OFFICE OF NAVAL RESEARCH

GRANT N00014-89-J-1178

R&T CODE 413Q001-05

TECHNICAL REPORT NO. #50

INVESTIGATION OF TRAP EMISSION KINETICS IN MOS CAPACITORS USING A PUMP-PROBE CHARGE INTEGRATING TECHNIQUE

BY

J.C. Poler and E.A. Irene
Department of Chemistry
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-3290

Submitted to:

Applied Physics Letters

Reproduction in whole or in part is permitted for any purpose of the United States Government.

This document has been approved for public release and sale; its distribution is unlimited.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No 0704 0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
5/4/93

3. REPORT TYPE AND DATES COVERED

4. TITLE AND SUBTITLE
Investigation of Trap Emission Kinetics in MOS Capacitors using a Pump-probe Charge Integrating Technique

5. FUNDING NUMBERS
#N00014-89-J-1178

6. AUTHOR(S)
J.C. Poler and E.A. Irene

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
The University of North Carolina
Chemistry Department
CB #3290 Venable Hall
Chapel Hill, NC 27599-3290

8. PERFORMING ORGANIZATION REPORT NUMBER
Technical Report #50

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research
800 N. Quincy Street
Arlington, VA 22217-5000

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
None

12a. *DISTRIBUTION/AVAILABILITY STATEMENT*
This document has been approved for public release and sale, distribution of this document is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

We have developed a Pump-Probe charge integrating measurement technique for studying the emission kinetics of traps in the $M/SiO_2/Si$ system. Essentially, an MOS capacitor is pumped by exposure to a charging pulse. The emission of the charge at short time scales ($<$10ms), can be measured using a delayed application of a probe pulse, that determines the remainder of the filled traps as a function of delay time. For MOS capacitors grown on a lightly doped p-Si(111) substrate, we observe an uncommon behavior of the emission kinetics in the initial time regime ($<$100ms). A possible explanations for this phenomena is the perturbation of the emission cross-section ofthe probed traps due to the presence of another state in communication with the trap site. Our results on this system will be presented along with a comparison to other substrate types and processing parameters.

14. SUBJECT TERMS
Emission kinetics in MOS

15. NUMBER OF PAGES

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT
Unclassified

18. SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

19. SECURITY CLASSIFICATION OF ABSTRACT
Unclassified

20. LIMITATION OF ABSTRACT

Investigation of trap emission kinetics in MOS capacitors using a pump-probe charge integrating technique

J.C. Poler and E.A. Irene, Department of Chemistry, Venable Hall, University of North Carolina Chapel Hill, NC 27599-3290

We have developed a Pump-Probe charge integrating measurement technique for studying the emission kinetics of traps in the M/SiO_2/Si system. Essentially, an MOS capacitor is pumped by exposure to a charging pulse. The emission of the charge at short time scales (<10ms), can be measured using a delayed application of a probe pulse, that determines the remainder of the filled traps as a function of delay time. For MOS capacitors grown on a lightly doped p-Si(111) substrate, we observe an uncommon behavior of the emission kinetics in the initial time regime (<100ms). A possible explanations for this phenomena is the perturbation of the emission cross-section of the probed traps due to the presence of another state in communication with the trap site. Our results on this system will be presented along with a comparison to other substrate types and processing parameters.

I. INTRODUCTION

The capture of electrons and holes at localized states in SiO_2 is a well studied topic[1]. The most common way to measure the quantity of filled traps is by analyzing the flat band voltage shift from a C(V) curve[2]. This method is only applicable when the charge stays trapped for a time longer than it takes to acquire the C(V) data. To study the trapping kinetics the trap occupation is monitored as a function of time during the trapping process. Traps can be filled by a variety of methods. Avalanche injection[3], internal photo-emission[4] and tunneling injection[5] can all be used to supply carriers that are trapped in the oxide. The emission of charge from traps has been studied using thermally[6], optically[7] and electrostatically[8] stimulated techniques. All of these techniques measure detrapping on a long time scale, typically one to several thousand seconds. Traps with deep energy levels with respect to the conduction bands require stimulation for emission. Capacitance transient spectroscopy[9] (CTS) and charge pumping[10] (CP) can probe trap capture and emission kinetics at much faster time scales. When the sample is held at low temperatures, filled traps will not emit their charge. The CTS applies short bias pulses to stimulate detrapping of the charge. The occupation of the traps is monitored by measuring the capacitance of the device junction. This technique can measure the emission kinetics on the microsecond time scale. CP is used to measure interface trap properties. This method can also access very short time scales but requires a MOSFET device structure and the data is difficult to interpret.

In the present study we have used tunneling injection to examine the trapping kinetics in ultra thin SiO_2 films on Si. By applying bias pulses of a constant level, the charge that is injected into the device is integrated. The system is allowed to relax to its equilibrium uncharged state before applying the next pulse. To this end we have developed a pump-probe (PP) pulsing technique to examine the trap emission kinetics of MOS capacitcrs on the millisecond time scales. This technique is similar to the methods that are used to study chemical kinetics of species in solution and the gas phase, but using electrostatic pulses instead of optical ones[11]. This technique measures the trap kinetics at room temperature and the emission of charge from the traps is self initiated without the aid of external stimuli. The observation of detrapping under these conditions indicates that the trap is very shallow, where the charge is thermally emitted over the trap's potential well. Alternatively, the trap may be located very close to a contact where the charge can tunnel through the thin potential barrier (i.e. an interface trap).

There are a variety of traps that have been identified. They are usually characterized by associating the measured capture cross section and emission energy with the impurity or physical process that is responsible for their creation. The capture cross section is evaluated by fitting the trapped charge concentration as a function of time, Q(t), to a first order kinetic rate equation[12],

$$Q(t) = Q_{\infty}\left(1 - e^{-t/\tau_c}\right) , \qquad (1)$$

where τ_c is the capture time constant and is inversely proportional to the capture cross section.

II. EXPERIMENTAL

A. Sample Preparation

All of the MOS capacitors were grown by thermal oxidation of silicon. Si substrates of <111> and <100> orientations and of various concentrations of p-type doping were examined. All of the substrates were RCA cleaned[13] followed by an hydrofluoric acid dip to remove the native oxide. Different samples with varying processing parameters have displayed markedly different trapping and emission kinetics. We will use the following abbreviations to refer to the samples that fall into categories with similar trapping and emission kinetic behavior.

Sample ID	Description
A1	n$^+$-poly/SiO$_2$/p-Si(100)[11-25Ω·cm] O$_2$ 2%HCl 800 °C oxidation with Post Metalization Anneal (PMA)
B1	n$^+$-poly/SiO$_2$/p-Si(100)[0.5-1Ω·cm] dry O$_2$ 1050 °C Rapid Thermal Oxidation (RTO) for 60s with PMA
C1	Al/SiO$_2$/p-Si(100)[1-2Ω·cm] dry O$_2$ 800 °C oxidation Thermal No PMA (NPMA)
D1	Same as C1 but substrate is degenerately doped to ~0.002Ω·cm p-type
C2	Same as C1 but substrate has the <111> orientation
D2	Same as D1 but substrate has the <111> orientation

All of the C and D samples were grown in a double walled tube furnace in dry O$_2$. Aluminum gate contacts were evaporated onto the oxide through a shadow mask. These samples did not initially receive a PMA. Back side contact to the Si substrate was via a GaIn eutectic paste. The A1 sample was also grown in a tube furnace but the O$_2$ contained 2%HCl. Gate contacts were made via degenerately doped poly-Si with lithographically defined areas. The B1 sample was grown using an ultra dry rapid thermal oxidation (RTO) system. Gate contacts were again poly-Si and back side contact for both of these types of samples was accomplished via a blanket deposition of aluminum. The A1 and B1 samples all received a standard PMA of ~20min at 400 °C in forming gas. The gate contacts on all of the samples were about 0.001cm^2 in area and the oxide thicknesses were typically 50Å.

B. Pump-Pulse Apparatus

Figure 1(a) illustrates the sequences of pulses applied to the MOS structure to determine the emission kinetics. The pump pulse initialized the system and filled the traps. The pump is characterized by its height (V_{pmp}) and width (t_{pmp}). The polarity of all of the pulses is positive on the substrate, driving the p-type Si surface potential toward accumulation. After pumping the system, the traps are allowed to emit while the device is shorted to ground. After a delay time (t_d) a probing pulse of height (V_{prb}) and width (t_{prb}) is applied and the resulting flow of charge is integrated using the pulsed I(V) acquisition module. This charge is a measure of the traps emission during t_d. Following the probe pulse, the system is allowed to equilibrate for a predetermined "infinite" delay time (t_∞), typically 20 to 120s. By varying the delay time from 1ms to t_∞, we probe the kinetics of the trap emission processes in the device.

