A NEW MOS PHOTON COUNTING SENSOR OPERATING IN THE ABOVE-BREAKDOWN REGIME

by

TIMOTHY PAUL LESTER

B.Sc., The University of Victoria, 1975

M.Sc., The University of British Columbia, 1977

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

Department of Electrical Engineering

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

1982

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of ELECTRICAL ENGINEERING

The University of British Columbia
2075 Wesbrook Place
Vancouver, Canada
V6T 1W5

Date 03/06/82

ABSTRACT

A MOS optical sensor that utilizes avalanche multiplication in silicon is proposed and investigated both theoretically and experimentally. The above-breakdown operating regime is discussed and it is shown how a MOS photosensor may be operated in a photon counting mode by pulsing it into very deep depletion, beyond the point where avalanche breakdown normally occurs. Avalanche discharges in such a MOS sensor are self-quenching due to the formation of an inversion layer. This self-quenching property suggests that a monolithic self-scanned array of MOS photon counting sensors should be possible. It is described how specially designed charge-coupled arrays (PC-CCD's) could be operated in this new regime. The high response of silicon in the visible and near infrared, compared with the responsive quantum efficiency of the commonly-used photocathode materials, gives the proposed imager a distinct advantage over presently-existing photon counting sensors in these spectral regions.

It is shown that a PC-CCD must be fabricated on a p-type silicon substrate and illuminated from the back side in order to obtain a high avalanche initiation probability for the photogenerated carriers. It is also shown that all thermally activated, steady-state dark generation of carriers can be reduced to a negligible level by cooling the sensor to 100 K or less, while the generation due to interband tunneling may be reduced to an acceptable level by ensuring that the peak fields within the depletion region remain below approximately 4.3×10^5 Vcm^{-1}. The dark generation due to band-to-band tunneling via trap states may make it necessary to restrict the peak fields to even lower values. Re-triggering following a breakdown pulse, due to charge trapping or impact ionization of these traps during the avalanche, is also analysed. Optical coupling due to light emission during the avalanche

discharges is discussed and two methods for the prevention of this coupling between the image elements in linear arrays are described.

MOS gates that break down either at the $Si-SiO_2$ interface, or in the bulk at a n-p junction created by a buried n-channel, have been fabricated and operated above breakdown. The surface breakdown devices were operated in a charge-injection mode while the bulk breakdown devices were operated in a charge transfer mode similar to that which would occur in a full PC-CCD imager. The surface breakdown devices exhibited excessive dark count rates that were attributed to the high electric fields at the $Si-SiO_2$ interface. The bulk breakdown detectors were found to be far superior. They had very sharply peaked pulse height distributions and considerably lower dark pulse rates. Operation up to 12 volts above breakdown with a corresponding avalanche initiation probability greater than 0.9 was possible with these devices.

Only a very weak temperature dependence of the dark pulse rate was observed, suggesting that a tunneling mechanism of dark carrier generation was limiting the performance of the bulk-breakdown devices. The magnitude of the dark count rate agreed with that expected for band-to-band tunneling through mid-gap states. These states, through a change in their occupancy during breakdown, were also believed to cause the re-triggering of avalanches that was observed when operating at high, above-breakdown voltages. These limitations on performance can be expected to be removed by employing improved processing techniques which would reduce the mid-gap trap levels by one or two orders of magnitude.

TABLE OF CONTENTS

TABLE OF CONTENTS cont'd

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES cont'd

LIST OF FIGURES cont'd

LIST OF FIGURES cont'd

LIST OF FIGURES cont'd

LIST OF FIGURES cont'd

LIST OF FIGURES cont'd

ACKNOWLEDGEMENT

I would like to acknowledge the following people for their help and encouragement during the research and preparation of this thesis:

Gordon Walker, for his assistance during the initial stage of this project.

My advisor, Dave Pulfrey, for his constructive criticism and support throughout the research.

Garry Tarr, for many helpful discussions relating to experimental and theoretical aspects.

The financial assistance provided by the University of British Columbia fellowships is also gratefully acknowledged.

Very special thanks go to my wife, Cathie, for typing this thesis and for her patience during the lengthy preparation.

1 INTRODUCTION

The research reported in this thesis involves the investigation of a new mode of operation for charge coupled device (CCD) imagers that enables the direct detection of individual photons as they arrive. Such photon counting imagery is of great interest to astronomers and those in the space sciences, especially for satellite-borne observatories.

One of the requirements of modern astrophysics is the detection of very low photon fluxes with optimum sensitivity, spatial resolution and spectral resolution. The detection of very faint radiation is fundamentally limited by the quantified nature of the radiation itself, so that the astronomer must accurately track the source and use long exposures to record enough photon events on each elemental image area, or "pixel", to achieve the desired photometric accuracy. In ground-based observations, atmospheric turbulence and sky background determine the faintest objects for which useful photometric and spectroscopic information can be obtained. Observing from a space platform avoids the degrading atmospheric effects, however, there is still a low level background due to integrated starlight, zodiacal light, and scattered light in the optics. In addition to these low level backgrounds, it is very often impossible for the astronomer to avoid localized bright regions within the field. In spectroscopy there may be strong emission lines next to the region of interest while in two-dimensional photometry there are often bright foreground stars. Therefore, in order to realize the full potential of observing from space and to obtain the utmost from ground-based observations, a detector is required that is capable of recording small contrast differences in ultra faint images while at the same time not being swamped by the signal from localized bright areas. The imager must have maximum sensitivity, the lowest possible sensor noise, a large

dynamic range, good saturation characteristics, a very low dark signal, adequate resolution and spectral coverage, and a highly stable linear response.

Various photoelectronic image sensors have been developed that attempt to fulfill the above requirements. These detectors can be divided into two general categories:

(1) detectors that integrate the image internally and give an analog output

(2) photon counting imagers that allow external digital integration.

Before discussing in detail the proposed photon counting detector, presently existing low light level imagers are reviewed briefly in chapter two. The detective quantum efficiency is derived for both analog CCD and photon counting sensors, and the present limitations of these two imaging techniques are examined.

In chapter three the basic operation of the proposed photon counting imager is described and it is shown to have the potential for nearly optimum low light level performance in the visible and near infrared. Several problem areas are identified that require experimental investigation before the fabrication of such an array is attempted, and the background theory necessary for a full understanding of these issues is presented. The experimental investigation is discussed in chapter four. The design and fabrication of discrete photon counting MOS detectors is described and the experimental results obtained with these test devices are discussed. A summary of the main body of the thesis together with concluding remarks is given in chapter five.

2 BRIEF REVIEW OF LOW LIGHT LEVEL IMAGE SENSORS

The field of low light level imaging has progressed rapidly in the past few years. Advances in silicon VLSI technology have made possible large, defect-free solid state image sensors, and the rapidly growing microcomputer industry has made available a variety of low cost microprocessors and large memories for the manipulation and storage of the vast amount of data one obtains from such large sensor arrays. It is unnecessary to review in detail all of the analog and photon counting imaging systems presently in operation, because, by discussing the general types of image sensors and intensifier-sensor combinations used, it is possible to present a fairly accurate picture of the state of the art for low light level imaging. It must, of course, be remembered that in any real imaging system the type of sensor housing or intensifier-sensor assembly used, as well as the readout, control, and data handling electronics, can have a marked influence on the overall performance.

Analog imaging systems are divided into two general categories:

(1) image sensors with no pre-detection gain

(2) sensors that integrate the signal after some form of image intensification.

The latter category is discussed along with photon counting systems since its implementation is very similar and its characteristics are determined primarily by the image intensifier.

2.1 ANALOG IMAGERS

Low light level imagers in this category include television pick-up tubes [1,2], self-scanned photo diode arrays [3,4], charge injection device (CID) arrays [5,6] and charge coupled device (CCD) arrays [7,8].

Television pick-up tubes have been in routine operation for the past three decades, however, it was not until the late 1960's that television techniques started to replace the relatively inexpensive and well established photographic methods for quantitative scientific imaging. Television pick-up tubes offered the advantages of higher possible signal to noise in the image, linear response, and the possibility for post detection image processing. These features made television techniques superior to photography for low light level imaging, and by 1973 many groups in astronomy were using television sensors for spectroscopy and two-dimensional photometry [9]. Very promising results were obtained with these television type imagers, however, the newly developed silicon monolithic line and X-Y sensor arrays were becoming available at this time and offered some significant improvements. Solid state imagers have more stable photometric porperties, a highly linear response, and a well defined absolutely stable geometry. The advantage of small physical size is negated by the requirement that silicon monolithic imagers be cooled below 150 K in order to reduce the dark leakage current to a negligible level. Such temperatures generally require evacuated detector housings and special chip carriers.

The first solid state imagers to be used were self-scanned photodiode arrays. These sensors utilize the depletion region capacitance of electrically isolated p-n junction diodes for the integration and storage of photogenerated charge [10]. The individual photodiodes are coupled to one or a few video lines via MOS switches operated by on-chip shift registers. Readout occurs when the diodes are rebiased sequentially through the video line,

resulting in a train of recharging current pulses. The number of electrons (or holes) collected per incident photon is referred to as the responsive quantum efficiency (RQE).

The wide aperture linear photodiode arrays of 1024 elements or more supplied by Reticon Corp. were particularly suitable for spectroscopy and are still in use today. These sensors have a high responsive quantum efficiency from 400 nm to 900 nm (Fig. 2.1), very stable photometric properties, a saturation level greater than 10^7 photogenerated carriers, and only slight image spreading above saturation. The most serious drawback to these arrays is the high level of readout noise and the large switching signal due to capacitive coupling between the video line and the shift register. Carefully designed, highly stable drive circuitry allows the switching transients to be largely removed, however, the Johnson-Nyquist noise of the reset switches and the large video line capacitance results in typical r.m.s. noise levels of 700 to 1000 equivalent photoelectrons per pixel. This high level of readout noise means that photodiode arrays are not well suited to very low light level imaging. However, the large saturation charge does allow one to obtain high signal to noise ratios for long exposures at moderate light levels, where one can accumulate large integrated signals. It also enables a dynamic range greater than 10^4 to 1 to be recorded in a single integration. In addition to high readout noise the two dimensional photodiode arrays suffer from considerable dead space losses since the many video lines and reset switches lie within the active sensor area.

In the mid 1970's imagers that utilized the newly discovered charge transfer principle became available. These sensors use arrays of MOS capacitors for the integration of photogenerated charge. The gates of each capacitor are biased so as to drive the bulk silicon underneath into deep depletion. The potential wells thus formed collect and store the photo-generated

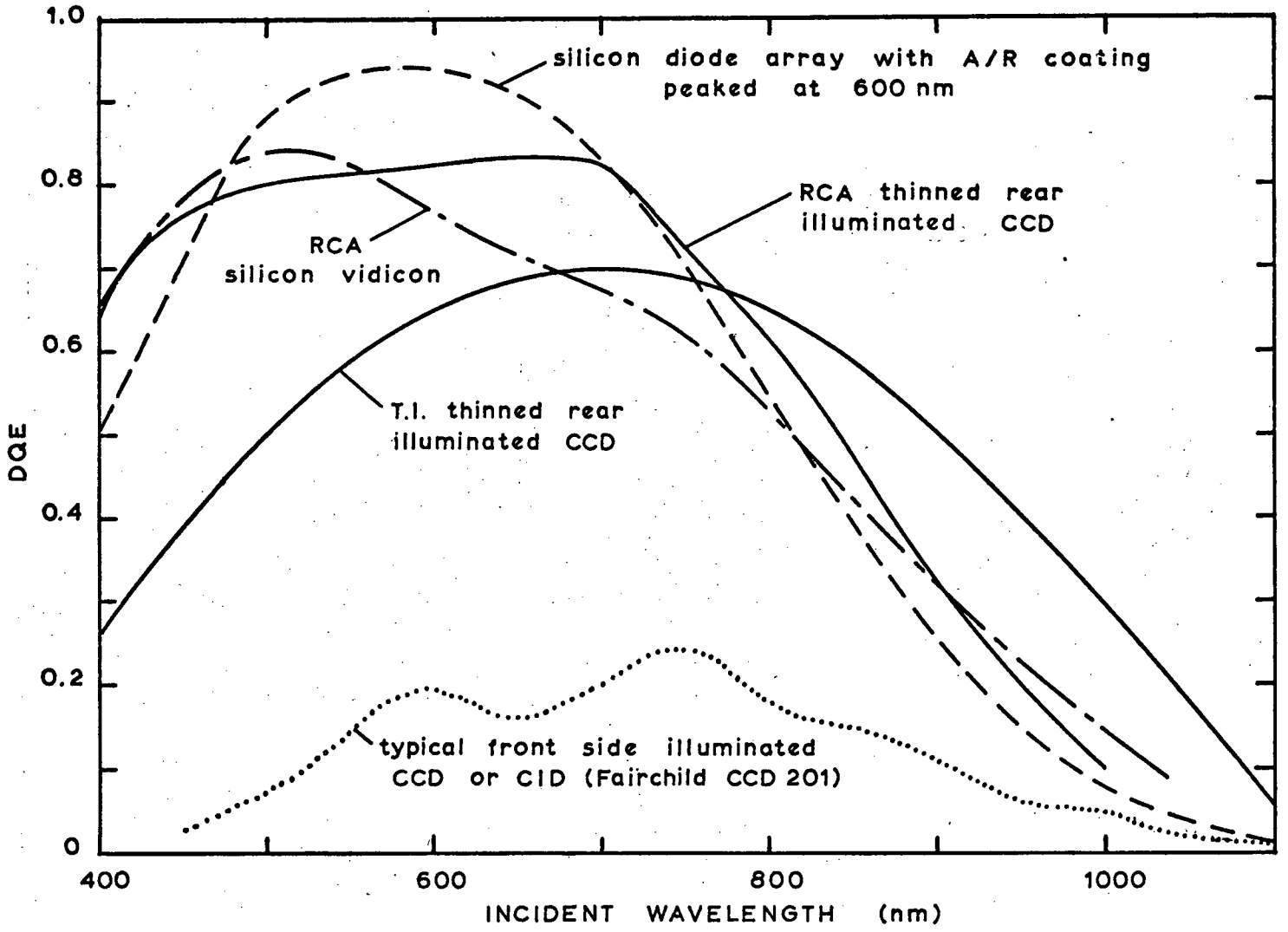


FIGURE 2.1 Responsive quantum efficiency of different analog sensor arrays as a function of wavelength.

minority carriers. Depending on the type of readout employed, charge transfer devices are of either the CCD or CID variety.

In CID imagers the individual sensing elements consist of a pair of gate electrodes that are connected in rows and columns. The signal from an individual pixel is obtained by pulsing the appropriate row of gates and sensing the voltage change on the appropriate column as the signal charge is transfered along the surface of the silicon from one gate to the other. The row and column selection is normally controlled by on-chip shift registers which results in a sequential readout. The above detection scheme leaves the signal charge unchanged so that multiple nondestructive readouts are possible. The CID imager is reset for a new integration by injecting the signal charge into the bulk of the semiconductor where it recombines with majority carriers.

The CID imager is not an inherently low noise device. Large switching spikes result from capacitve coupling between the row and column gates, and substantial signal attenuation by the column capacitance results in a typical rms readout noise of 800 to 1000 carriers per pixel. The noise may be greatly reduced by summing multiple nondestructive readouts. Unfortunately, the lowest noise level that can be achieved in this way is limited by variations of the large switching signal, which in turn requires very stable, low noise drive circuitry.

CID imagers have a linear response, however, there is a threshold effect which occurs after reset. The charge injection causes surface states to empty and subsequent signal charge is wasted refilling these states. To avoid this threshold probelm a small bias charge must be kept under the gates at all times. This bias signal must be read non-destructively and stored so that it can be subtracted from the subsequent data exposure. Another drawback to presently existing CID imagers is that they are front side illuminated which results in lower responsive quantum efficiencies, especially at short

wavelengths, due to absorption by the gate structure. In the case of semi-transparent electrodes such as polysilicon, loss results from absorption as well as from wavelength-dependent interferenece due to the multiple reflections at the various silicon-silicon dioxide interfaces. Consequently, the spectral transmission curve and hence the responsive quantum efficiency of the CID has a complex structure throughout the visible spectrum (Fig. 2.1).

The analog imager that presently has the most potential at low light levels is the CCD. In CCD's the charge transfer principle is used to transport the optically-created charge pattern over long distances across the silicon surface to an on-chip amplifier located at the edge of the sensor. This detection method results in a significant reduction in readout noise due to (1) signal detection at a common low capacitance node and (2) the elimination of switching spikes within the array. In addition, thinned, rear illuminated CCD sensors are available that have responsive quantum efficiencies approaching those of the self scanned photodiode arrays. The main difficulty with the CCD is the loss of charge due to interface-state trapping as the signal is transported along the interface. This problem may be partially overcome by introducing a bias charge or "fat zero", however, a better solution is to avoid surface states altogether by transporting the charge in a buried channel located away from the interface [11]. Because of the much lower density of bulk traps, such buried channel CCD's (BCCD's) are capable of very high charge transfer efficiencies.

2.1.1 The DQE For Analog CCD Sensors

A more quantitative appraisal of CCD imager performance at low light levels can be obtained by calculating the detective quantum efficiency, defined as:

$$DQE = \frac{(S/N)^2_{out}}{(S/N)^2_{in}} \tag{2.1}$$

where, $(S/N)_{out}$ = signal to noise ratio after detection

$(S/N)_{in}$ = signal to noise ratio before detection

A detector is performing optimally if its DQE is equal to 1. For all image sensors, the DQE depends on the temporal and spatial modulation transfer functions. Such a dependence results from the non-zero width in both the temporal and spatial response of the detector. At very low light levels one need only consider time stationary images and to simplify the analysis even further only the zero spatial frequency DQE will be calculated. If the DQE as a function of spatial frequency is desired it can be obtained from the spatial modulation transfer function (MTF) according to [12],

$$DQE(f) = \frac{DQE(0)\ MTF^2}{P(f)}$$

where, f is the spatial frequency

P(f) is the power spectrum of the noise.

For optical signals the lowest possible noise before detection is merely the statistical photon noise, i.e. no sky or scattered light background. In this case,

$$(S/N)_{in} = (R_p t)^{\frac{1}{2}} \tag{2.2}$$

where, R_p = mean rate of arrival of photons per pixel

t = integration time

For a CCD imager the signal to noise in the output after subtracting the dark background is,

$$(S/N)_{out} = \frac{RQE\ R_p t}{(RQE\ R_p t + \sigma^2 + 2R_d t)^{\frac{1}{2}}} \tag{2.3}$$

where, RQE = responsive quantum efficiency, equal to the ratio of the number of collected carriers over the number of incident photons. Included in this figure are any dead space losses.

σ = rms readout noise of the CCD expressed as an equivalent number of photoelectron charges per pixel.

R_d = dark leakage per pixel.

Substituting expressions (2.2) and (2.3) into (2.1) yields the zero spatial frequency DQE for a CCD,

$$DQE(0) = \frac{RQE^2}{RQE + (\sigma^2/R_p t) + 2(R_d/R_p)} \qquad (2.4)$$

Expression (2.4) for the DQE indicates that in addition to readout noise, excessive dark leakage also degrades the performance of CCD imagers. Fortunately, by cooling silicon monolithic arrays to 150 K or lower the dark generation of carriers may be reduced to a negligible level. For fixed integration time, and negligible dark generation, the DQE approaches its maximum value (equal to the detector's RQE) as the sensor irradiance increases. This is as expected since at high integrated signal levels the statistical photon noise dominates.

2.1.2 Performance of CCD Imagers Currently In Operation

Readout noise levels of 20 to 100 electrons rms have been achieved by several commercial CCD devices, however, most of these arrays are front side illuminated and the resultant responsive quantum efficiencies are low. Texas Instruments has produced large rear illuminated BCCD arrays specifically designed for low light level imaging. 400 x 400 pixel prototype arrays [13] have been in operation for some time and recently the full 800 x 800 element version has been completed. The 400 x 400 arrays are achieving readout noise levels of approximately 25 electrons rms [14] and have peak responsive quan-

tum efficiencies of 0.7 at 700 nm (Fig. 2.1). The sensor is thinned to a thickness of approximately 10μm over the active area leaving a thick rim for the support of the fragile membrane and for chip bonding. The processing-induced stresses at the front side silicon-silicon dioxide interface can cause the membrane to buckle significantly, creating some stability problems. A thinned rear illuminated 512 x 320 pixel BCCD that is bonded to a glass substrate-window combination has recently been produced by RCA. The bonding process largely eliminates membrane buckling and stability problems and reportedly enhances the responsive quantum efficiency. The RQE of the RCA CCD's is the highest yet achieved on a routine basis. Unfortunately, the quoted readout noise levels are 70 to 100 electrons rms.[15].

The DQE for the T.I. 400 x 400 BCCD is plotted in Fig. 2.2 versus total integrated incident signal at various wavelengths and assuming negligible dark leakage. Curves of constant output signal to noise are also shown. From these it is apparent that for integrated signal levels, such that $(S/N)_{out}$ is greater than approximately 70, the T.I. BCCD is photon noise limited with a DQE within 10% of the responsive quantum efficiency. The saturation level for BCCD's such as the T.I. 400 x 400 is typically 5×10^5 electrons. With a rms noise of 25 electrons the dynamic range becomes 2×10^4 to 1, and the signal to noise for a level near saturation would be 700 against signal quantum noise.

The CCD sensor appears to be a nearly ideal imager over a fairly wide spectral region around 700 nm. In practice, however, the internal charge integration and analog output can present some problems. In particular, nonlinearities introduced by the signal electronics before the analog data is digitized must be eliminated or calibrated out. Stability of the signal electronics is also of concern, especially if it is necessary to subtract one frame from another in order to remove a large background signal. The

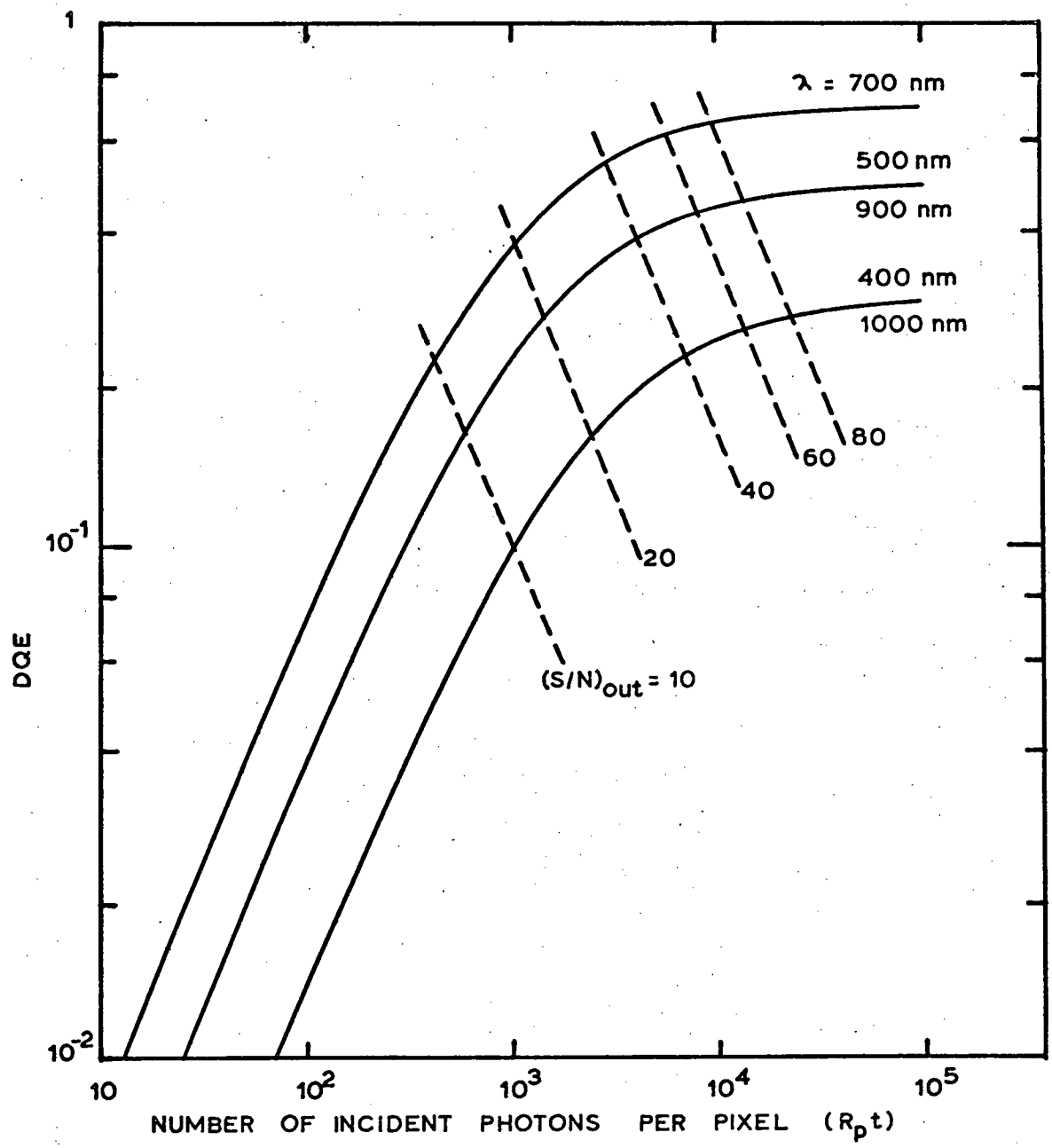


FIGURE 2.2 DQE for the T.I. 400 x 400 BCCD as a function of the total integrated incident signal, at several different wavelengths. Curves of constant signal to noise in the output are also indicated.

response of CCD imagers as well as the transfer characteristics of the on-chip pre-amplifier have been found to be highly linear, however, some CCD sensors show a threshold effect. For example, some of the recent RCA BCCD's show a zero exposure intercept that deviates by as much as 170 electrons from the level given by the readout of multiple empty frames [15]. This threshold may result from long-lifetime bulk traps and has serious consequences when trying to detect very faint images where little or no background is present.

2.2 PHOTON COUNTING IMAGERS

To avoid the problems of non-linearities and possible threshold effects associated with analog imagers and to obtain photon noise limited operation independent of the total integrated signal, astronomers pioneered the development of intensified CCD's that operate in a photon counting mode. The requirements for photon counting are straightforward. Enough pre-readout gain must be provided in order to reliably discriminate single photon events above readout noise in the detector and achieve a low background or dark count rate. Further, the imaging device selected must be capable of being scanned rapidly enough so that at the maximum sensor irradiance to be measured, rarely will more than one photon be incident on any one pixel during a frame time. Lastly, the intensified detector must be interfaced with a computer type memory to accumulate photoevents according to their location in the image.

Photon counting is obviously restricted to very faint images. In order to accommodate a large photon flux from an extended source the device must be scanned rapidly and there must be many small picture elements. However, this in turn dictates a large video signal bandwidth, which results in more readout noise per sample and necessitates higher intensifier gain so that the signal

pulses will be clearly above the noise. It is possible, however, to operate some intensified detectors in a charge integration mode so that higher photon fluxes can be measured. The integration time is set so as not to saturate the sensor and multiple frames may be summed without degradation due to sensor readout noise, provided there is sufficient gain in the image intensifier. The intensifier gain required is somewhat less than for photon counting, but the pulse height distribution should be narrow since the dispersion in pulse heights introduces noise which increases roughly with the square root of the number of photon events recorded. The photon shot noise also increases with the square root of the number of photon events so that, provided the relative dispersion in pulse heights is small, photon noise limited operation is possible. The noise contribution from cosmic ray and ion events, both of which produce very large pulses, can be a problem at very low light levels. Photon counting systems are able to reject these large events.

Many of the early intensified imagers did not have enough gain to discriminate single photon events above readout noise and were operated in the above intensified charge integration mode. Instability and non-linearities introduced by the image intensifier, as well as the noise due to cosmic ray and ion events, however, are limitations when operating in this mode and it is preferable, if possible, to photon count.

2.2.1. Linearity and DQE of Photon Counting Imagers

At low photon flux levels the DQE of an image photon counting system is constant and approximately equal to the responsive quantum efficiency of the intensifier photo-cathode. However, only one photon event may be detected per pixel per frame and as the photon flux increases, temporal sampling effects introduce non-linearities and lower the DQE. These effects are examined below for spatially and temporally invariant signals.

For an intensified imager operating in a photon counting mode, the mean event rate is,

$$n = R_p \, RQE \, \eta + n_d \, , \; n_d\tau << 1 \tag{2.5}$$

where, R_p = mean rate of arrival of incident photons per pixel

RQE = responsive quantum efficiency of the photocathode

η = probability that a photoelectron results in an output signal pulse above the discriminator level

n_d = mean number of events/sec/pixel due to noise and thermionic emission from the photocathode

τ = frame time

The arrival of incident photons and the occurence of dark events obey Poisson statistics, therefore, the probability that there will be a count in frame time τ is,

$$P = 1 - \exp(-n\tau)$$

The number of counts measured in N frames is, therefore,

$$C = PN = [1 - \exp(-n\tau)]N \tag{2.6}$$

and the mean square deviation is,

$$\overline{(\Delta C)^2} = \exp(-n\tau)[1 - \exp(-n\tau)]N \tag{2.7}$$

Equation (2.6) shows the departure from linearity due to the finite frame rate. After correcting for this nonlinearity the mean number of counts in N frames becomes,

$$C' = -N \, \ell n[1 - (C/N)] \tag{2.8}$$

and the mean square deviation becomes,

$$\overline{(\Delta C')^2} \simeq \left(\frac{dC'}{dC}\right)^2 \overline{(\Delta C)^2} = [\exp(n\tau) - 1]N \tag{2.9}$$

Therefore, after subtracting the dark count rate, the signal to noise in the output is,

$$(S/N)_{out} = \frac{(n - n_d)\tau N}{\{[\exp(n\tau) - 1]N + n_d\tau N\}^{\frac{1}{2}}} \tag{2.10}$$

The signal to noise in the input is,

$$(S/N)_{in} = (R_p\tau N)^{\frac{1}{2}} = \left(\frac{(n - n_d)\tau N}{RQE\ \eta}\right)^{\frac{1}{2}} \tag{2.11}$$

Therefore, the zero spatial frequency DQE becomes,

$$DQE(0) = RQE\ \eta\ \frac{(n - n_d)\tau}{\exp(n\tau) - 1 + n_d\tau} \tag{2.12}$$

Equation 2.12 shows the importance of a high RQE and η, and reveals the degradation due to the finite frame time τ. Figure 2.3 shows DQE/RQE η as a function of the event flux per frame $\eta\tau$(for n_d=0). In order for the DQE not to be degraded by more than 10% the frame rate must be more than five times the maximum count rate to be measured.

For event rates much lower than the frame rate ($n\tau \ll 1$), the DQE can be approximated by,

$$DQE(0) \simeq RQE\ \eta\ \frac{1 - (n_d/n)}{1 + (n_d/n)} \tag{2.13}$$

Equation 2.13 reveals the degrading effect of the dark event flux n_d, and shows the importance of choosing the optimum discriminator level for the photon flux to be measured. A higher discriminator level reduces the number of dark counts due to noise, but may also lower the probability η.

2.2.2 Photon Counting Imagers Presently in Use

Presently existing photon counting imagers are of three basic types: (1) A high gain image intensifier tube optically coupled to a television camera tube or a silicon monolithic array [16,17].

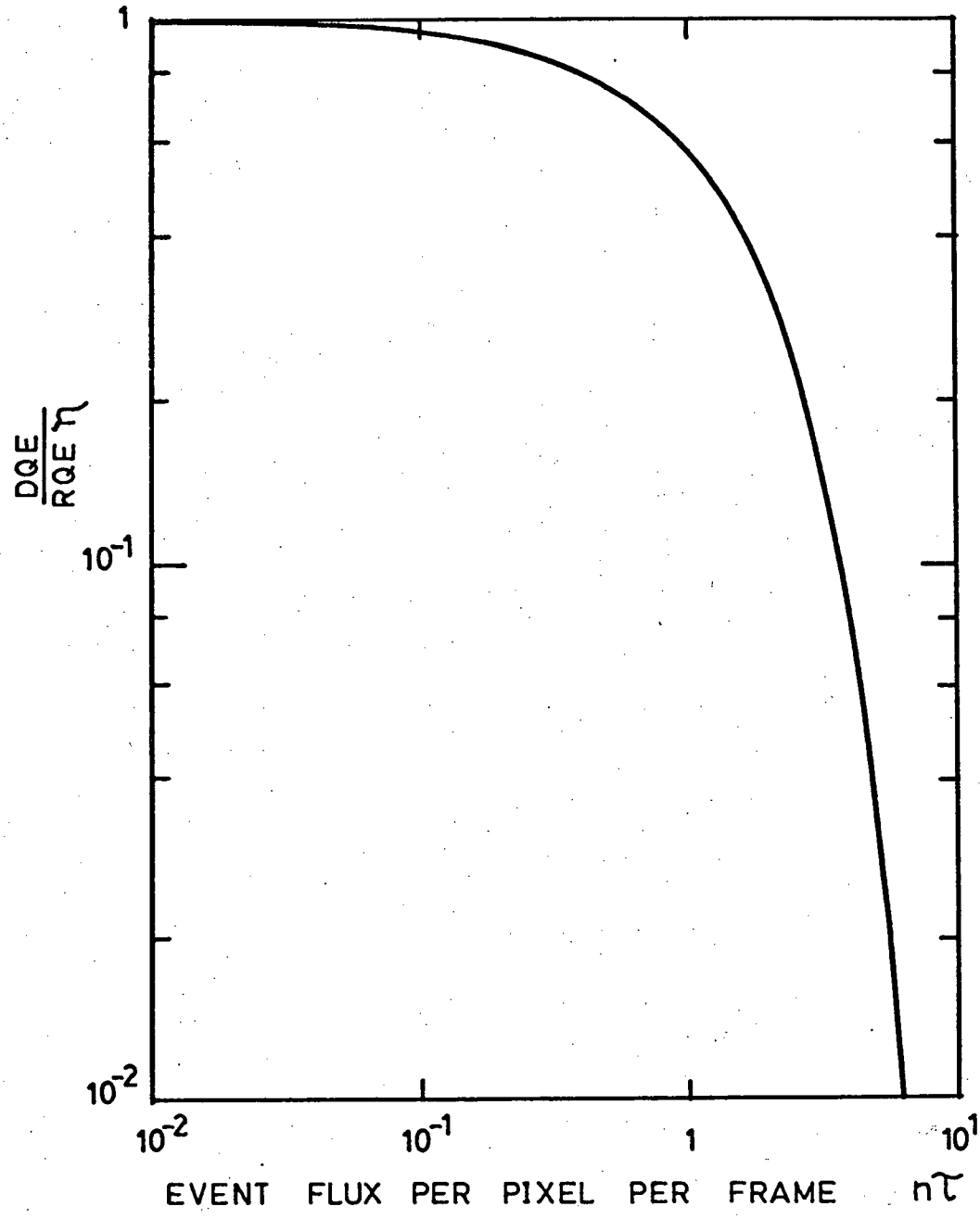


FIGURE 2.3 The effect of temporal sampling on the DQE of a photon counting detector.

(2) A special image tube that has as its anode either the silicon target of a vidicon or a solid state semiconductor array that directly detects photoelectron images from the photocathode [18,19,20,21].

(3) A microchannel plate (MCP) intensifier with a self-scanned multi-anode array [22], an x-y coincidence anode array or a resistive position sensitive anode [23,24,25].

In photon counting systems of the first category the image intensifier tubes can be either fibre-optically or lens coupled to the sensor. If they are lens coupled, multi-stage tubes or tubes employing a microchannel plate intensifier must be used to overcome the large coupling losses. The temporal and spatial spreading of photon events caused by these image intensifiers limits the maximum photon flux that can be measured. The temporal spread results from output phosphor lag and can cause a single event to appear in several successive frames. By subtracting the previous frame each time, all but the new events may be rejected, however, this also introduces a reduced sensitivity for subsequent detection of events in the same pixel. The spatial spreading can cause a single photon event to be recorded on several adjacent pixels in the detector. Not only does this lower the MTF but, because of variations in size, the photon events are recorded with different statistical weights. This is a source of noise and lowers the DQE. The MTF and DQE can be restored by processing each frame prior to storage so as to detect the event centers. However, more than one pixel per photon event is inhibited during a frame time so that the linearity and DQE are more quickly affected by increasing photon flux than was previously indicated (equations 2.6 and 2.12). The speed with which a frame can be processed for event center detection typically limits frame rates to less than 100/sec for 500 x 500 pixels or more.

In photon counting systems of the second category, the troublesome out-

put phosphor and optical coupling are eliminated by incorporating within the image tube a semiconductor array anode that directly detects the high energy photoelectrons. There is no event lag with these imagers and event center detection is generally not required. These image tubes are normally operated with accelerating potentials between 15 and 30 kV and typically give an electron-hole pair yield that is less than 4000 per incident photoelectron. In order to photon count, therefore, a low noise CCD sensor is required. Unfortunately, the lifetimes of CCD imagers are very short when operated in an electron-bombarded mode due to the damage introduced at the silicon-silicon dioxide interface. The transfer channels of a CCD are particularly sensitive to damage so that interline transfer CCD's with shielded transfer channels must be used, however, the 50% dead space caused by the interleaved channels reduces the photoelectron detection probability η. Rear-illuminated, electron-bombarded CCD's appear to have much longer lifetimes [21,26], however, very few CCD's have so far been operated in this mode and the data available on overall performance and lifetime is limited.

The microchannel plate intensifier is able to deliver an output pulse of 10^5 to 10^7 electrons, far in excess of the noise level of common electrical amplifiers and discriminators. This allows a multi-anode array to be placed in proximity focus at the output of the MCP, with each anode connected to an individual counting circuit. The individual anode electrodes are typically 0.5mm to 1.0mm square and are separated by interpixel screening electrodes to prevent cross-talk. Since all the pixels are effectively read out in parallel very high count rates per pixel can be accommodated. The MCP itself is limited to approximately 10^5 events $mm^{-2}sec^{-1}$ after which the gain falls off rapidly because of the charge lost by the high resistivity channel walls. An opaque photocathode deposited directly on the front face of the MCP can be used with these imagers. A field is established in front of the

MCP so as to collect the photoelectrons originating from the areas between channel openings.

The most serious limitation to discrete anode MCP's is the small number of pixels (less than 500) achievable with currently existing ceramic and electronic technologies. To overcome this limitation, coincidence techniques may be used to determine the spatial location of an event by the simultaneous arrival of charge upon two orthogonal sets of electrodes. In this way 2N circuits can handle an image having N^2 pixels. Large coincidence anode arrays may be fabricated, however, the pulse pair resolution time limits the maximum system count rate. If two photon events occur within the coincidence-resolving time, four counts are generated, two at the actual event locations, and two at false mirror image locations. The only way to avoid false counts is to remain well below the maximum count rate, which limits the use of these sensors to extremely faint, low contrast images that do not have any localized bright regions.

The photon counting system that has the greatest potential in terms of the maximum count rate per pixel (for large arrays) is a microchannel plate intensifier with a proximity focused, self-scanned anode array. The individual anodes store the charge from the MCP for a frame time and are interrogated using the same techniques as were developed for the self-scanned photodiode arrays. The high gain of the MCP makes very high frame rates possible. Unfortunately, this approach has not been pursued as actively as the coincidence anode technique and only small self-scanned anode arrays have so far been fabricated and tested. Spreading of the very large charge pulses leaving the MCP may make event center detection necessary for large arrays and thus limit the frame rate.

From the above discussion it is apparent that all of the presently existing photon counting imagers suitable for high resolution imaging have

rather low maximum frame rates. This does not affect their DQE or linearity when detecting very faint images, however, it does limit the dynamic range that can be recorded. A high dark count rate will lower the DQE and further limit the dynamic range but, by cooling the photocathodes and choosing the optimum discriminator setting (according to equation 2.13) this degradation may be reduced to a negligible level. All of the image photon counting systems utilize either a semi-transparent or opaque photoemissive surface as the initial light sensitive element and, as indicated by equation (2.12), it is the responsive quantum efficiency of this surface that ultimately limits the DQE at low light levels. Figure 2.4 shows the responsive quantum efficiency versus wavelength for the most widely used photocathode materials. Because of the low RQE's of available photocathodes longward of 600 nm, analog CCD imagers are able to offer superior performance in the red spectral regions, particularly for photon flux levels where a high signal to noise can be obtained. This is shown more clearly in Fig. 2.5 where the DQE versus wavelength is plotted for both the Texas Instruments rear-illuminated BCCD and a typical photon counting imager equipped with a tri-alk. photocathode. The DQE for an analog CCD also depends on the total integrated signal which, in turn, determines the output signal to noise. Curves of constant signal to noise are used in Fig. 2.5 to show the wavelength dependence of the DQE for the T.I. array. The curve for $(S/N)_{out}>1000$ represents the maximum DQE possible. The DQE for the pulse counting system is shown for photon event rates much lower than the frame rate and is therefore independent of the output signal to noise.

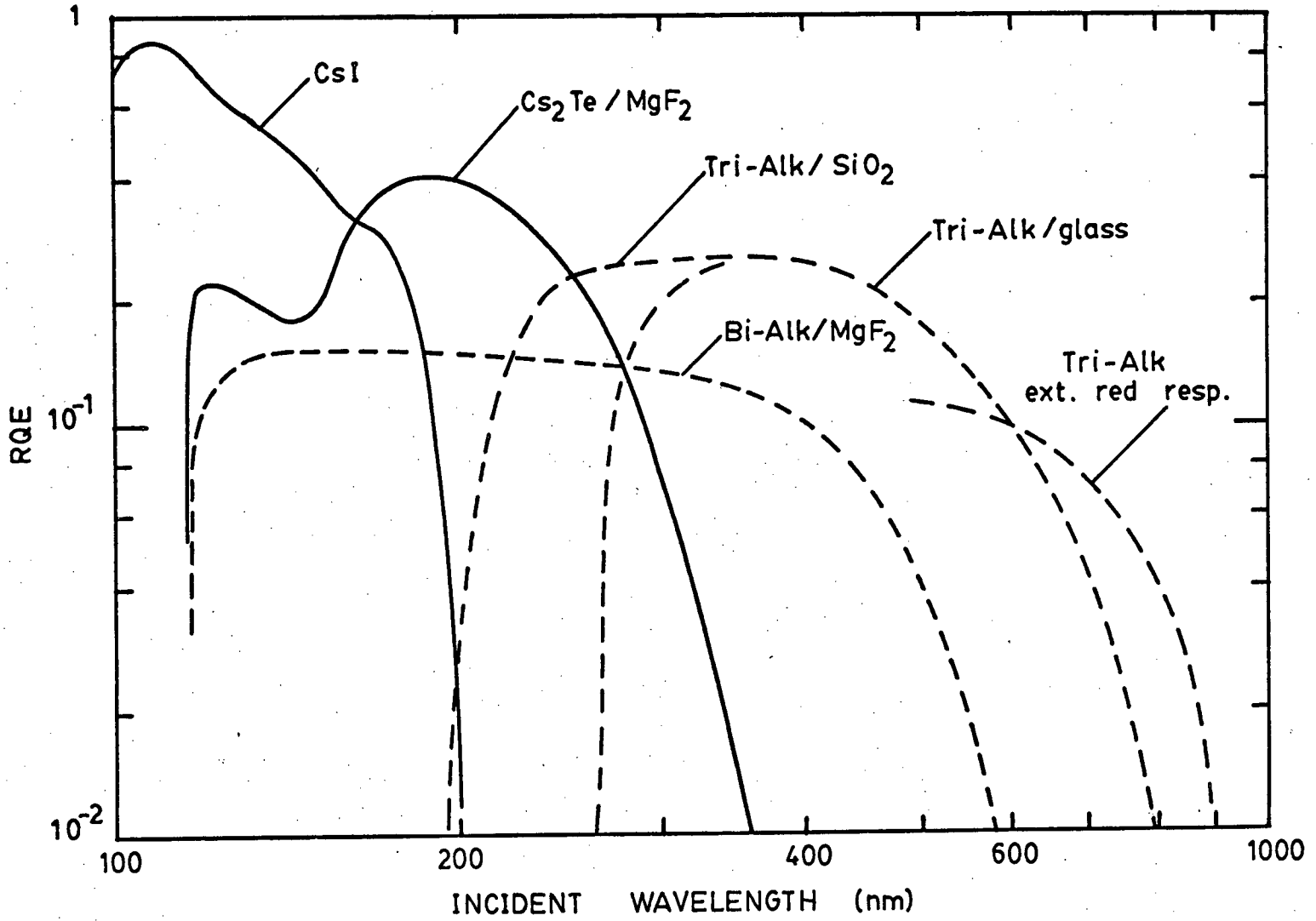


FIGURE 2.4 Responsive quantum efficiency of different photocathode/window combinations as a function of wavelength.

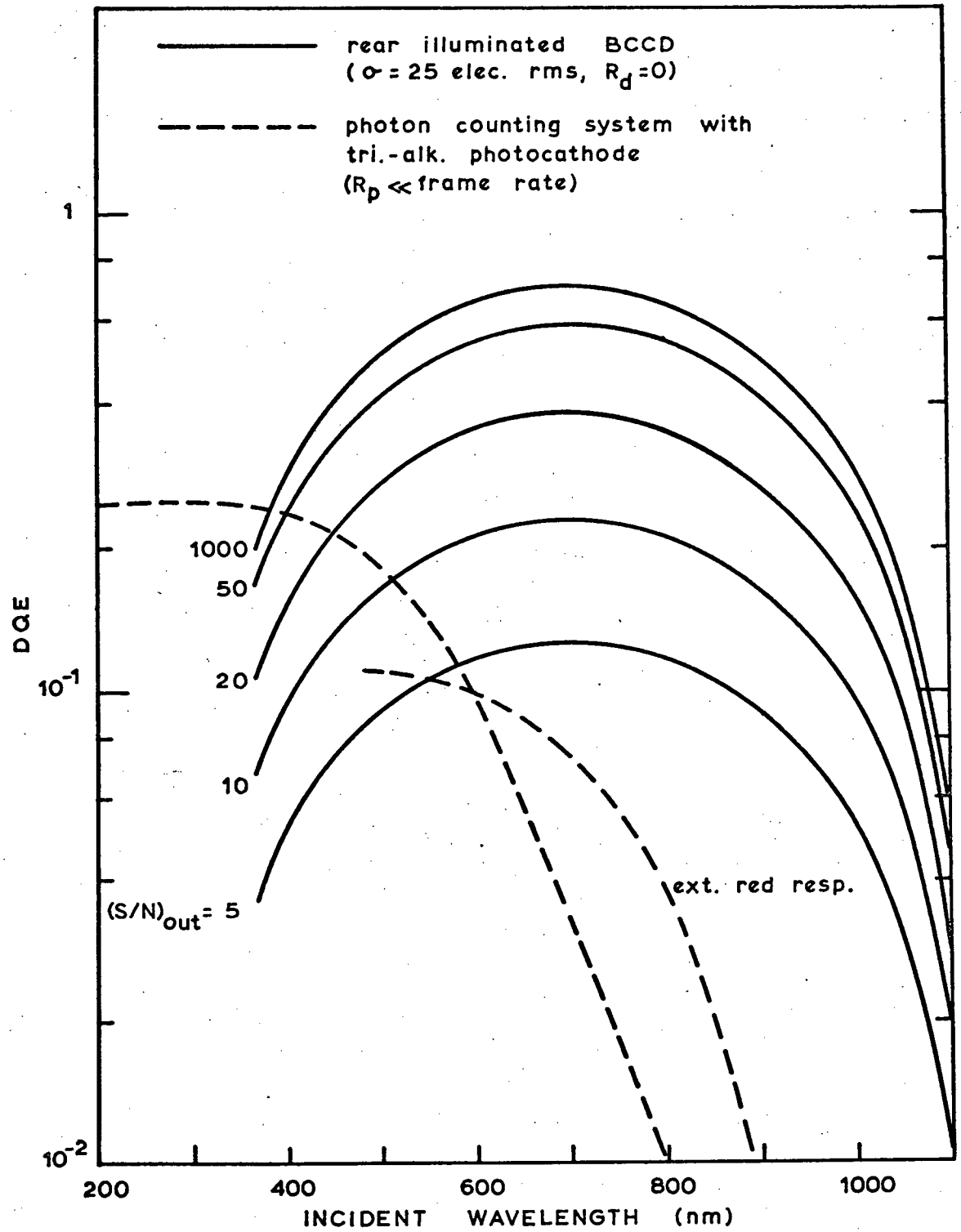


FIGURE 2.5 The DQE as a function of wavelength for a photon counting detector with a tri.-alk. photocathode, and for a rear-illuminated analog BCCD detector (T.I. 400 x 400 BCCD). The total integrated incident signal required to give the indicated values of output signal to noise for the BCCD can be obtained from Fig. 2.2.

3 THE PROPOSED PHOTON COUNTING SENSOR AND THEORY OF OPERATION

It has been demonstrated [27]-[31] that avalanche photodiodes can be used above the breakdown voltage in a photon counting (or Geiger tube) mode, provided the dark current is sufficiently small. When such an avalanche diode is suddenly biased above breakdown it is initially non-conducting. Not until a carrier is injected into or generated within the high field region of the depletion layer can an avalanche be initiated, so triggering the diode into reverse conduction. Furthermore, by cooling the avalanche diode to very low temperatures it should be possible to reduce the thermal generation rate of carriers to a negligible level so that virtually the only carriers available to initiate an avalanche will be those generated by incident photons. After a self-sustaining avalanche has been started the diode can be made ready to detect another photon, provided there is a suitable quenching circuit which momentarily reduces the diode bias and interrupts the avalanche.

The equivalent circuit for an avalanche diode operating above breakdown is shown in Fig. 3.1 [35]. V_b denotes the breakdown voltage, R_s the diode series impedance, R_c the space charge resistance during breakdown, and C_d the diode junction capacitance, including guard ring capacitance. The bias and detection circuit are represented by the applied voltage V_a, a stray shunt capacitance C_s and a load resistance R_L, which, if large enough, will also serve to quench the avalanche breakdowns. An avalanche discharge is represented by the closing of switch S. As soon as an avalanche is initiated, C_d discharges towards V_b with a time constant of approximately $R_c(C_d + C_s)$. With a large enough load resistor, however, the current becomes very small in the vicinity of V_b and statistical fluctuations in the number of carriers in the high field region of the diode depletion layer cause the avalanche to turn off. The mean time to turn off decreases as the circuit impedance in-

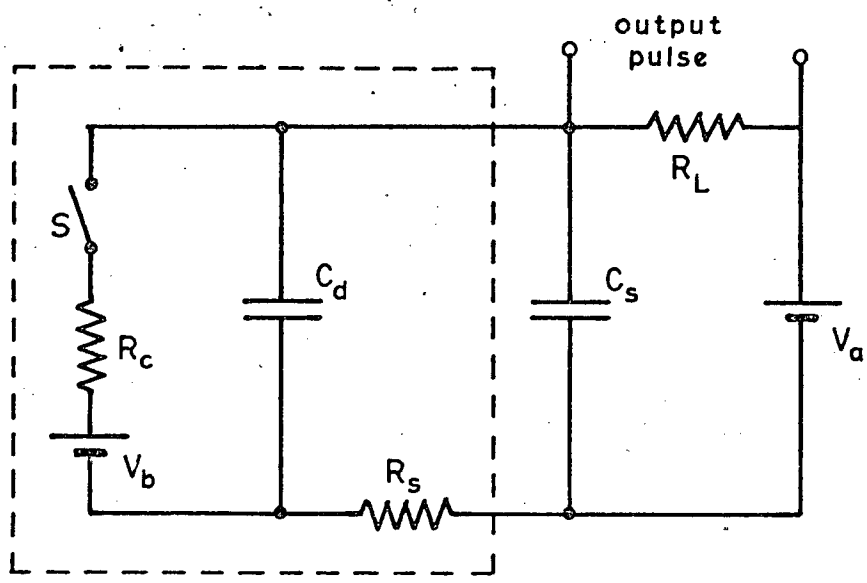


FIGURE 3.1 Equivalent circuit for an avalanche diode operating above breakdown (shown in dashed box) plus the bias and detection circuit. An avalanche discharge is represented by the closing of switch S.

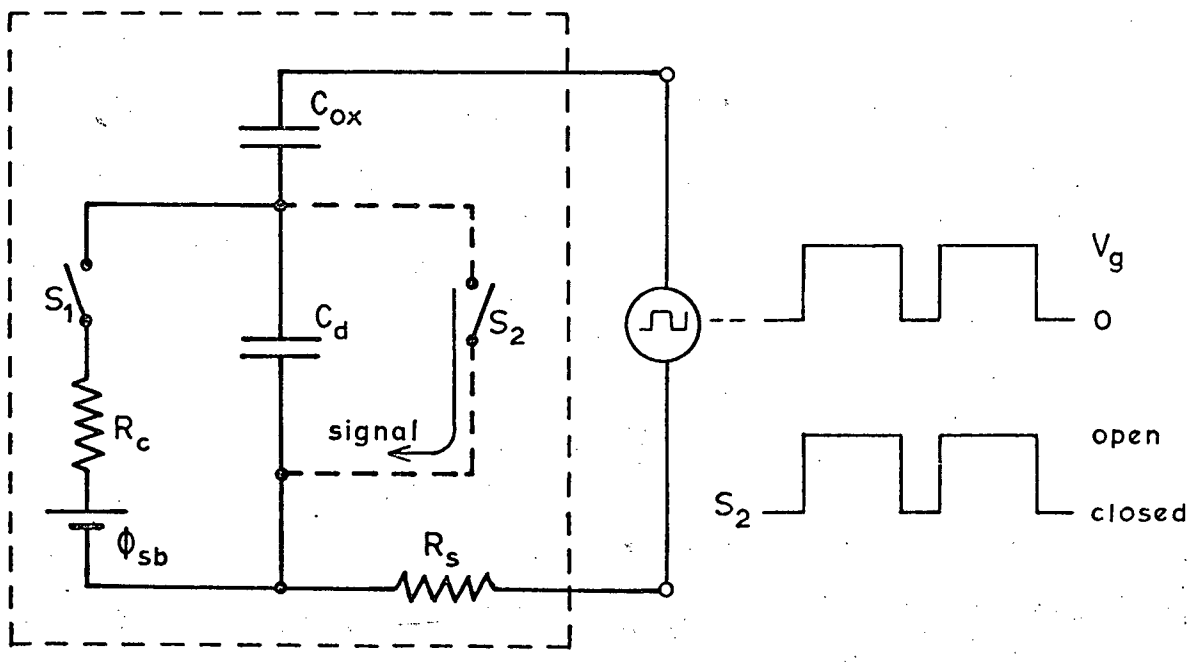


FIGURE 3.3 Equivalent circuit for an MOS gate operating above breakdown. Charge injection (or charge transfer) is represented by the dashed conduction path and switch S_2.

creases [35,36]. Typically a load resistor of 100 kΩ is used for small area (~100 μm^2) diodes. Once the avalanche has been quenched (S opens) C_d is recharged to V_a with a time constant of approximately $R_L(C_d + C_s)$. Since R_L is normally much larger than R_c, this time constant determines the dead time per pulse. Clearly there is a compromise between minimizing R_L and still maintaining a short quench time.

Not every photogenerated carrier that is generated within or diffuses to the depletion layer will trigger the diode into reverse conduction. A few will transit the high field region without suffering any ionizing collisions while others may initiate chains of ionizing collisions that die out after only a few carriers have been generated. In order to establish a self-sustaining avalanche, not only must the initiating carrier cause at least one ionization, but during each subsequent transit time some descendant must also cause at least one ionization. The probability that a carrier will have an infinite number of descendants (i.e., will trigger the diode into reverse conduction) has been referred to as the avalanche initiation probability [32]. The avalanche initiation probability is zero at the breakdown voltage and, as would be expected, eventually saturates to one as the diode is biased further above breakdown.

A silicon avalanche photodiode operated in the above mode offers the unique possibility of a solid state photon counting detector capable of a very high DQE. Furthermore, because of its small possible size, such an avalanche diode could be the basic element to an entirely solid state, high performance, photon counting imager. A small monolithic or hybrid array of avalanche photodiodes, each with its own pulse counting circuitry, would be of considerable interest. Ultimately, however, one would like to fabricate large self-scanned arrays capable of high frame rates. In this case the diodes must be operated in the charge integration mode (i.e., pulsed above breakdown and

then isolated for a frame time). Unfortunately, the requirement for high voltage reset switches and the need to prevent avalanche discharges during readout, severely complicates the design of such a self-scanned diode array.

One is not restricted, however, to using an array of junction diodes. The image element of a solid state photon counting sensor could also be a metal-insulator-semiconductor capacitor that is pulsed into very deep depletion, such that the field in the depleted semiconductor is higher than that which would normally cause breakdown and the subsequent formation of an inversion layer. One important advantage of such an MIS breakdown image element is its inherent quenching mechanism. Once an avalanche has been initiated and large numbers of carriers are being generated, an inversion layer will build very rapidly until the potential across the semiconductor depletion region has dropped to a value too small to sustain the avalanche (Fig. 3.2). The MIS capacitor will then remain in a partially discharged state until it is reset by removing the charge that is forming the inversion layer under the gate.

An equivalent circuit for such an MIS breakdown gate is shown in Fig. 3.3. V_g is the gate voltage above breakdown and ϕ_{sb} is the silicon surface potential at the onset of breakdown. R_s and R_c are the series and space charge impedance, C_d is the depletion region capacitance, and C_{ox} is the oxide capacitance. The closing of switch S_1 represents the initiation of an avalanche discharge while the closing of switch S_2 represents charge transfer and reset. MOS gates of this sort could be integrated into a charge coupled array so that the readout, and simultaneous reset, would proceed exactly as in a normal CCD imager. The difference, of course, is that instead of creating only one carrier pair, a single photon is now able to trigger a momentarily sustained avalanche, thereby inducing a sizeable charge packet under the detection gate.

Figure 3.4 illustrates this new operating regime for CCD imagers.

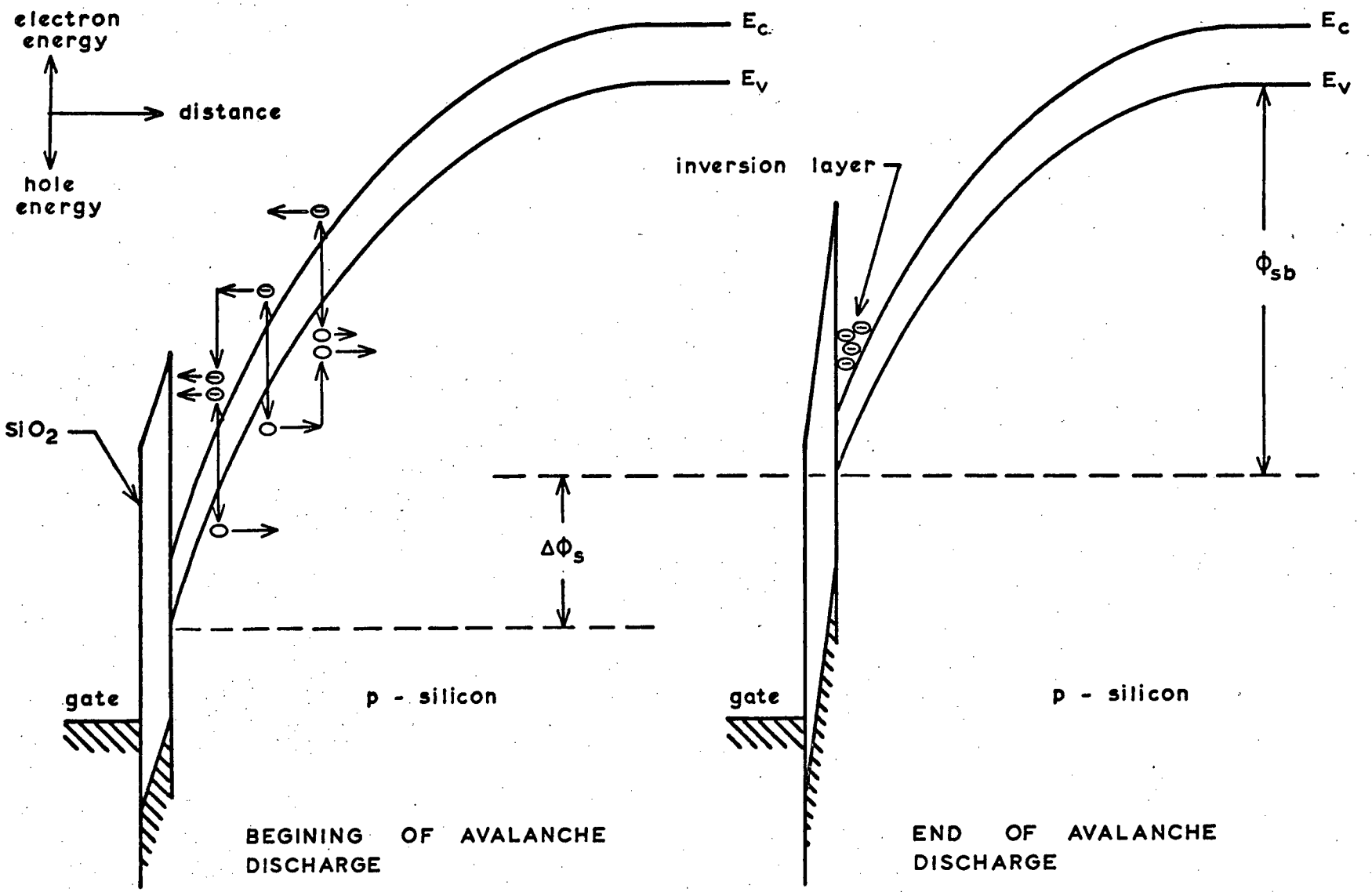


FIGURE 3.2 Energy band diagram for a p-substrate MOS gate at the beginning and end of an avalanche discharge. ϕ_{sb} = silicon surface potential at the onset of breakdown. $\Delta\phi$ = amount by which the surface potential exceeds breakdown initially.

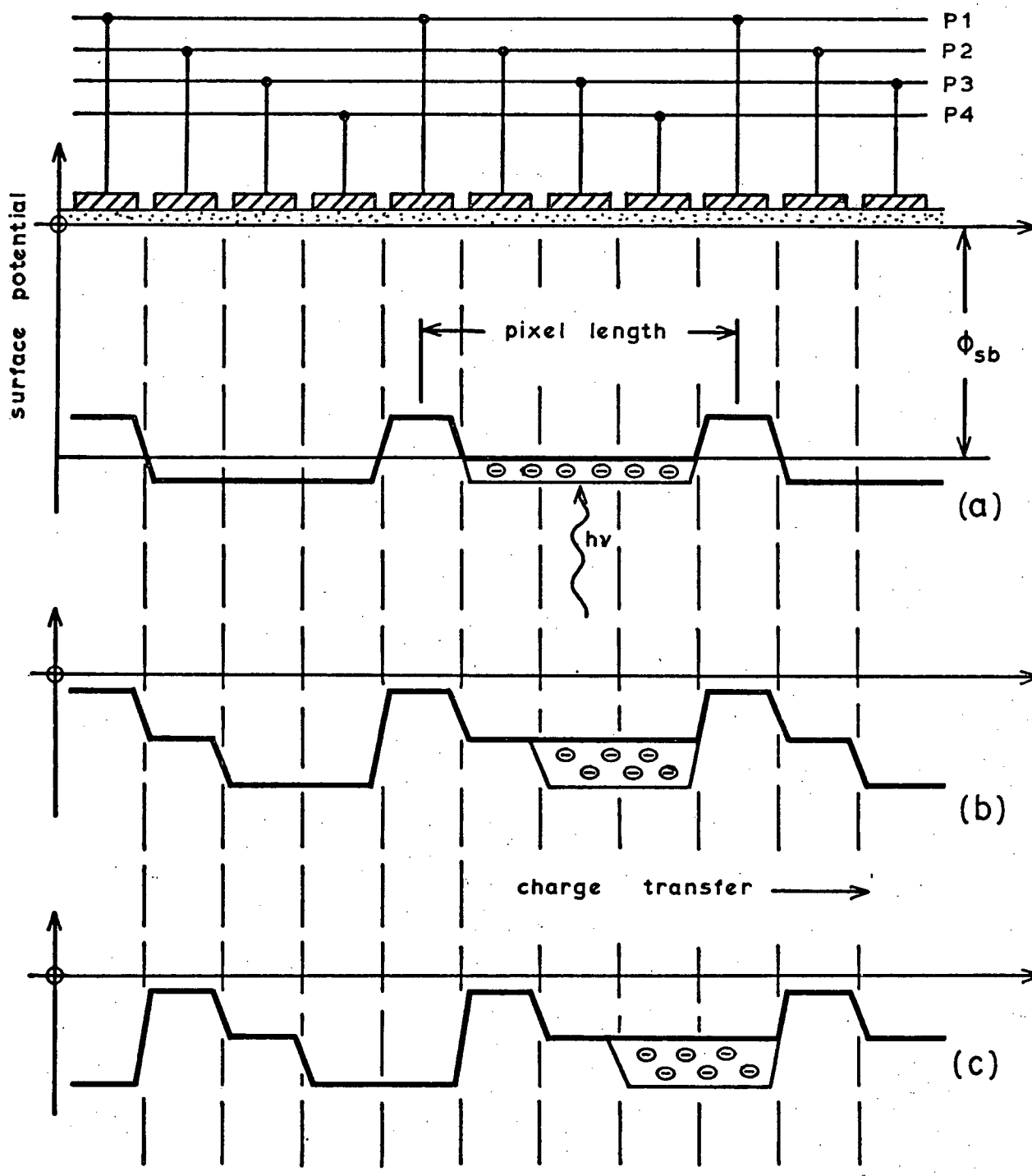


FIGURE 3.4 Potential well diagram illustrating the basic operation of a 4-phase PC-CCD. Heavy solid line indicates the potential at the $Si-SiO_2$ interface.

