a large array form. Specifically, an optical gain in the order of 100, which is not far from the present demonstrated gain of 80, should be easily obtained. Electrical power dissipation on the order of 100 μW per neuron should be achievable if the detector sensitivity is increased and the leakage current in the circuit is decreased. This allows the operating current in the neuron to be lowered without sacrificing optical gain. With 100 μW power dissipation per neuron, an array of 100 $\times$ 100 optoelectronic neurons can be inserted into an optical system in which practical optical interconnnects are demonstrated with holograms without being limited by the heat sinking capability of the chip. Finally, along with the increase in the input detector sensitivity, the optical switching power

is expected to decrease proportionally. With further optimization in minimizing the gate capacitance of the transistor, the overall optical switching energy is also expected to decrease. At that state, the limiting factor in expanding the size of the array will not be the heat dissipation issue, but rather and most likely the photolithography capability. Nevertheless, these neurons will be sufficiently good for use in most of the optical parallel computation systems.

Reference

[1] D. Psaltis, and N. H. Farhat, "Optical Information Processing Based on an Associative-Memory Model of Neural Nets with Thresholding and Feedback," Opt. Lett., Vol. 10, pp. 98, 1985.

[2] Y. S. Abu-Mostafa, and D. Psaltis, "Optical Neural Computers," Scientific American, Vol. 256, pp. 88, 1987.

[3] N. H. Farhat, "Optoelectronic Analogs of Self-Programming Neural Nets : Architecture and Methodologies for Implementing Fast Stochastic Learning by Simulated Annealing," Appl. Opt., Vol. 26, pp. 5093, 1987.

[4] Y. Owechko, G. J. Dunning, E. Marom, and B. H. Soffer, "Holographic Associative Memory With Nonlinearities in the Correlation Domain," Appl. Opt., Vol. 26, pp. 1900, 1987.

[5] D. Z. Anderson, "Coherent Optical Eigenstate Memory," Opt. Lett., Vol. 11, pp. 56, 1986.

[6] A. Yariv, and S. K. Kwong, "Associative Memories Based on Message-Bearing Optical Modes in Phase-Conjugate Resonators," Opt. Lett., Vol. 11, pp. 186, 1986.

[7] C. Guest, and R. Te Kolste, "Designs and Devices for Optical Bidirectional Associative Memories," Appl. Opt., Vol. 26, pp. 5055, 1987.

[8] R. A. Athale, H. H. Szu, and C. B. Friedlander, "Optical Implementation of Associative Memory with Controlled Nonlinearity in the Correlation Domain," Opt. Lett., Vol. 11, pp. 482, 1986.

[9] F. H. Mok, M. C. Tackitt, and H. M. Stoll, "Storage of 500 High-Resolution Holograms in a $LiNBO_3$ Crystal," Opt. Lett., Vol. 16, pp. 605, 1991.

[10] D. Psaltis, D. Brady, and K Wagner, "Adaptive Optical Networks Using Pho-

torefractive Crystals," Appl. Opt., Vol. 27, pp. 1752, 1988.

[11] E. G. Paek, and D. Psaltis, "Optical Associative Memory Using Fourier Transform Holograms," Opt. Eng., Vol 26, pp. 428, 1987.

[12] W. P. Bleha, L. T. Lipton, E. Wiener-Avnear, J. Grinberg, P. G. Reif, D. Cassasent, H. B. Brown, and B. V. Markevitch, "Application of the Liquid Crystal Light Valve to Real-Time Optical Data Processing," Opt. Eng., Vol. 17, pp. 371, 1978.

[13] L. K. Cotter, T. J. Drabik, R. J. Dillon, and M. A. Handschy, "Ferroelectric Liquid Crystal Silicon Integrated Circuit Spatial Light Modulator," Opt. Lett., Vol. 15, pp. 291, 1990.

[14] G. H. Haertling, "PLZT Electrooptic Materials and Applications - a Review," Ferroelectrics, Vol 75, pp. 25, 1987.