The pulse application and timing circuit is illustrated in Fig. 1(b). The probe pulse and charge integration is regulated by a pulsed I(V) module described elsewhere[14]. If a large current flows through the device under test during the pulses, there will be a drain on the battery cells. To avoid a drop in the pump pulse height due to current saturation of the batteries, a 1000μF electrolytic capacitor is charged by the battery source and acts as the constant voltage source for the system. The applied bias is carried to the substrate contact, in the Faraday black box, by a twisted conducting pair. A 0.1μF ceramic capacitor is connected in parallel with the pump source and in series with the probe source at the substrate contact. A 5V relay is used to switch the pump signal (at C2) from ground to the bias source, and back to ground at the completion of the pump pulse. A 8254 programmable interval timer chip was used to set and initiate the pump pulse width. The width of the pump pulse is determined by the count N_1, loaded into the counter and the clock frequency. The counter is operated in the read most significant byte only mode, so that we can vary N_1 from 256 to 65280. The clock frequency is varied, using a HCT153 multiplexer and the 12 stage counter from 8Mhz to 1953Hz. This combination allows pump pulse widths from 32μs up to 33.6s. Because of switch bouncing at the relay, we limit the pump pulse widths to greater than 0.5ms. The delay time between the pump and probe pulses can be adjusted to within 1ms with minimal software overhead time. The pulse shapes and timing sequences were characterized using a 400MHz digital oscilloscope.

To test the utility and accuracy of the PP acquisition design, we characterized the charging and discharging of a polystyrene capacitor with C = 500pF in series with a resistor of R = 1000MΩ. The emission of charge from the capacitor through the resistor can be described by a single exponential and the decay rate can be described by the RC time constant. With a constant probe pulse shape, the discharging kinetics was measured using various combinations of pump pulse heights and widths. After the pump pulse, the capacitor is initially fully charged. To immediately probe the device with the second pulse, there would not be any measured injected charge. However, the time between the pulses to increases, the device has time to release some of the charge that was collected during the pump pulse. Subsequent probing of the device, will re-charge the capacitor. The difference between the charge collected with the probe pulse, had the capacitor been fully discharged, Q_∞, and the amount collected with the probe after waiting a shorter delay time, t_d, is the amount of charge still on the capacitor. By varying t_d the charge on the capacitor as a function of time $Q(t)=Q_\infty - Q(t_d)$ is determined. By fitting the Q(t) versus t data to a single exponential function we determined the RC time constant to be 0.48-0.72s, which is about what is theoretically expected ($R \cdot C=0.5s$). The time constant did not depend on the characteristics of the pump pulse. With this result we are confident that this PP technique can accurately and reproducibly determine the residual charge left on or in a capacitor as a function of time.

III. RESULTS

A. Charge trapping

Figure 2 shows the injected charge versus pulse width for a 50Å A1 MOS capacitor (solid circle). The charge is shown to be logarithmically dependent on the injection time (i.e. $Q(t) = 61.88 + 1.68\text{Log}(t)$) and not in accord with the first order rate equation, Eqn. 1. Wallmark[15] shows that the charge trapping in an $Al/Si_3N_4/SiO_2$/p-Si gate of a MOSFET, does exhibit a logarithmic dependence on pulse width if the traps are spatially distributed into the insulator. Their conclusions are derived on the assumption that the film thickness and trap distribution are much greater than the electron wave length in the system. This approximation is not as valid for our samples. The results from a ~60Å C1 are shown in Fig. 2(solid square) along with a fit of the data (τ=5ms, Q_∞=380pC) (dashed) to Eqn. 1. The data is shown to have a sharper transition than would be predicted for a simple first order reaction. A quick survey of some of the simple low order rate equations[16] indicates that the observed trapping phenomenon is either convoluted with indeterminate experimental artifacts or exhibiting a non-trivial kinetic behavior.

B. Charge injection versus transport

The thin film MOS capacitor has a finite resistance due to tunneling of charge through the insulating film. We are proposing to measure charge injection and trapping in the oxide, not transport through the device. A charge integration technique, like the one we employ, can not differentiate between a charge traversing the entire device, and a charge moving from one electrode and becoming localized within the device. In addition, this technique can not determine the polarity or identity of the conducting species. However, we can use the characteristics of the injected charge versus pulse height and width to possibly support or contradict a charge transport model.

Figure 3 shows the results of the injected charge Q_{prb} as a function of pulse height V_{prb}. The pulse width was held constant at t_{prb}=5ms. The measurements were made on a ~50Å C2 MOS capacitor. The device was allowed to relax to its equilibrium uncharged state after each pulse. With the low bias limit of our measurement of V_{prb}=1.7V, in this range it seems as though the injected charge is insensitive to the bias of the pulse. As V_{prb} is raised above 2V the injected charge is linearly dependent on the pulse height (Q_{prb} = 322.6 + 13.1$\cdot V_{prb}$ pC). Tunneling currents are not linearly dependent on the applied bias in this voltage range and for these film thicknesses[17]. A valid interpretation of the linear dependence of Q_{prb} on V_{prb} is that the pulse is filling traps that are uniformly distributed in energy within the forbidden band of the oxide film. As the pulse height approaches 5V. FN conduction starts to dominate the integrated charge signal.

To test that we are measuring charge trapping and not some unknown conduction mechanism, we have explored the time dependence of the injected charge at constant applied biases. Figure 4 shows the results of the injected charge dependency as a function of pulse width. The pulse height is held constant at V_{prb}=1.7V and we allow the system to reach equilibrium after each pulse before acquiring the next data point (t_{∞}=20.0s). These results are for the same device measured in Fig. 3. We show that the injected charge is logarithmically dependent on the pulse width. If we were measuring transport through the device we would find a persistent current, but rather an injection of charge that decreases logarithmically with time is observed. We shall show below that there is not enough trapped charge to continuously lower the applied potential of the pulse as its width increases (discussed below), which could cause the decrease in the apparent current. Therefore, except for an initial transient from the pulse[14], the current should be independent of the pulse width, which it is not. As discussed above, a logarithmic dependence of trap filling as a function of time is not described by a simple first order reaction. Although we do not provide a quantitative model for charge trapping, our results do indicate the filling of trapping centers with the injected charge. The trapping kinetics of sample C2 in Fig. 4 differ from those of sample C1 in Fig. 2(solid square) suggesting some substrate orientation dependence on the charge trapping efficiencies. The results in Fig. 4 are characterized by three distinct regions. The short (<1ms) and long (>100ms) pulse regions have a similar slope (i.e. 3.2pC/decade) but different y-intercepts. The transition between these two regions indicates the additional filling of traps that are not accessible to the shorter pulses. This result is sample dependent and will be discussed below in more detail.

C. Pump-Probe analysis of MOS devices

In our analysis of the PP trap emission data we define the number of filled probed traps after a delay time t_d, as $\Delta Q(t_d) = Q_{\infty} - Q(t_d)$. This is the number of traps still full after a delay of t_d from the end of the pump pulse, that would have been otherwise filled by the probe pulse. We are particular with this definition so that we may describe the concept of filling other "non-probed" traps, and their possible effect on the emission dynamics of the probed traps. In Fig. 5 we compare the trap emission curves, $\Delta Q(t_d)$ versus $\log(t_d)$, for various pump pulse heights and all other parameters held constant. The logarithm of the delay time is used because of the excursion over four orders of magnitude in time. Although this has the effect of magnifying the small changes of $\Delta Q(t_d)$ in the short time regime, these observations are both experimentally significant and reproducible. The results are identical when remeasured on the same device, as are the kinetics on different devices and different samples when normalized to Q_∞. The results presented here are for ~50Å C2 MOS capacitors.

The emission rate of the charge from the probed traps is the slope of a $\Delta Q(t_d)$ versus t_d curve. This is not the coordinate system that we use to describe our data, therefore visual determination of the emission rate from the $\Delta Q(t_d)$ versus $\log(t_d)$ coordinates can be deceiving. However, it should be apparent from Fig. 5 that the emission kinetics does depend on the bias level of the pump pulse that fills the traps. If the interactions of the filled traps with either themselves or other centers in the "lattice" were negligible, and all of the traps were identical, the emission kinetics should obey a simple first order law as:

$$\Delta Q(t_d) = [T]_{full} - [T]_{empty} + q \quad , \tag{2}$$

where [T] represents the concentration of traps. The occupation of the traps, in the above reaction after the pump pulse can be described by a single exponential kinetic rate equation:

$$\Delta Q(t_d) = Q_\infty \cdot \exp[-t_d/\tau] \quad , \tag{3}$$

where τ represents the reaction rate constant. The PP data for the 0.5V pump pulse in Fig. 5(open circle) can be described by a single exponential decay with τ = 83.3ms. The long t_d tail of the data seems to be fit by a slightly different rate constant, but the experiment is at its resolution limit in this region. As the pump pulse level is increased, the number of probed traps also increases as expected. However, the emission kinetics changes as the pump is increased. For V_{pmp}<2V, the decay constant for the trap emission reaction remains nearly constant. The slope of the emission curves in the short time regime (<10ms) increases as more of the probed traps are filled. Above V_{pmp}=0.75V, the first order decay kinetic model begins to break down, and we are unable to fit the data to a single exponential rate constant. This indicates that as the concentration of the probed traps increase, the emission mechanism is altered. For pump pulses above 2V we show that the probed traps are filled to saturation. A comparison of the results between V_{pmp}=3V and V_{pmp}=2V for t_d<20ms shows identical trap emission kinetics. However at longer t_d, the behavior of the system is markedly different. Since we show that all of the probed traps are filled by the pump pulse, and that there are differences in the emission kinetics as we raise the pump level further, we suggest that there is something that is perturbing the probed traps within the device.