(a) detection part of cycle.
(b) end of detection phase, ready for charge transfer.
(c) charge transfer and reset.

During the readout phase the imager is operated exactly as a normal CCD. It is only during the detection part of the cycle in the interval between readouts that one or more phases is held at a potential above breakdown. Since no avalanche discharges can occur during readout, image smearing is not a problem. However, in order to minimize dead time this phase must last for only a small fraction of the total frame time. The array would, therefore, have to be organized for frame transfer into a separate storage and readout area. The large signal charge packets, and the fact that only their presence need be determined during readout, should make very high clocking rates possible during frame transfer and readout. In addition to dead time, a certain amount of dead space may also be unavoidable. In order to define the individual pixels and prevent crosstalk, the regions between transfer channels and under at least one clock phase must be kept below breakdown. Photogenerated carriers from these areas may not pass through a high field region before being collected in the potential wells under the active gates, and thus, may be unable to trigger an avalanche discharge.

The linearity and DQE for a photon counting CCD operating in the breakdown mode (hereafter referred to as a PC-CCD) are described by the same equations as were derived previously for photon counting imagers (equations 2.6 and 2.12), however, η now refers to the avalanche initiation probability, and the RQE describes the silicon response. It is reasonable to expect that the dead space in a rear-illuminated PC-CCD would be less than 50%, perhaps as low as 25%, and that the RQE for each pixel could be at least as high as for the RCA or Texas Instruments CCD's. Assuming also that the dark count rate is negligible and that the avalanche initiation probability is 0.9 or higher, the DQE versus wavelength for a PC-CCD would be approximately as shown in Fig. 3.5. It is clear that the high responsive quantum efficiency typical of silicon detectors would give the PC-CCD a distinct advantage over

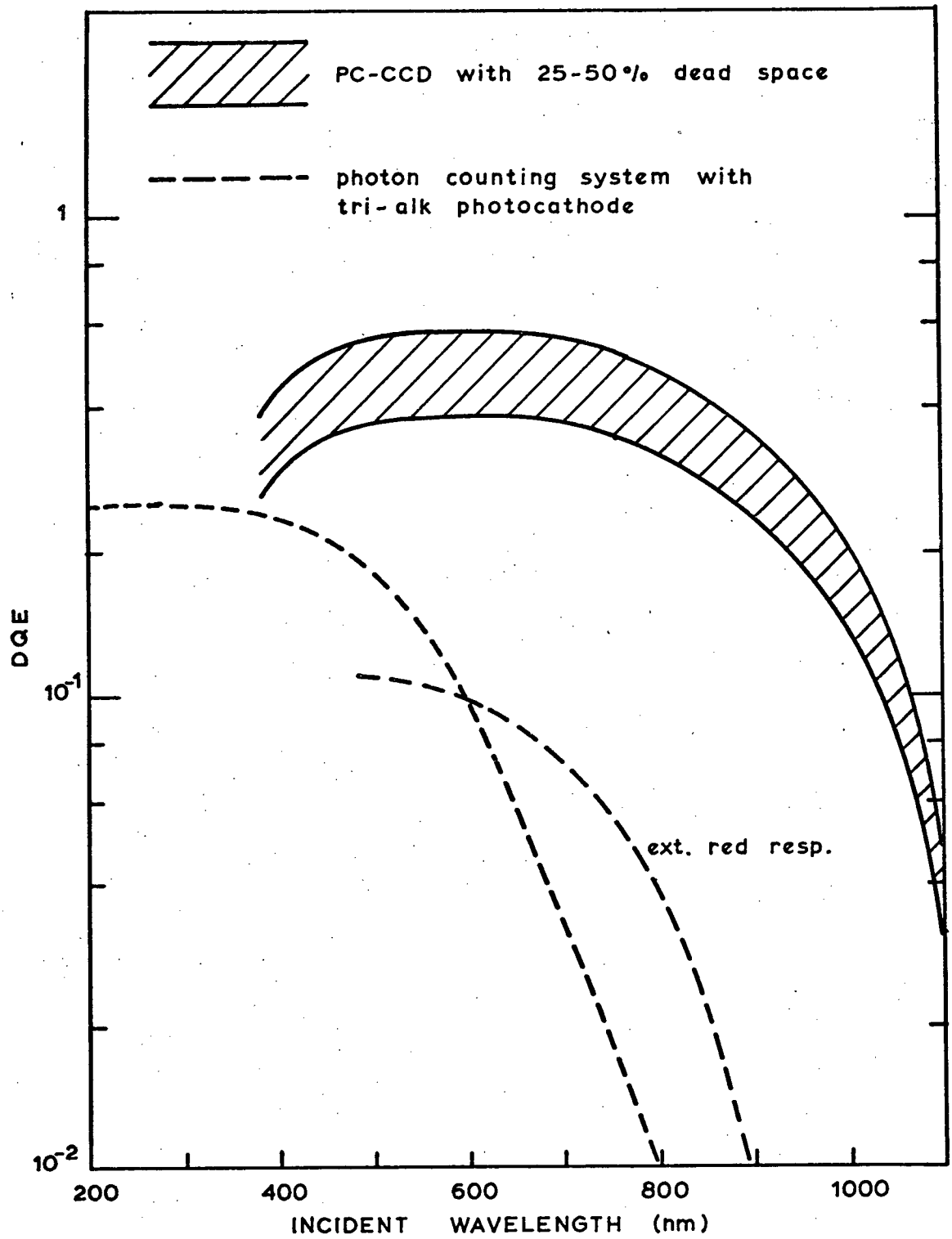


FIGURE 3.5 Expected performance of a PC-CCD compared to existing photon counting systems employing a tri.-alk. photocathode.

presently existing photon counting imagers in the visible and near infrared portion of the spectrum. The long wavelength response shown is based on room temperature RQE's and would be somewhat reduced at lower temperatures due to the decreased infrared absorption coefficients. This could be partly compensated for by optimizing the anti-reflection coating for these wavelengths and using a thicker substrate. By comparing Fig. 3.5 with Fig. 2.5 it can be seen that a PC-CCD could also compete very favorably with analog CCD's, especially at very low light levels where high output signal to noise ratios are not possible.

The new breakdown mode of operation for CCD's would appear relatively straightforward, however, before a successful PC-CCD imager can be developed the following important problems must be dealt with:

(1) Maximizing the avalanche initiation probability.

(2) Reduction of the dark count rate to a negligible level.

(3) Prevention of premature edge breakdown in the image elements, and achieving planar, microplasma-free discharges.

(4) Minimizing pixel cross-talk due to light emission during the avalanche discharges.

3.1 THE AVALANCHE INITIATION PROBABILITY

The avalanche initiation probability has been defined as the probability that a carrier, injected into or generated within the depletion layer of a diode biased above the breakdown voltage, will trigger a self-sustaining avalanche (i.e., one that would continue to grow indefinitely in the absence of any limiting mechanisms). Such a sustained avalanche is, of course, not possible with an MOS breakdown detector. Once an avalanche has been initiated in such a device, the buildup of space charge and the formation of the in-

version layer reduces the peak electric field and the corresponding values of the ionization coefficients, ultimately to the point at which the avalanche turns off. Before the electric field and ionization rates have been reduced significantly, however, the number of carriers in the high field region of the depletion layer will have grown to the extent that a statistical fluctuation to zero, prior to the generation of a detectable signal charge in the inversion layer, is highly improbable [32,33]. Thus, to a very good approximation, the probability that a carrier will trigger an avalanche that results in a detectable pulse, is the same as the avalanche initiation probability of an idealized device with no current limiting mechanisms, and for which the electric field and ionization coefficients retain their initial zero current values. This is an important simplification since with no current flowing, the electric field distribution in the depletion layer and the corresponding values of the ionization coefficients may be accurately calculated. The ionization rate data for silicon used in this investigation is given in Appendix A.

3.1.1 Triggering Probability Theory

Consider the depletion region of a planar avalanche diode (or MOS gate) that is biased above the breakdown voltage, (shown schematically in Fig. 3.6); $P_e(x)$ and $P_h(x)$ are the avalanche initiation probabilities, for electrons and holes respectively, that start at position x in the depletion layer. $P_p(x)$ is likewise the avalanche initiation probability for an electron-hole pair that has been generated at position x. The probability that an electron or hole will suffer an ionizing collision in the infinitesimal distance δx is simply given by $\alpha_e(x)\delta x$ and $\alpha_h(x)\delta x$ respectively, where α_e and α_h are the electron and hole ionization rates. The probability $(1 - P_p)$ that neither the electron or the hole of a pair generated at position x are able to trigger a breakdown

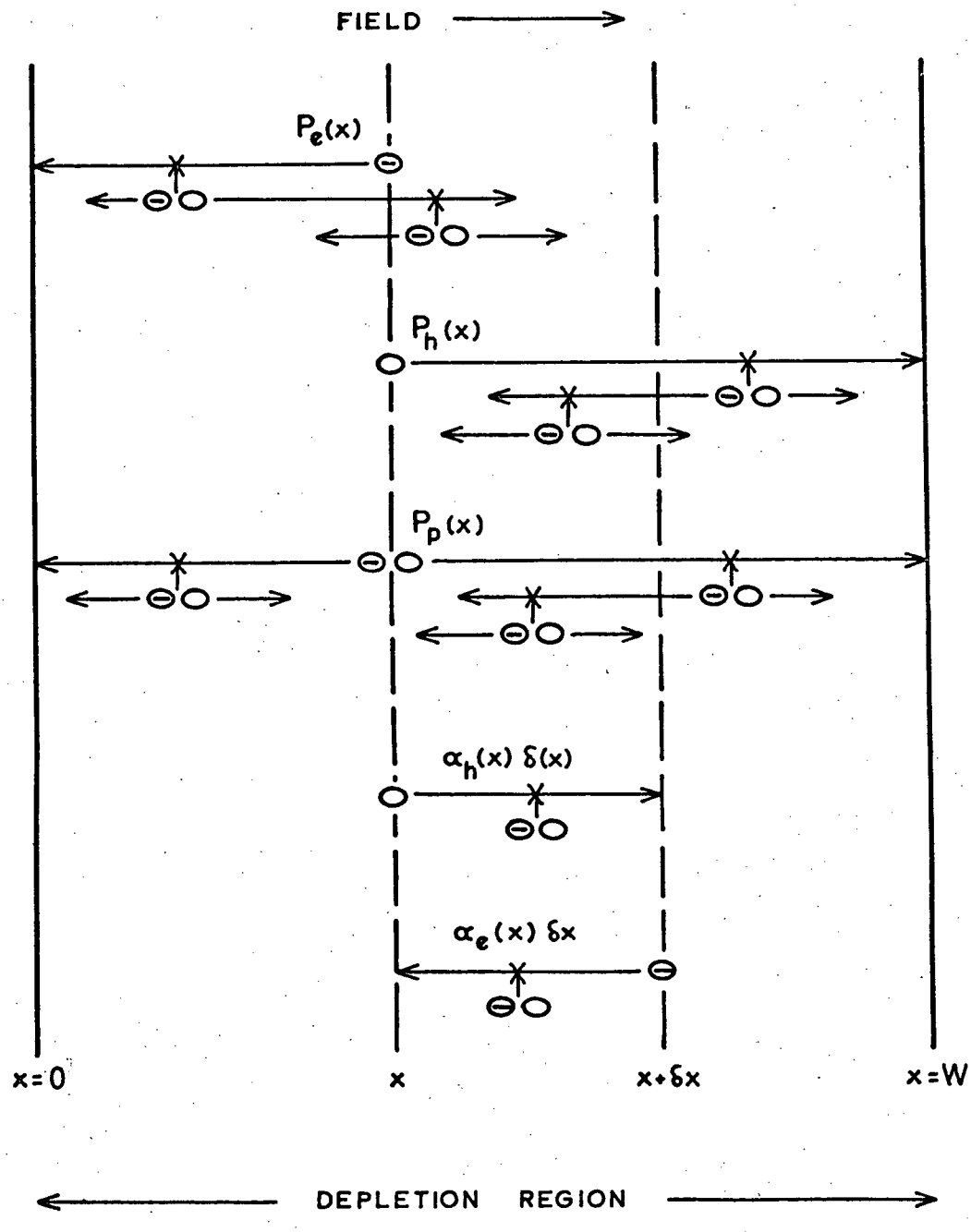


FIGURE 3.6 Model of the impact ionization that occurs subsequent to the introduction of a triggering carrier (or carrier pair) at position X in the depletion region. The probability of occurence is indicated in each case.

is given by the product,

$$(1 - P_p) = (1 - P_e)(1 - P_h)$$

therefore,

$$P_p = P_e + P_h - P_e P_h \tag{3.1}$$

Similarly, Oldham et.al. [31] show that the probability $P_e(x + \delta x)$ that an electron starting at position $x + \delta x$ will trigger a sustained avalanche can be expressed as,

$$P_e(x + \delta x) = P_e(x) + \alpha_e(x)\, \delta x\, P_p(x) - P_e(x)\, \alpha_e(x)\, \delta x\, P_p(x)$$

or in differential form, after substituting (3.1) for $P_p(x)$,

$$\frac{dP_e}{dx} = (1 - P_e)\alpha_e(P_e + P_h - P_e P_h) \tag{3.2}$$

A similar expression can be derived for holes,

$$\frac{dP_h}{dx} = -(1 - P_h)\alpha_h(P_e + P_h - P_e P_h) \tag{3.3}$$

In the case of an idealized diode with no current limiting mechanisms, the zero current values for the ionization coefficients $\alpha_e(x)$ and $\alpha_h(x)$ may be used. Therefore, provided the electric field distribution prior to breakdown is known, (3.2) and (3.3) can be integrated with the boundary conditions

$$P_e(0) = 0 \tag{3.4}$$

$$P_h(w) = 0 \tag{3.5}$$

Numerical techniques must, in general, be used to solve (3.2) - (3.5). The most straightforward method is to use an assumed initial value for $P_h(0)$ so that standard numerical techniques for the solution to systems of differential equations may be used. The value for $P_h(0)$ can then be modified in an iterative procedure until the extra boundary condition $P_h(w) = 0$ is satisfied. Examples of the solution for the avalanche initiation probabilities $P_e(x)$ and $P_h(x)$ for an MOS gate biased above breakdown are

shown in Fig. 3.7. The corresponding band diagram is that shown in Fig. 3.2. A uniformly doped substrate with $N_A = 5.5 \times 10^{15}$ cm^{-3} and room temperature ionization rates were used for this example. The parameter $\Delta\phi_s$ is the amount by which the silicon surface potential exceeds the breakdown voltage.

It is immediately obvious from Fig. 3.7 that, in silicon, electrons are far more effective than holes in triggering a sustained avalanche. This is simply due to the higher ionization coefficient for electrons (see Appendix A) and yields the well-known result that silicon avalanche detectors should be designed in such a way that it is predominantly the photogenerated electrons that initiate avalanche multiplication [34]. This means that photons must be absorbed in the p-region adjacent to the high field region and that the substrate of an MOS avalanche detector or PC-CCD would have to be p-type, as in the example shown in Fig. 3.7. Fortunately, this is standard practice with CCD imagers since electrons have a higher mobility in silicon and are the preferrred signal carrrier. The avalanche initiation probability for electrons entering the depletion layer from the neutral bulk (i.e., $P_e(w)$ in Fig. 3.7), is shown as a function of the surface potential in Fig. 3.8. The triggering probability for holes originating from the silicon/silicon dioxide interface is also shown for comparison.

It is apparent from the theoretical results shown in Fig. 3.8 that a large voltage above breakdown is required to reach saturation of the avalanche initiation probability. In the example shown, an excess bias of 10 V or more would be required in order to operate in the plateau region of the electron triggering probability. The hole triggering probability has not even begun to saturate and is in fact increasing slightly superlinearly at these excess voltages. It would be desirable to operate a PC-CCD at somewhat higher excess voltages to ensure that small variations in the breakdown voltage over the image area do not introduce large variations in response.

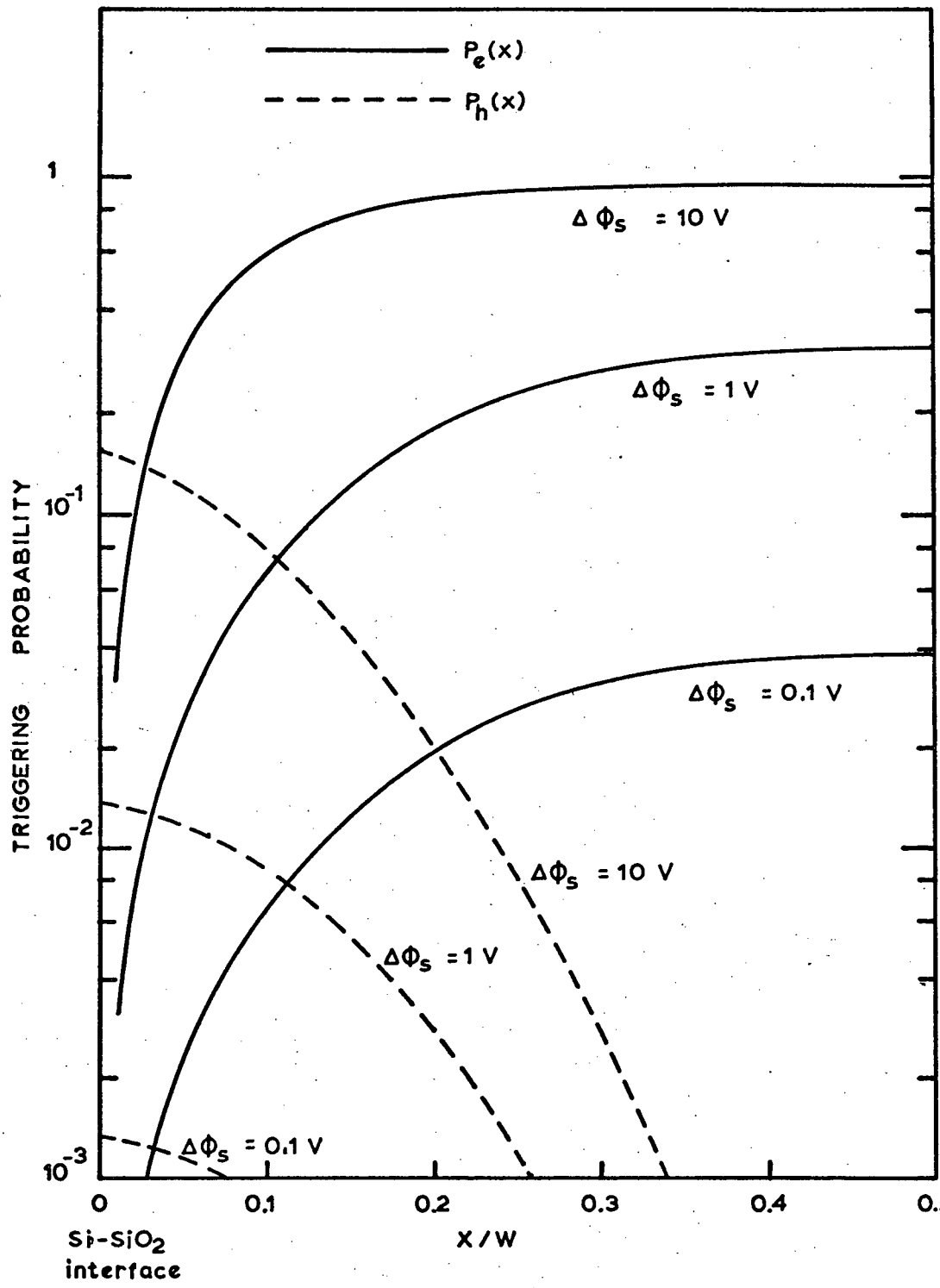


FIGURE 3.7 The avalanche initiation probabilities $P_e(x)$ and $P_h(x)$ for a p-substrate MOS gate. $\Delta\phi_s$ is the amount by which the surface potential exceeds breakdown. W is the depletion layer width. The ionization rate data used is given in Appendix A, $N_A = 5\times10^{15}cm^{-3}$, T = 300K.

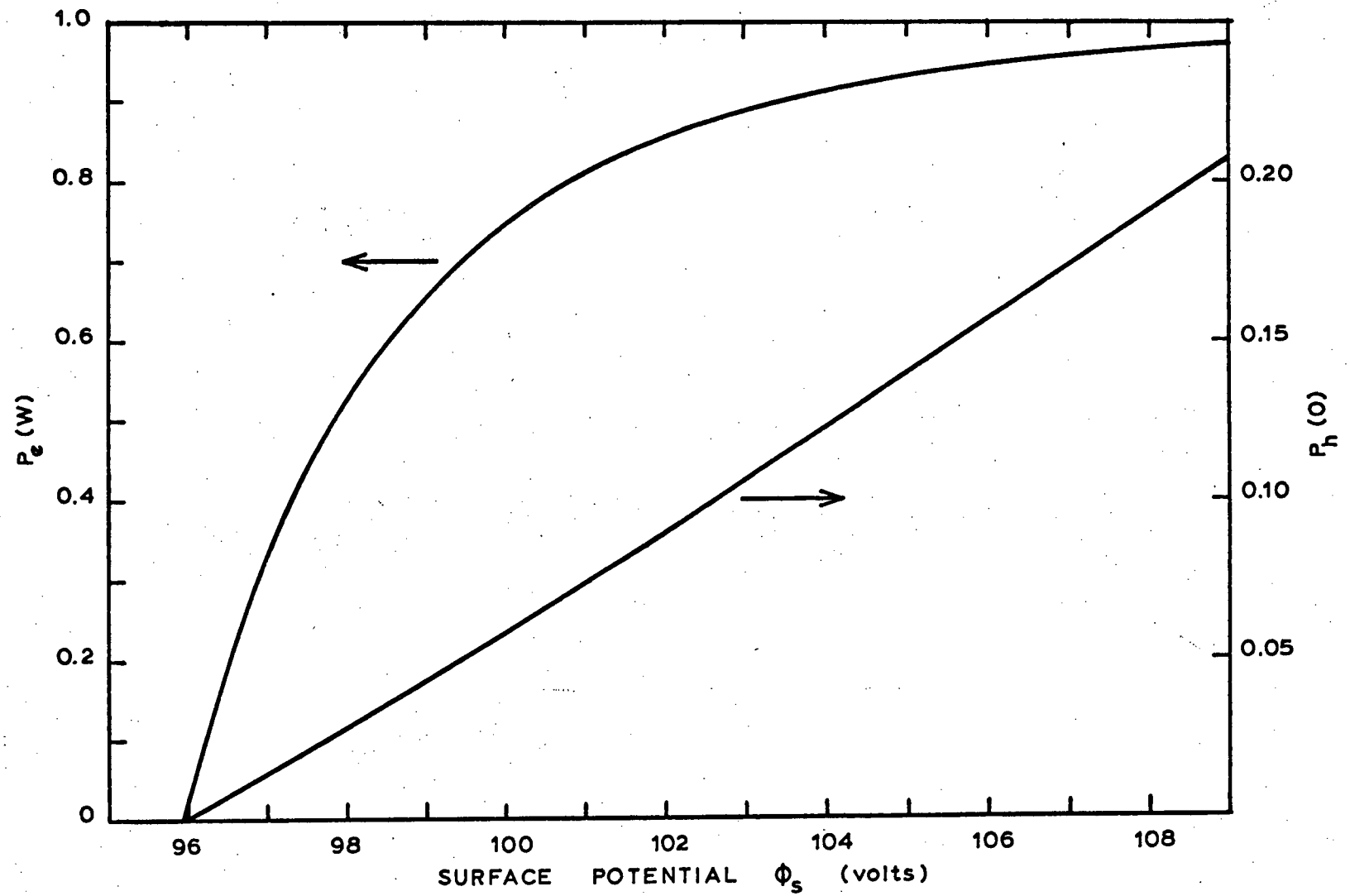


FIGURE 3.8 The avalanche initiation probability as a function of surface potential for electrons originating in the bulk and for holes originating at the Si-SiO_2 interface. Same parameters as for Fig. 3.7.

In order to solve for $P_e(x)$ and $P_h(x)$ using the coupled differential equations (3.2) and (3.3), the electric field distribution through the depletion layer must be known. McIntyre [32] has shown, however, that by making the approximation $\alpha_h = k\alpha_e$, where k is a constant independent of field, (3.2) and (3.3) may be combined to give integrable expressions for $P_h(0)$ and $P_e(w)$. The resultant expressions after integration are:

$$-\ln[1 - P_h(0)] = \left(\frac{k}{(1 - k)}\right) \ln[P_h(0) + f(w) + 1 - P_h(0)] \qquad (3.6)$$

$$1 - P_h(0) = [1 - P_e(w)]^k \qquad (3.7)$$

where $f(w) = \exp[(1 - k)\zeta]$, and $\zeta = \int_0^w \alpha_e(x)dx$.

From these equations, the probabilities $P_h(0)$ and $P_e(w)$ may be calculated as a function of the two parameters k and ζ without having to know the exact field distribution. Unfortunately, the approximation $\alpha_h = k\alpha_e$ is very poor for silicon, and it was found necessary to use equations (3.2) - (3.5) in order to obtain sufficiently accurate values for $P_e(w)$ and $P_h(0)$ when designing the MOS breakdown test structures and analyzing the experimental results.

3.1.2 Previous Experimental Investigations

The pulse rate of an avalanche photo diode biased above the breakdown voltage is determined by the product of the number of carriers transiting the depletion layer per unit time, and the appropriate avalanche initiation probability. By measuring the voltage dependence of the pulse rate under conditions of constant carrier injection, the avalanche initiation probability may be determined experimentally. Such measurements were made by Keil and Bernt [27] by observing the voltage dependence of the photon-induced pulse rate under constant illumination. They found evidence for saturation

of the triggering probability in some of the diodes they studied but the value of excess bias at which saturation occured is much lower than expected. The diodes they studied, however, were not free from microplasmas, and the diode they reported showing saturation had a dark count rate an order of magnitude larger than the photon-induced pulse rate, making their triggering probabilities somewhat suspect.

Oldham et al. [31] have operated small area, defect-free avalanche photodiodes above the breakdown voltage in order to provide experimental verification of their theoretically predicted avalanche initiation probabilities. The miniature avalanche diodes used were n^+p abrupt junction devices with a deeply diffused n^- guard ring to prevent premature edge breakdown and to define the smallest possible breakdown area (typically 3-5 μm in diameter). The diodes were illuminated from the n^+ side with 390 nm and 1050 nm radiation in order to provide hole and electron injection respectively. The high absorption coefficients at 390 nm ensured pure hole injection from the neutral region of the n^+ layer so that the photon induced pulse rate provided a direct measure of the avalanche initiation probability for holes entering the depletion region. Pure electron injection from the neutral p-type bulk was not possible in the case of illumination from the n^+ side. However, Oldham et al. argued that because of the small absorption coefficient at 1050 nm and the much larger electron ionization coefficients, the 1050 nm responses should be roughly proportional to the avalanche initiation probability for electrons entering the depletion region.

The measured response at 390 nm was found to follow the theoretical hole triggering probability very closely, however, only the linear region of the avalanche initiation probability versus excess bias was covered, and since only a very rough estimate of the absolute triggering probability was possible, the results do not, in fact, provide good confirmation of the

theory. The response at 1050 nm did show some saturation, but considerably less than was theoretically predicted. This was thought to be due to the voltage dependence of the effective collecting volume for electrons and to uncertainties in the available ionization rate data. Further investigations are required in order to verify that the triggering probabilities saturate with increasing overvoltage as predicted by the theory, and that it will be possible to operate an MOS gate in the plateau region of the electron triggering probability at reasonable excess biases.

3.2 DARK GENERATION OF TRIGGERING CARRIERS

Photogeneration must be the dominant mechanism for the production of triggering carriers if an MOS gate is to be operated above breakdown as a photon counting detector. The dark generation rate of triggering carriers must be made very small if one wants to detect low photon fluxes. How small the dark generation rate needs to be can be determined by examining its effect on the DQE of the sensor (Fig. 3.9). In order for the DQE not to be reduced by more than 10%, the dark count rate must be less than 5% of the event rate due to photogenerated carriers. In many cases it would be desirable to detect photon fluxes at least as low as 0.1 per second per pixel, for pixel areas on the order of 1000 μm^2, and hence the dark count rate should be maintained at a level below 500 $sec^{-1}cm^{-2}$ (corresponding to a dark leakage current of approximately 10^{-16} A cm^{-2}). A major effort will be required to achieve such low dark count rates.

Two basic mechanisms are responsible for the dark generation of carriers within or adjacent to the surface space charge region of an MOS gate. The first involves thermally activated processes that occur at bulk defect or impurity trapping levels and at the Si-SiO_2 interface states. The second mechanism, applicable only in the high field region, involves band to band tunneling or tunneling between one band and a bulk trapping level or surface state. These generation mechanisms are described in more detail below.

3.2.1 Review of Recombination and Generation at Bulk Defect or Impurity Centers

In a perfect (intrinsic) semiconductor there exists a forbidden gap between the valence and conduction bands which is free of states that can be occupied by electrons. Thermal generation of carriers in such a material

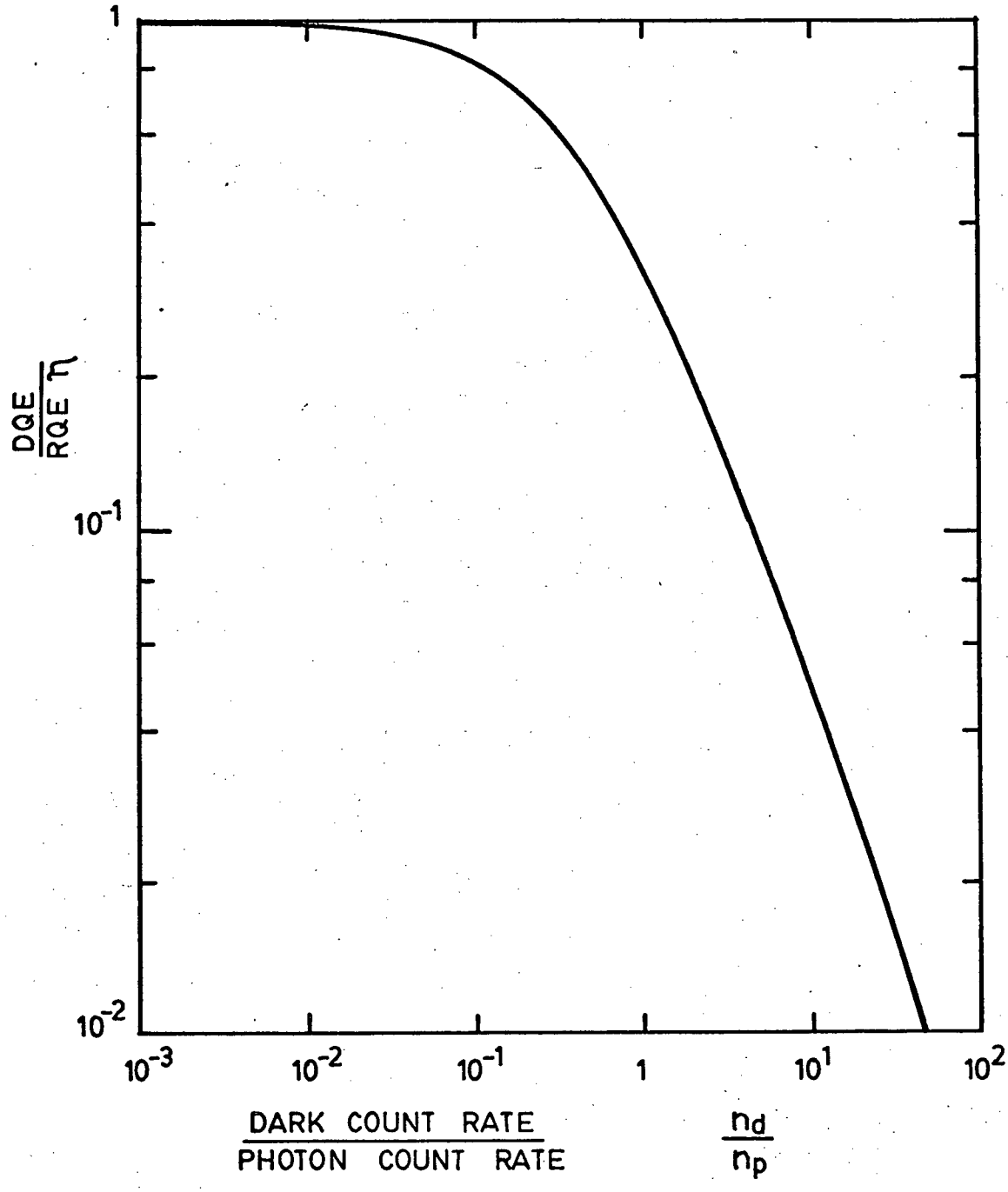


FIGURE 3.9 The degrading effect of the dark count rate on the detective quantum efficiency, plotted according to Eq. (2.13).

requires that an electron be raised directly from the valence to conduction band. Semiconductor devices on the other hand are made from extrinsic material that is doped with impurities having a low ionization energy for electrons and holes, to give predominantly n or p type conductivity. In addition to these dopant impurities, a number of other unwanted impurities with higher ionization energies are generally present or are unintentionally introduced during device processing. When the ionization energy of the impurity is higher it may not be completely ionized at normal operating temperatures and will act as a trap or recombination-generation center. These energy levels are not only caused by impurities but also by a large variety of crystal defects. The difference between shallow dopant levels, traps and recombination-generation centers is only qualitative. Which role the energy level will play depends on the temperature, the concentration of free carriers and the relative cross section for the capture of majority and minority carriers. Deep levels within the band gap are generally loosely referred to as simply "traps". The statistics of carrier capture and emission by such intermediate energy levels has been worked out by Shockley and Read [37] and by Hall [38]. For use in the analysis of dark generation mechanisms this theory will be briefly reviewed.

Following the basic concepts of Shockley and Read, four basic processes are defined for centers with a single energy level within the band gap (Fig. 3.10):

(a) Electron capture from the conduction band

(b) Electron emission into the conduction band

(c) Hole capture from the valence band

(d) Hole emission into the valence band

Processes (a) and (c) are described by two capture probabilities $c_n(sec^{-1})$ and $c_p(sec^{-1})$, while processes (b) and (d) are determined by the emission

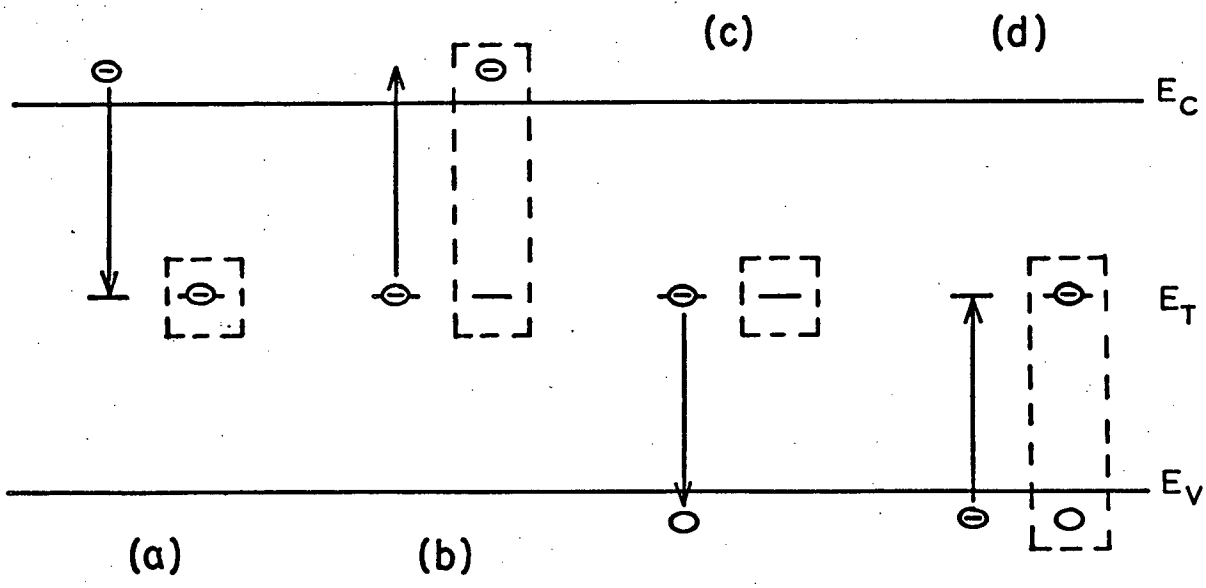


FIGURE 3.10 The four basic Shockley-Read-Hall processes that may occur at a trapping level. The final state is indicated in the dashed boxes, the arrow indicates the direction of the electron transition.

(a) electron capture
(b) electron emission
(c) hole capture
(d) hole emission

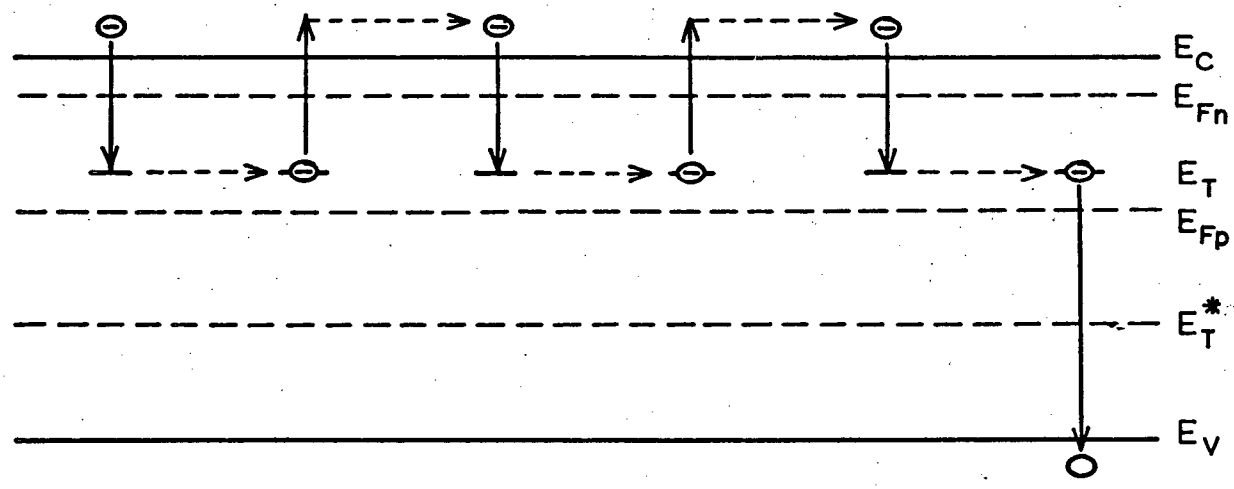


FIGURE 3.11 Action of an electron trap. The relative positions of E_{Fn}, E_T^* and E_{Fp} are also shown. Electrons are trapped and re-emitted several times before finally disappearing through recombination.

probabilities $e_n(sec^{-1})$ and $e_p(sec^{-1})$ respectively. Energy levels are further characterized by being donor levels or acceptor levels. A donor is neutral when occupied by an electron and positive when ionized, while an acceptor level is negatively charged when occupied by an electron and neutral when ionized. Doubly charged donors or acceptors are also possible.

In addition to the energy level, the capture cross-section is a basic property of a given trap. Experimentally, values for the cross-section are obtained by measuring the capture rate c. If the carriers all have the same velocity v_o the capture rate would be simply,

$$c_{n,p} = \sigma_{n,p} v_o$$

where $\sigma_{n,p}(cm^2)$ is the capture cross-section for electrons or holes. In reality the carriers have a thermal distribution of velocities and energy and furthermore, the capture cross-section may depend on electron energy. The cross-sections most frequently quoted are obtained by dividing the observed capture rate by the mean thermal velocity,

$$\sigma_{rms} = \frac{c}{(3kT/m^*)^{\frac{1}{2}}} = \frac{c}{v_{th}} \tag{3.8}$$

where m* is the conductivity effective mass. The experimental values of σ cover a considerable range, from 10^{-12} to 10^{-22} cm^2. This range of capture cross-sections can be qualitatively understood by considering that dependent on occupancy, a level may be either Coulombically attractive, neutral or repulsive.

If the concentration of centers in the semiconductor is N_T, a certain fraction f_T will be in the more negative state (e.g., ionized acceptors or neutral donors) and a fraction $(1 - f_T)$ will be in the more positive state (e.g., neutral acceptors or ionized donors). In thermal equilibrium the occupancy f_T is given by the Fermi-Dirac distribution,

$$f_T = \left\{1 + \frac{1}{g} \exp\left(\frac{E_T - E_F}{kT}\right)\right\}^{-1} \tag{3.9}$$

where E_F is the Fermi energy, E_T is the energy level of the center, and g is the spin degeneracy factor, usually assumed to be 2. The rate at which electrons and holes are captured depends on both the occupation of the centers and the density of free carriers. The rate at which carriers are emitted depends only on the occupation. The four rate equations corresponding to processes (a) - (d) are, therefore,

$$r_a = c_n n N_T(1 - f_T) \tag{3.10}$$

$$r_b = e_n N_T f_T \tag{3.11}$$

$$r_c = c_p p N_T f_T \tag{3.12}$$

$$r_d = e_p N_T(1 - f_T) \tag{3.13}$$

The electron concentration in the conduction band varies as

$$\frac{dn}{dt} = r_a - r_b = e_n N_T f_T - c_n n N_T(1 - f_T) \tag{3.14}$$

Similarly the rate of change of the hole concentration in the valence band is

$$\frac{dp}{dt} = r_d - r_c = e_p N_T(1 - f_T) - c_p p N_T f_T \tag{3.15}$$

It is interesting to calculate the limiting electron and hole lifetimes that occur when f_T is equal to zero and one respectively. From (3.14),

$$\tau_{no} = \frac{n}{\left|\frac{dn}{dt}\right|_{f_T=0}} = \frac{1}{c_n N_T} \tag{3.16}$$

similarly from (3.15),

$$\tau_{po} = \frac{1}{c_p N_T} \tag{3.17}$$

For more general values of f_T we need expressions for the emission rates e_n and e_p. By invoking the principle of detailed balance it is possible to obtain a relation between the capture and emission probabilities.

The principle of detailed balance states that in equilibrium

$$r_a = r_b \quad \text{and} \quad r_c = r_d.$$

For the case of Boltzmann statistics (i.e., for the Fermi level several kT below the conduction band edge) the electron concentration in the conduction band, in thermal equilibrium, is

$$\begin{aligned} n &= N_C \exp\left(\frac{E_F - E_C}{kT}\right) \\ &= n_i \exp\left(\frac{E_F - E_i}{kT}\right) \end{aligned} \tag{3.18}$$

where N_C is the effective density of states in the conduction band and E_C and E_i are the energy of the conduction band edge and the intrinsic level respectively. Equating r_a and r_b and using (3.9) for f_T and expressions (3.18) for n gives

$$\begin{aligned} \frac{e_n}{c_n} &= \frac{N_C}{g} \exp\left(\frac{E_T - E_C}{kT}\right) \\ &= \frac{n_i}{g} \exp\left(\frac{E_T - E_i}{kT}\right) \end{aligned} \tag{3.19}$$

The emission probability of holes can be obtained similarly from

$$\begin{aligned} p &= N_V \exp\left(\frac{E_V - E_F}{kT}\right) \\ &= n_i \exp\left(\frac{E_i - E_T}{kT}\right) \end{aligned} \tag{3.20}$$

where N_V is the effective density of states in the valence band and E_V is the energy level of the valence band edge. Thus,

$$\frac{e_p}{c_p} = N_V g \exp\left(\frac{E_V - E_T}{kT}\right)$$

$$= n_i g \exp\left(\frac{E_i - E_T}{kT}\right) \qquad (3.21)$$

As expected, the emission probabilities depend exponentially on the energy difference between the trap level and the respective band edge. Also, since the emission probabilities given by (3.19) and (3.21) are not functions of the Fermi energy E_F, the same expressions may be used in non-equilibrium situations.

For steady state, non-equilibrium conditions the rate of change of trapped charge is zero, therefore,

$$\frac{dn}{dt} = \frac{dp}{dt}$$

Equating (3.14) for dn/dt and (3.15) for dp/dt enables the steady state occupancy f_T to be determined.

$$f_T = \frac{e_p + c_n n}{e_n + e_p + c_n n + c_p p} \qquad (3.22)$$

If this is reinserted into either (3.14) or (3.15) the net steady state rate of recombination or generation is obtained.

$$U = \frac{dn}{dt} = \frac{dp}{dt} = \frac{(e_n e_p - c_n c_p np)\, N_T}{e_n + e_p + c_n n + c_p p} \qquad (3.23)$$

Another question of importance is whether a particular energy level will behave as a trap, a recombination, or a generation center. Any energy level may contribute to either recombination or generation. The difference between traps and recombination-generation centers lies in the relative capture and emission rates for electrons and holes. Carriers will be cap-

tured by traps, and reemitted into the band from which they came, several times before finally disappearing through recombination (Fig. 3.11). In terms of the four rate equations (3.10) - (3.13) the conditions for electron traps are:

$$r_a >> r_d \quad , \quad r_b >> r_c$$

Similarly for hole traps

$$r_c >> r_b \quad , \quad r_d >> r_a$$

For recombination centers the conditions are:

$$r_a >> r_d \quad , \quad r_c >> r_b$$

and for generation centers

$$r_d >> r_a \quad , \quad r_b >> r_c$$

Stockmann [39] has shown that it is possible to express these four sets of inequalities in the alternate form:

electron traps if, $E_{Fn}, E_{Fp} > E_T^*$ (3.24)

hole traps if, $E_T^* > E_{Fn}, E_{Fp}$ (3.25)

recombination centers if, $E_{Fn} > E_T^* > E_{Fp}$ (3.26)

generation centers if, $E_{Fp} > E_T^* > E_{Fn}$ (3.27)

Where E_T^* is defined by,

$$E_T^* \equiv 2\,E_i - E_T - kT\,\mathrm{Ln}\,(c_n/c_p) \tag{3.28}$$

and E_{Fn} and E_{Fp} are the non-equilibrium, quasi Fermi levels that describe the electron and hole concentrations according to:

$$n = n_i \exp\left(\frac{E_{Fn} - E_i}{kT}\right)$$

$$p = n_i \exp\left(\frac{E_i - E_{Fp}}{kT}\right)$$

If the capture rates (i.e., the capture cross-sections) for holes and electrons are equal, Eq. (3.28) reduces to

$$E_T^* - E_i = E_i - E_T$$

in which case the classification level E_T^* is in the mirror image position of the level E_T with respect to the intrinsic level E_i.

The treatment of intermediate levels thus far has been for centers with a single energy level only. In reality, many defect or impurity centers have several donor and/or acceptor levels. Shockley and Last [40] have shown, however, that equations (3.10) - (3.23) can be generalized to the case of a multilevel center, and that the behaviour is qualitatively very similar to the single level case. If the number of levels is not too great they are usually many kT apart at low temperatures and may be treated independently of each other.

3.2.2 Steady State Bulk Generation: Low Field Case

When defect or impurity centers are located within a reverse biased depletion region (where condition (3.27) holds) their occupancy is only determined by the emission probabilities e_n and e_p since capture is negligible due to the low carrier concentration in such a region. Setting $n = p = 0$ in (3.22) gives

$$f_T = \frac{e_p}{e_n + e_p} = \frac{1}{(e_n/e_p) - 1} \tag{3.29}$$

and from (3.23) the volume rate of generation of electron-hole pairs becomes

$$g_B = \frac{e_n e_p}{e_n + e_p} N_T \tag{3.30}$$

The centers must alternate between electron and hole emission. Eq. (3.29) relates the steady state occupancy to the ratio of the emission rates. From (3.19) and (3.21) this ratio can be expressed as

$$\frac{e_n}{e_p} = \frac{1}{g^2}\frac{c_n}{c_p}\exp 2\left(\frac{E_T - E_i}{kT}\right) \tag{3.31}$$

For levels away from midgap the exponential factor very rapidly takes over, forcing levels in the upper half of the band gap to be predominantly empty and in the lower half to be full of electrons. If many different levels are present within the forbidden gap, the steady state emission of electrons and holes will be dominated by those levels for which the electron and hole emission probabilities are equal. From (3.31) the energy of these levels is

$$E_{Tmax} = E_i + \frac{kT}{2}\ln\left(\frac{g^2 c_p}{kT}\right) \tag{3.32}$$

For similar electron and hole capture cross-sections, E_{Tmax} lies very close to E_i (i.e., mid gap).

Using expressions (3.19) and (3.21) for e_n and e_p in Eq. (3.30) for g_B with $E_T = E_{Tmax}$ gives

$$g_B = \frac{n_i}{2\tau_e} \tag{3.33}$$

where the effective lifetime τ_e is given by

$$\tau_e = \frac{1}{(c_n c_p)^{\frac{1}{2}} N_T} = \frac{1}{(\tau_{no}\tau_{po})^{\frac{1}{2}}}$$

The total rate of generation per unit area of electron-hole pairs is

$$G_B = \frac{n_i W}{2\tau_e}$$

where W is the depletion layer width.

The important thing to note from (3.33) is that the temperature dependence of g_B is determined mainly through n_i which varies as

$$n_i \propto T^{3/2} \exp\left(\frac{-E_g}{2kT}\right)$$

where $E_g \equiv E_C - E_V$

Lowering the temperature should, therefore, be a very effective means of reducing the steady state thermal generation of carriers in the depletion region of a wide band gap semiconductor such as silicon. Typical bulk lifetimes τ_{no} and τ_{po}, after CCD processing, are on the order of 10^{-4} sec. which gives $g_B = 8 \times 10^{13}$ sec $^{-1}$cm^{-3} at room temperature. This is reduced by 20 orders of magnitude upon cooling to 100 K. It will, therefore, clearly be possible to reduce this type of generation to well below 500 carriers sec^{-1} cm^{-2} even for very wide depletion regions (several tens of microns).

In addition to the generation occuring in the space charge region, one also has to consider the diffusion of minority carrier electrons from the neutral bulk. For thinned CCD's the diffusion length of these minority carriers is generally much greater than the substrate thickness, and it is the condition of the back surface that determines the size of the electron injection current. The injection current can be obtained by solving the continuity equation for minority carrier electrons in the neutral bulk, subject to the appropriate boundary conditions. The analysis may be simplified by making the following assumptions:

(1) the minority carrier diffusion length in the bulk away from the surface is very much greater than the substrate thickness L so that recombination/generation in the bulk can be neglected in comparison with recombination/generation at the rear surface (i.e., $L/\tau_n \ll s$).

(2) the rear silicon surface is a region of high electron hole recombination, characterized by a high value of surface recombination velocity s.

With these assumptions the continuity equation for electrons becomes

$$D_n \frac{d^2 n}{dx^2} = 0$$

subject to the boundary conditions

$$D_n \left.\frac{dn}{dx}\right|_{x=L} = s[n_o - n(L)] \quad , \quad \text{and } n(0) = 0$$

where: n_o = equilibrium electron concentration

D_n = electron diffusivity

x=0 = edge of the depletion region

L = thickness of the neutral region

The electron injection current is, therefore,

$$J_{ndiff} = qD_n \left.\frac{dn}{dx}\right|_{x=0} = qD_n \frac{sn_o}{(sL + D_n)} \tag{3.34}$$

With a very high rear-surface recombination velocity, such that $s >> D_n/L$, J_{ndiff} can be approximated by

$$J_{ndiff} = qD_n \frac{n_o}{L} = \frac{qD_n n_i^2}{L N_A}$$

while for a low rear-surface recombination velocity, $\frac{L}{\tau_n} << s < \frac{D_n}{L}$, J_{ndiff} becomes

$$J_{ndiff} = qsn_o = qs \frac{n_i^2}{N_A} \quad .$$

The n_i^2 dependence ensures that the minority carrier injection diffusion current can be reduced to an entirely negligible level by cooling.

3.2.3 Steady State Generation at the Silicon/Silicon Dioxide Interface: Low Field Case

The disruption of the periodic lattice structure at the silicon/silicon dioxide interface introduces a high density of available energy states in the forbidden gap near the interface. The density of the interface states de-

pends on the orientation of the silicon substrate and very critically on the oxidation and annealing processes that the silicon sample is subjected to. Silicon surfaces on a (100) crystal plane have been found to have the lowest interface state densities. For a properly annealed, thermally grown oxide on (100) silicon, the interface state density is in the range 10^{9} - 10^{11} $cm^{-2}eV^{-1}$ with a distribution in energy which is relatively flat in the middle of the band gap and which peaks towards the conduction and valence band edges [41].

Although the physical origin of the interface states is different from that of the bulk centers, surface generation can be treated in a manner sililar to bulk generation. In the absence of an inversion layer, the concentration of both electrons and holes at the interface is very low and the steady state generation rate of electron hole pairs is given by an expression similar to (3.30),

$$G_s = \int_{E_V}^{E_C} \left(\frac{e_n e_p}{e_n + e_p}\right) N_s(E_T)\, dE_T \tag{3.35}$$

where $N_s(E_T)$ is the distribution of interface states per unit area per unit energy. Using (3.19) and (3.21) for e_n and e_p (3.35) becomes

$$G_s = \int_{E_V}^{E_C} \frac{n_i c_n c_p N_s(E_T)}{c_n \exp\left(\frac{E_T - E_i}{kT}\right) + c_p \exp\left(\frac{E_i - E_T}{kT}\right)}\, dE_T \tag{3.36}$$

The major contribution to the integral comes from those interface states within a few kT of the intrinsic level E_i. By further assuming that the electron and hole capture cross-sections are equal (i.e., $\sigma_n = \sigma_p$, $c_n = c_p$), the integral can be evaluated approximately as

$$G_s \simeq \frac{1}{2}\pi\sigma v_{th} kT\, N_s(E_i)\, n_i$$

$$= \frac{1}{2} s\, n_i \qquad (3.37)$$

where s is the commonly used parameter called the surface recombination velocity. The assumption $\sigma_n = \sigma_p$ is not really justified because experimental data indicates that the electron and hole capture cross-sections differ by as much as an order of magnitude. However, the error made by setting $\sigma_n = \sigma_p$ is at most a factor of 2. Typically, the mid gap interface state density after CCD processing is 10^9 $cm^{-2}eV^{-1}$. The capture cross-sections for electrons and holes by mid gap interface states are about $10^{-14} cm^2$, therefore, the room temperature steady state generation of electron hole pairs at the interface is $G_s \simeq 8 \times 10^{10}$ $cm^{-2}sec^{-1}$, corresponding to a surface recombination velocity of approximately 10 cm sec^{-1}. In p-type substrates it is only the holes emitted from interface states that are able to traverse the depletion region and trigger a discharge. Since the temperature dependence of G_s is determined mainly by n_i, cooling should be an effective means for reducing G_s to well below 500 $sec^{-1}cm^{-2}$.

When an inversion layer is present, capture processes can no longer be neglected. For p-type substrates the effect of electron capture from the inversion layer is to keep the interface states filled up to a level considerably above the intrinsic level, and the hole emission rate, therefore, becomes extremely small. In this case the steady state occupancy can be shown to be

$$f_T = \left\{\exp\left(\frac{E_T - E_{Fn}}{kT}\right) + 1\right\}^{-1}$$

Thus the position of the quasi fermi level for electrons, E_{Fn}, determines the occupation of the interface states.

3.2.4 High Field Effects

The high electric fields within the space charge region of an avalanche diode or MOS gate, biased above breakdown, can substantially influence carrier emission from traps through either the Poole-Frenkel effect or tunneling. These field effects are illustrated schematically in Fig. 3.12 which shows the distortion of the coulomb potential well around a trapping center with increasing electric fields. At low fields carrier emission is thermally activated. As the field increases, the potential barrier is lowered, thus enhancing the probability of thermal emission. At higher fields tunneling begins to dominate the emission process and finally, at very high fields, the barrier is lowered below the ground state of the trap and it becomes delocalized. Delocalization generally need not be considered, however, since emission by tunneling increases very rapidly with increasing field, and empties the traps before delocalization occurs.

For a trap to experience the Poole-Frenkel effect it must be coulomb attractive to the emitted carrier. If it is neutral after carrier emission, barrier lowering will not occur due to the absence of the coulomb potential. The Poole-Frenkel effect can, therefore, only increase the emission rate of electrons from donor levels or holes from acceptor levels. The barrier lowering is similar to Schottky barrier lowering in a metal-semiconductor junction. The image charge, however, is fixed rather than mobile as in Schottky emission, resulting in a barrier lowering twice as great for the Poole-Frenkel effect. Referring to Fig. 3.12(b),

$$\Delta\phi_{PF} = \left(\frac{q^3\mathcal{E}}{\pi\epsilon}\right)^{\frac{1}{2}} = \beta\sqrt{\mathcal{E}} \qquad (3.38)$$

where $\mathcal{E}$ is the applied electric field and ϵ is the high frequency dielectric constant. For silicon it has been shown that the appropriate dielectric

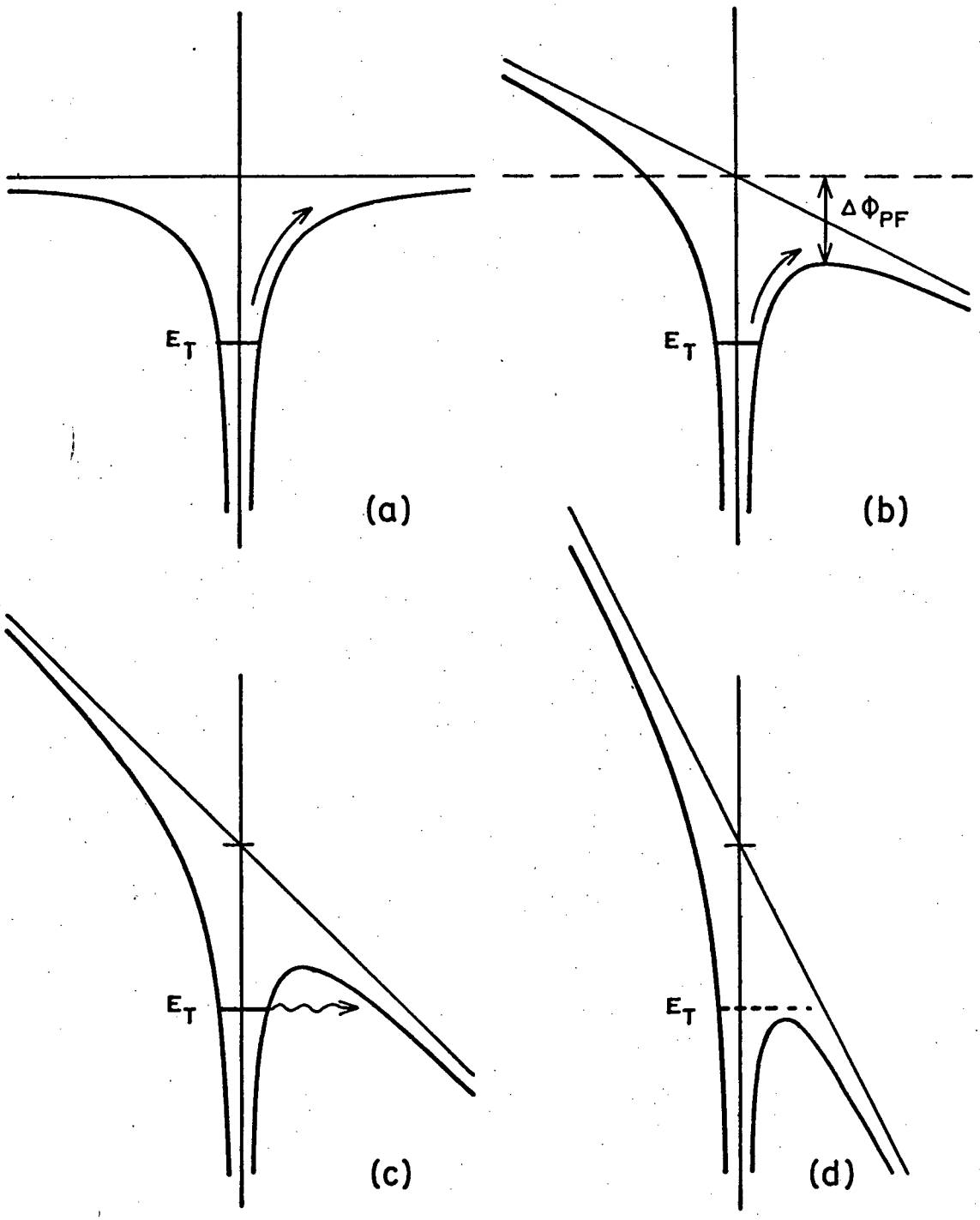


FIGURE 3.12 Schematic diagram showing the distribution of the coulomb potential well around a trapping center for different electric field strengths.

(a) low field - thermal emission dominates
(b) moderate field - thermal emission with reduced barrier
(c) high field - tunnelling dominates
(d) very high field - trap delocalization

constant is very close to the static value for fields up to at least 1.5×10^5 V cm^{-1} [42]. Eq. (3.38) gives the maximum barrier lowering in the direction of the applied field. In order to calculate the resulting emission probabilities, however, a three dimensional model is required.

The three dimensional treatment of Poole-Frenkel emission from traps has been developed by Hartke [43]. The resulting expression for the barrier lowering, in polar coordinates, is

$$\Delta\phi_{PF}(\theta) = \beta(\mathcal{E}\cos\theta)^{\frac{1}{2}} \tag{3.39}$$

where the barrier is lowered only for $0<\theta<\pi/2$. The ratio of the field enhanced emission rate e_{PF} to the zero field emission rate e_o, is approximated by Hartke to be

$$\frac{e_{PF}}{e_o} = \left(\frac{kT}{\beta\sqrt{\mathcal{E}}}\right)^2 \left\{1 + \left(\frac{\beta\sqrt{\mathcal{E}}}{kT} - 1\right)\exp\left(\frac{\beta\sqrt{\mathcal{E}}}{kT}\right)\right\} + \frac{1}{2} \equiv \gamma \tag{3.40}$$

In deriving (3.40) Hartke assumes a spherically symmetric, field independent, attempt-to-escape frequency of $\nu/4\pi$ per unit solid angle where ν is given by the relation

$$e_o = \nu \exp(-\Delta E/kT)$$

$$\Delta E = E_c - E_T \text{ for donor levels}$$

$$= E_T - E_v \text{ for acceptor levels}$$

For high fields or low temperatures where $\beta\sqrt{\mathcal{E}} >> kT$, (3.40) may be approximated by

$$\frac{e_{PF}}{e_o} \approx \frac{kT}{\beta\sqrt{\mathcal{E}}} \exp\left(\frac{\beta\sqrt{\mathcal{E}}}{kT}\right) \tag{3.41}$$

Fig. 3.13 shows this approximate form for e_{pF}/e_o along with the more accurate value given by (3.40), plotted as a function of $\beta\sqrt{\mathcal{E}}/kT$.