[15] H. K. Choi, G. W. Turner, J. C. C. Fan, J. M. Phillips, and B. Y. Tsaur, "Prospects for Monolithic GaAs/Si Integration," in Heteroepitaxy on Silicon II (J. C. C. Fan, J. M. Phillips, and B. Y. Tsaur eds.). Pittsburgh, PA : Material Research Society, pp. 213, 1987.

[16] D. R. Mayers, J. F. Klem, and J. A. Lott, "(AlGa)As/(InGa)As Strained-Quantum-Well FET's on Silicon Dioxide by Selective Device Lift-off as an Alternative to Heteroepitaxy," Tech. Dig. of International Electron Device Meeting, New York, pp. 704, 1988.

[17] N. Bar-Chaim, S. Margalit, A. Yariv, and I. Ury, "GaAs Integrated Optoelectronics," IEEE Trans. Electron Dev., Vol. ED-29, pp. 1372, 1982.

[18] S. M. Sze, "Physics of Semiconductor Devices," pp. 50, John Wiley & Sons, Inc., 1981.

[19] T. P. Lee, and A. G. Dentai, "Power and Modulation Bandwidth of GaAs-AlGaAs High-Radiance LED's for Optical Communication Systems," IEEE J. Quantum Electron., Vol. QE-14, pp. 150, 1978.

[20] H. Kressel, M. Ettenberg, J. P. Wittke, and I. Kadany, "Semiconductor Devices for Optical Communications," Top. Appl. Phys., Vol. 39, pp. 9, Ed. H. Kressel, Spring-Verlag, 1980.

[21] D. Marcuse, "LED Fundamentals : Comparison of Front- and Edge-Emitting Diodes," IEEE J. Quantum Electron., Vol.QE-13, pp. 819, 1977.

[22] J. Dziwior, and W. Schmid, "Auger Coefficients for Highly Doped and Highly Excited Silicon," Appl. Phys. Lett., Vol. 31, pp. 346, 1977.

[23] H. C. Casey Jr. and F. Stern, "Concentration Dependent Absorption and Spontaneous Emission of Heavily Doped GaAs," J. Appl. Phys., Vol. 47, pp. 631, 1976.

[24] G. A. Acket, W. Nijam, and H. Lam, "Electron Lifetime and Diffusion Constant in Germanium-Doped Gallium Arsenide," J. Appl. Phys., Vol. 45, pp. 3033, 1974.

[25] E. Hecht and A. Zajac, "Optics," Ch. 9, Addison & Wesley, 1979.

[26] B. Tell, Y. H. Lee, K. F. Brown-Goebeler, J. L. Jewell, R. E. Leibenguth, M. T. Asom, G. Livescu, L. Luther, and V. D. Mattera, "High-Power CW Vertical Cavity Top Surface-Emitting GaAs Quantum Well Lasers," Appl. Phys. Lett., Vol. 57, pp. 1855, 1990.

[27] S. H. Lin, J. H. Kim, J. Katz, and D. Psaltis, "Integration of High-Gain Double Heterojunction GaAs Bipolar Transistors with a LED for Optical Neural Network Applications," Proceedings of IEEE/Cornell Conference on Advanced Concepts in High Speed Semiconductor Devices and Circuits, pp. 344, Aug. 7-9, 1989.

[28] R. L. Thorton, R. D. Burnham, and W. Streifer, "High Reflectivity GaAs-AlGaAs Mirrors Fabricated by Metalorganic Chemical Vapor Deposition," Appl. Phys. Lett., Vol. 45, pp. 1028, 1984.

[29] F. W. Ostermayer, Jr., P. A. Kohl, and R. H. Burton, "Photoelectrochemical

Etching of Integral Lenses on InGaAsP/InP Light-Emitting Diodes," Appl. Phys. Lett., Vol. 43, pp. 642,1983.

[30] J. Heinen, "Preparation and Properties of Monolithically Integrated Lenses on InGaAsP/InP Light-Emitting Diodes," Electron. Lett., Vol. 18, pp. 831, 1982.

[31] O. Wada, S. Yamakoshi, M. Abe, Y. Nishitani, and T. Sakura, "High Radiance InGaAsP/InP Lensed LED's for Optical Communication Systems at 1.2-1.3 μm," IEEE J. Quantum Electron., Vol. QE-17, pp. 174, 1981.