Another possible explanation of these results is that the decrease of charge injection during the probe pulse is due to a voltage offset established by the trapped charge in the device from the pump pulse. However, this model fails when we consider, among other things, that the amount of injected charge is small (<400pC), and this could only produce a voltage shift of less than 0.4V in these samples. In Fig. 5 the difference in $\Delta Q(t_d)$ for V_{pmp}=3V compared to V_{pmp}=2V at t_d=100ms is 98pC. The results of Fig. 3 show that to decrease the injected charge by this amount there would have to be a 7.5V internal voltage shift due to the pump pulse, being sustained for ~100ms, which doesn't seem likely.

In Fig. 6 the effect that the width of the pump pulse has on the emission kinetics of a ~50Å C2 MOS capacitor (solid) is shown. The pump height was held constant at V_{pmp}=3.00V and the device was allowed to equilibrate for 30s between the acquisition of the data points. Pump pulse widths of 0.05s (circle), 1.00s (square) and 5.00s (diamond) are used. In the short t_d regime, the emission rate of charge from the traps is identical for all three pulse widths. But as t_d increases, so does the differences in the emission rates. The derivative of an exponential is a monotonically decreasing exponential. The derivative of the $\Delta Q(t_d)$ versus t_d is like an exponentially decreasing function, but there is a small peak near $t_d \sim \tau$. This peak is small compared to the rate of change of $\Delta Q(t_d)$ in the short t_d regime. To emphasize the local maximum in the emission rate, we plot the slope of the $\Delta Q(t_d)$ versus $\log(t_d)$ curve. The magnitude of the slope of the $\Delta Q(t_d)$ versus $\log(t_d)$ curves are shown for the various pump widths in Fig. 6(dashed). We show that as the pump width increases, the position of the local maximum in the trap emission rate is observed at longer delay times. If we interpret the position of the local maximum to approximate the decay time τ, then these results show that the pump pulse is perturbing the system so as to increase the decay time and decrease the emission rate of the traps.

Figure 7 summarizes the trap emission kinetics as a function of pump pulse shape. These results are specific for the C1 and C2 MOS capacitors. We observe similar behavior for samples with 50Å, 85Å and 217Å thick oxide films. This confirms our hypothesis that we are measuring charge trapping in the oxide and not charge transport through the oxide. The pH of a final buffered hydrofluoric acid (BHF) dip of the silicon before oxide growth has been shown to effect the structure of the surface defects[18]. If these defects are responsible for the observed phenomenon then varying the pH of the acid etch might perturb the observed kinetic behavior. Initial studies have shown that the qualitative nature of the trap emission kinetics is not dependent on the pH of the BHF wafer cleaning procedures.

Figure 7 shows that the emission kinetics are independent of the pump pulse width when V_{pmp} is less than 2V. Above a threshold pump bias (~2V) the emission kinetics is dependent on the width of the pump pulse. The thicker film samples have a higher pump threshold bias. As t_{pmp} increases, the emission of the charge from the traps is pushed to longer delay times. When the pump width is less than a threshold value (~10ms), the position of the maximum in the slope of the kinetics curve is independent of the applied bias. The results suggest that when the pump pulse is above a bias threshold and is wide enough, it not only fills the probed traps but it also perturbs the system. This suggest that there are (at least) two distinct types of trapping centers in the oxides grown on the C1 and C2 samples (possibly interface states and bulk oxide traps). Reviewing the results of Fig. 4, we observed a transition between two regions of the charge versus pulse width data. The slopes of the short t_{prb} and long t_{prb} regions of the Q_{prb} versus $\log(t_{prb})$ data are typically similar. This implies that the traps that are being filled in the short pulse region are filling with a similar efficiency in the longer pulse region. However, there is a vertical offset separating the regions were the longer pulses are injecting a constant amount of extra charge. For the data in Fig. 4 there is an additional 15.9pC of traps being filled that were not accessible to the shorter pulses. This supports our hypothesis that there is a second trap that is being filled for longer pump pulse widths. In general, our probe pulse widths are below the transition region of Fig. 4 and the pump pulses are above the transition region.

When we increase the probe pulse width to greater than the transition region width, we see an upward vertical offset of the emission kinetics curve. The shape of the emission curve is identical for most t_{prb}. Varying t_{prb} should not alter the shape of the emission curve, since the probe pulse only fills the traps that have been emptied during the delay time t_d. The transition region of Fig. 4 for probe bias levels from 1.7V (our minimum bias limit) and higher is observed. This confuses our argument since we only see a pump pulse dependence on the emission kinetics for V_{pmp}>2V. It may be that the kinetic effect is not sensitive enough to pick up the affect of the additional traps at the lower concentration levels. When we increase the pulse bias to 2V, the vertical offset of the transition regions increases by 80% and increases 160% for the 3V pulse heights.

These results support our model that the pump pulse is filling the probed traps and an additional "non-probed" trap. The non-probed traps have smaller capture cross sections and therefore take longer pulses to significantly populate. The non-probed traps are described by an energy level higher than that required to fill the probed traps. The non-probed traps appear to interact with the probed traps. This interaction is observed as a perturbation of the emission kinetics of the probed traps. The interaction is occupation dependent where a filled non-probed trap decreases the emission probability of the filled probed traps. Common to all of the emission kinetics curves, where the pump pulse is above the bias and width thresholds, is the observation of two distinct regions of trap emission. For short t_d, the emission rates are identical, independent of the pump pulse shape. After a characteristic delay time (τ_t), the emission kinetics goes through a transition where the probed trap occupation falls off rapidly. As the pump pulse width is increased, τ_t also increases.

D. Substrate and processing effects on emission kinetics

Figure 8 shows the emission kinetics for several samples. The results are normalized to Q_∞ for ease of comparison. The charge injection levels were different for the three samples shown. The C2 (circle) had a Q_∞=439pC where the A1 (triangle) and D2 (square) had a Q_∞ of 57pC and 14pC respectfully. The pump pulse bias was 3V and the pump pulse width was examined for t_{pmp}=1.0s and t_{pmp}=0.05s. All three of the samples exhibit different kinetic behavior. The A1 sample has a poly-Si gate contact where the other two have Al gate contacts. The B1 sample (not shown) looks similar to the A1 sample and it too has a poly-Si gate contact. The C1 and D1 samples, with Al gate contacts, have similar emission kinetics as the C2 and D2 samples respectfully. Therefore we do not observe any qualitative substrate orientation dependence on the trap emission processes. Comparing the lightly doped substrate, Al gate contact sample and the degenerately doped substrate, Al gate contact sample we observe markedly different levels of traps, and trap emission behavior. Since the only difference between these samples is the substrate doping concentration, we are confident that the traps we are measuring are very close to the oxide/substrate interface. Their proximity to the substrate contact would also explain the efficiency of the zero field emission that we have measured. Tunneling through the trap's potential well to the substrate contact is most likely the dominant emission mechanism.

It is shown in Fig. 8 that only the C1 and C2 samples have trap emission kinetics that depend on the shape of the pump pulse. The <111>/<100>DD samples do not show any pump pulse dependence of the emission kinetics, but we may be limited by the capabilities of the acquisition method. We can not measure the trap emission at delay times shorter than 1-2ms. Although we can measure the 14pC of charge, we are approaching the level of possible residual charges measured by the circuitry. The kinetics curve may be shifted to much shorter times and we are therefore measuring the long t_d tail, which is less sensitive to the pump pulse shape. In general, whenever the concentration of injected charge is less than 100pC (10nC/cm^2) we do not observe any pump pulse dependency of the emission kinetics. In addition, for the samples that do not exhibit this phenomenon, we do not measure a transition region in the Q_{pmp} versus log(t_d) data (see Fig. 4). This indicates that the filling of the extra traps during the longer pump pulse is responsible for slower emission kinetics of the probed traps.

If the emission kinetics of the probed traps for the C1 and C2 samples depend on the interactions with themselves and/or another non-probed trap, then this dependence should scale with the concentration of the traps. We do not know the mechanism that creates the traps in the oxide, nor do we understand why the concentration of the traps is dependent on the substrate doping levels. If the trap was associated with a dopant impurity, then the degenerately doped substrate samples should exhibit far more trapping than the lightly doped samples, which it doesn't.

IV. DISCUSSION

Our model implies that there are interactions within the oxide that alter the emission kinetics of the traps. Although we are not certain of the identity of the traps or the trapped species, we are proposing that the trap is located near the Si/SiO_2 interface. The trap should be near the Si interface, since the observed results are dependent on the doping concentration of the substrate. The trapped species are most likely holes injected from the substrate accumulation layer. This is consistent with the observation that we see similar trapping levels and emission kinetics for various film thicknesses. The incapacity of electrons to tunnel ~200Å from the Al contact to the trap, with an applied bias <3V, supports this conclusion.

While the identity of the trapping species is not clear from our results, we have shown that the charge injection into the traps increases linearly with applied bias (and band bending). This suggests that we are filling interface states, created from the unsatisfied bonds (dangling bond) of the trivalent Si at the Si/SiO_2 interface. However, the reported capture and emission times[19] of the interface traps are much faster than the rates we observe in our results, implying that these are not the same species that we observe.