The effect of Poole-Frenkel emission on the steady state generation in

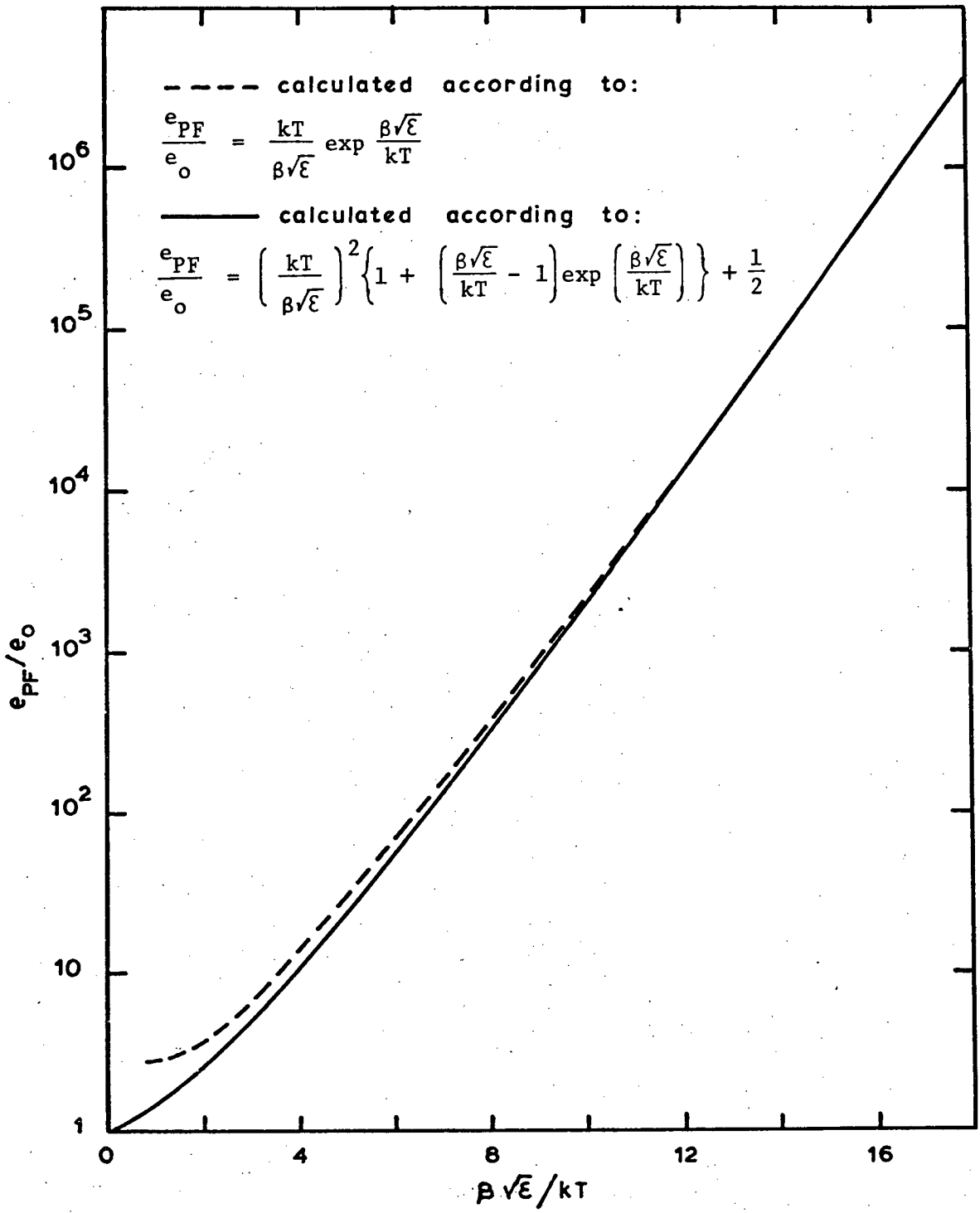


FIGURE 3.13 Ratio of the field enhanced emission rate e_{pF} to the zero field emission rate e_o as a function of $\beta\sqrt{\mathcal{E}}/kT$, $\beta = (q^3/\pi\epsilon_s)^{\frac{1}{2}}$.

the depletion layer can be estimated as follows. From (3.30) and (3.40) the revised steady state generation (for the case of donor levels) is

$$g_B = \frac{\gamma e_{no} e_{po}}{\gamma e_{no} + e_{po}} N_T \tag{3.42}$$

As before, the most effective levels for steady state emission are those for which $\gamma e_{no} = e_{po}$, i.e., those levels at

$$E_{Tmax} = E_i + \frac{kT}{2} \ln\left(\frac{g^2 c_p}{\gamma c_n}\right) \tag{3.43}$$

Using expressions (3.19) and (3.21) for e_{po} and e_{no} in (3.42), with E_T given by (3.43), gives

$$g_B = \frac{\gamma^{\frac{1}{2}} n_i}{2\tau_e} \tag{3.44}$$

and the rate of generation per unit area is

$$G_B = \frac{n_i}{2\tau_e} \int_0^w [\gamma(x)]^{\frac{1}{2}} dx \tag{3.45}$$

Identical expressions for g_B and G_B are obtained for acceptor levels.

The steady state generation rate g_B is, therefore, increased by a factor $\gamma^{\frac{1}{2}}$ when Poole-Frenkel high field effects are included. Using the approximate expression for γ given in (3.40) enables the temperature dependence of g_B to be expressed as

$$g_B \propto T^2 \exp\left(\frac{\beta\sqrt{\mathcal{E}} - E_g}{2kT}\right)$$

Therefore, provided the barrier lowering $\beta\sqrt{\mathcal{E}}$ is only a small fraction of the energy gap, lowering the temperature should still be a very effective means of reducing this type of steady state generation. As will be shown in the

next section, the peak field in the space charge region of silicon diodes or MOS gates must be kept below approximately 4.2×10^7 Vm^{-1} in order to avoid significant carrier generation due to interband tunneling. The Poole-Frenkel barrier lowering for a field of this magnitude is 0.14 eV, which is indeed only a small fraction of the silicon band gap.

Poole-Frenkel effects at the interface can be treated in precisely the same manner, with the result that the surface recombination velocity is increased by a factor $\gamma^{\frac{1}{2}}$,

$$G_s = \frac{1}{2}\gamma^{\frac{1}{2}}\pi\sigma v_{th} kT\, N_s(E_i)\, n_i$$
$$= \frac{1}{2} s_{PF}\, n_i \qquad (3.46)$$

At very high fields it becomes possible for electrons to make tunneling transitions between bulk trapping levels and either band, leading to electron and hole emission probabilities which are virtually independent of temperature. It also becomes possible for electrons to make tunneling transitions from the valence band into empty interface states. Under steady state conditions these tunneling processes may combine with the thermally activated emission from traps or interface states, or a two step tunneling process may occur from the valence band into a bulk trap then from the trap into the conduction band. These basic steady state tunneling generation mechanisms are illustrated in Fig. 3.14. Tunneling mechanisms involving more than one trap (examples of which are illustrated by the dashed lines in Fig. 3.14) are also possible but are expected to be relatively unimportant compared to the single level processes.

When tunneling and Poole-Frenkel effects are included, the rate of emission of holes from donor levels becomes

$$\frac{dp}{dt} = [e_{po} + e^{o}_{pt}(x)]N_T(1 - f_T) \qquad (3.47)$$

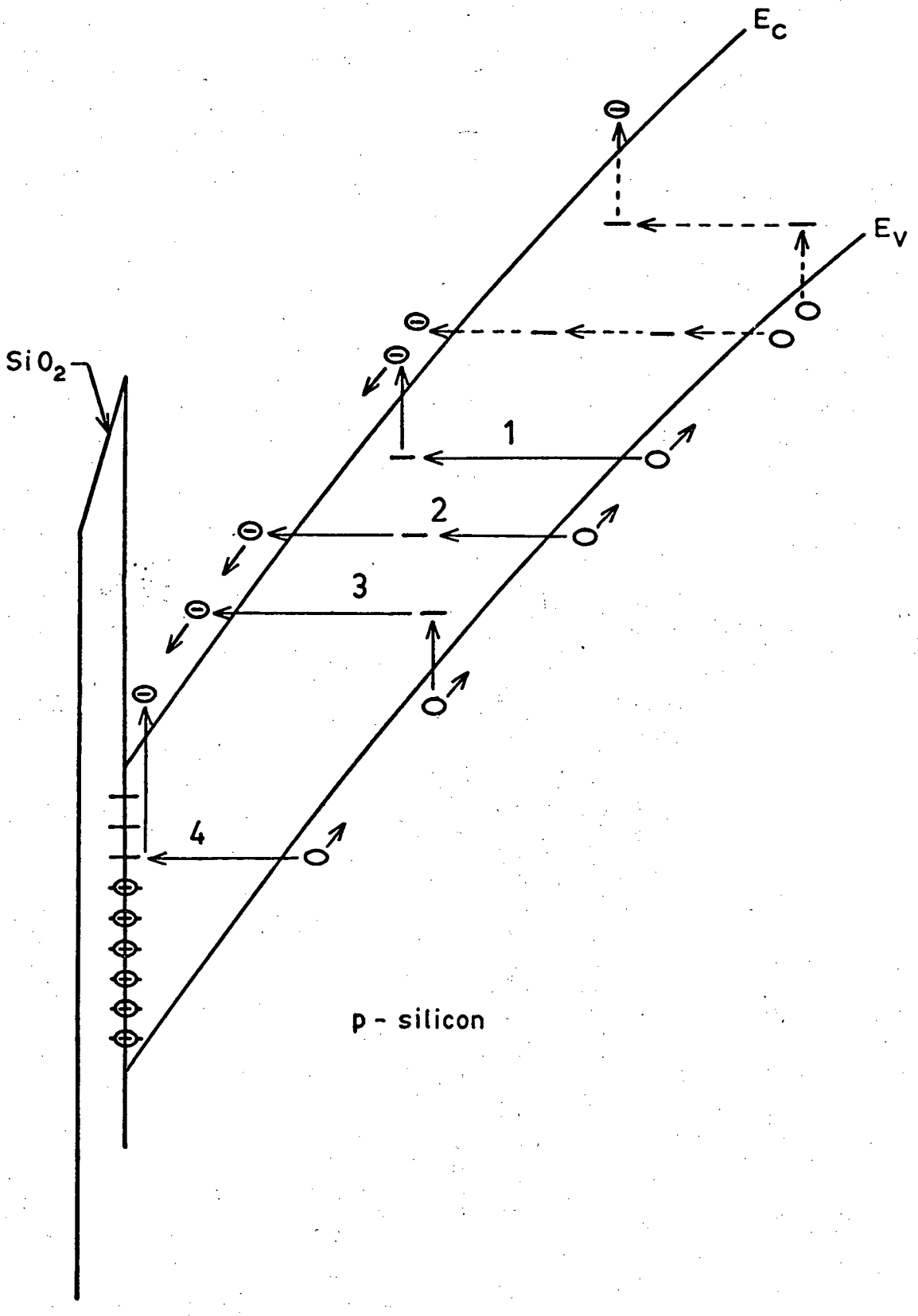


FIGURE 3.14 Energy band diagram illustrating the proposed steady state generation mechanisms involving the tunnelling emission of electrons and/or holes by mid gap levels.

where e_{pt}^{o} is a hole emission probability equal to the probability per unit time for tunneling from the valence band to an unoccupied donor level. The rate of emission of electrons becomes

$$\frac{dn}{dt} = [\gamma(x)e_{no} + e_{nt}^{+}(x)]N_T f_T \tag{3.48}$$

where e_{nt}^{+} is an emission probability corresponding to tunneling from an occupied donor level to the conduction band. Expressions (3.29) for f_T, and (3.30) for g_B, therefore, still hold provided the emission probabilities e_n and e_p are replaced by the new emission probabilities e_n^* and e_p^*, defined by:

$$e_n^*(x) = \gamma(x)e_{no} + e_{nt}^{+}(x) \tag{3.49}$$

$$e_p^*(x) = e_{po} + e_{pt}^{o}(x) \tag{3.50}$$

For acceptor levels the new emission probabilities are

$$e_n^*(x) = e_{no} + e_{nt}^{o}(x) \tag{3.51}$$

$$e_p^*(x) = \gamma(x)e_{po} + e_{pt}^{-}(x) \tag{3.52}$$

The appropriate emission probabilities to be used in expression (3.35) for G_s are $e_n^*(0)$ and $e_p^*(0)$. For p-type substrates $e_{nt}^{+}(0)$ is, of course, equal to zero.

For shallow donor levels the electron emission probability becomes very high, and the rate determining step in the steady state generation of electron-hole pairs is either thermally activated hole emission or tunneling from the valence band to the trap. At low temperatures the probability of hole emission by tunneling may exceed the thermal emission probability by many orders of magnitude in the high field region of the depletion layer. The tunneling distance, however, is close to that for band to band tunneling

and since the density of traps N_T is many orders of magnitude smaller than the density of states in the valence band, tunneling through shallow donors is expected to be unimportant compared to interband tunneling. A similar argument holds for shallow acceptor levels. The situation is quite different for deep bulk levels in the high field region. At low temperatures, thermal emission from these levels is generally negligible compared to the tunneling transitions. Thus, the generation of electrons and holes through deep levels in the high field region will proceed almost entirely by mechanism 2 in Fig. 3.14. Since the tunneling distances are approximately half that for band to band tunneling, this type of carrier generation can become significant. Tunneling through deep level traps is considered in more detail in section 3.2.5.

In the absence of an inversion layer, tunneling of electrons into unoccupied interface states, as in mechanism 4 in Fig. 3.14, can greatly increase the hole emission probability at low temperatures. This, together with the enhanced electron emission probability due to Poole-Frenkel barrier lowering, can substantially increase the effective surface recombination velocity. It may, therefore, be necessary to maintain the surface in strong inversion (i.e., E_{Fn} very close to the conduction band edge) in order to increase the tunneling distance for hole emission.

3.2.5 Dark Generation Due to Tunneling

When the electric field in an insulator or semiconductor is sufficiently high, the forbidden energy gap may be treated in the manner of a potential barrier of finite width w given by

$$w = E_g/q\mathcal{E},$$

and thus it is possible for the valence band electrons to make direct quantum mechanical tunneling transitions to the conduction band. Normally hot

electron effects, such as impact ionization and avalanche, precede tunneling and the field never reaches a value where tunneling becomes important. Observation of large tunneling currents is restricted to specially-made, narrow p-n junctions in heavily doped semiconductors, often referred to as Esaki diodes [44]. In this investigation, however, we are concerned with extremely small reverse bias currents (of the order of 10^{-16} A) and, furthermore, the maximum fields are higher than those which would normally cause avalanche breakdown. It is therefore necessary to consider interband tunneling as a possible dark generation mechanism even for junctions or MOS gates formed on relatively lightly doped substrates.

Many theoretical and experimental studies of interband tunneling have been reported in the literature. These studies, however, are generally directed at explaining the current voltage characteristics of Esaki tunnel diodes in which the built-in field is large enough to cause significant tunneling at zero bias. Such diodes generally have degenerately doped n and p regions. The only experimental results known to the author on the tunneling generation of triggering carriers in lightly doped avalanche diodes operating above breakdown, are those of Haitz [36]. Haitz measured the dark count rate for diodes known to have very low densities of bulk trapping levels and showed that the observed dependence of the count rate on temperature and peak field could be described by an interband tunneling generation of carriers. From Haitz's dark count rate data and the quoted junction area, the reverse bias tunnel currents for his 30V breakdown (one-sided step junction) diodes is estimated to be 2×10^{-12} A cm^{-2} (before multiplication) at breakdown increasing to 2×10^{-8} A cm^{-2} at 10 V above breakdown. Clearly, diodes or MOS structures that have lower peak fields must be considered if the dark generation is to be maintained below 10^{-16} A cm^{-2}.

In one-sided step junctions or MOS gates, the peak field at break-

down may be reduced by decreasing the substrate doping. Unfortunately, as regards fabrication of a PC-CCD, this also has the undesirable effect of increasing the breakdown voltage. In order to examine more closely the effects of substrate doping and different doping profiles on the dark generation due to tunneling, we need an accurate expression for the volume rate of electron-hole pair generation that my be integrated over the depletion region.

Interband tunneling calculations can be classified into two general categories according to whether the traditional Zener field-emission model [45] or the more recent Fredkin-Wannier junction potential model [46] is used. In both pictures of interband tunneling the externally applied field is treated as a pertubation on the atomic forces. This is justified in most cases since, at the highest fields normally encountered, the change in external potential V over a lattice constant is small compared to the amplitude of the periodic crystal potential. Electrons moving under the combined potentials are normally represented as Bloch wave packets constructed from the eigenstates of the Hamiltonian for the unperturbed crystal, and can be treated semiclassically as free particles affected only by the external potential V but responding according to effective dynamical laws. Considered classically, the electron then has, at any instant, a band index n, a crystal momentum k, and a position r. The effective dynamics represents a quantum treatment of the crystal potential and a classical treatment of the external potential. With a full quantum mechanical treatment there exists the possibility of a change in band index n by interband tunneling. WKB methods [47] or ordinary time-dependent perturbation theory may be used to include the slowly varying external potential in the quantum mechanical calculation and arrive at a tunneling probability.

The Fredkin-Wannier model for interband tunneling applies specific-

ally to Esaki diodes and considers the motion of Bloch electrons in an external potential that changes over a short distance from one constant value to another. In such a model the tunneling probability for an electron colliding with the junction barrier is essentially defined in terms of the formalism of scattering theory and is expressed as a cross-section for scattering across to the other side. Since we require the volume rate of generation of electron-hole pairs in a high field region that extends over a considerable distance, the theoretical results based on the older field emission model will be easier to apply. Here the external potential is taken to be $V(x) = -Fx$, representing a constant field $F = q\mathcal{E}$. Quasi-classically, under the influence of the constant electric field, the electrons cycle through the Brillouin zone at a constant rate $\dot{\mathbf{k}} = F/h$, with a period $\tau = hK/F$ where K is the width of the Brillouin zone. The rate of leakage into the adjacent band is greatest when the $\mathbf{k}$ vector is at the band edge. The volume rate of generation due to Zener field emission is obtained by calculating a transition probability per unit time which is then integrated over the Brillouin zone.

In indirect band gap materials, such as silicon, the tunneling calculations are complicated by the requirement for phonon cooperation. Tunneling across an indirect energy gap involves a change in electron momentum perpendicular to the tunneling direction. Since there is no force perpendicular to the tunneling direction, this may only occur by the emission or absorption of a phonon. The values of the fundamental phonon energies in silicon that would conserve momentum in a tunneling transition are the phonon energies that occur at the same position in the Brillouin Zone as the conduction band minimum, and are [48]:

the transverse acoustic (TA) at 17.9 meV

the longitudinal acoustic (LA) at 43.7 meV

the longitudinal optic (LO) at 53.2 meV

the transverse optic (TO) at 58.5 meV

The functional dependence of the phonon assisted tunneling probability with electric field is very similar to the result for direct tunneling. The main difference is an additional prefactor for the probability of phonon scattering that may reduce the overall tunneling probability by as much as three orders of magnitude at any given field.

Indirect phonon assisted tunneling has been considered by Kane [49] [50] using the constant electric field model. After summing the transition probability per unit time over all possible initial and final states, the resulting expression for the volume rate of generation of electron-hole pairs applicable at large reverse bias, is

$$g_t = A \frac{\mathcal{E}^{5/2}}{E_g^{7/4}} \sum_{E_p} [M(E_p)]^2 \{(u+1)\exp[\frac{-B}{\mathcal{E}}(E_g + E_p)^{3/2}] + u \exp[\frac{-B}{\mathcal{E}}(E_g - E_p)^{3/2}]\} \quad (3.53)$$

where A is a constant and $M(E_p)$ is the matrix element for phonon scattering. B depends on the reduced effective mass for tunneling and on the shape of the potential barrier in the forbidden gap. Kane assumes a triangular potential barrier, for which

$$B = \frac{4(2m_r^*)^{1/2}}{3qh} \quad (3.54)$$

E_p is the phonon energy and u is the phonon occupation, given by

$$u = \frac{1}{\exp(E_p/kT) - 1} \quad (3.55)$$

The summation should extend over the four fundamental phonon energies mentioned earlier, as well as over the phonon energies involved in any multiple phonon scattering events. In silicon, however, it has been established experimentally [51] [52] that the transverse acoustic and transverse optic phonons at 17.9 meV and 58.5 meV respectively, are the dominant scattering agents for indirect tunneling. We will make the further approximation that the matrix elements for phonon scattering are equal for these two phonons.

In an actual diode or MOS gate the field $\mathcal{E}$ is not constant but is a function of position x in the depletion region. The rate of tunneling generation per unit area therefore becomes

$$G_t = \int_0^W g_t(x)\,dx$$

$$= \frac{AM^2}{E_g^{7/4}} \int_0^W \mathcal{E}(x)^{5/2} \sum_{TA,TO} \{(u+1)\exp[\frac{-B}{\mathcal{E}(x)}(E_g + E_p)^{3/2}] + u\exp[\frac{-B}{\mathcal{E}(x)}(E_g - E_p)^{3/2}]\}\,dx \qquad (3.56)$$

In practice the values of AM^2 and B must be determined experimentally. Fairly reliable estimates for these values were obtained by fitting Haitz's [36] dark count rate data to (3.56). The following formulas and numerical values were used in the calculations:

w_1 (width constant for Haitz's step junction)
$= 0.23 \times 10^{-6}\ mV^{-1/2}$

T (substrate temperature) = 196 K

u (phonon occupation, at 196 K)
= 0.52 for TA phonon at 17.9 meV
= 0.03 for TO phonon at 58.5 meV

E_g (band gap, at 196 K) = 1.14 ev

$$\mathcal{E}(x) = 2/w_1 \left[w_1(V_a + V_i)^{\frac{1}{2}} - x\right]$$

where $(V_a + V_i)$ = applied voltage plus built-in voltage

The values of AM^2 and B for best fit (at large excess biases, where the avalanche initiation probability is close to one) are:

$$AM^2 = 5.1 \times 10^{18}\ eV^{7/4} V^{-\frac{1}{2}} sec^{-1}$$

$$B = 1.89 \times 10^{9}\ eV^{-3/2}\ V\ m^{-1}$$

By using (3.54) for B the reduced effective mass for tunneling becomes $m_r^* = 0.077\ m_o$, which is in reasonable agreement with what one would expect for tunneling from the light mass valence band in the light mass direction of the conduction band.

With the above values for AM and B, (3.56) may be used to predict the generation rate due to interband tunneling for different substrate doping levels. Figure 3.15 shows the calculated generation rate of electron hole pairs versus peak electric field, for MOS gates operating above breakdown at a temperature of 100 K. The calculated generation rate versus surface potential is shown in Fig. 3.16. The corresponding band diagram is that shown in Fig. 3.2. Although the generation rates shown are based on Haitz's data, they have been extrapolated to values eight orders of magnitude lower than the minimum generation rate he measured. Assuming the theoretical expression used for the extrapolation (Eq. 3.56) is accurate, the estimated error bars for Haitz's data lead to an error after extrapolation of approximately plus or minus one order of magnitude for the lowest generation rates indicated in Figures 3.15 and 3.16. It is apparent from the results shown in Fig. 3.15 that the peak electric field at the $Si\text{-}SiO_2$ interface should be kept below approximately $4.3 \times 10^{7}\ Vm^{-1}$ in order to ensure a dark generation rate less than $500\ sec^{-1}cm^{-2}$.

A substrate doping level below approximately $8 \times 10^{15} cm^{-3}$ is required

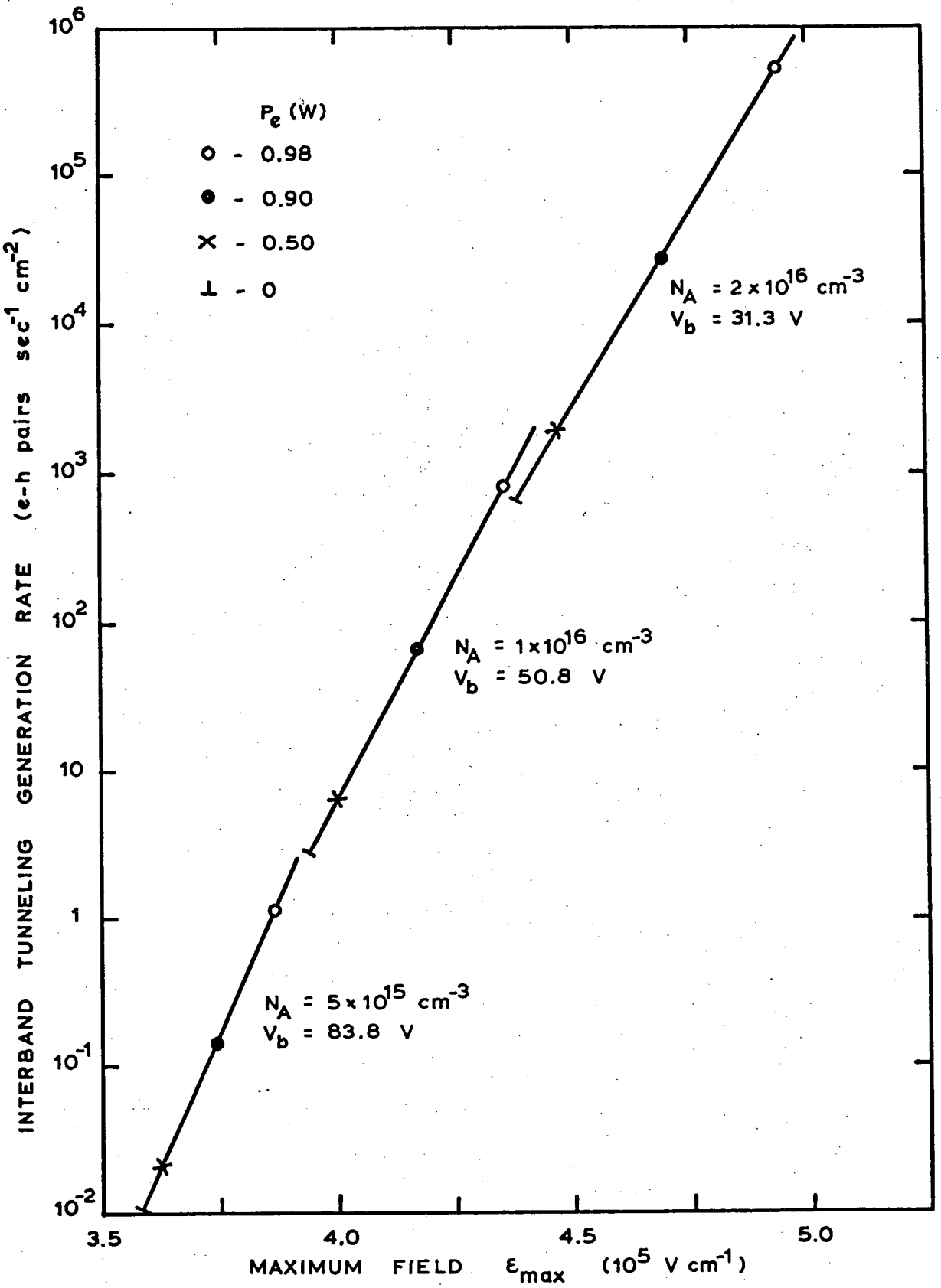


FIGURE 3.15 Interband tunnelling generation rate versus the peak electric field in n+p step junctions or p-substrate MOS gates, with different levels of substrate doping, T = 100K. The avalanche initiation probability $P_e(W)$ is also indicated.

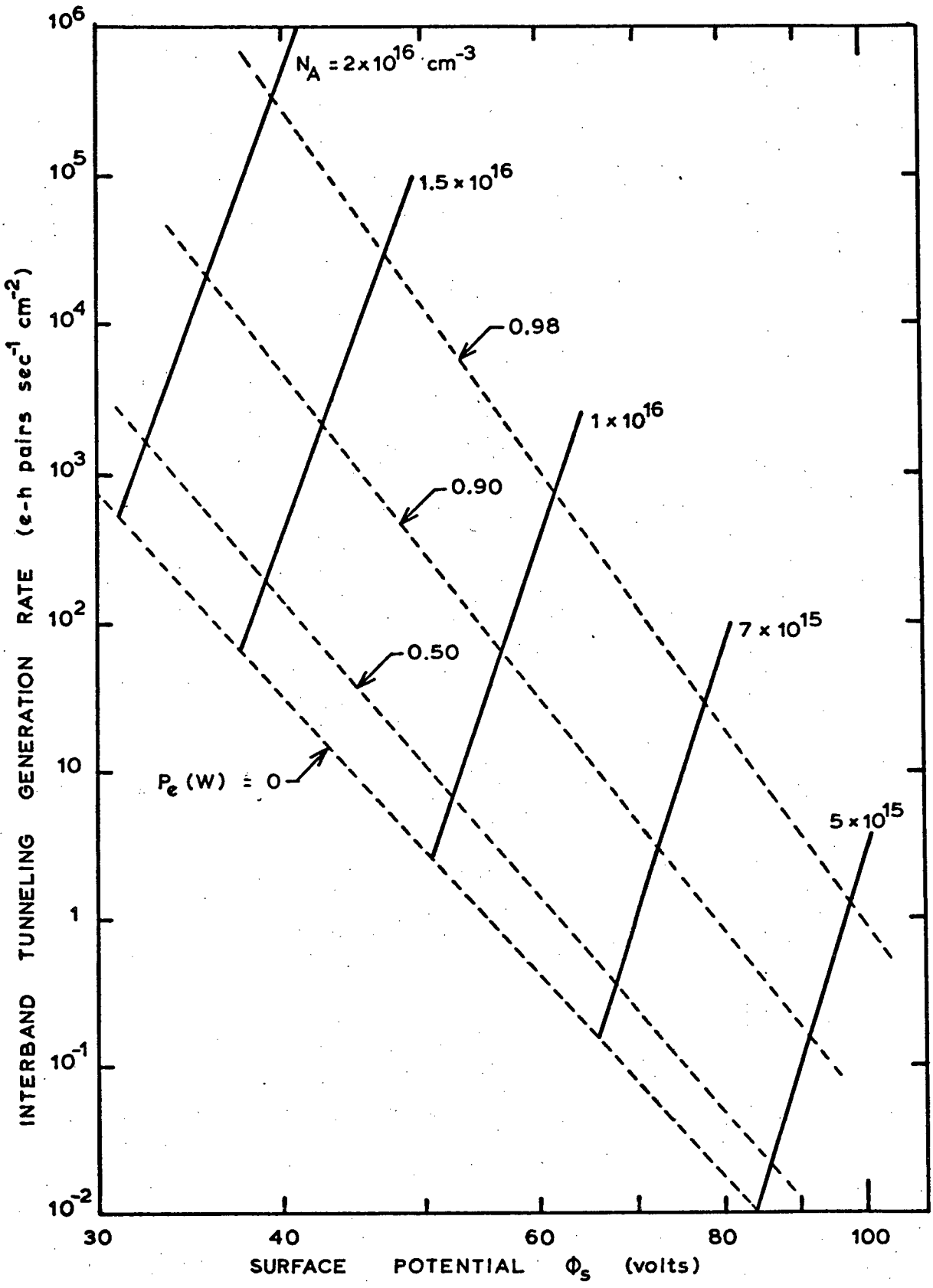


FIGURE 3.16 The interband tunnelling generation rate in MOS structures, plotted as a function of the silicon surface potential ϕ_s, T = 100K.

to maintain the peak electric field below this value and yet still enable operation in the plateau region of the avalanche initiation probability (above $P_e(w) = 0.9$). From Fig. 3.16 it can be seen that the operating surface potentials for such an MOS gate lie in the range 60-75 V. The actual gate operating voltages will of course be substantially higher than this due to the potential drop across the SiO_2 layer.

The interband tunneling discussed above is not the only tunneling dark current mechanism. As was pointed out in section 3.2.4, electrons may also tunnel from the valence to conduction band via mid gap states. Price [53] has derived an effective first order matrix element for the conventional field-emission process involving tunneling from defect or impurity levels within the band gap to the conduction band. The resulting probability per unit time that an electron, bound to a trap, is field ionized is found to have the same general form as that for interband tunneling transitions except that the energy gap is replaced by the ionization energy of the impurity, namely the height of the barrier through which the electron must tunnel (i.e., $E_c - E_T$). Using Price's matrix element for tunneling, Sah [54] determines the transition probability per unit time to be

$$e_{nt} = A_T \left(\frac{\mathcal{E}}{E_C - E_T}\right) M_T^2 \exp\left[\frac{-B_C}{\mathcal{E}_T} (E_C - E_T - \Delta E_T)^{3/2}\right] \qquad (3.57)$$

where A_T is a constant and M_T is the matrix element for tunneling from trap states. ΔE_T is the barrier lowering for coulomb attractive centers (equal to the Poole-Frenkel barrier lowering) and $\mathcal{E}_T$ is an effective field given by

$$\mathcal{E}_T = \mathcal{E}\left(\frac{E_T - \Delta E_T}{E_T}\right) \qquad (3.58)$$

The tunneling emission rate from coulomb attractive centers is designated by e_{nt}^+. For neutral centers $E_T = 0$ and $\mathcal{E}_T = \mathcal{E}$, and the designation is e_{nt}^o.

By analogy with the case of interband tunneling the effective mass in B_c may be regarded as a reduced effective mass for tunneling, the dominant component of which will be the electron mass in the conduction band.

For tunneling from the valence band into an empty trap or interface state, a similar expression to (3.57) is expected, except that the height of the energy barrier, through which the electron tunnels, is now the difference between the band gap and the ionization energy of the impurity level (i.e., $E_g - E_c + E_T = E_T - E_v$). The dominant component to the effective mass in this case is the hole mass in the valence band. Thus,

$$e_{pt} = A_T \left(\frac{\mathcal{E}}{E_T - E_V}\right) M_T^2 \exp\left[\frac{-B_V}{\mathcal{E}_T} \left(E_T - E_V - \Delta E_T\right)^{3/2}\right] \qquad (3.59)$$

The transition rates for coulomb attractive (to holes) and neutral centers are designated by e_{pt}^- and e_{pt}^o respectively.

The only experimental work on tunneling into and out of deep levels has been directed at explaining the excess current in forward biased Esaki diodes. Sah [54] has made a detailed study of the impurity induced excess current in gold doped silicon Esaki junctions. The observed excess current was correlated quantitatively with various two-step combinations of Shockley-Read-Hall and tunneling transitions using the theoretical tunneling rate calculation due to Price. The estimated value of $A_T M_T^2$ for the tunneling transitions valence band to trap to conduction band (mechanisms 2 in Fig. 3.14) obtained from Sah's data is given in table 3.1. Sah only reports experimental data for one value of field, therefore, no experimental value for B can be obtained from his results. The value of B used by Sah was calculated according to (3.54) using the transverse electron and light hole effective masses. These masses and the corresponding values for B are listed in table 3.1. Chynoweth et al. [55] have measured the field dependence of excess current in silicon Esaki diodes by varying the n^+ dopant con-

centration. Their estimate of the effective mass to be used in (3.54) is listed in table 3.1 along with the value for $A_T M_T^2$ required to obtain agreement with Sah's data.

The steady state dark generation rate per unit area due to tunneling through deep bulk levels is

$$G_{Bt} = \int_0^W g_{Bt}(x)\, dx \tag{3.60}$$

where $g_{Bt}(x)$ for donor levels is

$$g_{Bt}^{+}(x) = \frac{e_{pt}^{o}(x)\, e_{nt}^{+}(x)}{e_{pt}^{o}(x) + e_{nt}^{+}(x)}\, N_T \tag{3.61}$$

and for acceptor levels is

$$g_{Bt}^{-}(x) = \frac{e_{pt}^{-}(x)\, e_{nt}^{o}(x)}{e_{pt}^{-}(x) + e_{nt}^{o}(x)}\, N_T \tag{3.62}$$

TABLE 3.1 Experimental values for the effective mass and matrix element for tunneling through trap states.

	$A_T M_T^2$ ($eVmV^{-1}sec^{-1}$)	m_c^*/m_o	m_v^*/m_o	B_c [b] ($Vm^{-1}eV^{-3/2}$)	B_v [b] ($Vm^{-1}eV^{-3/2}$)
Sah	2.7×10^3	0.19	0.16	2.98×10^9	2.73×10^9
Chynoweth	6.5×10^3 [a]	0.3		3.74×10^9	

[a] Value required to fit Sah's data using 0.3 for m^*/m_o

[b] Calculated according to Eq. (3.54)

The generation rate is highest for those levels for which the transition probabilities e_{nt} and e_{pt} are equal. At field strengths where the Poole-Frenkel barrier lowering is small, these levels will lie close to mid band. The density of such deep level traps will depend on the quality of the starting silicon wafer and on the bulk gettering steps included in the processing. Typical deep level impurity/defect densities after CCD processing lie in the range 10^{10} - $10^{12}cm^{-3}$. Figure 3.17 shows the calculated generation rate for $N_T = 10^{10}cm^{-3}$ versus peak electric field in MOS gates with different substrate doping levels. The generation rate per unit area was calculated from (3.60) using the data in Table 3.1. The extrapolated generation rates are very uncertain for these field strengths, as indicated by the large difference in the rates predicted from Sah's data and those predicted using the effective mass estimate of Chynoweth et al. It is apparent, however, that tunneling through deep traps will be one of the dominant triggering mechanisms in the dark. Precisely how important this mechanism turns out to be will depend very much on the density and energy level of the deep impurity/defect centers present in the high field region near the interface and on the correct value for the tunneling effective mass.

3.2.6 Dark Generation of Triggering Carriers in the Non-Steady State

So far only steady state dark generation mechanisms have been considered. Immediately following the large depleting pulse above breakdown, however, there is a transient during which the interface states and bulk traps are relaxing to their new steady state occupancy. During this transient there is an enhanced emission of either electrons or holes depending on the previous occupancy of the traps, prior to the depleting pulse. There is also a transient following each avalanche discharge. During the discharge the occupancy of the bulk traps and interface states may be altered by either

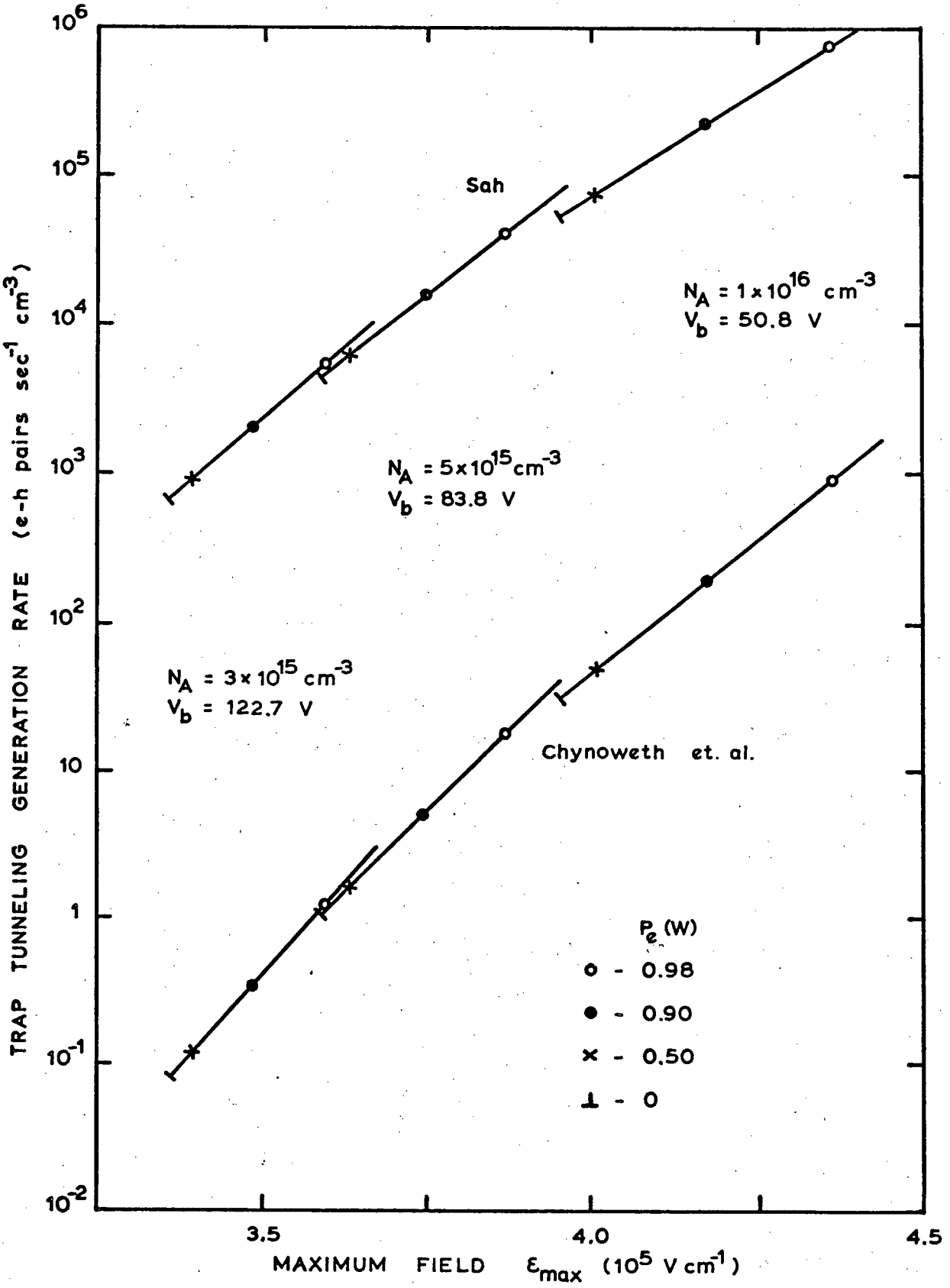


FIGURE 3.17 Tunnelling generation rate through traps as a function of the peak electric field in n^+p step junctions or p-substrate MOS gates, T = 100K. The results are extrapolated from the data given by Sah [54] and Chynoweth et.al. [55].

direct capture of the mobile electrons and holes created by the avalanche, or by impact ionization. For short reset times, complete relaxation of the trap occupancy following an avalanche discharge may not be possible before the next pulse above breakdown occurs.

Consider first the emission of holes from interface states. This is shown in Fig. 3.18 for two different reset conditions corresponding to surface accumulation and surface inversion. When the surface is in accumulation during reset the interface states empty down to the Fermi level by capturing holes. After the application of the depleting pulse above breakdown, the density of both electrons and holes is very low at the interface. Electron capture is, therefore, negligible and holes are emitted very rapidly as the levels nearest the valence band fill with electrons. These holes have avalanche initiation probability $P_h(0)$ and are immediately swept out through the depletion layer. Unless $P_h(0)$ is extremely small, an avalanche will be triggered within a very short time after the application of the depleting pulse.

In order to achieve a low dark count rate, the silicon surface must be inverted during reset. In this case the interface states fill up to the Fermi level by capturing electrons from the inversion layer. In strong inversion this occurs very rapidly due to the high surface concentration of electrons, and hence the occupancy of the levels immediately prior to the pulse above breakdown will be insensitive to any previous avalanche discharges. After the depleting pulse the inversion layer still remains. Electron capture and emission, therefore, continue to determine the occupancy f_T, forcing all the interface states below E_{Fn} to be full of electrons. The tunneling distance for hole emission is, thus, large, and at low temperatures hole emission will be a very rare event.

From the above discussion it is apparent that a PC-CCD will require a

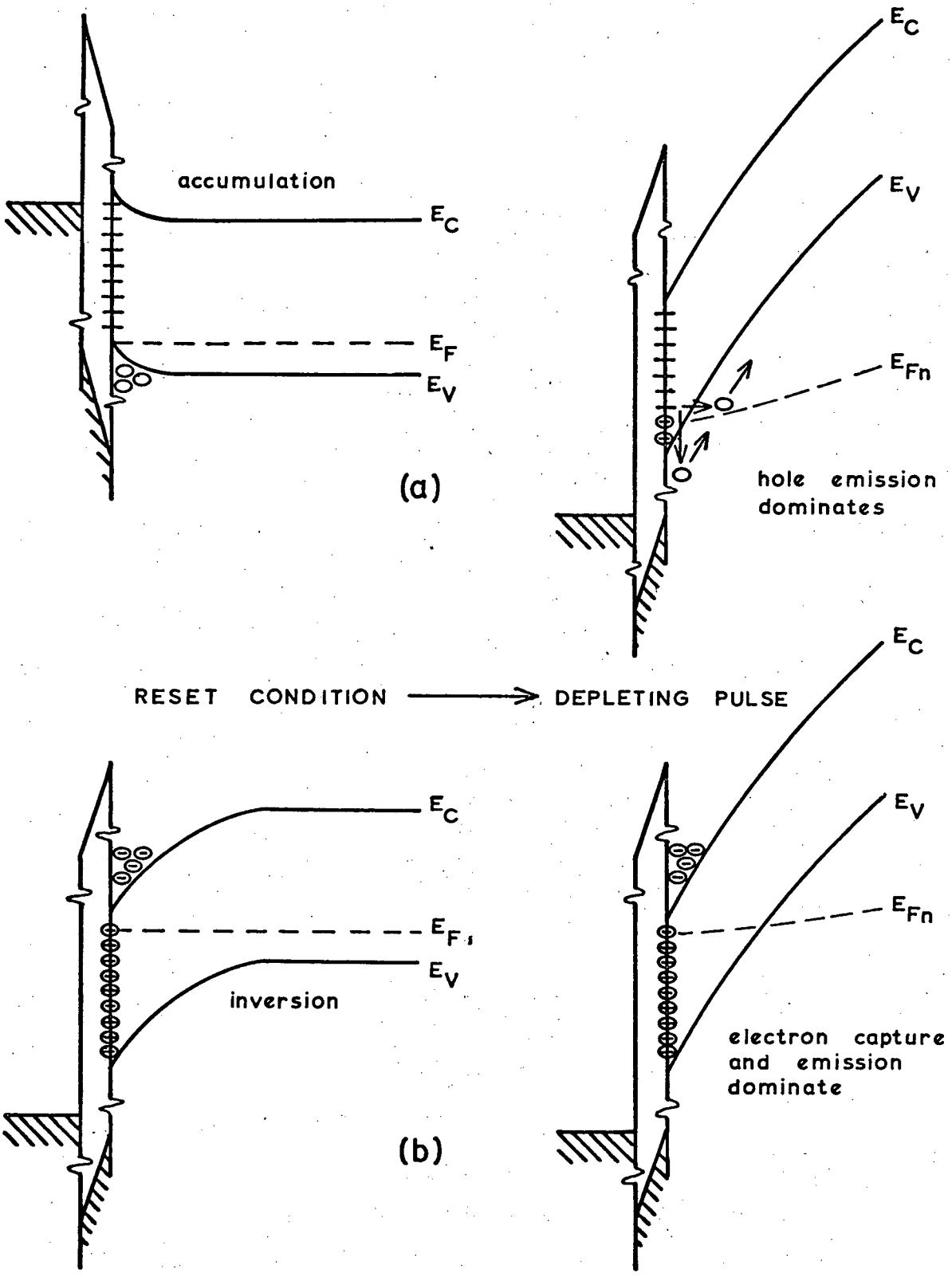


FIGURE 3.18 Energy band diagram for a p-substrate MOS gate illustrating the generation of triggering carriers at the interface, immediately following a depleting pulse.

(a) pulsed from accumulation during reset
(b) pulsed from inversion during reset.

circulating background charge or "fat zero" in order to suppress the dark generation of triggering carriers from interface states. In practice, however, the effectiveness of a circulating background charge may be limited by edge effects. At the boundaries of the breakdown area there is a transition region where the surface is not in inversion before or after the depleting pulse. The dark generation from interface states in this region will be dominated by the steady state emission of electrons and holes, which may become significant at the high surface fields encountered above breakdown. In addition, hot electrons incident on the interface during an avalanche discharge may impact ionize some of the interface states resulting in an increased hole emission rate during the transient following the discharge.

The non-steady state behaviour of the bulk trapping centers is more complex. Figure 3.19 shows the band diagrams for a typical PC-CCD gate during reset and after the depleting pulse above breakdown. After the depleting pulse, and in the absence of an avalanche discharge, the centers may only change their charge state by the emission of electrons or holes. In this case the rate of change of trapped charge can be expressed as

$$N_T \frac{d}{dt} f_T(x,t) = e_p^*(x)\, N_T - [e_p^*(x) + e_n^*(x)]\, N_T f_T(x,t) \qquad (3.63)$$

where e_n^* and e_p^* are defined by equations (3.49) to (3.52). The return to equilibrium is, therefore, exponential with a time constant $\tau(x) = [e_n^*(x) + e_p^*(x)]^{-1}$. The solution for $f_T(x,t)$ may be expressed as

$$f_T(x,t) = f_T(x,\infty) - [f_T(x,\infty) - f_T(x,0)] \exp -t/\tau(x) \qquad (3.64)$$

where $f_T(x,\infty)$ is the steady state occupancy given by

$$f_T(x,\infty) = \frac{e_p^*(x)}{e_n^*(x) + e_p^*(x)}$$

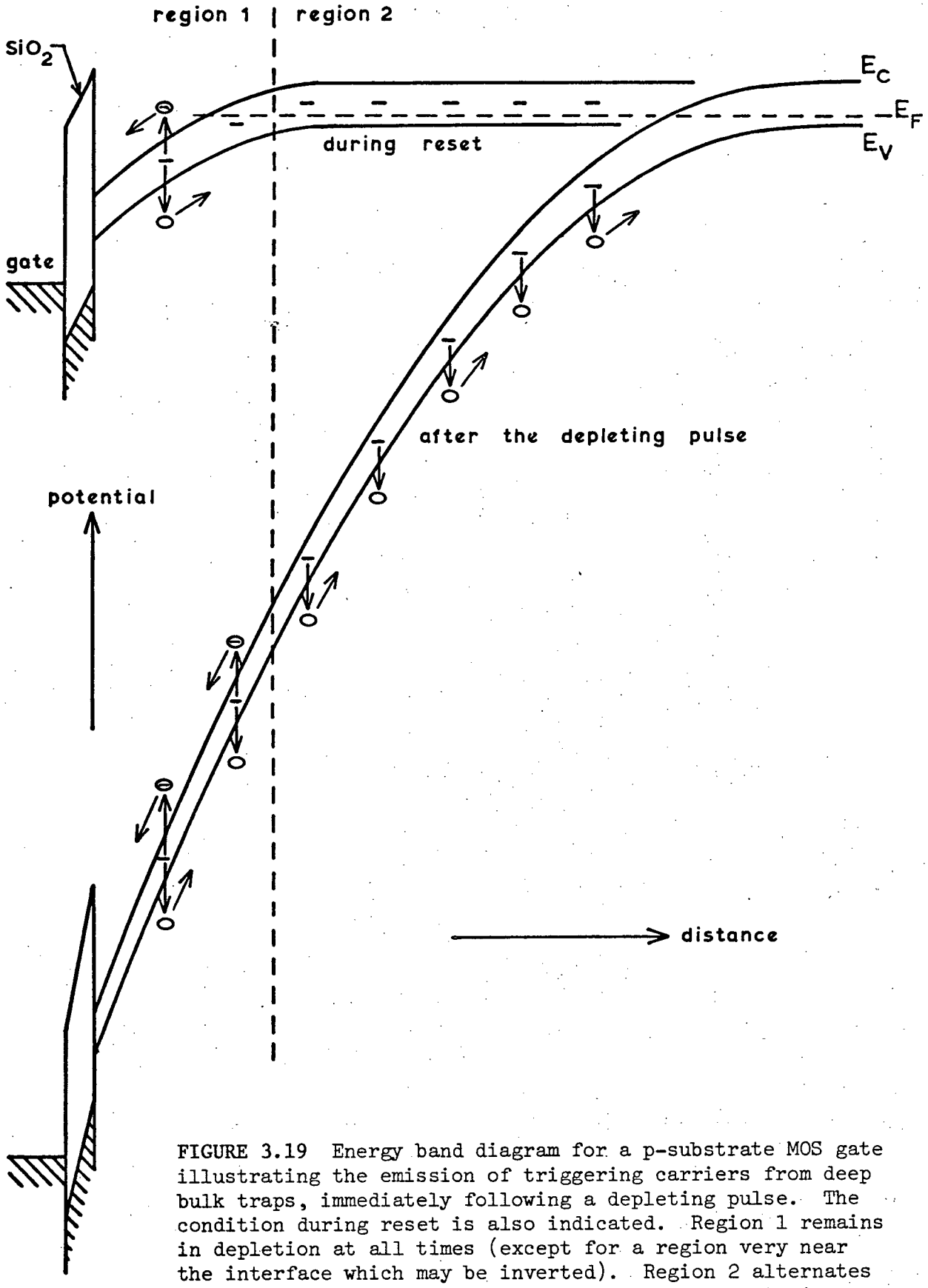


FIGURE 3.19 Energy band diagram for a p-substrate MOS gate illustrating the emission of triggering carriers from deep bulk traps, immediately following a depleting pulse. The condition during reset is also indicated. Region 1 remains in depletion at all times (except for a region very near the interface which may be inverted). Region 2 alternates between the equilibrium neutral condition and depletion.

and $f_T(x,0)$ is the initial trap occupancy at the instant the gate is pulsed. At time t after the depleting pulse above breakdown, the mean dark event rate due to bulk traps (i.e., the probability per unit time that an avalanche discharge will be triggered) is

$$R(t) = AN_T \int_W e_n^*(x)\, f_T(x,t)\, P_e(x) + e_p^*(x)[1 - f_T(x,t)]P_h(x)\, dx \qquad (3.65)$$

where $P_e(x)$ and $P_h(x)$ are the electron and hole avalanche initiation probabilities, and A is the area. The difference between the initial rate at t=0 and the rate that would be obtained with the steady state trap occupancy, $f_T(x,\infty)$, is

$$R(\infty) - R(0) = AN_T \int_W [f_T(x,\infty) - f_T(x,0)] \cdot [e_n^*(x)P_e(x) - e_p^*(x)P_h(x)]\, dx \qquad (3.66)$$

Equation (3.66) shows that a disturbance of the trap occupancy can cause either an increase or decrease in the breakdown rate, depending on the sign of the integrand. In order to interpret this result physically, the integral is divided into two parts. The first part is over those regions that remain in depletion during reset (region 1 in Fig. 3.19) and the second part of the integral is over those regions that alternate between depletion and the equilibrium neutral condition (region 2 in Fig. 3.19). During reset the traps within region 2 empty down to the equilibrium fermi level by capturing holes. This occurs very rapidly due to the high hole concentration in the p-type neutral bulk. Therefore, the trap occupancy in this region prior to each depleting pulse will be independent of any changes that may have occured during previous cycles. Following the depleting pulse, electrons and holes are emitted alternately. Initially holes are emitted at a higher rate than electrons ($f_T(x,\infty) > f_T(x,0)$) as the trap

occupancy adjusts to the new steady state in depletion. In general, this increased hole emission will result in a reduction of the breakdown rate since $P_h(x)$ is much smaller than $P_e(x)$ in region 2 (see Fig. 3.7). Only if the traps are close to the valence band but above the equilibrium Fermi level during reset (i.e., for which $e_p^*(x) >> e_n^*(x)$ and $f_T(x,0) \approx 0$), can the breakdown rate be increased.

Except during periods of avalanche discharge, region 1 remains depleted at all times. It is only during the transient following each avalanche, therefore, that the steady state occupancy is disturbed. Under avalanche breakdown conditions, free carrier concentrations of the order of 10^{14} cm^{-3} to 10^{16} cm^{-3} are reached. Consequently, electron and hole capture by traps becomes possible or, similarly, the hot electrons and holes may impact ionize the traps. In either case an exponential transition between the initial and final states would be expected, which may be described by an equation of the same form as (3.64). If the avalanche were to continue for time t_A then the fraction of centers in the more negative state, upon termination of the avalanche, would be

$$f_{AT}(x,t_A) = f_{AT}(x,\infty) - [f_{AT}(x,\infty) - f_T^*(x)]\exp -t_A/\tau_A(x) \qquad (3.67)$$

where:

$f_{AT}(x,\infty)$ = limiting trap occupancy during an extended avalanche.

$f_T^*(x)$ = occupancy of the traps immediately prior to the avalanche.

$\tau_A(x)$ = time constant for the transient.

For the case that the avalanche discharges are well separated in time, the initial occupancy is equal to the steady state occupancy in depletion (i.e., $f_T^*(x) = f_T(x,\infty)$).

Following a discharge, the trap relaxation in region 1 is given by

$$f_T(x,t) = f_T(x,\infty) - [f_T(x,\infty) - f_{AT}(x,t_A)]\exp -t/\tau(x) \qquad (3.68)$$

where t=0 is now taken to be the termination of the avalanche. Making this time transformation in (3.66) and using (3.68) gives

$$R_1(\infty) - R_1(t_s) = AN_T \int_{w_1} [f_T(x,\infty) - f_{AT}(x,t_A)]\exp -t_s/\tau(x) \cdot[e_n^*(x)P_e(x) - e_p^*(x)P_h(x)]\, dx \qquad (3.69)$$

where t_s is the time from the termination of the avalanche to the next pulse above breakdown (the integration here is over region 1 only). As before, there can be either an increase or decrease in the breakdown rate, however, the magnitude of the change now decays exponentially with increasing reset duration.

The field distribution, the variation of the ionization coefficients $\alpha_e(x)$ and $\alpha_h(x)$, and the distribution of free carriers within the space charge layer during breakdown are shown schematically in Figures 3.20(a)-(c). The boundary between regions 1 and 2 is also indicated. If charge trapping is the dominant process during breakdown then from Fig. 3.20(c) it can be seen that, subsequent to the discharge, there will be an increase in the emission of electrons from traps located near the $Si-SiO_2$ interface, and an increased hole emission rate for the remainder of the traps in region 1. This is shown schematically in Fig. 3.21(b). Figure 3.21(a) indicates the variation of $P_e(x)$ and $P_h(x)$ for reference. For deep traps ($e_n^*(x) \sim e_p^*(x)$) the overall effect of the increased electron or hole emission shown in Fig. 3.21(b) is a slight reduction in the triggering rate $R_1(t_s)$. If any charge is trapped by shallow levels during the avalanche discharge, the subsequent triggering rate may be increased. The shallowest levels, however, will relax to their steady state occupancy before the next pulse above breakdown occurs (i.e., $(e_n^* + e_p^*)^{-1} < t_s$).

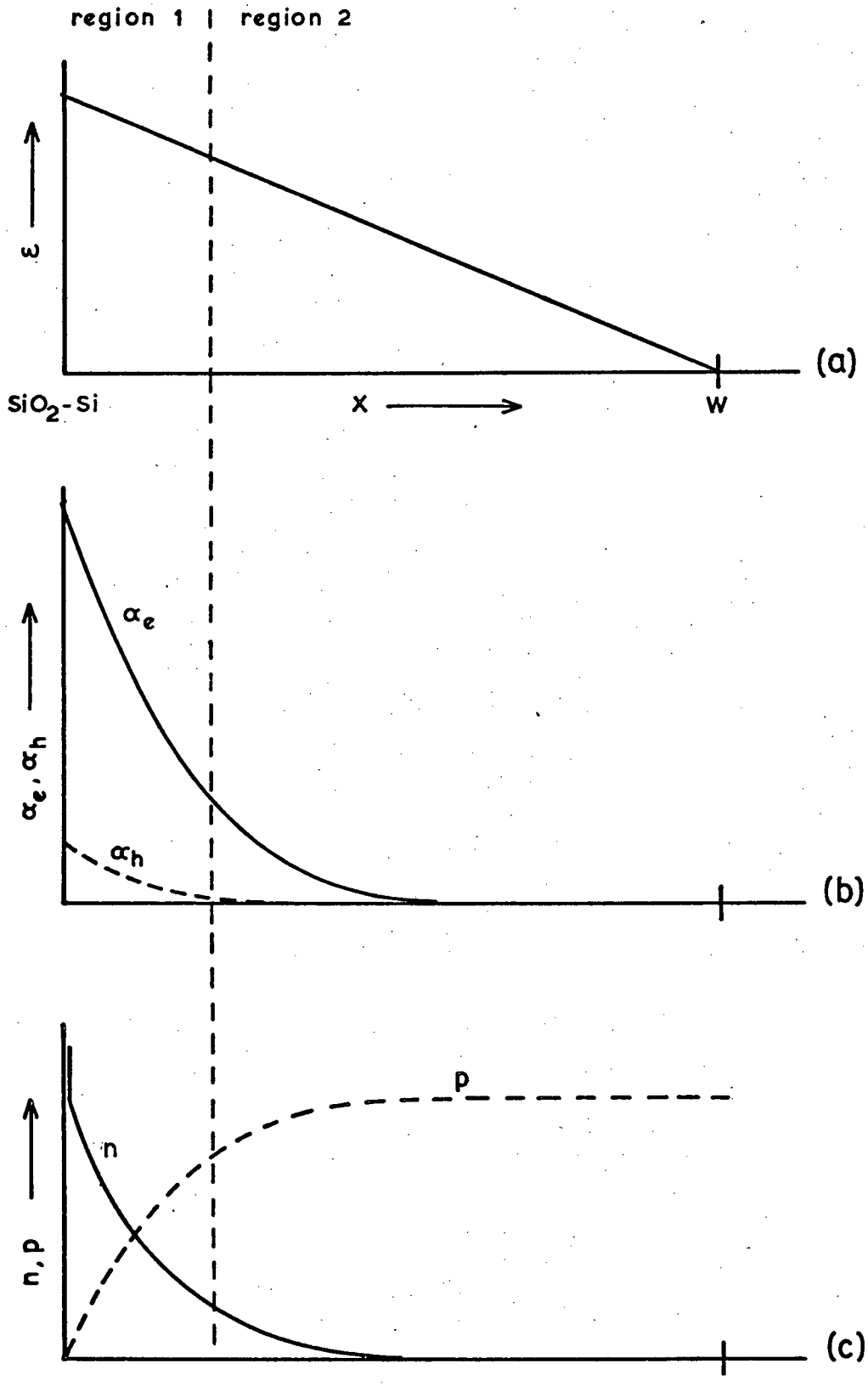
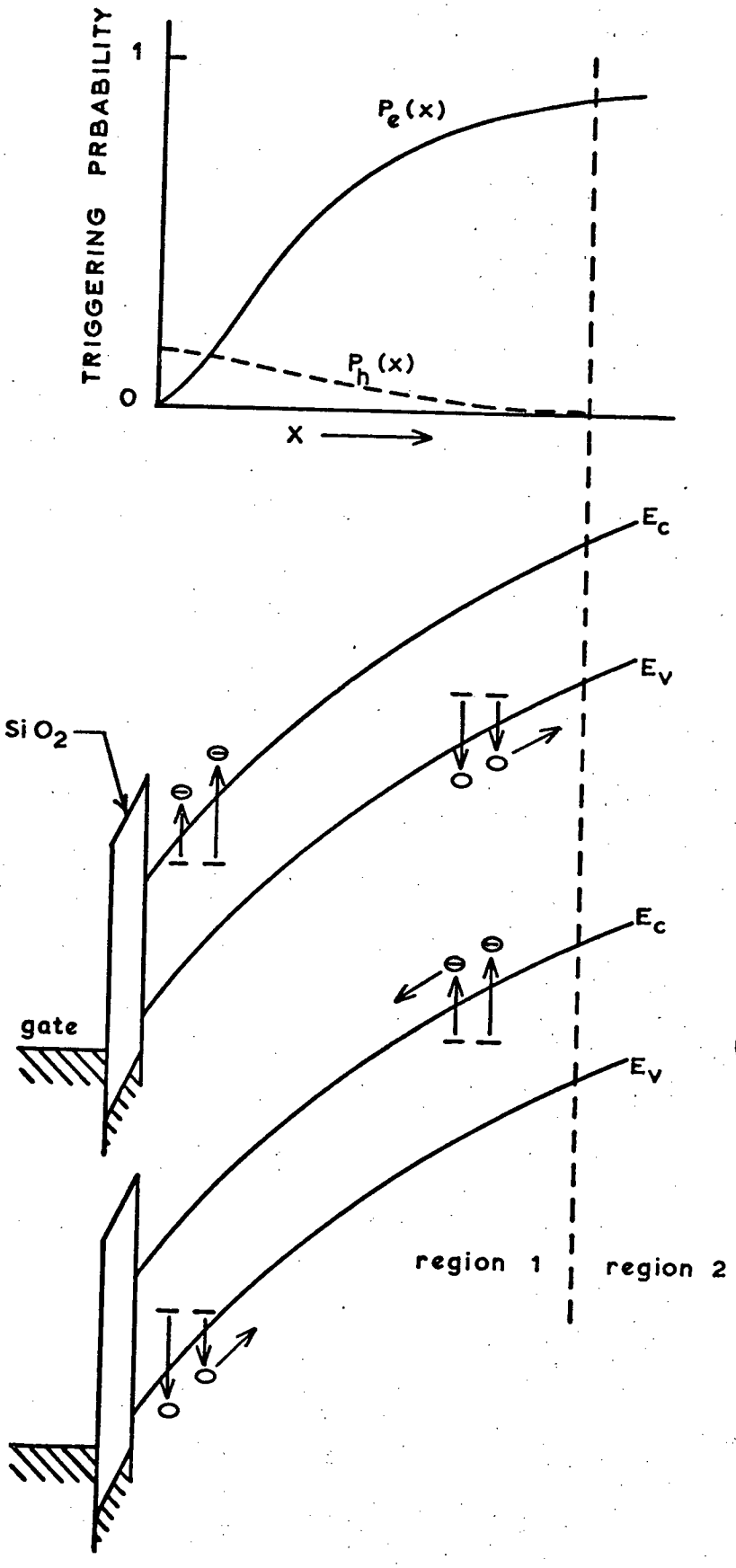


FIGURE 3.20 Model for the depletion region of a p-substrate MOS gate during breakdown. (a) is the field distribution, (b) is the variation of the electron and hole ionization coefficients and (c) is the density of hot electrons and holes.