[32] H. Kroemer,"Heterostructure Bipolar Transistors and Integrated Circuits," Proc. IEEE, Vol. 70, pp. 13, 1982.

[33] S. L. Su, O. Tejayadi, T. J. Drummond, R. Fisher, and H. Morkoc, "Double Heterojunction Al_xGa_{1-x} As/GaAs Bipolar Transistors (DHBJT's) by MBE with a Current Gain of 1650," IEEE Electron. Dev. Lett., Vol. 4, pp. 130, 1983.

[34] J. Temmyo, Y. Hasumi, and A. Kozen, "MOVPE Grown GaAlAs/GaAs DH Devices operating as Laser Diodes and Bipolar Transistors for Optoelectronic Integration," Proceeding of Device Research Conference, IIIA-6, 1987.

[35] Y. Chen, R. N. Nottenburg, M. B. Panish, R. A. Hamm, and D. A. Humphrey, "Subpicosecond InP/InGaAs Heterostructure Bipolar Transistors," IEEE Electron. Dev. Lett., Vol. 10, pp. 267, 1989.

[36] P. M. Asbeck, M. F. Chang, J. A. Higgins, N. H. Sheng, G. J. Sullivan, and K. Wang, "GaAlAs/GaAs Heterojunction Bipolar Transistors : Issues and Prospect for Application," IEEE Trans. Electron. Dev., Vol. 36, pp. 2032, 1989.

[37] P. M. Enquist, J. A. Hutchby, M. F. Chang, P. M. Asbeck, N. H. Sheng, and J. A. Higgins, "High-Frequency Performance of MOVPE npn AlGaAs/GaAs Heterojunction Bipolar Transistors," Electron. Lett., Vol. 25, pp. 1124, 1989.

[38] H. Ito, T. Ishibashi, and T. Sugeta, "Fabrication and Characterization of Al-

GaAs/GaAs Heterojunction Bipolar Transistors," IEEE Trans. Electron. Dev., Vol. 34, pp. 224, 1987.

[39] W. Lee and C. G. Fonstad, "$In_{0.52}Al_{0.48}As/In_{0.53}$ $Ga_{0.47}As/In_{0.52}Al_{0.48}As$ Abrupt Double-Heterojunction Bipolar Transistors," IEEE Electron. Dev. Lett., Vol. 7, pp. 683, 1986.

[40] J. R. Hayes, F. Capasso, A. C. Gossard, R. J. Malik, and W. Wiegmann, "Bipolar Transistors with Graded Band-gap Base," Electron Lett., Vol. 19, pp. 410, 1983

[41] M. F. Chang, P. M. Asbeck, D. L. Miller, and K. C. Wang, " GaAs/(GaAs)As Heterojunction Bipolar Transistors Using a Self-Aligned Substitutional Emitter Process," IEEE Electron. Dev. Lett., Vol. 7, pp. 8, 1986.

[42] T. Makimoto, and N. Kobayashi, "AlGaAs/GaAs Heterojunction Bipolar Transistors with Heavily C-Doped Base Layers Grown by Flow-Rate Modulation Epitaxy," Appl. Phys. Lett., Vol. 54, pp. 39, 1988.

[43] S. Adachi, and T. Ishibashi, "Collector-Up HBT's Fabricated by Be^+ and O^+ Ion Implantation," IEEE Electron Dev. Lett., Vol. 7, pp. 32, 1986.

[44] T. Izawa, T. Ishibashi, and T. Sugeta, "AlGaAs/GaAs Heterojunction Bipolar Transistors," International Electron. Dev. Meeting, pp. 328, 1985.

[45] J. R. Hayes, F. Capasso, R. J. Malik, A. C. Gossard, and W. Wiegmann, "Optimum Emitter Grading for Heterojunction Bipolar Transistors," Appl. Phys. Lett., Vol. 43, pp. 949, 1983.

[46] D. L. Miller, P. M. Asbeck, R. J. Anderson, and F. H. Eisen, "(GaAl)As/GaAs Heterojunction Bipolar Transistors with Graded Composition in the Base," Electron. Lett., Vol. 19, pp 367, 1983.