Assuming that there are two distinct (or related) hole traps near the Si/SiO_2 interface, how does the occupation of one trap interact with the emission kinetics of the other? The occurrence of random noise events in electronic signals has been attributed to the complex and random nature of traps capturing and emitting charge within the device. Random telegraph noise (1/f noise) and random telegraph signals (RTS) in a variety of devices and materials have been studied[20]. The work of Ralls *et al*[21] expanded the study of RTSs to very small MOSFET devices, where the detection of a single trapping event was possible. The trapping and emission of an electron in the gate oxide was shown to alter the channel conductance in the switch. The study of RTSs has been extended to the MOS tunneling diodes by Farmer *et al*[22]. They studied the changes in resistance of ~20Å thick MOS capacitors as the occupancy of electrons in the traps fluctuated as a function of time. They observed a switching between two distinct resistance levels designated two-level fluctuations (TLFs). This has been shown to be due to local barrier perturbation from the coulomb potential of the trapped charge[23]. Kirton and Uren[24] invoke electron-lattice interactions to explain the complex behavior of the RTS. Their model implies that the trap center alters its configuration through electron capture and multiphonon emission. Farmer argues that the defect configuration is thermally activated and switches between metastable states in which electrons can elastically tunnel in and out[25]. The transformation of the trap configuration is dependent on the local stresses in the oxide lattice.

Farmer's *et al* results have shown distinct characteristics of the TLFs as a function of temperature and applied bias. Some of their observations show two distinct TLFs where the occupation state of one is related to the occupation state of the other. It is argued that the interaction between the traps is through long range lattice deformations induced by trap reconfiguration forces. The lattice deformations of one trap effects the ability of another trap to reconfigure and trap or emit an electron. For some of the TLF interactions they found that the emission of an electron from one TLF, allowed the emission of an electron from the second TLF. They also observed the contrary effect, where emission of an electron from the first TLF enables the capture of an electron into the second TLF. The capture process in the second TLF was never observed (over a two hour period) when the first TLF was full. Another interesting observation of Farmer's work is that the transition frequency and magnitude of the resistance changes were orders of magnitude different than would be expected from single electron capture into a single trap. The magnitude of the fluctuation was typically three orders of magnitude too large. This can be explained if the switching was due to the correlated reconfiguration of an ensemble of trapping centers. The switching times are many orders of magnitude too slow, as would be expected from the fluctuations of an ensemble of particles instead of a single independent trap.

An analysis of the amplitude of some of the larger TLFs observed by Farmer from a $1\mu m^2$ device indicates the trapping of ~2000 electrons. For the emission kinetics of our C1 and C2 samples we observe the trapping of an additional 25pC of charge when the pump pulse is raised above the bias threshold. This is about the same density of traps observed in the RTS experiments (viz. $2 \cdot 10^{11}$traps/cm^2), and tempts us to draw analogy to the trap interaction model of the RTSs. However, it is counter-intuitive that $1 \cdot 10^9$ electrons (traps) could be interacting over distances of hundreds of microns at room temperature! It is however possible that the non-probed traps are interacting locally with the probed traps. If we assume that the traps are located on a plane and distributed uniformly on a two dimensional cubic lattice we can determine the inter-trap distances. The distances between the probed traps would be ~65Å and the distances between the non-probed traps is greater than 225Å. Considering distances between the randomly distributed SiO_2 tetrahedra are on the order of 5Å these distances seem large to support a lattice deformation, trap interaction model. However, if the traps were located in clusters near point and/or line defects in the substrate[26], these distances would be significantly reduced. Our observations indicate a trap interaction model. Although not quantitative, we believe that the occupancy of the non-probed trap distorts the local oxide lattice. The filled probed trap that is located within the lattice distortion will have a higher barrier for charge emission. This is observed in the shifting of the transition time τ_t to longer delay times in the trap emission data.

We are able to remove the trap interactions in our devices by annealing the samples in 10%H_2/N_2 at 400°C for 20mn. Figure 9 shows the results of the emission kinetics for the same sample used in Fig. 7 but after receiving the PMA. The most striking observation is that the emission kinetics are almost independent of the pump pulse width. The maximum in the slope of the $\Delta Q(t_d)$ versus $\log(t_d)$ curve is constant for all t_{pmp} and V_{pmp}. The emission kinetics for the low bias pump approximates a first order decay reaction and has a decay time constant of $\tau = 12.9$s, which is significantly longer than the decay constant measured before the PMA (τ_{NPMA}=83.3ms). The PMA reduced the concentration of the probed traps to levels similarly observed in the A1 samples which had also received a PMA. The decay constant from the A1 and B1 samples is similar to that measured in the PMA C2 samples. When we measure the injected charge trapping as a function of pulse width for the PMA C2 sample, we no longer observe the transition indicative of the extra non-probed traps. The anneal not only removed the non-probed traps (or shifted the threshold voltages much higher) and reduced the concentration of the probed traps, it also altered the emission kinetics of the probed traps. Since filling of the non-probed traps typically increased the decay time of the probed traps, we expected that removal of the non-probed traps with the

PMA should decrease the decay time. This is not what we observe. Therefore, something about the anneal must have altered the traps or their local environment. If the trap interaction/emission model is dependent on lattice deformations, stress at the interface could effect the ability of the traps to reconfigure. A PMA is known to both reduce trap concentrations and interfacial stress in the oxide film[27]. Stress induced strain in semiconducting materials has been shown to shift the energy levels of the conduction and valence bands[28]. The study of pressure dependence on deep trapping centers in Si show a shifting of the traps energy levels[29]. This work also showed that the electron thermal emission rate decreased with increasing compressive strain on the sample.

A more detailed analysis of the identity of the trapping species is needed for a more quantitative model of our results. The location and distribution of the traps in the oxide need to be addressed, if we are to quantify the trap emission mechanisms (i.e. tunneling). The study of the emission kinetics as a function of temperature should provide us with information on the energy distribution of the traps and their capture and emission cross sections. A more detailed study of the PMA processing conditions should elucidate a better understanding of the trap interaction model.

ACKNOWLEDGMENTS

This work was supported in part by the Office of Naval Research. We also appreciate partial funding for one of us (JCP) through a fellowship from the Semiconductor Research Corporation Education Alliance.

1. Y. Nissan-Cohen, J. Shappir and D. Frohman-Bentchkowsy, J. Appl. Phys. **60**, 2024 (1986).

2. S.M. Sze, Physics of Semiconductor Devices, (Wiley, New York, 1969), pp. 444-66.

3. E.H. Nicollian, A. Goetzberger and C.N. Berglund, Appl. Phys. Lett. **15**, 174 (1969).

4. R. Williams, Phys. Rev. **140**, A569 (1965).

5. M. Lenzlinger and E.H. Snow, J. Appl. Phys. **40**, 278 (1969).

6. L.D. Yau and C.T. Sah, Phys. Stat. Solidi A **6**, 561 (1971).

7. C.T. Sah, L.L. Rosier and L. Forbes, Appl. Phys. Lett. **15**, 161 (1969).

8. S.E. Thompson and T. Nishida, J. Appl. Phys. **70**, 6864 (1991).

9. G.L. Miller, D.V. Lang and L.C. Kimerling, Ann. Rev. of Materials Sci. **7**, edited by R.A. Huggins, (Annual Reviews, Palo Alto, 1977), pp. 377-448.

10. G. Groeseneken, H.E. Maes, N. Beltran and R.F. DeKeersmaecker, IEEE Trans. Elect. Devices **ED-31**, 42, (1984).

11. Creation and Detection of the Excited State, edited by A.A. Lamola (Marcel Dekker, New York, 1971) vol. **1**.

12. D.J. DiMaria, The Physics of SiO_2 and its Interfaces, edited by S.T. Pantelides, (Pergomon, New York, 1978), pp. 160-78.

13. W. Kern and D.A. Puotinen, RCA Rev. **31**, 187 (1970).

14. J.C. Poler ... RSI paper

15. E.C. Ross and J.T. Wallmark, RCA Review **30**, 366 (1969).

16. P.W. Atkins, Physical Chemistry, 3rd ed. (W.H. Freeman, New York, 1986), p. 696.

17. R. Stratton, J. Phys. Chem. Solids **23**, 1177 (1962).

18. G.S. Higashi, R.S. Becker, Y.J. Chabal and A.J. Becker, Appl. Phys. Lett. **58**, 1656 (1991).

19. E.H. Nicollian and J.R. Brews, MOS Physics and Technology, (Wiley, New York, 1982), p. 342.

20. M.J. Kirton, M.J. Uren, S. Collins, M. Schultz. A. Karmann and K. Scheffer, Semicond. Sci. Technol. **4**, 1116 (1989).

21. K.S. Ralls, W.J. Skocpol, L.D. Jackel, R.E. Howard, L.A. Fetter, R.W. Epworth, D.M. Tennant, Phys. Rev. Lett. **52**, 228 (1984).