(a) triggering probability as a function of position in the depletion region.

(b) carrier capture dominant during breakdown.

(c) impact ionization dominant during breakdown.

FIGURE 3.21 The increased emission of electrons and holes by bulk traps subsequent to an avalanche discharge.

The change in occupancy that occurs if impact ionization of the traps is dominant during breakdown is more difficult to determine. In this case there are two competing processes: 1) hot electrons or holes can knock off electrons from the traps, leading to a positive charge change or 2) hot carriers can knock off a hole, leading to a negative charge change. Figure 3.21(c) illustrates th worst possible situation following breakdown i.e., for the case that the traps are charged in the opposite sense to the free carrier density during the avalanche. As before, the only levels that will be actively emitting carriers at the beginning of the next cycle are those for which $(e_n^* + e_p^*)^{-1} > t_s$.

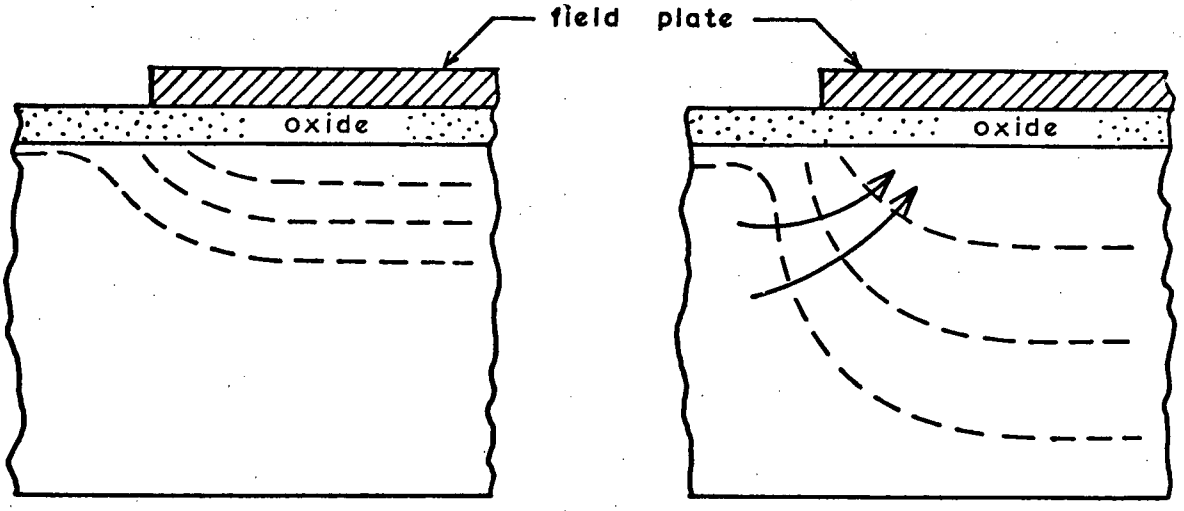
It is clear that the overall effect of an avalanche discharge will depend very much on the relative sizes (as a function of position in the depletion region) of the four impact ionization cross-sections and two capture cross-sections, and on the distribution of the hot carriers during breakdown. The magnitude of the effect will also depend on the duration of the avalanche, or more precisely, on the amount of charge transiting the depletion region during an avalanche discharge. Without this detailed information all that can be said is that the resulting change in occupancy is not expected to lead to a large decrease in the triggering rate at the start of the next pulse. How large the increase will be, if any, will depend on the reset duration, t_s.

Haitz [36], by varying the shunt capacitance in the high impedance reset circuit of his small area n^+p guard ring diodes, has measured the dark count rate as a function of the total charge crossing the junction during an avalanche. Initially the count rate increased linearly with increasing shunt capacitance, however, some saturation of the count rate was evident for the highest charge levels reported, indicating that the trap occupancy had

approached the steady state value during avalanche. By extrapolating the linear count rate versus shunt capacitance relation to zero capacitance, Haitz was able to determine the steady state dark generation rate for his diodes. The maximum count rates measured were a factor of 20 larger than this and corresponded to a charge of 1×10^{15} carriers cm^{-2} crossing the junction during the avalanche discharges. This result is encouraging in view of the fact that the size of the discharge pulse for an MOS gate is typically three orders of magnitude lower than this.

3.3 PREVENTION OF PREMATURE EDGE BREAKDOWN AND LOCALIZED MICROPLASMA BREAKDOWN

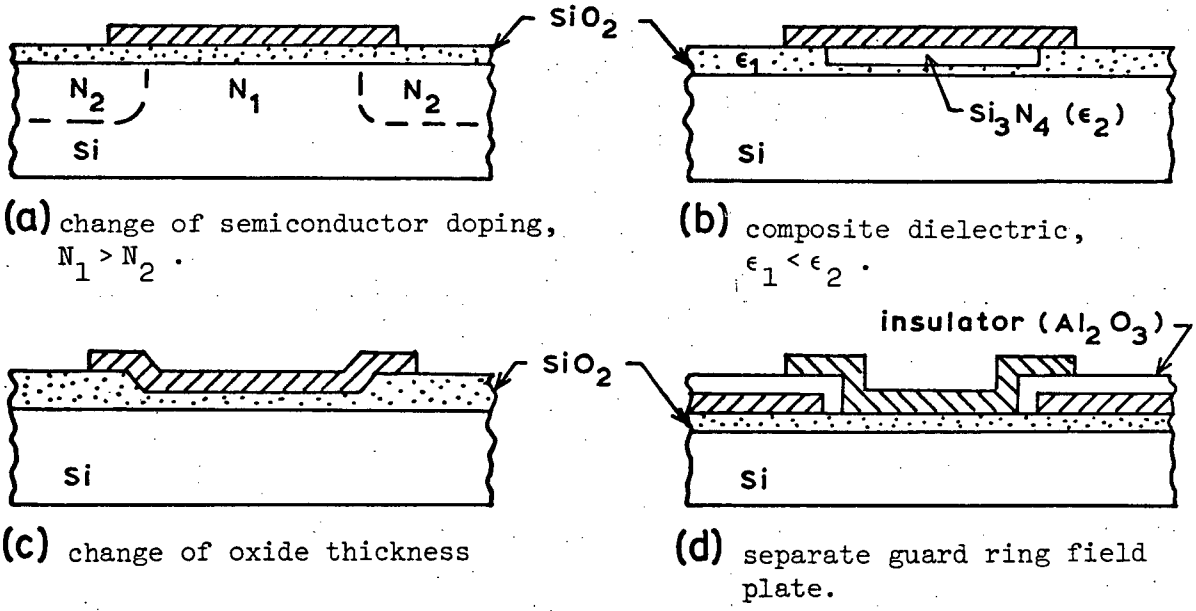
Goetzberger and Nicollian [56] have studied the edge breakdown effect for silicon dioxide/silicon MOS gates pulsed into deep depletion. With an oxide thickness of 1000Å they found that edge breakdown was unimportant for substrate doping levels greater than approximately $5 \times 10^{16} cm^{-3}$. Below this doping level the electric field at the oxide-silicon interface is enhanced at the edges of the metal gate causing a reduction in the breakdown voltage. Figure 3.22 illustrates this effect and shows the difference between high and low semiconductor doping. According to this explanation the edge breakdown effect should depend on the ratio of depletion layer width to oxide thickness. Rusu and Bulucea [57] have made computer-aided calculations of the breakdown voltage of deeply depleted MOS capacitors using the ionization integral method. The potentials were calculated from the two-dimensional Poisson equation using finite differences and successive over-relaxation. Lee's room temperature ionization rate data (Appendix A) and the constant K approximation ($\alpha_h = k\alpha_e$) were used to calculate the ionization integral. In addition to useful design plots of breakdown voltage versus



(a) High doping density, narrow depletion region, uniform fields.

(b) Low doping density, wide depletion region, high fringing field.

FIGURE 3.22 Cross section of MOS capacitor showing equipotential lines in the space charge region at the edge of the field plate. Edge breakdown occurs in (b) due to the high fringing field.



(a) change of semiconductor doping, $N_1 > N_2$.

(b) composite dielectric, $\epsilon_1 < \epsilon_2$.

(c) change of oxide thickness

(d) separate guard ring field plate.

FIGURE 3.23 Four different MOS structures to avoid premature avalanche breakdown at the outer edge of the field plate in high resistivity samples.

oxide thickness and substrate doping, Rusu and Bulucea obtain a universal criterion for field uniformity in terms of the ratio of oxide thickness to the maximum (breakdown) width of the silicon depletion region. According to their calculations this ratio should be larger than 0.3 in order not to have field concentration around the edges of the metal plate.

It was established in section 3.2.5 that the substrate doping of a PC-CCD should be less than approximately $8 \times 10^{15} cm^{-3}$ in order to avoid interband tunneling. On such lightly doped substrates the oxide thickness required to satisfy the above criterion for field uniformity becomes rather large, and it may prove better to use some sort of guard ring in order to prevent edge breakdown. Some of the possible guard ring structures that can be used on MOS devices are shown in Fig. 3.23 [56]. In each case the action of the guard ring is to cause a more gradual transition from the deeply depleted breakdown region to the undepleted (or slightly depleted) surrounding area, thereby eliminating the high fringing fields. In a PC-CCD, all but structure (a) in Fig. 3.23 can be used both to define the transfer channels and to prevent edge breakdown. High fringing fields in the charge transfer direction are avoided by adjusting the potential barriers between pixels to be only slightly below breakdown (see Fig. 3.4).

In addition to premature edge breakdown, it is also possible for there to be local preferred sites for breakdown (so-called microplasmas) due to the presence of crystal defects. These crystal imperfections may cause a reduction in the threshold energy required for ionization [58], or they may become decorated with impurities leading to a local enhancement of the electric field [59] [60]. Consequently, the breakdown voltage in the vicinity of crystal defects is lower than in the surrounding area. In the early work on avalanche diodes such microplasmas were found to dominate the breakdown characteristics, and they have been the subject of many investigations.

Subsequent progress in crystal growing and device fabrication technology, however, has led to the routine fabrication of microplasma-free junctions that breakdown uniformly over the entire junction area.

The fabrication of uniform breakdown MOS gates free from microplasmas may be more difficult. It is well known that high temperature thermal oxidation of dislocation-free silicon wafers frequently results in the formation of Frank-type stacking fault defects at the Si-SiO_2 interface [61]. These defects, especially if they become decorated with impurities, greatly increase the leakage currents in MOS and CCD structures [62] [63] and may also result in localized breakdown. It has been found that the origin of oxidation induced stacking faults (OSF's) is either residual saw and lapping damage or grown-in micro defects formed during crystal growth. Chemically etching the wafer before oxidation eliminates any mechanical damage but the grown-in defects are more difficult to eliminate. A number of procedures have been proposed to suppress these nucleation centers and to shrink existing stacking faults. Preoxidation gettering involving a phosphorous or silicon nitride deposition on the back side of the wafer [64] [65] is a well-known technique for the suppression of SF nuclei. It has also been found that adding small amounts of chlorine in the form of HCl or C_2HCl_3 to the oxidation atmosphere can suppress and shrink OSF's, and can eliminate grown-in defects [62] - [69]. More recently it has been shown that the nucleation of OSF's on the front side of the wafer can be greatly suppressed by the gettering action of mechanical damage on the back [70] [71], or by the gettering action of thermally induced microdefects in the inner part of oxygen rich Czochralski wafers (so-called intrinsic gettering) [72] [73].

By using a combination of the above techniques, particularly HCl oxidation and back-surface damage, large CCD arrays have been fabricated that are free from any dark current anomalies. The same techniques should, there-

fore, make possible the fabrication of large microplasma-free PC-CCD's

3.4 OPTICAL COUPLING BETWEEN IMAGE ELEMENTS

Light emission during avalanche breakdown is a well-known phenomenon and poses an important problem in the proposed PC-CCD imager. Even if only a few photons are emitted during the avalanche discharge of an image element, these photons would have a high probability of being re-absorbed and triggering the breakdown of adjacent pixels. These pixels would then also emit photons, thereby starting a chain reaction. In order to prevent such a chain reaction and reduce the optical coupling between pixels to a nigligible level, the number of photons emitted per discharge, times the probability that they will trigger another pixel, must be made very much less than one.

Breakdown radiation was first reported in silicon junctions as far back as 1955 [74], however, the mechanism for this radiation has not been definitely established. Existing explanations include radiative (phonon-assisted) recombination of the free electrons and holes present during the breakdown [75], and radiative intraband transitions of the hot electrons and holes [75] [76] including radiation from the bremsstrahlung of hot carriers in the coulomb field of charged impurity centers [77]. For all the semiconductor materials studied the spectrum is very broad, extending to energies both considerably greater than and less than the energy gap. The spectrum for silicon is shown in Fig. 3.24 [75]. The total light output has been shown to increase linearly with breakdown current, however, there are very few reported estimates of the emission efficiency. Initial estimates for silicon place the efficiency at 10^{-8} visible photons for every electron crossing the junction [75]. It is the near-infrared photons with absorption coefficients less than $10^{2}cm^{-1}$, however, that will be responsible

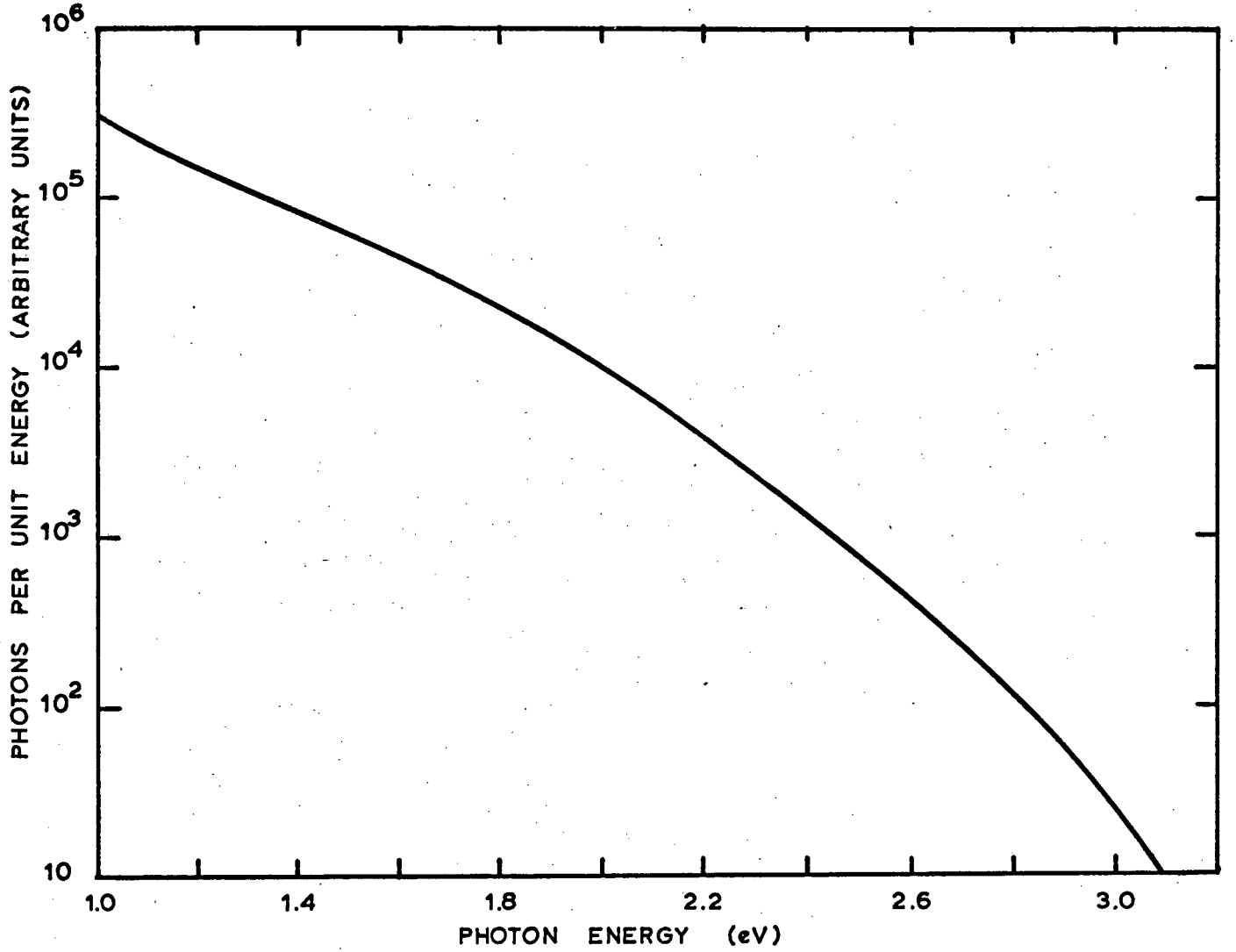


FIGURE 3.24 The emission spectrum during reverse bias avalanche breakdown in silicon, from [75].

for the majority of the photon coupling between pixels. As can be seen from Fig. 3.24, the emission efficiency for these photons is several orders of magnitude higher.

Photon energies less than the band gap energy are normally attributed to intraband transitions, although such photons could also result from the recombination of hot carriers through deep traps. The spectrum in Fig. 3.24 was obtained from junctions containing many microplasma light spots. Such microplasmas imply regions of crystal damage and it is well known that deep levels are associated with such structural defects. It is possible, therefore, that defect-free junctions would have a smaller emission efficiency for infrared photons. Unfortunately, there are no reports as to whether the spectral distribution of the uniform glow from perfect junctions is any different from that reported earlier on avalanche diodes with microplasmas.

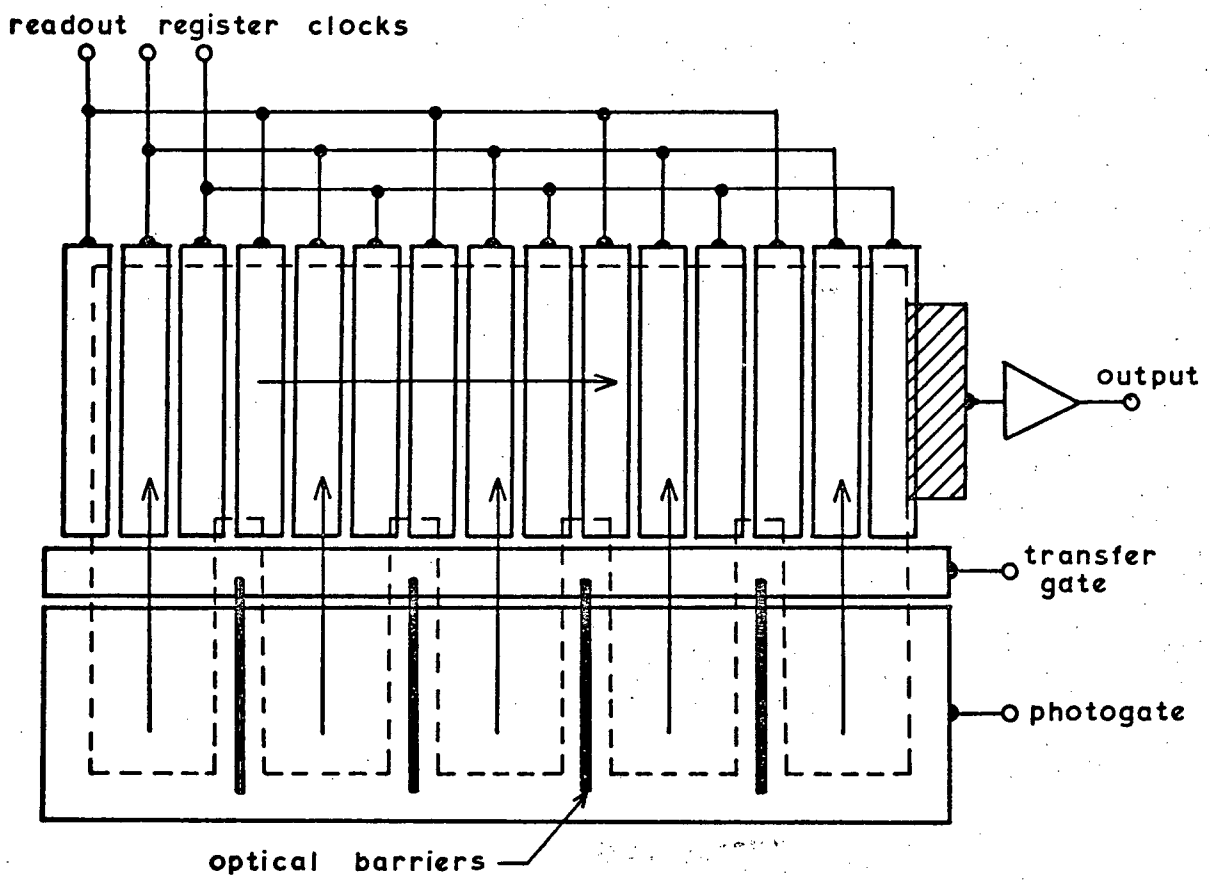
Haitz [78] has made a quantitative investigation of the photon coupling mechanism and has found that the pulse rate of a small area avalanche diode operating above breakdown (detector) is increased significantly by the reverse breakdown of another diode (emitter) on the same silicon wafer. The induced pulse rate was found to increase linearly with the breakdown current of the emitter, and to decrease with the square of the distance between emitting and detecting diodes. The range of diode separations measured was 0.06 - 0.6 cm. A square dependence over such distances implies that the interaction between the diodes was due to infrared radiation with an absorption coefficient in silicon of 1 cm^{-1} or less, corresponding to photon energies equal to or less than the band gap energy [79]. Such infrared photons can generate carriers either by phonon-assisted transitions from the valence band to the conduction band, or by generation from deep traps. The square dependence on distance also indicates that the coupling radiation

propogated directly from the emitter to the detector, and that total internal reflections at the top and bottom of the silicon slice were negligible. From an analysis of his photon coupling data, Haitz estimated the efficiency of coupling light generation to be 2×10^{-5} photons (with an energy equal to or less than the band gap) per electron crossing the junction.

The size of the avalanche discharge pulses in a PC-CCD exceeds 10^5 electrons even for very small area pixels $(25\ \mu m)^2$. It is, therefore, clear that some sort of optical barrier between pixels will be required in order to reduce the probability of an emitted photon triggering another discharge. How complete the optical isolation needs to be will depend on the size and center to center spacing of the pixels. It will also depend on the charge per unit area crossing the junction during an avalanche discharge and hence on the amount of overvoltage employed. Further studies on photon coupling will have to be made before the required degree of optical isolation can be determined.

Figure 3.25 illustrates how complete optical isolation might be achieved in a linear array organized for parallel transfer into a separate readout channel. The PC-CCD is fabricated on a 110 crystal plane, and an anisotropic etch [80] is used to cut deep vertical wall grooves between pixels. The grooves are etched to the required depth early on in the fabrication while the wafer is still full thickness. The groove walls can then be given the same oxidation and annealing steps as the gate oxide. The final step in the fabrication would be to fill in the grooves with an opaque material. A groove width of 3 to 4 microns and a depth to width ratio of 5 should be achievable by this method.

An alternate structure that may be fabricated on 100 silicon, uses deep V-grooves [81] anisotropically etched between the pixels, as shown in Fig. 3.26. As an automatic result of extending the gate across the V-grooves,



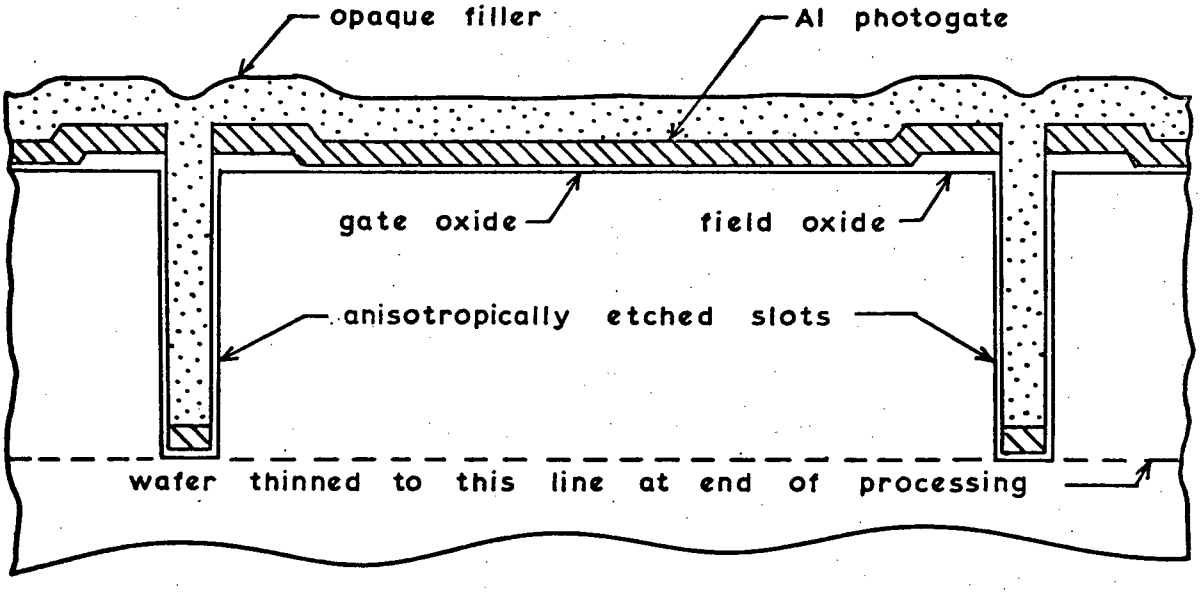
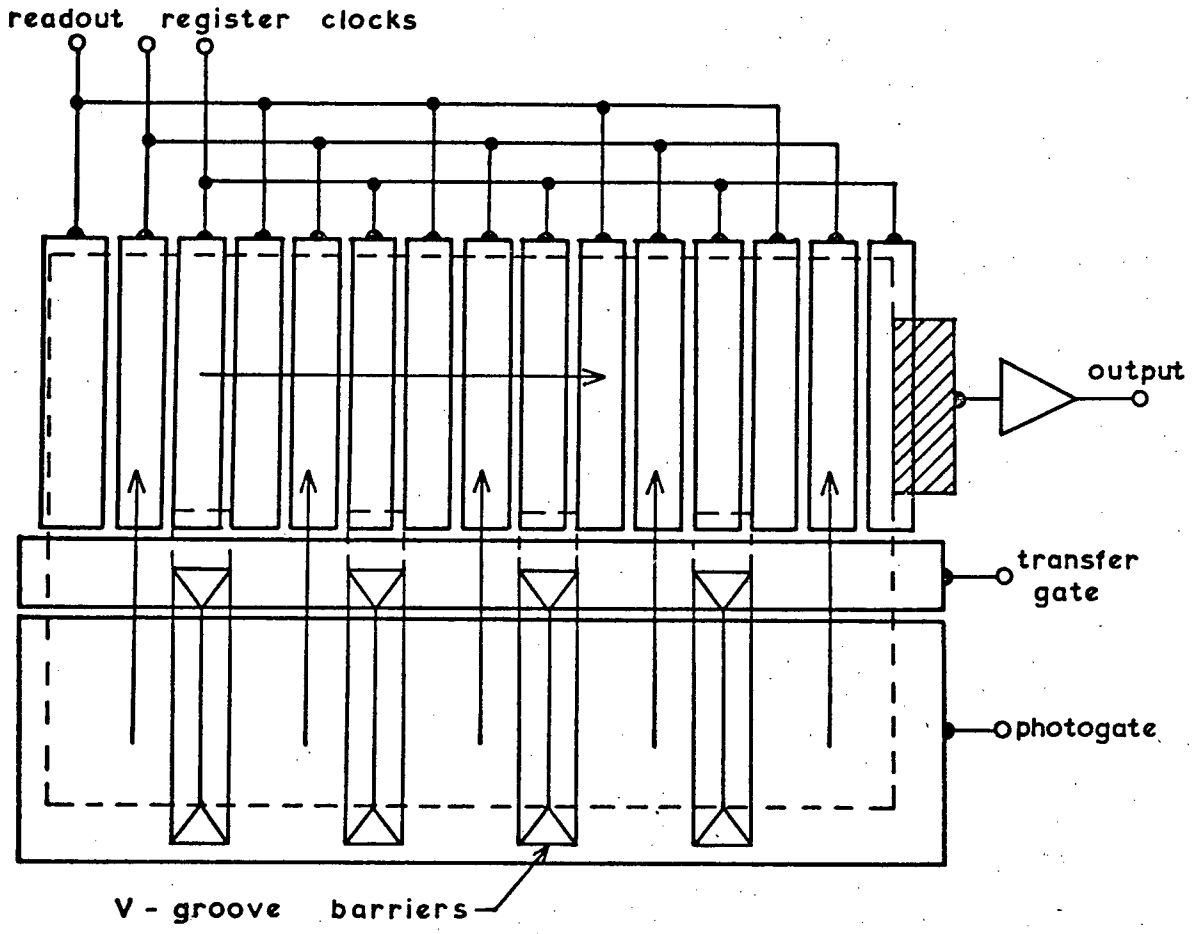


FIGURE 3.25 Illustration of how deep anisotropically etched slots may be used to optically isolate the individual pixels in a linear PC-CCD array fabricated on (110) silicon.



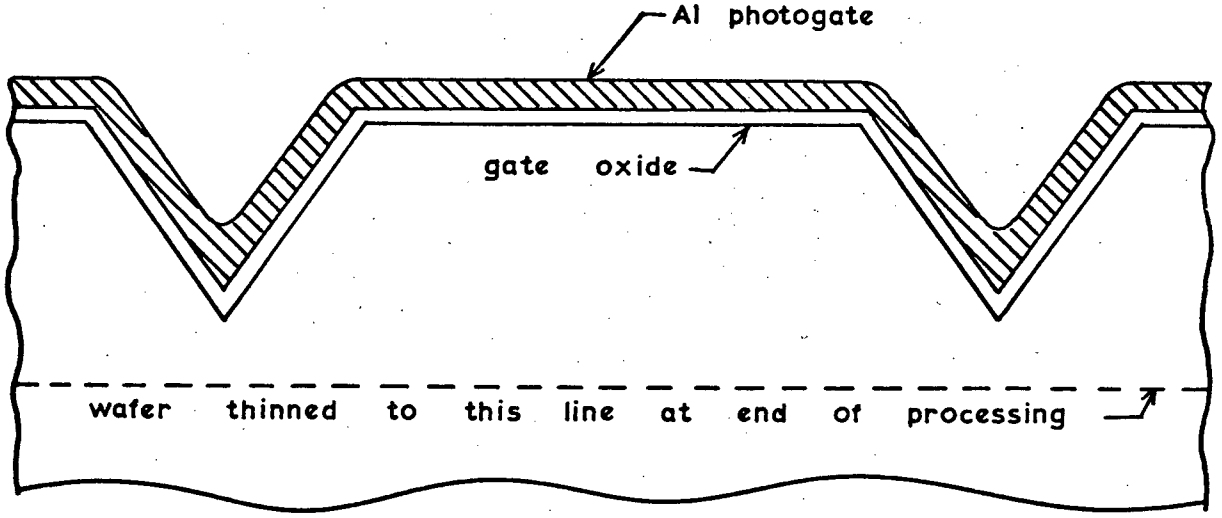


FIGURE 3.26 Illustration of how anisotropically etched v-grooves might be used to provide the required degree of optical isolation in linear arrays fabricated on (100) silicon.

potential barriers of small lateral dimension are generated at the apex of each groove [82]. These potential barriers serve to define individual pixels, however, the oxide must be thick enough to prevent field enhancement and localized breakdown near the groove apex. Complete optical isolation between pixels is not possible with this structure since a thickness of silicon equal to the depletion layer width must exist under the apex of the grooves.

Optical isolation of the pixels in a two dimensional array is more difficult. With two orthoganal sets of V-grooves, however, it may be possible to achieve the required degree of optical isolation and yet still enable charge transfer along the channels [82].

4 EXPERIMENTAL INVESTIGATION OF MOS STRUCTURES PULSED ABOVE BREAKDOWN

It is apparent from the discussion in chapter three that an experimental study on the above-breakdown operating regime of MOS structures is required before the development of a full PC-CCD array is attempted. In particular, it must be determined if the avalanche initiation probability saturates as expected, and if it is possible to obtain the very low dark count rates that are required. Data is also required on the degree of photon coupling as a function of the size of discharge pulse and pixel separation. In order to investigate these issues, small area discrete MOS devices, suitable for operation above breakdown and at low temperatures, have been fabricated and tested. The test devices were fabricated on p-type substrates and illuminated from the back side.

The investigation was conducted in two parts. For the initial investigation three very simple device structures were adopted that required a minimum of high temperature processing. Fabrication problems were encountered that prevented the operation of two of the three test structures, however, the results obtained with the remaining structure indicated that dark generation occuring at the $Si-SiO_2$ interface would be a major problem in surface channel PC-CCD's. In the second part of the investigation these detrimental surface effects were largely eliminated by going to a buried channel device structure that breaks down at a bulk n-p junction away from the interface. In a full PC-CCD array the added complexity of this structure is also justified by the higher clocking rates, and hence higher frame rates, that are permissible with buried channel CCD's.

4.1 SURFACE BREAKDOWN DEVICES

4.1.1 Design Considerations

Figure 4.1 shows the potential distribution perpendicular to the interface for a surface breakdown MOS gate, and defines the various potentials and distances that will be refered to in the text.

With these structures the peak field in the depletion region occurs at the $Si-SiO_2$ interface. It was established in section 3.2.5 that this peak field should be kept below approximately 4.3×10^5 Vcm^{-1} in order to avoid significant interband tunneling, and that a substrate doping less than approximately 8×10^{15} cm^{-3} is required if uniformly doped substrates are used. In order to reduce the dark generation due to the tunneling emission from traps and interface states to an acceptable level, the peak field at the interface, and corresponding substrate doping, may have to be even lower. As the substrate doping is lowered, however, the operating voltages increase rapidly (see Fig. 3.16) and the depletion layer widths become comparable to the lateral dimensions of the photogate, making it difficult to achieve uniform planar breakdowns. For this study a substrate doping of $N_A=7x10^{15}$ cm^{-3} was adopted, for which the calculated depletion layer width at breakdown (at 100K) is approximately 3.5 μm. The actual depletion layer widths will be somewhat wider, and the corresponding peak fields lower, than those calculated on the basis of a uniformly doped substrate. This is because boron doped substrates acquire a more lightly doped surface layer after thermal oxidation due to the out diffusion of boron [83], resulting in a structure analogous to an $n^+\pi p$ diode.

It is vital that the field in the oxide remain below the breakdown field strength for SiO_2. The breakdown field for thermal oxides grown in an O_2/HCl ambient ranges from $7 \times 10^6 Vcm^{-1}$ to $1 \times 10^7 Vcm^{-1}$ [84,85]. Prior to an avalanche discharge the field in the gate oxide is given by,

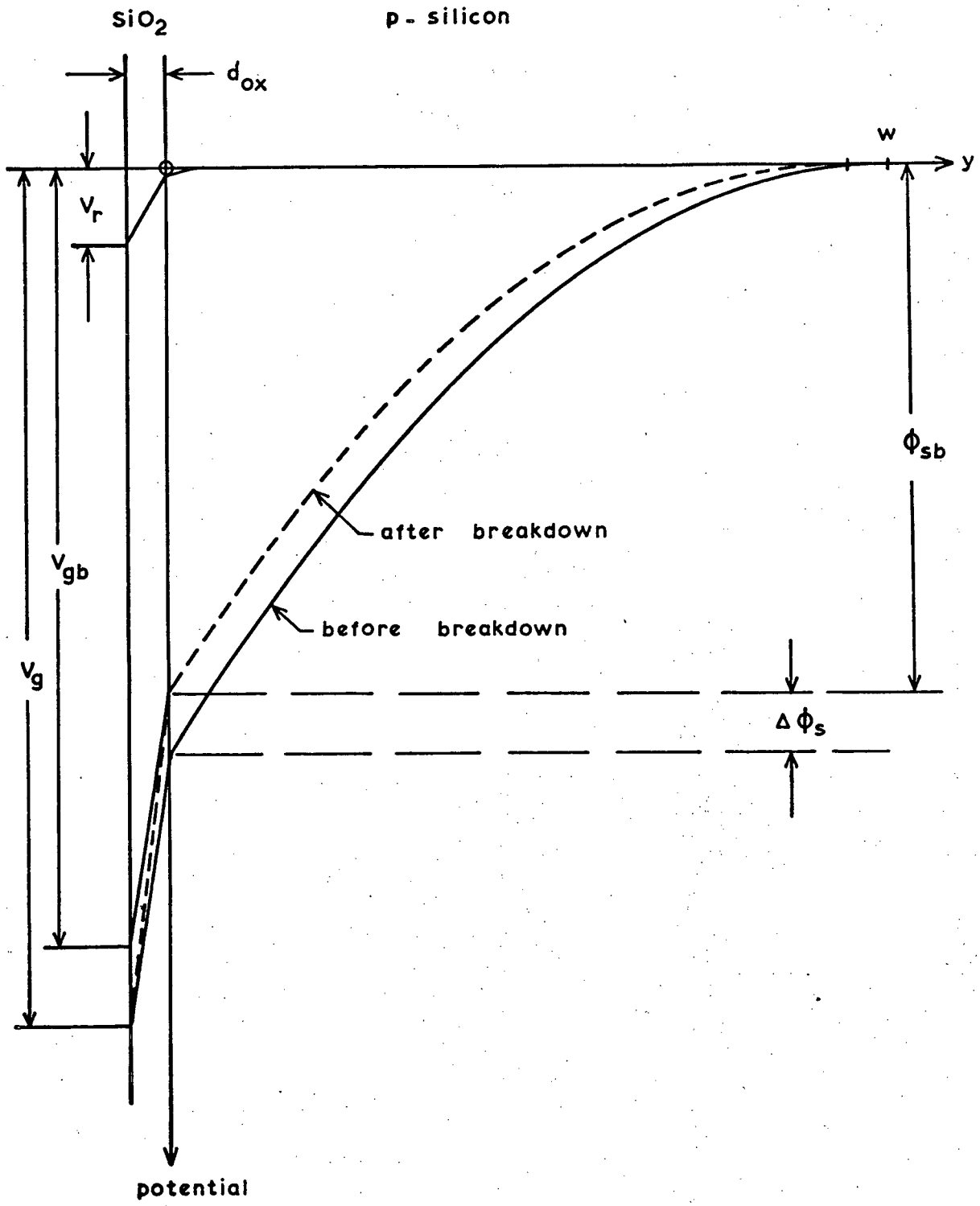


FIGURE 4.1 Potential distribution perpendicular to the interface for a surface-breakdown MOS gate, before and after breakdown, at breakdown, and during reset.

$$\mathcal{E}_{ox} = \frac{\epsilon_s}{\epsilon_i}\mathcal{E}_{max} + \frac{(Q_I - Q_{ss})}{\epsilon_i}$$

where $\mathcal{E}_{max}$ = peak field in the silicon (at the Si-SiO_2 interface)

Q_{ss} = positive fixed oxide charge

Q_I = charge in the inversion layer during reset.

Subsequent to an avalanche discharge the inversion layer charge is increased and the field in the oxide becomes,

$$\mathcal{E}_{ox} = \frac{\epsilon_s}{\epsilon_i}\mathcal{E}_{max} + \frac{(Q_I - Q_{ss})}{\epsilon_i} + \frac{\Delta\phi_s}{d_{ox}} \qquad (4.1)$$

For the substrate doping chosen, the charge $(Q_I - Q_{ss})$ required for strong inversion is approximately 10^{-7} coul cm^{-2}, while the peak field in the silicon at the interface will be close to 4×10^5 Vcm^{-1}. Putting these values in (4.1) gives,

$$\mathcal{E}_{ox} = 1.6 \times 10^6 + \frac{\Delta\phi_s}{d_{ox}} \quad Vcm^{-1}$$

For this investigation, a gate oxide thickness of 0.2μm was chosen so as to enable operation several tens of volts above breakdown and still remain sufficiently below the oxide breakdown field that leakage currents are minimal [86,87]. In addition, with a gate oxide of this thickness, the test devices will not be destroyed if they are inadvertently exposed to room light while deeply depleted. Further, there is evidence that the SiO_2 defect density at high electric field strengths, ($2 - 4 \times 10^6$ Vcm^{-1}) increases rapidly as the oxide thickness is reduced below 0.2μm [88].

During the periods of avalanche breakdown some of the hot electrons incident on the interface will be injected into the conduction band of the

adjacent SiO_2 layer [89,90]. The injection probability, however, should be very small ($<2 \times 10^{-3}$)[91], and in addition, it has been found that trapping of the injected electrons is negligible in dry thermal oxides, so that no drift or stability problems are to be expected.

4.1.2 Test Structure Designs and Mask Layouts

The three miniature surface breakdown devices used in the initial investigation are shown in Figures 4.2 - 4.4. Device 1 (Fig. 4.2) has the simplest possible structure, consisting only of an MOS field plate with a thick oxide guard ring, and an ohmic contact to the substrate. In operation the gate is pulsed from a condition of strong inversion to a potential above that required for breakdown, and held there for a frame time. The gate potential is then returned to the original reset value so as to inject into the substrate any additional charge collected in the inversion layer as the result of an avalanche discharge. The avalanche discharge pulses can be detected simply by monitoring the substrate current, however, the large additional displacement currents during the rising and falling edges of the drive pulse complicate the detection. These drive pulse transients are primarily due to capacitive coupling between the relatively large area bonding pad and the substrate. In device 2 (Fig. 4.3) the drive pulse transients are greatly reduced by providing a field shield underneath the bonding pad and interconnect line. Also, by omitting the thick field oxide, MOS devices with a field plate type guard ring may be fabricated, as in device 2b (Fig. 4.3(b)).

The charge injection mode of operation of devices 1 and 2 results in long reset times since the injected electrons must be given time to recombine with the majority carrier holes. At low temperatures the recombination time constant can be several miliseconds for long lifetime substrates. For

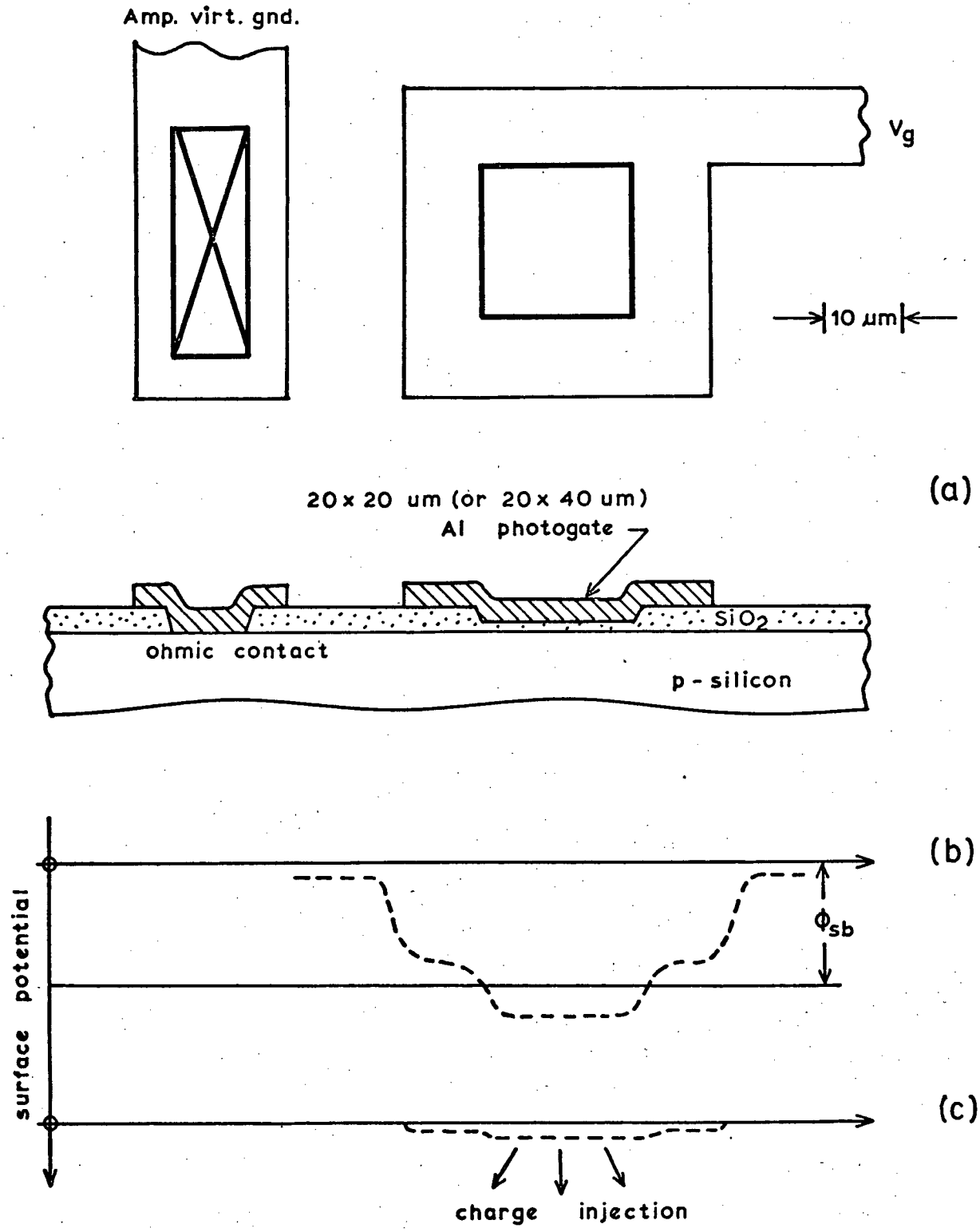


FIGURE 4.2 Charge injection test device, structure 1

(a) layout of the individual gates and vertical structure
(b) potential well diagram in deep depletion, before breakdown
(c) during reset, charge injection

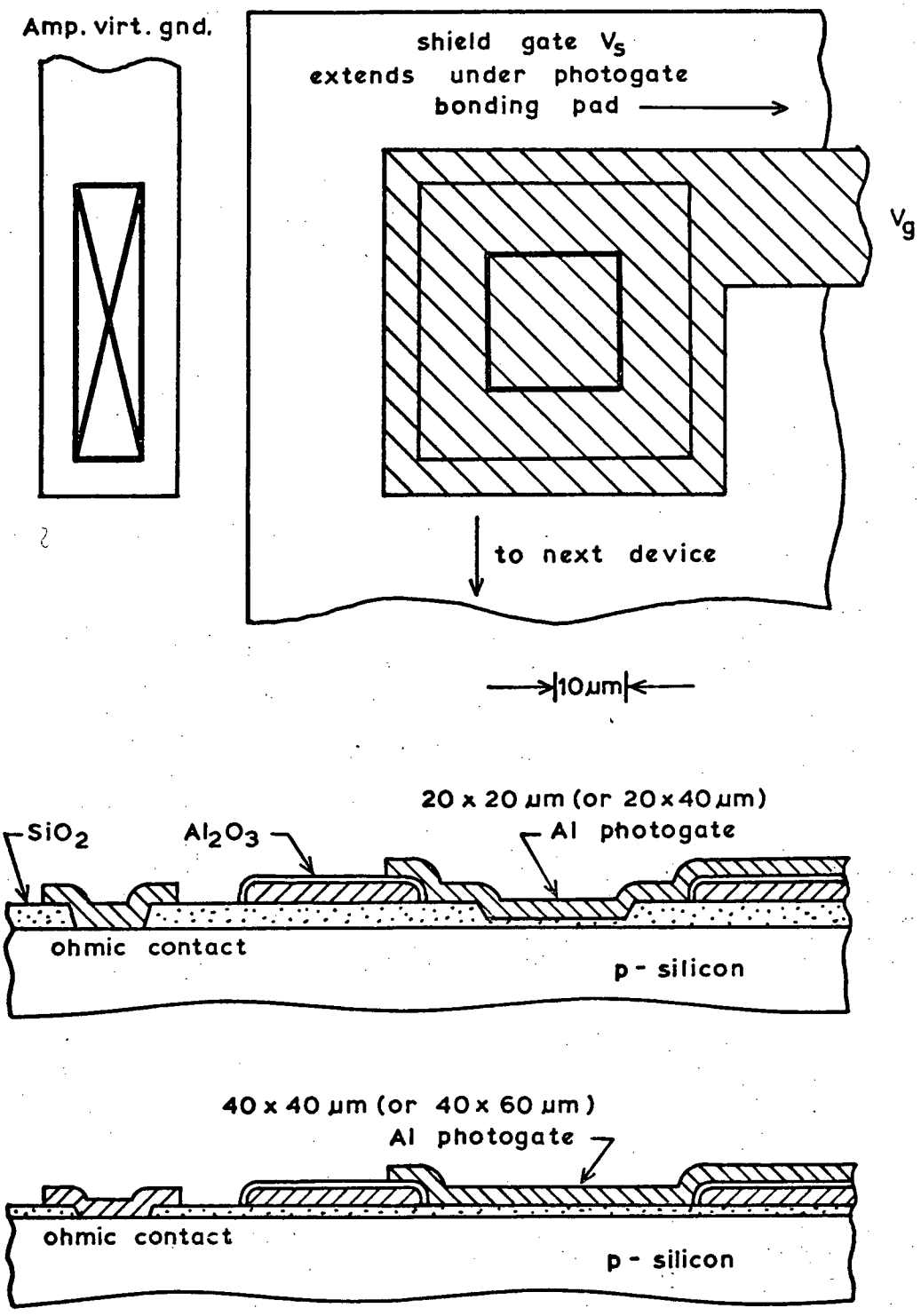


FIGURE 4.3 Charge injection test device, structure 2

(a) layout and vertical structure
(b) alternate structure obtained by omitting the thick field oxide

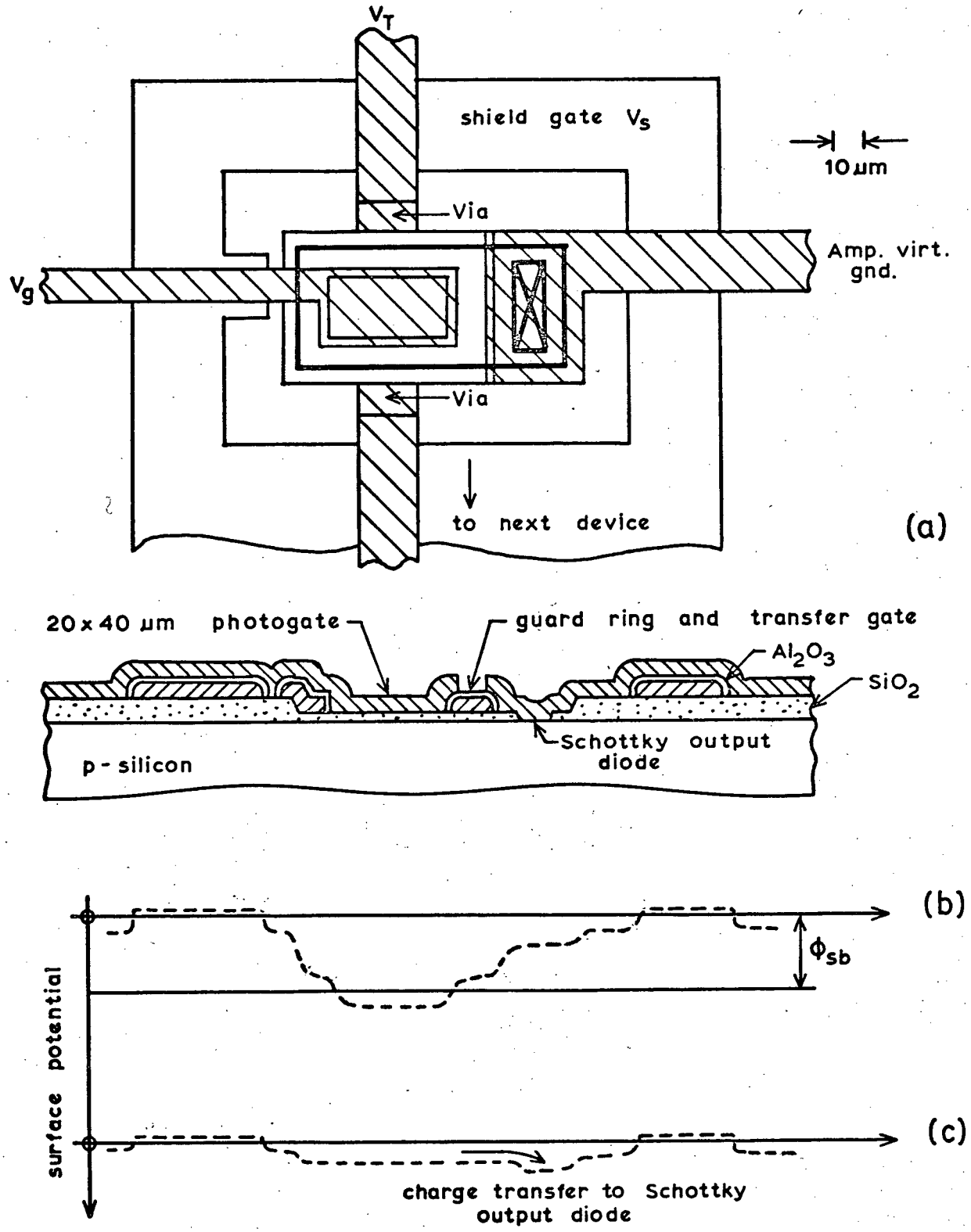


FIGURE 4.4 Charge transfer test device, structure 3

(a) layout of the individual gate and vertical structure
(b) potential well diagram in deep depletion before breakdown
(c) during reset, charge transfer

this reason a device that operates in the charge transfer mode was included as one of the initial test devices (Fig. 4.4). In these devices the signal charge is transferred laterally along the interface to an output diode. During readout the guard ring field plate serves as a transfer gate between the output diode and the photogate. A shield gate is provided around the breakdown gate and output diode in order to interrupt any surface channels that would otherwise connect the devices to the bonding pad depletion regions.

No diffusions other than those used for bulk gettering purposes are required to fabricate the three test structures. The only remaining high temperature steps involved are the thick field and gate oxidations. The gates and interconnect lines are aluminum. The first level undergoes an alloying step in order to make any necessary ohmic contacts to the substrate, and is then partially anodized to form the insulator between first and second level metal [92-94]. The second level of aluminum is not sintered so that MIS diodes may be formed [95]. This double level aluminum/Al_2O_3/aluminum metalization was chosen over the more conventional polysilicon/SiO_2/aluminum scheme, partly because of a lack of polysilicon deposition facilities, and also because there are no high temperature steps involved during the double level aluminum metalization. The minimal high temperature processing required to fabricate the test devices facilitates the maintenance of long bulk lifetimes.

Several test devices of each type were laid out onto a three part die so that they could be fabricated simultaneously using a single set of photomasks. The layouts were designed so as to provide a range of gate separations for photon coupling measurements, and to have an arrangement of bonding pads suitable for mounting in 16 pin DIP packages. Five photomasks are required to fabricate the composite test chip, they are:

1 - gate oxide mask

2 - substrate contact mask

3 - first level aluminum mask

4 - first to second level aluminum contact mask (vias)

5 - second level aluminum mask

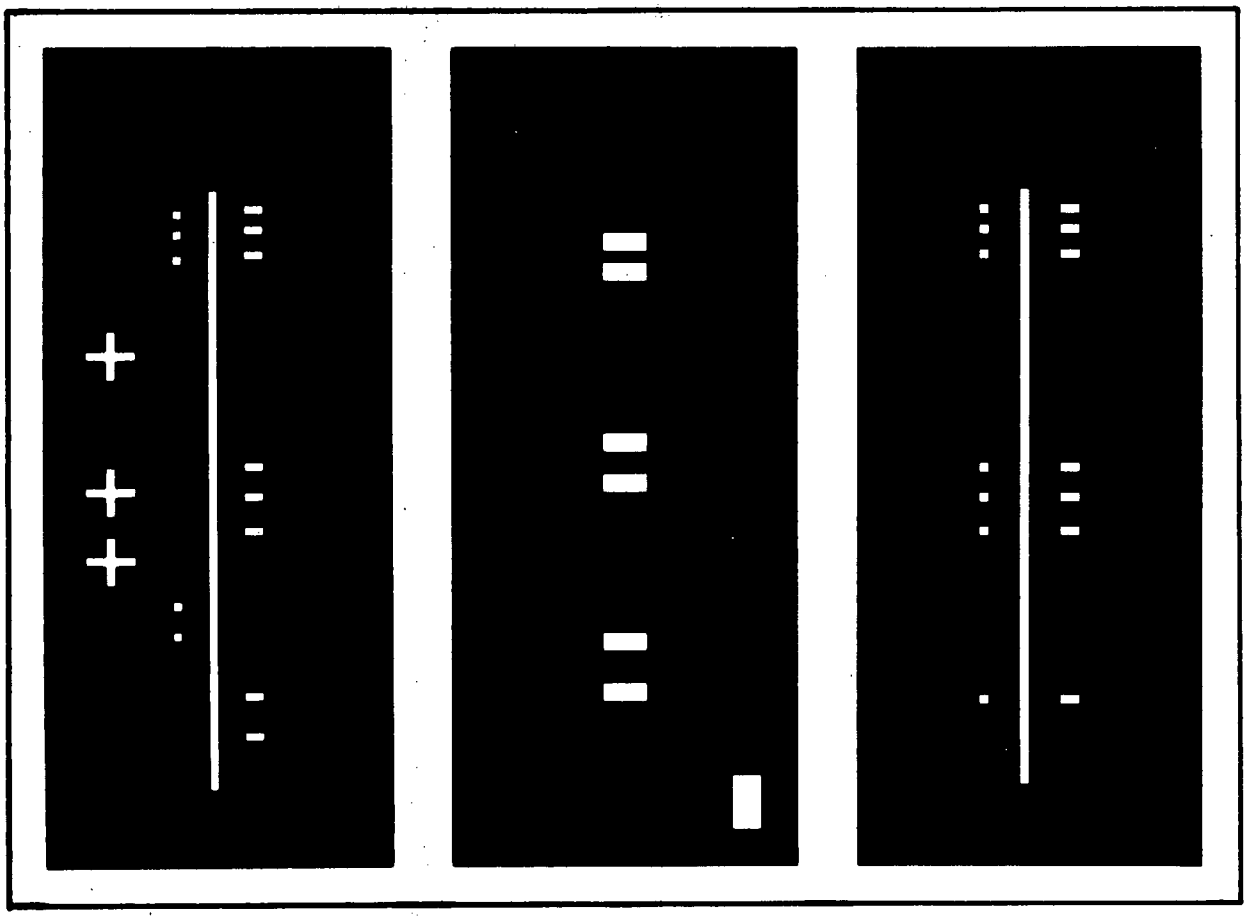
Rubylith masters were generated manually on a coordinograph at a scale of 250 times. The photographic reduction and step and repeat operations were performed by Precision Photomask[1]. The five mask levels are shown in Fig. 4.5(a) - (e).

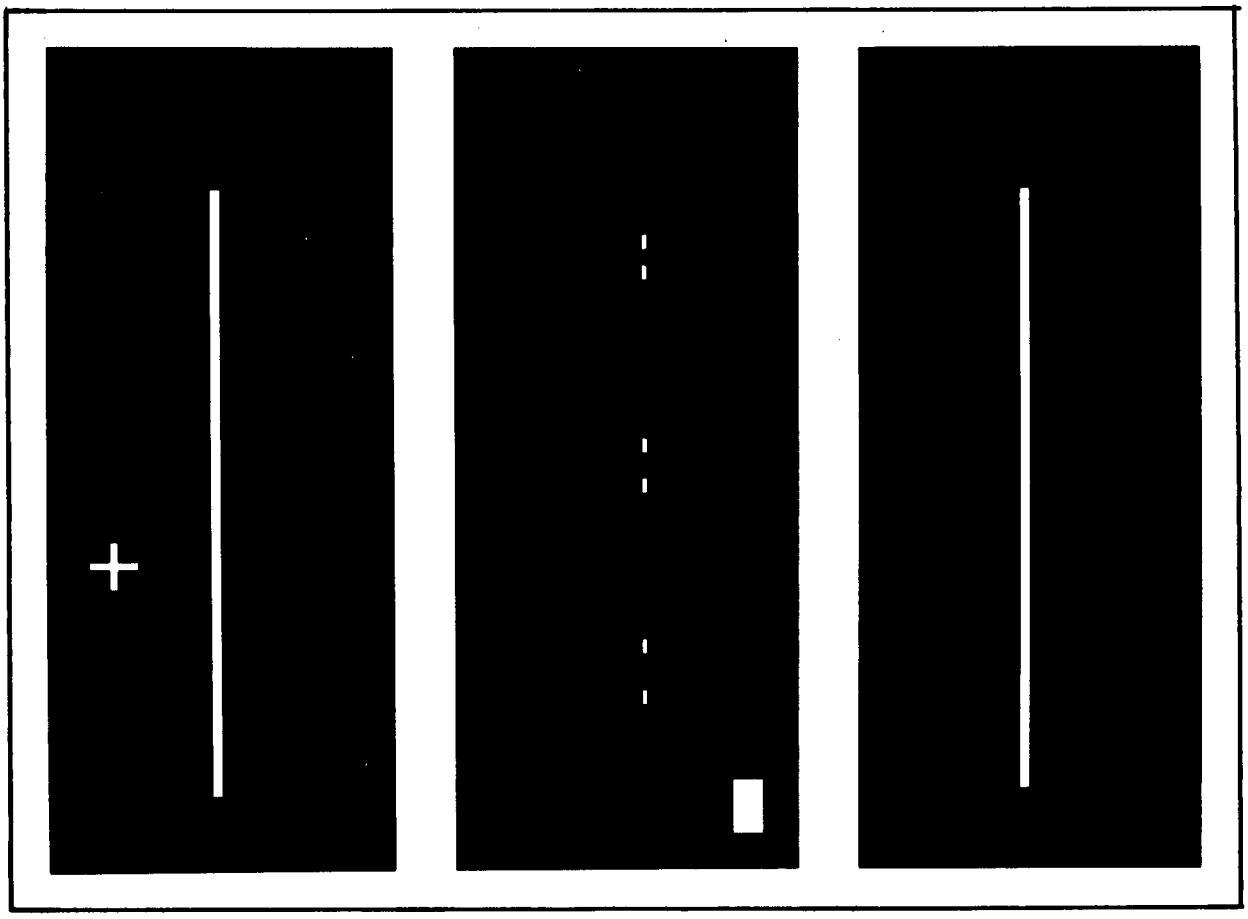
The required first level aluminum pattern consists of various isolated features, (gates, field shields etc.), however, the electrochemical anodization process requires electrical contact to each first level feature that is to be anodized. For this purpose mask 3 includes temporary anodizing contacts that are extended to an aluminum bus in the scribe line area. Contact is made to this bus through the substrate during the anodization process. A protective photoresist mask is used to prevent the temporary contacts and aluminum bus, together with the first to second level vias, from being anodized [92]. Photoresist is also used to protect bare areas of silicon from the electrolyte (in this case the MIS contact areas), as these would result in large leakage paths. All of these protected areas are included on mask 4. The anodic Al_2O_3 layer formed is inert to the phosphoric acid etchant used to pattern aluminum, therefore, the removal of the anodization contacts and bus does not require an additional photomask. These unprotected regions of first level aluminum are etched away while patterning the second level of aluminum. The entire fabrication sequence is illustrated in Fig. 4.6.