[47] S. Tiwari, "GaAlAs/GaAs Heterojunction Bipolar Transistors : Experiment and Theory," International Electron. Dev. Meeting, pp. 262, 1986.

[48] R. J. Malik, F. Capasso, R. A. Stall, R. A. Kiel, R. W. Ryan, R. Wunder, and

C. G. Bethea, "High-Gain, High-Frequency AlGaAs/GaAs Graded Band-gap Base Bipolar Transistors with a Be Diffusion Setback Layer in the Base," Appl. Phys. Lett., Vol. 46, pp. 600, 1985.

[49] A. Scavennec, D. Ankri, C. Besombes, C. Courbet, J. Riou, and F. Heliot, "High-Gain Low-Noise GaAlAs/GaAs Phototransistors," Electron. Lett., Vol. 19, pp. 394, 1983.

[50] D. Ankri, A. Scavennec, C. Besombes, C. Courbet, F. Heliot, and J. Riou, "Diffused Epitaxial GaAsAs/GaAs Heterojunction Bipolar Transistors for High-Frequency Operation," Appl. Phys. Lett., Vol. 40, pp. 816, 1982.

[51] T. R. Chen, Y. H. Zhuang, B. Chang, M. B. Yi, and A. Yariv, "A Stripe-Geometry InGaAsP/InP Heterojunction Bipolar Transistors Suitable for Optical Integration," IEEE Electron. Lett., Vol. 8, pp. 191, 1987.

[52] M. Konagai, and K. Takahashi, "(GaAl)As/GaAs Heterojunction Transistors with High Injection Efficiency," J. Appl. Phys., Vol. 46, pp. 2120, 1975.

[53] R. P. Tijburg, and T. Van Dongen, "Selective Etching of III-V Compounds with Redos Systems," J. Electrochem. Soc : Solid-State Science and Technology, Vol. 123, pp. 687, 1976.

[54] M. Konagai, K. Katsukawa, and K. Takahashi, "(GaAl)As/GaAs Heterojunction Phototransistors with High Current Gain," J. Appl. Phys., Vol. 48, pp. 4389, 1977.

[55] Y. S. Hiraoka, J. Yoshida, and M. Azuma, "Two-Dimensional Analysis of Emitter-Size Effect on Current Gain for GaAlAs/GaAs HBT's," IEEE Trans. Electron. Dev., Vol. 34, pp. 721, 1987.

[56] H. J. Lee, L. Y. Juravel, J. C. Wooley, and A. J. Spring-Thorpe, "Electron Transport and Band Structure of $Ga_{1-x}Al_xAs$ Alloys," Phys. Rev., Vol. B21, pp. 659, 1980.

[57] R. Dingle, "Confined Carrier Quantum States in Ultrathin semiconductor Het-

erostructure," Festkörperprobleme/Advances in Solid State Physics, Vol. 15, pp. 21, 1975.

[58] S. Tiwari, and D. J. Frank, "Analysis of the Operation of GaAlAs/GaAs HBT's," IEEE Trans. Electron. Dev., Vol. 36, pp. 2105, 1989.

[59] R. H. Saul, P. T. Lee, and C. A. Burrus, "Light-Emitting-Diode Device Design," Semiconductors and Semimetals, Vol. 22, Part C, Chapter 5, pp. 205, Academic Press, Inc., 1985.

[60] O. Nakajima, K. Nagata, H. Ito, T. Ishibashi, and T. Sugeta, "Emitter-Base Junction Size Effect on Current Gain H_{fe} of AlGaAs/GaAs Heterojunction Bipolar Transistors," Japan. J. Appl. Phys., Vol. 24, pp. L596, 1985.

[61] E. S. Yang, "Microelectronics Devices," pp. 104, McGraw-Hill, Inc., 1988.

[62] E. S. Yang, "Microelectronics Devices," pp. 125, McGraw-Hill, Inc., 1988.

[63] D. Ankri, R. Azoulay, E. Caquot, J. Dangla, C. Dubon, and J. F. Palmier, "Analysis of d.c. Characterization of GaAlAs/GaAs Double Heterojunction Bipolar Transistors," Solid State Electron., Vol. 29, pp. 141, 1986.