22. K.R. Farmer, C.T. Rogers and R.A. Buhrman, Phys. Rev. Lett. **58**, 2255 (1987).

23. F.W. Schmidlin, J. Appl. Phys. **37**, 2823 (1966).

24. M.J. Kirton and M.J. Uren, Advances in Physics **38**, pp. 433-36.

25. K.R. Farmer and R.A. Buhrman, Semicond. Sci. Technol. **4**, 1084 (1989).

26. G.E. Pike, Phys. Rev. B **30**, 795 (1984).

27. E.A. Irene, Phil. Mag. B **55**, 131 (1987).

28. T.P. Pearsall, CRC critical reviews in solid state and material sciences **15(6)**, 551 (1989).

29. G.A. Samara, Phys. Rev. B **39**, 764 (1989).

Figure 1: Pump-Pulse emission kinetics acquisition. The pulse sequence (a) is controlled by the timing circuit (b).

Figure 2: Charge per pulse versus pulse width for a 50Å IBM1 (●) and a <100>LD (■) MOS capacitors. Injected Charge is logarithmically dependent on the injection time. V=3V and equilibrium time is 30s(10s).

Figure 3: Charge per pulse as a function of pulse height injected into a <111>LD MOS capacitor. The linear dependence on V_{prb} is indicative of trap filling across the oxide's forbidden band.

Figure 4: Injected Charge per pulse versus pulse width at a constant injection bias. Results are from same device shown in Fig. 3. Charge is logarithmically dependent on pulse width.

Figure 5: Full probed traps after a 1.0s pump pulse vs t_d. The emission kinetics (dashed) are dependent on V_{pmp}. Exponential fits (solid) are shown for 0.5, 1.0 and 3.0V pumps.

Figure 6: Full probed traps after a 3.0V pump pulse vs t_d. The emission kinetics (solid) are dependent on t_{pmp}, as seen in the derivative of the data (dashed).

Figure 7: Full probed traps vs t_d. The emission kinetics (solid) are dependent on both t_{pmp} and V_{pmp}. The pump width does not effect the emission kinetics for V_{pmp}<2V.

Figure 8: Normalized trap occupancy versus t_d for <111>LD (●), <111>HD (■) and IBM1 (▲) 50Å MOS capacitors. Emission kinetics resulting from a 1.0s and a 0.05s pump pulse width.

Figure 9: Trap emission kinetics for a <111>LD sample after PMA. No pump pulse dependence on emission kinetics is observed.