[1] 5085 Isaballe street, St Hubert, Quebec, Canada.

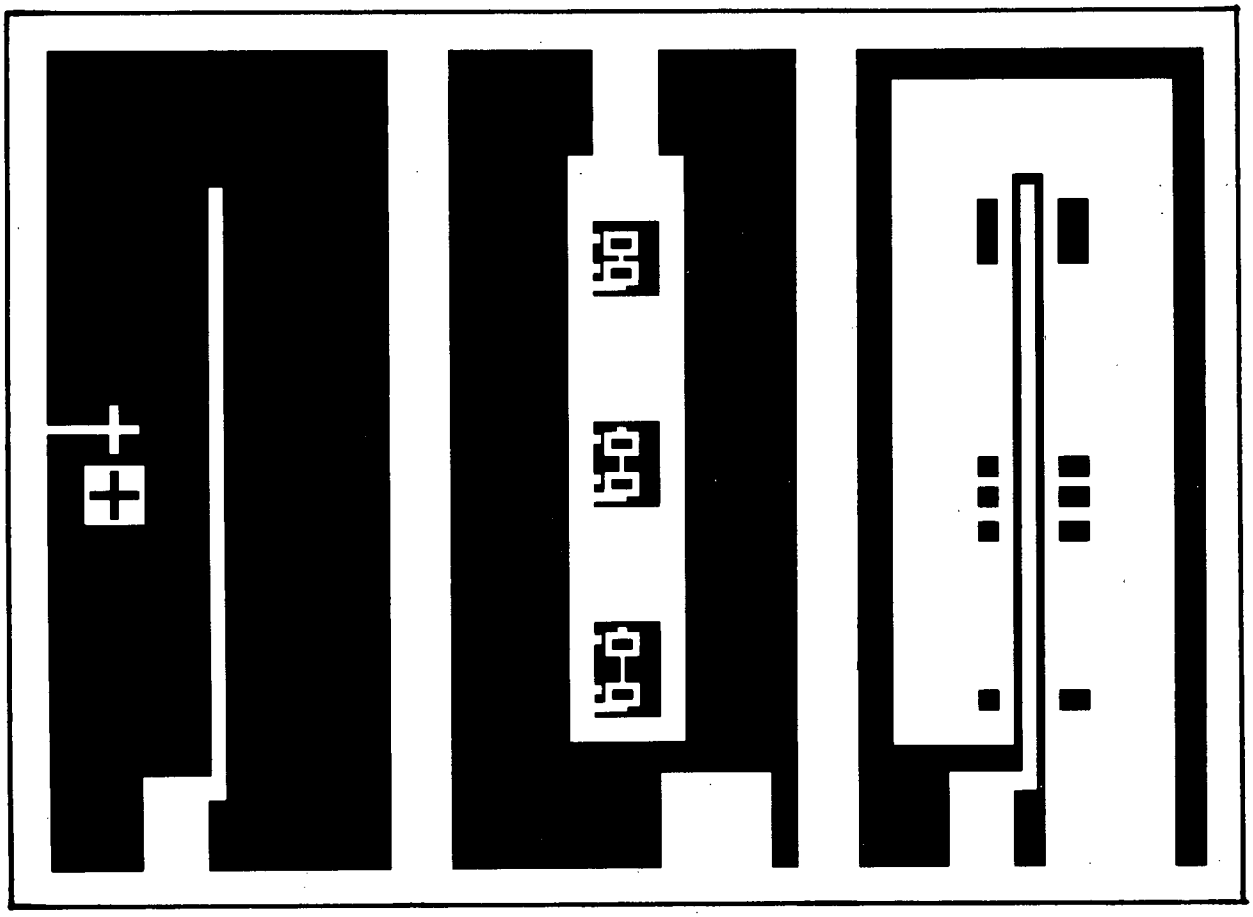
FIGURE 4.5 Copies of the five photomasks used to fabricate the devices. Figures (a), (b), and (d) are reversals of the actual masks used. The last number in the device designations refers to the gate numbers indicated in (e)

(a) gate oxide mask
(b) substrate contact mask
(c) first level aluminum mask
(d) via mask
(e) second level aluminum mask

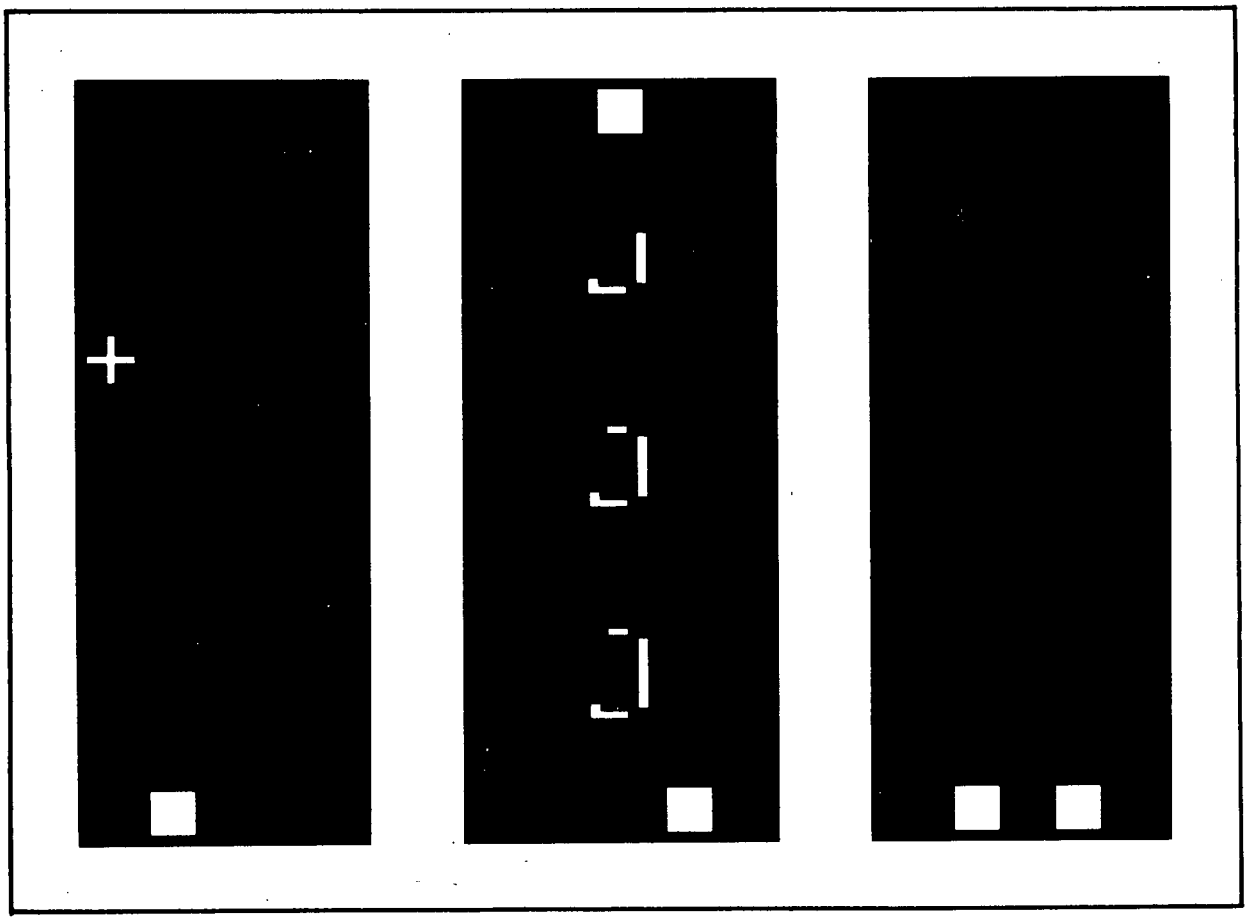




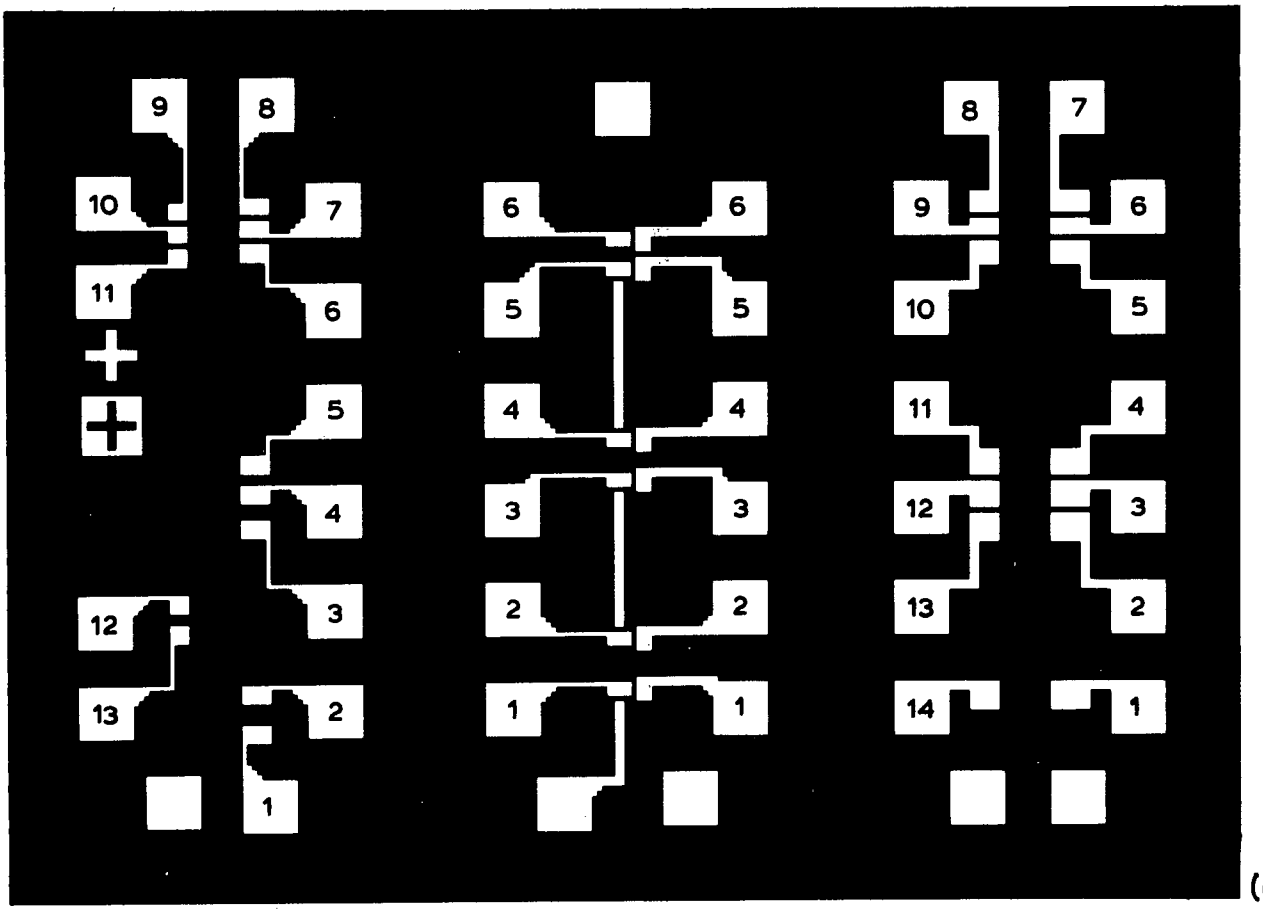
(b)



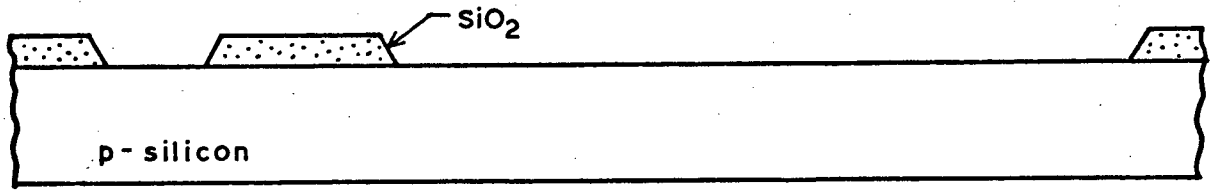
(c)



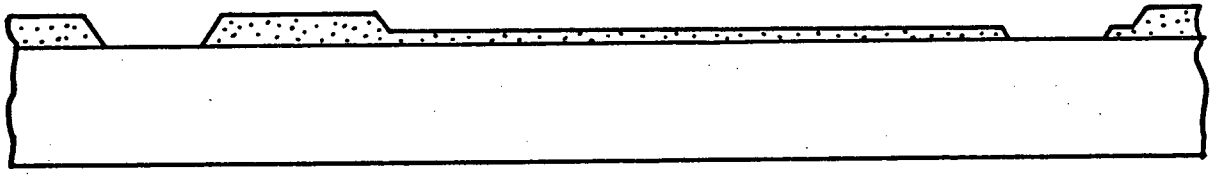
(d)



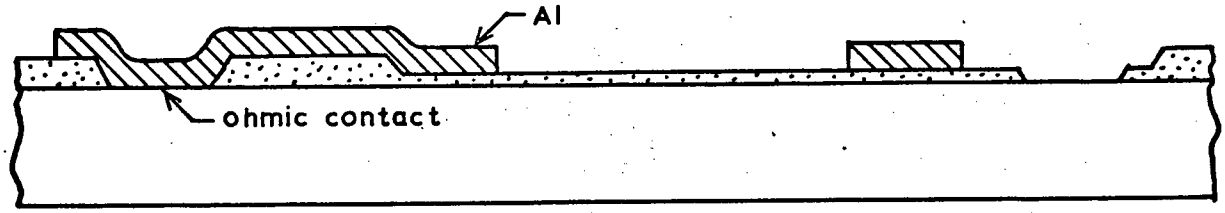
(e)



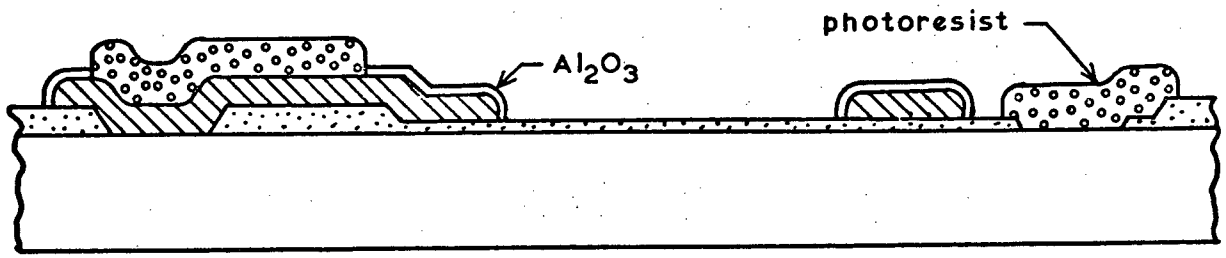
(1) field oxidation and gate oxide window etch



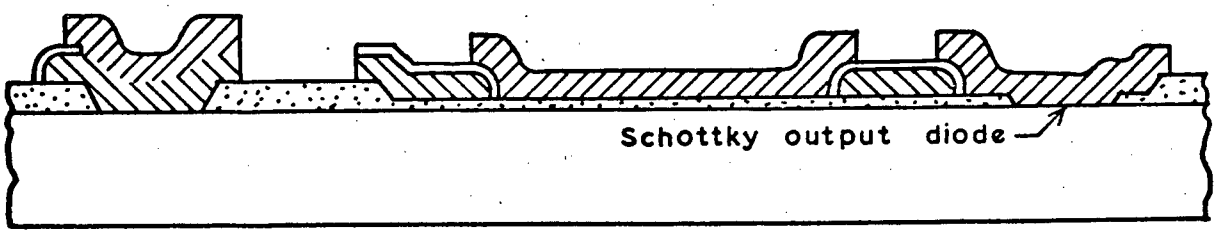
(2) gate oxidation and contact window etch



(3) first level Al deposition and pattern etch, plus H_2 anneal



(4) selective anodization of first level Al



(5) second level Al deposition and pattern etch

FIGURE 4.6 Fabrication sequence for the surface-breakdown test devices.

4.1.3 Device Fabrication

Two lots of devices were fabricated on 2.2 - 2.5Ω cm boron doped (100) Czochralski wafers, with chemo-mechanically polished front surfaces and bright etched rear surfaces. The processing details for the two fabrication runs are listed in Table 4.1 and 4.2. Table 4.3 lists the device wafers and test wafers used and gives the initial bulk resistivities and the final gate and thick field oxide thicknesses. In both fabrications extensive use was made of dry HCl oxidations in order to:

(1) - suppress oxidation induced stacking faults and eliminate grown-in defects [66-69].

(2) - getter lifetime-degrading impurities from the silicon [96-98].

(3) - passivate (i.e., neutralize) any mobile ionic sodium or potassium present in the oxide [99-101].

Complete passivation of all mobile alkali ions is not critical to the successful operation of the present devices, since these ions will be "frozen-in" at the low operating temperature (below 100K). In addition to the above benifits it has also been found that, subsequent to a hydrogen anneal (or aluminum "alneal"), HCl oxides have lower surface state densities than oxides grown in pure O_2 [102]. Further, the defect density at high electric field strengths is lower in HCl oxides [103].

TABLE 4.1 Processing Details For First Fabrication

Step	Operations	Details
	initial clean	- hot xylene with ultrasonic agitation - RCA clean
	surface clean-up	temp. 1100°C
	oxidation	cycle: 5 min O_2 400 min O_2 + 5% HCl 5 min N_2
	strip all oxide	- buffered HF - HCl/H_2O_2 part of RCA clean
1-1	field oxidation	temp. 1100°C cycle: 5 min O_2 140 min O_2 + H_2 [a] 400 min O_2 + 5% HCl 5 min N_2 oxide thickness, 1.00 μm
	gate window etch	- neg resist, mask 1 - buffered HF - strip resist, hot H_2SO_4/H_2O_2 - RCA clean
	gate oxidation	temp. 1100°C cycle: 5 min O_2 62 min O_2 + 5% HCl 5 min O_2 15 min N_2 slow pull in N_2 [b]
1-2		oxide thickness, 0.20 μm
	contact etch	- neg. resist, mask 2 - buffered HF - resist strip, hot H_2SO_4/H_2O_2 - RCA clean - 2% HF dip (20 sec)

TABLE 4.1 cont'd.

Step	Operations	Details
1-3	aluminum evap. I	electron beam evap.: substrate temp., 250°C evap. rate, 50-200 Åsec^{-1} film thickness, 1.0 μm
	aluminum etch I	- neg. resist, mask 3 - phosphoric/nitric etch, 60°C - resist strip, Microstrip
	contact sinter plus hydrogen anneal	temp. 450°C cycle: 10 min N_2 30 min N_2 + 50% H_2
1-4	via photoresist	- Waycoat LSI 295 positive, mask 4
	aluminum anodization	: electrolyte, 25w/o APB-EG [c] current, 0.5 mA cm^{-2} formation voltage, 85 V soak time at 85 V, 2 min - resist strip, hot acetone - Phosphoric etch, 40°C, 60 sec
1-5	aluminum evaporation II	tungsten filament evap.: substrate temp., 200°C evap. rate, 50-100 Åsec^{-1} film thickness 1.0 μm
	aluminum etch II	- neg. resist, mask 5 - phosphoric etch 60°C - resist strip, Microstrip

[a] Burnt hydrogen wet oxidation

[b] Pulled manually from the hot zone over a period of about 10 min.

[c] 25 weight % ammonium pentaborate dissolved in (hot) ethylene glycol (under an argon atmosphere to prevent oxidation).

TABLE 4.2 Processing Details For Second Fabrication

Step	Operations	Details
	initial clean	- hot xylene with ultrasonic agitation - RCA clean
	surface clean-up oxidation plus phosphorus diffusion mask	temp. 1125°C cycle: 5 min O_2 100 min $O_2 + H_2$ [b] 400 min O_2 + 5% HCl 5 min N_2
	strip oxide off	- HF vapor-etch (back side)
2-1	back	- HCl/H_2O_2 part of RCA clean
	phosphorus getter diffusion [a]	temp. 1050°C source $POCl_3$ pass over, 15°C cycle: 5 min N_2 + 3% O_2 30 min N_2 + 3% O_2 + source N_2 5 min N_2 + 3% O_2
	strip oxide off front	- HF vapor-etch (front side) - 10% HF, 60 sec. (phosphorous glaze strip on back side) - HCl/H_2O_2 part of RCA clean
2-2	field oxidation [a]	slices P2-0, P2-2: temp. 1125°C cycle: 10 min O_2 40 min $O_2 + H_2$ 5 min O_2 15 min N_2 Oxide thickness, 0.52 μm slices P2-1, P2-3: temp. 1125°C

TABLE 4.2 cont'd.

Step	Operations	Details
2-2 cont'd		cycle: 10 min O_2 73 min $O_2 + H_2$ 5 min O_2 15 min N_2 oxide thickness, 0.72 μm
	gate window etch	- neg. resist, mask 1 - buffered HF (back side protected) - strip resist, hot H_2SO_4/H_2O_2 - RCA clean
2-3	gate oxidation [a]	temp. 1125°C cycle: 10 min O_2 49 min O_2 + 5% HCl 5 min O_2 5 min N_2 slow pull in N_2 [c] oxide thickness, 0.18 μm
	contact window etch	- neg. resist, mask 2 - buffered HF - strip resist, hot H_2SO_4/H_2O_2 - RCA clean - 2% HF dip 20 sec.
2-4	aluminum evaporation	electron beam evap: substrate temp., 250°C evap. rate, ≈ 50 Åsec^{-1} film thickness, 1.0 μm

TABLE 4.2 cont'd.

Step	Operations	Details
2-4 cont'd	aluminum etch I	- neg. resist, masks 3 + 5 (double exposure) - phosphoric/nitric etch 60°C - strip resist, Microstrip
	aluminum etch II	- neg. resist, masks 4 + 5 (double exposure) - phosphoric/nitric etch, 60°C - strip resist, Microstrip
	contact sinter	temp. 450°C cycle: 30 min N_2
	hydrogen anneal	temp. 350°C cycle: 180 min H_2
	etch off back side n^+ layer	- white etch (front protected with wax) - hot trichloroethylene (to strip wax)

[a] Quartz furnace tube and boat pre-cleaned with O_2/HCl gas flow for 2 hours, prior to loading devices.

[b] Burnt hydrogen wet oxidation.

[c] Pulled manually from the hot zone very slowly, over a period of about 20 min., followed by a fast pull once reaching a temperature of approximately 600°C.

TABLE 4.3 Device And Test Wafer Data

wafer[a]	use	bulk[b] resistivity (Ωcm)	gate oxide thickness (μm)	gate oxide V_{FB} (V)	field oxide thickness[d] (μm)	field oxide V_{FB} (V)
P1-0 P1-1 P1-2 P1-3	anodization tests	~2.5	~0.20	-	-	-
P1-4*	devices	2.35	0.19	1.5	1.00	8.0
P1-5		2.45	0.19	-	1.00	-
P1-6		2.30	0.19	-	1.00	-
P1-7	surface state measurements	2.38	0.19	-	-	-
P2-0	devices	2.45	0.17	-	0.55	-
P2-1		2.32	0.17	-	0.74	-
P2-2		2.35	0.17	-	0.55	-
P2-3*		2.22	0.17	1.2	0.74	5.3
P2-4	profile and surface state measurements	2.24	0.17[c]	-	-	-
P2-5		2.48	0.17[c]	-	-	-

[a] Boron doped 2"(100) Czochralski wafers, chemo-mechanical polish on front surface, bright-etched rear surface. Wafer thickness ~280μm. *Device wafers tested.

[b] Measured with four point probe before processing.

[c] Thickness obtained from measured capacitance in strong accumulation (using $\epsilon_i/\epsilon_o=3.8$).

[d] Thickness obtained from colour, accuracy approximately ±0.02μm.

(1) First fabrication run

Prior to device processing the wafers were oxidized in an O_2/HCl ambient to remove any surface damage and provide some initial gettering of impurities. This oxide was then stripped and the thick field and gate oxides were grown, again using HCl. A field oxide thickness of 1 μm was chosen in order to ensure that the silicon beneath the edges of the metalization would remain well below breakdown during device operation. A 15 minute nitrogen anneal was included at the end of the gate oxidation in order to minimize the positive fixed surface-state charge Q_{ss} [104]. After opening both the ohmic and MIS contact windows the first level of aluminum was deposited (via electron beam evaporation) in a planetary evaporator. The substrates were heated to 250°C during the evaporation to further improve step coverage [105]. In order to achieve tapered metalization edges and alleviate step coverage problems during the second level metalization, nitric acid was added to the phosphoric acid etchant used to pattern the first level of aluminum [106]. The nitric acid causes the photoresist to lose adherence and gradually lift from the edge inwards during etching, resulting in the desired edge profile. After patterning the first level of aluminum the slices were given a 450°C contact sinter in order to make ohmic contact to the anodization bus and to anneal out any x-ray damage caused by the electron beam evaporation. A hydrogen anneal to reduce surface state densities was also included at this time, since an active metal "alneal" following the second metalization is not possible, as this would sinter the MIS contacts and destroy their rectifying property.

After applying the via photoresist the wafers were anodized individually, in an electrolyte of 25 w/o ammonium pentaborate in ethylene glycol, using the special teflon wafer holder shown in Fig. 4.7. The anodic oxide layers were formed at a low constant current density of 0.5 mA cm^{-2} and held

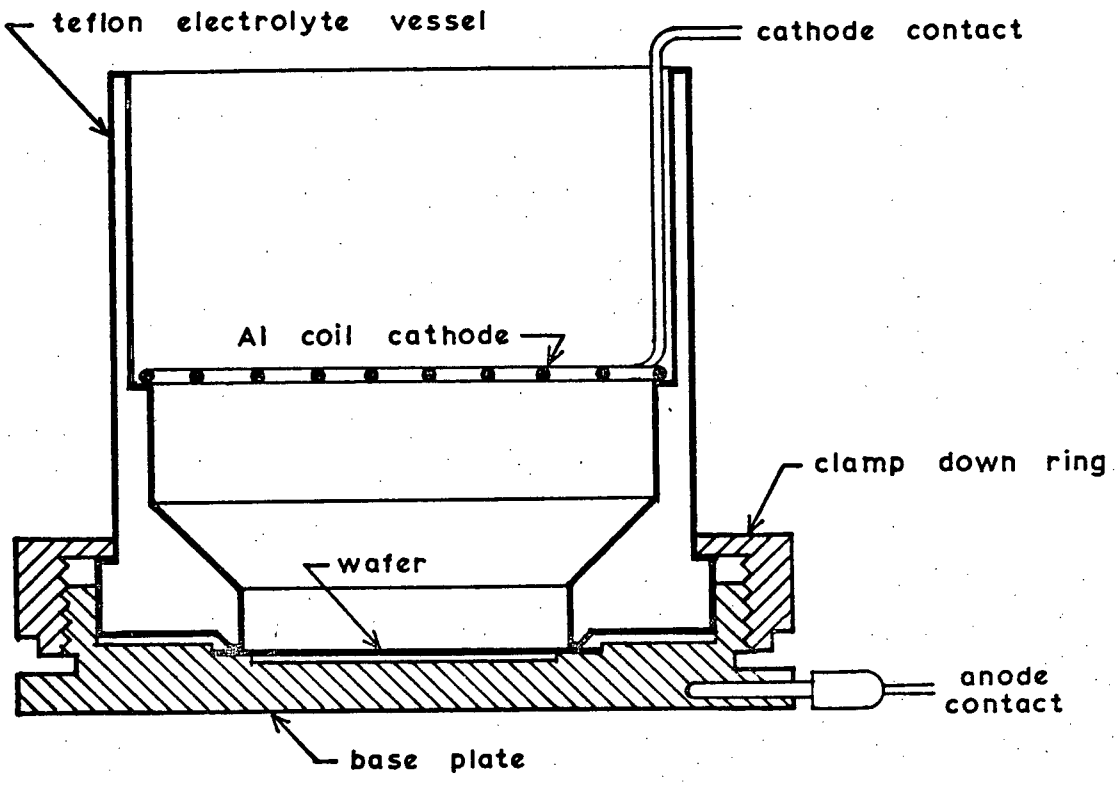


FIGURE 4.7 Teflon anodization cell used in the device fabrication.

for only a few minutes at the final voltage while the current decayed, since both high current density and a prolonged soak time at constant voltage are reported to result in pore formation [107]. The results of some initial anodization tests, on wafers with large photoresist patterns, had indicated that the anodic oxide could be formed to a voltage of 120 V before the resist mask started to break down and lift at the edges. Unfortunately, during the anodization of the device wafers it was found that the small photoresist features lifted at a much lower voltage. Parts of the via photoresist pattern lifted completely at approximately 60 V on the first device wafer anodized. Re-baking the photoresist at a higher temperature enabled the anodization to be extended to about 85 V on the remaining two wafers. However, after depositing and patterning the second level of aluminum it was found that the breakdown voltage of the Al_2O_3 layer was only 40 V, less than half the formation voltage. The second level of aluminum was evaporated from a tungsten filament since the MIS contacts preclude the annealing of x-ray damage resulting from an electron beam evaporation.

The cause of the low breakdown strength of the anodic Al_2O_3 layer is not certain, but is thought to be due to annealing hillocks present in the first level of aluminum prior to the anodization [108,109]. Such hillocks are the result of compressive stresses in the aluminum film following the contact sinter. Even if these features are properly anodized, the increased fields present at the apex of the sharp hillocks could cause premature breakdown. Because of the low breakdown voltage of the anodic oxide layer, only the simple MOS gates with thick oxide guard rings (device structure 1) were operational above breakdown.

(2) Second fabrication run

On this fabrication run the double level Al-Al_2O_3-Al metalization

scheme was abandoned and only device structure 1 was fabricated. In addition to the initial HCl oxidation a phosphorus getter diffusion [110,111] was also included prior to device fabrication, in order to try and improve bulk lifetimes. Furthermore, to help reduce the number of impurities introduced during processing, the wafer boat and furnace tube were cleaned with an O_2/HCl gas flow for two hours before each high temperature operation. Devices were fabricated with two different thicknesses of field (i.e., guard ring) oxide, 0.54μm and 0.74μm. The gate oxidation cycle was identical to the previous fabrication. After opening the contact windows a single level of aluminum was deposited via electron beam and patterned. In order to achieve the desired pattern in one level of aluminum using the existing masks, double exposures and two photoresist operations were required (see Table 4.2). After etching the aluminum pattern the slices were given a 30 minute contact sinter and aluminum "alneal" at 450°C in N_2, followed by a 350°C hydrogen anneal in pure H_2 for 3 hours. The back-side n^+ layer (phosphorus getter) was then etched away, completing the processing.

4.1.4 Test Chamber and Electronics

For testing, the individual die were mounted in modified ceramic 16 pin DIP packages. To enable back-side illumination of the devices the metal bottoms of the DIP packages were replaced by glass and the test chips were glued to this using a transparent epoxy (Araldite high vacuum epoxy). The package pins were then bent backwards to facilitate illumination of the devices through the glass

Figure 4.8 shows a cross-section of the cold chamber used for the low temperature device testing. Cooling is accomplished with a liquid nitrogen heat pipe that protrudes into an evacuated housing. The front (window) end of the housing is removable so that the packaged devices can be clamped to

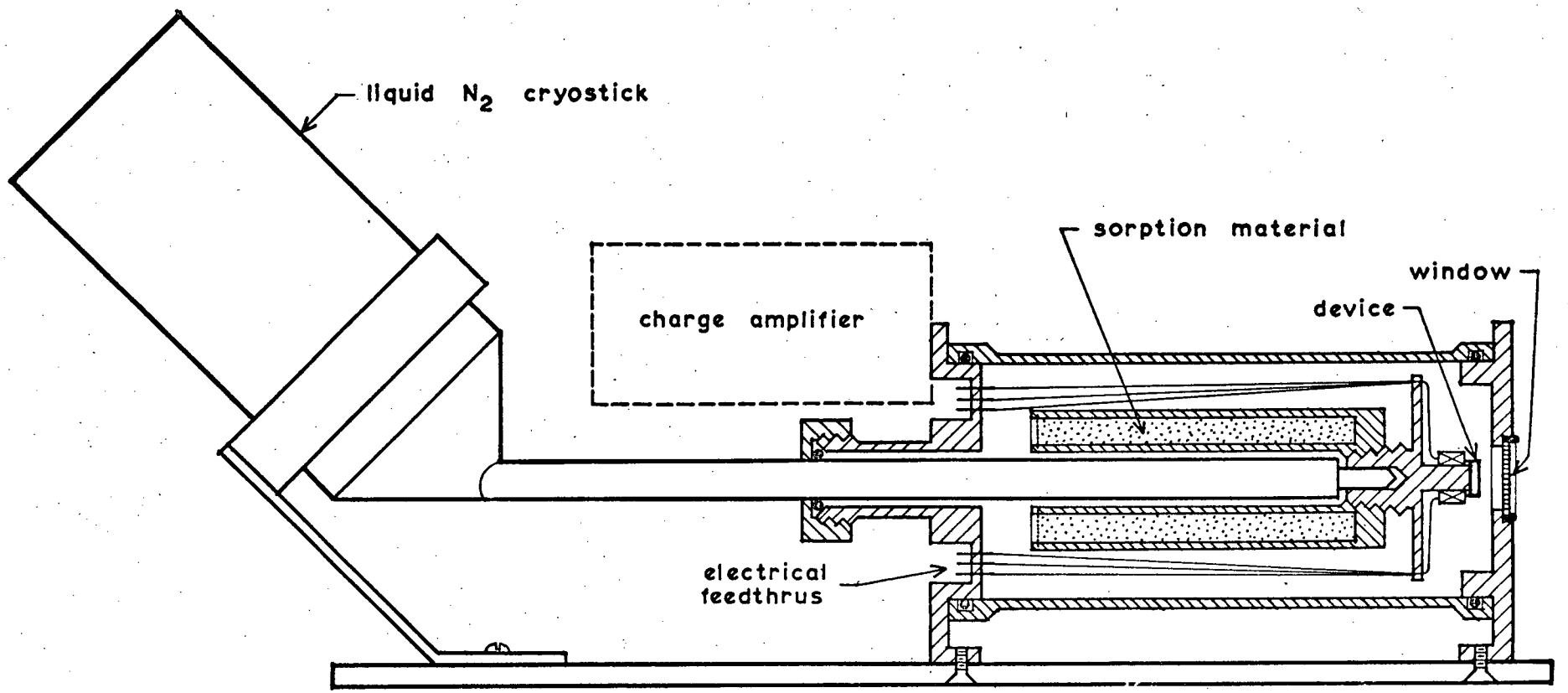


FIGURE 4.8 Cross-section of the cold chamber used for the low temperature device testing.

an aluminum cold finger on the end of the heat pipe. The test socket is connected to electrical feedthroughs on the back of the chamber with 3 mil teflon insulated copper wire. The liquid nitrogen heat pipe also cools a canister of zeolite sorption material within the test chamber. After inserting the test devices the chamber is rough-pumped and then sealed off. During the cool-down the sorption pump evacuates the chamber further and once cold will maintain the vacuum below 10^{-4} torr. The vacuum provides insulation and ensures that no moisture is able to condense and freeze on the devices. A copper constantan thermocouple is used to monitor the temperature of the cold finger. In operation the cold finger reaches a stable temperature of 80 K. A power resistor was added during the testing of the bulk channel devices, enabling the substrate temperature to be altered over a moderate temperature range from 80 K to 140 K.

A block diagram of the test electronics is shown in Fig. 4.9 . Schematics for the driver, amplifier, descriminator and timing circuitry are given in Appendix B. The driver supplies a very clean trapezoidal waveform with no overshoot or undershoot. The ramp rates on the leading and trailing edges may be varied independently from $10^{4} Vsec^{-1}$ to $5 \times 10^{6} Vsec^{-1}$. The upper level may be varied from +20 V to +150 V while the lower level may be varied from -90 V to +50 V. The amplifier used to detect the avalanche charge pulses is housed in an aluminum box on the back of the cold chamber. It consists basically of a low noise wide-band J-FET input op-amp (LF 356) operated as a current to voltage amplifier, followed by further voltage amplification and then a high speed sample and hold smplifier (LH0053), operated as a preset integrator. The output of the current amplifier is monitored on an oscilloscope and the integrator output is passed to a descriminator circuit and counter. The current integrator may be taken out of the preset mode to start integration anywhere along the positive or nega-

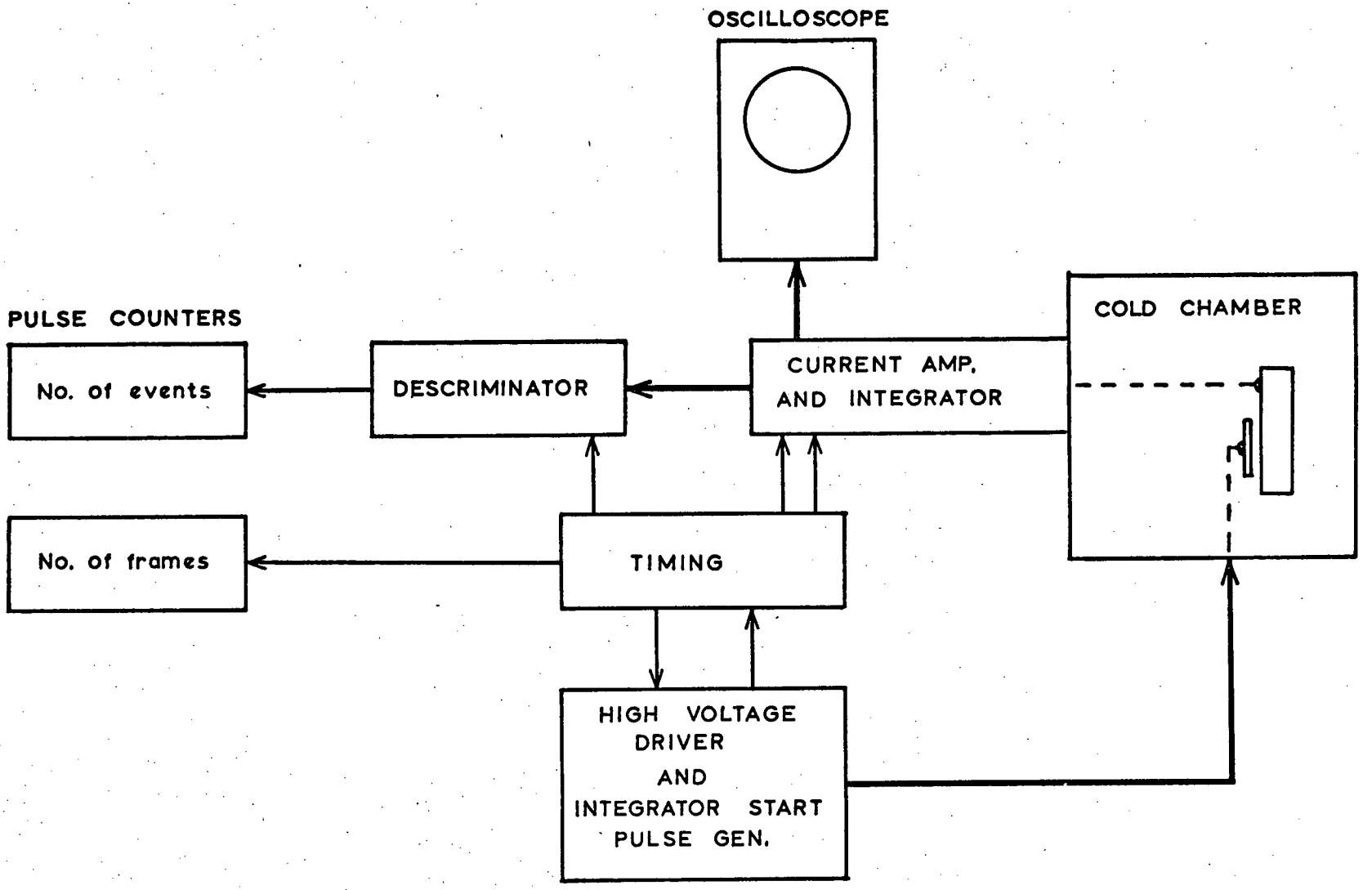


FIGURE 4.9 Block diagram of the test electronics.

tive going edge of the drive pulse and the duration of the integration may be varied independently of the driver waveform. Before being preset, the integrator goes into a hold mode momentarily while the descriminator is activated. The level of noise at the descriminator varied from 3×10^3 electrons r.m.s. for short (0.2msec) integrations to approximately 1.5×10^4 electrons r.m.s. for longer (2msec) integrations.

4.1.5 Modeling of the Completed Devices

In order to interpret the results and make comparisons with theory, an accurate doping profile through the depletion region under the gate, and accurate gate oxide and field oxide thicknesses are required. The field oxide thickness was determined to within ± 0.02 μm from its colour. The doping profile and gate oxide thickness were determined from C-V data according to the method outlined in Appendix C. These latter measurements were made on wafers that had received the same high temperature processing as the device slices and that had a nearly identical starting resistivity (see Table 4.3). The measured doping profiles for the two fabrication runs are shown in Figures 4.10 and 4.11. To simplify the calculation of electric field strength as a function of position in the depletion region, the doping profiles were approximated by the two straight line segments indicated. By making the depletion approximation it can then be shown that the electric field is given by:

$$\mathcal{E}_{1s}(y) = \frac{qN_1}{\epsilon_s}(w - y) - \frac{q(N_1 - N_0)}{\epsilon_s 2d_1}(d_1 - y)^2 , \quad 0 \leq y \leq d_1 \tag{4.2}$$

$$\mathcal{E}_{2s}(y) = \frac{qN_1}{\epsilon_s}(w - y) , \quad d_1 \leq y \tag{4.3}$$

$$\mathcal{E}_{ox} = \frac{\epsilon_s}{\epsilon_i}\mathcal{E}_{1s} \tag{4.4}$$

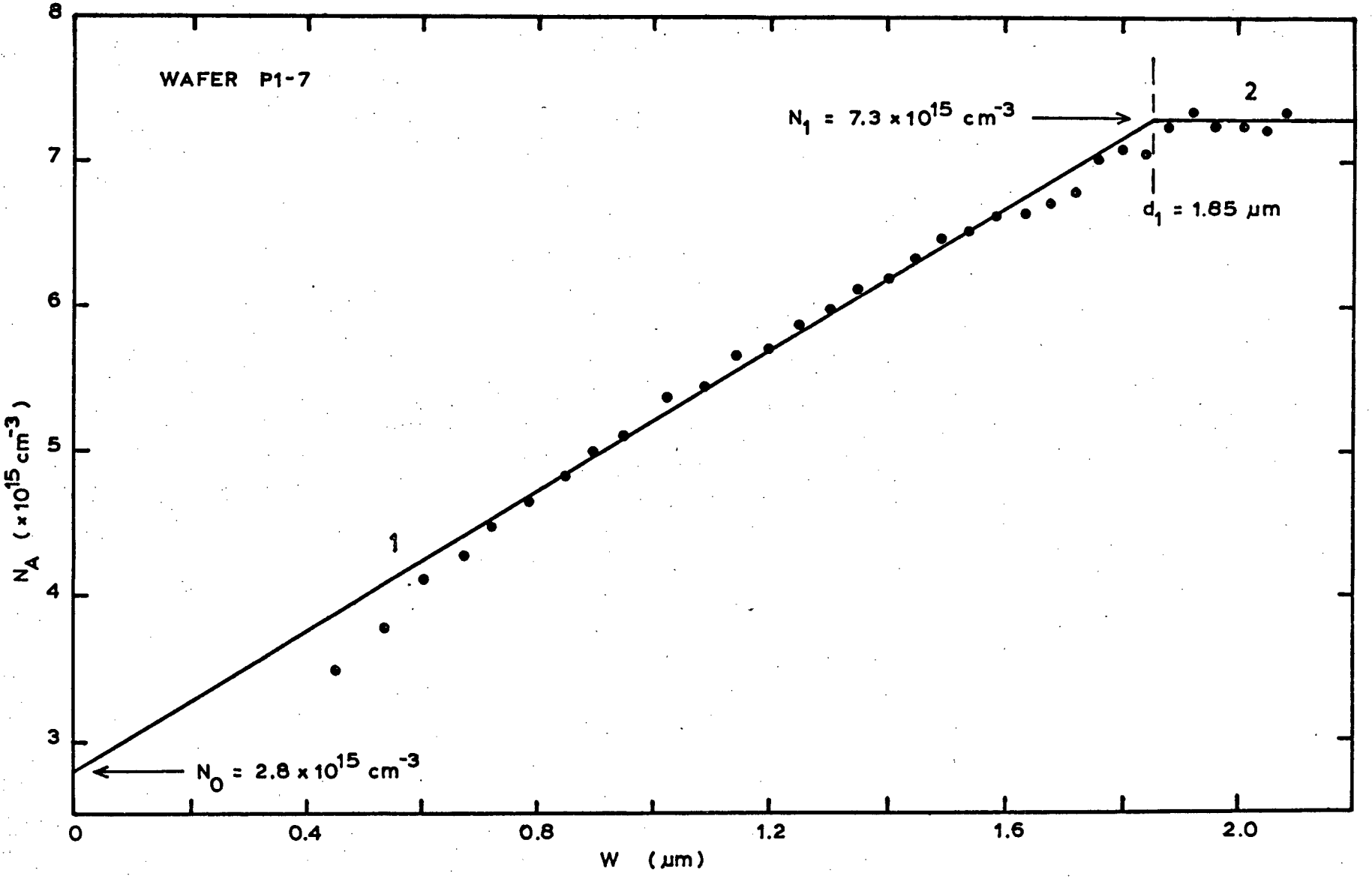


FIGURE 4.10 Surface doping profile for wafer P1-7, obtained from C(V) data by the method described in Appendix C. The solid line shows the approximate profile used to model the devices.

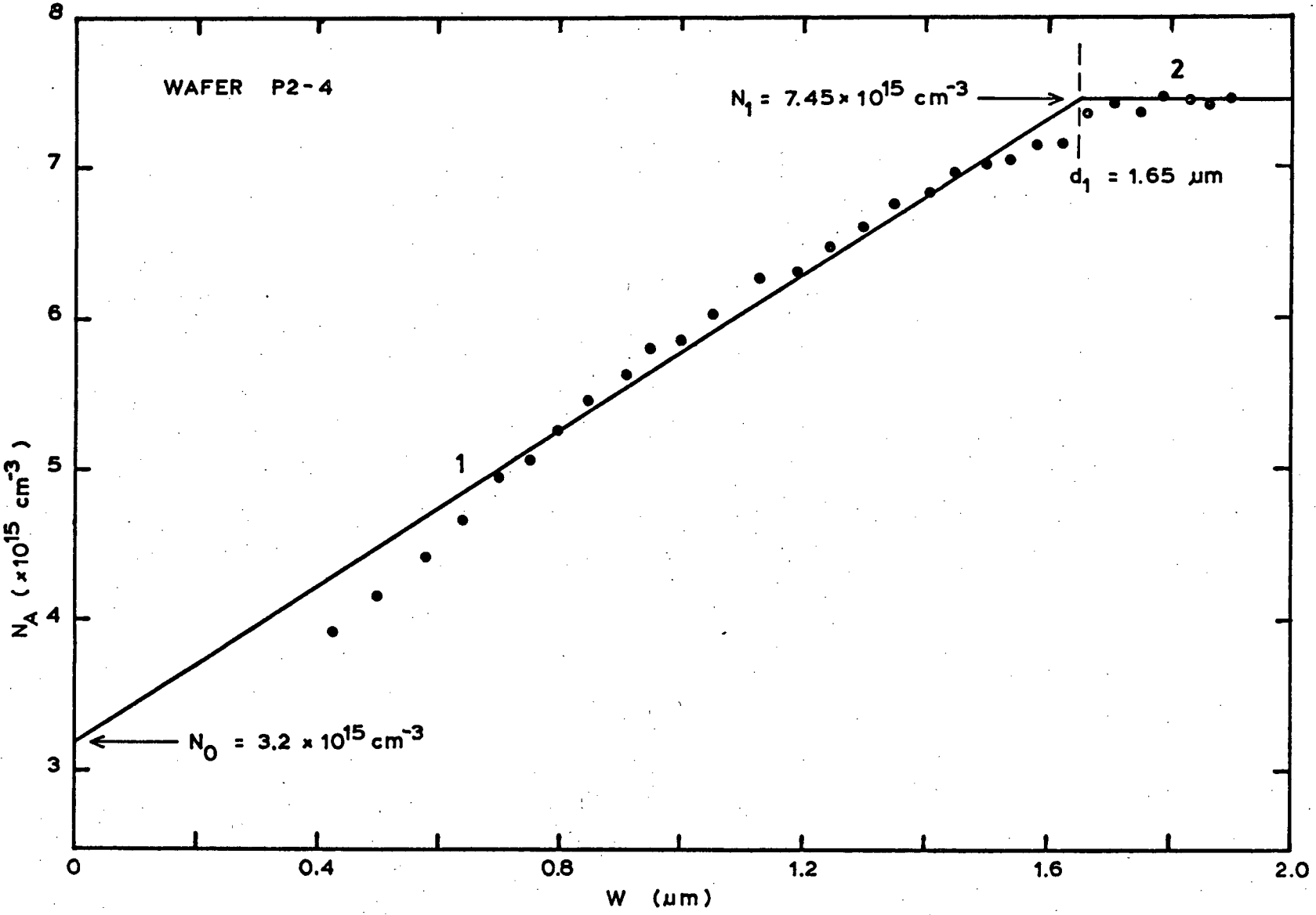


FIGURE 4.11 Surface doping profile for wafer P2-4 and the approximate profile used for device modeling.

where y is the position in the depletion region, measured from Si-SiO_2 interface, and N_0, N_1 and d_1 are the parameters given in Figures 4.10 and 4.11. The depletion layer width w is given by

$$w = -\frac{\epsilon_s}{\epsilon_i} d_{ox} + \left\{ \left(\frac{\epsilon_s}{\epsilon_i} d_{ox}\right)^2 + \frac{\epsilon_s (N_1 - N_0)}{\epsilon_i N_1} d_1 d_{ox} + \frac{(N_1 - N_0)}{3N_1} d_1^2 + \frac{2\epsilon_s Q_I d_1}{\epsilon_i q N_1} + \frac{2\epsilon_s (V_g - V_{FB})}{q N_1} \right\}^{\frac{1}{2}} \qquad (4.5)$$

where Q_I is the amount of charge present in the surface inversion layer and V_{FB} is the flat band voltage. The silicon surface potential is given by,

$$\phi_s = \frac{qN_1}{2\epsilon_s} w^2 - \frac{q(N_1 - N_0)}{6\epsilon_s} d_1^2 \qquad (4.6)$$

In addition to the device structure parameters, the interface state density and bulk lifetime were also determined for the devices from the second fabrication. These measurements were made on test wafer P2-5. The measured interface state density is shown in Fig. 4.12. The method used, described in Appendix C, limited the range over which the interface state density could be determined to those energies near mid-gap and towards the valence band edge, whereas, the interface states of interest are those nearer the conduction band. The densities of these interface states, however, are typically lower that those near the valence band [102]. The bulk lifetime was estimated by monitoring the substrate displacement current of the 1 mm dia MOS gates, following a depleting pulse at room temperature. The initial value of the decay current was measured on a Keithly model 602 electrometer (in the fast mode) and related to the bulk lifetime through Eq. 3.33. The depletion region width w was obtained from Eq. 4.5 using the doping profile

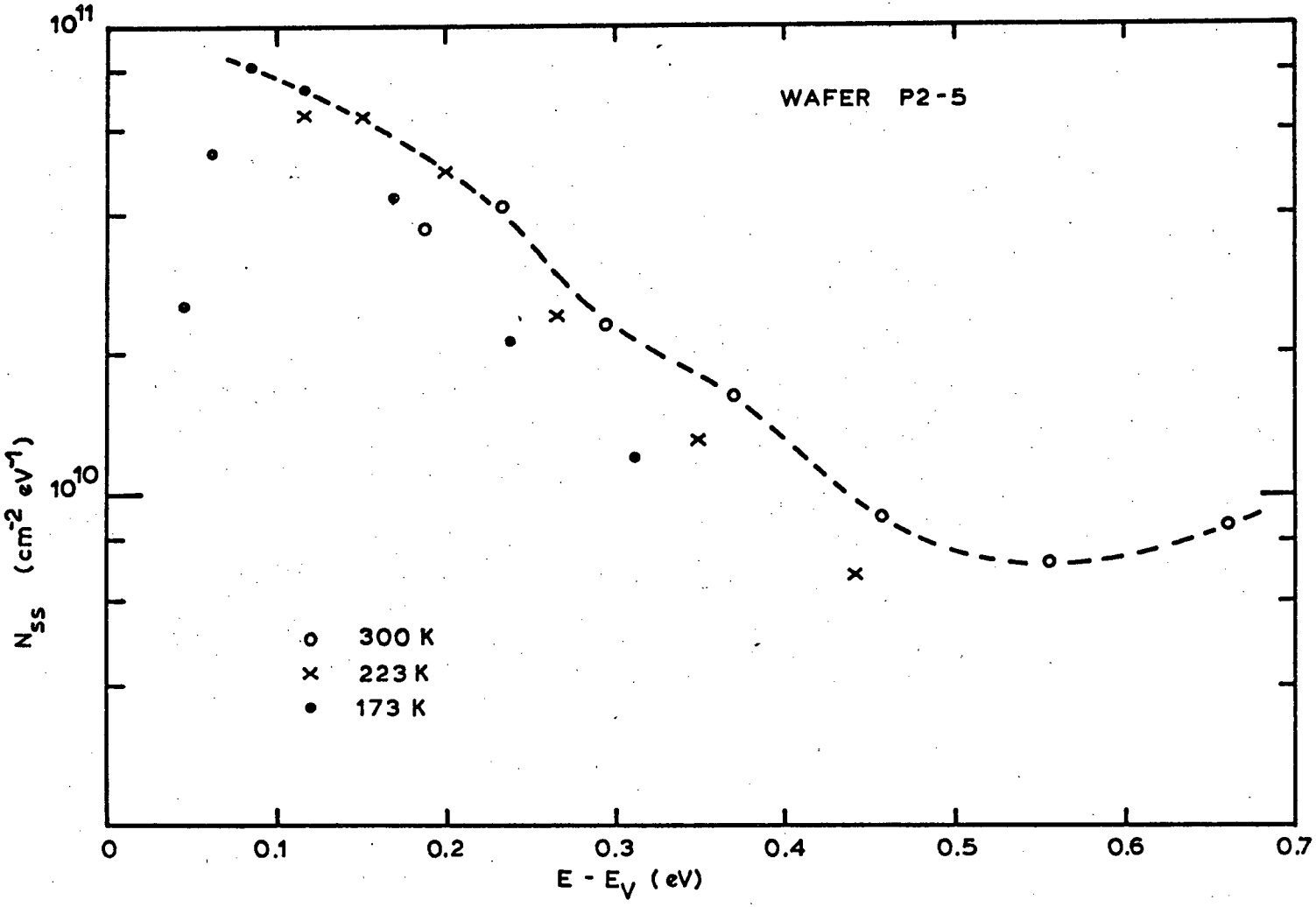


FIGURE 4.12 Interface state density for wafer P2-5 as a function of position in the band gap, obtained from C(V) measurements by the method outlined in Appendix C. The true interface state density is represented by the upper envelope of the three sets of data.

data from wafer P2-4. The contribution to the decay current from interface states was minimized by pulsing the gates from a condition of strong inversion. The measured bulk lifetime according to this method was 60 μsec.

Before any high voltage tests were conducted on the test devices the flat-band voltages for the gate and field oxides were determined by driving the photogate with a triangular waveform (±15 V) and observing the displacement current on an oscilloscope. Since the voltage ramps are linear the displacement current is proportional to the capacitance. The flat-band voltages for the thick and thin oxide regions may be determined separately by using strong illumination and measuring the voltages of the two inflections in the C-V curve, corresponding to inversion under the thick and thin oxides, respectively. To within the accuracy of the measurements, all of the devices tested, from a given wafer, were found to have the same flat band voltages. The gate and guard ring flat band voltages are given in Table 4.3. The flat band voltages did not change with time, even after the devices had been operated above breakdown, and cycled between room temperature and 80 K, many times.

4.1.6 Experimental Results and Discussion

The above breakdown testing was conducted at a temperature of 80 K. Examples of the driver waveform used and the typical output pulses obtained are illustrated in Fig. 4.13. The integration was started during the positive going ramp, slightly before crossing the breakdown voltage. The large drive pulse displacement currents due to capacitive coupling between the bonding pad and the substrate made it necessary to limit the ramp rate to 1×10^5 $Vsec^{-1}$ in order not to saturate the current amplifier.

In accordance with the theoretical discussion in chapter 3, it was found that when the devices were pulsed into deep depletion from a reset

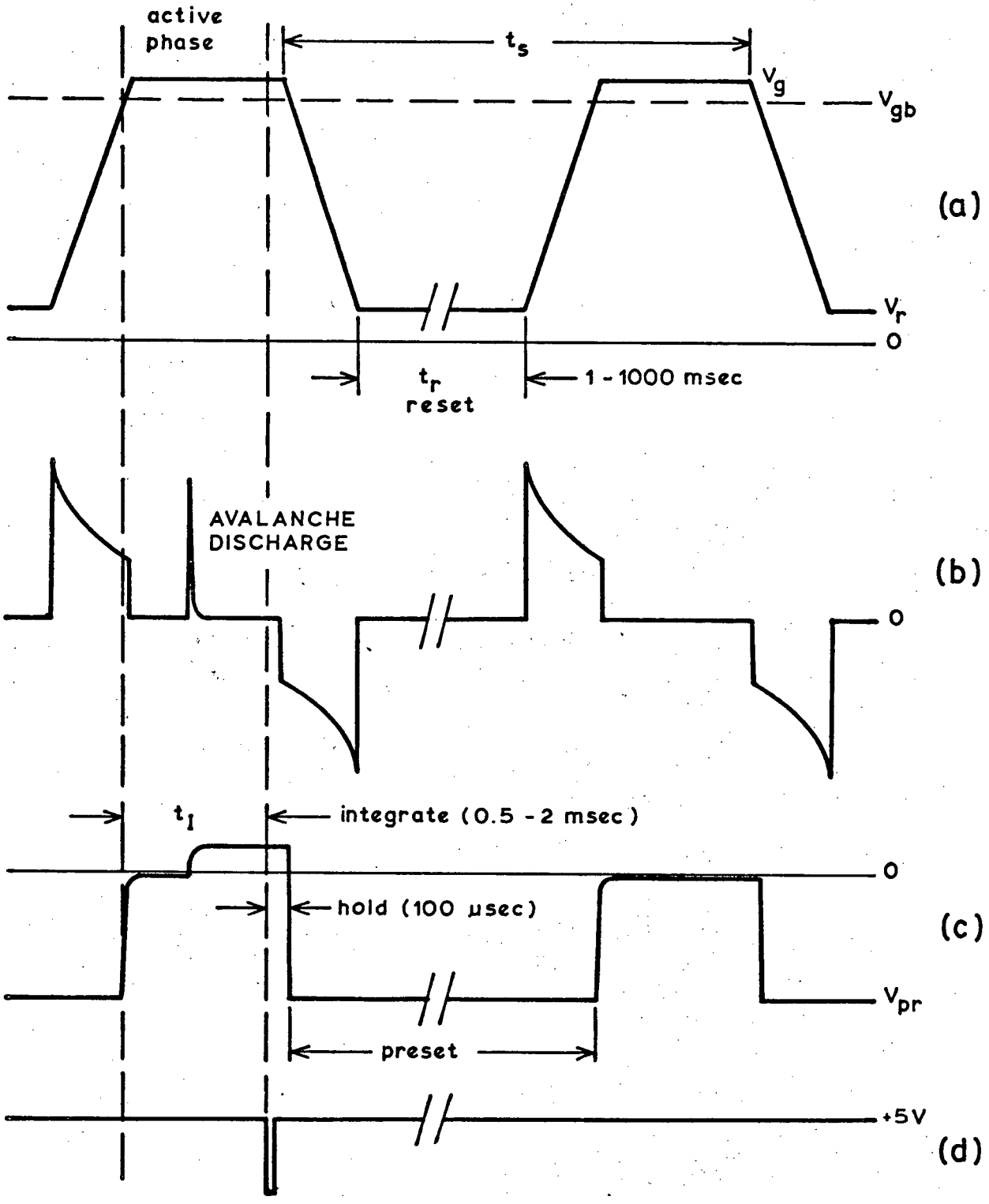


FIGURE 4.13 Test waveforms for the surface-breakdown, charge-injection devices

(a) high voltage driver
(b) output of the current to voltage amplifier. Substrate displacement current
(c) output of the preset integrator
(d) discriminator output, threshold = 0 V

condition corresponding to surface accumulation, breakdown always occurred during the voltage ramp, within a few microseconds of crossing the breakdown voltage. Once initiated, the breakdown continued for the remainder of the ramp duration. Pulsing the diodes from a reset condition corresponding to inversion (under the thin oxide region of the gate) increased the mean delay time to breakdown by about a factor of 3 and resulted in discrete breakdown pulses. However, not until the silicon under the thick oxide guard ring was also inverted during reset could appreciable delays to breakdown be achieved, enabling the devices to be pulsed several volts above breakdown.

When operating in the latter mode the reset inversion layer charge present under the guard ring, bonding pad and interconnect line is transferred to the thin oxide gate region during the depleting pulse, thereby, increasing the gate potential required to cause breakdown in the underlying silicon. Since no inversion layer charge remains under the guard ring after the depleting pulse, the increased gate potential causes the guard ring to be more deeply depleted. These effects are shown in Figures 4.14 and 4.15, for devices from the two fabrication runs. The silicon surface potential under the guard ring, calculated using (4.6) and the doping profile data in Figures 4.10 and 4.11 is also indicated. The change in breakdown voltage V_{gb} for a given change in reset voltage, when the guard ring is heavily inverted during reset, is given by

$$\Delta V_{gb} = \Delta V_r \left\{ 1 + \left(\frac{A_f d_g}{A_g d_f} \right) \right\} \qquad (4.7)$$

where d_g = gate oxide thickness

d_f = field oxide thickness

A_g = area of the thin oxide gate region

A_f = area of the guard ring, bonding pad and interconnect line

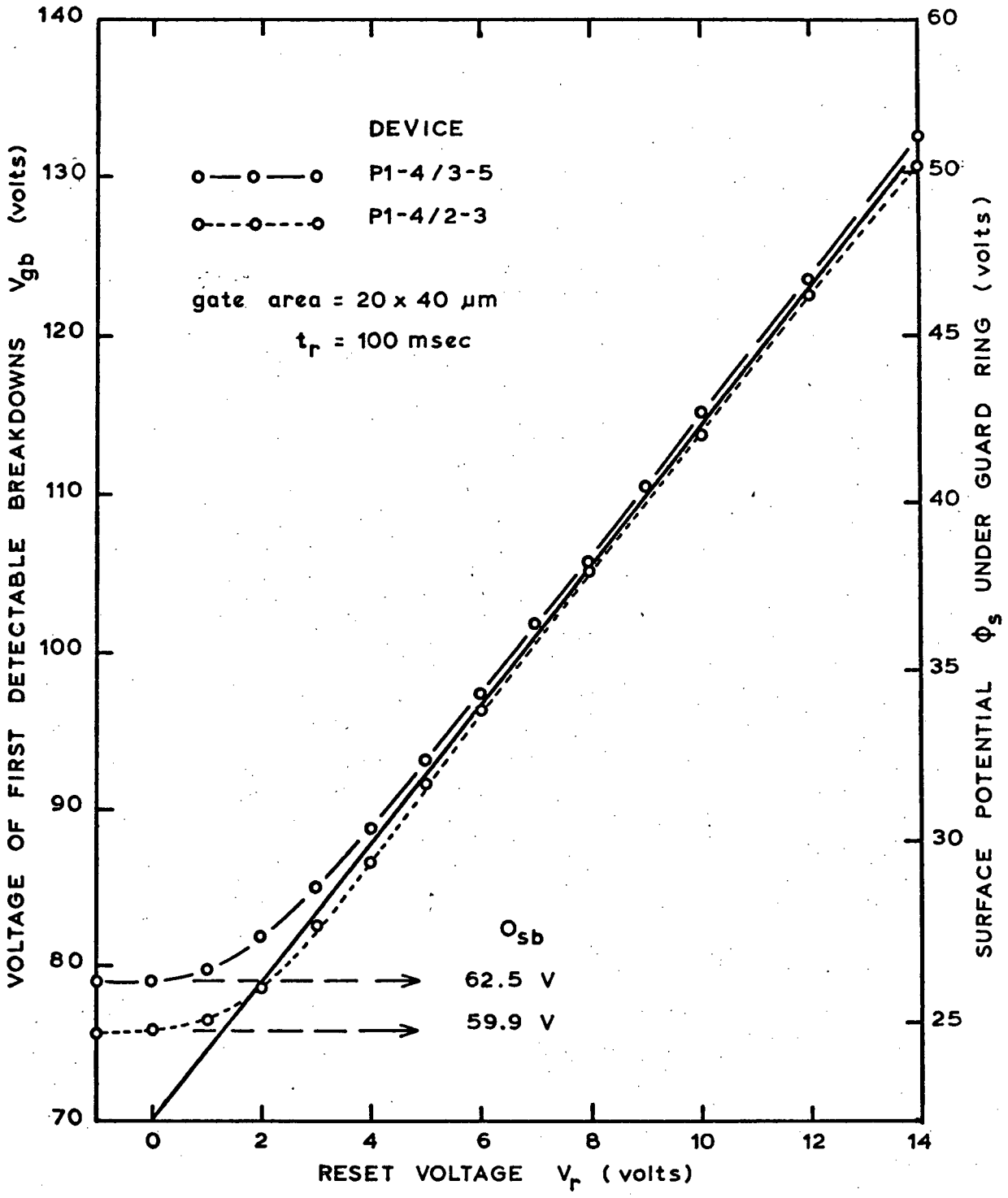


FIGURE 4.14 Voltage of first detectable breakdowns as a function of the reset voltage for devices from the first fabrication. The solid line indicates the calculated slope for the case that the guard ring is strongly inverted during reset. The silicon surface potentials under the guard ring and under the active region of the gate are also indicated.