[64] C. Blaauw, A. J. Spring-Thorpe, S. Szioba, and B. Emmerstorfer, "CVD-SiO_2 and Plasma-SiN_x Films as Zn Diffusion Masks for GaAs," J. Electron. Material, Vol. 13, pp. 251, 1984.

[65] W. Lee, D. Ueda, T. Ma, Y. Pao, and J. S. Harris, "Effect of Emitter-Base Spacing on the Current Gain of AlGaAs/GaAs Heterojunction Bipolar Transistors," IEEE Electron Dev. Lett., Vol. 10, pp. 200, 1989.

[66] H. Lin, and S. Lee, "Super-Gain AlGaAs/GaAs Heterojunction Bipolar Transistors Using an Emitter Edge-Thinning Design," Appl. Phys. Lett., Vol. 47, pp. 839, 1985.

[67] T. Won, S. Iyer, S. Agarwala, and H. Morkoc, "Collector Offset Voltage of Heterojunction Bipolar Transistors Grown by Molecular Beam Epitaxy," IEEE Electron. Dev. Lett., Vol. 10, pp. 274, 1989.

[68] The Handbook of Chemistry and Physics, 48th ed., Chemical Rubber Co., 1967.

[69] B. J. Baliga, and S. K. Ghandhi, "PSG Masks for Diffusion in Gallium Arsenide," IEEE Trans. Electron. Dev., Vol. 19, pp. 761, 1972.

[70] S. K. Ghandhi, "VLSI Fabrication Principles," pp. 195, John Wiley & Sons, Inc., 1983.

[71] S. M. Sze, "Physics of Semiconductor Devices," pp. 92, John Wiley & Sons, Inc., 1981.

[72] M. Y. Ghannam, and R. P. Mertens, "Surface Recombination Current with a Nonideality Factor Greater Than 2," IEEE Electron. Dev. Lett., Vol. 10, pp. 242, 1989.

[73] P. Enquist, G. W. Wicks, and L. F. Eastman, "Anomalous Redistribution of Beryllium in GaAs Grown by Molecular Beam Epitaxy," J. Appl. Phys., Vol. 58, pp. 4130, 1985.

[74] D. L. Miller, and P. M. Asbeck, "Be Redistribution During Growth of GaAs and AlGaAs by Molecular Beam Epitaxy," J. Appl. Phys., Vol. 57, pp. 1816, 1985.

[75] S. M. Sze, "Physics of Semiconductor Devices," pp. 37, John Wiley & Sons, Inc., 1981.

[76] R. E. Reed-Hill, "Physical Metallurgy Principles," pp. 254 and 287, D. Van Nostrand Co., 1973.

[77] Z. D. Jastrzebski, "The Nature and Properties of Engineering Materials," pp. 109 and 128, John Wiley & Sons, Inc., 1976.

[78] Private communication with Prof. M. A. Nicolet of Caltech.

[79] Private communication with Prof. Y. C. Tai of Caltech.

[80] J. Katz, N. Bar-Chaim, P. C. Chen, S. Margalit, I. Ury, D. Wilt, M. Yust, and A. Yariv, "Monolithic Integration of a GaAlAs Buried-Heterostructure Laser

and a Bipolar Phototransistor," Appl. Phys. Lett., vol. 37, pp. 211, 1980.

[81] Y. Hasumi, A. Kozen, J. Temmyo, and H. Asahi, "A GaAs/GaAlAs Double-Heterojunction Device Function as a Bipolar Transistor and Injection Laser for Optoelectronic Integrated Circuits," IEEE Electron. Dev. Lett., vol. EDL-8, pp. 10, 1987.

[82] A. Cazarre, J. Tasselli, A. Marty, J. P. Baibe, and G. Rey, "GaAlAs/GaAs Heterojunction Bipolar Phototransistors Grown by LPE with a Current Gain of 50000," Electron. Lett., vol. 21, pp. 1124, 1985.

[83] K Taira, C. Takano, H. Kawai, and M. Arai, "Emitter Grading in AlGaAs/GaAs Heterojunction Bipolar Transistor Grown by Metalorganic Chemical Vapor Deposition," Appl. Phys. Lett., vol. 49, pp. 1278, 1986.