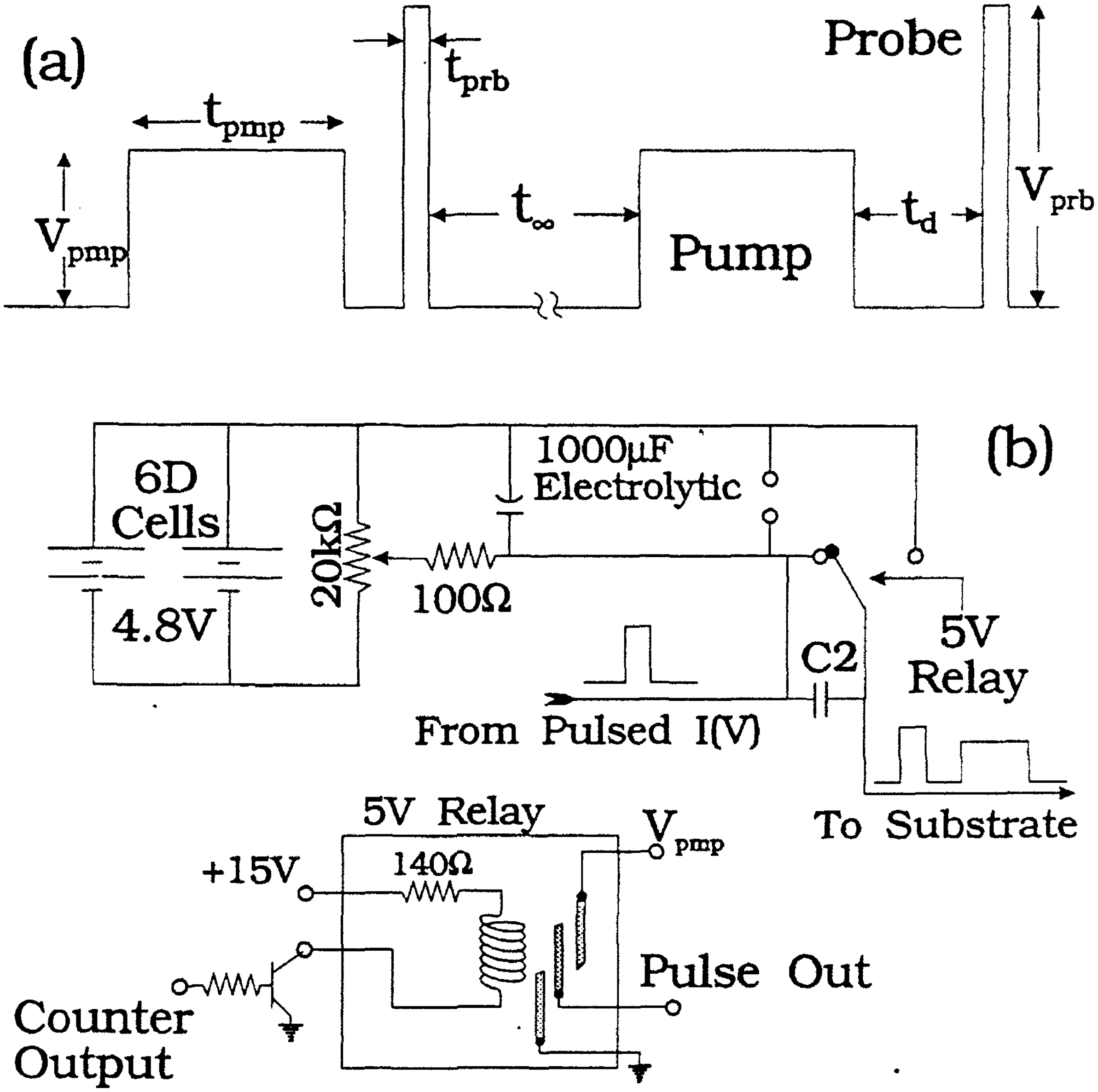

Fig 1 Jordan Poler

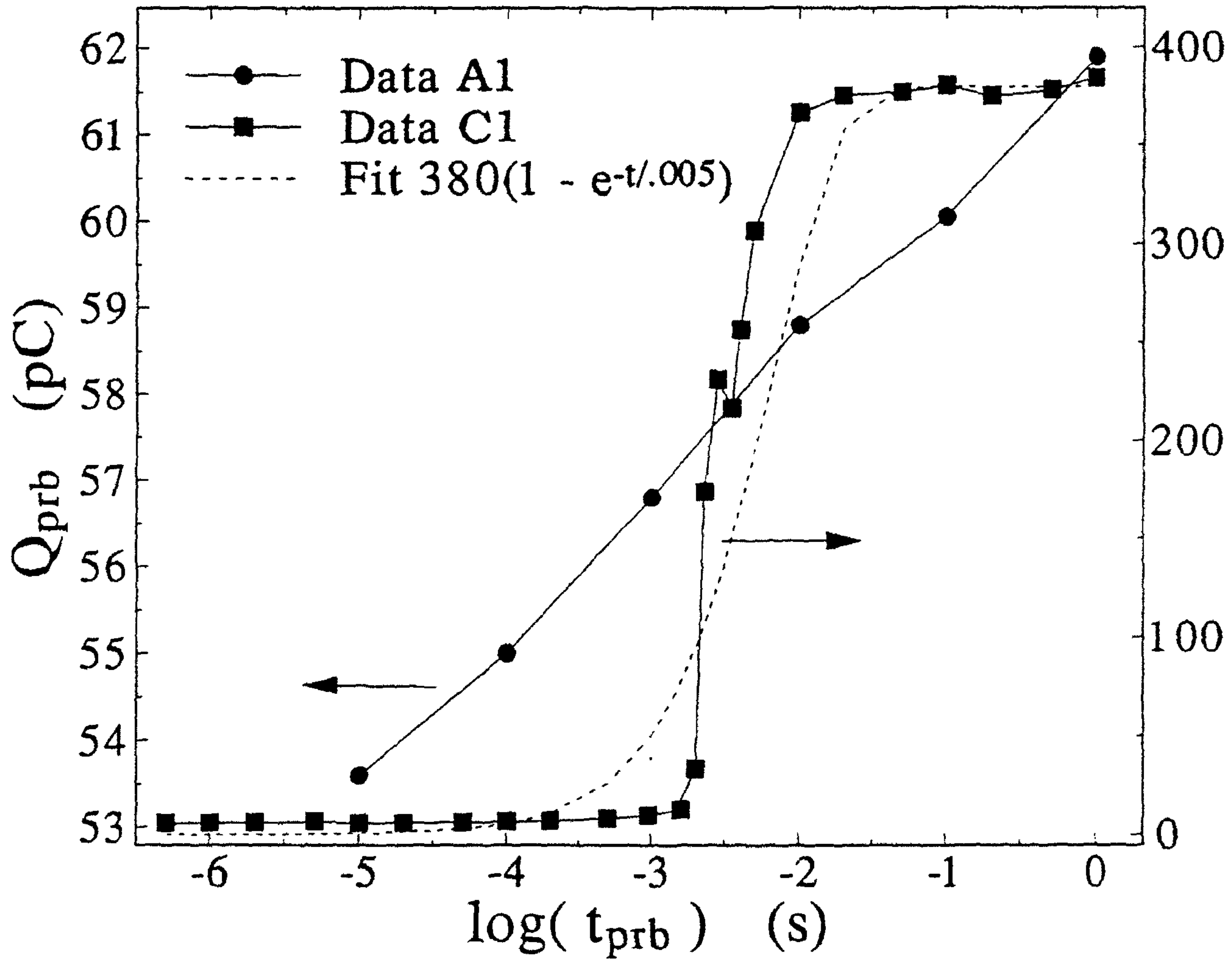

Fig 2 Jordan Poler

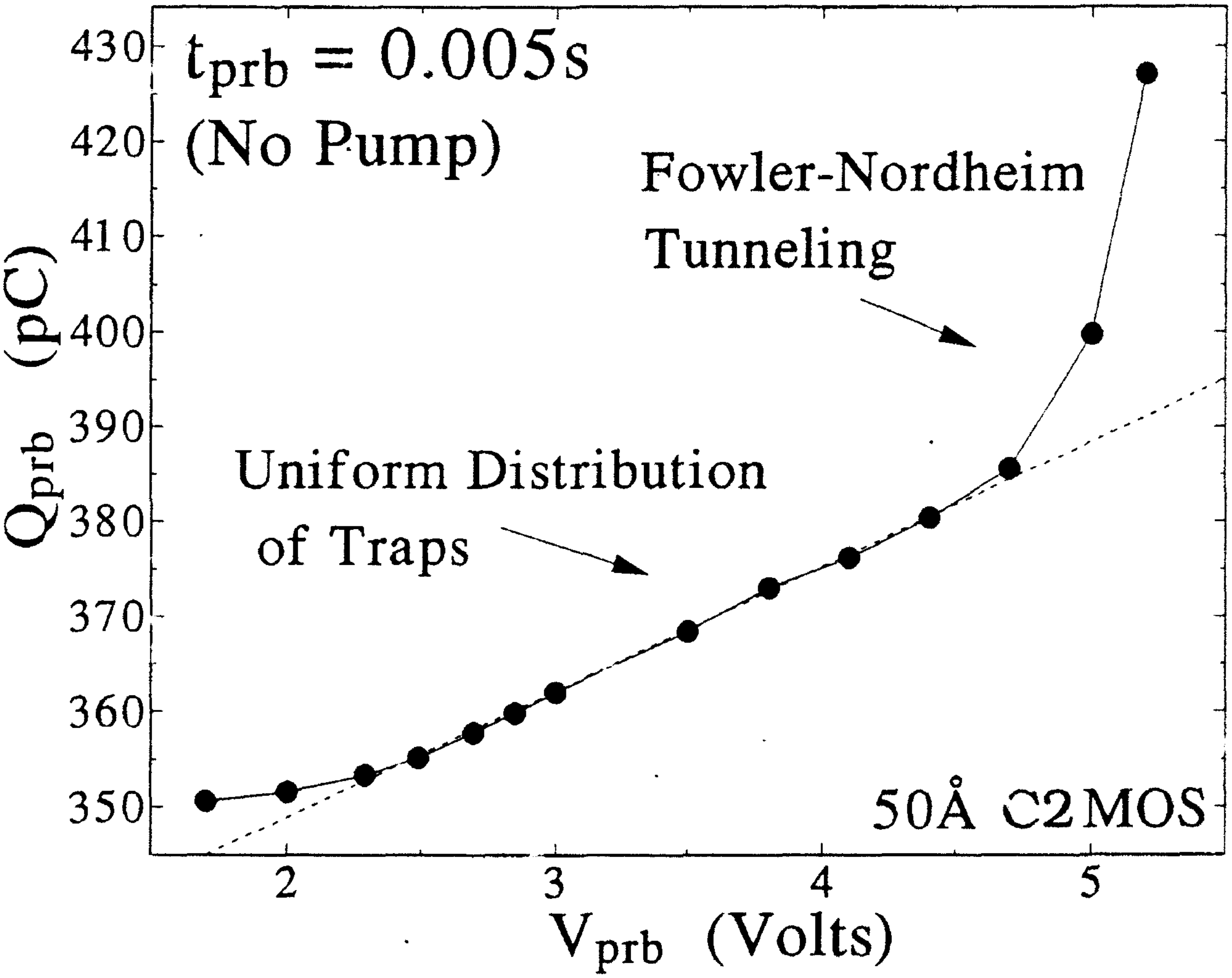

Fig 3 Jordan Poler