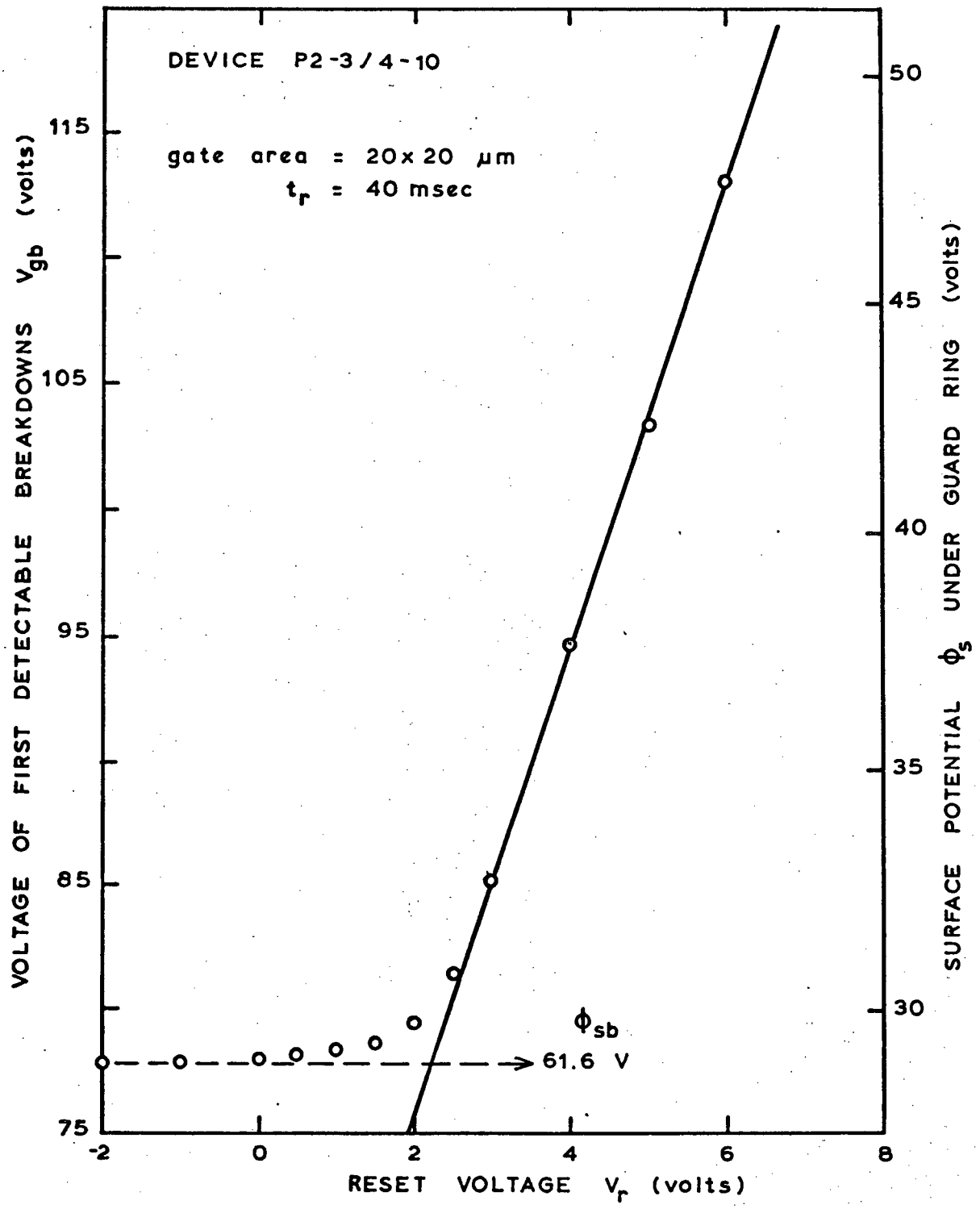


FIGURE 4.15 Voltage of first detectable breakdowns as a function of the reset voltage (second fabrication).

The calculated slope $\Delta V_{gb}/\Delta V_r$ from (4.7) is indicated by the solid line in Figures 4.14 and 4.15. The close agreement between the calculated and observed slopes indicates that premature edge breakdown was not occuring. If it were, one would expect the slope $\Delta V_{gb}/\Delta V_r$ to be greater than that calculated from (4.7), since the guard ring becomes more deeply depleted (i.e., more effective in preventing edge breakdown) as the reset voltage is increased. With further increases in reset voltage, or as the field oxide thickness is reduced, a point is eventually reached where the silicon under the guard ring, bonding pad, and interconnect line also break down. The thinner field oxide on wafers P2-0 and P2-2 prevented operation of these devices above the gate breakdown voltage when the guard ring was heavily inverted during reset.

Figure 4.16 shows the maximum charge per pulse as a function of the gate voltage for one of the 20 x 20 μm gates from wafer P2-3. The pulse height distribution (for $V_g - V_{gb} = 5$ V) is shown in Fig. 4.17. The charge measurements were obtained from the discriminator setting using the calculated gain of the current amplifier-integrator combination (9.36×10^5 elect. V^{-1}). The discharge pulses increase linearly with excess gate voltage, as expected, however, the maximum charge pulses are smaller than those predicted from

$$\Delta Q_I = A_g \frac{\epsilon_i}{d_{ox}} (V_g - V_{gb}) \tag{4.8}$$

and the pulse height distribution is considerably wider than anticipated. These effects are thought to be due to the build-up of positive space charge during an avalanche (e.g., holes drifting through the low field regions of the depletion region). The build-up of space charge limits the avalanche discharge current to the point where statistical fluctuations can cause the number of carriers in the high field region to drop to zero, thus terminat-

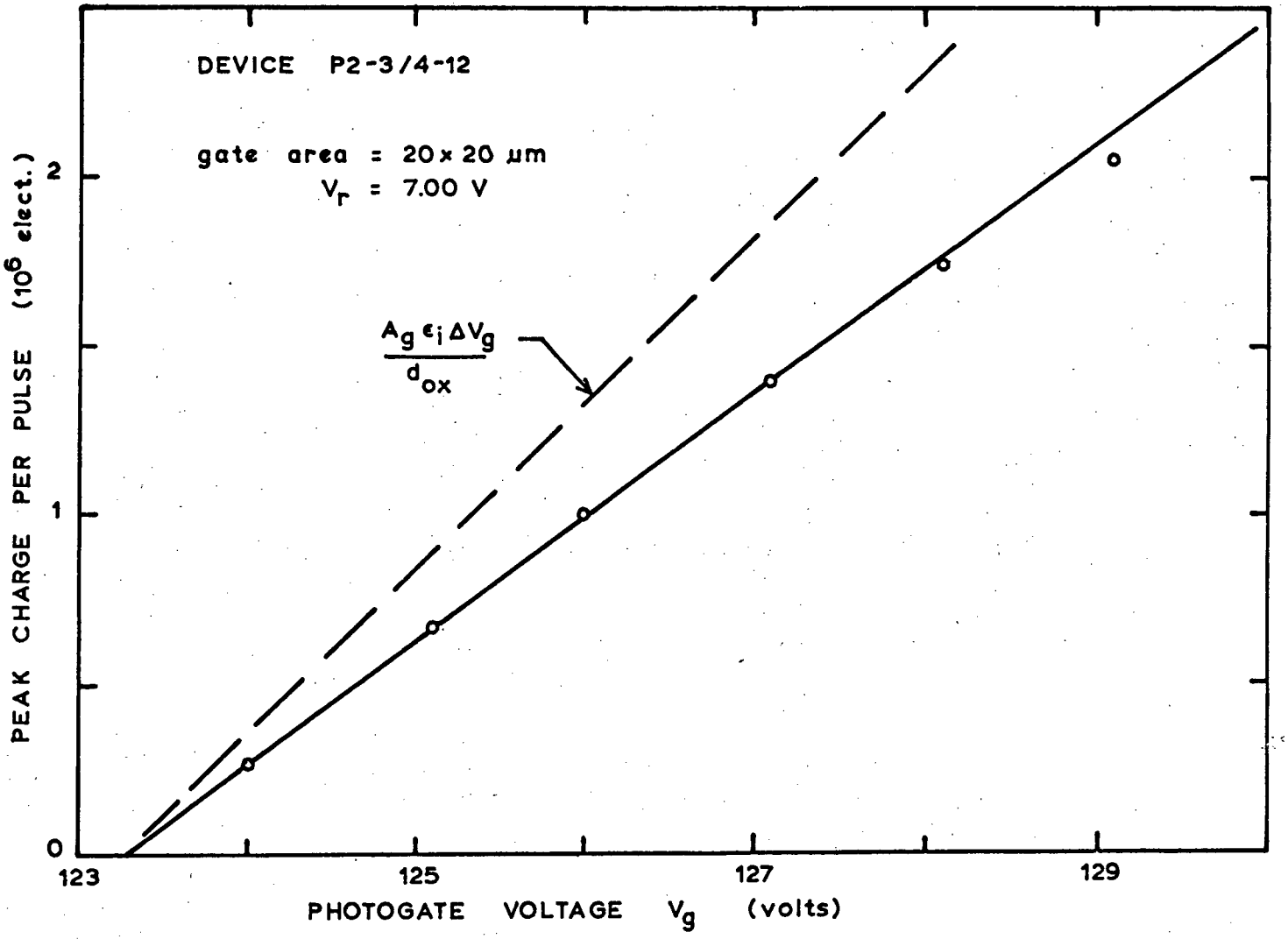


FIGURE 4.16 Maximum charge per pulse as a function of the photogate voltage. The dashed line indicates the expected variation, according to Eq. 4.8.

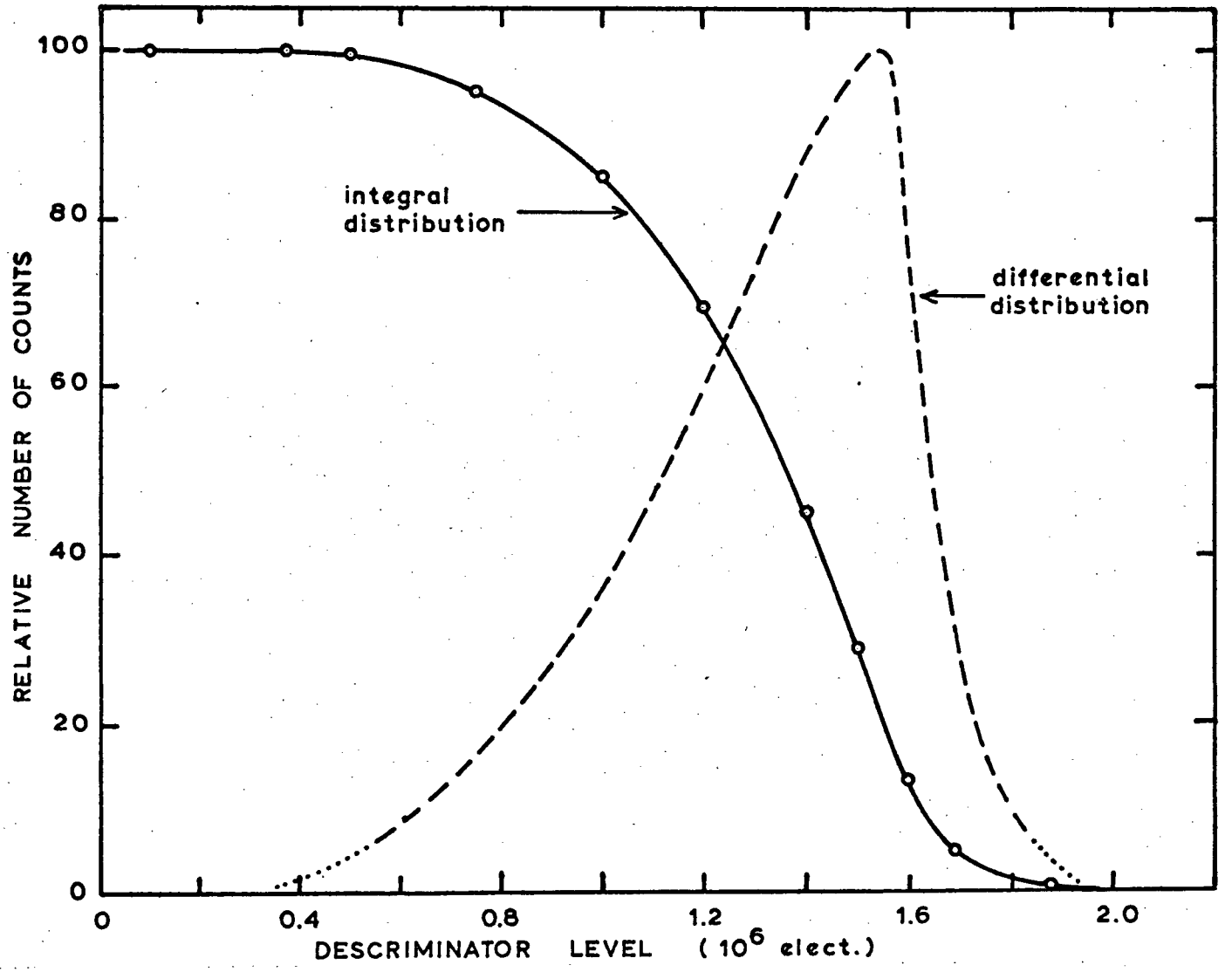


FIGURE 4.17 Typical pulse height distribution for the surface-breakdown devices. The differential distribution was obtained by graphically differentiating the integral distribution. $V_g - V_{gb} = 5V$.

ing the avalanche.

(1) Results From First Fabrication

Six test chips (78 devices) were examined from wafer P1-4. The dark count rates, at a given excess bias, were found to differ by less than a factor of 2, for all but a few devices which had anomalously high count rates. Except for these "bad" devices it was found that even after the guard ring was inverted during reset, the dark count rate at a given excess gate bias continued to decrease as the reset voltage (i.e., the inversion layer charge) increased. This decrease continued until the guard ring also started to break down. This effect is shown in Fig. 4.18 for one of the better devices from wafer P1-4. The silicon surface potential ϕ_s was calculated from (4.5) and (4.6). All of the count rates presented in this discussion have been corrected for dead time and temporal sampling effects (according to Eq. 2.7), and are expressed as counts/sec. The dark count rates shown in Fig. 4.18 are very high, increasing supralinearly as the gate bias is increased. 100 counts sec^{-1} for the 40 x 20 μm gate corresponds to 1.25×10^7 counts $sec^{-1}cm^{-2}$, more that 4 orders of magnitude larger than the desired maximum dark count rate of 500 $sec^{-1}cm^{-2}$.

In spite of the high dark event rate it was possible to operate these first devices under low level illumination in a photon counting mode. Fig. 4.19 shows a typical example of the count rate under illumination, plotted as a function of the excess gate bias. The devices were illuminated with a red gallium arsenide-phosphide LED (TIL 220, λ_{max}= 620 nm) run at very low current densities and pulsed on during the integration only. The entire back side of the chip was illuminated and no attempt was made to determine the level of illumination or to estimate the number of carriers generated, per light pulse, within a carrier diffusion length of the gate. Therefore,

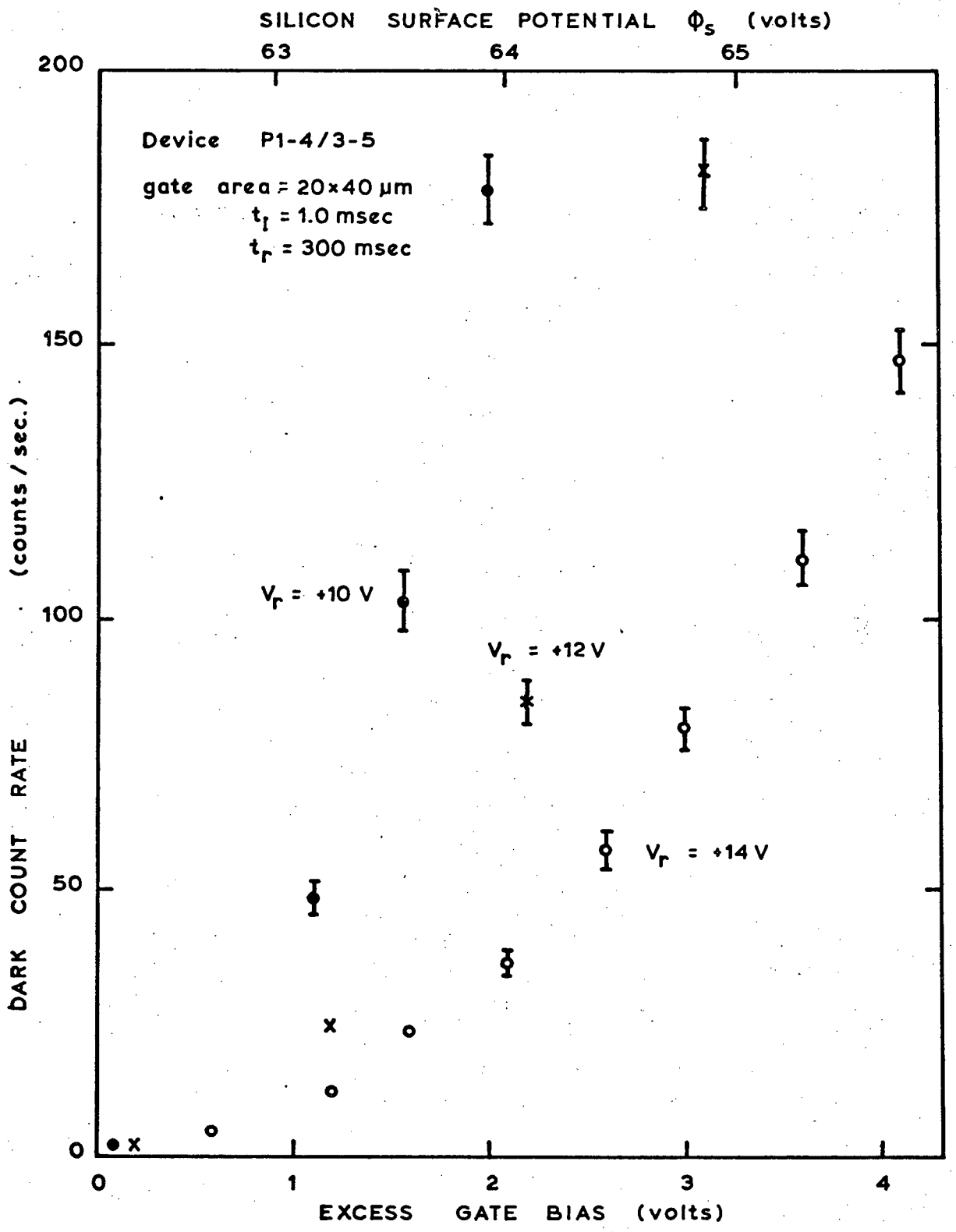


FIGURE 4.18 Dark count rate as a function of the excess photogate bias for the three reset inversion conditions, V_r = +10V, +12V, +14V. The calculated silicon surface potential before breakdown ϕ_s is also indicated.

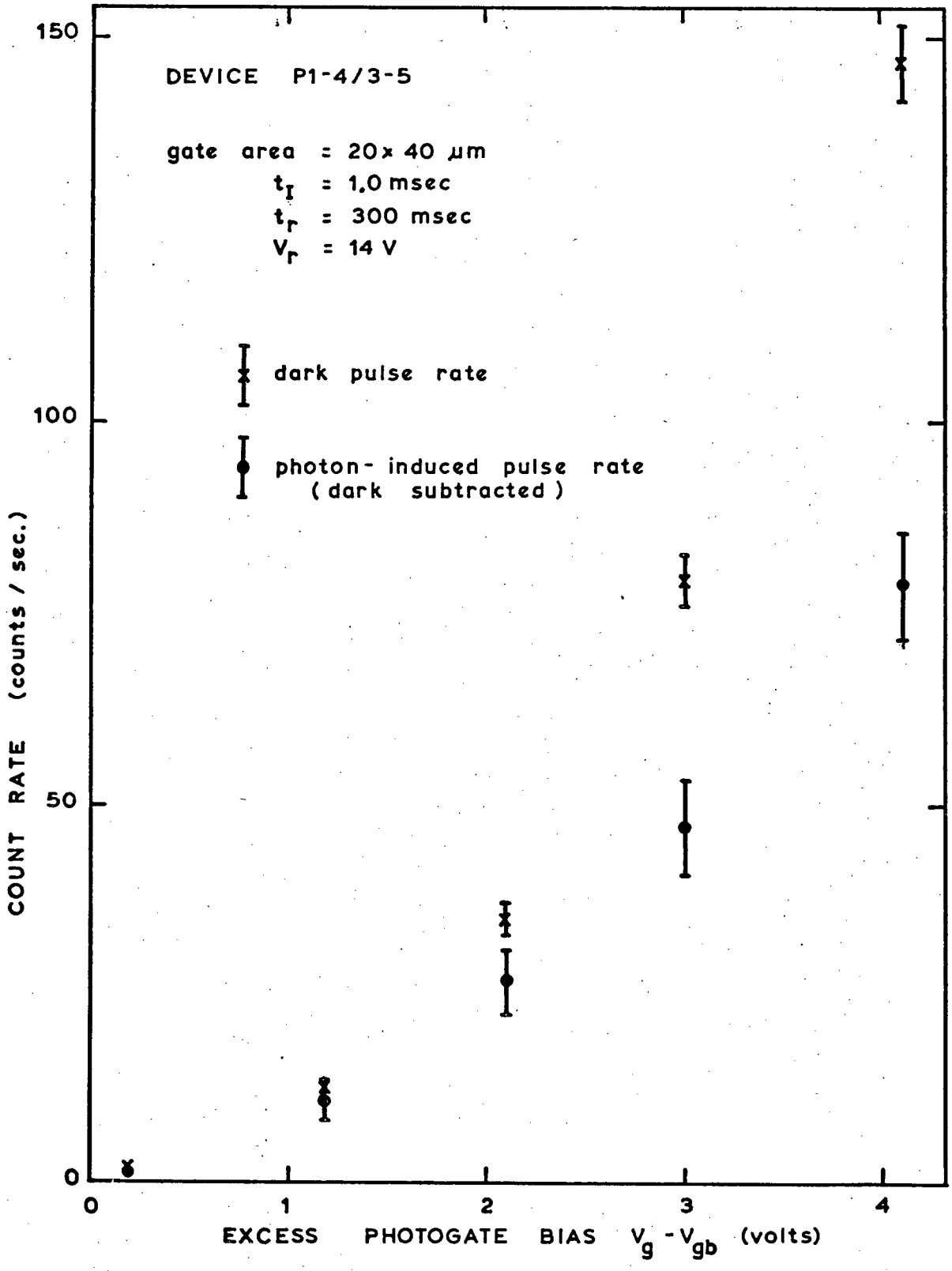


FIGURE 4.19 Dark and photon induced pulse rates as a function of excess photogate bias, for one of the better devices from the first fabrication.

no determination of the expected pulse rate could be made. The light-induced pulse rates were always observed to increase supralinearly with increasing gate bias, showing no sign of saturation.

In order to obtain consistent results when measuring the dark and photon induced pulse rates it was found necessary to operate the devices with a reset duration of 100 msec or longer. With shorter reset times there was a charging effect which caused a positive shift in the gate potentials required to generate a given size of breakdown pulse (i.e., to reach a given pre-breakdown surface potential ϕ_s). The shift in gate potential required to maintain a constant breakdown pulse size of 1×10^6 electrons, corresponding to an excess gate bias of approximately 1 volt, is plotted as a function of the cycle time t_s in Fig. 4.20. Sufficient illumination was used to ensure a breakdown every cycle. Assuming that a fixed amount of charge is trapped per cycle and that the amount of trapped charge decays exponentially with a time constant τ between avalanche discharge pulses, the shift in gate potential after many cycles is expected to be given by,

$$V_g(t_s) - V_g(\infty) = V_o \sum_{i=1}^{\infty} \exp -it_s/\tau \qquad (4.9)$$

where V_o is the shift in gate potential due to a single discharge. As indicated in Fig. 4.20 the observed shifts in gate voltage can be fit very closely by Eq. (4.9), with $V_o = 0.55$ V and $\tau = 40$ msec.

The charging effect shown in Fig. 4.20 was only observed when the devices were pulsed from a reset condition corresponding to depletion or inversion under the thin oxide region of the gate. When pulsed from accumulation no positive shift in the gate potential could be detected. Interface states cannot be responsible for this charge trapping since the inversion layer ensures that the only empty levels available to trap electrons lie very close to the conduction band. The detrapping time constant for these

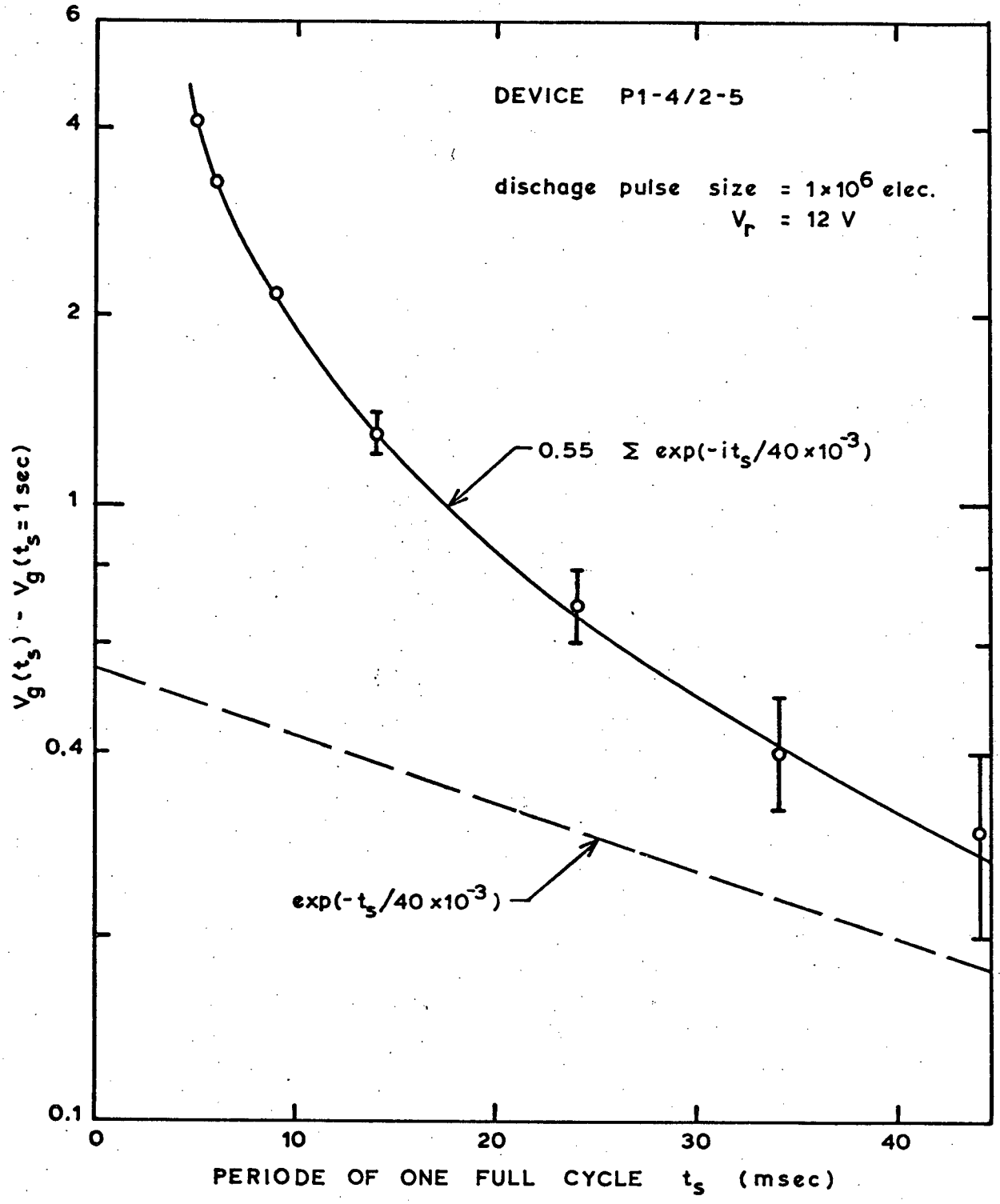


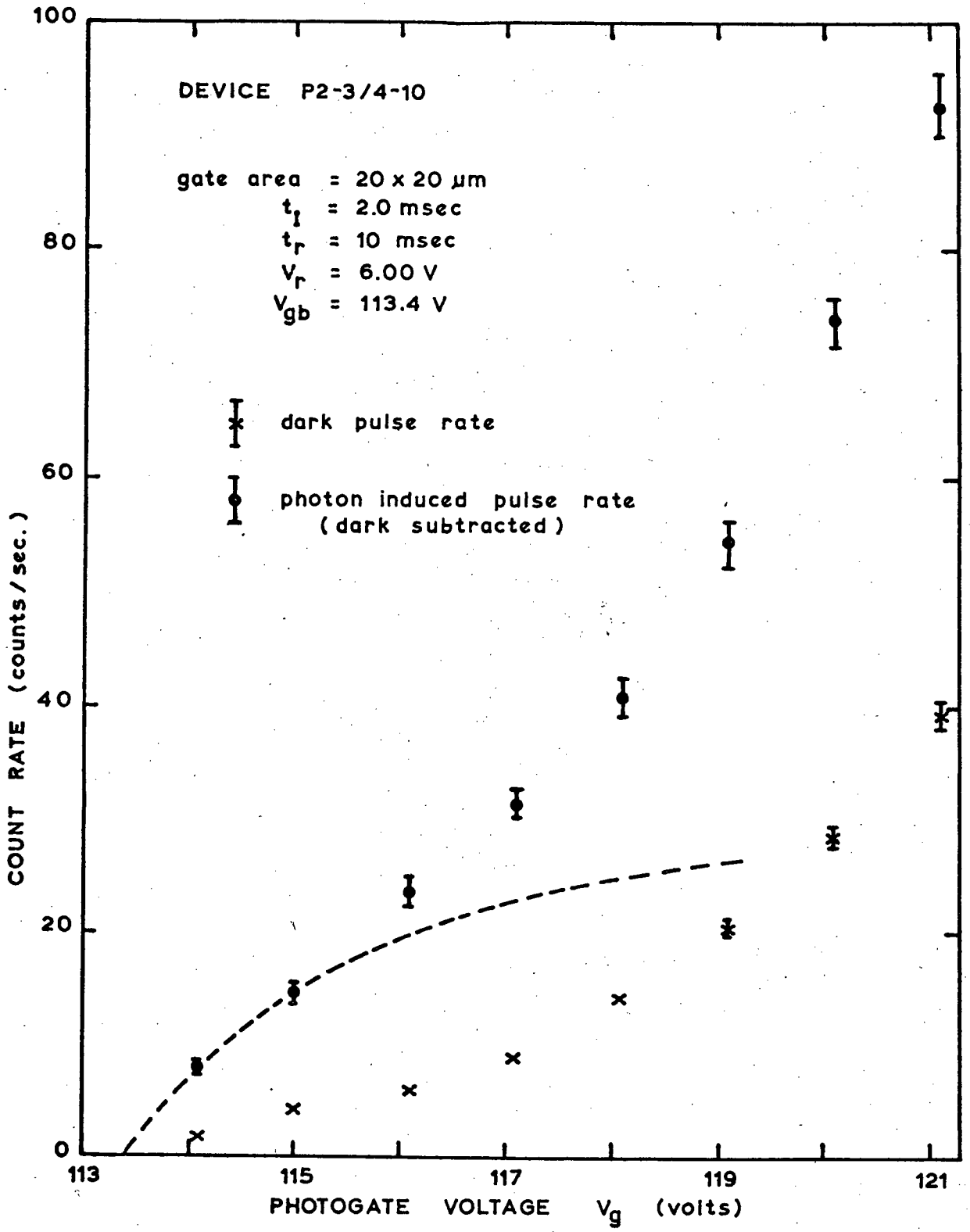
FIGURE 4.20 Shift in photogate potential required to maintain a constant output pulse size of 1 x 10^6 elect., as a function of the cycle time t_s. The solid line shows the best (visual) fit to Eq. 4.9. The dashed line shows the shift after one pulse.

levels is very short. This then suggests that the electrons are being trapped in the silicon near the interface, in the region that remains in depletion during reset (see section 3.2.6). However, with charge trapping of this magnitude, and a detrapping time constant of 40 msec, a large increase in the avalanche pulse rate would be expected as the reset time is decreased. Such an effect was not apparent. Very little, if any, increase in pulse rate could be detected when going from a cycle time of 1 sec to 5 msec. For this reason it is believed that the charging effect was due to electron injection into oxide traps near the $Si-SiO_2$ interface.

(2) Results From Second Fabrication

Five test chips (65 devices) were examined from wafer P2-3. The dark count rates at a given excess bias were approximately an order of magnitude lower than those obtained with the devices from wafer P1-4, however, the variations in dark count rate from device to device were larger. No devices were found with anomalously high count rates and a build-up of negative charge at the $Si-SiO_2$ interface, when operating above breakdown, was not observed with these devices. No shift in the gate potential required for a given pulse size could be detected for cycle times as short as 3 msec.

The dark and photon-induced pulse rates for two of the better devices are shown in Figures 4.21 and 4.22 . The devices had not been thinned so that the rear illumination provided virtually pure electron injection from the neutral bulk. The photon induced count rate in this case should follow the electron avalanche initiation probability $P_e(w)$. The theoretically determined probability $P_e(w)$, calculated according to (3.2) - (3.5) and the ionization rate data of Appendix A, is also shown in Figures 4.21 and 4.22. The electric field $\mathcal{E}(y)$, was calculated from equations (4.2) - (4.5),



FIGURES 4.21 and 4.22 Dark and photon induced pulse rates as a function of the photogate bias for two of the better surface-breakdown devices from the second fabrication. The dashed line indicates the calculated variation of $P_e(w)$, arbitrarily fit through the first two experimental points.

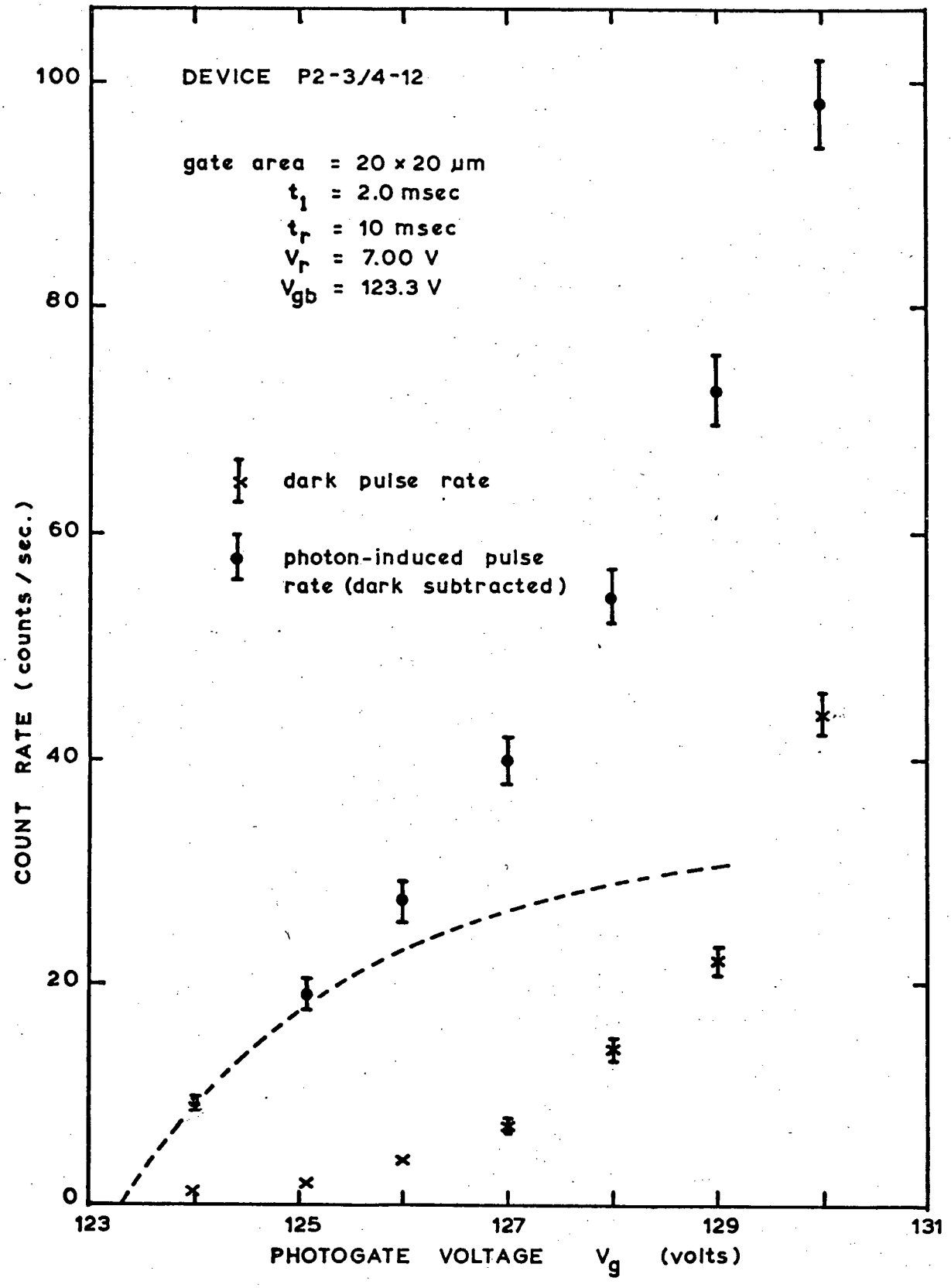

DEVICE P2-3/4-12
gate area = 20 x 20 μm
t_1 = 2.0 msec
t_r = 10 msec
V_r = 7.00 V
V_{gb} = 123.3 V
dark pulse rate
photon-induced pulse rate (dark subtracted)
COUNT RATE (counts / sec.)
0
20
40
60
80
100
123
125
127
129
131
PHOTOGATE VOLTAGE V_g (volts)

FIGURE 4.22

using the doping profile data in Fig. 4.11. Since no absolute experimental triggering probabilities were determined, the theoretical curves were arbitrarily fit from the observed breakdown voltage (voltage of first detectable breakdown pulses) through the first two experimental points. The remaining points deviate by an increasing amount from the theoretical curve. This indicates a large degree of re-triggering, due to either charge trapping or impact ionization of traps, during the periods of avalanche discharge, as discussed in section 3.2.6. Provided the traps are not charged to saturation the degree of re-triggering should be roughly proportional to the product of the count rate and the size of the discharge pulses, leading to the type of rapidly increasing count rate observed.

The estimated bulk lifetime of 60 μsec reflects predominantly the mid gap trap densities and thus, may not be an indication of the density of those traps responsible for the re-triggering. By taking count rate data at different temperatures it may have proved possible to characterize these traps, however, the high, and varied, fields present in the depletion region would greatly complicate the interpretation of such data. It was not thought that the surface breakdown devices warranted such an investigation, since the likelihood of a successful surface channel PC-CCD is small due to interface state effects. The need to maintain the periphery of the active region in inversion during reset, and the very large inversion layer charge required, severely complicates the design of a low dark count rate surface channel PC-CCD. The low mid gap interface state density of $7 \times 10^{9} cm^{-2} eV^{-1}$ measured on test wafer P2-5 indicates that the annealing treatment used was nearly optimum and that the observed behaviour was not due to an exceptionally large density of interface states. Any further reduction in the interface state density of a PC-CCD, using existing techniques, would be at best an order to magnitude.

4.2 BULK BREAKDOWN DEVICES

By introducing a thin layer of n-type silicon between the substrate and the insulator of a p-substrate MOS gate it is possible to form a potential minimum in the n-layer, away from the $Si-SiO_2$ interface (Fig. 4.23). In order to accomplish this it is essential for the p-n junction to be maintained in reverse bias so that the thin layer of n-type material would be fully depleted in the absence of any stored charge. This vertical structure forms the basis for a buried channel CCD, in which separate n-regions define the individual transfer channels. The reverse bias is provided by a d.c. contact at the output end of each n-type channel.

As was the case with a surface channel CCD it is possible to pulse the individual gates of a buried channel CCD so as to cause the p-type bulk to be deeply depleted, beyond the point where breakdown would normally occur. With a sufficiently thick n-layer the potential minimum will still be located away from the interface in this deeply depleted condition, so that, once triggered, the avalanche discharge will be confined to the bulk n-p junction. The fields at the surface of the semiconductor (where the bands bend upwards), and in the insulator, can be kept relatively low provided the n-layer has the correct thickness and doping density. Figure 4.23 shows the potential distribution perpendicular to the interface for a bulk breakdown MOS gate, and defines the various potentials and distances that will be referred to. Like the surface breakdown devices, this structure is self-quenching since the electrons generated during the avalanche collect in the potential well, thereby reducing the potential difference across the n-p junction.

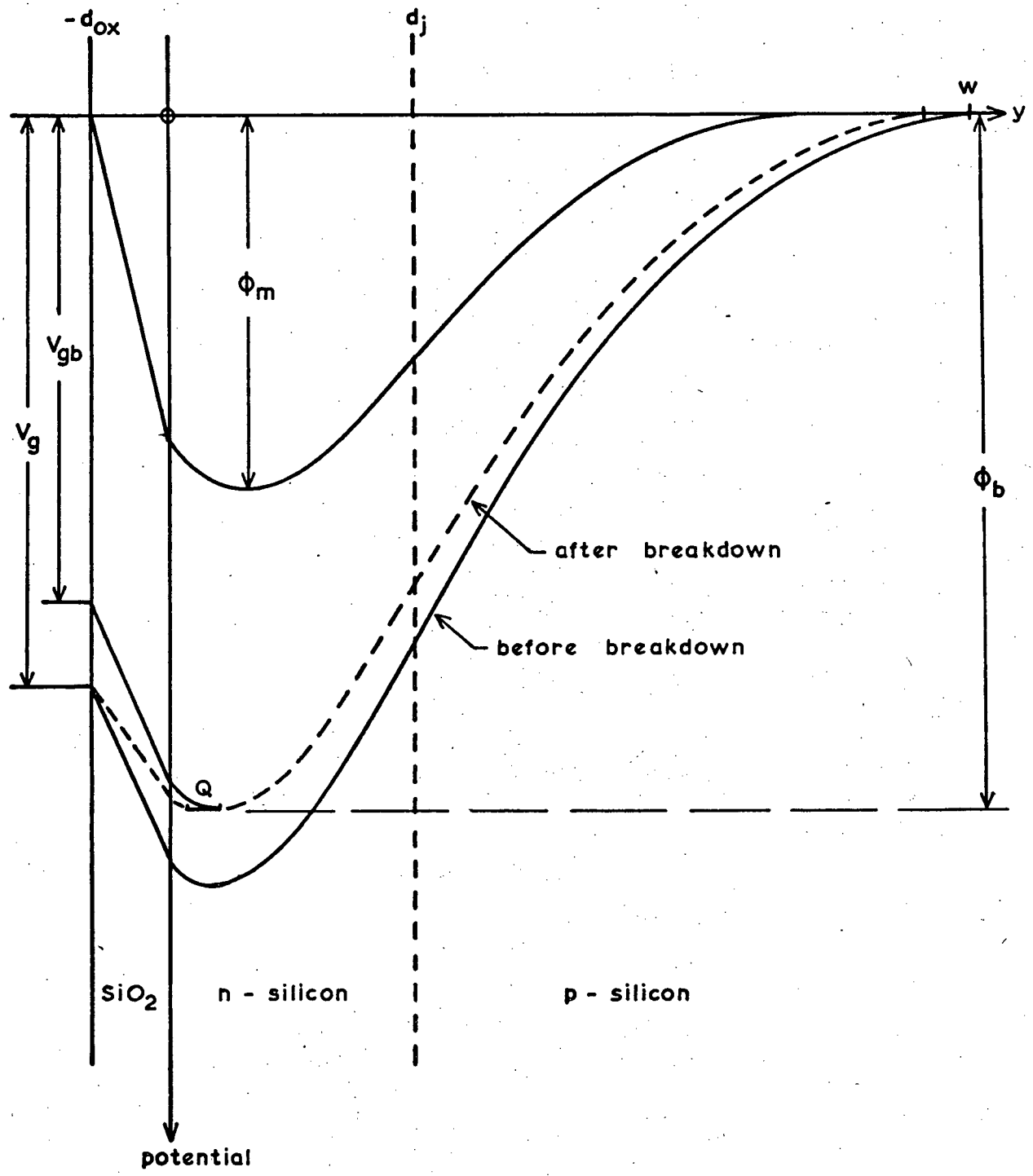


FIGURE 4.23 Potential distribution perpendicular to the surface for a bulk-breakdown MOS gate, before and after breakdown, at breakdown, and during reset. The compensated region containing the stored charge Q resulting from an avalanche discharge is also indicated.

In addition to the much lower oxide field strengths, and the fact that breakdown now occurs in the bulk of the semiconductor away from the interface, the bulk breakdown MOS structure has the further advantage that carriers generated via interface states can no longer trigger an avalanche. Holes generated at the interface migrate laterally along the silicon surface and escape out of the sides of the channel to the p-type bulk where they recombine. Electrons generated via interface states drift through the low field surface region and are collected in the potential well. Only if carriers are generated on the bulk side of the potential minimum can an avalanche be triggered. For this reason back-side illumination is mandatory for a buried channel PC-CCD.

Inherent in the buried channel CCD structure is the ability to control the bulk fringing fields and reduce premature edge breakdown. In order to fully deplete the n-channel the individual transfer gates must extend completely across the channel, and out over the p-type substrate. When pulsed above breakdown these gates are positive with respect to the substrate, and hence cause the p-silicon adjacent to the n-channel to be depleted, thereby lowering the bulk fringing fields. Reducing the oxide thickness increases the gate voltage required for breakdown, which in turn causes the p-type silicon on either side of the channel to become more deeply depleted. Provided the lateral n-p transition is not too abrupt, it should be possible to select an oxide thin enough to prevent edge breakdown along the sides of the n-channels, but still sufficiently thick to prevent breakdown in the p-type substrate at the edges of the gates. As with the surface channel CCD, high fringing fields in the charge transfer direction can be avoided by keeping the adjacent transfer gate only slightly below breakdown.

4.2.1 Design Consideration and Equations

Although the bulk n-p junction will generally not be abrupt, the restrictions on the peak electric field in the depletion region, and on the substrate doping, are roughly the same as were determined for the surface breakdown devices. A substrate doping of $N_A < 7 \times 10^{15} cm^{-3}$ was therefore also adopted for the bulk breakdown devices. Several factors determine the appropriate thickness and doping of the n-layer. It is desirable to minimize the dopant concentration (per cm^3) in this layer as far as possible in order to facilitate long minority carrier lifetimes and low trap densities, but, at the same time, the dopant dose (per cm^2) must be high enough that the potential minimum is located away from the interface when the gate is deeply depleted. On the other hand, if the dopant dose is too great the fields at the interface will be very high during charge transfer, and the silicon may break down at the surface. It is also desirable to minimize the junction curvature at the edges of the n-region (i.e., use a deep n-layer) as this relaxes the problem of controlling edge breakdown along the sides of the n-channel. Also a high junction curvature may cause the n-channel to breakdown prematurely at the output and, before the channel depletion voltage is reached.

In order to arrive at some actual quantitative specifications for the n-layer and the oxide thickness, initial guesses were made on the basis of the above considerations, and then modified according to the results of one dimensional calculations of the potential and field distributions normal to the surface. The required gate voltages in deep depletion were determined by calculating the avalanche initiation probability $P_e(w)$ according to equations (3.2) - (3.5) and the ionization rate data in Appendix A. The equations required to calculate the potential and field distributions were obtained as follows.

The potentials and distances used in the following derivation are defined in Fig. 4.23. Using the depletion approximation, the Poisson equations for the insulator, the n-type top layer and the p-type substrate become

$$\frac{d^2\phi_{ox}(y)}{dy^2} = 0 \quad , \quad -d_{ox} \leq y \leq 0 \tag{4.10}$$

$$\frac{d^2\phi_{si}(y)}{dy^2} = -\frac{\rho(y)}{\epsilon_s} \quad , \quad 0 \leq y \tag{4.11}$$

where $\rho(y)$ is the volume charge density in the silicon as a function of the distance from the $Si-SiO_2$ interface, and ϕ_{ox} and ϕ_{si} are the electrostatic potentials in the insulator and in the silicon, respectively. The electrostatic potential of the substrate is taken to be 0. The boundary conditions needed to solve (4.10) and (4.11) are

$$\phi_{ox}(-d_{ox}) = V_g \tag{4.12}$$

$$\phi_{ox}(0) = \phi_{si}(0) = \phi_s \tag{4.13}$$

$$\epsilon_i \frac{d\phi_{ox}(0)}{dy} = \epsilon_s \frac{d\phi_{si}(0)}{dy} \tag{4.14}$$

$$\phi_{si}(w) = 0 \tag{4.15}$$

$$\frac{d\phi_{si}(w)}{dy} = 0 \tag{4.16}$$

By using partial integration and the boundary conditions (4.12) - (4.16) $\phi_{ox}(y)$ and $\phi_{si}(y)$ can be expressed as

$$\phi_{ox}(y) = V_g - \frac{(y + d_{ox})}{\epsilon_i} \int_0^w \rho(y)\, dy \tag{4.17}$$

$$\phi_{si}(y) = \frac{y}{\epsilon_i}\int_w^y \rho(y)\,dy - \frac{1}{\epsilon_s}\int_w^y y\rho(y)\,dy \tag{4.18}$$

while the electric field in the silicon becomes

$$\mathcal{E}_{si}(y) = \frac{1}{\epsilon_s}\int_y^w \rho(y)\,dy \tag{4.19}$$

where w may be determined by an iterative procedure from boundary condition (4.13), which becomes

$$\frac{1}{\epsilon_s}\int_0^w y\rho(y)\,dy + \frac{d_{ox}}{\epsilon_i}\int_0^w \rho(y)\,dy - V_g = 0 \tag{4.20}$$

The position of the potential minimum y_m can likewise be determined from

$$\int_w^{y_m} \rho(y)\,dy = 0 \tag{4.21}$$

The normal procedure for generating the n-layer is to initially predope the surface by ion implantation (or a low temperature furnace predeposition) and then drive the dopant in at a high temperature. The resulting doping profile in this case is approximately Gaussian, and $\rho(y)$ becomes

$$\rho(y) = q\,N(0)\,\exp -(y/\beta)^2 - q\,N_A \tag{4.22}$$

where $N(0)$ is the surface concentration of the n-layer dopant (cm^{-3}) and N_A is the bulk acceptor doping density. The parameter β is related to the junction depth according to

$$\beta = y_j[\ln N(0) - \ln N_A]^{-\frac{1}{2}} \tag{4.23}$$

Approximating the doping profile by (4.22) enable the integrals appearing in (4.17) - (4.21) to be replaced by

$$\int_a^b \rho(y)\,dy = \left\{\frac{\sqrt{\pi}}{2} N(0)\,\beta\,\mathrm{erf}(y/\beta) - N_A y\right\}\Big|_a^b \tag{4.24}$$

$$\int_a^b y\rho(y)\,dy = \left\{N(0)\frac{\beta^2}{2}\left[1 - \exp -(y/\beta)^2\right] - \frac{N_A y^2}{2}\right\}\Big|_a^b \tag{4.25}$$

On the basis of the one dimensional calculations the following specificationswere determined for the oxide thickness, n-channel junction depth, and n-layer surface concentration to be used with a substrate doping of $N_A = 7 \times 10^{15}$ cm^{-3}:

$$d_{ox} = 0.5\ \mu m$$

$$y_j = 2.0\ \mu m$$

$$N(0) = 3.2 \times 10^{16} \text{ to } 4.0 \times 10^{16}\ \text{cm}^{-3}$$

An oxide thickness of 0.5 μm was chosen as this is the minimum thickness permissible due to processing constraints that are discussed in the next section. The junction depth was restricted to 2 μm because the maximum temperature of the drive-in furnace was limited to 1150°C, and to drive the n-layer in much further would have required very long drive-in times. Also, if the n-layer is deeper the surface concentration must be lower, and control over the uniformity of this diffusion becomes more difficult.

4.2.2 Test Structure Design and Fabrication

The bulk breakdown device used in this investigation is shown in Fig. 4.24. It is basically the same as the previous surface breakdown test structure shown in Fig. 4.4, except that a n-diffusion now defines the transfer channel and only a single thickness of oxide is used. The additional n+ and p+ diffusions are required in order to make good ohmic contacts to the n-channel and p-substrate. Because of the similarity of the two device structures it was possible to fabricate the bulk breakdown test

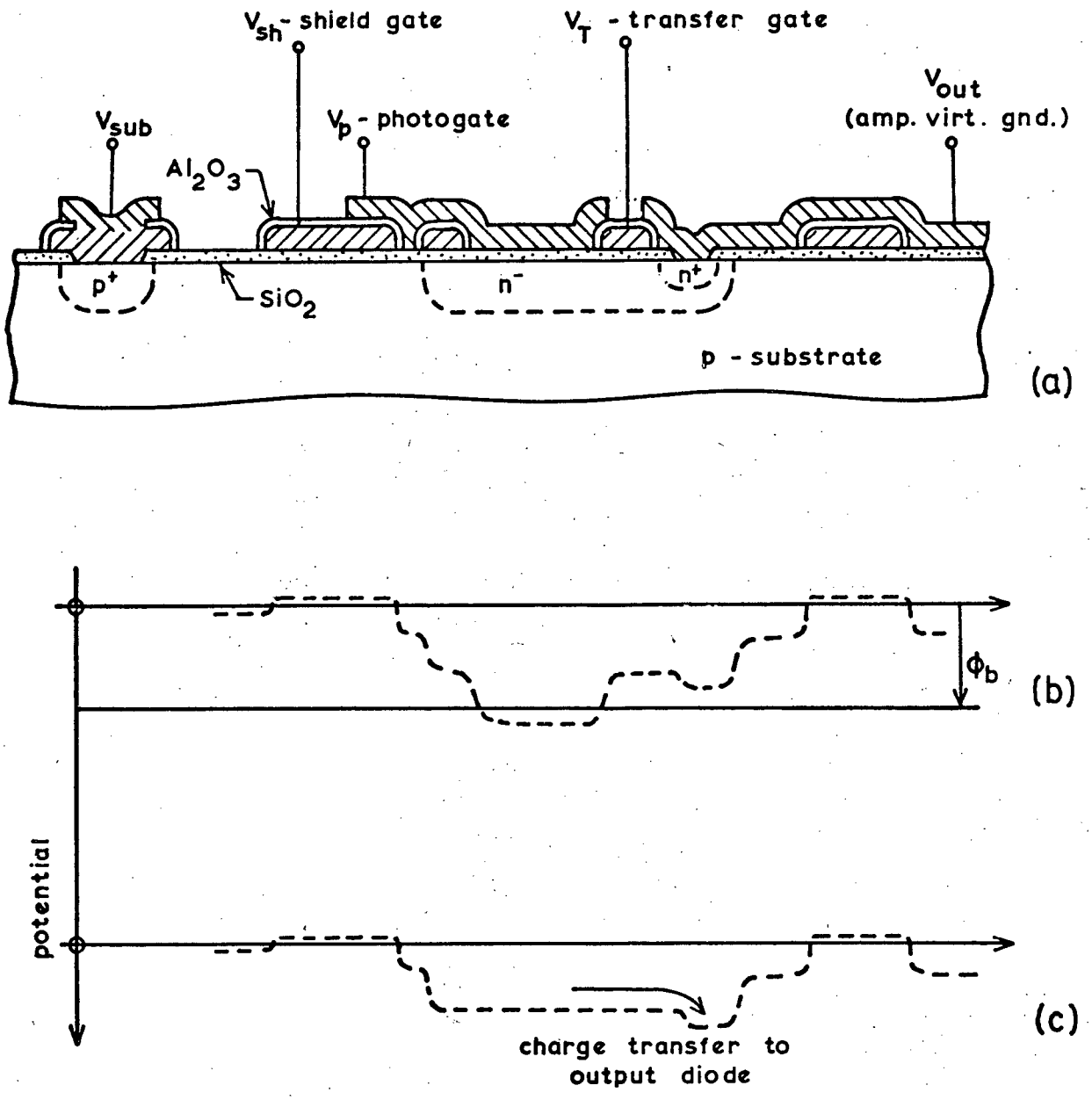


FIGURE 4.24 Bulk-breakdown, charge-transfer test device

(a) vertical structure. Layout is identical to the surface-breakdown test device shown in Fig. 4.4
(b) potential well diagram in deep depletion, before breakdown
(c) during reset, charge transfer

structure using the existing masks. Double exposures (to two different masks) were required to define the n+ and p+ diffusions and to etch the n+ and p+ contact windows. The fabrication sequence for the bulk breakdown devices is illustrated in Fig. 4.25 and the processing details are listed in Table 4.4. As before, the devices were fabricated on 2.2 - 2.5Ω cm boron doped (100) Czochralski wafers with chemo-mechanically polished front surfaces. The back sides, however, had a rough, sand-blasted surface finish. This heavily damaged layer of silicon getters lifetime-degrading impurities from the bulk silicon during device processing. The heavy phosphorus concentration introduced into the back side during the n+ contact predeposition also helps to getter impurities. All dry oxidations were carried out with HCl added to the oxidation atmosphere.

The fabrication again calls for an Al-Al_2O_3-Al double level metalization. However, there are no MIS contacts this time so that the contact sinter and hydrogen anneal may be performed as a last step in the device processing, thereby avoiding any anodization problems due to hillock formation in the first level of aluminum. Also, the interlevel vias are opened by selectively etching the Al_2O_3 anodic oxide layer [94], rather than masking these areas with positive resist during the anodization. The maximum formation voltage is then only restricted by the electrolyte used, and by the thickness of SiO_2 over the areas of the wafer that are exposed to the electrolyte but not covered with aluminum. Bare areas of silicon exposed to the electrolyte were avoided by opening the p+ and n+ contact windows in separate operations (see Fig. 4.25). The 25 w/o APB-EG electrolyte used can support anodization up to 350 V before side reactions start to occur [112]. A formation voltage of 220 V was chosen for device fabrication to ensure that the anodic oxide would support a 100 V potential difference between the two gate levels. The SiO_2 gate oxide must then be at least 0.5μm thick to pre-

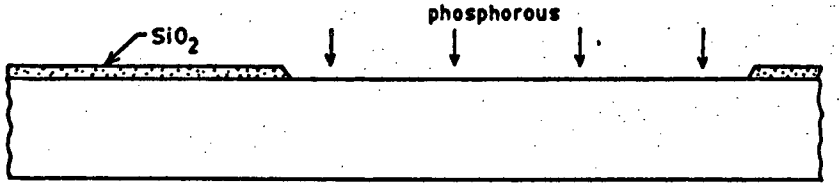
vent anodization of the silicon.

Considerable difficulty was encountered in controlling the low temperature phosphorous predeposition used to form the transfer channels. Initial tests, conducted to determine the best combination of predeposition temperature and time, indicated that achieving the desired dopant concentration would be somewhat of a hit or miss operation, due to the poor temperature stability of the phosphorous predeposition furnace after loading the wafer boat. Ideally the channel should have been pre-doped by ion implantation and then simultaneously annealed and driven in; however, an ion-implanter was not readily available. Instead an attempt was made to fabricate some successful test devices by using a furnace predeposition, and processing several wafers with different doping times.

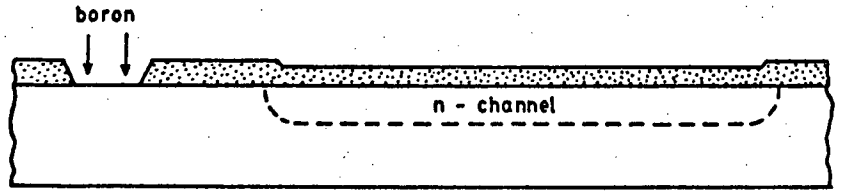
Nine wafers (M64 - M72) were processed. During the n-channel predeposition step the wafers were split into three groups and given 8, 12, and 20 minute predepositions as 780°C. Four point probe resistivity measurements, conducted after the drive-in and the removal of all masking oxides, indicated that only wafers M65, 66, 72 (20 min. predep.) had received sufficient phosphorous dopant, and that even these had a more lightly doped n-channel than desired. The incorrectly doped wafers were continued along with the good wafers and used for testing during the aluminum anodization and via etch. M72 was inadvertently destroyed while spin drying, during a later RCA clean.

Selectively etching the anodic Al_2O_3 to form the interlevel vias, also proved to be very difficult. The negative resist used (Waycoat 200 negative) could not withstand the 80°C CrO_3/phosphoric acid etch for more than about 10-15 minutes before lifting, while the Al_2O_3 etch rate was much lower than expected, requiring approximately 10 min. for complete removal of the 220 V films. Also, the end point could not be determined visually:

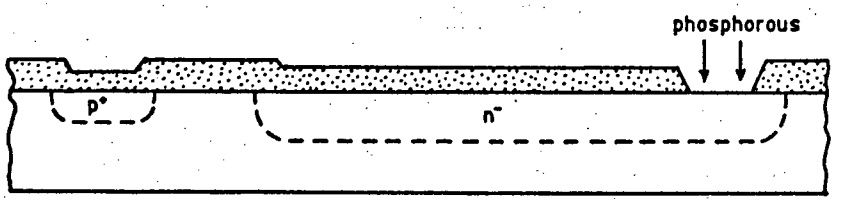
only by probing the via areas very gently and measuring the electrical resistance to the underlying aluminum was it possible to determine the completion of the etch. The good device wafers (M65,66) were given a 10 min via etch after which probing indicated that bare aluminum had been exposed. However, after depositing the second level of aluminum it was discovered that the Al_2O_3 had not been completely removed. Apparently the contact probe must have broken through a thin remaining layer of oxide. Unfortunately, since the n+ contact windows had already been opened, it was not possible to simply remove all the aluminum and Al_2O_3 and repeat the metalization procedure. Instead it was also necessary to remove the SiO_2 gate oxide and re-oxidize the wafers. This, however, consumed a significant portion of the n-channel and further reduced the channel doping (cm^{-2}). On the second attempt at etching the Al_2O_3 vias, the CrO_3/phosphoric acid etch was followed by an etch in straight phosphoric acid at $50^{o}C$ until bubbles could be observed (approximately 2 min), indicating that the underlying aluminum was being etched and that the Al_2O_3 had been completely removed. Only wafer M66 was successfully completed in this way. The resist lifted on M65 during the CrO_3/phosphoric acid etch. Table 4.5 lists the various thickness and doping parameters measured on wafer M66 at the completion of the processing. The final n-layer dopant dose (cm^{-2}) was considerably lower than desired but it was felt that the devices would still be operational above breakdown.