[84] J. H. Kim, S. H. Lin, J. Katz, and D. Psaltis, "Monolithically Integrated Two-Dimensional Arrays of Optoelectronic Threshold Devices for Neural Network Applications," SPIE Proc., vol. 1043-7, 1989.

[85] S. Sze, "Physics of Semiconductor Devices," pp. 190, John Wiley & Sons, Inc., 1981.

[86] L. F. Eastman, and M. S. Shur, "Substrate Current in GaAs MESFET's," IEEE Trans. Electron. Dev., Vol. ED-26, pp. 1359, 1979.

[87] S. Sze, "Physics of Semiconductor Devices," pp. 312, John Wiley & Sons, Inc., 1981.

[88] R. A. Sadler, and L. F. Eastman, "High-Speed Logic at 300K with Self-Aligned Submicrometer-Gate GaAs MESFET's," IEEE Electron. Dev. Lett., Vol. EDL-4, pp. 215, 1983.

[89] H. M. Levy, and R. E. Lee, "Self-Aligned Submicron Gate Digital GaAs Integrated Circuits," IEEE Electron. Dev. Lett., Vol. EDL-4, pp. 102, 1983.

[90] Private Communication with Mr. David Forgerson of Vitesse Semiconductor Corp., Camarillo, CA.

[91] R. E. Williams, "Gallium Arsenide Processing Techniques," pp. 101, Artech House, Inc., 1984.

[92] H. M. Macksey, F. H. Doerbeck, and R. C. Vail, "Optimization of GaAs Power MESFET Device and Material Parameters for 15-GHz Operation," IEEE Trans. Electron. Devices, ED-27, pp. 467, 1980.

[93] T. Sugeta, and Y. Mizushima, "High Speed Photoresponse Mechanism of a GaAs-MESFET," Jpn. J. Appl. Phys., Vol. 19, pp. L27, 1980.

[94] J. C. Gummel, and J. M. Ballantyne, "The OPFET : A New High Speed Optical Detector," Technical Digest of the International Electron Device Meeting, pp. 120, 1978.

[95] J. C. Gummel, and J. M. Ballantyne, Comments on "High Speed Photoresponse Mechanism of a GaAs-MESFET," Jpn. J. Appl. Phys., Vol. 19, pp. L273, 1980.

[96] A. L. Lentine, H. S. Hinton, D. A. B. Miller, J. E. Henry, J. E. Cunningham, and L. M. F. Chirovsky, "Symmetric Self-Electrooptic Effect Device : Optical Set-Reset Latch, Differential Logic Gate, and Differential Modulator/Detector," IEEE J. Quantum Electron., Vol. QE-25, pp. 1928, 1989.

[97] S. H. Lin, F. Ho, J. H. Kim, and D. Psaltis, "GaAs-Based Optoelectronic Neurons," Technical Digest of Optical Computing Conference, Salt Lake City, UT, March 4-6, pp. 295, 1991.

[98] S. H. Lin, F. Ho, J. H. Kim, and D. Psaltis, "Monolithic Integrated Optoelectronic Thresholding Devices for Neural Network Applications," Technical Digest of Conference on Lasers and Electro-Optics, Baltimore, MD, May 12-17, pp. 82, 1991.

[99] S. H. Lin, D. Psaltis, and J. H. Kim, "High-Gain GaAs Optoelectronic Thresholding Devices for Optical Neural Network Applications," Technical Digest of Integrated Photonic Research Conference, Monterey, CA, April 9-11, pp.8,

1991.

[100] K. Iga, S. Kinoshita, and F. Koyama, "Microcavity AlGaAs/GaAs Surface-Emitting Laser With $I_{th} = 6$ mA," Electron. Lett., Vol. 23, pp. 134, 1987.

[101] S. S. Pei, and N. J. Shah, "Heterostructure Field Effect Transistors," Introduction to Semiconductor Technology : GaAs and Related Compounds, Ch. 3, Ed. C. T. Wang, John Wiley & Sons, Inc., 1990.