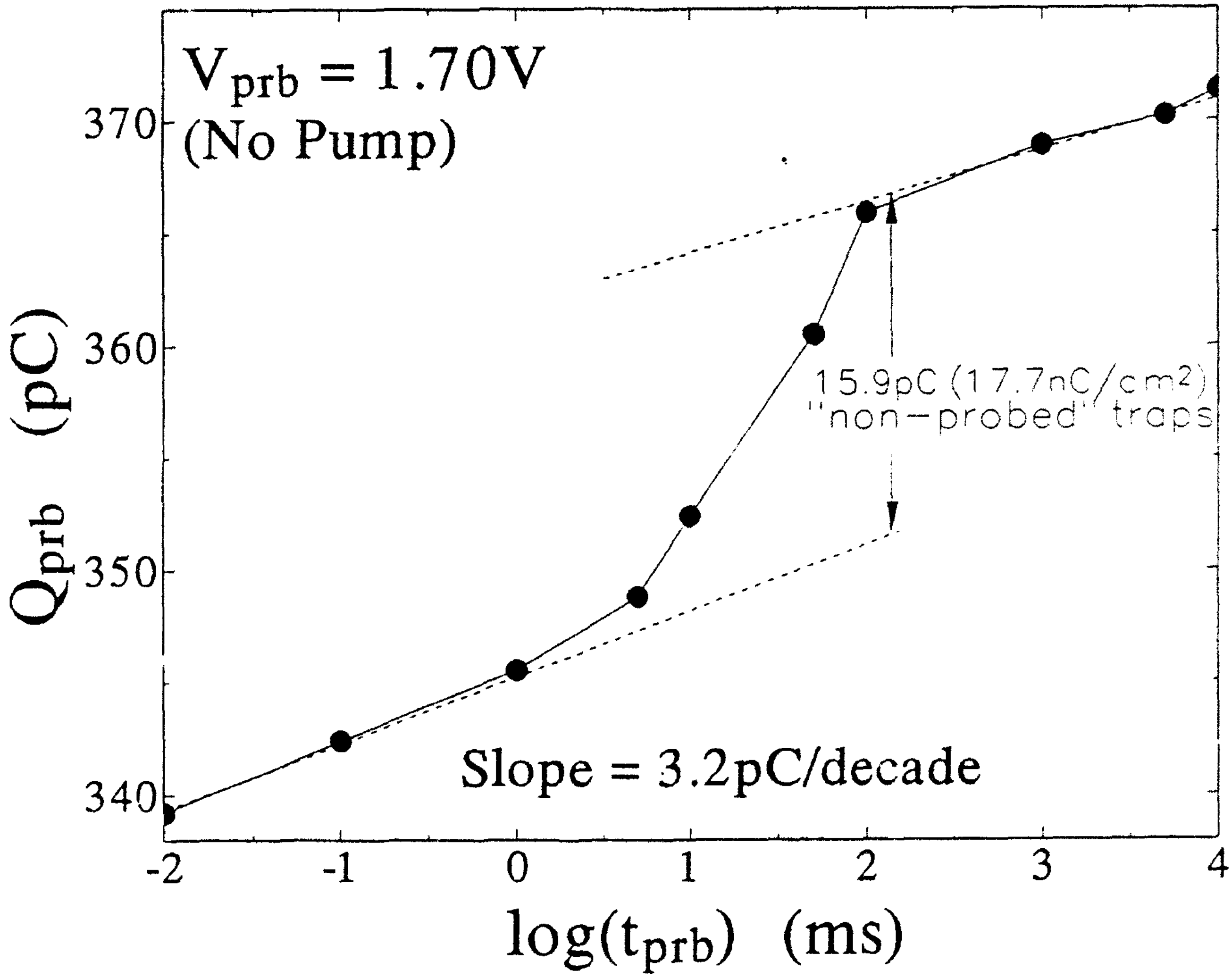

Fig 4 Jordan Poler

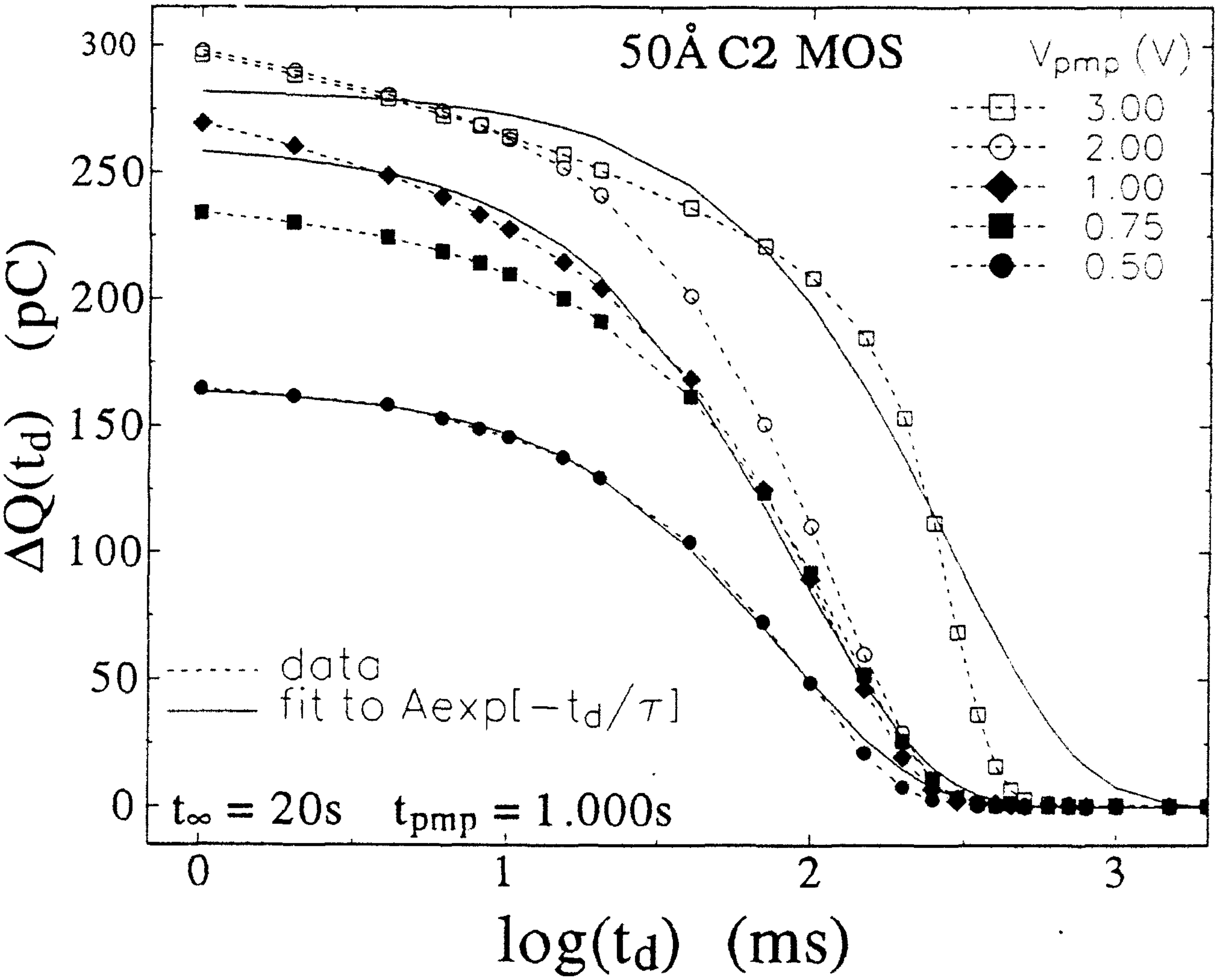

Fig 5 Jordan Poler

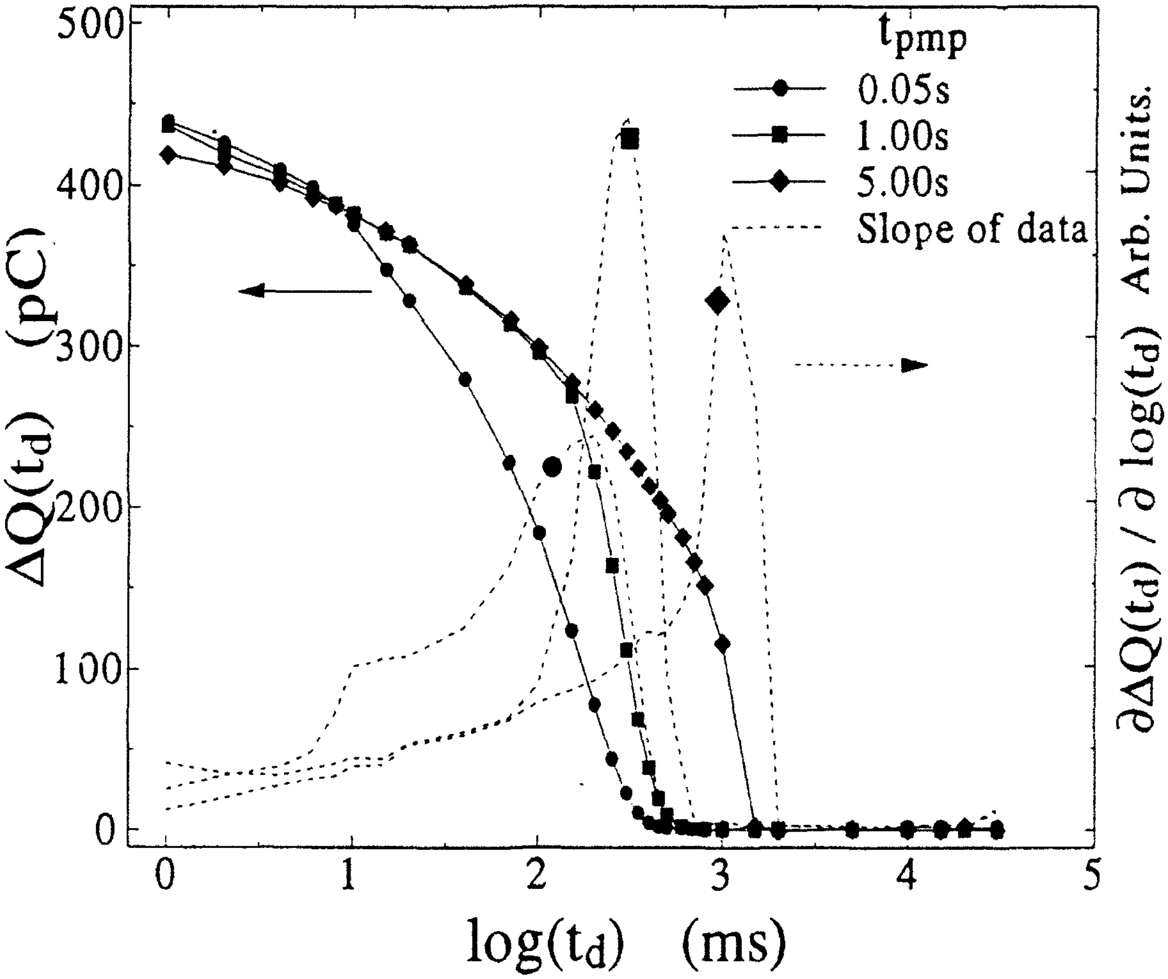

Fig 6 Jordan Poler