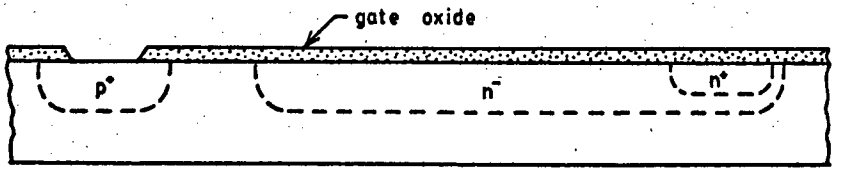
1. masking oxidation - phosphorous predep.



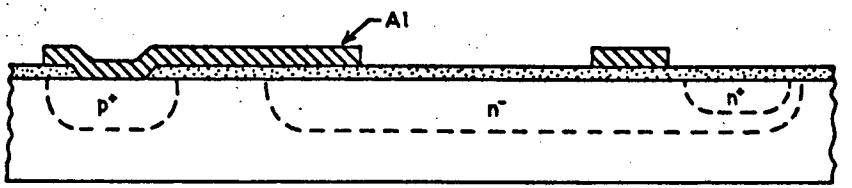
2. n-channel drive in an oxidation - boron predep.



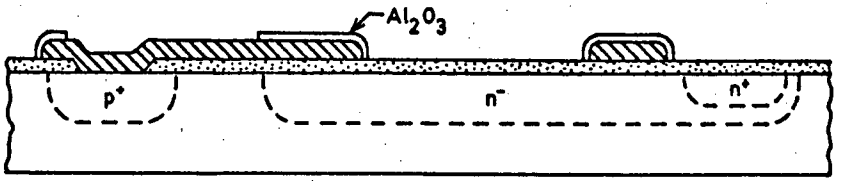
3. oxidation - phosphorus predep.



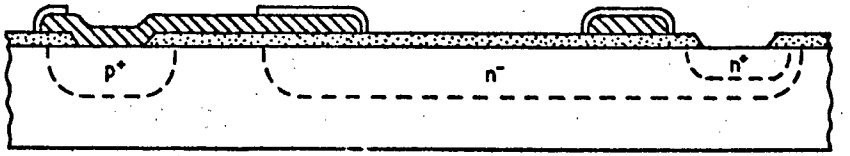
4. strip all oxides - gate oxidation - open substrate contacts.



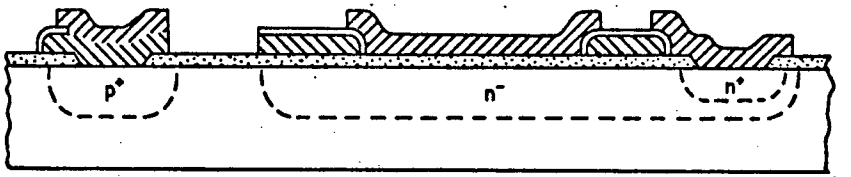
5. deposit and pattern first level of Aluminum



6. Aluminum anodization and via etch.



7. open n^+ channel contacts.



8. deposit and pattern second level of Aluminum plus H_2 anneal.

FIGURE 4.25 Fabrication sequence for the bulk-breakdown test devices

TABLE 4.4 Processing Details For The Bulk Breakdown Devices

Step	Operations	Details
	initial clean	- hot xylene with ultrasonic agitation - RCA clean
	n^- masking oxide	temp. 1150°C cycle: 30 min O_2 5 min N_2 oxide thickness, 0.10 μm
	channel window etch	- neg. resist, mask 1 - buffered HF - resist strip, hot H_2SO_4/H_2O_2 - RCA clean
1	channel predep.	temp. 780°C source $POCl_3$ pass over, 15°C boat and tube predoped at 1050°C for 1 h slices M67, 69, 70 cycle: 5 min N_2 + 3% O_2 8 min N_2 + 3% O_2 + Source N_2 5 min N_2 + 3% O_2 slices M64, 68, 71 cycle: 5 min N_2 + 3% O_2 12 min N_2 + 3% O_2 + Source N_2 5 min N_2 + 3% O_2 slices M65, 66, 72 cycle: 5 min N_2 + 3% O_2 20 min N_2 + 3% O_2 + Source N_2 5 min N_2 + 3% O_2 - phosphorus glaze strip, 10% HF, 15 sec. - HCl/H_2O_2 part of RCA clean

TABLE 4.4 cont'd.

Step	Operations	Details
	n^- drive-in and oxidation [a]	temp. 1150°C cycle: 245 min O_2 + 3% HCl 40 min O_2 + H_2 10 min N_2
2	p^+ diff. window etch	- neg. resist, masks 2 + 5 (double exposure) - buffered HF - resist strip, hot H_2SO_4/H_2O_2 - RCA clean
	p^+ pre-dep	temp. 1000°C source BBr_3 pass over, 15°C cycle: 5 min N_2 + 3% O_2 5 min N_2 + 3% O_2 + Source N_2 30 min N_2 + Source N_2 5 min N_2 - boron glaze strip, HF/H_2O(1:1), 30 sec - HCl/H_2O_2 part of RCA clean
	p^+ drive-in and oxidation [a]	temp. 1100°C cycle: 5 min O_2 120 min O_2 + H_2 5 min N_2
3	n^+ diff. window etch	- neg. resist, masks 2 + 3 (double exposure) - buffered HF - resist strip, hot H_2SO_4/H_2O_2 - RCA clean

TABLE 4.4 cont'd.

Step	Operations	Details
3 cont'd	n+ pre-dep.	temp. $1050^{o}C$ source $POCl_3$ pass over, $15^{o}C$ cycle: 5 min N_2 + 3% O_2 20 min N_2 + 3% O_2 + Source N_2 15 min N_2 + 3% O_2
	strip all masking oxides	- HF/H_2O(1:1) - HCl/H_2O_2 part of RCA clean
	gate oxidation [a]	temp. $1150^{o}C$ cycle 3 min O_2 40 min O_2 + H_2 30 min O_2 + 5% HCl 10 min N_2
4	p^+ contact window etch	- neg. resist, masks 2 + 5 (double exposure) - buffered HF - resist strip, hot H_2SO_4/ H_2O_2
	etch off back side n^+ layer	- white etch (front protected with wax) - hot trichlorethylene - hot H_2SO_4/H_2O_2 - RCA clean
5	aluminum evap. I	- tungsten filament evap. substrate temp. $250^{o}C$ evap. rate 150 $\mathring{A}sec^{-1}$ - film thickness 1.2 μm
	aluminum etch I	- neg. resist, mask 3 - phosphoric/nitric etch, $60^{o}C$ - resist strip, Microstrip

TABLE 4.4 cont'd

Step	Operations	Details
6	anodization	: electrolyte, 25 w/o APB-EG current, 1 mA cm^{-2} formation voltage, 220 V soak time at 220V, 5 min
	via etch	- neg. resist, mask 4 - CrO_3/H_3PO_4 etch, 80°C, 10 min - phosphoric etch, 50°C, 2 min - resist strip, Microstrip
7	n^+ contact window etch	- neg. resist, masks 2 + 3 (double exposure) - buffered HF - resist strip, Microstrip
8	aluminum evap. II	- tungsten filament evap. substrate temp. 250°C evap. rate 150 Åsec^{-1} - film thickness 1.0 μm
	aluminum etch II	- neg. resist, mask 5 - phosphoric etch, 60°C - resist strip, Microstrip
	contact sinter plus H_2 anneal	temp. 400°C cycle: 60 min H_2

[a] Quartz furnace tube and boat pre-cleaned with O_2/HCl gas flow for 2 hours, prior to loading devices.

TABLE 4.5 Data for Wafer M66

Bulk resistivity (before processing)	2.42 Ωcm
Substrate doping [a]	5.8×10^{15} cm^{-3}
Channel doping [b] (surface concentration)	3.0×10^{16} cm^{-3}
Junction depth (from bevel and stain)	1.63 μm
Oxide thickness (from colour)	0.52 μm
Channel breakdown voltage	~65 V

[a] Determined from low temperature deep depletion C-V data using the shield gate of the operational devices (see Appendix C).

[b] Determined from four point probe resistivity measurements and the junction depth, assuming a Gaussian profile for the n-layer.

4.2.3 Two Dimensional Modeling of the Completed Devices

Unlike the surface breakdown devices it was not certain that the bulk breakdown devices could be operated in such a way as to completely eliminate premature edge breakdown. This is due to the limited voltage range over which the transfer gate can be operated. It must, at all times, be sufficiently negative with respect to the channel output to ensure complete depletion of the n-layer. The degree of depletion in the bulk silicon under the transfer gate is, therefore, limited by the breakdown voltage of the channel output diode.

In order to compare the experimentally observed pulse rates with those predicted from the theory it is necessary to have some idea of the potential distribution in the bulk breakdown structure and the uniformity of the breakdown voltage. The approximate two dimensional model shown in Fig. D1 (Appendix D) was used for this purpose. A zero gate separation was used and the transfer gate and n-layer were assumed to be infinite in extent. The later assumption is justified because variations in the doping profile and degree of depletion, at distances greater than 10 μm (the transfer gate width) from the edge of the breakdown region, have a negligible effect on the potential distribution under the photogate. A zero gate separation is justified since the actual gate separations (0.25 μm) are more than an order of magnitude less than the depletion region width and are not expected to significantly alter the potential distribution or the peak breakdown fields away from the $Si-SiO_2$ interface. The use of zero gate separations does, however, result in higher lateral fields at the surface of the silicon, in the transition region from one gate to the other. The potential distribution under the photogate (across its 20 μm width) and under the transfer gate on either side was calculated by the method described in Appendix D. The depletion approximation was used in order to linearize the two dimen-

sional Poisson equations and make possible an analytical solution. A Guassian doping profile was assumed, enabling the integrals to be replaced by (4.24) and (4.25).

The following parameters were used to model the devices from wafer M66

$T = 80$ K

$N_A = 5.8 \times 10^{15}$ cm^{-3}

$N(0) = 2.7 \times 10^{16}$ cm^{-3}

$d_{ox} = 0.52$ μm

$y_j = 1.63$ μm

N(0) has been adjusted slightly from the measured value (Table 4.5) in order to obtain agreement between the calculated and observed n-layer depletion voltage. Figure 4.26 shows the resulting potential distribution for V_g = 100 V. A transfer gate voltage of V_T = 40 V relative to the substrate was used as this is close to the maximum allowed with a channel output breakdown voltage of 65 V. The avalanche initiation probabilities $P_h(o)$ and $P_e(w)$ as a function of position x under the photogate are shown in Fig. 4.27. These quantities were obtained by integrating equations (3.2) - (3.5) along the lines (a) - (g) shown in Fig. 4.26, as described in Appendix D. The position x refers to the starting position of the carrier involved. The avalanche initiation probabilities for photogate voltages of 103V and 108V are also shown.

These results indicate that premature edge breakdown has not been completely eliminated; however, the calculated increase in avalanche initiation probability towards the edge of the photogate corresponds to only a 3% increase in the peak field. Since the uniformity of the triggering probability improves as the gate is biased farther above breakdown, this degree of premature edge breakdown is not expected to be important in a

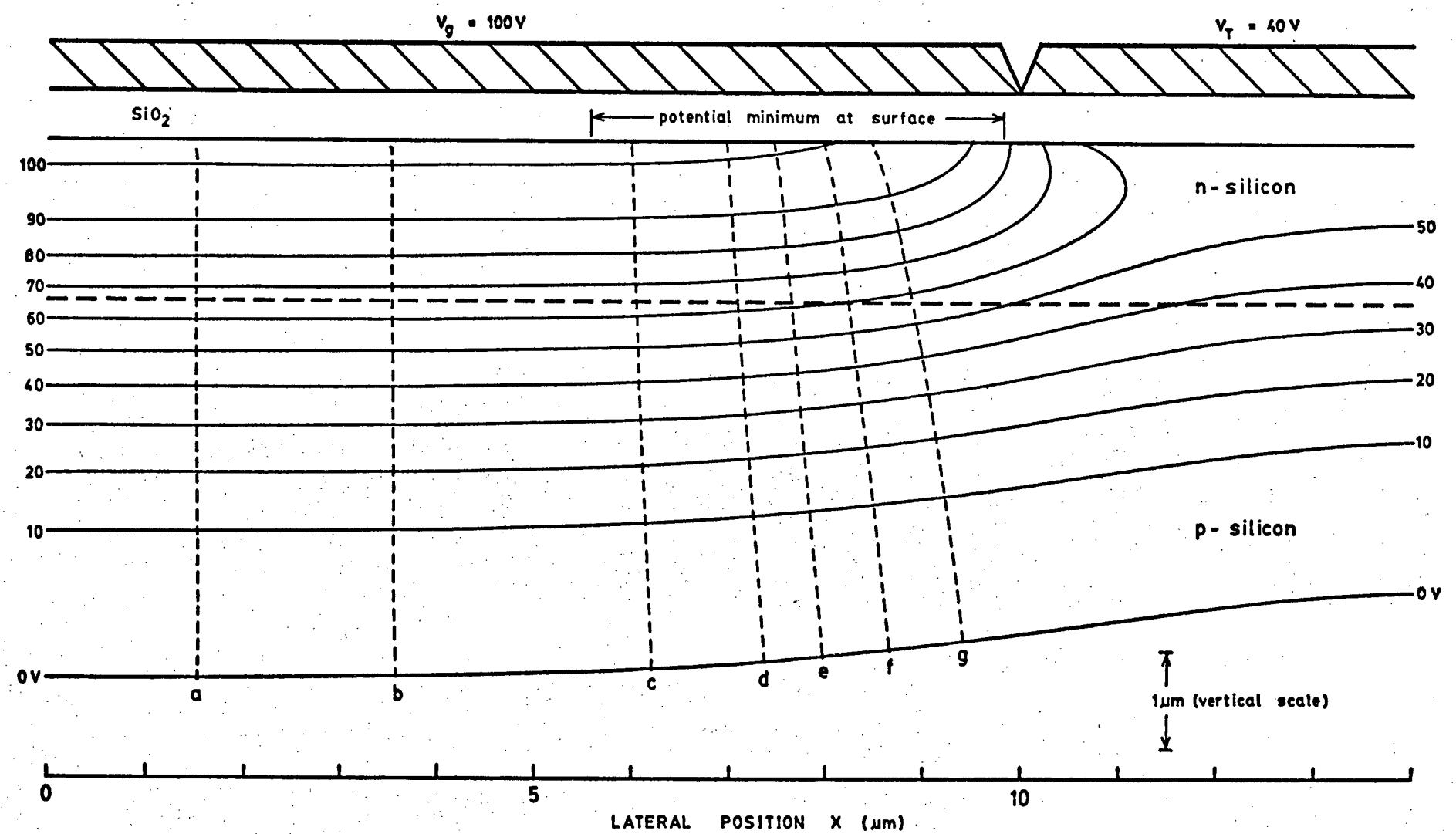


FIGURE 4.26 Two-dimensional potential distribution under the photogate for V_g = 100 V and V_T = 40 V. The lines along which the avalanche initiation probabilities shown in Fig. 4.27 were calculated, are also shown.

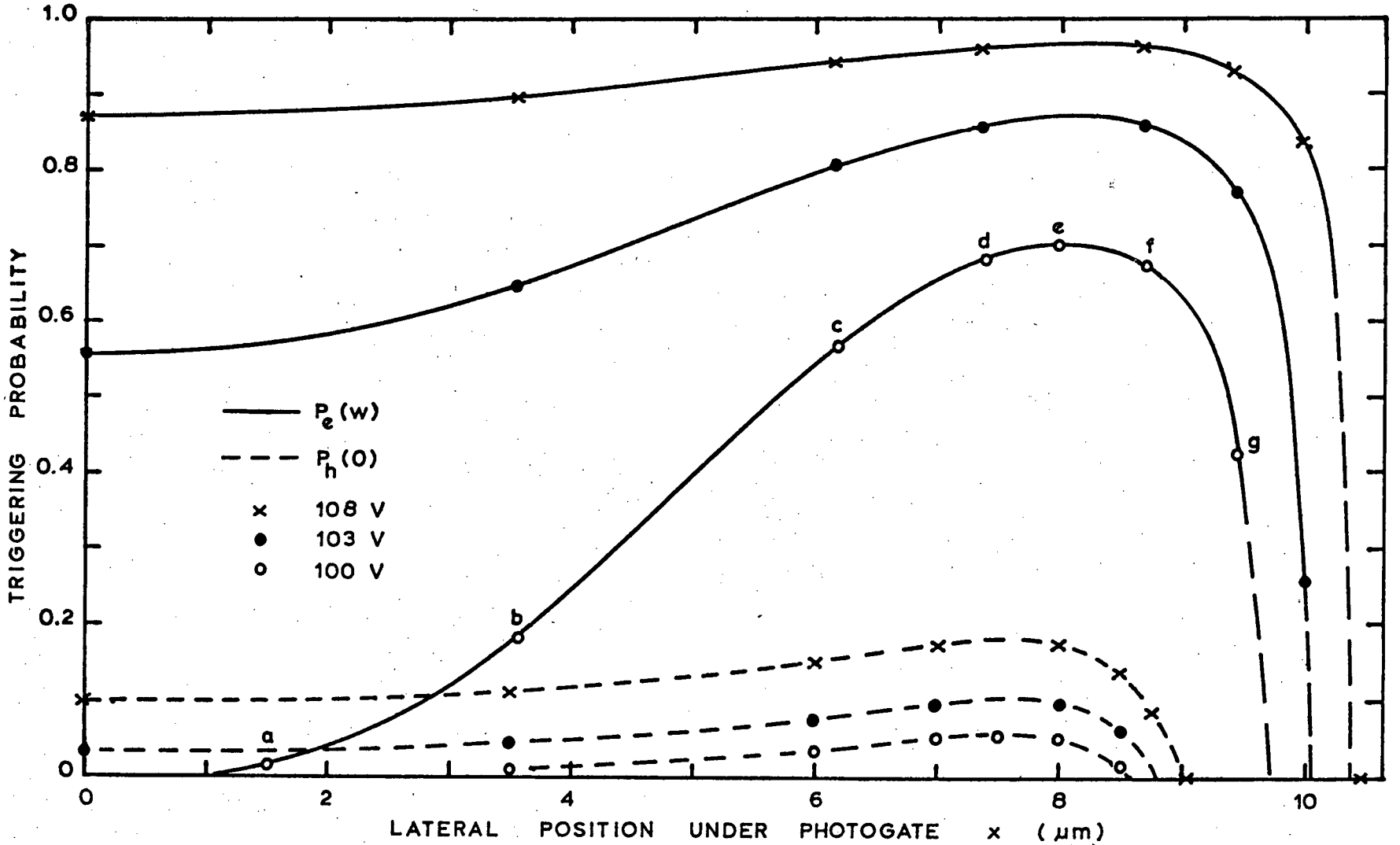


FIGURE 4.27 Results of the two-dimensional calculation of the variation of the avalanche initiation probability with position under the photogate, T = 80 K. Points a - g for the 100 V data correspond to the field lines a - g in Fig. 4.26.

PC-CCD. As regards the test devices, however, it means that the measured photon induced pulse rates can no longer be used to determine $P_e(w)$ versus excess bias directly. Initially the rapid increase in pulse rate is due primarily to the increasing effective triggering area. At higher excess biases the increase in photon induced pulse rate follows $P_e(w)$ more closely.

As was discussed in sention 4.2.1, the final n-layer dopant dose on wafer M66 was lower than desired. As a consequence of this there is a region around the edge of the photogate where the potential minimum is located at the interface rather than in the bulk. As the photogate is biased further above breakdown this region extends inwards. The calculated extent for V_g = 100 V and V_T = 40 V is indicated in Fig. 4.26, however, its width is sensitive to the n-channel doping parameters used so that it is not certain how wide this region is in the actual test devices.

Although holes generated by interface states are able to trigger breakdowns when the n-channel is too lightly doped, their contribution to the dark pulse rate should be very small. This is because the interface remains depleted of electrons and holes during reset so that the interface states remain filled to approximately mid gap. Since the fields in the silicon are very low at the interface, the steady state generation rate of holes should be well below 500 $sec^{-1}cm^{-2}$ (see section 3.2.3). Subsequent to an avalanche discharge the interface states above mid gap are also filled with electrons so that the hole emission rate is temporarily reduced even further. With a correctly doped n-channel there is still a region at the edge of the photogate where the potential minimum is at the interface, however, it is very narrow and lies well outside the area where $P_h(0)>0$.

4.2.4 Experimental Results and Discussion

The yield of operational devices on wafer M66 was very low. This was primarily due to misalignment between masking levels, particularly the two metalization masks and the via mask, and to pinholes in the SiO_2 and Al_2O_3 oxides. The misalignment was a result of the rather large ($\pm 4\mu m$) step and repeat errors on the masks themselves and not due to alignment errors during the photoresist operations. Out of the 80 test die 10 were found upon visual inspection to be correctly aligned. These 10 die were packaged and further tested for fatal defects such as interlevel shorts or shorts to the substrate. This reduced the number of good test die to 4. The first device tested at high voltages suffered a destructive breakdown of the Al_2O_3 layer, resulting in a short between the transfer gate and the field gate. The remaining 3 test die (18 devices) were fully operational above breakdown.

The following operating voltages were used for the above breakdown device testing (see Fig. 4.24):

V_{sub} = -60V

V_{out} = 0 V (Amp. virt. gnd.)

V_{sh} = -75 V

V_T = -30 V

V_p = -32 to +58 V

The measured flat band voltage was -2.0 V on both the p-substrate and the n-channel. Charge transfer to the channel output takes place at the end of the negative-going photogate reset pulse. Integration of the channel current was, therefore, started during the negative-going ramp, at V_p = -20 V. Figure 4.28 illustrates the timing used and the typical output pulses obtained. A ramp rate of 5×10^6 $Vsec^{-1}$ was used on both the positive and negative-going drive pulse edges.

The first detectable breakdowns under the photogate occurred at

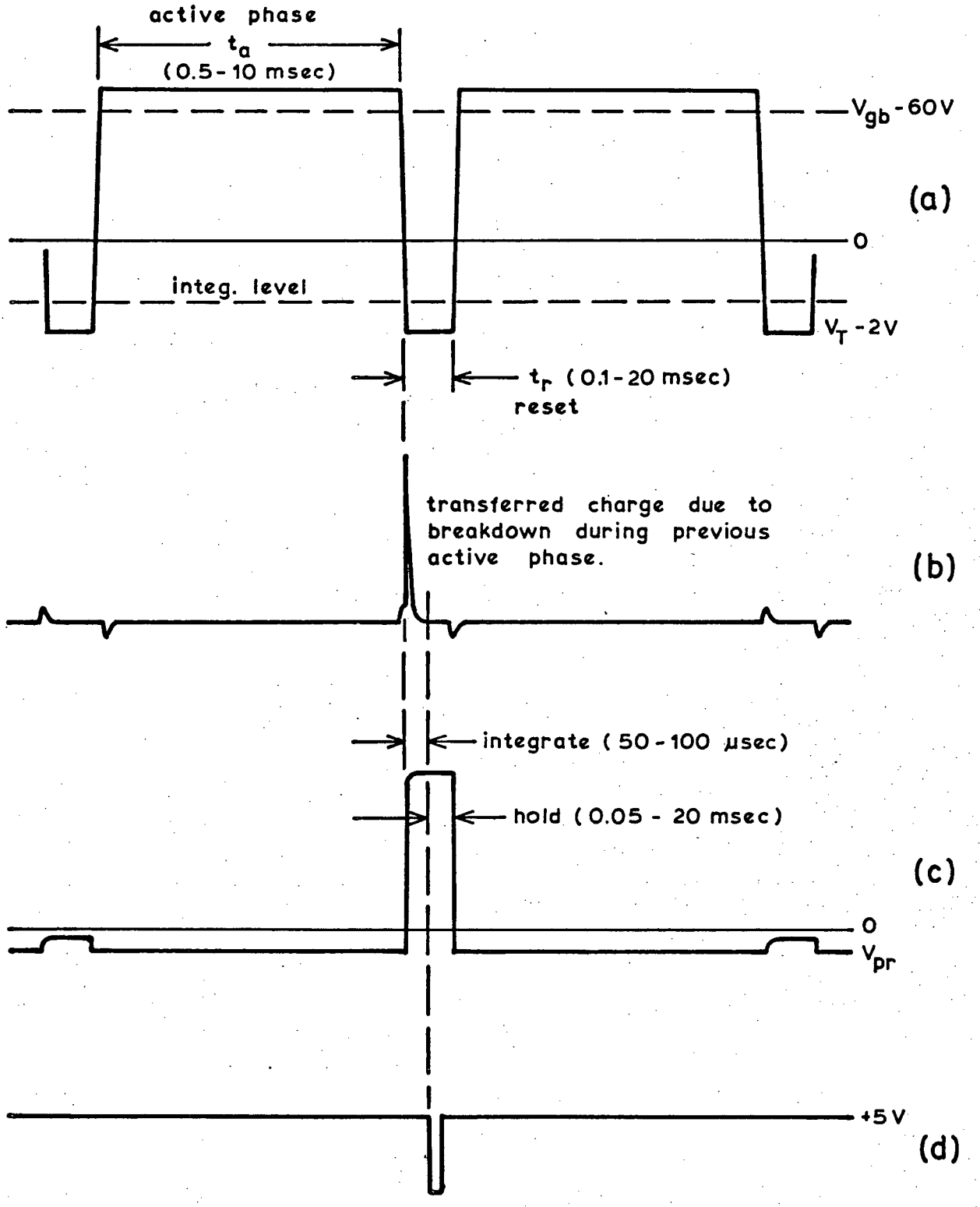


FIGURE 4.28 Test waveforms and timing for the bulk-breakdown, charge-transfer devices.

(a) high voltage driver (photogate)
(b) output of the current to voltage amplifier, Channel output current.
(c) output of the preset integrator
(d) discriminator output, threshold = 0 V

approximately V_p = +41 V (a gate potential of V_g = 101 V relative to the substrate), in reasonable agreement with the value for V_{gb} predicted from the two-dimensional calculations. The photogate voltage could be extended to approximately +58 V before the p-type bulk under the 10 μm wide line connecting the photogate to its bonding pad also started to break down. Because the shield gate is biased so as to accumulate the p-substrate, the photogate interconnect line first breaks down at the point where it steps up over the shield gate (see Fig. 4.4(a)).

In contrast to the surface breakdown devices, the pulse height distribution for the bulk devices was very sharply peaked at all photogate biases except those very close to V_{gb}. Typically the pulse heights varied by less than ±10% (total variation) except for the odd low pulse that was assumed to be due to a discharge occuring during the rising or falling edge of the drive pulse.

The dark pulse rates showed considerable variation over the 18 operational devices. All of the devices, however, had dark pulse rates that were considerably lower than those of the previous surface breakdown devices, in spite of the fact that the measured bulk lifetime (after processing) for wafer M66 was approximately 5 μsec, a factor of 10 lower than for the surface breakdown devices. The bulk lifetime was estimated from the room temperature leakage current of the n-p channel junction (60 nA cm^{-2} at 20 V reverse bias) and from room temperature charge collection measurements under the photogate. Both methods gave essentially the same bulk lifetime. Out of all 18 devices, two in particular exhibited dark pulse rates that were a factor of two lower than the rest, devices M66/2-1 and M66/4-3. The majority of the dark and photon-induced pulse rate measurements were made on device M66/2-1. As before, all of the pulse rates presented and referred to in the following discussion are those obtained after correcting for dead time

and temporal sampling effects.

At any given excess gate bias it was found that the dark pulse rate decreased if the duration at or above breakdown t_a (see Fig. 4.28) was increased while holding the reset duration t_r constant. This result is shown in Fig. 4.29. The dark pulse rates have been plotted as a function of $1/t_a$, i.e., the number of resets per second of active time. It is apparent that the dark pulse rate increases linearly with the number of resets. The limiting dark pulse rate at $(1/t_a)=0$ is approximately 2×10^4 $sec^{-1}cm^{-2}$ at $(V_g - V_{gb}) = 10$ V, only a factor of 40 larger than the desired maximum dark count rate. In accordance with the above result it was also found that the dark pulse rate increased to a limiting value as the reset duration was increased, while holding t_a constant, Fig. 4.30.

In order to obtain further insight into the mechanism responsible for the dark pulse rate, a heater was installed on the end of the liquid nitrogen heat pipe, enabling operation up to 140 K. The results of dark pulse rate measurements at 80K, 100K, 120K, and 140K are presented in Figures 4.31(a) and (b). Over this temperature range there was essentially no variation in the limiting dark count rate (i.e., when operated with short reset times, $t_r = 0.1$ msec, and long active times, $t_a = 10$ msec), as shown in Fig. 4.31 (a). Furthermore, the rapid increase in pulse rate with increasing gate bias always started at the same gate bias even though the breakdown voltage had increased by approximately 5 volts from 80K to 140K. This temperature independent supralinear behaviour of the dark pulse rate strongly suggests a tunneling mechanism for the dark generation of triggering carriers. The dark generation at $V_g = 117$ V, however, is approximately eight orders of magnitude higher than the rate obtained by extrapolating Haitz's [36] interband tunneling data according to Eq.(3.56), as described in section 3.2.5. For this reason it is believed that the tunneling genera-

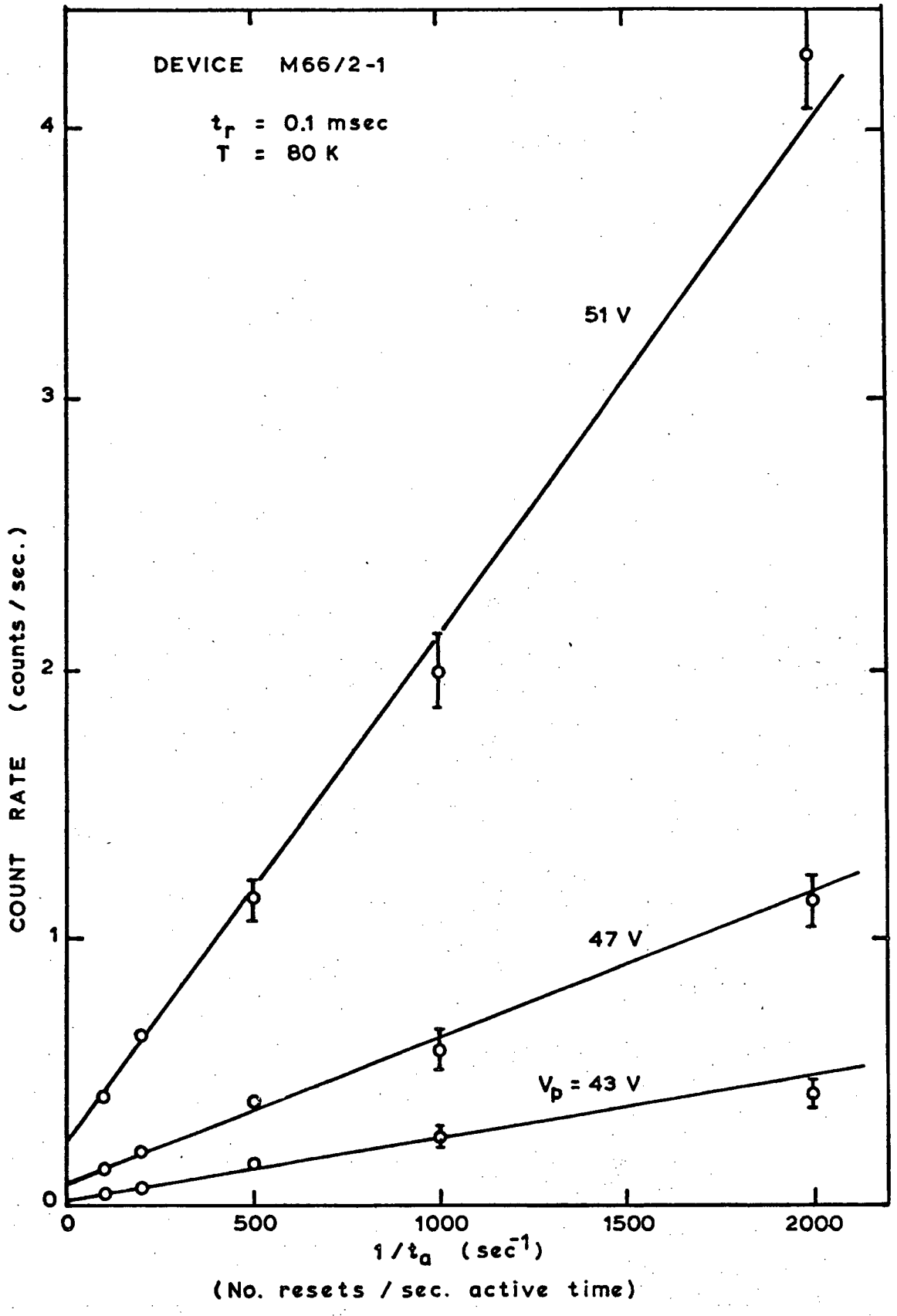


FIGURE 4.29 Dark pulse rate as a function of $1/t_a$. (i.e., as a function of the number of resets per sec. of active time) for a fixed reset duration of t_r= 0.1 msec.

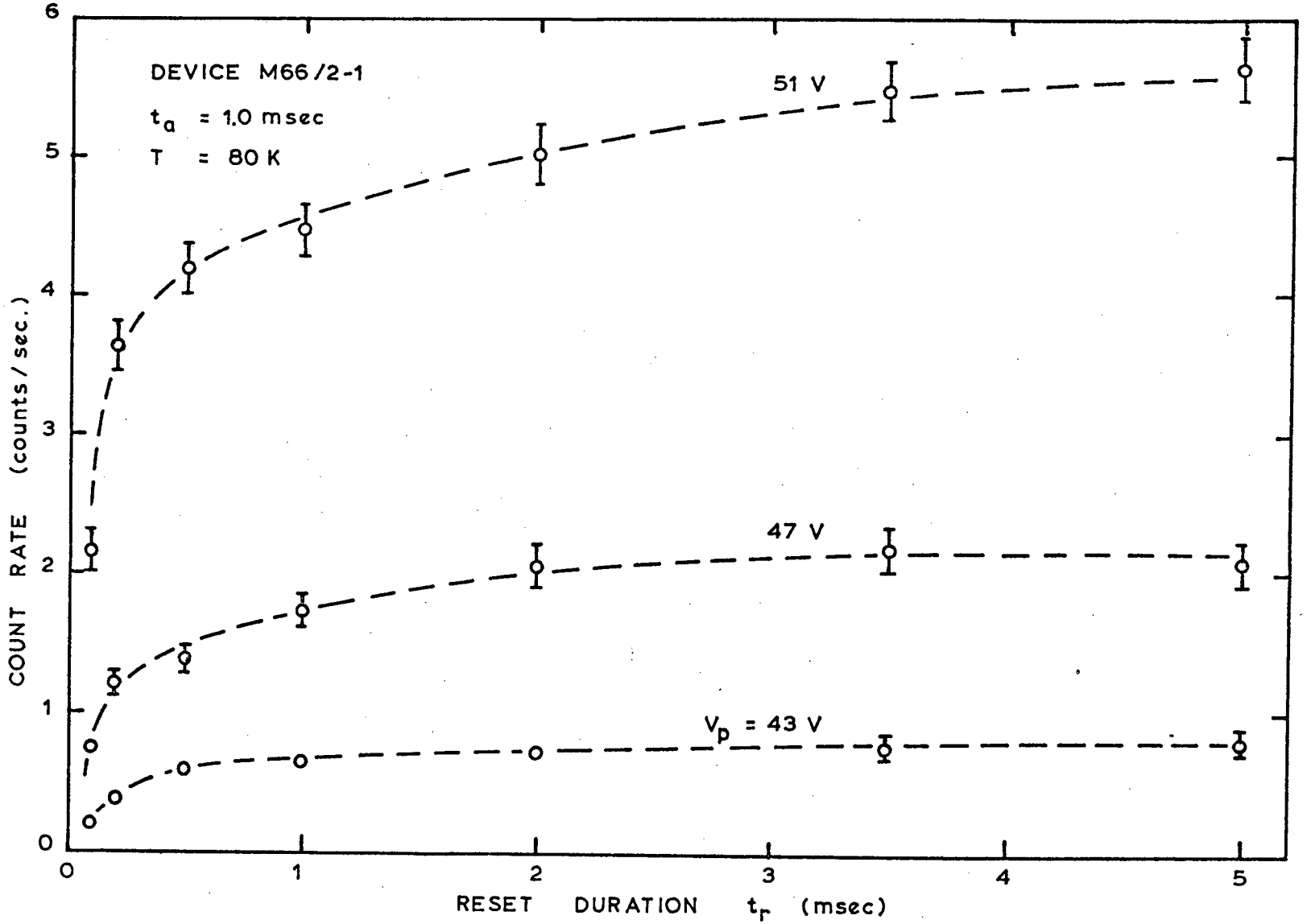


FIGURE 4.30 Dark pulse rate as a function of the reset duration for a fixed duration above breakdown of t_a = 1.0 msec.

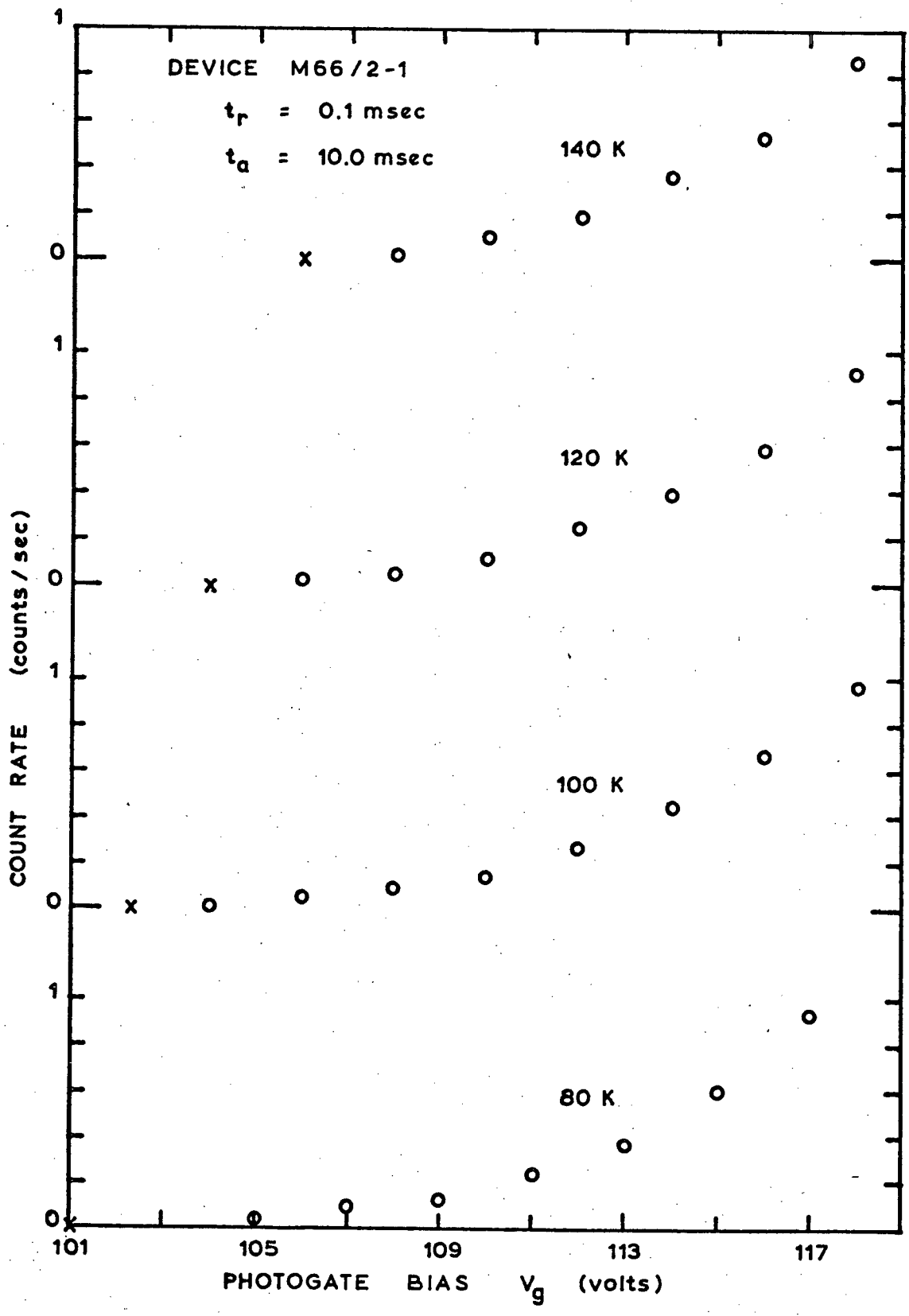


FIGURE 4.31 (a) Dark pulse rate as a function of the photogate bias and substrate temperature. The x's mark the approximate breakdown voltage.

t_r = 0.1 msec , t_a = 10.0 msec

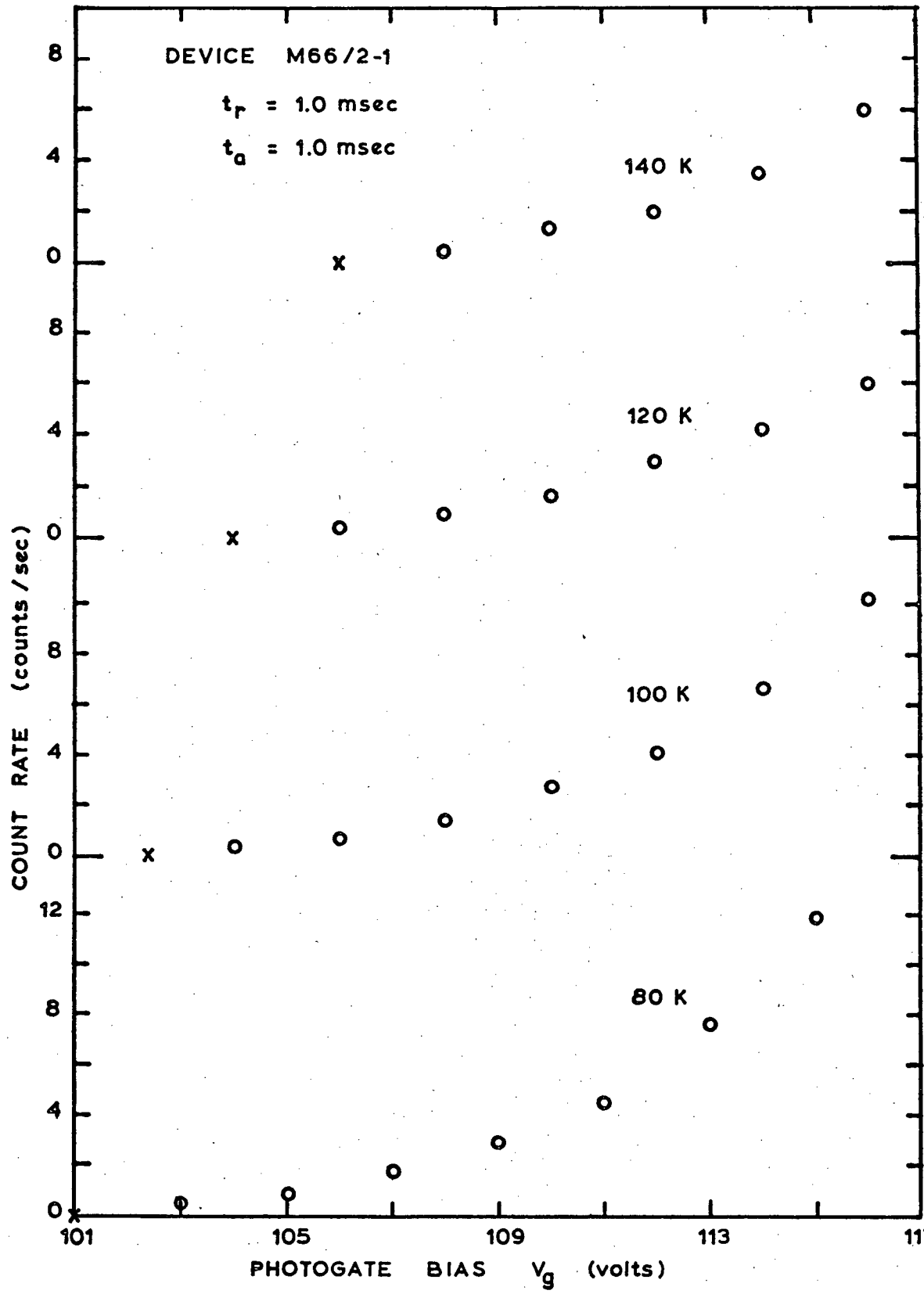


FIGURE 4.31 (b)

t_r = 1.0 msec , t_a = 1.0 msec

tion of triggering carriers is occurring through deep traps. The observed generation rates are in general agreement with those extrapolated from Sah's [54] data, according to equations (3.57)-(3.62), provided a mid gap trap density of 10^{12}-10^{13} cm^{-3} is assumed. Such a density of mid gap levels is consistent with the 5 μsec bulk lifetimes measured on wafer M66.

Referring back to Fig. 4.29 it can be seen that the increase in pulse rate with decreasing t_a is such as to maintain a constant ratio between the pulse rates at different gate biases. This implies that the increase in dark generation is due to a change in the occupancy of those traps involved in the tunneling, leading to an enhanced tunneling emission rate of either electrons or holes following the reset. This conclusion is substantiated by the lack of any strong temperature dependence of the dark pulse rate when operating with long (1 msec) reset times and shorter (1 msec) active times, Fig. 4.31(b). The slight decrease in dark pulse rate with increasing temperature, seen in Fig. 4.31(b), cannot be explained at present, nor has it been possible to envisage a mechanism whereby the occupancy of the traps involved in the tunneling generation changes during reset so as to increase the subsequent generation rate. The problem is basically that the conditions during reset are not sufficiently different from those above breakdown to make possible a large change in occupancy. The entire high field region remains in depletion during reset and the peak field is less than a factor of 1.5 lower than in the deeply depleted condition.

Figure 4.32 shows the red gallium arsenide phosphide LED source used to make the photon induced pulse rate measurements. Light emerging from the 500μm dia pinhole on the integrating sphere was imaged at 10 times reduction on to the rear of the 300μm thick devices, using an f/2 16 mm lens. A photodiode mounted in the integrating sphere was used to monitor the light intensity. An absolute calibration of the number of photogenerated electrons per

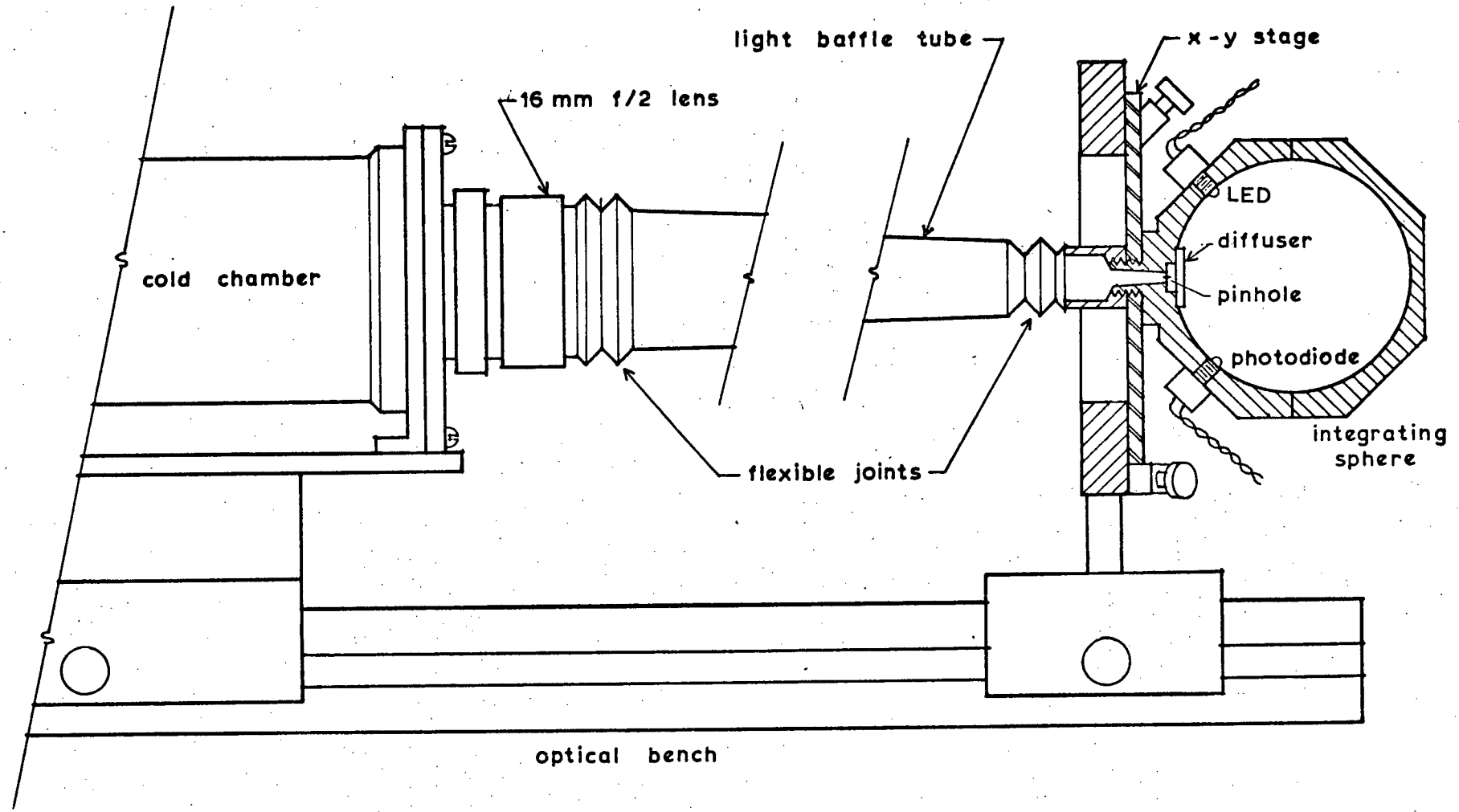


FIGURE 4.32 Red gallium arsenide phosphide LED source used for the photon induced pulse rate measurements.

second reaching the edge of the photogate depletion region from the neutral bulk was determined from charge integration measurements. For these measurements the devices were operated in the charge integration mode with the following gate potentials:

V_{sub} = -30V

V_{out} = 0 V (amp. virt. gnd.)

V_{sh} = -45V

V_T = -45V

V_p = -47 to +20V

By operating the transfer gate at -15V relative to the substrate the effective collection area is restricted to the area of n-channel under the photogate and transfer gate. The p-substrate around the n-channel is held in accumulation at the surface by the transfer gate, thus providing a barrier to the photogenerated carriers collected in the surface depletion region outside this area. The effective collecting area, measured under an optical microscope, was 45μm x 65μm. The measured area of the photogate was 22μm x 42μm, 31.6% of the effective collecting area. 60 min. integrations were made using the same level of illumination as was used for the above-breakdown tests. The measured signal voltages were converted to a charge measurement using the calculated gain of the current amplifier integrator combination. 60 min dark integrations resulted in zero measurable charge.

Figures 4.33(a) - (c) show typical examples of the photon induced pulse rate for the following three operating conditions:

(a) t_a = 1.0 msec, t_r = 0.2 msec

(b) t_a = 10.0 msec, t_r = 0.2 msec

(c) t_a = 1.0 msec, t_r = 20.0 msec

The injection level obtained from charge integration measurements, as described above, is indicated in each case. The indicated uncertainty in this

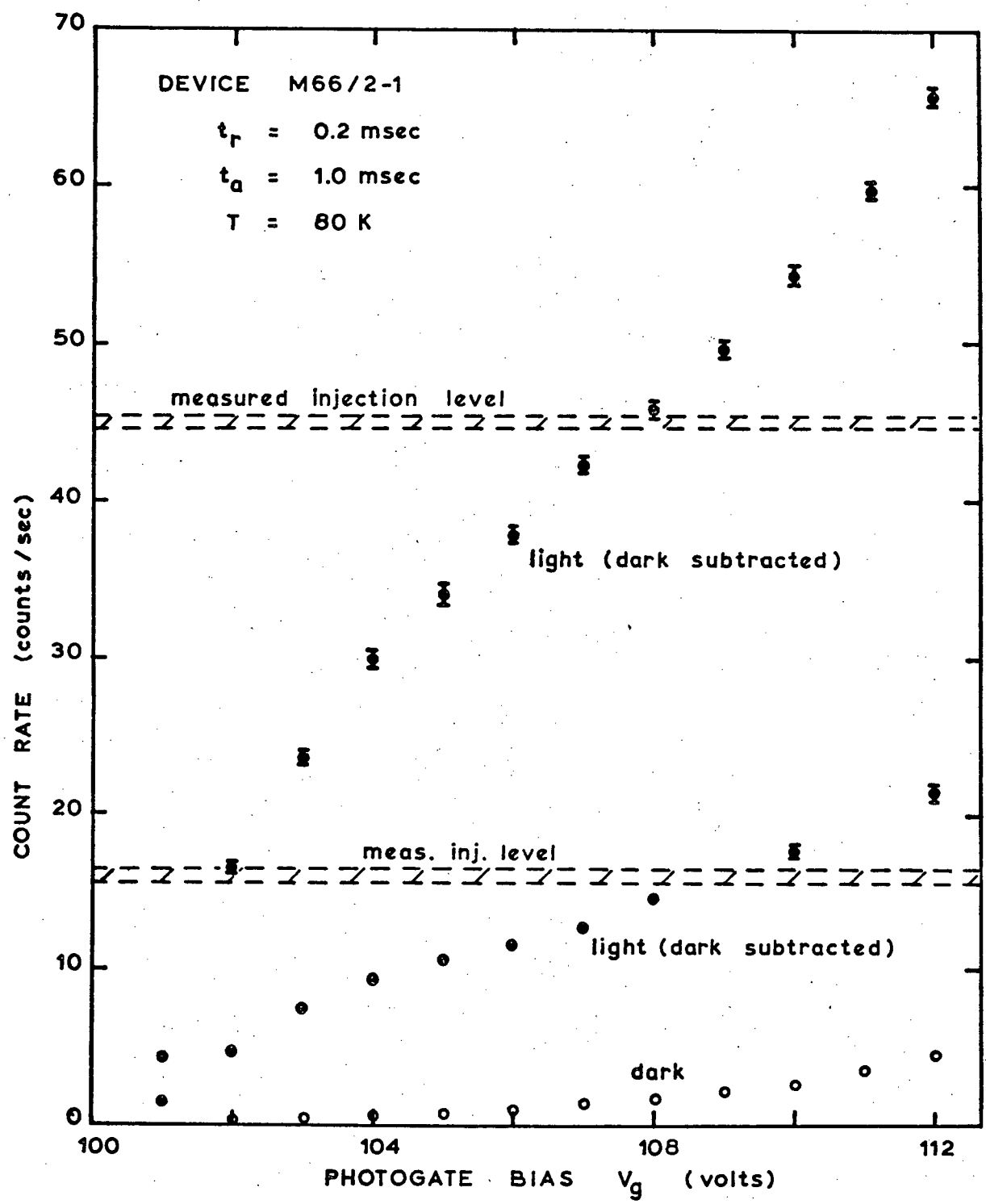


FIGURE 4.33 (a) Dark and photon induced pulse rates as a function of the photogate bias for device M66/2-1. The measured injection level is from charge integration measurements below breakdown.

t_a = 1.0 msec , t_r = 0.2 msec

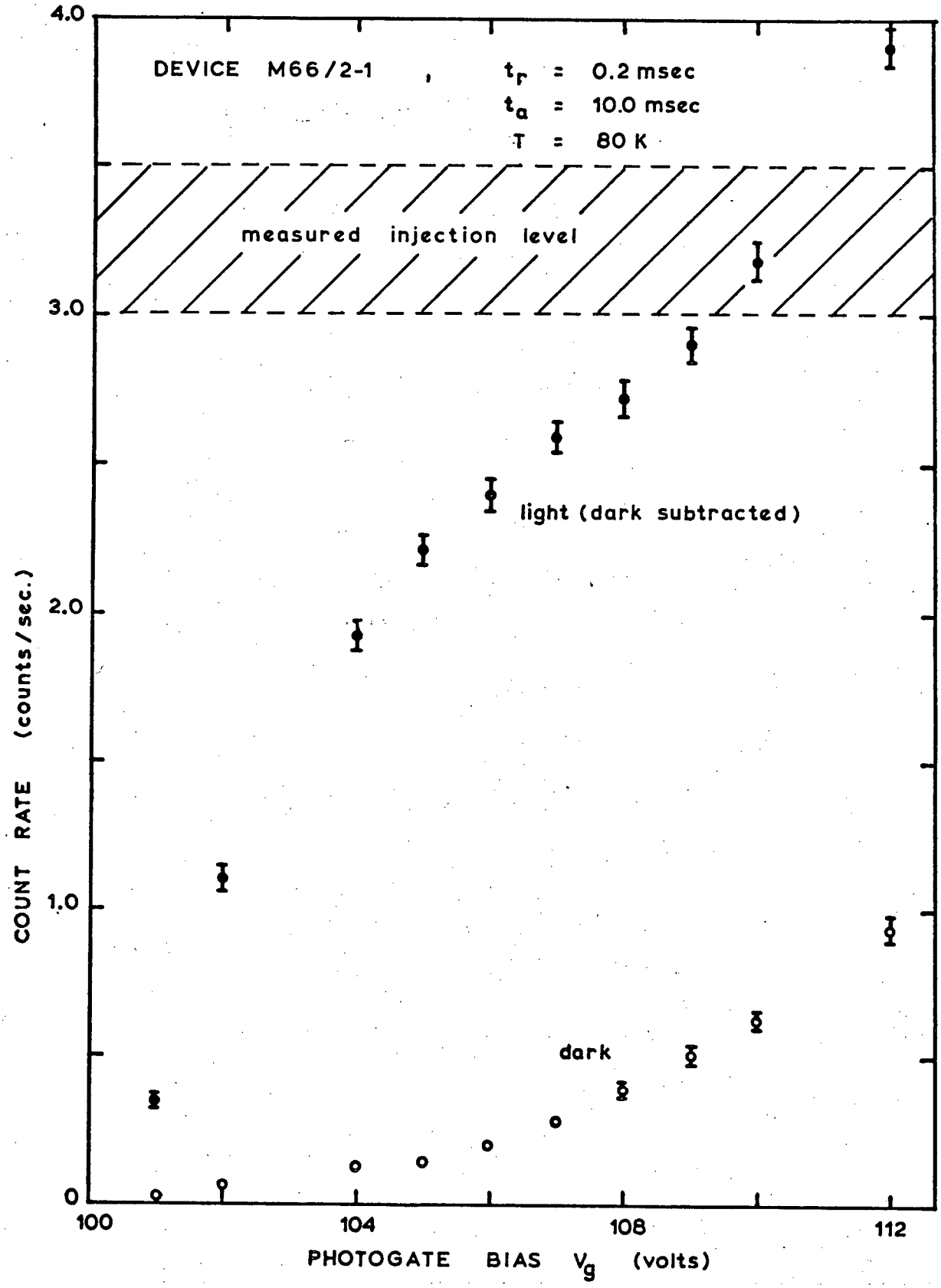


FIGURE 4.33 (b)

t_a = 10.0 msec , t_r = 0.2 msec

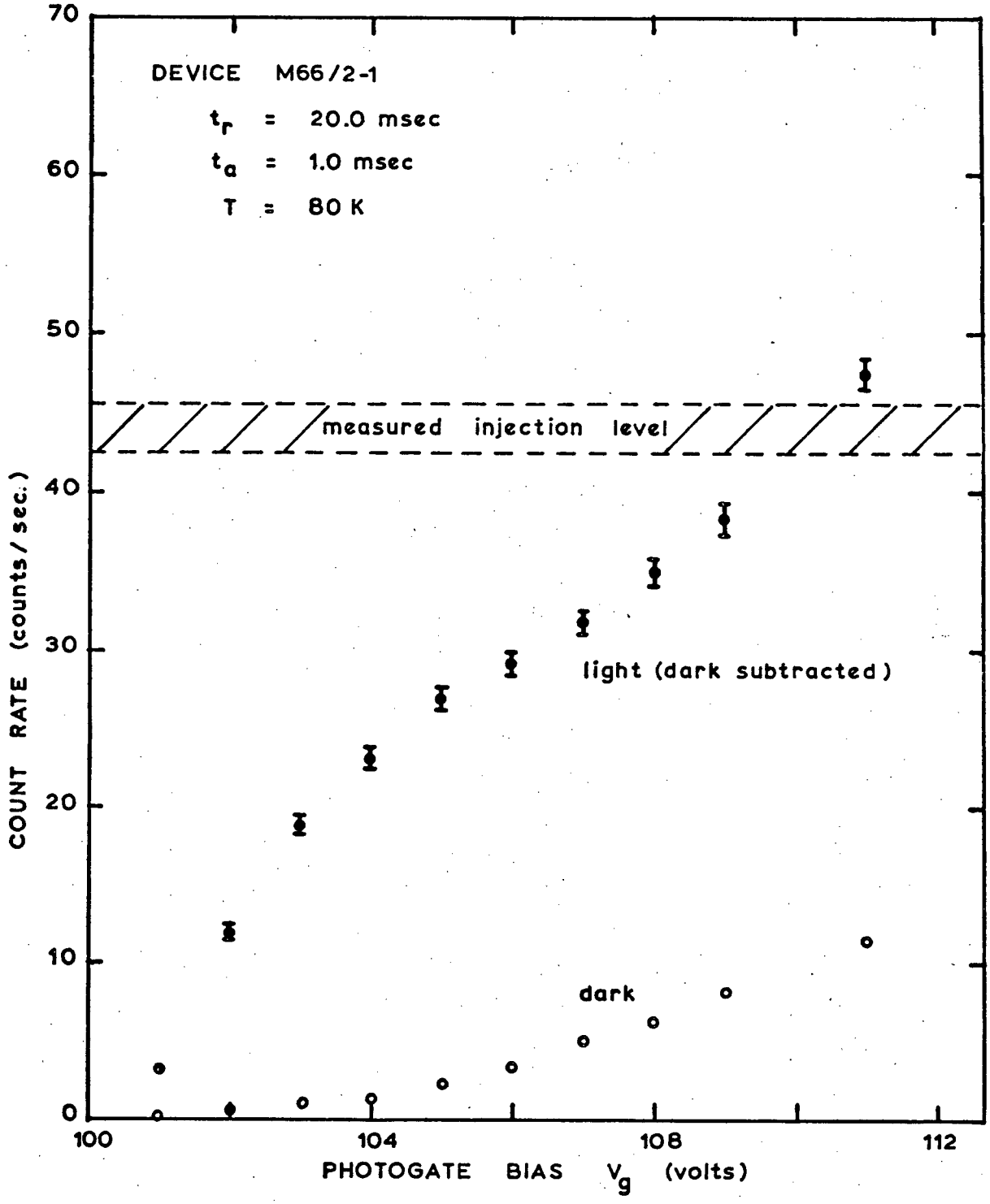


FIGURE 4.33 (c)

t_a = 1.0 msec , t_r = 20.0 msec

level reflects the level of noise in the charge amplifier only, and not the uncertainty in the estimated ratio of collecting area to photogate area. Initially the photon induced pulse rate begins to saturate with increasing photogate bias as expected, however, at approximately $V_g - V_{gb} = 7$ V the increase turns supralinear. Up to that point the level to which the pulse rate appears to be saturating is in reasonable agreement with the measured injection level.

As with the surface breakdown devices the supralinear behavior of the photon induced pulse rate at high excess gate biases is attributed to either carrier capture by traps or impact ionization of the traps, during the periods of avalanche discharge. By comparing Figures 4.33(a) and (c) it can be seen that the supralinear behavior of the photon induced pulse rate changes very little as the reset duration t_r is increased from 0.2 msec to 20.0 msec. This indicates that the detrapping time constant during reset (at 80K) is much longer than 20 msec. In Fig. 4.33(b) the supralinear behavior starts more abruptly and at a somewhat higher excess gate bias. It is felt that this is due to the longer active time t_a and a shorter detrapping time constant due to tunneling under the high field conditions that exist after the discharge (i.e., at V_{gb}).