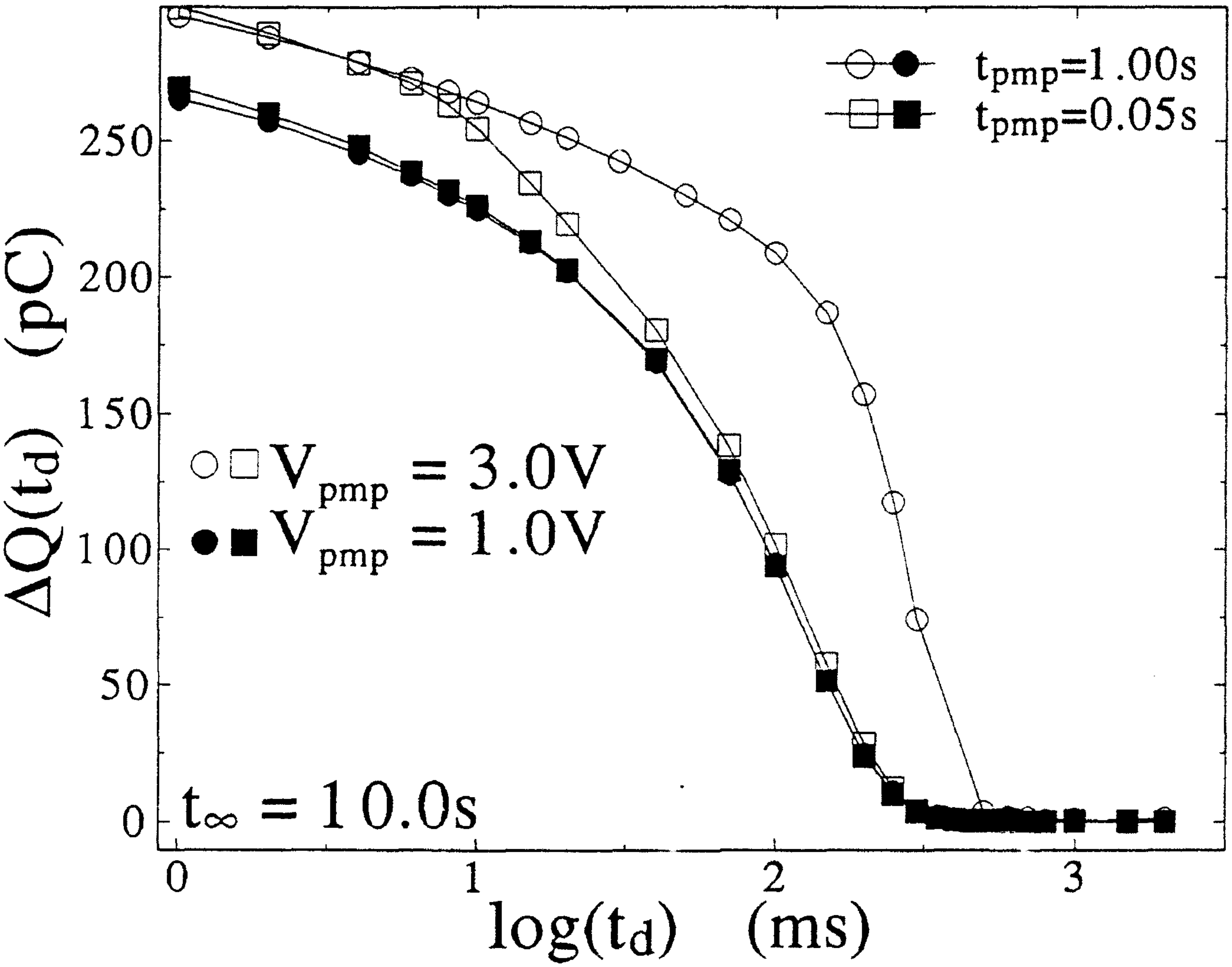

Fig 7 Jordan Poler

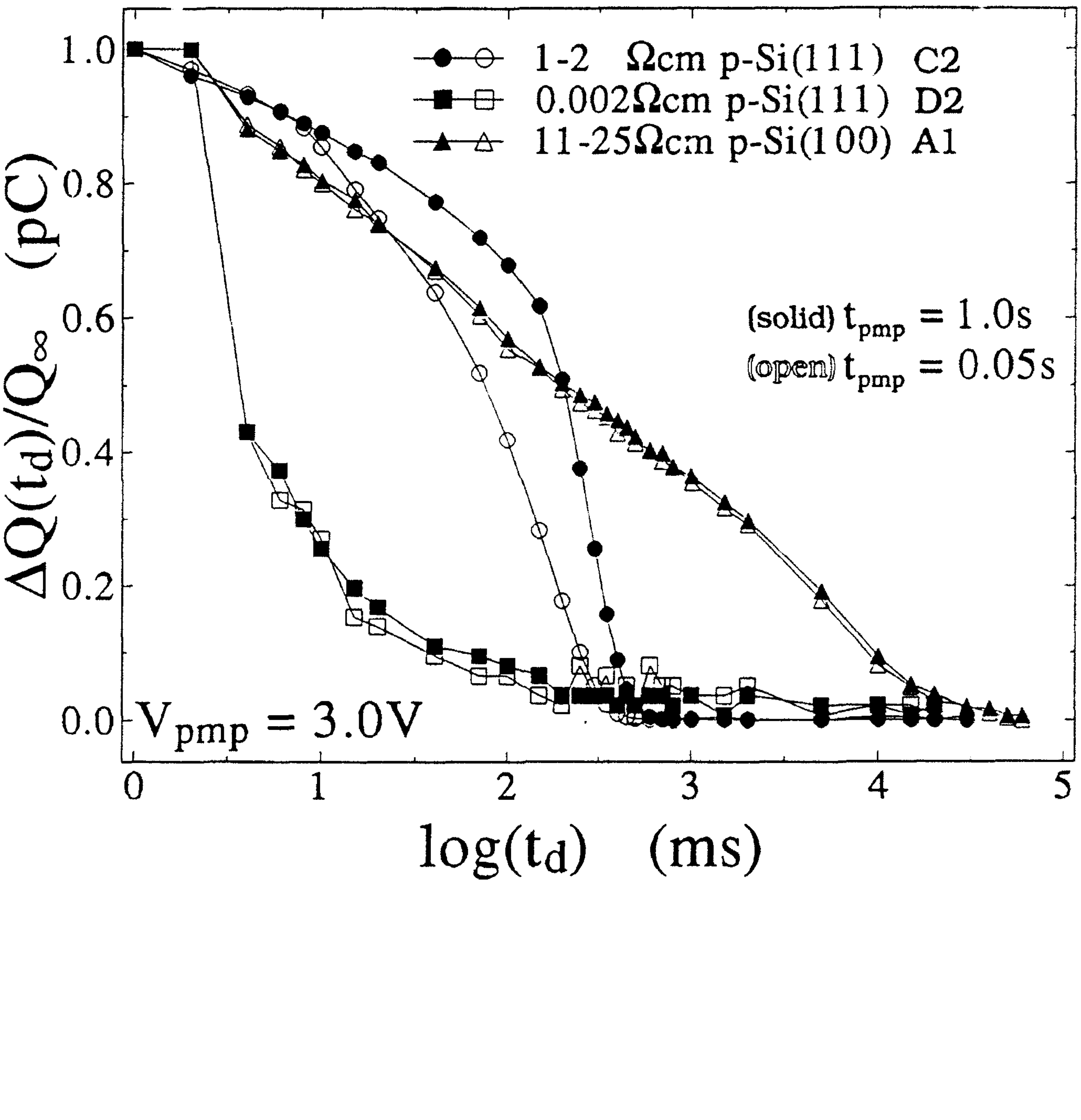

Fig 8 Jordan Poler

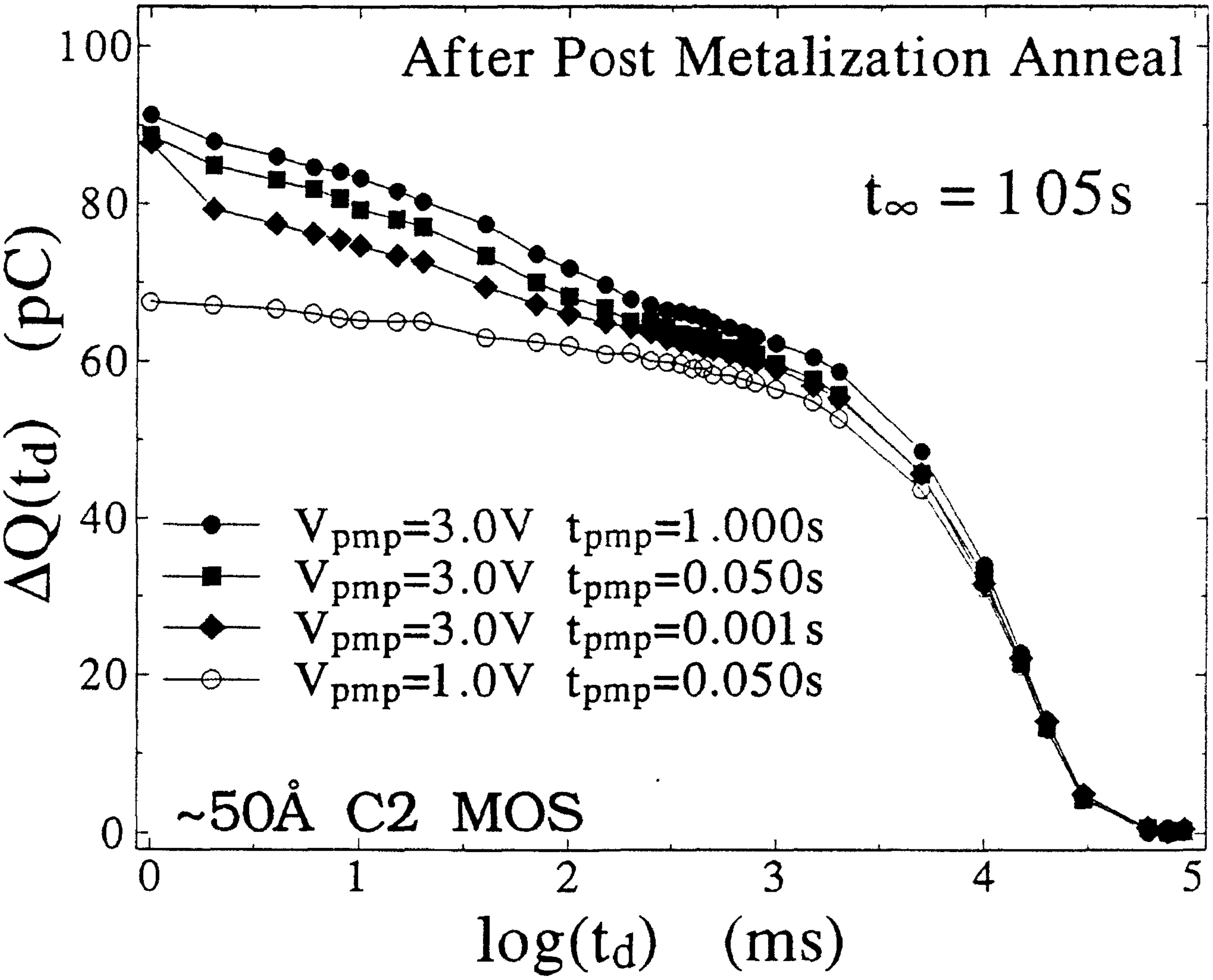

Fig 9 Jordan Poler