Since the detrapping time constant is much longer than 20 msec during reset and somewhat longer than 10 msec during the active time, the following experiment was devised to obtain a true measure of the number of pulses triggered by photogenerated carriers. The devices were operated with an active time of $t_a = 1.0$ msec and a reset time of $t_r = 2.0$ msec. The LED source was pulsed on during the active time of every second cycle and two separate counts were accumulated; one for those cycles with the LED on, and the other for those cycles with the LED off. The 2.0 msec reset allows time for the photogenerated carriers in the bulk to recombine

before the next pulse above breakdown. The counts accumulated with the LED off are, therefore, due only to the dark counts and the counts resulting from the detrapping following an avalanche discharge. These counts can then be subtracted from those obtained with the LED on to arrive at the counts due to photogenerated carriers. In order to minimize coincidence losses, the light level was adjusted so that fewer than 3% of the frames contained counts. The results of such measurements are shown in Fig. 4.34.

In order to compare these results with those predicted from the theory, a full three dimensional calculation of the potential distribution and variation of $P_e(w)$ under the photogate is required. Such calculations have not been made. Instead, the two dimensional results were used in such a way as to approximate a three dimensional solution. For these calculations a photogate 20μm wide by 44μm long, but with hemispherical ends, was assumed. This gives the same total area as a 20μm x 40μm gate. The variation of the avalanche initiation probability, along any line normal to the perimeter and extending into the longitudinal center line of the gate, was assumed to be the same as that given by the two dimensional solution shown in Fig. 4.27. Using this construction, the expected variation of the photon induced pulse rate with gate bias (for rear illumination, i.e., pure electron injection) was calculated from the double integral

$$\iint P_e(w)\, dx\, dz \quad .$$

The calculated variation of the photon induced pulse rate is shown by the solid line in Fig. 4.34. The calculated curve has been shifted by +1.85V, and has been arbitrarily fit through the experimental point at V_g = 110 V, as no measurement of the absolute injection level was made.

The deviation of the first two experimental points from the low voltage tail of the calculated curve is an expected result since space charge effects

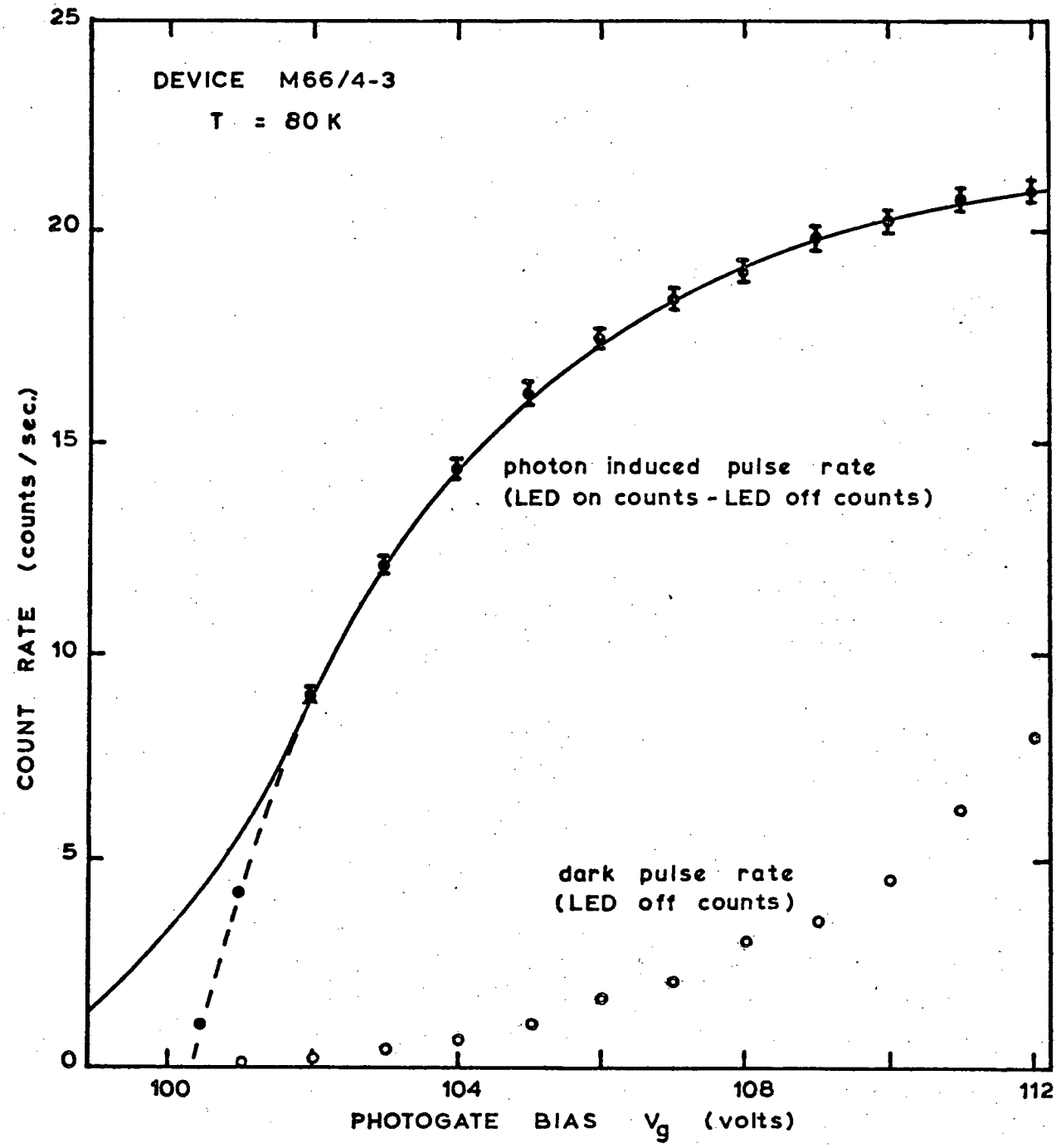


FIGURE 4.34 Photon-induced pulse rate as a function of the photogate bias when the dark subtraction includes those counts resulting from the detrapping following an avalanche discharge. The solid line shows the variation of the photon-induced pulse rate, calculated using the results of the two-dimensional modeling. The calculated curve has been shifted by +1.85 V and fitted through the experimental point at V_g = 110 V.

have been neglected. At small excess gate biases the build-up of space charge significantly reduces the number of discharges that result in pulses larger than the 5×10^4 elect. discriminator level used for these measurements. Also, the effective breakdown area is reduced in the low voltage tail which further lowers the size of the discharge pulses.

The excellent agreement between the calculated and observed variation of the photon induced pulse rate with gate bias for the remainder of the experimental points indicates that the avalanche initiation probability is accurately described by the theory due to Oldham et.al [31], equations (3.2) - (3.5). The upper experimental point at $V_g = 112$ V corresponds to an avalanche initiation probability for electrons, $P_e(w)$, that ranges from 0.9 at the center of the photogate to 0.97 at the edge.

5 SUMMARY AND CONCLUSIONS

A solid state photon counting sensor, based on the above-breakdown operating regime of MOS structures, has been proposed and investigated both theoretically and experimentally. It has also been described how specially designed charge-coupled arrays may be operated in this new regime, resulting in the realization of an entirely solid state high performance photon counting imager.

In order to demonstrate the need for such an imager and its potential advantages, the general properties of state-of-the-art analog and photon counting imagers are reviewed briefly in chapter two. It is shown that CCD imagers are superior to other types of analog detectors for ultra low light level imaging, largely as a result of the low levels of readout noise ($\sigma<25$ electrons r.m.s.) that are being achieved with current CCD sensors. A quantitative means of appraising the performance of low light level sensors (the DQE) is introduced and it is shown that analog CCD sensors are very nearly an optimum detector over a wide range of wavelengths, centered at 700 nm, provided the statistical photon noise dominates the signal to noise in the output. For current CCD imagers this condition is met for an output signal to noise greater than approximately 70 to 1.

At very low photon fluxes, photon counting is generally the preferred imaging technique. This not only makes possible a DQE that is independent of the total integrated signal but it avoids some of the linearity, stability, and threshold problems often encountered with analog imagers. It is shown, however, that temporal sampling effects introduce non-linearities and lower the DQE if the photon arrival rate per pixel is allowed to approach the frame rate of the imager. Many of the existing photon counting systems

require real-time frame processing to detect the photon event centers, resulting in a limited frame rate (typically 100 sec^{-1}) and a low dynamic range. All of the image photon counting systems utilize either a semi-transparent or opaque photoemmisive surface as the initial light sensitive element. It is the limited responsive quantum efficiency of the commonly used photocathode material (RQE≤0.2) that ultimately limits the DQE of photon counting imagers at very low light levels.

The proposed solid state photon counting sensor and its theory of operation are discussed in chapter three. The above-breakdown operating regime is discussed and it is shown how an MOS photosensor may be operated in a photon counting (or Gieger tube) mode by pulsing it into very deep depletion, beyond the point where avalanche breakdown normally occurs. Operation above the breakdown voltage has been previously demonstrated, however, only p-n photodiodes have been operated in this mode so far. Furthermore, there has been no attempt to optimize these devices as low light level photon counting sensors, nor has a monolithic imaging array of such detectors been considered. The MOS photon counting sensor is shown to have an inherent self-quenching property that greatly simplifies its incorporation into such an array. The proposed photon counting CCD (PC-CCD) has the potential for achieving a DQE that is independent of the total integrated signal and which is nearly equal to the photon-noise limited DQE's being achieved with analog CCD imagers. Also, the large size of the signal charge packets generated by the photon induced avalanche discharges should enable high clocking rates to be used during readout, and make possible a high frame rate.

The following major problem areas in the development of a successful PC-CCD imager were identified:

(1) maximizing the avalanche initiation probability
(2) reduction of the dark pulse rate to an acceptable level

(3) achieving planar microplasma-free avalanche discharges

(4) prevention of optical coupling due to light emission during the avalanche discharges.

In the remainder of chapter 3 the theoretical background required for a full understanding of these problems is introduced and the existing experimental data is reviewed. It is shown that a PC-CCD must be fabricated on a p-type silicon substrate and illuminated from the back side in order to obtain a high triggering probability for the photogenerated carriers.

All of the possible dark generation mechanisms have been discussed in detail. It is shown that all thermally activated steady state dark generation mechanisms that occur in the bulk can be reduced to a negligible level by cooling the sensor to low temperatures (100 K or lower). The generation of possible triggering carriers (holes) by interface states can be controlled by maintaining an inversion charge at the silicon surface. The generation of possible triggering carriers by interband tunneling may be reduced to an acceptable level by ensuring that the peak fields within the depletion region are below approximately 4.3×10^5 Vcm^{-1}. Operation above breakdown with peak fields as low or lower than this requires wide depletion regions (i.e., a low substrate doping) which in turn results in large (60-70V) potentials across this region. Due to the potential drop across the oxide the MOS gate used to generate these depletion regions must be operated at still higher voltages so that specially designed charge transfer arrays are required for operation in the above-breakdown regime.

The generation of triggering carriers by band to band tunneling through trap states is also examined. There is very little existing data upon which to base estimates for the dark generation due to this mechanism but by extrapolating data obtained with high field Esaki diodes, order-of-magnitude estimates can be made. Such estimates indicated that tunneling through

traps would likely be the dominant steady state dark generation mechanism and that, depending on the trap density, peak fields lower than 3×10^5 Vcm^{-1} may be required. Achieving the lowest possible trap density is essential for minimizing this type of dark generation.

Following the depleting pulse above breakdown, there is a transient situation during which the dark pulse rate may be either increased or decreased from its steady state value depending on the conditions during reset and on the energy and capture cross-sections of the trapping levels. A similar transient situation exists after each breakdown pulse, due to charge trapping or impact ionization of the traps during the avalanche discharge. Theoretical expressions are derived to describe the change in pulse rate during these transients. A low density of traps is required to minimize the possibility of re-triggering due to increased carrier emmission following an avalanche discharge.

The control of fringing fields and the prevention of premature edge breakdown is discussed briefly, along with the processing techniques used to prevent or eliminate lattice defects that might lead to localized microplasma breakdown.

A review of the existing data on light emission during avalanche breakdown indicates that some form of optical barrier between the individual pixels in an array will be required. Two methods for achieving a high degree of optical isolation between the pixels in linear arrays are described.

The experimental investigation of MOS structures operating in the above-breakdown regime is reported in chapter four. This investigation was directed primarily at problems (1) and (2) above.

MOS structures that breakdown at the silicon surface were studied initially. An extension of the gate metal over the thick field oxide was used to prevent edge breakdown and the devices studied were reset by inject-

ing the charge into the substrate. A gate oxide thickness greater than 0.2 μm was shown to be necessary. Difficulties encountered during the double level Al-Al_2O_3-Al metalization indicated that the aluminum anodization should precede any contact sinter or annealing treatments and that selective etching rather than selective anodization should be used to form the first to second level metal vias.

In order to obtain appreciable delays to breakdown the surface-breakdown devices had to be pulsed from a reset condition corresponding to inversion under both the active region of the gate and the guard ring. When the guard ring is not inverted during reset the interface states around the edge of the breakdown area are unoccupied above mid gap. In the deeply depleted condition the breakdown area extends outwards into this edge region, resulting in a large dark pulse rate which is due to the increased hole emission from the unoccupied interface states under the high field conditions that exist in deep depletion. The low interface state density measured on the test samples indicates that it will not be possible to reduce this form of dark generation to an acceptable level by further reductions in the interface state density.

Inverting the guard ring during reset results in a large reduction of the dark pulse rate but its effectiveness is limited because the inversion layer charge is transferred to the active (thin oxide) region of the photogate after pulsing above breakdown thus allowing the interface states in the edge region to empty. Also, this charge transfer increases the gate voltage required for breakdown under the photogate and results in very high oxide field strengths. The dark pulse rates of the test devices continued to decrease as the inversion layer charge during reset was increased, until the point where the guard ring also started to break down or the higher oxide fields resulted in a destructive breakdown of the gate oxide. The lowest

dark pulse rates measured on the surface breakdown devices, at excess biases sufficient to give a triggering probability larger than 0.5, were 3 orders of magnitude higher than the maximum desired dark pulse rate of 500 $sec^{-1}cm^{-2}$. For the above reasons MOS structures that breakdown at the silicon surface were ruled out as possible image elements for a PC-CCD. These problems can be avoided by going to an MOS structure that breaks down in the bulk away from the $Si-SiO_2$ interface. It is shown that such a self-quenching, bulk-breakdown MOS gate can be made by forming a lightly doped n-channel at the silicon surface, as in the buried channel CCD structure. Oxide field strengths are very low in the bulk-breakdown structure so that the dark generation by interface states is negligible at the low (<100 K) operating temperatures. Furthermore, with a correctly doped n-channel the potential minimum lies in the bulk of the n-channel away from the $Si-SiO_2$ interface so that the carriers emitted by interface states are unable to trigger a breakdown. The results of two-dimensional modeling of the completed test devices indicated that the n-channel had not been sufficiently doped and that the potential minimum was located at the $Si-SiO_2$ interface in the deeply depleted condition above breakdown. The results of the modeling did, however, confirm that the degree of premature edge-breakdown in the fabricated devices would not be excessive.

In general, the behavior of the bulk-breakdown MOS pulse counting detectors was found to be far superior to the surface-breakdown devices. The dark pulse rates were considerably lower, in spite of poorer bulk lifetimes and, in contrast to the surface devices, the pulse height distribution was very sharply peaked.

Only a very weak temperature dependence of the dark pulse rate was observed, suggesting a tunneling mechanism. The dark count rate was, however, eight orders of magnitude higher than that predicted for the direct band to

band tunneling generation of triggering carriers. The limiting dark generation mechanism for the bulk-breakdown devices studied (peak fields $\simeq 3.3 \times 10^5$ Vcm^{-1}) is, therefore, believed to be interband tunneling through trap states. The measured dark pulse rates are in general agreement with those predicted from the theory and existing data obtained with high field Esaki diodes, provided a mid gap trap density of 10^{12} to 10^{13} cm^{-3} is assumed. The measured bulk lifetime of 5 μsec is consistent with this density of mid gap generation centers.

The number of dark pulses detected was also found to be linearly related to the number of resets, suggesting that the tunneling emission of triggering carriers in the high field region is increased following a reset. There is, however, no apparent mechanism to explain the change in trap occupancy during reset required to cause this increased emission.

The photon induced pulse rate measurements indicated that retriggering following an avalanche discharge was still a problem. After making a simple dark subtraction the measured pulse rates did, however, show some initial signs of saturation with increasing gate bias. The level to which the pulse rates appeared to be saturating, before the increases turned supralinear, was in general agreement with the measured electron injection level.

By pulsing the LED light source on every second cycle it was possible to make a dark subtraction that included the counts due to re-triggering following an avalanche discharge. When this was done it was found that the photon induced pulse rate saturated precisely as predicted by the theory. These photon induced pulse rate measurements were extended up to gate biases 12V above breakdown, corresponding to an electron triggering probability greater than 0.9.

The re-triggering following the avalanche discharges constitutes a form of positive feedback and is an undesirable effect as it would cause

severe non-linearities in the response of a PC-CCD. It is believed, however, that with improved processing, the trap density can be reduced to a sufficiently low level that this non-linearity will not be a serious problem at excess gate biases sufficient to give triggering probabilities greater than 0.9. Such a decrease in trap densities would also lower the dark pulse rate below the desired maximum rate of 500 $sec^{-1}cm^{-2}$. Further improvements would also be expected with a bulk-breakdown MOS gate designed to operate with lower peak fields.

BIBLIOGRAPHY

[1] J.L. Lowrance, and P. Zucchino, in Methods of Experimental Physics, vol.12, Astrophysics, L. Marton, and N. Carleton, eds., Academic Press, Inc., New York (1974).

[2] J.A. Westphal, and J. Kristian, Proc. I.A.U. Colloquium 40, paper 19 (Sept. 1976).

[3] R.H. Dyck, and G.P. Weckler, IEEE Trans. on Electron Devices, ED-15, 196-201 (1968).

[4] M.H. White, in Solid State Imaging, P. Jespers, F. van de Wiele, and M. White, eds., Noordhoff (Leyden) (1976).

[5] G.J. Michon, and H.K. Burke, ISSCC Dig. Tech. Papers, pp. 138-139, Philadelphia (1973).

[6] G.J. Michon, and H.K. Burke, in Solid State Imaging, P. Jespers, F. van de Wiele, and M. White, eds., Noordhoff (Leyden) (1976).

[7] C.H. Sequin, and M.F. Tompsett, "Charge Transfer Devices," Advances in Electronics and Electron Physics, Supplement B, Academic Press Inc., New York (1975).

[8] G.S. Hobson, "Charge-Transfer Devices," A.H.W. Beck, and J. Lamb, eds., Edward Arnold (Publishers) Ltd., London (1978).

[9] Astronomical Observations with Television-Type Sensors, Proceedings of a symposium held at University of British Columbia, Vancouver 8, B.C., Canada, (15 May 1973), J.W. Glaspey, and G.A.H. Walker, eds.

[10] G.P. Weckler, IEEE Jour. of Solid-State Circuits, SC-2, 65 (1967).

[11] R.H. Walden, R.H. Krambeck, R.J. Strain, J. McKenna, N.L. Schryer, and G.E. Smith, Bell Syst. Tech. Jour., 51, 1635-1640 (1972).

[12] R. Shaw, J. Photo. Sciences, 11, 199 (1963).

[13] G.A. Antcliffe, L.J. Hornbeck, W.W. Chan, J.W. Walker, W.C. Rhines, and D.R. Collins, IEEE Trans. on Electron Devices, ED-23, 1225-1232 (1976).

[14] G.G. Fahlman, S.W. Mochnacki, C. Pritchet, A. Condal, and G.A.H. Walker, in "Adv. E.E.P." Vol. 52, p. 453 (1979).

[15] H. Gursky, J. Geary, R. Schild, T. Stephenson, and T. Weekes, Proc. Soc. Photo-Opt. Instr. Eng., 264, 14-18 (1980).

[16[A. Boksenberg, Proc. I.A.U. Colloquium 40, paper 13 (Sept. 1976).

[17] J.L. Geary, M.A. Beetz, and D. Fath, Proc. I.A.U. Colloquium 40, paper 29 (1976).

[18] E.A. Beaver, and C.E. McIlwain, Rev. of Sci. Instr., 42, No. 9 (1971).

[19] R.G. Tull, Proc. I.A.U. Colloquium 40, paper 23 (Sept. 1976).

[20] J.P. Choisser, Optical Eng., 16, No. 3, p. 262 (1977).

[21] J. Lowrance, and P. Zucchino, Proc. Soc. Photo-Opt. Instr. Eng., 172, 232-233 (1979).

[22] A.L. Broadfoot, and B.R. Sandel, Proc. I.A.U. Colloquium 40, paper 34 (1976)

[23] J.G. Timothy, and R.L. Bybee, Proc. Soc. Photo-Opt. Instr. Eng., 116, 24-32 (1977).

[24] M. Lampton, Proc. I.A.U. Colloquium 40, paper 32 (1976).

[25] J.G. Timothy, and G.H. Mount, Proc. Soc. Photo-Opt. Instr. Eng., 172, 199-206 (1979).

[26] J.P. Choisser, R.O. Ginaven, G.D. Hall, H.A. Naber, R.D. Smith, E.A. Beaver, and R.G. Hier, Proc. Soc. Photo-Opt. Instr. Eng., 172, 239-251 (1979).

[27] G. Keil, and H. Bernt, Solid State Electronics, 9, 321-325 (1966).

[28] H. Sigmund, Infrared Physics, 8, 259-264 (1968).

[29] P.P. Webb, and R.J. McIntyre, Bull. Amer. Phys. Soc. II, 15, 813 (1970).

[30] R.J. McIntyre, IEEE Trans. Electron Devices, ED-19, 703-713 (1972).

[31] W.G. Oldham, R.R. Samuelson, and P. Antognetti, IEEE Trans. Electron Devices, ED-19, 1056-1060 (1972).

[32] R.J. McIntyre, IEEE Trans. Electron Devices, ED-20, 637-641 (1973).

[33] R.J. McIntyre, J. Appl. Phys., 32, 983-995 (1961).

[34] G.E. Stillman, and C.M. Wolfe, in Semiconductors and Semimetals, Vol. 4, p. 357, R.K. Willardson, and A.C. Beer, eds., Academic Press, New York (1977).

[35] R.H. Haitz, J. Appl. Phys., 35, 1370-1375 (1964).

[36] R.H. Haitz, J. Appl. Phys., 36, 3123-3131 (1965).

[37] W. Shockley, and W.T. Read, Jr., Phys. Rev., 87, 835 (1952).

[38] R.N. Hall, Phys. Rev., 87, 387 (1952).

[39] F. Stockmann, Phys. Stat. Sol. (a), 20, 217 (1973).

[40] W. Shockley, and J. Last, Phys. Rev., 107

[41] K. Ziegler, and E. Klausmann, Appl. Phys. Lett., 26, 400 (1975).

[42] S.M. Sze, C.R. Crowell, and D. Kahng, J. Appl. Phys., 35, 2534 (1964).

[43] J.L. Hartke, J. Appl. Phys., 39, 4871 (1968).

[44] L. Esaki, Phys, Rev., 109, 603 (1958).

[45] C. Zener, Proc. Roy. Soc. (London), A 145, 523 (1934).

[46] D.R. Fredkin, and G.H. Wannier, Phys. Rev., 128, 2054 (1962).

[47] W. Franz, in "Tunneling Phenomena in Solids", p.13, E. Burstein, and S. Lundqvist, eds., Plenum Press, New York (1969).

[48] R.A. Logan, in "Tunneling Phenomena in Solids", p.149, E. Burstein, and S. Lundqvist, eds., Plenum Press, New York (1969).

[49] E.O. Kane, J. Phys. Chem. Solids, 12, 181-188 (1959).

[50] E.O. Kane, J. Appl. Phys., 32, 83-91 (1961).

[51] L. Esaki, and Y. Miyahara, Solid State Electron., 1, 13, (1960).

[52] A.G. Chynoweth, R.A. Logan, and D.E. Thomas, Phys. Rev., 125, 877 (1962).

[53] P. J. Price, Bull. Am. Phys. Soc., 5, 406 (1960).

[54] C.T. Sah, Phys. Rev., 123, 1594-1612 (1961).

[55] A.G. Chynoweth, W.L. Feldmann, and R.A. Logan, Phys. Rev., 121, 684 (1961).

[56] A. Goetzberger, and E.H. Micollian, J. Appl. Phys., 38, 4582-4588 (1967).

[57] A. Rusu, and C. Bulucea, IEEE. Trans. Electron Devices, ED-26, 201-205 (1979).

[58] A.G. Chynoweth, and G.L. Pearson, J. Appl. Phys., 29, 1103 (1958).

[59] W. Shockley, Solid State Elect., 2, 35 (1961).

[60] R.H. Haitz, A. Geotzberger, R.M. Scarlett, and W. Shockley, J. Appl. Phys., 34, 1581 (1963).

[61] S.M. Hu, J. Vac. Sci. Technol., 14, 17 (1977), and references therein.

[62] K. Tanikawa, Y. Ito, and H. Sei, Appl. Phys. Lett., 28, 285-287 (1976).

[63] C.L. Claeys, E.E. Laes, G.J. Declerck, and R.J. Van Overstraeten, in Semiconductor Silicon 1977, p.774, H.F. Huff, and E. Sirtl, eds., The Electrochemical Society, Inc., Princeton, New Jersey.

[64] G.A. Rozgonyi, P.M. Petroff, and M.H. Read, J. Electrochem. Soc., 122, 1725-1729 (1975).

[65] P.M. Petroff, G.A. Rozgonyi, and T.T. Sheng, J. Electrochem. Soc., 123, 565-570 (1976).

[66] H. Shiraki, Jap. J. Appl. Phys., 15, 1-10 (1976).

[67] H. Shiraki, in Semiconductor Silicon 1977, p.546, H.R. Huff, and E. Sirtl, eds., The electrochemical Society, Inc., Princeton, New Jersey.

[68] T. Hattori, Appl. Phys. Lett., 30, 312-314 (1977).

[69] T. Hattori, J. Appl. Phys., 49, 2994 (1978).

[70] S.P. Murarka, T.E. Seidel, J.V. Dalton, J.M. Dishman, and M.H. Read, J. Electrochem. Soc., 127, 716-724 (1980).

[71] A. Rohatgi, and P. Rai-Choudhury, J. Electrochem. Soc., 127, 1136-1139 (1980).

[72] T.Y. Tan, E.E. Gardner, and W.K. Tice, Appl. Phys. Lett., 30, 175 (1977)

[73] K. Yamamoto, S. Kishino, Y. Matsushita, and T. Iizuka, Appl. Phys. Lett., 36, 195 (1980).

[74] R. Newman, Phys. Rev., 100 700 (1955).

[75] A.G. Chynoweth, and K.G. McKay, Phys. Rev., 102, 369-376 (1956).

[76] P.A. Wolff, Phys. Chem. Solids, 16, 184 (1960).

[77] J. Shewchun, and L.Y. Wei, Solid State Elect., 8, 485-493 (1965).

[78] R.H. Haitz, Solid State Elect., 8, 417-425 (1965).

[79] G.G. Macfarlane, T.P. McLean, J.E. Quarrington, and V. Roberts, Phys. Rev., 11, 1245 (1958).

[80] A.I. Stoller, RCA REV., 31, 271 (1970).

[81] M.J. Declercq, L. Gerzberg, and J.D. Meindl, J. Electrochem. Soc., 122, 545-552 (1975).

[82] C.M. Parks, and C.A.T. Salama, Solid State Elect., 18, 1061-1067 (1975).

[83] A.S. Grove, O. Leistiko, and C.T. Sah, J. Appl. Phys., 35, 2695 (1964).

[84] N.J. Chou, and J.M. Eldridge, J. Electrochem. Soc., 117, 1287 (1970).

[85] C.M. Osburn, J. Electrochem. Soc., 121, 809 (1974).

[86] M. Lenzlinger, and E.H. Snow, J. Appl. Phys., 40, 278 (1969).

[87] C.M. Osburn, and E.J. Weitzman, J. Electrochem. Soc., 119, 603 (1972).

[88] Abel-Khalek Mohamed Zakzuk, J. Electrochem. Soc., 126, 1771 (1979).

[89] R. Poirier, and J. Olivier, Appl. Phys. Lett., 15, 364 (1969).

[90] T.H. Ning, Solid-State Elect., 21, 273 (1978).

[91] E.H. Nicollian, and C.N. Berglund, J. Appl. Phys., 41, 3052 (1970).

[92] D.R. Collins, S.R. Shortes, W.R. McMahon, R.C. Bracken, and T.C. Penn, J. Electrochem. Soc., 120, 521 (1973).

[93] R.K. Raymond, and M.B. Das, Solid-State Elect., 19 181 (1976).

[94] G.C. Schwartz, and V. Platter, J. Electrochem. Soc., 122, 1508 (1975).

[95] M.A. Green, F.D. King, and J. Shewchun, Solid State Elect., 17, 551-561 (1974).

[96] R.S. Ronen, and P.H. Robinson, J. Electrochem. Soc., 119 747 (1922).

[97] D.R. Young, and C.M. Osburn, J. Electrochem. Soc., 120, 1578 (1973).

[98] G. Baccarani, M. Severi, and Soncini, J. Electrochem. Soc., 120, 1436 (1973).

[99] R.J. Kriegler, Y.C. Cheng, and D.R. Colton, J. Electrochem. Soc., 119, 388 (1972).

[100] A. Rohatgi, S.R. Butler, and F.J. Feigl, J. Electrochem. Soc., 126, 149 (1979).

[101] J. Monkowski, J. Stach, and R.E. Tressler, J. Electrochem. Soc., 126, 1129 (1979).

[102] R.R. Razouk, and B.E. Deal, J. Electrochem. Soc., 126, 1573 (1979).

[103] C.M. Osburn, J. Electrochem. Soc., 121, 809 (1974).

[104] B.E. Deal, J. Electrochem. Soc., 121, 1980 (1974).

[105] A.J. Learn, Thin Solid Films, 20, 261 (1974).

[106] A.J. Learn, J. Electrochem. Soc., 123, 394 (1976).

[107] J.S.L. Leach, and P. Neuford, Corrosion Sci., 9, 413 (1969).

[108] P. Chaudhari, J. Appl. Phys., 45, 4339 (1974).

[109] C.J. Dell'Oca, and A.J. Learn, Thin Solid Films, 8, R47 (1971).

[110] T.E. Seidel, R.L. Meek, and A.G. Cullis, J.Appl. Phys., 46, 600 (1975).

[111] R.L. Meek, and T.E. Seidel, J. Phys. Chem. Solids, 36, 731 (1975).

[112] R.W. Santway, and R.S. Alwitt, J. Electrochem. Soc., 117, 1282 (1970).

[1A] G.A. Baraff, Phys. Rev., 128, 2507 (1962).

[2A] G.A. Baraff, Phys. Rev., 133, A26 (1964).

[3A] C.R. Crowell, and S.M. Sze, Appl. Phys. Lett., 9, 242 (1966).

[4A] G.E. Stillman, and C.M. Wolfe, in "Semiconductors and Semimetals vol.12" pp325-340, Academic Press, New York (1977), and references therein.

[5A] C.A. Lee, R.A. Logan, R.L. Bardorf, J.J. Kleimack, and W. Wiegmann, Phys. Rev., 134, A761 (1964).

[6A] J. Conradi, Solid State Electron., 17, 99 (1974).

[1C] S.M. Sze, "Physics of Semiconductor Devices," p.372, Wiley, New York (1969).

[2C] E.H. Nicollian, M.H. Hanes, and J.R. Brews, IEEE Trans Electron. Devices, ED-20, 380 (1973).

[3C] C.N. Berglund, IEEE Trans. Electron. Devices, ED-13, 701 (1966).

[4C] R. Castagne, C.R. Acad. Sci., B267, 866 (1968).

[5C] M. Kuhn, Solid-State Electron., 13, 873 (1970).

[1D] K.M. De Meyer, and G.J. Declerck, IEEE Trans. Electron. Devices, ED-28, 313 (1981).

[2D] H.W. Hanneman, and L.J.M. Esser, Philips Res. Rep., 30, 56 (1975).

APPENDIX A

Electron and hole ionization coefficients in silicon

For small electric field strengths the free carriers in a semiconductor are able to remain in thermal equilibrium with the lattice through impurity and acoustic phonon scattering, and conduction occurs at the band edges. At higher field strengths optical phonon emission dominates the scattering process and the drift velocity saturates. The current is no longer ohmic, however, conduction still occurs at the band edges. At still higher field strengths ($\sim 10^5$ V/cm in silicon) the carriers gain energy from the field faster than they can lose it by emitting phonons. At these and higher fields the carriers are no longer in equilibrium with the lattice and their energy relative to the band edge increases until they have acquired the threshold energy for impact ionization. The electron and hole ionization coefficients are used to describe the average distance a carrier will travel before generating an electron-hole pair by impact ionization.

The most general theory for the ionization rates is a modification of Baraff's theory [1A,2A] due to Crowell and Sze [3A]. The assumptions made by Baraff to obtain a numerical solution were:

(1) The energy bands in momentum space are parabolic.

(2) Only scattering by, and emission of, optical phonons is important, i.e., the lattice temperature was assumed to be very low. The resulting energy loss is, therefore, equal to the optical phonon energy E_r.

(3) The mean-free-path for optical phonon emission λ is independent of energy.

(4) The mean-free-path for impact ionization is constant for electron energies greater than the threshold ionization energy E_I.

(5) The scattering is spherically symmetric.

Crowell and Sze later improved upon this theory by including optical phonon absorption thereby enabling the temperature dependence of the ionization coefficients to be determined. They suggested that an average energy loss per collision $\langle E_r \rangle$ could be used in place of E_r in Baraff's Theory, and give the following expression for the temperature dependence of λ and E_r

$$\frac{\langle E_r \rangle}{E_r} = \tanh\left(\frac{E_r}{2kT}\right) = \frac{\lambda}{\lambda_o} \qquad \text{(A1)}$$

where λ_o is the low temperature limit to the mean-free-path for optical phonon generation. They also give the following approximation to Baraff's numerical solution for the ionization coefficients α (correct to within $\pm 2\%$ for $0.01<r<0.06$ and $5<x<16$),

$$\begin{aligned} \alpha\lambda = \exp \quad & (11.5r^2 - 1.17r + 3.9 \times 10^{-4})\, x^2 \\ & + (46r^2 - 11.9r + 1.75 \times 10^{-2})\, x \\ & + (-757r^2 + 75.5r - 1.92) \end{aligned} \qquad \text{(A2)}$$

where $r = \dfrac{\langle E_r \rangle}{E_I}$ and $x = \dfrac{E_I}{q\mathcal{E}\lambda}$

There is considerable uncertainty as to the exact magnitude and field variation of the electron and hole ionization rates in silicon [4A]. The ionization rates used in this work are those measured at room temperature by Lee et.al. [5A]. These are the most widely quoted ionization rates, and the only ones that can be fit to Baraff's theory with a reasonable value for the parameter λ, the accepted value of $E_I \simeq 3/2\ E_g$, and the measured optical phonon energy of $E_r = 0.063$ eV. The ionization rates were transformed to the appropriate temperature using Crowell's and Sze's modification to Baraff's theory, Equations (A1) and (A2). The parameters used in (A1) and (A2) that fit Lee's data are [6A]:

$$E_r = 0.063 \text{ eV}$$

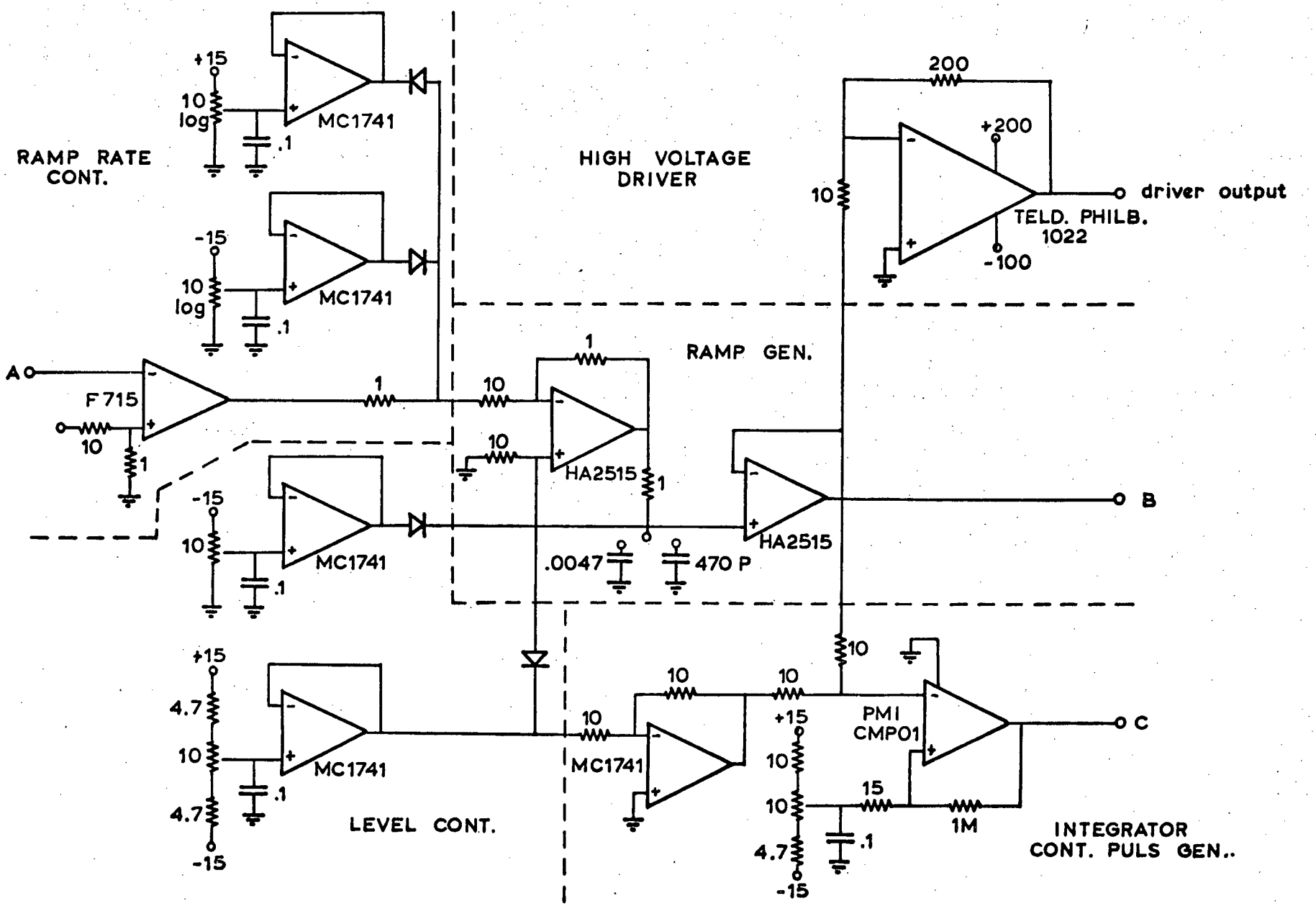
$$E_I \text{ (300K)} = 1.6 \text{ eV}$$

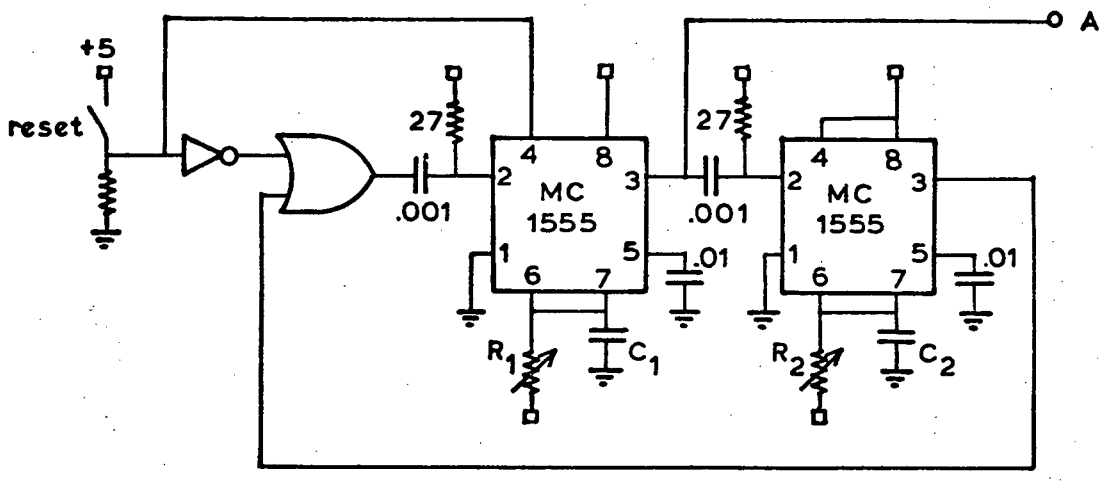
$$\lambda_o \text{ (electrons)} = 76 \text{ Å}$$

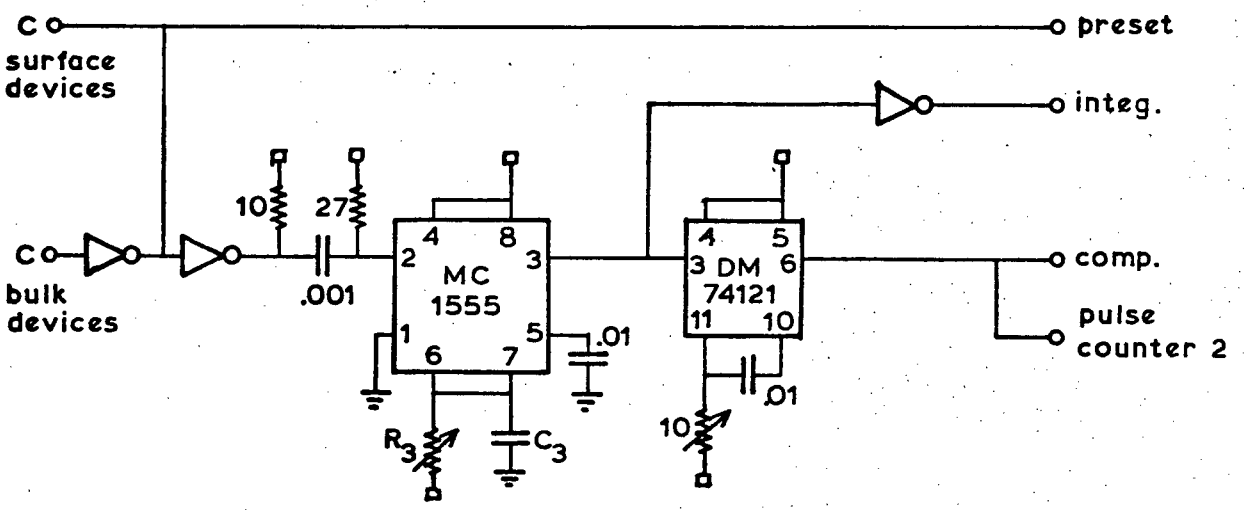
$$\lambda_o \text{ (holes)} = 49 \text{ Å}$$

APPENDIX B

Simplified schematics for the high voltage drive, timing circuitry, charge amplifier and discriminator.

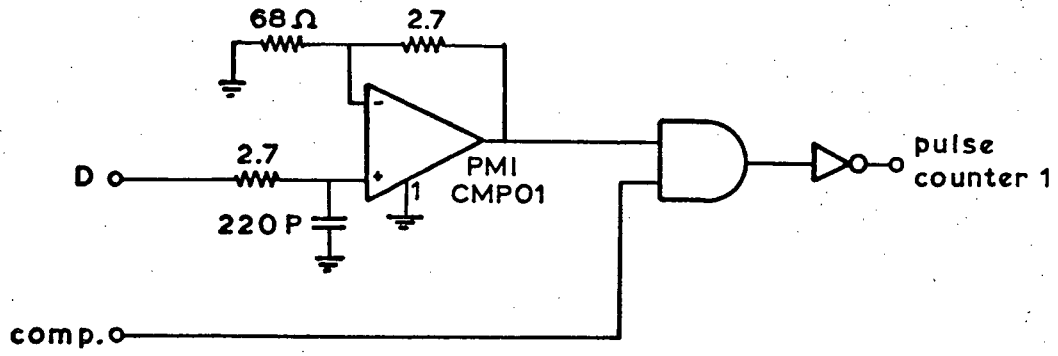

RAMP RATE CONT.
+15
10 log
MC1741
.1
-15
10 log
MC1741
.1
HIGH VOLTAGE DRIVER
200
+200
10
TELD. PHILB. 1022
-100
driver output
RAMP GEN.
A
F 715
10
1
1
10
1
10
HA2515
1
-15
10
MC1741
.1
.0047
470 P
HA2515
B
+15
4.7
10
4.7
-15
MC1741
.1
LEVEL CONT.
10
10
10
MC1741
10
+15
10
10
4.7
-15
.1
PMI CMP01
C
15
1M
INTEGRATOR CONT. PULS GEN..



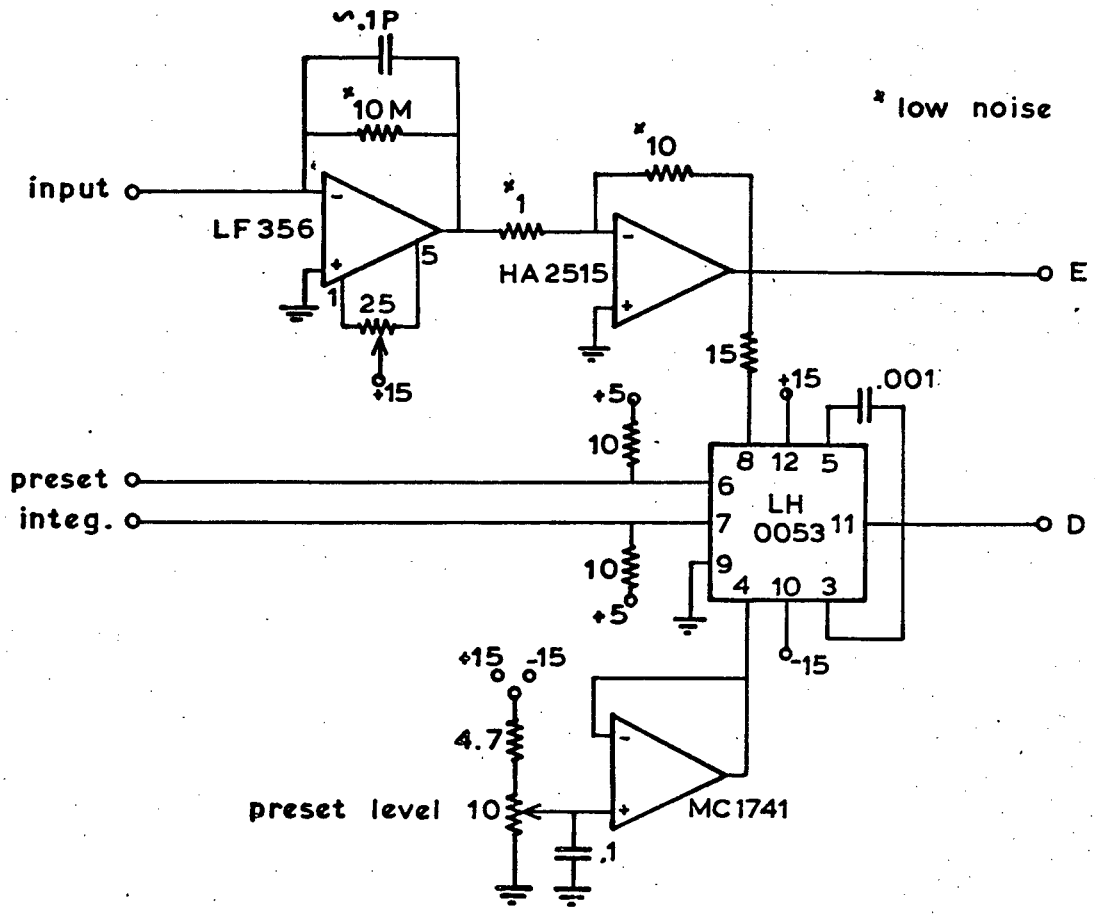


TIMING

	surface dev.	bulk dev.
R_1	10	100
R_2	100	10
R_3	10	10
C_1	.47	1
C_2	1	.1
C_3	.47	.1



DESCRIMINATOR



CURRENT AMP. AND
INTEGRATOR

APPENDIX C

Methods used to determine the doping profile and interface state density

(1) Doping Profile

The doping profiles were obtained by the well known dC/dV - method based on the following formulae (1C)

$$N(w) = \pm\frac{2}{q\epsilon_s}\left\{\frac{d(1/C_{sc}^2)}{d\phi_s}\right\}^{-1} \tag{C1}$$

$$w = \epsilon_s/C_{sc} \tag{C2}$$

where N = doping density

w = depletion layer width (i.e., distance from the semiconductor surface at which the doping density is determined)

C_{sc} = space charge capacitance per unit area

ϕ_s = semiconductor surface potential

The negative sign applies for n-type substrates and the positive sign for p-type. Equations (C1) and (C2) are based on the depletion approximation, which becomes invalid for small values of w where the majority carrier concentration can no longer be neglected in comparison to the doping density. This limits the determination of the doping density to values of w greater than 2 L_D where L_D is the extrinsic Debye length [2C].

$$L_D = \left\{\frac{2kT\epsilon_s}{qN}\right\}^{\frac{1}{2}}$$

In depletion the space charge capacitance is related to the measured high frequency MOS capacitanc C_{HF} according to

$$C_{sc} = \frac{C_{HF}C_{ox}}{C_{ox} - C_{HF}} \tag{C3}$$

provided the measuring frequency is high enough that the interface states do not contribute to the overall capacitance. C_{ox} is the oxide capacitance which, in the case of a low interface state density, can be approximated by the measured capacitance in strong accumulation. In order to use Eq. (C1) it is further necessary to obtain a relationship between the surface potential ϕ_s and the applied voltage V_g. Berglund (3C) has shown that the surface potential may be determined to within an additive constant by integrating the measured C(V) curve from a point corresponding to strong accumulation, V_{acc}, toward depletion,

$$\phi_s(V_g) = \int_{V_{acc}}^{V_g} \left\{1 - \frac{C_{HF}(V_g)}{C_{ox}}\right\} dV_g + K_1 \qquad (C4)$$

The additve constant K_1 drops out when performing the derivative in (C1).

It is customary to use pulsed C-V measurements in order to prevent the formation of an inversion layer and enable the doping profile to be obtained to greater depths. Because of the lack of such measuring equipment, a different method was used here. Samples with several 1mm dia MOS capacitors were mounted in 16 pin DIP packages and cooled to 80K in the cold chamber used for the device testing. Under total darkness the inversion layer charge collects so slowly that it was possible to obtain the C (V) curve point by point with a capacitance bridge. The measuring frequency used was 500 kHz.

(2) Interface State Density

The interface state density was obtained by combining quasi-static C-V measurements with high frequency C-V measurements, as discussed by Castagne [40]. The difference between the quasi-static (low frequency) capacitance C_{LF} and the high frequency capacitance C_{HF} is directly related to the sur-

face state density as follows

$$N_{ss}(\phi_s) = \frac{C_{ss}(\phi_s)}{q} = \frac{C_{LF}C_{ox}}{C_{ox} - C_{LF}} - \frac{C_{HF}C_{ox}}{C_{ox} - C_{HF}} \qquad (C5)$$

where N_{ss} is the interface state density per unit area per eV and C_{ss} is the capacitance due to interface states. The interface state densities obtained from Eq. (C5) are only valid when the MOS sample is in depletion. In inversion the minority carriers are not able to follow the high frequency ac signal used to measure C_{HF}, while in accumulation the interface states closest to the band edge are able to follow the ac signal, so that C_{ss} can no longer be determined. The constant in Eq. (C4), needed to determine $\phi_s(V_g)$, was obtained by plotting $1/C_{sc}^2$ versus $(\phi_s - K_1)$, which results in a straight line in depletion (with uniformly doped samples). $1/C_{sc}$ goes to 0 as ϕ_s goes to 0, therefore, the intercept gives K_1.

As with the doping profile measurements the interface state density was measured on 1 mm dia, p-substrate, MOS capacitors. The high frequency capacitance measurements were made with a 1 MHz (Boonton) capacitance meter. The quasi-static C(V) curves were obtained by measuring the MOS displacement current in response to a linear voltage ramp (10^{-2} V/sec), as described by Kuhn [5C]. The displacement current was measured on a Keithley model 602 electrometer (in the fast mode). In addition to the room temperature C(V) curves, quasi-static and high frequency C-V measurements were also made at 173K and 223K in order to obtain interface state densities closer to the valence band edge. At these lower temperatures the energy range over which the interface state information is valid becomes quite narrow because the deep levels can no longer follow the slow voltage ramp used for the quasi-static measurement (see Fig. 4.20).

APPENDIX D

Method used for the two-dimensional calculation of the potential and field distributions in the bulk breakdown devices

The structure for which the solutions will be derived is shown in Fig. D1

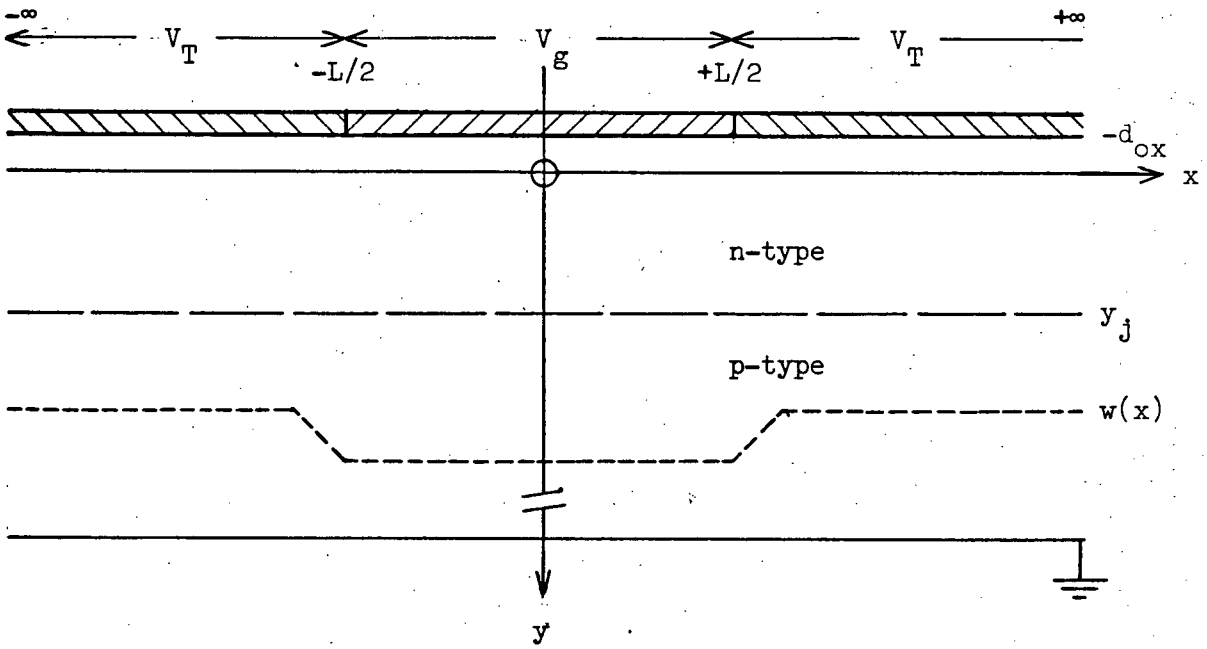


FIGURE D1 Device structure used for the two-dimensional model.

The method is analogous to that described by Meyer and Declerck [1D] for obtaining potential distributions in BCCD structures, and uses the superposition principle in such a way that all boundary conditions are satisfied. This is not the case in the method of Meyer and Declerck. They use the superposition principle in the oxide and n-layer only, and then match the solution in the substrate in such a way that the potential goes to zero at the edge of the depletion region. Their solution, however, does not satisfy the additional boundary condition grad $\phi(x,w) = 0$ and the potential distribution in the substrate cannot be accurately determined, particularly if

the substrate doping density is comparable to that of the n-layer as it is with the bulk breakdown devices.

In the method discussed below, the superposition principle is used throughout. This in fact results in a simpler solution as it eliminates the complicated matching condition between the n-layer and the substrate solutions.

A zero gate separation is assumed as shown in Fig. D1. By making the depletion approximation and using the superposition principle it is possible to obtain the solution for the electrostatic potential as the sum of two parts,

$$\phi_{si}(x,y) = \phi^{h}(x,y) + \phi^{p}_{si}(x,y) \tag{D1}$$

$$\phi_{ox}(x,y) = \phi^{h}(x,y) + \phi^{p}_{ox}(x,y) \tag{D2}$$

where $\phi^{h}(x,y)$ is the solution to the homogeneous problem obtained by setting the charge densities equal to zero, i.e., to the problem,

$$\nabla^2\phi^{h}(x,y) = 0 \ , \ -d_{ox} < y \tag{D3}$$

with boundary condition,

$$\phi^{h}(x,y_j) = V(x) \tag{D4}$$

and $\phi^{p}(x,y)$ is a particular solution for the case $V_g = V_T = 0$, i.e., to the one dimensional problem,

$$\frac{\partial^2\phi^{p}_{ox}(y)}{\partial y^2} = 0 \ , \ -d_{ox} < y < 0 \tag{D5}$$

$$\frac{\partial^2\phi^{p}_{si}(y)}{\partial y^2} = \frac{-\rho(y)}{\epsilon_s} \ , \ 0 < y \tag{D6}$$

with boundary conditions

$$\phi^{p}_{ox}(-d_{ox}) = 0 \tag{D7}$$

$$\phi^{p}_{ox}(0) = \phi^{p}_{si}(0) \tag{D8}$$

$$\epsilon_i \frac{\partial \phi^{p}_{ox}(0)}{\partial y} = \epsilon_s \frac{\partial \phi^{p}_{si}(0)}{\partial y} \tag{D9}$$

The additional boundary conditions needed to obtain ϕ_{ox} and ϕ_{si} are

$$\phi_{si}(x,w) = 0 \tag{D10}$$

$$\frac{\partial \phi_{si}(x,w)}{\partial y} = 0 \tag{D11}$$

$$\frac{\partial \phi_{si}(x,w)}{\partial y} = 0 \tag{D12}$$

The particular problem posed by Equations (D5)-(D9) is straight-forward. By using partial integration and the three boundary conditions, $\phi^{p}_{ox}(x,y)$ and $\phi^{p}_{si}(x,y)$ can be expressed as

$$\phi^{p}_{ox}(x,y) = C_1(y + d_{ox}) \ , \ -d_{ox} < y < 0 \tag{D13}$$

$$\phi^{p}_{si}(x,y) = \frac{y}{\epsilon_s}\int_0^y \rho(y)\,dy - \frac{1}{\epsilon_s}\int_0^y y\rho(y)\,dy + C_1\left(\frac{\epsilon_i}{\epsilon_s} y + d_{ox}\right) \ , \ 0 < y \tag{D14}$$

For the solution to the homogeneous problem posed by Equations (D3) and (D4), Meyer and Declerck use a method derived from the theorem of image forces [2D]. For the gate structure shown in Fig. D1, however, there is an easier solution based on conformal mapping techniques. By using the transformations,

$$x' = x \ , \ y' = y + d_{ox}$$

$$z'' = \ln\left\{\frac{z' - L/2}{z' + L/2}\right\} \ , \ z' = x' + iy' \tag{D15}$$

the electrostatics problem is transformed to that shown in Fig. D2, for which the following solution can be obtained by inspection,

$$\phi^h(x'',y'') = V_T + \frac{(V_g - V_T)}{\pi} y'' \qquad \text{(D16)}$$

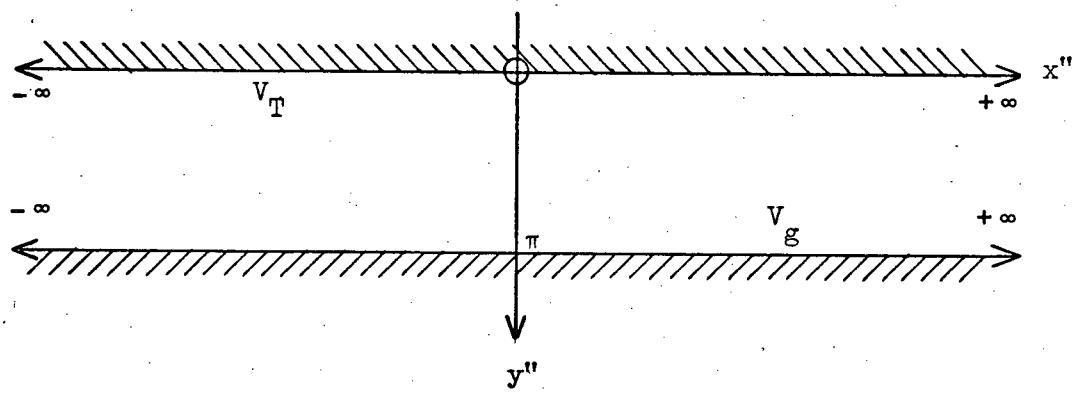


FIGURE D2 Device structure after conformal transformation.

After applying the inverse transformation the solution becomes

$$\phi^h(x,y) = V_T + \frac{(V_g - V_T)}{\pi} \left\{ \tan^{-1}\left[\frac{y + d_{ox}}{x - L/2}\right] - \tan^{-1}\left[\frac{y + d_{ox}}{x + L/2}\right] \right\} \qquad \text{(D17)}$$

The constant C_1 can now be obtained from B.C. (D11),

$$C_1 = -\frac{\epsilon_s}{\epsilon_i} \left\{ \frac{1}{\epsilon_s} \int_0^w \rho(y)\, dy + \frac{\partial \phi^h(x,w)}{\partial y} \right\} \qquad \text{(D18)}$$

and w is determined by an iterative procedure from B.C. (D10), which becomes

$$\phi^h(x,w) - \frac{1}{\epsilon_s} \int_0^w y\rho(y)\, dy - \frac{\epsilon_s}{\epsilon_i} d_{ox} \int_0^w \rho(y)\, dy - \left\{ w + \frac{\epsilon_s}{\epsilon_i} d_{ox} \right\} \frac{\partial \phi^h(x,w)}{\partial y} = 0 \qquad \text{(D19)}$$

where the derivative of (D17), with respect to y, is

$$\frac{\partial \phi^h(x,y)}{y} =$$

$$\frac{(V_g - V_T)}{\pi}\left\{\left[\frac{(x - L/2) + (y + d_{ox})^2}{(x - L/2)}\right]^{-1} - \left[\frac{(x + L/2) + (y + d_{ox})^2}{(x + L/2)}\right]^{-1}\right\} \qquad \text{(D20)}$$

Boundary condition (D12) follows automatically from B.C. (D10).

In order to solve for $P_e(w)$ and $P_h(0)$ equations (3.2) - (3.5) were integrated along a line starting from a point at the potential minimum (in the bulk or at the interface) and following the direction of the electric field vector (see Figures 4.26 in text). The x and y components of the electric field were obtained from the x and y derivatives of $\phi_{si}(x,y)$, as follows,

$$\mathcal{E}_x(x,y) = -\frac{\partial \phi_{si}(x,y)}{\partial x}$$

$$= \frac{\partial \phi^h(x,y)}{\partial x} + \left(\frac{\epsilon_i}{\epsilon_s} y + d_{ox}\right)\frac{\partial C_1}{\partial x} \qquad \text{(D21)}$$

$$\mathcal{E}_y(x,y) = -\frac{\partial \phi_{si}(x,y)}{\partial y}$$

$$= \frac{\partial \phi^h(x,y)}{\partial y} + \frac{1}{\epsilon_s}\int_0^y \rho(y)\,dy + \frac{\epsilon_i}{\epsilon_s} C_1 \qquad \text{(D22)}$$

$\partial C_1/\partial x$ can be obtained from B.C. (D12) as follows,

$$\frac{\partial C_1}{\partial x} = -\left(\frac{\epsilon_i}{\epsilon_s} w + d_{ox}\right)^{-1}\frac{\partial \phi^h(x,w)}{\partial x} \qquad \text{(D23)}$$

where the derivative of (D17) with respect to x is

$$\frac{\partial \phi^h(x,y)}{\partial x} =$$

$$\frac{(V_g - V_T)}{\pi}\left\{\left[\frac{(x + L/2)}{(y + d_{ox})} + (y + d_{ox})\right]^{-1} - \left[\frac{(x - L/2)}{(y + d_{ox})} + (y + d_{ox})\right]^{-1}\right\} \qquad \text{(D24)}$$